Customer Segmentation

Report

Andrew Wright

—

July 2022

Business Problem

In business, every deal might not be profitable to the customer, every client might not be interested to spend. They have asked that I look at past data and analyze how to segment the customers using Cohort Analysis, RFM, and K-Means Clustering to group the customers accordingly. The results will be presented to the company on the best way to group their customers and market to them more effectively.

Datasets

The dataset we will be using will be acquired from an online source: https://www.kaggle.com/datasets/jihyeseo/online-retail-data-set-from-uci-ml-repo

The dataset contains the following features: Invoice Number, Stock Code, Description of Item Sold, Quantity, Date of Invoice Unite Price, Customer ID and the Country of Sale.

Data Science Approach

1. Examine dataset for missing values, data types, outliers
2. Explore the data
3. Cohort Analysis
4. RFM
5. Prep Data for Clustering
6. Model with k-means
7. Document and report findings

Deliverables

1. Developed code via Jupyter notebooks
2. Final written report
3. Presentation Slide Deck
4. Dashboards

Data Wrangling

The raw dataset had 541, 909 rows and 9 columns

Dropped duplicated data

There were also negative values

I searched it and found that some of them were returns

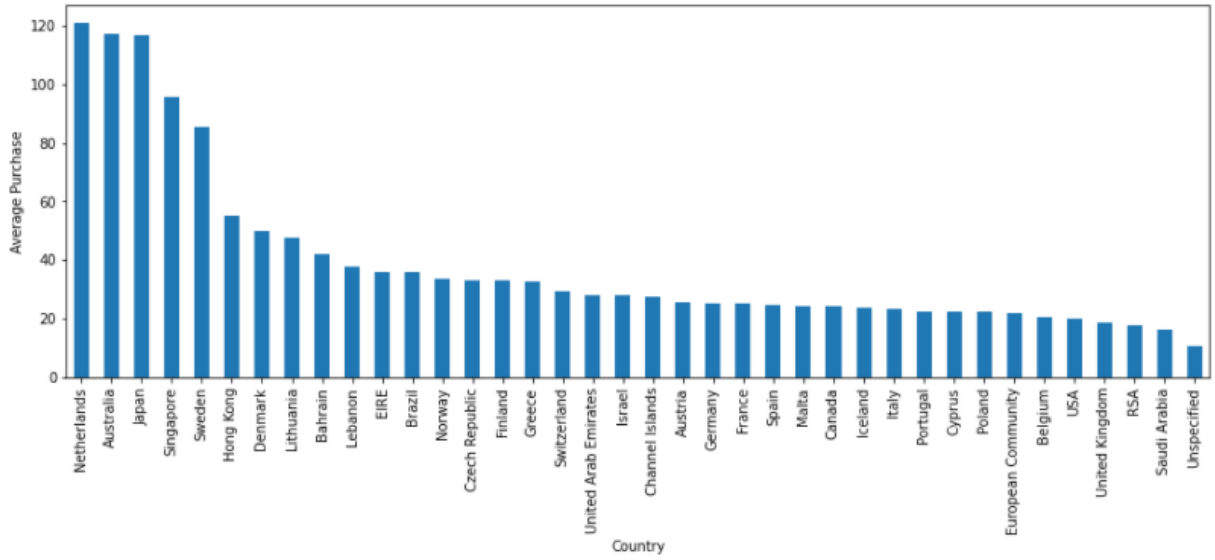
Removed rows where Quantity and Unit Price were below 0

	Quantity	UnitPrice	CustomerID	Total
count	536641.000000	536641.000000	536641.000000	536641.000000
mean	9.620029	4.632656	11435.904653	18.123861
std	219.130156	97.233118	6795.044250	380.656263
min	-80995.000000	-11062.060000	0.000000	-168469.600000
25%	1.000000	1.250000	0.000000	3.750000
50%	3.000000	2.080000	14336.000000	9.870000
75%	10.000000	4.130000	16241.000000	17.400000
max	80995.000000	38970.000000	18287.000000	168469.600000

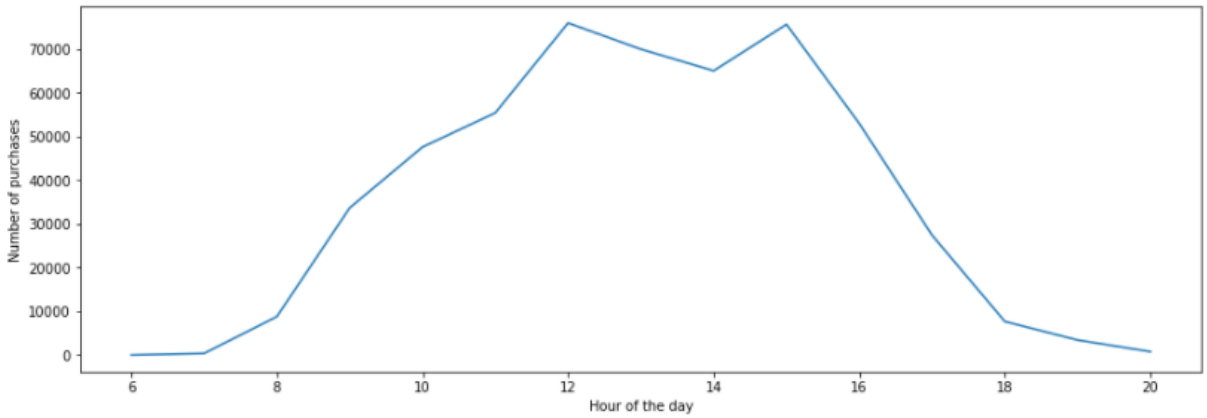
Exploratory Data Analysis

The dataset was analyzed for some trends.

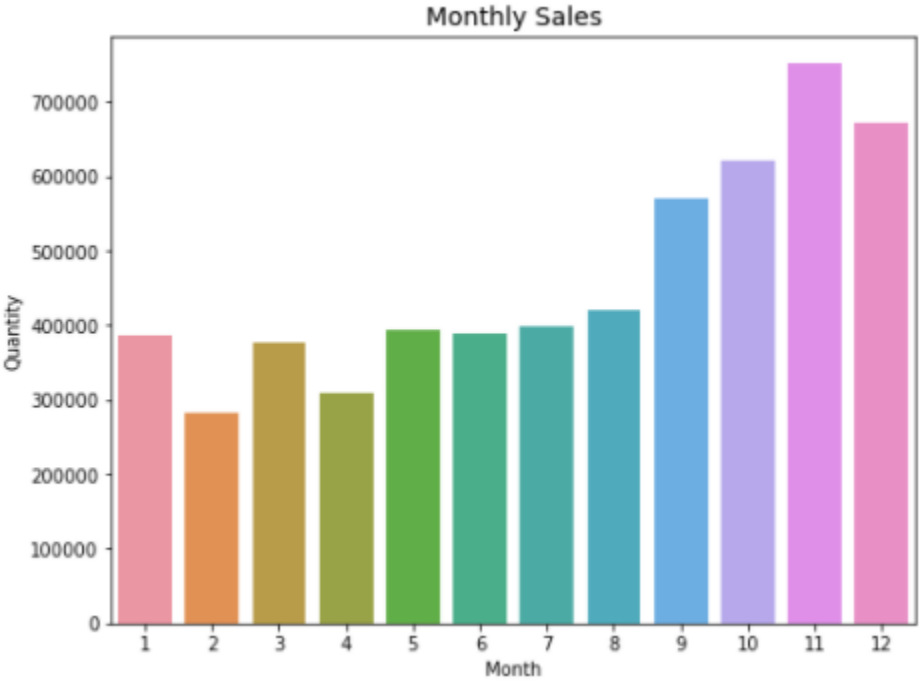
Average purchase was Netherlands



Most purchases were made during noon and 3pm

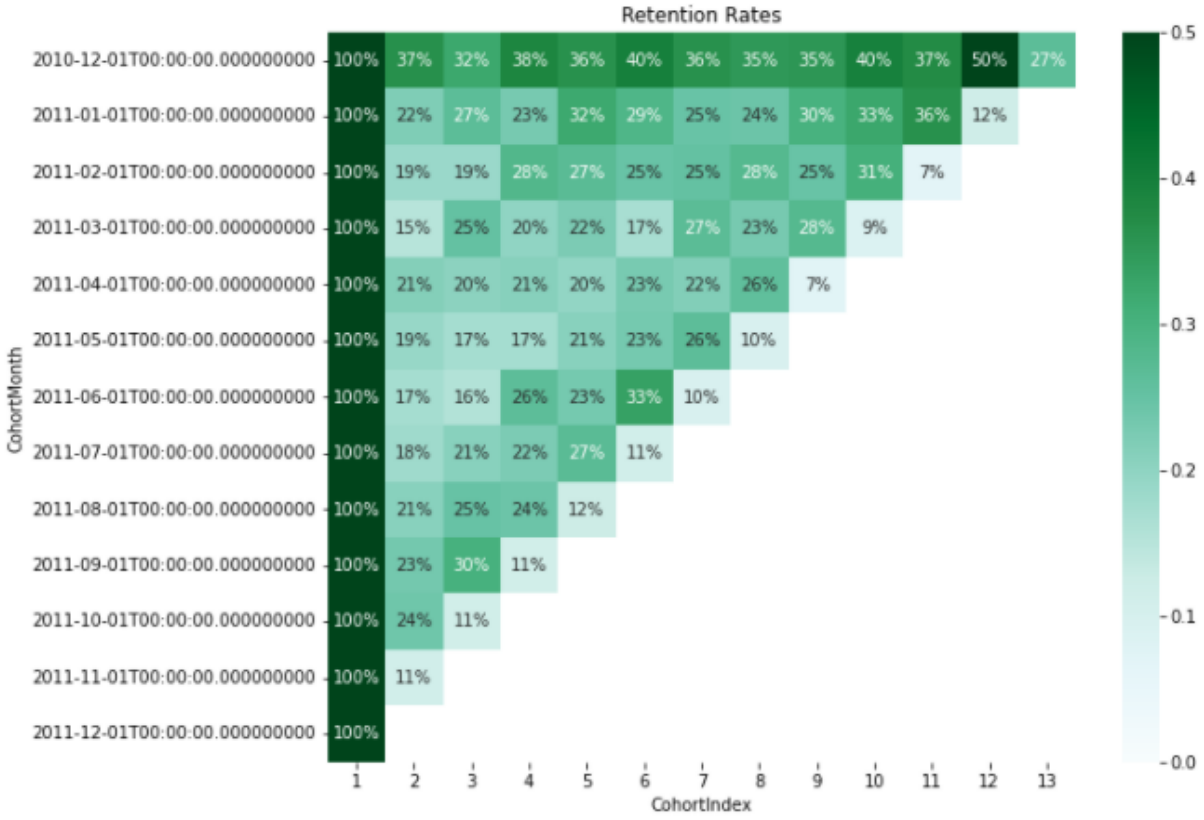


Most Purchases were made during the last 4 months of the year


Monthly Sales
Quantity
700000
600000
500000
400000
300000
200000
100000
0
1
2
3
4
5
6
7
8
9
10
11
12
Month

Cohort Analysis and RFM

A Monthly Cohort Analysis was created



The above heatmap shows the retention rates for the Monthly Cohort

The dataset was next broken into RFM and assigned labels

- RFM is an acronym of recency, frequency and monetary. Recency is about when was the last order of a customer. It means the number of days since a customer made the last purchase
- Frequency is about the number of purchases in a given period.
- Monetary is the total amount of money a customer spent in that given period.

Process of calculating percentiles:

1. Sort customers based on that metric

2. Break customers into a pre-defined number of groups of equal size
3. Assign a label to each group

CustomerID	Recency	Frequency	MonetaryValue	R	F	M	RFM_Score	RFM_Segment	RFM_Level
0	1	132186	1754901.91	3	3	3	9	3.03.03.0	Top
12346	326	1	77183.60	1	1	1	3	1.01.01.0	Low
12347	2	182	4310.00	3	3	3	9	3.03.03.0	Top
12348	75	31	1797.24	2	2	2	6	2.02.02.0	Middle
12349	19	73	1757.55	3	3	3	9	3.03.03.0	Top
...	...	...	...	...	...	...	...	...	...
18280	278	10	180.60	1	1	1	3	1.01.01.0	Low
18281	181	7	80.82	1	1	1	3	1.01.01.0	Low
18282	8	12	178.05	3	1	1	5	3.01.01.0	Low
18283	4	721	2045.53	3	3	3	9	3.03.03.0	Top
18287	43	70	1837.28	2	2	2	6	2.02.02.0	Middle

Pre-Processing and Training

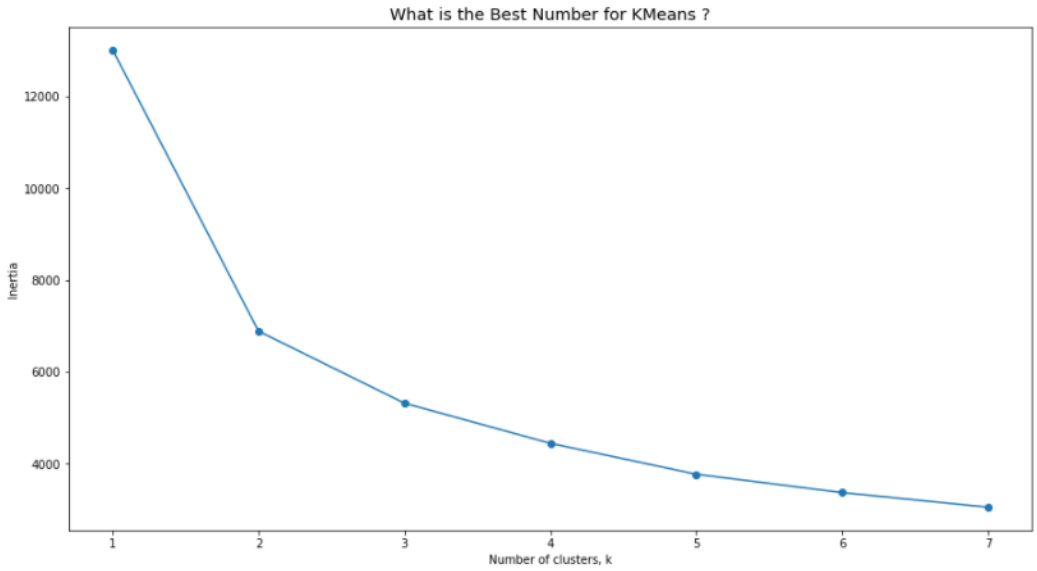
Monetary, Frequency, and Recency values were grouped by customer id. Plotting these values in a box-plot, showed we had outliers. Standard scalar was used to normalize the data.

K-Means Clustering

K-Means clustering was chosen to group the data.

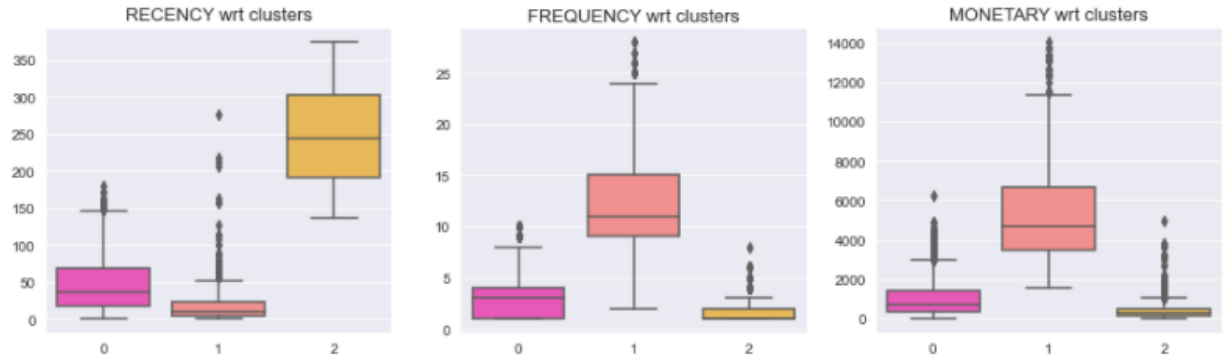
Kmeans algorithm tries to partition the dataset into K distinct clusters.

The elbow method was used to determine the optimal value of K



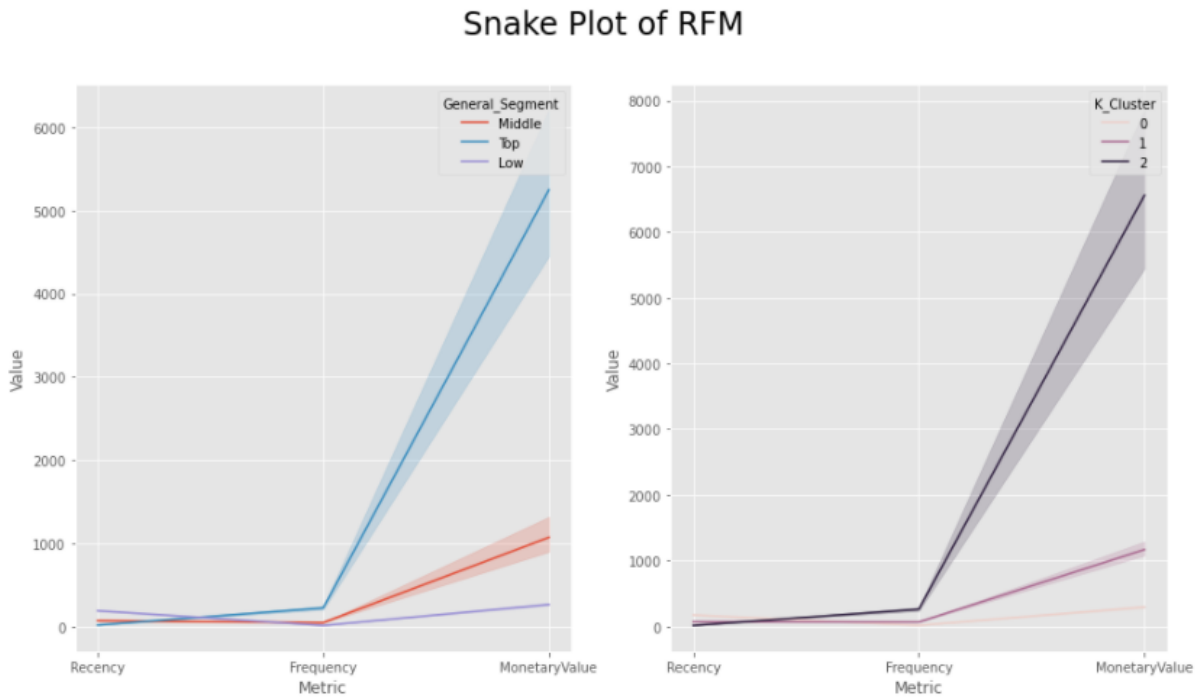
The optimal clusters appears to be 3

At this point the model was fitted to the data with 3 as the value of clusters



and now we can compare the results of using the RFM analysis and the K-Means clustering

Conclusion



As you can see from the snake plots above, the RFM and K-Means have a very similar result. The K-means definitely has a generally higher mean.

	Recency	Frequency	MonetaryValue	
	mean	mean	mean	count
RFM_Level				
Low	160.3	18.8	642.0	1926
Middle	52.9	84.5	1576.6	1591
Top	10.4	430.8	8390.9	822

	Recency	Frequency	MonetaryValue	
	mean	mean	mean	count
K_Cluster				
0	171.0	15.0	293.0	1523
1	69.0	65.0	1167.0	1859
2	13.0	260.0	6559.0	956

Both methods would work well to segment the customers into segments for marketing

Future Work

Create Tableau dashboard to model the results