

**Indian Bank Loan Analysis**

D3 Report

**Master in Data Science**

Multivariate Analysis

Group 11

October 30, 2023

**Data**:

Indian Bank Customers

**Team members**:

Casanova Lloveras, Adrià

Chimeno Sarabia, Alícia

García Pizarro, Víctor

Giurcoiu, Victor-George

Ji, Zhengyong

# **Index**

[**Index** 1](#_heading=h.44sinio)

[**Motivation of the work** 2](#_heading=h.z337ya)

[**Data Source presentation** 3](#_heading=h.1y810tw)

[**Data Preprocessing** 5](#_heading=h.4i7ojhp)

[**EDA - Exploratory Data Analysis** 8](#_heading=h.1ci93xb)

[**PCA - Principal Component Analysis - and Outliers Detection** 10](#_heading=h.2bn6wsx)

[**MCA - Multiple Correspondence Analysis** 26](#_heading=h.qsh70q)

[**Multiple Factorial Analysis** 31](#_heading=h.1pxezwc)

# 

# **Motivation of the work**

During the following project, we went through the entire data analysis process over a real-world problem, applying the technical skills acquired and the analytical point of view developed during the course.

At the very beginning of the project, during the selecting process there was a large amount of data sets, nevertheless, following the criteria of minimum amount required for numerical and binary variables, the quantity was reduced to a limited selection. Among these datasets, the India – Bank loan dataset was the most interesting and has a most practical use in the real-world context.

Additionally, this case study serves as a comprehensive and practical example as it covers all the required features. The records contain valuable information such as missing values, unbalanced data, outliers and data sets that require statistical transformations, that allow us to implement almost all the methods obtained during the session.

This data set has collected a set of socio-economic indicators from the clients who have applied for a specific loan to the bank entity. With their consent, these data can be implemented for a data analysis to leverage the assessment of loan application and to spot those collectives that tend to overdue the payment on time. With the study, the entity can implement a sounder model for risk analysis in the banking and financial service to minimize the risk of capital loss.

During the realization of this work, we have introduced Github in our working environment in order to synchronize all the R scripts and documentation and be able to work simultaneously and cooperate between all the members.

# 

# **Data Source presentation**

The database of interest contains socio-economic information about clients who applied for a loan. It originates from a study conducted by IIIT Bangalore, the International Institute of Information Technology Bangalore, that aimed to understand which customers fail to repay a loan, according to the author who uploaded the database on [Kaggle](https://www.kaggle.com/datasets/mishra5001/credit-card) (<https://www.kaggle.com/datasets/mishra5001/credit-card>).

More precisely the database contains 5000 rows. From all those individuals there are 8 numerical variables, 7 categorical and 4 booleans (see Table 1 for more details). Note that, if not specified, time units are in days and money is in rupee (INR).

In this context, this study will analyze the target variable, that states if an individual has payment difficulties or not, to increase our knowledge about the characteristics that makes individuals more susceptible to return their loans on time. Also, it should be remarked that from all the variables available only "ID of loan" will not be considered in the analysis as it does not provide any information apart from identifying an individual.

*Table 1. Metadata Presentation*

| **Variable** | **Full Name**  **// Short\_name** | **Meaning** | **Range** | **Missings (code)** |
| --- | --- | --- | --- | --- |
| ID of loan | SK\_ID\_CURR  //id | ID of loan | [1, 5000] |  |
| Target | TARGET | Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases) |  |  |
| Name of contract type | NAME\_CONTRACT\_TYPE  //contract | Identification if loan is cash or revolving (Cash loans, Revolving loans) |  |  |
| Gender of clients | CODE\_GENDER  //gender | Gender of the client (M - male, F - female) |  | XNA |
| Does the client own a car? | FLAG\_OWN\_CAR//car | Flag if the client owns a car (Y - yes, N - no) |  |  |
| Number of children the clients has | CNT\_CHILDREN  //n\_child | Number of children the client has when asking for a loan | [0, 6] |  |
| Income of the clients | AMT\_INCOME\_TOTAL  //income | Income of the client | [27000, 1350000] |  |
| Credit amount | AMT\_CREDIT  //credit | Credit amount of the loan | [45000, 2606400] |  |
| Loan annuity | AMT\_ANNUITY  //loan | Fixed amount of money that should be payed each month to return the loan | [3172, 129888] |  |
| Goods price | AMT\_GOODS\_PRICE  //price | Price of the goods for which the loan is given | [45000, 2250000] | NA |
| Income type of clients | NAME\_INCOME\_TYPE//  job\_stat | Clients income type (businessman, working, maternity leave,…) |  |  |
| Education type of clients | NAME\_EDUCATION\_TYPE  //studies | Level of highest education the client achieved (Higher education, Incomplete higher, Secondary / secondary special) |  |  |
| Family status of clients | NAME\_FAMILIY\_STATUS  //family | Family status of the client (Married, Single / not married, Widow, Civil marriage, Separated) |  |  |
| Housing type of clients | NAME\_HOUSING\_TYPE  //house | What is the housing situation of the client (renting, living with parents, ...) |  |  |
| Age in days | DAYS\_BIRTH  //age | Client's age in days at the time of application | [-25149, -7721] |  |
| Days of employment | DAYS\_EMPLOYED  //job\_duration | How many days before the application the person started current employment | [-15290, 365243] |  |
| Occupation type of clients | OCCUPATION\_TYPE  //occupation | What kind of occupation does the client have (Laborers, Managers, Core staff, …) |  | “” |
| Type of organization of clients | ORGANIZATION\_TYPE//  job\_type | Type of organization where client works (School, Business Entity Type 3, Industry: type 4, …) |  |  |
| Number of enquiries one month before | AMT\_REQ\_CREDIT\_BUREAU\_MON  //n\_enquiries | Number of enquiries to Credit Bureau about the client one month before application (excluding one week before application) | [0,24] | NA |
| Type of accompanying client | NAME\_TYPE\_SUITE  //companion | Who was accompanying the client when he was applying for the loan (Unaccompanied, Family, Spouse, partner, …) |  | “” |
| ***Table 1:*** The table contains numerical variables (blue), categorical (orange) and boolean (green). “Full name” is the initial name of the database and “short name” is the one used in the analysis. | | | | |

# **Data Preprocessing**

In the below section, we’ll introduce all the statistical transformations that we performed among the row data, and the justification of each step that we choose.

* **Basic transformations**

Firstly, it was necessary to rename the variables names (use of the “short\_name”, stated at Table 1) and some verbose categories to produce plots cleaner and more informative. For example, the category “Single / not married" in the “family” variable was renamed as “single”.

Secondly, for our binary target, which is initially presented as 0 for paid customers and 1 for overdue, we transform it to factor (“paid”, “overdue”). And the same activity was performed with the rest of the binary variable, such as ownership of a car or house.

Thirdly, we did a data quality analysis to ensure that the ranges of all variables were intuitive. For instance, initially the age had negative values as it was registered as “the days a customer spent in their lifetime until the moment they apply for the loan”. As this was not practical to gain knowledge from data, we converted this variable into the unit “years” and, to maintain coherence of units, “job\_duration” was converted too.

Fourth, according to the database some individuals had been working 365243 days (1000 years), this observation occurs only when the customer is not working at the moment of application. So, we declared this variable for this record as “NA” manually.

During the modality analysis, what has been observed is that for some categoric variable, there could be several mutations, e.g. for Industry job, there is Industry type 1, type 2 until type 12. As the meaning of each type is not specified in the documentation, nor a discriminant difference was observed between different types, we decided to perform a reduction of categories to a higher level.

* **Imputation**

Prior to the imputation we calculated the % of missing values and detected that there are three numerical features with missing values, namely price (0.12% missing), job\_duration (17.18% missing), and n\_enquires (13.78% missing). To handle missing values in the qualitative variables, we opted to substitute them by “Variablename\_Unknown”, adding a new modality for each qualitative variable with missing values (job\_type with 17.18% missing, occupation with 30.68% missing, and companion with 0.56% missing).

We first ran the Little Test using the mcar\_test function to check if the numerical features are missing completely at random (MCAR). We got a p-value of 0.0, thus we could reject the null hypothesis, indicating that **our variables are either missing at random or not missing at random**. However, it's important to note that there is no statistical test available to explicitly determine which of these scenarios is the case. Therefore, for our analysis, we have made the **assumption that the variables are missing at random**.

For the purpose of imputing the missing numerical values, we experimented with various imputation techniques, including MIMI[[1]](#footnote-0), MICE[[2]](#footnote-1), and KNN[[3]](#footnote-2). To assess the quality of these imputations, we examined density plots for each of the imputation methods, specifically focusing on the three features with missing data. After careful evaluation, we observed that the density plots were most alike when using the imputations generated by the MICE method (as *Figure 1*). As a result of this consistency, we opted to utilize the MICE imputations for subsequent analysis.



*Figure 1. Comparison of density plot before and after imputation for job\_duration feature.*



*Figure 2. Comparison of density plot before and after imputation for price feature.*



*Figure 3. Comparison of density plot before and after imputation for n\_enquiries feature.*

# **EDA - Exploratory Data Analysis**

In the project, the Exploratory Data Analysis (EDA) was done mainly automatically using the SmartDEA package in R before and after the imputation. This package produced a complete report of basic univariate and bivariate descriptive analysis that can be seen in (Annex 1).

Using those reports and statistical tests like Kolmogorov-Smirnov test or Shapiro–Wilk we can state that none of the variables follow a normal distribution. This suggests that in the sample there are different patterns which can be analyzed in further analysis like PCA and MCA. For example, the distribution of “price” shows different peaks, which means that people tend to ask for certain amounts of money when asking for a loan.



*Figure 4. EDA for the case study.*

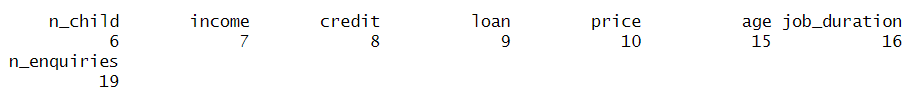
Another relevant observation is that there are some variables highly correlated, as shown in the figure above (left). Here, there are some trivial relations, as, for instance, it stands to reason that the amount of money people ask for a loan is related to the amount they spend on buying goods or services (R^2 = 0.99). However, there are some relations that are much lower than would be expected, for instance “loan” and “income” figure above, right. Looking at this figure, it seems clear that “income” is not such an important variable as one could imagine to determine the loan of a client.

Lastly, we determined that our sample was clearly unbalanced, especially in the target variable as 91% have paid the loan on time. As a consequence, when we extract information about debtors, we need to consider that their sample size was initially small so patterns observed may be extrapolated to the population with care. However, using the median and the mode, it is possible to describe the more frequent applicant for a loan in the database as *“a female of around 42 years, married, without children or cars, who has secondary studies and is working with 7 years of experience and earning around 450000. She has never asked for a loan before and will ask around 513531 INR / year (14% more than her earning) and the loan will be accepted and she will return the money”*. This description should be taken with careful consideration as it came from a basic analysis and more complex methods are required to understand the behavior of the population.

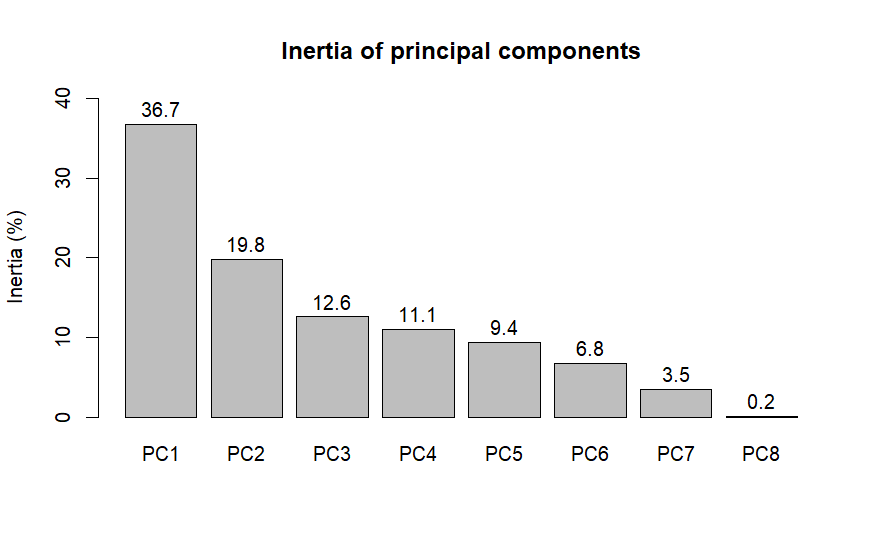
# 

# **PCA - Principal Component Analysis - and Outliers Detection**

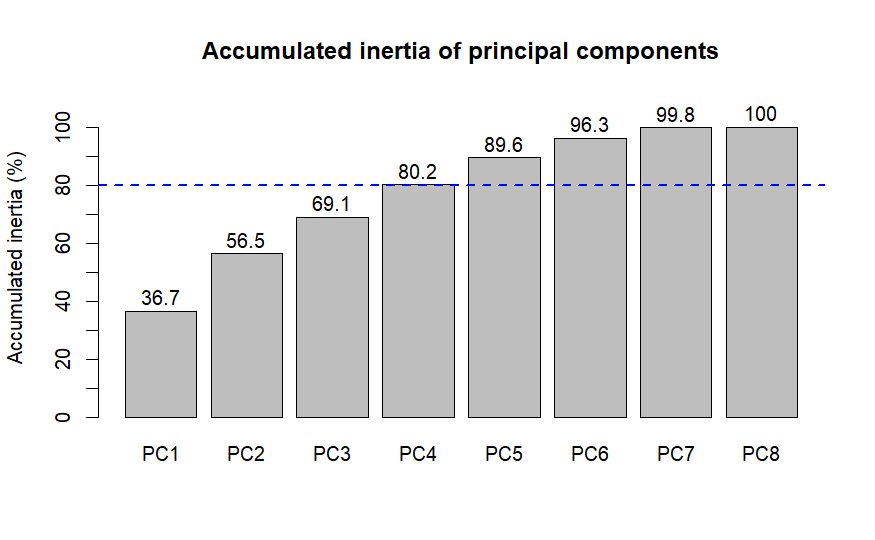
First of all, let us recall what are the numerical variables in our data frame.



We have obtained nine principal components with the following inertias:



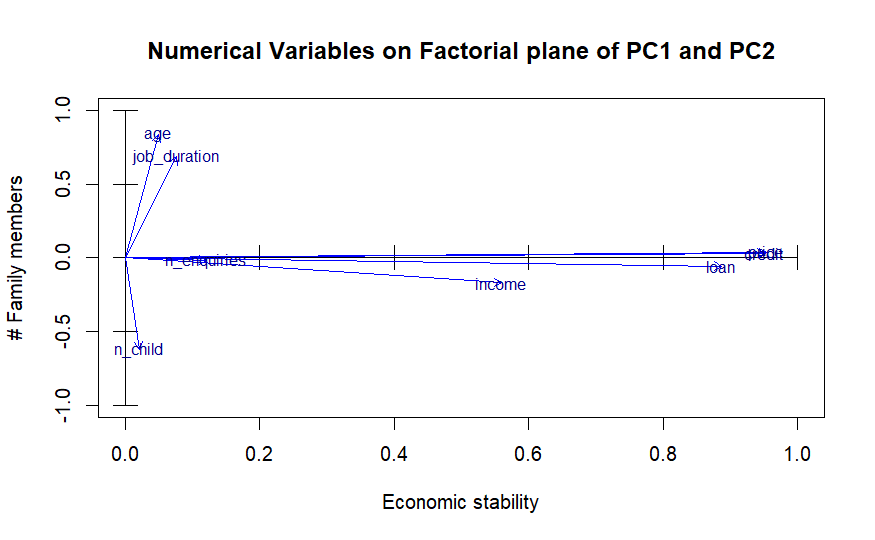
*Figure 5. Inertia (%) of each principal component*



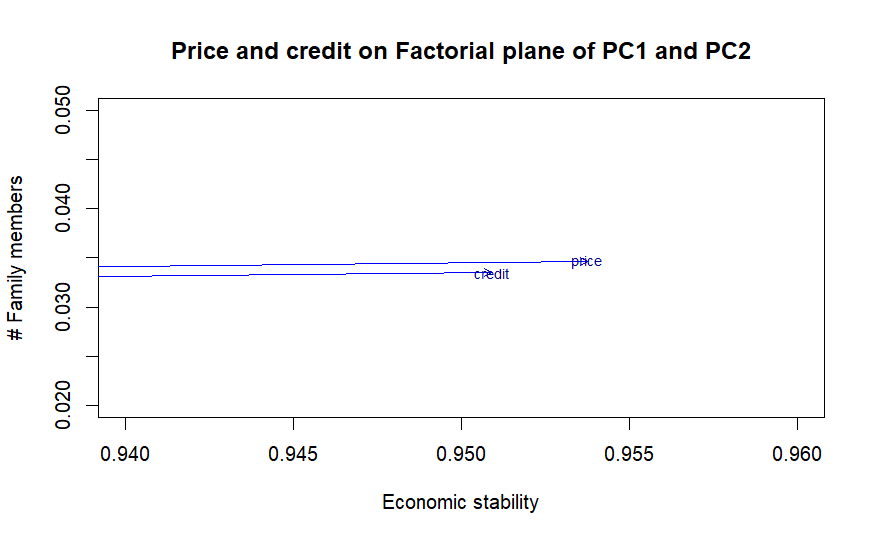
*Figure 6. Accumulated inertia of principal components to reach the 80%.*

We can see that PC1, PC2, PC3 and PC4 accumulate up to around 80% of the total inertia. Below, we will analyze the factorial planes generated by all possible pairs in {PC1, PC2, PC3} to try to figure out the latent meaning of these new variables. Moreover, our goal is to extract relevant information out of the factorial planes.

Our first factorial plane studied will be that generated by PC1 and PC2. PC2 is highly correlated with credit, price, loan and, to a lesser extent, income. This suggests that PC1 represents the economic stability of the client, which increases when we move right on the biplot. On the other hand, PC2 is correlated with age and job duration upwards and the number of children downwards. Even though it does not have such a clear meaning as PC1, this axis could describe the number of alive family members of the customer. Indeed, when we move down the plot, the number of children increases and hence, PC2 as well. Moving upwards increases age, so PC2 is expected to decrease at a proportional rate.

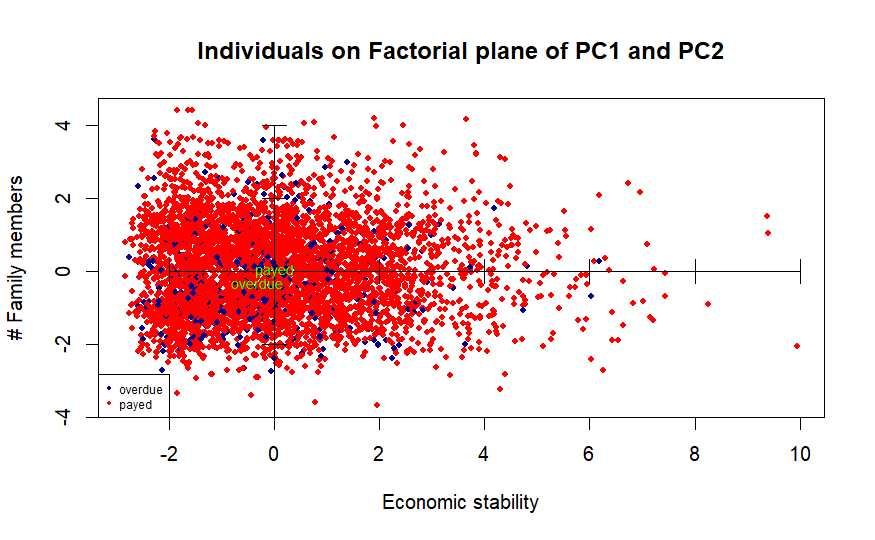


*Figure 7. Projection of Numerical variable among the factorial plan, considering plane PC1 and PC2*

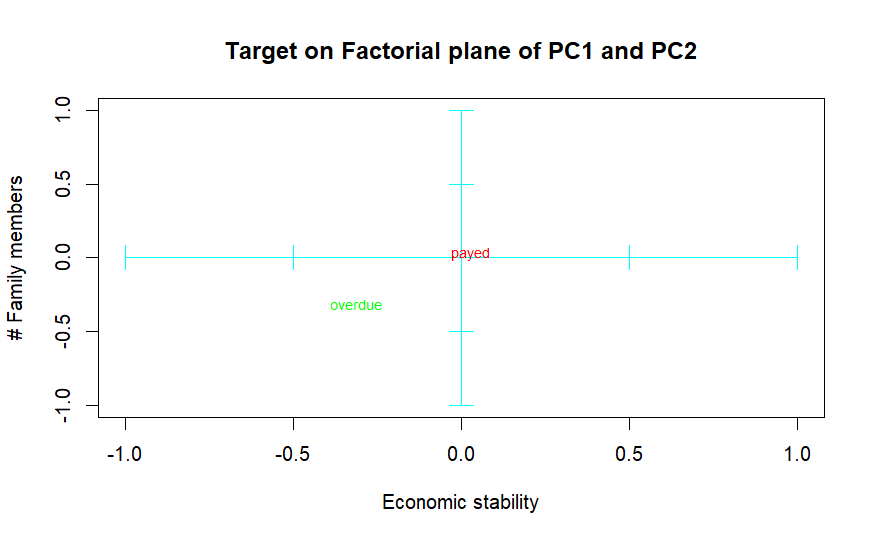


*Figure 8. Projection of Price and Credit on the factorial plan, so observe the highly correlation between the metrics.*

Next, we plot the individuals in two different colors depending on whether they have paid the loan on time or not. It is easy to observe that most people paid the loan on time. Looking at the centroids of the clouds of points, we also conclude that people less economically stable and with more alive family members tend to delay the payment of the loan, which seems reasonable. In fact, our population is skewed left, with just a few individuals possessing a large economic stability or, in other words, being wealthier.

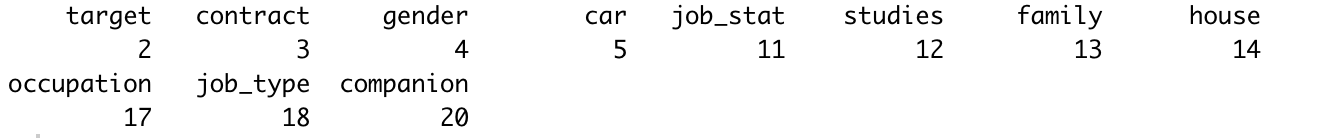


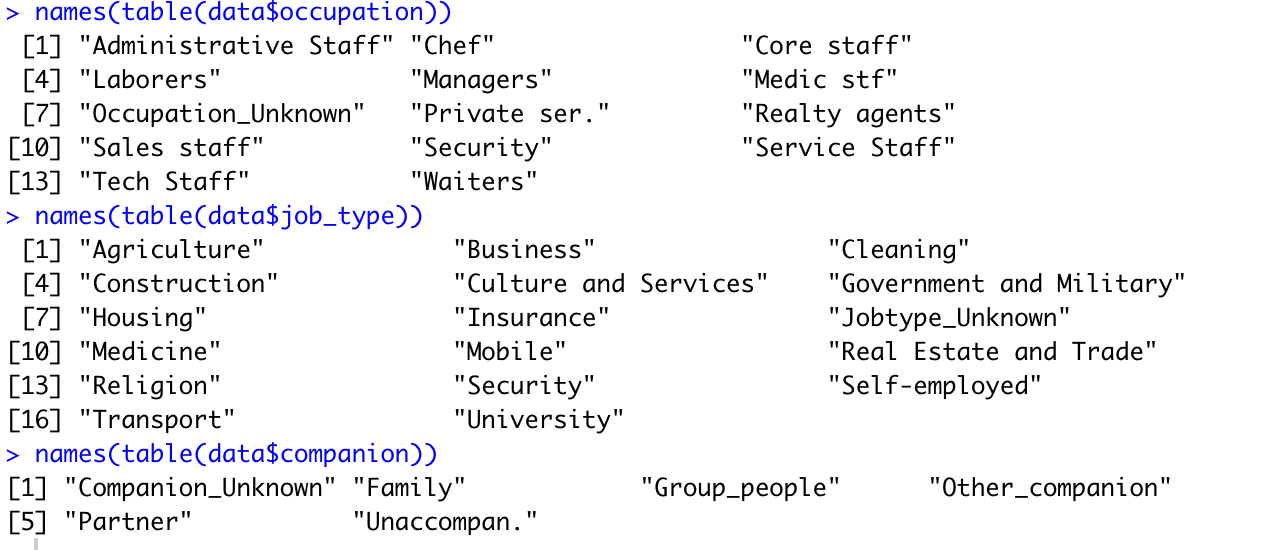
*Figure 9. Target (paid or overdue) observation projecting on Factorial plane of PC1 and PC2*

****

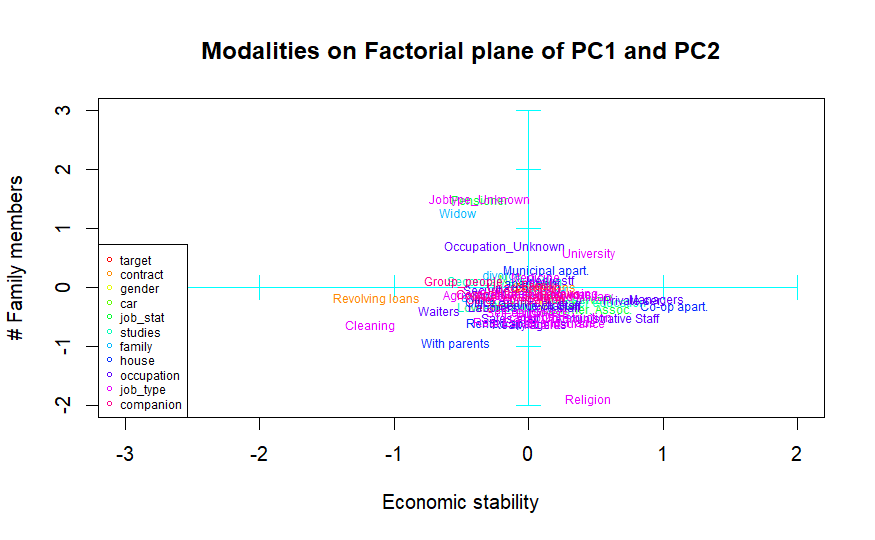
*Figure 10. Target (paid or overdue) projection, by their centroid, on the factorial plane of PC1 and PC2*

Now we will project the modalities of our qualitative variables on the factorial plane to spot relationships among them. Below we can see the variables and their modalities.





Since the clouds of points of the target modalities are very mixed, it is useless to plot them below the other modalities centroids, so we will only plot their centroids.

****

*Figure 11. Modalities projection on the factorial plane of PC1 and PC2*

At first, we plotted all modalities together, but there were so many that it was difficult to see any relationship. That is why we graph them in different biplots that can be found in the annex. In the first one, we can see the contract, gender and car. By proximity of the centroids, women usually don’t own a car and men do. Moreover, men tend to be more stable economically and keep more family members alive; a realistic conclusion. It is striking how fewer stable individuals usually ask for revolving loans instead of cash loans, that is, installment loans.

Afterwards we plot job status, studies and family. Here we can see many more relations. Some are evident, like the fact that widows tend to be pensioners, poor and with few family members alive. Or the fact that people with lower education have more family members and are less stable economically and highly educated individuals are richer. However, it is not so trivial that this last population has more alive family members and tends to work for the state or in commerce. More trivial relationships could be stated, but instead we will focus on the target variable. Individuals of the paid modality most of the time are married or divorced and have a secondary education. On the other hand, the overdue modality contains more single clients with a lower education who don’t work for the state nor in commerce.

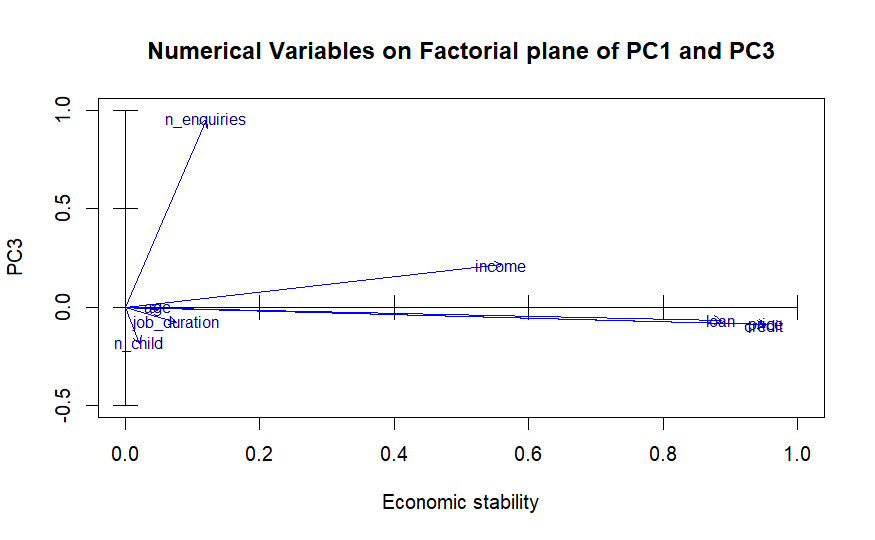
We will proceed by plotting the housing of our customers. Again, many obvious relations come up, like the fact that individuals living with their parents have more family members and are not stable economically. Clients that delay the payment of the loan tend to live in office apartments and, to a lesser extent, in rented apartments. Meanwhile, those who do not, live in standard or municipal apartments.

Plotting occupation, it can be seen that medic staff and chefs are the clients that pay the loan on time most often. Specially laborers, but also waiters, sales staff and realty agents, belong to the overdue modality. Note that we have splitted occupation modalities in two plots to improve visibility. Another observation is that most centroids are below the horizontal axis, while unknown data is way above it. That is, customers with small families do not tend to tell their occupation, maybe because they are unemployed.

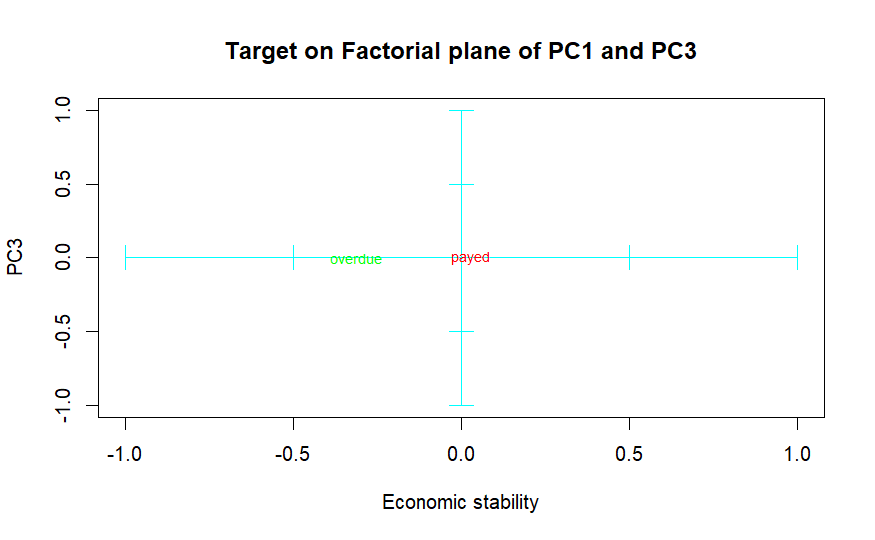
When we plot job types, we realize the same thing that has happened with occupation, unknown data comes from clients with few family members alive and known data belongs to medium or large-sized families. Relevant observations are also the same that the occupation plot has shown.

Finally, we plot companion modalities. It can be noticed that many clients that paid on time were either unaccompanied or with their families when asked for the loan. It is also remarkable that customers that asked for the loan together with a group of people were economically unstable.

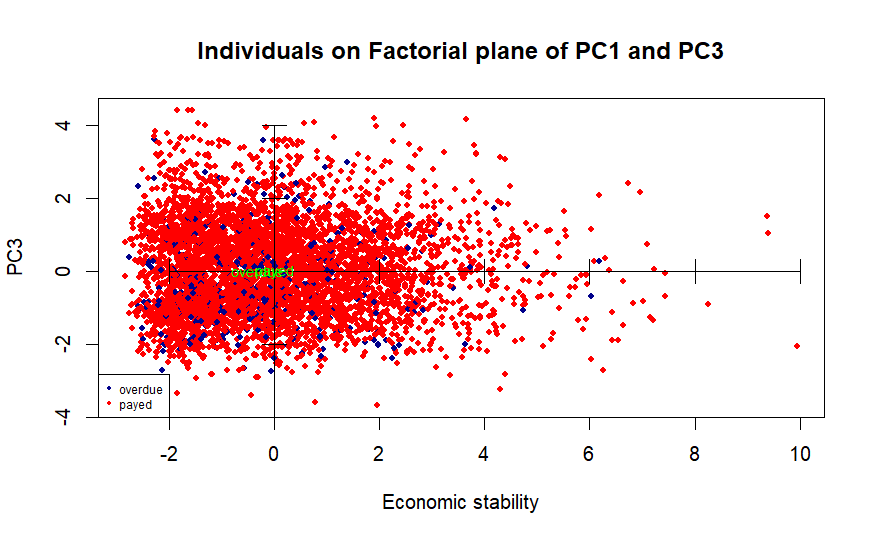
Now we will try to give a meaning to the third principal component by plotting the numerical variables and the individuals on the factorial planes generated by {(PC1, PC3), (PC2, PC3)}.

****

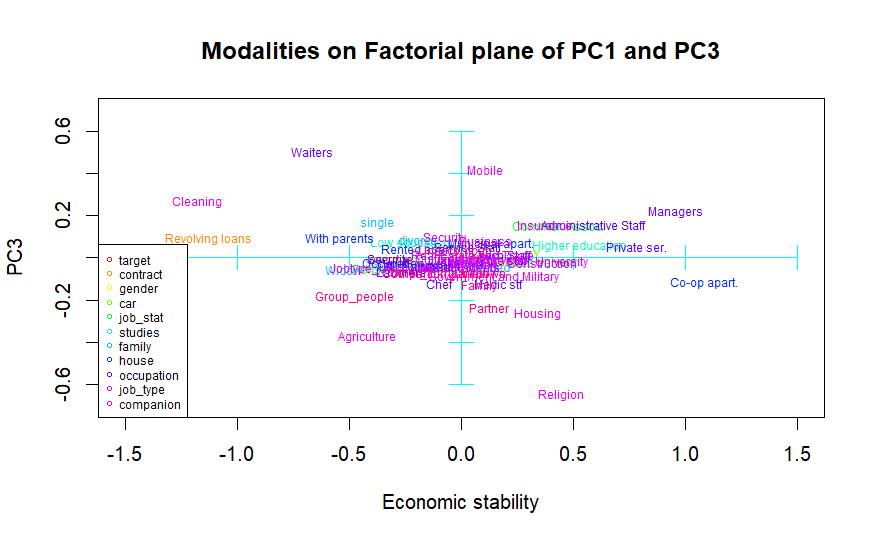
*Figure 12. Projection of Numerical variable among the factorial plan, considering plane PC1 and PC3*

****

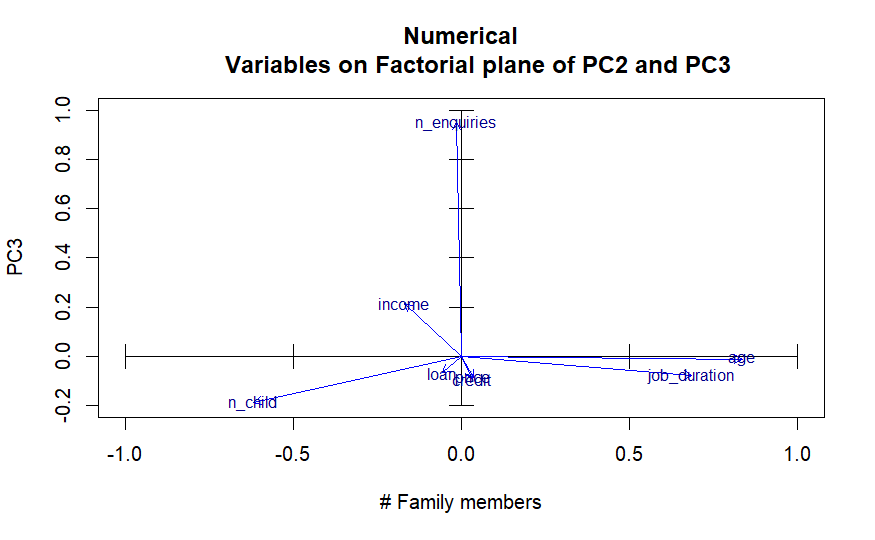
*Figure 13. Target (paid or overdue) projection, by their centroid, on the factorial plane of PC1 and PC3*

****

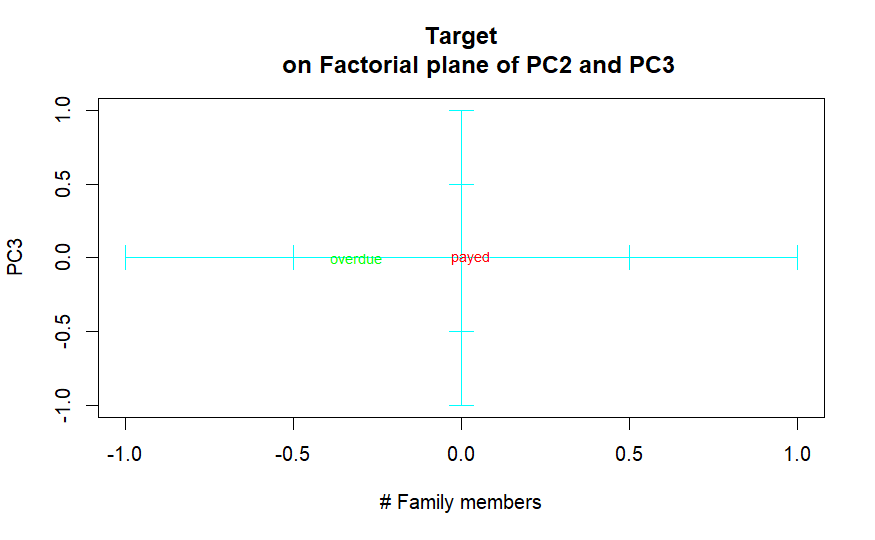
*Figure 14. Target (paid or overdue) observation projecting on Factorial plane of PC1 and PC3*

****

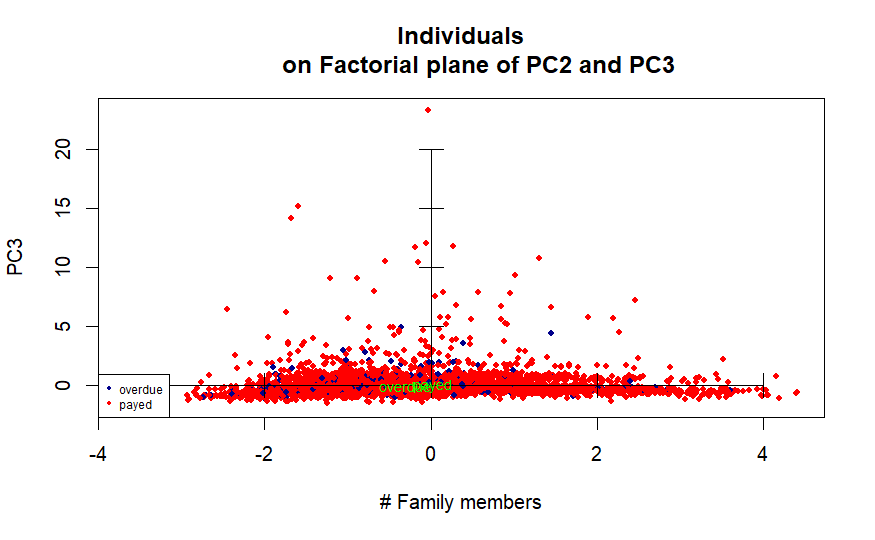
*Figure 15. Modalities projection on the factorial plane of PC1 and PC3*

****

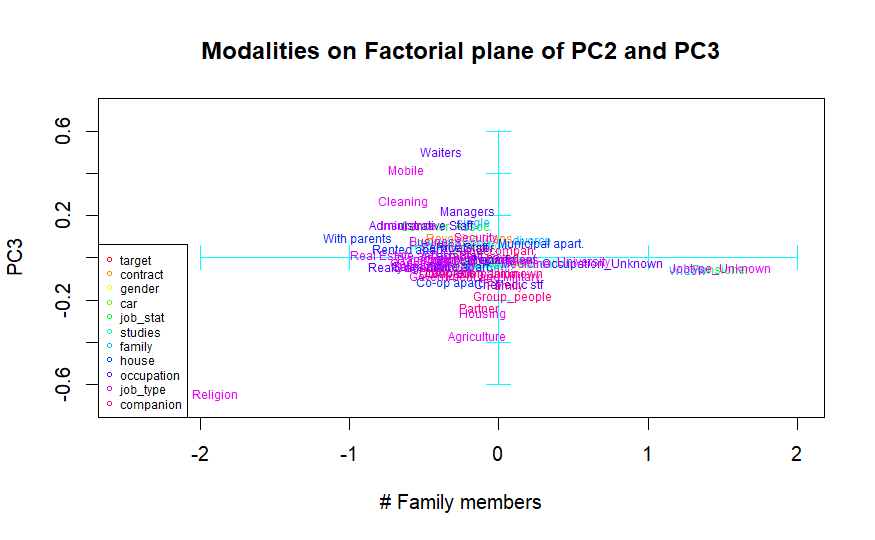
*Figure 16. Projection of Numerical variable among the factorial plan, considering plane PC2 and PC3*

****

*Figure 17. Target (paid or overdue) projection, by their centroid, on the factorial plane of PC2 and PC3*

****

*Figure 18. Target (paid or overdue) observation projecting on Factorial plane of PC2 and PC3*

****

*Figure 19. Modalities projection on the factorial plane of PC2 and PC3*

Looking at the biplots obtained we can only say that it is correlated with the number of enquiries clients due to the bank. Moreover, it does not affect the centroids of our target modalities much. In general, modalities do not seem to describe much PC3 neither, so we will not study it further.

After this thorough discussion of the Principal Component Analysis we have performed, we will extract its main conclusions. We have discovered two relevant latent variables: economic stability and number of alive family members. The first one is clearly a key descriptor of our target, but the second one allows us to see a series of relations between modalities. Most are trivial and could be guessed beforehand, but still we have been able to extract some new information that is summarized in the following points:

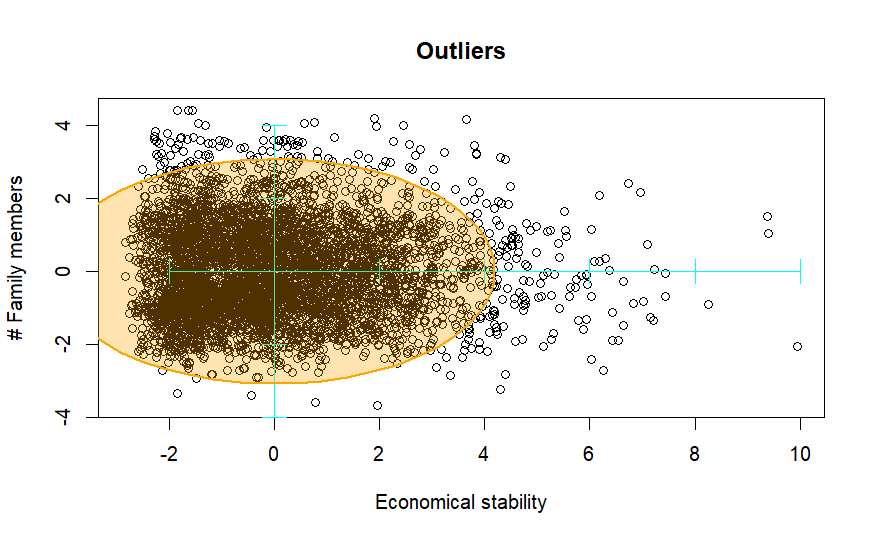
* Women usually don’t own a car and men do.
* Less stable individuals usually ask for revolving loans instead of installment loans.
* Highly educated individuals tend to work for the state or in commerce.
* Individuals of the paid modality most of the time are married or divorced and have a secondary education. The overdue modality contains more single clients with a lower education who don’t work for the state nor in commerce.
* Clients that delay the payment of the loan tend to live in office apartments and, to a lesser extent, in rented apartments. Meanwhile, those who do not, usually live in standard or municipal apartments.
* Medic staff and chefs are the clients that pay the loan on time most often. Specially laborers, but also waiters, sales staff and realty agents, normally belong to the overdue modality.
* In general, customers that paid on time were either unaccompanied or with their families when they asked for the loan.
* Individuals that asked for the loan together with a group of people habitually were economically unstable.

Outlier detection

In this section, we’ll perform the outlier detection and the elimination considering the Mahalanobis distance over the treated PCA data, taking into account the first and second component.

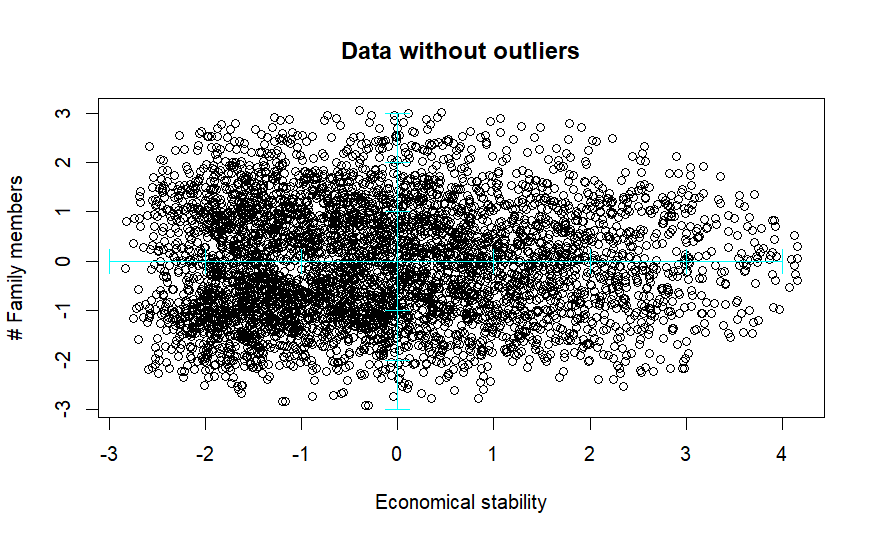
Due to the lack of timing, we were unable to perform an exhaustive review of outlier detection for each feature in the preprocessing process. Instead of that, we decided to apply it on the data after the PCA, with Mahalanobis distance, in order to find the outlier outside the 95% confidence interval.

Projecting all the records on the factorial plane of PCA, considering the first two components as principal, we have the below plot, where the orange ellipse covers the data with a confidence interval of 95%.



*Figure 20. Target projection on the 1rst and 2n components, applying Mahalanobis distance as an ellipse to cover the confidence interval of 95%.*

We have considered those observations outside the area as outliers, and removed them for further analysis, as the picture below shows.

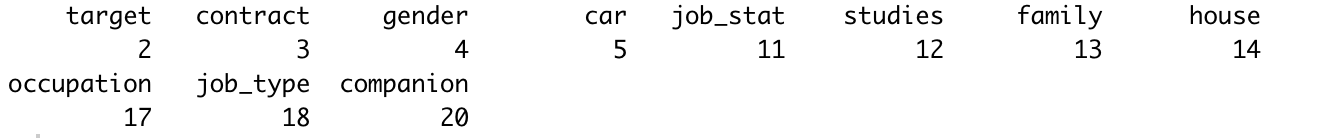


*Figure 21. Treated datasets removing the outliers with Mahalanobis distances.*

This data set will be used in further analysis.

# **MCA - Multiple Correspondence Analysis**

The next step to our analysis is MCA. Let’s recall our categorical variables:



We have the following modalities for each variable:



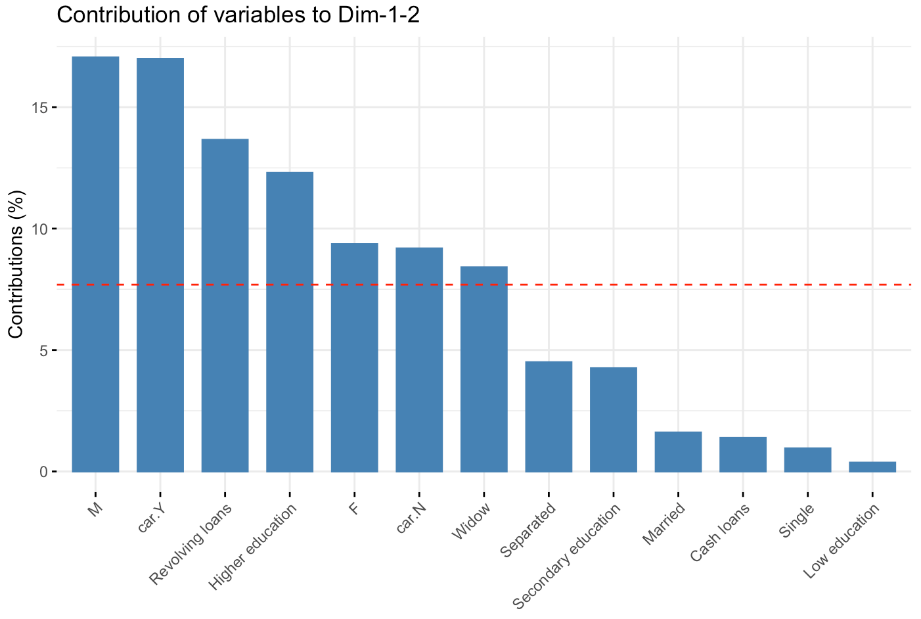
First, we tried an MCA analysis with all our features, but we realized that we needed a lot of dimensions (72) to have 80% of the data explained. We supposed that was because there are 2 features that have a lot of modalities. Therefore, to continue with our analysis, we made some considerations.

We have considered the following:

* The variables job\_type and occupation have too many modalities, for now, we are just going to use them as extra information.
* The house variable is too centered, we shall remove it also for the analysis.
* The target variable is also not considered for the MCA, only for extra information.

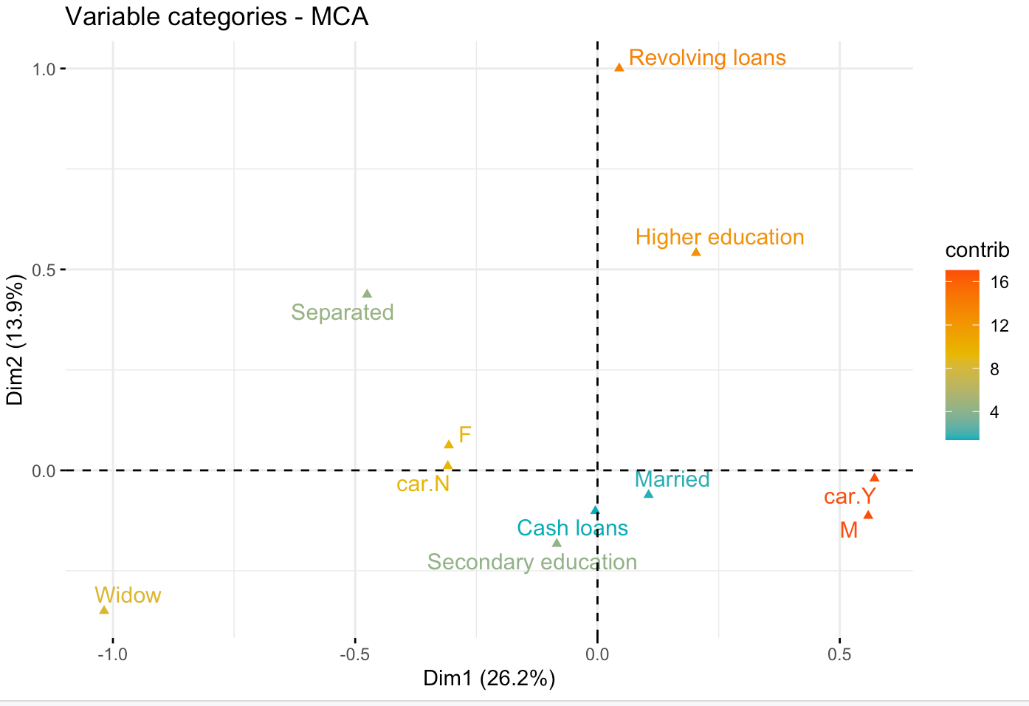
While performing the MCA we can see that:

* We need at least 6 dimensions to describe 80% of the data. With a variance in the first dimension of 18.5% and 13.5% the second.
* The most influential modality (in dimension 1 and 2) is being Male. The top contributions of modalities are:



*Figure 22. Histogram of contribution of each variable.*

With the biplot we can extract the following information:

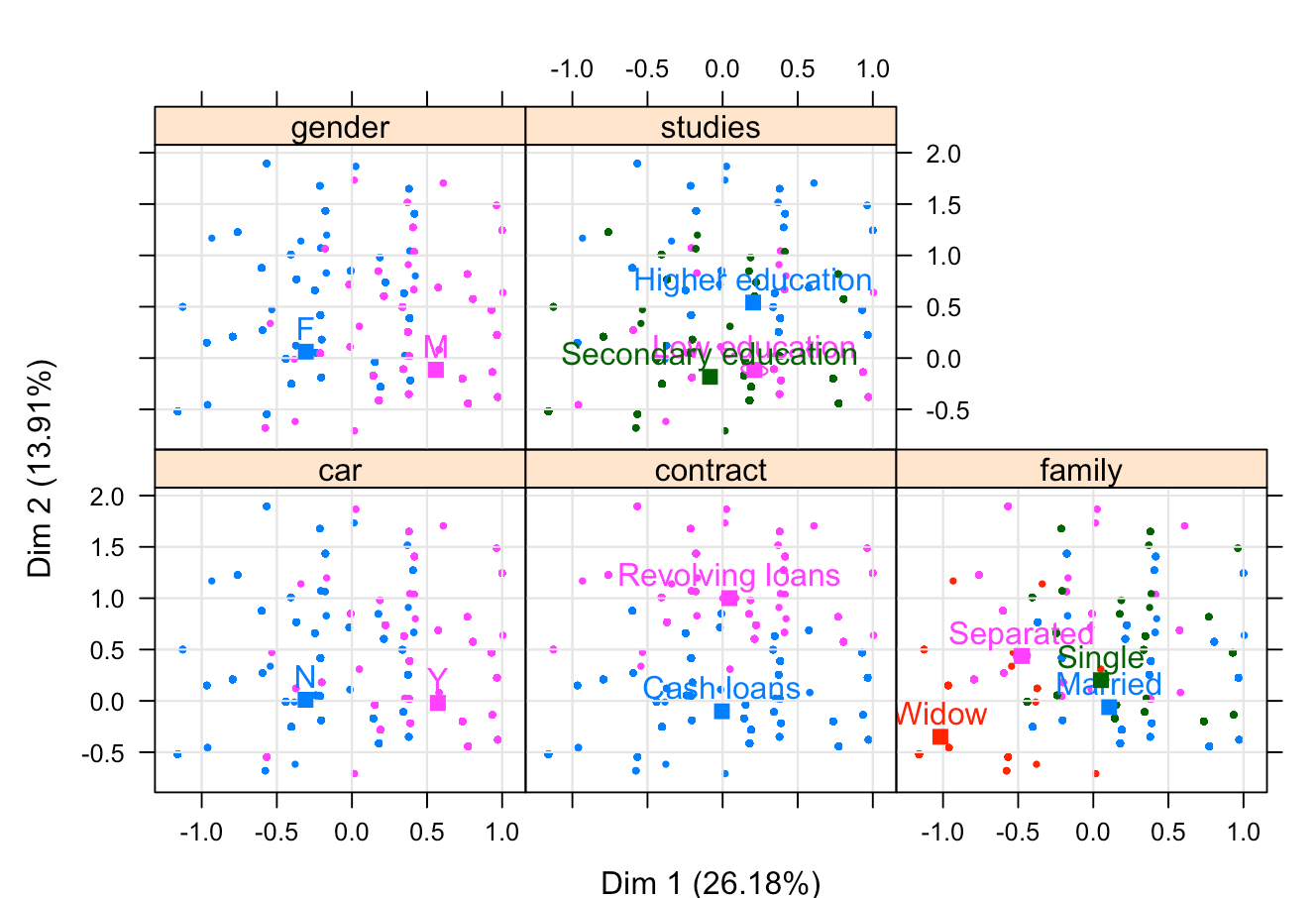


*Figure 23. MCA plot of previous modalities.*

* Modality car=YES and gender=Male are highly correlated with high contribution in dimension 1.
* Modality car=NO and gender=Female are very close and with relatively high contribution in dimension 1.

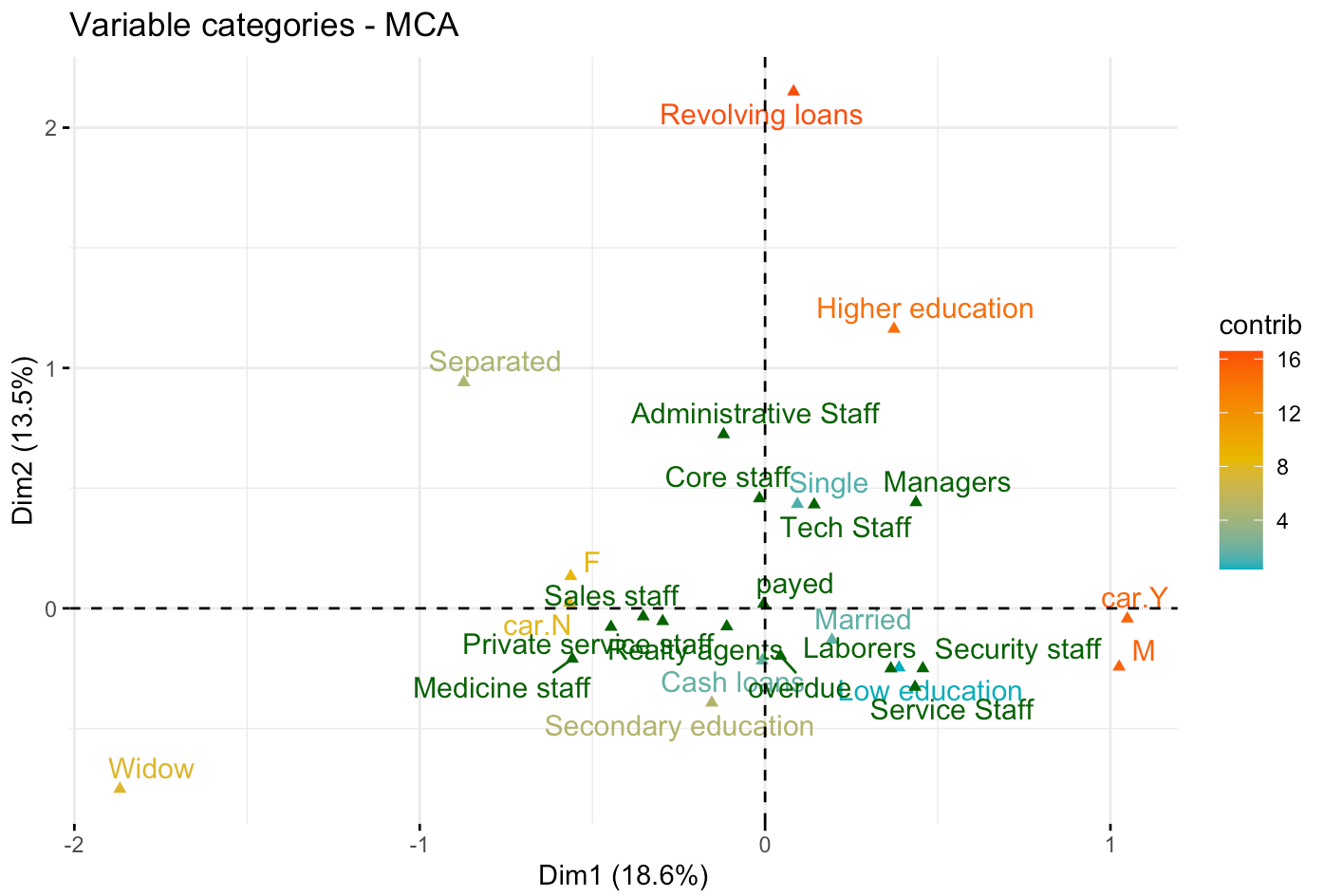
About the dimensions:

* Dimension 1 is about being a male and having a car vs being a female and having no car.
* Dimension 2 is about the type of the loan. Whether it is revolving or cash.



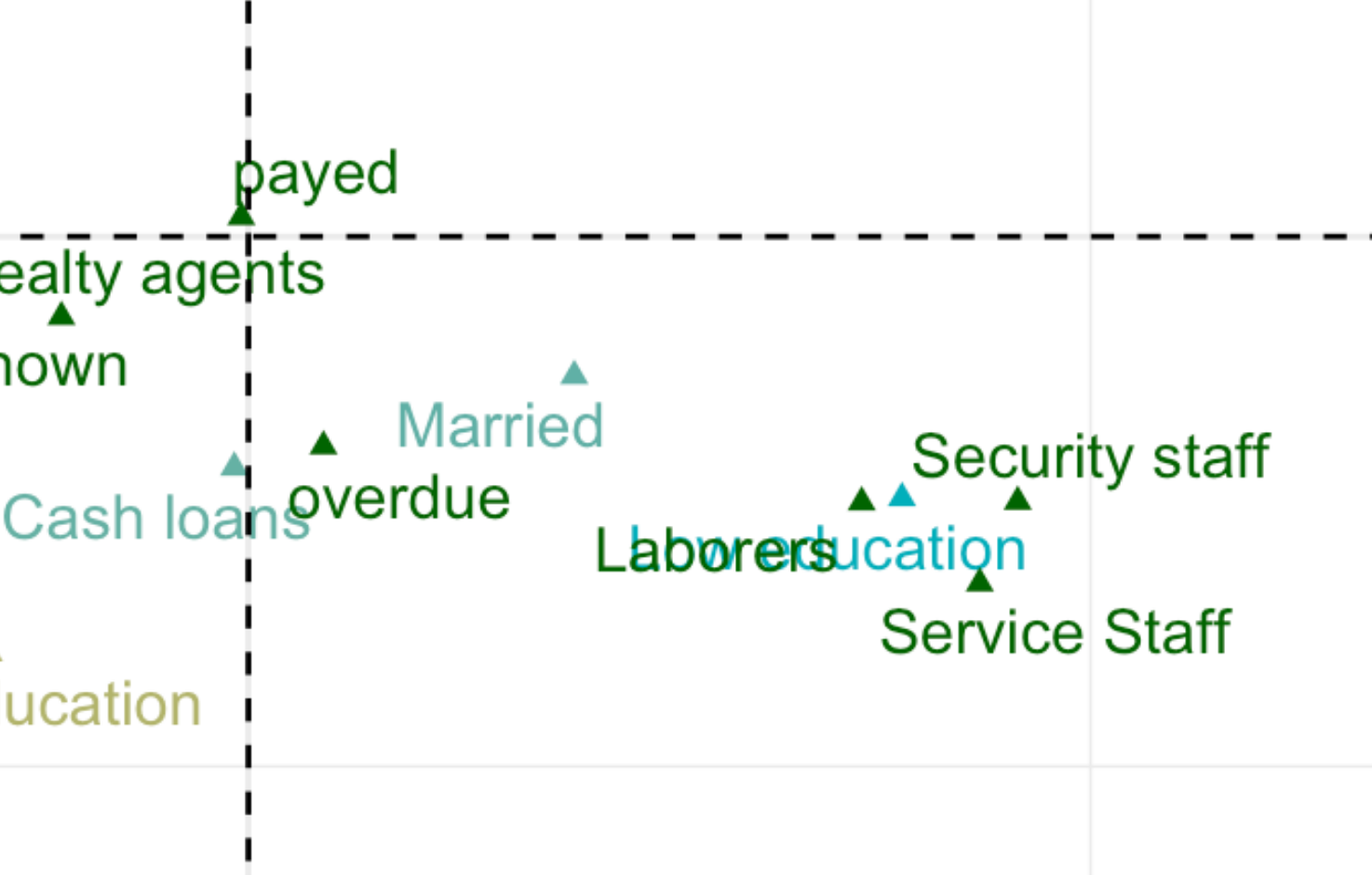
*Figure 24. An MCA study carry out for each modality separately.*

If we add the variable occupation and the target as extra information, we have the following result.



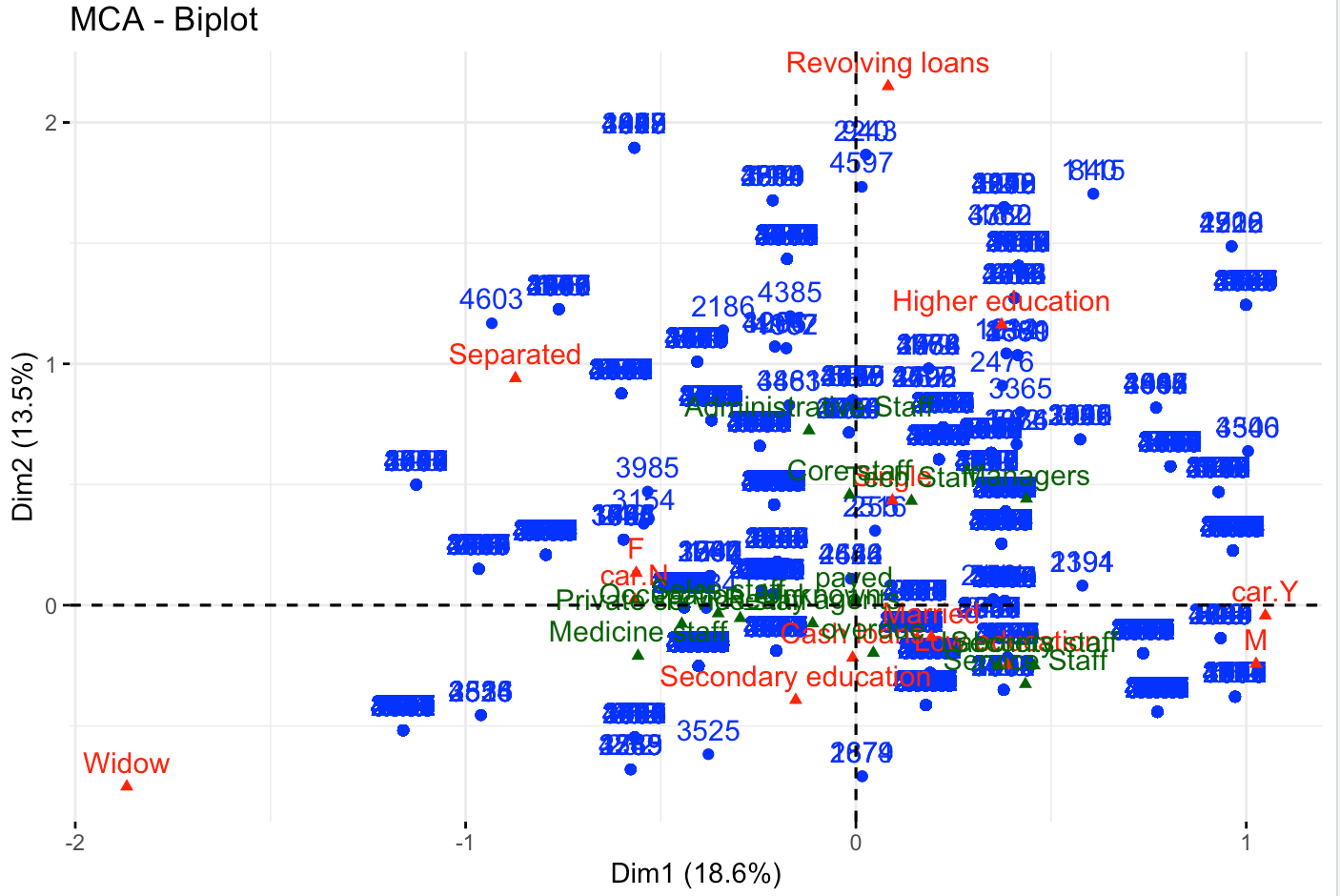
*Figure 25. MCA plot, adding the Job\_occupation modalities.*

We shall zoom in the fourth quadrant. We can see that a person having Low Education is related to working in Services or Security. Recall that the occupation “Service Staff” is a group of the elements: “Cleaning staff", "Cooking staff","Drivers","Waiters/barmen staff". This set is relatively closer to the delay of the payment of the loan than to pay on time.



*Figure 26. A zoom in of the previous MCP plot*

We can plot the individuals. We can see that we have a lot of points that are in the same place. This shows that the individuals are behaving in similar ways.



*Figure 27. MCA plot with all the observation.*

If we plot the occupation and the target variable, we can see that the following occupations: Administrative, Managers, Core Staff, Tech staff are positioned above the x axis and the Service, Security, Laborers are below. Same pattern as the target variable. The people that paid the loan on time are slightly higher than the origin and the people that didn’t pay on time are below. We can see here a pattern for the occupation and the target variable.



*Figure 28. MCA plot with their corresponding confidence area*

Also, we related the y axis with the **type** of the loan. Whether it is a revolving loan or a cash loan. We shall recall that the centrum of gravity of the paid modality is higher than the overdue modality. That would mean that the individuals that paid the loan out of time are more likely to have a cash loan than a revolving loan. This could be explained by:

* Urgent Financial Needs: People may opt for cash loans when they have immediate financial needs (unexpected or urgent expenses, medical emergencies). In such cases, individuals may not have thoroughly planned for repayment, leading to a higher risk of delinquency.
* Higher Interest Rates: Cash loans, particularly payday loans, can have higher interest rates compared to revolving loans.
* Borrower's Financial Stability: Individuals who resort to cash loans may be in a more financially precarious situation compared to those using revolving credit. This is consistent with our analysis, because the jobs that are located below the x axis tend to be more financially unstable (waiters, cleaners, drivers, cooks). They may need at some point urgently money to borrow from the bank and they would not be able to pay it on time.



# **Multiple Factorial Analysis**

Even though PCA has been our main source of conclusions in the analysis of the database, MCA has elucidated a core pattern that involves the target variable. Specifically, on the MCA plot of the two first dimensions, people that paid the loan on time are slightly higher than the origin and people that didn’t pay on time are below. Hence, we can, for example, tell what professions are more likely to pay the loan on time and what are not.

The remainder of conclusions made in the analysis are summarized at the end of the PCA section.

**Annex**

EDA Reports (after and before analysis)

The basic statistics of numerical data is shown in the following table:

id target contract gender

Min. : 1 Length:5000 Length:5000 Length:5000

1st Qu.:1251 Class :character Class :character Class :character

Median :2500 Mode :character Mode :character Mode :character

Mean :2500

3rd Qu.:3750

Max. :5000

car n\_child income credit

Length:5000 Min. :0.0000 Min. : 27000 Min. : 45000

Class :character 1st Qu.:0.0000 1st Qu.: 112500 1st Qu.: 270000

Mode :character Median :0.0000 Median : 144000 Median : 513531

Mean :0.4198 Mean : 166536 Mean : 598769

3rd Qu.:1.0000 3rd Qu.: 202500 3rd Qu.: 810000

Max. :6.0000 Max. :1350000 Max. :2606400

loan price job\_stat studies

Min. : 3172 Min. : 45000 Length:5000 Length:5000

1st Qu.: 16457 1st Qu.: 234000 Class :character Class :character

Median : 25083 Median : 450000 Mode :character Mode :character

Mean : 27071 Mean : 536690

3rd Qu.: 34911 3rd Qu.: 679500

Max. :129888 Max. :2250000

family house age job\_duration

Length:5000 Length:5000 Min. :21.00 Min. : 0.1041

Class :character Class :character 1st Qu.:33.00 1st Qu.: 2.2219

Mode :character Mode :character Median :42.00 Median : 4.8466

Mean :43.18 Mean : 7.3363

3rd Qu.:53.00 3rd Qu.: 9.5452

Max. :68.00 Max. :41.8904

occupation job\_type n\_enquiries companion

Length:5000 Length:5000 Min. : 0.0000 Length:5000

Class :character Class :character 1st Qu.: 0.0000 Class :character

Mode :character Mode :character Median : 0.0000 Mode :character

Mean : 0.2828

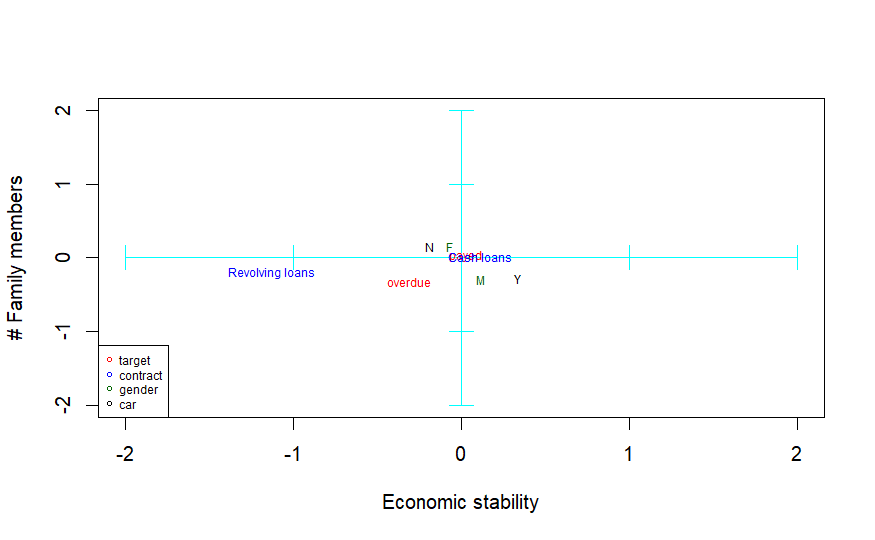
3rd Qu.: 0.0000

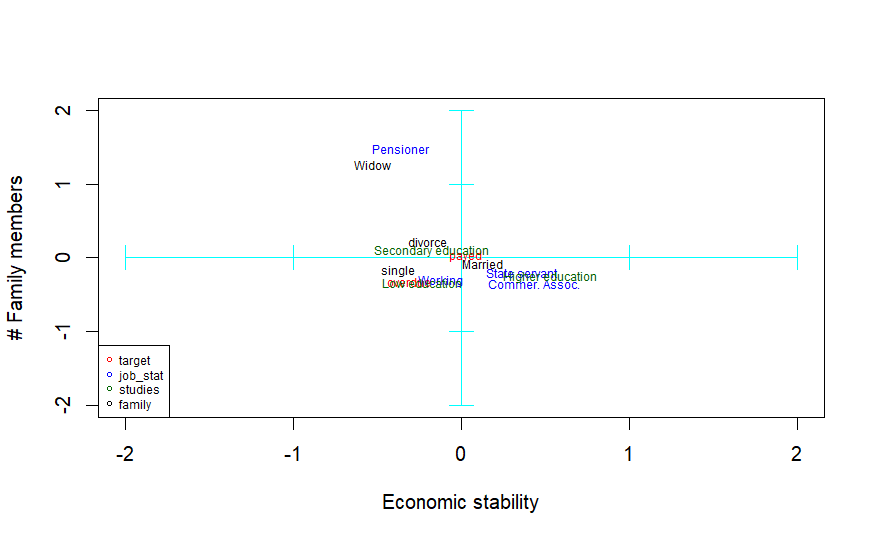
Max. :24.0000

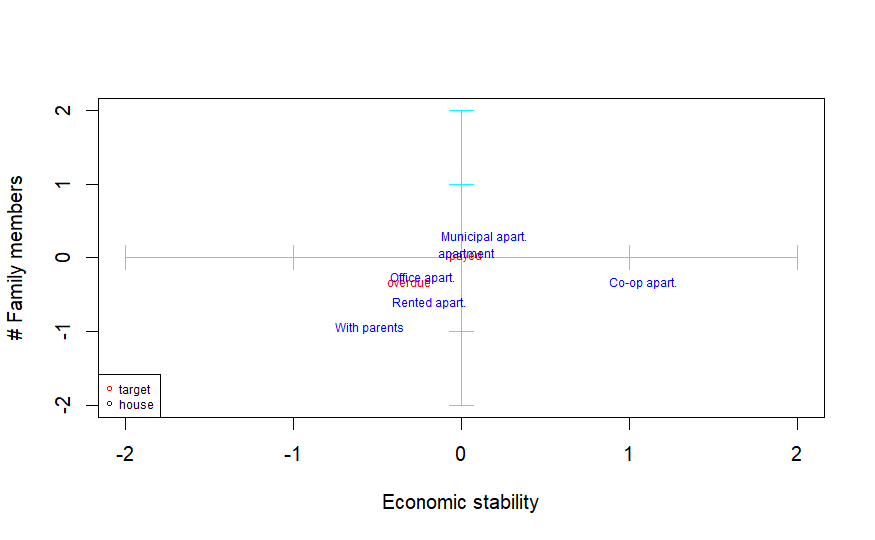
For a more detailed EDA analysis see the reports generated by “Smart EDA” package before the imputation, after imputation, and only for payers and debtors. To see the report it is recommended that you download them from the following link:

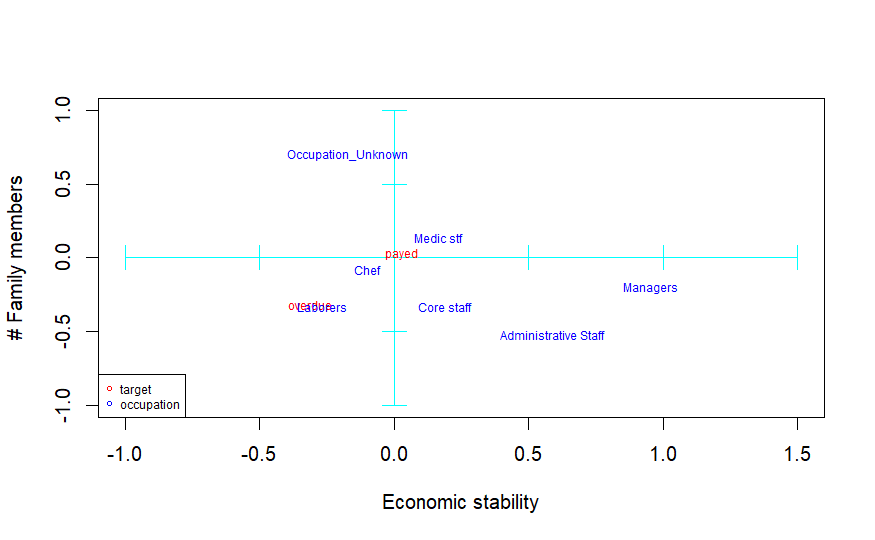
*https://drive.google.com/drive/folders/1g62S6VQS3HHU6jYw\_gR6PM7LAXpt6n3N?usp=drive\_link*

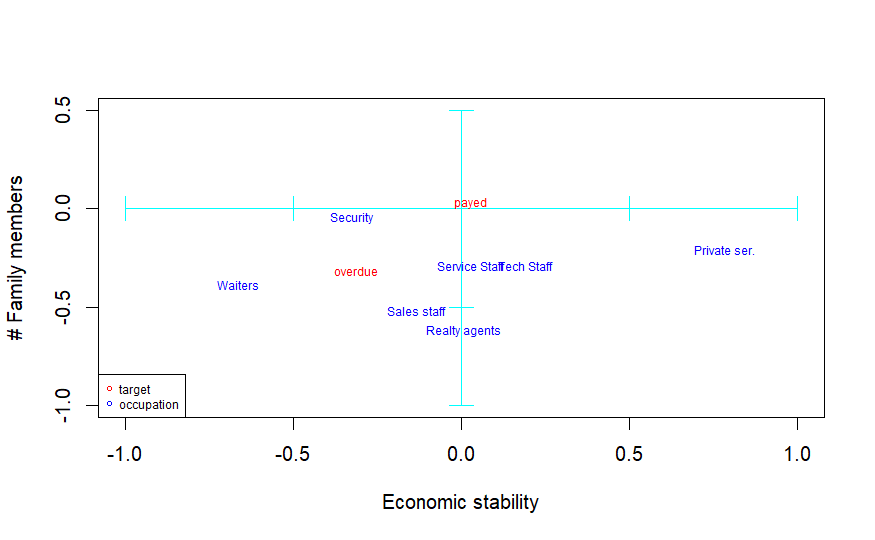
PCA Modalities

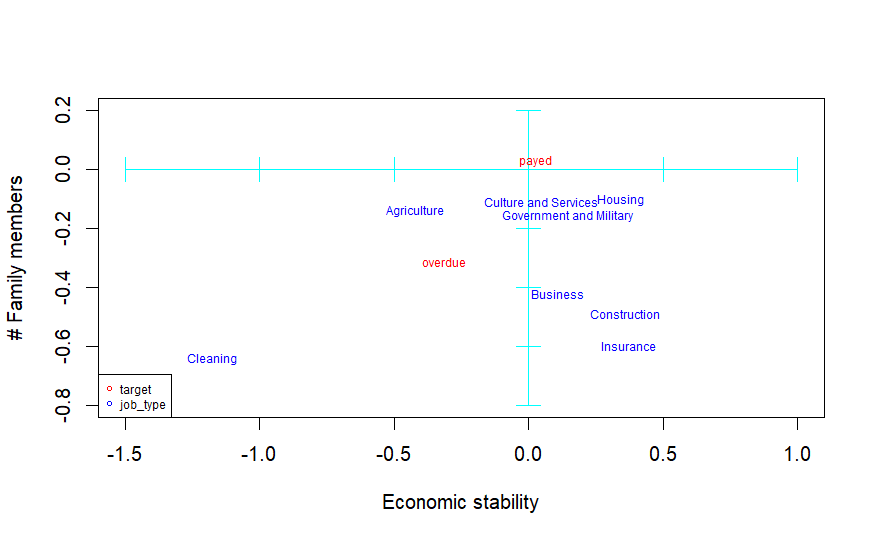
****

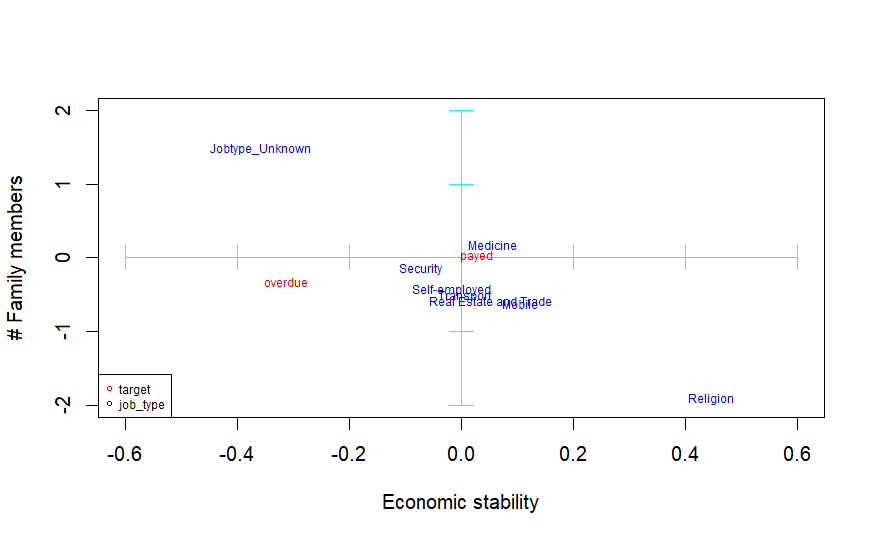
****

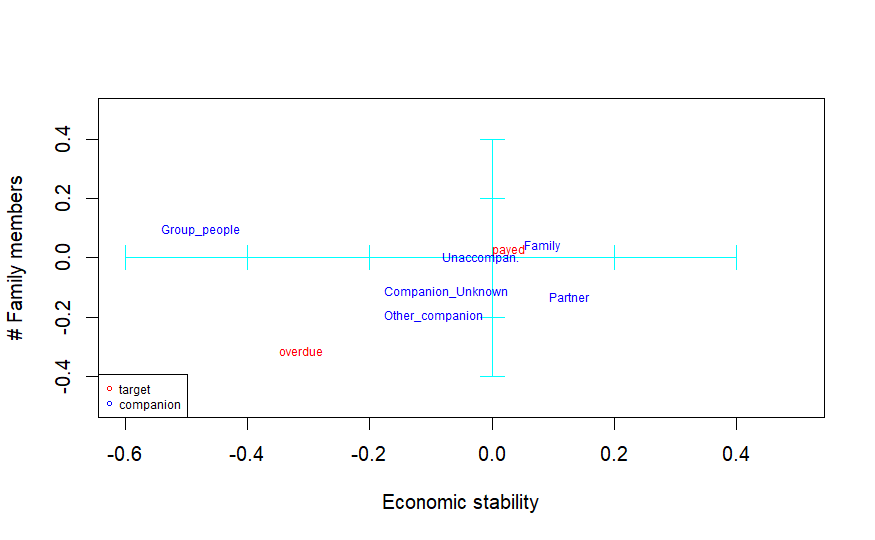
****

****

****

****

****

****

Gantt diagram and task distribution

| **Activity** | **Description** | **Starting date** | **Estimated Duration** | **Deadline** | **Adria C.** | **Alicia C.** | **Ji** | **Victor G.** | **Victor Geo** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Delivery D3** | | Wk-40 | 4 | **Wk-44** |  |  |  |  |  |
| **Activity 9** | **Project introduction** | Wk-40 | 1 | Wk-40 |  |  | **X** |  |  |
| **Activity 10** | **Index** | Wk-40 | 1 | Wk-41 |  |  | **X** |  | **X** |
| **Activity 11** | **Project Motivation** | Wk-40 | 1 | Wk-41 |  |  | **X** | **X** |  |
| **Activity 12** | **Data Source presentation** | Wk-40 | 1 | Wk-41 |  |  | **X** | **X** |  |
| **Activity 13** | **Description of Data structure & metadata** | Wk-40 | 1 | Wk-41 |  |  | **X** | **X** | **X** |
| **Activity 14** | **Preprocessing Justification** | Wk-40 | 1 | Wk-41 | **X** | **X** |  | **X** |  |
| **Activity 15** | **Data Overview with R** | Wk-40 | 1 | Wk-42 |  | **X** |  | **X** | **X** |
| **Activity 16** | **Univariate & bivariate analysis** | Wk-41 | 1 | Wk-42 |  |  | **X** | **X** |  |
| **Activity 17** | **PCA** | Wk-41 | 1 | Wk-42 | **X** | **X** |  |  | **X** |
| **Activity 18** | **MCA** | Wk-41 | 1 | Wk-43 |  | **X** |  |  |  |
| **Activity 19** | **MFA** | Wk-41 | 1 | Wk-43 | **X** |  |  |  |  |

Contingency risk table

| **Risk** | **How to Prevent** | **How to Manage** |
| --- | --- | --- |
| Team member leave | Communication. If someone is going to leave and the other members are aware of it the risk to the project is less | Readjustment of task distributions |
| Data Source untrustable | Ensure that the source of the database is trustable before making any further analysis | Be transparent at the report: write about what is known about the source and outline the reasons why it is not trustable |
| Work progress is lost | Each member will have at least one backup of all the project | Start again from a previous version |
| Inconsistency of versions in scripts | Use a version control platform like GitHub or a base script, where all the members start their work | Merge scripts |
| Communication lost with one member | We will have at least two ways to contact each of the members, like Gmail and WhatsApp | Contact the person from an alternative way |
| A team member is overworked | Have a good task distribution according to the workload of the semester | Readjustment of task distributions |
| Difficulties when applying a method (e.g., PCA) | Communication during all work giving tips and help when needed | Collaboration in the team. Other members can work in an activity initially planned for one if required |
| Imputation is done incorrectly (e.g., distributions before and after clearly differ) | The imputation is done by different methods and by different people | Repeat the imputation by another method, if possible, or leave missing values |
| Difficulties to explain any relation with the initial target variable in mining analysis (e.g., we do not manage to predict sales in the next 6 months) | Bibliography research or use of more advance mining techniques | Be transparent in the report |
| One member cannot attend the final presentation of the project | Members will know all the presentation | The explanation of the missing member is divided into the other members |
| Difficulties to explain the initial target variable in mining analysis (eg: we do not manage to predict sales in the next 6 months) | Perform several analysis | Be transparent in the report |
| One member cannot attend the final presentation of the project | Members will know all the presentation | The explanation of the missing member is divided into the other members |

1. Mixed Intelligent-Multivariate Missing Imputation [↑](#footnote-ref-0)
2. Multivariate Imputation By Chained Equations [↑](#footnote-ref-1)
3. K-Nearest Neighbors [↑](#footnote-ref-2)