

# Análise Exploratória SAEB 2019

Tailine J. S. Nonato

2024-06-29

## Análise Exploratória SAEB 2019

### Carregamento de pacotes e dados

```
if (!require(pacman)) install.packages("pacman")
pacman::p_load(vroom, tidyverse, knitr, openxlsx, kableExtra, lme4, gridExtra, clipr)
options(OutDec = ",")
setwd("C:/Users/User/Documents/GitHub/gradest-1/TCC/rel parcial")
```

### Filtragem dos dados para região Centro-Oeste

```
df_aluno <- vroom("TS_ALUNO_9EF_2019.csv", locale = locale(encoding = "Latin1"), delim = "
df_escola <- vroom("TS_ESCOLA_2019.csv", locale = locale(encoding = "Latin1"), delim = ";")

unique(df_escola$NIVEL_SOCIO_ECONOMICO)

dim(df_aluno)
dim(df_escola)

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 <- vroom("TS_ALUNO_9EF_2019_co.csv", locale = locale(encoding = "UTF-8"))
df_escola <- vroom("TS_ESCOLA_2019_co.csv", locale = locale(encoding = "UTF-8"))

df <- inner_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))

dim(df)
write.csv(df, "dados_tcc2.csv")

```

## Leitura dos dados

```

df <- vroom("dados_tcc2.csv", locale = locale(encoding = "UTF-8"), delim = ",")

var_df <- names(df)
dim(df)

df <- df %>%
  filter(IN_PREENCHIMENTO_LP == 1, IN_PREENCHIMENTO_MT == 1)

set.seed(123)
df_sampled <- df %>% sample_n(5000)

theme_tcc <- function() {
  theme_classic() +
    theme(
      axis.line = element_line(colour = "black"),
      panel.border = element_rect(colour = "black", fill=NA),
      panel.background = element_blank(),
      plot.background = element_blank(),
      axis.ticks = element_line(colour = "black"),
      panel.grid.major = element_blank(),
      panel.grid.minor = element_blank(),
      axis.text = element_text(colour = "black"),
      axis.title = element_text(colour="black", size=18)
    )
}

```

## Análise descritiva dos dados

### Quantidade de Escolas

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

### Quantidade de Alunos por Escola

```
df %>%  
  group_by(ID_ESCOLA) %>%  
  summarise(n = n()) %>%  
  ggplot(aes(x = n)) +  
  geom_histogram(binwidth = 10, fill = "white", color="black") +  
  labs(x = "Quantidade de alunos",  
       y = "Frequência") +  
  theme_tcc()  
ggsave("img/dist_alunos_escola.pdf", width = 8, height = 5, dpi = 300)  
  
# quadro resumo  
df %>%  
  group_by(ID_ESCOLA) %>%  
  summarise(n = n()) %>%  
  summarise(  
    Média = mean(n),  
    `Desvio Padrão` = sd(n),  
    Mínimo = min(n),  
    `Primeiro Quartil` = quantile(n, 0.25),  
    Mediana = median(n),  
    `Terceiro Quartil` = quantile(n, 0.75),  
    Máximo = max(n)  
  ) %>%  
  mutate(across(everything(), ~format(round(., 2), nsmall = 2))) %>%  
  pivot_longer(everything(), names_to = "Estatística", values_to = "Número de Alunos por Escola")  
kable(booktabs = TRUE, caption = "Summary Statistics for Number of Students per School",
```

Análise:

## Nível Socioeconômico (NSe) das Escolas

```
nse_freq <- df %>%
  mutate(NIVEL_SOCIO_ECONOMICO = as.factor(NIVEL_SOCIO_ECONOMICO)) %>%
  filter(!is.na(NIVEL_SOCIO_ECONOMICO)) %>%
  count(NIVEL_SOCIO_ECONOMICO) %>%
  mutate(prop = round(n / sum(n) * 100, 2)) %>%
  select(NIVEL_SOCIO_ECONOMICO, prop) # Select only the desired columns

kable(nse_freq, caption = "Frequência do Nível Socioeconômico das Escolas", booktabs = TRUE,
      col.names = c("Nível Socioeconômico", "Proporção (%)", align = "c", format = "latex"
kable_styling()
```

## Área das Escolas

```
area_freq <- df %>%
  mutate(ID_AREA = ifelse(ID_AREA == 1, "Capital", "Interior")) %>%
  count(ID_AREA) %>%
  mutate(prop = round(n / sum(n) * 100, 2)) %>%
  select(ID_AREA, prop)

kable(area_freq, caption = "Frequência da Área das Escolas", booktabs = TRUE,
      col.names = c("Área", "Proporção (%)", align = "c", format = "latex")
```

## Localização das Escolas

```
loc_freq <- df %>%
  mutate(ID_LOCALIZACAO = ifelse(ID_LOCALIZACAO == 1, "Urbana", "Rural")) %>%
  mutate(ID_LOCALIZACAO = factor(ID_LOCALIZACAO)) %>%
  count(ID_LOCALIZACAO) %>%
  mutate(prop = round(n / sum(n) * 100, 2)) %>%
  #select(ID_LOCALIZACAO, prop)

kable(loc_freq, caption = "Frequência da Administração das Escolas", booktabs = TRUE, align
```

## Distribuição da proficiência em Língua Portuguesa e Matemática

```
#lingua portuguesa (ggplot histogram)
df %>%
  ggplot(aes(x = PROFICIENCIA_LP_SAEB)) +
  geom_histogram(binwidth = 10, fill = "white", color="black") +
  labs(x = "Língua Portuguesa",
       y = "Frequência") +
  xlim(100,400)+
  ylim(0,13000)+
  theme_tcc()
ggsave("img/dist_lp.pdf", width = 5, height = 5, dpi = 300)

#lingua portuguesa (qqplot)
df_sampled %>%
  ggplot(aes(sample = PROFICIENCIA_LP_SAEB)) +
  stat_qq() +
  labs(x = "Quantis teóricos", y = "Quantis amostrais") +
  ylim(100,400)+
  theme_tcc()
ggsave("img/qqplot_lp.pdf", width = 5, height = 5, dpi = 300)

#lingua portuguesa (boxplot)
df %>%
  ggplot(aes(y = PROFICIENCIA_MT_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Língua Portuguesa", y="") +
  scale_x_discrete(labels = c("Língua Portuguesa")) +
  ylim(100,400)+
  theme_tcc()
ggsave("img/boxplot_lp.pdf", width = 5, height = 5, dpi = 300)

#matemática (ggplot histogram)
df %>%
  ggplot(aes(x = PROFICIENCIA_MT_SAEB)) +
  geom_histogram(binwidth = 10, fill = "white", color="black") +
  labs(x = "Matemática",
       y = "Frequência") +
  xlim(100,400)+
  ylim(0,13000)+
  theme_tcc()
```

```

ggsave("img/dist_mt.pdf", width = 5, height = 5, dpi = 300)

#matematica (qqplot)
df_sampled %>%
  ggplot(aes(sample = PROFICIENCIA_MT_SAEB)) +
  stat_qq() +
  geom_abline(slope = 1, intercept = 0, color = "red") +
  labs(x = "Quantis teóricos", y = "Quantis amostrais") +
  ylim(100,400)+
  theme_tcc()
ggsave("img/qqplot_mt.pdf", width = 5, height = 5, dpi = 300)

#matematica (boxplot)
df %>%
  ggplot(aes(y = PROFICIENCIA_MT_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Matemática", y="") +
  scale_x_discrete(labels = c("Matemática")) +
  ylim(100,400)+
  theme_tcc()
ggsave("img/boxplot_mt.pdf", width = 5, height = 5, dpi = 300)

```

## Quadro Resumo da proficiência em Língua Portuguesa e Matemática

```

lp_stats <- summarise(df,
  Média = mean(PROFICIENCIA_LP_SAEB, na.rm = TRUE),
  `Desvio Padrão` = sd(PROFICIENCIA_LP_SAEB, na.rm = TRUE),
  Mínimo = min(PROFICIENCIA_LP_SAEB, na.rm = TRUE),
  `Primeiro Quartil` = quantile(PROFICIENCIA_LP_SAEB, 0.25, na.rm = TRUE),
  Mediana = median(PROFICIENCIA_LP_SAEB, na.rm = TRUE),
  `Terceiro Quartil` = quantile(PROFICIENCIA_LP_SAEB, 0.75, na.rm = TRUE),
  Máximo = max(PROFICIENCIA_LP_SAEB, na.rm = TRUE)
) %>% pivot_longer(everything(), names_to = "Estatística", values_to = "Língua Portuguesa")

# Calculate summary statistics for Matemática
mt_stats <- summarise(df,
  Média = mean(PROFICIENCIA_MT_SAEB, na.rm = TRUE),
  `Desvio Padrão` = sd(PROFICIENCIA_MT_SAEB, na.rm = TRUE),
  Mínimo = min(PROFICIENCIA_MT_SAEB, na.rm = TRUE),
  `Primeiro Quartil` = quantile(PROFICIENCIA_MT_SAEB, 0.25, na.rm = TRUE),

```

```

Mediana = median(PROFICIENCIA_MT_SAEB, na.rm = TRUE),
`Terceiro Quartil` = quantile(PROFICIENCIA_MT_SAEB, 0.75, na.rm = TRUE),
Máximo = max(PROFICIENCIA_MT_SAEB, na.rm = TRUE)
) %>% pivot_longer(everything(), names_to = "Estatística", values_to = "Matemática")

# Combine the statistics into one data frame
combined_stats <- full_join(lp_stats, mt_stats, by = "Estatística")

# round to 2 decimal places and decimal mark as comma
combined_stats <- combined_stats %>% mutate(across(where(is.numeric), ~format(round(., 2),

combined_stats %>%
  kable(booktabs = TRUE, caption = "Summary Statistics for Língua Portuguesa and Matemática",
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"), full_width = F, fo
  column_spec(1, bold = T, border_right = TRUE) %>%
  scroll_box(width = "100%", height = "500px")

```

## Boxplots da proficiência em Língua Portuguesa e Matemática

### Por Nível Socioeconômico

```

df %>%
  filter(!is.na(NIVEL_SOCIO_ECONOMICO)) %>%
  ggplot(aes(x = as.factor(NIVEL_SOCIO_ECONOMICO), y = PROFICIENCIA_LP_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Nível Socioeconômico da Escola",
       y = "Língua Portuguesa") +
  ylim(100,400)+
  theme_tcc()
ggsave("img/boxplot_lp_nse.pdf", width = 8, height = 5, dpi = 300)

df %>%
  filter(!is.na(NIVEL_SOCIO_ECONOMICO)) %>%
  ggplot(aes(x = as.factor(NIVEL_SOCIO_ECONOMICO), y = PROFICIENCIA_MT_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Nível Socioeconômico da Escola",
       y = "Matemática") +
  ylim(100,400)+
  theme_tcc()
ggsave("img/boxplot_mt_nse.pdf", width = 8, height = 5, dpi = 300)

```

## Por Área

```
df %>%
  mutate(ID_AREA = ifelse(ID_AREA == 1, "Capital", "Interior")) %>%
  ggplot(aes(x = as.factor(ID_AREA), y = PROFICIENCIA_LP_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Área da Escola",
       y = "Língua Portuguesa") +
  ylim(100,400)+
  theme_tcc()
ggsave("img/boxplot_lp_area.pdf", width = 8, height = 5, dpi = 300)

df %>%
  mutate(ID_AREA = ifelse(ID_AREA == 1, "Capital", "Interior")) %>%
  ggplot(aes(x = as.factor(ID_AREA), y = PROFICIENCIA_MT_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Área da Escola",
       y = "Matemática") +
  ylim(100,400)+
  theme_tcc()
ggsave("img/boxplot_mt_area.pdf", width = 8, height = 5, dpi = 300)
```

## Por Administração

```
df %>%
  mutate(ID_LOCALIZACAO = ifelse(ID_LOCALIZACAO == 1, "Urbana", "Rural")) %>%
  ggplot(aes(x = as.factor(ID_LOCALIZACAO), y = PROFICIENCIA_LP_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Localização da Escola",
       y = "Língua Portuguesa") +
  ylim(100,400)+
  theme_tcc()
ggsave("img/boxplot_lp_loc.pdf", width = 8, height = 5, dpi = 300)

df %>%
  mutate(ID_LOCALIZACAO = ifelse(ID_LOCALIZACAO == 1, "Urbana", "Rural")) %>%
  ggplot(aes(x = as.factor(ID_LOCALIZACAO), y = PROFICIENCIA_MT_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Localização da Escola",
       y = "Matemática") +
```



```

ylim(100,400)+
theme_tcc()
ggsave("img/boxplot_mt_loc.pdf", width = 8, height = 5, dpi = 300)

```

## Por TX\_RESP\_Q002

Branca. Preta. Parda. Amarela. Indígena. Não quero declarar.

```

unique(df$TX_RESP_Q002)

df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q002 %in% c(".", "*")) %>%
  mutate(TX_RESP_Q002 = case_when(
    TX_RESP_Q002 == "A" ~ "Branca",
    TX_RESP_Q002 == "B" ~ "Preta",
    TX_RESP_Q002 == "C" ~ "Parda",
    TX_RESP_Q002 == "D" ~ "Amarela",
    TX_RESP_Q002 == "E" ~ "Indígena",
    TX_RESP_Q002 == "F" ~ "Não declarado",
    TRUE ~ TX_RESP_Q002
  )) %>%
  mutate(TX_RESP_Q002 = factor(TX_RESP_Q002, levels = c("Branca", "Preta", "Parda", "Amarela", "Indígena", "Não declarado")))
ggplot(aes(x = TX_RESP_Q002, y = PROFICIENCIA_LP_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Cor/raça autodeclarada",
       y = "Língua Portuguesa") +
  ylim(100,400)+
  theme_tcc()
ggsave("img/boxplot_lp_raca.pdf", width = 8, height = 5, dpi = 300)

df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q002 %in% c(".", "*")) %>%
  mutate(TX_RESP_Q002 = case_when(
    TX_RESP_Q002 == "A" ~ "Branca",
    TX_RESP_Q002 == "B" ~ "Preta",
    TX_RESP_Q002 == "C" ~ "Parda",
    TX_RESP_Q002 == "D" ~ "Amarela",
    TX_RESP_Q002 == "E" ~ "Indígena",
    TX_RESP_Q002 == "F" ~ "Não declarado",
    TRUE ~ TX_RESP_Q002
  )) %>%
  mutate(TX_RESP_Q002 = factor(TX_RESP_Q002, levels = c("Branca", "Preta", "Parda", "Amarela", "Indígena", "Não declarado")))

```

```

)) %>%
mutate(TX_RESP_Q002 = factor(TX_RESP_Q002, levels = c("Branca", "Preta", "Parda", "Amarela", "Indeterminada"))) +
ggplot(aes(x = TX_RESP_Q002, y = PROFICIENCIA_MT_SAEB)) +
geom_boxplot(fill = "white") +
labs(x = "Cor/raça autodeclarada",
      y = "Matemática") +
ylim(100,400)+
theme_tcc()
ggsave("img/boxplot_mt_raca.pdf", width = 8, height = 5, dpi = 300)

```

## Por TX\_RESP\_Q004

```

unique(df$TX_RESP_Q004)

df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q004 %in% c(".", "*")) %>%
  mutate(TX_RESP_Q004 = case_when(
    TX_RESP_Q004 == "A" ~ "< Fundamental",
    TX_RESP_Q004 == "B" ~ "Fundamental Incomp",
    TX_RESP_Q004 == "C" ~ "Fundamental",
    TX_RESP_Q004 == "D" ~ "Médio",
    TX_RESP_Q004 == "E" ~ "Superior",
    TX_RESP_Q004 == "F" ~ "Não sei",
    TRUE ~ TX_RESP_Q004
  )) %>%
  mutate(TX_RESP_Q004 = factor(TX_RESP_Q004, levels = c("Não sei", "< Fundamental", "Fundamental", "Fundamental Incomp", "Médio", "Superior"))) +
  ggplot(aes(x = TX_RESP_Q004, y = PROFICIENCIA_LP_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Escolaridade da mãe",
        y = "Língua Portuguesa") +
  ylim(100,400)+
  theme_tcc()
ggsave("img/boxplot_lp_mae.pdf", width = 8, height = 5, dpi = 300)

df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q004 %in% c(".", "*")) %>%
  mutate(TX_RESP_Q004 = case_when(
    TX_RESP_Q004 == "A" ~ "< Fundamental",
    TX_RESP_Q004 == "B" ~ "Fundamental Incomp",
    TX_RESP_Q004 == "C" ~ "Fundamental",

```

```

    TX_RESP_Q004 == "D" ~ "Médio",
    TX_RESP_Q004 == "E" ~ "Superior",
    TX_RESP_Q004 == "F" ~ "Não sei",
    TRUE ~ TX_RESP_Q004
  )) %>%
  mutate(TX_RESP_Q004 = factor(TX_RESP_Q004, levels = c("Não sei", "< Fundamental", "Fund
  ggplot(aes(x = TX_RESP_Q004, y = PROFICIENCIA_MT_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Escolaridade da mae",
        y = "Matemática") +
  ylim(100,400)+
  theme_tcc()
  ggsave("img/boxplot_mt_mae.pdf", width = 8, height = 5, dpi = 300)

```

## Por TX\_RESP\_Q011

```

df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q011 %in% c(".", "*")) %>%
  mutate(TX_RESP_Q011 = case_when(
    TX_RESP_Q011 == "A" ~ "Menos de 30min",
    TX_RESP_Q011 == "B" ~ "Entre 30min e 1h",
    TX_RESP_Q011 == "C" ~ "Mais de 1h",
    TRUE ~ TX_RESP_Q011
  )) %>%
  mutate(TX_RESP_Q011 = factor(TX_RESP_Q011, levels = c("Menos de 30min", "Entre 30min e 1
  ggplot(aes(x = TX_RESP_Q011, y = PROFICIENCIA_LP_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Tempo para chegar a escola", y = "Língua Portuguesa") +
  ylim(100,400)+
  theme_tcc()
  ggsave("img/boxplot_lp_tempo.pdf", width = 8, height = 5, dpi = 300)

df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q011 %in% c(".", "*")) %>%
  mutate(TX_RESP_Q011 = case_when(
    TX_RESP_Q011 == "A" ~ "Menos de 30min",
    TX_RESP_Q011 == "B" ~ "Entre 30min e 1h",
    TX_RESP_Q011 == "C" ~ "Mais de 1h",
    TRUE ~ TX_RESP_Q011
  )) %>%

```

```

mutate(TX_RESP_Q011 = factor(TX_RESP_Q011, levels = c("Menos de 30min", "Entre 30min e
ggplot(aes(x = as.factor(TX_RESP_Q011), y = PROFICIENCIA_MT_SAEB)) +
geom_boxplot(fill = "white") +
labs(x = "Tempo para chegar a escola",
      y = "Matemática") +
ylim(100,400)+
theme_tcc()
ggsave("img/boxplot_mt_tempo.pdf", width = 8, height = 5, dpi = 300)

```

## Por TX\_RESP\_Q015

```

df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q015 %in% c(".", "*")) %>%
  mutate(TX_RESP_Q015 = case_when(
    TX_RESP_Q015 == "A" ~ "Não",
    TX_RESP_Q015 == "B" ~ "1 vez",
    TX_RESP_Q015 == "C" ~ "2 ou mais",
    TRUE ~ TX_RESP_Q015
  )) %>%
  mutate(TX_RESP_Q015 = factor(TX_RESP_Q015, levels = c("Não", "1 vez", "2 ou mais"))) %>%
  ggplot(aes(x = TX_RESP_Q015, y = PROFICIENCIA_LP_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Reprovações",
       y = "Língua Portuguesa") +
  ylim(100,400)+
  theme_tcc()
ggsave("img/boxplot_lp_reprovacao.pdf", width = 8, height = 5, dpi = 300)

df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q015 %in% c(".", "*")) %>%
  mutate(TX_RESP_Q015 = case_when(
    TX_RESP_Q015 == "A" ~ "Não",
    TX_RESP_Q015 == "B" ~ "1 vez",
    TX_RESP_Q015 == "C" ~ "2 ou mais",
    TRUE ~ TX_RESP_Q015
  )) %>%
  mutate(TX_RESP_Q015 = factor(TX_RESP_Q015, levels = c("Não", "1 vez", "2 ou mais"))) %>%
  ggplot(aes(x = as.factor(TX_RESP_Q015), y = PROFICIENCIA_MT_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Reprovações",

```

```

      y = "Matemática") +
    ylim(100,400)+
    theme_tcc()
  ggsave("img/boxplot_mt_reprovacao.pdf", width = 8, height = 5, dpi = 300)

```

## Por TX\_RESP\_Q016

```

df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q016 %in% c(".", "*")) %>%
  mutate(TX_RESP_Q016 = case_when(
    TX_RESP_Q016 == "A" ~ "Não",
    TX_RESP_Q016 == "B" ~ "1 vez",
    TX_RESP_Q016 == "C" ~ "2 ou mais",
    TRUE ~ TX_RESP_Q016
  )) %>%
  mutate(TX_RESP_Q016 = factor(TX_RESP_Q016, levels = c("Não", "1 vez", "2 ou mais"))) %>%
  ggplot(aes(x = as.factor(TX_RESP_Q016), y = PROFICIENCIA_LP_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Abandono escolar",
       y = "Língua Portuguesa") +
  ylim(100,400)+
  theme_tcc()
  ggsave("img/boxplot_lp_abandono.pdf", width = 8, height = 5, dpi = 300)

df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q016 %in% c(".", "*")) %>%
  mutate(TX_RESP_Q016 = case_when(
    TX_RESP_Q016 == "A" ~ "Não",
    TX_RESP_Q016 == "B" ~ "1 vez",
    TX_RESP_Q016 == "C" ~ "2 ou mais",
    TRUE ~ TX_RESP_Q016
  )) %>%
  mutate(TX_RESP_Q016 = factor(TX_RESP_Q016, levels = c("Não", "1 vez", "2 ou mais"))) %>%
  ggplot(aes(x = as.factor(TX_RESP_Q016), y = PROFICIENCIA_MT_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Abandono escolar",
       y = "Matemática") +
  ylim(100,400)+
  theme_tcc()
  ggsave("img/boxplot_mt_abandono.pdf", width = 8, height = 5, dpi = 300)

```

Nas questões Q017[a-z] (a seguir)

- A: Não uso meu tempo para isso.
- B: Menos de 1 hora.
- C: Entre 1 e 2 horas.
- D: Mais de 2 horas.

### Por TX\_RESP\_Q017A

```
unique(df$TX_RESP_Q017A)
```

```
df %>%
```

```
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q017A %in% c(".", "*")) %>%
```

```
  mutate(TX_RESP_Q017A = case_when(
```

```
    TX_RESP_Q017A == "A" ~ "Não tem",
```

```
    TX_RESP_Q017A == "B" ~ "Menos de 1h",
```

```
    TX_RESP_Q017A == "C" ~ "Entre 1 e 2h",
```

```
    TX_RESP_Q017A == "D" ~ "Mais de 2h",
```

```
    TRUE ~ TX_RESP_Q017A
```

```
  )) %>%
```

```
  mutate(TX_RESP_Q017A = factor(TX_RESP_Q017A, levels = c("Não tem", "Menos de 1h", "Entre
```

```
ggplot(aes(x = TX_RESP_Q017A, y = PROFICIENCIA_LP_SAEB)) +
```

```
geom_boxplot(fill = "white") +
```

```
labs(x = "Tempo fazendo cursos",
```

```
     y = "Língua Portuguesa") +
```

```
ylim(100,400)+
```

```
theme_tcc()
```

```
ggsave("img/boxplot_lp_lazer.pdf", width = 8, height = 5, dpi = 300)
```

```
df %>%
```

```
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q017A %in% c(".", "*")) %>%
```

```
  mutate(TX_RESP_Q017A = case_when(
```

```
    TX_RESP_Q017A == "A" ~ "Não tem",
```

```
    TX_RESP_Q017A == "B" ~ "Menos de 1h",
```

```
    TX_RESP_Q017A == "C" ~ "Entre 1 e 2h",
```

```
    TX_RESP_Q017A == "D" ~ "Mais de 2h",
```

```
    TRUE ~ TX_RESP_Q017A
```

```
  )) %>%
```

```
  mutate(TX_RESP_Q017A = factor(TX_RESP_Q017A, levels = c("Não tem", "Menos de 1h", "Ent
```

```
ggplot(aes(x = TX_RESP_Q017A, y = PROFICIENCIA_MT_SAEB)) +
```

```

geom_boxplot(fill = "white") +
labs(x = "Lazer",
      y = "Matemática") +
ylim(100,400)+
theme_tcc()
ggsave("img/boxplot_mt_lazer.pdf", width = 8, height = 5, dpi = 300)

```

## Por TX\_RESP\_Q017B (Fazer cursos)

```
unique(df$TX_RESP_Q017B)
```

```
df %>%
```

```

filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q017B %in% c(".", "*")) %>%
mutate(TX_RESP_Q017B = case_when(
  TX_RESP_Q017B == "A" ~ "Não tem",
  TX_RESP_Q017B == "B" ~ "Menos de 1h",
  TX_RESP_Q017B == "C" ~ "Entre 1 e 2h",
  TX_RESP_Q017B == "D" ~ "Mais de 2h",
  TRUE ~ TX_RESP_Q017B)) %>%

```

```

mutate(TX_RESP_Q017B = factor(TX_RESP_Q017B, levels = c("Não tem", "Menos de 1h", "Ent
ggplot(aes(x = TX_RESP_Q017B, y = PROFICIENCIA_LP_SAEB)) +
geom_boxplot(fill = "white") +
labs(x = "Tempo fazendo cursos",
      y = "Língua Portuguesa") +
ylim(100,400)+
theme_tcc()
ggsave("img/boxplot_lp_cursos.pdf", width = 8, height = 5, dpi = 300)

```

```
df %>%
```

```

filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q017B %in% c(".", "*")) %>%
mutate(TX_RESP_Q017B = case_when(
  TX_RESP_Q017B == "A" ~ "Não tem",
  TX_RESP_Q017B == "B" ~ "Menos de 1h",
  TX_RESP_Q017B == "C" ~ "Entre 1 e 2h",
  TX_RESP_Q017B == "D" ~ "Mais de 2h",
  TRUE ~ TX_RESP_Q017B)) %>%

```

```

mutate(TX_RESP_Q017B = factor(TX_RESP_Q017B, levels = c("Não tem", "Menos de 1h", "Ent
ggplot(aes(x = TX_RESP_Q017B, y = PROFICIENCIA_MT_SAEB)) +
geom_boxplot(fill = "white") +
labs(x = "Tempo fazendo cursos",

```

```

      y = "Matemática") +
    ylim(100,400)+
    theme_tcc()
  ggsave("img/boxplot_mt_cursos.pdf", width = 8, height = 5, dpi = 300)

```

### Por TX\_RESP\_Q017C (Trabalhos domesticos)

```

df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q017C %in% c(".", "*")) %>%
  mutate(TX_RESP_Q017C = case_when(
    TX_RESP_Q017C == "A" ~ "Não tem",
    TX_RESP_Q017C == "B" ~ "Menos de 1h",
    TX_RESP_Q017C == "C" ~ "Entre 1 e 2h",
    TX_RESP_Q017C == "D" ~ "Mais de 2h",
    TRUE ~ TX_RESP_Q017C)) %>%
  mutate(TX_RESP_Q017C = factor(TX_RESP_Q017C, levels = c("Não tem", "Menos de 1h", "Entre 1 e 2h", "Mais de 2h")))
  ggplot(aes(x = TX_RESP_Q017C, y = PROFICIENCIA_LP_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Tempo fazendo trabalhos domésticos",
       y = "Língua Portuguesa") +
  ylim(100,400)+
  theme_tcc()
  ggsave("img/boxplot_lp_domesticos.pdf", width = 8, height = 5, dpi = 300)

df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q017C %in% c(".", "*")) %>%
  mutate(TX_RESP_Q017C = case_when(
    TX_RESP_Q017C == "A" ~ "Não tem",
    TX_RESP_Q017C == "B" ~ "Menos de 1h",
    TX_RESP_Q017C == "C" ~ "Entre 1 e 2h",
    TX_RESP_Q017C == "D" ~ "Mais de 2h",
    TRUE ~ TX_RESP_Q017C)) %>%
  mutate(TX_RESP_Q017C = factor(TX_RESP_Q017C, levels = c("Não tem", "Menos de 1h", "Entre 1 e 2h", "Mais de 2h")))
  ggplot(aes(x = TX_RESP_Q017C, y = PROFICIENCIA_MT_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Tempo fazendo trabalhos domésticos",
       y = "Matemática") +
  ylim(100,400)+
  theme_tcc()
  ggsave("img/boxplot_mt_domesticos.pdf", width = 8, height = 5, dpi = 300)

```



## Por TX\_RESP\_Q017D

```
df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q017D %in% c(".", "*")) %>%
  mutate(TX_RESP_Q017D = case_when(
    TX_RESP_Q017D == "A" ~ "Não tem",
    TX_RESP_Q017D == "B" ~ "Menos de 1h",
    TX_RESP_Q017D == "C" ~ "Entre 1 e 2h",
    TX_RESP_Q017D == "D" ~ "Mais de 2h",
    TRUE ~ TX_RESP_Q017D)) %>%
  mutate(TX_RESP_Q017D = factor(TX_RESP_Q017D, levels = c("Não tem", "Menos de 1h", "Ent
ggplot(aes(x = TX_RESP_Q017D, y = PROFICIENCIA_LP_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Tempo de estudo",
       y = "Língua Portuguesa") +
  ylim(100,400)+
  theme_tcc()
ggsave("img/boxplot_lp_estudo.pdf", width = 8, height = 5, dpi = 300)

df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q017D %in% c(".", "*")) %>%
  mutate(TX_RESP_Q017D = case_when(
    TX_RESP_Q017D == "A" ~ "Não tem",
    TX_RESP_Q017D == "B" ~ "Menos de 1h",
    TX_RESP_Q017D == "C" ~ "Entre 1 e 2h",
    TX_RESP_Q017D == "D" ~ "Mais de 2h",
    TRUE ~ TX_RESP_Q017D)) %>%
  mutate(TX_RESP_Q017D = factor(TX_RESP_Q017D, levels = c("Não tem", "Menos de 1h", "Ent
ggplot(aes(x = TX_RESP_Q017D, y = PROFICIENCIA_MT_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Tempo de estudo",
       y = "Matemática") +
  ylim(100,400)+
  theme_tcc()
ggsave("img/boxplot_mt_estudo.pdf", width = 8, height = 5, dpi = 300)
```

## Por TX\_RESP\_Q017E

```
df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q017E %in% c(".", "*")) %>%
  mutate(TX_RESP_Q017E = case_when(
    TX_RESP_Q017E == "A" ~ "Não tem",
    TX_RESP_Q017E == "B" ~ "Menos de 1h",
    TX_RESP_Q017E == "C" ~ "Entre 1 e 2h",
    TX_RESP_Q017E == "D" ~ "Mais de 2h",
    TRUE ~ TX_RESP_Q017E)) %>%
  mutate(TX_RESP_Q017E = factor(TX_RESP_Q017E, levels = c("Não tem", "Menos de 1h", "Ent
ggplot(aes(x = as.factor(TX_RESP_Q017E), y = PROFICIENCIA_LP_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Tempo de trabalho fora de casa",
       y = "Língua Portuguesa") +
  ylim(100,400)+
  theme_tcc()
ggsave("img/boxplot_lp_trabalho.pdf", width = 8, height = 5, dpi = 300)

df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q017E %in% c(".", "*")) %>%
  mutate(TX_RESP_Q017E = case_when(
    TX_RESP_Q017E == "A" ~ "Não tem",
    TX_RESP_Q017E == "B" ~ "Menos de 1h",
    TX_RESP_Q017E == "C" ~ "Entre 1 e 2h",
    TX_RESP_Q017E == "D" ~ "Mais de 2h",
    TRUE ~ TX_RESP_Q017E)) %>%
  mutate(TX_RESP_Q017E = factor(TX_RESP_Q017E, levels = c("Não tem", "Menos de 1h", "Ent
ggplot(aes(x = as.factor(TX_RESP_Q017E), y = PROFICIENCIA_MT_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Tempo de trabalho fora de casa",
       y = "Matemática") +
  ylim(100,400)+
  theme_tcc()
ggsave("img/boxplot_mt_trabalho.pdf", width = 8, height = 5, dpi = 300)
```

Em relação às questões Q018[a-z] (a seguir)

- A: Nunca ou quase nunca
- B: De vez em quando
- C: Sempre ou quase sempre

## Por TX\_RESP\_Q018A

```
unique(df$TX_RESP_Q018A)

df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q018A %in% c(".", "*")) %>%
  mutate(TX_RESP_Q018A = case_when(
    TX_RESP_Q018A == "A" ~ "Nunca ou quase nunca",
    TX_RESP_Q018A == "B" ~ "De vez em quando",
    TX_RESP_Q018A == "C" ~ "Sempre ou quase sempre",
    TRUE ~ TX_RESP_Q018A
  )) %>%
  mutate(TX_RESP_Q018A = factor(TX_RESP_Q018A, levels = c("Nunca ou quase nunca", "De vez em quando", "Sempre ou quase sempre")))
  ggplot(aes(x = TX_RESP_Q018A, y = PROFICIENCIA_LP_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Leitura de notícias",
       y = "Língua Portuguesa") +
  ylim(100, 400) +
  theme_tcc()
ggsave("img/boxplot_lp_noticias.pdf", width = 8, height = 5, dpi = 300)

df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q018A %in% c(".", "*")) %>%
  mutate(TX_RESP_Q018A = case_when(
    TX_RESP_Q018A == "A" ~ "Nunca ou quase nunca",
    TX_RESP_Q018A == "B" ~ "De vez em quando",
    TX_RESP_Q018A == "C" ~ "Sempre ou quase sempre",
    TRUE ~ TX_RESP_Q018A
  )) %>%
  mutate(TX_RESP_Q018A = factor(TX_RESP_Q018A, levels = c("Nunca ou quase nunca", "De vez em quando", "Sempre ou quase sempre")))
  ggplot(aes(x = TX_RESP_Q018A, y = PROFICIENCIA_MT_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Leitura de notícias",
       y = "Matemática") +
  ylim(100, 400) +
  theme_tcc()
ggsave("img/boxplot_mt_noticias.pdf", width = 8, height = 5, dpi = 300)
```

## Por TX\_RESP\_Q018B (Livros)

```
df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q018B %in% c(".", "*")) %>%
  mutate(TX_RESP_Q018B = case_when(
    TX_RESP_Q018B == "A" ~ "Nunca ou quase nunca",
    TX_RESP_Q018B == "B" ~ "De vez em quando",
    TX_RESP_Q018B == "C" ~ "Sempre ou quase sempre",
    TRUE ~ TX_RESP_Q018B
  )) %>%
  mutate(TX_RESP_Q018B = factor(TX_RESP_Q018B, levels = c("Nunca ou quase nunca", "De vez em quando", "Sempre ou quase sempre")))
ggplot(aes(x = TX_RESP_Q018B, y = PROFICIENCIA_LP_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Leitura de livros",
       y = "Língua Portuguesa") +
  ylim(100,400)+
  theme_tcc()
ggsave("img/boxplot_lp_livros.pdf", width = 8, height = 5, dpi = 300)

df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q018B %in% c(".", "*")) %>%
  mutate(TX_RESP_Q018B = case_when(
    TX_RESP_Q018B == "A" ~ "Nunca ou quase nunca",
    TX_RESP_Q018B == "B" ~ "De vez em quando",
    TX_RESP_Q018B == "C" ~ "Sempre ou quase sempre",
    TRUE ~ TX_RESP_Q018B
  )) %>%
  mutate(TX_RESP_Q018B = factor(TX_RESP_Q018B, levels = c("Nunca ou quase nunca", "De vez em quando", "Sempre ou quase sempre")))
ggplot(aes(x = TX_RESP_Q018B, y = PROFICIENCIA_MT_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Leitura de livros",
       y = "Matemática") +
  ylim(100,400)+
  theme_tcc()
ggsave("img/boxplot_mt_livros.pdf", width = 8, height = 5, dpi = 300)
```

## Por TX\_RESP\_Q018C (Quadrinhos)

```
df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q018C %in% c(".", "*")) %>%
  mutate(TX_RESP_Q018C = case_when(
    TX_RESP_Q018C == "A" ~ "Nunca ou quase nunca",
    TX_RESP_Q018C == "B" ~ "De vez em quando",
    TX_RESP_Q018C == "C" ~ "Sempre ou quase sempre",
    TRUE ~ TX_RESP_Q018C
  )) %>%
  mutate(TX_RESP_Q018C = factor(TX_RESP_Q018C, levels = c("Nunca ou quase nunca", "De vez
ggplot(aes(x = TX_RESP_Q018C, y = PROFICIENCIA_LP_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Leitura de quadrinhos",
       y = "Língua Portuguesa") +
  ylim(100,400)+
  theme_tcc()
ggsave("img/boxplot_lp_quadrinhos.pdf", width = 8, height = 5, dpi = 300)

df %>%
  filter(IN_PREENCHIMENTO_QUESTIONARIO == 1, !TX_RESP_Q018C %in% c(".", "*")) %>%
  mutate(TX_RESP_Q018C = case_when(
    TX_RESP_Q018C == "A" ~ "Nunca ou quase nunca",
    TX_RESP_Q018C == "B" ~ "De vez em quando",
    TX_RESP_Q018C == "C" ~ "Sempre ou quase sempre",
    TRUE ~ TX_RESP_Q018C
  )) %>%
  mutate(TX_RESP_Q018C = factor(TX_RESP_Q018C, levels = c("Nunca ou quase nunca", "De vez
ggplot(aes(x = TX_RESP_Q018C, y = PROFICIENCIA_MT_SAEB)) +
  geom_boxplot(fill = "white") +
  labs(x = "Leitura de quadrinhos",
       y = "Matemática") +
  ylim(100,400)+
  theme_tcc()
ggsave("img/boxplot_mt_quadrinhos.pdf", width = 8, height = 5, dpi = 300)
```

## Correlação entre as variáveis

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