

Análise Exploratória SAEB 2019

Tailine J. S. Nonato

2024-06-24

Análise Exploratória SAEB 2019

Carregamento de pacotes e dados

```
if (!require(pacman)) install.packages("pacman")

Loading required package: pacman

pacman::p_load(vroom,tidyverse, knitr, openxlsx, kableExtra, lme4, gridExtra)
setwd("C:/Users/User/Documents/GitHub/gradest-1/TCC/rel parcial")
```

Filtragem dos dados para região Centro-Oeste

```
df_aluno <- read.csv("TS_ALUNO_9EF_2019.csv", sep = ";", encoding = "latin1")
df_escola <- read.csv("TS_ESCOLA_2019.csv", sep = ";", encoding = "latin1")
unique(df_escola$NIVEL_SOCIO_ECONOMICO)

var_al <- names(df_aluno)
var_es <- names(df_escola)

df_aluno <- df_aluno %>% filter(ID_REGIAO == 5)
df_escola <- df_escola %>% filter(ID_REGIAO == 5)

write.csv(df_aluno, file = "TS_ALUNO_9EF_2019_co.csv")
write.csv(df_escola, file = "TS_ESCOLA_2019_co.csv")
```

```
df_aluno <- read.csv("TS_ALUNO_9EF_2019_co.csv", encoding="latin1")
df_escola <- read.csv("TS_ESCOLA_2019_co.csv", encoding="latin1")

df <- left_join(df_aluno, df_escola, by = c("ID_ESCOLA", "ID_REGIAO", "ID_SAEB", "ID_UF", "ID_MUNICIPIO"))
df <- df %>% mutate(ID_ESCOLA = as.factor(ID_ESCOLA))

write.csv(df, "dados_tcc.csv")
```

Leitura dos dados

```
df <- read.csv("dados_tcc.csv", encoding = "latin1")

var_df <- names(df)
dim(df)

[1] 203694      261
```

Análise descritiva dos dados

Quantidade de Escolas

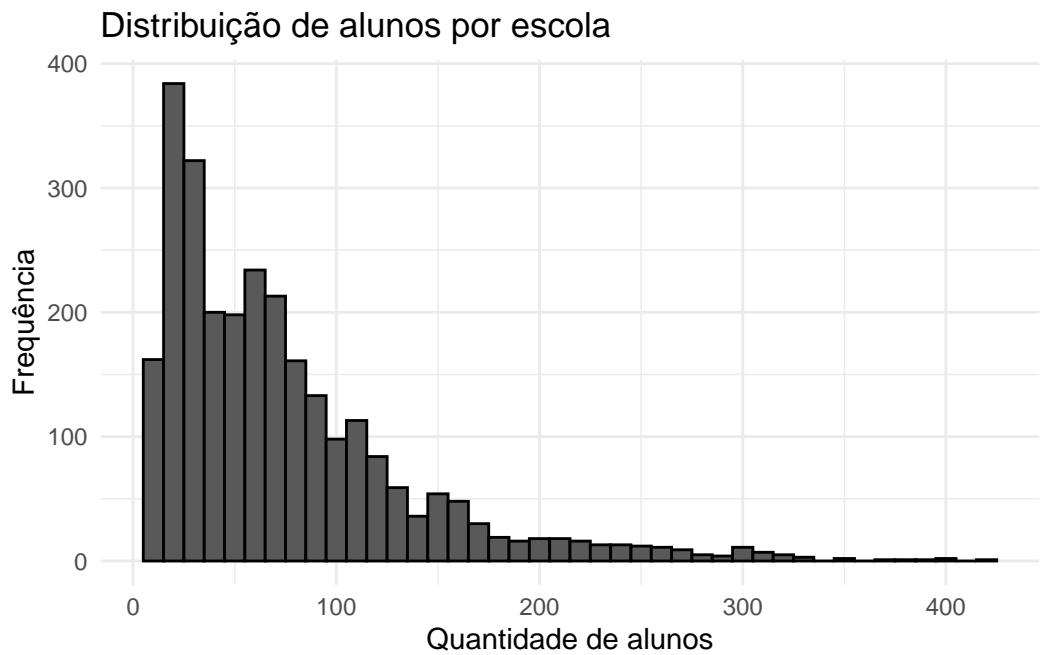
```
df %>%
  summarise(n = n_distinct(ID_ESCOLA))

n
1 2717
```

Quantidade de Alunos por Escola

```
df %>%
  group_by(ID_ESCOLA) %>%
  summarise(n = n()) %>%
  ggplot(aes(x = n)) +
  geom_histogram(binwidth = 10, color = "black") +
  labs(title = "Distribuição de alunos por escola",
       x = "Quantidade de alunos",
       y = "Frequência") +
```

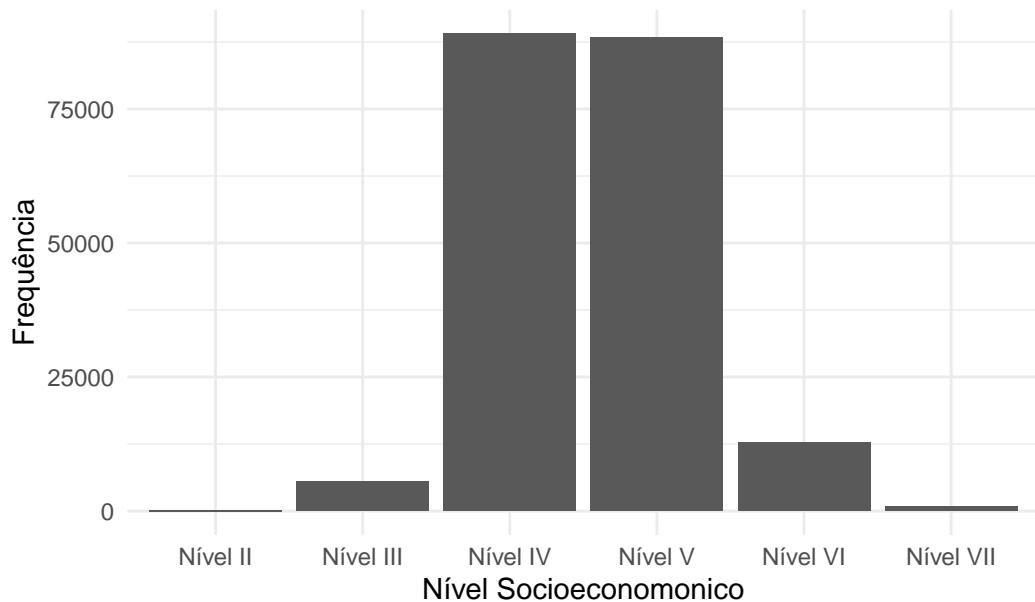
`theme_minimal()`



Nível Socioeconômico (NSe) das Escolas

```
#unique(df$NIVEL_SOCIO_ECONOMICO)
df %>%
  mutate(NIVEL_SOCIO_ECONOMICO = factor(NIVEL_SOCIO_ECONOMICO, levels = c("Nível I", "Nível II", "Nível III", "Nível IV", "Nível V")),
  filter(!is.na(NIVEL_SOCIO_ECONOMICO)) %>%
  ggplot(aes(x = NIVEL_SOCIO_ECONOMICO)) +
  geom_bar() +
  labs(title = "Distribuição de escolas por NSe",
       x = "Nível Socioecononomico",
       y = "Frequência") +
  theme_minimal()
```

Distribuição de escolas por NSe



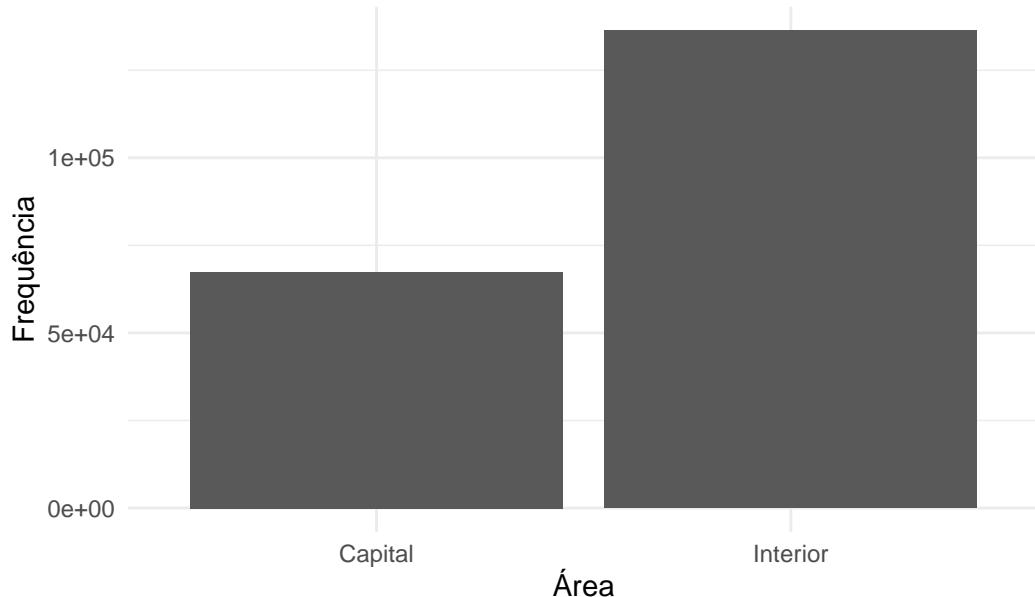
Área das Escolas

```
unique(df$ID_AREA)
```

```
[1] 2 1
```

```
df %>%
  mutate(ID_AREA = ifelse(ID_AREA==1,"Capital","Interior")) %>%
  ggplot(aes(x = ID_AREA)) +
  geom_bar() +
  labs(title = "Distribuição de escolas por área",
       x = "Área",
       y = "Frequência") +
  theme_minimal()
```

Distribuição de escolas por área

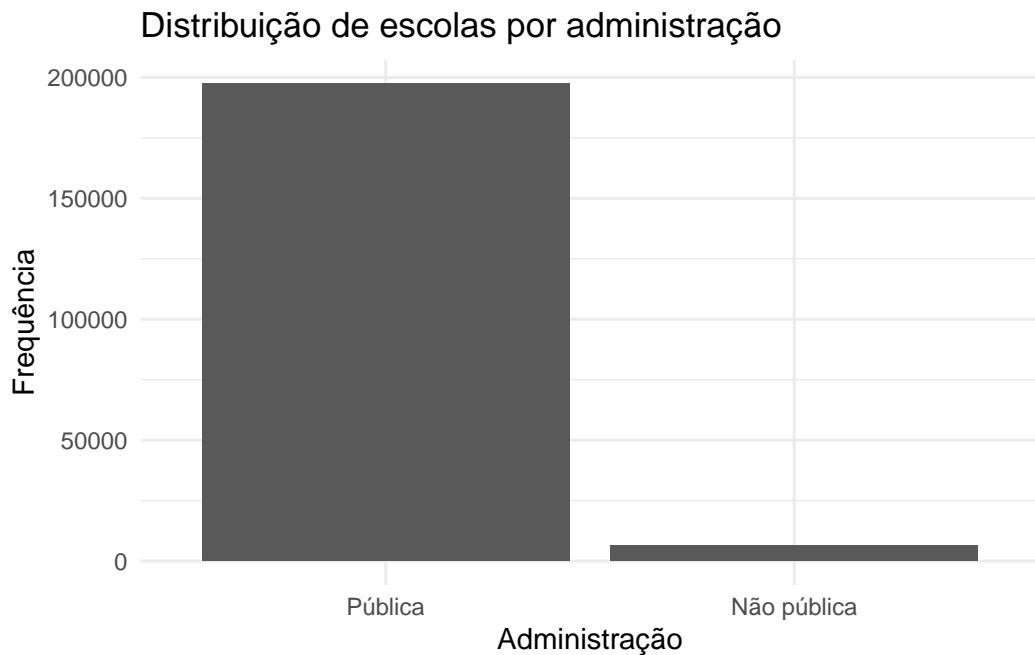


Administração das Escolas (Pública)

```
unique(df$IN_PUBLICA)
```

```
[1] 1 0
```

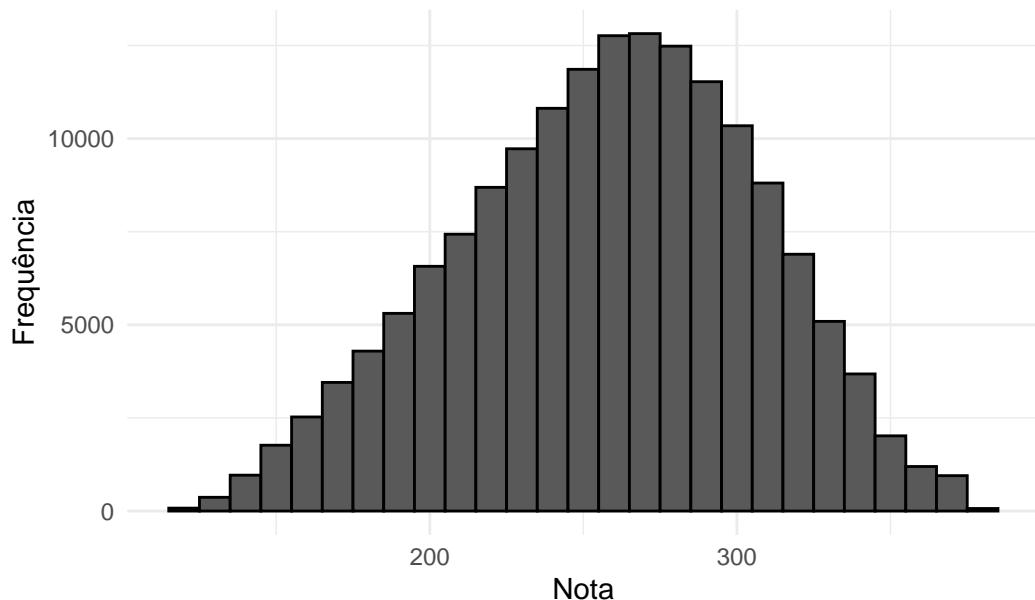
```
df %>%
  mutate(IN_PUBLICA = ifelse(IN_PUBLICA==1,"Pública","Não pública")) %>%
  mutate(IN_PUBLICA = factor(IN_PUBLICA, levels = c("Pública", "Não pública"))) %>%
  ggplot(aes(x = IN_PUBLICA)) +
  geom_bar() +
  labs(title = "Distribuição de escolas por administração",
       x = "Administração",
       y = "Frequência") +
  theme_minimal()
```



Distribuição da proficiência em Língua Portuguesa e Matemática

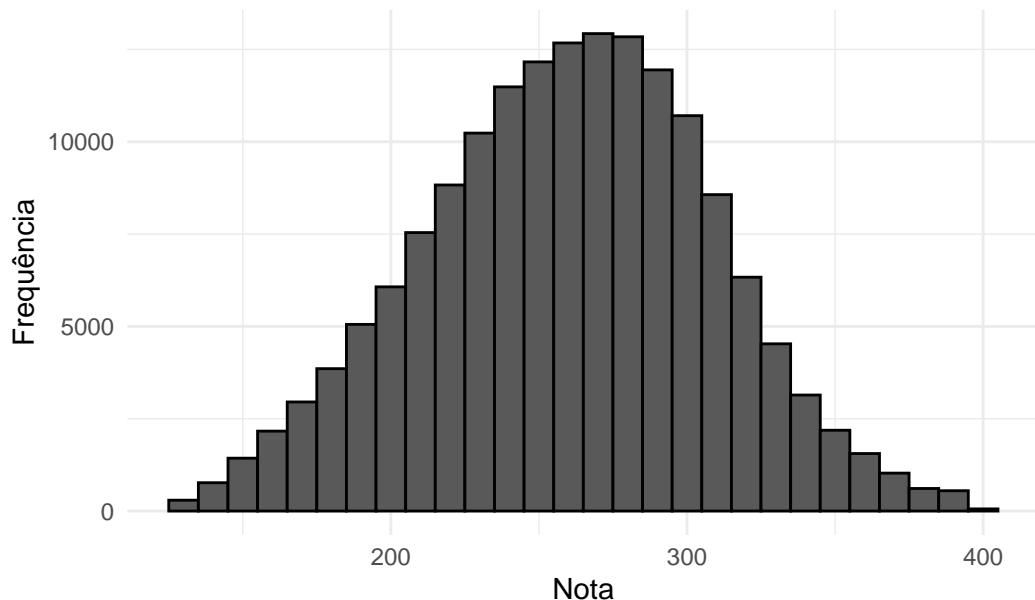
```
#lingua portuguesa (ggplot histogram)
df %>%
  ggplot(aes(x = PROFICIENCIA_LP_SAEB)) +
  geom_histogram(binwidth = 10, color = "black") +
  labs(title = "Distribuição de notas de LP",
       x = "Nota",
       y = "Frequência") +
  theme_minimal()
```

Distribuição de notas de LP



```
#matemática (ggplot histogram)
df %>%
  ggplot(aes(x = PROFICIENCIA_MT_SAEB)) +
  geom_histogram(binwidth = 10, color = "black") +
  labs(title = "Distribuição de notas de MT",
       x = "Nota",
       y = "Frequência") +
  theme_minimal()
```

Distribuição de notas de MT



Teste de normalidade

```
set.seed(123) # for reproducibility
sample_LP <- sample(df$PROFICIENCIA_LP_SAEB, 5000)
shapiro.test(sample_LP)
```

```
Shapiro-Wilk normality test

data: sample_LP
W = 0.99082, p-value = 2.602e-15

set.seed(123) # for reproducibility
sample_MT <- sample(df$PROFICIENCIA_MT_SAEB, 5000)
shapiro.test(sample_MT)
```



```
Shapiro-Wilk normality test

data: sample_MT
W = 0.99768, p-value = 1.105e-05
```

Correlação entre as variáveis

```
#cor_df <- df %>% cor()
```

Modelagem Multinível

Modelo nulo

```
#modelo_nulo
modelo_nulo <- lmer(PROFICIENCIA_LP_SAEB ~ 1 + (1|ID_ESCOLA), data = df)
summary(modelo_nulo)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: PROFICIENCIA_LP_SAEB ~ 1 + (1 | ID_ESCOLA)
Data: df

REML criterion at convergence: 1707692

Scaled residuals:
    Min      1Q  Median      3Q     Max 
-3.5874 -0.6716  0.0499  0.7084  3.1703 

Random effects:
 Groups   Name        Variance Std.Dev. 
ID_ESCOLA (Intercept) 348.9   18.68  
Residual           2070.0   45.50  
Number of obs: 162498, groups: ID_ESCOLA, 2681

Fixed effects:
            Estimate Std. Error t value
(Intercept) 256.29     0.39   657.2
```

Modelo com variáveis preditoras

```
#modelo_com_variaveis_preditoras
modelo <- lmer(PROFICIENCIA_LP_SAEB ~ NIVEL_SOCIO_ECONOMICO + ID_AREA + IN_PUBLICA + (1|ID_ESCOLA))
```

fixed-effect model matrix is rank deficient so dropping 1 column / coefficient

```
summary(modelo)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: PROFICIENCIA_LP_SAEB ~ NIVEL_SOCIO_ECONOMICO + ID_AREA + IN_PUBLICA +
          (1 | ID_ESCOLA)
Data: df
```

REML criterion at convergence: 1659140

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.6238	-0.6752	0.0486	0.7099	3.1280

Random effects:

Groups	Name	Variance	Std.Dev.
ID_ESCOLA	(Intercept)	189.3	13.76
	Residual	2076.5	45.57
Number of obs:	157964, groups:	ID_ESCOLA,	2530

Fixed effects:

	Estimate	Std. Error	t value	
(Intercept)	225.1612	4.1557	54.181	
NIVEL_SOCIO_ECONOMICO	nível II	-35.4749	8.8661	-4.001
NIVEL_SOCIO_ECONOMICO	nível III	-3.7496	4.0739	-0.920
NIVEL_SOCIO_ECONOMICO	nível IV	12.5577	3.8623	3.251
NIVEL_SOCIO_ECONOMICO	nível V	25.1541	3.8791	6.484
NIVEL_SOCIO_ECONOMICO	nível VI	44.0952	4.1463	10.635
NIVEL_SOCIO_ECONOMICO	nível VII	76.6305	8.1210	9.436
ID_AREA		6.4522	0.7984	8.081

```
Warning in abbreviate(rn, minlength = 11): abbreviate used with non-ASCII chars
Warning in abbreviate(rn, minlength = 11): abbreviate used with non-ASCII chars
Warning in abbreviate(rn, minlength = 11): abbreviate used with non-ASCII chars
Warning in abbreviate(rn, minlength = 11): abbreviate used with non-ASCII chars
Warning in abbreviate(rn, minlength = 11): abbreviate used with non-ASCII chars
Warning in abbreviate(rn, minlength = 11): abbreviate used with non-ASCII chars
Warning in abbreviate(rn, minlength = 11): abbreviate used with non-ASCII chars
Warning in abbreviate(rn, minlength = 11): abbreviate used with non-ASCII chars
Warning in abbreviate(rn, minlength = 11): abbreviate used with non-ASCII chars
```

Correlation of Fixed Effects:

Warning in abbreviate(rn, minlength = 6): abbreviate used with non-ASCII chars

	(Intr)		NIVEL_SOCIO_ECONOMICONvII
NIVEL_SOCIO_ECONOMICONvII	-0.400		
NIVEL_SOCIO_ECONOMICONIII	-0.869	0.408	
NIVEL_SOCIO_ECONOMICONIV	-0.926	0.430	
NIVEL_SOCIO_ECONOMICONvV	-0.941	0.428	
NIVEL_SOCIO_ECONOMICONvVI	-0.895	0.400	
NIVEL_SOCIO_ECONOMICONvII	-0.474	0.204	
ID_AREA	-0.384	0.000	
			NIVEL_SOCIO_ECONOMICONIII NIVEL_SOCIO_ECONOMICONIV

```

NIVEL_SOCIO_ECONOMICONvII
NIVEL_SOCIO_ECONOMICONIII
NIVEL_SOCIO_ECONOMICONIV  0.935
NIVEL_SOCIO_ECONOMICONvV  0.931          0.984
NIVEL_SOCIO_ECONOMICONvVI 0.871          0.922
NIVEL_SOCIO_ECONOMICONvII 0.445          0.472
ID_AREA                   0.000          0.024
                                NIVEL_SOCIO_ECONOMICONvV NIVEL_SOCIO_ECONOMICONvVI
NIVEL_SOCIO_ECONOMICONvII
NIVEL_SOCIO_ECONOMICONIII
NIVEL_SOCIO_ECONOMICONIV
NIVEL_SOCIO_ECONOMICONvV
NIVEL_SOCIO_ECONOMICONvVI  0.923
NIVEL_SOCIO_ECONOMICONvII 0.474          0.448
ID_AREA                   0.073          0.105
                                NIVEL_SOCIO_ECONOMICONvII
NIVEL_SOCIO_ECONOMICONvII
NIVEL_SOCIO_ECONOMICONIII
NIVEL_SOCIO_ECONOMICONIV
NIVEL_SOCIO_ECONOMICONvV
NIVEL_SOCIO_ECONOMICONvVI
NIVEL_SOCIO_ECONOMICONvII
ID_AREA                   0.098
fit warnings:
fixed-effect model matrix is rank deficient so dropping 1 column / coefficient

```

Análise de resíduos

```

#resíduos
resíduos <- residuals(modelo)
qqnorm(resíduos)
qqline(resíduos)

```

