Análise descritiva no RCommander (INI/FIOCRUZ)

Raquel de Vasconcellos Carvalhaes de Oliveira
Luan Nóe da Silva
Colaboração na elaboração: Fabiano Marcos

Com o banco de dados birthwt, vamos realizar uma análise descritiva das variáveis quantitativas e qualitativas por medidas-sumário, tabelas de frequência, gráficos e tabelas.

Resumos Estatísticos

Importante: Antes de realizar qualquer resumo do banco de dados verifique a estrutura do seu banco de dados pelo comando str(banco)!!!!!

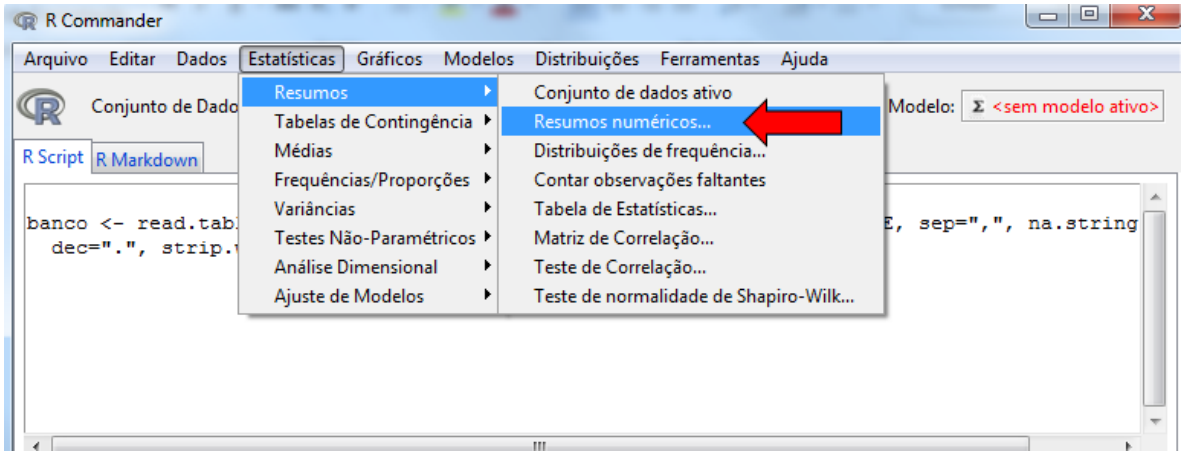
Obtendo o resumo da variável quantitativa “age” pelo script:

summary(banco$age)

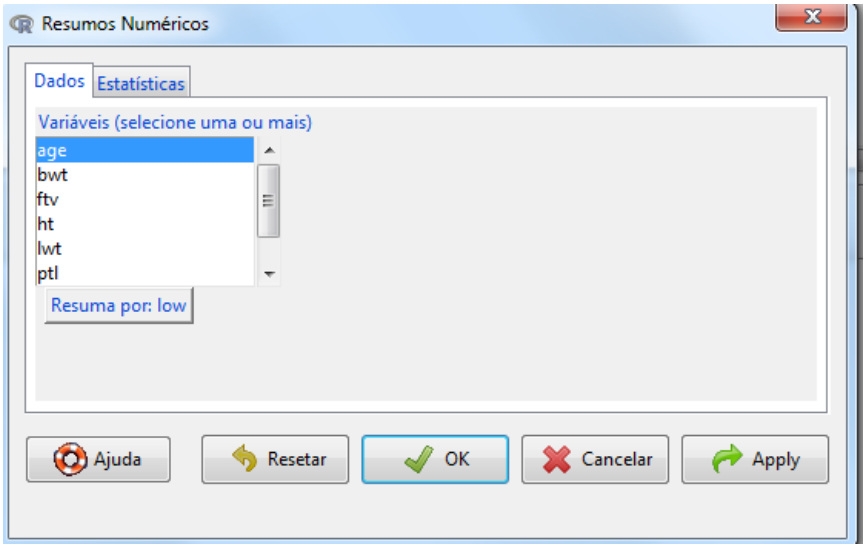
```
> summary(banco$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.00   19.00   23.00   23.24   26.00   45.00
```

Uma outra forma se obter resumos estatísticos de variáveis quantitativas pelo R Commander:

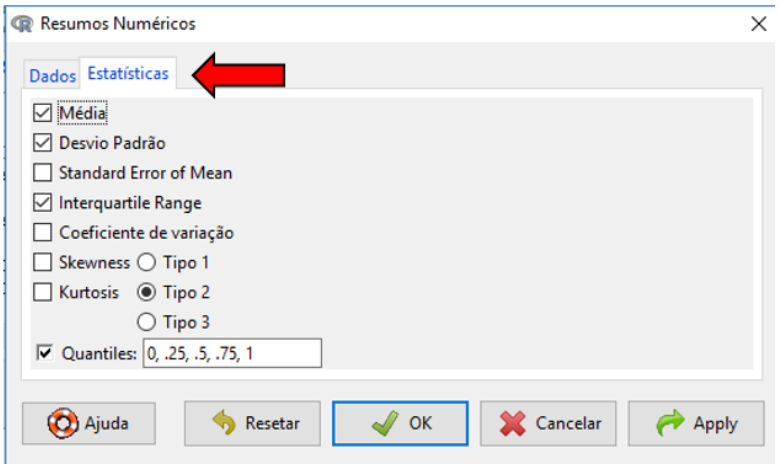
“Estatística -> Resumos -> Resumos Numéricos”



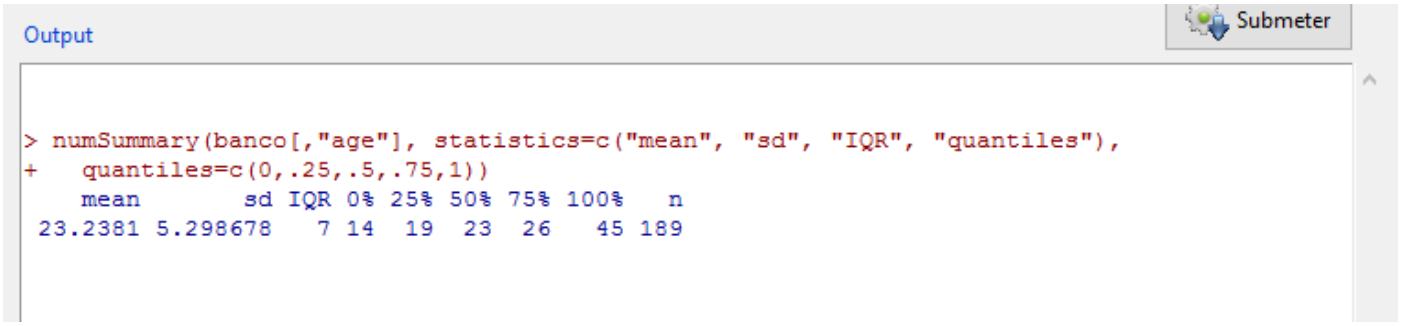
Selecione a(s) variável(is) que deseja realizar o resumo (Para selecionar mais de uma, utilize o CTRL).
OBS: Se desejar que os resultados saiam separados por alguma categoria de outra variável, clique no botão “Resuma por” e escolha a variável qualitativa correspondente a estratificação.



Clique na aba “Estatística” e escolha as medidas-sumário desejadas (exemplo: média, quartis, desvio-padrão) e dê OK.



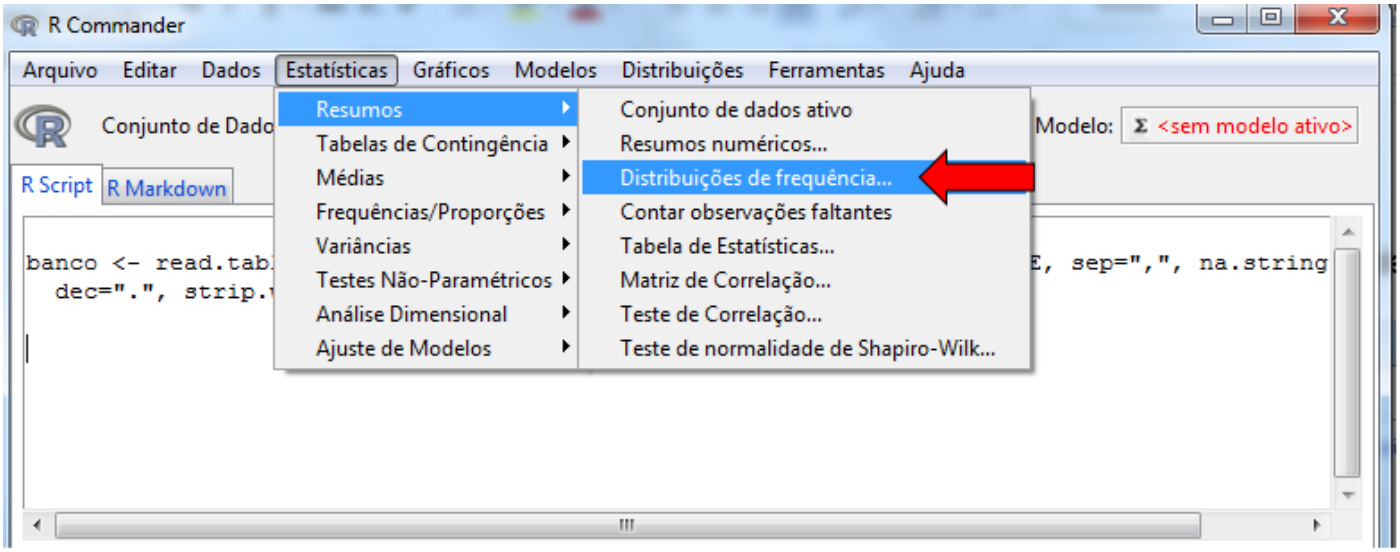
Com a variável “age” temos o seguinte resultado:



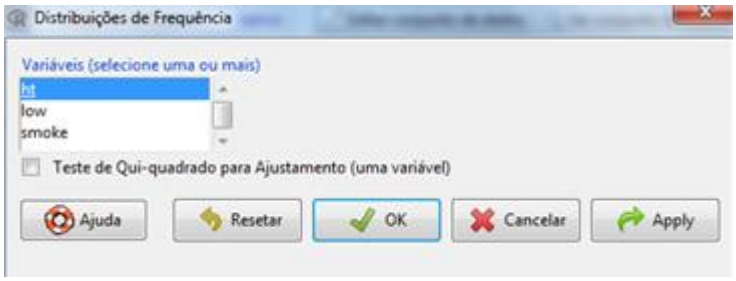
Tabelas de distribuição de frequências para variáveis qualitativas (ex:ht, low, smoke) podem ser realizadas pelo comando “table()” no script. Exemplo: O comando “table(banco$ht)“ cria uma distribuição de frequência para a variável “ht”.

No R Commander:

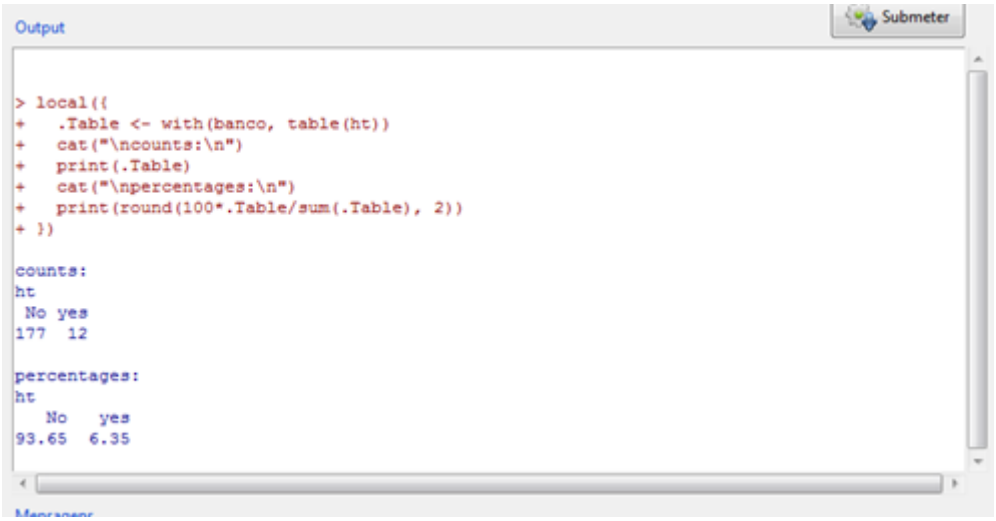
“Estatísticas-> Resumos -> Distribuições de frequência “



Feito isso escolhe-se a(s) variável(is) que se deseja realizar a distribuição de frequência. Vamos escolher a variável “ht” e dar OK.



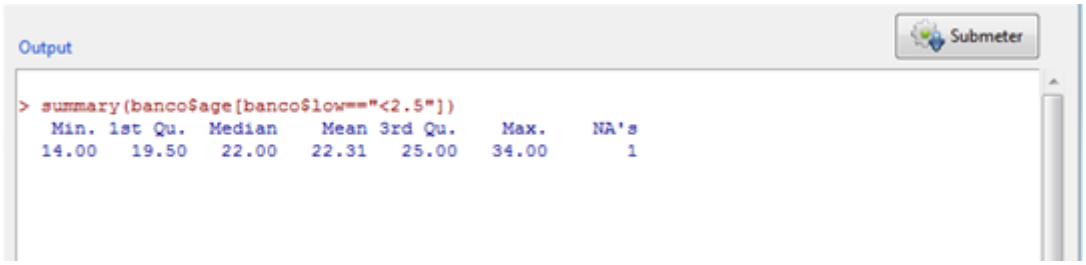
Observação: Observe que a caixa de diálogo só mostra as variáveis definidas como qualitativas (factor), visto que não faz sentido tirar % de variáveis quantitativas.



Outras medidas sumárias do banco de dados podem ser realizadas em subconjuntos de dados (filtros).

Exemplo: Vamos pedir o sumário da variável “age” das mães com filhos que nasceram com baixo peso.

summary(banco$age[banco$low=="<2.5"])



Vamos pedir o sumário da variável “age” das mães com filhos que nasceram com peso normal:

summary(banco$age[banco$low==">2.5"])

Vamos agora pedir o sumário da variável “age” das mães com filhos com baixo peso e que fumaram:

summary(banco$age[banco$low=="<2.5"&banco$smoke=="Sim"]) ou
summary(banco$age[banco$low=="<2.5"&banco$smoke=="Sim"])

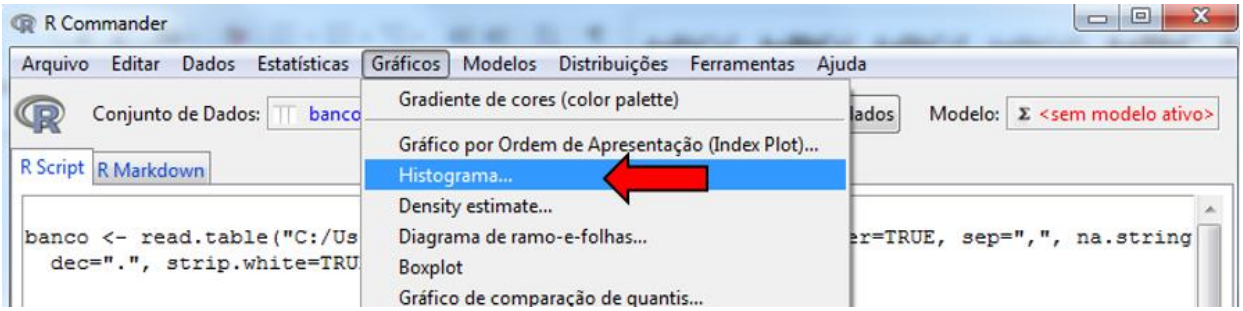
Gráficos

Histograma

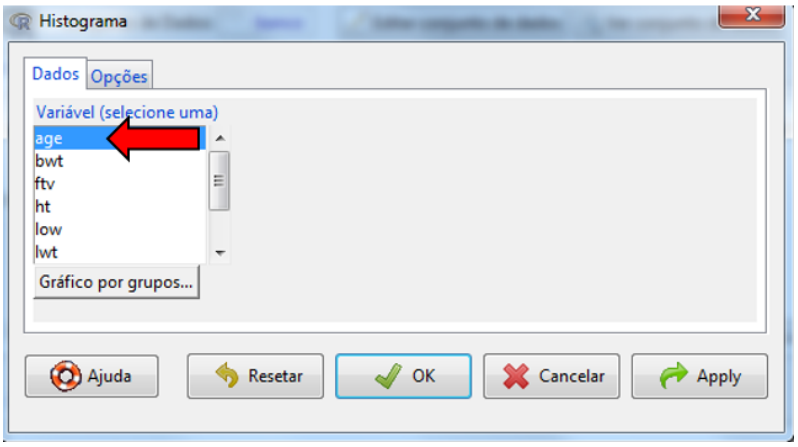
Para fazer um Histograma para a variável quantitativa “age” digite o comando na janela de script:

```
h1<-hist(banco$age,main="Histograma da variável idade")
```

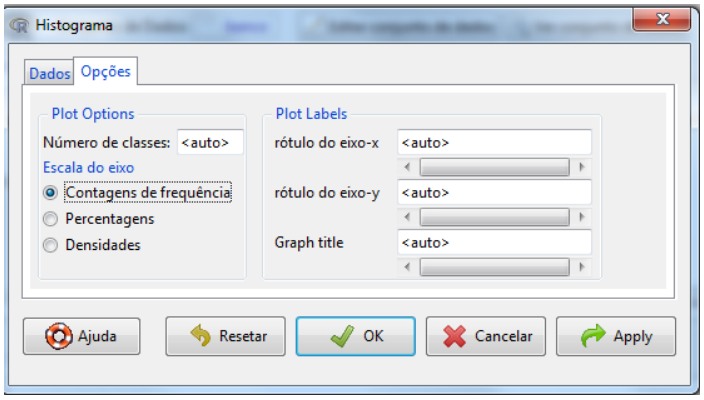
Ou faça o seguinte no menu do R Commander.



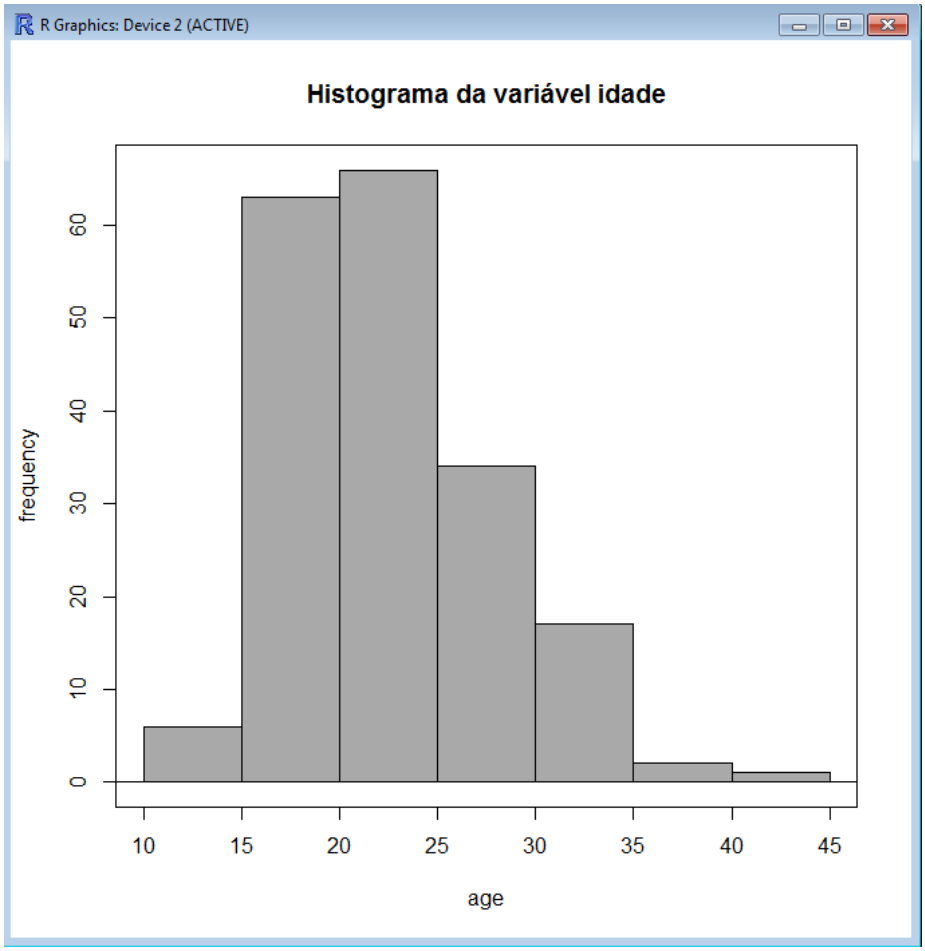
Ao clicar na opção do Histograma aparecerá a seguinte caixa de diálogo para realizar a escolha da variável, nosso exemplo, escolha a variável age.



Para criar uma legenda, modificar o rótulo dos eixos x e y ou mudar a escala do eixo, clique na aba “Opções” e parecerá a seguinte tela:



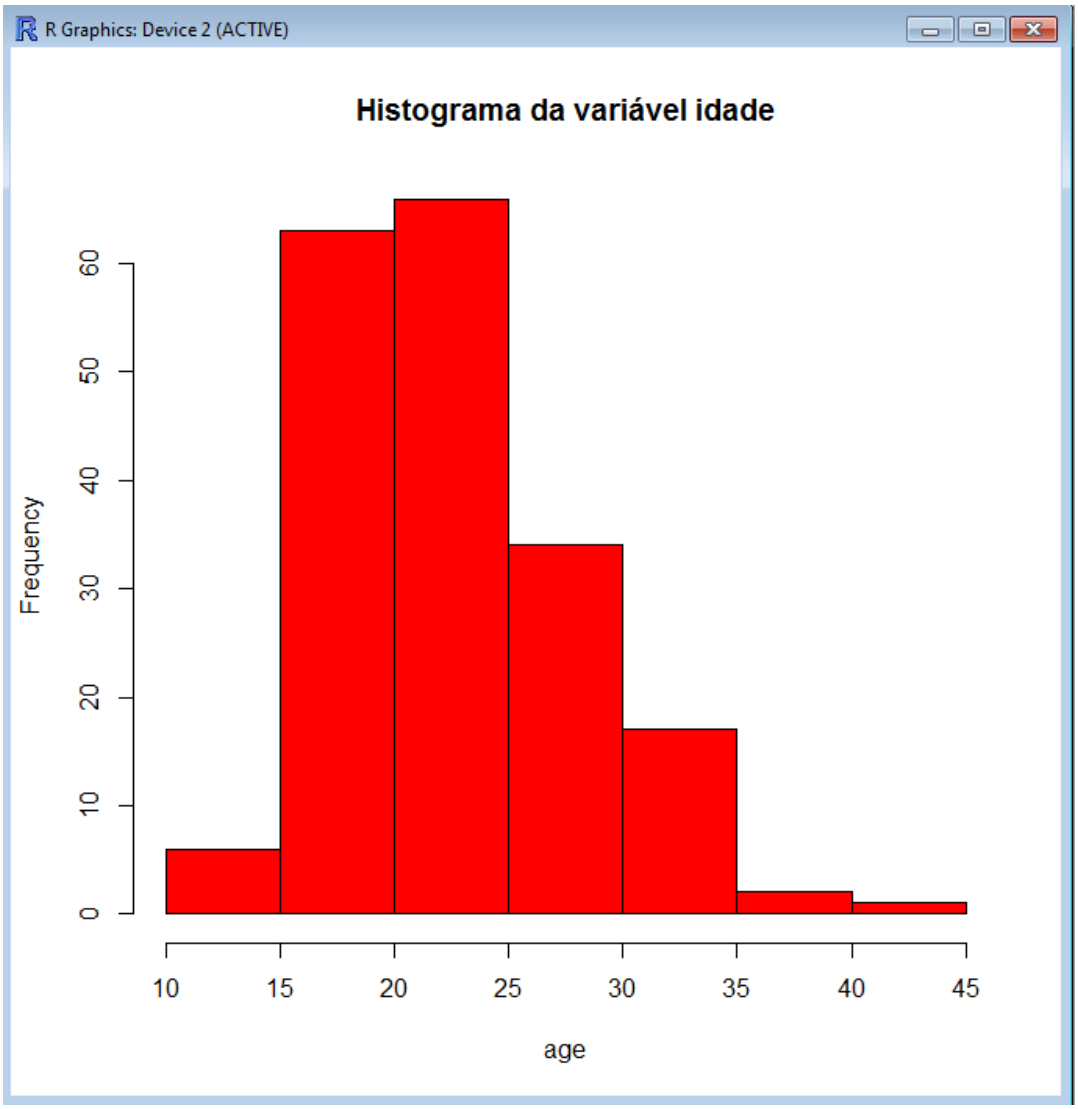
Clique em OK. Para o nosso exemplo aparecerá o seguinte gráfico:



OBS: Só é possível mudar as cores do gráfico modificando e executando o comando no script!

No exemplo abaixo, editamos o comando anterior para criar um histograma de cor vermelha.

```
h1<-hist(banco$age,main="Histograma da variável idade",xlab="age",col="red")
```



Histograma para variável age

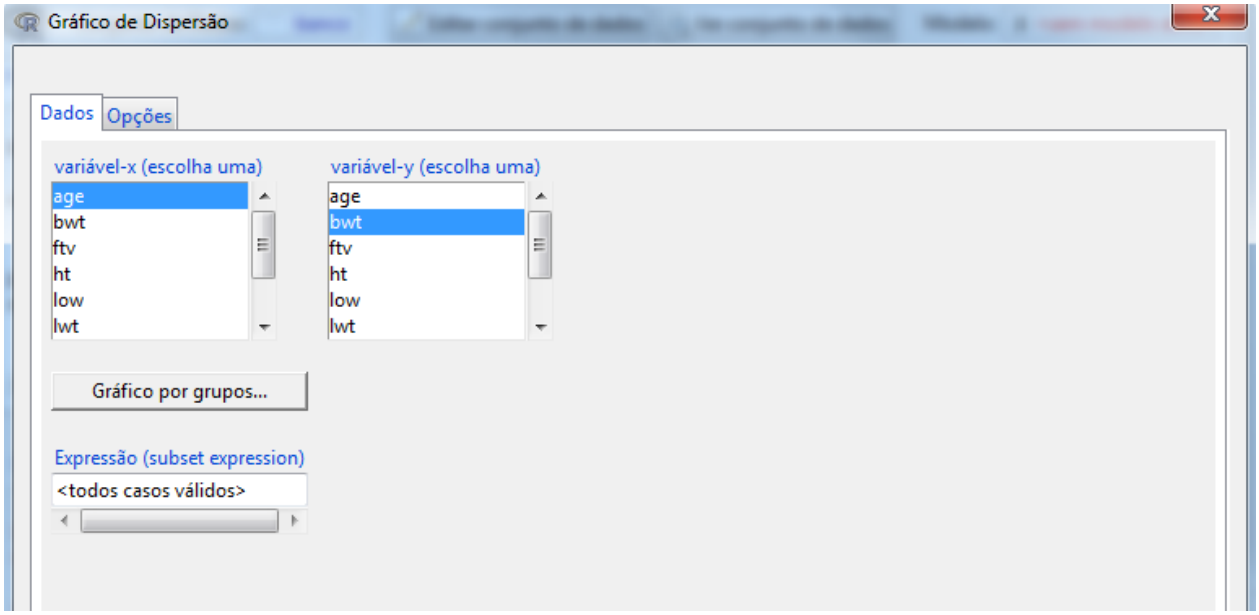
h2<-hist(banco$age,main="Histograma da variável idade", xlab=”age”,col="red",probability=TRUE)

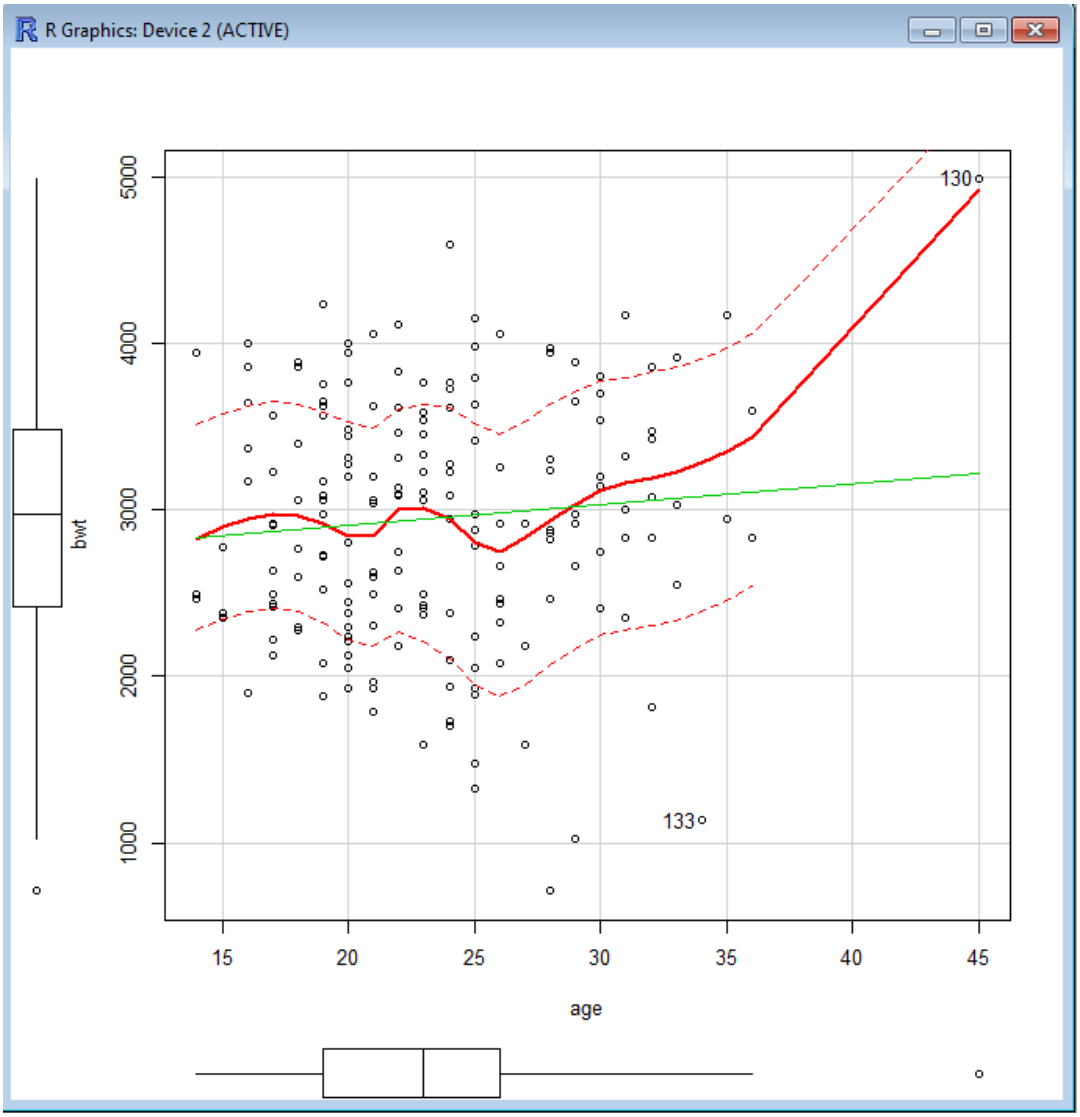
Gráfico de Dispersão

Digite no script plot(banco$age,banco$bwt)

ou clique no menu do R Commander: “Gráficos> Diagrama de dispersão”

Escolhas as variáveis desejadas para os eixos x e y e clique em Ok. No caso do nosso exemplo, queremos ver a relação entre as variáveis “age” e “bwt”.





Pergunta: O que acontece com o peso da criança quando a idade da mãe tende a crescer?

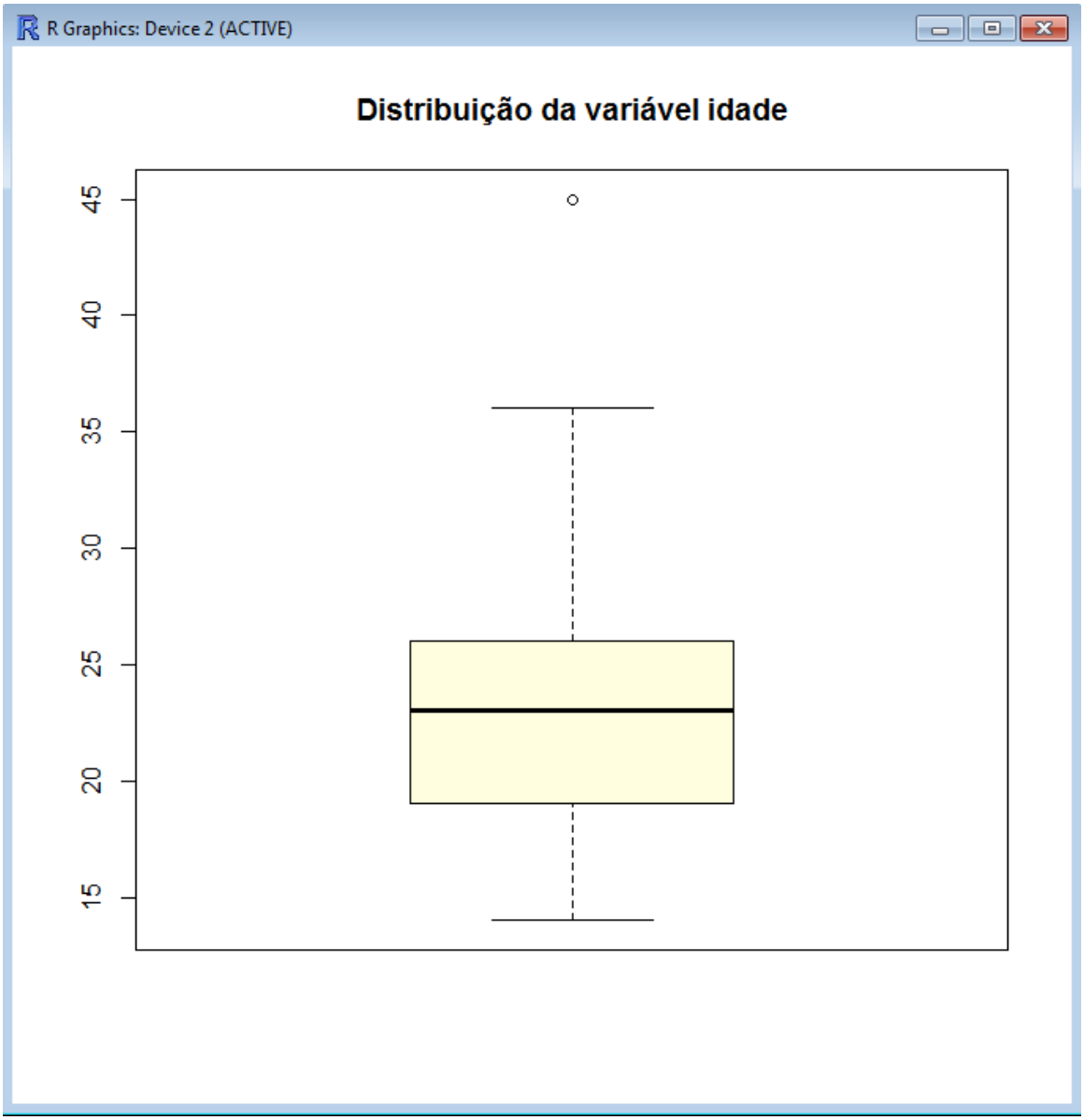
Box Plot

Para criar um boxplot da variável “age” por exemplo, faça o seguinte caminho no R Commander:

“Gráficos->Boxplot”

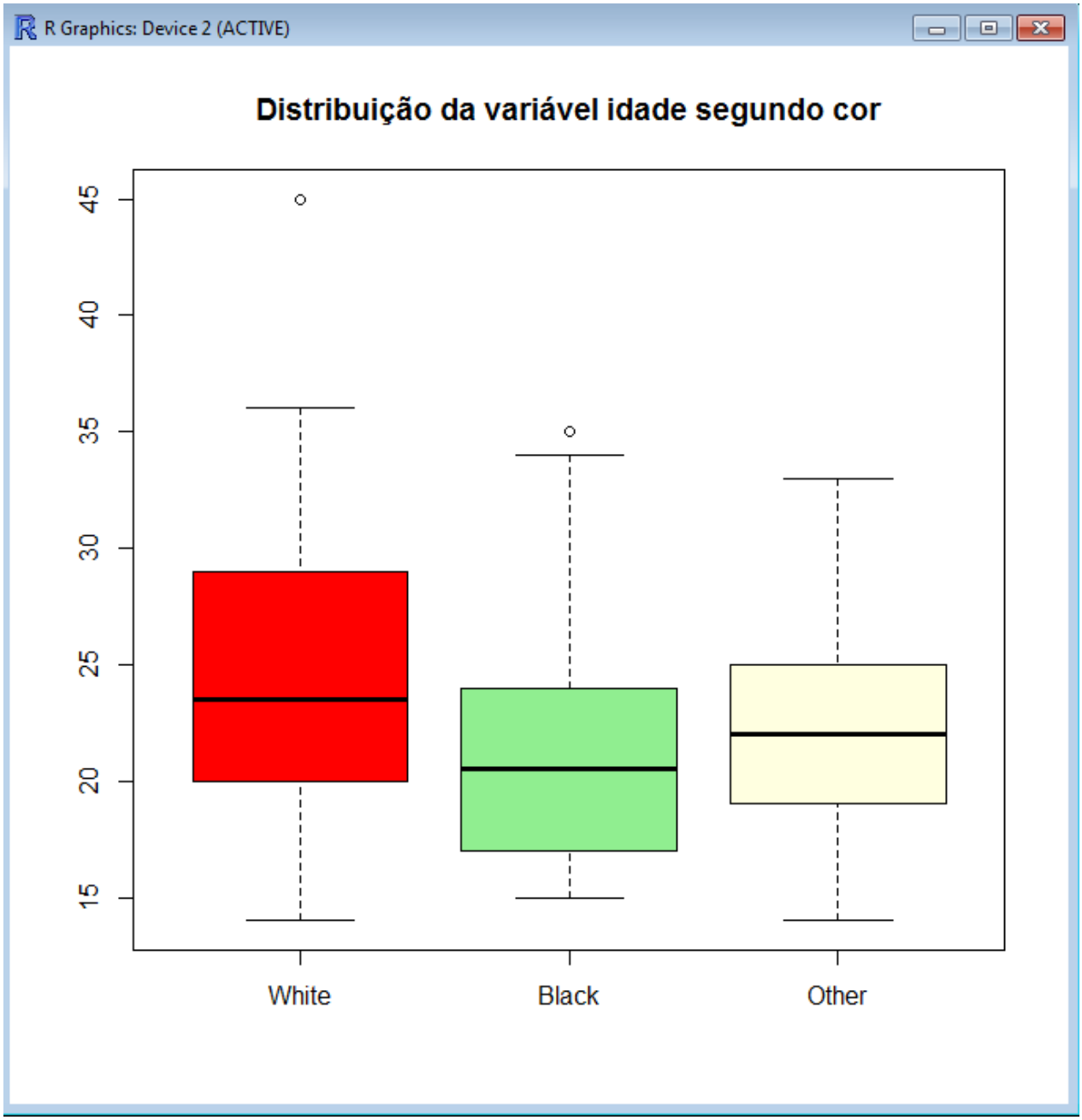
ou digite o seguinte comando no R script:

boxplot(banco$age,main="Distribuição da variável idade", col="lightyellow")



Boxplot da distribuição da variável quantitativa idade segundo a variável qualitativa cor. No R Script digite o seguinte comando:

```
boxplot(banco$age~banco$race,main="Distribuição da variável idade segundo cor", col=c("red","lightgreen","lightyellow"))
```



Pergunta: Parece ter diferença da idade nos grupos?

Gráfico de Setores (Pizza)

Para criar um gráfico de setores para variável “race” digite o comando no R Script:

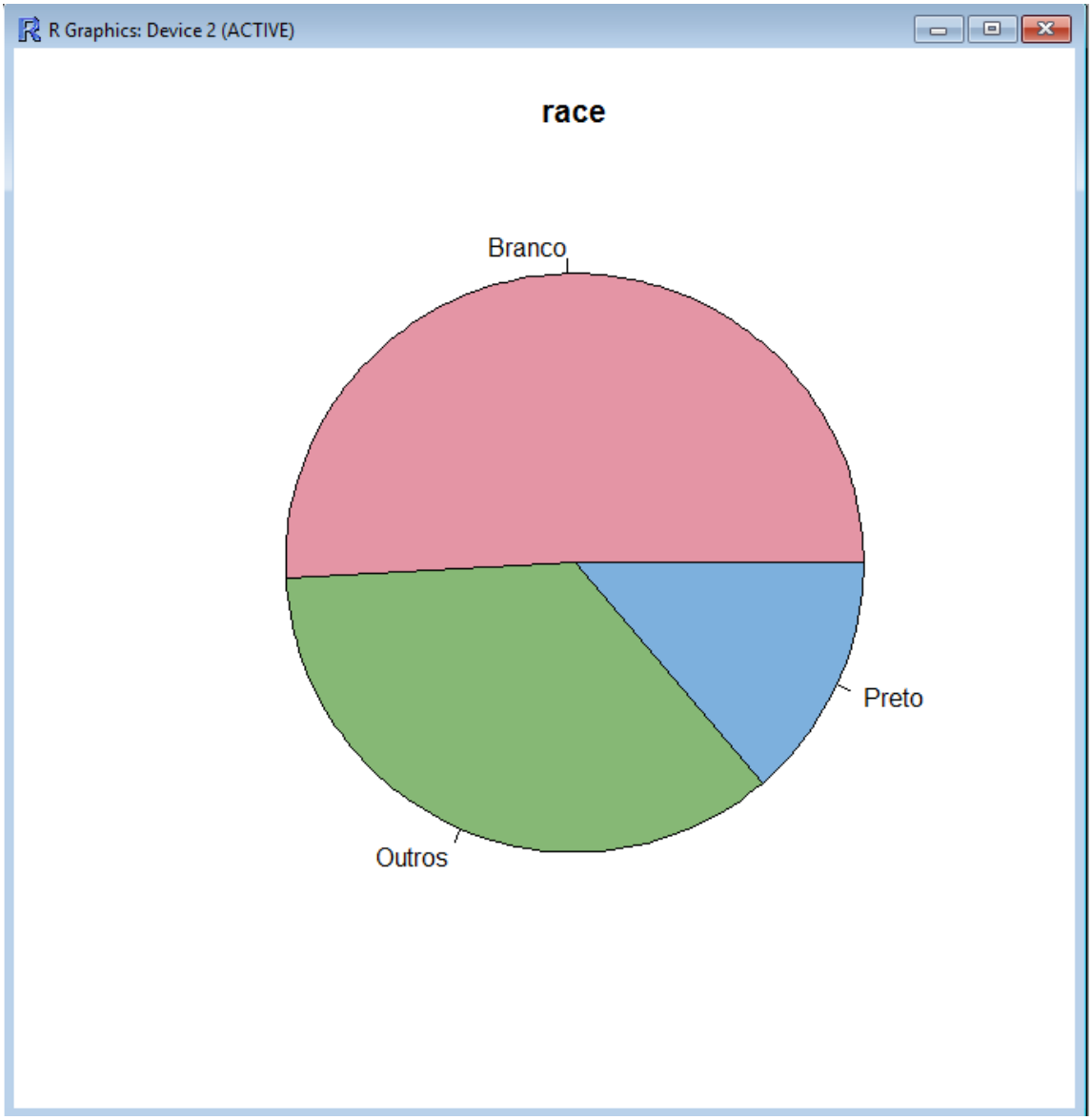
pie(table(banco$race))

ou faça o seguinte caminho no R Commander

“Gráficos > Gráfico de pizza “

Aparecerá a opção com todos os tipos de variáveis qualitativas, selecione a variável desejada e escolha as opções do gráfico.

Para o nosso exemplo o gráfico de setores para a variável “race” fica da seguinte forma:

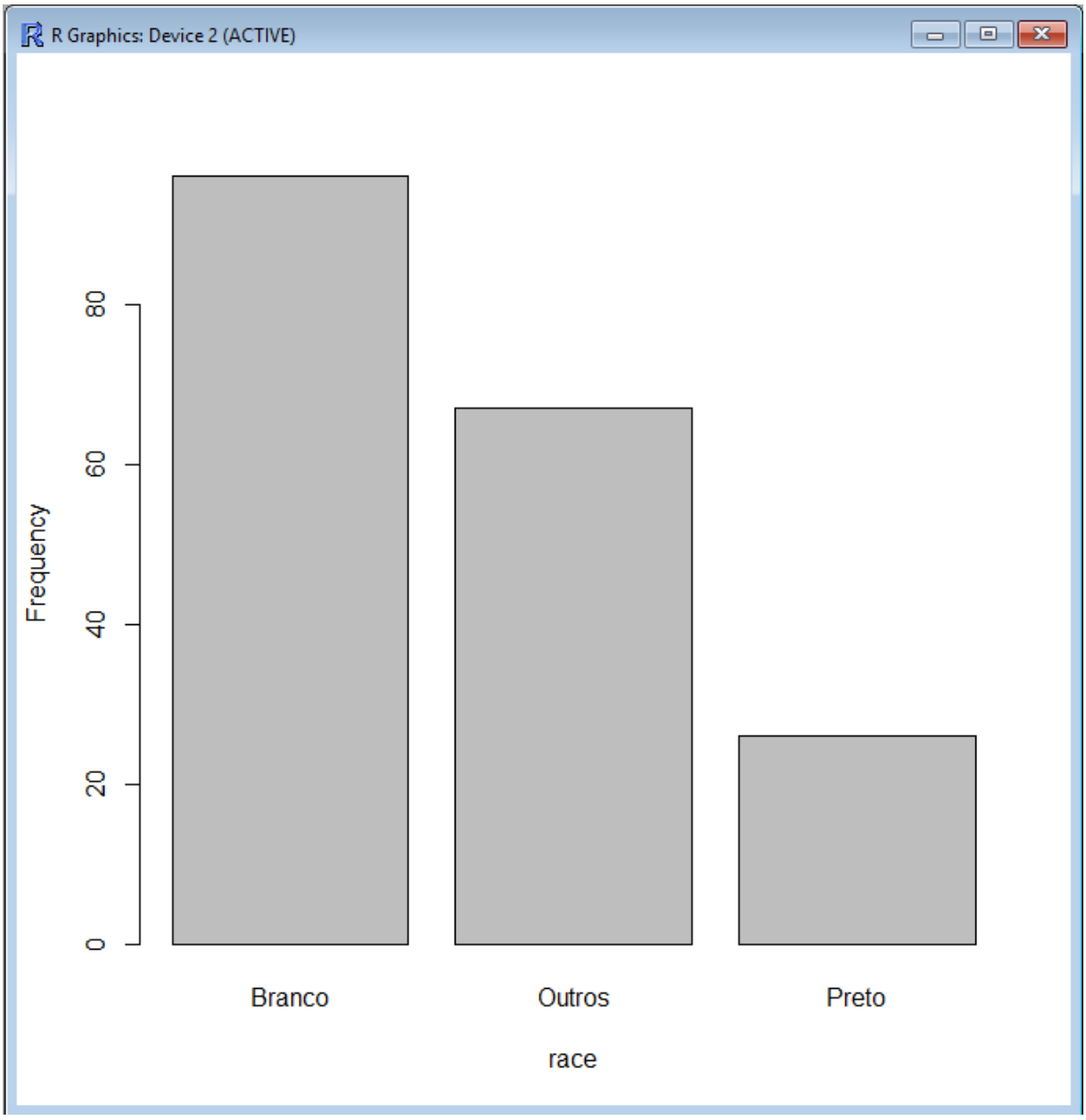


OBS: Só é possível colocar a porcentagem de cada categoria pelo R script!!! Editando os comandos anteriores:

```
pie(table(banco$race),main="Gráfico de setores: Raça")
text(locator(n=1),paste(round(prop.table(table(banco$race))[1],digits=2)*100,"%"))
text(locator(n=1),paste(round(prop.table(table(banco$race))[2],digits=2)*100,"%"))
text(locator(n=1),paste(round(prop.table(table(banco$race))[3],digits=2)*100,"%"))
```

Após executar o comando, clique em cada categoria para adicionar seus respectivos valores.

Da mesma maneira podemos criar um gráfico de barras para a variável “race” no R Commander



OBS: Caso no banco de dados não existe nenhuma variável do tipo fator (variáveis qualitativas) as opções de gráficos de setores e de barra ficam desabilitadas.

Gráfico de Linha

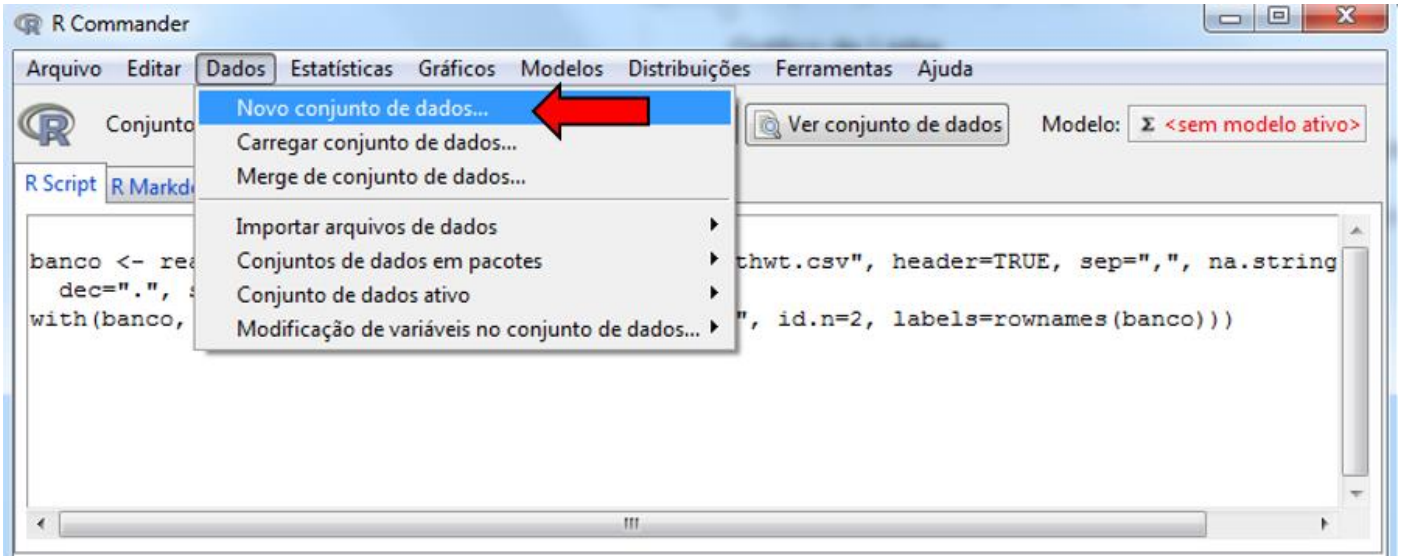
Esse banco de dados não possui dados com temporalidade. Assim, para fazer um gráfico de linha vamos criar dois vetores com informações de taxa de mortalidade e os anos de ocorrência, a fim de avaliar a evolução temporal de uma taxa ao longo do tempo. *OBS: Se os dados tivessem digitados previamente em uma planilha bastaria importar os dados.*

Na janela de script:

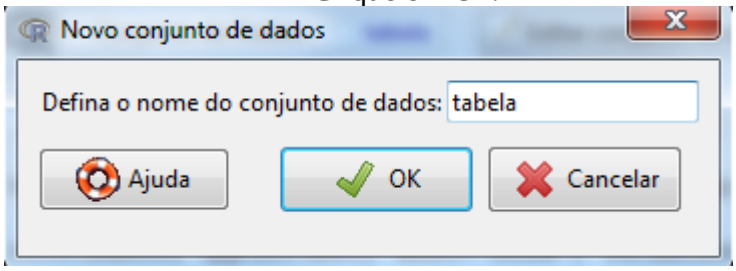
```
TM<-c(50,40,25,35,10,5)
anos<-c(1980,1985,1990,1995,2000,2005)
tabela<-as.data.frame(cbind(TM,anos))
tabela

plot(tabela$anos, tabela$TM, type="l")
```

No menu do R Commander:

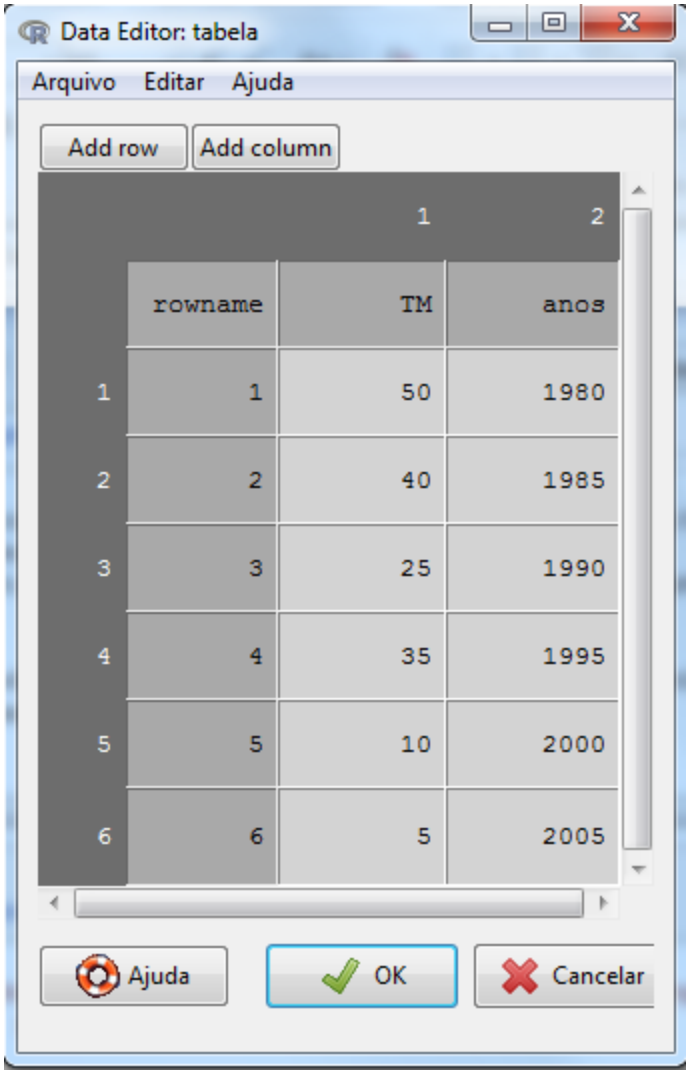
Clique em "Dados > Novo conjunto de dados "



Então aparecerá a opção para definir o nome do novo conjunto de dados.
Para o nosso exemplo vamos chamar de "tabela". Clique em Ok.

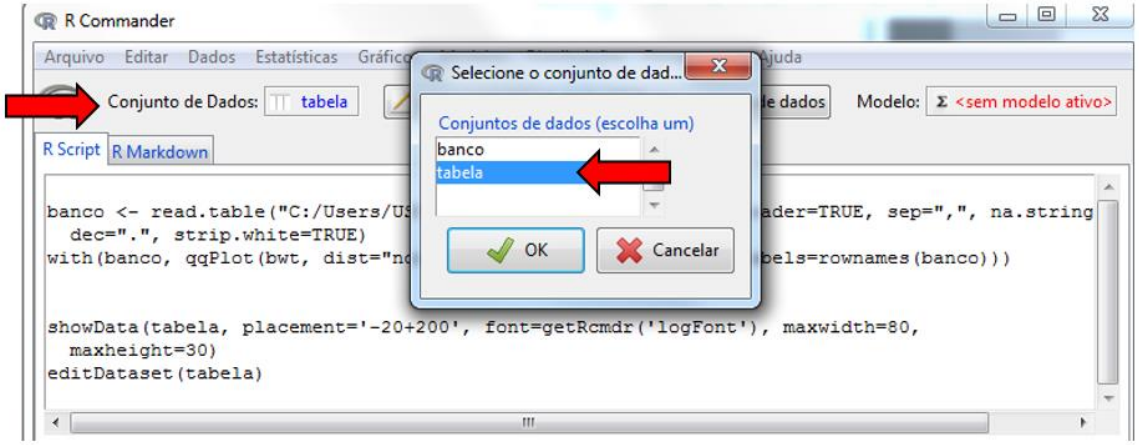


Aparecerá uma opção para adicionar linhas e colunas. Adicione linhas e colunas conforme dado em nosso exemplo:



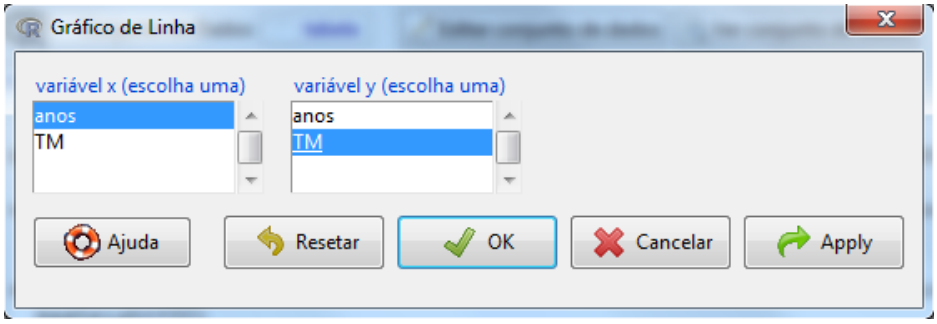
	rowname	TM	anos
1	1	50	1980
2	2	40	1985
3	3	25	1990
4	4	35	1995
5	5	10	2000
6	6	5	2005

Atenção: Após carregar mais de um banco, deve-se informar ao R qual o banco de dados está ativo para o menu. A troca do banco de dados ativo pode ser feita clicando no campo “Conjunto de dados: ” onde aparecerá as opções de banco. Para o nosso exemplo escolha o banco “tabela” que acabamos de criar.

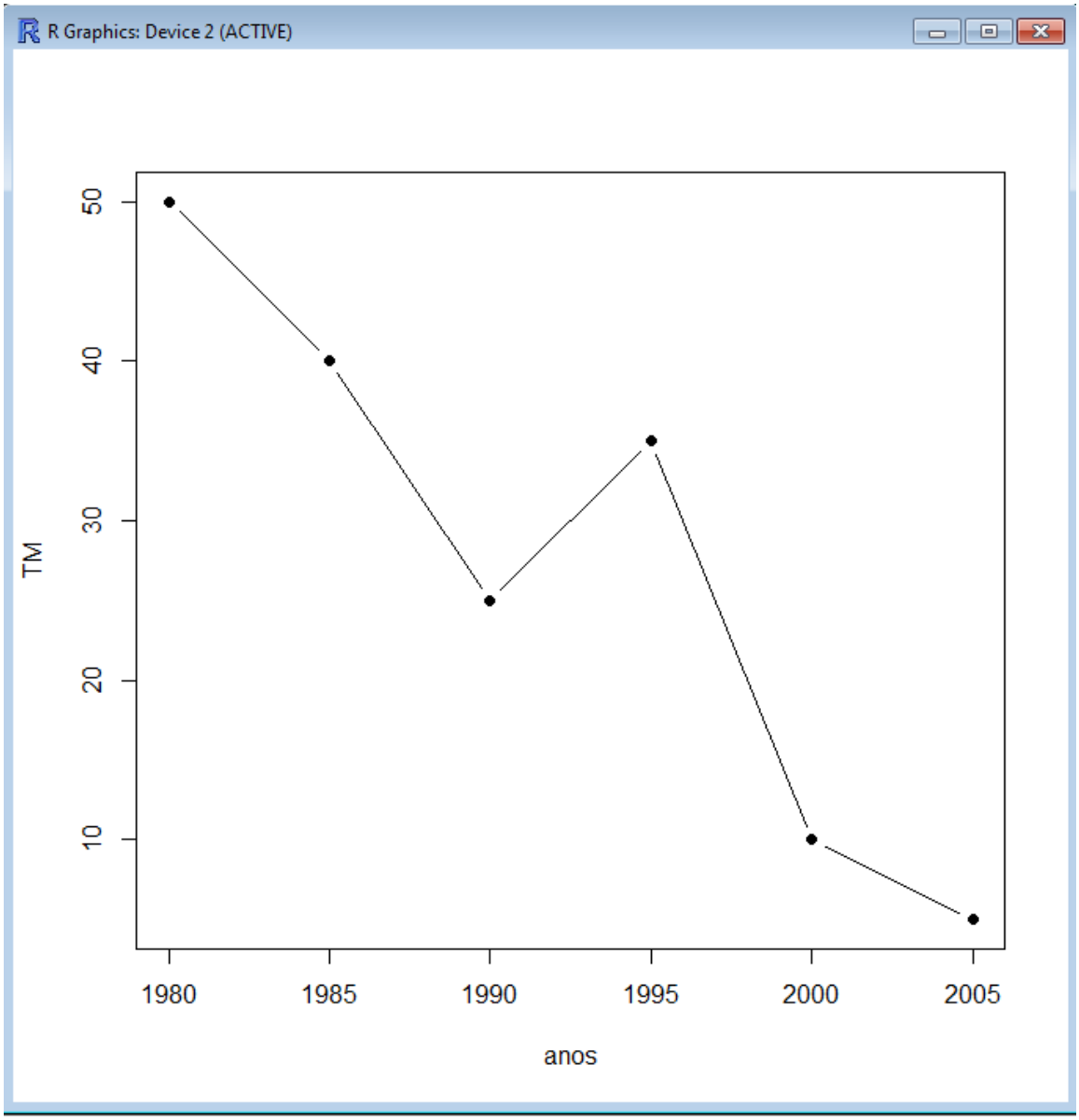


Agora vá em “Gráficos > Grafico de linha”

Selecione as variáveis desejadas, que nesse exemplo são “TM” e “anos”



Clique em ok e então parecerá o seguinte resultado:

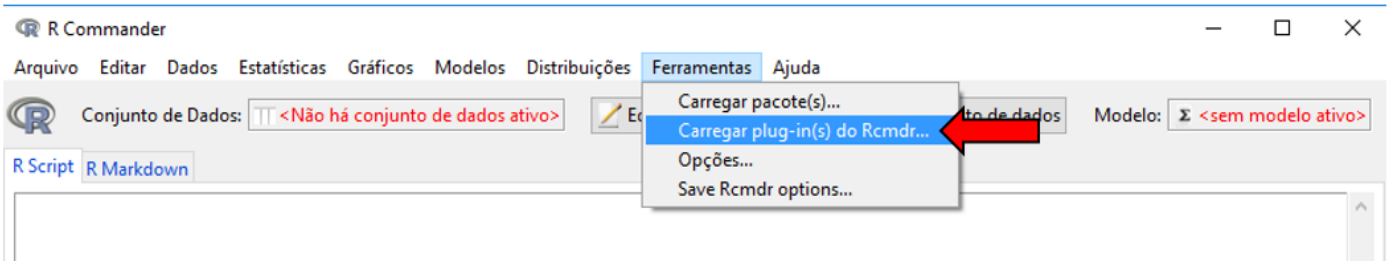


Carregando o Plugin EZR

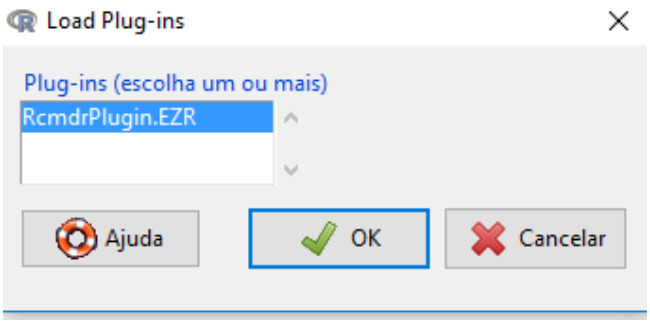
O plug-in “Easy R” EZR realiza diversas análises de forma mais fácil e com mais possibilidade de personalização. Além disso, permite a realização de comandos específicos à epidemiologia. O EZR modificará as opções de menu do RCommander. Antes de carregá-lo salve seu script e área de trabalho.

Após carregar o ambiente do R Commander, vamos carregar o “RcmdrPlugin.EZR” no seguinte caminho:

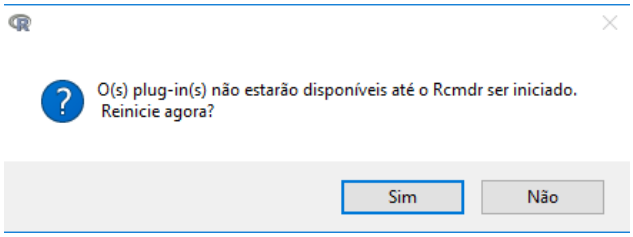
Ferramentas>Carregar plug-in(s) do Rcmdr..



Escolha o plugin desejado, no caso RcmdrPlugin.EZR e clique em OK



Aparecerá uma mensagem dizendo que o R Commander será reiniciado clique em “Sim”

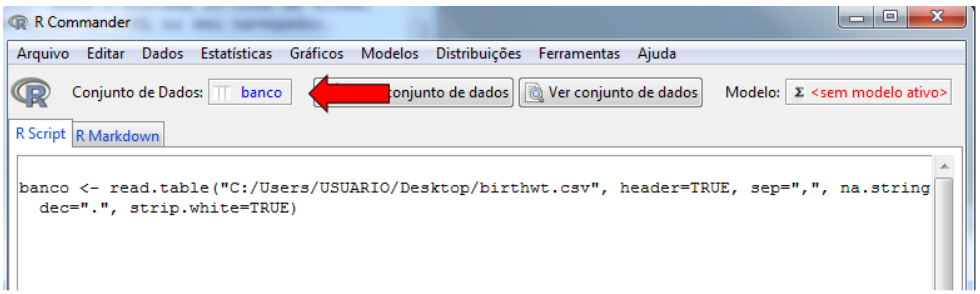


Então o R Commander irá reinicializar e abrir automaticamente. Clique no conjunto de dados ativo e verifique que todos os bancos de dados continuam na memória. Entretanto, o script é perdido, caso não esteja salvo previamente.

Dados para o R Commander no RcmdrPlugin.EZR

Note que antes de carregar ou criar um banco de dados, no campo conjunto de dados, está indicando a inexistência de qualquer informação ou dados a serem estudados.

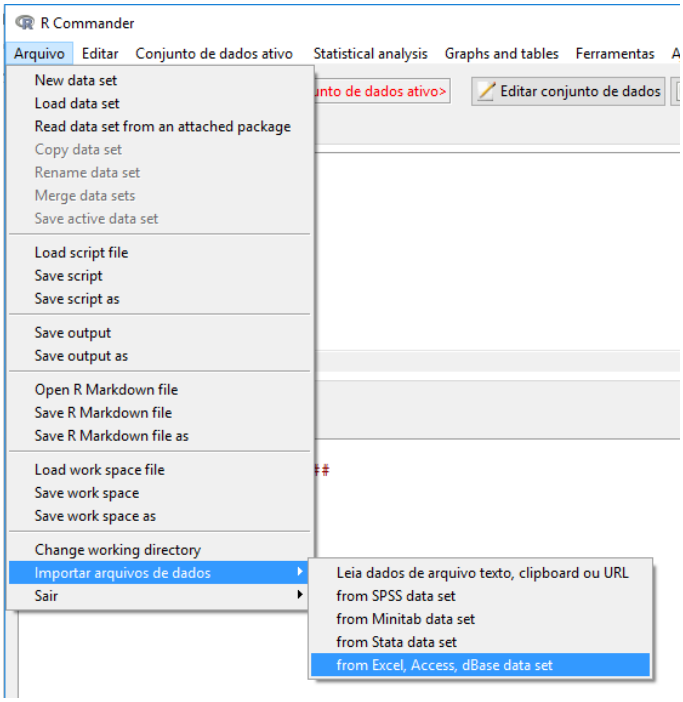
Esse status munda ao carregar/ou criar um banco conforme indicado no exemplo abaixo, em que o status do campo “Conjunto de Dados” mudou ao ser carregado um arquivo de dados nomeado como “banco”.



Para carregar um novo banco de dados, basta utilizar o menu do RCommander:

Arquivo> Importar arquivos de dados

E escolha o tipo de arquivo em que deseja importar.

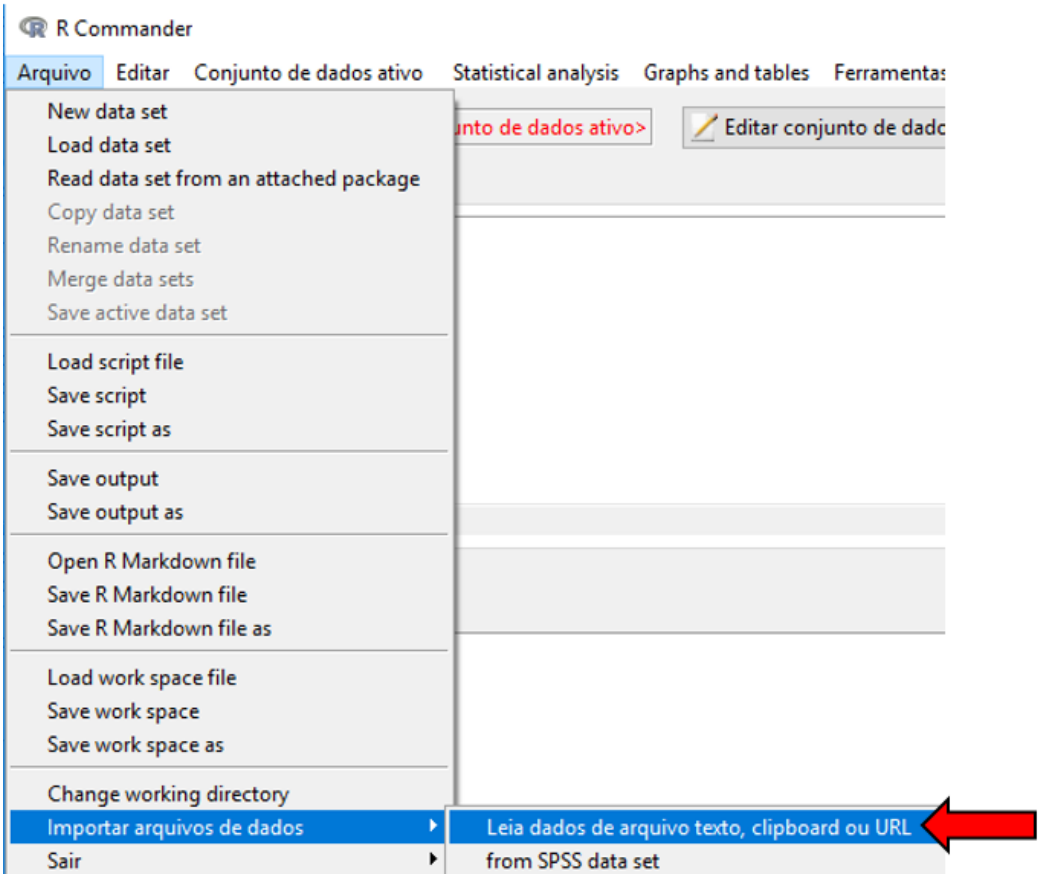

R Commander
Arquivo
Editar
Conjunto de dados ativo
Statistical analysis
Graphs and tables
Ferramentas
Editar conjunto de dados
New data set
Load data set
Read data set from an attached package
Copy data set
Rename data set
Merge data sets
Save active data set
Load script file
Save script
Save script as
Save output
Save output as
Open R Markdown file
Save R Markdown file
Save R Markdown file as
Load work space file
Save work space
Save work space as
Change working directory
Importar arquivos de dados
Sair
Leia dados de arquivo texto, clipboard ou URL
from SPSS data set
from Minitab data set
from Stata data set
from Excel, Access, dBase data set

Análise Descritiva no EZR

Vamos agora realizar uma análise descritiva do banco de dados *"birthwt.csv"* utilizando as ferramentas do EZR.

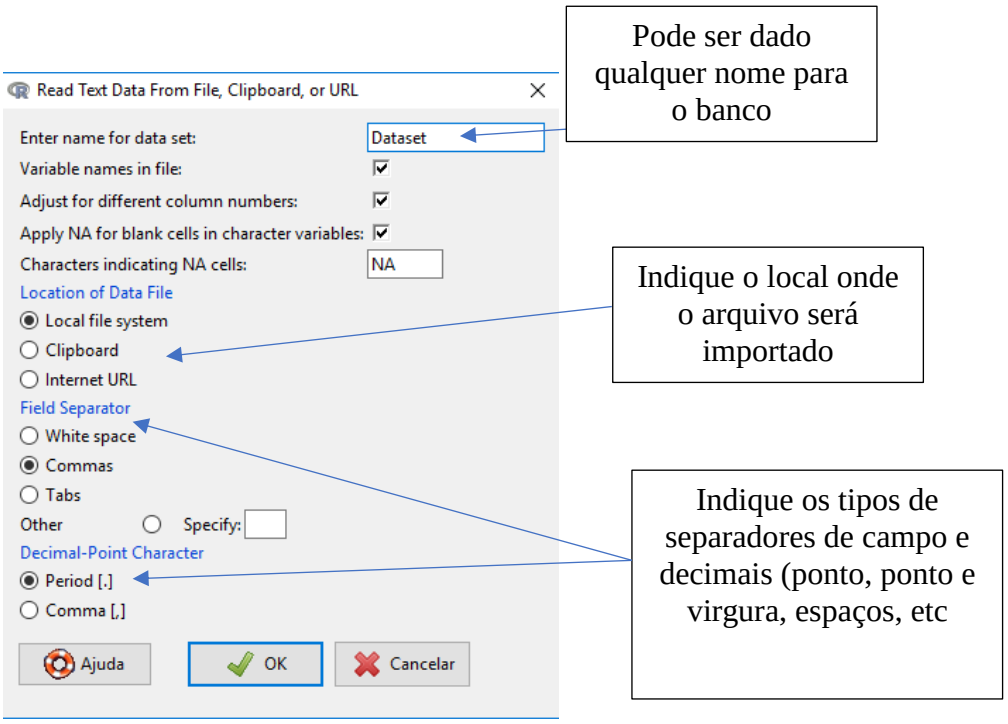
Primeiro vamos carregar o banco de dados para a nossa análise no EZR. Faça o seguinte caminho:

Arquivo > Importar arquivos de dados > Leia dados de arquivos texto, clipboard ou URL



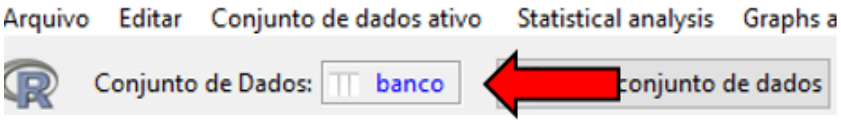
OBS: No exemplo acima esse caminho é utilizado para importação de arquivos do tipo csv, porém há outras extensões que pode ser importadas pelo mesmo caminho.

No caso do nosso exemplo aparecerá as seguintes opções para o arquivo:



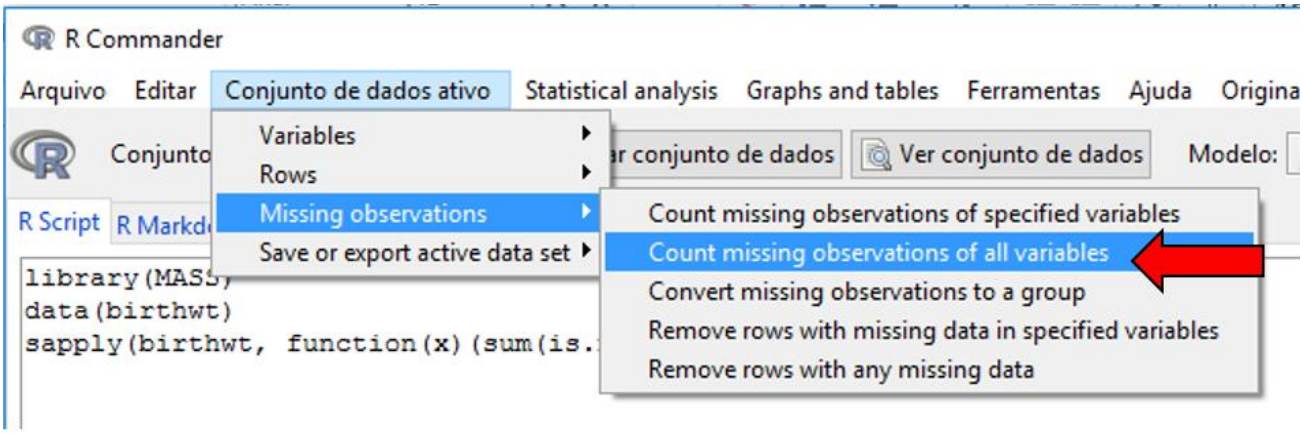
Selecione o nome que deseja dar ao seu banco, os separadores de campo e de decimais.
Após indique o diretório ondo o arquivo está salvo, selecione e clique em ok.

Para o nosso exemplo chamaremos o conjunto de dados de “banco” e após carregamento no o mesmo deve estar ativo no campo “Conjunto de dado: ” conforme figura abaixo.



Verificando se há dados faltantes:

Conjunto de dados ativos > Missing observation > Count missing observation of all varablle

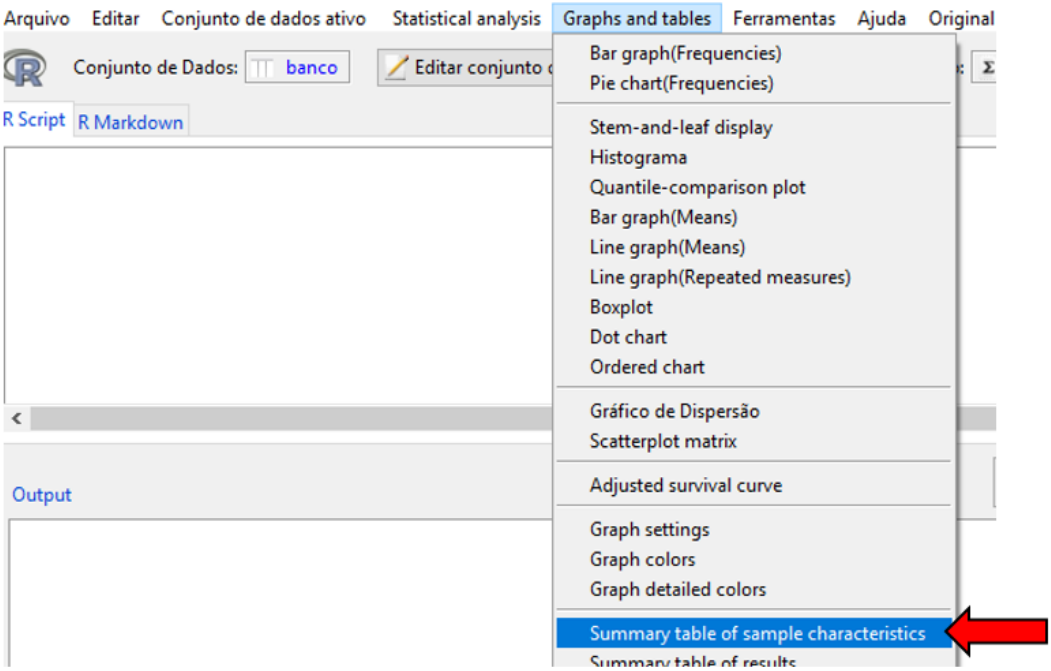


Aparecerá o seguinte resultado contabilizando se há ou não variáveis com dados faltantes (missings)

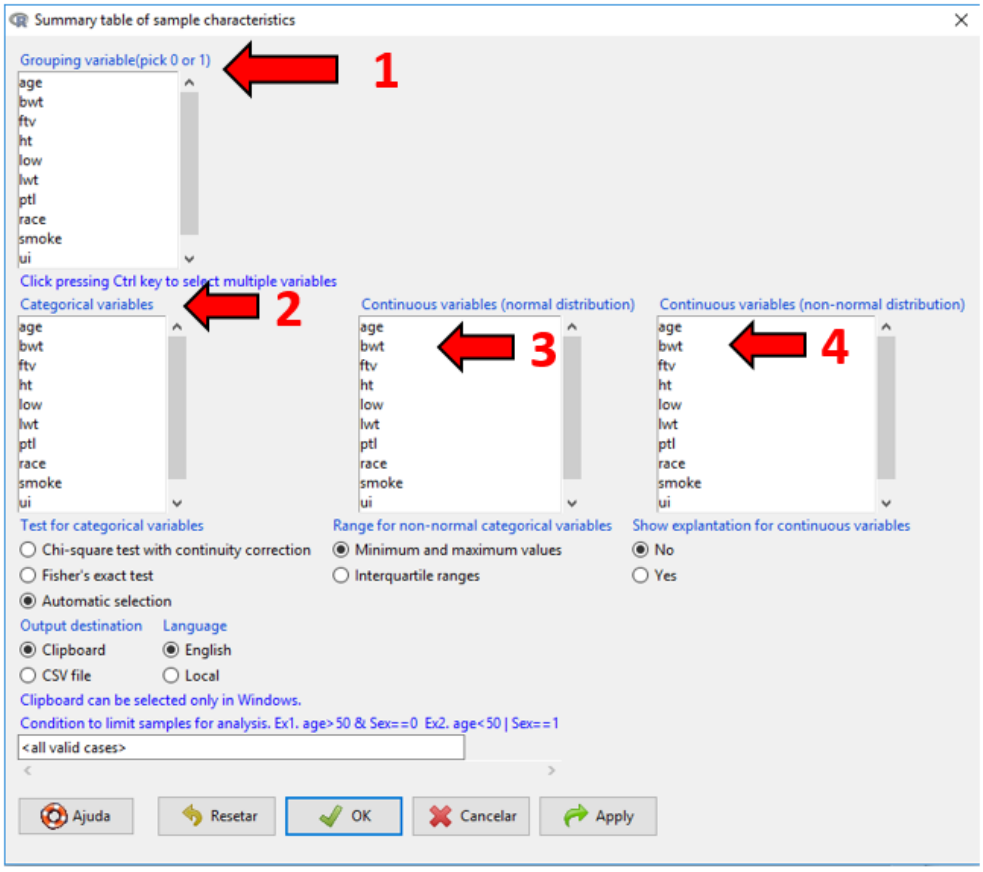
```
Output
> sapply(banco, function(x)(sum(is.na(x)))) # NA counts
                X               low               age               lwt              race             smoke
                0                 0                 0                 0                 0                 0
              ptl                ht                ui               ftv               bwt               cor
                0                 0                 0                 0                 0                 0
     faixaetaria relação.fumo.peso
                0                 0
```

Análise Descritiva

Clique no menu em: Graphs and tables > Summary table sample characteristic



Aparecerá as seguintes opções conforme tela abaixo:



1-Grouping variable: Variáveis de agrupamento são variáveis qualitativas utilizadas para agrupar ou estratificar observações. As variáveis de agrupamento são úteis para resumir ou visualizar dados por grupo. Neste campo, escolhe-se no máximo apenas uma variável.

2- Categorical variables – Para definir as variáveis qualitativas ou categóricas. Podem ser selecionadas uma ou mais variáveis.

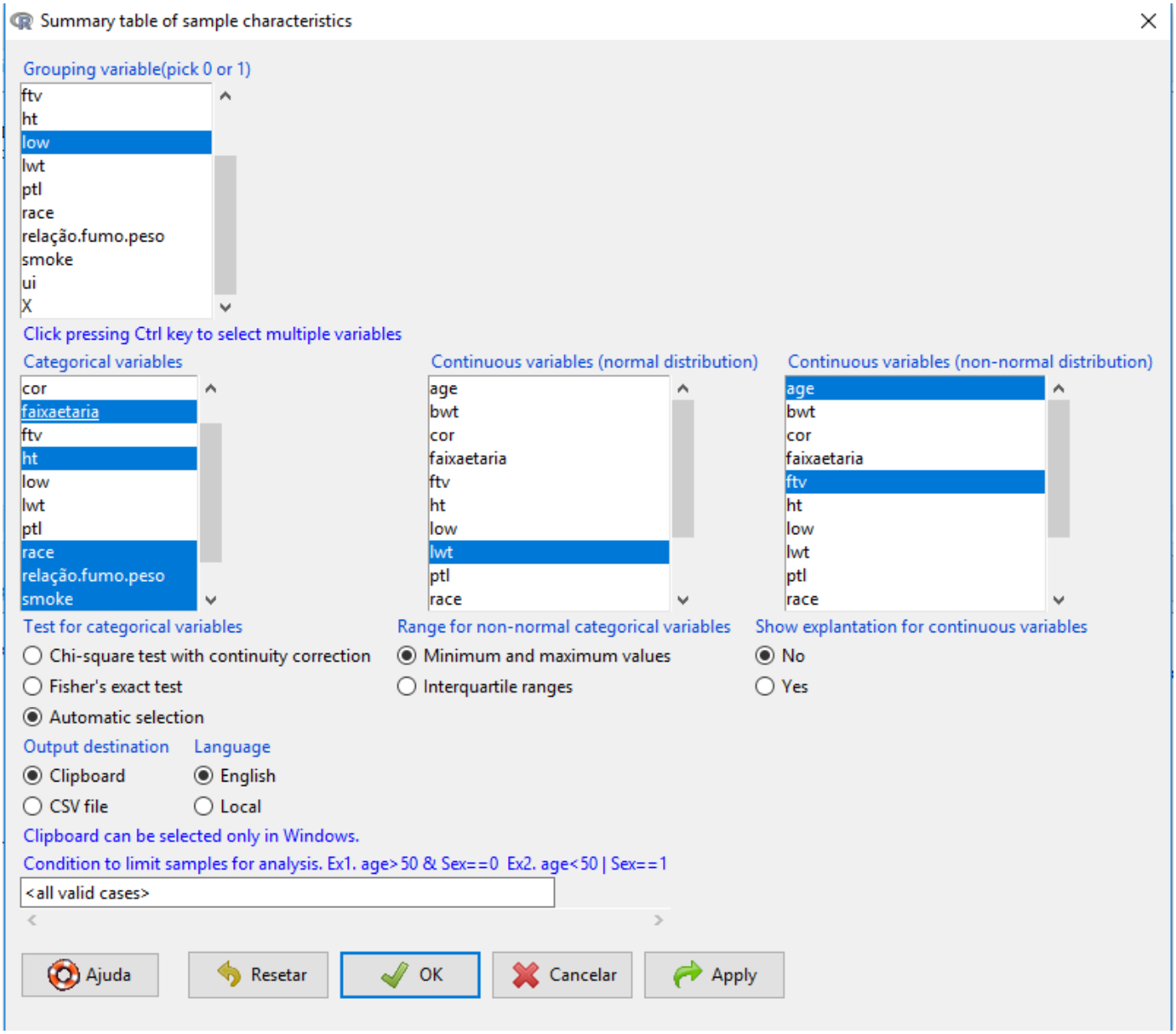
3- Continuous variables (normal distribution) – Local para definir as variáveis quantitativas do tipo contínua com distribuição normal. Podem ser selecionadas uma ou mais variáveis.

3- Continuous variables (non-normal distribution) – Local para selecionar as variáveis tipo qualitativa contínua que não supõe distribuição normal. Pode ser seleciona uma ou mais variáveis.

Atenção: Os campos não pré-definem os tipos de variáveis. Em cada campo são mostradas todas as variáveis sem discriminar seu tipo (categórica, continua com ou sem distribuição normal). O usuário que deve selecionar corretamente as variáveis de forma adequada à cada campo.

Exemplo: Suponha que que estejamos interessados em verificar a relação da variável *low* com as demais variáveis, ou seja, verificar quais os fatores estão associados ao baixo peso ao nascer.

No campo Grouping variable selecione a variável *low* e nos demais campos selecione as variáveis de acordo com seu tipo e clique em OK.



A análise será fará todos os testes, relacionado cada variável ao desfecho low (análises bivariadas). O resultado no output segue em forma de tabela conforme mostrado na figura abaixo.

Output

```
> FinalTable <- rbind(n=row1, FinalTable)

> print(FinalTable[,2:length(FinalTable[1,])], quote=FALSE)
                     Group                  <2.5                >2.5                p.value
                                            59                  130
faixaetaria (%)       Menor de 20           15 (25.4)            36 (27.7)           0.162
                     21 a 29                40 (67.8)            71 (54.6)
                     31 a 39                 4 ( 6.8)            22 (16.9)
                     40  ou mais             0 ( 0.0)             1 ( 0.8)
ht (%)               Não                    52 (88.1)           125 (96.2)           0.052
                     Sim                     7 (11.9)             5 ( 3.8)
race (%)             Branco                 23 (39.0)            73 (56.2)           0.079
                     Outros                 25 (42.4)            42 (32.3)
                     Preto                  11 (18.6)            15 (11.5)
relação.fumo.peso (%)  Não fuma e baixo peso 29 (49.2)             0 ( 0.0)           <0.001
                     Fuma e baixo peso      30 (50.8)             0 ( 0.0)
                     Fuma e peso normal      0 ( 0.0)            44 (33.8)
                     Não fuma e peso normal  0 ( 0.0)            86 (66.2)
smoke (%)            Não                    29 (49.2)            86 (66.2)           0.036
                     Sim                    30 (50.8)            44 (33.8)
lwt                                         122.14 (26.56)      133.30 (31.72)       0.020
age                                         22.00 [14.00, 34.00] 23.00 [14.00, 45.00] 0.247
ftv                                          0.00 [0.00, 4.00]    1.00 [0.00, 6.00]  0.238

> FinalTable <- rbind(row0, FinalTable)

> row0 <- rep("", length(colnames(FinalTable)))
```

O resultado também pode ser salvo para um arquivo csv, basta selecionar a opção “CSV file” antes de clicar em OK. Indique o diretório e o arquivo será salva automaticamente no local desejado.