Análise Exploratória De Dados

Karen Anne Aciole Alves, 119210934
Kilian Macedo Melcher, 120110391
Vinicius Sousa Azevedo, 120110338

08/08/2022

Análise Exploratória de Dados

Quais os estados que possuem notas mais alta?

O Sistema de Avaliação da Educação Básica (Saeb) é um conjunto de avaliações externas em larga escala que permite ao Inep realizar um diagnóstico da educação básica brasileira e de fatores que podem interferir no desempenho do estudante. Desde 1995, a avaliação bienal busca fornecer um panorama da Educação Básica e sofreu algumas mudanças metodológicas para aprimoramento. Juntamente com outros indicadores, as notas do SAEB estruturam a nota do Índice de Desenvolvimento da Educação Básica (Ideb).

Problema:

1. Analisar a média de notas de cada estado do Brasil no SAEB.
2. Classificar os estados com as maiores notas.

Importando pacotes necessários:

```
library(dplyr)
library(DataExplorer)
library(ggplot2)
```

Primeiro é necessário carregar o arquivo que contém a base de dados a ser analisada.

```
dados <- read.csv("base_de_dados.csv")
```

Conhecendo nossos dados, a seguir é mostrado as 6 primeiras linhas da nossa base de dados:

```
head(dados)
```

```
##    ano id_regiao sigla_uf id_municipio area id_escola rede localizacao id_turma
## 1 2019         1       AC      6311082    2  61289499    3           2  1184224
## 2 2019         1       AC      6311082    2  61289499    3           2  1184224
## 3 2019         1       AC      6311082    2  61289499    3           2  1184224
## 4 2019         1       AC      6311082    2  61289499    3           2  1184224
## 5 2019         1       AC      6311082    2  61289499    3           2  1184224
## 6 2019         1       AC      6311082    2  61289499    3           2  1184224
```

```
##   turno serie id_aluno situacao_censo disciplina preenchimento_caderno presenca
## 1     1     2 38981514              1         LP                     1        1
## 2     1     2 38981515              1         LP                     1        1
## 3     1     2 38981516              1         LP                     1        1
## 4     1     2 38981517              1         LP                     1        1
## 5     1     2 38981518              1         LP                     0        0
## 6     1     2 38981519              1         LP                     1        1
##   caderno bloco_1 bloco_2 bloco_1_aberto bloco_2_aberto respostas_bloco_1
## 1      10       3       5              3              5                NA
## 2      10       3       5              3              5                NA
## 3      10       3       5              3              5                NA
## 4      10       3       5              3              5                NA
## 5      10       3       5              3              5                NA
## 6      10       3       5              3              5                NA
##   respostas_bloco_2 conceito_q1 conceito_q2 resposta_texto conceito_proposito
## 1                NA           B           B             TX                  A
## 2                NA           B           B             TX                  A
## 3                NA           B           B             TX                  A
## 4                NA           A           A             TX                  A
## 5                NA           D           D             BR                  .
## 6                NA           B           B             TX                  A
##   conceito_elemento conceito_segmentacao texto_grafia indicador_proficiencia
## 1                 B                    A            A                     1
## 2                 B                    B            A                     1
## 3                 B                    B            A                     1
## 4                 A                    A            A                     1
## 5                 .                    .            .                     0
## 6                 A                    B            A                     1
##   amostra estrato peso_aluno proficiencia erro_padrao proficiencia_saeb
## 1       1   12322   6.391051    -0.460629    0.314986          722.7011
## 2       1   12322   6.391051    -0.134725    0.341066          740.7582
## 3       1   12322   6.391051    -0.790666    0.290154          704.4150
## 4       1   12322   6.391051     1.063636    0.503050          807.1550
## 5       1   12322         NA           NA          NA                NA
## 6       1   12322   6.391051    -0.712789    0.282118          708.7298
##   erro_padrao_saeb
## 1         17.45222
## 2         18.89721
## 3         16.07637
## 4         27.87215
## 5               NA
## 6         15.63112
```

Temos 38 variáveis, mas para essa análise iremos focar apenas na nota de cada estudante que respondeu o questionário e seu respectivo estado. Dessa forma, iremos selecionar apenas as variáveis na qual estamos interessados.

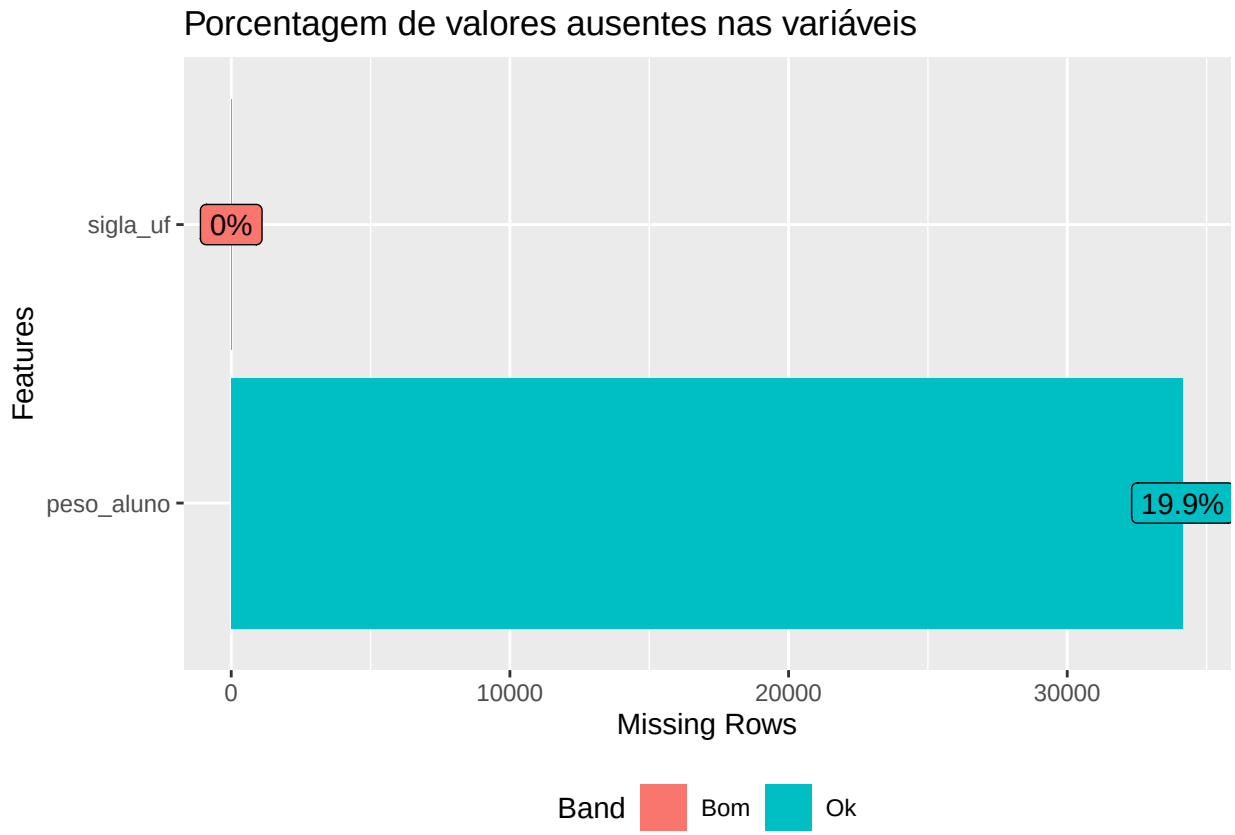
```
dados_filtrados <- dados %>%
  select(sigla_uf, peso_aluno)

knitr::kable(head(dados_filtrados), caption="Primeiras linhas da base de dados")
```

Table 1: Primeiras linhas da base de dados

sigla_uf	peso_aluno
AC	6.391051
AC	6.391051
AC	6.391051
AC	6.391051
AC	NA
AC	6.391051

O próximo passo é procurar por valores ausentes (NAs, zeros, etc) com o pacote "DataExplorer".

```
# O parâmetro group da função categoriza a variável de acordo com os limites
# superiores estipulados (valores padrões da biblioteca) foi alterado apenas os
# nomes para o português.

plot_missing(dados_filtrados,
             title="Porcentagem de valores ausentes nas variáveis",
             group = list(Bom = 0.05, Ok = 0.4, Ruim = 0.8, Remover = 1))
```



O gráfico demonstra que temos apenas 20% de valores ausentes na variável "peso_aluno", além disso, é considerada como uma margem "Ok" de valores ausentes pois é um valor pequeno de dados ausentes comparado ao tamanho da nossa base de dados.

Limpeza de dados

Retirando as linhas com valores ausentes:

```
nrow(dados_filtrados) # Antes de retirar os dados ausentes
```

```
## [1] 171576
```

```
dados_validos <- dados_filtrados %>% filter(peso_aluno != "none" & between(peso_aluno, 0, 500))
nrow(dados_validos) # Depois de retirar os dados ausentes
```

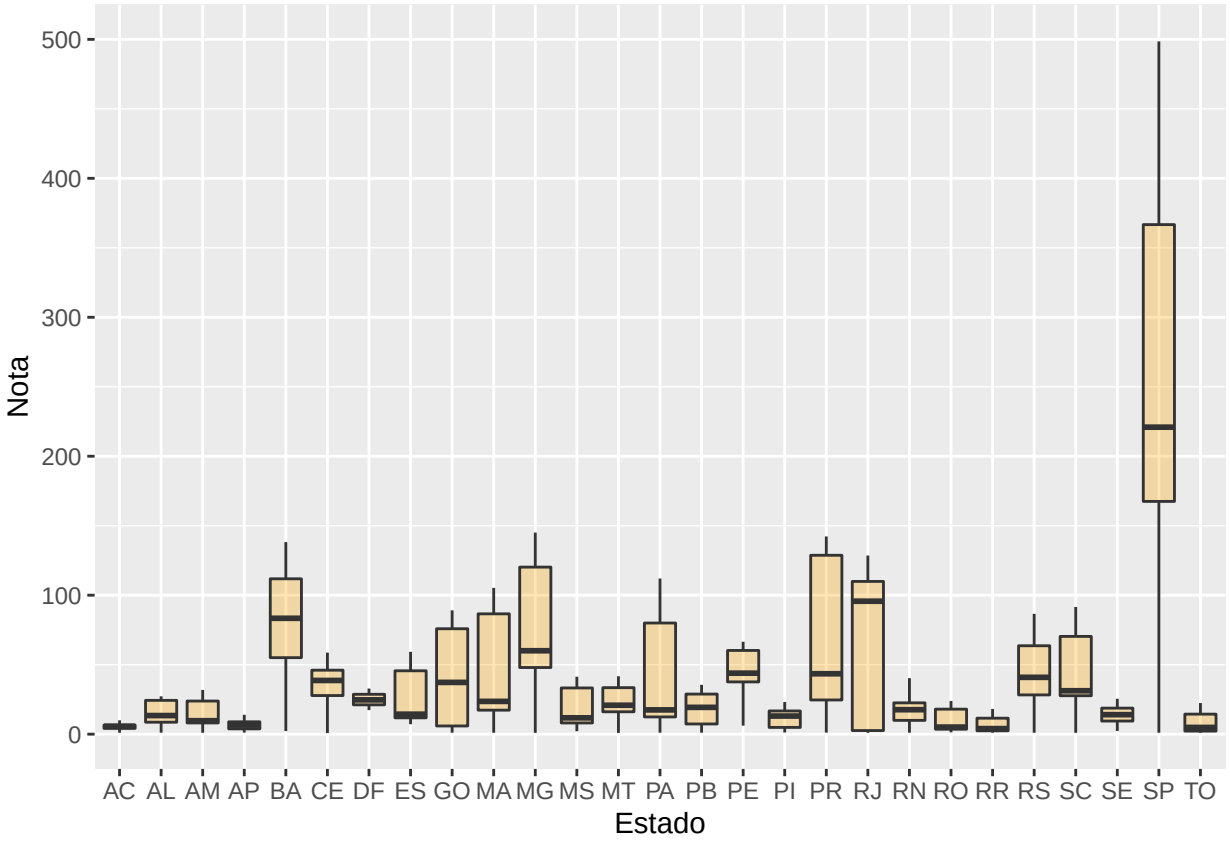
```
## [1] 137254
```

Foi feito um novo arquivo .csv com apenas as variáveis que serão analisadas:

```
#write.csv(dados_validos, file = "amostra_utilizada.csv", row.names=FALSE)
data <- read.csv("amostra_utilizada.csv")
```

Box-plot para visualizar a relação das notas entre os estados:

```
ggplot(as.data.frame(data), aes(x = sigla_uf, y = peso_aluno, fill=peso_aluno)) +
  geom_boxplot(outlier.shape=NA, fill="orange", alpha=0.3) +
  coord_cartesian(ylim =  c(0, 500)) +
  labs(x="Estado", y = "Nota")
```



De imediato vemos que o estado com maior nota é o estado de São Paulo.

Extraindo a média de notas de cada estado:

```
media_por_estado <- data %>%
  group_by(sigla_uf) %>%
  summarise_at(vars(peso_aluno),
               list(media_notas=mean))

knitr::kable(media_por_estado, digits=2, caption="Média de notas por estado")
```

Table 2: Média de notas por estado

sigla_uf	media_notas
AC	5.32
AL	15.12
AM	14.70
AP	6.65
BA	79.87
CE	35.08
DF	24.00
ES	26.32
GO	42.65
MA	43.64
MG	78.27
MS	19.68
MT	22.40
PA	39.13
PB	18.58
PE	44.90
PI	11.05
PR	73.85
RJ	73.19
RN	17.12
RO	9.57
RR	6.59
RS	44.01
SC	40.56
SE	13.98
SP	254.70
TO	8.13

Extraindo a quantidade de estudantes que participaram da avaliação de cada estado:

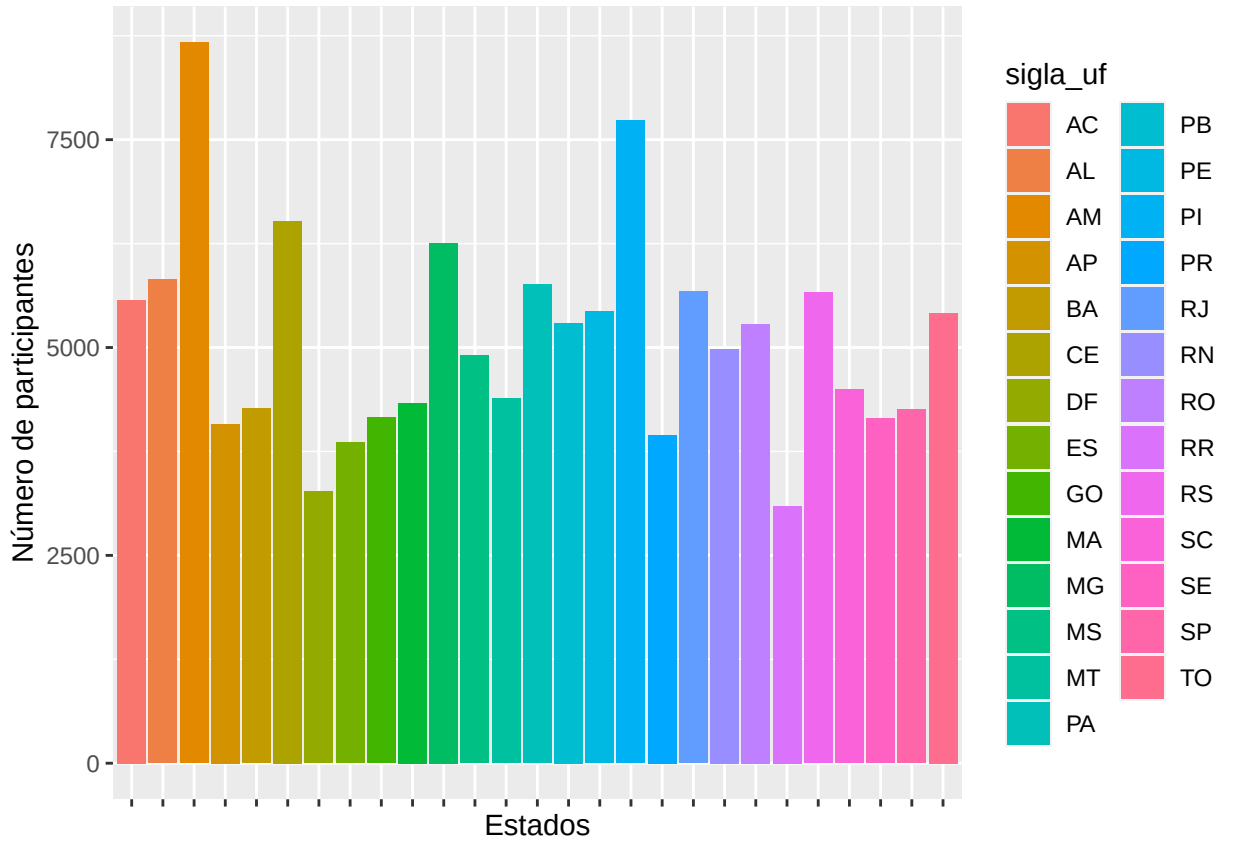
```
participacao_por_estado <- data %>%
  count(sigla_uf)

knitr::kable(participacao_por_estado, caption="Participação de estudantes por estado")
```

Table 3: Participação de estudantes por estado

sigla_uf	n
AC	5574
AL	5820
AM	8667
AP	4078
BA	4267
CE	6521
DF	3271
ES	3856
GO	4158
MA	4330
MG	6255
MS	4907
MT	4388
PA	5756
PB	5293
PE	5430
PI	7728
PR	3946
RJ	5676
RN	4981
RO	5282
RR	3092
RS	5664
SC	4492
SE	4149
SP	4257
TO	5416

Gráfico de barra para visualizar o número de estudantes que participaram referente a cada estado:

```
ggplot(as.data.frame(participacao_por_estado), aes(x=sigla_uf, y=n, fill = sigla_uf))+
  geom_col(position="dodge")+
  labs(x="Estados", y="Número de participantes")+
  theme(axis.text.x = element_blank())
```



Por outro lado, temos que, o estado do Amazonas teve a maior quantidade de estudantes participantes.

Analisando os 6 estados com as maiores notas:

```
maiores_medias <- media_por_estado[order(media_por_estado$media_notas, decreasing=TRUE),]
knitr::kable(head(maiores_medias), digits=2, caption="Estados com as maiores notas")
```

Table 4: Estados com as maiores notas

sigla_uf	media_notas
SP	254.70
BA	79.87
MG	78.27
PR	73.85
RJ	73.19
PE	44.90

Analisando os 6 estados com as menores notas:

```
menores_medias <- media_por_estado[order(media_por_estado$media_notas, decreasing=TRUE),]
knitr::kable(tail(menores_medias), digits=2, caption="Estados com as menores notas")
```

Table 5: Estados com as menores notas

sigla_uf	media_notas
PI	11.05
RO	9.57
TO	8.13
AP	6.65
RR	6.59
AC	5.32

Analisando os 6 estados com mais participantes:

```
maiores_participantes <- participacao_por_estado[order(participacao_por_estado$n, decreasing=TRUE),]
knitr::kable(head(maiores_participantes), caption="Estados com maior número de participantes")
```

Table 6: Estados com maior número de participantes

	sigla_uf	n
3	AM	8667
17	PI	7728
6	CE	6521
11	MG	6255
2	AL	5820
14	PA	5756

Analisando os 6 estados com menores participantes

```
menores_participantes <- participacao_por_estado[order(participacao_por_estado$n, decreasing=TRUE),]
knitr::kable(tail(maiores_participantes), caption="Estados com menor número de participantes")
```

Table 7: Estados com menor número de participantes

	sigla_uf	n
25	SE	4149
4	AP	4078
18	PR	3946
8	ES	3856
7	DF	3271
22	RR	3092

Conclusões:

Por meio da análise, foi possível identificar que temos uma grande disparidade entre a média de notas do estado de São Paulo em comparação aos demais estados. Dentre as maiores notas, temos São Paulo, Rio de Janeiro e Minas Gerais da região Sudeste; os estados da Bahia e Pernambuco da região Nordeste; e o estado do Paraná da região Sul. Curiosamente, os estados com maiores participantes são das regiões Norte e Nordeste, e com apenas Minas Gerais da região Sudeste. Por fim, os estados com as menores notas são da região Norte e Nordeste.

É possível estender a discussão dessa análise para o âmbito da desigualdade social do país relacionando com a qualidade do ensino nas escolas públicas de cada estado, sabendo que essa qualidade varia de acordo com as condições socioeconômicas de cada estado.

Análise de teste de hipótese

Nesse dado, a variável "**Peso aluno**" é uma nota que o aluno recebe que pode variar de 0 a 500. Assumindo que o peso médio de um aluno é de 250, vamos verificar se a proporção dos alunos que obtiveram peso maior ou igual ao peso médio é de pelo menos 50% nessa amostra.

h0: p = 0.5 (Hipótese nula).

h1: p < 0.5 (Hipótese alternativa).

```
library(dplyr)

# Definindo os parâmetros do teste
peso_medio <- 250
probabilidade_de_sucesso <- 0.5

# Lista contendo 1 para peso maior ou igual ao peso médio e 0 para menor.
verifica_peso_medio <- as.numeric(data$peso_aluno >= peso_medio)

tabela_peso_medio <- table(verifica_peso_medio)
tamanho_da_amostra <- dim(data)[1]
total_sucessos <- tabela_peso_medio[2]

# Calculando o teste de hipótese unilateral à esquerda
prop.test(
  x = total_sucessos,
  n = tamanho_da_amostra,
  p = probabilidade_de_sucesso,
  alternative = 'less',
  correct = FALSE
)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  total_sucessos out of tamanho_da_amostra, null probability probabilidade_de_sucesso
## X-squared = 130659, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is less than 0.5
## 95 percent confidence interval:
##  0.00000000 0.01265625
## sample estimates:
##          p
## 0.01215994
```

Resultados:

- Intervalo de confiança de 95%.
- 1,26% dos alunos nessa amostra possuem um peso igual ou maior ao peso médio.
- A proporção de crianças com peso maior ou igual ao peso médio é menor que 50%.
- Valor p = 0.01215994 = 1,2% (< 5%)
- Rejeita a hipótese nula.