Projeto Final: Análise dos Microdados do ENEM 2023 com R

UNIT - Universidade Tiradentes

Curso: Ciência de Dados e Inteligência Artificial

Disciplina: Linguagens Python e R com Foco em Análise de Dados

2025

Fernanda Amaral de Souza; Larissa Castor Ramos; João Antônio Silveira Matos; Samuel Moreira da Cruz; Tenisson José Andrade Fonseca

Prof. Msc. Bruna Martini Dalmoro

O presente relatório apresenta uma análise estatística detalhada dos microdados do Exame Nacional do Ensino Médio (Enem) referentes à região Nordeste, com foco na compreensão do perfil socioeconômico e do desempenho acadêmico dos participantes. Inicialmente, procedeu-se à classificação das variáveis em quantitativas e qualitativas, permitindo a definição de abordagens analíticas adequadas para cada tipo de dado.

Essa classificação referente ao tipo das variáveis é uma análise descritiva e comparativa fundamental no processo de exploração e preparação dos dados, ela define como cada variável deve ser tratada nas etapas posteriores de análise estatística. No caso dos microdados do ENEM, essa verificação permite separar as variáveis que representam medidas numéricas como notas, idades e tempos de prova, daquelas que representam atributos ou categorias como sexo, cor/raça e região de residência.

Leitura e visualização do dataset dos microdados do ENEM. Com o objetivo de identificar e classificar as variáveis presentes no conjunto de microdados do ENEM em dois grupos principais (quantitativas e qualitativas), foi realizada uma análise inicial utilizando a função ‘str’. Para proporcionar melhor visualização foi criada uma função condicional utilizando ‘if/else’ onde as variáveis numéricas ou inteiras foram classificadas como quantitativas e as variáveis do tipo texto ou fator foram classificadas como qualitativas.

Os resultados dessa classificação foram organizados na Tabela 1 associando o nome de cada variável ao seu respectivo tipo a partir da função ‘sapply’ que percorreu cada coluna identificando o tipo de dado (qualitativo ou quantitativo). Esse resultado pode ser visualizado em gráfico de pizza (Gráfico 1), elaborado a partir da biblioteca ggplot2, concluindo que no dataset analisado existe um maior percentual (75%) de variáveis qualitativas e na Tabela 2 onde estão apresentados os dados absolutos.

Tabela 1: Descrição dos Tipos de Variáveis presentes no Dataset Microdados do Enem Nordeste.

Variavel	Tipo
NU_INSCRICAO	Quantitativa
NU_ANO	Quantitativa
TP_FAIXA_ETARIA	Qualitativa
TP_SEXO	Qualitativa
TP_ESTADO_CIVIL	Qualitativa
TP_COR_RACA	Qualitativa
TP_NACIONALIDADE	Qualitativa
TP_ST_CONCLUSAO	Qualitativa
TP_ANO_CONCLUIU	Qualitativa
TP_ESCOLA	Qualitativa
TP_ENSINO	Qualitativa
IN_TREINEIRO	Qualitativa
CO_MUNICIPIO_ESC	Quantitativa
NO_MUNICIPIO_ESC	Qualitativa
CO_UF_ESC	Quantitativa
SG_UF_ESC	Qualitativa
TP_DEPENDENCIA_ADM_ESC	Qualitativa
TP_LOCALIZACAO_ESC	Qualitativa
TP_SIT_FUNC_ESC	Qualitativa
CO_MUNICIPIO_PROVA	Quantitativa
NO_MUNICIPIO_PROVA	Qualitativa
CO_UF_PROVA	Quantitativa

SG_UF_PROVA	Qualitativa
TP_PRESENCA_CN	Qualitativa
TP_PRESENCA_CH	Qualitativa
TP_PRESENCA_LC	Qualitativa
TP_PRESENCA_MT	Qualitativa
CO_PROVA_CN	Qualitativa
CO_PROVA_CH	Qualitativa
CO_PROVA_LC	Qualitativa
CO_PROVA_MT	Qualitativa
NU_NOTA_CN	Quantitativa
NU_NOTA_CH	Quantitativa
NU_NOTA_LC	Quantitativa
NU_NOTA_MT	Quantitativa
TP_LINGUA	Qualitativa
TP_STATUS_REDACAO	Qualitativa
NU_NOTA_COMP1	Quantitativa
NU_NOTA_COMP2	Quantitativa
NU_NOTA_COMP3	Quantitativa
NU_NOTA_COMP4	Quantitativa
NU_NOTA_COMP5	Quantitativa
NU_NOTA_REDACAO	Quantitativa
Q001	Qualitativa
Q002	Qualitativa
Q003	Qualitativa
Q004	Qualitativa
Q005	Quantitativa
Q006	Qualitativa
Q007	Qualitativa
Q008	Qualitativa
Q009	Qualitativa

Q010	Qualitativa
Q011	Qualitativa
Q012	Qualitativa
Q013	Qualitativa
Q014	Qualitativa
Q015	Qualitativa
Q016	Qualitativa
Q017	Qualitativa
Q018	Qualitativa
Q019	Qualitativa
Q020	Qualitativa
Q021	Qualitativa
Q022	Qualitativa
Q023	Qualitativa
Q024	Qualitativa
Q025	Qualitativa

Gráfico 1: Distribuição percentual dos tipos de variáveis do Dataset Microdados do Enem Nordeste.

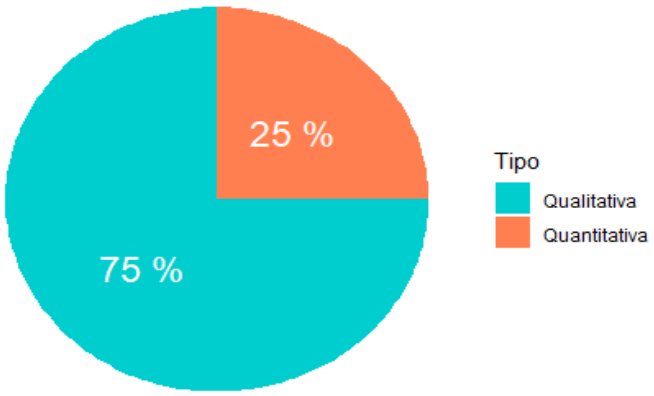
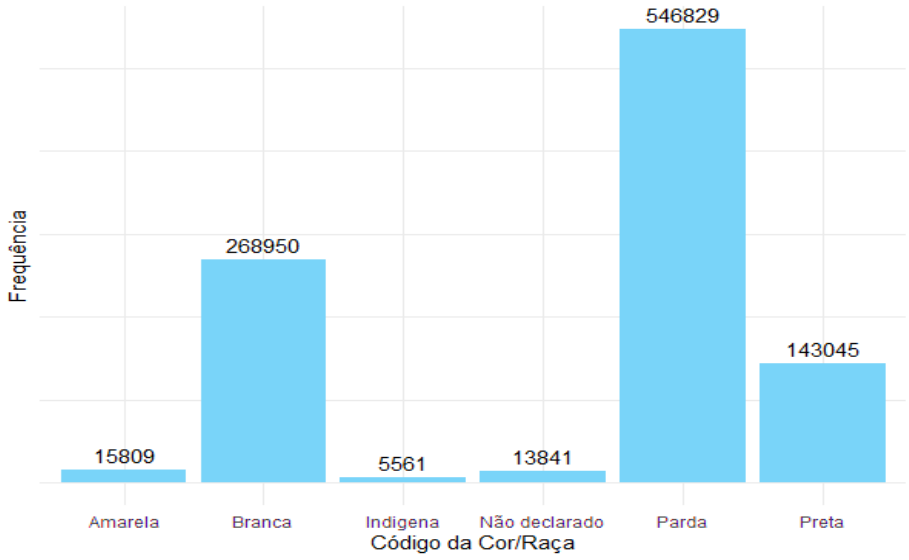


Tabela 2: Quantidade de variáveis quantitativas e qualitativas do Dataset Microdados do Enem Nordeste.

Tipo	Quantidade
Qualitativa	51
Quantitativa	17

Para compreender a composição do grupo avaliado em termos de diversidade e perfil demográfico foi avaliada a variável qualitativa nominal TP_COR_RACA. Para isso, foi gerado um gráfico de barras em que o eixo **x** corresponde às categorias oficiais referente a cor/raça e o eixo **y** indica o número de participantes em cada categoria (Gráfico 2). Essa abordagem é um recurso importante para análises estatísticas descritivas e subsidiar estudos sobre equidade, inclusão e políticas públicas na educação. É possível observar que a maior quantidade de participantes (546.829) se autodeclarou pardos, e apenas 5.561 participantes se autodeclararam indígenas, sendo a classificação com menor participação.

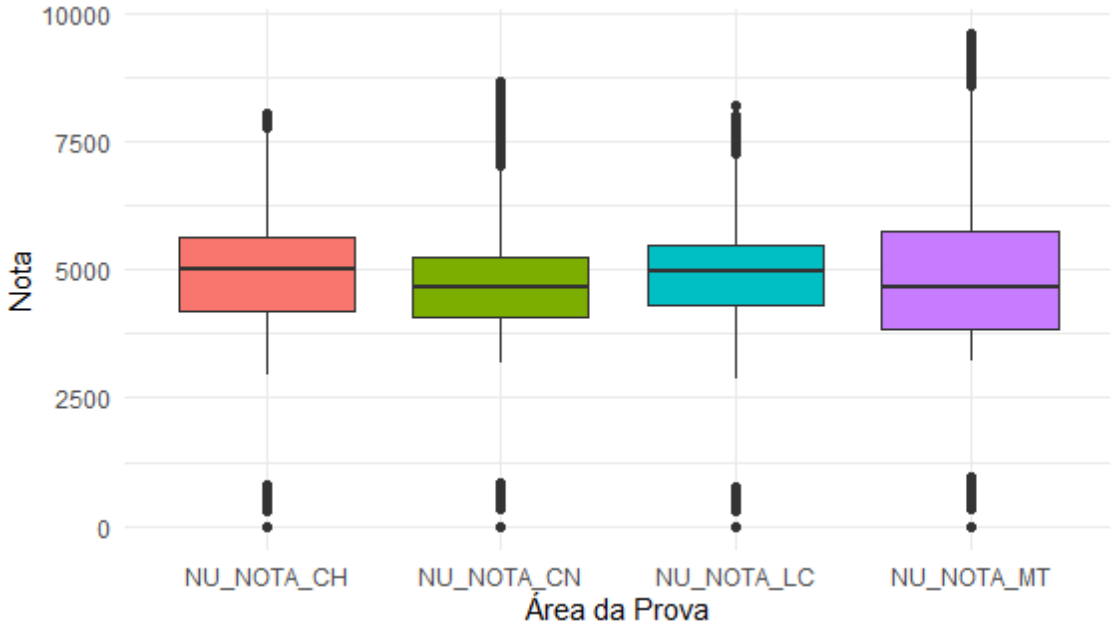
Gráfico 2: Frequência da variável TP_COR_RACA do Dataset Microdados do Enem Nordeste.



Para elaboração do gráfico de barras ilustrando a frequência da variável TP_COR_RACA, foi utilizada a função ‘as.factor(TP_COR_RACA)’ para converter a variável em fator, garantindo que seja interpretada como categórica pelo ‘ggplot2’. Em seguida, ‘geom_bar()’ foi empregada para calcular automaticamente as frequências e plotar as barras correspondentes a cada categoria. Já a função ‘geom_text()’ adicionou rótulos numéricos acima das barras, mostrando a contagem exata de observações para facilitar a leitura dos dados. E visando aprimorar a estética do gráfico, foi escolhido o tema ‘theme_minimal()’, oferecendo um visual limpo.

Para uma avaliação estatística das notas, foi gerado um gráfico de boxplot onde foi possível observar a distribuição das notas do ENEM nas quatro áreas de conhecimento: Ciências Humanas (NU_NOTA_CH), Ciências da Natureza (NU_NOTA_CN), Linguagens e Códigos (NU_NOTA_LC) e Matemática (NU_NOTA_MT). Cada caixa representa o intervalo entre o primeiro e o terceiro quartil (Q1 e Q3), com a linha interna indicando a mediana das notas. Observa-se que a mediana das notas de Ciências Humanas (4997) é levemente maior que a das demais áreas, seguida de Matemática (4672), Linguagens (4949) e Ciências da Natureza apresenta a menor mediana (4647). O comprimento dos bigodes sugere uma variabilidade moderada a alta nas notas, com Matemática apresentando a maior amplitude de variação.

Gráfico 3: BoxPlot referente à distribuição estatística das notas do Dataset Microdados do Enem Nordeste.



O código para gerar o gráfico também utilizou a biblioteca ‘ggplot2’. Inicialmente, os dados das notas (NU_NOTA_CH, NU_NOTA_CN, NU_NOTA_LC, NU_NOTA_MT) foram reorganizados com a função ‘pivot_longer()’ do pacote ‘tidyr’, transformando as colunas de cada área em uma única coluna chamada Prova e os valores das notas em outra chamada Nota. Em seguida, o comando ‘ggplot()’ foi utilizado para definir o conjunto de dados e os mapeamentos estéticos (aes), atribuindo Area ao eixo X e Nota ao eixo Y. A função ‘geom_boxplot()’ construiu os boxplots, exibindo mediana, quartis e valores extremos. A função ‘labs()’ adicionou o título e os rótulos dos eixos, e novamente uma aplicação limpa e sem grades utilizando o tema ‘theme_minimal()’. Por fim, ‘geom_boxplot()’ automaticamente tratou outliers como pontos individuais, o que facilitou a interpretação da dispersão das notas.

Uma análise de correlação (Tabela 3) entre as notas revelou padrões interessantes no desempenho dos participantes. Observou-se uma forte correlação entre as áreas de Linguagens e Códigos e Ciências Humanas (corr = 0,75), indicando que estudantes com bom desempenho em uma dessas áreas tendem a ter resultados igualmente elevados na outra, possivelmente devido a competências comuns, como interpretação e análise crítica de textos.

Também foi identificada uma correlação relevante entre Matemática e Ciências da Natureza (corr = 0,68), sugerindo que habilidades lógico-quantitativas e de raciocínio científico estão associadas. Já a nota da Redação apresentou baixa correlação com as demais áreas, reforçando seu caráter mais independente e possivelmente relacionado a competências específicas de produção textual.

Tabela 3: Correlação entre as notas das áreas abordadas no Enem.

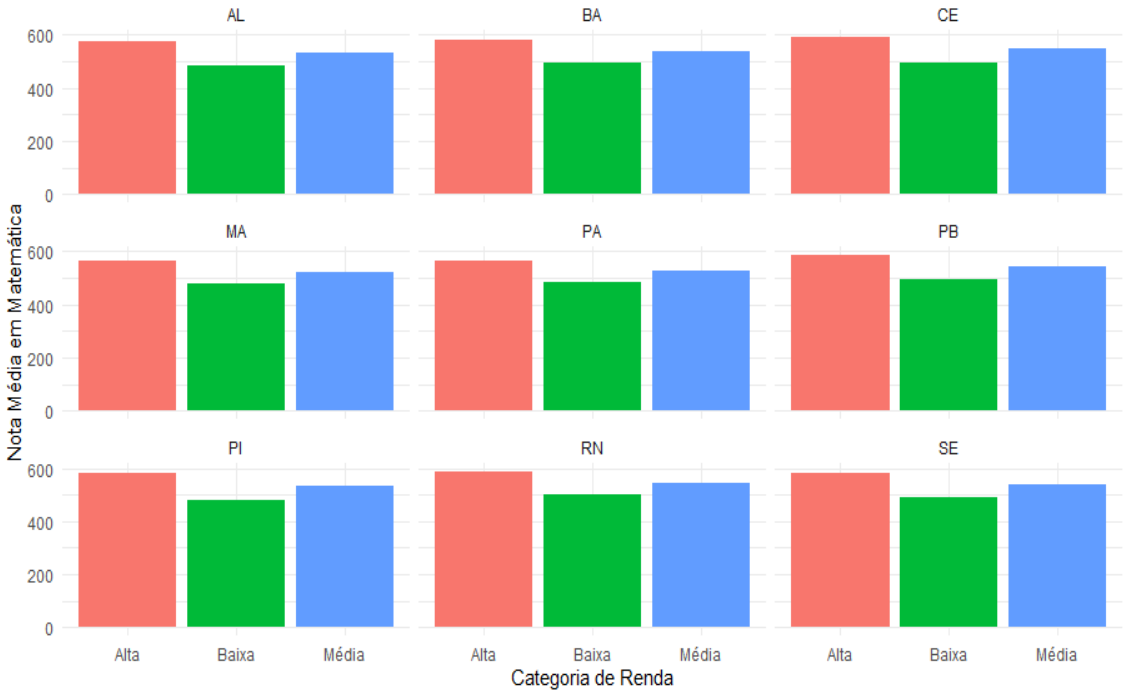
	NU_NOTA_CN	NU_NOTA_CH	NU_NOTA_LC	NU_NOTA_MT	NU_NOTA_REDACAO
NU_NOTA_CN	1.00	0.58	0.57	0.68	0.42
NU_NOTA_CH	0.58	1.00	0.75	0.61	0.48
NU_NOTA_LC	0.57	0.75	1.00	0.60	0.49
NU_NOTA_MT	0.68	0.61	0.60	1.00	0.49
NU_NOTA_REDACAO	0.42	0.48	0.49	0.49	1.00

Para realizar a análise de correlação, foi acionada a função select() do pacote dplyr, com o objetivo de selecionar apenas as colunas de interesse do conjunto de dados (NU_NOTA_LC, NU_NOTA_CH, NU_NOTA_MT, NU_NOTA_CN e NU_NOTA_REDACAO). Em seguida, aplicou-se a função cor() visando calcular a matriz de correlação entre as variáveis numéricas, utilizando por padrão o coeficiente de correlação de Pearson.

Para facilitar a interpretação dos resultados e apresentar valores mais claros, foi utilizado o comando round(2), que arredonda todos os coeficientes para duas casas decimais. Dessa forma, foi possível identificar de forma precisa o grau de associação linear entre as diferentes áreas de conhecimento avaliadas.

Para avaliar a relação entre rendimento familiar e desempenho em Matemática, foi calculada a média das notas de NU_NOTA_MT agrupando os participantes de acordo com as faixas de renda informadas na variável Q006. Foi percebido (Gráfico 4) que resultados evidenciaram um padrão consistente onde os estudantes pertencentes a faixas de maior renda apresentaram médias de Matemática superiores às daqueles com menor renda em todos os estados analisados, sugerindo uma possível associação positiva entre condição socioeconômica e desempenho nessa área do exame.

Gráfico 4: Nota média de Matemática por faixa de renda e estado do Dataset Microdados do Enem Nordeste.



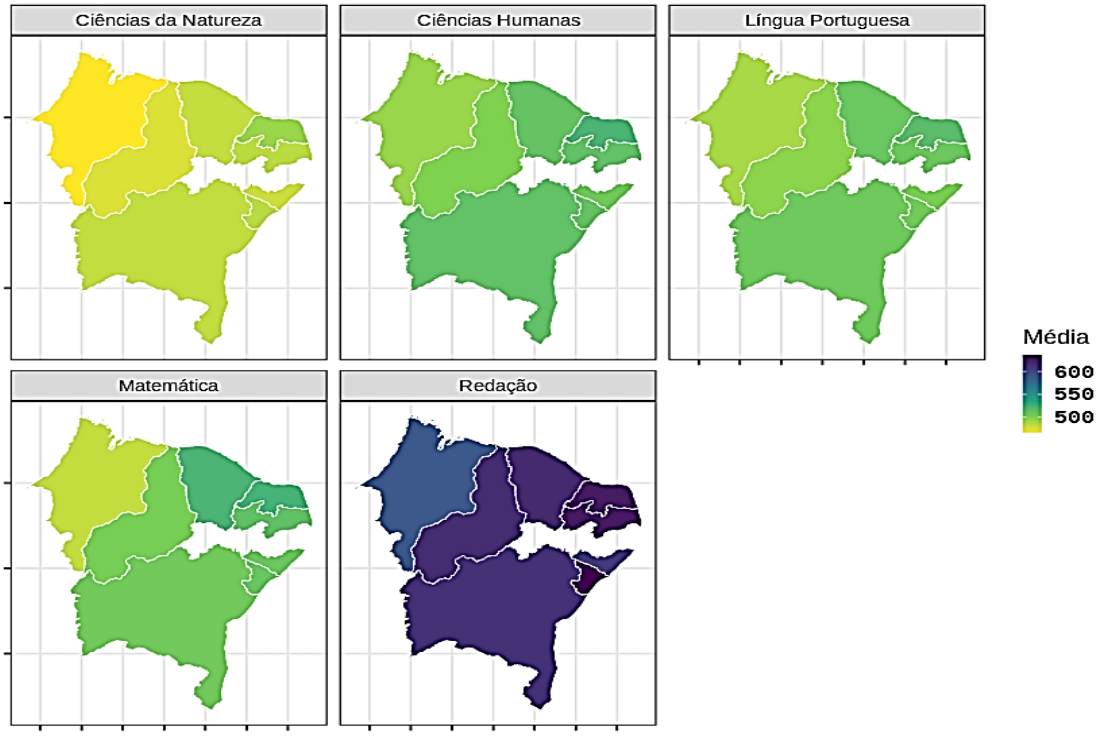
Para o desenvolvimento dessa análise, foi utilizada a função ‘unique()’ para identificar todos os valores distintos presentes na coluna Q006, que representa a faixa de renda familiar dos participantes. Em seguida, a função ‘mutate()’ para criar uma nova coluna denominada renda_cat, na qual as faixas de renda foram categorizadas em três grupos: "Baixa", "Média" e "Alta". Os dados, então, foram agrupados por estado (SG_UF_PROVA) e pela nova categoria de renda (renda_cat), utilizando funções de agrupamento, e calculou-se a média das notas de Matemática (NU_NOTA_MT) para cada grupo formado.

Para a visualização, foi construído um gráfico de barras, em que o eixo X representa as categorias de renda e o eixo Y apresenta a média das notas de Matemática. O argumento ‘fill’ foi utilizado para definir a cor de preenchimento das barras conforme a faixa de renda. A função ‘facet_wrap()’ dividiu o gráfico em painéis, permitindo visualizar separadamente os resultados de cada estado do Nordeste. Por fim, a função ‘theme_minimal()’ foi aplicada para simplificar e tornar a aparência do gráfico mais limpa e objetiva.

Realizando um apanhado geral sobre as médias dos participantes do Enem, o Gráfico 5 proporciona a visualização das médias por meio de mapas, onde cada município é colorido de acordo com sua pontuação, seguindo uma escala de cores que varia do amarelo (menor média) ao roxo escuro (maior média).

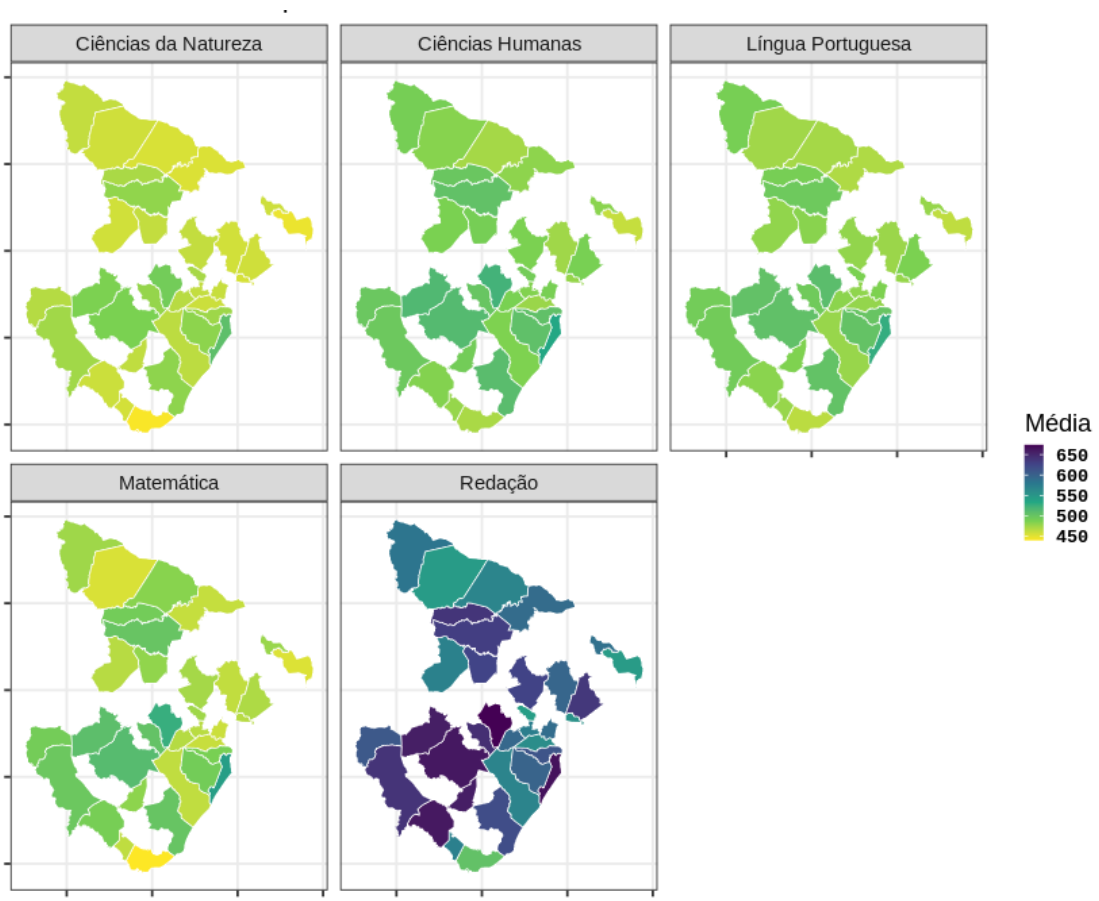
É possível notar que, nas disciplinas de Ciências da Natureza, Ciências Humanas, Linguagem e Matemática, as médias são relativamente homogêneas entre os estados. No entanto, na disciplina de Redação, o estado de Sergipe se destaca com uma das maiores médias da região, representado pela tonalidade mais escura no mapa, sugerindo um desempenho superior nesta área de conhecimento em comparação com os demais estados nordestinos. Vale a pena ressaltar que o estado de Pernambuco não teve seus dados exibidos devido à ausência de informações no conjunto de dados utilizado para a análise.

Gráfico 5: Mapa de calor referente às médias das notas de diferentes disciplinas nos estados do Dataset Microdados do Enem Nordeste.



Já no Gráfico 6, que ilustra o mapa de Sergipe, é possível observar que as maiores médias, especialmente na disciplina de Redação, tendem a se concentrar na região litorânea e na capital, Aracaju, indicadas pelas cores mais escuras nesses mapas. Isso sugere uma possível disparidade no desempenho entre os municípios do estado.

Gráfico 6: Mapa de calor referente às médias das notas de diferentes disciplinas do Dataset Microdados do Enem Nordeste no estado de Sergipe.



A aplicação da Ciência de Dados, aliada ao conhecimento em estatística e à utilização da linguagem R, oferece um poderoso conjunto de ferramentas para compreender padrões, identificar correlações e apoiar estratégias baseadas em evidências, potencializando a eficiência e a precisão das análises educacionais. Ressalta-se que ainda há potencial para a extração de outros insights relevantes para o dataset avaliado, direcionando outras necessidades informativas para tomadas de decisões específicas.

CÓDIGOS UTILIZADOS

```
#Bibliotecas utilizadas
library(dplyr)
library(writexl)
library(ggplot2)
library(tidyr)
library(geobr)
library(tidyverse)

#Avaliação dos tipos de dados / gerar função para descrever resultados
enem <- microdados_enem_nordeste_1_
str(microdados_enem_nordeste_1_)
classificar_variavel <- function(x) {
  if (is.numeric(x) || is.integer(x)) {
    return("Quantitativa")
  } else if (is.character(x) || is.factor(x)) {
    return("Qualitativa")
  } else {
    return("Outro tipo")
  }
}
tabela_variaveis <- data.frame(
  Variavel = names(enem),
  Tipo = sapply(enem, classificar_variavel)
)
write_xlsx(tabela_variaveis, "tabela_variaveis.xlsx")

#Contando a quantidade de variáveis quantitativas e qualitativas
contagem_tipos <- table(tabela_tipos$Tipo)
contagem_tipos_variaveis <- as.data.frame(contagem_tipos)
colnames(contagem_tipos_variaveis) <- c("Tipo", "Quantidade")
write_xlsx(contagem_tipos_variaveis, "contagem_tipos_variaveis.xlsx")
```

```
#criando um gráfico de pizza para visualização do percentual de variáveis quantitativas e qualitativas
contagem_tipos_variaveis$Percentual <- round(100 * contagem_tipos_variaveis$Quantidade / sum(contagem_tipos_variaveis$Quantidade), 1)
contagem_tipos_variaveis$Label <- paste0 (contagem_tipos_variaveis$Percentual, " %")
ggplot(contagem_tipos_variaveis, aes(x = "", y = Quantidade, fill = Tipo)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  theme_void() +
  geom_text(aes(label = Label),
            position = position_stack(vjust = 0.5),
            color = "white",
            size = 6) +
  ggtitle("Distribuição de Variáveis por Tipo") +
  scale_fill_manual(values = c("cyan3", "coral"))

#Avaliação da frequência da cor/raça entre os participantes
ggplot(enem, aes(x = as.factor(TP_COR_RACA))) +
  geom_bar(fill = "skyblue") +
  geom_text(stat = "count",
            aes(label = after_stat(count)),
            vjust = -0.5,
            size = 4) +
  labs(title = "Frequência da Variável TP_COR_RACA",
       x = "Código da Cor/Raça",
       y = "Frequência") +
  theme_minimal() +
  theme(axis.text.y = element_blank(),
      axis.ticks.y = element_blank())

#Avaliação estatística das notas
```

```
# Selecionar apenas as colunas desejadas
notas <- microdados_enem_nordeste_1_[, c("NU_NOTA_CN", "NU_NOTA_CH", "NU_NOTA_LC", "NU_NOTA_MT")] %>%
drop_na()  #acrescentado devido a ausência de algumas notas de candidatos faltantes


# Converter de formato largo para longo (necessário para ggplot)
notas_long <- notas %>%
  pivot_longer(cols = everything(),
               names_to = "Prova",
               values_to = "Nota")


# Criar boxplots comparativos
ggplot(notas_long, aes(x = Prova, y = Nota, fill = Prova)) +
  geom_boxplot() +
  labs(title = "Distribuição das Notas por Área do ENEM",
       x = "Área da Prova",
       y = "Nota") +
  theme_minimal() +
  theme(legend.position = "none")
```

#selecionando as colunas de notas no dataset

```
#selecionando as colunas de notas no dataset
notas <- dataset_enem %>%
  select(NU_NOTA_CN, NU_NOTA_CH, NU_NOTA_LC, NU_NOTA_MT, NU_NOTA_REDACAO)

#calculado a matriz de correlação das variaveis , use=complete.obs para retirar valores 'NA'
correlacao_notas <- cor(notas, use= 'complete.obs') %>% round(2)
```

#Avaliação da média da nota de matemática por renda de cada estado

```r
library(dplyr)
library(ggplot2)

dataset_enem %>%
  mutate(renda_cat = case_when(
    Q006 %in% c(
      "Nenhuma Renda",
      "Ate R$ 1.320,00",
      "De R$ 1.320,01 ate R$ 1.980,00."
    ) ~ "Baixa",

    Q006 %in% c(
      "De R$ 1.980,01 ate R$ 2.640,00.",
      "De R$ 2.640,01 ate R$ 3.300,00.",
      "De R$ 3.300,01 ate R$ 3.960,00.",
      "De R$ 3.960,01 ate R$ 5.280,00."
    ) ~ "Média",

    Q006 %in% c(
      "De R$ 5.280,01 ate R$ 6.600,00.",
      "De R$ 6.600,01 ate R$ 7.920,00.",
      "De R$ 7.920,01 ate R$ 9240,00.",
      "De R$ 9.240,01 ate R$ 10.560,00.",
      "De R$ 10.560,01 ate R$ 11.880,00.",
      "De R$ 11.880,01 ate R$ 13.200,00.",
      "De R$ 13.200,01 ate R$ 15.840,00.",
      "De R$ 15.840,01 ate R$19.800,00.",
      "De R$ 19.800,01 ate R$ 26.400,00.",
      "Acima de R$ 26.400,00."
    ) ~ "Alta",

    TRUE ~ NA_character_
  )) %>%
  group_by(SG_UF_PROVA, renda_cat) %>%
  summarise(media_mt = mean(NU_NOTA_MT, na.rm = TRUE), .groups = "drop") %>%
  ggplot(aes(x = renda_cat, y = media_mt, fill = renda_cat)) +
  geom_col() +
  facet_wrap(~ SG_UF_PROVA) +
  labs(
    title = "Nota média de Matemática por faixa de renda e estado",
    x = "Categoria de Renda",
    y = "Nota Média em Matemática"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```

#Análise das médias das notas com visualização em mapas de calor

Carregar dataframe ----------

ENEM_2023 <- read.csv2("microdados_enem_nordeste.csv")

Visualizar Dados ----------

Visualizar primeiros registros do dataframe

head(ENEM_2023)

Visualizar resumo dados

glimpse(ENEM_2023)

Verificar estados

distinct(ENEM_2023,SG_UF_PROVA)

OBSERVACAO: Verificamos que não consta no dataframe os dados do estado

#de PE - Pernambuco, consta o estado PA - Pará

Médias das disciplinas por estados do Nordeste

```
### Calcular as médias de cada disciplina para cada estado
medias_por_estado <- ENEM_2023 %>%
  group_by(CO_UF_PROVA) %>%
  summarise(media_cn = mean(NU_NOTA_CN, na.rm = TRUE),
            media_ch = mean(NU_NOTA_CH, na.rm = TRUE),
            media_lc = mean(NU_NOTA_LC, na.rm = TRUE),
            media_mt = mean(NU_NOTA_MT, na.rm = TRUE),
            media_re = mean(NU_NOTA_REDACAO, na.rm = TRUE))
### Visualizar médias por estados NE
print(medias_por_estado)
### Pivotar tabela - arrumar para exibir no gráfico
medias_tabela <- medias_por_estado %>%
  pivot_longer(cols = c(media_cn,media_ch,media_lc,media_mt,media_re),
               names_to = "disciplina",
               values_to = "medias")
### Obter os dados geográficos dos estados
estados <- read_state(code_state = "all", year = 2020)
### Filtrar estados do nordeste
estados_NE <- estados %>%
  filter(code_region == "2")
### Renomear coluna para facilitar união
estados_NE <- estados_NE %>%
  rename(CO_UF_PROVA = code_state)
### Unir os dados das médias com os dados geográficos dos estados
medias_geodata <- estados_NE %>%
  inner_join(medias_tabela, by = "CO_UF_PROVA")
### Renomear valores para exibição no gráfico
medias_geodata$disciplina[medias_geodata$disciplina == "media_cn"] <- "Ciências da
Natureza"
```

```
medias_geodata$disciplina[medias_geodata$disciplina == "media_ch"] <- "Ciências Humanas"
medias_geodata$disciplina[medias_geodata$disciplina == "media_lc"] <- "Língua Portuguesa"
medias_geodata$disciplina[medias_geodata$disciplina == "media_mt"] <- "Matemática"
medias_geodata$disciplina[medias_geodata$disciplina == "media_re"] <- "Redação"

# Gráfico mapa médias disciplinas estados do nordeste
### Criar mapa dos estados do Nordeste com ggplot2
### Observacao - não exibindo Pernambuco por falta de dados
ggplot(data = medias_geodata) +
  geom_sf(aes(fill = medias), color = "white") +
  facet_wrap( ~ disciplina) +
  #alterar a letra no parâmetro option para mudar as cores
  scale_fill_viridis_c(option = "D", direction = -1) +
  labs(title = "Médias ENEM - 2023 - Nordeste",
       subtitle = "Média das notas de diferentes disciplinas por estado.\nCores mais escuras representam médias maiores.",
       caption = "Fonte: INEP - Microdados Enem 2023.\nEstado de Pernambuco não exibido por falta de dados.",
       fill = "Média") +
  theme_bw() +
  #formatação dos campos
  theme(
    legend.text = element_text(family = "Ubuntu", face = "bold", color = "black", size = 10),
    legend.key.size = unit(0.3, "cm"),
    legend.position = "right",
    plot.title = element_text(family = "Ubuntu", face = "bold", size = 20),
```

```
    plot.subtitle = element_text(size = 12),
    plot.margin = margin(t = 20, r = 8, b = 7, l = 5),
    axis.text.x = element_blank(),
    axis.text.y = element_blank(),
    painel.grid = element_blank()
  )
#-------------------------------------------------------------------------------
## Médias das disciplinas por municípios de Sergipe
### Criar novo dataframe só com os dados de Sergipe
ENEM_2023_SE <- ENEM_2023 %>%
  filter(SG_UF_PROVA == "SE")
### Visualizar resumo dados
glimpse(ENEM_2023_SE)
### Calcular as médias de cada disciplina para cada município de Sergipe
medias_por_municipio <- ENEM_2023_SE %>%
  group_by(CO_MUNICIPIO_PROVA) %>%
  summarise(media_cn = mean(NU_NOTA_CN, na.rm = TRUE),
            media_ch = mean(NU_NOTA_CH, na.rm = TRUE),
            media_lc = mean(NU_NOTA_LC, na.rm = TRUE),
            media_mt = mean(NU_NOTA_MT, na.rm = TRUE),
            media_re = mean(NU_NOTA_REDACAO, na.rm = TRUE))
### Pivotar tabela - arrumar para exibir no gráfico
medias_tabela_SE <- medias_por_municipio %>%
  pivot_longer(cols = c(media_cn,media_ch,media_lc,media_mt,media_re),
               names_to = "disciplina",
               values_to = "medias")
### Obter os dados geográficos dos 75 municípios de Sergipe
municipios <- read_municipality("SE", year = 2020)
### Visualizar resumo dados
```

```
glimpse(municipios)
### Exibir nome dos municipios
distinct(municipios,name_muni)
### Converter para inteiro para permitir união
municipios$code_muni = as.integer(municipios$code_muni)
### Renomear coluna para falicitar união
municipios <- municipios %>%
  rename(CO_MUNICIPIO_PROVA = code_muni)
### Unir os dados das médias com os dados geográficos dos municipios
medias_geodata_SE <- municipios %>%
  inner_join(medias_tabela_SE, by = "CO_MUNICIPIO_PROVA")
### Renomear valores para melhor exibição no gráfico
medias_geodata_SE$disciplina[medias_geodata_SE$disciplina == "media_cn"] <-
"Ciências da Natureza"
medias_geodata_SE$disciplina[medias_geodata_SE$disciplina == "media_ch"] <-
"Ciências Humanas"
medias_geodata_SE$disciplina[medias_geodata_SE$disciplina == "media_lc"] <-
"Língua Portuguesa"
medias_geodata_SE$disciplina[medias_geodata_SE$disciplina == "media_mt"] <-
"Matemática"
medias_geodata_SE$disciplina[medias_geodata_SE$disciplina == "media_re"] <-
"Redação"

# Gráfico mapa médias disciplinas municipios de Sergipe
### Observacao - exibindo apenas municípios onde ocorreu prova
ggplot(data = medias_geodata_SE) +
  geom_sf(aes(fill = medias), color = "white") +
  facet_wrap( ~ disciplina) +
  #alterar a letra no parâmetro option para mudar as cores
  scale_fill_viridis_c(option = "D", direction = -1) +
```

```
  labs(title = "Médias ENEM - 2023\nMunicípios de Sergipe",
       subtitle = "Média das notas de diferentes disciplinas por municípios de
Sergipe.\nCores mais escuras representam médias maiores.",
       caption = "Fonte: INEP - Microdados Enem 2023.",
       fill = "Média") +
  theme_bw() +
  #formatação dos campos
  theme(
    legend.text = element_text(family = "Ubuntu", face = "bold", color = "black", size =
8),
    legend.key.size = unit(0.3, "cm"),
    legend.position = "right",
    plot.title = element_text(family = "Ubuntu", face = "bold", size = 20),
    plot.subtitle = element_text(size = 12),
    plot.margin = margin(t = 20, r = 8, b = 7, l = 5),
    axis.text.x = element_blank(),
    axis.text.y = element_blank(),
    painel.grid = element_blank()
  )
```