UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Indian Bank Loan Analysis

D4 Report

Master in Data Science

Multivariate Analysis

Group 11

January 16, 2024

Data:

Indian Bank Customers

Team members:

Casanova Lloveras, Adrià

Chimeno Sarabia, Alícia

García Pizarro, Víctor

Giurcoiu, Victor-George

Ji, Zhengyong

Index

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).

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], MICE[2], and KNN[3]. 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.

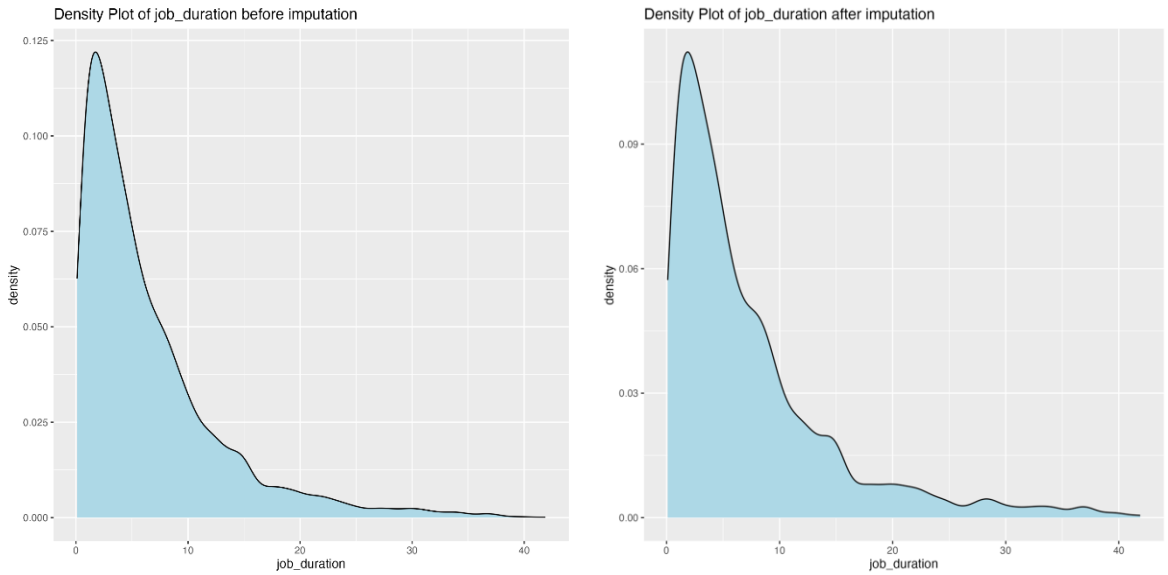


Figure 1. Comparison of density plot before and after imputation for job_duration feature.

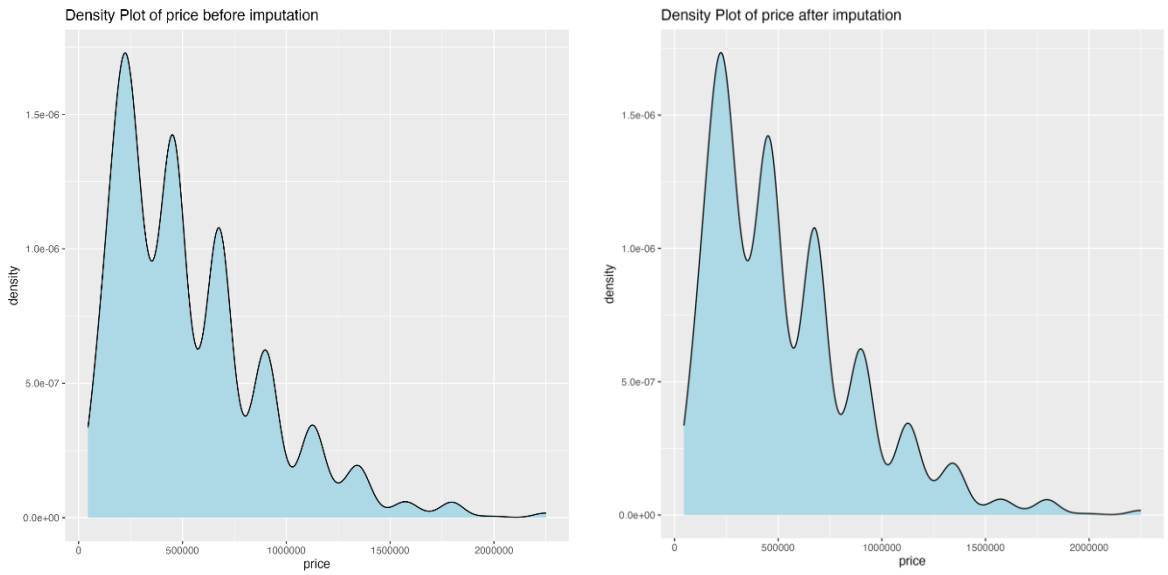


Figure 2. Comparison of density plot before and after imputation for price feature.

[1] Mixed Intelligent-Multivariate Missing Imputation
[2] Multivariate Imputation By Chained Equations
[3] K-Nearest Neighbors

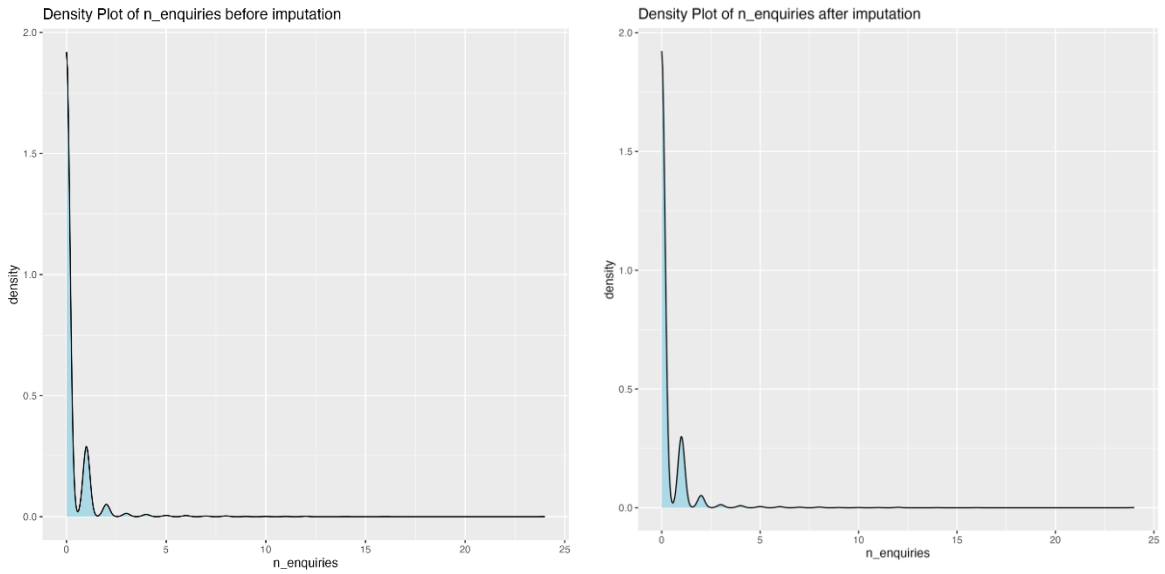


Figure 3. Comparison of density plot before and after imputation for n_enquiries feature.

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.

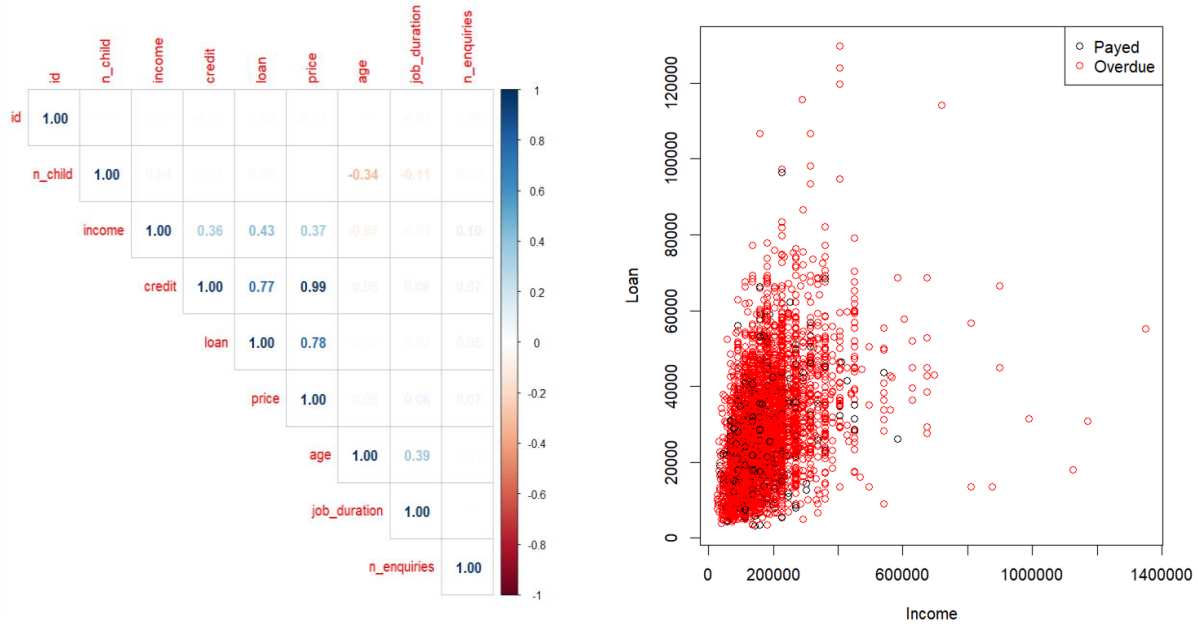


Figure 4. EDA for the case study.

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.

```
    n_child      income      credit        loan       price         age job_duration
          6           7           8           9          10          15           16
n_enquiries
         19
```

We have obtained nine principal components with the following inertias:

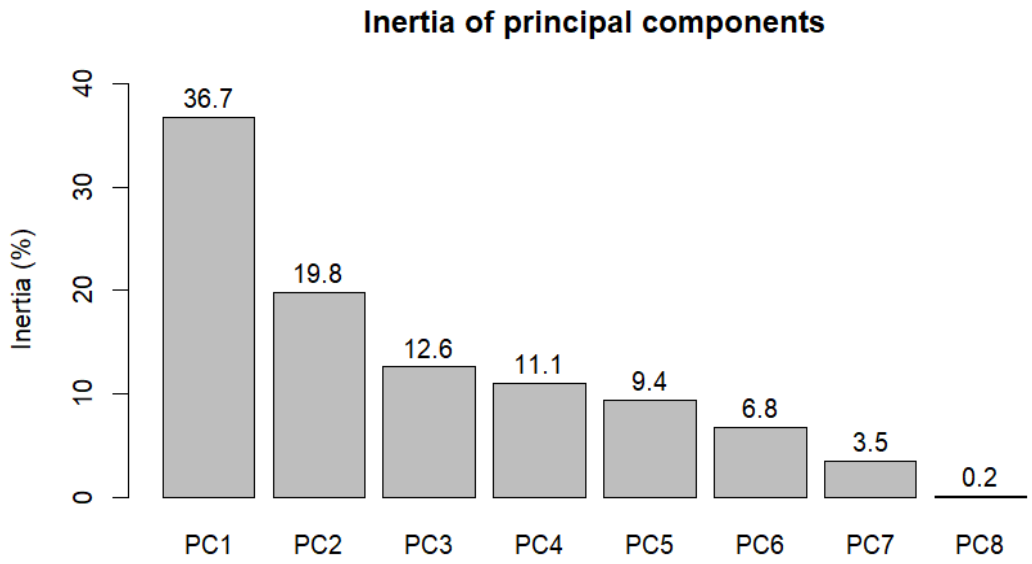


Figure 5. Inertia (%) of each principal component

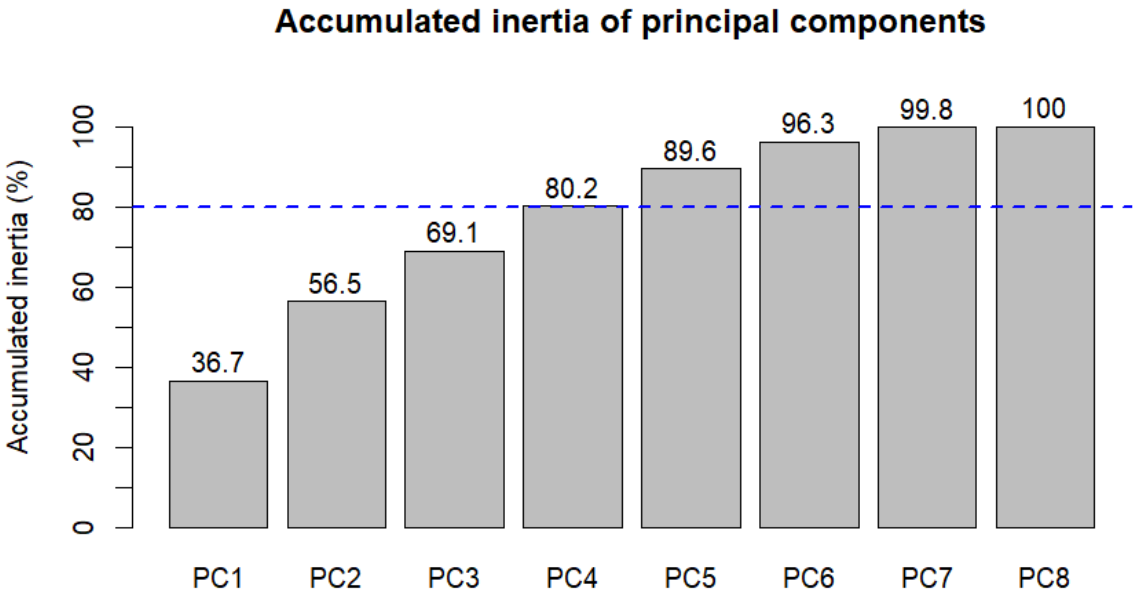


Figure 6. Accumulated inertia of principal components to reach the 80%.

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 client's wealth and his/her/their loan, 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. From a general point of view, this principal component represents the family status of the client and his/her/their sociological profile.

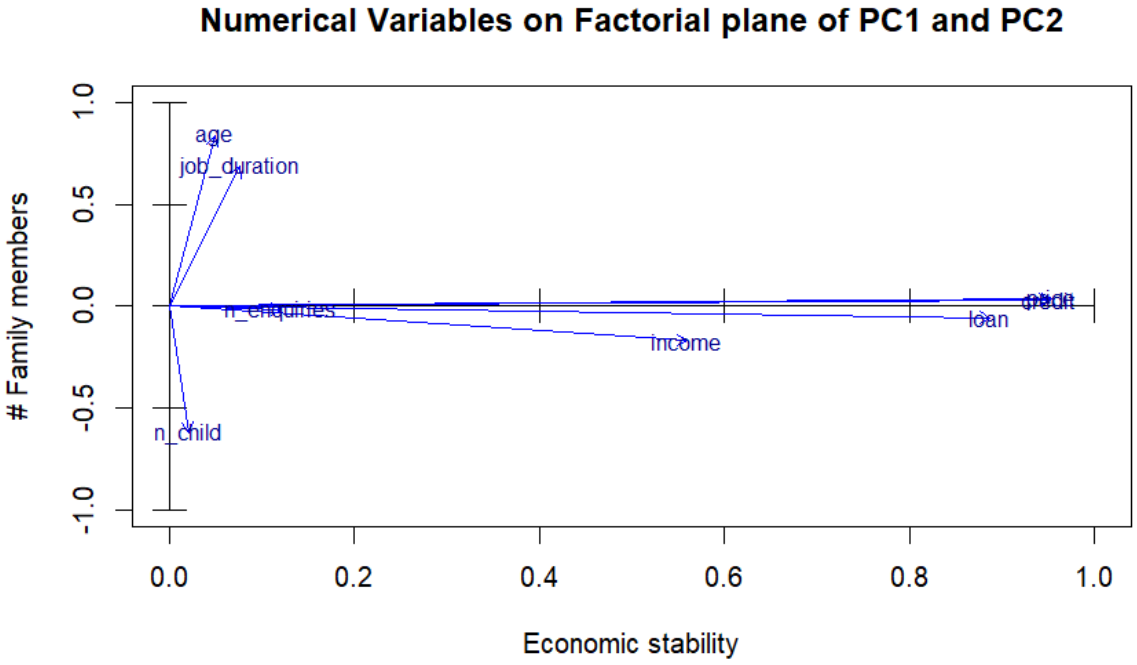


Figure 7. Projection of Numerical variable among the factorial plan, considering plane PC1 and PC2

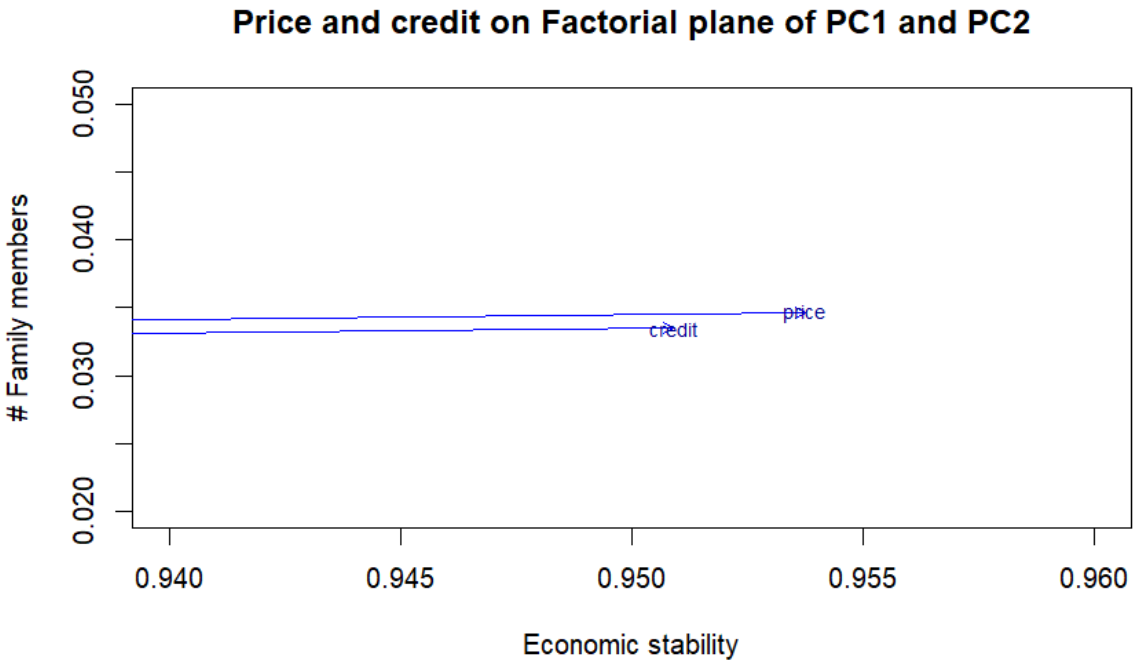


Figure 8. Projection of Price and Credit on the factorial plan, so observe the highly correlation between the metrics.

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.

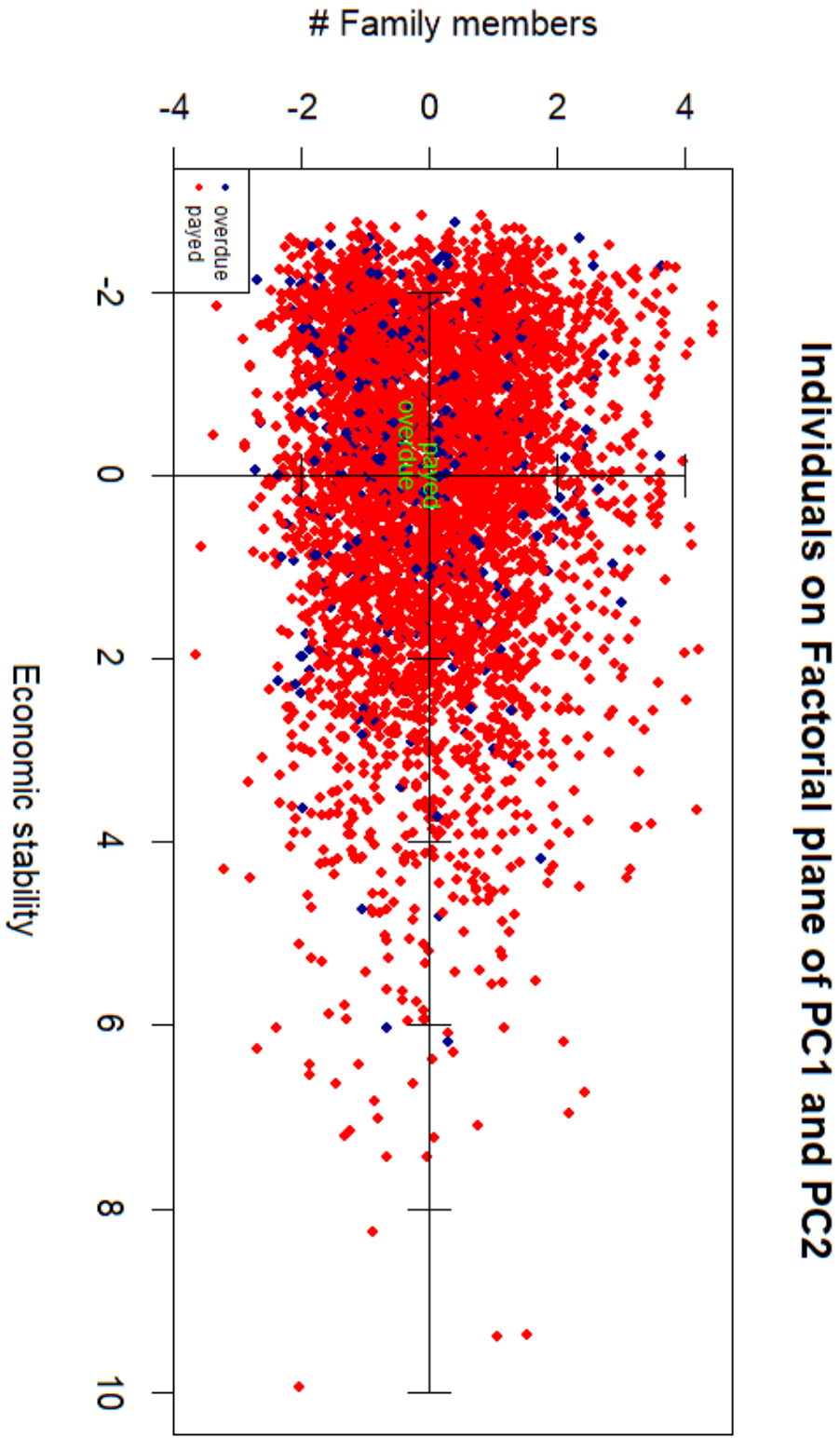


Figure 9. Target (paid or overdue) observation projecting on Factorial plane of PC1 and PC2

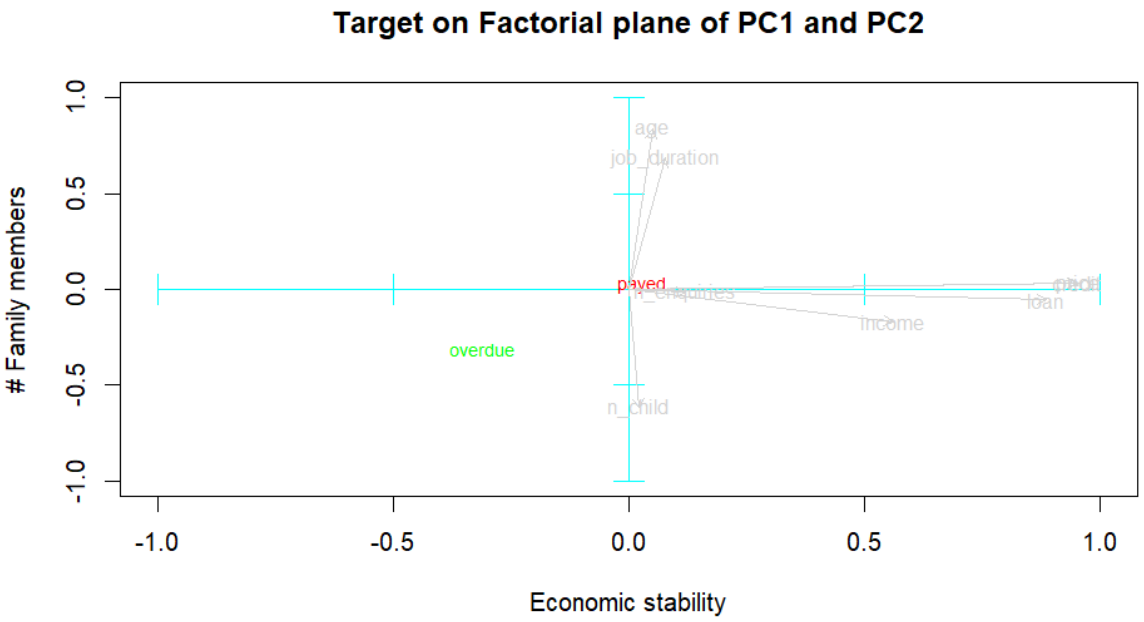


Figure 10. Target (paid or overdue) projection, by their centroid, on the factorial plane of PC1 and PC2

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.

```
    target    contract      gender         car    job_stat     studies      family       house
         2           3           4           5          11          12          13          14
occupation    job_type   companion
        17          18          20
```

```
> names(table(data$target))
[1] "payed"   "overdue"
> names(table(data$contract))
[1] "Cash loans"      "Revolving loans"
> names(table(data$gender))
[1] "F" "M"
> # check every modality
> names(table(data$target))
[1] "payed"   "overdue"
> names(table(data$contract))
[1] "Cash loans"      "Revolving loans"
> names(table(data$gender))
[1] "F" "M"
> names(table(data$car))
[1] "N" "Y"
> names(table(data$job_stat))
[1] "Commer. Assoc." "Pensioner"      "State servant"  "Working"
> names(table(data$studies))
[1] "Higher education"    "Low education"       "Secondary education"
> names(table(data$family))
[1] "Married" "divorce" "single"  "Widow"
> names(table(data$house))
[1] "Co-op apart."     "apartment"        "Municipal apart." "Office apart."
[5] "Rented apart."    "With parents"
> names(table(data$occupation))
 [1] "Administrative Staff" "Chef"                 "Core staff"
 [4] "Laborers"             "Managers"             "Medic stf"
 [7] "Occupation_Unknown"   "Private ser."         "Realty agents"
[10] "Sales staff"          "Security"             "Service Staff"
[13] "Tech Staff"           "Waiters"
> names(table(data$job_type))
 [1] "Agriculture"             "Business"                "Cleaning"
 [4] "Construction"            "Culture and Services"    "Government and Military"
 [7] "Housing"                 "Insurance"               "Jobtype_Unknown"
[10] "Medicine"                "Mobile"                  "Real Estate and Trade"
[13] "Religion"                "Security"                "Self-employed"
[16] "Transport"               "University"
> names(table(data$companion))
[1] "Companion_Unknown" "Family"            "Group_people"      "Other_companion"
[5] "Partner"           "Unaccompan."
```

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.

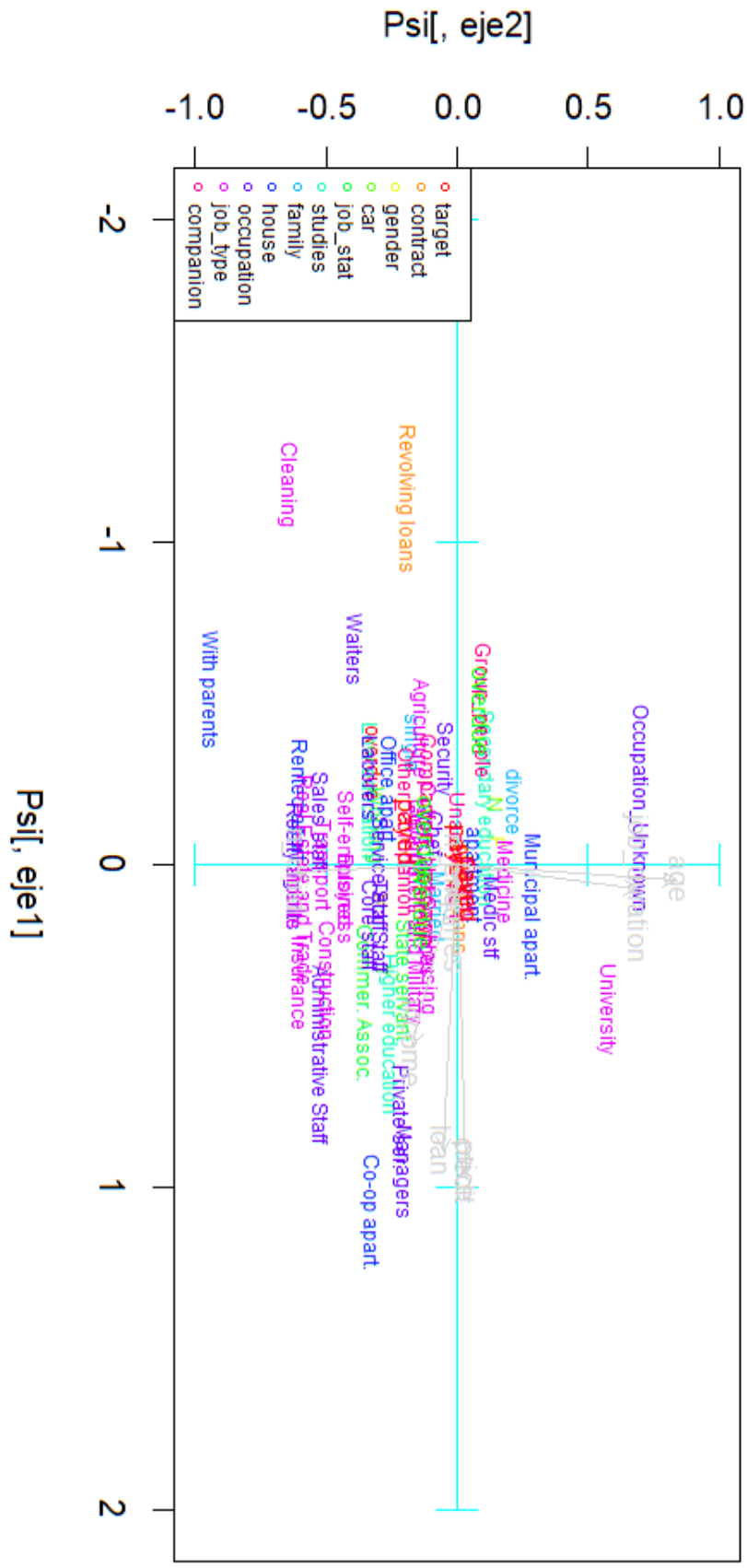


Figure 11. Modalities projection on the factorial plane of PC1 and PC2

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)}.

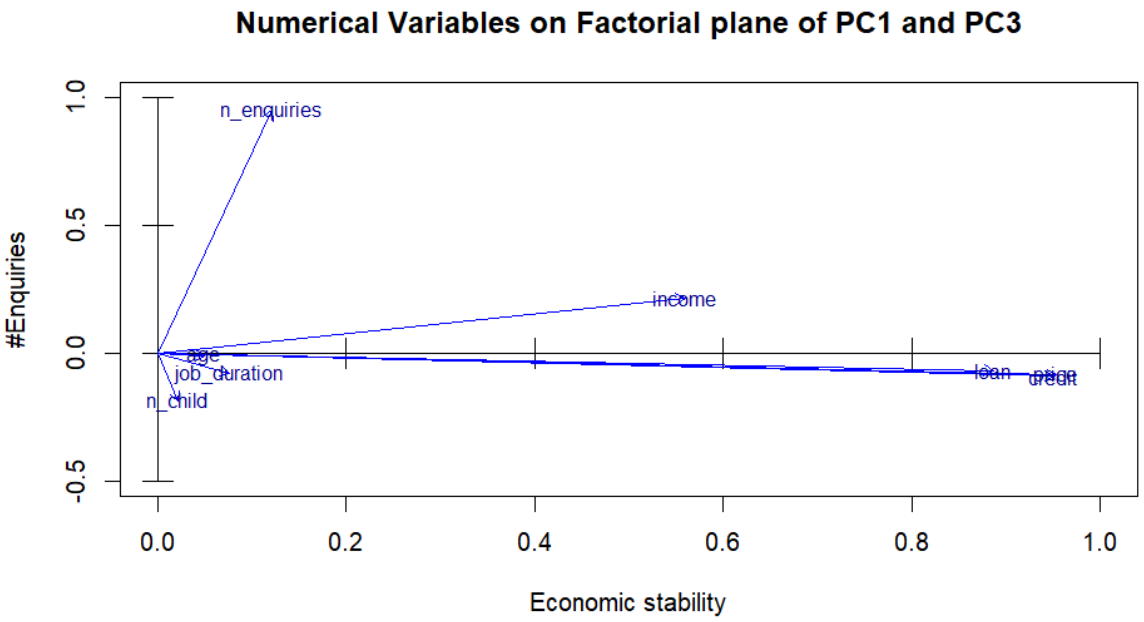


Figure 12. Projection of Numerical variable among the factorial plan, considering plane PC1 and PC3

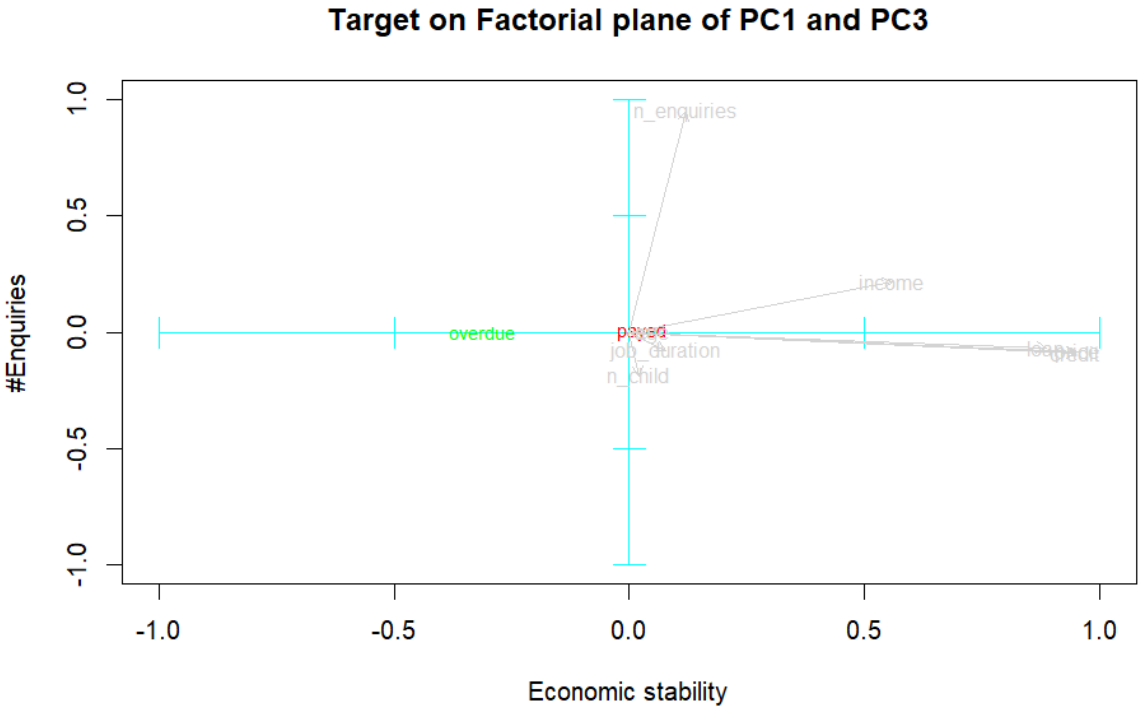


Figure 13. Target (paid or overdue) projection, by their centroid, on the factorial plane of PC1 and PC3

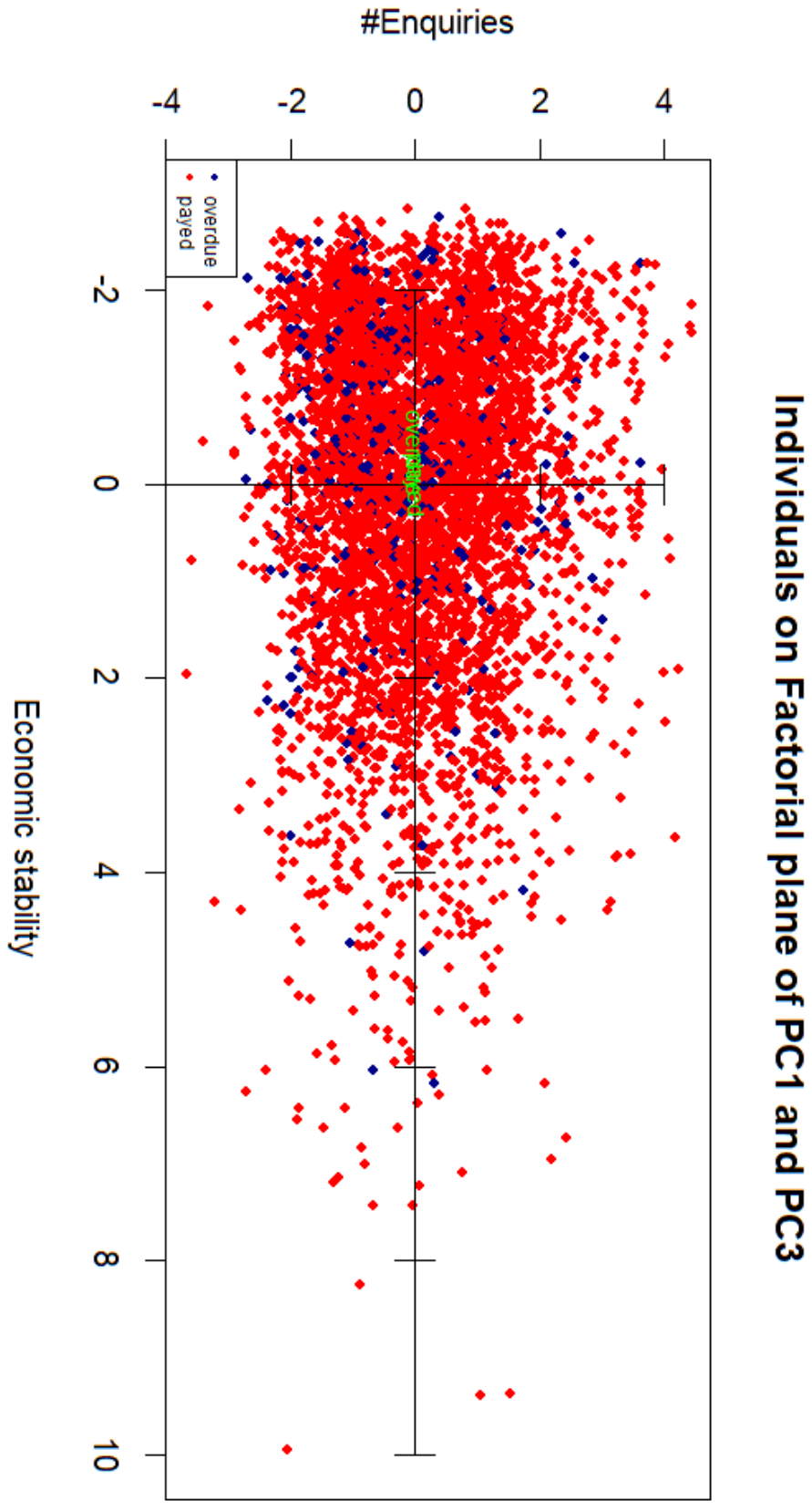


Figure 14. Target (paid or overdue) observation projecting on Factorial plane of PC1 and PC3

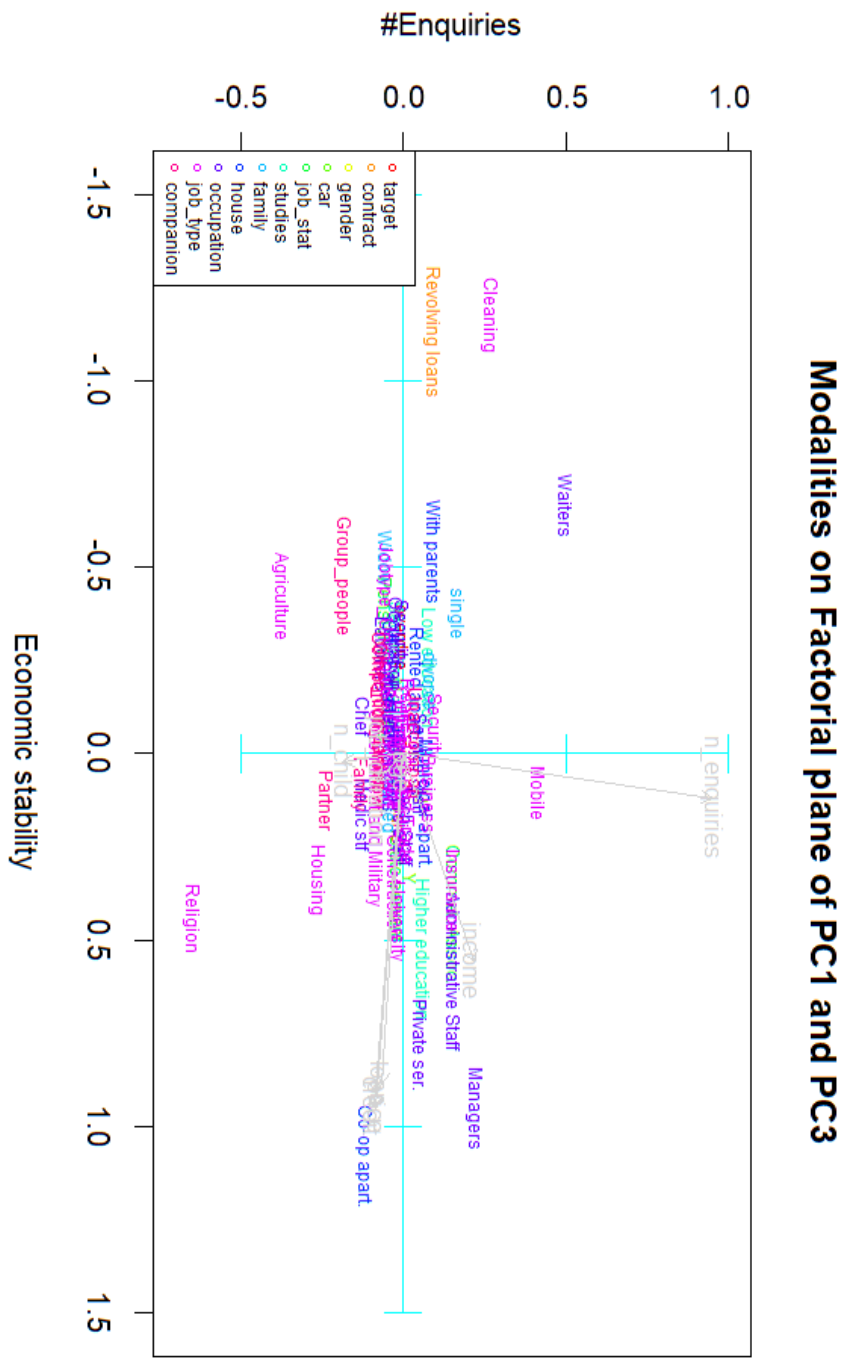


Figure 15. Modalities projection on the factorial plane of PC1 and PC3

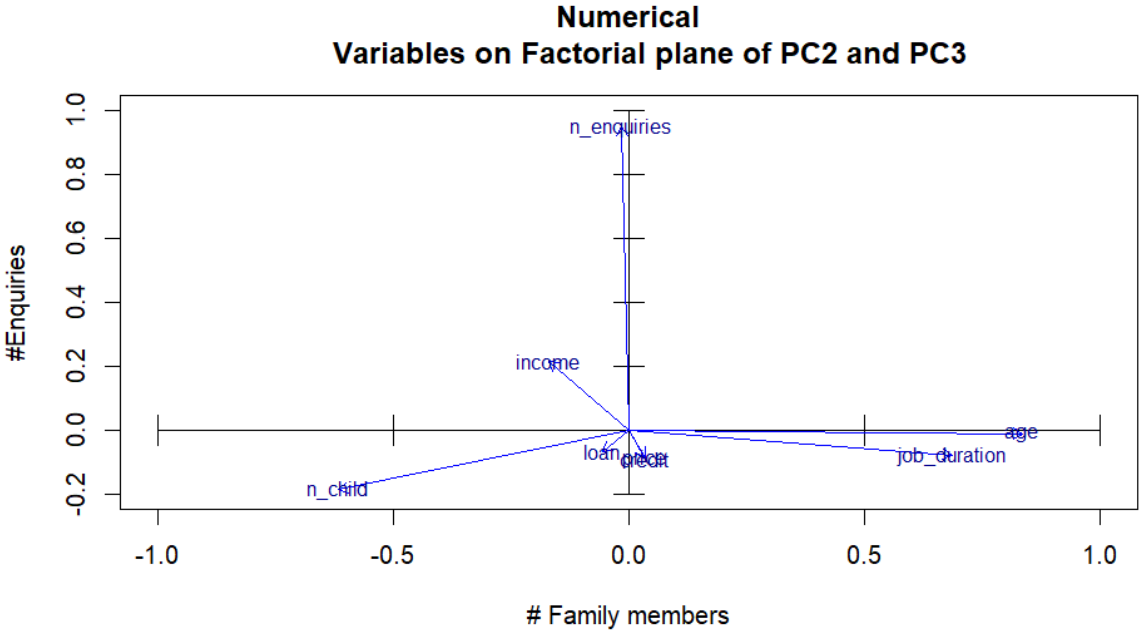


Figure 16. Projection of Numerical variable among the factorial plan, considesring plane PC2 and PC3

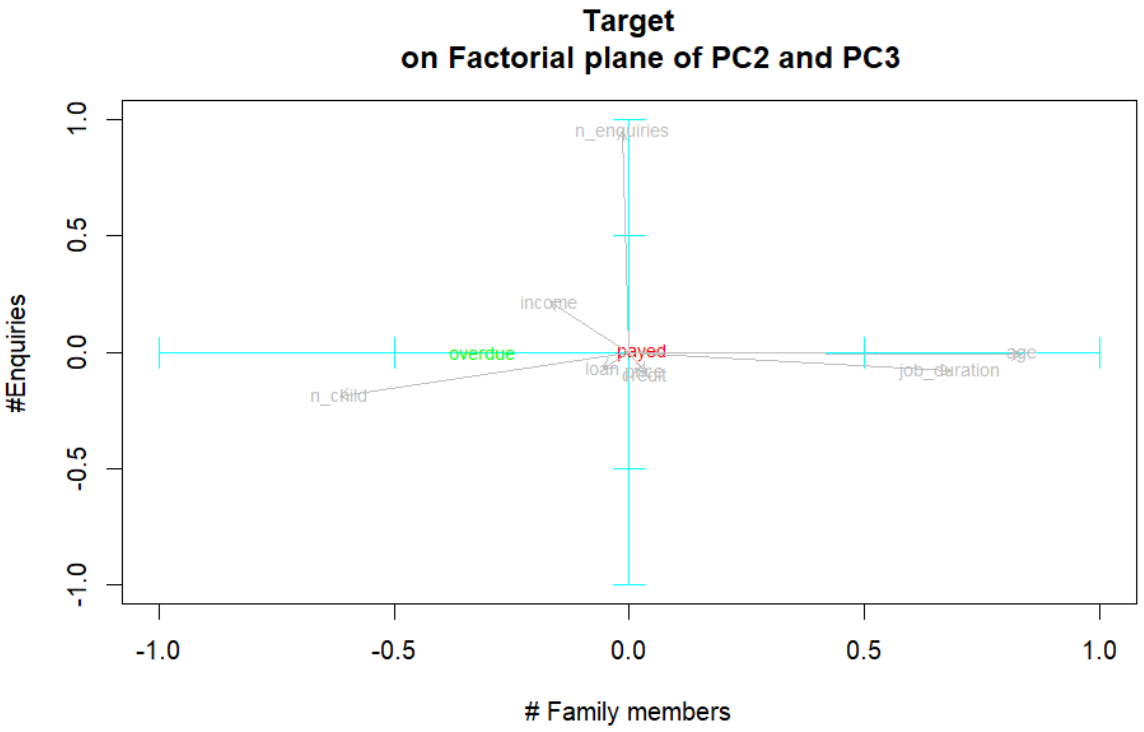


Figure 17. Target (paid or overdue) projection, by their centroid, on the factorial plane of PC2 and PC3

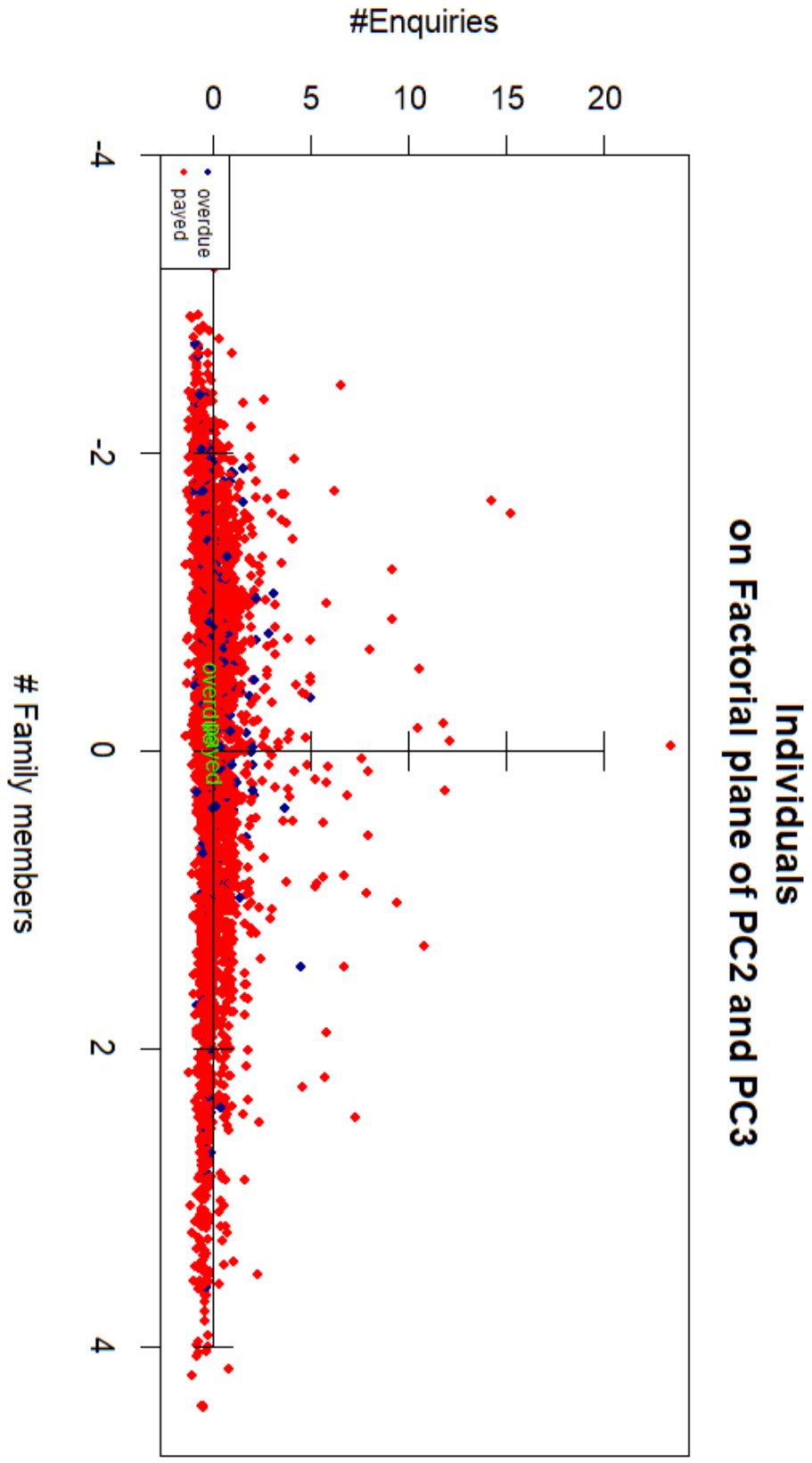


Figure 18. Target (paid or overdue) observation projecting on Factorial plane of PC2 and PC3

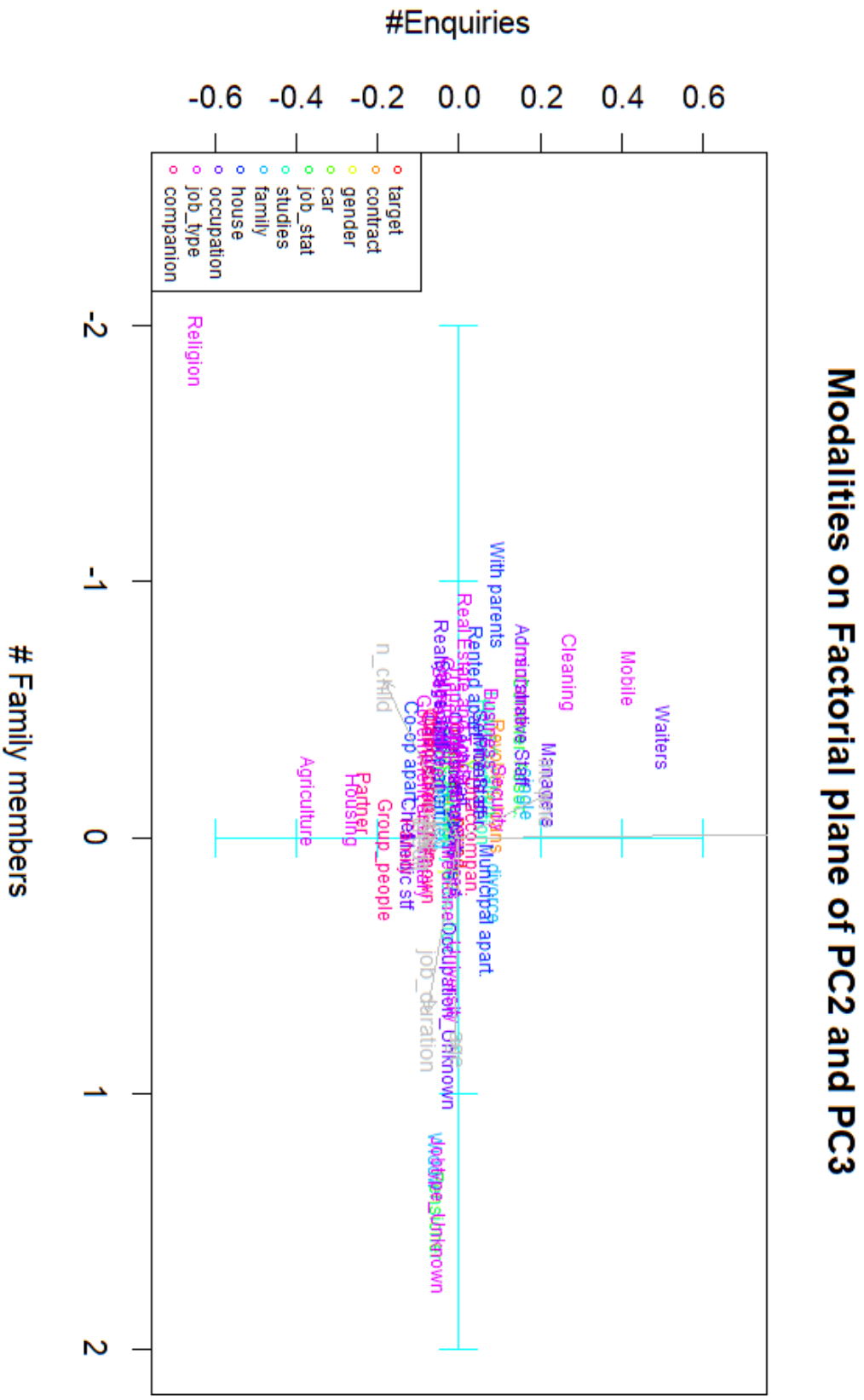


Figure 19. Modalities projection on the factorial plane of PC2 and PC3

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.

After this thorough discussion of the Principal Component Analysis we have performed, we will extract its main conclusions. We have discovered three relevant latent variables: wealth & loan, family & sociological status, and #enquiries. 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%.

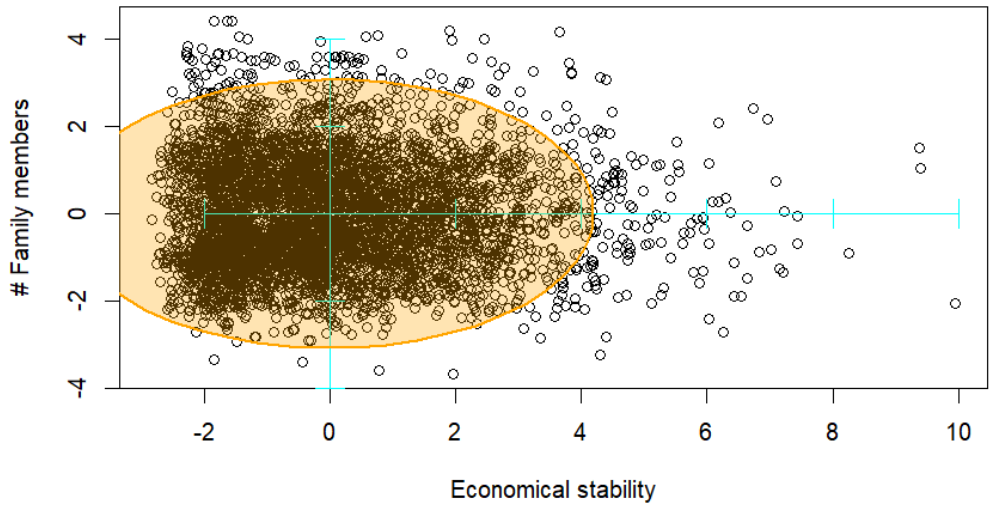


Figure 20. Target projection on the 1rst and 2n components, applying Mahalanobis distance as an ellipse to cover the confidence interval of 95%.

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

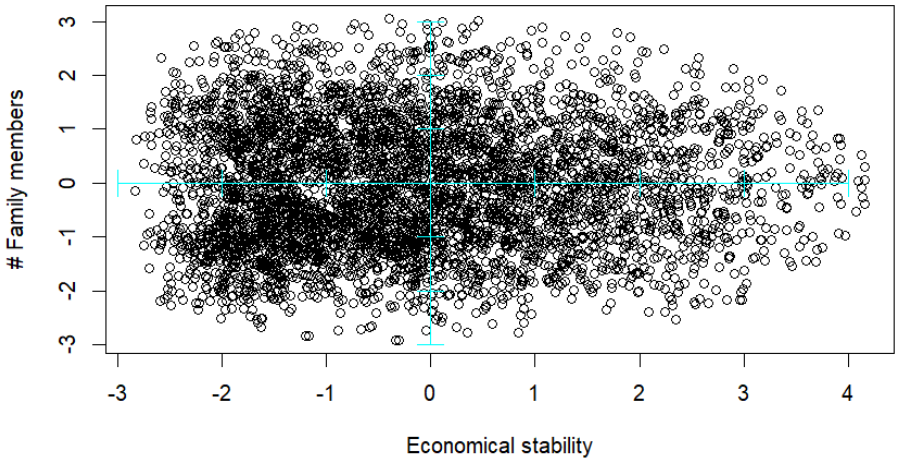


Figure 21. Treated datasets removing the outliers with Mahalanobis distances.

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:

target	contract	gender	car	job_stat	studies	family	house
2	3	4	5	11	12	13	14

occupation	job_type	companion
17	18	20

We have the following modalities for each variable:

```
> names(table(data$target))
[1] "payed"   "overdue"
> names(table(data$contract))
[1] "Cash loans"      "Revolving loans"
> names(table(data$gender))
[1] "F" "M"
> # check every modality
> names(table(data$target))
[1] "payed"   "overdue"
> names(table(data$contract))
[1] "Cash loans"      "Revolving loans"
> names(table(data$gender))
[1] "F" "M"
> names(table(data$car))
[1] "N" "Y"
> names(table(data$job_stat))
[1] "Commer. Assoc." "Pensioner"      "State servant"  "Working"
> names(table(data$studies))
[1] "Higher education"    "Low education"       "Secondary education"
> names(table(data$family))
[1] "Married" "divorce" "single"  "Widow"
> names(table(data$house))
[1] "Co-op apart."     "apartment"        "Municipal apart." "Office apart."
[5] "Rented apart."    "With parents"
```

```
> names(table(data$occupation))
 [1] "Administrative Staff" "Chef"                 "Core staff"
 [4] "Laborers"             "Managers"             "Medic stf"
 [7] "Occupation_Unknown"   "Private ser."         "Realty agents"
[10] "Sales staff"          "Security"             "Service Staff"
[13] "Tech Staff"           "Waiters"
> names(table(data$job_type))
 [1] "Agriculture"             "Business"                "Cleaning"
 [4] "Construction"            "Culture and Services"    "Government and Military"
 [7] "Housing"                 "Insurance"               "Jobtype_Unknown"
[10] "Medicine"                "Mobile"                  "Real Estate and Trade"
[13] "Religion"                "Security"                "Self-employed"
[16] "Transport"               "University"
> names(table(data$companion))
[1] "Companion_Unknown" "Family"            "Group_people"      "Other_companion"
[5] "Partner"           "Unaccompan."
```

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:

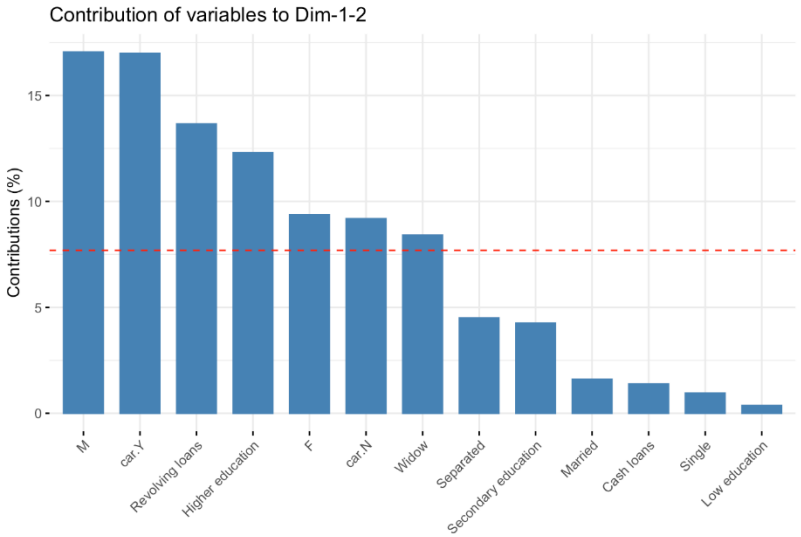


Figure 22. Histogram of contribution of each variable.

With the biplot we can extract the following information:

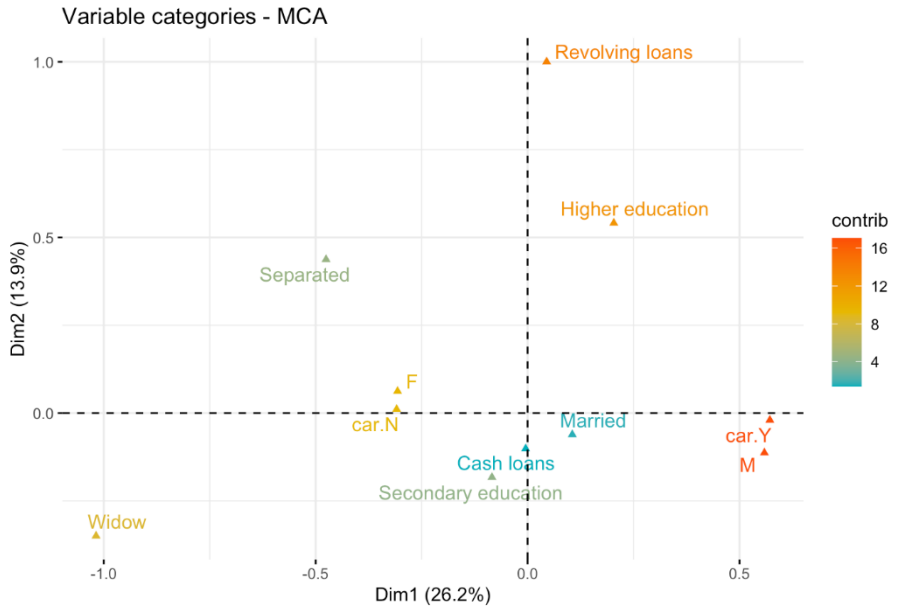


Figure 23. MCA plot of previous modalities.

- 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.

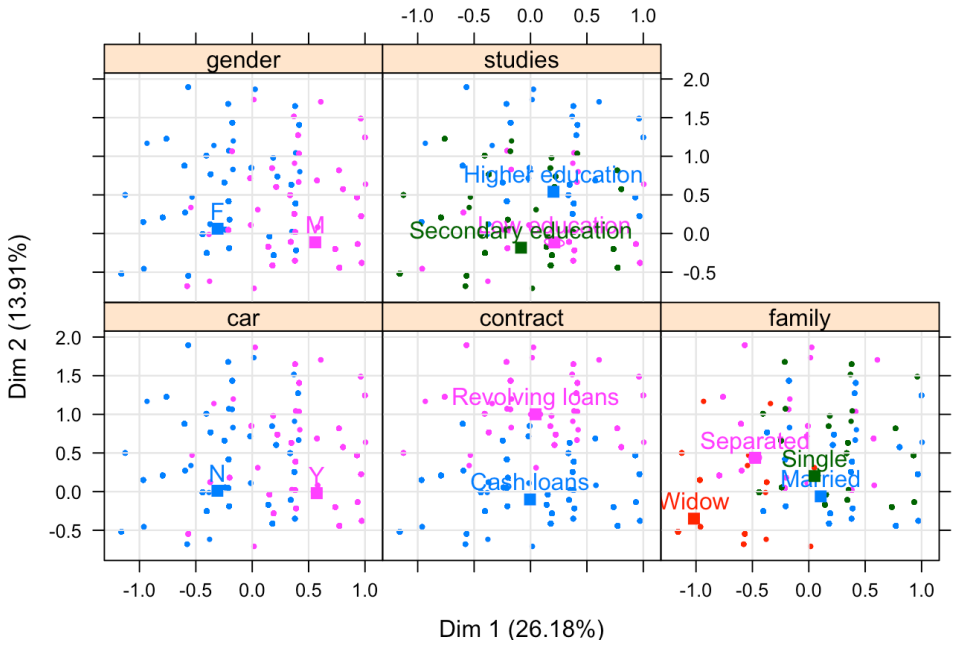


Figure 24. An MCA study carry out for each modality separately.

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

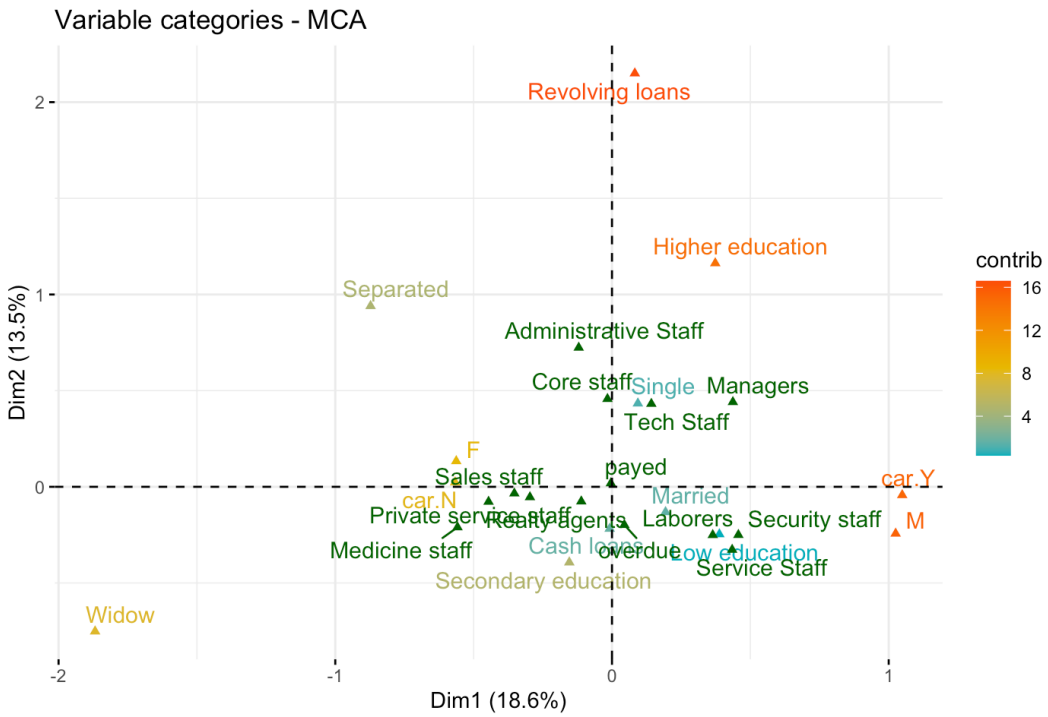


Figure 25. MCA plot, adding the Job_occupation modalities.

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.

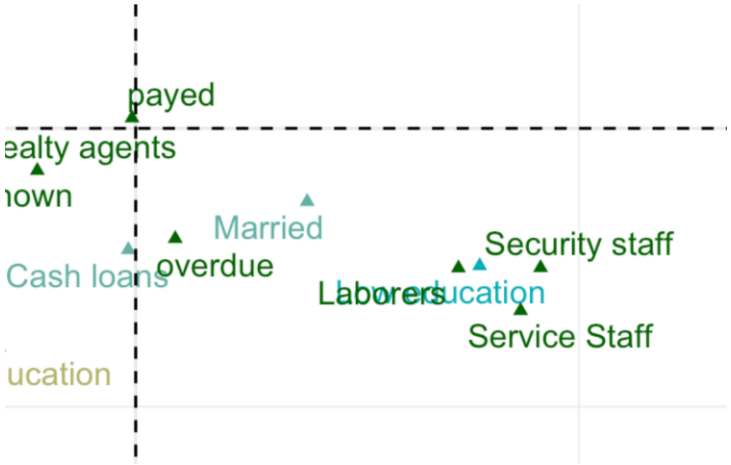


Figure 26. A zoom in of the previous MCP plot

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.

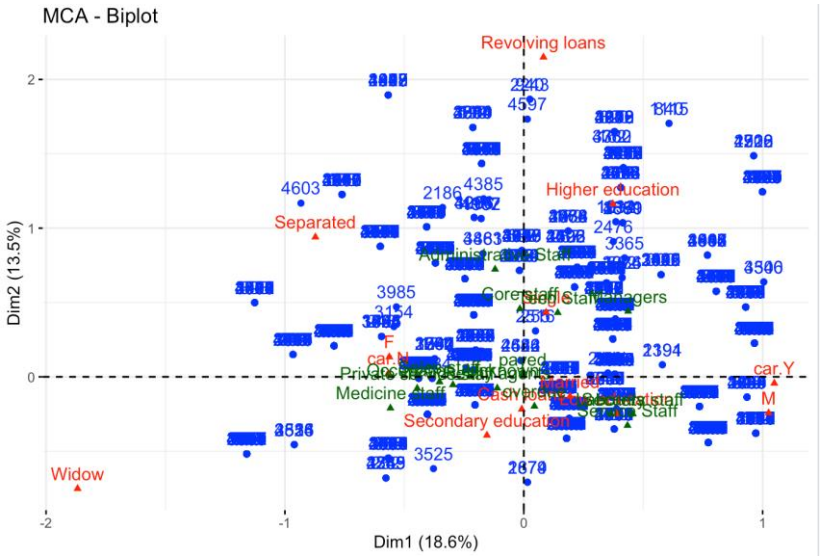


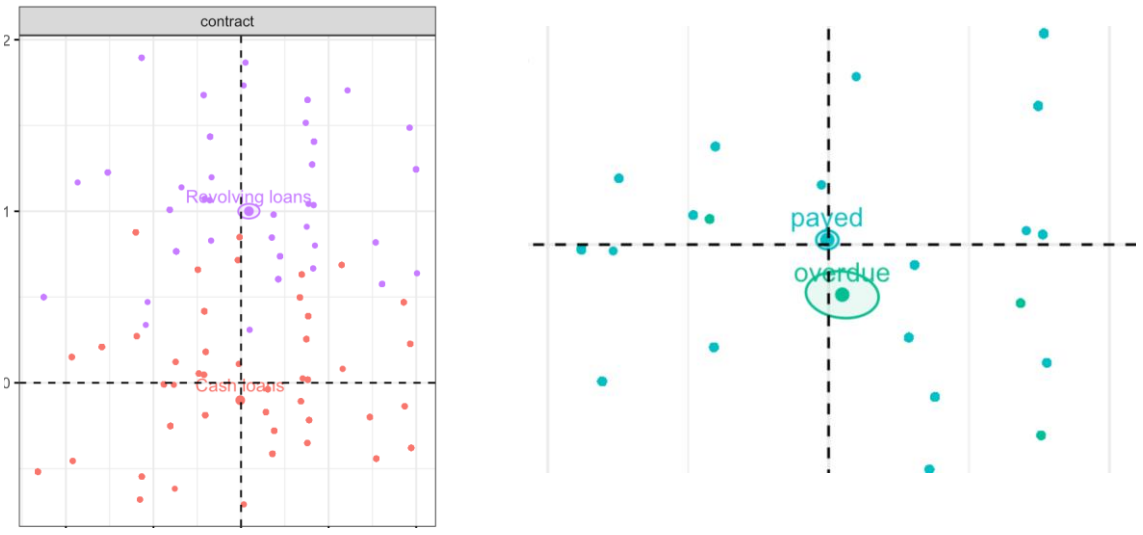
Figure 27. MCA plot with all the observation.

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.

Association rules mining analysis

In this part of the project, firstly, we analyzed the most common itemsets and discussed the threshold of lift, confidence, and support considered. Then, we analyzed the resulting rules and, lastly, we focused only on the rules that contained the target.

Analysis of the itemsets and parameters (confidence and support)

We perform association rules using ECLAT and apriori methods using all the categorical variables and obtaining 62 items (the figure can be seen in the Annex). Firstly, we analyzed the most common itemsets (Fig 29), and we saw that the two most common items were related to the loan, which is in 90% of the cases paid and of a type “cash” and is consistent with our database, that is highly unbalanced. Moreover, transactions normally have 4 elements; we will not have more than 8 items and less than 6 in 90% of the cases.

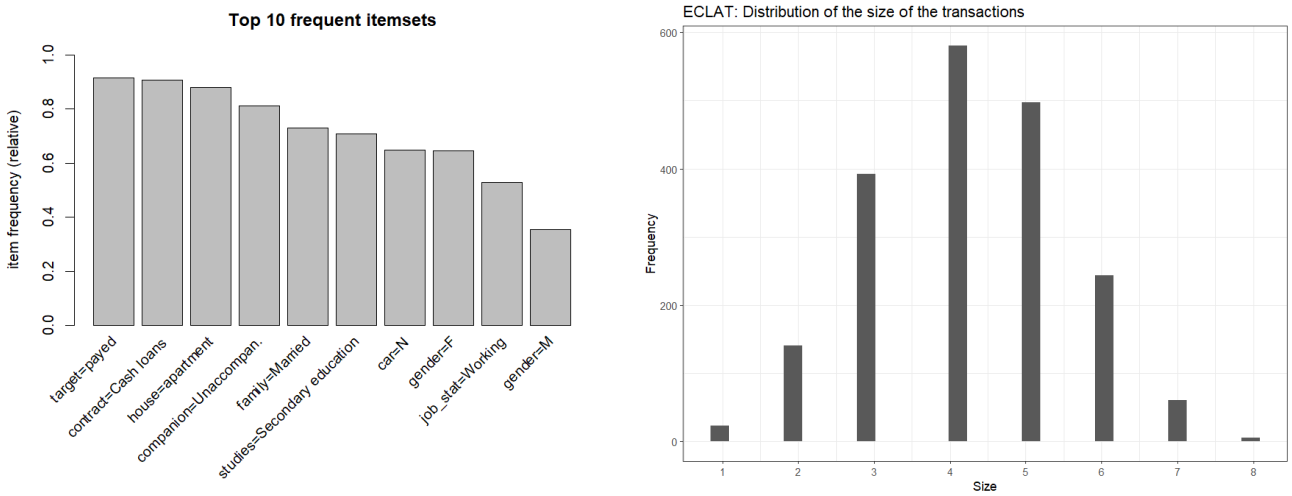


Figure 29. Most common itemsets, in the left, and size of the transactions using ECLAT.

Secondly, we focused on deciding the values of parameters to create the association rules. We decided to set high confidence (0.7), a typical value for making decisions in finances, and small support (0.01), a value that was obtained by plotting the numbers of rules and the minimum support (see Figure 30).

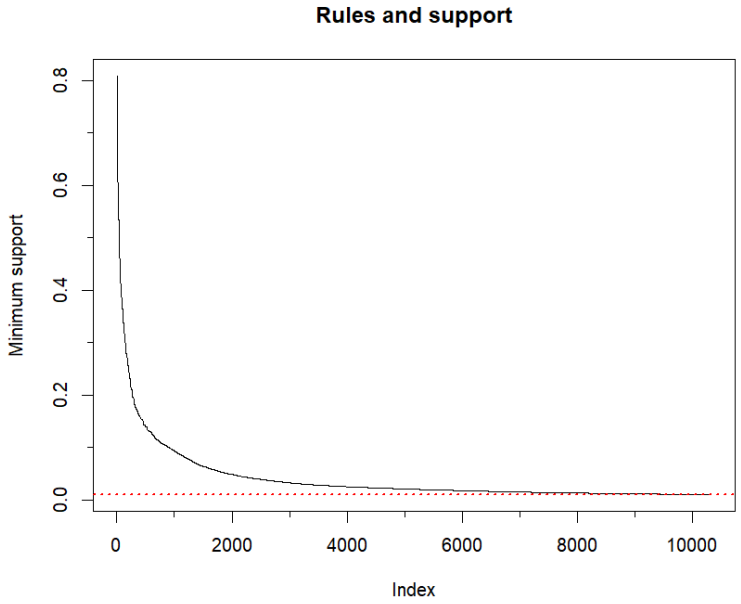


Figure 30. Total number of rules by reducing the minimum support and setting a confidence of 0.7. In red, you can see the minimum support used in the project (0.01).

General rules

After deciding the values of confidence and support we created rules and then selected only the ones that were not redundant and statistically correct, with a lift >= 3. This way we obtained 85 rules, described in Figure 31.

The rules with a higher lift (17) were trivial since they relate to people with the same occupation and job type (see rules with “medicine” in the RHS on the right of the figure). Some rules with perfect confidence were associated with “job_type = pensioners” and “occupation = unknown occupation”. Analyzing those rules we discovered that, in our dataset, all pensioners reported an *unknown_occupation*, which was reasonable as “working” is not an applicable variable for them. Note, however, that 675 individuals declared unknown occupations and were not pensioners, so not all unknowns in occupations are pensioners. The rest of the rules do not give us any additional information as they contain “job_type = pensioners” or “occupation = unknown” on the right-hand side.

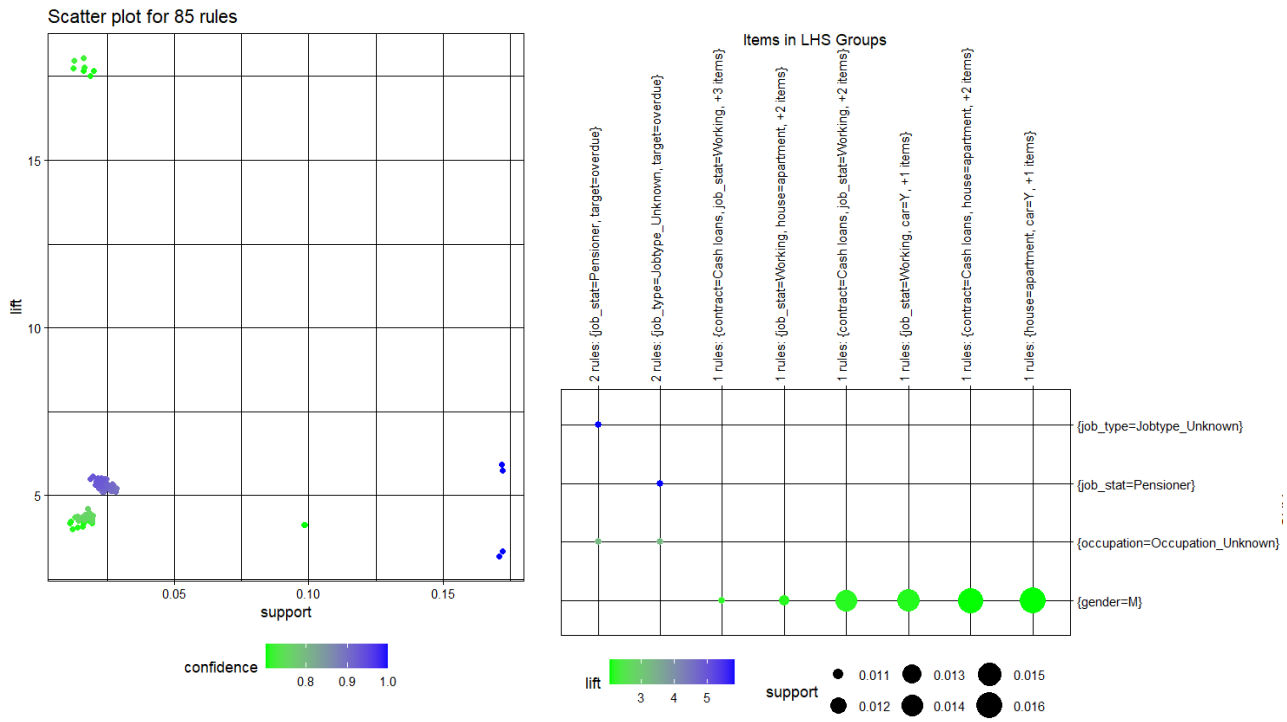


Figure 31. Rules that were not redundant with a lift bigger or equal to 3. On the right, there are represented the best 10 rules sorted by lift.

To gain more knowledge, we focused only on the simplest rules with 1 element on the right-hand side (Figure 32). Here, we saw that the best 4 rules were consistent with the fact that all pensioners have job_type unknown. The acceptable rules, with lift around 2, give us additional information on jobs that are normally done by men (security, service Staff, and construction). Additionally, we see that women tend to work in Medicine, Administrative, or as a chef, however, as the lift is too low we should not consider those rules.

Left-hand side ⇒	Right-hand side	Support	Confidence	Lift	Count
{job_stat=Pensioner}	{job_type=Jobtype_Unknown}	0.17	1.00	5.82	859
{job_type=Jobtype_Unknown}	{job_stat=Pensioner}	0.17	1.00	5.82	859
{job_stat=Pensioner}	{occupation=Occupation_Unknown}	0.17	1.00	3.26	859

{job_type=Jobtype_Unknown}	{occupation=Occupation_Unknown}	0.17	1.00	3.26	859
{occupation=Service Staff}	{gender=M}	0.06	0.78	2.21	306
{occupation=Security}	{gender=M}	0.02	0.74	2.08	78
{job_type=Construction}	{gender=M}	0.01	0.72	2.04	71
{occupation=Medic stf}	{gender=F}	0.03	0.97	1.50	152
{family=Widow}	{gender=F}	0.05	0.95	1.47	233
{occupation=Administrative Staff}	{gender=F}	0.04	0.94	1.46	176
{job_type=Construction}	{job_stat=Working}	0.02	0.77	1.45	75
{occupation=Laborers}	{job_stat=Working}	0.14	0.76	1.44	717
{job_type=Medicine}	{gender=F}	0.04	0.93	1.43	185
{occupation=Security}	{job_stat=Working}	0.02	0.75	1.43	80
{occupation=Sales staff}	{gender=F}	0.10	0.89	1.38	482
{occupation=Service Staff}	{job_stat=Working}	0.06	0.73	1.37	284
{job_type=Transport}	{job_stat=Working}	0.02	0.72	1.37	115
{occupation=Chef}	{gender=F}	0.01	0.88	1.36	74
{occupation=Chef}	{car=N}	0.01	0.87	1.34	73

{family=Widow}	{car=N}	0.04	0.87	1.33	212

Figure 32. Best 20 rules with a RHS of 2. The color denotes the quality of the rule: yellow (good), green (acceptable), and red (bad).

Rules of the target

Lastly, we focused only on the rules that were useful to predict the target. Firstly, we tried to predict the payers (Figure 33, right), but unfortunately, as the lift was too low those rules did not happen in reality. Secondly, we analyzed which variables were associated with payers (Figure 33, left), independently on the right-hand side of the rule. Here, even if the rules are correct, they do not give us any additional knowledge as only highlight trivial information about jobs or the relationship between pensioners and "Jobtype_Unknown".

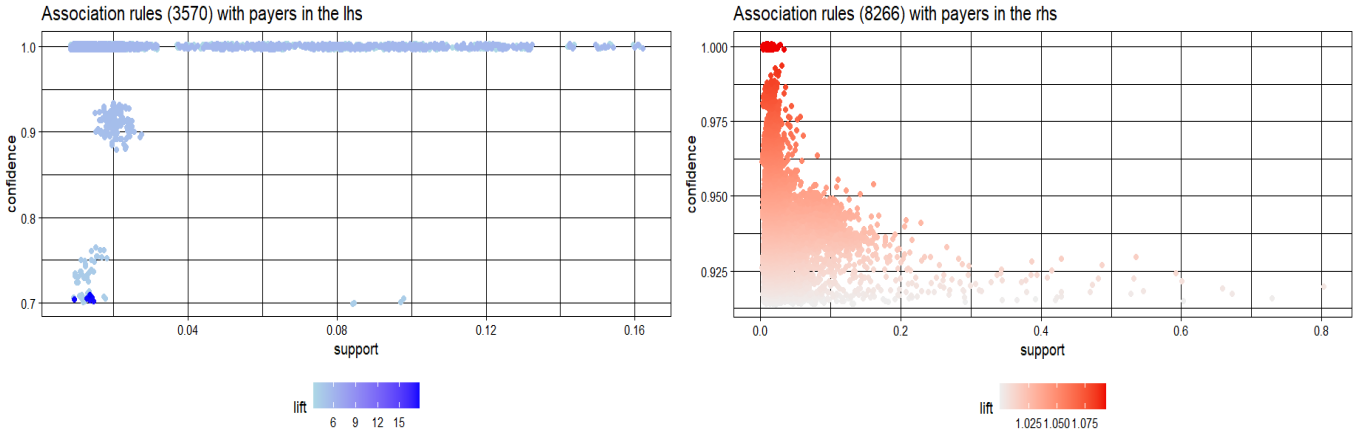


Figure 33: Rules with minimum support of 0.01 and setting target = "payers" on the right-hand side and the left-hand side.

Lastly, we tried to predict the people that did not pay the loan but there were no rules with a support of 0.01. To find some rules we decided to reduce the support to see how this increased the number of rules (Figure 34), here we see that to find some rules we need to set a support of at least 0.001. However, we decided not to report them as, even if some rules had a lift of more than 3, the rules only happened around 5 people in the database, so not very useful for predicting the debtors in reality.

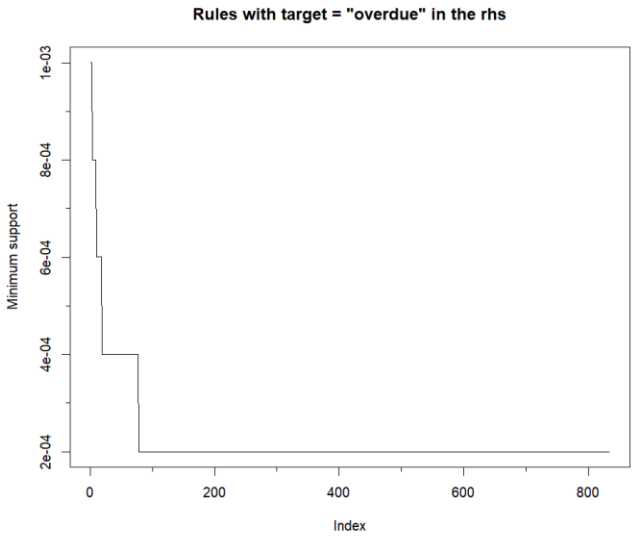


Figure 34: Total number of rules with target = "overdue" on the right-hand side by reducing the minimum support. Note that there are only 450 individuals that did not pay, so rules around 0.0001 are only seen in one individual, so can not be considered as a good association rule.

We also tried to put "overdue" in the left-hand side setting the minimum support back to 0.01. The result is Figure 35, which shows, again, trivial rules related to "pensioner" and "job type unknown", as we saw before. But more importantly, we discover that males tend to have a car, live in an apartment, work, and do not pay on time. Note, however, that those rules have a smaller lift, around 2, and a small support (0.02) which means that may be rare in reality even if those are correct.

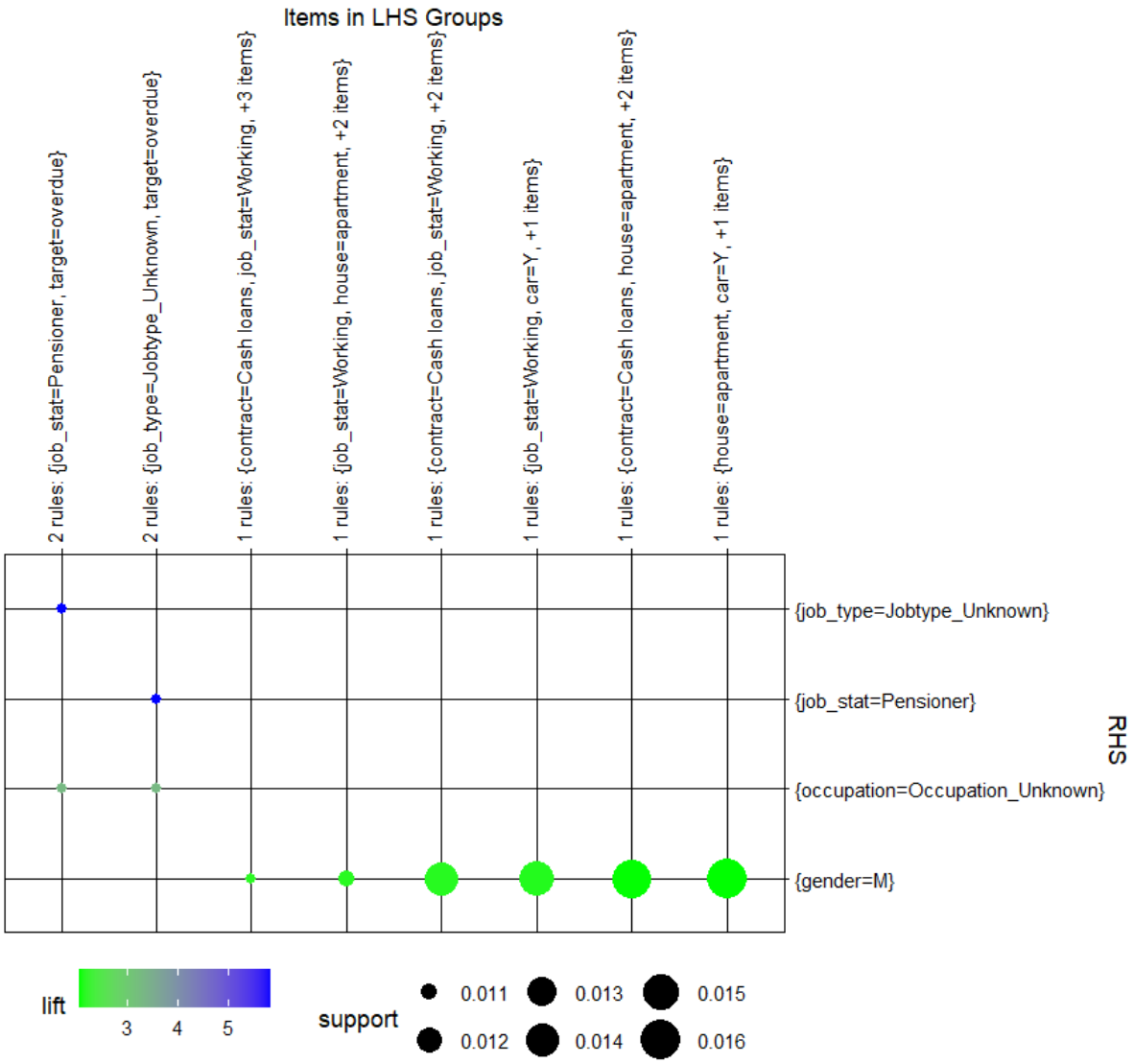


Figure 35: Rules with "overdue" in the left hand side

Hierarchical Clustering

In this part of the project, our initial step involves running the hierarchical clustering on the original dataset, using all the 19 variables. Gower distance is calculated for the variables, and the resulting dissimilarity matrix is squared. Subsequently, hierarchical clustering is performed

on the squared dissimilarity matrix. The clustering method employed is Ward's method. The resulting dendrogram is illustrated in Figure 37.

As captured in the figure, it is evident that the highest linkage height is observed for k=2, followed by k=4, and k=3 in third position. Firstly, we performed profiling for k=2, but we noticed a substantial loss of information when restricting the data on only two clusters. Thus, we decided to run additional profiling for k=3 and k=4 to capture a more comprehensive understanding of the inherent structures. Figure 36 presents a description of the cluster sizes for every used k.

Number of clusters	Size cluster 1	Size cluster 2	Size cluster 3	Size cluster 4
k=2	4141	858	-	-
k=3	3144	997	858	-
k=4	2206	997	938	858

Figure 36: Description of the cluster sizes for k=2,3,4.

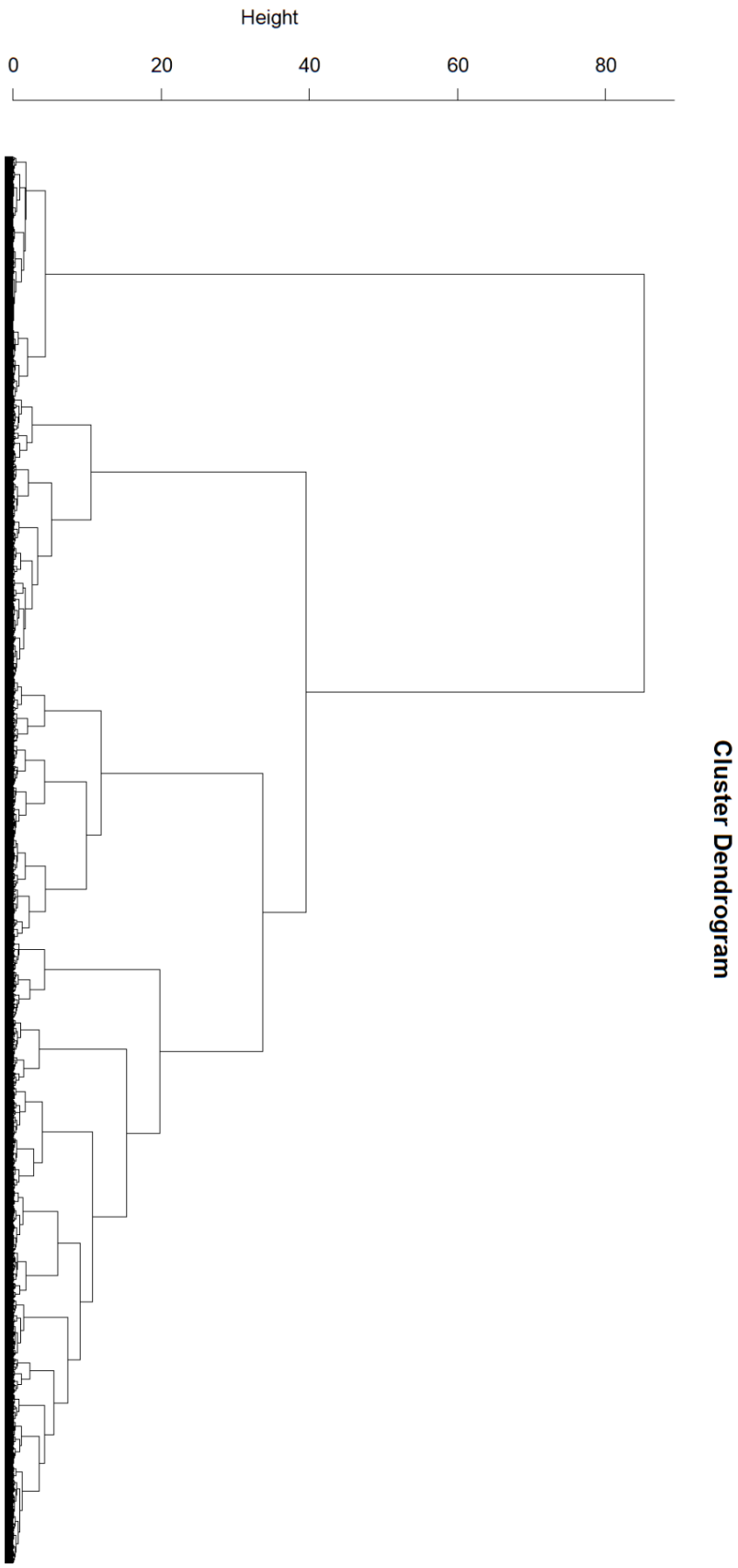


Figure 37: Dendrogram resulted from Hierarchical Clustering.

Profiling of Clusters

For k = 2, the clusters primarily differentiate based on the age of individuals, with cluster number 2 comprising pensioners with an unknown job occupation. Cluster number 1 includes young individuals and adults residing in apartments or with their parents. However, due to the broad nature of these clusters and the similar distribution of the target variable in both, we opted to focus our profiling analysis on a higher number of clusters.

When analyzing profiling with k = 3, we noticed that there is an aggregation of two individual types with different education levels (see Figure 38) on the same cluster 1, which led us to think that they could belong to two independent groups. Hence, we performed profiling analysis with k = 4. This resulted in the partition of individuals in cluster 1 into two different sub-clusters (clusters 1 and 3 in the right-side plot) according to their education levels, with higher educated individuals in cluster number 3.

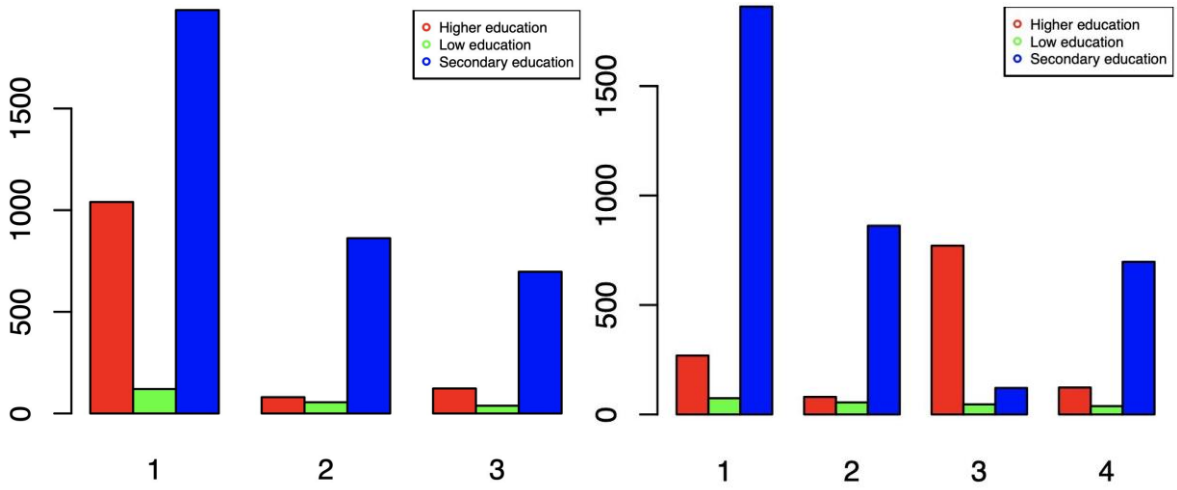


Figure 38: Comparison of barplots for the studies variable between k=3 and k=4.

As captured in the table below, we performed different statistical tests to analyze the conditional distribution of variables to clusters. We chose the Chi Square independent test for the qualitative variables and Kruskal-Wallis for the numerical variables. Analyzing the p-values, we can notice that all of them are significant.

Variable	Type	Test	p-value
target	qualitative	Chi Square	<2,2E-15
contract	qualitative	Chi Square	<2,2E-15
gender	qualitative	Chi Square	<2,2E-15
car	qualitative	Chi Square	<2,2E-15
n_child	numerical	Kruskal-Wallis	2,65E-85
income	numerical	Kruskal-Wallis	1,64E-111
credit	numerical	Kruskal-Wallis	4,35E-25

loan	numerical	Kruskal-Wallis	7,95E-53
price	numerical	Kruskal-Wallis	4,93E-30
job_stat	qualitative	Chi Square	<2,2E-15
studies	qualitative	Chi Square	<2,2E-15
family	qualitative	Chi Square	<2,2E-15
house	qualitative	Chi Square	<2,2E-15
age	numerical	Kruskal-Wallis	0,00E+00
job_duration	numerical	Kruskal-Wallis	2,54E-58
occupation	qualitative	Chi Square	<2,2E-15
job_type	qualitative	Chi Square	<2,2E-15
n_enquiries	numerical	Kruskal-Wallis	3,37E-03
companion	qualitative	Chi Square	2,15E-09

Figure 39: Variable type, test used and resulting p-value for all variables with k=4.

Right after that, we analyzed each variable one by one, and picked those that give more explainability among each cluster. Next, we performed individual statistical tests to assess which variables are significant in each cluster. These tests aimed to evaluate the null hypothesis, investigating whether the global mean is equal to the local mean within each cluster.

					P - values			
Variable	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
cluster size	2206	997	938	858	-	-	-	-
Target	Payed: 87,22%	Payed: 96,99%	Payed: 93,50%	Payed: 93,70%	0	0	0	0
	Overdue: 12,78%	Overdue: 3,01%	Overdue: 6,50%	Overdue: 6,30%				
Gender	Mostly female (87,08%)	Mostly male (97.69%)	Mostly female (62.90%)	Mostly female (80.77%)	0	0	0	0
Car	No (76.11%)	Yes (61.28%)	No (52.13%)	No (80.54%)	0	0	0	0
Child (mean)	0,4914	0,4447	0,5789	0,0303	3,26E-10	9,10E-02	5,75E-14	0
Income (mean)	152103	185769	210535	133146	0	9,04E-14	2,92E-58	0
Job status (commercial, pensioneer,	568	245	313	0	0	0	0	0
	1	0	0	858				

state servant, Working)	107	28	232	0				
	1530	724	393	0				
Study	Sec. edu (84.45%)	Sec. edu (86.46%)	High Edu (82.19%)	Sec. edu (81.23%)	0	0	0	0
Family stats.	Single (14%)	Single (25%)	Married (85%)	Widow (16%)	0	0	0	0
Age (year, mean)	40.59	38.81	39.17	59.25	0	0	0	0
Job duration (year, mean)	6.79	5.5	6.83	10.89	3,58E-05	0	2,54E-02	1,74E-06
Job occupation	Sales (19.72%)	Laborers (41.02%)	Core staff (22.7%)	Unknow (100%) (pensioner)	0	0	0	0

Figure 40: Statistics and p-values for every cluster for significant variables for k=4.

Analyzing Figure 40 and the profiling plots, we can try to extract some basic features that describe each cluster.

Cluster 1 and **Cluster 2** are from **similar social classes**, but **distinguished** by **gender.** This is illustrated in the left-side of Figure 41, where we notice that clusters 1 and 4 are mostly composed of women, whereas cluster 2 predominantly includes men. Moreover, in clusters 1 and 4 the majority of the individuals do not own a car (see Figure 41 right), which may be attributed to cultural factors, such as Indian women being less likely to own cars.

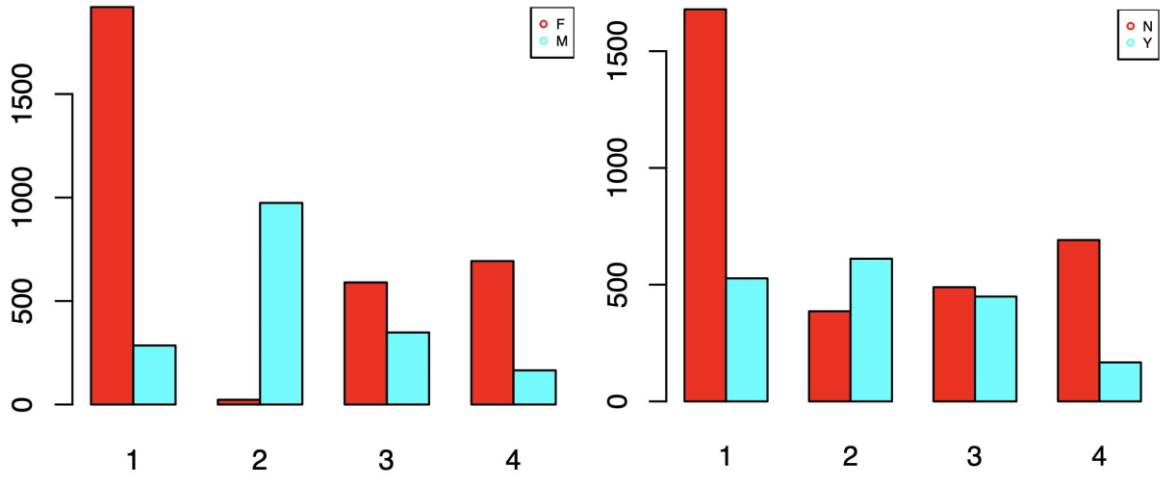


Figure 41: Gender (left) and car owning distribution (right) for k=4.

Their **education** levels are **similar** (Figure 44), with the vast majority having only secondary education (84% and 86% respectively). And even though their **income has some discrepancy** (152k vs. 185k) (Figure 42), this could be a result derived by gender inequality in their society.

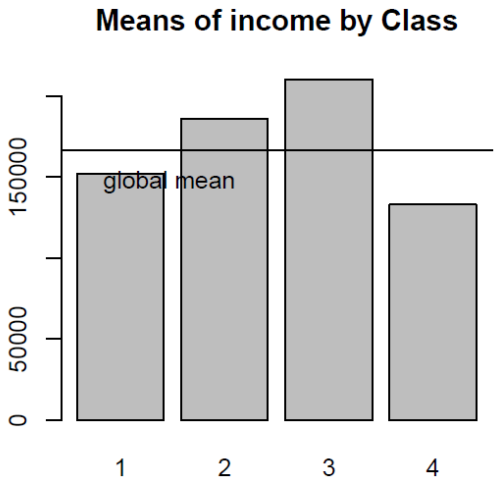
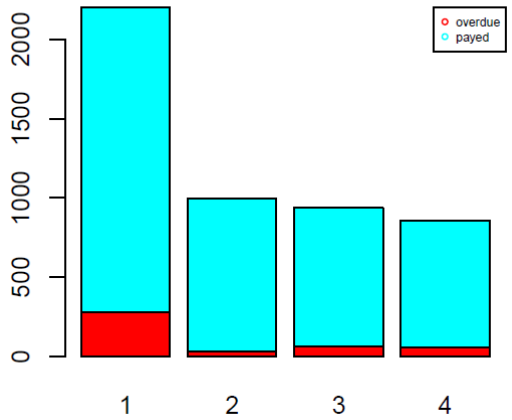


Figure 42: Average values of the income by class for k=4.

And in the target level, the **Cluster 1 tends to overdue the payment (12,78%) significantly more than the Cluster 2 (3%).**



Cluster	Payed [%]	Overdue[%]
1	87.22	12.78
2	96.99	3.01
3	93.50	6.50
4	93.70	6.30

Figure 43: Probability of payed and overdue by class for k=4.

In the **Cluster 3**, the **gender** is more **homogeneously distributed**. Even though there is 62.9% female (consequence of unbalanced gender distribution), the relative frequency is 18.28% of females and 19.64% of men belong to this cluster.

We believe that this cluster represents the grouping of individuals who belong to a **higher social-economic** atmosphere, with higher income (210.535) and education (82.19% High Education level) (Figure 44). This also can be observed by their job type, where core staff represents 22.7% of all the occupations. On the other hand, they have a more stable familiar situation (85% married, with highest nº child mean). In the target level, the percentage of overdue is 6.5%.

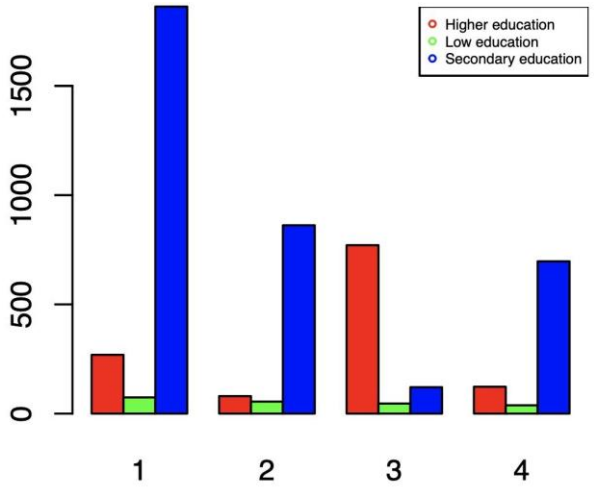


Figure 44: Education levels distribution of each class for k=4..

In the **Cluster 4**, it is a grouping of retired people, whose income is much lower (133k) and the jobs status are exclusively pensioners (below Figure 45). The average age of this cluster is 59 years old (Figure 46), and the percentage of widows is significantly higher than another cluster. In the target level, the percentage of overdue is 6.3%.

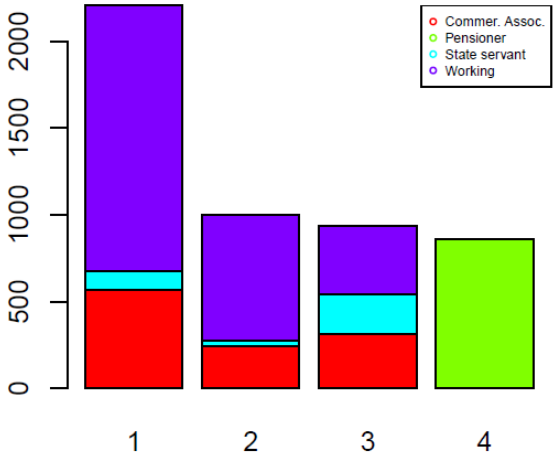


Figure 45: Job status distribution of each class for k=4..

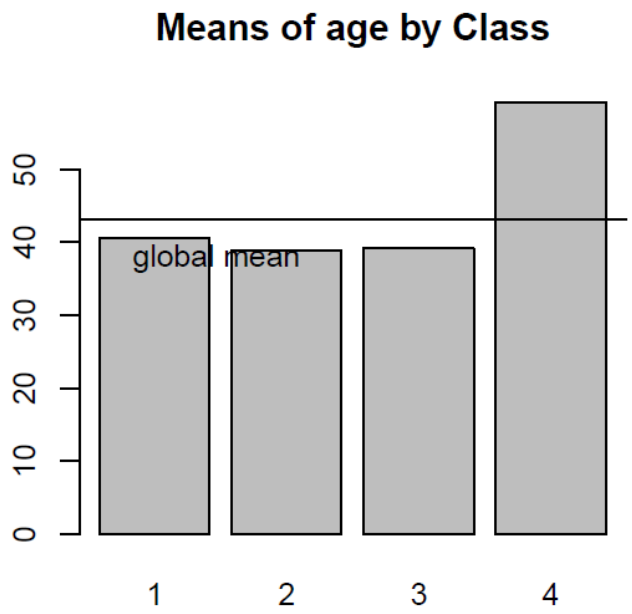


Figure 46: Mean of age by class for k=4.

And finally, extracting all these characteristics, we can conclude it at below table and synthesis.

Class	Description of the Class	Subjective Business Suggestion
C 1	Biggest cluster size (2206 observation) Mostly women who do not own a car. They have a relatively low income and a low education level. Their occupation types are mostly related to the service field as Sales, Laborer etc. **Overdue: 12.78%**	Be careful with accepting the application.
C 2	Mostly men with the second highest mean income and low education level. But they are reliable by returning the loan. **Overdue: 3%**	High-quality client.
C 3	Cluster with highest income and highest education level. Occupation types are mostly related to high prestige jobs. Socio-economically stable. **Overdue: 6.5%**	High-value clients, but are not very punctual with payment.
C 4	Oldest people with lowest income, and are innestable. Mostly women composed of retired pensioners who do not own a car. **Overdue: 6.3%**	Low-value clients. Do not recommend accepting high value loans.

Table 47: Summary description of each class and subjective business suggestion for k=4.

Decisions tree (CART-Classification And Regression Trees)

Constructing a decision tree model is particularly interesting due its **explicability** characteristics. We created a **classification tree** because our target variable is categorical, precisely binary.

Firstly, as we have observed, our target is **highly unbalanced** (91 '4% payed, 8,6% overdue), therefore we performed a balancing task, experimenting with three different techniques: undersampling, oversampling, and a combination of both. We opted for the combination of both techniques because it provided better results.

To create an optimal model, we must fine-tune the following **hyperparameters**: cp, maxdepth, minsplit, minbucket and determine which values yield the best results. For the best **cp** (complexity parameter) we performed a search to compute the most optimal one for the tree. Our algorithm determined that the best cp is 0.01305684. With all the other parameters predetermined, our classification tree is Figure 48.

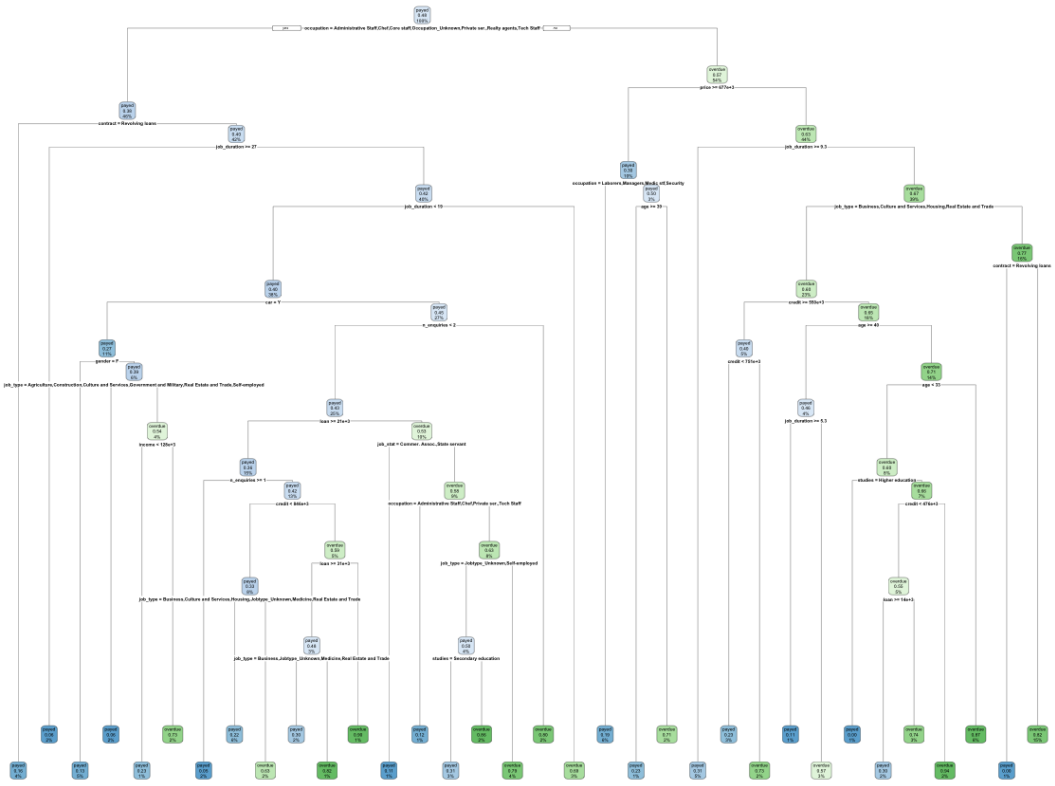


Figure 48: Optimal model with cp is 0.01305684

We can observe that most of the terminal nodes have a darker color (green and blue), indicating high purity. Additionally, the tree is quite large in depth, prompting consideration of the possibility of overfitting.

The most important variables are summarized in Table 49.

```
importance[importance >= 1]
    credit        price   occupation     job_type job_duration         loan          age
      13.8         11.0         10.0         10.0          9.6          8.8          8.3
  contract     job_stat      studies       income  n_enquiries        house          car
       5.2          4.8          4.5          3.4          3.2          1.9          1.8
    family       gender
       1.5          1.3
```

Figure 49: Most important variables

It is interesting to describe the **paths** that end with a high pure node and high percentage of data in the node. In this tree the most interesting paths are :

Path 1 (Figure 50): We have a 82 % pure node that contains 15% of the data.

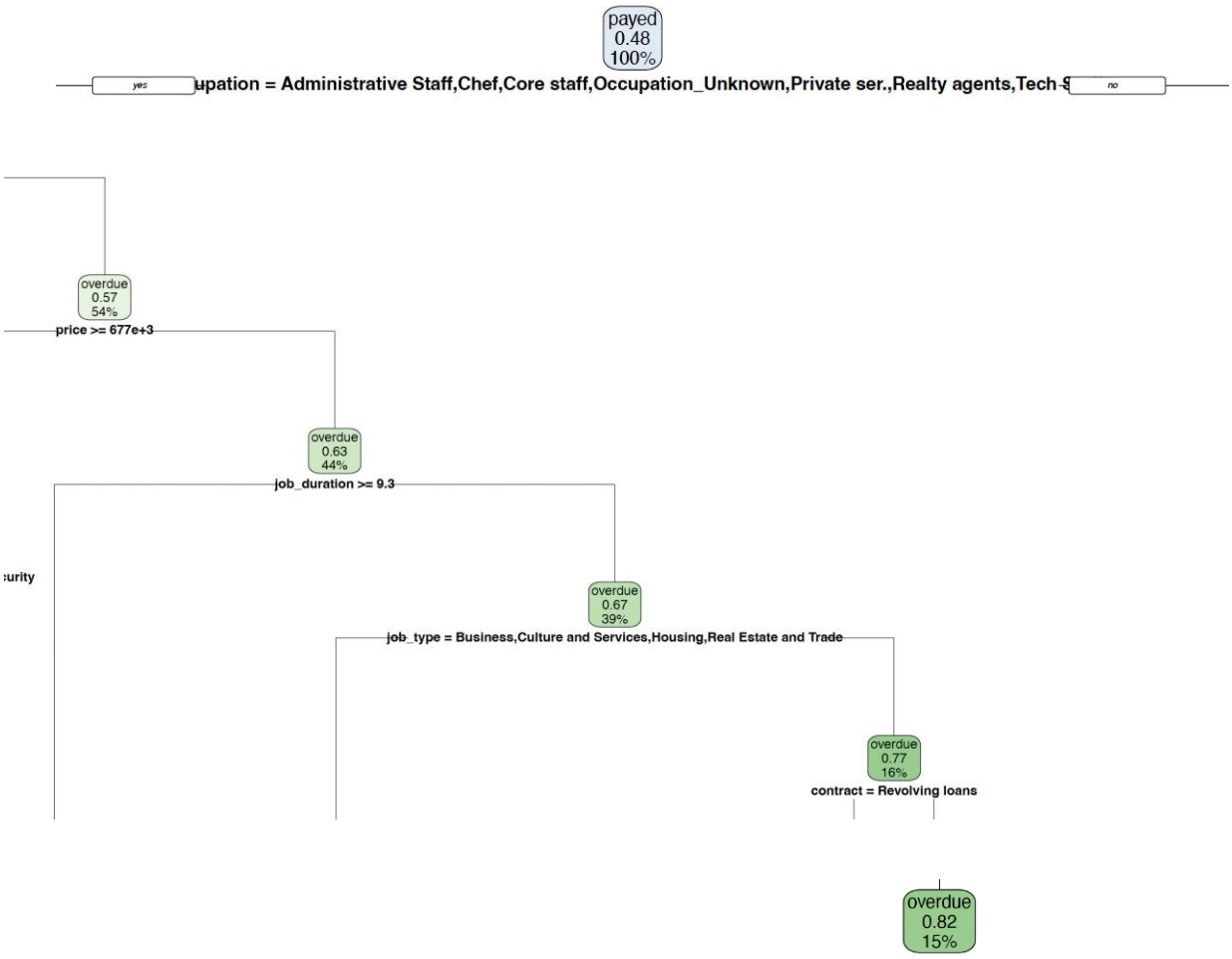


Figure 50: Path 1

This path contains the individuals that: respond *no* to all the dichotomies (the path on the right). These individuals are predicted to overdue the loan.

Path 2 (Figure 51): path that is 99% pure :

```
target is 0.99 when
    occupation is Laborers or Managers or Medic stf or Sales staff or Security or Service Staff or Waiters
    job_duration < 9.3
    job_type is Business or Culture and Services or Housing or Real Estate and Trade
    price < 677250
    credit is 476010 to 550422
    age < 33
    studies is Low education or Secondary education
```

Figure 50: Path 2

Certainly, we can affirm that young individuals employed in the business sector, with substantial bank savings (exceeding 550,000), and seeking a loan amounting to no more than 25% of their savings consistently fulfill their loan payments.

Path 3 (Figure 51):

```
target is 0.00 when
    occupation is Laborers or Managers or Medic stf or Sales staff or Security or Service Staff or Waiters
    job_duration < 9.3
    job_type is Construction or Government and Military or Medicine or Security or Self-employed or Transport
    contract is Revolving loans
    price < 677250
```

Figure 51: Path 3

Therefore individuals with occupations related to service, construction, medicine with low job duration that apply for revolving loans with prices lower than 677K tend to not pay on time the loan most of the time.

We shall compute the accuracy of the tree on both the training and test samples to assess its **predictive power** and determine if there is overfitting in the model. We have an accuracy of **79,4%** in the train sample.

With our classification tree model, we predict the values of the test sample and compute the accuracy.
We can see a decrease in accuracy from 79% to **64%**, indicating clear **overfitting** in our decision tree. Now, we will attempt **pruning** to create a more general tree that works good for other samples, not just our training set.

We need to reduce the length of the tree by adjusting the **'maxdepth'** parameter. We are looking for a value that:

- minimizes the difference between the accuracy of the train sample and the test
- maximizes each accuracy value

For maxdepth=6 we have

- train accuracy: 70%

- test accuracy: 68%

Although the purity of the terminal nodes is lower, it is a more general tree that works best for new possible different individuals wanting to apply to the loan. And also, has a better explicability (Figure 52).

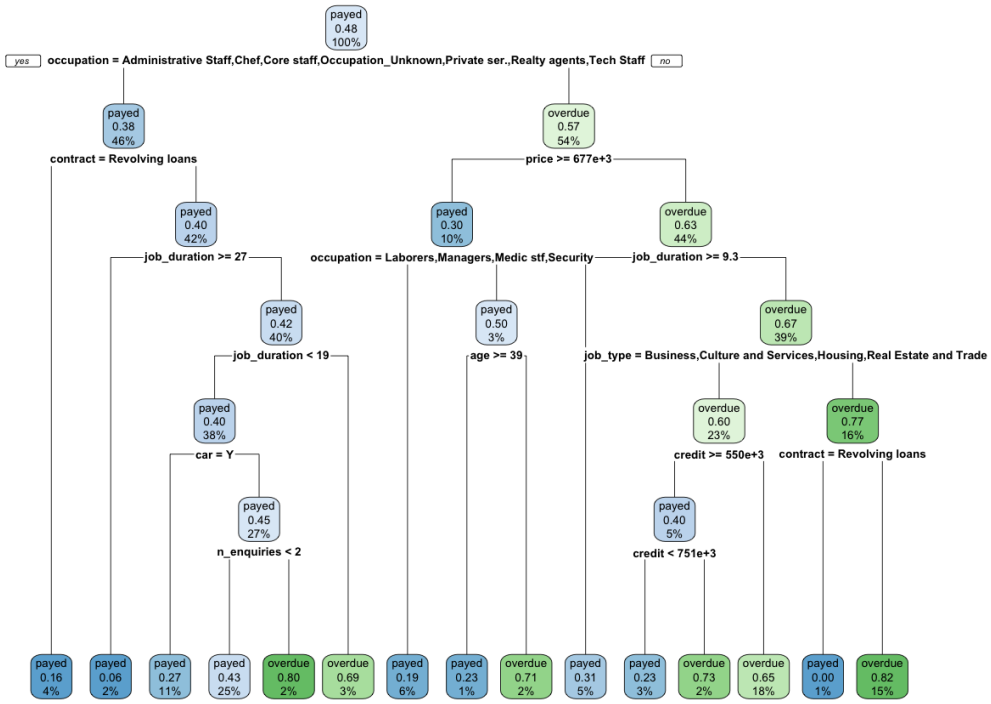


Figure 52: Path 3

- With the most specific tree, we have found several paths that are 100% accurate.
  - Young, rich business individuals applying for a loan that is less than 25% of their savings always pay the loan on time.
  - Service employees with low time on their job applying for a revolving loan lower than 677K are predicted to not pay the loan on time.
- Decision trees may not be the best model to predict a general case but
- - Have explicability
- - Can predict with high confidence rare / not very frequent groups of individuals.

In the realm of loans, certain countries require business owners to provide reasons for approving or denying a loan. Utilizing decision tree predictions would thus serve as an effective method to anticipate customer outcomes and understand the specific rules influencing the final decision.

Linear Discriminant analysis

In the below section, we have focused on the linear discriminant analysis among our data in order to validate if it's a suitable classification method. LDA is a classification model, where predicted output is the categorical target (payed or overdue) and predictors are the numerical variables in the dataframe.
As we've introduced previously, the dataset is unbalanced in terms of target. Overdue observation weights less than 10% among the total observation (Table 53), so a pre-treatment to balance them is needed.

Overdue	Payed
427	4572

Table 53. Target data distribution. Evidence that shows an unbalanced data set

The approach which we applied for balancing the dataset was **undersampling**, which reduces the size of payed observation to the same size as overdue.

After training the model, we have created a new feature (LD1) which stores the predictions. And using the barplot, we can observe that they follow a gaussian distribution (Figure 54), but the interclasses means are very close to each other and the intra-classes variance are high. These two characteristics show that this method has a bad performance among our dataset.

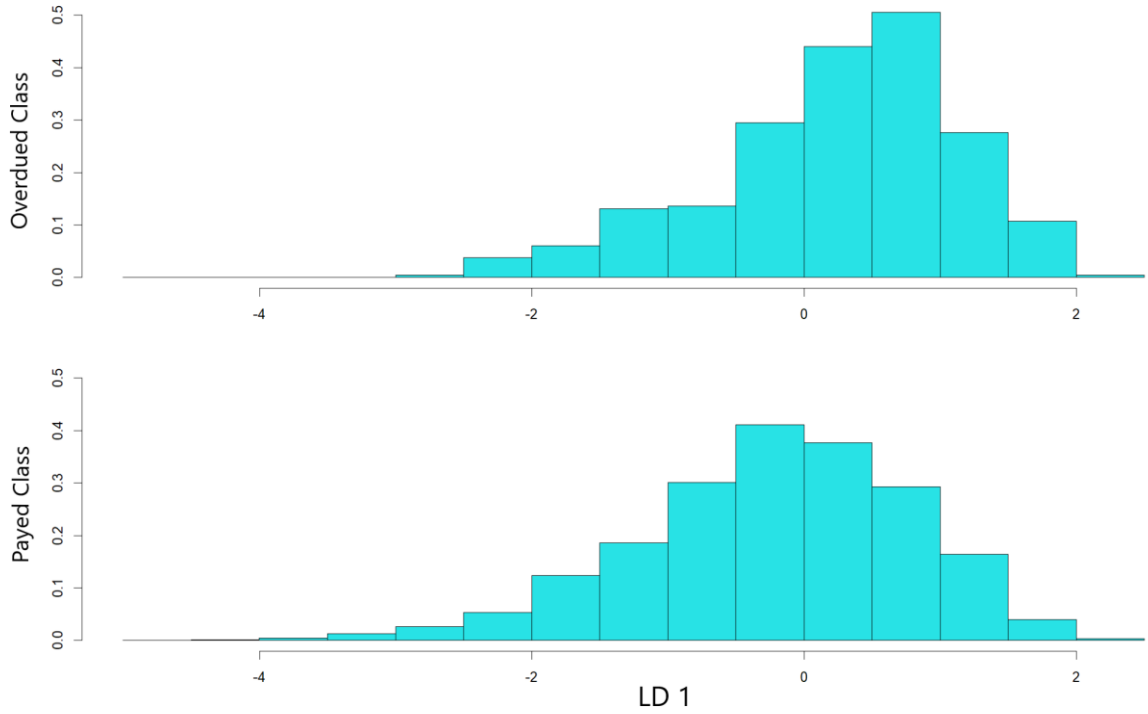


Figure 54. Normal distribution of LD1 compared between classified classes, with their mean and variance.

The predicted result (Figure 55) is expressed in the below confusion table, where we can observe an accuracy of 44.54%.

	Predicted Class		
Actual class	Class	Payed	Overdue
	Overdue	140	287
	Payed	2485	2087

Table 55. Confusion matrix for undersampling dataset

With this level of accuracy, we can conclude that the LDA model is not suitable for our dataset.

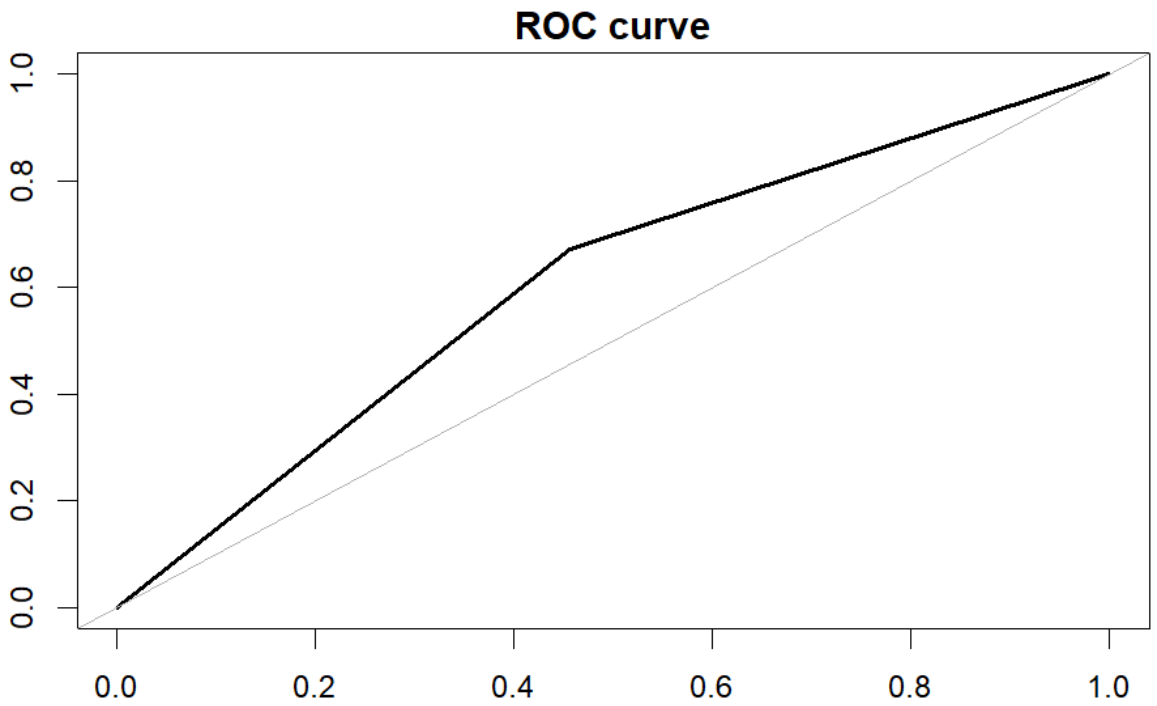


Figure 56. ROC curve for composed-method treated database

To remove possible inference due to inappropriate balancing techniques, we repeated the same process but combined both **undersampling method** and **oversampling method**.
In this case, using oversampling techniques, we can create an artificial dataset for overdue observation, and at the same time, reduce the payed observation with undersampling. We have considered a good approach of 1000 observations for each class.

	Predicted Class		
Actual class	Class	Payed	Overdue
	Overdue	161	266

	Payed	2686	1886

Table 57. Confusion matrix for composed-method treated database

After training the model, we repeated the same validation techniques, but the result remains the same, without any improvement.

Discussion and conclusions

The different models we have used in this project have raised different conclusions that we summarize in the following paragraphs.

PCA unveiled wealth's impact on loans, highlighted sociological factors in payment behavior, and distinguished timely from delayed payers, informing targeted strategies for loan management and risk assessment.

On the other hand, MCA showed that late loan payments are associated with a preference for cash loans over revolving loans, possibly driven by urgent financial needs, higher interest rates, and the financial instability of specific occupations like waiters and cleaners.

In association rules, we were not able to find any good rule, with enough lift, to predict when someone will pay the loan. Moreover, with a confidence of 0.01, we were not able to find any rule that predicts when someone will not pay the loan.

In the Hierarchical clustering section, we decided to use cluster size equal to 4 to split the data set into four different groups. This partitioning gives the best balance between sample size and variable explainability inter-clusters.

After that, we analyzed each variable one by one, and picked those that give more explainability among each cluster. The most significant variables that we considered are income, age, educational level and job status, which makes intuitive sense. With these variables, we were able to describe the characteristics of each four groups (Cluster 1 women with relatively low income, Cluster 2 men with low education level, Cluster 3 individuals with high income and education level, and Cluster 4 old people).

About decision trees, we have found several paths that are 100% accurate. Young, rich business individuals applying for a loan that is less than 25% of their savings always pay the loan on time. And another one: Service employees with low time on their job applying for a revolving loan lower than 677K are predicted to not pay the loan on time. Decision trees may not be the best model to predict a general case but: have explicability and can predict with high confidence rare / not very frequent groups of individuals.

In the Linear discriminant analysis section, we explore its application on an imbalance data set. Different data sampling methods were performed, but the LDA model shows poor performance with an accuracy of 44,54% in the best scenario. Hence we conclude that this model is not suitable for our data set.

Planned Gantt diagram and task distribution

Activity	Description	Starting date	Deadline	Adria C.	Alicia C.	Ji	Victor G.	Victor Geo	Status
Delivery D1		Wk-36 (11/09/2023)	**Wk-38 (18/09/2023)**						**Completed**
Activity 1	**Group creation**	Wk-36	Wk-37	**X**	**X**	**X**	**X**	**X**	**Completed**
Activity 2	**Data exploration**	wk-37	Wk-38	**X**	**X**	**X**	**X**	**X**	**Completed**
Activity 3	**Getting the Data**	wk-37	Wk-38	**X**	**X**	**X**	**X**	**X**	**Completed**
Activity 4	**Data description**	wk-37	Wk-38	**X**	**X**	**X**	**X**	**X**	**Completed**
Activity 5	**Data Structure overview**	wk-37	Wk-38	**X**	**X**	**X**	**X**	**X**	**Completed**
Delivery D2		Wk-39 (25/09/2023)	**Wk-40 (05/10/2023)**						**Completed**
Activity 6	**Group Consolidation**	Wk-39	Wk-40	**X**					**Completed**
Activity 7	**Design of the Grantt diagram & contingency plan**	Wk-39	Wk-40			**X**	**X**		**Completed**
Activity 8	**Creation of metadata file**	Wk-39	Wk-40	**X**	**X**			**X**	**Completed**
Delivery D3		Wk-40 (02/10/2023)	**Wk-44 (31/10/2023)**						

Activity 9	**Project introduction**	Wk-40	Wk-40	**X**	**X**	**X**	**X**	**X**	**Completed**
Activity 10	**Index**	Wk-40	Wk-41			**X**		**X**	**Completed**
Activity 11	**Project Motivation**	Wk-40	Wk-41			**X**			**Completed**
Activity 12	**Data Source presentation**	Wk-40	Wk-41		**X**	**X**	**X**		**Completed**
Activity 13	**Description of Data structure & metadata**	Wk-40	Wk-41		**X**		**X**	**X**	**Completed**
Activity 14	**Preprocessing Justification**	Wk-40	Wk-41	**X**	**X**	**X**			**Completed**
Activity 15	**Data Overview with R**	Wk-40	Wk-42		**X**		**X**	**X**	**Completed**
Activity 16	**Conclusion of univariate& bivariate analysis**	Wk-41	Wk-42				**X**		**Completed**
Activity 17	**PCA**	Wk-41	Wk-42	**X**	**X**			**X**	**Completed**
Activity 18	**MCA**	Wk-41	Wk-43		**X**				**Completed**
Activity 19	**MFA**	Wk-41	Wk-43	**X**					**Completed**
Delivery D4			**Wk-03 (15/01/2024)**						
Activity 20	**Association Rule Mining**	Wk-45	Wk-46	**X**			**X**		**Completed**

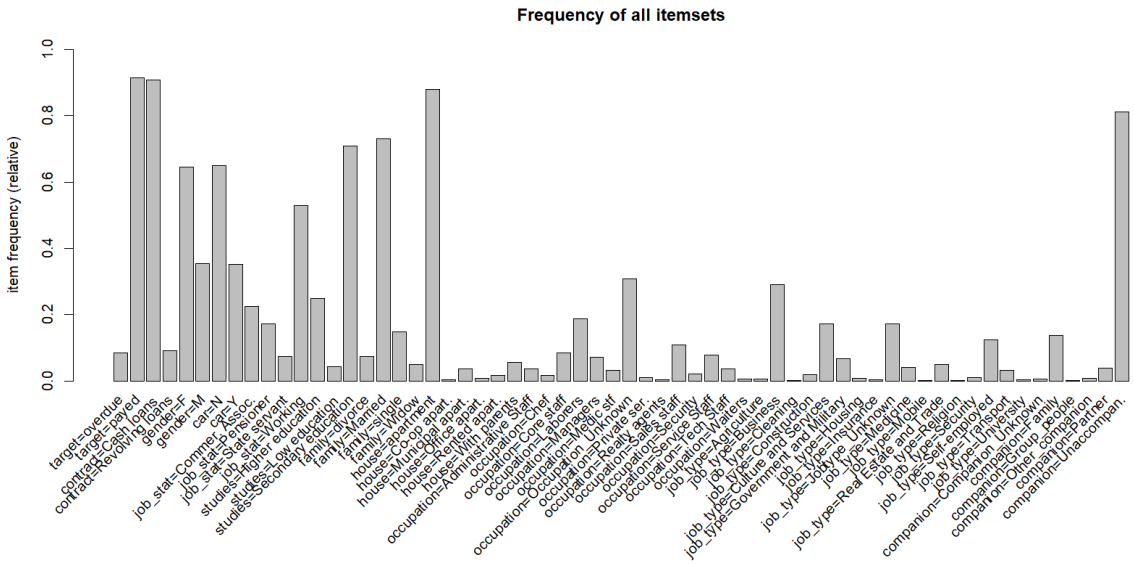
Activity 21	**Report writting**	Wk-45	Wk-46	**X**	**X**	**X**	**X**	**X**	**Completed**
Activity 22	**Hierarchical clustering**	Wk-45	Wk-46	X	**X**			**X**	**Completed**
Activity 22B	**Model Based Clustering**	Wk-45	Wk-46		**X**			**X**	**Completed**
Activity 23	**Profiling of Clusters**	Wk-45	Wk-46			**X**		**X**	**Completed**
Activity 24	**Decisions tree**	Wk-45	Wk-46		**X**			**X**	**Completed**
Activity 25	**Discriminant Analysis (LDA)**	Wk-45	Wk-46			**X**			**Completed**
Activity 26	**Discussion and Conclusion**	Wk-45	Wk-46	**X**	**X**	**X**	**X**	**X**	**Completed**
Activity 27	**Gantt Table**	Wk-47	Wk-50	**X**		**X**			**Completed**
Presentation			**Wk-03 (16/01/2024)**						
Activity 1	**Submit the Documentation**	Wk-02	Wk-02	**X**	**X**	**X**	**X**	**X**	**Completed**
Activity 2	**Making project Presentation**	Wk-51	Wk-01	**X**	**X**	**X**	**X**	**X**	**Completed**
Activity 3	**Practice oral presentation**	Wk-01	Wk-02	**X**	**X**	**X**	**X**	**X**	**Completed**

Planned 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 advanced 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

Annex



Frequency of all itemsets (66) in the transactional database.

EDA Reports (after and before analysis)

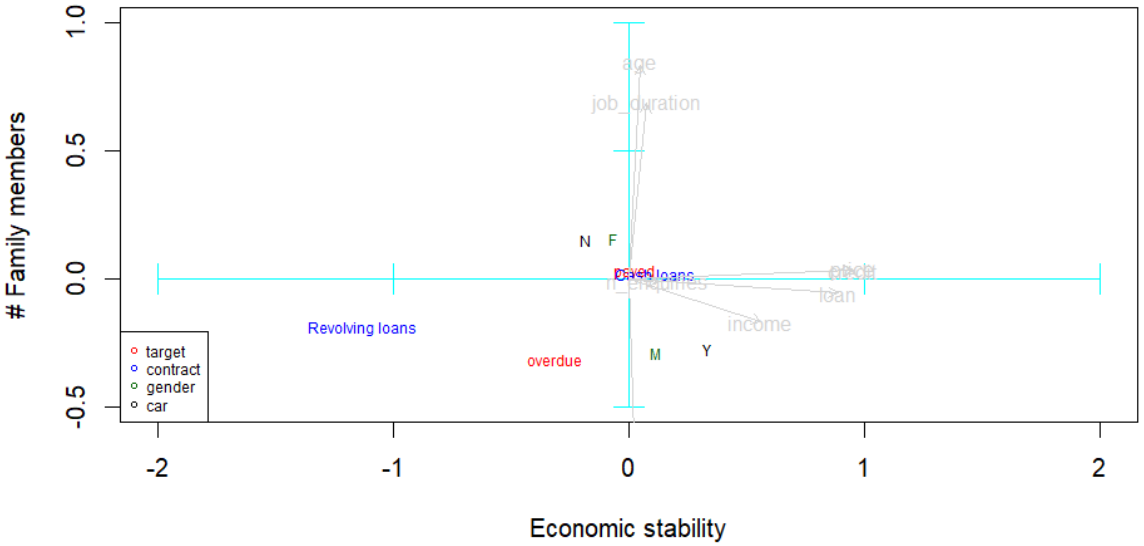
The basic statistics of numerical data is shown in the following table:

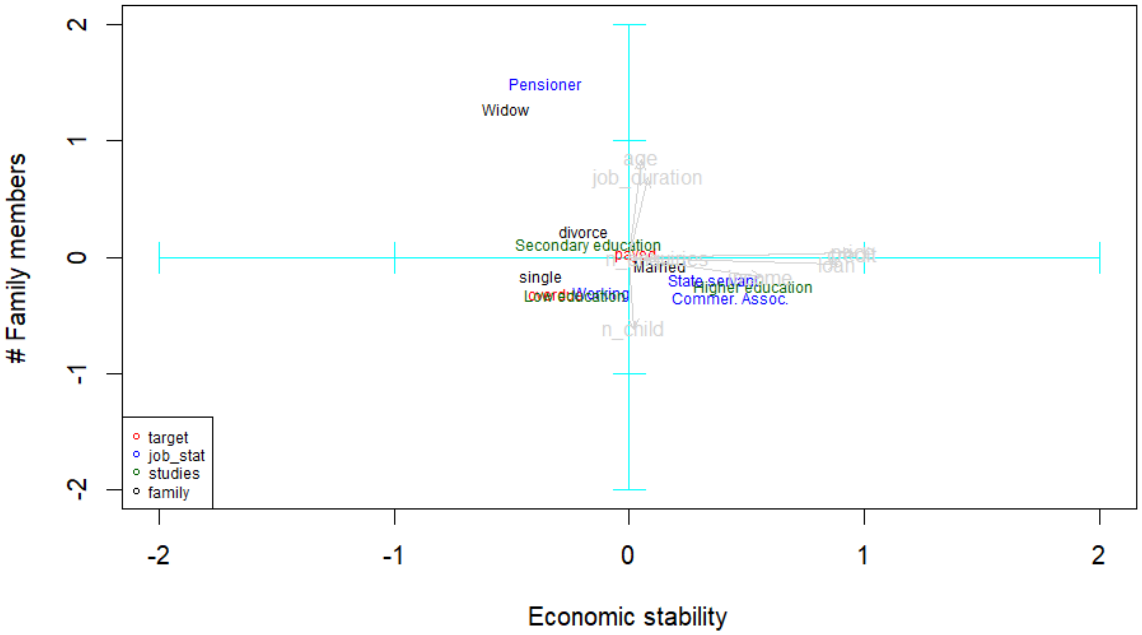
```
       id           target            contract            gender
 Min.   :   1   Length:5000        Length:5000        Length:5000
 1st Qu.:1251   Class :character   Class :character   Class :character
 Median :2500   Mode  :character   Mode  :character   Mode  :character
 Mean   :2500
 3rd Qu.:3750
 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
```

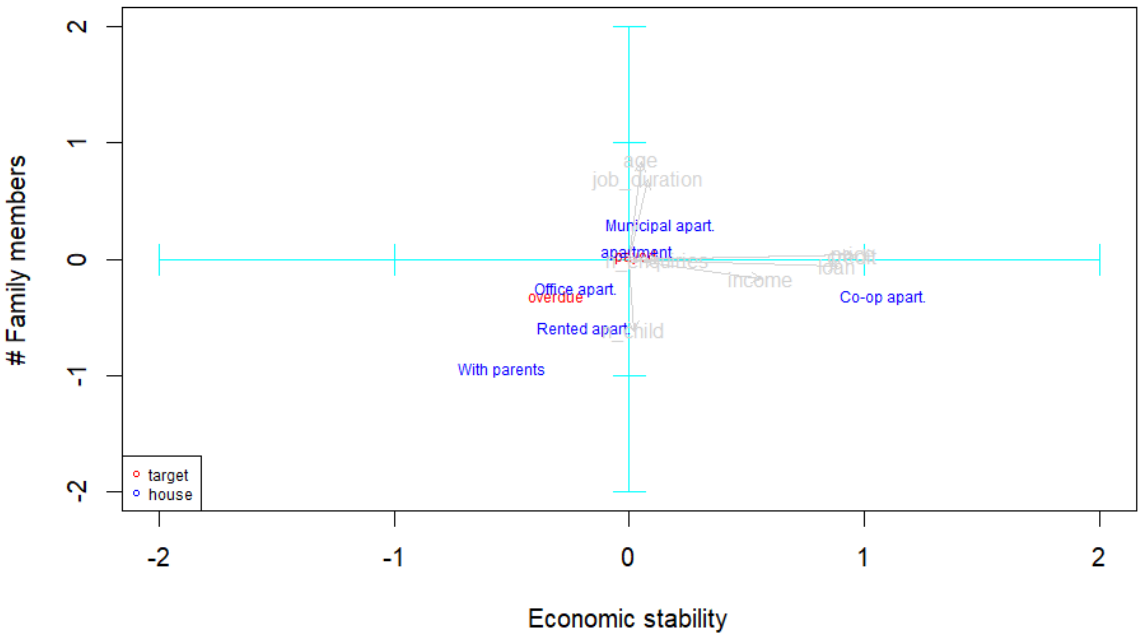
```
                                    Mean   :43.18   Mean   : 7.3363
                                    3rd Qu.:53.00   3rd Qu.: 9.5452
                                    Max.   :68.00   Max.   :41.8904
 occupation          job_type        n_enquiries       companion
Length:5000        Length:5000      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
```

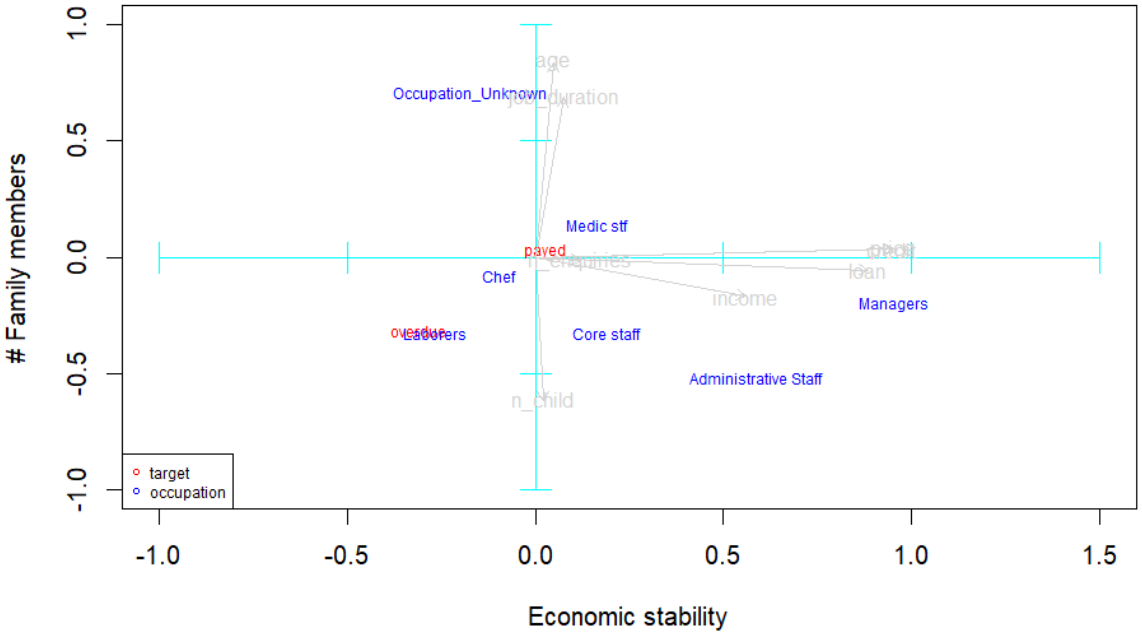
For a more detailed EDA analysis see the reports generated by “Smart EDA” package before the imputation, after imputation, and only for payers and debtors. To see the report it is recommended that you download them from the following link: *https://drive.google.com/drive/folders/1g62S6VQS3HHU6jYw_gR6PM7LAXpt6n3N?usp=drive_link*

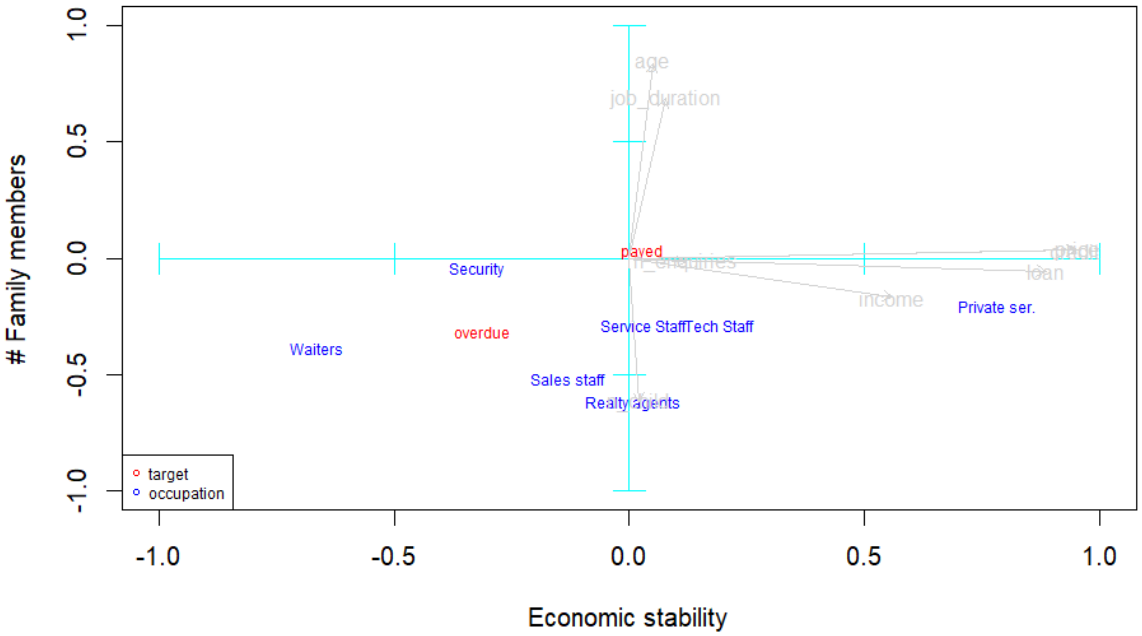
PCA Modalities on the plane generated by PC1 and PC2

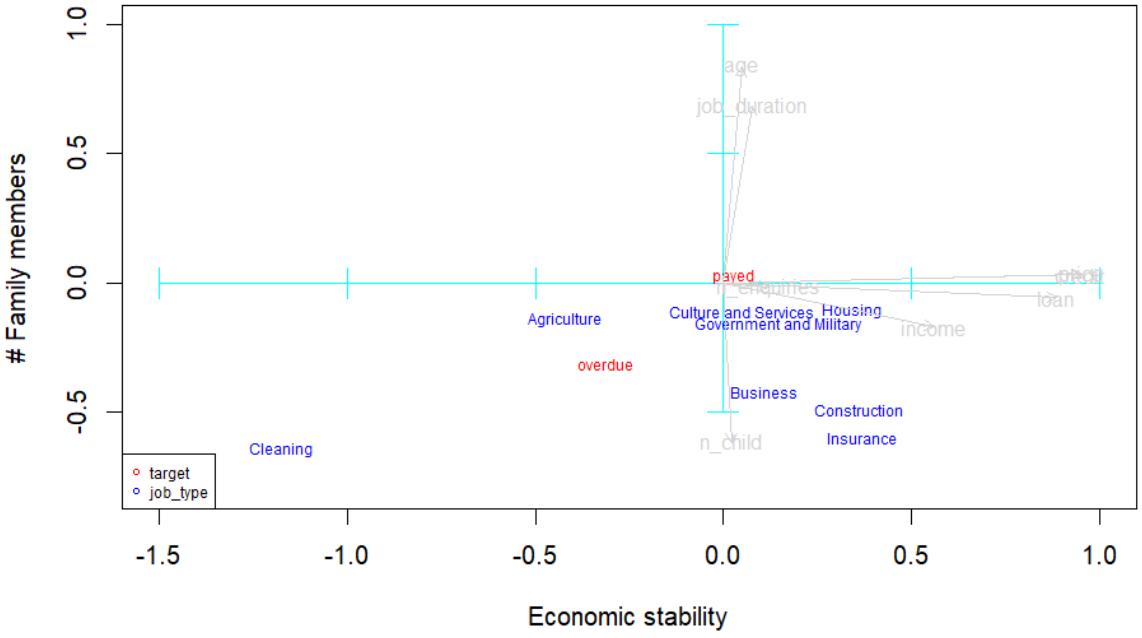


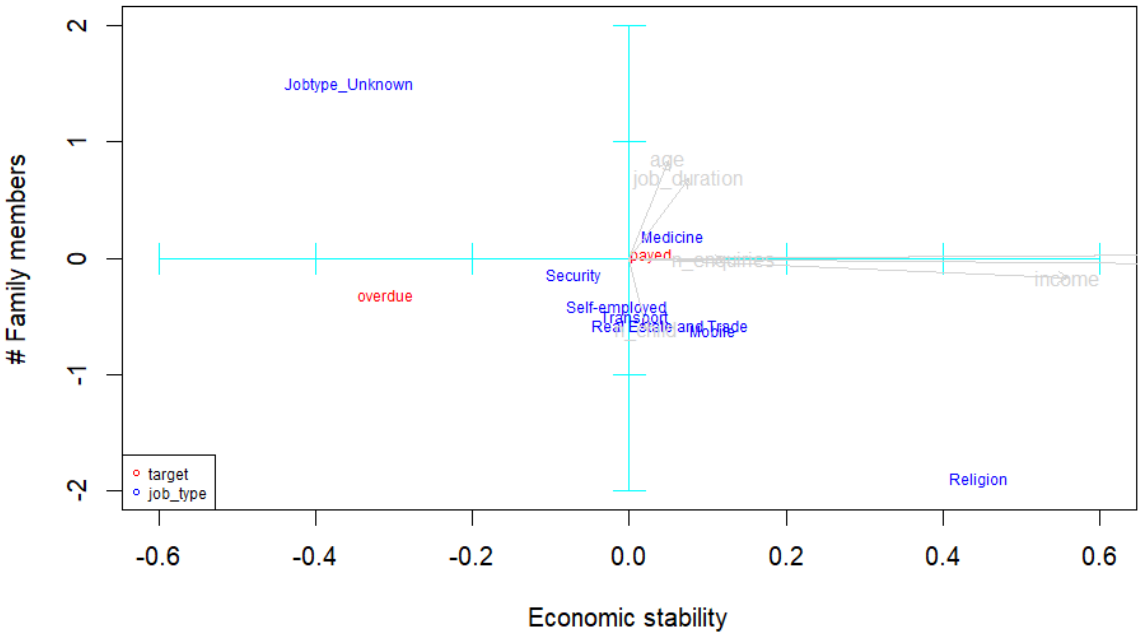

Pensioner
Widow
age
job_duration
divorce
Secondary education
Married
single
State servant
Higher education
Commer. Assoc.
n_child
target
job_stat
studies
family
Family members
Economic stability
-2
-1
0
1
2

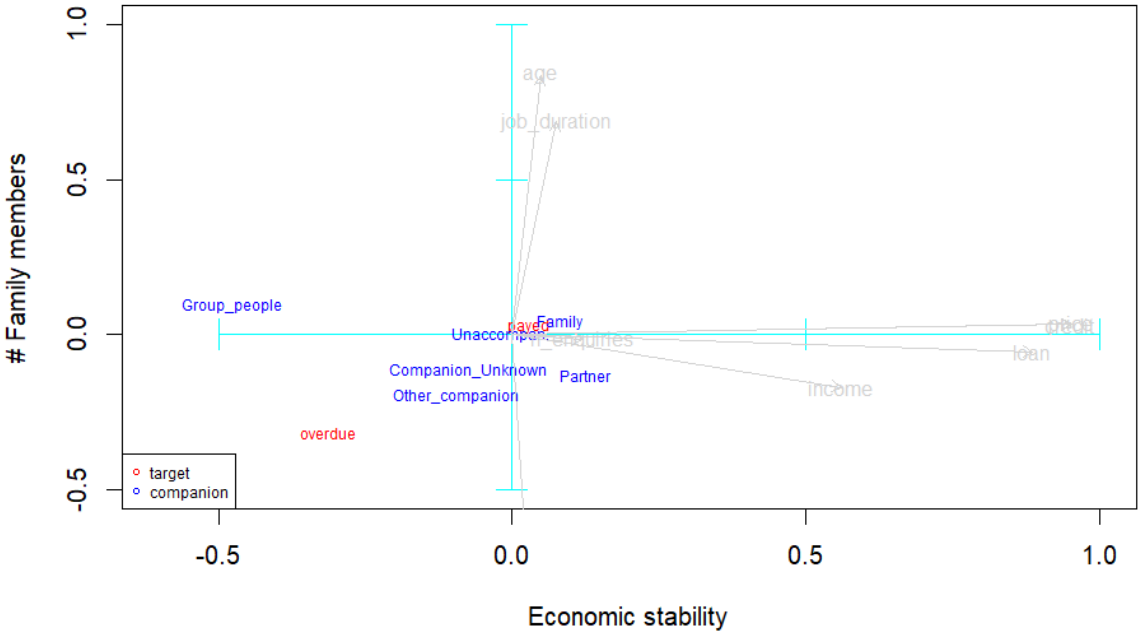

age
job_duration
Municipal apart.
apartment
Office apart.
overdue
Co-op apart.
Rented apart.
With parents
n_child
target
house
Family members
Economic stability
-2
-1
0
1
2


age
Occupation_Unknown
job_duration
Medic stf
payed
Chef
loan
income
Managers
overdue
Laborers
Core staff
Administrative Staff
n_child
target
occupation
Family members
Economic stability
1.0
0.5
0.0
-0.5
-1.0
1.5


age
job_duration
payed
Security
loan
income
Private ser.
Service StaffTech Staff
overdue
Waiters
Sales staff
Realty agents
target
occupation
Family members
Economic stability
1.0
0.5
0.0
-0.5
-1.0


Family members
Economic stability
age
job_duration
paved
n_enquiries
loan
Agriculture
Culture and Services
Housing
Government and Military
income
overdue
Business
Construction
Insurance
n_child
Cleaning
target
job_type


Family members
Economic stability
Jobtype_Unknown
age
job_duration
Medicine
paved
n_enquiries
income
Security
overdue
Self-employed
Transport
Real Estate and Trade
Mobile
Religion
target
job_type


1.0
0.5
0.0
-0.5
Family members
age
job_duration
Group_people
Family
Unaccompanied
Companion_Unknown
Partner
Other_companion
income
loan
overdue
target
companion
-0.5
0.0
0.5
1.0
Economic stability

Profiling with k = 3

Variable	Type	Test	p-value
target	qualitative	Chi Square	0
contract	qualitative	Chi Square	<2,2E-16
gender	qualitative	Chi Square	<2,2E-16
car	qualitative	Chi Square	<2,2E-16
n_child	numerical	Kruskal-Wallis	5,34E-84
income	numerical	Kruskal-Wallis	6,54E-46
credit	numerical	Kruskal-Wallis	4,52E-05
loan	numerical	Kruskal-Wallis	3,37E-17
price	numerical	Kruskal-Wallis	7,96E-05
job_stat	qualitative	Chi Square	<2,2E-16
studies	qualitative	Chi Square	<2,2E-16
family	qualitative	Chi Square	<2,2E-16
house	qualitative	Chi Square	1,45E-08
age	numerical	Kruskal-Wallis	0,00E+00
job_duration	numerical	Kruskal-Wallis	1,49E-47
occupation	qualitative	Chi Square	<2,2E-16
job_type	qualitative	Chi Square	<2,2E-16
n_enquiries	numerical	Kruskal-Wallis	3,50E-02
companion	qualitative	Chi Square	1,70E-03

Variable type, test used and resulting p-value for all variables with k=3.

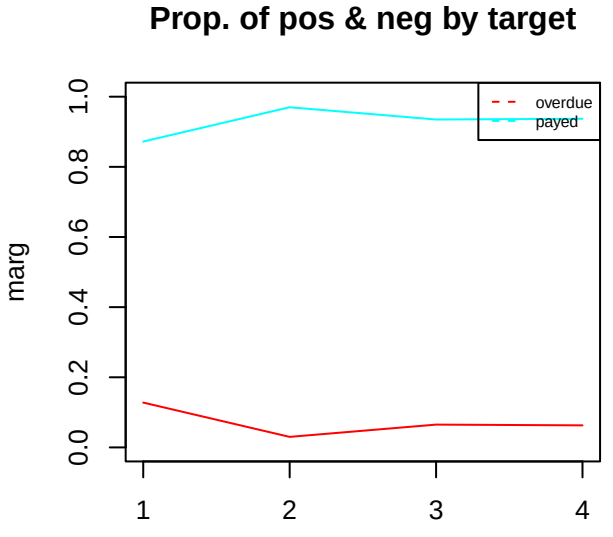
				P -values		
Variable	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 1**	**Cluster 2**	**Cluster 3**
cluster size	3144	997	858	-	-	-
Target	Payed: 89,09% Overdue: 10,91%	Payed: 96,99% Overdue: 3,01%	Payed: 93,70% Overdue: 6,30%	0	0	0
Gender	Mostly female (79,86%)	Mostly male (97.69%)	Mostly female (80.77%)	0	0	0
Car	No (68,95%)	Yes (61.28%)	No (80.54%)	0	0	0
Child (mean)	0,5175	0,4473	0,0300	5,57E-35	9,10E-02	0
Income (mean)	169536	185769	133146	1,31E-03	9,04E-14	0

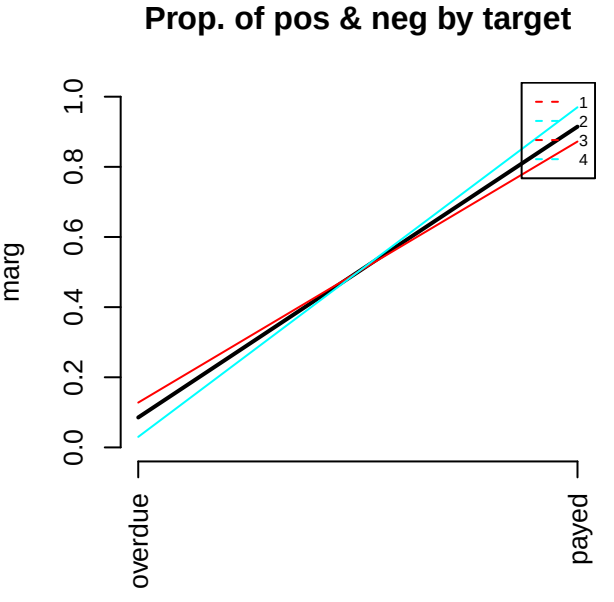
Job status (commercial, pensioneer, state servant, Working)	881	245	0	0	0	0
	1	0	858			
	339	28	0			
	1923	724	0			
Study	Sec. edu (63,10%)	Sec. edu (86,46%)	Sec. edu (81,23%)	0	0	0
Family stats (single, widow)	Single (12%) Widow (3%)	Single (25%) Widow (1%)	Single (10%) Widow (16%)	0	0	0
Age (year, mean)	40,17	38,81	59,25	0	0	0
Job duration (year, mean)	6,8	5,5	10,89	7,81E-09	0	1,74E-06
Job occupation	Unknown (17,36%), Laborers (16,85%), Sales staff (16,22%)	Laborers (41,02%)	Unknow (100%) (pensioner)	0	0	0

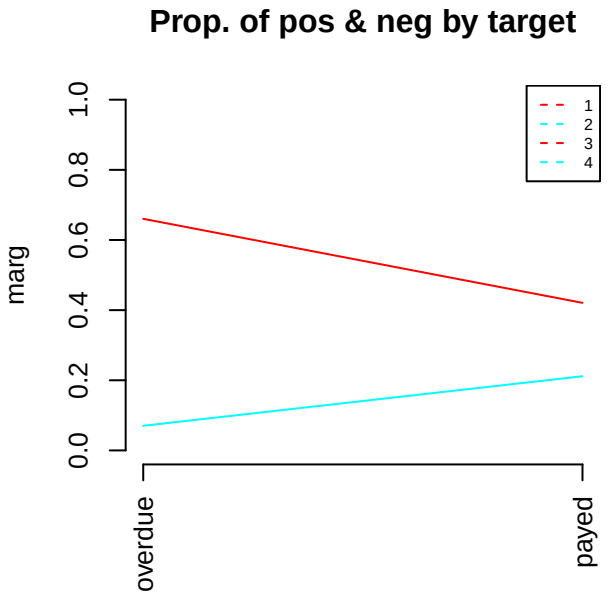
Statistics and p-values for every cluster for significant variables for k=3.

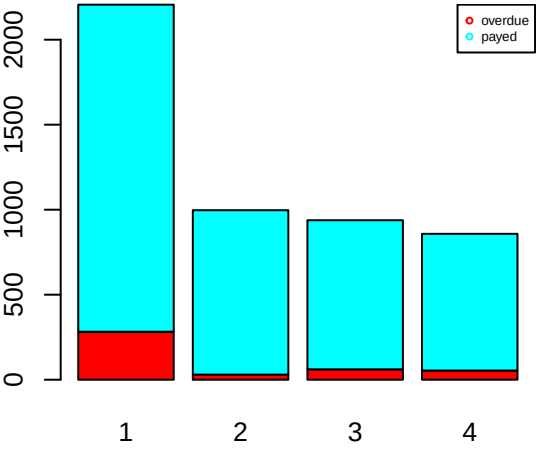
Class	**Description of the Class**	**Subjective Business Suggestion**
C 1	Biggest cluster size (3144 observation) Mostly women with relatively low income. Their occupation types are mostly related to the service field as Sales, Laborer etc. Overdue: 10,91%	Be careful with accepting the application.
C 2	Mostly men who own a car, with the second highest mean income and low education level. But they are reliable by returning the loan. Overdue: 3%	High-quality client.
C 3	Oldest people with lowest income, and are innestable. Mostly women composed of retired pensioners who do not own a car. Overdue: 6,3%	Low-value clients. DO NOT recommend a high value loan.

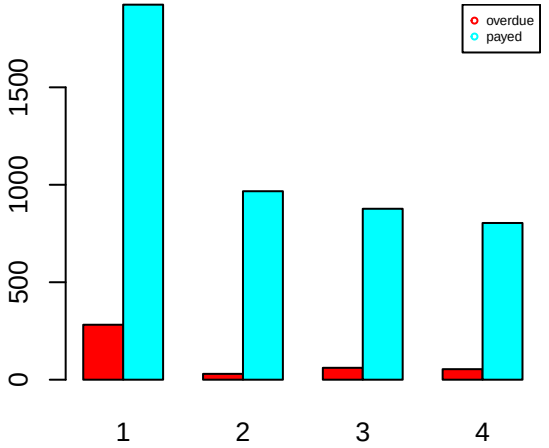
Summary description of each class and subjective business suggestion for k=3.

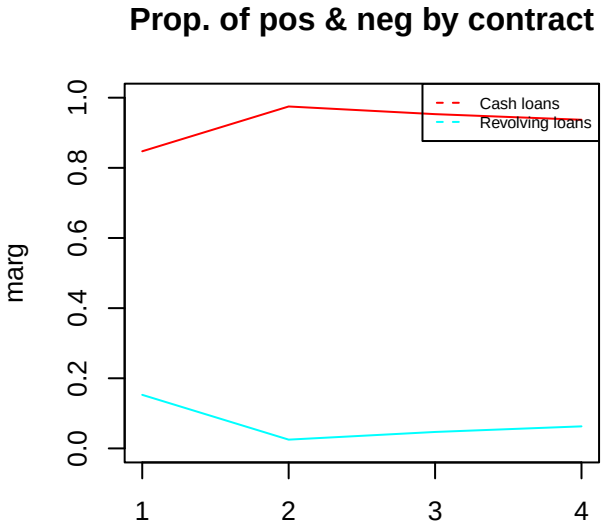

Prop. of pos & neg by target
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
overdue
payed

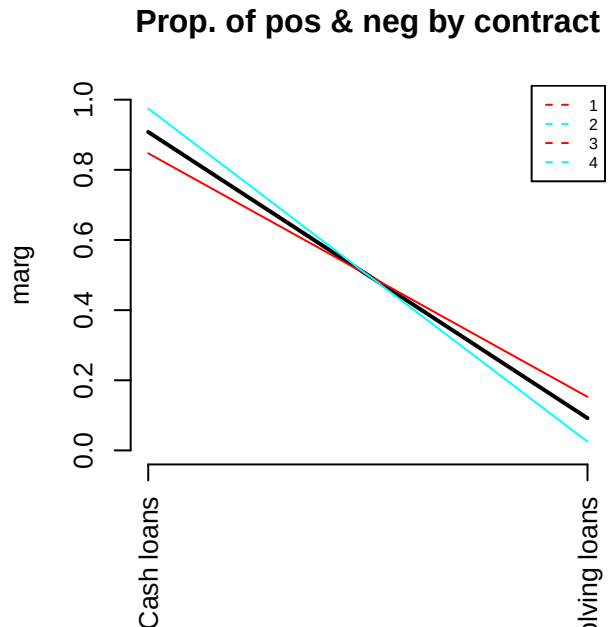

Prop. of pos & neg by target
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
overdue
payed

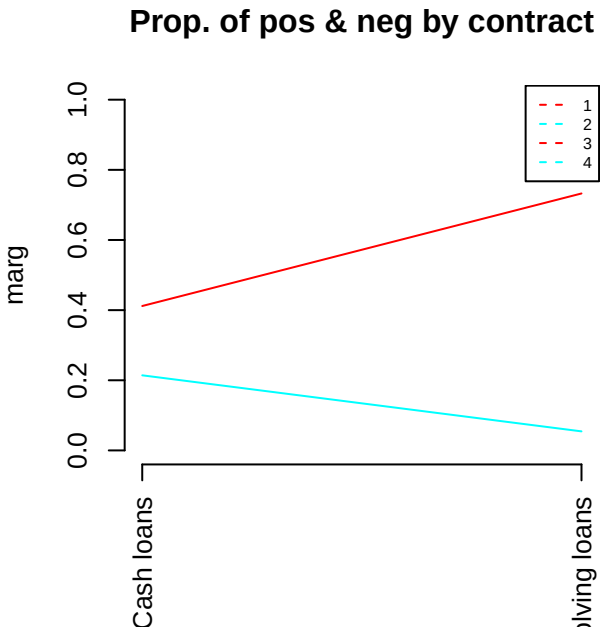

Prop. of pos & neg by target
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
overdue
payed

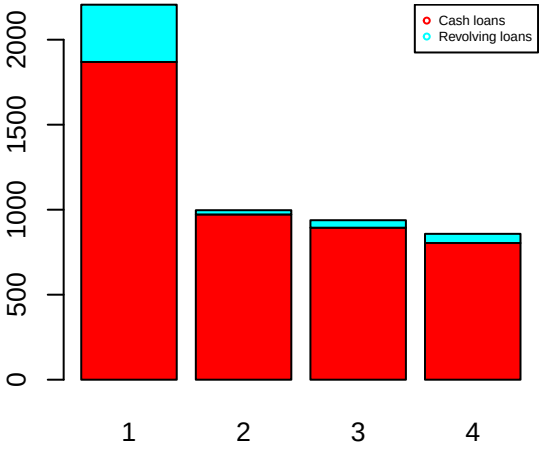

0
500
1000
1500
2000
overdue
payed
1
2
3
4

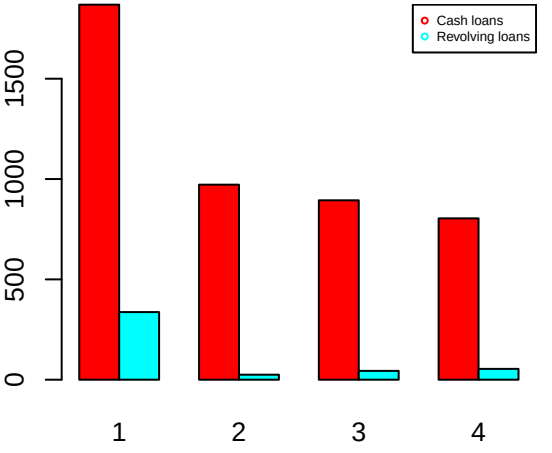

overdue
payed
0
500
1000
1500
1
2
3
4

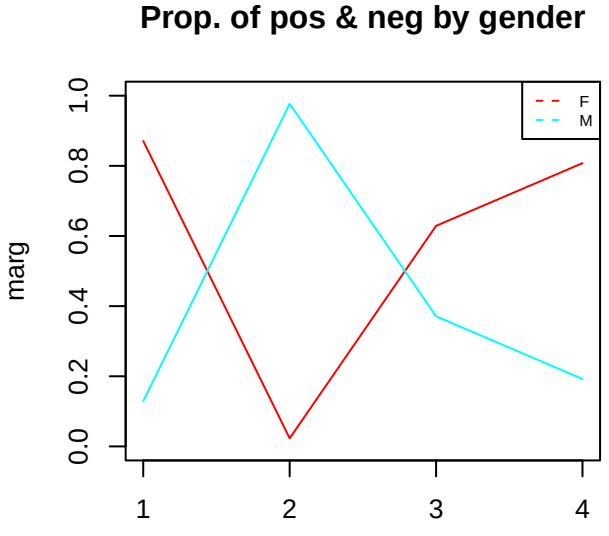

Prop. of pos & neg by contract
Cash loans
Revolving loans
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4

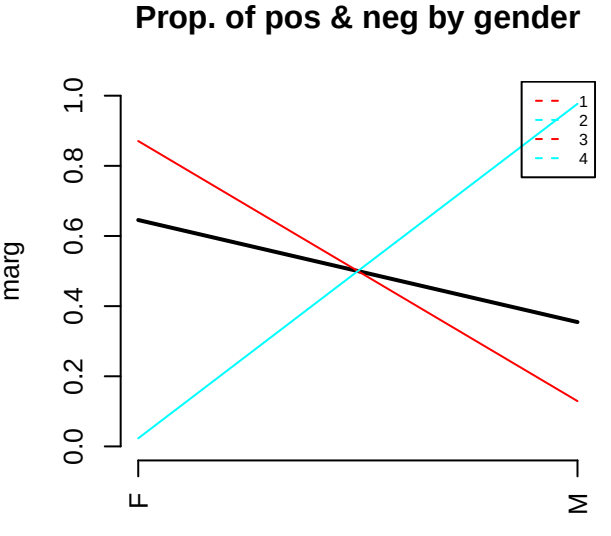

Prop. of pos & neg by contract
1
2
3
4
marg
0.0
0.2
0.4
0.6
0.8
1.0
Cash loans
olving loans

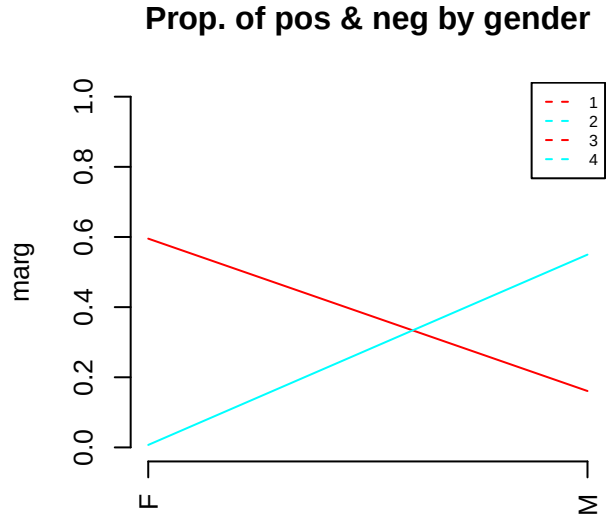

Prop. of pos & neg by contract
1
2
3
4
marg
0.0
0.2
0.4
0.6
0.8
1.0
Cash loans
olving loans

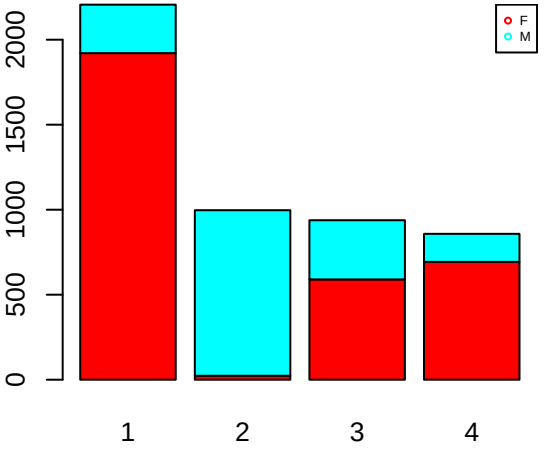

Cash loans
Revolving loans
0
500
1000
1500
2000
1
2
3
4

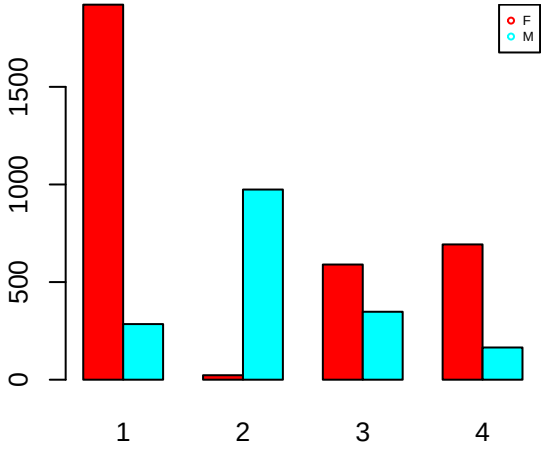

Cash loans
Revolving loans
0
500
1000
1500
1
2
3
4

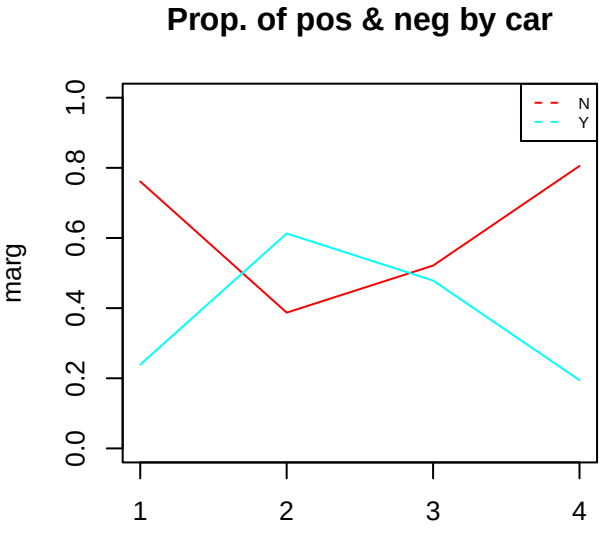

Prop. of pos & neg by gender
F
M
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4

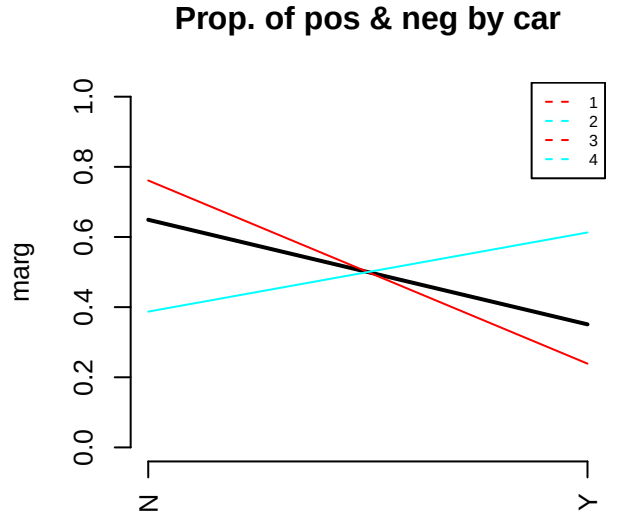

Prop. of pos & neg by gender
1
2
3
4
marg
0.0
0.2
0.4
0.6
0.8
1.0
F
M

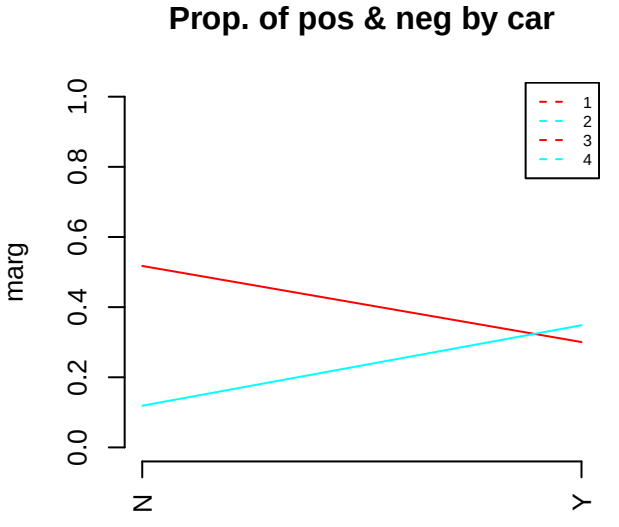

Prop. of pos & neg by gender
marg
0.0
0.2
0.4
0.6
0.8
1.0
F
M
1
2
3
4

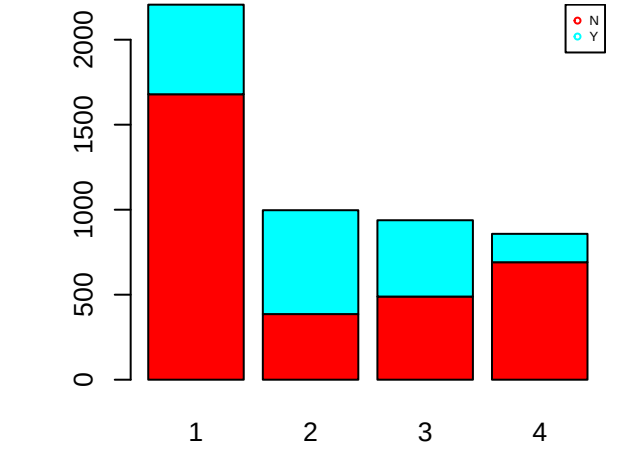

0
500
1000
1500
2000
1
2
3
4
F
M

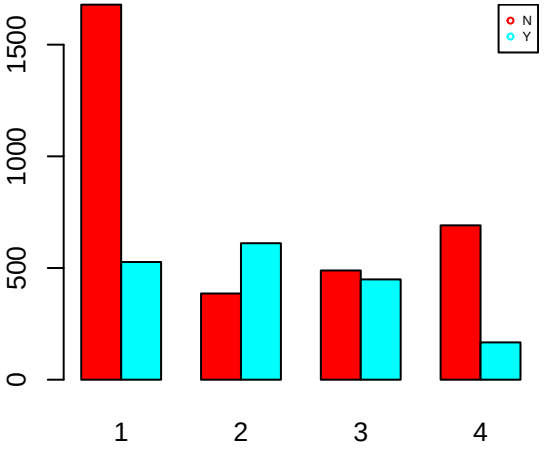

0
500
1000
1500
1
2
3
4
F
M

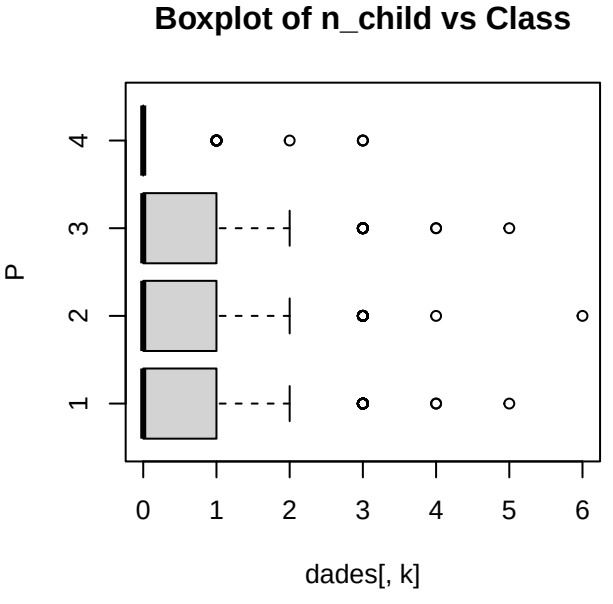

Prop. of pos & neg by car
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
N
Y

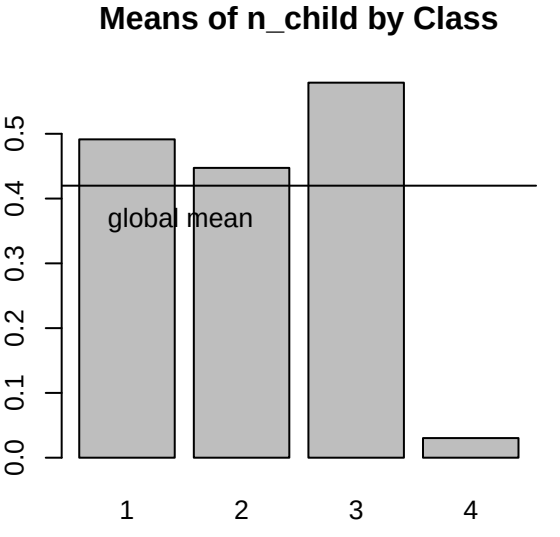

Prop. of pos & neg by car
marg
0.0
0.2
0.4
0.6
0.8
1.0
N
Y
1
2
3
4

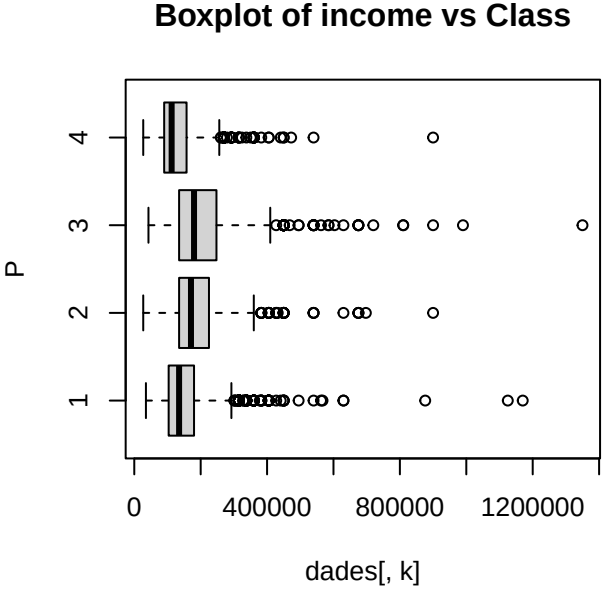

Prop. of pos & neg by car
marg
0.0
0.2
0.4
0.6
0.8
1.0
N
Y
1
2
3
4

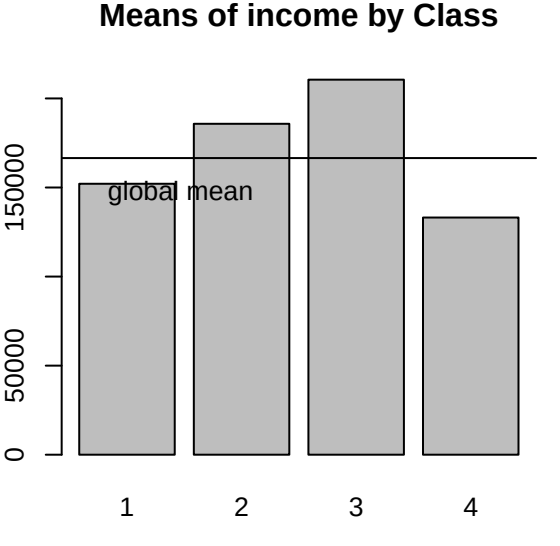

0
500
1000
1500
2000
N
Y
1
2
3
4

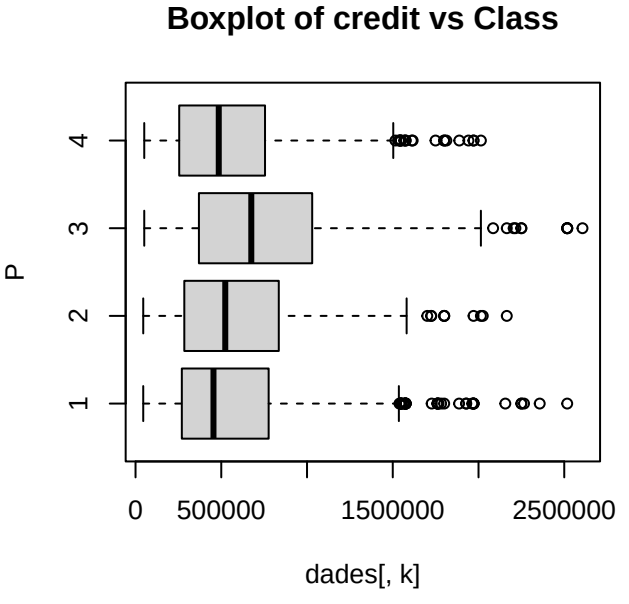

0
500
1000
1500
N
Y
1
2
3
4

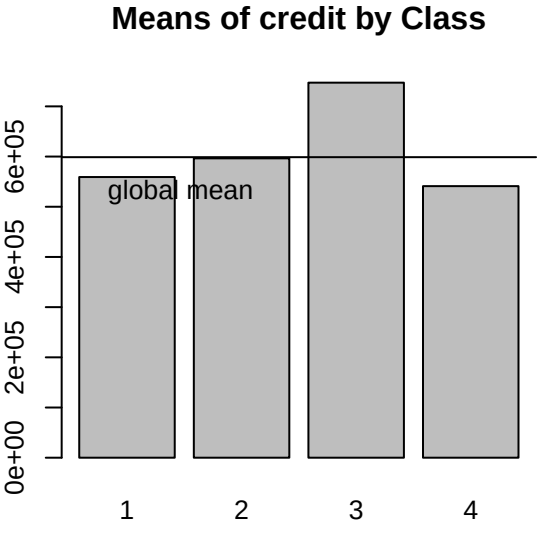

Boxplot of n_child vs Class
P
4
3
2
1
0
1
2
3
4
5
6
dades[, k]

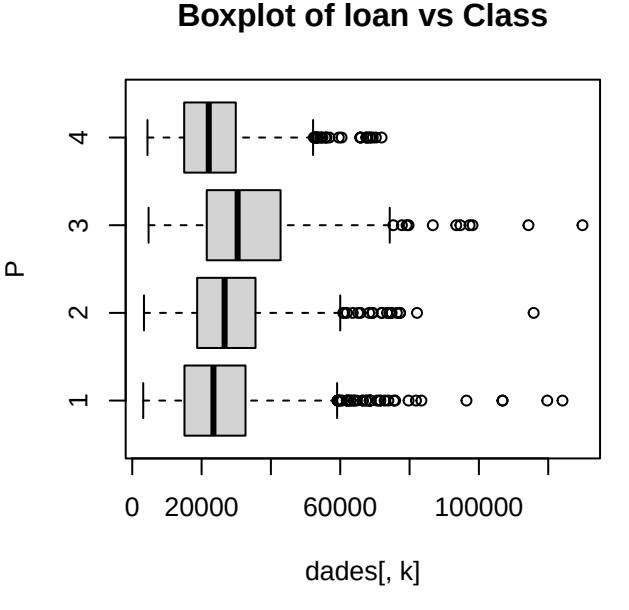

Means of n_child by Class
0.0
0.1
0.2
0.3
0.4
0.5
global mean
1
2
3
4

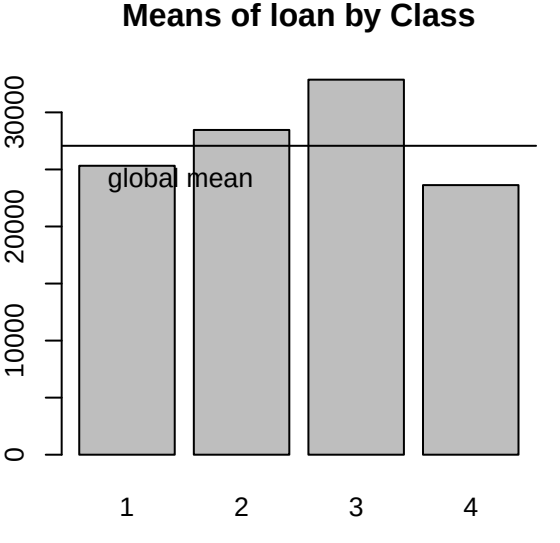

Boxplot of income vs Class
P
4
3
2
1
0
400000
800000
1200000
dades[, k]

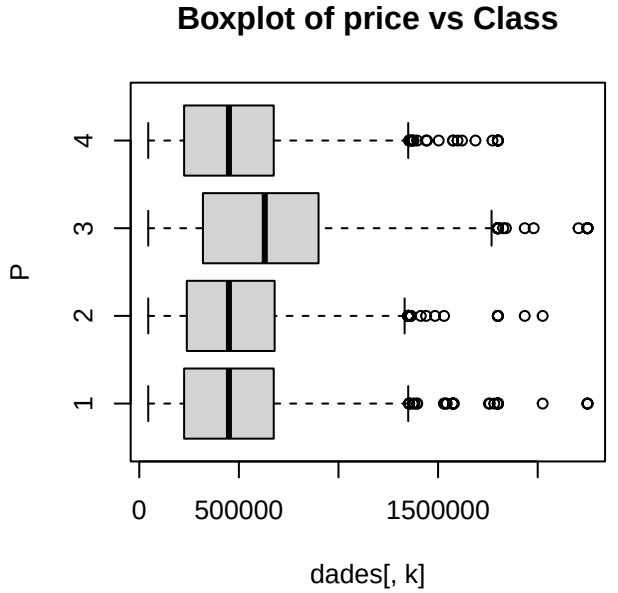

Means of income by Class
0
50000
150000
global mean
1
2
3
4

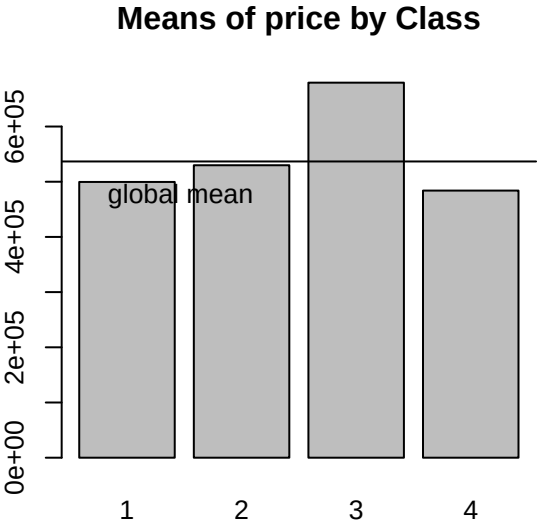

Boxplot of credit vs Class
P
4
3
2
1
0
500000
1500000
2500000
dades[, k]

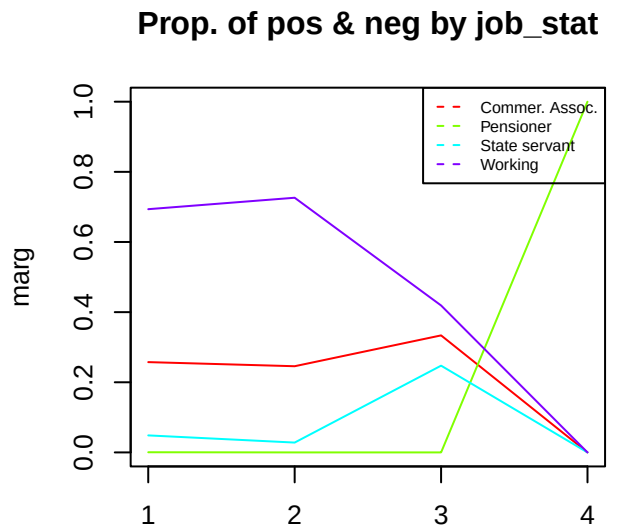

Means of credit by Class
0e+00
2e+05
4e+05
6e+05
global mean
1
2
3
4

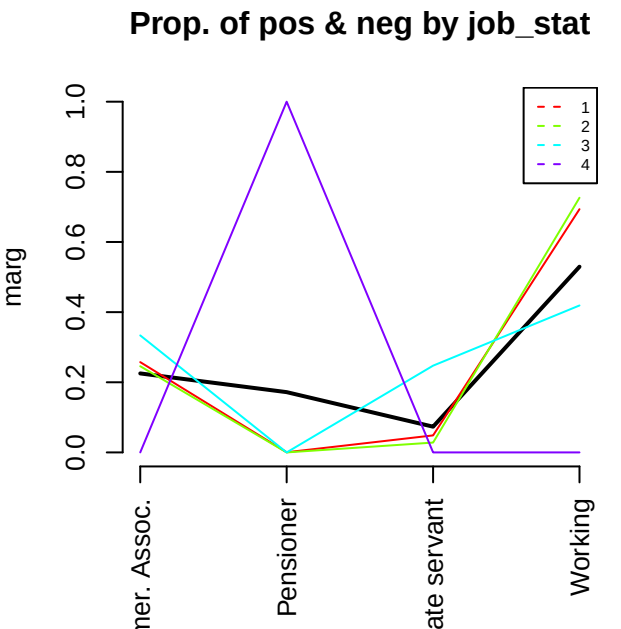

Boxplot of loan vs Class
P
4
3
2
1
0
20000
60000
100000
dades[, k]

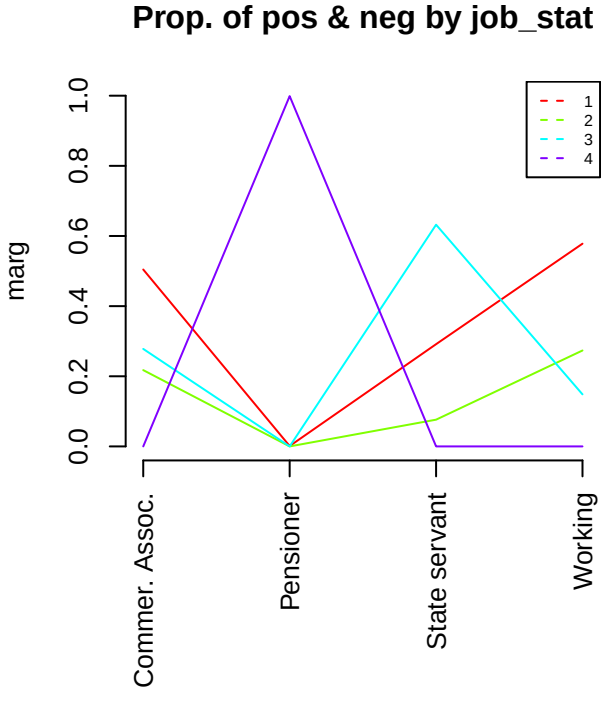

Means of loan by Class
0
10000
20000
30000
global mean
1
2
3
4

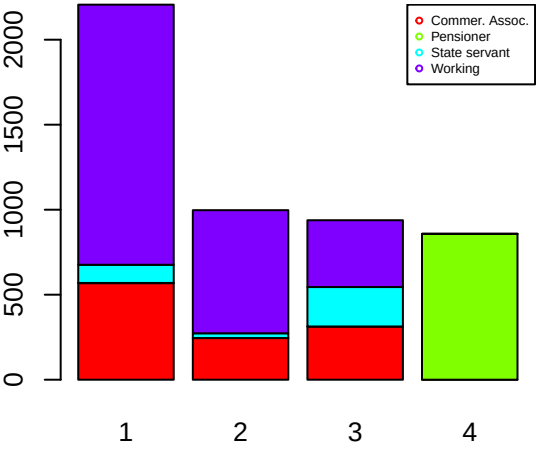

Boxplot of price vs Class
P
4
3
2
1
0
500000
1500000
dades[, k]

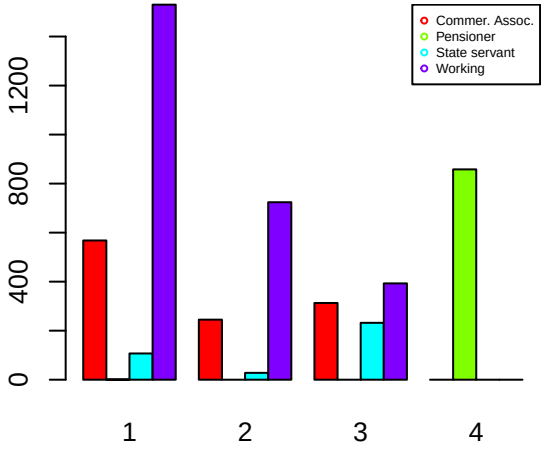

Means of price by Class
0e+00
2e+05
4e+05
6e+05
global mean
1
2
3
4

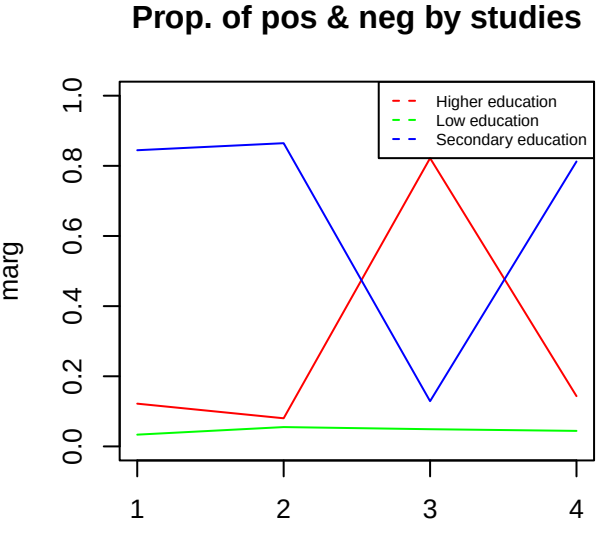

Prop. of pos & neg by job_stat
marg
0.0
0.2
0.4
0.6
0.8
1.0
Commer. Assoc.
Pensioner
State servant
Working
1
2
3
4

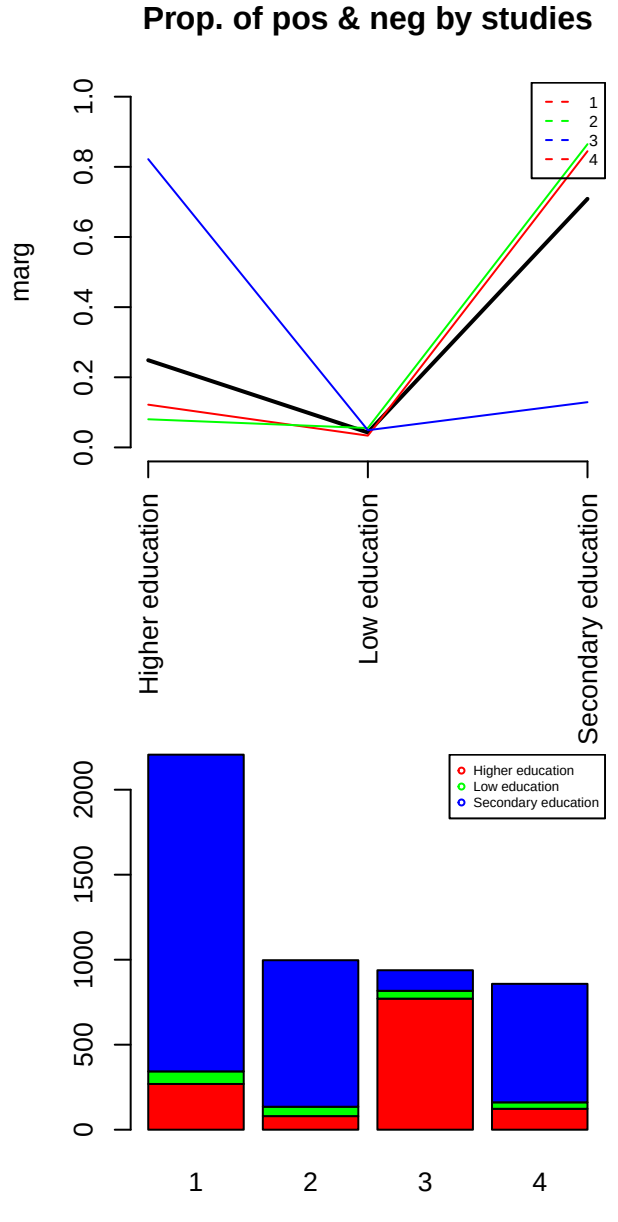

Prop. of pos & neg by job_stat
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
Pensioner
Working

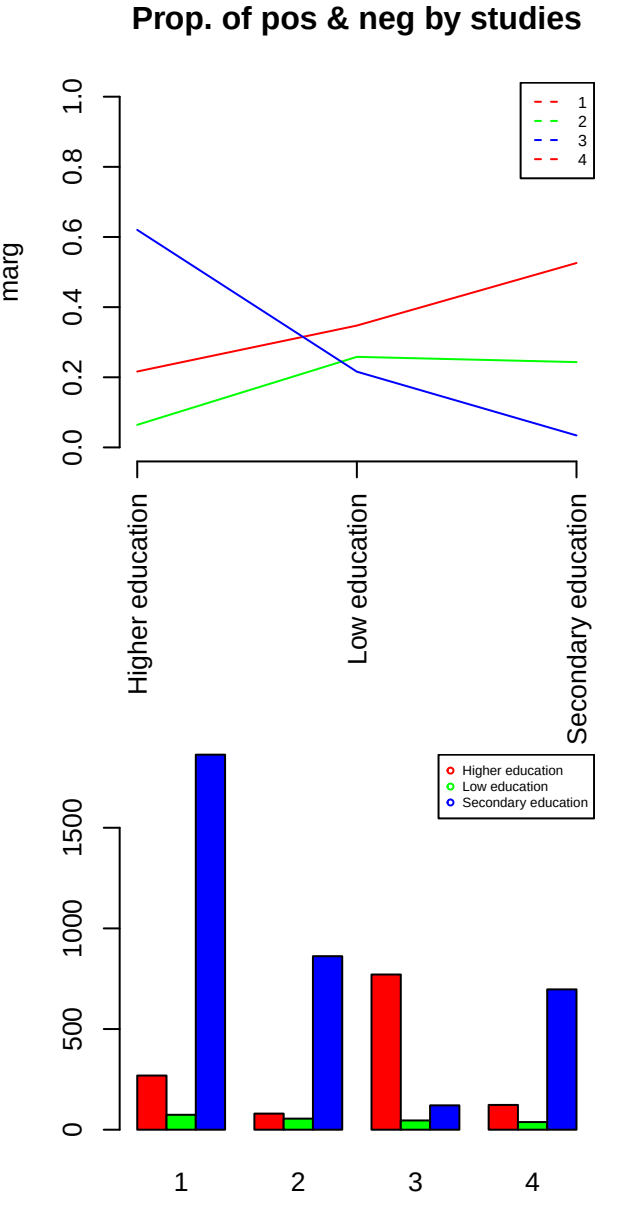

Prop. of pos & neg by job_stat
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
Commer. Assoc.
Pensioner
State servant
Working

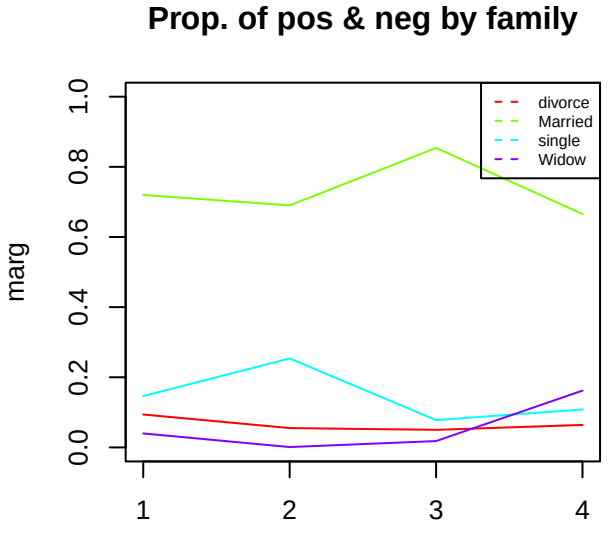

0
500
1000
1500
2000
Commer. Assoc.
Pensioner
State servant
Working
1
2
3
4

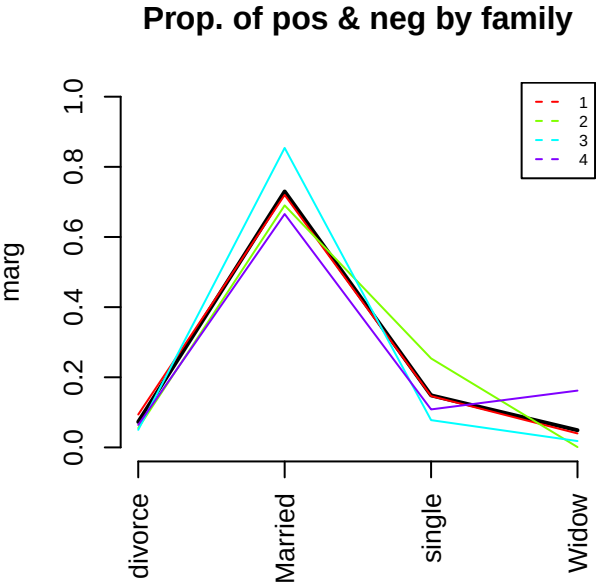

0
400
800
1200
Commer. Assoc.
Pensioner
State servant
Working
1
2
3
4

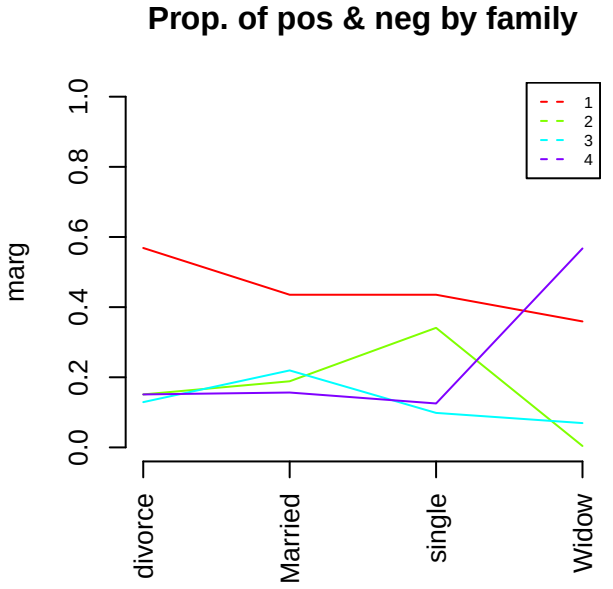

Prop. of pos & neg by studies
marg
0.0
0.2
0.4
0.6
0.8
1.0
Higher education
Low education
Secondary education
1
2
3
4

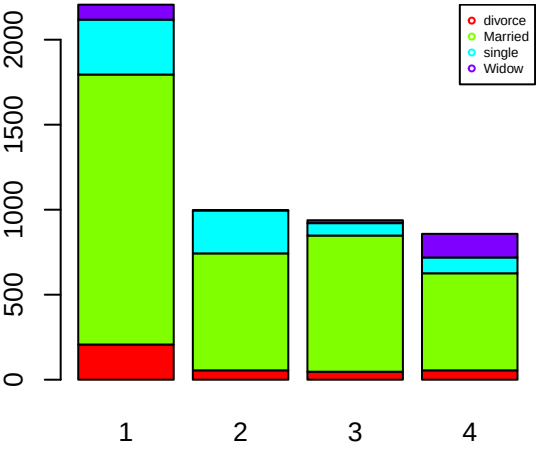

Prop. of pos & neg by studies
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
Higher education
Low education
Secondary education
0
500
1000
1500
2000
Higher education
Low education
Secondary education
1
2
3
4

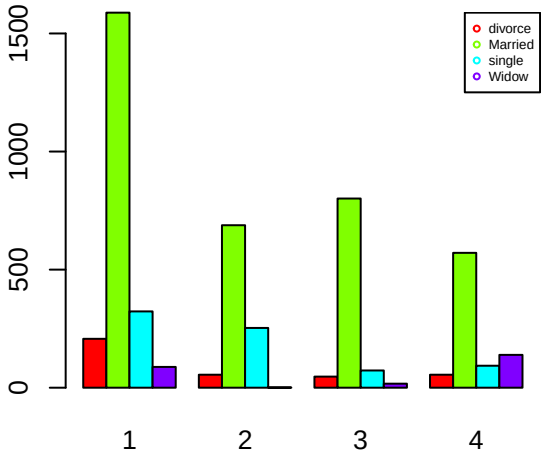

Prop. of pos & neg by studies
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
Higher education
Low education
Secondary education
0
500
1000
1500
Higher education
Low education
Secondary education
1
2
3
4

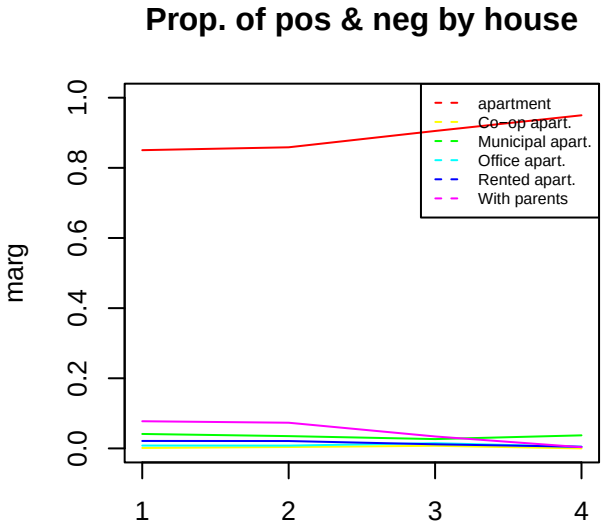

Prop. of pos & neg by family
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
divorce
Married
single
Widow

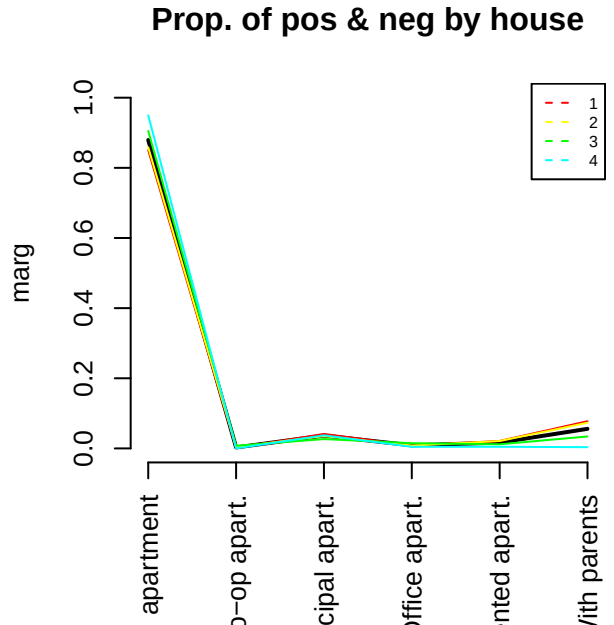

Prop. of pos & neg by family
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
divorce
Married
single
Widow

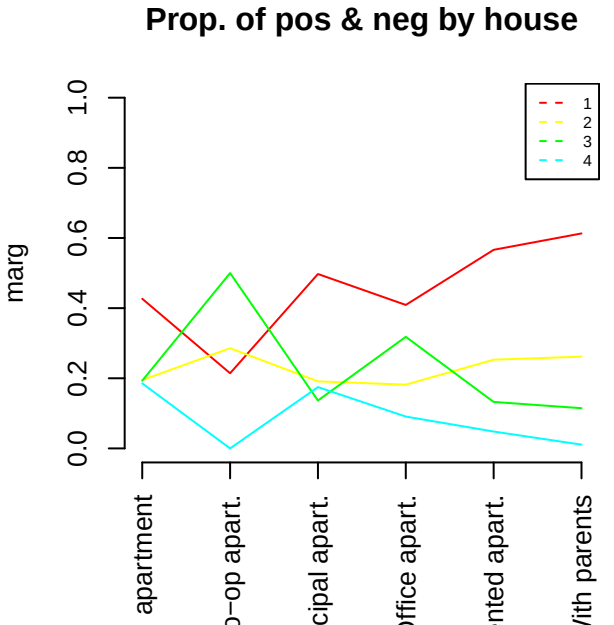

Prop. of pos & neg by family
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
divorce
Married
single
Widow

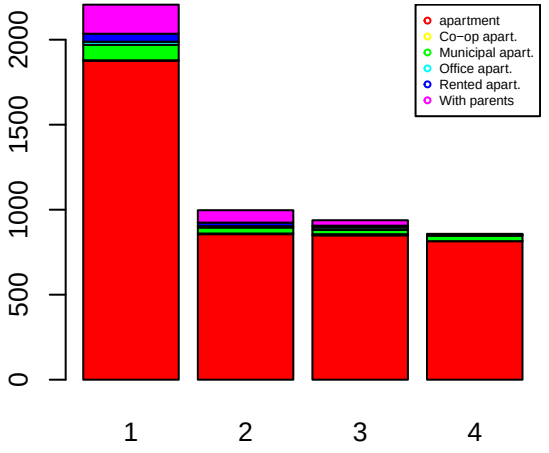

0
500
1000
1500
2000
divorce
Married
single
Widow
1
2
3
4

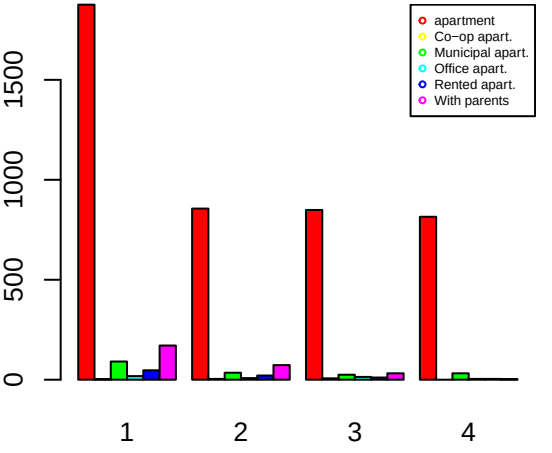

divorce
Married
single
Widow
0
500
1000
1500
1
2
3
4

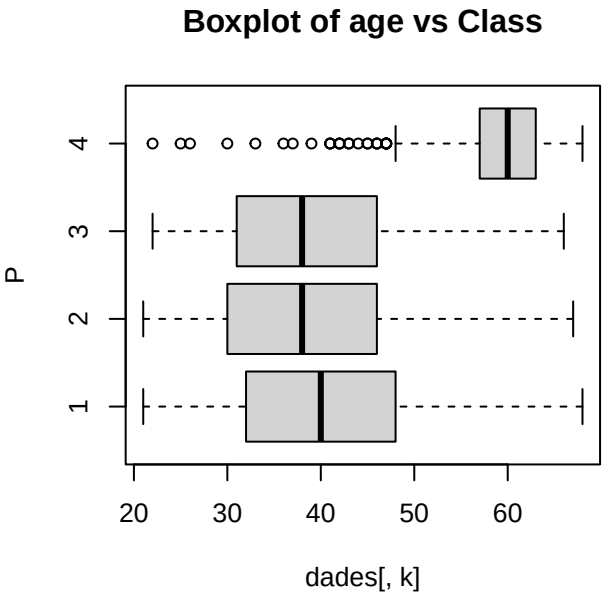

Prop. of pos & neg by house
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
apartment
Co−op apart.
Municipal apart.
Office apart.
Rented apart.
With parents


Prop. of pos & neg by house
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
apartment
o−op apart.
cipal apart.
ffice apart.
nted apart.
ith parents

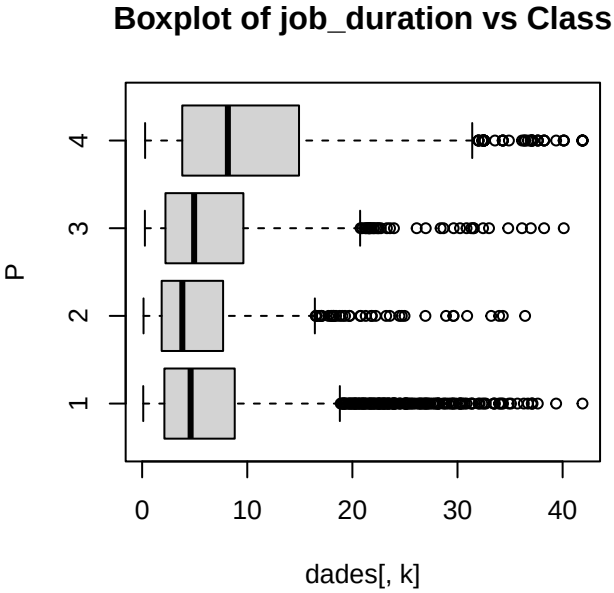

Prop. of pos & neg by house
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
apartment
o−op apart.
cipal apart.
ffice apart.
nted apart.
ith parents


apartment
Co−op apart.
Municipal apart.
Office apart.
Rented apart.
With parents
0
500
1000
1500
2000
1
2
3
4

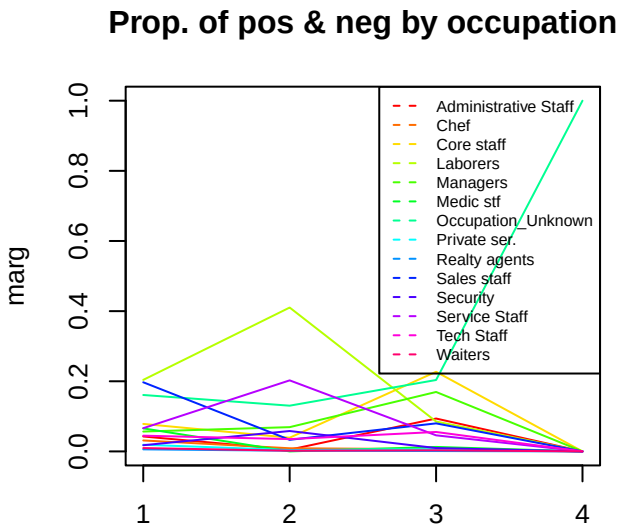

apartment
Co−op apart.
Municipal apart.
Office apart.
Rented apart.
With parents
0
500
1000
1500
1
2
3
4

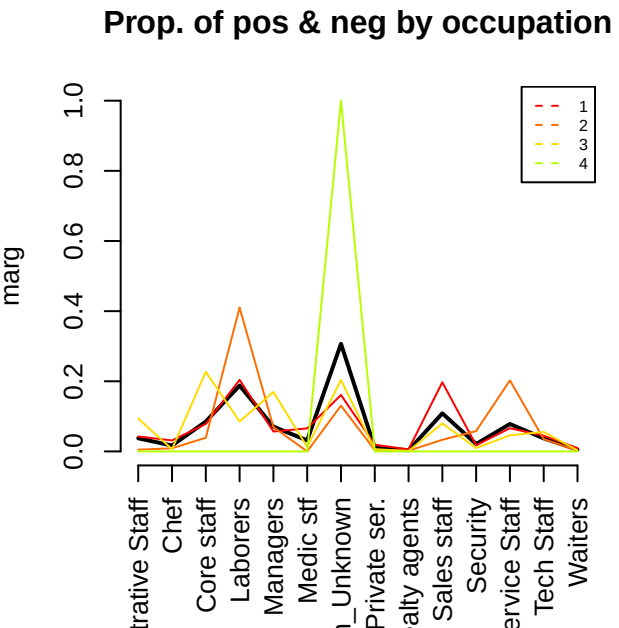

Boxplot of age vs Class
P
1
2
3
4
20
30
40
50
60
dades[, k]


Means of age by Class
global mean
0
10
20
30
40
50
1
2
3
4

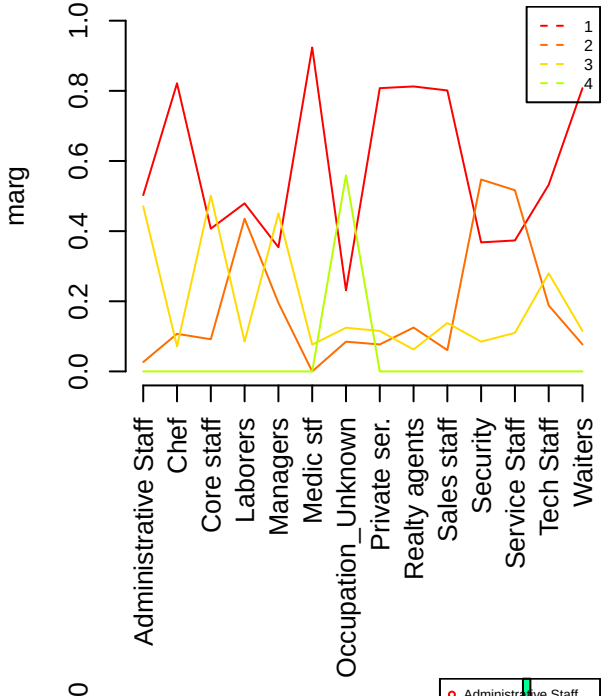

Boxplot of job_duration vs Class
P
4
3
2
1
0
10
20
30
40
dades[, k]

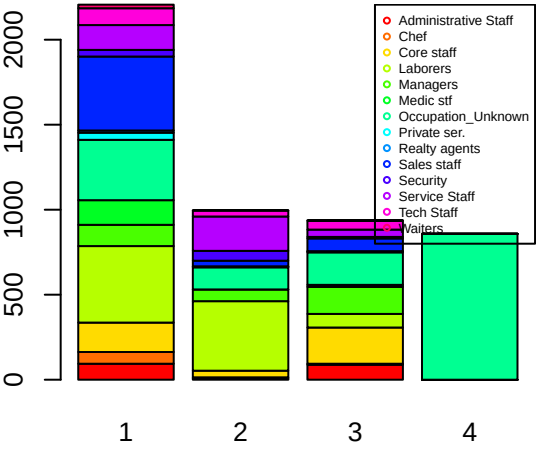

Means of job_duration by Class
0
2
4
6
8
10
global mean
1
2
3
4

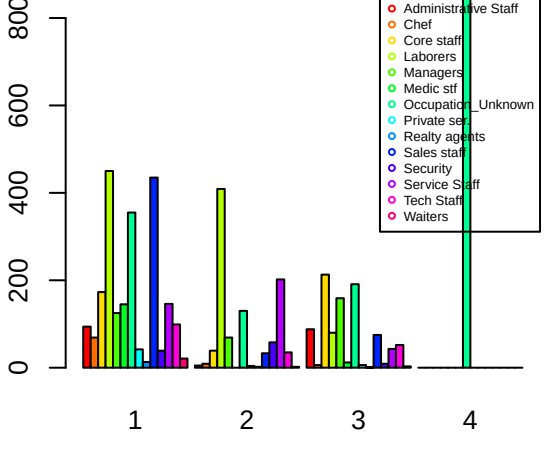

Prop. of pos & neg by occupation
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
Administrative Staff
Chef
Core staff
Laborers
Managers
Medic stf
Occupation_Unknown
Private ser.
Realty agents
Sales staff
Security
Service Staff
Tech Staff
Waiters


Prop. of pos & neg by occupation
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
Chef
Core staff
Laborers
Managers
Medic stf
Private ser.
Sales staff
Security
Tech Staff
Waiters

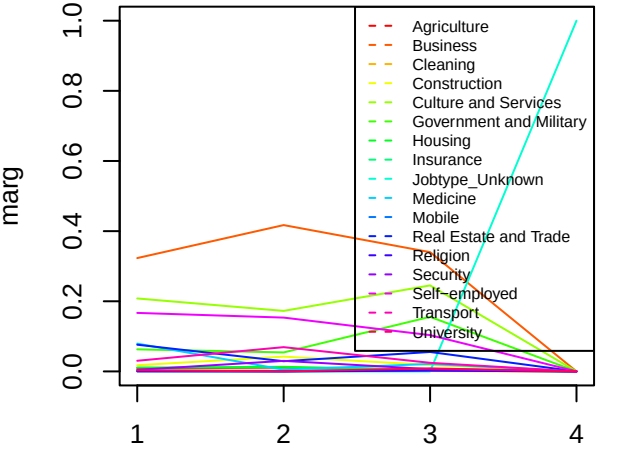
Prop. of pos & neg by occupation

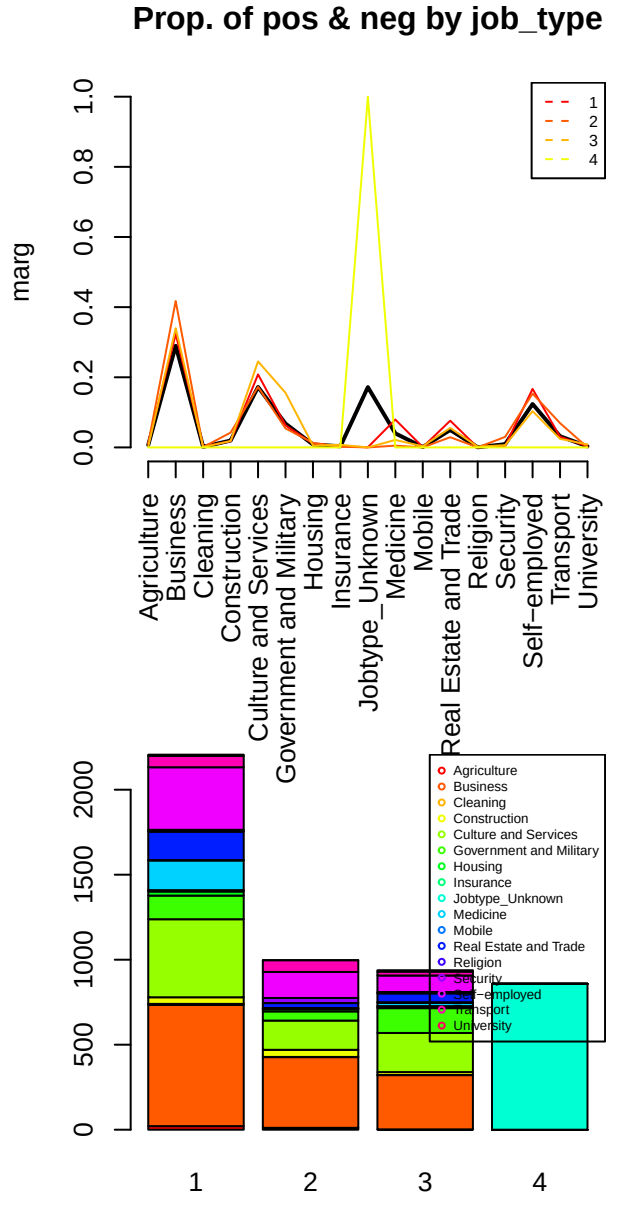


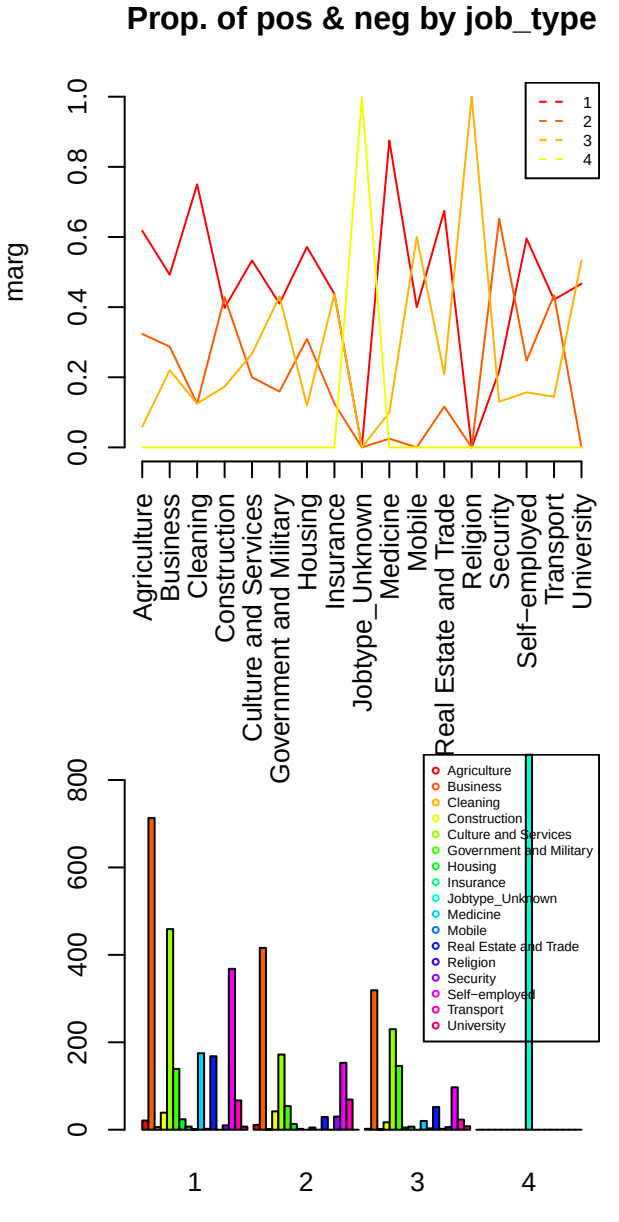


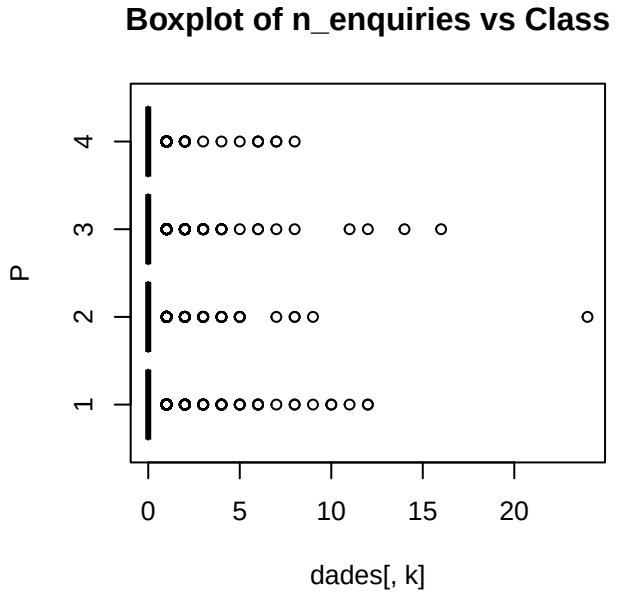


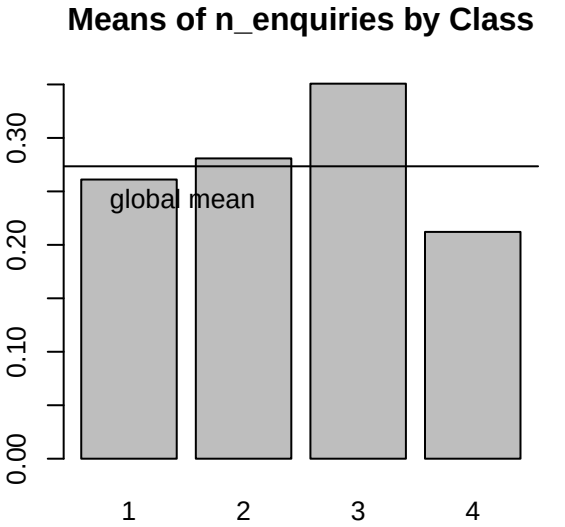
Prop. of pos & neg by job_type

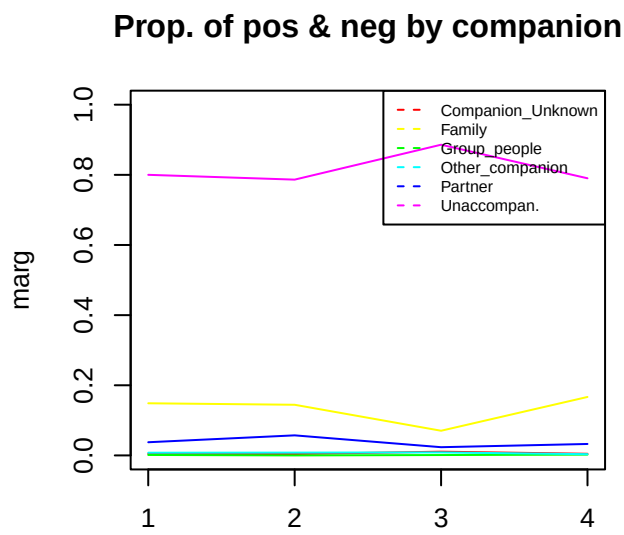


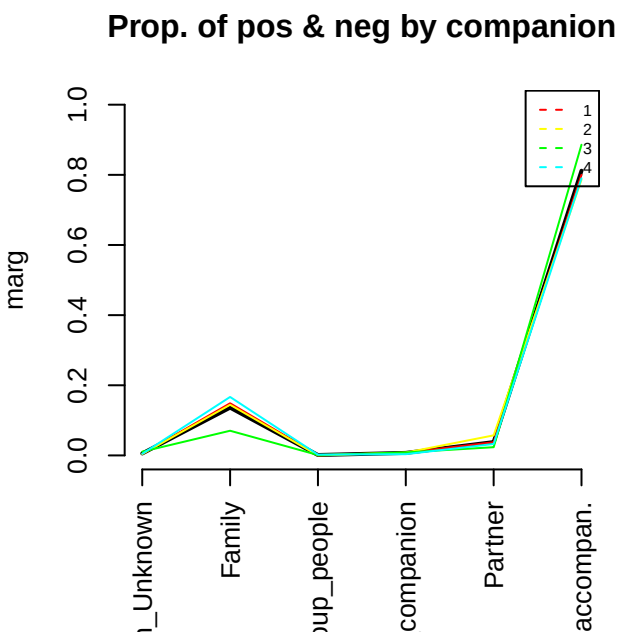

Prop. of pos & neg by job_type
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
Agriculture
Business
Cleaning
Construction
Culture and Services
Government and Military
Housing
Insurance
Jobtype_Unknown
Medicine
Mobile
Real Estate and Trade
Religion
Security
Self−employed
Transport
University
0
500
1000
1500
2000


Prop. of pos & neg by job_type
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
Agriculture
Business
Cleaning
Construction
Culture and Services
Government and Military
Housing
Insurance
Jobtype_Unknown
Medicine
Mobile
Real Estate and Trade
Religion
Security
Self−employed
Transport
University
0
200
400
600
800


Boxplot of n_enquiries vs Class
P
1
2
3
4
0
5
10
15
20
dades[, k]


Means of n_enquiries by Class
0.00
0.10
0.20
0.30
global mean
1
2
3
4


Prop. of pos & neg by companion
marg
0.0
0.2
0.4
0.6
0.8
1.0
Companion_Unknown
Family
Group_people
Other_companion
Partner
Unaccompan.
1
2
3
4


Prop. of pos & neg by companion
marg
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
Family
Partner

Prop. of pos & neg by companion