**Step 1: Dataset Structure Exploration**:

**Examine the Dataset Structure**:

* Look at the **number of rows** (records) and **columns** (features).
* Identify **column headers** to understand the meaning of each feature.
* Ensure that the data in each column makes sense (e.g., numerical data in numerical columns, categorical data in categorical columns).

**Calculate the Number of Rows and Columns**:

* In Excel, use the following formulas:
  + **Number of rows** (excluding the header):

=COUNTA(A2:A100) # Replace with the appropriate column

* + **Number of columns**:

=COLUMNS(A1:Z1) # Replace with the appropriate range

**Step 2: Basic Statistics and Data Summary**:

**Calculate Summary Statistics for Numerical Features**:

* For numerical columns (e.g., age, sepal length), use Excel functions to calculate:
  + **Mean**:

=AVERAGE(B2:B100) # Replace with your numerical column

* + **Minimum**:

=MIN(B2:B100)

* + **Maximum**:

=MAX(B2:B100)

* + **Standard Deviation**:

=STDEV.P(B2:B100)

**Frequency Count for Categorical Variables**:

* For categorical columns (e.g., gender, species), use **COUNTIF** to count the occurrences of each category.
* Example to count occurrences of "Male" in a "Gender" column:
* =COUNTIF(C2:C100, "Male")

**Step 3: Identify Missing Data**:

**Highlight Missing Values**:

* Use **Conditional Formatting** to highlight missing or blank cells:
* Select the data range, go to **Home > Conditional Formatting > New Rule**.
* Choose **Format only cells that contain** and select **Blanks**.
* Alternatively, use Excel formulas to count missing values in a column:

=COUNTBLANK(B2:B100) # Replace with your column range

**Handle Missing Data**:

* You can either:
* **Remove rows with missing data** by selecting and deleting them.
* **Fill missing values** by replacing them with the column mean, median, or mode.

**Step 4: Data Visualization**:

**Create Visualizations to Understand the Data**:

* **Histogram** for numerical data (e.g., sepal length, age).
* Select the data column, then go to **Insert > Chart > Histogram**.
* **Bar Chart** for categorical data (e.g., species, gender).
* Select the data column, then go to **Insert > Chart > Bar Chart**.

**Interpret the Visualizations**:

* Look for patterns, outliers, or imbalances in the data.
* Discuss any insights you can draw from the visualizations (e.g., which species is most common in the Iris dataset).

**Step 5: Prepare for Discussion and Submission**:

**Summarize your findings:**

* What are the key observations you made during your data exploration?
* Were there any missing values or outliers, and how did you handle them?
* What insights did you gain from your visualizations?

**Step-by-Step Data Cleaning Drill**

Open the Titanic dataset in Excel.

**Identify Missing Data**:

* Use **Conditional Formatting** to highlight missing values:
  + Select the data range (e.g., from A2 to J1000).
  + Go to **Home > Conditional Formatting > New Rule**.
  + Choose "Format only cells that contain" and select "Blanks".

**Handle Missing Values**:

* For **Age** (numerical data), replace missing values with the **mean** of the column:
  + Calculate the **mean** using =AVERAGE(E2:E1000) in Excel (assuming E is the "Age" column).
* Replace missing values with the calculated mean or median, depending on the dataset's distribution.
  + For **Embarked** (categorical data), replace missing values with the **mode** (the most frequent value):
  + Use =MODE(I2:I1000) to find the mode for "Embarked" and replace missing values with that mode.

**GOOGLE COLAB PYTHON CODE SUMMARY**

**# Import the pandas library as pd**

import pandas as pd

**# Load the file "train.csv" dataset using the df variable**

df = pd.read\_csv("train.csv")

**# Display the full content of the dataset**

print(df)

**# Display the first 10(Head) and last(Tail) 10 rows**

print(df.head(10))

print(df.tail(10))

**# Count or check the missing values per column**

print(df.isnull().sum())

**# Show the structure and details of the dataset**

print(df.info())

**# Display statistical data for numeric columns**

print(df.describe())

**# Count the number of rows and columns**

print(df.shape)

**# Show data types of each column**

print(df.dtypes)

**# List all column names**

print(df.columns)

**# Print all unique values in the 'Name' column**

print(df['Name'].unique())

**# Fill missing values in the 'Age' column using the median**

df['Age'].fillna(df['Age'].median(), inplace=True)

print(df.isnull().sum())

**# Calculate and print the mean of the 'Age' column**

mean\_age = df['Age'].mean()

print("Mean (Age):", mean\_age)

**# Calculate and print the median of the 'Fare' column**

median\_fare = df['Fare'].median()

print("Median (Fare):", median\_fare)

**# Calculate and print the mode of the 'Embarked' column**

mode\_embarked = df['Embarked'].mode()[0]

print("Mode (Embarked):", mode\_embarked)

**# Fill missing values in the 'Embarked' column using the mode**

df['Embarked'].fillna(mode\_embarked, inplace=True)