CAPSTONE PROJECT

ONLINE RETAIL – CUSTOMER SEGMENTATION

BY
MANASA S

TABLE OF CONTENTS

PROBLEM STATEMENT

A UK-based and registered non-store online retail is trying to understand the various customer purchase patterns for their firm to improve their sales. They want to focus on the customer requirement to increase the sales. Obtained data has the purchasing history of customers which contains all the transactions occurred between 01/12/2010 and 09/12/2011. The company mainly sells unique all-occasion gifts.

The aim is to give an understanding about the customers based on their purchasing trends.

PROJECT OBJECTIVE

1. It is required to give enough evidence based insights to understand the various customer purchase patterns for the company.
2. To find useful insights about the customer purchasing history that can be an added advantage for the online retailer.
3. The aim is to segment the customers based on their purchasing behavior using RFM so that the company can target its customers efficiently.

DATA DESCRIPTION

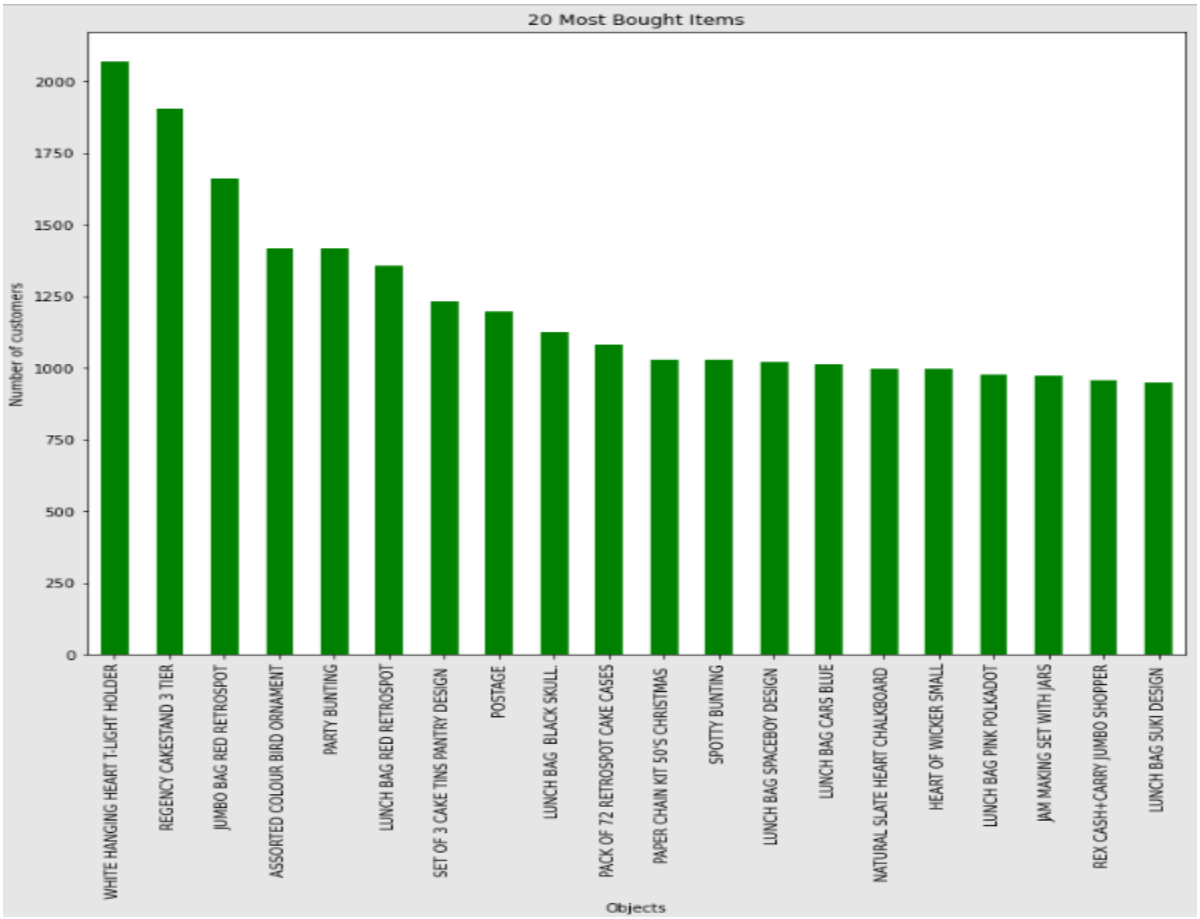
The dataset ‘ONLINE RETAIL.csv’ consists of 541909 rows and 8 columns.

S.No	FEATURE NAME	DESCRIPTION
1.	Invoice No	A 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.
2.	Stock code	Product (item) code. A 5-digit integral number uniquely assigned to each distinct product.
3.	Description	Product/ item description
4.	Quantity	The quantities of each product (item) per transaction.
5.	Invoice date	The date and time when a transaction was generated.
6.	Unit Price	Price of the product per unit in sterling.
7.	Customer ID	Customer number, a 5-digit integral number uniquely assigned to each customer.
8.	Country	The name of the country where a customer purchases.

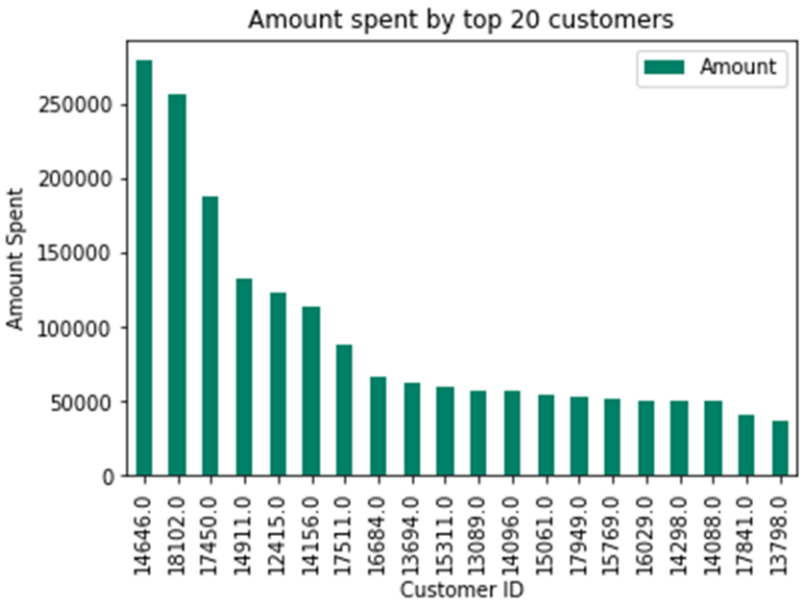
DATA PREPROCESSING STEPS AND INSPIRATION

1. **EDA:**

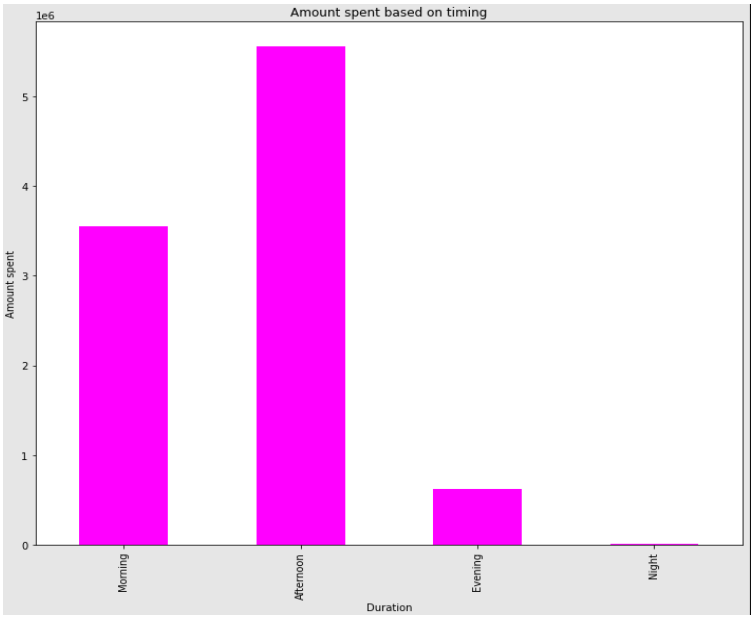
   a. Most popular products among customers.



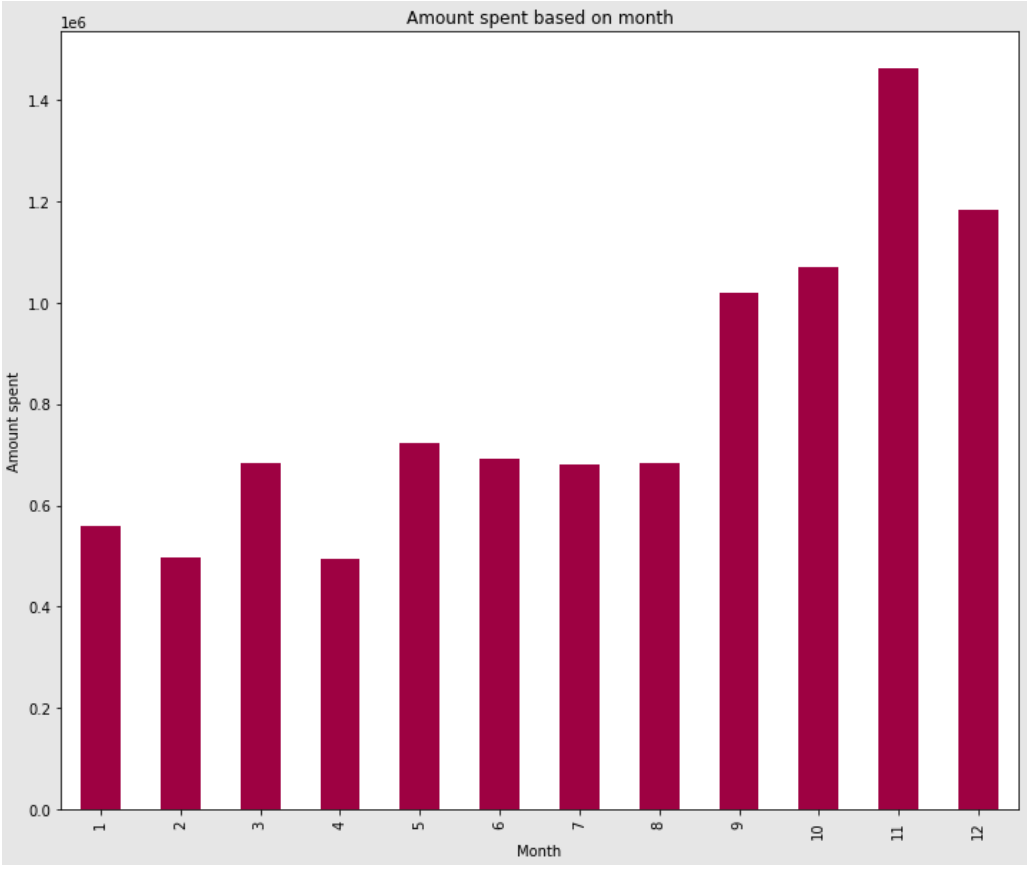
   b. The amount spent by top 20 customers' detail.



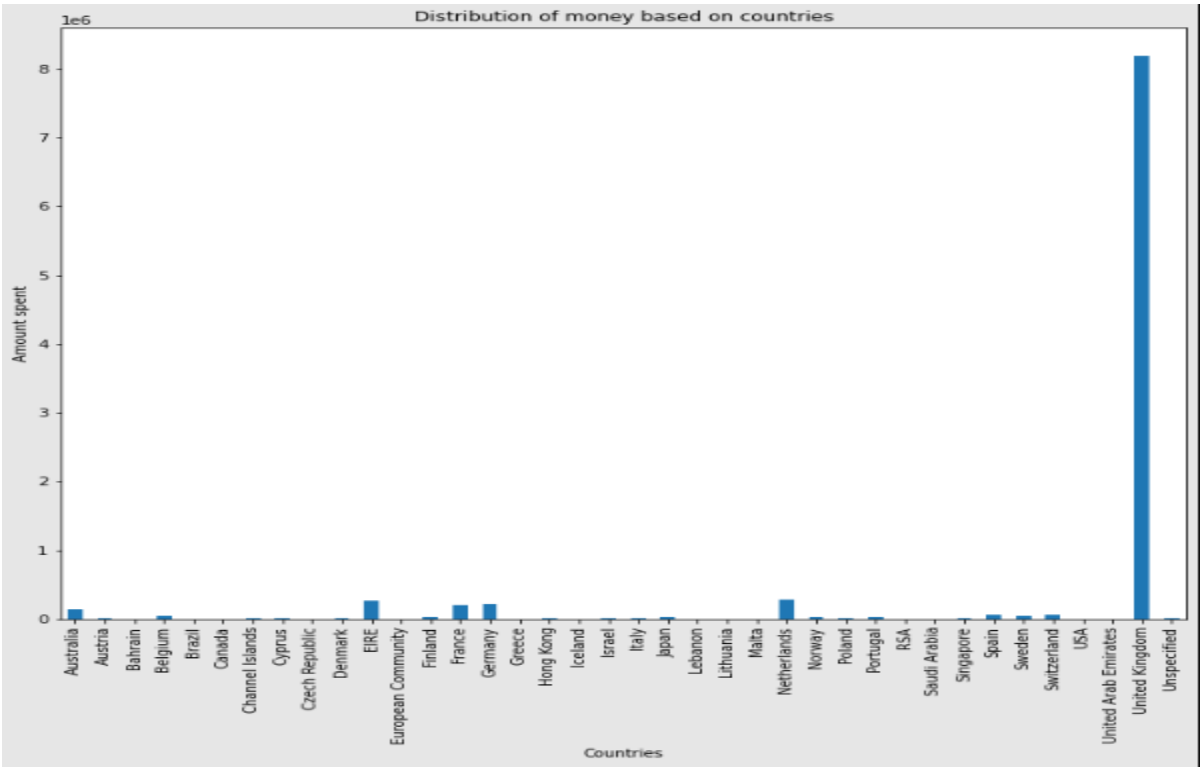
c. The maximum sales have happened during 11 - 16 hours (i.e. afternoon hours)



d. November month has seen very high sales than any other month followed by December.



e. UK is seen to have the highest sales.



2. Quantity and Unit Price are negative indicating that items are cancelled. So remove the data with stock code starting with 'C' as they indicate cancelled products.

	Quantity	UnitPrice	CustomerID
count	541909.000000	541909.000000	406829.000000
mean	9.552250	4.611114	15287.690570
std	218.081158	96.759853	1713.600303
min	-80995.000000	-11062.060000	12346.000000
25%	1.000000	1.250000	13953.000000
50%	3.000000	2.080000	15152.000000
75%	10.000000	4.130000	16791.000000
max	80995.000000	38970.000000	18287.000000

3. Amount is calculated by multiplying unit price and quantity.

	Amount
CustomerID	
14646.0	279489.02
18102.0	256438.49
17450.0	187482.17

4. Convert Invoice date to date time format and capture number of days, months, year and time separately. Based on time, we will further classify if it is morning, afternoon, evening or night.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Amount	day	year	month	hours	duration
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	15.30	1	2010	12	8	Morning
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34	1	2010	12	8	Morning
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	22.00	1	2010	12	8	Morning

5. While investigating the missing values it is noticed that customer id column alone has missing values. So it's better to drop them as the problem in hand is about customer segmentation.

```
InvoiceNo          0
StockCode          0
Description        0
Quantity           0
InvoiceDate        0
UnitPrice          0
CustomerID     132220
Country            0
Amount             0
dtype: int64
```

6. From the data the Recency, Frequency and Monetary aspects of the customers have been extracted.
   i. Recency : Number of days since last purchase
   ii. Frequency : Number of tracsactions
   iii. Monetary : Total amount of transactions (revenue contributed)

	CustomerID	Monetary	Frequency	Recency
0	12346.0	77183.60	1	325
1	12347.0	4310.00	182	1
2	12348.0	1797.24	31	74
3	12349.0	1757.55	73	18
4	12350.0	334.40	17	309

7. Different dataframes one with outliers removed, one with log normalized values and scaled using MinMax scaler and Standard Scaler have been prepared for the analysis.

CHOOSING THE ALGORITHM FOR THE PROJECT

Multiple algorithms were tested on different dataframes obtained by varied scalings.

1. KMeans :

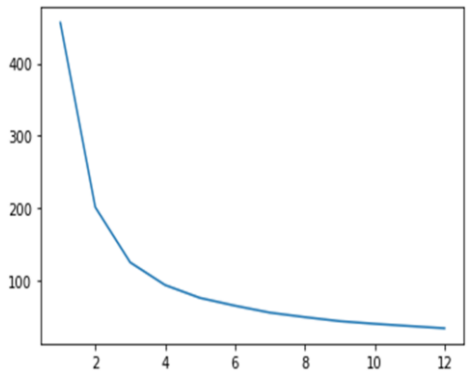
   Here data was clustered into 3 groups(found using elbow plot) and the algorithm was fit on different dataframes(one with standa rd scaled and min max scaled) to obtain the best silhouette and Calinski Harabasz score. It is observed that min max scaled data seems to perform better in light of both the scores.

Silhouette_score:

	Cluster	Silhouette_score
0	2	0.572607
1	3	0.553103
2	4	0.490754
3	5	0.442792
4	6	0.389862
5	7	0.394894
6	8	0.375654
7	9	0.376103
8	10	0.376082
9	11	0.381286
10	12	0.365618

Calinski Harabasz score:

	Cluster	CH score
0	2	5394.060086
1	3	5646.177370
2	4	5503.097589
3	5	5349.078348
4	6	5128.913824
5	7	5135.182784
6	8	5033.055300
7	9	5039.601909
8	10	4933.096051
9	11	4849.809287
10	12	4843.207922



2. Agglomerative clustering:

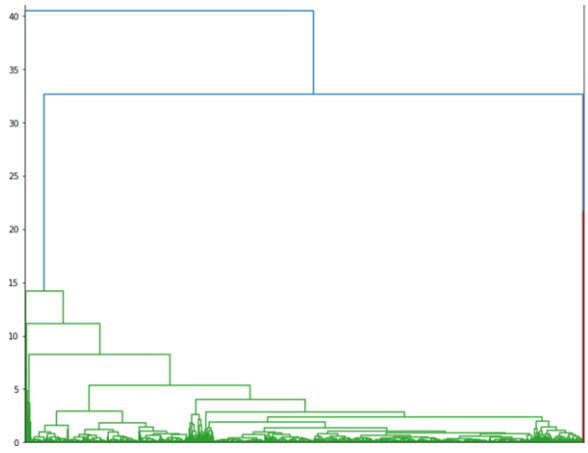
Here various dataframes were used to fit agglomerative clustering algorithm. The data which was standard normalized was seen to have a very high silhouette and Calinski – Harabasz score for n_clusters=3. But, more than 90% of the data points are in one cluster. Even with minmax scaled data and n_clusters = 4 there was no big improvement.

Silhouette score:

	cluster	Silhouette score
0	2	0.945553
1	3	0.929693
2	4	0.911286
3	5	0.771116
4	6	0.770826
5	7	0.770968
6	8	0.716933
7	9	0.716937
8	10	0.697164
9	11	0.694917
10	12	0.694754

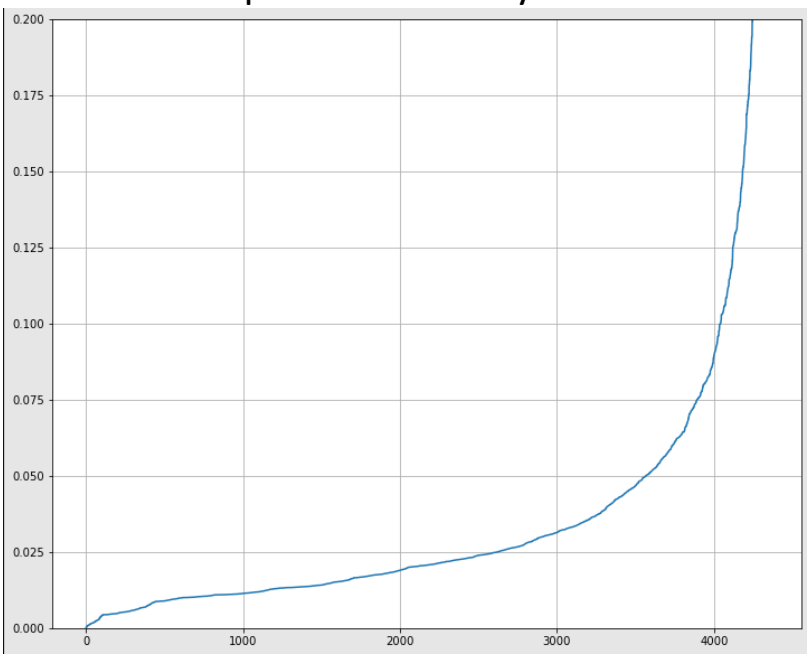
Calinski Harabasz score:

	cluster	CH score
0	2	1155.004312
1	3	1569.682455
2	4	1153.166374
3	5	1053.444119
4	6	865.703615
5	7	747.156494
6	8	755.511251
7	9	679.758396
8	10	613.105839
9	11	569.855096
10	12	522.572276



3. DBSCAN:

The parameter eps was found by k neighbors graph and minimum samples in each cluster was got by log(number of datapoints). Also RandomsizedsearchCV was performed(which did not return the best parameters). The silhouette scores suggests that dbscan was not able to cluster the points efficiently.




```
1 db=DBSCAN(eps=0.085,min_samples=9)
```

```
1 db.fit(data_std)
```

```
DBSCAN
DBSCAN(eps=0.085, min_samples=9)
```

```
1 set(db.labels_)
```

```
{-1, 0, 1, 2, 3}
```

```
1 silhouette_score(data_std,db.labels_)
```

```
-0.19595881465783263
```

I have chosen the KMeans algorithm for the data in hand as the optimal number of clusters obtained, the Silhouette score and Calinksi Harabasz score seems to be in place and inferable than the other models built. The elbow graph has suggested a cluster of 3 whose inferences have been recorded below.

The algorithm works by calculating the distance between any 2 points in the cluster and points which are closer to the centroid are grouped as a single cluster.

ASSUMPTIONS

1. K-Means:
   (a) The clusters are spherical.
   (b) Clusters are of similar size.
   (c) Variables are uncorrelated within clusters.
2. DBSCAN:
   (a) Clusters are dense regions in space separated by some space.

MODEL EVALUATION AND TECHNIQUE

The model was evaluated using:
(a) Silhouette score
(b) Calinski Harabasz score
(c) Boxplots to visualize the clusters

```
1 label=km_model_eval(3,data_out_rem_minmax)
```

```
Cluster Labels:
 [2 0 0 ... 0 2 0]
```

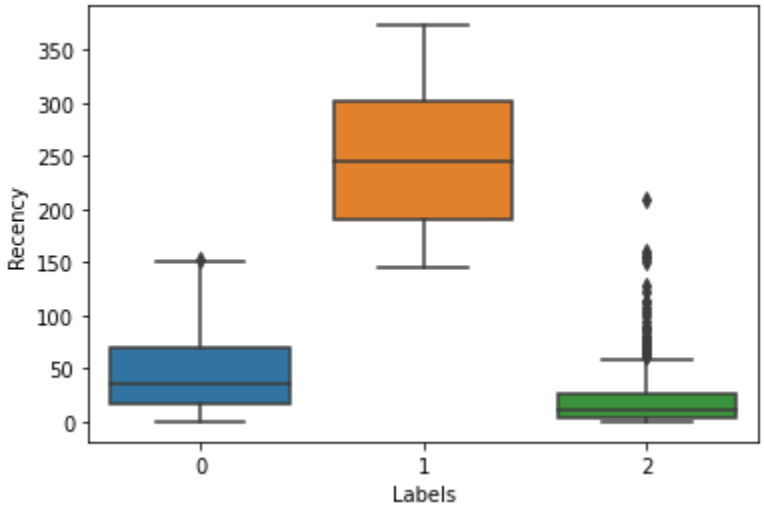
```
1 np.bincount(label)
```

```
array([2718, 1041,  496], dtype=int64)
```

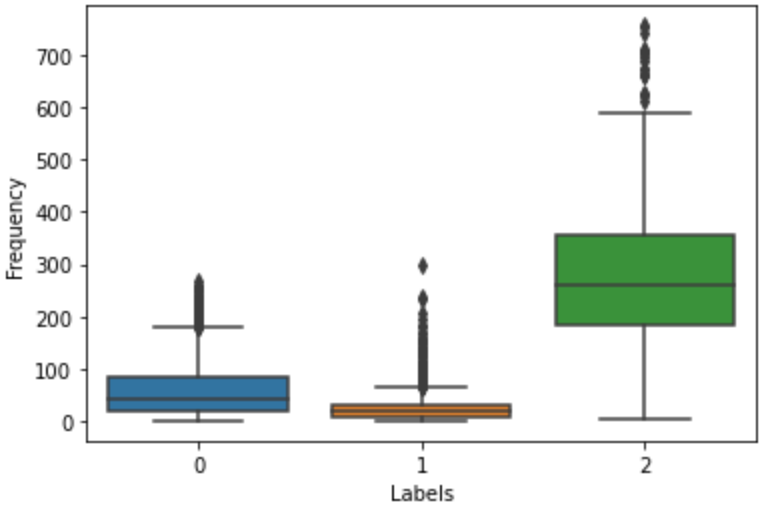
	CustomerID	Recency	Frequency	Monetary	Labels
1	12347.0	1	182	4310.00	2
2	12348.0	74	31	1797.24	0
3	12349.0	18	73	1757.55	0
4	12350.0	309	17	334.40	1
5	12352.0	35	85	2506.04	0

Inferences:

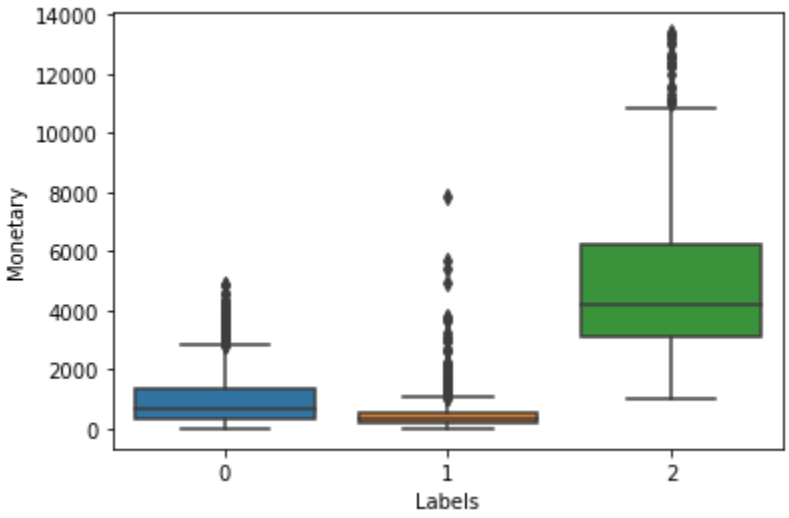
1. The customers in cluster 2 has the highest recency. The customers in cluster 0 have very low recency.



2. Cluster 0 has made the highest number of transactions. Cluster 2, on the other hand have made least transactions.



3. Cluster 0 customers have spent a good average amount towards the transaction. Cluster 2, on the other hand have not performed well towards the monetary aspect of the store, except a few outliers.



INFERENCES FROM THE PROJECT

1. Based on RFM scores obtained based on binning the Recency, Frequency and Monetary values, customers have been segmented into different categories like: Lost customers, Hibernating customers, Cannot Lose Them, At Risk, About to Sleep, Need Attention, Promising, New Customers, Potential Loyalist, Loyal, Champions.

	CustomerID	Monetary	Frequency	Recency	Monetary_score	Frequency_score	Recency_score	rfm_score	Status
0	12346.0	77183.60	1	325	5	1	1	115	Cannot Lose Them
1	12347.0	4310.00	182	1	5	5	5	555	Champions
2	12348.0	1797.24	31	74	4	3	2	234	At Risk
3	12349.0	1757.55	73	18	4	4	4	444	Loyal
4	12350.0	334.40	17	309	2	2	1	122	Hibernating customers

	Status	Number of Customers	Percentage of customers
0	About To Sleep	192	4.426003
1	At Risk	411	9.474412
2	Cannot Lose Them	89	2.051637
3	Champions	795	18.326418
4	Hibernating customers	810	18.672199
5	Lost customers	450	10.373444
6	Loyal	420	9.681881
7	Need Attention	226	5.209774
8	New Customers	302	6.961734
9	Potential Loyalist	510	11.756570
10	Promising	133	3.065929

2. At Risk customers:
   They are customers who purchased often and spent big amounts, but haven't purchased recently.
   - The customers at risk have not visited the shop for atleast 71 days and atmost 371 days.
   - There are totally 411 customers with an average spending of 1596.16 dollars.
   - On an average they have made 82 purchases.

Action : We can send them personalized reactivation campaigns to reconnect or offer renewals and helpful products to encourage another purchase.

	CustomerID	Monetary	Frequency	Recency	rfm_score
count	411.000000	411.000000	411.000000	411.000000	411.000000
mean	15228.559611	1596.165669	81.922141	147.827251	213.841849
std	1740.499624	1401.879791	60.912562	73.595280	48.201982
min	12348.000000	278.740000	14.000000	71.000000	124.000000
25%	13687.000000	781.190000	44.000000	88.000000	145.000000
50%	15303.000000	1145.840000	66.000000	123.000000	242.000000
75%	16730.000000	1832.565000	98.500000	189.000000	244.000000
max	18260.000000	11072.670000	543.000000	371.000000	255.000000

3. Cannot lose customers:
   They are customers who used to visit and purchase quite often, but haven’t been visiting recently. If we don't take any actions immediately, we will lose them.

- There are totally 89 customers who have not visited for atleast 73 days and atmost 371 days
- They have spent an average of 3335 dollars
- On an average they have made 52 purchases

Action : We can bring them back with relevant promotions, seasonal discounts to make them feel special or run surveys to find out what went wrong and avoid losing them to a competitor.

	CustomerID	Monetary	Frequency	Recency	rfm_score
count	89.000000	89.000000	89.000000	89.000000	89.000000
mean	15162.471910	3335.147652	51.932584	217.595506	147.101124
std	1769.721082	10035.792985	63.485288	75.476625	36.472606
min	12346.000000	516.420000	1.000000	73.000000	113.000000
25%	13715.000000	973.840000	5.000000	184.000000	114.000000
50%	15069.000000	1289.500000	12.000000	217.000000	144.000000
75%	16714.000000	1850.560000	91.000000	268.000000	155.000000
max	18239.000000	77183.600000	297.000000	371.000000	215.000000

4. New customers:

   They are customers who have a high overall RFM score but are not frequent shoppers.

- There are totally 302 customers who have visited at least 1 day ago and at most 71 days before
- On an average they have made 11 purchases
- They have spent 210 dollars on an average

Action: We can start building relationships with these customers by providing onboarding support and special offers to increase their visits.

	CustomerID	Monetary	Frequency	Recency	rfm_score
count	302.000000	302.000000	302.000000	302.000000	302.000000
mean	15311.711921	210.105861	11.102649	27.841060	400.248344
std	1724.493572	111.893360	6.907509	17.675778	66.162047
min	12367.000000	6.200000	1.000000	0.000000	311.000000
25%	13819.750000	125.362500	6.000000	16.000000	311.000000
50%	15277.500000	187.030000	10.000000	23.000000	411.500000
75%	16846.750000	284.632500	15.000000	38.000000	422.000000
max	18282.000000	486.720000	29.000000	71.000000	512.000000

5. Need attention customers:

   They don't frequently purchase, but they spend quite a good amount of money.

- There are 226 customers who have purchased atleast 15 days ago and atmost 121 days ago
- They have spent 1470 dollars on an average and have purchased 57 times on average

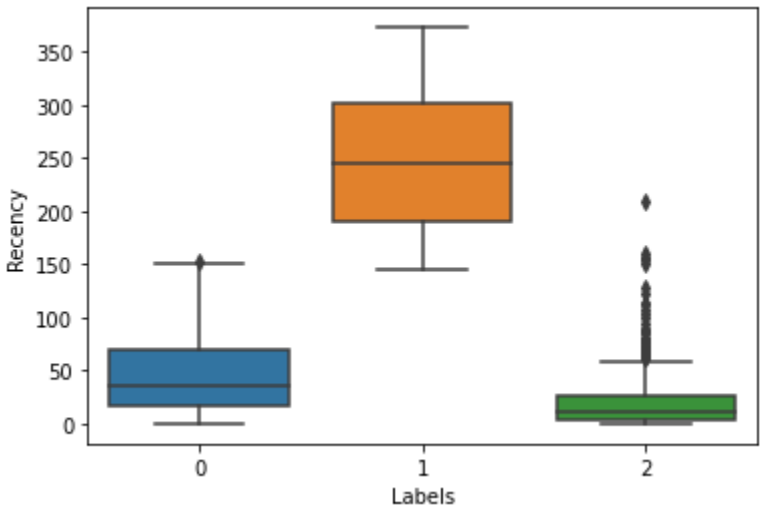
Action: In order to make them transform into a customer who purchases frequently, we can offer them some discount with a time limit of 30 days, so that they would revisit and purchase.

	CustomerID	Monetary	Frequency	Recency	rfm_score
count	226.000000	226.000000	226.000000	226.000000	226.000000
mean	15250.123894	1470.642389	57.252212	31.610619	414.238938
std	1719.928190	2533.784307	24.893428	20.336692	78.559223
min	12372.000000	490.520000	15.000000	0.000000	324.000000
25%	13730.250000	808.302500	40.000000	15.250000	334.000000
50%	15180.500000	1054.615000	51.000000	29.000000	434.000000
75%	16790.500000	1384.162500	72.000000	49.750000	443.000000
max	18252.000000	26879.040000	121.000000	71.000000	535.000000

6. Based on K-Means clustering the following labels were obtained:

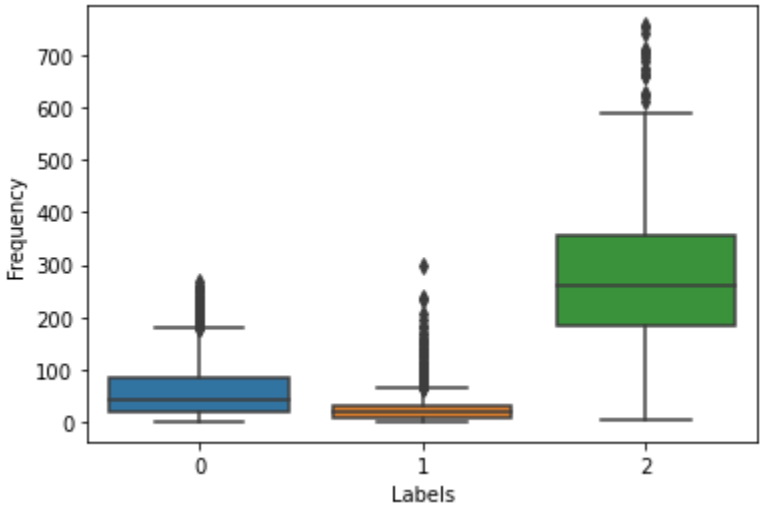
	CustomerID	Recency	Frequency	Monetary	Labels
1	12347.0	1	182	4310.00	2
2	12348.0	74	31	1797.24	0
3	12349.0	18	73	1757.55	0
4	12350.0	309	17	334.40	1
5	12352.0	35	85	2506.04	0

- The customers in cluster 2 has the highest recency which means they have to be concentrated upon to bring them back as loyal customers.
- Giving seasonal discounts to them or offering them coupons with a short time validity can make them come back and getting a review from them can help the turn to loyal customers
- The customers in cluster 0 have very low recency which suggests that they have been loyal customers so far. But the outliers in cluster 0 suggest that these customers are on the verge of being lost.

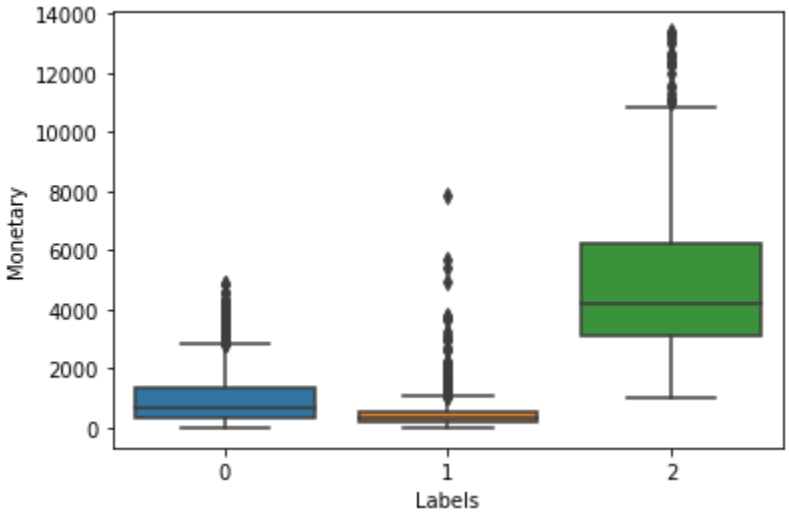


- Cluster 0 has made the highest number of transactions.
- Cluster 2 on the other hand; have made least transactions on an average and to bring them back.

- Ask feedback from them if they are in need of any new items in the store
- Give offers on specific products to increase the number of transactions and give a chance for them to buy back.



- Cluster 0 customers have spent a good average amount towards the transaction
- Cluster 2 on the other hand have not performed well towards the monetary aspect of the store, except a few outliers



FUTURE POSSIBILITIES

The models can be used with fresh incoming data points in future and the best algorithm can be chosen based on the Silhouette and Calinski Harabasz scores. However, centroids can be dragged by outliers, or outliers might get their own cluster instead of being ignored. K-means also has trouble clustering data where clusters are of varying sizes and density. The best model has to be selected each time based on its performance on the data.

CONCLUSION

In this project the customer segmentation is analyzed with a given data using multiple models and RFM techniques. Using RFM scores, customers were binned into different categories, hence giving the company an insight into the section of customers who has to be concentrated upon by the company. Also models like K-Means, Agglomerative clustering and DBSCAN were built on the data and checked for its ability to separate the data better.

REFERENCES

1. RFM Analysis for Customer Segmentation | CleverTap
2. Clustering Algorithms | Machine Learning | Google Developers