Capstone_report

Marta Tenconi

17 June 2019

1. Introduction

This project is about the customer segmentation of a mall. The data can be obtained from kaggle datasets: (https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python (https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python)) The original data set was created only for learning purpose.

The context:

The data set provides membership cards information of a mall. Basic data about the customers are provided as: Customer ID, age, gender, annual income, spending score.

The datasets:

The file “Mall_Customers.csv” is here opened in a data frame named “data”. It has 200 observations of 5 variables. Each line of the file refers to one customer, and describes the following attributes: - CustomerID = Unique ID assigned to the customer,
- Gender = Gender of the customer,
- Age = Age of the customer,
- Annual Income (k$) = Annual Income of the customers,
- Spending Score (1-100) = Score assigned by the mall based on customer behavior and spending nature.

Aim of the project:

To understand the behavior of the customers of a mall and to divide them into distinct segments, accordingly. Different segments have different habits and needs, therefore, after the segmentation process marketers will be able to develop different strategies to target each specific segment. In this way, the mall will be able to focus on particular target groups so as to provide the best experience for them.

Key steps:

First step: download of the data set and download of the required libraries.
Second step: some data exploration and data visualization.
Third step: analysis of the data: clustering using principal component analysis and K-means.
Fourth step: interpretation of the data.
Final step: conclusions.

2. Data set and libraries download

Download of the dataset from the website (https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python (https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python)) and import the file in R.

Download of the libraries needed for the project.

3. Exploration of the data and data visualization

‘data’ set details:

‘data’ is an object of class data frame composed by 5 variables:
- CustomerID: Unique ID assigned to each customer,
- Gender: Gender of the customers,
- Age: Age of the customers,
- Annual Income (k$): Annual Income of the customers, - Spending Score (1-100): Score assigned by the mall based on customer behavior and spending nature.

Class of ‘data’ data set:

```
## [1] "data.frame"
```

Structure of ‘data’ data set:

```
## 'data.frame':    200 obs. of  5 variables:
##  $ CustomerID            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Gender                : Factor w/ 2 levels "Female","Male": 2 2 1 1 1 1 1 1 2 1 ...
##  $ Age                   : int  19 21 20 23 31 22 35 23 64 30 ...
##  $ Annual.Income..k..    : int  15 15 16 16 17 17 18 18 19 19 ...
##  $ Spending.Score..1.100.: int  39 81 6 77 40 76 6 94 3 72 ...
```

Each row of data represents one costumer identified by the CustomerID, no missing values are present:

```
##   CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100.
## 1          1   Male  19                 15                      39
## 2          2   Male  21                 15                      81
## 3          3 Female  20                 16                       6
## 4          4 Female  23                 16                      77
## 5          5 Female  31                 17                      40
## 6          6 Female  22                 17                      76
```

Summary of the data:

Gender	Age	Annual.Income..k..	Spending.Score..1.100.
Female:112	Min. :18.00	Min. : 15.00	Min. : 1.00
Male : 88	1st Qu.:28.75	1st Qu.: 41.50	1st Qu.:34.75
NA	Median :36.00	Median : 61.50	Median :50.00
NA	Mean :38.85	Mean : 60.56	Mean :50.20
NA	3rd Qu.:49.00	3rd Qu.: 78.00	3rd Qu.:73.00
NA	Max. :70.00	Max. :137.00	Max. :99.00

Average and standard deviation of the data:

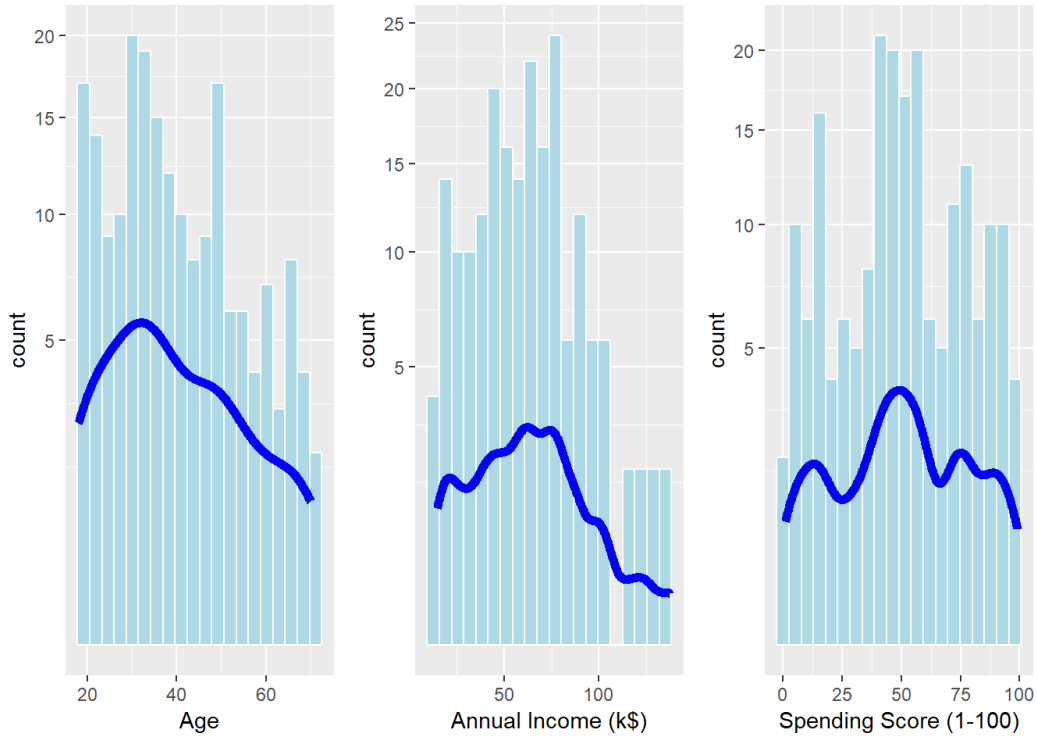
avg_Age	sd_Age	avg_Annual.Income	sd_Annual.Income	avg_Spending.Score	sd_Spending.Score
38.85	13.97	60.56	26.26	50.2	25.82

Customers age, annual income and spending score have high standard deviations.

Data Visualization

A. Histograms

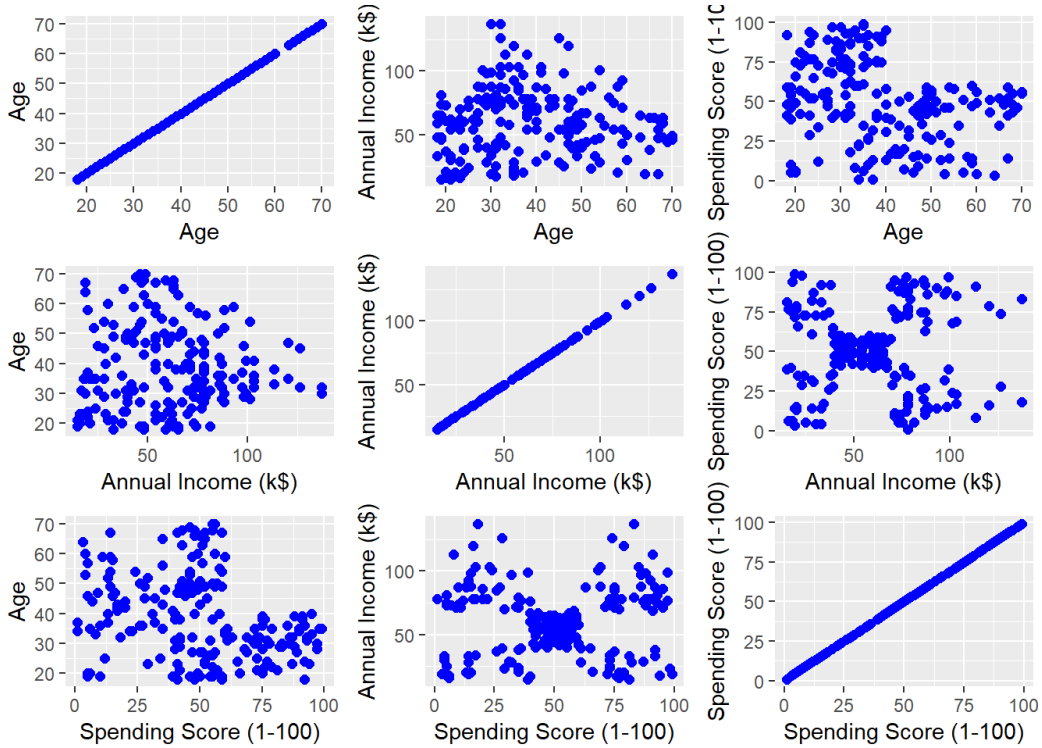
Histograms of the age, annual income and spending score, overlapped by the smooth curve:



Age, annual income and spending score do not present a normal distribution: age and annual income ares skewed righ, while spending score is symmetrical but it has too large tails to be normal.

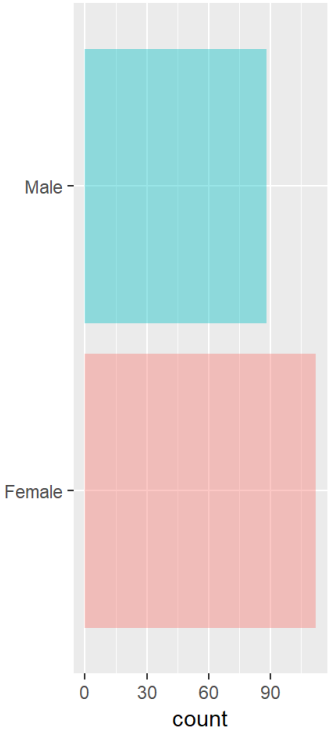
B. Relation between Age, Annual Income and Spending Score

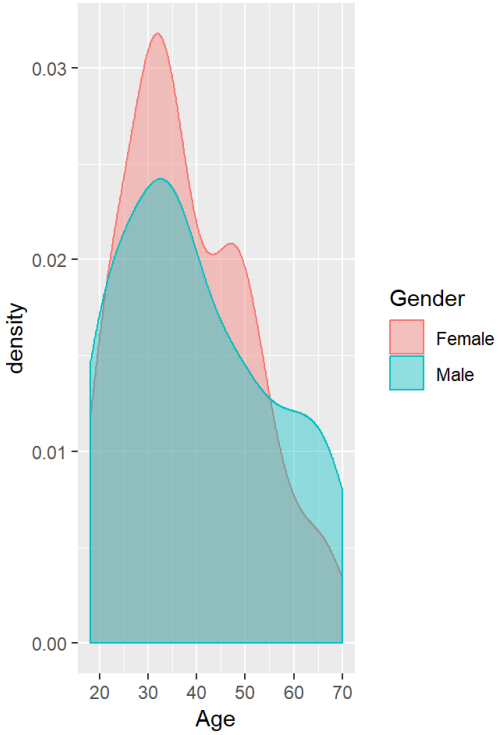
Pairplot for all columns after dropping non-informative ordinal data (CustomerID and gender)



Spending score vs annual income show interesting correlation

C. Count plots of gender and density plot of age distribution, by gender:



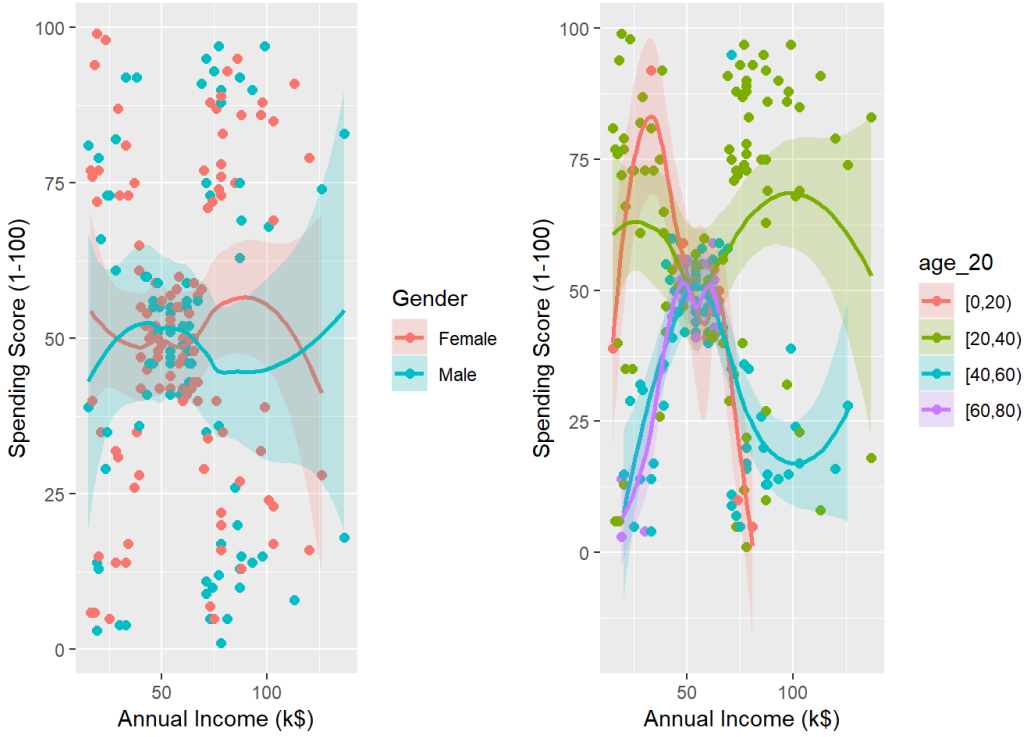


Percentage of male customers = 44%
Percentage of female customers = 56%
Gender and age are usually good indicators for distinguishing target groups. Here men are slightly more represented than women. There is a difference also in the activity of the two groups (males & females): both males and females are highly active between the ages 25-35 ca., however, women have a second pick of activity between the ages 45-50 ca. while after the age of 45, men's' activity declines.

some more investigation:

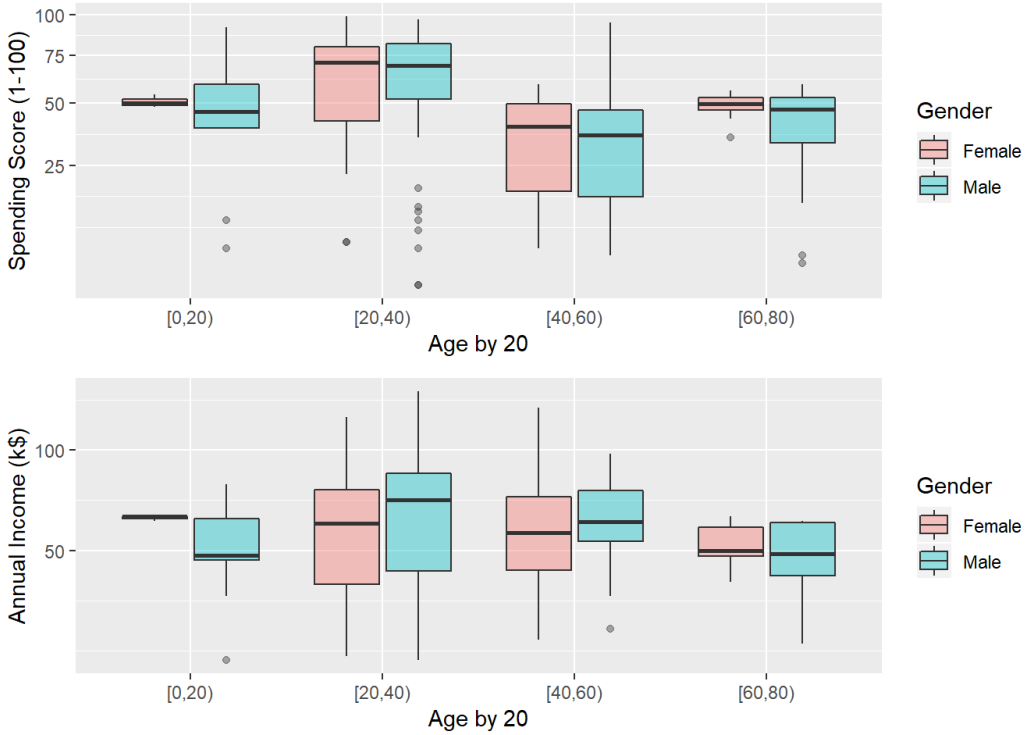
D. Scatter plot of the age vs annual income, according to gender, with indication of smooth conditional means:

E. Scatter plots of the annual income vs spending score, according to Gender and age, with indication of smooth conditional means:

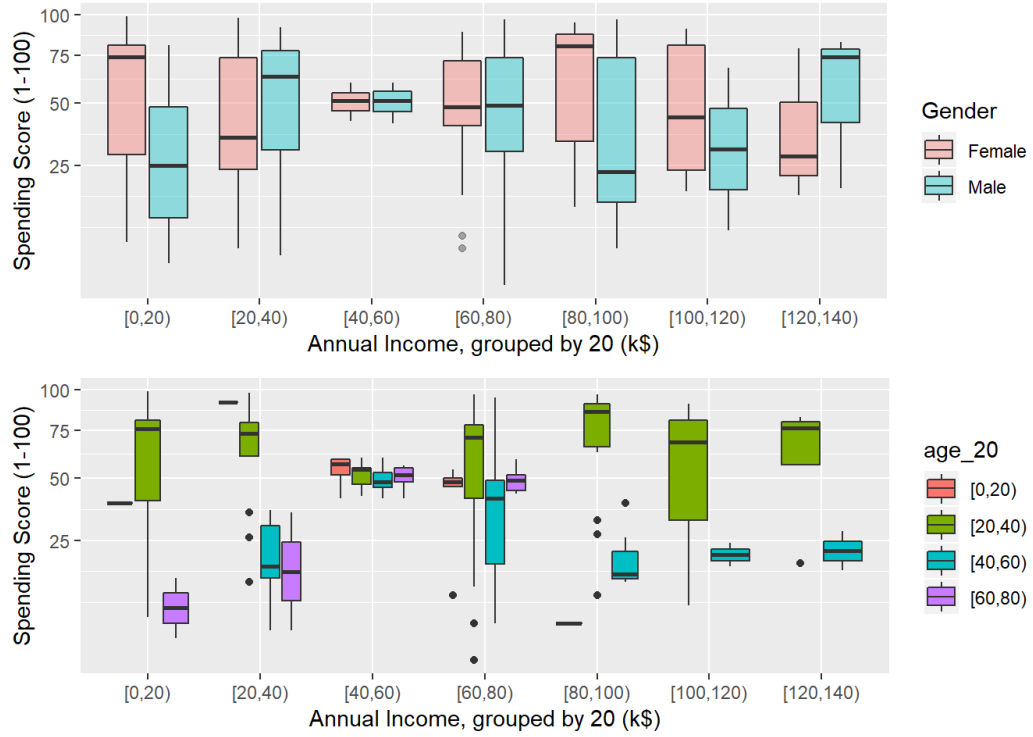


F. Distribution of the spending score vs annual income, grouped by gender and age:

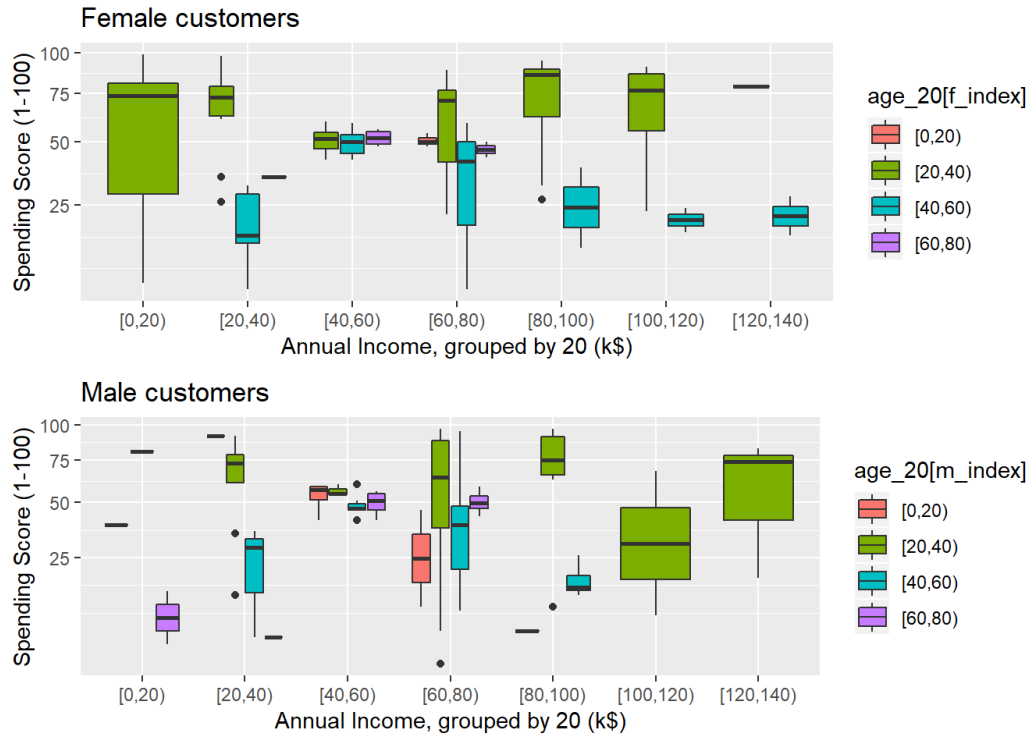
F1. Boxplots of spending score and boxplots of annual income stratified by age, grouped by gender:



F2. Boxplots of spending score stratified by annual income, grouped by age and gender:



F3. Box plots of spending score stratified by annual income, grouped by age; for female and male selectively:



Percentage of the customers representing each strata of annual income: - Annual income = 0-20k $: 6%
- Annual income = 20-40k $: 17% - Annual income = 40-60k $: 23% - Annual income = 60-80k $: 35% - Annual income = 80-100k $: 12% - Annual income = 100-120k $: 4%
- Annual income = 120-140k $: 3%

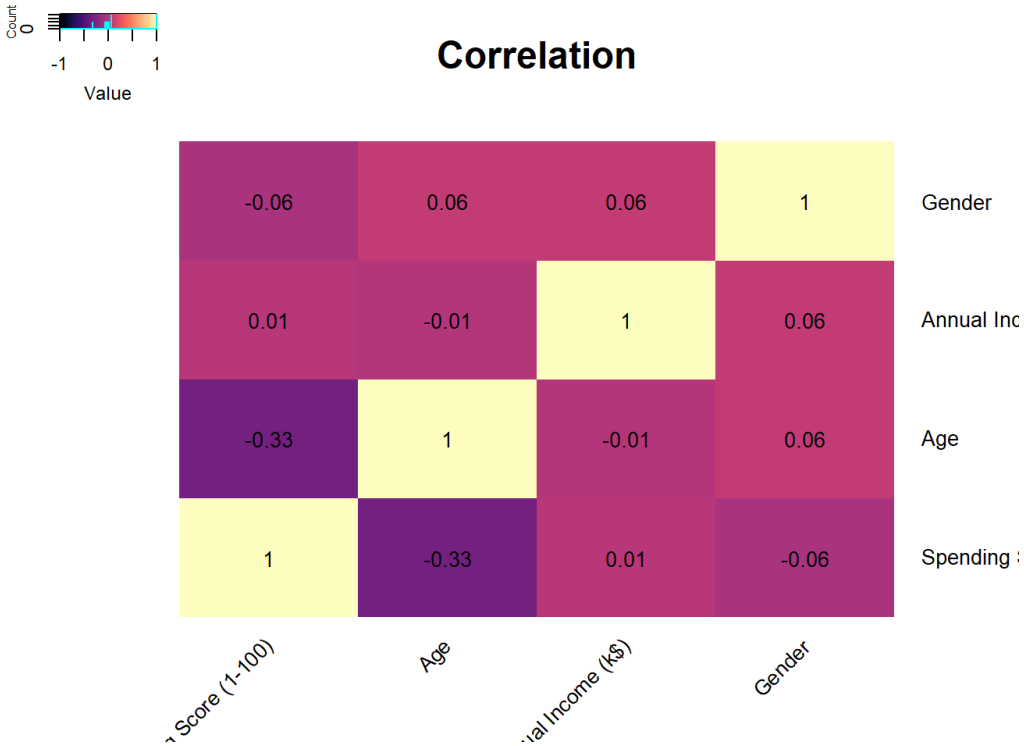
The two groups (males and female) seem to have a similar trend and there are not great anomalies, with exception of a few outliers in the males' spending score distribution (image F1). However, some interesting observations emerge when performing the boxplots of spending score stratified by annual income, and grouped by gender and age: in general some important differences between male and female behavior are evident, with exception for customers with an annual income between 40-80k; while customers younger than 40 years old usually have higher spending score, with exception of customers with an annual income between 40-60k, representing the 23% of the entire customers.

F. Heatmap

Correlation between the different attributes (gender, annual income, age, spending score) of the data:

```
##                        Gender   Age Annual.Income..k..
## Gender                   1.00  0.06               0.06
## Age                      0.06  1.00              -0.01
## Annual.Income..k..       0.06 -0.01               1.00
## Spending.Score..1.100.  -0.06 -0.33               0.01
##                        Spending.Score..1.100.
## Gender                                  -0.06
## Age                                     -0.33
## Annual.Income..k..                       0.01
## Spending.Score..1.100.                   1.00
```

Graph Showing the correlation between the different attributes (gender, annual income, age, spending score) of the data, The map reflects the most correlated features with yellow and least correlated features with viridis 'magma' palette.



The attributes do not have good correlation among them

4. Analysis: clustering

Clustering is a Machine Learning technique for grouping and finding patterns in the data. In clustering not labelled data are given to unsupervised algorithm, meaning that only the input variables(X) are given with no corresponding output variables. Here, I will use principal component analysis and K-means clustering.

4.1. Principal Component Analysis (PCA):

PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

PCA works best with numerical data, therefore, the variable gender is transformed in a numerical variable. Moreover, data needs to be standardized to make variables comparable

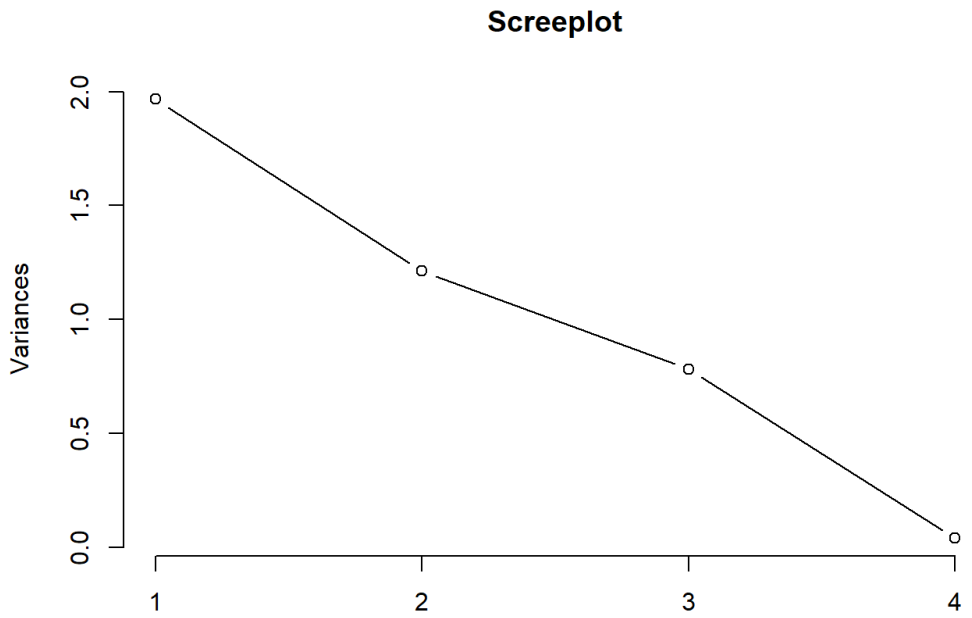
Pca performed using the prcomp() function. The variable CustomerID is excluded from the analysis

```
pca <- prcomp(log_nd[,-1], center = TRUE, scale = TRUE)
```

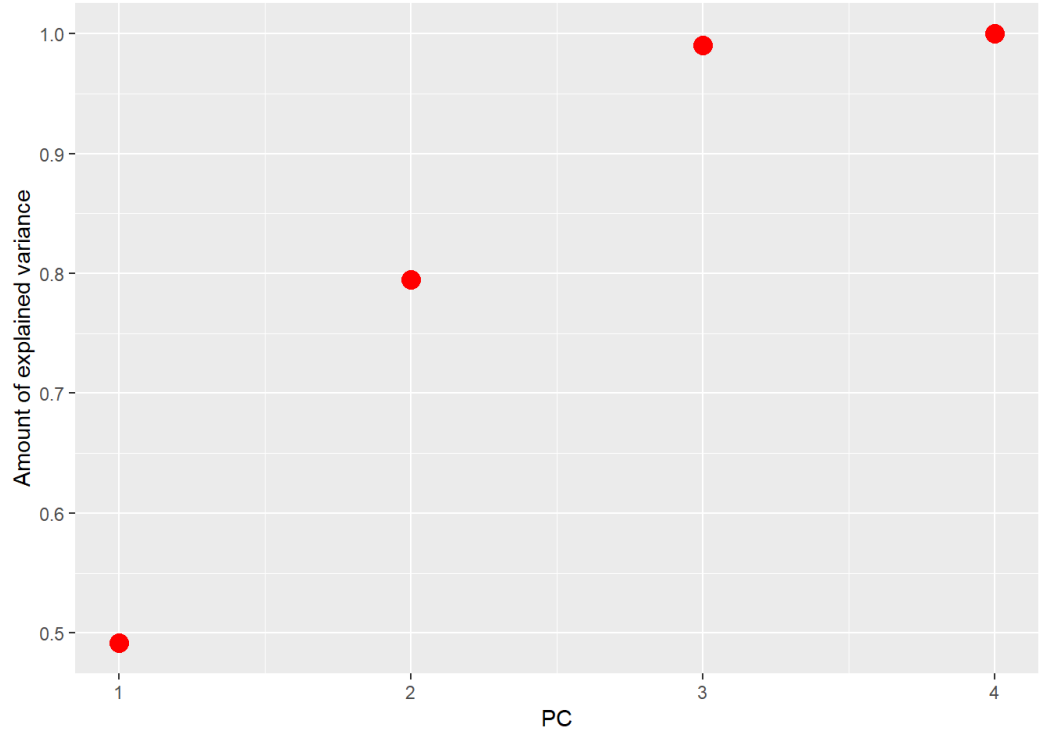
Analysis of the results:

```
## Standard deviations (1, .., p=4):
## [1] 1.4021770 1.1009913 0.8845651 0.1981472
##
## Rotation (n x k) = (4 x 4):
##                                PC1         PC2         PC3         PC4
## Age                     0.10068475  0.69584997 -0.70869571 -0.05835914
## Annual.Income..k..      0.70410578 -0.06588199 -0.02284691  0.70666303
## Spending.Score..1.100. -0.04027065 -0.71126275 -0.70007158 -0.04881971
## CustomerID              0.70176628 -0.07454990  0.08442862 -0.70344737
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4
## Standard deviation     1.4022 1.1010 0.8846 0.19815
## Proportion of Variance 0.4915 0.3030 0.1956 0.00982
## Cumulative Proportion  0.4915 0.7946 0.9902 1.00000
```



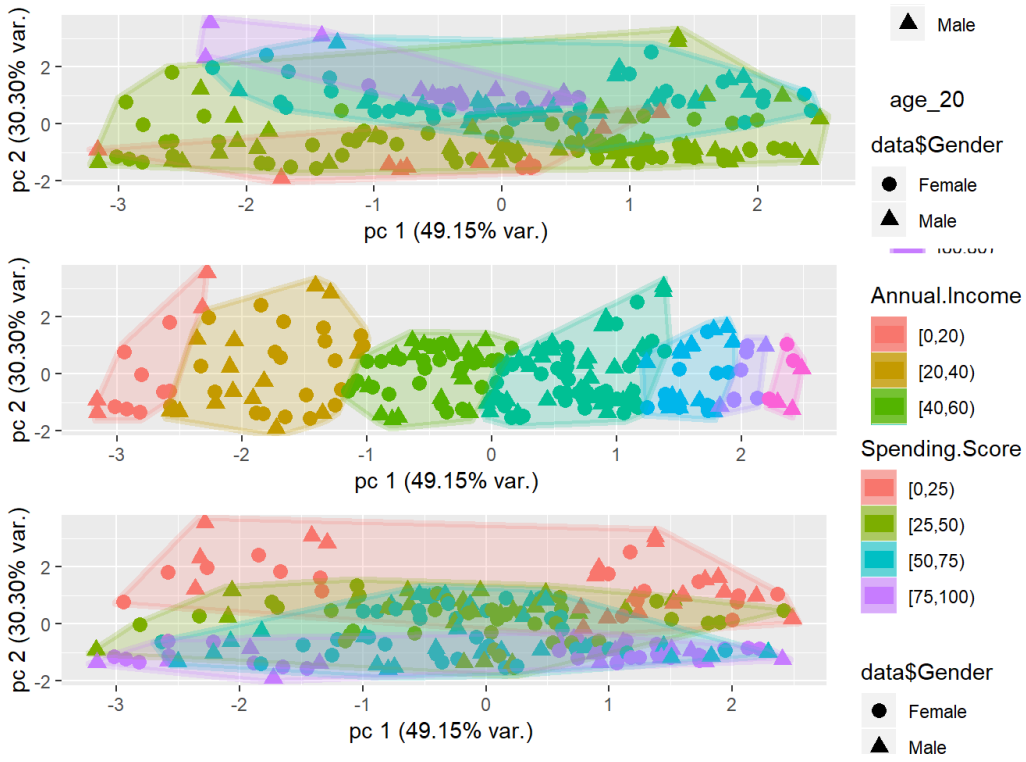
I obtain 4 principal components (pc1-pc4). Each of these explains a percentage of the total variation in the dataset:



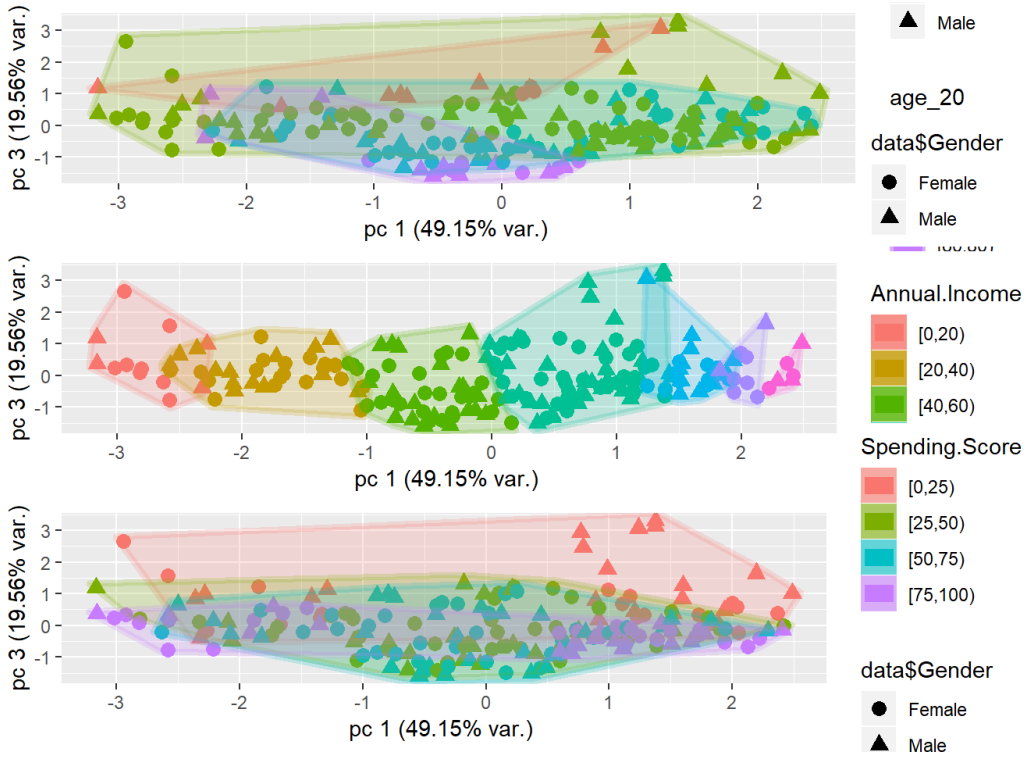
A new data set with pca results (pc1, pc2, pc3, pc4), age, annual income, spending score, gender) is needed for the following analysis.

```
##         pc1         pc2       pc3        pc4  age_20 Annual.Income
## 1 -3.164557 -0.91694575 1.1866815 -0.4735723  [0,20)        [0,20)
## 2 -3.160301 -1.35233793 0.3808946 -0.5444565 [20,40)        [0,20)
## 3 -2.946213  0.76696517 2.6601555 -0.3064813 [20,40)        [0,20)
## 4 -3.019081 -1.14761479 0.2478133 -0.4903654 [20,40)        [0,20)
## 5 -2.809295 -0.03192885 0.2204466 -0.4271637 [20,40)        [0,20)
## 6 -2.922255 -1.23117290 0.3448048 -0.4223964 [20,40)        [0,20)
##   Spending.Score Gender
## 1        [25,50)   Male
## 2       [75,100)   Male
## 3         [0,25) Female
## 4       [75,100) Female
## 5        [25,50) Female
## 6       [75,100) Female
```

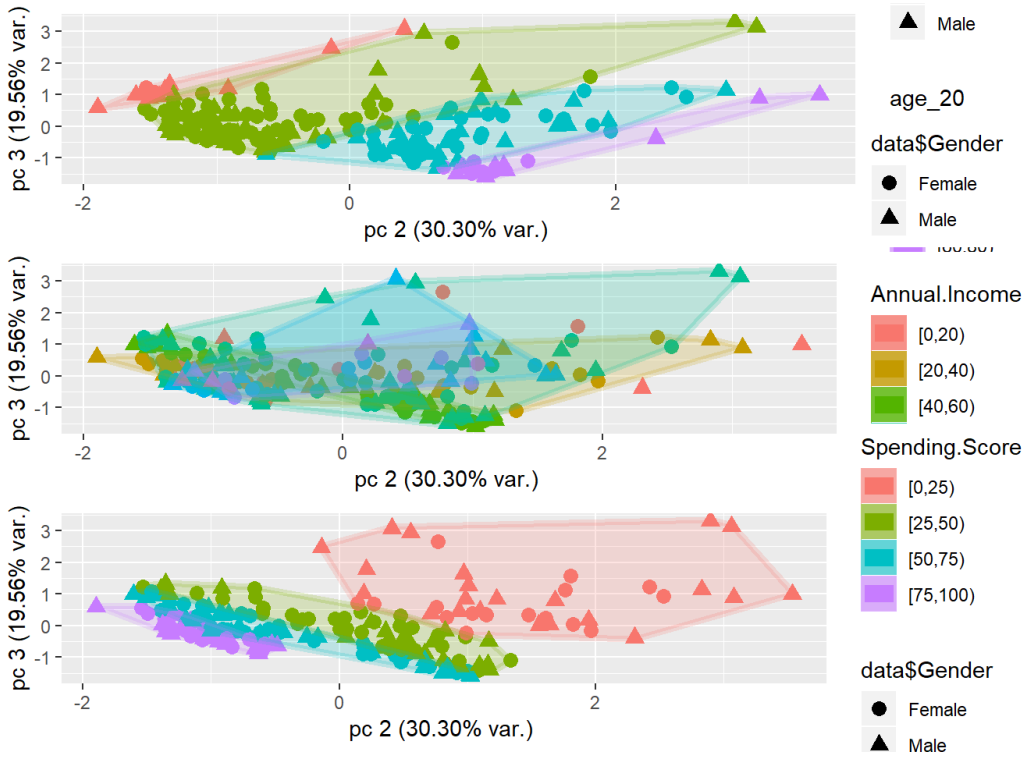
Scatter plot pf pc1 and pc2, grouped by age, annual income, and spending score, respectively:



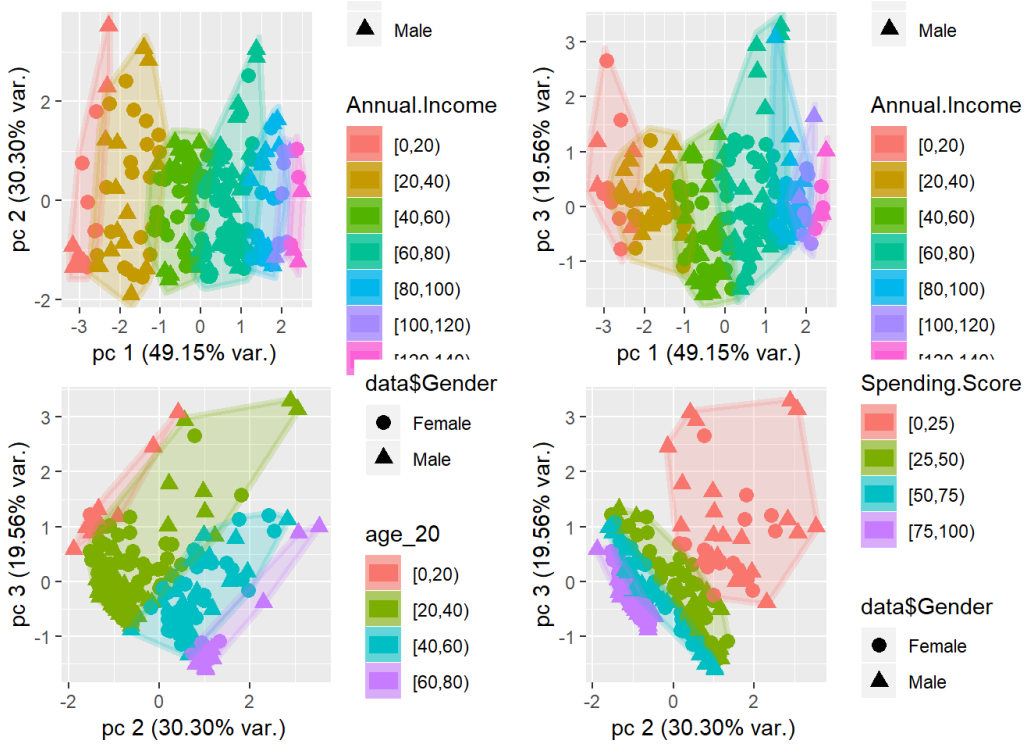
Scatter plot pf pc1 and pc3, grouped by age, annual income, and spending score, respectively



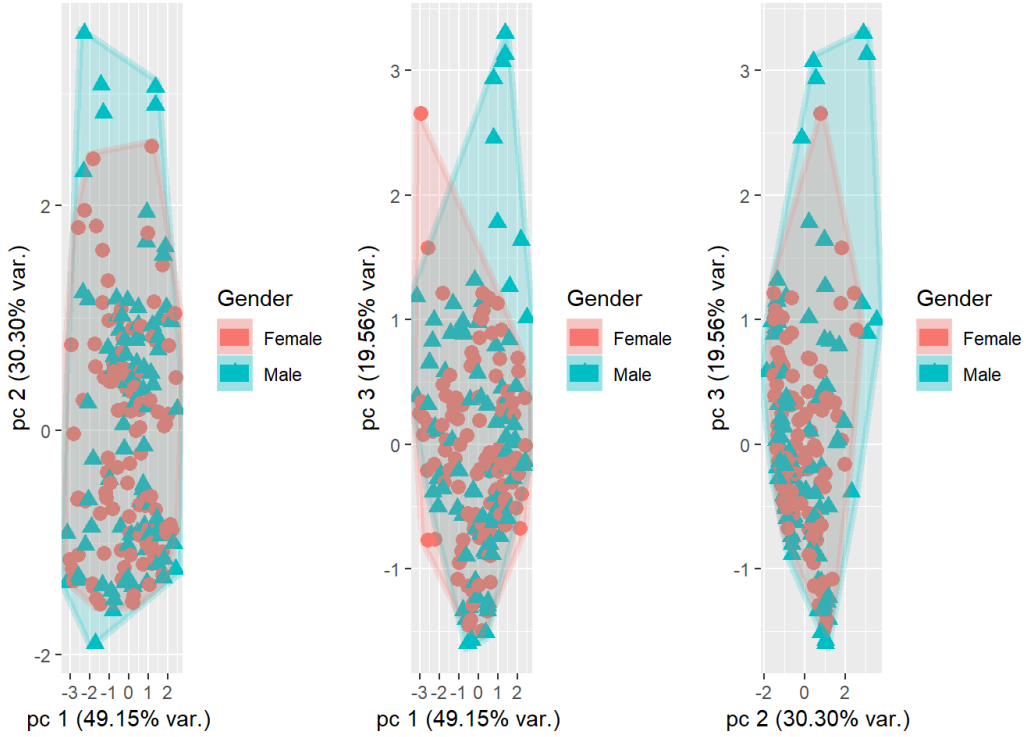
Scatter plot of pc2 and pc3, grouped by age, annual income, and spending score, respectively



Scatter plots of pc1, pc2, pc3 with the more informative clustering



Scatter plots of pc1, pc2, pc3, grouped by gender



The more informative variables seem to be the annual income and spending score. In the scater plots with gender grouping, the two groups (Male and Female) are completely overlapped, therefore they do not provide significant information.

4.2. K-means clustering:

K-means clustering is an unsupervised machine learning algorithm for partitioning a given data set into a set of k groups (clusters), where k represents the number of groups pre-specified by the analyst. It classifies objects in multiple clusters, such that objects within the same cluster are as similar as possible (high intra-class similarity), whereas objects from different clusters are as dissimilar as possible (low inter-class similarity). In k-means clustering, each cluster is represented by its center (i.e, centroid) which corresponds to the mean of points assigned to the cluster.

Main steps for k-means clustering:
1: scale the data, to make variables comparable. 2: determine the optimal number of Clusters with the ‘elbow method’: Elbow Method: compute clustering algorithm (k-means clustering) for different values of k. For each k, calculate the total within-cluster sum of square (wss). Plot the curve of wss according to the number of clusters k. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.
3: perform k-means function, using the chosen k.

The data have to be scaled, to make variables comparable.

According to PCA results, annual income and spending score are the two more informative variable, therefore, I perform k-means clustering using these two variables.

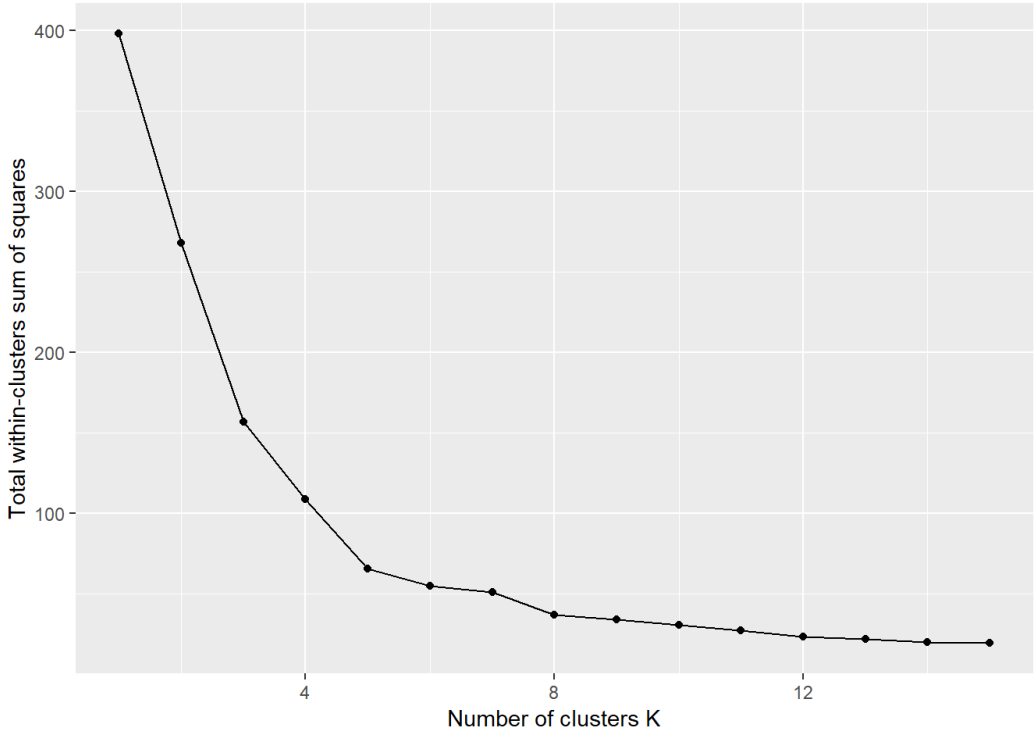
Function to compute total within-cluster sum of square (wss).

```
data2 <- data_scale %>% select(Annual.Income..k.., Spending.Score..1.100.)
wss <- function(k) {                          # within-cluster sum of square (wss)
  kmeans(data2, k, nstart = 10 )$tot.withinss
}
```

I compute and plot wss for k = 1 to k = 15 and extract wss for 2-15 clusters.

```
k.values <- 1:15
wss_values <- map_dbl(k.values, wss)
```

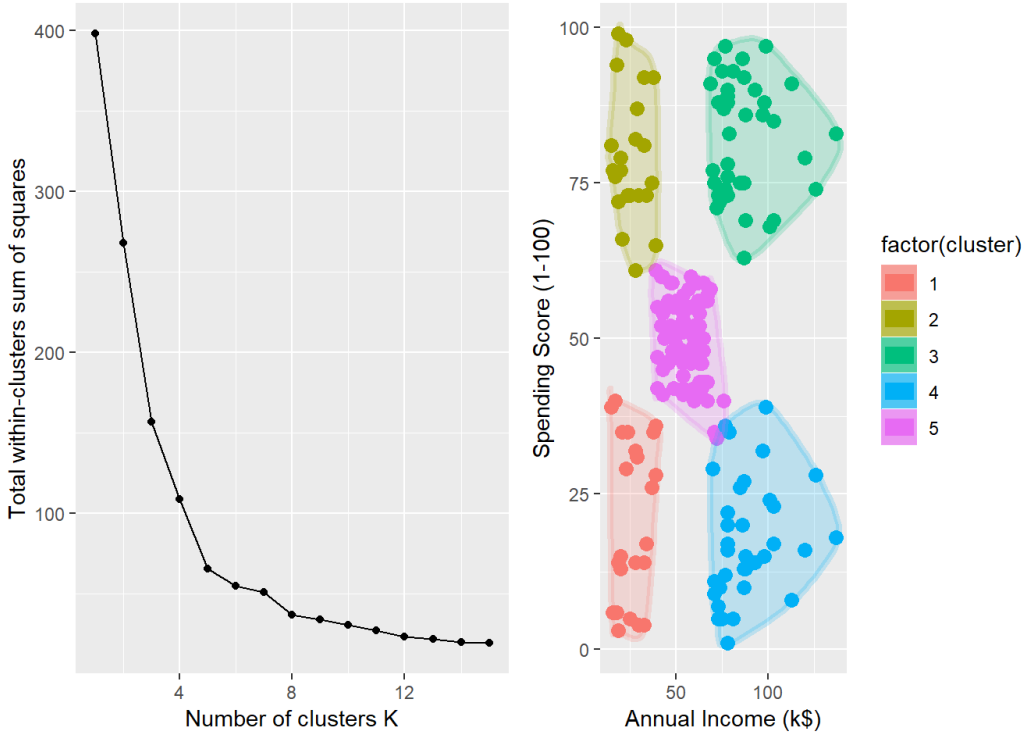
From the plot of the curve of wss according to the number of clusters k; I choose a k = 5 clusters (elbow method)



I compute k-means clustering with k = 5

```
set.seed(123)
final2 <- kmeans(data2, 5, nstart = 25)
new_data2 <- data %>% mutate(cluster = final2$cluster)
```

Curve of wss according to the number of clusters k and annual income vs spending score scatter plot showing the distribution of the 5 clusters:



Descriptive statistics at the cluster level:

```
## # A tibble: 5 x 7
##   Cluster Age_avg Annual.Income..~ Spending.Score.~ Age_sd Annual.Income..~
##     <int>   <dbl>            <dbl>            <dbl>  <dbl>            <dbl>
## 1       1    45.2             26.3             20.9  13.2              7.89
## 2       2    25.3             25.7             79.4   5.26             7.57
## 3       3    32.7             86.5             82.1   3.73            16.3
## 4       4    41.1             88.2             17.1  11.3             16.4
## 5       5    42.7             55.3             49.5  16.4              8.99
## # ... with 1 more variable: Spending.Score..1.100._sd <dbl>
```

Descriptive statistics at the cluster and gender level:

```
## # A tibble: 10 x 8
## # Groups:   Cluster [?]
##    Cluster Gender Age_avg Annual.Income..~ Spending.Score.~ Age_sd
##      <int> <fct>    <dbl>            <dbl>            <dbl>  <dbl>
##  1       1 Female    43.2             27.4             21.7  11.7
##  2       1 Male      48.3             24.7             19.7  15.5
##  3       2 Female    25.5             25.7             80.5   5.22
##  4       2 Male      25               25.8             77.7   5.61
##  5       3 Female    32.2             86.0             81.7   3.08
##  6       3 Male      33.3             87.1             82.7   4.39
##  7       4 Female    43.2             90.9             22.1   9.16
##  8       4 Male      39.3             85.9             12.9  12.9
##  9       5 Female    40.9             55.4             49    14.7
## 10       5 Male      45.4             55.2             50.3  18.6
## # ... with 2 more variables: Annual.Income..k.._sd <dbl>,
## #   Spending.Score..1.100._sd <dbl>
```

5. Interpretation of the data

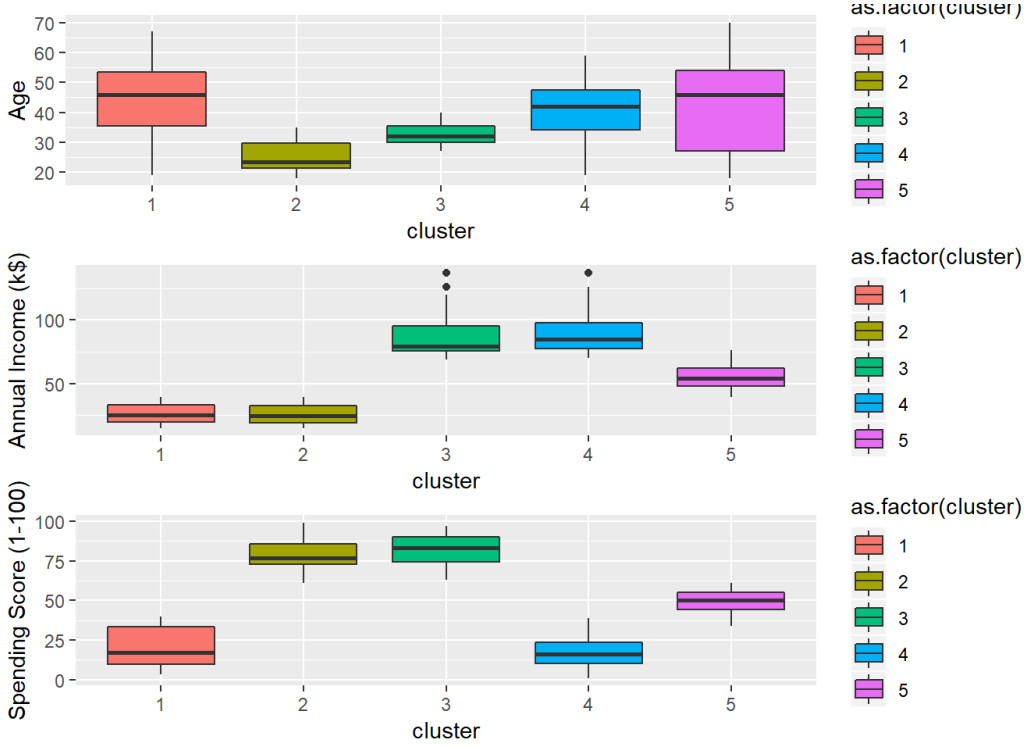
From the PCA and k-means I obtained five clusters with the following characteristics:

Cluster	Age_avg	Annual.Income..k.._avg	Spending.Score..1.100._avg	Age_sd	Annual.Income..k.._sd	Spending.Score..1.100._sd
1	45.21739	26.30435	20.91304	13.22861	7.893811	13.017167
2	25.27273	25.72727	79.36364	5.25703	7.566731	10.504174
3	32.69231	86.53846	82.12821	3.72865	16.312485	9.364489
4	41.11429	88.20000	17.11429	11.34168	16.399067	9.952154
5	42.71605	55.29630	49.51852	16.44782	8.988109	6.530909

Descriptive statistics at cluster level:

Cluster	Age_avg	Annual.Income..k.._avg	Spending.Score..1.100._avg
1	45.21739	26.30435	20.91304
2	25.27273	25.72727	79.36364
3	32.69231	86.53846	82.12821
4	41.11429	88.20000	17.11429
5	42.71605	55.29630	49.51852

The boxplot representation of the original data stratified by cluster shows that the groups are different and real:



Consideration on the gender factor:

Despite through some data exploration and data visualization (chapter 2) it seemed to emerge some interesting observations and important differences between male and female behavior; from the analysis of the clusters grouped by gender, it does not result any important difference and the two groups (female and male) can be treated together:

Cluster	Gender	Age_avg	Annual.Income..k.._avg	Spending.Score..1.100._avg
1	Female	43.21429	27.35714	21.71429
1	Male	48.33333	24.66667	19.66667
2	Female	25.46154	25.69231	80.53846
2	Male	25.00000	25.77778	77.66667
3	Female	32.19048	86.04762	81.66667
3	Male	33.27778	87.11111	82.66667
4	Female	43.25000	90.93750	22.06250
4	Male	39.31579	85.89474	12.94737
5	Female	40.89583	55.35417	49.00000
5	Male	45.36364	55.21212	50.27273

6. Conclusions

With the clutering analysis: principal component analysis and k-means, I reach meaningful insights about the mall customer and I segment them into five clusters: - Cluster 1: Customers with low annual income and low annual spending score,
- Cluster 2: Customers low annual income but high annual spending score,
- Cluster 3: Customers with high annual income and high annual spend, - Cluster 4: Customers with high annual income but low annual spend,
- Cluster 5: Customers with medium annual income and medium annual spending score.
Data on customers segments can help marketers to make better decisions and plan strategic strategic marketing approach targeted to the specific customers.