RFM Analysis:
Segmenting Customers Wisely

Shane McCallum, Springboard Data Science Career Track, March 2021

Smart Problems Require Smart Solutions

Client wants to tailor marketing campaigns to their best, average, and weakest customers:

- How do we measure “Best?”
- How can future data be added?
- What are the features needed to determine the tiers or “segments?”

SOLUTION:

- Measure across the three most important attributes:
  - Recency (last purchase date)
  - Frequency (amount of purchases)
  - Monetary Value (total amount customer spends)
- Maintainable and reproducible

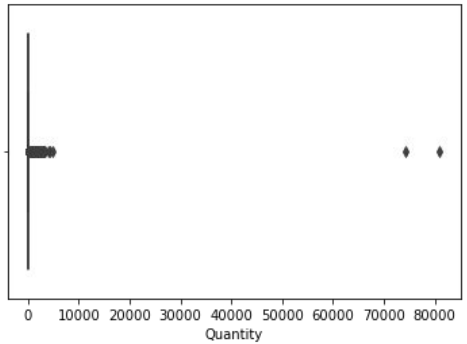
Who else wants this?

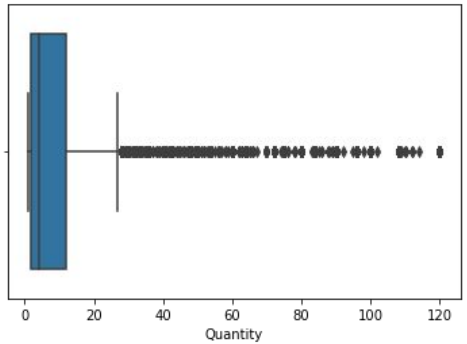
- Digital markets that record customer ID
- Physical stores that use loyalty programs
- Online magazines and subscriptions

Cleaning our Client's Data

From UCI Machine Learning Repository:

- Over 400,000 transactions;
  - 135,000 "guest" purchases - Dump it;
  - Returns - Dump it;
  - Extreme outliers - Dump it.
- Over a year of data;
  - Great for a Cohort Analysis
  - Easy to follow customers with IDs





Exploring the story of the Data

Purchases from Customers:

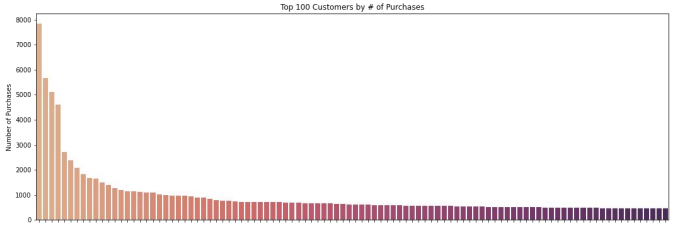
- Top 100 customers are likely businesses
- Nice gradient, not too drastic

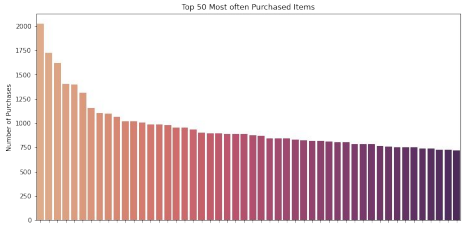
Most Purchase Products:

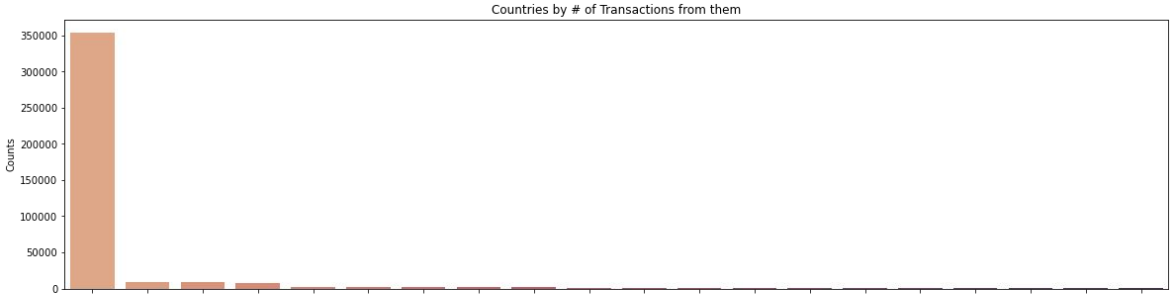
- Another clean gradient; not reliant on 1 “magic” product.

Countries by most transactions:

- Really can only focus on the UK, this year.





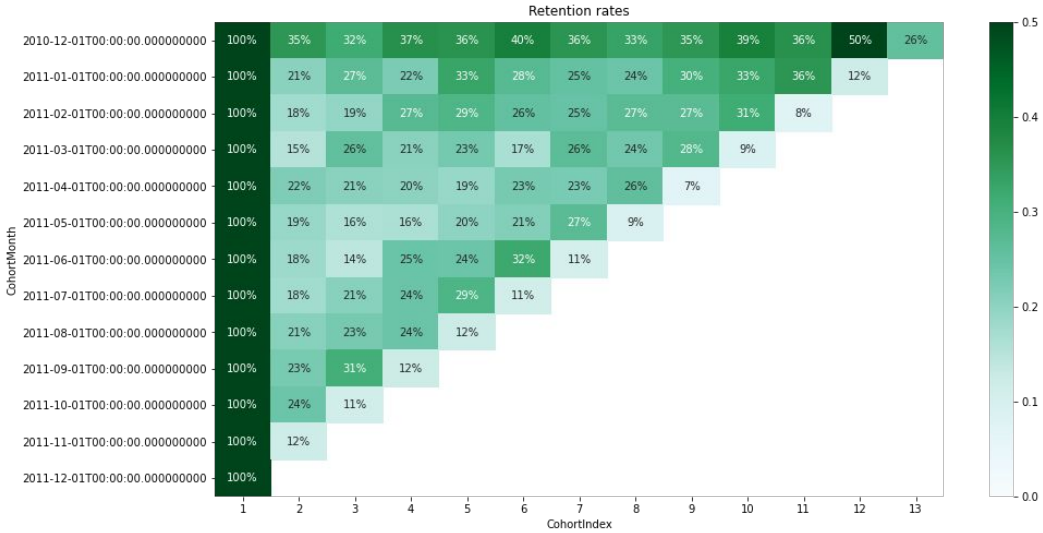


Cohort Analysis

Reveals retention rates of “cohorts.”

Can measure:

- Product lifespan
- Customer lifecycle
- Prior Marketing effectiveness



Segmenting into RFM

Customer ID is used as index

- Low Recency will mean a better score
- High Frequency and Monetary Value will mean a better score
- Levels of importance can be calculated
- RFM Score:
  - >=9 are Best
  - >=5, <9 are Average
  - <5 is Weak
- Outliers don't "overweight" the scoring

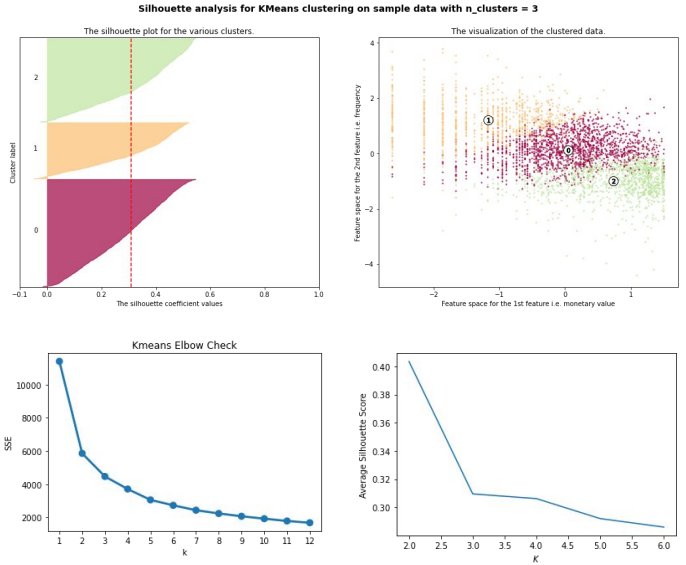
CustomerID	Recency	Frequency	MonetaryValue
12747.0	2	100	3694.06
12748.0	1	4543	28729.70
12749.0	4	198	4040.88
12820.0	3	59	942.34
12821.0	214	6	92.72

CustomerID	Recency	Frequency	MonetaryValue	R	F	M	RFM_Score	RFM_Level	RFM_Segment
12747.0	2	100	3694.06	4	4	4	12	Best Customer	444
12748.0	1	4543	28729.70	4	4	4	12	Best Customer	444
12749.0	4	198	4040.88	4	4	4	12	Best Customer	444
12820.0	3	59	942.34	4	3	3	10	Best Customer	433
12821.0	214	6	92.72	1	1	1	3	Weak Customer	111

Clustering with KMeans

Use KMeans to find the “clusters” in the data;

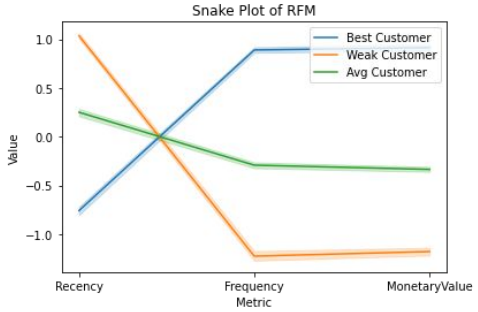
- Elbow check suggest there are 3 clusters
- Using Silhouette Score to visualize cluster overlap
- Confirming Silhouette Score Averages

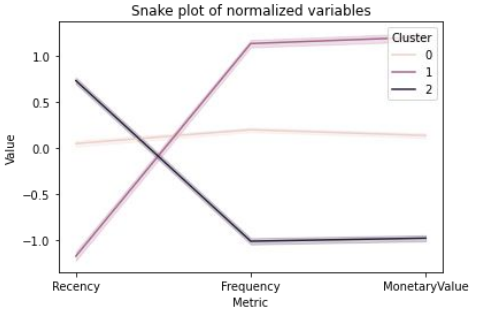


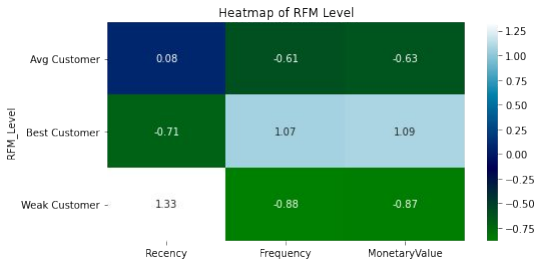
Comparison after Clustering:

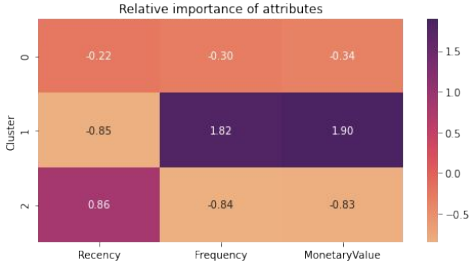
Customer Segmentation should be supported by clustering;

- Best Customers match Cluster 1
- Average Customer varies some from Cluster 0.
- Importance of Attributes for RFM and Cluster clearly align.









Summary

- Client has model for segmenting UK customers into 3 tiers.
- Future data from UK customers can be continually fed into model
  - Will automatically update clusters
  - Will reveal if new tiers are needed
  - Maintenance of model is minimal
- Future Considerations:
  - Other countries will require duplicate models
  - Outliers may average into data and change scores
  - Possibility to model and track various consumer markets exists.

Thank You!

Shane McCallum
Data Scientist & Sociologist

Contact me:
Email: McCallum.D.Shane@gmail.com
LinkedIn: https://www.linkedin.com/in/shane-mccallum/
GitHub: https://github.com/Shane-McCallum

Project Report available here.