Analytics Specialization and Application

Student No: 20325050

Module Code: BUSI 4370

Year: 2021-2022

Coursework I

1. An executive summary

Even though market segmentation has a lot of applications, it is vital for target marketing nowadays. Keeping this perspective in mind, the transactional data of the convenience store is used for segmenting customers based on their purchase behavior. Among four data sets, the report is based on the file with the name basket_sample. The file contains a record of 3000 customers who have purchased different goods over a period of six months. The dataset contains almost 195547 transactions. Each transaction is recorded in detail with purchase date and time, it's quantity, value, and the number of categories in the basket. RFM model based on the frequency, recency and monetary features seems appropriate to distinguish customers for target marketing. Since this is an unsupervised machine learning task, a whole process of data processing – where we take care of correlation and outliers by standardizing/scaling the data- and dimensionality reduction is applied before doing the segmentation through K-means algorithms. Silhouette score revealed that the value of k – optimal clusters- should be three. However, due to the data manager's demand and micro analysis of all customers, six segments are created with not very low silhouette score. Thus, all customers are divided into six segments: champions, loyalists, potential loyalists, promising, those need special attention, and those at the risk of churning.

2. Feature Description

I. Deciding Important features

To proceed with the goal, it is important to look at the available features. First and foremost, the identity of customer in the form of customer id is indispensable from the data set. Once we know the customer, we can see what he has purchased over the period of six months.

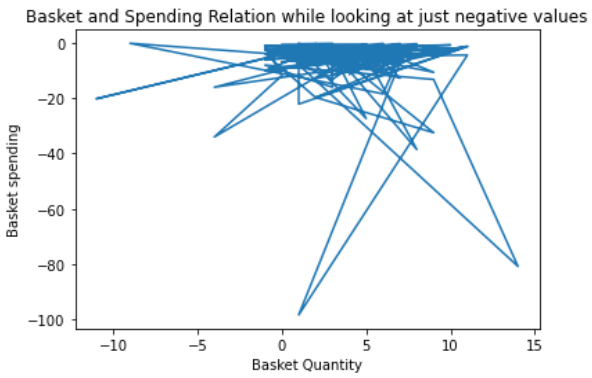
```
In [16]: basket_data.head()
```

Out[16]:

	customer_number	purchase_time	basket_quantity	basket_spend	basket_categories
0	11911	2007-03-01 07:06:00	7	£3.09	3
1	4047	2007-03-01 07:13:00	9	£7.99	5
2	3571	2007-03-01 07:27:00	9	£37.06	6
3	4079	2007-03-01 07:34:00	11	£11.91	5
4	6063	2007-03-01 07:36:00	3	£1.45	1

There are 195,547 transactions conducted over six months by 3000 customers in the data set. Purchase time is important because it helps us in analyzing the customer purchase pattern in each time period – it tells when and how frequent a customer came for shopping. Basket quantity and spend inform about a customer's purchase preferences. Customers with less items and high value are categorized as high-value buyers. Moreover, the businesses can ascertain who purchases products with high and low marginsData Preparation and Cleaning

Even though the data provided has no missing values, the presence of negative basket quantity and spend values raise some important questions. First and foremost, it was important to check correlation of negative basket to negative spending. Basket spending was negative 160 times whereas basket quantity was negative for just 19 times. Plotting negative values of spending against basket showed irregularity. Thus, the transactions with negative value of quantity and spending are dropped form the data set for better analysis.



II. Building RFM Model: Feature Engineering

One of the most widely used models for finding out customer behavior is the RFM model. It is based on three important parameters. Recency shows the time when the customer purchased last time form the time of the analy sis. F stands for frequency, which means the number of times a customer has purchased over a particular period. L astly, M stand for monetary value which is the accumulated sum of the customer's shopping over the same period time. The primary purpose for segmenting customers is to find out customers with their specific needs to advertise for.

The data shows that the collection of records started at 2007-03-01 07:06:00 and ended at 2007-08-31 21:55:00. We set the next day to be the time of analysis - 2007-09-01 - for our analysis. After applying group by function on customer number, we take the difference between analysis date and last purchase data of the customer to find recency of the customer. Similarly, frequency is calculated by counting the occurrences of the basket for all custom ers through their customer ids. Lastly, we add the cumulative sum of spending of each customer over the six mont hs period. Thus, the new RFM model for our original features look like this table:

The table on right hand side shows RFM data frame for our data set. Average recency of the customers is 8.12 days. Each customer has visited 65 tiems on average. Average spedning of each custoemr is 769 pounds. Therefore, RFM uncovers insightful information about custoemrs' shopping behavior.

The distribution plot of the three features is shown below. It can be seen that all the parameters are rightly skewed, i.e., their median is greater than their mean. Monetary value seems more normal as mean and median has a narrow difference.

```
In [67]: # Here reset index is applied to make all the columns look even in the data frame
         RFM_data.reset_index().head()
```

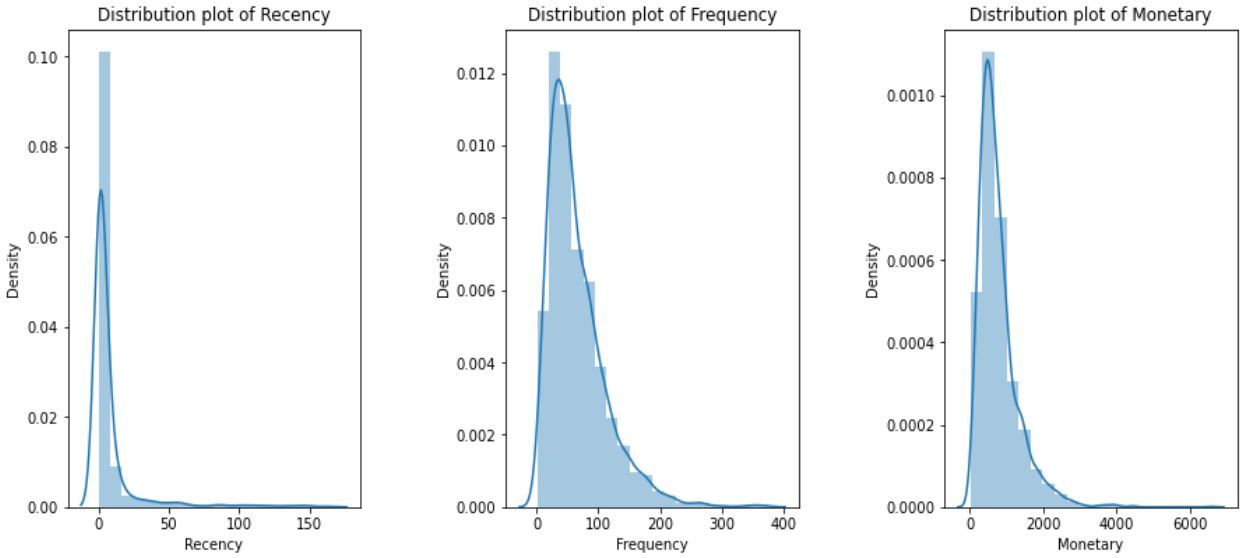
Out[67]:

	customer_number	Recency	Frequency	Monetary
0	14	1	56	675.72
1	45	1	33	585.73
2	52	2	59	222.18
3	61	3	37	547.87
4	63	7	48	293.34

```
In [68]: RFM_data.describe()
```

Out[68]:

	Recency	Frequency	Monetary
count	3000.000000	3000.000000	3000.000000
mean	8.121667	65.123667	769.758087
std	20.938489	47.392134	552.972322
min	0.000000	1.000000	7.280000
25%	0.000000	32.000000	406.707500
50%	2.000000	53.000000	627.170000
75%	6.000000	86.000000	958.660000
max	164.000000	374.000000	6588.650000



3. Segmentation Methodology: Appropriteness of feature selection/engineering

I. Applying standardization to make data more Normal

This preprocessing step is important because of the clustering algorithm. Since we will use K-means algorithm, the technique seeks globular clusters. Standardization will put all these variables on the same scale, i.e., from -1 to 1, and it will be easy for model to accurately separate scaled data. Normalization helps us in comparing different variables using the same scale.

II. Why do we not need factorization, i.e., dimensionality reduction, on our RFM Model?

Principal Component Analysis (PCA) is a technique to reduce the dimensionality of data. However, RFM model is derived from those features which cannot be eliminated, as they are indispensable for analysis. In fact, how can we ignore either recency, frequency, or monetary feature while categorizing customers based on their purchasing behavior. Therefore, PCA is unnecessary for RFM model.

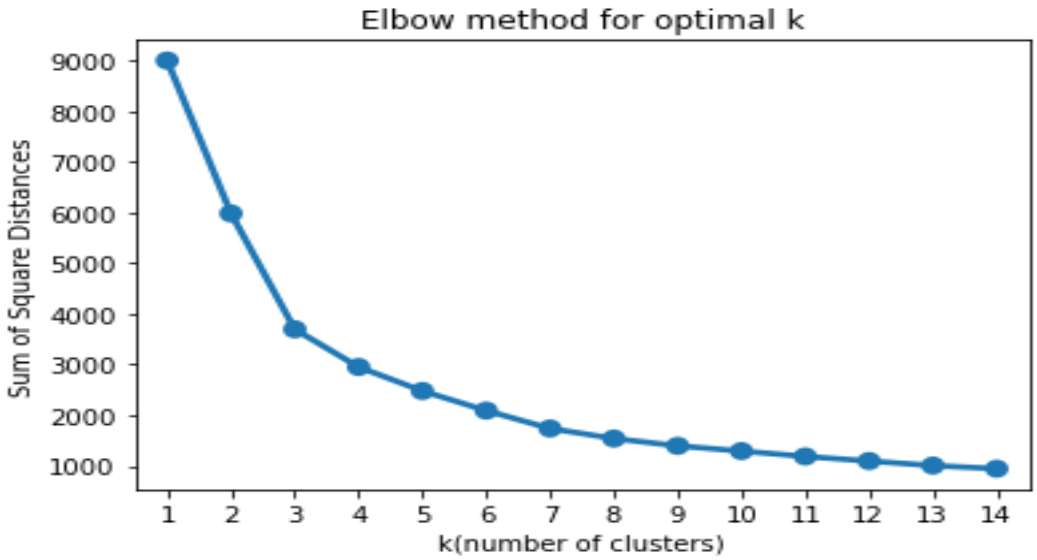
III. Cluster into segments using K-Algorithm

There are three main methods for clustering: partitioning (k-means), hierarchical and Density-based clustering (DBSCAN). Even though all of three techniques are important, K-means clustering seems to be appropriate for the transactional data due to it's fast and robust nature. Interpretability and efficiency are two important traits of this algorithm. With distinct data sets, it produces the best results even in the tighter clusters. Moreover, it uses initialization to find centroids and work best with the spherical clusters. Nevertheless, it has some limitations too. It needs us to specify number of clusters, i.e., k, which can be difficult to predict beforehand. It misinterprets overlapping data sets and can fail in distinguishing them. Randomly chosen centroids may sometimes do not produce accurate results. Moreover, k-means neither handles noisy nor non-linear data.

IV. Elbow method to determine the number of clusters

A technique to find optimal number of clusters in a data set, where the relations between number of clusters (k) and sum of square distances is measured. The shape of this curve is like an elbow. The elbow point shows the

optimal number of clusters. After applying this method on our data set, we found that the elbow point is 3, i.e., k=3. Therefore, after third cluster, sum of square distance is almost flat.



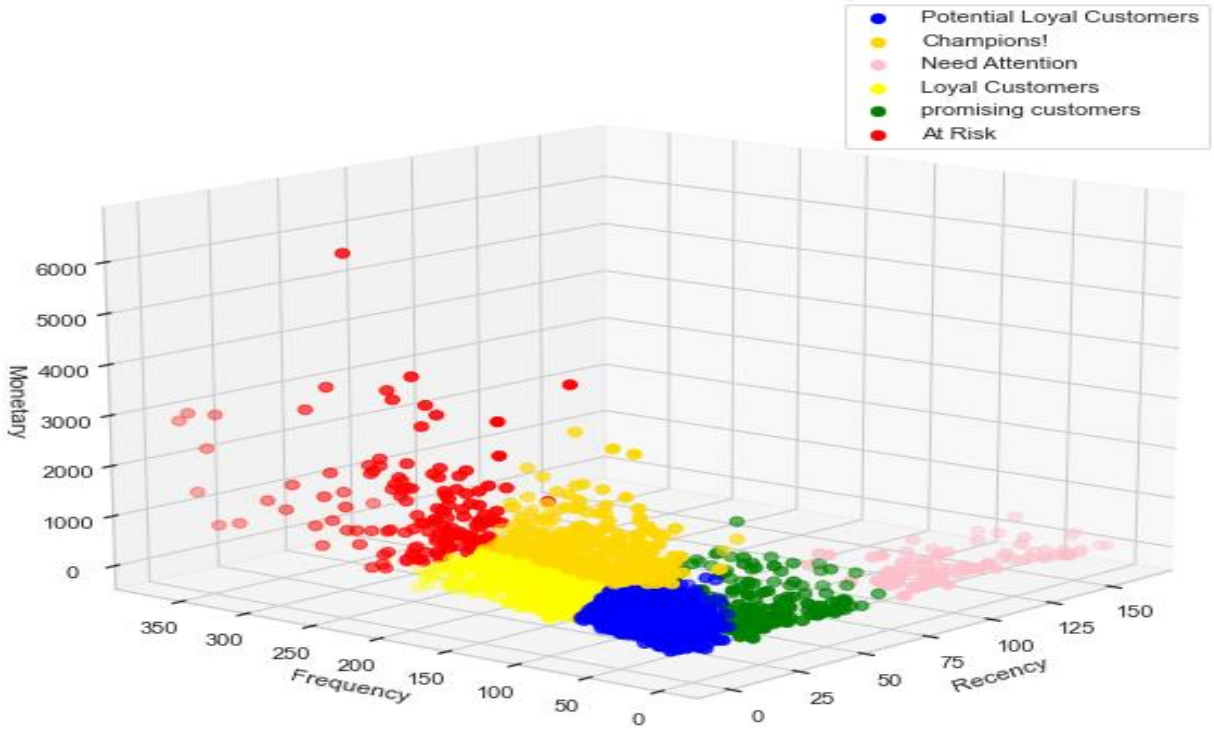
V. Evaluating K-means clustering algorithm using Silhouette Score

Choosing k is crucial for accurately separating data points. Another technique to know the optimal number of clusters is silhouette score. According to a senior data scientist, Tushar Joshi, "Silhouette Score is a metric to evaluate the performance of clustering algorithm. It uses compactness of individual clusters (***intra cluster distance***) and separation amongst clusters (***inter cluster distance***) to measure an overall representative score of how well our clustering algorithm has performed."

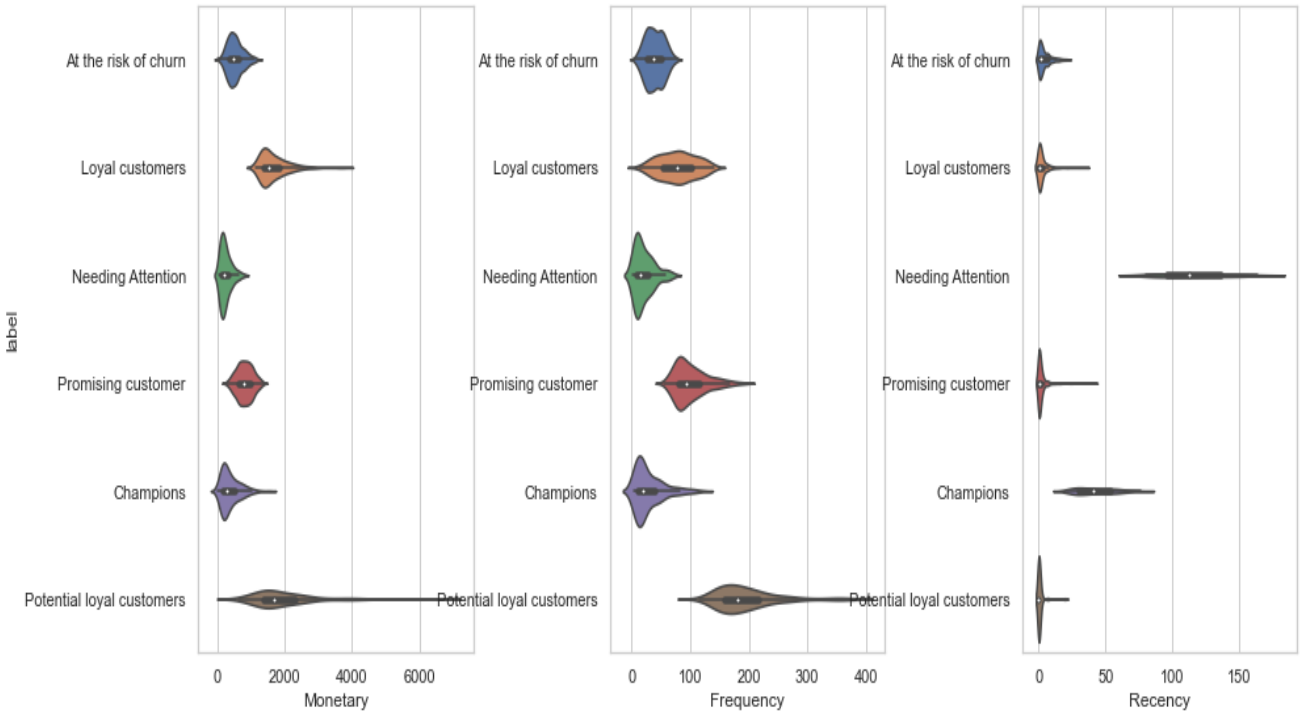
Silhouette score confirms the optimal number of clusters to be 3 with highest score of 0.50. However, we have produced six clusters to fulfill the demand of marketing manager. The accuracy of cluster has marginally decreased from 0.5 to 0.42.

```
For n_clusters = 2. The average silhouette_score is : 0.439041250116795)
For n_clusters = 3. The average silhouette_score is : 0.5074719882904812)
For n_clusters = 4. The average silhouette_score is : 0.38250601622816705)
For n_clusters = 5. The average silhouette_score is : 0.3944860717382693)
For n_clusters = 6. The average silhouette_score is : 0.4259609962902588)
For n_clusters = 7. The average silhouette_score is : 0.36426134122873655)
For n_clusters = 8. The average silhouette_score is : 0.3673591853918562)
For n_clusters = 9. The average silhouette_score is : 0.35227062063941944)
```

4. Result Section: 3D visualization of six clusters



i. Pan-portraits of clusters



II. Analysis and description of results

a. Segment 0: Potential Loyal customers

Potential loyal customers are those who buy frequently and that in massive amounts. However, their recency is little lower than those with loyal and champion categories. It is noticeable that they consist of more than 50 percent of all customers. Loyalty card option can increase their recency. With little bit of special treatment, this massive cluster of customers can turn into champions. The results can be substantial in sell and profit for the company.

Segment 0

	Cluster	customer_number	Recency	Frequency	Monetary
count	1533.0	1533.000000	1533.000000	1533.000000	1533.000000
mean	0.0	8263.380300	4.030659	38.714286	515.768904
std	0.0	4972.132628	4.644795	15.560479	228.048812
min	0.0	14.000000	0.000000	4.000000	18.360000
25%	0.0	3617.000000	1.000000	27.000000	350.800000
50%	0.0	8325.000000	2.000000	38.000000	484.720000
75%	0.0	12711.000000	6.000000	51.000000	647.150000
max	0.0	16316.000000	22.000000	78.000000	1222.340000

b. Segment 1: Champions!

This cohort is a group of the best customers. They not only bought recently, but also buy frequently and that too the most expensive items from the store. The management should give rewards to these customers. The higher the credibility of the store for them, the higher the chances of long customer life. Given their frequent visits (194 on average in just 6 months) and their average spending of they can promote new products better than any other cohort.

Segment 1

	Cluster	customer_number	Recency	Frequency	Monetary
count	151.0	151.000000	151.000000	151.000000	151.000000
mean	1.0	8218.940397	0.781457	194.953642	1907.346954
std	0.0	3673.662048	2.858657	51.487906	869.666041
min	1.0	263.000000	0.000000	116.000000	625.160000
25%	1.0	5252.500000	0.000000	158.500000	1389.400000
50%	1.0	8814.000000	0.000000	180.000000	1664.980000
75%	1.0	11072.000000	0.000000	216.000000	2298.930000
max	1.0	16292.000000	20.000000	374.000000	6588.650000

c. Segment 2: Loyal Customers

They are almost 24 percent (745) of the total customer base. Their average recency is second lowest, 1.73, and frequency second highest (100) in all the clusters. This category of customers has second highest monetary values as well. It's good news for the store are a sizeable segment. As prominent champions, they should be the target of constructive feedback and surveys. Store can lure them with extra rewards and can potentially quadruple the sale of the store.

Segment 2

	Cluster	customer_number	Recency	Frequency	Monetary
count	745.0	745.000000	745.000000	745.000000	745.00000
mean	2.0	7521.267114	1.731544	100.316779	803.22404
std	0.0	4390.004865	3.669899	27.697949	231.05510
min	2.0	119.000000	0.000000	55.000000	255.63000
25%	2.0	3953.000000	0.000000	80.000000	632.32000
50%	2.0	7131.000000	1.000000	94.000000	799.67000
75%	2.0	10950.000000	2.000000	115.000000	975.57000
max	2.0	16287.000000	42.000000	195.000000	1363.15000

d. Segment 3: At risk of churn

Segment 3 customers, which are just 80 in number, have higher chances of churn. Their last shopping with the store was 115 days on average. Besides, they were not the regular buyers when they used to buy. Monetary values of their purchase was lowest, 236 pounds on average, among all the segments. Therefore, they are at the risk of churning. Nevertheless, these are not the valueless customers. On average, they have spend 236 pounds in the six months. They might not be a lot in numbers, it worth sending them a reminder through text or an email for their last customer experience.

Segment 3

	Cluster	customer_number	Recency	Frequency	Monetary
count	80.0	80.000000	80.000000	80.000000	80.000000
mean	3.0	8563.925000	115.712500	20.237500	236.181500
std	0.0	5168.062722	24.395576	16.647685	162.946085
min	3.0	110.000000	80.000000	1.000000	48.830000
25%	3.0	3441.500000	95.500000	7.000000	114.865000
50%	3.0	9540.000000	113.000000	14.000000	193.745000
75%	3.0	12908.250000	136.500000	28.500000	318.187500
max	3.0	16100.000000	164.000000	70.000000	784.760000

e. Segment 4: Promising Customers (Affluent class)

These customers may not have a higher frequency, but they spent a decent amount of money when they have done shopping. Given their total size of 332 in the customer base and monetary value of 1667 pounds on average in the period of six months, their frequent visits will increase the total sales by a large amount. Thus, marketing manager can get in touch with them on call to improve their shopping experience. The store may offer combo products as an incentive. Their purchase experience can be improved by emailing them. Based on their advice, new quality- and maybe luxury - products can be introduced in new product line. Moreover, the store should make them aware of the new offering to this segment of customers.

Segment 4

	Cluster	customer_number	Recency	Frequency	Monetary
count	332.0	332.000000	332.000000	332.000000	332.000000
mean	4.0	7673.506024	2.454819	77.331325	1667.663886
std	0.0	3962.948368	4.428117	30.906718	423.740572
min	4.0	67.000000	0.000000	12.000000	1134.690000
25%	4.0	4454.750000	0.000000	52.000000	1363.647500
50%	4.0	7785.000000	1.000000	78.000000	1540.730000
75%	4.0	10782.250000	3.000000	101.250000	1854.652500
max	4.0	16203.000000	35.000000	140.000000	3764.810000

f. Segment 5: Need Attention

They can neither be included in churners nor in regular customers. Their average recency score tells that it has been 42 on average when they arrive for last spending. It's a short period to judge whether they will not come again. Past data shows that they have bought 28 times in the last six months, which cannot be neglected. It's not late for the advertisement manager to get in touch with them. With a little bit attention, they can return to buy again.

Segment 5

	Cluster	customer_number	Recency	Frequency	Monetary
count	159.0	159.000000	159.000000	159.000000	159.000000
mean	5.0	9699.930818	42.176101	28.647799	375.033333
std	0.0	4732.420399	14.161606	25.030902	278.998530
min	5.0	149.000000	21.000000	3.000000	7.280000
25%	5.0	5320.000000	30.000000	10.000000	167.610000
50%	5.0	11066.000000	41.000000	19.000000	275.760000
75%	5.0	13554.500000	53.500000	40.000000	528.400000
max	5.0	16184.000000	76.000000	120.000000	1536.300000

4. Efficacy of Insights and recommendations

Knowing your customer matters for higher sales, target marketing, brand awareness, and reputation of the brand in the market. An extensive 2017 study of Mulchy and Salon revealed, "No matter how brilliant the idea is, if your audience's needs are not met, you won't keep them as customers." Sales increase as you target the group of customers based on their previous transactional data. The most valuable group identified as champions will play a crucial role in brand awareness. As frequent visitors, they will be early adopters of new products. Secondly, loyal customers are key to upsell high value products. They should be engaged to increase the life-time value customers. Thirdly, potential loyalists need a little more personalized effort in marketing to turn them into loyalists. Any membership deal or discount will go a long way in term of sell volume as they consist of more than 50 percent of all 3000 customers. Fourthly, promising customers buy high value products, but they are not consistent. Giving them free trails can incentivize them to visit the store and buy complementary goods. At the risk of churning class should not be neglected. The data manger can reach out to them to better the customer journey for future clients. Those who have not come back can be crucial for introspection and improvement. Lastly, there are some customers who need attention because of they can be potential returners. Therefore, management should give limited time offers based on their previous basket categories. Last but not least, identified rich customers are promising in terms of sell numbers. Luxury items, new categories, affluent-base buying patterns will improve their frequency and sell numbers.

References:

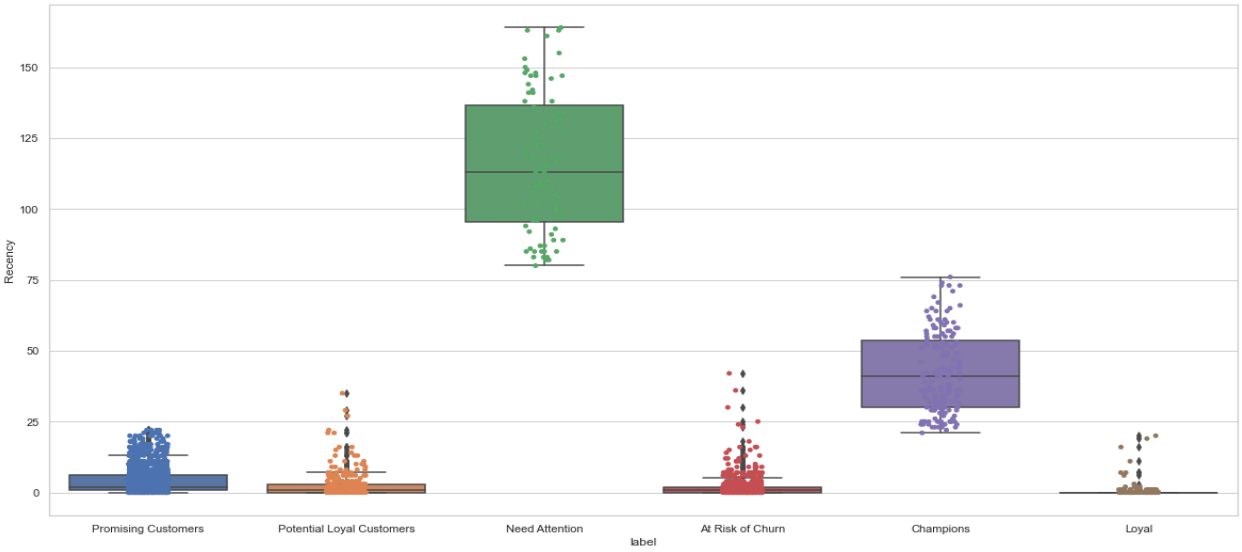
1. Joshi, T. (2021). *Silhouette Score*. [online] Medium. Available at: https://tushar-joshi-89.medium.com/silhouette-score-a9f7d8d78f29 [Accessed 21 Mar. 2022].

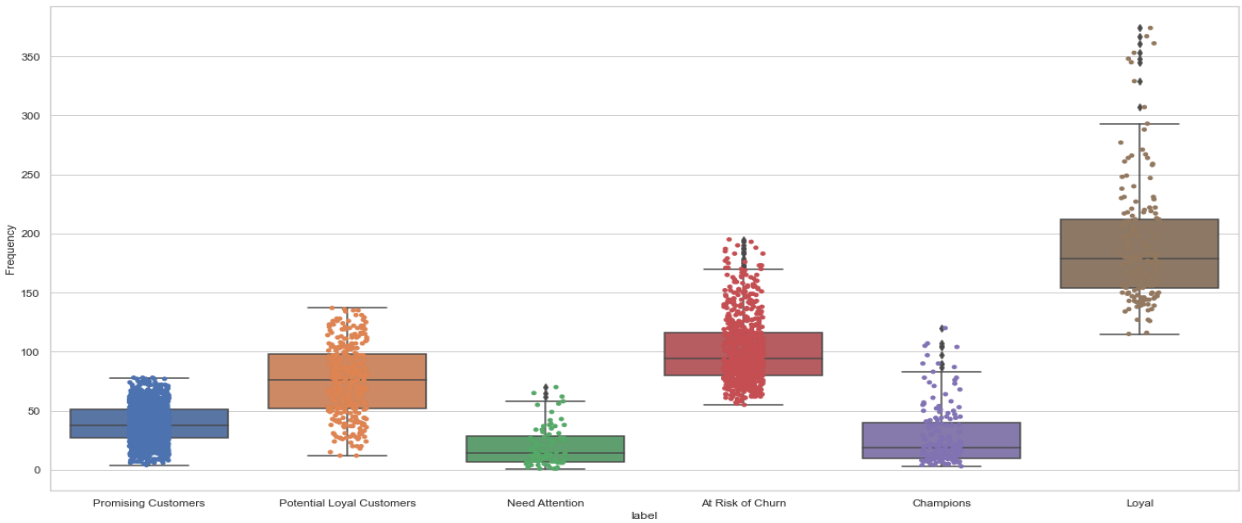
2. Anon, (2017). *RFM Analysis For Successful Customer Segmentation - Putler*. [online] Available at: https://www.putler.com/rfm-analysis/#Summary_of_RFM_segmentation_pros_cons_recommendations [Accessed 21 Mar. 2022].

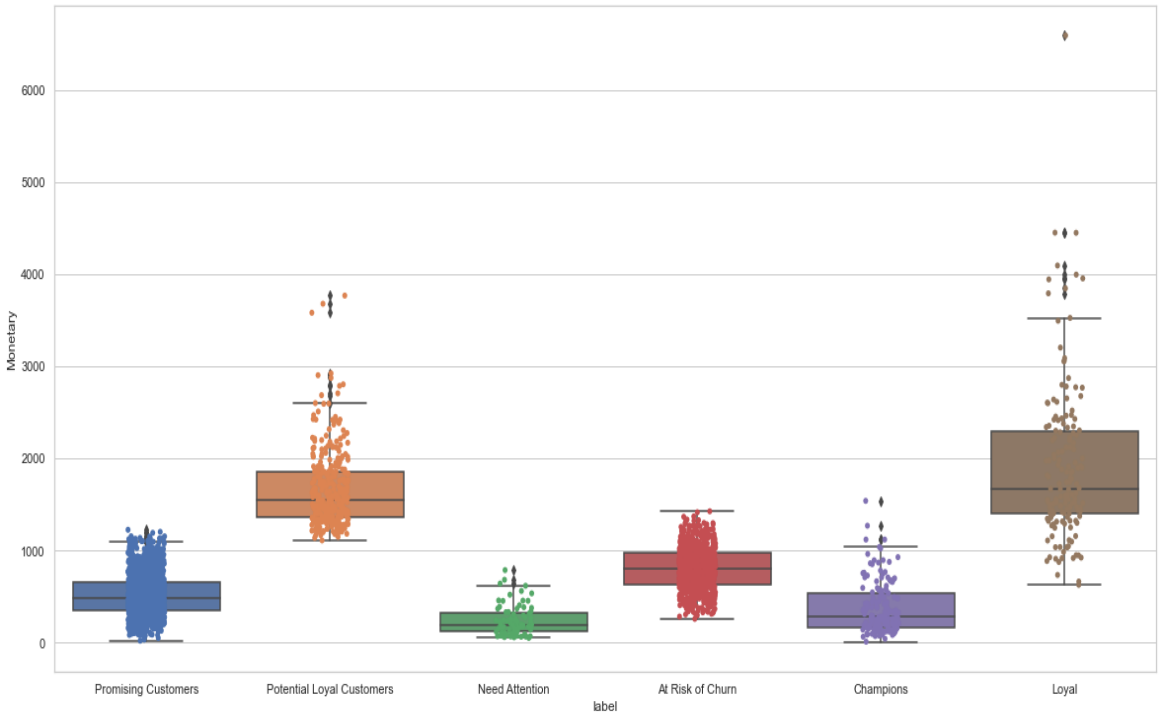
3. The AI University (2019). *Customer Segmentation using RFM K-Means & Python | Who are your Loyal Customers ? YouTube*. Available at: https://www.youtube.com/watch?v=fdUofaT8gUw.

Appendix

1. Box plots of customers






6000
5000
4000
3000
2000
1000
0
Monetary
Promising Customers
Potential Loyal Customers
Need Attention
At Risk of Churn
Champions
Loyal
label