Optimizing The Business Processes of Retail and E-Commerce Platforms with Data Analytics and Machine Learning

1st Kushal Patel
Data Analytics For Business,
Zemelman School of Business
and Information Technology,
Windsor, Canada
kp54@myscc.ca

2nd Viraj Modi
Data Analytics For Business,
Zemelman School of Business
and Information Technology,
Windsor, Canada
vm21@myscc.ca

3rd Vishakha Parab
Data Analytics For Business,
Zemelman School of Business
and Information Technology,
Windsor, Canada
vp30@myscc.ca

***Abstract*— Improving techniques of conducting business with the help of data science tools and techniques like Market Basket Analysis, Customer Segmentation using RFM analysis and Clustering, Sales Forecasting to improve overall business to increase sales and profits.**

Keywords*— *Basket Analysis, Support, Confidence, Lift, RFM analysis, Recency, Frequency, Monetary, Clusters, Customer Segment, Clustering, Forecasting, KPI's, Time Series, Analysis, Error.

I. INTRODUCTION

The retail industry has drastically changed from brick-and-mortar stores to e-commerce and has overall increased in the previous decade as the advancement of technology never stopped. We can see the online presence of a small convenience store near our homes or big-size stores, on platforms like Google or social media with their dedicated sites for e-commerce. So, in the new tech world now the business owners have changed their techniques of conducting business with the help of digitalized tools and with the help of software like CRM where they can record data for billing, shipping, taxes, sales, inventory, and many more. As the current world thrives on data and almost every possible problem can be solved with data so with the help of data science business owners can make decisions and can predict their business for the near future with help of data science. It is not limited to decision making or predicting the outcomes in near future, but the data can also help us in recommending the product placement to increase sales and forecast their sales, inventory requirements, make informed marketing strategies, and many more. The data can help both the retail store and e-commerce to advance their business in numerous ways. Our multiple data-oriented insights and predictions will help the business owners in decision making and planning for their business.

Market Basket Analysis is to make a recommendation system for an e-commerce website and to help a retail store with its product placement to increase sales. Customer Segmentation i.e., categorizing the customer based on RFM score to find loyal customers and to find the potential customers churn with the help of RFM Analysis. Sales Forecasting to understand the business direction. Inventory Management to stock freight in the future. Analytical Reporting to analyze business data with the help of business intelligence dashboards and visualizations to display the business's key performance indicators (KPIs), assess performance measures, and generate actionable insights.

In our project, before using the data we have followed the five data ethical principles like consent means the author of this data has made publicly open for analysis so directly it gives us authority to use it, clarity means we are very clear about goals that we want to achieve by using this data and how this data will be used, consistency means we will be not giving this data to any third party or combine it with other data, control means data owners have all authority to deny from not using their data for analysis at any point of time, and consequences means we are aware of the consequences if we misuse this data or do not comply with one of these principles.

II. RELATED WORK

Raorane A.A., Kulkarni R.V., and Jitkar B.D. in their paper "Association Rule – Extracting Knowledge Using Market Basket Analysis" showcase that how the power of advanced data mining tools can be used compared to the traditional approach for analyzing the huge amount of data which can help organizations to make the correct decision about their marketing strategy to increase sales that can help them to be ahead of their competitions. In this paper first, they explain the three-tier data warehouse architecture to understand how knowledge is extracted from the data warehouse for generating reports, performing analysis, and data mining. They have used one of the local supermarkets called Shetkari Bazar's (located in Kolhapur city in Maharastra) actual day-to-day

transactions dataset made by the customers for their Market Basket Analysis. In market basket analysis frequently purchased products and their combinations are found by generating different association rules. Each rule is determined based upon the level of support and confidence it generates. Support is the number of transactions where products A and B are bought together. Confidence is the number of transactions in which if product A is bought then product B is also purchased. We can set a controlling threshold value of these two measures to find the best combinations of products. In this paper, they suggest association rules which satisfy minimum support and confidence threshold are considered best. "the total number of possible association rules, R is exponential to the number of items, d" [1]. The results of their market basket analysis suggest that different threshold values will generate various association rules as we set a low threshold value it becomes more strict and if we set a high threshold value then it becomes wider. Market basket analysis is very helpful in the retail industry for building a strong product placement strategy to generate more profit and sales.

Loraine Charlet Annie M.C. and Ashok Kumar D.in their paper "Market Basket Analysis for a superstore based on Frequent Itemset Mining" shows the usage of the k-apriori algorithm for generating frequent itemset for a particular superstore in comparison to apriori algorithm. Data mining is finding interesting patterns from databases using different techniques like association rules, clusters, correlations, and sequence classifiers but association rule is one of the most popular technique. The association rule helps us to find correlations and patterns in a huge amount of data. "Association rules are derived from the frequent itemsets using support and confidence as threshold levels. The sets of items which have minimum support are known as Frequent Itemset. The support of an itemset is defined as the proportion of transactions in the data set which contain the itemset. Confidence is defined as the measure of certainty or trustworthiness associated with each discovered pattern. Association rules derived depends on confidence" [2]. In this paper, they have used Anantha Superstore located in Tirunelveli city data for doing market basket analysis using frequent itemset mining. This store has multiple categories of items like household, bakery, kitchen wares, and fruit-vegetables out which household has the highest number of products of different brands and quality. Their focus is to find the best association rules or product combinations for a household category by applying apriori and k-apriori algorithms. Apriori works based on different support and confidence generated using frequent itemset. K-apriori algorithm separates customers into various categories(groups) or clusters then finds frequent itemset and association rules for each cluster individually. If the number of the cluster (transactions) increases the number of association rules and frequent itemset also increases. In "clustering efficiency is measured using the popular metrics like Inter-cluster and Intra-cluster distances."[2]. Inter-cluster means the sum of the distance between clusters is maximize and intra-cluster means distance should be minimized. Their results show that for Anantha store data k-apriori algorithm gives better results compare to apriori because of their different groups of items and also this algorithm will help them to improve their product placement strategy to increase more sales. Also, it proves that the higher the number of support and confidence stronger the rule is.

Michael Schaidnagel, Christian Abele, Fritz Laux, Ilia Petrov in their paper "Sales Prediction with Parametrized Time Series Analysis"[3] exhibits the prediction of sales, here they include not only the historical sales data but also the future cost of products that will bring variations in the sales quantity. Sales prediction is an essential objective for any analysis depends upon the time series. The motive is to forecast the historical sales data to obtain the sales quantities for the next coming 14 days that can be achieved by expanding the time series calculations and analysis towards the future. Moreover, the dispute is, if the forecasting accuracy is determined to be low, then it is more important to find the data that can build "Similar" products or situations and form clusters of products for developing a prediction model. In this paper, they made utilization of two major analyzing multiple time series techniques like Vector Autoregressive (VAR) models which are also known as Vector ARIMA Model (Autoregressive Integrated Moving Average) and the Parametrized Prediction Model. In addition to this, the most essential factor is not only the sales history but also the price for looking into the prediction accuracy rate, also to forecast the results they need to take a long history, sufficient data and a variable parameter named, "price" that need to be correlated with the sales, so the dimensions used for ARIMA Model are price, product and time. They utilized the Eclipse with Java platform and for data storage used MYSQL Apache webserver. Further, when the ARIMA prediction gave an accuracy of only about 47%, they used another way and made a comparison between them. A new algorithm or method, The Parametrized Time Series model for Sales forecasting, which is also known as sales prediction algorithm Fr, while implementing this algorithm, 2 parameters were taken into considerations: hidden periodicity and predefined future price. In comparison to ARIMA, this Model provided an improvement of 26.7% in total. The results of both models suggest that Parametrized Time Series Model forecasted the sales quantities for a huge data which is 20% more accurate than ARIMA Model. Though the Parametrized Model is more suitable for future predictions of huge and large data, ARIMA Model

can also be accurately utilized and is considered as the best possible option for some finite amount of data and the limited amount of future forecasting values.

Manisha Gahirwal and Vijayalakshmi M. in their paper "Inter Time Series sales Forecasting"[4] They are presenting a research on the improvement of forecasting accuracy. Mainly their focus is on how the best model of only one series can be applied to similar frequency pattern series for forecasting using association mining, in the beginning, they have explained in detail the forecasting, its different variations or types, and the detailed usefulness of how to use them. In this paper, they are trying to combine the forecasting and its results for enhancing the quality of forecast from various models rather than utilizing a single prediction. The decomposition of the original series into multiple components and then implementation of separate forecasting over each component such as trend, seasonality, and an irregular component is been undertaken. In addition to this, Models like ARIMA, and Holt-Winter is been used, also the ARIMA predicted value is further compared with the Holt Winter Method and the respective results achieved are better than Holt Winter Method. Other than combining the forecast methods, Error Measures are calculated to estimate which method predicts the best. Time series Association rule mining is most essential in various fields and better research techniques of data mining, that extract the forecast patterns, interesting correlations, and much more which is applied over the series to obtain the best and the bad models along with combined forecast values that are calculated and MAPE is achieved. They are using 30-time series they first calculated the forecasted ARIMA MAPE value and then obtained the Holt-Winter MAPE value and further made a comparison between them, which showcases that the ARIMA MAPE has a better value than the Holt-Winter MAPE value. In addition to all of this comparison, they performed the inter-time series and proved that how the results of best models from one series have the same results as the other series of the same frequency.

"Data mining using RFM Analysis"[5] paper written by "Divya D. Nimbalkar, Asst Prof. Paulami Shah." Has performed similar work like ours as they have calculated the RFM Scores to find the loyalty of a customer through which the business owners can understand their customer behaviour. They have also used the CRM i.e. Customer Relationship Management data to perform this analysis. They have compared different techniques to find the customer relationship by using the customer value analysis where we can find the RFM Model. They have also used the clustering and classification method as well to find out through which method they can find the best results. In the classification technique, they have also explained about the tree-based algorithms how they work and what other algorithms can be used like LEM2, C4.5 decision trees, C5 decision trees, and TOPSIS. They have used proposed methods in different papers and find out which method is the best for customer segmentation and found out that RFM analysis is the best to find the customer values.

"Clustering optimization in RFM analysis based on k-means" paper written by "Rendra Gustriansyah, Nazori Suhandi, Fery Antony". This research is performed to find the optimal number of clusters that should be used in the K-means algorithm to perform RFM analysis. They have a real-world dataset of a pharmacy a year to perform this research. They have used the classic way of calculating the RFM Scores and divide them into 5 segments of 20% each of all customer data. Then they have used eight different types of indexes to validate the optimal number of clusters (k) and they are Elbow Method[13], Silhouette Index, Calinski-Harabasz Index, Davies-Bouldin Index, Ratkowski Index, Hubert Index, Ball-Hall Index, and Krzanowski-Lai Index. Then they have plotted them to check the different optimum values of clusters (k) provided by the different indexes then they have used the voting technique to select the number of clusters. To evaluate their results they have used the Cluster Quality Testing that is by evaluating the value of Variance (R) "R is the ratio value between the average distance of data in the same cluster (intra-cluster distance) and the average distance of data in the other clusters (inter-cluster distance)" [6] and when the value of R is close to 0 means our clusters are of good quality. This research was purely focused on performing better clustering.

III. METHODS

In this section of our report, we will try to explain how data was selected, a basic understanding of our data, how we have performed data wrangling to improve the usage of different features. In this section, you will see how we have performed Market Basket Analysis, Customer Segmentation, Sales Forecasting. This section will also show us that how different machine learning and time series forecasting models.

A. Data Selection

One of the crucial decision is data selection. As our topic suggests about business improvements then we should have data through which we can perform various analysis to project how business owners can grow. Hence we have used the open data of the superstore in the United States which is provided by a company called Strategy Titan [7]. Through our data, we can perform various analysis related to prediction, forecasting, data mining, data visualization.

B. Understanding Data

There are approximately 160k rows and 24 columns in the dataset where each row represents the order placed by a customer which includes Profit Ratio, Discount, Customer Demographics, Product Name, Manufacturer Name, etc. Apart from this, the other main features are Freight Bill, Freight Gain/Loss, Freight Recovery, and Channel from which the customer has purchased the product. These are very crucial features for making data-driven business decisions.

Also, with the help of this data, we were able to achieve our other analytical goals like by using a huge amount of product details to perform Market Basket Analysis, with customer details to perform RFM analysis, by sales we have done forecasting of future Sales.

C. Data Cleaning

In data cleaning, we are focusing on nulls, error values like if we have any alphabets in any integer features or any special characters are present in our data or inappropriate numerical values like if there was a sales are done but the value of that sale is 0 we have to take care of such problems, we check the data quality or it's data health by performing data profiling, checking nulls, checking the datatypes.

By performing the above techniques we found that our data types were perfect, no as such problem with data profiling, when we checked the nulls then we found that we were having no nulls. But when we have done the data profiling then we have realized that we have some sales values with 0. In the real world, when there is a sales performed then definitely we know that the value of sales cannot be 0. Then we have checked that in how many samples we have any sales value that is 0 or less than it that is negative values of sales. After filtering such samples we have realized that there is a total of 16 samples that have such a problem associated with it out of 159904. Then we have checked what is the percentage of such error based data with respect to our whole data. We found that data the weight of 0.1% to our whole data in that situation we can just drop such samples as it carries very low weight compared to our dataset.

D. Data Wrangling

To perform RFM Analysis and Market Basket Analysis we need certain details of Customer like Customer ID which is unique for every customer. We also need the unique Product ID for all different products. In our dataset, we were not having such information hence we have created Customer ID and Product ID for each unique customers and products. We have used excel to produce unique values of Customer ID and Product ID by using the PIVOT TABLES to create 2 different sheets in excel in which one had the unique customer name with its unique Customer ID. In the other sheet, we had the unique product name with its unique Product ID. Then we have used the joins in python to assign the respective values of Customer ID and Product ID to their customers and products respectively. Hence by using the joins in python we have added 2 new features to our data frame that are Customer ID and Product ID.

- **Note: Now we will explain each different analysis one by one where you will understand different processes and methods related to the respective analysis.**

E. Methods for Different Analysis

MARKET BASKET ANALYSIS

In the retail industry every day lots of customers come and buy which eventually generate tons of sales or transactional data. For example, in a superstore where the everyday customer comes to buy their groceries for each purchase unique bill gets generate. This type of sales data contains all purchase details such as like day of purchase, product names, product price, quantity, total sales, total profit, and other demographics data related to the store. Market basket analysis is one of the data mining technique which helps any organization to find hidden patterns in their sales data. Hidden patterns such as finding the best combination or mix of products which two products are bought together more frequently. Based upon this organization can make their product marketing strategies by making combos of two more products together or making buy one get one offers likewise that helps to increase the sales of the organization.

1. Data Selection

After doing data cleaning and wrangling as mentioned above we have prepared our data for applying market basket analysis. We have applied market basket analysis on different sets of a dataset like on overall data then filtering data based upon product categories like furniture, office supply, parts, and technology filtering data based upon the channel meaning the platform via customer made purchase like retail and e-commerce. This all selected data contains the only required columns like order id, order date, product id, product name, quantity, price sales which are very important for applying market basket analysis and other unnecessary columns are removed.

2. Data Visualization

Visualization helps us to explore our data better. Sometimes looking at just numbers we can not get insight into data whereby by plotting data using different charts we can get more idea of data and it's more appealing. In our project before applying basket analysis on all different data we have tried to plot this data by creating a bar chart and line chart. Where we are trying to plot the top most frequent

product in that particular dataset with the help of a bar chart, finding the trend of orders during the given period with the help of a line chart, and finding the yearly, monthly, weekly total number of orders placed with a bar chart. Some interesting insights were generated with the help of this for example, in the case of overall data we get to know that the most frequent product is Staple Envelope, order trend was the same over the four years and it has seasonality in it, and year 2019 is having the highest number of orders, the month of November is having the highest number of orders over the four years, and in case of weeks orders are almost constant on each day but highest orders are placed on Saturday.

3. Feature engineering

Feature engineering is a technique for transforming your data to make it suitable for analysis. There are two ways you can do feature engineering one is creating one new feature with the help of the existing feature and the second is to transform the existing feature. In our project for doing market basket analysis, we have transformed our data so that we can further use it for doing basket analysis. Using the order id, product name, and quantity we have done feature engineering first we are grouping all records with order id and product name by the sum of their quantity also replacing column names with product names and order id as an index column. Second, after applying the group by converting all the values less than or equal to zero as zero and greater than or equal to one as one. This whole process we have applied to our every data.

The reason for doing this whole process is to make it suitable for basket analysis. Group by will help us to create the basket of all products that have been placed in one particular order with their total quantity. For example,

Order ID	**Product Name**	**Quantity**
1	A	3
1	B	2
2	C	1

We have this type of data now with the help of group by we are putting all the products of one unique order into one basket with their quantity in it which will look like below.

Product Name	**A**	**B**	**C**
Order ID			
1	3	2	0
2	0	0	1

As we can see there are a total of two unique orders are placed now with the help of the group by we created a basket for both the unique orders with the name and quantity of each product they contain just like in the above example Order Id 1 has two products A and B with their quantity 3 and 2 respectively since it does not have product C it's quantity is shown as 0.

The reason for converting all the less than or equal to zero values as zero and values greater than or equal to one as one because to convert all the values into zeros and ones. Another most important reason is the algorithm which will help us to do market basket analysis only accepts the values of zeros and ones as input. So, after doing the conversion same above grouped data will look like as below.

Product Name	**A**	**B**	**C**
Order ID			
1	1	1	0
2	0	0	1

4. Method For Basket Analysis

There are various methods available for doing market basket analysis such as Apriori, K-Apriori, FP (Frequent Pattern) Growth, and Eclat. All methods help us to do data mining of our data to find hidden patterns in it. In our project, we have used FP Growth for doing market basket analysis.

[8]FP Growth is a tree-based depth-first search method. It is a very fast method for data mining because it does not need to find all products that exist in data and assign a unique key to them. Compare to Apriori methods FP Growth methods use less memory and computational power which is very important because in basket analysis you have to read the whole data. When data is big usage of memory and computational power becomes very crucial.

After applying feature engineering on data we have used the FP Growth method to find the hidden patterns or finding the best combination of products. While applying the method we have to pass one value to it which is called Support that tells us that we want to create combinations from those products whose frequency is desired by us. After passing this value our method will look into the whole data and give us all possible product combinations or rules with their support value. Apart from this, it will also give us other measures such as lift, confidence, leverage conviction. At last, we filter out the strongest product combinations or rules based upon the value of lift and confidence set by us.

Now let's understand this whole process with sample data for example,

Product Name	**A**	**B**	**C**
Order ID			
1	1	1	0
2	0	1	1
3	1	0	1
4	0	0	1
5	1	1	0

After this, we apply the FP growth method to this data where we pass the support value let's say 0.6 and this will generate below product combinations and rules.

Antecedents	Consequents	Antecedent Support	Consequents Support	Support	Confidence	Lift	Leverage	Conviction
A	B	0.49	0.47	0.61	0.76	1.95	0.68	1.88
A	C	0.45	0.40	0.63	0.66	0.89	0.62	1.27

Once product combinations or rules are created we filter the rules by setting the minimum confidence and lift value. If we take a minimum value of >=0.7 for confidence and >=1 for lift apply to sample data then rules will be filtered out as below.

Antecedents	Consequents	Antecedent Support	Consequents Support	Support	Confidence	Lift	Leverage	Conviction
A	B	0.49	0.47	0.61	0.76	1.95	0.68	1.88

The above rule is the strongest product combination which is generated by the FP growth method. We can tell that whenever product A is bought product B is also bought together. Organizations can put these two products together or make a combo of it or put a discount on them that will help them increase their sales.

Now let's understand that what all these measure values stand for and how they are calculated. Below we will understand them one by one.[9]

Support:

It means that we want to consider the products that come or frequent a certain number of times in data. If we take our above sample data example then we have taken the support value of 0.6 which means we are saying that give me product combination who have minimum support value of 0.6. Let's consider Product A then,

Support = number of times antecedents presents in transaction / Total Numer of Transactions

= 3/5

= 0.6

Confidence:

With this, we are saying that we want only those rules or product combinations whose confidence level is set by us. Generally, it should be 0.7 or 0.8 meaning 70% or 80% confident that product A and product B is coming together. Let's take the example of our sample data then if we want the combination of Product A to Product B then,

Confidence = number of times antecedents -> consequents in transaction/ support of antecedents number of times in transaction

= A->B/ support of A

Lift:

It helps us to filter the rules or product combinations based on its value set by us. It is calculated with the help of confidence and support. Generally, the best value is greater than or equal to 1. Let's take an example of our sample data then,

Lift = confidence of antecedents -> consequents in transaction / support of consequents number of times in transaction

= confidence of A -> B / support of B

Leverage:

The number of times antecedents and consequents both are coming together in the transaction to the probability of if both are independent to each other or not value of 0 means they are independent.

Conviction:

The number of times antecedents and consequents both are coming together in the transaction to the probability of if both are dependent on each other or not value of 1 means they are dependent and inf means infinity times dependent.

For creating strong rules or product combinations we only consider Support, Confidence, and Lift value.

CUSTOMER SEGMENTATION

This helps business owners to understand the value of their customer base. We have 2 different methods to perform customer segmentation. The first one is using RFM Analysis and the second one is Clustering using the RFM Score.

1. Feature Selection

We have used the titan sales dataset from which we have selected required features which are Customer ID, Order ID, Order Date, Sales. We have selected this feature as these features will help us to perform customer segmentation using RFM Analysis and Clustering.

2. Feature Engineering

Before performing feature selection, we just check the number of customers and find their count and we also find the time frame of our dataset. In our feature engineering firstly, we find the total amount of sales of all customers as monetary, to find the frequency of customer we use Order ID, to find the Recency we use the Order Date by subtracting it by present date by aggregating and group by function. Then we just rename all three features as recency, frequency

and monetary respectively. After that, we filter our data for 1 year that is 365 days.

3. Methods For Prediction

We have 2 different methods to perform customer segmentation one is using RFM Analysis and the other is Clustering using the RFM Score.

RFM Analysis

RFM analysis is a customer behavior segmentation technique that is driven by data. RFM stands for recency, frequency, and monetary value. The idea is to segment customers based on when their last purchase was, how often they have purchased in the past, and how much they have spent overall a time interval.[10]

We divide our data into 5 equal quantiles consisting of 20 % data into each quartile in increasing order. Then we create the function to assign the values from 1 to 5 where 1 being the least and 5 being the best score of recency, frequency and monetary scores. Once we assign the scores to recency, frequency and monetary as R, F, M columns then we add those columns to our data frame. After that, we combine the R, F, M scores into one column called RFM Score from which we can decide which customer goes to which segment. Once we are done with the previous step then we create a final column called segment by mapping different R and F scores to perform different customer segments as 'at-risk, 'can't loose', 'need attention, 'loyal customer', 'champions', 'new customers'. At last, we visualize the results to find in which segment we need more attention or can also find the best customers as well.

Clustering

Before performing clustering to find different customer segments we extract the RFM data frame and then again load it for our cluster. As we use Recency, Frequency and Monetary values to build our clusters.

We visualize Recency, Frequency, Monetary values then we find that our data is skewed then we plot the box plot to find if we have any outliers in our data. The box plot shows us that we do have outliers present in our data and then we use a z-score to handle outliers by considering 6 standard deviation that is we take 99% of data, so we use $z < 3$ to remove outliers. We have removed outliers to get better prediction results as we are applying the machine learning algorithm for clustering. Then we do apply our clustering algorithm that is K-means Clustering to our uncleaned data to have a look at how it performs the clustering on uncleaned data and then on cleaned data to check how outliers affect our analysis. Let's first understand what is [12]K-means clustering? K-means clustering is the unsupervised machine learning algorithm that helps us to label our data and to categorizes our data into different categories. It works iteratively to find the optimal centroid while computing the centroid. In this algorithm, we have to provide the number of clusters to compute centroids. Basically in this algorithm, it tries to find the minimum sum of squared distance between the data points and centroid, we know that less variation within the clusters will allow more similar data points within the same cluster. But now, here comes the main question of how to find the optimal number of clusters? To answer this question there are various techniques available to find the optimal number of clusters that is the best value of K. In our business problem we have used mainly 2 methods first which is the most common method used to find the optimum value of K is Elbow Method for K-means clustering. In the elbow method, we visualize the cost function produced by different values of K in a plot. We know that if the value of K increases then average distortion will decrease and by doing so each cluster will have fewer constituent instances and those instances will be closer to their centroids. So as the value of K increases then average distortion will increase. The value at which the improvement in the distortion declines the most will look like an elbow in the plot and is called elbow at which we should not divide our data into further clusters. In our business problem, we have used 2 different values of K in our analysis that is 3 and 4 to find the best results.

We have also used [11]Hierarchical Clustering to make customer segmentation on clean data only to check which clustering algorithm works better. The hierarchical clustering falls into 2 categories one is the Agglomerative Hierarchical Algorithms and the other is Divisive Hierarchical Algorithms. We have used the Agglomerative Hierarchical Algorithm in our use case so let's understand what is it? In agglomerative hierarchical algorithms, we are treating each data point as a single cluster and then use the agglomerate (bottom-up approach) or successively merge them as pair of clusters and the hierarchy which is formed in this process of clustering is represented as a dendogram or tree structure. Now we have the same question that how we can find the optimum number of clusters? Here comes the second method to find the optimum number of clusters by examining the dendrogram. It is very simple to find the number of clusters from the dendogram we just have to look around the longest vertical line this is formed by the biggest cluster. Then we draw a horizontal line through it and check that at how many point our horizontal line is passing. The number of points through which the horizontal line passes are the optimal number of clusters in the Agglomerative Hierarchical Algorithm. In our case, we have found 2 clusters from the dendogram.

SALES FORECASTING

1. Outlier Removal

Z-score is in general also known as Standard Score that is measured in terms of Standard Deviation from

the mean. Z-score is a statistical as well as a numerical measurement that determines the relationship amongst the value and means of a group of values. A Z-score 0 specifies that the data point's score is the same as the mean score. However, the Z-score 1.0 implies that the value is one standard deviation from the mean score. The basic Z-score formula:

$$z = \frac{X - \mu}{\sigma}$$

Fig.1 Z-score[15]

In our project, we calculated the Z-score for Titanized Sales Dataset. Boxplot[18] method is a graphical representation of numerical data through their quartiles. Boxplot can also have lines that are extended from the boxes known as, whiskers which implies the variability beyond the upper and lower quartiles. **Outliers** are plotted as individual points or data points. The spacing between the Boxplot specifies the degree of dispersion and skewness and shows outliers. Boxplots are helping in exhibiting the Outliers within a DataSet. An outlier is an observation that is numerically deviated from the rest of the data. When examining a boxplot, an outlier lies outside the whiskers of the boxplot. We removed the outliers from the Dataset for our Sales data.

Time-Series Models [19]

Time Series Analysis, known as Trend Analysis. Time-series forecasting models utilize the information regarding historical data and associated patterns for future predictions. It relates most frequently to the level, trend, analysis, cyclical fluctuation analysis, and seasonality issues. Here in our project, we are using various time series methods to make a future prediction on Sales value like Simple Moving Average, Exponential Smoothing Models, SARIMA Model, and Facebook Prophet. We will explore all the methods one-by-one in depth.

2. Forecasting KPIs [16]

Forecasting accuracy is a Key Performance Indicator. Forecasting KPI shows how well are the future predictions for future demands. This will improve the results and predictions to maximize the profit ratio. Some of the KPIs are explained below that will be utilized in our project:

MSE:

The Mean Absolute Error (MAE) is a really good KPI to calculate the forecasting accuracy level. As the term itself implies, it is the mean of the absolute error.

$$MAE = \frac{1}{n}\sum |e_t|$$

Fig.2 MAE

RMSE:

It is defined as the square root of the Mean Squared Error. It is an essential KPI for forecasting accuracy.

$$RMSE = \sqrt{\frac{1}{n}\sum e_t^2}$$

Fig.3 RMSE

Simple Moving Average [20]

A simple Moving Average is a stock indicator that is mostly utilized in technical analysis. Besides, it is a statistical calculation that takes the arithmetic mean of a given set of prices in the past. It is one of the simplest models, also known as **Rolling mean.** It is calculated by averaging the data of the time series over a specific period. Forecasting is the average of the demand during the last 'n' periods.

$$f_t = \frac{1}{n}\sum_{i=1}^{n} d_{t-i}$$

Fig.4 Simple Moving Average

n = Number of periods we use to take the average of the same.

d = it is the demand we observed during the period t.

f_t = Forecast we made for period t.

For Moving Average, demand will be denoted as d and f for the forecast. In our project, our first step was to implement the Moving Average by determining the essential parameters, that are necessary to accomplish the calculation. For our Titanized Sales Dataset, we have considered the Moving Average throughout **5**, which will consider the mean of every 5 columns and it will roll ahead by one step every time after calculating the means for every 5 columns. The Dataset already had 2 columns, Date and Sales which we require for the calculations. After applying the moving average, we will see one more column, that is Forecast. Then we will plot the Trend of Moving Average Model. In Error calculations for Moving Average, we have to subtract the Forecasted value from the actual Sales value. Then, we need to measure the KPIs for our predicted values. The comparison is based on the equation, MSE > RMSE > MAE. Which remains the same for all the Time Series Models. The Result after implementing this formulation.

3. Model Initialization

For simple initialization, the first forecast period = 0, which is the first demand observation.

$$f_0 = d_0$$

Fig.5 Initial Period

For the other set of periods, which is also called averaging, the forecasting is the average of the first n demand occurrences.

$$f_0 = \frac{1}{n}\sum_{t=0}^{n} d_t$$

Fig.6 Average

Exponential Smoothing Models

Exponential Smoothing is a "**Rule of Thumb**" technique for smoothing the time series data using the exponential function. It is often used for analyzing the data for future predictions. This model is significant for non-stationary data. For our project dataset, we will be using all the three models of Exponential smoothing for making a comparative analysis, like Simple Exponential smoothing, Double Exponential smoothing, and Triple Exponential Smoothing.

Simple Exponential Smoothing

Simple Exponential smoothing(SES) is a time series forecasting method for univariate data without a trend or seasonality. The most essential parameter that is required for performing SES is, called **alpha (a)**. The primary idea of the model is to assume that the future prediction will be more or less the same as the historical data. The Alpha parameter controls the rate. The alpha value is usually set between 0 and 1. The SES model will then forecast the future demand as its last estimation of the level. The Simple Exponential Smoothing formula:

$$f_t = \alpha d_{t-1} + (1-\alpha) f_{t-1}$$

$$0 < \alpha \leq 1$$

Fig.7 Simple Exponential Smoothing

In our analysis prediction for future Sales data, here we will take a **period** of **5** and the **alpha** value is **0.3**. Further, the SES function will be used for making the forecasting and calculating its relevant values. We applied the algorithm for making the predictions, by creating the required arrays, defining the required variables, and implementing the formulas, as necessary. The next step will be, assigning the actual values into the smoothing function. The KPIs for our predicted values are then measured. In the end, the values are plotted on the graph to show the actual data and the forecasted data over time.

Double Exponential Smoothing(Holt)

Double Exponential Smoothing is also known as Holt Method. DES not only includes Alpha but also Beta(β) is another essential parameter that controls the rate at which the influence of the observation at prior time steps decayed exponentially. Double Exponential Smoothing employs a level and a trend component at each period. And use two-weight as mentioned before, for updating the component at each period. The Double Exponential Smoothing Equations:

$$L_t = \alpha\, Y_t + (1-\alpha)\,[L_{t-1} + T_{t-1}]$$

$$T_t = \gamma\,[L_t - L_{t-1}] + (1-\gamma)\,T_{t-1}$$

$$\hat{Y}_t = L_{t-1} + T_{t-1}$$

Fig.8 Double Exponential Smoothing

In our analysis prediction for future Sales data, here we will take a **period** of **5** and the **alpha** value as **0.3** and **Beta** value as 0.4. Further, the DES function will be used for forecasting the values. For this algorithm, two loops are used one for forecasting the historical periods and the other for forecasting the extra periods i.e., 5. The next step will be, assigning the actual values into the smoothing function. The KPIs for our predicted values are then measured. In the end, the values are plotted on the graph to show the actual data and the forecasted data over time.

Triple Exponential Smoothing(Holt-Winters)

Triple Exponential smoothing models are also known as Holt-Winters Smoothing Model. Seasonality factors are calculated in TES along with a new learning rate called **gamma** (g), which controls the influence on the seasonal component. As with the trend, seasonality can be modeled as either an additive or multiplicative process. In our analysis prediction for Sales data, in TES the data is split into training and testing data. Where the Sales data is divided into two datasets Train, which contains 70% of the data, and Test, which occupies the remaining 30%. For Triple Exponential, the period is 5, Alpha value is 0.3, Beta value is 4.0, and the seasonal period is 12. Then the KPIs calculation was then executed. In the end, the values are plotted on the graph to show the actual data and the forecasted data over time.

SARIMA Model [17]

SARIMA Model stands for Seasonal Autoregressive Integrated Moving Average Model. It is one step different from an ARIMA Model based on the concept of the seasonal trend. In many time-series datasets, frequent seasonal effects have come into the picture. SARIMA is ARIMA Model with a Seasonal component. The formulation equation is SARIMA(p,d,q)x(P,D,Q,s). Initially, for the Sales dataset, we plotted the graph to visualize the data into the plot. This leads us towards the examination of different parameters by plotting data, decomposing the parameters, and study each residual. By using the 'sm.tsa.seasonal_decompose' command Time Series can be decomposed into three distinct components like Trend, Seasonality, and Noise.

4. Augmented Dickey-Fuller Test

Using this test, we can visualize the stationarity of data. It is also called as Unit Root Test. ADF tests the null hypothesis that a unit root is present in a time series data. The alternative hypothesis is different depending on which version of the test is utilized but is mostly stationary or non-stationary.

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t,$$

Fig.9 Augmented Dicky-Fuller Test

For Sale Dataset, ADF test algorithmic function was used to check the stationarity. The test is conducted on basis of two values, P-Value and ADF Statistics value. The Sales data is considered to be stationary if the p-value is low as per the Null-Hypothesis and the critical values at 1%, 5%, 10% confidence levels are as close as possible to the ADF Statistics. After, performing the test, it is seen that there is no need to perform extra steps to make the data stationary as it is already in the Stationary form. Further, the AIC combination model was calculated for fetching the best fit values for SARIMA Order and Seasonal Order parameters. In AIC value selection, according to Peterson T.(2014) the AIC(Akaike information criterion), the Lower the value of AIC, the better the combination. So, our test output suggests us the best value of ARIMA(1, 1, 1)x(0, 1, 1, 12)12 - AIC:4740.841, as our best combination amongst all in the listing. As we checked before, the best-fitted value for SARIMA Model is SARIMA(1, 1, 1)x(0, 1, 1, 12). We have used the SRIMA Model function and then fitted the model, stored the result, and printed the results in form of SIRIMA Results.

5. Diagnostic Plot

We plotted a diagnostics chart to show the results of the data after applying the SARIMA Model. There are four plots used to showcase the condition of the data. From the normal Q-Q plot, we can see that we almost have a straight line, which suggests no systematic departure from normality. Also, the Correlogram plot, which is the last plot on the bottom right suggests, that there is no autocorrelation in the residuals, and so they are effectively white noise. The value of the graph is near 1 demonstrate the Strongest Correlation.

6. One-step ahead Forecast

The one-step-ahead forecasting value demonstrates how we forecasted for the upcoming days for giving an appropriate Sales prediction for future utilization of the data. Further, we made the predictions and calculated the values 12 steps ahead of the Actual data. Where we can visualize the lower and upper values which the model indicates as boundaries for the forecasting. It can be observed that Forecasted Sales is increasing steadily along with some fluctuation.

Facebook Prophet Model

Facebook Prophet is an open-source software launched by Facebook's Core Data Science team. It is a procedure for forecasting time series data based on an additive model in which non-linear trends are fit with annual, weekly, and daily seasonality, as well as holiday effect. At its core, the prophet has general ideas that are similar to a generalized **additive regression model** with four main components. With Prophet, you are not stuck with the results of a completely automatic procedure if the forecast is not satisfactory. For Sale DataSet, Prophet also implies the strict condition that the input columns be named **ds**(the time column) and **y**(the metric column). We have to change the column names to this respective name in compulsion. Then, further, instantiate a new prophet object. Once the prophet model has been initialized, we can call its fit method with Dataframe. In the next step, the Prophet will generate the 165 datestamps in the future. Where we have to define these datestamps as periods =165 like an argument. The Dataframe of future dates is then used as an input to the prediction method of our fitted model. Besides, we have further plotted the fitted model representing the future forecasted values. Prophet relies on Markov chain Monte Carlo (MCMC) methods to generate its forecasts. MCMC is a stochastic process, so values will be slightly different each time. The data have no weekly seasonality. However, it has a Trend that shows a positively rising slope.

IV. RESULTS

MARKET BASKET ANALYSIS

Using the FP Growth method on a different set of datasets we have generated different rules or product combinations. Below is the table which shows the top product combination with given support, confidence, and lifts value.

Table 1. Results of Market Basket Analysis

Type Of Data	Support	Confidence	Lift	Rules Generated	Top Product Combination
Over All	0.001	>=0.9	>=6	5	(Pierce Parts - Replacement Kit 3, Pierce Parts - Replacement Kit 5) & (Pierce Parts -

					Replacement Kit 4)
Furniture Category	0.0004	>=0.9	>=6	319	(DAX Two-Tone Rosewood/ Black Document Frame, Desktop, 5 x 7, Global Deluxe Stacking Chair, Gray) & (Global Value Steno Chair, Gray)
Office Supply Category	0.0002	>=0.9	>=6	32,514	(Holmes Replacement Filter for HEPA Air Cleaner, Very Large Room, HEPA Filter, Eldon Fold 'N Roll Cart System) & (Acco Pressboard Covers with Storage Hooks, 9 1/2" x 11", Executive Red)
Parts Category	0.001	>=0.6	>=6	19	(Pierce Parts - Replacement Kit 3) & (Pierce Parts - Replacement Kit 4)
Technology Category	0.0005	>=0.8	>=6	304	(Apple iPhone 5, Logitech LS21 Speaker System - PC Multimedia - 2.1-CH - Wired) & (SanDisk Cruzer 64 GB USB Flash Drive)
Retail Channel	0.0009	>=0.8	>=6	2	(Hoses, Water Filters) & (Replacement Element)
E-Commerce Channel	0.002	>=0.6	>=6	10	(Car Light) & (Replacement Element)

From the above results, we can say that the highest number of rules are generated for the Office Supply product category, and the strongest rules are generated for overall data because we set the support value of 0.001, the confidence of more than 90%, and lift value more than of 6.

CUSTOMER SEGMENTATION

The below table will show us the results of customer segmentation performed by RFM analysis. Where we will understand the business performance and future business techniques to improve the customer base. This table describes us that this company has an alarming situation as the maximum of their customer does not belong from "champions" and "loyal customers" as they are less than 50% among all customers, and if we talk about "at risk" customers they are maximum in the proportion that is 33% among all. Their "need attention" group of customers are just 12% less than loyal customers which do not sound great for the company.

Table 2. Results of RFM Analysis

Customer Segments	**Customer Count**	**Percentage of Total Customers**
Can't Loose	42	33%
New Customers	43	26%
Need Attention	86	15%
Loyal Customers	106	12%
Champions	183	6%
At Risk	236	6%

Now we will understand the results of clustering for customer segmentation. We have used two different unsupervised machine learning models to make segments of our customer base. Firstly, we will discuss the results of the K-means clustering technique we had used the elbow method to find the optimal number of customer segments. By using it we found that we can use an optimal number of 3 clusters to segment our customers. We have also used hierarchical clustering to perform segmentation in which we have used the dendogram to find the optimal number of clusters where we found that we have our model suggests us 2 clusters as the optimum number of clusters. We have also visualized the results of both the clustering techniques and realized that we can easily differentiate the clusters formed for Recency & Monetary, Frequency & Monetary but cannot differentiate the clusters between the Recency & Frequency. The reason behind not able to differentiate between Recency & Frequency is due to the way the data is distributed in these 2 features. We have also performed model evaluation by using the [14]Silhouette Score which indicates that the optimal score for the number of clusters is 3. Hence we can say that the K-means clustering works better on our data to perform customer segmentation.

SALES FORECASTING

Once the formation of all Time Series Models is done, a comparative analysis is made amongst them. Based on this analysis, we have compared the RMSE value and MAE value for the Models, and as mentioned before the lower the KPI values the Better the Model for making future predictions and Generating the Forecasting values for the Sales data.

Table 3. Results of Model Error comparisons

Index	Models	RMSE	MAE
0	Moving Average	74078.34	56091.44
1	Single Exponential Smoothing	87203.09	65446.89
2	Double Exponential Smoothing	102494.12	81070.49
3	Triple Exponential Smoothing	400494.72	362597.82
4	Facebook Prophet	54650.41	231496.39

Earlier we have shown the combined plot for comparative analysis between moving average and exponential smoothing methods, which saw that moving average and double exponential smoothing is the good fit model providing us with good forecasting values. However, after performing the SARIMA Model and Facebook Prophet Models, we can also observe that both models are better fitted than the previous ones. The SARIMA results are showing a clear picture in the image, why it is the best fit model.

SARIMAX Results

Dep. Variable:	Sales	No. Observations:	213
Model:	SARIMAX(1, 1, 1)x(0, 1, 1, 12)	Log Likelihood	-2366.420
Date:	Mon, 12 Apr 2021	AIC	4740.841
Time:	16:45:32	BIC	4753.744
Sample:	12-27-2015 - 01-19-2020	HQIC	4746.069
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.7541	0.111	6.817	0.000	0.537	0.971
ma.L1	-0.9687	0.063	-15.445	0.000	-1.092	-0.846
ma.S.L12	-1.0335	0.053	-19.613	0.000	-1.137	-0.930
sigma2	1.046e+10	5.68e-13	1.84e+22	0.000	1.05e+10	1.05e+10

Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	9.09
Prob(Q):	0.94	Prob(JB):	0.01
Heteroskedasticity (H):	1.43	Skew:	0.26
Prob(H) (two-sided):	0.16	Kurtosis:	3.95

Fig.10 SARIMA Results

As their results are comparatively higher in terms of KPIs measurement as well as the future prediction for forecasted values. In addition to this, we have seen all the plots and charts for different analytical models, and it makes it easy for us to predict and find which model suits them best and which is worst, and for what reason the respective models are worst and best.

TABLEAU REPORTING

we have used the Titanized Sales data and the data output from the RFM analysis together to answer various business questions. We have built different visualization charts representing the direction of the business from the different factors which are features. We have also built different dashboards which help business owners constantly monitor and data-driven decisions for their business improvements. From our dashboards, anyone can understand the business on the micro-level scale. Our analysis help to understand the importance of different customers and their needs based on their needs of geography.

V. DISCUSSION

The results which we have achieved in market basket analysis are good. One can make use of our analysis and methods to solve their business problem. They can make various product placement strategies for improving sales. After researching other works we can say that we have achieved an optimum level of results in our project. The main objective of optimizing the business processes as a whole with the help of our basket analysis one process is optimized.

There are a lot of failures and challenges you face when you want to achieve something big. We also faced a few failures and challenges while doing market basket analysis such as with this big data using Apriori methods was making the whole process very slow. We tried to improve it but were not able to make it fast. To overcome this failure we learned about a new method FP Growth through which we eventually able to improve the speed of market basket analysis with more accurate results.

We believe that the results obtained for customer segmentation using the RFM analysis are pretty

important as we can understand which customers are driving our business in which direction based on the RFM Score and if we talk about the clustering then we should further work and perform more research on finding that how we can find better clusters for the Recency and Frequency of the customer as our both clusters were not able to differentiate the clusters in this case. This is one of the nodes from which this project can be extended to perform better customer segmentation. There are many more opportunities to perform different analysis on this data as there are many more features which can be explored to performed data mining algorithms to find insights. We can say that what every we have learned in our academic we have implemented in an organized and appropriate way in terms of Customer Segmentation.

In the beginning, after the data preprocessing, when we started with Model implementation, we noticed one issue with all the plots and charts of Time Series Models. All the graphs were showing the inappropriate flow of the data which was not clean and having some sort of error values. Which we solved using the Normalization method, by performing the Z-score to remove outliers. Moreover, we were experiencing some problems with SARIMA Model, where we were not getting the RSME and MAE values for the same. Later, after performing extensive research on how to find different KPI measures for the SARIMA Model with a stationary dataset. We found that it is very difficult to find those KPIs while performing SRIMA on the Stationary dataset. Hence, we have used old-school methodology to find the best performing model by observing the graphs of various models. It is not impossible to find those values, but due to time constraints, we were not able to find those values.

VI. CONCLUSION

In this project, we have found that how businesses can optimize and improve their ways of conducting business in different ways. Our analysis shows the business owners that how they can use the product placement techniques and make recommendations for their e-commerce platforms by performing market basket analysis to find the best combination of products through which they can increase their sales.

In the case of market basket analysis, no result is good or bad. It depends on the organization how they want to build their marketing strategies. In general, your rules or product combinations should be strong enough that if you use them in the real world that will help to generate profits or revenues.

We have also used customer segmentation analysis through which we have shown that how customers can be divided into different segments by shopping habits. RFM Analysis helps us to find customer segments by their RFM score which have performed manually using logic and we found that we have 5 different segments. Clustering helps us to find the segments based on the values of recency, frequency and monetary values based on the clustering technique where we found that the optimum number of clusters should be 3 that is 3 different customer segments and the best model will be K-means Clustering. This analysis will help the marketing team to how they should target their customer base.

We have also performed sales forecasting by using time-series models to predict the sales of our business and found that the SARIMA model is the best model to forecast the sales in our business problem. The SARIMA model is most suitable based upon the analysis results in comparison to other models, which can be observed from the graphical plotting. Moreover, it is the best model for implementing analysis and forecasting on large datasets. The sales forecasting will also help the business owners to think about how to plan their supply chain, inventory, and men power at their business at different times of the year.

VII. CONTRIBUTIONS

During Project:

1. **Kushal Patel:** Identified business problems then worked on data selection. Created analytical goals for customer segmentation problems. Found related work regarding Customer Segmentation using RFM Analysis and Clustering. Then I have performed Data Wrangling to add the Customer ID and Product ID features to the data. After that, I have worked in RFM Analysis and Clustering to perform customer segmentation. In the end, I have also performed the reporting and answering the business questions by using visualization and an interactive dashboard. I have also merged the customer segment data extracted from RFM Analysis with the original data to merge with the titanized sales data which is our original data to find more insights and answer business questions.
2. **Viraj modi:** Identified business problems then worked on data selection. Researching about market basket analysis. How we can perform market basket analysis in our project. Understanding the various methods of doing market basket analysis and its measures like support, confidence, lift. Created analytical goals for market basket analysis. Performed the data cleaning part in which I have tried to handle the null and garbage values. Executed the whole market basket analysis on different sets of data with the basic exploration of each data with the help of data visualization. Applied trial and error approach for selecting best support, confidence, and lift value to generate strong rules or product combinations. Gave suggestions and feedback for the reports and dashboards generated with the help of tableau also helped in formatting each

chart and dashboard properly. Equally contributed in all other assessments of the project.

3. **Vishakha Parab:** Identified business problems then worked on data selection. Created analytical goals for sales forecasting. Performed extensive research to understand time series forecasting and its associated different forecasting models. Found related work to understand how different people have used forecasting techniques to predict upcoming sales. I have also handled the outliers to normalize the data required for time series forecasting. Implemented different statistical and time series forecasting models to predict future sales. I have also performed a comparative analysis of different models which I have used to find the best forecasting model.

Report Writing:

1. **Kushal Patel:** Related Work related to customer segmentation where I have summarized two research papers related to customer segmentation, Problem Statement & Objective, Proposed solutions and methods, results, discussion, conclusion, and references related to customer segmentation. I have worked on all the parts of the report that belongs to customer segmentation.
2. **Viraj Modi:** Related Work related to market basket analysis where I have summarized two research papers related to market basket analysis, Problem Statement & Objective, Proposed solutions and methods, results, discussion, conclusion, and references related to market basket analysis. I have worked on all the parts of the report that belongs to market basket analysis.
3. **Vishakha Parab:** Related Work related to sales forecasting where I have summarized two research papers related to sales forecasting, Problem Statement & Objective, Proposed solutions and methods, results, discussion, conclusion, and references related to sales forecasting. I have worked on all the parts of the report that belongs to sales forecasting.

Note: On Title, Abstract, Problem statement, and objective every group member has worked together.

VIII. REFERENCES

[1] Raorane A.A., Kulkarni R.V., and Jitkar B.D, " Association Rule – Extracting Knowledge Using Market Basket Analysis", International Science Community Association, January 2012.

[2] Loraine Charlet Annie M.C., Ashok Kumar D, "Market Basket Analysis for a superstore based on Frequent Itemset Mining", International Journal of Computer Science, September 2012.

[3] Michael Schaidnagel, Christian Abele, Fritz Laux, Ilia Petrov, "Sales Prediction with Parametrized Time Series Analysis", Research Gate, January 2013.

[4] Manisha Gahirwal, Vijayalakshmi M., "Inter Time Series sales Forecasting", Research Gate, March 2013.

[5] Divya D. Nimbalkar, Asst Prof. Paulami Shah, "Data mining using RFM Analysis", Research Gate, December 2013.

[6] Rendra Gustriansyah, Nazori Suhandi, Fery Antony, "Clustering optimization in RFM analysis based on k-means", Indonesian Journal of Electrical Engineering and Computer Science, 2020.

[7] Krantz, J. (2020, December 12). Strategy Titan. Retrieved from www.strategytitan.com: https://www.strategytitan.com/blog/titanized-real-world-dataset-to-develop-your-analytics-muscle

[8] Kumar, A. (2020, July 3). Unfold Data Science Retrived From www.youtube.com: https://www.youtube.com/watch?v=qaTC0Y_evbk

[9] Kumar, A. (2020, March 10). Unfold Data Science Retrived From www.youtube.com: https://www.youtube.com/watch?v=4OORU17o1GY

[10] Makhija, P. (2020, August 20). clevertap Retrieve from www.clevertap.com https://clevertap.com/blog/rfm-analysis/

[11] Bock, T. (2018, August 12). displayr Retrieve from www.displayr.com https://www.displayr.com/what-is-hierarchical-clustering/

[12] Dr. Garbade J, M. (2018, September 12). towardsdatascience Retrieve from www.towardsdatascience.com https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

[13] Dangeti, P. (2017, July 21). oreilly Retrieve from www.oreilly.com https://www.oreilly.com/library/view/statistics-for-machine/9781788295758/c71ea970-0f3c-4973-8d3a-b09a7a6553c1.xhtml

[14] Bhardwaj, A. (2020, May 26). towardsdatascience Retrieve from www.towardsdatascience.com https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c

[15] Medic, S. (2017, January 13). statsmedic Retrieve from www.statsmedic.com https://www.statsmedic.com/post/interpret-the-z-score-like-it-s-your-job

[16] Vandeput, N. (2019, July 5). towardsdatascience Retrieve from www.towardsdatascience.com https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d

[17] Graves, A. (2020, January 7). towardsdatascience Retrieve from www.towardsdatascience.com https://towardsdatascience.com/time-series-forecasting-with-a-sarima-model-db051b7ae459#:~:text=As%20a%20quick%20overview%2C%20SARIMA,lags%20of%20the%20stationarized%20series

[18] Mcleod, S. (2019). simplypsychology Retrieve from www.simplypsychology.org https://www.simplypsychology.org/boxplots.html#:~:text=Box%20plots%20are%20useful%20as%20they%20show%20outliers%20within%20a,whiskers%20of%20the%20box%20plot

[19] Hayes, A. (2021, April 21). investopedia Retrieve from www.investopedia.com https://www.investopedia.com/terms/t/timeseries.asp#:~:text=Time%20series%20forecasting%20uses%20information,analysis%2C%20and%20issues%20of%20seasonality

[20] Fernando, J. (2021, January 17). investopedia
Retrieve from www.investopedia.com
https://www.investopedia.com/terms/m/movingaverage.asp#:~:text=A%20simple%20moving%20average%20(SMA,%2C%20100%2C%20or%20200%20days

IX. APPENDICES

Our whole project is in one python notebook file where you can just use run all and you will be able to find every piece of code running and giving you the results.
DAB402_Group_7_CAP_Project.ipynb

Our data reporting and dashboards can be viewed in our single tableau file where you must locate the data to look our work.
DAB402_Group_7_CAP_Project_Final.twb