Lumini's Hire Test

Thiago Pereira

22 de fevereiro de 2020

Este projeto é uma análise de uma fração dos dados dos candidatos do ENEM 2016 como teste para a vaga de Cientista de Dados Sênior da Lumini. Este documento está dividido em:

1.Objetivo

2.Conclusões

3.Configuração de ambiente de trabalho

4.Análise exploratória

`-Por região`

`-Por cor`

`-Por sexo`

`-Por classe social e poder aquisitivo`

`-Por motivação`

`-Por escola de origem`

Além disso, um aplicativo em Shiny para análise interativa dos dados está localizada na pasta Dashboard.

Objetivo

Segmentar os inscritos de forma clara e objetiva com o intuito de justificar investimentos em educação destinados a certas parcelas de alunos.

Conclusões

Pelos dados vistos na análise exploratória, podemos concluir que maiores investimentos em educação deveriam ser destinados a:

1-Escolas estaduais do Ceará, visto que é o estado com maior desigualdade de notas entre os 10% melhores e os 10% piores e com há uma proporção preocupante de alunos com baixa habilidade escrita.

2-A escolas de ensino fundamental em povos indígenas e quilombolas, que obtiveram notas médias muito baixas. No caso dos indígenas, há precedentes de alunos estudaram posteriormente em escolas particulares e obtiveram bons resultados.

3-A filhos de mãe solteira, pois se verificou que não saber a profissão do pai (pai ausente) impacta fortemente a nota do candidato.

4-A alunos de classe média baixa que obtiverem notas altas no ENEM, visto que os alunos com as maiores notas não são necessariamente filhos de pais com nível superior, este incentivo poderia florescer futuros profissionais promissores.

Configuração de ambiente de trabalho

Primeiramente, o diretório de trabalho foi selecionado para ficar igual ao diretório localizado no GitHub(https://github.com/lumini-it-solutions/lumini-hire-test/). Em seguida, foram carregadas as bibliotecas necessárias para realizar a análise.

```
## [1] "C:/Users/valej_000/Documents/Lumini/lumini-hire-test"
```

```
## corrplot 0.84 loaded
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   NU_INSCRICAO = col_character(),
##   NO_MUNICIPIO_RESIDENCIA = col_character(),
##   SG_UF_RESIDENCIA = col_character(),
##   TP_SEXO = col_character(),
##   NO_MUNICIPIO_NASCIMENTO = col_character(),
##   SG_UF_NASCIMENTO = col_character(),
##   NO_MUNICIPIO_ESC = col_character(),
##   SG_UF_ESC = col_character(),
##   NO_ENTIDADE_CERTIFICACAO = col_character(),
##   SG_UF_ENTIDADE_CERTIFICACAO = col_character(),
##   NO_MUNICIPIO_PROVA = col_character(),
##   SG_UF_PROVA = col_character(),
##   CO_PROVA_CN = col_character(),
##   CO_PROVA_CH = col_character(),
##   CO_PROVA_LC = col_character(),
##   CO_PROVA_MT = col_character(),
##   TX_RESPOSTAS_CN = col_character(),
##   TX_RESPOSTAS_CH = col_character(),
##   TX_RESPOSTAS_LC = col_character(),
##   TX_RESPOSTAS_MT = col_character()
##   # ... with 40 more columns
## )
```

```
## See spec(...) for full column specifications.
```

Análise exploratória

A primeira análise foi realizada lendo o data.frame com a função View. Como esta não foi suficiente para ler todas as colunas, utilizei a função fix para ler. (está comentada para evitar problemas)

Por região geográfica

A primeira segmentação foi por região e por estado. Não existe uma coluna região, mas ela está implícita no primeiro dígito das colunas CO_UF_* deste modo: *1=N 2=NE 3=SE 4=S 5=CO* Logo, foi feita uma função para criar as colunas regiões no data.frame.

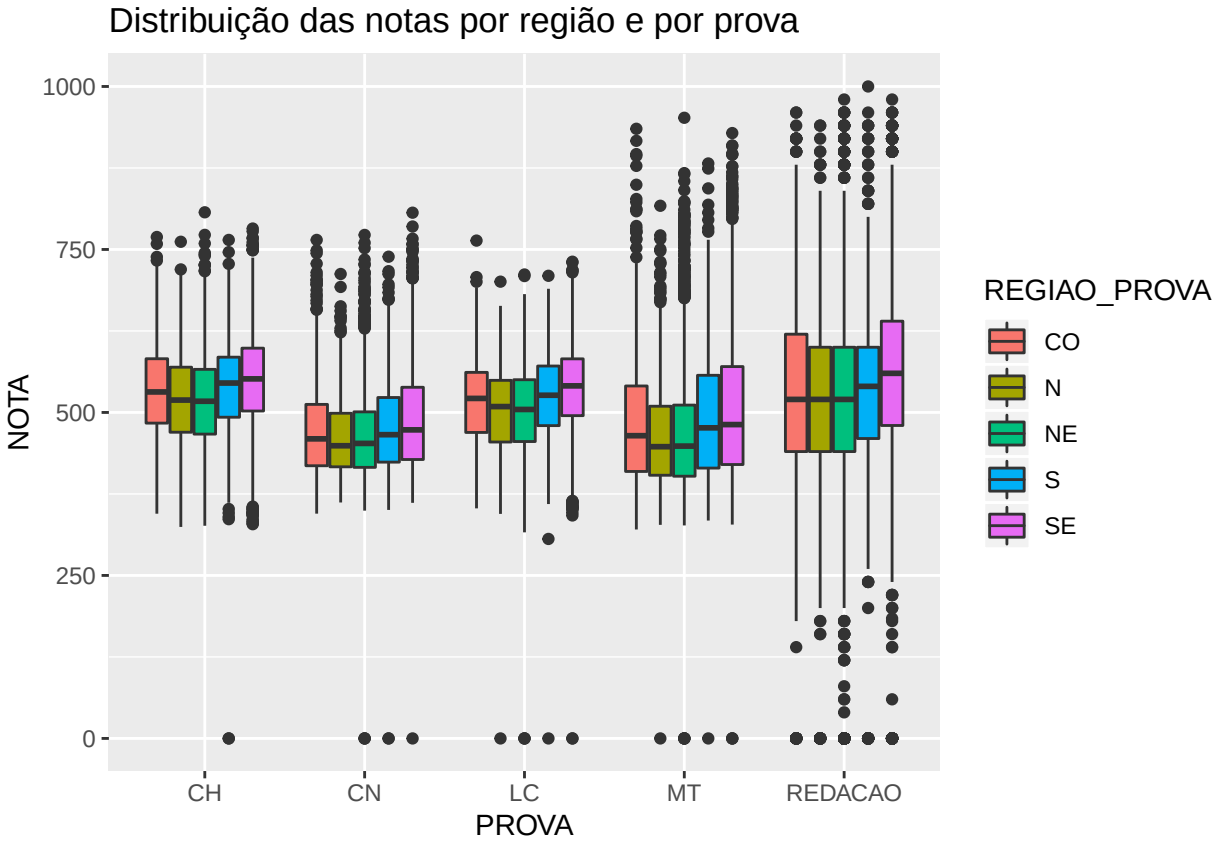
```
## [1] ""   "CO" "N"  "NE" "S"  "SE"
```

```
## [1] "CO" "N"  "NE" "S"  "SE"
```

```
## [1] "CO" "N"  "NE" "S"  "SE"
```

Existem 5 notas no ENEM: Ciências da Natureza(CN), Ciências Humanas(CH), Linguagens e Códigos(LC), Matemática(MT) e redação(REDACAO). Foi criado um no data.frame chamado notas com todas as notas consolidadas na mesma coluna. Como há alunos que realizaram apenas a prova em um dos dias de prova, foi criado o data.frame notas_PRESENTE, que é um filtro da tabela notas excluindo todos os alunos que se ausentaram da prova (reprovação imediata). Este filtro foi feito pois medidas como média e mediana ficariam prejudicadas caso levássemos em consideração estes alunos.

Muitas das análises realizadas neste estudo serão feitas com boxplots ou diagrama de caixa. Fazendo o boxplot



por região:
O boxplot é uma caixa com limites superior e inferior representando o primeiro e terceiro quartil, respectivamente, e um traço mostrando a mediana das notas. Além disso, o gráfico apresenta duas retas que se expandem até o fim do intervalo de confiança. Quaisquer pontos que estejam fora do intervalo de confiança são "*outliers*".

O resultado é consistente: as regiões Sul e Sudeste apresentaram melhores notas que o resto do país e as regiões Norte e Nordeste apresentaram as notas mais baixas. A tabela abaixo mostra métricas das notas de redação: a média por região, os percentis 10 e 90, adiferença de nota entre esses percentis (Gap90_10) e a **EscritaDeficiente** que indica a proporção de alunos com nota de redação abaixo de 250 e acima de 0. Zerar a prova pode acontecer por questões diversas da habilidade escrita, como, por exemplo, fugir do tema. Entretanto, uma nota baixa não zerada indica dificuldade dos alunos de escreverem um texto. A tabela mostra que a região mais com as melhores notas de redação é a Sudeste, enquanto a com os piores é a Nordeste, tanto na média, quanto na nota dos percentis.

```
## # A tibble: 5 x 6
##   REGIAO_PROVA media DezPiores DezMelhores EscritaDeficiente Gap90_10
##   <fct>        <dbl>     <dbl>       <dbl>             <dbl>    <dbl>
## 1 CO            525.       360         720           0.00840      360
## 2 N             515.       360         680           0.00760      320
## 3 NE            512.       340         700           0.0113       360
## 4 S             529.       380         680           0.00344      300
## 5 SE            555.       400         740           0.00591      340
```

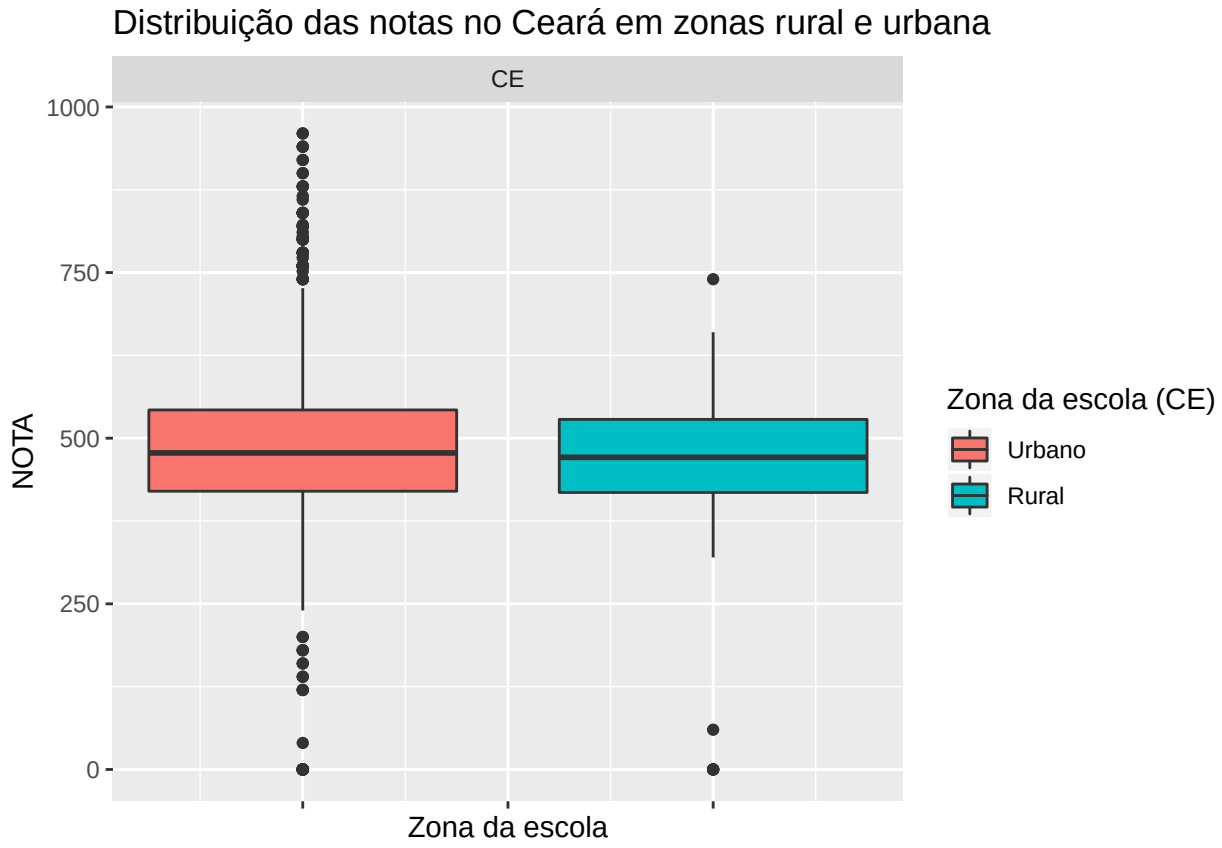
Em seguida, foi feita a mesma análise para as unidades federativas.

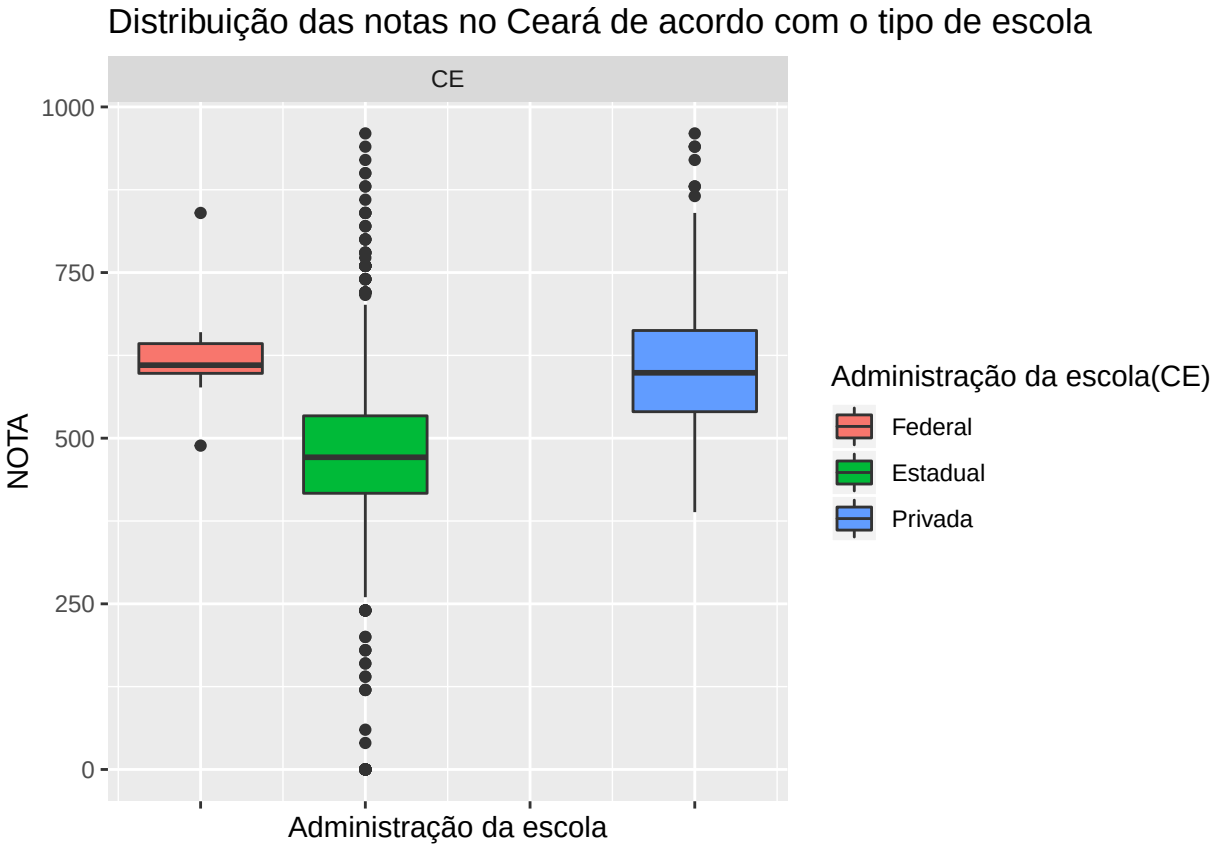

```
## # A tibble: 27 x 6
##    SG_UF_PROVA media DezPiores DezMelhores EscritaDeficiente Gap90_10
##    <chr>       <dbl>     <dbl>       <dbl>             <dbl>    <dbl>
##  1 CE           494.       300         700             0.02       400
##  2 GO           527.       360         738.            0.004      378.
##  3 RR           496.       328         704             0          376
##  4 SE           544.       376         740             0.005      364
##  5 DF           553.       400         760             0.008      360
```

```
##  6 MG           558.       400        760             0.006     360
##  7 MT           516.       320        680             0.018     360
##  8 PA           526.       360        720             0.003     360
##  9 SP           549.       380        740             0.007     360
## 10 AL           532.       380        732.            0.005     352.
## # ... with 17 more rows
```

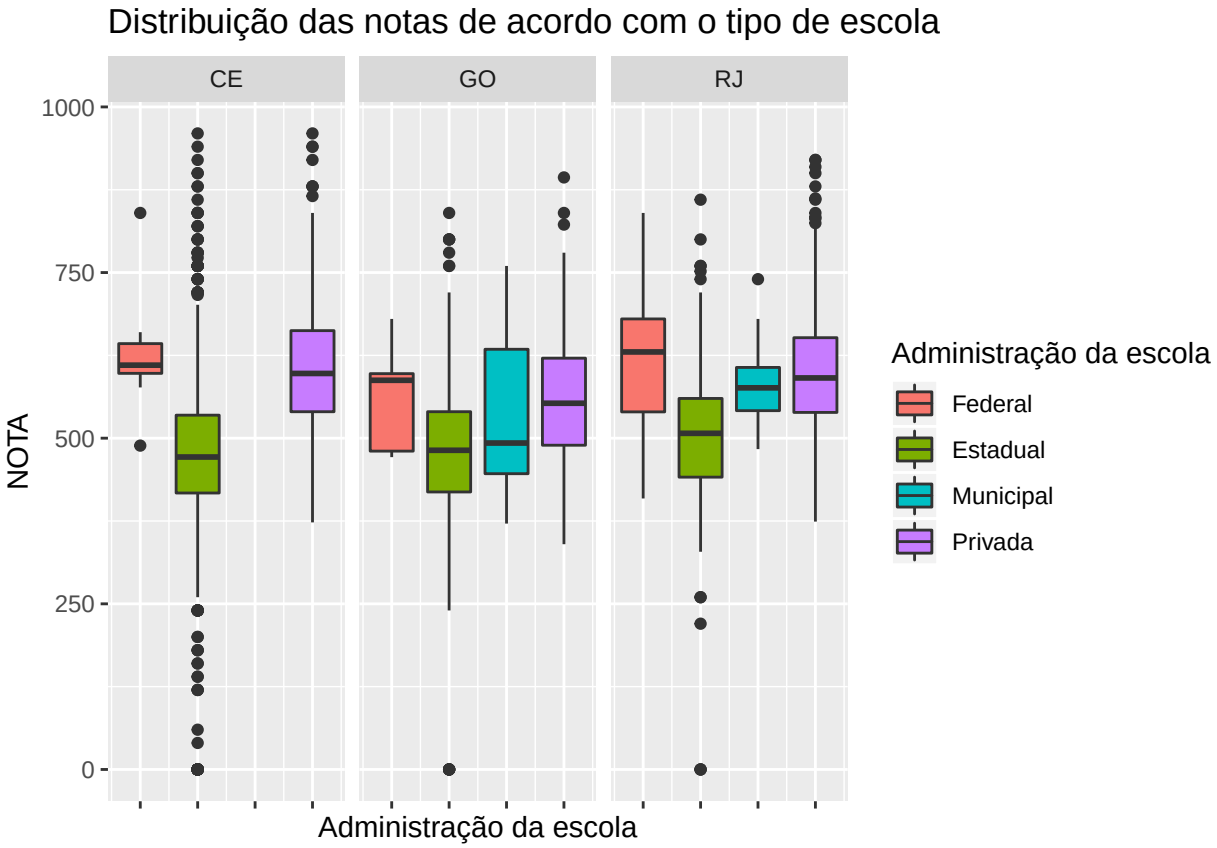
Um estado que chamou a atenção foi o Ceará. Ele possui nota média mais baixa e uma proporção preocupante elevada, entretanto, os melhores alunos alcançaram notas boas, o que pode indicar forte desigualdade social. Os estados de Goiás e Roraima mostraram uma diferença entre melhores e piores preocupante também.

Iremos nos debruçar um pouco mais sobre o Ceará, segmentando as escolas entre zonas rurais e urbanas e depois em públicas e privada.





A diferença de notas entre escolas rurais e urbanas não é tão grande. Já a diferença entre escolas estaduais versus escolas federais e privadas é grande. Todas as notas baixas do estado vieram de escolas estaduais. Claro que este padrão também aparece em outros estados, mas não é tão grave.

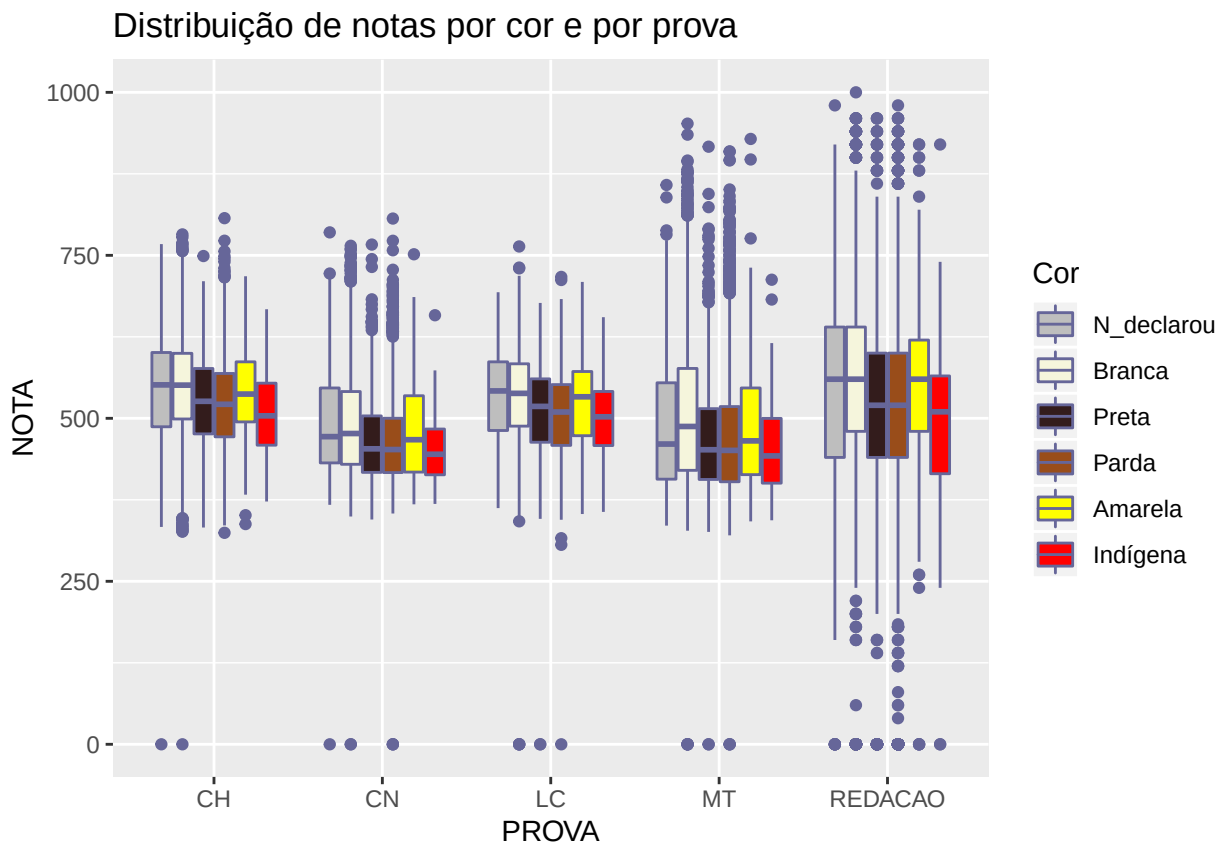


Apesar de ser normal que as escolas públicas sejam um pouco piores do que as escolas particulares, a diferença do Ceará é de 15 desvios padrões. Comparado, por exemplo, com o Goiás (2º pior estado com o maior gap entre os 10% mais e os 10% menos) temos o dobro de desvios padrões.

Deste modo, podemos afirmar que algumas escolas estaduais do Ceará precisam de uma atenção mais urgente, sendo necessário um maior nível de investimento, visando minimizar esta diferença entre escolas privadas e estaduais.

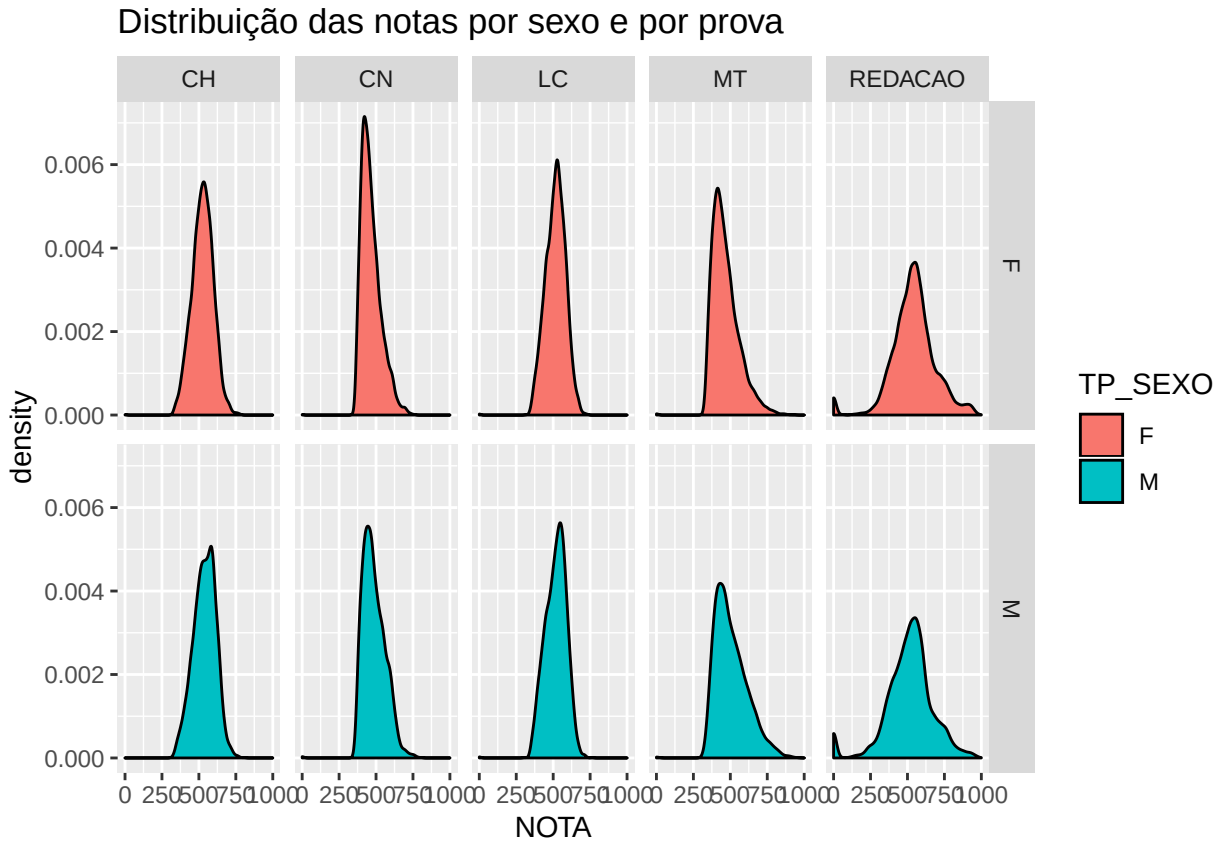
Por cor

Os dados mostraram que os candidatos de cor branca tendem a obter uma nota melhor, seguidos dos de cor amarela. Enquanto isso, os candidatos indígenas obtiveram, em média, as piores notas em todas as provas. Assim, vemos ser necessária uma maior atenção em escolas indígenas.



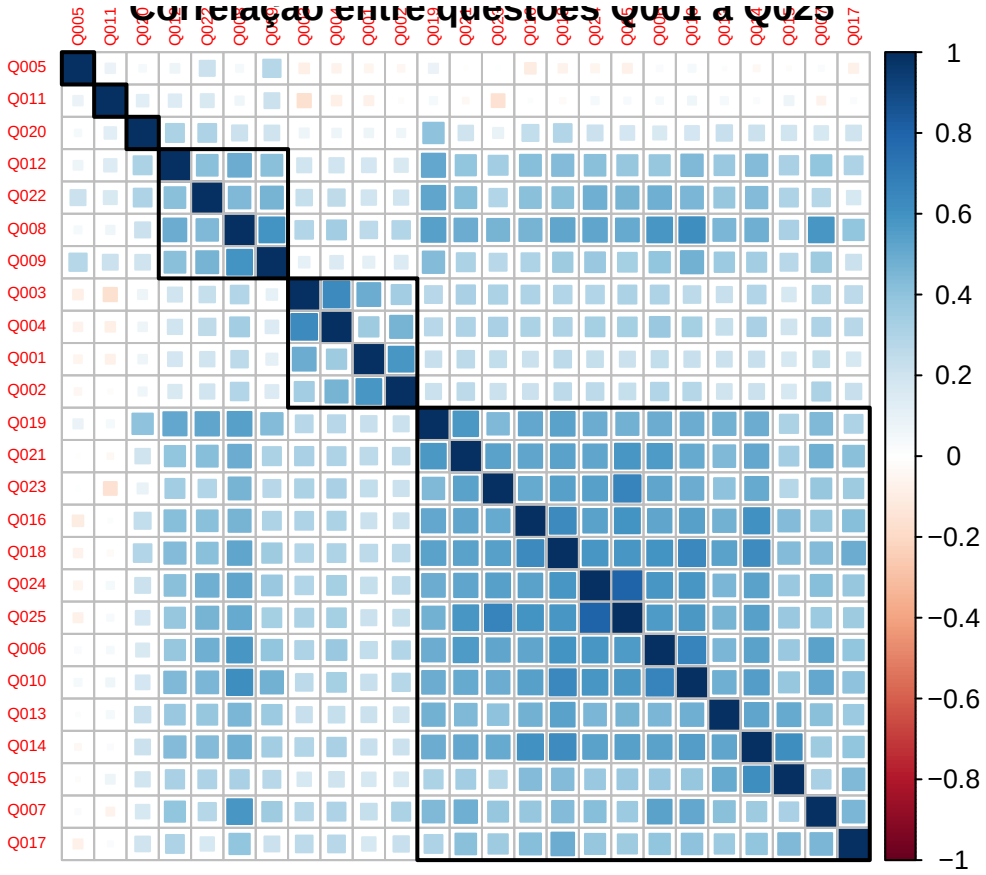
Nota por sexo

A diferença entre sexos se mostrou muito pequena. Os homens foram levemente melhores nas questões de múltipla escolha enquanto as meninas foram levemente melhores em redação. Esta diferença não é, no entanto, suficiente para propor uma mudança nos investimentos e nas políticas de ensino.



Classe social e poder aquisitivo

As perguntas das questões 001 a 025 se referem à classe social e ao poder aquisitivo dos candidatos. Antes de analisarmos as respostas às perguntas, vamos filtrar as mais relevantes e eliminar perguntas redundantes.



As 25 perguntas feitas tem respostas muito parecidas e podem ser reduzidas a 6 grupos. Continuaremos o estudo com as perguntas Q005, Q011, Q020, Q009, Q003 e Q025. Será testado agora se elas possuem impacto relevante nas notas de matemática e redação dos alunos.

```
##             P-valor Redação P-valor MT Relevante?
## (Intercept)           0.000      0.000          1
## Q005                  0.000      0.000          1
## Q011B                 0.043      0.009          1
## Q011C                 0.990      0.047          0
## Q011D                 0.478      0.696          0
## Q011E                 0.251      0.706          0
## Q020B                 0.135      0.571          0
## Q009B                 0.242      0.364          0
## Q009C                 0.756      0.131          0
## Q009D                 0.387      0.006          0
## Q009E                 0.146      0.004          0
## Q003B                 0.000      0.000          1
## Q003C                 0.000      0.000          1
## Q003D                 0.000      0.000          1
## Q003E                 0.000      0.000          1
## Q003F                 0.002      0.032          1
## Q025B                 0.000      0.000          1
```

Destas 6 perguntas escolhidas, as com impacto mais relevante sobre as notas de redação e matemática são: *Q003*: A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação do seu pai ou do homem responsável por você. *Q005*: Incluindo você, quantas pessoas moram atualmente em sua residência? *Q025*: Na sua residência tem acesso à

Internet?

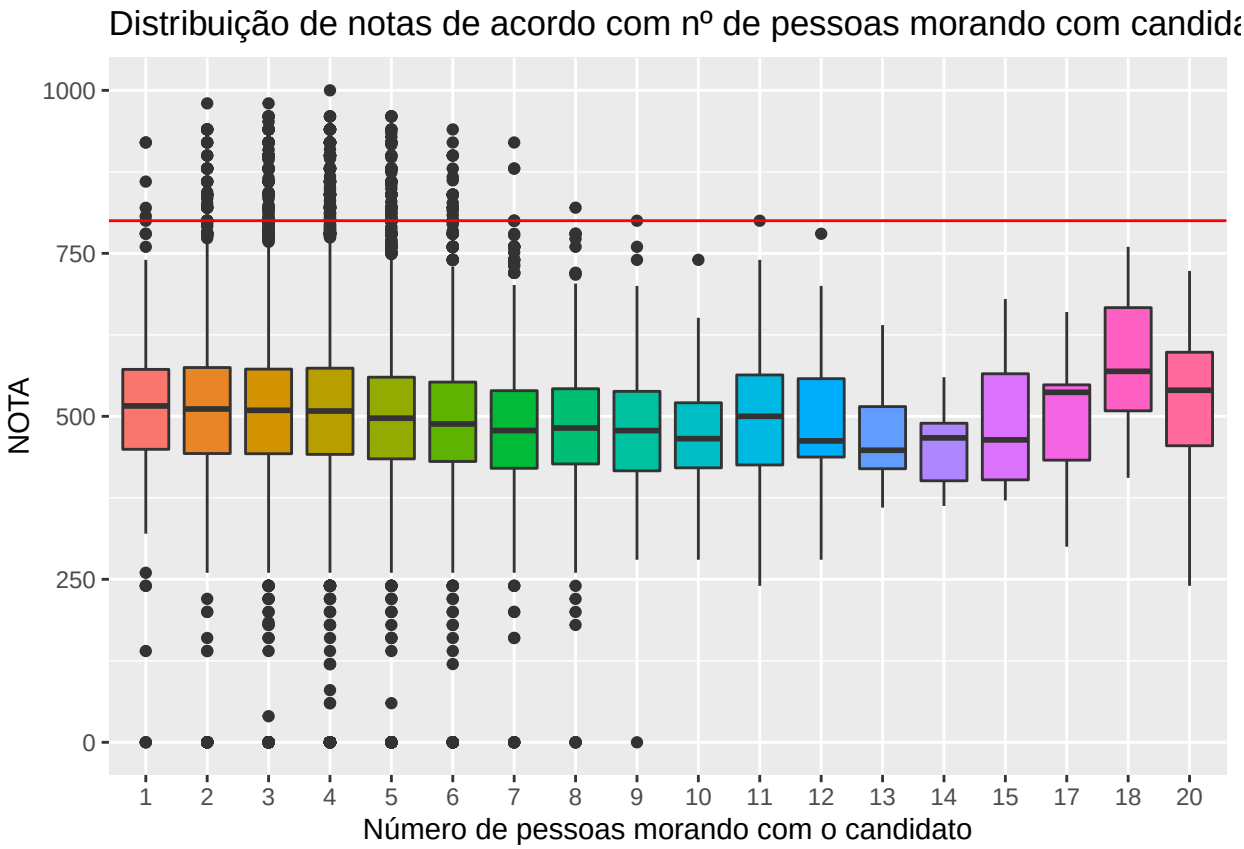
Q003: Grau de instrução/trabalho do pai

As categorias contemplam desde trabalhadores braçais (categoria A), cargos auxiliares (B), cargos técnicos(C e D), cargos que exigem nível superior (E) e desconhecido (F). É bem visível a gradação das notas de acordo com o grau de instrução paterno. A categoria F ficou com uma média de nota muito baixa, o que é um reflexo do abandono parental na educação dos filhos. Outro dado interessante foi que as maiores notas não foram alcançadas pelos filhos da categoria E, mas da categoria D.

Destes dados, abrem-se duas possibilidades de investimento em educação: Primeiro, uma política que desenvolva filhos de mãe solteira. Segundo, uma política que beneficie jovens de classe média baixa com grande potencial ao longo da faculdade, desenvolvendo-os academicamente, visando que se tornem professores universitários.

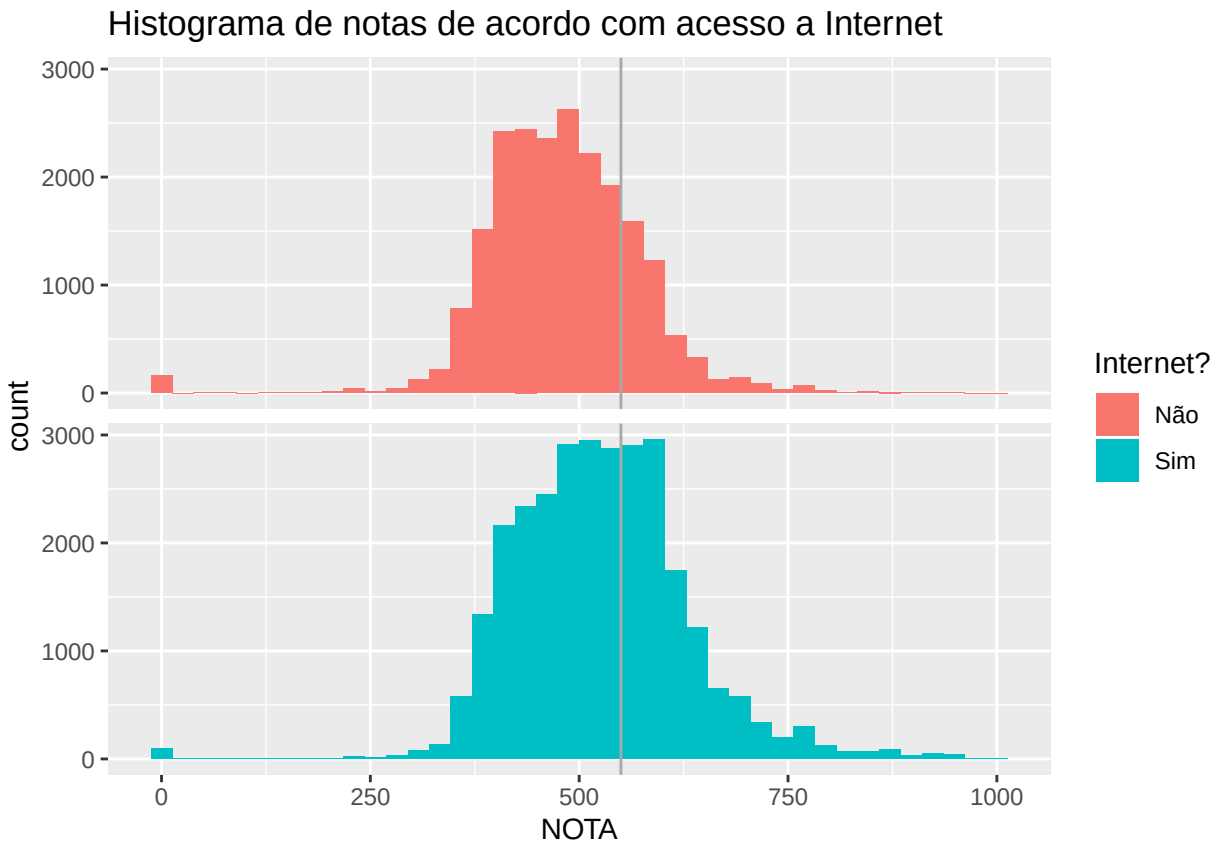
Q005: Número de pessoas na residência

O número de pessoas na casa não interfere tanto com a nota do aluno em média,entretanto, um número elevado de pessoas em casa o poderia atrapalhar a obter uma nota superior a 800.



Q025: Internet

A presença de internet aumenta a nota média consideravelmente.Segundo as regressões realizadas previamente, é uma diferença de cerca de 40 e 30 pontos para redação e matemática, respectivamente.Esta pergunta pode ser entendida como um demonstrativo de poder aquisitivo, assim como as outras perguntas estão no agrupamento da pergunta Q025.



Motivação

As perguntas de Q029 a Q41 são perguntas buscando compreender a motivação do candidato a prestar o ENEM. São perguntas que pedem uma nota de zero a cinco, sendo cinco a motivação exata. Como as motivações está arranjada em notas de 0 a 5, é possível quantificar sua importância através de uma regressão linear multivariada.

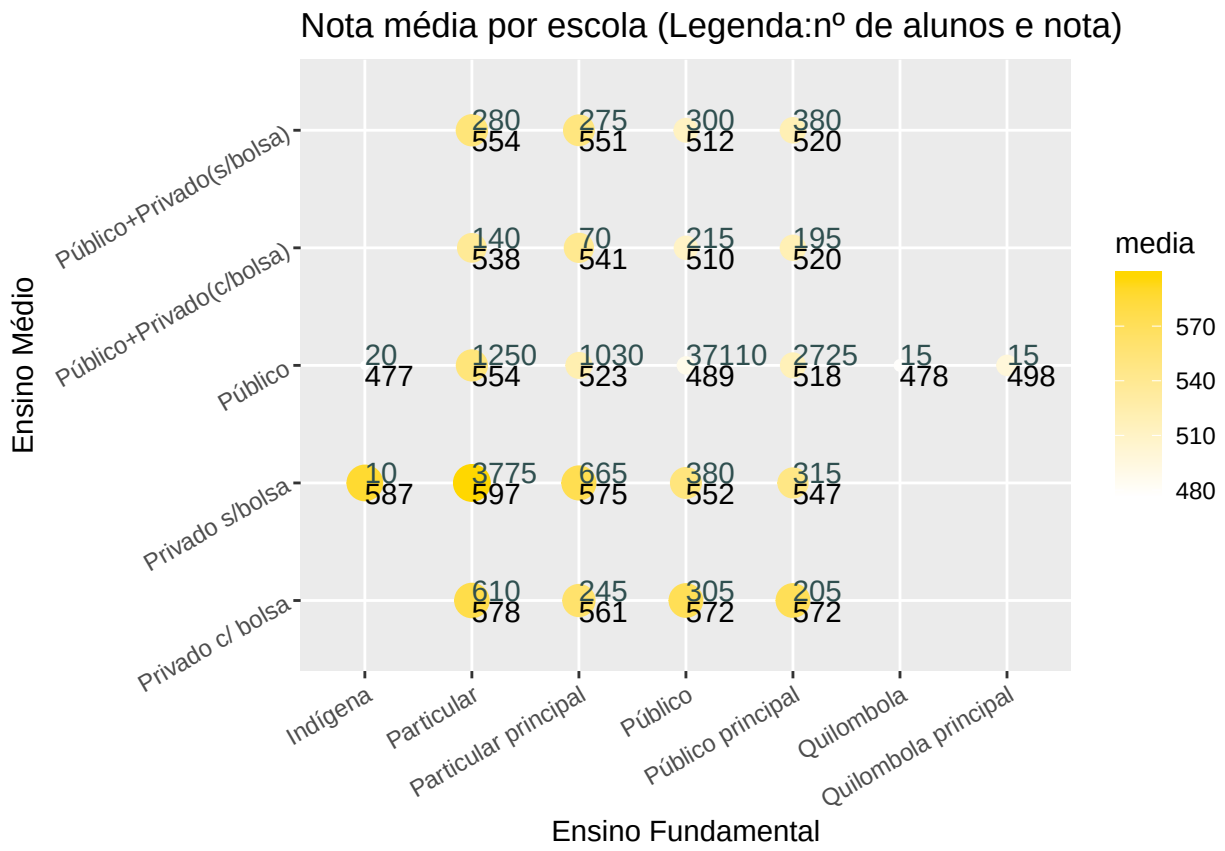
```
##
## Call:
## lm(formula = NU_NOTA_MT ~ Q029 + Q030 + Q031 + Q032 + Q033 +
##     Q034 + Q035 + Q036 + Q037 + Q038 + Q039 + Q040 + Q041, data = Q_mot)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -502.96  -65.35  -13.43   53.05  380.08
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 597.07520   16.70457  35.743  < 2e-16 ***
## Q029         -2.04899    1.33922  -1.530 0.126190
## Q030         -1.96214    0.97092  -2.021 0.043433 *
## Q031          0.42045    2.08637   0.202 0.840312
## Q032         -0.05718    2.45729  -0.023 0.981436
## Q033          0.62017    1.48120   0.419 0.675486
## Q034         -9.44794    2.39422  -3.946 8.24e-05 ***
## Q035          5.25181    2.69925   1.946 0.051848 .
```

```
## Q036          1.26390    1.34542   0.939 0.347643
## Q037         -9.64270    2.15801  -4.468 8.36e-06 ***
## Q038         -5.55152    1.44111  -3.852 0.000121 ***
## Q039          1.74076    1.34489   1.294 0.195703
## Q040         -6.10030    1.62557  -3.753 0.000180 ***
## Q041         -2.83387    1.30733  -2.168 0.030311 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 88.98 on 1850 degrees of freedom
##   (8269 observations deleted due to missingness)
## Multiple R-squared:  0.08163,    Adjusted R-squared:  0.07518
## F-statistic: 12.65 on 13 and 1850 DF,  p-value: < 2.2e-16
```

O modelo linear em si não explica suficientemente as notas, pois seu R^2 é muito baixo. Contudo, existem coeficientes lineares de motivações com alguma interferência estaticamente relevantes no desempenho do aluno.São perguntas Q030, Q034, Q037, Q038, Q040 e Q041. Uma pergunta com grande influência negativa na pontuação foi a Q037:Conseguir uma bolsa de estudos(ProUni, outras). Esta motivação pode diminuir a nota média do aluno em até 45 pontos, outro indicador do efeito da desigualdade social. As perguntas Q030 e Q38 também se referem a motivações de alunos de menor renda. Outra negativamente relevante é a pergunta Q034: Testar meus conhecimentos, que não é interessante para o objetivo do trabalho, pois indica que o candidato não está realizando a prova com menos garra. As perguntas Q040 e Q041 se referem a motivações de pessoas que já estão trabalham ou buscam emprego, o que cai no mesmo problema do questão 034.

Escola de origem

Por fim, foram segmentadas as notas médias por escola de origem de modo a olhar o ensino fundamental e o ensino médio dos candidatos.



Novamente, aparece as baixas notas dos povos indígenas. No entanto, alguns alunos terem tido um resultado claramente melhor tendo realizado o ensino médio em escolas privadas. Uma política pública para melhorar o nível dos alunos indígenas e quilombolas no ensino fundamental ou que levasse estes alunos para escolas privadas de qualidade poderia causar um impacto positivo sobre essas comunidades.