**TRABALHO PRÁTICO**

# DGT2823 - Tecnologias para Desenvolvimento de Soluções de Big Data

**UNIVERSIDADE:** Estácio

**CURSO:** Desenvolvimento Full Stack

**DISCIPLINA:** DGT2823 - Tecnologias para Desenvolvimento de Soluções de Big Data

**ALUNO:** Alex Barroso Paz

**MATRÍCULA:** 2023 0615 1781

**PROFESSOR/TUTOR:** Jhonatan Alves

**DATA:** 12/06/2025

**SUMÁRIO**

DGT2823 - Tecnologias para Desenvolvimento de Soluções de Big Data ........................................................... 1

1. INTRODUÇÃO ........................................................................................................................................................................ 3
2. OBJETIVOS .............................................................................................................................................................................. 4
   1. Objetivos Gerais ............................................................................................................................................................. 4
   2. Objetivos Específicos.................................................................................................................................................... 4
3. METODOLOGIA ..................................................................................................................................................................... 4
   1. Ambiente de Desenvolvimento ................................................................................................................................ 4
   2. Dataset Utilizado ........................................................................................................................................................... 5
   3. Problemas Identificados no Dataset....................................................................................................................... 5
4. MICROATIVIDADES .............................................................................................................................................................. 5
   1. Microatividade 1: Leitura de CSV............................................................................................................................. 5
   2. Microatividade 2: Subconjunto de Dados ............................................................................................................ 6
   3. Microatividade 3: Configuração de Visualização ............................................................................................... 7

4.5 Microatividade 5: Informações Gerais ................................................................................................................... 9

Trabalho Prático Final - DGT2823 ......................................................................................................................................11 Passo 1: Preparação do Dataset.........................................................................................................................................12

Passos 2-4: Leitura do CSV ..................................................................................................................................................13

Passo 5: Verificação dos Dados ..........................................................................................................................................14

Passo 6: Cópia do Dataset ...................................................................................................................................................14

Passo 7: Tratamento de Valores Nulos em 'Calories'..................................................................................................15

Passo 8: Tratamento Inicial da Coluna 'Date' ................................................................................................................16

Passo 9: Correção do Primeiro Erro ..................................................................................................................................16

Passo 10: Correção de Formato Inconsistente .............................................................................................................16

Passo 11: Conversão Final ....................................................................................................................................................16

Passo 12: Remoção de Registros Nulos ..........................................................................................................................16

Passo 13: Verificação Final ...................................................................................................................................................17

# 1. INTRODUÇÃO

Este trabalho prático tem como finalidade demonstrar o conhecimento adquirido na disciplina DGT2823 Tecnologias para Desenvolvimento de Soluções de Big Data, com foco específico na manipulação e limpeza de dados utilizando a biblioteca Pandas da linguagem Python.

O trabalho foi desenvolvido no ambiente Google Colab, uma plataforma baseada em nuvem que oferece notebooks Jupyter gratuitos, permitindo a execução de código Python de forma interativa e colaborativa.

O dataset utilizado contém informações sobre exercícios físicos, incluindo duração, data, pulso, pulso máximo e calorias queimadas. Este conjunto de dados apresenta propositalmente inconsistências e valores nulos, representando um cenário realista de dados que necessitam de tratamento antes de serem utilizados em análises.

# 2. OBJETIVOS

## 2.1 Objetivos Gerais

Aplicar técnicas de manipulação de dados utilizando a biblioteca Pandas

Realizar limpeza e tratamento de dados inconsistentes

Demonstrar proficiência em análise exploratória de dados

## 2.2 Objetivos Específicos



Descrever como ler arquivos CSV usando Pandas



Criar subconjuntos de dados a partir de conjuntos existentes



Configurar opções de

visualização de dados



Exibir informações estatísticas e estruturais de datasets



Tratar valores nulos e inconsistentes



Converter tipos de dados adequadamente



Validar a qualidade dos dados após tratamento

# 3. METODOLOGIA

## 3.1 Ambiente de Desenvolvimento

**Plataforma:** Google Colab

**Linguagem:** Python 3.x

**Biblioteca Principal:** Pandas

**Formato de Dados:** CSV (Comma-Separated Values)

## 3.2 Dataset Utilizado

O dataset contém 32 registros iniciais com as seguintes colunas:

**ID:** Identificador único do exercício

**Duration:** Duração do exercício em minutos

**Date:** Data do exercício

**Pulse:** Frequência cardíaca durante o exercício

**Maxpulse:** Frequência cardíaca máxima atingida

**Calories:** Calorias queimadas

**3.3**

**Problemas Identificados no Dataset**



Valores nulos (NaN) nas colunas Calories e Date



Formato inconsistente de data (linha 26: "20201226")



Necessidade de conversão de tipos de dados

# 4. MICROATIVIDADES

## 4.1 Microatividade 1: Leitura de CSV

**Objetivo:** Demonstrar a leitura de arquivos CSV utilizando a biblioteca Pandas.

**Código Implementado:**

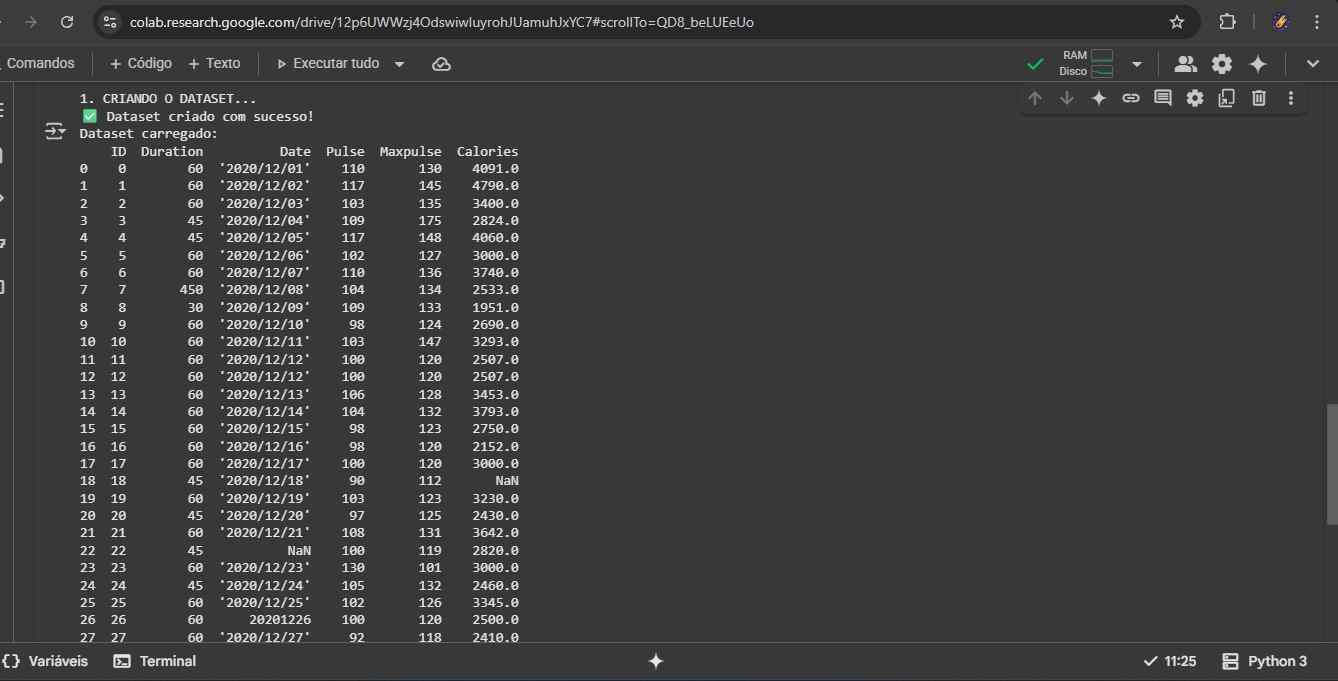
python import pandas as pd

df\_original =

pd.read\_csv('dados\_exercicio.csv', sep=';', engine='python', encoding='utf-8')

**Resultado:**

Screenshot mostrando o dataset carregado



**Análise:** A leitura foi realizada com sucesso, especificando o separador de colunas (;), a engine Python para maior flexibilidade e a codificação UTF-8.

## 4.2 Microatividade 2: Subconjunto de Dados

**Objetivo:** Criar um subconjunto contendo apenas 3 colunas do dataset original.

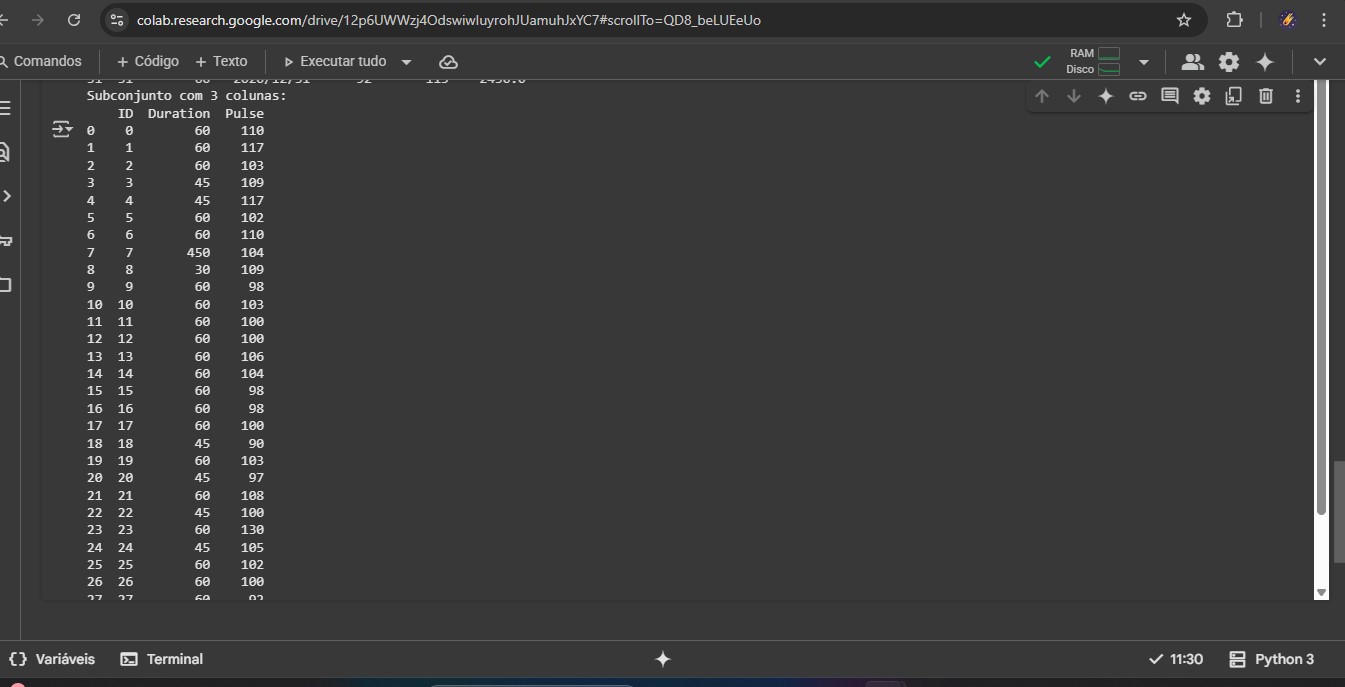
**Código Implementado:**

python df\_subconjunto = df\_original[['ID', 'Duration',

'Pulse']]

**Resultado:**

Screenshot do subconjunto com 3 colunas



**Análise:** O subconjunto foi criado com sucesso, mantendo apenas as colunas ID, Duration e Pulse, demonstrando a capacidade de seleção de colunas específicas.

## 4.3 Microatividade 3: Configuração de Visualização

**Objetivo:** Configurar o número máximo de linhas exibidas pelo Pandas.

**Código Implementado:**

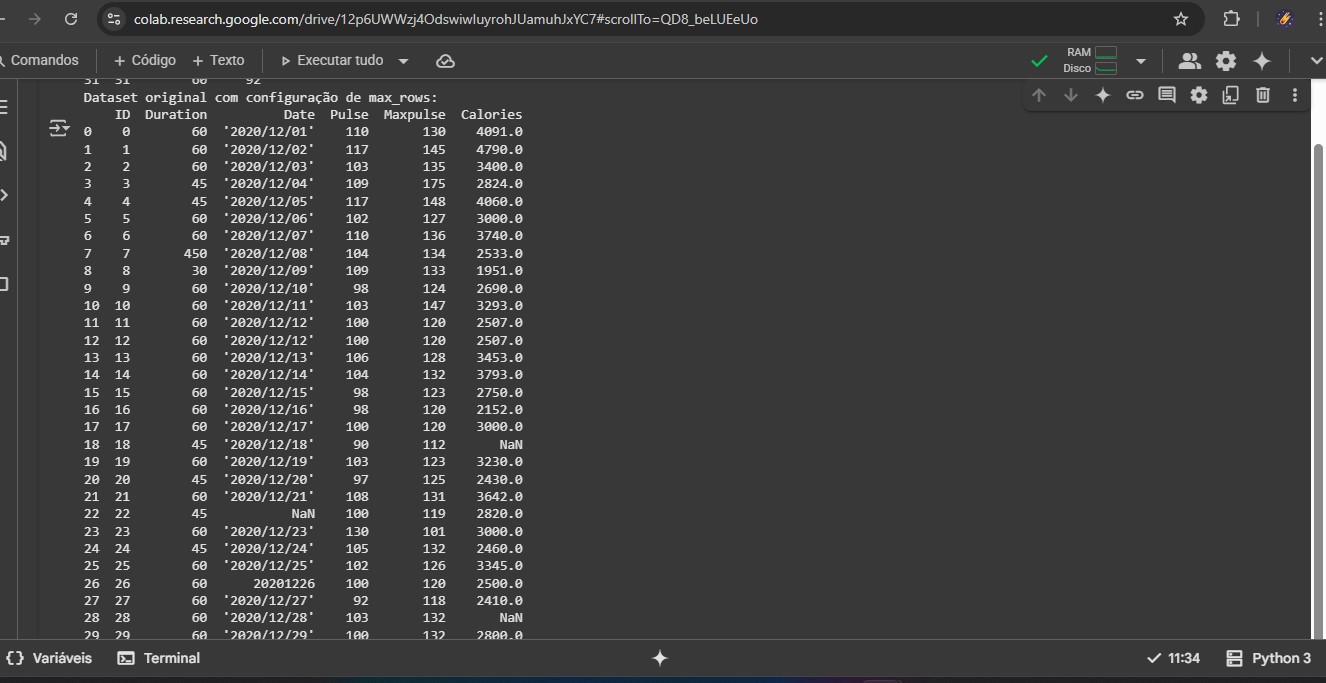
python

pd.set\_option('display.max\_rows',

9999) print(df\_original.to\_string())

**Resultado:**

Screenshot mostrando todas as linhas sem truncamento



**Análise:** A configuração permitiu visualizar todo o dataset sem truncamento, facilitando a análise completa dos dados.

**4.4 Microatividade 4: Primeiras e Últimas Linhas Objetivo:** Exibir as primeiras e últimas 10 linhas do dataset.

**Código Implementado:**

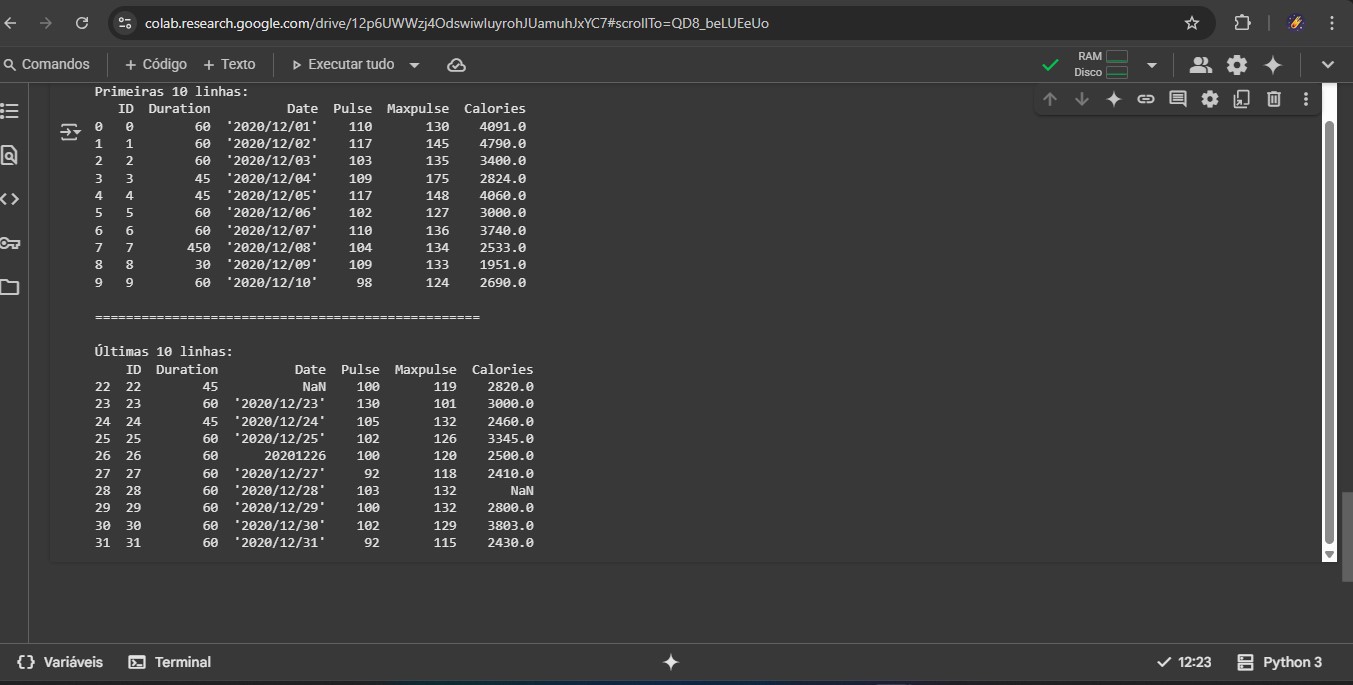
python print("Primeiras 10 linhas:") print(df\_original.head(10))

print("Últimas 10 linhas:") print(df\_original.tail(10))

**Resultado:**

Screenshot das primeiras 10 linhas

Screenshot das últimas 10 linhas



**Análise:** Os métodos head() e tail() facilitam a visualização rápida da estrutura e conteúdo do dataset, sendo fundamentais para análise exploratória.

## 4.5 Microatividade 5: Informações Gerais

**Objetivo:** Extrair informações estruturais e estatísticas do dataset.

**Código Implementado:**

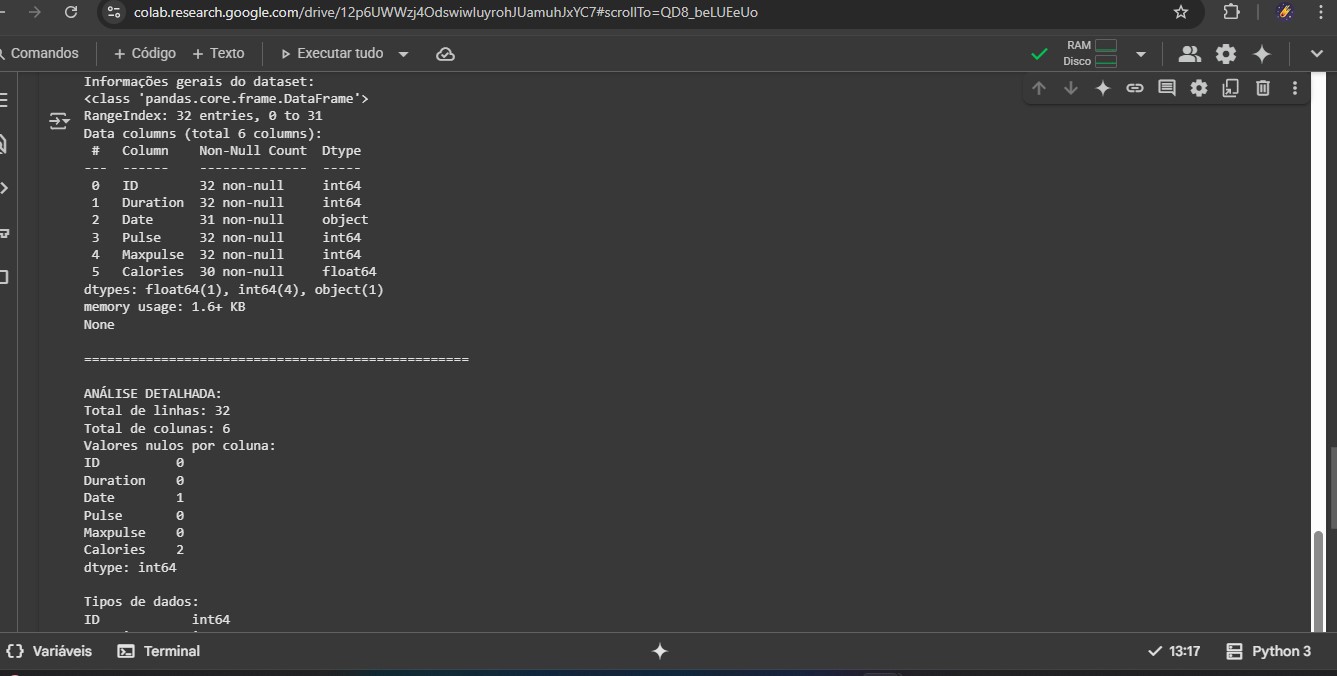
python print("Informações gerais:") print(df\_original.info())

print("Valores nulos:") print(df\_original.isnull().sum())

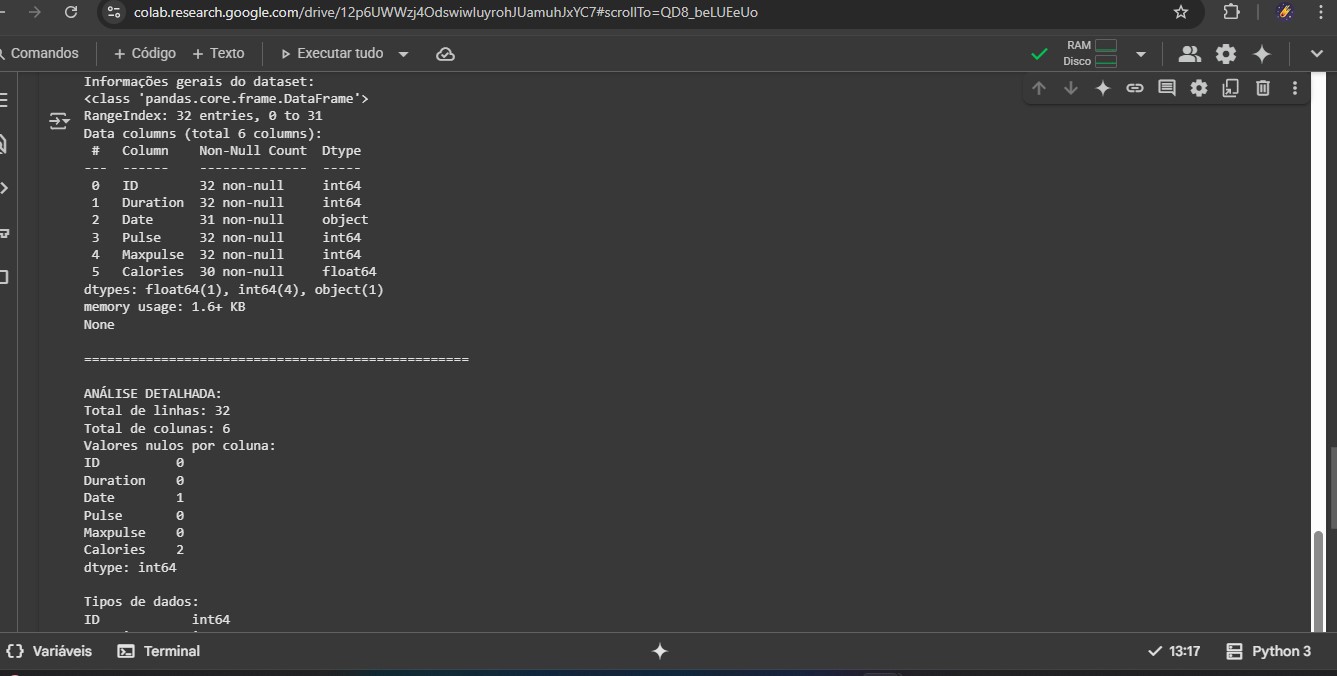
print("Tipos de dados:") print(df\_original.dtypes)

**Resultado:**

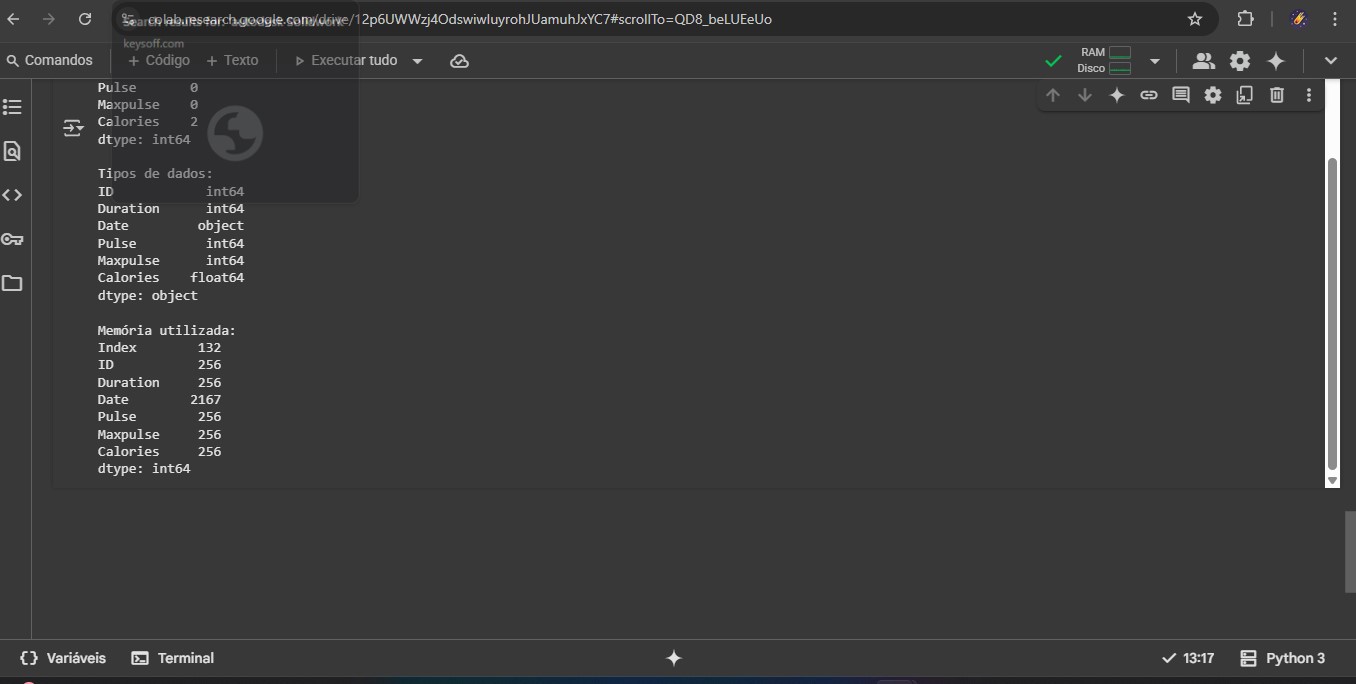
Screenshot das informações gerais



PRINT - Screenshot dos valores nulos



PRINT - Screenshot dos tipos de dados



**Análise:**



**Total de linhas:**



32



**Total de colunas:**



6



**Valores nulos identificados:**



3

registros (Calories: 2, Date:

1)



**Tipos de dados:**



Majoritariamente object, necessitando conversões

# Trabalho Prático Final - DGT2823

Este documento apresenta o código-fonte comentado para o trabalho prático da disciplina DGT2823

- Tecnologias para Desenvolvimento de Soluções de Big Data. O foco é a limpeza de dados utilizando a biblioteca Pandas no Python. Os principais trechos de código estão acompanhados de explicações resumidas. Espaços foram reservados para a inserção de prints das etapas mais relevantes.

# Passo 1: Preparação do Dataset

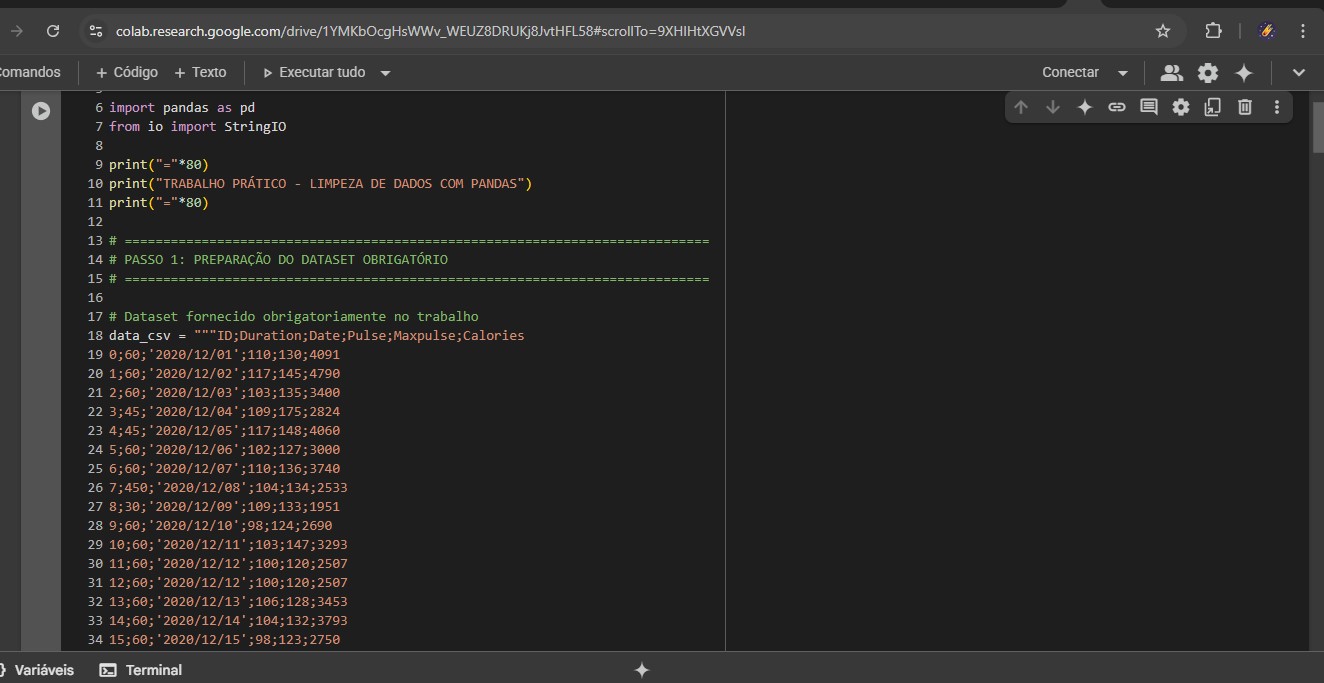
Criação do dataset obrigatório e salvamento em arquivo CSV.

data\_csv = """ID;Duration;Date;Pulse;Maxpulse;Calories ... (demais linhas do CSV) ...

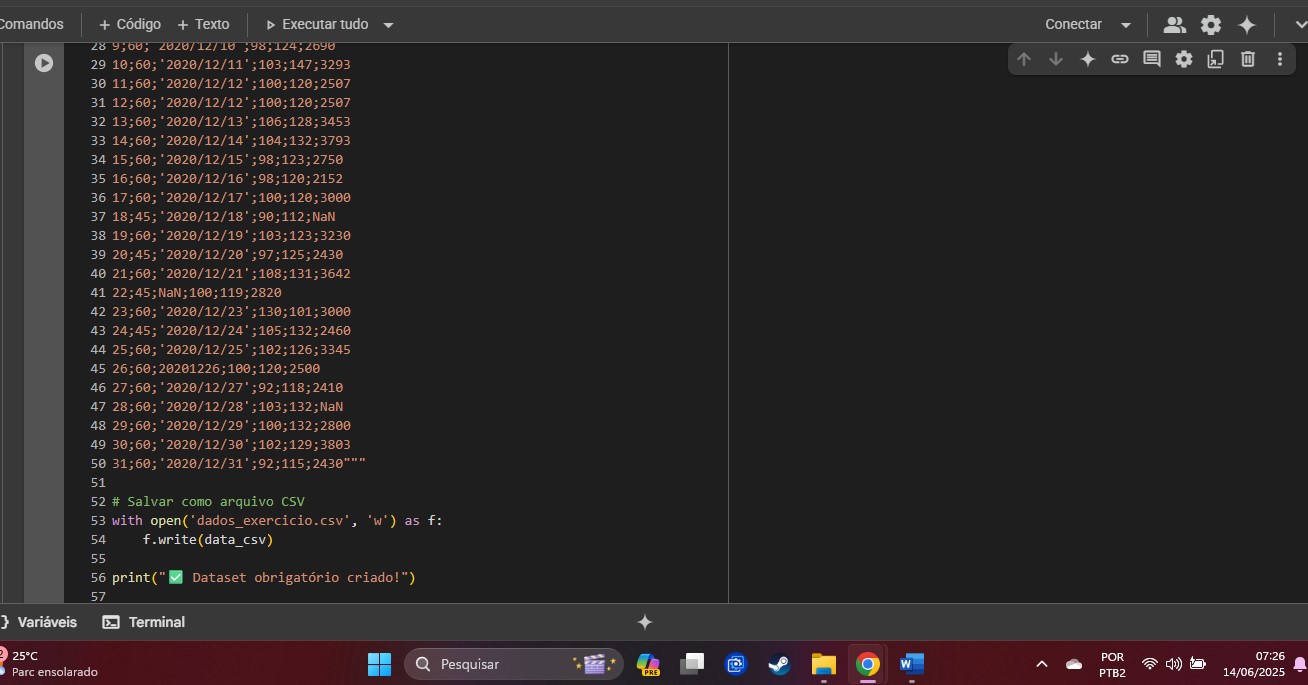
""" with open('dados\_exercicio.csv', 'w') as f:

f.write(data\_csv)

PRINT - Screenshot Criação do dataset



PRINT - Screenshot Criação do dataset

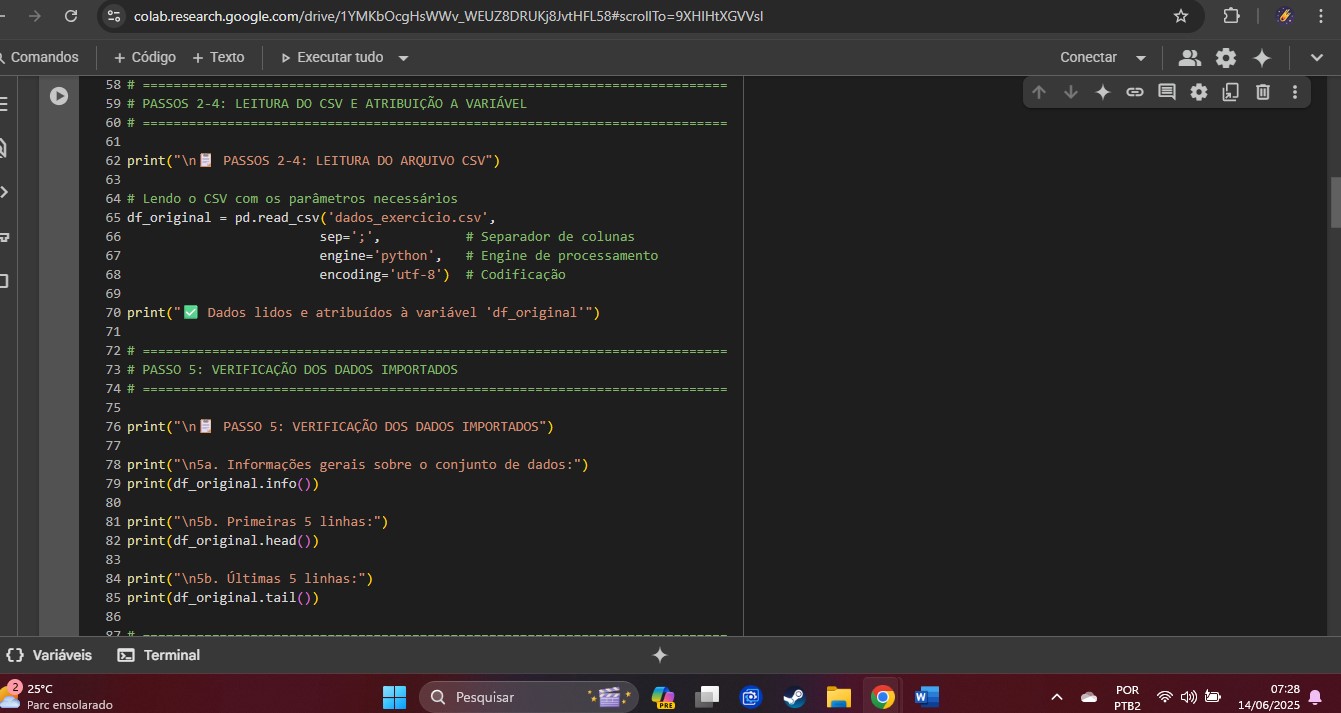


# Passos 2-4: Leitura do CSV

Leitura do arquivo CSV e atribuição à variável df\_original.

df\_original = pd.read\_csv('dados\_exercicio.csv', sep=';', engine='python', encoding='utf-8')

PRINT - Screenshot Leitura do arquivo CSV

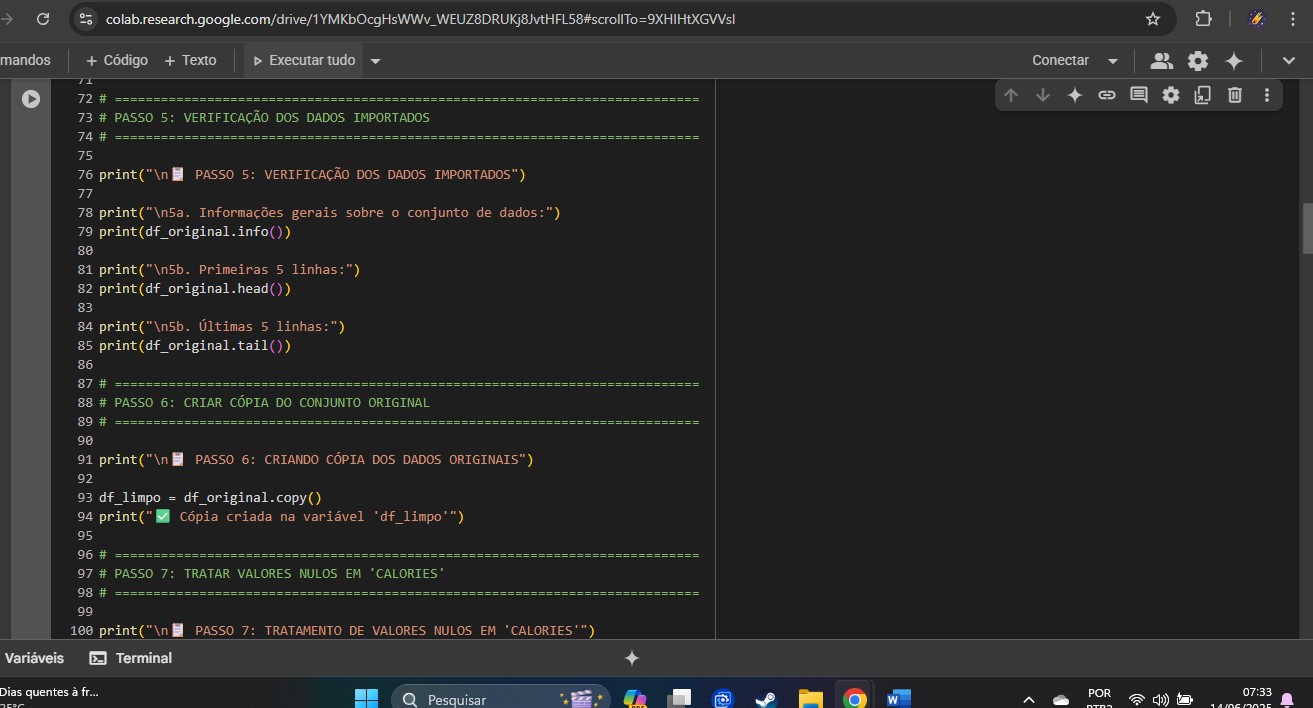


# Passo 5: Verificação dos Dados

Exibição das informações gerais e primeiras/últimas linhas.

print(df\_original.info()) print(df\_original.head()) print(df\_original.tail())

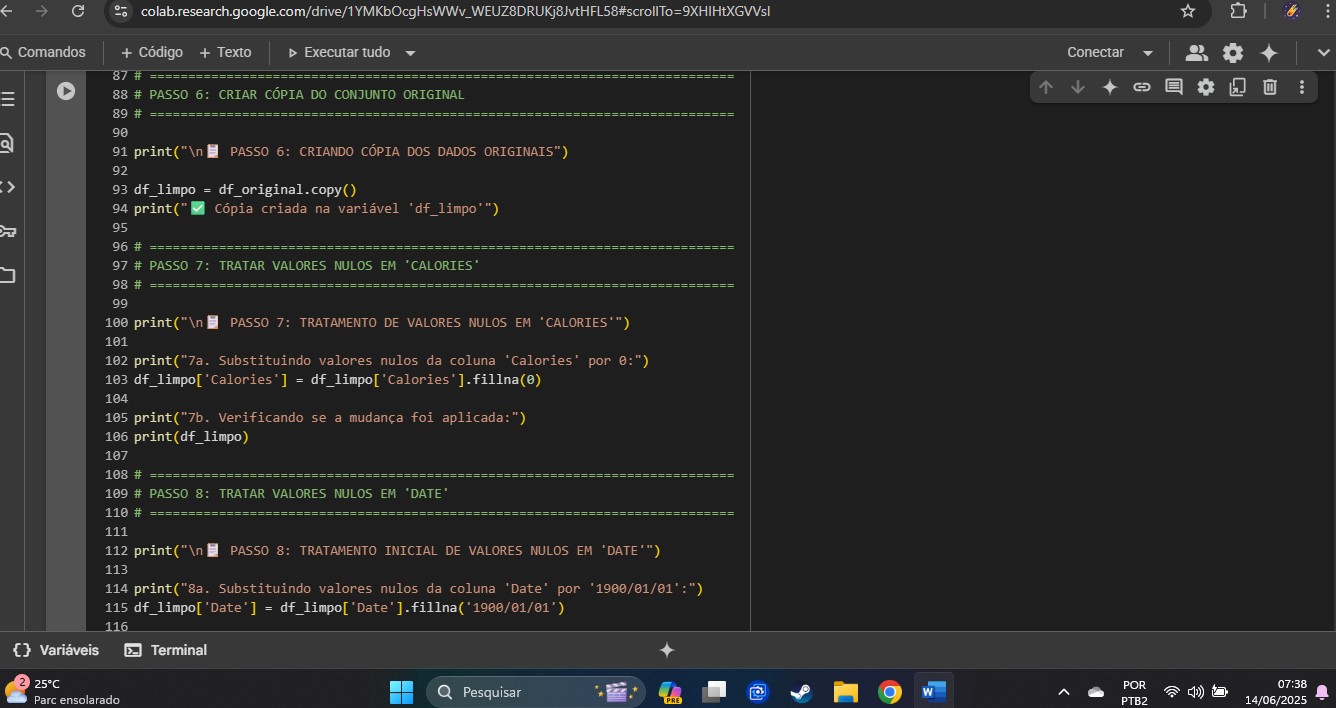
PRINT - Screenshot Verificação de dados



# Passo 6: Cópia do Dataset

Criação de uma cópia dos dados para futuras alterações. df\_limpo = df\_original.copy()

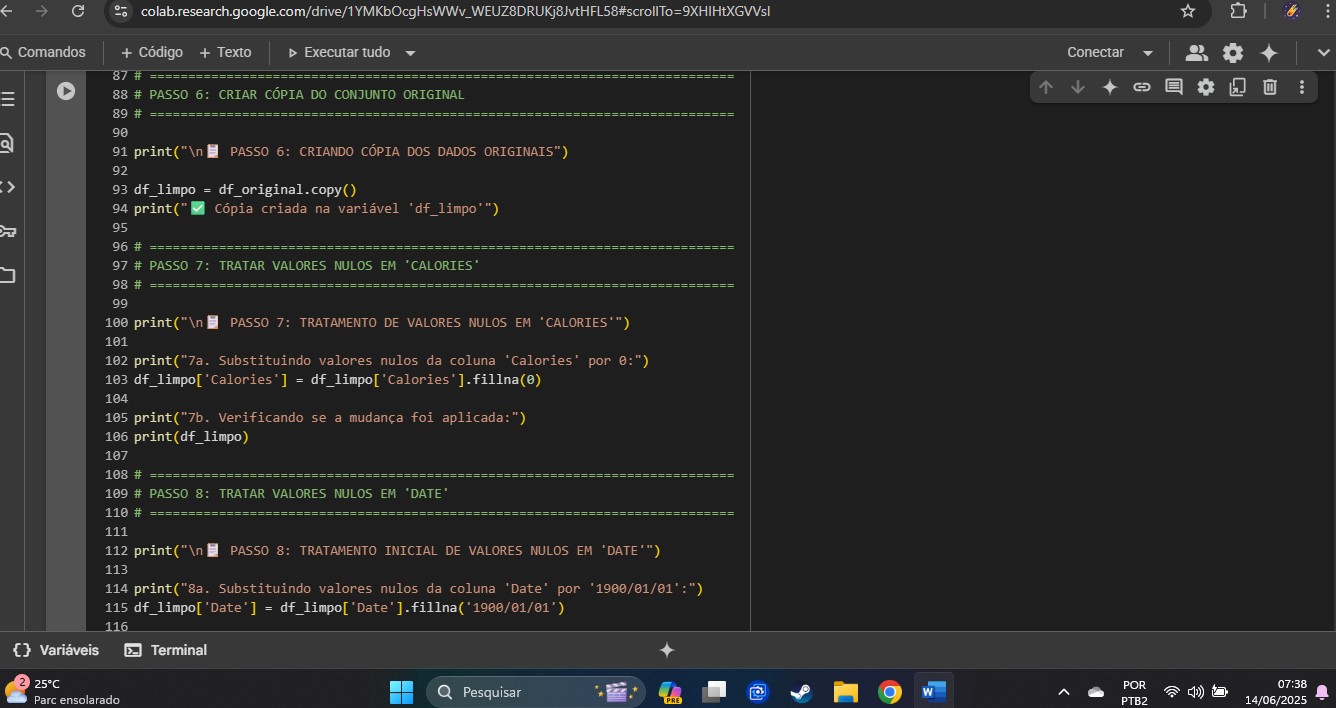
PRINT - Screenshot Cópia do Dataset



# Passo 7: Tratamento de Valores Nulos em 'Calories'

Substituição de valores nulos por 0 na coluna 'Calories'. df\_limpo['Calories'] = df\_limpo['Calories'].fillna(0)

PRINT - Screenshot Valores nulos em Calories



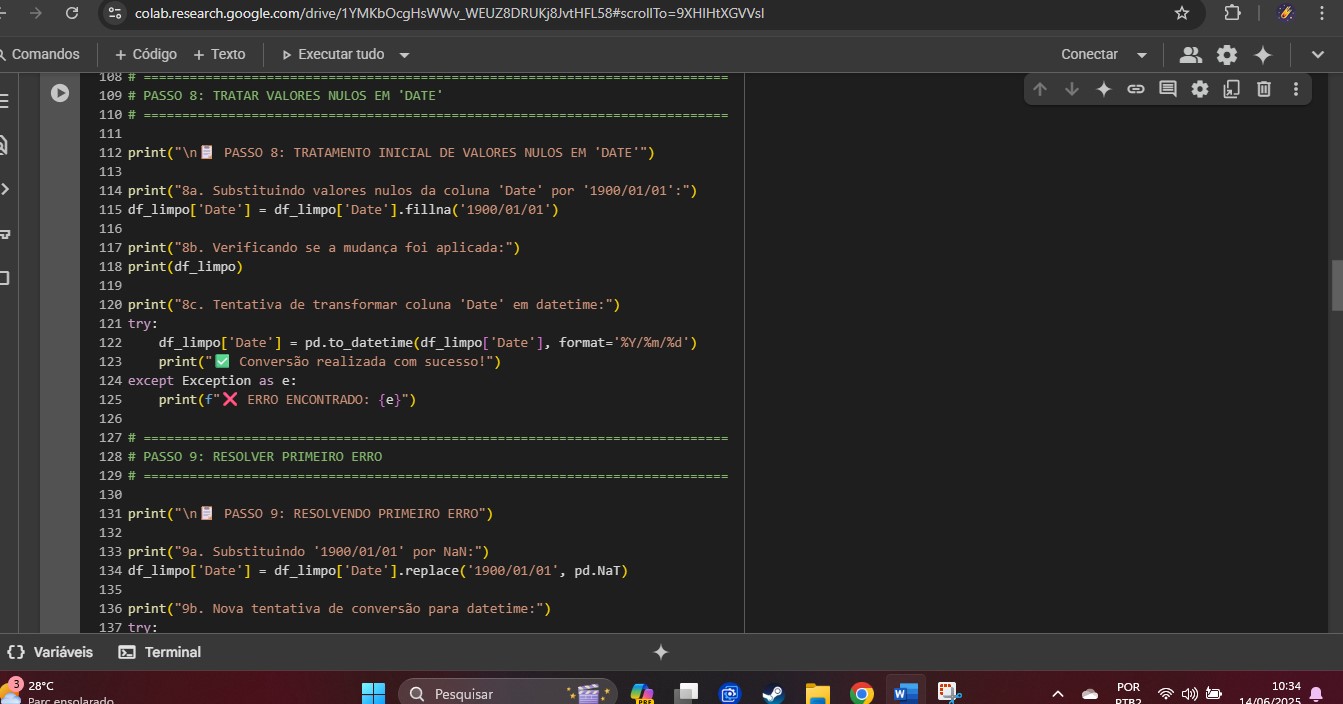
# Passo 8: Tratamento Inicial da Coluna 'Date'

Substituição de nulos por string temporária e tentativa de conversão.

df\_limpo['Date'] = df\_limpo['Date'].fillna('1900/01/01')

df\_limpo['Date'] = pd.to\_datetime(df\_limpo['Date'], format='%Y/%m/%d')

PRINT - Screenshot Incial da coluna Date



# Passo 9: Correção do Primeiro Erro

Substituição de '1900/01/01' por NaT e nova tentativa de conversão.

df\_limpo['Date'] = df\_limpo['Date'].replace('1900/01/01', pd.NaT)

df\_limpo['Date'] = pd.to\_datetime(df\_limpo['Date'], format='%Y/%m/%d', errors='coerce')

# Passo 10: Correção de Formato Inconsistente

Correção da entrada '20201226' com replace e nova conversão.

df\_limpo['Date'] = df\_limpo['Date'].replace('20201226', '2020/12/26')

df\_limpo['Date'] = pd.to\_datetime(df\_limpo['Date'], format='%Y/%m/%d', errors='coerce')

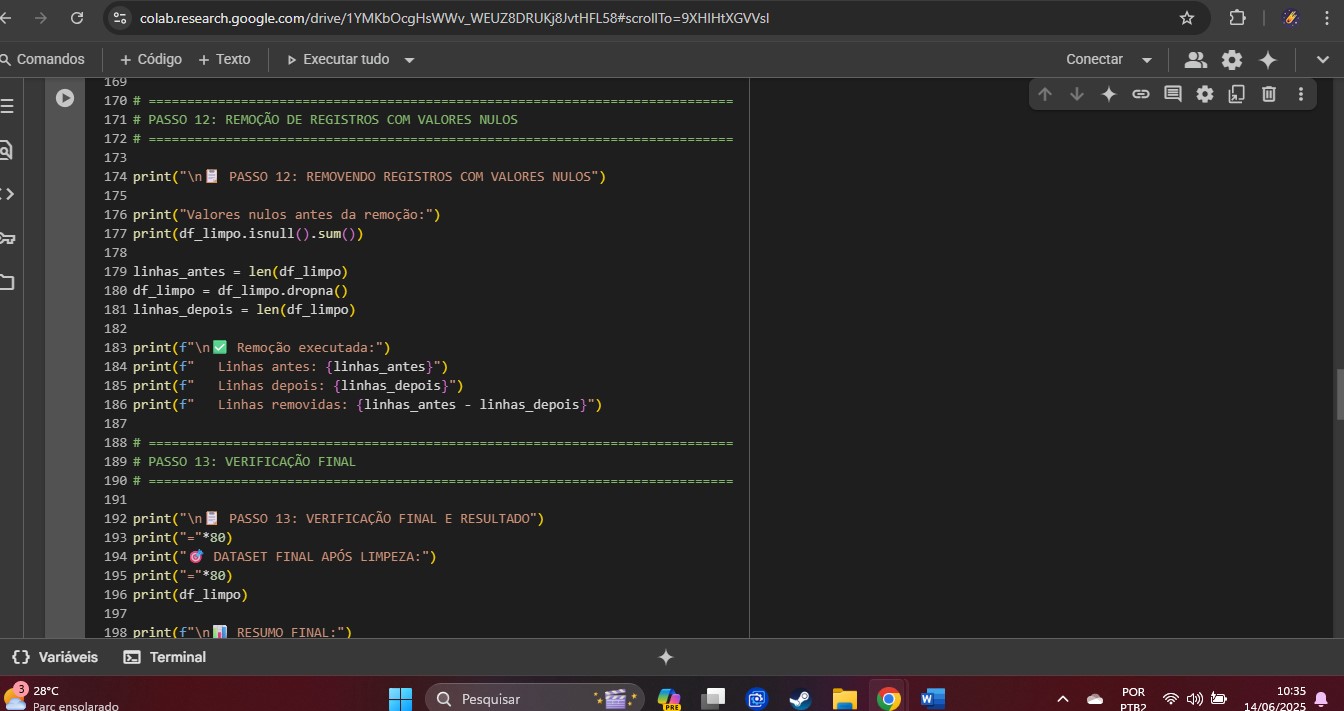
# Passo 11: Conversão Final

Verificação do tipo da coluna 'Date'. print(df\_limpo['Date'].dtype)

# Passo 12: Remoção de Registros Nulos

Remoção de linhas com valores nulos restantes. df\_limpo = df\_limpo.dropna()

PRINT - Screenshot Remoção registro nulos



# Passo 13: Verificação Final

Exibição do dataset final limpo e estatísticas descritivas.

print(df\_limpo.info()) print(df\_limpo.describe())

PRINT - Screenshot Verificação final

