

# AMAL MAJI

Certificate Program in Generative AI: Batch 13 (Tech | Data | Finance)

I am Choosing the DATA set of Tips. So I am analysing the basic details of below inquiry.

- 1.Basic Statistics of 5 important columns(Mean,Median,Mode,Std Dev,Min,Max)
- 2.Check for Missing Values and Outlies
- 3.Univariate Analysis(Visualize the distribution of a column)
- 4.Bi-variate Analysis"

- 1.Basic Input-Screen Shoot

ChatGPT ▾

 tips (2) (1) (1) (1).csv  
Spreadsheet

You are an expert Data Scientist with key expertise in Python and Data Analysis and Visualization. I am going to execute a data analysis task in Google colab, and you need to support me in writing the code so that i can execute it in Google Colab . Also provide how the output of the code would look like for me to visualize if that is what i want. I am a beginner, so please explain the things in details. I am also attached the CSV file for your reference on which i am performing the analysis.

Perfect 🙌 Thanks for sharing your dataset. I'll guide you step by step so you can run everything smoothly in **Google Colab**.

We'll start with some basics: loading the dataset, exploring it, and then performing **data analysis & visualization**.

#### ◆ Step 1: Import Required Libraries

In Google Colab, you'll need to import the standard Python data science libraries.

```
python 📄 Copy code  
  
# Import essential Libraries  
import pandas as pd # For data analysis  
import numpy as np  # For numerical operations  
import matplotlib.pyplot as plt # For visualization  
import seaborn as sns # For advanced visualization  
  
# Make plots Look better  
sns.set(style="whitegrid")  
⬇
```

- 2. Basic Statistics of 5 important columns (Mean, Median, Mode, Std Dev, Min, Max) - Screen Shoot

1. Basic Statistics of 5 important columns (Mean, Median, Mode, Std Dev, Min, Max)

Got it 🧐 You want to focus only on 5 important columns and compute their basic statistics:

👉 Mean, Median, Mode, Std Dev, Min, Max

Let's assume the important columns from your dataset are:

```
["total_bill", "tip", "size", "day", "sex"]
```

(Because "day" and "sex" are categorical, we'll include them just for Mode. For numeric stats, we'll focus on numeric columns only).

Here's the code:

#### ◆ Step 7: Basic Statistics of 5 Columns

python

Copy code

```
# Select 5 important columns
columns_to_analyze = ["total_bill", "tip", "size", "day", "sex"]

# Create dictionary to store results
stats = {}

for col in columns_to_analyze:
    if df[col].dtype in ['int64', 'float64']: # Numeric columns
```

+ Ask anything



- 3.Check for Missing Values and Outlies-Screen Shoot

## 2.Check for Missing Values and Outlies


Excellent step  After basic stats, the next part of any data analysis is checking for:

1. **Missing Values** (to see if any data is incomplete).
2. **Outliers** (values that are unusually far from the rest).

Let's handle them one by one in **Google Colab**.

### ◆ Step 9: Check for Missing Values

python


 Copy code

```
# Check missing values count in each column
missing_values = df.isnull().sum()

print("==== Missing Values in Each Column ====")
print(missing_values)
```

### 👉 Expected Output (for Tips dataset)

sql

 Copy code

```
==== Missing Values in Each Column ====
total_bill    0
tip           0
```

- 4.Univariate Analysis(Visualize the distribution of a column)-Screen Shoot

### 3.Univariate Analysis(Visualize the distribution of a column)

Perfect 🍌 Now we'll move to **Univariate Analysis**, which means analyzing **one variable (one column)** at a **time**.

This helps us understand the **distribution** of values in that column.

#### ◆ Step 11: Univariate Analysis

We can do this in two cases:

##### (a) For Numeric Columns

We'll use **Histogram** + **KDE curve** to see the distribution.

python

Copy code

```
# Example: Distribution of "total_bill"
plt.figure(figsize=(8,5))
sns.histplot(df["total_bill"], bins=30, kde=True, color="skyblue")
plt.title("Distribution of Total Bill", fontsize=14)
plt.xlabel("Total Bill Amount")
plt.ylabel("Frequency")
plt.show()
```

##### 🍌 Expected Output:

A histogram showing how **total\_bill** values are spread



## • 5.Bi-variate Analysis)-Screen Shoot



### 4.Bi-variate Analysis

Excellent 🏆 Now that we've done **Univariate Analysis**, let's move to **Bivariate Analysis** → studying the relationship between **two variables (columns)**.

This helps us answer questions like:

- Do higher **bills** lead to higher **tips**?
- Do **smokers** tip differently than **non-smokers**?
- Do **men and women** behave differently in tipping?

### ◆ Step 12: Bivariate Analysis

#### (a) Numeric vs Numeric

Example: Total Bill vs Tip (scatter plot)

python

Copy code

```
plt.figure(figsize=(8,5))
sns.scatterplot(x="total_bill", y="tip", data=df, hue="sex", palette="Set1")
plt.title("Total Bill vs Tip (colored by Gender)", fontsize=14)
plt.xlabel("Total Bill ($)")
```