PUBLIC TRANSPORTATION EFFICIENCY ANALYSIS

PHASE II

7145\_UNITEDINSTITUTEOFTECHNOLOGY\_Proj\_212990\_Team\_1

Team Mentor: [kavitha@uit.ac.in](mailto:kavitha@uit.ac.in)

Member Name: Mithun V

rajakuttimithun07@gmail.com

Abstract

In this phase we are going to load and preprocess the dataset.

Loading the dataset

Since our Datset is of CSV format, we go with pandas library to load the dataset

The below code simply defines the loading of the dataset.

The path for the dataset is mentioned below as “/content/Dataset.csv”

**Source Code:**

import numpy as np

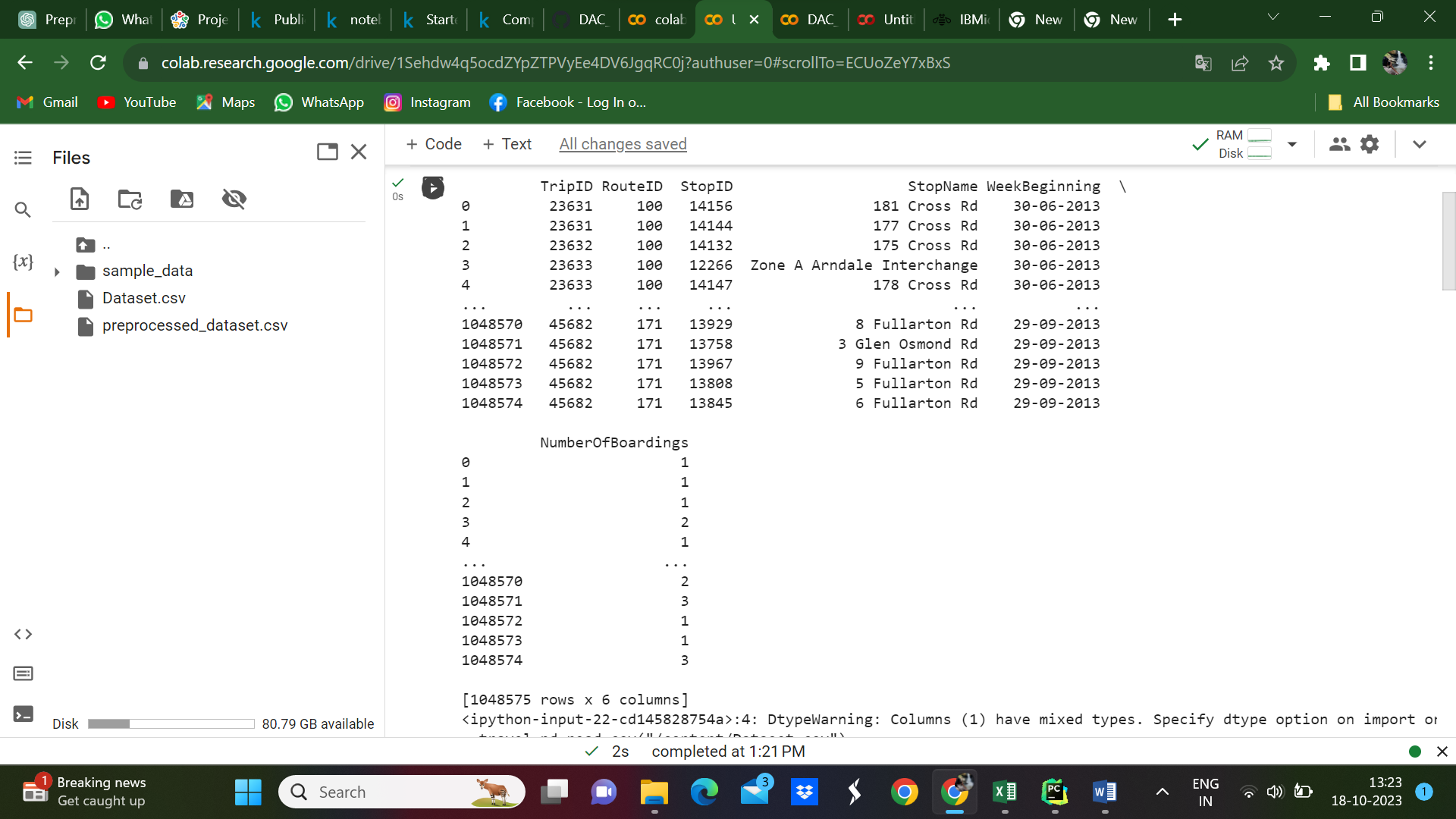
import pandas as pd

import matplotlib.pyplot as plt

travel=pd.read\_csv("/content/Dataset.csv")

print(travel)

**Output:**

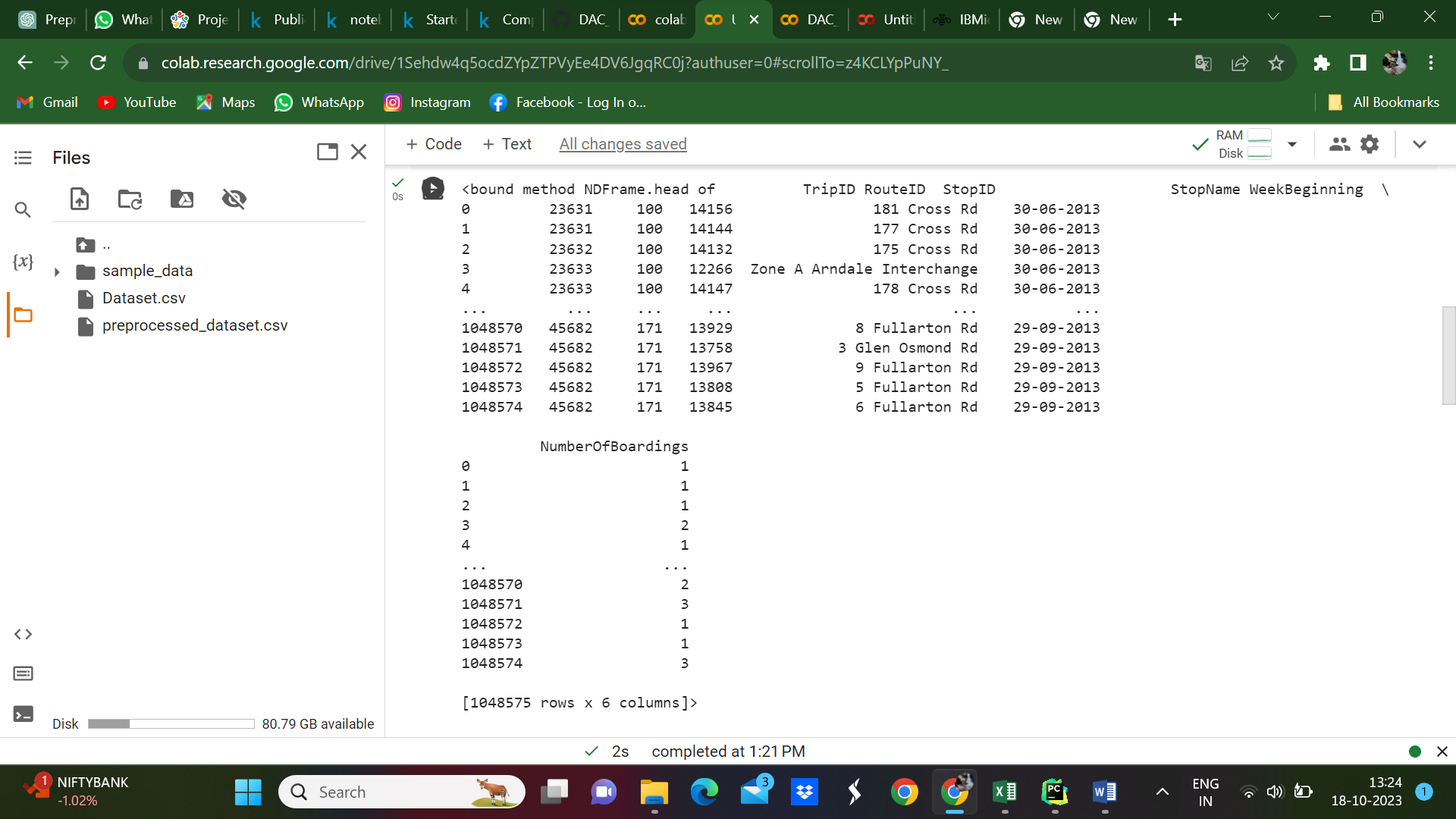


By using **print(travel.head())**, you'll see the first few rows of your dataset displayed in your console or Jupyter Notebook, which can help you get a sense of what your data looks like, the column names, and some example data points. This is a common step in the data exploration process to ensure that your data is loaded correctly and to have a preliminary view of its structure.

**Source Code:**

print(travel.head)

**Output:**



Remove rows with missing values:

The code `display(travel.drop\_duplicates())` is used to remove duplicate rows from the `travel` DataFrame and then display the resulting DataFrame

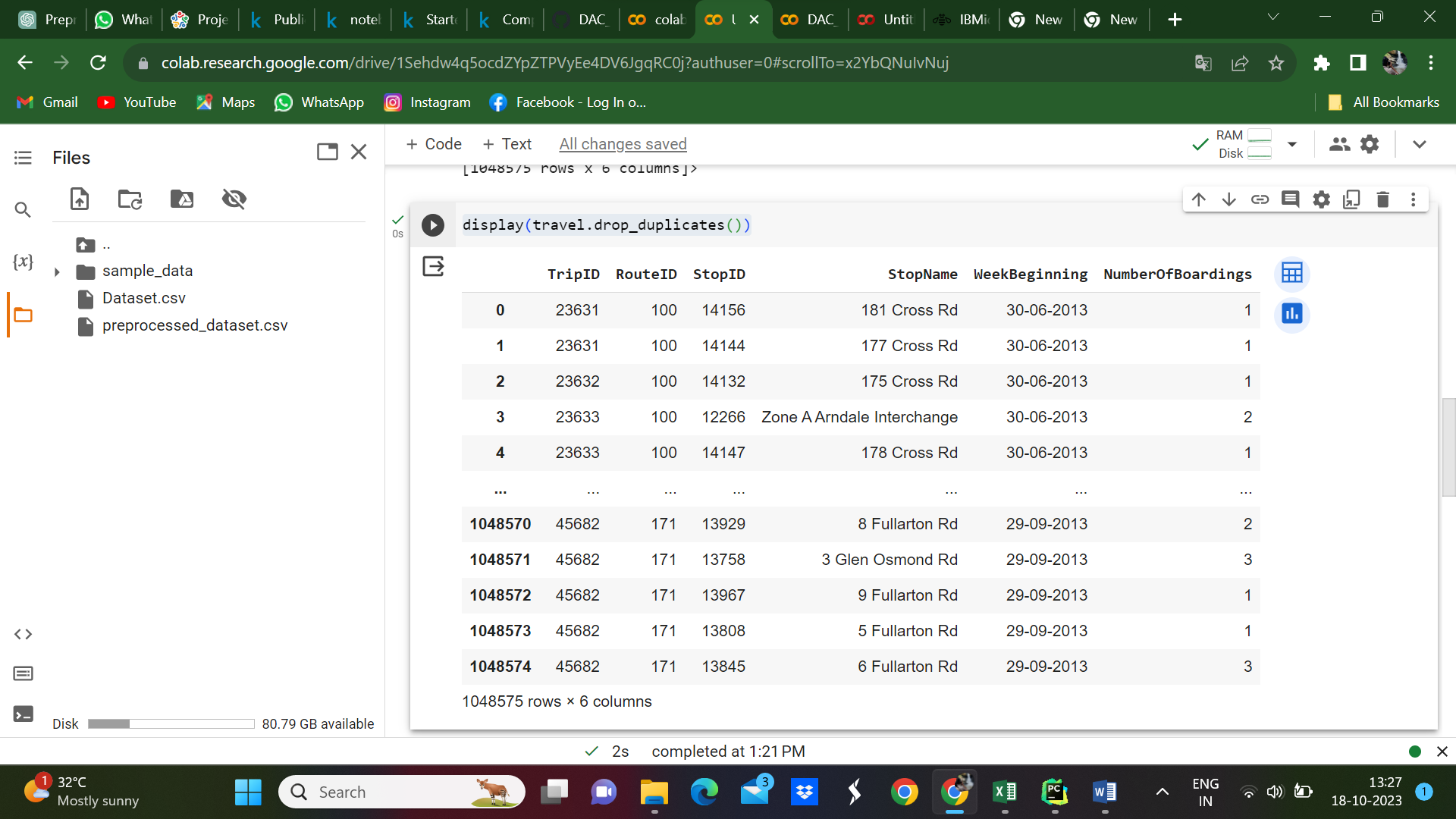
**drop\_duplicates() Method:** `drop\_duplicates()` is a built-in method in pandas that removes duplicate rows from a DataFrame. Rows are considered duplicates when all columns have the same values.

**display() Function**: The `display()` function is often used in Jupyter Notebook environments to render the output of a particular DataFrame, plot, or other data structure in a visually appealing way. The code will first remove duplicate rows from the `travel` DataFrame and create a new DataFrame with duplicates removed. Then, it will display the resulting DataFrame, which is now free of duplicate rows.This operation is significant when you want to ensure data integrity and cleanliness by eliminating redundant or erroneous data. Removing duplicates is a common data preprocessing step to ensure that your analysis or modeling is based on reliable and non-repetitive data.

**Source Code:**

display(travel.drop\_duplicates())

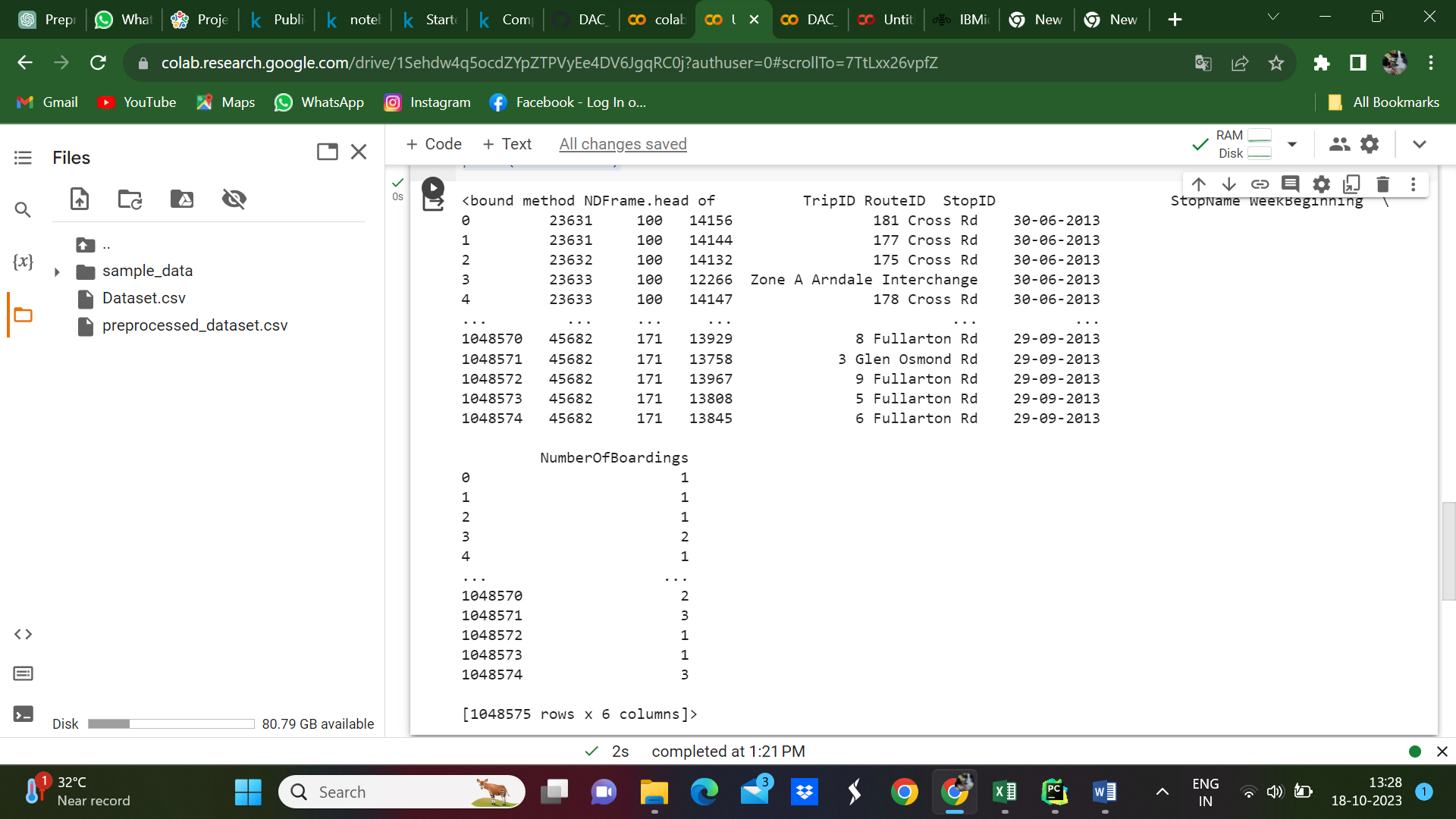
**Output:**



**Source Code:**

print(travel.head)

**Output:**



Missing Values:

The code `travel.isnull().sum()` is used to determine the count of missing (null) values in each column of the `travel` DataFrame. Here's the significance of this code:

isnull() Method: `isnull()` is a method provided by pandas for DataFrames that checks for missing values. It returns a DataFrame of the same shape as the original, where each element is `True` if the corresponding element in the original DataFrame is `NaN` (missing), and `False` otherwise.

sum() Method: After calling `isnull()`, you are applying the `sum()` method. This is used to calculate the sum of `True` values in each column. Since `True` is treated as 1 and `False` as 0 when summed, this effectively counts the number of missing values in each column.

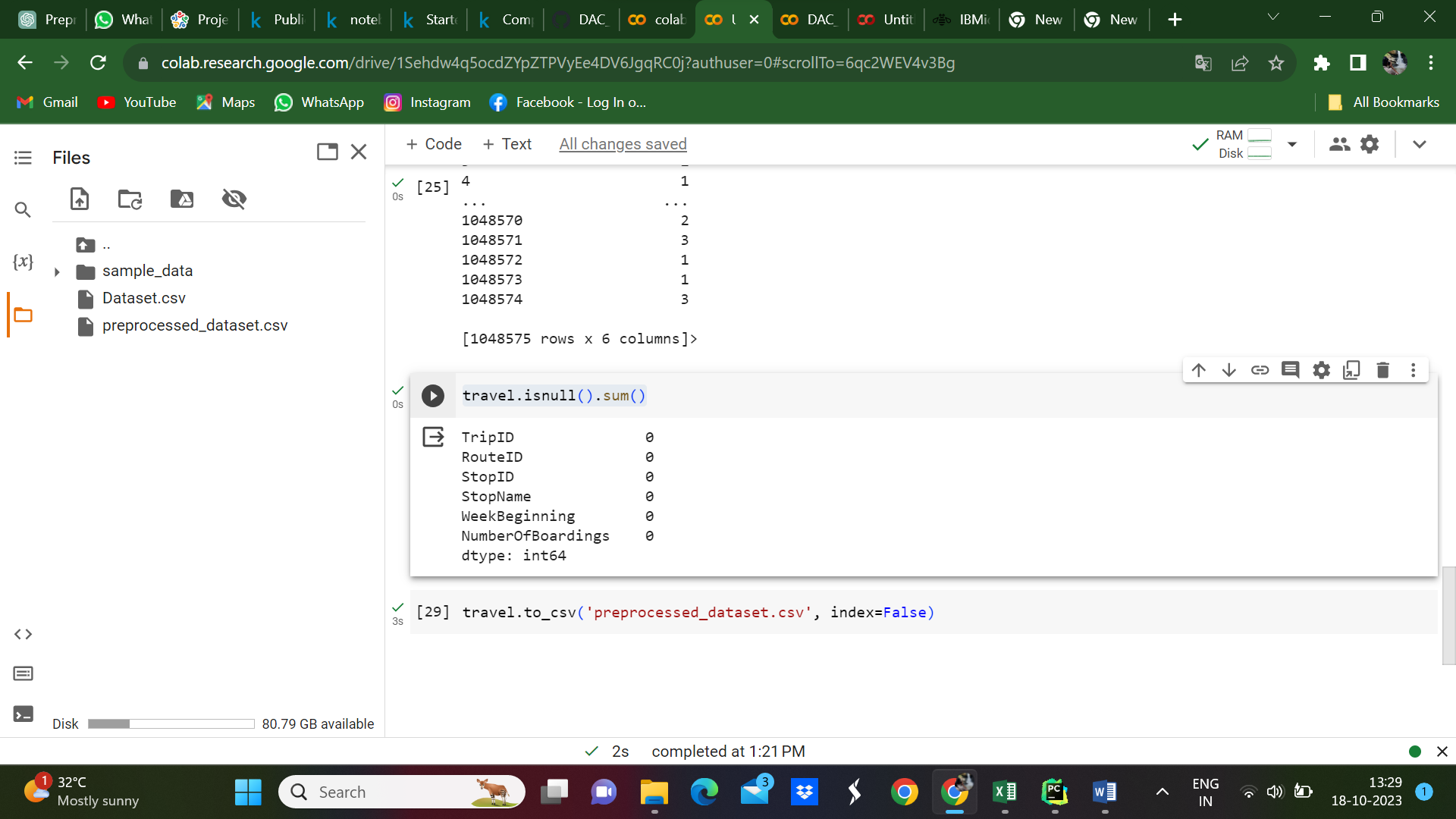
The output of `travel.isnull().sum()` will be a Series where the index corresponds to the column names of your DataFrame, and the values represent the count of missing values in each respective column.

This is a significant step in data preprocessing and data quality assessment. It helps you identify which columns have missing data and to what extent. You can then decide how to handle these missing values, whether by removing rows with missing values, imputing values, or using other strategies based on the context of your analysis or modeling.

**Source Code:**

travel.isnull().sum()

**Output:**



Save the Pre-Processed dataset:

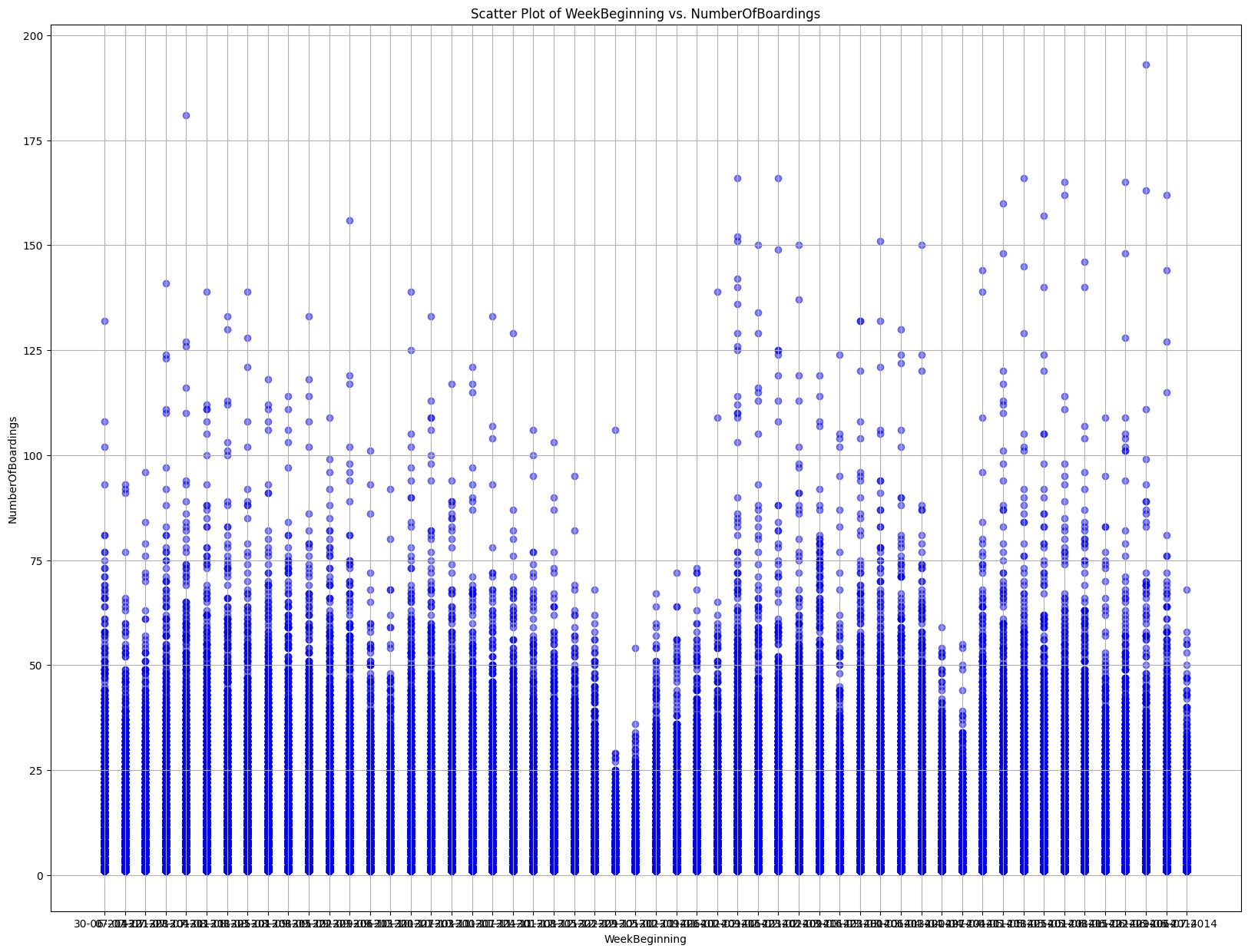
**Source Code:**

travel.to\_csv('preprocessed\_dataset.csv', index=False)

**Output:**

Preprocessed dataset created

Visualizing:



Conclusion:

In conclusion, data preprocessing and visualization are integral steps in the data analysis pipeline. They are essential for turning raw data into actionable insights and empowering decision-makers to make informed choices. A well-executed data preprocessing and visualization strategy can be a key driver of success in various domains, from business analytics to scientific research. Data preprocessing and visualization are fundamental steps in the data analysis process. They play a crucial role in making data more accessible, understandable, and usable for various analytical tasks. In this document, we have discussed the importance and key aspects of data preprocessing and visualization.