Graduation Project Plan

For

DEPI

Training Profile Name

Google Data Analyst Specialist

Group Code

HRV482A_DKH2_DAT1_G1_DEPI2

By

Asmaa Mahgoub

Hamdy Awad

Saad El-Mohamady

Sarah El-Alfy

Seif El-Nakieb

Super Vised By

Eng. Mohamed El-Alfy

Project Idea: Store Sales Dataset Analysis

Project Scope:

The objective of this project is to analyze a store sales dataset to identify trends, forecast future sales, and provide actionable insights to decision-makers. The project will follow a structured approach, including data preprocessing, analysis, forecasting, and visualization.

Technical Framework:

- **Tools & Technologies:** Python (pandas, Matplotlib, Seaborn), Power Bi.
- **Project Duration:** 4 weeks.

Phase 1: Data Preprocessing (one week)

- Build the data model and clean/preprocess the dataset.
- Deliverables: Cleaned dataset and preprocessing notebook.

Phase 2: Analysis Questions Phase (one week)

- Define key analytical questions relevant to business decision-making.
- Deliverables: List of analysis questions derived from data.

Phase 3: Forecasting Questions Phase (one week)

- Identify forecasting questions and generate predictive insights.
- Deliverables: Visualization plots for forecasting insights.

Phase 4: Visualization & Reporting (one week)

- Build a Power Bi dashboard and prepare a final presentation.
- Deliverables: Interactive dashboard, final report, and presentation.

Challenges & Risk Management:

- **Data Quality Issues:** Handling missing values, inconsistencies, and duplicates.
- **Scalability:** Ensuring the analysis framework can be extended for future datasets.
- **Visualization Complexity:** Choosing the right visuals for effective communication.

Final Deliverables:

1. **Cleaned Dataset** – Ready for analysis.
2. **Preprocessing Notebook** – Documenting data transformation steps.
3. **Set of Analysis Questions** – Answered with data insights.
4. **Forecasting Visualizations** – Graphical representation of predictive insights.
5. **Power Bi Dashboard** – Interactive data visualization tool.
6. **Final Report & Presentation** – Summary of project findings and methodologies.

Phase 1: Data Preprocessing

The initial phase of this data analysis project focuses on **preparing and refining the cleaned sales dataset** to ensure its suitability for in-depth exploration. By implementing a structured preprocessing approach, we aim to **enhance data quality, establish meaningful relationships, and uncover actionable insights** for decision-makers.

Key Steps in Data Preprocessing:

Data Model Design:

- The dataset will be organized into **fact and dimension tables** to optimize query performance.
- Logical **relationships between sales transactions, product details, and regional data** will be defined to support multidimensional analysis.

Preprocessing & Cleaning Strategy:

- **Standardize date formats** and **normalize categorical variables** for consistency.
- **Detect and treat outliers** using statistical methods (e.g., IQR, Z-score) to minimize bias.
- **Validate data integrity** across different sources to ensure uniformity

Tools:

- **Python (pandas)**: For data manipulation and analysis
- **Matplotlib & Seaborn**: For creating visualizations to support findings.

Deliverables:

- Comprehensive list of analysis questions that can be answered with the dataset
- Preliminary findings from initial data exploration
- Recommendations for deeper analysis based on initial insights

Phase 2: Analysis Questions

After carefully examining the cleaned sales dataset, we identified several key analysis questions that would provide valuable insights for decision makers:

1. **Sales Performance by Product Category**
    - Which product categories generate the highest sales?
    - How do sub-categories perform within each main category?
2. **Geographic Sales Distribution**
    - Which regions and states contribute most to overall sales?
    - Are there specific cities that outperform others in sales?
3. **Customer Segmentation Analysis**
    - How do sales differ across customer segments?
    - Which segment is most profitable?
4. **Shipping Analysis**
    - How does shipping mode affect sales performance?
    - What's the relationship between shipping time (Order Date to Ship Date) and sales?
5. **Temporal Sales Patterns**
    - Are there monthly/quarterly trends in sales performance?
    - How do sales vary by day of the week?
6. **Product Performance**
    - What are the top 10 best-selling products by revenue?
    - Are there products that consistently underperform?
7. **Customer Behavior**
    - Which customers make the highest value orders?
    - What's the average order value by customer segment?

Tools Used

- **Python (pandas)**: For data manipulation and analysis
- **Matplotlib & Seaborn**: For creating visualizations to support findings.

Deliverables

1. **Comprehensive list of analysis questions** that can be answered with the dataset
2. **Preliminary findings** from initial data exploration
3. **Recommendations for deeper analysis** based on initial insights

Phase 3: Determining Forecasting Questions

Based on the sales trends identified in Week 2, we developed these forecasting questions:

1. **Sales Trend Forecasting**
    - What will be the sales for each product category in the next 6 months based on historical trends?
2. **Seasonal Patterns**
    - Are there predictable seasonal patterns we can use to forecast peak sales periods?
3. **Regional Growth Projections**
    - Which regions are likely to see the highest sales growth in the coming year?
4. **Customer Segment Growth**
    - How will sales distribution across customer segments change in the next year?
5. **Product Performance Forecasting**
    - Which products/sub-categories are likely to see increased demand?

Tools Used

- **Python (pandas)**: For data preparation and time series analysis
- **scikit-learn**: For implementing forecasting models
- **Matplotlib/Seaborn**: For visualizing trends and forecasts

Deliverables

- **Time series visualizations** showing historical trends and forecasts
- **Forecasted values** for key metrics (sales by category, region, etc.)
- **Confidence intervals** for forecasts to understand prediction reliability
- **Recommendations for inventory planning** based on forecasted demand

The analysis provides a solid foundation for data-driven decision-making regarding product focus, regional strategies, and customer segment target

Phase 4: Visualization & Reporting

This phase focuses on transforming analytical insights into actionable business intelligence through interactive dashboards and comprehensive reporting. The goal is to communicate effectively to stakeholders, enabling data-driven decision-making.

Key Activities:

- Interactive Dashboard Development (Power BI):
- Design an intuitive, user-friendly dashboard highlighting:
- Sales trends by product, region, and customer segment.
- Forecasting results with visual comparisons (actual vs. predicted).
- Key performance indicators (KPIs) for quick decision-making.
- Implement drill-down capabilities for granular analysis.

Final Report Preparation: Document the methodology, findings, and recommendations in a structured report includes:

- Summary of preprocessing steps.
- Key insights and forecasting.
- Visualizations (charts, heatmaps, time-series plots).
- Business implications and actionable strategies.

Presentation Design: Create a concise, visually engaging slideshow for stakeholders, Focus on:

- Problem statement and objectives.
- High-impact insights (e.g., top-performing products, seasonal trends).
- Forecasted opportunities and risks.

Tools & Deliverables:

- Tools: Power BI, Python (Matplotlib/Seaborn), PowerPoint

Deliverables:

- Power BI Dashboard (Interactive visualization tool).
- Final Report (PDF/documentation of all phases).
- Presentation (Stakeholder briefing slides).

"This document outlines the structured approach for the Store Sales Dataset Analysis project, ensuring alignment with business objectives and methodological. By following this plan, the team will deliver actionable insights to drive strategic decision-making."

Project Screens

Cleaning phase Screens:

1-

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

2-

```
# Convert relevant columns to appropriate data types
df['Order Date'] = pd.to_datetime(df['Order Date'], format='%d/%m/%Y')
df['Ship Date'] = pd.to_datetime(df['Ship Date'], format='%d/%m/%Y')
df['Postal Code'] = df['Postal Code'].astype(object)
```

```
## Standaralize Column Names:
# replace space in column headers with _
df.columns = df.columns.str.replace(' ', '_')
# replace - in column headers with _
df.columns = df.columns.str.replace('-', '_')
# Check Changes
df.info()
```

3-

```
# inspecting row with the NaN value
df[df.isna().any(axis=1)]
```

	Row_ID	Order_ID	Order_Date	Ship_Date	Ship_Mode	Cust
2234	2235	CA-2018-104066	2018-12-05	2018-12-10	Standard Class	C
5274	5275	CA-2016-162887	2016-11-07	2016-11-09	Second Class	S

4-

```
# show duplicate rows
df.duplicated()
```

	0
0	False
1	False
2	False
3	False

5-

```
# Standardize 'Ship Mode' categories
df['Ship_Mode'] = df['Ship_Mode'].str.strip().str.title()

# Standardize 'Country' categories
df['Country'] = df['Country'].str.strip().str.title()
```

```
[ ]
# Count Each category of ship mode
df['Ship_Mode'].value_counts()
```

Ship_Mode	count
Standard Class	5859
Second Class	1902
First Class	1501
Same Day	538

dtype: int64

6-

Investigating Outliers

```
[ ] # Finding Outliers Using IQR (Interquartile Range) Method
Q1 = df['Sales'].quantile(0.25)
Q3 = df['Sales'].quantile(0.75)
IQR = Q3 - Q1

# Define outlier Bounds
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Identify outliers
iqr_outliers = df[(df['Sales'] < lower_bound) | (df['Sales'] > upper_bound)]
print(f"Total transactions: {len(df)}")
print(f"Outliers detected: {len(iqr_outliers)} ")
print(f"Outliers Percentage: {len(iqr_outliers)/len(df):.1%}")
print(f"Outlier Bounds: > ${upper_bound:.2f} or < ${lower_bound:.2f}")
```

```
Total transactions: 9800
Outliers detected: 1145
Outliers Percentage: 11.7%
Outlier Bounds: > $500.64 or < $-272.79
```

7-

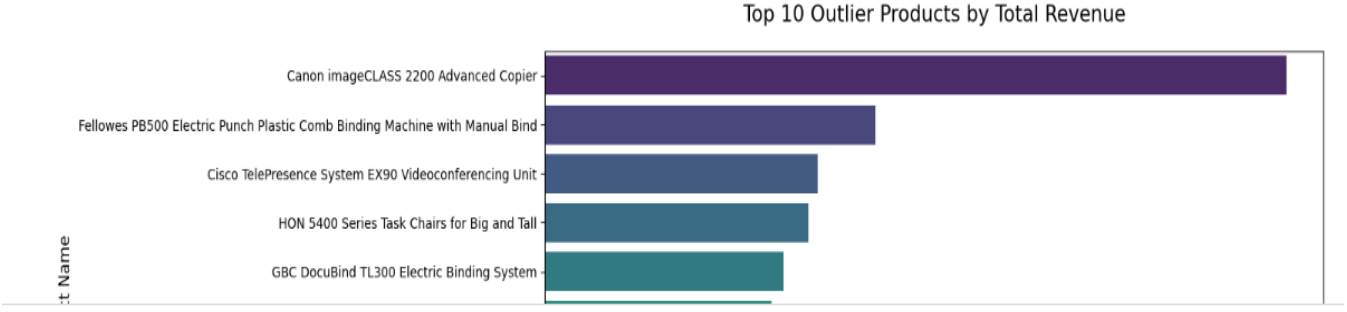
```
[ ] # Get top 10 outlier products by Count

    # Visualize 10 outlier products by Count
    top_outlier_products = iqr_outliers.groupby('Product_Name')['Sales'].sum()\
        .sort_values(ascending=False)\
        .head(10)

    # Create visualization
    plt.figure(figsize=(12, 6))
    sns.barplot(x=top_outlier_products.values, y=top_outlier_products.index, hue=top_outlier_products.index, palette="viridis", legend=False)

    # Add annotations and formatting
    plt.title('Top 10 Outlier Products by Total Revenue', fontsize=16, pad=20)
    plt.xlabel('Total Revenue ($)', fontsize=12)
    plt.ylabel('Product Name', fontsize=12)
```

```
Text(0, 0.5, 'Product Name')
```



8-

Findings Regarding Outliers

-Outliers is not Focused in a Certain Time or Region

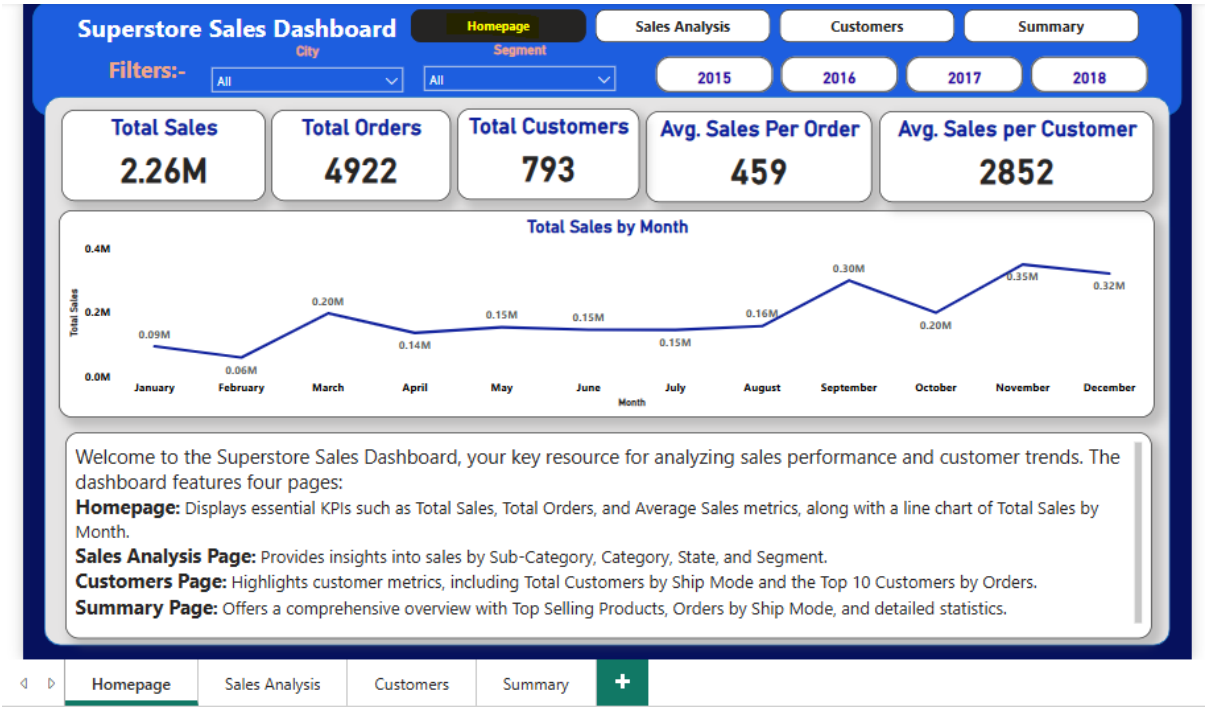
-Furniture Category has largest number of Outliers While Office Supplies has the least

-Consumer Segment has largest number of Outliers While Home Office has the least

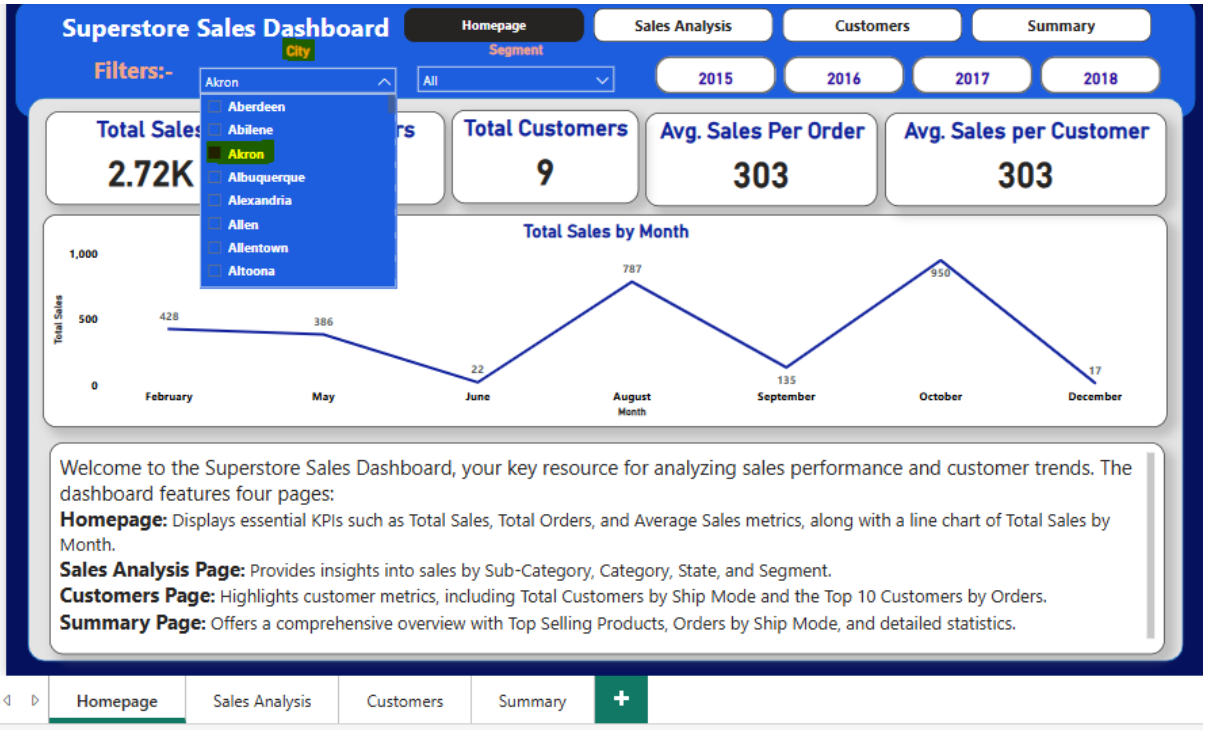
```
[ ] #Export the Cleaned Data
    df.to_csv("Superstore Sales Dataset- Cleaned.csv", index=False)
```

Dashboard Screens:

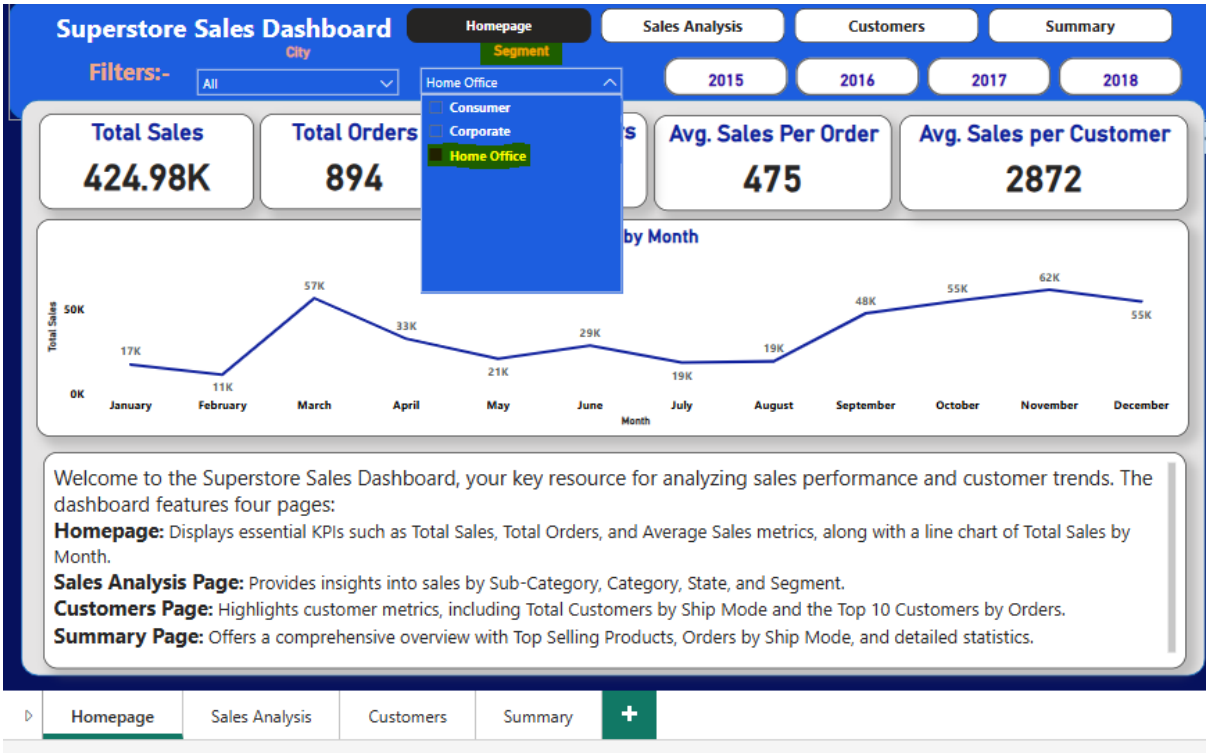
1-



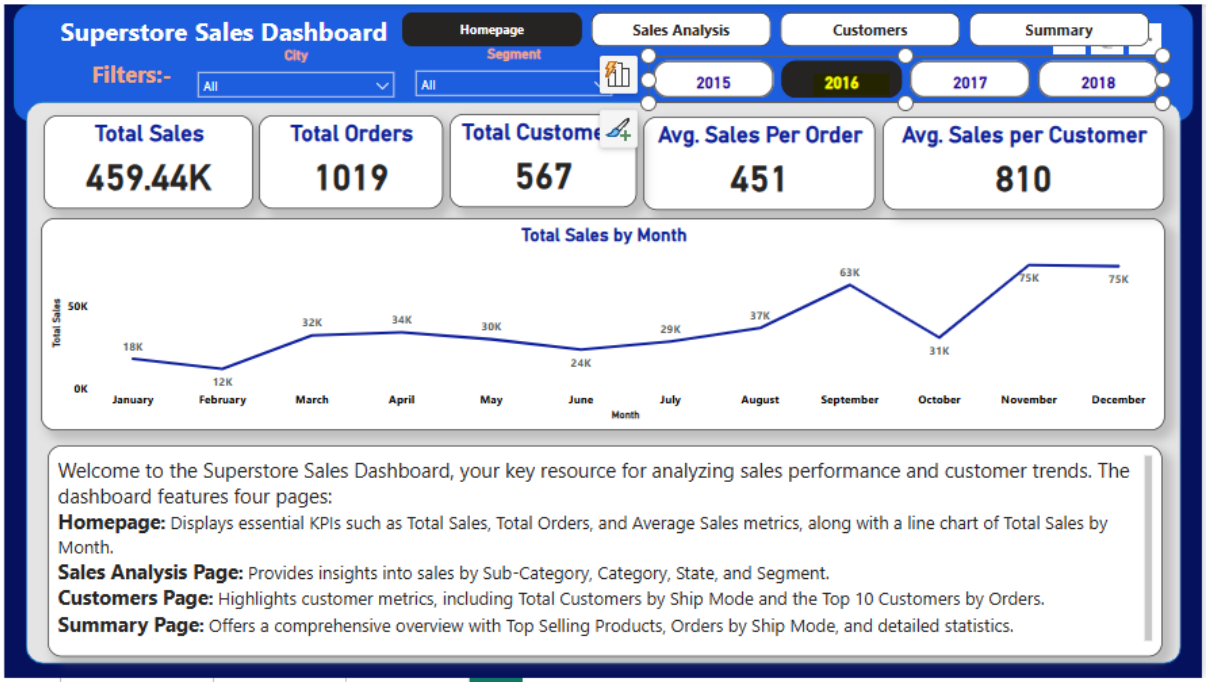
2-



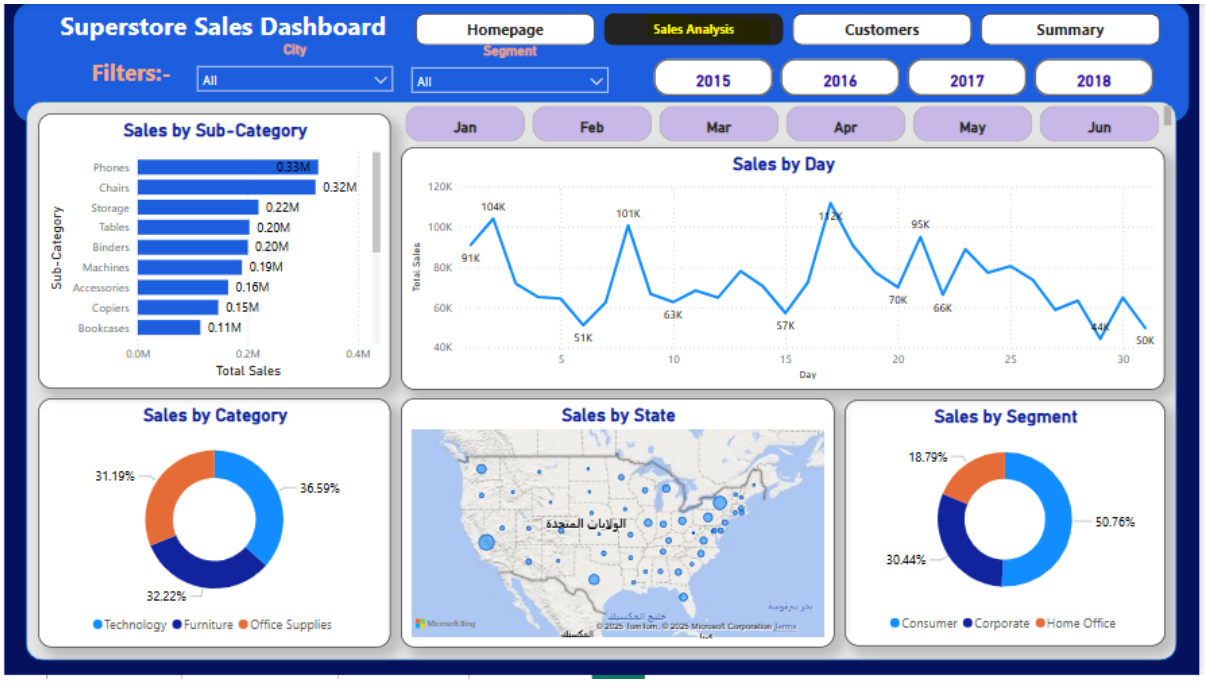
3-


Superstore Sales Dashboard
Homepage
Sales Analysis
Customers
Summary
City
Segment
Filters:-
All
Home Office
Consumer
Corporate
Home Office
2015
2016
2017
2018
Total Sales
424.98K
Total Orders
894
Avg. Sales Per Order
475
Avg. Sales per Customer
2872
by Month
Total Sales
50K
0K
17K
11K
57K
33K
21K
29K
19K
19K
48K
55K
62K
55K
January
February
March
April
May
June
July
August
September
October
November
December
Month
Welcome to the Superstore Sales Dashboard, your key resource for analyzing sales performance and customer trends. The dashboard features four pages:
Homepage: Displays essential KPIs such as Total Sales, Total Orders, and Average Sales metrics, along with a line chart of Total Sales by Month.
Sales Analysis Page: Provides insights into sales by Sub-Category, Category, State, and Segment.
Customers Page: Highlights customer metrics, including Total Customers by Ship Mode and the Top 10 Customers by Orders.
Summary Page: Offers a comprehensive overview with Top Selling Products, Orders by Ship Mode, and detailed statistics.
Homepage
Sales Analysis
Customers
Summary

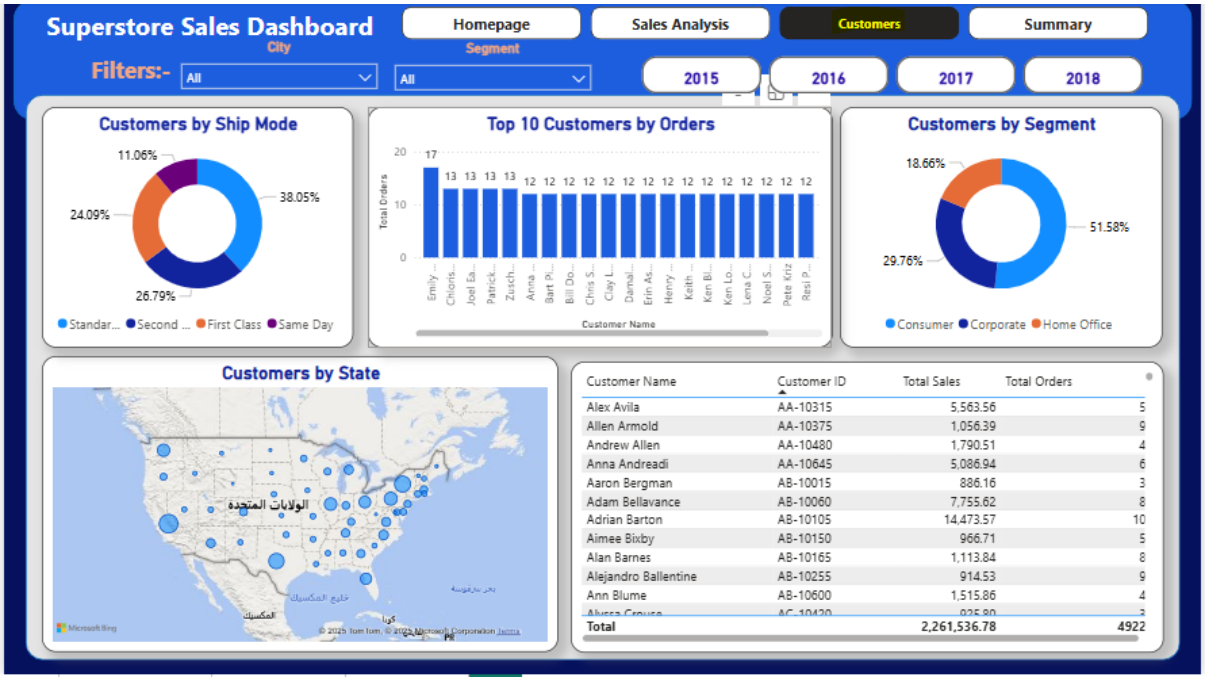
4-


Superstore Sales Dashboard
Homepage
Sales Analysis
Customers
Summary
City
Segment
Filters:-
All
All
2015
2016
2017
2018
Total Sales
459.44K
Total Orders
1019
Total Custome
567
Avg. Sales Per Order
451
Avg. Sales per Customer
810
Total Sales by Month
Total Sales
50K
0K
18K
12K
32K
34K
30K
24K
29K
37K
63K
31K
75K
75K
January
February
March
April
May
June
July
August
September
October
November
December
Month
Welcome to the Superstore Sales Dashboard, your key resource for analyzing sales performance and customer trends. The dashboard features four pages:
Homepage: Displays essential KPIs such as Total Sales, Total Orders, and Average Sales metrics, along with a line chart of Total Sales by Month.
Sales Analysis Page: Provides insights into sales by Sub-Category, Category, State, and Segment.
Customers Page: Highlights customer metrics, including Total Customers by Ship Mode and the Top 10 Customers by Orders.
Summary Page: Offers a comprehensive overview with Top Selling Products, Orders by Ship Mode, and detailed statistics.

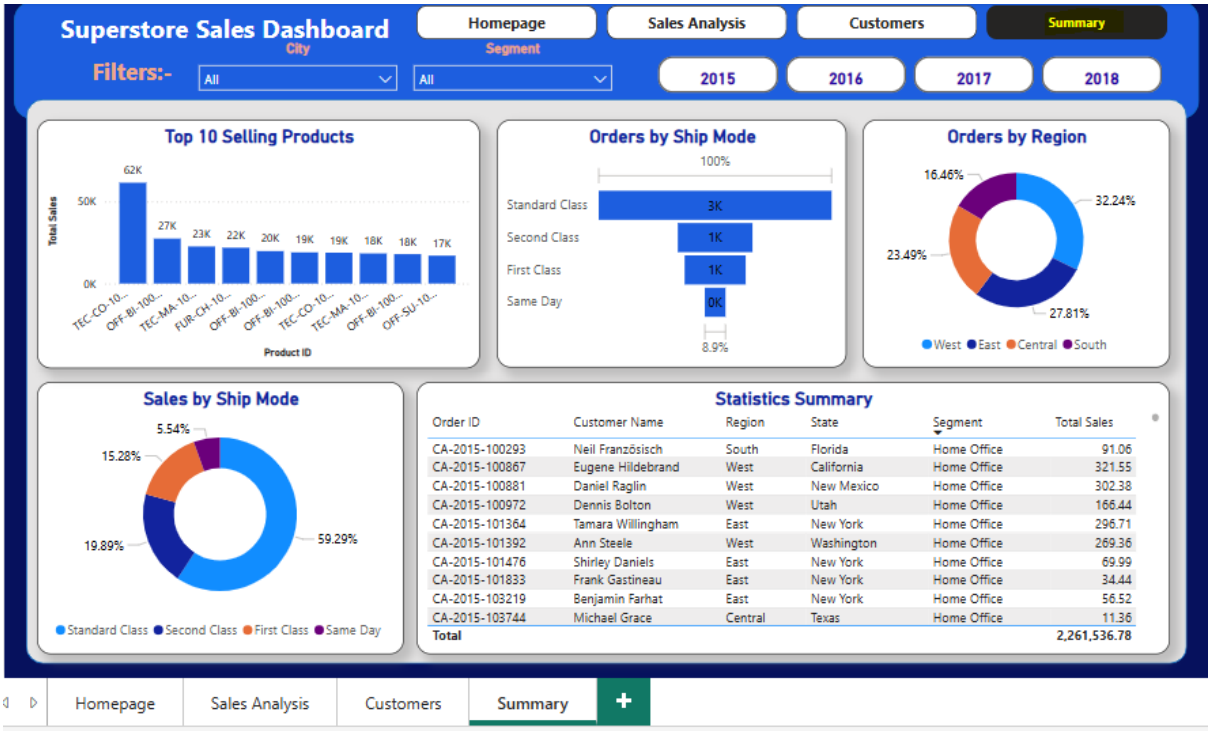
5-


Superstore Sales Dashboard
City
Filters:-
All
Segment
All
Homepage
Sales Analysis
Customers
Summary
2015
2016
2017
2018
Jan
Feb
Mar
Apr
May
Jun
Sales by Sub-Category
Phones 0.33M
Chairs 0.32M
Storage 0.22M
Tables 0.20M
Binders 0.20M
Machines 0.19M
Accessories 0.16M
Copiers 0.15M
Bookcases 0.11M
Sub-Category
Total Sales
Sales by Day
Total Sales
Day
Sales by Category
31.19%
36.59%
32.22%
Technology
Furniture
Office Supplies
Sales by State
Sales by Segment
18.79%
50.76%
30.44%
Consumer
Corporate
Home Office

6-


Superstore Sales Dashboard
City
Filters:-
All
Segment
All
Homepage
Sales Analysis
Customers
Summary
2015
2016
2017
2018
Customers by Ship Mode
11.06%
38.05%
24.09%
26.79%
Standar...
Second ...
First Class
Same Day
Top 10 Customers by Orders
Total Orders
Customer Name
Customers by Segment
18.66%
51.58%
29.76%
Consumer
Corporate
Home Office
Customers by State
Customer Name
Customer ID
Total Sales
Total Orders
Alex Avila AA-10315 5,563.56 5
Allen Armold AA-10375 1,056.39 9
Andrew Allen AA-10480 1,790.51 4
Anna Andreadi AA-10645 5,086.94 6
Aaron Bergman AB-10015 886.16 3
Adam Bellavance AB-10060 7,755.62 8
Adrian Barton AB-10105 14,473.57 10
Aimee Bixby AB-10150 966.71 5
Alan Barnes AB-10165 1,113.84 8
Alejandro Ballentine AB-10255 914.53 9
Ann Blume AB-10600 1,515.86 4
Total 2,261,536.78 4922

7-


Superstore Sales Dashboard
City
Segment
Homepage
Sales Analysis
Customers
Summary
Filters:-
All
All
2015
2016
2017
2018
Top 10 Selling Products
Total Sales
62K
50K
0K
27K
23K
22K
20K
19K
19K
18K
18K
17K
TEC-CO-10...
OFF-BI-100...
TEC-MA-10...
FUR-CH-10...
OFF-BI-100...
OFF-BI-100...
TEC-CO-10...
TEC-MA-10...
OFF-BI-100...
OFF-SU-10...
Product ID
Orders by Ship Mode
100%
Standard Class
3K
Second Class
1K
First Class
1K
Same Day
0K
8.9%
Orders by Region
16.46%
32.24%
23.49%
27.81%
West
East
Central
South
Sales by Ship Mode
5.54%
15.28%
19.89%
59.29%
Standard Class
Second Class
First Class
Same Day
Statistics Summary
Order ID
Customer Name
Region
State
Segment
Total Sales
CA-2015-100293
Neil Französisch
South
Florida
Home Office
91.06
CA-2015-100867
Eugene Hildebrand
West
California
Home Office
321.55
CA-2015-100881
Daniel Raglin
West
New Mexico
Home Office
302.38
CA-2015-100972
Dennis Bolton
West
Utah
Home Office
166.44
CA-2015-101364
Tamara Willingham
East
New York
Home Office
296.71
CA-2015-101392
Ann Steele
West
Washington
Home Office
269.36
CA-2015-101476
Shirley Daniels
East
New York
Home Office
69.99
CA-2015-101833
Frank Gastineau
East
New York
Home Office
34.44
CA-2015-103219
Benjamin Farhat
East
New York
Home Office
56.52
CA-2015-103744
Michael Grace
Central
Texas
Home Office
11.36
Total
2,261,536.78
Homepage
Sales Analysis
Customers
Summary