Experiment 2

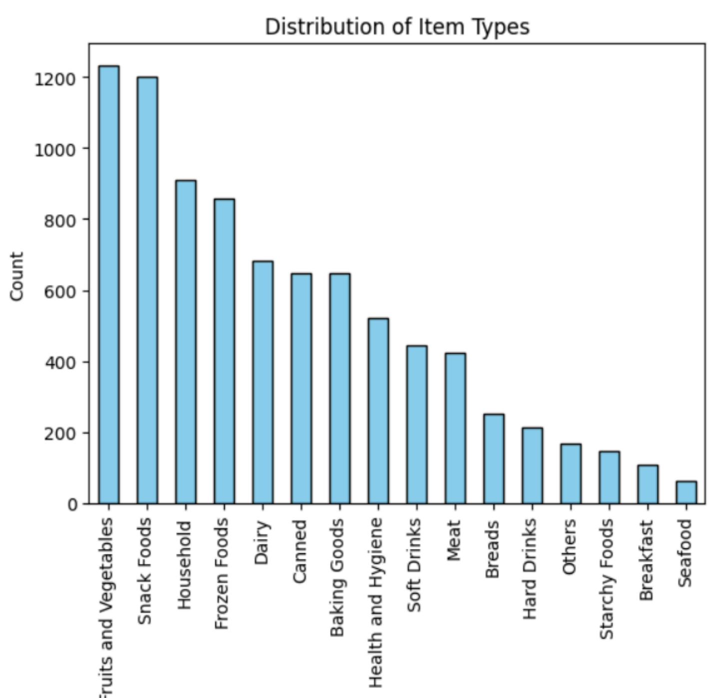
Aim: Data Visualization/ Exploratory data Analysis using Matplotlib and Seaborn.

Theory:
Data visualization and exploratory data analysis (EDA) using Matplotlib and Seaborn help uncover patterns, trends, and relationships within data. Matplotlib is a flexible, low-level library for creating static, animated, and interactive plots, while Seaborn is built on top of Matplotlib and provides a high-level interface for visually appealing statistical graphics. EDA involves techniques like histograms, scatter plots, box plots, and heatmaps to understand data distributions, detect outliers, and identify correlations.

1. **Bar graph and contingency table using any two features:**

   a. Bar Graph for distribution of item type.

```
import matplotlib.pyplot as plt

df['Item_Type'].value_counts().plot(kind='bar', color='skyblue', edgecolor='black')

plt.xlabel("Item Type")
plt.ylabel("Count")
plt.title("Distribution of Item Types")
plt.xticks(rotation=90)
plt.show()
```



From the graph, we observe that the majority of items fall under the category of "Fruits and Vegetables," with a count of approximately 1,200. Similarly, the "Snack Foods" category also comprises a comparable number of items, indicating a balanced distribution between these two categories. Rest of the items have a count of less than 1000. “Seafood” accounts for less than approximately 100 items, which is the lowest.

b. Contingency Table for Item type and Outlet type
   A contingency table (or cross-tabulation table) displays the frequency distribution of two categorical variables in a dataset. It helps in understanding the relationship between the two features

```python
import pandas as pd

contingency_table = pd.crosstab(df['Item_Type'], df['Outlet_Type'])
print(contingency_table)
```

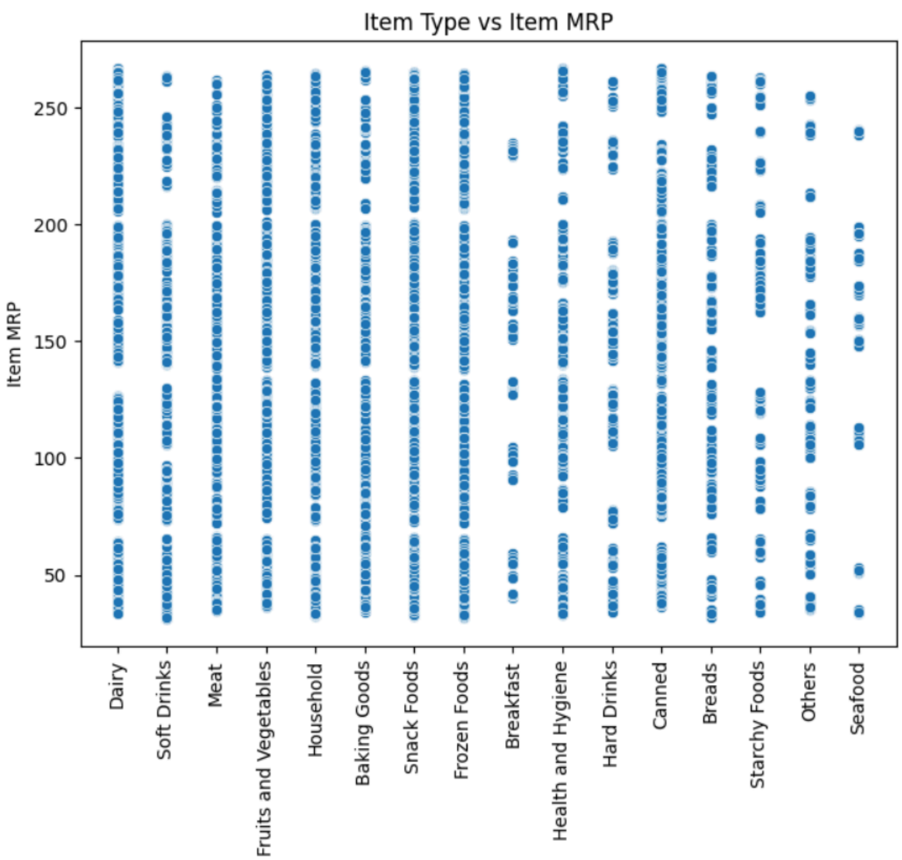
```
Outlet_Type            Grocery Store  Supermarket Type1  Supermarket Type2  \
Item_Type
Baking Goods                      85                426                 68
Breads                            33                160                 27
Breakfast                         19                 68                 12
Canned                            73                426                 78
Dairy                             92                450                 73
Frozen Foods                     103                572                 92
Fruits and Vegetables            152                805                135
Hard Drinks                       24                145                 22
Health and Hygiene                67                335                 58
Household                        119                597                 95
Meat                              66                257                 46
Others                            27                107                 20
Seafood                           10                 40                  7
Snack Foods                      146                785                132
Soft Drinks                       54                300                 46
Starchy Foods                     13                104                 17

Outlet_Type            Supermarket Type3
Item_Type
Baking Goods                          69
Breads                                31
Breakfast                             11
Canned                                72
Dairy                                 67
Frozen Foods                          89
Fruits and Vegetables                140
Hard Drinks                           23
Health and Hygiene                    60
Household                             99
Meat                                  56
Others                                15
Seafood                                7
Snack Foods                          137
Soft Drinks                           45
Starchy Foods                         14
```

From the table, it is evident that Supermarket Type 1 has the highest variety and number of items across all categories compared to other outlet types. Fruits and Vegetables and Snack Foods are the dominant categories in most outlets, indicating their popularity. Smaller categories like Seafood and Starchy Foods are less common across all outlet types, suggesting limited demand or availability.

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

a. Scatter Plot for Item Type vs Item MRP

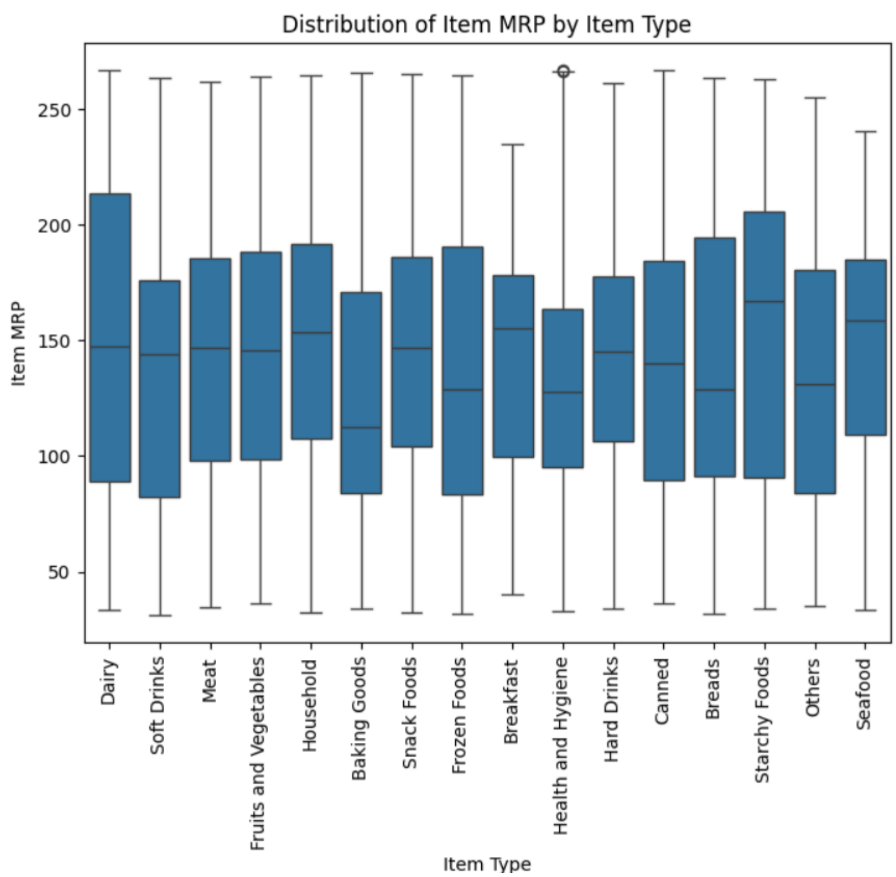
```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(8, 6))
sns.scatterplot(data=df, x="Item_Type", y="Item_MRP")
plt.title("Item Type vs Item MRP")
plt.xlabel("Item Type")
plt.xticks(rotation=90)
plt.ylabel("Item MRP")
plt.show()
```



From the visualization, it is evident that certain categories, such as Fruits and Vegetables, Snack Foods, and Meat, have a wide range of MRPs, spanning from low to high values. On the other hand, categories like Seafood and Breakfast seem to have fewer data points and a narrower MRP range, suggesting limited representation in the dataset.

b. Box Plot

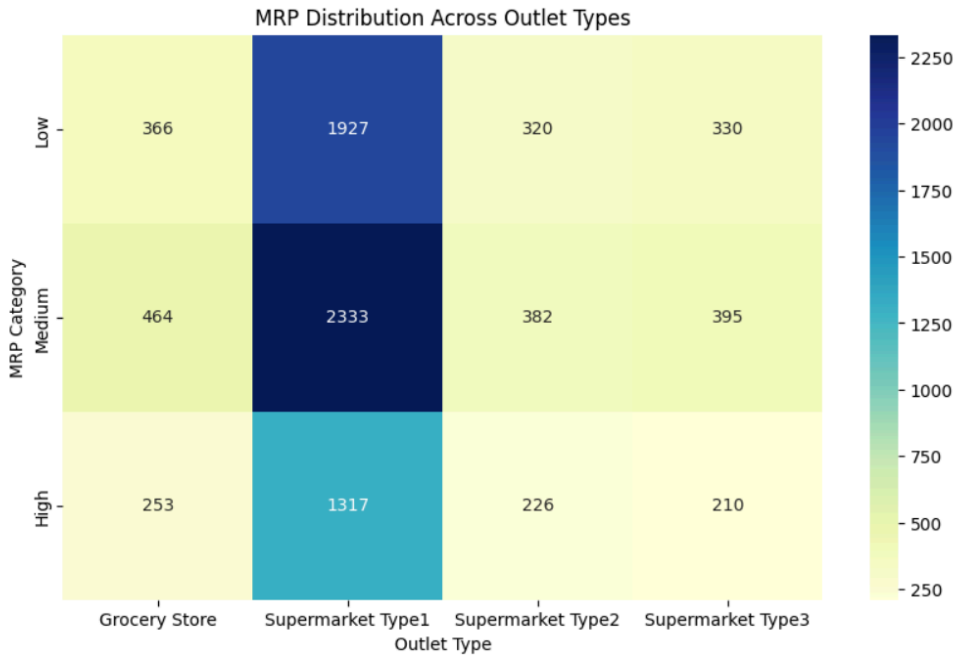
```
plt.figure(figsize=(8, 6))
sns.boxplot(data=df, x="Item_Type", y="Item_MRP")
plt.title("Distribution of Item MRP by Item Type")
plt.xlabel("Item Type")
plt.ylabel("Item MRP")
plt.xticks(rotation=90)
plt.show()
```



The box plot highlights the distribution of Item MRP across various Item Types, with noticeable variations in medians, such as higher values for Dairy and Health and Hygiene compared to Starchy Foods and Baking Goods.

c. Heatmap for Item MRP distribution

```
df["MRP_Bin"] = pd.cut(df["Item_MRP"], bins=3, labels=["Low", "Medium", "High"])
heatmap_data = pd.crosstab(df["MRP_Bin"], df["Outlet_Type"])

plt.figure(figsize=(10, 6))
sns.heatmap(heatmap_data, annot=True, cmap="YlGnBu", fmt="d")
plt.title("MRP Distribution Across Outlet Types")
plt.xlabel("Outlet Type")
plt.ylabel("MRP Category")
plt.show()
```

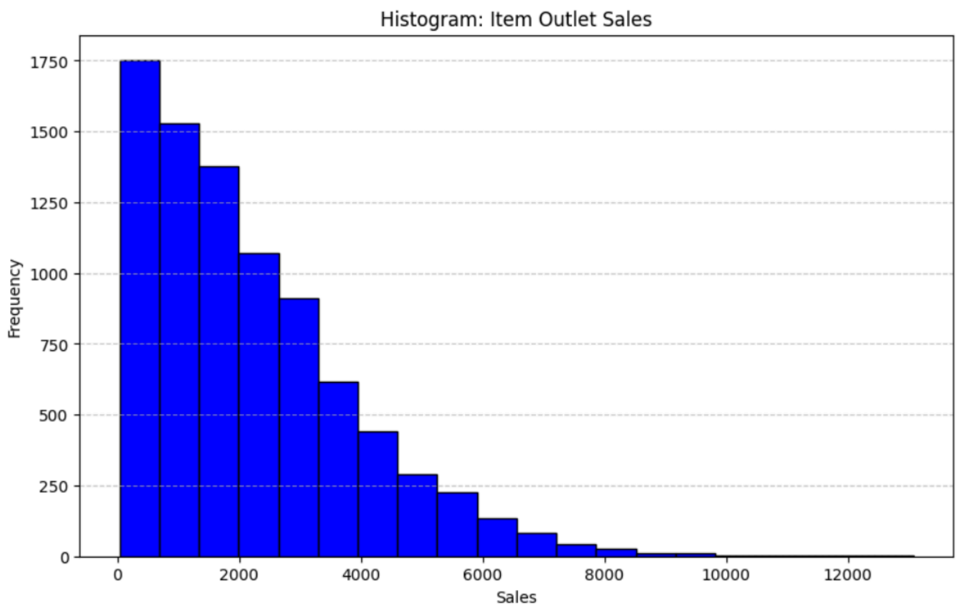


Supermarket Type 1 dominates in all MRP categories, particularly in medium MRP, followed by high MRP. Grocery Stores have a stronger presence in low and medium MRP categories but contribute minimally to high MRP. Supermarkets Types 2 and 3 show a balanced but smaller distribution across all categories, with a slight focus on medium and low MRP.

3. Histogram and normalized Histogram.

a. Histogram for Item Outlet Sales

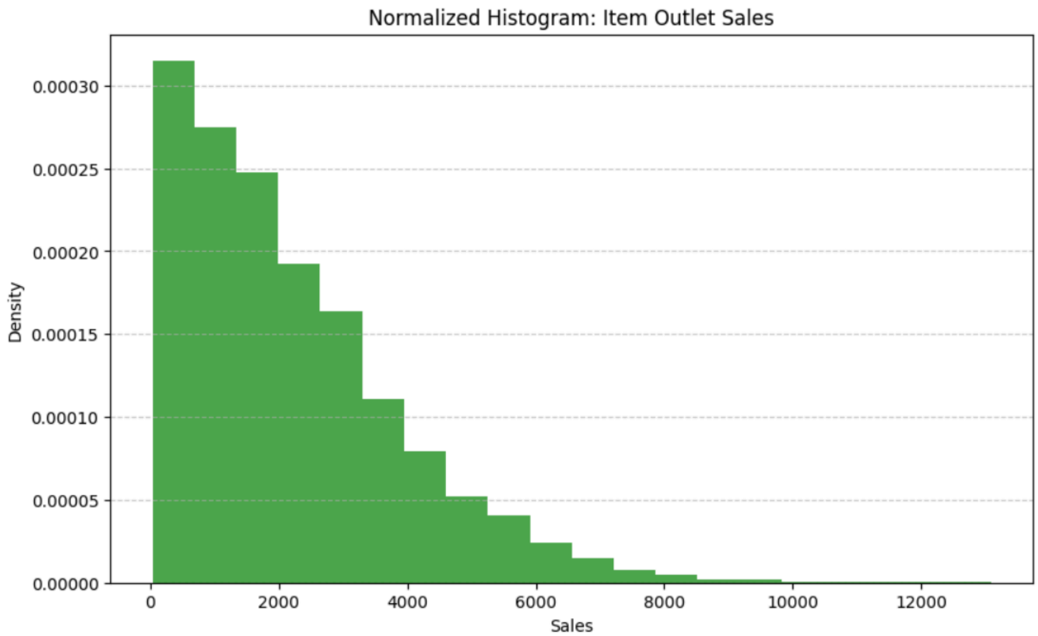
```
sales_data = df["Item_Outlet_Sales"]

plt.figure(figsize=(10, 6))
plt.hist(sales_data, bins=20, color='blue', edgecolor='black')
plt.title("Histogram: Item Outlet Sales")
plt.xlabel("Sales")
plt.ylabel("Frequency")
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```



The histogram reveals the distribution of item outlet sales, which is heavily skewed to the right. The majority of sales are concentrated in the lower range, with the highest frequency occurring between 0 and 2,000. As the sales value increases, the frequency decreases significantly, indicating that higher sales amounts are less common.

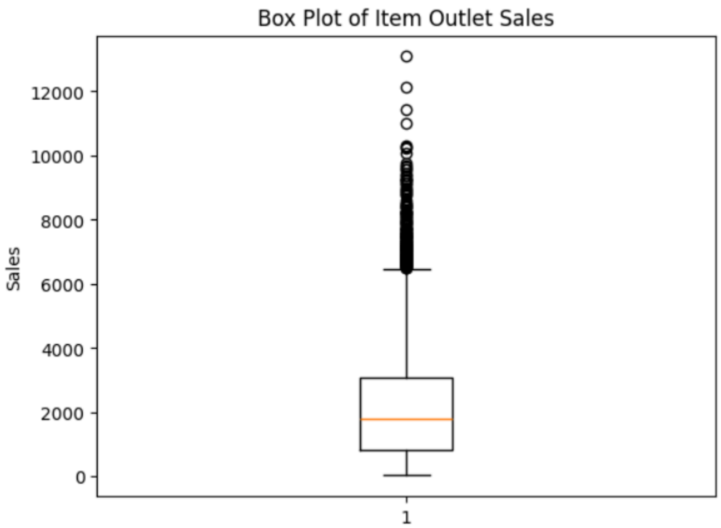
b. Normalized Histogram

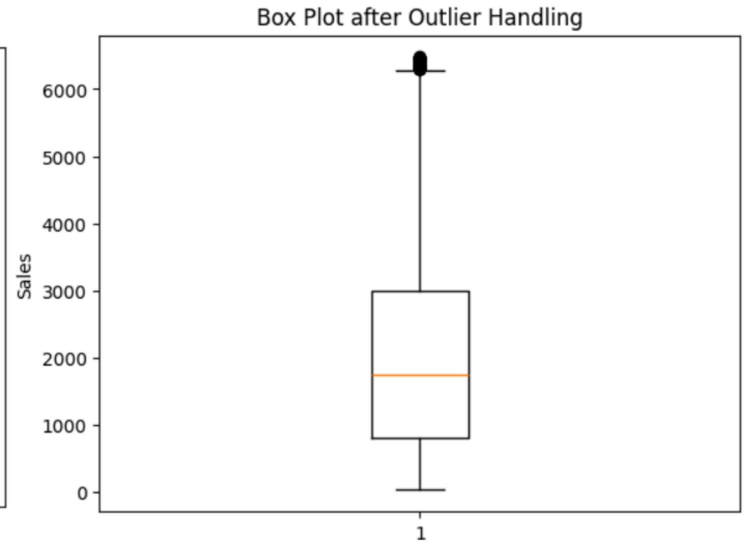
```
plt.figure(figsize=(10, 6))
plt.hist(sales_data, bins=20, color='green', alpha=0.7, density=True)
plt.title("Normalized Histogram: Item Outlet Sales")
plt.xlabel("Sales")
plt.ylabel("Density")
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```



4. Outliers using box plot and Inter quartile range.

```
sales = df['Item_Outlet_Sales']

plt.boxplot(sales)
plt.title('Box Plot of Item Outlet Sales')
plt.ylabel('Sales')
plt.show()

Q1 = sales.quantile(0.25)
Q3 = sales.quantile(0.75)
IQR = Q3 - Q1

lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

filtered_sales = sales[(sales >= lower_bound) & (sales <= upper_bound)]

df['Item_Outlet_Sales_Capped'] = np.where(
    sales < lower_bound, lower_bound,
    np.where(sales > upper_bound, upper_bound, sales)
)

plt.boxplot(filtered_sales)
plt.title('Box Plot after Outlier Handling')
plt.ylabel('Sales')
plt.show()
```





The first box plot highlights that the dataset has several outliers, with Item Outlet Sales extending well beyond the upper whisker (above ~6000). These are visible as individual points above the main plot area.After applying the interquartile range (IQR) method for outlier handling, the second box plot demonstrates that extreme outliers have been removed. The data now falls within a more compact range, capped approximately at the upper whisker (~6000)

Conclusion:

This experiment demonstrated the effective use of Matplotlib and Seaborn for data visualization and exploratory data analysis (EDA). By visualizing data through bar graphs, contingency tables, scatter plots, box plots, and heatmaps, we uncovered patterns, distributions, and relationships within the dataset. Normalized histograms provided deeper insights into data density, while the interquartile range (IQR) method successfully identified and removed outliers for improved data quality.