CASE STUDY: ANALYSIS OF SUPERSTORE

PROJECT

By - Mrunmayi Ram Kinhikar

PROJECT TOPIC

The topic "Analysis of Superstore" involves a comprehensive study and examination of a dataset related to a retail superstore.

The ultimate goal of the "Analysis of Superstore" is to provide actionable insights and recommendations to the superstore's management for optimizing sales, profits, and overall performance.

AGENDA

The main objective of this analysis is to gain insights and understand various aspects of the superstore's performance, including sales, profits, customer behavior, and regional trends. By employing data analysis techniques, the project aims to extract valuable information from the dataset and use it to make data-driven decisions for the superstore's operations.

PROJECT OVERVIEW

- In this project, we conduct a comprehensive analysis of a retail superstore's dataset to gain valuable insights into its performance and optimize various aspects of its operations. The dataset includes information on sales, profits, and customer transactions over a specific period.

- Data Cleaning and Preprocessing: We begin by cleaning the dataset to ensure data accuracy and consistency. Any missing or erroneous data will be addressed to ensure the integrity of the analysis.

- Exploratory Data Analysis (EDA): Through EDA, we visualize the data and identify patterns and trends. Graphs and charts will be employed to showcase sales and profit distributions, uncover seasonality, and highlight significant correlations.

PROJECT OVERVIEW

- Category and Subcategory Analysis: Bar charts and pie charts will illustrate the highest-selling categories and subcategories. This information will guide decision-making to focus on high-performing product lines and optimize the product mix.

- Regional Analysis: Bar graphs and heatmaps will display sales figures across different cities and states. By identifying regions with the highest sales, we can devise localized strategies to enhance market penetration.

- Recommendations: Based on the analysis, we will provide data-driven recommendations to improve sales and profitability. These recommendations may include product promotions, marketing campaigns, and supply chain optimizations.

WHO ARE THE END USERS?

"Superstore" which is a term commonly used to refer to retail stores that offer a wide range of products, the end users can be categorized into **two** main groups:

- **Retail Customers**
- **Store Staff and Employees**

WHO ARE THE END USERS?

- **Retail Customers**

These are the individuals or consumers who visit and make purchases from the Superstore. Retail customers are the primary end users of the products and services offered by the store. They are the ones who browse the aisles, select items, and complete transactions.

- **Store Staff and Employees**

The second group of end users includes the employees and staff members who work in the Superstore. These employees use various systems and tools, such as point-of-sale (POS) systems, inventory management software, and customer service interfaces, to facilitate smooth operations and provide assistance to the retail customers.

WHO ARE THE END USERS?

The end users of the "Analysis of Superstore" project can be various stakeholders involved in the management and decision-making processes of the retail superstore. These stakeholders may include:

- Store Managers
- Marketing Team
- Sales and Operations Team
- Supply Chain and Logistics Team
- Executives and Decision-makers
- Finance Team
- Customer Service Team
- Investors and Stakeholders

SOLUTION AND PRESENTATION

- Our analysis followed a data-driven approach, starting with data cleaning and preprocessing to ensure data accuracy and reliability.
- Exploratory Data Analysis (EDA) was conducted to visualize sales, profit, and other key metrics to identify trends and patterns.
- We performed category and subcategory analysis to determine the highest-selling product lines and subcategories.
- Regional sales analysis was conducted to understand the performance of different cities and states.

SOLUTION AND PRESENTATION

The solution for the "Analysis of Superstore" project involves a series of data analysis steps -

- Data Cleaning and Preprocessing
- Exploratory Data Analysis (EDA)
- Category and Subcategory Analysis
- Regional Sales Analysis
- Customer Segmentation
- Correlation Analysis
- Data-Driven Recommendations

WOW IN SYSTEM

1. Informed Decision-Making
2. Improved Efficiency and Productivity
3. Inventory Management
4. Customer Behavior Analysis:
5. Optimized Pricing Strategies
6. Efficient Store Layout
7. Promotional Campaign Evaluation
8. Loss Prevention and Fraud Detection
9. Supply Chain Optimization
10. Real-Time Decision-Making

MODELLING

1. Imported a data
2. Cleaned the data and make a new dataset named as 'superstore_cleaned_dataset.csv'

Importing the data set

```
In [1]: import pandas as pd
        df = pd.read_csv('train.csv')
```

```
In [2]: import numpy as np

        # Display the first few rows of the DataFrame
        print("first few rows of the DataFrame")
        print(df.head())
        print("-----------------------------------------------------------------")

        # Check the basic statistics of the DataFrame
        print("Basic statistics of the DataFrame")
        print(df.describe())
        print("-----------------------------------------------------------------")

        # Check the data types and non-null counts of each column
        print("Data types and non-null counts of each column")
        print(df.info())
        print("-----------------------------------------------------------------")

        # Check for missing values in each column
        print("missing values in each column")
        print(df.isnull().sum())
        print("-----------------------------------------------------------------")
```

```
first few rows of the DataFrame
   Row ID        Order ID  Order Date   Ship Date       Ship Mode Customer ID  \
0       1  CA-2017-152156  08-11-2017  11-11-2017    Second Class    CG-12520
1       2  CA-2017-152156  08-11-2017  11-11-2017    Second Class    CG-12520
2       3  CA-2017-138688  12-06-2017  16-06-2017    Second Class    DV-13045
3       4  US-2016-108966  11-10-2016  18-10-2016  Standard Class    SO-20335
4       5  US-2016-108966  11-10-2016  18-10-2016  Standard Class    SO-20335

     Customer Name    Segment        Country             City       State  \
0      Claire Gute   Consumer  United States        Henderson    Kentucky
1      Claire Gute   Consumer  United States        Henderson    Kentucky
2  Darrin Van Huff  Corporate  United States      Los Angeles  California
3   Sean O'Donnell   Consumer  United States  Fort Lauderdale     Florida
4   Sean O'Donnell   Consumer  United States  Fort Lauderdale     Florida

   Postal Code Region       Product ID         Category Sub-Category  \
0      42420.0  South  FUR-BO-10001798        Furniture    Bookcases
```

Removing duplicate rows

```
In [5]: df.drop_duplicates(inplace=True)
```

```
In [6]: df.to_csv('superstore_cleaned_dataset.csv', index=False)
```

Importing new dataset - Cleaned dataset

```
In [7]: import pandas as pd
        df = pd.read_csv('superstore_cleaned_dataset.csv')

        #importing the new dtaaset - which is our cleaned data set
```

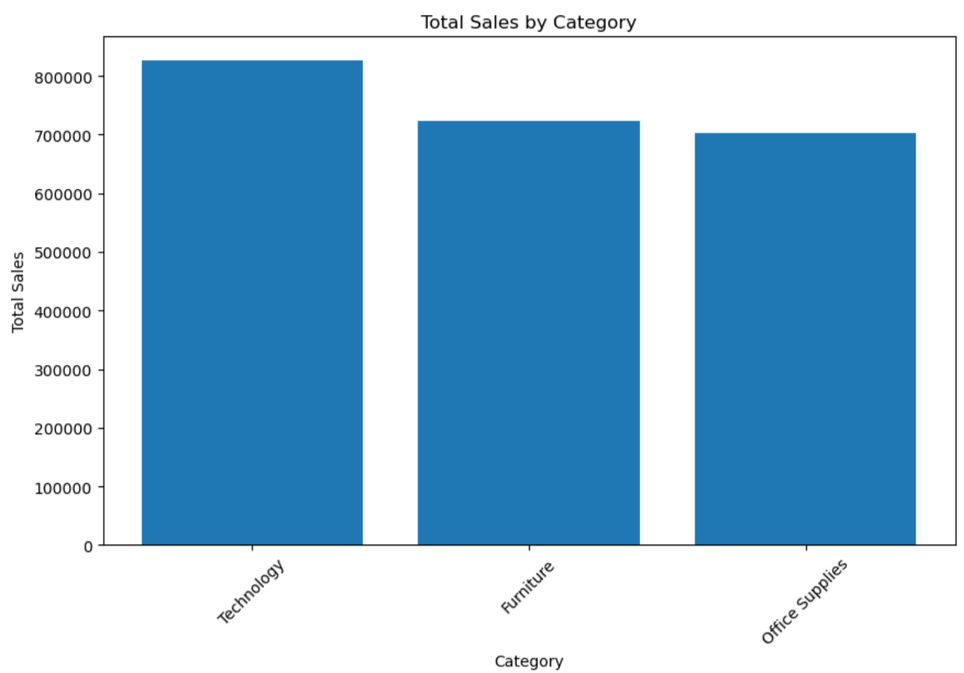
MODELLING

After claning the data, we are going to check the following things

1. Which category had the most sells

In [8]:
```
# Group the data by category and calculate the total sales for each category
category_sales = df.groupby('Category')['Sales'].sum().reset_index()

# Sort the categories based on total sales in descending order
category_sales = category_sales.sort_values(by='Sales', ascending=False)

# Get the category with the most sales (the first row after sorting)
most_sold_category = category_sales.iloc[0]['Category']

print("Category with the most sales:", most_sold_category)
```

```
Category with the most sales: Technology
```

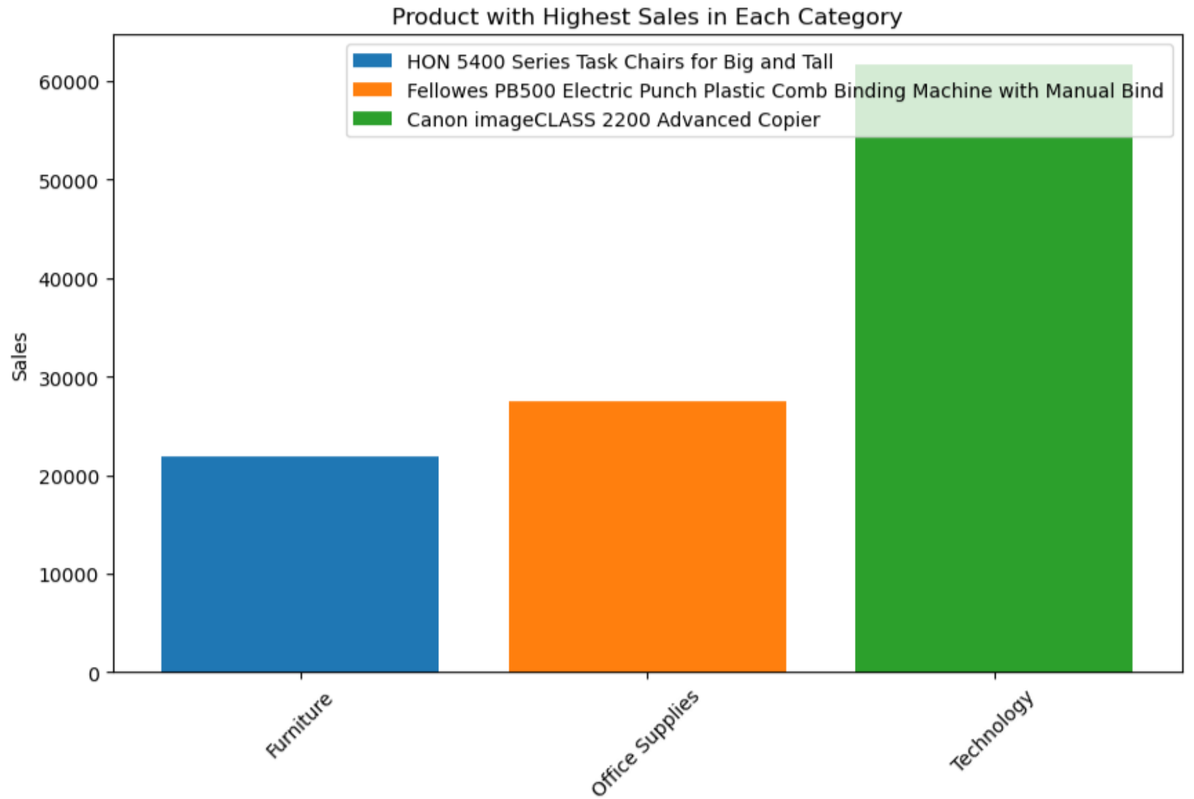


MODELLING

2. From each category, which product is sold majorly?

[10]:
```
# Group the data by category and product name and calculate the total sales for each combination
category_product_sales = df.groupby(['Category', 'Product Name'])['Sales'].sum().reset_index()

# For each category, find the product with the highest sales
major_product_in_category = category_product_sales.groupby('Category').apply(lambda x: x.loc[x['Sales'].idxmax()])

# Display the result
print("Product sold majorly from each category:")
print(major_product_in_category[['Category', 'Product Name', 'Sales']])
```

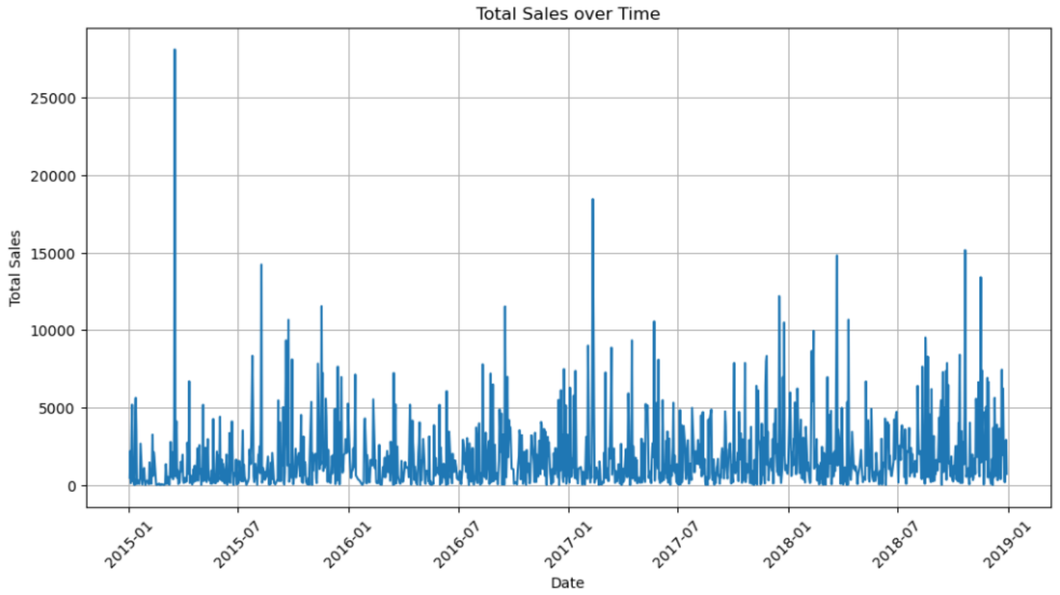
```
Product sold majorly from each category:
                     Category  \
Category
Furniture             Furniture
Office Supplies  Office Supplies
Technology            Technology

                                              Product Name      Sales
Category
Furniture            HON 5400 Series Task Chairs for Big and Tall  21870.576
Office Supplies  Fellowes PB500 Electric Punch Plastic Comb Bin...  27453.384
Technology                 Canon imageCLASS 2200 Advanced Copier  61599.824
```



MODELLING

3. Date with the maximum sells of that superstore

In [12]:
```
import pandas as pd
import matplotlib.pyplot as plt

# Read the cleaned dataset into a pandas DataFrame
df = pd.read_csv('superstore_cleaned_dataset.csv')

# Convert the "Order Date" column to a datetime format
df['Order Date'] = pd.to_datetime(df['Order Date'])

# Group the data by "Order Date" and calculate the total sales for each date
date_sales = df.groupby('Order Date')['Sales'].sum().reset_index()

# Find the date with the maximum total sales
date_with_max_sales = date_sales.loc[date_sales['Sales'].idxmax()]

# Extract the date and maximum sales value
max_sales_date = date_with_max_sales['Order Date']
max_sales_value = date_with_max_sales['Sales']

# Plot the total sales over time
plt.figure(figsize=(12, 6))
plt.plot(date_sales['Order Date'], date_sales['Sales'])
plt.xlabel('Date')
plt.ylabel('Total Sales')
plt.title('Total Sales over Time')
plt.xticks(rotation=45)
plt.grid(True)
plt.show()

print("Date with the maximum sales:", max_sales_date)
print("Maximum sales value on that date:", max_sales_value)
```




```
Date with the maximum sales: 2015-03-18 00:00:00
Maximum sales value on that date: 28106.716
```

MODELLING

4. Average difference between the Order Date and Ship Date

[13]:
```
import pandas as pd

# Read the cleaned dataset into a pandas DataFrame
df = pd.read_csv('superstore_cleaned_dataset.csv')

# Convert the "Order Date" and "Ship Date" columns to datetime format
df['Order Date'] = pd.to_datetime(df['Order Date'])
df['Ship Date'] = pd.to_datetime(df['Ship Date'])

# Calculate the time difference between "Ship Date" and "Order Date"
df['Time Difference'] = df['Ship Date'] - df['Order Date']

# Calculate the average time difference
average_time_difference = df['Time Difference'].mean()

print("Average difference between Order Date and Ship Date:", average_time_difference)
```

```
Average difference between Order Date and Ship Date: 9 days 04:23:54.262948207
```

5. Commonly used shipping method

14]:
```
import pandas as pd

# Read the cleaned dataset into a pandas DataFrame
df = pd.read_csv('superstore_cleaned_dataset.csv')

# Count the occurrences of each ship mode
ship_mode_counts = df['Ship Mode'].value_counts()

# Find the most commonly used ship mode (the one with the highest frequency,
most_common_ship_mode = ship_mode_counts.idxmax()

print("The most commonly used ship mode:", most_common_ship_mode)
```

```
The most commonly used ship mode: Standard Class
```

MODELLING

6. Customer who buys maximum products

[15]:
```python
import pandas as pd

# Read the cleaned dataset into a pandas DataFrame
df = pd.read_csv('superstore_cleaned_dataset.csv')

# Group the data by "Customer Name" and calculate the total number of products purchased by each customer
customer_product_count = df.groupby('Customer Name')['Product ID'].count().reset_index()

# Find the customer who buys the maximum number of products
customer_with_max_products = customer_product_count.loc[customer_product_count['Product ID'].idxmax()]

# Extract the customer name and the maximum product count
max_products_customer_name = customer_with_max_products['Customer Name']
max_product_count = customer_with_max_products['Product ID']

print("Customer who buys the maximum number of products:", max_products_customer_name)
print("Maximum number of products purchased:", max_product_count)
```

```
Customer who buys the maximum number of products: William Brown
Maximum number of products purchased: 35
```

7. City in which maximum sell has done

[16]:
```python
import pandas as pd

# Read the cleaned dataset into a pandas DataFrame
df = pd.read_csv('superstore_cleaned_dataset.csv')

# Group the data by "City" and calculate the total sales for each city
city_sales = df.groupby('City')['Sales'].sum().reset_index()

# Find the city with the maximum total sales
city_with_max_sales = city_sales.loc[city_sales['Sales'].idxmax()]

# Extract the city name and the maximum sales value
max_sales_city = city_with_max_sales['City']
max_sales_value = city_with_max_sales['Sales']

print("City with the maximum sales:", max_sales_city)
print("Maximum sales value in that city:", max_sales_value)
```

```
City with the maximum sales: New York City
Maximum sales value in that city: 252462.547
```

MODELLING

8. State in which maximum sell has done

17]:

```
import pandas as pd

# Read the cleaned dataset into a pandas DataFrame
df = pd.read_csv('superstore_cleaned_dataset.csv')

# Group the data by "State" and calculate the total sales for each state
state_sales = df.groupby('State')['Sales'].sum().reset_index()

# Find the state with the maximum total sales
state_with_max_sales = state_sales.loc[state_sales['Sales'].idxmax()]

# Extract the state name and the maximum sales value
max_sales_state = state_with_max_sales['State']
max_sales_value = state_with_max_sales['Sales']

print("State with the maximum sales:", max_sales_state)
print("Maximum sales value in that state:", max_sales_value)
```

```
State with the maximum sales: California
Maximum sales value in that state: 446306.4635
```

RESULTS

I made an analysis on the data set of the superstore that in given superstore in United States-

1. Category with the most sales is 'Technology'
2. In furniture category, the maximum sell is of 'HON 5400 Series Task Chairs for Big and Tall' and the sells are - 21870.576
3. In Office Supplies category, the maximum sell is of 'Fellowes PB500 Electric Punch Plastic.....' and the sells are - 27453.384
4. In Technology category, the maximum sell is of 'Canon imageCLASS 2200 Advanced Copier' and the sells are - 61599.824
5. Date with the maximum sales: 2015-03-18, Maximum sales value on that date: 28106.716

RESULTS

6. Average difference between Order Date and Ship Date: 9 days 04:23:54.262948207

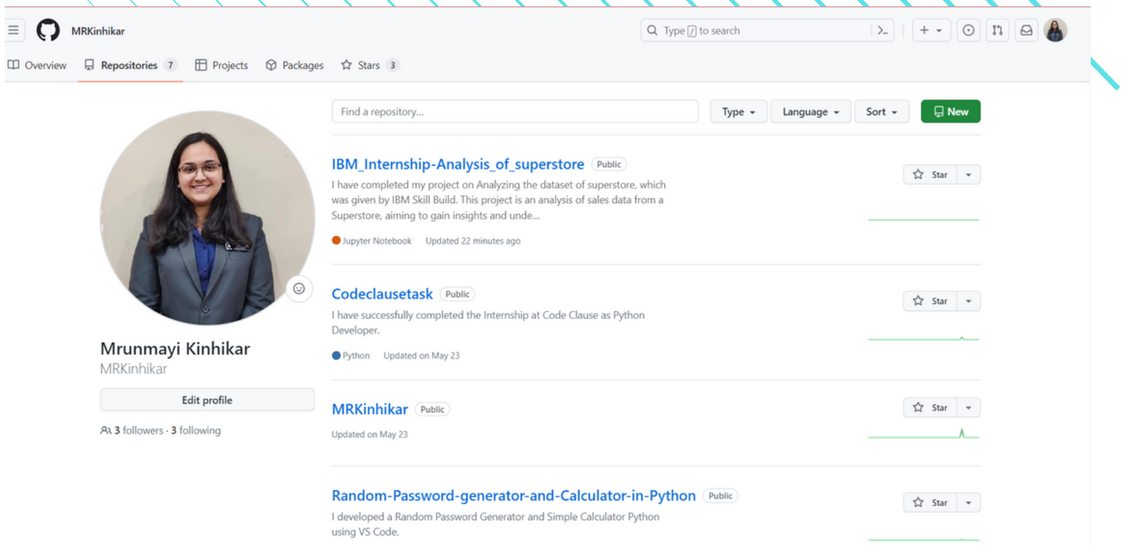
7. The most commonly used ship mode: Standard Class

8. Customer who buys the maximum number of products: William Brown
Maximum number of products purchased: 35

9. City with the maximum sales: New York City
Maximum sales value in that city: 252462.547

10. State with the maximum sales: California
Maximum sales value in that state: 446306.4635

LINKS-



Github repository link -

https://github.com/MRKinhikar/IBM_Internship-Analysis_of_superstore/blob/main/IBM%20Internship%20-%20Analysis%20of%20superstore.ipynb

Thank you !