STATISTICS' PAPER:

Tommaso Cecchellero (matricola: 2203926)

INTRODUCTION:

My work has been divided into three parts.

In the first part, I started my analysis by computing the most common statistical values for each variable included in the "Bank_customers_data." First of all, I calculated the absolute and relative frequencies. After that, in order to create a boxplot representation for each variable in the dataset, I computed the mean, median, mode, and, finally, the quartiles (Q1 and Q3).
Additionally, in this part, I included the variability analysis to understand the spread of the data around a central value. To assess this, I computed the variance, standard deviation, and Gini Impurity Index.

In the second part of the analysis, I aimed to highlight the association between some variables by computing the Chi-Square Index (χ^2) and Pearson's Correlation Coefficient.

Last but not least, in the third part, I performed a Cluster Analysis, using both methods we covered during the lectures: k-Means and Hierarchical clustering. I also conducted a Principal Components Analysis (PCA).

DATA AND METHODS:

PART 1:

The functions that I used in this part are:

>> table(...) → to compute the absolute frequencies,

>> prop.table(...) → to compute the relative frequencies,

>> mean(...) → to compute the means,

>> median(...) → to compute the medians,

>> I discovered that in R there isn't a built in function to compute the mode in the terms that we learned during the past lectures (the most frequent value in a vector), but the function mode(...) returns the storage mode of an object (e.g. "numeric"). So, to compute the most frequent value, I had to search for a custom-made function (see the programme's script).

>> quantile(..., 0.25) and quantile(..., 0.75) → to compute the first quartile (Q1) and the third quartile (Q3),

>> boxplot(...) → to create the boxplot of every single variable,

>> var(...) → to compute the variances,

>> sd(..) → to compute the standard deviations,

>> to compute the Gini Impurity Index for each class of the dataset, I had to calculate the probability pi of that class. The probability is simply the proportion of elements that belong to class i. So, in this way I had all the elements to solve this equation:

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2$$

Where: G is the Gini impurity, pi is the probability of class i in the dataset and n is the number of classes. After that, I had to select a set of data (table(...)) and divide it by the length (length(...)) of the same set of data. In this way I could found the proportions of each data, so I could solve the equation "1 - sum(class_proportions_...^2)" to find the value of the Gini Impurity Index.

PART 2:

The functions that I used in this part are:

>> In order to compute the Chi-Square Index ($\chi 2$), I had to consider a couple of variables to understand their association. So, I had to create a table with the 2 specific variables and perform a chisq.test(...) of the table to find the value.
I did that for these couples of variables:
→ age + income,
→ experience + income,
→ education + income,
→ family + ccavg,
→ family + loan.

>> In order to compute the Pearson's Correlation Coefficient, I had to consider a couple of variables in order to understand their association. It was a little bit easier because I had only to choose the variable and to define the method in the cor(...) function.
I did that for these couples of variables:
→ age + income,
→ experience + income,
→ education + income,
→ family + ccavg,
→ family + loan.

PART 3:

The functions that I used in this part are:

>> In order to perform the k-means analyse on each variable, I just used the kmeans(...) built in function. To represent it graphically, I used the plot(...) function.

>> In order to perform a hierarchical cluster analyse, first of all, I had to create a distance matrix for each variable using dist(...). After that, I used the hclust(...) function, where I loaded the distance matrix and defined the linkage method to calculate the distances between the clusters. Finally, to show graphically the result, I used the function plot(...) to project the dendrogram of each specific case.

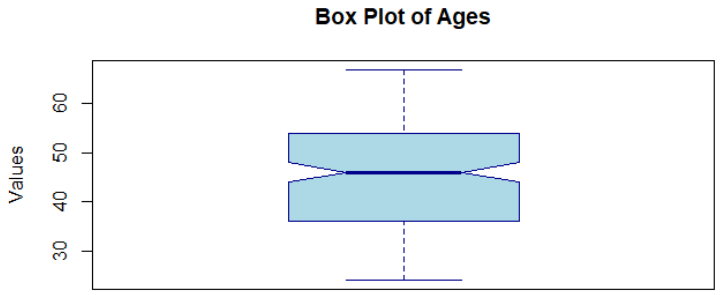
>> In order to compute the PCA, I had to create a table with alle the values of every single variable (deleting the ID column because it wasn't useful for my purpose) using data.frame(...) function. After, I had to scale all the data, using the scale(...) function, and then, using the prcomp(...) function, I could perform the analysis. Finally, to show the result, I created a biplot using the biplot(...) function.

RESULTS:

In this part I'm going to show graphically the results of my analyse:

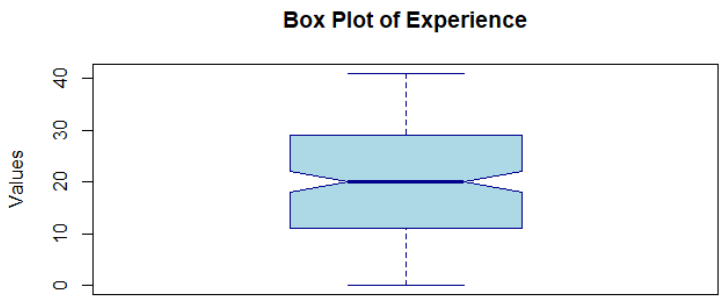
PART 1:

Here, you can see graphically the meanings of the first quartile, third quartile, and the median. I have also included a quick summary table with all the values I computed in the first part.

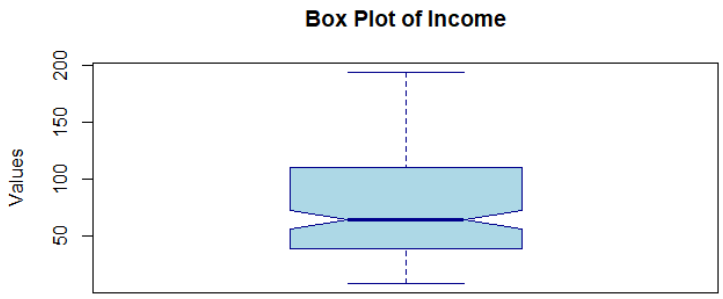


Statistic	Value
Q1	36
Median	46
Q3	54
Mean	45.015
Mode	53
Variance	128.055
St.Dev	11.316
Gini Index	0.9723

NOTE: The Bonus Note at the end of this part explains why I don't have a summary table for this variable.

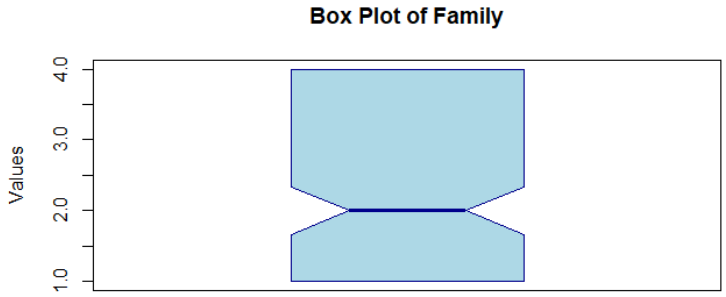


Statistic	Value
Q1	11.00000
Median	20.00000
Q3	29.00000
Mean	19.89500
Mode	23.00000
Variance	126.63716
St.Dev	11.25332
Gini Index	0.97215



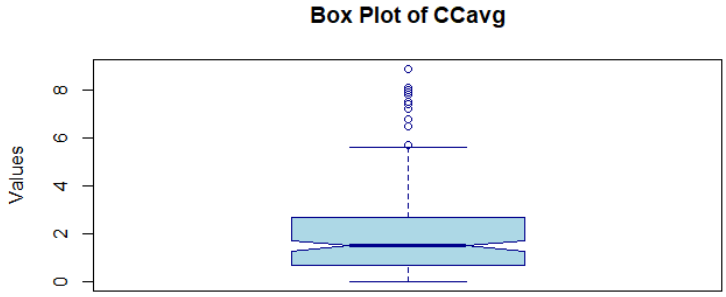
Statistic	Value
Q1	39.00000
Median	64.50000
Q3	109.75000
Mean	73.83500
Mode	45.00000
Variance	2129.67616
St.Dev	46.14841
Gini Index	0.98605

NOTE: These 3 boxplots represent the classical boxplot, they don't have any specific characteristic. In fact, we can notice that they have a quite high values in terms of variability (see Variance and St.Dev), which means that the data are spread in a relatively large range.



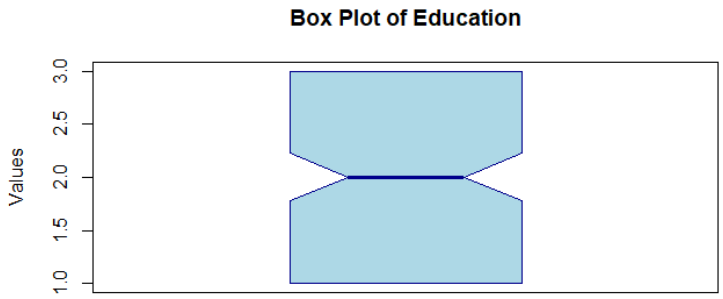
Statistic	Value
Q1	1.000000
Median	2.000000
Q3	4.000000
Mean	2.475000
Mode	1.000000
Variance	1.376256
St.Dev	1.173139
Gini Index	0.746150

NOTE: This boxplot shows that the data are very concentrated within a small range (from 1 to 4). It's hard to identify a significant outlier for this variable because, typically, families are composed of a relatively consistent number of people. The values in terms of variability are low (see Variance and St. Dev).



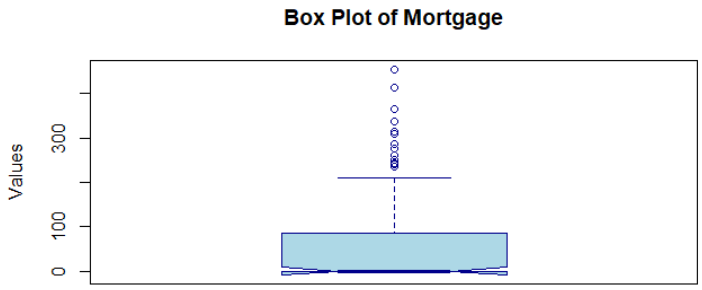
Statistic	Value
Q1	0.700000
Median	1.500000
Q3	2.670000
Mean	2.060300
Mode	1.000000
Variance	3.419156
St.Dev	1.849096
Gini Index	0.970450

NOTE: In this boxplot, we can see that there are some outliers that fall outside the distribution range. This may be because families composed of more people tend to spend more than the average. We will discuss later whether there is a connection between the variable 'family' and the variable 'CCAvg.'

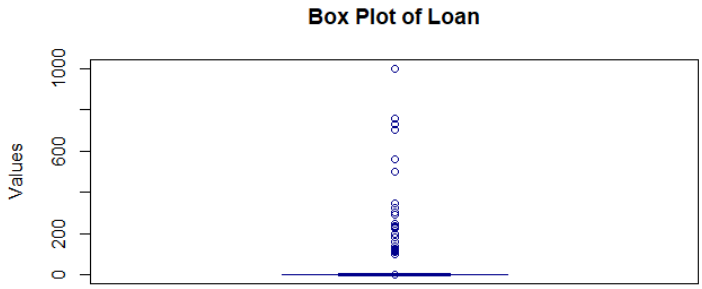


Statistic	Value
Q1	1.0000000
Median	2.0000000
Q3	3.0000000
Mean	1.8350000
Mode	1.0000000
Variance	0.7213819
St.Dev	0.8493420
Gini Index	0.6438500

NOTE: This boxplot shows that the data are very concentrated within a small range (from 1 to 3). It's hard to identify a significant outlier for this variable because, typically, the time a person spends studying is relatively consistent. In fact, the values in terms of variability are low (see Variance and St.Dev), so it's unlikely to find an outlier given the education levels represented in this dataset.



Statistic	Value
Q1	0.00000
Median	0.00000
Q3	83.00000
Mean	46.82500
Mode	0.00000
Variance	8302.06470
St.Dev	91.11567
Gini Index	0.44345



Statistic	Value
Q1	0.00000
Median	0.00000
Q3	0.00000
Mean	57.32000
Mode	0.00000
Variance	23822.17849
St.Dev	154.34435
Gini Index	0.40445

NOTE: In these last two boxplots, we can see that there are some outliers that fall outside the distribution range. We can conclude that there are more people who did not have to pay a monthly instalment related to a mortgage or a personal loan.

(*) BONUS NOTE: During my analyse, I wanted to create a table that contains all the values that I calculated with the functions for each variable, but the console reminds me this kind of error:

```
Errore in data.frame(Statistic = c("Q1", "Median", "Q3", "Mean", "Mode",  :
  argomento non specificato e senza un valore predefinito
```

I couldn't figure out the reason behind this.

PART 2:

>> Ages + Income:
Chi-Square Index = 4146
Pearson' Correlation Coefficient = -0.08579

>> Experience + Income:
Chi-Square Index = 4136.4
Pearson' Correlation Coefficient = -0.08267

>> Education + Income:
Chi-Square Index = 210.05
Pearson' Correlation Coefficient = -0.05660

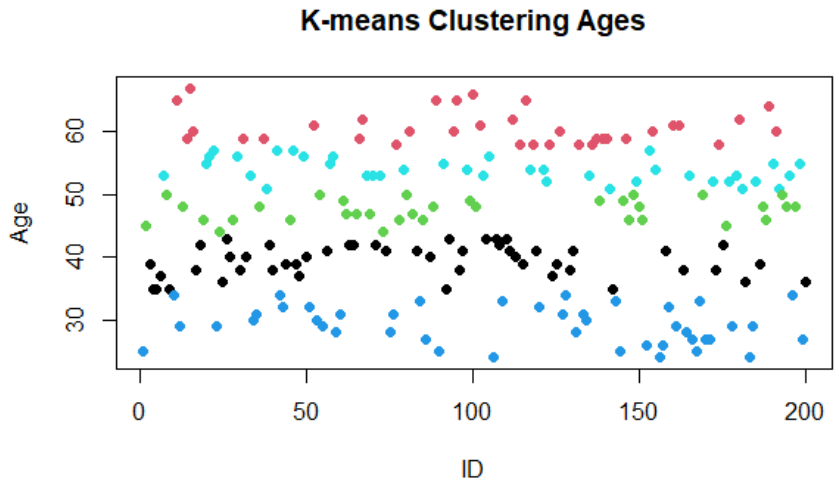
>> Family + CCavg:
Chi-Square Index = 209.25
Pearson' Correlation Coefficient = -0.18260

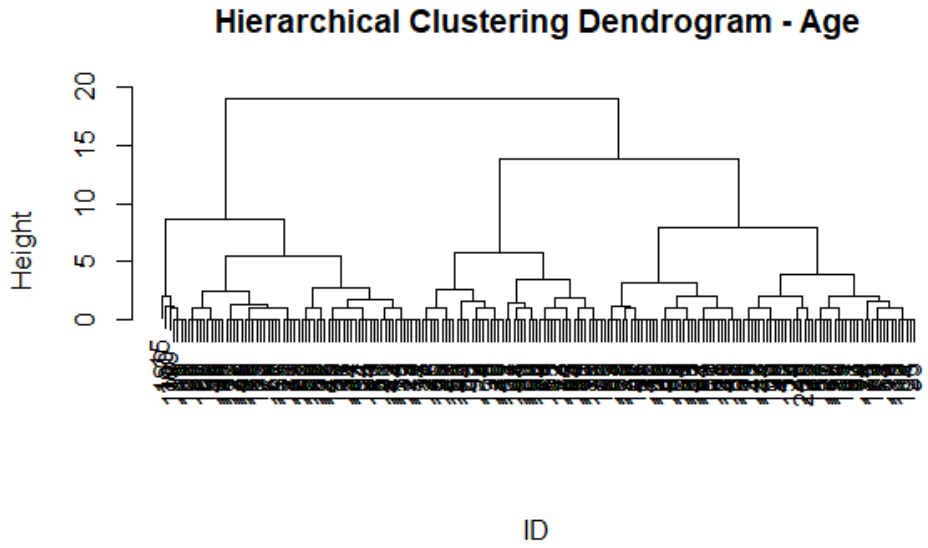
>> Family + Loan:
Chi-Square Index = 87.9
Pearson' Correlation Coefficient = 0.10611

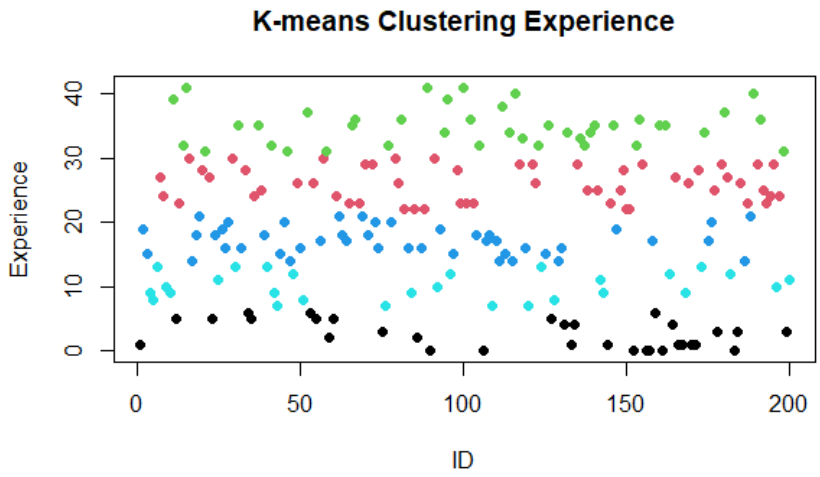
NOTE: Considering the Chi-Square Index, we can say that higher values indicate a significant difference between the observed data and the expected data, while smaller values indicate that the observed data is close to the expected data. Meanwhile, regarding the Correlation Coefficient, we can say that there are no variables that are uncorrelated. However, the results are quite surprising because I would expect a positive correlation among all the combinations that I decided to analyse. The reason behind this is that I thought, for example, that if someone is older, they would have a higher income.

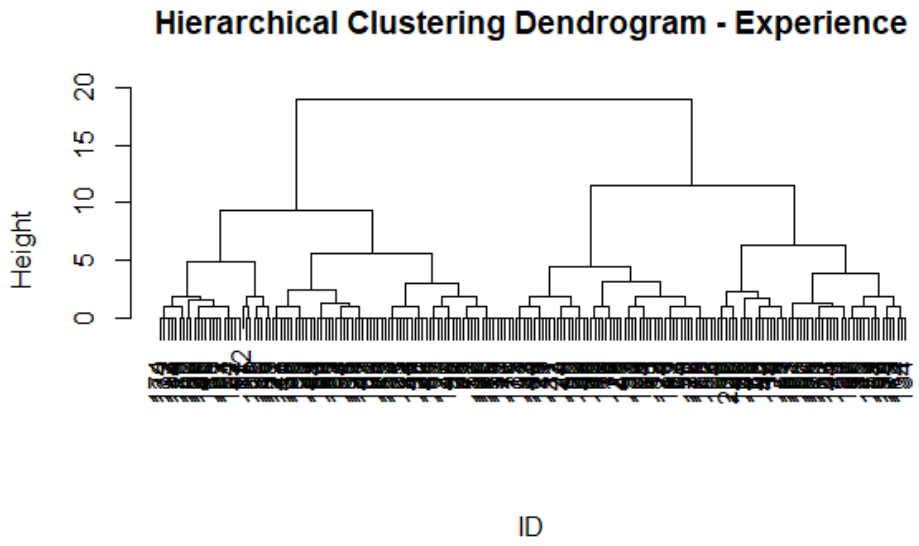
PART 3:

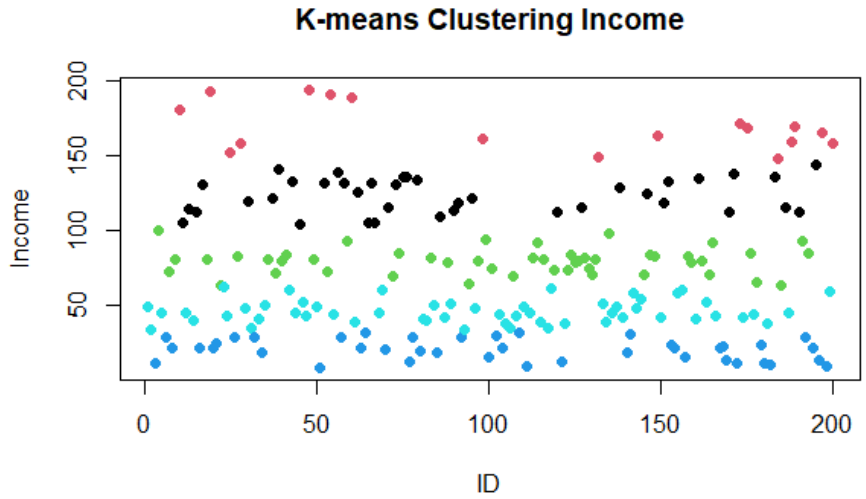
Here, we have the plots showing the results of the k-Means Analysis. The graph illustrates how the algorithm divided the data into clusters based on similarity. Additionally, we have the dendrograms displaying the results of the Hierarchical Analysis. The graph indicates that the clusters positioned at the bottom are the most similar, while the clusters located at the top of the graph are the most dissimilar.

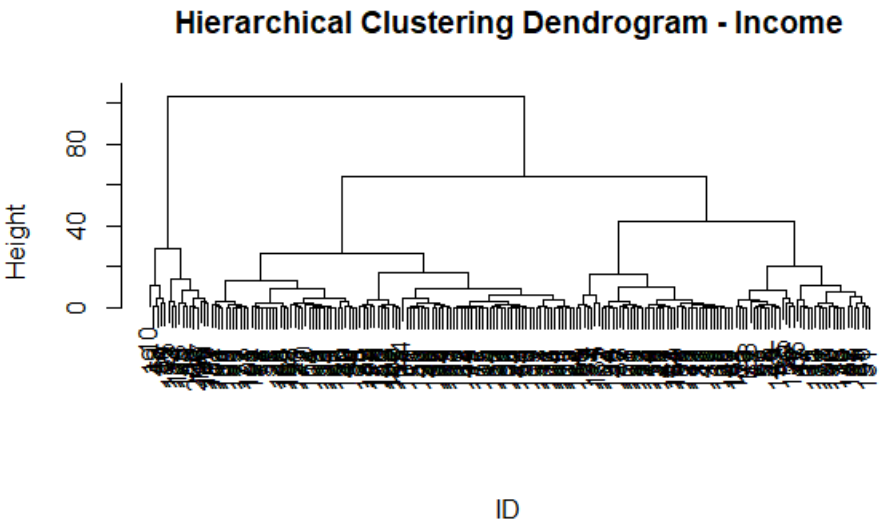


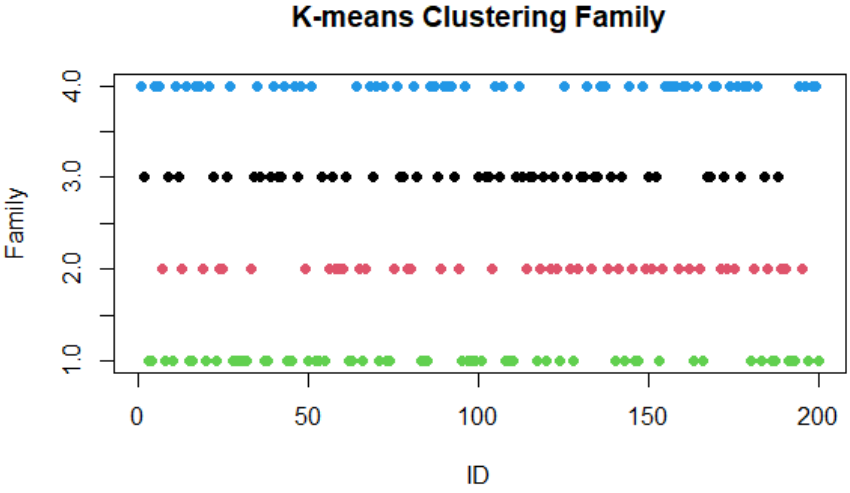



K-means Clustering Experience
Experience
0
10
20
30
40
0
50
100
150
200
ID

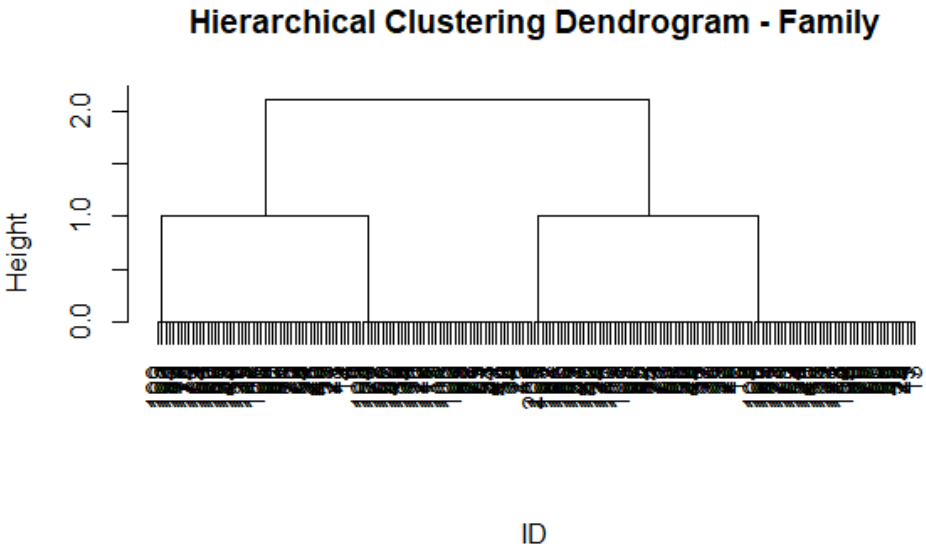

Hierarchical Clustering Dendrogram - Experience
Height
0
5
10
15
20
ID

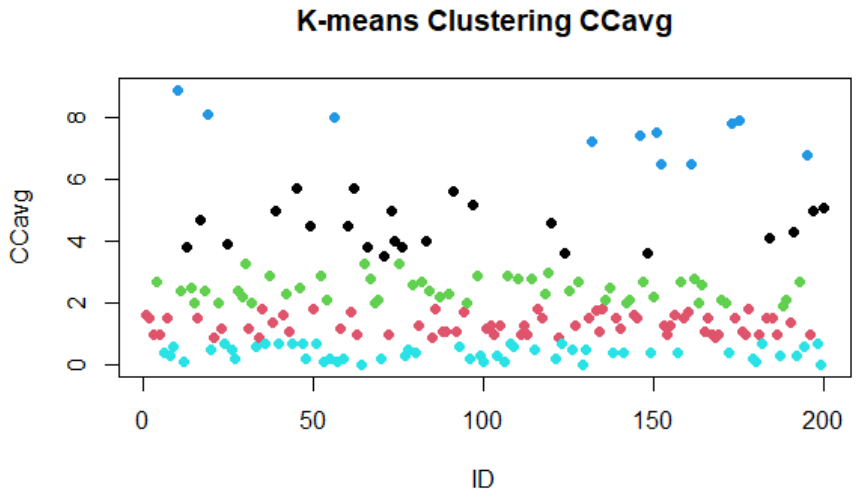

K-means Clustering Income
Income
50
100
150
200
0
50
100
150
200
ID

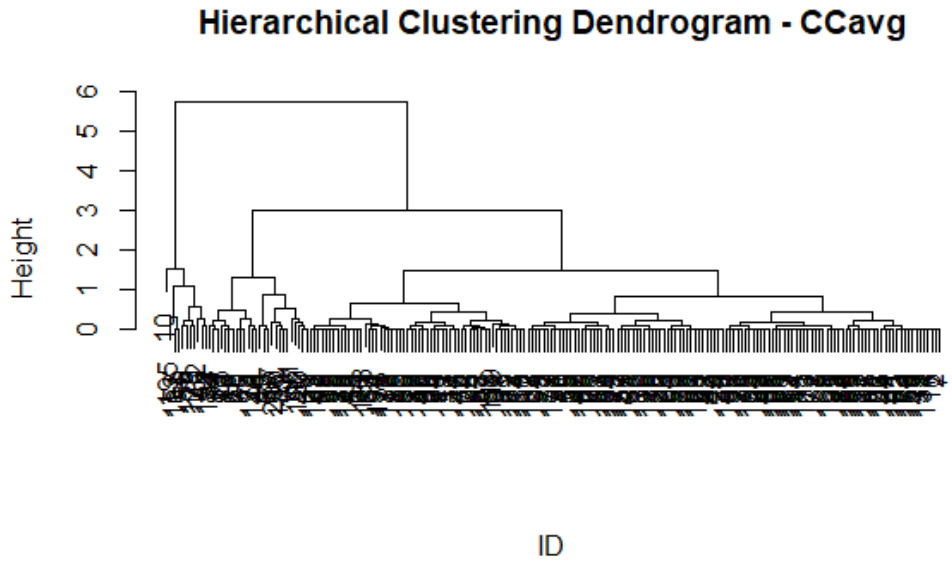


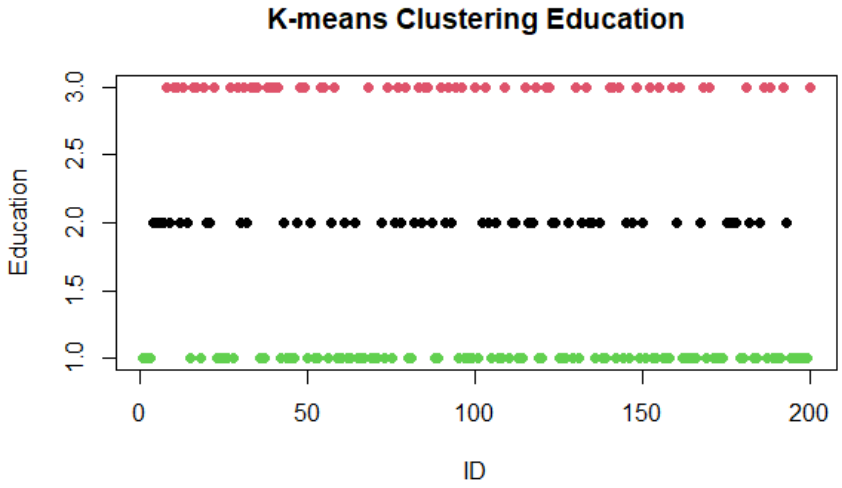


NOTE: There are only 4 clusters because there are only 4 possible values.

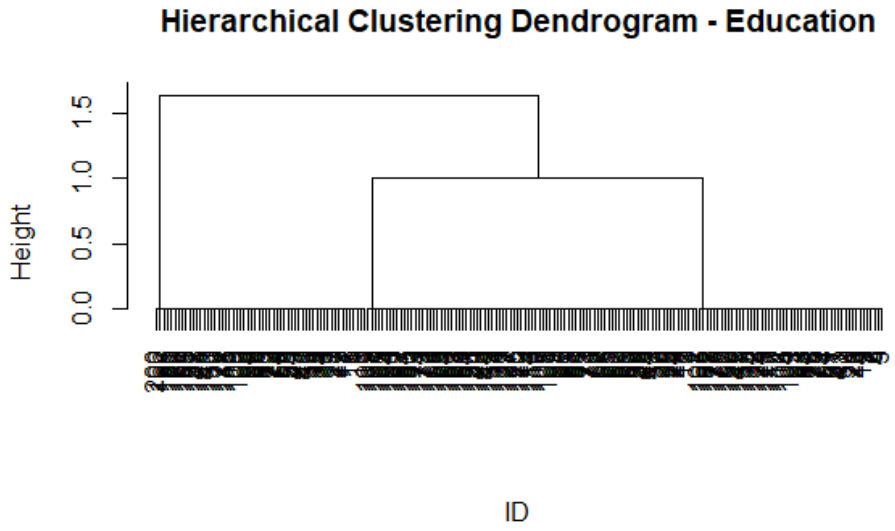


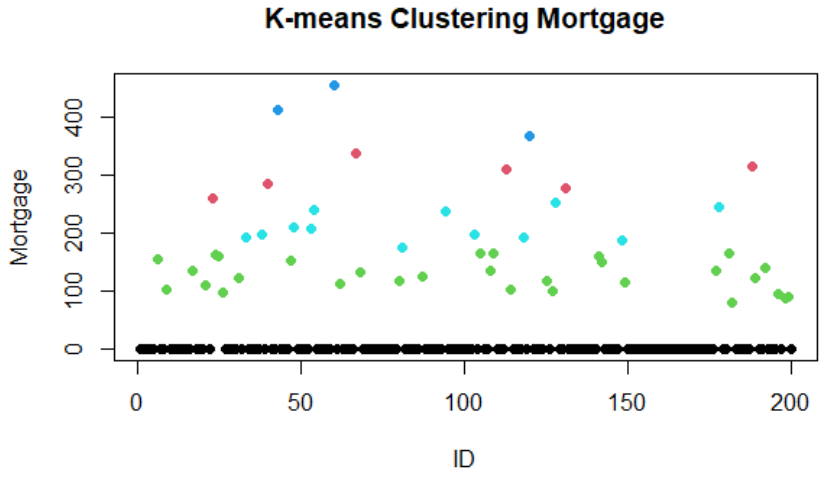


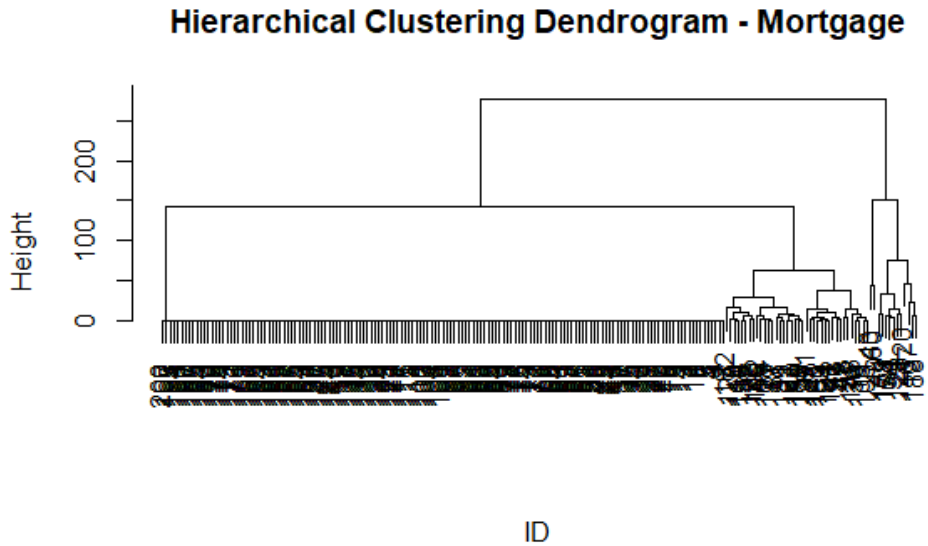


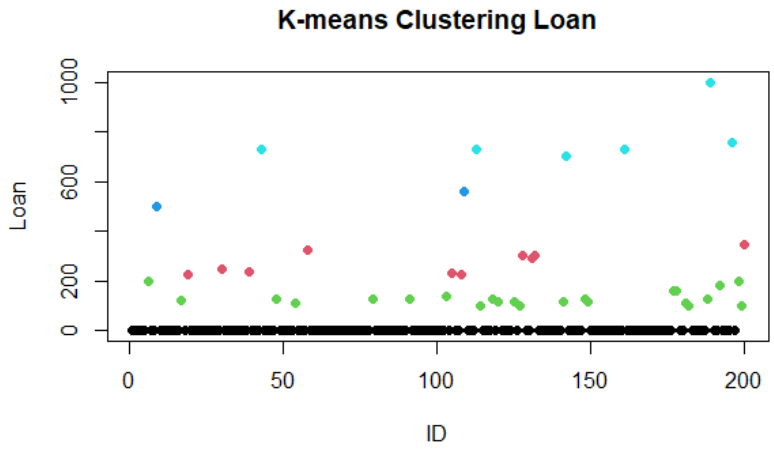


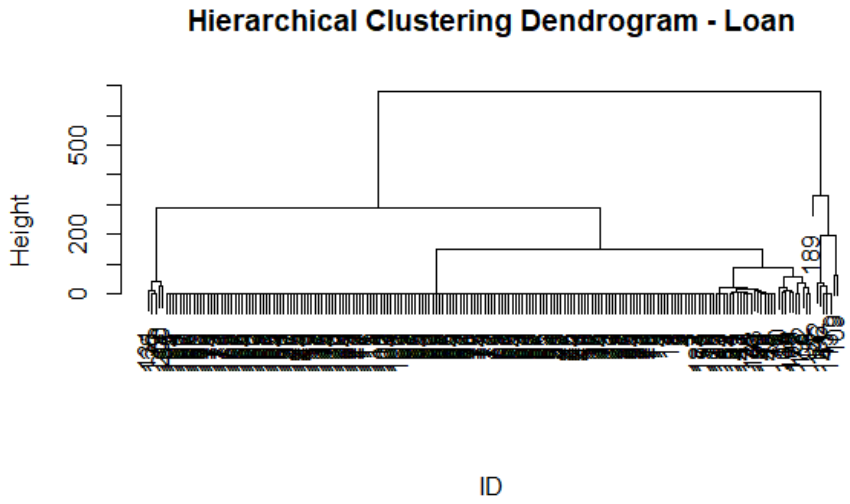
NOTE: There are only 3 clusters because there are only 3 possible values.


Hierarchical Clustering Dendrogram - Education
Height
0.0
0.5
1.0
1.5
ID

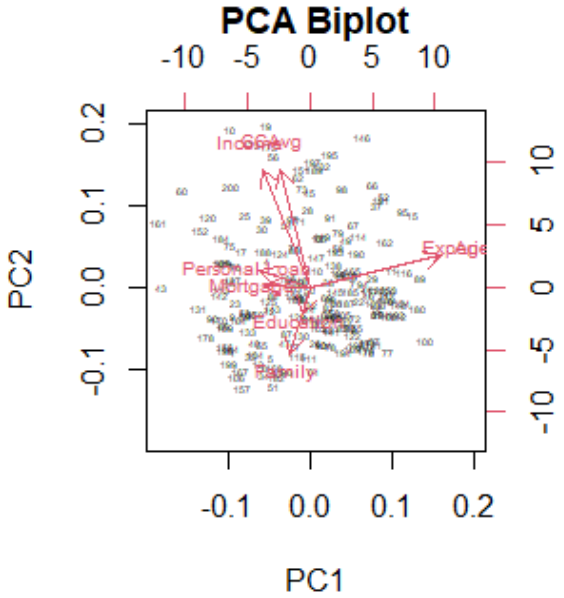

K-means Clustering Mortgage
Mortgage
0
100
200
300
400
0
50
100
150
200
ID


Hierarchical Clustering Dendrogram - Mortgage
Height
0
100
200
ID





Next, there is the graph related to the PCA analysis:



NOTE: The biplot represents the first and second principal components. They explain the maximum variance of the data while reducing complexity. All axes indicate how much variance is explained by the principal components. The dots in the graph represent the observations, while the vectors show the original variables. The angles between the vectors are important because they help us understand the correlation between the variables. There are three types of correlation:

1. Angle close to zero = positive correlation,
2. Angle of 90° = no correlation,
3. Angle close to 180° = negative correlation.

For example, considering Experience and Age, we see a positive correlation; this makes sense because older individuals are likely to have more work experience.