**Project Report: Data Cleaning and Exploratory Data Analysis (EDA)**

**1. Introduction**

This project involves cleaning a raw dataset and performing Exploratory Data Analysis (EDA) to uncover meaningful insights. The dataset contains sales information categorized by order IDs, customer segments, and shipping modes. The objective is to:

* Clean and transform the dataset into a structured format.
* Perform EDA to analyze sales patterns and customer behavior.
* Visualize the insights using Power BI.

**2. Data Cleaning Process**

The raw dataset contained inconsistencies such as merged headers, missing values, and unstructured columns. The cleaning process involved:

**2.1 Loading the Data**

import pandas as pd

# Load the raw data

file\_path = "C:\\Users\\Admin\\OneDrive\\Desktop\\Data analyst\\Project 2\\Ditry Data Sample.xlsx"

df\_dirty = pd.read\_excel(file\_path, sheet\_name=0)

**2.2 Renaming Columns and Dropping Unnecessary Rows**

# Rename columns for clarity

df\_dirty.columns = ["Segment", "Consumer", "Consumer\_2", "Consumer\_3", "Consumer\_4", "Consumer Total",

"Corporate", "Corporate\_2", "Corporate\_3", "Corporate\_4", "Corporate Total",

"Home Office", "Home\_Office\_2", "Home\_Office\_3", "Home\_Office\_4", "Home Office Total"]

# Drop irrelevant header rows

df\_dirty = df\_dirty.iloc[2:].reset\_index(drop=True)

**2.3 Transforming Data Using Melt Function**

# Normalize dataset using melt

df\_melted = df\_dirty.melt(id\_vars=["OrderID"], var\_name="Category", value\_name="Sales")

**2.4 Extracting Customer Segment and Ship Mode**

category\_mapping = {

"Consumer": ("Consumer", "First Class"), "Consumer\_2": ("Consumer", "Same Day"),

"Consumer\_3": ("Consumer", "Second Class"), "Consumer\_4": ("Consumer", "Standard Class"),

"Corporate": ("Corporate", "First Class"), "Corporate\_2": ("Corporate", "Same Day"),

"Corporate\_3": ("Corporate", "Second Class"), "Corporate\_4": ("Corporate", "Standard Class"),

"Home Office": ("Home Office", "First Class"), "Home\_Office\_2": ("Home Office", "Same Day"),

"Home\_Office\_3": ("Home Office", "Second Class"), "Home\_Office\_4": ("Home Office", "Standard Class"),

}

# Assign segment and ship mode

df\_melted["Segment"] = df\_melted["Category"].map(lambda x: category\_mapping.get(x, (None, None))[0])

df\_melted["Ship Mode"] = df\_melted["Category"].map(lambda x: category\_mapping.get(x, (None, None))[1])

**2.5 Finalizing Cleaned Data**

# Select required columns

df\_final = df\_melted[["Segment", "Ship Mode", "OrderID", "Sales"]].dropna()

# Convert Sales column to numeric

df\_final["Sales"] = pd.to\_numeric(df\_final["Sales"], errors="coerce")

**2.6 Saving Cleaned Data**

# Save cleaned data to an Excel file

cleaned\_file\_path = "C:\\Users\\Admin\\OneDrive\\Desktop\\Data analyst\\Project 2\\Cleaned\_Data.xlsx"

df\_final.to\_excel(cleaned\_file\_path, sheet\_name="Cleaned Data", index=False)

**3. Exploratory Data Analysis (EDA)**

After cleaning, we perform EDA to derive insights from the structured dataset.

**3.1 Data Summary**

# Check dataset structure

df\_final.info()

**3.2 Sales Distribution**

import seaborn as sns

import matplotlib.pyplot as plt

plt.figure(figsize=(10, 5))

sns.histplot(df\_final["Sales"], bins=30, kde=True, color='blue')

plt.title("Sales Distribution")

plt.show()

**3.3 Sales by Customer Segment**

sns.boxplot(x="Segment", y="Sales", data=df\_final, palette="Set2")

plt.title("Sales by Customer Segment")

plt.show()

**3.4 Sales by Ship Mode**

sns.barplot(x="Ship Mode", y="Sales", data=df\_final, estimator=sum, palette="muted")

plt.title("Total Sales by Shipping Mode")

plt.show()

**4. Data Visualization in Power BI**

The cleaned dataset was imported into Power BI for visualization. Key insights were represented using:

* **Gradient Backgrounds:** Used for aesthetics and readability.
* **Sales Trend Charts:** Displayed sales growth over time.
* **Category-wise Comparisons:** Used to analyze revenue by segment.
* **Interactive Filters:** Enabled users to explore data dynamically.

**5. Conclusion**

This project successfully cleaned a raw dataset and performed an in-depth analysis to extract valuable business insights. The Power BI dashboard presents a user-friendly, interactive representation of these findings. Future improvements can involve integrating live sales data for real-time analytics.