****

Individual Project Proposal

ALY 6980 – Capstone

Janos Mako

Prashant Munoth Lalit

Northeastern University

**Introduction and Background**

Small associations may not be in a circumstance to satisfy the aggregate of their customers, as a rule. It may exhibit hard to meet the necessities of each customer. People do not have vague tendencies, so rarely does one thing satisfy everyone. In this way, various associations may normally get a framework that is known as target advancing. This framework incorporates apportioning the market into pieces and making things or organizations to these segments. A goal exhibiting framework revolved around the customers' necessities and requirements. Hereafter, a fundamental for the progression of this customer-driven framework is the detail of the target business areas that the associations will try to serve. The exhibiting chiefs who may consider using target promoting will regularly isolate the market into social affairs (areas). By then, they center around the most gainful ones. They may change their promoting mix parts, including things, expenses, channels, and restricted time systems to suit the essentials of individual social occasions of clients. (Camilleri, 2017)

In e-commerce, where rivalry is intense and clients' inclinations can change rapidly, organizations need to section clients and target promoting activities adequately. The process of segmentation and targeting is successful if the clients gathered into a similar portion show similar conduct and response to promoting efforts. Nonetheless, the connection between segmentation and targeting is regularly absent. Some exploration commitments have as of late tended to this issue, by proposing ways to deal with fabricating client conduct models in each section. Be that as it may, clients' conduct can change with the specific circumstance, for example, in numerous e-commerce business applications. In these cases, building relevant models of conduct would give better prescient execution and thus, better focusing on. Nonetheless, the issue of remembering setting for a segmentation model, and building a predictive behavior model of each fragment reliably is yet an open issue. This examination targets giving a response to the accompanying exploration issue: how to remember setting for a segmentation model to assemble a viable predictive model of client conduct of each fragment. (Faraone, 2012)

**Problem Statement**

In this project we will discover the moving items for the eCommerce business, at that point, we will investigate the techniques to make segments that will be utilized for client division of the clients, and afterward, we will zero in on the best way to showcase the items to those sections. We expect to build up a task that can be utilized by small companies to satisfy their customer's prerequisites which is to offer clients various assortments of related items to browse dependent on their buying history to make their experience more gainful and personalized. For this project, the data is provided by our sponsor Quantum Analytica. The data contains details about the products, orders, customers, and inventory.

This venture can be utilized to construct a dashboard that shows the moving items and related subtleties for the association to comprehend the market and item drifts. Further, this undertaking can likewise be applied to incorporating a model that portions clients into various sections which could be utilized to comprehend the client bunches instead of an individual and spotlight on those gatherings for focused advertising.

**Literature Review**

Big Data of online item buys is an arising hotspot for getting clients' inclinations of an item includes for new item advancement. Jian Zhang, Alessandro Simeone, Peihua Gu, and Bo Hong have proposed a framework and related technique for item feature portrayal and clients' inclination expectation dependent on online item buying information. Details and segments of items are initially investigated and the connections between item particulars and segments were settled for characteristics categorization. The clients favored spec, characteristics, and mixes were anticipated for the advancement of new items. (Zhang, 2018)

Segmentation is a basic empowering agent to accomplish business goals and acknowledge advantages to fulfill and address the issues of the clients. The principle point of the client division is to partition the objective market into groups that have similar comparative attributes, necessities, and needs. Himanshu Singh and Dr. Shubhangi Neware have played out a predictive neural network approach dependent on the boundaries like item audits, items, purchasing behaviors, seeing patterns and time-sensitive fragments, and clustering strategies in electronic trade(Singh & Neware, 2020)**.**

Most customer segmentation approaches.

based on customer value fail to account for the factor of time and the trend of value changes

in their analysis. In this article, we classify customers based on their value using the RFM.

model and K-means clustering method. Then, an assessment of changes over several

periods is carried out. The originality of this research lies in its incorporation of time.

and trend of customer value changes in improving the accuracy of predictions based on the

past behavior of customers. For this purpose, we used the POS customer transactions.

Most customer segmentation approaches.

based on customer value fail to account for the factor of time and the trend of value changes

in their analysis. In this article, we classify customers based on their value using the RFM.

model and K-means clustering method. Then, an assessment of changes over several

periods is carried out. The originality of this research lies in its incorporation of time.

and trend of customer value changes in improving the accuracy of predictions based on the

past behavior of customers. For this purpose, we used the POS customer transactions.

Most client division approaches dependent on client esteem neglect to represent the factor of time and the pattern of significant worth changes in their investigation. Monireh Hosseini and Mostafa Shabani characterize clients dependent on their worth utilizing the RFMmodel and K-means clustering strategy. At that point, an appraisal of changes for more than a few periods is done. The inventiveness of this exploration lies in its fuse of time and pattern of client esteem changes in improving the exactness of forecasts dependent on the past conduct of clients. For this reason, they utilized the POS client exchanges (Hosseini & Shaban, 2015)

**Method**

**Data collection and Data processing**

We have been provided with the data from our sponsor Quantum Analytica. The data is distributed into four CSV files, each file denotes a table- orders, product, customer, inventory. Each of these tables has many features and information which will be used in our project to analyze the data and develop models. This will allow us to have a preliminary understanding of the models that have better accuracy and have a basic feel of what kind of data cleaning we would need. As our objective is customer segmentation, we will be focusing on customer and order data sets.

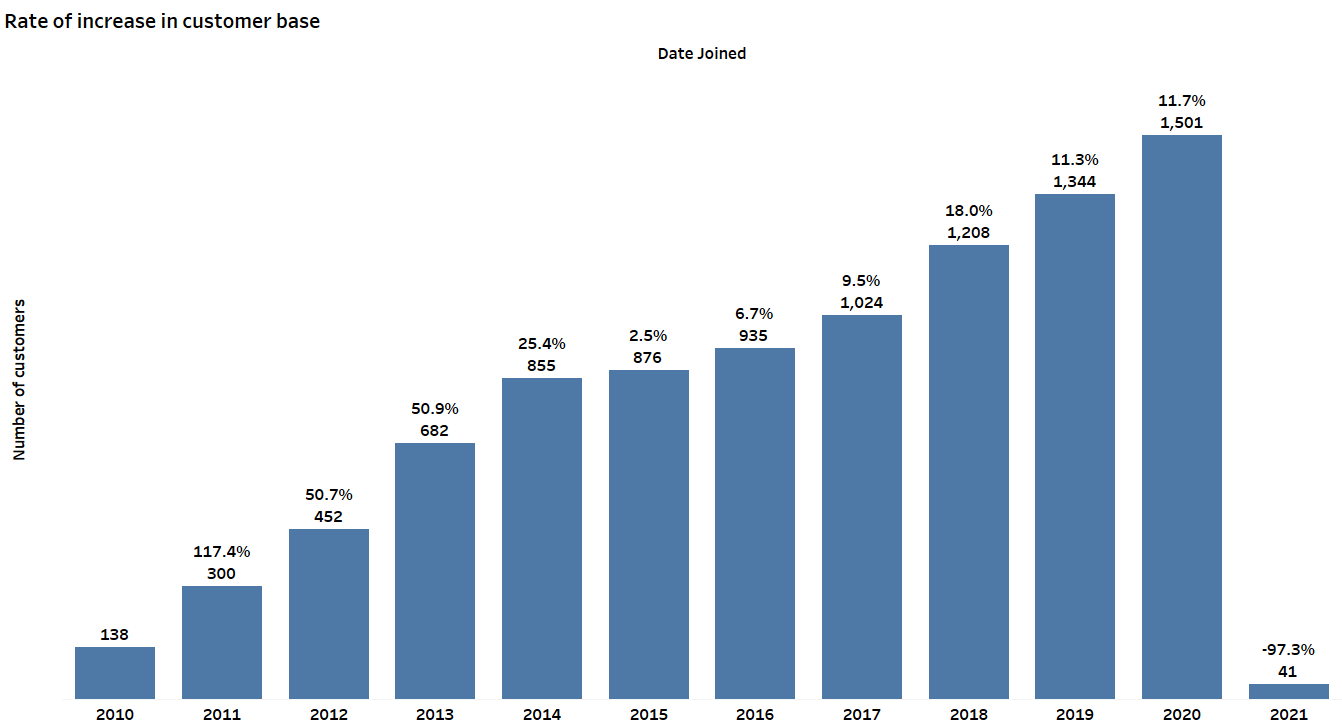
**Exploratory Data Analysis**

In this section, we will analyze the datasets and provide the required visualization which will help in further analysis and modeling. While analyzing the datasets we found that the customer data set contains 9356 observations and 9 features. The orders table contains 17936 observations and 55 features. The product table contains 1231observations and 35 features. The datasets can be combined by the common columns such as the orders dataset has customer id which can be used to join with the customer dataset and product id which can be used to join with products dataset.

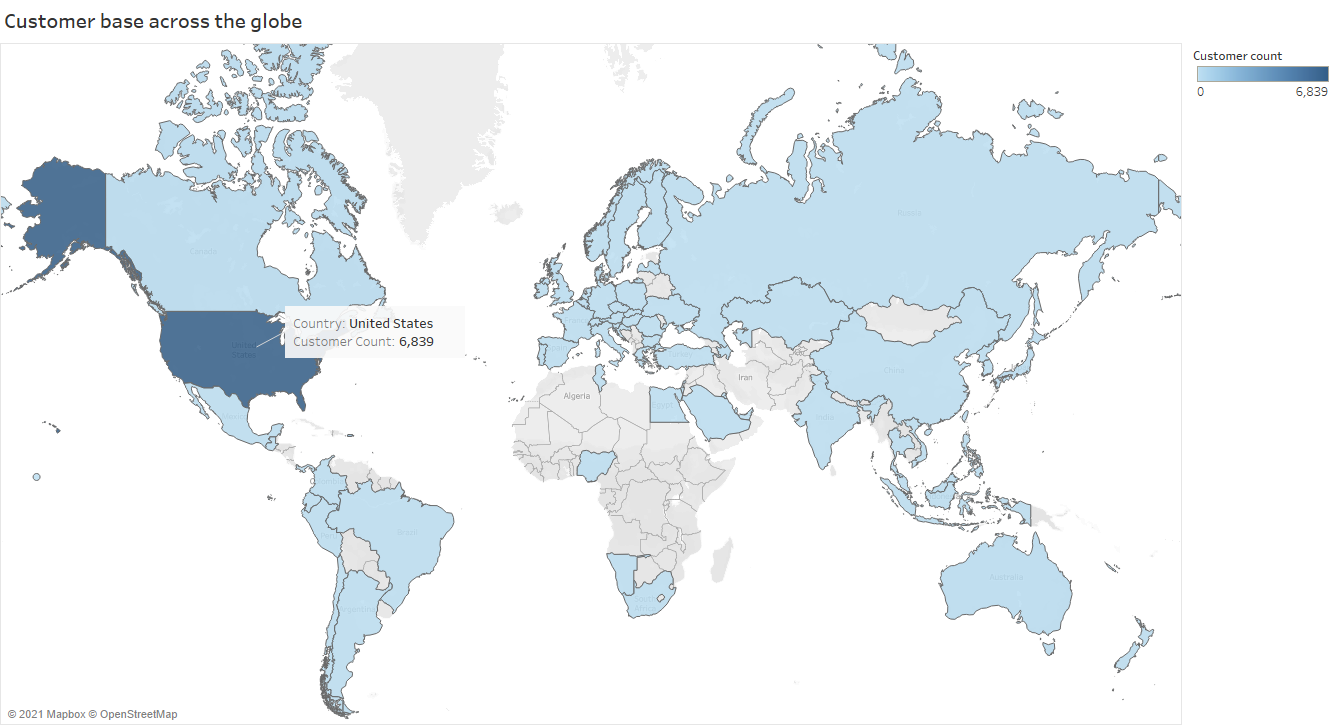
**Feature Engineering and Data Cleaning**

We will start by analyzing the feature by extracting the ones which will be needed for our analysis and model development which provide us with better results. In the customer dataset, we found many columns that were not applying to our objective such as Notes, customer groups, Tax-exempt category, further there are columns with very few values such as company, so we removed them. We also have a column called to address that contains details about the customer but in a comma-separated format such as the first name, the last name so on, we have separated these data into multiple columns each column describing the customer and have removed the columns that were not giving a value to our analysis or the columns that had very less information. In the orders and product dataset, we applied the same process, we have removed those columns that do not give any value to our objective. Then further we need to perform data cleaning and data processing to make the data clean, consistent, and usable. For numerical columns, we have imputed the missing values using the mean and for categorical columns, we have used the mode. For some observations, we had the value of zip code or city, but we did not have the value of the city, state, or country so using the lookup method we have filled the missing values. We have removed rows that had more than 60% missing values and for which values could not be filled. Further, we have corrected the data type for the columns which had the wrong data type. Since the tools we use are case sensitive we have converted some columns (such as payment methods) data to lowercase. Some columns such as payment method have multiple values some, we have taken only the first value and removed the others.

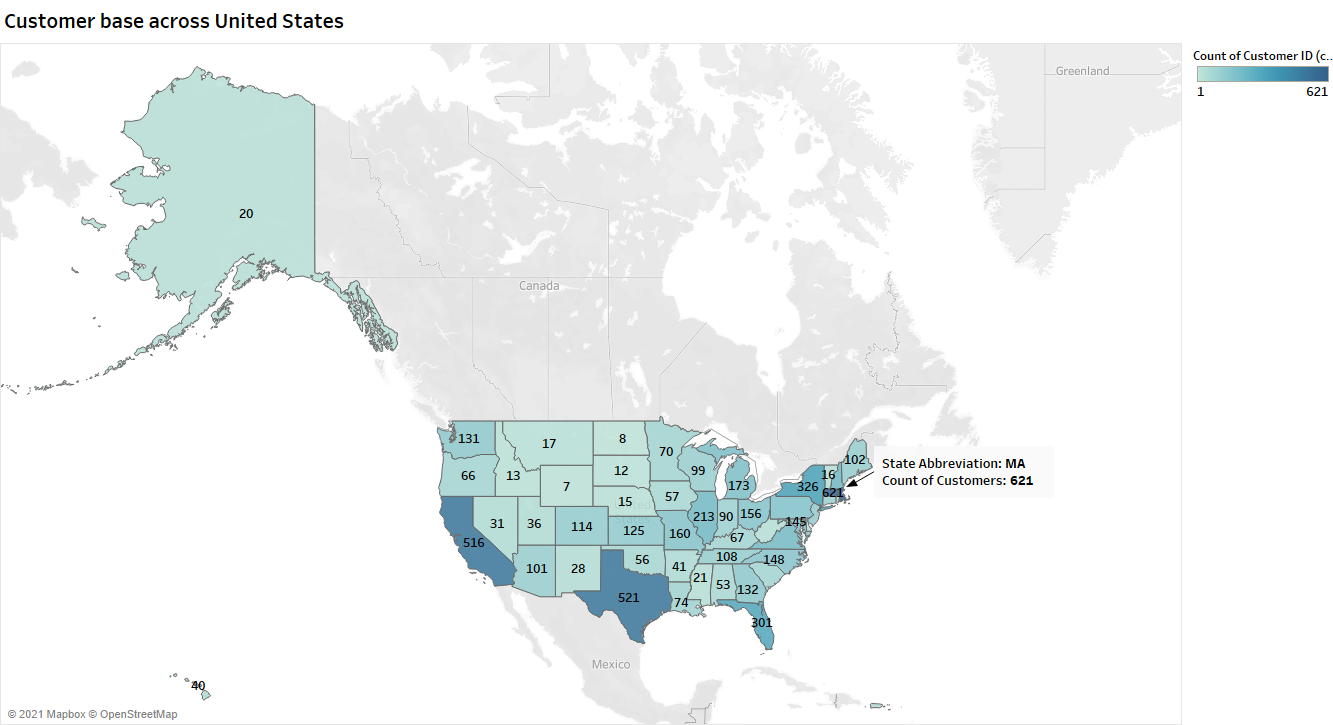
**Analysis**



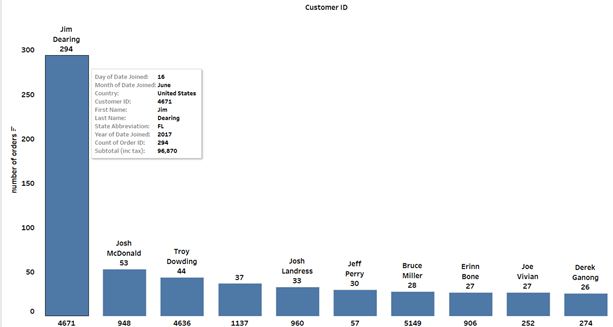
The above graph provides information about the increase in the customer base for the ACB company. There are two values for each bar in the graph one denoting the number of customers joined in that year and another is the percentage difference to the previous year denoting the rate of increase concerning the previous year. As the number of customers has increased every year, but the rate of increase is varying.



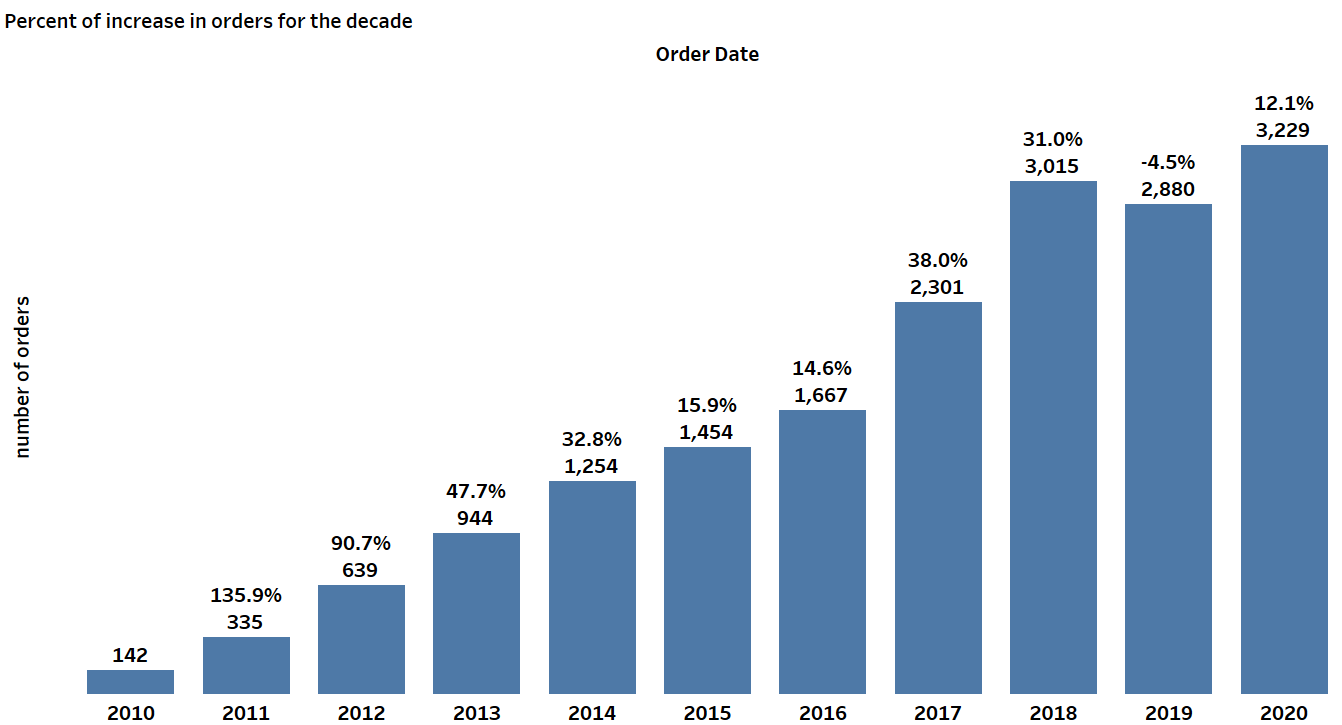
The above map shows the customer base across the globe, the count of the customers is represented by a blue color scale where the lighter part denotes lower customer count and going higher to darker one denoting a high customer base. From the map, we can observe that across the globe, the United States of America is ACB’s largest customer base.



The above map is a further analysis of the previous map. This map shows the customer base for the United States. In Massachusetts, ACB has the largest customer base of about 621 customers.

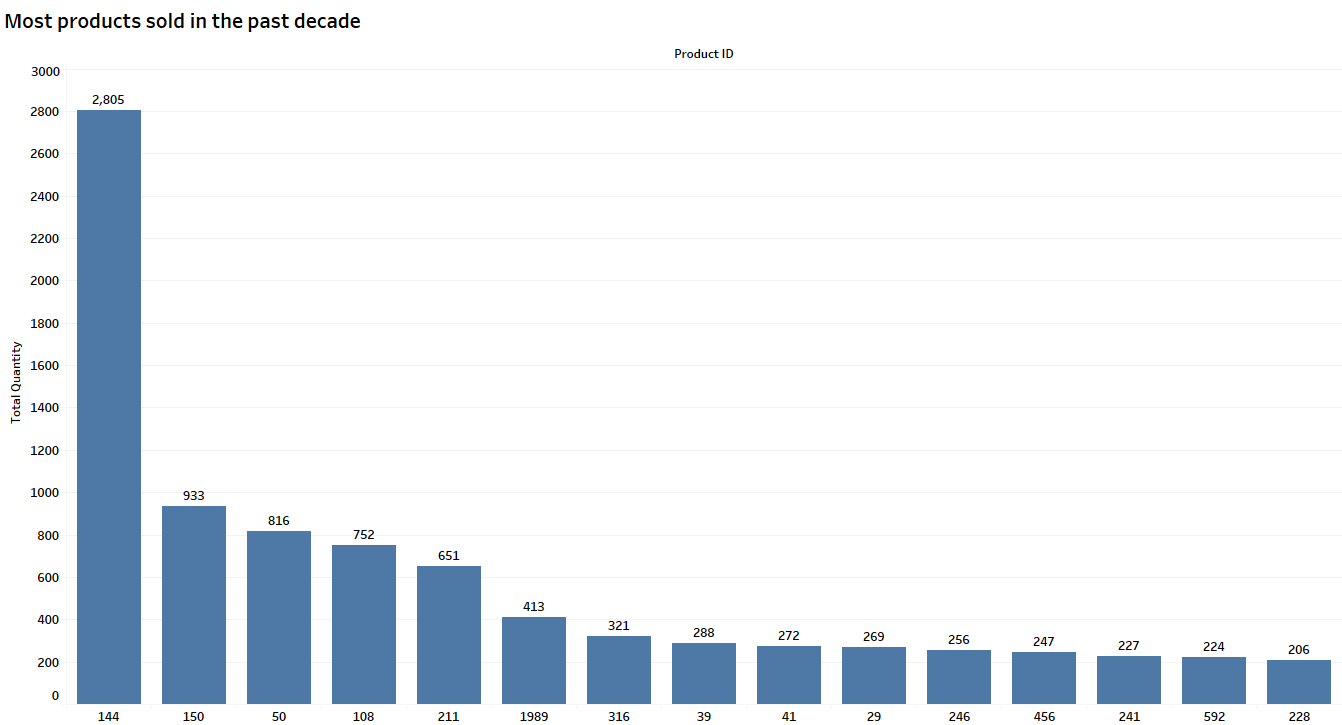


The above graph shows the top 10 customers who have ordered the most in the decade and it can be viewed that Jim Dearing from The United States of America, started ordering in 2016 but joined in 2017, has ordered about 294 times and has paid a subtotal of $96,870. From the further analysis, we found that most of his orders are either shipped or completed, few of his orders are canceled. He ordered the most in 2018 of about 132 orders.

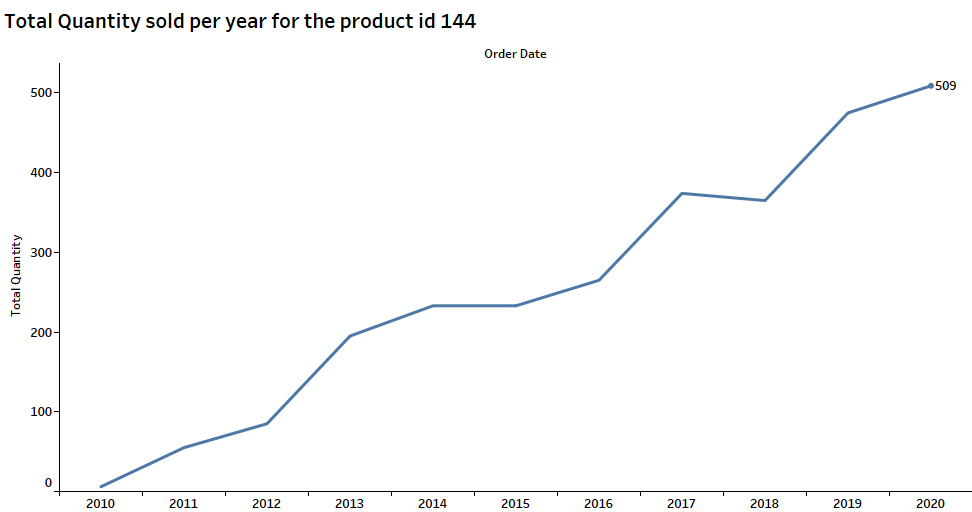


The above graph provides information about the increase in the order for the ACB company over the decade. There are two values for each bar in the graph one denoting the number of customers joined in that year and another is the percentage difference to the previous year denoting the rate of increase to the previous year. As the number of orders has been increasing every year except 2019 but the rate of increase is varying. Although in 2019 the orders dropped by 135 compared to 2018, they got back on track in 2020 with an increase of about 349 orders compared to 2019.

Next, we have created a chart to visualize the most sold products in the last decade and we have selected the top 15 products that were most sold. The product with id 144 (product name: Austin Custom Brass Standard Series Trumpet Mouthpieces) was the most sold.

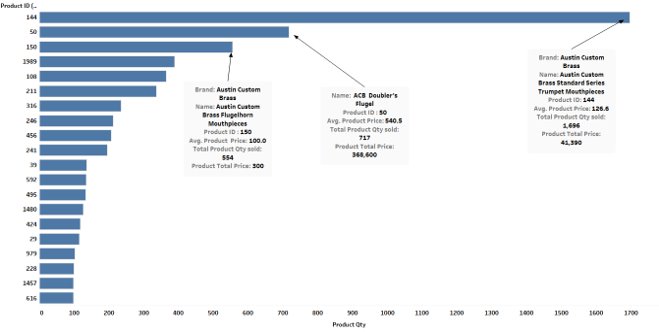


Further, we wanted to visualize the sale of productid 144 (product name: Austin Custom Brass Standard Series Trumpet Mouthpieces) each year, so we have created the following graph.

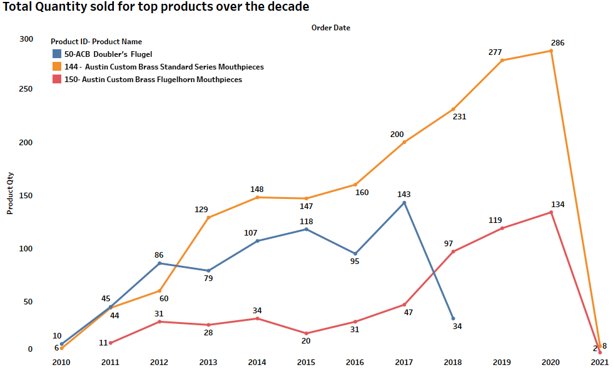


So, from the chart, it can be observed that the total quantity sold for the productid 144 was eventually increasing each year and it peaked in 2020 with a value of 509.

**Top Trending Products**

****

From the above chart, the Austin Custom Brass Standard Series Trumpet Mouthpieces (product id:144) is the top trending product with 1,696 pieces sold over the decade. The other two top products are also shown in the chart. Further, in the below chart we will observe these three products sold over the decade and which year had the highest sale for the product id 144.

****

* Products with id -144 and 150 were the highest sold in the year 2020.​
* As per the graph, there were no sales of product id-50 after 2018. ​
* Austin Custom Brass Standard Series Mouthpieces were highest sold (1418 pieces) in the United States over the decade with 152 pieces alone sold in New York.​

**Modeling**

One of the business methodologies in e-commerce is customer segmentation. Through the segmentation, the expectation of the client practices and examples can be improved, and hence, firms can make suitable promoting procedures as indicated by various customer groups. (Wang, 2020). We will be focusing on creating customer segmentation using the segmentation techniques and tools, metrics to evaluate the models and further improve the models.

**RFM Model**

RFM segmentation model permits advertisers to target explicit bunches of clients that are substantially more applicable for their specific behavior – and hence produce a lot higher paces of reaction, in addition to increased loyalty and client lifetime esteem. Like other grouping strategies, RFM segmentation is an amazing method to recognize segments of clients for unique treatment. RFM represents recency, frequency, and monetary. It segments the customers based on their transaction history:

* Recency(R): How recent was the last purchase?
* Frequency (F): How often do they purchase?
* Monetary (M): How much do they purchase?

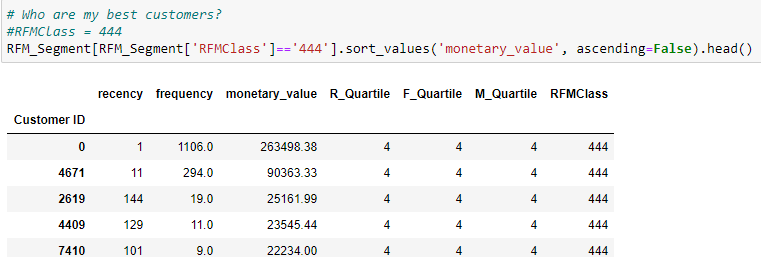
The resulting groups can be ordered from the most valuable (high R, F, M value) to the least valuable (least R, F, M value). For this analysis we used the following details of the customer:

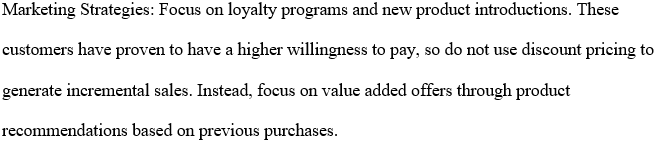
* Customer id- to identify the customer.
* Order\_Date: To find the recent order of the customer and calculate the number of days ago the purchase was made. This field is used for Recency.
* For Frequency, we will be counting the number of orders for each customer.
* For Monetary, we are finding the amount spent by this customer for that we will total up the amount spent for all the transactions for each customer.

Using the Model, we have answers for the following questions:

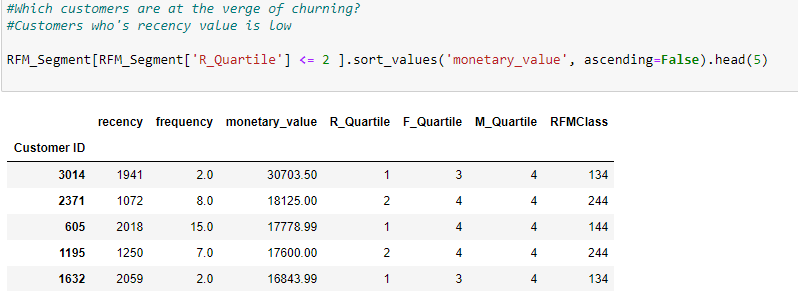
Who are the best customers?

The customers with RFM class as 444 means that their recent purchase was in the first quartile (recent purchase) and the frequency of purchase made is in the final quarter (number of purchases are high) and the total amount purchased for is also in the last quarter (large total amount). So below is the head of data for the best customers.





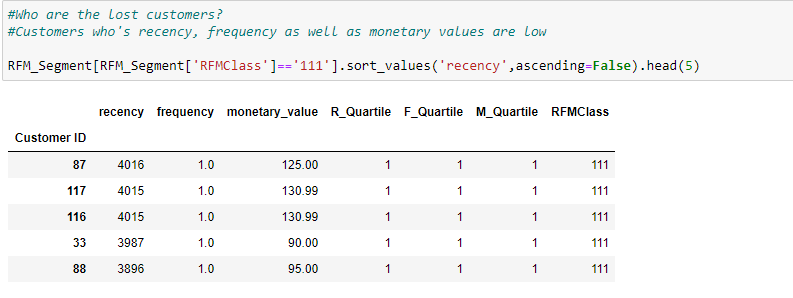
Which customers are on the verge of churning?



So, the customers whose recency is low, but their money is high and so is the frequency, then these customers are on the verge of churning. ACB can focus on these customers to reactivate the engagement and reduce the churn.

Who are the lost customers?

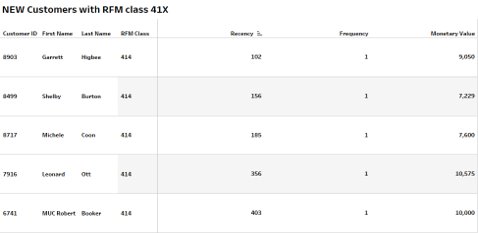
These are the customers whose recency, frequency, and monetary values are low. ACB can take measures to win back these lost customers.





Who are all the Rookies- new customers?

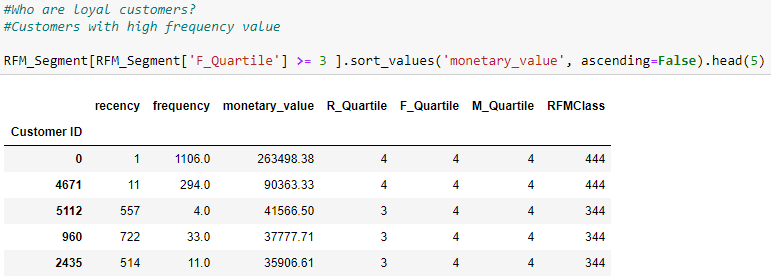
These are the customers who have high recency but low frequency as they are new customers and have not made many purchases.





Who are all the loyal customers?

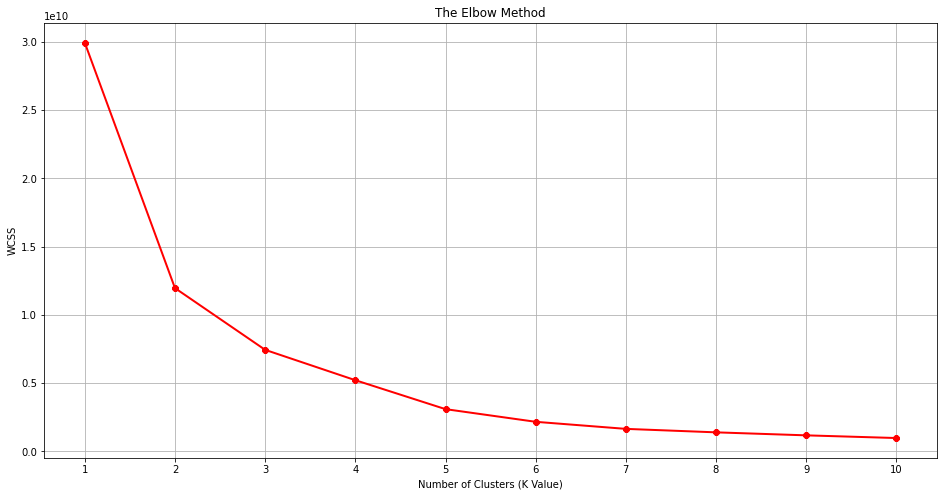
Finally, we have arrived at the most important part- loyal customers. These are customers who have a high frequency, that is, they keep buying from ACB. (Bajaj, 2019)



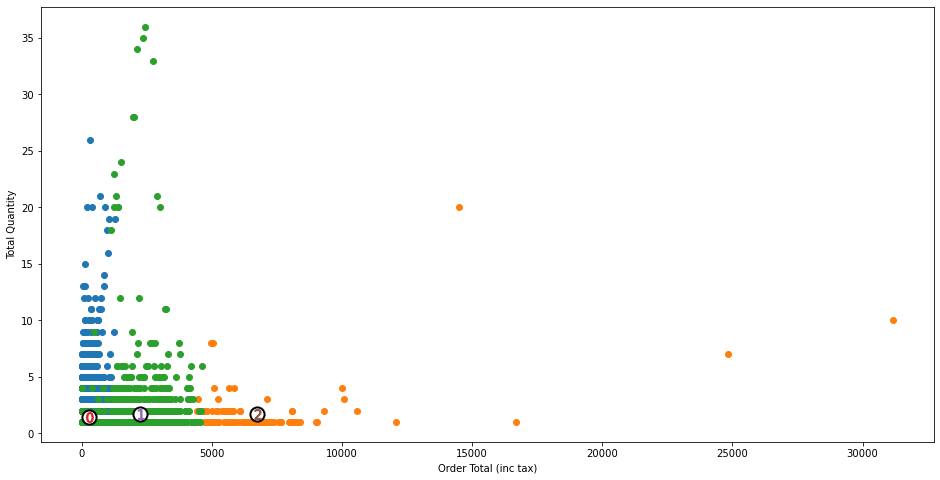
Normally, you need to speak with these clients routinely by phone, mail, email, online media, and that is just the beginning. These individuals are the ones who can and should impact your purchasing and promoting choices. Nothing will cause an unwavering client to feel better compared to requesting their info and showing them the amount, you esteem it. The truth is you can never do enough for them. Frequently, the more you do, the more they will prescribe you to other people. What is more, positive verbal exchange is gold for business. (HUNTER, 2020)

**K-MEANS Clustering**

Using the K-means algorithm we have segmented the customers based on the following features Order\_total, Total\_quantity, subtotal, shipping cost, handling cost, total shipped, country (encoded), and more. To find the optimal K value to be used for the clustering we have used the elbow and the graph is provided below.



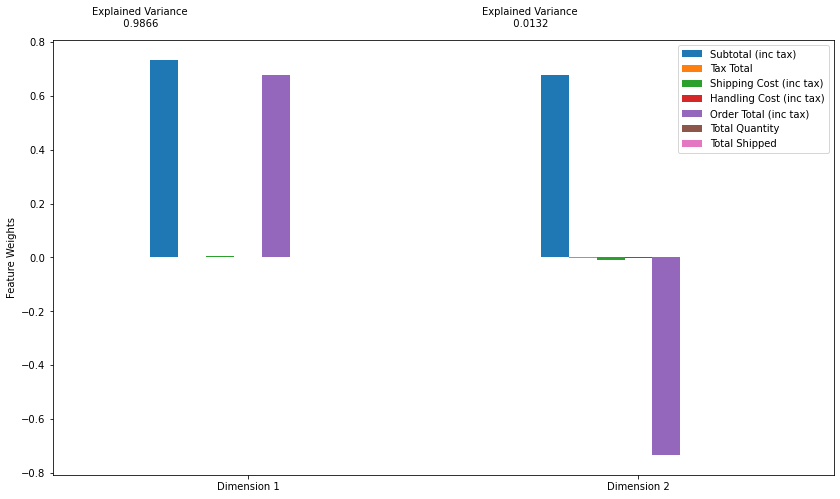
Using the elbow method, we found that the elbow is at 3 and hence that is our optimum number of clusters. So, we have finally created three clusters.



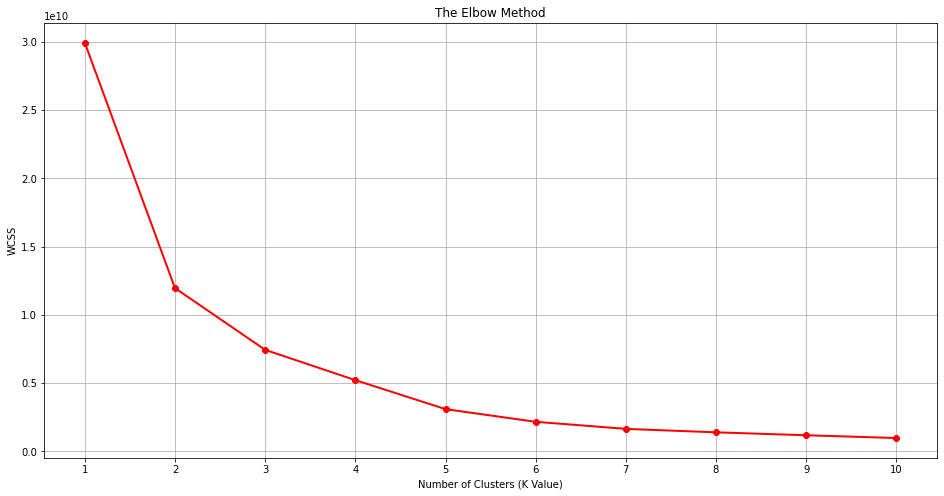
The above graph is created which shows the clusters formed with blue indicating cluster 0, orange – cluster1, and green – cluster 2. The centroid for each cluster is marked by the number.

**K-means using PCA.**

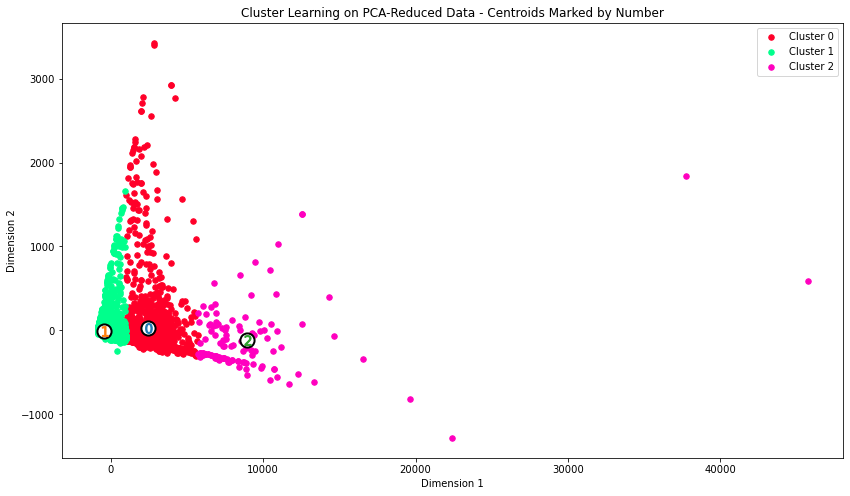


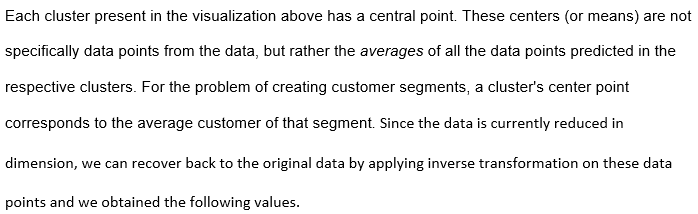
****

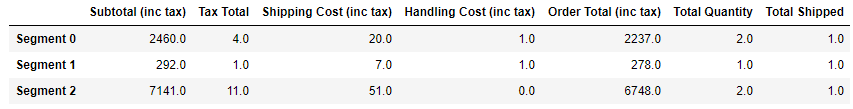
Then applied the K-mean on the PCA applied data and have got the optimal K value as 3 based on the elbow method.

****

So we have selected 3 as the optimal number of clusters and have obtained the following clusters

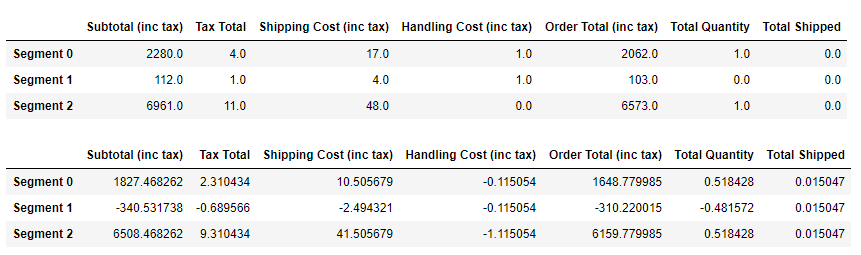






These are the average values for each segment and can be used to understand how the data is segmented. Most of the data points are segmented into segment 0.

Next, we will see how the cluster centers deviating from the median of the data set.



**DashBoard**

We have used the results found and have visualized them in a dashboard. In this dashboard we have four sections the first section represents the national dashboard, the second section represents the worldwide dashboard these two dashboards can be used to observe the performance, sales, and we have filters that can be used to filter the data for specific visualizations. The third dashboard is customer segmentation and is the visual representation of the RFM model, in this dashboard we have two graphs, in the first graph we can select a particular segment to observe the top products bought by the customers of that segment, the second graph can be used to find the segment of customers who have purchased the most for a specific product or product category. The fourth dashboard is product segmentation.

**Requirements and Tools**

The main requirement is the data and resources needed for the project. The data is provided by our sponsor Quantum Analytica. The tools we will be using are.

* Dataset files in .csv format.
* Jupyter is a free, open-source, intuitive web apparatus known as a computational notebook, which will be used to consolidate programming code, computational yield, informative content, and sight and sound assets in a solitary report.
* Tableau - a data visualization and data analytics tool
* Python – a coding language
* Libraries – pandas, NumPy, scipy, scikit-learn, matplotlib, seaborn.
* Presentation tool

**Conclusion**

The general population and customer population will be compared and segmented using an Unsupervised learning algorithm. We have created two models one RFM model and another K-means. We have also tried the PCA method to reduce the dimensions and improve the K-means model. We have found the best, loyal, highest paying, lost customers from the RFM model and have provided few marketing strategies that can be applied. We have also created an interactive dashboard that summarizes and visualizes our findings and results. Future work can be to obtain more information such as demographic and behavioral information which can lead to better segmentation and targeted marketing. Further, we can also focus on personalized recommendations which could improve customer retention, decrease cart abandonment rate, increase average order value, increase session time, sales, profit, and revenue.

# References

Bajaj, A. (2019, MAY 27). *RFM Analysis For Successful Customer Segmentation using Python*. Retrieved from Medium: https://aainabajaj39.medium.com/rfm-analysis-for-successful-customer-segmentation-using-python-6291decceb4b

Camilleri, M. (2017). Market Segmentation, Targeting and Positioning. In M. Camilleri, *Travel Marketing, Tourism Economics and the Airline Product: An Introduction to Theory and Practice.* Switzerland: Springer, Cham, Switzerland. DOI: 10.1007/978-3-319-49849-2\_4

Faraone, M. (2012). Using context to improve the effectiveness of segmentation and targeting in e-commerce. *Expert Systems with Applications, 39*(9), 8439-8451. Retrieved from https://doi.org/10.1016/j.eswa.2012.01.174.

Hosseini, M., & Shaban, M. (2015). New approach to customersegmentation based on changes in customer value. *Journal of Marketing Analytics, 3*. doi:10.1057/jma.2015.10

HUNTER, M. (2020, January 05). *The 5 Types of Customers (And How to Get Them to Buy More)*. Retrieved from the balance small business: https://www.thebalancesmb.com/the-5-types-of-customers-2948073

Singh, H., & Neware, D. (2020). Improving Customer Segmentation in E-Commerce using. *International Journal of Advanced Trends in Computer Science and Engineering, 9*(2), 2326-2331. doi:https://doi.org/10.30534/ijatcse/2020/215922020

Wang, S.-C. (2020). A hybrid big data analytical approach for analyzing customer patterns through an integrated supply chain network. *Journal of Industrial Information Integration, 20*. Retrieved from https://doi.org/10.1016/j.jii.2020.100177

Zhang, J. (2018). Product features characterization and customers’ preferences prediction based on purchasing data. *CIRP Annals, 67*(1), 149-152. doi:https://doi.org/10.1016/j.cirp.2018.04.020