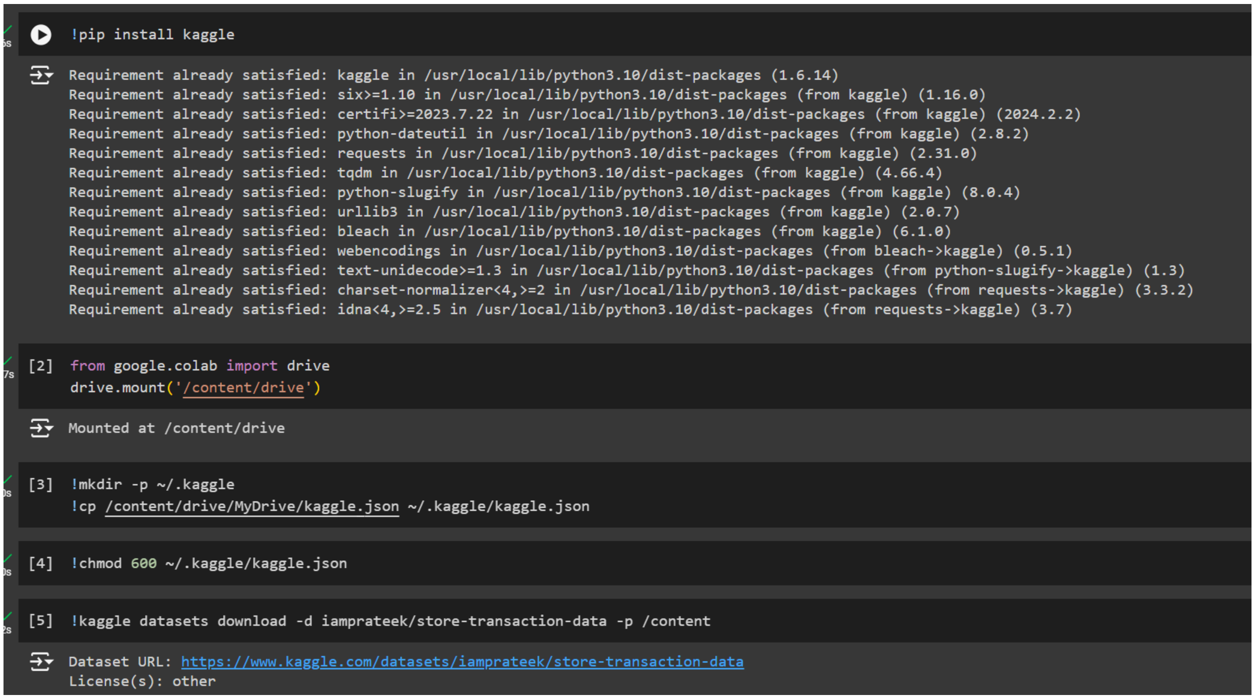
Comprehensive Analysis of social media data to understand user behavior and engagement

Introduction

This project aims to perform Exploratory Data Analysis (EDA) to derive insights into customer preferences, purchasing patterns, and overall behavior using the store transaction dataset. The dataset includes various details such as sales unit, sales value, product categories, brands, and transaction information. By analyzing this data, we can understand how users interact with content, identify patterns, and derive actionable insights for improving social media strategies.

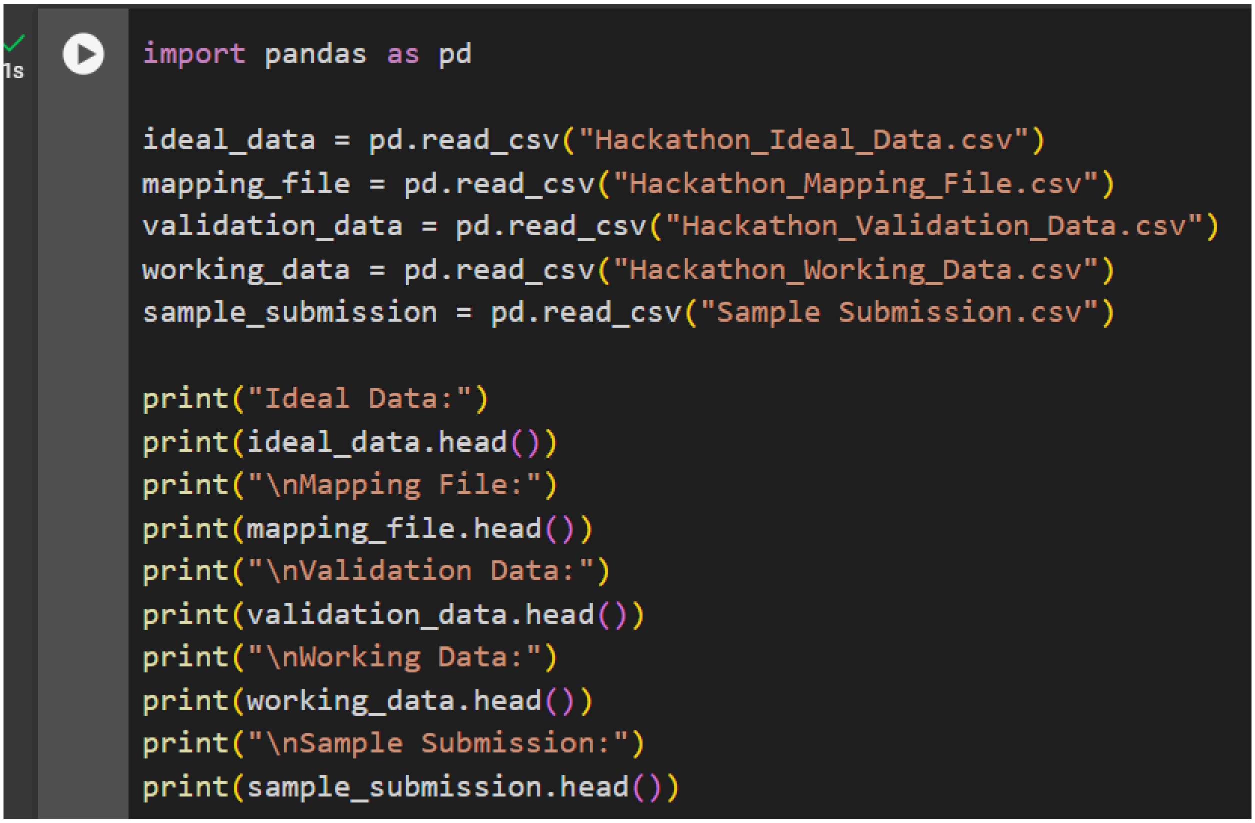
Data Preprocessing

Loading the Data



```
!pip install kaggle

Requirement already satisfied: kaggle in /usr/local/lib/python3.10/dist-packages (1.6.14)
Requirement already satisfied: six>=1.10 in /usr/local/lib/python3.10/dist-packages (from kaggle) (1.16.0)
Requirement already satisfied: certifi>=2023.7.22 in /usr/local/lib/python3.10/dist-packages (from kaggle) (2024.2.2)
Requirement already satisfied: python-dateutil in /usr/local/lib/python3.10/dist-packages (from kaggle) (2.8.2)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from kaggle) (2.31.0)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from kaggle) (4.66.4)
Requirement already satisfied: python-slugify in /usr/local/lib/python3.10/dist-packages (from kaggle) (8.0.4)
Requirement already satisfied: urllib3 in /usr/local/lib/python3.10/dist-packages (from kaggle) (2.0.7)
Requirement already satisfied: bleach in /usr/local/lib/python3.10/dist-packages (from kaggle) (6.1.0)
Requirement already satisfied: webencodings in /usr/local/lib/python3.10/dist-packages (from bleach->kaggle) (0.5.1)
Requirement already satisfied: text-unidecode>=1.3 in /usr/local/lib/python3.10/dist-packages (from python-slugify->kaggle) (1.3)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->kaggle) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->kaggle) (3.7)

[2] from google.colab import drive
    drive.mount('/content/drive')

Mounted at /content/drive

[3] !mkdir -p ~/.kaggle
    !cp /content/drive/MyDrive/kaggle.json ~/.kaggle/kaggle.json

[4] !chmod 600 ~/.kaggle/kaggle.json

[5] !kaggle datasets download -d iamprateek/store-transaction-data -p /content

Dataset URL: https://www.kaggle.com/datasets/iamprateek/store-transaction-data
License(s): other
```

I first inserted the dataset directly into colab using the API token from kaggle to my drive and then downloading and unzipping it. The 5 csv files thus obtained, I printed the first 5 elements of them using the head() function

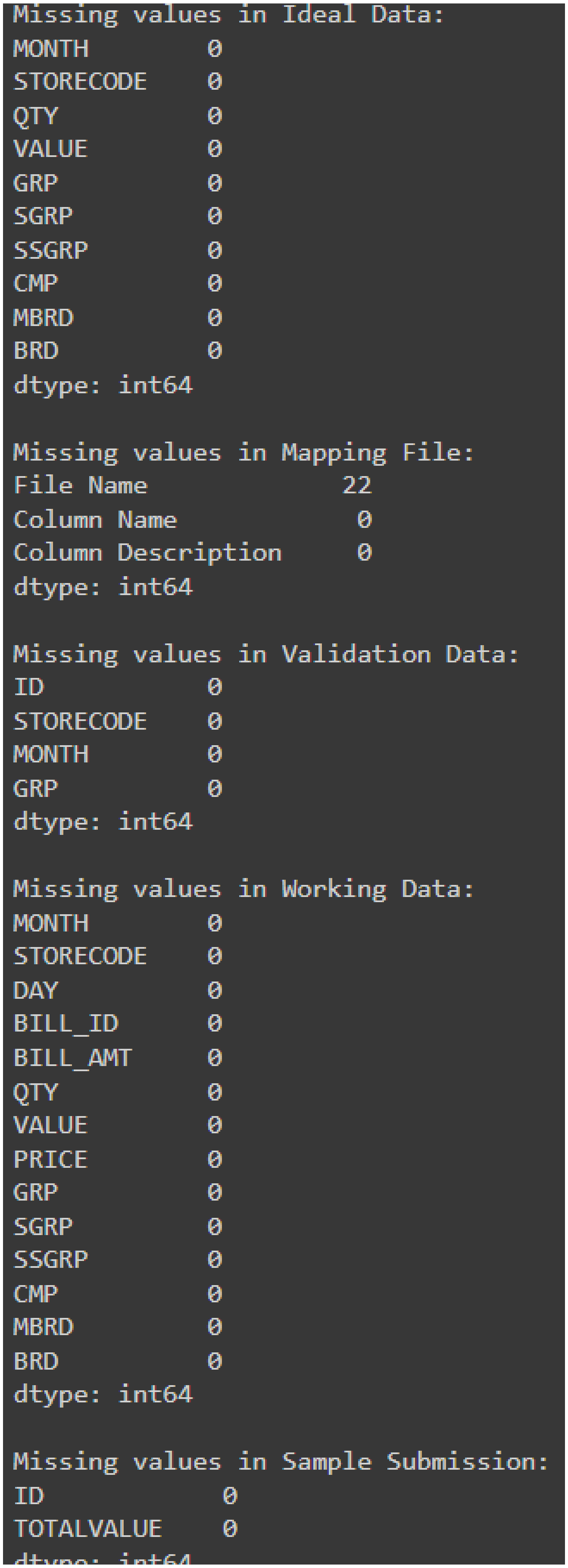


```python
import pandas as pd

ideal_data = pd.read_csv("Hackathon_Ideal_Data.csv")
mapping_file = pd.read_csv("Hackathon_Mapping_File.csv")
validation_data = pd.read_csv("Hackathon_Validation_Data.csv")
working_data = pd.read_csv("Hackathon_Working_Data.csv")
sample_submission = pd.read_csv("Sample Submission.csv")

print("Ideal Data:")
print(ideal_data.head())
print("\nMapping File:")
print(mapping_file.head())
print("\nValidation Data:")
print(validation_data.head())
print("\nWorking Data:")
print(working_data.head())
print("\nSample Submission:")
print(sample_submission.head())
```

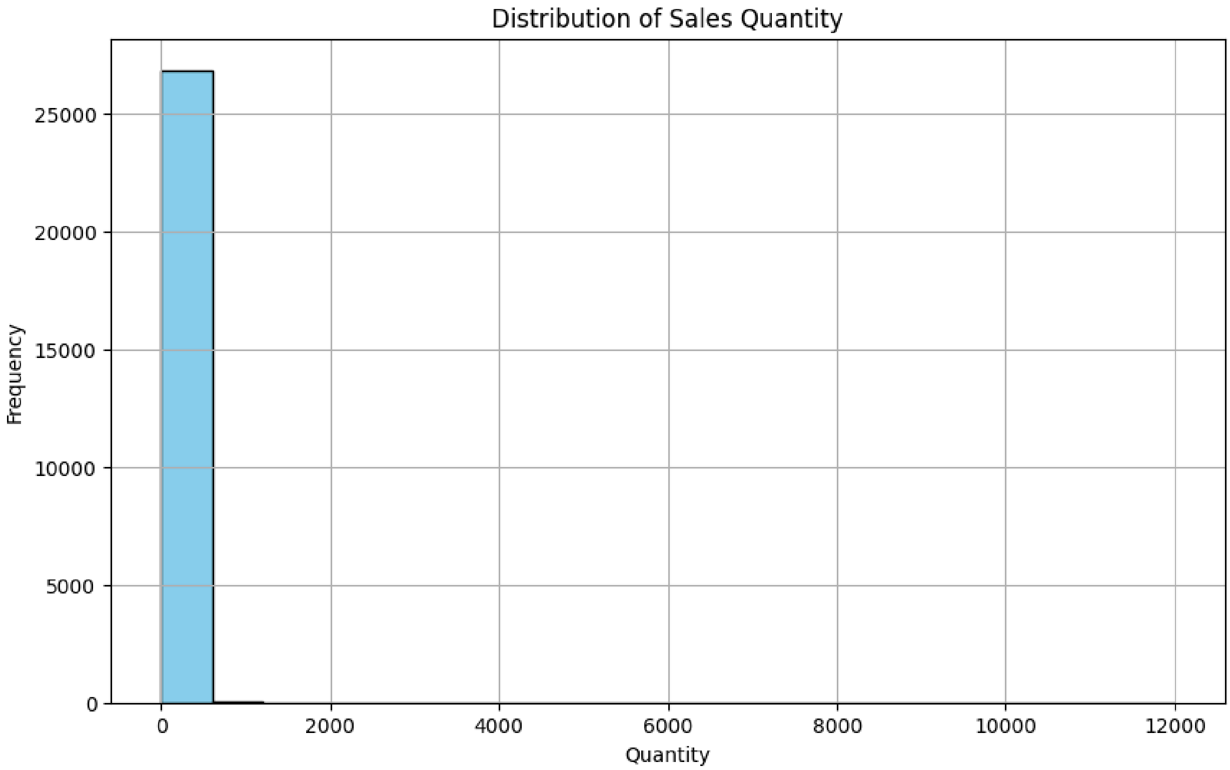
Handling Missing Values

I then checked for missing values in each dataset and handled them. Fortunately, there were no missing values in the dataset.

Also, I handled all the duplicate values in the dataset, by removing them. Just for my knowledge, I manually checked the values in the dataset next as well.



```
Missing values in Ideal Data:
MONTH         0
STORECODE     0
QTY           0
VALUE         0
GRP           0
SGRP          0
SSGRP         0
CMP           0
MBRD          0
BRD           0
dtype: int64

Missing values in Mapping File:
File Name              22
Column Name             0
Column Description      0
dtype: int64

Missing values in Validation Data:
ID            0
STORECODE     0
MONTH         0
GRP           0
dtype: int64

Missing values in Working Data:
MONTH         0
STORECODE     0
DAY           0
BILL_ID       0
BILL_AMT      0
QTY           0
VALUE         0
PRICE         0
GRP           0
SGRP          0
SSGRP         0
CMP           0
MBRD          0
BRD           0
dtype: int64

Missing values in Sample Submission:
ID             0
TOTALVALUE     0
dtype: int64
```



Next, there are some histograms that I have printed using matplotlib. This one represents the distribution of sales quantities in the dataset.

Distribution Shape:

- The shape of the histogram gives insights into the distribution of sales quantities. For example, if the histogram is long tailed to the right, it means that most transactions involve smaller quantities, but there are a few transactions with very large quantities.
- Conversely, if the histogram is the opposite, it means most transactions involve larger quantities.

Central Tendency:

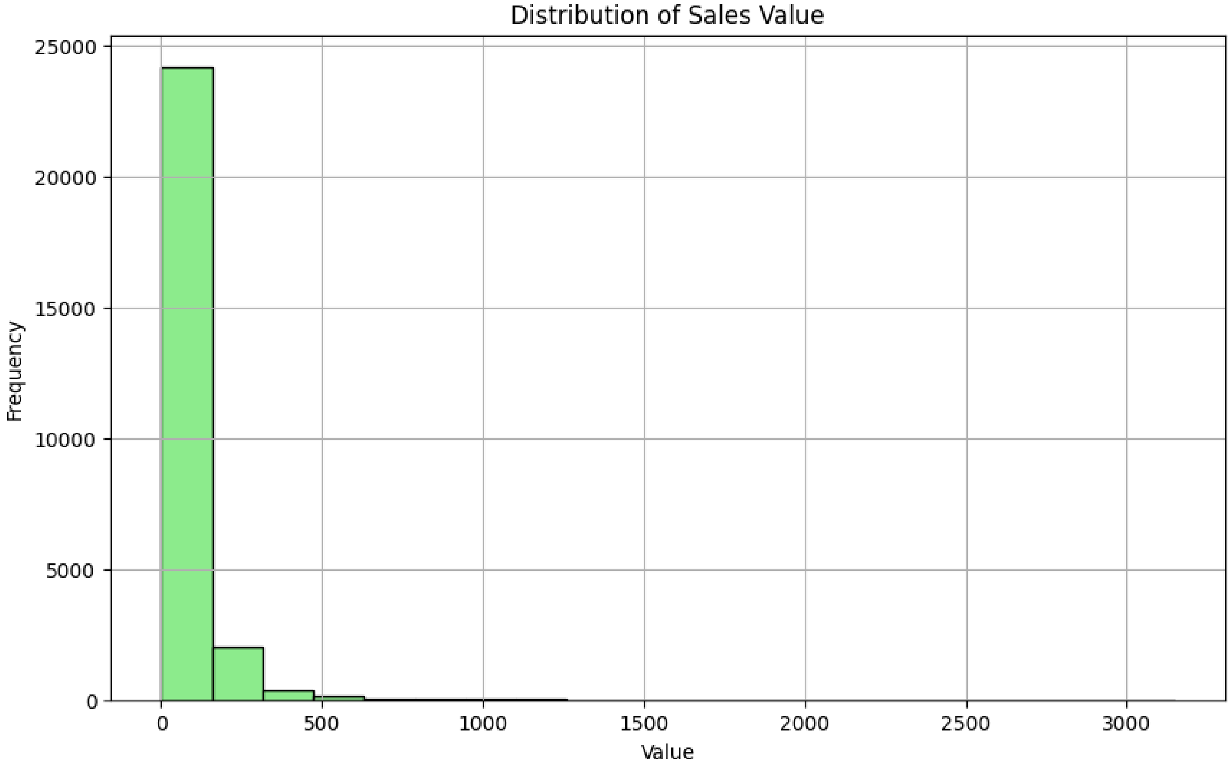
- The position of the tallest bars indicates the most common range of quantities sold. This gives an idea of the typical sales quantity in transactions.

Variability:

- The spread of the bars along the x-axis indicates the variability in sales quantities. A wider spread suggests more variation in the quantities sold across transactions.

Outliers:

- Bins that are significantly separated from the others might indicate outliers, i.e., transactions with unusually high or low sales quantities compared to the rest of the data.



This graph represents the distribution of sales value within the dataset.

Sales Value Distribution:

- The x-axis represents different ranges or bins of sales value. Each bin groups a range of sales values together.
- The y-axis represents the number of transactions that fall within each range of sales value.

Frequency of Sales:

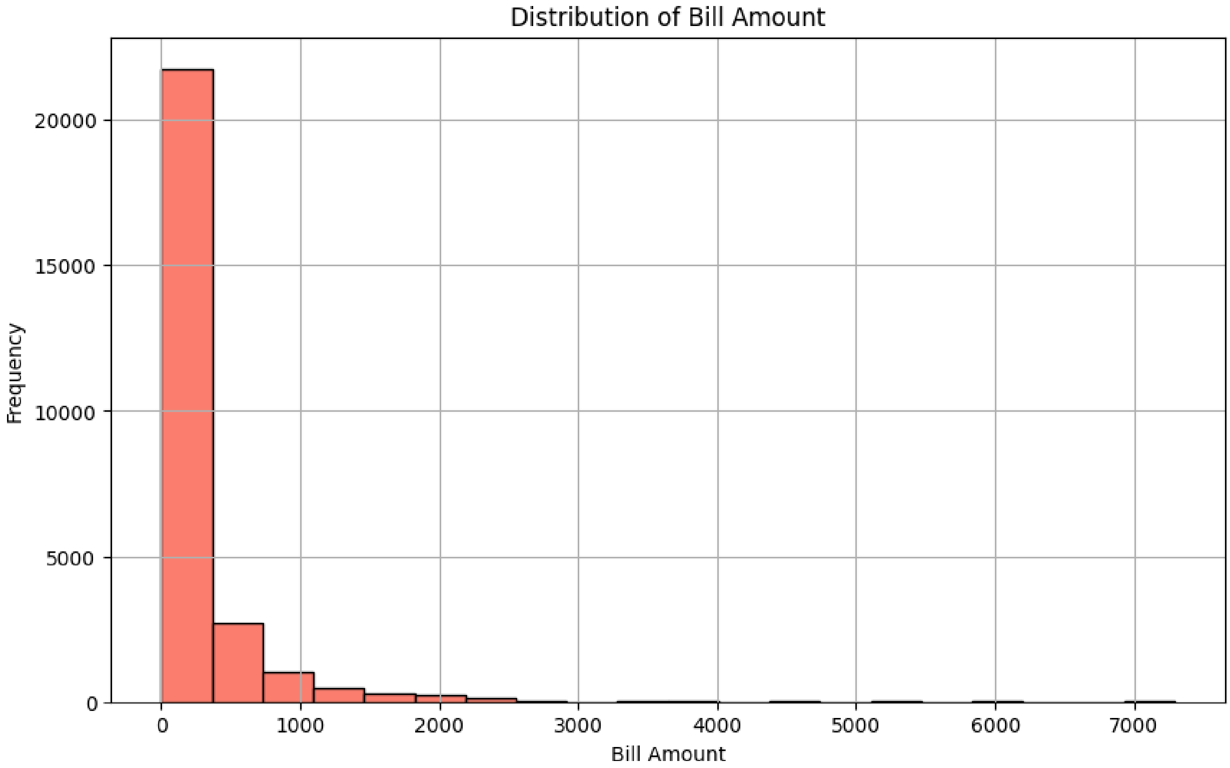
- The height of each bar indicates the frequency of transactions that had a sales value within the corresponding range. Taller bars indicate a higher frequency of transactions with those sales values.

Understanding Sales Behavior:

- By examining the shape of the histogram, we see how sales values are distributed across transactions.
- For example, if most of the bars are clustered towards the lower end of the sales value range, it indicates that most transactions have relatively low sales values.
- Conversely, if the bars are more spread out or clustered towards the higher end, it indicates that there are significant numbers of higher-value transactions.

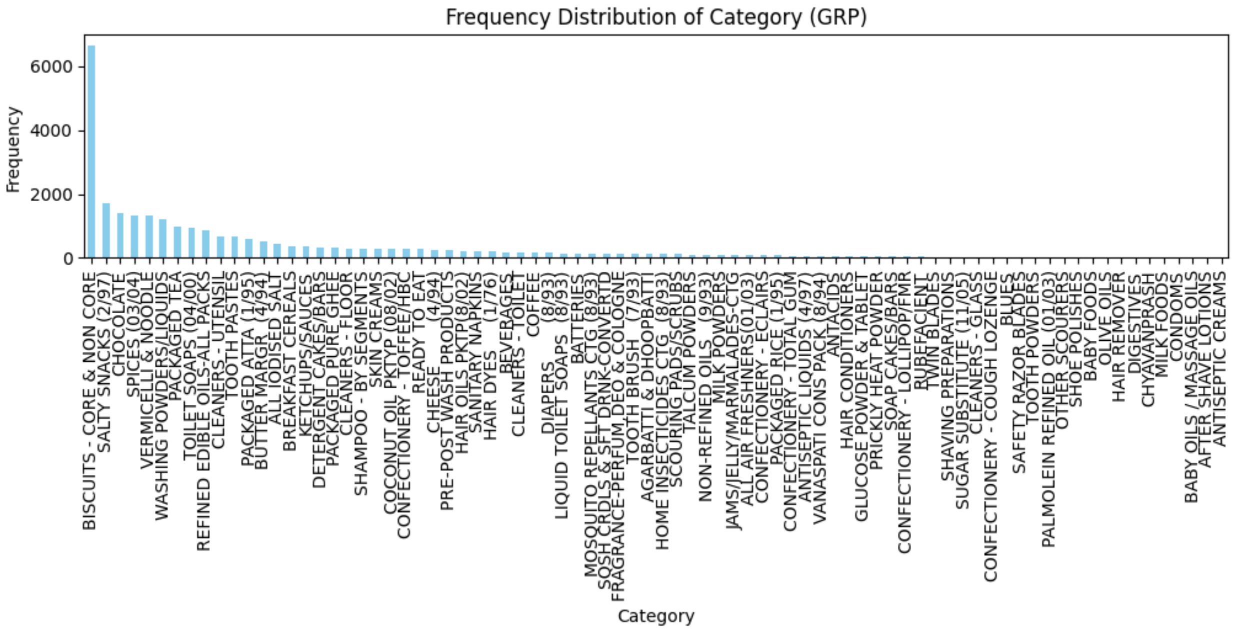
Identifying Patterns and Anomalies:

- The histogram helps in identifying any patterns or anomalies in the sales data. For example, if there is a bar that is significantly higher than the others, it may indicate a common sales value or a promotional period where many transactions occurred at that value.
- Gaps or bars with very low heights could indicate sales values that are less common or even errors in the data.



This graph represents the distribution of the bill amount variable in the dataset. Specifically, it is a histogram that visualizes how bill amounts are distributed across different ranges.

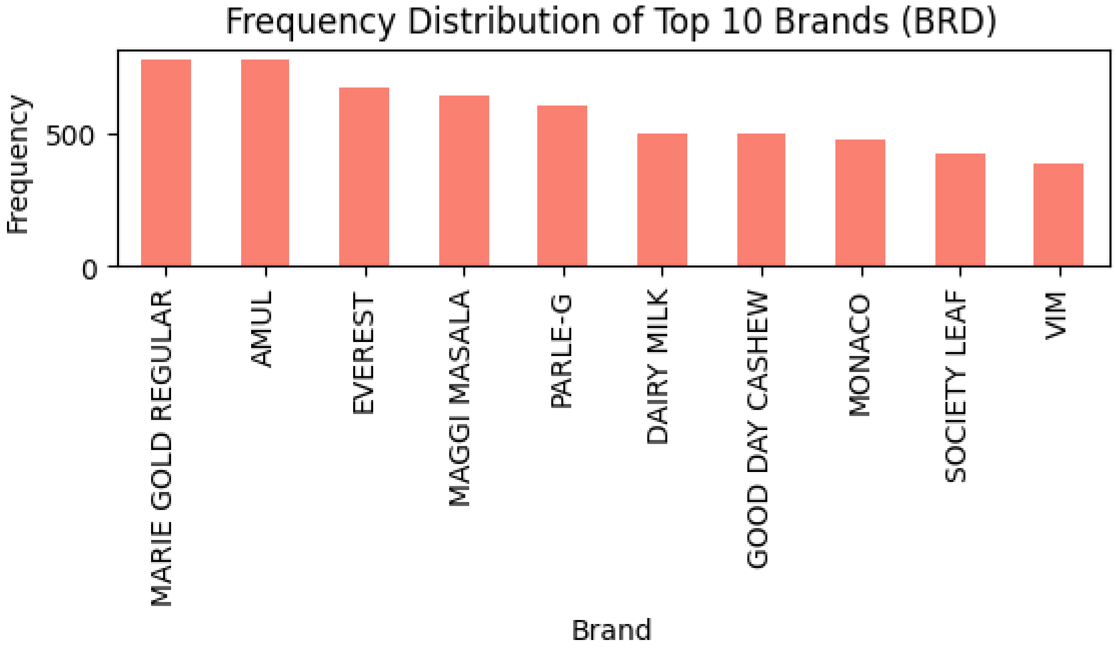
1. **Shape of Distribution**: By examining the shape of the histogram, we can infer the distribution pattern of bill amounts. For example, if the histogram has a longer tail on the right side, it indicates that there are a few transactions with very high bill amounts, but most transactions have lower bill amounts.
2. **Central Tendency**: The peak of the histogram indicates the most common range of bill amounts. This can help identify the typical bill amount most customers tend to have.
3. **Spread of Data**: The width of the distribution shows the range of bill amounts. A wider distribution indicates more variability in bill amounts, while a narrower distribution suggests that bill amounts are more consistent.
4. **Outliers**: Any bins that are isolated from the rest of the data may indicate outliers or exceptionally high bill amounts.



The graph represented by the code snippet is a bar chart showing the frequency distribution of the GRP column in the dataset.

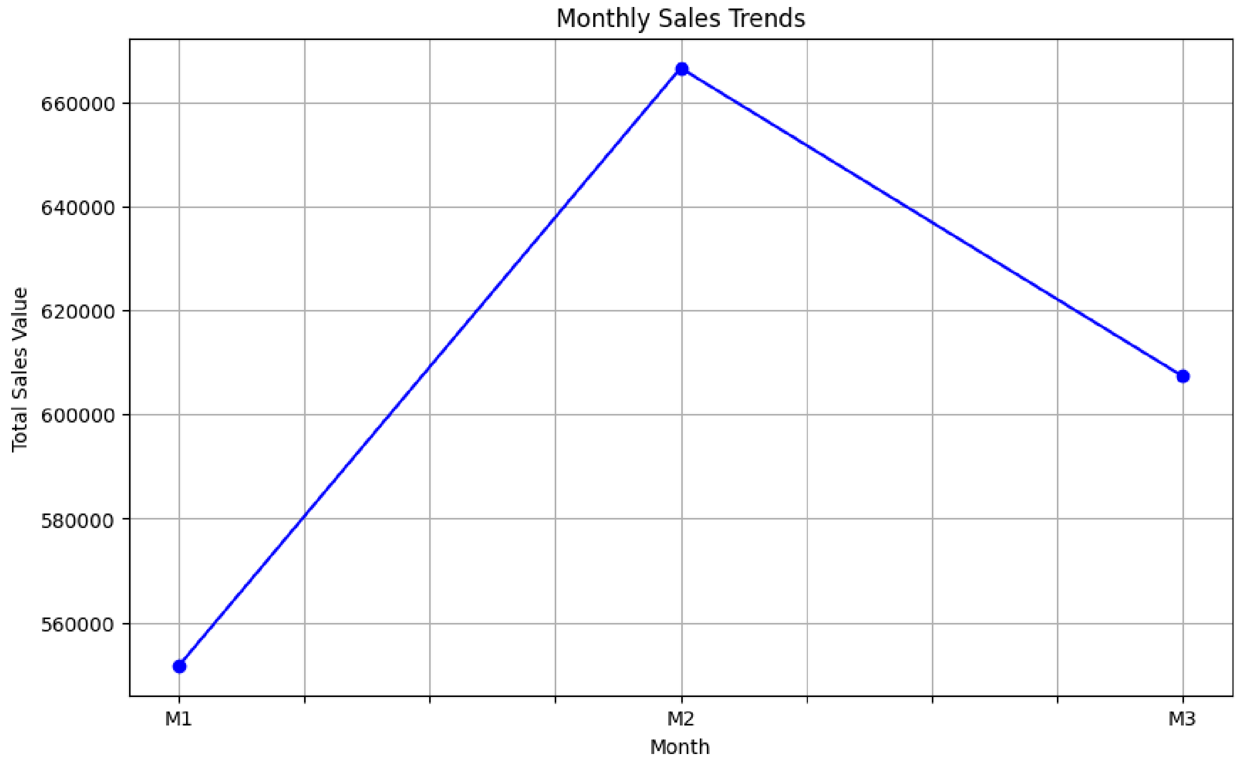
1. **X-axis**: This axis represents different product categories GRP present in the dataset.

2. **Y-axis**: This axis represents the number of occurrences or the count of each category in the dataset.
3. **Bars**: Each bar represents a different product category, and the height of the bar indicates the number of times that category appears in the dataset.



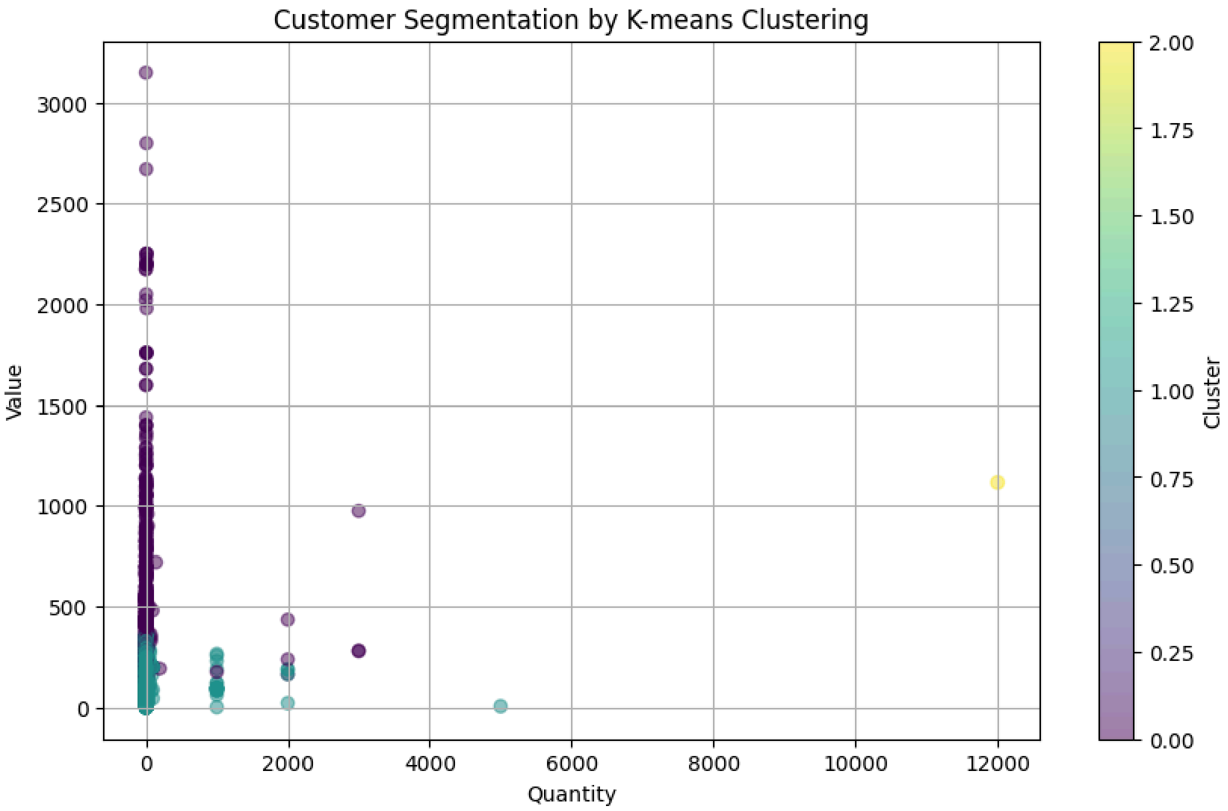
This graph represents the frequency distribution of the top 10 brands BRD in the dataset.

1. **Most Frequent Brands**: The graph highlights which brands are the most frequently occurring in the dataset. This can indicate popular brands or those that are sold most often in the store.
2. **Comparison**: By comparing the height of the bars, can quickly tell us which brands are more popular relative to others within the top 10.
3. **Distribution**: The graph provides a visual representation of the distribution of brand frequencies, if the market is more evenly distributed among several brands.



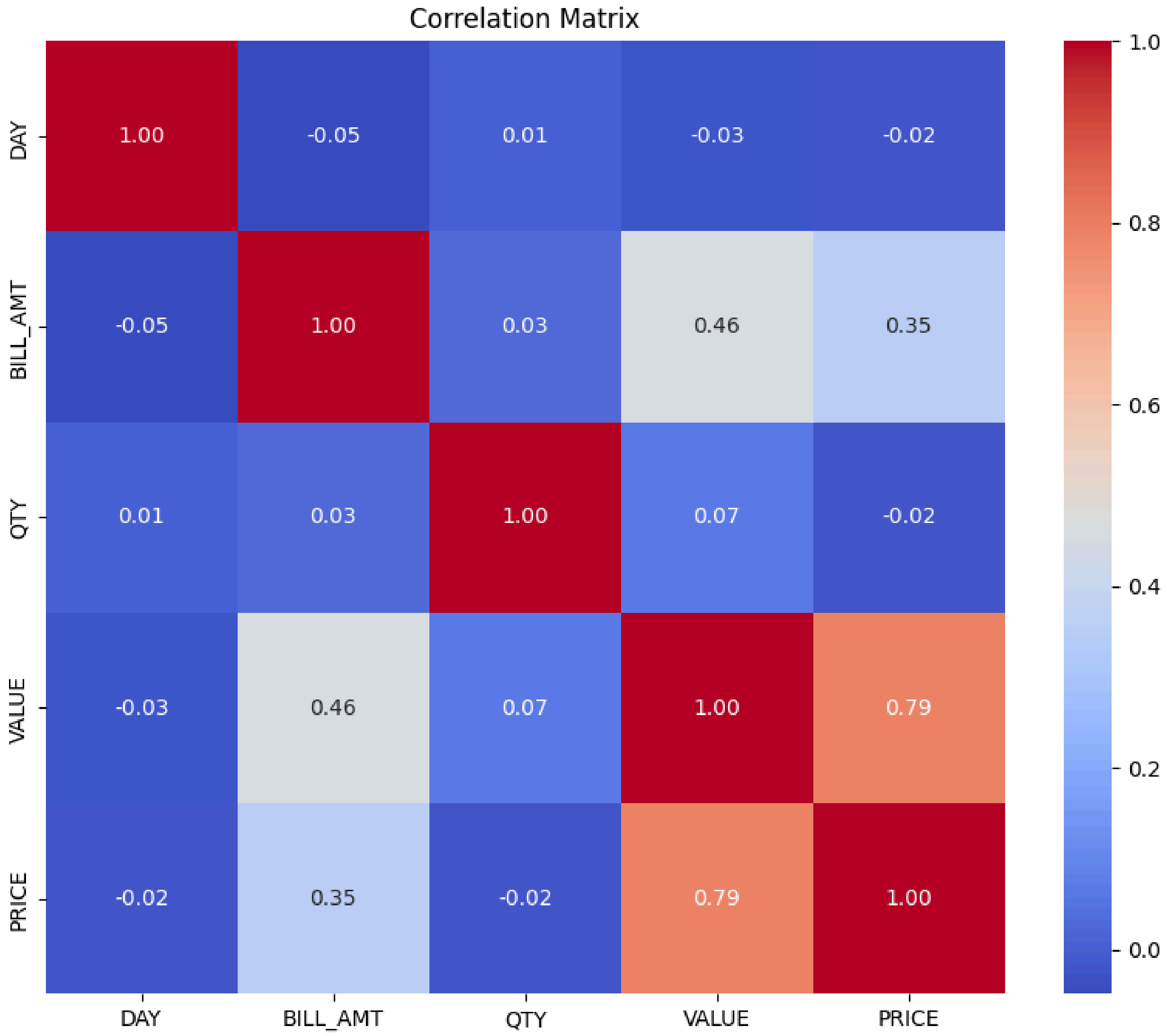
The graph represents the trends in total sales value over the different months.

1. **X-axis**: The x-axis represents the different months, denoted as 'M1', 'M2', etc.
2. **Y-axis**: The y-axis represents the total sales value for each month.
3. **Line Plot**: The line plot indicates the actual data points.
4. **Trends in Sales Over Time**: The graph shows how the total sales value changes from one month to the next.
    a. If the line is trending upwards, it indicates an increase in sales value over the months.
    b. If the line is trending downwards, it indicates a decrease in sales value.
    c. Fluctuations in the line indicate variability in sales from month to month.
    d. Peaks could represent high sales periods, potentially due to promotions, holidays, or seasonal demand.
    e. Troughs could indicate low sales periods.



1. **Axes**:
   a. The x-axis represents the quantity of items sold.
   b. The y-axis represents the sales value.
2. **Color Coding**:
   a. Each point represents a transaction.
   b. Points are color-coded based on the cluster they belong to (determined by the K-means algorithm).
3. **Clusters**:
   a. The different colors in the plot represent different clusters, indicating groups of transactions with similar patterns in terms of quantity and value.
   b. The K-means algorithm has divided the transactions into three clusters, each representing a distinct customer segment based on their purchasing behavior.
4. **Insights**:
   a. The clusters can help identify different customer segments. For example, one cluster might represent high-value, low-quantity purchases, another might represent low-value, high-quantity purchases, and a third might represent average-value, average-quantity purchases.

b. This segmentation can be useful for targeted marketing strategies, personalized promotions, and understanding customer behavior patterns.



The correlation matrix heatmap represents the pairwise correlation coefficients between the numerical variables in the dataset.

1. **Diagonal Elements**: The diagonal elements of the matrix represent the correlation of each variable with itself, which is always 1.0.
2. **Off-Diagonal Elements**: The off-diagonal elements represent the correlation between different pairs of variables.
3. **Color Coding**: The color of each cell represents the strength and direction of the correlation. For example:
    a. Dark red cells may represent high positive correlation (close to 1).

b. Dark blue cells may represent high negative correlation (close to -1).
c. White or light-colored cells may represent low or no correlation (close to 0).

Conclusion

In this project, we conducted a comprehensive analysis of social media data to gain insights into user behavior and engagement. We started with data preprocessing, ensuring data quality and consistency. We then performed exploratory data analysis (EDA) to understand the distribution and relationships between variables.

Additionally, we conducted advanced analyses such as customer retention analysis, price elasticity analysis, and customer lifetime value (CLV) analysis. These analyses provided valuable insights into customer preferences, purchasing patterns, and overall behavior, helping to inform data-driven strategies for improving social media engagement and business performance.