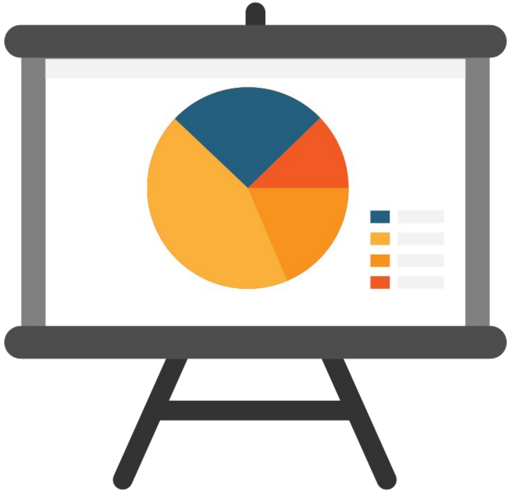
Analysis of Superstore Dataset

Submitted By,
SHIBIN T A
shibinashraf36@gmail.com

Github Link:
https://github.com/shibinashraf/superstore-da-ibmskillsbuild

Agenda

1. Introduction
2. Project Overview
3. End Users
4. Dataset
5. Data analysis
6. Conclusion



Introduction

We will be exploring the **Superstore Dataset**, which provides a comprehensive collection of sales and profit data from a fictional superstore. The dataset encompasses various variables such as sales, profit, region, category, and more. Our objective is to perform **Exploratory Data Analysis (EDA) and analysis** on this dataset to gain valuable insights and draw meaningful conclusions.

By analyzing this dataset, we aim to **uncover patterns, trends, and relationships** that can help us understand the performance of the superstore better. We will examine the sales and profit distribution, analyze regional performance, delve into category-specific insights, and explore the time series patterns of sales and profit.

Project Overview

- **Data Cleaning:** This process involves handling missing values, correcting inconsistent or inaccurate data, and removing duplicates. Cleaning the data ensures that it is in a suitable format for analysis and prevents misleading results.
- **Exploratory Data Analysis (EDA):** EDA involves examining and visualizing the data to understand its main characteristics. It includes tasks such as summarizing the data, identifying patterns and trends, detecting outliers, and exploring relationships between variables.
- **Statistical Analysis:** Statistical analysis involves applying statistical methods and techniques to analyze the data. This may include calculating summary statistics, conducting hypothesis tests, performing regression analysis, or running other statistical models to uncover relationships and make data-driven inferences.

Project Overview

- **Data Visualization:** Data visualization is the process of representing data visually through charts, graphs, or other visual elements. Visualization helps in understanding patterns, trends, and relationships in the data, making it easier to communicate findings to stakeholders effectively.
- **Modeling and Prediction:** Modeling involves building statistical or machine learning models to predict outcomes, classify data, or gain deeper insights. This process includes tasks like model selection, training, evaluation, and validation to create models that can accurately predict or explain the data.
- **Conclusion and Recommendations:** Based on the insights and findings obtained through the analysis, conclusions can be drawn, and actionable recommendations can be made. Conclusions should be supported by evidence from the data, and recommendations should provide practical guidance for decision-making or problem-solving.

End Users

- **Store Management:** The management team of the superstore can use the analysis results to gain a deeper understanding of their sales and profitability. They can make data-driven decisions regarding pricing strategies, inventory management, marketing campaigns, and resource allocation based on the insights obtained from the data analysis.
- **Sales and Marketing Teams:** The sales and marketing teams can leverage the analysis findings to optimize their strategies. They can identify the most profitable product categories, target specific customer segments for marketing campaigns, and refine their sales techniques based on the patterns and trends identified in the dataset.
- **Operations Team:** The operations team can utilize the analysis results to streamline operations and improve efficiency. They can identify regions or product categories that require attention, optimize supply chain management, and identify areas for cost reduction or process improvement.

End Users

- **Financial Analysts:** Financial analysts can use the analysis findings to assess the financial performance of the superstore. They can evaluate the profitability of different regions or product categories, identify cost drivers, and make recommendations for financial planning and forecasting.
- **Business Consultants:** Business consultants or external advisors can use the analysis insights to provide strategic guidance to the superstore. They can identify areas of improvement, recommend market expansion strategies, and offer insights on competitive positioning based on the analysis findings.
- **Researchers and Academics:** Researchers and academics interested in retail or sales analysis can utilize the Superstore Dataset and analysis results for further study or as a benchmark for their research. The dataset can contribute to the understanding of retail trends, consumer behavior, and market dynamics.

Dataset

- The Superstore Dataset is a comprehensive collection of sales and profit data from a fictional superstore.
- The dataset includes information on various aspects, such as sales, profit, region, category, product sub-category, customer segment, and shipping details.
- With a significant number of records, each representing a unique sale, the dataset provides a wealth of transactional data for analysis. The dataset covers a specific time period, enabling the examination of trends and patterns over time, such as monthly or yearly performance.
- Geographical information is also available, allowing for regional analysis and comparison of sales and profitability across different areas.

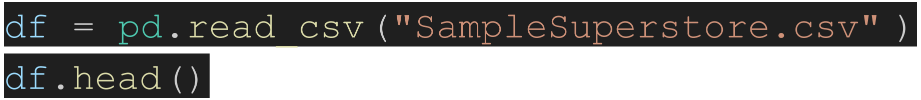
Dataset

- The dataset includes metrics like sales and profit, providing a foundation for evaluating the superstore's performance and identifying high-performing or underperforming areas.
- Moreover, product categorization facilitates the analysis of sales and profit trends based on different product categories and sub-categories.
- The dataset offers customer-related information, such as customer segment or ID, which can be leveraged for customer segmentation and analysis of customer behavior and preferences.

Data Analysis

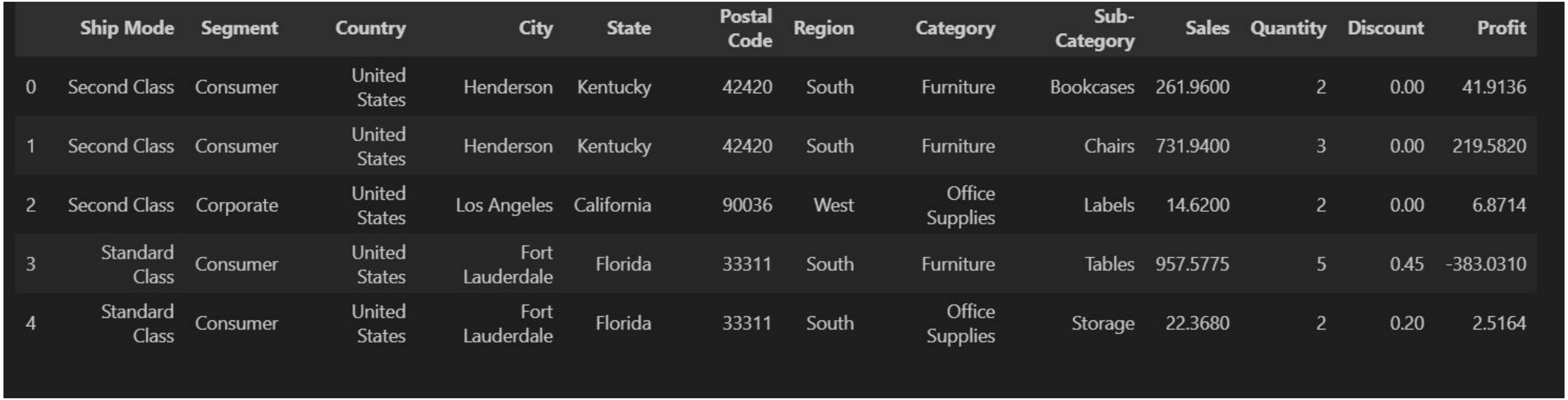
- The data analysis for this project was conducted using the **Visual Studio Code** (VS Code) integrated development environment. VS Code provides a convenient and versatile environment for writing and executing Python code, making it an ideal choice for performing data analysis tasks.
- To begin the analysis, Python and the required libraries were installed on the system. Python serves as the programming language of choice for data analysis due to its rich ecosystem of libraries and its flexibility in handling data. Libraries such as pandas, matplotlib, and others were installed using the pip package manager within the VS Code environment. These libraries provide powerful tools for data manipulation, analysis, and visualization.

The following code snippet reads the Superstore Dataset using pandas library and displays the first few rows of the dataset:



```
df = pd.read_csv("SampleSuperstore.csv")
df.head()
```

Output :

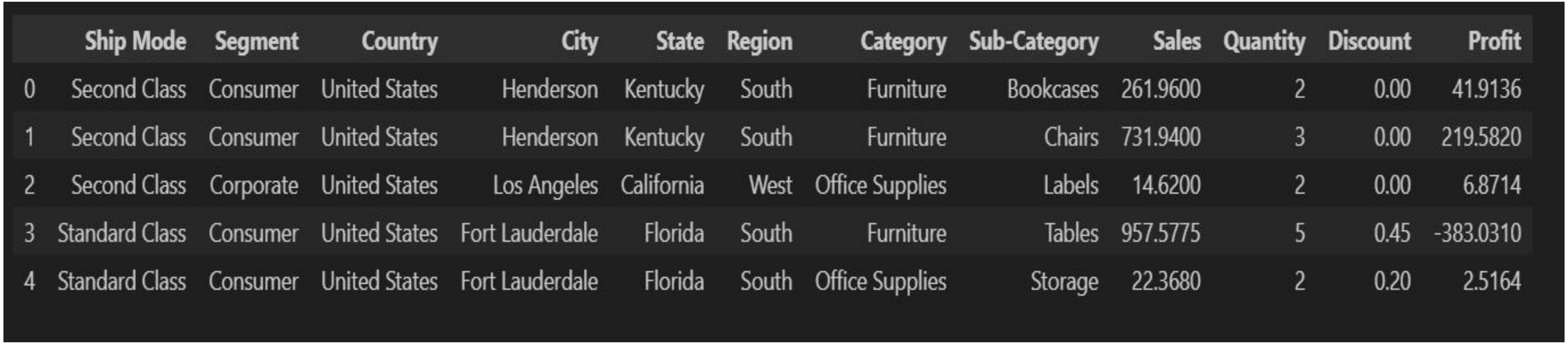


	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
0	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases	261.9600	2	0.00	41.9136
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs	731.9400	3	0.00	219.5820
2	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels	14.6200	2	0.00	6.8714
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables	957.5775	5	0.45	-383.0310
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage	22.3680	2	0.20	2.5164

The code snippet provided drops the "Postal Code" column from the DataFrame df using the drop() function in pandas:

```
df.drop(columns = "Postal Code", inplace = True)
df.head()
```

Output :

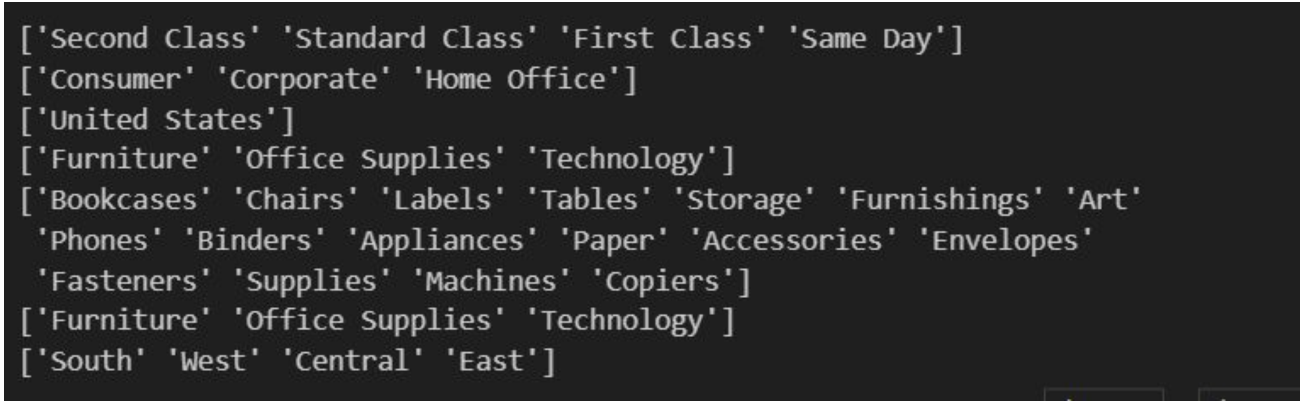


	Ship Mode	Segment	Country	City	State	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
0	Second Class	Consumer	United States	Henderson	Kentucky	South	Furniture	Bookcases	261.9600	2	0.00	41.9136
1	Second Class	Consumer	United States	Henderson	Kentucky	South	Furniture	Chairs	731.9400	3	0.00	219.5820
2	Second Class	Corporate	United States	Los Angeles	California	West	Office Supplies	Labels	14.6200	2	0.00	6.8714
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	South	Furniture	Tables	957.5775	5	0.45	-383.0310
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	South	Office Supplies	Storage	22.3680	2	0.20	2.5164

The following code snippets will print the unique values for each specified column in the DataFrame df :

```
print(df["Ship Mode"].unique())
print(df["Segment"].unique())
print(df["Country"].unique())
print(df["Category"].unique())
print(df["Sub-Category"].unique())
print(df["Category"].unique())
print(df["Region"].unique())
```

Output :

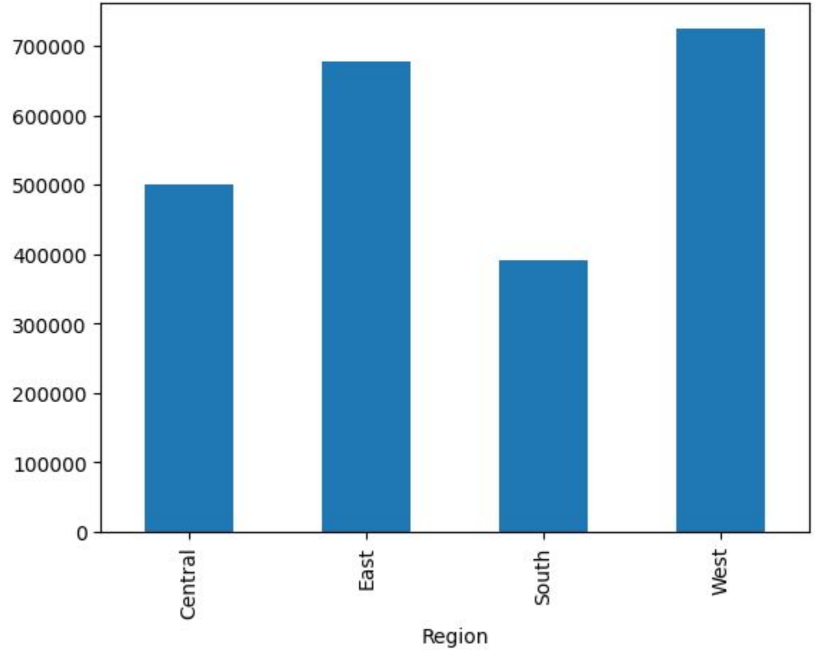


```
['Second Class' 'Standard Class' 'First Class' 'Same Day']
['Consumer' 'Corporate' 'Home Office']
['United States']
['Furniture' 'Office Supplies' 'Technology']
['Bookcases' 'Chairs' 'Labels' 'Tables' 'Storage' 'Furnishings' 'Art'
 'Phones' 'Binders' 'Appliances' 'Paper' 'Accessories' 'Envelopes'
 'Fasteners' 'Supplies' 'Machines' 'Copiers']
['Furniture' 'Office Supplies' 'Technology']
['South' 'West' 'Central' 'East']
```

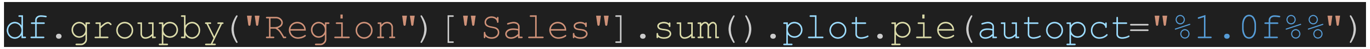
The following code snippet groups the data in the DataFrame df by the "Region" column and calculates the sum of sales for each region. It then creates a bar plot to visualize the total sales per region:

```
df.groupby("Region")["Sales"].sum().plot.bar()
```

Output :

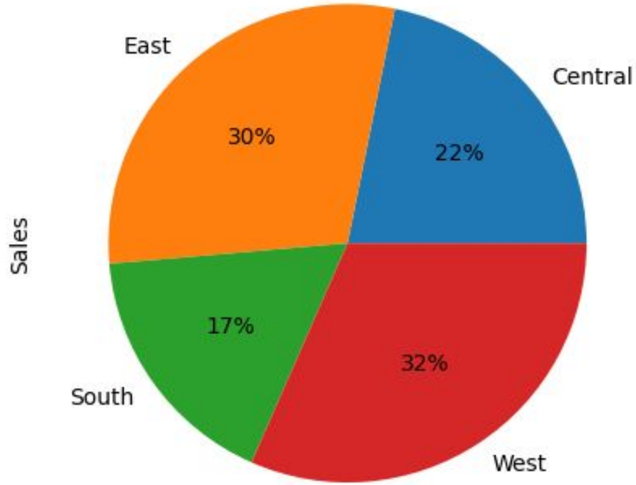


The following code snippet groups the data in the DataFrame df by the "Region" column and calculates the sum of sales for each region. It then creates a pie chart to visualize the proportion of total sales contributed by each region:



```
df.groupby("Region")["Sales"].sum().plot.pie(autopct="%1.0f%%")
```

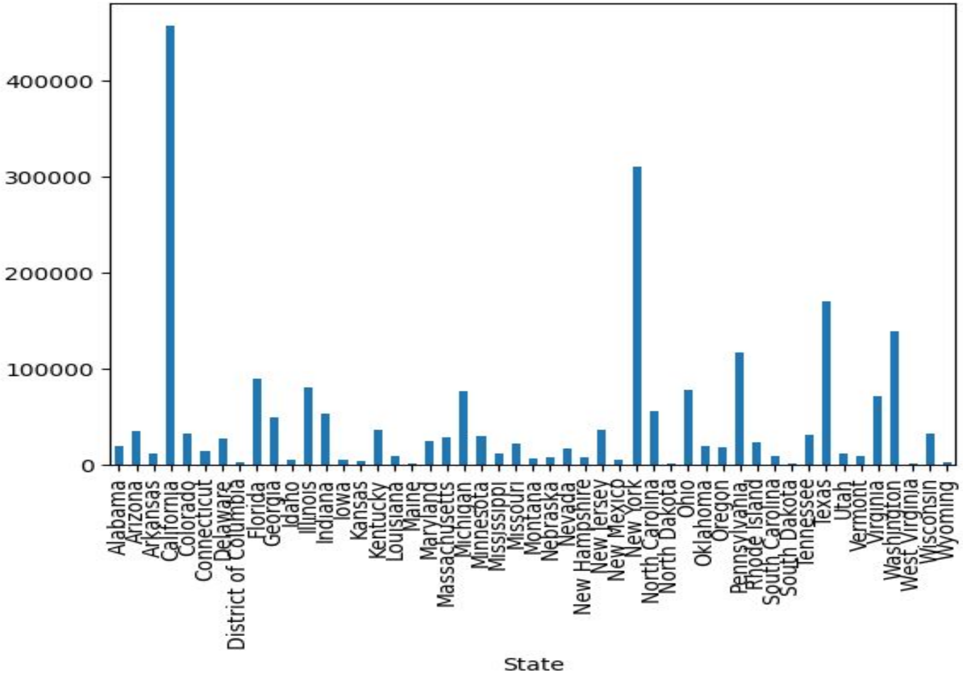
Output :



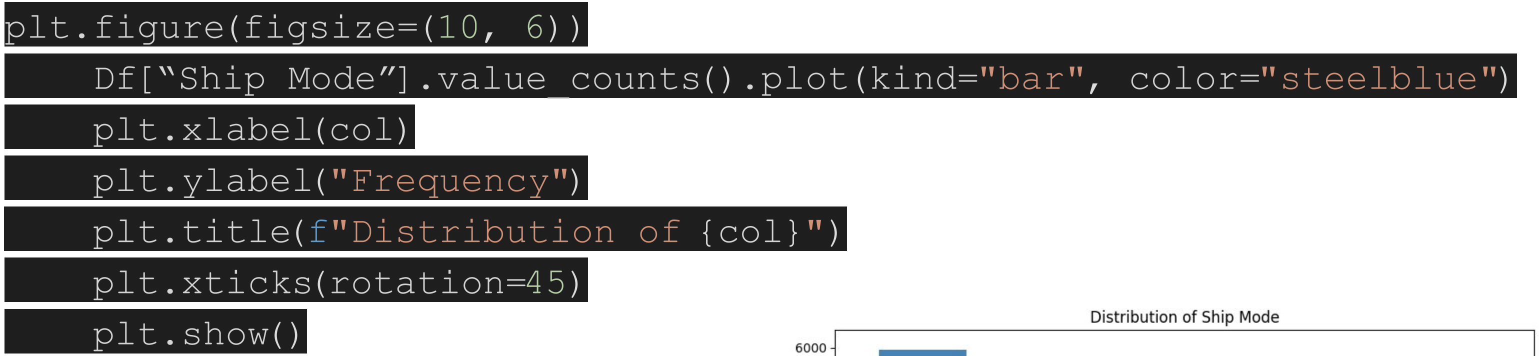
The following code snippet groups the data in the DataFrame df by the "State" column and calculates the sum of sales for each state. It then creates a bar plot to visualize the total sales per state:

```
df.groupby("State")["Sales"].sum().plot.bar()
```

Output :

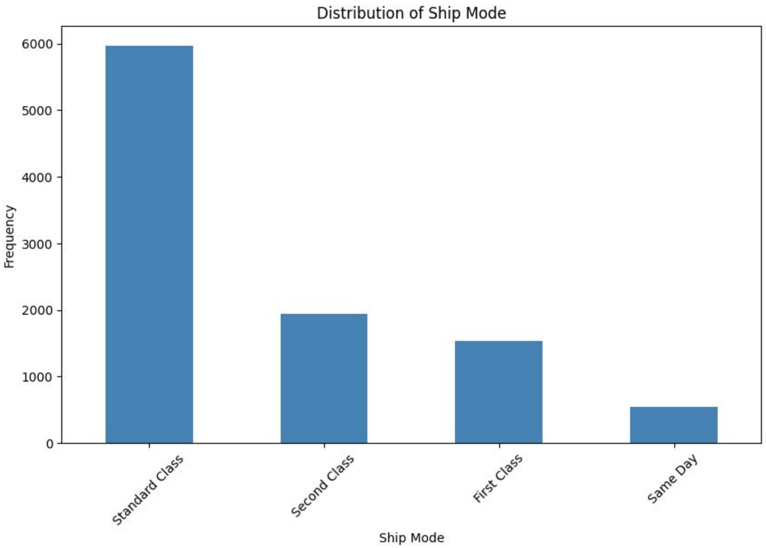


The following code snippet shows distribution of categorical attribute “Ship Mode” :



```
plt.figure(figsize=(10, 6))
    Df[“Ship Mode”].value_counts().plot(kind="bar", color="steelblue")
    plt.xlabel(col)
    plt.ylabel("Frequency")
    plt.title(f"Distribution of {col}")
    plt.xticks(rotation=45)
    plt.show()
```

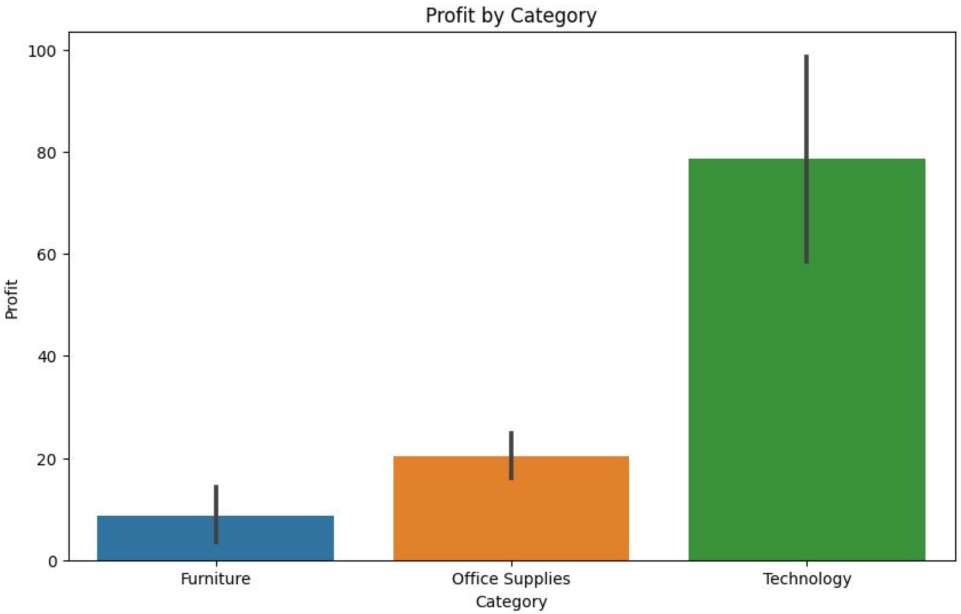
Output :



The following code snippet used for plotting a bar chart to visualize the profit by category using the seaborn library. :

```
import seaborn as sns
plt.figure(figsize=(10, 6))
sns.barplot(data=df, x="Category", y="Profit")
plt.xlabel("Category")
plt.ylabel("Profit")
plt.title("Profit by Category")
plt.show()
```

Output :



Conclusion

In conclusion, our analysis of the Superstore dataset has provided valuable insights into the sales and profit performance of the store. We observed a wide range of sales and profit values, with some instances of high performance. The "Standard Class" ship mode emerged as the most frequently used shipping method, while the "Consumer" segment appeared to be the largest customer group. Regions such as the East and West contributed significantly to overall sales. By examining different categories and sub-categories, we identified variations in sales and profit, with some categories showing higher profitability than others. The analysis also revealed a positive correlation between sales and profit, indicating that increasing sales can lead to higher profits. These findings can guide strategic decision-making, such as optimizing shipping methods and focusing on profitable categories, to further enhance the store's performance and profitability.

Thank You