Visualization of Superstore Data

ANJU TYAGI, ASTHA JAIN, MADHU BANDRU AND VUTHEJ KRISHNA REDDY

DREXEL UNIVERSITY

Abstract– Superstores are becoming more and more popular in The United States of America and studying the underlying trends in the sales of a superstore helps in smooth and successful functioning of the entity amongst competitors. This report utilizes Tableau as a tool for information visualization to analyze the trends in sales data of a superstore from 2015 to 2018. Data is then visualized to understand the impact of discounts on the sales and profits.

Index Terms – Superstore, sales visualization, information visualization, profits, losses, discounts.

1 INTRODUCTION

A Superstore can be defined as an extremely large retail store that caters wide range of merchandise offering groceries, apparel, household items, games and toys, furniture, electronic gadgets, medical supplies and many more at one single place to the customers. Now to effectively maintain such a store without running into any supply and demand, marketing and profitability issues, one has to understand the requirements of the customers and at the same time think innovatively in order to withstand the competitors.

Data Analytics leverages current and past data to forecast future trends and provides meaningful insights about the underlying trends and user behaviors. It also helps greatly in understanding which parts of the country is contributing towards higher sales and which are having the least sales.

Current dataset that we are using for visualizing the superstore data is from the Tableau community which is offered on Kaggle platform [1]. This dataset provides relevant information regarding the orders and sales of superstore ranging in the timeframe of 2015 through 2018. We wanted to understand the variations in yearly sales from 2015 to 2018 of a superstore and identify the locations with maximum and minimum number of sales. This helps in taking appropriate decisions by the superstore management to boost their sales in those identified locations. In this paper we will be focusing on investigating below mentioned points.

1. Regions with highest profits and sales from 2015 to 2018.
2. Categories and segments contributing towards yearly losses.
3. States-wise profits and losses in relation to the discounts offered.

Analyzing such information will prove to be crucial in successfully maintaining a superstore even in the presence of fierce competitors.

1.1 LITERATURE REVIEW

Prior studies on similar datasets focused on visualizing various aspects of sales data of the superstore which included the analysis of variations in number of orders from each individual state over

the years, highest and least orders across the regions of United States (Central, East, South, West), data distributions of the quantity of products ordered, percentage of sales by category, correlation between sales, quantity, discount and profit, visualizing statistical summary of sales, in-depth visualization of a particular state over the years to understand the trends to name a few [2]. It is **noteworthy** that these visualizations were easy to understand and did not include higher level of complexities in their plots. **However**, some of the visualizations were lacking effectiveness in conveying complete information at once i.e. multiple visualizations were required in order to convey the trends in profits across all the categories and sub-categories which spanned across 4 years. In this paper, we tried to generate visualizations which conveys all information at once rather than using multiple plots. For instance, **Figure 3- Profit for each category/segment over the years** conveys profits in each category and segment across 4 years in a single visualization. We further attempted to understand the role played by discounts in bringing profits to the superstore.

2 METHODOLOGY

In article [2] author focused majorly on

- Distribution of orders across all states
- How the frequency of quantity ordered
- Percentage of sales
- Author tried to plot co-relation among features sales, quantity, discount and profit

Here, the author concentrated more on the distribution of data and sales percentage, but clear identification of areas or categories where a business is profitable, or loss is not specified. Unless business owners identify real strengths and weaknesses, critical decisions are not made.

In our approach, we would like to cover few critical visualizations which are more beneficial to the superstore. The inferences we are willing to draw help superstores manage the products, sales and performing effective business. The problems we choose are more focused on knowing the truth about sales at the ground level. We assume that for understanding sales, profits and losses are the best areas to concentrate on. We also want to explore these profits and losses in both regions and category/segment-wise. We also try to link these profits and losses with discounts as well.

Below are the problems we wish to present and visualize in this report through which we would like to overcome the pitfalls from previous report [2].

- Identify regions with the highest profit and sales in a particular year.
- Identify categories and segments contributed to losses.
- Locate region/state with highest & lowest profits concerning discounts.

Below is the list of processes that followed to perform our study.

- Understanding the data
- Data preparation
- Literature review on data
- Identifying questions for visualizations
- Choosing the right visualizations
- Evaluation

2.1 DATASET

The superstore data has been collected from Kaggle [1] for year 2015 to 2018. The data set consist of 9994 instances with 21 columns. There are 6

numerical, 2 date and 13 string features in the data set which is detailed below.

S.No	Feature	Data Type	Description	Sample Data
1	Row ID	Numerical	Sequence of the record in dataset	13
2	Order ID	String	It describes about the customer's order number	CA-2018-114412
3	Order Date	Date	It describes the date when customer placed order	04/15/2018
4	Ship Date	Date	It describes the data when ordered item is shipped	04/20/2018
5	Ship Mode	String	It describes about the type of shipping mode selected	Second Class
6	Customer ID	String	This is the customer unique identification number	AA-10480
7	Customer Name	String	Name of the customer	Andrew Allen
8	Segment	String	It describes about the customer type	Consumer
9	Country	String	It describes the country to which customer belongs to	United States
10	City	String	It describes the city to which customer belongs to	Concord
11	State	String	It describes the state to which customer belongs to	North Carolina
12	Postal Code	Numerical	It describes the destination or area code of customer	28027
13	Region	String	It describes under which region customer belongs to	South
14	Product ID	String	This is the unique product identification number	OFF-PA-10002365
15	Category	String	It talks about the category the product comes under	Office Supplies
16	Sub-Category	String	It talks about the sub-category the product comes under	Paper
17	Product Name	String	It is the description about the product	Xerox 1967
18	Sales	Numerical	This is the amount collected for customer order	15.552
19	Quantity	Numerical	It describes number of items ordered by customer	3
20	Discount	Numerical	Amount reduced on actual price of the product	0.2
21	Profit	Numerical	This is the profit amount gained by super store	5.4432

Table 1- Data Dictionary of US Superstore Data

2.2 DATA CLEANING

Under the data cleaning process, we looked manually for junk data and missing columns. We could see 11 records do not have postal codes, and this feature does not show much impact on our visualization. So, we have dropped the postal code before starting visualization.

2.3 TOOLS

For this study, Tableau is used to generate visualizations of the data [4]. Tableau is a simple drag and drop data visualization tool widely used for business intelligence, but it is not limited to. It also helpful in creating interactive graphs and charts in the form of dashboards and worksheets to gain business insights. Tableau provides different types of charts like text tables, heat maps, highlight tables, symbol maps, maps, pie charts, line graphs, and many more. Tableau can load data from various sources and multiple forms like text, pdf, CSV, and even from the server.

For our problem, the data set is in excel format. Below are the steps followed to load and visualize the superstore dataset.

1. Load the data using "more" option under "To a File" section on the home screen.
2. Once the dataset is selected and loaded, Tableau provides a preview of the data in the "data source" tab.
3. When a worksheet is opened, the data set is segregated into "Dimensions" and

"Measures," which holds categorical and numerical features respectively.

4. Drag and drop necessary features into rows and columns tab for generating visualizations.
5. "Show Me" on the top right corner facilitates different types of graphs that are compatible with the selected features.
6. "Marks" section provides facilities like adding colors, size, labels, and many more for the selected features. These cosmetic changes are also used as sort and filter functions for particular selected features from the dataset.
7. Once the required visualization is generated, options are available to easily export the graph.

3 VISUALIZATION & DISCUSSION

3.1 REGIONS WITH HIGHEST PROFIT AND SALES IN PARTICULAR YEAR

Result: This visualization in **Figure 1- Profit in relation to region-wise yearly sales** was generated in Tableau. We first started by plotting direct statistical sales variance on the y-axis and then spread this out over the United States regions and years (2015,2016,2017,2018) on the x-axis. To understand the distribution of profit and sales, we included a color scale with a sequential palette for profit, the colors in the graph represent the degree of the profit, darker the color more the profit.

Significance: The resulting graph showed the distribution of profit and sales for different regions over the years. As observed, the West region showed the highest sales and profit followed by the East in 2018. In 2016, the South region seems to have the lowest sales.

Inference: At a glance, we can spot the highest and the lowest sales regions and also how profit is distributed.

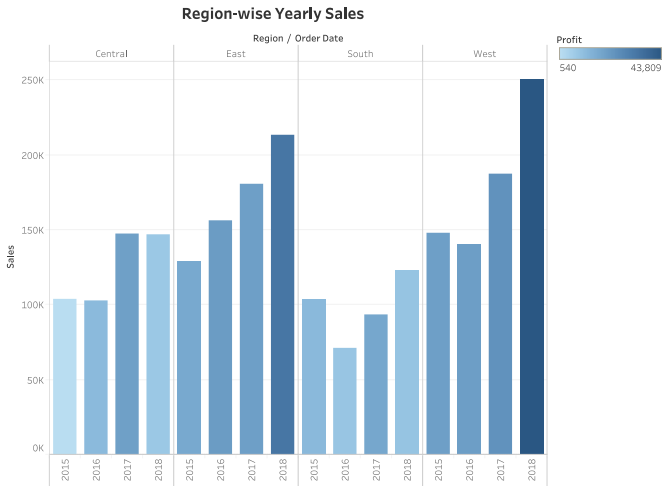

Figure 1- Profit in relation to region-wise yearly sales

3.2 REGION/STATE WITH HIGHEST & LOWEST PROFITS IN RELATION TO DISCOUNTS

Result: Our second data visualization is bar graphs using Tableau (**Figure 2.1- Region-wise profit and discount distribution** and **Figure 2.2- State-wise profit in relation to discounts**). Both **Figure 2.1- Region-wise profit and discount distribution** and **Figure 2.2- State-wise profit in relation to discounts** have been helpful to interpret how regions' profit are affected by the discounts offered. We plotted graphs **Figure 2.1- Region-wise profit and discount distribution** and **Figure 2.2- State-wise profit in relation to discounts** to find which states are contributing to loss with higher discounts. **Figure 2.1- Region-wise profit and discount distribution** Charted profit on the y-axis and then spread this out over the United States regions on the x-axis. To understand the distribution of profit and discounts,

we included a color scale with a sequential palette for discounts and labeled discounts on top of each bar, the colors in the graph represent the degree of the discounts, the darker the color higher the discounts. **Figure 2.2- State-wise profit in relation to discounts** further branched regions out to states on the x-axis.

Significance: In **Figure 2.1- Region-wise profit and discount distribution**, despite higher discounts in the central region, the profits are low comparing other regions. **Figure 2.2- State-wise profit in relation to discounts**, the Central region especially states like Texas and Illinois gave high discounts and ended in loss. Also, a similar trend was seen in the Eastern region for Pennsylvania and Ohio, but interestingly states with high discounts made losses most of the time except for New York and California.

Inference: High discounts kept the sales and profit for states like New York and California on a positive scale. Intriguingly, no state suffered losses with zero discount.

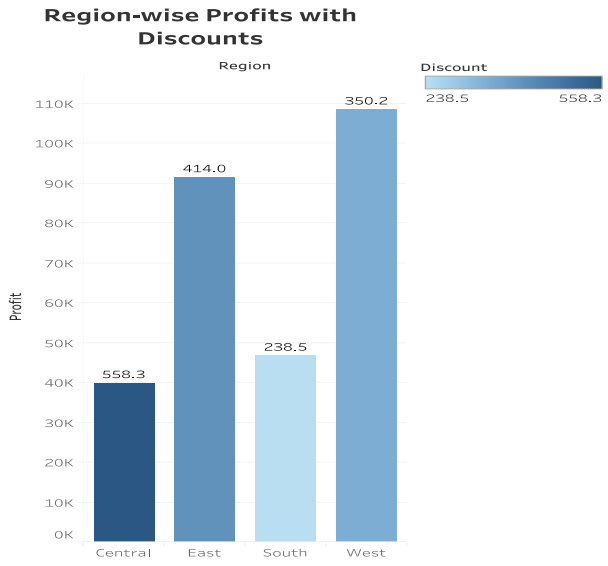

Figure 2.1- Region-wise profit and discount distribution.

Figure 2.2- State-wise profit in relation to discounts.

3.3 CATEGORY AND SEGMENT CONTRIBUTED TO LOSSES YEARLY

Result: Figure 3- Profit for each category/segment over the years shows an interactive information visualization that is created using Bar chart style. This bar chart takes direct statistical profit variance on the y-axis and then extended this out over the Categories (Furniture, Office Supplies, and Technology) and years on the x-axis. It also uses Categorical palettes containing distinct color codes to represent different segments (Consumer, Corporate, and Home Office) which help in visualizing the ratio of profit/loss made by each segment. The area of the segment in each bar represents the profit/loss made and gives a clear picture of which segment/category contributed to profits/losses in what year.

Significance: We found it striking that only the Furniture category contributed to losses in the year 2016, 2017, and 2018 for both Consumer and Home Office segments. Losses in 2018 are relatively lower than in 2016 and 2017.

Inference: Technology and Office supplies have always been on incremental profit over the years. The category of concern is Furniture for profit generation.

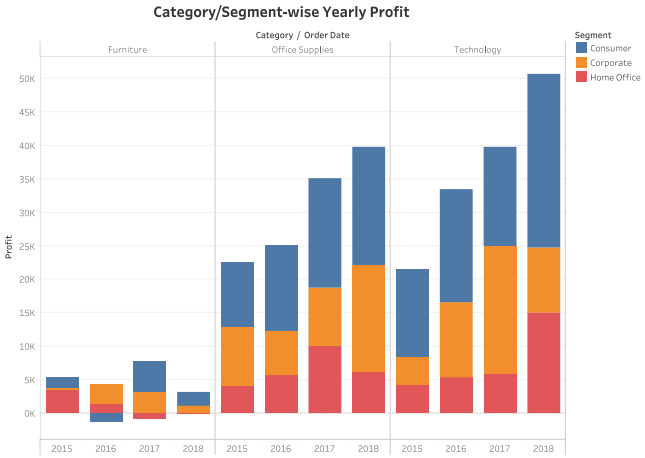

Figure 3- Profit for each category/segment over the years.

4 CONCLUSION

In this project, we primarily focused on sales, profit, and discounts in four regions of the United States and further in states of every region from 2015 to 2018. Secondarily, we showcased profit over categories and segments across all the years. While the eastern region presented gradual sales growth, it plunged for the western region in the year 2016. Nonetheless, the West region indicated the most tremendous sales and profit in 2018. For some states profit dropped when high discounts were accorded. Despite including great discounts states like New York and California kept the profit high. Intriguingly, no state suffered losses with zero discount.

With Tableau and raw statistical data, we may be able to draw some sort of inference based on our findings. Fortunately, with the assistance of information visualization tools, we are better prepared for what is to come.

5 ACKNOWLEDGEMENTS

The authors wish to thank Tableau community for developing the superstore data and Kaggle community for making it available.

6 REFERENCES

[1]. Kaggle Community - Superstore Dataset. Retrieved from https://www.kaggle.com/keyizhang14/superstore/version/1?select=Sample+-+Superstore.xlsx

[2]. Analysis on Superstore data using R https://www.slideshare.net/MonikaMishra15/superstore-data-analysis-using-r

[3]. Tableau Tutorial for beginners https://www.analyticsvidhya.com/blog/2017/07/data-visualisation-made-easy/

[4]. Tableau Software https://www.tableau.com

[5]. EDA on Superstore Data https://www.kaggle.com/swarnimapandey/superstore-analysis