

Experiment No 2

Aim: Data Visualization/ Exploratory data Analysis using Matplotlib and Seaborn.

Theory:

Bar graph & contingency table using any two features :

A bar graph is a visual representation of categorical data, where rectangular bars display the count or frequency of a particular variable. The bar graph illustrates the distribution of acceptable street percentages across different boroughs in New York City. Each bar represents a borough, and its height indicates the number of acceptable streets recorded. This visualization helps compare borough-wise street conditions and identify areas with higher or lower acceptable street percentages.

1. Create bar graph, contingency table using any 2 features.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

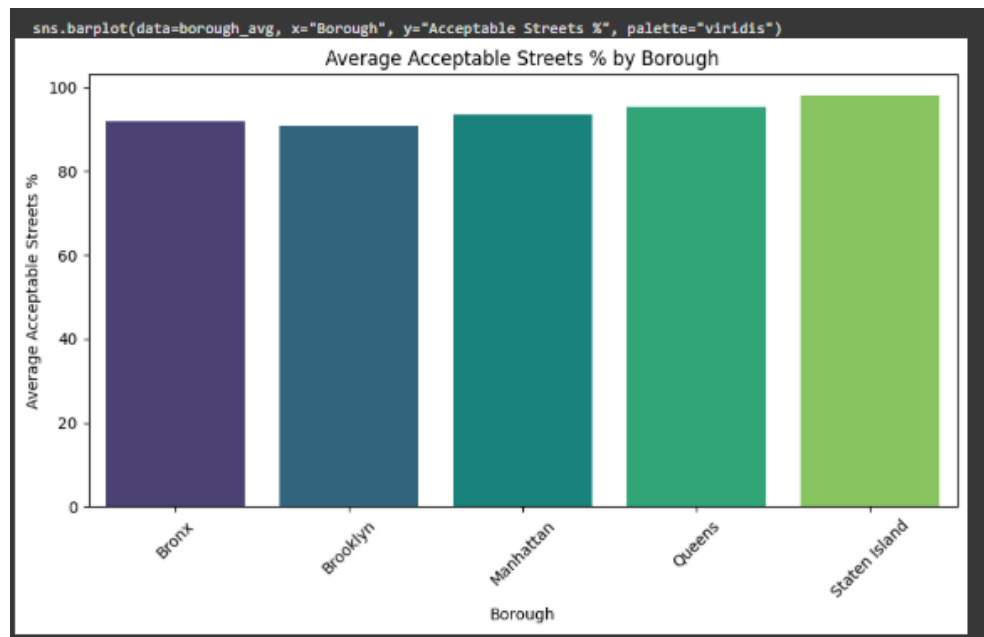
# Load dataset
file_path = "/content/scorecard.csv"
df = pd.read_csv(file_path)

# Drop missing values
df_clean = df.dropna(subset=["Borough", "Acceptable Streets %"])

# Calculate mean values separately
borough_avg = df_clean.groupby("Borough")["Acceptable Streets %"].mean().reset_index()

# Bar plot: Average "Acceptable Streets %" by Borough
plt.figure(figsize=(10, 5))
sns.barplot(data=borough_avg, x="Borough", y="Acceptable Streets %", palette="viridis")
plt.xticks(rotation=45)
plt.title("Average Acceptable Streets % by Borough")
plt.xlabel("Borough")
plt.ylabel("Average Acceptable Streets %")
plt.show()

# Contingency table: Borough vs. District
contingency_table = pd.crosstab(df["Borough"], df["District"])
print(contingency_table)
```



Contingency table: A contingency table provides a cross-tabulated view of data by showing the frequency of one variable against another. It displays the relationship between boroughs and their District numbers. This allows for an in-depth comparison of borough-wise street quality and highlights potential discrepancies.

District	BKN01	BKN02	BKN03	BKN04	BKN05	BKN08	BKN09	BKN16	BKN17	\
Borough										
Bronx	0	0	0	0	0	0	0	0	0	
Brooklyn	1020	816	1020	612	816	612	612	770	1020	
Manhattan	0	0	0	0	0	0	0	0	0	
Queens	0	0	0	0	0	0	0	0	0	
Staten Island	0	0	0	0	0	0	0	0	0	

District	BKS06	...	QW01	QW02	QW03	QW04	QW05	QW06	QW09	SI01	\
Borough											
Bronx	0	...	0	0	0	0	0	0	0	0	
Brooklyn	1021	...	0	0	0	0	0	0	0	0	
Manhattan	0	...	0	0	0	0	0	0	0	0	
Queens	0	...	1224	612	612	612	1020	408	816	0	
Staten Island	0	...	0	0	0	0	0	0	0	816	

District	SI02	SI03
Borough		
Bronx	0	0
Brooklyn	0	0
Manhattan	0	0
Queens	0	0
Staten Island	816	1598

[5 rows x 59 columns]

2. Plot Scatter plot, box plot, Heatmap using seaborn.

Scatter plot : The scatter plot visualizes the relationship between Acceptable Streets % and Acceptable Sidewalks %. Each point represents a data entry where the x-axis shows the percentage of acceptable streets, and the y-axis represents the percentage of acceptable sidewalks. The plot indicates a high concentration of data points in the upper-right region, suggesting that many areas have high percentages of both

acceptable streets and sidewalks. However, there is noticeable variance in sidewalk quality even when street conditions are lower.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load dataset
file_path = "/content/scorecard.csv" # Update with correct path
df = pd.read_csv(file_path)

# Drop missing values for relevant columns
df_numeric = df.dropna(subset=["Acceptable Streets %", "Acceptable Sidewalks %"])

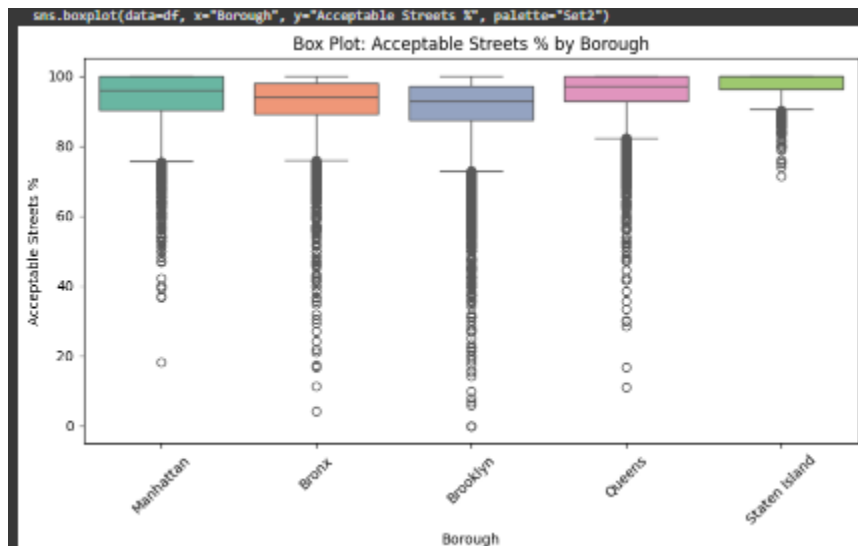
# Scatter Plot: Acceptable Streets % vs Acceptable Sidewalks %
plt.figure(figsize=(8, 5))
sns.scatterplot(data=df_numeric, x="Acceptable Streets %", y="Acceptable Sidewalks %", alpha=0.5)
plt.title("Scatter Plot: Acceptable Streets % vs Acceptable Sidewalks %")
plt.xlabel("Acceptable Streets %")
plt.ylabel("Acceptable Sidewalks %")
plt.show()

# Box Plot: Distribution of Acceptable Streets % by Borough
plt.figure(figsize=(10, 5))
sns.boxplot(data=df, x="Borough", y="Acceptable Streets %", palette="Set2")
plt.xticks(rotation=45)
plt.title("Box Plot: Acceptable Streets % by Borough")
plt.xlabel("Borough")
plt.ylabel("Acceptable Streets %")
plt.show()

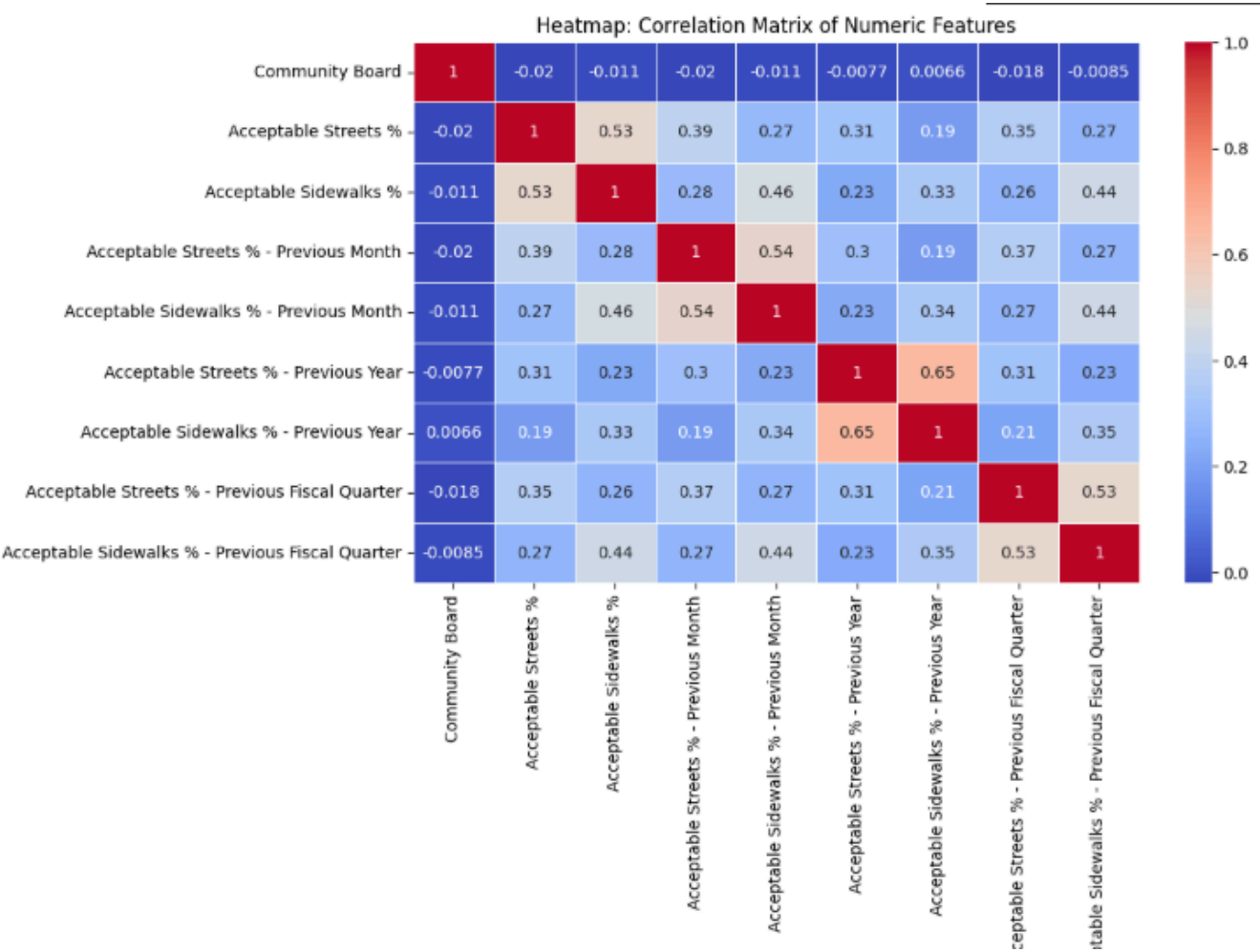
# Heatmap: Correlation matrix of numeric columns
plt.figure(figsize=(10, 6))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap="coolwarm", linewidths=0.5)
plt.title("Heatmap: Correlation Matrix of Numeric Features")
plt.show()
```



Box plot: The box plot shows the distribution of Acceptable Streets % across boroughs. Most have a high median near 100%, but Brooklyn and Bronx show more variability with lower outliers. Staten Island has the most consistent street quality.



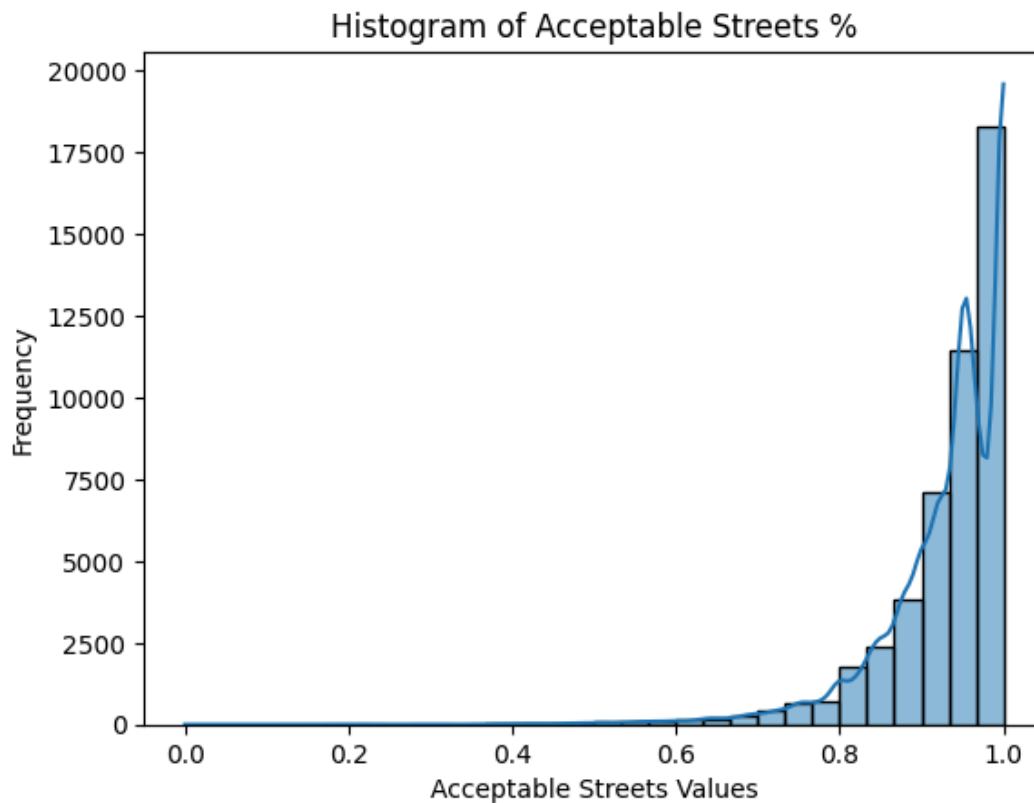
Heat map: The heatmap shows the correlation between various numeric features related to street and sidewalk conditions. Acceptable Streets % and Acceptable Sidewalks % have a moderate positive correlation (0.53). Past values, such as previous month, year, and fiscal quarter, show strong correlations with current values, especially Acceptable Streets % - Previous Year (0.65). The Community Board has almost no correlation with other features.



3. Create histogram and normalized Histogram.

```
import seaborn as sns
import matplotlib.pyplot as plt

# Example: Histogram of 'Acceptable Streets %'
sns.histplot(df['Acceptable Streets %'], bins=30, kde=True)
plt.title("Histogram of Acceptable Streets %")
plt.xlabel("Acceptable Streets Values")
plt.ylabel("Frequency")
plt.show()
```



4. Describe what this graph and table indicates.

Table Description:

1. The table represents data for districts across five New York City boroughs: Bronx, Brooklyn, Manhattan, Queens, and Staten Island.
2. Each borough is represented in rows, and each district within the boroughs is represented in columns (e.g., **BKN01**, **QW01**, **SI01**).
3. The **values** in the cells (e.g., **0**, **612**, **1020**) correspond to a specific metric (e.g., counts, measurements, or scores).
4. **Zero values** in several boroughs (Bronx, Manhattan, Queens, Staten Island) may indicate the absence of data or lack of measurement in those districts.
5. **Brooklyn** shows significant values (e.g., **1020**, **816**) in several districts, potentially indicating higher activity, population, or other metrics.

Graph Description:

1. The graph likely visualizes the distribution of these values across boroughs and districts.
2. **Bar graphs** or **heat maps** could be used to display the varying values across the districts of each borough.
3. In a **heat map**, Brooklyn's districts with higher values (e.g., **BKN01**, **BKN02**) would be represented with **high-intensity colors**, while other boroughs would have **lower-intensity or neutral colors** (e.g., Bronx, Manhattan, Staten Island).
4. If the graph is a **bar graph**, it might show multiple bars for each borough, with **Brooklyn** having the tallest bars in its districts, while other boroughs (like the Bronx or Queens) may have shorter or no bars.
5. The graph would visually highlight the stark differences between Brooklyn and the other boroughs, reinforcing the dominance of Brooklyn in this dataset.

5. Handle outlier using box plot and Interquartile range.

```
File Edit View Insert Runtime Tools Help
+ Code + Text

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
df = pd.read_csv("scorecard.csv")

# Create a box plot to visualize outliers
plt.figure(figsize=(12, 6))
sns.boxplot(data=df)
plt.title('Box Plot for Scorecard Data')
plt.show()

# Calculate Q1 (25th percentile) and Q3 (75th percentile)
Q1 = df.quantile(0.25)
Q3 = df.quantile(0.75)

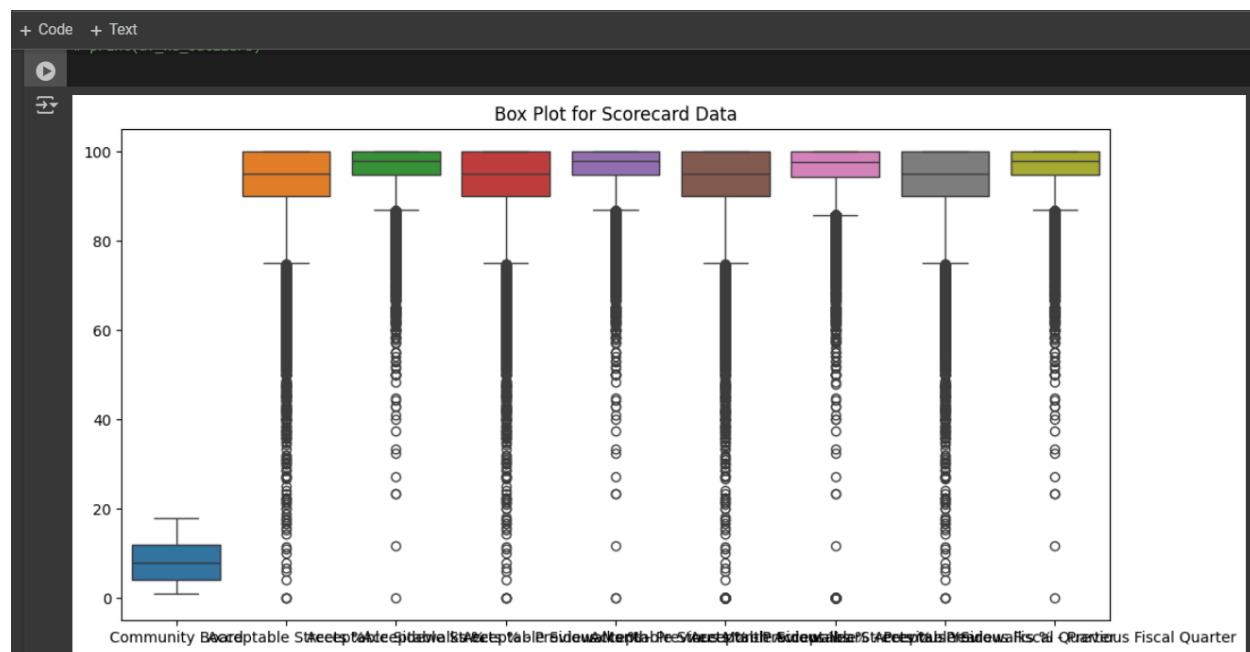
# Calculate the IQR (Interquartile Range)
IQR = Q3 - Q1

# Determine the lower and upper bounds for outliers
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Remove outliers from the dataset
df_no_outliers = df[(df >= lower_bound) & (df <= upper_bound)].dropna()

# Create a box plot after removing outliers
plt.figure(figsize=(12, 6))
sns.boxplot(data=df_no_outliers)
plt.title('Box Plot After Removing Outliers')
plt.show()

# Display the cleaned dataset (if needed)
print(df_no_outliers)
```



Conclusion:

- **This experiment successfully explored and visualized data using multiple graphs and tables.**
- **The table and graphs highlight variations in borough data, with Brooklyn showing significantly higher values.**
- **Outliers were handled using the box plot and IQR method, ensuring a cleaner dataset.**
- **The heatmap and histograms provided insights into data distribution, while scatter plots and box plots helped identify trends and anomalies.**