Name – Ashutosh Raturi

ashu.raturi621@gmail.com

Organization – DGT

College – Graphic Era Deemed to be University

State – Uttarakhand

Domain – Data Analytics

Internship Start and End date – From 12/06/23 till 24/07/23

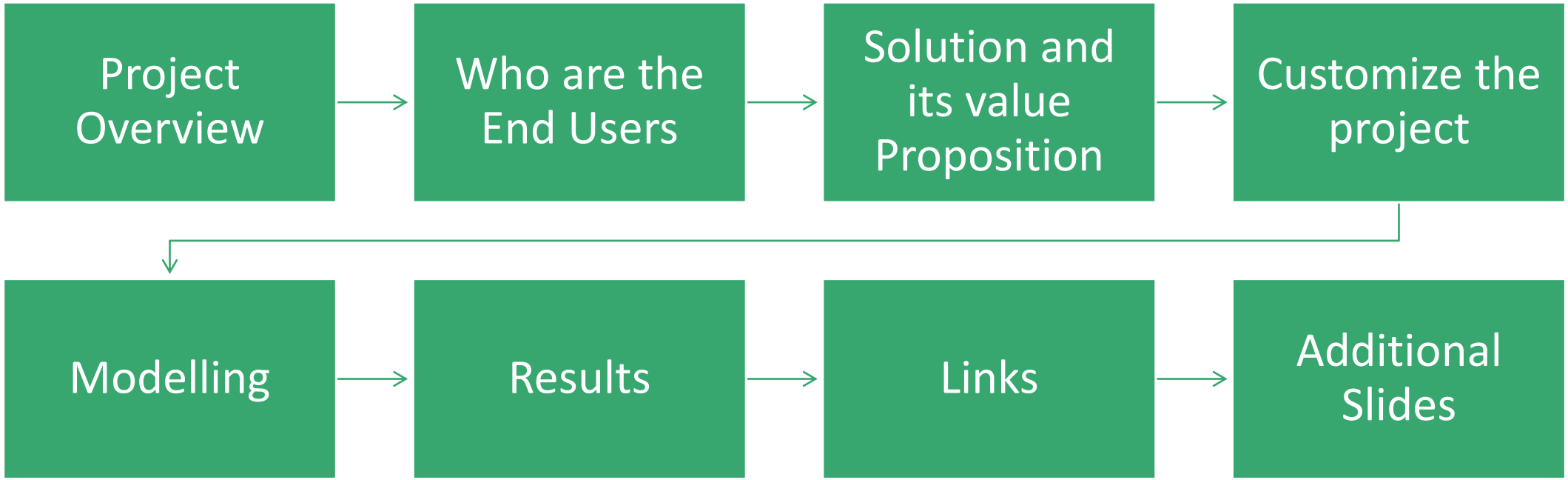
PROJECT TITLE : Analysis of Superstore Dataset

Introduction :

This project focuses on the optimization of retail store performance using data analysis techniques, with a specific case study based on the Superstore dataset. The dataset contains information about various aspects of the store's operations, including sales, customer demographics, product categories, and geographical regions. By analyzing this dataset, we aim to identify key trends, patterns, and factors that impact the store's overall performance.

By conducting a detailed analysis of the Superstore dataset, we will provide valuable recommendations to the store management on how to enhance sales, improve customer experience, and streamline operational processes. We aim to identify opportunities for improvement and make data-driven recommendations to optimize store performance.

AGENDA


Project Overview
Who are the End Users
Solution and its value Proposition
Customize the project
Modelling
Results
Links
Additional Slides

PROJECT OVERVIEW

Purpose: **The purpose of this project is to utilize data analysis techniques to optimize the performance of a retail store. By analyzing the Superstore dataset, the project aims to provide valuable insights and recommendations to improve sales revenue, enhance customer satisfaction, and streamline operational efficiency.**

Scope: **The scope of this project revolves around the Superstore dataset, a comprehensive simulation of a retail company's operations. The analysis covers sales metrics, customer behavior, and operational processes. The project will explore data distribution, trends, and correlations, focusing on customer segmentation, sales performance, and operational efficiency.**

Objective: **The primary objective of this project is to conduct a detailed data analysis of the Superstore dataset and extract actionable insights. By analyzing sales metrics to identify factors influencing sales fluctuations. Studying customer demographics and behavior to personalize marketing strategies. Evaluating operational processes to optimize resource allocation and reduce costs. Potentially applying predictive modeling to forecast future sales trends and customer behavior. Providing data-driven recommendations to enhance overall store performance and customer experience.**

WHO ARE THE END USERS

Retail Store Management: Decision-makers seeking data-driven insights to optimize sales, improve customer satisfaction, and streamline operations.

Characteristics: Decision-makers, industry expertise.

Need: Optimize sales, enhance efficiency, stay competitive.

Benefit: Informed decisions, improved profitability, sustainable success.

Marketing and Sales Teams: Customer-oriented teams looking for customer segmentation data and sales trends to create personalized marketing strategies.

Characteristics: Customer-focused, creative.

Need: Customer segmentation, personalized marketing.

Benefit: Increased conversions, customer loyalty.

SOLUTION AND ITS VALUE PROPOSITION

Solution:

The solution for the "Analysis of Superstore dataset" project involves conducting a comprehensive analysis of the Superstore dataset to gain insights into sales trends, customer behavior and operational efficiency.

Value Proposition:

Enhance Sales & Profitability: Optimize sales strategies, identify high-value customers, and popular products for increased revenue and profitability.

- **Improve Customer Satisfaction: Personalize marketing, understand preferences, and foster loyalty for better customer experiences.**
- **Streamline Operations: Identify inefficiencies, optimize resource allocation, and reduce costs for streamlined operations.**
- **Stay Competitive: Data-driven decisions adapt to market trends and customer demands for a competitive edge.**
- **Drive Sustainable Growth: Empower with data-backed solutions, ensuring long-term success in the retail market.**

Customize the project and make it my own

Modify Data Preprocessing: Adapted the data preprocessing to suit my project requirement by handling missing values, outliers, and data formatting. This help me get the more precise and accurate data which can be used for the analysis purpose.

Descriptive Analytics: Utilize descriptive analytics techniques to summarize and present key information about sales trends, customer behaviour, and operational performance within the Superstore dataset. This includes calculating summary statistics, generating frequency distributions, and identifying important patterns or trends.

Interactive Visualization: Using interactive and visually appealing data visualizations that allow stakeholders to explore insights intuitively and my solution stands out by utilizing the powerful libraries Matplotlib and Seaborn. These libraries offer extensive customization options, allowing for the creation of visually appealing and insightful charts, graphs, and plots. By leveraging the capabilities of Matplotlib and Seaborn, my solution presents data in a visually engaging manner, enhancing the understanding of complex patterns and relationships within the Superstore dataset.

MODELLING

Descriptive Modeling: By applying descriptive modeling to the Superstore dataset, we understand data's current state, trends, and patterns, providing valuable insights for decision-making.

Methodology :

- **Data Collection and Preprocessing: We gather the retail store's historical data, including sales records, customer details, and operational metrics. Data preprocessing involves handling missing values, outliers, and data formatting to ensure data integrity and quality.**
- **Exploratory Data Analysis (EDA): We perform EDA to gain initial insights into the dataset's characteristics. Through data visualization and summary statistics, we identify trends, distributions, and potential relationships between variables.**

Framework Used: For this project, we utilize popular data analysis libraries in Python, such as Pandas, NumPy, Matplotlib and Seaborn. These libraries provide a robust and efficient framework for data manipulation and data analysis.

By adhering to these technology principles, the solution deliver actionable insights and improved decision-making for optimizing retail store performance.

Importing necessory library

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

Loading the Dataset

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

	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

```
df.dtypes
```

```
Ship Mode        object
Segment          object
Country          object
City             object
State            object
Postal Code       int64
Region           object
Category         object
Sub-Category     object
Sales           float64
Quantity          int64
Discount        float64
Profit          float64
dtype: object
```

Checking for the Duplicate values

```
df.duplicated().sum()
```

```
17
```

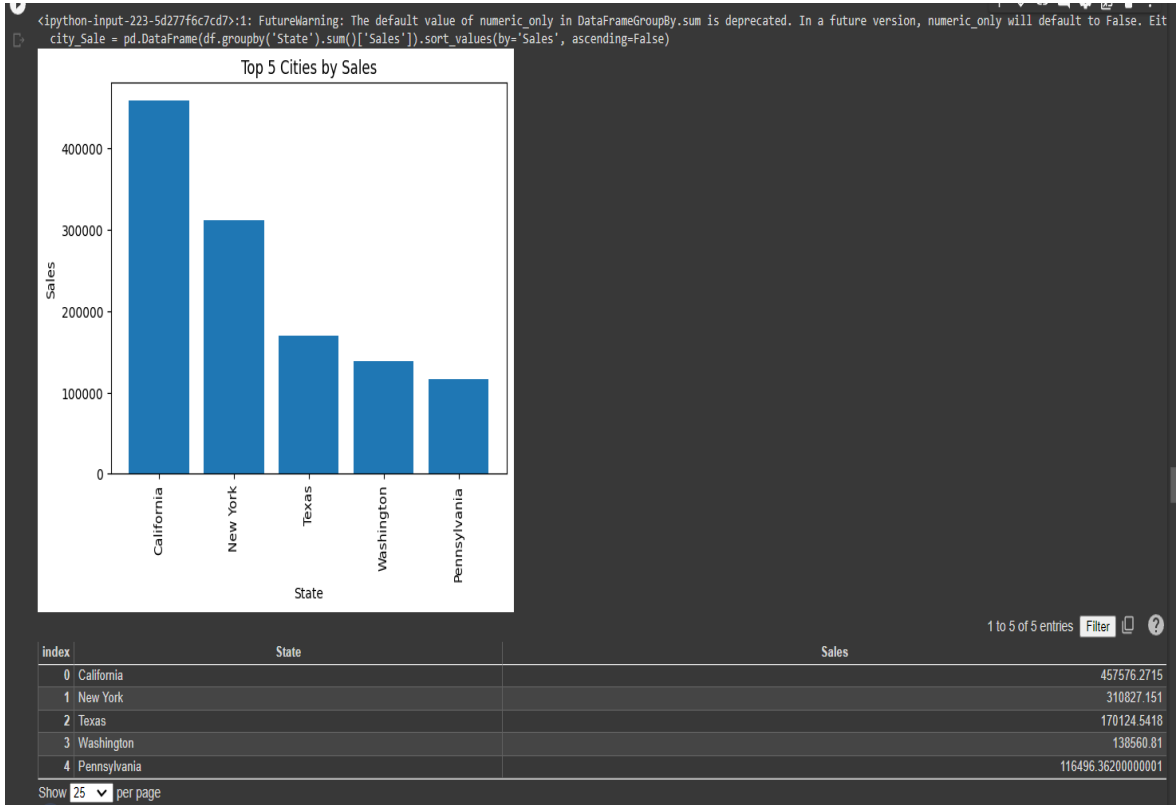
Droping the duplicate values

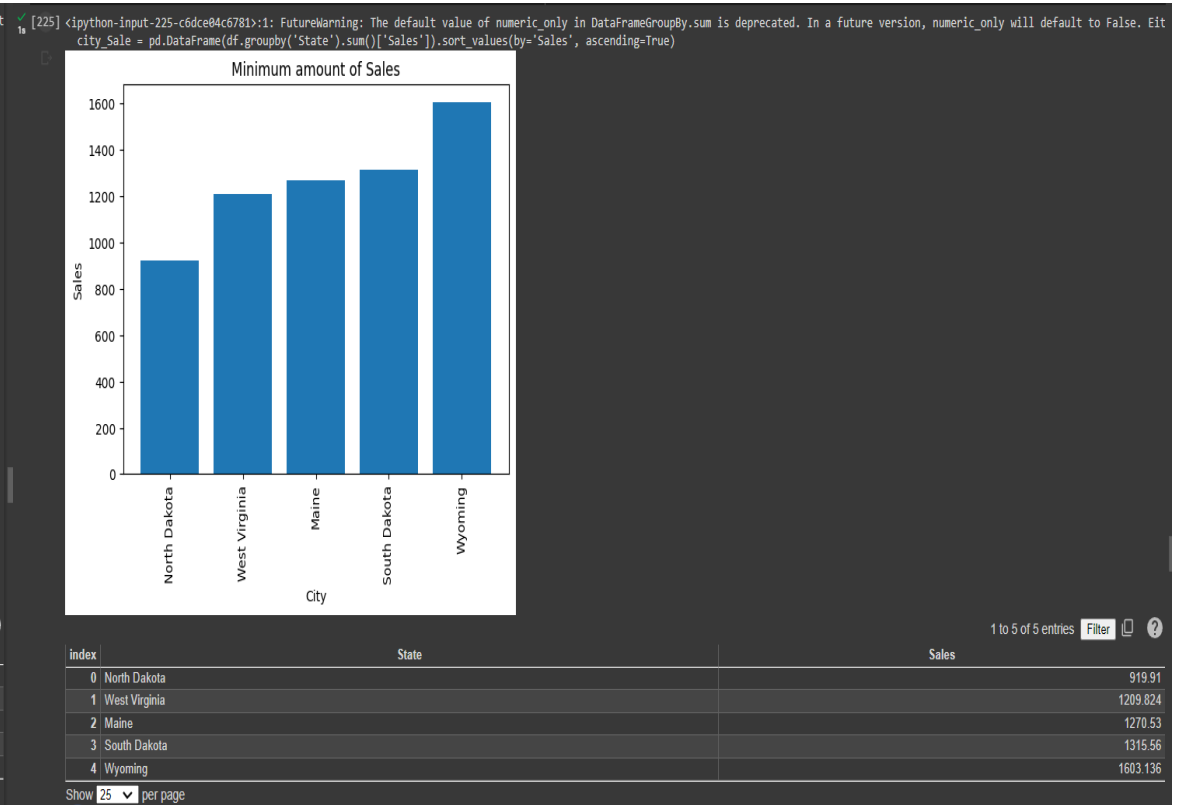
```
df=df.drop_duplicates()
```

```
df.duplicated().sum()
```

```
0
```

Results






```
<ipython-input-223-5d277f6c7cd7>:1: FutureWarning: The default value of numeric_only in DataFrameGroupBy.sum is deprecated. In a future version, numeric_only will default to False. Eit
  city_Sale = pd.DataFrame(df.groupby('State').sum()['Sales']).sort_values(by='Sales', ascending=False)
```


1 to 5 of 5 entries Filter

index	State	Sales
0	California	457576.2715
1	New York	310827.151
2	Texas	170124.5418
3	Washington	138560.81
4	Pennsylvania	116496.36200000001

Show 25 per page


```
[225] <ipython-input-225-c6dce04c6781>:1: FutureWarning: The default value of numeric_only in DataFrameGroupBy.sum is deprecated. In a future version, numeric_only will default to False. Eit
  city_Sale = pd.DataFrame(df.groupby('State').sum()['Sales']).sort_values(by='Sales', ascending=True)
```

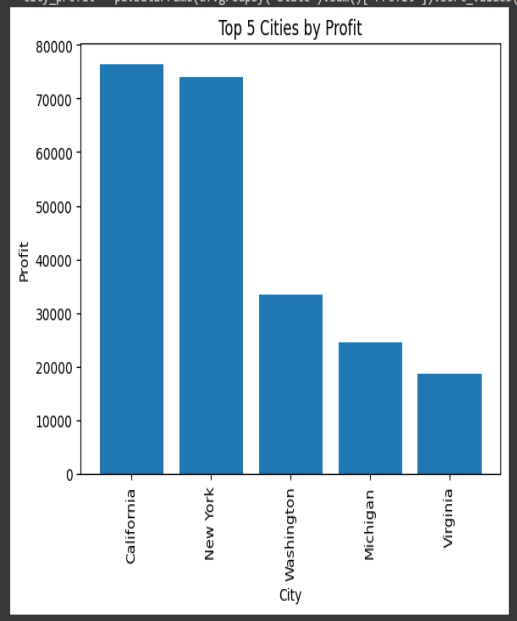

1 to 5 of 5 entries Filter

index	State	Sales
0	North Dakota	919.91
1	West Virginia	1209.824
2	Maine	1270.53
3	South Dakota	1315.56
4	Wyoming	1603.136

Show 25 per page

- **The company has highest sales in the state of California which is around 450k.**
- **New York is the state that constitutes the second highest sales for the company of around 300k.**
- **Lowest amount of sale is recorded in North Dakota 919.**

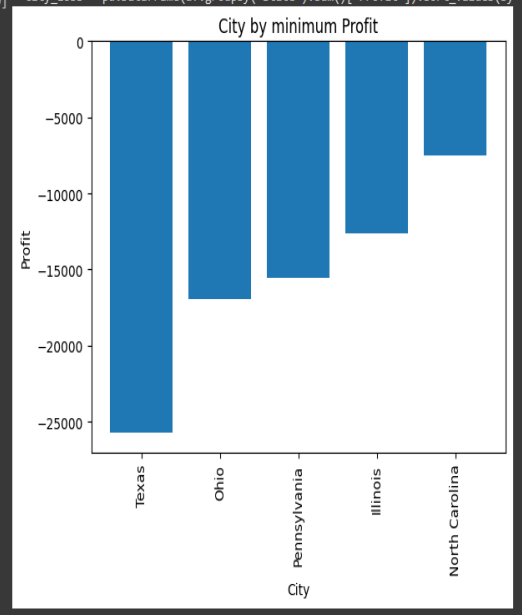
```
<ipython-input-228-c720afa0d616>:1: FutureWarning: The default value of numeric_only in DataFrameGroupBy.sum is deprecated. In a future version, numeric_only will default to False. Eit
  city_profit = pd.DataFrame(df.groupby('State').sum()['Profit']).sort_values(by='Profit', ascending=False)
```



1 to 5 of 5 entries Filter

index	State	Profit
0	California	76330.7891
1	New York	74015.4622
2	Washington	33368.2375
3	Michigan	24428.0903
4	Virginia	18597.9504

```
 city_loss = pd.DataFrame(df.groupby('State').sum()['Profit']).sort_values(by='Profit', ascending=True)
```

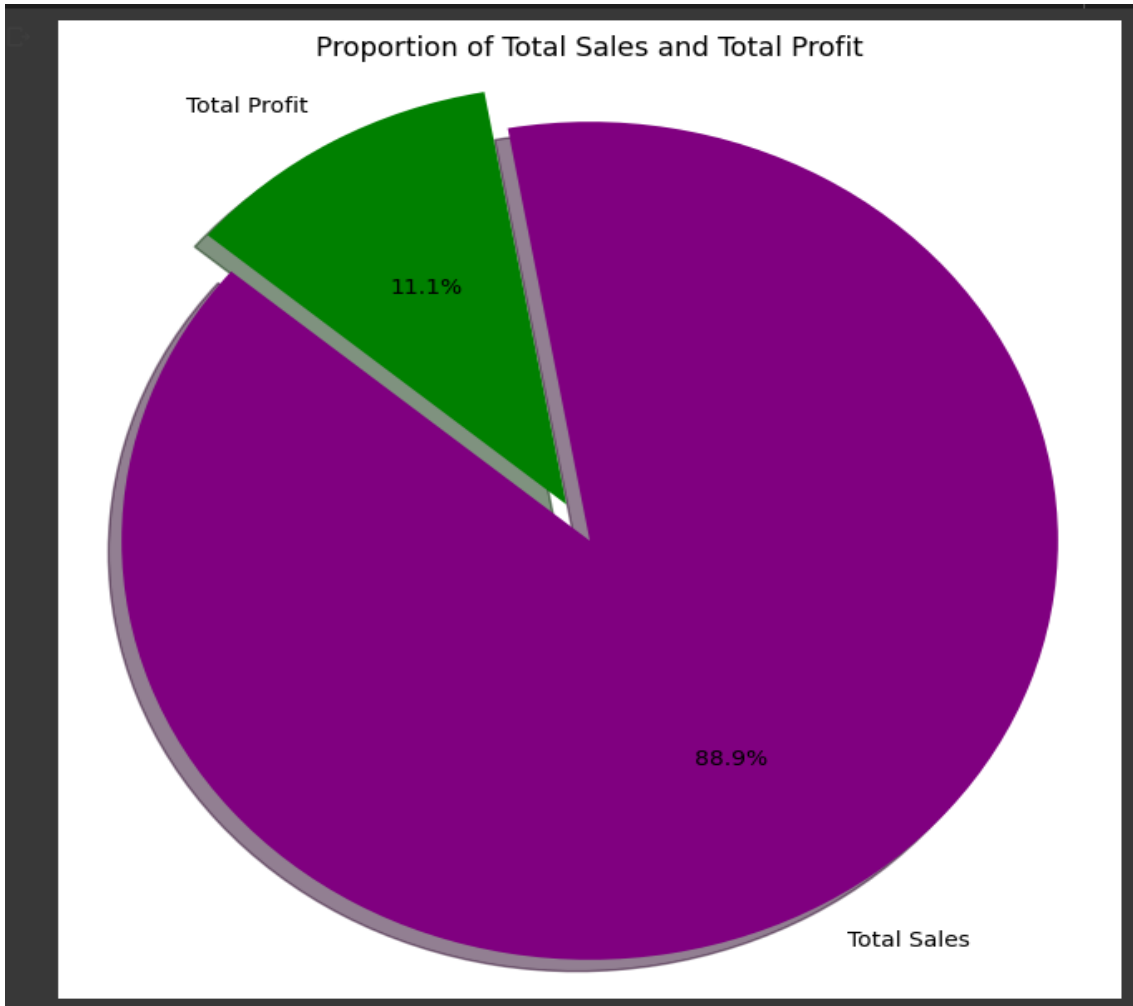


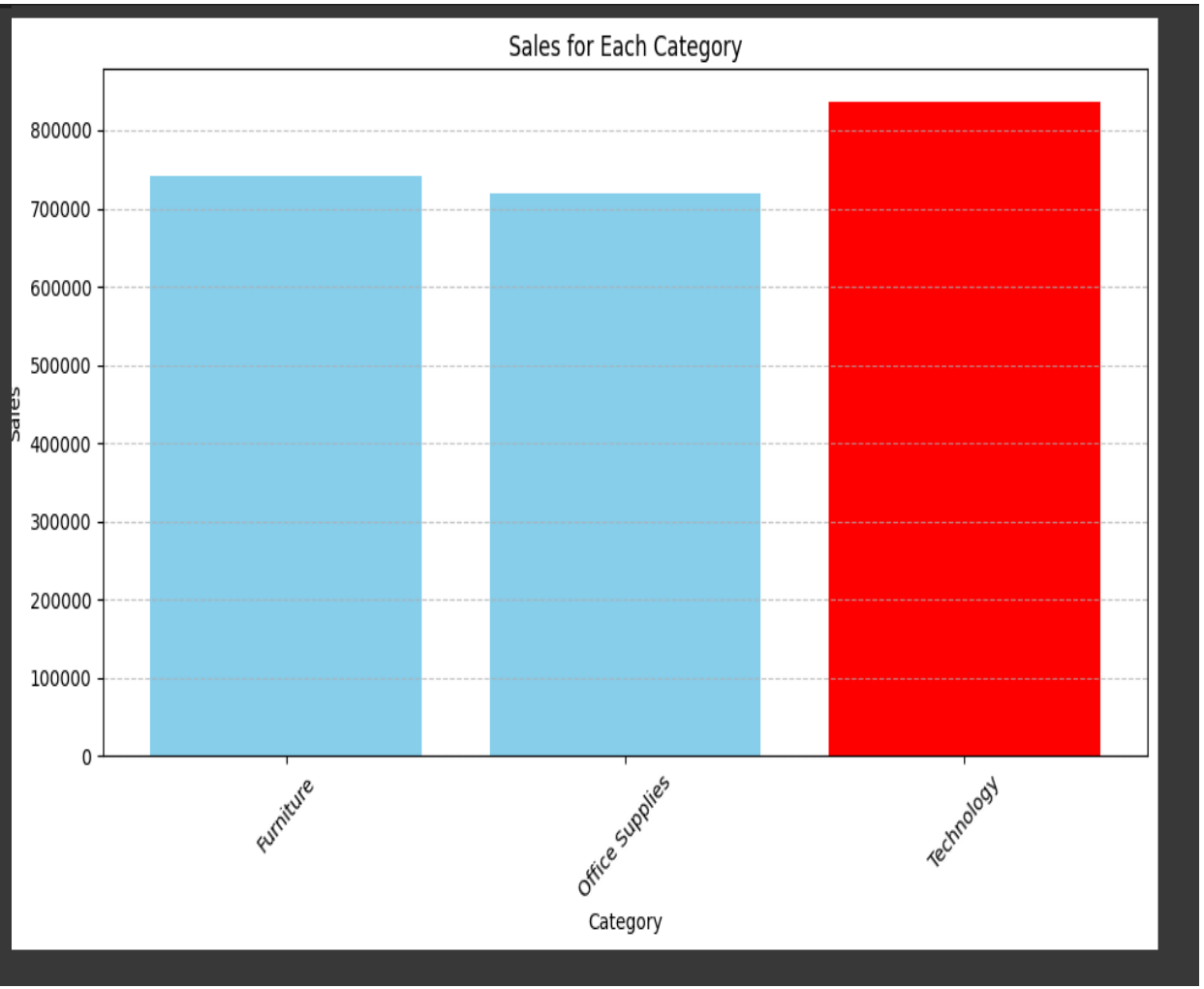
1 to 5 of 5 entries Filter

index	State	Profit
0	Texas	-25750.9833
1	Ohio	-16959.3178
2	Pennsylvania	-15565.4035
3	Illinois	-12601.65
4	North Carolina	-7490.9122

Show 25 per page

- **Cities like California, New York, Washington has recorded highest amount of profit.**
- **Company must focus on cities which made loss like Texas, Ohio, Pennsylvania etc.**
- **They should analyze their resource usage and the categories which are being sold in those states.**





- **Company has highest sales in the category of Technology.**
- **Proportion of total sales and total profit is shown above which need to be improved more.**

links

Git hub link:

IBM_Data_Analytics/superstore_data.ipynb at main · Ashutosh621/IBM_Data_Analytics (github.com)

Research Paper:

Here are some references for sales analysis on Superstore dataset:

• Chakraborty, M. (2020). Sales Analysis of Superstore using Power BI. Kaggle.

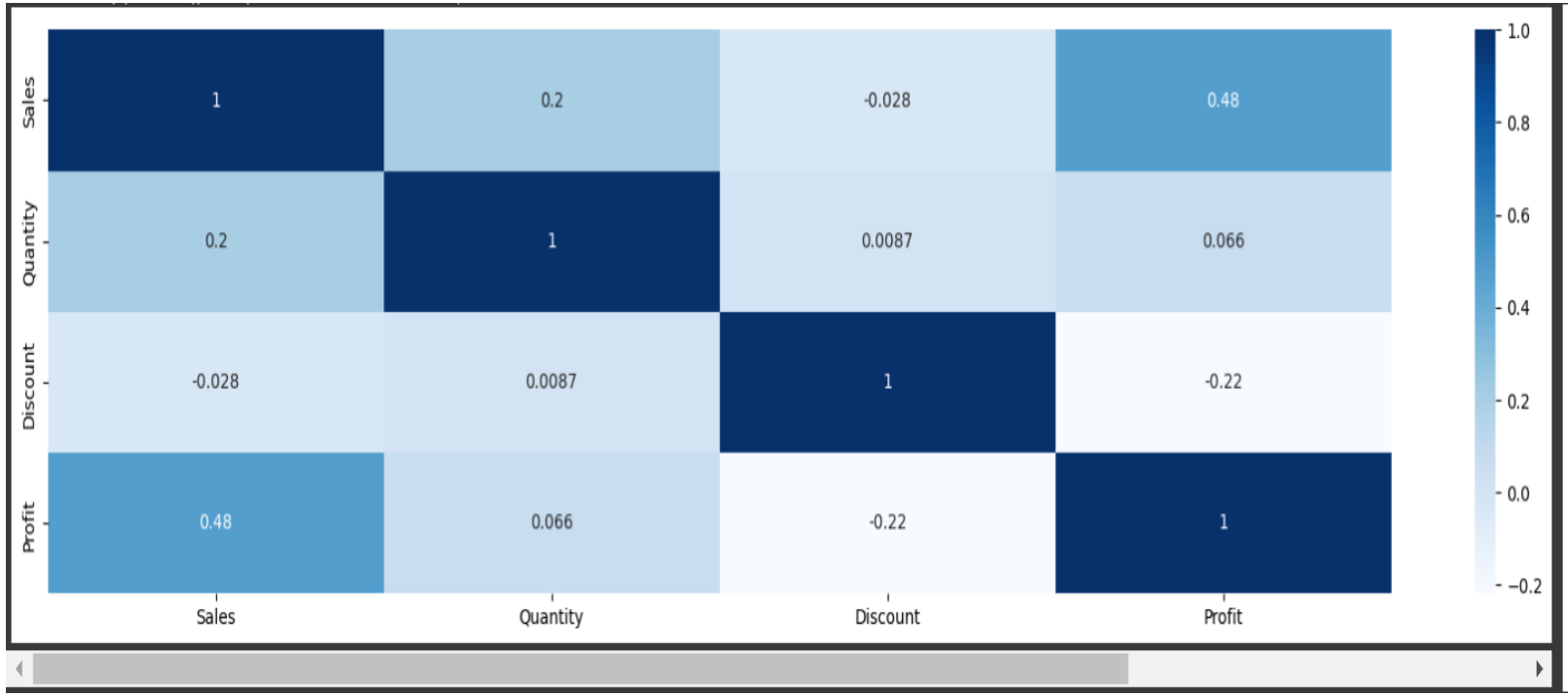
https://www.kaggle.com/moumoyesh/sales-analysis-of-superstore-using-power-bi

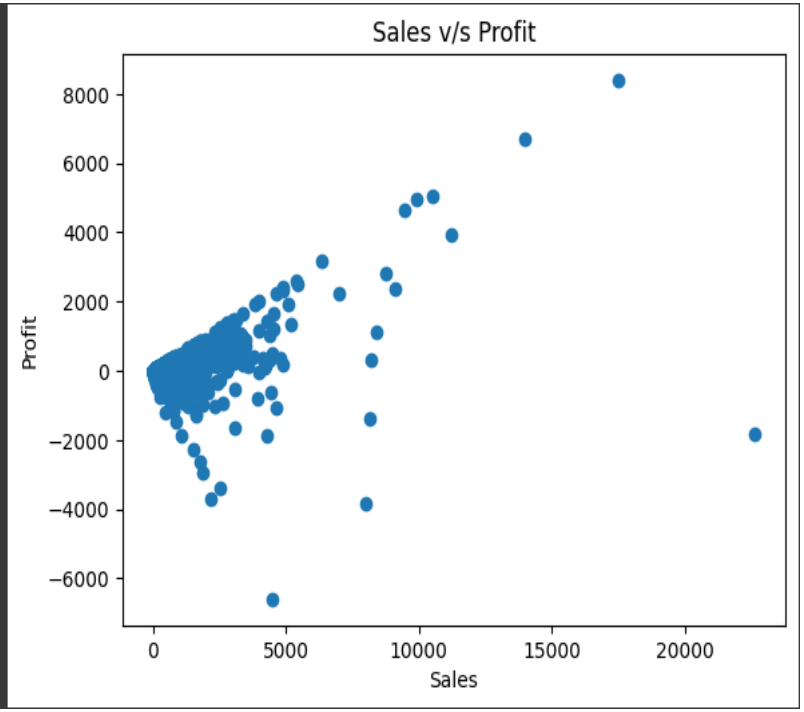
Microsoft. (n.d.). Analyse and visualize Superstore data in Power BI. https://powerbi.microsoft.com/enus/tutorials/analyze-and-visualize-superstore-data/

https://towardsdatascience.com/sales-analysis-of-superstore-dataset-using-power-bi-1432f74fa62e

• Pranav, B. (2021). Sales Analysis of Superstore Data using Power BI. Analytics Vidhya.

Additional Slide





- **At lower sales, a business either gets profit or loss. But, when the sales are high, a business mostly gets profit.**
- **It seems Profit is highly correlated with Sales.**

THANK YOU..