

# Análise Exploratória de Dados - AcemogluGlobalData

Cleyton Fernandes

12 de abril de 2025

## Introdução

Nessa disciplina, aprofundamos nossos conhecimentos na linguagem R e em estatística para realizar análises descritivas de bases de dados, uma tarefa essencial para o dia-a-dia de um cientista de dados. Este relatório apresenta uma análise exploratória da base de dados AcemogluGlobalData, utilizando o RMarkdown para documentar o processo e os resultados.

A base de dados foi disponibilizada pelo professor, a meu pedido, pois não consegui localizar uma base adequada no Kaggle que atendesse aos requisitos do projeto. A análise será realizada com o objetivo de explorar as variáveis, identificar padrões, tratar dados faltantes e criar um dashboard interativo com Shiny. Todos os códigos serão exibidos no relatório, conforme solicitado.

Os códigos e o relatório final estão disponíveis no repositório GitHub: <https://github.com/CleytonFernandes/Analiseexploratoria2025>.

## Escolha da Base de Dados

Escolhi a base de dados `AcemogluGlobalData.csv`, que foi disponibilizada pelo professor a meu pedido, pois não consegui localizar uma base adequada no Kaggle que atendesse aos requisitos do projeto. A base contém dados de diversos países ao longo de vários anos, com variáveis relacionadas a democracia, economia e demografia. As variáveis de interesse são:

- `dem_ind`: Índice de democracia (numérico, entre 0 e 1).
- `log_gdppc`: Logaritmo do PIB per capita (numérico).
- `log_pop`: Logaritmo da população (numérico).
- `age_1` a `age_5`: Proporções de faixas etárias (0-14, 15-29, 30-44, 45-59, 60+, todas numéricas).
- `educ`: Nível de educação (numérico).
- `age_median`: Idade mediana da população (numérico).

A base contém 13 variáveis no total, sendo 10 delas numéricas, o que atende ao requisito de ter pelo menos 4 variáveis numéricas. Além disso, a base possui dados faltantes em várias variáveis (como `log_gdppc` e `log_pop`), o que é necessário para a análise de completude e imputação de dados. A escolha dessa base é relevante porque permite explorar relações entre democracia, desenvolvimento econômico e características demográficas, temas importantes na atualidade.

## Objetivo da Análise e Resultados Esperados

O objetivo desta análise exploratória é entender as características da base `AcemogluGlobalData`, identificar padrões e relações entre as variáveis, verificar se as variáveis seguem uma distribuição normal, tratar dados faltantes e criar um dashboard interativo com Shiny. Especificamente, pretendo:

- Calcular estatísticas descritivas para entender a distribuição das variáveis.
- Identificar correlações entre variáveis, como a relação entre o índice de democracia e o PIB per capita.
- Verificar se as variáveis seguem uma distribuição normal, usando histogramas, gráficos Q-Q e testes estatísticos.
- Tratar dados faltantes para melhorar a qualidade da base.
- Criar um dashboard interativo que permita visualizar as variáveis de forma dinâmica.

Espero encontrar correlações significativas, como uma associação positiva entre o índice de democracia e o PIB per capita, e identificar quais variáveis se aproximam de uma distribuição normal. Além disso, espero que o dashboard facilite a exploração visual dos dados.

## Carregamento dos Dados

Carrego a base de dados, ajusto os tipos de dados e renomeio as variáveis para nomes mais descritivos.

```
dados <- read.csv("C:/Users/cleyt/OneDrive/Área de Trabalho/MBA em Data Science INFNET/Análise explorat
                sep = ";", dec = ",", stringsAsFactors = FALSE, encoding = "UTF-8")

# Converter colunas para numérico, primeiro para character para evitar problemas com fatores
dados$dem_ind <- as.numeric(as.character(dados$dem_ind))
dados$log_gdppc <- as.numeric(as.character(dados$log_gdppc))
dados$log_pop <- as.numeric(as.character(dados$log_pop))
dados$age_median <- as.numeric(as.character(dados$age_median))

# Renomear as variáveis
names(dados) <- c("País", "Ano", "Índice de Democracia", "Log do PIB per Capita", "Log da População",
                  "Proporção Idade 0-14", "Proporção Idade 15-29", "Proporção Idade 30-44",
                  "Proporção Idade 45-59", "Proporção Idade 60+", "Nível de Educação",
                  "Idade Mediana", "Código do País")

head(dados)
```

```
##      País  Ano Índice de Democracia Log do PIB per Capita Log da População
## 1 Andorra 1960                NA                NA                NA
## 2 Andorra 1965                NA                NA                NA
## 3 Andorra 1970                0.5                NA                NA
## 4 Andorra 1975                NA                NA                NA
## 5 Andorra 1980                NA                NA                NA
## 6 Andorra 1985                NA                NA                NA
##  Proporção Idade 0-14  Proporção Idade 15-29  Proporção Idade 30-44
## 1                NA                NA                NA
## 2                NA                NA                NA
## 3                NA                NA                NA
## 4                NA                NA                NA
## 5                NA                NA                NA
## 6                NA                NA                NA
##  Proporção Idade 45-59  Proporção Idade 60+  Nível de Educação  Idade Mediana
## 1                NA                NA                NA                NA
## 2                NA                NA                NA                NA
## 3                NA                NA                NA                NA
## 4                NA                NA                NA                NA
## 5                NA                NA                NA                NA
```

```
## 6          NA          NA          NA          NA
##  Código do País
## 1          1
## 2          1
## 3          1
## 4          1
## 5          1
## 6          1
```

## Matriz de Espalhamento

Crio uma matriz de espalhamento para visualizar correlações entre as variáveis Índice de Democracia, Log do PIB per Capita, Log da População e Idade Mediana.

```
pairs(dados[, c("Índice de Democracia", "Log do PIB per Capita", "Log da População", "Idade Mediana")],
main = "Matriz de Espalhamento")
```

Através da inspeção visual, as variáveis Log do PIB per Capita e Idade Mediana parecem mais correlacionadas, pois os pontos formam uma linha ascendente, indicando uma correlação positiva. Isso sugere que países com maior PIB per capita tendem a ter uma idade mediana mais alta, possivelmente devido a melhores condições de vida e maior expectativa de vida.

## Estatísticas Descritivas

Obtenho as estatísticas descritivas das variáveis numéricas usando a função `descr()` do pacote `summarytools`.

```
# Selecionar apenas colunas numéricas
dados_numericos <- dados[, c("Índice de Democracia", "Log do PIB per Capita", "Log da População",
                             "Proporção Idade 0-14", "Proporção Idade 15-29", "Proporção Idade 30-44",
                             "Proporção Idade 45-59", "Proporção Idade 60+", "Nível de Educação",
                             "Idade Mediana")]

descr(dados_numericos)
```

```
## Descriptive Statistics
## dados_numericos
## N: 1369
##
##          Idade Mediana  Índice de Democracia  Log da População  Log do PIB per Capita
## -----
##          Mean          22.40                  0.50                  8.67                  8.16
##          Std.Dev        6.47                  0.37                  1.87                  1.02
##          Min           14.40                  0.00                  3.71                  5.77
##          Q1            17.50                  0.17                  7.66                  7.29
##          Median         19.30                  0.50                  8.75                  8.14
##          Q3            27.30                  0.83                  9.83                  8.97
##          Max           39.70                  1.00                 14.00                 10.45
##          MAD            3.71                  0.49                  1.61                  1.25
##          IQR            9.80                  0.67                  2.18                  1.68
##          CV             0.29                  0.74                  0.22                  0.13
##          Skewness        0.96                  0.10                 -0.14                  0.06
##          SE.Skewness      0.07                  0.07                  0.07                  0.08
##          Kurtosis       -0.49                 -1.52                  0.05                 -0.99
```

```

##          N.Valid          1214.00          1266.00          1202.00          966.00
##          N          1369.00          1369.00          1369.00          1369.00
##          Pct.Valid          88.68          92.48          87.80          70.56
##
## Table: Table continues below
##
##
##
##          Nível de Educação    Proporção Idade 0-14    Proporção Idade 15-29
## -----
##          Mean          4.43          0.37          0.26
##          Std.Dev          2.86          0.09          0.03
##          Min          0.04          0.15          0.19
##          Q1          2.02          0.29          0.24
##          Median          4.00          0.41          0.26
##          Q3          6.62          0.45          0.27
##          Max          12.18          0.52          0.36
##          MAD          3.24          0.08          0.02
##          IQR          4.59          0.16          0.03
##          CV          0.65          0.25          0.10
##          Skewness          0.47          -0.66          -0.06
##          SE.Skewness          0.09          0.07          0.07
##          Kurtosis          -0.73          -0.97          0.41
##          N.Valid          780.00          1268.00          1268.00
##          N          1369.00          1369.00          1369.00
##          Pct.Valid          56.98          92.62          92.62
##
## Table: Table continues below
##
##
##
##          Proporção Idade 30-44    Proporção Idade 45-59    Proporção Idade 60+
## -----
##          Mean          0.17          0.11          0.08
##          Std.Dev          0.03          0.04          0.05
##          Min          0.09          0.06          0.02
##          Q1          0.15          0.09          0.05
##          Median          0.17          0.10          0.06
##          Q3          0.19          0.15          0.12
##          Max          0.36          0.22          0.24
##          MAD          0.03          0.02          0.02
##          IQR          0.04          0.06          0.07
##          CV          0.18          0.32          0.59
##          Skewness          1.21          0.89          1.12
##          SE.Skewness          0.07          0.07          0.07
##          Kurtosis          2.93          -0.49          -0.03
##          N.Valid          1268.00          1268.00          1268.00
##          N          1369.00          1369.00          1369.00
##          Pct.Valid          92.62          92.62          92.62

```

A tabela acima mostra as estatísticas descritivas, como média, desvio padrão, mínimo e máximo, para cada variável numérica. Por exemplo, podemos observar que o Índice de Democracia varia de 0 a 1, com uma média que indica o nível médio de democracia nos países da base.

## Explicação da Distribuição Normal

A distribuição normal, que também é chamada de curva de Gauss, é uma distribuição de probabilidade que tem o formato de um sino. Aprendi que ela é muito importante em estatística porque muitos dados na natureza seguem esse padrão. Ela é definida pela média, que é o valor central, e pelo desvio padrão, que mostra o quanto os dados estão espalhados. A maior parte dos dados (cerca de 68%) fica a 1 desvio padrão da média, 95% a 2 desvios padrão, e 99,7% a 3 desvios padrão. Além disso, a curva é simétrica, ou seja, os dados se distribuem igualmente dos dois lados da média.

## Histogramas

Crio histogramas para visualizar a distribuição de todas as variáveis numéricas usando o pacote `ggplot2`. Escolhi um `binwidth` inicial de 0,1 para variáveis como `Índice de Democracia`, que varia de 0 a 1, mas uso `scales = "free"` para ajustar automaticamente os intervalos e escalas para variáveis com ranges diferentes, como `Log do PIB per Capita` e `Idade Mediana`.

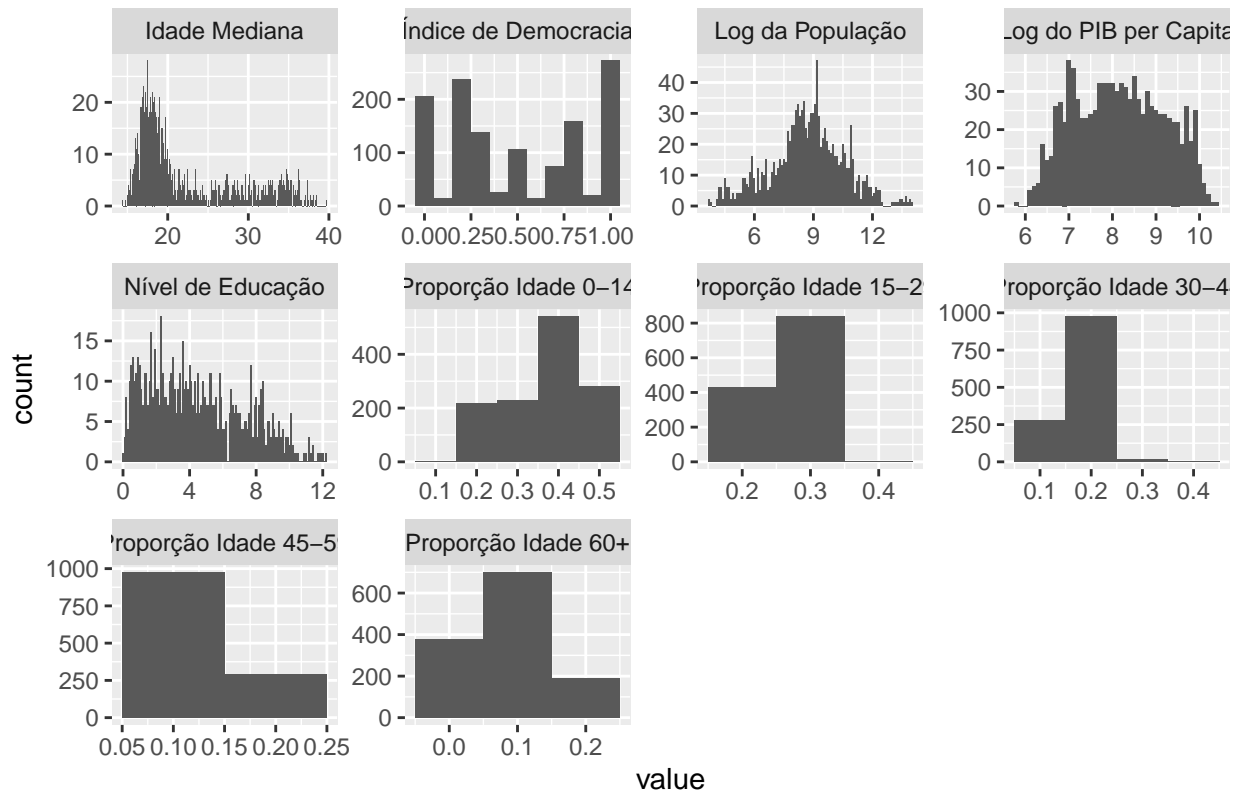
```
# Carregar o pacote tidyr para transformar os dados
library(tidyr)

# Converter o dataframe numérico para o formato longo (long format)
dados_numericos_long <- pivot_longer(dados_numericos, cols = everything(), names_to = "variable", values_to = "value")

# Criar histogramas para todas as variáveis numéricas
ggplot(dados_numericos_long, aes(x = value)) +
  geom_histogram(binwidth = 0.1) +
  facet_wrap(~variable, scales = "free") +
  labs(title = "Histogramas das Variáveis Numéricas")

## Warning: Removed 1922 rows containing non-finite outside the scale range
## ('stat_bin()').
```

## Histogramas das Variáveis Numéricas



Os histogramas mostram a distribuição de cada variável. Por exemplo, o histograma de Índice de Democracia parece assimétrico, com muitos valores próximos de 0 ou 1, enquanto Idade Mediana tem um formato mais próximo de uma curva em sino.

## Gráficos Q-Q

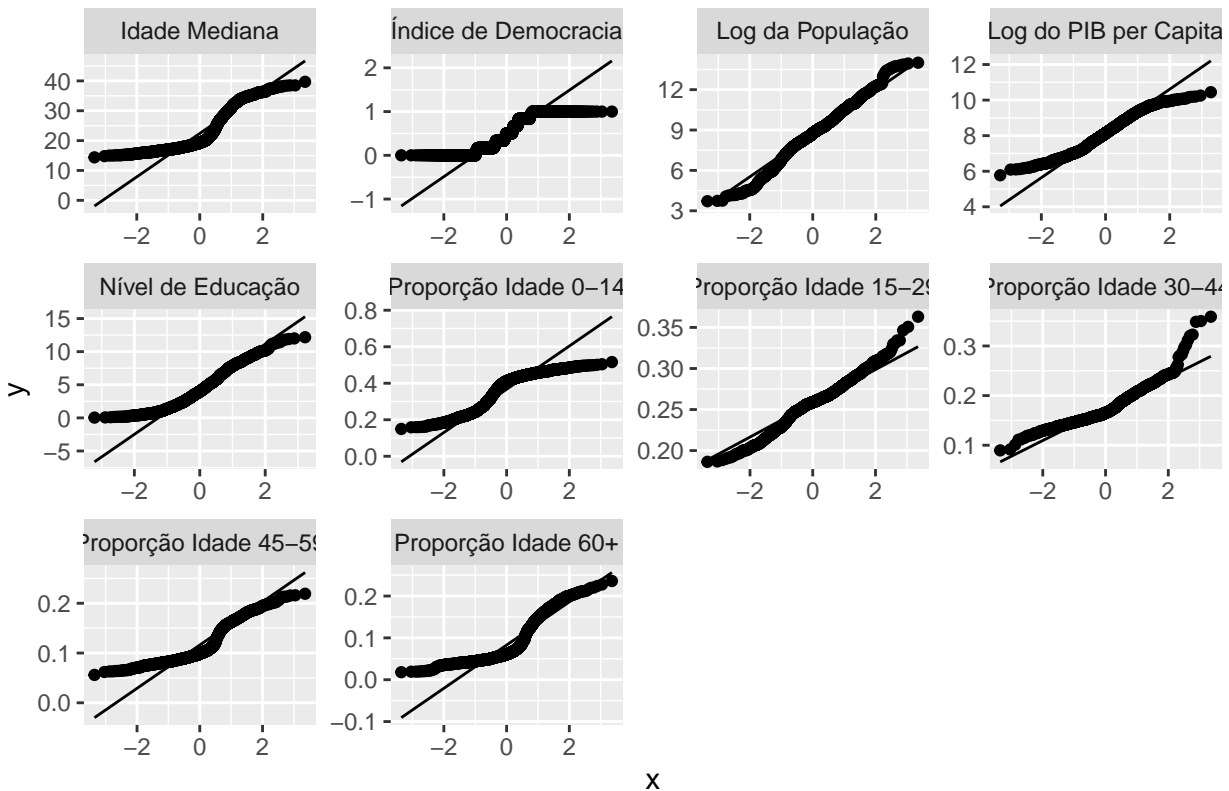
Crio gráficos Q-Q para avaliar a normalidade das variáveis numéricas usando o pacote `ggpubr`.

```
# Criar gráficos Q-Q para todas as variáveis numéricas
ggplot(dados_numericos_long, aes(sample = value)) +
  stat_qq() +
  stat_qq_line() +
  facet_wrap(~variable, scales = "free") +
  labs(title = "Gráficos Q-Q das Variáveis Numéricas")
```

```
## Warning: Removed 1922 rows containing non-finite outside the scale range
## ('stat_qq()').
```

```
## Warning: Removed 1922 rows containing non-finite outside the scale range
## ('stat_qq_line()').
```

## Gráficos Q–Q das Variáveis Numéricas



Os gráficos Q–Q mostram como os quantis das variáveis se comparam aos quantis de uma distribuição normal. Se os pontos seguem a linha diagonal, a variável é aproximadamente normal. Observo que variáveis como Índice de Democracia desviam bastante da linha, enquanto Idade Mediana parece mais próxima de uma distribuição normal.

## Teste de Normalidade

Realizo o teste de Shapiro-Wilk para verificar a normalidade das variáveis numéricas.

```
# Aplicar o teste de Shapiro-Wilk a cada variável numérica
for (var in names(dados_numericos)) {
  cat("Teste de Shapiro-Wilk para", var, "\n")
  # Remover NAs antes do teste
  valores <- na.omit(dados_numericos[[var]])
  # O teste Shapiro-Wilk só pode ser aplicado a amostras com 3 a 5000 observações
  if (length(valores) >= 3 & length(valores) <= 5000) {
    print(shapiro.test(valores))
  } else {
    cat("Amostra muito pequena ou muito grande para o teste Shapiro-Wilk.\n")
  }
  cat("\n")
}
```

```
## Teste de Shapiro-Wilk para Índice de Democracia
##
## Shapiro-Wilk normality test
```

```

##
## data:  valores
## W = 0.87365, p-value < 2.2e-16
##
##
## Teste de Shapiro-Wilk para Log do PIB per Capita
##
## Shapiro-Wilk normality test
##
## data:  valores
## W = 0.9753, p-value = 9.589e-12
##
##
## Teste de Shapiro-Wilk para Log da População
##
## Shapiro-Wilk normality test
##
## data:  valores
## W = 0.99183, p-value = 3.272e-06
##
##
## Teste de Shapiro-Wilk para Proporção Idade 0-14
##
## Shapiro-Wilk normality test
##
## data:  valores
## W = 0.88731, p-value < 2.2e-16
##
##
## Teste de Shapiro-Wilk para Proporção Idade 15-29
##
## Shapiro-Wilk normality test
##
## data:  valores
## W = 0.98759, p-value = 6.537e-09
##
##
## Teste de Shapiro-Wilk para Proporção Idade 30-44
##
## Shapiro-Wilk normality test
##
## data:  valores
## W = 0.9237, p-value < 2.2e-16
##
##
## Teste de Shapiro-Wilk para Proporção Idade 45-59
##
## Shapiro-Wilk normality test
##
## data:  valores
## W = 0.8682, p-value < 2.2e-16
##
##
## Teste de Shapiro-Wilk para Proporção Idade 60+

```



```
##
## Shapiro-Wilk normality test
##
## data:  valores
## W = 0.82573, p-value < 2.2e-16
##
##
## Teste de Shapiro-Wilk para Nível de Educação
##
## Shapiro-Wilk normality test
##
## data:  valores
## W = 0.95559, p-value = 1.327e-14
##
##
## Teste de Shapiro-Wilk para Idade Mediana
##
## Shapiro-Wilk normality test
##
## data:  valores
## W = 0.83498, p-value < 2.2e-16
```

Os resultados mostram o valor-p para cada variável. Se o valor-p for menor que 0,05, rejeitamos a hipótese de normalidade. Por exemplo, para Índice de Democracia, o valor-p é muito baixo, indicando que a variável não é normal.

## Conclusão sobre Normalidade

Com base nos histogramas, gráficos Q-Q e testes de Shapiro-Wilk, concluo que a maioria das variáveis não segue uma distribuição normal. Os histogramas de variáveis como **Índice de Democracia** e **Log da População** não apresentam formato de sino, os gráficos Q-Q mostram desvios significativos da linha diagonal para essas variáveis, e os valores-p dos testes de Shapiro-Wilk são menores que 0,05 para quase todas as variáveis, rejeitando a hipótese de normalidade. No entanto, a variável **Idade Mediana** parece mais próxima de uma distribuição normal, com um histograma mais simétrico e um gráfico Q-Q que segue mais de perto a linha diagonal, embora o teste de Shapiro-Wilk ainda indique um valor-p baixo.

## Completeness de Dados

Completeness de dados significa que todas as informações esperadas em uma base de dados estão presentes, ou seja, não há valores faltantes. Por exemplo, se uma variável como **Log do PIB per Capita** deveria ter um valor para cada país e ano, mas alguns estão em branco (NA), a base não está completa para essa variável. A completeness é importante porque dados faltantes podem prejudicar a análise, levando a resultados incorretos ou incompletos.

## Impacto dos Dados Faltantes

Dados faltantes podem ter vários impactos na análise exploratória. Primeiro, eles reduzem o número de observações disponíveis para análise, o que pode diminuir a precisão dos resultados. Por exemplo, se muitos países não têm valores para **Log do PIB per Capita**, não posso analisar a relação entre essa variável e o **Índice de Democracia** para esses países. Segundo, dados faltantes podem introduzir viés, especialmente se os valores ausentes não forem aleatórios (por exemplo, se países mais pobres tendem a não reportar o PIB).

Por fim, eles dificultam a interpretação de gráficos e testes estatísticos, como a matriz de espalhamento, que ignora linhas com valores ausentes.

## Índice de Completude

Calculo o índice de completude para cada variável numérica da base.

```
# Número total de observações
n_total <- nrow(dados_numericos)

# Calcular o número de valores ausentes por variável
valores_ausentes <- colSums(is.na(dados_numericos))

# Calcular o índice de completude (porcentagem de valores não ausentes)
completude <- (1 - valores_ausentes / n_total) * 100

# Exibir o resultado
data.frame(Variável = names(dados_numericos), Completude = completude)
```

##		Variável	Completude
##	Índice de Democracia	Índice de Democracia	92.47626
##	Log do PIB per Capita	Log do PIB per Capita	70.56245
##	Log da População	Log da População	87.80131
##	Proporção Idade 0-14	Proporção Idade 0-14	92.62235
##	Proporção Idade 15-29	Proporção Idade 15-29	92.62235
##	Proporção Idade 30-44	Proporção Idade 30-44	92.62235
##	Proporção Idade 45-59	Proporção Idade 45-59	92.62235
##	Proporção Idade 60+	Proporção Idade 60+	92.62235
##	Nível de Educação	Nível de Educação	56.97589
##	Idade Mediana	Idade Mediana	88.67787

A tabela acima mostra a porcentagem de completude para cada variável. Variáveis como Log do PIB per Capita e Log da População têm muitos valores ausentes, o que indica baixa completude.

## Imputação de Dados

Realizo a imputação de dados faltantes usando o pacote `mice` com o método PMM (Predictive Mean Matching).

```
# Imputação de dados com o pacote mice
library(mice)
dados_imputados <- mice(dados_numericos, m = 5, method = "pmm", maxit = 50, seed = 123)
```

```
##
## iter imp variable
## 1 1 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propor
## 1 2 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propor
## 1 3 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propor
## 1 4 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propor
## 1 5 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propor
## 2 1 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propor
```

[illegible]

[illegible]

[illegible]



```
## 45 3 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 45 4 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 45 5 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 46 1 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 46 2 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 46 3 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 46 4 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 46 5 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 47 1 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 47 2 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 47 3 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 47 4 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 47 5 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 48 1 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 48 2 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 48 3 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 48 4 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 48 5 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 49 1 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 49 2 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 49 3 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 49 4 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 49 5 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 50 1 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 50 2 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 50 3 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 50 4 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
## 50 5 Índice de Democracia Log do PIB per Capita Log da População Proporção Idade 0-14 Propo
```

```
## Warning: Number of logged events: 1235
```

```
dados_completos <- complete(dados_imputados)
```

```
# Verificar se ainda há valores ausentes
colSums(is.na(dados_completos))
```

```
## Índice de Democracia Log do PIB per Capita Log da População
## 0 0 0
## Proporção Idade 0-14 Proporção Idade 15-29 Proporção Idade 30-44
## 0 0 0
## Proporção Idade 45-59 Proporção Idade 60+ Nível de Educação
## 0 0 0
## Idade Mediana
## 0
```

Após a imputação, não há mais valores ausentes no dataframe `dados_completos`, que será usado nas próximas análises.

## Dashboard Shiny

Criei um dashboard interativo com o pacote Shiny, que permite selecionar uma variável, escolher a cor da linha e ajustar os limites dos eixos X e Y. Abaixo está um print da tela do dashboard:



Figure 1: Print do Dashboard Shiny

## Referências

1. Wikipédia - Distribuição Normal: [https://en.wikipedia.org/wiki/Normal\\_distribution](https://en.wikipedia.org/wiki/Normal_distribution)  
Utilizei esta página para entender melhor as características da distribuição normal e escrever a explicação na seção correspondente.
2. Repositório GitHub do Projeto: <https://github.com/CleytonFernandes/Analiseexploratoria2025>  
Todos os códigos do projeto, incluindo o arquivo RMarkdown (`relatorio.Rmd`) e o aplicativo Shiny (`app.R`), estão disponíveis neste repositório.
3. Documentação do Pacote MICE: <https://cran.r-project.org/web/packages/mice/mice.pdf>  
Consultei a documentação oficial do pacote MICE para entender como realizar a imputação de dados faltantes.
4. Documentação do Pacote Shiny: <https://shiny.rstudio.com/>  
Usei a documentação oficial do Shiny para aprender a criar o dashboard interativo.