

**DIPLOMA IN COMPUTER SCIENCE**

**COURSEWORK**

**SUBJECT** **:**  **BIG DATA ANALYTICS**

**SUBJECT CODE** **:**  **DDS1313**

**DEADLINE** **:**  **12/8/2023**

|  |  |
| --- | --- |
| **NAME** | **STUDENT ID** |
| Albert Pang Kian Cheng | 22040943 |
| Chin Yung Xuan | 22057590 |
| Eng Zheng Yu | 22070544 |
| Tye Khan Wei | 22063051 |
| Tan Yann Bin | 22071674 |

# Grading Rubric

**Part 1:@25%**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Criteria | Competency Level & Marks | | | | |
| Poor | Below Average | Average | Above Average | Excellent |
| Phase 1 | Very limited background study  State the domain and business only with no reference to the information gained  (1-3 mark) | Limited background study  Demonstrate a background study in brief with no reference to the information gained  (4-5 marks) | Adequate background study  Demonstrate a background study of the domain and businesses with no reference to the information gained  (6-7 marks) | Strong background study  Demonstrate a background study of the domain and businesses Provide any references related to the information gained  (8-9 marks) | Outstanding background study  Demonstrate an extended background study of the domain and businesses Provide more than 1 references such as yearly report, white papers etc. |
|  |  |  |  |  | (10 marks) |
| Phase 2 \*individual parts | Data is loaded with no cleaning steps mentioned Frequent errors in code with no proper comments explaining the code  (1-3 marks) | Data is cleaned  No description on the process  No EDA  Some errors in code with limited comments explaining the code  (4-6 marks) | Data is cleaned and some steps are shown At least 1 EDAs  Few errors in code with adequate comments explaining the code  (7-8 marks) | Data is cleaned Walkthrough of the processes are described At least 2 EDAs  Very few errors in code with a lot of comments explaining the code  (9 marks) | Data is cleaned thoroughly Walkthrough of the processes are well described More than 3 EDAs  No errors in code with a lot of comments explaining the code.  Code is properly indented  and properly sequenced |
|  |  |  |  |  | (10 marks) |
|  | Marks |
| Albert Pang Kian Cheng |  |
| Chin Yung Xuan |  |
| Eng Zheng Yu |  |
| Tye Khan Wei |  |
| Tan Yann Bin |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Overall documentation, Formatting and References | No proper page numbers Errors in format across the document  inconsistent references A lot of errors in labelling/formatting  (1 marks) | Document have page number  Errors in format across the document inconsistent references A lot of errors in labelling/formatting  (2 marks) | Document have page number, table of content and figures Predominantly consistent format across the document inconsistent references Adequate errors in labelling/formatting  (3 marks) | Document have page number, table of content and figures, and properly labelled tables and figures  Consistent format across the document  Proper and consistent references  Few errors in  labelling/formatting (4 marks) | Document have page number, table of content and figures, and properly labelled tables and figures  Consistent format across the document  Proper and consistent references  (5 marks) |

# Part 2:@45%

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Criteria | Competency Level & Marks | | | | |
| Poor | Below Average | Average | Above Average | Excellent |
| Phase 3 | No explanation on the model chosen  No feature selection process explanation  Not understanding on the whole process of model planning  (1-3 marks) | Provide some explanation on the model chosen  Least feature selection process explanation Demonstrate a little understanding on the whole process of model planning  (4-5 marks) | Provide adequate explanation on the model chosen  Feature selection process are explained Demonstrate adequate understanding on the whole process of model planning  (6-7 marks) | Provide good explanation on the model chosen Feature selection process are well explained Demonstrate understanding on the whole process of model planning  (8-9 marks) | Complete and concise explanation on the model chosen  Feature selection process are well explained  Demonstrate high level of understanding on the whole process of model planning (10 marks) |
|  | Frequent errors in code with no | Some errors in code with | Few errors in code with | Very few errors in code | No errors in code with a |
|  | proper comments explaining | limited comments | adequate comments | with a lot of comments | lot of comments |
|  | the code | explaining the code | explaining the code | explaining the code | explaining the code |
| Phase 4 | (1-3 marks) | (4-5 marks) | (6-7 marks) | (8-9 marks) | Code are properly |
|  |  |  |  |  | indented and properly |
|  |  |  |  |  | sequenced |
|  |  |  |  |  | (10 marks) |
| Phase 5 | No screenshots of results based | Includes few screenshots | Includes some | Complete screenshots of | Complete screenshots of |
|  | on the algorithm of choice, | of results based on the | screenshots of results | results based on the | results based on the |
|  | visualization/graph | algorithm of choice | based on the algorithm of | algorithm of choice | algorithm of choice |
|  | No hypothesis mentioned | Includes | choice | Includes | Includes |
|  | Limited analysis explanation | visualization/graph | Includes | visualization/graph | visualization/graph |
|  | and relate to the hypothesis | Limited research to | visualization/graph | Evidence of research to | Evidence of extended |
|  | Demonstrate no understanding | justify hypothesis | Evidence of research to | justify results | research to justify results |
|  | on the analysis | Limited analysis | justify results | Strong analysis | Demonstrate clear and |
|  | (1-3 marks) | explanation and relate to | Strong analysis | explanation and relate to | outstanding understanding |
|  |  | the hypothesis | explanation and relate to | the hypothesis | on the analysis and |
|  |  | Demonstrate poor | the hypothesis | Demonstrate proper | hypothesis is/are met |
|  |  | understanding on the | Demonstrate partially | understanding with minor | (10 marks) |
|  |  | analysis | understanding on the | gaps on the analysis |  |
|  |  | (4-5 marks) | analysis | (8-9 marks) |  |
|  |  |  | (6-7 marks) |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Phase 6  \*individual parts | Summary of the analysis work (1 -3marks) | Summary of the analysis work  Address the difficulties facing during the analysis (4-5 marks) | Summary of the analysis work  Address the difficulties facing during the analysis Provide at least 1 improvement to the project  (6-7 marks) | Proper summary of the analysis work  Address the difficulties facing during the analysis Provide at least 1 improvement to the project  Provide at least 1 suggestion for future works that can be done (8-9 marks) | Complete and concise summary of the analysis work  Address the difficulties facing during the analysis Provide more than 1 improvements to the project  Provide more than 1 suggestion for future works that can be done  (10 marks) |
|  | | | |  | Marks |
| Albert Pang Kian Cheng |  |
| Chin Yung Xuan |  |
| Eng Zheng Yu |  |
| Tye Khan Wei |  |
| Tan Yann Bin |  |
| Overall documentation, Formatting and References | No proper page numbers Errors in format across the document  inconsistent references A lot of errors in labelling/formatting  (1 marks) | Document have page number  Errors in format across the document inconsistent references A lot of errors in labelling/formatting  (2 marks) | Document have page number, table of content and figures Predominantly consistent format across the document  inconsistent references Adequate errors in labelling/formatting  (3 marks) | Document have page number, table of content and figures, and properly labelled tables and figures Consistent format across the document  Proper and consistent references  Few errors in labelling/formatting  (4 marks) | Document have page number, table of content and figures, and properly labelled tables and figures Consistent format across the document  Proper and consistent references  (5 marks) |

Contents

1. Introduction
   1. Background----------------------------------------------------------------------------------10-12
   2. Problem Statement----------------------------------------------------------------------------13
   3. Objectives------------------------------------------------------------------------------------------14
2. Work Done
   1. Extract, Transform, Load (ETL) and Data Understanding---------15-18
   2. Exploratory Data Analysis (EDA)-----------------------------------------------18-31
   3. Data Preparation and Modelling (Predictive Analytics) --------31-37
   4. Evaluation------------------------------------------------------------------------------------37-38
   5. Deployment----------------------------------------------------------------------------------39-44
3. Conclusions and Future Work
   1. Primary Findings-------------------------------------------------------------------------------45
   2. Strengths and Weaknesses of this Project-----------------------------------------46
4. References----------------------------------------------------------------------------------------------------47
5. **Introduction**
   1. Background

* History of Walmart

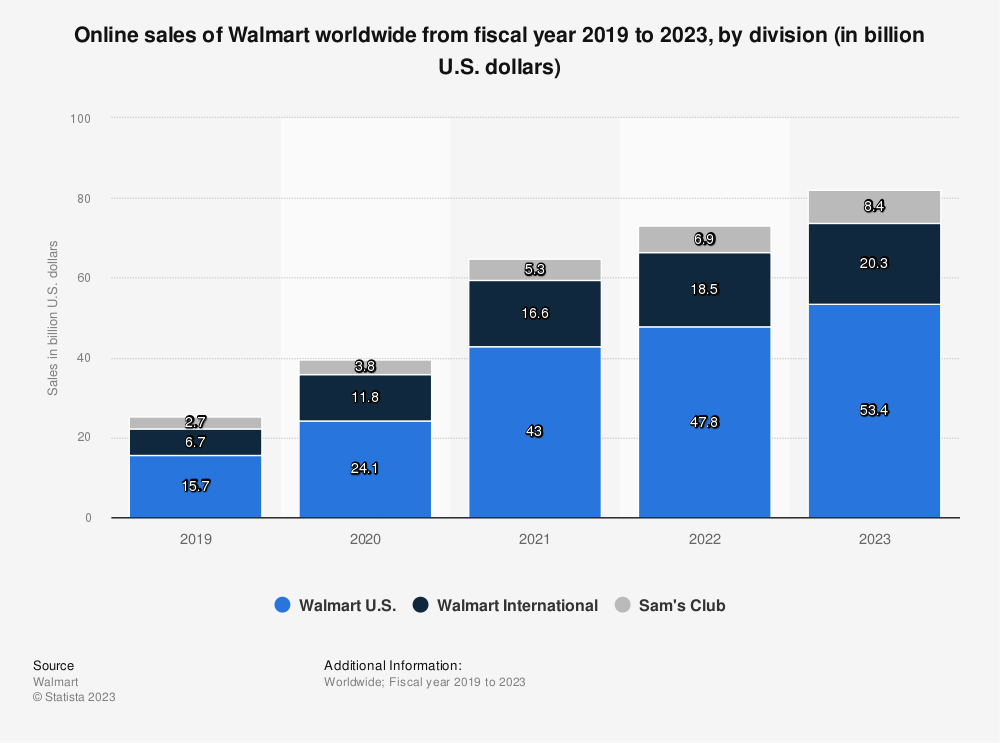
Walmart is an American multinational retail corporation that operates a chain of hypermarkets, discount department stores, and grocery stores in the United States, headquartered in Bentonville, Arkansas (Wikipedia, 1962). The company was founded by the Sam and James “Bud” Walton brothers nearby Rogers, Arkansas during 1962. They also own and operate Sam’s Club retail warehouses. They were inspired by the vision of offering affordable prices and excellent customer service through a discount retail store. Therefore, the beginning of Walmart’s rapid expansion in the United States is marked throughout the 1970s. Walmart stores are established in small towns and rural area across the United States to begin with. In the 1980s, within a decade of opening the combination of grocery and merchandise Supercentres, Walmart became one of the largest grocers in the United States. Walmart introduced its warehouse store concept with Sam's Club and went public, becoming a major player in the retail industry. Then, during the 1990s, the company internationally expand into countries like Canada, Mexico, and the UK, along with the "Everyday Low Prices" philosophy. Despite the low-pay employees’ controversies in the early 2000s, Walmart's growth persisted, and it focused on enhancing its e-commerce presence in response to changing consumer habits. They acquire numerous e-commerce business, including Jet.com and Moosejaw (Tikkanen, 2023). Through the 2010s, Walmart continued adapting to evolving markets while emphasizing sustainability and social responsibility. Walmart has over 10,000 stores and clubs in 24 countries as of October 31, 2022, with under 46 distinct names. For example, in the United States and Canada, the business is known as Walmart, whereas in Mexico and Central America, it is known as Walmart de México y Centroamérica, and in India, it is known as Flipkart Wholesale. Additionally, it also operates entirely on its own in South Africa, Canada, and Chile.

* Walmart’s Competitors

Walmart faces competition from a variety of retailers and companies operating in different segments of the market. Some of its main competitors include Amazon, Target, and Costco. Amazon is one of the biggest competitors to Walmart in the realm of e-commerce. With its vast online marketplace, fast shipping options, and extensive product range, Amazon has challenged Walmart's traditional retail dominance. As for Target, it is a major competitor with a similar retail model, offering a combination of affordable products and a wide range of goods. It often targets a slightly more upscale and trend-conscious demographic compared to Walmart. Furthermore, Costco is a competitor to Walmart’s warehouse store concept (Sam’s Club), that operates on a membership-based model. It offers bulk goods at discounted prices and focuses on appealing to businesses and families looking to buy in larger quantities.

* E-commerce

In this modern digital era, e-commerce is a part of people’s lives that cannot be taken away. Using the convenience of the Internet, many people choose to shop using e-commerce platforms rather than physical stores. Therefore, Walmart also started to shift their focus towards e-commerce business during the year 2020 (McKinnon, 2023). Walmart operates an extensive online marketplace where customers can purchase a wide range of products, including electronics, clothing, household goods, and more. In an effort to bolster its e-commerce capabilities, Walmart acquired several online retailers, including Jet.com in 2016. The acquisition of Jet.com brought expertise in e-commerce and technology to Walmart and helped accelerate its online growth. Besides, Walmart also joined forces with Shopify in 2020 to open the Walmart Marketplace to more sellers to be involved. Furthermore, Walmart has invested heavily in its online grocery shopping service, allowing customers to order groceries online and have them delivered or prepared for pickup. This service has gained popularity, especially during the COVID-19 pandemic, as more customers sought contactless shopping options. Overall, Walmart's e-commerce strategy focuses on leveraging its existing retail strength, offering a variety of delivery and pickup options, and providing customers with a seamless shopping experience whether they choose to shop in physical stores or online. The company's e-commerce efforts are a crucial component of its overall growth and competitiveness in the evolving retail landscape. However, many organizations still lack the data analytics expertise needed to build predictive models that would aid in the continued success of enterprises. Businesses should acquire information about online consumers’ behaviours and data to improve the client experience. As a result, this might increase the e-commerce’s sales, making it more successful.



As seen in the graph above, Walmart's online sales have been on steady growth for its online sales for a few years.

* Predictive modelling

In this project, we will be analysing the e-commerce sales data of Walmart. There are two main objectives for this project. The first objective is to understand the purchase intention of the online customers, and the next objective is to predict the profitability of Walmart. We are required to create a predictive model to satisfy these two objectives. We thought that by applying predictive modelling, it would be possible to make right adjustments and significantly help Walmart’s e-commerce business by allowing them to make right business decisions.

* 1. Problem Statement

With the increasing popularity of online businesses, many established organizations are keen to expand their businesses using e-commerce platforms. However, many companies that handle their own e-commerce applications do not have sufficient knowledge on data analytics skills, causing them to not have a predictive model to help them gain insight about their e-commerce business. This will greatly affect their sales and ability to improve their business in the online sector.

Besides that, building customers’ trust towards your e-commerce platform is not easy, and it is also critical. E-commerce businesses must address concerns towards customers’ data security, privacy, and the potential for fraud to ensure customers feel confident making purchases and making transactions online. They also must include return-refund policy for all purchases customers make. Besides security and privacy, most customers also value high quality user experience and website performance. Therefore, providing a seamless and user-friendly online shopping experience is crucial. Slow-loading websites, confusing navigation, and technical glitches can lead to customers’ frustration and abandoned shopping carts.

Furthermore, due to the e-commerce landscape being highly competitive, with numerous competitors vying for customers’ attention and loyalty, to stand out in a crowded market and differentiating themselves from competitors is very challenging, especially for saturated products. Therefore, most e-commerce platforms will try to provide discounts to customers to attract them to purchase more. For example, they will give free shipping vouchers to customers when they make a purchase above RM15. This act will increase the sales, but at the same time, reduce the profit due to having to cut off the shipping fees for the customers.

* 1. Objectives

A machine learning model will be created in this project, to improve and provide support to our problem statement above.

This project is carried out to identify areas of improvement for the business. Whether it be geographic locations that have high potential for the company to be expanded into, or identifying product categories that can be improved on. The EDA can let the company gain insight onto customer behaviour that can be identified and used to expand the company’s business operations and to improve efficiency as well.

The objective of this machine learning model is to segment customers based on demographics, purchase history, and behaviours to target marketing efforts more effectively. This can lead to higher conversion rates and improved customer loyalty. Walmart can also understand their online shoppers’ intention better with this result. Walmart can also use the model created to predict the profitability for a given order and can take preventative measures against the orders with negative profitability.

Besides, the machine learning model also helps to differentiate the priority of each order. Therefore, Walmart can ship out the critically prioritized order first, to maintain efficient deliveries and obtain good customer reviews. We can use the above qualities to predict the profitability of Walmart in several aspects. The model can also be used to analyse any existing and ongoing trend when it comes to sales in a certain region, whether it be a continent, country or state.

1. **Work Done**

In this report, we will be dealing with Descriptive and Predictive Analysis, with are concerned with extracting descriptive statistics from the dataset and using the data to build a machine learning model. To get a deeper understanding of our data, we will also touch on the “Why?” behind some of the statistics and observations we make throughout this report. However, we will not be using any sophisticated Diagnostic Analytics approaches. In addition, we will provide guidance and recommendations based on the insights we learn during this research to assist our chosen organization in leveraging the data of their online users.

* 1. **Extract, Transform, Load (ETL) and Data Understanding**

A screenshot of a computer

Description automatically generated

Firstly, we will be conducting ETL process, which is Extract, Transform and Load. This is a process whereby we take the dataset and transform it from a messy and raw dataset into a clean and classified dataset. It will then be more of a reliable and usable resource for us to load them into the machine learning model for analysis.

We will be using Databricks, which is an Apache Spark Technology as our ETL tools for this project. We will mainly be using Scala programming language for accomplishing the objectives of this project. The first step is to import the dataset in CSV format into the data frame of Databricks to allow the machine to forecast the purchase information of customers based on their shipping mode and city. We named the data frame ‘shopperDF’ and display it to get a clearer and more accessible tables.

A screenshot of a computer code

Description automatically generated

Next, we will display the schema of the dataset in a tree format alongside with all the column names and the data types. There are a total of 51290 records in the dataset, with a total of 23 attributes. These attributes are divided into categorical and numerical columns. Therefore, there are 14 categorical columns and 9 numerical columns. The columns we have in this dataset are “RowID”, “OrderID”, “OrderDate”, “ShipDate”, “ShipMode”, “CustomerID”, “CustomerName”, “Segment”, “City”, “State”, “Country”, “PostalCode”, “Market”, “Region”, “ProductID”, Category”, “Sub-Category”, “Sales”, “Quantity”, “Discount”, “Profit”, “ShippingCost”, and “OrderPriority”.

When the online shoppers make a purchase or order, the values of these columns will be updated in a real-time basis.

|  |  |  |
| --- | --- | --- |
| Columns | Description | Data type |
| RowID | The row number of every order. | Integer |
| OrderID | The unique code given to every order made by online shoppers. | String |
| OrderDate | The date when the online shoppers made the purchase. | Date |
| ShipDate | The date when Walmart ships out the online shopper’s orders. | Date |
| CustomerID | The unique code given to every online shopper. | String |
| CustomerName | Online shoppers’ names. | String |
| Segment | One shopper type. | String |
| City | City of order shipment. | String |
| State | State of order shipment. | String |
| Country | Country of order shipment. | String |
| PostalCode | Postal code of the address of the order shipment. | String |
| Market | General area of the order shipment. | Integer |
| Region | General region of the order shipment. | String |
| ProductID | The unique code given to every product. | String |
| Category | Categories of products. | String |
| Sub-Category | Sub-categories of the products. | String |
| Sales | The total sales of that order. | Double |
| Quantity | Item sold in that order. | Integer |
| Discount | Discount percentage of that order. | Double |
| Profit | Profit made by Walmart from that order. | Double |
| ShippingCost | Shipping fees needed to be paid by the buyers. | Double |
| OrderPriority | Which shipment is more urgent than the others. | String |

|  |  |
| --- | --- |
| Data Type | Function |
| Integer | It stores whole numbers that are either positive or negative values. |
| Double | It stores floating-point numbers that can hold up to 15 significant digits which can be fractional, decimal values, positive and negative infinity, and whole number. |
| String | It stores a sequence of elements and characters such as space, alphabets, and numbers. |
| Date | It shows the date with the format of DD/MM/YYYY. |

* 1. **Exploratory Data Analysis (EDA)**

In this section, we will be answering business questions by performing EDA. We will primarily use SQL programming language for this section, while still using the Apache Spark analytics engine.

A close up of a text

Description automatically generated

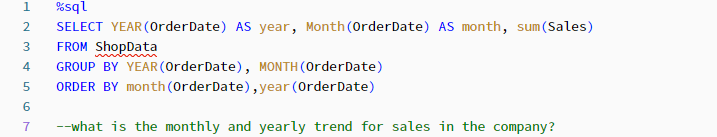
We created a temporary view of the data frame object to answer business questions. The data frame object will allow us to use the spark SQL statements.

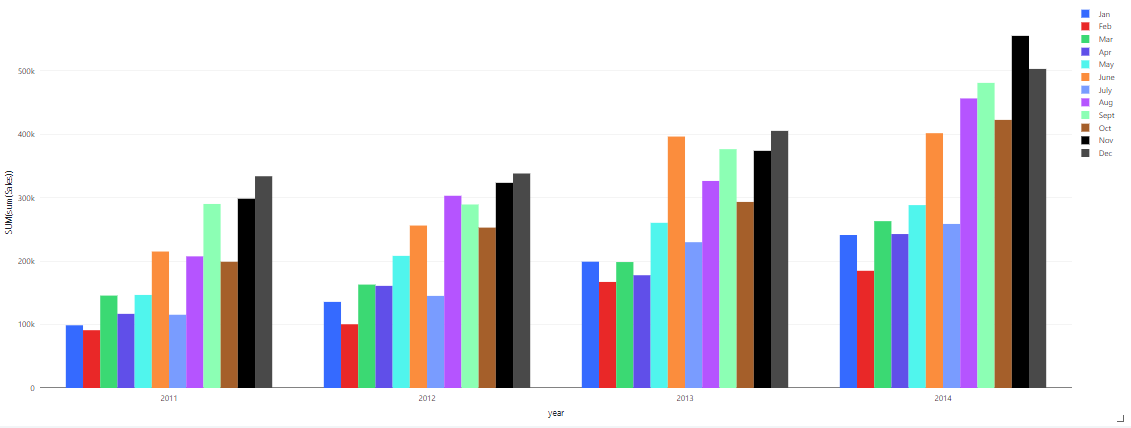
**Member 1 – Albert Pang Kian Cheng**

**Business Question 1**

What is the monthly and yearly trend, if any, for the sales in the company? Are there any periods of time where there are consistently high and low sales?

We will first query the SQL to have all the sales data, then arrange the sales by month and year. Then the data generated we will use to create a bar graph. The x column will have year data, while y columns is the sum of sales. For the x axis we will group by month so that all the months each year are arranged from January to December.



  
From the graph, we can see that there is a slight increase for total sales from the year 2011 to 2014. Comparing the same month across all years, there is an increase for the sales per month.

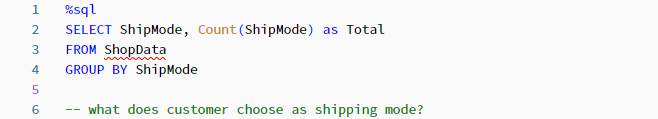
There can be a trend seen for the same year as well, where most of the sales come from the last few months of the year such as November and December. While for months February and July, there is a pattern of lower sales for the month compared to previous and next months.

Therefore, the business can focus on preparing for high sale periods to ensure operation remains smooth throughout the peaks, and efforts can be made to attract more sales during the non-peak periods.

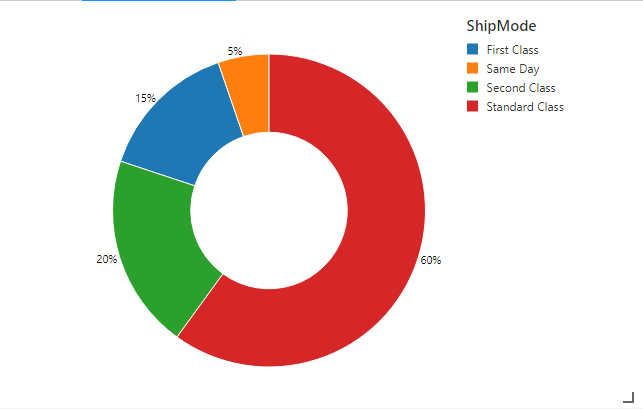
**Member 2 – Chin Yung Xuan**

Business Question 2

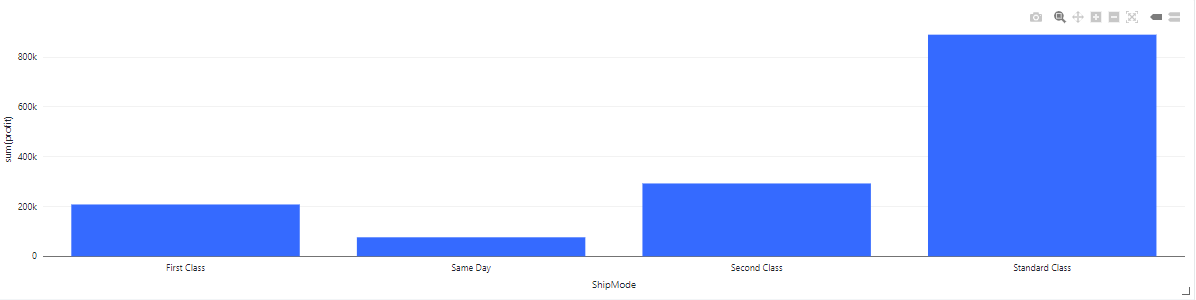
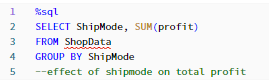
What type of shipping modes does customer choose? Why? Is there any effect on total profit and average profit?



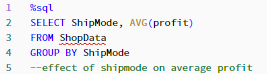
We query the SQL to show the different types of Shipping mode, as well as count the occurrence of each type of shipping mode to determine each of their percentages in the whole dataset. After counting each type of shipping mode, we group by their respective types to show the total. While the data is in tabular form, we visualise it using pie chart.

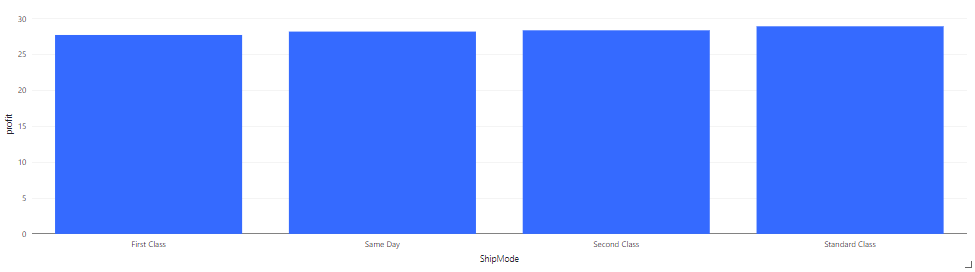


From the pie chart above, we can see that most orders are shipped with standard class at 60%, while second class constitutes 20%, then first class has a percentage of 15% and lastly same day shipping accounts for 5%. Standard class shipping modes are used by most of the online shoppers because the fees for this shipping method is more affordable. The delivery of these orders will take a significantly longer time; therefore, Walmart could not charge as high as other shipping modes. On the other hand, The Same Day delivery mode is only used by 5% of the online shoppers because it is pricier compared to others. This is understandable as the shipping company must arrange their order shipments using express transportations, allowing them to receive their orders on the same day.



From the bar graph it is seen that profits come mostly from standard class, followed by second class, first class and same day shipping modes. The total profit seems to correlate to the amount seen in the pie chart above, meaning that since most orders are using standard class shipping, most of the profit also comes from the same shipping type.



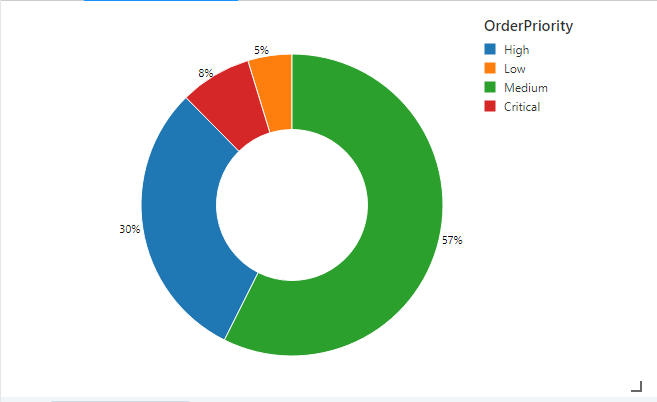
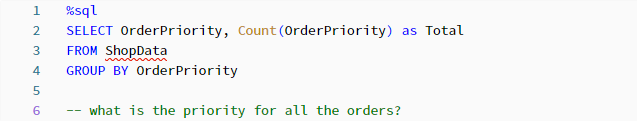


It seems that the shipping mode does not affect the average profitability for a given order, so we can estimate that an order’s shipping mode will not correlate to the profits in the order.

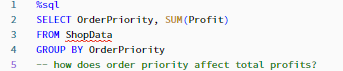
**Member 3 – Eng Zheng Yu**

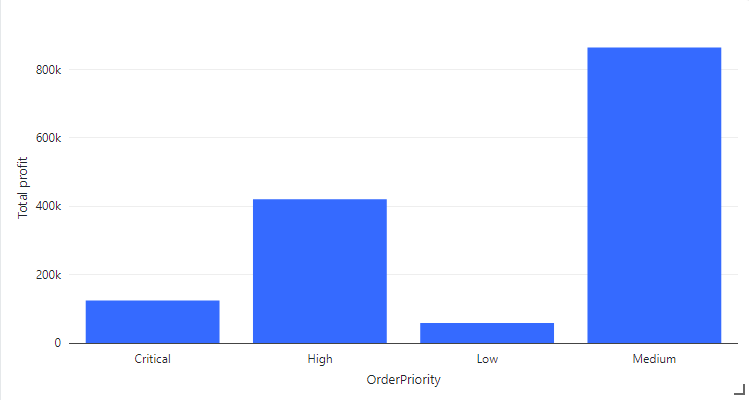
Business Question 3

What is the ratio for order priority? And does it affect average and total profits? Why?

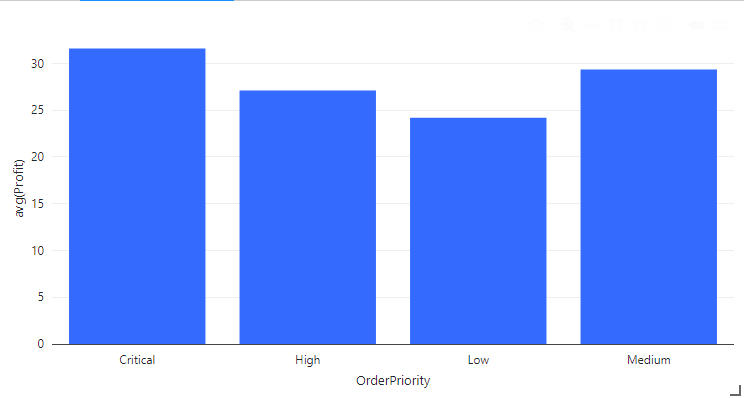
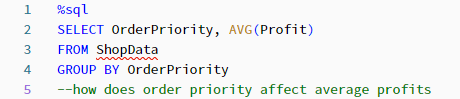


From the pie chart it is seen that most order are of medium priority, which is 57%. Then it is followed by high priority which stands at 30%, after that is critical at 8% and finally at 5% is low priority. We assume that since most customers are not in a rush to obtain the product, the order priority will be mostly medium, while only small percentage will have instructions to ship quickly for high priority and then critical priority. Whatever percentage is left will be low priority for those who don’t mind need the items soon or waiting for items to restock. For example, a person who preordered an out-of-stock product, or arranged for future delivery will be categorized as low priority.





From the bar chart it is seen that there is a similarity to the pie chart above. Since most orders are of medium priority, most of the company’s profits also come from medium priority orders and so on. Therefore, order priority is correlated with total profit.

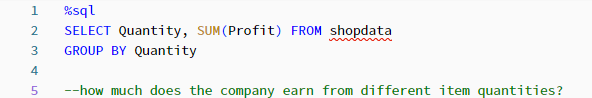


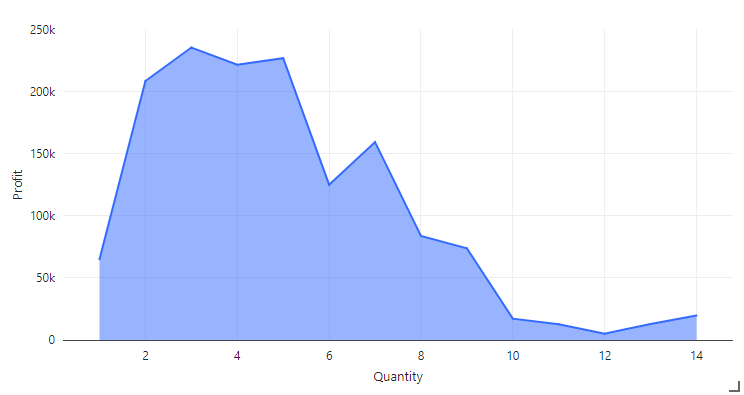
We can see that order priority does influence average profitability for a given order. On average, an order that is deemed critical will have higher profitability over the other types. This means that online shoppers are willing to pay more when their orders are critically needed. But the next highest profitability comes from order with medium priority, then followed by high and low priority orders. We assume that people who choose critical items will have high needs for the item, so such customers will buy more than required for their current needs to avoid the same problem occur again.

**Member 4 – Tan Yann Bin**

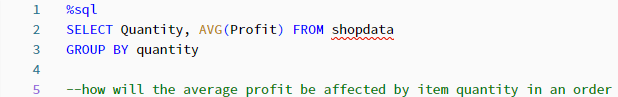
Business Question 4

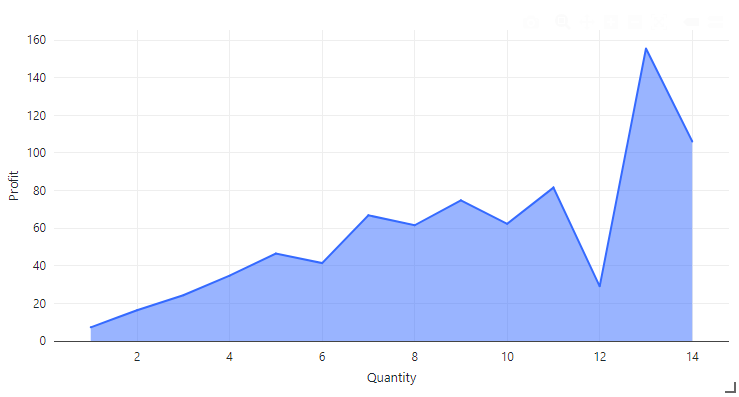
How many items quantity is in each order? And is there any effect on the total and average profits in the company?





From the line graph we can see that most of the profits come from orders with around 2 to 5 quantities, then falling off on either side. This is probably due to the high amount of order that have quantity around 2 to 5 items. Most customers will be the average person who just need a few units of the product and choose to buy a few of the products only, which would constitute a large sum for the profit values.



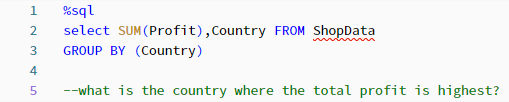


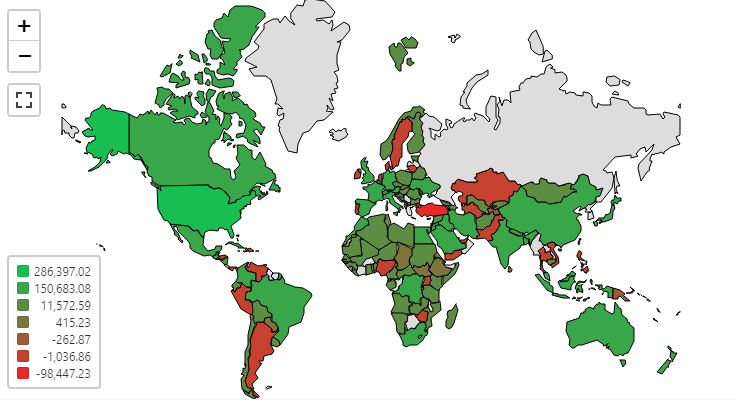
From the line graph, there is a trend that goes upward from left to right on the x axis. Since there are more items ordered for a given order, then the profit is multiplied for the given item, thus increasing the average profitability for a given order. For those who purchase lots of items, we presume is businesses who are fulfilling their business needs and don’t mind spending higher amount for a larger quantity of items.

**Member 5 – Tye Khan Wei**

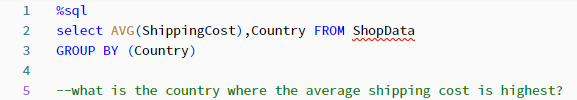
Business Question 5

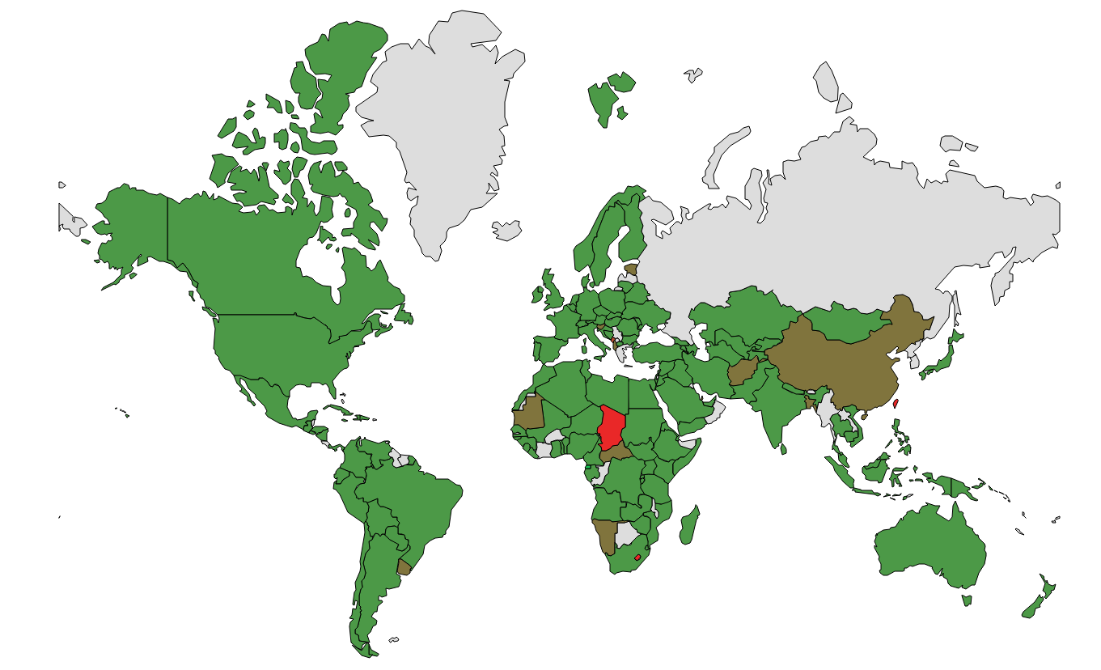
What are the regions where there are high profits for the company? Is there any region that causes high loss for the company? How about the shipping costs for each country?





From the map, most of the countries bring in profit for the company. There are a few regions that bring the highest profit to the company, namely USA, China, India, United Kingdom and France. There are regions that also bring the most losses , namely Turkey, Argentina, Sweden, Kazakhstan and Nigeria. We can assume that these countries mainly purchase mainly products that do not bring profit. But overall, most countries do return a profit, for the 147 countries in the dataset, 118 returns profit.



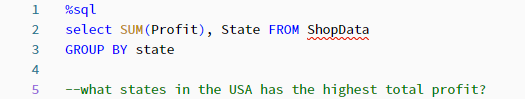


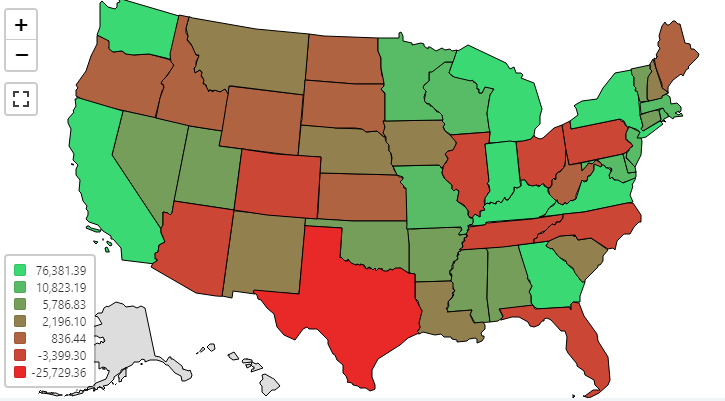
We can see that the red regions correlate towards higher shipping costs to the customer, there are some regions such as China, Taiwan and Chad that are the highest. This may be due to local regulations that call for stricter customs inspection, which may contribute to the final cost of shipping.

**Member 1 – Albert Pang Kian Cheng**

Business Question 6

What are the regions in the USA that bring profit to the company? Is there any that cause losses for the company?



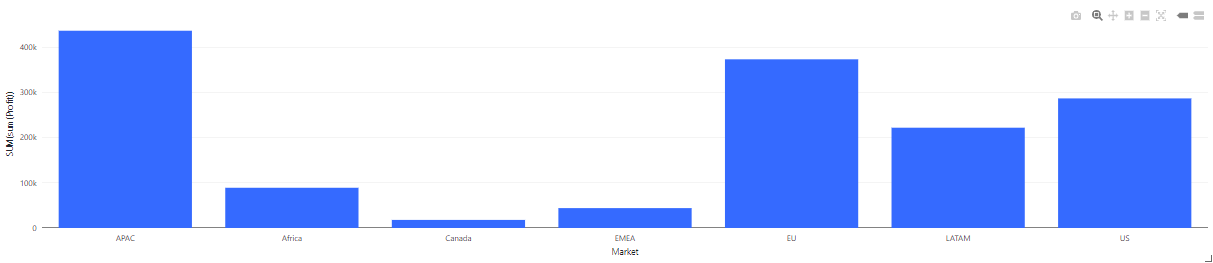
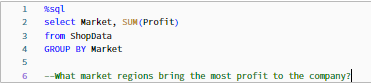


We can see from the map that most of the states in the US return a profit to the company. While the dataset did not include data from states of Hawaii and Alaska, out of the remaining 49 states that have data, 39 return profit. Some of the states with highest profit are California, New York and Washington. The states with highest sales seem to be of states where there is high population density, meaning states that are mostly of developed and are states with the big cities. While the states with highest losses are Texas, Ohio and Pennsylvania.

**Member 2 – Chin Yung Xuan**

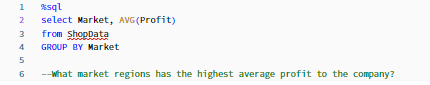
Business Question 7

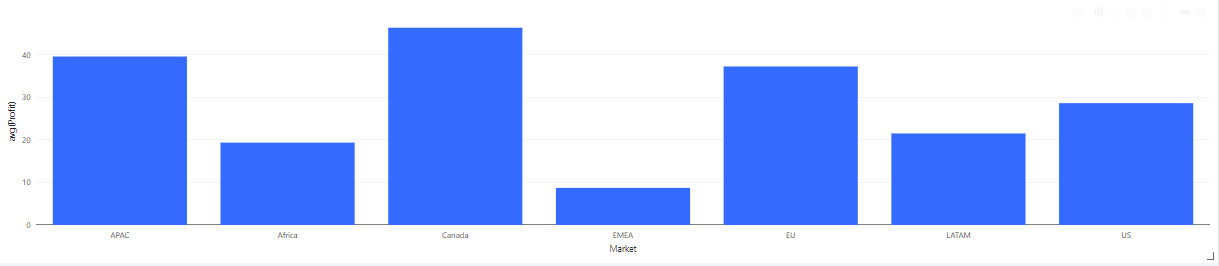
What are the regions that bring the highest and lowest profit to the company? How about the average transaction profit for each region?



We can see that the bar chart vividly illustrates that there are some high total profit Walmarts across various regions. For example, the US, Latin America and the European Nation. Meanwhile, the lowest sum of profit comes from the Canada region, emerging as the outlier. It is worth emphasizing that the presented bar chart encapsulates the cumulative impact of total sales, indicating that the number of transactions in each region plays a pivotal role in shaping the resultant profit figures.

Regions with high total profits might signify robust market demand, effective sales strategies, or even higher-priced products. On the other hand, regions with lower total profits could indicate untapped potential, competition challenges, or specific market preferences that might not align with the current product offerings. By scrutinizing factors such as average transaction value, customer retention rates, and product preferences, the company can unearth opportunities to optimize its approach for both high-profit and low-profit regions.





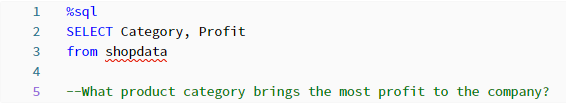
When we look at average order profitability for the regions, Canada comes out on top as the highest average profit, this means that while the average is high, the number of orders from Canada is not as high as the other markets, which leads to a low total profit. The company can consider expanding operations or offering more competitive options in Canada since there is high potential in the given market. This can lead to a more targeted marketing strategies to resonate well with the Canadian audience.

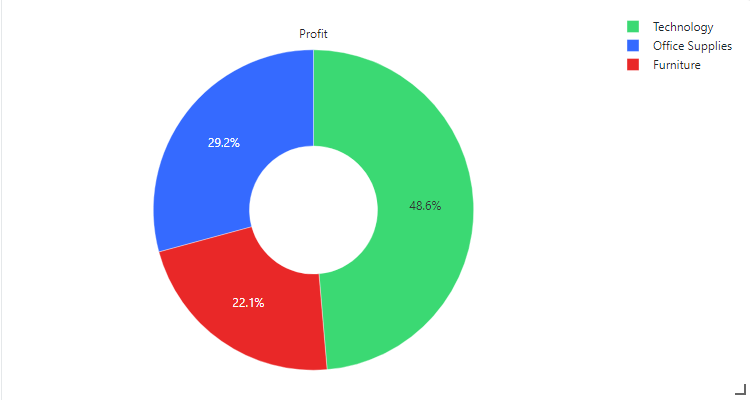
By embarking on a journey of expansion and targeted growth within the Canadian market, the company can transform high individual order profitability into a resounding success story marked by both average and total profitability.

**Member 3 – Eng Zheng Yu**

Business Question 8

What product category brings the most profit to the company?



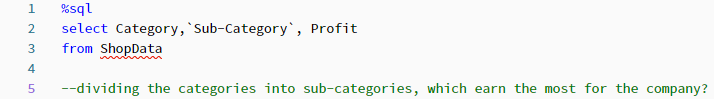


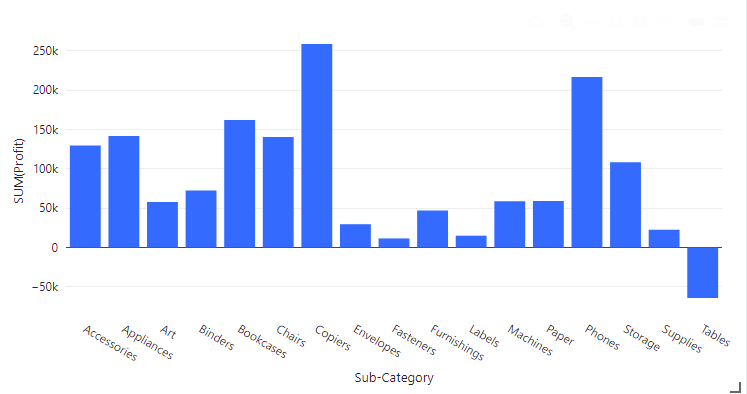
From the pie chart we can see that most of the profits come from the technology product category, which is at 48.6%, followed by office supplies category at 29.2% and lastly furniture category which is at 22.1%. We will explore the products that contribute to the profit of the company, it could be that there are more products in the technology category that bring up the average profits, or the office supplies and furniture categories have products that are not as fast-selling and contribute to lower sales.

**Member 4 – Tan Yann Bin**

Business Question 9

What is the sub-category that earns the company the highest profit? Is there any that cause losses to the company?



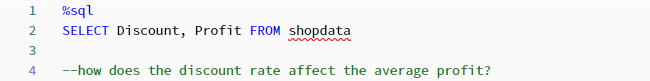


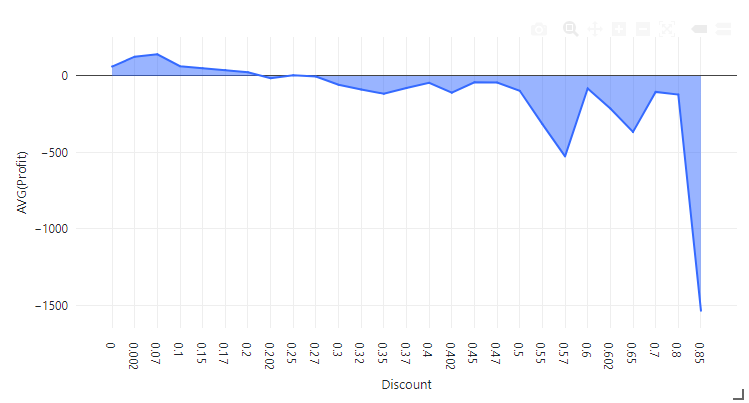
From the bar graph, we can see that most sub-categories bring in profit for the company, with the exception being tables that have negative profit for the company. We can find ways to minimize the losses coming from the sale of tables, we could revise our pricing for the products or finding suppliers that offer similar products for lower costs. From this graph we can see that quite a few sub-categories are from the technology, which brings up the total profit for the technology category. On the other end, table sub-category is the only one that has total loss, which means that it does bring down the total sales for the furniture category, which can explain why furniture category returned the least profit for the company.

**Member 5 – Tye Khan Wei**

Business Question 10

What is the correlation for discount rate to the profits of company?



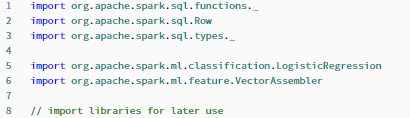


From the line graph we can see that for discount rates between 0.002 and 0.1 which are 2% to 10% percent discount, the average profit for those orders is still positive, and after 0.2 or 20% discount, most of the average profit is going to negative with 0.85 or 85% discount having the lowest profit at –1500. For the higher end of the x-axis, it makes sense that if the order has high discounts around 0.8 or 80%, the company is selling the items at a loss.

2.3 Data Preparation and Modelling (Predictive Analytics)

What is the predicted profit for a given order?

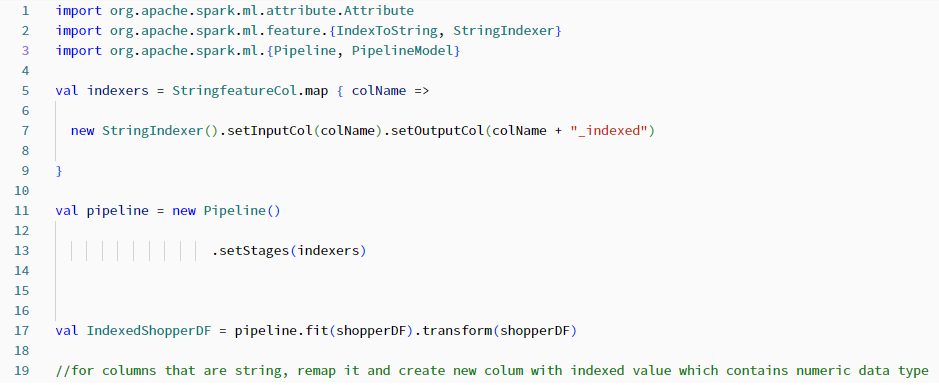
To answer this question, we will need to apply data testing and data training on profit prediction of the given order by using a Logistic Regression Model.



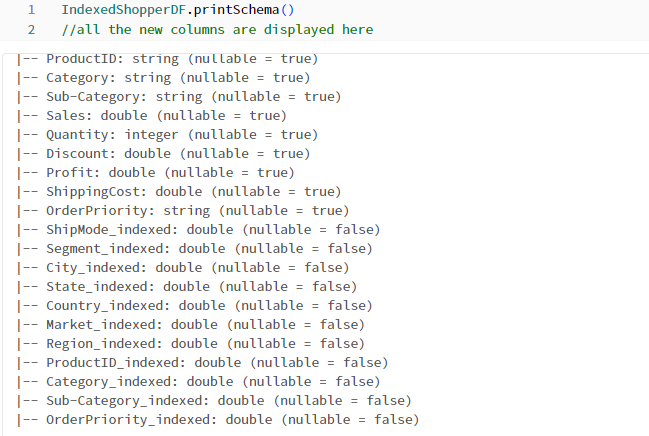
From the above, we import some libraries to be used to construct the logistic regression model.



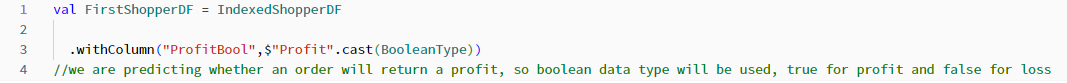
We will create new variable called StringFeaturCol that will store string data type columns that will then join using array function.



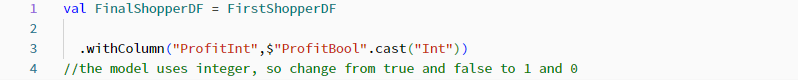
We import some more libraries for later use. From the StringfeatureCol, we create new column names that has the original name and add on “\_indexed” to represent the new columns. A pipeline variable is created, then IndexedShopperDF is created with the “Indexed” added in front to differentiate from the original dataset which is ShopperDF



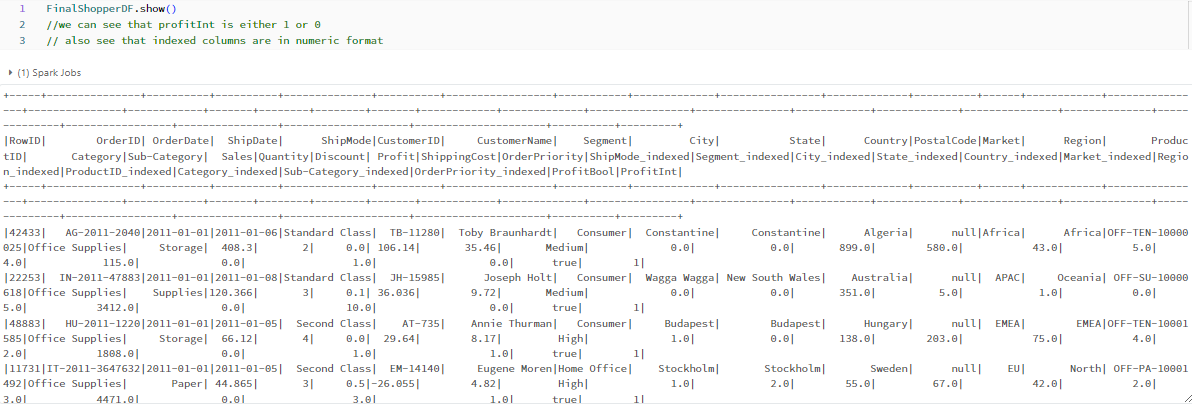
As displayed above, all the original string data type columns have new columns in the original schema with the same datatype.



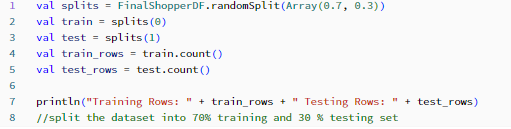
Since we want to predict whether an order will return a profit, we will first convert the profit column into boolean, which means if an order has positive profit, boolean will be true, and if order has negative profit, boolean will be false. We rename the dataframe to become FirstShopperDf since this will not be the final version used in the model training.



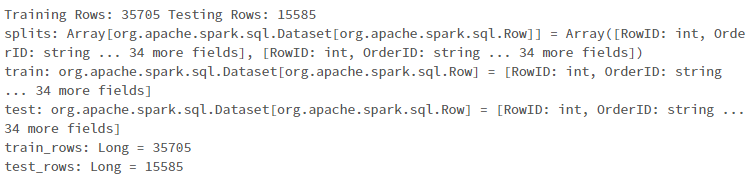
Then, we convert the boolean data type to integer meaning true will be 1, and false will be 0, since it is used during the predictive model creating later. The model does not recognize boolean values which is true and false, so we will change it to numeric 1 and 0. We then rename the dataframe to FinalShopperDF which is derived from FirstShopperDF.



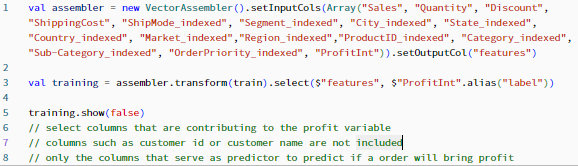
We can check that the FinalShopperDF does contain ProfitInt that contains either 1 or 0.



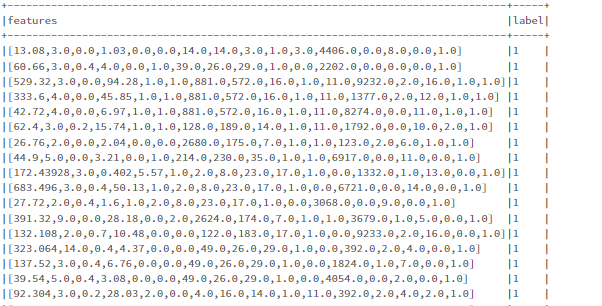
Then, we will split the data into two parts for training and testing the data, to ensure how well the data can predict the outcome for profit based on how much they train and learning the patterns from the training data. We split the data into 70% training and 30% testing.



From the results above, we can see that 35705 rows of data will be used as training data, while 15585 test rows will be used as testing set.

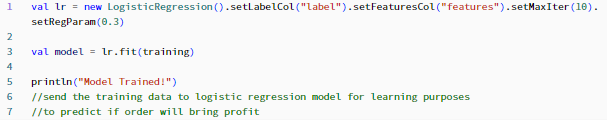


For the feature column that we created, we only choose columns of data that is likely to influence the result, which is profit column, so columns such as CustomerID and CustomerName is not chosen. The vector assembler combines the input column into “features” which is the left side in the table. The “ProfitInt ” column will not be included in “features” group but will be used as “label” and then the output profit is the right side in the table. We then display the two columns “features” and “label”.

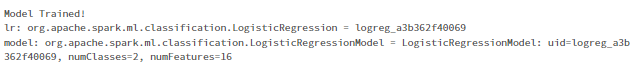


We can see whether an order will return a profit based on the input columns.

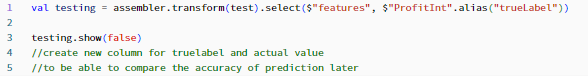
Modelling



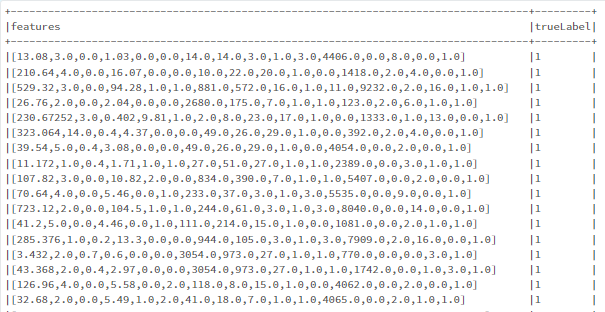
We chose to use logistic regression model as our prediction algorithm. Since we are predicting the profit for an order to be either yes or no, we would use logistic regression model for this type of prediction. The algorithm for prediction is created. The logistic regression needs parameters such as label column which is the input columns and features column which is the ProfitInt column. Once the model is trained, we then print out “Model Trained!”



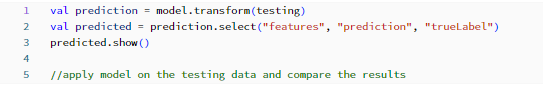
From the result, we see that the model has been created.

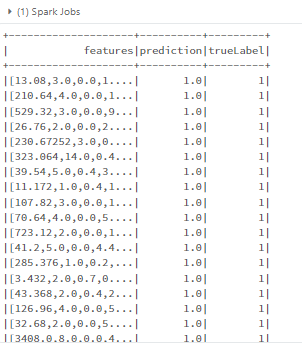


Since we have a trained model, we us it to test the 30 percent of the remaining test data that we set aside. We arrange the input columns into “features” and renamed the “ProfitInt” column into “trueLabel” for testing purposes.



Here we can see both the columns features and trueLabel are arranged together.

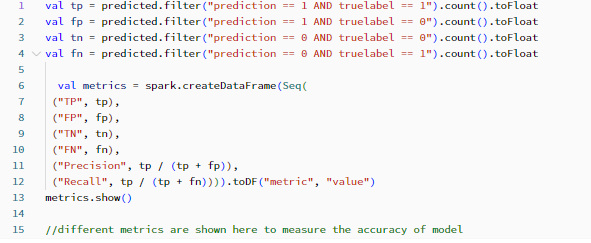


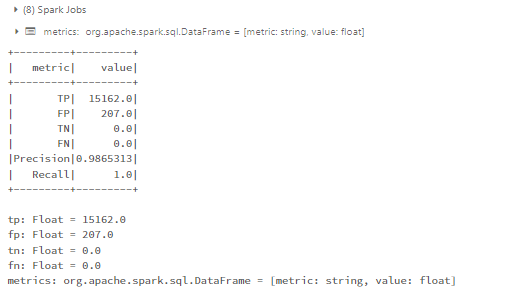


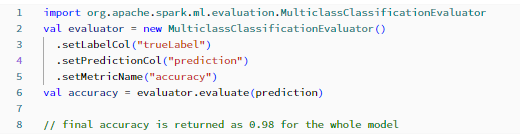
We showed the predicted profit that is generated by the logistic regression model in the prediction column, while the actual profit is then shown in the trueLabel column. These two columns side by side shows the accuracy and consistency of the predicted result compared to the actual result.

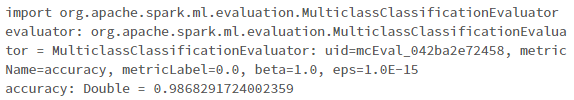
**2.4 Evaluation**

Evaluating the accuracy of the prediction model whether will profit or not for a given order



We display some metrics that correlate to the precision of the model, TP which is true positive, FP which is false positive, TN which is true negative and FN which is false negative, precision at 0.98 and recall. We had the highest numbers of true positives followed by some false positives.



We import a library that is meant to evaluate the accuracy of a prediction model. Creating a Multiclassification Evaluator which requires the columns that we have created earlier which are the truelabel and prediction columns, and we measure the accuracy of the predicted values compared to the actual values. The evaluation is completed and returns the final accuracy returned is 0.98

**2.5 Deployment (Prescriptive analytics)**

The prediction model in this project is aimed to predict the profitability for certain orders. Customers that have not purchased immediately will save their order into the online cart system. From there, the model could predict the profits earned if the customer chose to check out with the items. To ensure profitability on most transactions, we have come up with several recommended actions.

Member 1 – Albert Pang Kian Cheng

Understanding the profitability of a certain order can bring a strategic insight for the company. If the model is deployed to analyse a certain combination of products in the same order, preventative measures can be deployed to prevent any negative orders that will affect the company’s bottom line. Even for orders that have are in a good standing for the company in terms of profitability, it will not bring adverse effects to recommend customers to purchase a greater amount and greater variety of products to gauge customer interest for certain products. If the customer is recommended products that they are interested, it increases sales for the company, as well as satisfaction to the customer as they can find all items that they are looking for at the same place.

For example, the model can predict the outcome when a customer has a cart that has yet to be checked out, to predict total profits, recommend items that are more profitable to reduce loss by the company, or even increase the profitability of the order to maximize profits to the company. For example, one product that consistently causes loss is tables and if the cart of a customer contains the table product, we can let the system recommend other products from more profitable categories such as storage or appliances.

The recommended product can be dependent on the customer’s online activity to recommend similar products that the customer is interested in. Some product categories include technology which includes phones and copiers. The recommended product can also be a product that is popular among the customer’s area or region, we can leverage our existing product info to better serve the customers in different geographic locations.

Member 2 – Chin Yung Xuan

Understanding customer behaviour and maximizing profitability are essential in the dynamic world of e-commerce. An exceptional potential to maximize income streams is provided by the capacity to forecast profitability for pending orders. Businesses can strategize on how to convert these potential purchases into real earnings by studying shoppers’ behaviour, notably the tendency to save orders in online carts. A significant discovery from the exploratory data analysis points to the possibility for enhancing profitability by customizing promotions based on certain area features.

Exploring and analysing the data reveals a fascinating trend, which is certain regions exhibit high average profits coupled with low total sales. This intriguing phenomenon indicates an untapped potential for targeted interventions. One such example is the Canadian market region. Despite its modest total sales figures, it boasts a disproportionately high average profit per transaction. This phenomenon presents an excellent opportunity to capitalize on regional preferences and drive profitability through various customized promotions.

To illustrate this strategy, let us delve deeper into the Canada case study. While traditional approaches might focus solely on boosting sales volumes, a deeper analysis unveils a more subtle strategy. By aligning promotions with the preferences of the Canadian market, the company can effectively convert the existing customer interest into actual revenue. Below are some examples of promotions the organization can consider:

The first promotion can be a tailored product promotion. We have to identify products that resonate particularly well with the Canadian customers. These could be items with a higher profit margin that have historically performed exceptionally well in the region. Offering discounts, bundle deals, or exclusive offers on these products can entice hesitant customers to complete their purchases.

Besides that, Walmart can also come up with regional campaigns. They can launch targeted marketing campaigns that highlight the benefits of purchasing from the Canada region. Emphasize perks such as free shipping, localized customer support, or even region-specific packaging. This not only entices local customers but also adds a personal touch to the shopping experience.

Furthermore, an incentivized loyalty program specific to the Canadian market can be implemented. By offering rewards, points, or cashback for repeat purchases, they can also easily form returning customers and solidify their engagement with the brand.

In addition, Walmart should increase the amount of limited time offers in the Canadian region to create a sense of urgency of customers to purchase the items in their shopping carts. This strategy leverages the psychology of scarcity and prompts customers to take immediate action. Moreover, they should also opt for cross selling and upselling. This can be done by analysing the historical data of customers’ purchases to identify complementary products that are often purchased alongside popular Canadian items. By suggesting and recommending these products during checkout, they can boost the overall order value and, consequently, the profit margin.

Overall, the knowledge acquired by examining regional profit trends is an intriguing potential in the world of e-commerce, where data-driven decisions are the key to business success. Businesses should strategically focus their efforts on targeted promotions that appeal to local shoppers by realising the potential of high profit, low sale regions like Canada. These regions may develop into vibrant centres of profitability through tailored product offers, regional campaigns, and customer-focused incentives. In the end, the blending of data analysis and strategic marketing equips companies to maximise the potential of their clientele and promote ongoing profitability.

Member 3 – Eng Zheng Yu

Another insight that we have found is that there are regions with high sales but also high shipping costs. This means that customers in the region are generally interested in purchasing products from the company and some are even willing to pay the high shipping costs to receive the item. While there always be some difficulty in having products from the company’s warehouses to customers in all parts of the world, we can take some measures to lower the barrier to entry for customers who want to purchase. Two main criteria for the company are cost and shipping duration. There will be customers who prefer to wait but have lower shipping costs, there will also be customers who prefer fast delivery without worrying about the prices.

The company can arrange to collaborate with a regional logistics company to reduce shipping costs. A local logistic company will have more experience serving the delivery needs in their area since they are the ones that have been operating for a long time. The company should arrange for business agreements or contracts that seek to benefit both companies. In exchange for lower shipping costs, the company can choose to only work with a single logistics company for all their shipments going to a region or country. A mutual business agreement can bring benefits to our company and gain access to an untapped market, as well as the logistics company that stands to profit from the increased businesses from the sales and shipping of the company’s product.

The company can provide flexible shipping options by acknowledging the various priorities of its customers, such as cost-consciousness and expedited delivery. Express Shipping satisfies customers who want quick delivery even though it comes at a higher cost, while Standard Shipping caters to those who value affordability and patience with its affordable rates and extended delivery windows. The business can strategically place warehouses in areas with high shipping demand and strong sales, reducing costs and delivery times. Additionally, focused marketing initiatives can take advantage of the local appeal by highlighting faster shipping and limited-time regional deals. Negotiating carrier contracts for better rates and considering a minimum order threshold for free shipping, which encourages larger purchases, can both help allay concerns about rising shipping costs. A positive customer experience across all selected shipping options is guaranteed by open communication regarding anticipated delivery times and real-time tracking updates. The company can improve customer satisfaction, reduce costs, and enhance its standing as a customer-centric enterprise through these coordinated strategies.

For example, regions like China and Taiwan have high shipping costs. This business arrangement saves the customer’s shipping costs, we then pass on the savings to consumers, to increase their customer preference and to lower the associated cost for customers that want to purchase from us. Customers will then see that shipping is convenient and quick, and they will be more satisfied, and the company can gain the local customer’s support.

Member 4 – Tan Yann Bin

We noticed from the 10th business question, we can afford to offer slight discount to items to entice customer, while still maintaining profit margins. Discounts entice customers to act quickly and purchase while the discount period is still available. Customers also are always happy to purchase products that they want at a discount. The presence of discount can attract more customers to browse as well as purchase from the company. This also increases customer return rate as customers will be willing to revisit the company’s stores to see if there are new items on discount.

We can see that form discount ranges from 0.02 to 0.1 which is 2 percent to 10 percent we still have a stable profit margin to work with. This also increases sales volume and with this, we are not holding on to products for longer periods. For products that are always renewing such as our technology category, it will be unwise to hold on to existing inventory for too long as it may be replaced with newer technology, and when that happens, our inventory value could drop quickly. We ensure high inventory turnover and restock only products that are still relevant and sought after by customer purchasing trends.

This also incentivise customers to spend with us. Customers who are interested might not be in a rush to purchase, so by offering a slight discount, they are more willing to spend since there is a discount offered.

Member 5 – Tye Khan Wei

In the dynamic landscape of modern commerce, fostering customer loyalty is not just a strategy but a necessity for sustained business growth. One potent approach to achieve this is by offering enticing perks that not only retain existing customers but also entice them to become loyal advocates of your brand. By strategically implementing a range of incentivized programs, companies can tap into the psychology of consumer behaviour and create a win-win situation for both the customers and the business.

At the forefront of this strategy lies the idea of offering tangible benefits to repeat customers. One effective incentive is to provide reduced shipping costs or alluring product discounts for customers who make repeat orders of high-margin items. This dual-pronged approach ensures that customers are not only encouraged to make repeat purchases but are also inclined to select products that promise a higher return for the company. To achieve this, the concept of a loyalty program can be employed. By enrolling in this program, customers gain access to a tiered system of benefits that become progressively attractive as their purchasing frequency increases. For instance, certain benefits could remain locked until a customer reaches a specific threshold of item purchases. This ingenious mechanism not only incites continued shopping but also maximizes the allure of the perks, thereby amplifying their perceived value in the eyes of the customer.

However, the allure of a loyalty program need not be limited to the confines of individual transactions. The strategic launch of dedicated campaigns around these programs can further solidify customer loyalty. For example, by rewarding customers for referring friends and family, a company transforms shopping into a social activity. The existing customers, now turned into brand advocates, not only expand the company's potential customer base but also derive a sense of satisfaction from being part of a community that values their contribution. To elaborate, a successful referral could trigger benefits such as a 5% discount or free shipping on the next purchase, thus fostering a sense of camaraderie and mutual reward.

Taking a slightly different route, a cash-back system can also be incorporated into the loyalty strategy. By offering customers a percentage of their purchase amount as 'cashback,' a company provides an intriguing proposition that keeps customers engaged over time. Though the accumulated cashback might not be redeemable as real money, its perceived value to save on future purchases encourages customers to maintain their shopping momentum. This, in turn, fosters repeat business while heightening the sense of financial prudence among customers.

In today's saturated market, the battle for customer attention and retention is relentless. The creative implementation of loyalty-boosting strategies becomes the bridge that transforms sporadic shoppers into devoted brand enthusiasts. Through perks that range from enticing discounts to ingenious referral programs, companies wield a powerful tool to shape consumer behaviour while nurturing a lasting connection with their clientele. By tapping into the innate human desire for reward and value, businesses can craft a narrative where every purchase becomes a step toward not just acquisition, but a meaningful relationship with the brand.

1. **Conclusions and future work**
   1. **Primary Findings**

From what we have found from the project, we are getting closer to achieving our objectives of predicting the profitability of any order. We also correlated the effect on order priority on the profitability, as well as get a wide view on certain aspects in the business operation, the breakdown on profitable areas as well as potentially untapped markets.

We set out to train a prediction model to predict the profitability of a given order, whether it will bring positive profits or negative which is considered as a loss. We can predict the profitability based on various factors such as order quantity, productID as well as discount offered for the order.

Some additional information that is visualised include shipping cost to most of the countries and states in our customer base. We identified areas to focus on where there are high shipping costs which we can try to reduce to increase customer popularity.

We also identified product categories and sub-categories that seem to negatively influence the profitability. From this, we can make efforts to reduce the losses coming from these product categories, either by recommending other more profitable products for the customer or even changing our pricing for said products.

For the second part of the project, the model is trained to identify orders that have high probability to either bring profit or loss. The model had a final accuracy score of 0.98 which is sufficient for our use case.

Our visualizations created during EDA phase should also be taken into consideration as it has created some insight into the effects of various factors into the profitability of a given order.

With all this, we are confident that the company has lots to gain following this to act and to increase their efficiency and profitability in their business operations.

* 1. **Strengths and Weaknesses**

One strength that can be found is that the model successfully identified regions with high potential for the company to expand into, the company can just make efforts to better understand customer behaviour in the region and plan their expansion accordingly. The model also identified certain products that are negatively affecting the bottom line of the company, so the company can find ways to increase the profitability of the certain items that are identified. Furthermore, the model created has factored in all the different criteria for calculating the profitability of the order. So, the modelling portion of this project can mainly be seen as a success.

Meanwhile, there are still some weaknesses of the project. The dataset that is deployed can have more EDA done to find more insights for the company, there could be more ways to arrange the data to have more clarification and specifics to be more in-depth. More visualizations can be created to visualize the more in-depth portions of the data. There are also other business questions that can be explored aside from the examples above. One weakness that is also found is that for geographical visualizations, we did not visualize it more specifically than the state level for the countries. Large companies that have the potential to increase their growth for each city and we have yet to explore further on that concept.

1. **References**

Walmart: eCommerce sales worldwide FY2019-FY2023, by division,T. Ozbun, Mar 20, 2023

<https://www.statista.com/statistics/1109330/walmart-ecommerce-sales-by-division-worldwide/#:~:text=Walmart%3A%20eCommerce%20sales%20worldwide%20FY2019%2DFY2023%2C%20by%20division&text=In%202023%2C%20Walmart%20U.S.%20had,in%20comparison%20to%202022's%20figure>.

*Walmart - Wikipedia* (1962) *Walmart - Wikipedia*. Available at: <https://en.wikipedia.org/wiki/Walmart>.

*Sam Walton | Biography & Facts* (no date) *Encyclopedia Britannica*. Available at: <https://www.britannica.com/biography/Sam-Walton>.

*6 Reasons Walmart’s eCommerce Strategy is Winning* (2023) *Indigo9 Digital Inc.* Available at: <https://www.indigo9digital.com/blog/4-secrets-to-walmarts-ecommerce-sucess>.