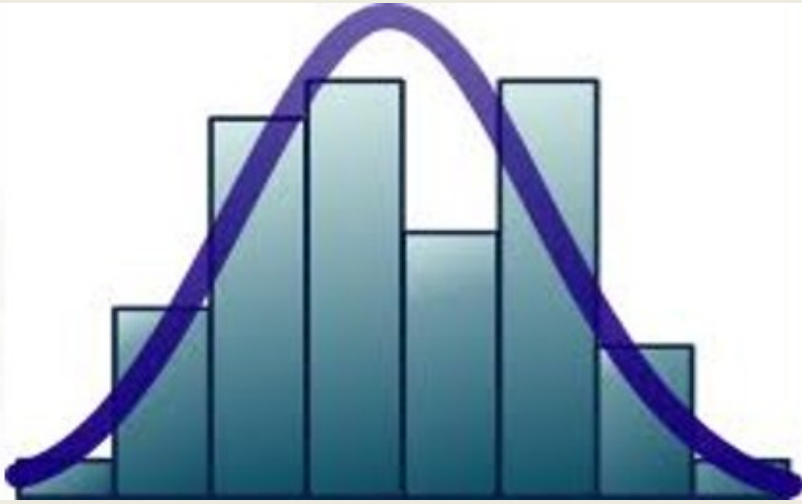
Probabilidade e Estatística



INSTITUTO FEDERAL
Catarinense
Campus Blumenau

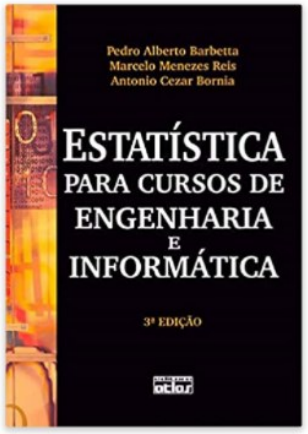
Professor Jeovani Schmitt

Probabilidade e Estatística

Aula 4

Análise Exploratória de Dados (AED)

- ✓ Distribuição de frequências: Tabelas e Gráficos para variáveis **quantitativas**
- ✓ Atividade prática – Uso do R e Rstudio para fazer uma análise de dados



Capítulo 3

Estatística Descritiva

- Tabelas
- Gráficos (Barras, Setores Circulares, Histograma, Linha, Dispersão)
- Medidas de posição (Média, Mediana, Percentis, Moda)
- Medidas de dispersão (Variância, Desvio Padrão, Coeficiente de Variação)

Exemplo: BASE SALÁRIO DE FUNCIONÁRIOS QUE TRABALHAM COM DATA SCIENCE

BASE DE DADOS salarios.csv **(no SIGAA – Aula 4)**

A	B	C	D	E	F	G	H	I
ano	experienci	emprego	cargo	salario_US	pais_empr	trab_remo	pais_empr	tam_empresa
2020	MI	FT	Data Scien	79833	DE	0	DE	L
2020	SE	FT	Machine L	260000	JP	0	JP	S
2020	SE	FT	Big Data E	109024	GB	50	GB	M
2020	MI	FT	Product Da	20000	HN	0	HN	S
2020	SE	FT	Machine L	150000	US	50	US	L
2020	EN	FT	Data Analy	72000	US	100	US	L
2020	SE	FT	Lead Data	190000	US	100	US	S
2020	MI	FT	Data Scien	35735	HU	50	HU	L
2020	MI	FT	Business D	135000	US	100	US	L
2020	SE	FT	Lead Data	125000	NZ	50	NZ	S
2020	EN	FT	Data Scien	51321	FR	0	FR	S
2020	MI	FT	Data Scien	40481	IN	0	IN	L
2020	EN	FT	Data Scien	39916	FR	0	FR	M
2020	MI	FT	Lead Data	87000	US	100	US	L
2020	MI	FT	Data Analy	85000	US	100	US	L
2020	MI	FT	Data Analy	8000	PK	50	PK	L
2020	EN	FT	Data Engin	41689	JP	100	JP	S

https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries

Exemplo: BASE SALÁRIO DE FUNCIONÁRIOS QUE TRABALHAM COM DATA SCIENCE

Descrição dos Dados

Variável	Descrição
ano	O ano em que o salário foi pago.
experiencia	O nível de experiência no cargo durante o ano com os seguintes valores possíveis: **EN** (Nível básico / Junior), **MI** (Nível médio / Intermediário), **SE** (Nível sênior / Expert), **EX** (Nível executivo / Diretor)
emprego	O tipo de emprego para a função: **PT** (Part-time), **FT** (Full-time), **CT** (Contract), **FL** (Freelance)
cargo	A função exercida durante o ano
salario_USD	O salário em USD (taxa de câmbio dividida pela taxa média em USD para o respectivo ano via fxdata.foorilla.com).
pais_empreg	O país de residência do funcionário durante o ano de trabalho como um código de país ISO 3166.
trab_remoto	O tempo total de trabalho feito remotamente, os valores possíveis são os seguintes: 0 Nenhum trabalho remoto (menos de 20%), 50 Parcialmente remoto, 100 Totalmente remoto (mais de 80%)
pais_empresa	O país da sede do empregador ou da filial contratante como um código de país ISO 3166.
tam_empresa	O número médio de pessoas que trabalharam para a empresa durante o ano: S menos de 50 funcionários (pequeno), M 50 a 250 funcionários (médio), L mais de 250 funcionários (grande)

Tabela e Gráfico

Exemplo 1: Salário em USD (salario_USD) por nível de experiência (experiencia = EN, MI, SE, EX)

Construção de **tabela e gráfico** para uma variável **qualitativa** e uma **quantitativa**

```
tabela.medias <- aggregate(base$salario_USD,
by=list(base$experiencia), FUN="mean")
colnames(tabela.medias) <- c("Experiencia","Sal_Medio")
tabela.medias

library(ggplot2)
ggplot(tabela.medias, aes(x=Experiencia, y=Sal_Medio)) +
  geom_bar(stat="identity")
```

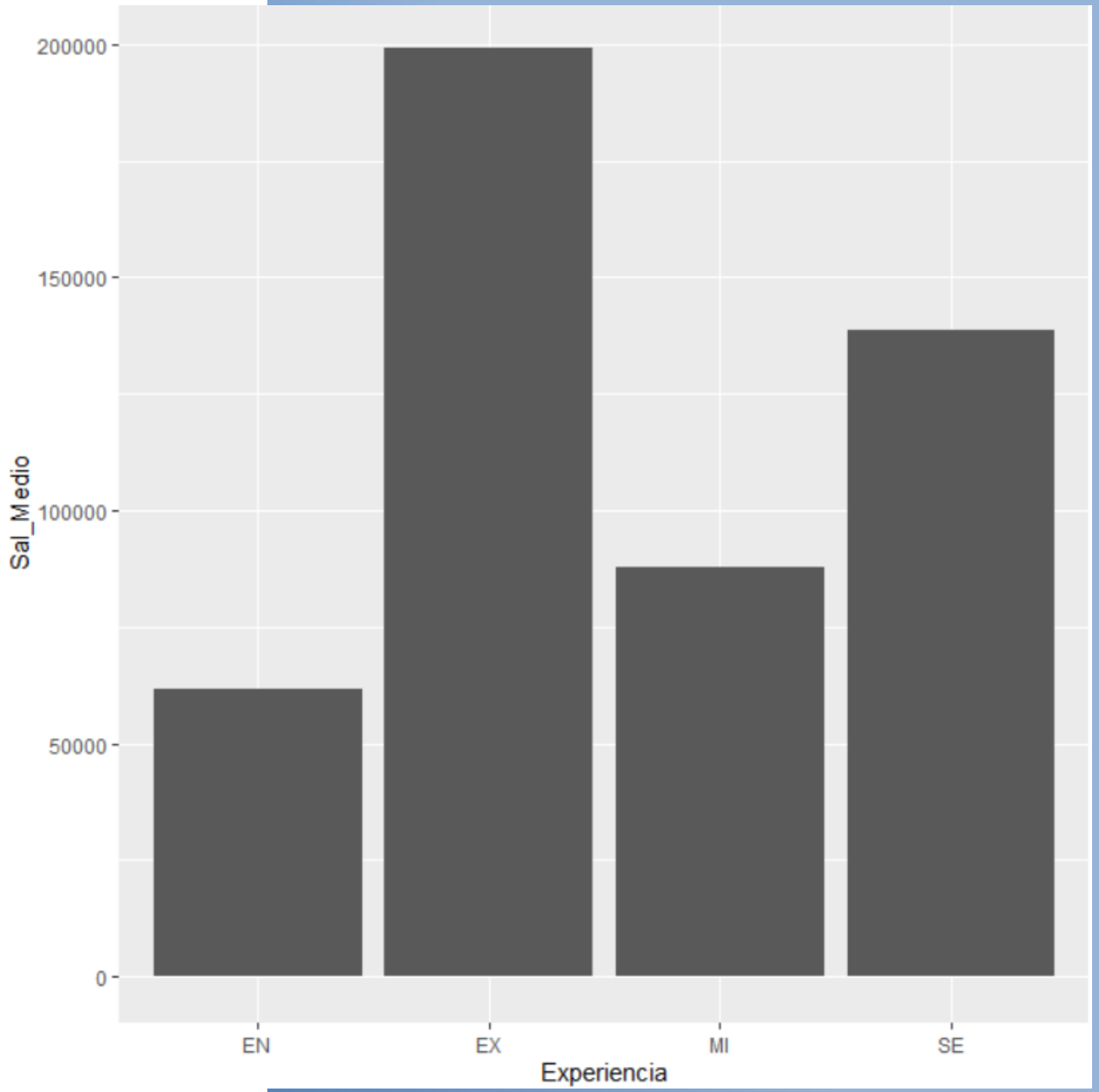
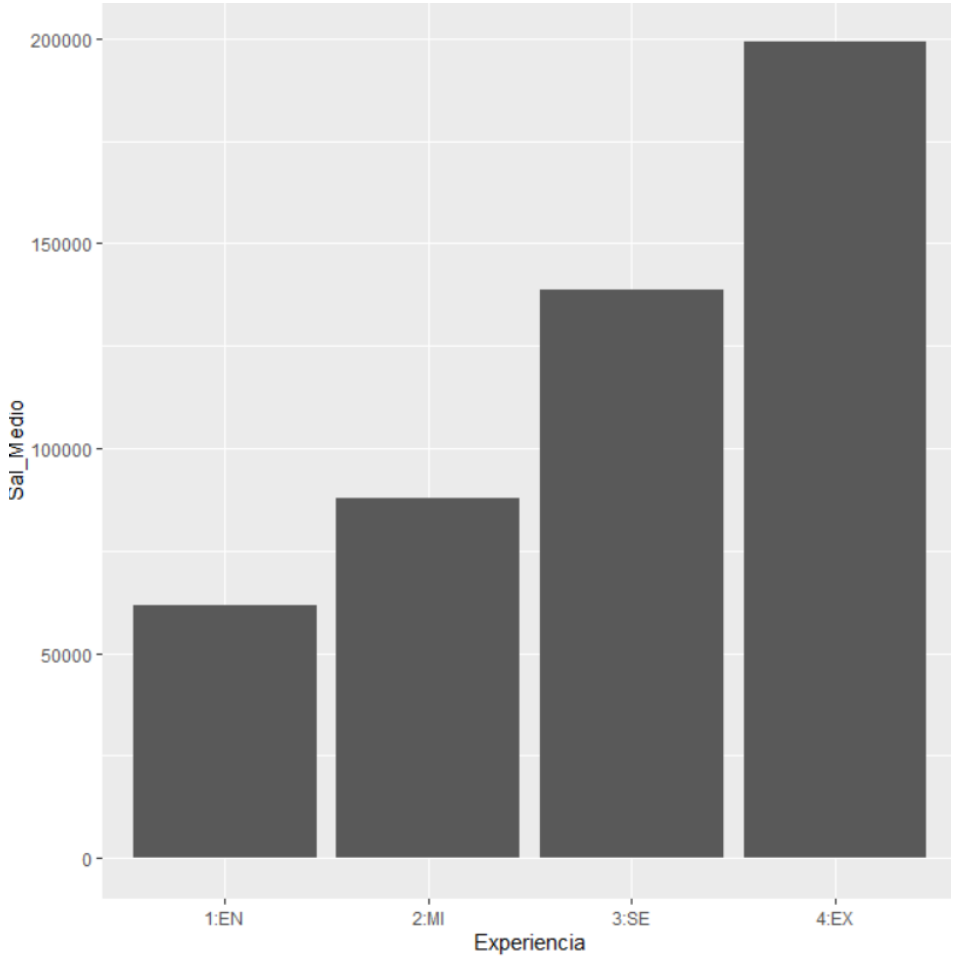

200000
150000
100000
50000
0
Sal_Medio
EN
EX
MI
SE
Experiencia

Gráfico em colunas e Tabela de médias




```
> tabela.medias
  Experiencia Sal_Medio
1        1:EN  61643.32
2        2:MI  87996.06
3        3:SE 138617.29
4        4:EX 199392.04
```

Tabela e Gráfico

Exemplo 2: Resposta dos pacientes a dois tipos de drogas em 5 níveis de dosagem

Construção de **gráfico** para duas variáveis **quantitativas**

Table 3.1 Patient responses to two drugs at five dosage levels

Dosage	Response to Drug A	Response to Drug B
20	16	15
30	20	18
40	27	25
45	40	31
60	60	40

Fonte: Kabacoff, R. I. R in action – Data analysis and Graphic with R. p. 49

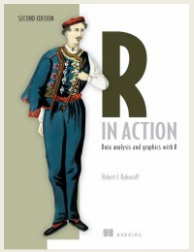
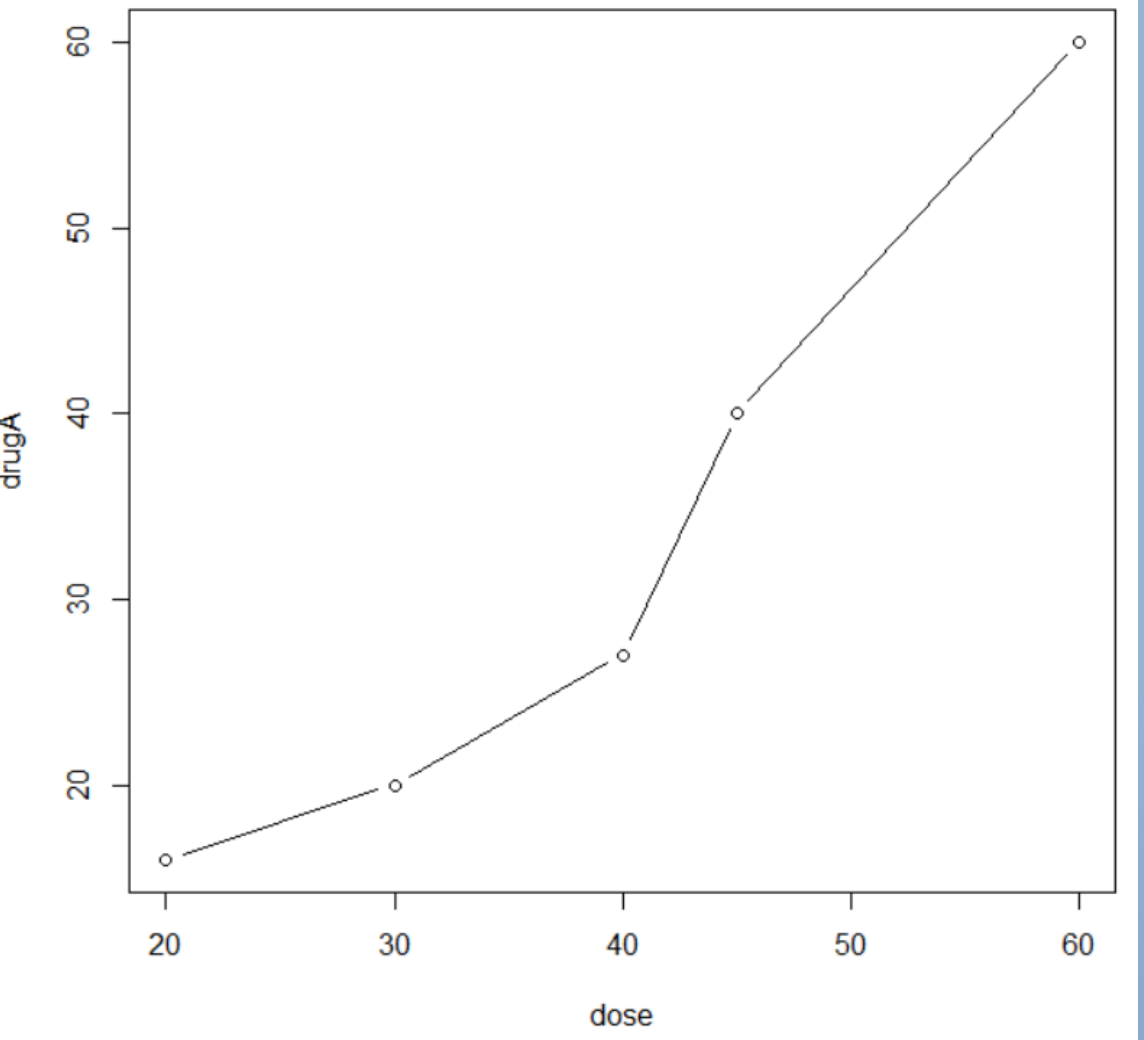


Gráfico de linha relacionando dose para resposta da droga A




```
plot(dose, drugA, type="b")
plot(dose, drugA, type="b", lty=2, pch=17)
```

Tabela e Gráfico

Exemplo 3: Comparando a resposta dos pacientes a dois tipos de drogas por dose

Construção de **gráfico** para duas variáveis **quantitativas**

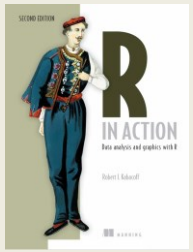
```
dose  <- c(20, 30, 40, 45, 60)
drugA <- c(16, 20, 27, 40, 60)
drugB <- c(15, 18, 25, 31, 40)

opar <- par(no.readonly=TRUE)

par(lwd=2, cex=1.5, font.lab=2)

plot(dose, drugA, type="b",
     pch=15, lty=1, col="red", ylim=c(0, 60),
     main="Drug A vs. Drug B",
     xlab="Drug Dosage", ylab="Drug Response")

lines(dose, drugB, type="b",
      pch=17, lty=2, col="blue")

abline(h=c(30), lwd=1.5, lty=2, col="gray")

library(Hmisc)
minor.tick(nx=3, ny=3, tick.ratio=0.5)

legend("topleft", inset=.05, title="Drug Type", c("A","B")
       lty=c(1, 2), pch=c(15, 17), col=c("red", "blue"))

par(opar)
```

Increases line, text, symbol, and label size

Generates the graph

Adds minor tick marks

Adds a legend



Fonte: Kabacoff, R. I. R in action – Data analysis and Graphic with R. p. 61

Gráfico de linha relacionando dose para resposta da droga A

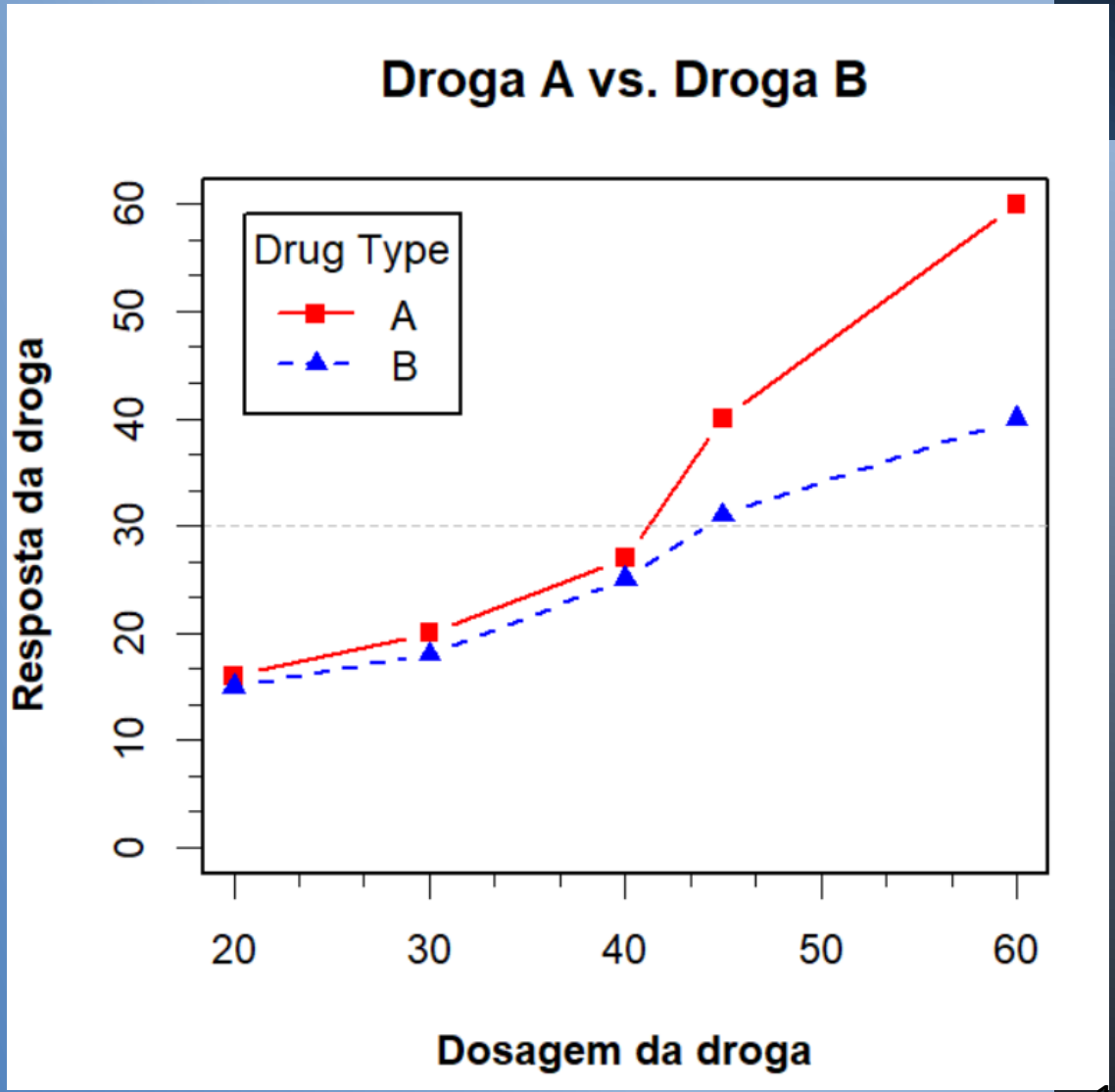


Tabela e Gráfico

Exemplo 4: Construir uma **tabela** para a variável salário (salario_USD)

Construção de tabela para uma variável **quantitativa**

```
library(psych)
# install.packages("psych")


dados$Cat_Salario[dados$salario_USD <100000]  = "G1"
dados$Cat_Salario[dados$salario_USD >=100000 & dados$salario_USD <200000]  = "G2"
dados$Cat_Salario[dados$salario_USD >=200000 & dados$salario_USD <300000]  = "G3"
dados$Cat_Salario[dados$salario_USD >=300000 & dados$salario_USD <400000]  = "G4"
dados$Cat_Salario[dados$salario_USD >=400000]  = "G5"
freq.tabela <- table(dados$Cat_Salario, useNA = "ifany")
freq.tabela
```

Tabela e Gráfico

Exemplo 4: Construir uma **tabela** para a variável salário (salario_USD)

Construção de tabela para uma variável **quantitativa**

G1	G2	G3	G4	G5
287	256	54	3	7

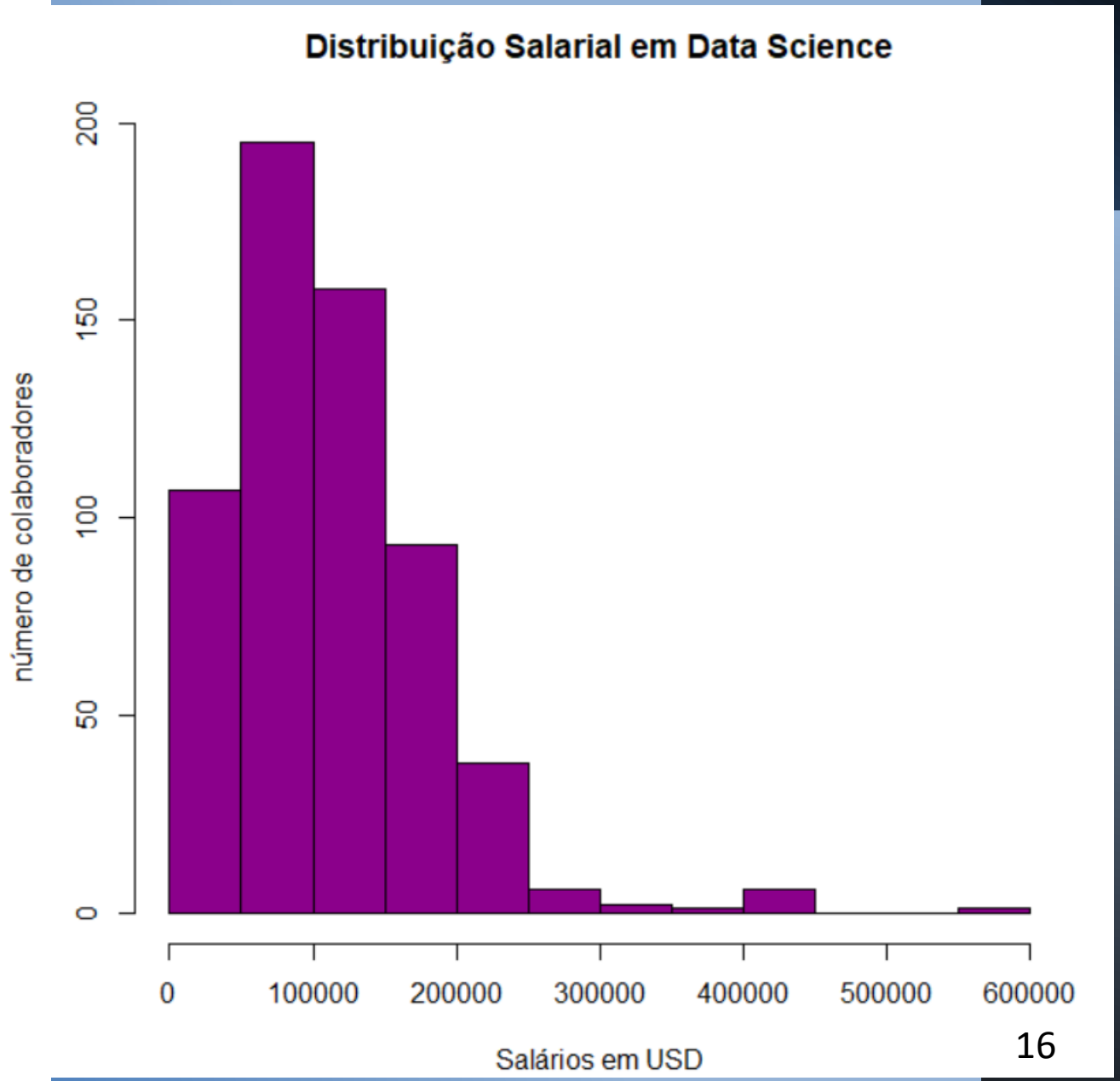
Tabela e Gráfico

Exemplo 5: Construir um gráfico para a variável salário (salario_USD) (histograma)

```
# Histograma
library(dplyr)
library(ggplot2)
hist(dados$salario_USD)
hist(dados$salario_USD,
    breaks=6)
hist(dados$salario_USD,
    main="Distribuição Salarial em Data Science",
    xlab="Salários em USD",
    ylab = "número de colaboradores",
    col="darkmagenta",
    xaxt = 'n',
    freq=TRUE)
myTicks = axTicks(1)
axis(1, at = myTicks, labels = formatC(myTicks, format = 'd'))
```

Gráfico

Histograma da distribuição salarial dos colaboradores



Explorando bases nativas do R

Exemplo 6 - base mtcars

```
data()      # mostra os conjuntos de dados disponíveis

library(help = "datasets")

?mtcars
```

mtcars é uma base disponível no R

Informações sobre marcas de 32 carros em 11 variáveis

No console do R digite ?mtcars

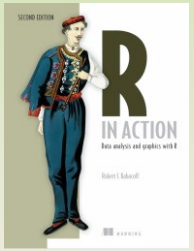
para entender a base

Combinação de gráficos

Exemplo 7: Quatro gráficos combinados em duas linhas e duas colunas para variáveis da base mtcars

```
attach(mtcars)
opar <- par(no.readonly=TRUE)
par(mfrow=c(2,2))
plot(wt,mpg, main="Scatterplot of wt vs. mpg")
plot(wt,disp, main="Scatterplot of wt vs. disp")
hist(wt, main="Histogram of wt")
boxplot(wt, main="Boxplot of wt")
par(opar)
detach(mtcars)
```

Fonte: Kabacoff, R. I. R in action – Data analysis and Graphic with R. p. 65



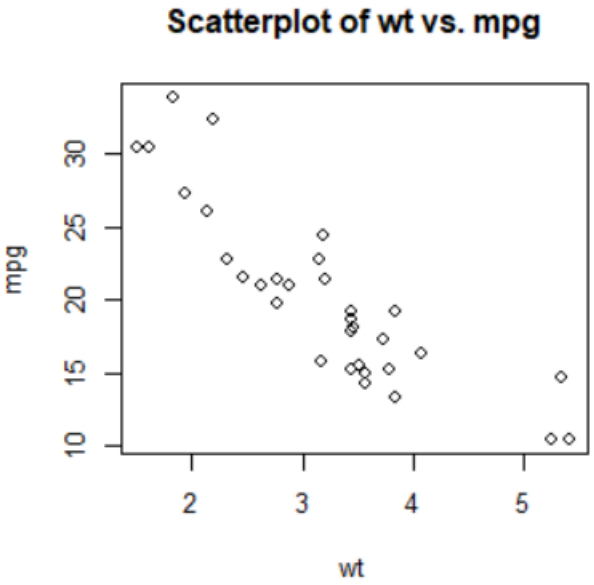
Combinação de 4 gráficos para variáveis da base mtcars

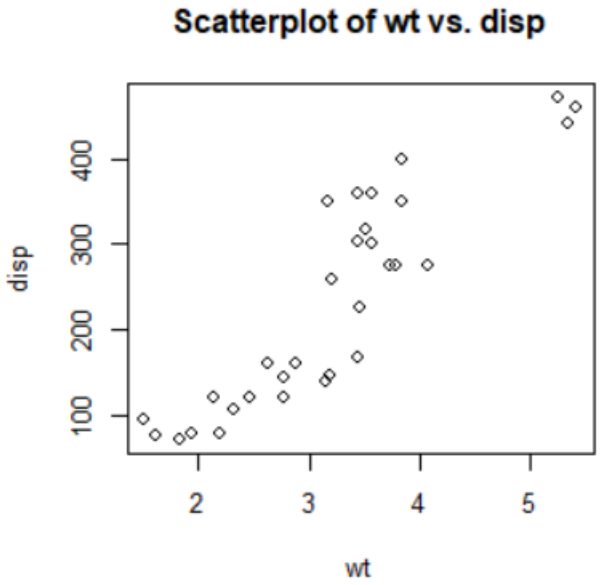
mtcars é uma base disponível no R

Informações sobre marcas de 32 carros em 11 variáveis

No console do R digite ?mtcars

para entender a base





Histograma

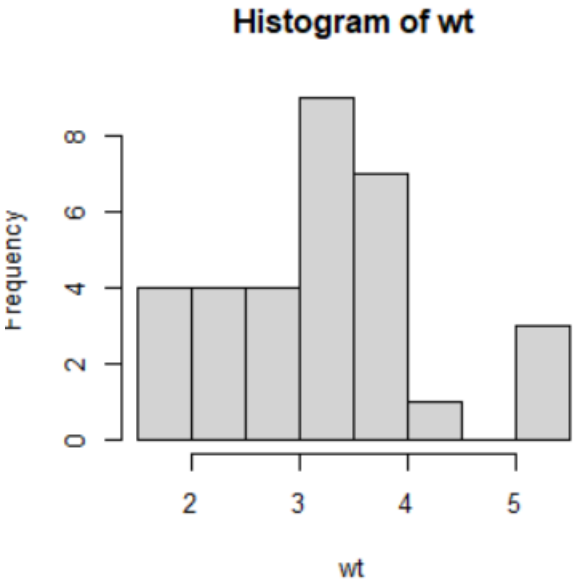
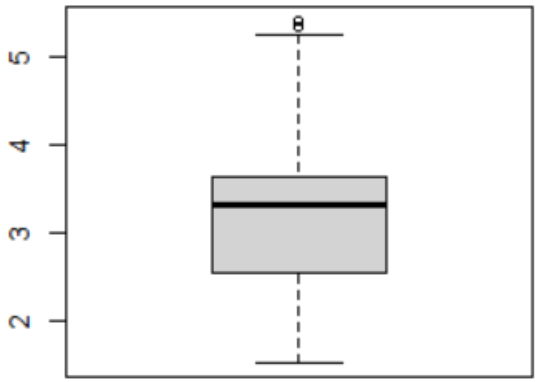


Diagrama de caixa



ATIVIDADE

No LABORATÓRIO:

Baixar do SIGAA (Aula 14/03) os arquivos:

Aula 4 - Slides - variáveis quantitativas - gráficos e tabelas

base de dados salário - Data Science

salarios

código para executar no R Studio

sintaxe_R_aula4

Reproduzir no R o código dos 7 exemplos da sintaxe