Online Retailer: Customer Segmentation

Introduction

Buying a car, a house, groceries or even a toy for a year old has changed from what it was 30 years back. From the electronic system called the Electronic Data Interchange that companies used back in 1960 to facilitate the transfer of documents, to first sale of CD on website in 1994, to 2020 surge in disinfectant wipes and hand sanitizer sales, Ecommerce has changed our lives.

Ecommerce is a business model that allows companies and individuals to buy and sell goods and services over the Internet. This has led to rise of businesses who use Internet as their market place to sell goods and services like Amazon, Netflix, e-bay, Etsy, Alibaba, and many more. These big e-tailers (or electronic retailers) brought convenience, expanded selection, and affordability to Consumers. The huge success of these large e-tailers have not only disrupted many brick and mortar retailers but also smaller e-tailers, forcing them to drive efficiency in Online retailing path.

Objective

A UK-based and registered, non-store online retail company is looking to increase efficiencies and profits. It mainly sells unique all-occasion gift-ware. Many customers of the company are wholesalers.

The business objective is to understand customer to enhance Product offering and maximize profit. Understanding their customers will assist Online Retailer to design specific Marketing plan to reach to individual Customer segment and improve profits.

Customer segmentation is a method of dividing customers into groups or clusters on the basis of common characteristics. Segments can be based on

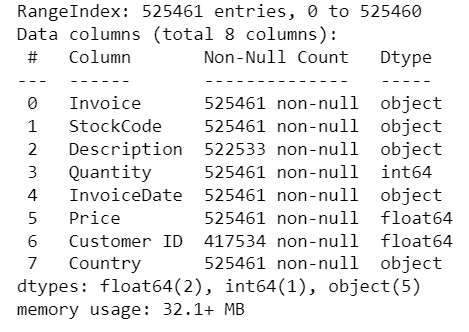
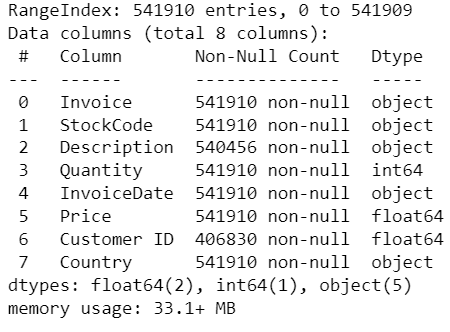
1. Demographic characteristics such as occupation, gender, age, location, and marital status.
2. Geographic characteristics such as location, region, urban/rural
3. Psychographic characteristics such as social class, lifestyle and personality characteristics
4. Behavioral characteristics such as spending, consumption habits, product/service usage, and previously purchased products

The available historical dataset for Online Retailer can be used to understand the behavioral characteristics of customers and efficiently build targeted group of the audience for tailored communication.

Data Analysis

Data Collection: Our Online Retail data set contains all the transactions occurring for company between 01/12/2009 and 09/12/2011. The attributes are

* InvoiceNo: Invoice number. A 6- digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation. (Categorical)
* StockCode: Product (item) code. A 5-digit integral number uniquely assigned to each distinct product. (Categorical)
* Description: Product (item) name. (Categorical)
* Quantity: The quantities of each product (item) per transaction. (Numeric)
* InvoiceDate: Invoice date and time. The day and time when a transaction was generated (Numeric)
* UnitPrice: Unit price. Product price per unit in sterling (Â£). (Numeric)
* Customer ID: Customer number. A 5-digit integral number uniquely assigned to each customer. (Categorical)
* Country: Country name. The name of the country where a customer resides. (Categorical)

1. b. Data - Worksheet 2

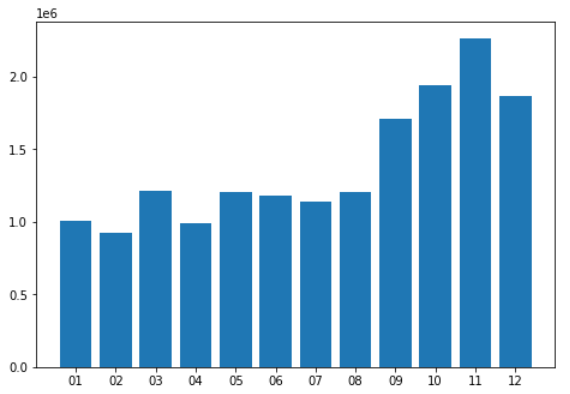
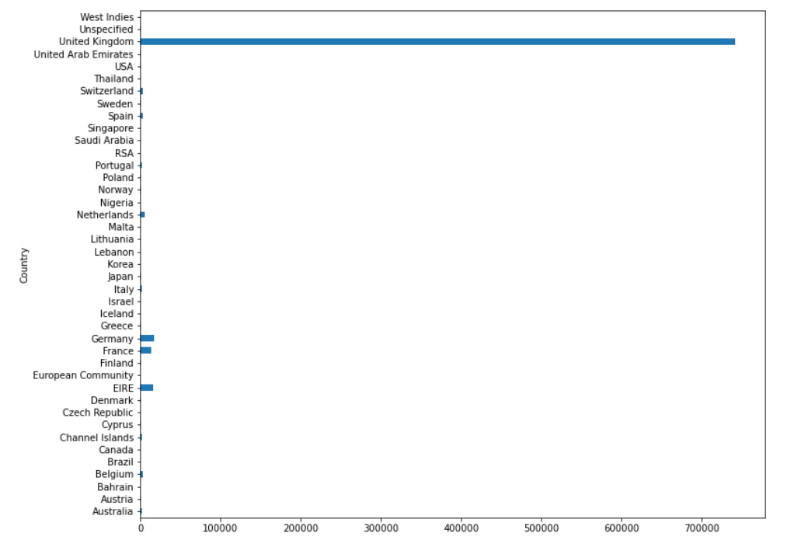
1. a. Data - Worksheet 1

Data Cleaning: To prepare data for cluster analysis,

* Removed transactions with missing Customer ID, or null data
* Removed Outliers
  + Negative Quantity and Price, Cancelled orders
    - Order Cancels have Negative Quantity. Also ‘C’ on Invoice number
    - Negative Price marked for bad debt adjustments (5 entries)
    - 2 Bad debt adjustments (=22k amount) same date 12/8/11
  + Removed Non UK
* Normalized and scaled RFM using Standard Scaler library

Exploratory Data Analysis: On exploring dataset with few time series and distribution graphs, noticed

* 5942 customers spent on average 2801.80 sterling
* Each customers had purchased on average approx. 7.5 times in 2 years, at least once and at most 508 times
* On average, 293.97 sterling spent by a customer on a purchases
* 75% of customer purchased for under 357.41 sterling



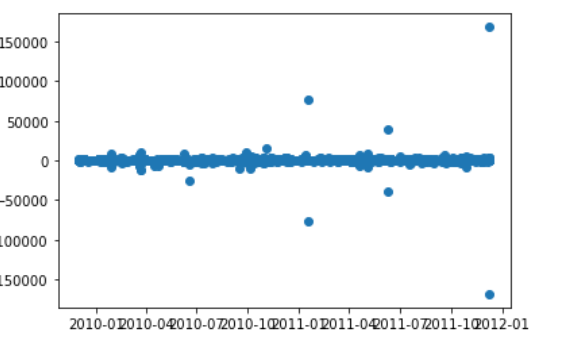
3. Transactions by month

2. Transactions by Country

* Most Customers and most revenue generated from United Kingdom
  + Ireland, Netherlands, Germany, France, and Australia - top Revenue generating countries/ markets
* Seasonality in revenue generated
  + Peak months: Q4 – Oct to Dec

Baseline: Customer Segments can be identified using Clustering analysis. Clustering is an unsupervised machine learning technique that devices the population into several clusters or groups in such a way that data points in a cluster are similar to each other, and data points in different clusters are dissimilar.

On reviewing only customer spent amount, 3 clusters can be identified as on scatter plot below.

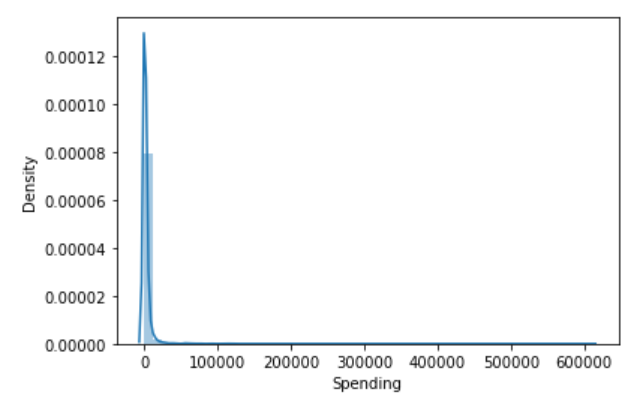
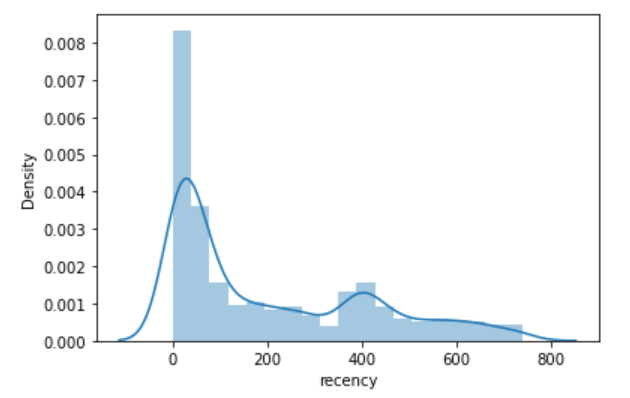


4. Purchase by Invoice date

* Cluster 0: Orders with moderate revenue generation
* Cluster 1: Orders that were cancelled and resulted in loss of revenue
* Cluster 2: Orders with high revenue generation

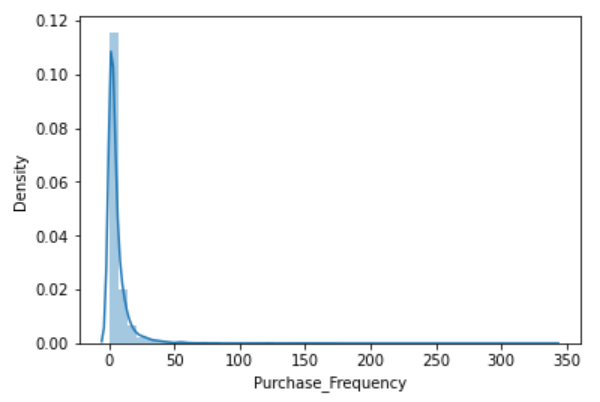
However, there are segment of customer who is the big spender but what if they purchased only once or have not purchased in last 10 years? In order to segment more logically, I have used RFM analysis.

RFM Model: RFM (Recency, Frequency, and Monetary) analysis is a behavior-based approach grouping customers into segments. It groups the customers on the basis of their previous purchase transactions. How recently, how often, and how much did a customer buy.



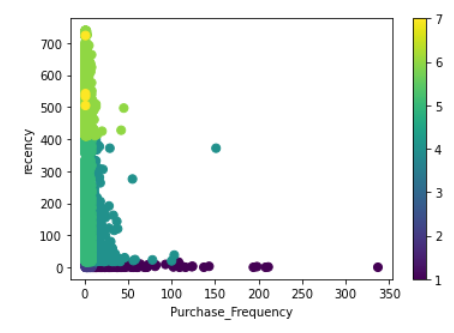
6. Spending

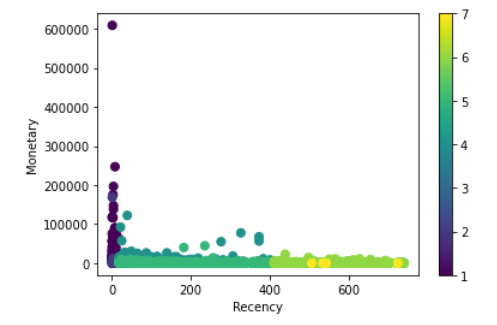
5. Last purchase or Recency

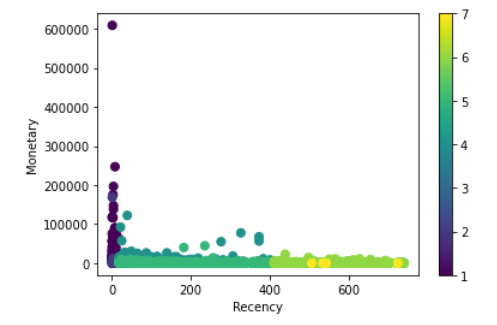


7. Frequency of Purchase

* Step 1: Calculated
  + The last purchase date or how recently did customer purchase
  + Amount spent by the customer
  + how frequently or the number of times a customer purchased
* Step 2: Assigned scores 3, 2, 1 based on quartiles for RFM
* Step 3: Combine the quartile scores and assign cluster segments based on scores, like 333 score for RFM is segment- Champion.

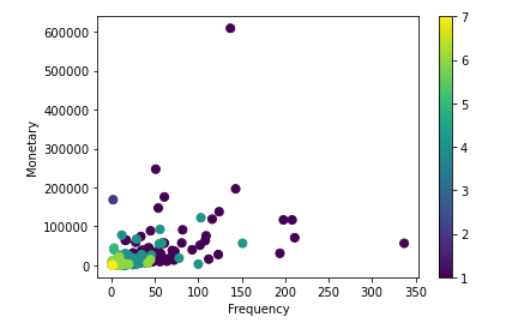




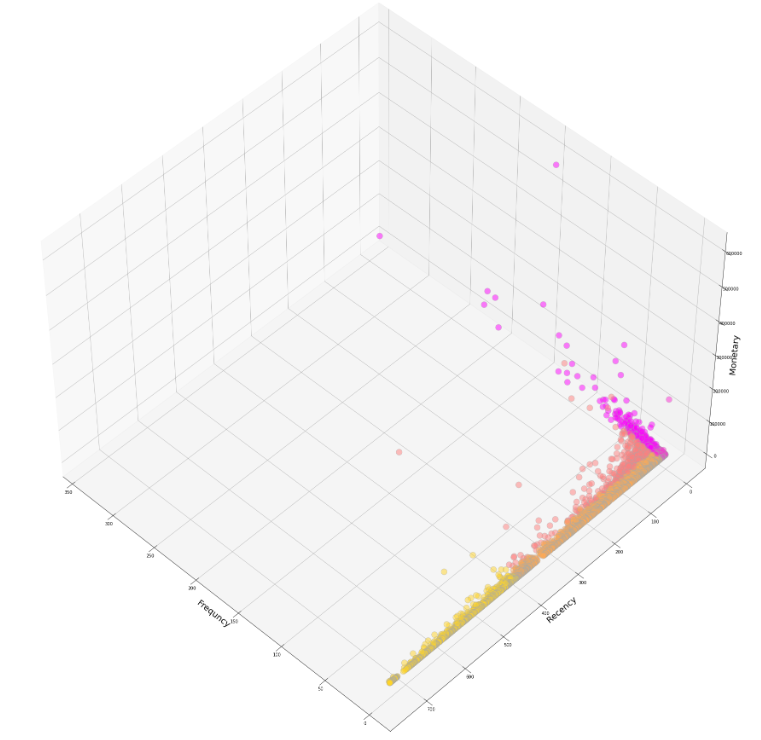


9. Scatter plot: Frequency Vs. Recency

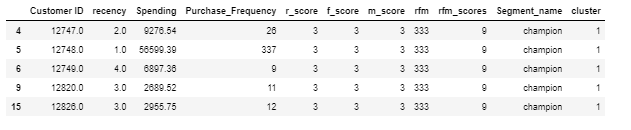
8. Scatter plot: Recency Vs. Monetary



10. Scatter plot: Frequency Vs. Monetary



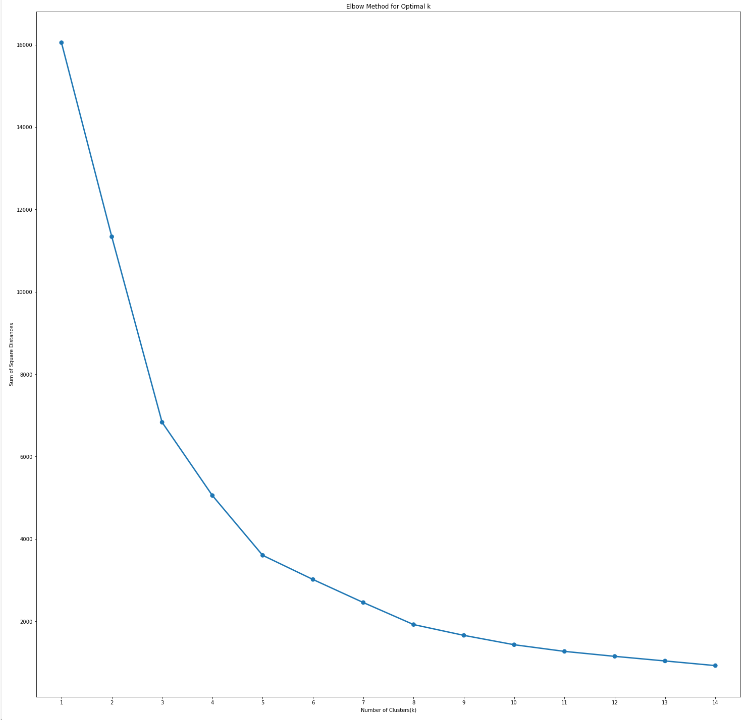
11. 3D plot: Frequency Vs. Recency Vs. Monetary



12. RFM score and Segments

Elbow Method: It is a heuristic used in determining the optimal number of clusters in a data set. It consists of plotting the explained variation for a range of number of clusters, and picking the elbow of the curve as the number of clusters to use. The perception is that increasing the number of clusters will naturally improve the fit, since there are more parameters until some point that results in over-fitting. The elbow reflects this optimal point.

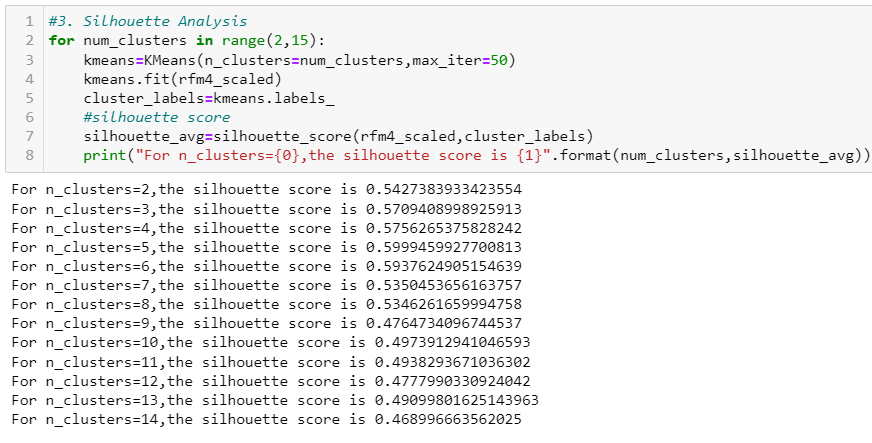
For Online retailer data, elbow of curve indicated 5 clusters.



13. Elbow method for Optimal K Clusters

Silhouette Coefficient or silhouette score: It is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

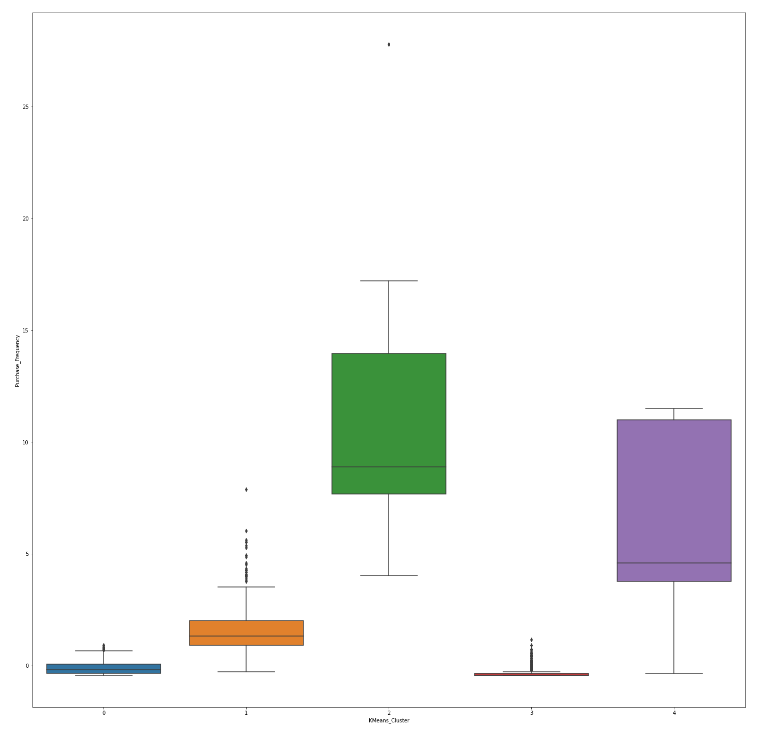
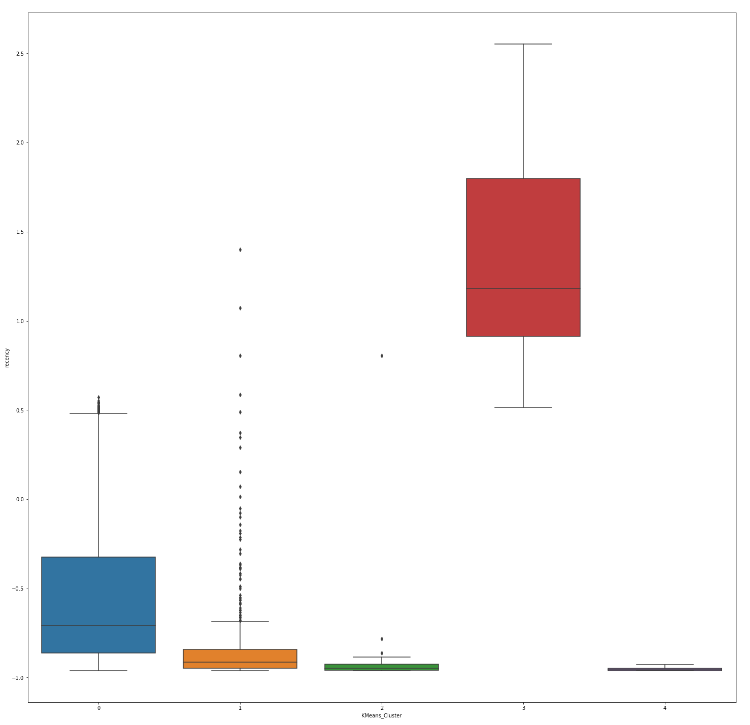
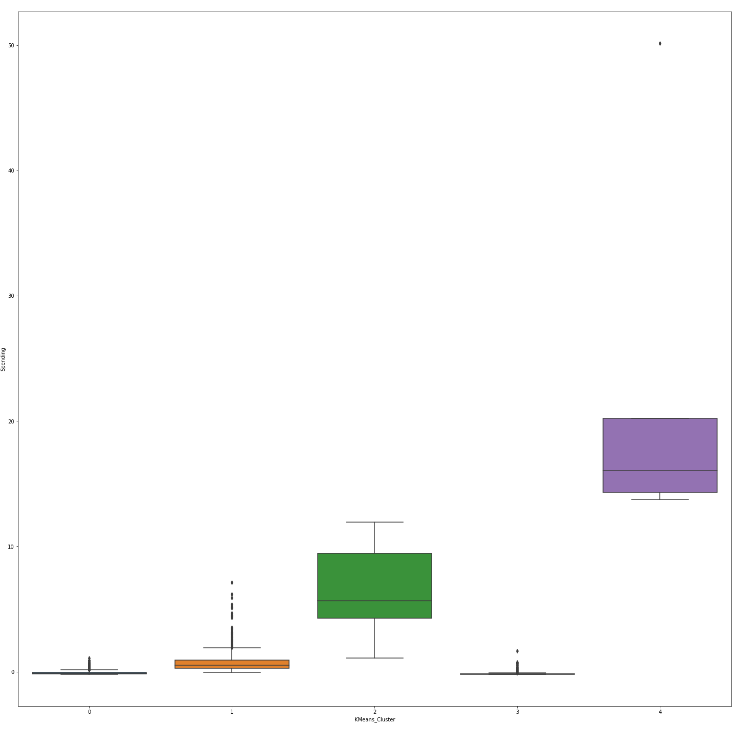
* 1: Means clusters are well apart from each other and clearly distinguished.
* 0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.
* -1: Means clusters are assigned in the wrong way



14. Silhouette Coefficient Optimal K Clusters

5 clusters has the highest silhouette score

K Means Clustering:

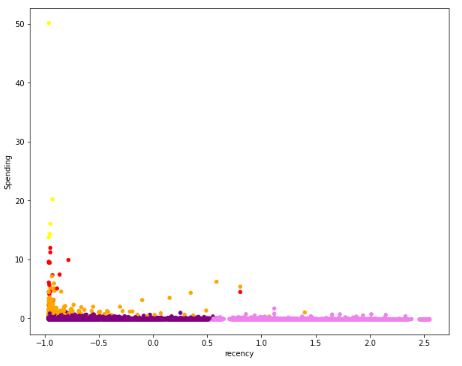
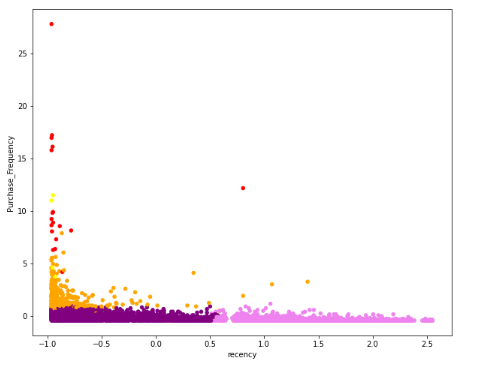


16. K Means Clusters by Recency

17. K Means Clusters by Purchase Frequency

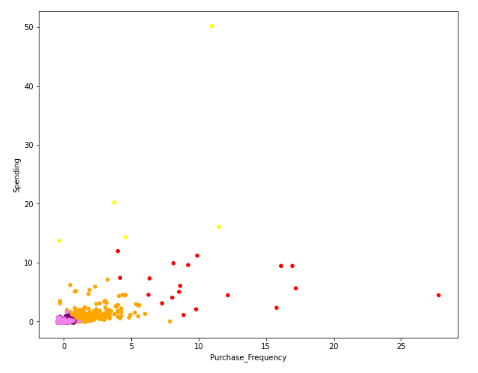
15. K Means Clusters by Spending

* 5 clusters identified from the data are
  + Cluster 0: Customer group who bought fairly recently, but fewer number of times and spending lower compared to others
  + Cluster 1: Customers who have recently purchased fairly high value products and did not repeat purchasing frequently
  + Cluster 2: Customers who have very high repeat purchases and higher spending, and recently purchased
  + Cluster 3: These Customers have not purchased in a very long time and have high churn while spending very little
  + Cluster 4: This is the best group of customers who buy very frequently and spent the most. Also they bought very recently



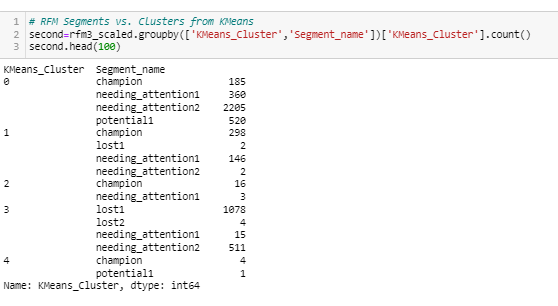
19. Scatter plot: Recency Vs. Spending

18. Scatter plot: Recency Vs. Purchase Frequency



20. Scatter plot: Purchase Frequency Vs. Spending

RFM Segment vs. K Means Clusters:



21. RFM Segment vs. K Means Clusters

* Optimal Clusters identified with Elbow or Silhouette scores different compared to RFM segments identified with quartiles scores
  + Also overlap of segments and K-Means clusters
* 61% of customers are in K-Means Cluster 0
  + Champions: high Recency, high Frequency, high Spending spread across 4 K-Means clusters
    - 59% in K Means Cluster 1
    - 37% in Cluster 0
    - 3% in Cluster 2
    - 1% in Cluster 4

Conclusion:

* Online Retailer has over 61% of Customers in Cluster 0. A Marketing plan to drive this customer base to buy more frequently like utilizing recommenders, coupon offerings for related recent purchases, will increase assist in growing Online retailer
* 30% of customers have not purchased in recently. Online retailer is likely to lose this segment. Nurture campaigns may assist in bringing back this group
* Cluster 1 makes up 8% of customers who did not purchase repeatedly. But they spent fairly recent and bought high value products. This group may be the dissatisfied customers. A campaign for feedback and acting upon feedback, will help to keep this customer base with Online retailer

Next Steps:

The customer segmentation model can be implemented by gradually adding more countries after initially deploying with UK dataset.

For the clusters like Customer 0 and Cluster 2, recommender based on previous items purchased, can help Online retailer to increase product sales and retain customers.

References:

Clustering validation:

https://www.eecis.udel.edu/~compbio/thesis/JessShenThesis.pdf

General model deployment:

https://towardsdatascience.com/the-machine-learning-lifecycle-in-2021-473717c633bc

https://neptune.ai/blog/model-deployment-strategies

Silhouette:

https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c

Customer Segmentation:

https://medium.com/web-mining-is688-spring-2021/exploring-customers-segmentation-with-rfm-analysis-and-k-means-clustering-118f9ffcd9f0

https://towardsdatascience.com/simple-customer-segmentation-using-rfm-analysis-1ccee2b6d8b9

https://www.youtube.com/watch?v=fdUofaT8gUw

https://www.youtube.com/watch?v=VE0m2FLoJCk

https://www.youtube.com/watch?v=JCV7pPxJDdk