University of Warsaw
Faculty of Economic Sciences

Clustering of Customers using K-means algorithm

Subject: Unsupervised Learning
Supervisor: Dr. Jacek Lewkowicz
Authors: Noam Shmuel, Lashari Gochiashvili

Warsaw, 2019

Preface:

We based our analysis on the article *Measuring Customers Satisfaction of E-Commerce Sites Using Clustering Techniques: Case Study of Nyazco Website* by Rezaeian, A., Shokouhyar, S., & Dehghan. The very same article was presented in our class by myself, Noam, and Jakub Byler. As part of our presentation, Jacub and I brought up a lot of criticism on the article, which motivated Lashari Gochiashvili and I to redo the case study with changes and fix ups according to the issues found in the original paper.

Introduction:

Nowadays customers are heavy users of online shopping. For online shoppers, customer data is much more available now than it has been even before. By applying data science methods one can find key trends and apply different marketing strategies to better capture customers and have successful long term relationships. In our paper, we used the dataset of physical retail shoppers. We aim to analyze data and apply a K-means clustering method to better group customers and derive insights from each cluster. Which in the end, it will be value-added for the shopping platforms to enhance their marketing strategies which lead to satisfied customers and shareholders.

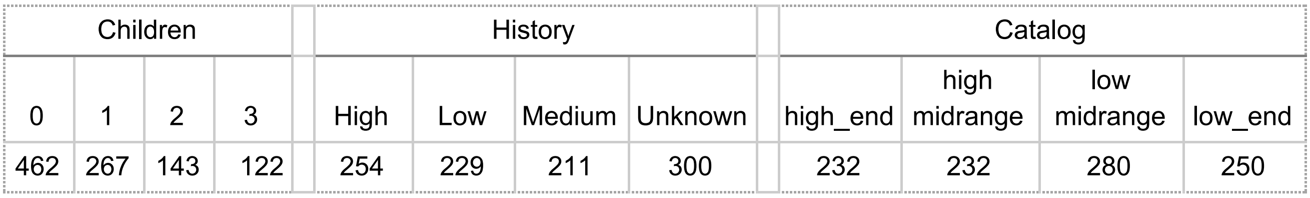
The dataset:

The dataset contains information about customers of a retail shopping site. The dataset has 10 variables and 1000 records (before data clean up).
To prepare the dataset for clustering we applied data cleaning manipulations.
Firstly, we removed a variable due to a significant amount of missing values, left us with 9 variables. Also, we changed the 'Catalog' variable value notations with more intuitive notations. Moreover, we cleaned 3 additional records with missing values in the 'Money Spent' variable, leaving us with 9 variables and 997 records to analyze.
After the first step of cleaning and rearranging the data, we could move forward to get to know the dataset better. The variables consist of 7 factors and 2 integers in the dataset. Factor discrete variables: Age, Gender, Own Home, Married, Location, Children and Catalogs. Continuous variables: Salary and Amount spent.

Exploratory Data Analysis:

The initial step in every analysis is the Exploratory Data Analysis (EDA) and summary statistics. Although it is not the aim of the research, this is a crucial step towards in order to gain a holistic view and understanding of our data, which would lead to the best results at the very end. We shall start with exploring our discrete variables.

Age				Gender			Own Home			Married			Location	
Middle	Old	Young		M	F		Own	Rent		Married	Single		Close	Far
504	205	285		493	501		514	480		500	494		706	288



Children					History					Catalog			
0	1	2	3		High	Low	Medium	Unknown		high_end	high midrange	low midrange	low_end
462	267	143	122		254	229	211	300		232	232	280	250

Figure 1. Table of distribution of categorical variables

The distribution and relation between the continuous variables: Amount Spent and Salary with Gender will be presented in the distributions below.

Most of the customers are in middle age 504 vs 205 old and vs 285 young. Gender is distributed evenly. Same with Own Home and Married variables distributions. The majority lives close 706 vs 288 far. Most customers, 462, in the dataset do not have children. ‘Catalog’ indicated the type of products the customer has bought, and it’s distributed evenly as well.

Analyzing the correlation matrix of the 8 variables reveals us to some intuitive insights and some which are surprising. Both Marriage and Salary, as well as, Money-Spent and Salary are highly correlated positively. Number of children and age are negatively correlated, in that dataset, older people have a small amount of children- either 0 or 1. Another surprising insight coming from the correlation matrix is that marriage and number of children are not correlated! After looking into it we found out that married and singles have about the same amount of children. Less surprising and more intuitive of no correlation is apparent between Gender and Age. More about the relation:

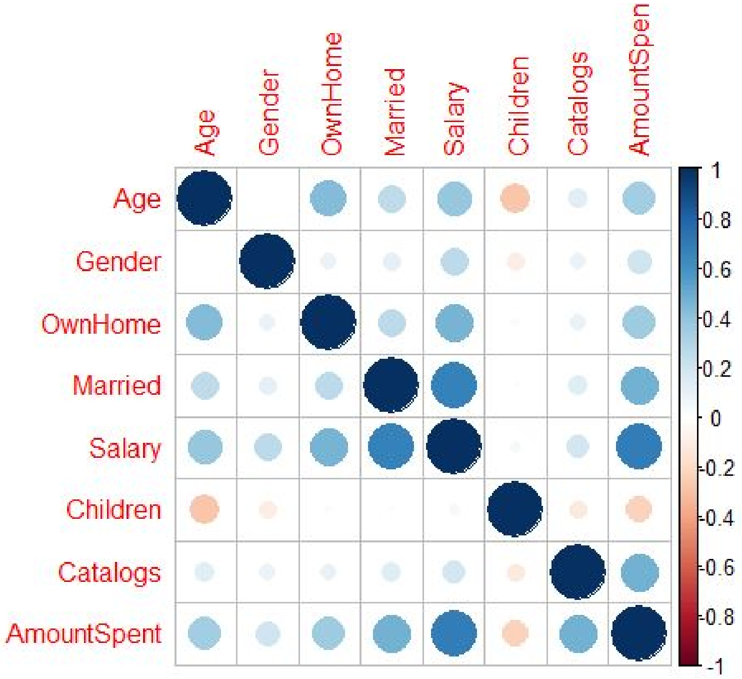


Figure 2. Correlation Matrix

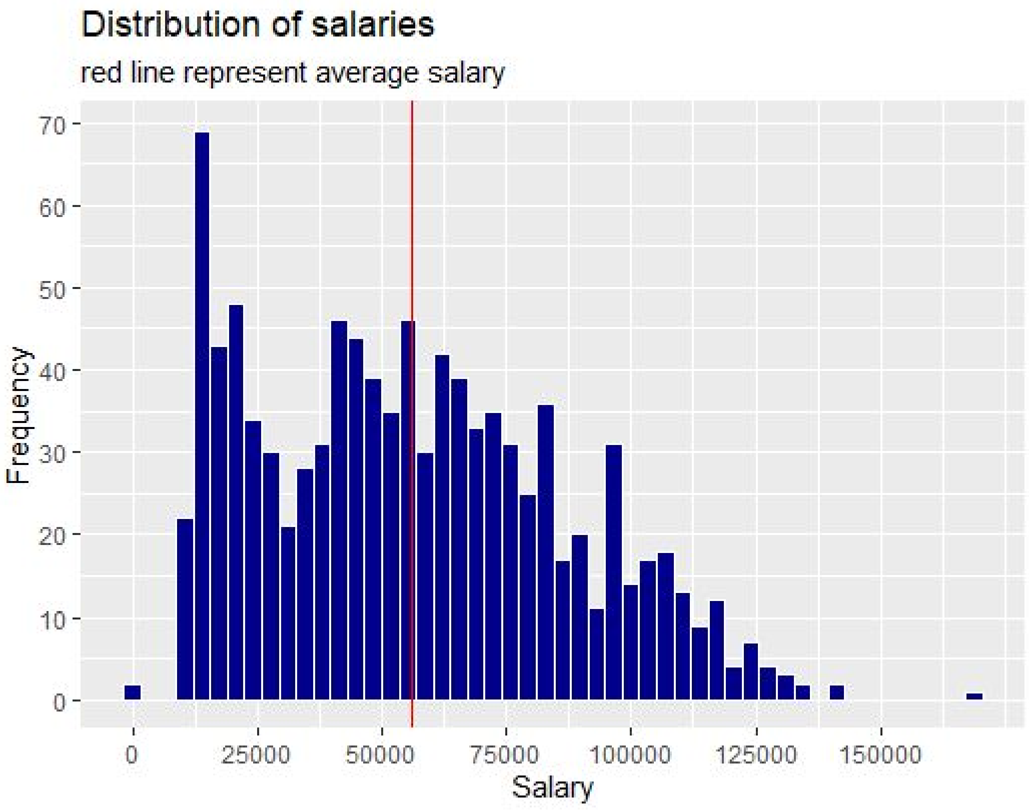


Figure 3. Distribution of salaries

Salary distribution is skewed to the right, with an average salary of 56032.

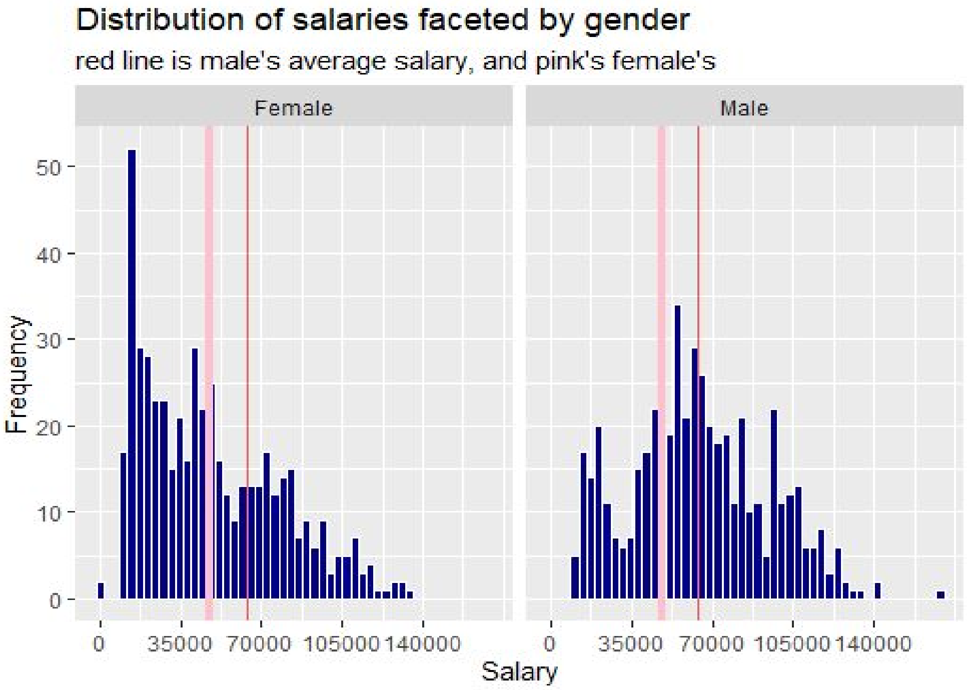


Figure 4. Distribution of salaries for Female (left) and Male (right)

Distribution of salaries for Male is close to a normal distribution, while the distribution of salaries for Female has a heavy tail and positive skewness. As we already have seen in the correlation matrix, males have higher average salaries.

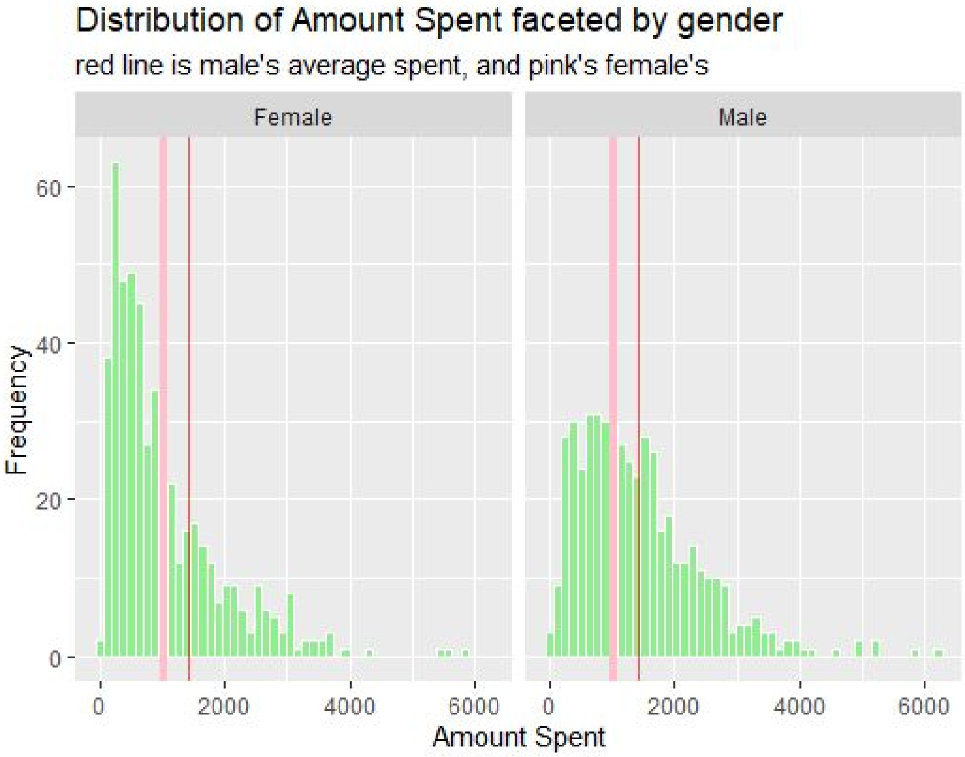


Figure 5. Distribution of Amount Spent by Female (left) and by Male (right)

Looking at the distribution of Amount Spent is close to the distribution of Salaries by gender (and as we have seen- positively correlated). Males spend on average 37.3% more than Females.

Summary statistics and exploratory data analysis and visualization of the data are an integral part of any analysis. It is extremely important for Unsupervised Learning as in the upcoming classes we will get familiar with dimensionality reduction techniques, and the decision of which dimensions to reduce is manly appointed by EDA.

As we are familiar with our data, the next step would be the clustering analysis using K-means.

Clustering Techniques for Segmentation of Customers

Clustering is a data mining method which is using customer data to segment customers into groups in a way that members of one group have big similarities within the group members while they do not have many similarities with other group members. One of the most widespread methods in clustering is K-means method. This method, in simple words, is taking K numbers from all observations. These K numbers are the centers of clusters. Then the calculation is run to identify to which cluster each member of observation belongs by using Euclidean distance. Every added observation new centers of clusters are calculated and new observation is assigned to the relevant cluster.

Methodology

After cleaning up and rearranging the data, we were left with 996 records and 9 variables. Using K-mean we defined 4 clusters, 4 different types of customer segments, each one with its unique characteristics of customers.

The decision of choosing 4 clusters was backed by different validity measures: Total Within-Sum-Squares (‘Elbow method’), silhouette score and Calinski-Harabasz index. Every single one of these measures plays a role in the decision of picking the optimal number of clusters.

The within-sum-squares (WSS) depicted in the figure below as the ‘Elbow method’. The idea behind the visualization is very simple indeed, the total WSS in a low number of clusters analysis is high due to heterogeneity. When we raise the number of clusters, it reduces the number of total WSS due to higher similarity. One would think that it would be best to raise the number of clusters to the maximum and reduce the total WSS. That’s true, to some extent, however, there is a trade-off between the meaning of a number of clusters to the total WSS. That’s the Achilles heel of the method.

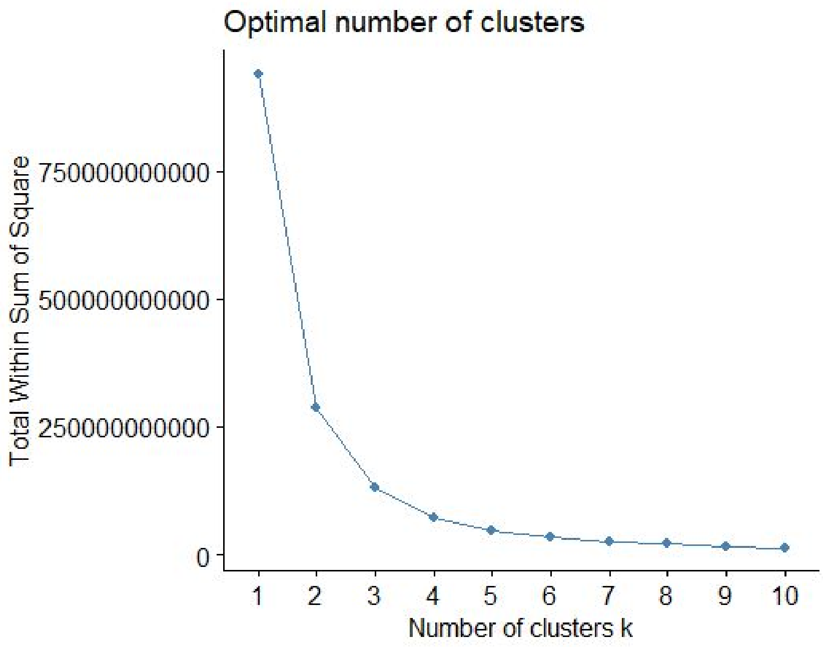


Figure 6. Elbow method for the optimal number of clusters

By looking at the figure above we can see that indeed the total WSS is being reduced with each incremented number of clusters. However, it's also very visible to notice that the change is less significant after, some would say the 3rd, some would say the 4th number of clusters.

As the 'Elbow method' did not provide conclusive evidence on how many clusters would be optimal for our data, it did give us the impression that it would be either 2 or 4.

Moved to the next validation method, the silhouette score.

$$S(x) = \frac{(b(x)-a(x))}{\max\{b(x),a(x)\}}$$

Formula 1. Silhouette Score

Unlike the total WSS, the result of the silhouette score is global for every dataset and it ranges between -1 to +1.

Firstly, the interpretation of the silhouette score formula above is as follows: b(x) would be the minimum average distance between x and the closed neighbor cluster, while a(x) would be the average distance within the cluster. The difference between these two averages normalized by the maximum of two. In simple words, if all points were assigned optimally, the difference between b(x) and a(x) would be great and the score would be close to +1. On the other hand, if all the points were assigned to the wrong cluster, we would get a score close to -1. A score of 0 simply means that there is a similar cluster that would be as good as the clustered originally assigned. More specifically to our data:

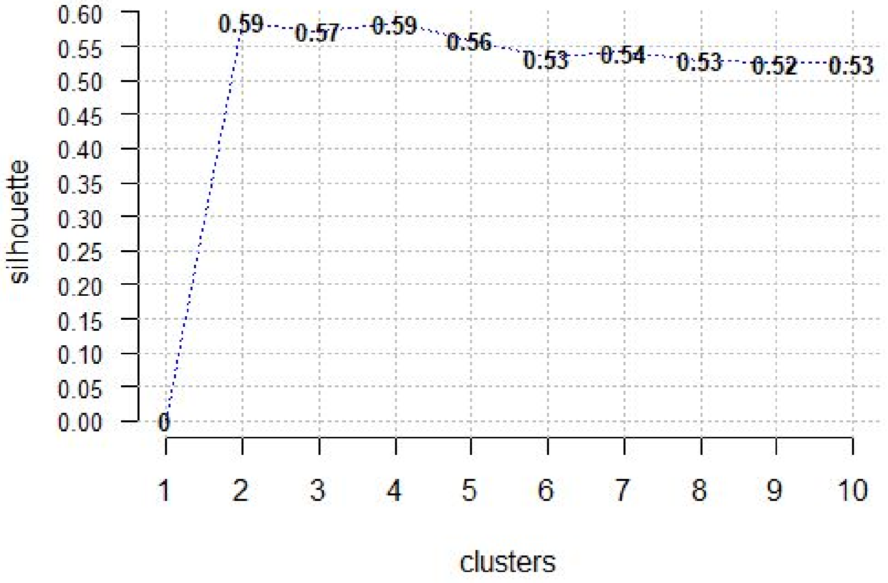


Figure 7. Silhouette score

The intuition which given to us by the total WSS, that would be best to pick either 2 or 4 clusters is backed by similar silhouette score.

Confident with a relatively high score of 5.9 we move on to deciding what would be the best number of clusters, 2 or 4. This is where the Calinski-Harabasz index comes into the picture.

Calinski-Harabasz index, unlike the silhouette score, is not global rather relative. Meaning, the Calinski-Harabasz index is best used to compare for the same data different number of clusters, which is exactly the situation we have encountered. The formula is as follows:

$$\frac{SSB / k - 1}{SSW / N - k}$$

Formula 2. Calinski-Harabasz index

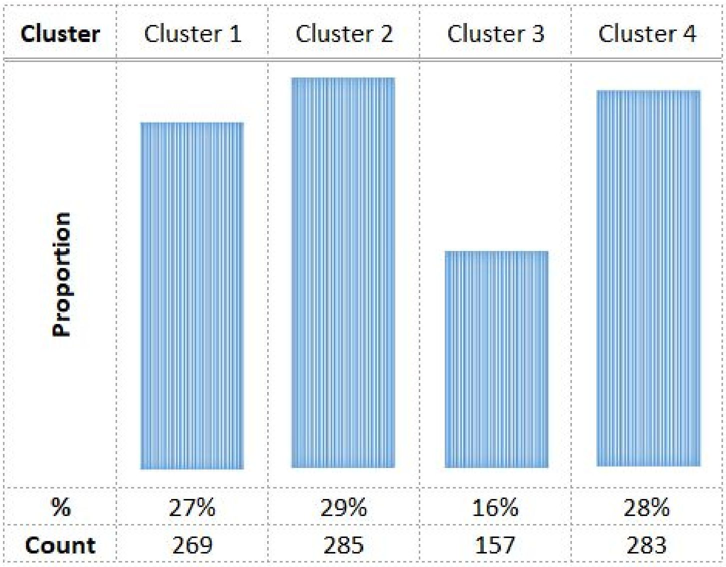
SSB denotes the sum of squares between clusters, while SSW is within the sum of squares (N is the number of observations and K is the number of clusters). In simple words, high variation between clusters- SSB divide by low variation within a cluster is our goal.

Hence, the higher the index the better results. In our data, the Calinski-Harabasz index of 2 clusters is: 2257.6, while for 4 clusters is 4018.9

After conducting 3 independent validity measures, total within the sum of squares, silhouette score, and Calinski-Harabasz index, there is no doubt in our mind that the optimal number of clusters for our data is 4.

The results

The distribution of customers between the four cluster is as following:



Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4
%	27%	29%	16%	28%
Count	269	285	157	283

Figure 8. Distribution of customers within four clusters

Clusters 1,2,4 are distributed almost evenly with 269, 285 and 283 clients respectively, while cluster number 3 has 157 clients. The results of the clustering are stated below in **figure 9**.

			Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
		Appearances	269	285	157	283
Gender	Female	501	189	145	53	114
	Male	493	80	140	104	169
Age	young	285	208	55	0	22
	middle	504	10	168	130	196
	old	205	51	62	27	65
OwnHome	Rent	480	213	154	26	87
	Own	524	56	131	131	196

Married	Single	494	247	183	0	64
	Married	500	22	102	157	219
Location	Far	288	90	74	39	85
	Close	706	179	211	118	198
Children	0	462	116	148	54	144
	1	267	78	75	53	61
	2	143	41	34	22	46
	3	122	34	28	28	32
Catalog	low-end	250	98	67	22	63
	mid-range low	180	79	86	50	65
	mid-range high	232	46	70	38	78
	high-end	232	46	62	47	77
Salary	Min		0	33,000	90,100	59,400
	Mean		19,516	46,154	106,653	72,609
	Max		32,700	59,300	168,800	89,500
AmountSpent	Min		0	105	213	177
	Mean		401	1,006	2,261	1,629
	Max		1,320	3,044	6,217	5,209

Figure 9. Results of the clustering

Cluster number-1 mostly young single women with no children, who live in rent. The cluster has the lowest average salary and the lowest average amount spent.

Cluster number-2 middle age and old men and women, with mostly no or single children who buy mid-range products. The second-lowest average salary and amount spent.

Cluster number-3 mostly middle-aged men, own homes. Every single person in the cluster is married. Relative to the other clusters they have the highest ratio of 2 and 3 children. They made the highest salaries and spend the most.

Cluster number-4 middle age who mostly own homes and are married. Buy high-end products and spend the second-highest amount.

The characteristics of each cluster are highly distinctive and create almost homogeneous segments. It's easy to notice that cluster number 1 has the "least valuable" customers when it comes to generating money, however, we have no data about the purchasing frequency of this segment. But the picture that depicted from this segment is of a young female student, who doesn't make a lot of money and doesn't spend it either. Cluster number 3, on the other hand, are middle-aged men, who have a steady high income, owns children and spend the highest amount. Cluster number 2 is middle age, and old customers who don't make a lot of money and don't spend much, almost similar to cluster number 4 but these middle-age do have high salaries and do spend a lot, mostly on high-end products.

Conclusions

K-means has proven itself as a simple yet robust classification method. The results are clear and each clustered segment has different distinguishing characteristics, and that's no surprise why these methods are highly used in the marketing industry for the purpose we have shown in our case study- segmentation of clients but also for matching the best products similar to what the clients might have bought. The results are more satisfied clients, satisfied clients tend to use our service/product more often, which generates more income. All parties are satisfied, and the social surplus increases.

References:

Rezaeian, A., Shokouhyar, S., & Dehghan, F. 2016. Measuring Customers Satisfaction of E-Commerce Sites Using Clustering Techniques: Case Study of Nyazco Website. International Journal of Management, Accounting and Economics, 3(1), 61-74.

Customer Segmentation Using K Means Clustering, a blogpost. https://towardsdatascience.com/customer-segmentation-using-k-means-clustering-d33964f238c3 (accessed 19.11.12)

K-means clustering, a wikipedia note. https://en.wikipedia.org/wiki/K-means_clustering (accessed 19.11.14)

R package description on cran.r-project.org: An Introduction to corrplot Package. https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html (accessed 19.11.14)

R package description on r-blogger: RFM Analysis in R https://www.r-bloggers.com/rfm-analysis-in-r/ (accessed 19.11.15)

Skewness, a wikipedia note. https://en.wikipedia.org/wiki/Skewness (accessed 19.11.15)

Data Science Project – Customer Segmentation using Machine Learning in R. https://data-flair.training/blogs/r-data-science-project-customer-segmentation/ (accessed 19.11.18)

The Data set was taken from Kaggle.com: https://www.kaggle.com/arawind/retail-marketing