CIÊNCIA DE DADOS COM LINGUAGEM R

Richard Guilherme dos Santos

Contents

Bem vindo!

Me chamo Richard, sou matemático e atualmente faço mestrado em estatística pelo programa PIPGEs.

Caso queira me avisar sobre algum problema no livro, erro de digitação ou dúvida, aqui estão minhas redes sociais:

Chapter 1

Introdução

Este livro tem como objetivo servir como guia para as aulas do curso Ciência de Dados com R. Nele apresentaremos os conceitos de:

1. **Estatística Básica:** Nesta parte do curso abordaremos conceitos de estatística como variáveis, tipos de distribuições discretas e contínuas, medidas descritivas e distribuição normal.
2. **Manipulação de dados no R:** Neste tópico serão abordados as principais formas de manipulação de dados utilizando a linguagem R, com ênfase nas bibliotecas dplyr e tidyr. Além disso, abordaremos a criação de gráficos pelo pacote ggplot2.
3. **Modelos de Regressão Linear:** Parte final do curso, onde o aluno aprenderá sobre diagrama de dispersão, coeficiente de correlação linear, regressão linear simples, múltipla e regressão logística, ganhando a capacidade de começar a criar modelos utilizando a linguagem R.

Chapter 2

Introdução ao R

Aqui introduziremos alguns comandos da linguagem `R`, onde utilizamos funções para realizar operações que vão desde leitura e manipulação de dados a operações matemáticas.

Comecemos criando um vetor de números:

```
x <- c(1,3,2,5)
# x = c(1,3,2,5) # Também podemos utilizar "=" para atribuir variáveis
x
```

```
## [1] 1 3 2 5
```

O comando acima combina os números 1,3,2 e 5 em um vetor de números e os salva em um objeto denominado x. Escrevemos x para recebermos os atributos do vetor.

A partir disto podemos utilizar outras funções para calcularmos informações destes atributos, como o tamanho de um vetor:

```
length(x)
```

```
## [1] 4
```

sua média:

```
mean(x)
```

```
## [1] 2.75
```

também podemos realizar operações entre os vetores:

```
a <- c(1,2,3)
b <- c(2,3,4)
a+b
```

```
## [1] 3 5 7
```

Há outros tipos de objetos que podem ser criados quando trabalhamos com R. Dentre os mais importantes para manipulação de dados estão as matrizes:

```
mat = matrix(data = c(1,2,3,4), nrow = 2, ncol = 2,
             byrow = TRUE)
mat
```

```
##      [,1] [,2]
## [1,]    1    2
## [2,]    3    4
```

muitos devem já estar familiarizados com estas. A linguagem R fornece as mais diversas operações entre matrizes:

```
a = matrix(data = 1:9, nrow = 3, ncol = 3)
b = matrix(data = 1:9, nrow = 3, ncol = 3, byrow = T)
# a + b # Soma de matrizes
# a * b # Multiplicação dos elementos das matrizes termo a termo
# a %*% b # Multiplicação de matrizes
# t(a) # Transposta da matriz
# det(a) # Determinante de uma matriz
# solve(a) # Inversa da matriz
sqrt(a) # Raiz quadrada dos elementos da matriz
```

```
##          [,1]     [,2]     [,3]
## [1,] 1.000000 2.000000 2.645751
## [2,] 1.414214 2.236068 2.828427
## [3,] 1.732051 2.449490 3.000000
```

> Funções aceitam os mais diversos tipos de argumentos. Para termos uma ideia de quais utilizarmos e seus respectivos atributos devemos fazer consultas em suas bibliotecas:

```
help(matrix)
```

Além disso, para armazenamento de dados temos os data.frames, tabelas que aceitam dados de tipos distintos:

```
nomes = c('Carol', 'Alfredo', 'Godoberto')
idade = c(18, 23, 19)
peso = c(69, 75, 80)
altura = c(1.70, 1.80, 1.85)
ICM = peso/altura^2
df = data.frame(nomes, idade, peso, altura, ICM)
df
```

```
##       nomes idade peso altura      ICM
## 1     Carol    18   69   1.70 23.87543
```

```
## 2   Alfredo    23   75   1.80 23.14815
## 3 Godoberto    19   80   1.85 23.37473
```

Chapter 3

Medidas Descritivas

> **Importante:** A partir deste capítulo utilizaremos a função `kable` do pacote `knitr` para visualização de conjuntos de dados. Na prática isto não é necessário, apenas o realizamos para efeitos de visualização.

3.1 Tipos de Variáveis

Antes de analisarmos conjuntos de dados, é necessário termos um conhecimento sobre tipos de variáveis. Para isto, consideremos a seguinte tabela:

```
nome = c('Djoko','Wilson','Leon', 'Nilce')
est_civil = c('Solteiro','Casado', 'Casado', 'Casado')
escolaridade = c('Pós-graduação',
                 'Ensino médio completo',
                 'Pós-graduação',
                 'Superior completo')
n_filhos = c(0, 0, 0, 0)
salario = c(4500, 3000, 2000, 5500)
idade = c(29, 33, 39, 32)
df_youtubers = data.frame(nome, est_civil, escolaridade, n_filhos, salario, idade)
kable(df_youtubers, align = 'c',
      caption = 'Dados sobre Youtubers.') # Melhor visualização dos dados para este PDF
```

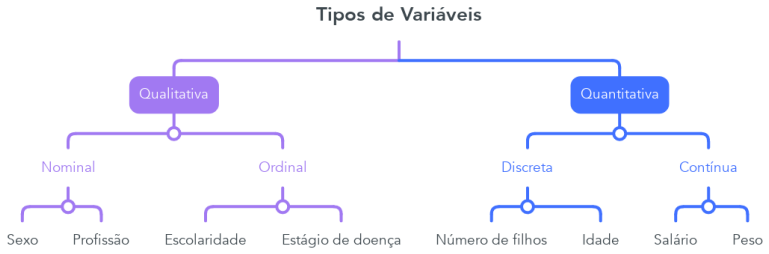
Variáveis como sexo, escolaridade e estado civil apresentam realizações de uma qualidade ou atributo do indivíduo pesquisado, enquanto outras como número de filhos, salário e idade apresentam números como resultados de uma contagem ou mensuração. Chamamos as do primeiro tipo de **qualitativas** e as do segundo de **quantitativas**

Cada uma das duas ainda pode ser dividida em dois tipos:

Table 3.1: Dados sobre Youtubers.

nome	est_civil	escolaridade	n_filhos	salario	idade
Djoko	Solteiro	Pós-graduação	0	4500	29
Wilson	Casado	Ensino médio completo	0	3000	33
Leon	Casado	Pós-graduação	0	2000	39
Nilce	Casado	Superior completo	0	5500	32

- **Variável qualitativa nominal:** atributos não apresentam uma ordem lógica;
- **Variável qualitativa ordinal:** atributos apresentam uma ordem lógica bem estabelecida;
- **Variável quantitativa discreta:** dados de contagem, assumem apenas valores inteiros;
- **Variável quantitativa contínua:** dados que podem assumir qualquer tipo de valor.



Muitas vezes queremos resumir estes dados, apresentando um ou mais valores que sejam representativos da série toda. Neste contexto entram às **medidas de posição** e **dispersão**.

3.2 Medidas de Posição

Usualmente utilizamos uma das seguintes medidas de posição (ou localização): **média, mediana ou moda**. Vamos as suas definições:

- A uma variável atribuiremos a letra X enquanto para seus elementos os valores $x_1, \ldots, x_n$, sendo n o seu total de elementos.
- **Moda:** valor mais frequente do conjunto de valores observados.
- **Mediana:** valor que ocupa a posição central das observações quando estas estão ordenadas em ordem crescente.

  - Quando o número de observações for **par**, usa-se como mediana a média aritmética das duas observações centrais.

Na tabela 3.1 temos a seguinte mediana para a coluna salário:

```
median(df_youtubers$salario)
```

```
## [1] 3750
```

- Matematicamente ordenamos os dados do menor para o maior: $2000, 3000, 4500, 5500$, selecionamos as observações centrais 3000 e 4500. Por fim, calculamos a média artimética de ambas, $\frac{3000+4500}{2}$, para obtermos a mediana.

Além disso, podemos calcular a mediana para todas as colunas:

```
# apply: aplica uma função a um conjunto de dados
# MARGIN = 2: 1 para aplicar a função a todas as linhas e 2 a todas as colunas
# FUN: função a ser aplicada ao conjunto de dados
apply(df_youtubers[, c('n_filhos','salario','idade')], MARGIN = 2, FUN = median)
```

```
## n_filhos  salario    idade
##      0.0   3750.0     32.5
```

- **Média:** soma de todos os elementos do conjunto dividida pela quantidade de elementos do conjunto

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Na tabela 3.1 temos a seguinte média para o salário:

```
mean(df_youtubers$salario)
```

```
## [1] 3750
```

Podemos calcular para todas as colunas que possuam valores numéricos:

```
colMeans(df_youtubers[, c('idade', 'salario')])
```

```
##   idade salario
##   33.25 3750.00
```

3.3 Medidas de Dispersão

O **resumo** de um conjunto de dados por uma única medida representativa de posição esconde toda a informação sobre a variabilidade de um conjunto de observações. Consideremos que cinco alunos realizaram cinco provas, obtendo as seguintes notas:

```
nomes = c('alunoA', 'alunoB', 'alunoC',
          'alunoD', 'alunoE')
notas = matrix(c(3,4,5,6,7,
                 1,3,5,7,9,
                 2,5,5,5,8,
                 3,5,5,5,7,
                 0,0,5,10,10), nrow = 5, ncol = 5, byrow = T)
df_alunos = data.frame(notas, row.names = nomes)
colnames(df_alunos) = c('P1', 'P2', 'P3', 'P4', 'P5')
kable(df_alunos, align = 'c')
```

	P1	P2	P3	P4	P5
alunoA	3	4	5	6	7
alunoB	1	3	5	7	9
alunoC	2	5	5	5	8
alunoD	3	5	5	5	7
alunoE	0	0	5	10	10

Temos as seguintes médias para os alunos:

```
# Podemos ver a média de cada coluna utilizando colMeans(df_alunos)
rowMeans(df_alunos)
```

```
## alunoA alunoB alunoC alunoD alunoE
##      5      5      5      5      5
```

Cada aluno possui a mesma média de notas, porém, isto não informa nada sobre a diferença na **variabilidade das notas.** A partir disto, são criadas medidas que sumarizam a **variabilidade** de um conjunto de observações.

Uma primeira ideia é considerar a soma das diferença dos dados em relação a média:

$$x_1 - \overline{x} + x_2 - \overline{x} + \cdots + x_n - \overline{x}$$

Porém, podemos mostrar que em qualquer conjunto a soma destes desvios é igual a zero. Uma alternativa é então adicionar o valor absoluto em cada diferença:

$$|x_1 - \overline{x}| + |x_2 - \overline{x}| + \cdots + |x_n - \overline{x}|$$

Apesar de possuir uma boa interpretabilidade, tal métrica não possui propriedades matemáticas interessantes. Assim, estatísticos trabalham com a diferença dos dados em relação a média ao quadrado:

$$(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2$$

Como muitas vezes queremos comparar conjuntos de dados de diferentes tamanhos, realizamos a divisão desta soma pelo total de elementos em uma amostra e a este número chamamos de **variância**:

$$\text{var}(X) = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n}$$

E a partir disto, definimos **desvio padrão** como sendo a raiz da variância:

$$\text{dp} = \sqrt{\text{var}(X)}$$

Realizamos isto pois caso os dados estejam em uma certa unidade de medida, como cm , ao calcularmos a variância passamos a trabalhar com cm^2, o que dificulta a interpretabilidade dos resultados. Utilizando o valor na raiz quadrada, voltamos a trabalhar com a unidade de medida utilizada.

3.4 Quantis Empíricos

Tanto a **média** como o **desvio padrão** podem não ser medidas adequadas para representar um conjunto de dados, uma vez que:

- São afetados por **valores extremos;**
- Apenas os dois valores não dão informação sobre a **simetria** ou **assimetria** da distribuição dos dados

Vimos que a **mediana** define uma divisão dos dados em duas metades. Além dela existem medidas chamadas de **quantil de ordem p** ou **p-quantil** indicado por $q(p)$ onde p é uma proporção qualquer, $0 < p < 1$ tal que 100% das observações sejam menores do que $q(p)$.

Abaixo temos alguns dos nomes dos quantis mais utilizados:

- $q(0.25) = q_1$: **1° Quartil** ou **25° Percentil**
- $q(0.50) = q_2$: **2° Quartil**, **Mediana** ou **50° Percentil**
- $q(0.75) = q_3$: **3° Quartil** ou **75° Percentil**
- $q(0.40)1$: **4° Decil**
- $q(0.95)$: **95° Percentil**

No `R` podemos visualizar os quartis da seguinte forma:

```
quantile(df_alunos$P1)
```

```
##   0%  25%  50%  75% 100%
##    0    1    2    3    3
```

Em várias colunas:

```
apply(df_alunos, 2, quantile, seq(0,1,.2))
```

```
##       P1  P2 P3   P4   P5
## 0%   0.0 0.0  5  5.0  7.0
## 20%  0.8 2.4  5  5.0  7.0
## 40%  1.6 3.6  5  5.6  7.6
## 60%  2.4 4.4  5  6.4  8.4
## 80%  3.0 5.0  5  7.6  9.2
## 100% 3.0 5.0  5 10.0 10.0
```

3.5 Box Plot

A informação contida nos quantis pode ser confusa quando estamos observando vários conjuntos de dados. A partir disto a traduzimos em um diagrama, qual é chamado de **box plot:**

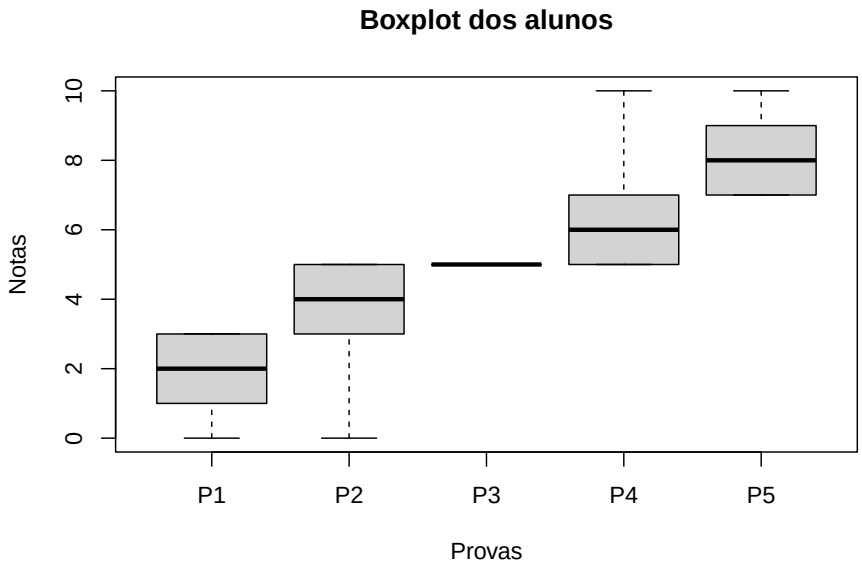
Para construção dessa gráfico definimos por **intervalo interquartil** o valor:

$$\text{IQ}(X) = q_3 - q_1$$

Desenhamos um retângulo que parte do primeiro quartil até o terceiro, com a mediana sendo representada por uma linha em seu interior. A partir do retângulo desenhamos uma linha até o maior ponto que não exceta o valor $q_3 + 1.5 \cdot \text{IQ}(X)$, chamado de limite superior. De modo análogo fazemos o mesmo procedimento até a parte inferior do retângulo considerando o valor $q_1 + 1.5 \cdot \text{IQ}(X)$ chamado de limite interior. As observações que estiverem acima do limite superior ou abaixo do limite superior são chamados de pontos exteriores e representadas por asteriscos. Essas observações podem ser chamaas de outliers ou valores atípicos.

Modo simples de como realizar um boxplot pelo R:

```
boxplot(df_alunos, xlab = "Provas", ylab = "Notas",
        main = "Boxplot dos alunos")
```



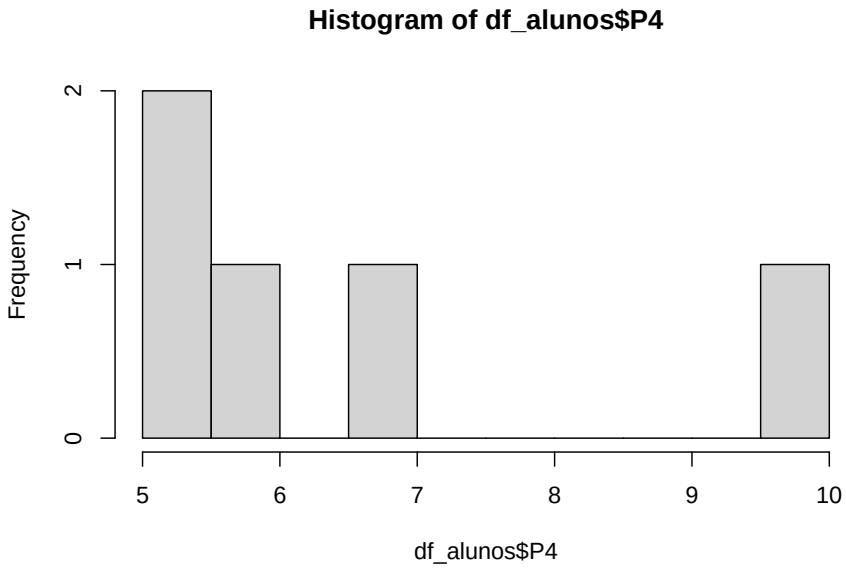
O aluno mais atento pode se perguntar: porque alguns dos boxplots não possuem a linha superior e/ou inferior? Isto ocorre quando temos muitos dados em uma mesma categoria, com o primeiro ou terceiro quartil tendo o mesmo valor que o mínimo ou máximo do conjunto de dados:

```
apply(df_alunos, 2, quantile)
```

```
##      P1 P2 P3 P4 P5
## 0%    0  0  5  5  7
## 25%   1  3  5  5  7
## 50%   2  4  5  6  8
## 75%   3  5  5  7  9
## 100%  3  5  5 10 10
```

O **box plot** dá uma ideia de posição, dispersão, assimetria dos dados.

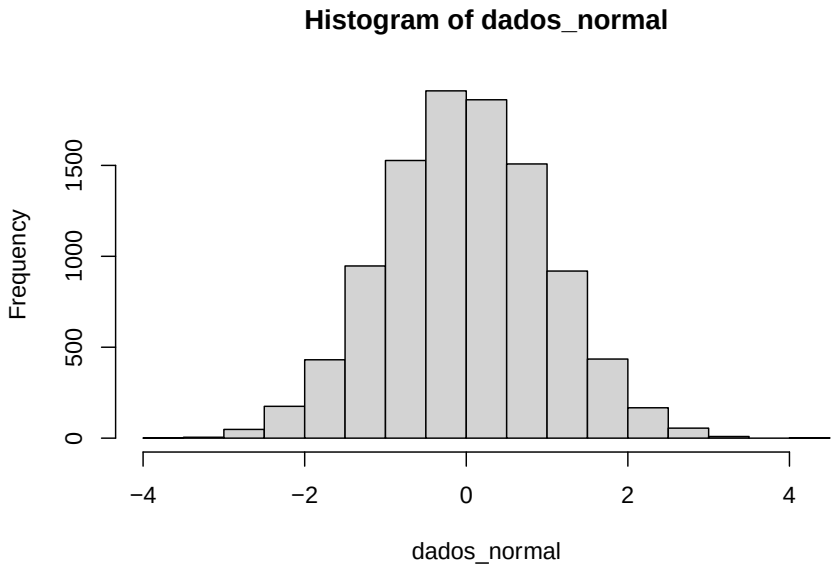
```
hist(df_alunos$P4, breaks = seq(5, 10, 0.5))
```



3.6 Transformações

Vários procedimentos estatísticos são baseados na posição que os dados possuem uma distribuição em forma de sino (distribuição normal) ou que a distribuição seja mais ou menos simétrica:

```
# Simula 500 dados de uma distribuição normal
dados_normal <- rnorm(n = 10000)
# Gráfico de suas frequências
hist(dados_normal)
```

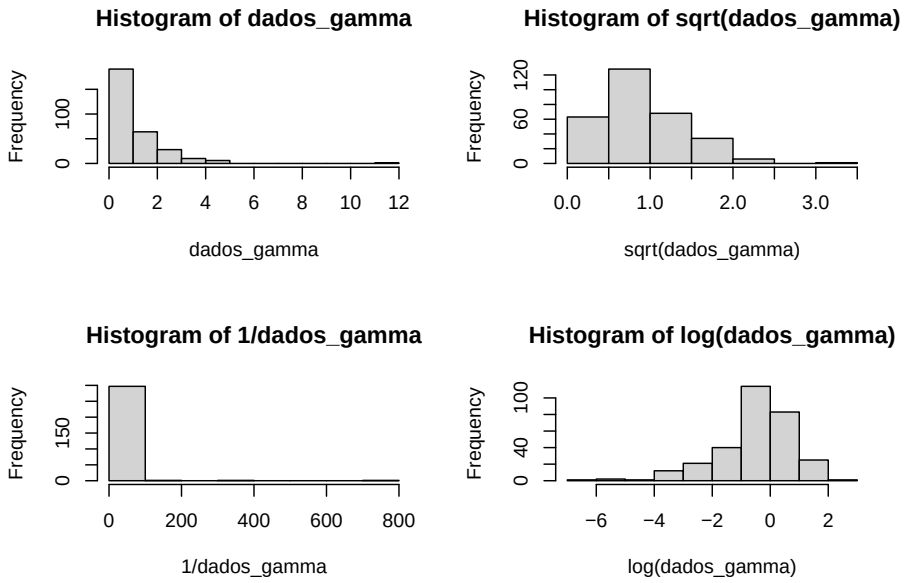


Se quisermos utilizar tais procedimentos podemos efetuar transformações nas observações, de modo a se obter uma distribuição mais simétrica e próxima da normal. As transformações mais frequentemente utilizadas são:

$$x = \begin{cases} \sqrt{x} \\ \ln(x) \\ \frac{1}{x} \end{cases}$$

para cada transformação obtemos gráficos apropriados para os dados originais e transformados, de modo a escolhermos o valor mais adequado de p.

```
dados_gamma <- rgamma(n = 300, shape = 1)
par(mfrow = c(2,2)) # MultiFrame rowwise layout
hist(dados_gamma)
hist(sqrt(dados_gamma))
hist(1/dados_gamma)
hist(log(dados_gamma))
```



3.7 Lab 01 - Conjunto de dados Iris

O conjunto de dados Iris é um dos mais utilizados quando introduzimos conceitos de ciência de dados. Este pode ser encontrado em UCI Machine Learning Repository. Tal conjunto consiste de 150 amostras de 4 tipos de espécies de flores distintas contendo os atributos:

- SepalLengthCm
- SepalWidthCm
- PetalLengthCm
- PetalWidthCm

Podemos acessá-lo no R sem nenhum carregamento prévio da seguinte forma:

```
# A função head() mostra os cinco primeiros itens de data.frame:
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Há certas boas práticas ao carregar um conjunto de dados, dentre elas temos:

- Visualização de sua dimensão:

```
# O primeiro valor é a quantidade de linhas do conjunto de dados
# e o segundo a sua quantidade de atributos
dim(iris)
```

```
## [1] 150   5
```

- Visualização do tipo de cada atributo:

```
str(iris) # Structure of an Arbitrary R Object
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

- Sumário de seus atributos:

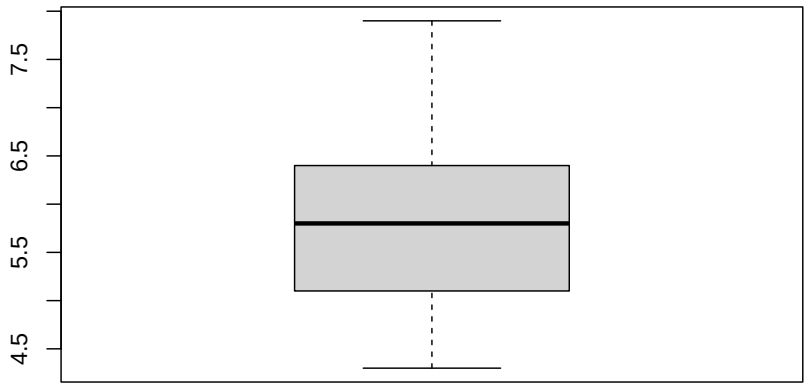
```
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

Dessa maneira poderemos contatar valores errôneos no conjunto de dados, distribuições de variáveis categóricas e ter um melhor contato com o conjunto de dados.

Há ainda diversas maneiras de realizarmos visualizações desse conjunto no `R`, observemos o boxplot da variável Sepa.Length:

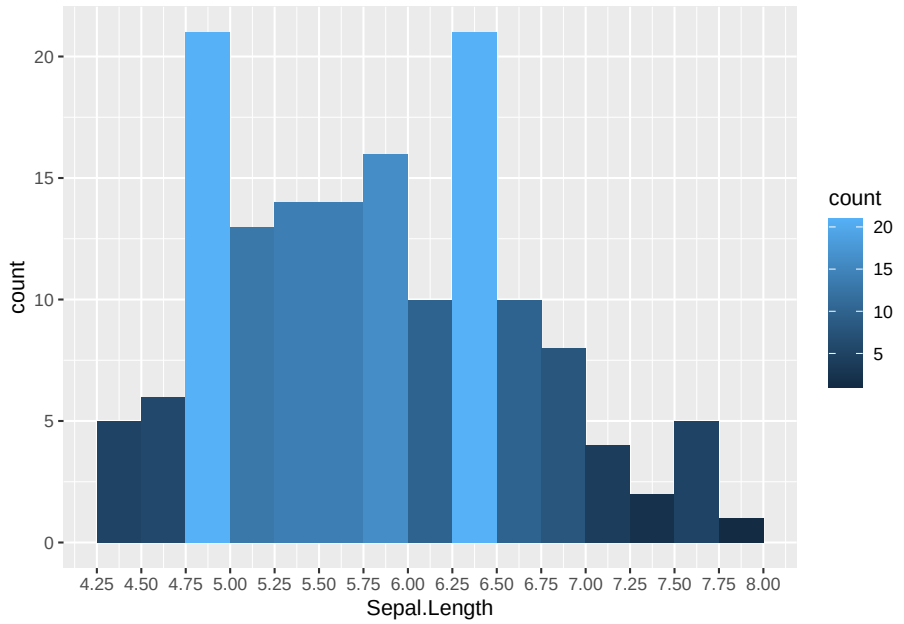
```
boxplot(iris$Sepal.Length)
```



Observamos que não há presença de outliers, além disso, como a parte debaixo do retângulo separado pela linha que representa a mediana é menor, isto indica que a distribuição dos dados é ligeiramente assimétrica, o qual é confirmado pelo histograma:

```
hist(iris$Sepal.Length)
```

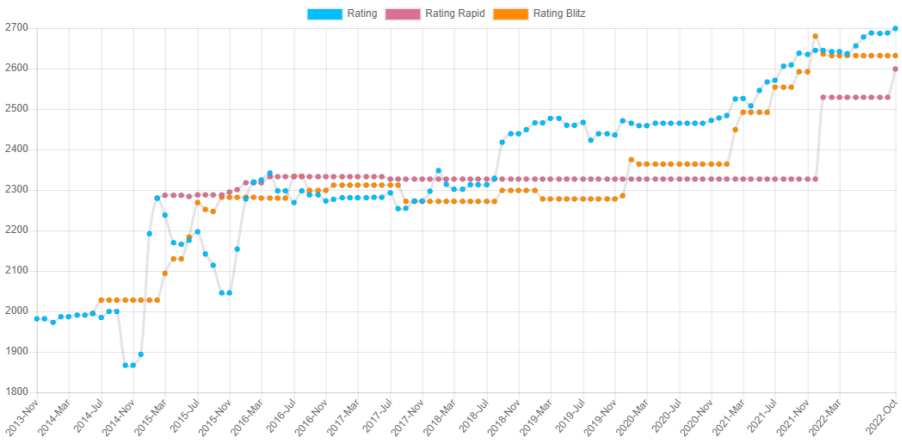

```
ggplot(data = iris, aes(x = Sepal.Length, fill = ..count..)) +
  geom_histogram(binwidth = 0.25, boundary = 0) +
  scale_x_continuous(breaks = seq(1, 10, by = 0.25))
```



3.8 Lab 02 - Xadrez Brasil

Em 2022 houve uma polêmica no universo de xadrez entre o atual campeão mundial de xadrez Magnus Carlsen e o jovem grande mestre Hanns Niemann, qual possui um histórico de trapaças e foi acusado de repetir estes atos no torneio. Magnus chegou a abandonar o torneio motivado por acreditar que seu oponente utilizava de engines em sua partida.

Baseado nisso vários estatísticos se debrulharam entre as partidas de Niemann com o objetivo de encontrar evidências de sua rápida ascensão no xadrez, qual bateu recordes como grande mestre mais rápido da história.



Uma destas analises foi realizada pelo canal brasileiro de youtube Xadrez Brasil, e nela podemos observar como medidas de posição e dispersão podem ser utilizadas para fortalecer ou não a hipótese de que Niemann estava trapaçeando.

- https://www.youtube.com/watch?v=60QPEGsOCyw

3.9 Projeto 01 - Machine Learning from Disaster

Todo mundo já assistiu, ou pelo menos ouviu falar, sobre o desastre do navio Titanic.

Incrivelmente este caso também pode ser estudado utilizando aprendizado de máquina! Na verdade este é um dos primeiros desafios que trabalhamos quando estudamos nossos primeiros algoritmos. O conjunto de dados e suas informações pode ser encontrado no site Kaggle, um site que hospeda diversos conjuntos de dados e competições de machine learning.

Na aula aprenderemos como baixamos e analisamos as observaçõs contidas nesse conjunto de dados, qual pode ser visualizado na tabela abaixo:

```
library(readr)
titanic_train <- read_csv("G:/Meu Drive/Dados/titanic_train.csv")
```

```
## Rows: 891 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Name, Sex, Ticket, Cabin, Embarked
## dbl (7): PassengerId, Survived, Pclass, Age, SibSp, Parch, Fare
##
## 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.
```

```
kable(head(titanic_train), align = 'c')
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp
1	0	3	Braund, Mr. Owen Harris	male	22	1
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1
3	1	3	Heikkinen, Miss. Laina	female	26	0
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1
5	0	3	Allen, Mr. William Henry	male	35	0
6	0	3	Moran, Mr. James	male	NA	0

O objetivo do trabalho é prever se um passageiro sobreviveu ou não no naufrágio do Titanic. Para começarmos trabalhando podemos utilizar a função `table` para observar a distribuição dos passageiros que sobreviveram ao acidente:

```
table(titanic_train$Survived)
```

```
##
##   0   1
## 549 342
```

Observar a média de idade entre os passageiros:

```
mean(titanic_train$Age, na.rm = TRUE)
```

```
## [1] 29.69912
```

Quais são os tipos de variáveis que eu tenho no conjunto de dados?

```
str(titanic_train)
```

```
## spec_tbl_df [891 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ PassengerId: num [1:891] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : num [1:891] 0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : num [1:891] 3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr [1:891] "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Br
##  $ Sex        : chr [1:891] "male" "female" "female" "female" ...
##  $ Age        : num [1:891] 22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : num [1:891] 1 1 0 1 0 0 0 3 0 1 ...
```

```
##  $ Parch      : num [1:891] 0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr [1:891] "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num [1:891] 7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr [1:891] NA "C85" NA "C123" ...
##  $ Embarked   : chr [1:891] "S" "C" "S" "S" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   PassengerId = col_double(),
##   ..   Survived = col_double(),
##   ..   Pclass = col_double(),
##   ..   Name = col_character(),
##   ..   Sex = col_character(),
##   ..   Age = col_double(),
##   ..   SibSp = col_double(),
##   ..   Parch = col_double(),
##   ..   Ticket = col_character(),
##   ..   Fare = col_double(),
##   ..   Cabin = col_character(),
##   ..   Embarked = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Por fim, vamos selecionar

```
kable(summary(titanic_train), align = 'c')
```

	PassengerId	Survived	Pclass	Name	Sex	Age
	Min. : 1.0	Min. :0.0000	Min. :1.000	Length:891	Length:891	Min. : 0
	1st Qu.:223.5	1st Qu.:0.0000	1st Qu.:2.000	Class :character	Class :character	1st Qu.:2
	Median :446.0	Median :0.0000	Median :3.000	Mode :character	Mode :character	Median :2
	Mean :446.0	Mean :0.3838	Mean :2.309	NA	NA	Mean :29
	3rd Qu.:668.5	3rd Qu.:1.0000	3rd Qu.:3.000	NA	NA	3rd Qu.:3
	Max. :891.0	Max. :1.0000	Max. :3.000	NA	NA	Max. :80
	NA	NA	NA	NA	NA	NA's :1

Chapter 4

Tipos de Distribuições Discretas

Para atender a situações mais práticas, é necessário expandir os conceitos relacionados a probabilidade de forma que tenhamos modelos probabilísticos que representem todos os tipos de variáveis. Neste capítulo trabalharemos com variáveis quantativas discretas.

Exemplo (Bussab):

Chamamos de **variável aleatória discreta** uma função X definida no espaço amostral Ω que assume valores em um conjunto de números finito.

Neste contexto vimos como associar a cada valor x_i da variável aleatória X a sua probabilidade de ocorrência. Matematicamente, escrevemos

Além disso, chamamos de **função de probabilidade** da variável aleatória discreta X a função que a cada valor de x_i associa a sua probabilidade de ocorrência

$$p(x_i) = PX = x_i) = p_i, i = 1, 2, ...$$

4.1 Valor Médio de uma Variável Aleatória

Dada uma variável aleatórai X discreta, assumindo os valores $x_1,\dots, x_n$, chamamos de valor médio ou esperança de X o valor

$$E[X] = \sum_{i=1}^{n} x_i P(X = x_i) = \sum_{i=1}^{n} x_i p_i.$$

Chamamos de variância da variável aleatória X o valor

$$\text{var}[X] = \sum_{i=1}^{n} [x_i - E[X]]^2 p_i$$

Chapter 5

Tipos de Distribuições Contínuas

Chapter 6

Introdução as bibliotecas do R

6.1 Dplyr

6.2 Tidyr

6.3 GGPlot2

Chapter 7

Regressão Linear