**Analysis and Visualization of SuperStore Sales**

## Introduction

SuperStore deals in products across the consumer, corporate and home office segments with categories across technology, furniture and office supplies. Analysis was carried out on the dataset emanating from Superstore across the years 2014 – 2017. The goal was to explore sales and profit over the years across segments and categories.

## Data Structure

The original dataset is a 21-column dataset containing records of sales from 2014 – 2017. Columns include Order ID, Order Date, Ship Date, Ship Mode, Customer ID, Customer Name, Segment, Country, City, State, Postal Code, Region, Product ID, Category, Product Name, Sales, Quantity, Discount and Profit.

## Analysis Process/Methodology

I went through some procedures to make sense of the data and ensure readiness of the data for detailed analysis and actionable information.

1. Data Gathering
2. Data Assessment and Cleaning
3. Data Visualization
4. Communication and Insights

## Data Gathering

Raw data was sourced from Kaggle and downloaded to my local system.

## Data Assessment and Cleaning

To ensure data is well prepared for visualization, I had to properly assess and clean. The goal of this was to ensure that I identified any issues and clean and required. To do this, I leveraged pandas (python library for data analysis and visualization). Steps I took in this process are outlined below:

- Using Anaconda and Jupyter notebook, I imported the pandas library
- I read the downloaded csv dataset and sampled the head and tail of the data. From the sample, most columns seemed to be in order per data quality standards. However, I noticed varying number of digits for postal code.

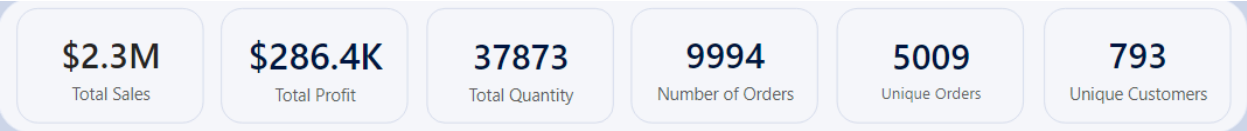
- To proceed, I checked the shape of the data to confirm the number of rows and columns present in the dataset. There were 9,994 rows and 21 columns present.
- Using the .info() method, I checked for null values and datatypes per column. There were no null values in any column however, postal code was stored as int64 which explained why it had varying number of digits.
- I proceeded to get a further description of the dataset, checked for duplicates and for unique values. There were no duplicates in the dataset.
- To begin cleaning, a copy of the initial dataset was made to ensure that the raw data remains available.
- I dropped the Row ID column as it was not necessary for my analysis
- I converted the Postal Code datatype to string from int64 and standardized it to five digits.
- I then converted the dataset to Excel format and saved on my local system for analysis with Power BI.

**Data Visualization**

To carry out data visualization, I imported the excel file into power BI and built a report to communicate findings.

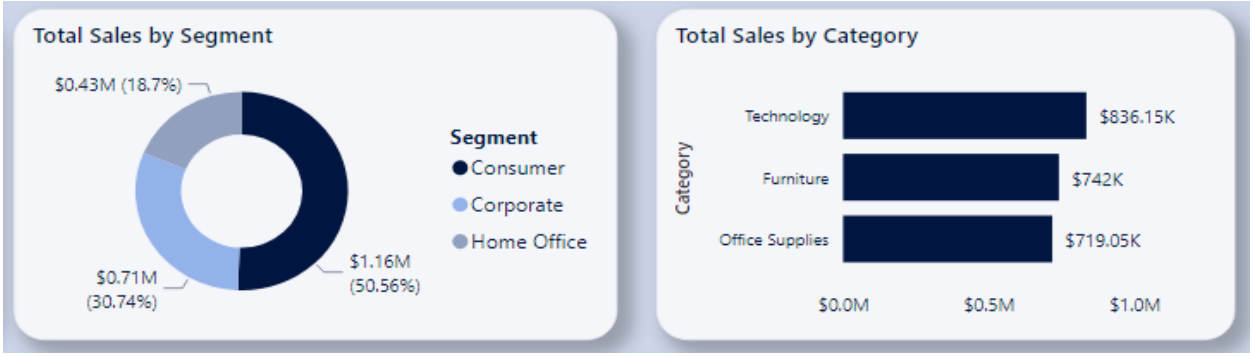
**Communication and Insights**

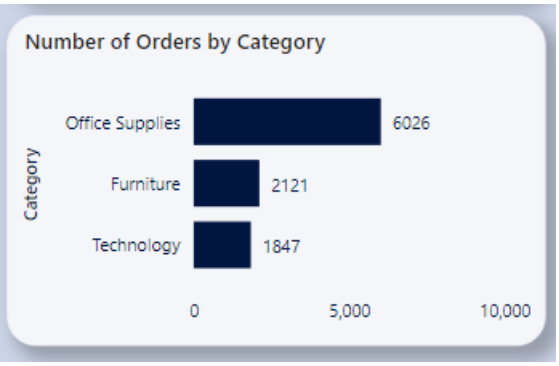
Total sales generated between 2014 – 2017 was $2,297,200.86 while a profit of $286,397.02 was made by the store. A total quantity of 37,873 goods was sold in 9,994 orders by 793 unique customers.




Asides the decline in 2015, value of sales has grown progressively over the years with 2016 having the highest year on year increase in sales.

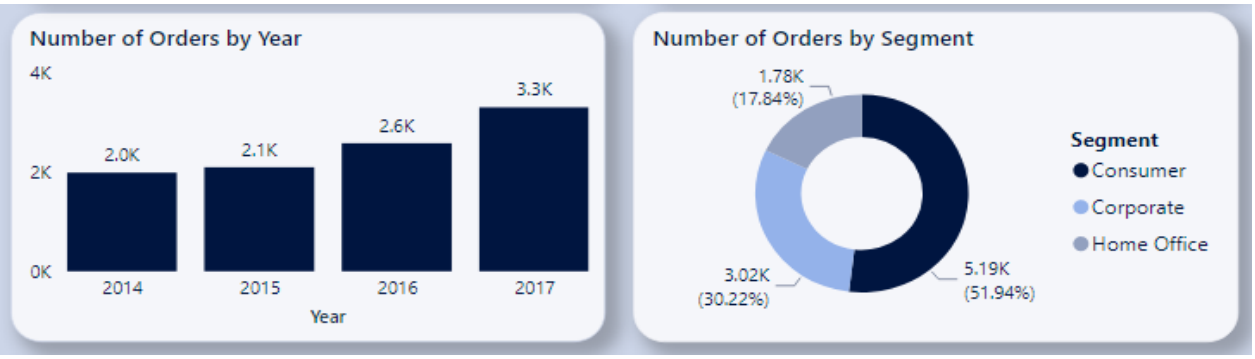

Consumer is the segment with the highest value of sales while technology is the category with is the highest value of sales.

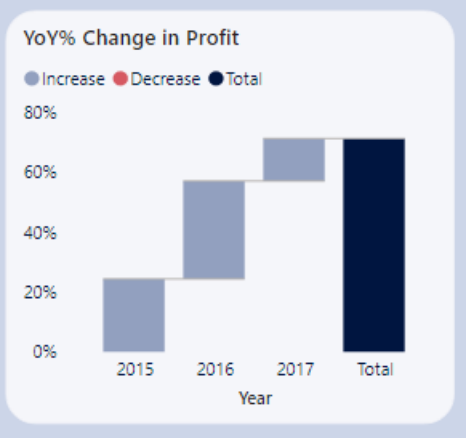





What's interesting however is that by number of orders, technology is the category with the least number of orders.

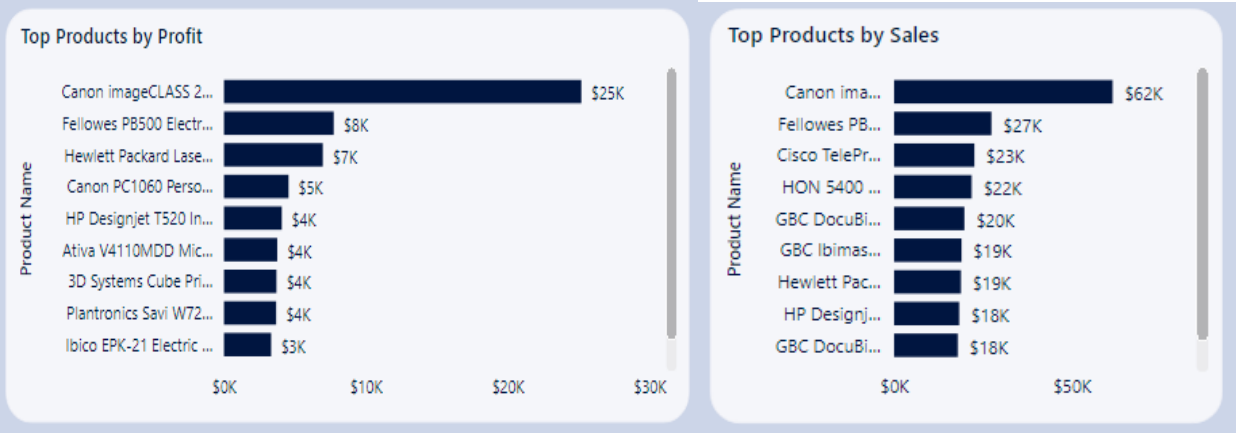
Consumer is the segment with the highest number orders. Year on year, number of orders has grown progressively.



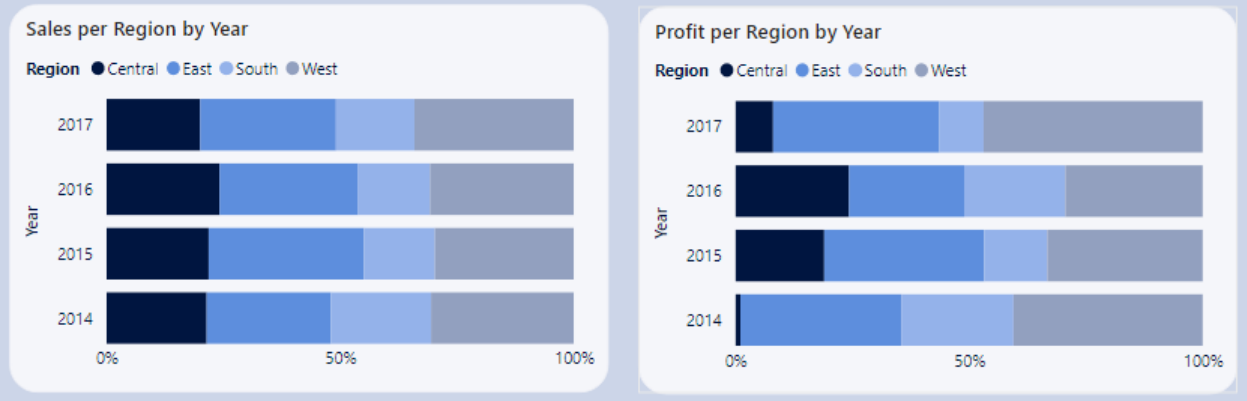




Profit has increased progressively over the years with 2016 being the year with the highest increase in profit year on year. Consumer and technology remain the most profitable segment and category respectively.

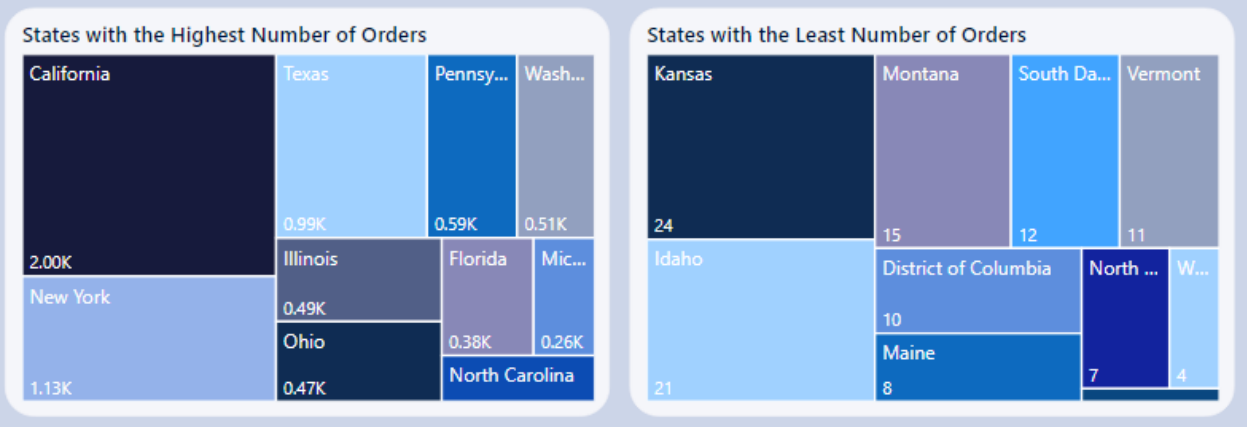
Canon image class 2200 advanced copier is the top grossing product by sales and profit.



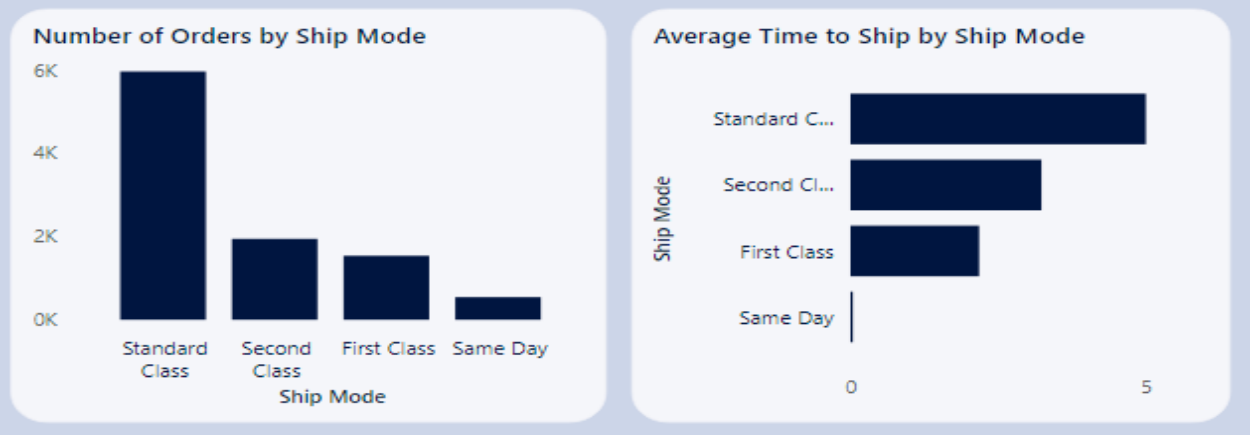

Asides 2015, West is the region with the highest value of sales and profit. In 2015, east was the region with the highest value of sales and profit.




California is the state with the highest number of orders (2001 orders) while Wyoming is the state with the least number of orders (1 order).



Standard class is the most popular shipping mode with an average shipping time of five days. Second class and first class have an average shipping time of 3 and 2 days respectively.




**Recommendation**

- Carry out further analysis to determine what drives high sales value of the technology category despite having the least number of orders. Is it the pricing, marketing etc. and can this be implemented across all categories?
- Further analysis to find out the reason for the decline in sales and profit in the west region in 2015. Was this decline also responsible for the decline in value of sales in 2015?