WQD7001 PRINCIPLES OF DATA SCIENCE

GROUP PROJECT PART 1 (G2.9)

PROJECT TITLE:

Analyzing Customer Preferences and Effective Marketing with a Data-Driven Approach to Retail Strategies

Team Member:

No	Matric ID	Name	Roles	E-Portfolio Link
1	S2192852	Jia Hui Wong	Generalist	https://jhwong97.github.io
2	S2191926	Kar Hong Sam	Leader	https://karhong-sam.github.io/
3	22060214	Mei Zhu	Secretary	https://22060214.wixsite.com/christine
4	S2177044	Wing Hong Cheah	Maker	https://winghongjason.github.io/
5	S2158054	Yuan Wei Kam	Oracle	https://yuanweiagatha.wixsite.com/yuan-wei-kam-s-e-por

Table of Contents

1 Project Title

Analyzing Customer Preferences and Effective Marketing with a Data-Driven Approach to Retail Strategies

2 Project Abstract

The COVID-19 pandemic has accelerated the growth of e-commerce, leading to increased competition among online retailers. To remain competitive in this industry, online retailers must adopt data-driven decision making. This project involves the analysis of a superstore dataset using the CRISP-DM cycle and conducting exploratory data analysis. The objective of this analysis is to provide end users or business owners with insights derived from the available data, enabling them to identify the most profitable products and determine the best and worst performing sales regions. By gaining an understanding of customer preferences through these insights, online retailers can develop effective retail strategies to optimize their profit margins. Additionally, this project aims to develop a predictive model for forecasting profitable products and improving profit margins.

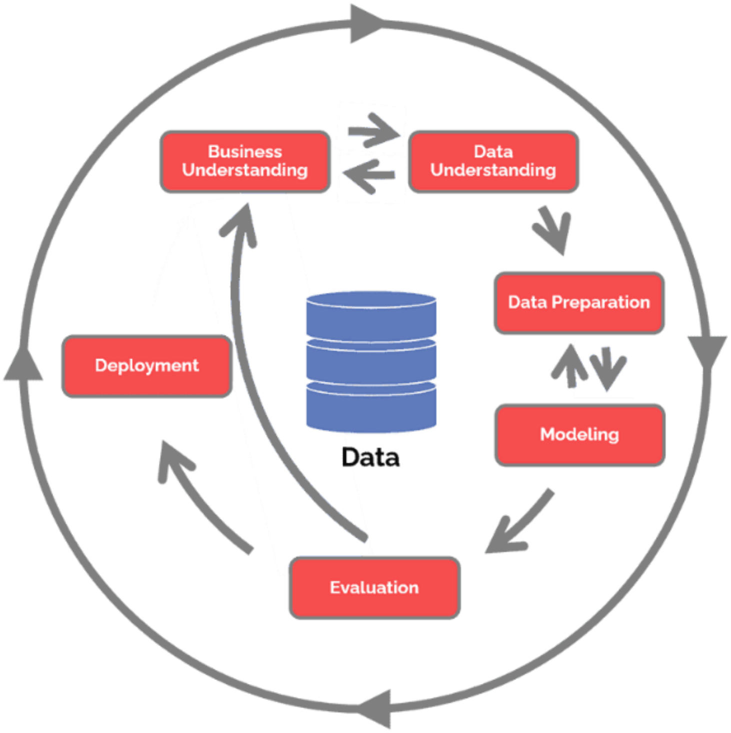
3 Introduction

The COVID-19 pandemic has accelerated the growth of the e-commerce or online retail industry, resulting in an increase in online sales and a shift in consumer behavior towards online shopping (United Nations Conference on Trade and Development, 2017). Therefore, many physical retailers have started to explore the possibilities in online retailing as the pandemic has forced many brick-and-mortar shops to temporarily or permanently close. Besides, some retailers have also adopted both online and offline modes to provide a better shopping experience for their existing customers. The growth of e-commerce results in an increase in competition between retailers, which potentially induces some negative impacts, such as struggling to retain existing customers and a loss of market share (Organisation for Economic Cooperation and Development, 2018).

Online retailers are lacking the ability to use big data, and only a few e-commerce enterprises can gain insights from big data, such as reducing the cost of market promotion and setting the business direction (Xia & Lv, 2021). The reasons stopping online retailers from using big data are mainly due to a lack of awareness, a lack of proficiency in using big data technology, and a lack of financial support. However, online retailers can remain competitive without the aid of big data technology by practicing data-driven decision making based on their currently available data.

As stated in an article from Harvard Business School online, data-driven decision making is the process of using data to inform your decision-making process and validate a course of action before committing to it. The key benefits of implementing it are making the right decision, identifying business opportunities and potential threats, and promoting cost savings by optimizing business operations (Tim, 2019).

In this project, the Cross Industry Standard Process for Data Mining (CRISP-DM) is implemented to analyze a selected superstore dataset. Figure 1 shows the diagram of the CRISP-DM workflow which consists of six (6) main phases. The selected superstore data set is used for analyzing customer preferences and developing sustainable marketing strategies. Through the data analysis, the end users are able to gain insights on their product sales and performance of sales in different regions. By understanding their current business's performance, they are able to develop a business plan to resolve any potential problems. In addition, this project also aims to develop a machine learning model to predict profitable products and improve profit margins.



4 Problem Statement

The e-commerce industry has experienced rapid growth in recent years, leading to increased competition among online retailers. However, despite the abundance of available data, many e-commerce businesses fail to fully utilize this information to gain a competitive advantage. In order for online retailers to stay competitive in this era, data-driven decision making plays a crucial role. With the vast amount of data available, business must use sophisticated data analysis tools to extract insights that can drive effective decision making. Those who fail to do so may be at a disadvantage in the marketplace. To succeed in the highly competitive e-commerce industry, online retailers need to analyze the available data to understand their customer preferences and develop sustainable marketing strategies. In this project, data analysis is carried out to address these challenges by analyzing a given superstore dataset to analyze customer preferences and promote effective marketing.

5 Project Objective

The objectives of this project are:

- To identify the most profitable products or categories to optimize sales and revenue.
- To identify the best and worst performing regions/markets to optimize sales and revenue.
- To develop a machine learning model to predict profitable products and improve margins.

6 Scope and Domain

The scope of this project is focused on data-driven strategies for online retailers which covers three specific areas of focus:

- The analysis part of this project focus on sales data for all products and categories sold by Superstore.
- Carry out analysis on the sales and profit data across different geographic regions/markets where the Superstore operates.
- Collect and analyse historical sales and profit data for all products/categories sold by the Superstore and use the collected data to train a machine learning model for classifying if a product is profitable or not.

7 Summarize Literature Review

Superstore analysis is a critical aspect of modern retail management, as it enables retailers to gain valuable insights into consumer behaviour, product trends, and overall market dynamics. Two recent articles, "Data Science Innovation in Supermarkets: A Case Study of Woolworth" and "Data Science in Supermarkets," highlight the importance of using data science techniques to analyse the massive amounts of data generated by superstores, to better understand consumer needs and preferences. The former article examines how Woolworths, one of Australia's largest retailers, uses data science to enhance their operations and customer experience, while the latter provides an overview of the role of data science in the retail industry and highlights its key applications in supermarket settings.

Superstores offer several advantages over traditional supermarkets, such as a wider range of products and more personalized customer experiences. However, superstore analysis faces some challenges, such as the lack of standardization in data collection and difficulties in identifying patterns in large datasets. Data science can help bridge this gap by providing practical insights through the analysis of data collected from various sources. Proper data preparation and cleaning, along with the use of advanced analytical tools, can aid in identifying trends and patterns in consumer behaviour, leading to better decision-making and increased profitability for superstores. All in all, data science is an essential tool for superstores to gain a competitive advantage and make informed decisions in today's retail landscape.

8 Data Science Life Cycle:

8.1 Business Understanding

As mentioned in the introduction of this project, the impact of the Covid-19 pandemic results in the booming on online retail industry and causes increased competition among online retailers. It is a challenging task for online retailers to remain competitive in the industry without data-driven decision making. Therefore, it is crucial for online retailers to practice or adopt data-driven decision making based on any existing available data such as customer data and product sales. By practicing data-driven decision making, it helps to develop better retail strategies via analyzing customer preferences and effective marketing.

8.2 Dataset Preparation (Obtain and Clean)

The Superstore dataset, which can be found on Kaggle, is a fictitious sales dataset that provides transactional data for a retail company. This dataset is divided into three main categories, namely furniture, office supplies, and technology. Each category contains several sub-categories that include bookcases, chairs, tables for furniture and phones, accessories, and machines for technology. The dataset consists of 9994 observations and 21 variables, including customer names, product categories, shipping details, order dates, sales amounts, profit margins, quantity sold, discount offered, order priority, shipping mode, region, and sub-category. The sales data covers four regions (Central, East, South, and West) in the United States and spans over a period of four years from 2014 to 2017.

Both descriptive and predictive techniques were utilized to analyze the dataset in alignment with the project objectives. In order to prepare the dataset for analysis, unnecessary columns such as 'Row ID', 'Ship Date', 'Ship Mode', 'Customer Name', 'Country', and 'Postal Code' were eliminated as they did not provide any significant information for the analysis. The remaining columns were also checked for null values, and it was confirmed that none were present. Additionally, duplicated observations were identified and removed to ensure the accuracy of the analysis. The data cleaning efforts resulted in a more concise and purposeful dataset, which improved the effectiveness of the analysis while reducing its memory usage.

Additionally, a new column called "profit margin" was introduced to facilitate the analysis process. The original dataset included a "profit" column, which was in dollars and made it challenging to make meaningful comparisons across various products. By introducing the "profit margin" column, expressed as a percentage, the profits were standardized, allowing for more insightful comparisons.

8.3 Collaborator/ End Users (Asking Questions)

The targeted end users for this project be any business owners that are involved in online retailers or physical stores.

The potential questions that we can answer for the end users from EDA will be as follows: -

1. Which products/categories are the most profitable?
   (Can be concluded from section ***8.4.2*** under EDA)

2. Which regions/markets are performing the best/worst?
   (Can be concluded from section ***8.4.3*** under EDA)

3. Are there any specific customer segments that we should focus on?
   (Can be concluded from section ***8.4.6*** under EDA)

4. How can we reduce losses and improve margins using machine learning?
   (Can be concluded from the second part of this report)

From the perspective of business owners, they will be interested in obtaining answers to the aforementioned questions as they directly impact the profitability and sustainability of their business. Knowing which products or categories generate the highest profits helps businesses focus their marketing efforts on these lucrative areas. Understanding the performance of different regions/markets allows for effective resource allocation and expansion of the customer base in promising areas. Furthermore, identifying specific customer segments enables businesses to tailor their products and services to meet customer needs, resulting in increased customer satisfaction and loyalty. Finally, the implementation of machine learning in data analysis enables business owners to build predictive models for forecasting business growth, particularly in terms of profit margins. By addressing those questions, end users can enhance their competitive advantage, drive revenue growth, and ensure long-term success in their respective markets.

EDA (Exploratory Data Analysis) plays a key role in answering questions about profitability, market performance, and customer segments. Through data visualization techniques such as bar charts and pie charts, we can identify the most profitable products and categories, assess the performance of different regions or markets, and pinpoint the customer segments with the highest potential. By leveraging these insights, business owners can focus their marketing efforts, optimize region resource allocation, boost sales in profitable regions, and tailor strategies to retain and attract valuable customer segments. EDA empowers businesses to make data-driven decisions, optimize profitability, drive growth, and gain a competitive edge in the market.

8.4 Exploratory Data Analysis (EDA)

8.4.1 Histogram graph to check the distribution of each feature.

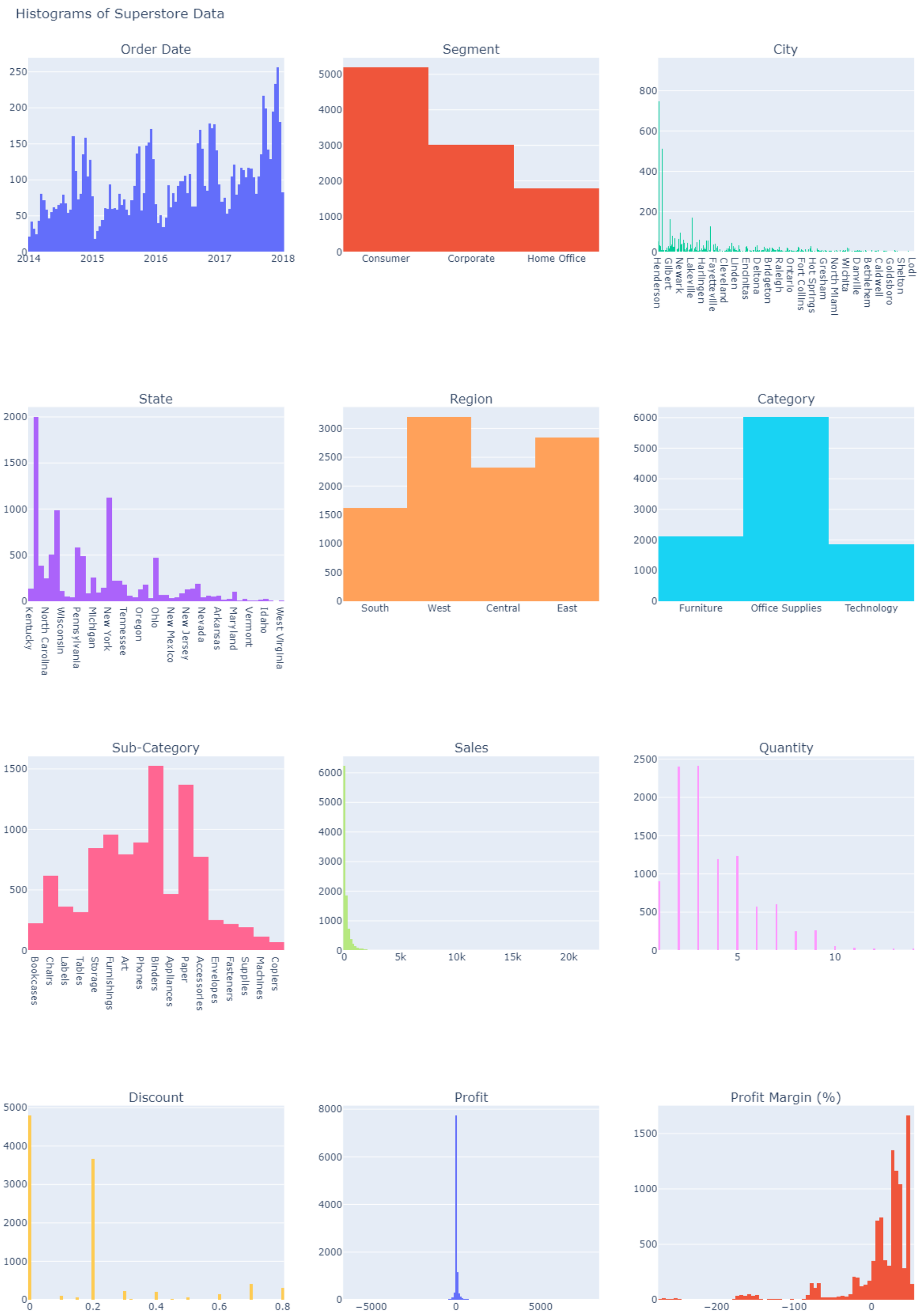


Figure 9.1: Histogram on Superstore data.

Findings:

- From the histogram, the distribution of each feature could be examined.
- For Segment, it can be observed that the highest frequency of values for this variable is found in the consumer segment, followed by corporate and home office.
- For City, Henderson is having the highest frequency of values for this variable.
- For State, Kentucky is having the highest frequency and followed by New York.
- For Region, West Region is having the highest frequency of values, followed by East, Central and South region.
- For Category, office supplies are having highest frequency of values, followed by furniture and technology.
- For Sub-Category, binders are having the highest frequency of value and followed by accessories.
- For Sales, the distribution is right-skewed, where sales that approaching 0 is having the highest frequency.
- For Quantity, the distribution is right-skewed, where quantity that is in between of 0 and 5 is having the highest frequency.
- For Discount, 0 is having the highest frequency of value and followed by 0.2.
- For Profit, profit that is approaching 0 is having the highest frequency of value.
- For Profit Margin (%), profit margin (%) in the positive region is having higher frequency in compared with the negative region.

8.4.2 Identify the sales based on category in bar chart.

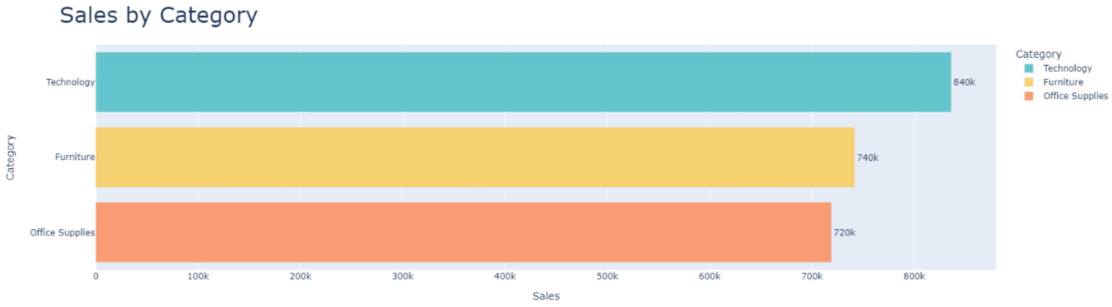


Figure 9.2: Sales by Category.

Findings:

From the bar chart, it can be observed that the technology products are the having the highest sales, accounting for 840,000, followed by office supplies at 740,000 and furniture products (720,000) is having the least sales.

8.4.3 Identify the profit based on category in bar chart.

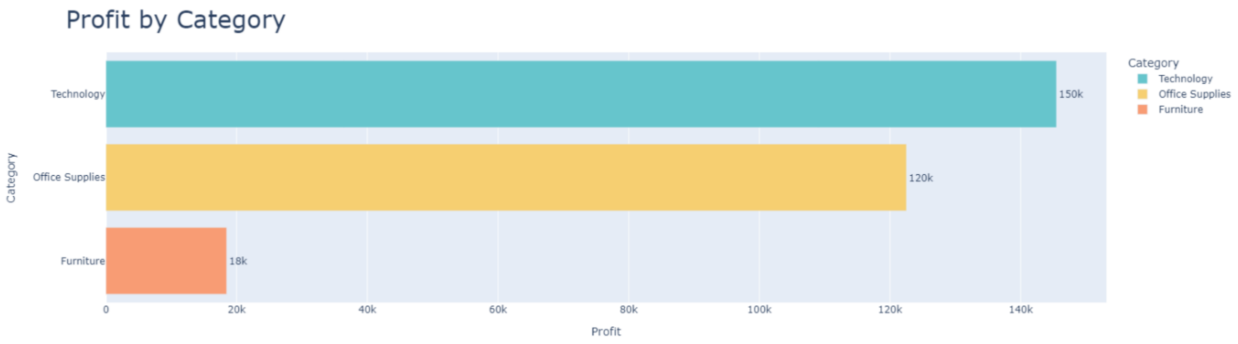


Figure 9.3: Profit by Category.

Findings:

From the bar chart, it can be observed that the technology products are the having the highest profit, accounting for 150,000, followed by office supplies at 120,000 and furniture products (18,000) is the least profitable. By using this profit by category bar chart, it can be revealed that Technology products are the most profitable category in compared with products from other categories.

8.4.4 Identify the profit based on region in bar chart.

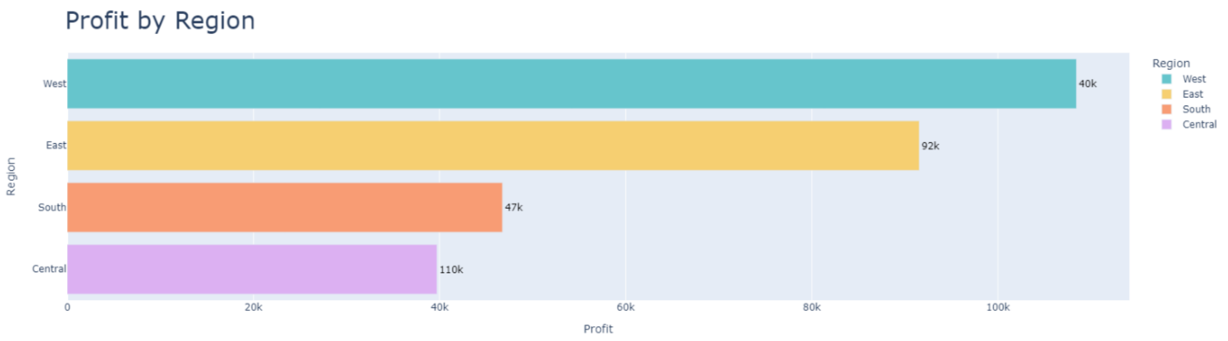


Figure 9.4: Profit by Region.

Findings:

From the bar chart, it can be observed that the West Region is the most profitable accounting for 110,000, followed by East Region and South Region at 92,000 and 47,000 respectively, whereas Central region is the least profitable by only accounting for 40,000. Therefore, we can conclude that West Region performed the best whereas Central region performed the worst.

8.4.5 Distribution of (category feature) and Profit/Sales in cluster column chart.

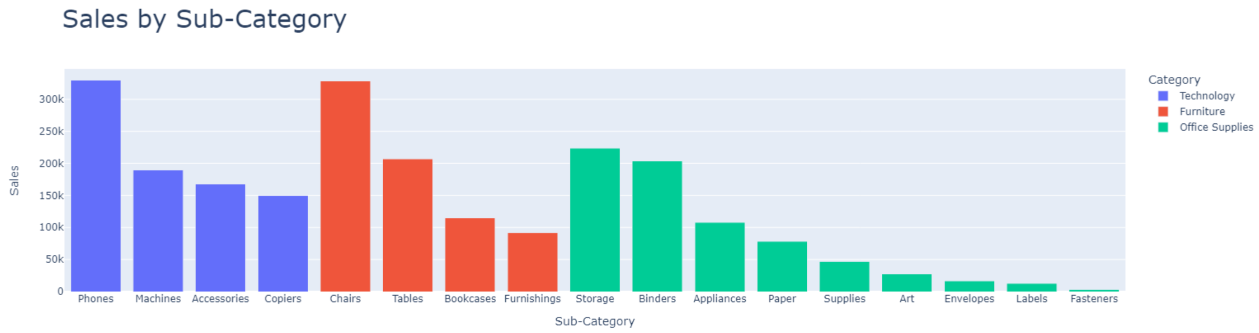


Figure 9.5: Sales by Sub-Category.

Findings:

- By observing the cluster column chart, it becomes evident that technology products led in sales, followed by furniture and office supplies
- For the sales in technology category, phones ranked at first, followed by machines, accessories and copiers.
- For the sales in furniture category, chairs are having the most sales, tables ranked second and followed by bookcase, whereas furnishings ranked last.
- For the sales in office supplies, storage ranked first, followed by binders, appliances, paper, supplies, art, envelopes, labels and fasteners.

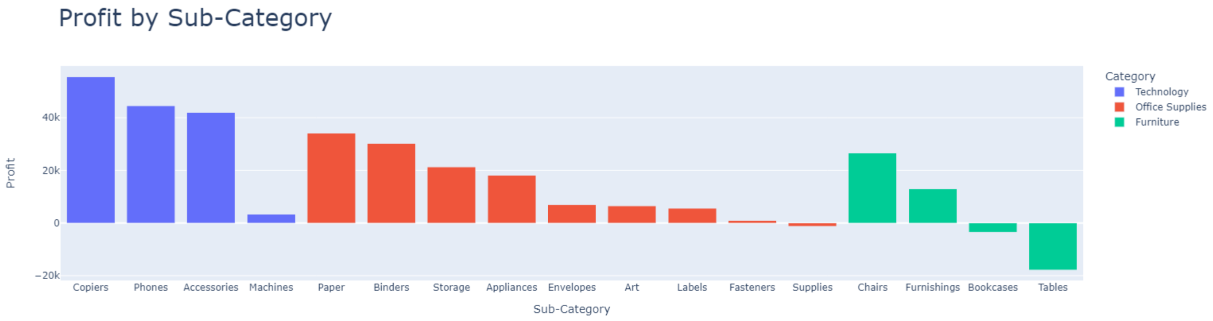


Figure 9.6: Profit by Sub-Category.

Findings:

- By observing the cluster column chart, it becomes evident that technology products led in profit, followed by office supplies and furniture.
- For the profit in technology products, copiers are the most profitable, followed by phones, accessories and machines is the least profitable.
- For office supplies and furniture category, the products can be categorised into two categories: profitable and loss.

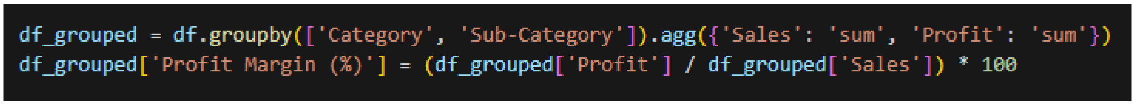
- For office supplies, paper is the most profitable, followed by binders, storage, appliances, envelopes, art, labels, fasteners and supplies. Out of these products, supplies are making loss whereas all other products are profitable.
- For furniture, chairs are the most profitable, followed by furnishings, bookcases and tables. Out of these products, bookcases and tables are making loss, whereas chairs and furnishings are profitable.
- In a nutshell, it can be revealed that Copiers in the Technology category were the most profitable products.

8.4.6 New column “Profit Margin (%)” has been created to identify the profit margin in category features (“Category”, “Sub-category)

Profit Margin can be defined as:

$$Pr\,o\,fit\ M\arg i\,n\ (\%) = \frac{Pr\,o\,fit}{Sales}$$

Therefore, new column “Profit Margin (%)” is created using the following codes:



```
df_grouped = df.groupby(['Category', 'Sub-Category']).agg({'Sales': 'sum', 'Profit': 'sum'})
df_grouped['Profit Margin (%)'] = (df_grouped['Profit'] / df_grouped['Sales']) * 100
```

Figure 9.7: Code to create new column “Profit Margin (%)

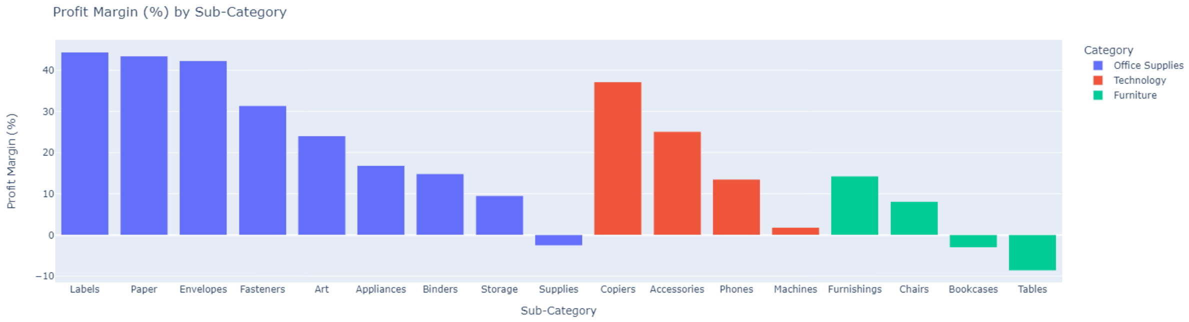


Figure 9.8: Profit Margin by Sub-Category.

Findings:

- From the cluster column chart, it can be observed that the office supplies are leading in profit margin, followed by technology products and furniture.
- For office supplies, labels are generating the highest profit margin, followed by paper, envelopes, fasteners, art, appliances, binders, storage and supplies. Supplies is the only products in the office supplies category that undergo losses.
- For technology product, all the products are having positive profit margin. Copiers is having the highest profit margin, followed by accessories, phones and machines.
- For furniture, furnishings are having the highest profit margin, follow by chairs, bookcases and tables. Furnishings and chairs are having the positive profit margin while bookcases and tables are having negative profit margin.

8.4.7 Identify the distribution of profit margin on customer segment in pie chart.

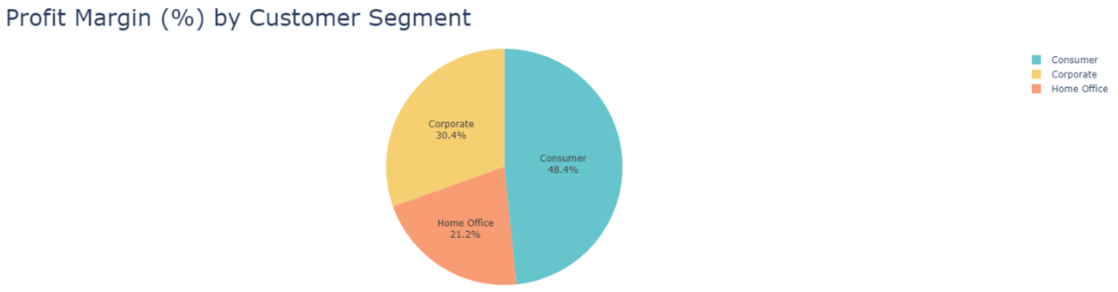


Figure 9.9: Profit Margin by Customer Segment.

Findings:

- From pie chart, it can be noticed the consumer segment is generating the highest profit margin, accounting for 48.4%, followed by corporate at 30.4% and home office (21.2%).
- Therefore, by visualising the profit margin by customer segment, it can be revealed that consumers segment is the most preferable segment in compared with the other two customer segments.

9 Conclusion:

In this project, the CRISP-DM methodology was implemented, specifically focusing on the exploratory data analysis (EDA) phase. The analysis yielded significant insights. Notably, it was found that copiers categorized under Technology were the most profitable items. While the West region exhibited strong performance, the Central region requires attention to enhance sales. Furthermore, the consumer segment emerged as the most lucrative, making a substantial contribution to overall revenue. Three out of four research questions were successfully addressed, with the remaining question concerning the development of a predictive model, which will be pursued in the subsequent project phase.

References & Appendixes

Organisation for Economic Cooperation and Development (2018). Implications of E-commerce for Competition Policy. https://www.oecd.org/daf/competition/e-commerce-implications-for-competition-policy.htm

United Nations Conference on Trade and Development (2021). How Covid-19 triggered the digital and e-commerce turning point. https://unctad.org/news/how-covid-19-triggered-digital-and-e-commerce-turning-point

Su, E. (2020, August 19). Data Science Innovation in Supermarkets: A Case Study of Woolworth. Medium. https://medium.com/@erransu/data-science-innovation-in-supermarkets-a-case-study-of-woolworth-f1756d9ac956

Xia, L. & Lv, X. (2021). Problems and Countermeasures Existing in E-Commerce Enterprise Network Marketing under the Background of Big Data. *Mathematical Problems in Engineering Volume 2021, Article ID 4786318, 8 pages*. https://doi.org/10.1155/2021/4786318

Tim S. (2019, August 26). The Advantages of Data-Driven Decision Making. https://online.hbs.edu/blog/post/data-driven-decision-making

Prismaretail.ai. (n.d.). Data Science in Supermarkets: Everything You Need to Know. Prismaretail.ai. Retrieved April 28, 2023, from https://prismaretail.ai/data-science-in-supermarkets.html

Vivek Kumar. (2020). Superstore Dataset Final. Kaggle. https://www.kaggle.com/vivek468/superstore-dataset-final