DATA ANALYSIS & VISUALIZATION

U18AIE0219L

EXPERIMENT -1

Pre-lab Questions:

1. Define the term 'descriptive statistics'.

Descriptive statistics refers to a branch of statistics that focuses on summarizing and describing the main features of a dataset. It provides simple numerical and graphical summaries that help to understand the data's overall structure, distribution, and key characteristics.

2. What are the key measures of central tendency and variability?

• The key measures of central tendency are:

o Mean: The average of all values

o Median: The middle value when data is sorted

o Mode: The most frequently occurring value

• The key measures of variability are:

o Range: The difference between the highest and lowest values

o Variance: Measures the average squared differences from the mean

o Standard Deviation: The square root of variance, indicating how much values deviate from the mean

3. Why are mean, median, and mode important in summarizing data?

The mean, median, and mode are important because they summarize the central tendency of a dataset:

- Mean:
  - Represents the average
  - Useful for normally distributed data
  - Sensitive to outliers
- Median:
  - The middle value, splitting the data into two halves
  - Resistant to outliers
  - Ideal for skewed data
- Mode:
  - The most frequent value
  - Useful for categorical data

    - Highlights patterns in the dataset

4. How would you detect an outlier in a dataset using basic statistics?

Outliers in a dataset can be detected using the following basic statistical methods:

- Z-Score:
    - Calculate the z-score: Z=x−μσZ = \frac{x - \mu}{\sigma}Z=σx−μ.
    - Any value with |Z|>3|Z| > 3|Z|>3 is considered an outlier.
- Interquartile Range (IQR):
    - Compute Q1 (25th percentile) and Q3 (75th percentile).
    - Calculate IQR: IQR=Q3−Q1IQR = Q3 - Q1IQR=Q3−Q1.
    - Outliers are values:
    - Below Q1−1.5×IQRQ1 - 1.5 \times IQRQ1−1.5×IQR,
    - Above Q3+1.5×IQRQ3 + 1.5 \times IQRQ3+1.5×IQR.
- Visualization:
    - Use boxplots to visually identify points outside the whiskers.

5. Provide an example of when you would use descriptive statistics in real world applications.

Descriptive statistics are used in business sales analysis, where measures like mean sales, median profit, and standard deviation are calculated to understand performance, identify trends, and detect variability across regions or time periods. For example, a company may analyze quarterly sales data to determine the average sales (mean), the most frequent sales figure (mode), and variability in sales (standard deviation) to improve strategies.

Descriptive Statistical Analysis

Aim:

To perform descriptive statistical analysis on a dataset to compute central tendency, variability, and distribution

Resources used:

Programming Language – Python

Platform-Google Colab

Dataset- sales_data.csv

Code:

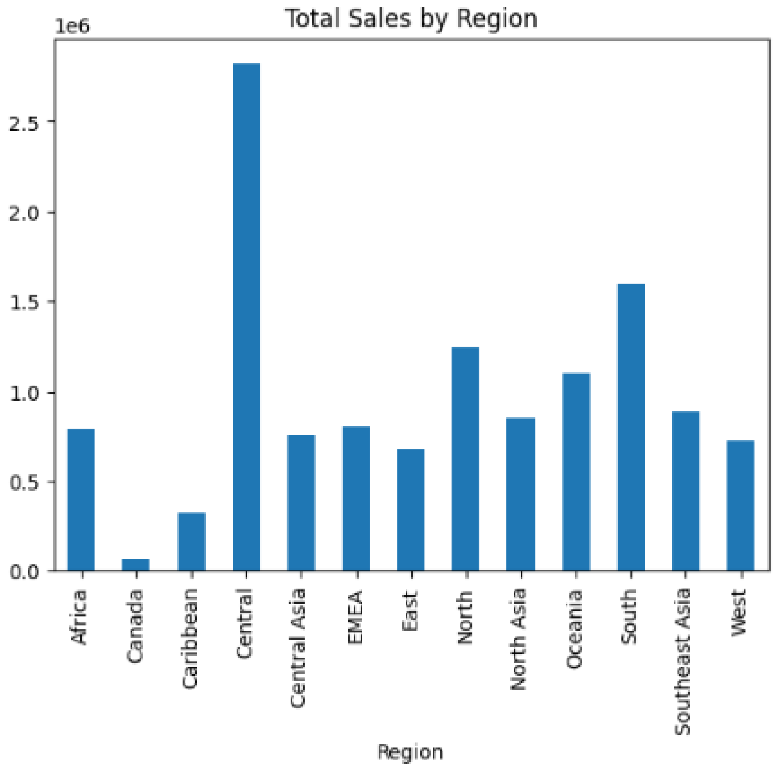
```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

# Load the dataset

data=pd.read_csv(r"sales_data.csv",encoding="latin1")

# Mean

mean_sales = data['Sales'].mean()

mean_profit = data['Profit'].mean()

# Median

median_sales = data['Sales'].median()

median_profit = data['Profit'].median()

# Mode

mode_sales = data['Sales'].mode()[0]

mode_profit = data['Profit'].mode()[0]


print(f"Mean Sales: {mean_sales}, Mean Profit: {mean_profit}")

print(f"Median Sales: {median_sales}, Median Profit: {median_profit}")

print(f"Mode Sales: {mode_sales}, Mode Profit: {mode_profit}")
```

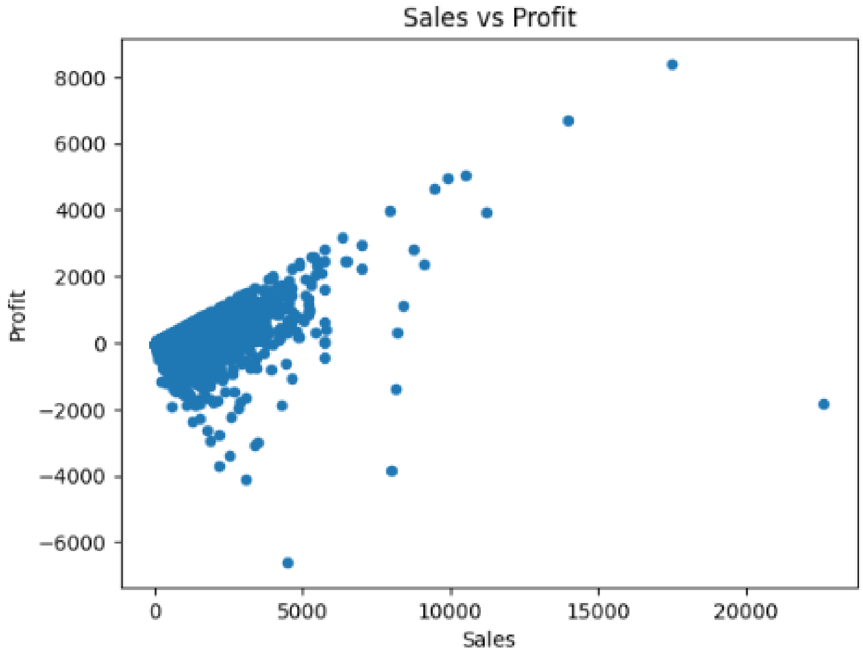
```
# Standard Deviation
std_sales = data['Sales'].std()
std_profit = data['Profit'].std()
# Variance
var_sales = data['Sales'].var()
var_profit = data['Profit'].var()
# Range
range_sales = data['Sales'].max() - data['Sales'].min()
range_profit = data['Profit'].max() - data['Profit'].min()


print(f"Standard Deviation (Sales): {std_sales}, (Profit):
{std_profit}")
print(f"Variance (Sales): {var_sales}, (Profit): {var_profit}")
print(f"Range (Sales): {range_sales}, (Profit): {range_profit}")


# bar chart- Compare total sales or profit across regions
data.groupby('Region')['Sales'].sum().plot(kind='bar', title='Total
Sales by Region')
# scatter plot- Explore the relationship between sales and profit
data.plot(kind='scatter', x='Sales', y='Profit', title='Sales vs
Profit')
# pie chart- Show regional contribution to total sales
data.groupby('Region')['Sales'].sum().plot(kind='pie',
autopct='%1.1f%%', title='Sales Distribution by Region')
```

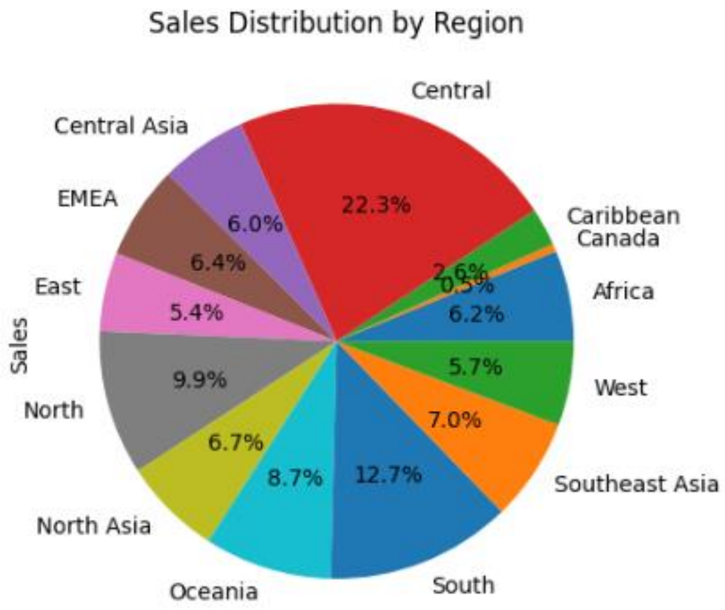
Google Colab Link:

https://colab.research.google.com/drive/1p9Xgqbynk_5oc-qBzgEoznEfwpf7_jwQ?usp=sharing

Output:

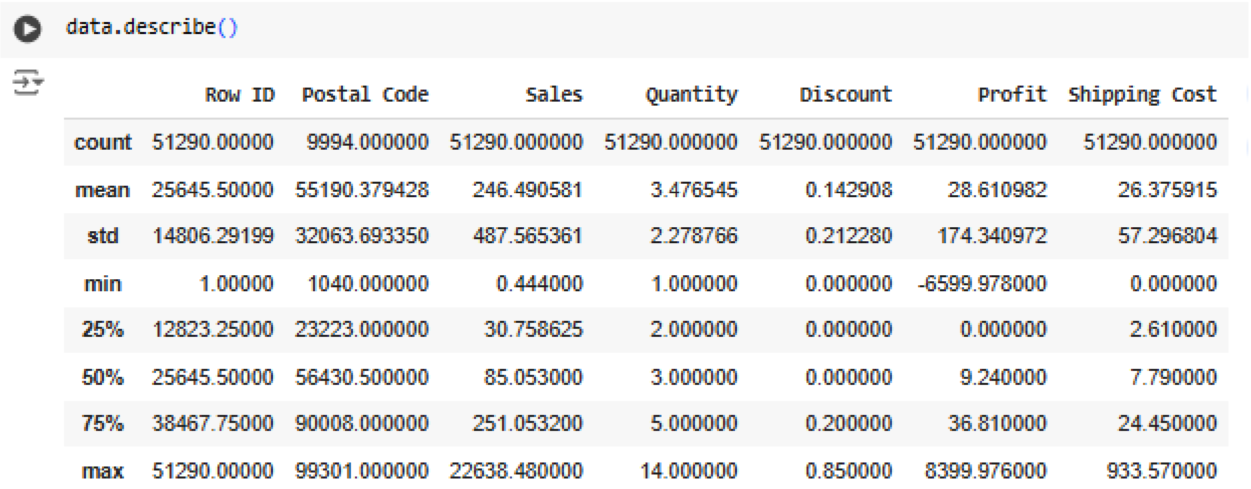

Total Sales by Region
1e6
2.5
2.0
1.5
1.0
0.5
0.0
Africa
Canada
Caribbean
Central
Central Asia
EMEA
East
North
North Asia
Oceania
South
Southeast Asia
West
Region


Sales vs Profit
8000
6000
4000
2000
0
−2000
−4000
−6000
Profit
0
5000
10000
15000
20000
Sales



Post-lab Questions:

1. **What is the output of data.describe() in Pandas, and what does it summarize?**



```
data.describe()
```

	Row ID	Postal Code	Sales	Quantity	Discount	Profit	Shipping Cost
count	51290.00000	9994.000000	51290.000000	51290.000000	51290.000000	51290.000000	51290.000000
mean	25645.50000	55190.379428	246.490581	3.476545	0.142908	28.610982	26.375915
std	14806.29199	32063.693350	487.565361	2.278766	0.212280	174.340972	57.296804
min	1.00000	1040.000000	0.444000	1.000000	0.000000	-6599.978000	0.000000
25%	12823.25000	23223.000000	30.758625	2.000000	0.000000	0.000000	2.610000
50%	25645.50000	56430.500000	85.053000	3.000000	0.000000	9.240000	7.790000
75%	38467.75000	90008.000000	251.053200	5.000000	0.200000	36.810000	24.450000
max	51290.00000	99301.000000	22638.480000	14.000000	0.850000	8399.976000	933.570000

The data.describe() function in Pandas provides a summary of descriptive statistics for numerical columns in a dataset. It includes the following metrics:

- Count: The number of non-missing values in each column
- Mean: The arithmetic average of the values
- Standard Deviation (std): Measures the spread of the data
- Minimum (min): The smallest value in the column

- 25% (Q1): The first quartile, where 25% of the values are below this point
- 50% (Median): The middle value, dividing the dataset into two equal halves
- 75% (Q3): The third quartile, where 75% of the values are below this point
- Maximum (max): The largest value in the column

2. Explain why the median is often preferred over the mean in datasets with outliers.

The median is often preferred over the mean in datasets with outliers because it is resistant to extreme values**.** While the mean takes into account all values, including outliers, which can skew the result significantly, the median represents the middle value when the data is ordered, making it less influenced by unusually high or low values. This makes the median a more accurate measure of central tendency in datasets with outliers or skewed distributions.

3. How does the .groupby() method enhance data analysis?

The .groupby() method in Pandas enhances data analysis by allowing you to group data based on specific criteria (such as a column or multiple columns) and then apply aggregation or transformation functions to each group. This enables more granular analysis, such as calculating summary statistics (mean, sum, count) for different subsets of the data. It simplifies the process of analyzing patterns and relationships within groups, for example, comparing sales by region or analyzing average profits by product category.

4. If the mode of the sales column is 200, what does this indicate about the dataset?

If the mode of the sales column is 200, it indicates that 200 is the most frequently occurring sales value in the dataset.

Done by:

Deepika Ganesan- 22BCS019

Durga S- 22BCS026

Nandhini Priya- 22BCS073

Nithiyaa T- 22BCS082

Priya Dharshini- 22BCS094

Rithika- 22BCS103

Sandhiya G- 22BCS109

Santhiya P 22BCS111