PRML Minor Project

Team Members: Arman Garg (B21CS013)

Samyak Haribhakta (B21CS067)

Project 4

Q: A company that sells some of the product, and you want to know how well the selling performance of the product. You have the data that we can analyze, but what kind of analysis can we do? Well, we can segment customers based on their buying behavior on the market. Your task is to classify the data into the possible types of customers which the retailer can encounter.

The data used is: ‘https://archive.ics.uci.edu/ml/datasets/online+retail’

Soln:

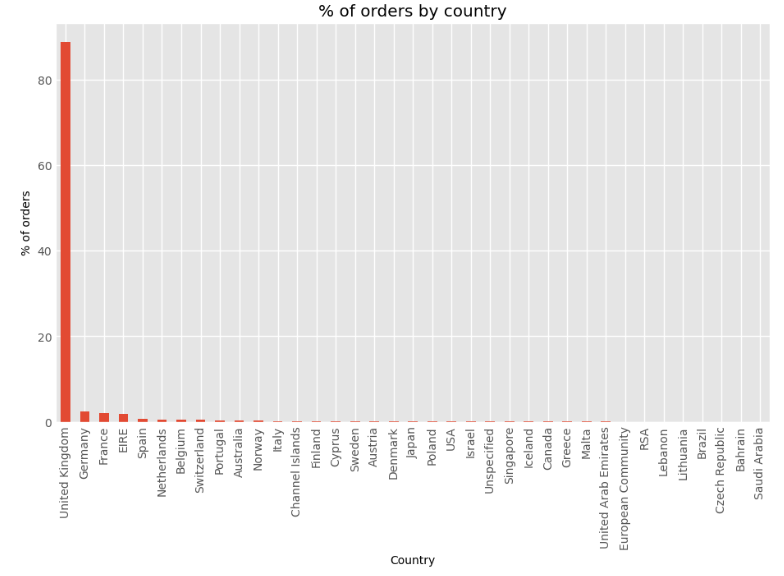
Data analysis and preprocessing:

For this project we started by analyzing the data. The data was huge with more than 5 lakh entries divided over different countries.

Firstly we Preprocessed the data, imported the data as csv, then removed the missing, duplicate and na values. Then we dropped the Description column as it was not in use, other columns like quantity and unit price already described the type of order so there was no need of order description.

Then for analyzing more we find out the no. of unique customers which came out as 4372, no. of unique transactions which came out to be 22190 and unique products 3684.

Then we thought and divided the data into different countries and made a bar graph for it. From that graph we found out that United Kingdom has maximum no. of orders with more than 95% of orders. So then we thought of analyzing it for UK first.



After setting the data for only UK we first found out the total no. of cancelled orders which came out as 7501 in number.

After removing all the false, duplicate, NA, and cancelled orders the data left was about 3.5 lakh which was earlier close to 6 lakh.

Now the Unique customer = 3921,

Unique transactions = 16649,

Unique products = 3645

After preprocessing we first created columns for invoice month. The main goal for this was to find out how old each customer is so the first month of their order is noted.

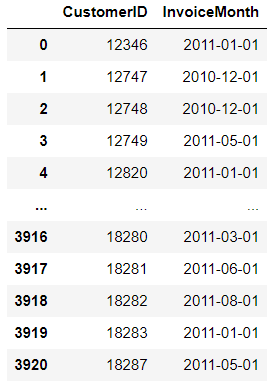
For analyzing we thought to use Cohort analysis.

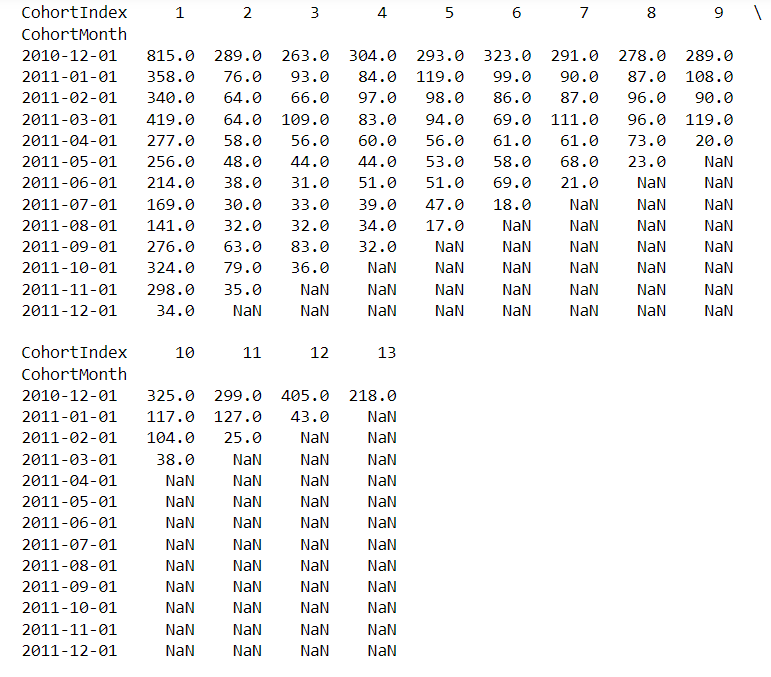
Cohort analysis involves grouping of data based on their common characteristics.

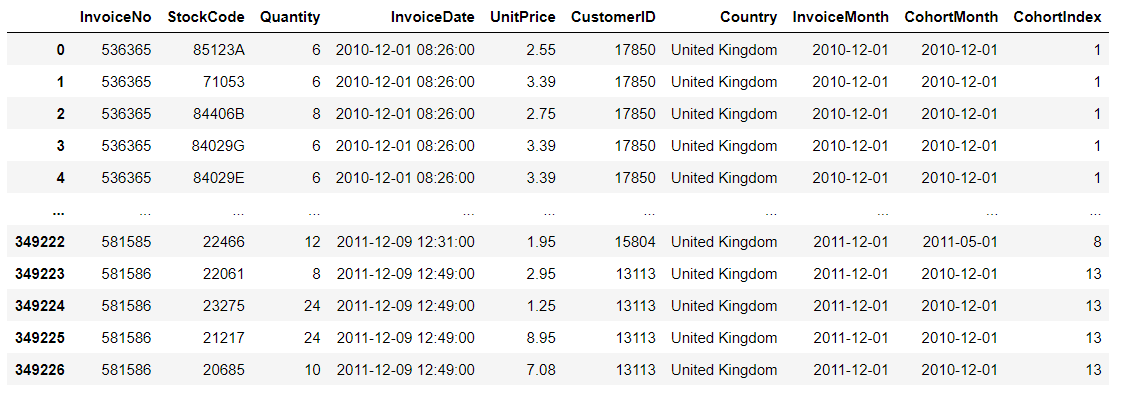
Cohort analysis based on the month of order involves grouping customers who made their first purchase in the same month and analyzing their behavior over time. This type of analysis can help to understand how customers who joined the business at the same time are behaving, and how they differ from other cohorts.

For cohort analysis, there were few labels that we created:

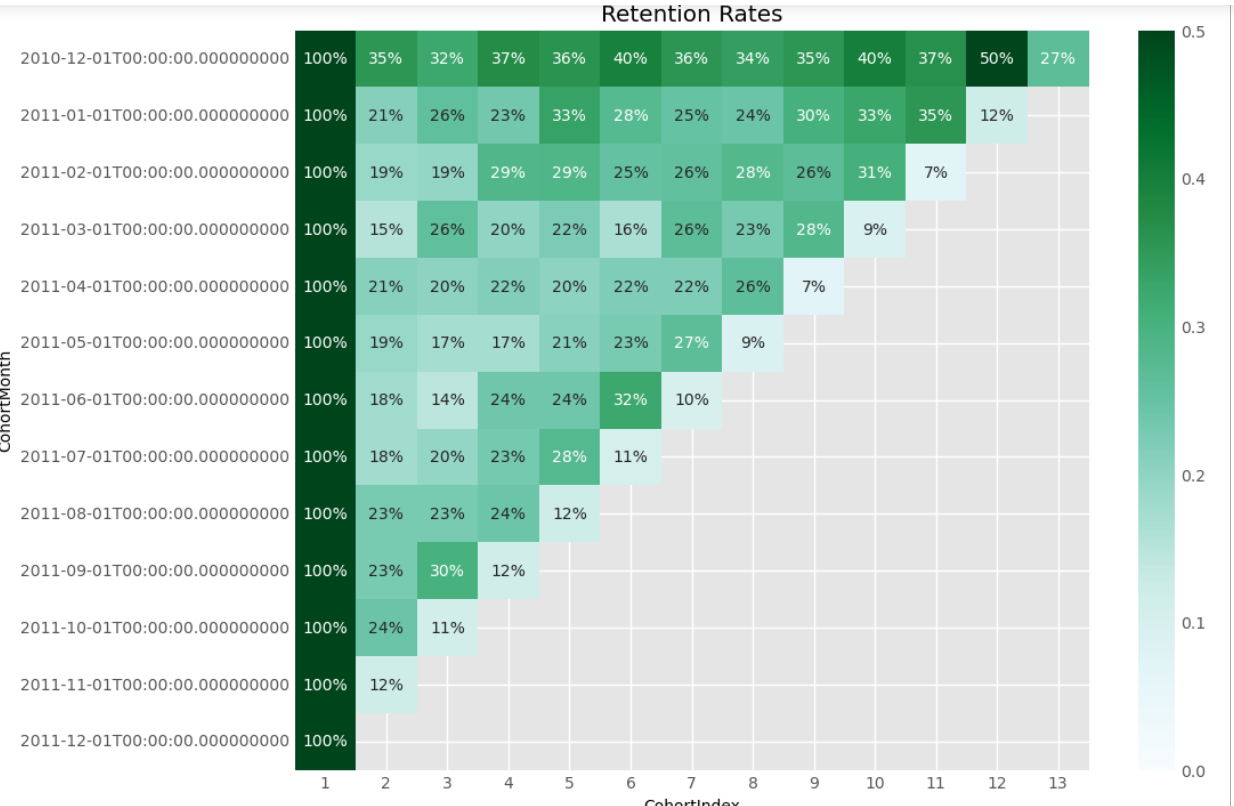
1. Invoice period: A string representation of the year and month of a single transaction/invoice.
2. Cohort group: A string representation of the year and month of the customer’s first purchase. This label is common across all invoices for a particular customer.
3. Cohort period / Cohort Index: A integer representation a customer’s stage in its “lifetime”. The number represents the number of months passed since the first purchase.



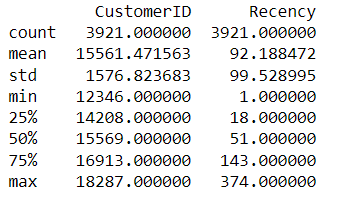




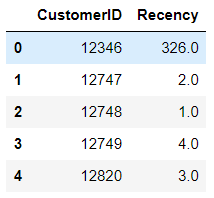
Then we created a table for retention rates for the customers.



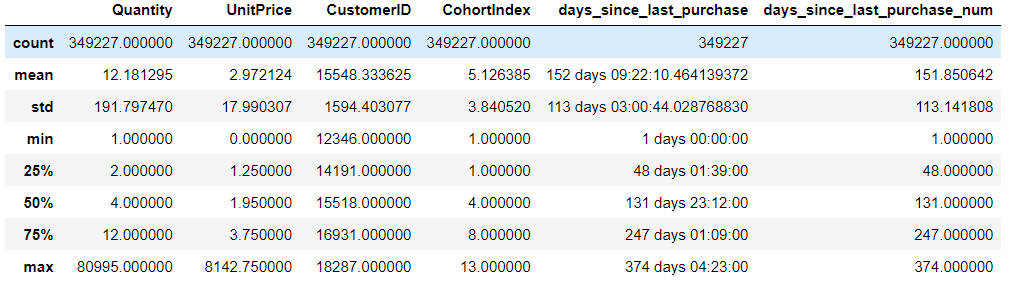
Then we find out customer I’d with recency:



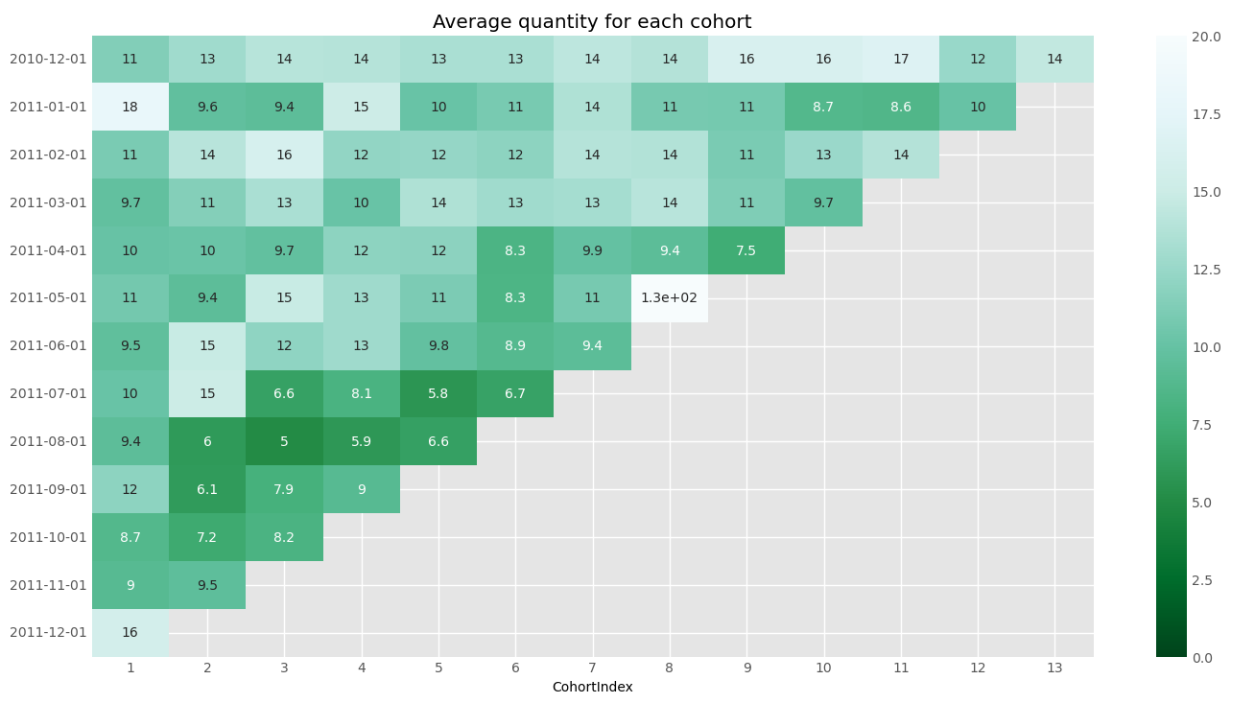
An overview how it will look:



Here recency defines no. of days from last purchase.



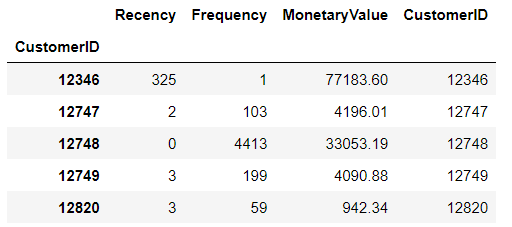
Then we created heatmap for average quantity for each cohort:



From this analysis we can find how many customers have recent order, which one are the oldest customer, which used to order early and now don’t order and which are the new ones to order.

After this we calculated total sum per order by multiplying order quantity and unit price. We renamed this total sum to MonetaryValue.

An overview how the table looks after adding monetary value with recency, customer id and frequency.



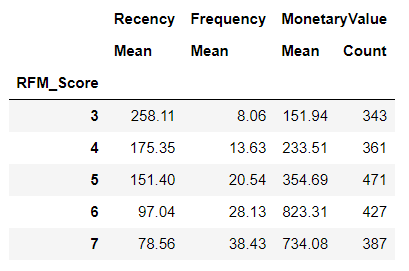
From we thought of analyzing the data with RFM.

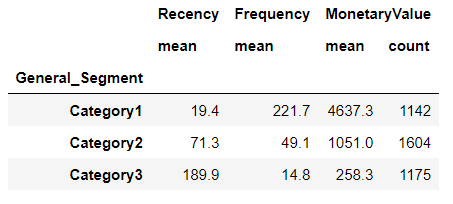
**RFM** is an acronym of recency, frequency and monetary. Recency is about when was the last order of a customer. It means the number of days since a customer made the last purchase. If it’s a case for a website or an app, this could be interpreted as the last visit day or the last login time.

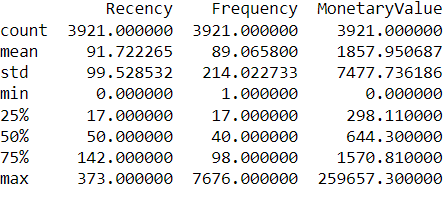
**Frequency** is about the number of purchases in a given period. It could be 3 months, 6 months or 1 year. So we can understand this value as for how often or how many a customer used the product of a company. The bigger the value is, the more engaged the customers are. Could we say them as out VIP? Not necessary. Cause we also have to think about how much they actually paid for each purchase, which means monetary value.

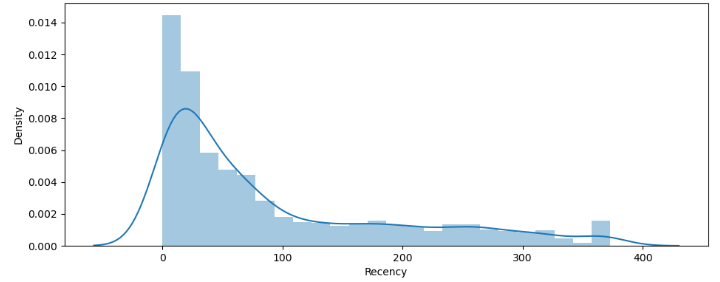
**Monetary** is the total amount of money a customer spent in that given period. Therefore big spenders will be differentiated with other customers such as MVP or VIP.

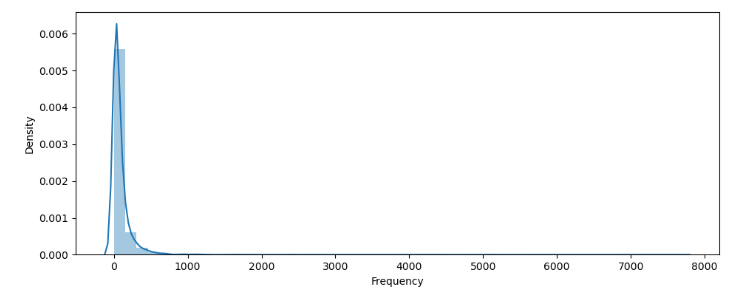
There are 3 RFM metrics, Recency, Frequency, Monetary.

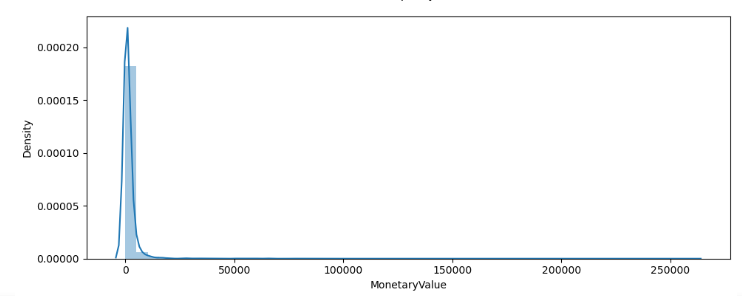


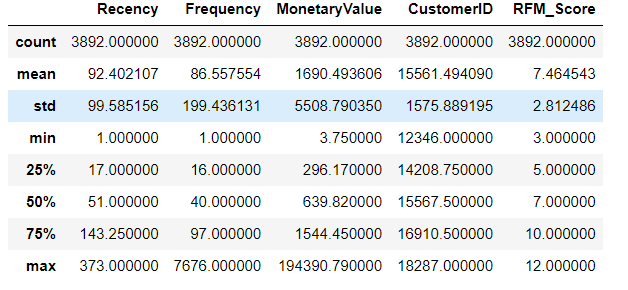




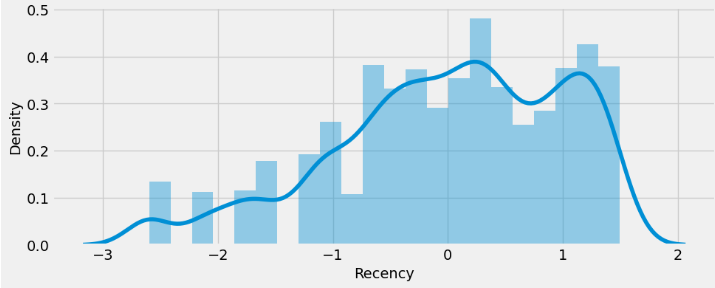


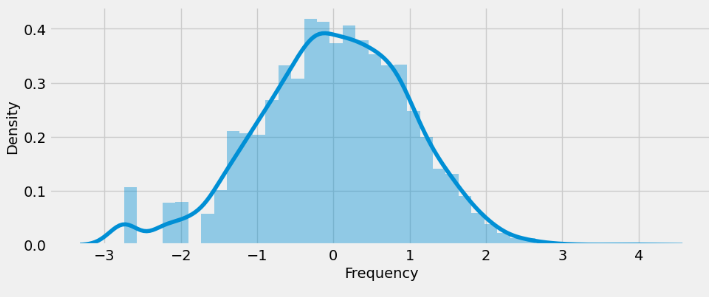


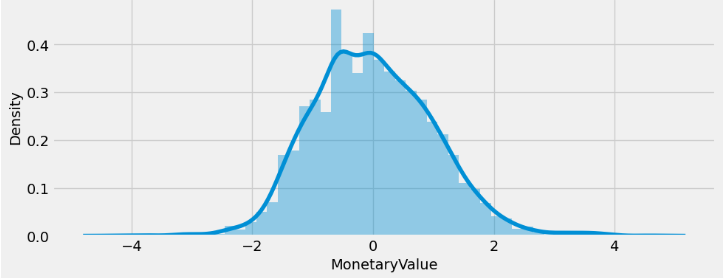




From this table, we find that mean and variance are not equal so for better visualization we standard scaled the data and then created the graphs again.



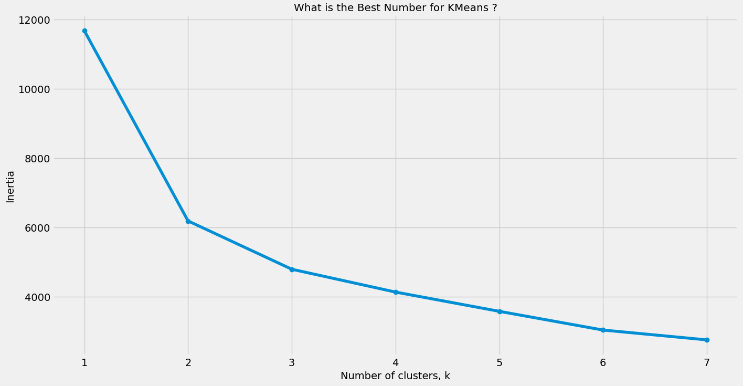




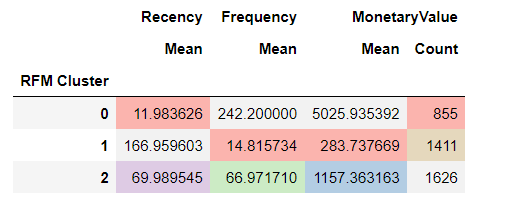
Now we thought of using KMeans clustering for more deeper analysis of the data and solving the problem. For this we created KMeans from scratch.

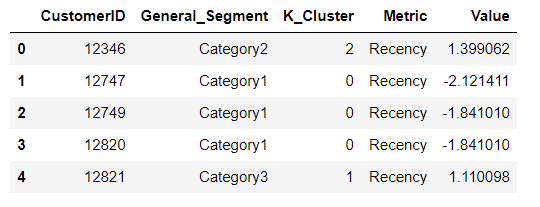
Key steps:

1. Data pre-processing
2. Choosing a number of clusters
3. Running k-means clustering on pre-processed data
4. Analyzing average RFM values of each cluster



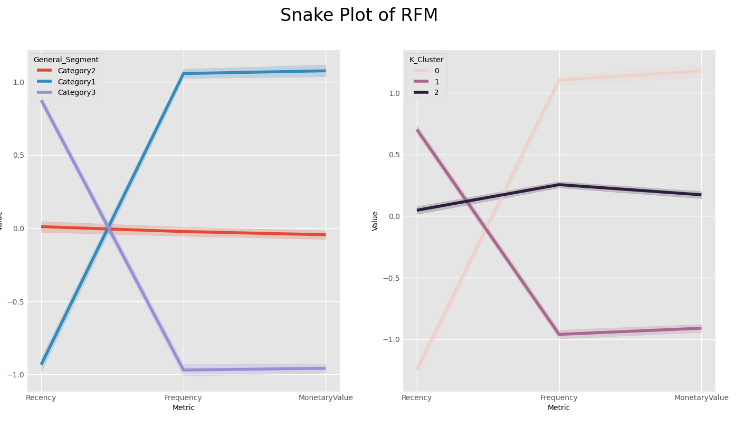
From this graph we used 3 as the number of cluster using the elbow method as it will give the best results.



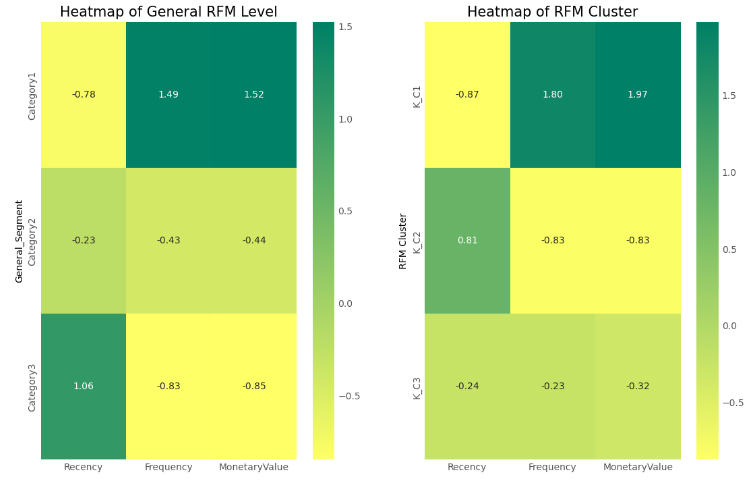


Then we created snake plots to understand and compare segments

* Market research technique to compare different segments
* Visual representation of each segment’s attributes
* Need to first normalize data (center & scale)
* Plot each cluster’s average normslized value of each attribute



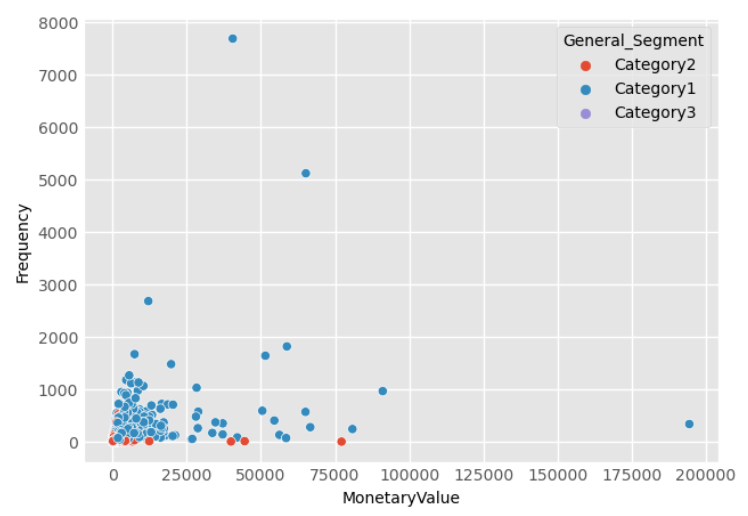
For better understanding we created heat maps. Heat maps are a graphical representation of data where larger valures are colored in darker scales and smaller values in lighter. We can compare the variance between groups quite intuitiverly by colors.



From these heat plots we can infer that:

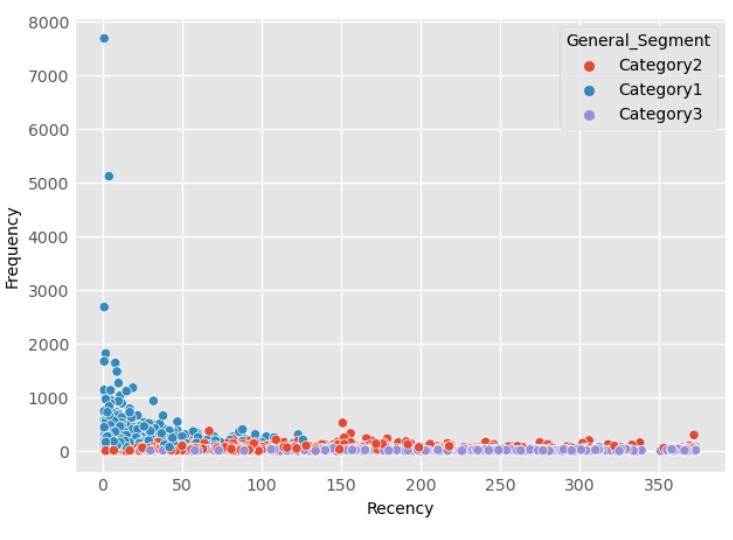
In general RFM level Category 3 has max value for Recency, while in both Requency and Monetary Category 1 has max value.

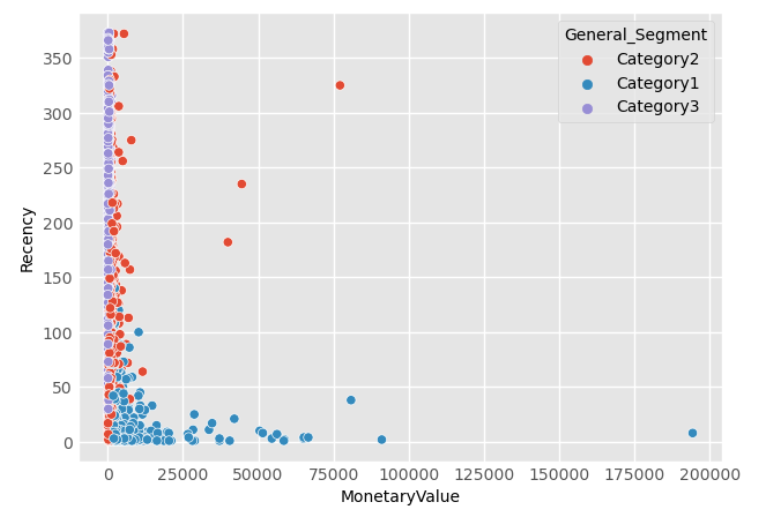
In RFM cluster map, max recency value is in Category 2 while same other 2 values are max in Category 2.



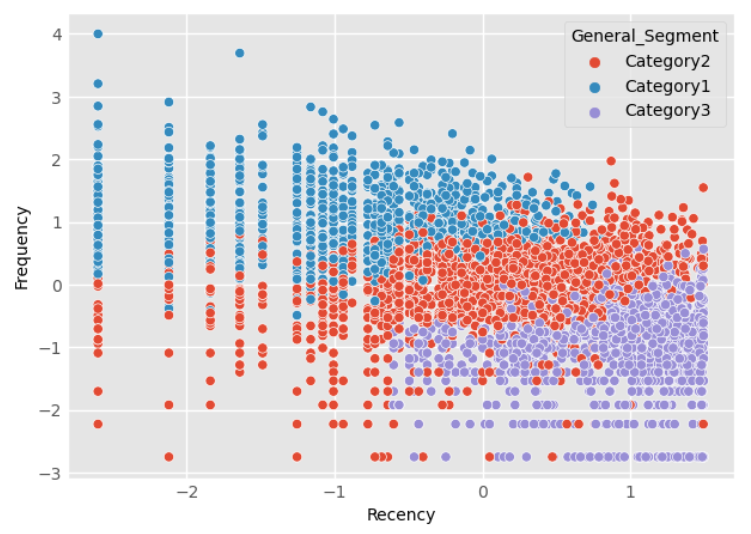
This is the unscaled general segmentation graph between requency and monetarvalue. Here the data is not scaled so it is clusterized at 1 point except that there are only ouliers. So it was not quite good for KMeans clustering.

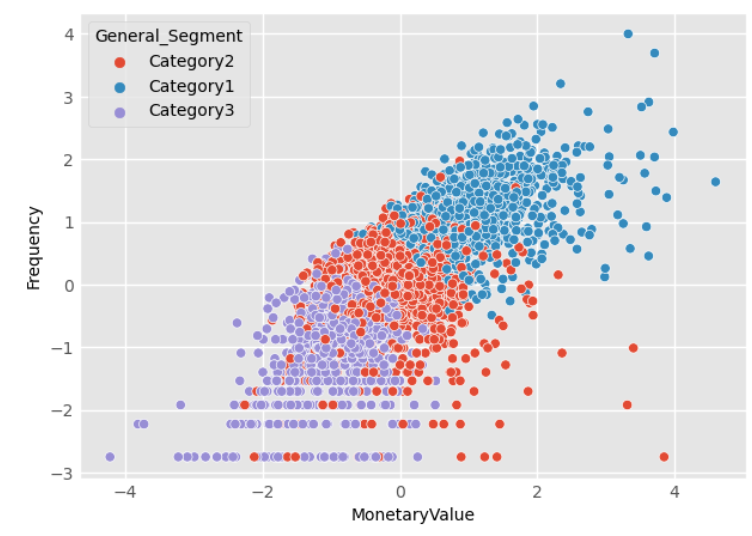
Similarly for other graphs





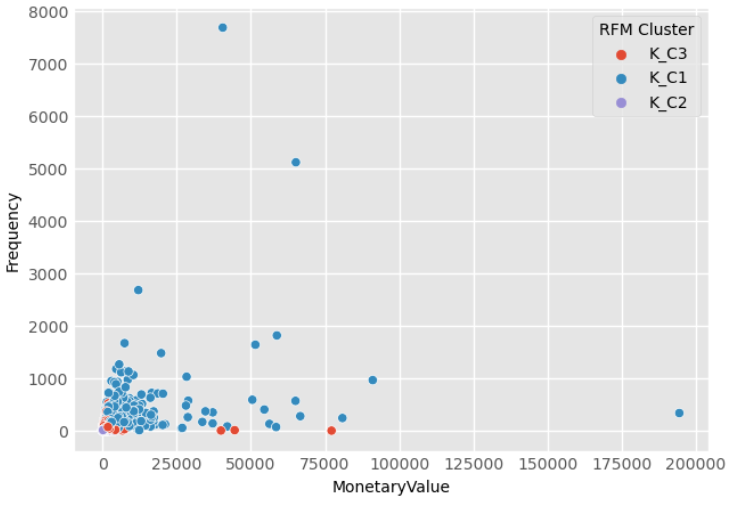
We scaled the data to apply KMeans clustering and for better visulisation.

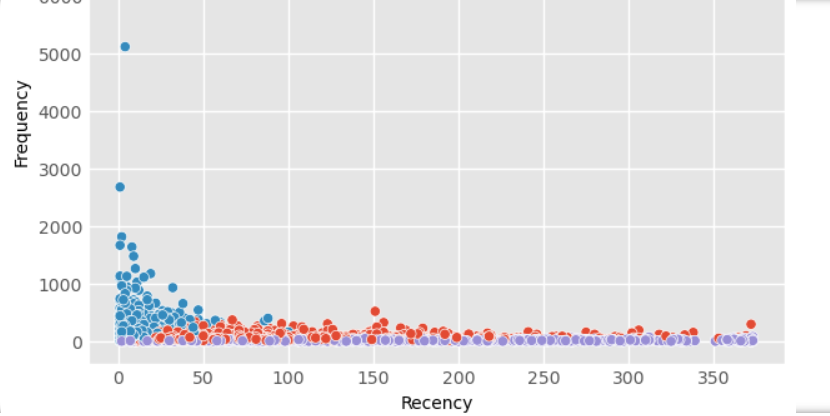


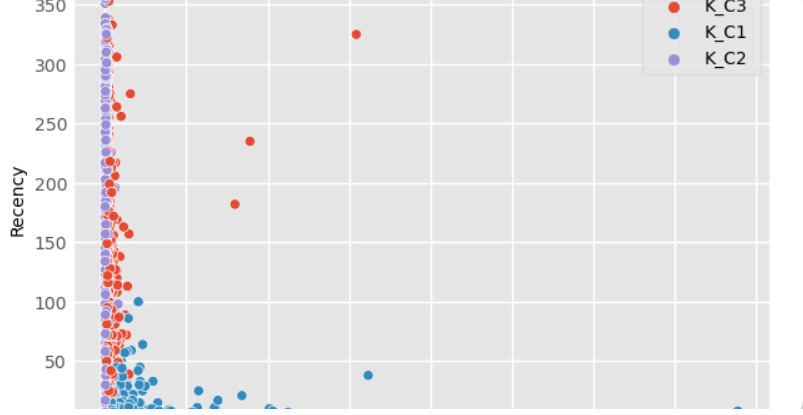




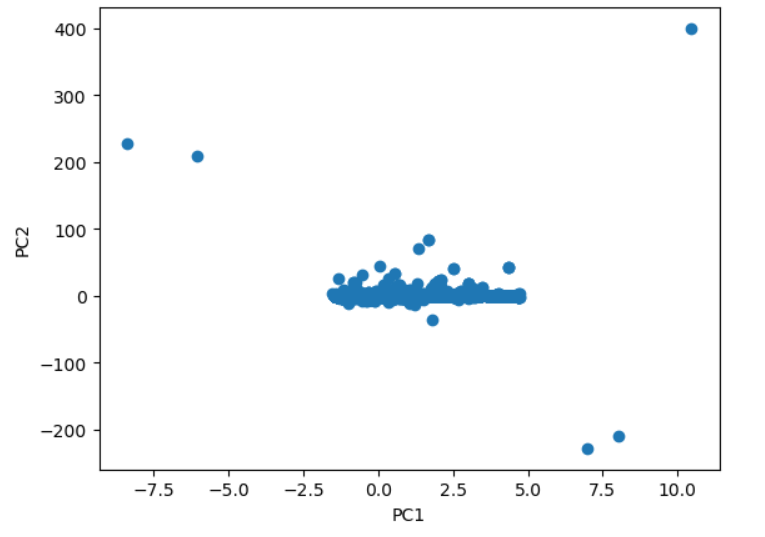
#Unscaled-KMeans







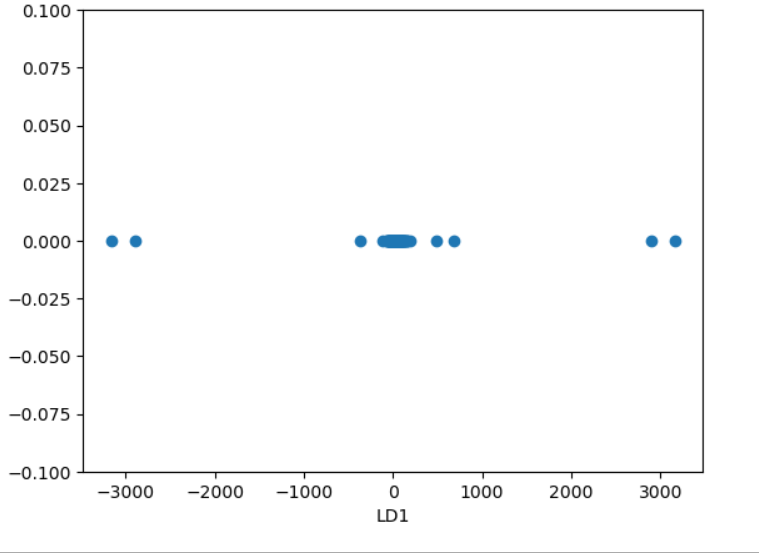
Now we apply PCA on it:



This it the graph we get using PCA.

We applied PCA and plotted the first two PCs on the scatter plot. The scatter plot helps us visualize the distribution of the data points in the reduced feature space, where each point represents a unique combination of the original features.

Now we apply LDA on it for further analysis



This is the graph we get for LDA

By applying LDA we obtain a set of liner combinations of the original features that can best separate the data into distinct groups based on the target variable.

We reduced the dimensionality of the dataset to one dimension using LDA, and obtained a single discriminant that maximally separates the transactions based on the InvoiceNo. The scatter plot help us visulize the relationship between transactions and the InvoiceNo variable. We may observe clusters of transctions that belong to the same InvoiceNo, which can indicate that these transactions are part of the same order or transaction.