**Source Code Documentation**

This document provides a detailed description of the data analysis process conducted in the **Superstore Business Analysis** project.

🔗 **GitHub Link:** [Project Repository](https://github.com/Maianh2510/Project--Superstore-Business-Analysis.git)

The script performs data loading, cleaning, processing, and visualization to support analysis in **Power BI**.

**1. Import Required Libraries**

The script utilizes the following libraries for data processing and visualization:

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

**2. Load Data**

The dataset is loaded from an **Excel file** using pandas. The script displays the first five rows and general information about the dataset.

df = pd.read\_excel('/content/Superstore.xlsx')

df.head()

df.info()

**3. Data Cleaning**

The cleaning steps include duplicate checks and data type conversion.

**Check for Duplicates**

df.duplicated()

**Convert Specific Columns to String Type**

columns\_string = ['Customer\_ID', 'Customer\_Name', 'Postal\_Code', 'Product\_ID', 'Product\_Name']

df[columns\_string] = df[columns\_string].astype('string')

df[columns\_string].dtypes

**Check Unique Values in Categorical Columns**

df['Ship\_Mode'].unique()

df['Segment'].unique()

df['Country'].unique()

df['City'].unique()

df['State'].unique()

df['Region'].unique()

df['Category'].unique()

df['Sub\_Category'].unique()

**Convert Categorical Columns to Category Type**

columns\_category = ['Ship\_Mode', 'Segment', 'Country', 'City', 'State', 'Region', 'Category', 'Sub\_Category']

df[columns\_category] = df[columns\_category].astype('category')

df[columns\_category].dtypes

**Convert Date Columns to DateTime Format**

df['Order\_Date'] = pd.to\_datetime(df['Order\_Date'])

df['Order\_Date'].head()

df['Ship\_Date'] = pd.to\_datetime(df['Ship\_Date'])

df['Ship\_Date'].head()

**Recheck data**

df.info()

**4. Handling Missing Values**

Missing values in **'Ship\_Mode'** and **'Country'** columns are filled using the mode value.

df.isnull().sum()

df['Ship\_Mode'] = df['Ship\_Mode'].fillna(df['Ship\_Mode'].mode()[0])

df['Country'] = df['Country'].fillna(df['Country'].mode()[0])

df.describe()

**5. Handling Outliers**

Box plots are used to detect outliers in the **'Sales'** and **'Profit'** columns. The **Interquartile Range (IQR) method** is applied to remove them.

**Visualizing Outliers with Box Plots**

plt.figure(figsize=(8,5))

sns.boxplot(data=df[['Sales']])

plt.title("Boxplot for Sales")

plt.show()

plt.figure(figsize=(8,5))

sns.boxplot(data=df[['Profit']])

plt.title("Boxplot for Profit")

plt.show()

**Function to Remove Outliers**

def rm\_outlier(col\_name):

Q1 = df[col\_name].quantile(0.25)

Q3 = df[col\_name].quantile(0.75)

IQR = Q3 - Q1

lower\_bound = Q1 - 1.5 \* IQR

upper\_bound = Q3 + 1.5 \* IQR

df\_result = df[(df[col\_name] >= lower\_bound) & (df[col\_name] <= upper\_bound)]

return df\_result

df = rm\_outlier(col\_name='Sales')

df = rm\_outlier(col\_name='Profit')

**6. Correlation Analysis**

A **correlation matrix** is generated to assess relationships between variables and visualized using a **heatmap**.

numerical\_columns = ['Sales', 'Quantity', 'Discount', 'Profit']

correlation\_matrix = df[numerical\_columns].corr()

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm')

plt.title('Sales Correlation Matrix')

plt.show()

**7. Export Processed Data**

After cleaning and processing, the dataset is exported as a **CSV file** for visualization in **Power BI**.

df.to\_csv("sales.csv", index=False)