DMBI Lab

EXPERIMENT NO. 4

Name : Kaushik Kotian **Roll no.** 30 **Div :** D15B **Batch:**B

AIM : To perform exploratory data analysis and data visualization using python.
1. Descriptive analysis - statistical measures of data (Central tendency) 2.
Descriptive analysis - statistical measures of data (Dispersion)
3. Correlation between attributes
4. Different Visualization techniques and use of it.
Inferences derived after every analysis.

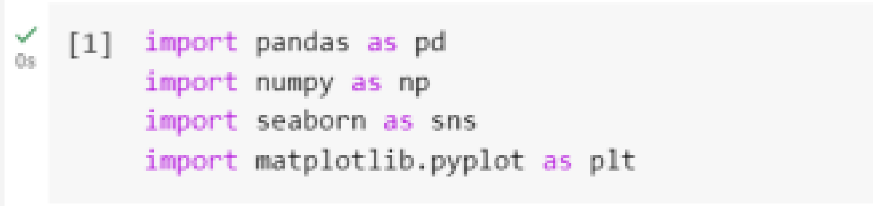
Theory :

Exploratory Data Analysis (EDA) is a pivotal initial step in data analysis, aimed at comprehensively understanding dataset characteristics through statistical measures and visualizations. Central tendency metrics like mean, median, and mode offer insights into typical data values, while dispersion metrics such as standard deviation and variance indicate data variability. Correlation analysis, utilizing Pearson and Spearman coefficients, reveals associations between variables. Visualization techniques such as histograms, box plots, scatter plots, and heatmaps provide intuitive representations of data distributions, outliers, and relationships. Through EDA, analysts derive valuable insights into dataset structures, trends, and potential anomalies, guiding subsequent analytical decisions and modeling processes.

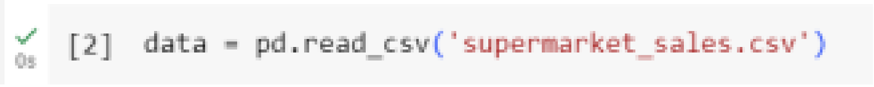
Link to the dataset : https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales

Steps:

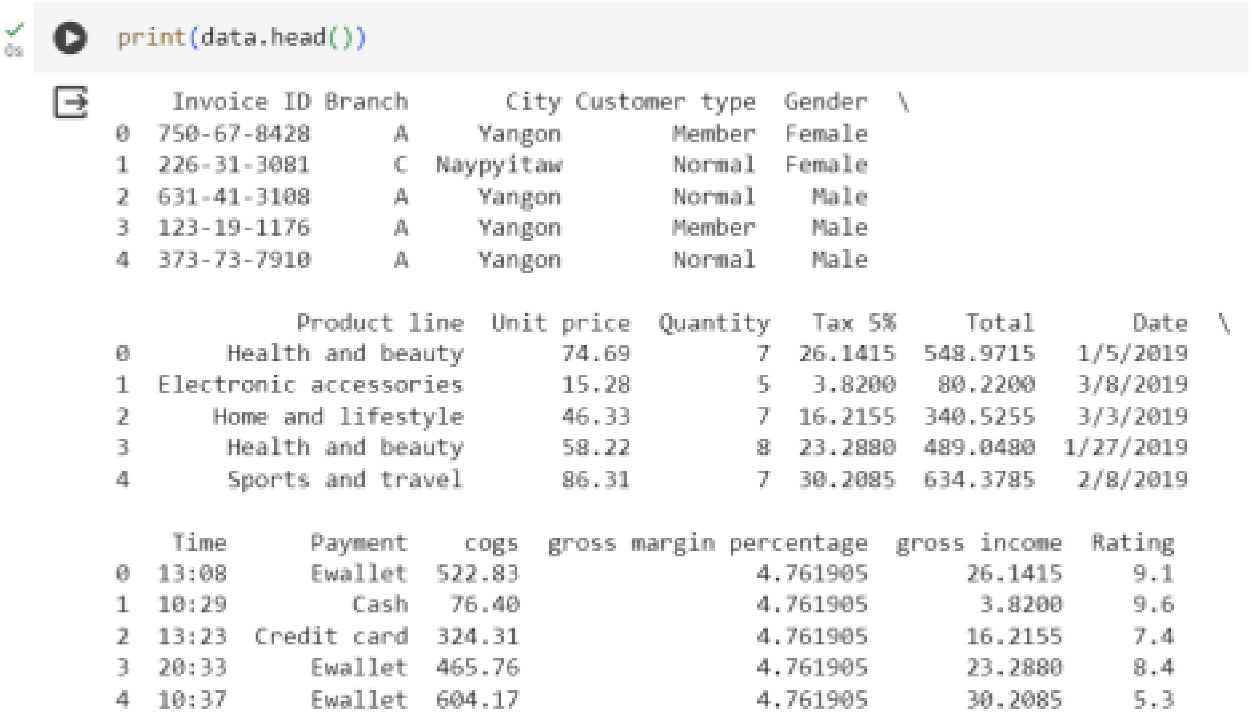
1. **Import necessary libraries and load the dataset.**



```
[1] import pandas as pd
    import numpy as np
    import seaborn as sns
    import matplotlib.pyplot as plt
```



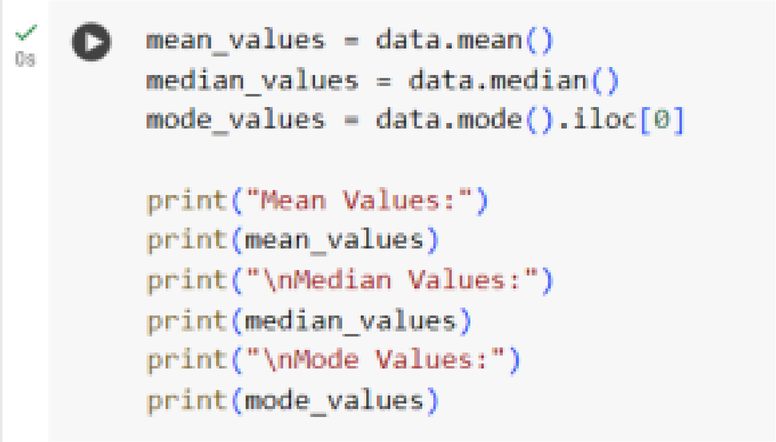
```
[2] data = pd.read_csv('supermarket_sales.csv')
```



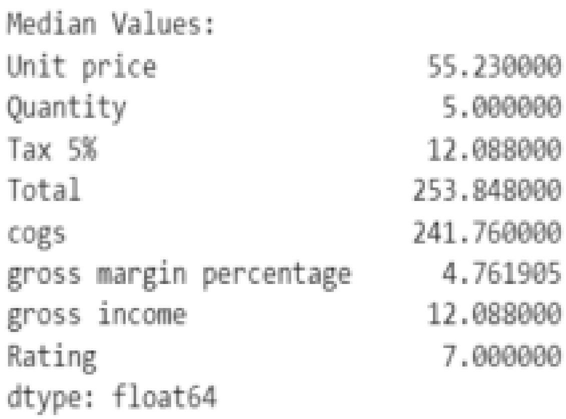
```
print(data.head())
```

```
    Invoice ID Branch       City Customer type  Gender  \
0  750-67-8428      A     Yangon        Member  Female
1  226-31-3081      C  Naypyitaw        Normal  Female
2  631-41-3108      A     Yangon        Normal    Male
3  123-19-1176      A     Yangon        Member    Male
4  373-73-7910      A     Yangon        Normal    Male

             Product line  Unit price  Quantity   Tax 5%     Total       Date  \
0       Health and beauty       74.69         7  26.1415  548.9715   1/5/2019
1  Electronic accessories       15.28         5   3.8200   80.2200   3/8/2019
2      Home and lifestyle       46.33         7  16.2155  340.5255   3/3/2019
3       Health and beauty       58.22         8  23.2880  489.0480  1/27/2019
4       Sports and travel       86.31         7  30.2085  634.3785   2/8/2019

    Time      Payment    cogs  gross margin percentage  gross income  Rating
0  13:08      Ewallet  522.83                 4.761905       26.1415     9.1
1  10:29         Cash   76.40                 4.761905        3.8200     9.6
2  13:23  Credit card  324.31                 4.761905       16.2155     7.4
3  20:33      Ewallet  465.76                 4.761905       23.2880     8.4
4  10:37      Ewallet  604.17                 4.761905       30.2085     5.3
```

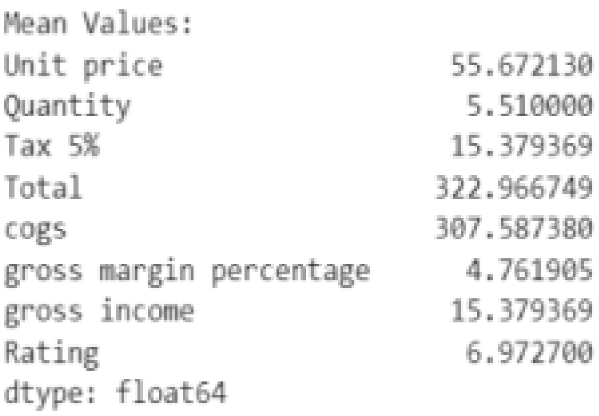
2. Descriptive analysis - Central tendency: Mean, Median, and Mode.



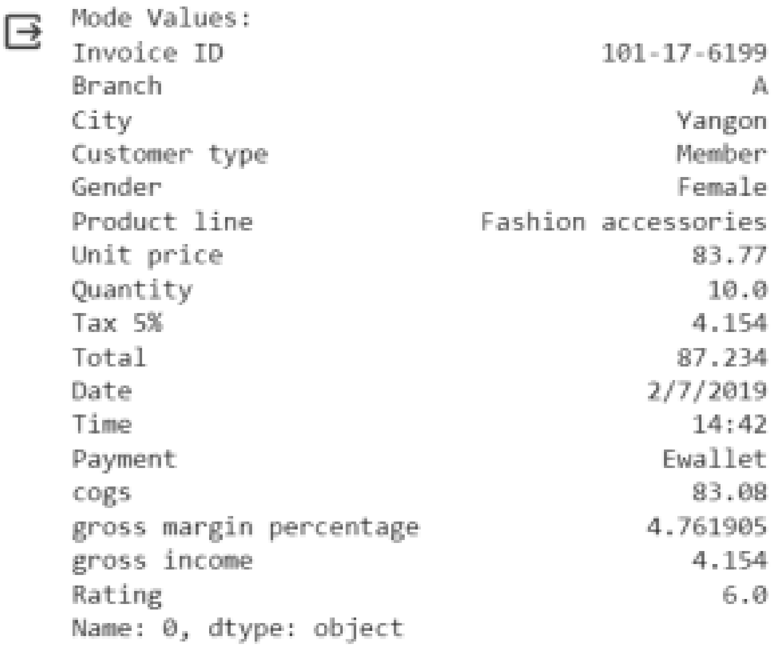
```
mean_values = data.mean()
median_values = data.median()
mode_values = data.mode().iloc[0]

print("Mean Values:")
print(mean_values)
print("\nMedian Values:")
print(median_values)
print("\nMode Values:")
print(mode_values)
```



```
Median Values:
Unit price                  55.230000
Quantity                     5.000000
Tax 5%                      12.088000
Total                      253.848000
cogs                       241.760000
gross margin percentage      4.761905
gross income                12.088000
Rating                       7.000000
dtype: float64
```

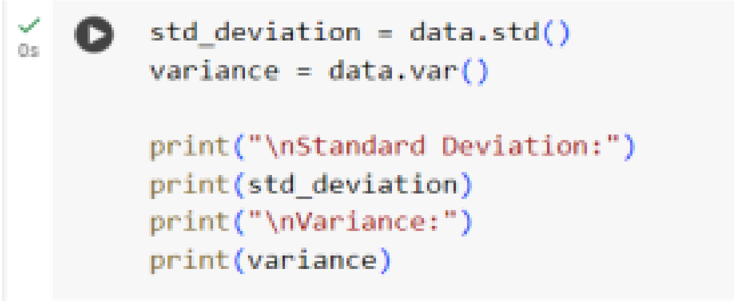


```
Mean Values:
Unit price                  55.672130
Quantity                     5.510000
Tax 5%                      15.379369
Total                      322.966749
cogs                       307.587380
gross margin percentage      4.761905
gross income                15.379369
Rating                       6.972700
dtype: float64
```

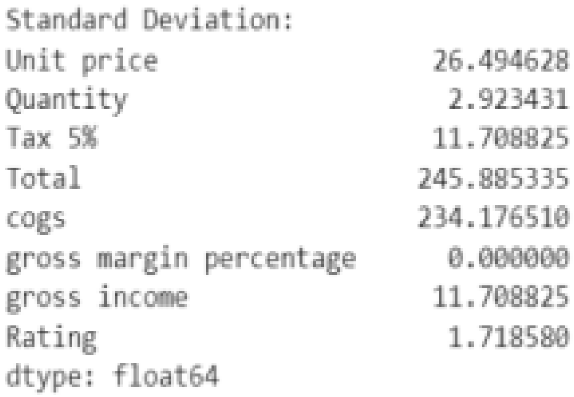


```
Mode Values:
Invoice ID                         101-17-6199
Branch                                       A
City                                    Yangon
Customer type                           Member
Gender                                  Female
Product line               Fashion accessories
Unit price                               83.77
Quantity                                  10.0
Tax 5%                                   4.154
Total                                   87.234
Date                                  2/7/2019
Time                                     14:42
Payment                                Ewallet
cogs                                     83.08
gross margin percentage               4.761905
gross income                             4.154
Rating                                     6.0
Name: 0, dtype: object
```

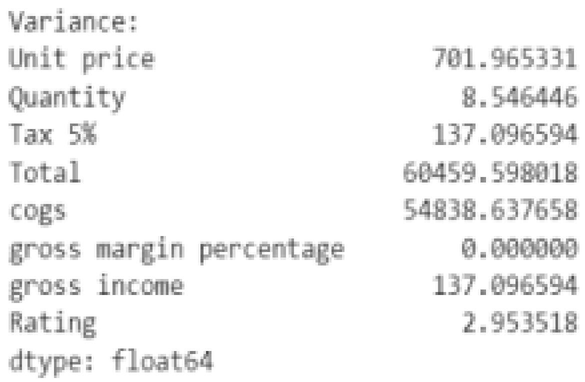
3. Descriptive analysis - Dispersion : Standard deviation and Variance



```
std_deviation = data.std()
variance = data.var()

print("\nStandard Deviation:")
print(std_deviation)
print("\nVariance:")
print(variance)
```



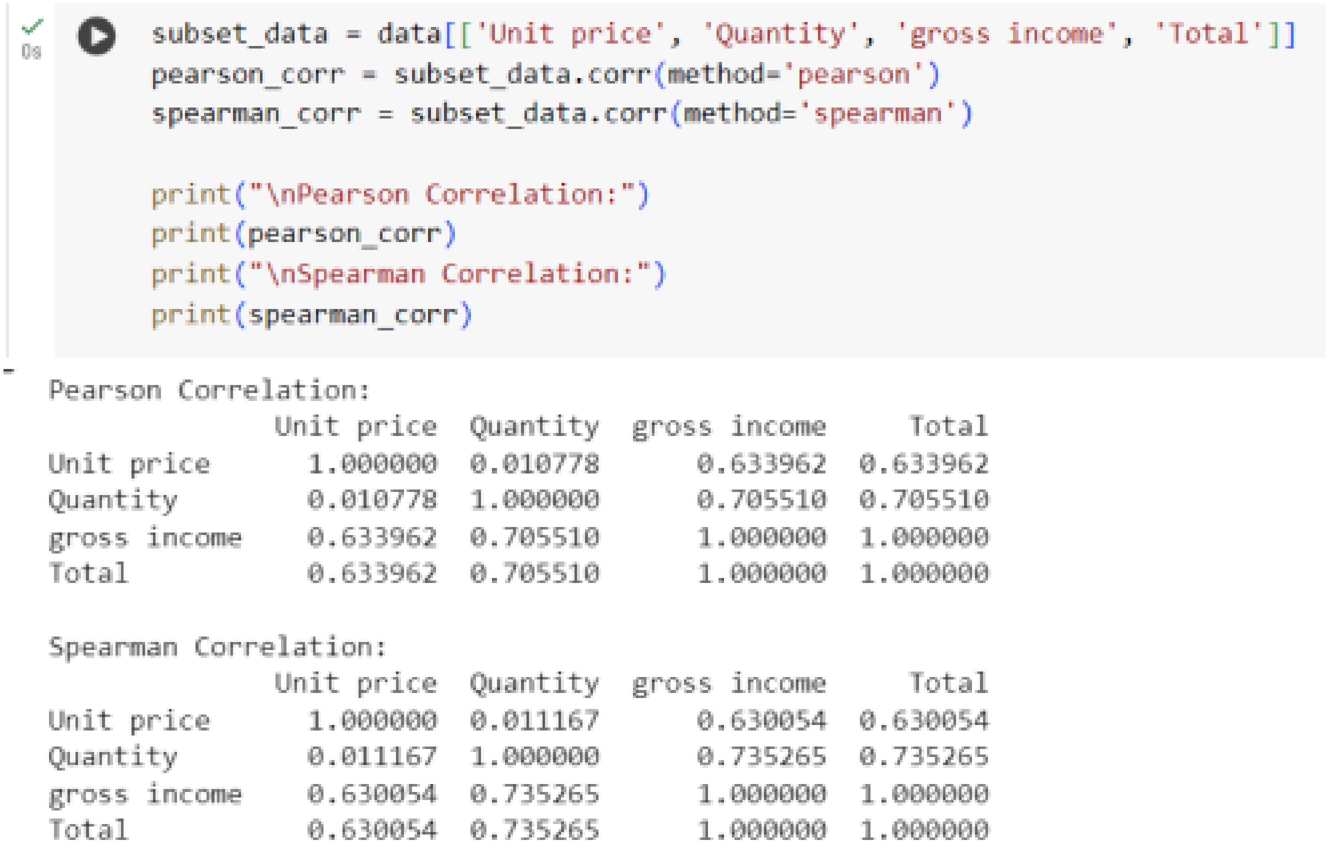
```
Standard Deviation:
Unit price                  26.494628
Quantity                     2.923431
Tax 5%                      11.708825
Total                      245.885335
cogs                       234.176510
gross margin percentage      0.000000
gross income                11.708825
Rating                       1.718580
dtype: float64
```



```
Variance:
Unit price                    701.965331
Quantity                        8.546446
Tax 5%                        137.096594
Total                       60459.598018
cogs                        54838.637658
gross margin percentage         0.000000
gross income                  137.096594
Rating                          2.953518
dtype: float64
```

Inference :
Standard Deviation: Indicates the amount of variation or dispersion from the mean.Here, Total price values are spread out over a wider range while Rating value shows little spreading.
Variance: Represents the average squared deviation from the mean. Here, gross margin percentage has same value as mean, while Total value has large deviation from mean.

4. Correlation between attributes : Pearson and Spearman correlation



```
subset_data = data[['Unit price', 'Quantity', 'gross income', 'Total']]
pearson_corr = subset_data.corr(method='pearson')
spearman_corr = subset_data.corr(method='spearman')

print("\nPearson Correlation:")
print(pearson_corr)
print("\nSpearman Correlation:")
print(spearman_corr)
```

```
Pearson Correlation:
              Unit price  Quantity  gross income     Total
Unit price      1.000000  0.010778      0.633962  0.633962
Quantity        0.010778  1.000000      0.705510  0.705510
gross income    0.633962  0.705510      1.000000  1.000000
Total           0.633962  0.705510      1.000000  1.000000

Spearman Correlation:
              Unit price  Quantity  gross income     Total
Unit price      1.000000  0.011167      0.630054  0.630054
Quantity        0.011167  1.000000      0.735265  0.735265
gross income    0.630054  0.735265      1.000000  1.000000
Total           0.630054  0.735265      1.000000  1.000000
```

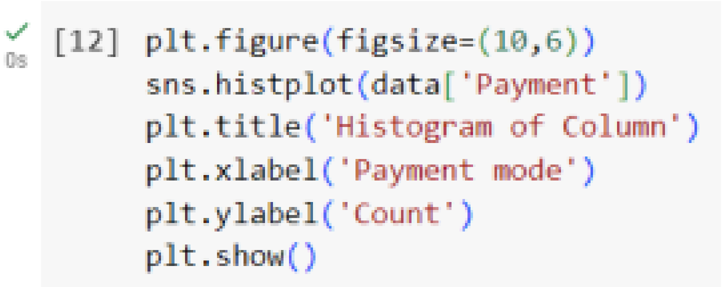
Inference :

Pearson Correlation: Measures the linear correlation between two continuous variables. Unit price and gross income are highly correlated while unit price and quantity are less correlated.

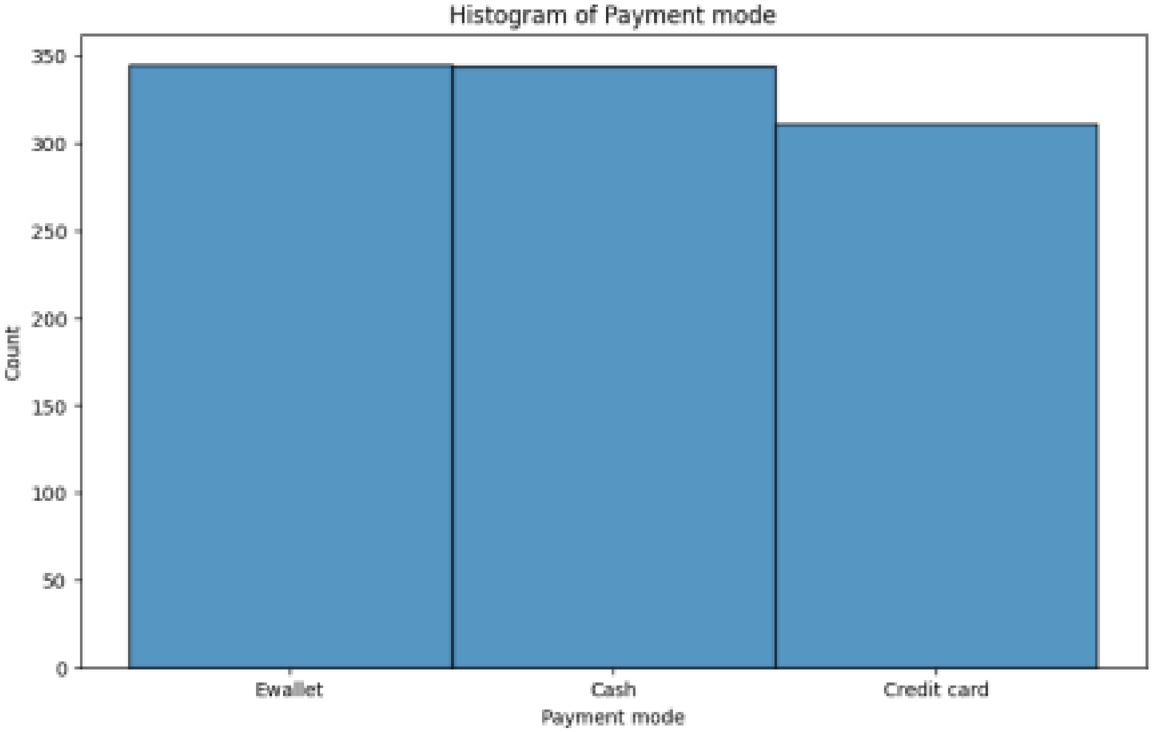
Spearman Correlation: Measures the strength and direction of association between two ranked variables. Here, gross income and total are closely related while unit price and quantity are not.

5. Visualization techniques

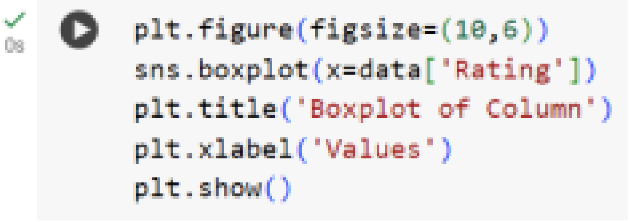
1. Histogram :



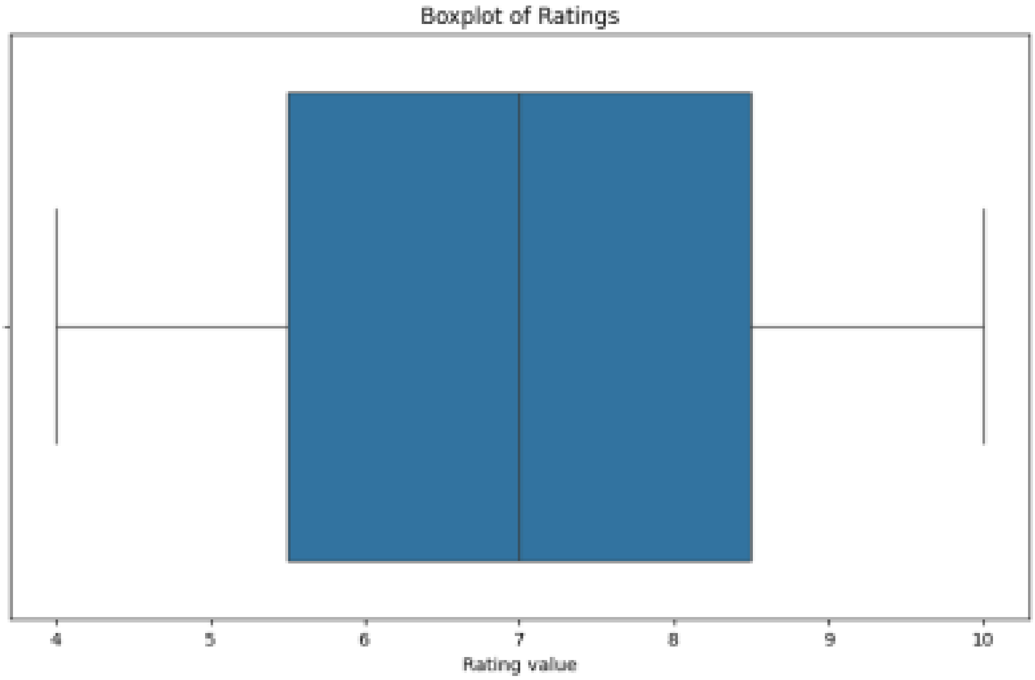
```
[12] plt.figure(figsize=(10,6))
     sns.histplot(data['Payment'])
     plt.title('Histogram of Column')
     plt.xlabel('Payment mode')
     plt.ylabel('Count')
     plt.show()
```



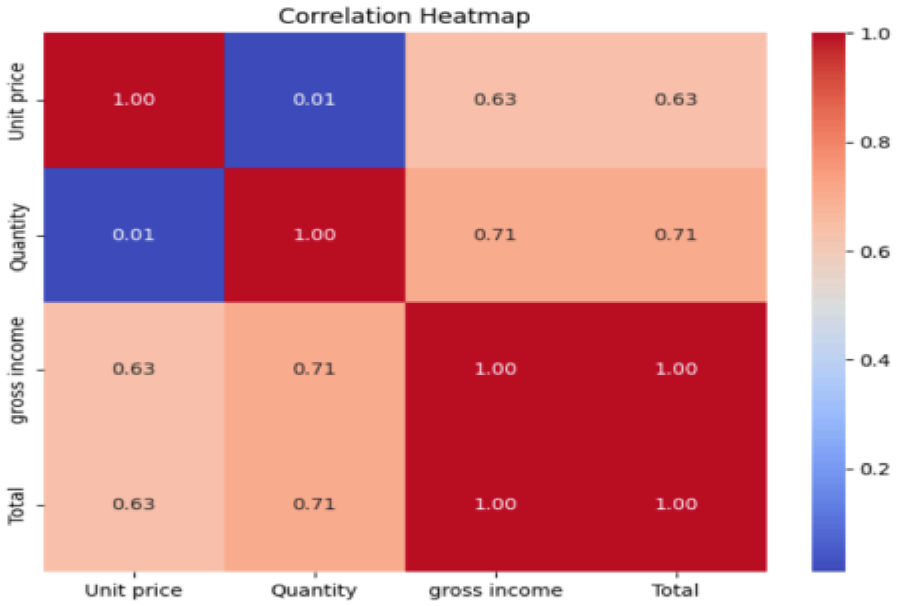
2. Box plot :



```
plt.figure(figsize=(10,6))
sns.boxplot(x=data['Rating'])
plt.title('Boxplot of Column')
plt.xlabel('Values')
plt.show()
```



3. Heatmap:



Inference :

Histogram: Provides a graphical representation of the distribution of numerical data. Here, It helps to understand the frequency distribution of Payment mode type.

Box plot: Displays the distribution of Rating data through quartiles. It's useful for detecting outliers and comparing distributions between different groups.

Heatmap: Visualizes the correlation matrix between variables. Here, It identifies the correlation between unit price,gross income, total,and quantity values.

CONCLUSION : Hence we have performed Exploratory data analysis (EDA) on our chosen dataset of Supermarket Sales, and also performed Data visualization using Python on the dataset.