UNIVERSIDADE FEDERAL DO PARANÁ

PPGECon
Disciplina: Estatística
Professor: Adalto Acir Althaus Junior

Lista de Exercícios 01

Item 1 (dataset 1)

A planilha DEMO traz informações de 1.000 respondentes quanto à sua idade em anos, o seu estado civil (1- casado , 0- não casado), quanto tempo (em anos) vive no endereço atual, sua renda anual (em milhares de reais), o preço do carro principal (em milhares de reais), sua escolaridade (1- primeiro grau, 2- segundo grau, 3- terceiro grau, 4- Pós graduação especialização, 5- mestrado/doutorado), quanto tempo, em anos, está no emprego atual (t_emp_atual), se é (1) ou não (0) aposentado, o sexo (m- masc e f- femin) e sua satisfação no trabalho (de 1- Nada satisfeito a 5- Muito satisfeito).

APURAÇÃO

```
df1 <- read_csv2("C:/Users/DELL/OneDrive/R/Rprojetos/ufpr_ppgecon/estatistica/data/Exerc_1_descritiva_da
```

```
## i Using "',' as decimal and "'.'" as grouping mark. Use `read_delim()` for more control.
```

```
## Rows: 1000 Columns: 10
## -- Column specification --------------------------------------------------------------------------
## Delimiter: ";"
## chr (1): sexo
## dbl (9): idade, est_civil, endereco, renda, carro, escolaridade, t_empr_atual, aposentado, satisf_tra
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df1
```

```
## # A tibble: 1,000 x 10
##    idade est_civil endereco renda carro escolaridade t_empr_atual aposentado satisf_trabal sexo
##    <dbl>     <dbl>    <dbl> <dbl> <dbl>        <dbl>        <dbl>      <dbl>         <dbl> <chr>
##  1    55         1       12    72  36.2            1           23          0             5 f
##  2    56         0       29   153  76.9            1           35          0             4 m
##  3    28         1        9    28  13.7            3            4          0             3 f
##  4    24         1        4    26  12.5            4            0          0             1 m
##  5    25         0        2    23  11.3            2            5          0             2 m
##  6    45         1        9    76  37.2            3           13          0             2 m
##  7    42         0       19    40  19.8            3           10          0             2 m
##  8    35         0       15    57  28.2            2            1          0             1 f
##  9    46         0       26    24  12.2            1           11          0             5 f
## 10    34         1        0    89  46.1            3           12          0             4 m
## # i 990 more rows
```

```
summary(df1)
```

```
##      idade         est_civil        endereco          renda             carro          escolaridade
##  Min.   :18.00   Min.   :0.000   Min.   : 0.000   Min.   :   9.000   Min.   : 4.4000   Min.   :1.000
##  1st Qu.:31.75   1st Qu.:0.000   1st Qu.: 4.000   1st Qu.:  28.000   1st Qu.:13.9750   1st Qu.:2.000
##  Median :41.00   Median :1.000   Median : 9.000   Median :  43.000   Median :21.6000   Median :2.000
```

```
##  Mean   :41.42   Mean   :0.511   Mean   :11.382   Mean   :  72.911   Mean   :30.3036   Mean   :2.564
##  3rd Qu.:50.00   3rd Qu.:1.000   3rd Qu.:17.000   3rd Qu.:  80.000   3rd Qu.:39.6000   3rd Qu.:4.000
##  Max.   :77.00   Max.   :1.000   Max.   :50.000   Max.   :1116.000   Max.   :98.8000   Max.   :5.000
##      sexo
##  Length:1000
##  Class :character
##  Mode  :character
##
##
##
```

QUESTÕES:

a) Classifique cada variável em ESCALAR, ORDINAL ou NOMINAL
Resp:

b) Represente as variáveis categóricas graficamente para resumir as informações da melhor maneira possível
Resp:

c) Para as variáveis escalares faça um resumo de todas as medidas estudadas (média, mediana, desvio-padrão, etc)
Resp:

d) Examine a possibilidade das variáveis possuirem distribuição normal de probabilidades

```
shapiro.test(df1$idade)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df1$idade
## W = 0.98118328, p-value = 0.0000000004500873
```

```
shapiro.test(df1$est_civil)
```

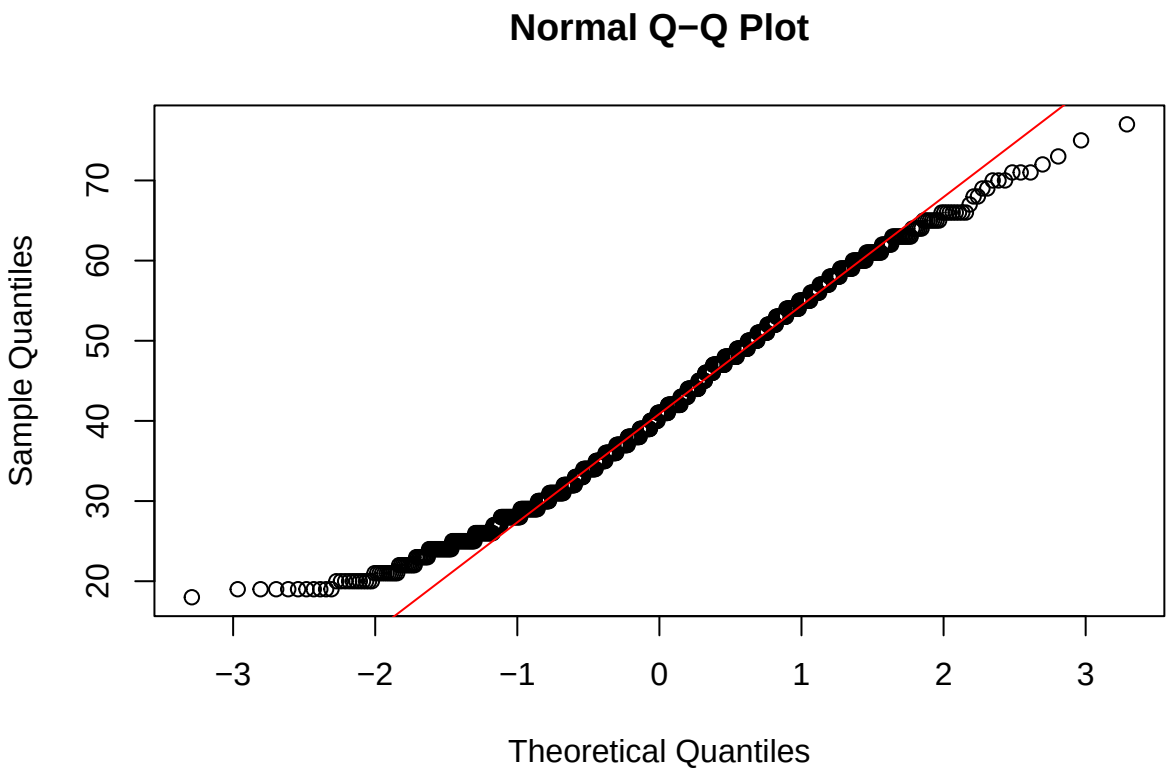
```
##
##  Shapiro-Wilk normality test
##
## data:  df1$est_civil
## W = 0.6364148, p-value < 2.2204e-16
```

```
ks.test(df1$idade, "pnorm", mean = mean(df1$idade), sd = sd(df1$idade))
```

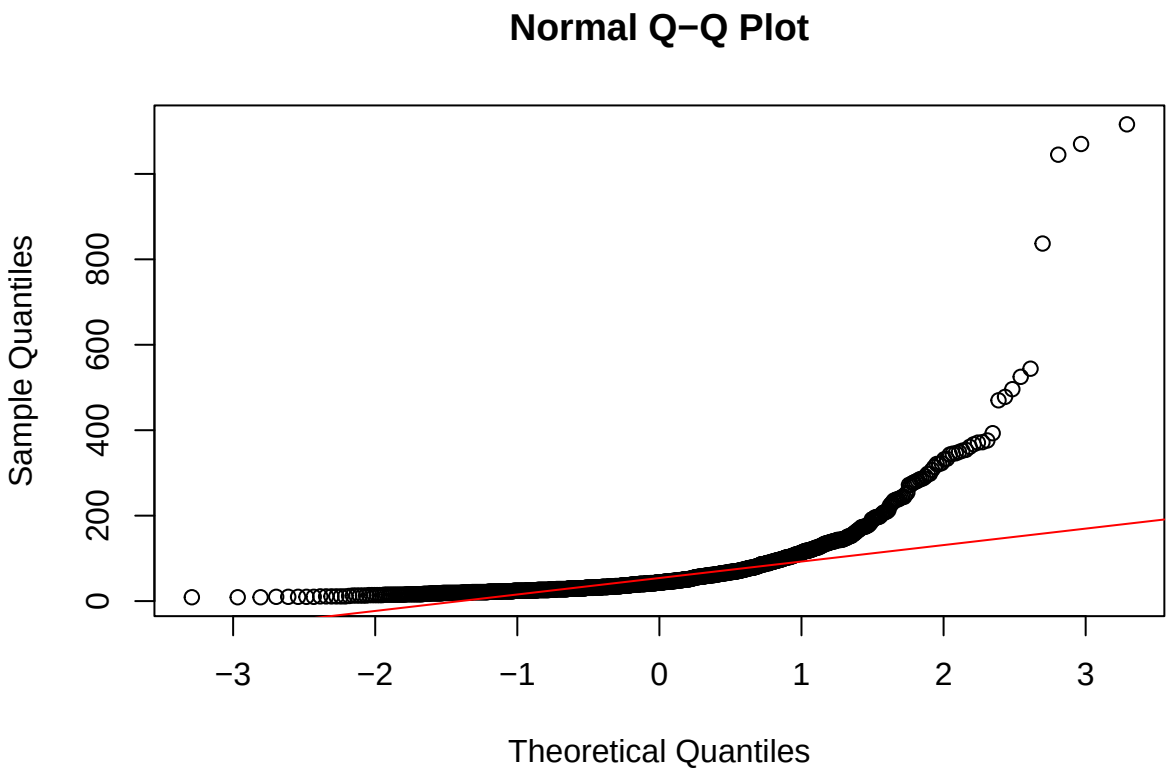
```
## Warning in ks.test.default(df1$idade, "pnorm", mean = mean(df1$idade), sd = sd(df1$idade)): ties shou
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  df1$idade
## D = 0.055819969, p-value = 0.003932065
## alternative hypothesis: two-sided
```

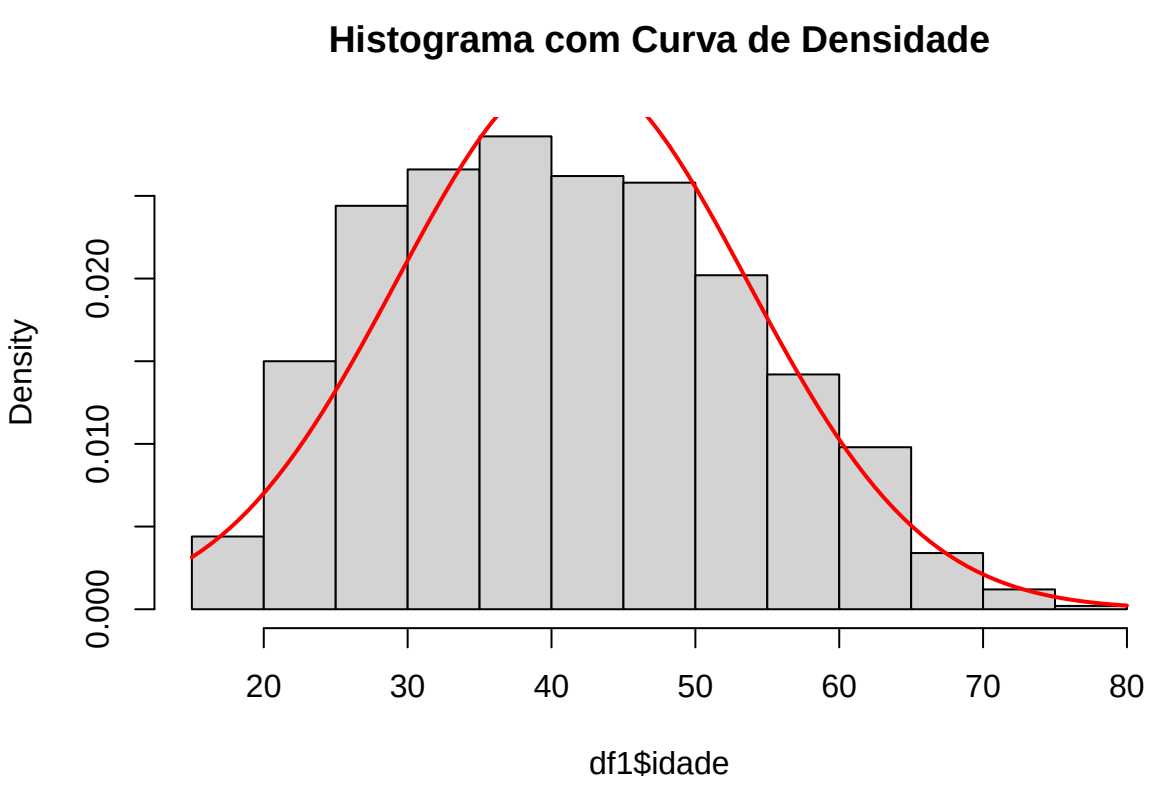
```
qqnorm(df1$idade)
qqline(df1$idade, col = "red")
```



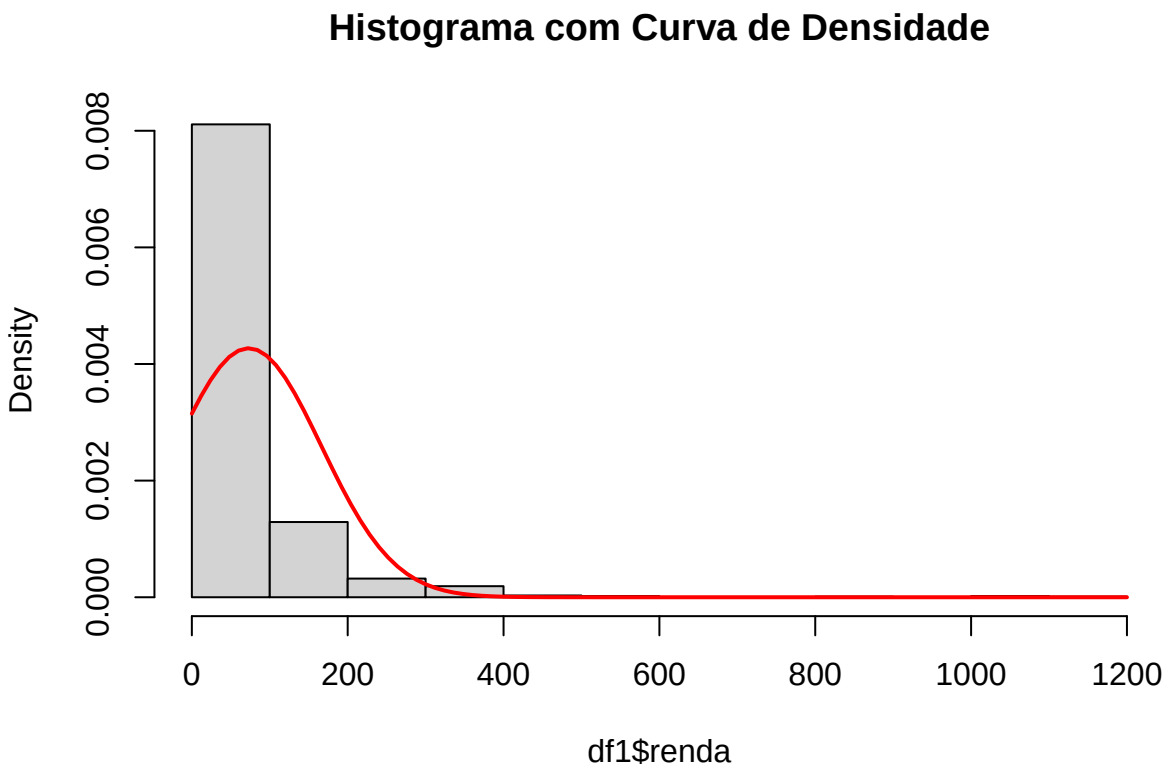

```
qqnorm(df1$renda)
qqline(df1$renda, col = "red")
```




```r
hist(df1$idade, probability = TRUE, main = "Histograma com Curva de Densidade")
curve(dnorm(x, mean = mean(df1$idade), sd = sd(df1$idade)),
      add = TRUE, col = "red", lwd = 2)
```




```
hist(df1$renda, probability = TRUE, main = "Histograma com Curva de Densidade")
curve(dnorm(x, mean = mean(df1$renda), sd = sd(df1$renda)),
      add = TRUE, col = "red", lwd = 2)
```

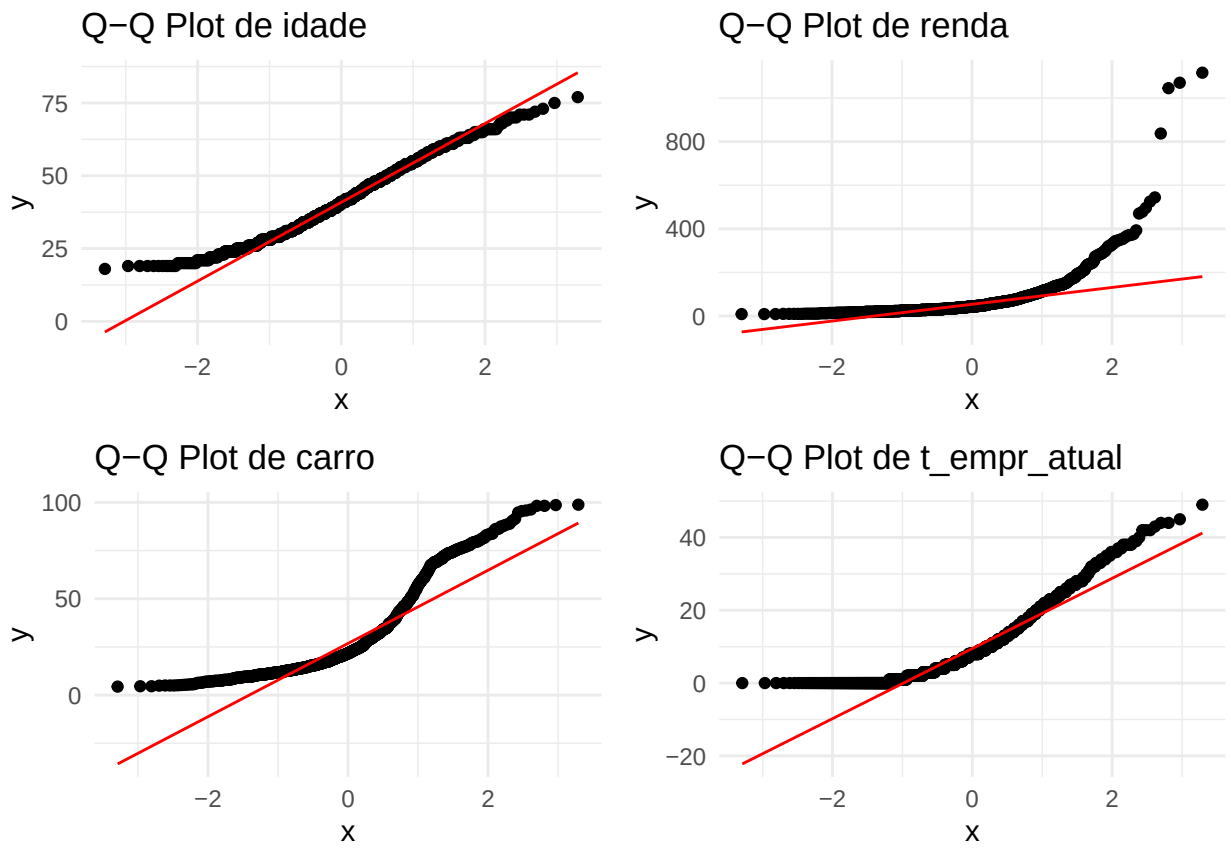



```r
# Função para criar um gráfico Q-Q com ggplot2
make_qqplot <- function(data, var_name) {
  ggplot(data, aes(sample = .data[[var_name]])) +
    stat_qq() +
    stat_qq_line(colour = "red") +
    ggtitle(paste("Q-Q Plot de", var_name)) +
    theme_minimal()
}

# Lista de variáveis numéricas (excluindo variáveis categóricas)
var_names <- c("idade", "renda", "carro", "t_empr_atual")

# Criar uma lista de gráficos Q-Q para cada variável
plots <- lapply(var_names, function(v) make_qqplot(df1, v))

# Organizar os gráficos em uma grade de 2 colunas
grid.arrange(grobs = plots, ncol = 2)
```



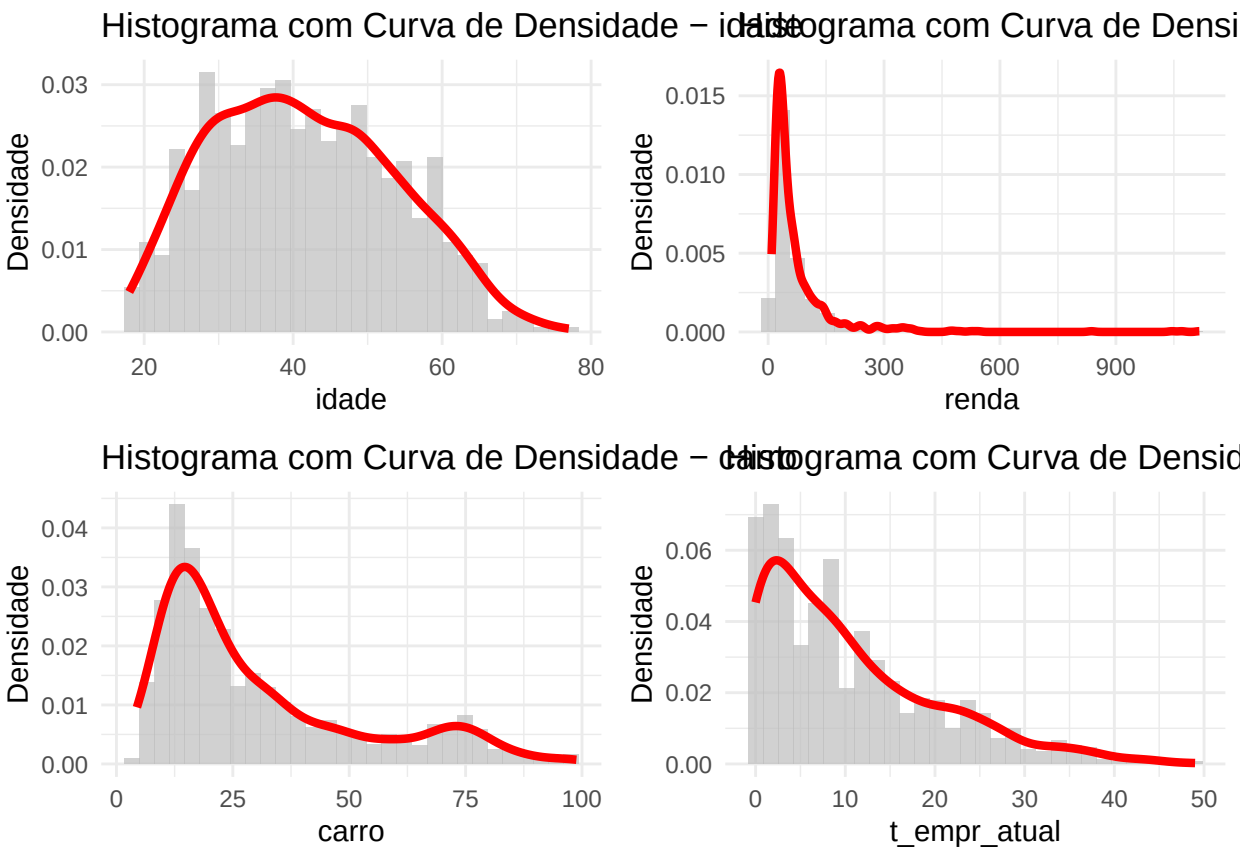

```r
# Função para criar um histograma com curva de densidade
make_histogram <- function(data, var_name) {
  # Verifica se a variável é numérica; caso contrário, retorna NULL
  if (!is.numeric(data[[var_name]])) {
    return(NULL)
  }

  p <- ggplot(data, aes_string(x = var_name)) +
    geom_histogram(aes(y = ..density..), bins = 30, fill = "gray", alpha = 0.7) +
    geom_density(color = "red", size = 1.5) +
    labs(title = paste("Histograma com Curva de Densidade -", var_name),
         x = var_name,
         y = "Densidade") +
    theme_minimal()
  return(p)
}

# Lista de variáveis numéricas
var_names <- c("idade", "renda", "carro", "t_empr_atual")

# Criar uma lista de gráficos para cada variável numérica
plots <- lapply(var_names, function(v) make_histogram(df1, v))
plots <- plots[!sapply(plots, is.null)]  # Remove NULLs caso alguma variável não seja numérica

# Organizar os gráficos em uma grade de 2 colunas
grid.arrange(grobs = plots, ncol = 2)
```



Item 2 (dataset 2)

Ao lado são apresentados dados de gastos per capita, em milhares de dólares, para cada estado americano em 20xx.

APURAÇÃO

```
df2 <- read_csv2("C:/Users/DELL/OneDrive/R/Rprojetos/ufpr_ppgecon/estatistica/data/Exerc_1_descritiva_da
```

```
## i Using "','" as decimal and "'.'" as grouping mark. Use `read_delim()` for more control.
```

```
## Rows: 50 Columns: 2
## -- Column specification --------------------------------------------------------------------------------
## Delimiter: ";"
## chr (1): estado
## dbl (1): gasto
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df2
```

```
## # A tibble: 50 x 2
##    estado      gasto
##    <chr>       <dbl>
##  1 Alabama      8.62
##  2 Alasca      12.9
```

```
##  3 Arizona      7.31
##  4 Arkansas     7.08
##  5 Califórnia   6.47
##  6 Colorado     6.53
##  7 Connecticut  8.65
##  8 Delaware     6.33
##  9 Flórida      7.01
## 10 Georgia      6.25
## # i 40 more rows
```

```r
summary(df2)
```

```
##     estado              gasto
##  Length:50          Min.   : 5.46900
##  Class :character   1st Qu.: 6.36175
##  Mode  :character   Median : 7.26300
##                     Mean   : 7.52730
##                     3rd Qu.: 8.20725
##                     Max.   :12.88500
```

QUESTÕES:

a) Faça um resumo das estatísticas descritivas desses dados

b) Decida se os dados apresentados podem estar aproximadamente normalmente distribuídos

Item 3

Suponha que o volume de negócios diários comercializados na Bolsa de Nova York (NYSE) seja uma variável normalmente distribuída com média de 1,8 bilhão e desvio-padrão de 0,15 bilhão.

APURAÇÃO

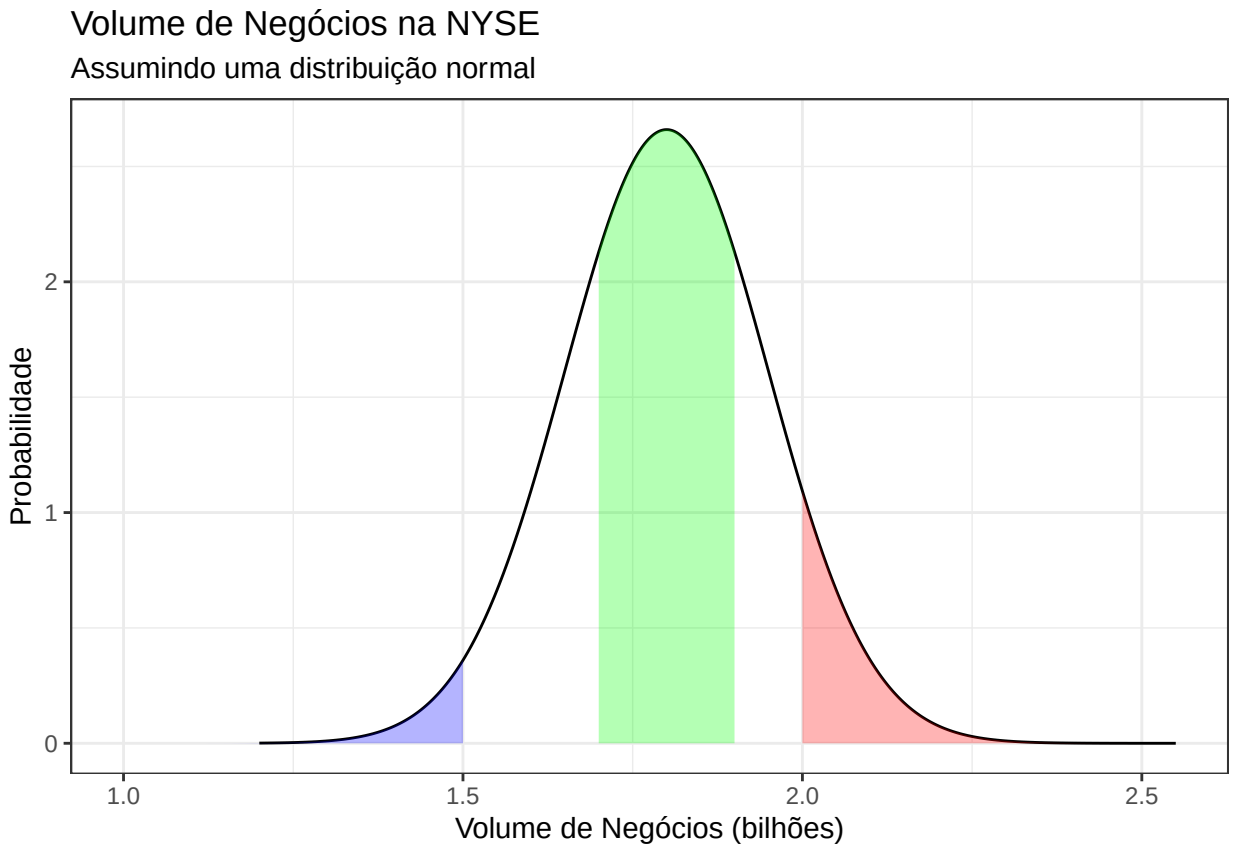
```r
# parametros
mean3 <- 1.8
sd3 <- 0.15
```

```r
# Calculando as probabilidades
## Usando 'pnorm': probabilidade de ser menor ou igual a um valor (dist normal)
prob_a <- pnorm(1.5, mean = mean3, sd = sd3)
prob_b <- 1 - pnorm(2, mean = mean3, sd = sd3)
prob_c <- pnorm(1.9, mean = mean3, sd = sd3) - pnorm(1.7, mean = mean3, sd = sd3)

# Gerando um gráfico
## Criando um data frame para o gráfico
x <- seq(mean3 - 4*sd3, mean3 + 5*sd3, length.out = 1000)
y <- dnorm(x, mean = mean3, sd = sd3)
data <- data.frame(x, y)

## Gerando o gráfico
ggplot(data, aes(x, y)) +
  geom_line() +
  stat_function(
    fun = dnorm,
```

```
  args = list(mean = mean3, sd = sd3),
  geom = "area",
  fill = "blue",
  xlim = c(1, 1.5),
  alpha = 0.3) +
stat_function(
  fun = dnorm,
  args = list(mean = mean3, sd = sd3),
  geom = "area",
  fill = "red",
  xlim = c(2, 2.5),
  alpha = 0.3) +
stat_function(
  fun = dnorm,
  args = list(mean = mean3, sd = sd3),
  geom = "area",
  fill = "green",
  xlim = c(1.7, 1.9),
  alpha = 0.3) +
labs(title = 'Volume de Negócios na NYSE',
     subtitle = 'Assumindo uma distribuição normal',
     x = 'Volume de Negócios (bilhões)',
     y = 'Probabilidade') +
theme_bw()
```



QUESTÕES:

Para um dia aleatoriamente escolhido, qual a probabilidade do volume estar:
a) abaixo de 1,5 bilhão?
Resp: 0.0227501319
b) acima de 2 bilhões?
Resp: 0.0912112197
c) entre 1,7 e 1,9 bilhão?
Resp: 0.4950149249

Item 4

Uma análise estatística de 1.000 chamadas telefônicas de longa distância originadas dos escritórios da Bricks and Clicks Computer Corporation indicam que a duração dessas chamadas estão normalmente distribuídas. Sendo a média e o desvio-padrão da duração das chamadas 240 segundos e 40 segundos, respectivamente.

APURAÇÃO

```
mean4 <- 240
sd4 <- 40
```

QUESTÕES:

a) Calcule a probabilidade de uma chamada durar menos de 180 segundos.

b) Qual a probabilidade de uma chamada durar entre 200 e 300 segundos?

c) Um empregado realizou diversas chamadas com duração acima de 350 segundos.
Você pode aceitar que esse é um fato casual?

Item 5 (dataset 5)

Uma pesquisa realizada entre instituições financeiras da América Latina apresentou os resultados descritos na tabela abaixo. Você diria que existe associação entre o tempo de atuação e o número de clientes?

APURAÇÃO

```
df5 <- read_csv2("C:/Users/DELL/OneDrive/R/Rprojetos/ufpr_ppgecon/estatistica/data/Exerc_1_descritiva_da
```

```
## i Using "',' " as decimal and "'.'" as grouping mark. Use `read_delim()` for more control.
```

```
## Rows: 5 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## chr (1): Instituição
## dbl (2): Tempo de atuação, Número de clientes
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df5
```

```
## # A tibble: 5 x 3
```

```
##   Instituição `Tempo de atuação` `Número de clientes`
##   <chr>                     <dbl>                 <dbl>
## 1 A                            25                   102
## 2 B                            32                   121
## 3 C                            28                    80
## 4 D                            53                   181
## 5 E                            44                   132
```

QUESTÕES:

a) Construa o diagrama de dispersão dos dados.

b) Calcule a covariância e o coeficiente de correlação.

Item 6 (dataset 6)

Os preços de fechamento de diversos ativos negociados na BOVESPA aparecem listados na planilha portfolio.

APURAÇÃO

```
df6 <- read_csv2("C:/Users/DELL/OneDrive/R/Rprojetos/ufpr_ppgecon/estatistica/data/Exerc_1_descritiva_da
```

```
## i Using "','" as decimal and "'.'" as grouping mark. Use `read_delim()` for more control.
```

```
## Rows: 599 Columns: 15
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## chr  (1): date
## dbl (14): PETR4, VALE5, CSNA3, GGBR4, USIM5, JBSS3, MRFG3, BEEF3, LAME4, AMBV4, NATU3, ITUB4, BBDC4,
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df6
```

```
## # A tibble: 599 x 15
##    date       PETR4 VALE5 CSNA3 GGBR4 USIM5 JBSS3 MRFG3 BEEF3 LAME4 AMBV4 NATU3 ITUB4 BBDC4 BBAS3
##    <chr>      <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 20/05/2008  47.2  54.1  37.7  40.0  42.6  8.31  20.2  10.3  12.8  117.  18.3  32.3  30.4  25.9
##  2 21/05/2008  48.0  52.6  36.7  38.9  41.1  8.56  20.5  10.2  12.5  114.  17.2  31.5  29.7  25.1
##  3 23/05/2008  46.2  51.9  36.7  38.6  41.0  8.46  20.6   9.8  12.5  111.  16.6  31.4  29.7  24.4
##  4 26/05/2008  47.1  51.6  37.1  39.1  41.3  8.42  20.8  10.1  12.5  111.  16.7  31.4  29.3  24.4
##  5 27/05/2008  45.6  50.5  36.5  39.2  40.5  8.61  21.4  10.3  12.2  110.  16.4  32.3  29.9  25.0
##  6 28/05/2008  46.6  51.6  37.6  40.1  41.4  9.26  22.6  10.6  12.6  107.  17.2  33.4  31.0  26.5
##  7 29/05/2008  45.1  49.3  36.2  38.5  40.4  9.64  22.5  10.8  12.8  105.  17.7  33.2  30.4  26.7
##  8 30/05/2008  44.8  50.2  35.6  39.0  40.5  9.84  23.0  11.2  12.9  105.  17.1  34.5  30.9  28.0
##  9 02/06/2008  45.5  50.2  35.5  38.5  40.2  9.64  22.5  11.2  12.4  103.  16.7  33.0  30.2  27.2
## 10 03/06/2008  43.3  48.5  34.6  38.3  39.8  9.13  22.7  11.3  12.5  100.  16.5  32.1  29.5  26.2
## # i 589 more rows
```

QUESTÕES:

a) Se você fosse comprar somente um dos papéis dentre os listados, qual seria a melhor escolha com base no período analisado?

b) Calcule o risco e o retorno de uma carteira formada com 50% de PETR4 e 50% de VALE5. Simule o resultado para diversos níveis de correlação.

c) Encontre os papéis com menor correlação.

d) Com base nesses dois papéis, qual percentual do seu capital você aplicaria em cada ação para obter a carteira de menor risco?

e) Os retornos parecem seguir uma curva normal?