

INTRODUÇÃO AO R COMMANDER: UMA ABORDAGEM COMPUTACIONAL VOLTADA AO ENSINO DE ESTATÍSTICA

**Calvin Rodrigues
Semana da Estatística
2020**



R Commander

R Commander é um pacote desenvolvido por John Fox que nos permite operar o R de maneira simplificada através de um menu de navegação, não havendo necessidade de conhecer e digitar comandos.

Com ele é possível obter, por exemplo, os principais resumos numéricos (média, mediana, desvio padrão, quantis, etc) além de resumos tabulares e gráficos, sem necessidade de conhecimento prévio sobre programação ou estatística.



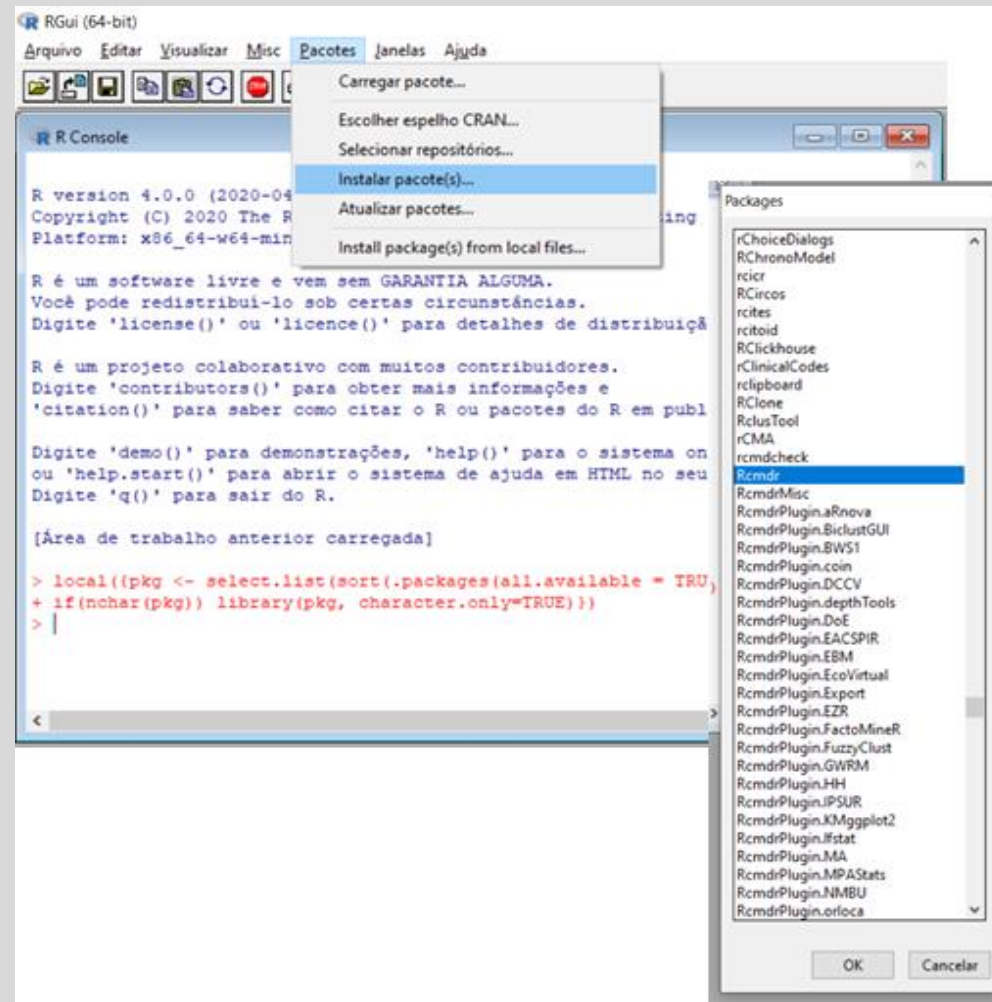
Conteúdo

- Instruções de instalação do R Commander (instruções para instalação do R em vídeo presente no [sympla](#) do minicurso)
- Conjunto de dados presentes no R
- Comandos aritméticos e matemáticos básicos
- Principais resumos numéricos (médias, mediana, desvio-padrão, quantis, correlação, etc.)
- Resumos tabulares de variáveis qualitativa e quantitativas. Tabelas de contingência
- Principais resumos gráficos (histograma, gráfico de barras, gráfico de setores, gráfico de dispersão, etc.)
- Instruções para importação de conjuntos de dados.
- Exercício dados Saeb 1999



Instalação e Carregamento

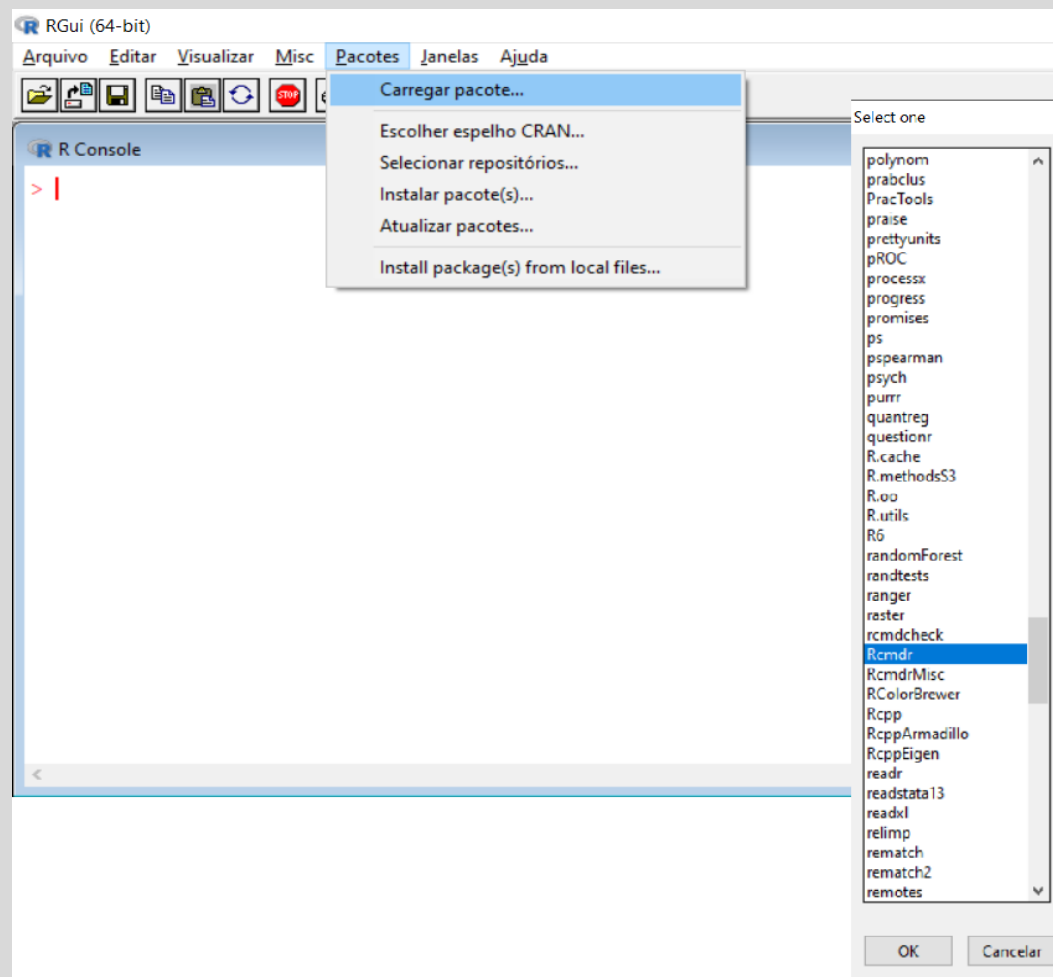
Para começarmos a usar o R Commander, precisamos instalar e carregar o respectivo pacote, chamado Rcmdr, para isso, temos na barra superior do R a opção *Pacotes*.





Instalação e Carregamento

Após a instalação precisaremos carregá-lo, para que a interface apareça, usando a mesma opção *Pacotes*. Pode ser necessária a instalação de outros pacotes ao tentar carregar o Rcmdr, nesse caso o R pedirá permissão para instalá-los automaticamente.





Instalação e Carregamento

R Commander

Arquivo Editar Dados Estatísticas Gráficos Modelos Distribuições Ferramentas Ajuda

Conjunto de Dados: Editar conjunto de dados Ver conjunto de dados Modelo:

R Script R Markdown

Output

Mensagens

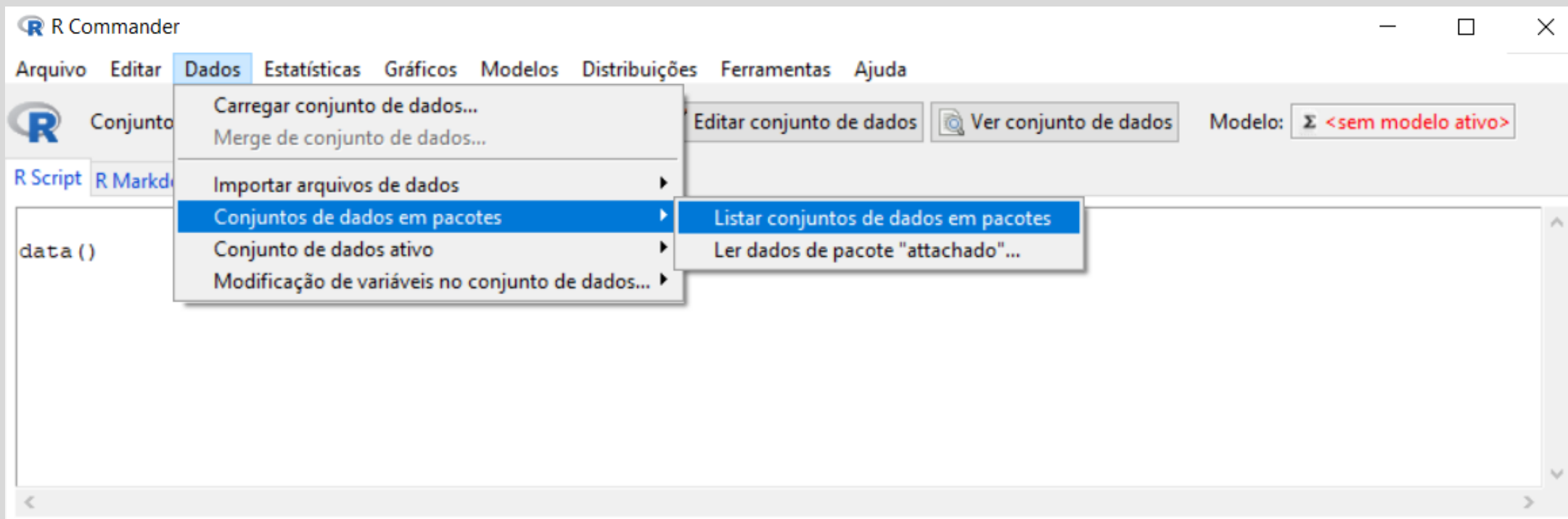
```
[2] AVISO: The Windows version of the R Commander works best under
RGui with the single-document interface (SDI); see ?Commander.
```



Conjuntos de Dados

Para começarmos, precisamos escolher um conjunto de dados, nesse primeiro momento vamos usar um conjunto de dados já presente no R (sem importação ou instalação de novos pacotes).

Podemos ver todos os conjuntos de dados disponíveis na aba *Dados* da interface do R Commander.





Conjuntos de Dados

Agora que sabemos quais bancos de dados temos disponíveis, além do que cada um representa, vamos importar o banco *Salaries* do pacote *mtcars* (instalado automaticamente junto com o Rcmdr), que representa 9 meses de salário de 397 professores de uma determinada faculdade dos Estados Unidos. As variáveis são:

- rank: Prof – professor titular;
AssocProf – professor associado;
AsstProf – professor assistente
- discipline: A – departamentos “teórico”;
B – departamentos “aplicados”
- yrs.since.phd: anos desde a formação PHD
- yrs.service: anos de serviço na faculdade
- sex: sexo do professor
- salary: soma dos primeiros 9 meses de salário em 2008 (dólares)

	rank	discipline	yrs.since.phd	yrs.service	sex	salary
1	Prof	B	19	18	Male	139750
2	Prof	B	20	16	Male	173200
3	AsstProf	B	4	3	Male	79750
4	Prof	B	45	39	Male	115000
5	Prof	B	40	41	Male	141500
6	AssocProf	B	6	6	Male	97000
7	Prof	B	30	23	Male	175000
8	Prof	B	45	45	Male	147765
9	Prof	B	21	20	Male	119250
10	Prof	B	18	18	Female	129000
11	AssocProf	B	12	8	Male	119800
12	AsstProf	B	7	2	Male	79800
13	AsstProf	B	1	1	Male	77700
14	AsstProf	B	2	0	Male	78000
15	Prof	B	20	18	Male	104800
16	Prof	B	12	3	Male	117150
17	Prof	B	19	20	Male	101000
18	Prof	A	38	34	Male	103450
19	Prof	A	37	23	Male	124750
20	Prof	A	39	36	Female	137000
21	Prof	A	31	26	Male	89565
22	Prof	A	36	31	Male	102580
23	Prof	A	34	30	Male	93904
24	Prof	A	24	19	Male	113068
25	AssocProf	A	13	8	Female	74830



Conjuntos de Dados

Precisamos colocar o conjunto de dados em evidência na interface do R Commander, para isso usaremos novamente a opção *Dados*.

The screenshot shows the R Commander interface with the 'Dados' menu open. The menu options are:

- Carregar conjunto de dados...
- Merge de conjunto de dados...
- Importar arquivos de dados
- Conjuntos de dados em pacotes
- Conjunto de dados ativo
- Modificação de variáveis no conjunto de dados...

The 'Conjuntos de dados em pacotes' option is selected, and a sub-menu is open with the following options:

- Listar conjuntos de dados em pacotes
- Ler dados de pacote "attachado"...

The 'Ler dados de pacote "attachado"...' option is selected, and a dialog box titled 'Leia dados do pacote' is open. The dialog box has two columns of data sets:

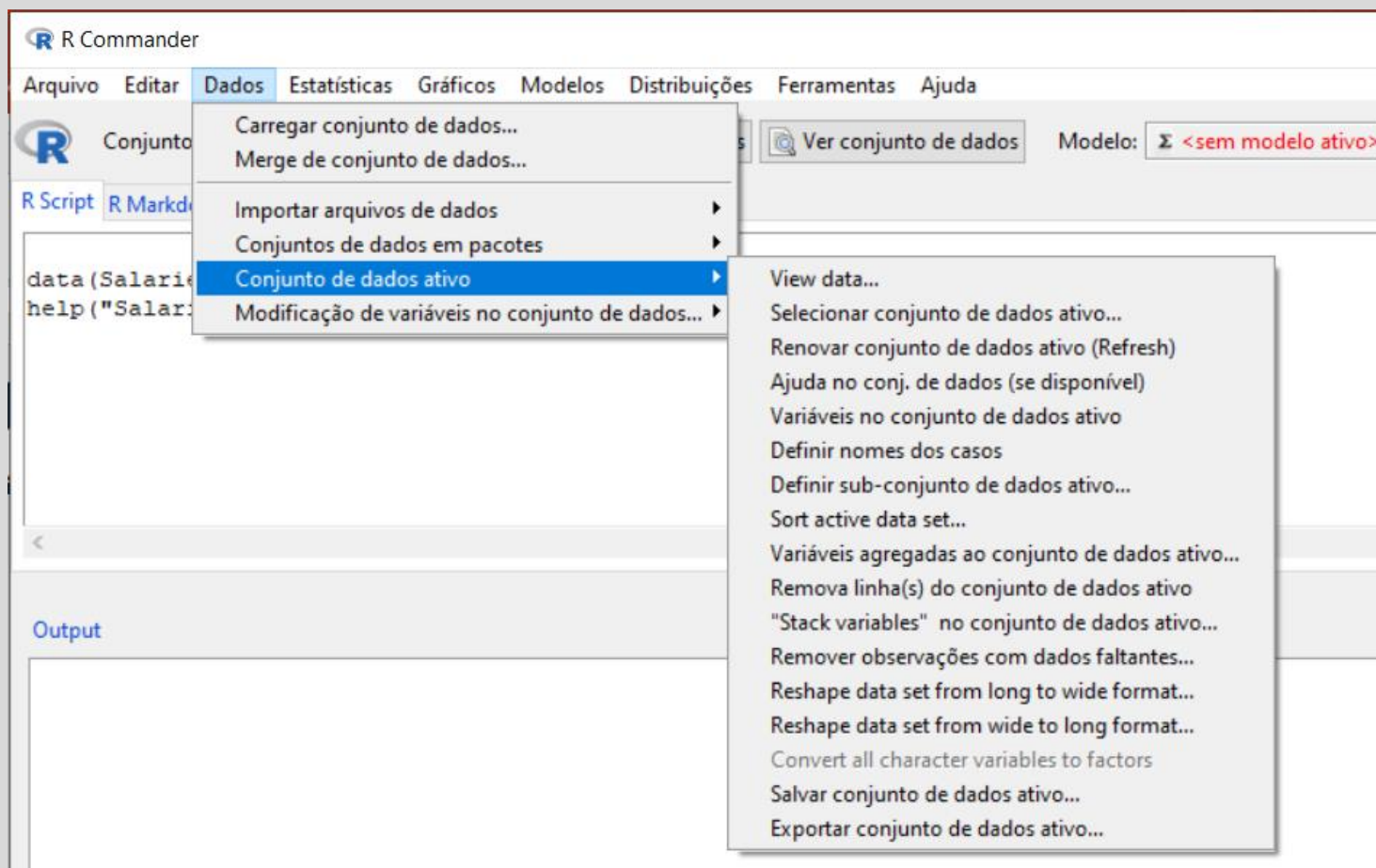
Pacote (clique-duplo para selecionar)	Conjunto de dados (clique-duplo para selecionar)
carData	Sahlins
datasets	Salaries
sandwich	Soils
	States
	TitanicSurvival
	Transact

Below the columns, there is a text input field for 'Defina o nome do conjunto de dados:' and a button for 'Ajuda no conjunto de dados selecionado'. At the bottom, there are three buttons: 'Ajuda', 'OK', and 'Cancelar'.



Conjuntos de Dados

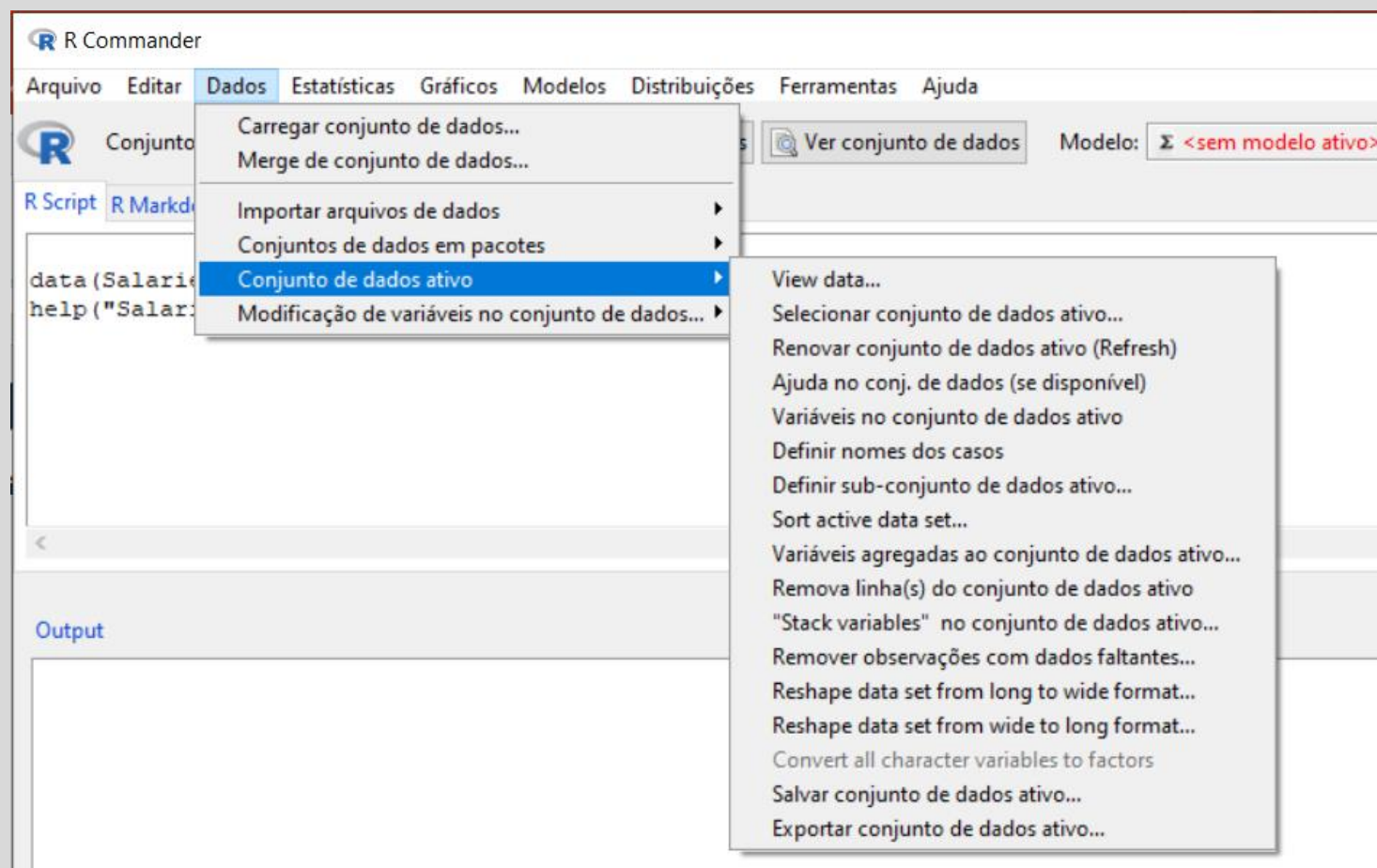
Agora que já temos o conjunto de dados, podemos realizar as primeiras interações, como selecionar quais variáveis queremos ver, abrir a página web que contém a descrição do conjunto, editá-lo, exportá-lo, entre outras opções contidas na parte “Conjunto de dados ativo” da aba *Dados*.





Conjuntos de Dados

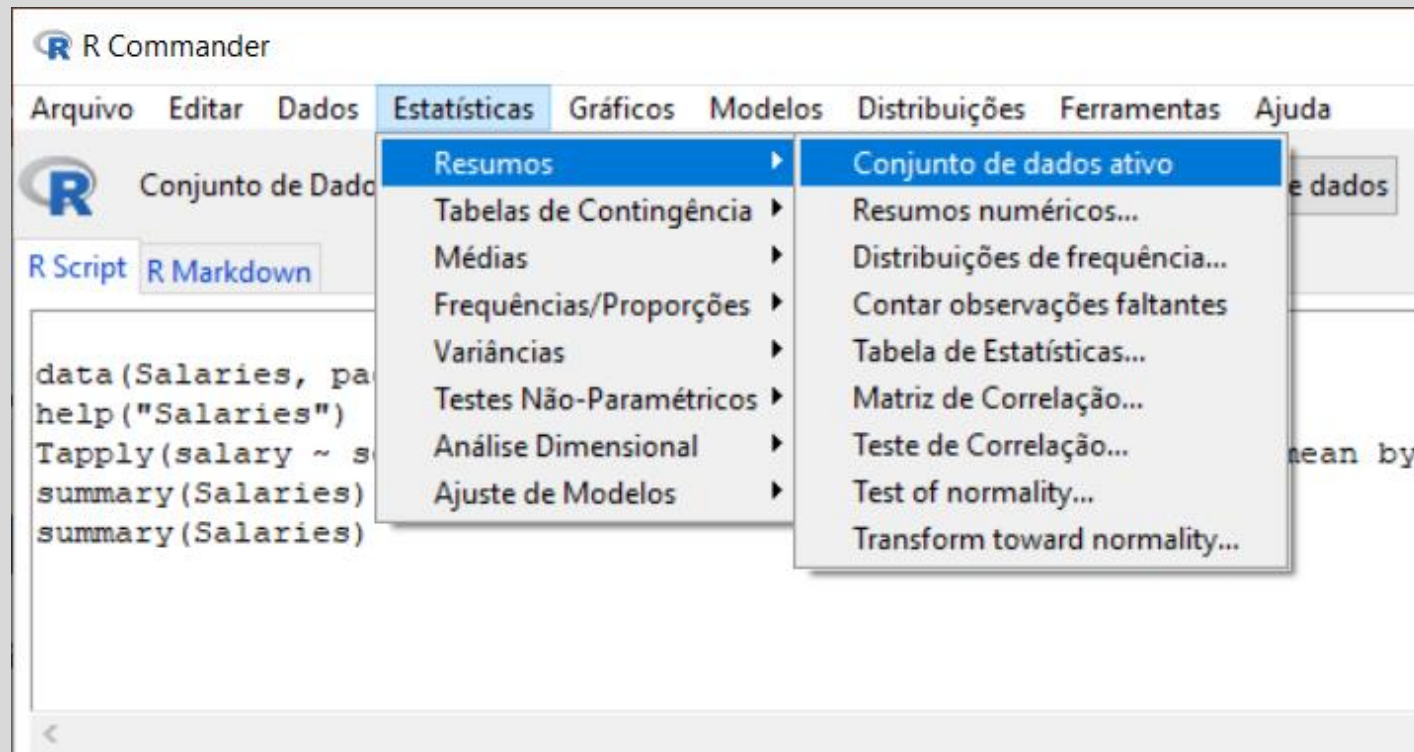
Agora que já temos o conjunto de dados, podemos realizar as primeiras interações, como selecionar quais variáveis queremos ver, abrir a página web que contém a descrição do conjunto, editá-lo, exportá-lo, entre outras opções contidas na parte “Conjunto de dados ativo” da aba *Dados*.





Resumos Numéricos

Com o conjunto de dados ativo, podemos começar a análise estatística dos dados, primeiramente, faremos o resumo do conjunto de dados considerando todas as variáveis (chamado de sumário) usando a opção *Estatísticas*.





Resumos Numéricos

A saída, dada pela parte *Output* da interface nos mostra o mínimo e máximo, primeiro e terceiro quartil (partição de 25 e 75% dos dados), mediana (partição de 50% dos dados) e média nas variáveis quantitativas além do número de observações para cada fator nas variáveis qualitativas.

```
> summary(Salaries)
      rank  discipline yrs.since.phd  yrs.service      sex      salary
AsstProf : 67  A:181      Min.      : 1.00    Min.      : 0.00  Female: 39  Min.      : 57800
AssocProf: 64  B:216     1st Qu.:12.00  1st Qu.: 7.00  Male   :358  1st Qu.: 91000
Prof      :266     Median :21.00  Median :16.00                Median :107300
                                     Mean   :22.31  Mean   :17.61                Mean   :113706
                                     3rd Qu.:32.00  3rd Qu.:27.00                3rd Qu.:134185
                                     Max.   :56.00  Max.   :60.00                Max.   :231545
```

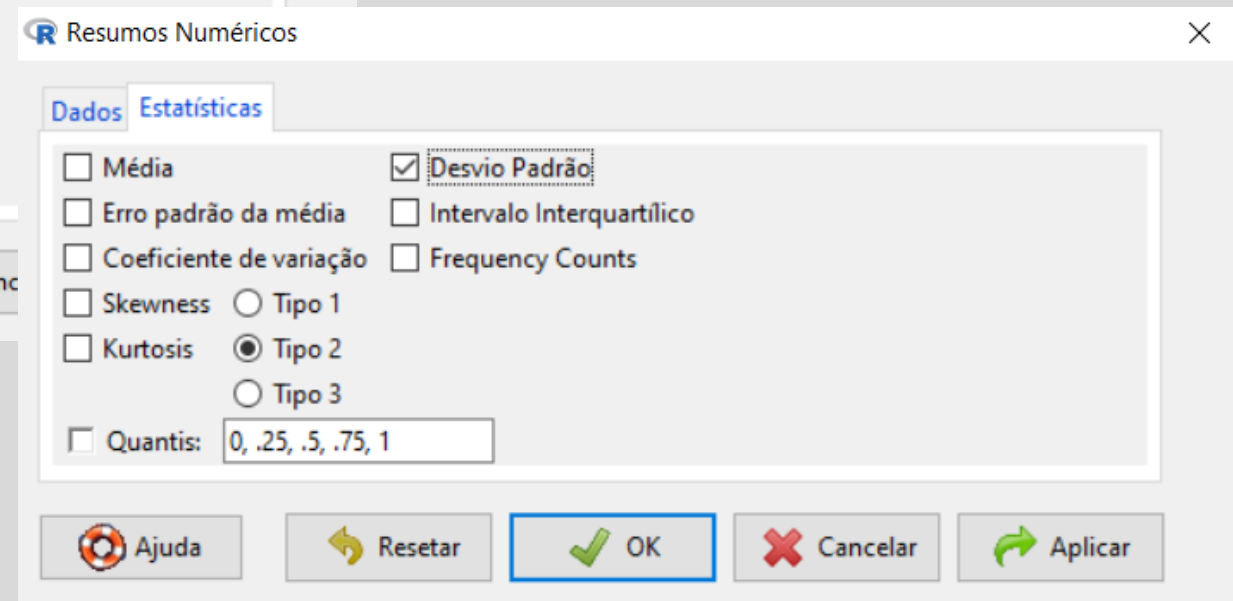
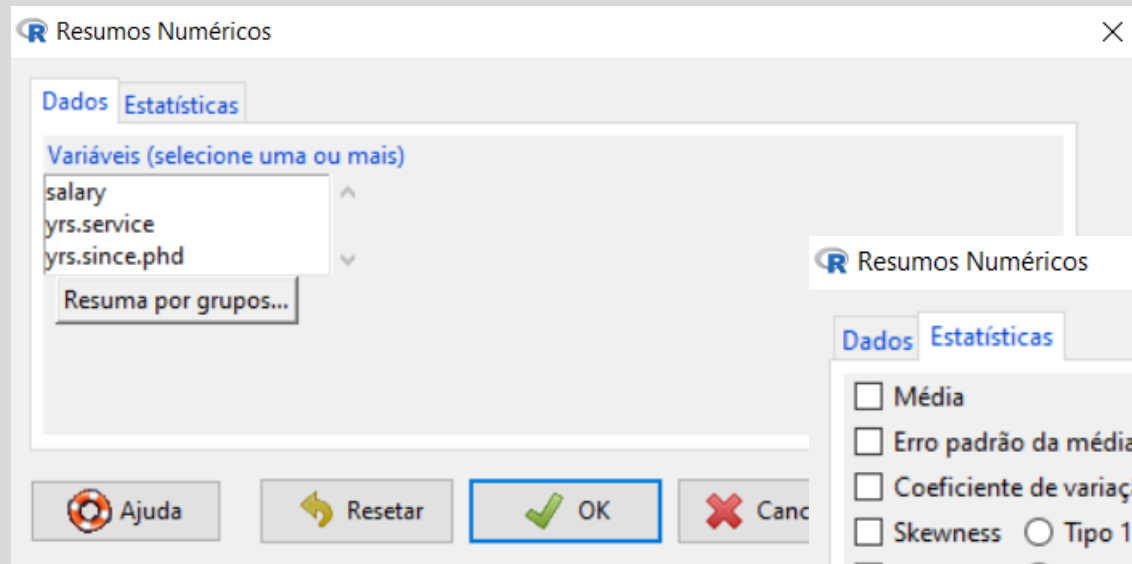
Observamos que, por exemplo, há mais professores em disciplinas aplicadas que em disciplinas teóricas, além de somente 39 dos 397 professores serem do sexo feminino, e que a média dos 9 meses de salário dos professores é \$113706 e que o maior salário acumulado é de \$231545. Através do primeiro quartil temos a informação de que 25% dos professores receberam menos de \$91000 em 9 meses.



Resumos Numéricos

Falta ainda analisar a variação dos valores nas variáveis numéricas, ainda em *Estatísticas - Resumos* temos a opção “resumos numéricos” que só aceita variáveis quantitativas e permite resumo por grupos, porém, nesse primeiro momento vamos analisar as variáveis separadamente.

Vamos usá-la para calcular o desvio padrão (raiz quadrada da variância) da variável salário.





Resumos Numéricos

Agora que já sabemos como calcular os principais resumos numéricos para cada variável individualmente, vamos analisar as variáveis em conjunto, começando pela matriz de correlação do conjunto de dados, que mede o grau de associação entre duas variáveis quantitativas (entre -1 e 1).

Essa matriz pode ser calculada em *Estatísticas – Resumos – Matriz de correlação*

The image shows a dialog box titled "Matriz de correlação" (Correlation Matrix) from the R software interface. The dialog box has a white background and a grey border. At the top left, there is an R logo and the title "Matriz de correlação". At the top right, there is a close button (X). The main content area is divided into several sections:

- Variáveis (escolha 2 ou mais)**: A list box containing three variables: "salary", "yrs.service", and "yrs.since.phd". The "yrs.service" variable is currently selected and highlighted in blue.
- Tipos de Correlações**: Three radio button options: "Produto-momento de Pearson" (selected), "Spearman (rank-order)", and "Parcial".
- Observações para Usar**: Two radio button options: "Observações completas" (selected) and "Observações pareadas completas".
- P-valores pareados**: A checkbox that is currently unchecked.

At the bottom of the dialog box, there are five buttons: "Ajuda" (Help), "Resetar" (Reset), "OK" (highlighted with a blue border), "Cancelar" (Cancel), and "Aplicar" (Apply).



Resumos Numéricos

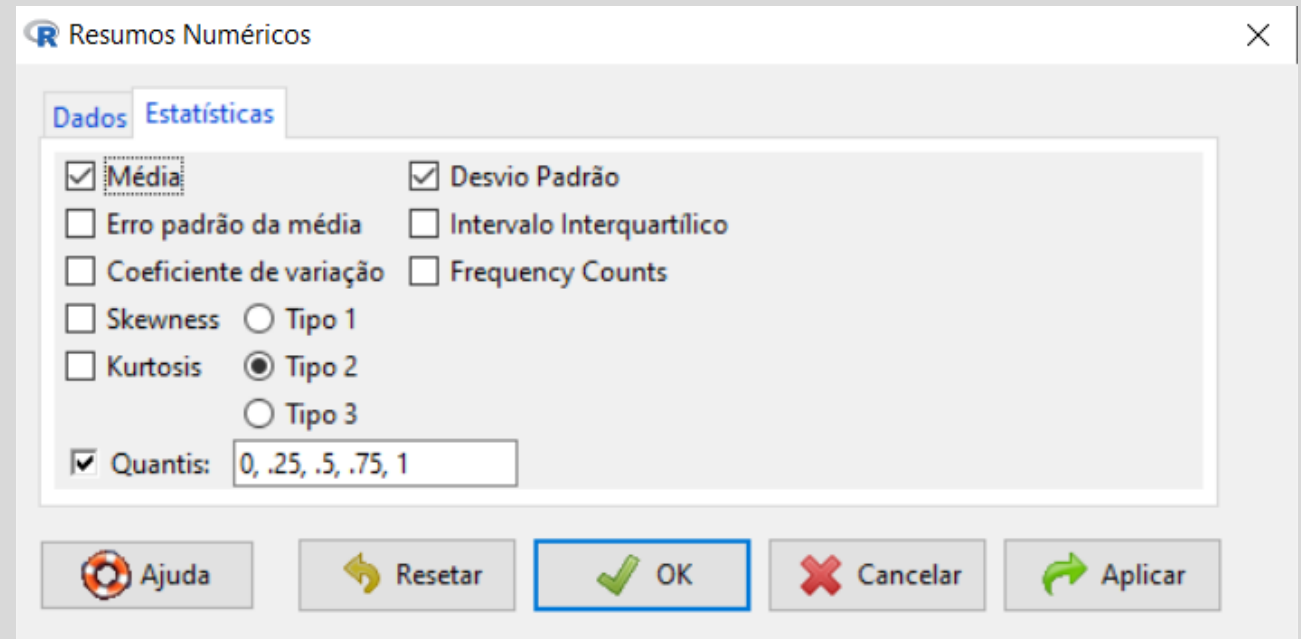
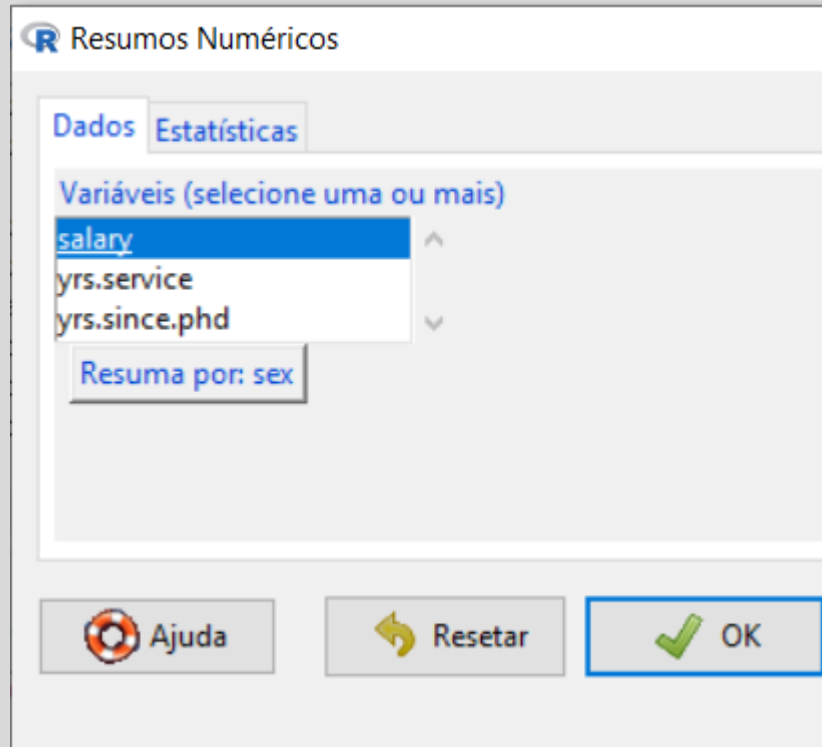
Podemos ver que as variáveis anos de serviço e anos desde a formação PHD são fortemente e positivamente correlacionadas (valor próximo de 1), a variável salário não tem forte associação (valor maior que 0,70) com nenhuma outra variável quantitativa nesse conjunto de dados.

```
> cor(Salaries[,c("salary", "yrs.service", "yrs.since.phd")], use="complete")
          salary yrs.service yrs.since.phd
salary      1.0000000  0.3347447  0.4192311
yrs.service 0.3347447  1.0000000  0.9096491
yrs.since.phd 0.4192311  0.9096491  1.0000000
```




Resumos Numéricos

Como mencionado anteriormente, em *resumos numéricos* podemos separar as variáveis quantitativas por grupos (dados pelas variáveis qualitativas), vamos então obter os resumos numéricos (média, mediana, desvio padrão e quartis) dos salários por sexo.





Resumos Numéricos

Como mencionado anteriormente, em *resumos numéricos* podemos separar as variáveis quantitativas por grupos (dados pelas variáveis qualitativas), vamos então obter os resumos numéricos (média, mediana, desvio padrão e quartis) dos salários por sexo.

```
> numSummary(Salaries[, "salary", drop=FALSE], groups=Salaries$sex, statistics=c("mean", "sd", "quantiles"),
+   quantiles=c(0, .25, .5, .75, 1))
      mean      sd   0%   25%   50%   75%  100% salary:n
Female 101002.4 25952.13 62884 77250 103750 117002.5 161101      39
Male   115090.4 30436.93 57800 92000 108043 134863.8 231545     358
```

Vemos que a média dos salários entre os sexos diverge em cerca de \$14000 e que as mulheres têm valores de salário acumulado em 9 meses mais próximos da média que os homens (desvio padrão menor).

E, como a média é menor, vemos que os quartis também apresentam valores menores para o sexo feminino.



Resumos Numéricos

Podemos também analisar as duas outras variáveis quantitativas pelos sexos.

R Resumos Numéricos

Dados Estatísticas

Variáveis (selecione uma ou mais)

- salary
- yrs.service
- yrs.since.phd

Resuma por: sex

Ajuda Resetar OK

```
Variable: yrs.service
      mean      sd 0% 25% 50% 75% 100%  n
Female 11.56410  8.813252  0  4  10 17.5  36 39
Male   18.27374 13.226234  0  7  18 27.0  60 358
```

```
Variable: yrs.since.phd
      mean      sd 0% 25% 50% 75% 100%  n
Female 16.51282  9.784176  2  10  17 23.5  39 39
Male   22.94693 13.036470  1  12  22 33.0  56 358
```



Resumos Numéricos

Resumindo pela variável disciplina:

```
Variable: salary
      mean      sd    0%    25%    50%    75%   100%   n
A 108548.4 30538.15 57800 83000.00 104350.0 125192.0 205500 181
B 118028.7 29459.14 67559 94905.25 113018.5 139836.5 231545 216

Variable: yrs.service
      mean      sd 0%  25% 50% 75% 100%   n
A 19.95028 13.67816 0 8.00 19 30 57 181
B 15.65741 12.10317 0 5.75 14 23 60 216

Variable: yrs.since.phd
      mean      sd 0% 25% 50% 75% 100%   n
A 25.38122 13.11799 2 14 27.0 36 56 181
B 19.74537 12.13547 1 10 18.5 28 56 216
```



Resumos Numéricos

Resumindo pela variável rank:

```
Variable: salary
      mean      sd    0%    25%    50%    75%    100%    n
AsstProf  80775.99  8174.113 63100  74000.0  79800.0  88597.5  97032  67
AssocProf 93876.44 13831.700 62884  82475.0  95626.5 104226.2 126431  64
Prof      126772.11 27718.675 57800 105975.2 123321.5 145080.5 231545 266

Variable: yrs.service
      mean      sd 0% 25% 50% 75% 100%    n
AsstProf  2.373134  1.495811  0  1  3  3  6  67
AssocProf 11.953125 10.100180  1  7  8 11 53  64
Prof      22.815789 11.590493  0 15 21 30 60 266

Variable: yrs.since.phd
      mean      sd 0% 25% 50% 75% 100%    n
AsstProf  5.104478  2.541381  1  3.5  4  7.00  11  67
AssocProf 15.453125  9.652584  6 10.0 12 17.25  49  64
Prof      28.300752 10.108830 11 20.0 28 36.75  56 266
```



Resumos Tabulares

Como primeiro método, e mais simples, de criação de tabelas temos as tabelas de frequência das variáveis qualitativas, que são obtidas a partir das opções *Estatísticas – Resumos – Distribuições de frequência*, podemos selecionar e obter para as três variáveis de uma vez.

```
counts:
discipline
  A  B
181 216

percentages:
discipline
  A  B
45.59 54.41
```

```
counts:
rank
  AsstProf AssocProf   Prof
         67         64    266

percentages:
rank
  AsstProf AssocProf   Prof
    16.88    16.12    67.00
```

```
counts:
sex
Female  Male
     39   358

percentages:
sex
Female  Male
    9.82 90.18
```



Resumos Tabulares

Já para as variáveis quantitativas o caminho é diferente, faremos o mesmo que anteriormente para os resumos numéricos, porém, selecionaremos a opção *Frequency Counts*, vemos, como exemplo, as frequências da variável anos de serviço na faculdade.

```
> binnedCounts(Salaries[, "yrs.service", drop=FALSE])
Binned distribution of yrs.service
      Count Percent
[0, 5]      82  20.65
(5, 10]     73  18.39
(10, 15]    37   9.32
(15, 20]    58  14.61
(20, 25]    40  10.08
(25, 30]    37   9.32
(30, 35]    22   5.54
(35, 40]    27   6.80
(40, 45]    12   3.02
(45, 50]     5   1.26
(50, 55]     2   0.50
(55, 60]     2   0.50
Total     397  99.99
```

Obs: Em variáveis quantitativas, temos que as frequências são contadas por intervalos de valores

Vemos que quase metade (48,36%) dos professores têm menos de 15 anos de serviço na faculdade



Resumos Tabulares

Para as outras duas variáveis quantitativas (salário e anos desde a formação PHD) temos as seguintes frequências:

Binned distribution of salary		
	Count	Percent
[40000, 60000]	1	0.25
(60000, 80000]	50	12.59
(80000, 100000]	90	22.67
(100000, 120000]	114	28.72
(120000, 140000]	60	15.11
(140000, 160000]	48	12.09
(160000, 180000]	23	5.79
(180000, 200000]	8	2.02
(200000, 220000]	2	0.50
(220000, 240000]	1	0.25
Total	397	99.99

Binned distribution of yrs.since.phd		
	Count	Percent
[0, 5]	42	10.58
(5, 10]	45	11.34
(10, 15]	52	13.10
(15, 20]	54	13.60
(20, 25]	46	11.59
(25, 30]	47	11.84
(30, 35]	35	8.82
(35, 40]	41	10.33
(40, 45]	20	5.04
(45, 50]	10	2.52
(50, 55]	3	0.76
(55, 60]	2	0.50
Total	397	100.02



Resumos Tabulares

Para a análise de variáveis categóricas temos como principal ferramenta a tabela de contingência, que mostra a frequência das observações por duas variáveis. Podemos criá-las através das opções *Estatísticas – Tabelas de Contingência – Tabela de dupla entrada* e então escolhemos quais as duas variáveis qualitativas usaremos na análise.

Primeiramente faremos para as variáveis disciplina e rank

The image displays two screenshots of the R software interface for creating contingency tables. The first screenshot shows the 'Dados' (Data) tab of the 'Tabelas de dupla entrada' (Contingency Tables) dialog box. It features two dropdown menus: 'Variável linha (escolha uma)' (Row variable) and 'Variável coluna (escolha uma)' (Column variable). Both are set to 'discipline'. Below these is a text field for 'Expressão (subset expression)' containing '<todos casos válidos>'. At the bottom are buttons for 'Ajuda' (Help), 'Resetar' (Reset), 'OK', and 'Cancelar' (Cancel). The second screenshot shows the 'Estatísticas' (Statistics) tab of the same dialog box. It has two sections: 'Computar Percentagens' (Compute Percentages) with radio buttons for 'Percentual nas linhas' (Percentages in rows), 'Percentual nas colunas' (Percentages in columns), 'Percentagens do total' (Percentages of total) (which is selected), and 'Sem percentual' (No percentages); and 'Testes de Hipótese' (Hypothesis Tests) with checkboxes for 'Teste de independência de Qui-Quadrado' (Chi-square test of independence), 'Componentes da estatística do Qui-quadrado' (Chi-square test components), 'Apresente frequências esperadas' (Show expected frequencies), and 'Teste exato de Fisher' (Fisher's exact test). At the bottom are buttons for 'Ajuda', 'Resetar', 'OK', 'Cancelar', and 'Aplicar' (Apply).



Resumos Tabulares

Temos como retorno as frequências absolutas e relativas e vemos que a maior diferença em quantidade de rank de professores por disciplina se dá nos professores assistentes, 43 dos 67 lecionam disciplinas com departamentos aplicados.

Vemos que a maior parte dos professores (54,4%) são de departamentos aplicados e 67% dos professores da faculdade são titulares.

```
Frequency table:
      rank
discipline AsstProf AssocProf Prof
      A         24         26  131
      B         43         38  135

Total percentages:
      AsstProf AssocProf Prof Total
A         6.0         6.5  33  45.6
B         10.8         9.6  34  54.4
Total     16.9         16.1  67 100.0
```



Resumos Tabulares

Temos para as outras variáveis qualitativas as tabelas de contingência com rank como variável linha X sexo como variável coluna e disciplina como variável linha X sexo como variável coluna, respectivamente

```
Frequency table:
      sex
rank   Female Male
AsstProf    11   56
AssocProf    10   54
Prof         18  248

Total percentages:
      Female Male Total
AsstProf    2.8 14.1 16.9
AssocProf    2.5 13.6 16.1
Prof         4.5 62.5 67.0
Total       9.8 90.2 100.0
```

```
Frequency table:
      sex
discipline Female Male
A           18  163
B           21  195

Total percentages:
      Female Male Total
A           4.5 41.1 45.6
B           5.3 49.1 54.4
Total       9.8 90.2 100.0
```



Resumos Gráficos

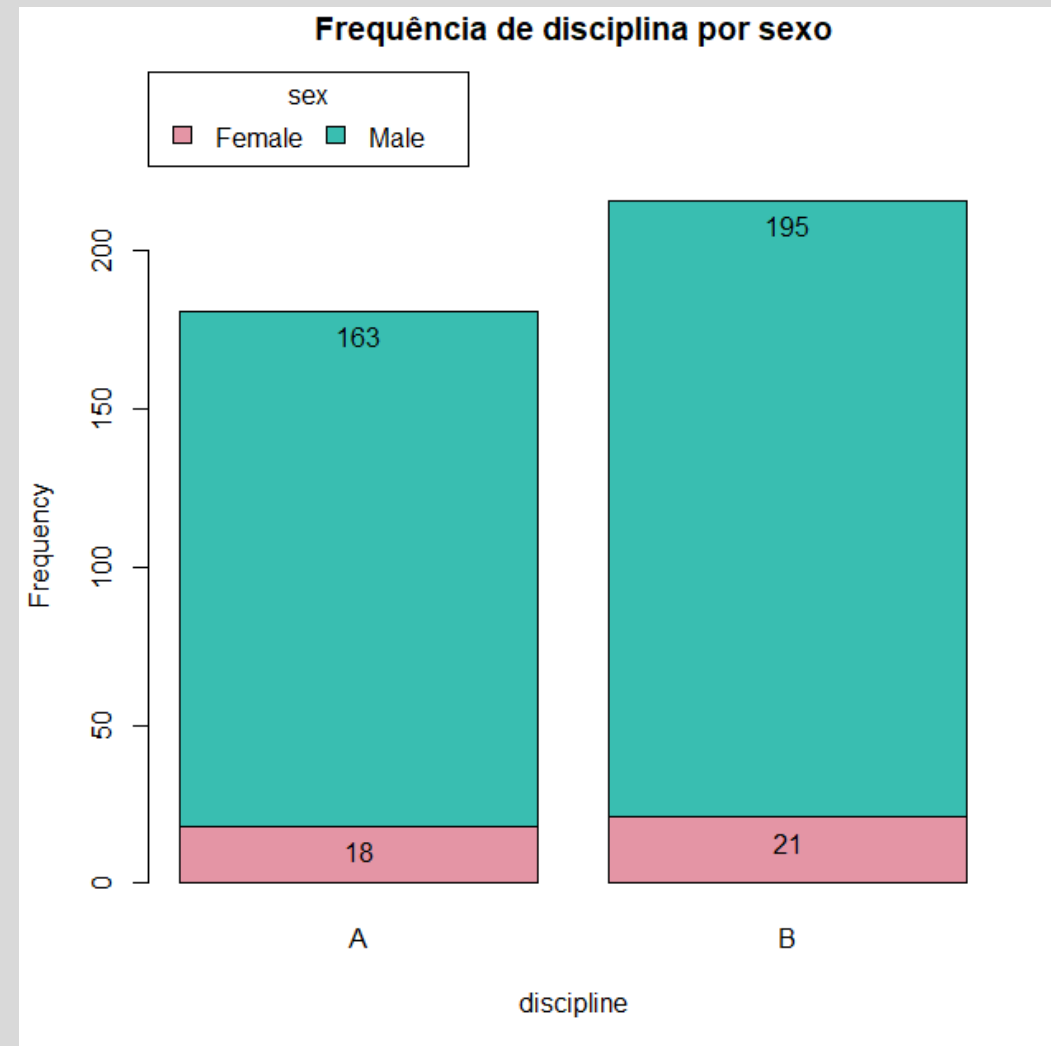
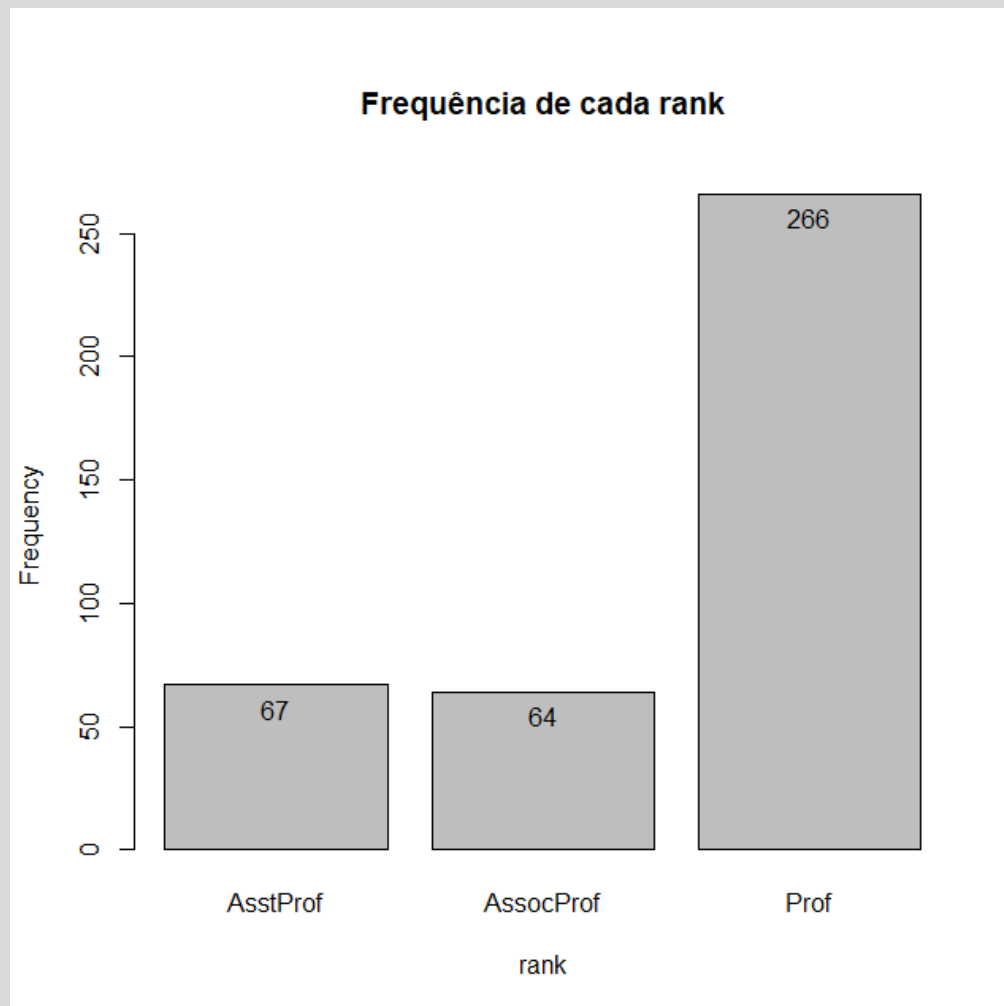
A interface do R Commander contém uma opção inteiramente focada em visualização gráfica de dados na qual podemos fazer desde gráficos simples como gráfico de setores e de pontos até gráficos 3D.

Similarmente ao que já fizemos, os gráficos quando de variáveis quantitativas podem conter todos os valores ou podem ser separados em função de uma variável qualitativa (como salário por sexo, por exemplo).



Resumos Gráficos: Gráfico de Barras

O gráfico de barras é útil quando queremos saber o número de observações para cada possível resposta de uma variável qualitativa. Temos, como exemplo, o gráfico de rank dos professores e o gráfico de disciplinas condicionado ao sexo.

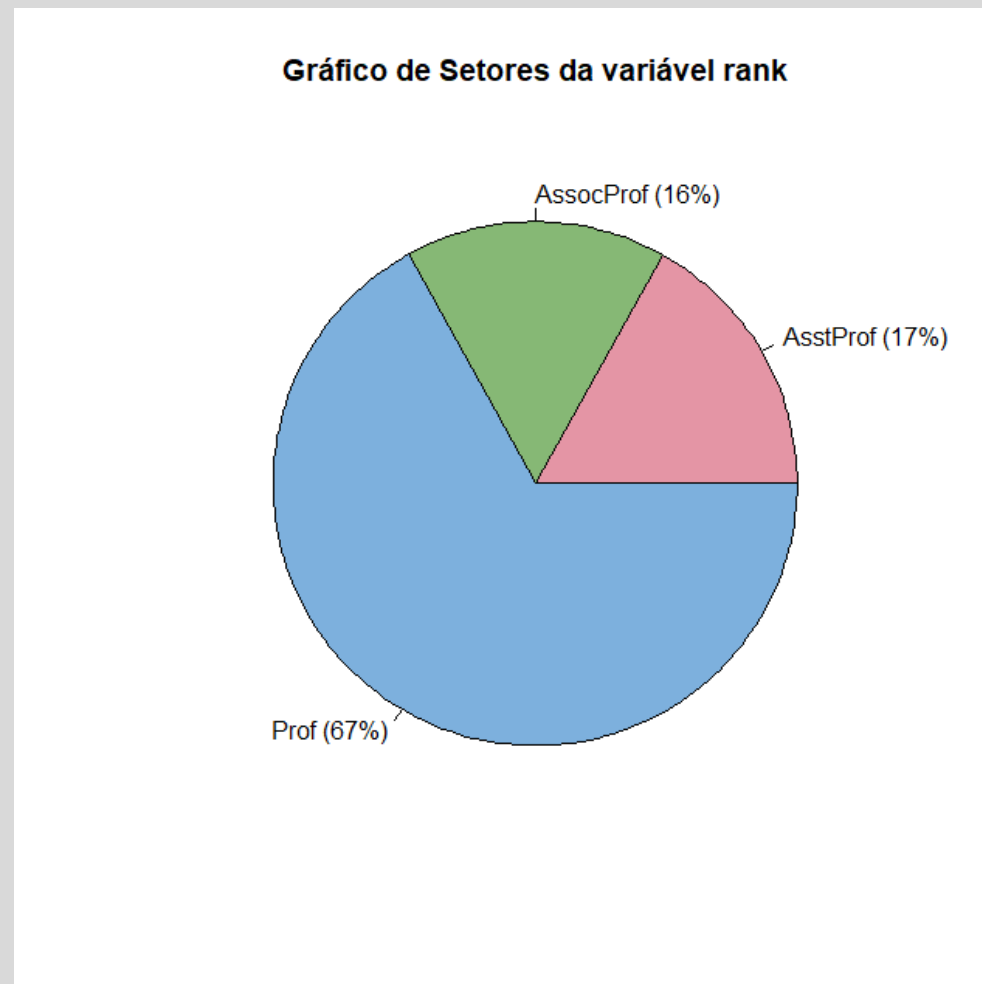
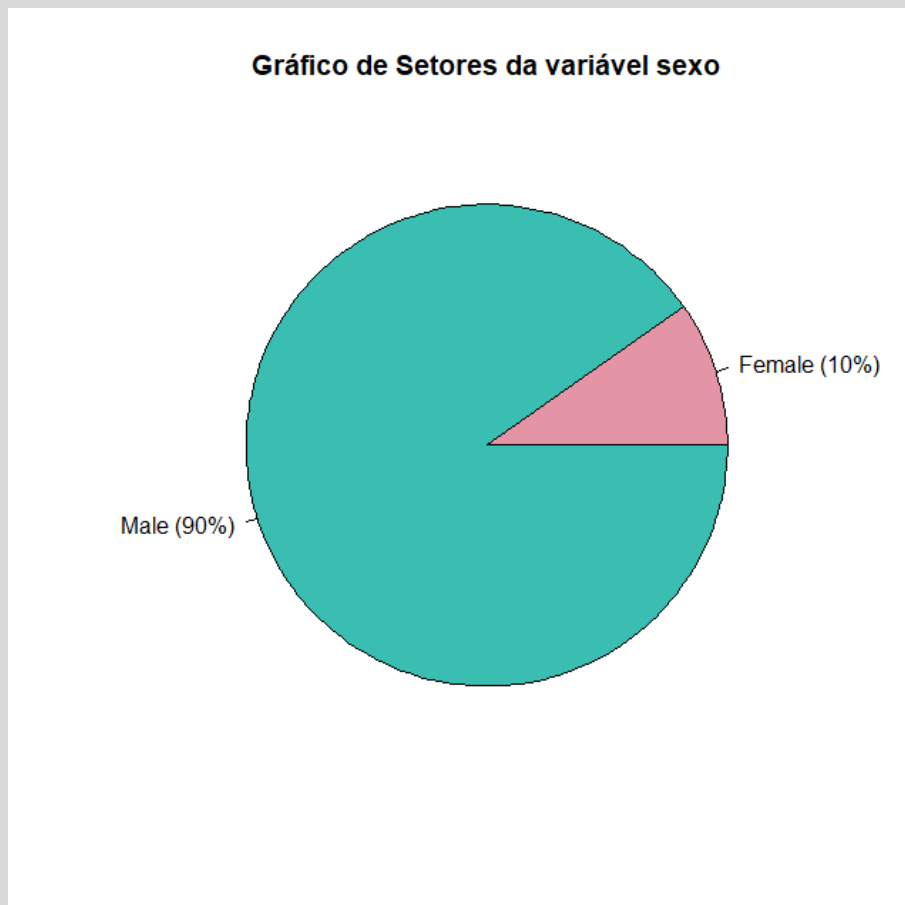




Resumos Gráficos: Gráfico de Setores

O gráfico de setores, também chamado gráfico de pizza ou torta, é um diagrama circular em que os valores de cada categoria em uma variável qualitativa são proporcionais às medidas dos ângulos, esses valores podem ser dados em frequência (contagem) ou porcentagem (mais comum).

Temos o gráfico de setores para a variável sexo e rank:

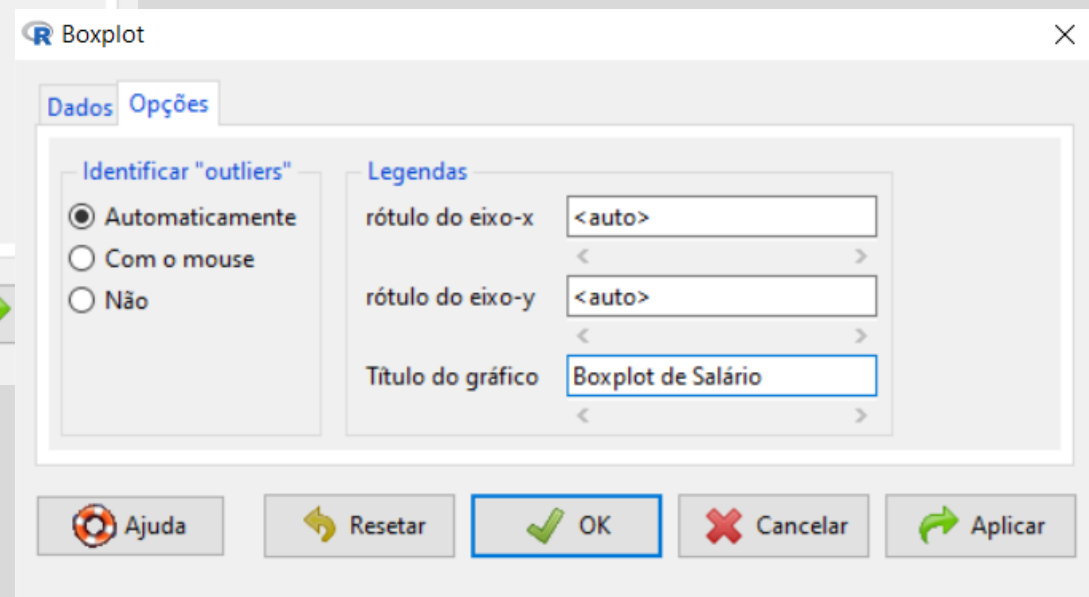
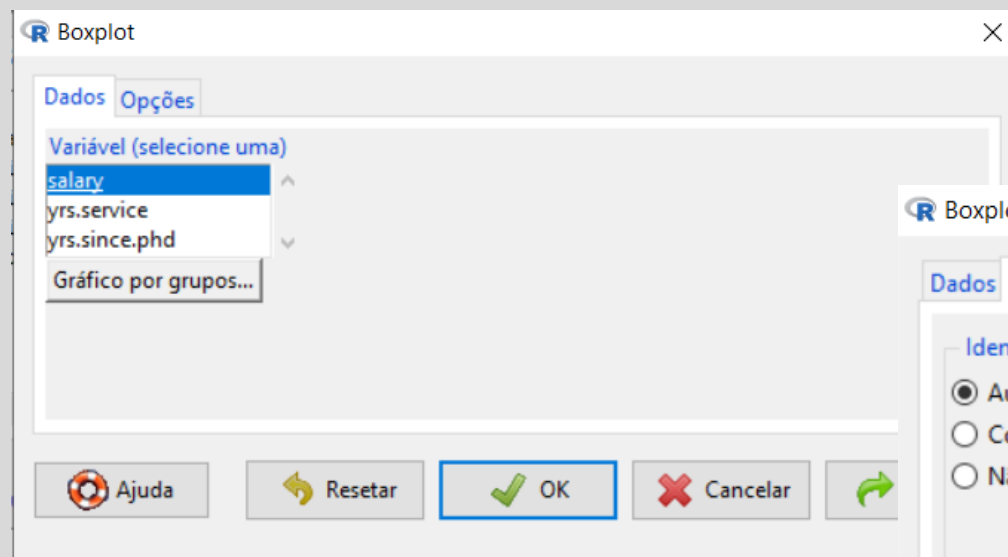




Resumos Gráficos: Boxplot

O boxplot, ou diagrama de caixa, é uma ferramenta gráfica criada para representar a variação de dados observados de determinada variável numérica, nele temos a disposição dos quartis e dos valores máximo e mínimo observados na variável, além de podermos ver a presença (ou não) de pontos discrepantes, ou outliers, e ter ideia sobre a simetria da distribuição das observações.

Primeiramente, vamos fazer o boxplot da variável salário (*Gráficos – Boxplot*).



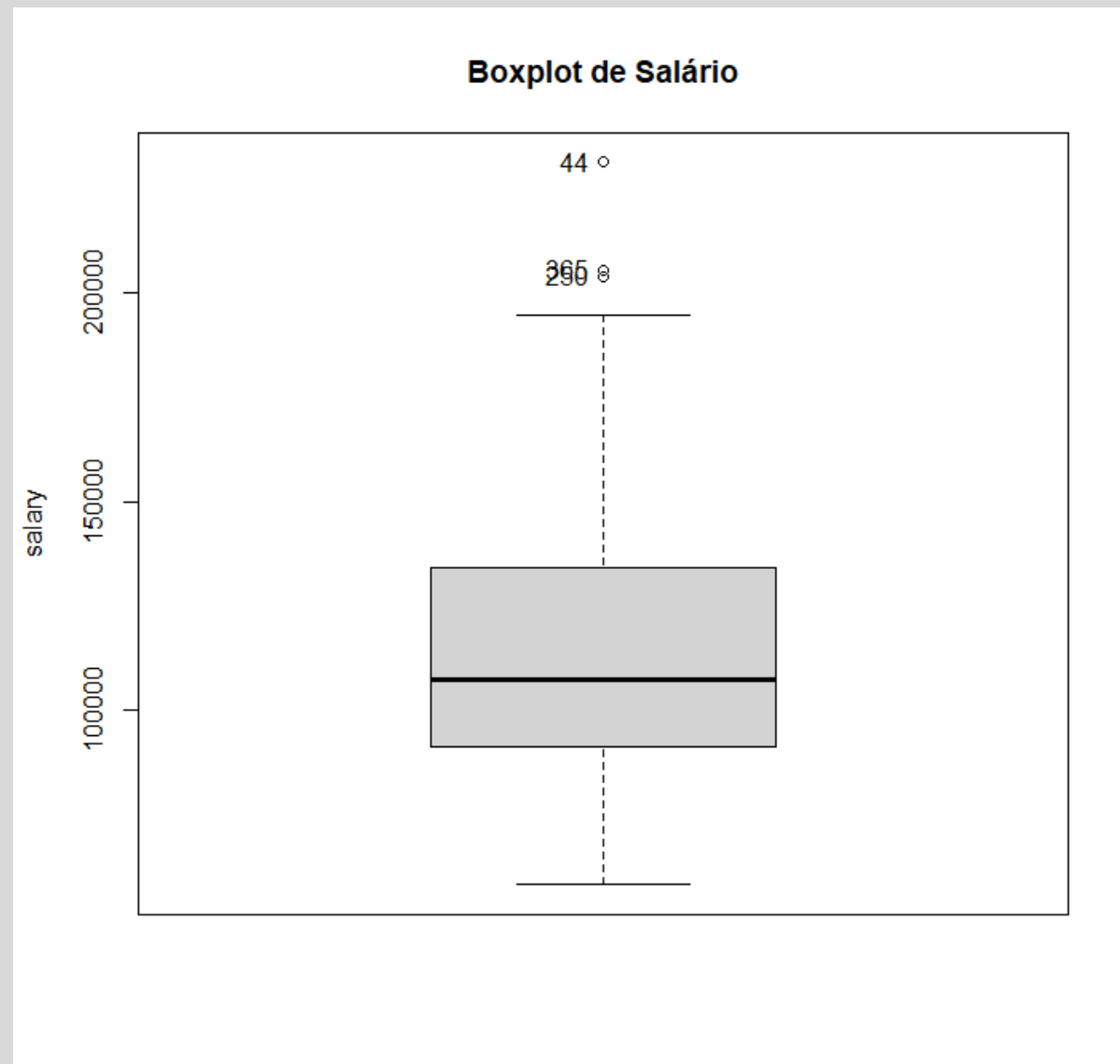


Resumos Gráficos: Boxplot

Através do boxplot podemos destacar os seguintes pontos:

- A distribuição das observações não é simétrica (valores igualmente distribuídos ao redor da média/mediana)
- Presença de 3 pontos discrepantes na parte superior do boxplot, ou seja, 3 professores ganham muito mais que o esperado em normalidade (simetria).
- LI = \$26222,5 | LS = \$198962,5
- Professor 44: \$231545
250: \$204000
365: \$205500

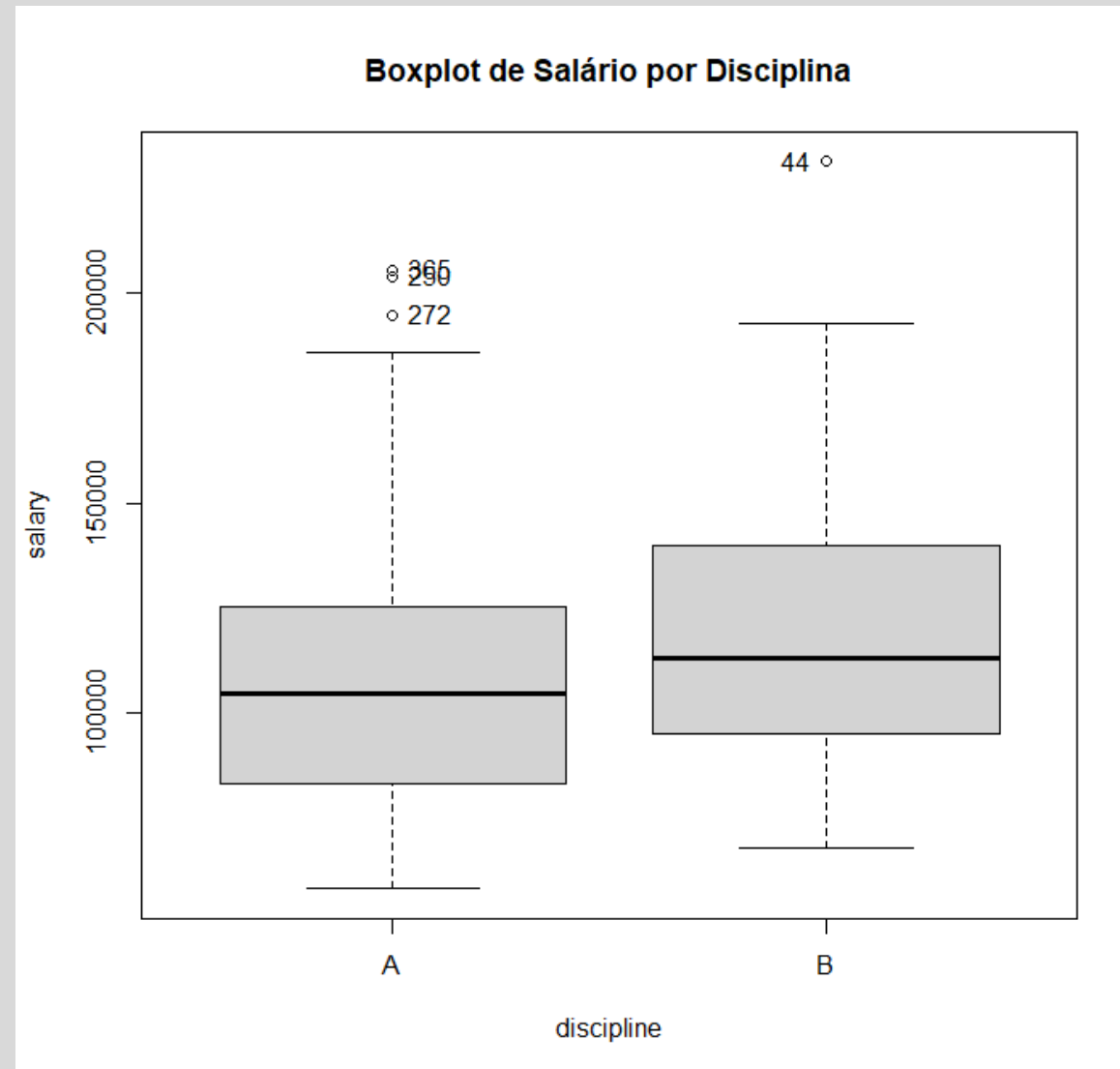
Obs: O gráfico é retornado na interface do R.





Resumos Gráficos: Boxplot

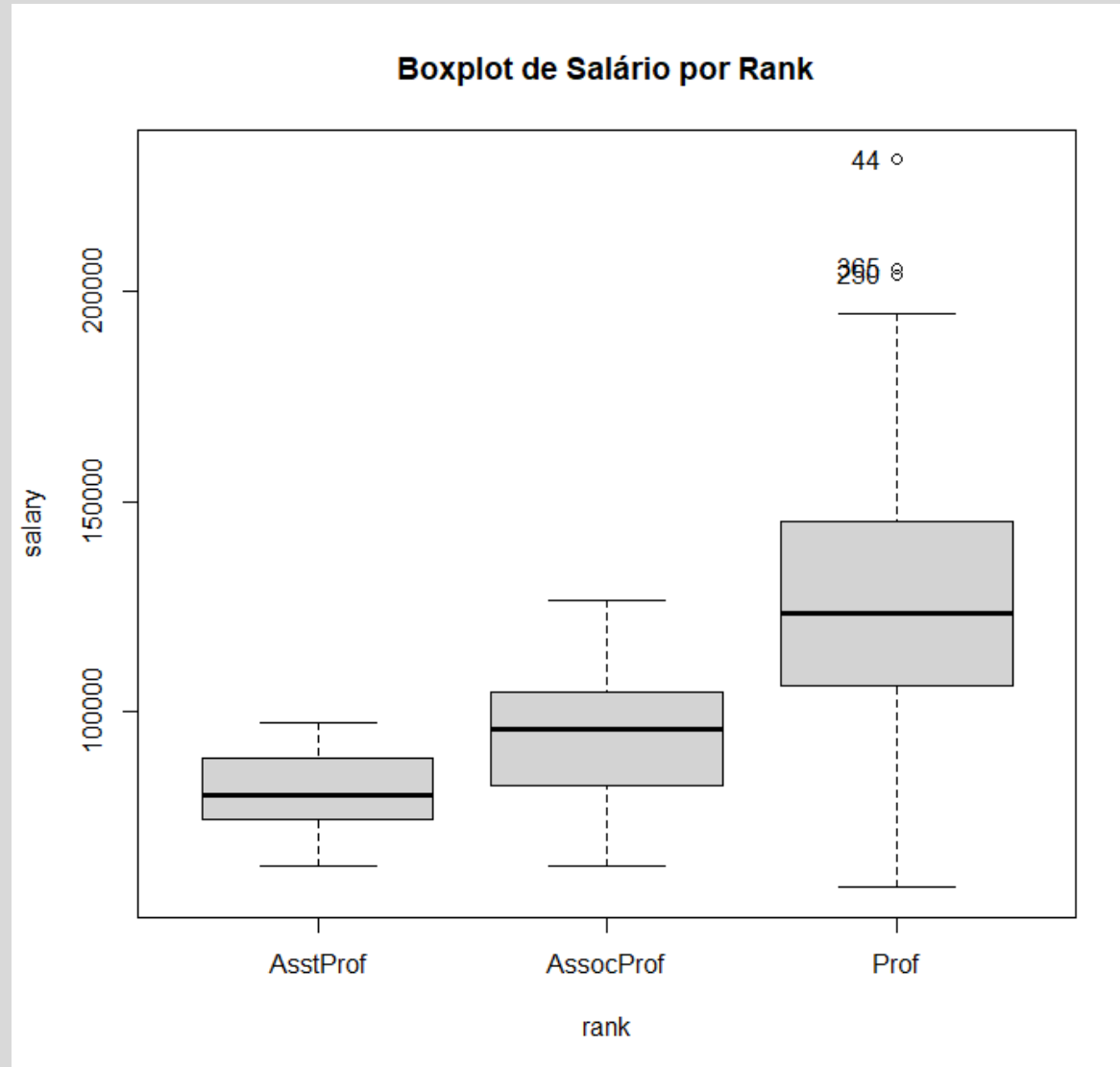
O que dizer sobre o boxplot de salários por disciplina?





Resumos Gráficos: Boxplot

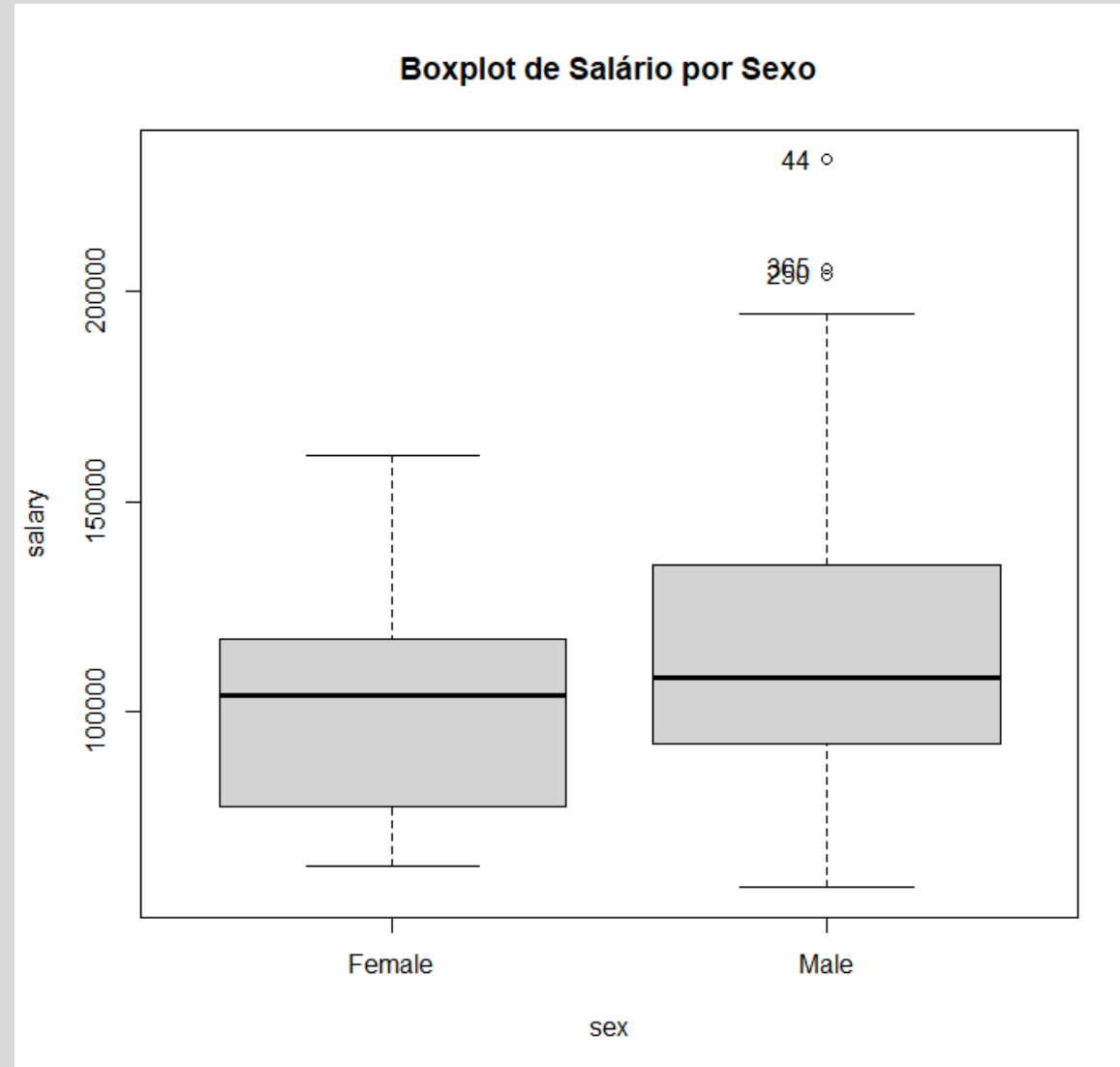
O que dizer sobre o boxplot de salários por rank?





Resumos Gráficos: Boxplot

O que dizer sobre o boxplot de salários por sexo?

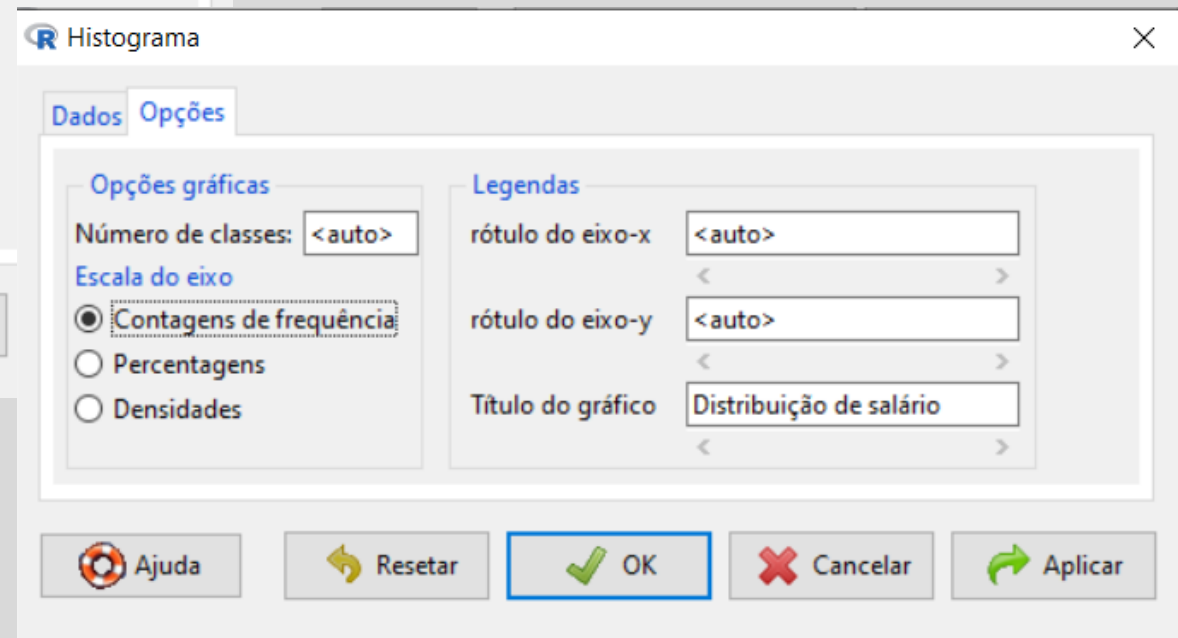
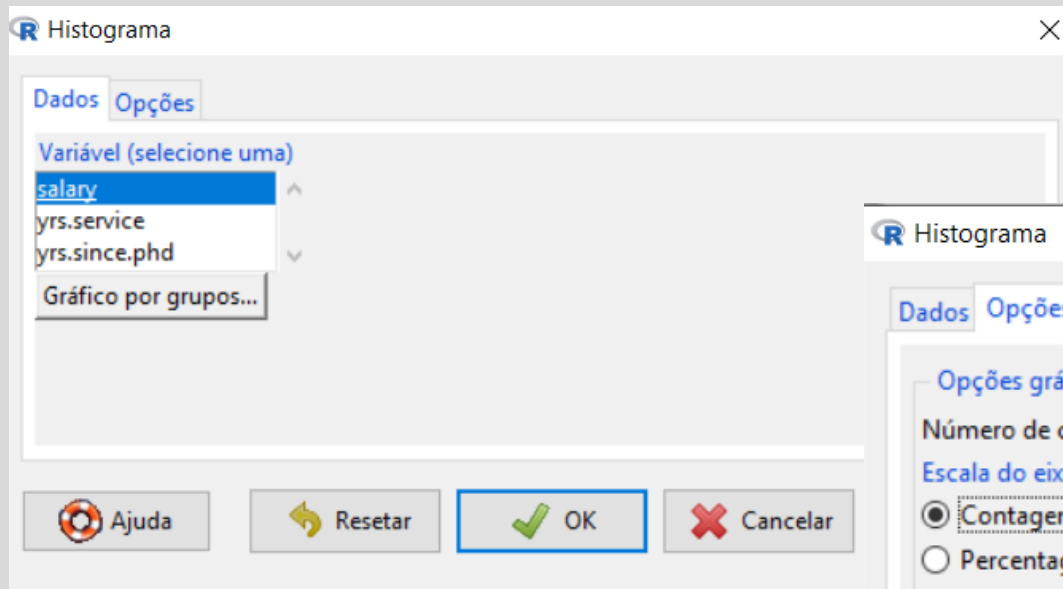




Resumos Gráficos: Histograma

O histograma, também conhecido como gráfico de frequências, é a representação gráfica da distribuição das observações de certa variável quantitativa, ou seja, com ele vemos como as observações se comportam (distribuem) em relação a localização do valor central.

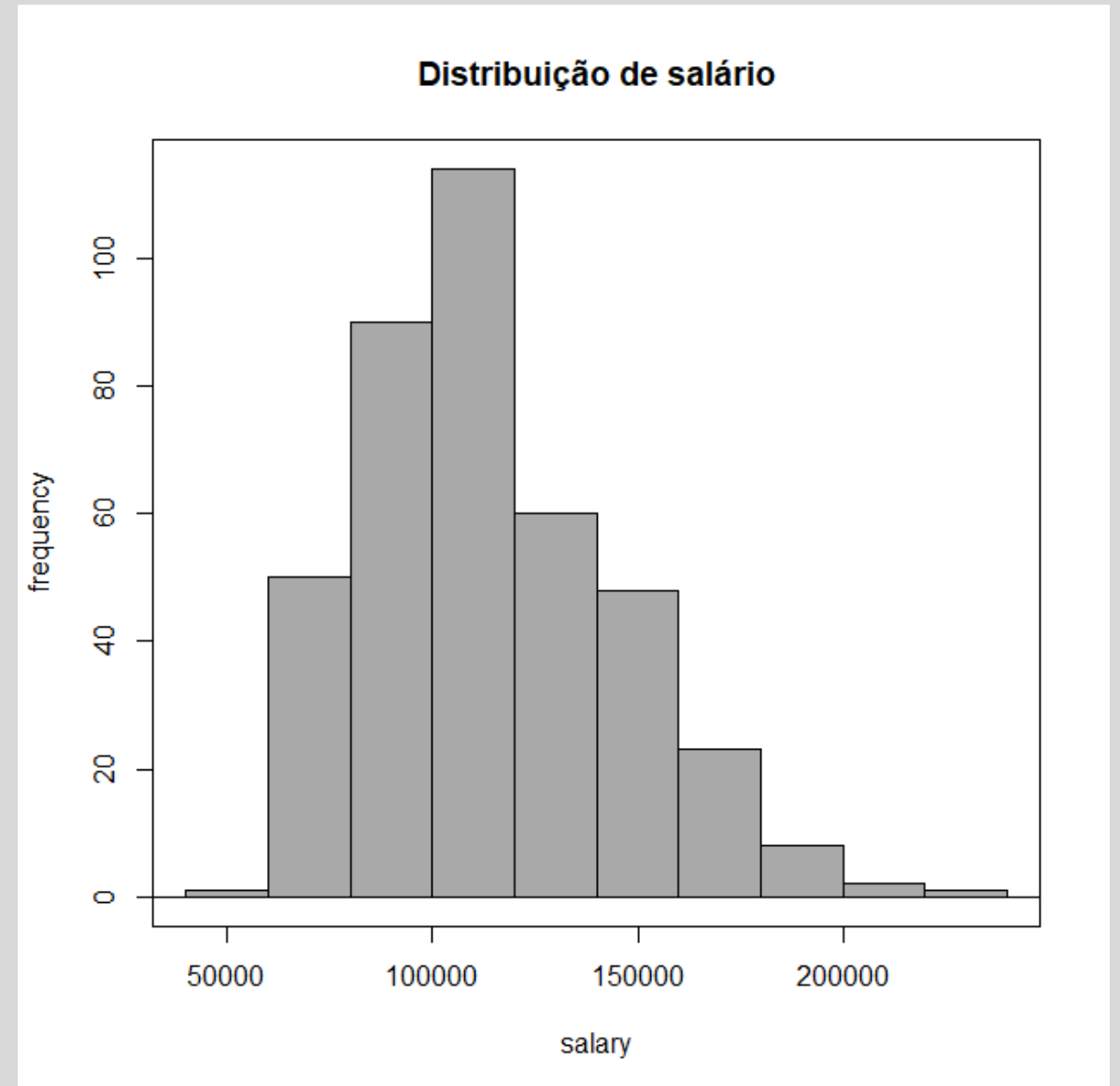
Tomando novamente a variável salário, temos:





Resumos Gráficos: Histograma

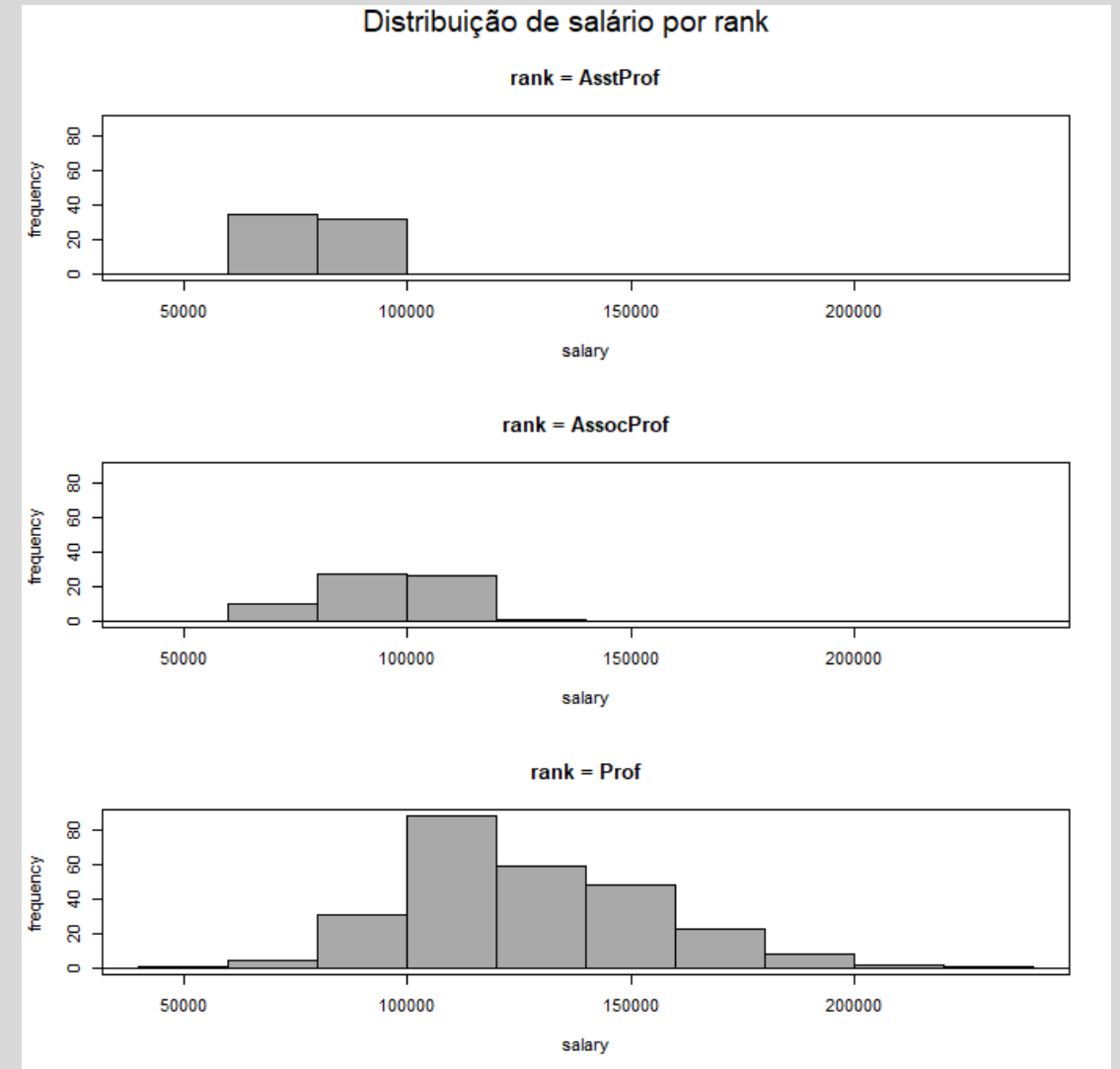
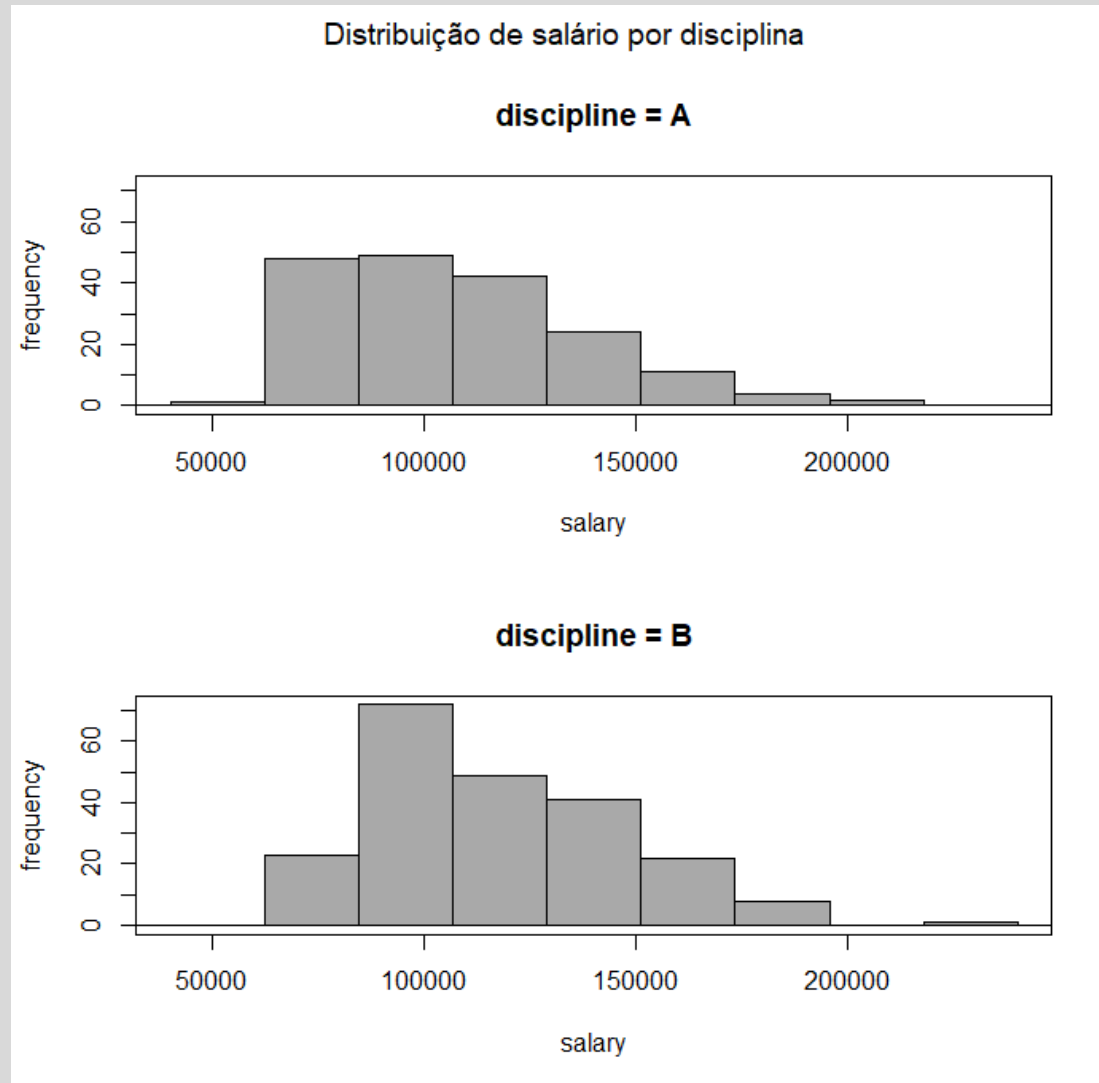
- Assimetria positiva
- Maior frequência nos valores próximos a \$100000





Resumos Gráficos: Histograma

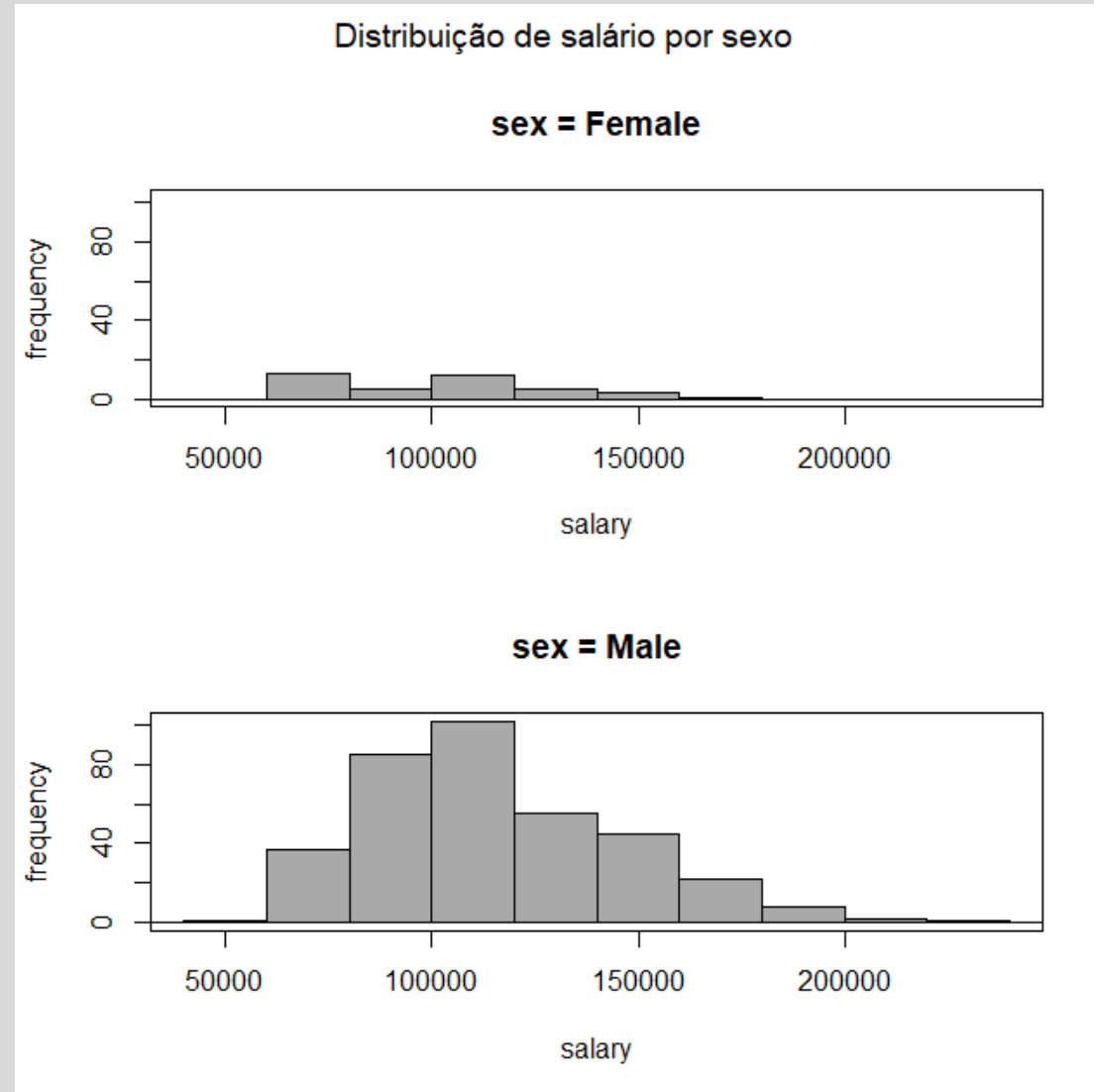
O que dizer sobre o histograma de salários por disciplina? E por rank?





Resumos Gráficos: Histograma

O que dizer sobre o histograma de salários por sexo?

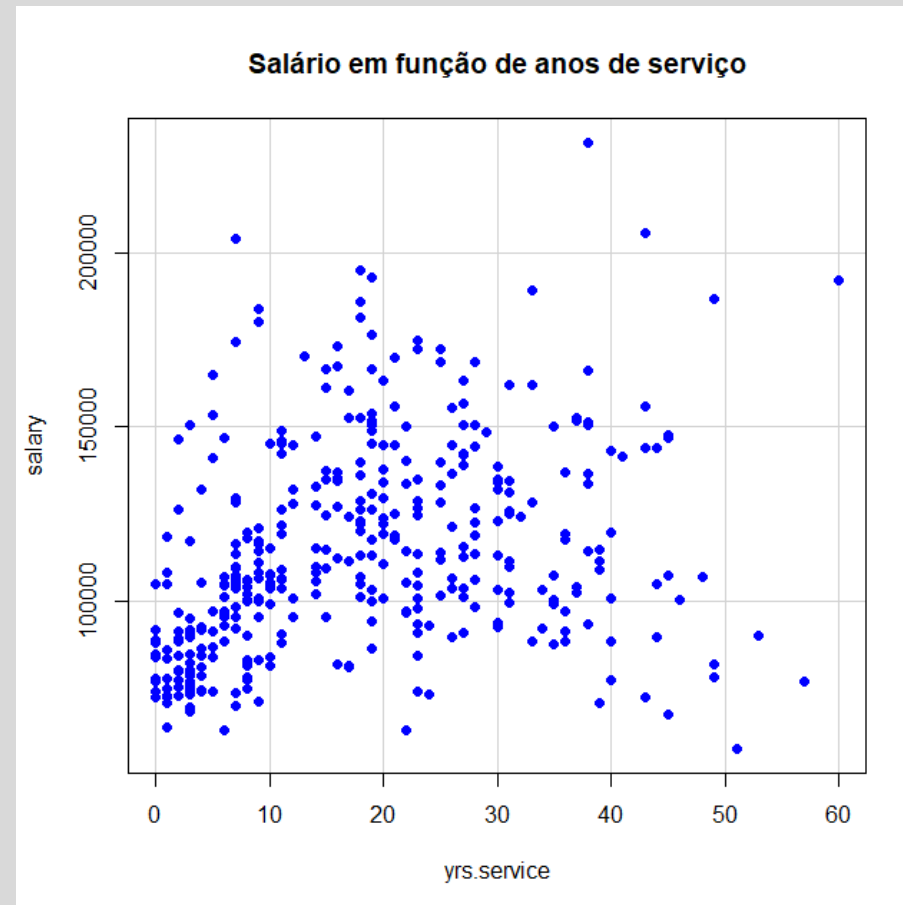




Resumos Gráficos: Gráfico de Dispersão

O gráfico de dispersão, ou diagrama de dispersão, é uma ferramenta poderosa para o estudo da relação entre duas variáveis quantitativas (normalmente contínuas), nele podemos ver a correlação entre duas variáveis através da exibição dos valores em coordenadas cartesianas.

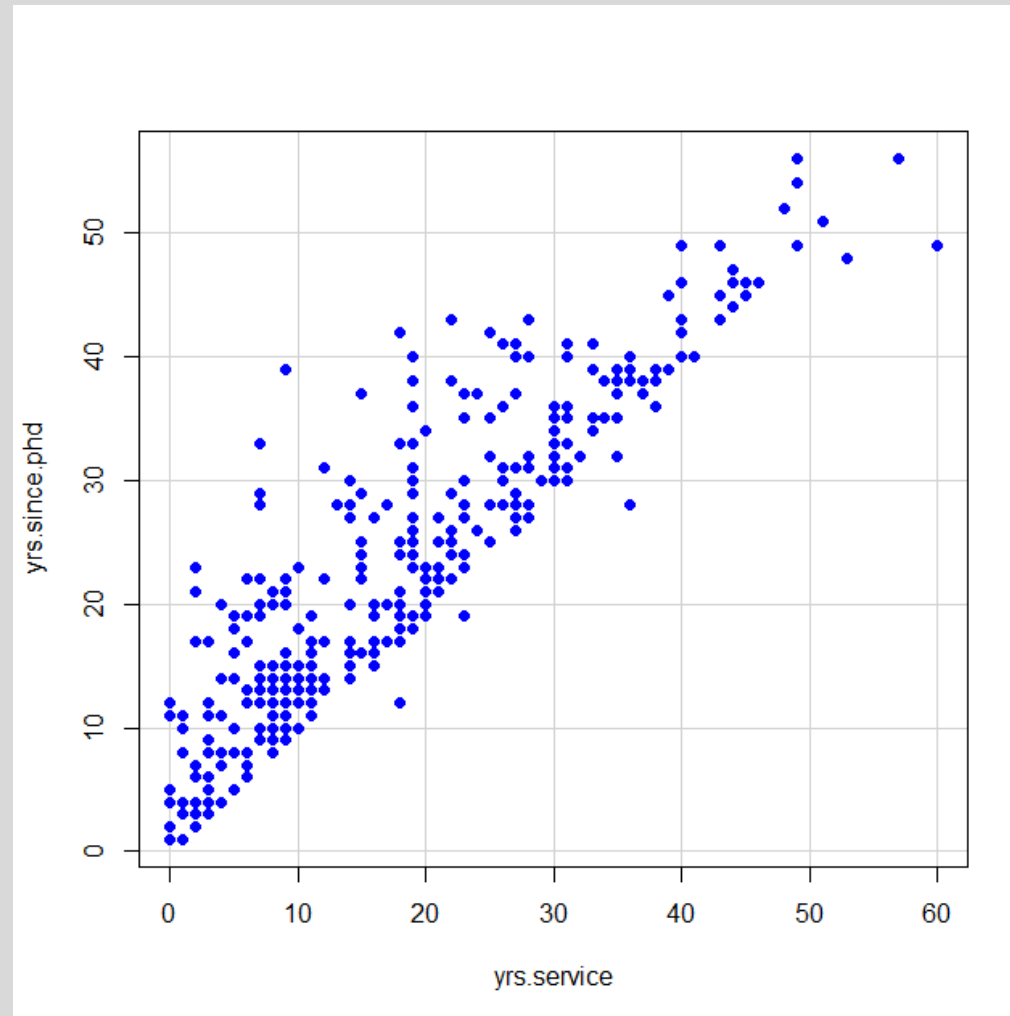
Primeiramente, como exemplo, vamos fazer o gráfico de dispersão da variável salário (eixo y) em função da variável anos de serviço (eixo x)





Resumos Gráficos: Gráfico de Dispersão

Sabemos que anos de serviço e anos desde a formação PHD têm alta correlação positiva (aproximadamente 0,91), portanto, o gráfico de dispersão dessas duas variáveis é facilmente identificável como uma reta.





Importação de conjuntos de dados

Obviamente não podemos nos restringir a conjunto de dados contidos no R. O R Commander, tal como o R, permite a importação de dados externos do tipo .txt, .csv, .sav, .xls, entre outros.

Para importamos dados externos, primeiramente temos que mudar o diretório do R Commander para o mesmo lugar onde se encontra o conjunto de dados em questão.

The screenshot shows the R Commander application window. The 'Arquivo' (File) menu is open, displaying options such as 'Alterar o diretório de trabalho...', 'Abrir arquivo c/ script...', and 'Salvar script...'. The main window displays R code for data import. A file explorer window titled 'Selecionar pasta' is overlaid, showing the directory structure: 'Usuários > Calvin > Documentos > UFJF'. The folder 'Amostragem II' is selected. The file explorer table is as follows:

Nome	Data de modificaç...	Tipo	Tamanho
ADC	30/11/2020 11:34	Pasta de arquivos	
Amostragem II	30/11/2020 11:29	Pasta de arquivos	
Computacional II	04/11/2019 20:04	Pasta de arquivos	
Curso JS	11/10/2019 19:04	Pasta de arquivos	
D3	25/10/2020 11:37	Pasta de arquivos	
ED	24/11/2019 12:02	Pasta de arquivos	
FlowingData	27/04/2019 11:23	Pasta de arquivos	
IC	25/10/2020 12:14	Pasta de arquivos	
INP	24/11/2020 09:30	Pasta de arquivos	
java script	28/04/2019 12:21	Pasta de arquivos	
Multivariada	30/11/2020 11:24	Pasta de arquivos	
R Commander	06/12/2020 13:22	Pasta de arquivos	

The file explorer window shows the path 'Pasta: Amostragem II' and buttons for 'Selecionar pasta' and 'Cancelar'.

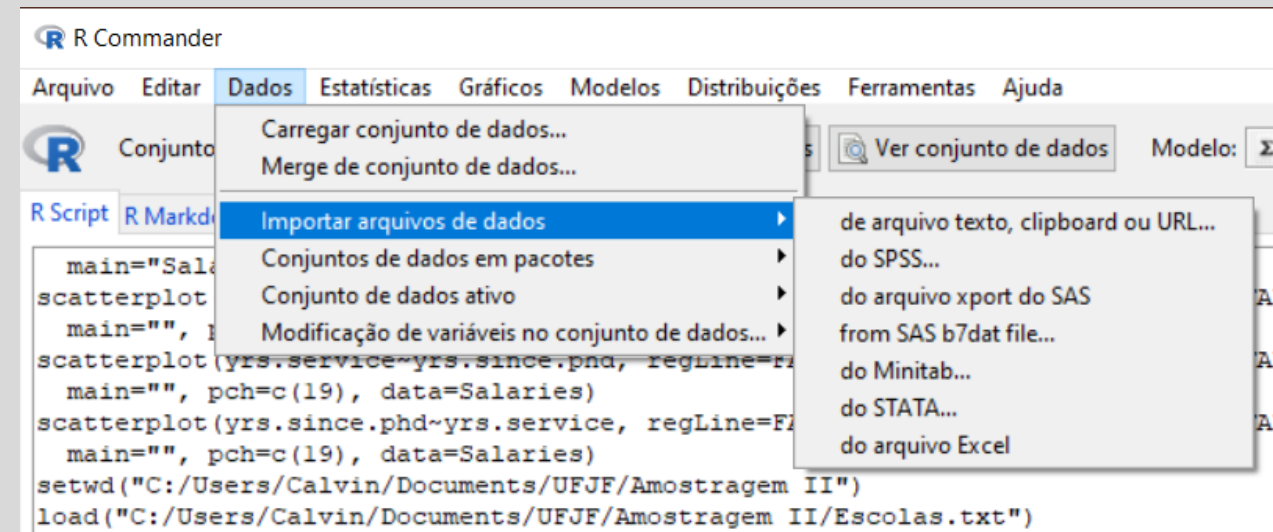


Importação de conjuntos de dados

Tendo mudado o diretório, podemos usar a opção *Dados – Importar conjunto de dados*, é muito importante nesse caso saber qual a extensão do arquivo que estamos importando, tal como o separador de campos e de decimais.

Arquivos de formato livre mais usados

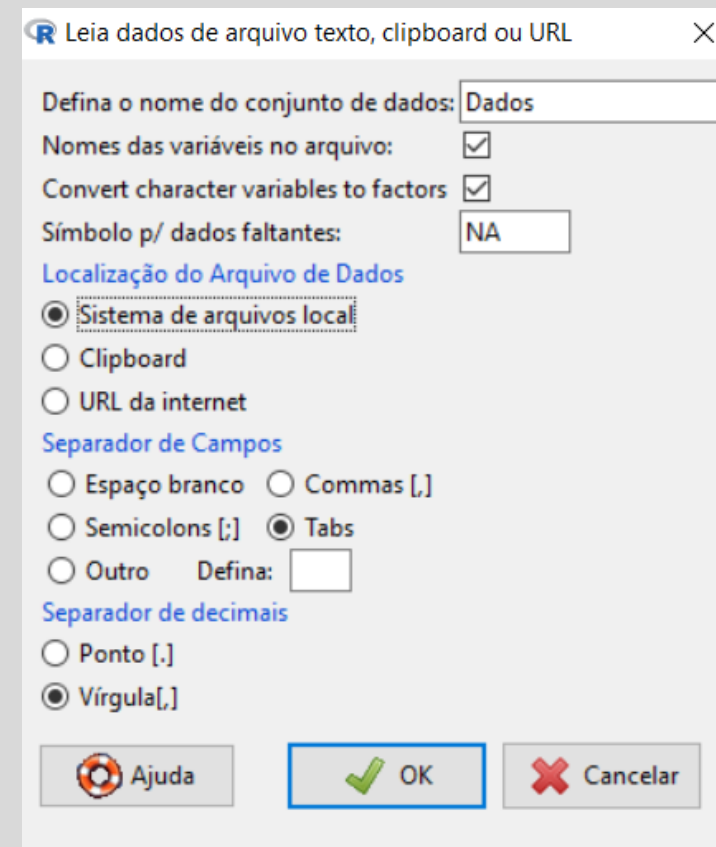
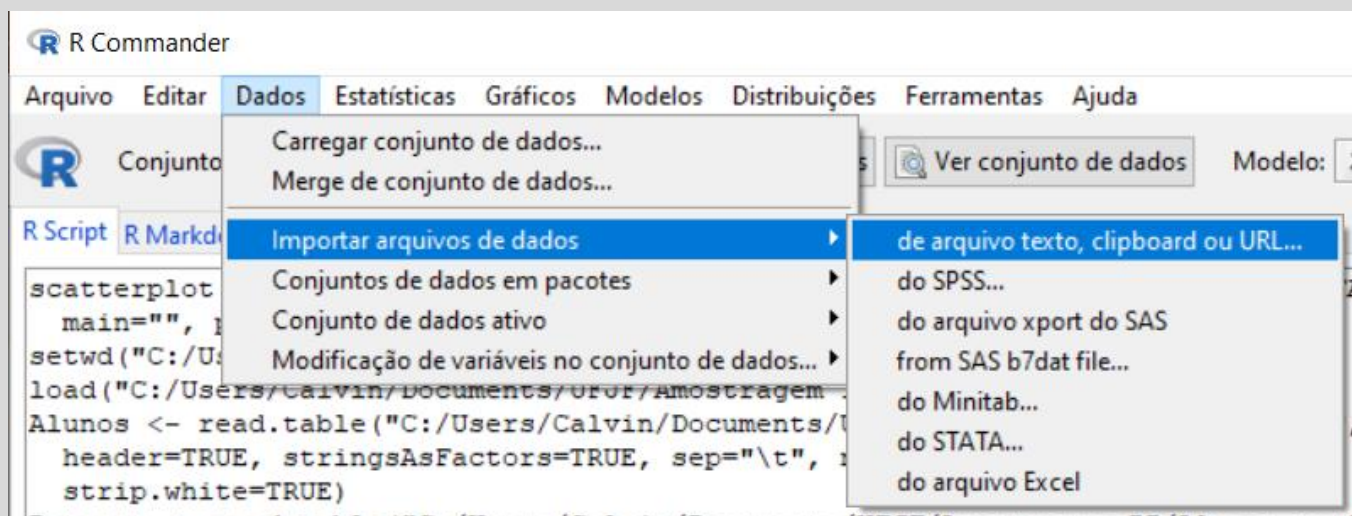
- extensão.txt
 - Separador de campos: Tab
 - Separador de decimais: Vírgula (BR)
- extensão.csv
 - Separador de campos: Ponto e vírgula
 - Separador de decimais: Vírgula (BR)





Importação de conjuntos de dados

No caso, por exemplo, de um arquivo de extensão.txt, temos o seguinte



Obs: No caso brasileiro o separador de decimais é vírgula, porém, caso o banco de dados for de outro país usamos ponto, portanto é importante saber de onde o conjunto de dados foi retirado, indico abrir sempre anteriormente em um bloco de notas.



Dados Saeb

Foi disponibilizado o banco de dados Saeb99.csv (sympla) para exercitar o que foi feito até aqui. Esse banco contém uma subamostra de 722 alunos do terceiro ano do ensino médio e 4 variáveis listadas abaixo

- Variáveis Qualitativas:
 - ufesc_c: estado onde se localiza a escola
31 – MG | 35 – SP
 - q13_3: Qual seu sexo?
1 – Masculino | 2 - Feminino
 - q26_3: Você gosta de física?
1 – Não gosto | 2 – Gosto mais ou menos |
3 – Gosto muito
- Variável Quantitativa:
profic99: proficiência em física

```
Saeb99 - Bloco de Notas
Arquivo  Editar  Formatar  Exibir  Ajuda
ufesc-c;q13-3;q26-3;profic99
31;2;2;280,40
31;1;3;308,43
31;2;1;303,59
31;2;2;338,41
31;2;3;400,53
31;1;2;384,92
31;1;2;340,11
31;1;3;352,89
31;2;3;375,12
31;2;2;307,44
31;1;2;381,71
31;2;2;318,72
31;2;3;343,45
31;2;1;335,78
31;1;1;364,69
```



Exercício

Tente responder as perguntas seguintes, é importante para fixar o conhecimento adquirido!

1. Em qual dos dois estados o desempenho em física parece ser melhor?
2. Alunos que gostam de física parecem ter um desempenho melhor que os que não gostam?
3. Mulheres parecem gostar mais de física que homens?
4. Em qual estado parece ter mais alunos que não gostam de física?

R Leia dados de arquivo texto, clipboard ou URL

Defina o nome do conjunto de dados:

Nomes das variáveis no arquivo:

Convert character variables to factors

Símbolo p/ dados faltantes:

Localização do Arquivo de Dados

Sistema de arquivos local

Clipboard

URL da internet

Separador de Campos

Espaço branco Commas [,]

Semicolons [;] Tabs

Outro Defina:

Separador de decimais

Ponto [.]

Vírgula [,]

Ajuda OK Cancelar



Referências

- <http://portal.inep.gov.br/educacao-basica/saeb>
- http://www.ufjf.br/lupercio_bessegato/2018/09/02/ensino-de-estatistica-2/
- <https://www.rcommander.com/>
- <https://www.r-project.org/>