

**The Background**

The largest Ecommerce Data Set on the Pakistani ecommerce industry, encompassing years 2016 till 2018, was selected in wake of the growing tilt towards shopping online. This was before COVID hit, so essentially the industry was under trying waters and the trend can only be extrapolated over the years, given the veracity with which the masses have shifted to making their purchases online. However, uncertainty remains as to which products are preferred when making an online purchase and which ones consumers would prefer to buy in person. Additionally, ideas related to whether a person with a bad experience tends to shop online again or do people with good experiences become routine online buyers remain. How much role do prices play when categorizing online sales, is it that people prefer buying low priced products online so the risk is minimal? Which categories are preferred and which skus pertaining to them specifically, are the products which do need to be tried before buying or are they. Which payment methods have been the most preferred, which month witnesses the most sales - is it seasonal or cyclical or a one time happening. All these are imperative for business to grow and for new businesses to enter the market with good acumen about the industry and customer preferences.

**Literature Review**

Over the years, there has been an increased focus on using big data analysis in e-commerce to understand dynamics and forces of the business model. As e-commerce continues to increase, it is important to understand the various factors at play behind the platform’s massive success over the last few years and its prospects in the future. More and more businesses are incorporating e-commerce now because of its immense business benefits. However, due to large data traffic, it becomes important to analyze the data in a way that benefits the business. This analysis can then be used to reap greater benefits such as boost in overall business value, identifying new products or services, diversifying into different markets or improving sales and revenue (Akter & Wamba 2016). There is also an added benefit in terms of the non-financial factors resulting from using big data analysis, such as customer satisfaction, loyalty and retention. For example, United Parcel Service (UPS) was able to improve its customer retention rate by identifying and understanding the complaints and patterns of customer defection (Akter & Wamba 2016).

Moreover, the businesses are able to modify their marketing and sales strategies as analytics would help them gather specific insights through e-commerce. Additionally, greater preferences about customers, such as their most preferred item, least preferred brand can also be generated as a part of the e commerce analysis. Analytics in e-commerce would enable companies in producing customized products and services which would ultimately lead to increased customer satisfaction and retention. As pointed out by one study, 56% of customers today expect some form of personalisation from companies (Mileva 2023).

Lastly, predictive modeling can be used to determine optimal prices by gauging demand and competitor pricing. This in turn can be used to increase profits and lower costs (Gogineni).

**Problem**

The problem is twofold. Firstly, the goal is to minimize the costs incurred as a part of e-commerce operations so that prices can be optimized and a competitive advantage is gained. Secondly, the company wants to improve customer retention and therefore, data analysis will be used to determine the solution to these problems.

**Raw Dataset**

The data was obtained from an open source. It is the largest retail e-commerce orders dataset encompassing over half a million records. It has 28 columns, 585,525 rows of data, 500,000 values of which were null and required cleaning. The variables it is made up of include the following:

"item\_id", "status", "created\_at", "sku", "price", "qty\_ordered", "grand\_total", "increment\_id", "category\_name\_1", "sales\_commission\_code", "discount\_amount", "payment\_method", "Working Date", "BI Status", "MV", "Year", "Month", "Customer Since", "M-Y", "FY", "Customer ID"

**Data Processing**

**Removing Data**

The last 5 columns of the data were removed as they were blank as well as columns titled increment\_id, sales\_commission\_code, working\_date, BT.status, MV, M.Y, FY as they offered no real value to the analysis, The first four provided no value addition, while the rest were variations of month and date.

**Conversion of Data to Correct Formats**

Created\_at and Customer.Since columns were converted into date formats respectively by using the as.Date function and setting the format as %m/%d/%Y, they were previously in chr format. Customer.Id was changed to integer format from character format.

Variables like category\_name\_1, payment\_method and status were converted into factors as they are all categorical variables which can only assume certain levels. As.factor was used to implement the change. Before this however, tables of category\_name\_1, status, payment\_metho and sku were made so an understanding of the variables could be achieved to then decide how to move forth with cleaning them. After converting them into factors, tables were recreated. Skus were not converted because they had too much variation to be classified into categories.

**Reordering Columns**

The columns were reordered to have them placed in positions that make the most sense relative to each other and provide ease of understanding.

**Dealing with NAs and Null Characters**

NA values for each numerical column were counted and omitted using na.omit. After that, null characters were removed from columns with categorical variables. ‘ ‘ as a part of the subset was used to undertake this and remove these characters from status, payment\_method, category\_name\_1 and sku.

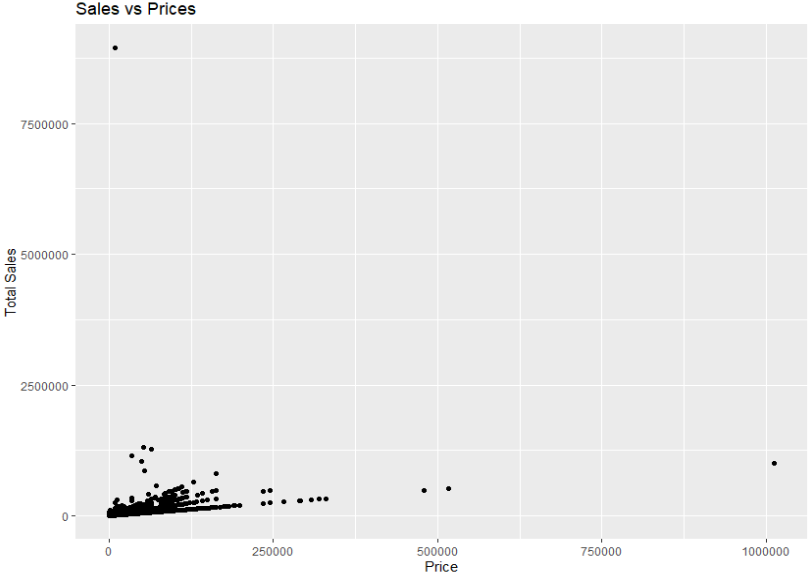
**Feature Engineering**

\\N was removed from status as there were only 4 occurrences. \\Other from category\_name\_1 was removed. Status required more cleaning than the rest as there was too much variation and splitting in between complete, refund and canceled. A named vector of replacements was made (Exhibit 9). The values were then replaced in the status column of the data in the form of order canceled, order complete and order refund.

After this, payment\_method was dealt with for the same reason and the same manner (Exhibit 10). Ones with very small amounts were categorized as others.

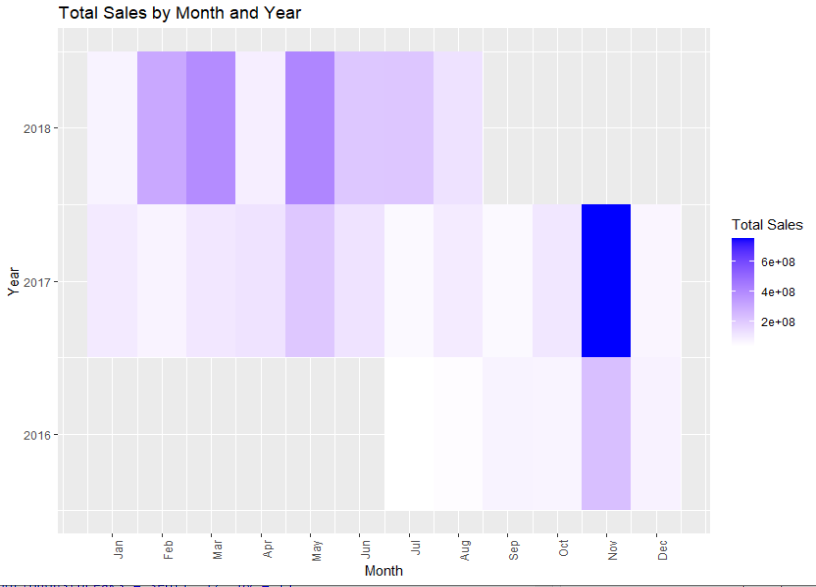
**Exploratory Data Analysis**

EDA was undertaken to identify underlying trends and potential pain points or helpful insights before a more rigorous manner of analysis was executed.

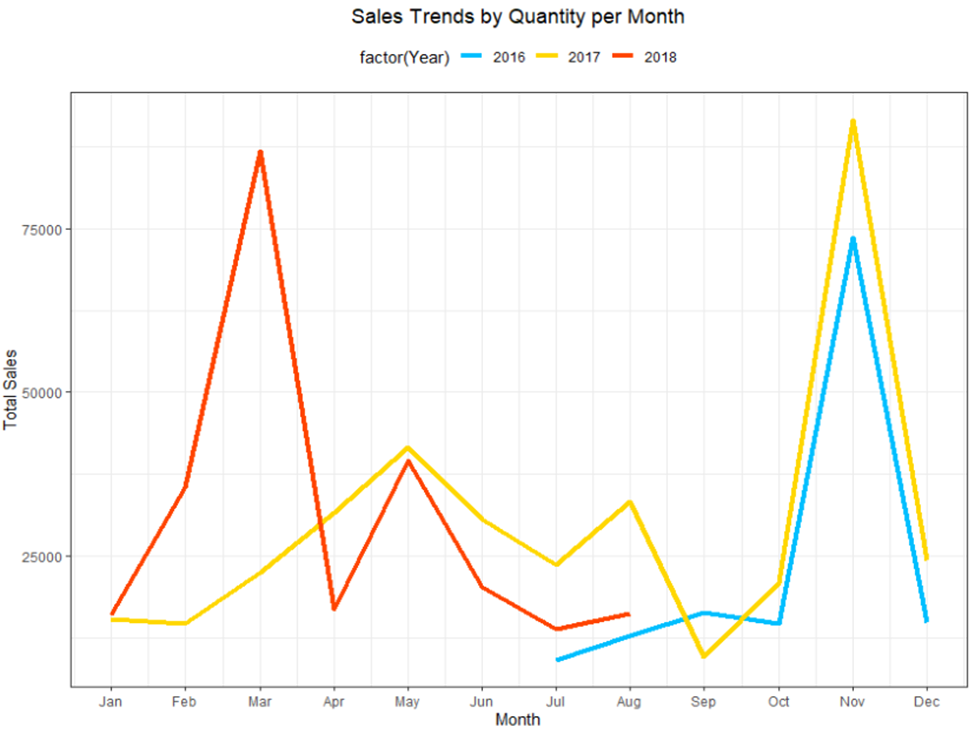


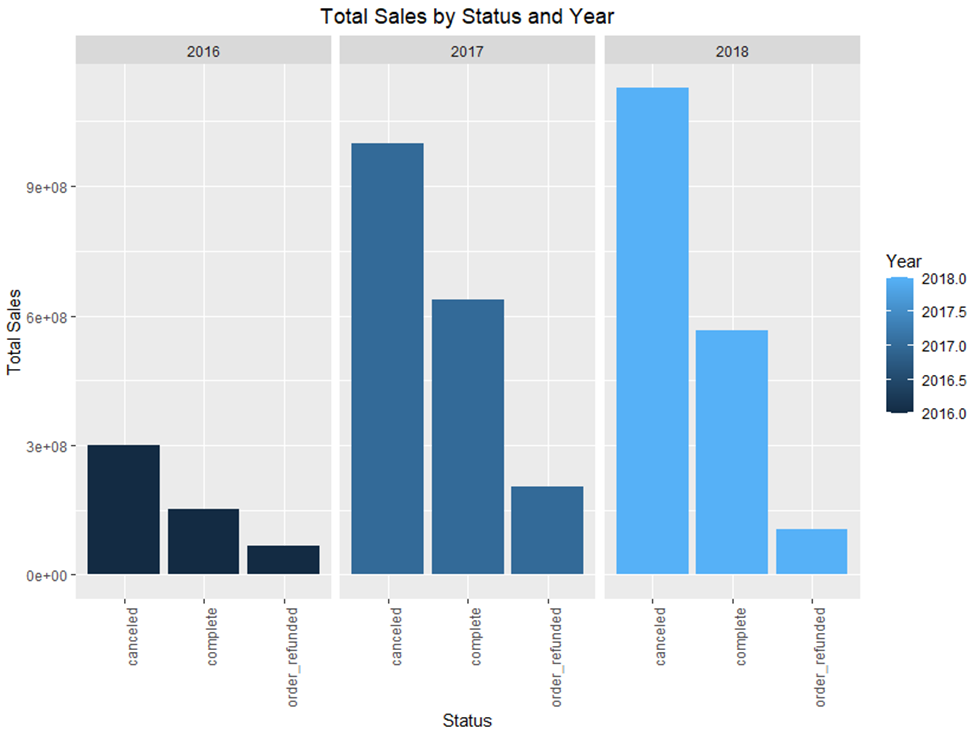
We conducted an analysis of sales vs prices to understand the impact of prices on total sales and how prices can be altered in order to achieve maximum sales. It can be seen that most of the sales are concentrated till a price of 250000 (accumulated).

Additionally, total sales by month would help us determine which months should be focused on in order to retain our customers or increase customers.

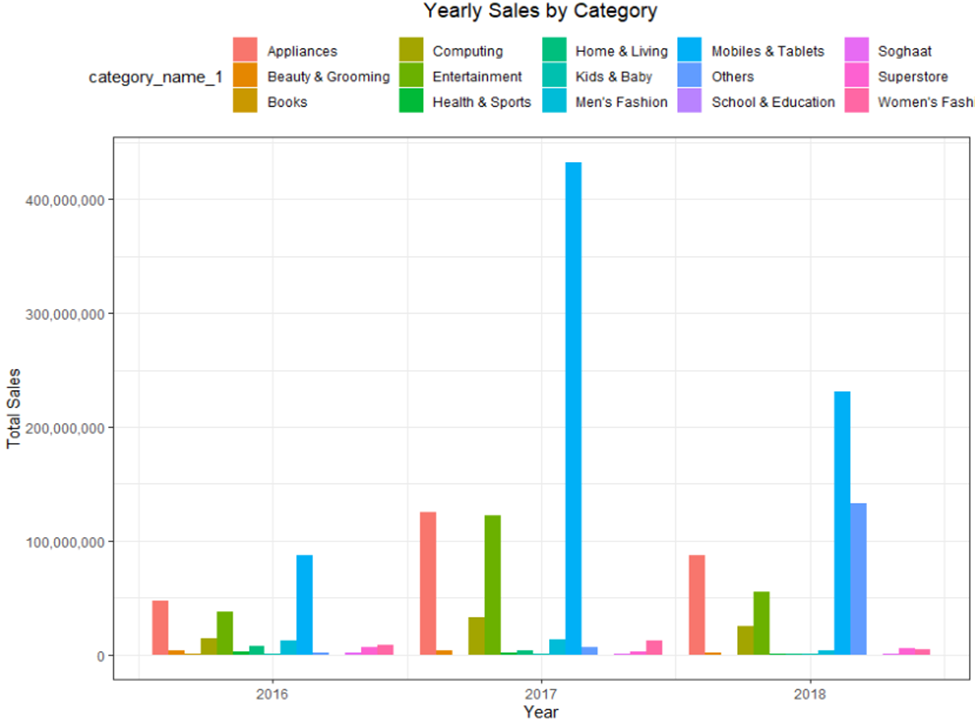
Additionally, a heat map was plotted to understand how sales were divided amongst the years and months as an expectation about dead months was set. Most of the sales were made during the November of 2017 while the least sales were during July of 2018 which might be because Eid was celebrated in June that year.

Upon an analysis of the least popular skus and items, the tail(10) function resulted in 10 skus that were never sold and they all belonged to the superstore category. The skus were written with numbers so a deeper analysis was not possible, the assumption thus is that they might be perishable products which consumers often find unfeasible to order because of their limited life (Exhibit 1).

A plot to understand sales trends by Quantity per month was made to understand the effect of seasonality on sales, perhaps hotter months resulted in more online purchases. As can be seen, however, in 2018 March had the most sales, compared to November in 2016 and 2017. July was the lowest for 2016 and 2018 while September was lowest for 2017.



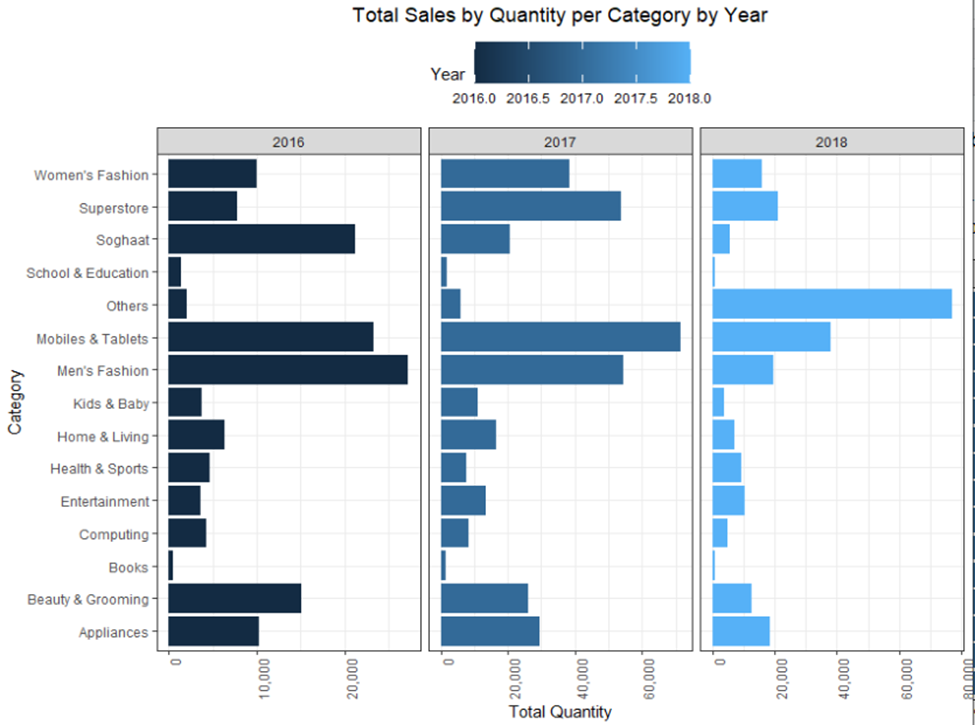
The frequency of canceled orders has only increased over the years, completed orders peaked in 2017 however they too witnessed a fall in 2018.

To understand which categories sell the most, a plot against sales and years was made. A growing trend towards ordering Mobiles and Tablets online can be seen while the trend for appliances peaked during 2017 but has been falling, likewise with health and sports as their sales too fell. The sales in the computer category too increased slightly before falling in 2018. A deeper analysis into the months too was conducted (Exhibit 3) wherein most of the sales were made during June and July during 2018 compared to November 2016.

Chart, bar chart

Description automatically generated

To better understand the stigma associated with paying online, an expectation of COD being the most preferred mode was expected however, Easypaisa made for the most sales. It is important to note here that total sales are taken as quantity ordered multiplied by the price, so it is possible that only payments for more pricier products were made this way. Credit cards were the least used mode of payment.

Additionally, categories per Year too in terms of quantities were studied and an evident fall in the sales on Mobiles & Tablets, Men’s fashion, Soghaat, Women’s fashion, and Beauty and Grooming can be seen. Sales in the Others category have increased significantly, sales in the entertainment category have remained stable as have health and sports.

Top ten items ordered categorized by their quantity too were found to better understand which items within the top categories sold well (Exhibit 4) Additionally, top 10 customers based on their total purchases too were classified to better understand their contribution to sales, which products they preferred and from a business perspective how to better target them with promotional products and offerings (Exhibit 5). Median value of purchase is relatively higher in COD and credit cards as compared to easypaisa, The range of the former two is also more indicating greater variability. (Exhibit 6). Most orders lie between 0-100 while some are above 500. There are very few orders with a high value, indicating that mostly items sold are not that expensive (Exhibit 7)..

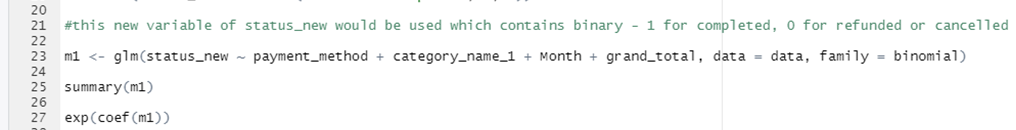
**Segmentation**

Status was set to factor given its categorical existem=nce as was payment\_method. The dataset was sampled to 20,000 rows randomly by a pseudo random number by using set.seed(), as that was the maximum limit given the limited computational power. To check if it was representative of the actual data summaries of both were compared which were similar, so the subset was then used. **Gower Method** and **Daisy function** were used as they automatically calculate distance and make a dissimilarity matrix for both numeric and categorical variables by using **euclidean** and **manhattan** distance respectively. The resulting **Distance** was then converted into a matrix, called distance\_matrix.

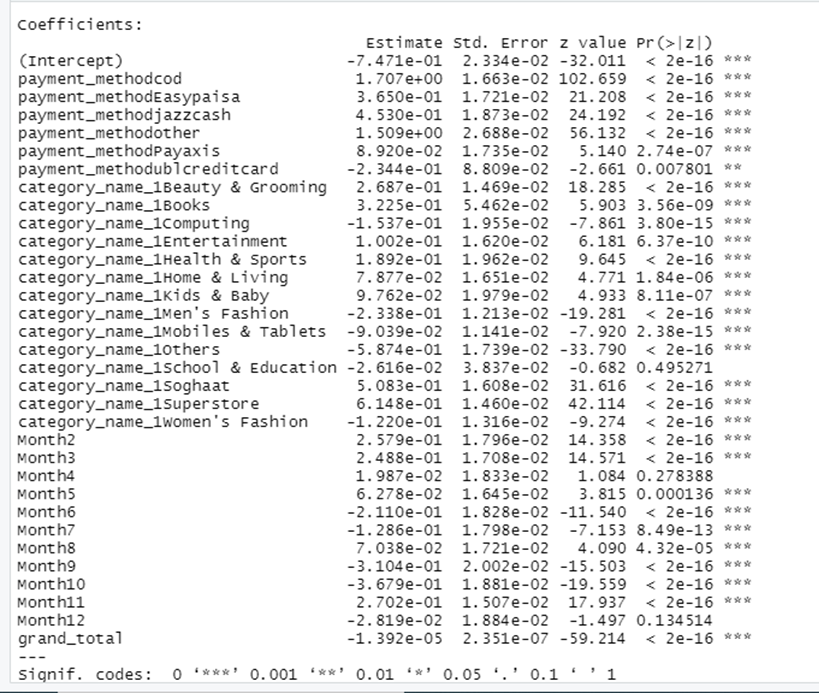
In place of k means, **PAM method - Kmediod Clustering**, was used as it only works for numerical data with a known distribution, and is not fit for categorical data. The **silhouette score** for each respective cluster was then calculated and the one with the maximum score was then selected (Exhibit 11 and 18). The larger the silhouette score the better, and it falls under the -1 to 1 range. Clusters were made from 2 till 7. The silhouette score decreased from cluster 2 to 3 and 4 too. It increased in cluster 5 and in 6 it increased to 0.3, it then decreased in 7 so k=6 was then used. The cluster sizes were noted. The prices for cluster 2 through 7 were 17684.77, 13834.77, 2295.20, 1143.27, 3677.47, 7957.88, respectively. The average quantity ordered was 1+ in all while the discount\_amount was 793.3465, 1621.7487, 482.6851, 100.7455, 135.4630 and 625.5754, respectively.

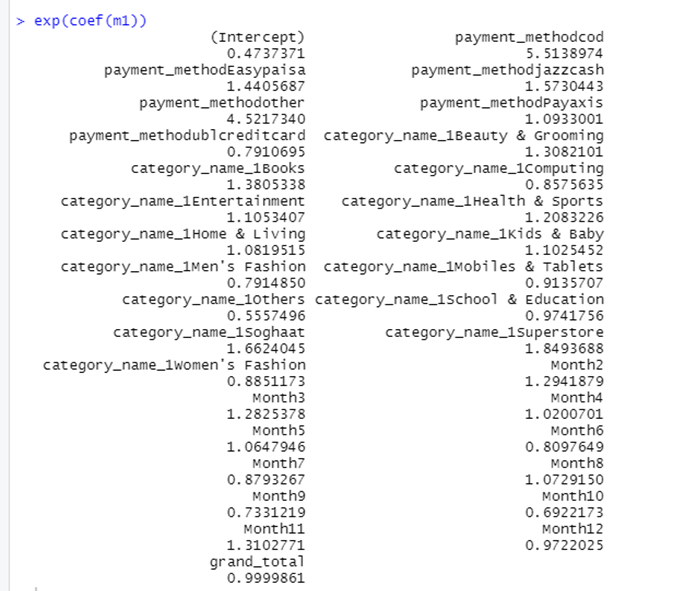
Cbind was then used to attach the defined clusters with the data set so each row could be linked to their respective cluster. Aggregates for numerical data were then calculated for the numerical data (Exhibit 12). **Mode** was used as a measure of central tendency for categorical data. It was defined as a function and run for category\_name\_1, status, and payment\_method as they were the only relevant categorical variables (Exhibit 11-17).

**Logistic Regression**

Also known as Choice models, are useful when the customers have a choice, and we can model their past choices to predict the future. Hence, with applications in predictive modelling and segmenting customers, this tool can be a handy one.  
In the context of this problem, the question we are solving is what can be one of the reasons customers cancel or refund their orders. For simplicity, every completed order is labeled as 1, and every canceled and refunded order is labelled 0. Next, appropriate variables are converted into factors such as quantity ordered, payment method, category name etc. The model looks like this:

Here, the effects of each of the dependent variables on the completion of order is measured. The results are to the right.

As it can be seen, most of the variables are statistically significant. To analyze, the coefficient of -1.39e-05 for grand total can be read as for every 1 PKR increase in grand total, the log likelihood of completion decreases by -1.39e-05. To make a better sense of the coefficients, exponents of all the coefficients are taken.



By exponents of the coefficients, their meaning can be easily derived. For example, with Month 1 (January) taken as base, the completion likelihood of orders are 1.294 times higher than month 1, or completion likelihood (odds) increase by 29.4% when the month is changed from January to February.

**Important insights**

For the payment method, with a Bank Alfalah card as the base, we could see that the completion likelihood is the maximum at Cash on Delivery (COD – 5.513). And by UBL credit card, it is the least with completion likelihood decreasing by 21% if payment method switched from bank alfalah to UBL.

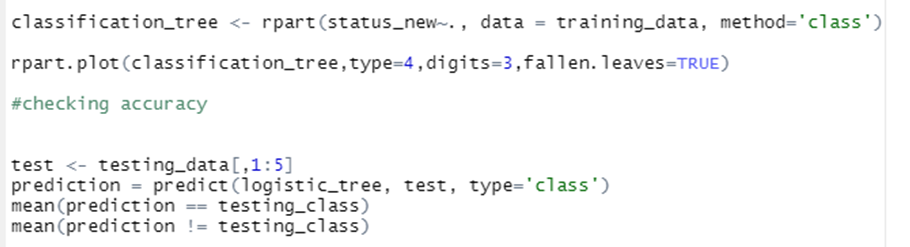
For categories with Appliances as base, we could see the completion likelihood increasing for every other category except women’s and men’s fashion. (decreases by 12% and 21% respectively). Hence it can be seen that fashion items have a higher probability to get canceled.

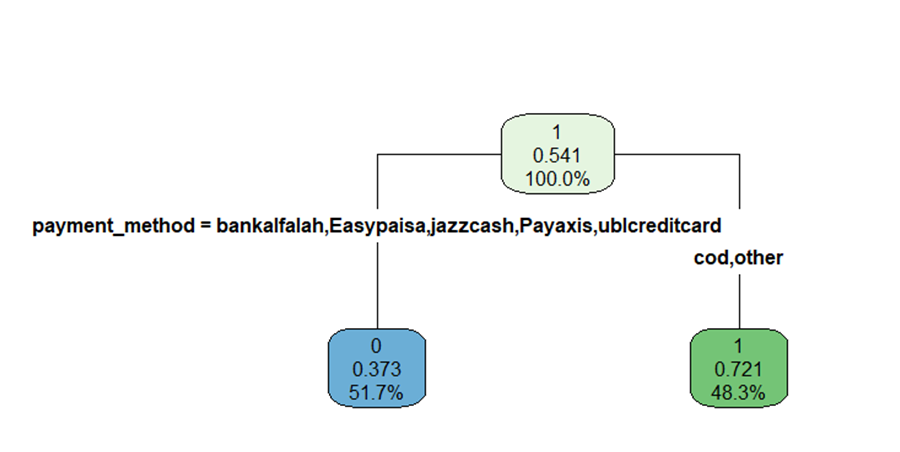
Similarly for months, June, July, September, October and December the cancellation seems to increase. Considering this data is for 2016-2018, these months can correlate to Eid and Christmas, which are major proponents of online purchase behavior.

Hence, an online business could view this result, and then find out which prospective customer is most likely to refund and cancel their order so that they could target their efforts towards the one with higher probability of completion of order.

**Classification Tree**

A simple classification tree can also be set up, considering the response variable (Completion) is binary and not continuous. Here, the data set was first split into training (65%) and testing data. Then the following code was run:



The output was a classification decision tree.   
Starting at the root node, completion of order accounts for 54.1% of the data. This node asks whether the payment method is Alfalah, Easypaisa, JazzCash, Payaxis, UBL or COD/Other. Which can simply be considered online vs cash on delivery. The left most lead node accounts for order cancellation/refund, and illustrates that there is a 37.3% chance that an order will be canceled, with online payment methods accounting for 51.7% of the population. Whereas there is a 72.1% chance of completion of order, with 48.3% of the customers using COD method.

**Limitations and Conclusion**

**Assumptions**

1. For ARules the frequency of repeated orders in the same categories can not be measured, so no specific recommendations are possible for the same categories, only different ones.
2. For Logistic regression and classification decision tree, it was important to change the dependent variable into a binary form, with 1 as completed and 0 as refunded/canceled. Refunded/canceled are both taken 0 for the sake of simplicity.
3. Another assumption which is taken here is that the cancellation of the order can be due to poor quality of the product as well, due to the type of data available, this was not catered.

**Limitations**

1. The dataset is limited to two years which makes a lot of the results collected probable to extrapolate only, so sudden changes in the trends like COVID’s occurrence may not be accounted for properly.
2. The data has limited details about the skus, the companies they belong to, their origins and their brands, so the analysis is relatively at a surface level.
3. The data also has limited information about customer satisfaction so deriving conclusions based on popularity alone in terms of quantity sold or total sales is not reflective of the true perception about the product.
4. Arules becomes outdated very quickly as new products and shopping trends emerge every year or even month, so it has to be updated and processed constantly

**Conclusion**

Conclusion points for logistic and classification:

* Focus on COD as these have a high chance of completion
* Improve online payment gateways

**Appendix**

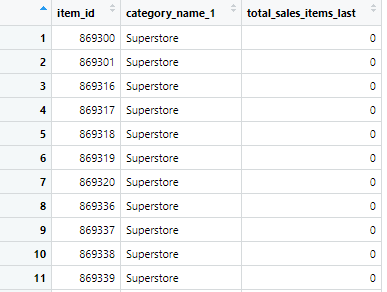


Exhibit 1

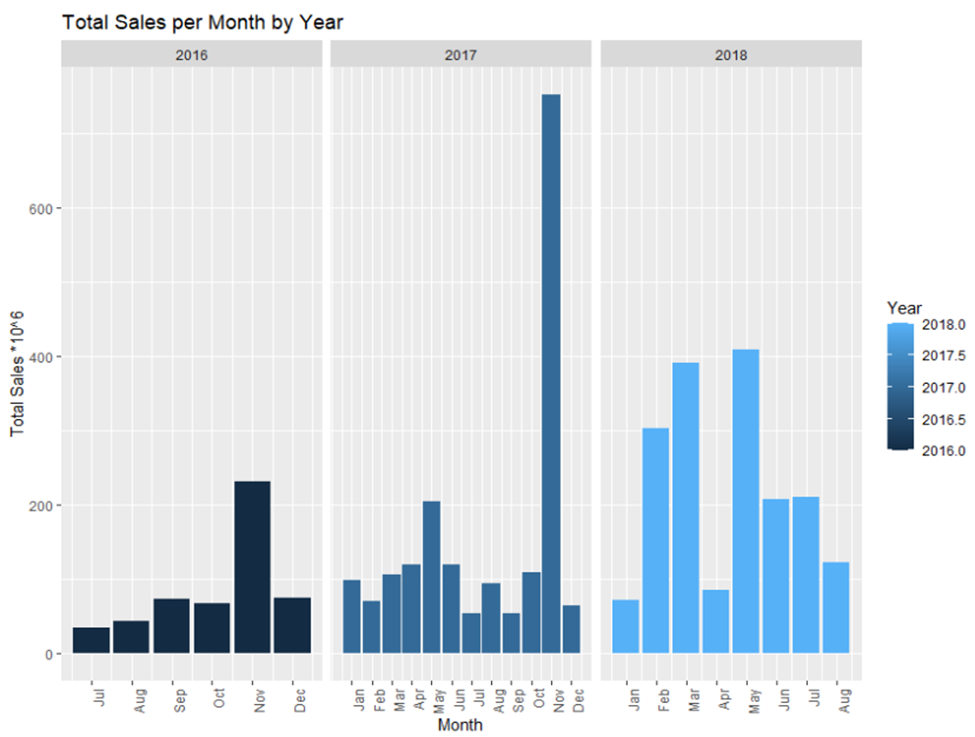


Exhibit 2

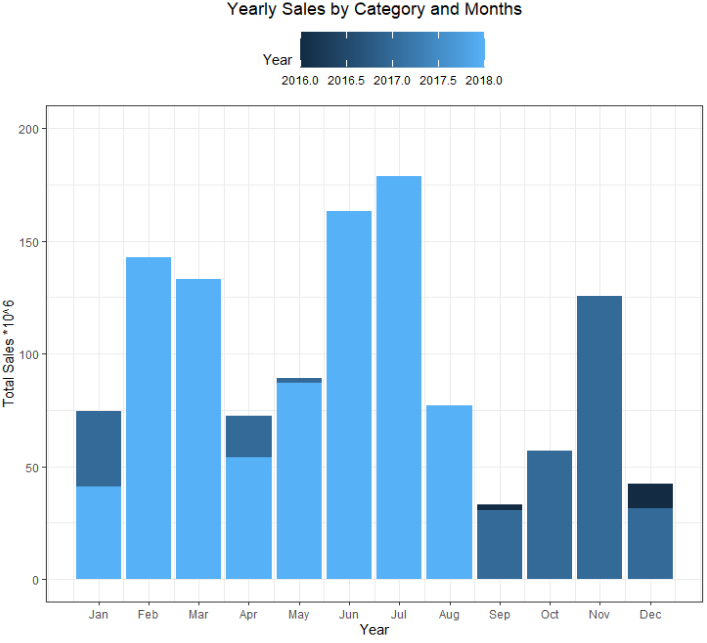


Exhibit 3

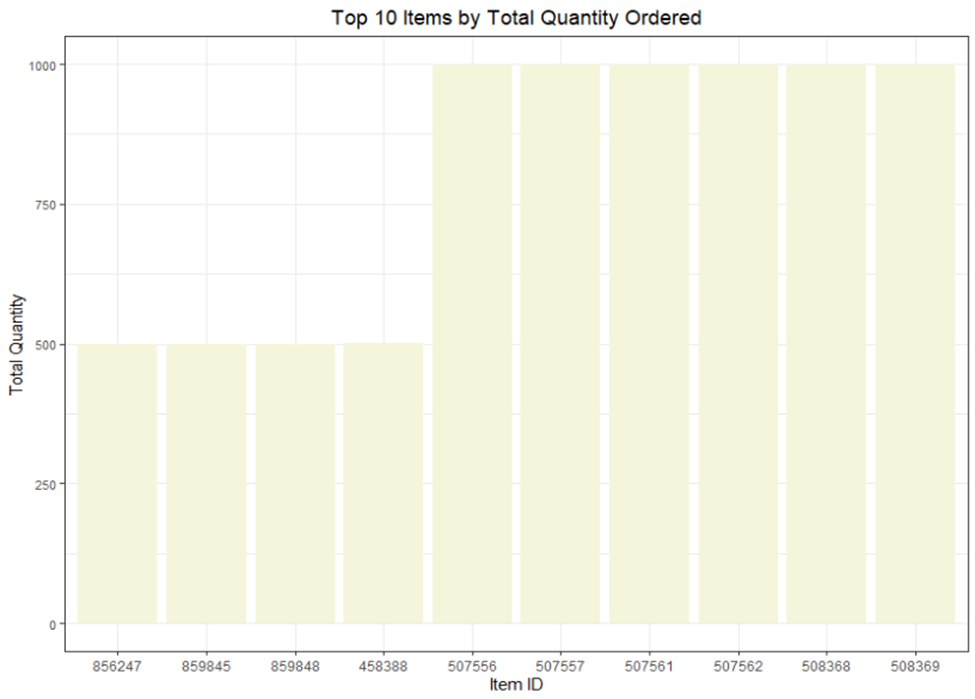


Exhibit 4

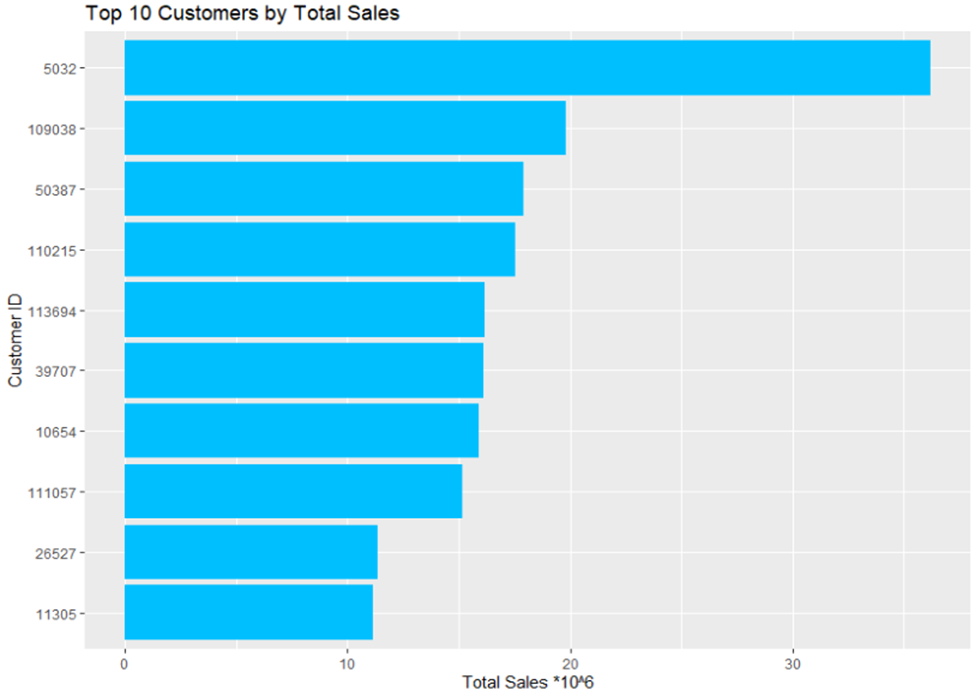


Exhibit 5

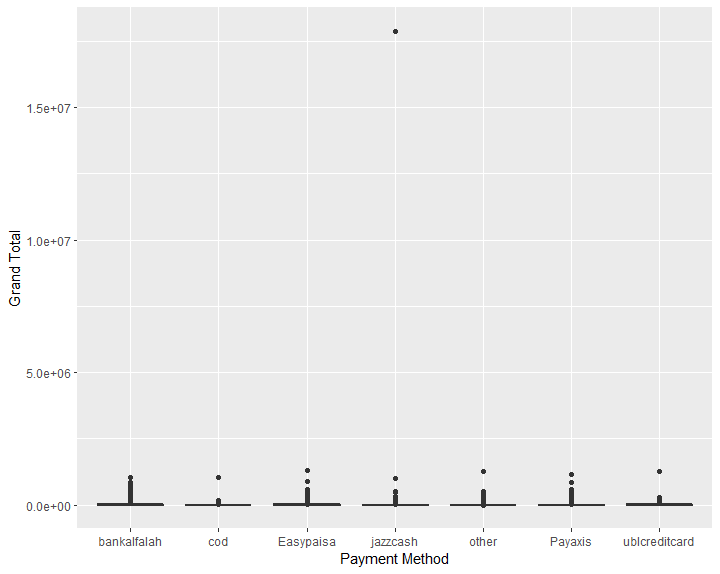


Exhibit 6

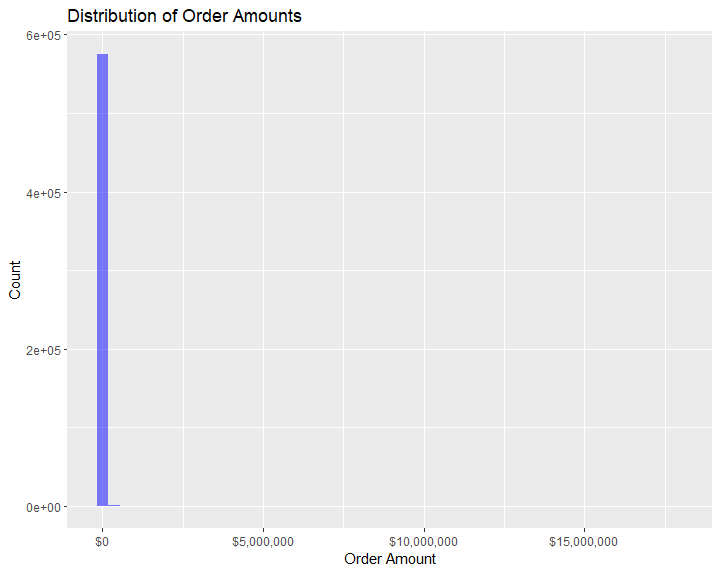


Exhibit 8

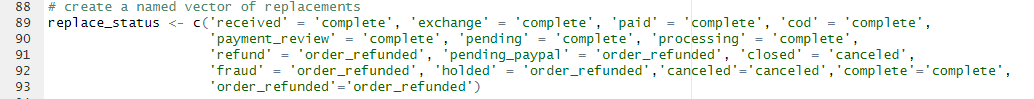


Exhibit 9

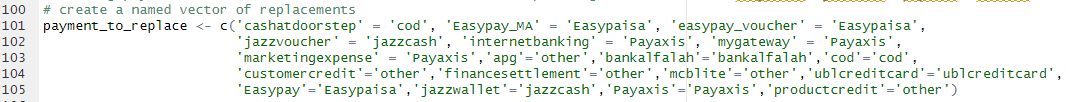


Exhibit 10

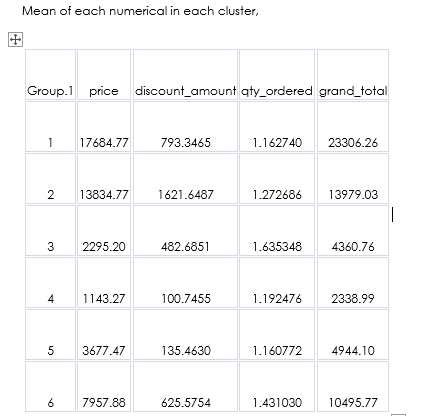


Exhibit 11

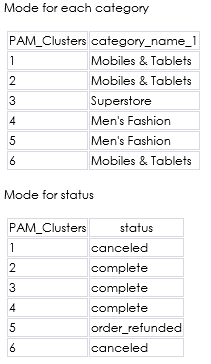


Exhibit 12 and 13

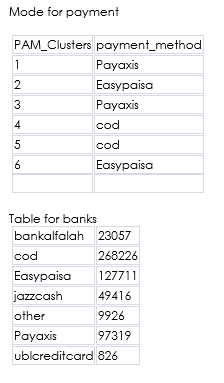


Exhibit 14 and 15

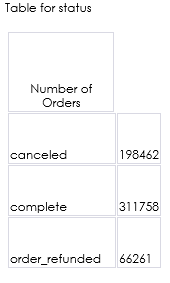


Exhibit 16

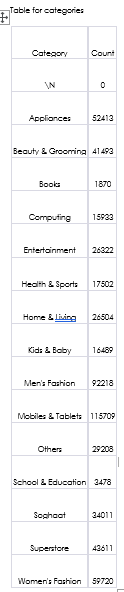


Exhibit 17

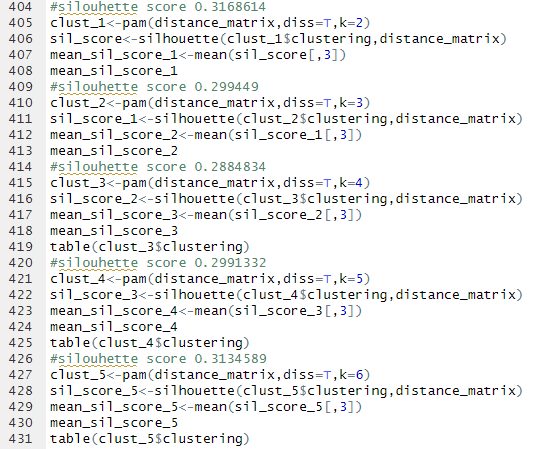


Exhibit 18