Projeto Integrador

Felipe A. & Vinícius N.

21/11/2021

Introdução

A motivação para o seguinte trabalho constitui-se de poder analisar e ter o primeiro contato com a área de análise de dados, aperfeiçoar habilidades com estatística descritiva e programação voltada para Data Science. Para tanto, nós pegamos uma base do site Kaggle, com informações sobre a pontuação em testes com mil alunos. Ao longo do trabalho, nós tentaremos responder as seguintes perguntas divididas em dois grupos:

Grupo A:

- Que tipos de características mais impactam a pontuação geral?
- Podemos, de alguma forma, prever essa pontuação com base nas informações prévias?

Grupo B:

- A relação da nota de um aluno em Escrita, tem a ver com a nota que ele obteve em Leitura e seu gênero?
- A relação da nota de um aluno em Matemática, tem a ver com a nota que ele obteve com a Escrita e seu gênero?

Dessa forma ao longo do trabalho, vamos poder definir porque a base de dados não nos ajuda a explicar as perguntas do Grupo B

Além disso, o presente trabalho tem como foco o aprendizado de uma forma nova de pensar, embasado na criação de hipóteses, o teste das mesmas, e na procura de conhecer a fundo o problema disposto.

Material e métodos

Coleta de dados:

A base obtida apresenta informações de nota de mil alunos em diversas matérias, gênero e até mesmo qual foi o tipo de alimentação que os estudantes tiveram durante o período que foi registrado. Esta base foi obtida através do site Kaggle, podemos imaginar que o intuito da construção dessa base de dados, foi analisar diversas variáveis que possam contribuir para o desempenho de um indivíduo que está cursando a escola.

```
library(knitr)
getwd()

## [1] "C:/Users/vinic/Desktop"

df<-read.csv('StudentsPerformance.csv')

kable(head(df,6)) # Criando uma tabela com as primeiras informações
```

gender	race.ethnicity	parental.level.of.education	lunch	test.preparation.course	math.score	reading.score	writing.score
female	group B	bachelor's degree	standard	none	72	72	74
female	group C	some college	standard	completed	69	90	88
female	group B	master's degree	standard	none	90	95	93
male	group A	associate's degree	free/reduced	none	47	57	44
male	group C	some college	standard	none	76	78	75
female	group B	associate's degree	standard	none	71	83	78

Variáveis contidas na base de dados.

```
## [1] "gender"                      "race.ethnicity"
## [3] "parental.level.of.education" "lunch"
## [5] "test.preparation.course"     "math.score"
## [7] "reading.score"               "writing.score"
```

Nossa base de dados está dividida em 8 variáveis, que são:

1. Gênero
2. Etnia
3. Nível de educação dos pais

4. Almoço
5. Preparação para o teste
6. Pontuação em matemática
7. Pontuação em leitura
8. Pontuação em escrita

As variáveis Gênero, Etnia, Nível de educação dos pais, Almoço e preparação para o teste, são variáveis qualitativas, já as variáveis que contemplam a pontuação dos estudantes nas matérias, são variáveis quantitativas.

```
str(df)

## 'data.frame':    1000 obs. of  8 variables:
##  $ gender                     : chr  "female" "female" "female" "male"
...
##  $ race.ethnicity             : chr  "group B" "group C" "group B"
"group A" ...
##  $ parental.level.of.education: chr  "bachelor's degree" "some
college" "master's degree" "associate's degree" ...
##  $ lunch                      : chr  "standard" "standard" "standard"
"free/reduced" ...
##  $ test.preparation.course    : chr  "none" "completed" "none" "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 ...
```

Inicialmente não realizamos nenhum tipo de alteação na base de dados, a mesma já não apresentava divergências necessárias que precisavamos corrigir em seu corpo. A decisão mantida, foi de permanecer com a base de dados do jeito que obtivemos ela para iniciar com as análises.

Porém, agregamos uma nova variável, a variável de média das notas obitdas pelos alunos, para podermos modelar conforme esta variável, as desvantagens podem ser a perda de visualização do “Big picture” porém, nos vai permitir visualizar a modelagem completa das notas, visualizando suas interferências.

```
length(df)

## [1] 8

dim(df)

## [1] 1000    8
```

Nós contamos com 1000 observaçoes para realizarmos as análises propostas e buscar as respostas para as questões levantadas.

Modelagem estatística:

Para a conclusão do trabalho, usaremos diversas técnicas/análises aprendidas ao longo desses dois anos de curso, entre elas:

- Técnicas exploratórias/Análise descritiva

Nosso objetivo durante essa fase é desenvolver uma análise descritiva e exploratória, em busca de visualizarmos melhor as variáveis apresentadas realizando uma análise inicial.

Durante esse ponto buscamos:

Avaliar as variáveis e possíveis mudanças na base de dados. Possíveis mudanças em variáveis quantitativas ou acrescentar novas variáveis com base nas já desenvolvidas.

- Técnicas/Análises prescritivas

Nosso objetivo durante essa fase do processo é avaliar e analisar os resultados obtidos com as inferências realizadas.

O estudo nesse ponto é necessário para tomarmos medidas e compreender como estão funcionando os dados desenvolvidos e recomendar novas medidas.

- Se possível Técnicas/Análises diagnósticas

Nosso objetivo durante essa fase das análises e verificar a exatidão de 4 fatores: Ajuste do modelo Variabilidade dos resíduos A normalidade dos dados E possíveis observações que tenham poder de alavancagem

Se necessário for, avaliar o impacto da exclusão de alguns dados que possam afetar o ajuste dos dados, pois se violar qualquer um dos pontos acima, podemos ter um modelo que não se ajusta corretamente e precisamos fazer modificações ideiais para corrigir os problemas que sejam apontados acima.

Resultados e discussões

Análise exploratória:

```
library(tidyverse)

df%>%
  str() # Dando uma olhada mais a fundo na estrutura dos nossos dados

## 'data.frame':    1000 obs. of  8 variables:
##  $ gender                     : chr  "female" "female" "female" "male"
...
##  $ race.ethnicity             : chr  "group B" "group C" "group B"
"group A" ...
##  $ parental.level.of.education: chr  "bachelor's degree" "some
college" "master's degree" "associate's degree" ...
##  $ lunch                      : chr  "standard" "standard" "standard"
"free/reduced" ...
##  $ test.preparation.course    : chr  "none" "completed" "none" "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 ...

df%>%
  summary()%>%kable() # Um sumário dos nossos dados
```

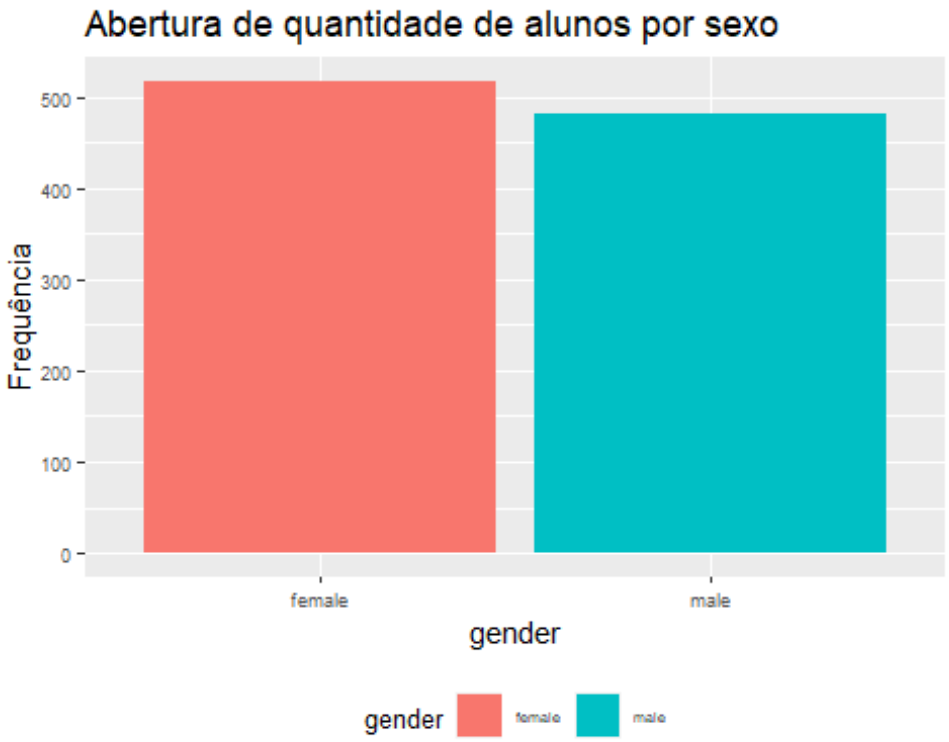
gender	race.ethnicity	parental.level.of.education	lunch	test.preparation.course	math.score	reading.score	writing.score
Length:1000	Length:1000	Length:1000	Length:1000	Length:1000	Min. : 0.00	Min. : 17.00	Min. : 10.00
Class :character	Class :character	Class :character	Class :character	Class :character	1st Qu.: 57.00	1st Qu.: 59.00	1st Qu.: 57.75
Mode :character	Mode :character	Mode :character	Mode :character	Mode :character	Median : 66.00	Median : 70.00	Median : 69.00
NA	NA	NA	NA	NA	Mean : 66.09	Mean : 69.17	Mean : 68.05
NA	NA	NA	NA	NA	3rd Qu.: 77.00	3rd Qu.: 79.00	3rd Qu.: 79.00
NA	NA	NA	NA	NA	Max. :100.000	Max. :100.00	Max. :100.00

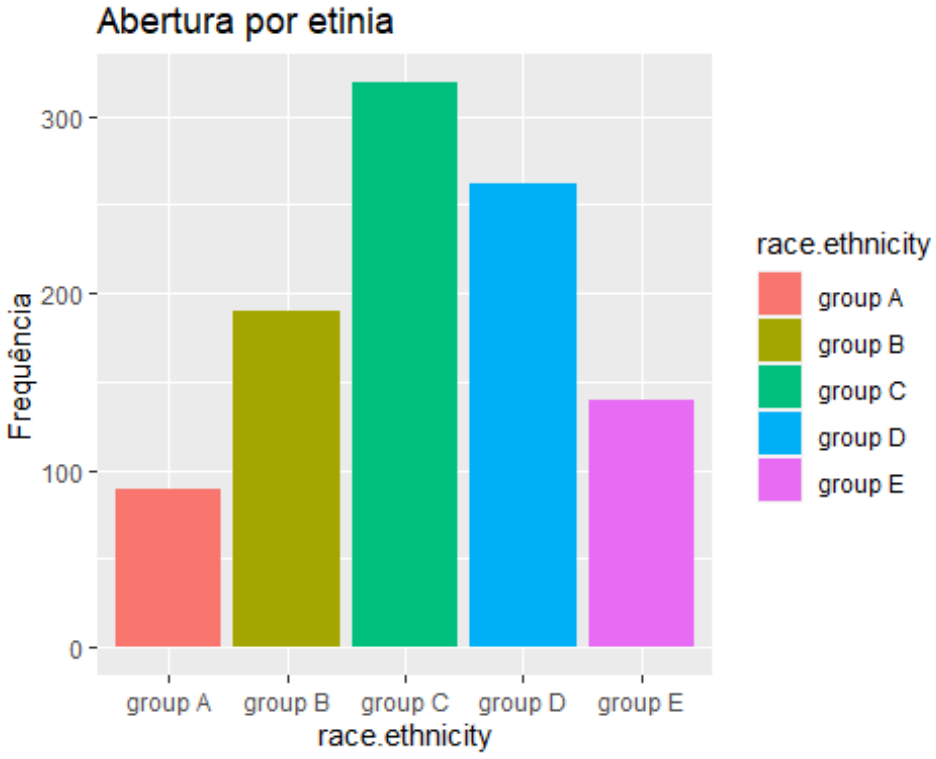
Gráficos

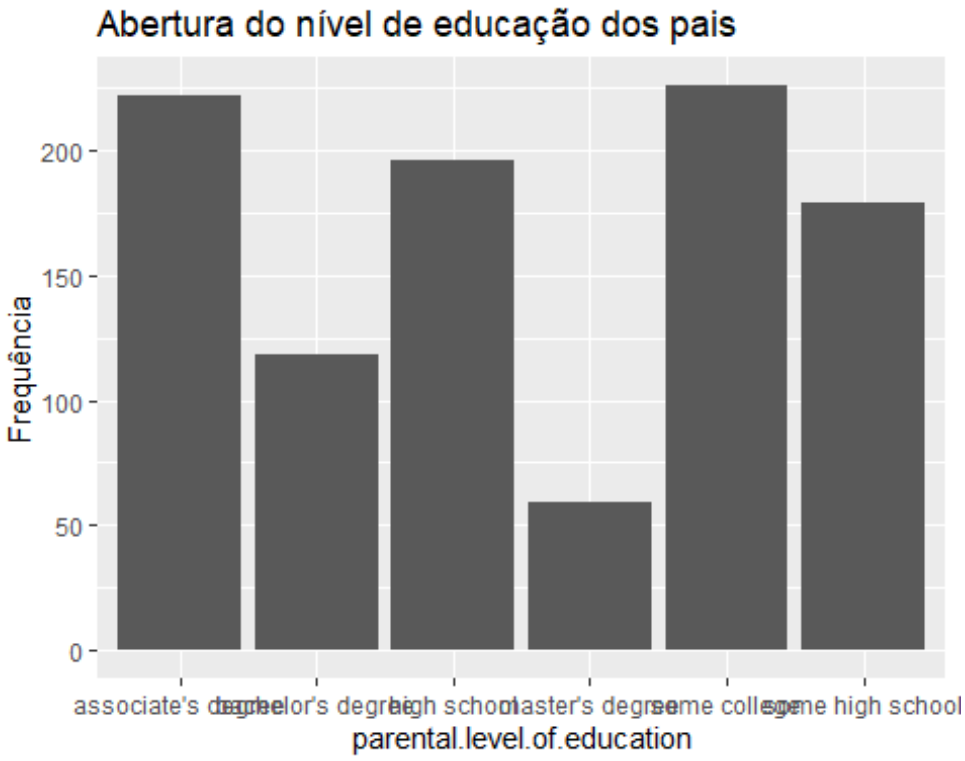
Uma parte importante de qualquer análise, é a criação de gráficos, para melhor análise.

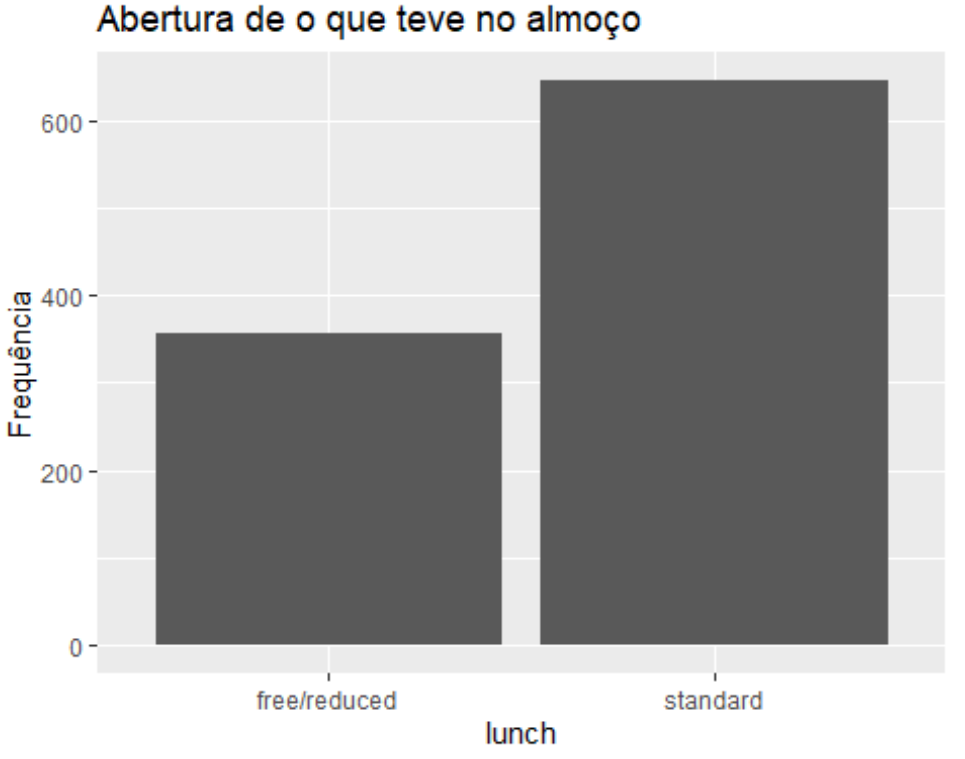
O motivo de plotarmos os gráficos e entender como exatamente a variável se comporta em relação à outras, com base nessa visualização podemos tomar decisões importes para como vamos construir o modelo.

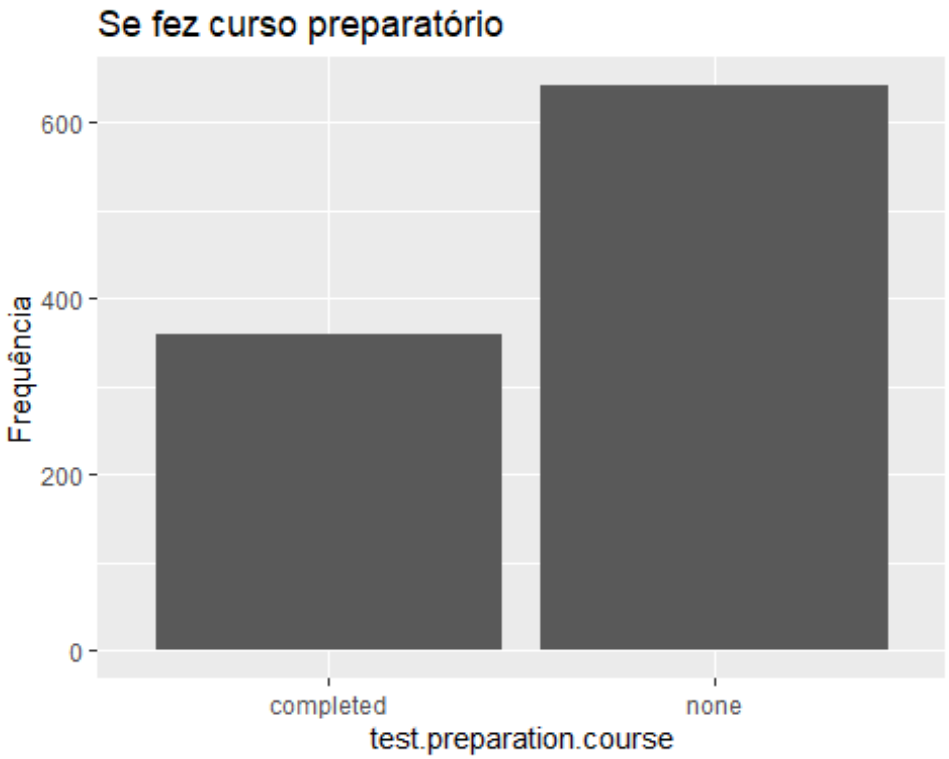
Com essas plotagens podemos evitar erros que futuramente e que necessitem de Análises de resíduos para configurar corretamente.

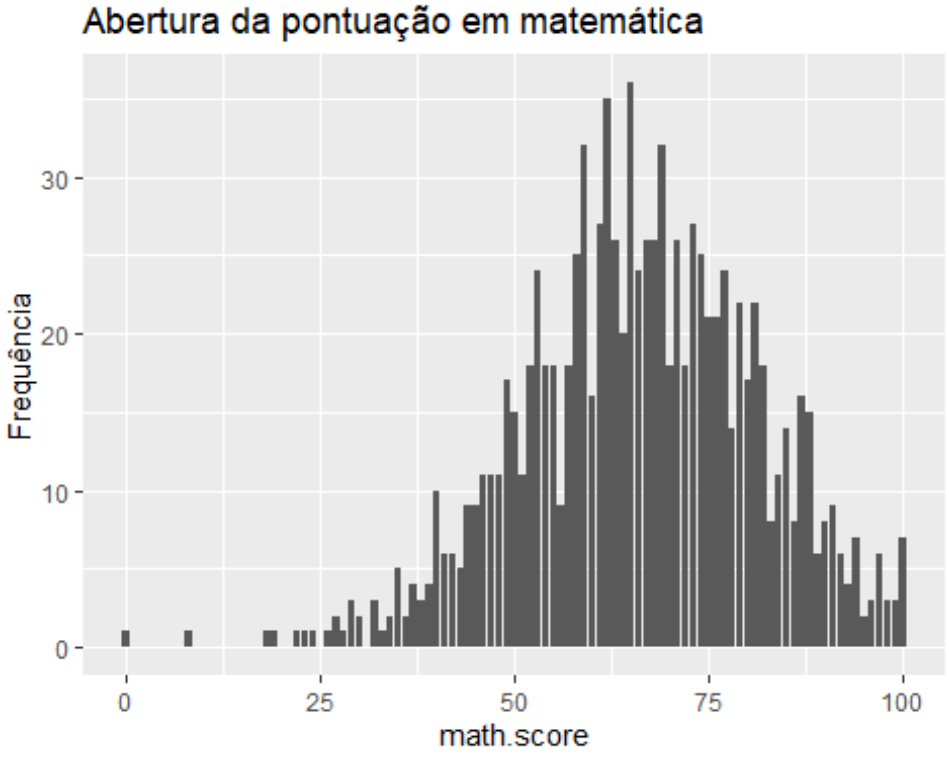

Abertura de quantidade de alunos por sexo
Frequência
500
400
300
200
100
0
female
male
gender
gender
female
male

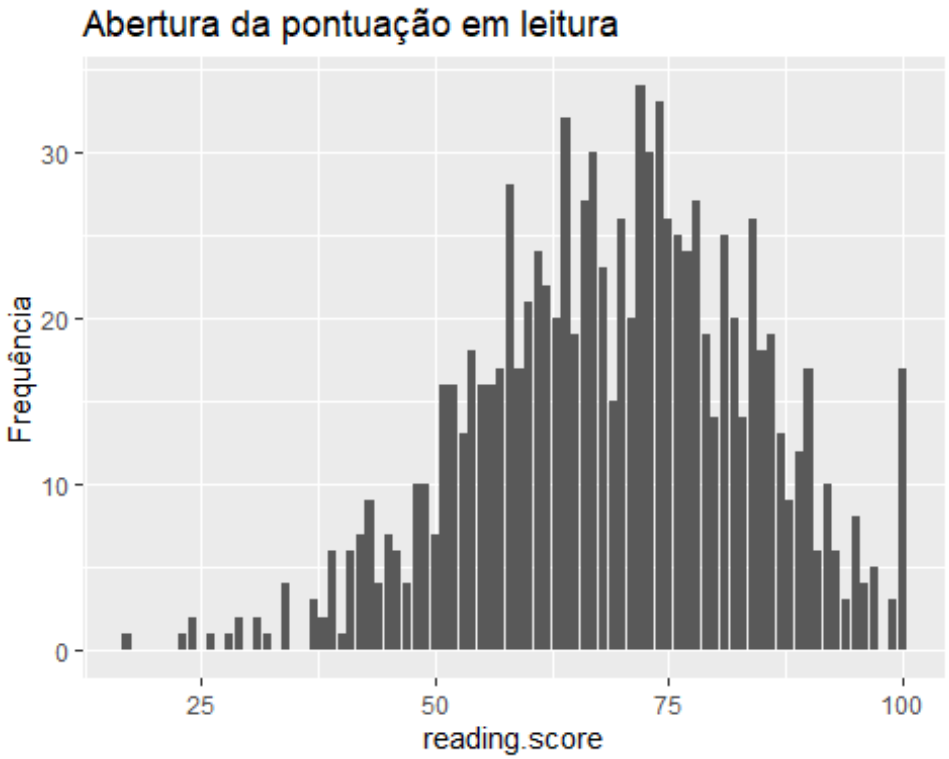

Abertura por etinia
Frequência
300
200
100
0
group A
group B
group C
group D
group E
race.ethnicity
race.ethnicity
group A
group B
group C
group D
group E

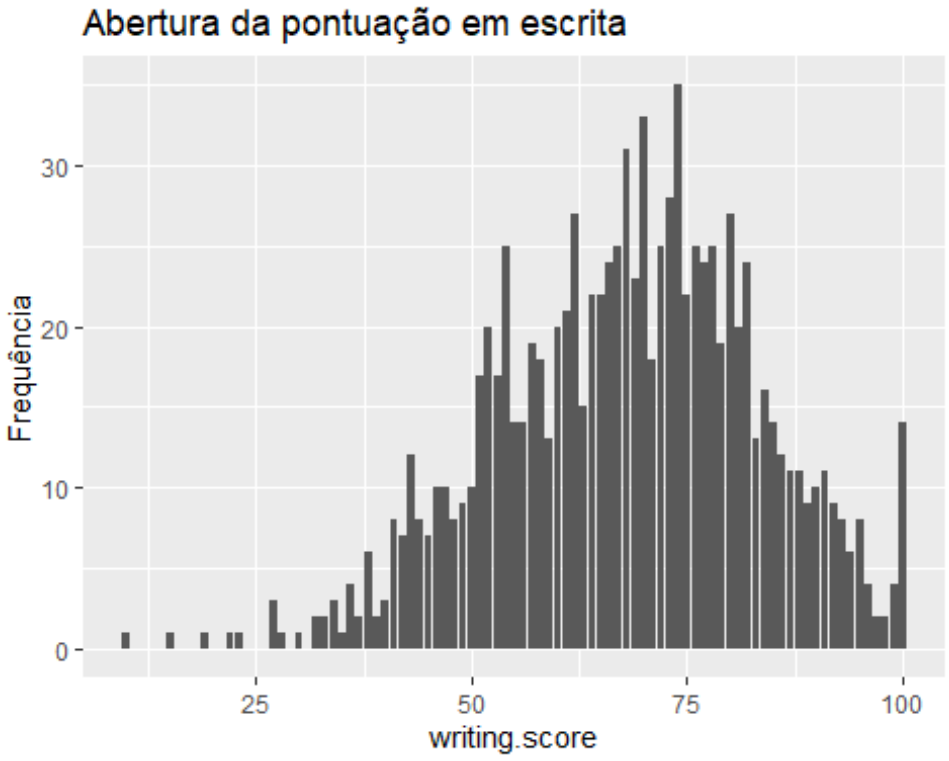

Abertura do nível de educação dos pais
Frequência
200
150
100
50
0
associate's degree
bachelor's degree
high school
master's degree
some college
some high school
parental.level.of.education


Abertura de o que teve no almoço
Frequência
600
400
200
0
free/reduced
standard
lunch


Se fez curso preparatório
600
400
200
0
Frequência
completed
none
test.preparation.course

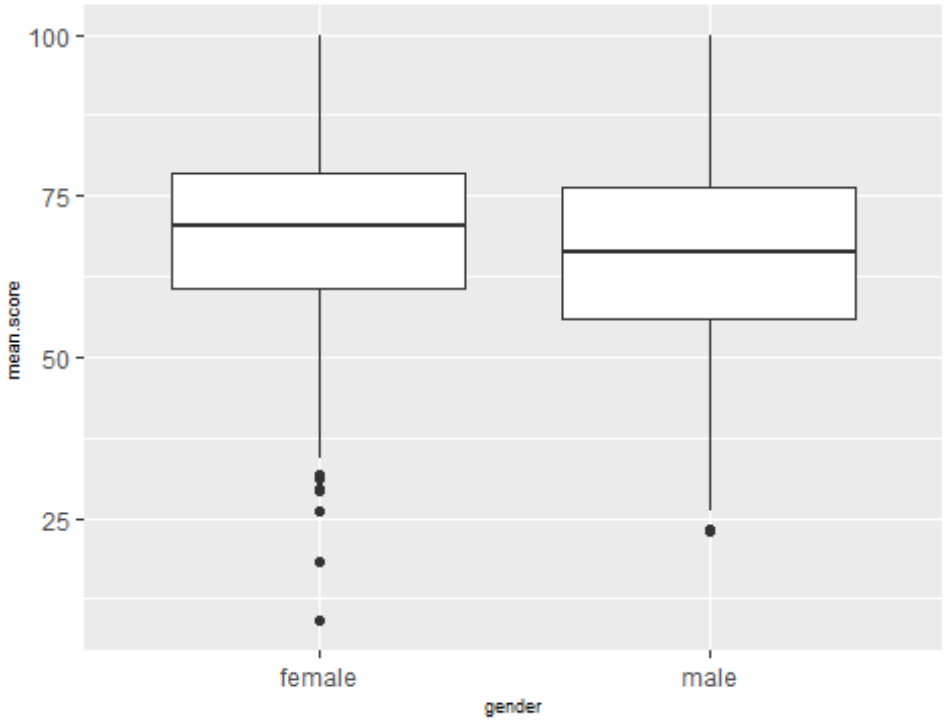

Abertura da pontuação em matemática
30
20
10
0
Frequência
0
25
50
75
100
math.score



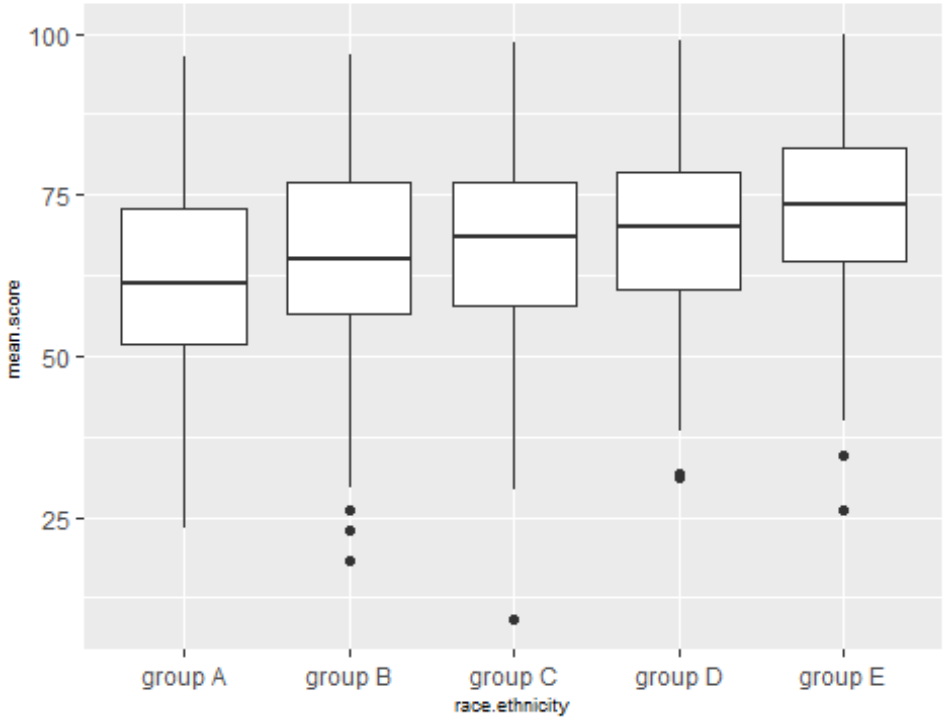


Dado o objetivo em questão (entender quais variáveis impactam na pontuação média dos testes), vamos criar uma variável com essa pontuação média e alguns gráficos mais detalhados, para entender a relação entre as variáveis e essa nova variável.

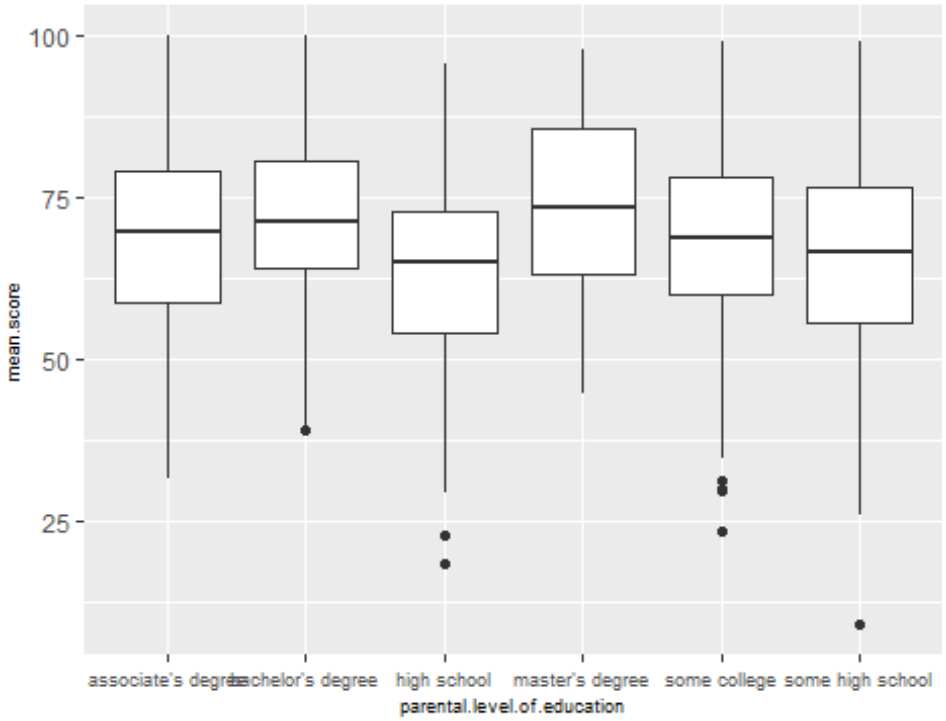
Boxplot do género com a pontuação média



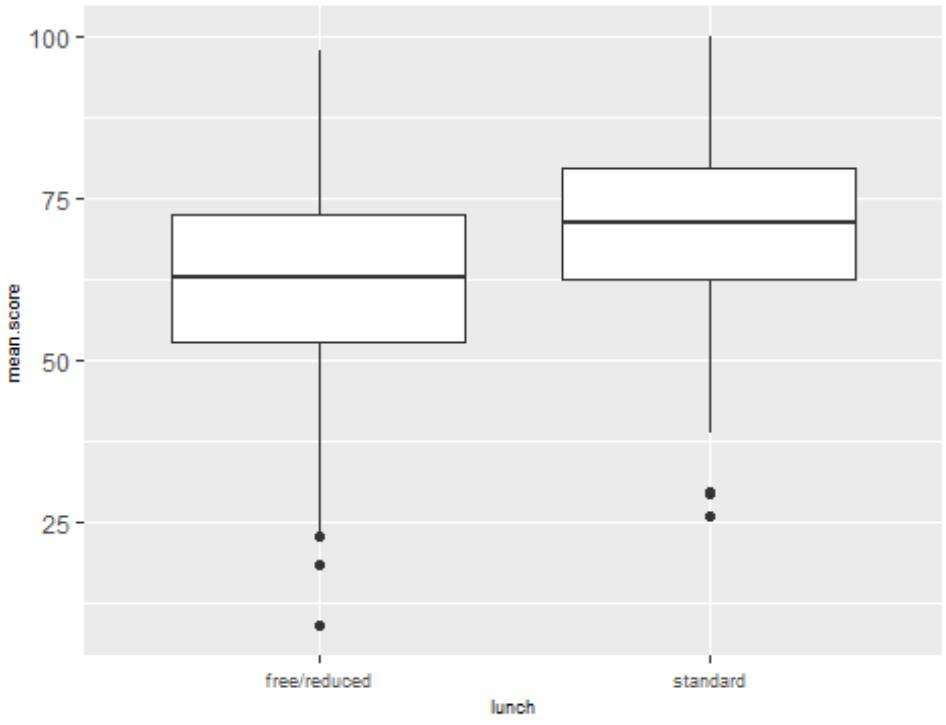
Boxplot da etnia com a pontuação média

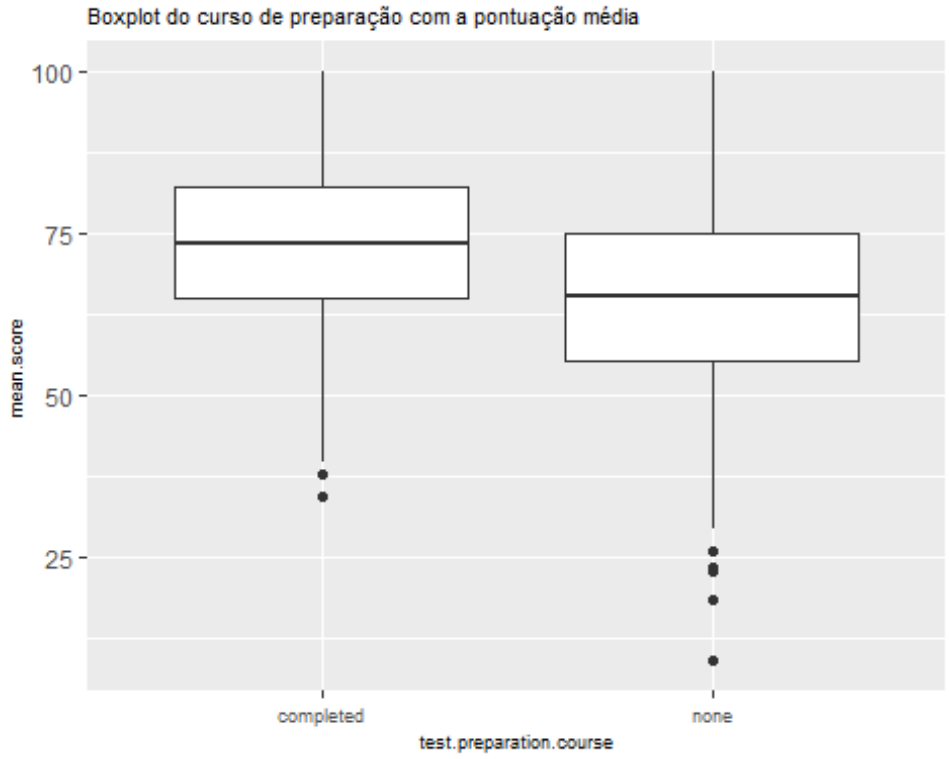


Boxplot do nível de educação dos pais com a pontuação média



Boxplot de o que a pessoa teve de almoço com a pontuação média





Ajuste do modelo

Vamos criar um modelo inicial, e como primeira versão, vamos adicionar todas as covariáveis ao mesmo.

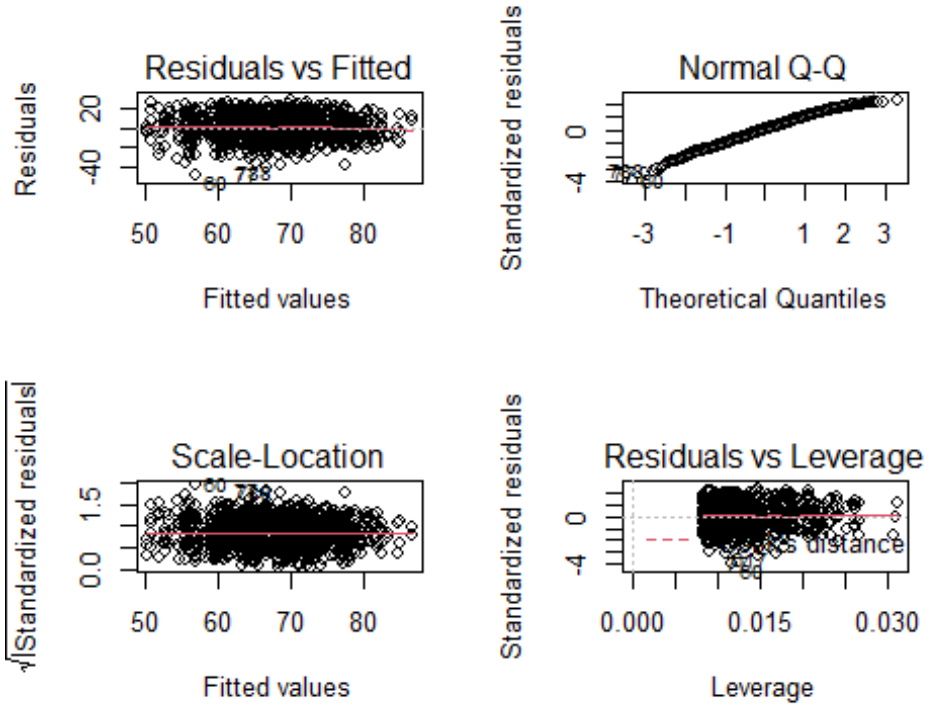
A estratégia principal, é ver o impacto que podemos ter com todas as variáveis trabalhando em conjunto, caso seja necessário algum tipo de ajuste, vamos verificar através dos gráficos de falta de ajuste, variabilidade, normalidade e potencial de observações ter alavancamento na reta.

```
##
## Call:
## lm(formula = mean.score ~ gender + race.ethnicity +
parental.level.of.education +
##     test.preparation.course + lunch, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.148  -8.298   0.646   8.736  27.522
##
## Coefficients:
##                                            Estimate Std. Error t
value
## (Intercept)                                 66.9408     1.7750
37.714
## gendermale                                  -3.7242     0.7955  -
4.682
```

```
## race.ethnicitygroup B                             1.5290     1.6116
0.949
## race.ethnicitygroup C                             2.3855     1.5093
1.581
## race.ethnicitygroup D                             5.1258     1.5398
3.329
## race.ethnicitygroup E                             6.9285     1.7081
4.056
## parental.level.of.educationbachelor's degree      2.5356     1.4240
1.781
## parental.level.of.educationhigh school           -5.1725     1.2298  -
4.206
## parental.level.of.educationmaster's degree        4.0922     1.8377
2.227
## parental.level.of.educationsome college          -0.9275     1.1823  -
0.785
## parental.level.of.educationsome high school      -4.5400     1.2639  -
3.592
## test.preparation.coursenone                      -7.6386     0.8302  -
9.201
## lunchstandard                                     8.7751     0.8275
10.605
##                                                 Pr(>|t|)
## (Intercept)                                      < 2e-16 ***
## gendermale                                      3.24e-06 ***
## race.ethnicitygroup B                           0.342983
## race.ethnicitygroup C                           0.114296
## race.ethnicitygroup D                           0.000904 ***
## race.ethnicitygroup E                           5.38e-05 ***
## parental.level.of.educationbachelor's degree    0.075287 .
## parental.level.of.educationhigh school          2.84e-05 ***
## parental.level.of.educationmaster's degree      0.026185 *
## parental.level.of.educationsome college         0.432934
## parental.level.of.educationsome high school     0.000344 ***
## test.preparation.coursenone                      < 2e-16 ***
## lunchstandard                                    < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.49 on 987 degrees of freedom
## Multiple R-squared:  0.2423, Adjusted R-squared:  0.2331
## F-statistic:  26.3 on 12 and 987 DF,  p-value: < 2.2e-16
```

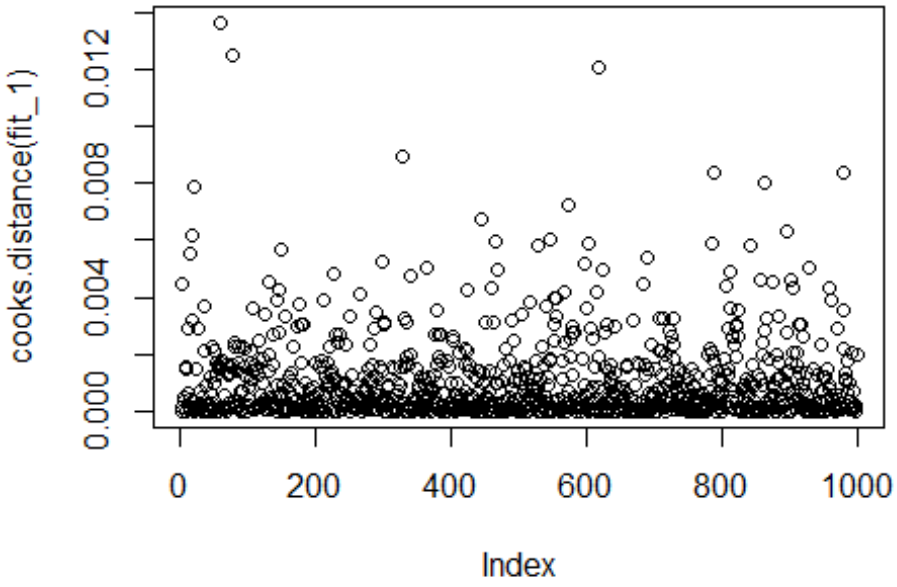
Pudemos ver que nossas covaríaveis foram estatísticamente significantes para explicar a resposta. No entanto, algo que podemos perceber, que o modelo explica pouca variabilidade da resposta, apenas 24%.

Vejamos os gráficos do modelo:



Com os gráficos obtidos podemos notar que não necessitamos de ajustes no modelo construído, isso prova que a estratégia adotada para a construção do modelo de regressão foi bem definida por essa razão nós não necessitamos de nenhum tipo de ajuste com a base de dados, ou transformações nas variáveis para readequar melhor o modelo.

```
plot(cooks.distance(fit_1))
```



Já com o modelo bem definido e sem problemas de qualquer tipo, avaliamos os valores de cooks e podemos observar que não possuímos valores que sejam necessários a intervenção de algum tipo de investigação. Pois as influências das observações parecem ser bem homogêneas.

Poderíamos ter um problema grande com as observações se possuíssimos alta alavancagem juntamente com um alto resíduo, isso nos iria garantir certamente problemas de ajuste que posteriormente deveríamos resolver com tratamentos específicos, o que não vem ao caso em nosso modelo.

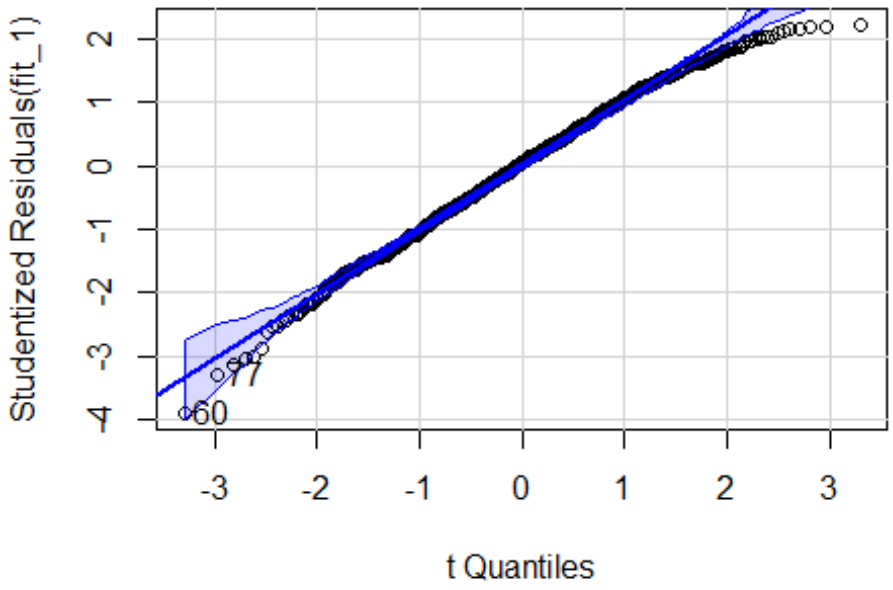
```
library(car)

## Warning: package 'car' was built under R version 4.1.2

## Carregando pacotes exigidos: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.1.2
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
qqPlot(fit_1)
```




```
## [1] 60 77
```

Através da biblioteca Car, podemos plotar o gráfico qqPlot que nos apresenta isoladamente o gráfico que investiga a nornmalidade mais a fundo e é possível notar a fuga da normalidade bem nos pontos extremos, destacando para o ponto mais extremo positivo, podemos avaliar que não existe um impacto tão significativo para a necessidade de reajuste do modelo.

```
x<-covratio(fit_1)
length(x[(1.05> x) & (x >1.00)])
```

```
## [1] 831
```

```
length(x[x>1.05])
```

```
## [1] 0
```

Através da função covratio, podemos observar também as variâncias totais de cada observação das 1000 que temos disponíveis na base de dados, a importância é que uma variância acima de 1 leva a crer que a mesma deve apresentar resíduos grandes, o que pode ser preocupante. Para avaliarmos esse impacto, foi plotado a contagem de quantas observações possuímos em dois intervalos. Tivemos 831 contagens de valores que estão entre 1.05 e 1.03 e nenhum valor acima de 1.05.

Infelizmente não é possível a utilização da função p.adjust para o modelo que foi selecionado, a razão é por sintaxe da função. Porém é fácil observar através dos boxplots já plotados anteriormente que não possuímos nenhum tipo de outlier significativo.

##Testando modelos diferentes, avaliando o impacto das variáveis.

```
fit_3 <- lm(mean.score ~ gender + race.ethnicity +
parental.level.of.education +
              test.preparation.course, data= df)

summary(fit_3)

##
## Call:
## lm(formula = mean.score ~ gender + race.ethnicity +
parental.level.of.education +
##     test.preparation.course, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.967  -8.769   0.264   9.515  29.897
##
## Coefficients:
##                                             Estimate Std. Error t
value
## (Intercept)                                  71.9515     1.8049
39.866
## gendermale                                   -3.5437     0.8389  -
4.224
## race.ethnicitygroup B                         1.9256     1.6996
1.133
## race.ethnicitygroup C                         2.8591     1.5914
1.797
## race.ethnicitygroup D                         5.5336     1.6239
3.408
## race.ethnicitygroup E                         7.9678     1.7989
4.429
## parental.level.of.educationbachelor's degree  2.3442     1.5021
1.561
## parental.level.of.educationhigh school       -5.2325     1.2973  -
4.033
## parental.level.of.educationmaster's degree    3.5976     1.9380
```

```
1.856
## parental.level.of.educationsome college       -0.9342     1.2472  -
0.749
## parental.level.of.educationsome high school   -4.3959     1.3333  -
3.297
## test.preparation.coursenone                    -7.4475     0.8756  -
8.506
##                                                Pr(>|t|)
## (Intercept)                                     < 2e-16 ***
## gendermale                                     2.62e-05 ***
## race.ethnicitygroup B                          0.257494
## race.ethnicitygroup C                          0.072708 .
## race.ethnicitygroup D                          0.000682 ***
## race.ethnicitygroup E                          1.05e-05 ***
## parental.level.of.educationbachelor's degree 0.118944
## parental.level.of.educationhigh school         5.92e-05 ***
## parental.level.of.educationmaster's degree     0.063695 .
## parental.level.of.educationsome college        0.454039
## parental.level.of.educationsome high school    0.001012 **
## test.preparation.coursenone                     < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.17 on 988 degrees of freedom
## Multiple R-squared:  0.1559, Adjusted R-squared:  0.1465
## F-statistic: 16.59 on 11 and 988 DF,  p-value: < 2.2e-16

fit_4 <- lm(mean.score ~ gender + race.ethnicity +
parental.level.of.education +
               + lunch, data= df)

summary(fit_4)

##
## Call:
## lm(formula = mean.score ~ gender + race.ethnicity +
parental.level.of.education +
##     +lunch, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.629  -8.694   0.637   9.018  31.679
##
## Coefficients:
##                                                Estimate Std. Error t
value
## (Intercept)                                     62.0108     1.7623
35.187
## gendermale                                      -3.6517     0.8284  -
4.408
```

```
## race.ethnicitygroup B                             1.6911     1.6783
1.008
## race.ethnicitygroup C                             2.6158     1.5716
1.664
## race.ethnicitygroup D                             4.9137     1.6035
3.064
## race.ethnicitygroup E                             7.6287     1.7772
4.293
## parental.level.of.educationbachelor's degree      2.7125     1.4830
1.829
## parental.level.of.educationhigh school           -5.7761     1.2790  -
4.516
## parental.level.of.educationmaster's degree        3.9397     1.9138
2.059
## parental.level.of.educationsome college          -1.1033     1.2312  -
0.896
## parental.level.of.educationsome high school      -3.9981     1.3149  -
3.041
## lunchstandard                                     8.6099     0.8616
9.993
##                                                 Pr(>|t|)
## (Intercept)                                      < 2e-16 ***
## gendermale                                      1.16e-05 ***
## race.ethnicitygroup B                            0.31389
## race.ethnicitygroup C                            0.09635 .
## race.ethnicitygroup D                            0.00224 **
## race.ethnicitygroup E                           1.94e-05 ***
## parental.level.of.educationbachelor's degree     0.06768 .
## parental.level.of.educationhigh school          7.06e-06 ***
## parental.level.of.educationmaster's degree       0.03980 *
## parental.level.of.educationsome college          0.37042
## parental.level.of.educationsome high school      0.00242 **
## lunchstandard                                    < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13 on 988 degrees of freedom
## Multiple R-squared:  0.1773, Adjusted R-squared:  0.1681
## F-statistic: 19.35 on 11 and 988 DF,  p-value: < 2.2e-16

fit_5 <- lm(mean.score ~ gender + race.ethnicity +
test.preparation.course + lunch, data= df)

summary(fit_5)

##
## Call:
## lm(formula = mean.score ~ gender + race.ethnicity +
test.preparation.course +
##     lunch, data = df)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.499  -8.248   0.161   9.150  30.829
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   65.2752     1.6169  40.370  < 2e-16 ***
## gendermale                    -4.0110     0.8136  -4.930 9.64e-07 ***
## race.ethnicitygroup B          1.4720     1.6479   0.893 0.371942
## race.ethnicitygroup C          2.9489     1.5399   1.915 0.055780 .
## race.ethnicitygroup D          5.7458     1.5720   3.655 0.000271 ***
## race.ethnicitygroup E          7.8218     1.7400   4.495 7.77e-06 ***
## test.preparation.coursenone   -7.7255     0.8464  -9.127  < 2e-16 ***
## lunchstandard                  8.6320     0.8474  10.187  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.79 on 992 degrees of freedom
## Multiple R-squared:  0.2004, Adjusted R-squared:  0.1947
## F-statistic: 35.51 on 7 and 992 DF,  p-value: < 2.2e-16
```

```
fit_6 <- lm(mean.score ~ gender + parental.level.of.education +
              test.preparation.course + lunch, data= df)

summary(fit_6)
```

```
##
## Call:
## lm(formula = mean.score ~ gender + parental.level.of.education +
##     test.preparation.course + lunch, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.663  -8.126   0.402   9.334  31.008
##
## Coefficients:
##                                              Estimate Std. Error t
value
## (Intercept)                                   70.3041     1.1950
58.832
## gendermale                                    -3.6458     0.8015  -
4.549
## parental.level.of.educationbachelor's degree   2.3893     1.4404
1.659
## parental.level.of.educationhigh school        -5.5783     1.2415  -
4.493
## parental.level.of.educationmaster's degree     4.4828     1.8536
2.418
## parental.level.of.educationsome college       -0.8459     1.1947  -
```

```
0.708
## parental.level.of.educationsome high school   -4.9326     1.2709  -
3.881
## test.preparation.coursenone                    -7.7081     0.8380  -
9.198
## lunchstandard                                   8.9821     0.8361
10.742
##                                                Pr(>|t|)
## (Intercept)                                     < 2e-16 ***
## gendermale                                     6.07e-06 ***
## parental.level.of.educationbachelor's degree 0.097472 .
## parental.level.of.educationhigh school         7.84e-06 ***
## parental.level.of.educationmaster's degree     0.015769 *
## parental.level.of.educationsome college        0.479071
## parental.level.of.educationsome high school    0.000111 ***
## test.preparation.coursenone                     < 2e-16 ***
## lunchstandard                                   < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.64 on 991 degrees of freedom
## Multiple R-squared:  0.2202, Adjusted R-squared:  0.2139
## F-statistic: 34.98 on 8 and 991 DF,  p-value: < 2.2e-16
```

```
fit_7 <-  lm(mean.score ~ race.ethnicity + parental.level.of.education +
              test.preparation.course, data= df)

summary(fit_7)
```

```
##
## Call:
## lm(formula = mean.score ~ race.ethnicity + parental.level.of.education
+
##     test.preparation.course, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.425  -8.782   0.489   9.125  31.677
##
## Coefficients:
##                                             Estimate Std. Error t
value
## (Intercept)                                  69.8295     1.7482
39.943
## race.ethnicitygroup B                         2.4417     1.7096
1.428
## race.ethnicitygroup C                         3.4159     1.5994
2.136
## race.ethnicitygroup D                         5.8201     1.6362
3.557
```

```
## race.ethnicitygroup E                            8.2667     1.8127
4.560
## parental.level.of.educationbachelor's degree   2.4034     1.5148
1.587
## parental.level.of.educationhigh school        -5.3856     1.3078  -
4.118
## parental.level.of.educationmaster's degree     3.9312     1.9528
2.013
## parental.level.of.educationsome college       -0.9143     1.2578  -
0.727
## parental.level.of.educationsome high school   -4.4080     1.3446  -
3.278
## test.preparation.coursenone                   -7.4126     0.8830  -
8.395
##                                                Pr(>|t|)
## (Intercept)                                     < 2e-16 ***
## race.ethnicitygroup B                          0.153529
## race.ethnicitygroup C                          0.032944 *
## race.ethnicitygroup D                          0.000393 ***
## race.ethnicitygroup E                          5.75e-06 ***
## parental.level.of.educationbachelor's degree 0.112920
## parental.level.of.educationhigh school       4.14e-05 ***
## parental.level.of.educationmaster's degree   0.044374 *
## parental.level.of.educationsome college      0.467449
## parental.level.of.educationsome high school  0.001080 **
## test.preparation.coursenone                     < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.28 on 989 degrees of freedom
## Multiple R-squared:  0.1407, Adjusted R-squared:  0.132
## F-statistic: 16.19 on 10 and 989 DF,  p-value: < 2.2e-16
```

O objetivo desse teste é visualizarmos qual o impacto que cada variável tem dentro do modelo escolhido estratégicamente.

Podemos notar que as variáveis que menos garantem impacto no modelo são: -Race.ethnicity -Parental.level.of.education

Instintivamente podemos notar que são variáveis que com certeza possuem algum tipo de influência nas notas que são obtidas pelos alunos.

Segundo Tania Regina da Silva, graduada em Pedagogia em 1996 pela UFPR, o incentivo dos pais é fundamental para o desenvolvimento de um estudante em uma escola e o grau de escolaridade dos pais é importantíssimo para eventualmente incentivar os filhos a se desenvolver mais, o que o resultado obtido se torna uma contradição.

Isso nos incentivou a buscar mais respostas com outros modelos, dessa vez desvinculando a variável “mean.score” e buscando respostas com base nas variáveis

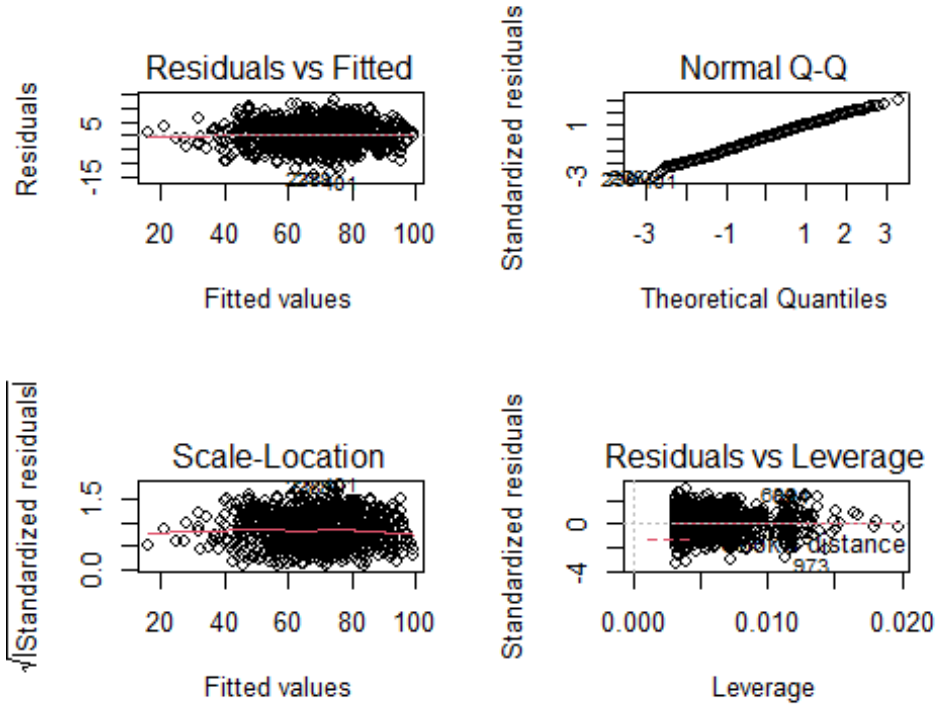
das notas isoladas dos alunos, math score, reading score e writing score. Ocasionando então na evolução do grupo B de questionamentos:

- A relação da nota de um aluno em Escrita, tem a ver com a nota que ele obteve em Leitura e seu gênero?
- A relação da nota de um aluno em Matemática, tem a ver com a nota que ele obteve com a Escrita e seu gênero?

```
fit_read <- lm(reading.score ~ writing.score  + race.ethnicity, data= df)

summary(fit_read)
```

```
##
## Call:
## lm(formula = reading.score ~ writing.score + race.ethnicity,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.3421  -2.8501   0.2019   3.0352  12.5028
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               7.01282    0.73073   9.597  < 2e-16 ***
## writing.score             0.92002    0.00910 101.096  < 2e-16 ***
## race.ethnicitygroup B    -0.01335    0.55416  -0.024  0.98078
## race.ethnicitygroup C    -0.31195    0.51871  -0.601  0.54771
## race.ethnicitygroup D    -1.51697    0.53305  -2.846  0.00452 **
## race.ethnicitygroup E     0.31991    0.58958   0.543  0.58752
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.309 on 994 degrees of freedom
## Multiple R-squared:  0.9133, Adjusted R-squared:  0.9129
## F-statistic:  2095 on 5 and 994 DF,  p-value: < 2.2e-16
```

```
#Plotando os gráficos de pressupostos
par(mfrow=c(2,2))
plot(fit_read)
```

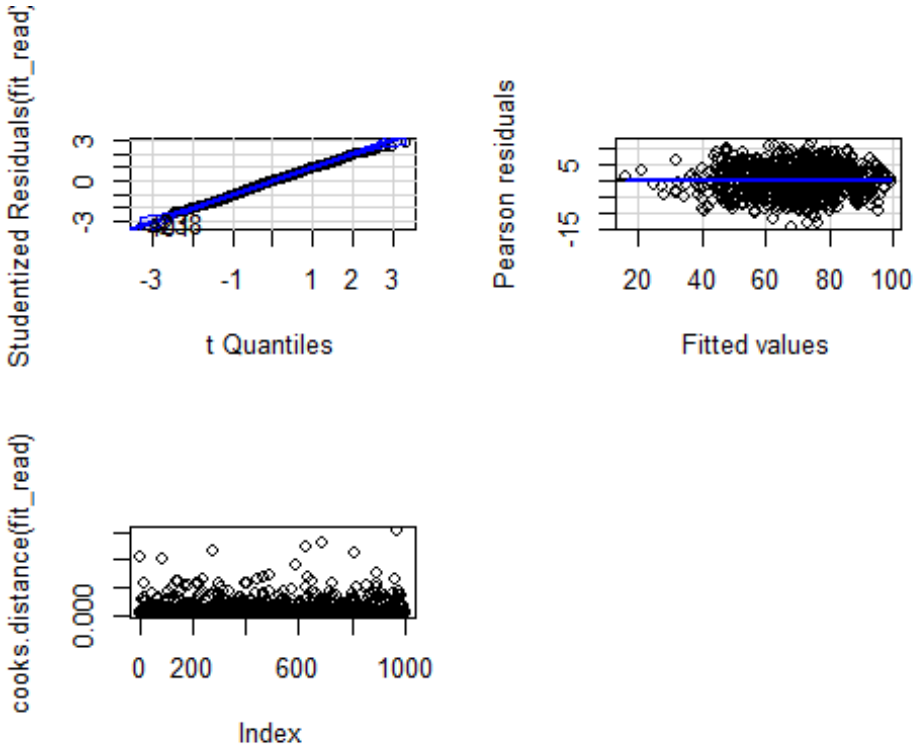



```
library("car")
#-----------------------------------

#Verificando a integridade do Modelo.

#Plotando o a normalidade do modelo.
qqPlot(fit_read)

## [1] 238 401

#Plotando o ajuste do modelo
residualPlot(fit_read)

#Plotando as observações mais influentes avaliando valor DFfits.

x<-influence.measures(fit_read)

summary(x)

## Potentially influential observations of
##   lm(formula = reading.score ~ writing.score + race.ethnicity,
data = df) :
##
##     dfb.1_ dfb.wrt. dfb.rc.B dfb.rc.C dfb.rc.D dfb.rc.E dffit   cov.r
cook.d
## 4    0.22  -0.09    -0.19    -0.20    -0.19    -0.17     0.25_*  0.99
0.01
```

```
## 60   0.02  -0.03     0.00     0.01     0.00     0.00     0.03
1.02_*  0.00
## 62   0.02  -0.01    -0.01    -0.01    -0.01    -0.01     0.02
1.02_*  0.00
## 71   0.01  -0.01     0.00     0.00    -0.07     0.00    -0.14
0.98_*  0.00
## 73   0.00   0.00     0.00     0.00     0.00     0.00     0.00
1.02_*  0.00
## 77  -0.03   0.04     0.00     0.00     0.00    -0.02    -0.05
1.02_*  0.00
## 83  -0.18   0.04     0.20     0.21     0.20     0.18    -0.25_*  0.99
0.01
## 222 -0.03   0.04     0.10     0.00    -0.01    -0.01     0.19
0.97_*  0.01
## 238  0.02  -0.03     0.00     0.00    -0.09     0.00    -0.20
0.95_*  0.01
## 278  0.02  -0.02     0.00     0.00     0.00    -0.16    -0.27_*
0.96_*  0.01
## 300 -0.02   0.03     0.00     0.00     0.09     0.00     0.18
0.96_*  0.01
## 304  0.01  -0.01     0.10     0.00     0.00     0.00     0.17
0.98_*  0.00
## 315  0.03  -0.04     0.00    -0.07     0.01     0.01    -0.16
0.97_*  0.00
## 328 -0.05   0.03     0.03     0.03     0.03     0.02    -0.05
1.03_*  0.00
## 388  0.00   0.00     0.00     0.06     0.00     0.00     0.13
0.98_*  0.00
## 396  0.01   0.00    -0.01    -0.01    -0.01     0.00     0.01
1.02_*  0.00
## 401  0.00   0.01     0.00    -0.09     0.00     0.00    -0.19
0.94_*  0.01
## 487 -0.05   0.07     0.00    -0.01    -0.08    -0.01    -0.16
0.98_*  0.00
## 490  0.00   0.00     0.01     0.01     0.01     0.01    -0.01
1.02_*  0.00
## 547  0.01   0.06    -0.08    -0.09    -0.09    -0.08     0.11
1.02_*  0.00
## 551  0.01  -0.01     0.00     0.06     0.00     0.00     0.12
0.98_*  0.00
## 597  0.06  -0.08     0.03     0.01     0.01     0.01     0.10
1.02_*  0.00
## 615  0.00   0.01    -0.01    -0.01    -0.01    -0.01     0.01
1.02_*  0.00
## 624  0.06   0.11    -0.21    -0.23    -0.23    -0.21     0.27_*  0.99
0.01
## 669  0.03  -0.04     0.00     0.07     0.00     0.01     0.15
0.97_*  0.00
## 689  0.24  -0.10    -0.21    -0.22    -0.21    -0.19     0.28_*  0.98
0.01
```

```
## 732 -0.01   0.01     0.01     0.01     0.01     0.01     -0.02
1.02_*  0.00
## 811 -0.24   0.12     0.18     0.19     0.18     0.16     -0.26_*  0.99
0.01
## 821 -0.01  -0.02     0.04     0.04     0.04     0.04     -0.05
1.02_*  0.00
## 836  0.03  -0.03     0.00    -0.06     0.00     0.00     -0.14
0.97_*  0.00
## 864  0.04  -0.06     0.00    -0.05     0.01     0.01     -0.13
0.98_*  0.00
## 888  0.04  -0.05     0.00     0.07     0.01     0.01      0.15
0.97_*  0.00
## 899 -0.03   0.03     0.00     0.00    -0.07     0.00     -0.14
0.98_*  0.00
## 912  0.01   0.01    -0.03    -0.04    -0.03    -0.03      0.04
1.02_*  0.00
## 928  0.00   0.00     0.00     0.00    -0.07     0.00     -0.14
0.98_*  0.00
## 973 -0.20   0.02     0.25     0.27     0.26     0.23     -0.31_*
0.97_*  0.02
##     hat
## 4    0.01
## 60   0.02_*
## 62   0.01
## 71   0.00
## 73   0.01
## 77   0.02_*
## 83   0.01
## 222  0.01
## 238  0.00
## 278  0.01
## 300  0.00
## 304  0.01
## 315  0.00
## 328  0.02_*
## 388  0.00
## 396  0.01
## 401  0.00
## 487  0.00
## 490  0.01
## 547  0.02
## 551  0.00
## 597  0.02
## 615  0.02
## 624  0.01
## 669  0.00
## 689  0.01
## 732  0.01
## 811  0.01
## 821  0.02
```

```
## 836  0.00
## 864  0.00
## 888  0.00
## 899  0.00
## 912  0.01
## 928  0.00
## 973  0.01

plot(cooks.distance(fit_read))
```



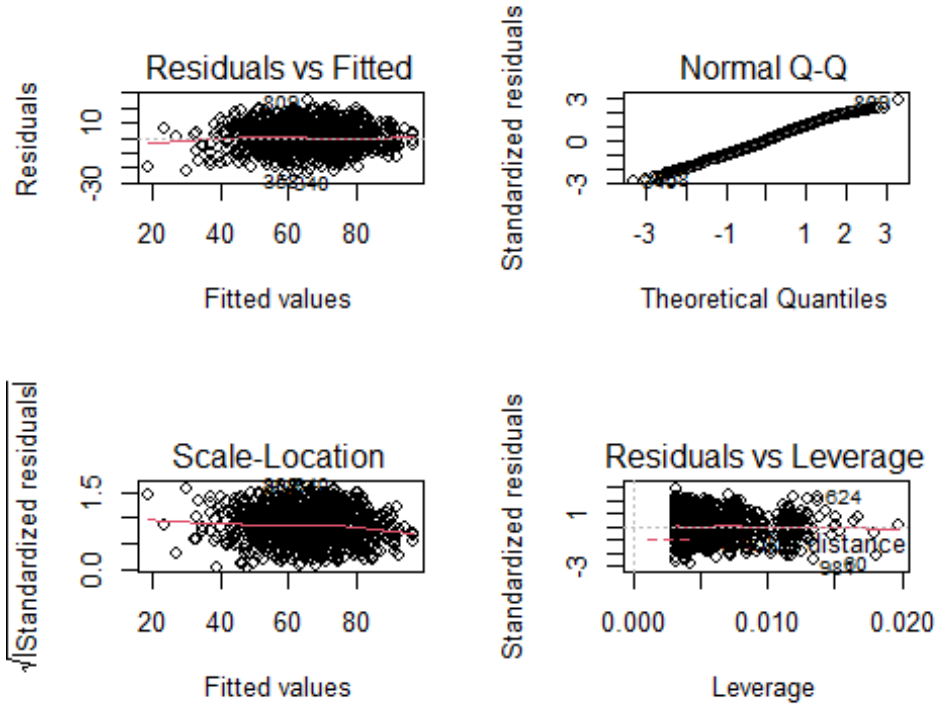
Inicialmente já podemos notar a diferença de explicação do modelo, esse modelo possui um nível de explicação de mais de 90%. O que nos faz acreditar que a nota de um aluno em Leitura, é explicado pela nota que obteve em escrita. Porém, a Etnia não é bem explicativa nesse modelo, apenas apresentando que a etnia D é mais significativa.

Podemos notar também a integridade do modelo, variância, ajuste bem definida, assim como através da função qqPlot, podemos notar a normalização do modelo.

```
fit_math <- lm(math.score ~ writing.score  + race.ethnicity, data= df)

summary(fit_math)

##
## Call:
## lm(formula = math.score ~ writing.score + race.ethnicity, data = df)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.357  -6.294  -0.118   6.454  24.611
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              12.16973    1.49561   8.137 1.20e-15 ***
## writing.score             0.78915    0.01863  42.368  < 2e-16 ***
## race.ethnicitygroup B    -0.48552    1.13422  -0.428    0.669
## race.ethnicitygroup C    -1.23211    1.06166  -1.161    0.246
## race.ethnicitygroup D    -0.16228    1.09102  -0.149    0.882
## race.ethnicitygroup E     5.30056    1.20671   4.393 1.24e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.82 on 994 degrees of freedom
## Multiple R-squared:  0.6634, Adjusted R-squared:  0.6617
## F-statistic: 391.7 on 5 and 994 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2, 2))
plot(fit_math)
```




```
qqPlot(fit_math)
```

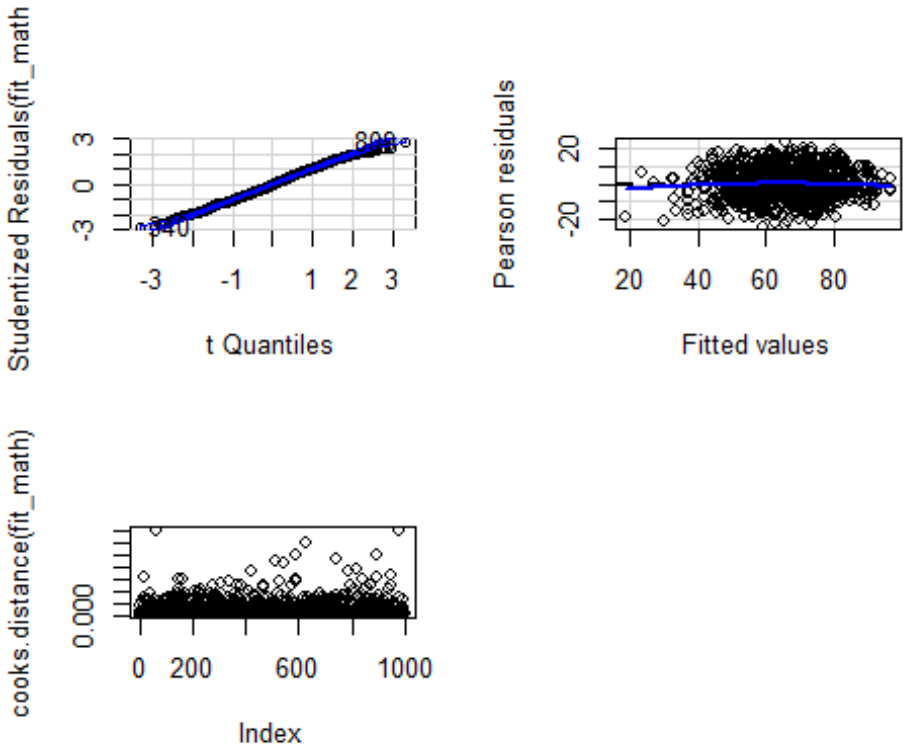
```
## [1] 340 809
```

```
residualPlot(fit_math)

y<-influence.measures(fit_math)

summary(y)
```

```
## Potentially influential observations of
##   lm(formula = math.score ~ writing.score + race.ethnicity, data = df)
:
##
##     dfb.1_ dfb.wrt. dfb.rc.B dfb.rc.C dfb.rc.D dfb.rc.E dffit   cov.r
cook.d
## 4    0.00   0.00     0.00     0.00     0.00     0.00     0.00
1.02_*  0.00
## 60  -0.21   0.27    -0.01    -0.08    -0.03    -0.04    -0.29_*  1.00
0.01
## 62   0.00   0.00     0.00     0.00     0.00     0.00     0.00
1.02_*  0.00
## 77  -0.05   0.06     0.00    -0.01    -0.01    -0.04    -0.07
1.02_*  0.00
## 297  0.00   0.00     0.00     0.00     0.00     0.00     0.00
1.02_*  0.00
## 328  0.01  -0.01    -0.01    -0.01    -0.01    -0.01     0.01
1.03_*  0.00
## 340 -0.05   0.06     0.00    -0.01    -0.09    -0.01    -0.18
0.96_*  0.01
## 358  0.01  -0.01     0.00    -0.07     0.00     0.00    -0.15
0.97_*  0.00
## 372  0.01  -0.01     0.00    -0.06     0.00     0.00    -0.14
0.97_*  0.00
## 379 -0.01  -0.01     0.02     0.02     0.02     0.02    -0.02
1.02_*  0.00
## 396  0.04  -0.02    -0.03    -0.04    -0.03    -0.03     0.05
1.02_*  0.00
## 415  0.05  -0.06     0.00    -0.06     0.01     0.01    -0.15
0.97_*  0.00
## 469 -0.01  -0.01     0.02     0.03     0.03     0.02    -0.03
1.02_*  0.00
## 470 -0.03   0.04     0.00     0.06    -0.01    -0.01     0.13
0.98_*  0.00
## 490  0.01   0.01    -0.02    -0.02    -0.02    -0.02     0.03
1.02_*  0.00
## 512  0.20  -0.09    -0.18    -0.18    -0.18    -0.16     0.23_*  0.99
0.01
## 529 -0.09   0.11    -0.01    -0.01    -0.09    -0.02    -0.18
0.98_*  0.01
## 547  0.00   0.03    -0.03    -0.04    -0.04    -0.03     0.05
1.02_*  0.00
## 572  0.01   0.04    -0.06    -0.07    -0.07    -0.06     0.09
1.02_*  0.00
```

```
## 592  0.19  -0.06    -0.19    -0.20    -0.19    -0.17     0.24_*  0.99
0.01
## 597  0.06  -0.08     0.03     0.01     0.01     0.01     0.10
1.02_*  0.00
## 615 -0.01  -0.03     0.04     0.04     0.04     0.04    -0.05
1.02_*  0.00
## 624  0.06   0.11    -0.21    -0.22    -0.22    -0.20     0.27_*  0.99
0.01
## 742 -0.17   0.03     0.19     0.20     0.20     0.18    -0.24_*  0.99
0.01
## 809 -0.01   0.01     0.00     0.07     0.00     0.00     0.16
0.96_*  0.00
## 821  0.00   0.00     0.00     0.00     0.00     0.00     0.00
1.02_*  0.00
## 838  0.00   0.00     0.00    -0.01    -0.01     0.00     0.01
1.02_*  0.00
## 845 -0.03   0.03     0.00     0.00    -0.08     0.00    -0.16
0.97_*  0.00
## 860 -0.02   0.02     0.00     0.06     0.00     0.00     0.12
0.98_*  0.00
## 893 -0.11  -0.05     0.20     0.22     0.21     0.19    -0.25_*  0.99
0.01
## 943  0.01  -0.02     0.00     0.06     0.00     0.00     0.13
0.98_*  0.00
## 951  0.00   0.00     0.00     0.00     0.00     0.12     0.20
0.98_*  0.01
## 981 -0.18   0.23    -0.11    -0.02    -0.03    -0.03    -0.29_*  0.98
0.01
## 984 -0.01  -0.04     0.06     0.07     0.07     0.06    -0.08
1.02_*  0.00
##     hat
## 4    0.01
## 60   0.02_*
## 62   0.01
## 77   0.02_*
## 297  0.01
## 328  0.02_*
## 340  0.00
## 358  0.00
## 372  0.00
## 379  0.01
## 396  0.01
## 415  0.00
## 469  0.01
## 470  0.00
## 490  0.01
## 512  0.01
## 529  0.01
## 547  0.02
## 572  0.02
```

```
## 592  0.01
## 597  0.02
## 615  0.02
## 624  0.01
## 742  0.01
## 809  0.00
## 821  0.02
## 838  0.01
## 845  0.00
## 860  0.00
## 893  0.01
## 943  0.00
## 951  0.01
## 981  0.01
## 984  0.01

plot(cooks.distance(fit_math))
```



Com uma taxa de explicação de 66% do modelo, podemos ver que a explicação drásticamente cai. Um aluno que possui habilidades em humanas, pode ter uma disposição a não performar tão bem em uma matéria de exatas. Não podemos ignorar a variável de identidade de raça, onde a raça étnica E, tem uma maior explicação nas notas de matemática.

Conclusão

Com os resultados obtidos através de toda a análise desenvolvida, podemos concluir que os objetivos destacados primordialmente durante as fases iniciais, os resultados que obtivemos é que o modelo escolhido através da estratégia inicial atendeu os requisitos necessários para as análises, porém não foi explicativo o suficiente para prever porque se deve as distribuições das notas obtidas pelos alunos. Modelos diferentes foram testados para tentar atingir uma maior taxa de explicação, mas nada chegou tão próximo do modelo descrito no presente trabalho. O que nos traz a hipótese de que apenas essas variáveis não são o suficiente para definirmos a explicação do que realmente afeta as notas dos alunos, por isso é necessário inclusão de novas pesquisas que possam explicar com maior exatidão qual é a distribuição e o que realmente influencia nessas pontuações, por isso as variáveis que visualmente mais impactam não tem um sentido correto de explicação, pois o impacto não são tão consideráveis.