UNSUPERVISED LEARNING-RETAIL CUSTOMER DATASET

ABSTRACT

This project uses 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.

GOALS AND METHODS OF THE REPORT

The main aim of the project is to analyse the dataset using unsupervised method - Clustering which is a data mining method which is used in 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.

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.

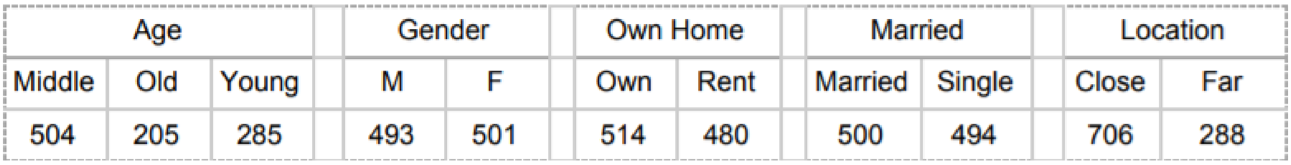
The study will include the theoretical background of the unsupervised learning method   used for the analysis.

FINDING AND KEYS

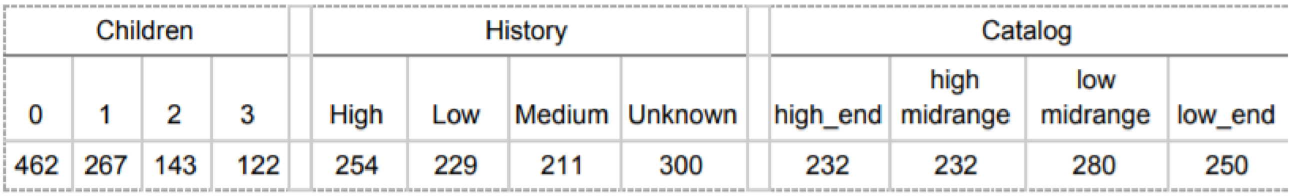
- Distribution of salaries for Male is close to a normal distribution, while the distribution of salaries for Female has a heavy tail and positive skewness
- Money spent and salary were correlated to each other
- Number of children and Age were negatively correlated
- Segmentation of customers is done to increase the revenue of the company
- There were 4 clusters in k means clustering

DATA CLEANING AND PREPROCESSING

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



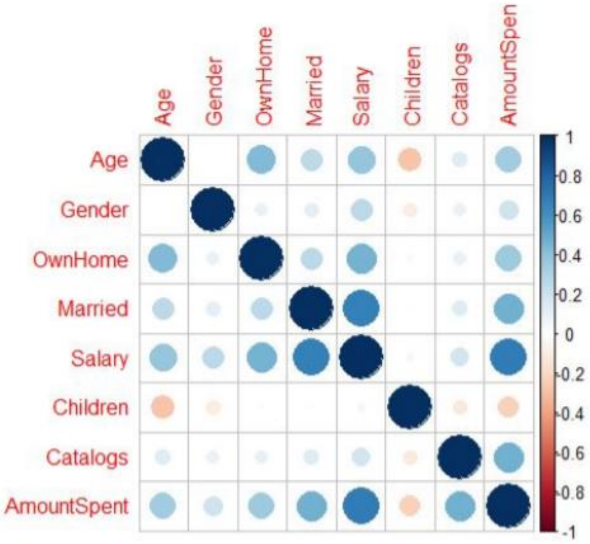
Age			Gender		Own Home		Married		Location	
Middle	Old	Young	M	F	Own	Rent	Married	Single	Close	Far
504	205	285	493	501	514	480	500	494	706	288



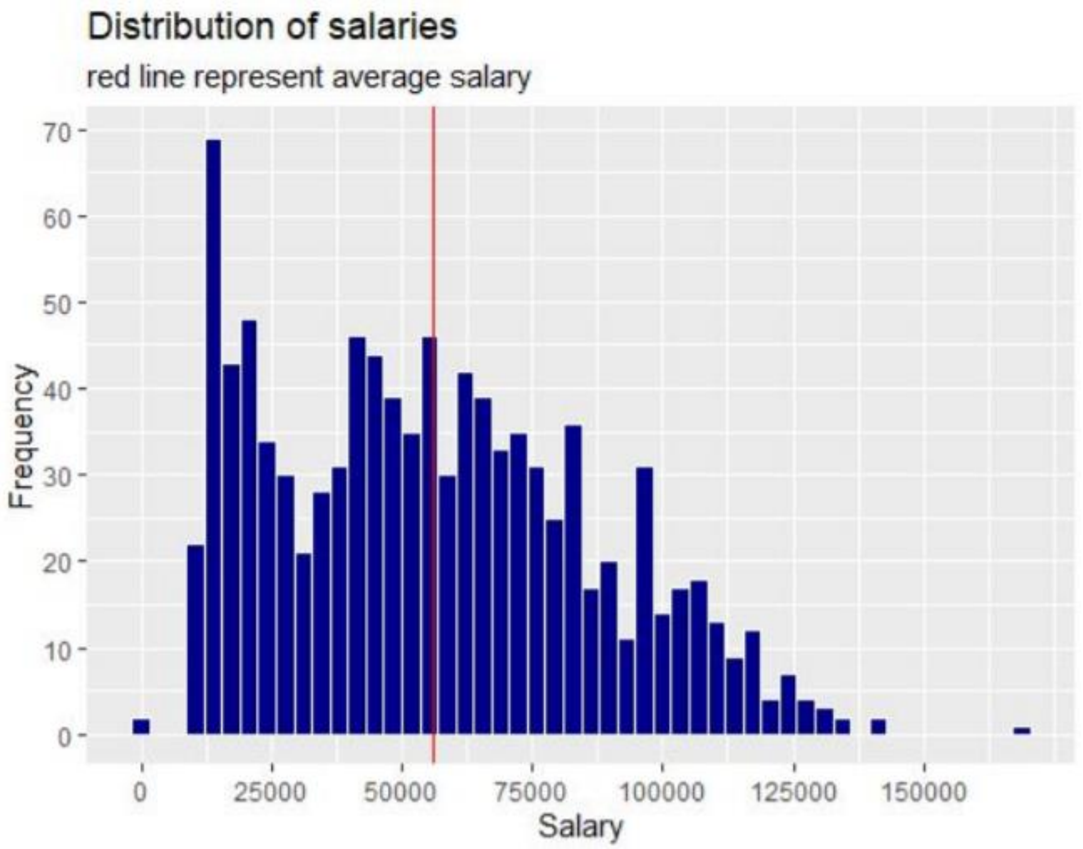
Children				History				Catalog			
0	1	2	3	High	Low	Medium	Unknown	high_end	high midrange	low midrange	low_end
462	267	143	122	254	229	211	300	232	232	280	250

The distribution between amount spent and salary is first found along with gender

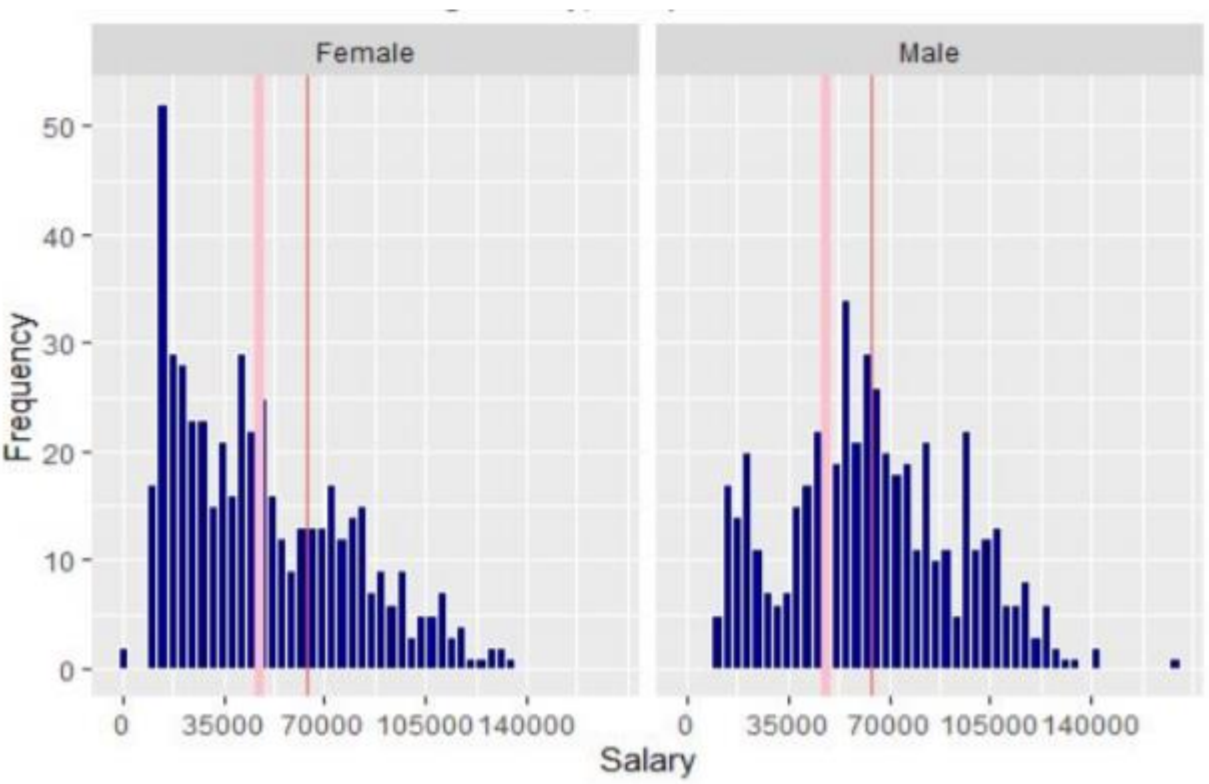
Most customers aged between middle age i.e. 504 of them ,young people who are 205 of them and young people that is 285 of them based on gender are evenly distributed. 706 of them were living close to each other while 288 were living far.Most customers about 462 of them do not have children.-Catalog- indicated the type of products the customers bought and it is evenly distributed as well.



After analysing the correlation matrix its found that 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 less number of children- either 0 or 1. Another factor 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. There was found to be no correlation is apparent between Gender and Age.

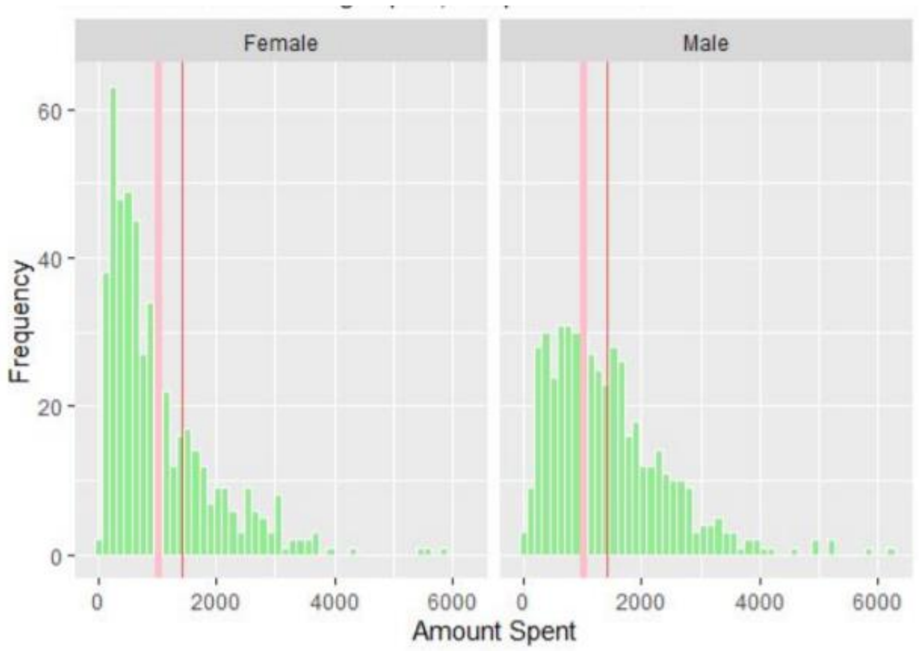


The salary distribution is skewed to the right with an average salary of 56032



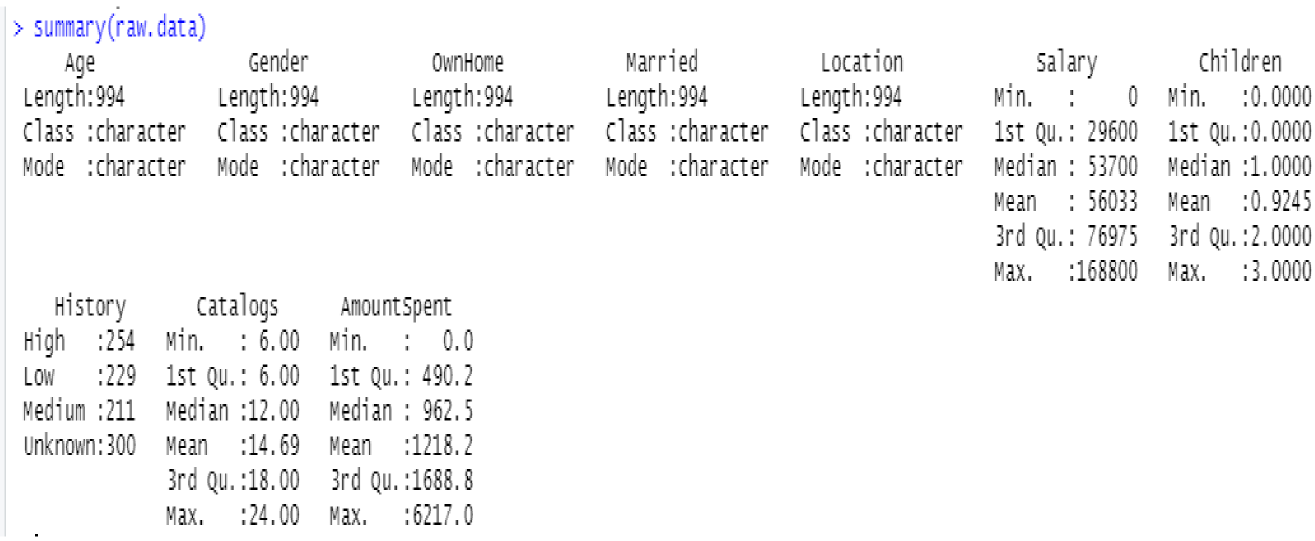
The salary distribution among male and female. Red line indicates male’s average salary and red line indicates the female’s average salary.

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.



If we look at the amount spent by each gender it looks close to the distribution of salary which is actually positively correlated. It is found that Male spends 37.3% more than female.

Summary of the data is as follows.



```
> summary(raw.data)
     Age               Gender            OwnHome             Married            Location             Salary          Children     
 Length:994         Length:994         Length:994         Length:994         Length:994         Min.   :     0   Min.   :0.0000  
 Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.: 29600   1st Qu.:0.0000  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median : 53700   Median :1.0000  
                                                                                                Mean   : 56033   Mean   :0.9245  
                                                                                                3rd Qu.: 76975   3rd Qu.:2.0000  
                                                                                                Max.   :168800   Max.   :3.0000  
    History       Catalogs      AmountSpent    
 High   :254   Min.   : 6.00   Min.   :   0.0  
 Low    :229   1st Qu.: 6.00   1st Qu.: 490.2  
 Medium :211   Median :12.00   Median : 962.5  
 Unknown:300   Mean   :14.69   Mean   :1218.2  
               3rd Qu.:18.00   3rd Qu.:1688.8  
               Max.   :24.00   Max.   :6217.0  
```

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

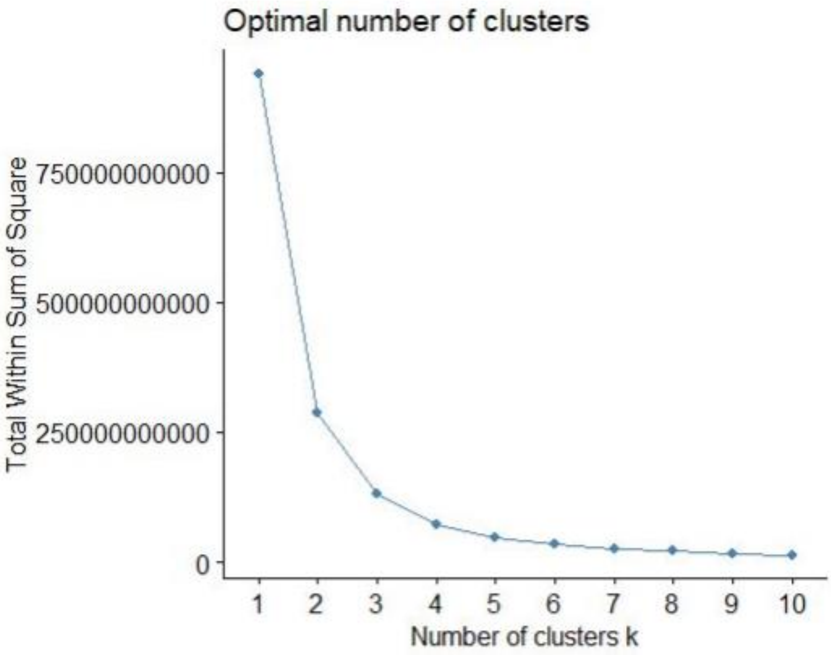
After cleaning up and rearranging the data, we were left with 996 records and 9 variables. Using K-mean we defined 4 clusters, 4 different types of customer segments, each one with its unique characteristics of customers. The decision of choosing 4 clusters was backed by different validity measures: 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.

ELBOW METHOD

In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.

If the line chart looks like an arm, then the "elbow" (the point of inflection on the curve) is the best value of k. The "arm" can be either up or down, but if there is a strong inflection point, it is a good indication that the underlying model fits best at that point.

The within-sum-squares (WCSS) depicted in the figure below is the 'Elbow method'. The idea behind the visualization is very simple indeed, the total WCSS in a lower when number of clusters is high due to heterogeneity. When we raise the number of clusters, it reduces the number of total WCSS 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.



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.

We could assume it as 2 or also 4 so since there is no proper result and since there is inconclusive evidence on how many clusters would be optimal we considered next method that is the silhouette score method

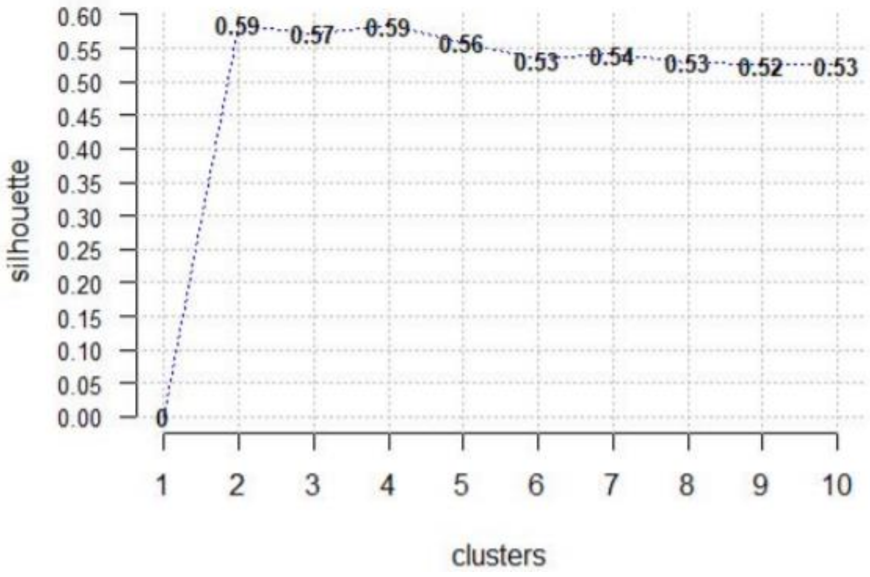
SILHOUETTE SCORE METHOD

Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object has been classified.

It is considered which is given by the formula.

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

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 neighbour 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:



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. We then try considering another method also that is the Calinski-Harabasz index

CALINSKI-HARABASZ INDEX

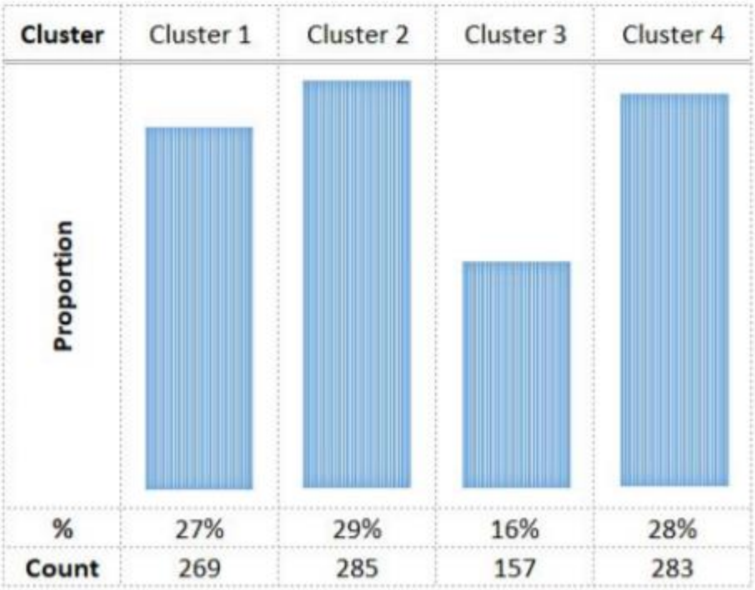
Calinski-Harabasz index, unlike the silhouette score, is not global rather relative. The Calinski-Harabasz index also known as the Variance Ratio Criterion, is the ratio of the sum of between-clusters dispersion and of inter-cluster dispersion for all clusters, the higher the score , the better the performances. 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

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

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.

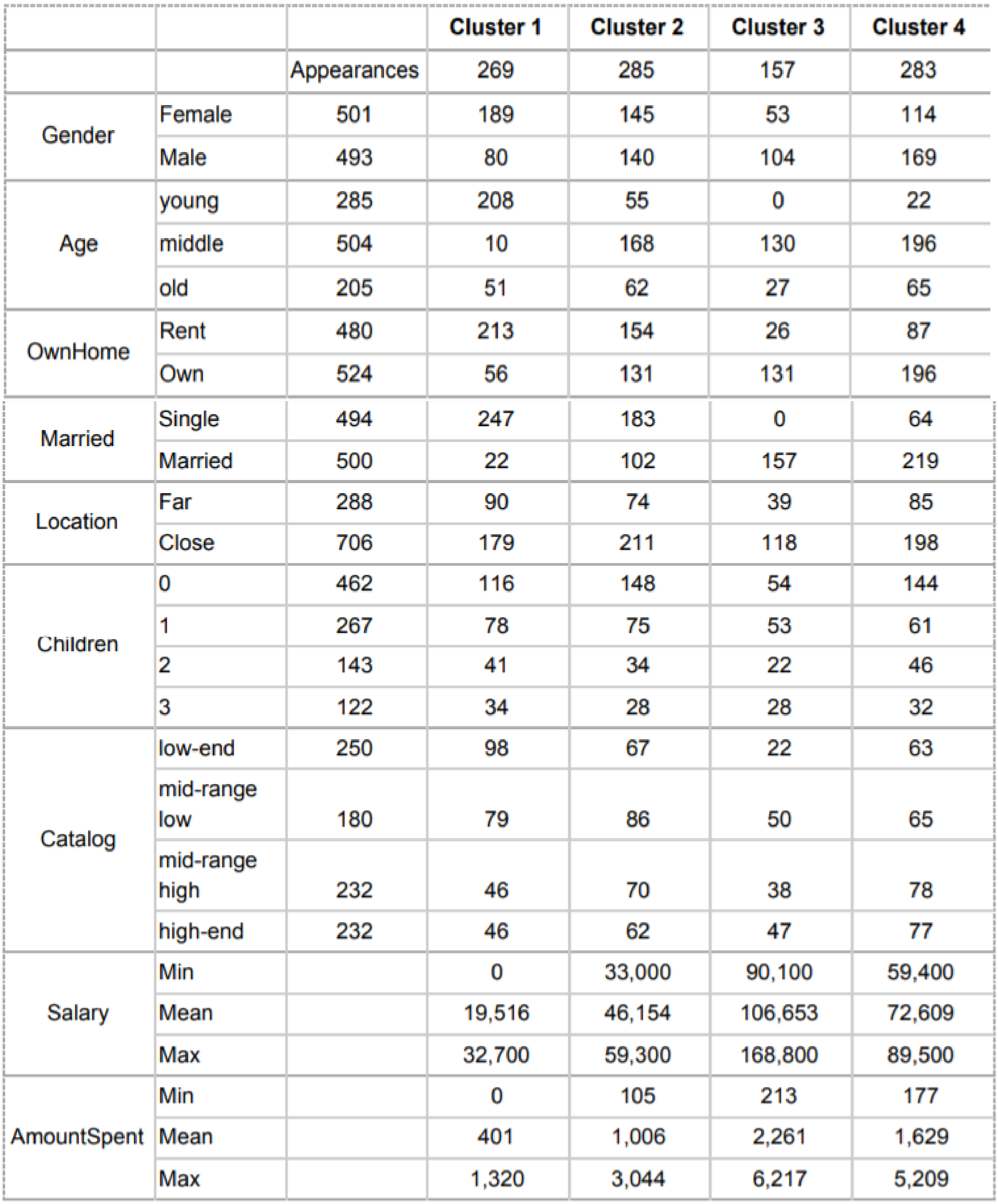
THE RESULTS

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

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



			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

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. After carrying about the clustering with three different methods of clustering the optimal number of clusters was found to be 4. The people belonging to the cluster 3 were the least and found to be 157

APPENDIX

```
library(dplyr)
library(ggplot2)
library(data.table)
library(ggcorrplot)
library(pastecs)
library(textir)
library(cluster)
library(factoextra)
library(flexclust)
library(fpc)
library(clustertend)
library(ClusterR)
library(data.table)
library(smacof)
library(corrplot)



raw.data <- read.csv("https://raw.githubusercontent.com/LashaGoch/Clustering-
with-K-means-in-R/master/retailMarketingDI.csv")

View(raw.data)



str(raw.data)

#There are 10 variables: ...

#and 1000 records (before cleaning the data)





########################## 1st PART : cleaning and orginazing the data
##########################

str(raw.data)



table(is.na(raw.data$Age)) #no NA

table(is.na(raw.data$Gender)) #no NA

table(is.na(raw.data$OwnHome)) #no NA

table(is.na(raw.data$Married)) #no NA

table(is.na(raw.data$Location)) #no NA

table(is.na(raw.data$Salary)) #no NA

table(is.na(raw.data$Children)) #no NA

table(is.na(raw.data$History)) #There are 303 NAs, which I will replace with
'Unknown'

table(is.na(raw.data$Catalogs)) #no NA
```

```r
table(is.na(raw.data$AmountSpent)) ##There are 6 NAs, records which I can remove OR do forcast on them -


#first I will replace the NAs of History with 'Unknown':

raw.data$History <- as.character(raw.data$History)

raw.data$History[is.na(raw.data$History)] <- 'Unknown'

raw.data$History <- factor(raw.data$History)

table((raw.data$History)) # worked successfully.


# Remove the 6 NAs with no amount spent

retail.df <- raw.data[!is.na(raw.data$AmountSpent),]


#  Factorize the Children veriable

retail.df$Children <- factor(retail.df$Children)




View(retail.df%>%

       group_by (Catalogs) %>%

       summarise (mean_of_amount = mean(AmountSpent),numebr_of_appirances = n()))

#By looking at the table above we can see that variabele Catalogs is actually a factor variable where 6 is the 'low_end' prices and 24 is 'high_end' products

#Where 12 and 16 are the mid_range products

#therefore I will change the notation to more intuitive notation (althoth there is no real change in the content) :


retail.df<- (retail.df %>%

               mutate(Catalog = ifelse (Catalogs ==6, 'low_end',

                                        (ifelse(Catalogs == 12, "low_midrange",

                                                (ifelse(Catalogs == 18, "high_midrange", "high_end")))))))


#I will factorize the new variable

retail.df$Catalog <- as.factor(retail.df$Catalog)
```

```
#And remove the old one:
retail.df$Catalogs <- NULL


str(retail.df) # we are left with 10 variables, 8 of them are factors and 2 are
integers (salary + amount spent)


############# 2nd PART : getting to know the Data (summary statistics and EDA)
################


# showing the distribution of each categorical veriables
lapply( retail.df %>%
          select(c("Age", "Gender", "OwnHome", "Married", "Location",
"Children", "History","Catalog"))
        ,table)


ggplot(data = retail.df, aes(x = Salary))+
  geom_histogram(bins = 50, colour = 'white', fill = 'darkblue')+
  scale_x_continuous(breaks = seq(0,150000,25000))+
  scale_y_continuous(breaks = seq(0,70,10))+
  xlab("Salary")+
  ylab("Frequency")+
  ggtitle("Distribution of salaries")+
  geom_vline(xintercept = mean(retail.df$Salary), color = 'red')+
  labs(subtitle  = 'red line represent average salary')



mean_salary_female <- mean(retail.df$Salary[retail.df$Gender =="Female"])
mean_salary_male <- mean(retail.df$Salary[retail.df$Gender =="Male"])



ggplot(data = retail.df, aes(x = Salary))+
  geom_histogram(bins = 50, colour = 'white', fill = 'darkblue')+
  scale_x_continuous(breaks = seq(0,150000,35000))+

  scale_y_continuous(breaks = seq(0,70,10))+

  xlab("Salary")+

  ylab("Frequency")+

  ggtitle("Distribution of salaries faceted by gender")+

  geom_vline(xintercept = mean_salary_female, color = 'pink',size=1.5)+

  geom_vline(xintercept = mean_salary_male, color = 'red', alpha= 0.6)+

  labs(subtitle  = "red line is male's average salary, and pink's female's")+

  facet_wrap(~Gender)


#explain the distributions, males is more normally distributed, while womens is
right skewed


mean_AmountSpent_female <- mean(retail.df$AmountSpent[retail.df$Gender
=="Female"])

mean_AmountSpent_male <- mean(retail.df$AmountSpent[retail.df$Gender =="Male"])


ggplot(data = retail.df, aes(x = AmountSpent))+

  geom_histogram(bins = 50, colour = 'white', fill = 'lightgreen')+

  scale_x_continuous()+

  scale_y_continuous()+

  xlab("Amount Spent")+

  ylab("Frequency")+

  ggtitle("Distribution of Amount Spent faceted by gender")+

  labs(subtitle  = "red line is male's average spent, and pink's female's")+

  facet_wrap(~Gender)+

  geom_vline(xintercept = mean_AmountSpent_female, color = 'pink',size=1.5)+

  geom_vline(xintercept = mean_AmountSpent_male, color = 'red', alpha= 0.6)


#Again explain the distributions, males is more normally looking distributed,
while womens is right skewed
#installed.packages("corrplot")

raw.data <- raw.data[!is.na(raw.data$AmountSpent),]
head(raw.data$Age, 10)
```

```
cor.data <- raw.data
levels(raw.data$Age)
cor.data$Age <- ifelse(cor.data$Age == 'Young', 0,
                       ifelse(cor.data$Age == 'Middle',1,2))

levels(raw.data$Gender)
cor.data$Gender <- ifelse(cor.data$Gender == "Female", 0 ,1)


levels(raw.data$OwnHome)
cor.data$OwnHome <- ifelse(cor.data$OwnHome == "Rent", 0 ,1)


levels(raw.data$Married)
cor.data$Married <- ifelse(cor.data$Married == "Single", 0 ,1)

levels(raw.data$Location)
cor.data$Location_close <- ifelse(cor.data$Location == "Far", 0 ,1)

cor.data$History<- NULL
cor.data$Location<- NULL


str(cor.data)
library(corrplot)

cor.maxtrix<- cor(cor.data, method = "pearson", use = "complete.obs")

corrplot(cor.maxtrix)
#correlation matrix


library(ggplot2)

par(mfrow=c(1,7))

barplot(table(raw.data$Age), main="Age", col = "#69b3a2")
barplot(table(raw.data$Gender), main="Gender", col = "#A9A9A9")
barplot(table(raw.data$OwnHome), main="Own Home?", col = "#69b3a2")
barplot(table(raw.data$Married), main="Married", col = "#A9A9A9")
barplot(table(raw.data$Location), main="Location", col = "#69b3a2")
barplot(table(raw.data$Children), main="Children", col = "#A9A9A9")
barplot(table(raw.data$Catalog), main="Catalog", col = "#69b3a2")

par(
  mfrow=c(1,2),
  mar=c(4,4,1,0)
)
hist((raw.data$AmountSpent), xlab="", main="Amount Spent", col = "#69b3a2")
hist((raw.data$Salary), xlab="", ylab="", main="Salary", col = "#A9A9A9")
#=========================================================================correl
ation
  retail.df <- raw.data[!is.na(raw.data$AmountSpent),]
clustering.df <- cor.data #make sure you have the current correlation_script
before you run this line
dim(clustering.df)[2] # make sure that you get 9 after running this line


# Choosing optimal number of clusters ----------------------------
#first before we run k-Means, let's decide how many clusters we want to generate
#we can do it in many various ways, I will start with the elbow method:
```

```r
#let's decide the maximum K to cluster. Say 10:
k.max <- 10

#we will create a vector of the total within sum of squars, in order to visulize
it
wss <- sapply(1:k.max, function(k){kmeans(clustering.df, k, nstart=50,iter.max =
1000 )$tot.withinss})

options("scipen"=999)
ggplot()+ aes(x = 1:k.max, y = wss) + geom_point() + geom_line()+
  labs(x = "Number of clusters K", y = "Total within-clusters sum of squares")+
  scale_x_continuous(breaks = seq(0,10,1))+
  ggtitle("The Elbow Method")

#We can use the built in function persented to us in class, fviz_nbclust:
fviz_nbclust(clustering.df, FUN = kmeans,method = "wss" ,nstart = 50)


#When looking at the Elbow Method, one cannot tell for sure what's the optimal
# number of clusters K. could be either 3 or 4
#(some would say only 2), therefore we shall look into the silhouette score
#using the built-in function Optimal_Clusters_KMeans:


#silhouette method

opt.k.sil<- Optimal_Clusters_KMeans(clustering.df, max_clusters=10,
plot_clusters=TRUE,
                                    criterion="silhouette")
#both 2 and 4 number of clusters generated a high silhouette score of 5.9
#combining that with the WSS output we can conclude that the optimal number of
clusters would be 4.

#Calinski-Harabasz index

#the final nail in the coffin would be Calinski-Harabasz index between 2 and 4
clusters
km_2k <- kmeans(clustering.df, 2)
km_4k <- kmeans(clustering.df, 4)

b=round(calinhara(clustering.df,km_2k$cluster),digits=1)
c=round(calinhara(clustering.df,km_4k$cluster),digits=1)
#It is obvious now that 4 clusters would be best and we can move on

# Custering ---------------------------------------------------
#We can start our clustering
retail.df$History <- NULL
retail.df <- raw.data[!is.na(raw.data$AmountSpent),]

KMC <- kmeans(clustering.df,centers = 4,iter.max = 999, nstart=50)

retail.clustered <- (cbind(retail.df, cluster= KMC$cluster))
# Create new DF, # consisted with the original DF
# with the cluster number for each observation

table_of_cluster_distribution <- table(retail.clustered$cluster) # the result:
# 1   2   3   4
# 157 285 283 269

barplot(table_of_cluster_distribution, xlab="Clusters",
        ylab="# of customers", main="# of customers in each cluster",
        col="#69b3a2")
```

```r
retail.clustered <- data.table(retail.clustered)
retail.clustered[, avg_AmountSpent_in_cluster :=
mean(AmountSpent),by=list(cluster)]
retail.clustered[, avg_SalarySpent_in_cluster := mean(Salary),by=list(cluster)]

retail.clustered  <-  retail.clustered[, c("Age", "Gender", "OwnHome",
"Married",
                                           "Location", "Children", "Catalogs",
"Salary","AmountSpent",
                                           "avg_AmountSpent_in_cluster",
"avg_SalarySpent_in_cluster", "cluster" )]

cluster_1 <- retail.clustered[retail.clustered$cluster==1,]
cluster_2 <- retail.clustered[retail.clustered$cluster==2,]
cluster_3 <- retail.clustered[retail.clustered$cluster==3,]
cluster_4 <- retail.clustered[retail.clustered$cluster==4,]

#View(cluster_1)
lapply(retail.clustered[,1:7],table)

data.with.clustering <- cbind(clustering.df, retail.clustered)
#View(data.with.clustering)


clustering.df.n<-scale((clustering.df))

scale((  range = c(0, 1)))



#View(clustering.df.n)
library(smacof)

dis<-dist(clustering.df.n) # dissimilarity matrix
dis2<-sim2diss(dis, method=1, to.dist = TRUE)

#mantel.test(as.matrix(dis), as.matrix(dis2))

mds_fit <-cmdscale(dis,eig=TRUE,k=5) # We use the type of "ordinal",
#since most of the vars in these Dataset are categorical, redcued 5 vars
```

DATA SHEET:

retailMarketingDI.csv