Digital Egypt Pioneers

Data Analysis Superstore Project

Content

- Introduction
- Data Cleaning and Normalization
- Questions and KPIs metrics
- Create charts and visualizations
- Creating Dashboard

Digital Egypt Pioneers

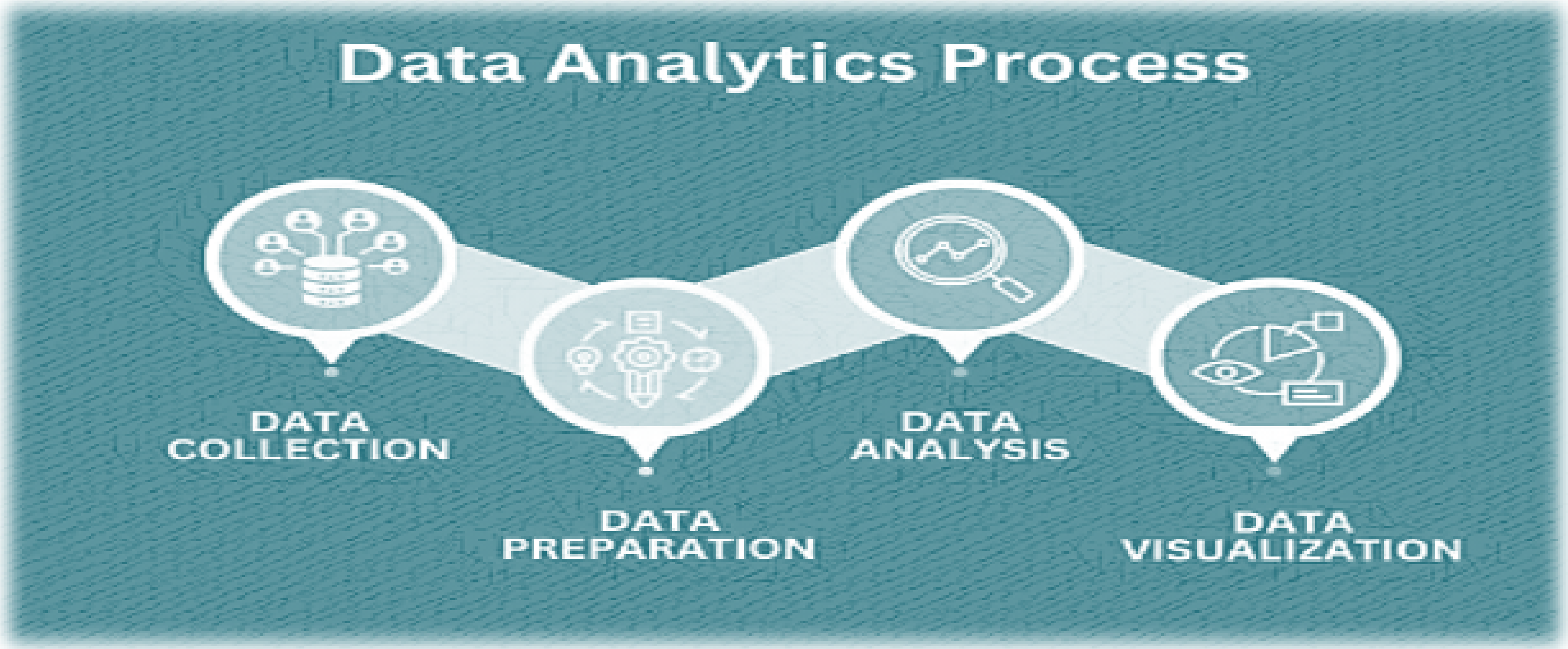
Team members: Almoataz M. Gad | Abddelrahman Ali | Ahmed Adel | Ahmed M. Elsafy

Digital Egypt Pioneers

➤ Introduction

- In today’s competitive retail environment, data-driven decision-making is essential for optimizing sales performance and enhancing customer satisfaction. This project focuses on analyzing the sales data of a superstore to uncover key insights and trends that can drive strategic improvements. By leveraging tools such as Excel, SQL Server, and Tableau, we have examined various factors like product categories, customer demographics and sales performance. The goal is to identify patterns and opportunities that can inform decisions on inventory management, pricing strategies, and customer targeting to increase profitability and market reach.

➤ Introduction



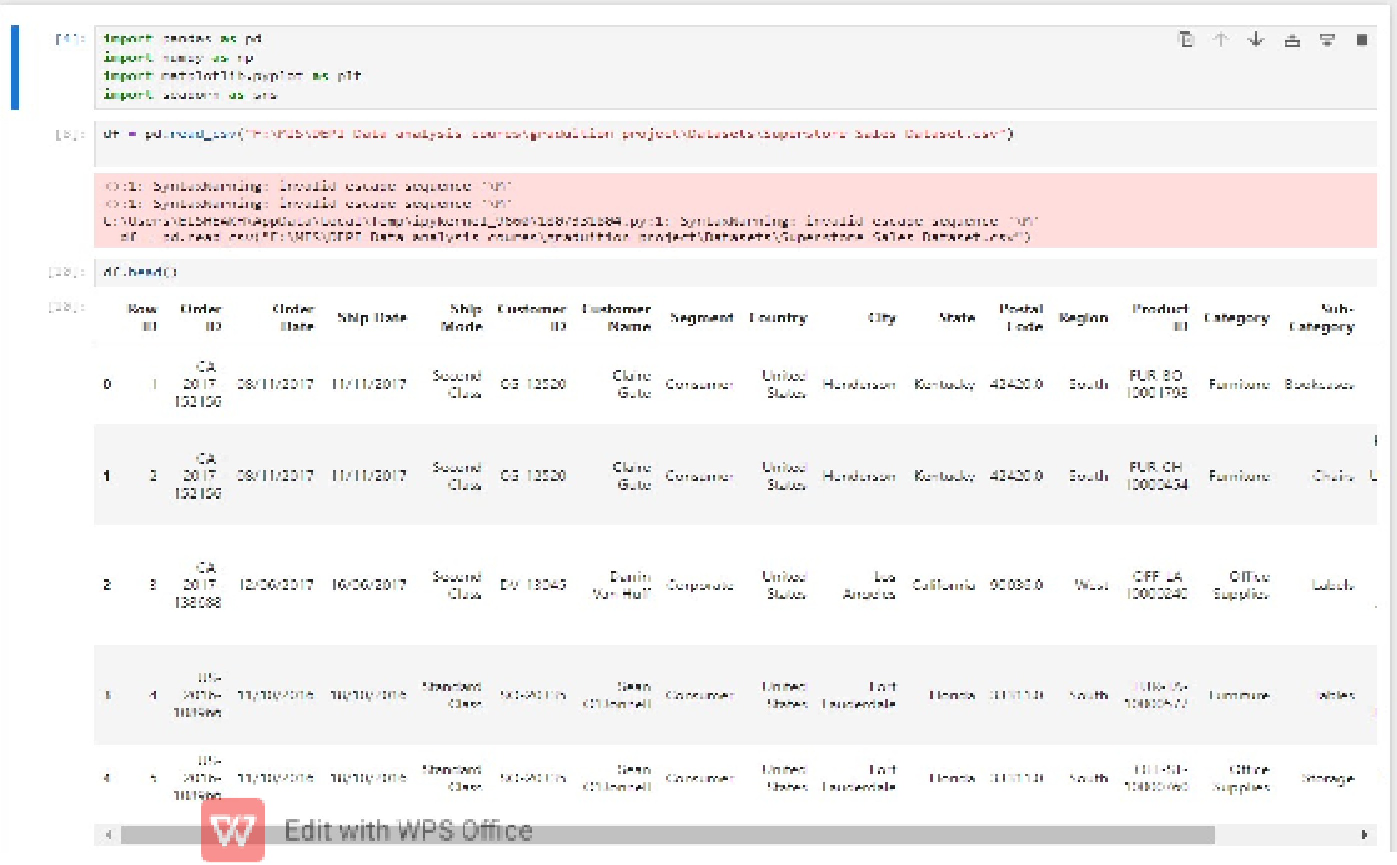
➤ Data Cleaning and Normalization

- Data Cleaning is the first step in the project. And in my team’s opinion data cleaning process it the most important step.
- Cleaned data is important to build high quality insights.
- We used python programming language to achieve cleaned data.
- We used python libraries like : Pandas ,Numpy and matplotlip.
- Next screen show data cleaning process by python.

Digital Egypt Pioneers

➤ Data Cleaning and Normalization

Data cleaning step can be done by many tools, our team used python programming language. By using libraries like pandas and numpy as shown.

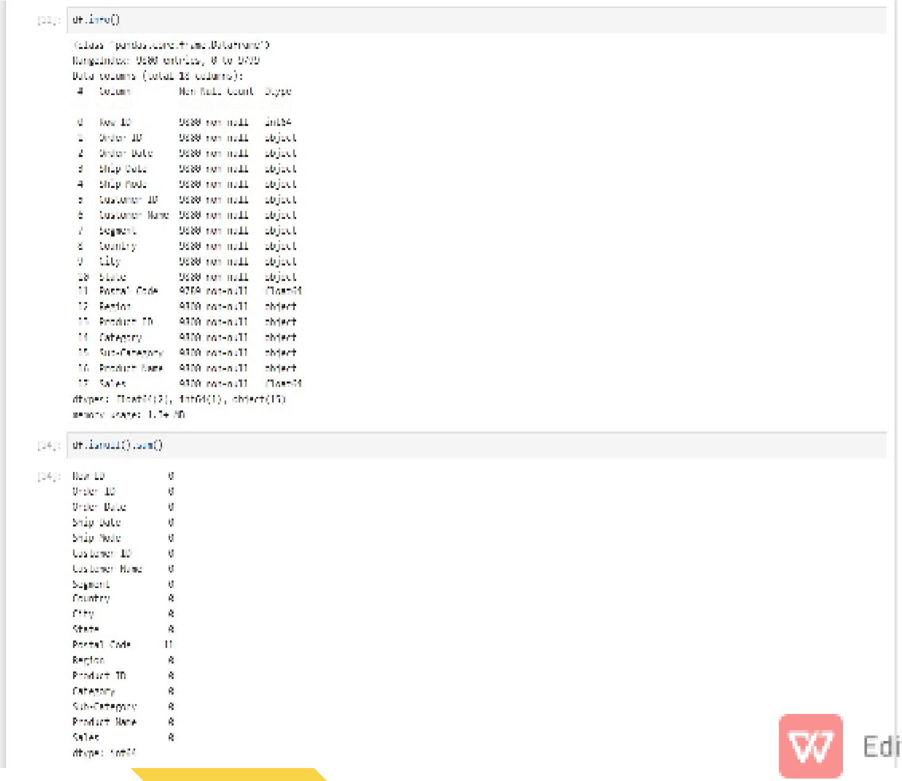


```
[4]: import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import seaborn as sns

[5]: df = pd.read_csv("F:\NTS\DEPI Data analysis course\graduation project\Datasets\Superstore Sales Dataset.csv")

<>:1: SyntaxWarning: invalid escape sequence '\N'
<>:1: SyntaxWarning: invalid escape sequence '\N'
C:\Users\ELSHEIKH\AppData\Local\Temp\ipykernel_[illegible]\[illegible].py:1: SyntaxWarning: invalid escape sequence '\N'
  df = pd.read_csv("F:\NTS\DEPI Data analysis course\graduation project\Datasets\Superstore Sales Dataset.csv")

[38]: df.head()
```

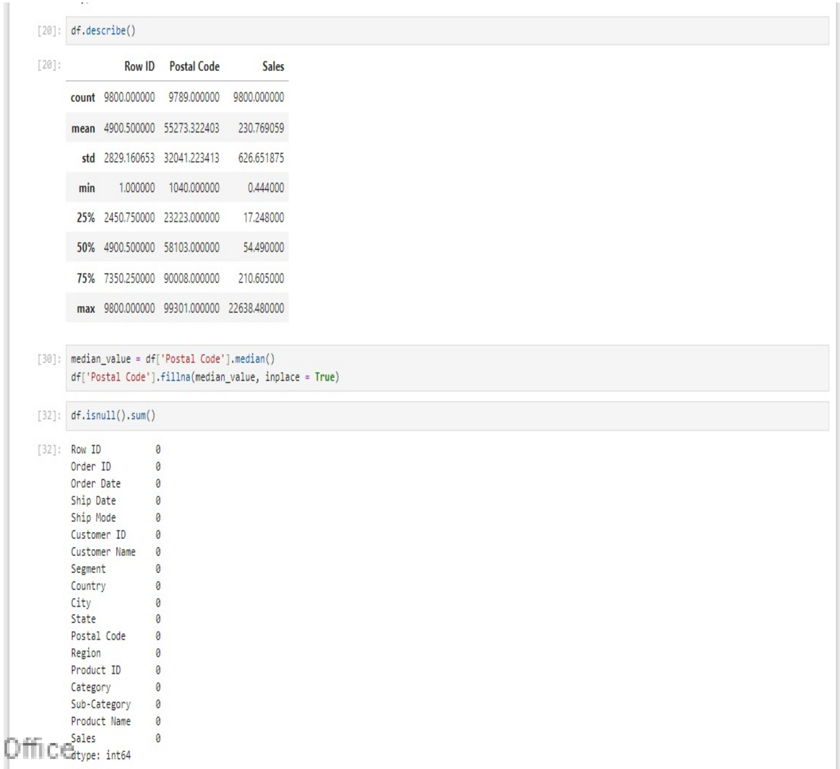
[38]:

	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category
0	1	CA-2017-152156	08/11/2017	11/11/2017	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420.0	South	FUR-BO-10001798	Furniture	Bookcases
1	2	CA-2017-152156	08/11/2017	11/11/2017	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420.0	South	FUR-CH-10000454	Furniture	Chairs
2	3	CA-2017-138688	12/06/2017	16/06/2017	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	California	90036.0	West	OFF-LA-10000240	Office Supplies	Labels
3	4	US-2015-108966	11/10/2015	18/10/2015	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311.0	South	FUR-TA-10000577	Furniture	Tables
4	5	US-2015-108966	11/10/2015	18/10/2015	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311.0	South	OFF-ST-10000760	Office Supplies	Storage

➤ Data Cleaning and Normalization



```
[23]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9800 entries, 0 to 9799
Data columns (total 18 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Row ID         9800 non-null   int64  
 1   Order ID       9800 non-null   object 
 2   Order Date     9800 non-null   object 
 3   Ship Date      9800 non-null   object 
 4   Ship Mode      9800 non-null   object 
 5   Customer ID    9800 non-null   object 
 6   Customer Name  9800 non-null   object 
 7   Segment        9800 non-null   object 
 8   Country        9800 non-null   object 
 9   City           9800 non-null   object 
 10  State          9800 non-null   object 
 11  Postal Code    9789 non-null   float64
 12  Region         9800 non-null   object 
 13  Product ID     9800 non-null   object 
 14  Category       9800 non-null   object 
 15  Sub-Category   9800 non-null   object 
 16  Product Name   9800 non-null   object 
 17  Sales          9800 non-null   float64
dtypes: float64(2), int64(1), object(15)
memory usage: 1.3+ MB

[24]: df.isnull().sum()

[24]: Row ID            0
Order ID          0
Order Date        0
Ship Date         0
Ship Mode         0
Customer ID       0
Customer Name     0
Segment           0
Country           0
City              0
State             0
Postal Code      11
Region            0
Product ID        0
Category          0
Sub-Category      0
Product Name      0
Sales             0
dtype: int64
```



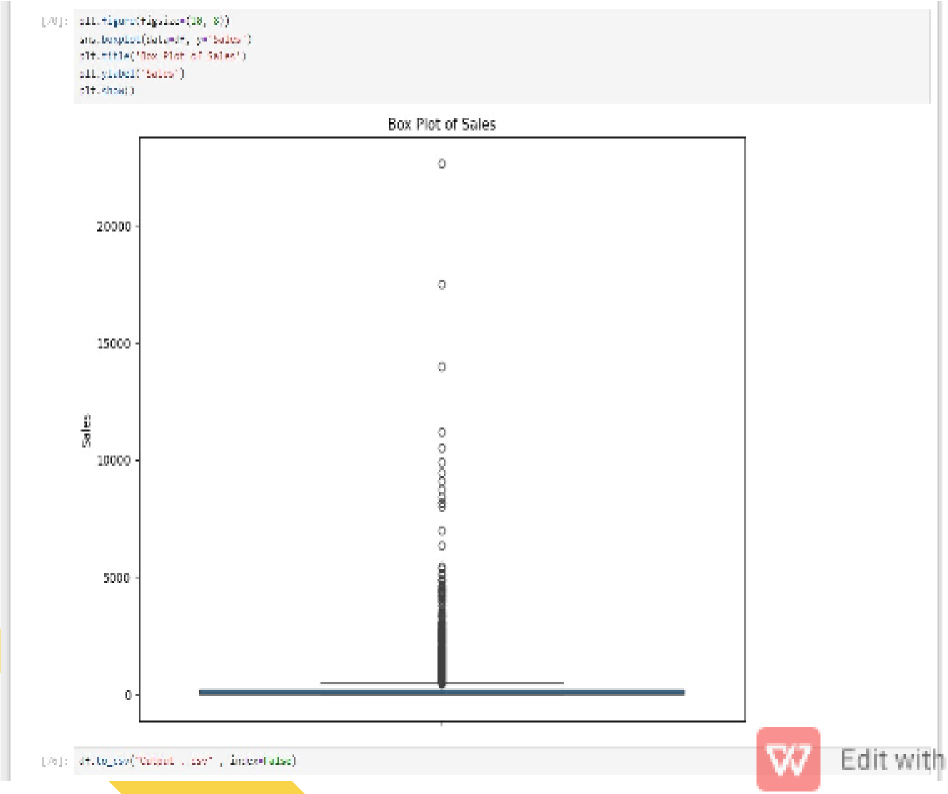
```
[20]: df.describe()
```

[20]:

	Row ID	Postal Code	Sales
count	9800.000000	9789.000000	9800.000000
mean	4900.500000	55273.322403	230.769059
std	2829.160653	32041.223413	626.651875
min	1.000000	1040.000000	0.444000
25%	2450.750000	23223.000000	17.248000
50%	4900.500000	58103.000000	54.490000
75%	7350.250000	90008.000000	210.605000
max	9800.000000	99301.000000	22638.480000

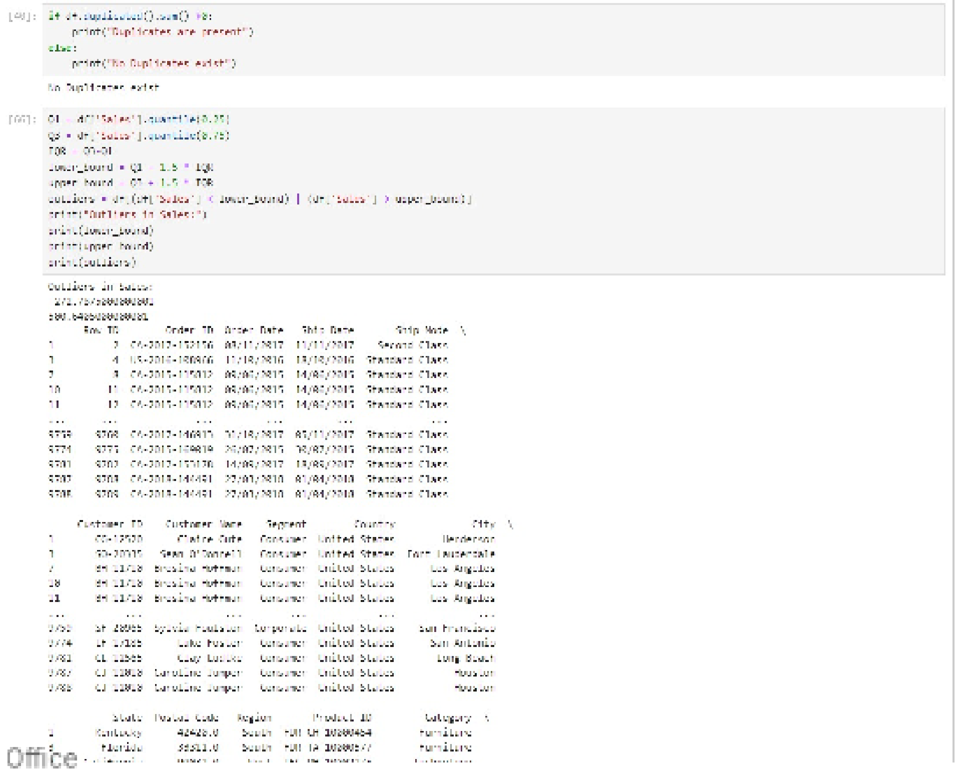
```
[30]: median_value = df['Postal Code'].median()
      df['Postal Code'].fillna(median_value, inplace = True)

[32]: df.isnull().sum()

[32]: Row ID           0
Order ID         0
Order Date       0
Ship Date        0
Ship Mode        0
Customer ID      0
Customer Name    0
Segment          0
Country          0
City             0
State            0
Postal Code      0
Region           0
Product ID       0
Category         0
Sub-Category     0
Product Name     0
Sales            0
dtype: int64
```

➤ Data Cleaning and Normalization



```
plt.figure(figsize=(10, 8))
sns.boxplot(data=df, y='Sales')
plt.title('Box Plot of Sales')
plt.ylabel('Sales')
plt.show()
```

```
df.to_csv("Output . csv" , index=False)
```



```
if df.duplicated().sum() >0:
    print("Duplicates are present")
else:
    print("No Duplicates exist")
```

```
No Duplicates exist
```

```
Q1 = df['Sales'].quantile(0.25)
Q3 = df['Sales'].quantile(0.75)
IQR = Q3-Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
outliers = df[(df['Sales'] < lower_bound) | (df['Sales'] > upper_bound)]
print("Outliers in Sales:")
print(lower_bound)
print(upper_bound)
print(outliers)
```

```
Outliers in Sales:
 -271.7575300000001
500.6425500000001
      Row ID        Order ID  Order Date   Ship Date       Ship Mode  \
1          2  CA-2017-152156  08/11/2017  11/11/2017    Second Class
3          4  US-2016-108966  11/10/2016  18/10/2016  Standard Class
7          8  CA-2015-115812  09/06/2015  14/06/2015  Standard Class
10        11  CA-2015-115812  09/06/2015  14/06/2015  Standard Class
11        12  CA-2015-115812  09/06/2015  14/06/2015  Standard Class
...      ...             ...         ...         ...             ...
9759    9760  CA-2017-140913  31/10/2017  03/11/2017  Standard Class
9774    9775  CA-2015-169019  26/07/2015  30/07/2015  Standard Class
9781    9782  CA-2017-173126  14/05/2017  18/05/2017  Standard Class
9787    9788  CA-2018-144491  22/01/2018  01/04/2018  Standard Class
9788    9789  CA-2018-144491  22/01/2018  01/04/2018  Standard Class

     Customer ID    Customer Name    Segment        Country             City  \
1       CC-12370       Claire Gute   Consumer  United States        Henderson
3       SO-20335   Sean O'Donnell   Consumer  United States  Fort Lauderdale
7       BH-11710  Brosina Hoffman   Consumer  United States      Los Angeles
10      BH-11710  Brosina Hoffman   Consumer  United States      Los Angeles
11      BH-11710  Brosina Hoffman   Consumer  United States      Los Angeles
...          ...              ...        ...            ...              ...
9759    SF-20065  Sylvia Foulston  Corporate  United States    San Francisco
9774    LF-17185      Luke Foster   Consumer  United States      San Antonio
9781    CL-12565      Clay Ludtke   Consumer  United States       Long Beach
9787    CJ-11875  Caroline Jumper   Consumer  United States          Houston
9788    CJ-11875  Caroline Jumper   Consumer  United States          Houston

            State  Postal Code   Region       Product ID         Category  \
1        Kentucky      42420.0    South  FUR-CH-10000454        Furniture
3         Florida      33311.0    South  FUR-TA-10000577        Furniture
```

Questions and KPIs metrics

Digital Egypt Pioneers

Questions and KPIs metrics

1. Customer Insights

- How many unique customers are there?
- What is the distribution of customers by segment (Customer, Corporate, Home Office)?
- Which segment generates the highest sales?
- How many orders does each customer place on average?

Charts:

- Bar chart: Number of customers by segment.
- Pie chart: Sales contribution by customer segment.
- Bar chart: Number of orders per customer.

2. Order Insights

- How many orders were placed, and what is the average order amount?
- What is the average time between the order date and the ship date?
- Which customers place the highest number of orders?

Charts:

- Line chart: Number of orders over time.
- Bar chart: Average ship time by customer.
- Line chart: Number of orders by month.

3. Product Insights

- Which products are the most and least sold by quantity?
- What are the best-selling product categories and sub-categories?
- How does product sales vary across different regions?

Charts:

- Bar chart: Sales by product category.
- Pie chart: Sales by sub-category.
- Heatmap: Product sales by region.

4. Sales and Revenue Insights

- What is the total sales amount, and how does it distribute across customer segments and regions?
- Which regions contribute the most to sales?
- What is the trend of sales over time?

Charts:

- Line chart: Total sales over time.
- Stacked bar chart: Sales by region.
- Pie chart: Sales distribution by region.

5. Geographical Insights

- How are sales distributed across different regions and cities?
- Which postal codes are the most active in terms of sales and orders?
- What is the regional contribution to overall revenue?

Charts:

- Map chart: Sales by postal code.
- Bar chart: Orders by region.
- Pie chart: Sales by city/state.

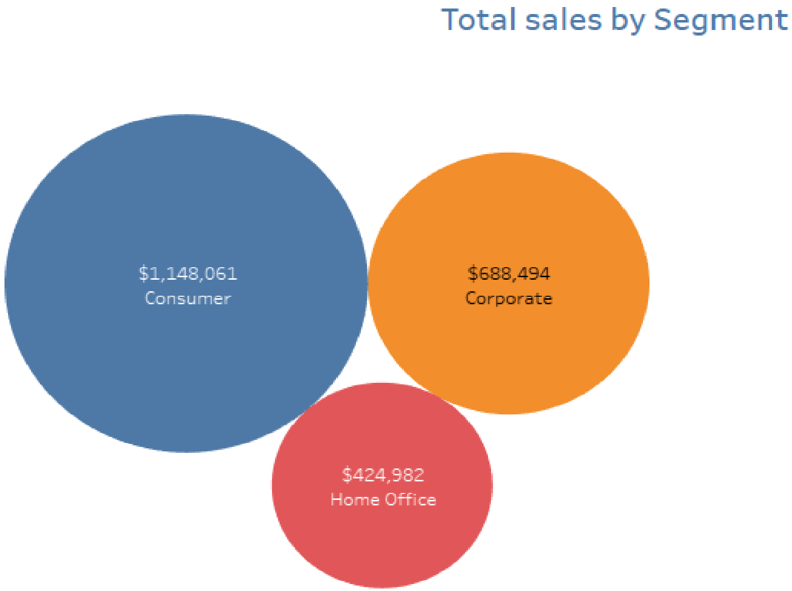
Summary of Potential Questions:

1. How many customers, orders, and products are there?
2. What is the sales distribution across customer segments?
3. What are the top-selling products?
4. Which customers contribute the most to revenue?
5. What are the geographical patterns of sales?

Digital Egypt Pioneers

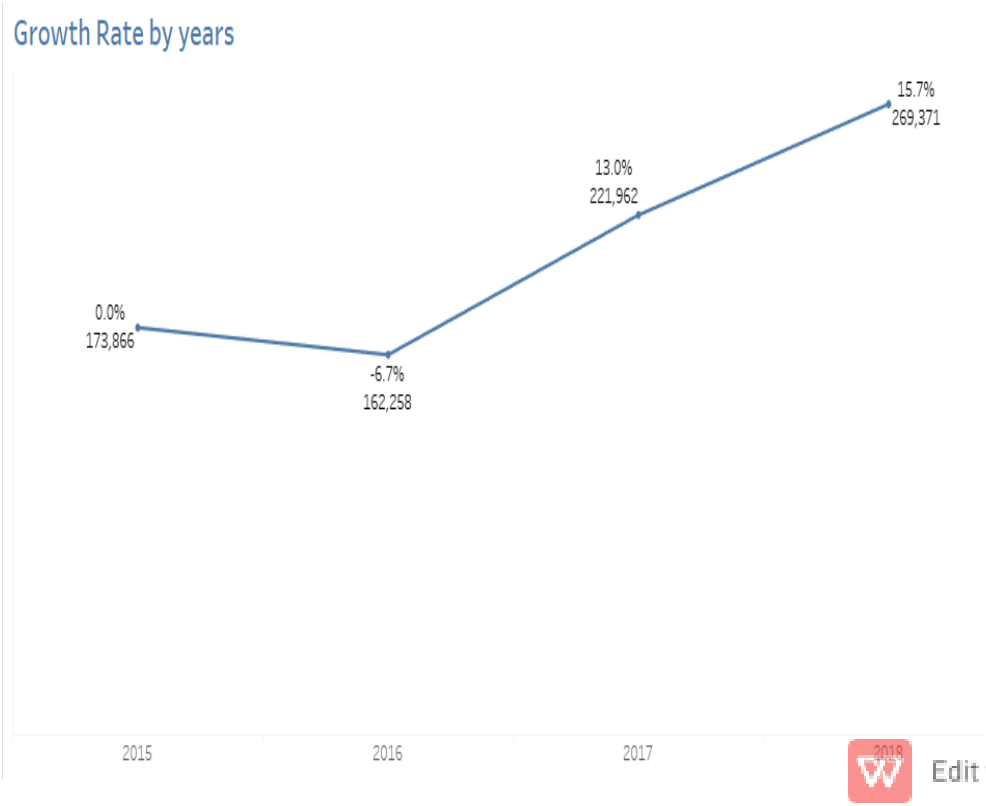
Create charts and visualizations

- In this Process our Team used **tableau desktop.**
- Next, a few examples of charts
- This chart shows amount of sales by Segment and it conclude that Consumer



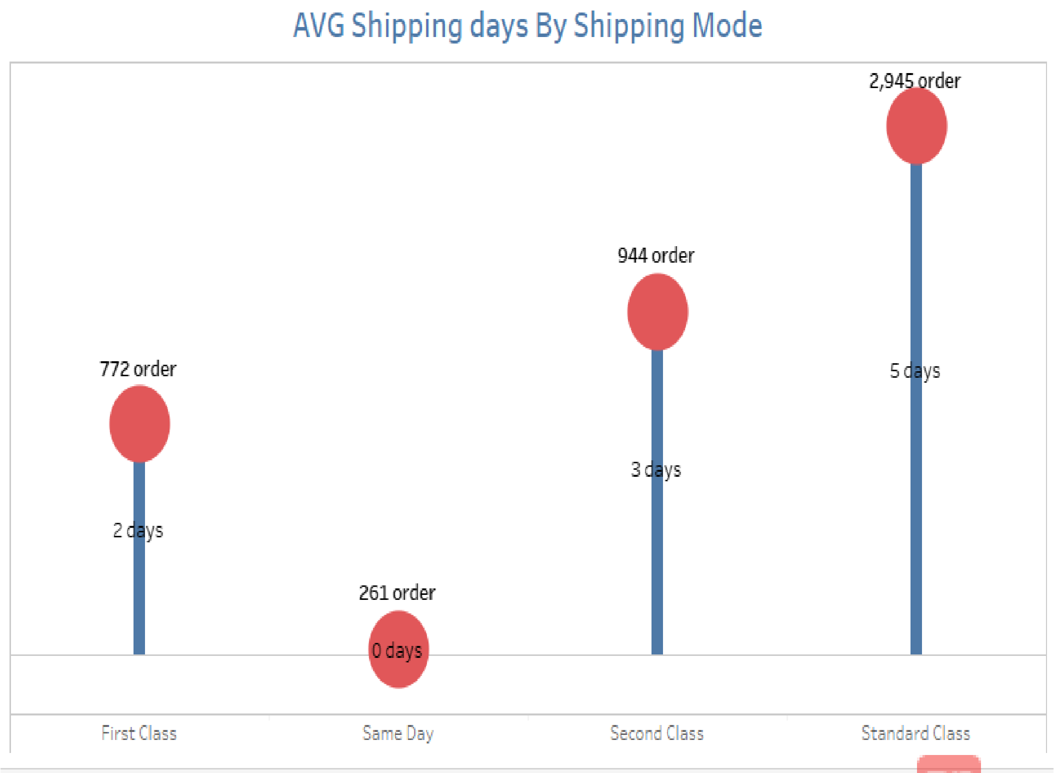
charts

- The growth rate is one of the most important KPI that give us an idea about the performance of the store by years.



charts

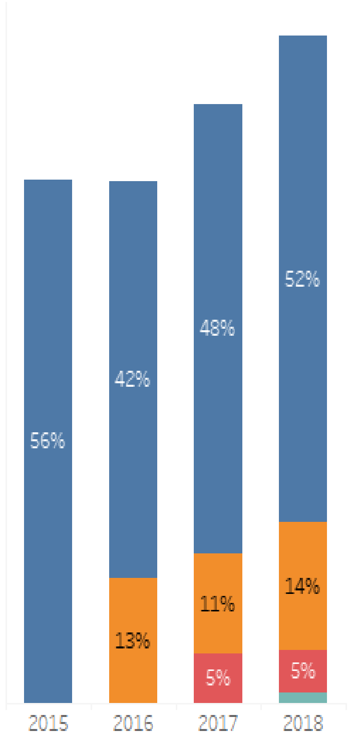
- Lollipop chart is amazing figure to show the relation between number of orders and shipping mode and average days top ship.



charts

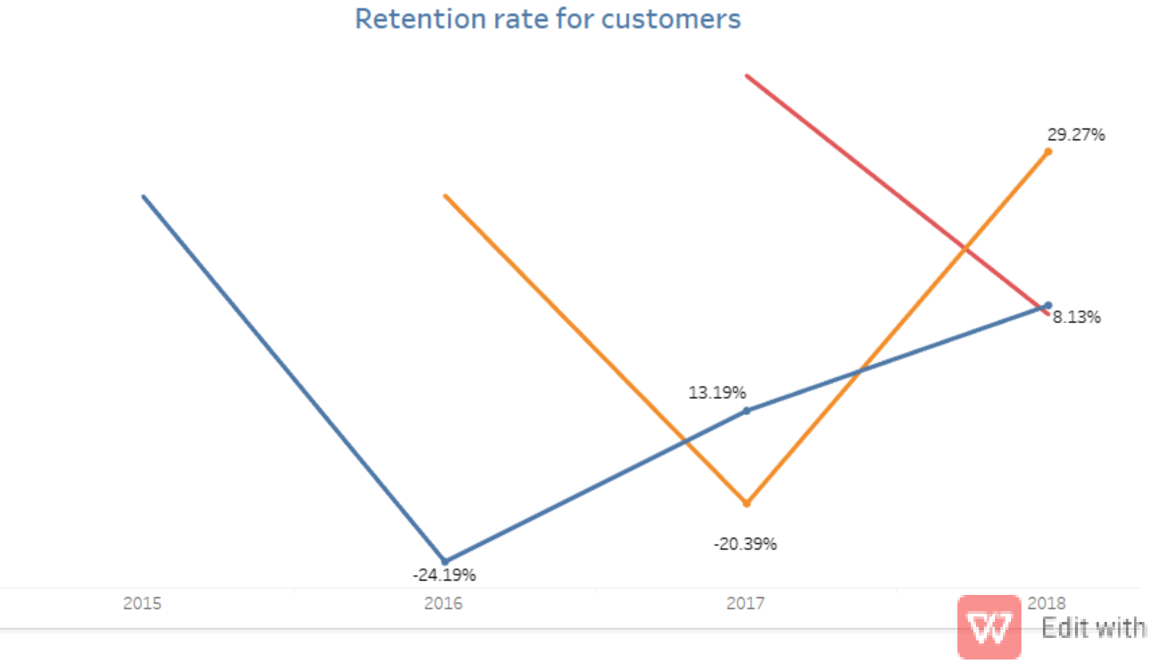
- This chart explains that old customers represented by blue color have the large scale in sales over years and represent the **loyalty** of the old customers

Performance for old customers on sales



charts

- Retention Rate also considering important metric to measure how satisfied the customers
- If retention rate is high it means the same customers come again and make orders

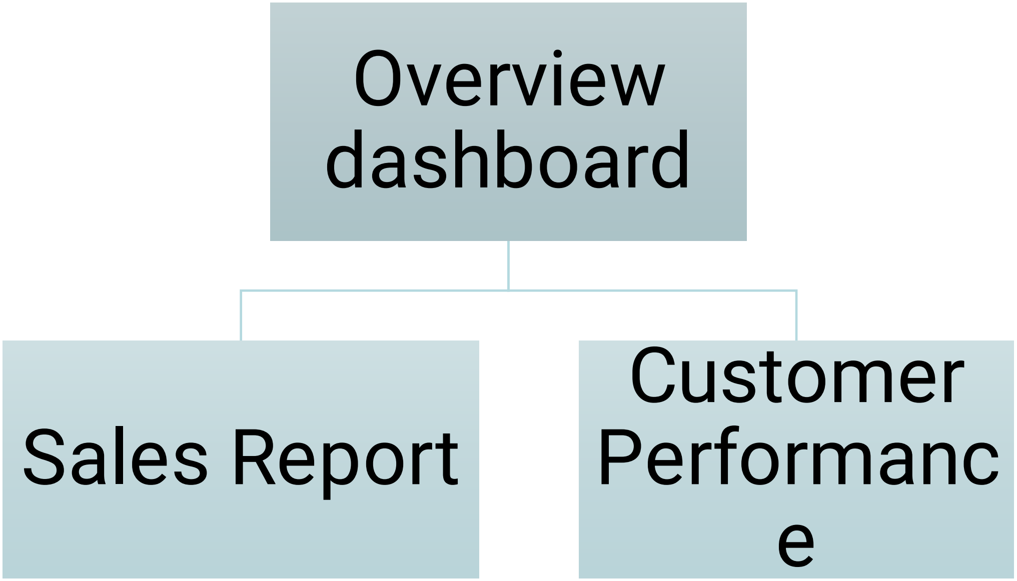


charts

Digital Egypt Pioneers

Creating Dashboard

- The dashboard phase was divided into three reports.



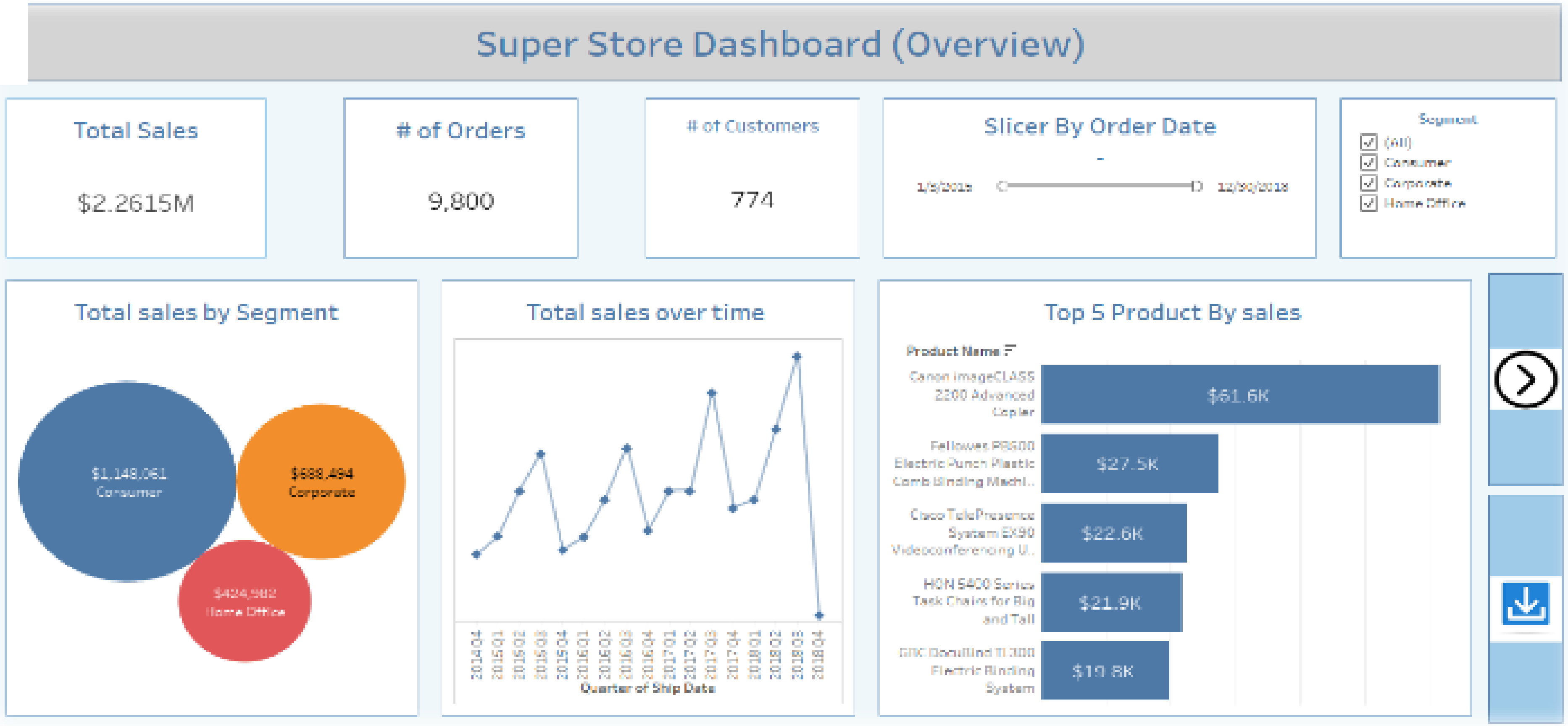
Overview dashboard

Sales Report

Customer Performance

Creating Dashboard

Digital Egypt Pioneers



Super Store Dashboard (Overview)

Total Sales

$2.2615M

# of Orders

9,800

# of Customers

774

Slicer By Order Date

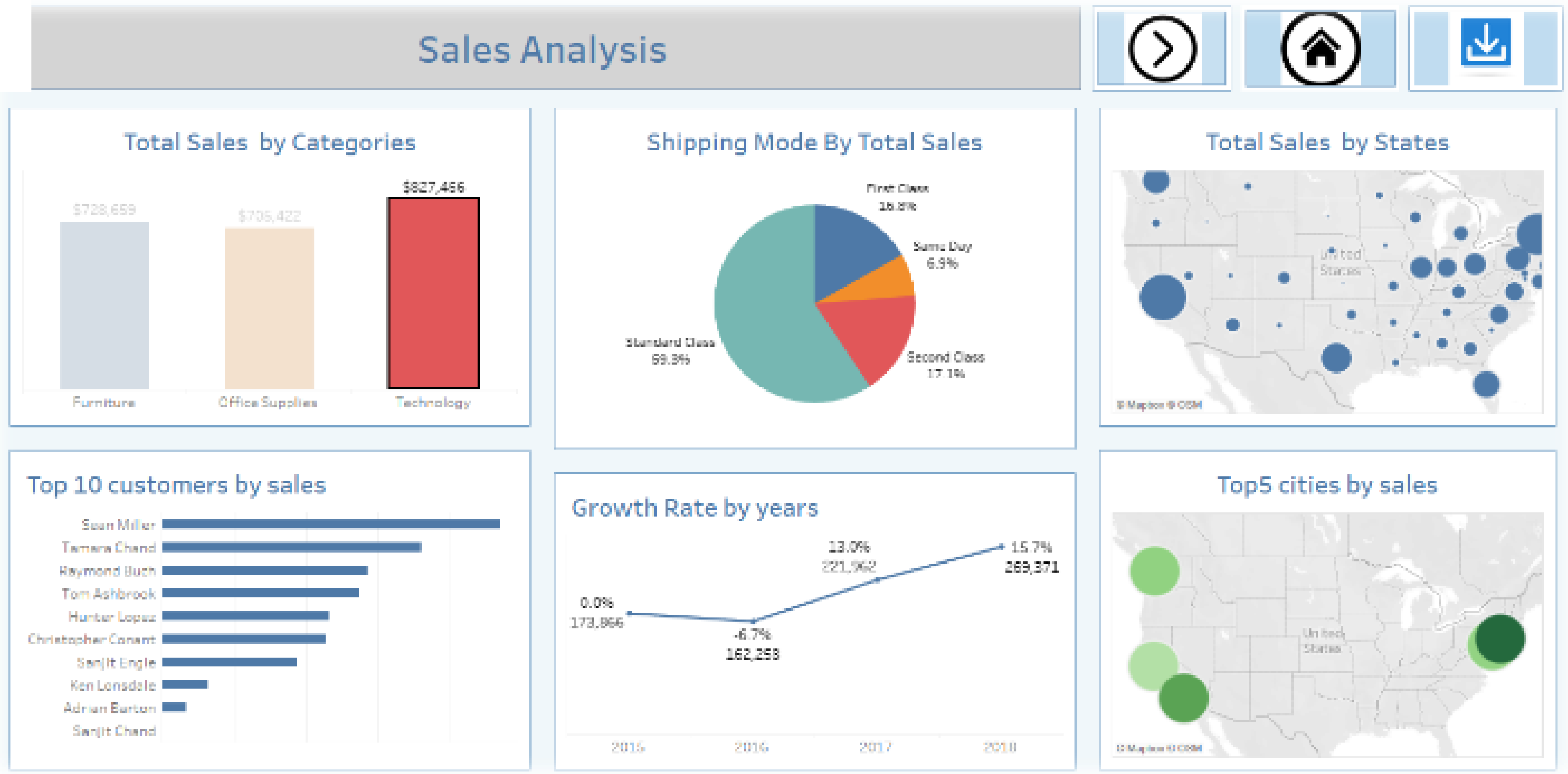
1/3/2015 – 12/30/2018

Segment

☑ (All)
☑ Consumer
☑ Corporate
☑ Home Office

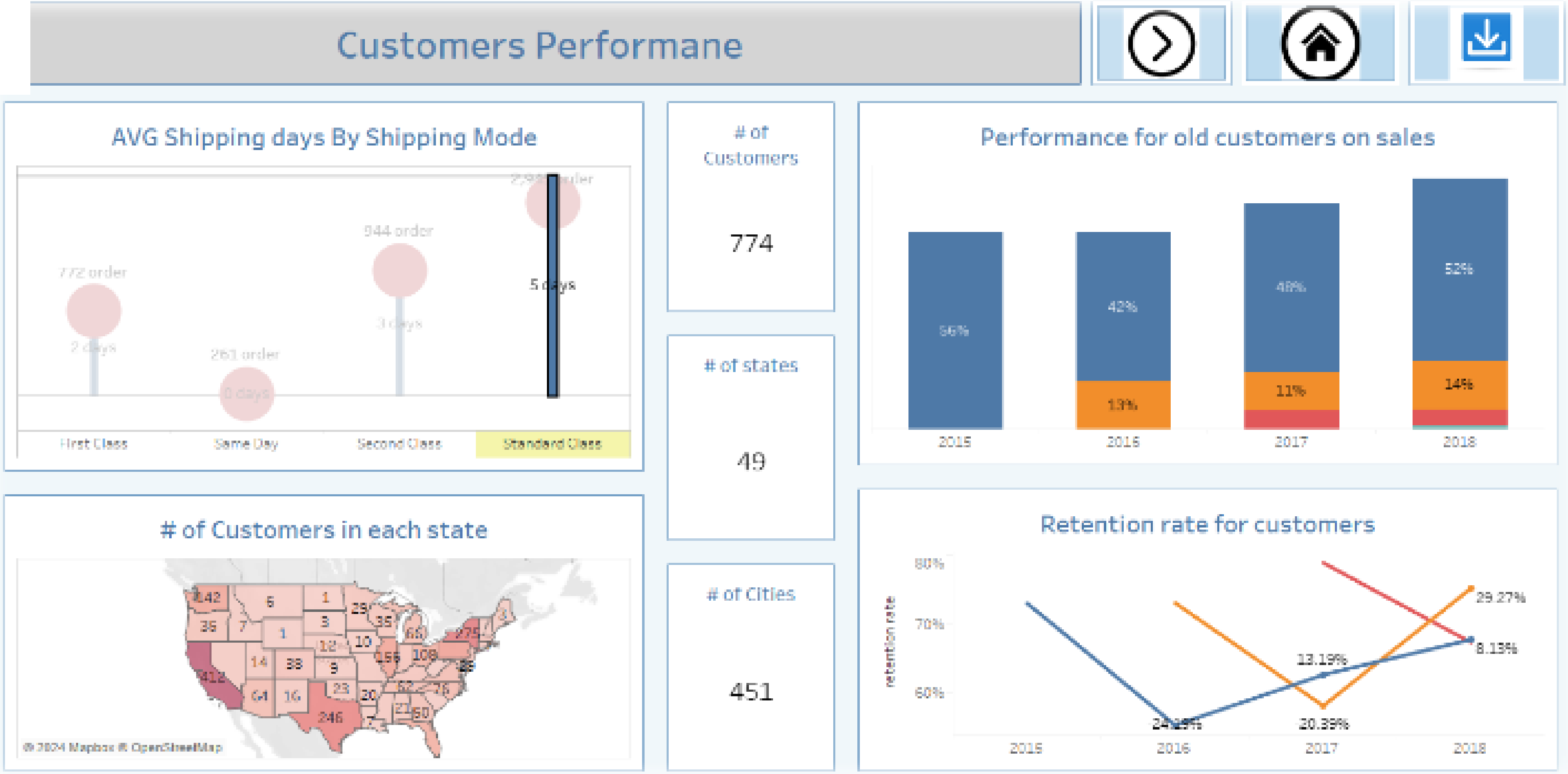
Creating Dashboard

Digital Egypt Pioneers



Creating Dashboard

Digital Egypt Pioneers



Creating Dashboard

Digital Egypt Pioneers

Thank you