Visual Analytics Workflow Design

Customer behavioural segmentation using data mining techniques

Problem definition

The main goal is to divide customers into groups which share certain characteristics. The specific goal is to perform the analysis of a dataset, group the transactions and make conclusions based on them.

Customers segmentation depends on the business objective but the common goal is to identify high and low value customers for marketing purposes. One of the technique to group customers for marketing purpose is RFM method which use information about last purchase, frequency of purchases and monetary value of all purchases.

Method description

To solve the problem I use clustering technique. I've decided to use k-Means clustering because is efficient in segmenting large datasets, quickly iterates to good results and it is easy to interpret.

Dataset description

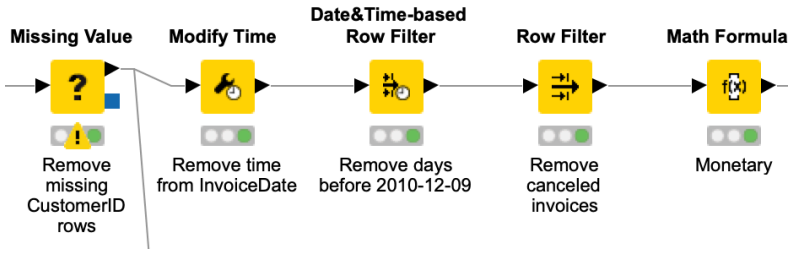
The Online Retail II data set contains all the transactions occurring for a UK-based and registered, non-store online retail. The company mainly sells unique all-occasion gift-ware.

The dataset contains only purchase behaviour information including: invoice number, product code, product name, invoice date and time, quantities of each product per transaction, unit price, country and customer number.

Row ID	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
Row0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01T08:...	2.55	17850	United Kingdom
Row1	536365	71053	WHITE METAL LANTERN	6	2010-12-01T08:...	3.39	17850	United Kingdom
Row2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01T08:...	2.75	17850	United Kingdom
Row3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01T08:...	3.39	17850	United Kingdom
Row4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01T08:...	3.39	17850	United Kingdom
Row5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01T08:...	7.65	17850	United Kingdom

The file containing data is an Excel file with transactions between 01/12/2009 and 09/12/2011 divided into two sheets. For the purpose of my analysis I picked only most recent transactions between 1/12/2010 and 9/12/2011 from the second sheet.

Data pre-processing



Firstly using statistic node I identify missing values in CustomerID and Description columns. There is around 25% missing values in CustomerID column. I have two options, I can treat all of them as one retail customer but it might disrupt the grouping process and give noise to our dataset or I can remove all rows with missing data. I've decided to remove all rows with missing CustomerID as it would affect my analysis. Description data is not going to be used in further processing, so I've left rows with missing values.

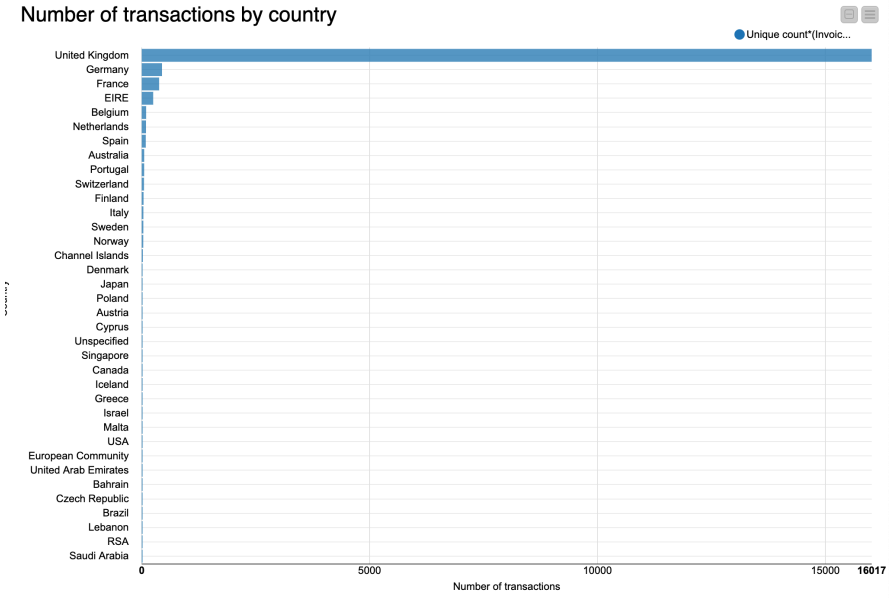
I remove time from InvoiceDate because I need day level data. In RFM method usually we consider only 12 months data because older transactions cannot say anything about current situation of customer, seller or about the product. So I remove older transaction than 9/12/2010.

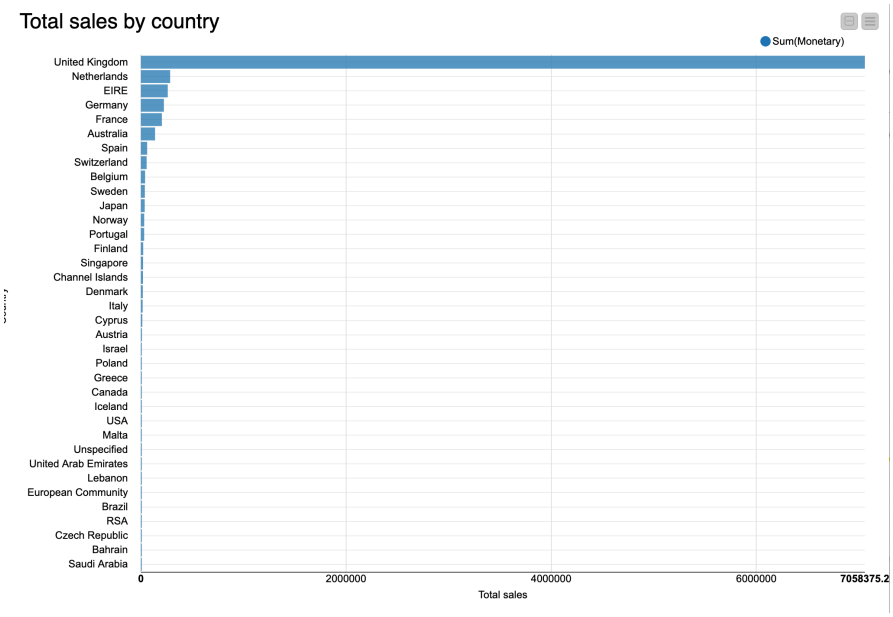
I know from dataset description that InvoiceCode which begins with "C" means that this transaction was cancelled. I decided to remove all cancelled invoices due to our RFM model contain Monetary column and it is hard to speak about value using cancelled invoices.

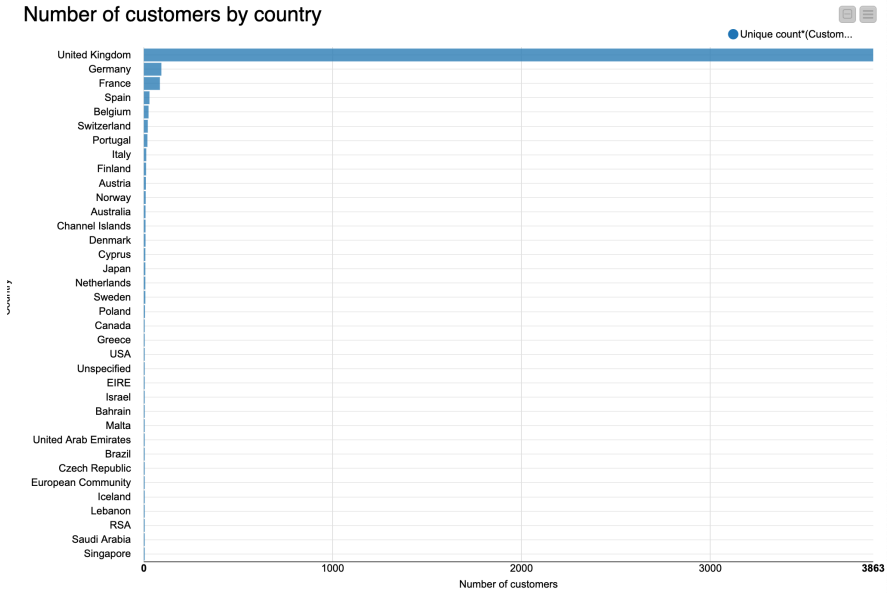
I create Monetary column by multiplying quantity and unit price which is the amount that the customer spent during the year.

Exploratory data analysis

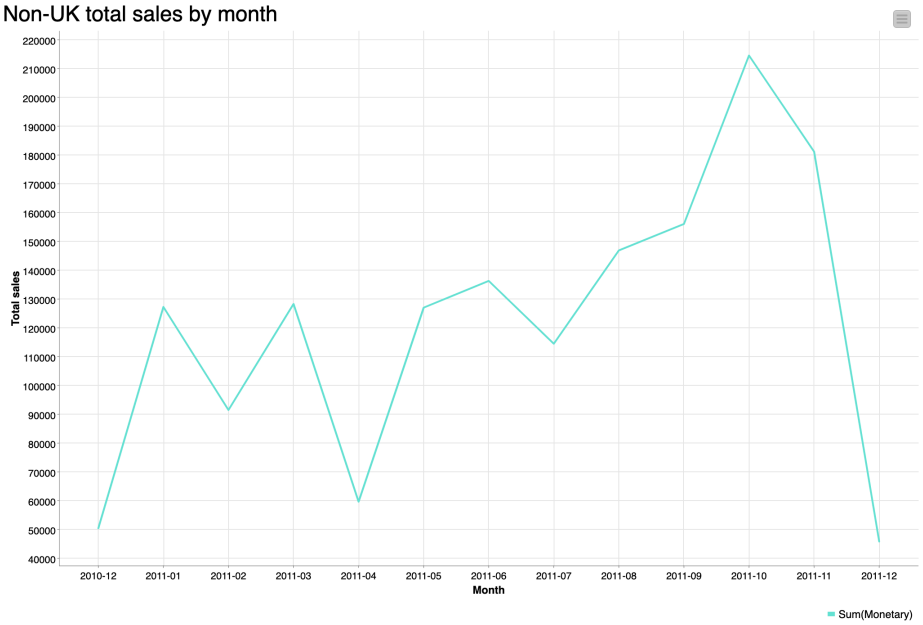
In the next step I do exploratory data analysis which can be seen below:



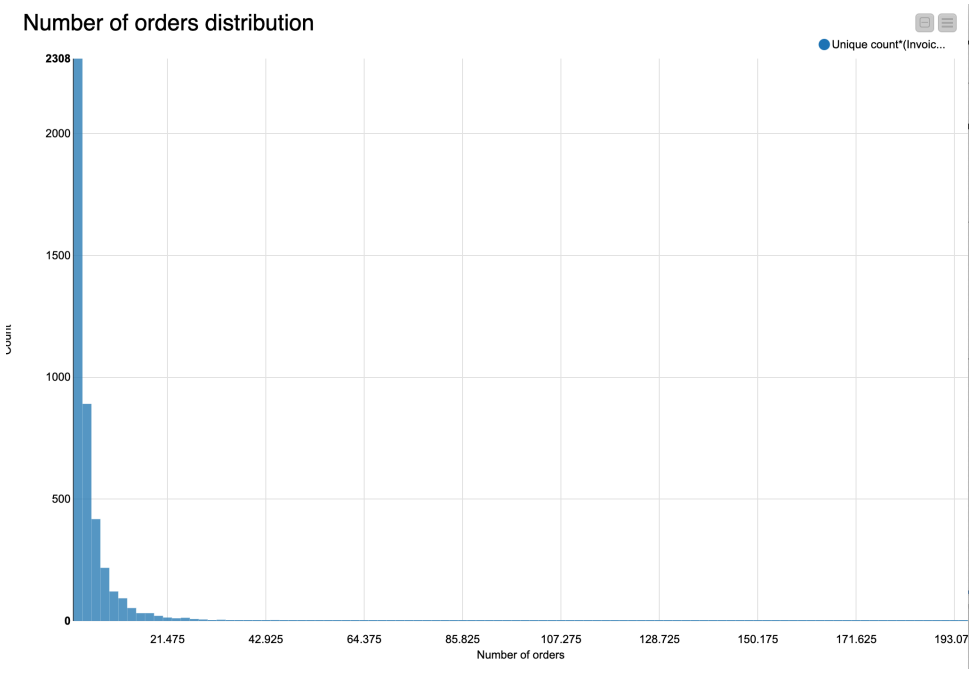




The plots above shows that most of the sales, transactions and customers are from the United Kingdom.



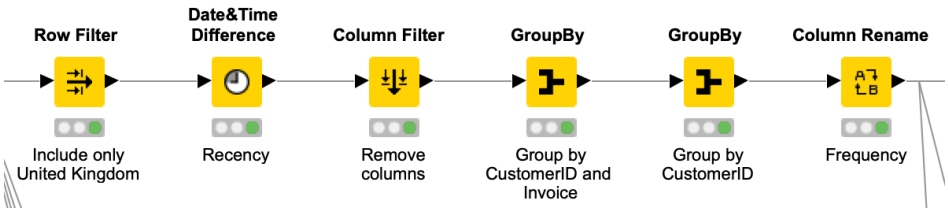
The plots above indicates that in November 2011 there was the highest sales for all the cuntries however for non-UK countries the highest peak accured in October 2011.



The plot above indicates that the most of the customers did only few orders through whole year.

Further pre-processing

I decided to perform analysis only for United Kingdom as we could see on plots above the most of the operations, customers and total sales come from that country. Also according to other researches customer clusters vary in different geographical locations.



In the next step I compute Recency column which refers to the number of days that have elapsed since the customer last made an order.

I remove column which are not needed: StockCode, Description, Quantity, InvoiceDate and Price.

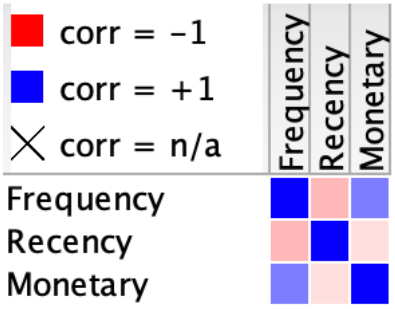
Each invoice in the dataset is split into multiple rows (each for a single product). I group the data by Invoice and Customer ID to have a single row for each invoice. In that step I get the number of days from last purchase and total Monetary values for the invoice. Then I group again only by CustomerID to get number of invoices per customer.

I change Invoice column as Frequency column which refers to the number of invoices with purchases during the year per customer.

All this steps give me Recency, Frequency and Monetary columns for customer level and it is our input dataset which I can now process with k-Means clustering.

Input data analysis

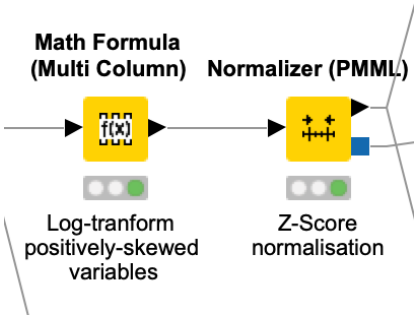
I checked correlation matrix between the variables where we can see negative correlation between Frequency and Recency and positive correlation between Frequency and Monetary variables.



Then I use statistic node to check the distribution of RFM variables and I see that all three variables are positively skew.

Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
Customer ID	12,346	15,555.4385	?	18,287	1,575.7529	-0.0073	-1.187	0	0	0	12.346 18.287
Frequency	1	4.1463	?	196	6.9368	10.4762	203.004	0	0	0	1 196
Recency	1	88.5374	?	366	94.1825	1.2356	0.4124	0	0	0	1 366
Monetary	0.0	1,827.1745	?	233,736.93	7,207.9574	19.3161	499.0171	0	0	0	0 233.737

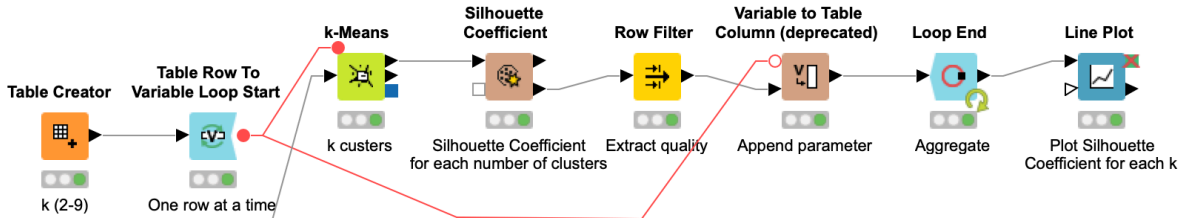
Input data transformation



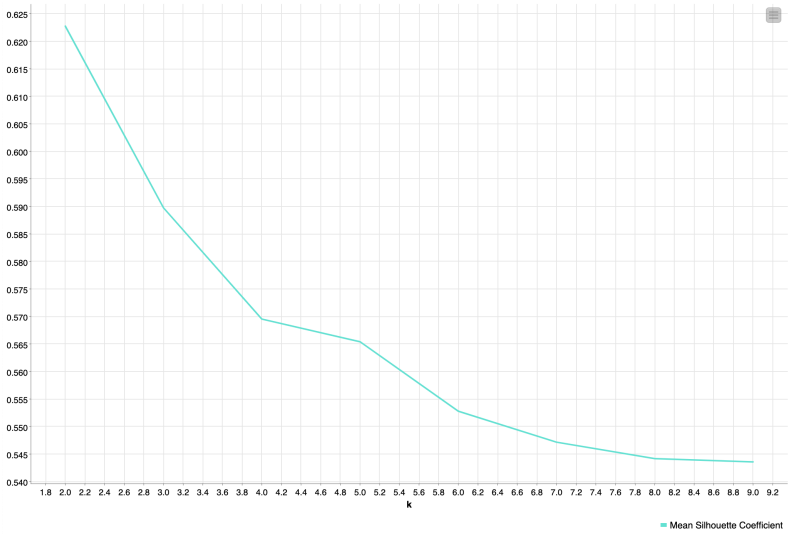
k-Means algorithm requires all variables to be continuous, have normal distribution and be standardised. It is very important because variables with higher variance will, have greater impact on the results. I deal with that by applying log-transformation and then Z-cores normalisation.

Clustering

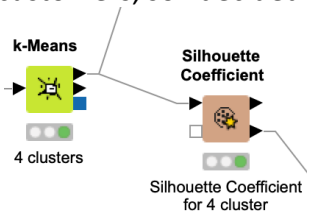
The next step is to determine number of clusters. I decided to use average silhouette coefficient approach which measure the quality of a clustering. It determines how closely each object matched to data within its cluster and how loosely it is matched to data of the neighbouring cluster. A high average silhouette value indicates a good clustering. The optimal number of clusters k is the one that maximizes the average silhouette over a range of possible values for k.



I create a loop which check average silhouette for each k between 2 and 9 and then plot it.



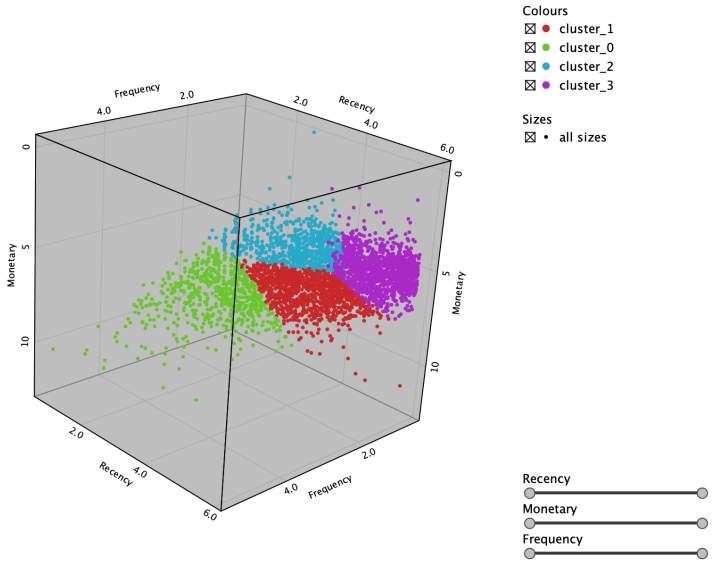
The highest average of silhouette coefficient is for 2 clusters however our decision should be based upon how the business plans to use the results, and the level of granularity they want to see in the clusters. I want to use the results to understand a range of customer behaviour from high-to-low value customers, so I decided to proceed with 4-cluster solution.



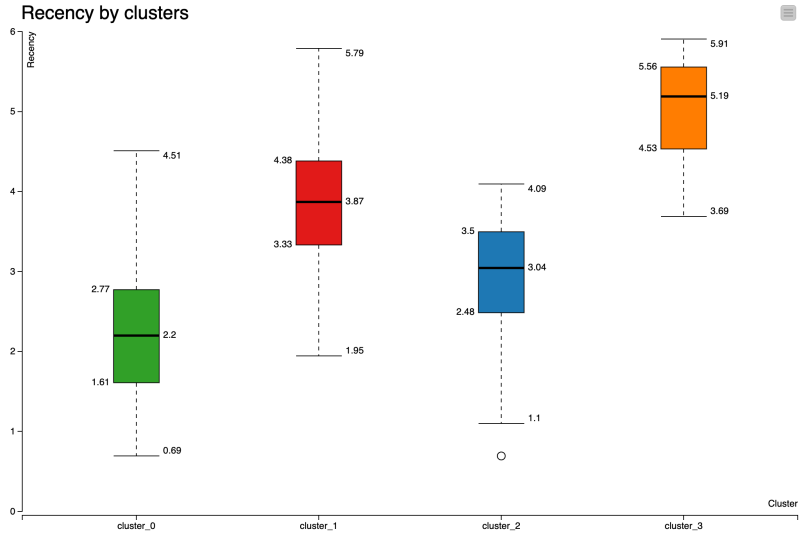
Then I evaluate 4 clusters by computing silhouette coefficient to describe cohesion of points that are within a cluster and separation between clusters. 4 clusters still indicates good separation.

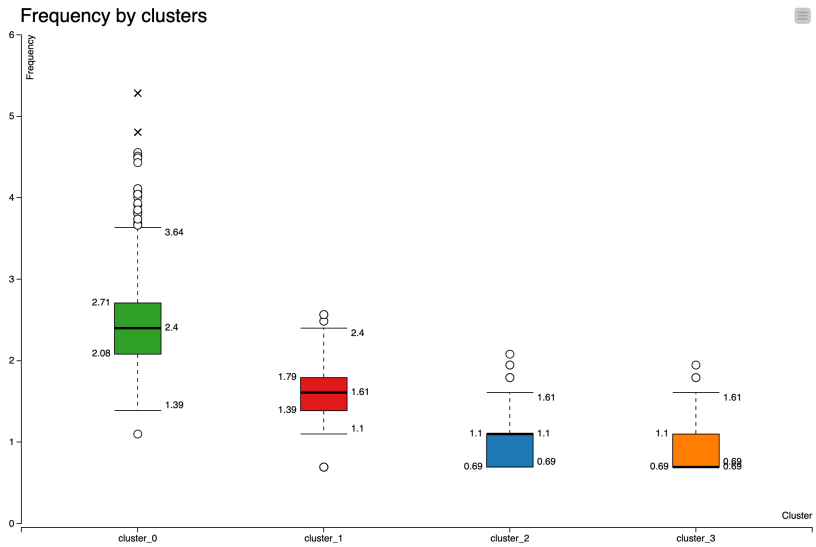
Results visualisation

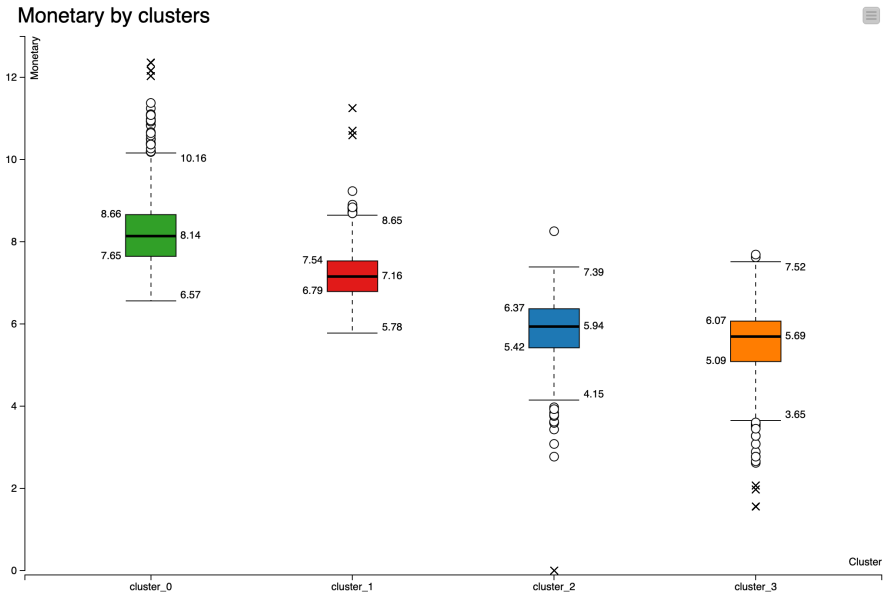
At the end I denormalise data to get again only log-transformed data. I keep log scale for better visualisation purpose. The results of clustering can be seen below in 3D:



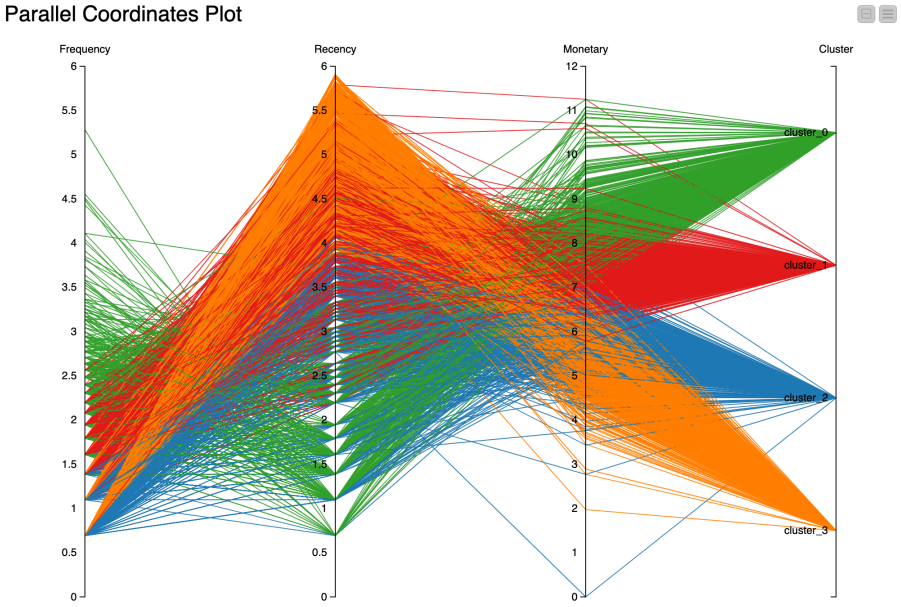
The box plots below show distribution of values for each variable in each cluster.







The parallel plot shows relationship between each variable in each cluster.



In the plots above we can clearly see the pattern in each cluster and the difference between them. This indicates that 4 cluster is a good choice to group the customers.

Conclusions

Below in the table I showed mean values for each variable by clusters.

Interpretation of results:

Cluster_0 – VIP customers. Customers who buy very often, purchaced recently and spent high amount of money.

Cluster_1 – old VIP customers. Customers who buy quite often, purchased quite long time ago and spent quite high amount of money.

Cluster_2 – casual retail customers. Customers who buy very rare, purchased recently and spent relatively small amount of money.

Cluster 3 – old casual retail customers. Customers who buy very rare, purchased very long time ago and spent relatively small amount of money.

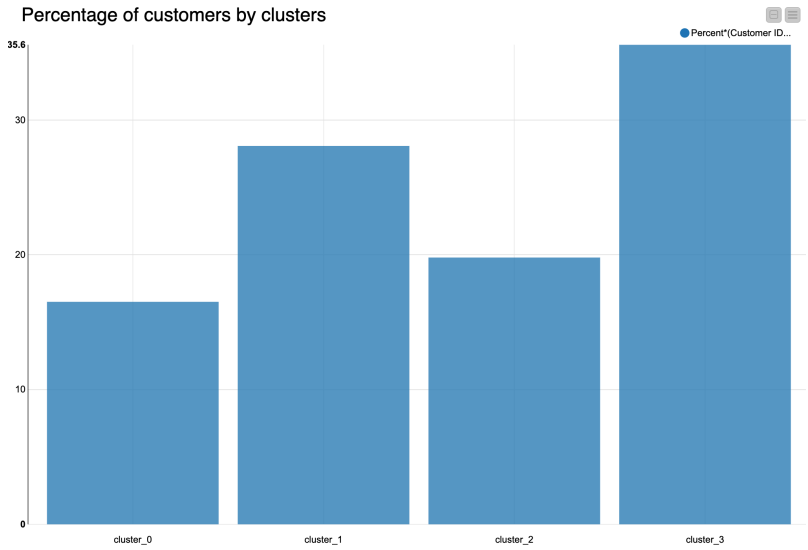
Mean RFM values by clusters

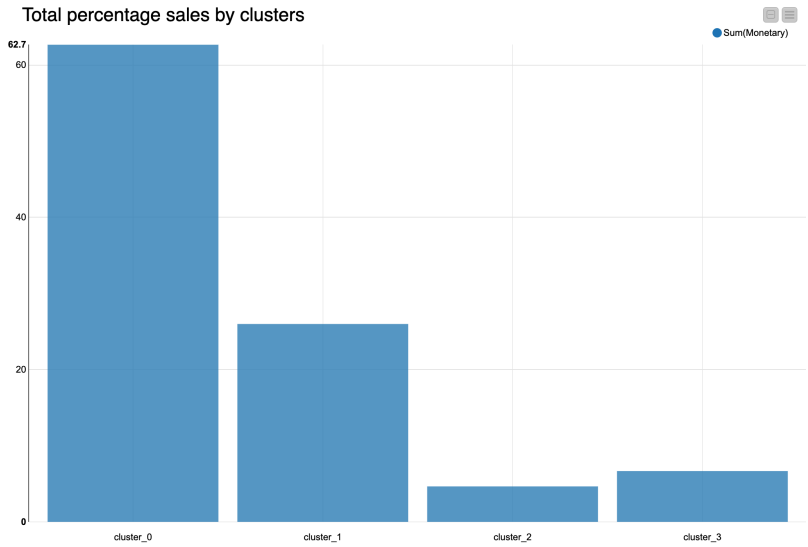
Show 10 entries Search:

RowID	Cluster	Mean(Frequency)	Mean(Recency)	Mean(Monetary)
Row0	cluster_0	13.059561128526637	11.327586206896564	6933.73244514104
Row1	cluster_1	4.105069124423961	61.74654377880186	1690.9276894009188
Row2	cluster_2	1.8209150326797379	22.46797385620918	429.9092941176467
Row3	cluster_3	1.3367272727272725	182.2618181818179	342.6302989090901

Showing 1 to 4 of 4 entries Previous 1 Next

The plots below show that **cluster_0 – VIP** customers, around 17% of all customers, contributed to over 62% of the total sales. On the other hand the highest number of customers, from **cluster_3** which is over 35% of all customers, contributed to only 6% of the total sales.





Natalia Kusek M00736050 – CST4060 Visual Analytics Workflow Design

Customer behavioural segmentation using data mining techniques