



# ANÁLISE PREDITIVA: O QUE É?

Prof. José Gonçalves Vicente



# Sumário

## Sumário

01 Introdução	4
02 Passos de uma Análise Preditiva	5
03 Questionário estruturado	5
04 Análise Estatística	7
<b>IBM SPSS Statistics for Windows</b>	<b>10</b>
01 Capítulo 1 : Introdução	11
02 Capítulo 2: Criar um banco de dados	19
03 Capítulo 3: Transformação de dados	37
04 Capítulo 4: Comandos do IBM SPSS	50
05 Capítulo 5: Manipulação de dados	62
06 Arquivos e janelas	83
07 União de dois bancos de dados	93
08 Gráficos	101

# 01 Introdução

Prever o futuro é um desejo comum entre as pessoas. Se fôssemos capazes de saber, com antecedência de seis meses ou mais, que uma grande crise econômica iria assolar o Brasil (considerando que não estamos vivendo um bom momento neste ano de 2018), o que poderíamos fazer? Será que conseguiríamos criar um plano para prevenir ou diminuir o impacto daquele grande problema? Poderíamos mudar o rumo da história? Com a ajuda de análises preditivas, poderíamos dizer que sim.

Análise preditiva não é “bola de cristal”, mas sim o trabalho de analisar um cenário específico e traçar possíveis tendências e mudanças capazes de afetar um planejamento estratégico.

Com o avanço de tecnologias que possibilitam uma mineração de dados eficiente, a análise preditiva conta, cada vez mais, com segurança e consolidação para descobrir padrões e avaliar as probabilidades de um resultado ou um acontecimento futuro, diferentemente da simples análise descritiva de dados, que apenas mede e apresenta resultados passados.

Um exemplo prático e didático foi o que ocorreu em 1984/1985 com relação às “diretas já”. Uma equipe técnica constituída por uma Instituição Federal, da qual fomos chamadas a fazer parte, para uma análise da população brasileira com relação ao movimento nas ruas pelas “diretas já”. Combinando algoritmos, cientistas de dados, e uma equipe multidisciplinar, o grupo foi capaz de analisar a movimentação nas ruas (não existiam ainda internet e redes sociais como nós as temos hoje), para prever quais movimentações sociais ficariam cheias; quais teriam um contexto mais violento e quais seriam pacíficas.

Com esse insumo analítico, todas as ações da Unidade Federal durante o evento deixaram de ser intuitivas. Cada passo era extremamente analítico e as chances de sucesso foram inimaginavelmente maiores.

Em trabalhos que fazem uso de técnicas de análise preditiva, o intercruzamento de fórmulas matemáticas e estatísticas possibilita a descoberta de padrões ocultos nos dados que especialistas humanos dificilmente enxergariam.

UMA ANALOGIA: Vamos imaginar que você pretenda viajar de Maringá até Londrina, dirigindo o seu carro. O primeiro passo ao chegar em seu carro é abrir a porta e, em seguida, entrar no carro: (1) coloca o cinto de segurança, (2) coloca em ponto morto, (3) liga o carro, (4) verifica o painel de controle do carro, (5) pressão dos pneus, (6) combustível, (7) se ajusta perfeitamente ao banco, (8) destrava a alavanca dos freios, (9) acende a luz baixa do veículo, (10) engata as marchas. Vai começar a viagem. A partir de agora é futuro: (11) Olha nos espelhos retrovisores, (12) verifica o limite de velocidade, (13) sinais nas placas de trânsito, (14) semáforos, (15) ultrapassagem proibida, (16) sinais de alerta nas estradas, (17) permitir ultrapassagem de outros veículos mais rápidos, (18) movimento dos pedestres etc.

O objetivo é fazer uma viagem confortável e chegar a Londrina feliz e salvo.

Os itens de 1 a 10 se referem a dados existentes e se repetem sempre que você vai viajar, ou seja, você sempre os tem. Os dados de 11 a 18 são dados que vão ocorrer no futuro em sua viagem. O sucesso de sua viagem depende sempre dos dados já coletados e conferidos e dos dados preditivos a serem alcançados.

Análise Preditiva “é um conjunto de técnicas matemáticas/estatísticas aplicado a um conjunto de dados para determinar a probabilidade de que algum cenário é provável que aconteça ou seja verdadeiro” (MATIS, 2006).

## 02 Passos de uma Análise Preditiva

Conheça bem o seu passado para ter uma melhor visão do que vai acontecer no futuro.

Etapa 1:

Extração de dados	Limpeza de dados	Identificar padrões	Fazer previsões
Primários, secundários	Analisar	Conhecimento	Previsões

Etapa 2:

Informação Analítica		Análise avançada		Decisão
Descritiva	Diagnóstica	Preditiva	Prescritiva	Cognitiva
O que está acontecendo?	Por que ocorreu?	O que provavelmente vai ocorrer?	Como podemos garantir que realmente aconteça?	De acordo com o que aprendemos, qual a melhor decisão?

Dados → informações → conhecimento → ideias → sabedoria → ação

## 03 Questionário estruturado

Dados

Neste curso, iremos utilizar dados obtidos de um questionário estruturado exclusivamente para fins didáticos. A coleta de dados foi realizada junto a professores e alunos da UniCesumar entre os dias 21 e 29 de março de 2018 e os dados envolvem itens demográficos, de atitude e comportamento. Dentre esses, estão opiniões sobre satisfação em relação a diversos aspectos sobre o cotidiano. Cerca de 500 adultos foram incluídos nessa simulação. Não fica afastada a proposta de, junto aos alunos desse curso, produzir um questionário estruturado de situações reais que se apresenta como sem respostas satisfatórias, como por exemplo: Por que as mulheres ganham salários inferiores aos homens nas mesmas funções?

### Questionário Estruturado – Pesquisa Social Geral

01 - Sexo do entrevistado:

1 [ ] masculino      2 [ ] feminino

02 - Estado civil do entrevistado:

1 [ ] casado      2 [ ] solteiro      3 [ ] união estável      4 [ ] outra

03 - Data de nascimento: \_\_\_\_/\_\_\_\_/\_\_\_\_

04 - Peso do entrevistado em quilogramas: \_\_\_\_\_

05 - Altura do entrevistado em metros: \_\_\_\_\_

06 - Grau de Escolaridade da mãe:

1 [ ] 1º grau completo/incompleto      2 [ ] 2º grau incompleto  
3 [ ] 2º grau completo      4 [ ] superior      5 [ ] Não se aplica

07 - Renda média familiar: R\$ \_\_\_\_\_

08 - Religião do entrevistado:

1 [ ] católico      2 [ ] evangélico      3 [ ] outra      4 [ ] não tem

09 – Você acredita na vida após a morte:

1 [ ] sim                      2 [ ] não                      3 [ ] não sei

10 – Qual a sua situação de felicidade de modo geral:

1 [ ] ótima                      2 [ ] boa                      3 [ ] regular

4 [ ] ruim                      5 [ ] péssima

11 – Quais as suas condições de saúde:

1 [ ] ótima                      2 [ ] boa                      3 [ ] regular

4 [ ] ruim                      5 [ ] péssima

12 – Na sua opinião hoje, quanto o governo deve fazer para reduzir a distância entre ricos e pobres?

1 [ ] muito menos                      2 [ ] um pouco menos                      3 [ ] o mesmo que já tem feito

4 [ ] um pouco mais                      5 [ ] muito mais                      6 [ ] não sei

13 – “Quando um negócio ou empresa ganha muito dinheiro, todos se beneficiam, inclusive os pobres”. A respeito dessa frase você:

1 [ ] discorda totalmente                      2 [ ] discorda parcialmente                      3 [ ] fica neutro

4 [ ] um pouco mais                      5 [ ] muito mais                      6 [ ] não sei

14 – Na sua opinião, qual o número de filhos ideal numa família: \_\_\_\_\_ filhos

15 – Você é favorável ao armamento dos cidadãos brasileiros que quiserem?

1 [ ] sim                      2 [ ] não                      3 [ ] não sei

16 – O quanto você valoriza a honestidade das decisões do congresso nacional brasileiro?

1 [ ] nada                      2 [ ] pouco                      3 [ ] muito

17 – Quantas horas você gasta, em média, na internet por dia? Resp. \_\_\_\_\_ horas

18 – Você oferece trabalho voluntário ou contribui financeiramente para organizações de caridade?

1 [ ] sim                      2 [ ] não                      3 [ ] às vezes

19 – Qual a sua opinião sobre o casamento ou união estável da comunidade LGBT?

1 [ ] favorável                      2 [ ] não favorável                      3 [ ] não sei

20 – Você acredita na existência do diabo?

1 [ ] sim                      2 [ ] não                      3 [ ] não sei

21 – Você acha que a maconha deveria ser legalizada?

1 [ ] sim                      2 [ ] não                      3 [ ] não sei

22 – Você tem o hábito de acompanhar, pela imprensa nacional (TV, Radio, Jornal, revistas), as notícias nacionais e internacionais?

1 [ ] sim                      2 [ ] não                      3 [ ] às vezes

23 – Quanto ao voto nas eleições presidenciais, você se considera um eleitor de:

1 [ ] direita                      2 [ ] centro                      3 [ ] esquerda                      4 [ ] não sabe

## 04 Análise Estatística

Vivemos constantemente rodeados por potenciais fontes de dados que, entre tantas, podem ser: *funcionários de uma empresa* (se estivermos interessados em avaliar seu desempenho antes e após um treinamento), *pacientes de uma clínica* (se quisermos estudar tempo de reação a uma determinada droga), *objetos em uma linha de produção* (se quisermos avaliar quantidades de peças defeituosas produzidas), *índices da bolsa de valores* (se nosso alvo são projeções econômicas) ou, até mesmo, rendimento das mulheres (se quisermos comparar salários de homens e mulheres nas mesmas funções).

Por hora, vamos supor que estamos interessados em estimar o peso médio dos brasileiros do sexo feminino e com mais de 25 anos de idade. Em geral, não temos recursos financeiros ou mesmo tempo suficiente para entrevistarmos cada pessoa e pesar cada uma delas, por isso, é preciso selecionar um **grupo** (ou uma **amostra**) de pessoas. Essa etapa é chamada de **coleta de dados**, que é seguida do devido **armazenamento** deles.

Uma vez reunidas todas as informações necessárias para a análise, esses dados precisam passar, frequentemente, por alguma transformação. Se a altura das pessoas tivesse sido registrada em gramas, seria conveniente transformá-la para quilogramas. Esta é a etapa de **transformação de dados**, que poderia incluir cálculos bem mais complexos.

Finalmente, os dados estão prontos para a análise, que será dividida em duas fases. A primeira é chamada de **análise descritiva**, na qual - como o próprio nome diz - descreveremos os dados. Alguns procedimentos utilizados nessa etapa são: *construção de tabelas e gráficos*, cálculo de medidas como *média, mediana, desvio padrão etc.*

No estudo sobre o peso dos brasileiros, por exemplo, poderíamos ter como primeiros resultados: o peso médio na amostra entrevistada é de 68 kg com um desvio padrão de 3 kg cm. Mas daí a dizer que este é o peso médio do brasileiro do sexo feminino com mais do que 25 anos é um passo grande e que deve ser dado com muito cuidado. É então que entra a segunda parte da análise, a **análise inferencial**: com a ajuda de testes estatísticos, poderemos mais do que simplesmente estimar qual o peso médio do grupo em questão: poderemos, também, associar uma certa *confiabilidade* a tal conclusão.

Depois disso, é hora de relatar as conclusões do estudo. Agora, toda informação - seja ela numérica ou gráfica - é bem-vinda, para que alguém que não participou do processo de análise possa compreender as conclusões sobre o estudo por meio de **relatórios e gráficos**.

Mas, afinal, por que devemos realizar uma análise estatística? Simplesmente porque precisamos da confiabilidade que essa análise oferece em um momento de tomada de decisão.

O **IBM SPSS Statistics** é considerado um dos melhores *softwares* estatísticos do mercado, justamente porque participa de grande parte do processo que engloba uma análise estatística. Logo na primeira etapa, é possível armazenar dados de forma rápida e prática no próprio IBM SPSS Statistics. Durante o processo de transformação e análise de dados, o IBM SPSS Statistics tem recursos considerados dos mais poderosos. Até mesmo em um dos momentos mais delicados da análise, a elaboração de relatórios, o IBM SPSS Statistics tem recursos excelentes, mas, e na hora de tomar decisões? Mesmo nesse momento crítico da sua análise, você pode contar com os recursos do IBM SPSS Statistics.

### Variáveis

Qualitativas		Quantitativas	
Ordinais	Nominais	Discretas	Contínuas

As variáveis nos estudos estatísticos são os valores que assumem determinadas características dentro de uma pesquisa e podem ser classificadas em **qualitativas ou quantitativas**.

### Variáveis Qualitativas

As **variáveis qualitativas** não podem ser expressas numericamente, pois relacionam situações como a cor da pele, cor dos olhos, marca de refrigerante, marca de automóvel, preferência musical entre outras. Elas podem ser divididas em *ordinais e nominais*. As **variáveis qualitativas ordinais**, apesar de não serem numéricas, obedecem a uma relação de ordem, por exemplo: conceitos como ótimo, bom, regular e ruim, classe social, grau de instrução, etc. Já as **variáveis qualitativas nominais** não estão relacionadas à ordem, elas são identificadas apenas por nomes, por exemplo, as cores: vermelho, amarelo, preto, azul, rosa, verde, etc. Também como exemplo de nominais temos as marcas de carros, nome de bebidas, local de nascimento, entre outros.

### Variáveis Quantitativas

No caso das **variáveis quantitativas**, usamos a representação numérica. Elas podem ser classificadas em discretas e contínuas. As **variáveis quantitativas discretas** acontecem relacionadas a situações limitadas, por exemplo: número de revistas vendidas, quantidade de consultas médicas, número de filhos de um casal. No caso das **variáveis quantitativas contínuas**, a abrangência pertence a um intervalo que se caracteriza por infinitos valores, como exemplo podemos citar: o peso de um produto, altura dos alunos de uma escola, velocidade de objetos, entre outras situações.

Um exemplo prático.

Consideremos uma pesquisa sobre saúde de 332 pessoas com 11 variáveis.

01 - Idade em anos completos.

02 - Gênero:

[ 0 ] = feminino      [ 1 ] = masculino

03 - Altura em cm \_\_\_\_\_.

04 - Peso em kg \_\_\_\_\_.

05 - Fumante:

[ 0 ] não fumante      [ 1 ] fumante

06 - Opinião sobre a própria condição física:

[ 1 ] ruim      [ 2 ] fraca      [ 3 ] regular      [ 4 ] boa      [ 5 ] ótima

07 - Nota final no curso de graduação.

08 - Salário mensal R\$\_\_\_\_\_.

09 - Prática regular de atividades físicas semanais:

[ 1 ] nenhuma      [ 2 ] 1 a 2 vezes      [ 3 ] 3 a 4 vezes      [ 4 ] mais de 5 vezes

10 - Grau de instrução da mãe:

[ 1 ] Analfabeta/1ºgrau incompleto      [ 2 ] 1ºgrau completo  
[ 3 ] 2ºgrau completo      [ 4 ] superior

11 - Número de irmãos: \_\_\_\_\_.

Variáveis	Casos
Ordinais	06, 10
Nominais	02, 05, 09
Discretas	01, 11
Contínuas	03, 04, 07, 08

No caso do **IBM SPSS Statistics** o programa apresenta 9 tipos de variáveis e 3 formas de medidas, que veremos mais adiante.

# IBM SPSS STATISTICS FOR WINDOWS

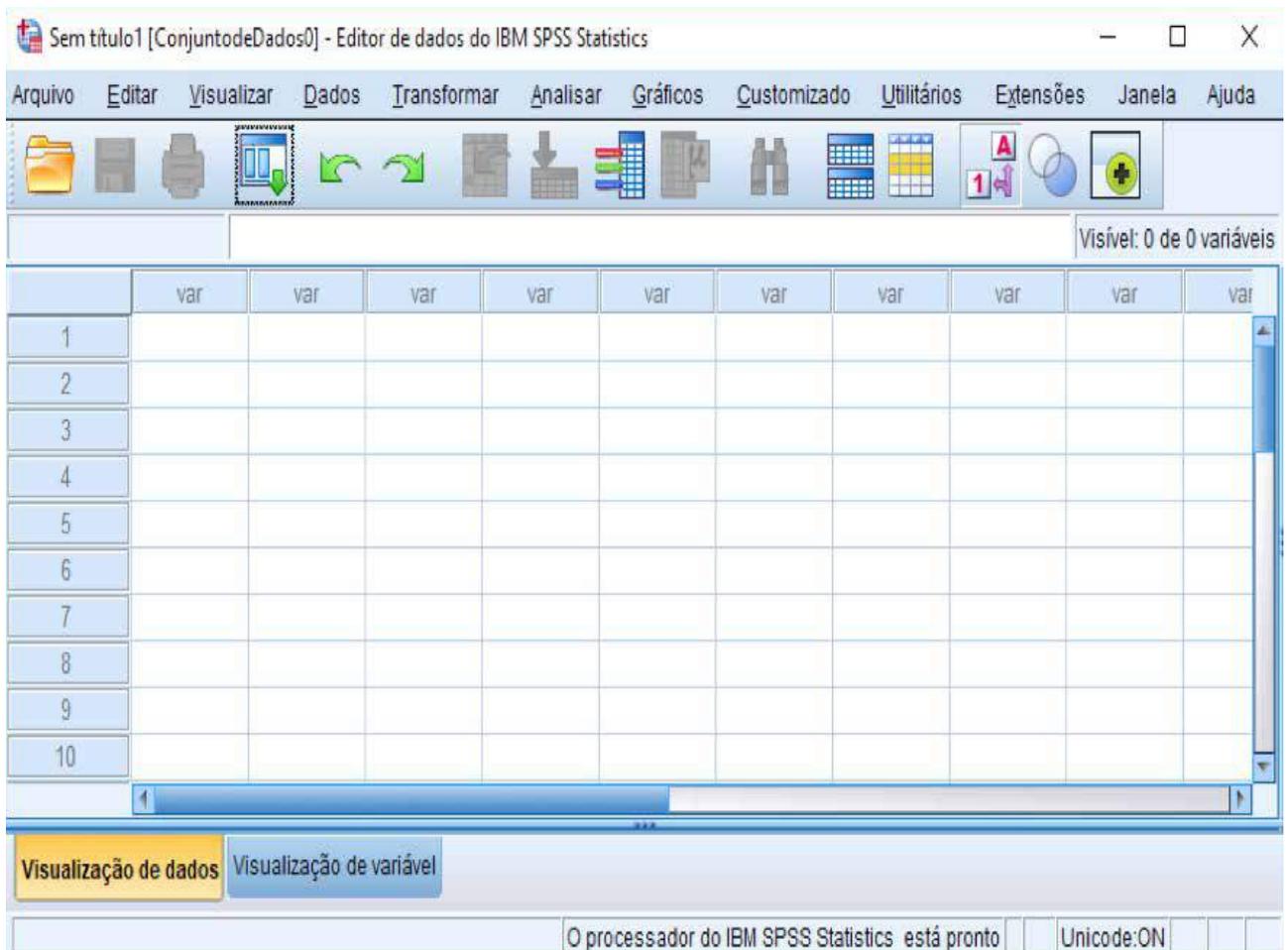
# 01 Introdução

## 1.01 Passos de uma Análise Estatística

### Janela Inicial

Inicialmente, o IBM SPSS Statistics abre uma janela para começarmos a trabalhar com os nossos dados: a janela sem título1 [ConjuntoDeDados0] – Editor de dados do IBM SPSS Statistics (veja Figura 1). Ao longo do trabalho, outras janelas podem ser abertas. Cada janela aberta é um arquivo em potencial, ou seja, pode ser armazenado, impresso, copiado para outras janelas etc.

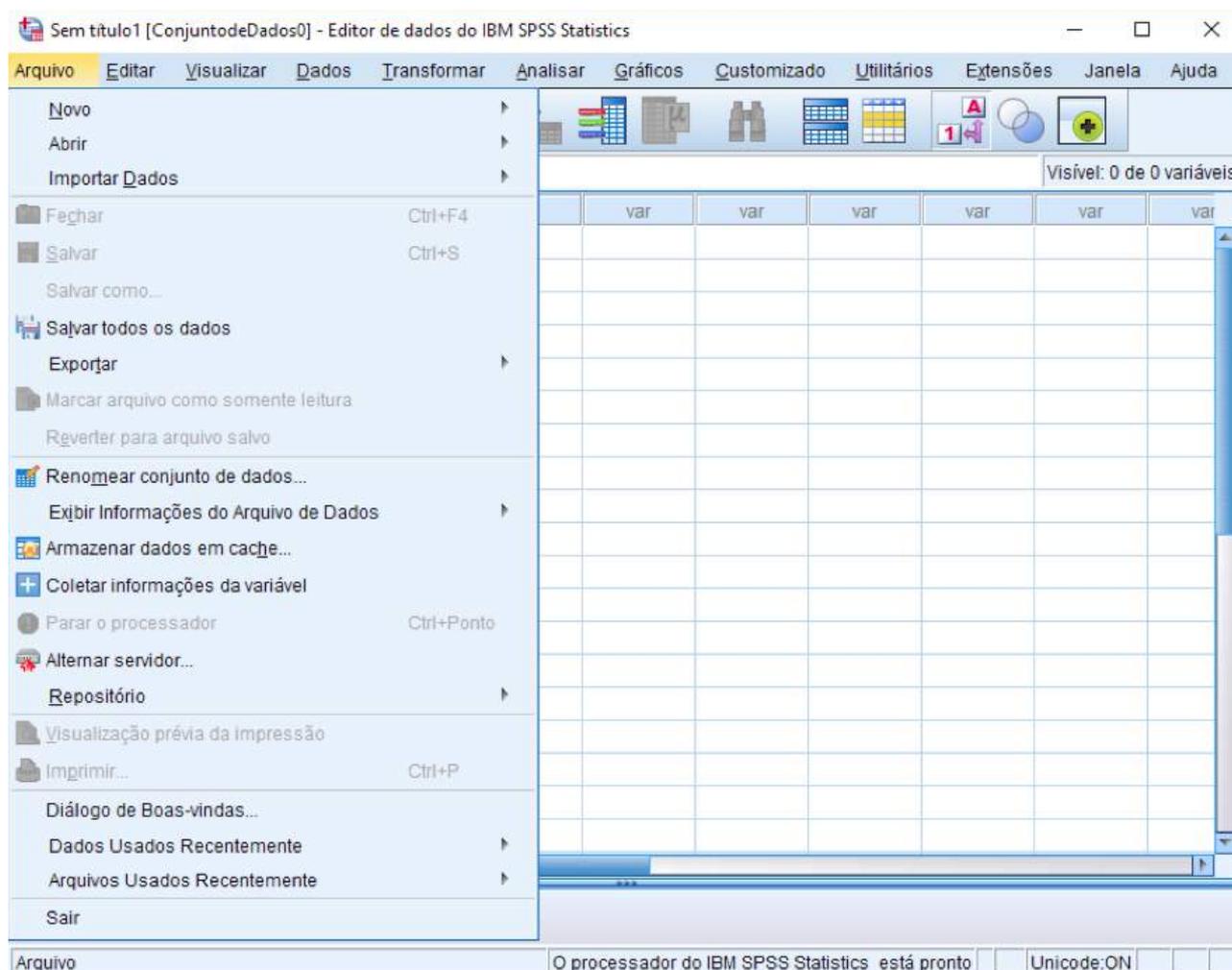
Figura 1 | Aspecto inicial do IBM SPSS Statistics



## 1.02 Arquivo

O menu **Arquivo** (ver Figura 2) permite que você tenha controle sobre tudo que se refere ao gerenciamento de arquivos, como importação, exportação, impressão e salvamento.

Figura 2 | O menu arquivo



O primeiro bloco desse menu destina-se à abertura de arquivos, seja ele um banco de dados, uma janela de saídas produzida anteriormente ou uma janela de sintaxe. Os tipos de bancos de dados que podem ser lidos diretamente pelo IBM SPSS Statistics são os mais diversos, como por exemplo, ASCII (texto), Microsoft Excel, dBase, entre outros e, é claro, os arquivos produzidos pelo próprio IBM SPSS Statistics nas suas diferentes versões. A opção **Novo** abre uma nova janela de dados (que chamamos de dataset ou conjunto de dados) ou outros tipos de janelas do IBM SPSS Statistics (saída, sintaxe ou script); a opção **Abrir** abre um arquivo diretamente; a opção **Importar Dados** permite a abertura de arquivos com formatos específicos – veremos dois exemplos de importação de arquivos (texto e Excel, respectivamente) a partir desta opção no final do Capítulo.

O segundo bloco do menu Arquivo tem por função salvar qualquer janela do IBM SPSS Statistics; assim, se estivermos na janela de saídas, podemos salvar todos os resultados de uma análise estatística utilizando a opção **Salvar**. O software permite ainda salvar um banco de dados criado no IBM SPSS Statistics em outro formato, assim como no próprio formato IBM SPSS Statistics. Você poderá também salvar a base de dados que estiver ativa no IBM SPSS Statistics por meio da opção **Exportar** – veremos esta opção no final do Capítulo II. A opção **Marcar Arquivo Como Somente Leitura** serve para evitar que um arquivo original seja modificado ou deletado

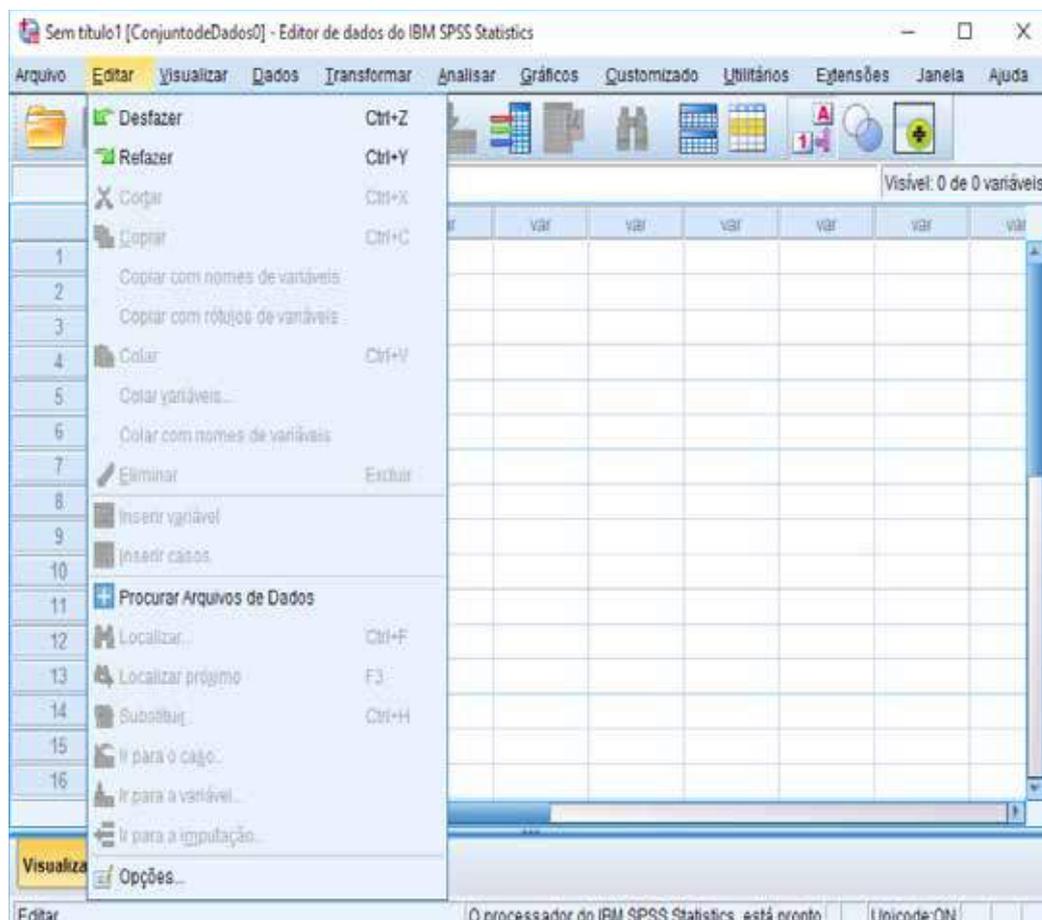
acidentalmente: com essa opção selecionada, o seu arquivo ficará marcado como Somente Leitura. (Caso essa opção esteja selecionada, para salvar o arquivo será necessário escolher outro nome ou entrar novamente no menu Arquivo e escolher a opção Marcar Arquivo Como Somente Leitura.)

No terceiro bloco do menu Arquivo, você pode renomear um dataset usando a opção **Renomear Conjunto de Dados**; obter informações sobre o conteúdo do seu banco de dados usando a opção **Exibir Informações do Arquivo de Dados**; forçar a criação de um arquivo de dados temporário em **Armazenar Dados em Cache**; parar o processador do IBM SPSS Statistics através de **Parar o Processador** ou conectar-se a um servidor para executar análises em bancos de dados maiores (desde que você tenha o IBM SPSS Statistics Server) através do **Alternar Servidor**. No quarto bloco de Arquivo está localizada a opção de impressão (**Imprimir**) de qualquer janela que esteja ativa, e no quinto bloco do menu Arquivo, temos a lista de arquivos recentemente utilizados e a opção Sair para você fechar o IBM SPSS Statistics.

### 1.03 Editar

O menu Editar (ver Figura 3) traz comandos relacionados à edição, que podem ser diferentes de acordo com a janela que estiver ativa. Se estivermos na janela de dados, podemos utilizar os recursos **Cortar e Colar** para cortar e colar, respectivamente, um conjunto de dados de um lugar para outro. Os comandos **Localizar e Ir Para o Caso** ajudam a encontrar um registro no banco de dados. No último bloco do menu Editar, temos o comando **Opções...**, onde podemos definir os principais parâmetros do IBM SPSS Statistics, como o padrão para o tipo da variável (o padrão do IBM SPSS Statistics é Numeric 8.2, isto é, variável numérica com oito dígitos, sendo que os dois últimos são casas decimais), ou ainda o tamanho e fonte dos títulos na janela de saídas, entre muitas outras. Algumas dessas opções do comando Opções... serão estudadas detalhadamente no final do curso.

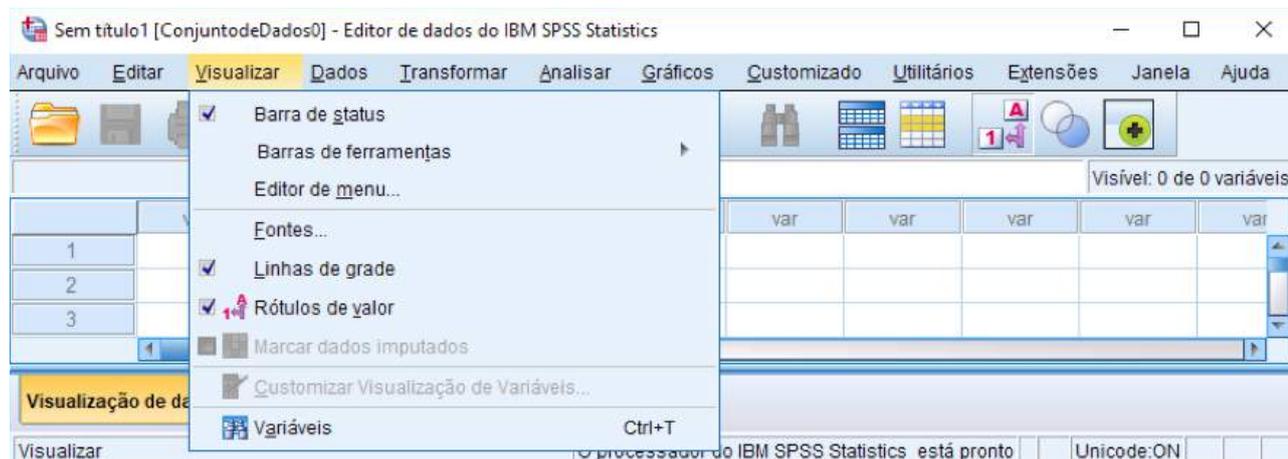
Figura 3: O menu editar



## 1.04 Visualizar

Pelo menu **Visualizar** (ver Figura 4), você escolhe como serão mostrados os dados na janela Editor de Dados (usando as opções **Fontes**, **Linhas de Grade** e **Rótulos de Valor**). Por esse menu, também é possível disponibilizar ou não (pela opção **Barra de Status**) a barra inferior, que contém mensagens importantes sobre o IBM SPSS Statistics, e, finalmente, pela opção **Barra de Ferramentas**, você pode customizar a(s) barra(s) de ferramentas, inclusive, criando uma barra totalmente nova.

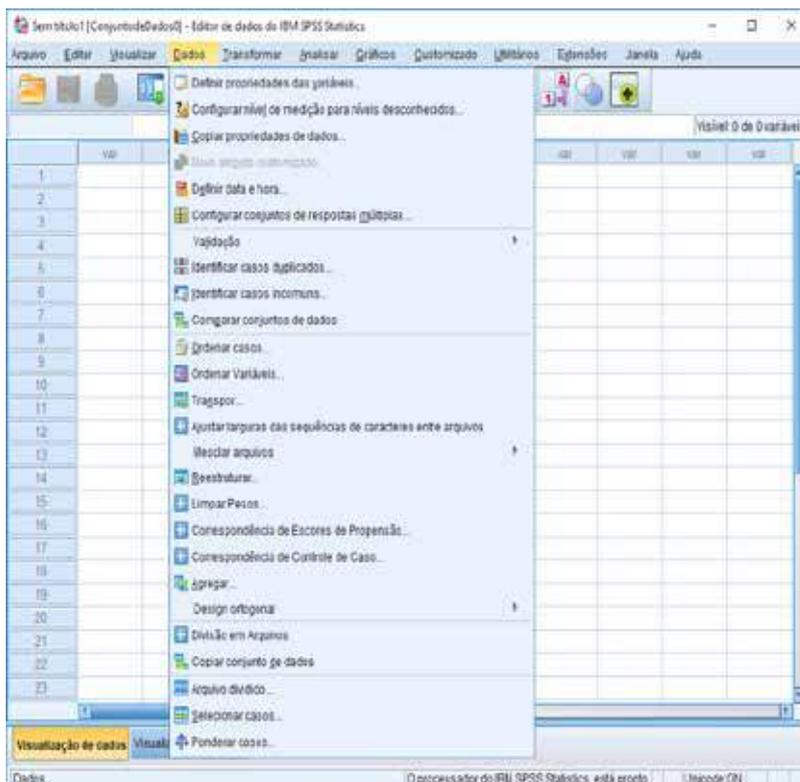
Figura 4: O menu visualizar



## 1.05 Dados

O menu **Dados** (Figura 5) traz funções relacionadas à manipulação dos dados. Comandos como definição de datas, ordenação do banco de dados, junção de arquivos em formato IBM SPSS Statistics, seleção de casos do arquivo de dados para trabalho e ponderação de valores fazem parte das opções desse menu (ver Figura 5).

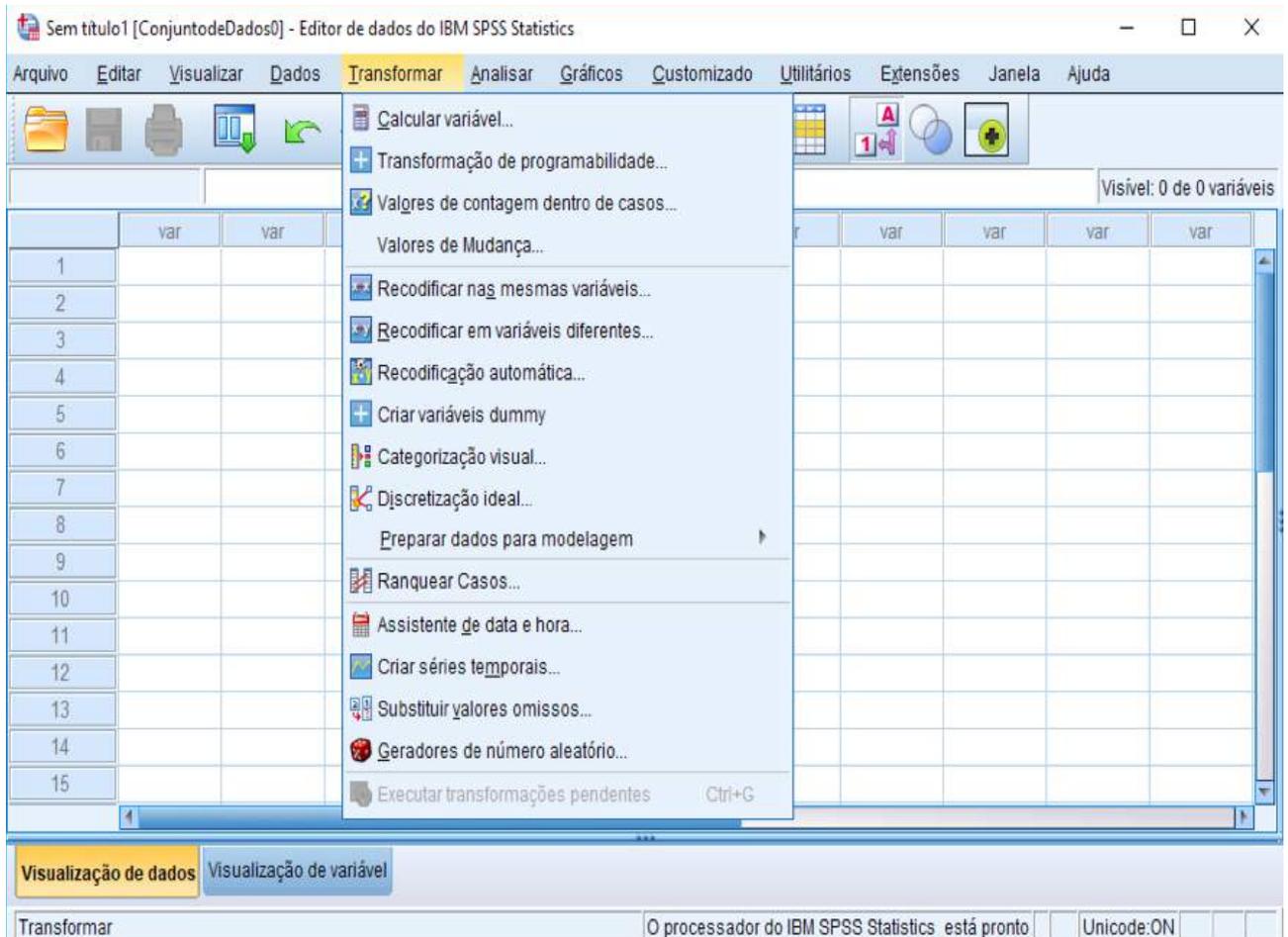
Figura 5 | O menu dados



## 1.06 Transformar

A etapa de transformação dos dados (Figura 6) concentra-se basicamente no menu **Transformar**. Recursos como criação de novas variáveis a partir das já existentes, recodificação de valores do banco de dados e, até mesmo, um assistente (wizard) para auxiliar no trabalho com as variáveis no formato data são algumas das opções desse menu (ver Figura 6).

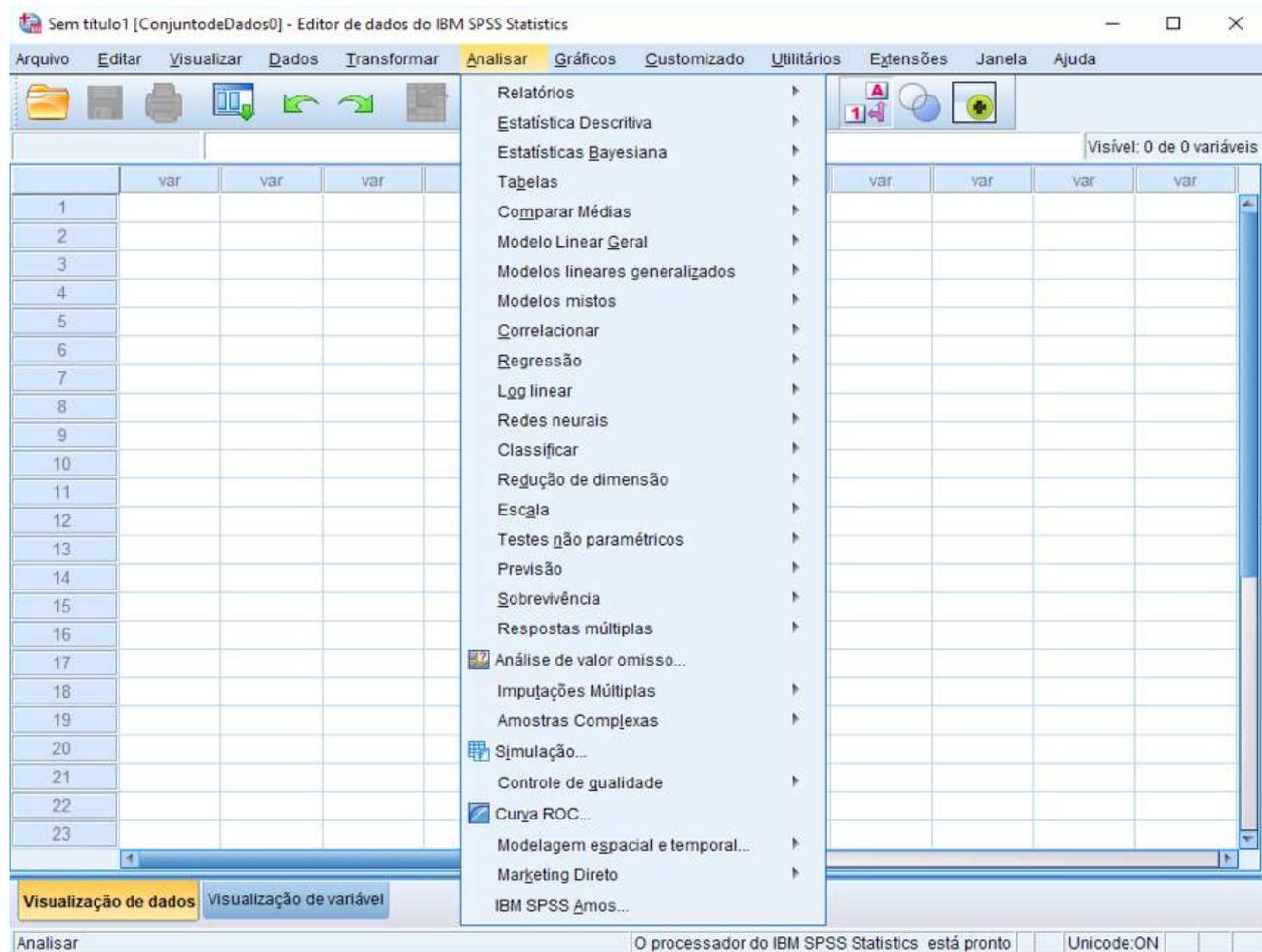
Figura 6 | O menu transformar



## 1.07 Analisar

O menu **Analisar** (ver Figura 7) reúne todos os procedimentos estatísticos disponíveis no IBM SPSS Statistics. Este é o principal menu que sofre alterações com a inclusão de outros módulos. À medida que são incluídos módulos opcionais, o número de opções (linhas) existentes nesse menu vai aumentando, ou seja, a quantidade de recursos estatísticos vai se tornando cada vez maior.

Figura 7 | O menu analisar

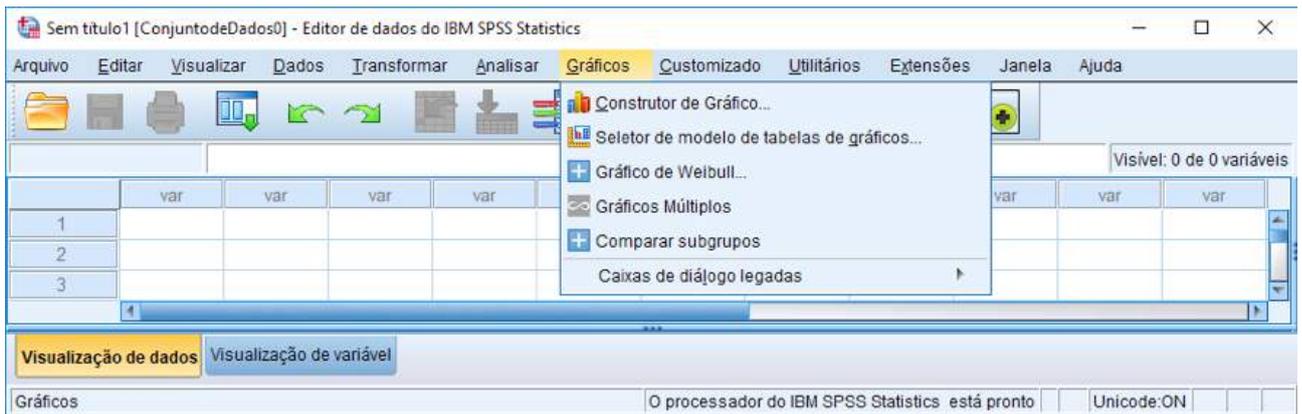


A opção **Marketing Direto** do menu Analisar (disponível apenas com o módulo Direct Marketing instalado) reúne técnicas para ajudar a entender o relacionamento com os clientes. Algumas análises que estão presentes neste módulo: Modelo RFV, Análise de Cluster e Análise de Perfis de Prospecção. Até a versão 24 do IBM SPSS Statistics, as opções deste módulo podiam ser acessadas em um menu à parte na barra de menus.

## 1.08 Gráficos

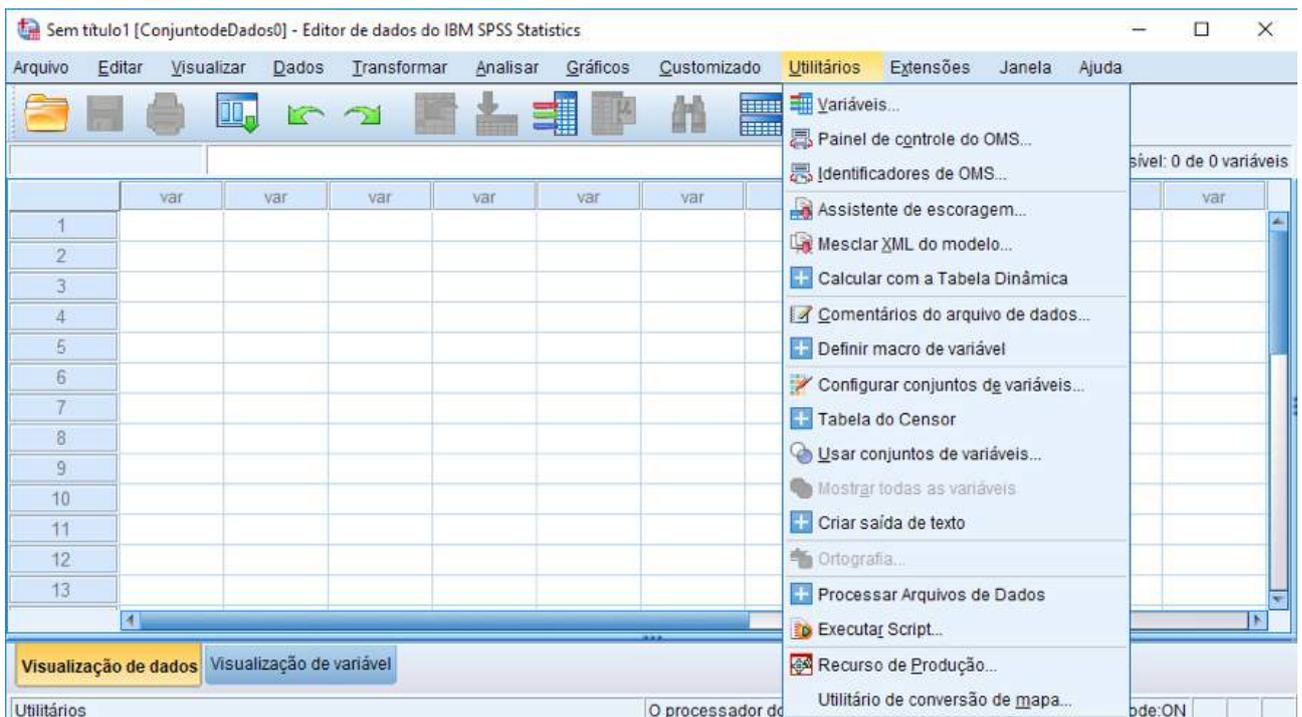
O menu **Gráficos** (ver Figura 8) contém todos os tipos de gráficos que o IBM SPSS Statistics oferece. Gráficos de barras, área, pizza, boxplot, histograma e cartas de controle são apenas alguns deles. Eles podem ser obtidos através da opção Construtor de Gráfico. Através da opção Seletor de Modelo de Tabelas de Gráficos, é possível visualizar o resultado das análises geograficamente com os modelos de mapas disponíveis no IBM SPSS Statistics.

Figura 8 | O menu Gráficos



## 1.09 Utilitários

**Utilitários** (ver Figura 9) é um menu que oferece recursos que não são imprescindíveis, mas certamente “quebram alguns galhos”. Um exemplo é obter informações sobre as variáveis do seu arquivo clicando na opção Variáveis. Além disso, temos a opção Executar Script..., onde podemos escolher algum script do IBM SPSS Statistics para ser executado. (Scripts são programas em Visual Basic que permitem a automatização de muitas tarefas no IBM SPSS Statistics).



## 1.10 Extensões

No menu **Extensões** (ver Figura 10), você pode baixar extensões adicionais para incrementar suas análises (a partir do Hub de Extensão) ou personalizar os menus do IBM SPSS Statistics, criando, até mesmo, um menu totalmente novo (a partir do Construtor de Diálogo Personalizado).

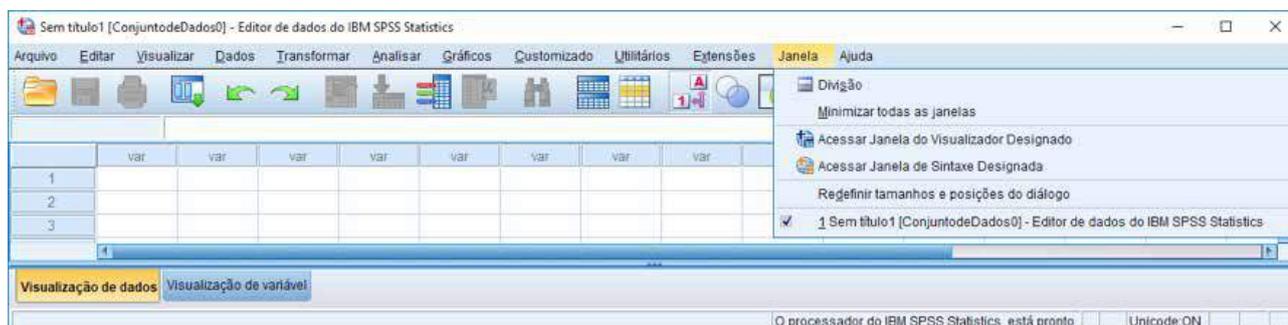
Figura 10



## 1.11 Janela

Utilizando o menu **Janela** (Figura 11), você pode navegar pelas janelas criadas pelo IBM SPSS Statistics, bastando escolher a janela desejada e clicar em seu nome; ou você pode minimizar o IBM SPSS Statistics com todas as suas janelas pela opção Minimizar Todas as Janelas. Por meio do comando Divisão, você divide a janela de dados em painéis para visualizar várias partes do banco de dados de uma só vez.

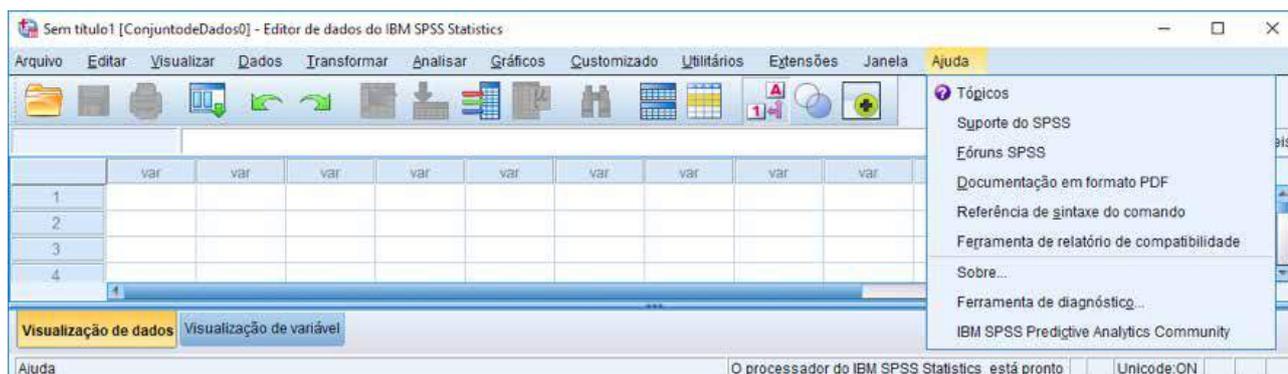
Figura 11: O menu Janela



## 1.12 Ajuda

O último menu que aparece é o menu de **Ajuda** (ver Figura 12). A ajuda no IBM SPSS Statistics é um dos pontos fortes do software, já que ela é muito mais do que uma cópia do manual. Quando clicamos em Ajuda, a primeira opção Tópicos lhe dá a chance de escolher entre esclarecimentos do significado de algum termo técnico ou de algum comando e um índice remissivo. Na opção Referência de Sintaxe do Comando, você terá todos os manuais de sintaxe do IBM SPSS Statistics.

Figura 12 | O menu Ajuda



Um tutorial do IBM SPSS Statistics (em Inglês) está disponível na janela de ajuda que é aberta quando clicamos na opção Tópicos do menu Ajuda (ver Figura I-13). Com ele, você pode obter informações básicas sobre o uso do IBM SPSS Statistics. A opção Case Studies (Estudos de Caso) apresenta exemplos de aplicações das análises oferecidas por cada módulo do IBM SPSS Statistics a partir de amostras de conjuntos de dados que ficam na pasta Samples, dentro do diretório em que o software está instalado. Já a opção Statistics Coach (também em Inglês) funciona como um técnico estatístico: ele irá fazer a você algumas perguntas sobre seus dados e o tipo de análise que você gostaria de obter e, então, baseado nas suas respostas, ele indicará o procedimento do IBM SPSS Statistics mais apropriado para o seu caso.

## 02 Criar um Banco de Dados

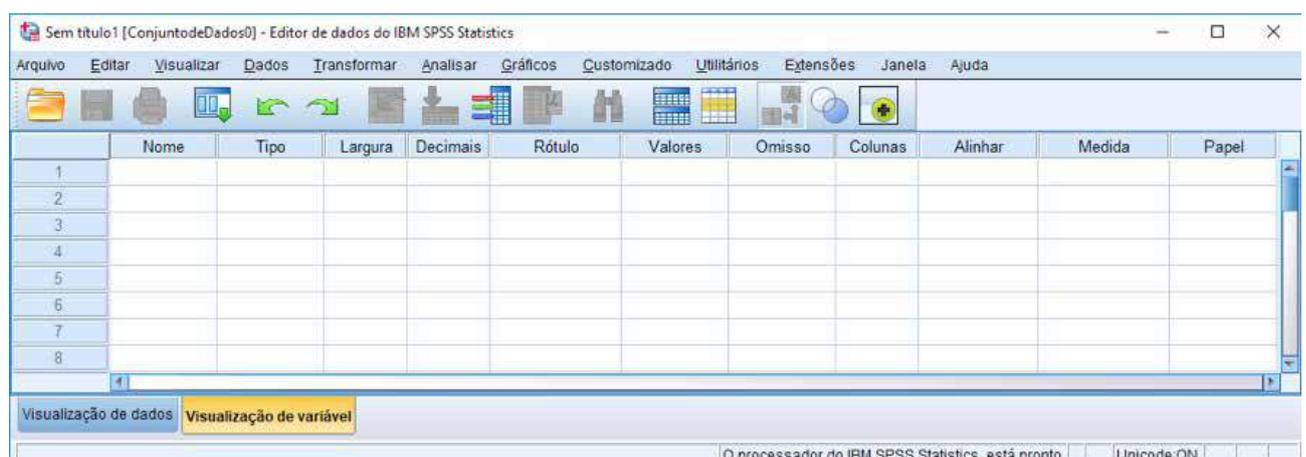
### 02.1 Como criar seu próprio banco de dados

Antes de começarmos a inserir valores na janela de dados, é preciso atentar ao fato de que o IBM SPSS Statistics tem, por regra geral, trabalhar com as variáveis (campos) definidas em colunas e casos (registros) em linhas. Neste treinamento, vamos trabalhar com as variáveis organizadas pela ordem em que aparecem no arquivo: no menu **Editar**, clique em **Opções** e, na aba **Geral**, selecione as opções **Exibir Nomes** e **Arquivo** na área **Listas de Variáveis**.

Digamos que estamos interessados em armazenar informações de uma pesquisa, sobre a programação das emissoras de televisão, realizada em São Paulo e Rio de Janeiro. Vamos criar uma coluna para a cidade do respondente.

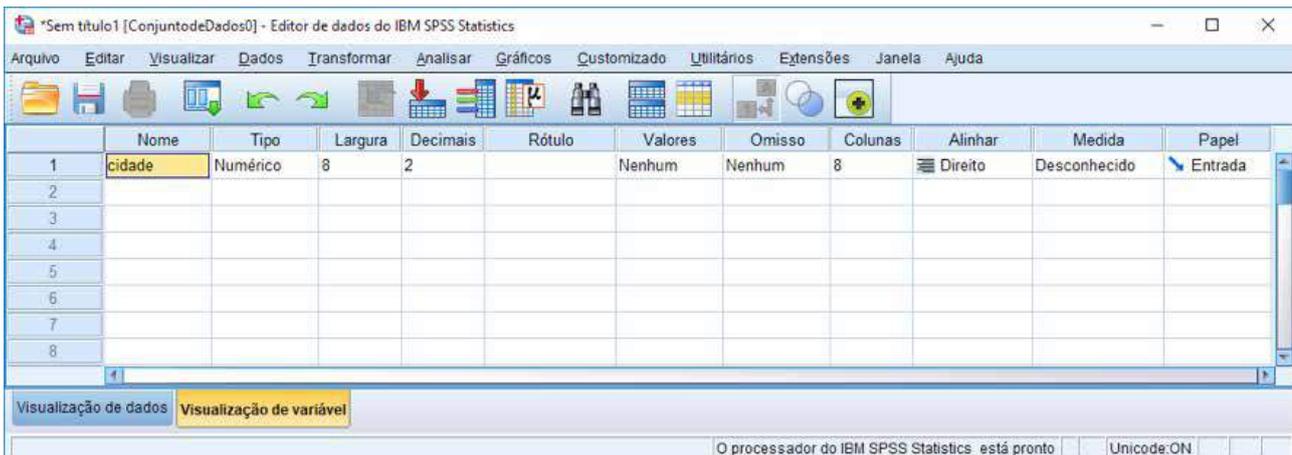
Para definir uma variável, clique duas vezes no nome da variável no topo da coluna ou, na parte inferior da janela de dados, clique em **Visualização de Variável**. Dessa maneira, o modo de definição de variáveis será exibido na janela de dados. Nesse modo, temos a opção de definir todas as características que fazem parte de uma variável: nome, tipo, rótulos, valores ausentes, formato da coluna e o nível de medida (ver Figura 13).

Figura 13 | Modo de definição de variáveis



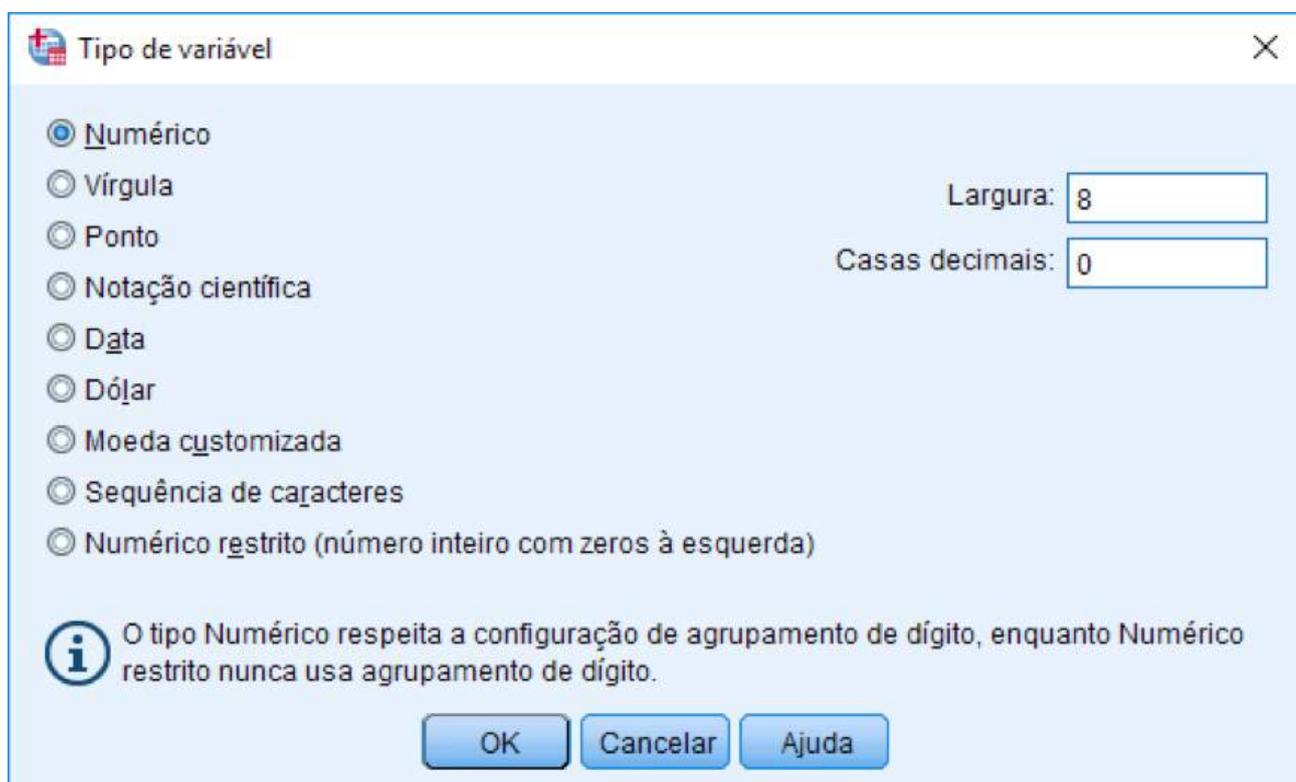
Para dar um nome à variável, basta digitá-lo no espaço designado por **Nome**; o nome da variável pode ter, no máximo, 64 caracteres, sem que haja caracteres especiais no início do nome, como +, &, \, -, \$, etc. (porém podemos acentuar); no nosso caso, o nome será *cidade* (ver Figura II-2). As palavras ALL, AND, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO e WITH são reservadas para os comandos do IBM SPSS Statistics e, por isso, não poderão ser usadas como nomes de variáveis. No nosso exemplo, vamos dar o nome *cidade* para a variável (ver Figura 14)

Figura 14 | Definição de variáveis: Nome



Para alterarmos o tipo da variável, basta clicar em **Tipo**, onde podemos definir a variável como numérica (*numeric*), moeda (*dollar*), data (*date*), alfanumérica (*string*, ou sequência de caracteres), entre outros tipos. Usaremos, no nosso caso, uma codificação para a variável *cidade*, onde o número 1 corresponderá à cidade de São Paulo e o número 2 ao Rio de Janeiro, pois estas foram as duas únicas cidades pesquisadas. Portanto, a variável *cidade* será definida como numérica (ver Figura 15). Ainda nessa janela, você pode especificar quantos caracteres a variável ocupa (em **Largura**) e o número de casas decimais (em **Casas decimais**). Vamos deixar 8 em **Largura** e 0 em **Casas decimais** (dado que só usamos os números 1 e 2, não precisamos de casas decimais) e clicar em **OK**.

Figura 15 | Definição de variáveis: Tipo, Tamanho e Casas Decimais



Você deve estar se perguntando: Por que não definir a variável *cidade* como alfanumérica (*string*) e digitar as palavras "São Paulo" e "Rio de Janeiro" ao invés dos números 1 e 2? A resposta é simples: muitos dos procedimentos estatísticos não conseguem trabalhar com variáveis desse tipo (*string*). A digitação de números, mesmo para variáveis categóricas, diminui o tempo e a

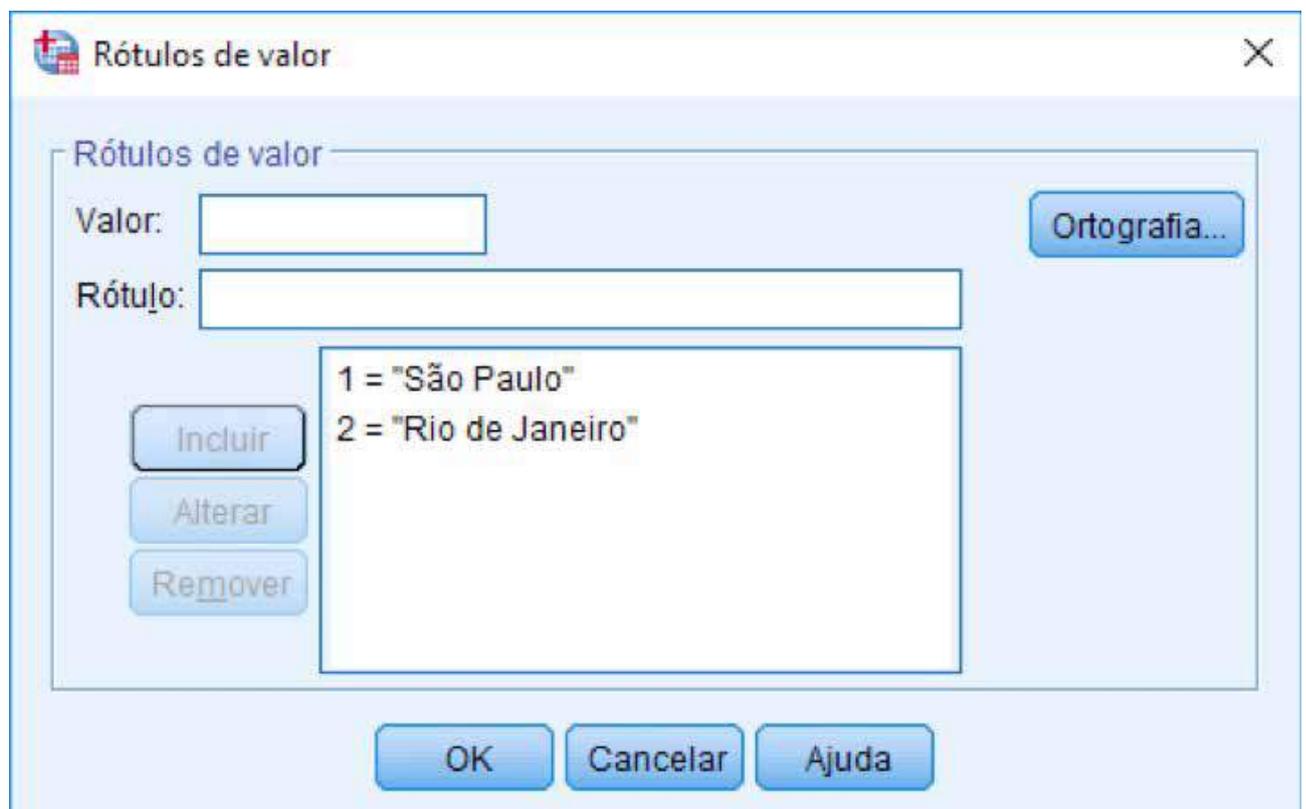
ocorrência de erros na digitação e, também, economiza espaço e, conseqüentemente, tempo no processamento dos casos.

Um recurso muito útil no IBM SPSS Statistics é a possibilidade de relacionarmos um texto (rótulo) ao nome das variáveis. Basta selecionar a coluna **Rótulo** (*Label*) e digitar o texto desejado. Vamos colocar "Cidade do entrevistado" como rótulo da variável cidade.

Podemos também definir rótulos para os valores que uma determinada variável pode assumir. A coluna **Valores** é reservada para esse tipo de definição. Clique no botão que aparece ao selecionarmos essa coluna: no espaço Valor, dentro da janela **Rótulos de Valor**, digitamos o valor que a variável assume e, em **Rótulo**, o rótulo (texto) para aquele valor. Feito isso, basta clicar em Incluir para armazenar o rótulo do primeiro valor e passar para a segunda definição.

No caso da variável *cidade*, definimos "São Paulo" como rótulo para o valor 1 e "Rio de Janeiro" para o valor 2 (ver Figura 16). A vantagem de definirmos rótulos para variáveis ou para os valores que uma variável pode assumir é que, toda vez que pedimos qualquer estatística envolvendo essas variáveis, eles aparecerão junto com os resultados na [janela de saídas](#). Dessa maneira poderemos identificar qual é a cidade referente a cada número (código).

Figura 16 | Definição de variáveis: Rótulos para os valores



Geralmente, não temos condições de obter um banco de dados totalmente completo (vamos supor que uma das perguntas da pesquisa é a idade do respondente; algumas pessoas não gostam de mencionar a idade). Se nada for informado nesse campo, o IBM SPSS Statistics coloca o separador decimal (. ou ,) no local e interpreta todos os pontos (ou vírgulas) como valores ausentes (*missing*). Tais valores, definidos como **system missing**, não serão considerados como valores válidos em qualquer tipo de análise a ser realizada.

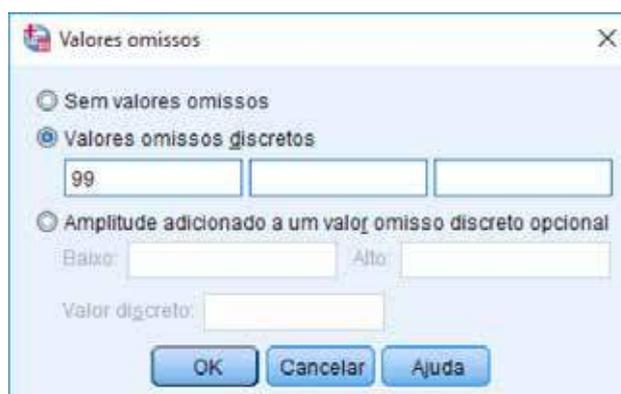
Podemos, porém, querer definir valores específicos como *missing* (no caso, **user missing**, valores ausentes definidos pelo usuário). Basta selecionar a coluna Omisso, clicar no botão que aparecerá e escolher a opção mais conveniente (ver Figura 17). Por exemplo, se escolhermos

a opção **Valores omissos discretos**, podemos definir o número 99 como valor user missing e cada vez que o IBM SPSS Statistics encontrar o número 99 no meio de uma análise estatística que envolve essa variável, ele saberá que aquele valor se refere a um valor ausente.

Caso a variável seja alfanumérica (*string*), qualquer valor - incluindo valor em branco - será considerado como válido. Para que isso não ocorra, basta definir como *missing* um espaço em branco: automaticamente, o IBM SPSS Statistics entenderá como valor *missing* qualquer resposta em branco dessa variável.

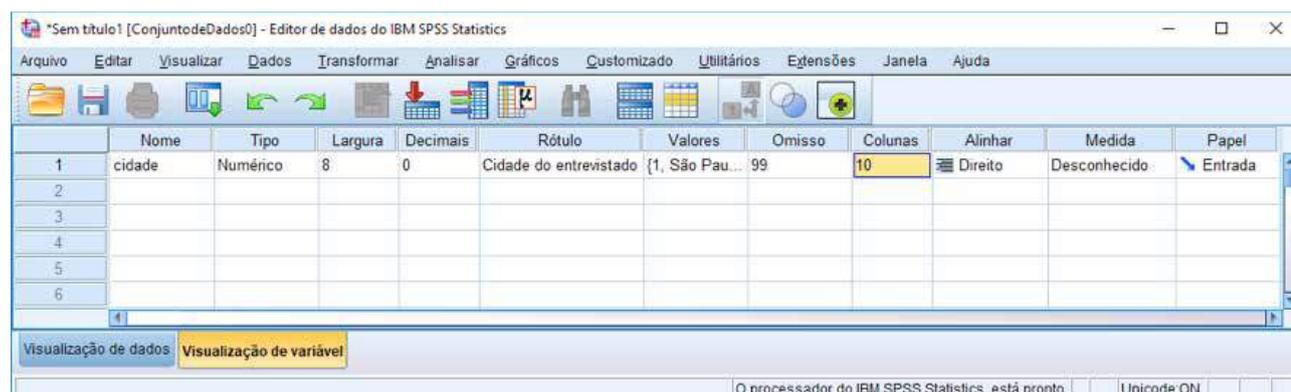
A definição do formato da coluna é um aspecto meramente estético. Através da coluna **Colunas**, você poderá determinar a largura da coluna, e na coluna **Alinhar**, o alinhamento do texto que pode ser: à esquerda, à direita ou centralizado.

Figura 17 | Definição de variáveis: Valores Omissos (*Missing Values*)



No caso da variável *cidade*, definiremos o tamanho da largura da coluna como 10 (dez) e o alinhamento à direita (ver Figura 18).

Figura 18 | Definição de variáveis: Formato da Coluna



E, finalmente, na coluna **Medida**, devemos escolher o nível de medida da variável. As possíveis escolhas são:

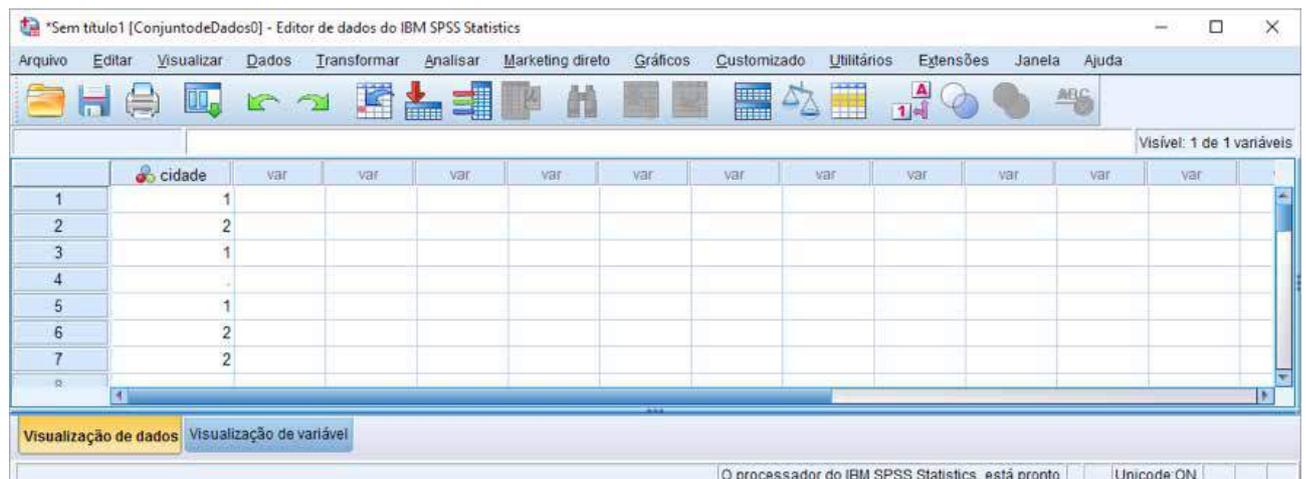
- **Escala (Scale)**: Use esta opção quando os dados são valores numéricos numa escala intervalar ou de razão (ex: idade, renda, peso, altura). Obs: A variável deverá ser numérica.
- **Ordinal**: Quando os dados representam categorias com uma ordem intrínseca (ex: pequeno, médio, grande; concordo totalmente, concordo, discordo, discordo totalmente). Podem ser variáveis alfanuméricas (*string*) ou numéricas com valores representando as categorias distintas (ex: 1 = "pequeno", 2 = "médio", 3 = "grande").

• **Nominal**: Quando os dados representam categorias sem ordem intrínseca (ex: as categorias de estado civil, ocupação, nacionalidade). Podem ser variáveis alfanuméricas (string) ou numéricas com valores representando as categorias distintas (ex: 1 = masculino; 2 = feminino).

No nosso caso, escolhemos a opção **Nominal**, uma vez que não existe ordem intrínseca nas categorias da variável *cidade*.

Pronto! Agora é só voltar para o modo de exibição dos dados (selecione **Visualização de Dados** na parte inferior da [janela de dados](#)) e digitar os valores coletados, lembrando que cada caso (registro) deve ser digitado em uma linha (ver Figura 19). Desse modo, o primeiro respondente é de São Paulo, o segundo do Rio de Janeiro e assim por diante.

Figura 19 | Janela de dados após definição de uma variável

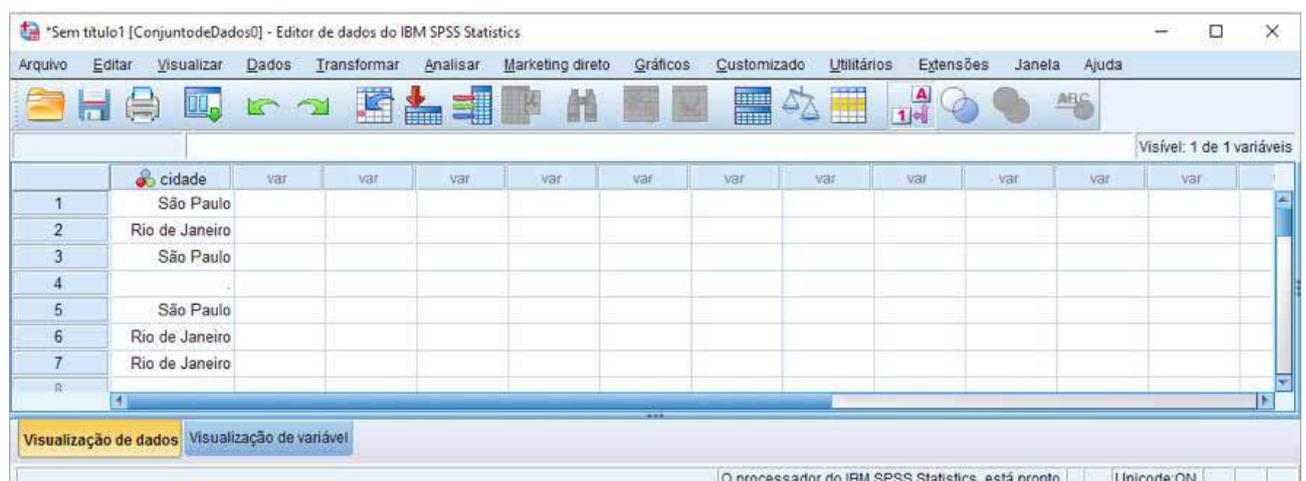


Note que apenas os números 1 e 2 aparecem na [janela de dados](#) e não os rótulos (nome das cidades). Para ativá-los basta clicar no ícone **Rótulos de Valor** (na barra de ferramentas) ou na opção **Rótulos de Valor** do menu **Visualizar**.

A Figura 20 mostra a janela de dados agora com os rótulos ativos. Para desativá-los, basta clicar novamente no mesmo ícone.

Para salvar as informações armazenadas nessa janela no formato IBM SPSS Statistics, basta clicar no menu **Arquivo** → **Salvar** (ou **Salvar Como**) e digitar o nome que queremos dar ao novo arquivo no quadro **Nome do Arquivo**. Note que a extensão dos arquivos do tipo IBM SPSS Statistics é .sav.

Figura 20 | Janela de dados após a ativação dos rótulos via ícone Rótulos de Valor



## 02.2 Definindo propriedades de variáveis categóricas

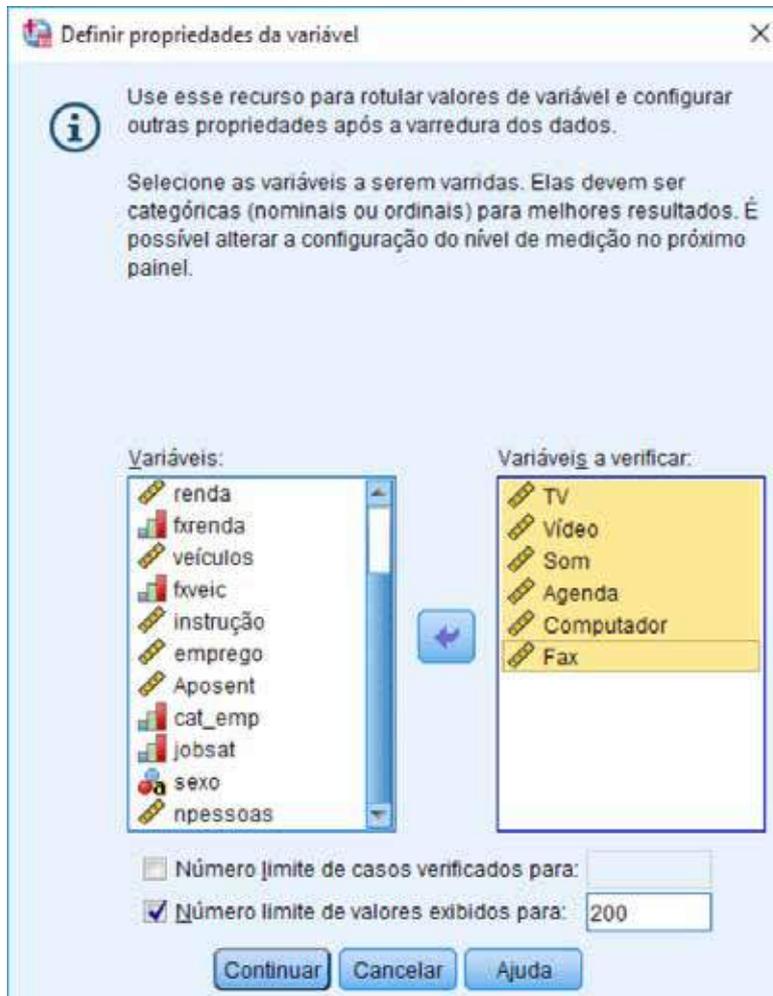
No menu **Dados**, temos a opção **Definir Propriedades das Variáveis**, que facilitará a definição dos rótulos dos valores e outras propriedades das variáveis categóricas (nominais ou ordinais). Esse comando faz a leitura dos valores atuais dos dados e lista todos os únicos valores para cada variável selecionada; identifica valores que estão sem rótulo e promove uma rotulação automática e, ainda, pode copiar os rótulos dos valores de uma variável para uma ou mais variáveis selecionadas.

No nosso exemplo, utilizaremos o arquivo *pesquisa.sav*: para abrir o arquivo no formato IBM SPSS Statistics (.sav) basta clicar no menu **Arquivo** → **Abrir** → **Dados** e escolher o caminho e o nome do arquivo, no nosso caso, *pesquisa.sav*.

Na variável *computador* (Possui computador?) temos duas respostas válidas: 0 ("Não") e 1 ("Sim"); porém, se olharmos o primeiro caso do banco de dados veremos que temos uma resposta 99 para essa variável. Nós poderíamos utilizar os conceitos vistos anteriormente para definir este valor 99 como *missing* e adicioná-lo também nas demais variáveis; porém, para fazer isso, utilizaremos o comando **Definir Propriedades das Variáveis**.

No menu **Dados**, selecione a opção **Definir Propriedades das Variáveis**. O IBM SPSS Statistics abrirá a janela inicial do comando (Figura 21). Nesta janela, devemos selecionar as variáveis que queremos definir em algum rótulo e/ou outras propriedades e colocá-las no campo **Variáveis a verificar**. Escolheremos as seguintes variáveis: *computador*, *fax*, *agenda*, *som*, *tv* e *vídeo*. (Pelas escolhas que fizemos nas **Opções** do IBM SPSS Statistics, estas variáveis aparecem no final da lista de variáveis, o que facilita a inclusão das mesmas na lista de variáveis a serem verificadas.)

Figura 21 | Janela para a escolha das variáveis que serão lidas



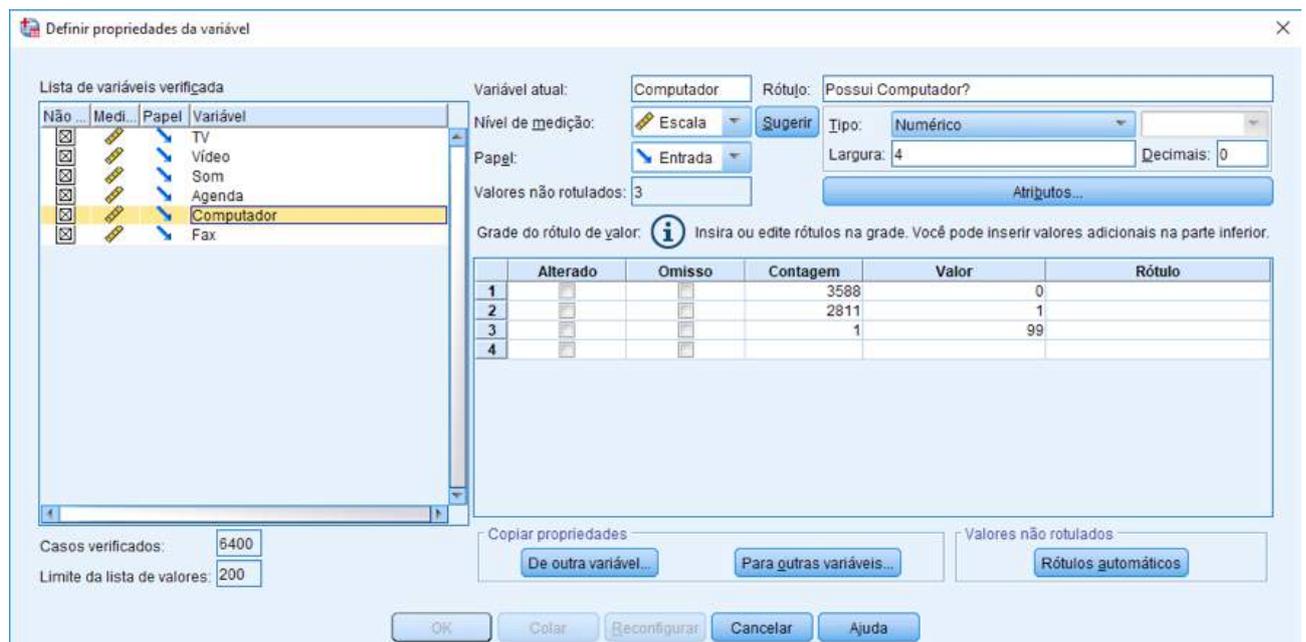
Nos bancos de dados com uma grande quantidade de casos, podemos selecionar a opção **Número limite de casos verificados** para limitar em  $n$  casos o número de casos que serão lidos em cada variável, reduzindo assim o tempo de processamento. A opção **Número limite de valores exibidos para:** será usada para limitar em  $n$  respostas o número de valores que serão mostrados para cada variável, evitando a presença de muitos valores, principalmente no caso de variáveis escalares. Mantenha as duas opções como estão e clique em **Continuar**.

No próximo passo, teremos todas as variáveis que foram lidas e suas respectivas propriedades (Figura 22).

Na lista de variáveis, selecione a variável *computador*. O nível de medida para essa variável é escalar (**Escala**). Nós podemos mudar esse nível de medida selecionando uma das opções (nominal, ordinal, escalar) presentes no campo **Nível de medição** ou clicando no botão **Sugerir**: nessa opção, teremos uma janela onde podemos transformar automaticamente o nível de medida da variável selecionada, seguindo a sugestão apresentada pelo *software* de acordo com os valores encontrados na variável. Escolha a opção **Nominal** e, em seguida, clique em **Continuar**.

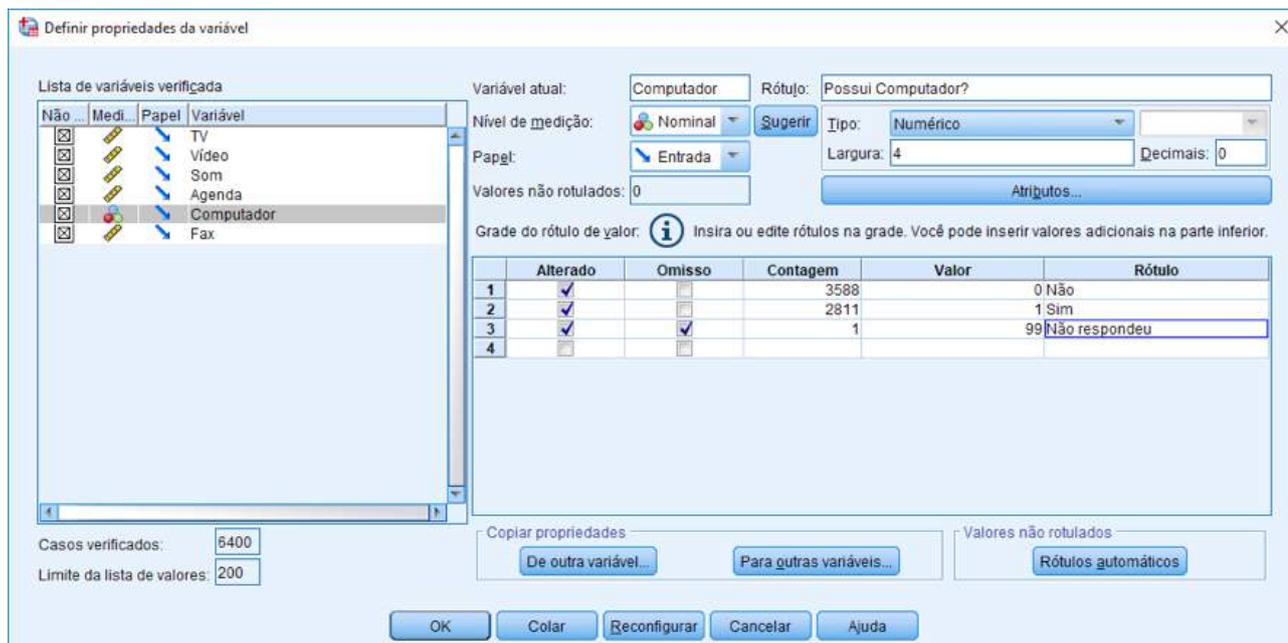
Na janela **Definir Propriedades de Variável** temos, também, a opção **Grade do rótulo de valor**, onde podemos adicionar novos valores e rótulos ou alterar os já existentes, visualizar a frequência de cada valor e definir qual valor será *missing*. Qualquer valor que sofra alguma alteração nessa janela ficará marcado como **Alterado**.

Figura 22 | Janela para definir as propriedades das variáveis (primeira parte)



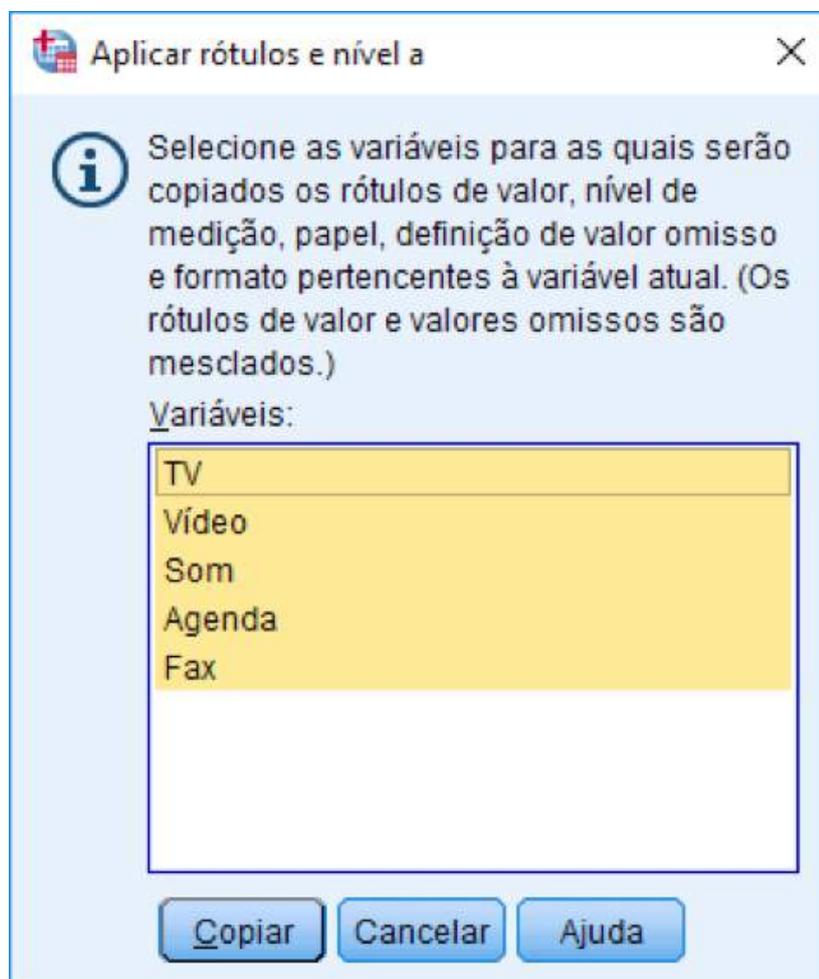
Ainda com a variável *computador* selecionada, faremos algumas alterações para o valor de *user missing* '99': na área **Rótulo**, escreveremos 'Não Respondeu' e clicaremos também na opção **Omisso** para essa resposta (Figura 23): desta maneira, estamos definindo que, sempre que o software encontrar o valor '99' na variável *computador*, ele será tratado como um valor *missing*.

Figura 23 | Janela para definir as propriedades das variáveis (segunda parte)



Podemos aplicar as mesmas definições da variável *computador* para as outras variáveis listadas. Para isso, basta clicar na opção **Para Outras Variáveis** do quadro **Copiar Propriedades** e em seguida, na janela **Aplicar Rótulos e Nível a** (Figura 24), selecione todas as variáveis e, em seguida, clique em **Copiar**.

Figura 24 | Janela 'Aplicar Rótulos e Nível a'



Podemos observar que, agora, todas as variáveis listadas receberam as mesmas definições que a variável *computador*, incluindo o valor '99' (Figura 25). Clique em **OK** para completar o procedimento. O resultado final está na Figura 26.

Outra opção muito útil para ajudá-lo na definição das características de suas variáveis chama-se **Copiar Propriedades dos Dados** e está presente no menu **Dados**. Ao escolher essa opção, será aberta uma janela para você selecionar um arquivo no formato IBM SPSS Statistics, chamado de arquivo modelo. Após a escolha, o IBM SPSS Statistics compara os nomes das variáveis do arquivo ativo com as do arquivo modelo e para aquelas variáveis que tiverem o mesmo nome, ele copia todas as definições dessas variáveis (como tipo, rótulo, valores ausentes, etc) do arquivo modelo para o arquivo ativo. Este comando substitui o antigo comando **Aplicar Dicionário aos Dados**.

Figura 25 | Janela de definição das propriedades das variáveis após as alterações

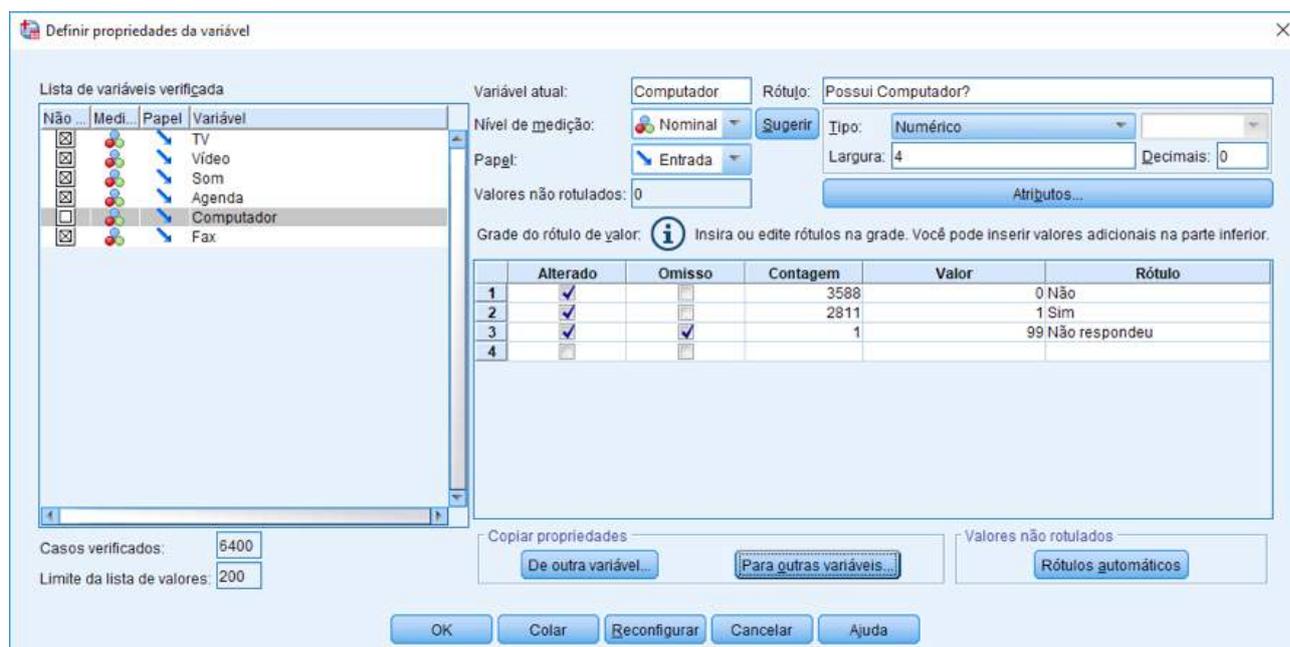
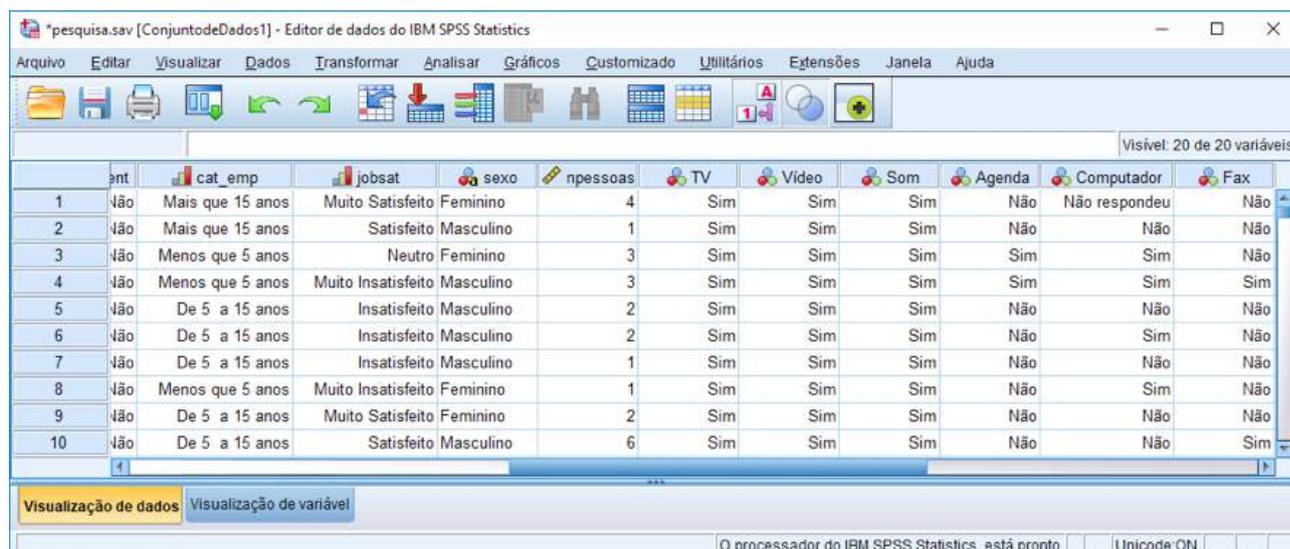


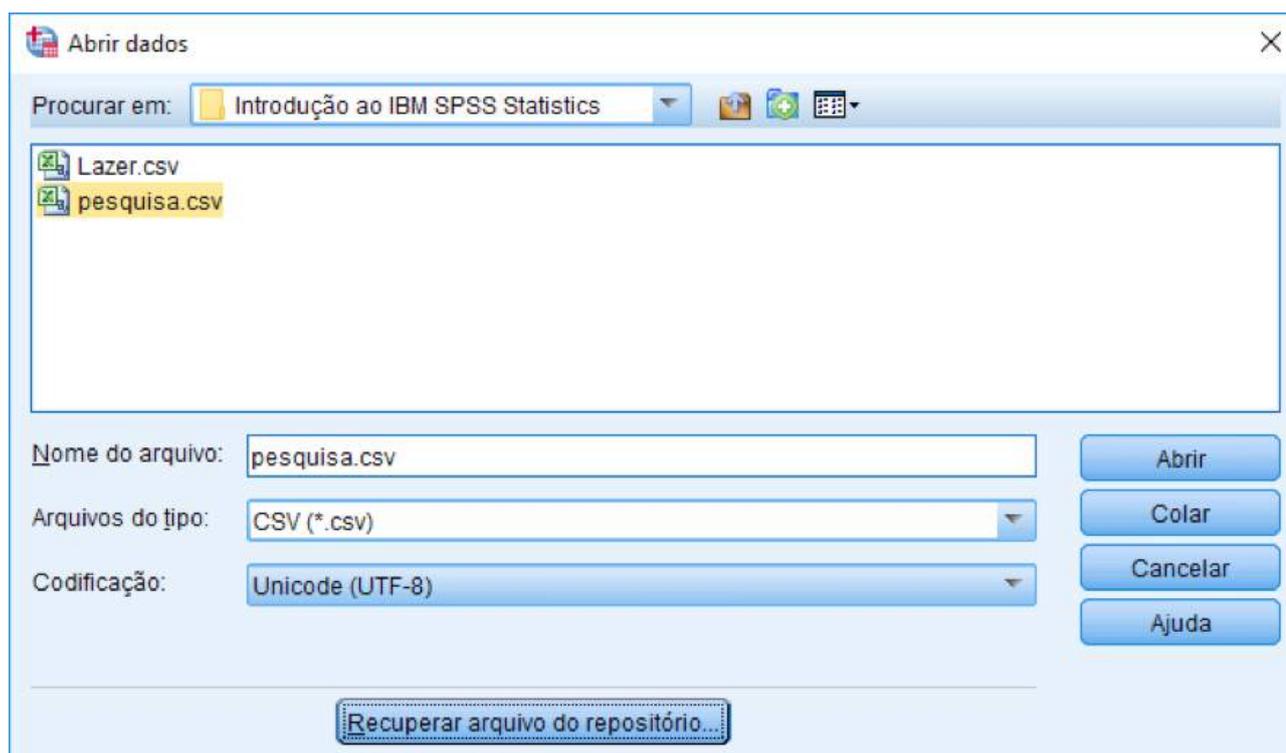
Figura 26: Arquivo *pesquisa.sav* após a definição das propriedades das variáveis



## 02.3 Formas de importação e exportação de dados

Conforme já mencionamos, o IBM SPSS Statistics tem a capacidade de importar bancos de dados gerados nos mais diversos formatos, entre eles Lotus 123, Microsoft Excel, Dbase e ASCII. Para ter acesso aos diferentes formatos de banco de dados, basta clicar no menu **Arquivo** → **Abrir** → **Dados** (ou em **Arquivo** → **Importar Dados** → **Dados de Texto**). Aparece então uma tela (ver Figura 27) na qual determinamos disco e subdiretórios, o nome do arquivo e a forma sob a qual o arquivo foi gerado (opção Arquivos do tipo, no canto inferior esquerdo). O padrão é o formato IBM SPSS Statistics (extensão *.sav*).

Figura 27 | Janela para a abertura de bancos de dados



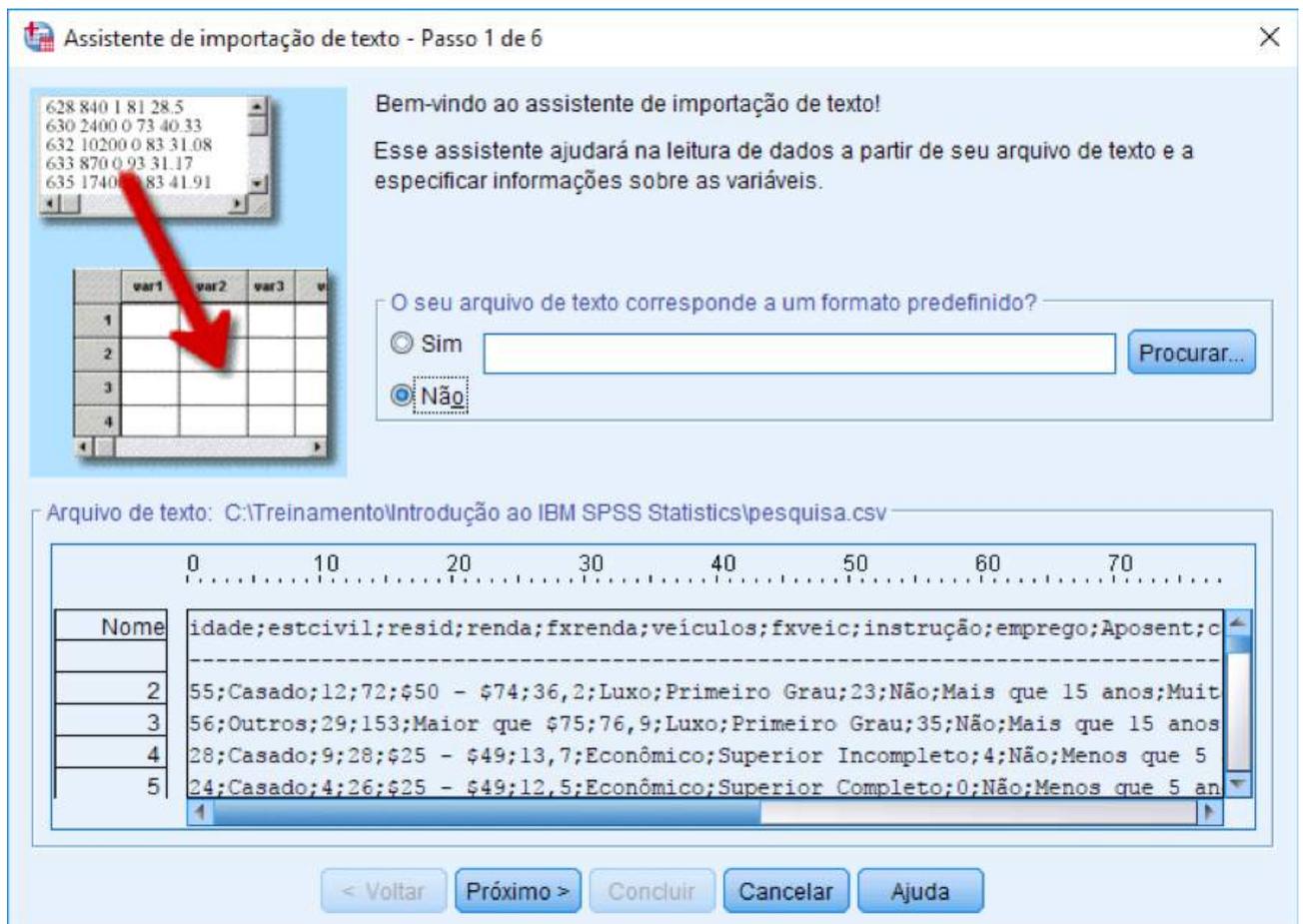
Para a abertura de arquivos do tipo texto (extensões *.txt*, *.dat*, *.tab* e *.csv*, por exemplo), o IBM SPSS Statistics possui um assistente próprio, uma vez que alguns destes arquivos podem ter valores ou nomes de variáveis separados por tabulações (extensão *.dat* ou *.tab*), vírgulas (extensão *.csv*) ou mesmo por espaços simples ou de tamanho fixo, o que pode gerar um problema grave na hora de ler variáveis alfanuméricas com espaços (por exemplo, nome completo de uma pessoa entrevistada em uma pesquisa).

Para o nosso primeiro exemplo de importação, selecione a opção **Texto** no campo **Arquivo do tipo**; apenas arquivos texto (extensões *.txt*, *.csv*, *.dat* e *.tab*) estarão disponíveis para a importação. Vamos escolher o arquivo *pesquisa.csv* para nosso exemplo. Certifique-se de que a opção **Unicode (UTF-8)** esteja selecionada na área **Codificação**, uma vez que o arquivo está codificado nesse padrão. Após a seleção, abre-se a janela do **Assistente de Importação de Texto (Text Import Wizard)**, o assistente do IBM SPSS Statistics para a importação de arquivos texto. Repare que o IBM SPSS Statistics possui a opção **Importar Dados** no menu **Arquivo** e, a partir dela, poderíamos ter escolhido a subopção **Dados CSVe**, assim, teríamos uma janela de abertura de arquivos *.csv*.

No primeiro passo (ver Figura 28), é possível visualizar uma amostra do arquivo e como as variáveis estão dispostas no mesmo; esse é um passo fundamental para determinarmos como os dados serão importados para o formato SPSS, especialmente quando estão em arquivos

de texto, com colunas de tamanho fixo, ou se os nomes das variáveis estão na primeira linha. A pré-visualização é um recurso que estará presente durante todos os passos do assistente. Além disso, no canto superior direito, há a possibilidade de escolher um formato predefinido para a importação (ao clicarmos em **Sim** na caixa **O seu arquivo de texto corresponde a um formato predefinido?**), algo útil quando abrimos arquivos com a mesma formatação de colunas ou com o mesmo separador entre valores (por exemplo, vírgulas). No nosso exemplo, mantenha as opções padrão e clique em **Próximo**.

Figura 28 | Assistente de Importação de Texto – Passo 1



No passo 2 (Figura 29), definimos se o arquivo possui algum tipo de separador (**Delimitado**, que é o nosso caso) ou colunas de tamanhos fixos (**Largura fixa**) e se o nome das variáveis está na primeira linha do arquivo (clique em **Sim** e mantenha o número padrão 1 para a linha que contém o nome das variáveis). Definidas estas opções, clique em **Próximo** para entrar no terceiro passo (Figura 30).

Agora, você tem a opção de escolher em qual linha do arquivo está o primeiro caso (quando definimos que a primeira linha contém os nomes das variáveis, o padrão é a linha 2, que é o que manteremos), se cada linha completa representa um caso ou apenas um conjunto de variáveis (manteremos a primeira opção, **Cada linha representa um caso**) e quantos casos você importará: todos (**Todos os casos**, opção que manteremos); uma quantidade qualquer (**Os primeiros n casos**) ou; uma amostra aleatória cuja porcentagem aproximada de casos selecionados é definida pelo usuário (**Uma porcentagem aleatória dos casos**). Clique em **Próximo**.

Figura 29 | Assistente de Importação de Texto – Passo 2

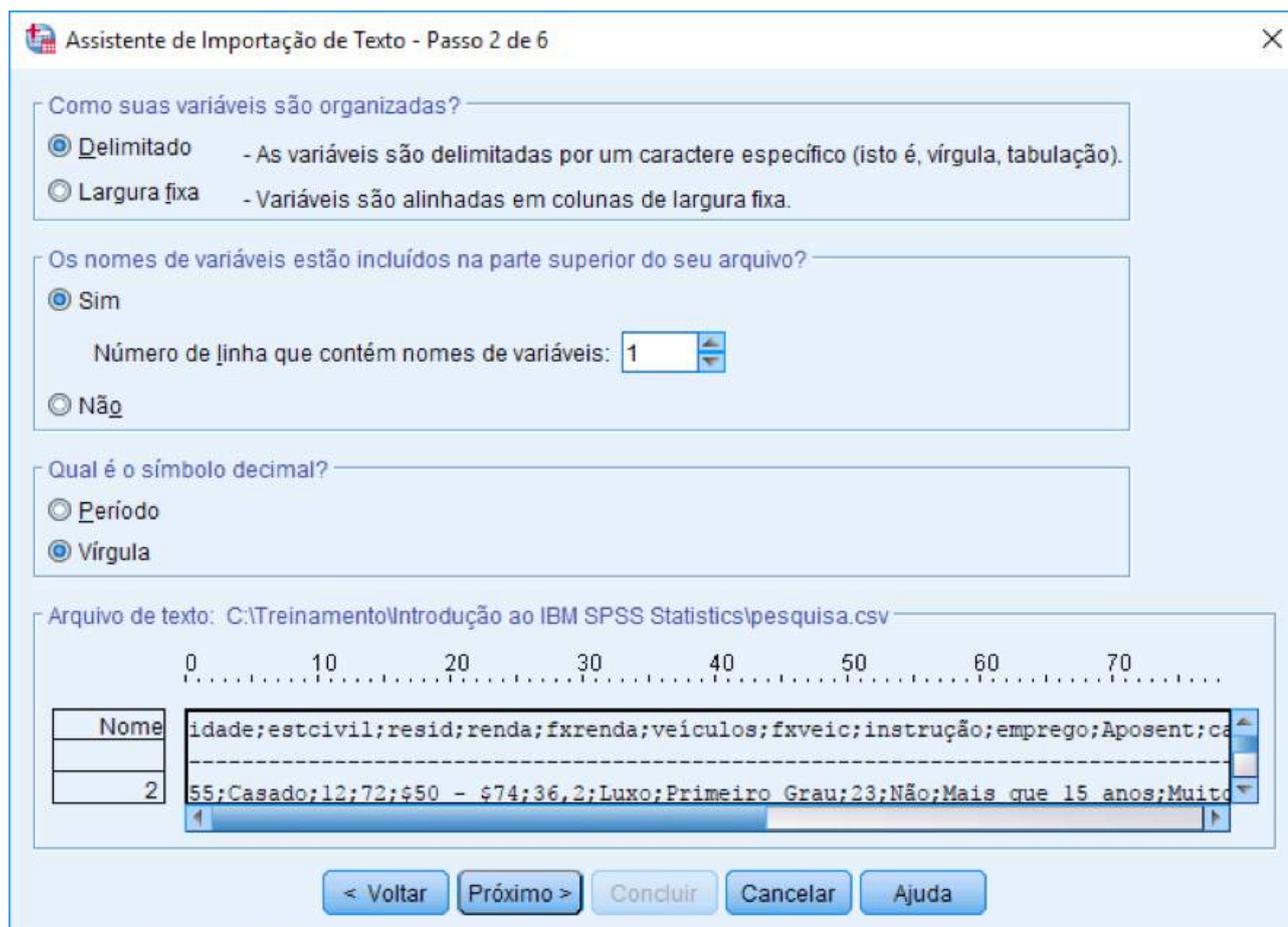


Figura 30 | Assistente de Importação de Texto – Passo 3

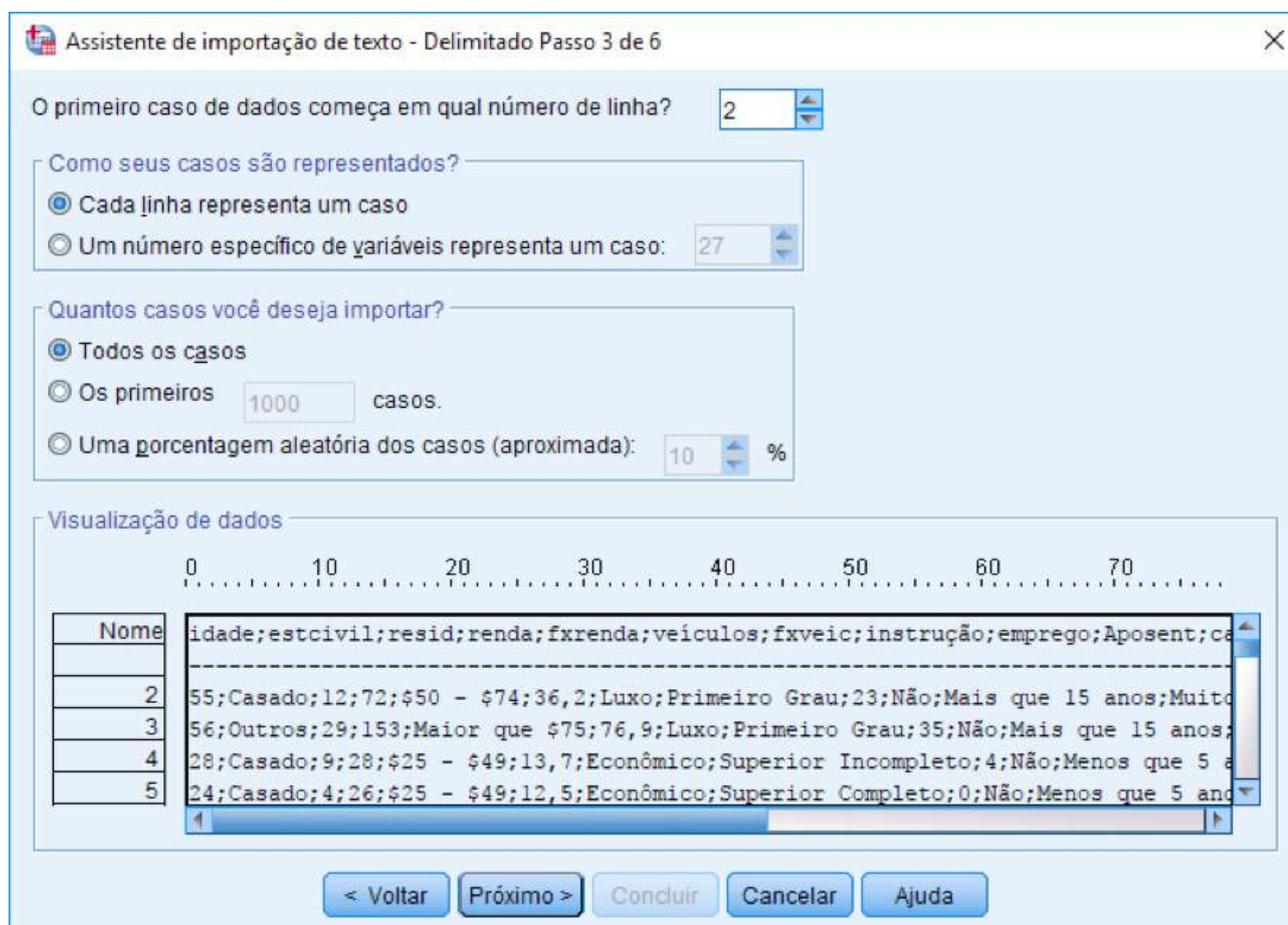




Figura 32 | Assistente de Importação de Texto – Passo 5

Assistente de Importação de Texto - Passo 5 de 6

Especificações para variáveis selecionadas na visualização de dados

Nome de variável:  Nome original: idade

Formato de dados:

Porcentagem de valores que determinam o formato de dados automático

Visualização de dados

idade	estcivil	resid	renda	fxrenda	veículos	fxveic	instrução	err
55	Casado	12	72	\$50 - \$74	36,2	Luxo	Primeiro Gr...	23
56	Outros	29	153	Maior que ...	76,9	Luxo	Primeiro Gr...	35
28	Casado	9	28	\$25 - \$49	13,7	Econômico	Superior In...	4
24	Casado	4	26	\$25 - \$49	12,5	Econômico	Superior C...	0
25	Outros	2	23	Menor que ...	11,3	Econômico	Segundo g...	5

< Voltar Próximo > Concluir Cancelar Ajuda

No último passo (Figura 33), podemos escolher se faremos a importação imediatamente ou colaremos o comando para uma janela de sintaxe. (Falaremos mais sobre sintaxe no Capítulo 6): manteremos marcada a opção **Não** para a caixa **Gostaria de colar a sintaxe?** Além disso, temos a possibilidade de salvar todas as opções definidas anteriormente em um novo arquivo (extensão *.tpf*, de *TextWizard Predefined Format*) que pode ser aberto no primeiro passo do *Text Import Wizard* em uso futuro. Por fim, clique no botão **Concluir**. Na Figura 34, temos o resultado da importação; agora, os dados estão prontos para serem preparados dentro do IBM SPSS Statistics.

Assistente de Importação de Texto - Passo 6 de 6

Você definiu com sucesso o formato do seu arquivo de texto.

Gostaria de salvar este formato de arquivo para uso futuro?

Sim  Não

Gostaria de colar a sintaxe?  Armazenar dados em cache localmente

Sim  Não

Pressione o botão Concluir para concluir o assistente de importação de texto.

Visualização de dados

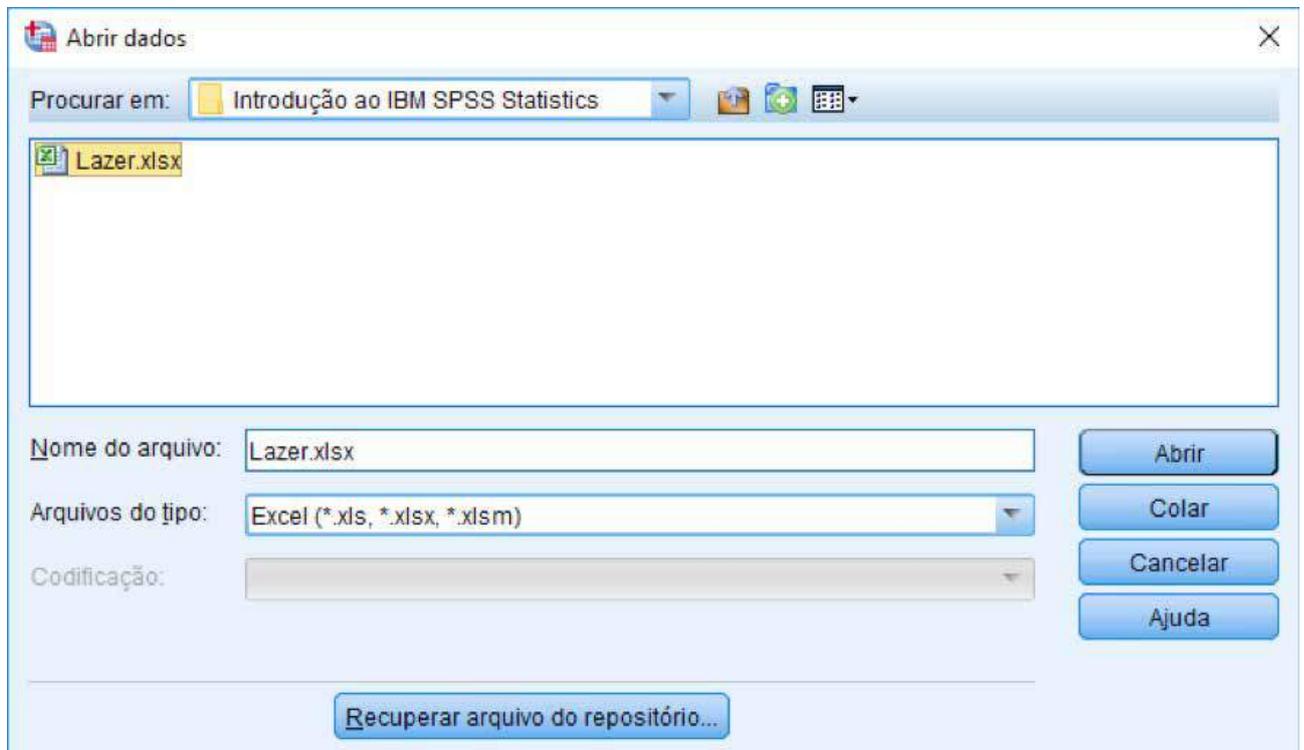
idade	estcivil	resid	renda	fxrenda	veículos	fxveic	instrução	err
55	Casado	12	72	\$50 - \$74	36,2	Luxo	Primeiro Gr...	23
56	Outros	29	153	Maior que ...	76,9	Luxo	Primeiro Gr...	35
28	Casado	9	28	\$25 - \$49	13,7	Econômico	Superior In...	4
24	Casado	4	26	\$25 - \$49	12,5	Econômico	Superior C...	0
25	Outros	2	23	Menor que ...	11,3	Econômico	Segundo g...	5
45	Casado	9	76	Maior que ...	37,2	Luxo	Superior In...	13
42	Outros	19	40	\$25 - \$49	19,8	Standard	Superior In...	10

< Voltar Próximo > Concluir Cancelar Ajuda

Figura 34 | Resultado final da importação do arquivo pesquisa.csv

	idade	estcivil	resid	renda	fxrenda	veiculos	fxveic	instrução
1	55	Casado	12	72	\$50 - \$74	36,2	Luxo	Primeiro Grau
2	56	Outros	29	153	Maior que \$75	76,9	Luxo	Primeiro Grau
3	28	Casado	9	28	\$25 - \$49	13,7	Econômico	Superior Incompleto
4	24	Casado	4	26	\$25 - \$49	12,5	Econômico	Superior Completo
5	25	Outros	2	23	Menor que \$25	11,3	Econômico	Segundo grau
6	45	Casado	9	76	Maior que \$75	37,2	Luxo	Superior Incompleto
7	42	Outros	19	40	\$25 - \$49	19,8	Standard	Superior Incompleto
8	35	Outros	15	57	\$50 - \$74	28,2	Standard	Segundo grau
9	46	Outros	26	24	Menor que \$25	12,2	Econômico	Primeiro Grau
10	34	Casado	0	89	Maior que \$75	46,1	Luxo	Superior Incompleto
11	55	Casado	17	72	\$50 - \$74	35,5	Luxo	Superior Incompleto
12	28	Outros	3	24	Menor que \$25	11,8	Econômico	Superior Completo
13	31	Casado	9	40	\$25 - \$49	21,3	Standard	Superior Completo
14	42	Outros	8	137	Maior que \$75	68,9	Luxo	Superior Incompleto
15	35	Outros	8	70	\$50 - \$74	34,1	Luxo	Superior Incompleto

Como um último exemplo típico de importação de dados, vamos abrir uma planilha de um arquivo do Microsoft Excel. Como alternativa à escolha do menu **Arquivo** → **Abrir** → **Dados**, vamos escolher o menu **Arquivo** → **Importar Dados** → **Excel** e, a seguir, escolha o arquivo *Lazer.xlsx* (ver Figura 35). A seguir, clique em **Abrir**.

Figura 35 | Importação do arquivo *Lazer.xlsx* (parte 1)

Agora, abre-se uma nova caixa de diálogo (ver Figura 36), na qual devemos definir em qual planilha estão os dados que serão importados (no caso em que o banco de dados possua duas ou mais planilhas diferentes) e o intervalo no qual os mesmos estão na referida planilha, incluindo eventuais nomes das variáveis. Selecionaremos a planilha LAZER na opção **Planilha**. Repare que uma opção padrão é a de ler os nomes das variáveis a partir da primeira linha do

intervalo de dados que foi definido (**Ler nomes de variáveis a partir da primeira linha de dados**): manteremos esta opção ativada.

Finalmente, repare que a pré-visualização dos dados nos mostra que a importação será feita de forma incorreta, uma vez que os nomes das variáveis estão originalmente na quarta linha da planilha LAZER (o cabeçalho da planilha, "Pesquisa de Opinião e Lazer", virou nome de uma das variáveis, o que é um indicativo claro de que o intervalo de dados a serem importados não está correto). Sendo assim, devemos especificar para o SPSS que os dados devem ser lidos no intervalo entre as células A4 (nome da primeira variável) e P35 (valor da variável MÚSICA para o último dos 31 casos presentes no banco de dados). Logo, preencheremos o campo **Amplitude** com o intervalo A4:P35 (ver Figura 37). Agora, repare que a pré-visualização indica que os dados serão importados corretamente. Por fim, basta clicar em **OK** e observar que o arquivo foi importado corretamente para uma nova janela de dados do IBM SPSS Statistics (ver Figura 38).

Figura 36 | Importação do arquivo Lazer.xlsx (parte 2)

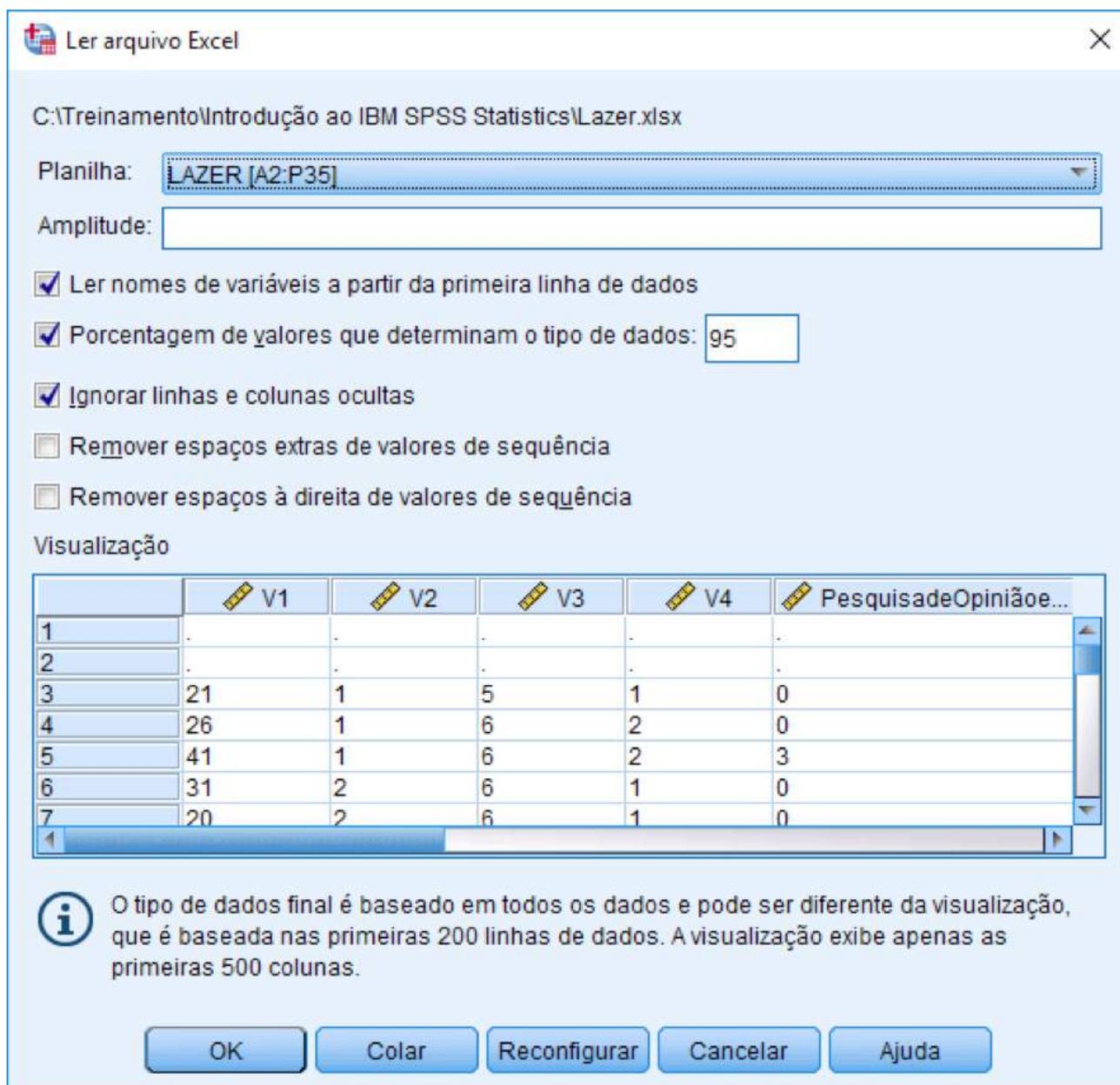


Figura 37 | Importação do arquivo Lazer.xlsx (parte 3)

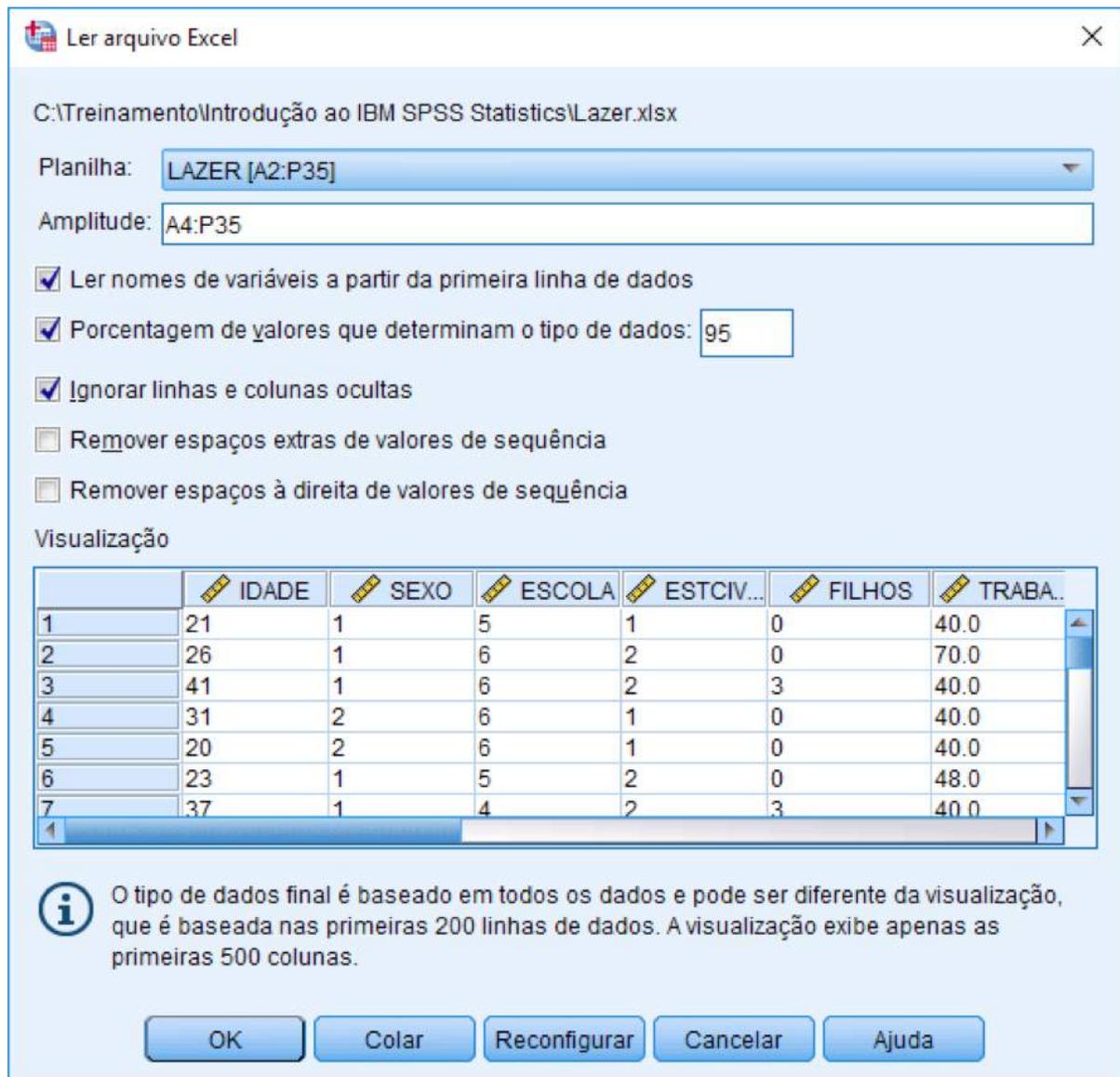
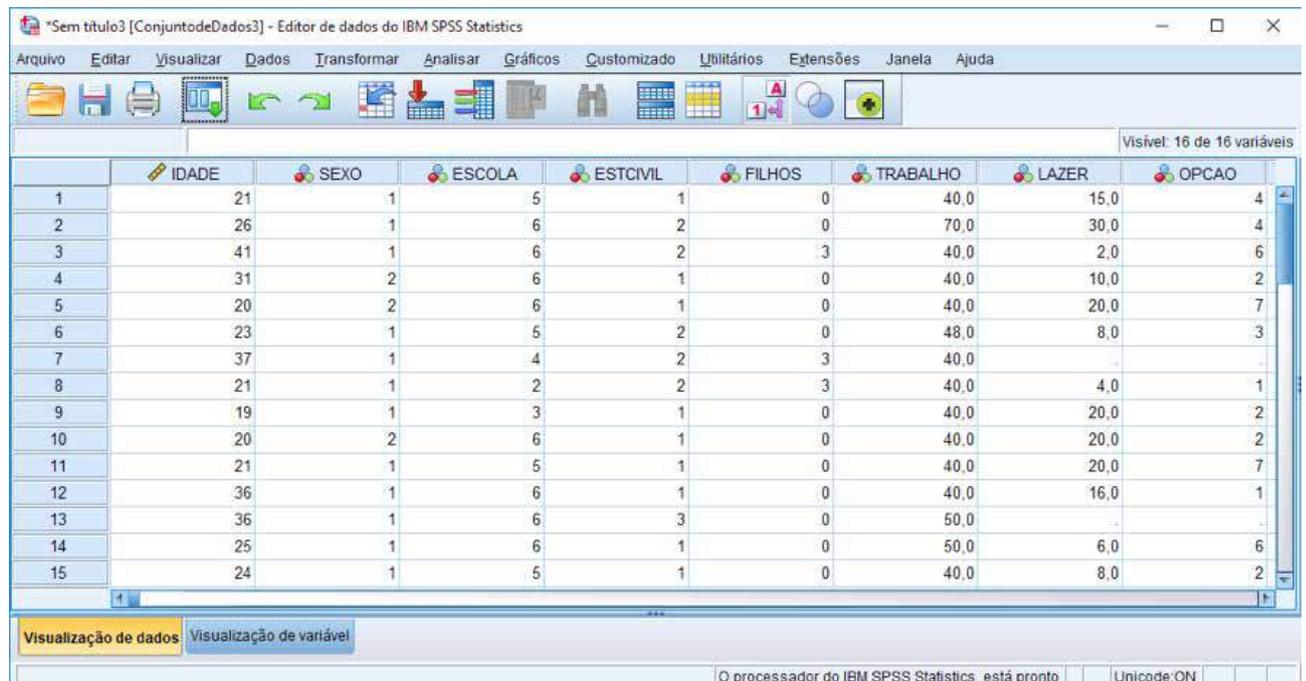


Figura 38 | Resultado final da importação do arquivo Lazer.xlsx



Em termos de exportação, os dados da janela de dados podem ser armazenados sob diversos formatos, tenha ele sido gerado através do IBM SPSS Statistics ou de qualquer outro aplicativo; para isso, a partir do *dataset* no qual o arquivo (*pesquisa.csv* ou *Lazer.xlsx*) foi importado, entramos no menu Arquivo e clicamos em **Salvar** ou **Salvar Como...** Aparecerá uma tela (ver Figura 39) semelhante à janela para abertura de arquivos, onde especificamos novamente disco e subdiretórios, o nome do arquivo, e a forma sob a qual o arquivo deve ser salvo abrindo o menu **Salvar como tipo** (canto inferior esquerdo). O padrão é o formato IBM SPSS Statistics (extensão *.sav*).

Figura 39 | Janela para salvar arquivos de dados



Outra opção, por exemplo, é salvar o arquivo no formato do Microsoft Excel (extensões *.xls* ou *.xlsx*, que podem ser escolhidas no menu **Salvar como tipo**), com a opção de salvar apenas os valores numéricos ou apenas seus rótulos (no caso de variáveis categóricas): basta selecionar a caixa **Salvar rótulos de valor onde definido em vez dos valores de dados**. Esta última opção é a ideal para a conversão do formato SPSS para o Excel, uma vez que as variáveis categóricas serão salvas apenas com seus rótulos e não com sua codificação numérica. Alternativamente, você pode utilizar a opção **Exportar** do menu **Arquivo** para salvar diretamente os dados em um formato específico da sua preferência.

Para maiores detalhes sobre importação e exportação de bancos de dados, bem como mais detalhes das outras opções do Assistente de Importação de Texto, consulte a **Ajuda** do IBM SPSS Statistics.

## 03 Transformação de dados

### 03.1 Recodificação utilizando o procedimento Categorização Visual

Utilizaremos aqui um recurso muitas vezes necessário e que antecede a análise estatística propriamente dita. A recodificação de variáveis pode ser usada, entre outras coisas, para criar faixas de valores baseadas em variáveis já existentes. Um exemplo típico é ter interesse em estudar a faixa etária das pessoas em um estudo onde a informação disponível originalmente é a idade. Para isso, vamos abrir o arquivo *Lazer.sav*.

No nosso exemplo, criaremos essa variável representando a faixa etária a partir da variável idade. Nossa ideia é construir tal variável de acordo com a seguinte divisão por faixas:

Valores Atuais (Idade)	Valores Futuros (Faixa Etária)
<= 25 anos	1
De 26 a 35 anos	2
De 36 a 45 anos	3
> 45 anos	4

Para fazer uma recodificação no IBM SPSS Statistics, entre no menu **Transformar** e escolha a opção **Categorização Visual (Visual Binning)**. O comando **Categorização Visual** foi criado com o objetivo de auxiliar o processo de criação de novas variáveis em categorias (faixa de valores) baseadas no agrupamento de variáveis numéricas (escalares ou ordinais) já existentes.

No primeiro passo da **Categorização Visual**, devemos informar uma variável numérica a ser recodificada (no caso, *idade*). Para isso, basta clicar na variável *idade* e movê-la, clicando na seta logo ao lado, para o quadro **Variáveis para categorização** (ver Figura 40).

Opcionalmente, podemos limitar a quantidade de casos que serão lidos pelo IBM SPSS Statistics utilizando a opção **Número limite de casos verificados para:** Clique em **Continuar**.

Devemos, também, dar um nome para a nova variável que será criada através do quadro **Variável categorizada** no canto superior direito. No espaço reservado ao nome da nova variável (**Nome**), digitamos *faixaet*. Podemos ainda criar um rótulo para a nova variável *faixaet* (por exemplo, "Faixa Etária"), preenchendo o quadro **Rótulo** ao lado do nome da variável (ver Figura 41).

Figura 40 | Seleção de variáveis para recodificação

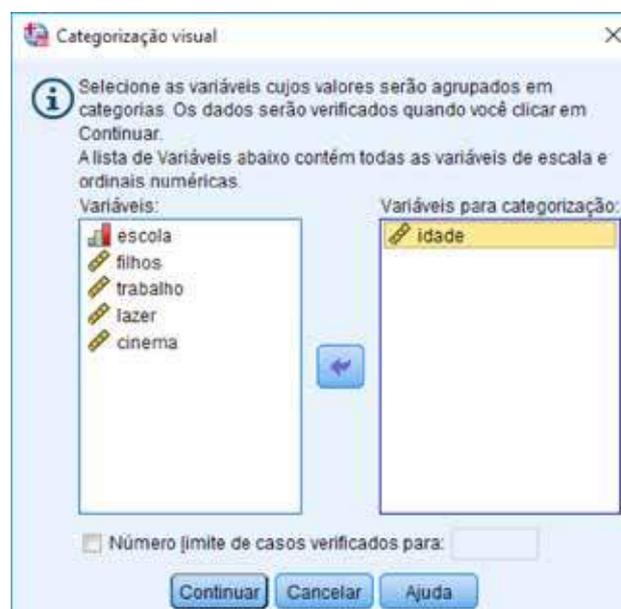


Figura 41 | Recodificação da variável IDADE

**Categorização visual**

Lista de variável verificada: idade

Variável atual: idade Nome: idade Rótulo: Idade dos Entrevistados

Variável categorizada: Variável categorizada: Idade dos Entrevistados (Categorizado)

Mínimo: 14 Valores não omissos Máximo: 60

Insira pontos de corte de intervalo ou clique em Criar pontos de corte para intervalos automáticos. Um valor de ponto de corte de 10, por exemplo, define um intervalo começando acima do intervalo anterior e terminando em 10.

Grade:

	Valor	Rótulo
1		HIGH
2		

Casos verificados: 31

Valores omissos: 0

Copiar categorias

De outra variável...

Para outras variáveis...

Terminais superiores

Incluído (<=)

Excluído (<)

Criar pontos de corte...

Criar rótulos

Inverter escala

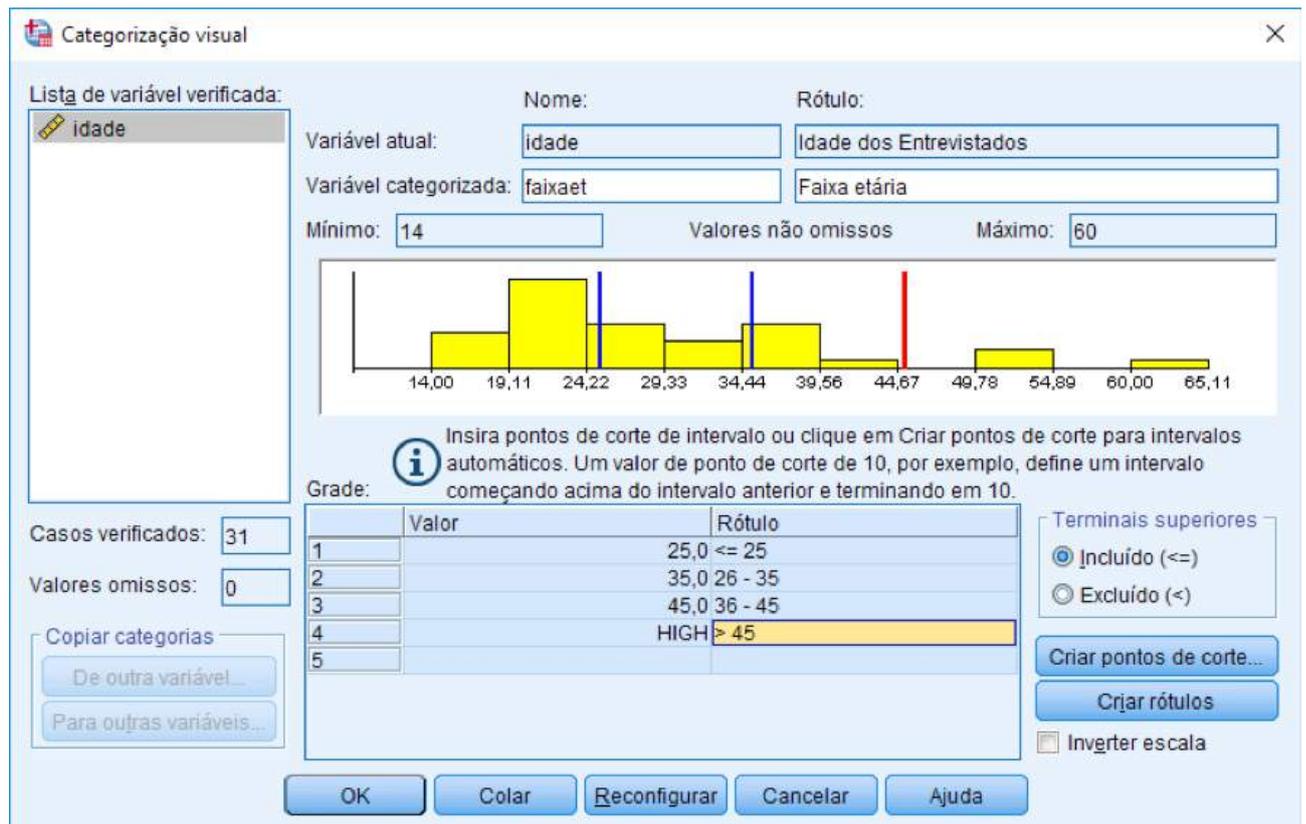
OK Colar Reconfigurar Cancelar Ajuda

Falta agora informar ao IBM SPSS Statistics as condições da recodificação. No quadro **Grade**, temos que especificar os valores da antiga variável que determinarão os valores da nova variável, para isso digitamos os valores finais de cada faixa (ponto de corte) no campo **Valor** (como padrão, o ponto de corte **HIGH** é automaticamente incluído). A opção **Terminais superiores** deverá ser utilizada em conjunto com o campo Valor para determinar a faixa que o ponto de corte será colocado. Por exemplo: Se colocarmos o ponto de corte 25 na primeira faixa e escolhermos a opção **Incluído (<=)**, a primeira faixa conterá todos os casos menores ou iguais a 25. Caso escolhêssemos a opção **Excluído (<)**, a primeira faixa corresponderia a todos os números estritamente menores do que 25; portanto, o valor 25 seria incluído na próxima faixa.

No nosso caso, entraremos com os valores (ponto de corte) logo abaixo do valor padrão **HIGH** (que corresponde a todos os valores acima do último ponto de corte inserido). Certifique-se que a opção **Incluído (<=)** esteja selecionada. Na posição 1 da coluna **Valor**, (quadro **Grade**), digitamos o valor 25, na posição 2 digitamos o valor 35 e na posição 3 digitamos o valor 45. Repare que a resposta **HIGH**, a cada novo valor que é inserido, é colocado na última posição.

Após a digitação dos valores (ponto de corte) de cada faixa, podemos clicar em **Criar Rótulos** para que o IBM SPSS Statistics insira os rótulos de cada faixa de valores automaticamente (ver Figura III-3). Mesmo utilizando os rótulos automáticos, eles poderão ser editados a qualquer momento, bastando para isso editar a coluna **Rótulo** com os rótulos de sua preferência.

Figura 42 | Rótulos inseridos automaticamente através da opção Criar Rótulos



Desta maneira, a definição da nova variável *faixaet* está encerrada. Clicando em **OK**, o IBM SPSS Statistics exibirá um aviso que indica a quantidade de variáveis que serão criadas pela **Categorização Visual** - basta clicar em **OK** para que o comando seja concluído. A nova variável passa a existir e aparece na última coluna do arquivo. Ela assume os valores 1, 2, 3 e 4 com os respectivos rótulos associados a estes valores.

### 03.2 Recodificação automática de variáveis alfanuméricas (string)

Outro recurso para recodificação presente no IBM SPSS Statistics é a **Recodificação Automática (Automatic Recode)**. Através deste procedimento, podemos transformar variáveis do tipo alfanuméricas (*string*) ou numéricas em variáveis numéricas consecutivas já codificadas (com rótulos associados). O IBM SPSS Statistics ordena em ordem alfabética (crescente ou decrescente) os valores da variável selecionada, relacionando o número 1 ao primeiro valor e assim por diante, até o último valor. A nova variável será numérica, assumindo desde o valor 1 até o número total de valores (ou cadeia de caracteres) existentes, além de assumir os antigos valores (alfanuméricos ou numéricos originais) como rótulos da nova variável numérica.

Por exemplo: suponha que desejamos transformar a variável *cidade* (que é alfanumérica) do nosso banco de dados *Lazer.sav* numa variável numérica. Se quisermos fazer determinadas análises usando a cidade de origem, teremos que executar esse procedimento.

Para isso, basta usar a opção **Recodificação Automática** presente no menu **Transformar**. Uma janela (ver Figura 43) é aberta e a primeira coisa a fazer é mover a variável a ser recodificada para o quadro central **Variável** → **Novo nome**. No nosso caso, moveremos a variável *cidade*. Como outra coluna será criada para armazenar a nova variável, é preciso dar um nome para essa coluna - digamos, *cidade2*. Devemos colocar essa informação no retângulo logo abaixo do quadro central e clicar em **Incluir Novo Nome**.

Devemos definir também a ordem da recodificação, que pode ser em ordem alfabética crescente ou decrescente, usando uma das opções: **Recodificar iniciando em: Valor mais baixo** (ordem crescente) ou **Valor mais alto** (ordem decrescente). No nosso exemplo, manteremos o padrão **Valor mais baixo**.

Com a opção **Usar o mesmo esquema de recodificação para todas as variáveis**, podemos aproveitar a mesma recodificação para outras variáveis (todas as variáveis selecionadas ficarão com os mesmos rótulos de valores). Com a opção **Tratar valores da sequência de caracteres em branco como omissos de usuário**, os valores brancos ou nulos nas variáveis do tipo *string* serão tratados como *missing* e não entrarão na recodificação. Além disso, podemos salvar um modelo (*template*) da recodificação (opção **Salvar modelo como**) e aproveitá-lo em outras recodificações (opção **Aplicar modelo como**). Finalmente, clique em **OK**.

Feito isso, o resultado da recodificação será mostrado na janela de saídas (ver Figura 44). Note que, antes da recodificação, a variável *cidade* assumia os valores 'São Paulo', 'Rio de Janeiro' e 'Brasília'. Após a recodificação, a nova variável criada (*cidade2*) assume os valores 1, 2 e 3 já com os rótulos relacionados (lado direito).

Figura 43 | Recodificação automática de variáveis do tipo string

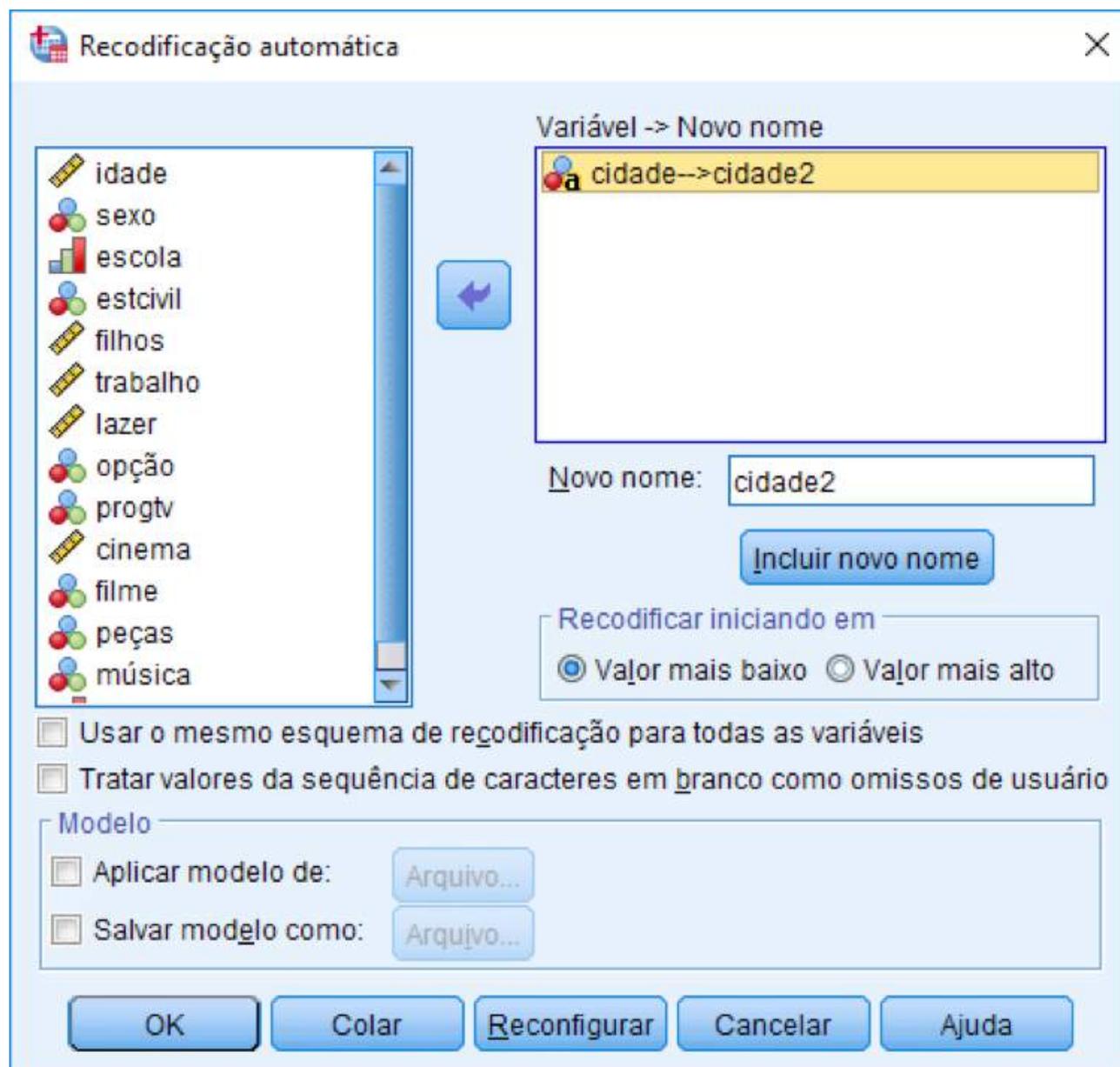


Figura 44 | Resultado da Recodificação Automática

cidade into cidade2 (Cidade dos Entrevistados)		
Old Value	New Value	Value Label
Brasília	1	Brasília
Rio de Janeiro	2	Rio de Janeiro
São Paulo	3	São Paulo

### 03.3 Cálculo de novas variáveis usando o procedimento Calcular Variável

Vamos agora estudar outro recurso de transformação de dados: a criação de uma nova variável como função de variáveis já existentes através do comando **Calcular Variável** (Compute Variable). Com este comando, é possível criar variáveis numéricas ou alfanuméricas (*string*) baseadas em outras variáveis da base de dados. Por exemplo, podemos criar uma variável que traga apenas o primeiro nome de um cliente (caso exista a variável nome na base de dados) ou até mesmo calcular o percentual de aumento a partir de variáveis que indiquem salário anterior e salário atual, entre outros. Consulte a **Ajuda** do IBM SPSS Statistics para conhecer todas as funções existentes no comando **Calcular Variável**.

No nosso exemplo, vamos utilizar as variáveis *trabalho* e *lazer*, presentes no arquivo *Lazer.sav*. Nosso interesse é investigar qual a porcentagem de lazer em relação às horas gastas com trabalho mais lazer. Por exemplo, se a pessoa trabalha 60 horas por semana e gasta 20 horas com lazer, ela gasta 25% do tempo dela de trabalho ou lazer, por semana, com lazer. Para chegar a este número, a operação que fizemos foi muito simples: somam-se as horas semanais de trabalho e lazer (*trabalho+lazer*); divide-se o número de horas semanais de lazer por essa soma (*lazer/(trabalho+lazer)*) e multiplica-se o resultado por 100.

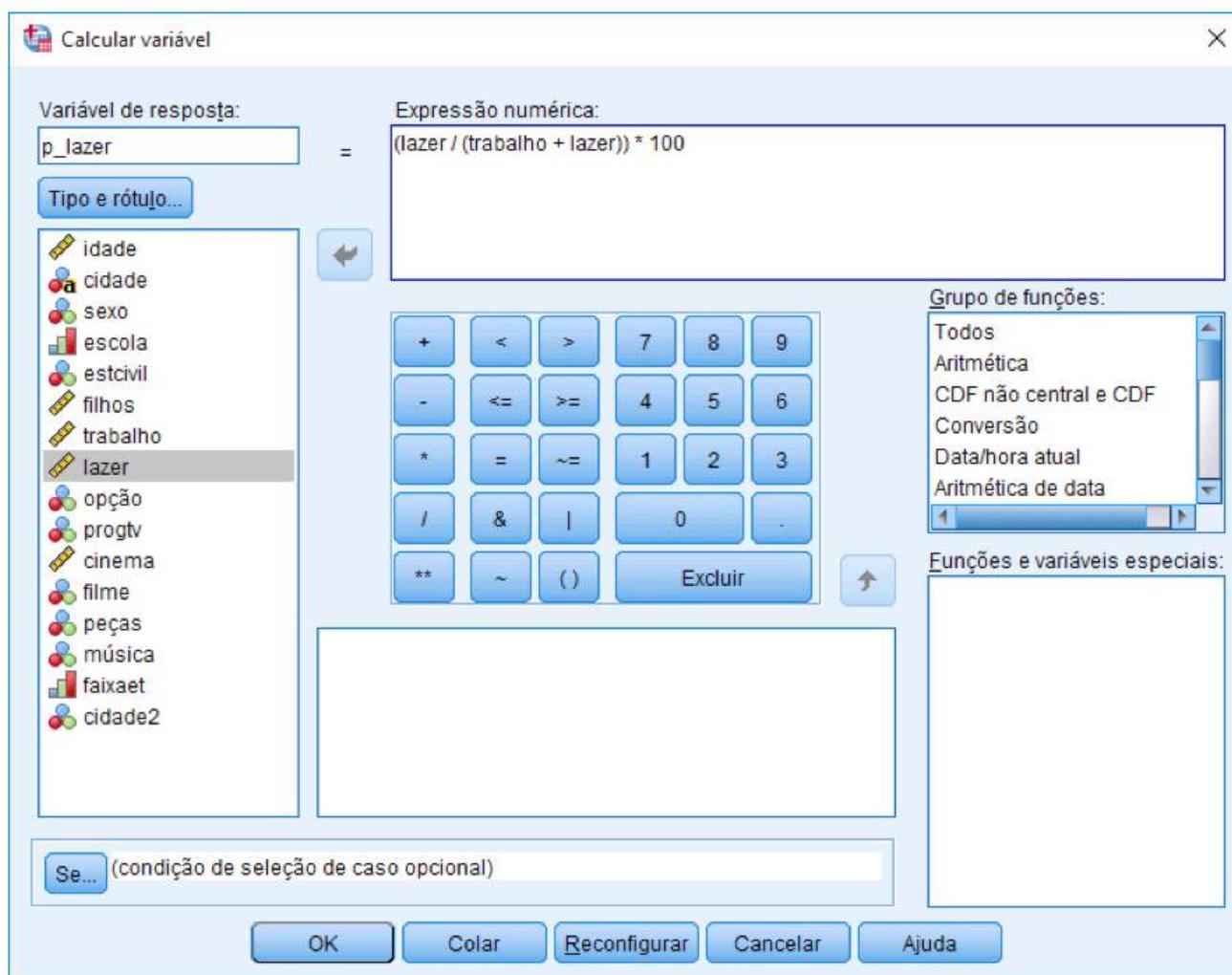
Nosso objetivo agora é criar no banco de dados uma nova variável, com nome *p\_lazer*, que contenha a informação desejada. Para isso, basta entrar no menu **Transformar** e clicar na opção **Calcular Variável**. Uma janela (ver Figura 45) é aberta com as seguintes características: um quadro no canto esquerdo apresenta uma lista com todas as variáveis existentes no banco de dados que podem ser usadas para o cálculo da nova variável, para selecioná-las basta clicar a seta ao lado; um quadro no canto superior esquerdo designado por **Variável de resposta**, onde digitaremos o nome da variável a ser criada; um quadro à direita na parte superior designado por **Expressão numérica**, onde devemos escrever a função (envolvendo as variáveis já existentes no arquivo) que permitirá o cálculo da nova variável numérica ou *string*.

No quadro central, existe uma série de botões que contêm desde funções matemáticas (+, -, \*, etc.) até funções lógicas (&, |, etc.); e, no quadro inferior direito, as funções divididas em **Grupo de funções e Funções e variáveis especiais**: o primeiro mostra uma lista de grupos de funções semelhantes e o segundo lista todas as funções existentes no IBM SPSS Statistics. Há uma grande quantidade de funções, desde soma e multiplicação até arco-tangente, logaritmo neperiano, entre outras que também podem ser utilizadas na construção de qualquer expressão numérica. Para as variáveis alfanuméricas (*string*), o comando **Calcular Variável** oferece funções para concatenar, substituir e outras.

Também, há a opção de fazer o cálculo apenas se determinada condição for satisfeita (a partir do botão *Se...*); discutiremos a seleção de casos com mais detalhes no Capítulo V. Caso a condição não for verificada para algum caso, o valor da nova variável para esse caso será um *system missing* (se for numérica) ou uma célula vazia (se for alfanumérica).

No nosso caso, a expressão a ser criada é  $(lazer/(trabalho+lazer))*100$ . Selecionamos a variável *lazer* do quadro esquerdo, clicamos o símbolo / presente no quadro central da janela; em seguida clicamos no símbolo ( ), selecionamos a variável *lazer*, o símbolo + e a variável *trabalho*; agora basta clicar o símbolo \* e o número 100. Poderíamos, ainda, ter escrito a expressão digitando diretamente no quadro **Expressão numérica**. A Figura III-6 mostra o resultado final da digitação na janela.

Figura 45 | Menu Transformar → Calcular Variável

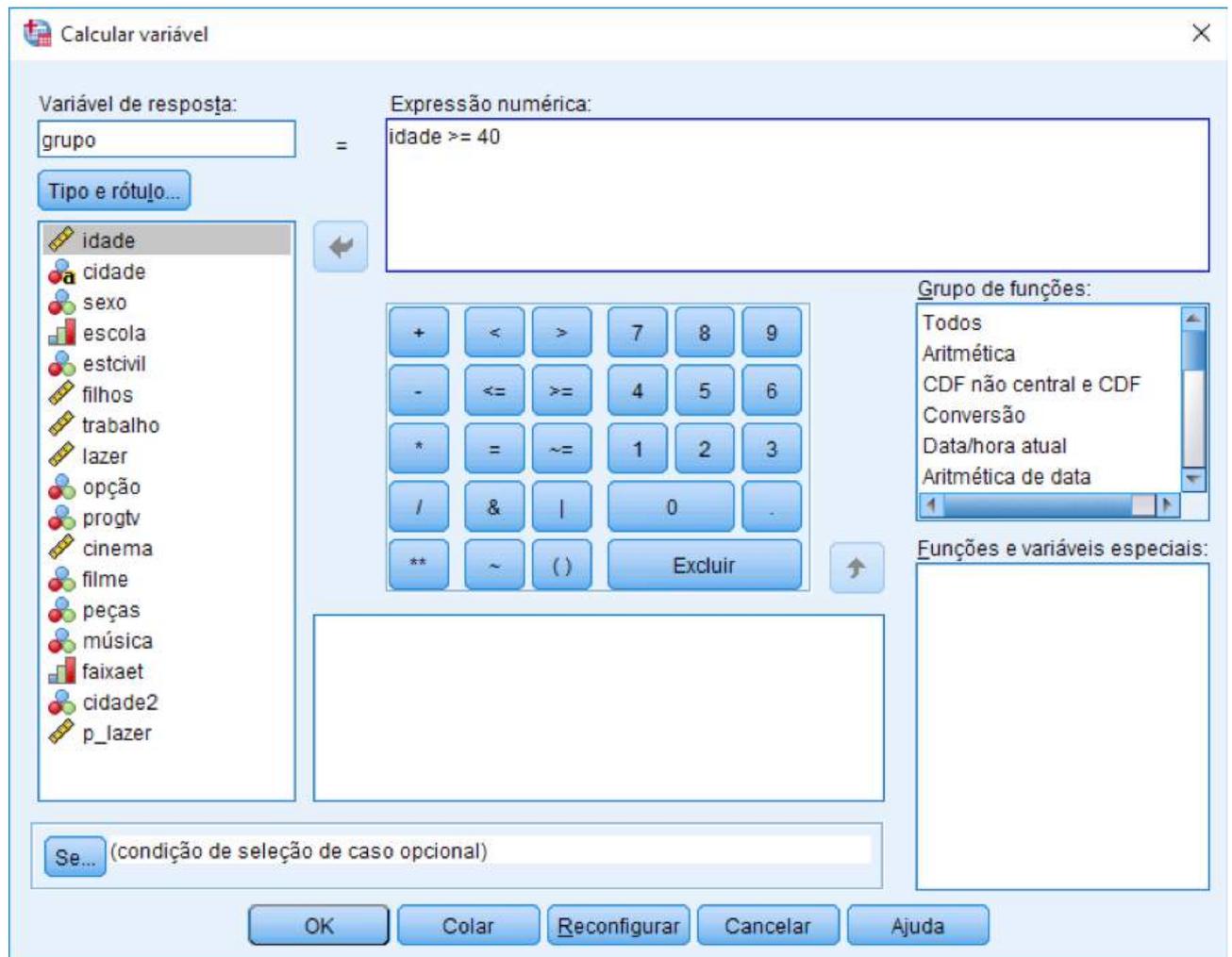


Vale observar que os símbolos especiais \*\* (potência), > (“maior que”), < (“menor que”), >= (“maior ou igual a”), <= (“menor ou igual a”) e ~= (“diferente de”) também podem ser utilizados para expressões numéricas, seja para cálculos aritméticos ou para comparação de valores numéricos.

Agora que nossa função já está pronta, devemos dar um nome à nova variável. Digitamos *p\_lazer* no quadro superior esquerdo **Variável de resposta**, e clicamos em OK (ver Figura III-6). Uma nova variável com nome *p\_lazer* será adicionada à janela de dados na primeira coluna vazia.

O comando **Calcular Variável** também pode ser utilizado para criar novas variáveis a partir de sentenças condicionais. Poderíamos, por exemplo, criar uma nova variável para identificar se a pessoa entrevistada tem pelo menos 40 anos de idade. Para isso, acesse novamente o comando via menu **Transformar**, clique no botão **Reconfigurar (Reset)** para limpar todos os campos, digite a condição **idade >= 40** no campo **Expressão numérica** e, finalmente, digite o nome grupo no campo **Variável de resposta** (ver Figura III-7). A seguir, clique em **OK** e repare que a nova variável foi criada e está no final do arquivo *Lazer.sav*.

Figura 46 | Criação da variável GRUPO via Calcular Variável



No IBM SPSS Statistics, uma sentença condicional recebe o valor numérico 1 (um) se for verdadeira (TRUE) ou o valor numérico 0 (zero) se for falsa (FALSE). Sendo assim, se um indivíduo tiver menos de 40 anos de idade, seu valor para a variável idoso será 0. No nosso exemplo, apenas um indivíduo recebeu o valor 1 para idoso. Vale observar que, caso o valor de uma variável envolvida em determinada condição seja **system missing**, o comando **Calcular Variável** criará um valor **system missing** para o(s) respectivo(s) caso(s) com valores ausentes. O resultado final da criação das duas variáveis (*p\_lazer* e *grupo*) está na Figura 47.

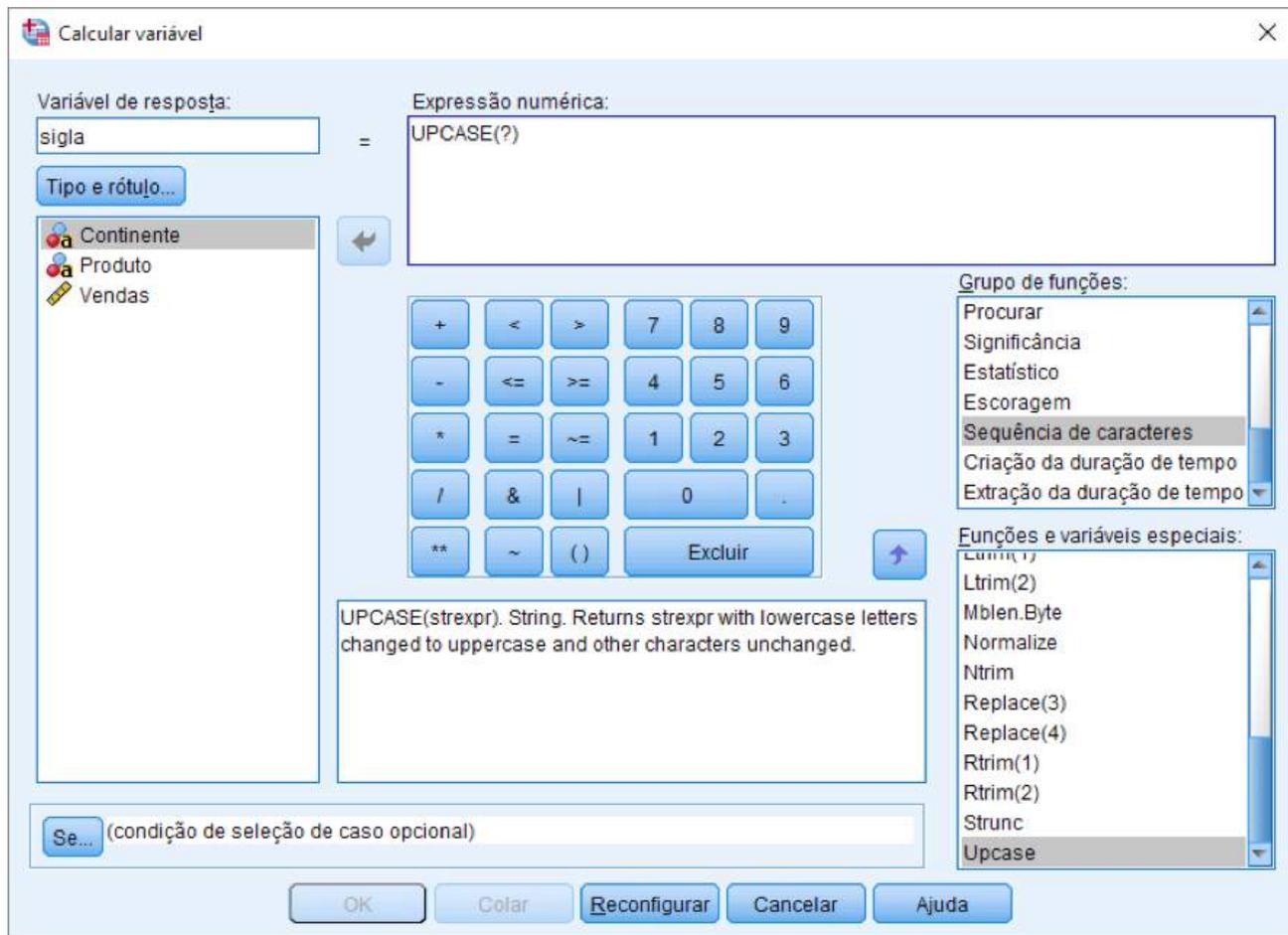
Como último exemplo desta seção, trabalharemos com variáveis alfanuméricas (*string*) a partir de duas funções simples, mas de relativa importância para quem manipula sequências de caracteres com frequência. Para tanto, vamos utilizar o arquivo de dados *Worldsales.sav*, que possui uma variável que identifica em qual continente cada um dos três tipos de produto (A, B ou C) foi comprado – no caso, a variável *Continente*.

Suponhamos que haja interesse em criar uma variável que represente uma sigla de três letras para cada continente, a partir das primeiras três letras de cada nome; portanto, a Europa ('Europe') receberia a sigla EUR; a América do Sul ('South America') receberia a sigla SOU, e assim por diante. Repare também que as três letras que formam cada valor desta variável *sigla* devem estar em **caixa alta** (upper case).

Para isso, usaremos novamente o comando **Calcular Variável**. Digite o nome *sigla* no campo **Variável de resposta**. A seguir, escolha a opção **Sequência de caracteres** na caixa **Grupo de funções** e clique duas vezes na função **Uppcase**, localizada dentro da caixa **Funções e variáveis especiais** (veja Figura 47). A função UPCASE converte todas as letras minúsculas de seu argumento (que deve ser uma variável alfanumérica) em maiúsculas. Repare que há uma

breve explicação sobre o comando no canto inferior da janela e um ponto de interrogação entre os parênteses, do lado direito da função UPCASE, indicando que apenas um argumento é necessário.

Figura 47 | Criação da variável SIGLA via Calcular Variável – Passo 1



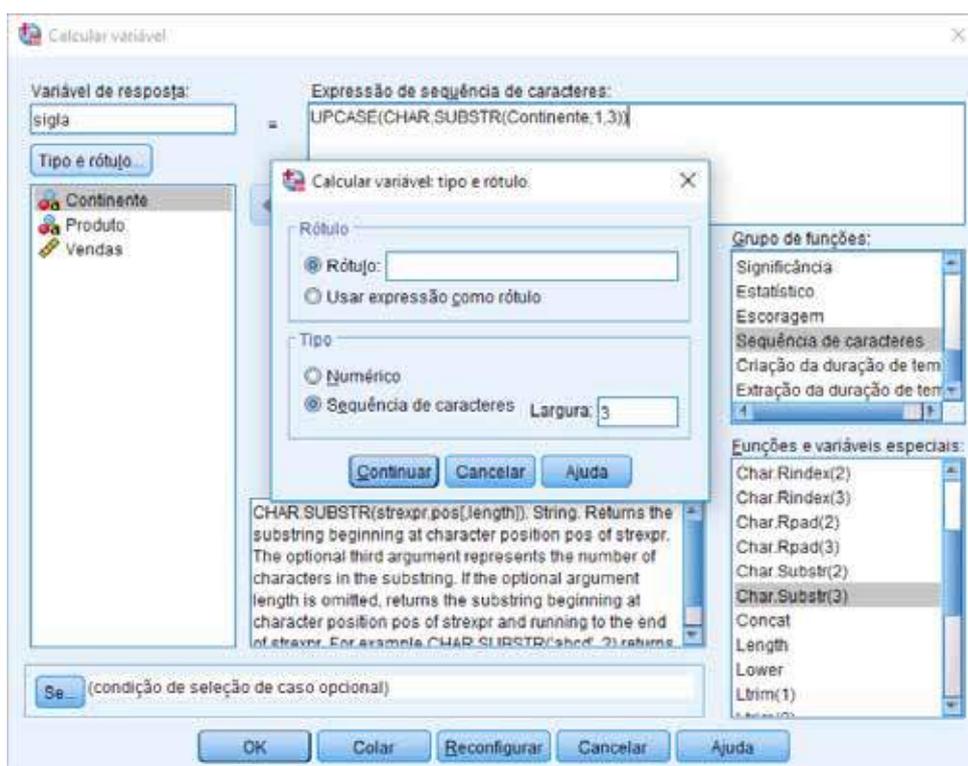
Agora, vamos escolher a função que “corta” a variável da forma que queremos: entre os parênteses da função UPCASE, elimine o ponto de interrogação e inclua a função **Char.Substr** (3) – clique duas vezes no nome da função, localizado dentro da caixa **Funções e variáveis especiais** (ver Figura 48). Repare que a função CHAR.SUBSTR que selecionamos requer três argumentos: a variável alfanumérica da qual extrairemos uma subsequência de caracteres; a posição inicial dessa subsequência e o tamanho desta subsequência.



Como retiraremos as primeiras três letras da variável *Continente*, nada mais natural que estes três argumentos sejam, nesta ordem: "Continente" (nome da variável), "1" (posição inicial da subsequência) e "3" (tamanho da subsequência). Na Figura 50, mostramos como os argumentos devem ser colocados na função.

Antes de clicar em OK, assegure-se de que a opção **Sequência de caracteres** esteja marcada na tela do botão **Tipo e Rótulo**, bem como o tamanho máximo da nova variável (*Largura*) esteja ajustado para qualquer valor maior ou igual a 3. Clique em **Continuar** e, depois, em **OK**. O resultado final para a variável criada está na Figura 51.

Figura 49 | Criação da variável SIGLA via Calcular Variável – Passo final



Para mais detalhes sobre funções de variáveis alfanuméricas (string functions) e como funciona a codificação de arquivos dentro do IBM SPSS Statistics, consulte a ajuda do IBM SPSS Statistics ou a seção **String Functions** da **Referência de Sintaxe do Comando**, localizada dentro do menu **Ajuda**.

Figura 50 | Arquivo Lazer.sav com as duas variáveis criadas via Calcular Variável

	ogtv	cinema	filme	peças	música	faixaet	cidade2	p_lazer	grupo
1	telejornais	0	comédia	comédia	mpb	> 45	São Paulo	17,65	1,00
2	filmes	0	terror	.	rock	36 - 45	São Paulo	.	,00
3	entrevista	10	suspense	comédia	rock	<= 25	São Paulo	25,37	,00
4	entrevista	2	comédia	comédia	rock	<= 25	São Paulo	25,00	,00
5	novelas	1	comédia	comédia	samba	36 - 45	São Paulo	.	,00
6	filmes	1	aventura	musical	clássica	26 - 35	São Paulo	20,00	,00
7	filmes	4	suspense	comédia	outros	<= 25	São Paulo	33,33	,00

Figura 51 | Arquivo Worldsales.sav com a variável criada (SIGLA) via Calcular Variável

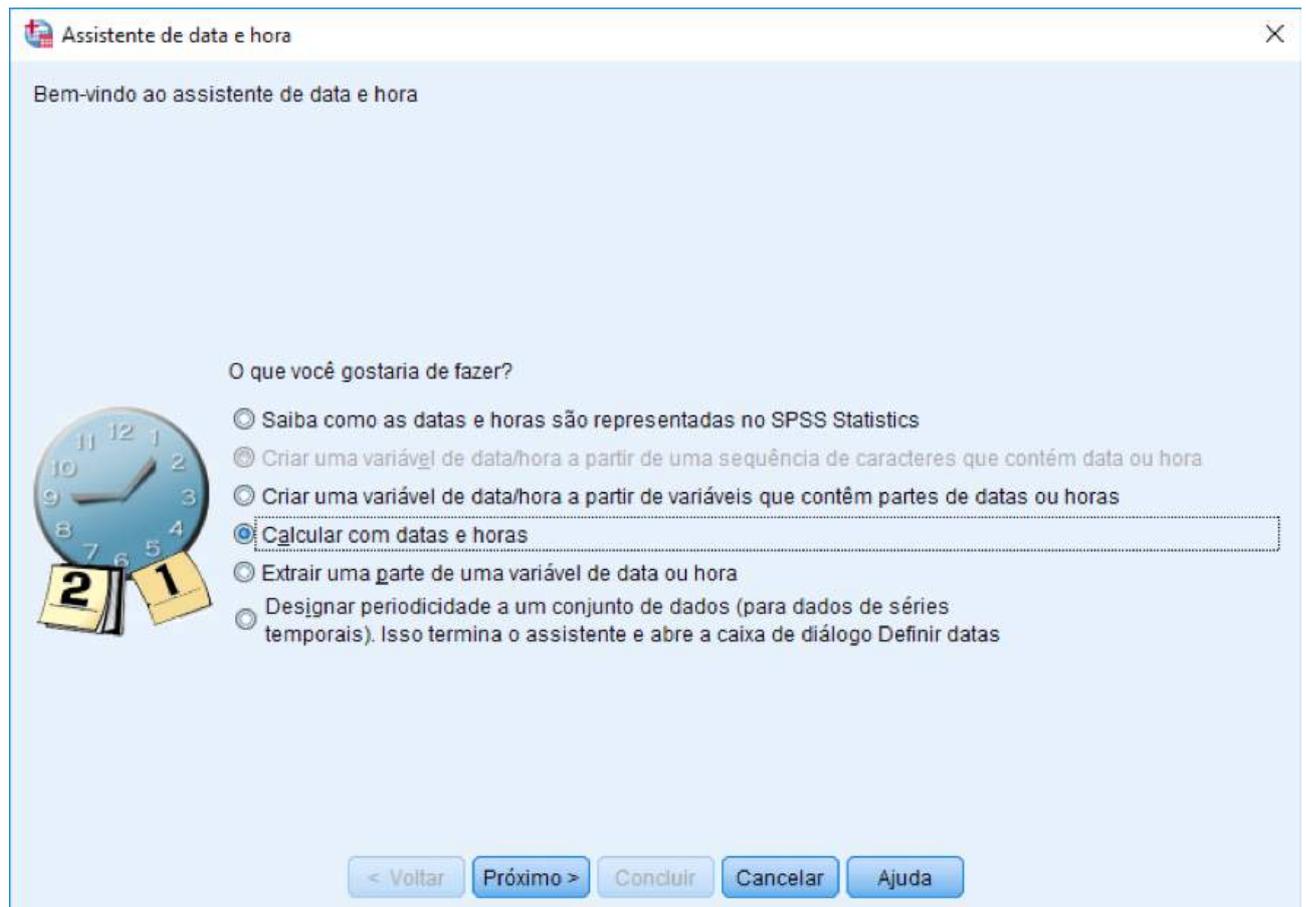
	Continente	Produto	Vendas	sigla	var	var	var	var	var	var
1	South America	Produto C	719,10	SOU						
2	Asia	Produto C	833,56	ASI						
3	South America	Produto A	468,49	SOU						
4	Australia	Produto A	873,52	AUS						
5	South America	Produto A	543,55	SOU						
6	Europe	Produto A	941,46	EUR						
7	Australia	Produto B	1277,05	AUS						
8	Africa	Produto B	534,50	AFR						
9	Asia	Produto C	789,00	ASI						
10	South America	Produto B	412,86	SOU						
11	Europe	Produto B	752,55	EUR						
12	Africa	Produto B	340,30	AFR						
13	Africa	Produto A	647,34	AFR						

### 03.4 Cálculo de datas através do Assistente de Data e Hora

Outro recurso bem interessante no IBM SPSS Statistics é o **Assistente de Data e Hora (Date and Time Wizard)**, um assistente que simplifica a criação de novas variáveis e os cálculos envolvendo variáveis no formato data. No nosso exemplo, utilizaremos o arquivo *Banco.sav*.

Para acessar o assistente, basta entrar no menu Transformar e clicar na opção **Assistente de Data e Hora** (ver Figura 52). Na tela inicial, você poderá escolher o que você deseja realizar e, na sequência, seguir os passos indicados para concluir a tarefa.

Figura 52 | Assistente de Data e Hora – Tela inicial



Utilizando esse assistente, é possível aprender como as datas e as horas são representadas no IBM SPSS Statistics; criar uma variável no formato data/hora a partir de uma variável alfanumérica (*string*) ou numérica que contenha data ou hora; realizar cálculos envolvendo data ou hora; extrair parte de uma data ou hora, ou criar uma variável que indique periodicidade, elemento fundamental nas análises de Séries Temporais.

Uma das tarefas mais comuns envolvendo variáveis de data é o cálculo da idade a partir da data de nascimento da pessoa. O arquivo *Banco.sav* contém diversas informações de funcionários de uma empresa, incluindo a data de nascimento, e o nosso objetivo nesse caso é calcular a idade de cada funcionário utilizando o **Assistente de Data e Hora**.

Para isso, basta entrar no menu **Transformar** e clicar na opção **Assistente de Data e Hora**. Na janela inicial do assistente, escolheremos a opção **Calcular com datas e horas** e, em seguida, clicaremos em **Próximo**. No Passo 1, escolhemos a segunda opção, **Calcular o número de unidades de tempo entre duas datas** (ver Figura 53), e, a seguir, clicamos em **Próximo**.

Na sequência, no Passo 2, devemos informar ao IBM SPSS Statistics as duas variáveis que serão utilizadas para calcular a idade e, também, a unidade de duração que você precisa no resultado. No nosso caso, utilizaremos a data de nascimento *datanasc* (que será colocada no campo **menos Data2**) e, para representar a data atual (de acordo com o que está ajustado no seu computador), utilizaremos a variável *\$TIME* (que será colocada no campo **Data1**). Os resultados podem ser representados em unidades de anos, meses, dias, horas, minutos ou segundos; vamos escolher a opção **Anos** na área Unidade (ver Figura 54). A seguir, clicamos em **Próximo**.

Figura 53 | Assistente de Data e Hora – Passo 1

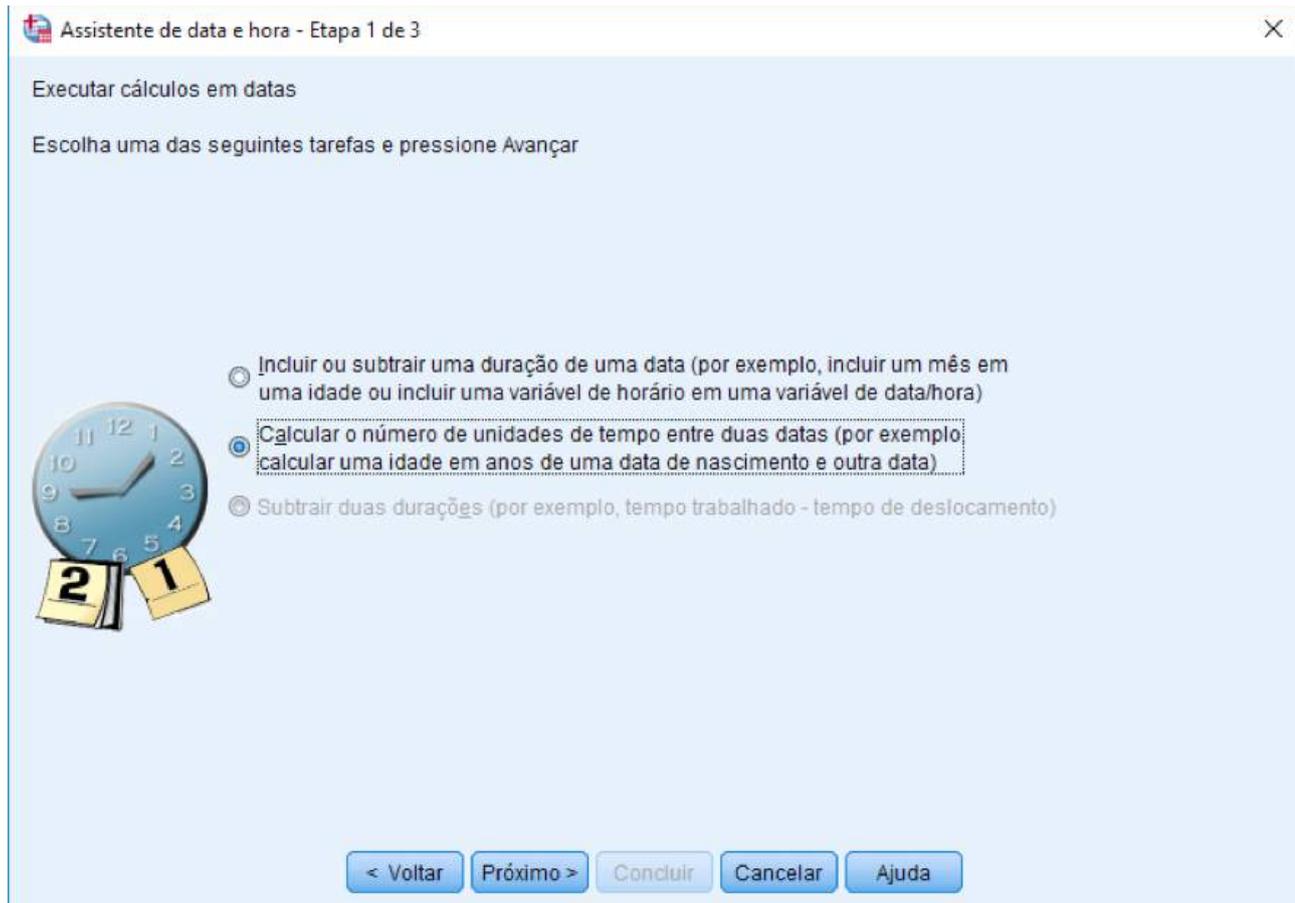
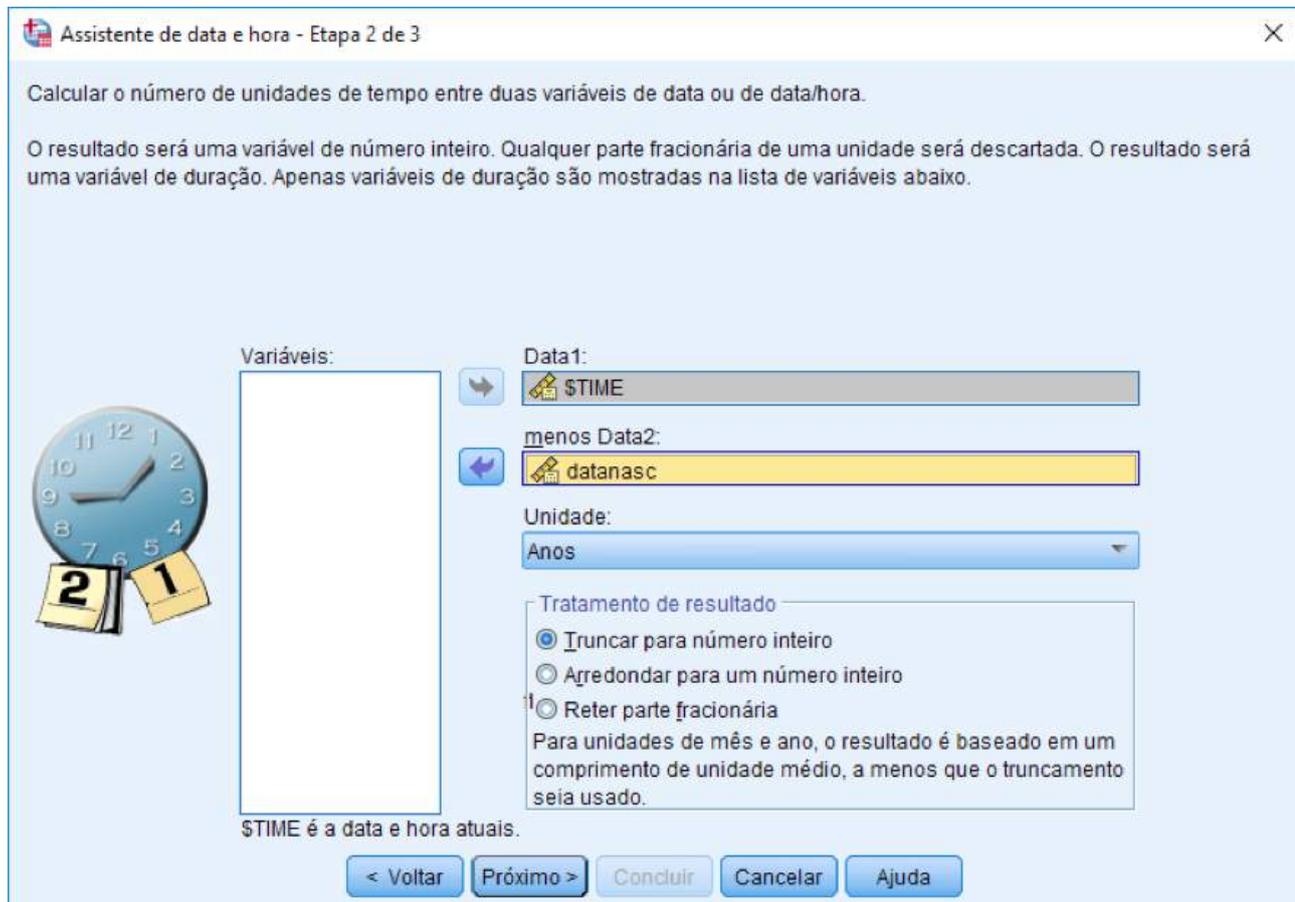


Figura 54 | Assistente de Data e Hora – Passo 2



No Passo 3, entre com o nome da variável que indicará o resultado - nesse caso *idade* - e, caso você queira, também pode ser definido um rótulo para a nova variável. Caso você não queira criar a variável na base de dados automaticamente (opção **Criar a variável agora**), você pode escolher a opção **Colar a sintaxe na janela de sintaxe** para criar a sintaxe que representa o comando que acabamos de definir para futuro uso ou modificação (ver Figura 55). Vamos escolher a opção **Criar a variável agora** e em seguida clicar em **Concluir**.

Figura 55 | Assistente de Data e Hora – Passo 3

Assistente de data e hora - Etapa 3 de 3

Cálculo: \$TIME - datanasc

Variável de resultado: idade Unidades: Anos

Rótulo de variáveis: Idade do funcionário

Execução

Criar a variável agora  Colar a sintaxe na janela de sintaxe

< Voltar Próximo > Concluir Cancelar Ajuda

A variável *idade* será criada no final da base de dados (Figura 56). Salve a base de dados com o nome *Banco1.sav*.

Figura 56 | Resultado do cálculo da variável IDADE

	id	datanasc	sexo	estudo	catemp	salário	salarin	temp_ser	exp_ant	minoria	idade
1	1	03.02.1952	Masculino	15	A	\$57,000	\$27,000	98	144	Não	65
2	2	23.05.1958	Masculino	16	C	\$40,200	\$18,750	98	36	Não	59
3	3	26.07.1929	Feminino	12	C	\$21,450	\$12,000	98	381	Não	88
4	4	15.04.1947	Feminino	8	C	\$21,900	\$13,200	98	190	Não	70
5	5	09.02.1955	Masculino	15	C	\$45,000	\$21,000	98	138	Não	62
6	6	22.08.1958	Masculino	15	C	\$32,100	\$13,500	98	67	Não	59
7	7	26.04.1956	Masculino	15	C	\$36,000	\$18,750	98	114	Não	61
8	8	06.05.1966	Feminino	12	C	\$21,900	\$9,750	98	0	Não	51
9	9	23.01.1946	Feminino	15	C	\$27,900	\$12,750	98	115	Não	71
10	10	13.02.1946	Feminino	12	C	\$24,000	\$13,500	98	244	Não	71
11	11	07.02.1950	Feminino	16	C	\$30,300	\$16,500	98	143	Não	67
12	12	11.01.1966	Masculino	8	C	\$28,350	\$12,000	98	26	Sim	51
13	13	17.07.1960	Masculino	15	C	\$27,750	\$14,250	98	34	Sim	57

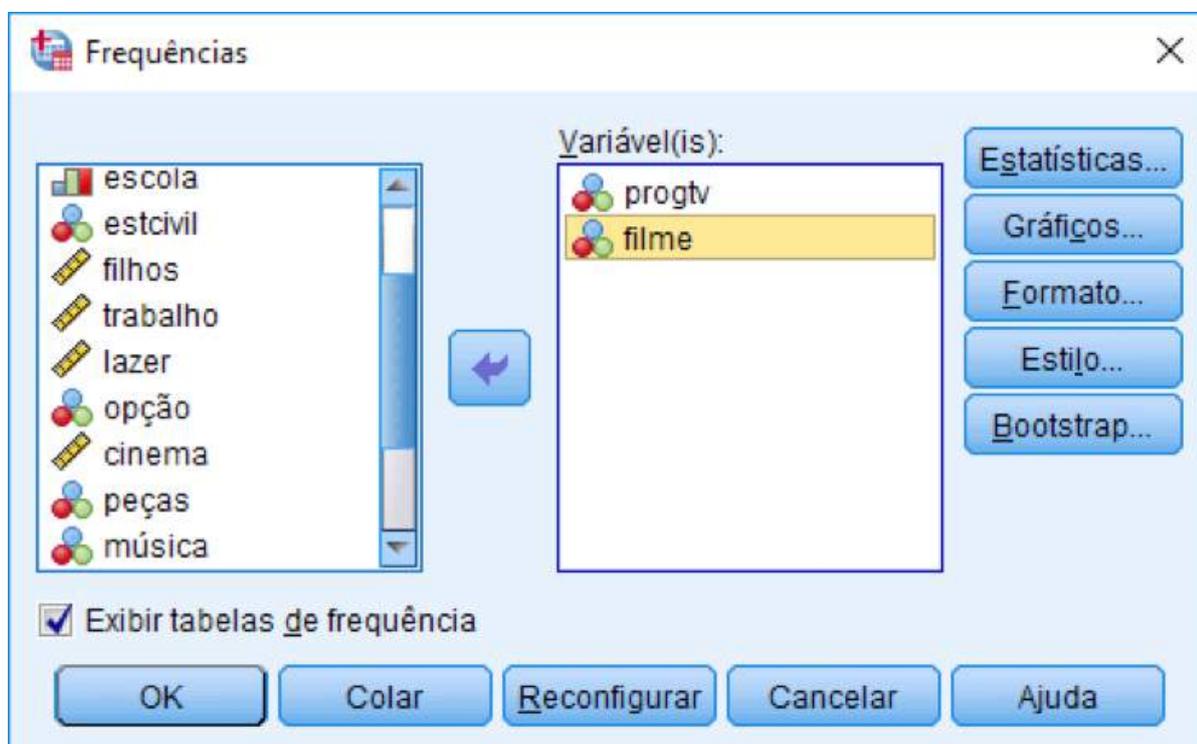
Vale observar, como último comentário deste capítulo, que a variável \$TIME, assim como outras variáveis internas do IBM SPSS Statistics (como \$DATE, \$CASENUM etc.), não podem ser criadas pela aba **Visualização de Variável**, mas podem ser acessadas via Assistente de Data e Hora, como vimos nesta seção, e via programação na janela de sintaxe.

## 04 Comandos do IBM SPSS

### 04.1 Frequências

Para utilizar a primeira opção de **Estatística Descritiva**, clique em **Frequências**. A seguir, aparecerá uma janela onde você pode incluir as variáveis para as quais deseja ter a distribuição de frequências. Vamos selecionar *progtv* (Qual seu tipo de programa de TV favorito?) e *filme* (Qual seu estilo de filme preferido?), ambas do arquivo *Lazer.sav*, como na Figura 57.

Figura 57 | Procedimento Frequências



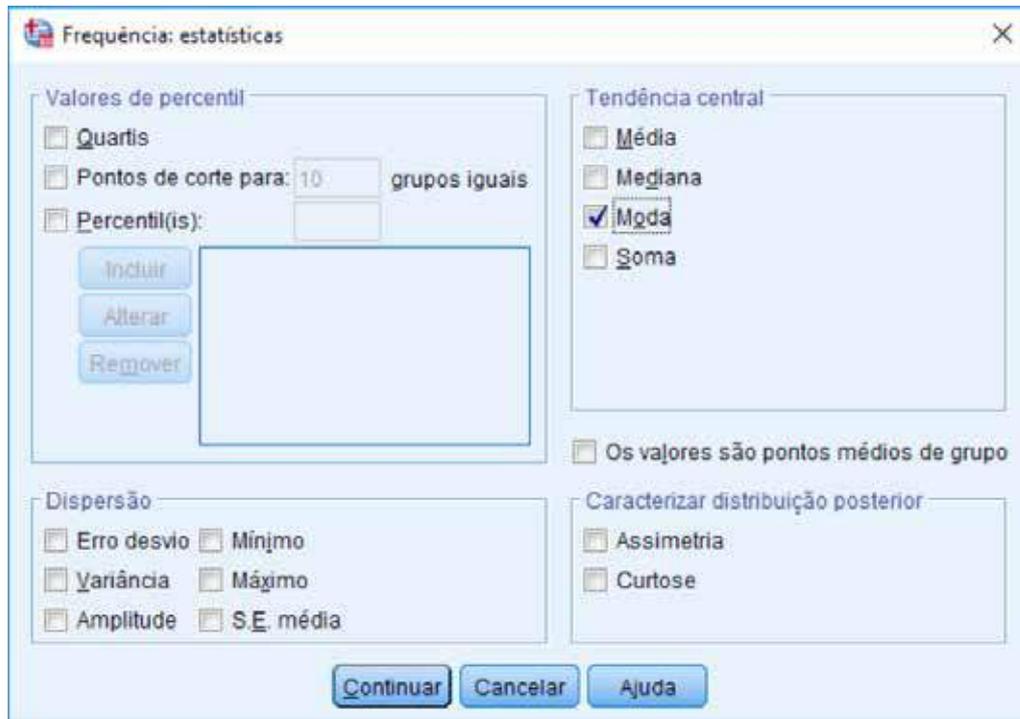
Na parte superior direita da janela, temos um quadro chamado **Estatísticas**. Esse quadro apresenta várias opções de medidas estatísticas (ver Figura 58). Devemos tomar cuidado, uma vez que nem todas as medidas apresentadas trazem informações relevantes para o estudo das variáveis em questão. As medidas estão divididas em quatro grupos:

- **Valores de Percentil** - No canto superior esquerdo, essa opção possibilita o cálculo dos quartis (Quartis), decis (**Pontos de corte para 10 grupos iguais** - você pode mudar o número 10 para qualquer divisão que desejar) ou quaisquer percentis (**Percentil(is)**) de interesse;
- **Tendência Central** - No canto superior direito, esta opção possibilita o cálculo da **Média**, **Mediana**, **Moda** e **Soma**;
- **Dispersão** - No canto inferior esquerdo, essa opção possibilita o cálculo do **Desvio Padrão (Erro Desvio)**, **Variância**, **Amplitude**, valor mínimo (**Mínimo**), valor máximo (**Máximo**) e erro padrão da média (**S.E. Média**);

• **Distribuição** - No canto inferior direito, essa opção possibilita o cálculo de duas medidas de distribuição: **Curtose (Kurtosis)** e **Assimetria (Skewness)**. A curtose é uma medida do achatamento da distribuição e a assimetria é um índice que reflete a simetria da distribuição em torno de um valor central.

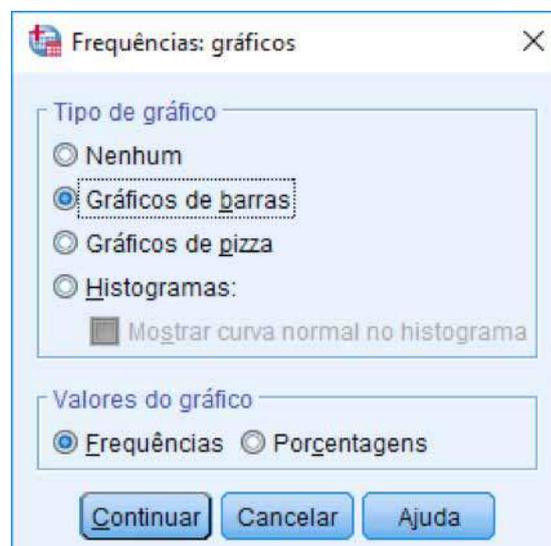
Devemos decidir agora, dentre todas essas medidas, aquelas que são adequadas para as variáveis *progtv* e *filme*. Como ambas são nominais, apenas a moda trará alguma informação relevante. Após selecionar a moda, clique em **Continuar**.

Figura 58 | Opções de medidas estatísticas no procedimento Frequências



Outro quadro que aparece na parte inferior da janela principal é o **Gráficos**. Este quadro apresenta três gráficos como opção: **Histograma** com ou sem uma curva normal, **Gráfico de barras (Bar Chart)** e **Gráfico de pizza (Pie Chart)**, como mostrado na Figura 59. Note que, apesar de apresentar um menu exclusivamente para gráficos (menu **Gráficos**), o IBM SPSS Statistics também possibilita a construção de gráficos dentro de procedimentos estatísticos. Nesse exemplo, vamos escolher a opção **Gráficos de barras** e clicar em **Continuar**.

Figura 59 | Opções de gráficos para o procedimento Frequências



O último quadro que aparece na parte inferior da janela principal é o **Formato**, que inclui opções de ordem estética (ver Figura 60). Através do quadro Ordenar por, controlamos a ordem em que as categorias da variável são apresentadas na janela de saídas. As opções disponíveis são por ordem crescente (ou decrescente) dos valores que a variável assume e crescente (ou decrescente) de frequências.

Se você marcar a opção **Suprimir tabelas com muitas categorias**, o IBM SPSS Statistics não irá fazer nenhuma tabela que tiver mais do que um certo número de categorias, sendo que o padrão é 10. Finalmente, no quadro **Variáveis múltiplas**, caso você tenha pedido estatísticas para suas variáveis, você pode escolher que essas estatísticas saiam numa só tabela (**Comparar variáveis**) ou uma tabela para cada variável separadamente (**Organizar saída por variáveis**). No nosso exemplo, vamos escolher a opção **Contagens decrescentes** do quadro **Ordenar por**.

Figura 60 | Opções de Formato para o procedimento Frequências



Escolhidas todas as opções, podemos clicar em **OK**. Feito isso, o IBM SPSS Statistics ativará a janela de saídas e apresenta o resultado para as frequências solicitadas (ver Figura 61).

Você deve perceber que duas distribuições de frequências são apresentadas: uma para a variável *progtv* e outra para *filme*. A distribuição de frequências apresenta os rótulos dos valores; o número de observações em cada categoria (**Frequência**); a porcentagem que a frequência representa no total de observações levando-se em conta valores ausentes (**Porcentagem**); a porcentagem que a frequência representa no total de observações desconsiderando os valores ausentes (**Porcentagem Válida**) e o percentual válido acumulado (**Porcentagem Cumulativa**).

Acima da distribuição de frequências, a moda também é apresentada na tabela de estatísticas (lembre-se que escolhemos a opção **Moda** no quadro **Estatísticas**).

Figura 61 | Resultado obtido a partir do procedimento Frequências

## Qual seu tipo de programa de TV favorito?

### Estatísticas

Qual seu tipo de programa de TV favorito?

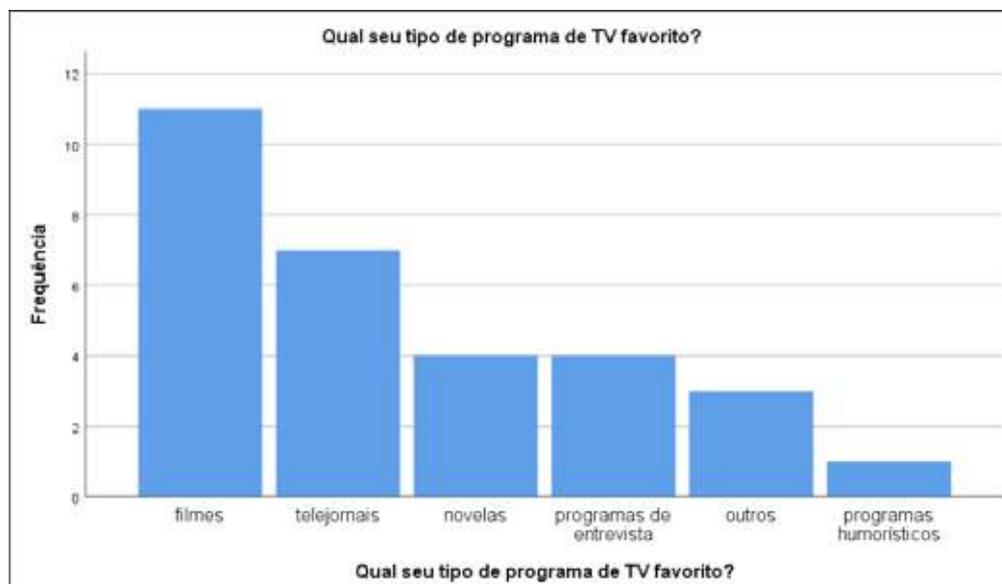
N	Válido	30
	Omisso	1
Modo		4

### Qual seu tipo de programa de TV favorito?

		Frequência	Porcentagem	Porcentagem válida	Porcentagem acumulativa
Válido	filmes	11	35,5	36,7	36,7
	telejornais	7	22,6	23,3	60,0
	novelas	4	12,9	13,3	73,3
	programas de entrevista	4	12,9	13,3	86,7
	outros	3	9,7	10,0	96,7
	programas humorísticos	1	3,2	3,3	100,0
	Total		30	96,8	100,0
Omisso	Sistema	1	3,2		
Total		31	100,0		

Temos, ainda, os gráficos de barras localizados na janela de saídas. Olhando o gráfico apresentado referente à variável *progtv*, por exemplo, notamos com facilidade que o tipo de programa preferido é filme (barra maior) e que o menos preferido é programa humorístico (menor barra) como mostra a Figura 62. É fácil perceber que a informação contida em um gráfico é muito mais simples de ser interpretada do que a informação numérica. Por este motivo, os gráficos são tão importantes na etapa de análise exploratória de dados.

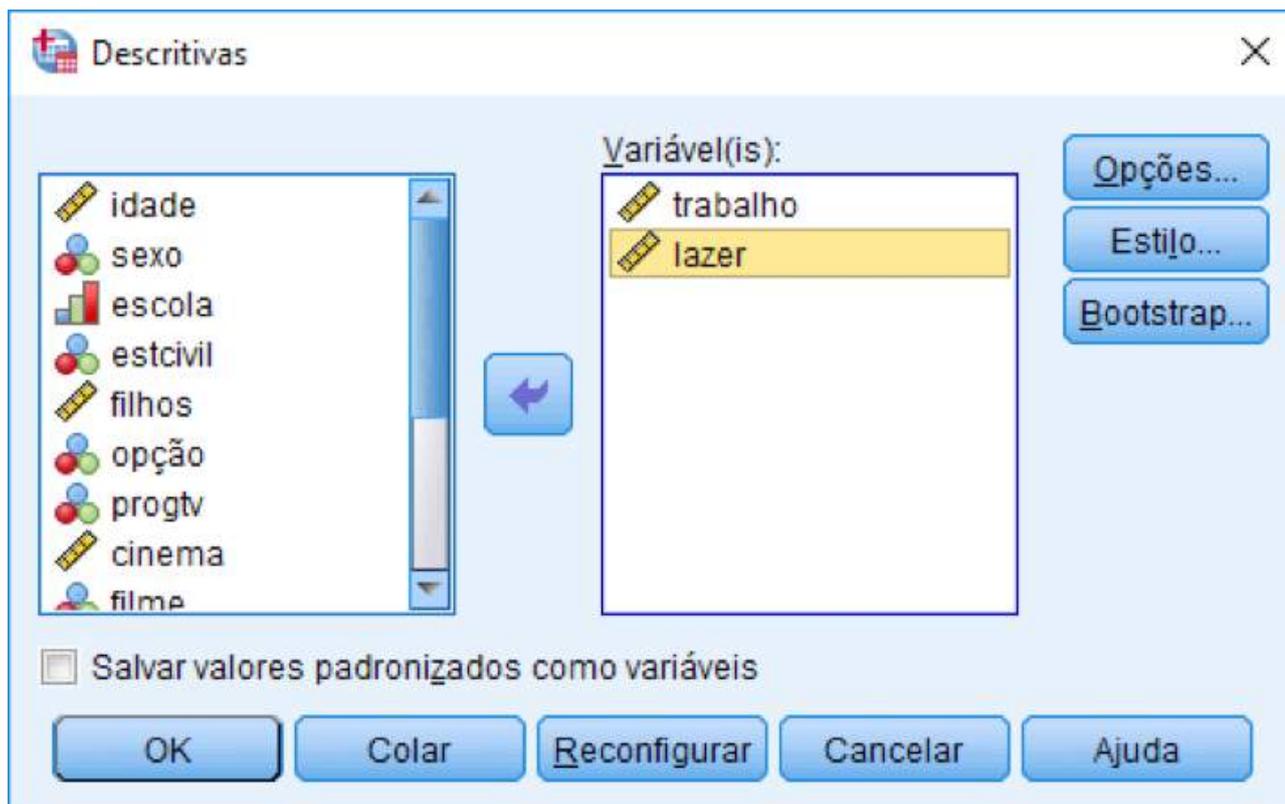
Figura 62 | Gráfico de barras gerado pelo procedimento Frequências



## 04.2 Descritivas

A segunda opção de Estatística Descritiva é o procedimento **Descritivas**. De maneira similar ao procedimento **Frequências**, temos a opção de incluir as variáveis que julgarmos necessárias (vamos colocar *trabalho* e *lazer*) ao quadro central **Variável(is)** (ver Figura 63). Esta opção geralmente é usada quando temos interesse em estudar variáveis escalares de maneira rápida.

Figura 63 | Procedimento Descritivas



O quadro superior direito, **Opções**, apresenta medidas estatísticas que podem ser usadas para a descrição dos dados (ver Figura 64). Existem dois grupos de medidas idênticos aos já discutidos anteriormente na opção **Estatísticas** do procedimento **Frequências**: **Dispersão e Distribuição**. Além dessas medidas, podemos pedir para o IBM SPSS Statistics apresentar a média e a soma das variáveis em estudo. O último quadro, **Ordem de Exibição**, permite escolher a ordem com que os resultados serão apresentados na janela de saídas: crescente ou decrescente de médias, alfabética ou a ordem em que as variáveis aparecem na lista. Este quadro é útil quando estamos estudando mais do que uma variável.

No nosso caso podemos escolher, por exemplo, as medidas: mínimo, máximo, média, desvio padrão e erro padrão, na ordem que as variáveis foram listadas (**Lista de Variáveis**). Agora, basta clicar **OK** na janela principal e analisar os resultados que são apresentados na janela de saídas (ver Figura 65).

Em média, as pessoas entrevistadas nesse estudo trabalham 37,9 horas por semana e têm 16,85 horas de lazer. Além disso, o máximo de horas trabalhadas foi 70 e o mínimo 0, enquanto que o máximo de horas de lazer foi 40 e o mínimo foi duas horas.

Figura 64 | Opções para o procedimento Descritivas

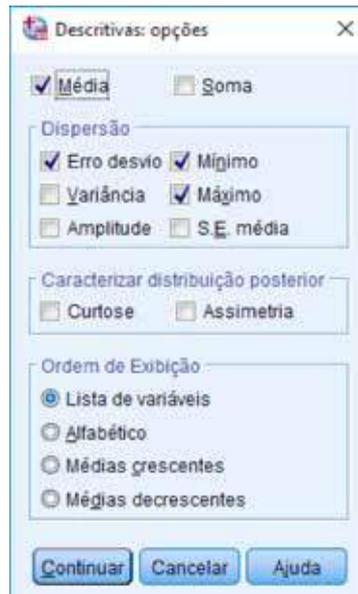


Figura 65 | Resultado obtido a partir do procedimento Descritivas

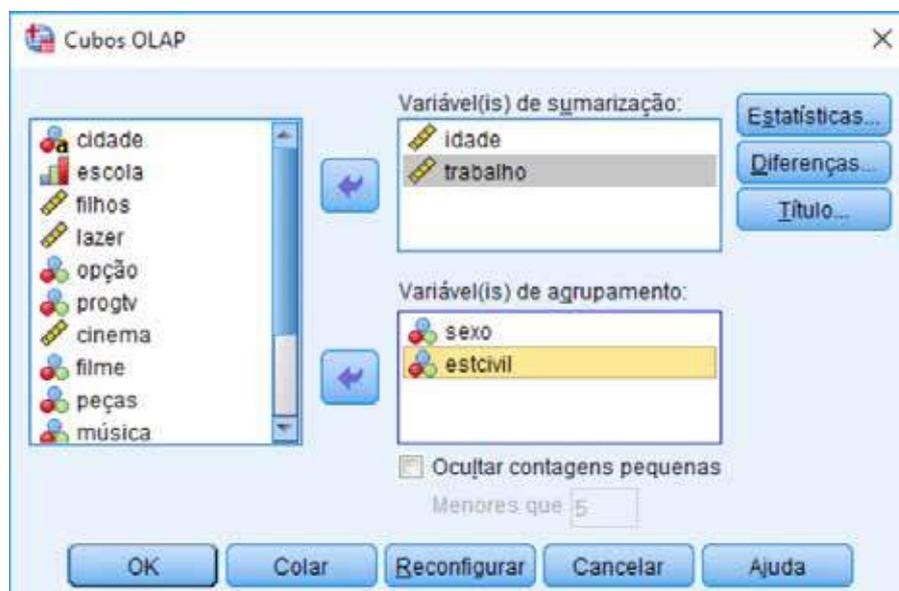
### Estadística Descritiva

	N	Mínimo	Máximo	Média	Erro Desvio
Horas de Trabalho por Semana	30	,0	70,0	37,900	13,4609
Horas Gastas com Lazer por Semana	27	2,0	40,0	16,852	8,1416
N válido (de lista)	26				

## 04.3 Cubos OLAP

A primeira opção do menu **Analisar** → **Relatórios** é **Cubos OLAP**. Este procedimento é utilizado para a construção de relatórios e podem se tornar extremamente flexíveis graças ao uso dos ícones de organização. A construção dos relatórios é simples: basta preencher os quadros da janela mostrada na Figura 66 com as variáveis desejadas. A seguir, detalharemos os quadros.

Figura 66 | Resultado obtido a partir do procedimento Descritivas

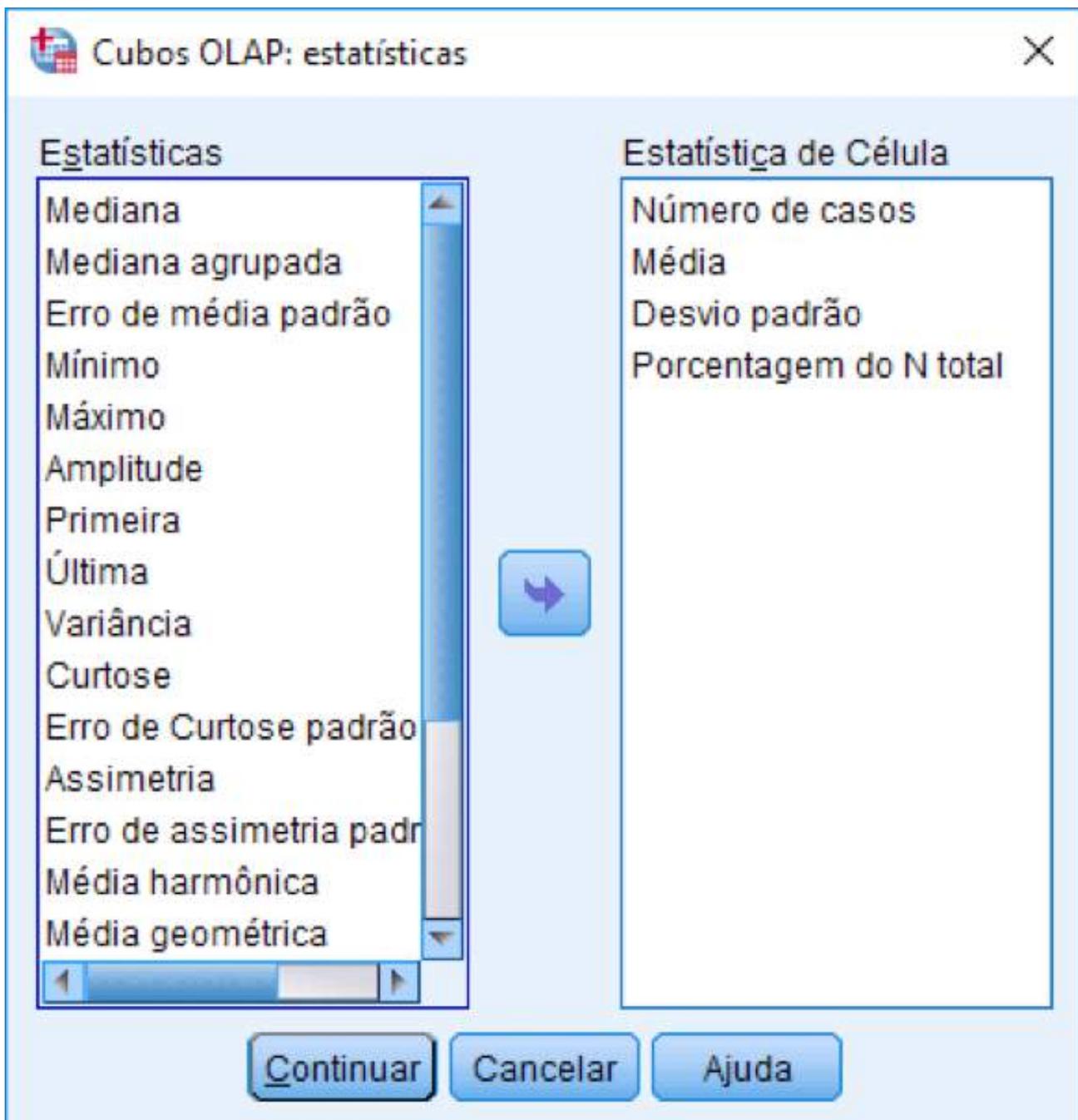


No quadro **Variável(is) de sumarização**, coloque as variáveis para as quais você deseja que ele calcule medidas descritivas como médias, somas, percentuais, mínimos, máximos, etc.; geralmente as variáveis nesse quadro são quantitativas. Como exemplo, vamos colocar as variáveis *idade* e *trabalho* do arquivo *Lazer.sav*.

No quadro abaixo, **Variável(is) de agrupamento**, coloque variáveis categóricas (nominais ou ordinais) para as quais você gostaria que fossem criados subgrupos. No nosso exemplo, colocaremos as variáveis *sexo* e *estcivil*.

Clicando no botão **Estatísticas**, você escolhe quais as estatísticas que serão calculadas. Como padrão, temos: soma, número de casos, média, desvio-padrão, percentual sobre a soma total, percentual sobre o N total; várias outras estatísticas como mediana, mínimo, máximo e variância, podem ser escolhidas (veja Figura 67). No nosso exemplo, removeremos a Soma e o percentual sobre a soma total (**Porcentagem da Soma Total**). Depois, basta clicarmos em **Continuar**.

Figura 67 | Opções de estatísticas para Cubos OLAP



Ao clicar no botão **Título** (ver Figura 68), você pode escolher um título e uma legenda para a tabela, que poderão ser modificados posteriormente. Após a digitação do título e da legenda, clique em **Continuar**. Agora, basta clicar no botão **OK** e teremos nosso relatório como na Figura 69.

Inicialmente, temos o número de casos, a média, o desvio-padrão e o percentual sobre o número de casos para as variáveis *idade* e *trabalho*, considerando todos os casos, independentemente de sexo, estado civil e grau de instrução. Assim, por exemplo, temos para a variável *idade* que:

- O número de pessoas que responderam a idade é 31;
- A idade média dessas pessoas é 28,9 anos e o desvio-padrão é 11,31 anos;
- Elas representam 100% das pessoas que responderam a idade (por enquanto, esta coluna é inútil).

Repare que o procedimento **Cubos OLAP**, de certa forma, é uma combinação interessante dos procedimentos **Frequências** e **Descritivas**, uma vez que há a possibilidade de analisar medidas resumo de variáveis escalares a partir do cruzamento de variáveis categóricas, formando grupos de interesse a partir do cruzamento de categorias destas variáveis.

Figura 68: Escolha de título e legenda para o procedimento Cubos OLAP

Figura 69 | Resultado gerado pelo procedimento Cubos OLAP

### Exemplo de Tabela OLAP

Sexo: Total | Estado civil: Total

	N	Média	Erro Padrão	% de N total
<b>Idade dos Entrevistados</b>	31	28,94	11,314	100,0%
<b>Horas de Trabalho por Semana</b>	30	37,900	13,4609	100,0%

Vamos editar a tabela e mostrar subgrupos baseados em *sexo* e *estado civil*: clique com o botão direito do mouse na tabela e escolha a opção **Editar Conteúdo...No Visualizador** (ou dê um clique duplo sobre a tabela) para ativá-la. Primeiramente, vamos selecionar apenas os homens; para isso, clique na seta ao lado da palavra **Total** associada a *sexo* e selecione *masculino* na lista que irá aparecer. De maneira similar, vamos selecionar a categoria *solteiro* em *estcivil*. Agora nossa análise para a idade ficaria assim (ver Figura IV-14):

- O número de homens solteiros que responderam a idade é 7;
- A idade média desses homens é 20,29 anos e o desvio-padrão é 3,63 anos;
- Eles representam 22,6% das pessoas que responderam a idade.

Figura 70 | Resumindo a análise aos homens solteiros

### Exemplo de Tabela OLAP

Sexo: masculino  
Estado Civil: solteiro

	N	Média	Erro Desvio	% de N total
Idade dos Entrevistados	7	20,29	3,638	22,6%
Horas de Trabalho por Semana	7	37,143	20,5866	23,3%

## 04.4 Tabelas de Referência Cruzada (Crosstabs)

A quarta opção de **Estatística Descritiva** é o procedimento **Tabelas de Referência Cruzada** (ou **Crosstabs**), que é utilizado para a construção de tabelas simples de dupla entrada. Imagine que temos duas variáveis qualitativas e estamos interessados em observar a distribuição de frequências de ambas em conjunto. Para isso, construímos uma tabela onde uma variável entra com suas categorias em linhas e a outra em colunas. Podemos então fazer um estudo descritivo sobre o comportamento conjunto das variáveis. Porém, muitas vezes queremos mais do que simplesmente observar. Queremos, também, verificar se existe efetivamente associação entre tais variáveis. Então, precisamos utilizar testes estatísticos (análise inferencial). Alguns desses testes estatísticos estão disponíveis no procedimento **Crosstabs**.

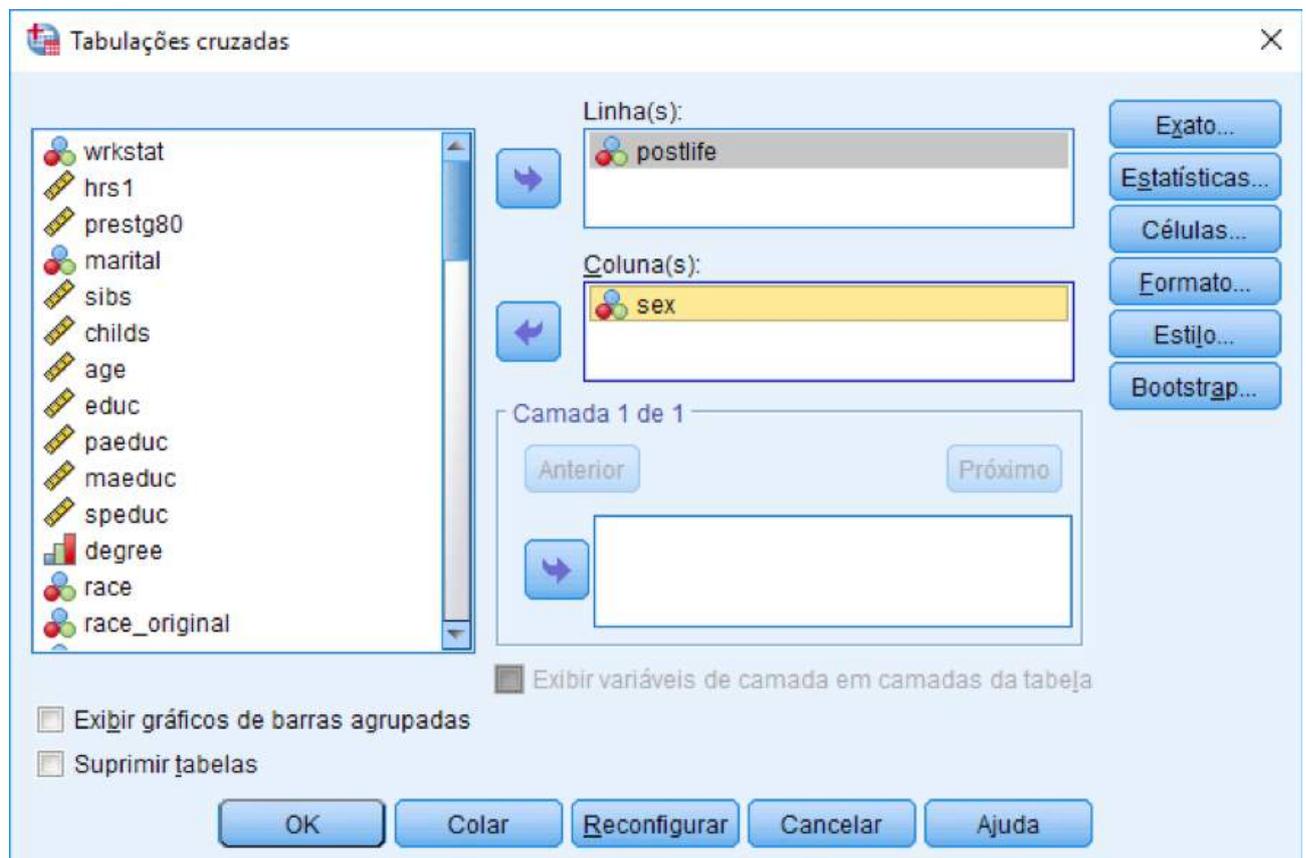
A Figura 71 mostra as opções do procedimento **Crosstabs**. No canto esquerdo, estão disponíveis todas as variáveis do banco de dados. Para construção da tabela, basta movermos a variável cujas categorias desejamos, dispostas nas linhas para o quadro **Linha(s)** localizado no centro da janela, e a variável cujas categorias desejamos, dispostas nas colunas para o quadro **Coluna(s)**. Note que podemos colocar mais do que uma variável nos quadros **Coluna(s)** e **Linhas(s)**. Entretanto, por meio do comando **Crosstabs** serão construídas várias tabelas de dupla entrada a partir da combinação das variáveis duas a duas.

Se você deseja construir uma tabela com mais de duas entradas (envolvendo três variáveis ou mais), você pode usar o quadro **Camada(Layer)**; porém, você terá opções limitadas de estatísticas e customização. Neste caso, o módulo *IBM SPSS Custom Tables* é indicado, já que ele possui recursos avançados para a criação e personalização de tabelas.

Suponha que estamos interessados em verificar se existe alguma associação entre a crença na vida após a morte e o gênero de um indivíduo norte-americano. Para isso, abriremos o arquivo *Gss2002.sav*. Moveremos a variável *postlife* (Crença na vida após a morte) para o quadro **Linha(s)** e a variável *sex* (Sexo do respondente) para o quadro **Coluna(s)**.

Se você marcar a opção **Exibir gráficos de barras agrupadas**, serão feitos gráficos de barras com as variáveis pedidas e a opção **Suprimir tabelas** fará com que nenhuma tabela seja feita – ou seja, apenas as estatísticas, testes e gráficos pedidos serão mostrados na janela de saídas.

Figura 71 | Procedimento Tabulações Cruzadas (Crosstabs)

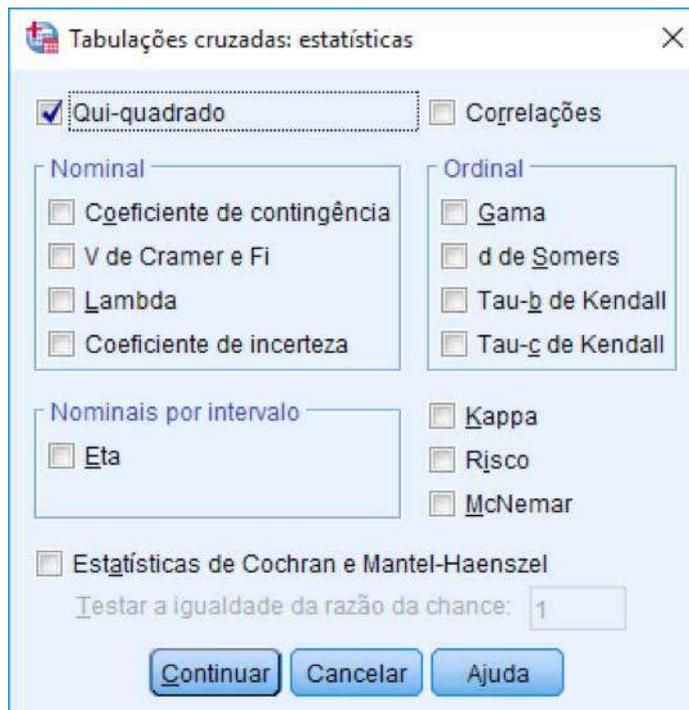
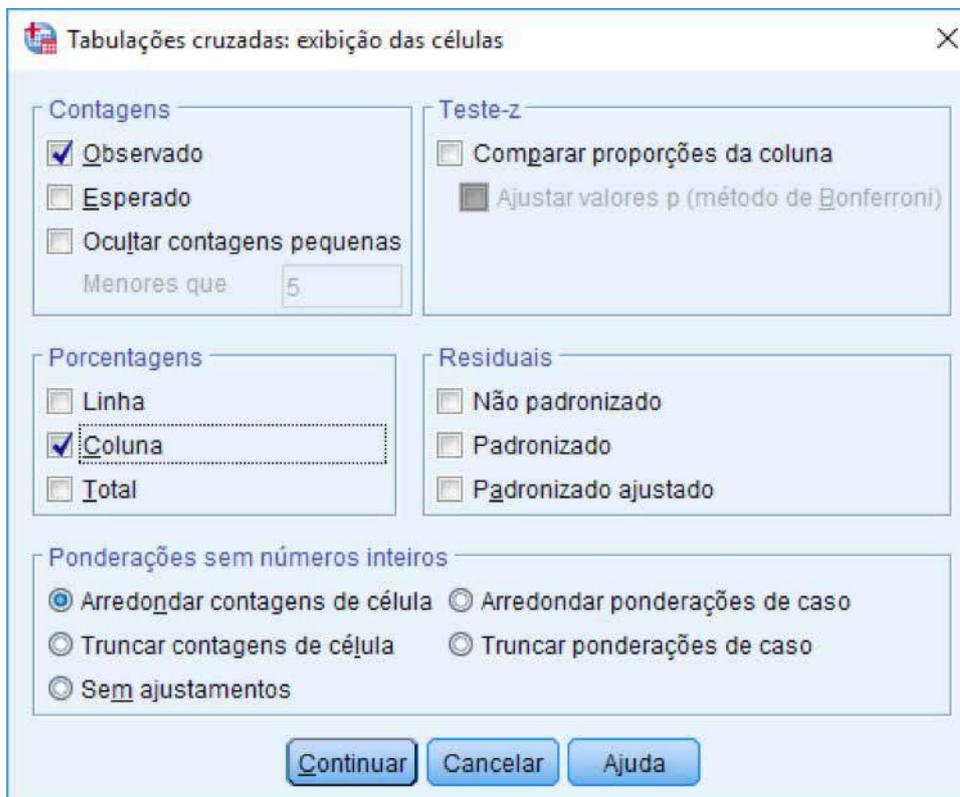


- **Estatísticas** - Deve ser usada para selecionar os testes estatísticos que serão realizados; note que são apresentados nesse quadro catorze testes estatísticos diferentes; cada um deles é apropriado para verificarmos suposições diferentes sobre as variáveis em questão (ver Figura 72), vamos selecionar a estatística **Qui-quadrado (Chi-square)**, que é utilizada para testar a hipótese de que as variáveis que definem as linhas e colunas são independentes;

- **Células** - Essa opção é usada para escolha das medidas que serão apresentadas em cada uma das células da tabela (ver Figura 73, as possíveis medidas apresentadas são: frequências ou valores observados (Contagens... Observado); valores esperados, isto é, frequências calculadas assumindo que as variáveis são independentes (**Contagens → Esperado**); porcentagens por linha, coluna e total (**Porcentagens, Linha, Coluna e Total** respectivamente); vários tipos de resíduos, que são a diferença entre o valor observado e esperado (quadro **Residuais**); até mesmo como os números não-inteiros serão apresentados na tabela (quadro **Ponderações sem números inteiros**); pediremos, para nosso exemplo, os valores observados e as porcentagens por coluna;

- **Formato** - Usada apenas para escolher a ordem em que as categorias (valores) da variável apresentada nas linhas da tabela serão apresentadas (**Ordem da linha**), que pode ser crescente ou decrescente; no nosso exemplo, escolheremos a opção **Crescente**;

O passo mais importante agora é saber interpretar todos os resultados presentes na janela de saídas, senão todo o trabalho que tivemos até agora foi apenas perda de tempo!

Figura 72 | Opções de Estatísticas do procedimento *Crosstabs*Figura 73 | Opções de exibição das Células do procedimento *Crosstabs*

A Figura 74 apresenta o resultado da tabela de dupla entrada feita utilizando o procedimento **Crosstabs**. Note que, dentro de cada célula, são apresentados os dois valores pedidos: *Contagem*, que representa a frequência, e % em *SEXO DO RESPONDENTE*, representando a porcentagem sobre a coluna (a variável *sex*). Esta disposição pode ser facilmente modificada usando os ícones de organização como já visto; note na Figura 75 a diferença em relação à

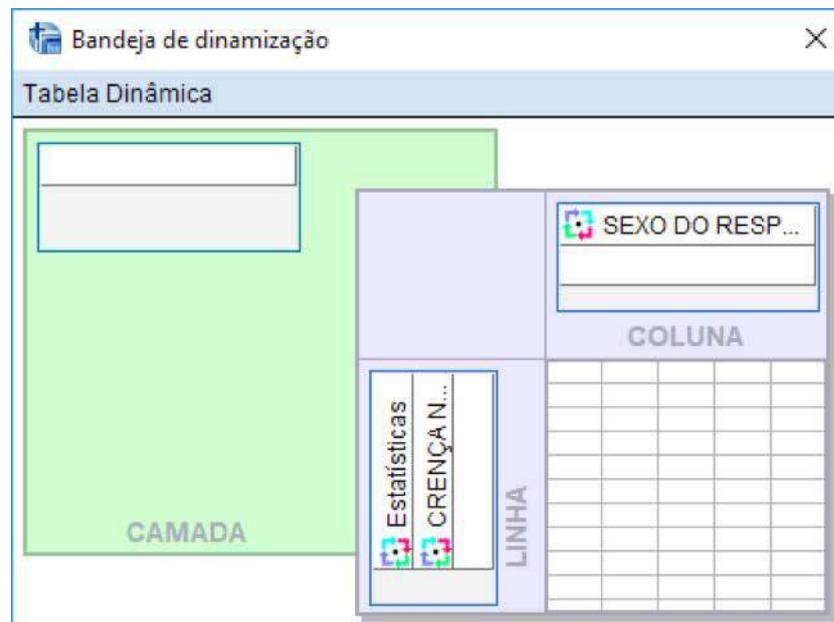
tabela que você tem na tela do seu computador. A Figura 75 mostra como alterar a tabela utilizando o comando **Bandeja de Dinamização (Pivoting Trays)**, presente no menu **Tabela Dinâmica (Pivot Table)** na tela de edição da tabela.

Figura 74 | Saída gerada pelo procedimento *Crosstabs*

**Tabulação cruzada CRENÇA NA VIDA APÓS A MORTE \* SEXO DO RESPONDENTE**

		SEXO DO RESPONDENTE		Total	
		MASCULINO	FEMININO		
CRENÇA NA VIDA APÓS A MORTE	SIM	Contagem	414	532	946
		% em SEXO DO RESPONDENTE	76,0%	85,3%	80,9%
	NÃO	Contagem	131	92	223
		% em SEXO DO RESPONDENTE	24,0%	14,7%	19,1%
Total		Contagem	545	624	1169
		% em SEXO DO RESPONDENTE	100,0%	100,0%	100,0%

Figura 75 | Opções da Bandeja de Dinamização (Pivoting Trays)



Além da tabela, também é apresentada, na janela de saídas, a estatística do teste qui-quadrado. Através dos valores apresentados, saberemos se existe alguma evidência para concluir que as variáveis são estatisticamente relacionadas (dependentes). Valores das estatísticas de qui-quadrado de Pearson e da razão de verossimilhança são apresentados; a mais usada e conhecida é a de Pearson, que no nosso caso tem valor 16,275. Como o nível de significância, *Significância Assintótica (Bilateral)* (ver Figura 76), associado a essa estatística é muito baixo (menor que 5%, valor de referência utilizado na maioria dos testes estatísticos), temos evidência suficiente para afirmar que as variáveis estão relacionadas; portanto, indivíduos de gêneros diferentes apresentam opiniões diferentes sobre a crença na vida após a morte.

Figura 76 | Teste de qui-quadrado para o cruzamento POSTLIFE x SEX

**Testes qui-quadrado**

	Valor	gl	Significância Assintótica (Bilateral)	Sig exata (2 lados)	Sig exata (1 lado)
Qui-quadrado de Pearson	16,275 <sup>a</sup>	1	,000		
Correção de continuidade <sup>b</sup>	15,679	1	,000		
Razão de verossimilhança	16,270	1	,000		
Teste Exato de Fisher				,000	,000
Associação Linear por Linear	16,261	1	,000		
Nº de Casos Válidos	1169				

a. 0 células (0,0%) esperavam uma contagem menor que 5. A contagem mínima esperada é 103,96.

b. Computado apenas para uma tabela 2x2

Um ponto importante é que, **no máximo**, 20% das células da tabela cruzada podem ter valores esperados menores que 5 (o que é indicado na nota de rodapé a). Caso esta condição não seja satisfeita, o teste qui-quadrado perde poder, ou seja, a probabilidade de detectar associação entre as variáveis diminui mesmo quando a hipótese de associação é verdadeira; logo, seu uso para tirar conclusões sobre a associação entre duas variáveis deve ser feito com cautela.

A terceira linha da tabela nos mostra uma estatística (**Associação Linear por Linear**) que mede a associação linear entre as variáveis da linha e da coluna. Esta estatística não deve ser usada para variáveis nominais (nosso caso, por exemplo) e também é chamada de estatística Qui-quadrado de Mantel-Haenszel.

Você deve ter percebido que as opções disponíveis em **Relatórios e Estatística Descritiva** para análise descritiva são usadas para obter diferentes informações sobre os dados. É importante lembrar que, embora o IBM SPSS Statistics seja um *software* muito fácil de operar, o bom senso do usuário é fundamental; em outras palavras, a cada passo que pretendemos dar, é preciso pensar antes se a análise faz sentido ou não, pois se pedirmos ao IBM SPSS Statistics que calcule a média para a variável *estcivil* do nosso arquivo *Lazer.sav*, ele o fará, mas a resposta obtida certamente não fará sentido algum.

## 05 Manipulação de dados

### 05.1 Ordenação de casos

Antes de realizarmos qualquer análise, é comum termos que “preparar” os dados para que possamos obter a informação desejada. Por exemplo: Uma das necessidades na hora da manipulação dos dados em nosso dia a dia é a ordenação dos casos segundo uma ou mais variáveis. Para fazer isso no IBM SPSS Statistics, você pode usar o procedimento **Ordenar Casos**, presente no menu **Dados**.

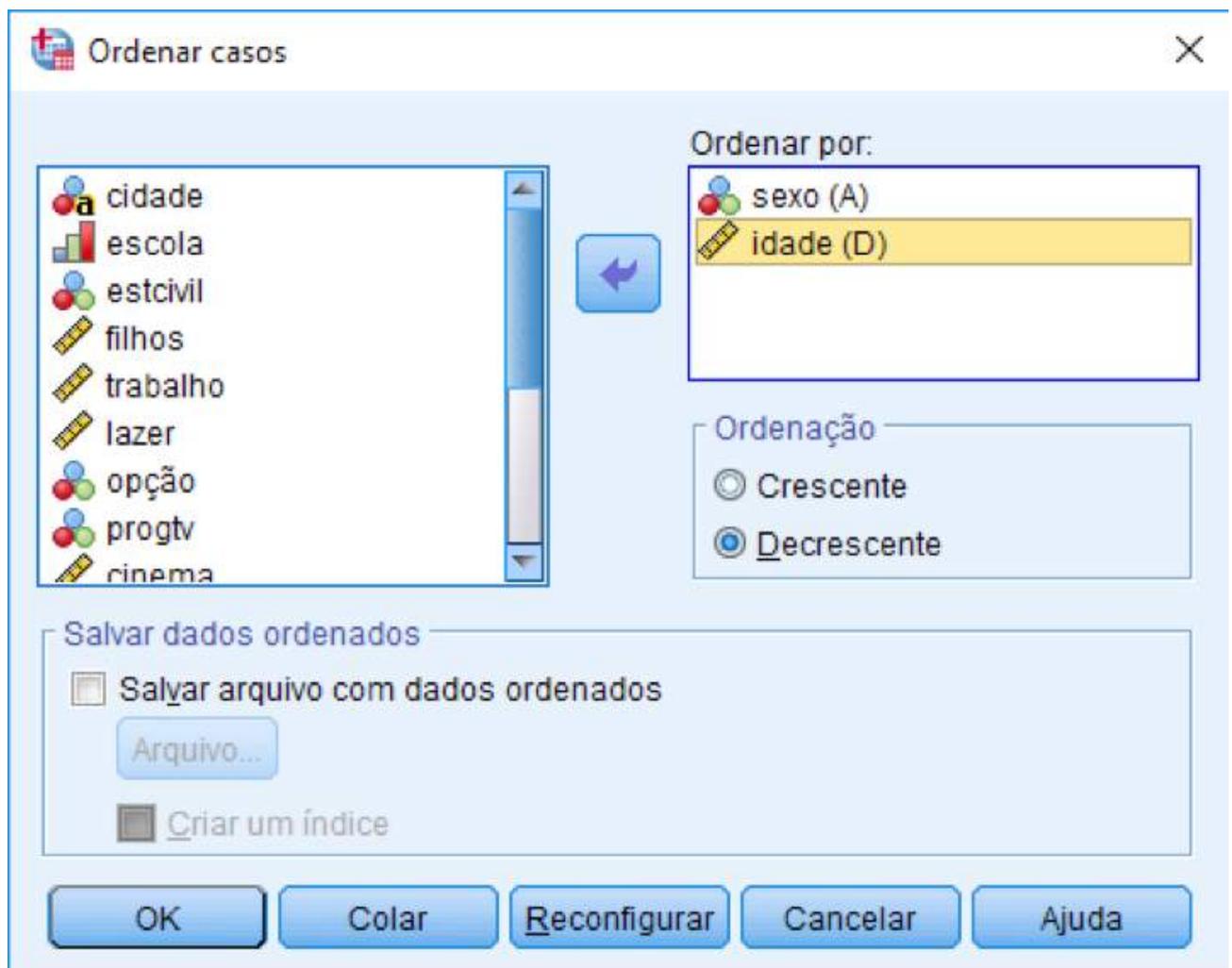
Após clicar no menu **Dados**, opção **Ordenar Casos**, uma janela é aberta (ver Figura 77). Movemos para o quadro **Ordenar por**: a variável segundo a qual o arquivo deve ser ordenado. Podemos mover para esse quadro mais do que uma variável. Nesse caso, o arquivo é ordenado, em primeiro lugar, pelos valores da primeira variável presente no quadro e em segundo lugar pela

segunda variável presente no quadro; a segunda ordenação é feita para os valores comuns da primeira variável e assim por diante.

Podemos escolher também entre ordem crescente ou decrescente de ordenação para cada uma das variáveis. Isso é feito através do quadro **Ordenação**, opções **Crescente** (que é a opção padrão do comando) ou **Decrescente**.

Vamos fazer uma ordenação segundo sexo (ordem crescente) e idade (ordem decrescente) a partir do arquivo *Lazer.sav*. Para isso, movemos a variável *sexo* para o quadro **Ordenar por**: e escolhemos a opção **Crescente** no quadro **Ordenação**. Movemos em seguida a variável *idade* para o quadro Ordenar por: e escolhemos a opção **Decrescente** no quadro **Ordenação**. Agora, basta clicar **OK** para validar a ordenação.

Figura 77 | Opção Ordenar Casos do menu Dados



Você pode notar na Figura 77 que aparece a letra **A** (de **A**scending, ou *Crescente*) na frente da variável *sexo* e a letra **D** (de **D**escending, ou *Decrescente*) na frente da variável *idade*. Isso indica a ordem pela qual será feita a ordenação para cada uma das variáveis.

A Figura 78 mostra a janela de dados após a ordenação. Note que nas 19 primeiras linhas aparecem informações referentes a pessoas do sexo masculino (*sexo* = 1); além disso, a primeira linha do arquivo contém as informações do homem mais velho e a linha 19 do mais jovem. As demais linhas contêm as informações das mulheres (*sexo* = 2) também ordenadas pela idade. De fato, a ordenação segue a ordem na qual as variáveis foram incluídas na janela **Ordenar Casos**.

Figura 78 | Janela de dados após ordenação

	idade	cidade	sexo	escola	estcivil	filhos	trabalho	lazer	opção	progtv	cinema
1	60	São Paulo	1	6	2	6	35,0	14,0	7	1	0
2	41	Rio de Janeiro	1	6	4	3	40,0	2,0	6	1	0
3	37	Rio de Janeiro	1	4	2	3	40,0				0
4	36	São Paulo	1	6	3	0	50,0			4	0
5	36	São Paulo	1	6	3	0	40,0	16,0	1	1	0
6	31	Brasília	1	6	4	3	50,0	5,0	1	4	2
7	29	Rio de Janeiro	1	6	4	0	40,0	3,0	7	1	0
8	26	Rio de Janeiro	1	6	1	0	70,0	26,0	4	4	10
9	25	São Paulo	1	6	3	0	50,0	17,0	6	5	10
10	24	São Paulo	1	5	3	0	40,0	15,0	2	4	10
11	23	São Paulo	1	5	2	0	48,0	16,0	3	5	2
12	22	São Paulo	1	5	1	0	30,0		3	1	10
13	21	São Paulo	1	2	2	3	40,0	10,0	1	4	1
14	21	Rio de Janeiro	1	5	1	0	40,0	18,0		6	3
15	21	Brasília	1	5	1	0	40,0	20,0	7	1	10
16	20	Brasília	1	5	3	0	20,0	18,0	7	2	2
17	19	São Paulo	1	5	1	0	40,0	22,0	1	4	2
18	19	Rio de Janeiro	1	3	1	0	40,0	20,0	2	6	5
19	14	Brasília	1	1	1	0	0	40,0	6	2	0
20	54	São Paulo	2	6	2	4	24,0	13,0	1	4	1
21	53	São Paulo	2	6	2	6	56,0	12,0	1	1	0
22	37	Brasília	2	5	2	0	45,0	8,0	1	4	12

Dica sobre ordenação: Você pode ordenar os dados direto da janela de dados sem usar a opção **Ordenar Casos** do menu Dados. Para isso, basta selecionar as variáveis que definem a ordenação (você pode usar as teclas CTRL ou SHIFT para selecionar várias variáveis), clicar com o botão direito do mouse no nome de qualquer uma das variáveis selecionadas e escolher **Ordenar em Ordem Ascendente** ou **Ordenar em Ordem Descendente**.

As variáveis também podem ser ordenadas utilizando o comando **Ordenar Variáveis** presente no menu Dados. Essa ordenação poderá ser feita por nome da variável, tipo, tamanho etc. Maiores detalhes poderão ser vistos no menu de **Ajuda** do IBM SPSS Statistics.

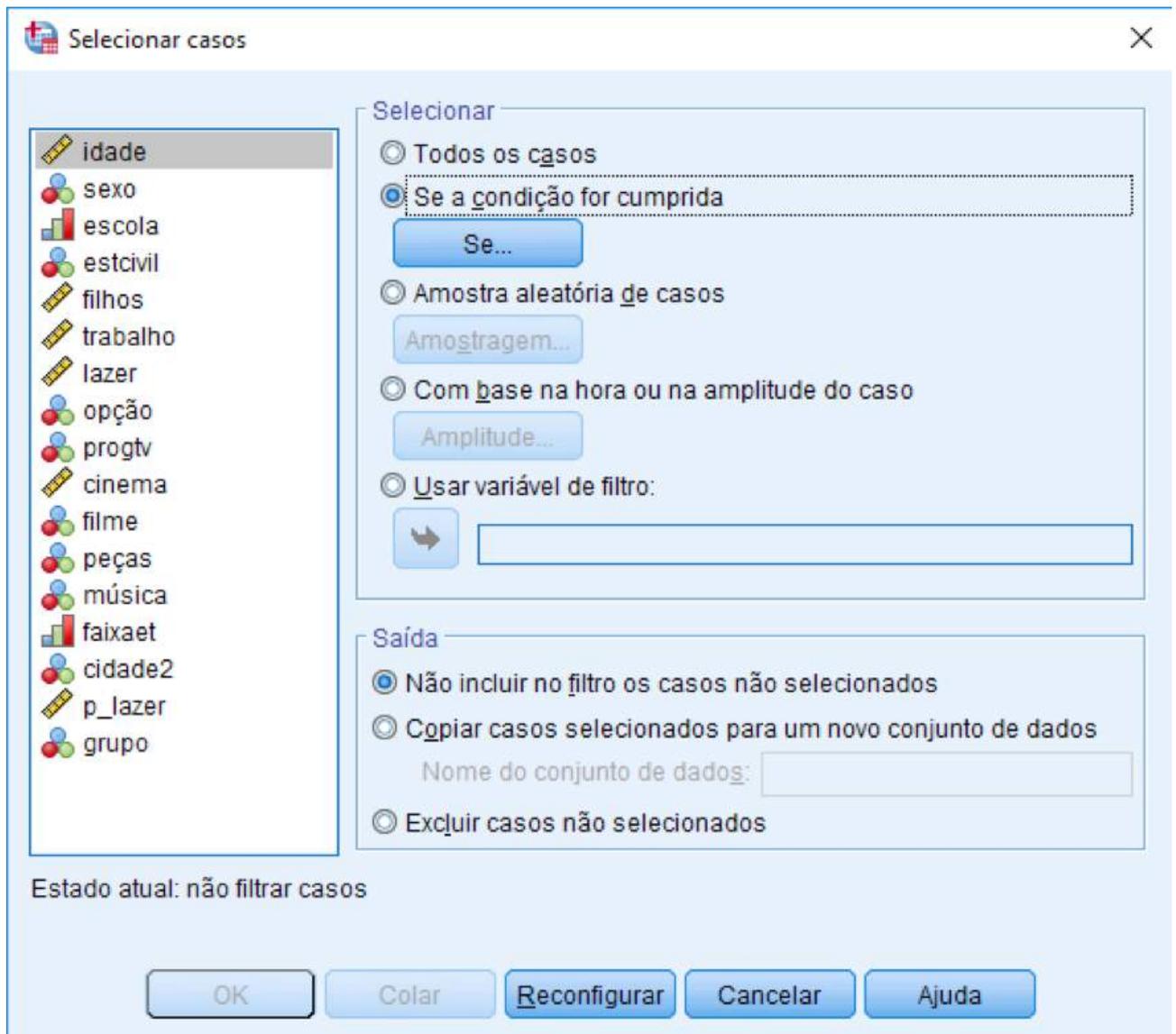
É possível também salvar o arquivo no momento da ordenação clicando em **Salvar arquivo com dados ordenados**. Porém, neste caso, é possível salvar o arquivo apenas na terminação do IBM SPSS Statistics (.sav).

## 05.2 Seleção de casos

Outra grande necessidade é a seleção (temporária ou permanente) de parte do arquivo de dados. Digamos que estamos interessados em estudar um segmento específico da amostra. O IBM SPSS Statistics possui várias formas de seleção de dados. Falaremos nessa seção de todas elas, mas discutiremos detalhadamente a mais usada de todas. Para maiores detalhes sobre as demais formas de seleção, recomendamos a utilização da **Ajuda** do IBM SPSS Statistics.

Para fazer qualquer tipo de seleção, devemos clicar no menu **Dados**, opção **Selecionar Casos**. A janela aberta é apresentada na Figura 79.

Figura 79 | Opção Selecionar Casos do menu Dados



No quadro central **Selecionar**, estão presentes cinco opções diferentes para seleção. Faremos a seguir uma breve descrição de cada uma delas:

- **Todos os casos** - Opção usada como padrão; utiliza todas as observações do banco de dados;
- **Se a condição for cumprida** - Através desta opção podemos definir expressões condicionais para seleção de casos; estudaremos esta opção detalhadamente mais adiante;
- **Amostra aleatória de casos** - Podemos selecionar uma porcentagem aproximada ou um número exato de casos dos  $n$  primeiros casos; a seleção é feita aleatoriamente;
- **Com base na hora ou na amplitude do caso** - Usamos essa opção quando estamos interessados em selecionar uma faixa específica de valores, por exemplo, os casos de número 100 até 200; também utilizada para fazer seleções baseadas em datas (por exemplo, selecionando todos os dados do mês de fevereiro até outubro);
- **Usar variável de filtro** - Escolhemos uma variável no banco de dados e usamos como filtro; todos os casos para os quais a variável de filtro assume o valor 0 (ou "ausente", "Not Selected") não serão selecionados.

Você tem três opções para o tratamento dos casos que não serão selecionados. É por meio do quadro Saída que podemos fazer a escolha:

- **Não incluir no filtro os casos não selecionados** - Os casos (linhas) que não são selecionados não são incluídos nas análises posteriores, mas permanecem na janela de dados; caso você mude de ideia e queira usar os casos não selecionados na mesma sessão do IBM SPSS Statistics, basta “desligar” o filtro (usando a opção Todos os casos no quadro Selecionar) ou deletar a variável *filter\_\$*;
- **Copiar casos selecionados para um novo conjunto de dados** - Os casos (linhas) selecionados são copiados para uma nova janela de dados sem que seja alterada a janela de dados atual;
- **Excluir casos não selecionados** - Os casos (linhas) não selecionados são apagados da janela de dados; caso você mude de ideia e queira usar os casos não selecionados, você deverá ler novamente o arquivo de dados original.

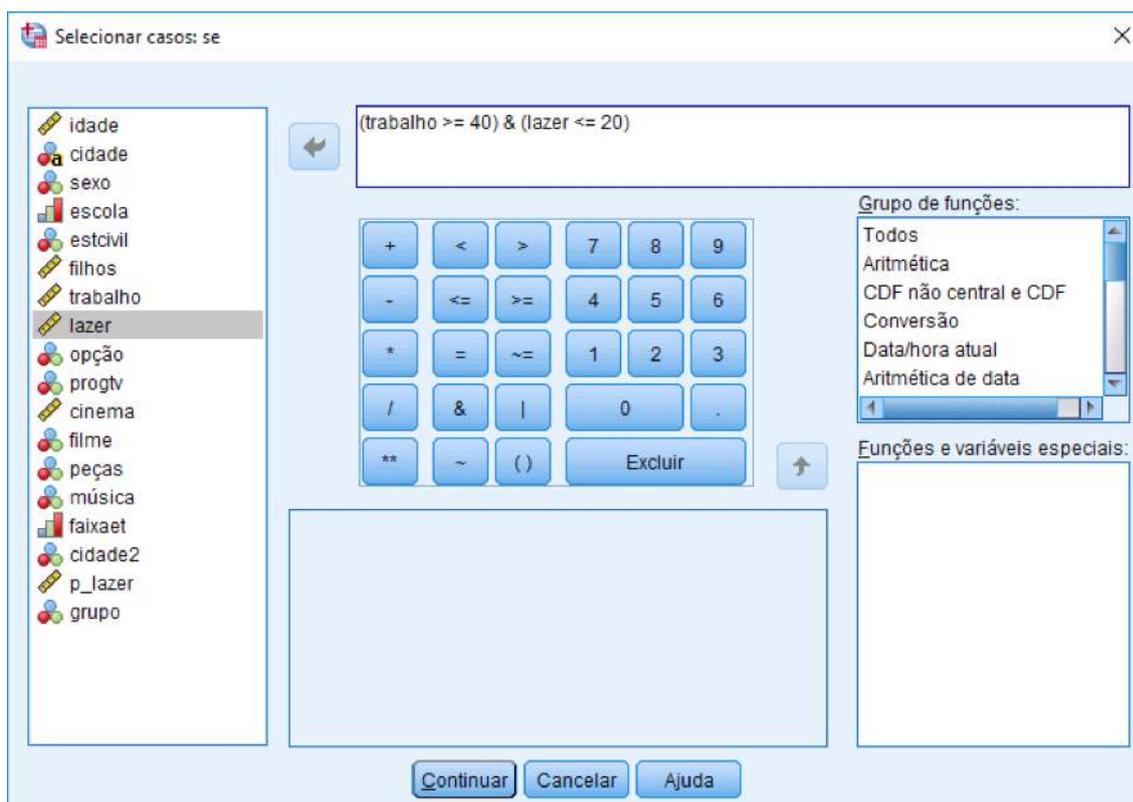
Suponha que, a partir do arquivo *Lazer.sav*, estamos interessados em selecionar as pessoas que trabalham pelo menos 40 horas por semana e que têm até 20 horas de lazer. A função condicional para seleção nesse caso é dada por:

trabalho >= 40 AND lazer <= 20

Portanto, o tipo de seleção de dados que faremos deve possibilitar a criação de sentenças matemáticas lógicas para seleção dos casos. Para isso, clicamos em **Se a condição for cumprida**, e entramos na caixa de diálogo **Se**.

Através da janela que é aberta (ver Figura 80), usamos o quadro superior para escrever uma função lógica na qual a seleção vai ser baseada. Para a construção da função, podemos usar todas as variáveis que estão no quadro à esquerda e as funções disponíveis no quadro inferior direito.

Figura 80 | Opção 'Selecionar Casos...Se a Condição For Cumprida' do menu Dados



Uma vez escrita a função que determina a regra de seleção dos casos, clique em **Continuar** e você voltará à janela anterior. No quadro inferior (**Saída**), vamos optar por **Não incluir no filtro os casos não selecionados** (ou seja, os casos não selecionados permanecem na tela de dados, porém não serão utilizados em análises futuras) e clicar em **OK**.

Você pode perceber que depois de feita a seleção, a janela de dados sofre algumas alterações (ver Figura 81). As linhas (casos) que não foram selecionadas apresentam uma listra no canto esquerdo da janela de dados. A barra de status, localizada na parte inferior direita da janela, apresenta a mensagem **Filtro Ativado**. Além disso, uma variável de nome filter\_\$ é adicionada à janela de dados. Essa nova variável apresenta valor 0 para as linhas que não foram selecionadas e valor 1 para as linhas que foram selecionadas. Apesar de você conseguir ver os casos que não foram selecionados, qualquer análise efetuada daí para frente não levará em conta esses casos.

A Figura 82 apresenta o resultado de uma análise descritiva feita através do menu **Analisar** → **Estatística Descritiva** → **Descritivas** para a variável trabalho. Note que o número médio de horas trabalhadas por semana é 43,06, o mínimo é 40 horas semanais e apenas 16 observações foram usadas para esses cálculos. Isso parece óbvio, já que usamos a condição trabalho >=40 como parte do critério de seleção.

Figura 81 | Janela de dados com seleção de casos ativa

	cinema	filme	peças	música	faixaet	cidade2	p_lazer	grupo	filter_\$
1	0	6	1	2	4	3	28,57	1,00	0
2	0	2	1	2	3	2	4,76	1,00	1
3	0	.	.	.	3	2	.	,00	.
4	0	5	.	4	3	3	.	,00	.
5	0	2	1	2	3	3	28,57	,00	1
6	2	6	1	1	2	1	9,09	,00	1
7	0	2	.	6	2	2	6,98	,00	1
8	10	2	1	4	2	2	27,08	,00	0
9	10	4	1	4	1	3	25,37	,00	1
10	10	4	.	4	1	3	27,27	,00	1
11	2	6	1	4	1	3	25,00	,00	1
12	10	6	1	4	1	3	.	,00	0
13	1	3	.	3	1	3	20,00	,00	1

Figura 82 | Análise descritiva da variável TRABALHO após a seleção de casos

### Estatística Descritiva

	N	Mínimo	Máximo	Média	Erro Desvio
Horas de Trabalho por Semana	16	40,0	56,0	43,063	5,1312
N válido (de lista)	16				

Podemos mudar de ideia e querer usar todas as observações para o cálculo das estatísticas. Temos duas maneiras de cancelar a seleção de casos se o filtro de casos estiver ativado. A primeira delas é ativar a opção **Todos os Casos** da janela de seleção de casos (opção **Selecionar Casos** do menu **Dados**) e clicar **OK**. A segunda maneira é apagar a variável *filter\_\$* do banco de dados; entretanto, recomenda-se de uma forma geral que as variáveis de filtro sejam mantidas e renomeadas, uma vez que o nome *filter\_\$* é o padrão para variáveis de filtro do SPSS e, a cada nova seleção de casos, a nova variável sobrescreverá a anterior caso o nome não seja alterado. Altere o nome da variável para *filtro\_trab\_lazer* na **Visualização de Variável**.

Usando agora o procedimento **Descritivas** para a variável trabalho, o resultado apresentado na Figura 83 é bem diferente do apresentado anteriormente, quando a seleção de casos estava ativa: note que o total de casos usados nos cálculos passou de 16 para 30, a média de horas trabalhadas baixou para 37,90 e agora o mínimo de horas trabalhadas é igual a 0 horas.

Figura 83 | Resultado da análise descritiva para a variável TRABALHO com todos os casos

Estatística Descritiva					
	N	Mínimo	Máximo	Média	Erro Desvio
Horas de Trabalho por Semana	30	,0	70,0	37,900	13,4609
N válido (de lista)	30				

Vamos supor, agora, que estejamos interessados em analisar somente os entrevistados que são procedentes de São Paulo em relação a horas de trabalho semanais. Usaremos a variável *cidade* e o comando **Selecionar Casos**. Na tela **Selecionar Casos: Se** (veja Figura 84), basta digitarmos a condição

`cidade = 'São Paulo'`

uma vez que a variável é alfanumérica (*string*) e, portanto, os valores precisam ser digitados entre aspas simples (ou duplas). Clique em **Continuar** e, depois, em **OK**. O resultado do filtro está na Figura V-9 e, como esperado, uma nova variável *filter\_\$* foi criada.

Figura 84 | Seleção de casos a partir da variável CIDADE

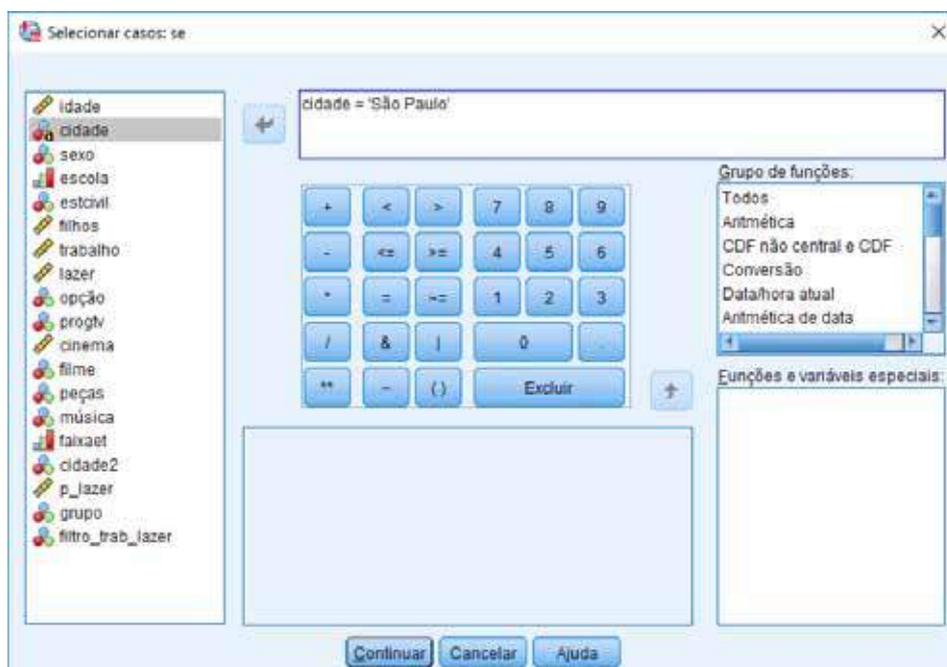


Figura 85 | Janela de dados com seleção de casos ativa para a variável CIDADE

Visualização de dados

	filme	peças	música	faixaet	cidade2	p_lazer	grupo	filtro_trab_lazer	filter_\$
1	comédia	comédia	mpb	> 45	São Paulo	28,57	1,00	Not Selected	Selected
2	aventura	comédia	mpb	36 - 45	Rio de Janeiro	4,76	1,00	Selected	Not Selected
3	.	.	.	36 - 45	Rio de Janeiro	.	,00	.	Not Selected
4	terror	.	rock	36 - 45	São Paulo	.	,00	.	Selected
5	aventura	comédia	mpb	36 - 45	São Paulo	28,57	,00	Selected	Selected
6	comédia	comédia	clássica	26 - 35	Brasilia	9,09	,00	Selected	Not Selected
7	aventura	.	gospel	26 - 35	Rio de Janeiro	6,98	,00	Selected	Not Selected
8	aventura	comédia	rock	26 - 35	Rio de Janeiro	27,08	,00	Not Selected	Not Selected
9	suspense	comédia	rock	<= 25	São Paulo	25,37	,00	Selected	Selected
10	suspense	.	rock	<= 25	São Paulo	27,27	,00	Selected	Selected
11	comédia	comédia	rock	<= 25	São Paulo	25,00	,00	Selected	Selected
12	comédia	comédia	rock	<= 25	São Paulo	.	,00	Not Selected	Selected
13	o científica	.	samba	<= 25	São Paulo	20,00	,00	Selected	Selected
14	suspense	comédia	rock	<= 25	Rio de Janeiro	31,03	,00	Selected	Not Selected
15	suspense	comédia	gospel	<= 25	Brasilia	33,33	,00	Selected	Not Selected
16	aventura	comédia	rock	<= 25	Brasilia	47,37	,00	Not Selected	Not Selected
17	comédia	comédia	rock	<= 25	São Paulo	35,48	,00	Not Selected	Selected

Visualização de variável

Alternativamente, você pode utilizar a variável *cidade2*, construída na recodificação automática do Capítulo III, para fazer a seleção. Todavia, a condição acima precisa ser alterada para

$cidade2 = 3$

já que a variável é **numérica**, embora possua os respectivos rótulos de valores: o número 3 está codificado como 'São Paulo'. É obrigatório que toda variável numérica que possua rótulos de valores (nominal ou ordinal) seja incluída em um filtro de seleção a partir de seus valores numéricos.

Finalmente, a partir do procedimento **Descritivas**, temos as medidas-resumo para a variável *trabalho*, levando-se em conta apenas os habitantes da cidade de São Paulo na análise (ver Figura 86). Ao todo, são 15 indivíduos paulistanos que trabalham, em média, 39,6 horas por semana, com valores que variam entre 20 e 56 horas semanais.

Figura 86 | Estatísticas descritivas para a variável TRABALHO após a nova seleção de casos

Estatística Descritiva					
	N	Mínimo	Máximo	Média	Erro Desvio
Horas de Trabalho por Semana	15	20,0	56,0	39,533	9,6278
N válido (de lista)	15				

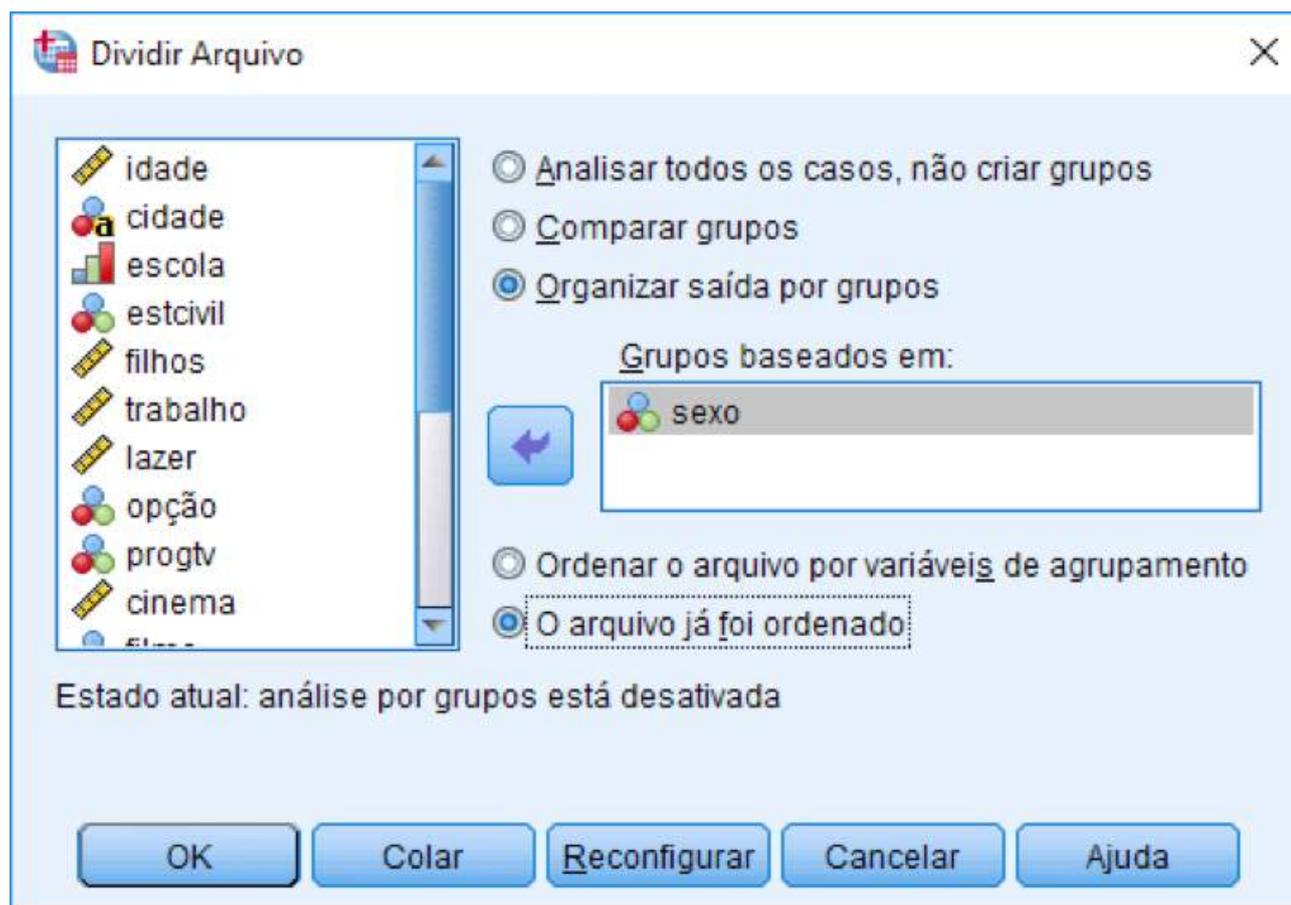
### 05.3 Procedimento de Divisão do Arquivo (Split File)

Vamos supor agora que, após uma série de análises, chegamos à conclusão de que o comportamento dos homens e das mulheres é completamente diferente com relação às preferências para horas de trabalho. Não faz sentido, portanto, apresentar a análise do questionário de opinião sobre lazer com os homens e mulheres juntos. No fundo, o que pretendemos fazer, daqui para frente, são duas análises idênticas, uma para cada gênero.

Para esse tipo de situação, podemos utilizar o procedimento de divisão de arquivo (ou **Split File**), presente no menu **Dados** → **Arquivo Dividido** (ver Figura 87). Como padrão, analisamos todos os casos juntos, sem separação por grupos. Por esse motivo, a opção selecionada na janela é **Analisar todos os casos**. Podemos separar a análise de acordo com as categorias da variável escolhida (ou combinação das categorias de várias variáveis, se for o caso), escolhendo **Comparar grupos** ou **Organizar saída por grupos**. A opção **Comparar grupos** faz com que os grupos sejam apresentados juntos para comparação e a opção **Organizar saída por grupos** mostra os resultados de cada procedimento do IBM SPSS Statistics organizados separadamente para cada grupo.

Após escolher uma das duas últimas opções, moveremos para o quadro **Grupos baseados em**: a variável (ou variáveis) que definirão os grupos para os quais a análise deve ser repetida. Se mais do que uma variável for selecionada, os grupos serão definidos pela combinação das categorias de todas as variáveis. Podemos ainda escolher se o banco de dados deve ser ordenado pela variável que definirá os grupos (**Ordenar o arquivo por variáveis de agrupamento**) ou se nenhuma ordenação deve ser feita (**O arquivo já foi ordenado**). Tenha em mente que, se o arquivo não estiver ordenado por essas variáveis, o procedimento Dividir Arquivo não funcionará corretamente.

Figura 87 | Estatísticas descritivas para a variável TRABALHO após a nova seleção de casos



No nosso caso, selecionamos **Organizar saída por grupos**, escolhemos a variável *sexo* do arquivo *Lazer.sav* e a movemos para o quadro **Grupos baseados em**: e, finalmente, clicamos em **OK**. A única mudança que acontece na janela de dados é a mensagem **Divisão por sexo** na barra de status, no canto inferior direito. Porém, para qualquer análise ou gráfico feito a partir de

agora, dois resultados serão gerados, um para os homens e outro para as mulheres. A Figura 88 apresenta o resultado de uma análise descritiva feita através do menu **Analisar** → **Estatística Descritiva** → **Descritivas** para a variável *trabalho* usando a opção **Organizar saída por grupo**.

Note que os resultados são apresentados para cada categoria da variável *sexo*.

Figura 88 | Resultado da análise descritiva para a variável TRABALHO após a divisão por SEXO

## Descritivos

### Sexo = masculino

#### Estatística Descritiva<sup>a</sup>

	N	Mínimo	Máximo	Média	Erro Desvio
Horas de Trabalho por Semana	19	,0	70,0	39,632	13,7162
N válido (de lista)	19				

a. Sexo = masculino

### Sexo = feminino

#### Estatística Descritiva<sup>a</sup>

	N	Mínimo	Máximo	Média	Erro Desvio
Horas de Trabalho por Semana	11	9,0	56,0	34,909	13,0878
N válido (de lista)	11				

a. Sexo = feminino

Podemos mudar de ideia e querer usar todas as observações para o cálculo das estatísticas. Para cancelar o procedimento **Dividir Arquivo**, basta ativar a opção **Analisar todos os casos** presente na janela de definição da opção **Arquivo Dividido** no menu **Dados**. Repare que a mensagem **Divisão por sexo** na barra de menus desaparecerá, indicando que as análises não estão mais divididas por alguma variável.

## 05.4 Procedimento Agregar (Aggregate)

O procedimento **Agregar** combina grupos de casos em um único caso e cria um novo banco de dados agregados. Os casos são agregados com base nos valores de uma ou mais variáveis.

O novo arquivo conterá um caso para cada grupo. Por exemplo, suponha que você tenha um banco de dados onde cada caso representa um município do Brasil; você poderia agregar esses dados criando um novo arquivo com uma linha para cada Estado.

Como exemplo, usaremos o arquivo *hospital.sav*, que contém dados sobre pacientes que foram internados em um hospital: a data de internação e o valor que custou essa internação. Nosso objetivo é criar um arquivo que contenha uma linha para cada paciente, tendo como variáveis a data da primeira e da última internação, além do valor total e valor médio das internações.

Para isso, devemos selecionar a opção **Agregar...** do menu **Dados**. Na janela aberta (ver Figura 89), temos que escolher uma ou mais variáveis que definirão os grupos (**Variável(is) de Quebra**); no nosso caso, escolheremos paciente, que é a variável que contém o código de cada *paciente*. No quadro **Variáveis Agregadas**, colocamos as variáveis que queremos no novo arquivo; note que a função padrão usada para agregar os dados para variáveis escalares é a média (*mean*), mas podemos mudar essa função para qualquer variável, clicando em **Função** (ver Figura 90). Além disso, você pode colocar a mesma variável diversas vezes no quadro **Variáveis Agregadas**, o que significa que você pode ter, no novo arquivo, várias colunas representando funções diferentes dessa variável.

No nosso exemplo, colocaremos a variável *data\_int* duas vezes no quadro **Variáveis agregadas** e selecionaremos, respectivamente, as funções **Mínimo e Máximo** dentro da opção **Função** (para selecionar a função escolhida em definitivo, clique em **Continuar**). Repare que o banco de dados já está ordenado pelas variáveis paciente e *data\_int*, o que significa que as funções **Mínimo e Primeiro** (assim como o par **Máximo e Último**) são equivalentes neste caso. Entretanto, de forma geral, a função **Primeiro** mostra o primeiro registro do banco de dados que está associado à(s) variável(is) de quebra, enquanto que a função **Último** mostra o último registro associado.

Da mesma forma, adicionaremos a variável *valor\_int* duas vezes no quadro **Variáveis agregadas** e selecionaremos a função **Soma** a partir do quadro **Função** para a primeira variável *valor\_int*, uma vez que não há a necessidade de solicitar novamente a média para a segunda variável. Repare também que você pode alterar o nome e o rótulo de cada variável agregada a partir da opção **Nome e Rótulo**.

Você pode ainda criar uma variável no novo arquivo que contenha o número de casos que foram agregados por grupo. No nosso caso, chamaremos essa variável de *N\_Int*, representando o número de internações que estão sendo estudadas para cada paciente.

Finalmente, deve ser escolhido como será mostrado o resultado do procedimento **Agregar**, que poderá ser de três maneiras diferentes: adicionar o resultado do **Agregar** ao arquivo atual (opção **Adicionar variáveis agregadas ao conjunto de dados ativo**), mostrar o resultado em uma nova janela de dados sem que essa nova janela seja salva (opção **Criar um novo conjunto de dados contendo apenas as variáveis agregadas**) ou salvar o resultado em um novo arquivo de dados (opção **Gravar um novo arquivo de dados contendo apenas as variáveis agregadas**). O conjunto de dados que será criado usando a opção **Criar um novo conjunto de dados contendo apenas as variáveis agregadas** será nomeado como *agregado*.

No quadro **Opções para conjuntos de dados muito grandes**, podemos pedir para que o conjunto de dados seja ordenado ou informar que o arquivo já está ordenado (opção marcada na Figura 89) pela(s) variável(is) de quebra. O procedimento **Agregar** precisa ordenar o conjunto de dados antes de gerar as variáveis agregadas, o que pode aumentar significativamente o tempo de execução do processo se há muitos casos. O resultado final do procedimento pode ser visualizado na Figura 91.

Figura 89 | Janela para definição do procedimento Agregar

**Agregar dados**

Variável(is) de Quebra: paciente

Variáveis agregadas

Resumos de variável(is):

data\_int\_min = MIN(data\_int)  
 data\_int\_max = MAX(data\_int)  
 valor\_int\_sum = SUM(valor\_int)  
 valor\_int\_mean = MEAN(valor\_int)

Número de casos Nome: N\_Int

Salvar

Adicionar variáveis agregadas ao conjunto de dados ativo

Criar um novo conjunto de dados contendo apenas as variáveis agregadas

Nome do conjunto de dados: agregado

Gravar um novo arquivo de dados contendo apenas as variáveis agregadas

Arquivo... C:\Users\estefano\OneDrive\DMSS - Arquivos\Ap...laggr.sav

Opções para conjuntos de dados muito grandes

O arquivo já foi classificado em variável(is) de quebra

Ordenar arquivo antes de agregar

OK Colar Reconfigurar Cancelar Ajuda

**Agregar dados: função agregada**

Estadísticas básicas	Valores específicos	Número de casos
<input type="radio"/> Média	<input type="radio"/> Primeiro	<input type="radio"/> Ponderado
<input type="radio"/> Mediana	<input type="radio"/> Último	<input type="radio"/> Ponderado omisso
<input type="radio"/> Soma	<input checked="" type="radio"/> Mínimo	<input type="radio"/> Não ponderado
<input type="radio"/> Desvio padrão	<input type="radio"/> Máximo	<input type="radio"/> Não ponderado omisso

Porcentagens, Frações, Contagens

Porcentagens  Frações  Contagens

Acima

Abaixo Valor:

Dentro

Fora Baixo:  Alto:

Continuar Cancelar Ajuda

Figura 91 | Resultado do procedimento Agregar para o arquivo hospital.sav

	paciente	data_int_min	data_int_max	valor_int_sum	valor_int_mean	N_int
1	1	02-Jan-96	01-Jan-97	2290,00	572,50	4
2	2	09-Jan-95	02-Feb-96	1198,00	399,33	3
3	3	01-Jan-97	01-Jan-97	741,00	741,00	1
4	4	04-Dec-96	03-Jan-97	800,00	266,67	3
5	5	04-Jul-96	04-Jul-96	789,00	789,00	1
6	6	04-Jun-96	21-Oct-96	857,00	428,50	2
7	7	01-Feb-96	03-May-96	666,00	333,00	2
8	8	08-Aug-96	12-Oct-96	1876,00	469,00	4
9	9	01-Apr-96	04-Jan-97	1391,00	463,67	3
10	10	09-Dec-95	01-Jan-97	1211,00	403,67	3
11	11	02-May-95	03-Jan-97	410,00	205,00	2
12	12	03-Apr-96	03-Apr-96	238,00	238,00	1
13	13	01-Apr-96	01-Apr-96	320,00	320,00	1
14						
15						
16						

## 05.5 Procedimento Reestruturar

O procedimento **Reestruturar** é utilizado para reorganizar seus dados. Por intermédio, você poderá transformar as variáveis selecionadas em casos, os casos selecionados em variáveis ou transpor todos os dados (as linhas viram colunas e as colunas viram linhas).

No nosso exemplo, vamos utilizar o comando **Reestruturar**, presente no menu **Dados**, para reorganizar o arquivo *hospital.sav*, transformando os casos selecionados em variáveis. Esse arquivo contém informações sobre 13 pacientes de um hospital, contendo uma variável chamada *paciente* (representando o código do paciente), uma variável chamada *data\_int* (data da internação do paciente) e uma variável chamada *valor\_int* (valor da internação), no total de 30 linhas. Cada linha do arquivo representa uma internação do paciente, ou seja, cada paciente pode aparecer mais que uma vez, dependendo do número de vezes que ele foi internado. O nosso objetivo é transformar esse banco de dados de modo que ele fique com apenas um paciente por linha e, nas colunas, devemos ter cada data de internação e valor da internação desse paciente.

Logo de cara, muita gente pode imaginar que isso seria facilmente resolvido pelo comando **Agregar**; porém, pelo **Agregar**, não conseguimos reproduzir todas as datas de internação - por isso, devemos escolher o comando **Reestruturar...** no menu **Dados**. Na janela aberta do **Assistente de Reestruturação de Dados** (ver Figura 92), devemos escolher o tipo de reestruturação que queremos fazer nos dados. No nosso caso, escolheremos a opção **Reestruturar casos selecionados em variáveis** para transformar os casos selecionados em variáveis e em seguida escolheremos a opção **Próximo**.

No próximo passo (ver Figura 93, definiremos como as variáveis do arquivo original poderão ser utilizadas no novo arquivo. Temos que definir uma ou mais variáveis que definirão os grupos (**Variável(is) Identificadora(s)**): no nosso caso, escolheremos a variável *paciente*, que é a variável que contém o código de cada paciente que foi internado - dessa forma cada código de paciente será agrupado (consolidado) em uma única linha. Quando necessário, podemos também escolher uma variável *índice* (**Variável(is) de índice**). Após selecionarmos as variáveis que iremos trabalhar, podemos clicar em **Próximo**.

Para utilizar o comando **Reestruturar**, devemos ter o banco de dados ordenado pelas variáveis que definirão os grupos (**Variável(is) identificadoras**); no nosso exemplo, escolhemos a variável *paciente*. Se o banco de dados já estiver ordenado por essa variável podemos escolher a opção **Não – usar os dados como classificados no momento** na janela **Ordenando Dados** (passo 3 do comando **Reestruturar**); caso você não tenha certeza de que o arquivo esteja ordenado, escolha a opção **Sim – os dados serão classificados pelas variáveis identificadoras** (ver Figura 94), para o comando fazer essa ordenação. No nosso caso, escolheremos **Não** e, em seguida, clicaremos em **Próximo**.

Figura 92 | Janela inicial do comando Reestruturar

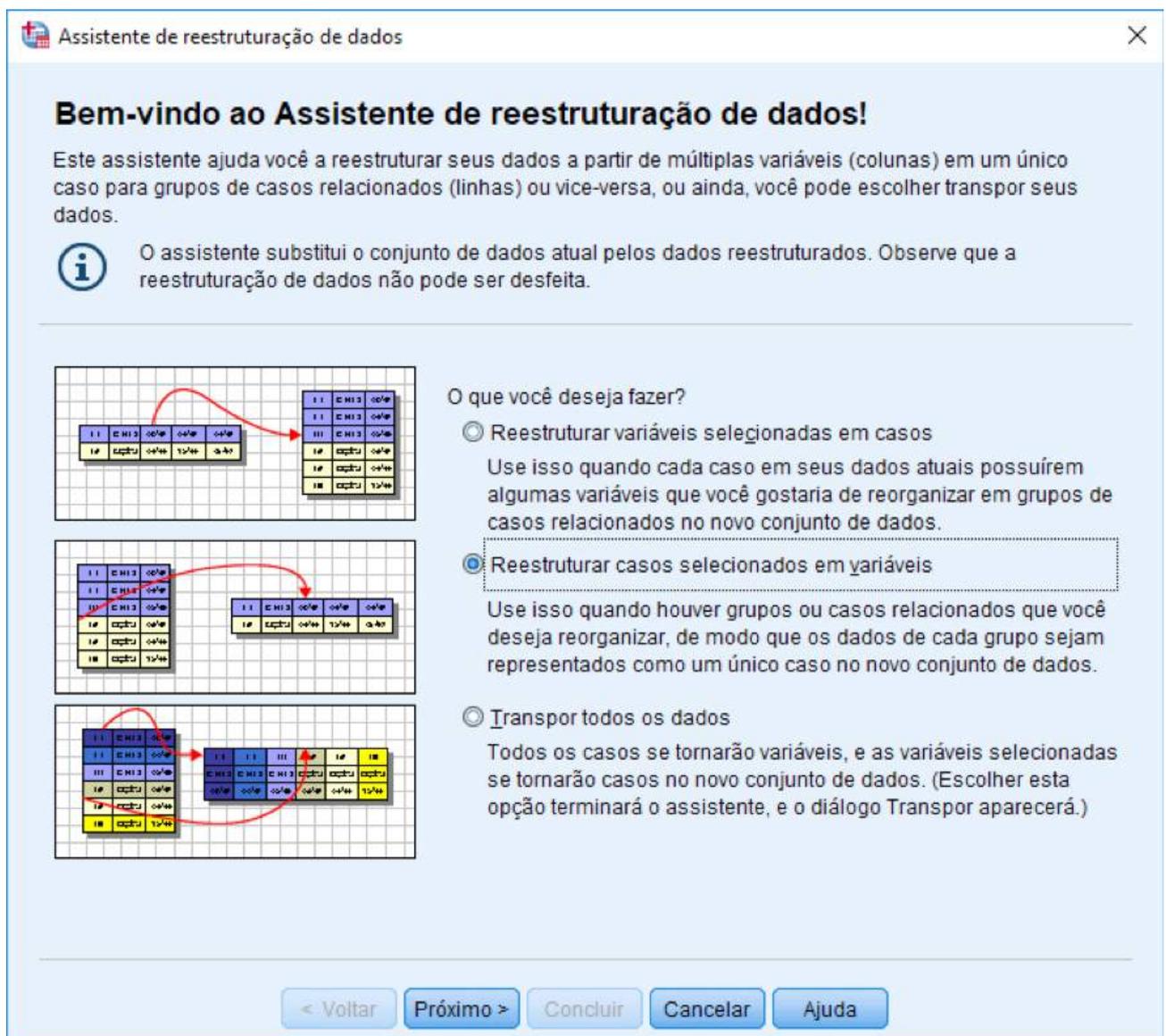
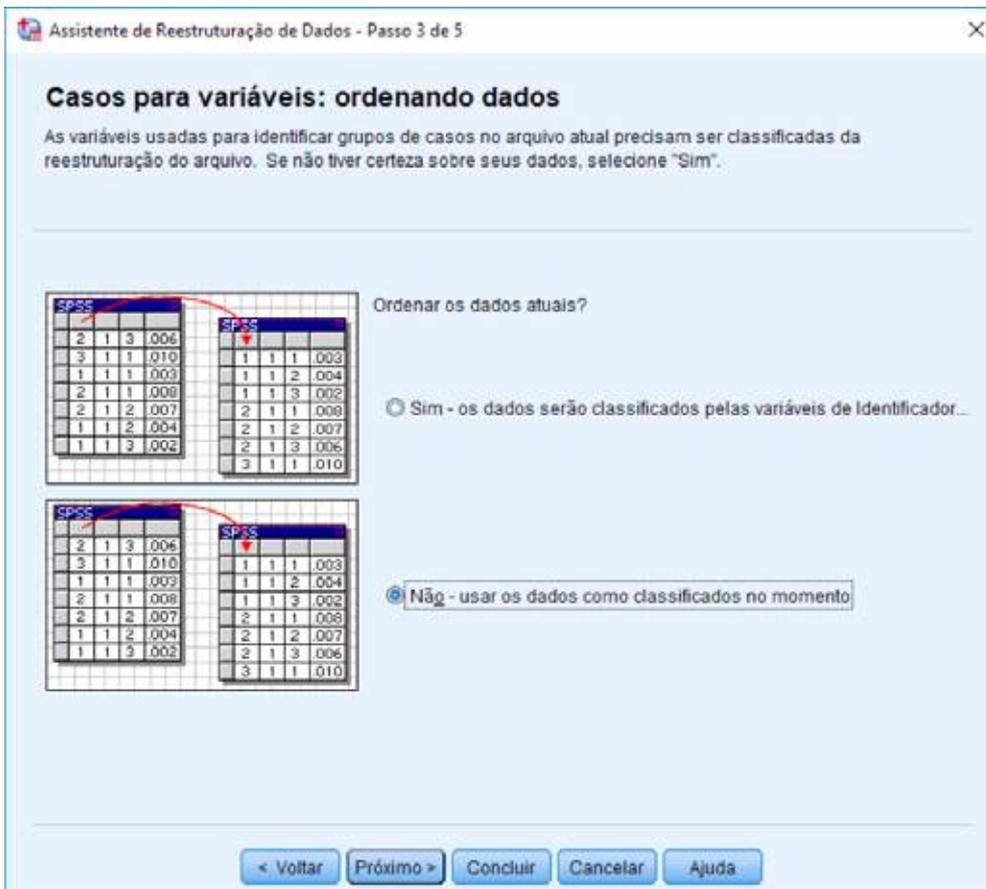


Figura 93 | Passo 2 do comando Reestruturar – Selecionar Variáveis



Figura 94 | Passo 3 do comando Reestruturar – Ordenando Dados



No passo 4 (ver Figura 95), podemos estabelecer algumas opções que serão aplicadas ao arquivo reestruturado tais como ordem que ficarão os novos grupos de variáveis (**Ordem dos novos grupos de variáveis**), uma variável frequência que indicará o número de linhas do arquivo original que foram utilizadas para criar uma linha do arquivo reestruturado (**Variável de contagem de caso**) e uma variável indicadora (**Variáveis indicadoras**). Nessa etapa, podemos escolher a opção **Agrupar por índice** e, em **Variável de contagem de caso**, definiremos a variável *N\_int* para representar o número de internações de cada paciente. Em seguida, clicaremos em **Próximo**.

Finalmente, chegamos ao último passo do comando **Reestruturar**. Nessa janela (ver Figura 96), devemos definir se vamos reestruturar os dados (**Reestruturar os dados agora**) ou se queremos criar uma sintaxe com as opções que foram escolhidas no assistente do comando **Reestruturar** (**Colar a sintaxe gerada pelo assistente em uma janela de sintaxe**). Escolheremos a primeira opção e, em seguida, clicaremos em **Concluir**.

O novo arquivo aparecerá na janela de dados do IBM SPSS Statistics mostrando apenas um único paciente por linha e, nas variáveis, todas as datas de internação de cada paciente (ver Figura 97).

Figura 95 | Passo 4 do comando Reestruturar – Opções

Assistente de Reestruturação de Dados - Passo 4 de 5

### Casos para variáveis: opções

Nesta passo, você pode definir as opções que serão aplicadas ao arquivo de dados reestruturado.

Ordem dos novos grupos de variáveis

Agrupar por variável original (por exemplo: w1 w2 w3, h1 h2 h3)

Agrupar por índice (por exemplo: w1 h1, w2 h2, w3 h3)

Variável de contagem de caso

Faça a contagem do número de casos nos dados atuais usados para criar um novo caso

Nome:

Rótulo:

Variáveis indicadoras

Criar variáveis indicadoras

Nome raíz:

< Voltar Próximo > Concluir Cancelar Ajuda

Figura 96 | Janela final do comando Reestruturar

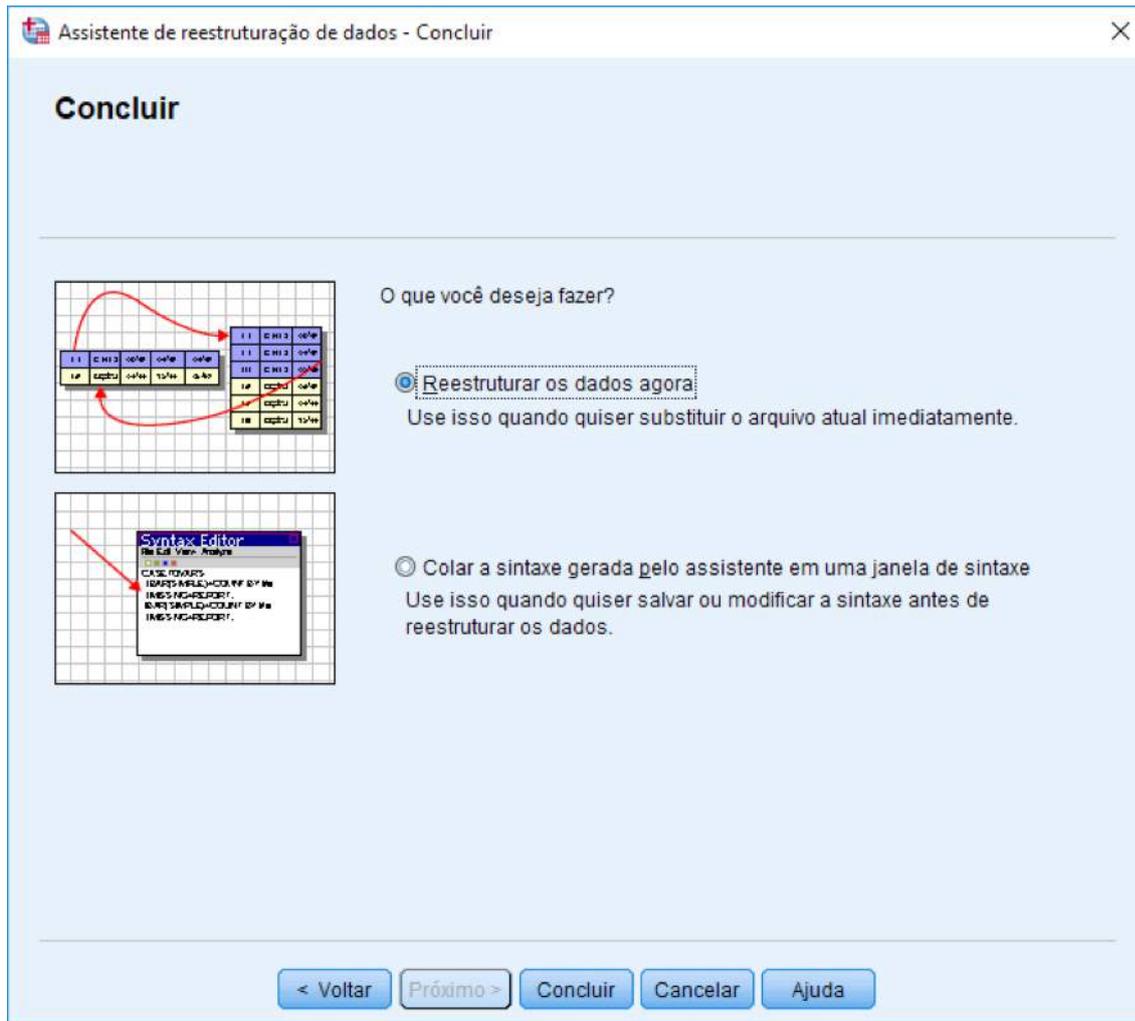


Figura 97 | Banco de dados criado através do comando Reestruturar

	paciente	N_int	data_int.1	valor_int.1	data_int.2	valor_int.2	data_int.3
1	1	4	02-Jan-96	520,00	03-Mar-96	230,00	05-Nov-96
2	2	3	09-Jan-95	350,00	04-Mar-95	390,00	02-Feb-96
3	3	1	01-Jan-97	741,00	.	.	.
4	4	3	04-Dec-96	351,00	12-Dec-96	326,00	03-Jan-97
5	5	1	04-Jul-96	789,00	.	.	.
6	6	2	04-Jun-96	536,00	21-Oct-96	321,00	.
7	7	2	01-Feb-96	124,00	03-May-96	542,00	.
8	8	4	08-Aug-96	123,00	05-Sep-96	324,00	07-Oct-96
9	9	3	01-Apr-96	280,00	20-Nov-96	370,00	04-Jan-97
10	10	3	09-Dec-95	321,00	05-Jul-96	236,00	01-Jan-97
11	11	2	02-May-95	170,00	03-Jan-97	240,00	.
12	12	1	03-Apr-96	238,00	.	.	.
13	13	1	01-Apr-96	320,00	.	.	.
14							
15							
16							

Em algumas situações, o banco de dados já vem nesse formato e seria interessante reestruturá-lo para um formato no qual cada internação (com sua data e valor) represente um caso, sempre identificada pelo número do paciente. Faremos o comando **Reestruturar** novamente no conjunto de dados para voltarmos ao formato original do mesmo, no qual havia apenas duas variáveis (*data\_int* e *valor\_int*) e cada internação representava um caso. Para isso, acesse novamente o assistente de reestruturação a partir do comando **Reestruturar** no menu **Dados**, certifique-se de que a primeira opção (**Reestruturar variáveis selecionadas em casos**) esteja selecionada e, a seguir, clique em **Próximo**.

O segundo passo do assistente (ver Figura 98) é dedicado à definição do número de variáveis que serão criadas a partir da reestruturação: no nosso caso, serão duas (*data\_int* e *valor\_int*). Então, clique no botão **Mais de um** (grupo de variáveis) e certifique-se de que o número 2 esteja digitado na respectiva caixa. Clique em **Próximo** a seguir.

A parte crítica do assistente reside no terceiro passo (ver Figura 99), no qual devemos definir a variável que identifica os casos (**Identificação de grupo de caso**), quais variáveis serão transpostas (**Variáveis a serem transpostas**) e, se necessário, quais variáveis não serão alteradas (**Variável(is) controlada(s)**). Na caixa **Identificação de grupo de caso**, selecione a opção **Usar variável selecionada** e selecione a variável paciente. Repare que, na caixa **Variáveis a serem transpostas**, há dois grupos (*trans1* e *trans2*) de forma que, em cada grupo, colocaremos as respectivas variáveis de data e valor das internações.

Figura 98 | Passo 2 da nova execução do comando Reestruturar

Assistente de Reestruturação de Dados - Passo 2 de 7

### Variáveis para casos: número de grupos de variáveis

Você decidiu reestruturar as variáveis selecionadas em grupos de casos relacionados no novo arquivo.

Um grupo de variáveis relacionadas, chamado de grupo de variável, representa as medições de uma v...  
 Por exemplo, a variável pode ser a largura. Se for registrada em três medições separadas, cada uma representando um ponto diferente no tempo, -w1, w2 e w3, então os dados serão reorganizados em um grupo de variáveis.

Se frequentemente houver mais de uma variável no arquivo, ele também será registrado em um grupo de variável, por exemplo, altura, registrada em h1, h2 e h3.

Quantos grupos de variável você deseja reestruturar?

Um (por exemplo, w1, w2 e w3)

Mais de um (por exemplo, w1, w2, w3 e h1, h2, h3, etc.)

Quantos?

< Voltar   Próximo >   Concluir   Cancelar   Ajuda

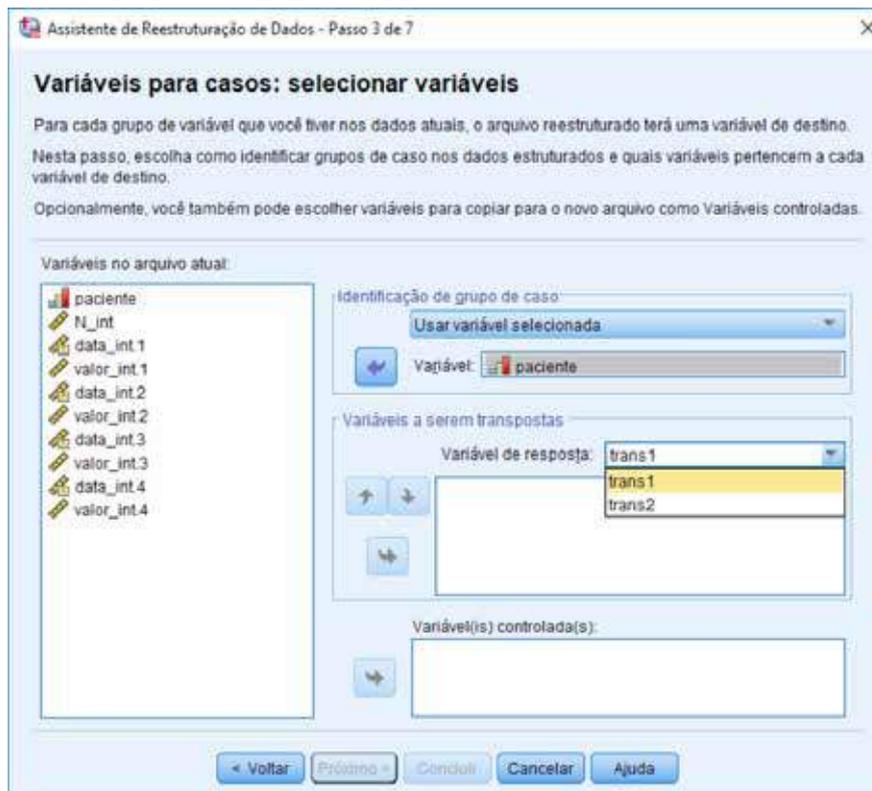
1	2	3	4
1	8	4	3
2	5	6	7

1
8
4
3
2
5
6
7

1	2	3	4	5	6
1	8	4	0.3	0.9	0.4
2	5	6	0.7	0.1	0.7

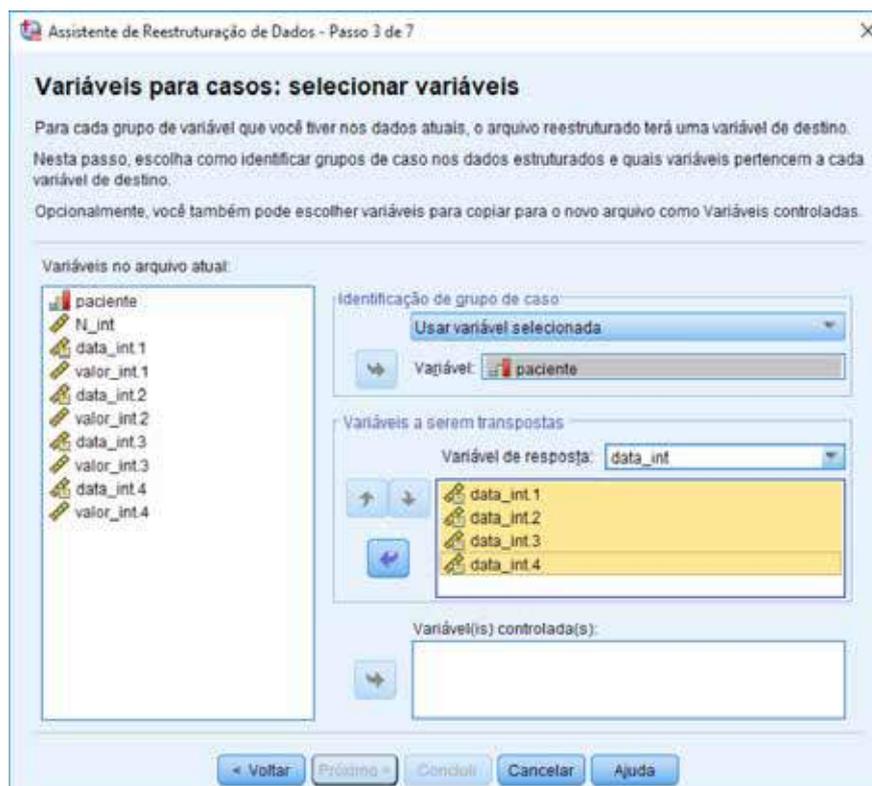
1	0.3
8	0.9
4	0.4
2	0.7
5	0.1
6	0.7

Figura 99 | Passo 3 da nova execução do comando Reestruturar



Na Figura 100, mostramos que o nome da variável pode ser alterado diretamente na caixa - mude o nome de *trans1* para *data\_int* no campo **Variável de resposta** - e, a seguir, selecionamos as quatro variáveis de data (*data\_int.1* a *data\_int.4*, nesta ordem) para a respectiva caixa. Repita o procedimento para *trans2*, mudando seu nome para *valor\_int* e selecionando as variáveis de valor (*valor\_int.1* a *valor\_int.4*, nesta ordem) para a respectiva caixa. Como não exportaremos mais variáveis, clique em **Próximo**.

Figura 100 | Criação da nova variável DATA\_INT



No quarto passo (ver Figura 101), é possível criar variáveis de índice, que são úteis para identificar casos por fatores ou tratamentos. No nosso exemplo, poderíamos criar uma variável de índice para indicar o número da internação associada a determinado paciente; contudo, clique em **Nenhuma** e, a seguir, em **Próximo**.

No penúltimo passo (ver Figura 102), clique em **Soltar variável(is) do novo arquivo de dados** na caixa **Tratamento de variáveis não selecionadas**: esta opção faz com que toda variável que não foi selecionada no terceiro passo seja excluída do banco reestruturado – no nosso caso, a variável *N\_int* será excluída. Na caixa **Valores em branco ou omissos do sistema em todas as variáveis transpostas**, temos a opção de manter os valores ausentes das variáveis originais ou simplesmente apagá-los, o que é a opção recomendada na maioria dos casos. Clique em **Descartar os dados** e, a seguir, clique em **Próximo**.

Figura 101 | Passo 4 da nova execução do comando Reestruturar

Assistente de Reestruturação de Dados - Passo 4 de 7
✕

### Variáveis para casos: criar variáveis de índice

Nos dados atuais, os valores para um grupo de variáveis aparecem em um único caso em variáveis múltiplas. Por exemplo, um único caso contém os valores para w1, w2 e w3.

Nos novos dados, os valores para um grupo de variáveis aparecerão em múltiplos casos em uma única variável. Por exemplo, haverá três casos, um para cada w1, w2 e w3.

Um índice é uma nova variável que identifica o grupo de novos casos que foi criado a partir do caso original. Por exemplo, um índice chamado "w" teria os valores 1, 2 e 3.

1	1	1	0.07
1	1	2	0.11
1	1	3	0.05
2	1	1	0.08
2	1	2	0.04
2	1	3	0.06

1	1	1	1	0.07
1	1	1	2	0.11
1	1	1	3	0.05
1	1	2	1	0.08
1	1	2	2	0.04
1	1	2	3	0.06

1	1	0.08	2	0.07
2	1	0.11	2	0.11
3	1	0.07	2	0.05
4	1	0.06	2	0.08
5	1	0.09	2	0.04
6	1	0.02	2	0.06

Quantas variáveis de índice você deseja criar?

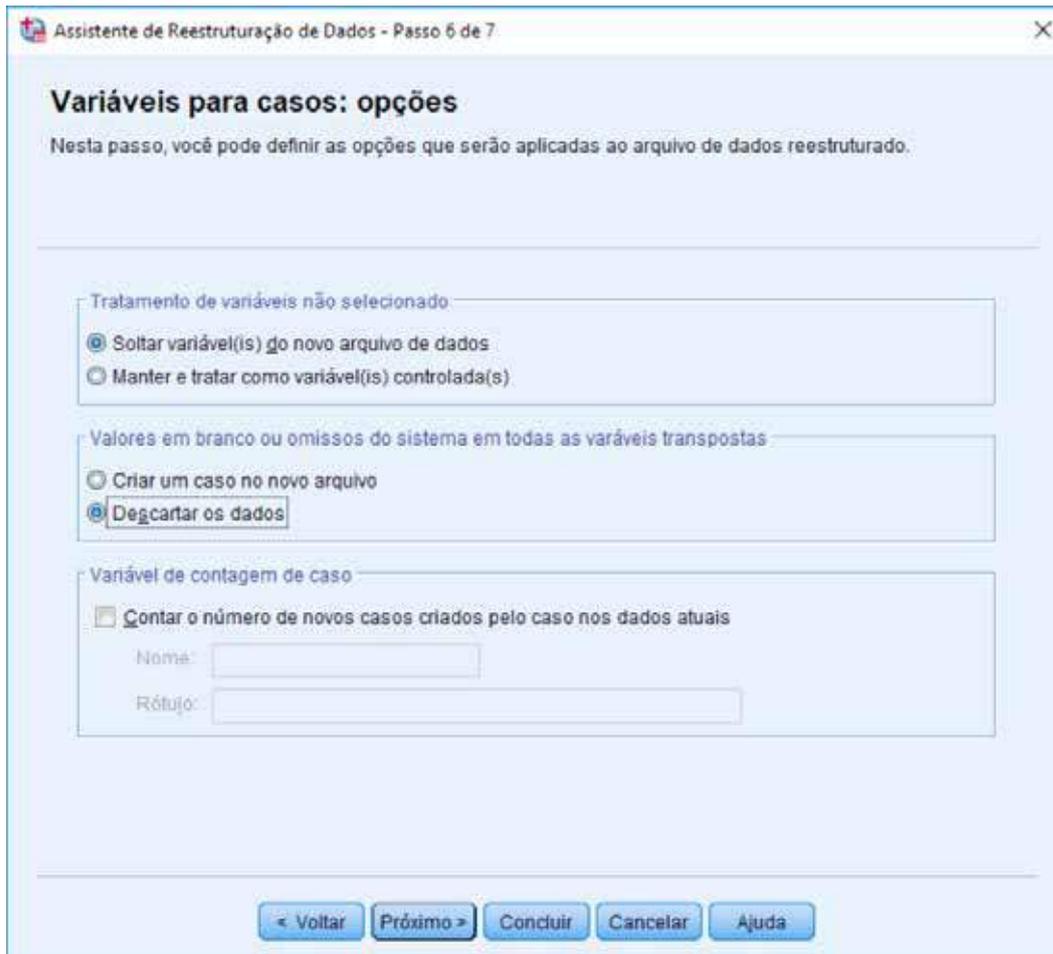
Uma  
Use isso quando um grupo de variável registrar os efeitos de um único fator, tratamento ou condição.

Mais de um    Quantos?   
Use isso quando um grupo de variável registrar os efeitos de mais de um fator, tratamento ou condição.

Nenhuma  
Use isso se as informações do índice forem armazenadas em um dos conjuntos de variáveis a serem transpostos.

< Voltar
Próximo >
Concluir
Cancelar
Ajuda

Figura 102 | Penúltimo passo da nova execução do comando Reestruturar



Mais uma vez, você terá a opção de reestruturar os dados imediatamente ou de salvar a sintaxe em uma janela de sintaxe. Escolha a primeira opção, **Reestruturar os dados agora**, e clique em **Concluir**. O banco de dados criado, que é idêntico ao arquivo *hospital.sav* original, está na Figura 103.

Figura 103 | Banco de dados reestruturado, igual ao arquivo *hospital.sav* original

\*hospital.sav [ConjuntodeDados1] - Editor de dados do IBM SPSS Statistics

Arquivo Editar Visualizar Dados Transformar Analisar Gráficos Customizado Utilitários Extensões Janela Ajuda

Visível: 3 de 3 variáveis

	paciente	data_int	valor_int	var	var	var	var	var	var
1	1	02-Jan-96	520,00						
2	1	03-Mar-96	230,00						
3	1	05-Nov-96	840,00						
4	1	01-Jan-97	700,00						
5	2	09-Jan-95	350,00						
6	2	04-Mar-95	390,00						
7	2	02-Feb-96	458,00						
8	3	01-Jan-97	741,00						
9	4	04-Dec-96	351,00						
10	4	12-Dec-96	326,00						
11	4	03-Jan-97	123,00						
12	5	04-Jul-96	789,00						
13	6	04-Jun-96	536,00						
14	6	21-Oct-96	321,00						
15	7	01-Feb-96	124,00						
16	7	03-May-96	542,00						

Visualização de dados Visualização de variável

O processador do IBM SPSS Statistics está pronto | Unicode:ON

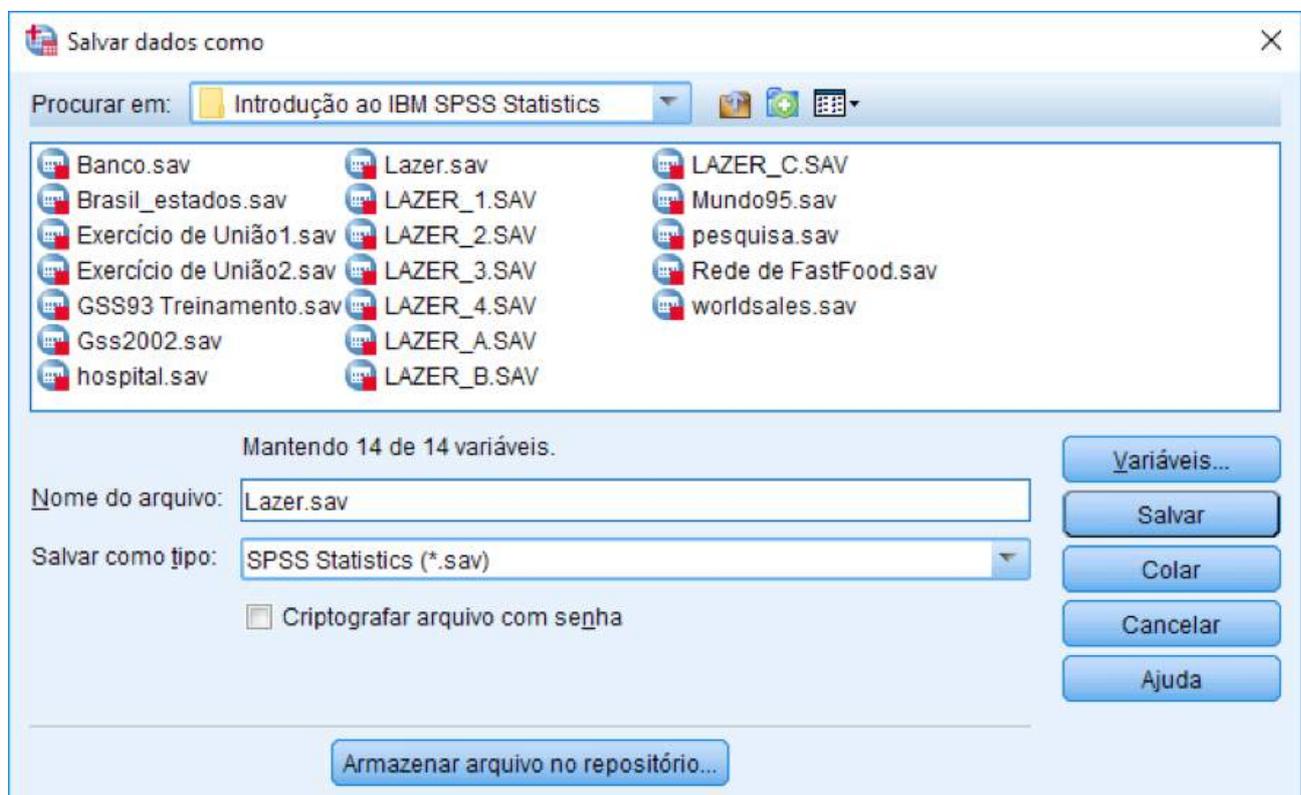
## 06 Arquivos e Janelas

### 06.1 Janela de dados

A janela **Editor de Dados** (ou simplesmente **janela de dados**) é destinada a armazenar os dados que serão usados para efetuar as análises. Você pode abrir várias janelas de dados em uma mesma sessão do IBM SPSS Statistics, mas isso não significa que se você quiser analisar dados, de dois arquivos diferentes, basta abrir uma janela de dados para cada arquivo. Para isso ser possível, os arquivos precisam estar presentes na mesma janela e, por isso, o IBM SPSS Statistics apresenta a opção **Mesclar Arquivos** no menu **Dados**. Esta opção será estudada em detalhes posteriormente.

Como já mencionado anteriormente, o IBM SPSS Statistics faz importação de arquivos de dados no formato Excel, Lotus, Dbase, Sylk, SAS, Stata, arquivo texto com delimitador (vírgulas, tabulações, etc.) e ASCII (arquivo texto com coluna de tamanho fixo), além do próprio formato IBM SPSS Statistics (versões DOS, Windows e *portable*) e através de ODBC.

Figura 104 | Salvando a janela de dados



Na digitação dos resultados de seus estudos na janela de dados, você pode criar seu banco de dados dentro do próprio IBM SPSS Statistics. O arquivo criado pode ser exportado em todos os formatos, nos quais podem ser importados diretamente (sem uso de ODBC). O formato de exportação padrão é o próprio formato IBM SPSS Statistics para Windows (extensão *.sav*).

Para salvar a janela de dados, basta clicar no menu Arquivo, opção Salvar (ver Figura 104). No canto inferior esquerdo, Salvar como tipo, o IBM SPSS Statistics apresenta uma lista com todos os formatos de exportação disponíveis. Basta escolher o de sua preferência (o padrão é o formato SPSS), nomeá-lo na área Nome do arquivo e clicar em OK

## 06.2 Janela de saídas

Quando você pedir uma análise estatística, um gráfico etc., o IBM SPSS Statistics automaticamente abrirá uma janela de saídas chamada **Saída1 [Documento1] – Visualizador do IBM SPSS Statistics**. Os resultados são armazenados de maneira sequencial na ordem em que os mesmos são gerados, porém você pode usar as facilidades da janela de saídas para: procurar saídas, mostrar ou esconder tabelas ou gráficos, mudar a ordem de visualização das saídas movendo itens e mover itens entre o IBM SPSS Statistics e outros programas.

A sua janela de saídas está dividida em dois painéis (veja a Figura 105): O painel da esquerda contém uma visão geral do conteúdo das saídas e é chamado de **painel geral** e o painel da direita contém as tabelas, gráficos e saídas de texto e é chamado **painel dos conteúdos**. Você pode usar as barras de movimentação (**scroll bars**) para procurar as saídas; clicar no item correspondente no painel geral para ir diretamente para o item procurado; clicar e arrastar a borda direita do painel geral para mudar a largura desse painel. Clique num ícone em formato de livro para mostrar ou esconder um item do painel geral.

## 06.3 Formas de exportação da janela de saída

É importante lembrar que o IBM SPSS Statistics permite que você abra quantas janelas de saídas quiser. Cada janela de saídas aberta é um arquivo em potencial que pode ser salvo no formato **Arquivos do Visualizador**, que é o formato próprio do IBM SPSS Statistics para suas janelas de saídas (extensão *.spv*). Para salvar uma janela de saídas, basta clicar no menu **Arquivo**, estando na janela que você quer gravar, e selecionar **Salvar** (ver Figura 106). Temos que escolher o driver e diretório no qual queremos armazenar o arquivo e digitamos o nome com o qual o arquivo será armazenado.

Além de abrir esse arquivo no IBM SPSS Statistics, você também pode exportar as saídas do IBM SPSS Statistics (*output*) como formato HTML, TXT, XLS (ou XLSX), DOC, PPT e PDF (a última opção está disponível desde a versão 15.0 do IBM SPSS Statistics) usando a opção **Exportar...** do menu **Arquivo** (ver Figura 107).

Figura 105 | A janela de saídas e seus painéis

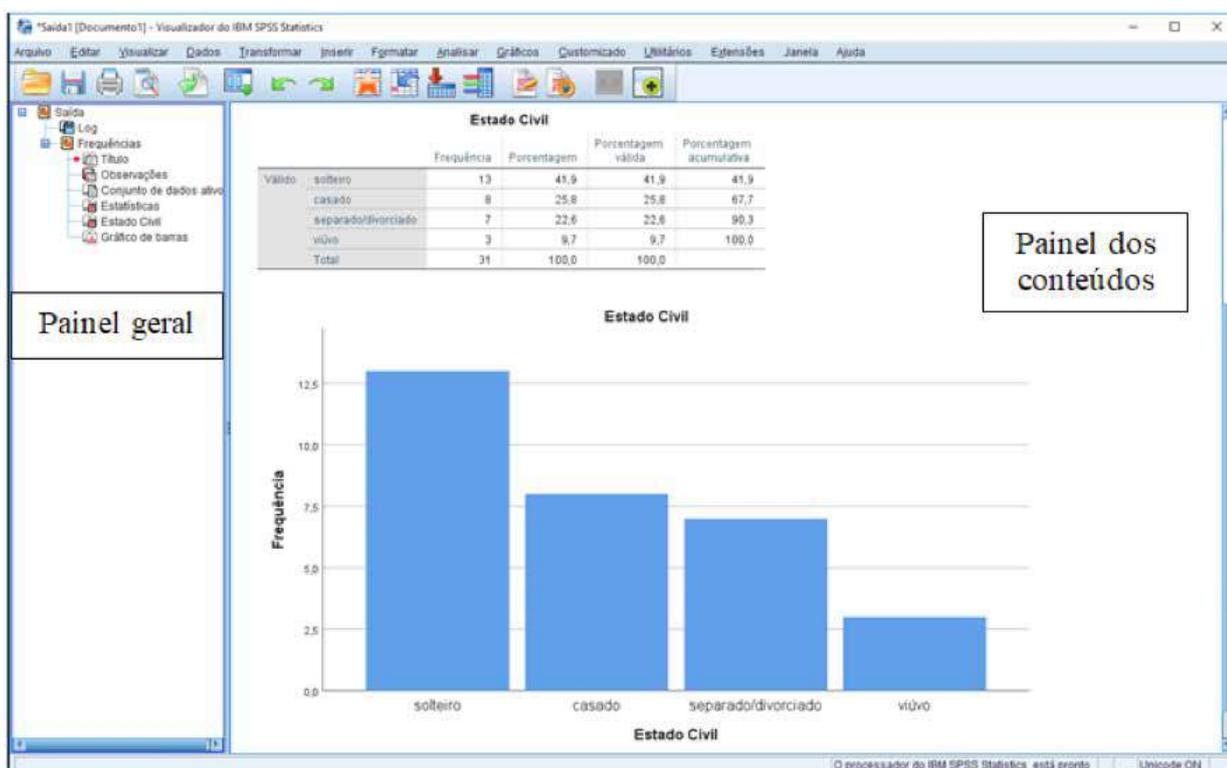
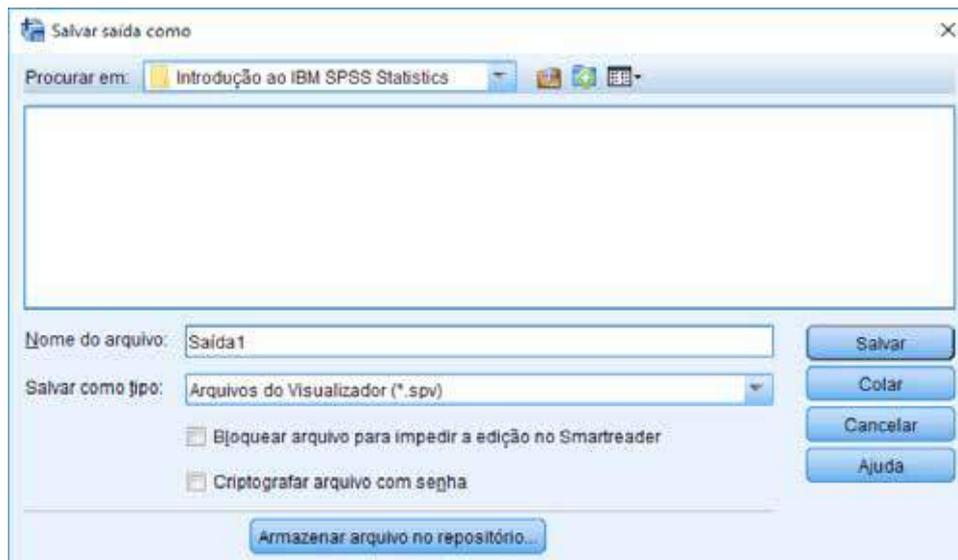


Figura 106 | Salvando uma janela de saídas



Se você escolher exportar as saídas do IBM SPSS Statistics, você tem na opção **Objetos a exportar** a possibilidade de exportar todos os objetos possíveis (**Todos**), apenas os visíveis (aqueles cujo ícone no painel geral é um livro aberto) em **Todos visíveis**, ou apenas objetos que você selecionou (**Selecionado**).

No quadro **Documento**, definimos o formato da exportação, que poderá ser HTML, *Web Report* (HTML5, relatório da web), TXT, Word, Excel, Power Point, PDF ou exportar somente os gráficos que estiverem no *output*. Na opção Gráficos, definimos o formato dos gráficos que estamos exportando (EPS, JPEG, TIFF, PNG, BMP ou EMF). Na Figura VI-5, você tem um exemplo de exportação de uma saída do IBM SPSS Statistics no formato de **Relatório da Web** (HTML5).

Figura 107 | Exportando uma janela de saídas

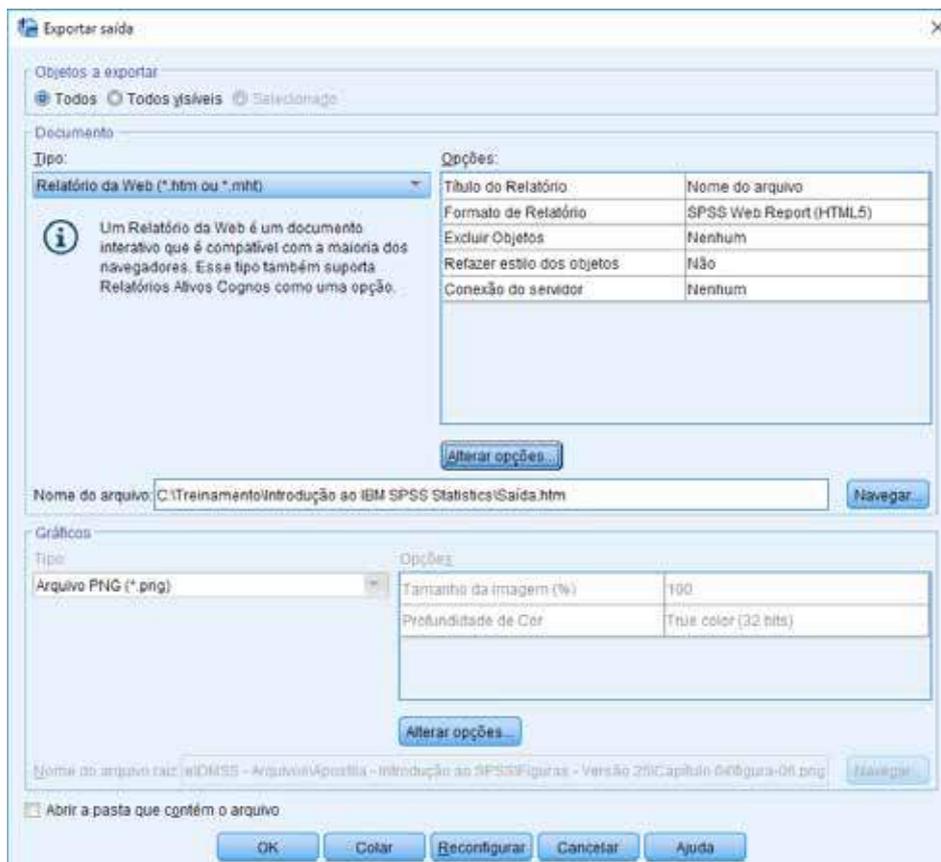


Figura 108 | Saída do IBM SPSS Statistics exportada no formato HTML5 (Relatório da Web)



## 06.4 Janela de sintaxe

As escolhas feitas dentro dos menus do IBM SPSS Statistics podem ser salvas na forma de comandos dentro de uma janela de programação. A vantagem de procedermos dessa maneira é que os procedimentos armazenados nessa janela podem ser executados novamente sem que precisemos entrar nos menus. Por exemplo: se você precisa rodar uma mesma análise todo mês com um arquivo atualizado, você pode tirar proveito dessa facilidade.

É bom lembrar que o IBM SPSS Statistics possui uma linguagem de programação própria e aqueles que a conhecem podem escrever os comandos sem a necessidade de usar o recurso **Color**. Através deste recurso, você não precisa digitar o código referente ao procedimento desejado: o próprio IBM SPSS Statistics escreve o código para você.

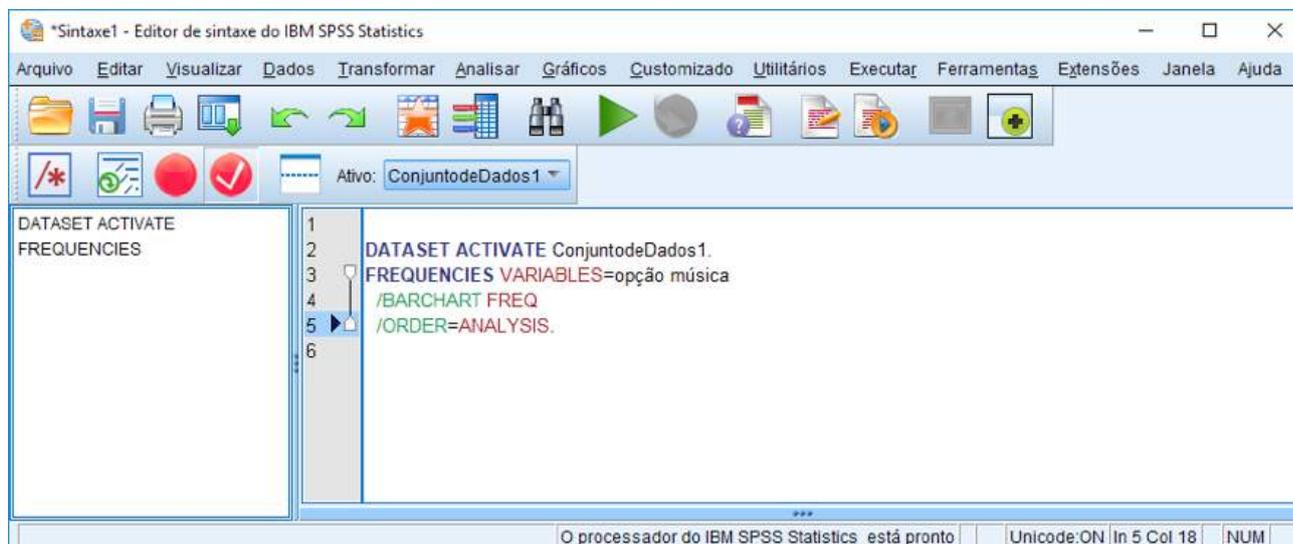
Vamos reproduzir um exemplo simples: a criação de uma janela de programação com a sintaxe referente à criação de uma tabela de frequências a partir do arquivo *Lazer.sav*. Devemos clicar o menu **Analisar**, escolher **Estatística Descritiva** e, depois, **Frequências**. Você pode modificar as opções dessa janela, como na Figura 109 com a opção de incluir um gráfico de barras com **frequências** para cada variável a partir do botão **Gráficos**; porém, agora, ao invés de clicar **OK**, clique em **Color**.

O resultado desse procedimento é a criação de uma janela chamada Sintaxe1 – Editor de Sintaxe do IBM SPSS Statistics, já com o código referente à criação de tabelas de frequências para as variáveis opção e música (ver Figura 110).

Figura 109 | Janela de definição de uma tabela de frequências



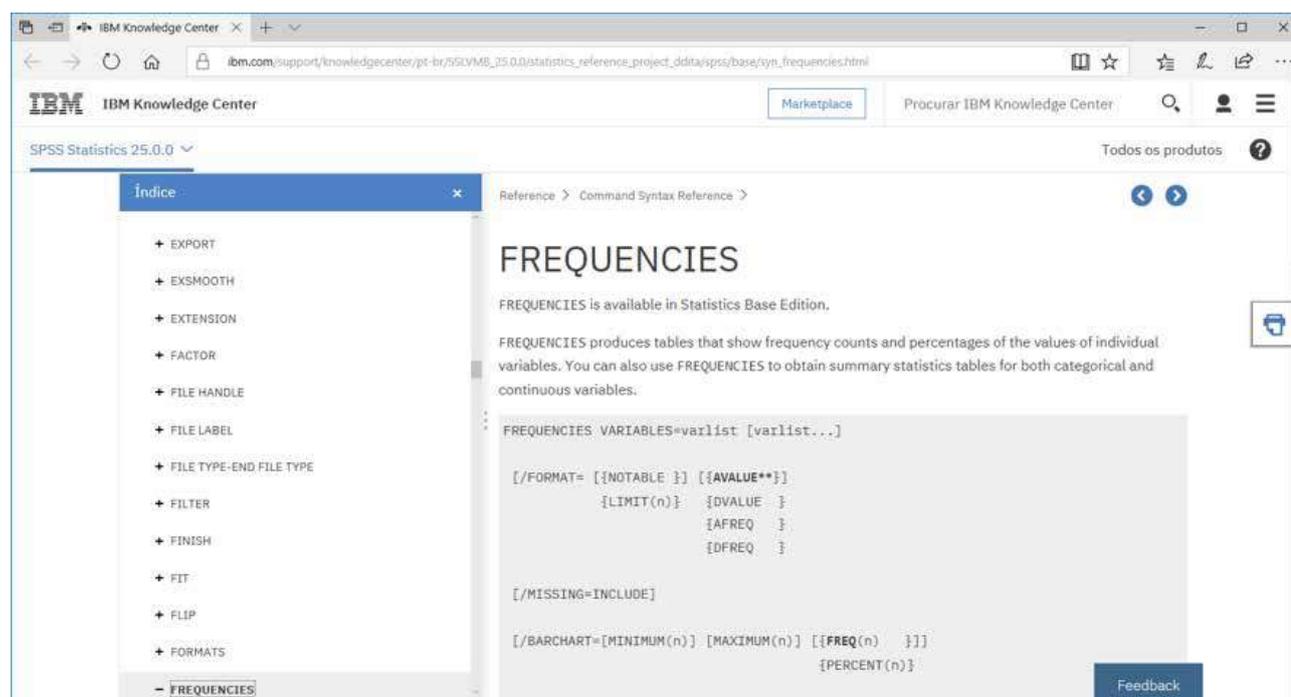
Figura 110 | Comando Frequências na janela de programação



Uma vez que só é possível trabalhar com um *dataset* de cada vez no IBM SPSS Statistics, precisamos declarar qual deles será utilizado: isso é feito a partir do comando DATASET ACTIVATE, no qual declaramos, deste ponto em diante, que toda manipulação dentro da janela de sintaxe terá efeitos no ConjuntoDeDados1 que, no nosso exemplo, corresponde ao arquivo *Lazer.sav*. (Você pode renomear o conjunto de dados a partir da opção **Renomear Conjunto de Dados** do menu **Arquivo** a partir do Editor de Dados).

Para saber quais são as demais especificações disponíveis para o comando FREQUENCIES, marcamos qualquer parte do comando FREQUENCIES e com um clique no ícone **Ajuda de Sintaxe** da barra de ferramentas (ou pressionando a tecla F1 do teclado), obtemos a informação desejada (ver Figura 111).

Figura 111 | Ajuda para definição de comandos na janela de sintaxe

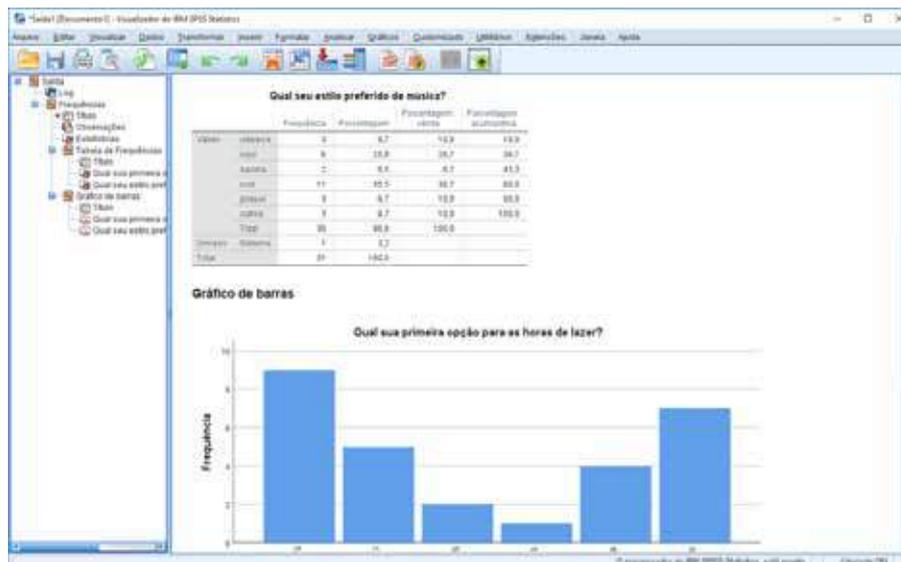


A maioria das opções do comando pode ser acessada através do menu (por exemplo, os subcomandos /STATISTICS e /HISTOGRAM). Todos os códigos entre colchetes são opcionais, e os entre chaves indicam opções que podem ser escolhidas. Ressaltamos que essa janela é parte do sistema de ajuda do IBM SPSS Statistics e, embora mostre todas as opções existentes para cada comando, não apresenta nenhuma explicação sobre o que cada uma faz. Para informações mais detalhadas sobre os comandos do IBM SPSS Statistics, o leitor pode consultar o manual **IBM SPSS Statistics 25 Command Syntax Reference** (disponível no menu **Ajuda... Referência de Sintaxe do Comando**).

O comando FREQUENCIES presente na janela de sintaxe vai produzir exatamente a mesma tabela de frequências feita por meio do menu **Frequências** clicando **OK**. Até agora, apenas o código foi gerado. Para rodar o procedimento basta selecionar o comando desejado com o mouse e clicar no ícone **Executar Seleção**. Note, que no menu Executar, você tem mais opções para rodar seu programa (como, por exemplo, a opção **Todos**, que roda todos os comandos sem que você precise selecionar nenhum anteriormente). Lembre-se que todo comando na janela de sintaxe do IBM SPSS Statistics precisa ser encerrado com um **ponto final**, independentemente do número de linhas necessárias para os subcomandos e outros argumentos na sintaxe.

Podemos continuar nosso programa selecionando outros procedimentos dos menus do IBM SPSS Statistics e depois clicando em **Colar**. A Figura 112 apresenta a janela de saídas com as tabelas de frequência.

Figura 112 | Resultado do comando FREQUENCIES presente na janela de saídas



Uma opção muito útil para quem precisa fazer recodificação das mesmas variáveis com frequência (devido à importação de novos casos a partir de formatos diferentes do padrão do IBM SPSS Statistics) é manter um **dicionário de variáveis**, com seus formatos e rótulos. Na janela de sintaxe, isso é feito com o comando DISPLAY DICTIONARY. A opção SORTED dispõe a lista de variáveis e seus rótulos de valores (no caso de variáveis categóricas) em ordem alfabética: no nosso exemplo, isso seria feito executando o comando DISPLAY SORTED DICTIONARY (ver Figura 113) e o resultado final está na Figura VI-11. Alternativamente, podemos selecionar algumas variáveis para nosso dicionário (por meio do subcomando /VARIABLES) e as mesmas que aparecerão na ordem definida pelo usuário: no nosso exemplo, isso seria feito executando o comando.

DISPLAY DICTIONARY /VARIABLES= idade cidade sexo trabalho lazer.

Ver Figura 113 e o resultado final do nosso exemplo está na Figura 115.

Figura 113 | Janela de sintaxe com os dois comandos DISPLAY DICTIONARY

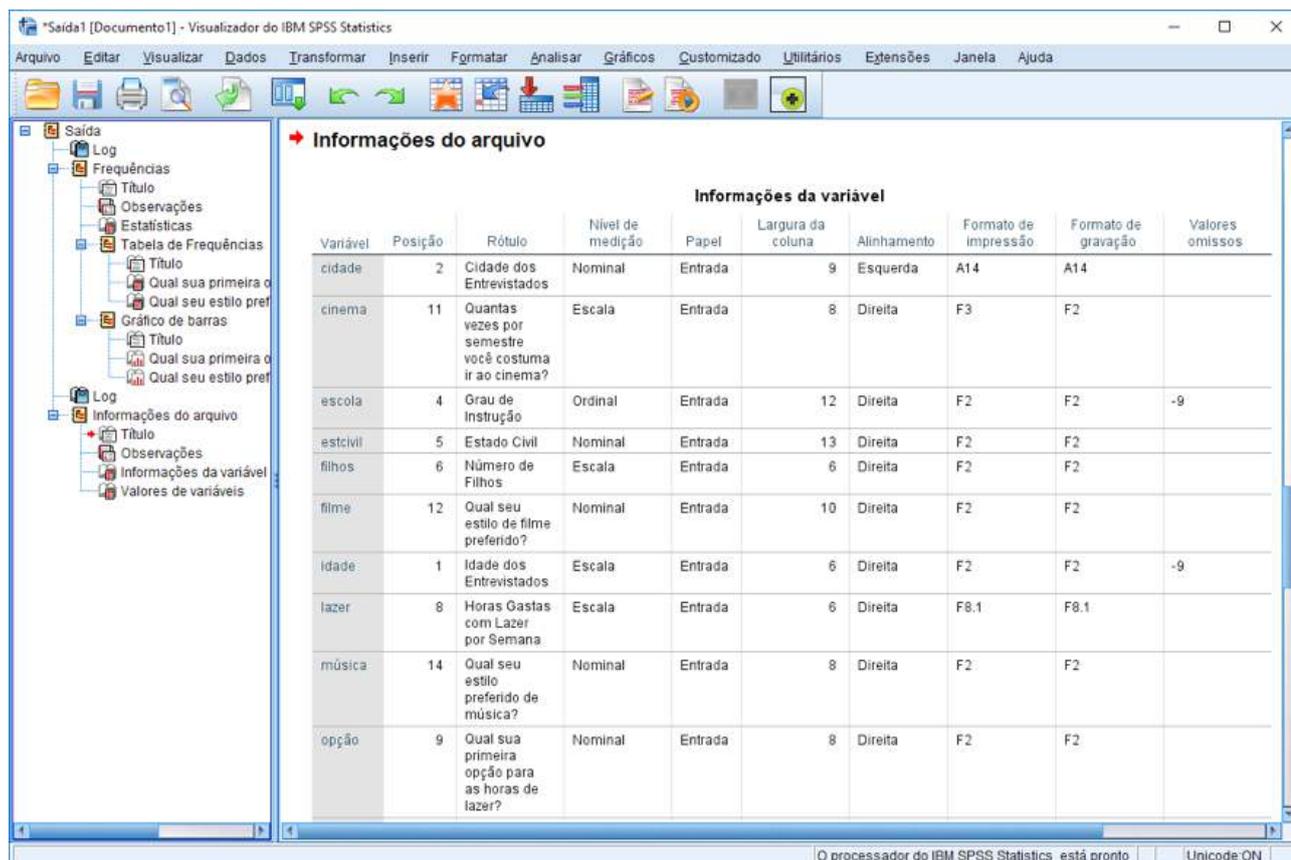
The screenshot shows the IBM SPSS Statistics Syntax Editor window. The window title is "\*Sintaxe1 - Editor de sintaxe do IBM SPSS Statistics". The menu bar includes Arquivo, Editar, Visualizar, Dados, Transformar, Analisar, Gráficos, Customizado, Utilitários, Executar, Ferramentas, Extensões, Janela, and Ajuda. The active dataset is "ConjuntodeDados1". The syntax editor shows the following commands:

```

1 DATASET ACTIVATE
2 FREQUENCIES
3 DISPLAY
4 DISPLAY
5
6
7
8
9 DISPLAY DICTIONARY /VARIABLES= idade cidade sexo trabalho lazer.
10

```

Figura 114 | Dicionário de variáveis ordenadas do arquivo *Lazer.sav*



A seguir, podemos exportar a saída para qualquer formato que seja conveniente (ver Figura 116). Como o dicionário de variáveis é exibido em formato de tabela, é recomendável que a saída selecionada seja exportada para o formato padrão do Microsoft Excel (extensões *.xls* ou *.xlsx*).

Figura 115 | Dicionário das variáveis selecionadas no arquivo *Lazer.sav*

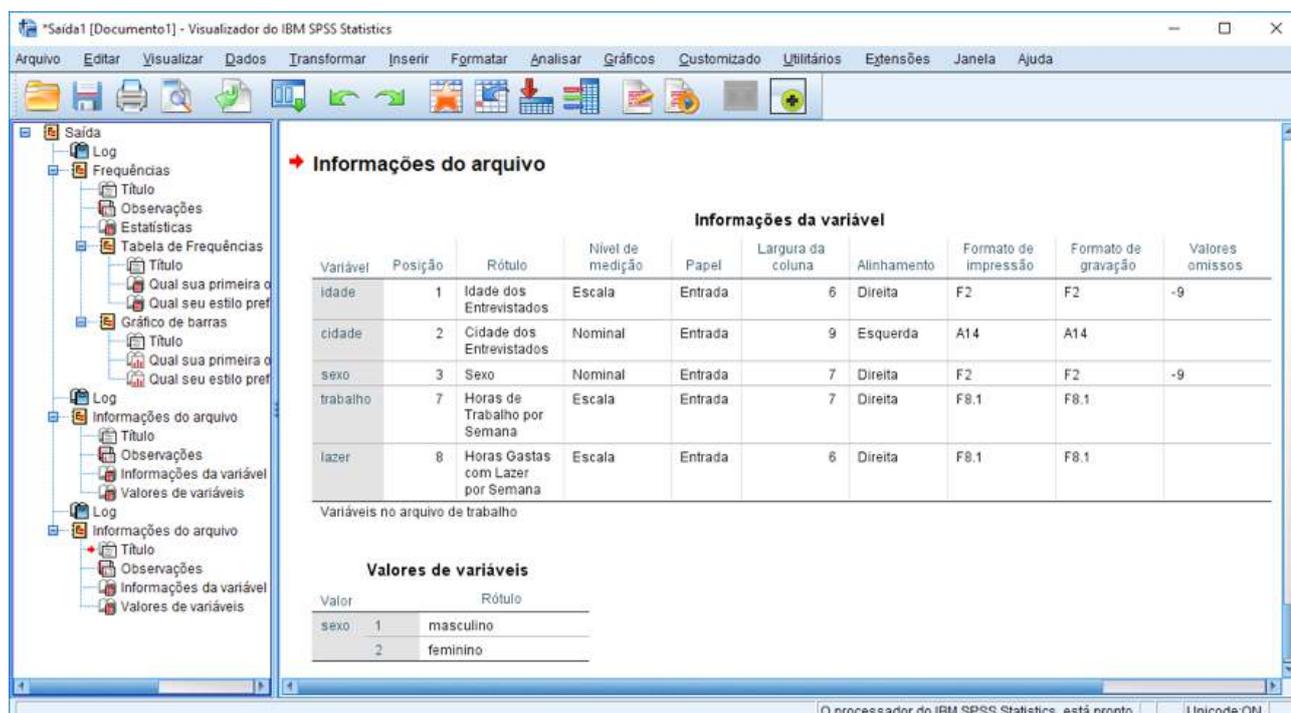
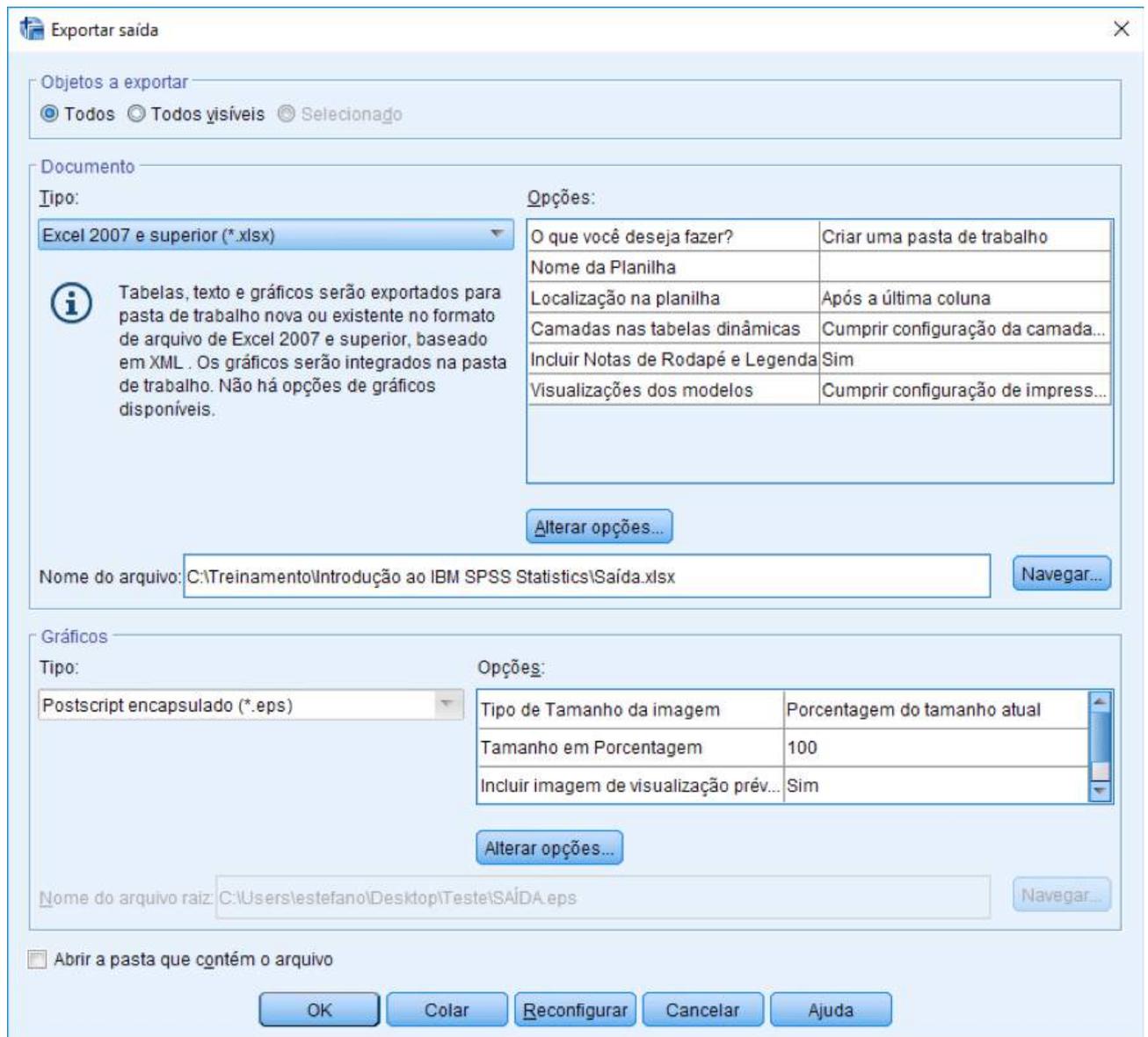


Figura 116 | Janela de exportação de uma saída para o formato Excel



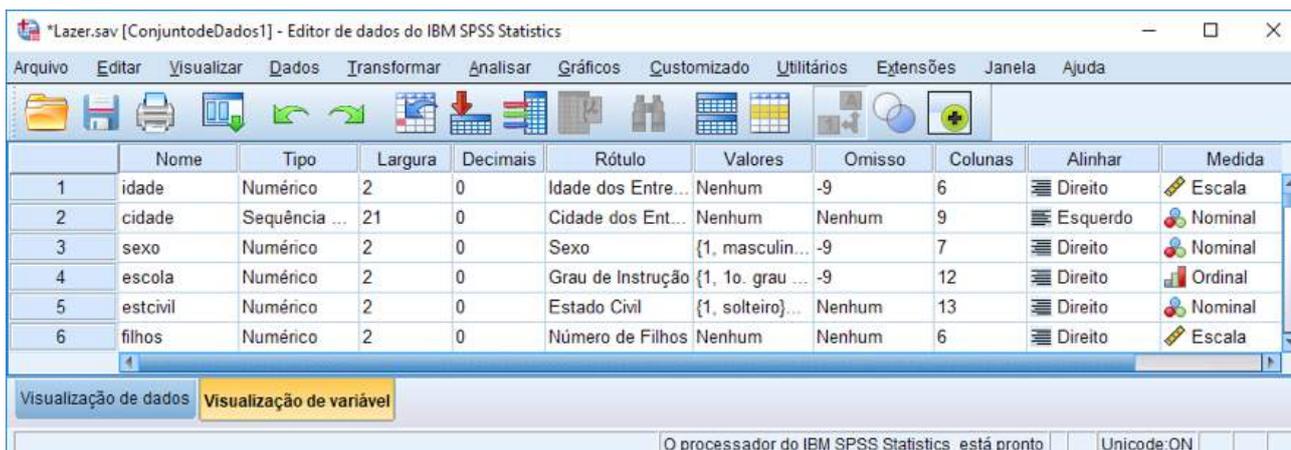
Um último comando que pode ser útil em transformação de variáveis é o ALTER TYPE, que altera o formato original de uma variável para um formato diferente, incluindo conversões de sequências alfanuméricas (*string*) para números ou datas. Por exemplo: Suponha que uma nova cidade seja incluída no arquivo *Lazer.sav*, digamos, São José do Rio Preto. Repare que este nome tem 21 caracteres (incluídos os espaços), mais do que os 14 caracteres do nome "Rio de Janeiro"; com isso, temos que atualizar o número máximo de caracteres (*width*) da variável alfanumérica *cidade*. Isso é feito na janela de programação a partir do comando

ALTER TYPE cidade (A21).

A sigla A21 remete ao formato alfanumérico ("A") com tamanho máximo (*width*) de 21 caracteres (ver Figura 117). Com isso, podemos incluir nomes maiores de cidades, como São José do Rio Preto. Consulte a ajuda do comando ALTER TYPE para mais informações e exemplos de conversão de variáveis.

**OBSERVAÇÃO:** Na codificação UNICODE (padrão do IBM SPSS Statistics, que pode ser alterado na aba Idioma do menu de **Opções**), todo símbolo acentuado é representado por dois caracteres. Portanto, você teria que digitar a sigla A23 para poder digitar "São José do Rio Preto" no Editor de Dados. Consulte a **Ajuda** do IBM SPSS Statistics para mais informações sobre codificação dos dados e dos arquivos gerados pelo *software*.

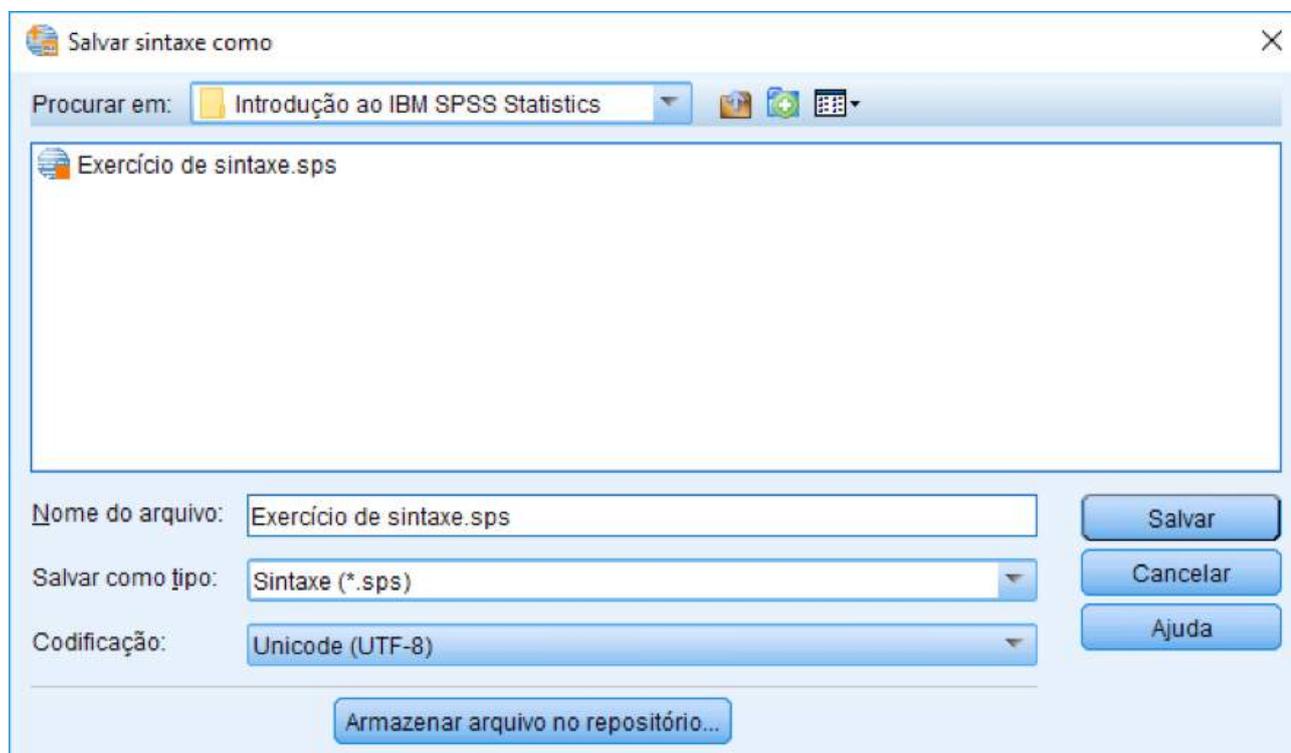
Figura 117 | Tamanho máximo da variável cidade (21) após a execução do ALTER TYPE



	Nome	Tipo	Largura	Decimais	Rótulo	Valores	Omisso	Colunas	Alinhar	Medida
1	idade	Numérico	2	0	Idade dos Entre...	Nenhum	-9	6	Direito	Escala
2	cidade	Sequência ...	21	0	Cidade dos Ent...	Nenhum	Nenhum	9	Esquerdo	Nominal
3	sexo	Numérico	2	0	Sexo	{1, masculin...	-9	7	Direito	Nominal
4	escola	Numérico	2	0	Grau de Instrução	{1, 1o. grau ...	-9	12	Direito	Ordinal
5	estcivil	Numérico	2	0	Estado Civil	{1, solteiro}...	Nenhum	13	Direito	Nominal
6	filhos	Numérico	2	0	Número de Filhos	Nenhum	Nenhum	6	Direito	Escala

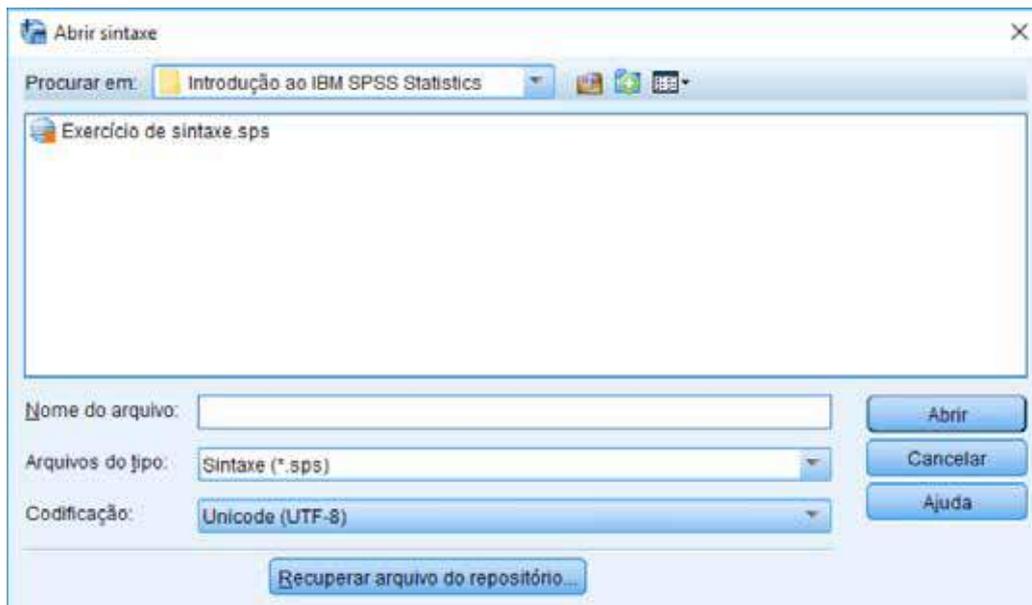
Cada janela de sintaxe aberta é um arquivo em potencial que pode ser salvo em formato texto. Para salvar uma janela de sintaxe, basta clicar no menu **Arquivo** e na opção **Salvar Como** (ver Figura 118). Escolhemos o disco e diretório no qual queremos armazenar o arquivo e digitamos o nome com o qual ele será armazenado (quadro **Nome do arquivo**). Para a janela **Sintaxe1**, podemos gerar um arquivo texto de nome Exercício de *sintaxe.sps*. Note que a extensão usada para os arquivos de sintaxe é *.sps*. Como esses arquivos são gerados como formato texto, podemos modificá-los em qualquer editor de textos (como, por exemplo, o Bloco de Notas do Windows).

Figura 118 | Salvando uma janela de sintaxe



Um arquivo de sintaxe salvo anteriormente pode ser aberto dentro de qualquer sessão do IBM SPSS Statistics, modificado, se necessário, e executado. Para abrir um arquivo desse tipo, basta clicar no menu **Arquivo**, na opção **Abrir** e escolher o tipo **Sintaxe (\*.sps)** no menu **Arquivos do tipo** (ver Figura 119). Após escolher o arquivo e clicar **OK**, uma janela de sintaxe é aberta e você poderá executar ou modificar esse arquivo.

Figura 119 | Janela para abertura de um arquivo com comandos (tipo sintaxe)



## 07 União de dois bancos de dados

### 07.1 Adicionando dados

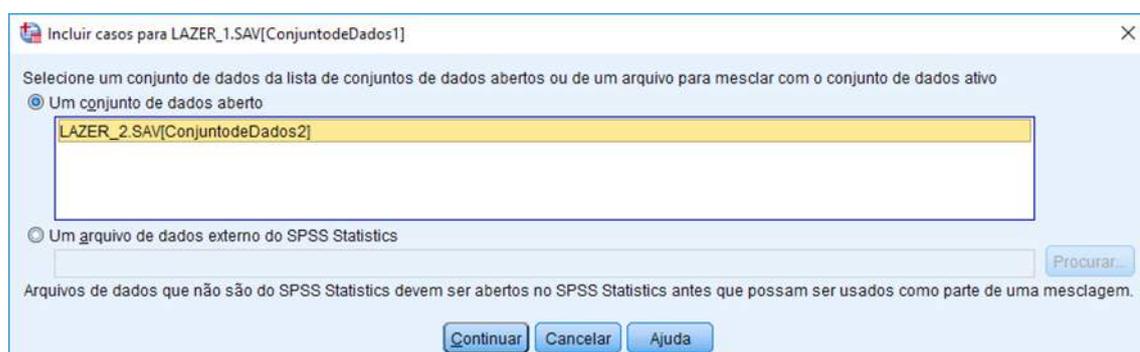
Essa opção é adequada quando os dois arquivos que desejamos unir tiverem as mesmas variáveis (ou pelo menos um grupo de variáveis em comum) e diferentes casos.

Por exemplo, suponha que a mesma pesquisa de opinião sobre lazer foi efetuada em São Paulo e no Rio de Janeiro. Vinte pessoas foram entrevistadas em São Paulo e onze pessoas no Rio de Janeiro. Ao final da pesquisa, dois arquivos foram gerados com as mesmas informações, um com os resultados de São Paulo (*LAZER\_1.sav*) e outro com os resultados do Rio de Janeiro (*LAZER\_2.sav*).

Após analisarmos os resultados regionalmente, vamos comparar os resultados das duas pesquisas. Para isso, vamos unir os dois arquivos. O arquivo final terá, portanto, 31 casos, sendo vinte de São Paulo e onze do Rio de Janeiro.

Em primeiro lugar devemos abrir um dos arquivos, por exemplo, *LAZER\_1.sav*. Em seguida, devemos clicar no menu **Dados** → **Mesclar Arquivos** → **Incluir Casos**. Na janela aberta (ver Figura 120) devemos indicar, na opção **Um arquivo de dados externo do SPSS Statistics**, o nome do arquivo a ser unido ao *LAZER\_1.sav* - que no nosso caso é *LAZER\_2.sav* - e clicar em **Continuar**. Caso o segundo arquivo já esteja aberto, escolha a opção **Um conjunto de dados aberto**.

Figura 120 | Especificando o arquivo a ser unido

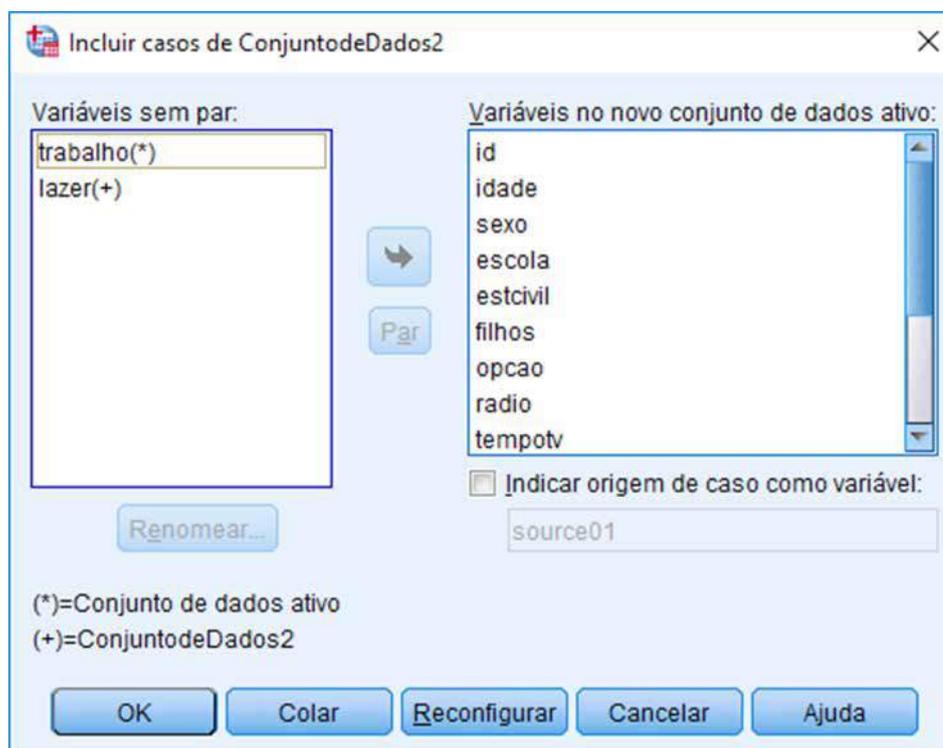


A nova janela apresentada contém uma série de informações sobre as variáveis presentes nos arquivos *LAZER\_1.sav* e *LAZER\_2.sav* (ver Figura 121). O quadro, no lado direito, de nome **Variáveis no novo conjunto de dados ativo** lista as variáveis que são comuns aos dois arquivos (variáveis pareadas). Como padrão, todas elas serão incluídas no “novo” arquivo que será criado após a união.

O quadro, no lado esquerdo, de nome **Variáveis sem par** contém todas as variáveis que não são comuns aos dois arquivos (variáveis não-pareadas). No nosso caso, duas variáveis aparecem nesse quadro: *trabalho(\*)* e *lazer(+)*. O sinal na frente da variável *trabalho* indica que ela está presente no arquivo ativo (*LAZER\_1.sav*), basta observar a legenda no canto inferior esquerdo: **(\*)=Conjunto de dados ativo**. O sinal na frente da variável *lazer* indica que essa variável está presente no arquivo *LAZER\_2.sav*. A princípio, todas as variáveis listadas nesse quadro não serão incluídas no novo arquivo.

Você tem a opção de incluir ou excluir quaisquer variáveis do novo arquivo. Para excluir variáveis comuns aos dois arquivos, basta movê-las para o quadro **Variáveis sem par**. Para incluir uma variável não-pareada, basta movê-la para o quadro **Variáveis no novo conjunto de dados ativo** usando o botão em formato de flecha. Se você quiser incluir uma variável que tenha nomes diferentes nos dois arquivos (por exemplo: *es\_civil* e *estcivil*), selecione as duas e use o botão **Par**; assim, essas duas variáveis irão passar para o quadro **Variáveis no novo conjunto de dados ativo** como uma só.

Figura 121 | Janela de definição para união de casos



Incluiremos as variáveis *trabalho* e *lazer*. No canto inferior direito, podemos marcar o quadro **Indicar origem de caso como variável** quando quisermos incluir uma variável que receberá valor 0 quando o caso pertence ao arquivo ativo e valor 1 quando o caso pertence ao arquivo que está sendo unido. Como padrão, o nome dessa nova variável é *source01*; para mudá-lo, basta clicar com o cursor no quadro que contém o nome e digitar o novo nome. Podemos usar o nome *cidade* para essa variável, pois a mesma terá valor 0 para casos provenientes de São Paulo e valor 1 para casos provenientes do Rio de Janeiro.

A Figura 122 apresenta a janela de definição de união de casos já com as mudanças mencionadas acima. Clicamos **OK** e a janela de dados é preenchida com o arquivo resultante da união dos arquivos *LAZER\_1.sav* e *LAZER\_2.sav* (ver Figura 123).

Notamos que a variável *cidade* esta posicionada na última coluna e tem valor 0 até a linha 20 e valor 1 a partir da linha 21. A variável *trabalho* tem apenas valores ausentes a partir da linha 21 e a variável *lazer* tem valores ausentes até a linha 20.

É importante ressaltar que, quando queremos unir mais do que dois arquivos, o procedimento é idêntico ao já descrito aqui. Unimos os dois primeiros arquivos, e, depois, ao arquivo resultante, unimos o terceiro e assim por diante. Casos de mais de dois arquivos também podem ser unidos através da janela de sintaxe com o comando **ADD FILES**, de forma que os arquivos são unidos na ordem em que são declarados na sintaxe.

Figura 122 | Janela de definição para união de casos após as modificações

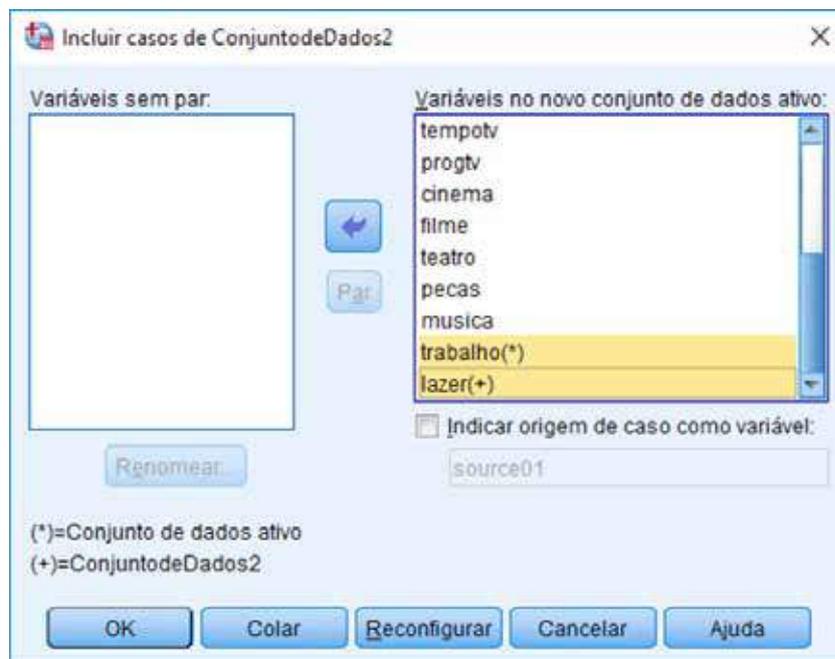


Figura 123 | Janela de dados com o arquivo resultante da união de casos

The screenshot shows the IBM SPSS Statistics data editor window for the file 'LAZER\_1.SAV [ConjuntodeDados1]'. The window title bar indicates it is the 'Editor de dados do IBM SPSS Statistics'. The menu bar includes 'Arquivo', 'Editar', 'Visualizar', 'Dados', 'Transformar', 'Analisar', 'Gráficos', 'Customizado', 'Utilitários', 'Estadísticos', 'Janela', and 'Ajuda'. The toolbar contains various icons for file operations and data manipulation. The main area displays a data grid with 17 columns and 32 rows. The columns are: 'filhas', 'trabalho', 'opcao', 'radio', 'tempotv', 'progrtv', 'cinema', 'filme', 'teatro', 'pecas', 'musica', and 'lazer'. The data shows a mix of values across the rows, with some missing values (indicated by empty cells). The status bar at the bottom indicates 'O processador do IBM SPSS Statistics está pronto' and 'Unicode: UTF-8'.

	filhas	trabalho	opcao	radio	tempotv	progrtv	cinema	filme	teatro	pecas	musica	lazer
10	0	40,0		.50	4,00	filmes	0	teatro	0	rock		
11	0	40,0	esportes	3,00	4,00	programas	10	suspense	2	comedia	rock	
12	0	35,0	cinema	.00	3,00	filmes	10	suspense	0	rock		
13	0	30,0	esportes	4,00	3,00	novelas	0	comedia	0	comedia	rock	
14	0	24,0	teatro	2,00	.00	telejornais	10	comedia	10	comedia	rock	
15	0	20,0	outros	3,00	4,00	novelas	2	aventura	1	comedia	rock	
16	0		tv/video	2,00	4,00	filmes	2	comedia	0	comedia	rock	
17	6	70,0	outros	.00	1,00	telejornais	0	comedia	0	comedia	mpb	
18	0	40,0	outros	.30	.00	telejornais	0	aventura	0		gospel	
19	3	40,0	tv/video	.00	1,00	filmes	2	comedia	1	comedia classica		
20	0	40,0	cinema	.00	.50	filmes	1	aventura	3	musical	classica	
21	0		outros	.50	2,00	filmes	4	suspense	3	comedia	outros	20,0
22	0		cinema	1,00	3,00	novelas	12	suspense	3	comedia	mpb	21,0
23	0		tv/video	.50	1,50	filmes	12	outros	4	comedia	gospel	8,0
24	0		outros	.50	2,00	programas	4	drama	1	comedia	mpb	16,0
25	6		tv/video	.00	2,00	telejornais	0	comedia	0	comedia	mpb	12,0
26	0		tv/video	4,00	3,00	filmes	1	aventura	0	comedia	outros	20,0
27	0		esportes	1,50	1,50	programas	15	suspense	4	musical	mpb	18,0
28	3		outros	.00	.50	outros	2	comedia	1	tragedia	classica	25,0
29	4		tv/video	.50	3,00	filmes	1	suspense	0	tragedia	mpb	13,0
30	0		cinema	4,00	4,00	programas	10	comedia	1	comedia	rock	30,0
31	3		tv/video	3,00	3,00	novelas	1	comedia		comedia	samba	

## 07.2 Adicionando variáveis

Esta opção é adequada quando os dois arquivos contêm os mesmos casos (ou, pelo menos, parte deles em comum), mas não as mesmas variáveis. O arquivo resultante terá as variáveis provenientes dos dois arquivos. Nesse tipo de união, é necessário que os arquivos a serem unidos estejam no formato IBM SPSS Statistics e é aconselhável que contenham, no mínimo, uma variável-chave.

A variável-chave (**key variable**) é aquela que identifica a correspondência dos casos nos dois arquivos. Essa variável deve ter o mesmo valor nos dois arquivos para casos equivalentes e o arquivo deve estar ordenado por essa variável antes que a união seja efetuada. Você pode adicionar variáveis sem usar uma variável-chave, mas, para que o resultado da união seja correto, os dois arquivos terão que conter os mesmos casos e exatamente na mesma ordem.

No nosso exemplo, vamos unir os arquivos *LAZER\_A.sav* e *LAZER\_B.sav*; alguns casos desses arquivos são apresentados nas Figura 124 e Figura 125, respectivamente. Os dois arquivos têm 30 casos (não necessariamente os mesmos) e uma variável em comum, *id*. É esta variável que será usada como variável-chave na união. Além disso, os dois arquivos já estão ordenados segundo a variável *id*. As demais variáveis estão apenas em um dos dois arquivos: *idade*, *sexo*, *escola*, *estcivil* e *filhos* no arquivo *LAZER\_A.sav*; *trabalho*, *lazer*, *opção*, *radio*, *tempotv*, *progtv*, *cinema*, *filme* e *teatro* no arquivo *LAZER\_B.sav*.

Figura 124 | Listagem de alguns casos do arquivo *LAZER\_A.sav*

	id	trabalho	lazer	opções	radio	tempotv	progtv	cinema	filme	teatro	peças	musica
1	1	40.0	10.0		2.00	1.00	outros	3	suspense	1	comedia	rock
2	2	70.0	26.0	shows/con	1.00	2.00	filmes	10	aventura	6	comedia	rock
3	3	40.0	2.0	esportes	2.00	1.00	telejornais	0	aventura	2	comedia	mpb
4	4	48.0	16.0	teatro	4.00	2.00	programas	2	comedia	8	comedia	rock
5	5	40.0			2.00	2.00		0		0		
6	6	40.0	10.0	tvvideo	2.00	5.00	filmes	1	ficcao cien	0		samba
7	7	40.0	20.0	cinema	2.00	3.00	outros	5	aventura			outros
8	9	40.0	16.0	tvvideo	2.00	4.00	telejornais	0	aventura	0	comedia	mpb
9	10	50.0			5.00	4.00	filmes	0	terror	0		rock
10	11	50.0	17.0	esportes	3.00	4.00	programas	10	suspense	2	comedia	rock
11	12	40.0	15.0	cinema	0.00	3.00	filmes	10	suspense	0		rock
12	13	0	40.0	esportes	4.00	3.00	novelas	0	comedia	0	comedia	rock
13	14	30.0		teatro	2.00	.00	telejornais	10	comedia	10	comedia	rock

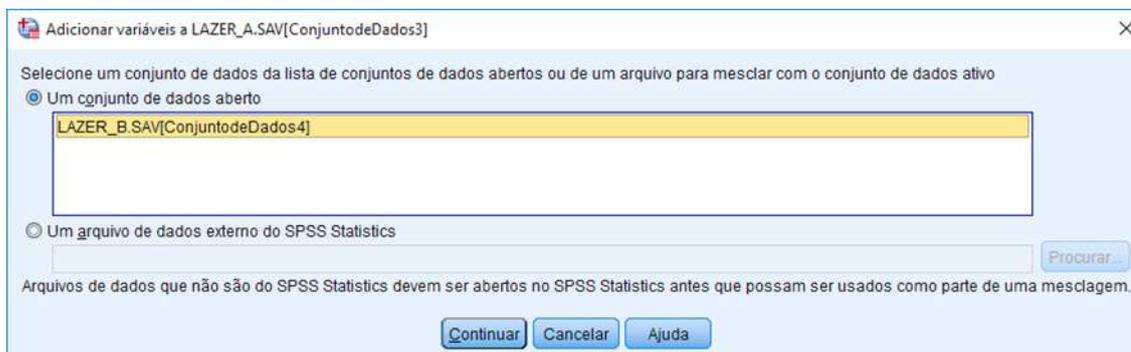
Figura 125 | Listagem de alguns casos do arquivo *LAZER\_B.sav*

	id	trabalho	lazer	opções	radio	tempotv	progtv	cinema	filme	teatro	peças	musica
1	1	40.0	10.0		2.00	1.00	outros	3	suspense	1	comedia	rock
2	2	70.0	26.0	shows/con	1.00	2.00	filmes	10	aventura	6	comedia	rock
3	3	40.0	2.0	esportes	2.00	1.00	telejornais	0	aventura	2	comedia	mpb
4	4	48.0	16.0	teatro	4.00	2.00	programas	2	comedia	8	comedia	rock
5	5	40.0			2.00	2.00		0		0		
6	6	40.0	10.0	tvvideo	2.00	5.00	filmes	1	ficcao cien	0		samba
7	7	40.0	20.0	cinema	2.00	3.00	outros	5	aventura			outros
8	9	40.0	16.0	tvvideo	2.00	4.00	telejornais	0	aventura	0	comedia	mpb
9	10	50.0			5.00	4.00	filmes	0	terror	0		rock
10	11	50.0	17.0	esportes	3.00	4.00	programas	10	suspense	2	comedia	rock
11	12	40.0	15.0	cinema	0.00	3.00	filmes	10	suspense	0		rock
12	13	0	40.0	esportes	4.00	3.00	novelas	0	comedia	0	comedia	rock
13	14	30.0		teatro	2.00	.00	telejornais	10	comedia	10	comedia	rock

O arquivo final, após a união, deverá ter 15 variáveis, sendo cinco provenientes do arquivo *LAZER\_A.sav* e nove provenientes do arquivo *LAZER\_B.sav*, além da própria variável *id*.

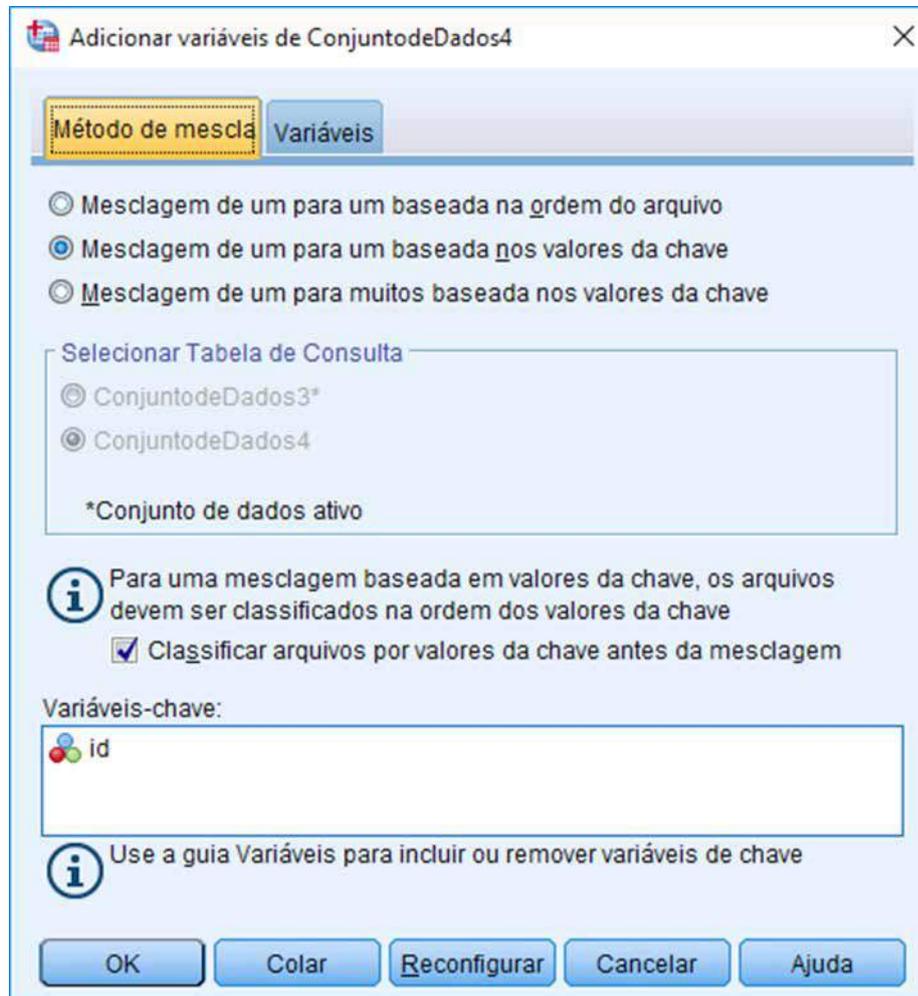
Da mesma forma que na união de casos, devemos abrir um dos dois arquivos, por exemplo, *LAZER\_A.sav*. Em seguida, devemos clicar o menu Dados... Mesclar Arquivos... Incluir Variáveis. Uma janela é aberta, e nela devemos indicar o nome do arquivo a ser unido a *LAZER\_A.sav*, que no nosso caso é *LAZER\_B.sav*, e clicar em Continuar (ver Figura 126).

Figura 126 | Janela de definição para união de variáveis



A próxima janela aberta (ver Figura 127) possui duas abas, onde a primeira aba define o método de união dos bancos de dados (**Método de Mescla**) e a segunda define quais variáveis serão incluídas ou excluídas no arquivo completo, além das variáveis-chave que serão utilizadas na união (**Variáveis**).

Figura 127 | Janela de definição para união de variáveis – aba Método de Mescla



A aba **Método de Mescla** possui três opções distintas de união de variáveis em um único banco de dados:

- Mesclagem de um para um baseada na ordem do arquivo: Os bancos de dados serão unidos de acordo com a ordem das linhas (casos) dos respectivos arquivos, ou seja, o primeiro caso do primeiro arquivo será unido com o primeiro caso do segundo arquivo, e assim por diante.
- Mesclagem de um para um baseada nos valores da chave: Os casos dos dois bancos de dados serão unidos de acordo com os valores da(s) variável(is)-chave definidas na aba **Variáveis**. Esta opção é equivalente à opção **Ambos os arquivos fornecem dados** do procedimento em versões do IBM SPSS Statistics anteriores à versão 25.
- Mesclagem de um para muitos baseada nos valores da chave: A união é baseada nos valores da(s) variável(is)-chave, mas o arquivo que é definido como **Tabela de Consulta (Query Table)** contribui apenas com os valores das variáveis para o novo arquivo, ou seja, os casos da tabela de consulta que não estão na outra tabela são ignorados na união de bancos de dados. Esta opção é equivalente às opções **O conjunto de dados não ativo é uma tabela vinculada** e **O conjunto de dados ativo é tabela vinculada** do procedimento em versões do IBM SPSS Statistics anteriores à versão 25.

Vamos manter a opção padrão selecionada (**Mesclagem de um para um baseada nos valores da chave**), porque queremos que todos os valores de *id* (variável-chave) apareçam no arquivo resultante da união (que está na Figura VII-10). De forma automática, o procedimento define que as variáveis que estão nos dois arquivos serão as variáveis-chave; no nosso exemplo, a única variável-chave é *id*, como mostra o quadro na aba **Método de Mescla**, mas você pode escolher quais são as variáveis-chave na aba **Variáveis**.

A união de bancos de dados falhará se os arquivos não estiverem ordenados pela(s) variável(is)-chave. Por este motivo, e por padrão, a caixa de verificação **Classificar arquivos por valores da chave antes da mesclagem** está selecionada, o que fará com que o IBM SPSS Statistics ordene os dois bancos de dados antes de fazer a união. Como os nossos dois arquivos já estão ordenados pelos valores de *id*, não seria necessário selecionar essa opção, mas vamos mantê-la por enquanto. Agora, clique na aba **Variáveis** (ver Figura 128).

Figura 128 | Janela de definição para união de variáveis – aba Variáveis



Dentro desta aba, você pode escolher quais variáveis serão incluídas ou excluídas na união de bancos de dados. Assim como na união por casos, as variáveis marcadas com um asterisco (\*) pertencem ao conjunto de dados ativo (no nosso exemplo, o arquivo *LAZER\_A.sav*) e as variáveis marcadas com um sinal de soma (+) pertencem ao conjunto de dados não-ativo (o arquivo *LAZER\_B.sav*). Como foi dito anteriormente, você pode definir o conjunto de variáveis-chave a partir desta aba.

Clique em **OK** para fazer a união dos arquivos. O resultado está na Figura 129.

Figura 129 | Janela de dados com o arquivo resultante da primeira união

id	trabalho	lazer	opcao	radio	tempo	prog	cinema	filme	teatro	novelas	musica		
1	0	40,0	18,0	2,00	1,00	outros	3	suspense	1	comedia	rock		
2	0	70,0	26,0	shows/con	1,00	2,50	filmes	10	aventura	6	comedia	rock	
3	3	40,0	2,0	esportes	2,00	1,00	telejornais	0	aventura	2	comedia	mp3	
4	ado	48,0	16,0	teatro	4,00	2,50	programas	2	comedia	8	comedia	rock	
5	ado	3	40,0		2,00	2,00		0					
6	ado	3	40,0	10,0	tv/video	2,00	5,00	filmes	1	ficcao can		sertao	
7	0	40,0	20,0	cinema	2,00	3,00	outros	5	aventura	0		outros	
8	0												
9	ado	0	40,0	16,0	tv/video	2,00	4,00	telejornais	0	aventura	0	comedia	mp3
10	ado	0	50,0		.50	4,00	filmes	0	terror	0		rock	
11	ado	0	50,0	17,0	esportes	3,00	4,00	programas	10	suspense	2	comedia	rock
12	ado	0	40,0	15,0	cinema	.00	3,00	filmes	10	suspense	0	rock	
13	0	0	40,0	esportes	4,00	3,00	novelas	0	comedia	0	comedia	rock	

Como exemplo final, mostraremos o resultado final da união dos mesmos bancos de dados, caso a opção **Mesclagem de um para muitos baseada nos valores da chave** tivesse sido selecionada (ver Figura 130). Neste caso, nós manteríamos a seleção padrão do ConjuntoDeDados4 (*LAZER\_B.sav*) como tabela de consulta, o que faz com que apenas os casos do conjunto de dados **ativo** (*LAZER\_A.sav*) estejam disponíveis após a união dos dois bancos. Neste exemplo, o caso com *id* = 17, que aparece na tabela de consulta, não está no arquivo unido porque ele não aparece no conjunto de dados ativo (ver Figura 131). De fato, a tabela de consulta somente contribui com os valores das variáveis para os casos que estão presentes no conjunto de dados ativo (identificados pelos valores de *id*).

Figura 130 | Janela de união de variáveis

Adicionar variáveis de ConjuntoDeDados4

Método de mescla Variáveis

Mesclagem de um para um baseada na ordem do arquivo  
 Mesclagem de um para um baseada nos valores da chave  
 Mesclagem de um para muitos baseada nos valores da chave

Selecionar Tabela de Consulta

ConjuntoDeDados3\*  
 ConjuntoDeDados4

\*Conjunto de dados ativo

Para uma mesclagem baseada em valores da chave, os arquivos devem ser classificados na ordem dos valores da chave  
 Classificar arquivos por valores da chave antes da mesclagem

Variáveis-chave:

id

Use a guia Variáveis para incluir ou remover variáveis de chave

OK Cotar Reconfigurar Cancelar Ajuda

Figura 131 | Janela de dados com o arquivo resultante da segunda união

	id	idade	sexo	escola	est civil	sítes	trabalho	lazer	opcao	radio
10	10	36	masculino	superior completo	separado/divorciado	0	50,0			50
11	11	25	masculino	superior completo	separado/divorciado	0	50,0	17,0	esportes	3,00
12	12	24	masculino	superior incompleto	separado/divorciado	0	40,0	15,0	cinema	0,00
13	13	14	masculino	1a. grau incompleto	solteiro	0	,0	40,0	esportes	4,00
14	14	22	masculino	superior incompleto	solteiro	0	30,0		teatro	2,00
15	15	20	masculino	superior incompleto	separado/divorciado	0	20,0	18,0	subtos	3,00
16	16	19	masculino	superior incompleto	solteiro	0	40,0	22,0	tv/video	2,00
17	18	29	masculino	superior completo	viuvo	0	40,0	3,0	subtos	3,00
18	19	21	masculino	superior completo	viuvo	3	50,0	5,0	tv/video	0,00
19	20	31	feminino	superior completo	solteiro	0	40,0	10,0	cinema	0,00
20	21	20	feminino	superior completo	solteiro	0	40,0	20,0	subtos	1,00
21	22	20	feminino	superior completo	solteiro	0	40,0	21,0	cinema	1,00
22	23	37	feminino	superior incompleto	casado	0	45,0	8,0	tv/video	5,00

O comando para unir bancos de dados com variáveis diferentes via programação é o **MATCH FILES**. A maior vantagem de utilizar este comando e o **ADD FILES** é que, em vez de unir apenas dois bancos pela janela do **Mesclar Arquivos**, é possível unir até 50 arquivos de uma vez, de modo que cada conjunto de dados que traz apenas as variáveis deve ser indicado em um novo subcomando /TABLE. Observe que poderíamos acrescentar um novo arquivo à união a partir do diretório no qual o mesmo se localiza, sem a necessidade de abrir um novo conjunto de dados no IBM SPSS Statistics (vide Figura VII-13, com o arquivo *LAZER\_C.sav*). Esta sintaxe pressupõe que os dados dos três arquivos já estejam ordenados pelos valores de id.

Vale lembrar que no nosso exemplo:

- ConjuntoDeDados3 é o nome do conjunto de dados referente ao arquivo *LAZER\_A.sav* (conjunto de dados ativo; marcado com um \*; já aberto no IBM SPSS Statistics);
- ConjuntoDeDados4 é o nome do conjunto de dados referente ao arquivo *LAZER\_B.sav* (conjunto de dados não-ativo; usado como tabela de consulta; já aberto no IBM SPSS Statistics) e que deve ser declarado entre aspas simples;
- *LAZER\_C.sav* (usado como tabela de consulta) é incluído na união a partir de seu diretório original, sem estar aberto no IBM SPSS Statistics, e tal diretório deve ser declarado entre aspas simples.

Lembre-se de que você pode renomear os conjuntos de dados abertos no IBM SPSS Statistics, mas quaisquer alterações que sejam feitas também precisarão ser feitas na sintaxe. Consulte a **Ajuda** dos comandos **ADD FILES** e **MATCH FILES** para mais informações sobre união de bancos de dados.

Figura 132 | Janela de sintaxe para a união de variáveis com três bancos de dados

```

1
2 DATASET ACTIVATE ConjuntoDeDados3
3 MATCH FILES /FILE=*
4 /TABLE='ConjuntoDeDados4'
5 /TABLE='C:\Treinamento\Introdução ao IBM SPSS Statistics\LAZER_C.sav'
6 /BY id
7 EXECUTE.
8

```

## 08 GRÁFICOS

### 08.1 Construção de gráficos através do Construtor de Gráfico

A interface **Construtor de Gráfico** (*Chart Builder*) permite que você construa o seu gráfico através de uma galeria de gráficos predefinidos ou de elementos individuais que compõem o gráfico (tais como eixos, pontos, barras, linhas etc.) e visualize o gráfico enquanto ele é construído. Existe ainda a opção de inserir elementos opcionais como títulos, subtítulos e rodapés.

Neste capítulo, criaremos um gráfico de barras para indicar a distribuição da variável *estcivil*, com as barras indicando cores diferentes para cada categoria da variável *sexo*. Utilizaremos para isso o arquivo *Lazer.sav*.

Para utilizar o **Construtor de Gráfico**, basta acessar o menu **Gráficos > Construtor de Gráfico**. Na janela aberta (ver Figura 133), vamos escolher o tipo de gráfico clicando na aba **Galeria** e, em seguida, escolhemos a opção **Barra** em **Escolha entre**.

Nesta janela, vários tipos de gráficos de barras estão disponíveis (simples, agrupado, empilhado e 3-D), para saber o nome de cada um basta passar o mouse em cima do gráfico.

Nós utilizaremos o gráfico do tipo **Barra em Cluster** (ou **barras agrupadas**): para isso, arraste o ícone correspondente ao gráfico tipo *cluster* (segundo ícone) para a tela acima ou clique duas vezes em cima do ícone do gráfico tipo *cluster* para começar a construção do gráfico (ver Figura 134).

Figura 133 | Janela para criar um gráfico através do Construtor de Gráfico



Uma prévia do gráfico de barras aparecerá na tela (ver Figura 134). A aba **Propriedade do Elemento** permite que você edite elementos, categorias, estatísticas e estilo das barras no seu gráfico. As barras mostradas na tela não representam os seus dados; elas são apenas para exemplificação.

Figura 134 | Gráfico de barras construído através do Construtor de Gráfico



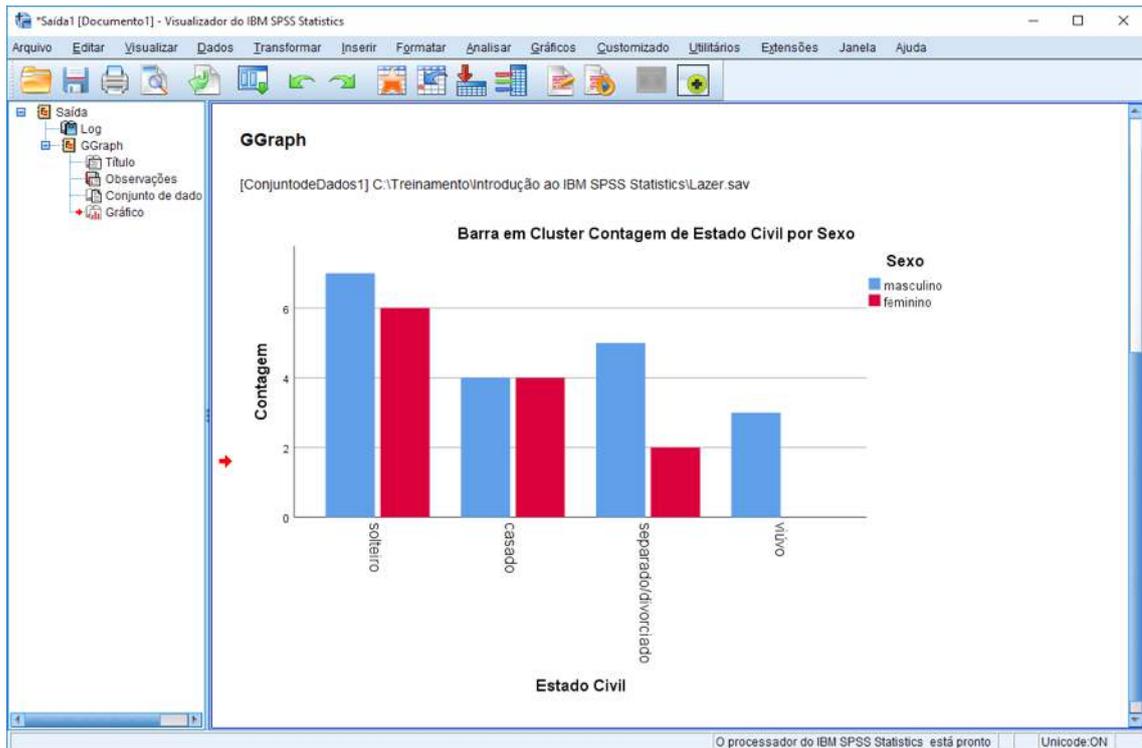
Agora nós colocaremos as variáveis que desejamos mostrar no gráfico de barras nas categorias correspondentes. No eixo X, colocamos a variável que desejamos definir no gráfico. Se nenhuma variável for especificada no eixo Y, o mesmo representará a altura de cada barra em frequência ou percentual. Caso você queira mostrar uma estatística no eixo Y como, por exemplo, a média de idade, será necessário colocar essa variável no eixo Y. Alguns tipos de gráficos como o diagrama de dispersão, por exemplo, necessitam de uma variável no eixo Y. Vamos colocar a variável *estcivil* no eixo X (área **Eixo X?**) e a variável *sexo* em **Cluster em X: Definir Cor** (ver Figura 134).

O padrão para o eixo Y é mostrar a contagem ou frequência para cada categoria de *estcivil*. Podemos facilmente alterar a estatística para percentual na janela de propriedades (**Propriedades do Elemento**). Por exemplo, caso fosse necessário criar um gráfico de barras representando o percentual de casos nas categorias de estado *civil* devemos escolher **Barra1** na seção **Editar propriedades de:** da janela de propriedades, clicar em **Porcentagem(?)** na lista de estatísticas e na opção **Configurar parâmetros**, devemos selecionar **Total para cada categoria do eixo X** na área **Denominador para calcular a porcentagem**.

Na aba **Elementos Básicos**, escolhemos o tipo de coordenada e um elemento (barra, linha, *boxplot*, etc.) para compor um gráfico. Na aba **Títulos/Notas de Rodapé**, podemos determinar títulos, subtítulo e rodapés para o nosso gráfico. A aba **Aparência do Gráfico** permite que você edite elementos do gráfico, como as cores padrão das barras e a inclusão ou exclusão de linhas de grade. Na aba **Opções** (localizada no canto superior direito da caixa de diálogo) controlamos algumas opções gerais nos gráficos, como: o tratamento dado aos valores ausentes (*missing*), escolha de um modelo (*template*) para o gráfico, determinar o tamanho do gráfico em percentual e até mesmo criar um gráfico dividido em vários painéis.

Clique em **OK** na janela que definimos o gráfico. A saída está na Figura 135.

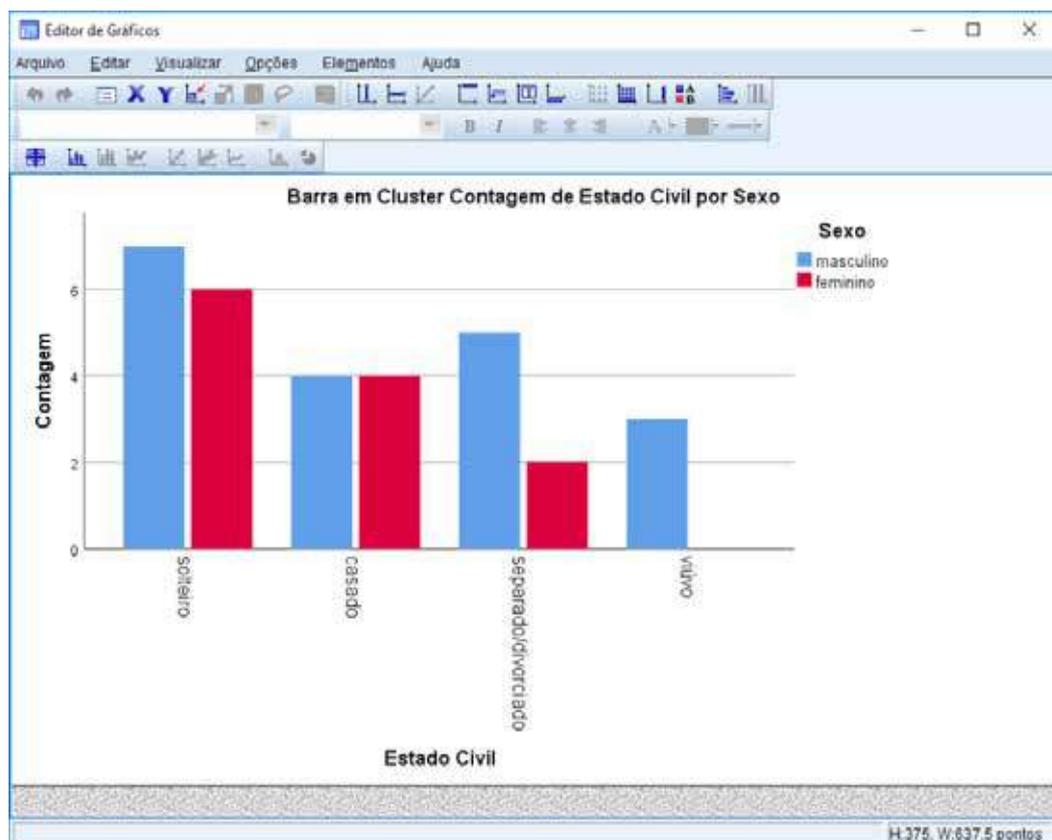
Figura 135 | Gráfico de barras criado através do Chart Builder



## 08.2 Editando o gráfico

Para editar um gráfico, devemos clicar duas vezes sobre o gráfico que está na janela de saída. Automaticamente, uma janela de edição de gráficos ficará aberta: trata-se do **Editor de Gráficos** (ver Figura 136).

Figura 136 | Janela de edição do gráfico (Editor de Gráficos)



No menu **Editar**, clique na opção **Propriedades** para acessar as opções de edição disponíveis para o nosso gráfico. Na janela de propriedades do gráfico, a primeira opção disponível para edição é a aba **Categorias** (ver Figura 137): nela, controlamos as categorias que desejamos incluir/excluir do nosso gráfico para cada variável que foi utilizada e, caso seja necessário, conseguimos alterar a ordem como as categorias estão aparecendo na tela.

Na aba **Opções da Barra** (ver Figura 138), alteramos a largura das barras e dos grupos, e, no caso dos gráficos do tipo **Boxplot**, determinamos o estilo das caixas.

Em **Espessura e Ângulo** (ver Figura 139), escolhemos o efeito que desejamos às nossas barras. Podemos deixar as barras com sombreado ou colocá-las em 3D.

Em **Variáveis** (ver Figura VIII-8), podemos alterar toda a estrutura do gráfico, pois é possível alterar o tipo de gráfico, trocar as variáveis de eixo, etc.

Figura 137 | Opção Categorias – Propriedades do Gráfico

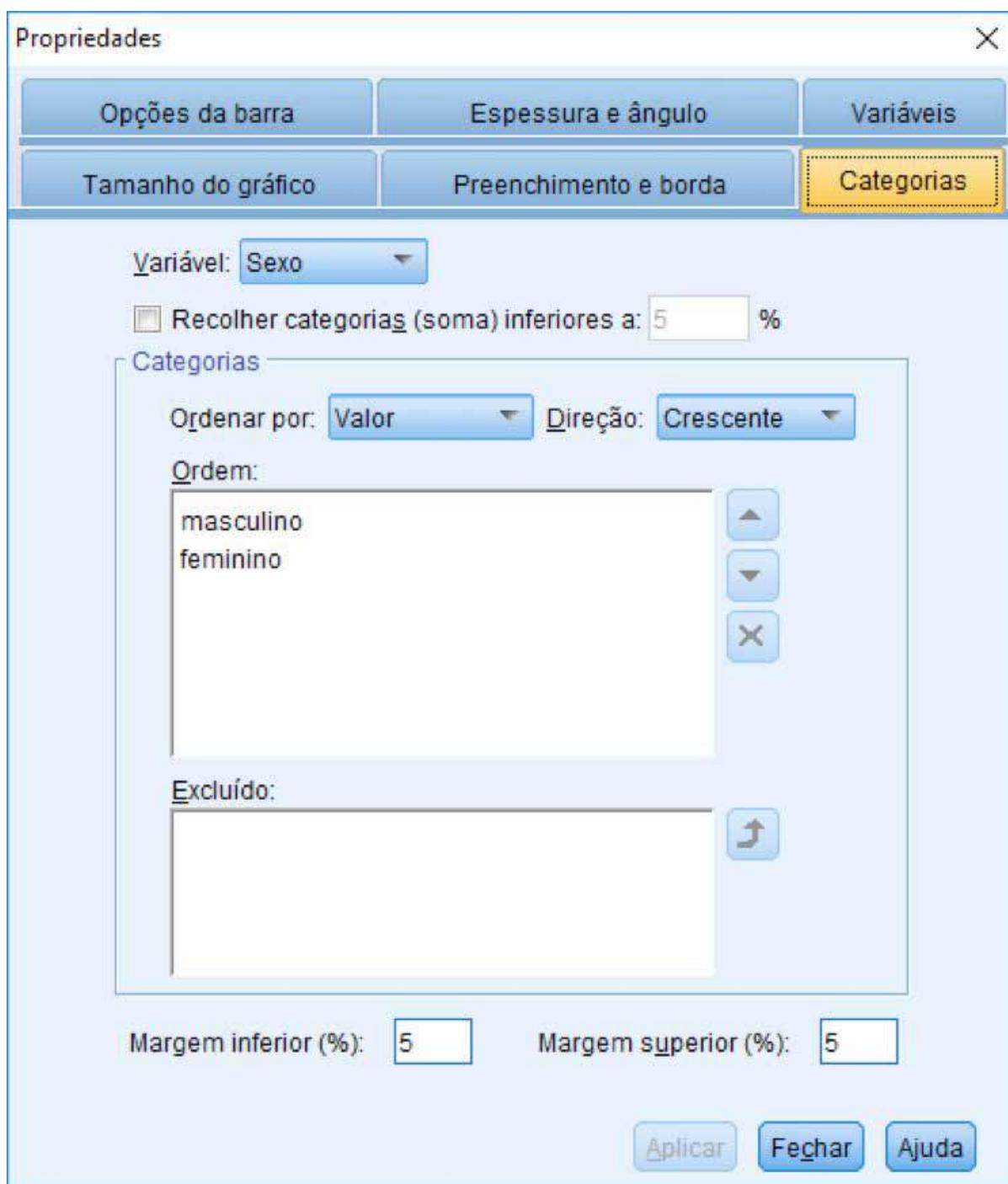


Figura 138 | Opção Opções da Barra – Propriedades do Gráfico

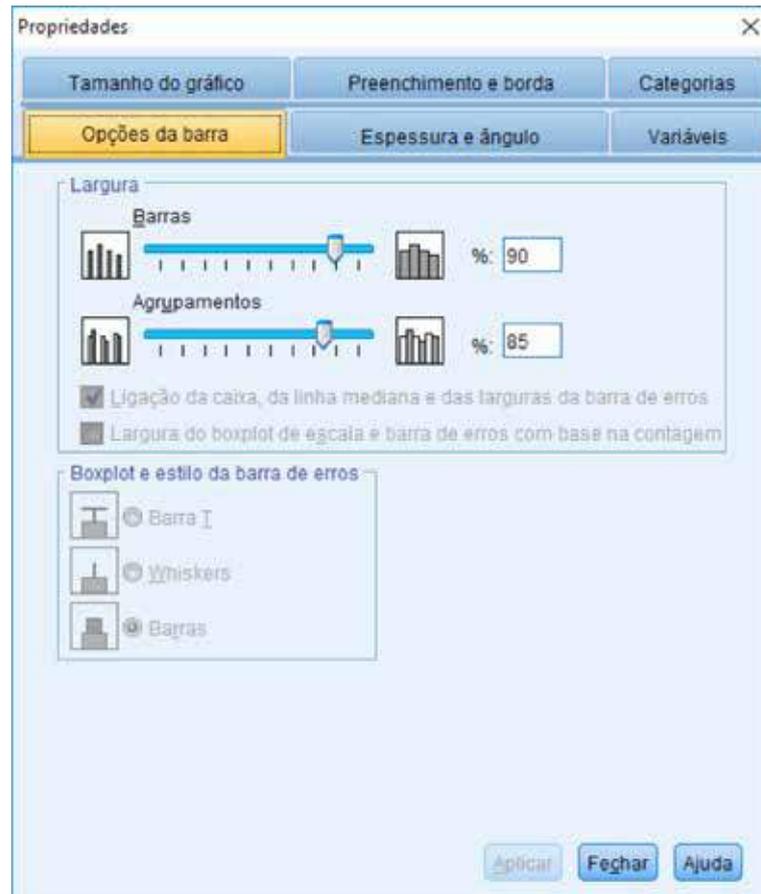


Figura 139 | Opção Espessura e Ângulo – Propriedades do Gráfico

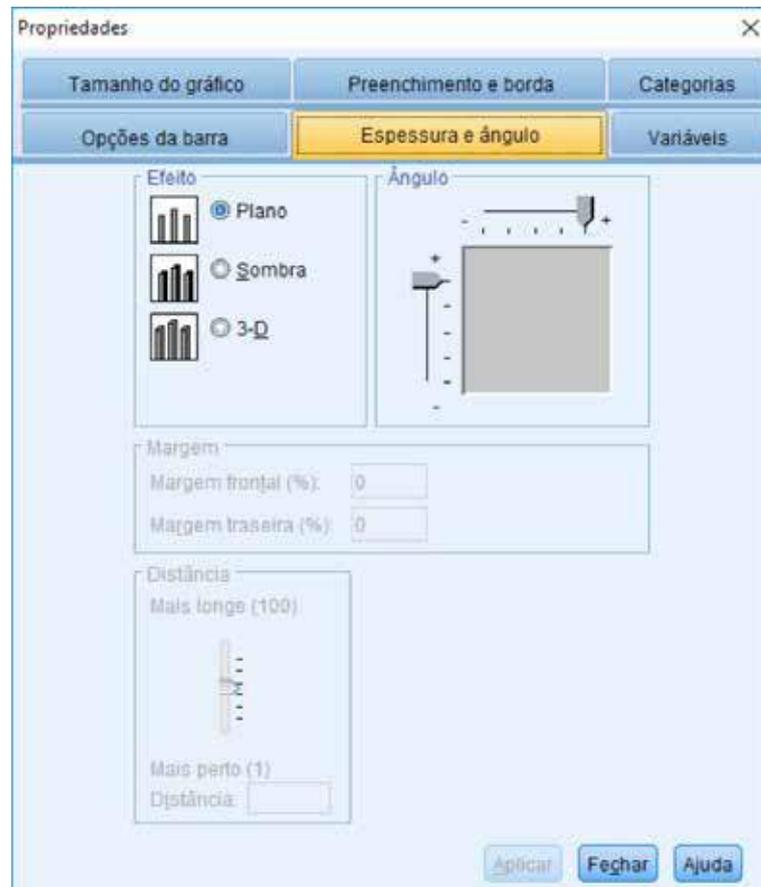
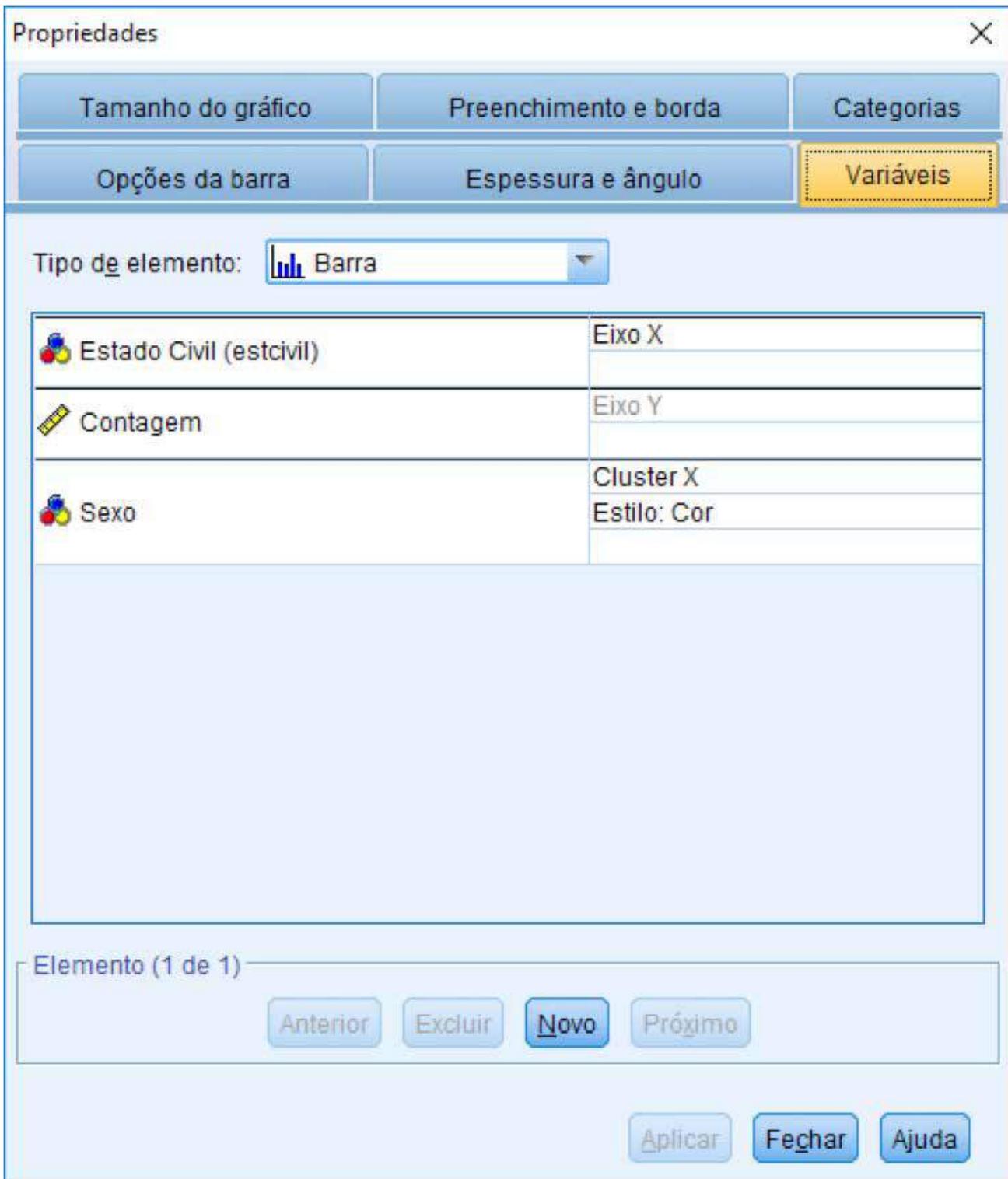


Figura 140 | Opção Variáveis – Propriedades do Gráfico



Em **Tamanho do Gráfico**, alteramos o tamanho do gráfico, mantendo ou não a proporção. E em **Preenchimento e Borda** (ver Figura 141), alteramos as cores das barras que estiverem selecionadas e também o tipo de borda de cada barra.

Outras opções de edição do gráfico estão disponíveis nos menus **Editar**, **Opções** e **Elementos**, tais como inserir um título, inserir texto, mostrar rótulo dos valores, esconder legenda, omitir a legenda, etc. Estas opções não estão descritas nesta apostila, mas você pode encontrar detalhes sobre cada uma delas na **Ajuda** do IBM SPSS Statistics.

Figura 141 | Opção Preenchimento e Borda – Propriedades do Gráfico



### 08.3 Mapas

A interface **Seletor de Modelo de Tabelas de Gráficos** (*Graphboard Template Chooser*) permite visualizar os dados geograficamente, criando diversos tipos de visualização como coropléticos (mapas coloridos), mapas com mini gráficos e mapas de sobreposição. O IBM SPSS Statistics trabalha com mapas no formato SMZ ou, ainda, em formato *shape* (ESRI) através de uma conversão, sendo que este precisa ter um arquivo DBF que o acompanhe na mesma pasta. Para utilizar esta conversão, basta acessar o menu **Utilitários** → **Utilitário de Conversão de Mapa** (ver Figura 142).

Além de converter arquivos no formato *shape* (ESRI), o Utilitário de Conversão de Mapa permite a modificação do nível de detalhe do mapa, como alteração dos rótulos, mesclar características e recursos de movimentos, entre muitas outras mudanças opcionais. Além disto, é possível usar o Utilitário para modificar um arquivo SMZ existente (incluindo os arquivos pré-instalados).

Figura 142 | Utilitário de Conversão de Mapa



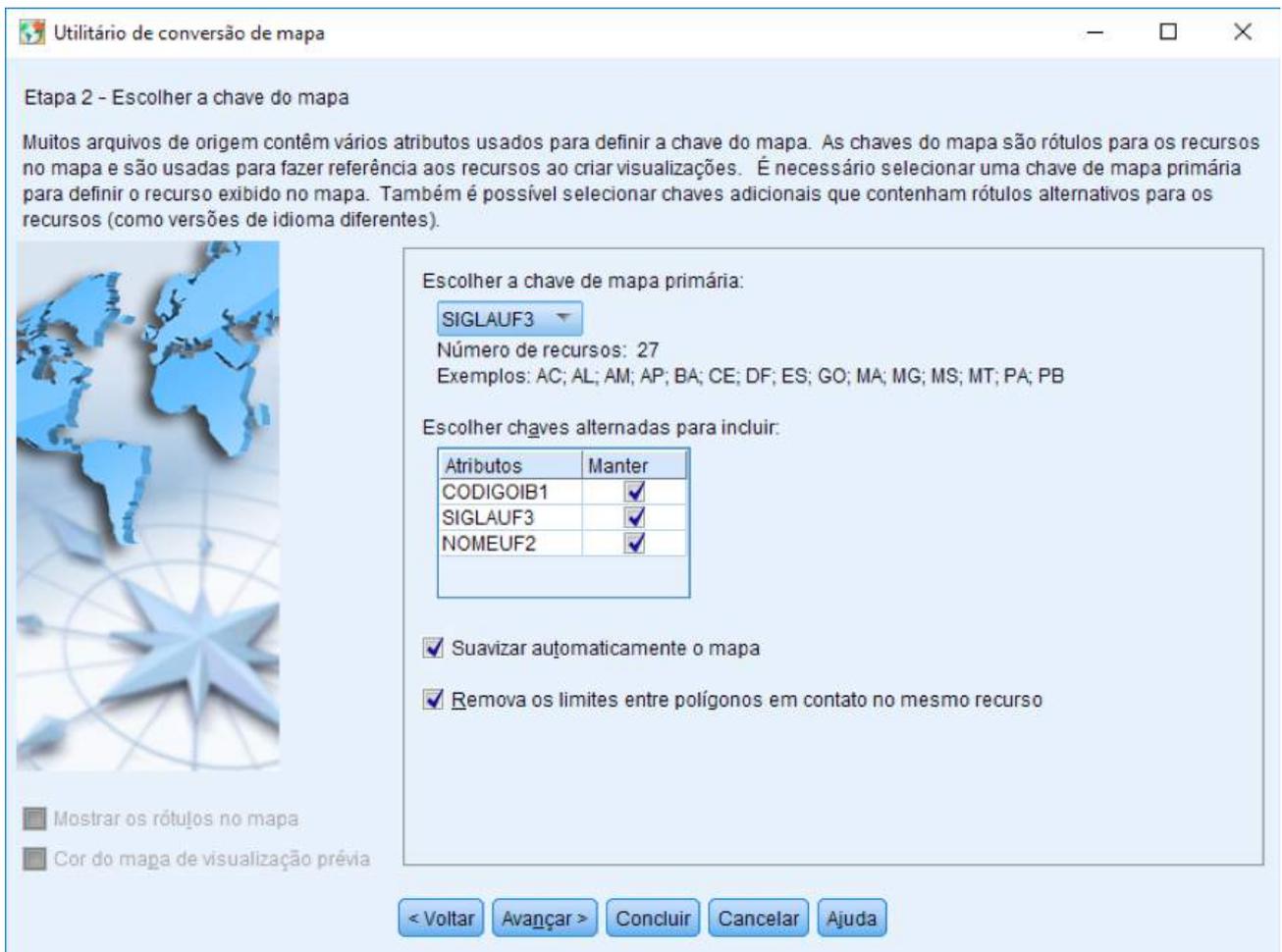
Para o nosso primeiro exemplo de utilização de mapas, precisaremos importar um arquivo de mapa para o IBM SPSS Statistics antes de abrirmos o arquivo de dados. O arquivo de mapa se chama *Brasil\_estados.smz* e ele deve ser adicionado na área **Selecione um arquivo .shp (ESRI) ou .smz para conversão** (clique no botão Navegar, localize o arquivo na pasta onde os arquivos do treinamento estão armazenados e, então, clique em **Abrir** – veja o resultado final na Figura 143). Quando um arquivo de mapa é exportado para o IBM SPSS Statistics, o Usuário de Conversão de Mapa gera um arquivo compatível com o IBM SPSS Statistics em um diretório diferente daquele onde o arquivo original está localizado (especifique um diretório e nome de arquivo diferente na área **Defina um nome de arquivo e destino para o arquivo de mapa convertido** – veja a Figura 144). No nosso exemplo, colocaremos o arquivo convertido na **Área de Trabalho** do computador (ele pode ser apagado posteriormente após o fim do procedimento de conversão).

Figura 143 | Arquivos utilizados no Utilitário de Conversão de Mapa



Clique em **Avançar** para seguir com a conversão. Na próxima etapa (veja a Figura 144), vamos definir as **chaves do mapa**, ou seja, os códigos que determinam a identificação de cada Estado brasileiro. Você pode especificar até três chaves de mapa neste arquivo: *CODIGOIB1* (código numérico dos Estados), *NOMEUF2* (nome por extenso dos Estados) e *SIGLAUF3* (sigla dos Estados). Selecione as três chaves de mapa e certifique-se de que *SIGLAUF3* seja a **chave primária** (ou seja, aquela que será a chave padrão no Seletor de Modelo de Tabelas de Gráficos quando o mapa *Brasil\_estados* for selecionado) – veja o resultado final na Figura 144.

Figura 144 | Definição das chaves do mapa e da chave primária (SIGLAUF3)



Agora, estamos prontos para a conversão do arquivo. Clique em **Concluir** para encerrar o procedimento. (Observação: Se você tivesse clicado em **Avançar**, você teria mais opções de edição do mapa, como suavizá-lo ou editar os valores das chaves.)

## 08.4 Construção de mapas através do Graphboard Template Chooser

A interface do **Graphboard Template Chooser** permite a construção de mapas através de modelos pré-instalados ou de arquivos convertidos.

Nesta seção, veremos dois exemplos de como criar um mapa pré-existente. O primeiro deles será construído a partir dos dados do arquivo *Brasil\_estados.sav*, que contém informações sobre o número de habitantes de cada Estado no Brasil em 2010 (ver Figura 145). Fica a cargo do usuário verificar os outros tipos de mapas existentes - consulte a **Ajuda** do IBM SPSS Statistics para conhecer todas as funções disponíveis.

Figura 145 | Visão geral do arquivo *Brasil\_estados.sav*

	NOMEUF2	SIGLAUF3	CODIGOIB1	pop_2010	var						
1	Acre	AC	12	707.125							
2	Alagoas	AL	27	3.093.994							
3	Amapá	AP	16	648.553							
4	Amazonas	AM	13	3.350.773							
5	Bahia	BA	29	13.633.969							
6	Ceará	CE	23	8.180.087							
7	Distrito Federal	DF	53	2.469.489							
8	Espírito Santo	ES	32	3.392.775							
9	Goiás	GO	52	5.849.105							
10	Maranhão	MA	21	6.424.340							
11	Mato Grosso	MT	51	3.033.991							
12	Mato Grosso do Sul	MS	50	2.404.256							
13	Minas Gerais	MG	31	19.159.260							

Para criar o mapa, basta acessar o menu **Gráficos**, opção **Seletor de Modelo de Tabelas de Gráficos**. Na aba **Básico**, você pode escolher quais variáveis serão utilizadas e, automaticamente, dependendo do nível de medida delas, a tela habilitará opções diferentes de gráficos. Em qualquer situação envolvendo gráficos com mapas é obrigatório que uma variável-chave (*map key*) seja selecionada. No nosso exemplo, vamos selecionar as variáveis *SIGLAUF3* (que será a nossa variável-chave de mapa) e *pop\_2010* e, também, certifique-se de que o gráfico **Coroplético de Médias** seja selecionado (ver Figura 146).

Em seguida, clique na aba **Detalhado** e certifique-se de que a variável *pop\_2010* esteja no quadro **Cor** e a variável *SIGLAUF3* esteja em **Chave de dados** (ver Figura 147). Neste tipo de gráfico, a estatística a ser adotada é a média da variável definida no quadro Cor para casos com o mesmo valor da variável-chave. Como só há um caso para cada Estado, este mapa é o ideal para representar a população total por Estado.

Agora, resta definir o mapa a ser utilizado. Clique na opção **Selecionar um Arquivo de Mapa** na caixa **Arquivos de mapa** para abrir a janela **Selecionar Mapas** e, em seguida, selecione *Brasil\_estados* no quadro **Mapa** e *SIGLAUF3* nos quadros **Chave do mapa** e **Chave de dados** (ver Figura 148). Repare que é possível validar a chave de mapa do banco de dados (*data key*) a partir do arquivo de mapa selecionado: basta clicar no botão Comparar. Se as chaves de mapa e dos dados estiverem corretamente pareadas, um *check* azul aparecerá na tela. Clique em OK.

Na aba **Títulos**, você tem a opção de definir o nome do mapa, bem como um subtítulo ou uma nota de rodapé caso julgue necessário. Finalmente, clique em OK para gerar o mapa (veja o resultado final na Figura 149).

Figura 146 | Graphboard Template Chooser – aba Básico

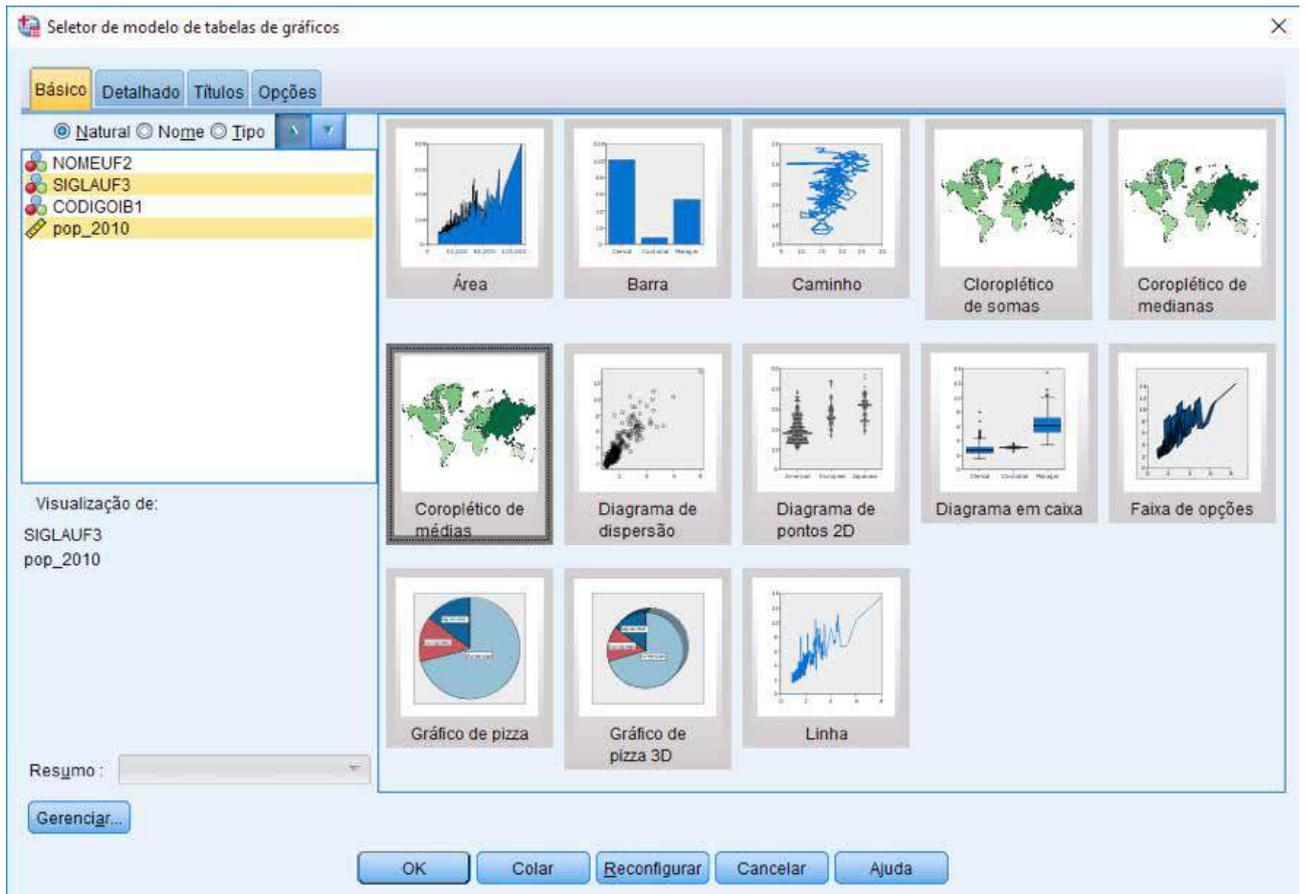


Figura 147 | Graphboard Template Chooser – aba Detalhado

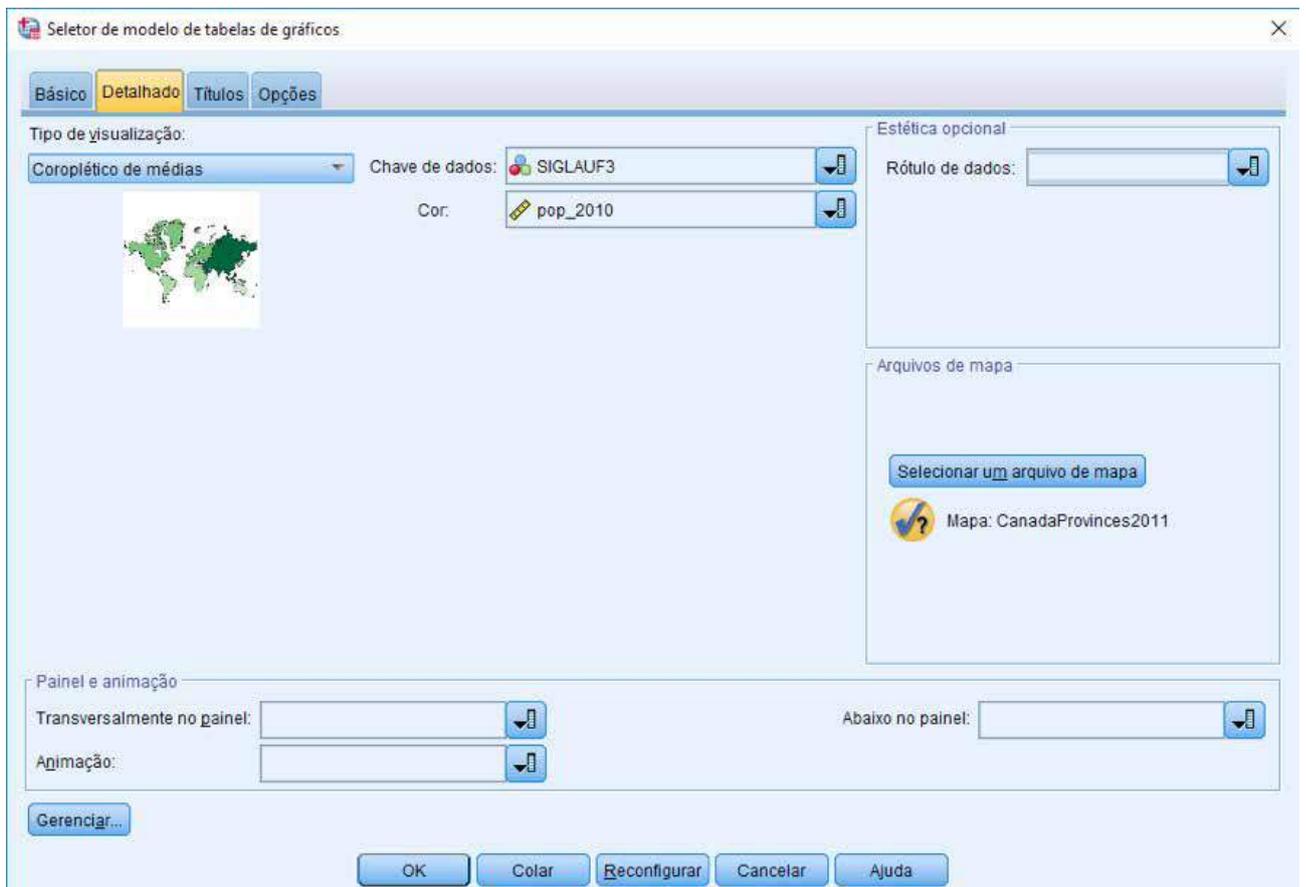


Figura 148 | Janela de seleção de mapas e validação da chave do mapa

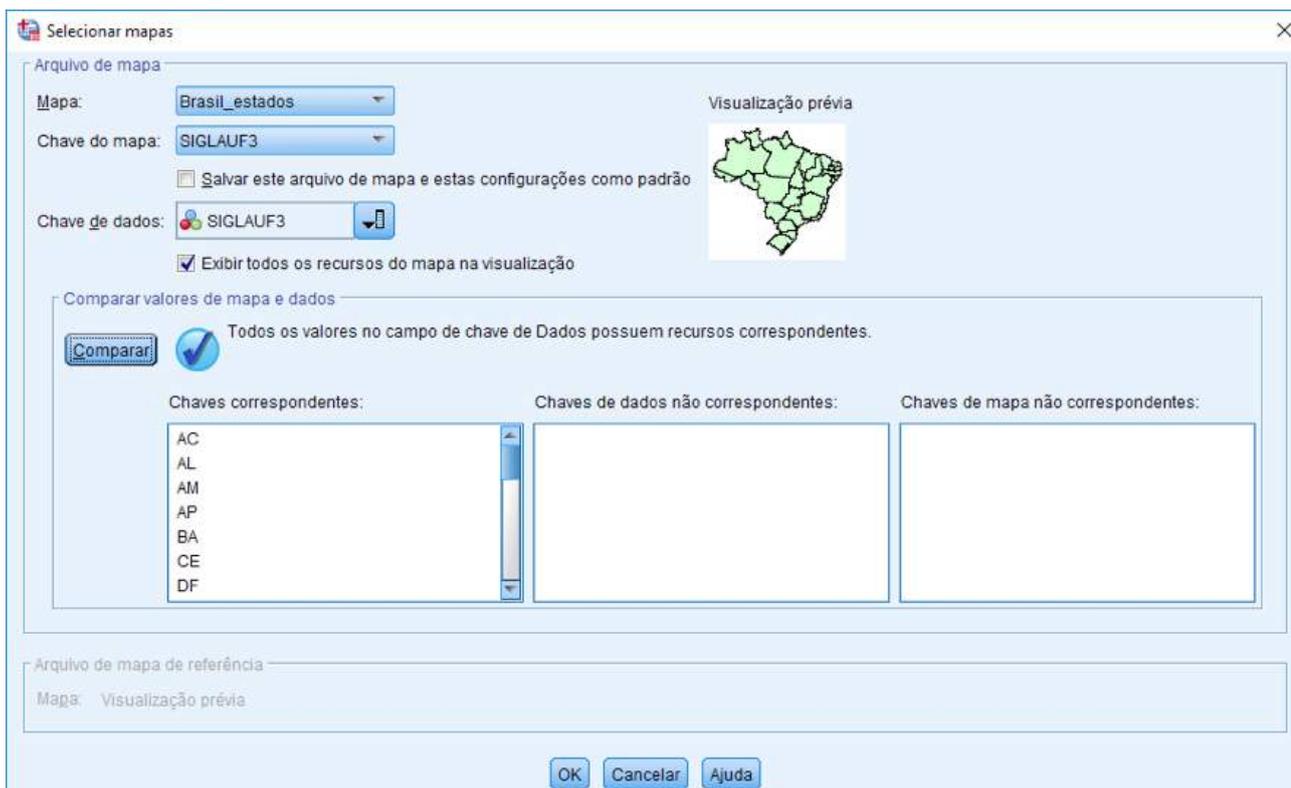
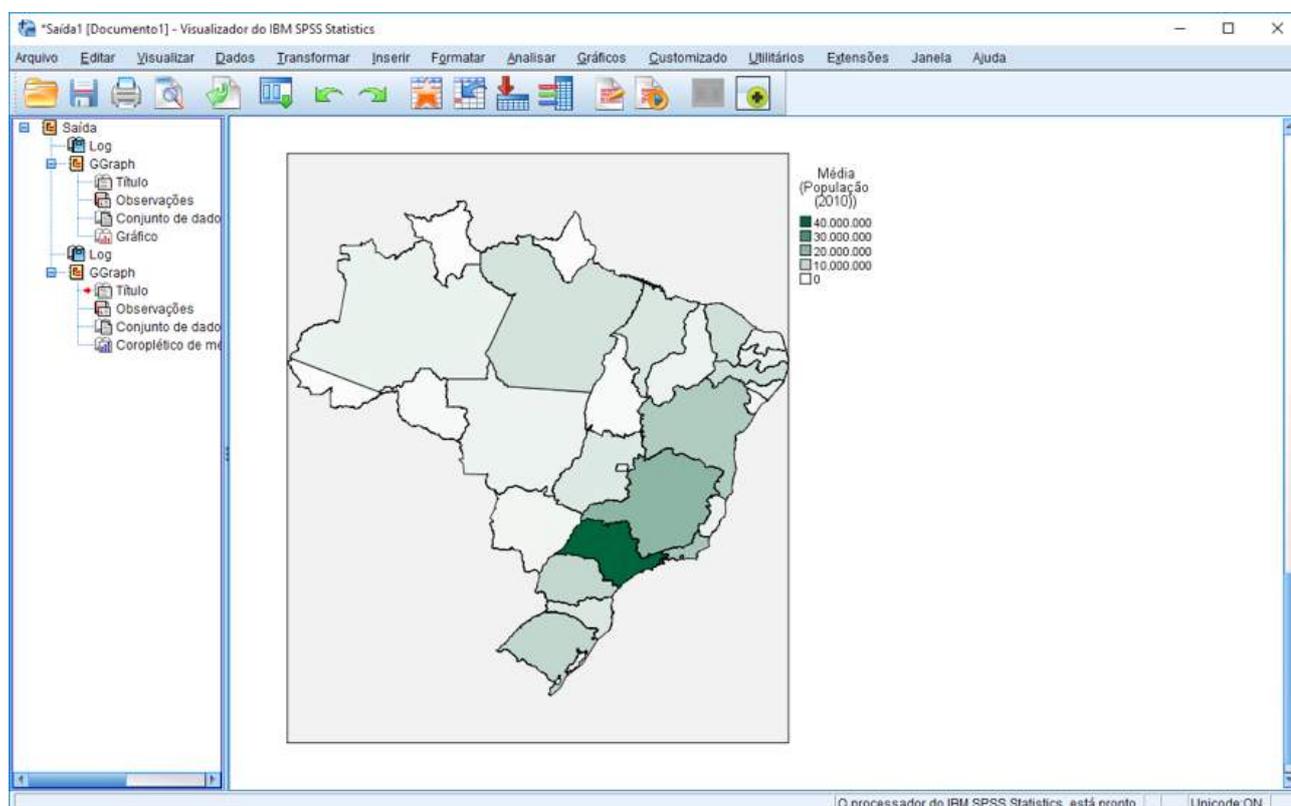


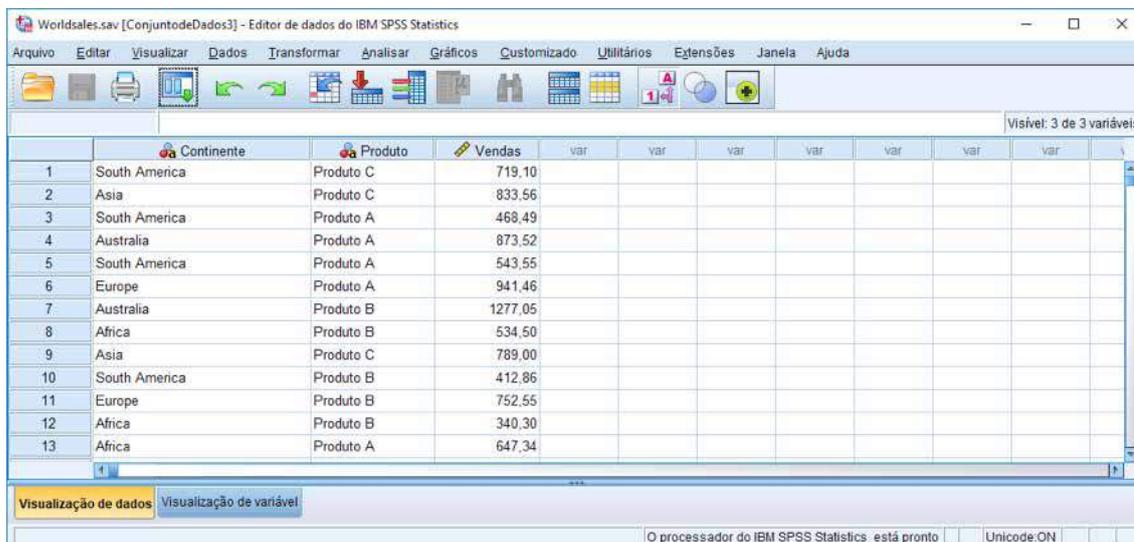
Figura 149 | Mapa gerado via Graphboard Template Chooser



No nosso segundo exemplo, abriremos o arquivo Worldsales.sav, que envolve vendas de três produtos por continente (ver Figura 150). Estamos interessados em construir um mapa com gráficos de barras para o valor médio das vendas dos três produtos (A, B e C) por continente. Para criar o mapa, basta acessar novamente o menu Gráficos, opção Seletor de Modelo de

Tabelas de Gráficos. Selecione as três variáveis (Continente, Vendas e Produto) e marque o gráfico Barras em um Mapa. Certifique-se de que a opção Média esteja selecionada no quadro Resumo (ver Figura 151).

Figura 150 | Visão geral do arquivo Worldsales.sav



	Continente	Produto	Vendas	var						
1	South America	Produto C	719.10							
2	Asia	Produto C	833.56							
3	South America	Produto A	468.49							
4	Australia	Produto A	873.52							
5	South America	Produto A	543.55							
6	Europe	Produto A	941.46							
7	Australia	Produto B	1277.05							
8	Africa	Produto B	534.50							
9	Asia	Produto C	789.00							
10	South America	Produto B	412.86							
11	Europe	Produto B	752.55							
12	Africa	Produto B	340.30							
13	Africa	Produto A	647.34							

Em seguida, clique na aba **Detalhado** e certifique-se de que a variável *Produto* esteja no quadro **Categorias**, a variável *Vendas* em **Valores** e a variável *Continente* em **Chave de dados**. Em **Arquivos de mapa**, clique na opção **Selecione um Arquivo de Mapa** (ver Figura VIII-20) e, em seguida, na janela **Selecionar Mapas**, escolha *Continents* em **Mapa** e a variável *Continente* em **Chave de dados**. Clique no botão Comparar para verificar se as chaves de mapa e dos dados estão pareadas (ver Figura 152).

Clique em **OK** para fechar a janela de seleção e, finalmente, em **OK** novamente para gerar o mapa com os gráficos de barras representando a média de vendas por produto em cada continente (ver Figura 153).

Figura 151 | Graphboard Template Chooser – aba Básico

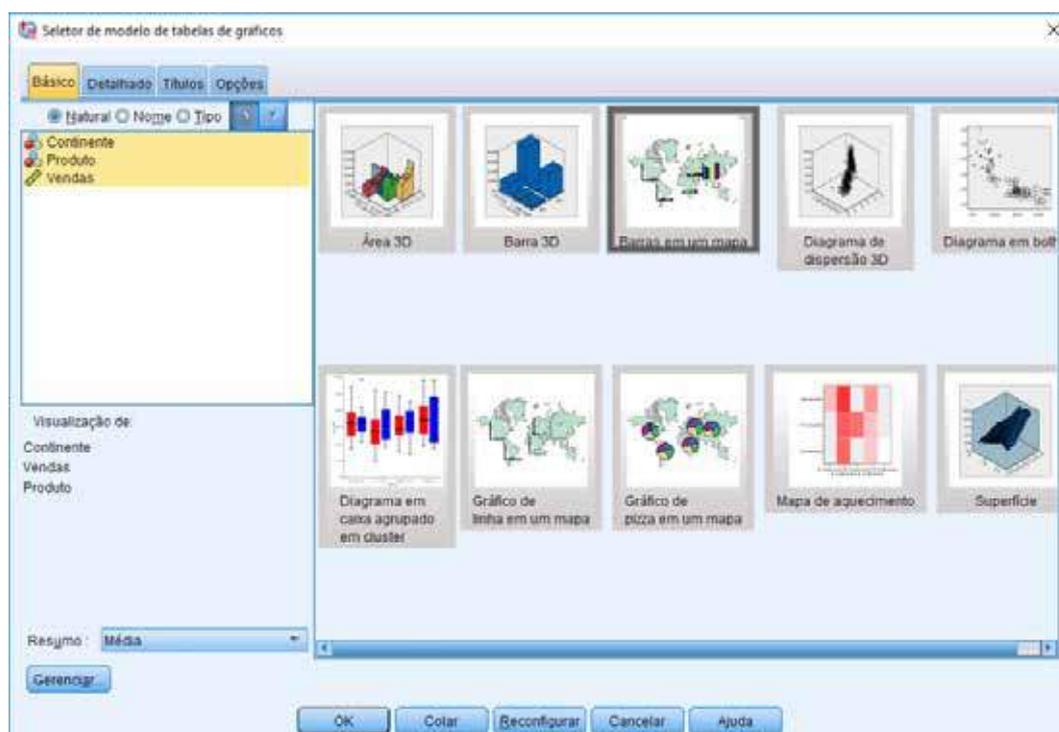


Figura 152 | Graphboard Template Chooser – aba Detalhado

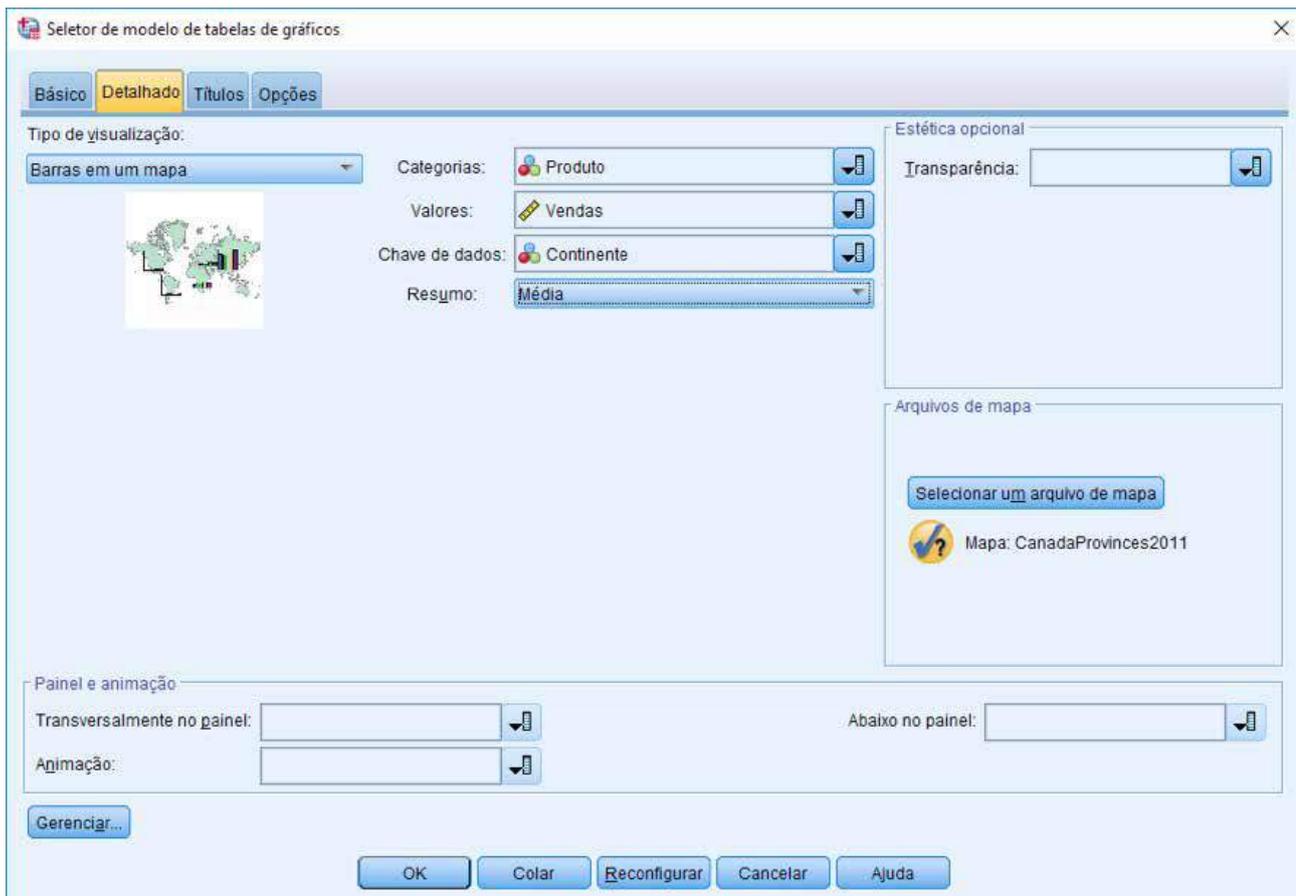


Figura 153 | Janela de seleção de mapas e validação da chave do mapa

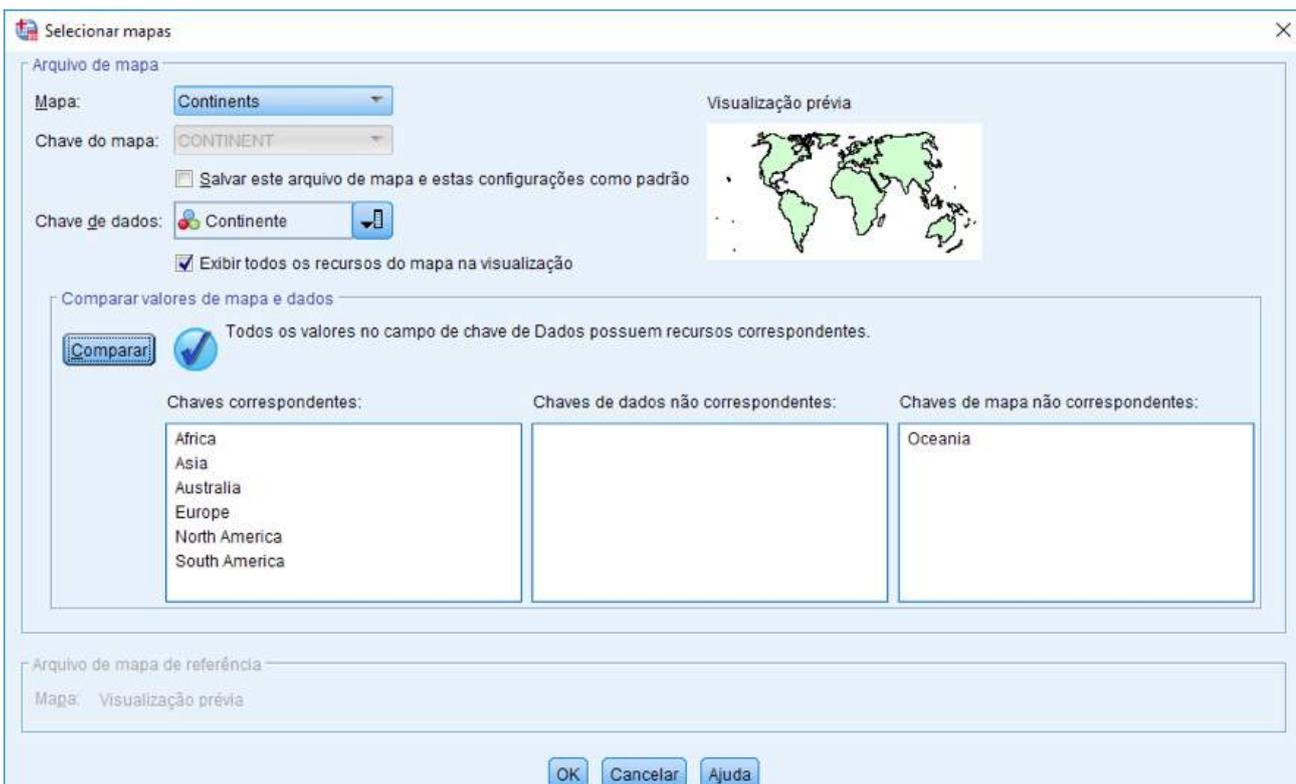
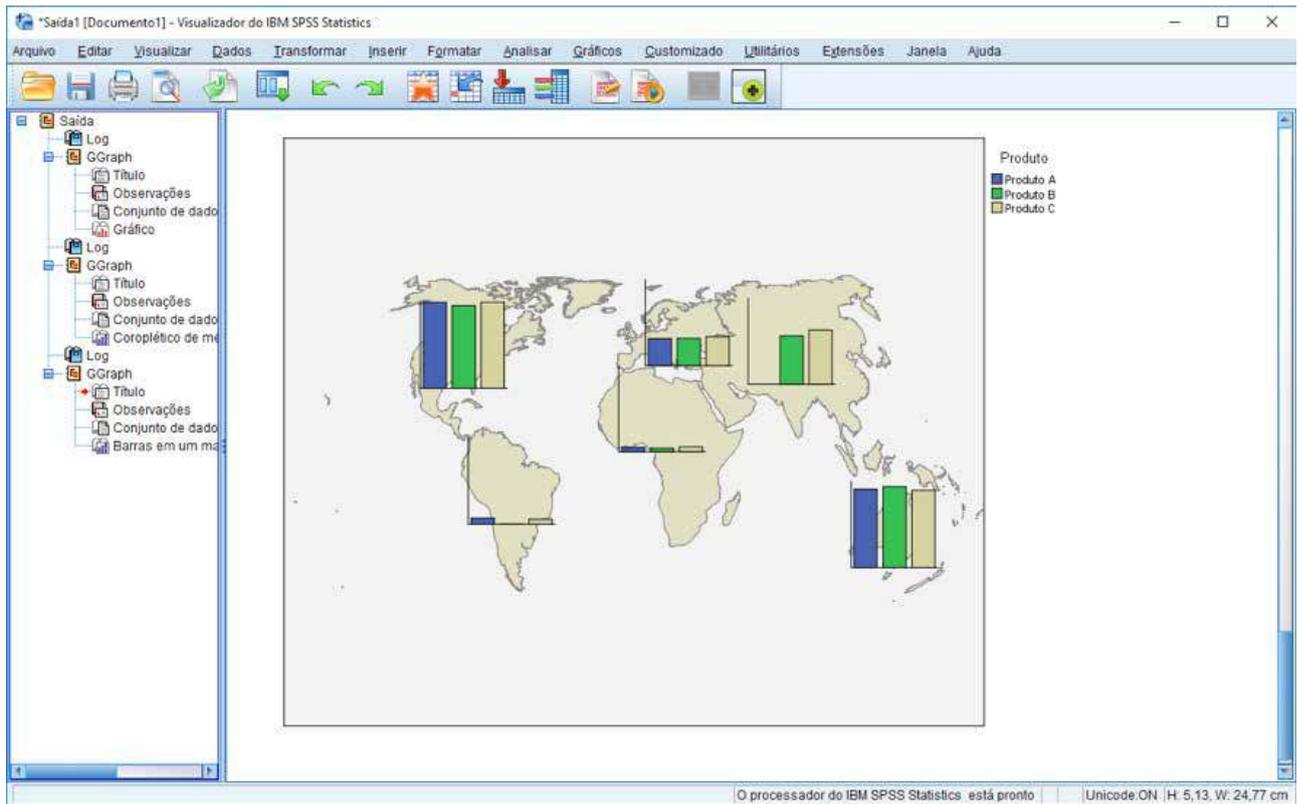


Figura 154 | Mapa gerado via Graphboard Template Chooser













**UniCesumar**

---

Av. Guedner, 1610 - Jd. Aclimação - Maringá - PR  
44. 3027 6360 • 0800 604 4422  
[www.unicesumar.edu.br](http://www.unicesumar.edu.br)