DATAMINING

Bank Term Deposit

Final deliverable

28 October 2021

Motivation of the work and general description of the problem to be analysed

A term deposit is a savings product often supplied by banks to their customers in which the customer would deposit money in exchange for interest paid on their deposited money. This can be seen as the customer making a loan to the bank for a set amount of time. There are economic incentives for the banks too in this arrangement. The deposited money increases the bank's assets and is in the end intended to increase the money supply. Thus there are economic benefits for both parties, if managed right. However, if the customer would change their mind and want to withdraw the money before the set out date, that would result in an economic penalty. Because of the potential economic benefits for the banks it is in their interest to be able to predict if a customer would make a term deposit.

In order to make these predictions, a number of variables are used that are connected to different individuals' financial health. Additionally, there are other variables storing information about previous interactions between bank and customer from previous marketing campaigns. As it is in the banks' interest to attract more term deposits they use, for example, marketing campaigns over phone, which can be seen in this dataset. These different variables and their impact can be explored with PCA, to facilitate further analysis of this multidimensional data, and clustering methods to try to discover patterns.

This report sets out to explore the bank term deposit dataset to see if there are any interesting patterns and characteristics between different groups of customers. Any potential findings are of use for the banks to increase their margin of profit but also to wider society, because the behaviours seen in this project could tell us much of how society works. On the other hand, if banks were to base their decision making on any potential findings presented in this study for their marketing campaigns, it could be discriminatory towards people whose scores do not lead to a positive outcome in the term deposit prediction. These people could then miss out on the opportunity to earn interest on their savings, which would be an economic loss. Therefore, any potential findings must be used with caution and the ethical implications must be discussed.

Data Source presentation

The data were retrieved from Kaggle, which is an online community of machine learning practitioners and data scientists. The URL to the website is https://www.kaggle.com/faviovaz/bank-term-deposit. The data was downloaded from the website as a CSV file to be further processed. The data consist of a collection of variables that could be used to predict whether a bank customer would subscribe to a term deposit. To make this prediction, the variables hold relevant information about individuals. This information ranges from personal details, to information about the customer's personal finance. There are also variables holding data about the interactions between the customers and the bank. In the following section we set out to further describe the data.

Formal description of Data structure and metadata

a) What rows of data matrix contain?

The rows of the data matrix each correspond to an individual that is in some way interesting to the bank. These individuals could already be customers of the bank or potential ones that are interesting to the marketing campaigns of the bank. The columns of the data matrix then each hold relevant information about these individuals. Further information about these variables is given in the following section with the Metadata Table appended.

b) Metadata Table

#	Variable	Description	Type	Values	Rephrasing Values	Measuring unit	Missing code	Range	Role
1	age	Age of customer	Numerical Discrete			years	NA	[19,89]	explanatory
2	job	Type of job	Categorical Nominal	admin, blue-collar, entrepreneur, housemaid, management, retired,self-employed, services, student, technician, unemployed, unknown	adm, bcol, ent, hous, mana, ret, semp, serv, stu, tech, unem, unk		unknown		explanatory
3	marital	Marital status	Categorical Nominal	divorced, married, single	div, marr, sing				explanatory
4	balance	Average yearly balance, in euros	Numerical Discrete				NA	[-3313,71188]	explanatory
5	education	Level of education	Categorical Ordinal	primary, secondary, tertiary, unknown	prim, sec, ter, unk		unknown		explanatory
6	default	Has credit in default?	Categorical Binary	yes, no	yes, no				explanatory
7	housing	Has housing loan?	Categorical Binary	yes, no	yes, no				explanatory
8	loan	Has personal loan?	Categorical Binary	yes, no	yes, no				explanatory

9	contact	Contact communication type	Categorical Nominal	cellular, telephone, unknown	cell, tel, unk		unknown		explanatory
10	month	Last contact month of year	Date						explanatory
11	day	Last contact day of the week	Date						explanatory
12	duration	Last contact duration, in seconds	Numerical Discrete			seconds		[4,3025]	explanatory
13	campaign	Number of contacts performed during this campaign and for this client	Numerical Discrete					[1,50]	explanatory
14	pdays	Number of days that passed by after the client was last contacted from a previous campaign	Numerical Discrete					[-1,871]	explanatory
15	previous	Number of contacts performed before this campaign and for this client	Numerical Discrete					[0,25]	explanatory
16	poutcome	Outcome of the previous marketing campaign of employees	Categorical Nominal	failure, other, success, unknown	fail, oth, suc, unk				explanatory
17	y	Has the client subscribed to a term deposit?	Categorical Binary	yes, no	yes, no				response

c) Final scope of the study with inclusion and exclusion criteria for both rows and columns

Since the dataset, when retrieved from the data source, contains variables that are interesting for the different methods used in this study, it was decided to maintain the dataset as it was. As will be outlined in the outliers section of the report, no outliers were identified and were thus not excluded. However, an omission that was made was that of the variable contact. The reason why is discussed in the preprocessing section.

Complete Data Mining process performed

The complete data mining process that was performed centered around the sequence of steps specified in lectures. More specifically, the steps aligned with the workflow outlined by Fayyad et al. 1996 and can be seen in the following figure.

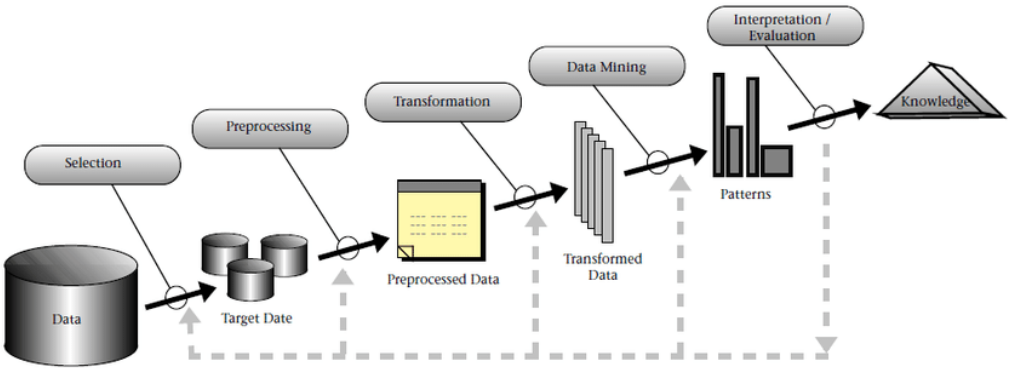


Fig. 0 Data Mining process

The first step of the data mining process was to define the problem. We decided to investigate if there were any identifiable clusters in the bank deposit data and the different characteristics of these different variables. The data, as written in the data source presentation, were collected from a dataset found on Kaggle. The data cleaning steps will be outlined in the preprocessing section, where it is explained how missings were treated with the KNN method and the reasoning behind the omissions made. Steps taken for dimensionality reduction are described in the PCA analysis section. The data mining technique used in this study is hierarchical clustering, which was deemed appropriate for the purpose. More specifically, it is performed with Gower distance calculation, which is compatible with mixed numeric and categorical variables. More information about how the clustering was performed and the observations made is provided in the hierarchical clustering section. The data mining step involved interpreting the clusters and profiling them, which will be done in the profiling of clusters section. The last step of the data mining process will be dealt with in the global discussion and general conclusions section, where interpretations and evaluations of the results will take place.

Evidently, the data mining process sets up very ambitious goals. These different steps have been completed under strict deadlines and have resulted in various previous documents. Parts of these previous documents have been reused for this final delivery,

which in essence is a detailed description of how the data mining process defined by Fayyad et al. was used to analyse the factors behind bank term deposits.

Detailed description of Preprocessing and data preparation.

The only variables that had missing data were age, job, balance, education and contact. We decided to treat them in three distinct ways.

First of all, we used KNN to approximate the missing ages and balances. Because these are the two most correlated variables the first step was to remove any entry that had both of them missing. Thus a total of 55 rows were deleted in this step.

After that we could start to use KNN on the variable age but we could not yet do that on the main dataset because the KNN function needs full numerical variables to work. In order to substitute the missing values of age we therefore needed to create an auxiliary dataset where we removed all entries with missing balances. Using the auxiliary set we computed KNN and substituted the missing ages with their approximations.

With that done we were able to recover the rows we had previously removed for having missing balance values and use KNN again to estimate them.

Secondly, we preprocessed the missing values in job and education and concluded that they were neither very abundant nor problematic so we decided to keep them and interpret them later while analysing the data.

Finally, we observed that the variable contact is missing in almost a third of all entries. Additionally, we did not consider it to have that much value for our intended predictions. Therefore we decided to remove it altogether from the dataset.

Outliers

We analysed all the numerical variables by making plots in order to remove all the data that could be an error.

Firstly, we analysed each numerical variable to see all possible outliers:

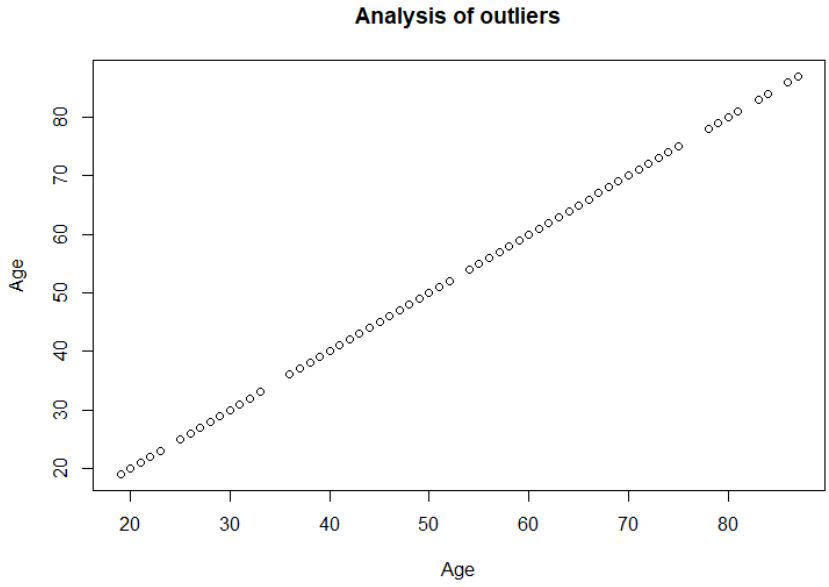


Fig. 1.1 Analysis of outliers - Age

The minimum and maximum age that we observed in Figure 1.1 were 19 and 87 years old respectively. This is a reasonable age span for potential bank customers. Therefore we concluded that there were no outliers for the variable age.

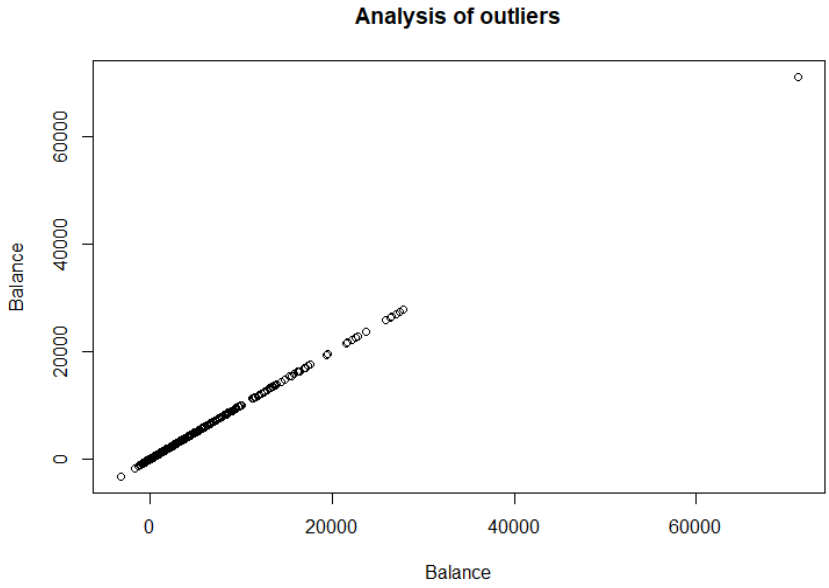


Fig. 1.2 Analysis of outliers - Balance

For the balance variable we were able to observe in Figure 1.2 that there appeared to exist an individual who had a much higher balance in their bank account. In order to determine if this was an outlier, we checked the value for the balance variable in this row, which was €71188. This did not appear to be an unreasonable amount for someone to have in their bank account, nor did it appear to be a matter of human error. Therefore we did not discard this row.

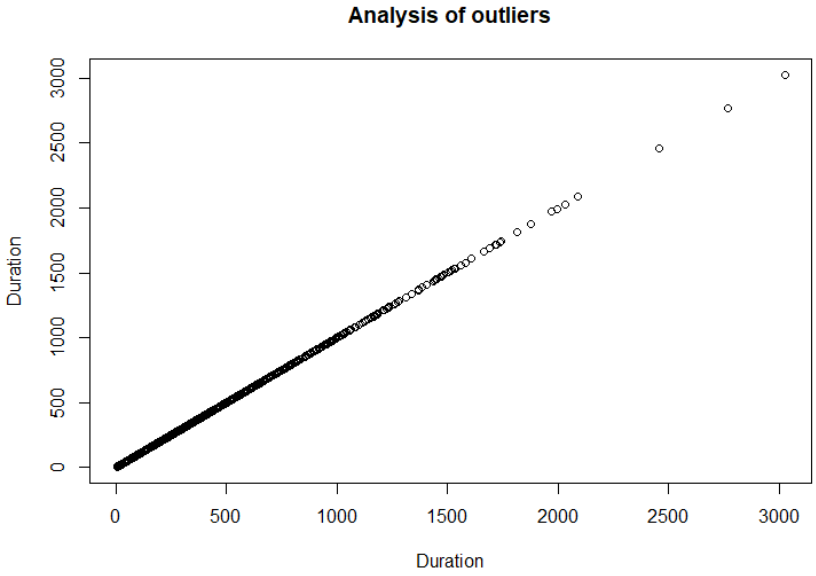


Fig. 1.3 Analysis of outliers - Duration

The maximum value we saw in Figure 1.3 for the duration variable, which is the duration of the last contact, was 3025 seconds (50.42 minutes). It is a reasonable time for a conversation, so it was not determined to be an outlier.

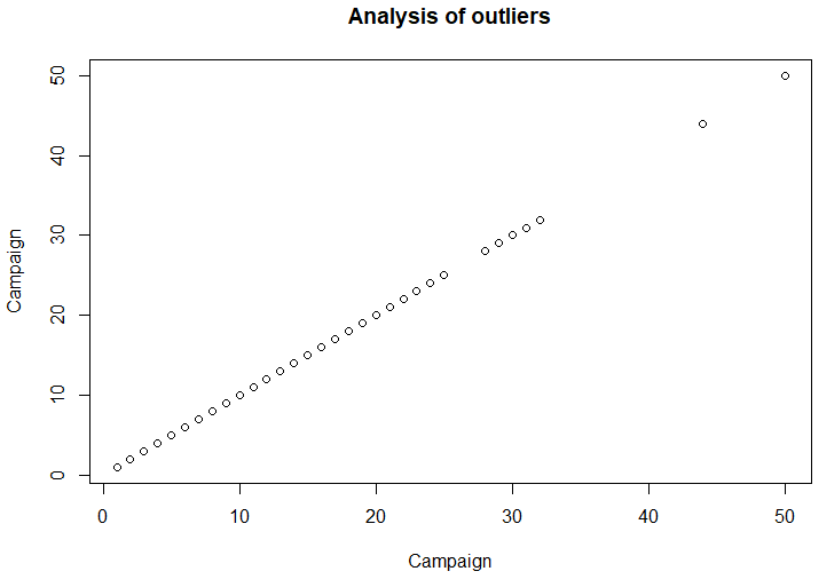


Fig. 1.4 Analysis of outliers - Campaign

In the campaign variable, which represents the number of times a single customer has been contacted during the campaign, we could observe two potential outliers in Figure 1.4, 44 and 50. That meant that two individuals would have been interacted with 44 and 50 times respectively, during the same campaign. We considered that to be a high number, but in comparison with the other individuals we did not consider these two instances to be outliers, as there were other individuals that were contacted more than 30 times. These two cases were not that far away from those.

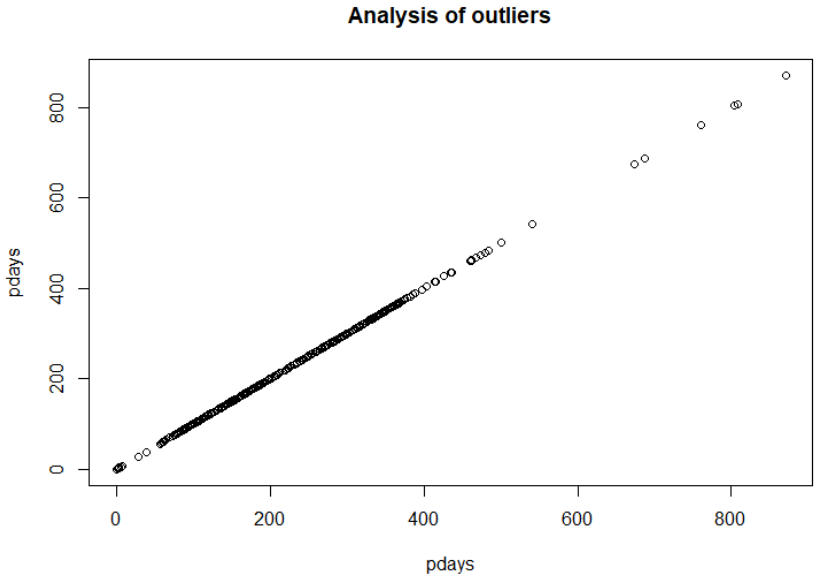


Fig. 1.5 Analysis of outliers - pdays

For the pdays variable, which represents the number of days since the client was last contacted from a previous campaign, we did not identify any outliers. Although there were some higher values, as can be seen in Figure 1.5, we did not identify these as outliers, since there are individuals that do not want to be contacted by a marketing campaign. These individuals would therefore have very large pdays values.

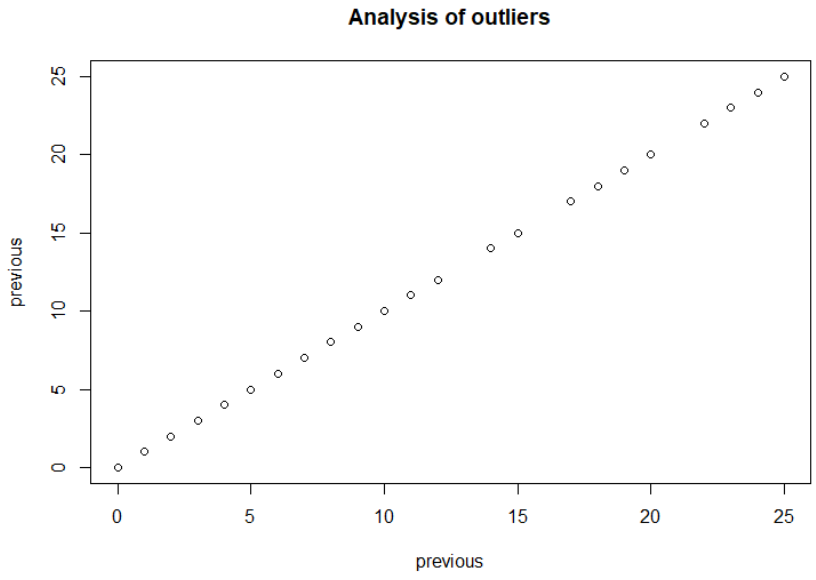


Fig. 1.6 Analysis of outliers - previous

By looking at Figure 1.6 we concluded that there were no outliers for the variable previous. To summarise this section, we did not identify nor discard any outliers.

Basic statistical descriptive analysis

a) Univariate for all the variables included in the study

Name of the variable:	**Age**
Minimum value:	19 years
Maximum value:	87 years
Mean:	41.46 years
Median:	40 years
Variance:	0.256 years^2
Standard deviation:	10.642 years

Histogram and box plot to see the type of distribution:

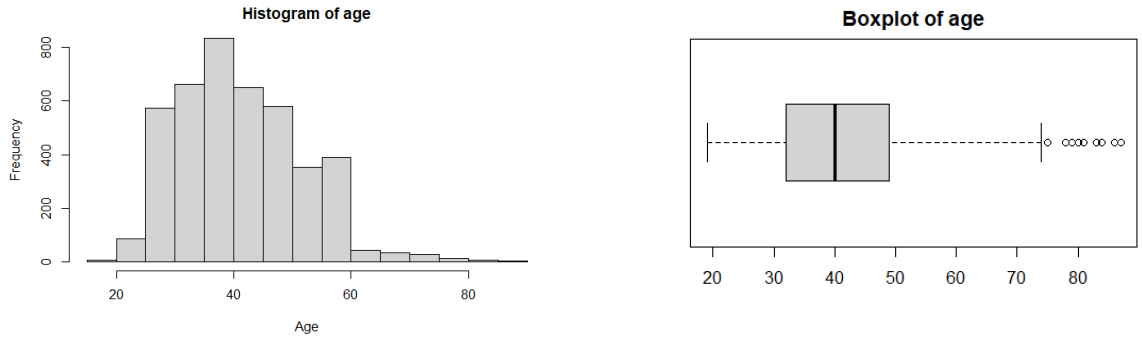
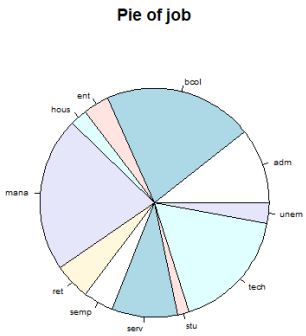


Fig. 2.1 & 2.2 Histogram and bar plot of variable age

Here we can conclude that our data follows a normal distribution for the age variable. Also we can observe that we have a higher frequency between the ages of 25 and 60 years old.

Name of the variable:	**Job**

Number of modalities: **11**



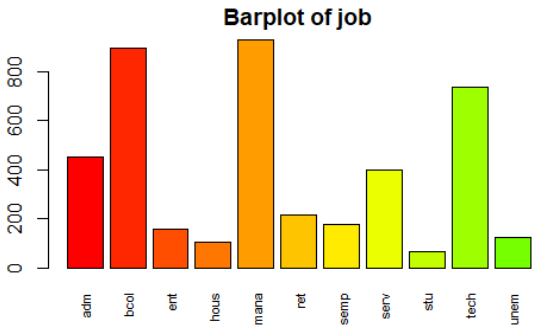


Fig. 2.3 & 2.4 Pie-chart and bar plot of variable job

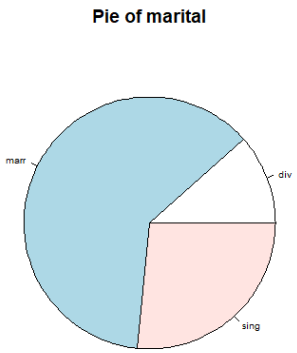
Frequency (and relative) table sorted

mana	bcol	tech	adm	serv	ret	semp	ent	unem	hous	stu
930	897	734	453	398	215	179	156	124	104	68
21.84%	21.07%	17.24%	10.64%	9.35%	5.05%	4.20%	3.66%	2.91%	2.44%	1.60%

We can observe that there are three predominant kind of jobs: managers, blue-collars and technicians

Name of the variable: **Marital**

Number of modalities:	**3**



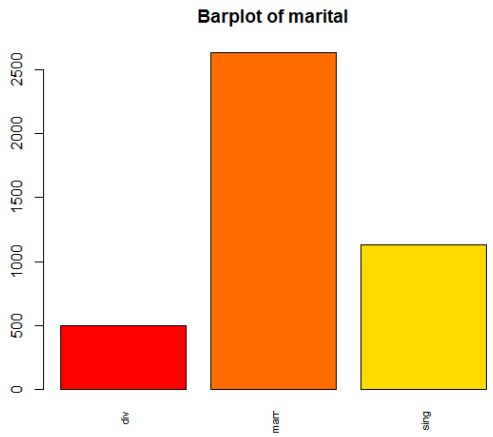


Fig. 2.5 & 2.6 Pie-chart and bar plot of variable marital

Frequency (and relative) table sorted

marr	sing	div
2632	1131	495
61.81%	26.56%	11.63%

The predominant marital status is married

Name of the variable:	**Balance**
Minimum value:	-3313
Maximum value:	71188
Mean:	1115.174
Median:	296
Variance:	2.421
Standard deviation:	2700.754

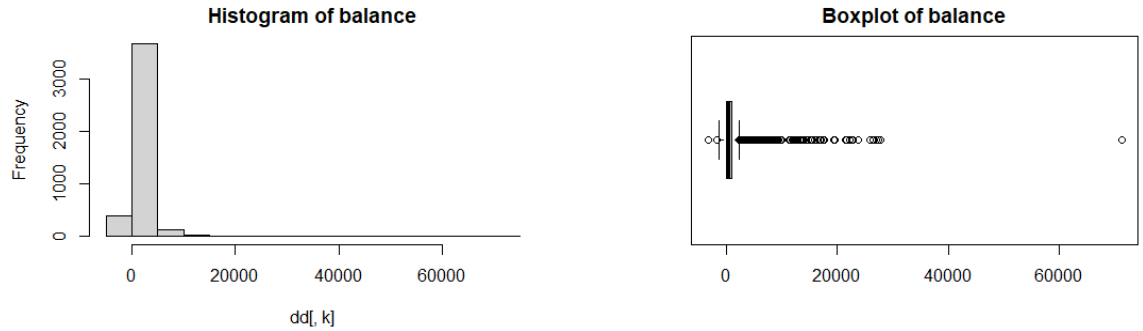


Fig. 2.7 & 2.8 Histogram and bar plot of variable balance

Balance is one of the most important variables in the dataset, as we can observe the mean is around 1000.

Name of the variable:	**Education**
Number of modalities:	**3**

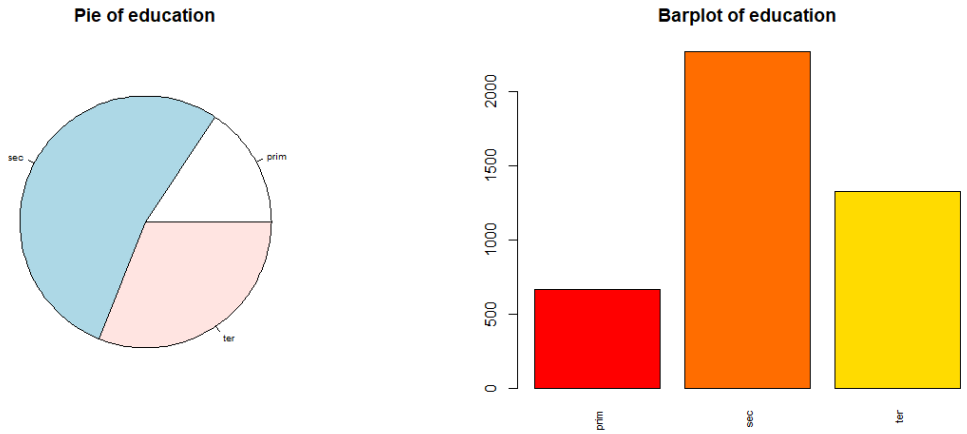


Fig. 2.9 & 2.10 Pie-chart and bar plot of variable education

Frequency (and relative) table sorted

sec	ter	prim
2265	1325	668
53.19%	31.12%	15.69%

Secondary education is the most common in the dataset

Name of the variable: **Default**

Number of modalities: **2**

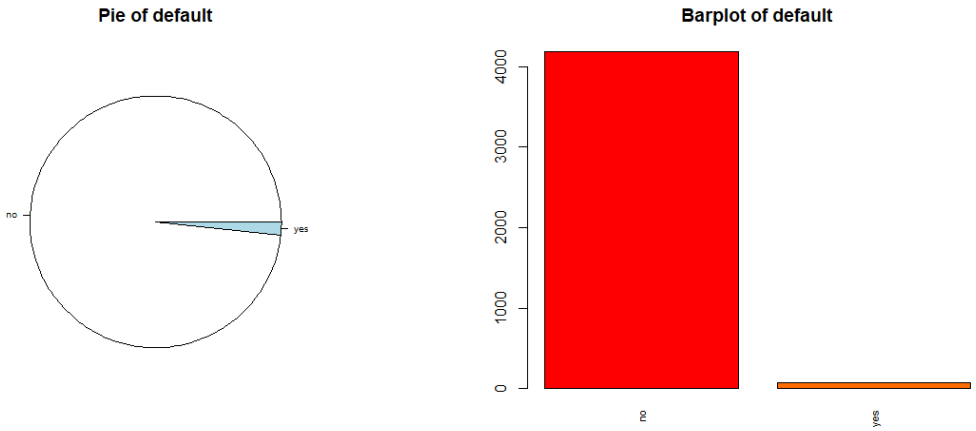


Fig. 2.11 & 2.12 Pie-chart and bar plot of variable default

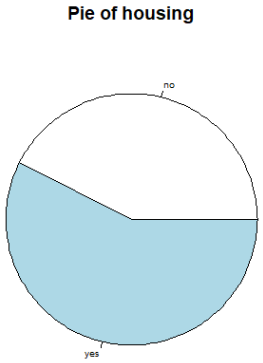
Frequency (and relative) table sorted

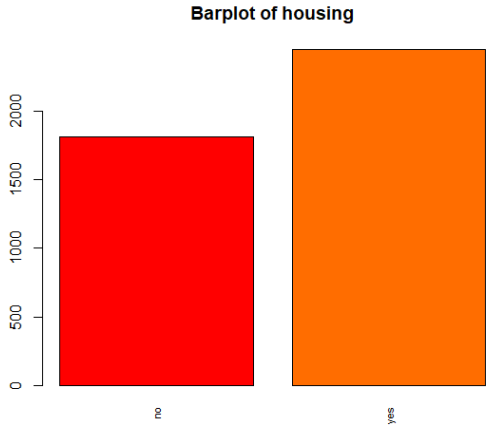
no	yes
4186	72
98.31%	1.69%

Practically none of the individuals have defaulted on their credit.

Name of the variable: **Housing**

Number of modalities: **2**





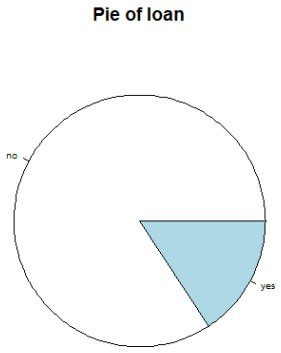
Frequency (and relative) table sorted

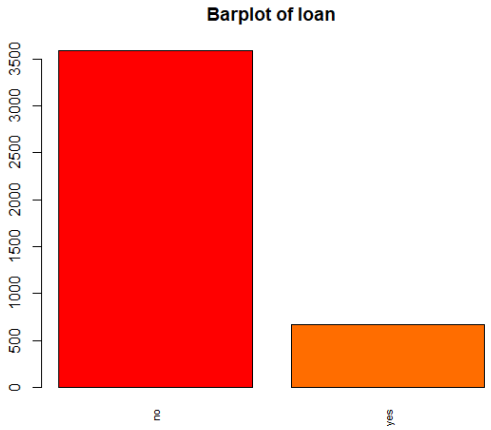
yes	no
2447	1811
57.47%	42.53%

We can observe that the majority of the individuals in the dataset have a mortgage.

Name of the variable: **Loan**

Number of modalities: **2**



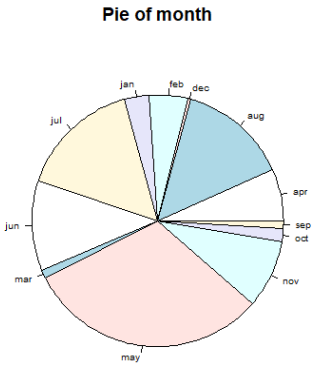


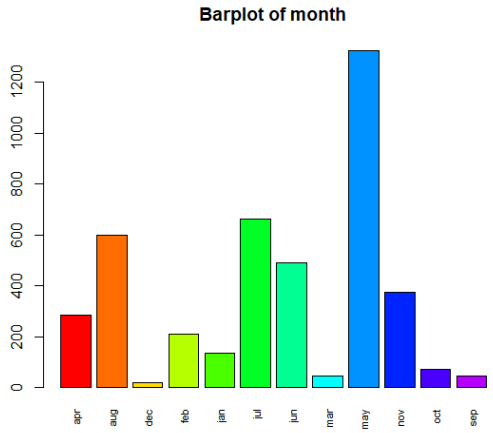
Frequency (and relative) table sorted

no	yes
3586	672
84.22%	15.78%

Around 84% of the individuals have no personal loan

Name of the variable:	**Month**
Number of modalities:	**12**



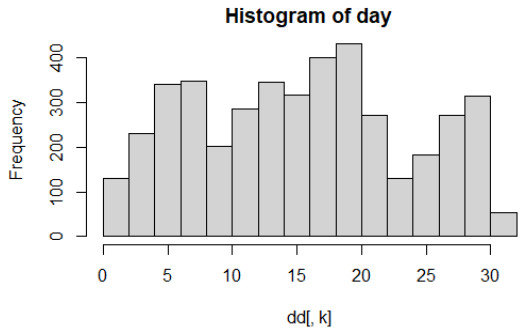


Frequency (and relative) table sorted

may	jul	aug	jun	nov	apr	feb	jan	oct	mar	sep	dec
1324	664	600	492	373	283	210	134	71	45	44	18
31.09%	15.59%	14.09%	11.55%	8.76%	6.65%	4.93%	3.15%	1.67%	1.06%	1.03%	0.42%

We can observe that the most of the contacts are on May

Name of the variable:	**Day**
Minimum value:	1
Maximum value:	31
Mean:	15.89
Median:	16
Variance:	0.516
Standard deviation:	8.204



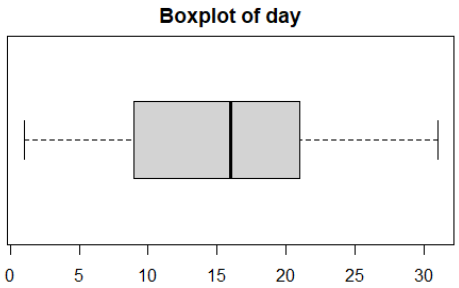
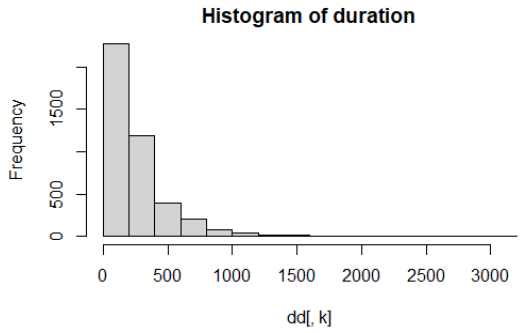


Fig. 2.13 & 2.14 Histogram and bar plot of variable day

Day variable, we can observe that the contacts are made around day 5, 15-20 and 26-29.

Name of the variable:	**Duration**
Minimum value:	4
Maximum value:	3025
Mean:	265.1
Median:	187
Variance:	0.985
Standard deviation:	261.2



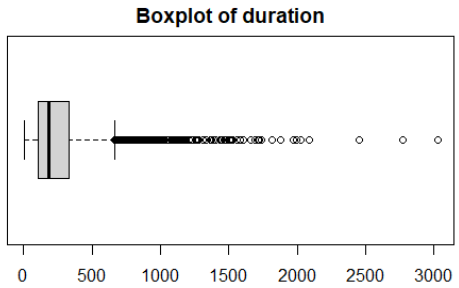
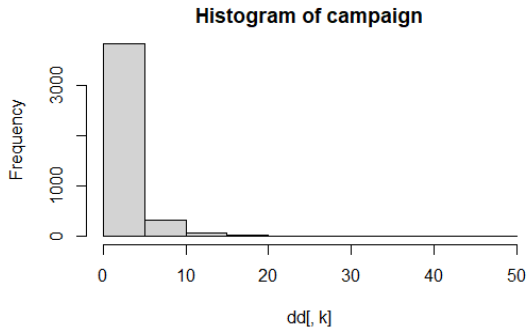


Fig. 2.15 & 2.16 Histogram and bar plot of variable duration

The common duration of a contact is around four minutes

Name of the variable:	**Campaign**
Minimum value:	1
Maximum value:	50
Mean:	2.806
Median:	2
Variance:	1.119
Standard deviation:	3.1402



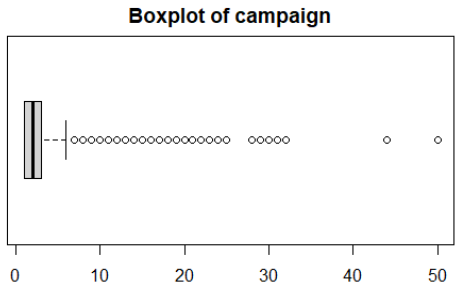
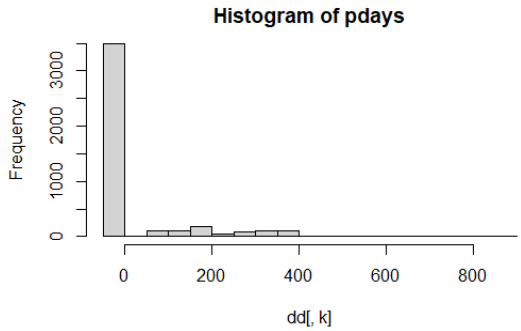


Fig. 2.17 & 2.18 Histogram and bar plot of variable campaign

The predominant number of contacts for the campaign is two. As we can observe, and detailed in the outlier section, there are a few individuals that have been contacted more than 30 times.

Name of the variable:	**Pdays**
Minimum value:	-1
Maximum value:	871
Mean:	40.01
Median:	-1
Variance:	2.509
Standard deviation:	100.417



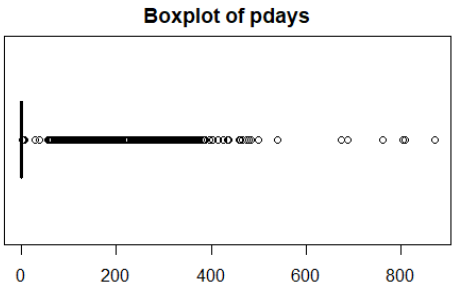
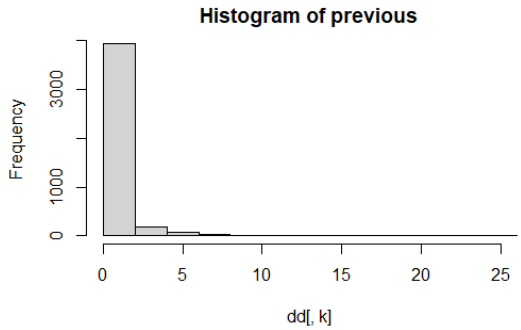


Fig. 2.19 & 2.20 Histogram and bar plot of variable pdays

The mean of days since a customer has been contacted from the last campaign are 40

Name of the variable:	**Previous**
Minimum value:	0
Maximum value:	25
Mean:	0.546
Median:	0
Variance:	3.130
Standard deviation:	1.709



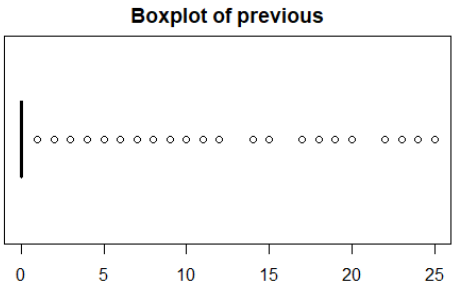


Fig. 2.21 & 2.22 Histogram and bar plot of variable previous

The mean of the previous contacts before this campaign is a half day

Name of the variable: **Poutcome**

Number of modalities: **4**

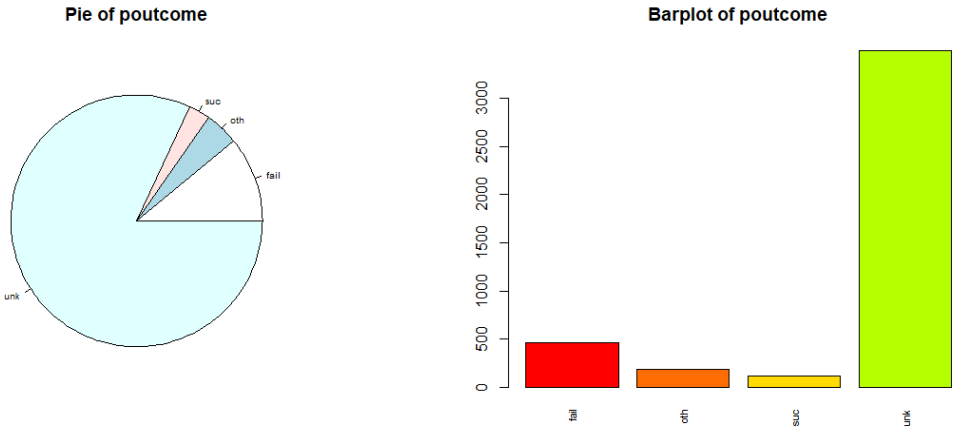


Fig. 2.23 & 2.24 Pie-chart and bar plot of variable poutcome

Frequency (and relative) table sorted

unk	fail	oth	suc
3490	468	186	114
81.96%	10.99%	4.37%	2.68%

Around 80% of individuals has unknown outcome of the previous campaign

Name of the variable: **Y**

Number of modalities: **4**

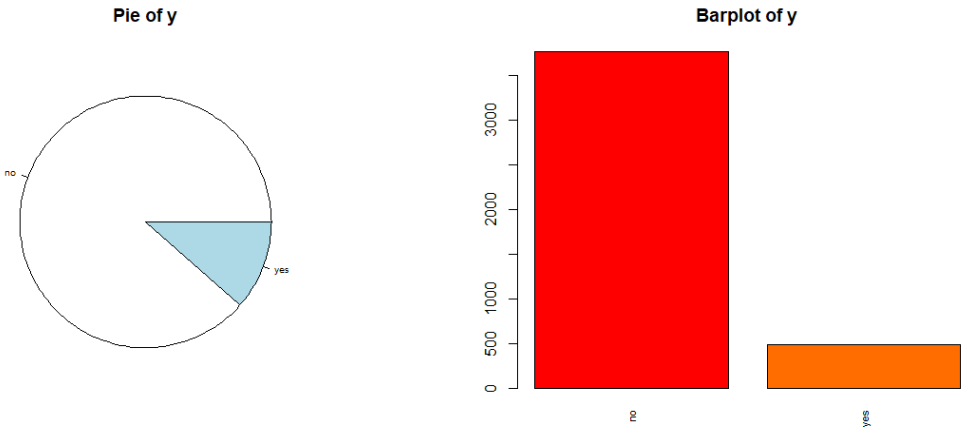


Fig. 2.25 & 2.26 Pie-chart and bar plot of variable y

Frequency (and relative) table sorted

no	**yes**
3768	490
88.49%	11.51%

Around 88% of the individuals has not been subscribed to a term deposit

b) Bivariate

Firstly we will see the correlation between all the numerical variables:

	age	balance	duration	campaign	pdays	previous
age	1	0.0644928	-0.001795	-0.012527	-0.006661	0.0084311
balance	0.0644928	1	-0.022096	-0.004631	0.0079317	0.0087489
duration	-0.001795	-0.022096	1	-0.070887	0.0011966	0.0142899
campaign	-0.012527	-0.004631	-0.070887	1	-0.092269	-0.064955
pdays	-0.006661	0.0079317	0.0011966	-0.092269	1	0.5752533
previous	0.0084311	0.0087489	0.0142899	-0.064955	0.5752533	1

As we can observe in the above table, we have weak direct or inverse relation between the variables. All the correlations are moving around 0, the only two variables that are strongly correlated are pdays and previous.

To more clearly describe all of the numerical variables we produced the following scatterplot matrix:

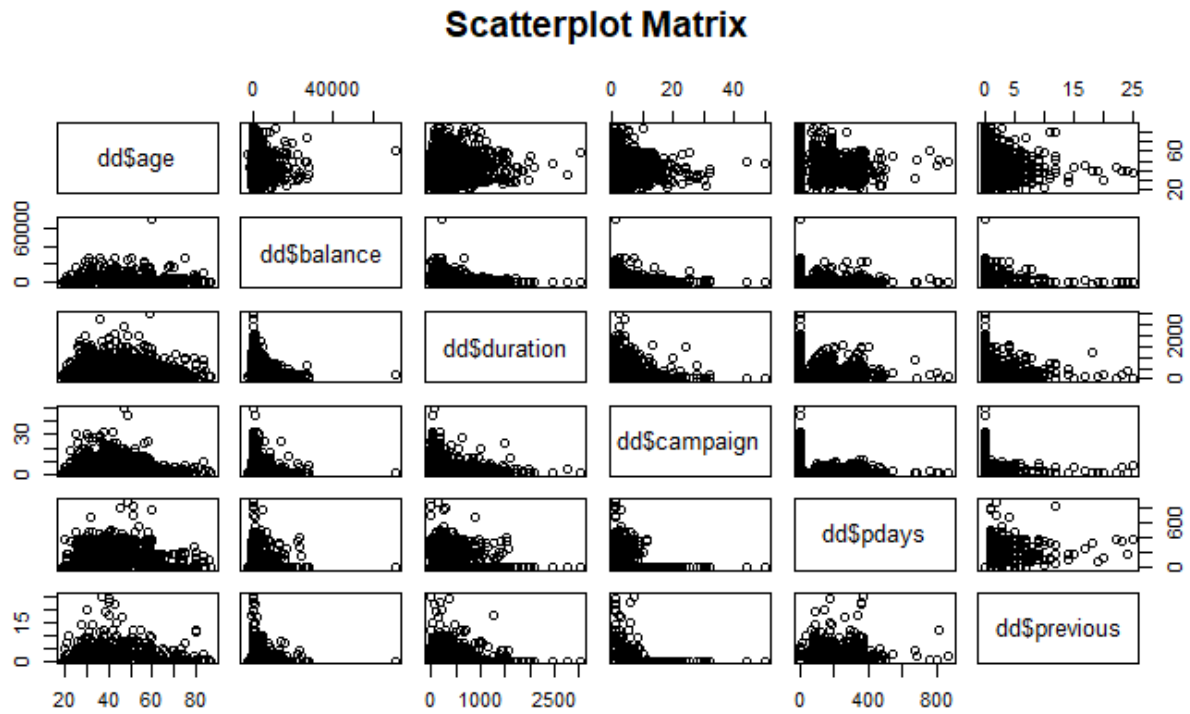


Fig. 2.27. Scatterplot matrix for age, balance, duration, campaign, days and previous

- Among age and balance variables, we can see that people with more balance are in an age between 25 to 60 years. Following a Gaussian form with the highest peak at 40-45 years.
- Among balance and duration variables, with more balance less duration of the contacts.

d) Conclude the section with one paragraph describing how is your data

The variable “age” follows a Gaussian distribution form.
We can observe that the variables have weak direct or indirect relationships, then when applying the clusters; they will not be biased towards specific types of variables.
With no clear correlation between variables, this could be more difficult to extract conclusions.

PCA analysis for numerical variables

We used Principal Component Analysis (PCA) to compute the principal components which were further used to perform a change of basis on the data.

For this to work, we had to be sure that we had no missing values in our variables, because of this, we used this method after treating missing values with KNN. Once the missing were treated we started with the PCA analysis. We have to know how many components we need to keep at least an 80% of total inertia.

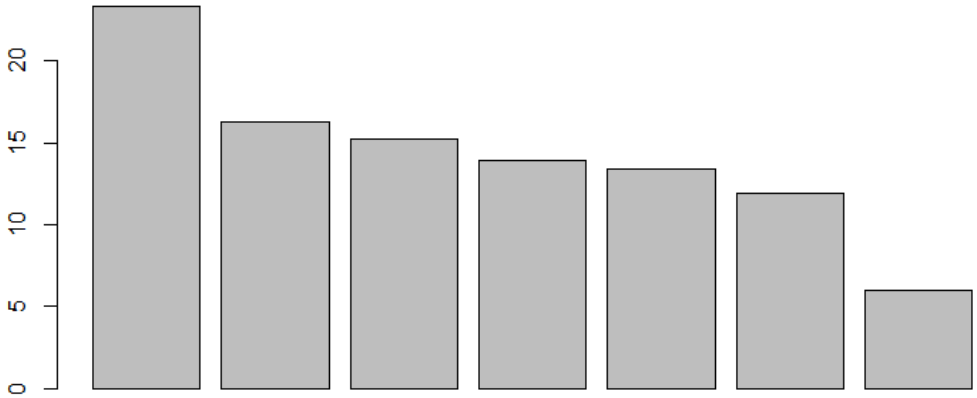


Fig. 3.1 Inertia of each PC

In Figure 3.1 we have each principal component with its total inertia while in Figure 3.2 we will be able to see the accumulated inertia in subspaces.

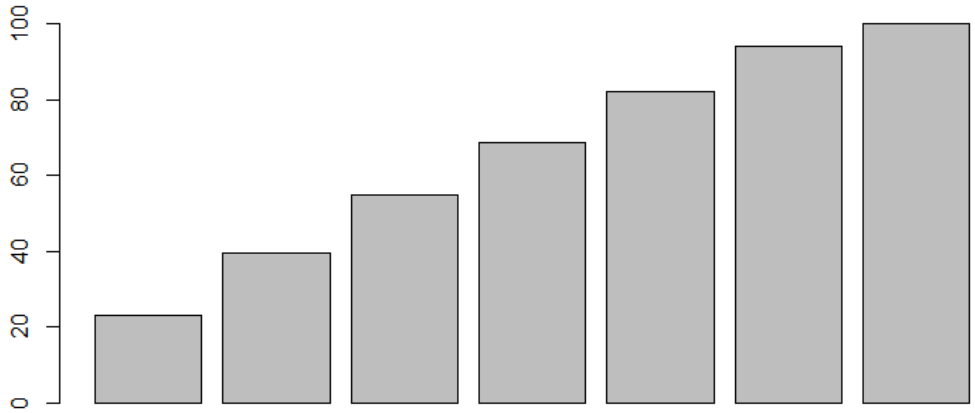


Fig. 3.2 Accumulated inertia on PC

If we take a closer look at Figure 3.2, we can see that using only five of the principal components will be enough to keep the required inertia.

The next step is to see how the data is placed in the plot with this new configuration, as we can see in the next figure, Figure 3.3.

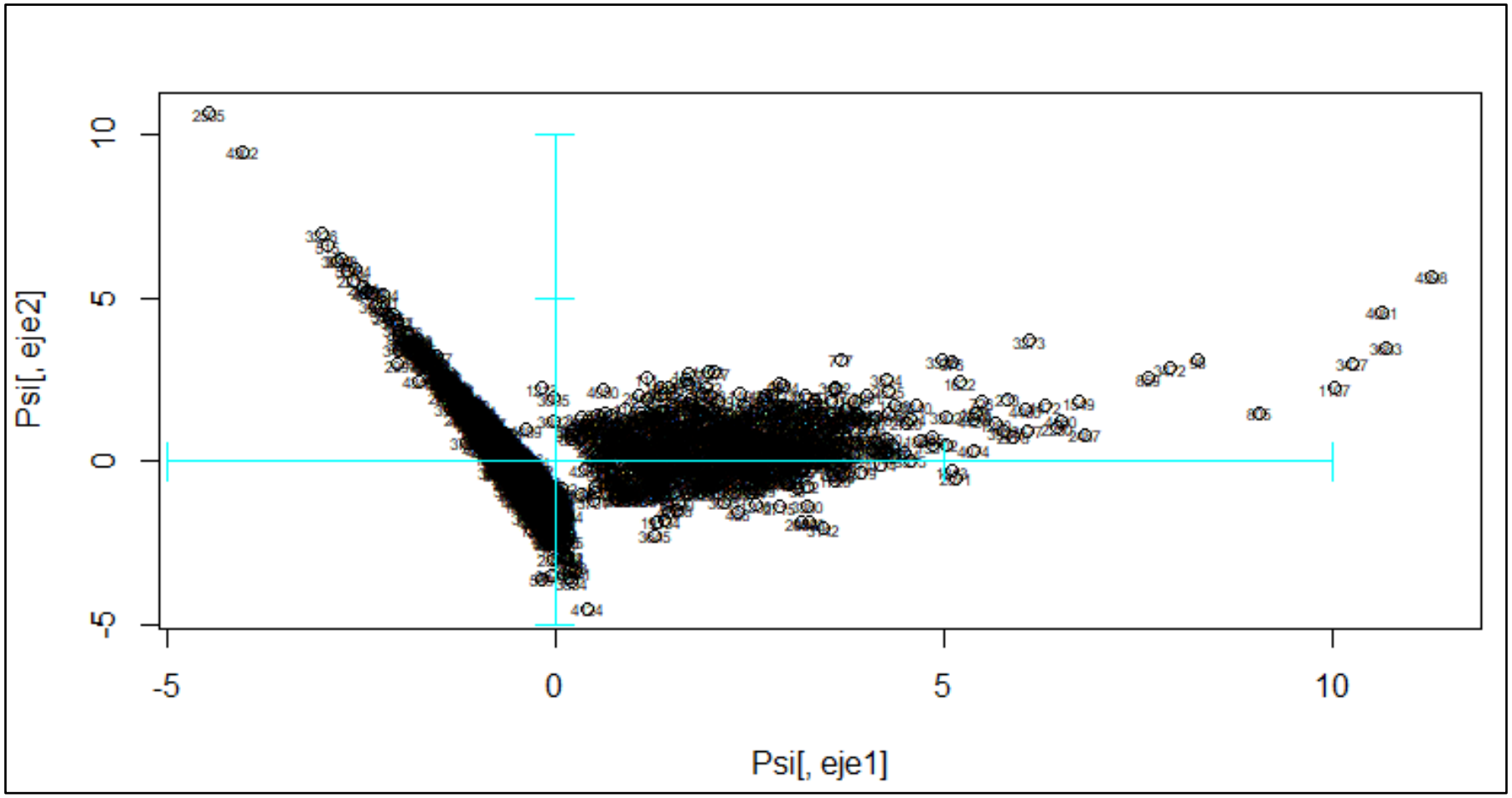


Fig. 3.3 Variables plotted in the new subspace

We can observe two different centroids, the reason why is for now unclear, so we have to find what influence the variables have and if there is any relationship between them. In the next figure we can see how these are plotted.

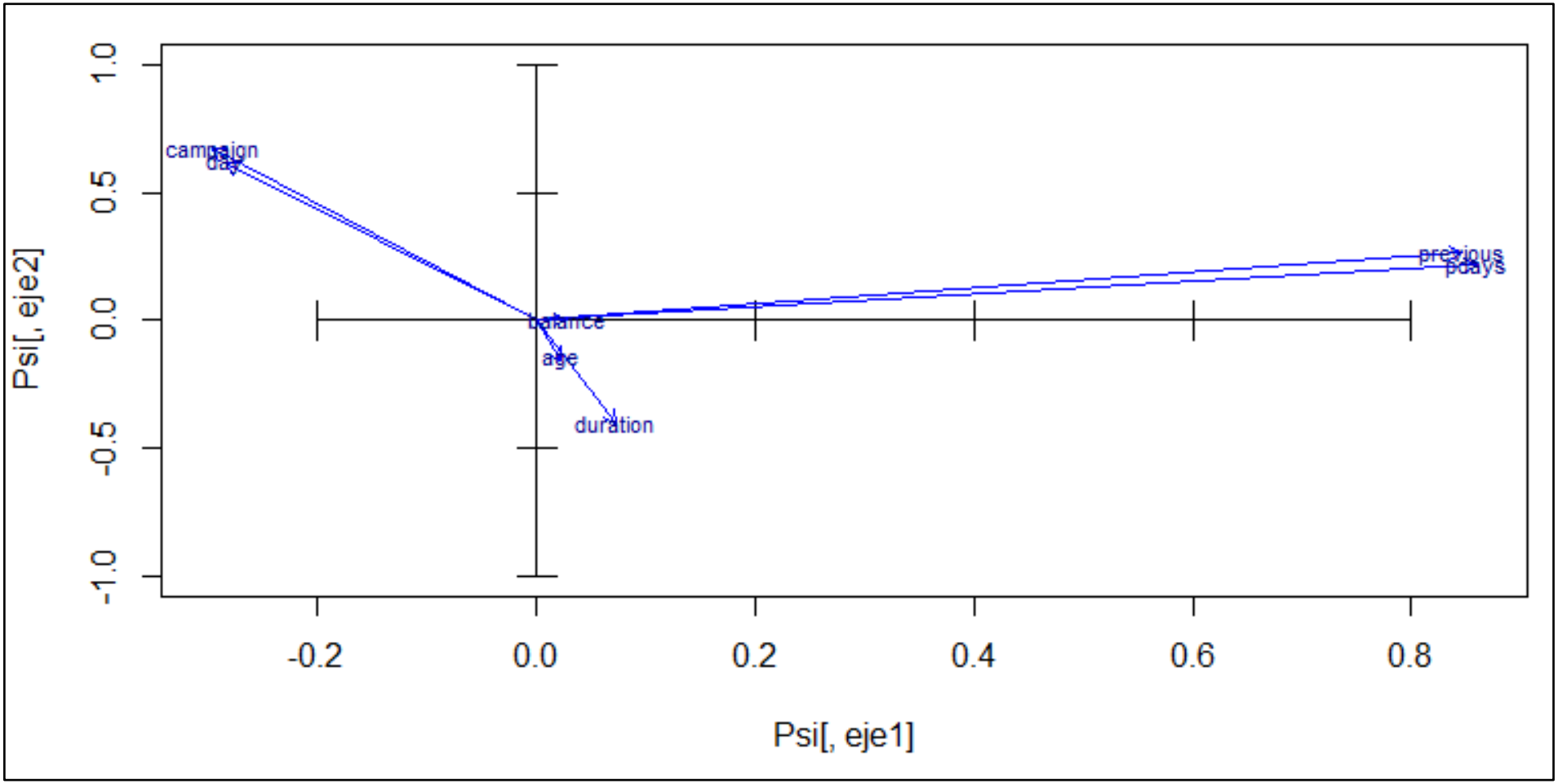


Fig. 3.4 representation of the PC in the new subspace

As we know the ones with a longer arrow and the ones that are closer to the axis are the ones with higher contribution. In this case, as can be seen in Figure 3.4, that would be the variables “previous” and “pdays”.

Using Figure 3.4 we can also try to read which variables are more closely connected. We could say that “previous” and “pdays” or “campaign” and “day” are correlated but to be sure we have to use a specific method to detect these correlations.

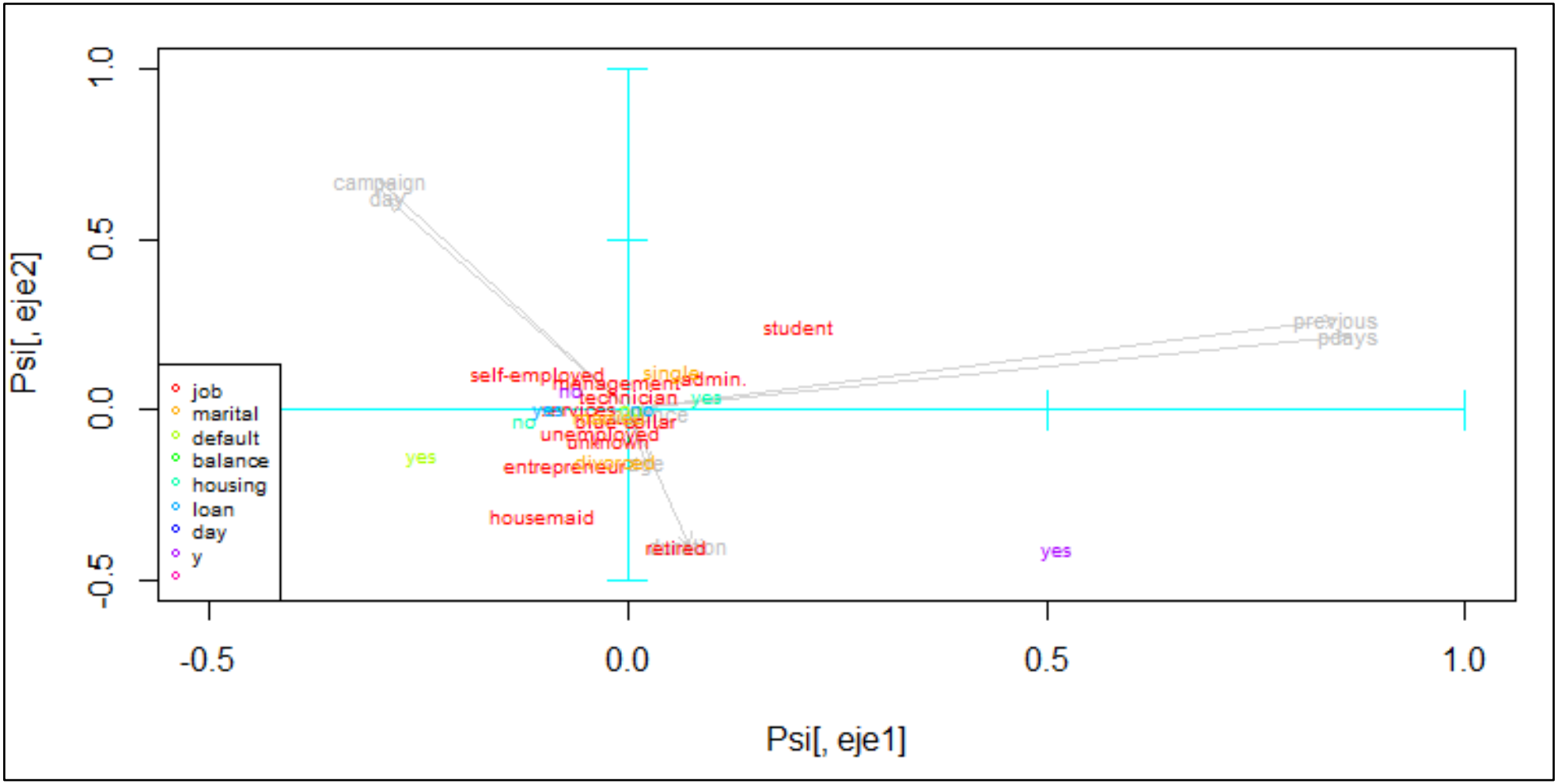


Fig. 3.5 representation of the PC in the new subspace and centroids

Figure 3.5 shows us how the categorical variables in our plot are distributed in centroids over the arrows in a light grey tone that represent our numerical variables.

Finally we had to try to divide our qualitative variables in the plot, after trying each one the only one which has given us the best performance to comprehend better our dataset and specifically this PCA analysis has been the “poutcome”.

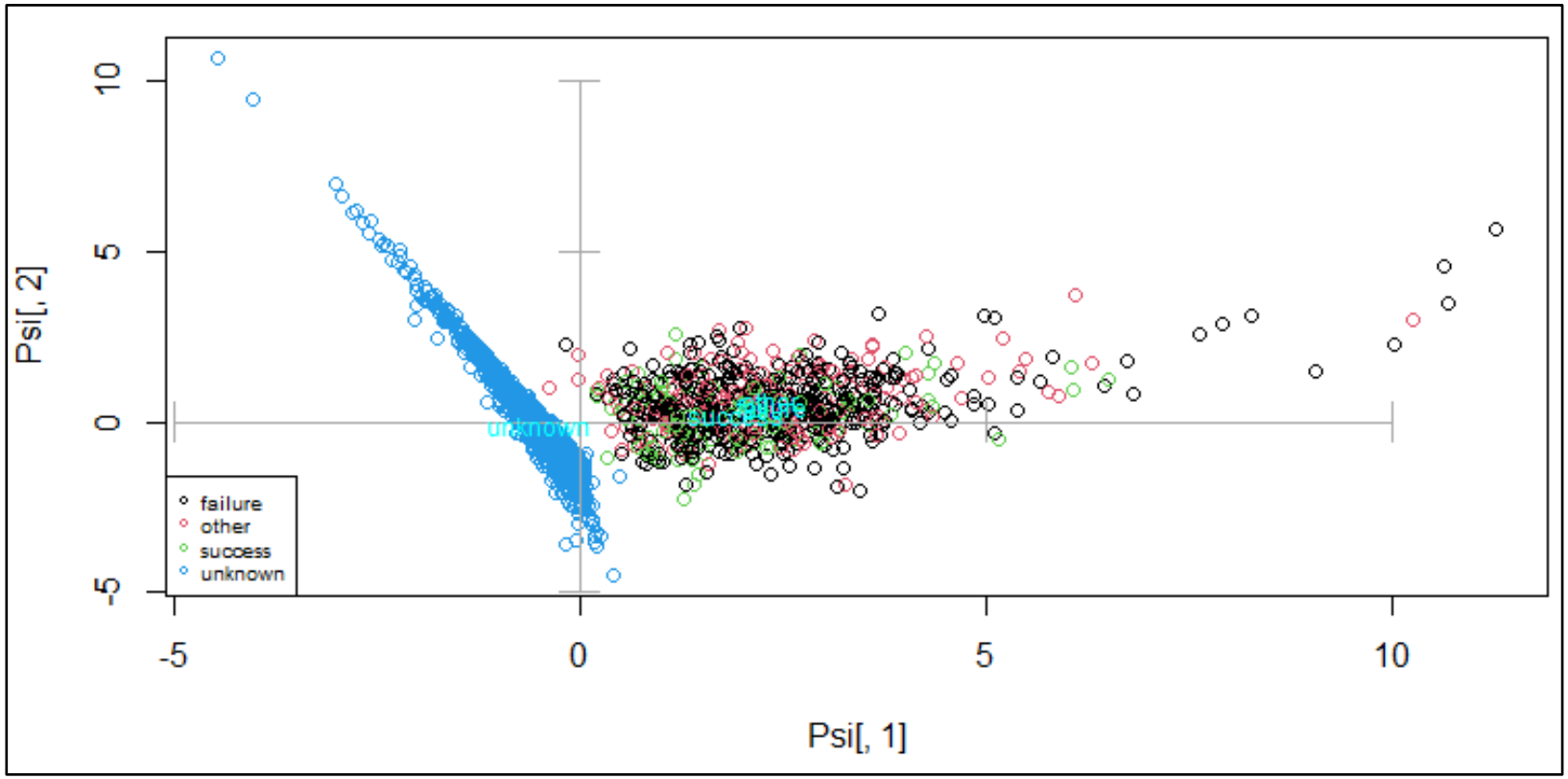


Fig. 3.6 Variables plotted in the new subspace sorted by its value in “poutcome”

As can be seen in Figure 3.6, all the variables with an “unknown” value are condensed in the blue section at the left of the plot while all the other sample is distributed in a centroid in the mid-right section.

We can assume that the variables with “unknown” as “poutcome” value follow the “campaign” and “day” Principal Components while the ones with all the other options are more attracted to their value in “previous” and “days” variables

Hierarchical Clustering on original data

In order to perform hierarchical clustering all the variables have been selected with the exception of the *contact* variable. It has been removed from the original dataset because it consisted mainly of *unknown* values and contributed very little information. That is, the variables used in clustering are: *age*, *job*, *marital*, *education*, *default*, *balance*, *housing*, *loan*, *day*, *month*, *duration*, *campaign*, *pdays*, *previous*, *poutcome* and *subscribed*.

To carry out the cluster analysis we have used Ward's method which is an agglomerative, hierarchical clustering procedure. The aggregation criteria that this method uses consists of merging two clusters if the increase of combined variance over the sum of the clusters' specific variances is the minimum compared to alternative merging options.
As for the metric, we employed the square of the Gower's dissimilarity coefficient because it can be used both with numerical and categorical variables.

After running the clustering procedure we have obtained the following dendrogram shown in Fig. 4.1.

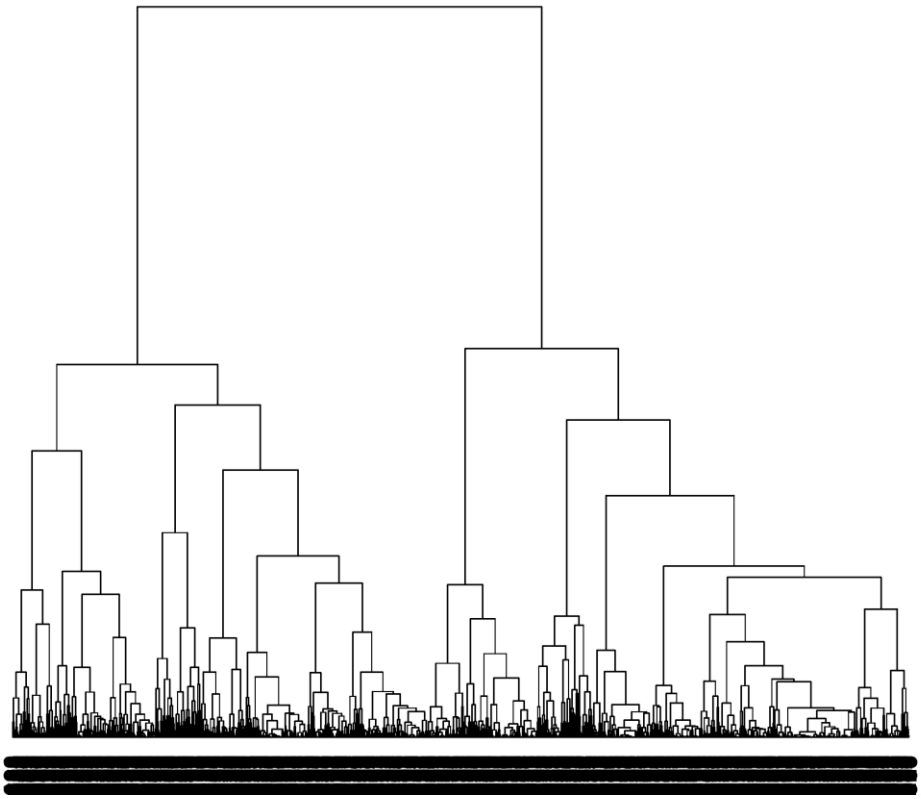


Fig. 4.1 Cluster dendrogram.

By inspecting the dendrogram and other graphics we have concluded that a good choice would be to cut the tree so that we end up with four clusters.
The resulting clusters and the cut line are represented in the Fig. 4.2.

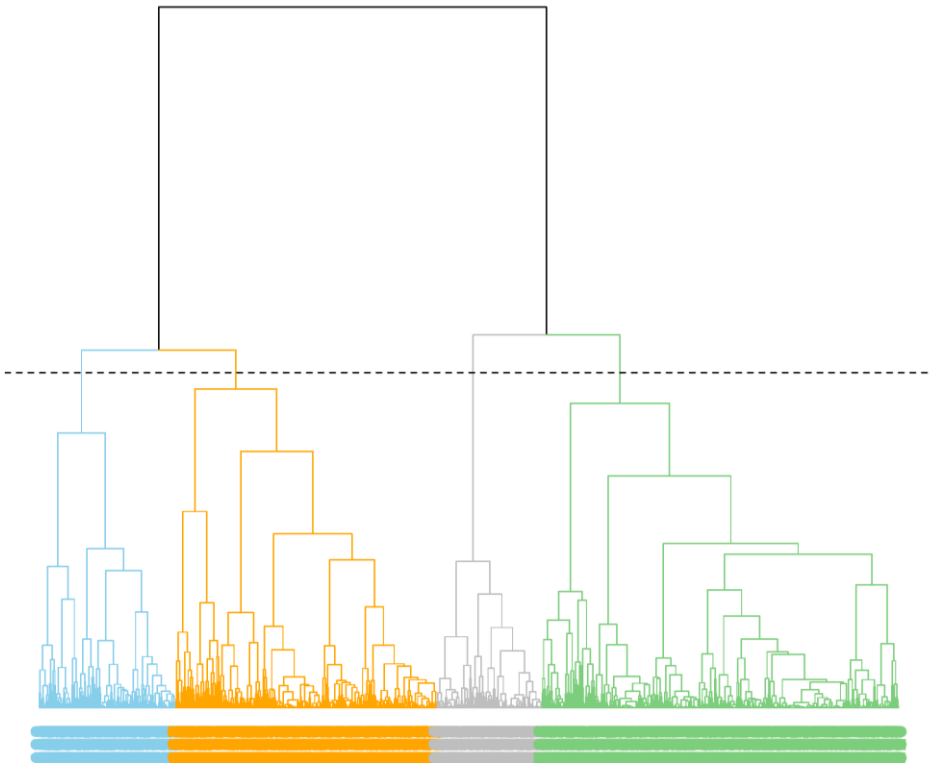


Fig. 4.2 Cluster dendrogram with marked classes and cutline.

— Class1, — Class 2, — Class 3, — Class 4, --- Cut line

The number of individuals in each cluster can be seen in table 4.1.

Cluster ID	Size
1	679
2	520
3	1766
4	1293

Table 4.1 Size of the clusters

Profiling of clusters

In order to identify which variables are significant in each cluster we used statistical tests and plots. For numerical variables, we used ANOVA for those that follow a normal distribution and Kruskal-Wallis otherwise. For categorical variables we used the χ^2 test.

Variable *age*

For the variable age we got the following barplot where it can be seen that there is some overlap but we think it is different enough to consider that it varies among classes.

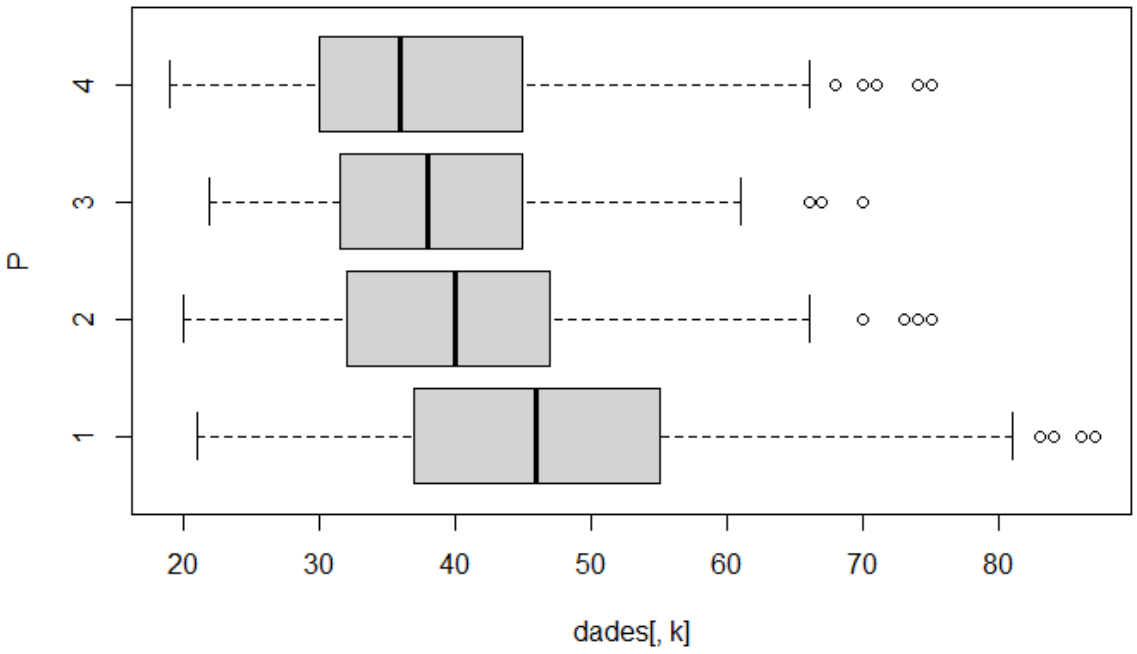


Fig. 5.1 Boxplot between ages and classes as result of clustering

Notice that the classes go from oldest to younger (the first class is the oldest and the fourth class the youngest).

Variable *job*

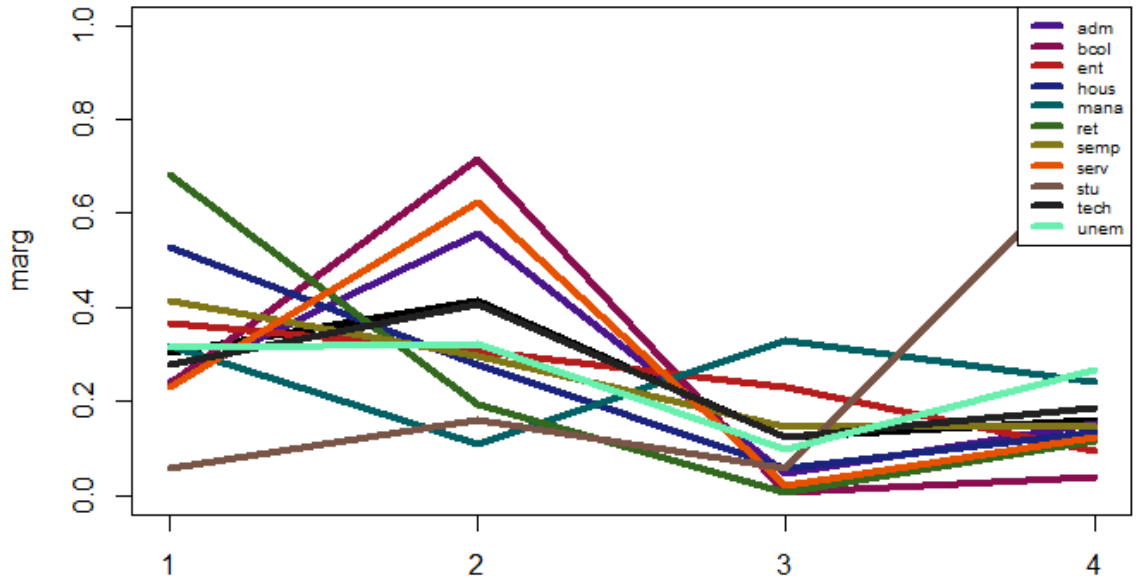


Fig. 5.2 Profiled of variable job without being class-conditioned

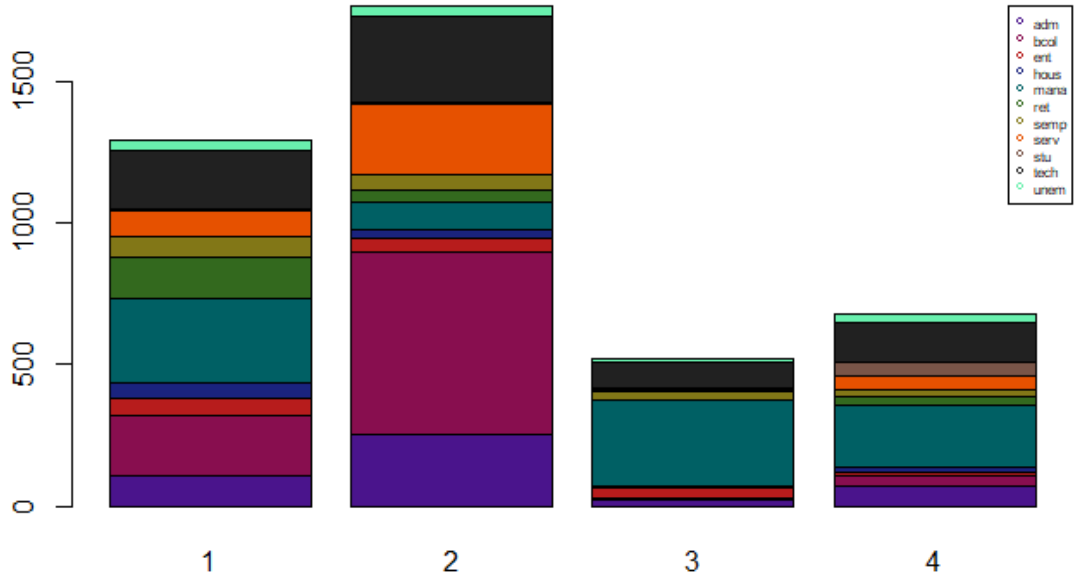


Fig. 5.3 Bar plot of clustering showing the variable job

Notice that in the first group most people's occupation is: retired or housemaid, but we also have a significant number of technicians, managers and blue-collar workers. For the second group most people work as blue-collar, service and administration. For the third class we have approximately the same number of managers as in the first group. And finally, for the fourth group we have almost all the students and many managers.

Variable *marital*

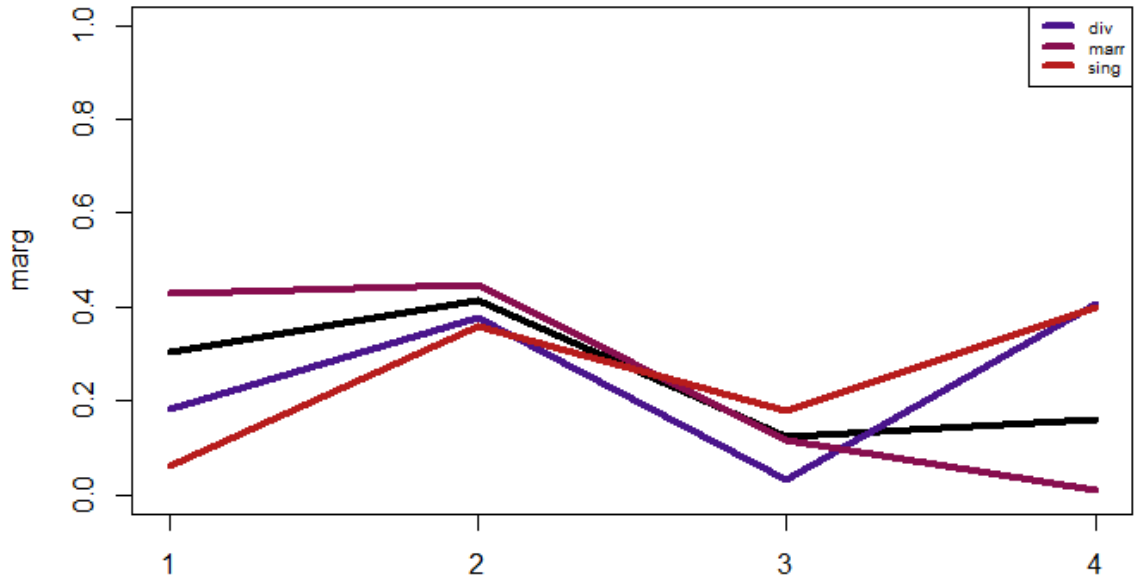


Fig. 5.4 Profiled of variable marital without being class-conditioned

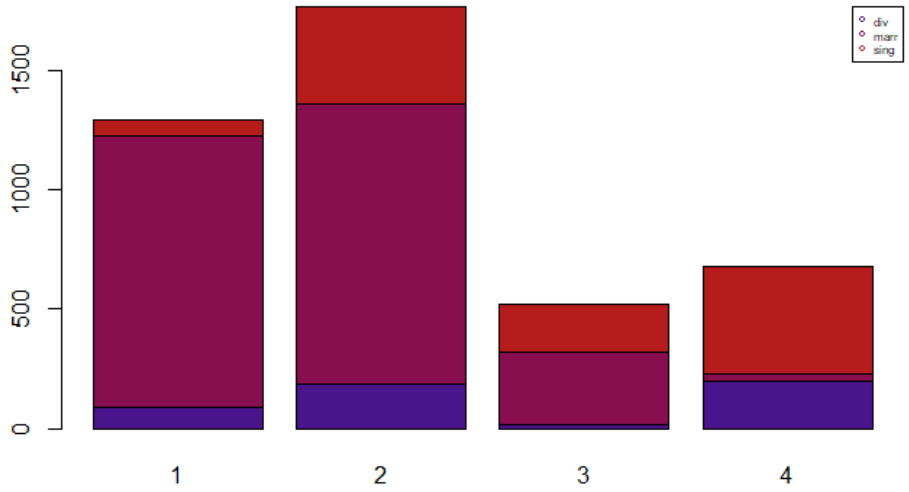


Fig. 5.5 Bar plot of clustering showing the variable marital

Notice that in the first class most people are married and in the last class most people are separated or single.

Variable *education*

As the χ^2 results show and from the barplot below we can say that the variable *education* is not the same for all the classes.

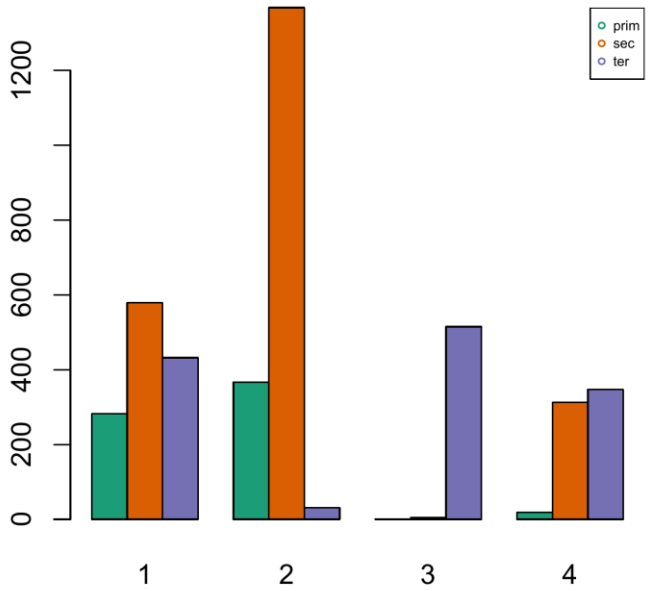


Fig. 5.6 Bar plot of clustering showing the variable education

From the plots below and the barplot we cannot conclude anything in particular about the first cluster. As for the second cluster, we could say that there are half of the persons from the total who have primary education and the majority of the persons that have a secondary education. In the third cluster the persons have almost exclusively a tertiary education and very few with primary or secondary education. In the fourth cluster there are very few people with a primary education and almost equal quantities of persons with secondary and tertiary education

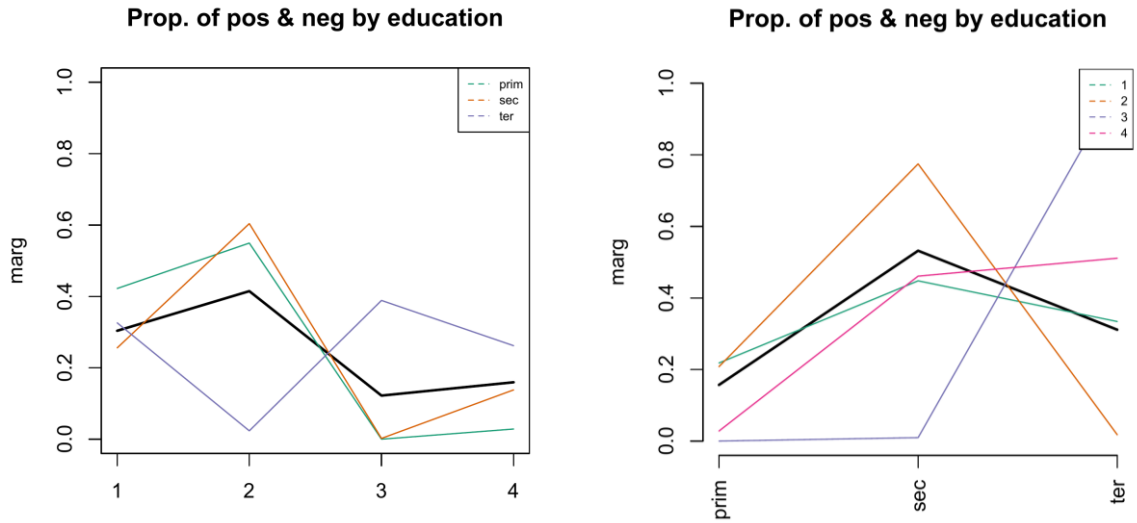


Fig. 5.7 Snake plots for the variable *education*

Variable *housing*

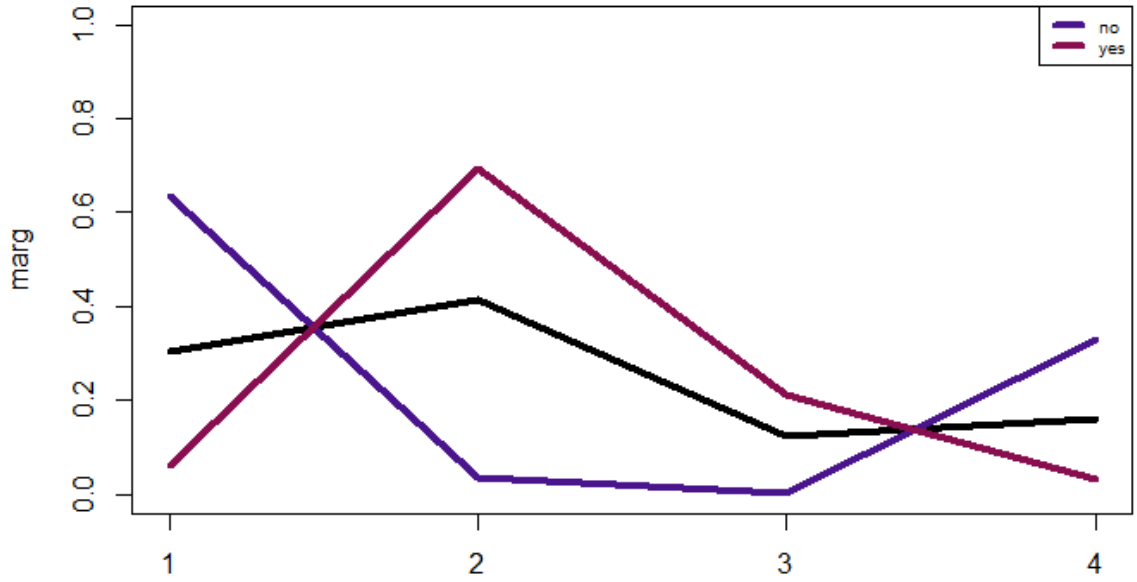


Fig. 5.8 Profiled of variable housing without being class-conditioned

In the first and fourth class almost no individuals have a housing loan and in the second and third class most of the people have it.

Variable *poutcome*

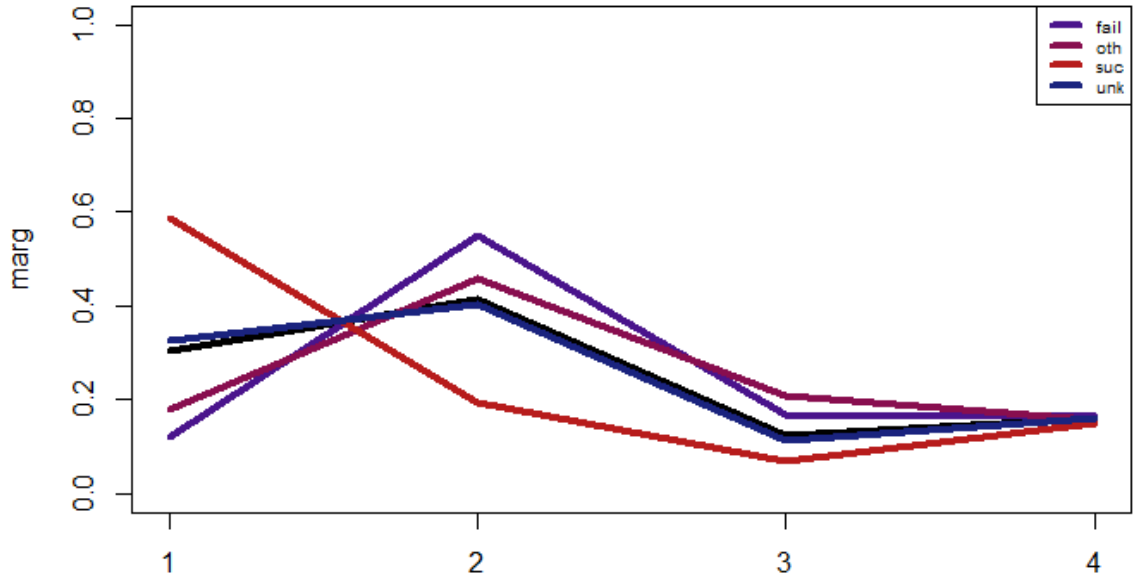


Fig. 5.9 Profiled of variable poutcome without being class-conditioned

Most of the people of the first class successfully subscribed to a bank term deposit in the previous campaign. For the second class the previous campaign resulted in a failure or other result that we do not know.

Relevant bivariate

Variables *balance & age*

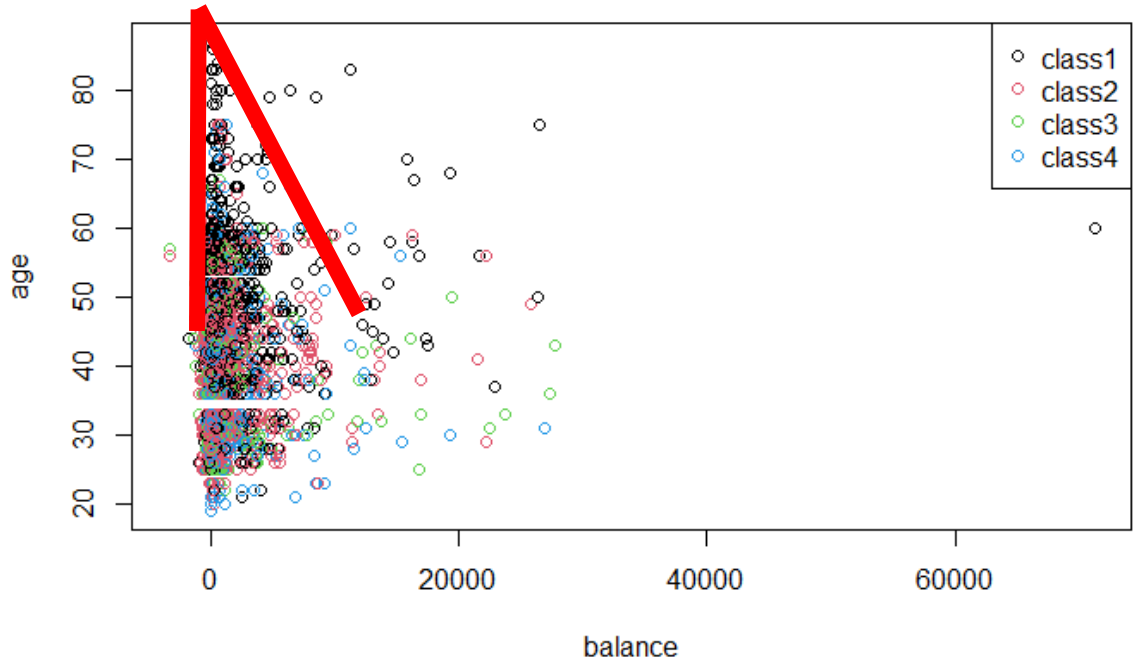


Fig. 5.10 Clustering Plot comparing balance with age

As we described above, the first cluster is an older set and when we compare with the variable balance, we can say that a big part of them tend to have a moderate balance.

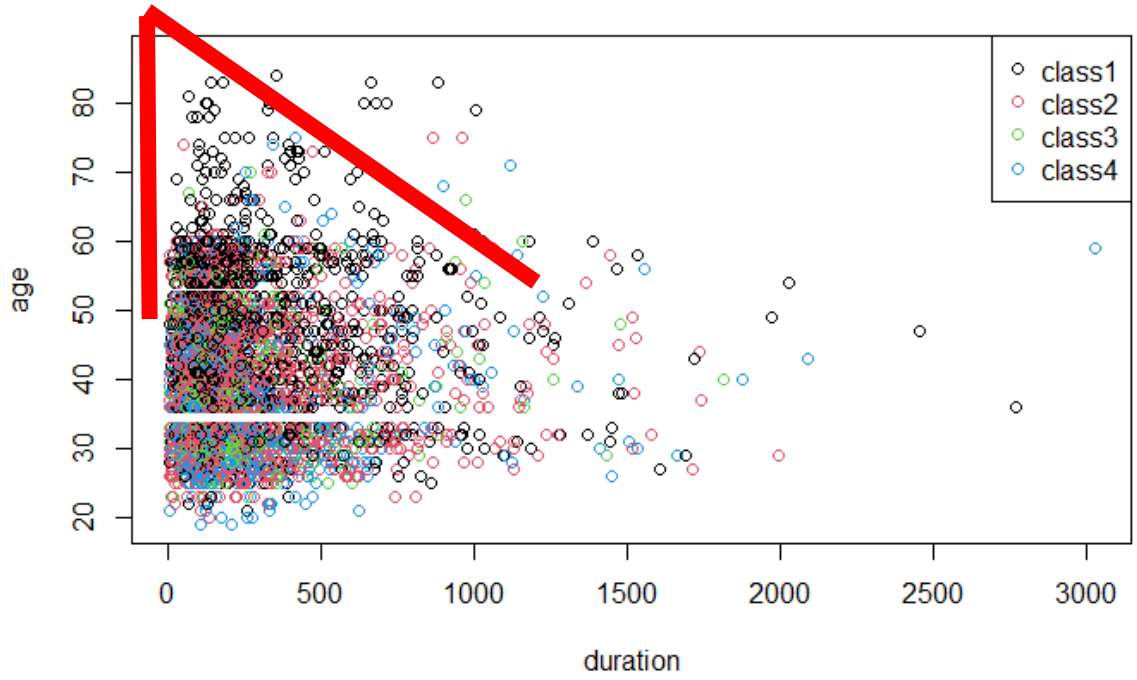


Fig. 5.11 Clustering Plot comparing duration with age

If we take a look at the top-left part of the plot we can notice that there is an accumulation of older people that they older they are the less contact lasts

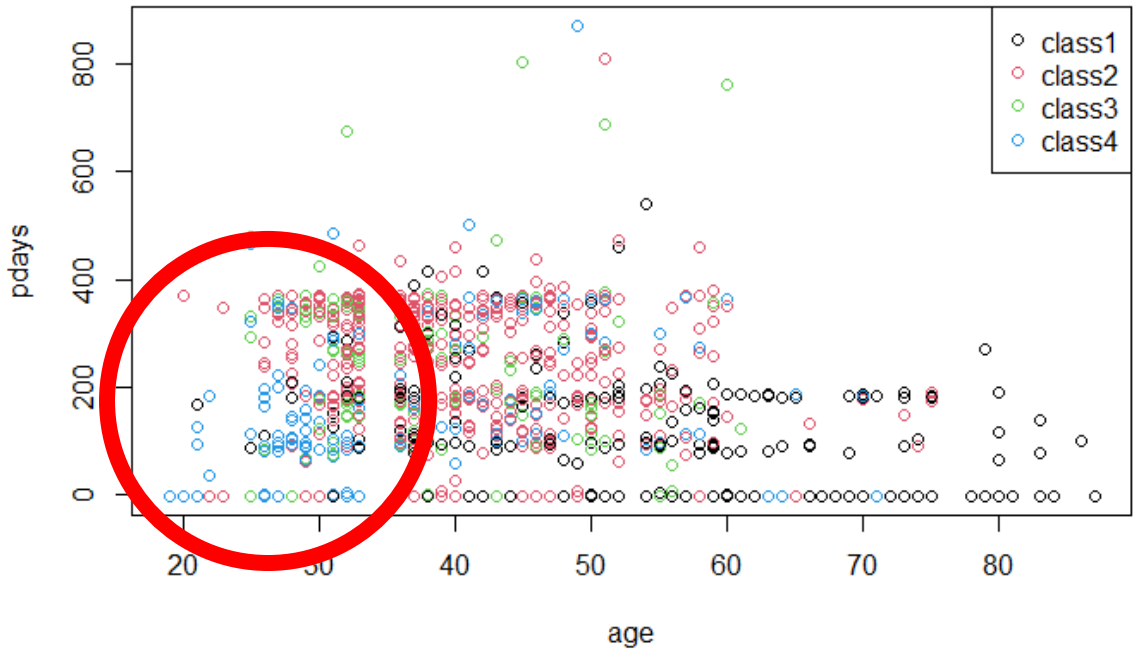


Fig. 5.12 Clustering Plot comparing age with pdays

At the left-bottom part of this plot we can see that there is a concentration of individuals of class 4, that are younger people and also mostly the number of days that passed by after they were contacted from the previous campaign is lower.

c) Synthesize the result of the classes' interpretation process into a set of templates characterizing the clusters, one template per cluster

Table 5.1

Class 1:

- Range of age:	Older
- Occupation:	Retired, housemaid, technician, or blue-collar worker
- Marital status:	Married
- Education:	(Cannot extract any conclusion)
- Has a housing loan?	(Cannot extract any conclusion)
- Outcome of previous campaign:	Successful

Class 2:

- Range of age:	Middle age
- Occupation:	Service, administration or blue-collar worker
- Marital status:	Married

- Education: Secondary education
- Has a housing loan? Yes
- Outcome of previous campaign: Failure/other/unknown

Class 3:

- Range of age: Middle age
- Occupation: Manager or technician
- Marital status: (Cannot extract any conclusion)
- Education: Tertiary education
- Has a housing loan? Yes
- Outcome of previous campaign: (Cannot extract any conclusion)

Class 4:

- Range of age: Younger
- Occupation: Manager or technician and have almost all the students
- Marital status: Single or divorced
- Education: Secondary or tertiary
- Has a housing loan? No
- Outcome of previous campaign: (Cannot extract any conclusion)
- Extra: The number of days that passed by after the previous campaign was lower than in other classes

Global discussion and general conclusions of the whole work

Analyse coincidences and divergences between PCA, AMC, Clustering

As can be seen in this report and throughout the working process, a lot of different methods have been applied to the dataset with different outputs as a result. In this section the results of these different methods will be discussed and compared.

The biggest takeaway from the PCA section, where PCA was applied to the bank term deposit data, is Figure 3.6 which is the projection that retained the most inertia and where it is interesting to see how the blue “unknown” values were grouped together, separate from the others. Other than this clear and separated group of blue unknowns, the other values seem to be overlapping in the center. In a similar manner, if we look at Figure 3.5, which shows how the categorical variables are distributed over the principal component axes, most of the values seem to be clustered together about the origin. The categorical variables that do however seem to be displaced from the center are, for the “job” variable: “student”, “retired” and “housemaid”, for the variable “default”: “yes”, and for the variable “y”: “yes”.

After using clustering on the dataset there appeared to be four clusters. When exploring and profiling these with different statistical tests we were able to some degree distinguish the characteristics within the groups and summarise these in Table 5.1. The clustering allows for distinguishing and finding commonalities between points that can be hard to make out by the human eye, as was seen in the discussion about PCA.

The fact that it is possible to cluster the individuals into groups with certain characteristics, through first reducing the number of dimensions with PCA and then clustering with Gower’s dissimilarity coefficient, is the key finding in this report. Finding patterns and being able to categorise data is in many ways fundamental for all sorts of further explorations. With the knowledge that these four groups exist, it helps anyone interested in profiling a potential bank customer to do so. Other than being advantageous knowledge for the banks, studying why these groups exist can also be interesting to discuss in a wider societal context.

The next steps that would be interesting to explore in a further report are many. A complementary study that would be interesting to perform is to seek answers from actual customers why they would or would not set up a bank term deposit. That would allow for studying the reasoning behind certain decisions made and could help with finding hidden factors that influence how individuals are grouped together. It would also be interesting to be able to predict, given a new customer, if they would sign up for a bank term deposit. This would be very useful for the banks as they would be able to target these customers in their efforts to expand their customer base.

To conclude this report, studying the bank term deposit dataset is very much an exploration of complex socioeconomic structures in society. The data mining process has given structure and the different tools for managing the tasks, such as the Gantt diagram, have facilitated the group work. Applying methods and techniques from lectures, such as preprocessing, PCA and clustering, helped in reaching the conclusion that there are four groups distinguishable in the data. This knowledge is useful for making sense of the data

and is a crucial stepping board into future explorations and predictions. That is why the work done here is important and a good base for many other related research projects.

Working plan

Initial Gantt diagram

Task	22 Sep	23 Sep	29 Sep	30 Sep	2 Oct	3 Oct	7 Oct	14 Oct	21 Oct	22 Oct	27 Oct	28 Oct
Definition and projects assignment												
SearchDDBB	1											
Basic structure of data matrix	1											
Delivery document		1										
Project kick-off												
email with group members		1										
Project development												
Planning			2									
Risk plan			4									
Missing processing				1								
Metadata				1								
Basic Descriptive statistics				3								
Additional descriptive statistics				3								
Delivery second document						1						
Final delivery												
PCA analysis for numerical variables						16						
Hierarchical Clustering on original data						16						

Profiling of clusters						16						
Conclusion										1		
Prepare the report										1		
Revision of the report											1	
Preparing presentation resources											1	
Revision of the ppt											1	
Final presentation												1

Final Gantt diagram

Task	22 Sep	23 Sep	29 Sep	30 Sep	2 Oct	3 Oct	7 Oct	14 Oct	21 Oct	22 Oct	25 Oct	26 Oct	27 Oct	28 Oct
Definition and projects assignment														
SearchDDBB	1													
Basic structure of data matrix	1													
Delivery document		1												
Project kick-off														
email with group members		1												
Project development														
Planning			2											
Risk plan			4											
Missing processing				1										
Metadata				1										
Basic Descriptive statistics				3										
Additional descriptive statistics				3										
Delivery second document						1								
Final delivery														
PCA analysis for numerical variables						16								
Hierarchical Clustering on original data						16								
Profiling of clusters						16								

Conclusion											2			
Prepare the report										1				
Revision of the report												2		
Preparing presentation resources												2		
Revision of the ppt													1	
Final presentation														1

Our working plan has been strictly followed, so for this reason does not change at all.

Final tasks assignment grid

Task	DCM	ACC	FGD	JLR	AM	XSTB
Definition and projects assignment						
SearchDDBB	**x**	x	x	x	x	x
Basic structure of data matrix		x			**x**	
Delivery document	**x**	x	x	x	x	x
Project kick-off						
email with group members	**x**					
Project development						
Planning		**x**	x	x	x	
Risk plan		**x**	x			
Missing processing	x					**x**
Metadata		x		**x**		
Basic Descriptive statistics		**x**				
Additional descriptive statistics	**x**	x	x	x	x	x
Delivery second document	**x**	x	x	x	x	x
Final delivery						
PCA analysis for numerical variables	**x**					x
Hierarchical Clustering on original data		**x**	x			
Profiling of clusters				x	**x**	
Conclusion	x	x	x	**x**	x	**x**
Prepare the report			**x**			
Revision of the report	x	x	**x**	x	x	x
Preparing presentation resources	x	x	**x**	x	x	x
Revision of the ppt			**x**			
Final presentation	x	x	**x**	x	x	x
Participant - x						
Coordinator - **x**						

Discussion about deviances of final scheduling

We did some changes about the tasks assignments, but nothing to mention.