A Case Of Super Store Sales - A Visual Exploration

UTHMAN BELLO MAIGARI

Abstract— In this paper, I will concentrate my efforts on analysing the superstore sales data. I will also be building a dashboard to provide a visual representation of overall sales at the superstore. A data analysis of the super store data is carried out in order to find trends over the past and present. Data visualisation is integrated with the dashboard so that the reader can see the results of the analytics within the dashboard. Some key findings were the sales made over time, growth rate within each year, etc.

1 Problem Statement

In the United States, Super Store operates a retail business. Consumers, corporations, and home offices buy furniture, office supplies, and technology products. Analysing and identifying weak areas and trends within Super Store's sales data will help me identify opportunities to track and monitor business growth over time. Knowing what works best for the company is essential with growing demands and cut-throat competition in the market. Additionally, effective strategic planning is an imperative aspect of the company's survival, to compete with other big superstore chains.

Visual analytics will be used as a tool to assist the superstore. This will be able to analyse sales of products over a period of time and identify trends over time based on sales. This study is designed to determine if virtual analytics can be used to deduce any phenomena, for example, a region, consumer, or product, in this case. In my paper, I will attempt to answer some of the following questions:

- How many sales will be made annually?
- What were the sales each agent produced?
- What are the sales by category and sub-category?
- What are the top products sold based on sales?

2 State of the Art

A) Sales Analysis and Performance of Super Store Using Qlik Geo-Analysis.

An analytical tool called Qlik is presented in this paper. Basically, Qlik provides comprehensive geospatial and mapping operations. A variety of refined geoanalytic use cases can be addressed using geo-operations such as period of time, within, and highest square measure, combined with made mapping options.

Although the use cases for mapping and geo-analytics are virtually endless, and they can vary widely depending upon the type of business that is utilising them, there are still one or two that nearly any company can find helpful right out of the box. Including:

- Visualisation: Agglomeration and warmth maps, KML layers, thematic maps, and routes.
- Analytics: Detailed filtering, numeric aggregates, dataset management, and charting.
- Management: Territory management, machine-controlled assignment plans, mass updates, and custom markers and shapes.

The main reason they chose Qlik in this paper was because of its Ability to develop information visualisations, Ability to create dashboards on ad-hoc information sources, and Ability to transform data into stories.

Their data included sales, profit, and geographical information about Super Store. Based on their visualisation and graphs some of the questions they addressed are:

- Performance of a store based on sales and consumers.
- The customer's behaviours and patterns.
- The sales and profit for each region.
- The potential reach of each store.

In this paper, the authors compiled data, built interactive information visualisations, created reports and dashboards. This was the way they approached their analysis.

B) Data Analysis, Sales Of Super Store In Chicago.

Based on my reading the purpose of the study is to perform an inference statistical analysis of the dataset and selected variables. This project will allow them to learn how to calculate correlations, regressions, and ANOVAs between all variables that make up the superstore record. The outcome of the project is that every superstore will be able to increase profits and sales.

According to their specified data, they used the Superstore Record in 2019 dataset in the United States.Because of the large amount of data, they only chose the Superstore record in Chicago that contained only 113 sample sizes. Data from this dataset shows every Superstore item in the Chicago area that has already been sold online. In this data, we can see the category of each item that has been sold along with the date of order. Based on their visualisation and graphs the main questions they addressed was the sales and profit margin of products. In this paper, the authors' analysis approach was using R-Studio where they can get the precise analysis and also learn implementing R-Studio by doing the project also gave them great experience to manage data set and having high critical thinking to sort data variables to use for each analysis such as Regression, Anova , Correlation.

The first paper is applicable to my problem. I would incorporate a similar tool called ‘tableau’ to carry out some of my visualisation analysis and also incorporate a dashboard. The main takeaway from what I learned from these papers was the approach each author took in solving their specific problem or unique question.

3 Properties of the Data

The dataset is based on Superstore, a well-known retailer in Canada. Sales agents are placed in every state in the USA as part of their business model. Each sales agent is responsible for bringing in sales for the state to which they are assigned. There are 7385 entries in the data, and there are 20 columns with 3 float types, 2 integer types, and 15 object types. In this case, there is no missing value, so the results or graphs will be unbiased. Datasets were obtained through Kaggel, an online community for data analysts to extract data.

This is the structure of the dataset:

Row ID: Unique row ID.

Order ID: Unique identifier of each order.

Order Date: Date of order.

Ship Date: Date that the product was shipped.

Ship Mode: Shipping types (e.g., Second Class, Standard Class, First Class, Same Day)

Customer ID: Unique identifier for each customer.

Customer Name: Name of the customer.

Segment: Customer segment (e.g., Consumer, Corporate, Home Office)

Country: Country where the customer ordered from.

City: City where the customer is from.

State: State where the customer is from.

Postal Code: Postal code where the customer is from.

Sales Agent ID: This is an identification for each sales agent.

Region: Region where the customer is from.

Product ID: Unique identifier of each product.

Category: Overall category that the product fits into (for e.g., Furniture, Office Supplies, Technology)

Sub-Category: Subcategories of categories (for e.g., Bookcases, Chairs, Labels)

Product Name: The name of the product.

Sales: Gross sales per order.

Quantity: The amount of product in each order.

The other dataset contains information about sales agents. Sales agents are responsible for generating sales in the state they are assigned to. In both datasets, sales agents IDs are shared, which is helpful for identifying which sales belong to which agent. There are 49 sales agents in total, listed by their names, level (junior, mid-level, executive), their assigned state, and age. The data contains 49 entries, 6 columns in total.

It was with the help of my Jupyter notebook that I carried out my data pre-processing and some data cleaning in order to ensure that the data quality and the detection of problems was accurate. In terms of missing data, my dataset had none which was a positive in my case.

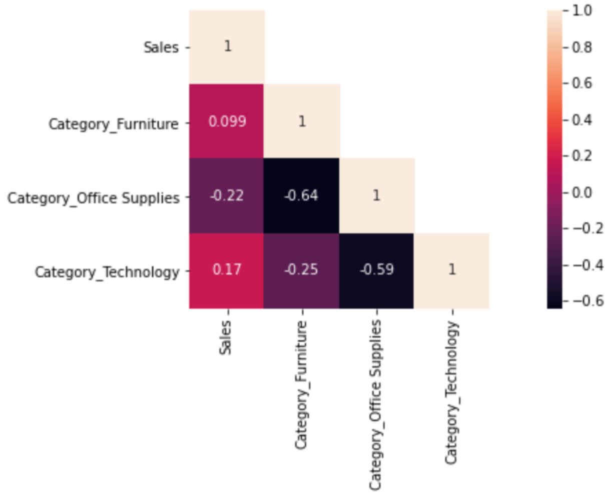


Figure 1: shows the correlation matrix.

We can gain a comprehensive understanding of the features' relationships by plotting a correlation matrix as shown in figure 1. I use this term to describe a relationship between categories based on products and sales of those products within those categories.

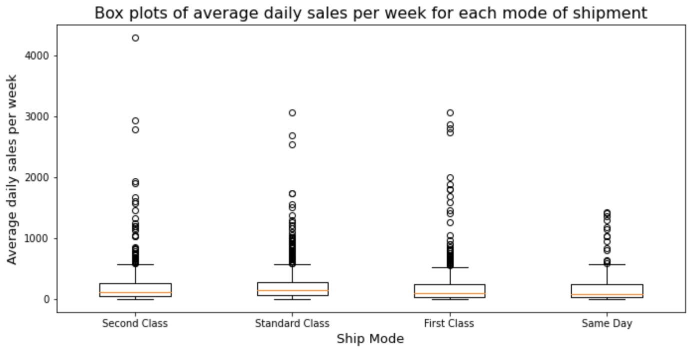


Figure 2: shows the daily sales for each mode of shipment.

Figure 2 shows, for each mode of shipment, the daily sales figures which are presented to indicate if there are any outliers within the daily sales data.

4 Analysis

4.1 Approach

In this report, *figure 3* depicts the analytical steps that were selected as a method of solving this problem - their purpose as well as the methods used to solve it. The following sections describe in detail each step of the process, including the relationship between the human reasoning process and the computational methods used in the process. Throughout the analysis process, both are essential. Only by combining them effectively can we be able to come up with a comprehensive analysis that can be applied to each step.

DATA PREPARATION

During the initial steps of the problem domain analysis, it is essential to use human reasoning to formulate hypotheses relevant to the problem domain. This is to determine if the available data is appropriate, and to identify the features most relevant to solving the problem. It is possible to gather a robust dataset cost-effectively using appropriate engineering methods. Furthermore, human reasoning is vital for understanding the properties of data by exploring its domain, such as structure, precision, and quality, and referring back to the domain. In order to support human reasoning, metadata information must be provided, including statistics summarising the data, histograms and scatter plots to determine its completeness, meaning, and transformation needs. As for data preparation, human reasoning allows the selection of a subset of meaningful columns. Computational methods allow the identification and filling in of missing values with PythonCode or Excel.

STEP 1: To begin with, I decided to perform a data cleansing process. This basically involves exploring existing features, as well as correlations between the features in the data. Following that, I continued to identify outliers to see if there are any features that are of significance to me. As soon as I have observed the results I am looking for, I then generate visual representations that illustrate them.

STEP 2: The second step in my analysis is performing an exploratory analysis. This is the purpose of which is to visually examine existing data and discover if there are any possible relationships between variables that might otherwise remain unnoticed or overlooked. Exploratory analysis will attempt to identify such relationships. As a result, it can be very helpful for setting up hypotheses for further testing if we are able to discover new connections. During an exploratory analysis, you try to get an overall picture of the data you are working with and let the information speak for itself. In a Jupyter notebook, I create visualisations exploring these ideas. For example, shipment modes, regions, customers, state revenue, etc. (Graphs such as a heat map, correlation matrix, scatter plot, etc.))

STEP 3: In my third step, I conducted a descriptive data analysis, which was the approach I used, since the data is related to business. Using historical data, it attempts to explain what happened. As a result, it precisely serves as a tool for tracking KPIs (Key Performance Indicators) and evaluating the company's performance. A dashboard that visually summarises leads acquired or closed deals relies on descriptive analysis, as do revenue or website visitor reports. The approach combines data from multiple sources to provide valuable insights into the past. I create visualisations in Tableau showing these ideas (i.e dashboard, sales, profits, growth rate etc) (Graphs such as bar, histogram, donut chart, etc)

STEP 4: At the end of this step, I intend to bring the information I have gained from steps 2 and 3 into an answer to the problem statement. In my scenario, I will be generating visualisations based on the analysis results, which will equate to a dashboard in terms of my analysis. My main tool for analysing data is going to be Tableau. It will be my go-to tool for the majority of my data analyses

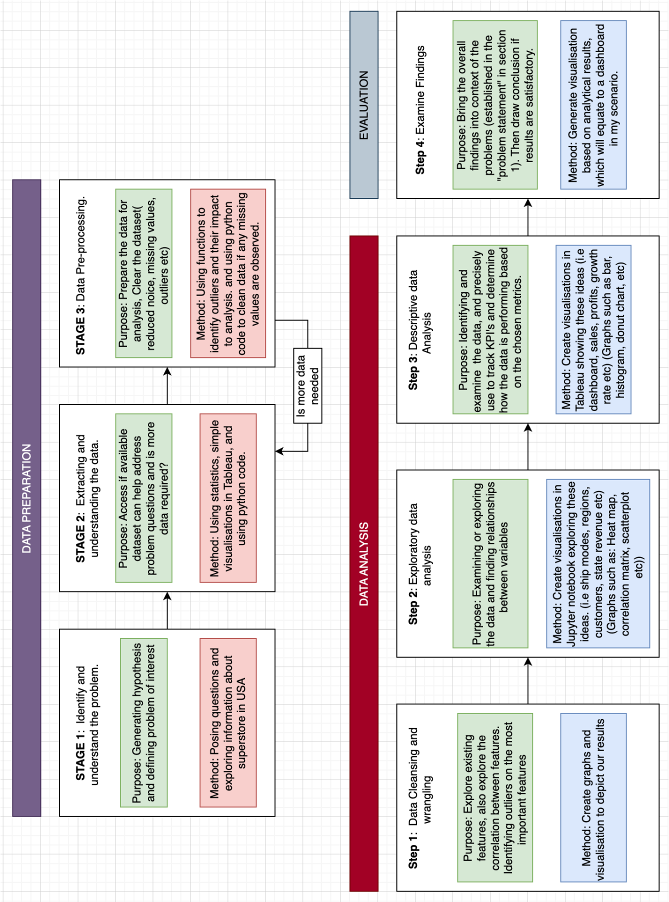

DATA PREPARATION
STAGE 1: Identify and understand the problem.
Purpose: Generating hypothesis and defining problem of interest
Method: Posing questions and exploring information about superstore in USA
STAGE 2: Extracting and understanding the data.
Purpose: Access if available dataset can help address problem questions and is more data required?
Method: Using statistics, simple visualisations in Tableau, and using python code.
STAGE 3: Data Pre-processing.
Purpose: Prepare the data for analysis, Clear the dataset(reduced noice, missing values, outliers etc)
Method: Using functions to identify outliers and their impact to analysis. and using python code to clean data if any missing values are observed.
Is more data needed
DATA ANALYSIS
Step 1: Data Cleansing and wrangling
Purpose: Explore existing features, also explore the correlation between features. Identifying outliers on the most important features
Method: Create graphs and visualisation to depict our results
Step 2: Exploratory data analysis
Purpose: Examining or exploring the data and finding relationships between variables
Method: Create visualisations in Jupyter notebook exploring these ideas. (i.e ship modes, regions, customers, state revenue etc) (Graphs such as: Heat map, correlation matrix, scatterplot etc))
Step 3: Descriptive data Analysis
Purpose: Identifying and examine the data, and precisely use to track KPI's and determine how the data is performing based on the chosen metrics.
Method: Create visualisations in Tableau showing these ideas (i.e dashboard, sales, profits, growth rate etc) (Graphs such as bar, histogram, donut chart, etc)
EVALUATION
Step 4: Examine Findings
Purpose: Bring the overall findings into context of the problems (established in the "problem statement" in section 1). Then draw conclusion if results are satisfactory.
Method: Generate visualisation based on analytical results, which will equate to a dashboard in my scenario.

Fig 3. Approach to visualization and human interaction.

Green: indicates human reasoning.

Blue: indicates visualization.

4.2 Process

In this section, we will follow the 2 main **steps** mentioned in the previous section and will try to answer our four research questions:

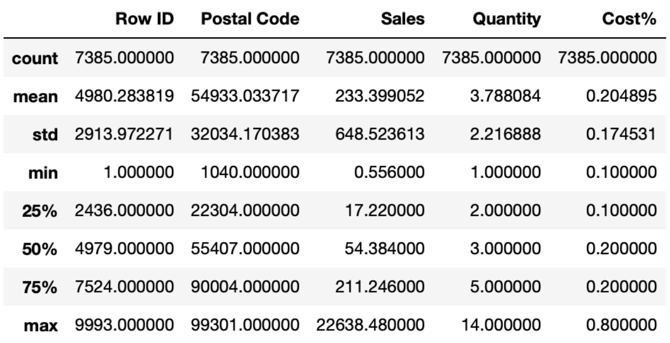
- How many sales will be made annually?
- What were the sales each agent produced?

- What are the sales by category and sub-category?
- What are the top 10 products sold based on sales?

4.2.1 Data Exploration Analysis.

The purpose of this section is to drill deeper into the data in order to gain more insight into it. As a starting point, we must determine what the problem is that we wish to solve and what questions we would like to answer. As part of the learning process, I need to be familiarised with the information of the data in terms of the number of columns, column labels, columns data types, memory usage, range index, and the number of cells in each column (non-null values).

I will be using the describe function to encapsulate more information about the numerical data within the data frame.



	Row ID	Postal Code	Sales	Quantity	Cost%
count	7385.000000	7385.000000	7385.000000	7385.000000	7385.000000
mean	4980.283819	54933.033717	233.399052	3.788084	0.204895
std	2913.972271	32034.170383	648.523613	2.216888	0.174531
min	1.000000	1040.000000	0.556000	1.000000	0.100000
25%	2436.000000	22304.000000	17.220000	2.000000	0.100000
50%	4979.000000	55407.000000	54.384000	3.000000	0.200000
75%	7524.000000	90004.000000	211.246000	5.000000	0.200000
max	9993.000000	99301.000000	22638.480000	14.000000	0.800000

Figure 4: shows the nominal data.

Count: The number of not-empty values, mean: The average (mean) value, std: the standard deviation, min: the minimum value, 25%: The 25% percentile, 50%: The 50% percentile, 75%: The 75% percentile, max: the maximum value. Percentile is a measure of how many of the values are less than the percentage given.

When I explored the dataset, I discovered that another feature (Region) would need to be visualised in order to make the data more meaningful. In a nutshell, the region refers to the place where each state is located geographically. In this dataset, there are four regions that are represented, namely the west, east, central, and south regions. This finding piqued my curiosity, and I decided to dig deeper into it as a result of my curiosity. According to the graph shown in *figure 5*, I have chosen to extract the sales data for each region (using the sort value function) which will give me a more complete picture of the situation based on the sales data. According to the data, we can see that the west region has the highest sales with $539,710, the east region has $493,981, the central region has $398,054 and the south region has the least with $291,905. There is a tendency for the west to have the highest sales in the United States due to the fact that it is the largest region.

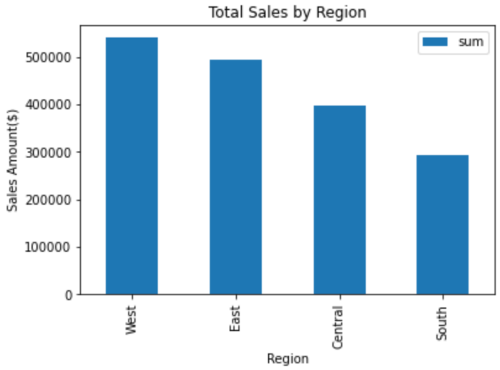


Figure 5: shows the Regions with their sales count.

The Analysis takes into account the overall sales from the 2017 to 2020.

As I was analysing the dataset, I had the idea that once I had identified the regions, I would want to get all the states that were contained in them and explore them further. According to the data there are 20 different states in total. The states that are in the data are California, New York, Texas, Washington, Pennsylvania, Florida, Virginia, Michigan, Georgia, and others. As a result, I decided to identify the amount of revenue that each state will accumulate during the period between 2017 and 2020.

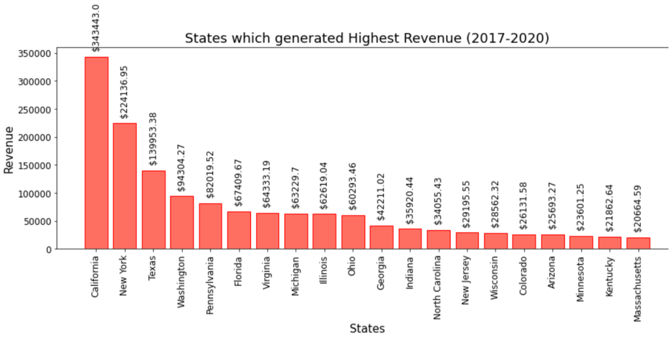


Figure 6: shows the states generating the highest revenue.

In the graph above, it can be seen that people from California contribute the highest number of revenue (with $343,443) to Superstore and the least state to generate revenue is Massachusetts (with $20,664). It might be due to the Superstore's location in California and it's a big city, which might be a contributing factor. In order to increase the sales of the Superstore, however, it is crucial to reach those states that make the smallest contribution to the sale of the product. They might not be able to access the superstore because of their location, which renders them inaccessible to it. As a result of solving this issue, the store might be able to boost its sales if it offers a delivery or advertising service to the targeted state.

To further my curiosity, I decided to construct a correlation matrix in order to gain a deeper understanding of it. This correlation matrix is an important data analysis metric that is computed to summarise data in order to gain a better understanding of the relationship between various variables and to make the appropriate decisions based on that understanding. To compute the matrix, Pandas DataFrame uses the **corr()** method. A correlation coefficient of Pearson's is calculated by default. The parameter 'method' could also be set to another method, such as Spearman's coefficient or Kendall Tau correlation coefficient. A **heatmap()** method was also used to plot the matrix.

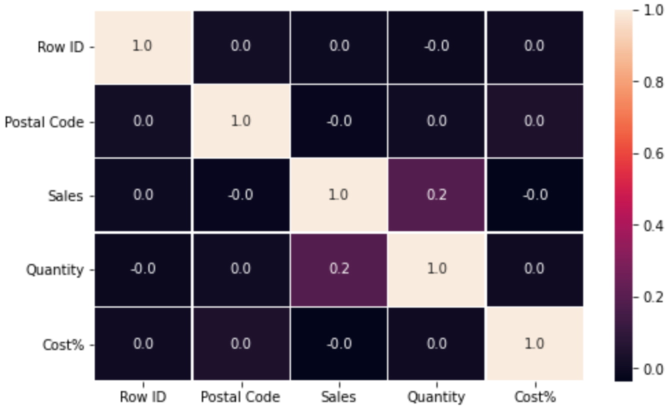


Figure 7: shows the correlation matrix.

Whenever I look at a correlation matrix, I keep in mind the following: All diagonal elements are 1. A diagonal element represents the correlation of a variable with itself, so it will always equal 1. An increase in one variable's value will increase the value of the other variable if the value of one increases (near 1.0). In the absence of any correlation between two variables (positive or negative), a value close to zero indicates that those variables are independent. Observing what I have seen so far, I believe that Sales and Quantity are positively correlated. In this situation, it would be reasonable to assume that a product's price would vary according to the quantity of products being sold.

4.2.2 Descriptive Data Analysis

This section is devoted to the identification and examination of the data, and how it can be accurately used to track KPIs and analyse how the data is performing based on the chosen metrics and how it can be improved. Also, the main analysis of the data will be performed using the tableau software tool, which is effective in analysing data.

It was my intention to observe and visualise the other dataset that was included in this analysis, which was the data from the sales agent. It is the responsibility of sales agents to generate sales in the state in which they are assigned. Interestingly, both datasets share a common set of sales agents' identifiers, which is helpful in identifying which sales are the responsibility of which sales agents. There are 49 sales agents in total, listed by their names, level (junior, mid-level, executive), their assigned state, and age. In total, there are 49 entries in the data, and there are 6 columns in total.

Next, we had to visualise the sales agents in the form of the names of the agents, which is a drag and drop feature in Tableau. In order to see if there were any trends in their sales, I'd then like to add their sales next to the sales agent. To show each sales agent with his/her sales based on the selected year, I created a bar graph of the agent's sales based on that selected year (which will be shown below). Also opted to add a feature which will be the growth rate, which is the percentage increase of the sales done in the previous years.

The growth formula:

GROWTH RATE = PRESENT - PAST / PAST

It consists of taking our present value, subtracting it from our past value, and dividing it by our past value. To do this in tableau, I navigate to the calculated field. In a calculated field, a value is created by calculating database fields rather than directly storing them in the database

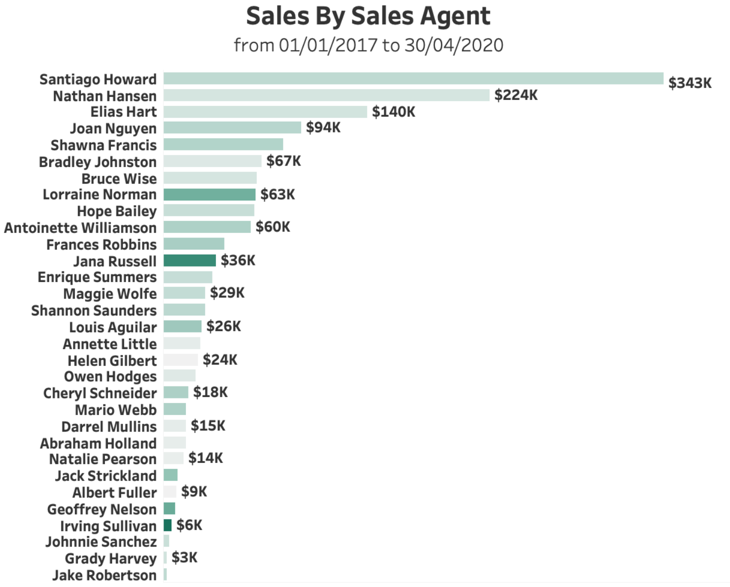


Figure 8: shows the sales each sales agent made overall.

Based on figure 8 we can see that Santiago Howard has the highest number of sales with $343k (2017 to 2020) and Grady Harvey has the lowest number of sales with $3k (2017 to 2020). Green is a colour that essentially signifies growth. In general, the darker the green, the more likely it is that sales agents will grow in the future. The research question has now been answered based on my first section (the 'Problem Statement').

Once I observed the data again, I continued to analyse it. In my analysis, I am looking for the category and subcategory feature within the data. I would like to point out that the category contains furniture items, office supplies, and technological items. This sub-category consists of a variety of items, including phones, chairs, racks, machines, tables, binders, accessories, art, and many others. It is the purpose of this study to analyse both features regarding their impact on sales. In addition, it will pinpoint which features have a higher level of revenue than others with respect to sales.

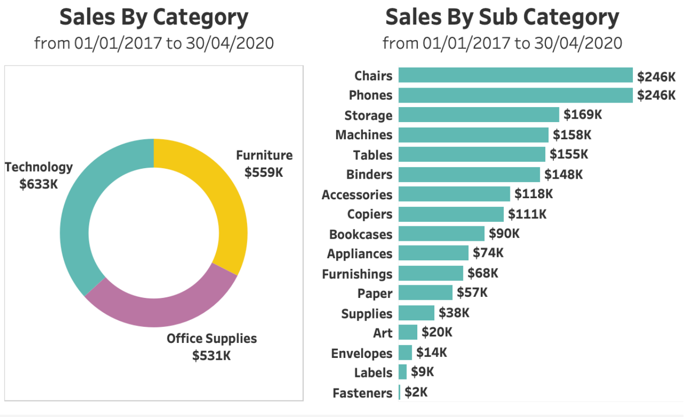


Figure 9: shows the sales By category and sub-category.

According to *figure 9*, the superstore's technology category contributes the highest sales ($633k) in profits to the superstore, while the furniture category contributes the second highest sales ($559k). Office supplies make up about the same amount ($521k) as furniture, but at a lower rate. As a result, it is recommended that the superstore concentrate more on office supplies sales.

The top two best-selling subcategories are phones and chairs. Compared to other areas, Art, Envelopes, Fasteners, and Label make close to zero margin to losses. Products such as these may be dropped from the product catalogue or their prices and profit margins may be increased. In addition, bargains may be made with suppliers for a lower price. I've taken into account the overall years of the superstore, but based on when I switch the years, certain categories or sub-categories are more dominant than others. The research question has now been answered based on my first section (the 'Problem Statement').

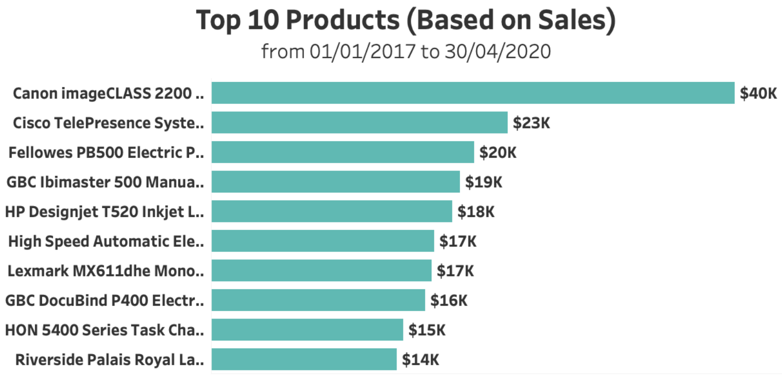


Figure 10: shows the sales of the top 10 products.

The figure shown above illustrates the highest-selling items in the top superstores during the entire period of 2017-2020, taking into account the overall amount of sales across the superstores. With sales of $40k, the Canon image camera is the top product in terms of profits that generates the most sales. The research question has now been answered based on my first section (the 'Problem Statement').

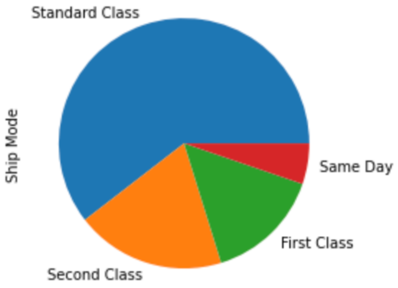


Figure 11: shows the shipping mode.

As shown in the pie chart above, shipping methods are categorised based on the selections made by customers. There are 4,466 customers who opt for the standard class, 1,427 customers for the second class, 1,103 customers for the first class, and 389 customers opt for same day service.

4.3 Analysis Results

There are a few significant findings in this study, including the income of sales agents, the states that generated the highest revenue, and sales by subcategory and category. The low-income states, like Massachusetts, appear to have the lowest sales profits out of all the states. Due to the fact that the west region consists of many states, it accounts for the sales. In terms of revenue sales from the period of 2017-2020, Sean Miller was the customer who brought in $24,508 in revenue sales which makes him a valuable customer.

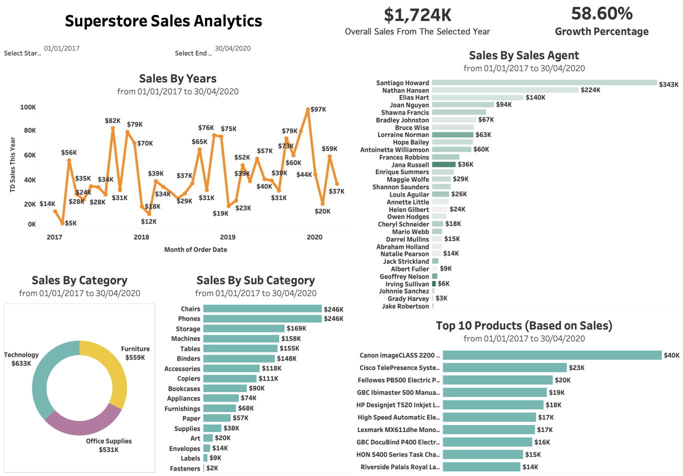


Figure 12: shows the Dashboard in tableau.

In order to provide a visualisation of the analytical question that I had in section 1, I have built interactive analytics in tableau based on the dashboard above. The key graph I want to highlight is the one that shows 'sales by year' over a period of time. A graph like this shows the annual growth rate of the company over the past few years, taking into account the monthly sales made during the past 4 years. Accordingly, it can be assumed that the profit for Superstore will be a fairly bullish trend over the next few years based on the current trends. On the dashboard, you can also see an overview of sales by year, as well as percentage growth by year for the selected years. In terms of overall sales, the company generated $1,724,000 and the growth rate was 58.60%.

5 Critical Reflection

Dashboards are powerful executive reports that are easy to create with data visualization skills. Excel dashboards are flexible, allowing you to design them according to your requirements. In addition, one might want to implement it in accordance with the prototype implementation. Furthermore, Excel dashboards are one of the cheapest tools for business intelligence. There was a limitation to this course work in that there was limited data. A major constraint was that a particular attribute could not have more than two conditions. If the data were more numerical, more functions could be applied. In turn, this results in a more predictable and effective sales process for teams, regions, and local companies. Users are able to view the whole picture that exists within their data, allowing them to make better decisions. It offers a comprehensive overview of the analytical capabilities that are used in sales management to unlock the power of information.

References

The list below provides examples of formatting references.

[1} Saddhartha Ghosh, Kandula Neha. Sales Analysis and Performance of superstore. April 2020

[5]Richard Veryard , Pragmatic Data Analysis. Oxford : Blackwell Scientific Publications. ISBN 0-632-01311- 7, 1984.

[6]Adèr, Herman J. , "Chapter 14: Phases and initial steps in data analysis". In Adèr, Herman J.; Mellenbergh, Gideon J.; Hand, David J. Advising on research methods : a consultant's companion. Huizen, Netherlands: Johannes van Kessel Pub. Pp. 333–356. ISBN 9789079418015. OCLC 905799857, 2008.

[7]Adèr, Herman J ,"Chapter 15: The main analysis phase". In Adèr, Herman J.; Mellenbergh, Gideon J.; Hand, David J. Advising on research methods : a consultant's companion. Huizen, Netherlands: Johannes van Kessel Pub. Pp. 357–386. ISBN 9789079418015. OCLC 905799857, 2008.

[8]Adèr, H.J. & Mellenbergh, G.J. (with contributions by D.J. Hand) . Advising on Research Methods: A Consultant's Companion. Huizen, the Netherlands: Johannes van Kessel Publishing, 2008.

[9]ASTM International (2002). Manual on Presentation of Data and Control Chart Analysis, MNL 7A, ISBN 0- 8031-2093-1

Websites

[1] My note books and tableau folder: https://cityuni-my.sharepoint.com/:f:/g/personal/uthman-bello_maigari_city_ac_uk/EjHJA5wylK5FuiXku_HWFX4BaAi0DpAxiFNyTLOfKGViKQ?e=BvcrPV

[2] https://medium.com/@dee_75726/how-to-visualise-your-data-better-d57b1e3203eb

[3] https://medium.com/swlh/product-sales-analysis-using-python-863b29026957

[4] https://www.velvetech.com/blog/types-of-data-analysis/