Probabilidade e Estatística

Professor Petrúcio Barros

Aluna: **Lívia De Maria Calado Machado Soares**
Aluno: **Pedro Henrique Mesquita Isidoro**

Questão 01)

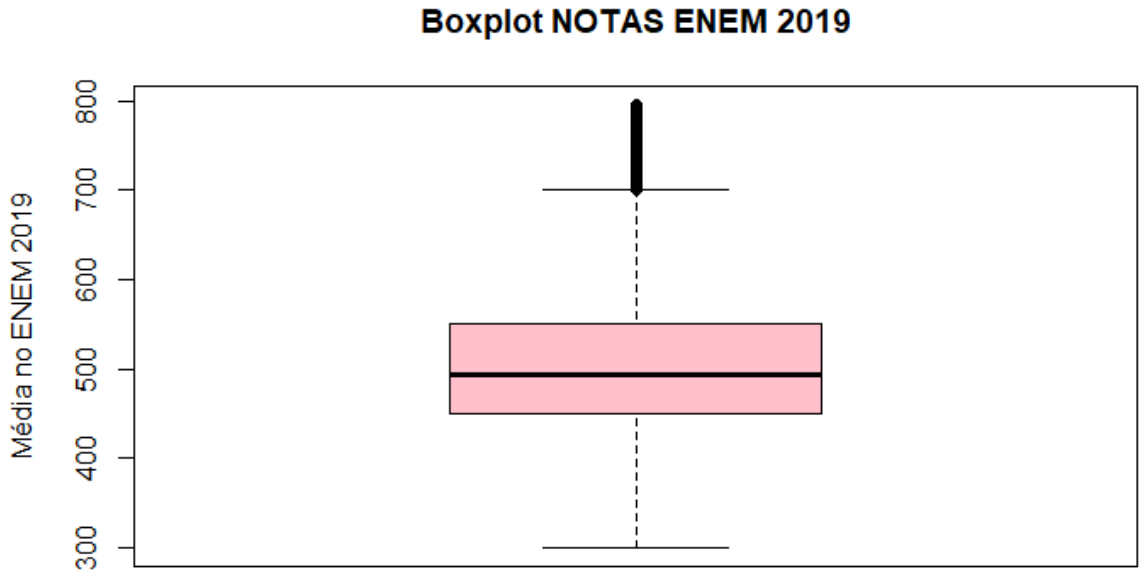
- **Gerar um gráfico boxplot e construir tabela com as frequências absolutas e relativas e suas respectivas acumuladas.**

IMPLEMENTAÇÃO

```
6
7  #Chamar alguns pacotes para serem usados
8  library(psych)
9  library(gridExtra)
10 library(formattable)
11 library(dplyr)
12
13
14 #Criar um boxplot com as NOTA_ENEM
15 boxplot(ENEM$NOTA_ENEN,col = "pink", ylab = "Média no ENEM 2019", main = "Boxplot NOTAS ENEM 2019")
16
```

No início do código, foi chamada determinadas bibliotecas para serem usadas ao decorrer do programa.

GRÁFICO



TABELA

```
# Tabela de frequências:

#Criação da tabela frequência relativa
FreqRel <- table(cut(ENEM$NOTA_ENEN, seq(300, 800, l = 6)))
FreqRel
#Criação da tabela frequência absoluta
FreqAbs <-prop.table(FreqRel)
FreqAbs
FreqAbs <-percent(c(FreqAbs))
FreqAbs

#Criar uma tabela com as frequências
Tabela_frequencias <- data.frame(
  Frequencia_Relativa = c(FreqRel),
  Frequencia_Absoluta =  c(FreqAbs))
#Exibir a tabela
Tabela_frequencias
#Plotar a tabela
formattable(Tabela_frequencias)
```

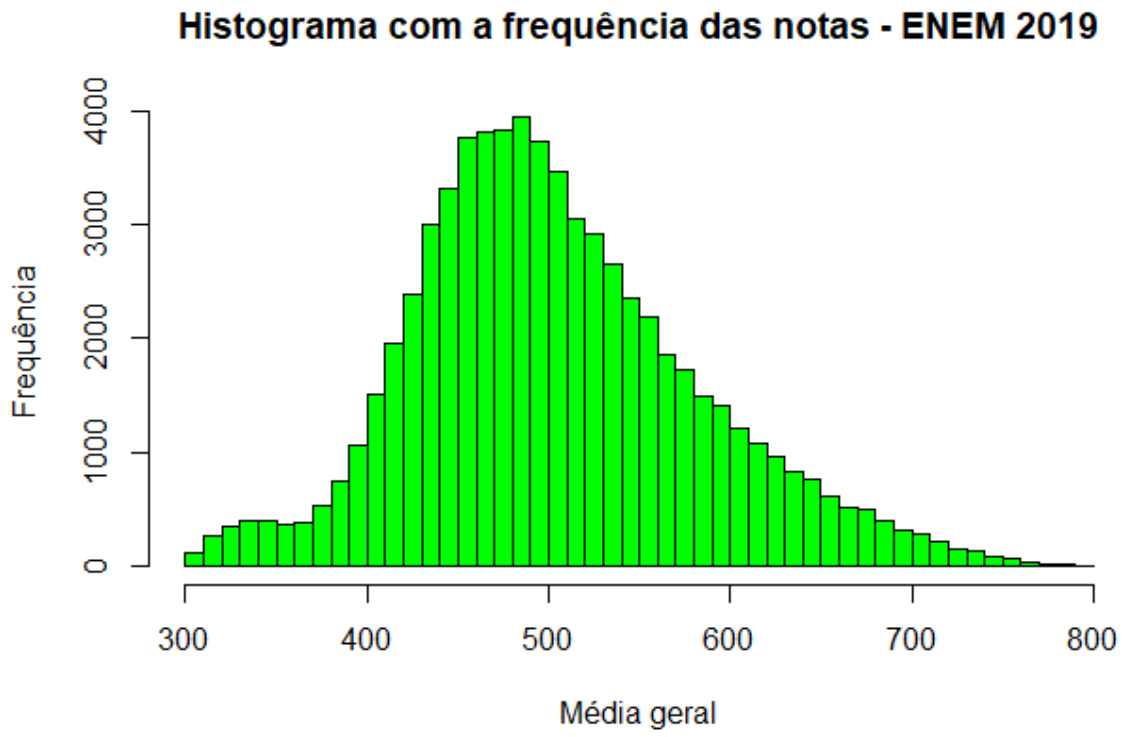
	Frequencia_Relativa	Frequencia_Absoluta
(300,400]	4618	6.87%
(400,500]	31270	46.54%
(500,600]	23136	34.43%
(600,700]	7189	10.70%
(700,800]	979	1.46%

- **Gerar um Histograma com a frequência das notas.**

IMPLEMENTAÇÃO

```
36
37  #Criar um histograma com as NOTA_ENEM
38  #800 - 300 = 500
39  #500 / 10 = 50
40
41  hist(ENEM$NOTA_ENEN, breaks = 50,
42       col = "green",xlab = "Média geral",
43       ylab = "Frequência",
44       main = "Histograma com a frequência das notas - ENEM 2019")
45
```

GRÁFICO

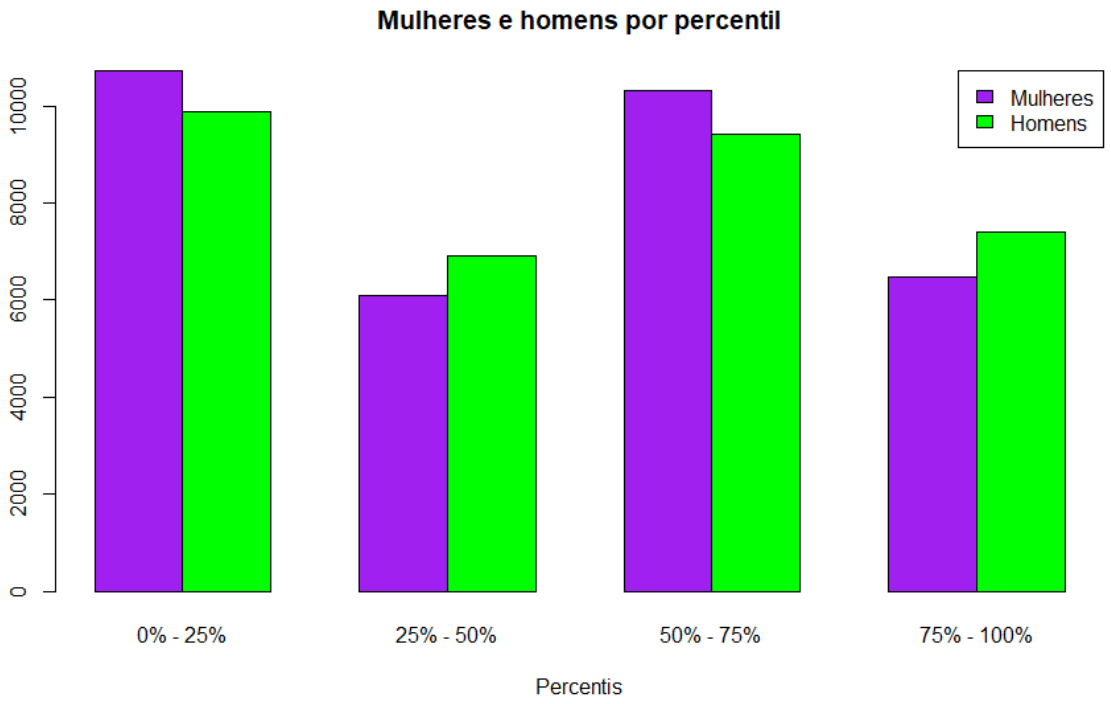


- **Gerar um gráfico de barras com as NOTA_ENEN agrupado pelos quartis e sexo e interpretar os valores.**

IMPLEMENTAÇÃO

```
48
49  #Gera os quartis da das notas
50  Quartil_Nota <- c(quantile(ENEM$NOTA_ENEN))
51  Quartil_Nota
52
53  #Transformar Feminino em 1 e masculino em 0 para uma melhor manipulação dos dados
54  GENERO <- ENEM$NOTA_ENEN
55  for(i in 1:67192){
56    if(ENEM$TP_SEXO[i] == 'Feminino'){
57      GENERO[i] = 1
58    }else{
59      GENERO[i] = 0
60    }
61  }
62  #Criar variáveis para armazenar a quantidade feminina(m) e masculina (n) em cada quartil
63  m1 = 0
64  m2 = 0
65  m3 = 0
66  m4 = 0
67  n1 = 0
68  n2 = 0
69  n3 = 0
70  n4 = 0
71  #Calcular a quantidade de cada um
72  for(i in 1:67192){
73    if(i < 16798){
74      if(GENERO[i] == 1){m1 = m1 + 1}
75      else{n1 = n1 + 1}
76      }
77    else{if(i < 33596){
78        if(GENERO[i] == 1){m2 = m2 +1}
79        else{n2 = n2 + 1}
80      }
81      else{if(i < 50394){
82          if(GENERO[i] == 1){m3 = m3 + 1}
83          else{n3 = n3 + 1}
84        }
85        else{if(GENERO[i] == 1){m4 = m4 + 1}
86          else{n4 = n4 + 1}
87        }
88      }
89    }
90  }
91  Percentil_MF<-c(m1,n1,m2,n2,m3,n3,m4,n4)
92  Per_Gen <-matrix(data = Percentil_MF, ncol = 4, byrow = TRUE,
93              dimnames = (list(c("Feminino","Masculino"),
94                               c("0% - 25%","25% - 50%","50% - 75%","75% - 100%"))))
95  #Gerar o gráfico com os respectivos quartis
96  barplot(Per_Gen,
97          main = "Mulheres e homens por percentil",
98          xlab = "Percentis",
99          col = c("purple","green"),
100         beside = TRUE
101 )
102 legend("topright",
103        fill = c("purple","green"),
104        c("Mulheres","Homens")
105
106 )
```

GRÁFICO



COMENTÁRIOS

Analisando o gráfico gerado com os percentis por homens e mulheres, é possível observar que houveram mais pessoas no primeiro e terceiro quartil e com isso mais mulheres. No segundo e quarto quartil, com menos pessoas, houve uma presença maior masculina. Essa falta de padrão no percentil mostra que não exisate também um padrão relacionando o gênero.

- **Escolher duas variáveis (colunas) e gerar os gráficos mais adequados para tais colunas.**

IMPLEMENTAÇÃO 01

```
#Escolher duas variáveis (colunas) e gerar os gráficos mais adequados para tais colunas.

#gráfico de pontos contagem  x nota redação x idade
#carrega os pacotes
library(ggplot2)
library(tidyverse)

#código
ENEM %>%
  group_by(NU_NOTA_REDACAO, NU_IDADE) %>%
  summarise(
    contagem = n()
  ) %>%
  ggplot(aes(x = NU_NOTA_REDACAO, y = contagem, fill = NU_IDADE, label = contagem)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Nota de redação por idade",
       subtitle = "Conjunto de dados ENEM 2019",
       x = "Nota redação", y = "Contagem")
```

IMPLEMENTAÇÃO 02

```
#Escolher duas variáveis (colunas) e gerar os gráficos mais adequados para tais colunas.

#gráfico de pontos contagem  x nota redação x Município
#carrega os pacotes
library(ggplot2)
library(tidyverse)

#código
ENEM %>%
  group_by(NU_NOTA_REDACAO, NO_MUNICIPIO_PROVA) %>%
  summarise(
    contagem = n()
  ) %>%
  ggplot(aes(x = NU_NOTA_REDACAO, y = contagem, fill = NO_MUNICIPIO_PROVA, label = contagem)) +
  geom_bar(stat = "identity") +
  labs(title = "Nota de redação por município",
       subtitle = "Conjunto de dados ENEM 2019",
       x = "Nota redação", y = "Contagem")

```

GRÁFICO 1(REDAÇÃO X IDADE)

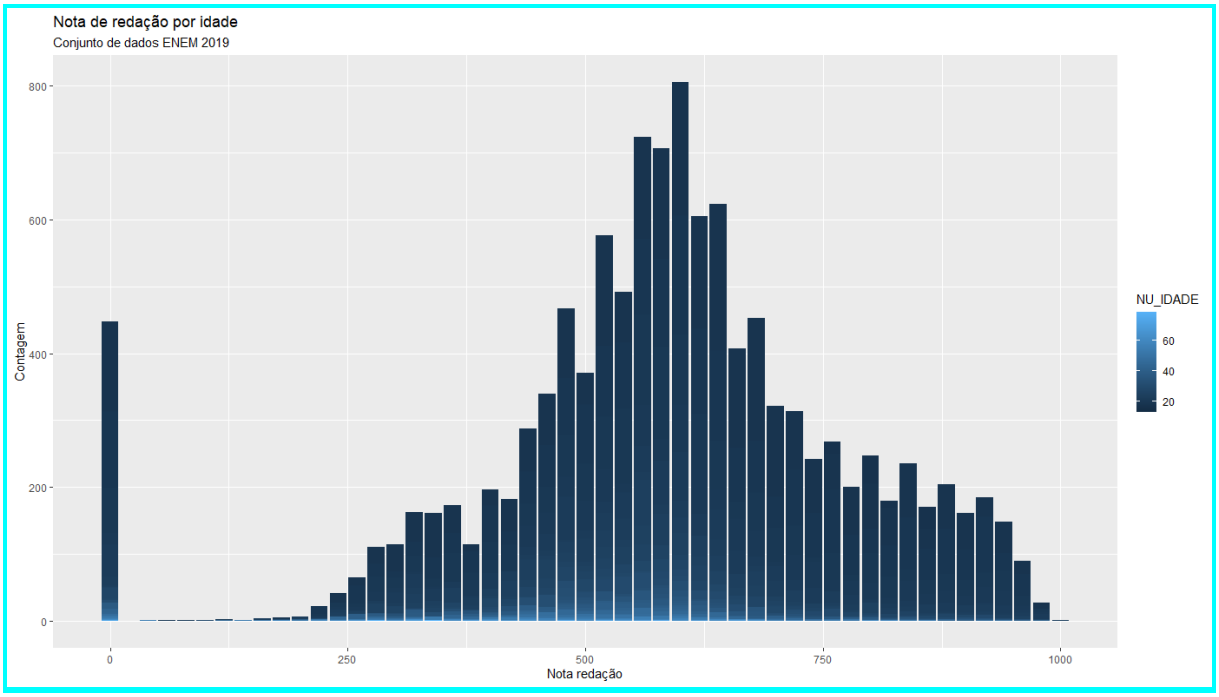
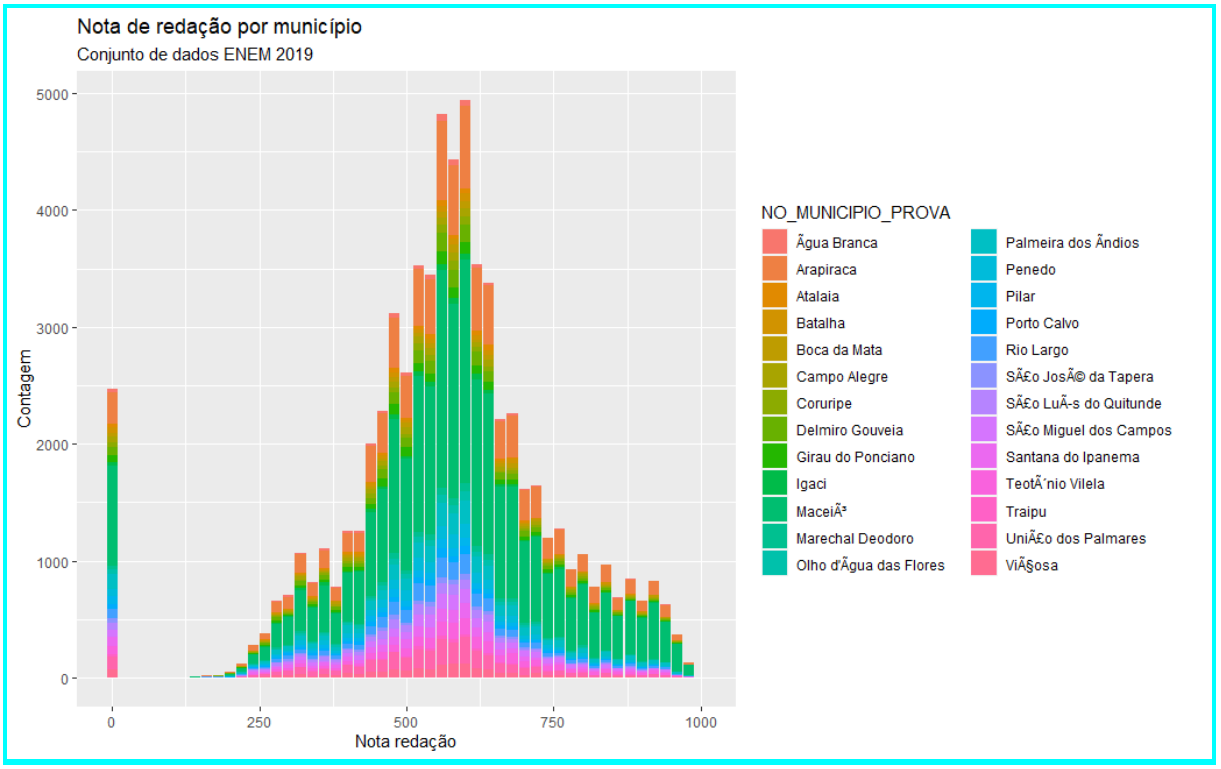


GRÁFICO 2(RDEÇÃO X MUNICÍPIO)

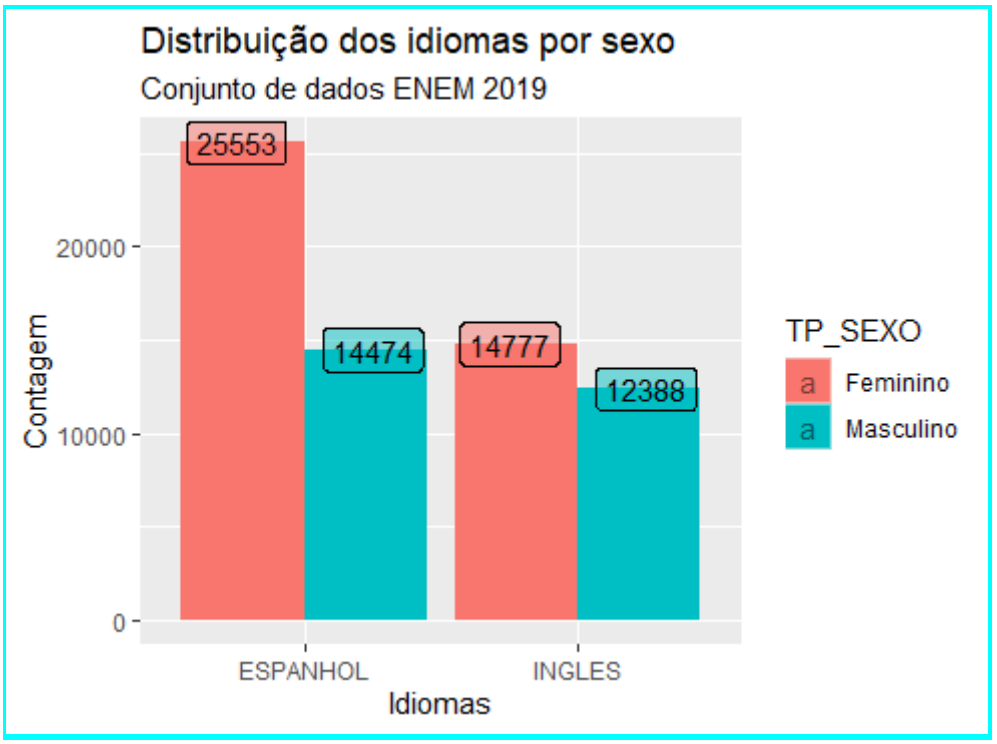


- **Escolher duas variáveis (colunas) qualitativas, gerar gráficos adequados e interpretar os resultados.**

IMPLEMENTAÇÃO

```
49
50  #gráfico de pontos contagem  x idioma x sexo
51  #carrega os pacotes
52  library(ggplot2)
53  library(tidyverse)
54  #acha o arquivo no diretório
55  dados = read.csv2(file.choose())
56  dados
57
58  #código
59  dados %>%
60    group_by(TP_LINGUA, TP_SEXO) %>%
61    summarise(
62      contagem = n()
63    ) %>%
64    ggplot(aes(x = TP_LINGUA, y = contagem, fill = TP_SEXO, label = contagem)) +
65    geom_bar(stat = "identity", position = "dodge") +
66    geom_label(position = position_dodge(width = 1), alpha = 0.5) +
67    labs(title = "Distribuição dos idiomas por sexo",
68         subtitle = "Conjunto de dados ENEM 2019",
69         x = "Idiomas", y = "Contagem")
70
```

GRÁFICO



COMENTÁRIOS

Pode-se analisar a partir do gráfico que há uma grande discrepância entre o número de participantes homens e mulheres, com as mulheres sendo a maioria. Além disso, é perceptível que a grande parte dos indivíduos optam por escolher o espanhol para a prova de idiomas do vestibular. Outro análise importante presente

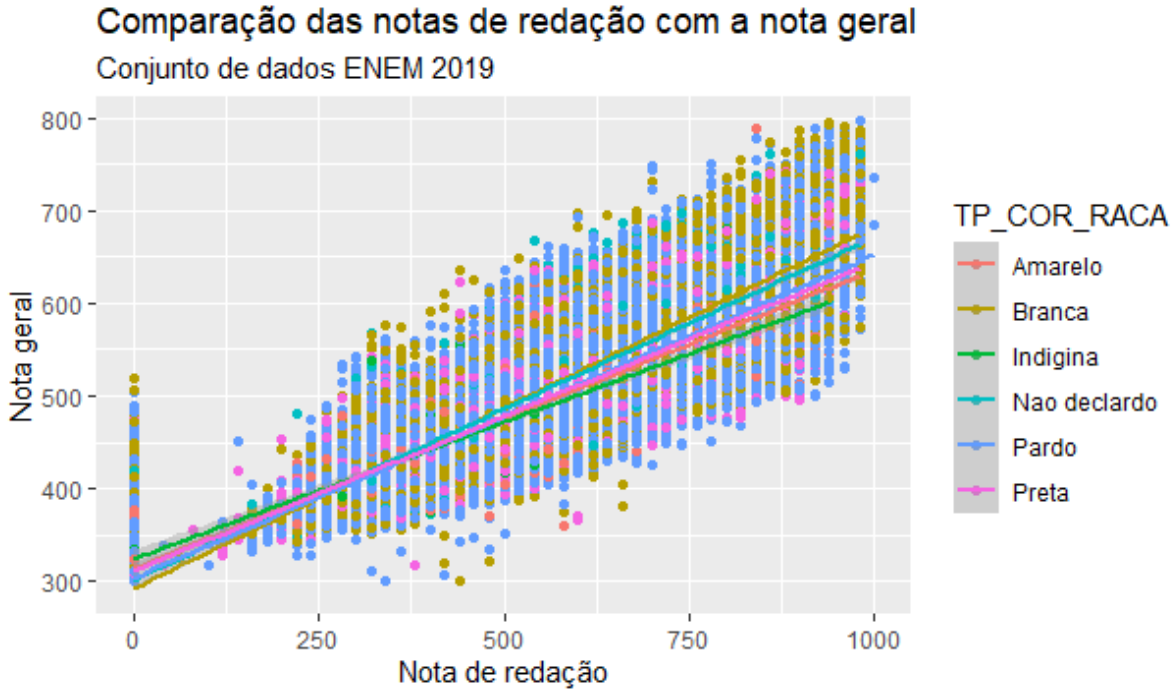
no gráfico, é a esmagadora diferença entre homens e mulheres na escolha da língua espanhola, sendo 25553 mulheres e 14474 homens.

- **Escolher duas variáveis, sendo uma qualitativa e outra quantitativa, combinar e interpretar os resultados.**

IMPLEMENTAÇÃO

```
32
33 ############################################################
34 #gráfico de pontos nota de redação  x nota geral
35 #carrega os pacotes
36 library(ggplot2)
37 library(tidyverse)
38
39 #acha o arquivo no diretório
40 dados = read.csv2(file.choose())
41 dados
42 #CÓDIGO
43 dados %>%
44   ggplot(aes(x = NU_NOTA_REDACAO, y = NOTA_ENEN, color = TP_COR_RACA)) + geom_point()+
45   geom_smooth(method = "lm") + labs(title = "Comparação das notas de redação com a nota geral",
46                                     subtitle = "Conjunto de dados ENEM 2019",
47                                         x = "Nota de redação", y = "Nota geral")
48
49
```

GRÁFICO



COMENTÁRIOS

A partir da análise de dados do gráfico de pontos, pode-se afirmar que quanto maior for a nota de redação, maior será a tendência de obter uma média geral alta. Além disso, é perceptível que existe uma tendência de notas variando conforme a cor/raça. Os traços lineares com cores distintas caracterizam a tendência de determinadas raças/cores a seguirem um padrão de equivalência entre as notas de redação e a média geral. A tendência de ter notas mais altas é de pessoas brancas, logo após temos as

pessoas que não declararam, depois pessoas pretas, seguido de amarelos e ,por fim, os indígenas. Porém, não podemos fazer uma simples associação a partir de uma lógica mecanicista e chegar a uma conclusão determinista, visto que o gráfico não leva em consideração a questão socioeconômica e a formação histórica do país. Para uma melhor interpretação dos dados apresentados, seria importante analisar como a questão racial se relaciona com o contexto social e econômico.

- **Escolher duas variáveis quantitativas, combinar e interpretar os resultados.**

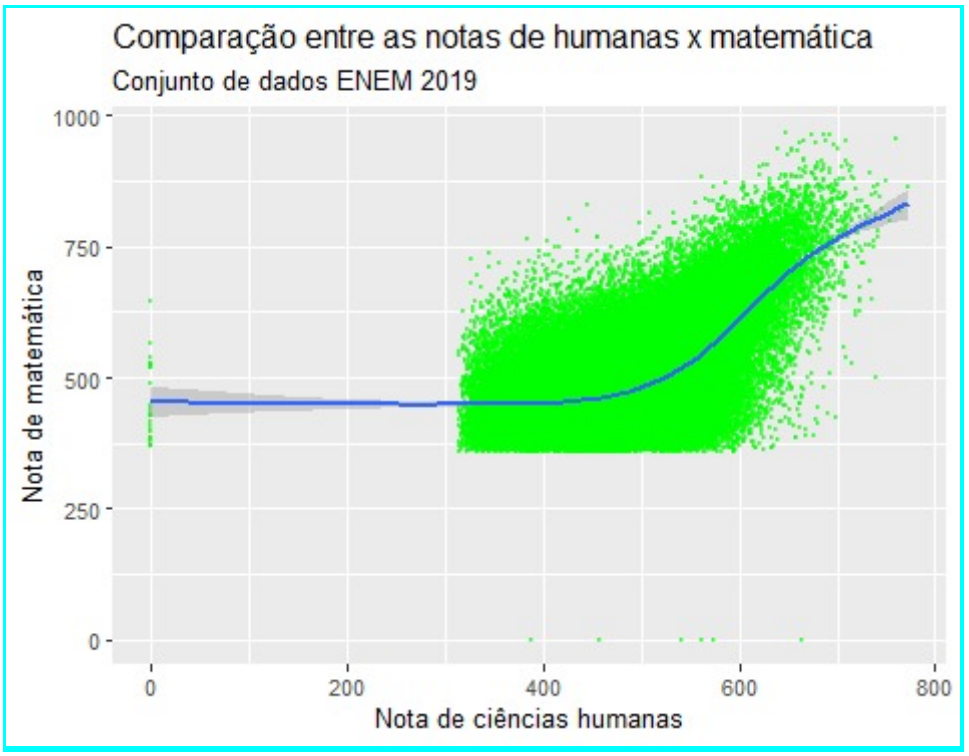
IMPLEMENTAÇÃO

```
#gráfico de pontos nota de matemática  x ciências humanas
#carrega os pacotes
library(ggplot2)
library(tidyverse)

#acha o arquivo no diretório
dados = read.csv2(file.choose())
dados
#CODIGO
dados %>%
  ggplot(aes(x = NU_NOTA_CH, y = NU_NOTA_MT, color = NU_NOTA_CH)) + geom_point(color = "green", size = 0.5, alpha = 0.5)+
  geom_smooth() +
  labs(title = "Comparação entre as notas de humanas x matemática",
                                  subtitle = "Conjunto de dados ENEM 2019",
                                  x = "Nota de ciências humanas", y = "Nota de matemática")

```

GRÁFICO



COMENTÁRIOS

Dado o gráfico de pontos que realiza a comparação entre uma relação das notas de ciências humanas com as notas de matemática, é interessante analisar que há uma tendência das notas das duas provas serem diretamente proporcionais, existindo apenas uma pequena diferença em certos momentos na linha de tendência. Ademais, percebe-se que há uma maior facilidade de chegar perto da nota 1000 em matemática, em contrapartida em ciências humanas não tiveram notas que passaram de 800.

- **Questão 02: Utilizando a base de dados, escolha um município, agrupe algumas variáveis a sua escolha, gere os gráficos que achar adequados e faça uma síntese dos resultados obtidos.**

O município escolhido foi Igaci

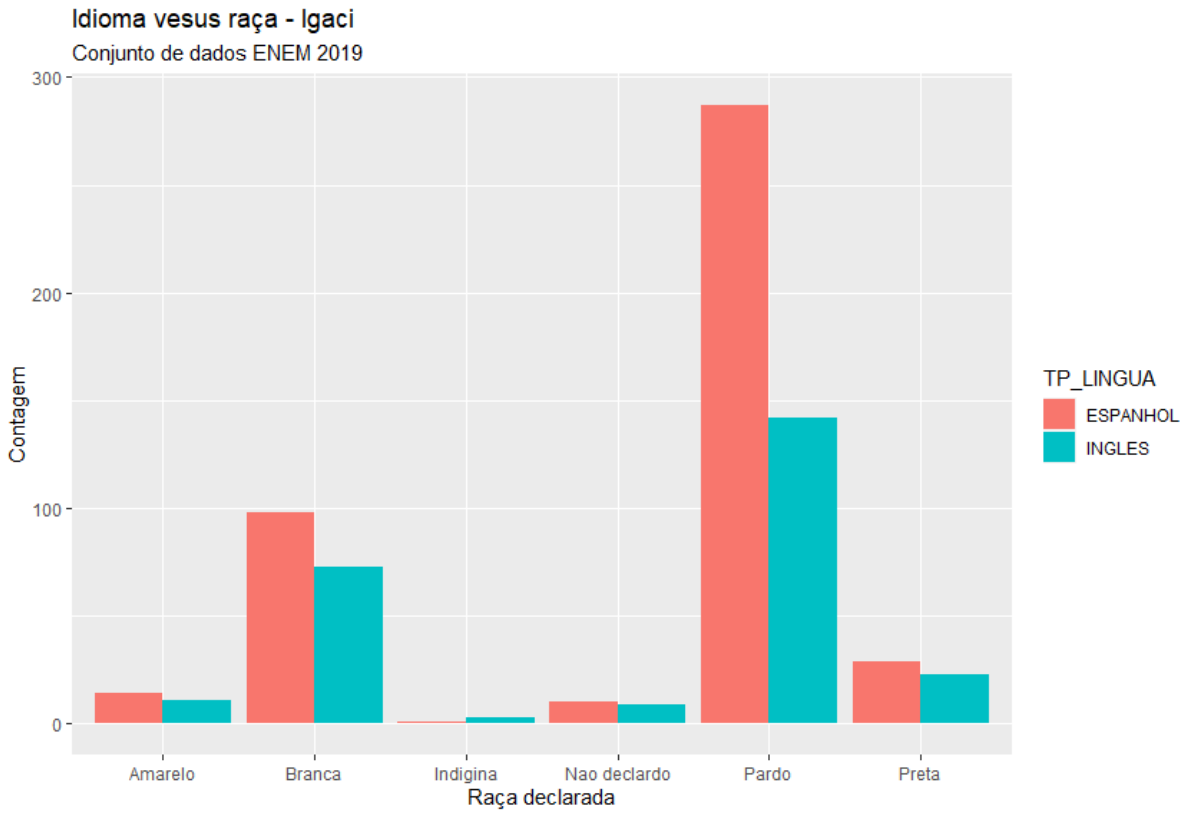
IMPLEMENTAÇÃO

```
#Primeiro foi separado igaci do resto do municípios de prova
igaci <- filter(ENEM, ENEM$NO_MUNICIPIO_PROVA == "Igaci")
igaci %>%
  group_by(TP_COR_RACA, TP_LINGUA) %>%
  summarise(
    contagem = n()
  ) %>%
  ggplot(aes(x = TP_COR_RACA, y = contagem, fill = TP_LINGUA, label = contagem)) +

  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Idioma vesus raça - Igaci",
       subtitle = "Conjunto de dados ENEM 2019",
       x = "Raça declarada", y = "Contagem")
```

GRÁFICO

- **Idioma x raça**



COMENTÁRIOS

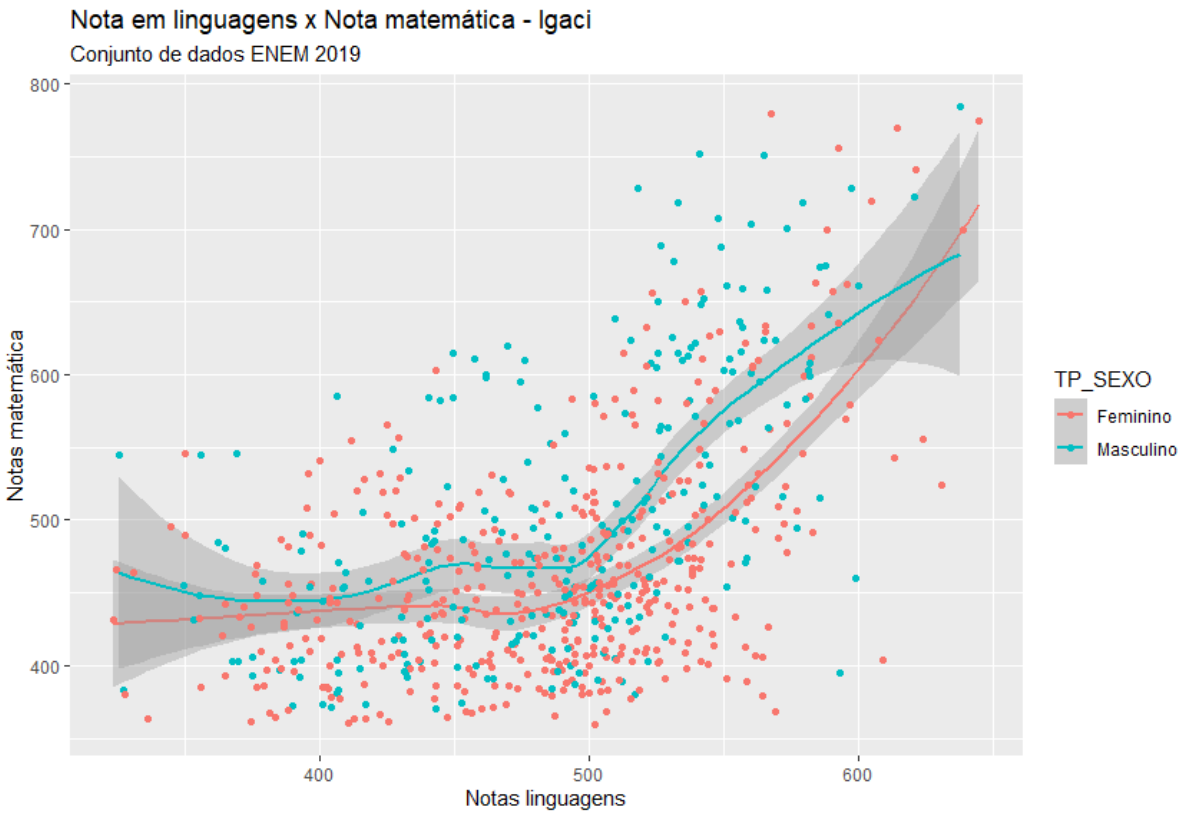
No gráfico acima, é possível notar a preferência geral da população pelo espanhol do que o inglês. Isso talvêz se dê pela similariedade do idioma com o nosso idoma nativo. A única divergência do dado se dá com os declarados indígenas, os quais escolheram inglês ao espanhol.

- **Nota em linguagens X Nota matemática**

IMPLEMENTAÇÃO

```
igaci %>%
  group_by(NU_NOTA_LC, TP_LINGUA) %>%
  ggplot(aes(x = NU_NOTA_LC, y = NU_NOTA_MT, color = TP_SEXO)) +
  geom_point() +
  labs(title = "Nota em linguagens x Nota matemática - Igaci",
       subtitle = "Conjunto de dados ENEM 2019",
       x = "Notas linguagens", y = "Notas matemática")
```

GRÁFICO



COMENTÁRIOS

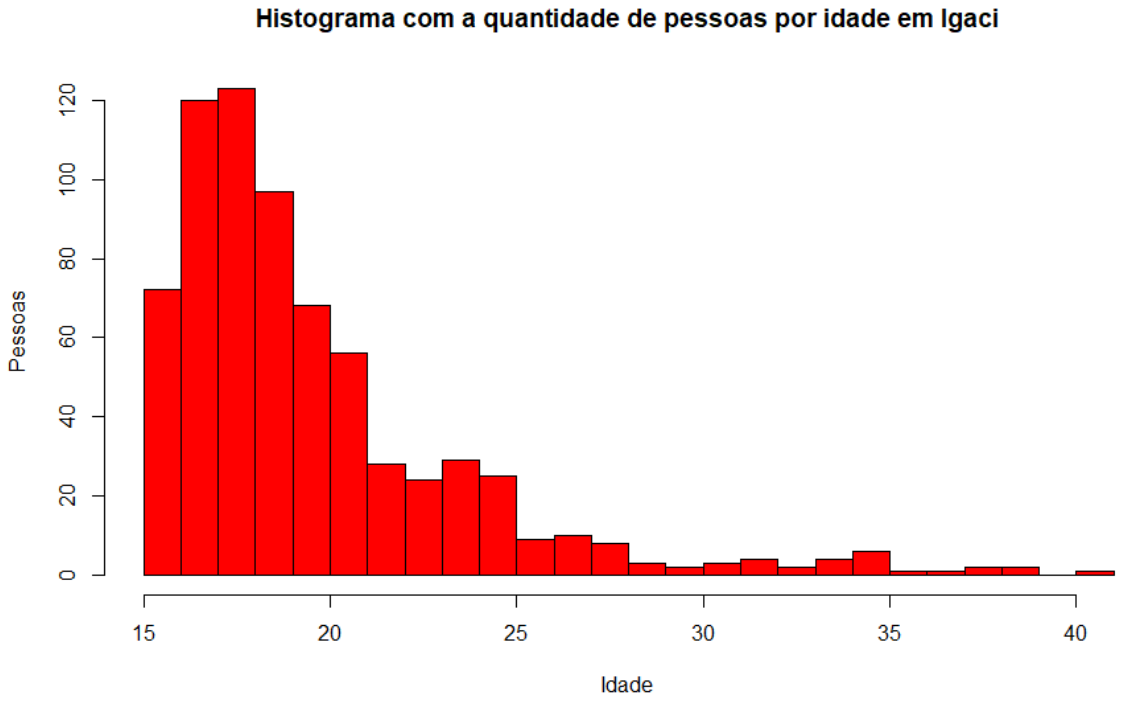
Dado o gráfico acima, é possível notar que há uma comparação entre as notas de linguagens e as notas de matemática visualizando cada indivíduo pelo sexo. Percebe-se que existe uma tendência de indivíduos do sexo masculino a obterem notas maiores que indivíduos do sexo feminino, porém em um determinado momento do gráfico, a tendência inverte e as mulheres são responsáveis pela grande parte das notas mais altas em ambas as disciplinas. Vale salientar que a maioria das notas de linguagens estão no intervalo de 400 - 600 e as notas de matemática encontram-se no intervalo 300 - 500.

- **Quantidade de pessoas x Idade**

IMPLEMENTAÇÃO

```
hist(igaci$NU_IDADE, breaks = 20,
     col = "red",xlab = "Idade",
     ylab = "Pessoas",
     main = "Histograma com a quantidade de pessoas por idade em Igaci")
```

GRÁFICO



COMENTÁRIOS

No gráfico acima, é possível observar a grande presença de pessoas de 15-20 anos realizando a prova. Isso provavelmente se dá pois é nessa idade que as pessoas realizam o ENEM para treinar ou para ingressar na faculdade. Dos 20 aos 25 anos, uma média de 25 pessoas por idade se mantém. A partir dos 25 essa média cai, ao ponto de que a pessoa mais velha a fazer o ENEM em Igaci ter 40 anos.