

# Relatório sobre a base de dados Student's Performance

---

Charles Barros 14/09/2021

## Banco de dados - StudentsPerformance

---

### Contextualização

---

O banco de dados analisado é um conjunto de dados fictício e deve ser usado apenas para fins de treinamento de ciência de dados. Este conjunto de dados inclui pontuações de três exames e uma variedade de fatores pessoais, sociais e econômicos que têm efeitos de interação sobre eles.

### O intuito da análise

---

O intuito desse relatório é demonstrar alguns conhecimentos com a linguagem de programação R e aplicar alguns conhecimentos adquiridos durante minha jornada acadêmica.

### Pacotes utilizados

---

```
library(tidyverse)
library(corrplot)
```

O **pacote Tidyverse** é uma coleção de pacotes R projetados para ciência de dados. Todos os pacotes compartilham uma filosofia de design, gramática e estruturas de dados subjacentes. Para mais informações [Saiba Mais em] (<https://www.tidyverse.org/>) .


O pacote **corrplot** é utilizado para visualizar uma matriz de correlação. Para mais informações [Saiba mais em] (<https://www.rdocumentation.org/packages/corrplot/versions/0.2-0/topics/corrplot>) .

## Codigo R

---

### Carregando a base de dados

```
dados <- read.csv(file = "StudentsPerformance.csv",header = TRUE,sep = ";",
head(dados, 5)
```



```
##  gender race.ethnicity parental.level.of.education      lunch
## 1 female      group B      bachelor's degree    standard
## 2 female      group C      some college      standard
## 3 female      group B      master's degree    standard
## 4  male      group A      associate's degree free/reduced
## 5  male      group C      some college      standard
##  test.preparation.course math.score reading.score writing.score
## 1              none          72           72           74
## 2      completed          69           90           88
## 3              none          90           95           93
## 4              none          47           57           44
## 5              none          76           78           75
```

### Verificando as variáveis do banco de dados

```
str(dados)
```

```
## 'data.frame':    1000 obs. of  8 variables:
##  $ gender          : Factor w/ 2 levels "female","male": 1
##  $ race.ethnicity   : Factor w/ 5 levels "group A","group B"
##  $ parental.level.of.education: Factor w/ 6 levels "associate's degree
##  $ lunch            : Factor w/ 2 levels "free/reduced",...:
##  $ test.preparation.course : Factor w/ 2 levels "completed","none":
##  $ math.score       : int  72 69 90 47 76 71 88 40 64 38 ...
##  $ reading.score    : int  72 90 95 57 78 83 95 43 64 60 ...
##  $ writing.score     : int  74 88 93 44 75 78 92 39 67 50 ...
```

## Verificando se a base apresenta valores ausentes

```
colSums(is.na(dados))
```

```
##                gender                race.ethnicity
##                0                            0
## parental.level.of.education                lunch
##                0                            0
##      test.preparation.course                math.score
##                0                            0
##                reading.score                writing.score
##                0                            0
```

## Sumario da base dados

```
summary(dados)
```

```
##      gender      race.ethnicity      parental.level.of.education
## female:518    group A: 89    associate's degree:222      free/redu
## male  :482    group B:190    bachelor's degree :118      standard
##                                     group C:319    high school      :196
##                                     group D:262    master's degree   : 59
##                                     group E:140    some college      :226
##                                     some high school :179
## test.preparation.course  math.score    reading.score    writing.scc
## completed:358           Min.   : 0.00    Min.   : 17.00    Min.   : 10
## none      :642           1st Qu.: 57.00    1st Qu.: 59.00    1st Qu.: 57
##                                     Median : 66.00    Median : 70.00    Median : 69
##                                     Mean   : 66.09    Mean   : 69.17    Mean   : 68
##                                     3rd Qu.: 77.00    3rd Qu.: 79.00    3rd Qu.: 79
##                                     Max.   :100.00    Max.   :100.00    Max.   :100
```

## Tabela dos dados

```
table(dados$gender,dados$race.ethnicity)
```

```
##
##           group A group B group C group D group E
##   female      36    104    180    129    69
##   male        53     86    139    133    71
```

```
table(dados$gender,dados$test.preparation.course)
```

```
##
##           completed none
##   female      184  334
##   male        174  308
```

```
table(dados$race.ethnicity,dados$test.preparation.course)
```

```
##
##           completed none
##   group A      31   58
##   group B      68  122
##   group C     117  202
##   group D      82  180
##   group E      60   80
```

```
table(dados$gender,dados$parental.level.of.education)
```

```
##
##           associate's degree bachelor's degree high school master's deg
##   female                116                63                94
##   male                  106                55                102
##
##           some college some high school
##   female                118                91
##   male                  108                88
```

Através das tabelas conseguimos visualizar a proporção dos gêneros ou etnias por variáveis.

## Verificando a média das matérias por gênero e etnia

```
Means_SEX <- dados %>%  
  group_by(gender)%>%  
  summarise_at(vars(math.score,reading.score,writing.score),list(mean = n
```



```
## # A tibble: 2 x 4  
##   gender math.score_mean reading.score_mean writing.score_mean  
##   <fct>         <dbl>         <dbl>         <dbl>  
## 1 female          63.6          72.6          72.5  
## 2 male            68.7          65.5          63.3
```

```
Means_ethnicity <- dados %>%  
  group_by(race.ethnicity)%>%  
  summarise_at(vars(math.score,reading.score,writing.score),list(mean = n
```



```
## # A tibble: 5 x 4  
##   race.ethnicity math.score_mean reading.score_mean writing.score_mear  
##   <fct>         <dbl>         <dbl>         <dbl>  
## 1 group A          61.6          64.7          62.7  
## 2 group B          63.5          67.4          65.6  
## 3 group C          64.5          69.1          67.8  
## 4 group D          67.4          70.0          70.1  
## 5 group E          73.8          73.0          71.4
```

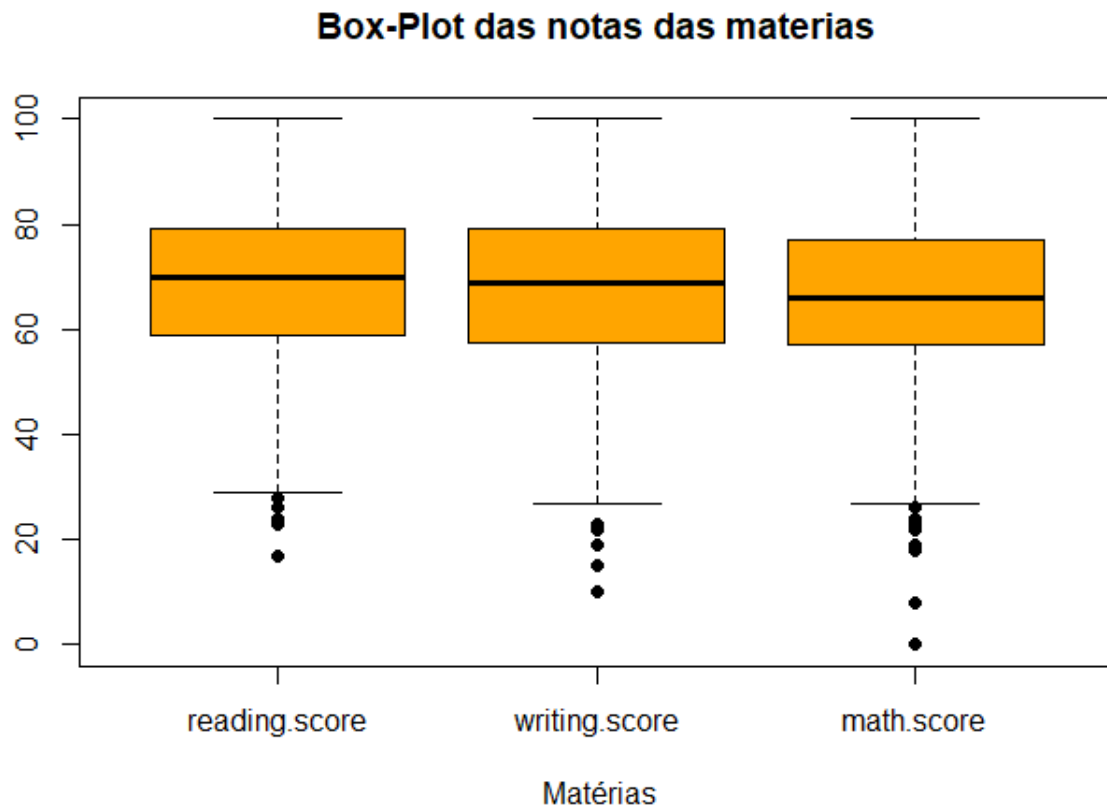


Utilizando o sub pacote (**dplyr**) contido no pacote **tidyverse** conseguimos selecionar as variáveis Scores e agrupar por gênero e etnia, sendo assim conseguimos selecionar a média por gênero e etnia.

## BOX-PLOT

```
dados_scores <- select(dados,reading.score,writing.score,math.score)
```

```
box_scores <- boxplot(dados_scores,  
  main = "Box-Plot das notas das materias",  
  xlab = " Matérias",  
  pch = 16,  
  col = "orange",  
  horizontal = FALSE)
```



Quando criamos um boxplot conseguimos verificar a distribuição dos dados. Sendo assim conseguimos visualizar onde se encontra o centro dos dados (a média ou mediana), amplitude dos dados (máximo-mínimo). Também conseguimos visualizar se os dados apresentam uma distribuição simétrica ou assimétrica e outliers presentes no conjunto de dados (ou distribuição dos dados).

```
box_scores
```

```
## $stats  
##      [,1] [,2] [,3]  
## [1,]  29 27.0  27
```

```
## [2,] 59 57.5 57
## [3,] 70 69.0 66
## [4,] 79 79.0 77
## [5,] 100 100.0 100
## attr(,"class")
## reading.score
## "integer"
##
## $n
## [1] 1000 1000 1000
##
## $conf
##      [,1]      [,2]      [,3]
## [1,] 69.00072 67.92577 65.00072
## [2,] 70.99928 70.07423 66.99928
##
## $out
## [1] 17 26 28 23 24 24 10 22 19 15 23 18 0 22 24 26 19 23 8
##
## $group
## [1] 1 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 3 3 3
##
## $names
## [1] "reading.score" "writing.score" "math.score"
```

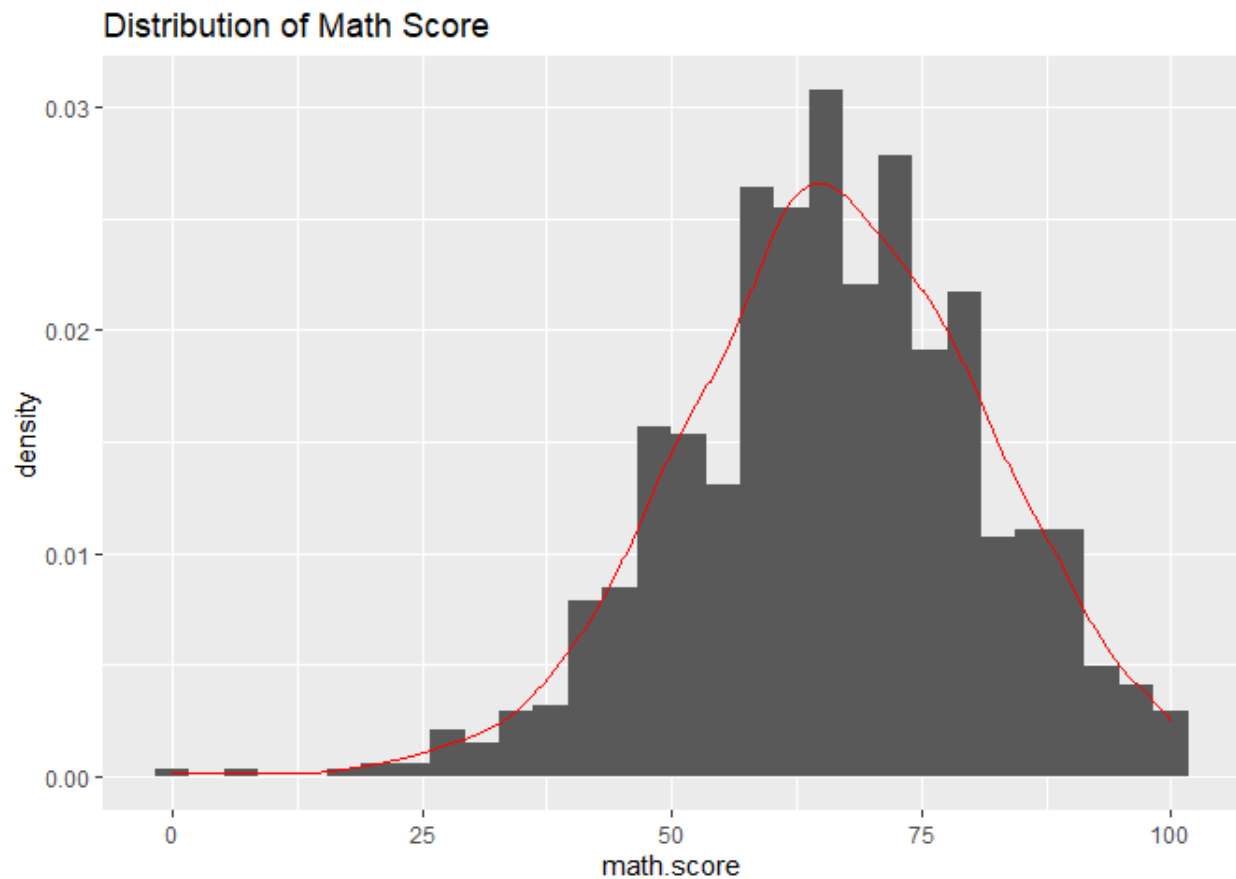
Quando executamos a variável **box\_scores** conseguimos visualizar as características de cada boxplot referente as notas.

## Verificando a distribuição das notas

Para realizarmos qualquer teste (ou técnica) estatística, devemos verificar a distribuição do conjunto de dados que estamos analisando. Sendo assim podemos utilizar alguns conceitos para verificarmos a distribuição dos dados, podemos utilizar histograma ou até mesmo o boxplot para visualizar a distribuição dos dados através de uma figura, porém não devemos apenas tomar uma decisão através de figura, sendo assim devemos utilizar o **teste de Shapiro** para verificarmos se os dados apresentam uma distribuição normal.

```
ggplot(dados, aes(math.score)) +
  geom_histogram(aes(y=..density..)) +
  geom_density(col = "red")+
  ggtitle('Distribution of Math Score')
```

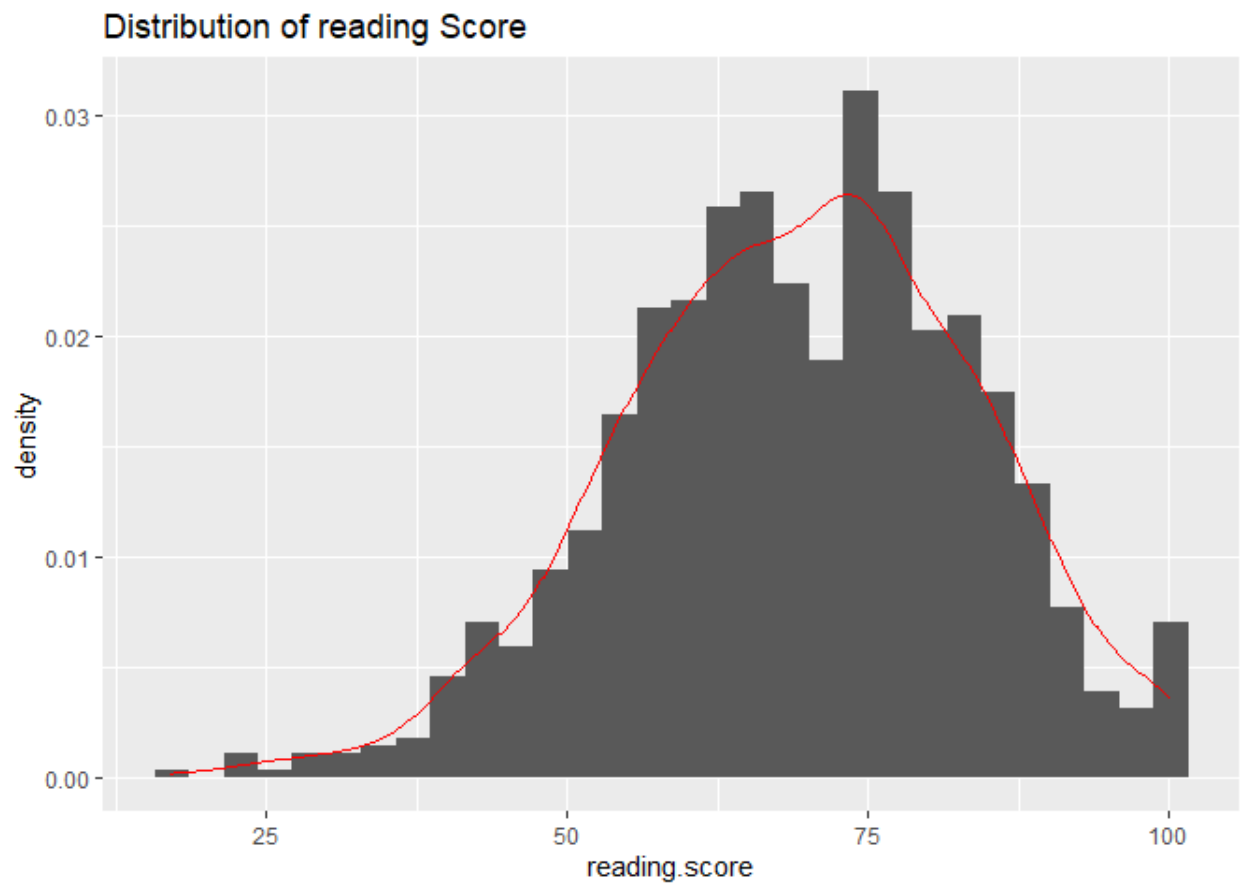
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(dados, aes(reading.score)) +  
  geom_histogram(aes(y=..density..)) +  
  geom_density(col = "red")+  
  ggtitle('Distribution of reading Score')
```

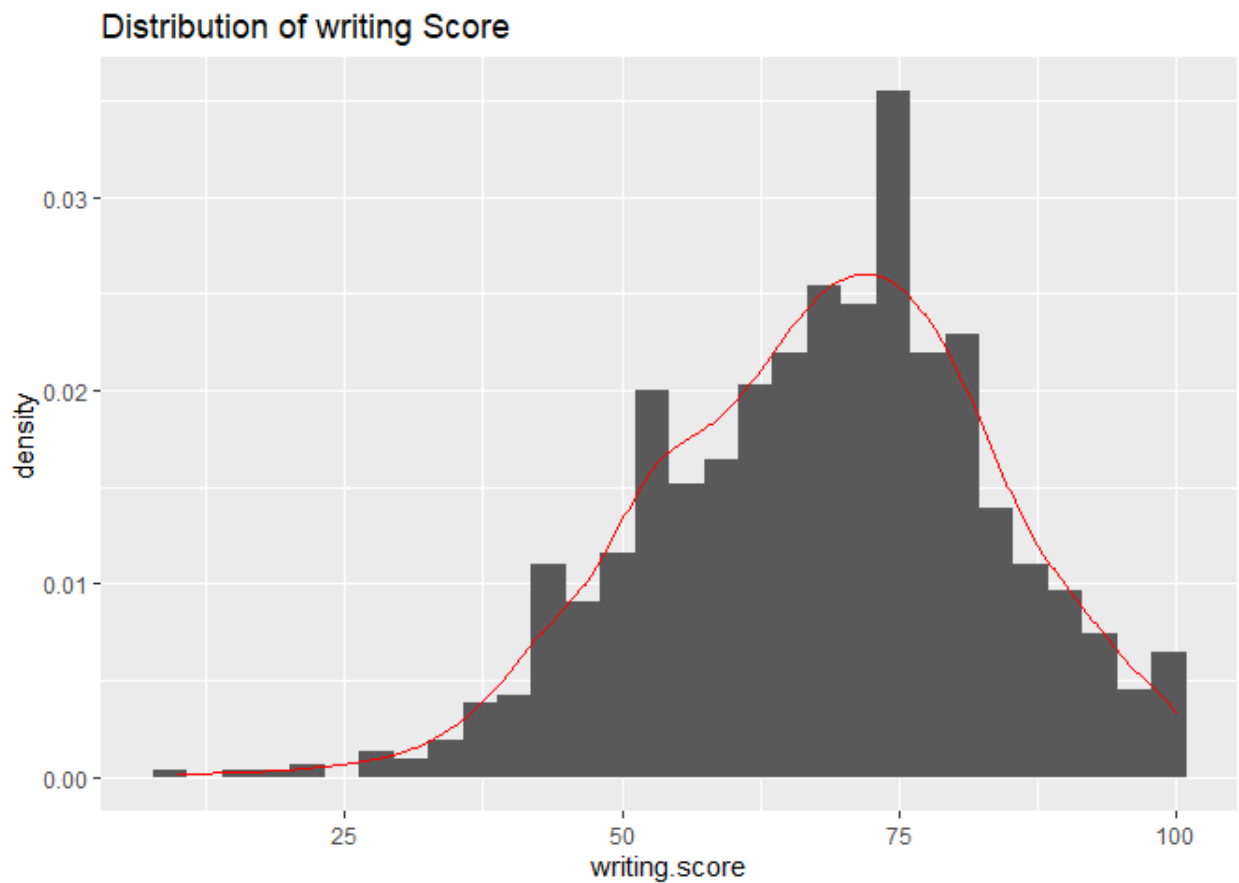
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





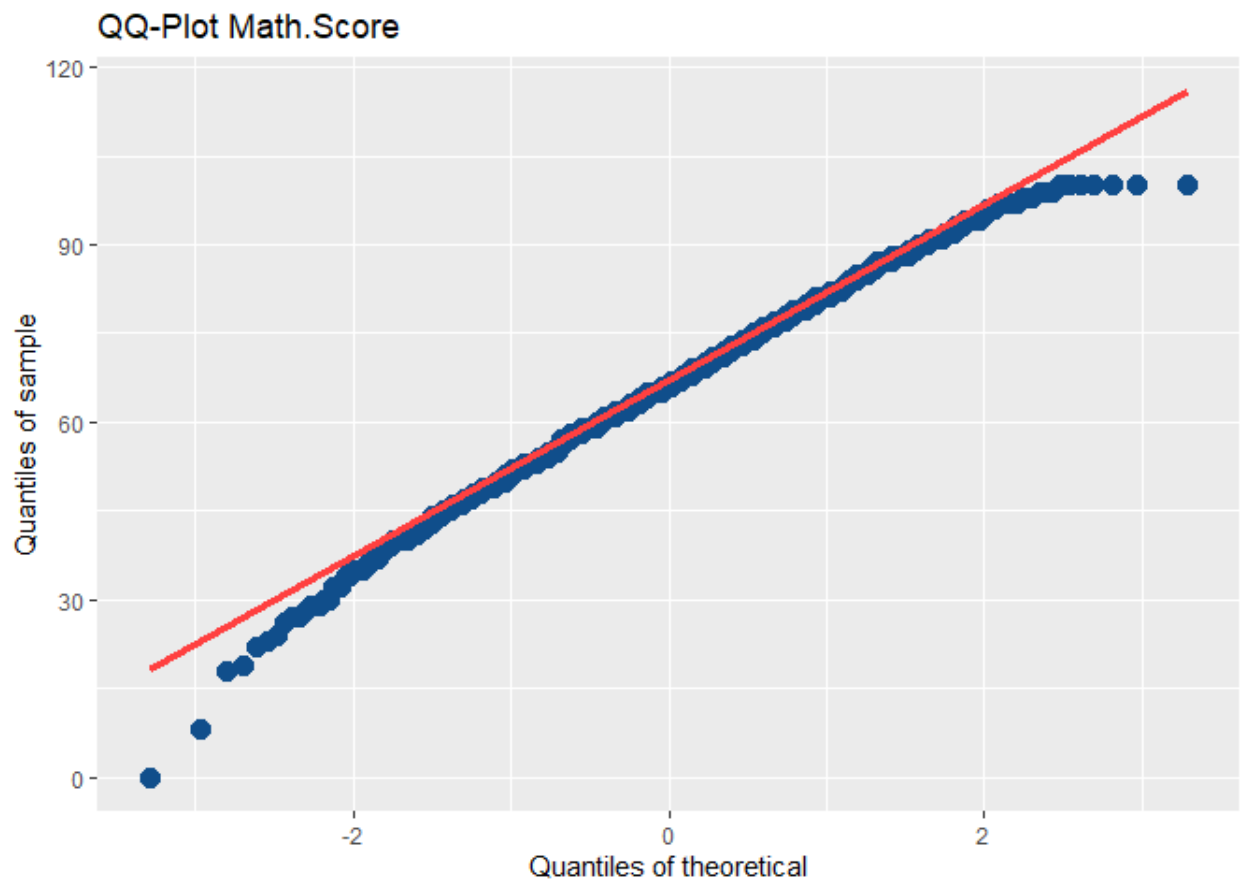
```
ggplot(dados, aes(writing.score)) +  
  geom_histogram(aes(y=..density..)) +  
  geom_density(col = "red")+  
  ggtitle('Distribution of writing Score')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

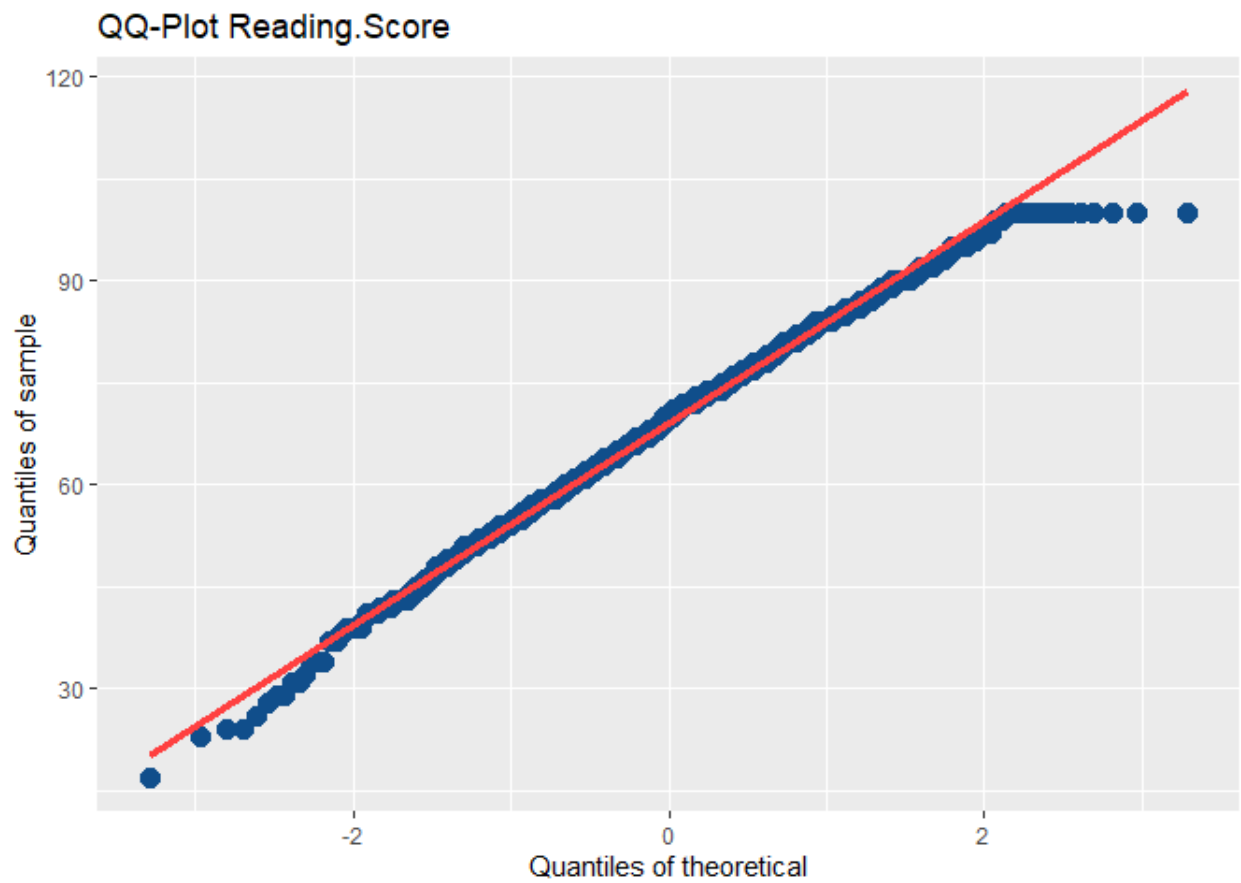


O histograma, também conhecido como distribuição de frequências, é a representação gráfica em colunas ou em barras de um conjunto de dados previamente tabulado e dividido em classes uniformes ou não uniformes. A base de cada retângulo representa uma classe.

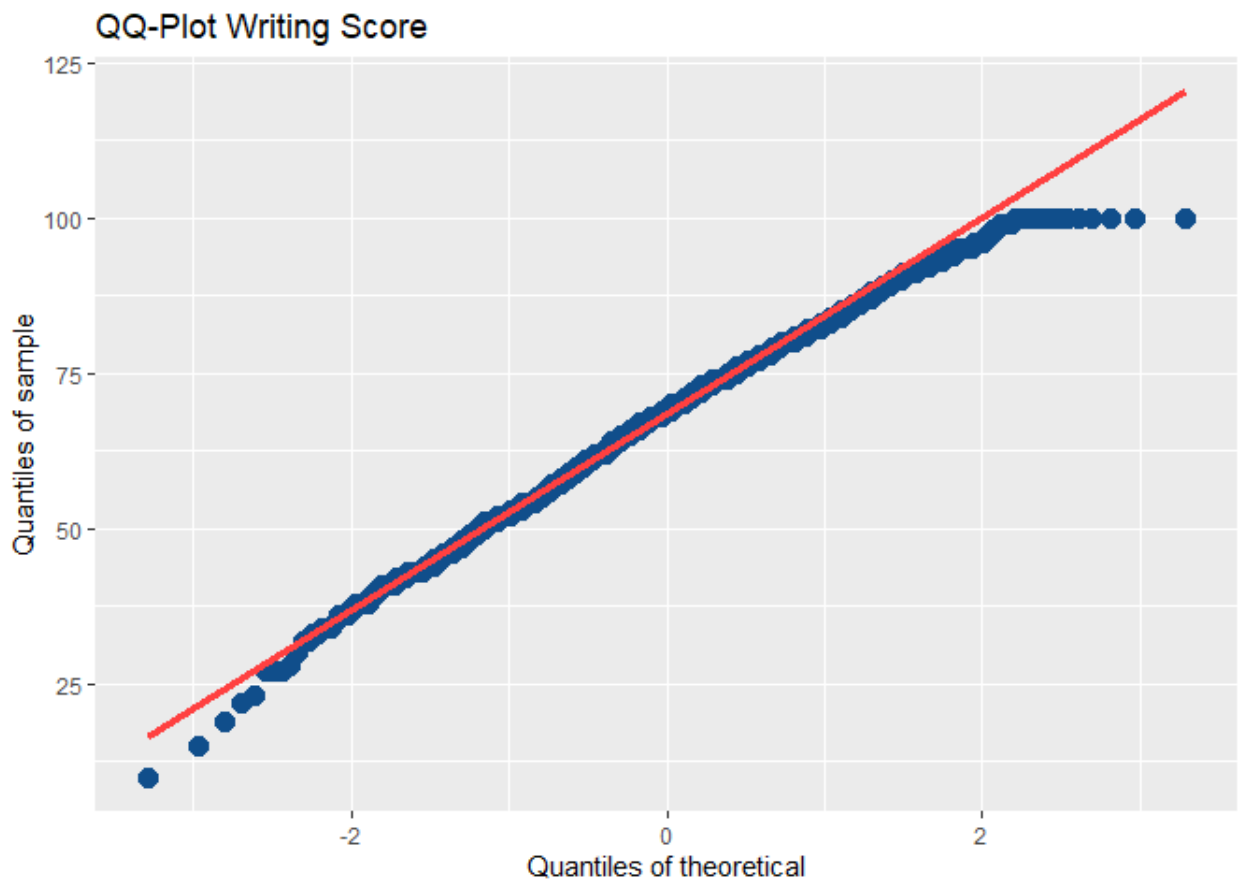
```
ggplot(dados_scores,aes(sample=math.score))+  
  stat_qq(shape=19,size=3.5,col='dodgerblue4')+  
  stat_qq_line(lwd=1.5,col='brown1')+  
  labs(y='Quantiles of sample')+  
  labs(x='Quantiles of theoretical')+  
  ggtitle("QQ-Plot Math.Score")
```



```
ggplot(dados_scores,aes(sample=reading.score))+  
  stat_qq(shape=19,size=3.5,col='dodgerblue4')+  
  stat_qq_line(lwd=1.5,col='brown1')+  
  labs(y='Quantiles of sample')+  
  labs(x='Quantiles of theoretical')+  
  ggtitle("QQ-Plot Reading.Score")
```



```
ggplot(dados_scores,aes(sample=writing.score))+  
  stat_qq(shape=19,size=3.5,col='dodgerblue4')+  
  stat_qq_line(lwd=1.5,col='brown1')+  
  labs(y='Quantiles of sample')+  
  labs(x='Quantiles of theoretical')+  
  ggtitle('QQ-Plot Writing Score')
```



O QQ-Plot é um método gráfico para comparar duas distribuições de probabilidade. Sendo assim é utilizado quando queremos comparar graficamente uma distribuição.

## Realizando Teste de Shapiro-Wilk

```
a <- shapiro.test(dados$math.score)$p.value  
b <- shapiro.test(dados$reading.score)$p.value  
c <- shapiro.test(dados$writing.score)$p.value
```

```
shapiro_score <- c(a,b,c)
```

```
shapiro_score < 0.05
```

```
## [1] TRUE TRUE TRUE
```

Após utilizar o QQ-plot para verificar se os dados apresentam um comportamento de uma distribuição normal, decidi utilizar o **teste de Shapiro-Wilk** para verificar se os dados apresentam ou não uma distribuição normal. Sendo assim decidi criar uma variável onde contém os **P-Valores** do teste de

Shapiro de cada nota, pôs caso o valor do teste seja menor que **5%** (ou 0,05) **\*\*** então teremos um resultado que demonstra que o conjunto de dados não apresenta uma distribuição normal, dessa forma conseguimos visualizar no output do código as palavras **\*\*(TRUE TRUE TRUE) \*\***.

## Realizando Teste Mann Whitney

Uma breve introdução sobre o teste de Mann Whitney é um teste não paramétrico aplicado para duas amostras independentes. Podemos dizer que é a versão do teste t (ou t student) para um conjunto de dados não paramétricos. Sendo assim queremos comparar dois grupos e verificar se existe uma diferença entre esses grupos.

As hipóteses do teste de Mann Whitney são:

H0: Não existe diferença  $p > 0,05$

H1: existe diferença  $P < 0,05$

A hipótese irá verificar se a mediana dos grupos analisados é diferente.

```
wilcox_math.score <- wilcox.test(math.score ~ gender, alternative = "two.s
wilcox_reading.score <- wilcox.test(reading.score ~ gender, alternative =
wilcox_writing.score <- wilcox.test(writing.score ~ gender, alternative =
```

O código acima foi criado 3 variáveis onde são aplicados os testes de Mann Whitney, com o objetivo de verificar a diferença das notas entre os gêneros com base na matéria.

```
wilcox_p.values <- c(wilcox_math.score, wilcox_reading.score, wilcox_writ
wilcox_p.values < 0.05
```

```
## [1] TRUE TRUE TRUE
```

Quando realizado os testes verificamos que os **P-Valores** do teste realizado para cada respectiva variável, verificamos que os **P-Values** de todos os testes

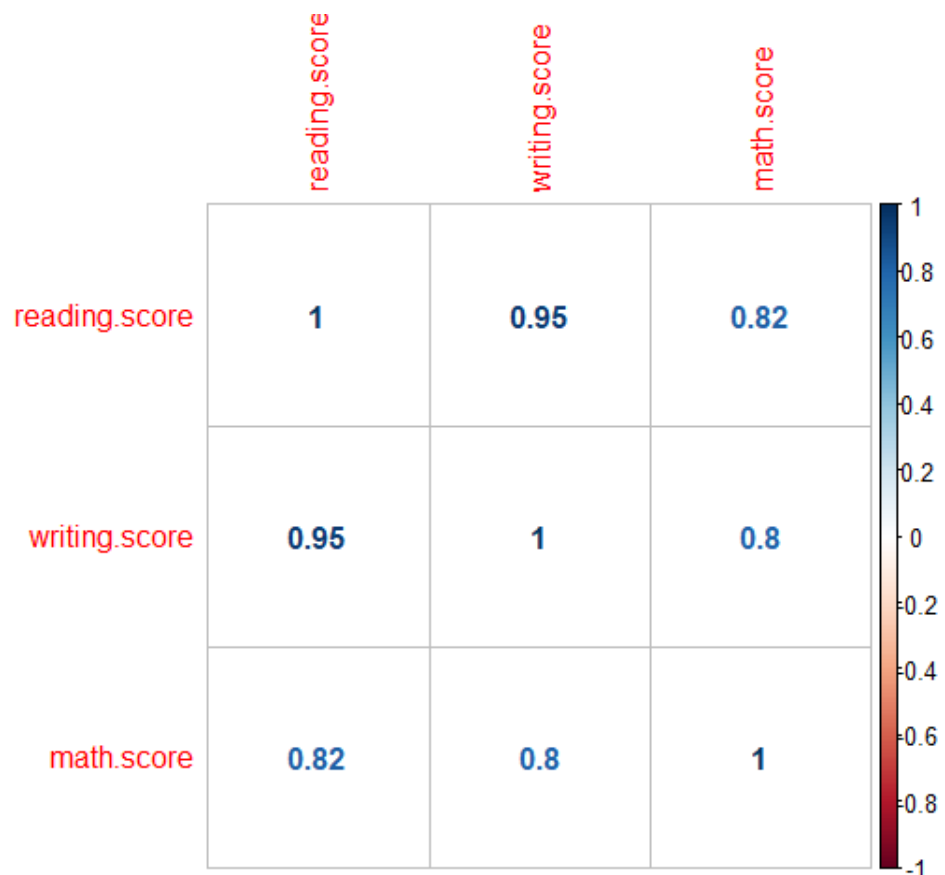
foram abaixo dos \*\*5% (ou 0,05) \*\*, sendo assim os grupos analisados apresentaram uma diferença entre as notas.

## Matriz de Correlação

Uma introdução sobre o tema **correlação** é determinar o grau de relacionamento entre duas variáveis. Vale ressaltar que relacionamento entre duas variáveis não significa casualidade entre uma variável(X) e variável(Y).

```
corre <- cor(dados_scores,method="pearson")
```

```
corrplot(corre,method = 'number')
```



Através da biblioteca **Corrplot** realizamos uma matriz de correlação onde é possível ver a correlação entre as variáveis notas. Dessa forma conseguimos visualizar do lado da matriz correlação uma barra de escala onde é representado o grau de correlação entre as variáveis. Sendo assim quando o valor da matriz de correlação estiver com a cor azul isso que significa que as variáveis comparadas apresentam uma mesma direção de relação. Já quando o valor da matriz de correlação apresentar uma cor vermelha isso significa

que as variáveis analisadas apresentam uma direção diferente das relações. Vale ressaltar novamente que correlação não é causalidade e que quaisquer conjuntos de dados que estivermos analisando sempre irá apresentar uma correlação o queremos visualizar e o grau de relação entre essas variáveis.

## Contato

---

Caso o leitor tenha encontrado algum erro ou queira sugerir alguma mudança ou sugestão entre em contato através do e-mail abaixo.

Email : [charles.b.ribeiro@gmail.com](mailto:charles.b.ribeiro@gmail.com)

" Não tenha medo de cometer erros,tenha medo de não aprender com eles - Peter Jones "