

**FACULTY OF SCIENCE AND TECHNOLOGY**

**CST3340**

**Business Intelligence**

**Coursework 2 (Part B)**

STUDENT NAME:

**SUBMISSION DATE:**

# Introduction

## Brief Overview of Dataset

Due to the need of every man to purchase things so as to meet their needs, superstore is one of the stores human needs can be met. It is very orderly to buy things in a store than going into a market where there can be different issues on organization and coordination, it will be easy and orderly to get things from the superstore. This dataset offers a thorough examination of the ordering, profit and sales procedures of a superstore. This dataset offers countless opportunity to identify customer behavior, needs and level of engagement with the products. The extensive dataset will enable us to acquire precious insights with the firm when and where the trade is coming from, to identifying the types of items that made the greatest income from the company and so on.

This dataset is made up of over two million sales records.

**Features**

Order ID – Order Identification number.

Order Date – Order Date.

Ship Date – Ship Date.

Ship Mode – Ship Mode.

Customer ID – Customer Identification number.

Customer Name – Customer Name.

Segment – Segment of the ordered Product.

Country – Country where order originated from.

City – City where order originated from.

State – State where order originated from.

Postal Code – Postal Code of City.

Region – Region of Country.

Product ID – Product Identification Number.

Category – Category of ordered Product.

Sub-Category – Sub-Category of ordered Product.

Product Name – Product Name

Sales – Price of Product.

Quantity – Quantity of Product Order.

Discount – Discount of Customer.

Profit – Profit from Product.

## Discussion of Cleaning Undertaken

After a comprehensive check for

* Missing Data.
* Contaminated Outliers.
* Data Inconsistencies.
* Invalid Data.
* Duplicate Data.
* Date Data Type Issues.
* Structural Errors.

In this dataset, it was so fortunate that the data is clean; there is no missing data, no contaminated outliers, invalid data, and duplicate data. But during the visualization, I noticed there was a misplacement of the data types and inconsistency of the data. I also found out that during the geospatial visualizing, there were some data that was null which I have to find their classifications and the state and especially the postal zip/ code.

# Data Analysis and Visualization

Having accessible means to examine and comprehend data is more crucial than ever in our increasingly data-driven environment. This is where data visualization is useful. The preferred method for many firms to analyze and disseminate information is data visualization in the form of dashboards, charts, tables, graphs, geospatial and infographic visual elements with the aim of making data more accessible and meaningful.

The graphic display of information and data is known as data visualization. Data visualization tools offer an easy approach to observe and analyze trends, outliers, and patterns in data by utilizing visual components like charts, graphs, and maps. Also, it presents an excellent way for staff members or business owners to clearly deliver data to non-technical audiences.

Some advantages include easy sharing of information, ability to visualize patterns and relationships and interactively explore opportunities.

## Diagrams and Interpretations

**Highest and lowest Profits per State chart**

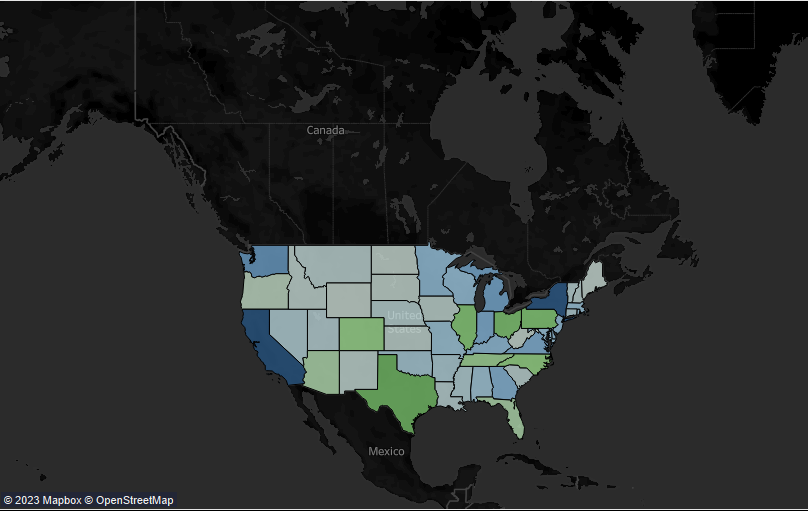
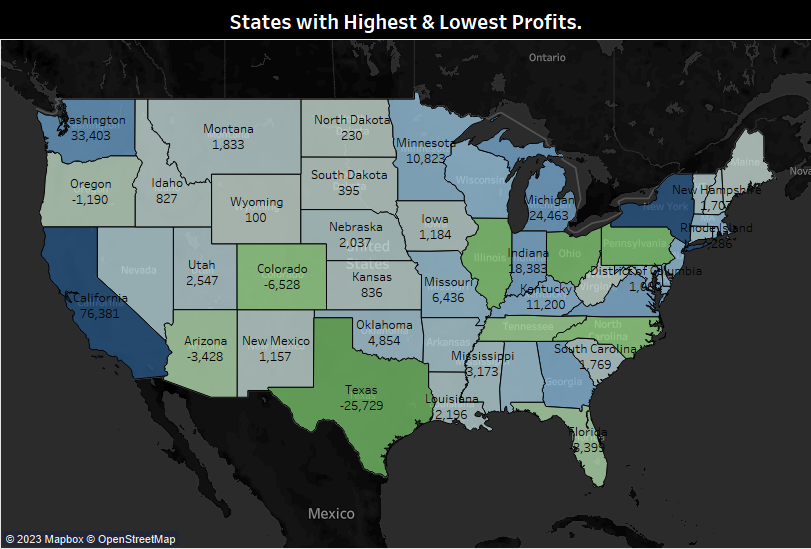


Fig1.1 Highest and lowest per state chart (complete view of the map).



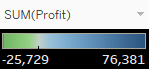


Fig1.2 Highest and lowest per state chart (View of the most populated part of the map).

**Discussion and interpretation**

This map is based on generated latitude (rows) and longitude (columns). The marking is automatic, and the states and the sum of revenue are labeled. Each state’s total profit is represented by a different hue, and the state and total profit are displayed in text. On this sheet, there are total number of 49 members for the state. The temperature diverging scheme is a combination of green-blue diverging palette, with green representing the state with the least profit and blue representing the state with most profit. According to the charts above, California generates the most profit, with a total of 76,381, while Texas generates the least, with total of -25,729. In light of this, California, a state in the west part of the country, can be said to be the one that produces the greatest profit.

**Sales vs Profit per Year**

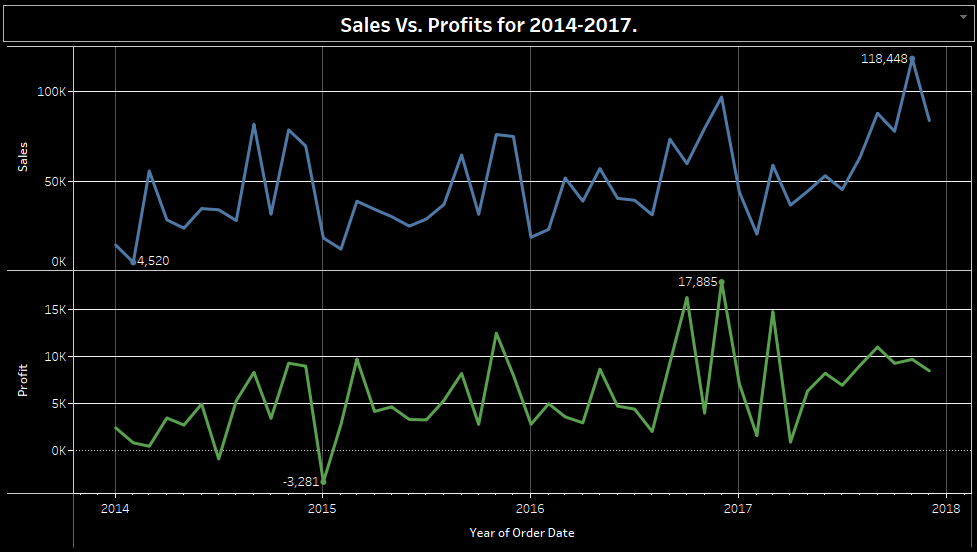


Fig 2 Sales vs Profit per Month chart

**Discussion and Interpretation**

This line chart depicts the total sales vs profit generated on each category of product sold in each state from 2014 to 2018. The color represents the sum of totals per year, and the marks are labelled with the sum of total. The row contains the total sum of sales vs profit, whereas the column tab has the years. Here, we conclude that in sales, 2017 is the busiest year, also in profit, 2016 is the busiest year for customer activities.

**Sub-Categories based on Quantities.**

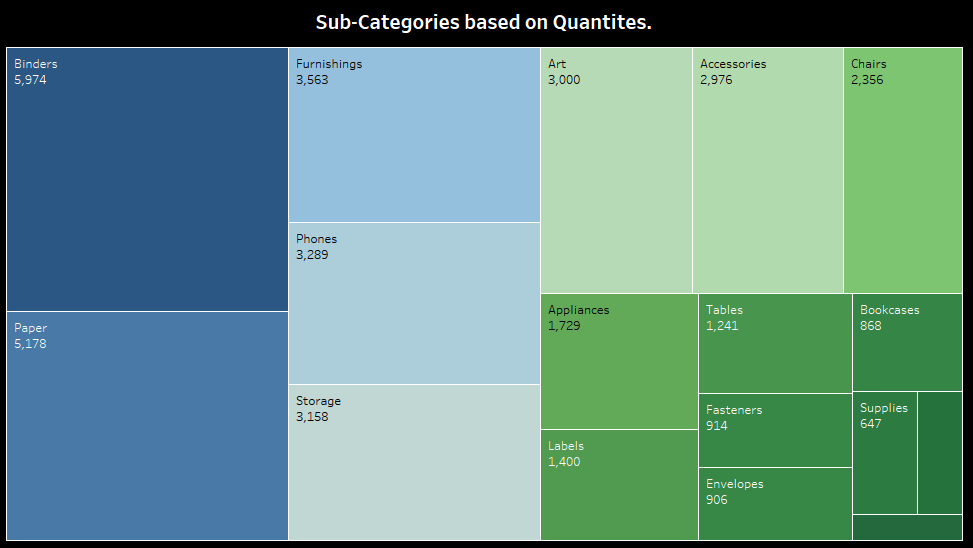


Fig 3 Sub-Categories based on Quantities

**Discussion and Interpretation**

This tree chart depicts the total quantity generated on each sub-category of items for sales that occurred that year, in order to achieve this, a Sub-Category per Quantity chart was formed. The color represents the overall quantity, and the marks are labelled with the Sub-Category and quantities. The total quantities ranges from 5,974 to 234. As a result, we can claim that the Sub-Category with the highest quantities is Binders while the least is Copiers.

**Sales of each sub-category compare to profits.**

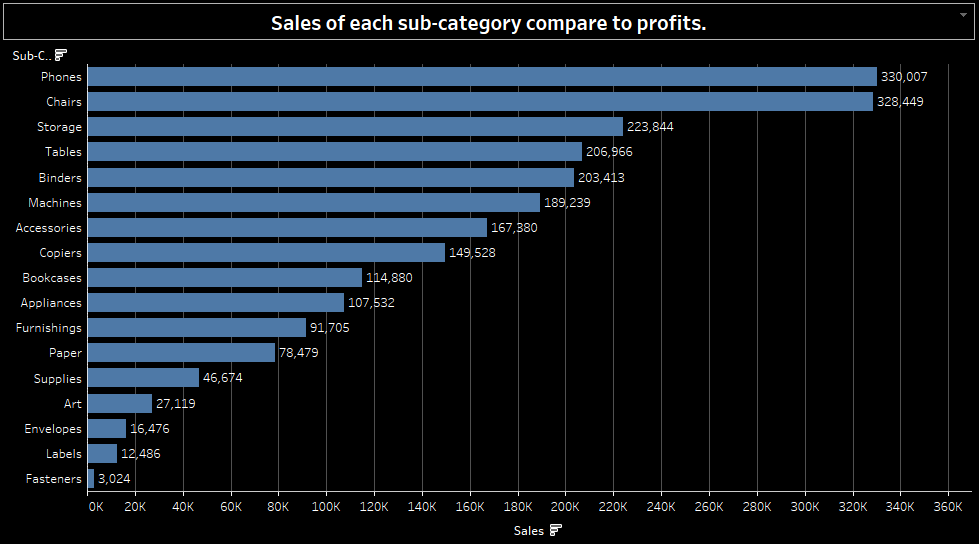


Fig 4 Sales of each sub-category compare to profits

**Discussion and Interpretation**

This bar chart depicts the total sales by various Sub-Category of items that occurred that year. According to “Chart junks” the more colors you put in a bar chat the harder it is to read the visualization and actually grasping the meaning. Sub-Category serves as the row here, while Sales serves as the column. On this chart, the sum of Sales runs from 3,024 to 33,007. They were sorted in descending order by the Sales of the Sub-Category. The Sub-Category has 17 members. According to the figure above, the Sub-Category product with the highest Sales is Phones with profit of 33,007 while the one with the least Sales is Fasteners with profit of 3,024.

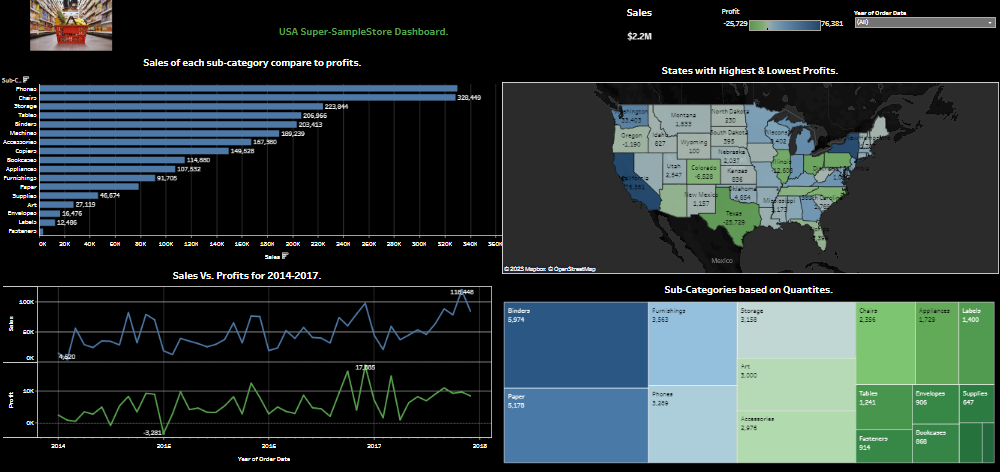
**USA Super-Sample store Dashboard **

Fig 5 USA Super-Sample Store Dashboard

# Overall trends and patterns observed

* According to Fig 1, California generated the most profit, with a total of 76,381, while Texas generated the least, with total of -25,729. In light of this, California, a state in the west part of the country, can be said to be the one that produces the greatest profit.
* There is a great peak in sales and profits in 2017 and 2016 respectively as depicted in Fig 2. The years had a total sales and profits of 118,448 and 17,885 respectively.
* The year 2014 and 2015 has the least sales and profits respectively generated which was a total of 4,520 and -3,281 respectively as sees in Fig 2 chart.
* The Sub-Category with the highest in Quantity is Binder as seen in Fig 3. Binder generated a Quantity of 5,974.
* Copiers had the least generated Quantity for the store. They are generated a total of 234.
* Phone in the Sub-Category has the highest sales as seen in Fig 4 with a total of 33,007
* Fasteners in the Sub-Category has the lowest sales as seen in Fig 4 with a total of 3,024.

# Selection of Data Mining Algorithm and Data Pre-processing

Data mining is the process of finding hidden and significant patterns in a large dataset, and it is used to evaluate enormous amounts of data. Due to the abundance of data mining techniques, selecting the appropriate one for a given data mining assignment can be challenging (Tang and MacLennan, 2005). Yet, the selection procedure becomes much more seamless when the analysis's priorities are straightforward. The analyst will need to select a data mining technique that can find the necessary trends in the data, but he or she will also need to make sure that the outcome is understandable.

With massive volumes of data being collected and stored on a constant basis, several businesses are becoming interested in data mining to discover patterns in their databases. The revelation of "Correlation Relationships" amid massive volumes of transaction records can aid in several decision-making processes, including catalogue design, cross-marketing, and customer shopping analysis. These mining algorithms are utilized in recommender systems.

For further analysis of this dataset, we’ll be using the A priori algorithm considering that a market basket analysis will be performed to better understand consumer behavior. The A priori Algorithm is a well-known Association Rule algorithm that is commonly used in market basket analysis. It is also thought to be more accurate than AIS and SETM algorithms. It aids in the discovery of frequent item sets in transactions and the identification of association rules between these items. The market basket analysis approach identifies client purchasing behaviors by identifying associations between the many goods that customers place in their "shopping baskets,". The identification of this type of association can aid retailers and marketers in developing marketing strategies by providing insight into the things customers regularly purchase together, whether from a grocery store or online retail. Looking at the series of visualizations carried out we have a lot of customers with different people who purchase from different categories and we can also see how some categories of goods make more revenue for the store, understanding the different behaviors the customers have and leveraging on this patterns will help the store to organize their ledge space for the best product placement, cross-sell, create store layouts based on consumer buying habits and purchases. Doing all this will result in an increase in sales, Improvement of customer experience, optimize marketing plans and campaigns and aid in demographic data analysis.

# Data Mining using Rules of Association

The use of an a priori algorithm to search for frequently occurring patterns in stores has increased steadily since 2000, reaching a high peak in 2018. The goal of this data mining effort is to look for correlations between certain stores that have increased in popularity since 2000 and have reached a high peak at specific locations. This is vital for fast order fulfillment because it brings the stores closer together.

Tests of encouragement and trust are used to compare which stores have the highest rating for products and which stores have the highest rate for stores that are the most commonly occurring in various regions. Once the rules of association are defined, it is easier for the company to work proactively in a way that anticipates the region's association. The absolute frequency of a simple store rent combination is measured using a 60 percent assist. Correlative frequency, or the probability of anything happening, can be determined with a confidence level of 75%.

# Conclusion

Data Science plays a critical role in the majority of internet utilities, not just in attracting new users but also in retaining existing ones. The explanation for this is that Data Science provides a more accurate view of your customers' preferences in the form of graphs and maps that use several metrics as feedback. This vital piece of knowledge aids you in molding your goods and services in a way that appeals to your clients as exclusive, drawing them to your website.

When a business-like USA superstore has an abundance of data, it's definitely a smart idea to bring the data to good use.

New technology also carries with them their own set of problems, but USA superstore has been addressing those problems head-on and regularly by soliciting audience feedback. USA superstore hopes to boost not only itself but also other businesses by open-sourcing some of its repositories and systems. Finally, it would be inaccurate to claim that USA superstores makes any of its choices based on Data Science insights because they do depend on a large number of human inputs.

The aim of this project is to create a product recommendation system and perform data analysis within USA. USA provided the data I used in this study.