E-commerce transactional data analysis

Clarence San

https://github.com/csanry/machinelearning/blob/main/ecomm_rfm/CODE/e-commerce-cohort-analysis-rfm-clustering.ipynb

Background and Context

This is a transnational data set which contains all the transactions occurring between Dec’10 and Dec’11 for a UK-based non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Typically, e-commerce datasets are proprietary and consequently hard to find among publicly available data. However, The UCI Machine Learning Repository has made this dataset containing actual transactions from 2010 and 2011. The dataset is maintained on their site, where it can be found by the title "Online Retail".

5	542K	3
Categorical Variables	**Recorded Transactions**	**Continuous Variables**

Why transactional data?

Business objectives

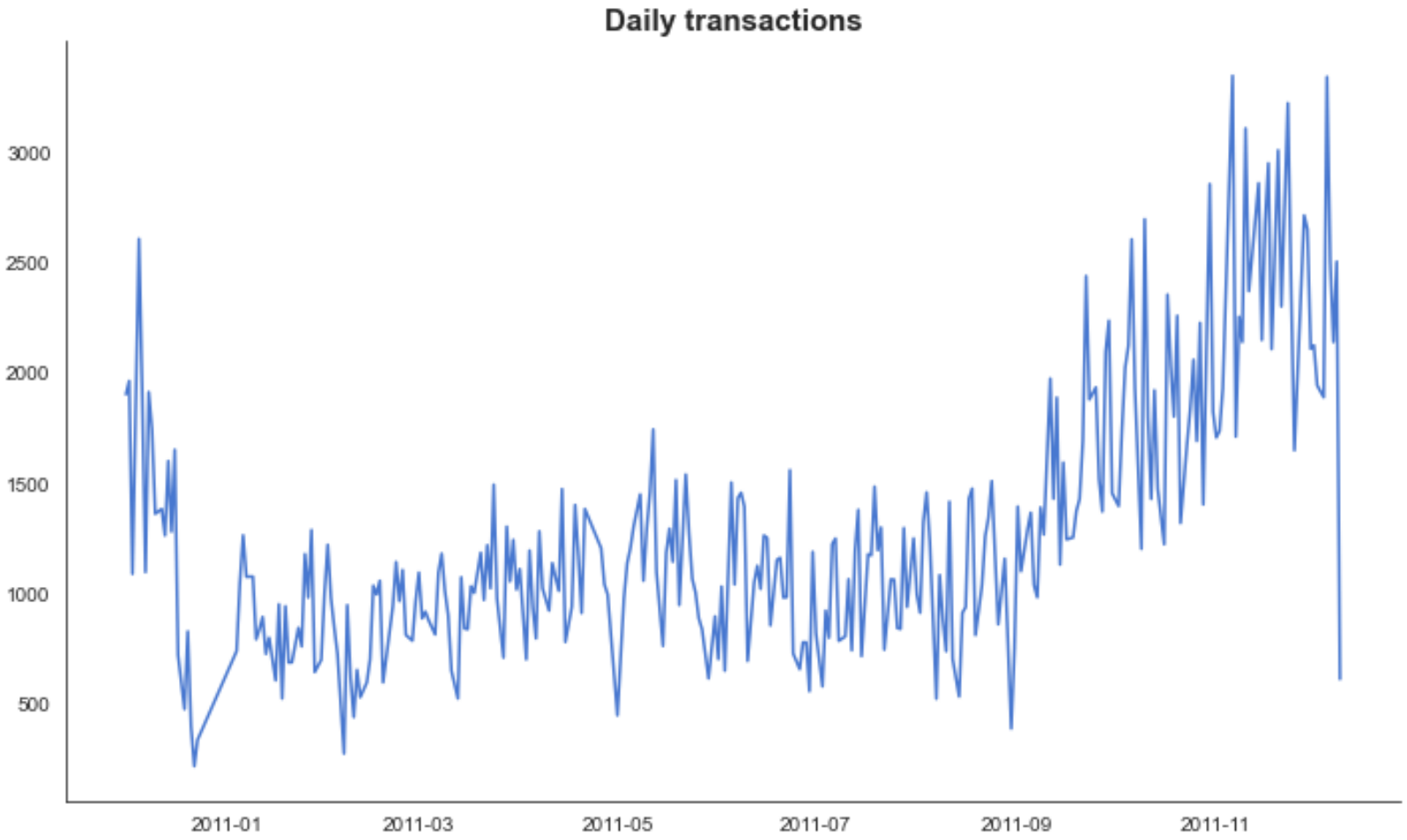
Widely applicable for various forms of analyses (Cohort, RFM, Clustering)

Central Questions

Who are our loyal customers?

What is the profile of our high-value customers?

Orders peak around the holiday season



	Quantity	Unit Price
Mean	13.1	3.1
Std	180.5	22.2
Min	1.0	0.001
Max	80995.0	8142.8

Origin of Orders are primarily from the UK

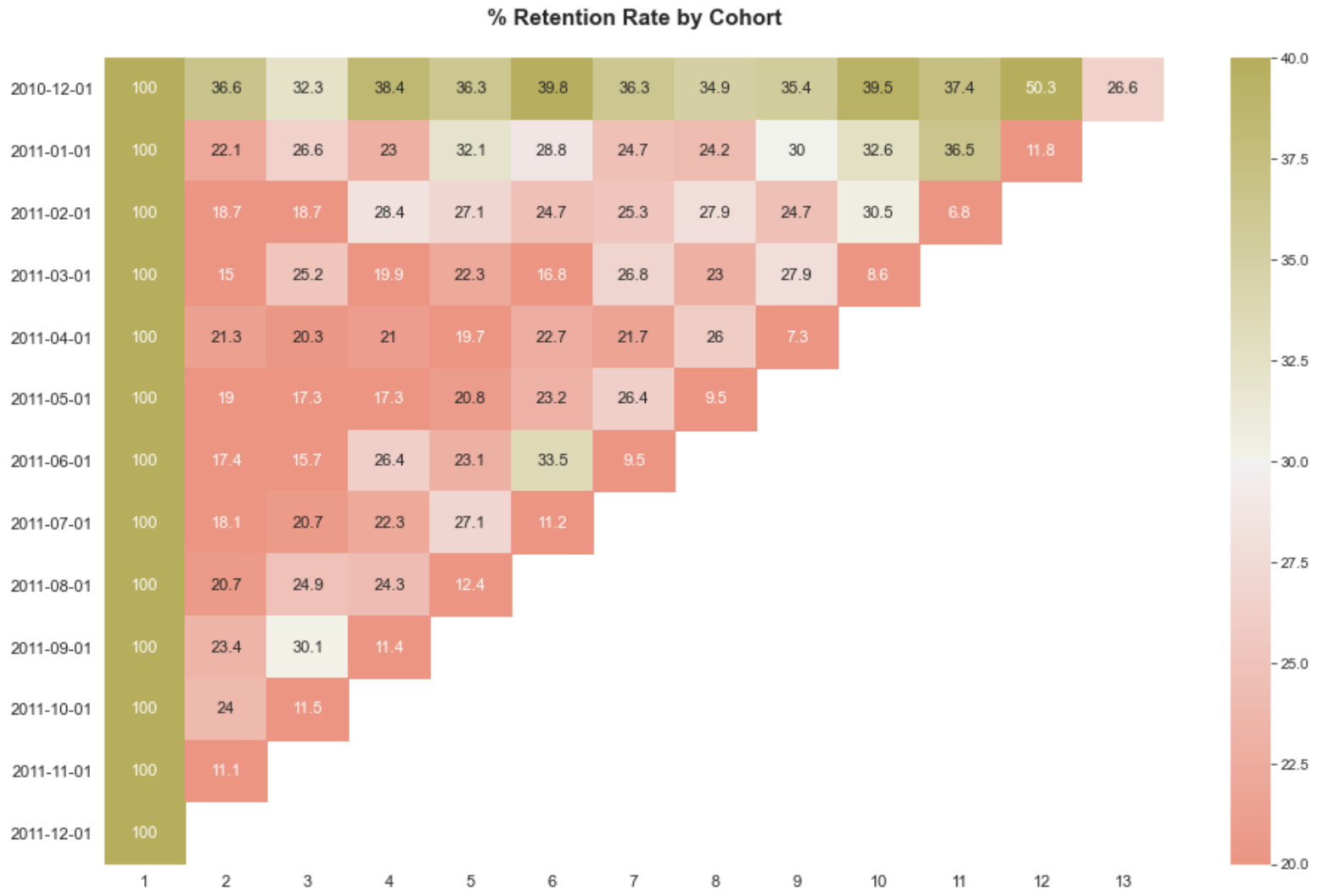
Cohort Analysis

Process

Counting the monthly active users

Calculating retention rate, average price, and quantity per cohort

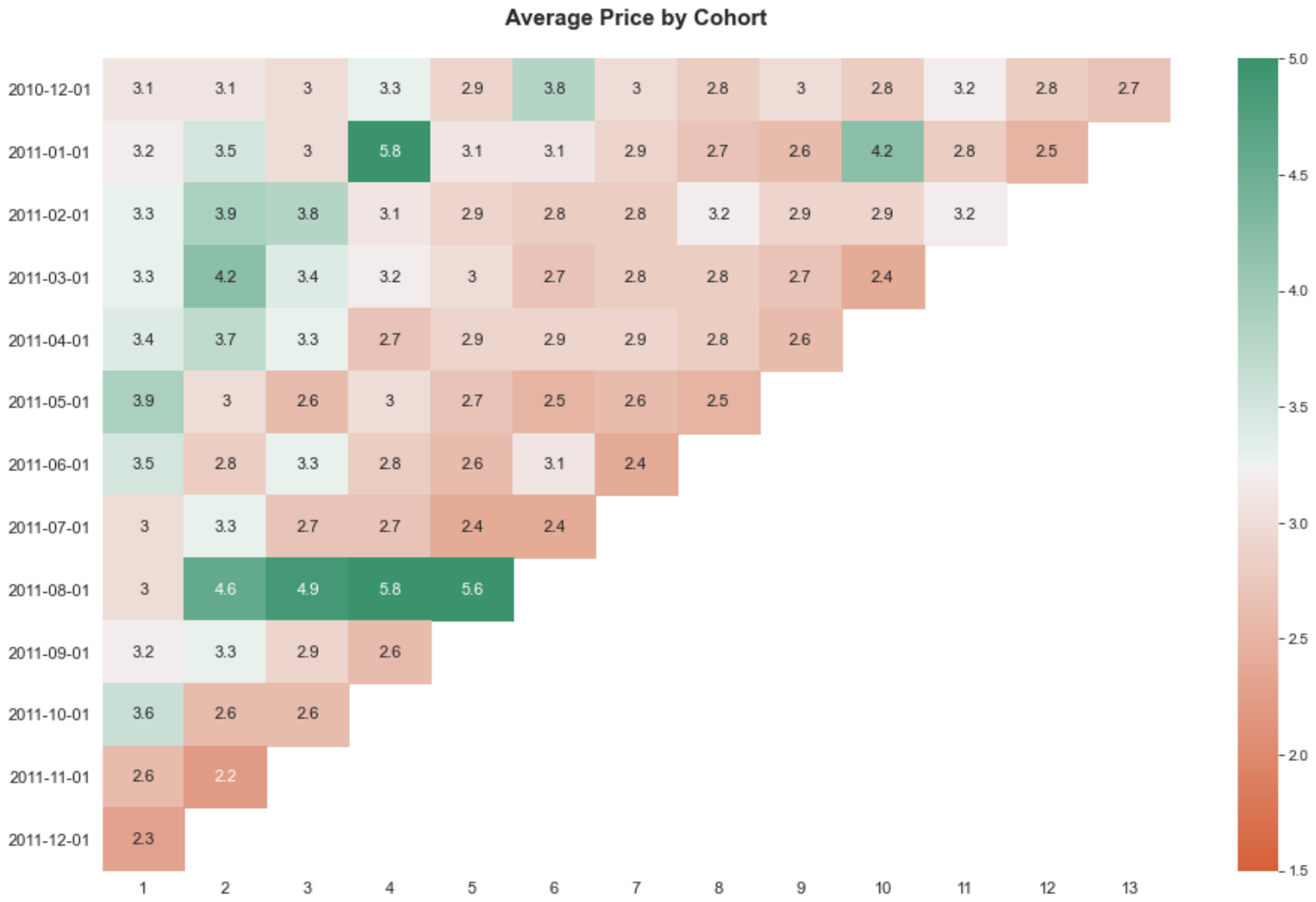
Retention Rate by Cohort



The Dec-2010 cohort shows good retention rate when compared against other cohorts, typical retention rates after a few months seems to be in the range of 18 - 24%

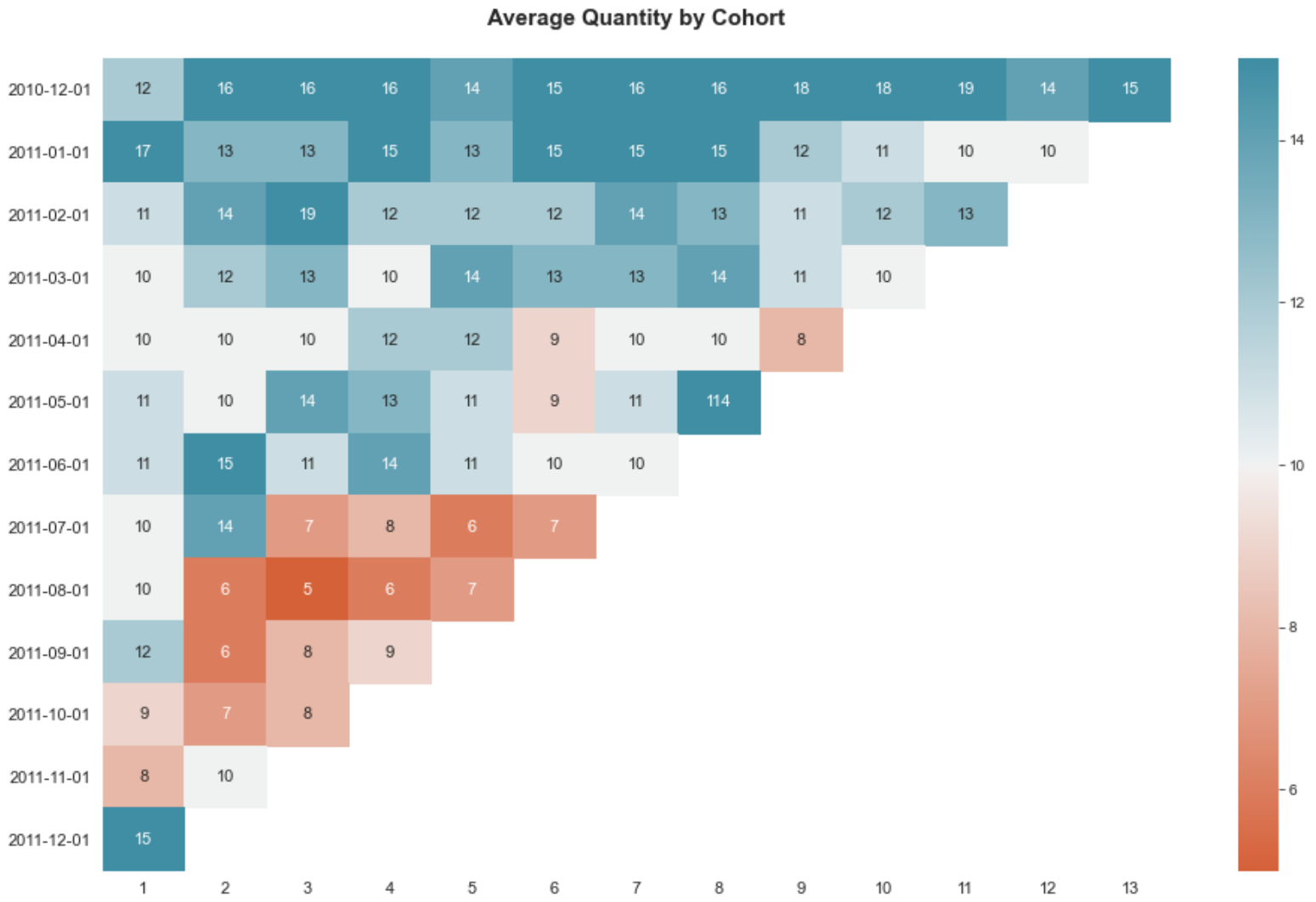
[1] Where cohort start date represents the month of customer's first transaction
[2] With the given period of a year, breaking down customers into monthly cohorts is conducive

Average Price by Cohort



The Aug-2011 cohort seemed to experience a spike in average price after the first month, while subsequent cohorts experienced lower average prices

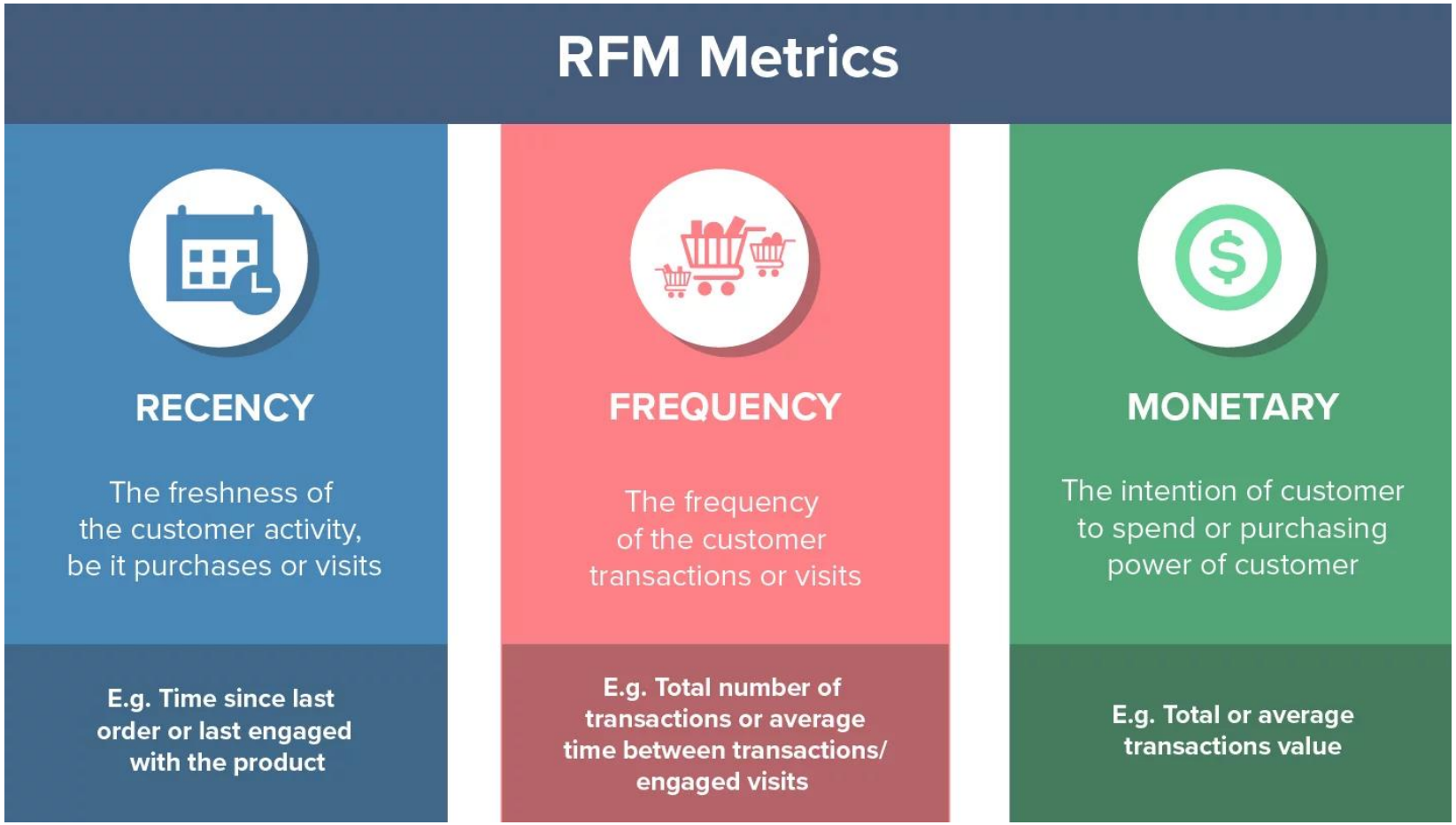
Average Quantity by Cohort



Later cohorts don't seem to perform as well in terms of quantity – it is possible that they could be buying more expensive items

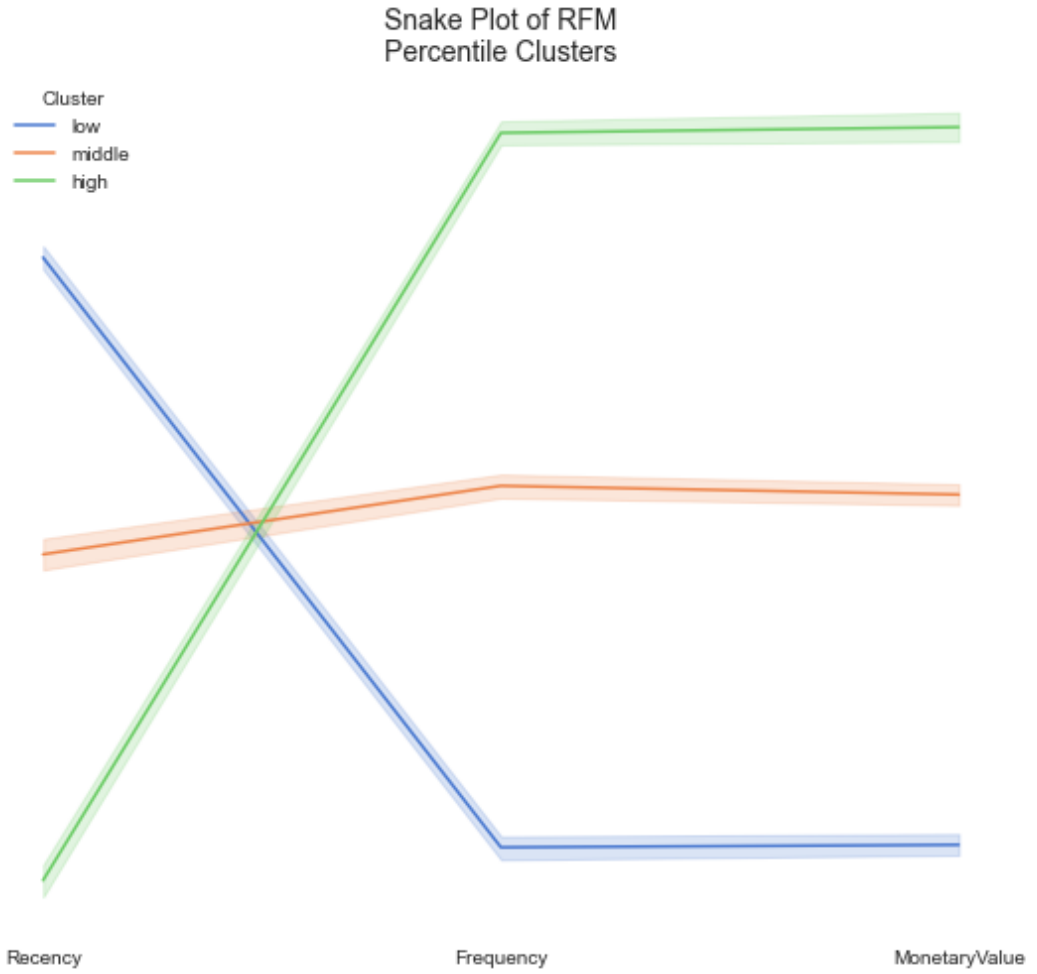
Recency – Frequency – Monetary Analysis

Sorting customers based on relationships



Snake Plot Results

Sorting customers based on relationships



RFM Cluster	Quantity Mean	Frequency Mean	Monetary Value Mean
Low	166.9	18.4	411.8
Middle	64.4	57.0	1163.1
High	20.6	225.9	5253.6

[1] Lower recency values are scored as better, while higher frequency and monetary values are scored as better

K-Means Clustering

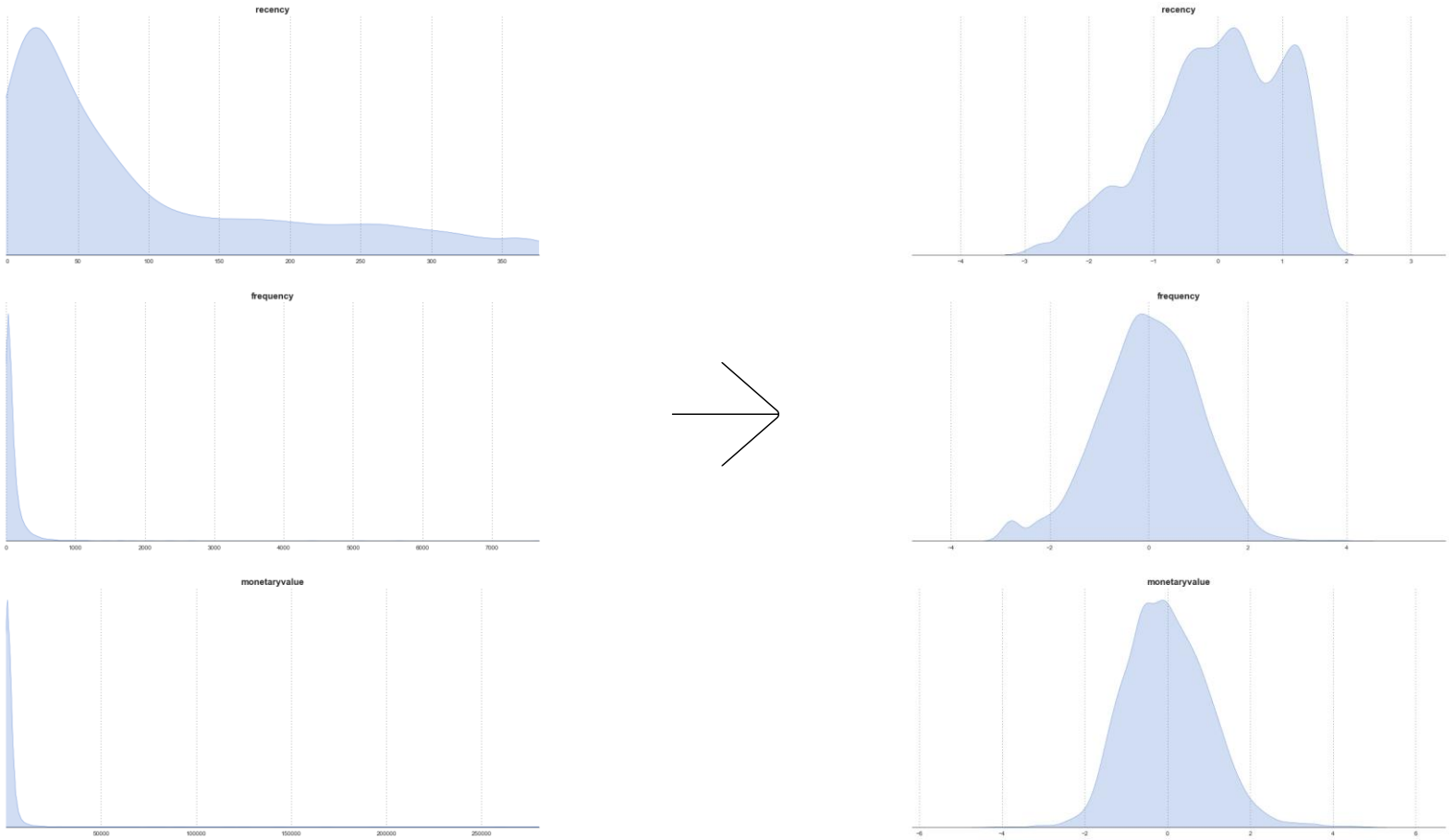
Process

Identifying possible clusters that exist within our dataset

Analyse average RFM values per cluster

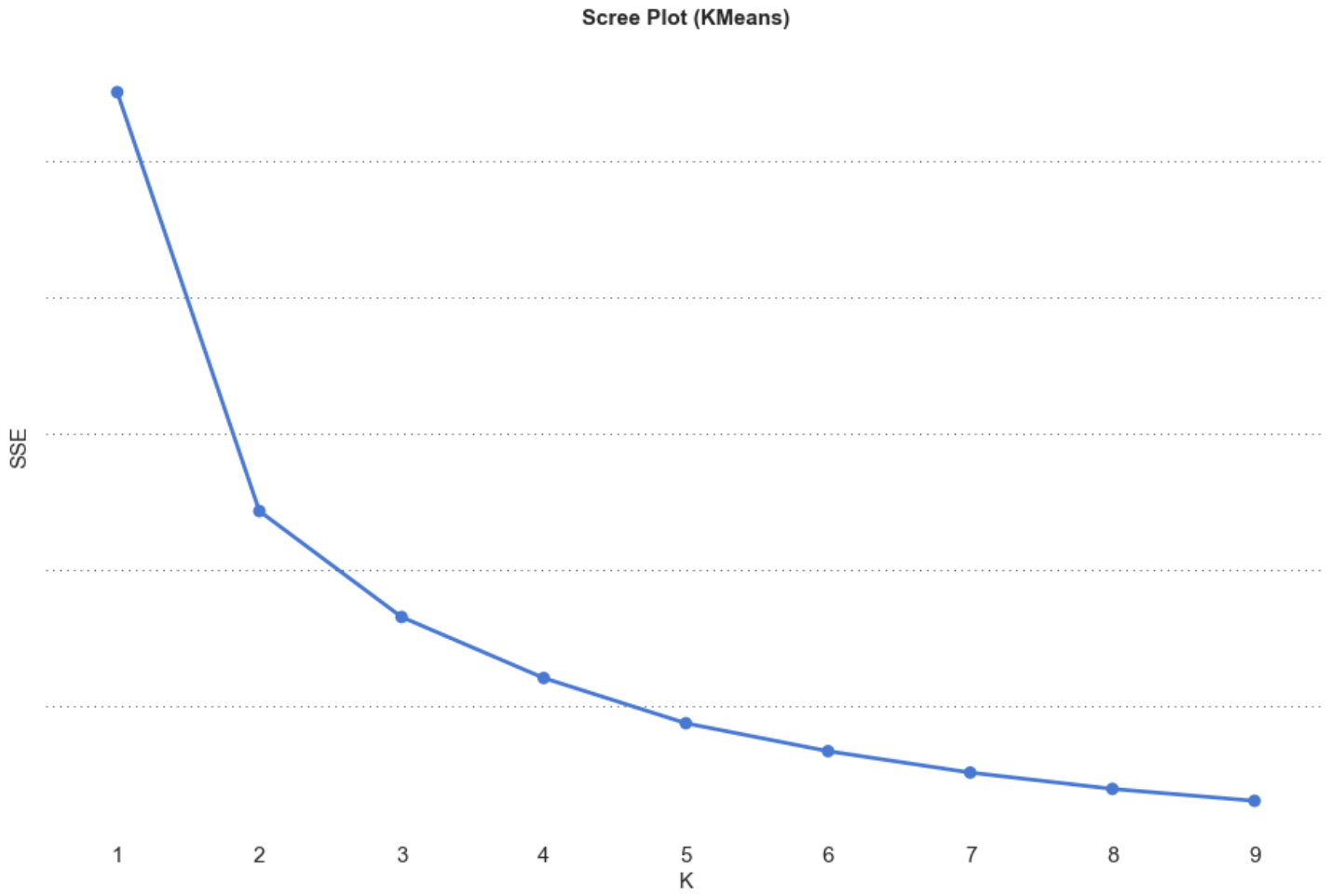
Data Preprocessing

Involves log transformations and normalization to approximate symmetric distribution



Scree Plot

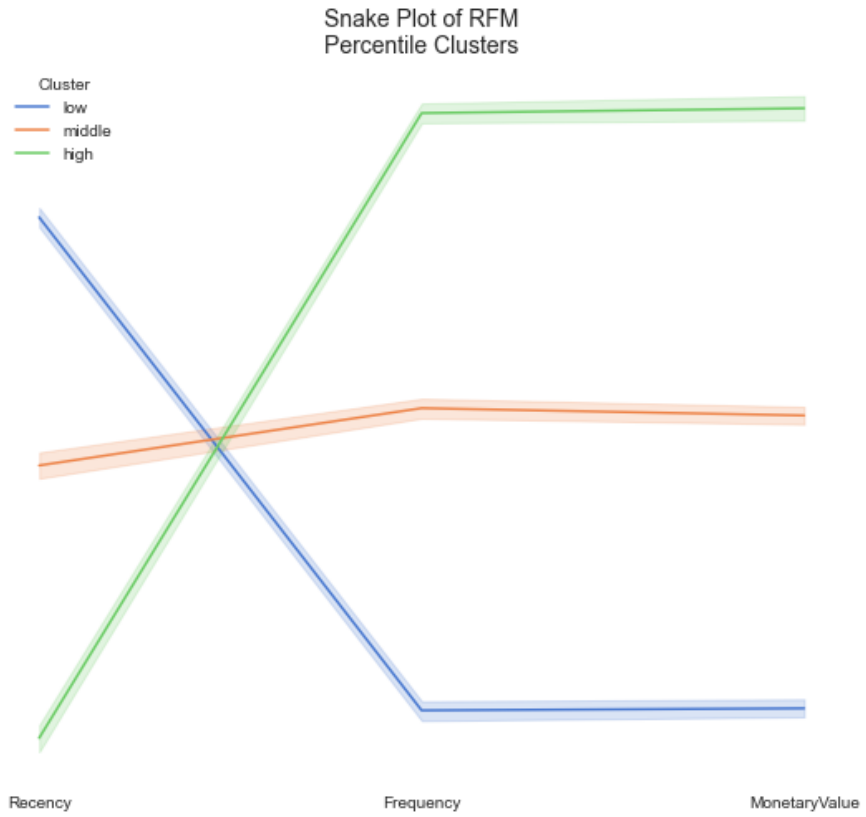
“Elbow” represents the “optimal” clusters

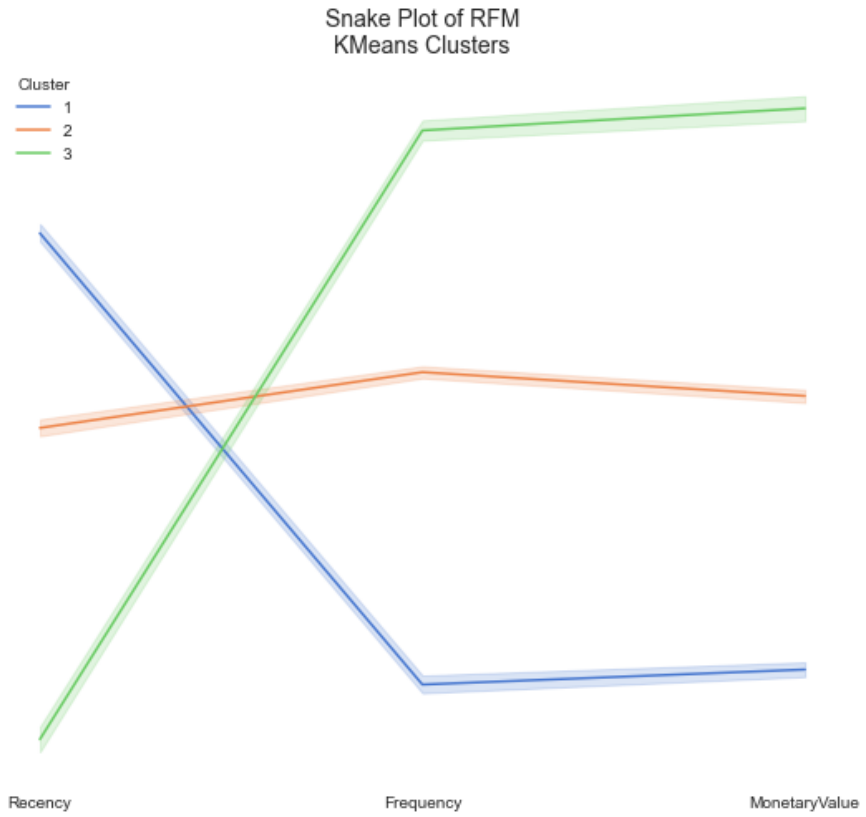


[1] Various validation techniques exist – the elbow criterion represents a more intuitive way to mathematical alternatives

Snake Plots

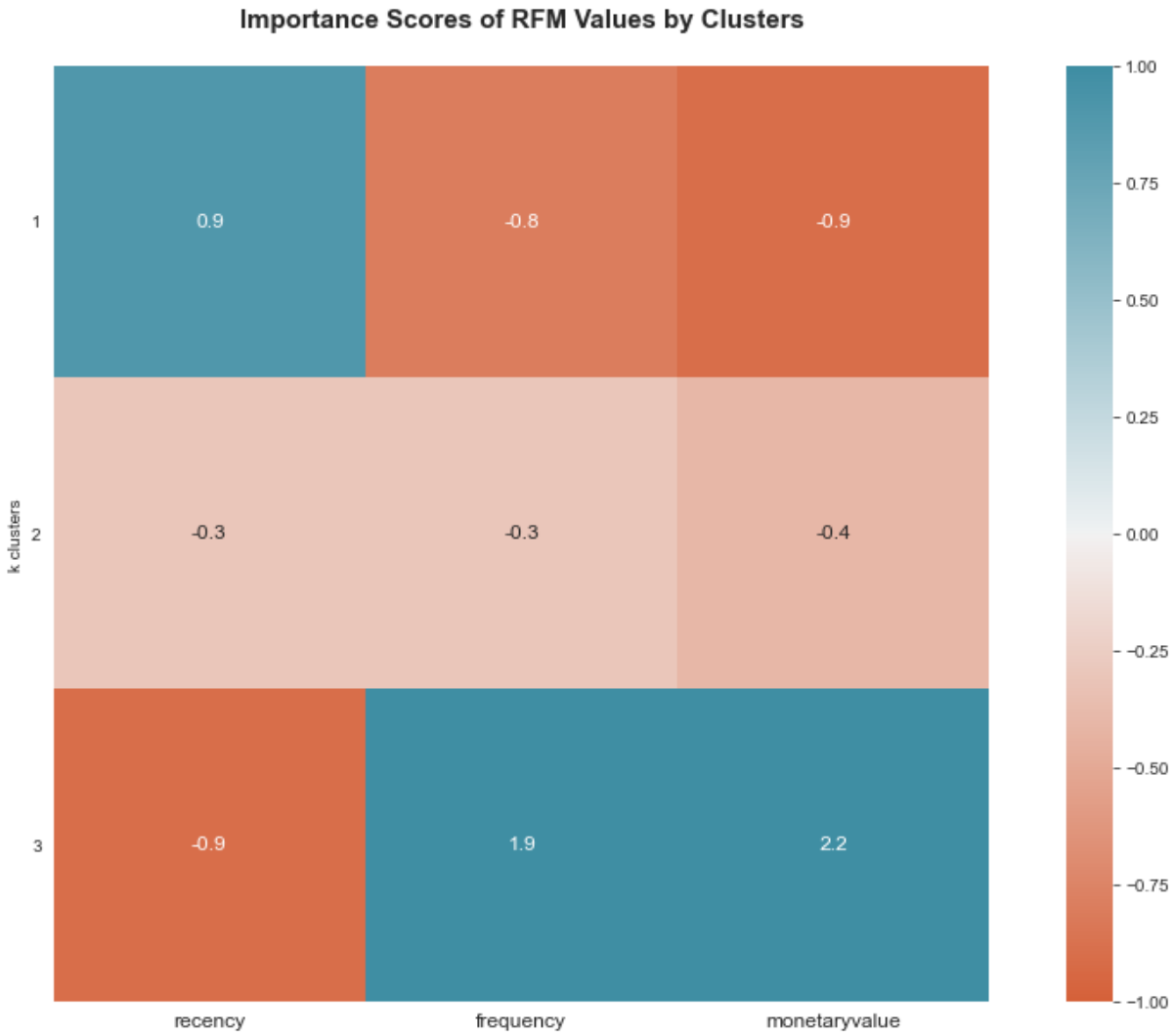
K=3 clusters are very similar to the results of RFM analysis





[1] High Value Customers can be identified as cluster 3 from K-Means

Importance scores



High frequency and monetary value transactions are what are more important to defining high value customers

[1] Importance scores are useful to determine the relative importance of each segment's attribute

Conclusion

Coverage

Cohort Analysis

RFM Metrics

RFM Levels

K-Means Clustering

Next Steps

Alternative groupings of k

Integrating existing dataset with demographic data

Market Basket Analysis of Customers

https://github.com/csanry/machinelearning/blob/main/ecomm_rfm/CODE/e-commerce-cohort-analysis-rfm-clustering.ipynb