LABORATÓRIO DE ESTUDOS E PESQUISAS EM EDUCAÇÃO E ECONOMIA SOCIAL

FRANCISCO RENATO GRACIANO FREIRE

DOCUMENTO DE PROTOCOLO DE LIMPEZA DE BASE DE DADOS

SOBRAL-CE

2024

Introdução

O objetivo deste documento é detalhar o protocolo de limpeza de dados a ser seguido para garantir a qualidade e a integridade dos dados utilizados na análise. A limpeza de dados é uma etapa crucial em qualquer projeto de análise, pois assegura que os dados sejam precisos, consistentes e confiáveis. Este protocolo descreve as etapas necessárias para identificar e corrigir problemas nos dados, como valores ausentes, erros de digitação, formatação inconsistente e duplicações.

A base de dados foi disponibilizada via planilha Google Docs e contém dados sobre as notas do Saeb aplicado aos alunos do ensino médio dos estados do Ceará e Pernambuco.

Fonte de dados

Base de dados: notas_saeb_2017_2023.xlsx - Planilhas Google

Descrição: A base de dados em questão é uma planilha contendo informações como município, código de município, unidade federativa, notas médias de Matemática, Português e Padronizada do Saeb dos municípios nos anos de 2017 e 2023.

Formato: .xlsx

Variáveis:

- sigla_uf,
- codigo_do_municipio,
- nome_do_municipio,
- rede,
- nota_matematica_2017,
- nota_lingua_portuguesa_2017,
- nota_media_padronizada_2017,
- nota_matematica_2019,
- nota_lingua_portuguesa_2019,
- nota_media_padronizada_2019,
- nota_matematica_2021,
- nota_lingua_portuguesa_2021,
- nota_media_padronizada_2021,
- nota_matematica_2023,
- nota_lingua_portuguesa_2023,
- nota_media_padronizada_2023

Qualidade de dados: completude avariada, afetada por dados ausentes.

Método de coleta: importação virtual online via rstudio utilizado o pacote googledrive

Procedimento de Limpeza

Importação dos dados:

```r
library(googlesheets4)
library(googledrive)
library(data.table)
library(readxl)
library(ggplot2)

gs4_auth()
drive_auth()
url <- "https://docs.google.com/spreadsheets/d/1uX4jipdyuYQSOWQkAPIFf177uL6XPRqt/edit?gid=778707309#gid=778707309"

#Criando um novo diretório para baixar a tabela de dados
novo_diretorio <- "C:/caseLepesR"
# Verificar se a pasta já existe e, se não, criá-la
if (!dir.exists(novo_diretorio)) {
  dir.create(novo_diretorio)
  print("Pasta criada com sucesso!")
} else {
  print("A pasta já existe.")
}

#Baixando a planilha no diretorio que acabou de ser criado
drive_download(as_id(url), path = "C:/caseLepesR/saeb.xlsx", type = "xlsx", overwrite = TRUE)

#Importando dado Local
dados <- read_excel("C:/caseLepesR/saeb.xlsx")

#Transformando dados em data table
dt_saeb <- as.data.table(dados)
```

Tratamento dos valores ausentes

Foi utilizado o método de imputação completando a base com a média dos valores de cada coluna levando em consideração a variável ano.

Após os dados serem importados foram transformados em um data table. Foi utilizado um for para percorrer o data table e imputar as médias nas colunas

```r
#Realizando a limpeza de dados ausentes
for (col in names(dt_saeb)) {
  if (is.numeric(dt_saeb[[col]])) {
    if(any(is.na(dt_saeb[[col]]))){
      dt_saeb[[col]][is.na(dt_saeb[[col]])] <- mean(dt_saeb[[col]], na.rm = TRUE)
    }
  }
}
```

Validação

Foi utilizada a função any(is.na( )) para verificar a existência de valores ausentes e quando ela retornou false, foi concluído a limpeza.