

MVA PROJECT: BANKING

30/10/2023

Project by: Andreja Andrejic, David Candela, Nayara Costa, Julie Oppedal,
Alexandros Tremopoulos

Index

1.	INTRODUCTION	2
	DATA.....	2
	DATA SOURCE PRESENTATION.....	2
	DATA STRUCTURE AND METADATA	2
	SCOPE OF STUDY	4
2.	PREPROCESSING	4
	MISSING VALUES	4
	Imputation of missing values	8
	OUTLIERS	13
3.	DESCRIPTIVE ANALYSIS.....	19
	Numerical Columns	19
	Categorical Columns.....	23
4.	PCA ANALYSIS.....	25
5.	MCA ANALYSIS.....	39

1. INTRODUCTION

DATA

The Bank Marketing dataset compiles call information from a Portuguese banking institution with the objective of predicting whether a client will subscribe to a term deposit. This classification task relies on various categorical features, such as the client's education level, occupation, and marital status. Additionally, numerical features like the client's age and average account balance are considered in the analysis. We selected this dataset as it is a dataset that covers our requirements for the project, with the correct number of categorical and numerical values.

DATA SOURCE PRESENTATION

How to download: The files can be downloaded directly through this link or going through the [Bank Marketing](#) link and pressing the ‘Download’ button. This will download a zip folder called ‘bank+marketing’. When the contents are extracted, there is a zip folder called ‘bank’. Inside this folder, there is the file ‘bank.csv’ that we are using. ‘bank.csv’ contains 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).

Description: The Bank Marketing dataset compiles call information from a Portuguese banking institution with the objective of predicting whether a client will subscribe to a term deposit. This classification task relies on various categorical features, such as the client's education level, occupation, and marital status. Additionally, numerical features like the client's age and average account balance are considered in the analysis. For more information visit the Bank Marketing web page.

DATA STRUCTURE AND METADATA

- 4521 records in the sample (bank.csv)
- 17 total variables of which:
- 7 are numerical ('age', 'balance', 'day', 'duration', 'campaign', 'pdays', 'previous')
- 10 are categorical ('job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'poutcome', 'y')
- 4 of the categorical are binary ('default', 'housing', 'loan', 'y')
- 6 of the categorical are qualitative ('job', 'marital', 'education', 'month', 'contact', 'poutcome')
- 6.83% of cells with missing.

MVA Banking

Full name	Variable name	Description	Type	Units	Modalities	Missing markers	Missing percentage
Age	age	Client's age	Integer	years			0
Job	job	Client's job	Categorical		'admin.', 'unemployed', 'management', ... ^a 'married', 'divorced' ^b , 'single'	'unknown'	0.85
Marital status	marital	Client's marital status	Categorical				0
Educational level	education	Client's level of education	Categorical		'secondary', 'primary', 'tertiary', 'unknown'	'unknown'	4.13
Credit default	default	Does the client have credit in default?	Binary		'yes', 'no'		0
Account balance	balance	Client's average account balance, in euros	Integer	euros			0
Has housing loan	housing	Does the client have a housing loan?	Binary		'yes', 'no'		0
Has personal loan	loan	Does the client have any other kind of loan?	Binary		'yes', 'no'		0
Contact method	contact	Method used to communicate with the client	Categorical		'telephone', 'cellular', 'unknown'	'unknown'	29.7
	day	Day of the month of the last contact	Date	days			0
	month	Month of the last contact	Date	months	'jan', 'feb', 'mar', ..., 'nov', 'dec'		0
Last contact duration	duration	Duration of the last contact with the client	Integer	seconds			0
Total times contacted	campaign	Number of times the client was contacted during the campaign	Integer				0
Days since last contact	pdays	Time since the client was last contacted from a previous campaign	Integer	days		'-1' ^c	81.9
Previous total contacts	previous	Number of times the client was contacted previous to the campaign	Integer				0
Previous result	poutcome	Outcome of the previous marketing campaign	Categorical		'other', 'failure', 'success', 'unknown'	'unknown'	81.9
Has subscribed	y	Outcome of this marketing campaign	Binary		'yes', 'no'		0

SCOPE OF STUDY

Following our initial data assessment, we made the decision not to exclude any rows or columns from the dataset. This strategic choice was made with the intention of maintaining the data's integrity and comprehensiveness. This approach allows us to analyse the dataset in its entirety, providing us with a broader perspective on the dataset.

2. PREPROCESSING

Ten out of seventeen variables contain characters, so they are converted to factors to be easily handled. Four of the variables are binary with ‘yes’ or ‘no’ as options. They are also converted to factors.

MISSING VALUES

There are five features with missing values. Four of them are categorical and one is numeric. Note that in the dataset missing values were replaced with ‘unknown’ or -1 in categorical and numeric variables respectively. So, to explore missing values replacement takes place again and ‘unknown’ and -1 values are converted to NAs.

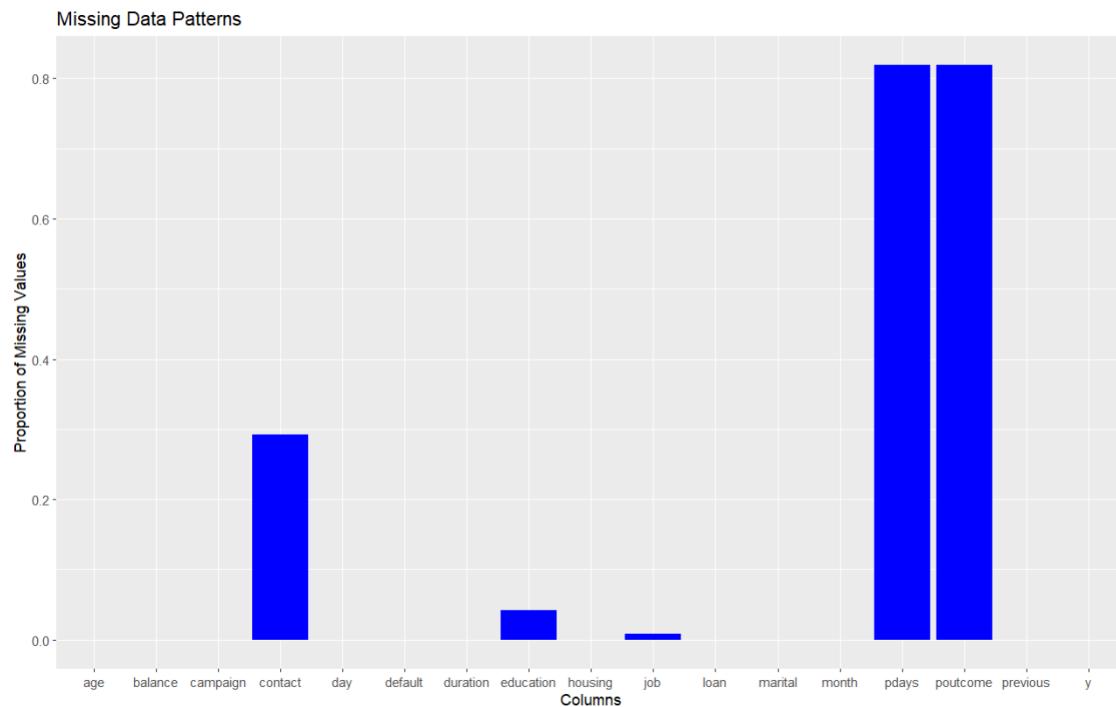
The percentage of the missing values in all cells is ~11.65%. Below in the figure the amount and the percentages of variables which miss values are displayed. Here the numeric variable is ‘pdays’ and the other four are categorical.

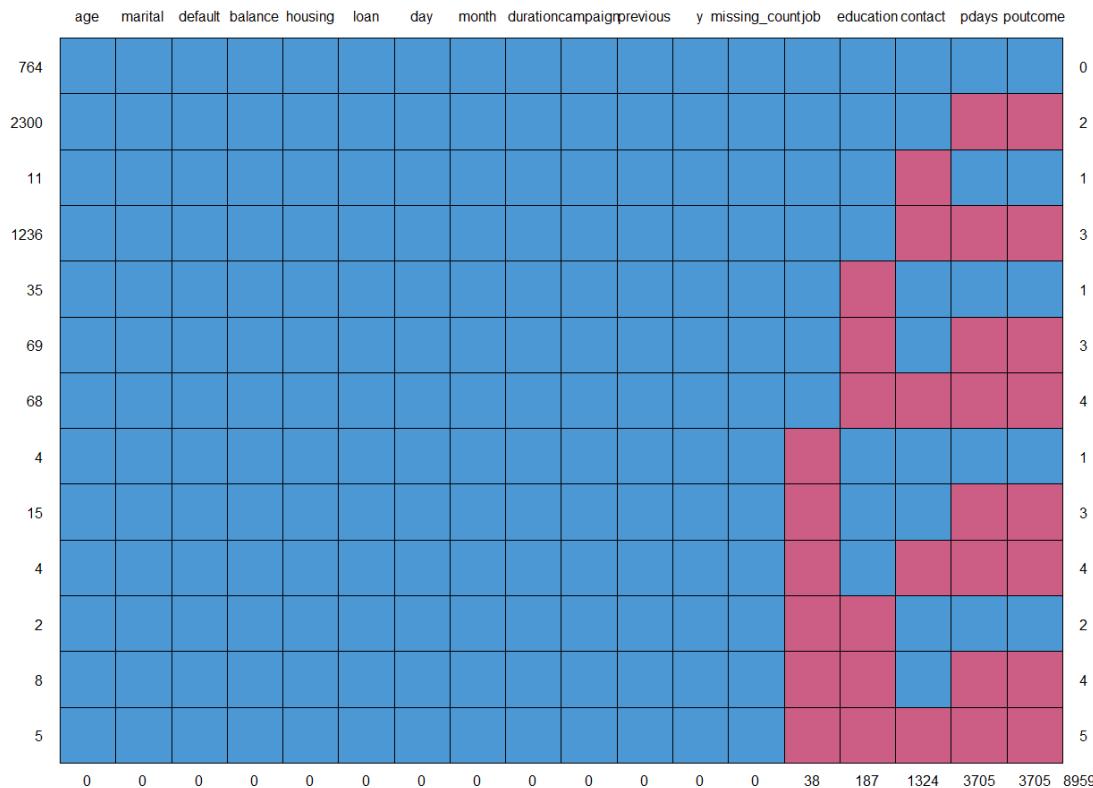
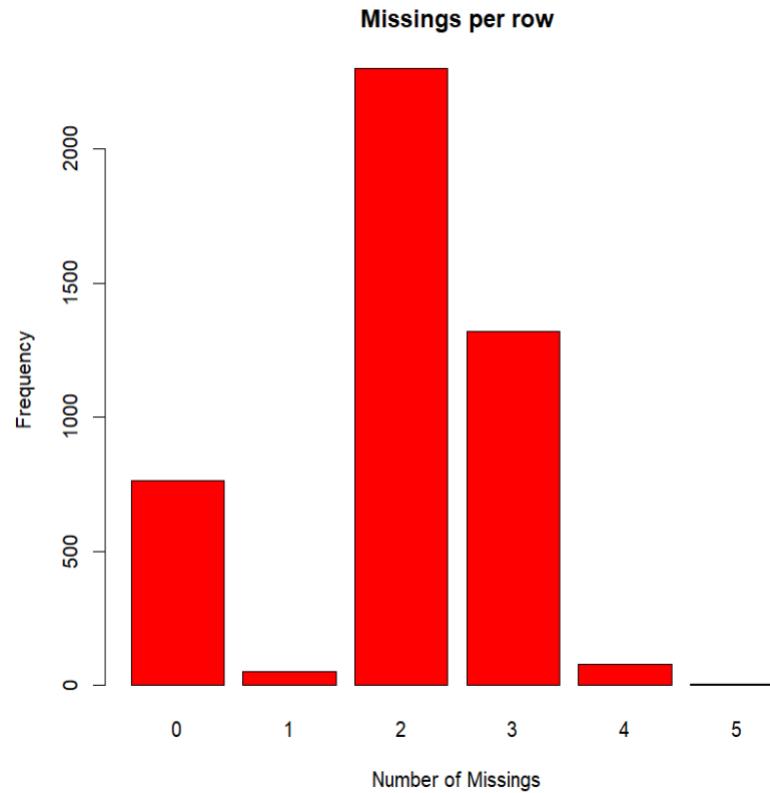
Variable <chr>	Missing_Count <dbl>	Percentage <dbl>
pdays	3705	81.950896
poutcome	3705	81.950896
contact	1324	29.285556
education	187	4.136253
job	38	0.840522

The next step is to check if the values are missing completely at random (MCAR) or not. Thus, little’s MCAR test is performed. However, this can be performed only in numeric variables and there is only one such which misses values. So, MCAR test is only performed to the numeric features (11 in total) and the result is a p-value equal to 0, which

indicates that missing values of ‘pdays’ variable are not completely at random and should be explored more in depth.

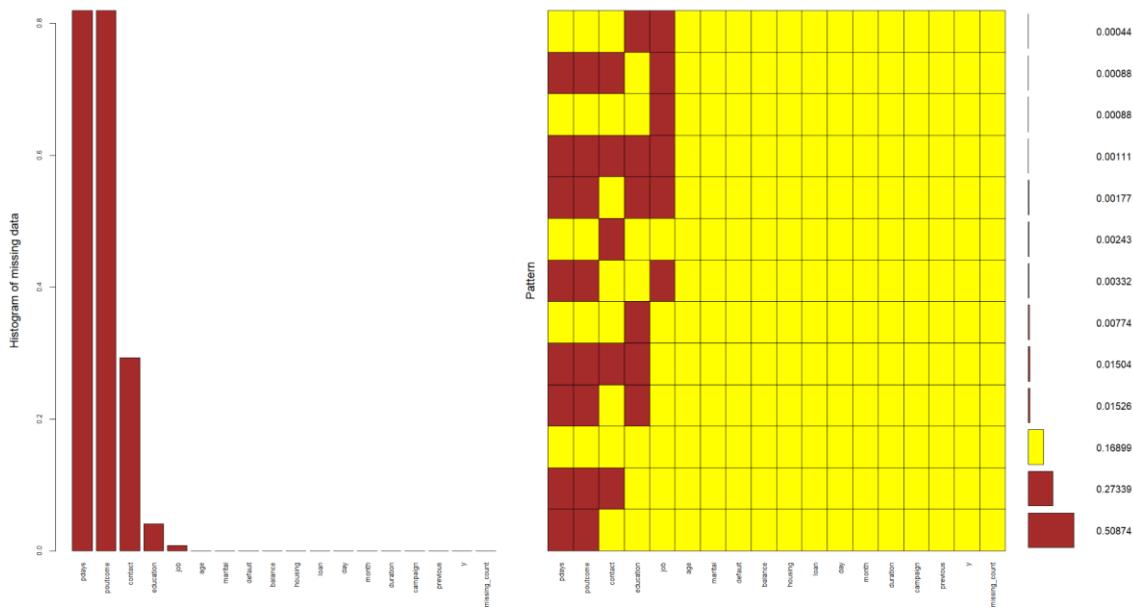
There are some plots that could help explore patterns in the missing values. Initially there is a plot with the frequency of missing values in each feature. Then in the next figure the frequency of missing values in rows can be observed. There are more than 2000 rows with at least two missing values.





In the figure just above there is the plot of the ‘md.pattern()’ function. There is a header (first row) that contains all the variable names, and the rows represent missing data. The left column indicates the number of times each pattern exists. The right column represents

the number of missing values in each pattern and the bottom row shows the number of missing values for each feature. The next plot shows again the percentages of missing values in each feature and the pattern table with the difference that instead of a number that indicates how many times a pattern exists; it shows the percentage of each pattern.

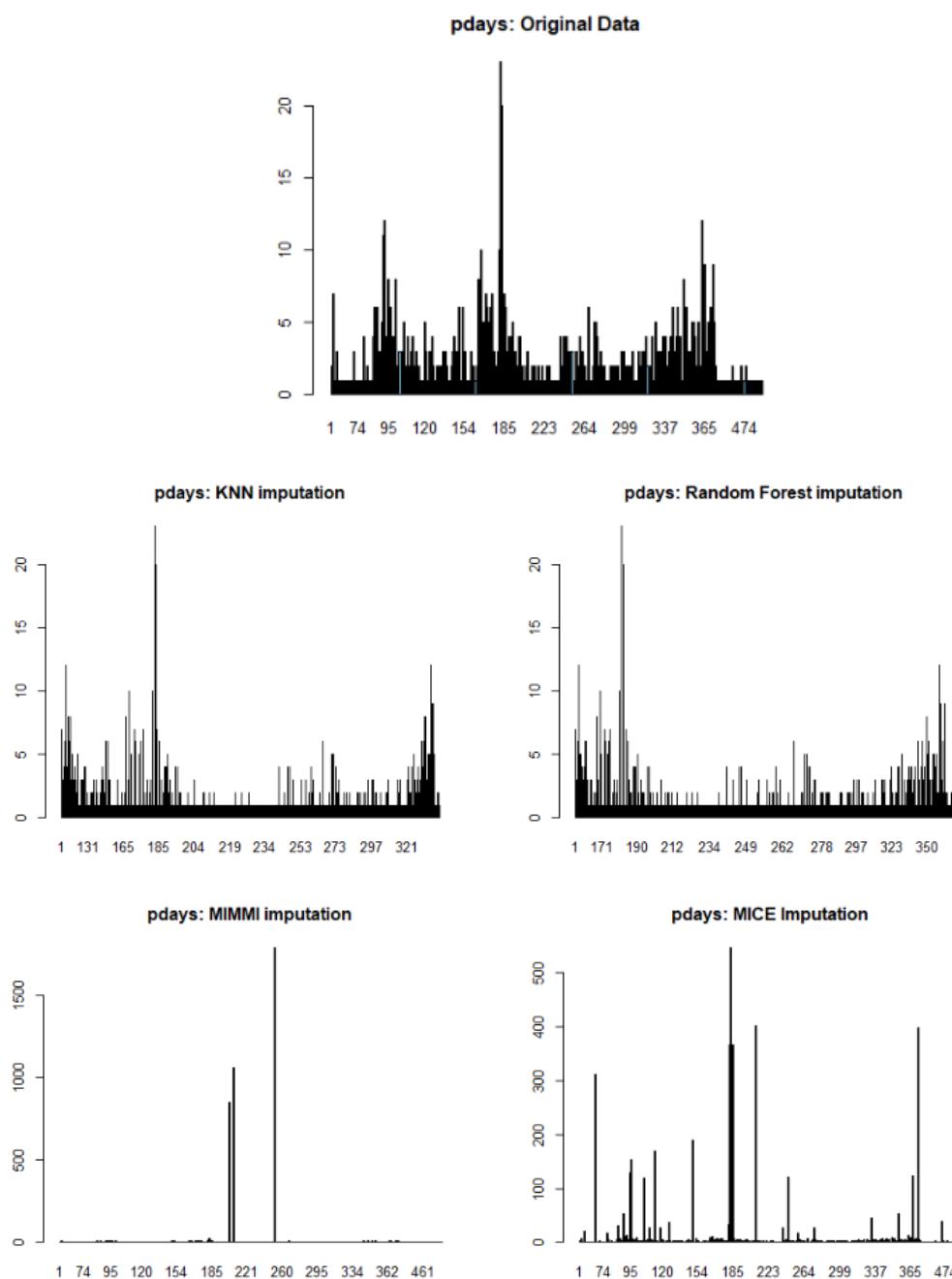


So, from these two plots we come to some conclusions. First of all ‘pdays’ and ‘poutcome’ go together i.e. they only miss values in the exact same rows. A reminder that ‘pdays’ are the time since the client was last contacted from a previous campaign while ‘poutcome’ is the outcome of the previous marketing campaign. From that, we conclude that these values are not missing out of an intention, so they are not MNAR (Missing Not at Random). It is also observed that ‘contact’ variable is missing 99,17% of the time (1313 out of 1324 missing values) along with ‘pdays’ and ‘poutcome’. Eleven times it is the only value missing. Contact is the communication method with the client, and it makes sense when days from the previous campaign, outcome of the campaign and the contact are missing altogether. We conclude that those three features have a pattern of missing values in the same rows (same clients). Thus, we consider ‘contact’ not being MNAR as well. The ‘education’ variable misses’ values ~80.2% of the time while ‘pdays’ and ‘poutcome’ miss as well and 60.07% when these two along with ‘contact’ miss values. In addition, 84,2% of the times ‘job’ values are NAs, ‘pdays’ and ‘poutcome’ are NAs as well. Although there may see patterns of ‘education’ and ‘job’ missing along with ‘pdays’ and ‘poutcome’, we will consider them as MNAR, because firstly ‘pdays’ and ‘poutcome’ are missing almost 82% of the time, which is a lot and may not be related, but most importantly due to chances that someone did not give his job and education status to the bank out of intention and personal reason. Consequently, we replace NAs in ‘job’ and ‘education’ again with ‘unknown’.

Imputation of missing values

Four different imputation methods were used. Those are KNN [k=5, k=67 (67 is the length of dataframe squared)], MIMMI(k=25, k=5) , Random Forest and MICE. KNN is only tested on ‘pdays’ , which is the only numerical feature that misses values and KNN is supposed to only handle numeric variables. The other 3 methods were used, for all 3 variables (‘pdays’, ‘poutcome’ and ‘contact’). After inputting them with all the methods, we have to decide which one performs/imputes better. Thus, there are some plots and some absolute numbers comparison, for before and after imputation.

Initially for ‘pdays’ the summary for each method is compared to the original data. We deduce that MICE is the better solution, because quartiles and mean are the closest to the original data among the 4 methods.:.

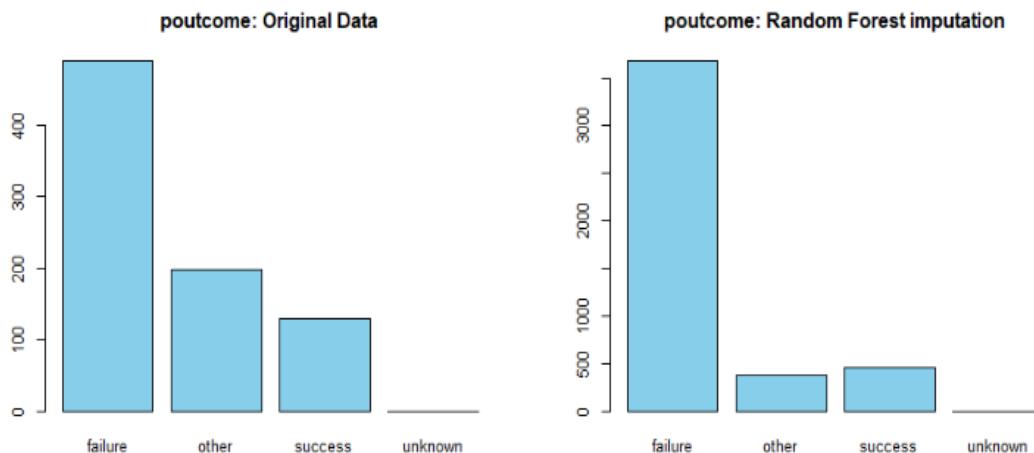


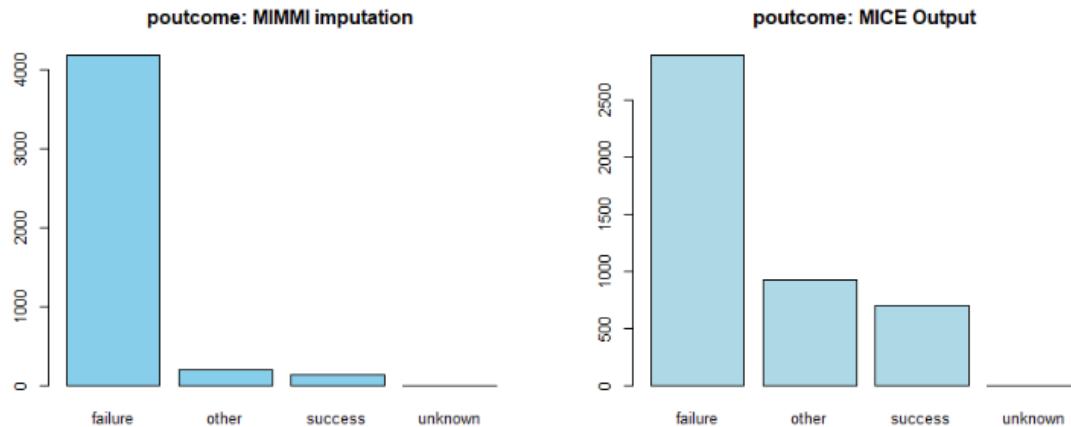
```

> print("pdays")
[1] "pdays"
> print("Original:")
[1] "Original:"
> summary(df$pdays)
   Min. 1st Qu. Median     Mean 3rd Qu.     Max.    NA's
      1.0    136.0   189.0   224.9   330.0   871.0    3705
> print("KNN:")
[1] "KNN:"
> summary(df_imp_knn$pdays)
   Min. 1st Qu. Median     Mean 3rd Qu.     Max.
      1.0    178.1   225.4   229.7   280.8   871.0
> print("Random Forest:")
[1] "Random Forest:"
> summary(df_imp_rf$pdays)
   Min. 1st Qu. Median     Mean 3rd Qu.     Max.
      1.0    199.0   252.9   251.5   299.0   871.0
> print("MIMMI:")
[1] "MIMMI:"
> summary(df_imp_mimmi$imputedData$pdays)
   Min. 1st Qu. Median     Mean 3rd Qu.     Max.
      1.0    198.2   202.9   226.2   253.9   871.0
> print("MICE:")
[1] "MICE:"
> summary(mice.output$pdays)
   Min. 1st Qu. Median     Mean 3rd Qu.     Max.
      1.0    114.0   183.0   198.6   247.0   871.0

```

Then for ‘poutcome’ we observe from plots that MICE barplots patterns are the nearest to the original data. That can be verified with the absolute numbers and the percentages of each distinct value of this feature. MICE imputed results have by far the closest percentages of values with the original data.





```
> print("poutcome")
[1] "poutcome"
> print("original:")
[1] "original:"
> table(df$poutcome)

failure other success unknown
490     197     129      0
> prop.table(table(df$poutcome)) * 100

failure other success unknown
60.04902 24.14216 15.80882 0.00000
> print("Random Forest:")
[1] "Random Forest:"
> table(df_imp_rf$poutcome)

failure other success unknown
3687     381     453      0
> prop.table(table(df_imp_rf$poutcome)) * 100

failure other success unknown
81.552754 8.427339 10.019907 0.000000
> print("MIMMI:")
[1] "MIMMI:"
> table(df_imp_mimmi$imputedData$poutcome)

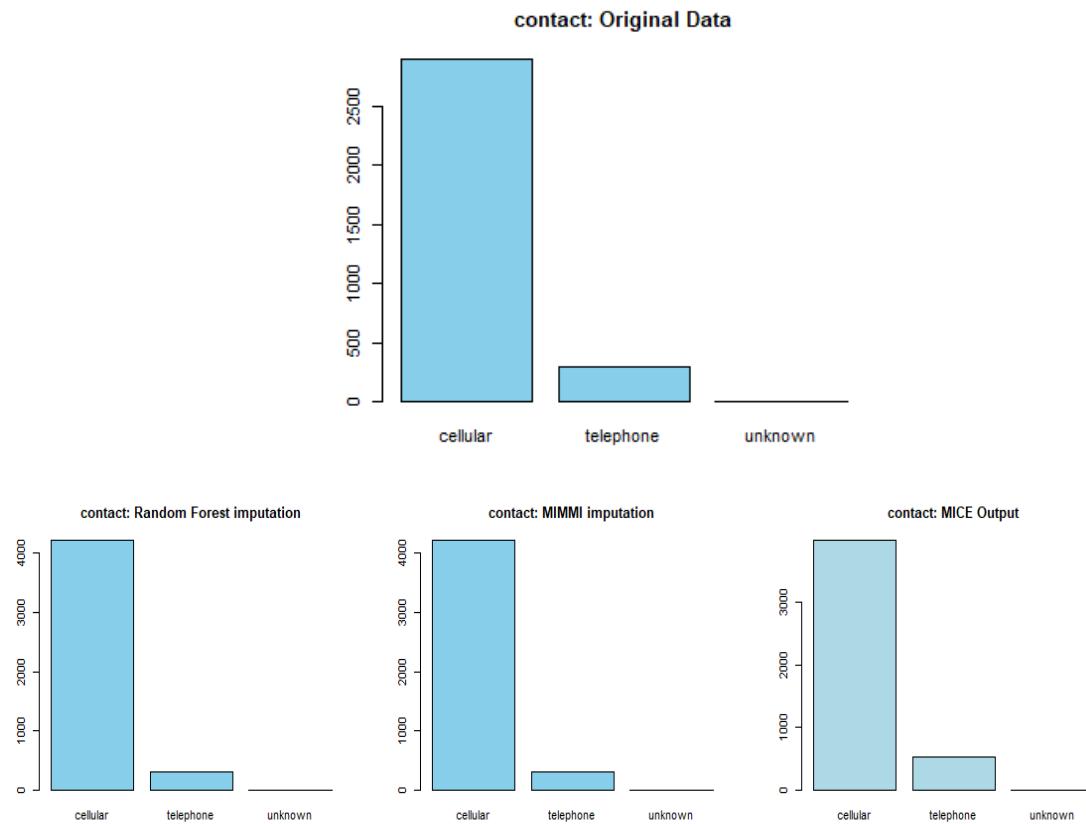
failure other success unknown
4195     197     129      0
> prop.table(table(df_imp_mimmi$imputedData$poutcome)) * 100

failure other success unknown
92.789206 4.357443 2.853351 0.000000
> print("MICE:")
[1] "MICE:"
> table(mice.output$poutcome)

failure other success unknown
2896     924     701      0
> prop.table(table(mice.output$poutcome)) * 100

failure other success unknown
64.05662 20.43796 15.50542 0.00000
>
```

Lastly, for ‘contact’ variable the case is similar with ‘poutcome’ in the barplots, but as well in the absolute numbers and the percentages. Here the percentages of MICE output are not by far closer to the original data, but they are clearly closer.



```
> print("contact")
[1] "Contact"
> print("original:")
[1] "original:"
> table(df$contact)

  cellular telephone  unknown
    2896           301       0
> prop.table(table(df$contact)) * 100

  cellular telephone  unknown
90.584923  9.415077  0.000000
> print("Random Forest:")
[1] "Random Forest:"
> table(df_imp_rf$contact)

  cellular telephone  unknown
    4218           303       0
> prop.table(table(df_imp_rf$contact)) * 100

  cellular telephone  unknown
93.297943  6.702057  0.000000
> print("MIMMI:")
[1] "MIMMI:"
> table(df_imp_mimmi$imputedData$contact)

  cellular telephone  unknown
    4220           301       0
> prop.table(table(df_imp_mimmi$imputedData$contact)) * 100

  cellular telephone  unknown
93.342181  6.657819  0.000000
> print("MICE:")
[1] "MICE:"
> table(mice.output$contact)

  cellular telephone  unknown
    3996           525       0
> prop.table(table(mice.output$contact)) * 100

  cellular telephone  unknown
88.38752  11.61248  0.000000
```

To conclude the imputation of missing values, it seems that MICE performs better than the other three methods, for each of the three variables we wanted to impute. Thus, these are the results we are going to keep in our analysis.

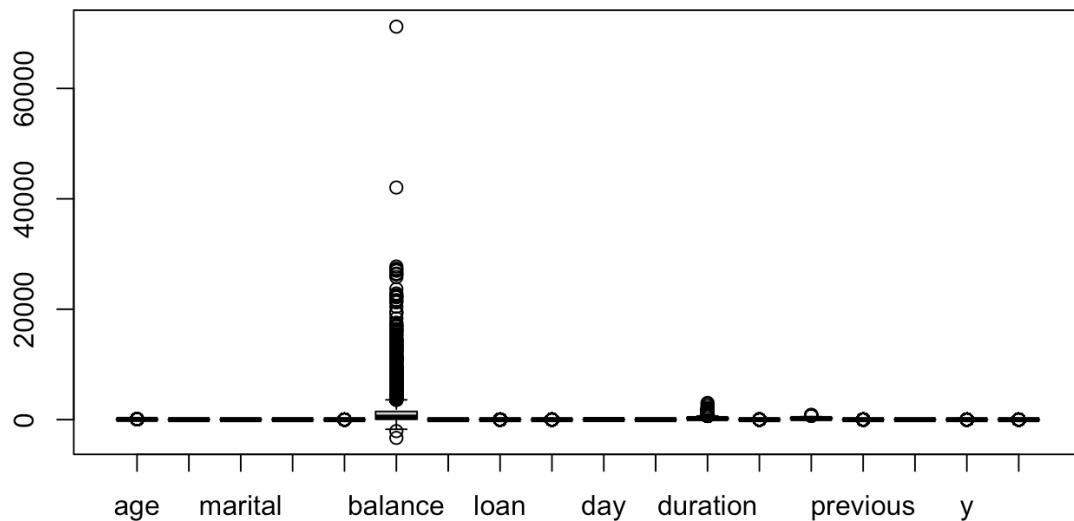
During working the deliverable and after this imputation took place, we observed something very important. There is one variable called ‘previous’ that indicates the number of times the client was contacted before the campaign. We tested this variable and saw that when it has value 0, ‘pdays’ has value -1 and ‘poutcome’ has value ‘unkown’. That makes perfect sense because if the bank has never contacted the client before, there are no passed days since last spoke with the client and not a previous outcome of the campaign. Thus, the dataset has -1 and ‘unknown’ respectively to indicate this. So, ‘pdays’ and ‘poutcome’ should not be imputed. A reminder that ‘job’ and ‘education’ were not imputed as well because they might be MNAR. That leaves us with only the ‘contact’ variable to be imputed. It is going to be imputed with mice and the

results are shown below in the figure. In addition, the ‘pdays’ column is going to be dropped, since it is almost 82% with -1 value, and we cannot use it. Even though our previous imputations will not be used eventually they remain in our report, to show the flow and the process of our dataset until we reach this conclusion.

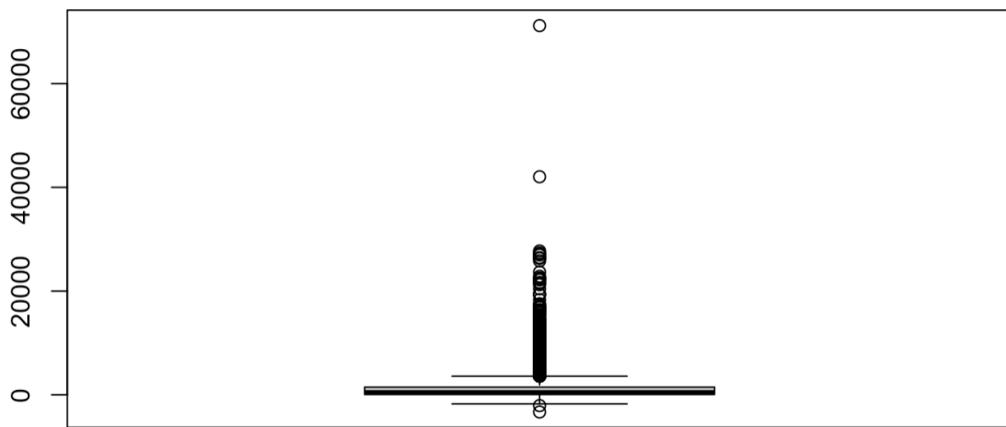
```
[1] "Original:"  
  
  cellular telephone unknown  
      2896          301          0  
  
  cellular telephone unknown  
90.584923  9.415077  0.000000  
[1] "MICE:"  
  
  cellular telephone unknown  
      4100          421          0  
  
  cellular telephone unknown  
90.687901  9.312099  0.000000
```

OUTLIERS

The provided data and charts help us see outliers in the bank's records. Most of these outliers are in the 'balance' section. Other areas like 'age', 'day', 'duration', 'previous', and 'y' have a few of these unusual values too, but not as many as 'balance'. When we look closer at 'balance', there are some people who have way more money than most. We also looked at a small part of the data that showed people with high balances, their age, job, and other details. From this, we can tell things like if they have missed any loan payments. In short, our main observation is that the 'balance' section has some very high values which can affect our analysis, so we should be careful with them.



Univariate Outliers Detection: balance

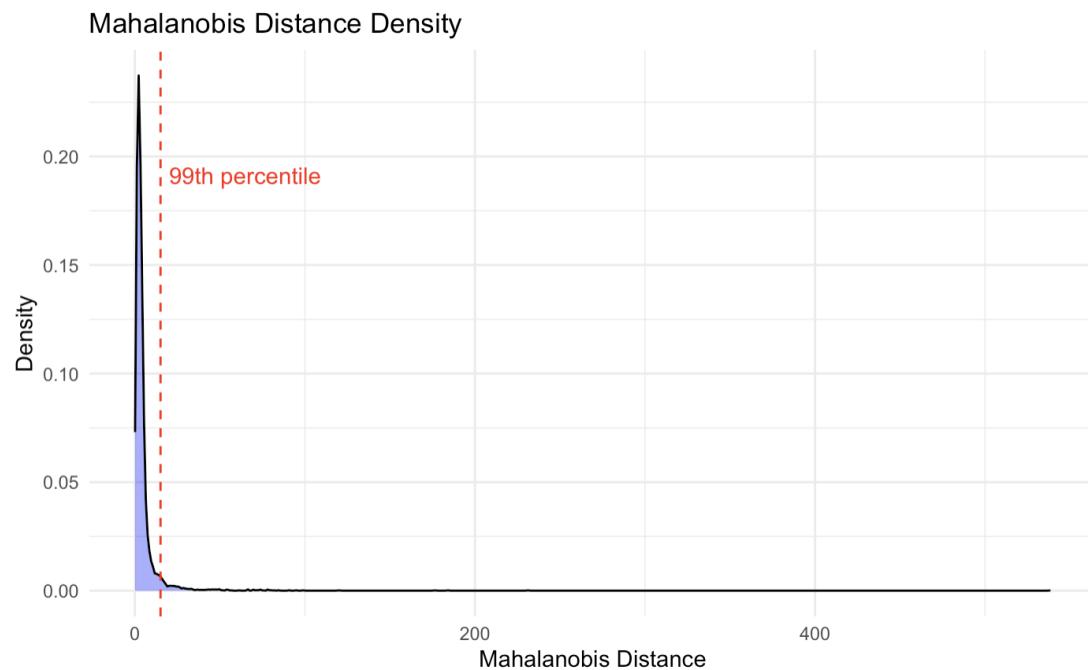


Since the goal is to predict a binary outcome based on the features, the relevance of outliers becomes crucial. Outliers might indicate rare but potentially important scenarios, like high-value clients or clients with unique attributes that sway their decision in favour or against subscribing.

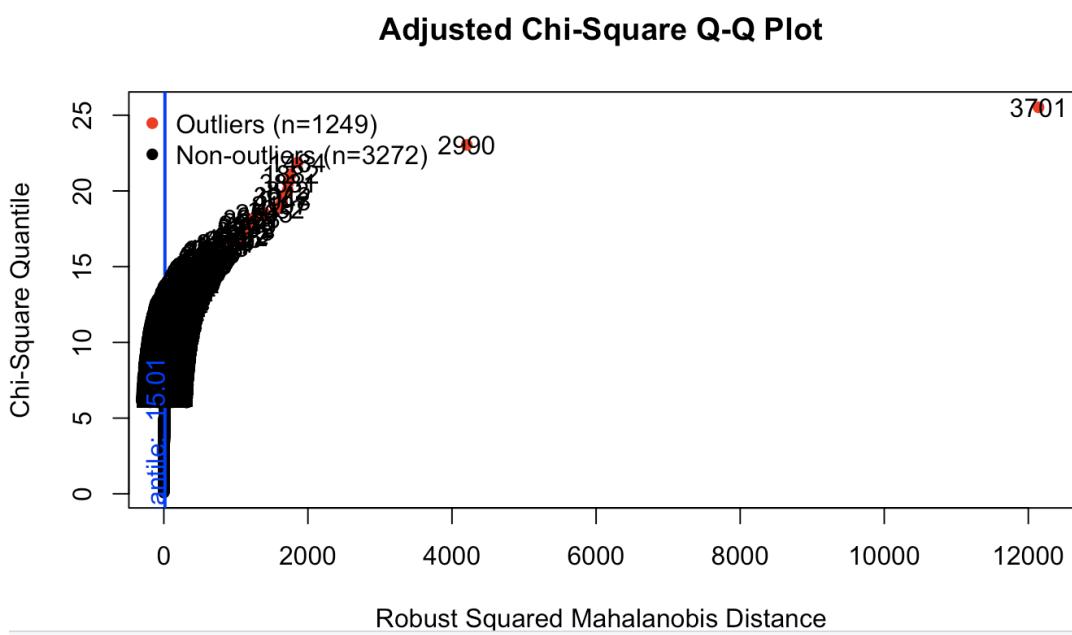
Given the nature of banking datasets, outliers, especially in financial metrics, can represent high-value or unique clients. Removing these outliers might result in losing insights about a crucial segment of the client base. As the dataset's objective is to predict term deposit subscription, understanding the behaviour of high-value clients might be vital. Furthermore, if we consider the wealth distribution in real-world scenarios, it's skewed with a smaller fraction of individuals holding a significant portion of the wealth.

However, we must use robust methods. While the decision is to keep the outliers, it's essential to ensure that the predictive model is not overly sensitive to these outliers. For a more in-depth insight, consider segmenting the data based on certain thresholds. This

way, we can analyse the behaviour of outliers, ensuring that unique patterns aren't weakened in broader analysis.



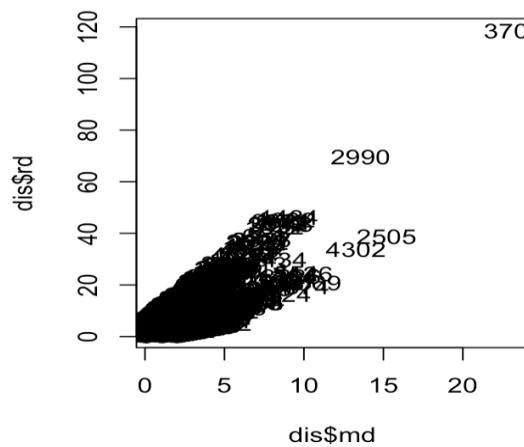
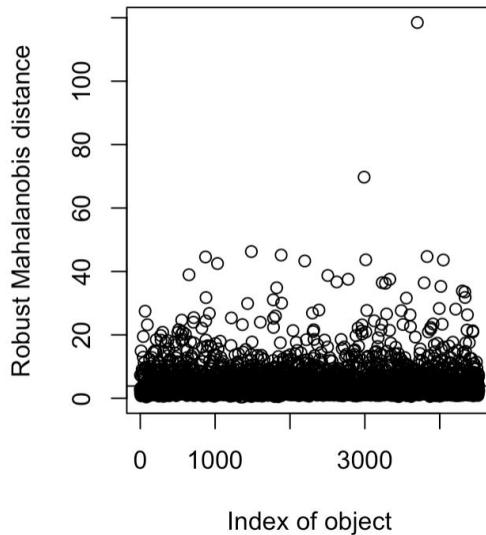
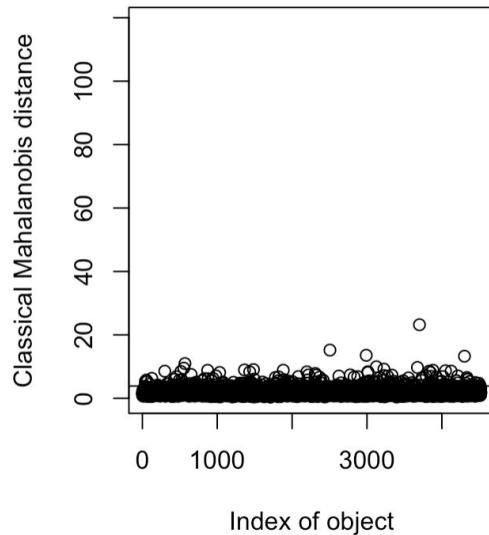
We visualized the density distribution of the Mahalanobis Distance, a measure used to determine the distance between a point and a distribution. From the plotted graph, it can be observed that the majority of data points cluster closely near the origin, with a sharp decline in density as the Mahalanobis Distance increases. Notably, the 99th percentile is indicated with a red dashed line, highlighting where only 1% of the observed data points exceed this distance. This analysis assists in understanding the spread and relationships of the data points in the given dataset.



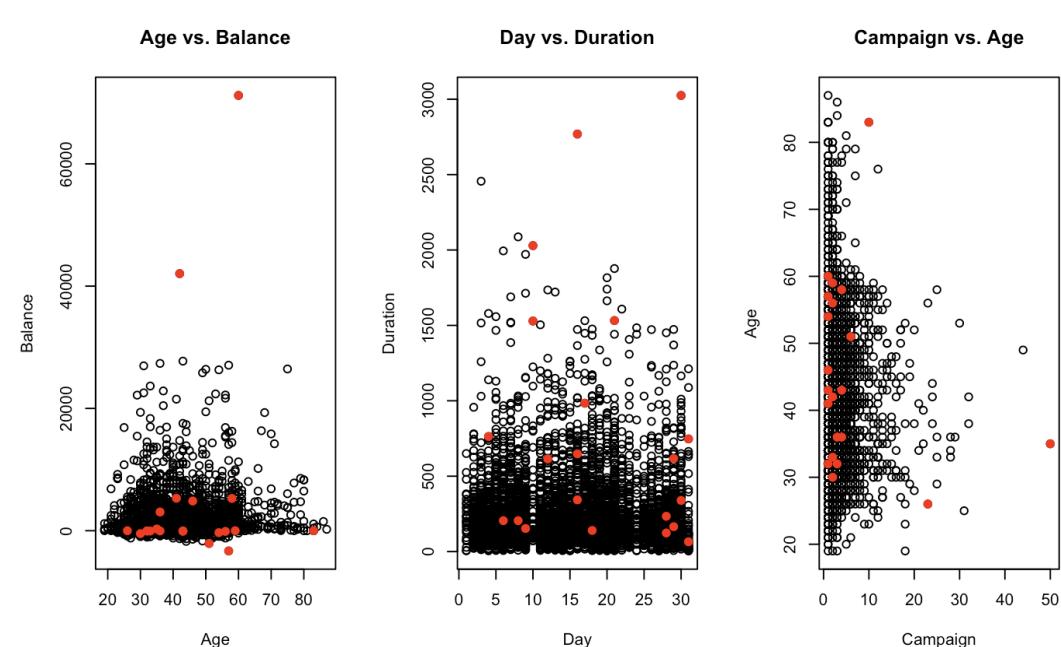
The presented plot showcases an "Adjusted Chi-Square Q-Q Plot," which is a graphical technique used to determine if the data follows a multivariate normal distribution. This plot places the quantiles of the observed distribution of the Robust Squared Mahalanobis Distance against the expected quantiles of the chi-square distribution. If the data is multivariate normal, most of the data points should lie close to the straight 45-degree reference line.

From the plot, we notice a clear deviation from the reference line, especially for higher quantiles. This deviation is indicative of potential outliers in the data. In this plot, outliers are highlighted in red and marked as "Outliers" with a count of 1,249, whereas the non-outliers are presented in black and have a count of 3,272.

In conclusion, the "Adjusted Chi-Square Q-Q Plot" coupled with the `mvn` function helps in identifying multivariate outliers in the dataset. In this analysis, 1,249 data points were flagged as potential outliers. This is significant as multivariate outliers can greatly impact statistical analyses and modelling and identifying them is a crucial step in data preprocessing.



In this analysis, the chemometrics package in R was employed to detect multivariate outliers in the dataset, specifically within the columns "age", "balance", "day", "duration", and "campaign". The primary metric utilized was the Mahalanobis distance, both in its classical and robust forms. This distance measures how far a particular data point is from the center of the distribution, accounting for the data's correlation structure. From the visualizations, it's evident that data points with high Mahalanobis distances, in both classical and robust interpretations, could be potential outliers. Particularly, in the scatter plot comparing these two distances, points distant from the primary cluster are highlighted as potential outliers. However, it's vital to emphasize that being labelled as an outlier doesn't warrant immediate removal; understanding the context and source of the data is paramount. This approach provides a systematic way to identify potential outliers in multivariate datasets, where univariate analyses might miss the intricacies of multivariate outlier patterns.



Top Outliers LOF Scores:
[1] 36.140392 16.299055 4.887982

Corresponding Rows for Top Outliers:

	age <int>	balance <int>	day <int>	duration <int>	campaign <int>
3701	60	71188	6	205	1
2990	42	42045	8	205	2
4518	57	-3313	9	153	1

3 rows

We employed the Local Outlier Factor (LOF) method to detect potential outliers within our dataset. The LOF scores quantify the degree of abnormality of each data point relative to its neighbours. From the analysis, we identified three top outliers with LOF scores of 36.140392, 16.299055, and 4.887982, respectively. A visualization of the data, with these

outliers highlighted, indicates that they deviate significantly from the main cluster of data points in variables such as "Age vs. Balance", "Day vs. Duration", and "Duration vs. Age". Further examination of the specific rows corresponding to these outliers in the dataset revealed the following:

- Row 3701: A 60-year-old individual with a balance of 71,188, contacted on the 6th day for a campaign duration of 205.
- Row 2990: A 42-year-old individual with a balance of 42,045, contacted on the 8th day for a campaign duration of 205.
- Row 4518: A 57-year-old individual with a balance of -3,313, contacted on the 9th day for a campaign duration of 153.

These data points can be considered as anomalies due to their distinct values in the above variables when compared to the general trend in the dataset. As a next step, it may be important to further investigate the reasons behind these deviations and decide whether to retain or exclude these outliers from subsequent analyses, depending on the specific objectives of the study.

In analysing the provided methods for outlier detection, it becomes evident that multiple approaches are adopted to cater to different data structures. Univariate outlier detection using boxplots serves as a straightforward approach, focusing on each variable individually. This is simple and effective for single-variable anomalies but doesn't account for the multidimensional nature of some datasets.

The Mahalanobis Distance offers a metric to identify multivariate outliers by considering the covariance between variables. Its graphical representation makes it easier to visualize and discern potential outliers. Similarly, the MVN package's multivariate outlier detection considers the multivariate nature of the data and is based on robust statistical measures. The Chemometrics package, too, offers multivariate outlier detection but might be better suited for specific datasets.

The Local Outlier Factor (LOF) algorithm, on the other hand, analyses the local density deviation of an observation concerning its neighbours. It can differentiate between clusters of varying densities, making it particularly effective for datasets with multiple clusters.

Given the results and the diverse nature of these methods, the best approach depends on the specific characteristics and requirements of the dataset in question. For purely univariate data, the boxplot method is effective and intuitive. For multivariate datasets where relationships between variables matter, the Mahalanobis Distance or MVN package could be ideal. However, if the dataset is expected to have clusters of varying densities, the LOF algorithm might be the most suitable. It is also worth considering a approach employing a combination of methods to ensure comprehensive outlier detection

3. DESCRIPTIVE ANALYSIS

In this chapter, we present a detailed descriptive analysis of a dataset using an R script. The dataset is subjected to various examinations, such as data dimension exploration, summary statistics, and a range of visualizations. This comprehensive analysis aims to provide insights into the trends within the dataset. For this analysis, we first segregated the data between the numerical and categorical variables.

```
> summary(df)
      age          job        marital       education      default      balance      housing
Min. :19.00 management:969 divorced:528 primary :678 no :4445 Min. :-3313 no :1962
1st Qu.:33.00 blue-collar:946 married :2797 secondary:2306 yes: 76 1st Qu.: 69 yes:2559
Median :39.00 technician:768 single  :1196 tertiary :1350
Mean  :41.17 admin.   :478 unknown  :187
3rd Qu.:49.00 services  :417
Max.  :87.00 retired  :230
(Other) :713

      loan         contact      day        month      duration      campaign      pdays
no :3830 cellular:2896 Min. : 1.00 may :1398 Min. : 4 Min. : 1.000 Min. : -1.00
yes: 691 telephone: 301 1st Qu.: 9.00 jul : 706 1st Qu.: 104 1st Qu.: 1.000 1st Qu.: -1.00
           unknown :1324 Median :16.00 aug : 633 Median : 185 Median : 2.000 Median : -1.00
           Mean   :15.92 jun : 531 Mean   : 264 Mean   : 2.794 Mean   : 39.77
           3rd Qu.:21.00 nov : 389 3rd Qu.: 329 3rd Qu.: 3.000 3rd Qu.: -1.00
           Max.  :31.00 apr : 293 Max.  :3025 Max.  :50.000 Max.  :871.00
(Other): 571

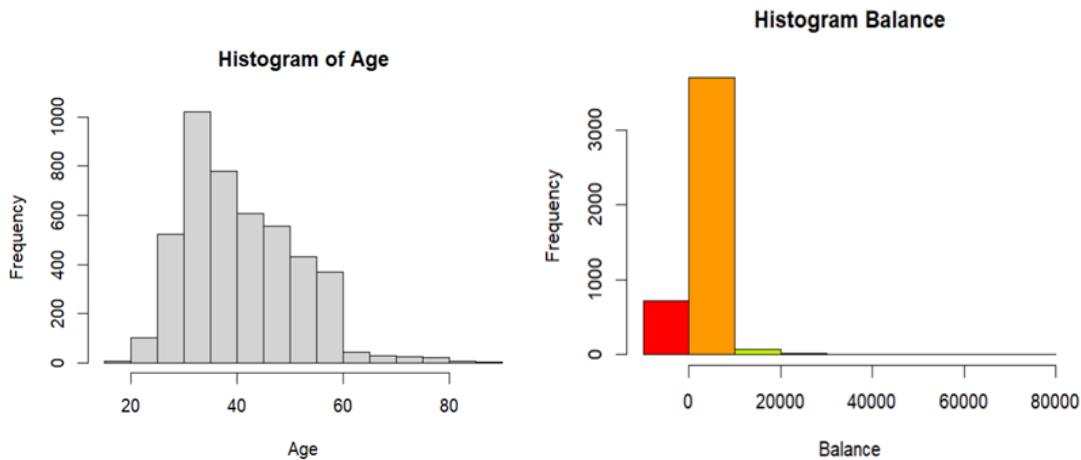
      previous      poutcome      y
Min. : 0.0000 failure: 490 no :4000
1st Qu.: 0.0000 other: 197 yes: 521
Median : 0.0000 success: 129
Mean  : 0.5426 unknown:3705
3rd Qu.: 0.0000
Max.  :25.0000
```

Numerical Columns

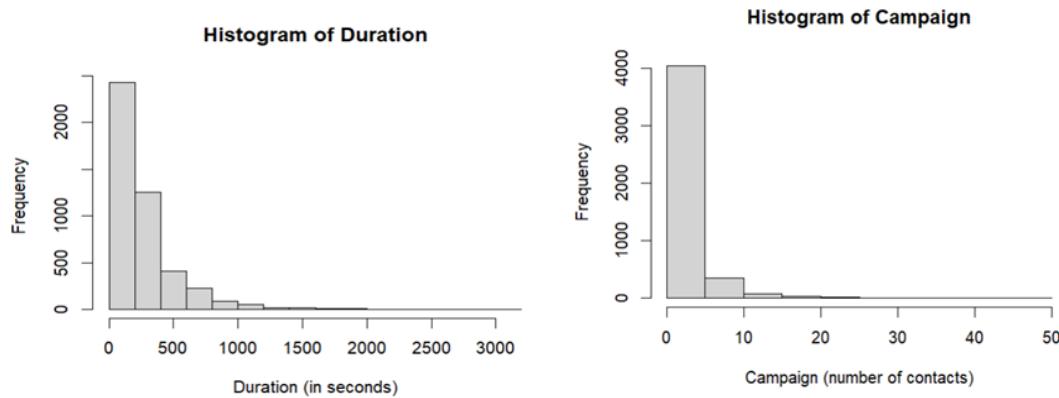
Focusing on the numerical variables, we first collected basic statistics to offer a more detailed view of the data. These statistics, including mean, median, and standard deviation, provide a clearer understanding of the data's central characteristics and variation.

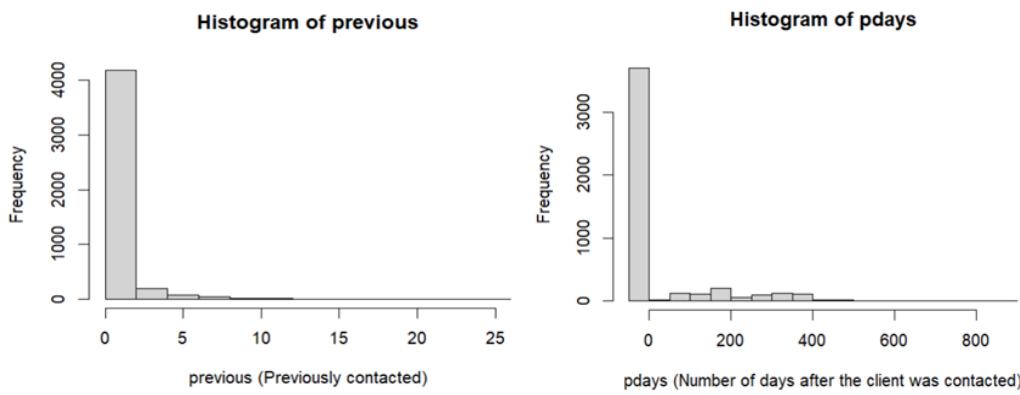
```
> basicStats(data.frame(age,balance,day,duration,campaign,pdays,previous))
      age        balance      day      duration      campaign      pdays      previous
nobs 4521.000000 4521.000000 4521.000000 4521.000000 4521.000000 4521.000000 4521.000000
NAs  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
Minimum 19.000000 -3313.000000 1.000000 4.000000 1.000000 -1.000000 0.000000 0.000000
Maximum 87.000000 71188.000000 31.000000 3025.000000 50.000000 871.000000 25.000000
1. Quartile 33.000000 69.000000 9.000000 104.000000 1.000000 -1.000000 0.000000 0.000000
3. Quartile 49.000000 1480.000000 21.000000 329.000000 3.000000 -1.000000 0.000000 0.000000
Mean 41.170095 1422.657819 15.915284 263.961292 2.793630 39.766645 0.542579
Median 39.000000 444.000000 16.000000 185.000000 2.000000 -1.000000 0.000000 0.000000
Sum 186130.000000 6431836.000000 71953.000000 1193369.000000 12630.000000 179785.000000 2453.000000
SE Mean 0.157294 44.760716 0.122663 3.864707 0.046250 1.489047 0.025187
LCL Mean 40.861721 1334.904929 15.674805 256.384577 2.702956 36.847384 0.493199
UCL Mean 41.478469 1510.410709 16.155764 271.538007 2.884303 42.685905 0.591959
Variance 111.856238 9057921.748594 68.024016 67525.469519 9.670897 10024.239560 2.868153
Stdev 10.576211 3009.638142 8.247667 259.856633 3.109807 100.121124 1.693562
Skewness 0.699037 6.592054 0.094564 2.770580 4.740767 2.715269 5.871361
Kurtosis 0.345583 88.250899 -1.040576 12.508007 37.108750 7.942161 51.912098
```

Creating histograms, we visually represent the distributions of the most relevant numerical variables:



For a better study of the target population, we focused on the age and balance variables in plots above. With this we can conclude that the target population stands between 30 and 45 years old and has an average yearly balance of 15000 \$.



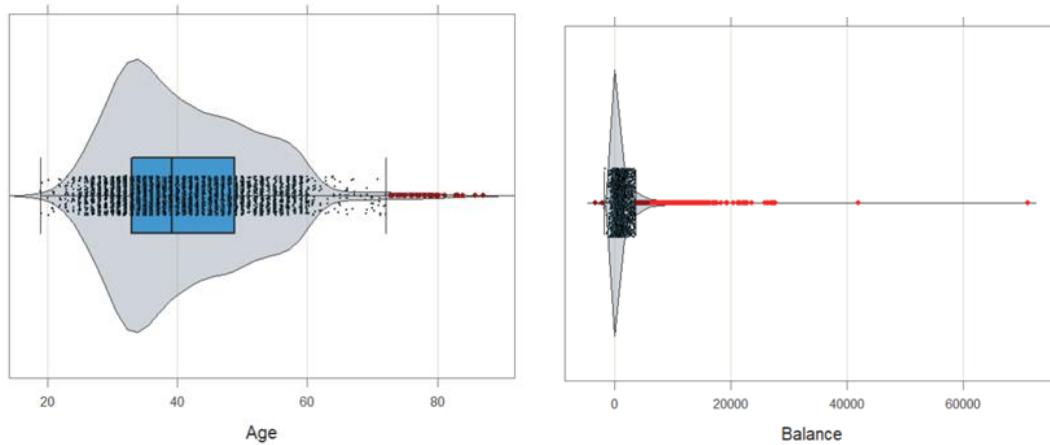


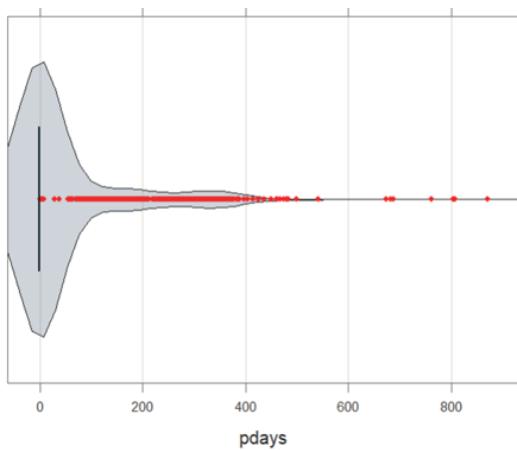
Whereas, with the analysis of the variables related to calls ('campaign,' 'duration,' 'previous'), we got the following conclusions:

- Call Duration:** Most calls were short in duration. Many conversations with customers were brief.
- Number of Contacts:** The data suggests that there were relatively few contact attempts with customers.
- Previous Contacts:** A significant portion of the target population had not been contacted previously.

The histograms of these datasets exhibit a striking concentration towards the left. This is primarily due to the fact that the datasets primarily capture call-related information, resulting in most variables displaying minimal variance. Characteristics such as call duration and the frequency of prior contacts tend to be consistently skewed towards lower values.

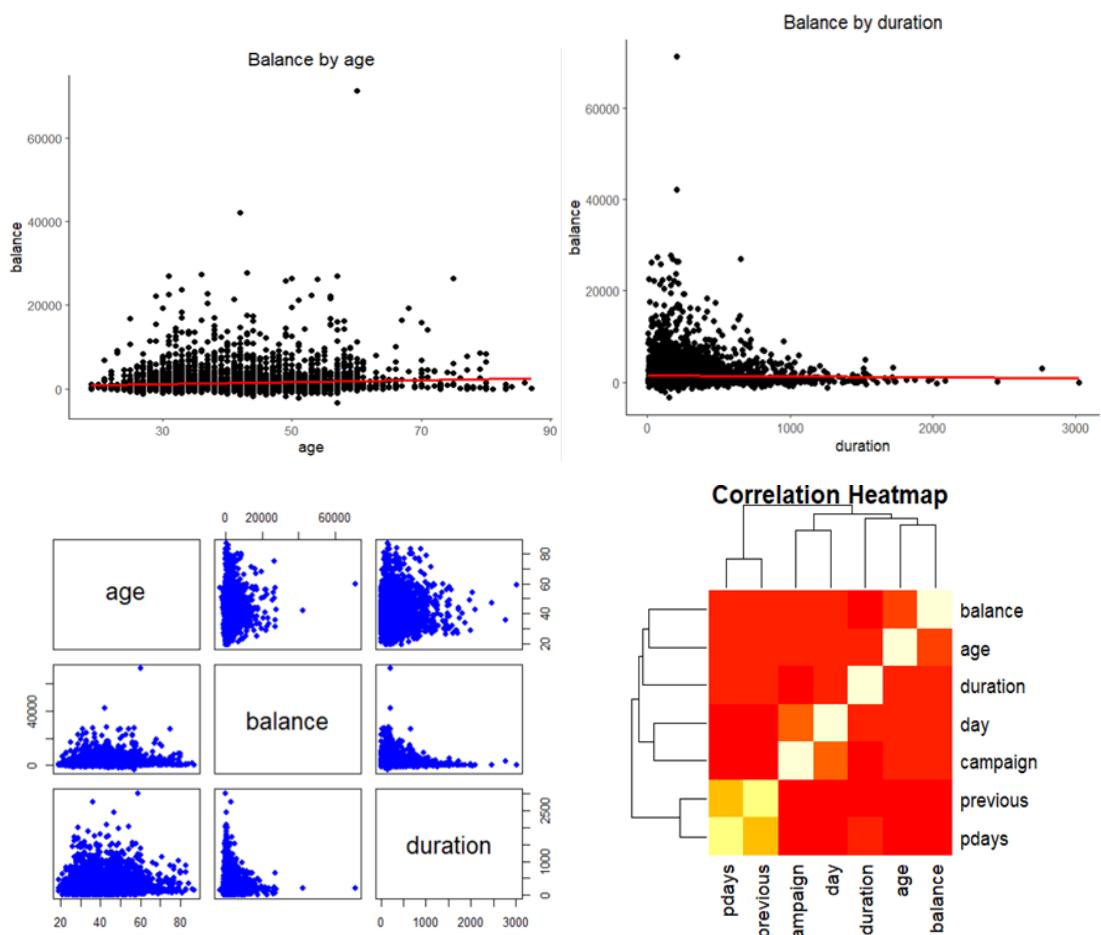
For identifying outliers in the dataset, we have used VBS plots, that integrates visualizations of the box plot, density distribution, individual data values and quick detection of outliers (dots in red):





With these plots, it is easily noticeable the number of outliers that the numerical variables have, and the density distribution of the data. Also, we identify that, since most of the target population was not previously contacted, the boxplot plot of pdays is compacted to -1, this, together with the previous pre-processing evaluation, helps us identify that this column is not good for analysis.

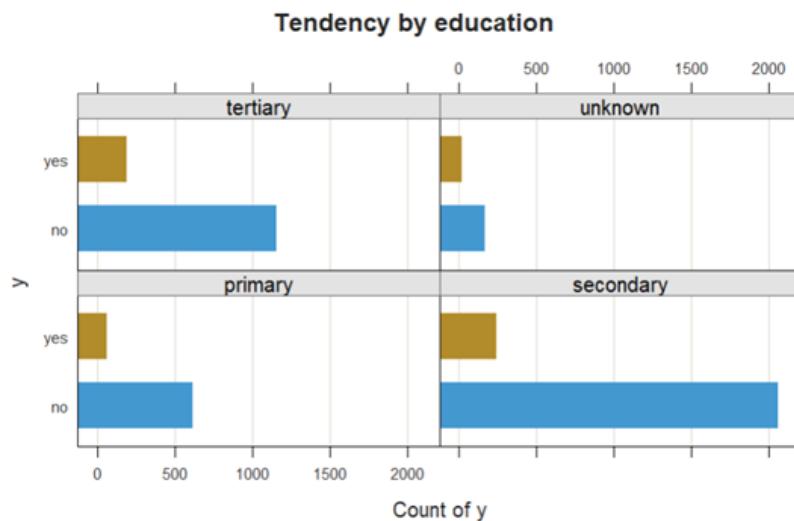
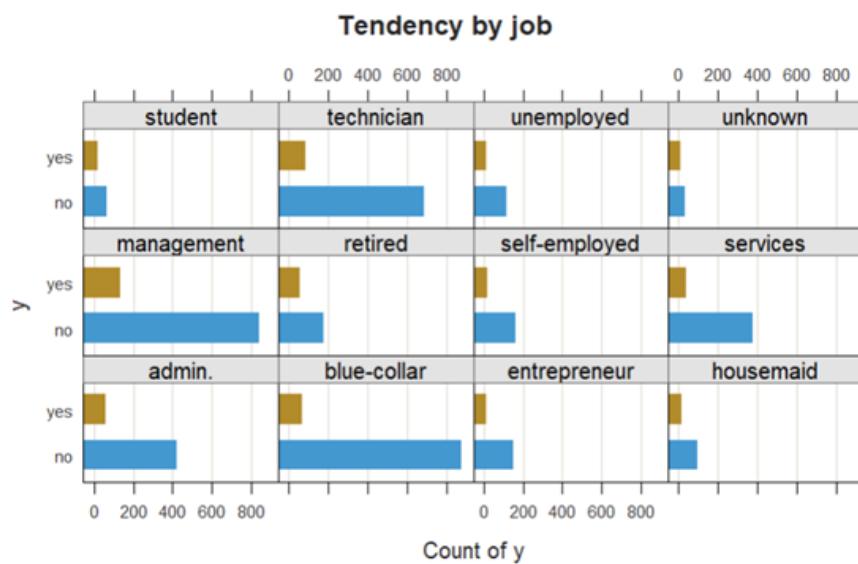
To explore relationships between numerical variables, we've included scatterplots, pairs and heatmap plots, helping us identify trends and correlations between these variables.



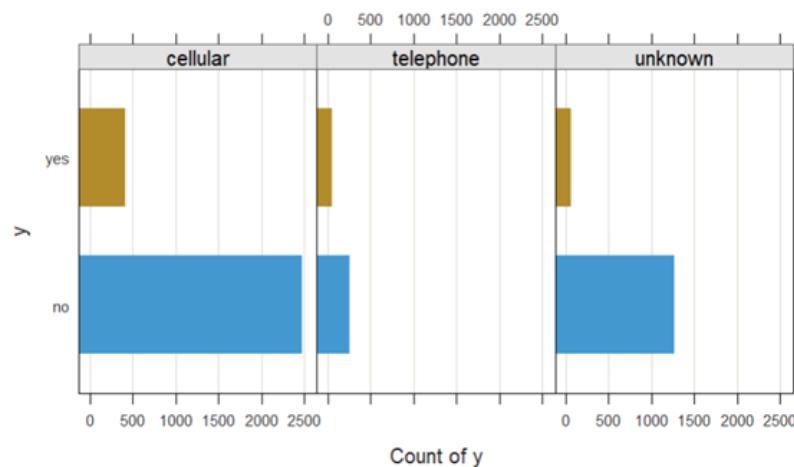
We create pairs and heatmap plots to visualize the correlations between 'age,' 'balance,' and 'duration.' While these plots hint at correlations, it's important to note that the presence of outliers can impact the strength of these relationships.

Categorical Columns

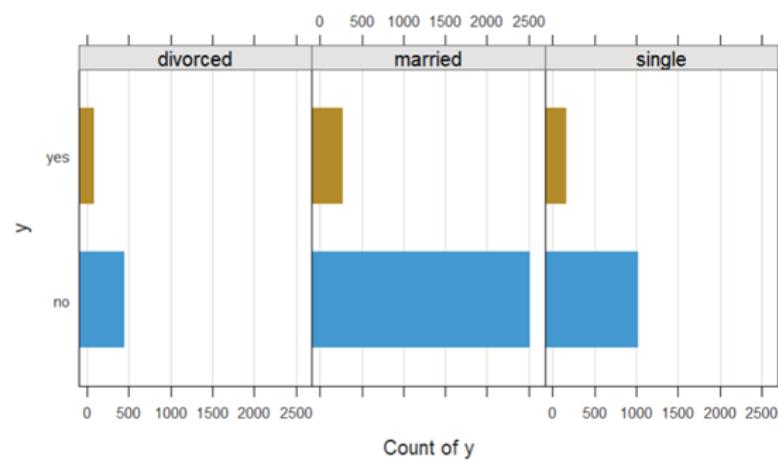
The categorical variables allow us to have a deeper analysis over the target population. For this analysis, we counted the occurrences of different categories within the most important columns: job, education, contact, and marital. Using these, we've plotted the count of subscriptions for a term deposit.



Tendency by contact



Tendency by marital status



In analysing the visualizations, we draw conclusions of the population that is more likely to subscribe to a term deposit:

- Marital Status:** Individuals who are not married (single or divorced) are more inclined to subscribe to a term deposit. This suggests that marital status plays a significant role in the decision-making process.
- Contact Method:** Subscribers are more likely to have been contacted by telephone. This communication channel seems to be more effective in encouraging term deposit subscriptions.
- Education Level:** It appears that individuals with only a primary education are more receptive to subscribe to term deposit.

By understanding the demographics and preferences of potential subscribers, it is possible to optimize and better select which variables to use when creating models.

4. PCA ANALYSIS

This chapter provides the Principal Component Analysis of our dataset. We used R scripts to execute PCA on the dataset and acquire the best 2D plane to visualize its data. Firstly, we split the dataframe into numerical and categorical parts, since PCA is conducted only on numerical data. We are left with 6 numerical columns as shown:

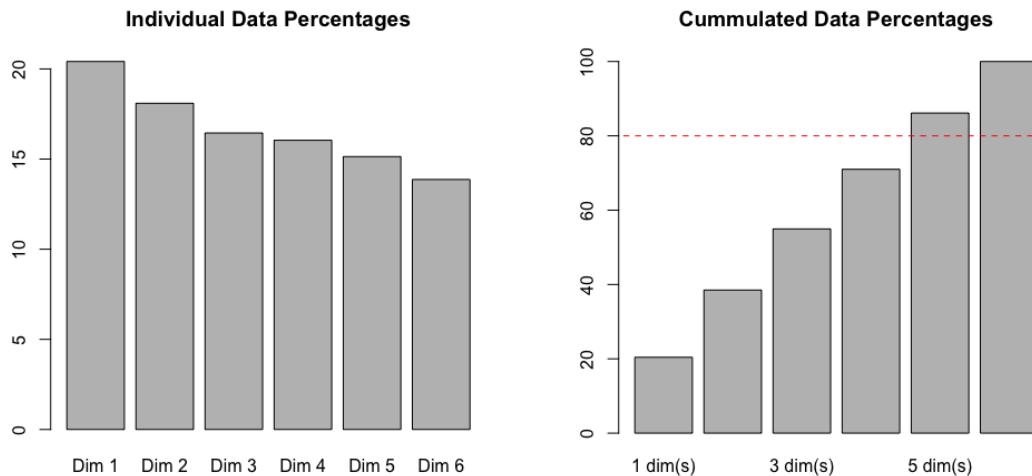
```
> summary(df[numerical])
      age          balance         day        duration       campaign      previous
Min. :19.00   Min. :-3313   Min. : 1.00   Min. : 4   Min. :1.000   Min. : 0.0000
1st Qu.:33.00  1st Qu.: 69    1st Qu.: 9.00   1st Qu.:104  1st Qu.:1.000   1st Qu.: 0.0000
Median :39.00  Median : 444   Median :16.00   Median :185  Median :2.000   Median : 0.0000
Mean   :41.17  Mean   :1423   Mean   :15.92   Mean   :264   Mean   :2.794   Mean   : 0.5426
3rd Qu.:49.00  3rd Qu.:1480   3rd Qu.:21.00   3rd Qu.:329  3rd Qu.:3.000   3rd Qu.: 0.0000
Max.  :87.00   Max. :71188   Max. :31.00   Max. :3025  Max. :50.000   Max. :25.0000
```

After executing the PCA, we get the eigenvalues and the rotation matrix for the 6 different dimensions:

```
> print(pca)
Standard deviations (1, ..., p=6):
[1] 1.1066900 1.0419491 0.9933714 0.9811107 0.9529797 0.9121644

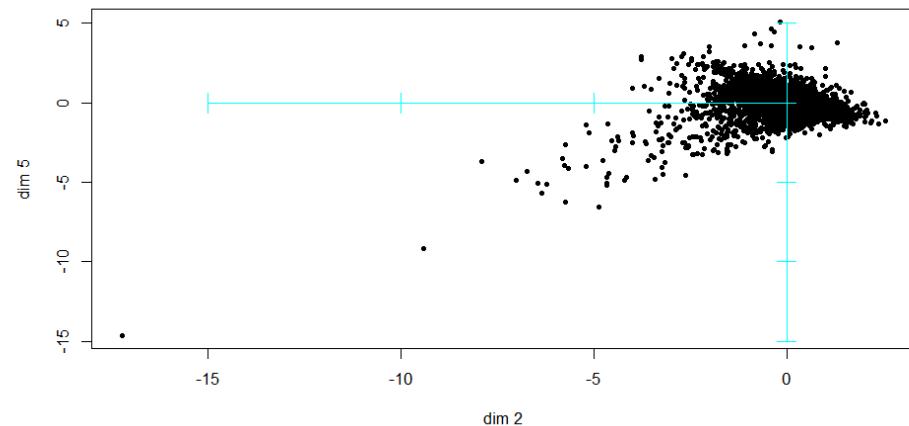
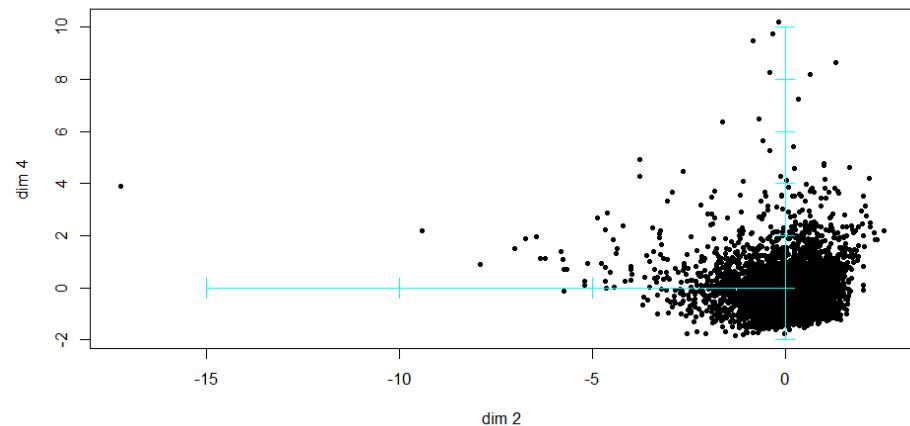
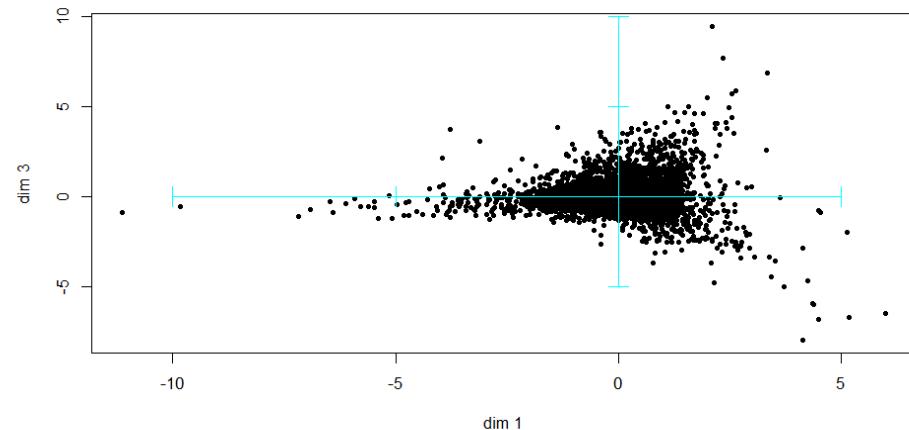
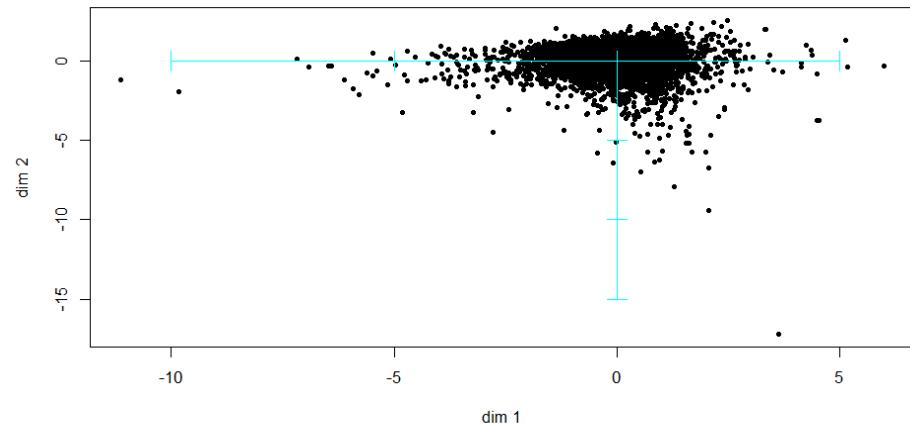
Rotation (n x k) = (6 x 6):
            PC1        PC2        PC3        PC4        PC5        PC6
age      0.09421272 -0.67118571  0.31800700 -0.1679807  0.63697552  0.07449768
balance  0.11083673 -0.69397914 -0.01840454  0.2233769 -0.67245040 -0.06068480
day     -0.59532770 -0.03303515  0.14779614  0.4384002  0.01836680  0.65583372
duration 0.27964444  0.23451275  0.80040989  0.4279559 -0.05770063 -0.19917613
campaign -0.63296663 -0.10458839 -0.03343592  0.2168777  0.14165457 -0.72124518
previous  0.38154468 -0.02960409 -0.48468049  0.7067590  0.34400570 -0.02799720
```

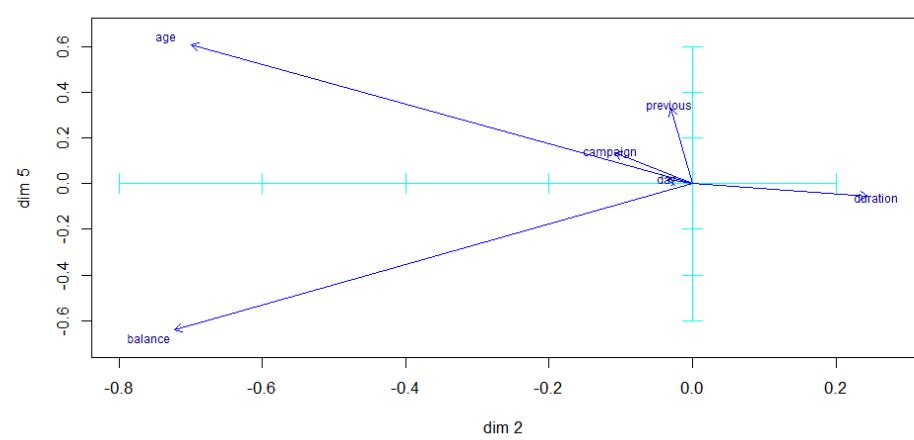
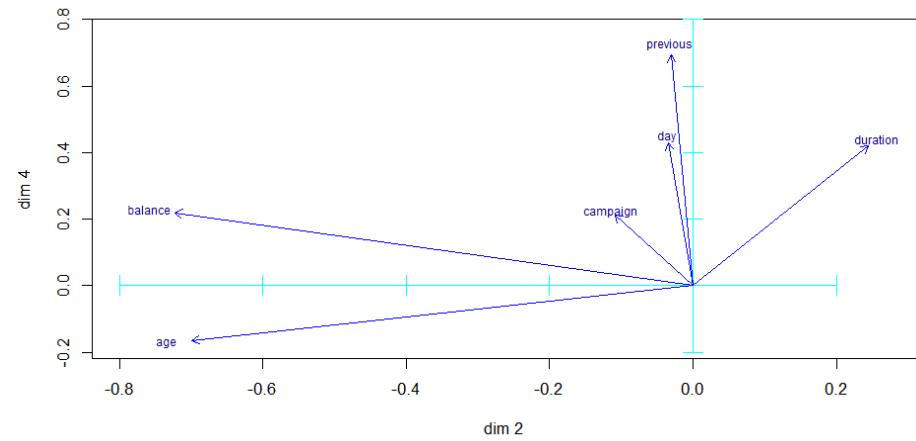
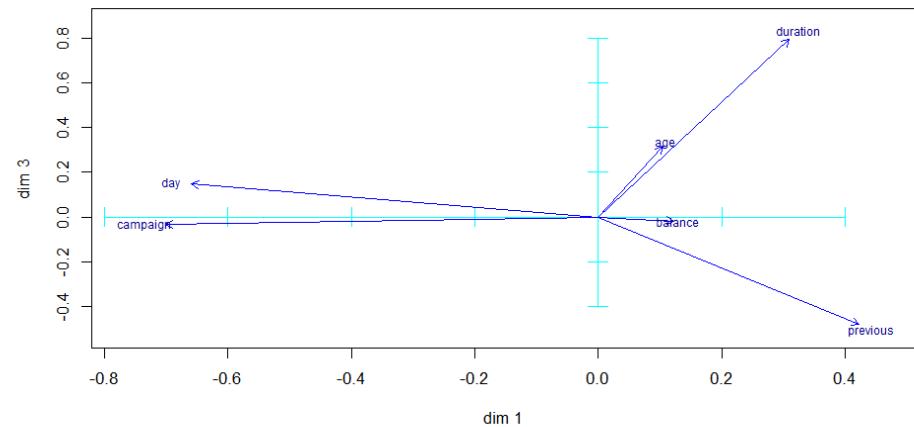
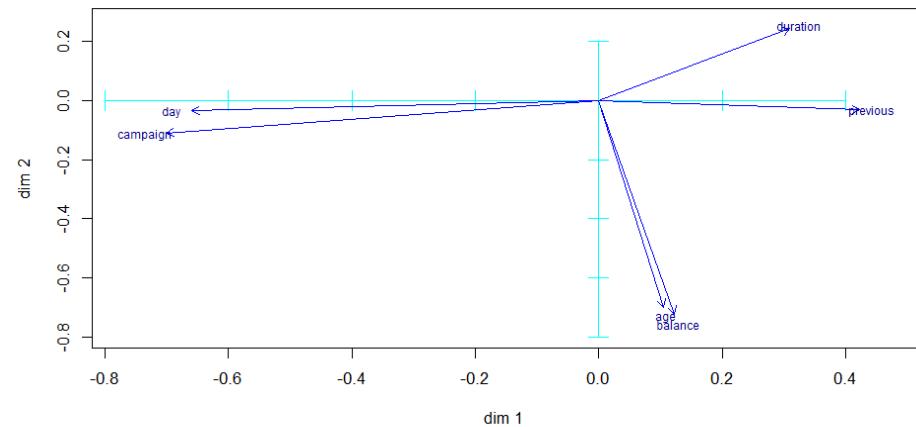
First, the meaning of the eigenvalues (standard deviations) is how much of the variance does each dimension of the transformation explain, and the matrix displays how each of the variables affect the new dimensions. To get a better view of what the deviations mean, we can show their respective values as a percentage, also referred to as data inertia, as well as its cumulative function.



The red dashed line on the cumulative graph represents the 80% threshold of the total data. We use it to find the number of dimensions which represent the majority of the variance, which in our case is 5, and the only dimensions we'll use for the PCA analysis. We also observe that the PCA can't reduce the dimensions very much as all numerical variables seem not to be greatly correlated.

If we plot our data in some of the combinations of these new dimensions, we see some tails appear. That is probably because, even though the covariance is 0, the variables aren't independent, and thus not normally distributed.





Additionally, if we use the original variables in those plots instead of the individuals, we can get a rough feeling for what each of the new dimensions mean. In those plots we see:

- The first dimension seems to represent if the number of calls decreased, and the final call was a bit longer and towards the start of the month.
- The second dimension is slightly about longer calls but mainly about young clients or with less money.
- The third dimension is for older people that hadn't been called much with a focus on long calls.
- The fourth dimension is for long calls at the end of the month, and most importantly, of the clients that had received many calls before.
- The fifth dimension is for older people without too much money.

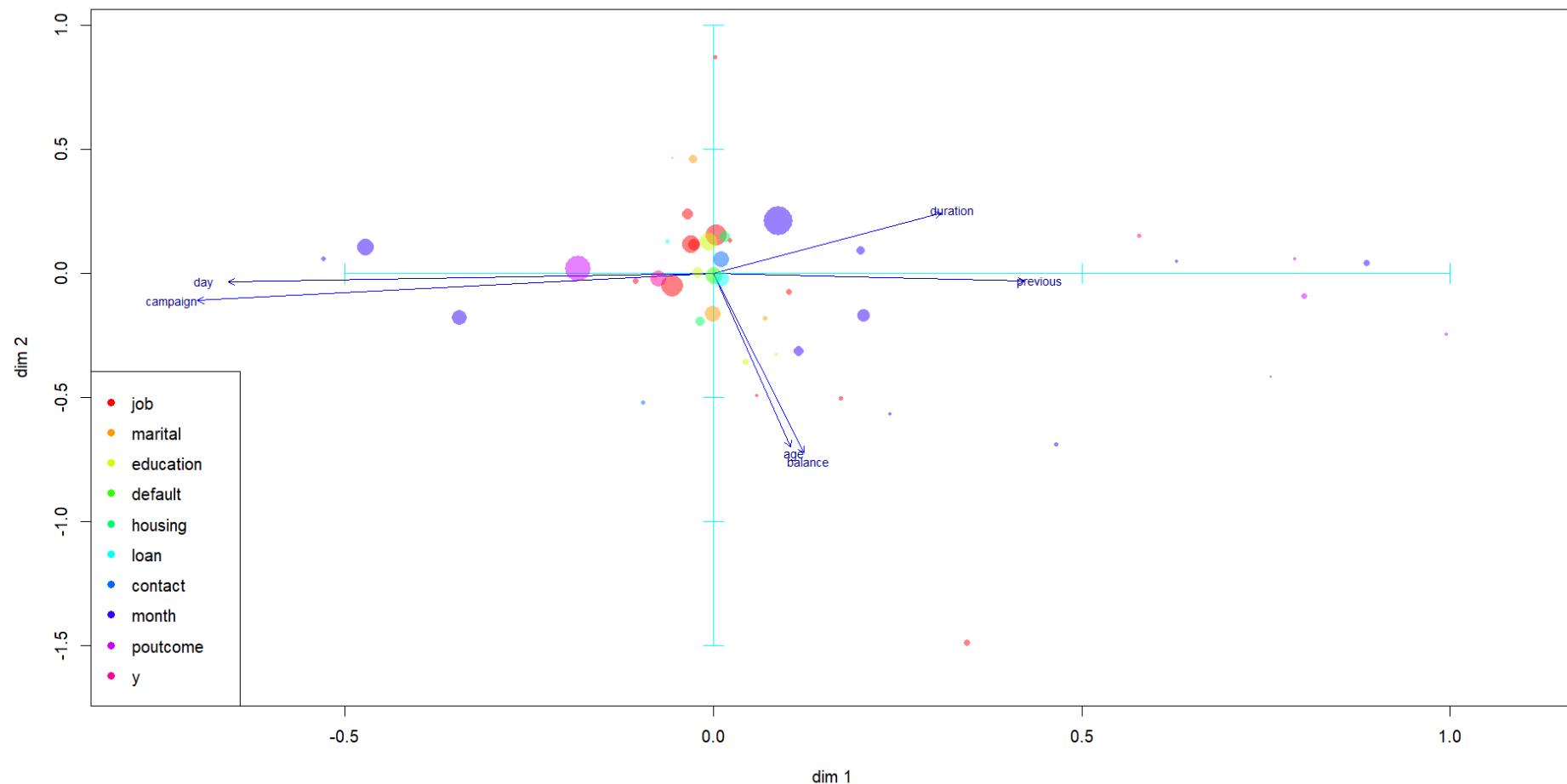
In short, the different dimensions represent the contact frequency, the client profile, the call's length, when and how the call went and the client profile again.

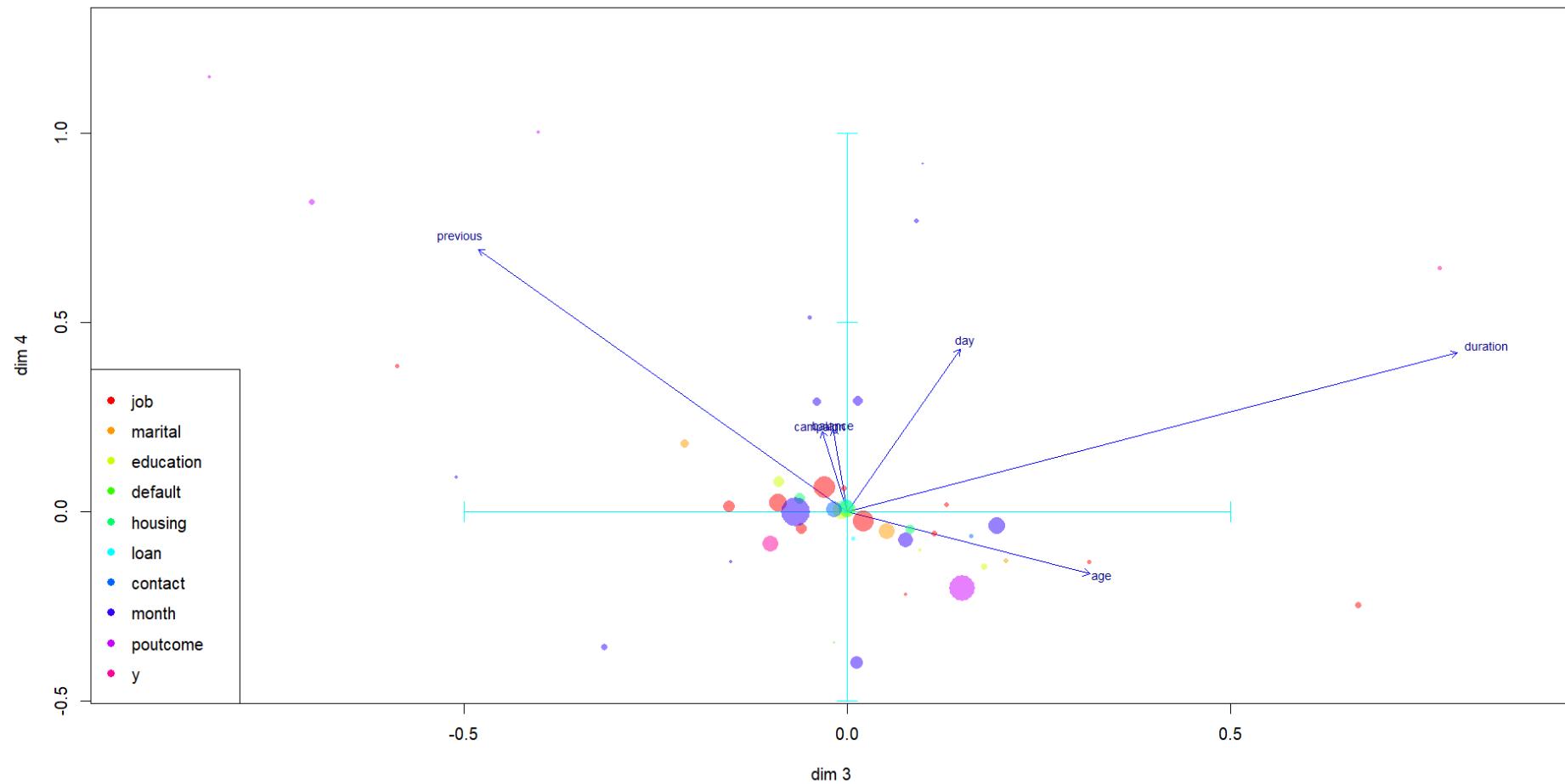
Apart from the numerical variables we also have the categorical ones. We can add their centroids to the plots to check how they relate to the numerical and how useful they may be. After that we can check individually one by one each of the ones with the most relation to the numerical variables (the centroids furthest from the origin).

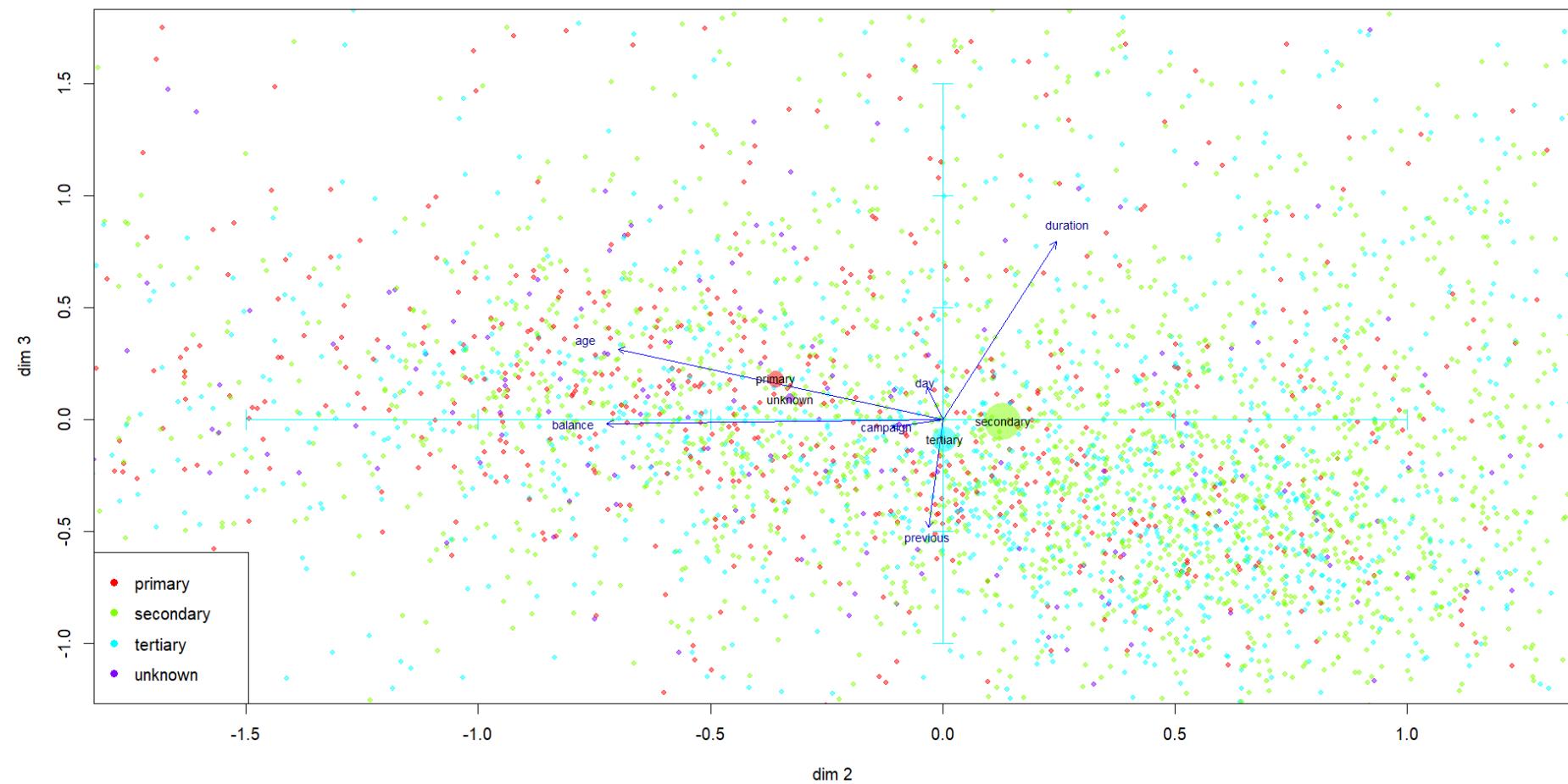
From those plots we see that the most related are:

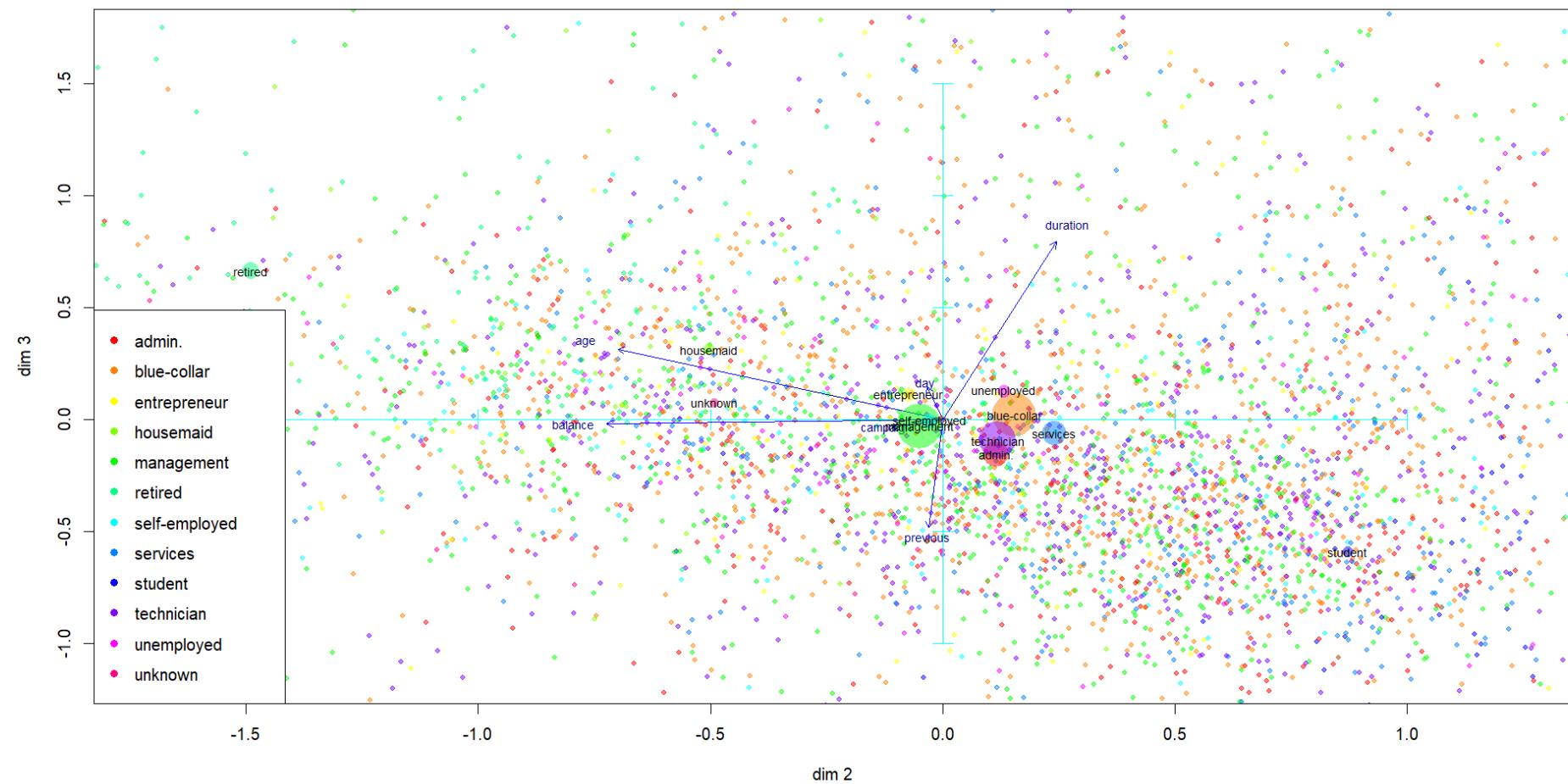
- Education with the second and third dimensions
- Job with the second and third dimensions
- Month with the first and fourth dimensions
- Outcome of the previous campaign (poutcome) with the first and fourth dimensions
- Marital status (marital) with the second and third dimensions

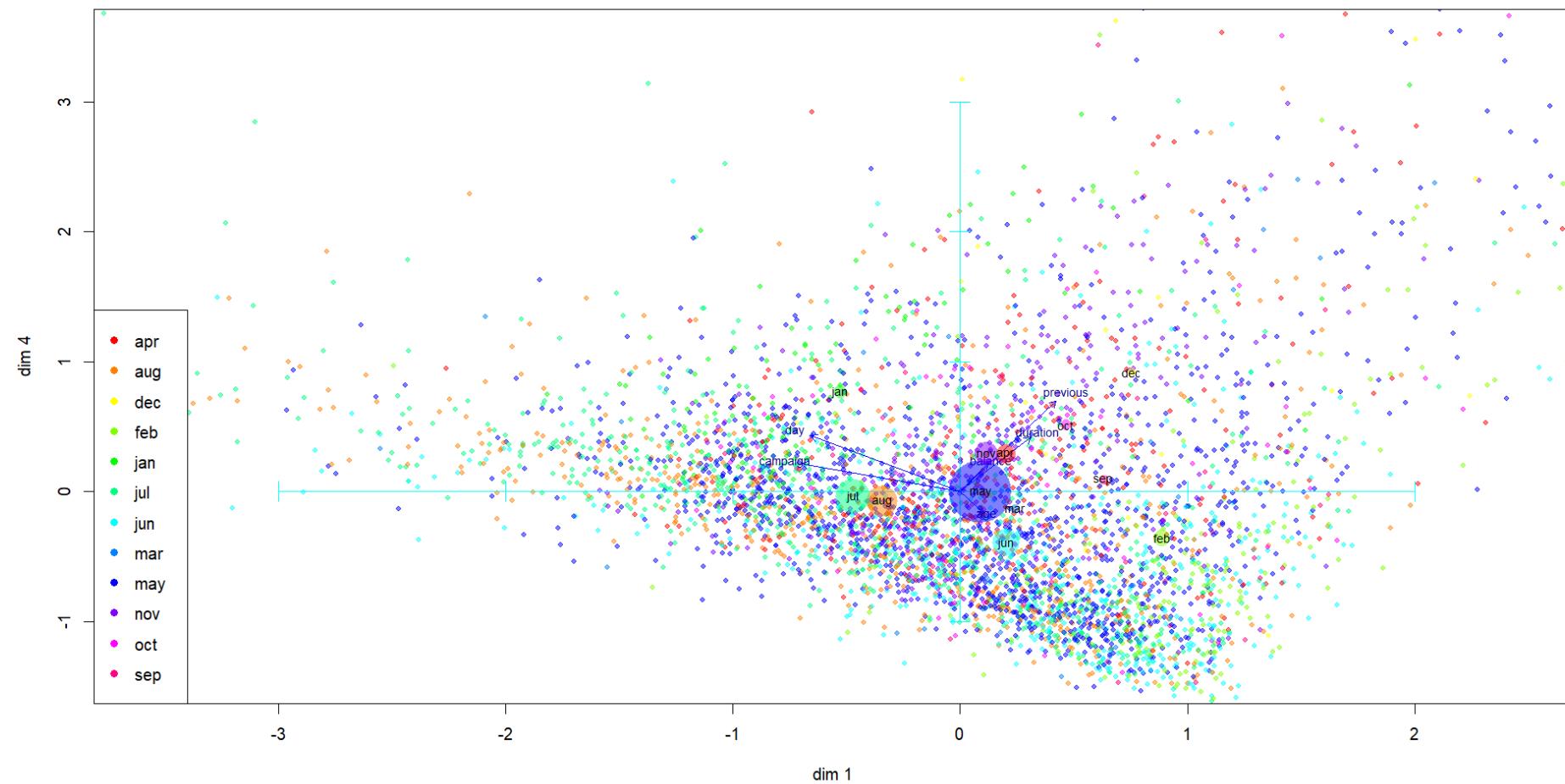
We can now plot those categories on their most prominent axis to see the most influential modalities and if the values are clustered. Additionally, plot the y category (our target for the analysis) to check which concept (PCA dimension) represents it the best.

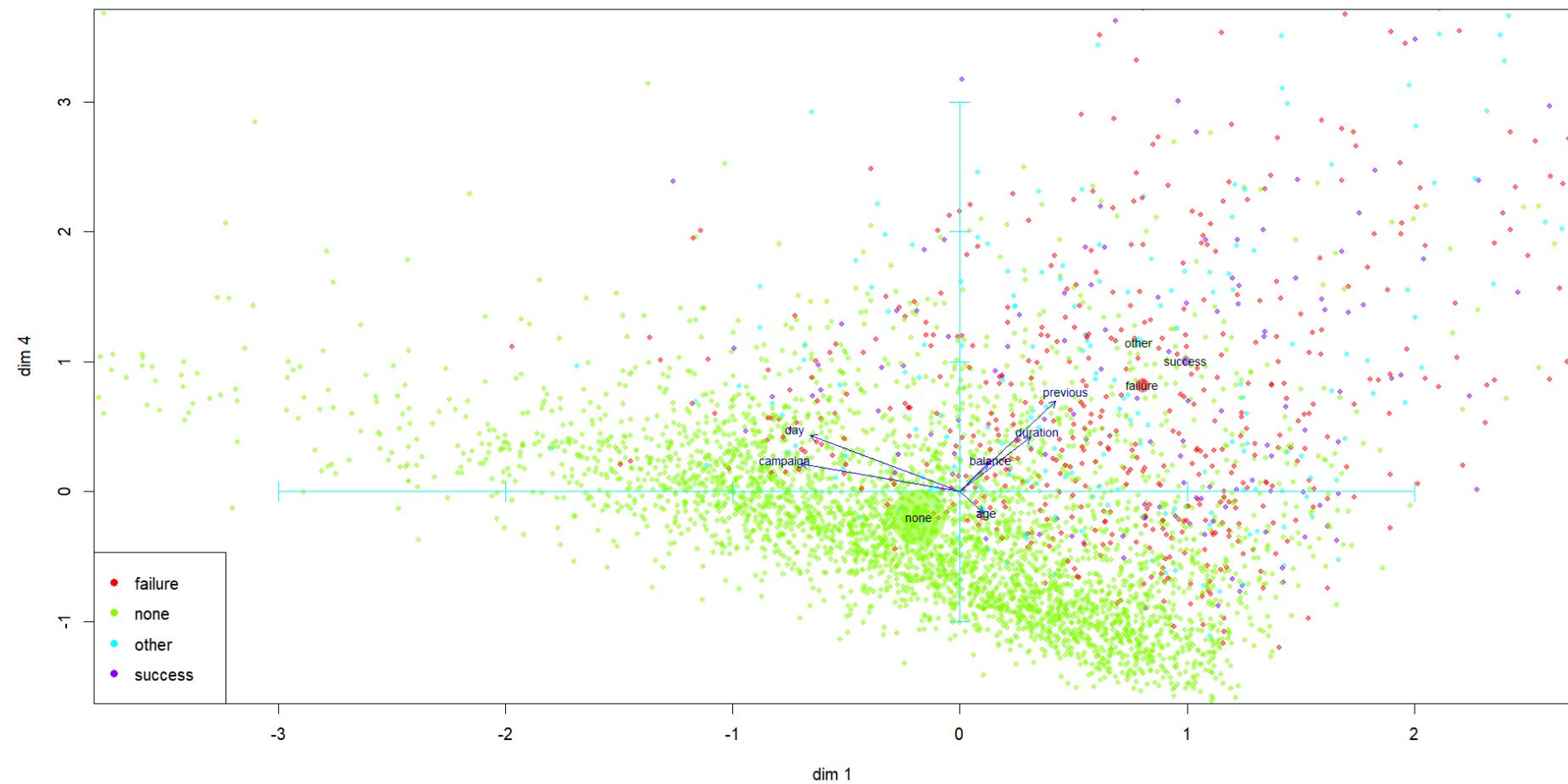


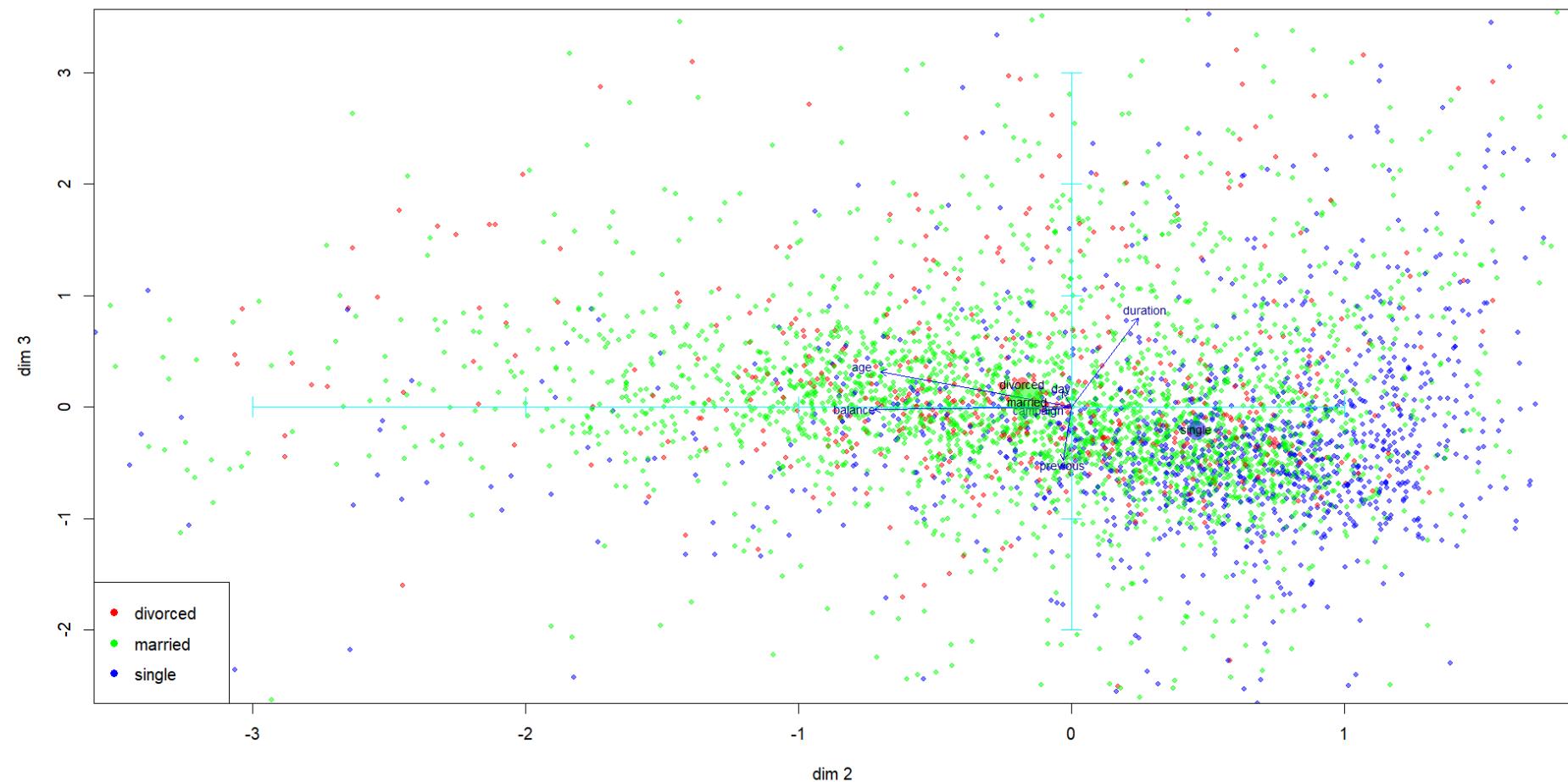


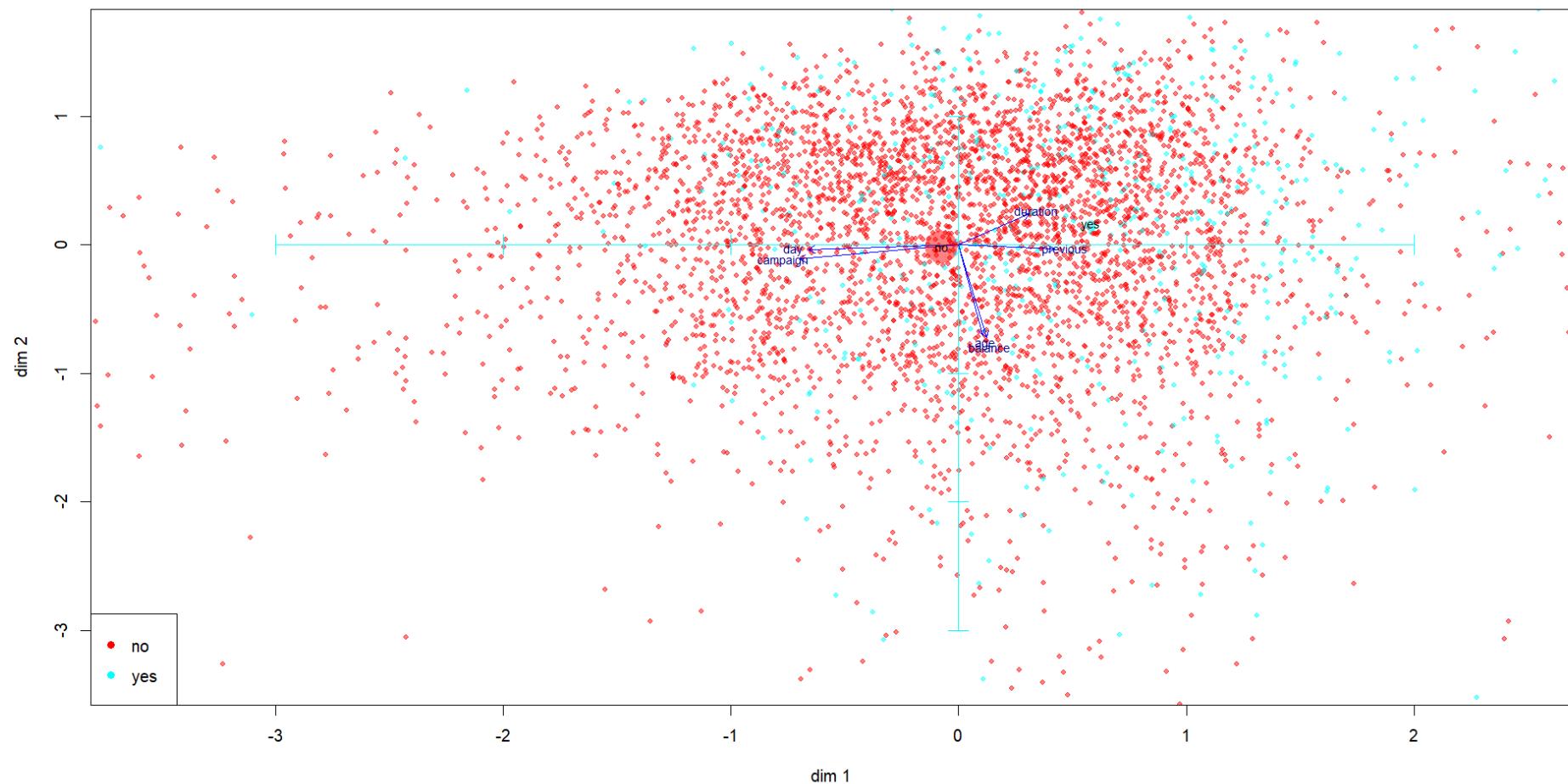


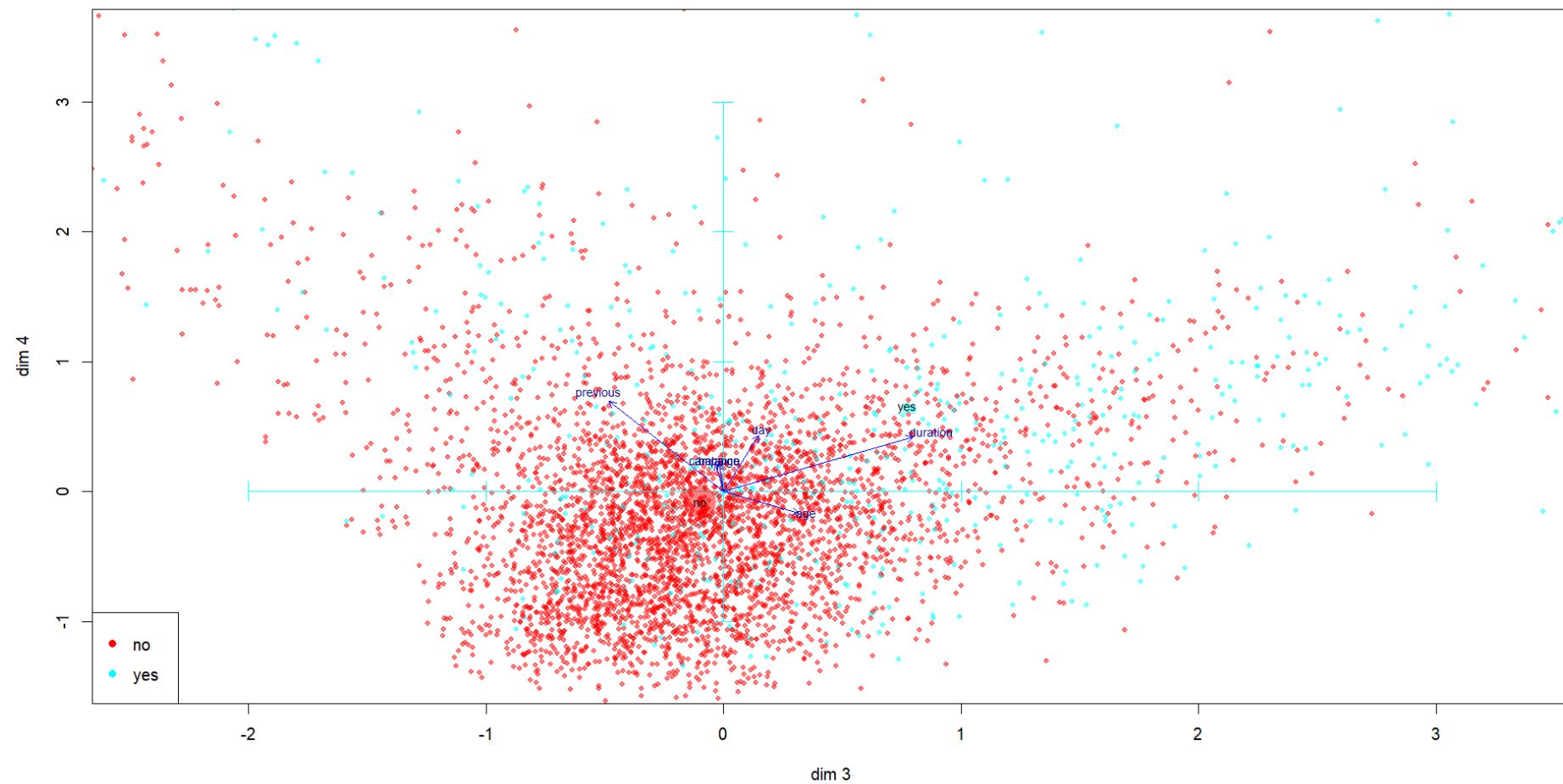










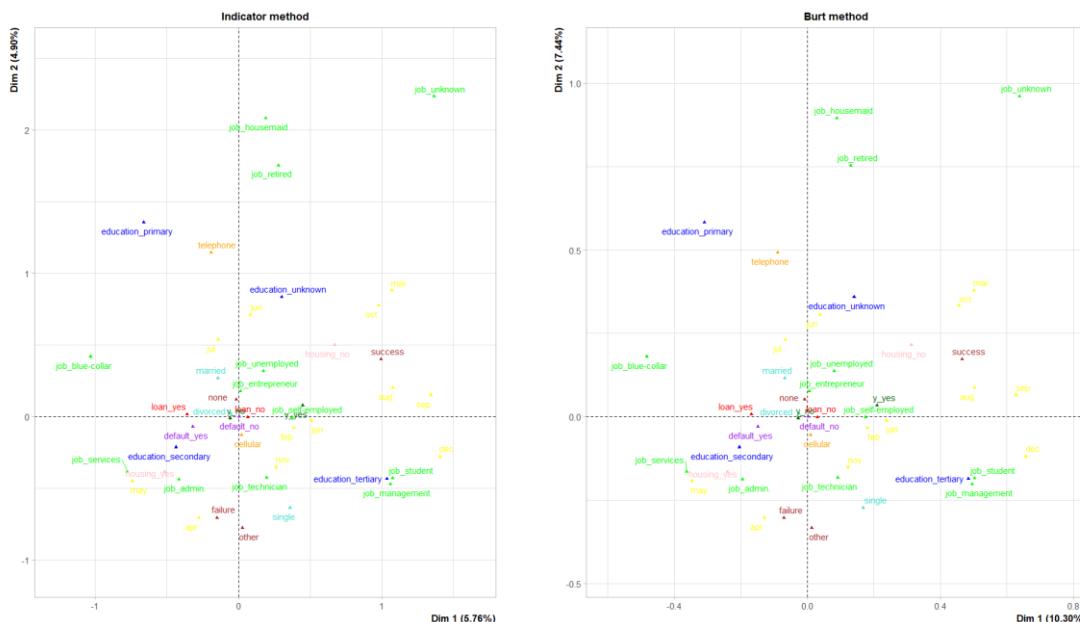


For all those factors we can conclude the following:

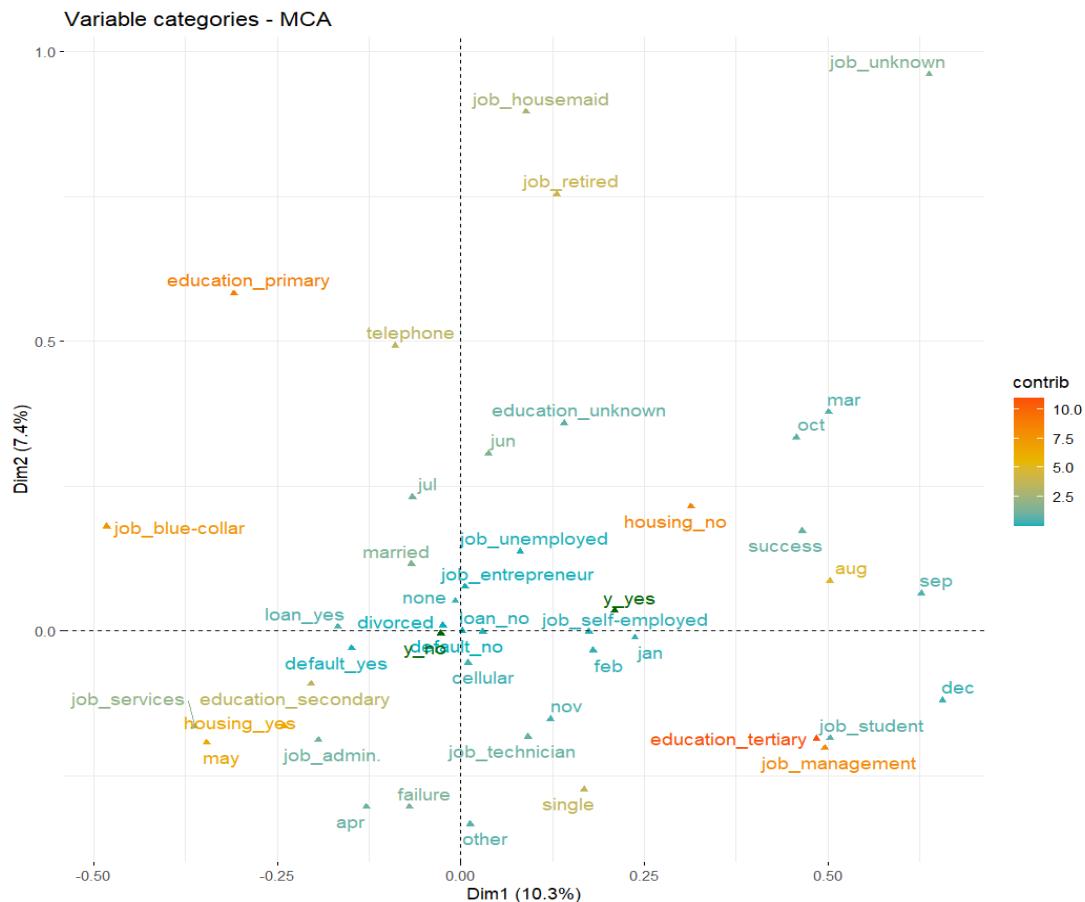
- Educational: There is a noticeable difference between the primary education and the other two levels according to the client profile.
 - Job: There are three modalities that display a unique behavior, from the rest. Those are students, housemaids and retired.
 - Month: There is a slight separation between the months of autumn and winter versus spring and summer.
 - Previous outcome: Although there is the noticeable cluster of the none modality on one of the sides aligned with previous, duration and balance. This was already expected as poutcome will always be none when the client wasn't contacted previously. (previous = 0)
 - Marital status: There is a noticeable difference between people that have and haven't been married according to the client profile with two partially overlapping clusters.
 - Has subscribed (y): Is most notable on call frequency and duration. And a positive result is more aligned with more previous contact, less now and longer calls.

5. MCA ANALYSIS

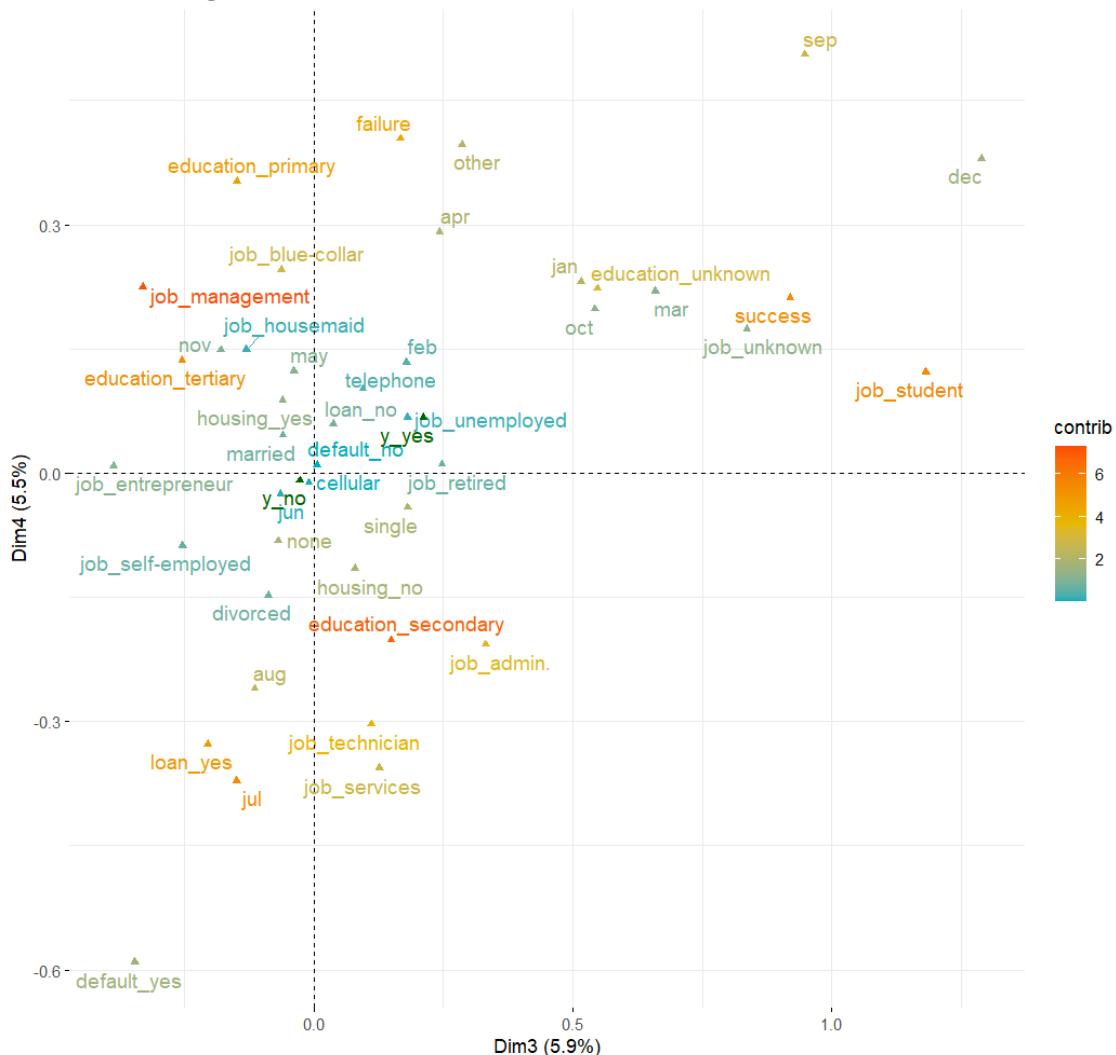
After completing the PCA which gave us insights about the correlations of numerical variables, this chapter will go through the Multiple Correspondence Analysis to gain knowledge on the relationships between categorical variables. First, we split the dataframe into numerical and categorical parts, as the MCA uses only categorical values, which leaves us with 9 explanatory variables and the response. We used both Indicator Analysis (Logic Table) and Burt Analysis (Burt Table) to check how they compare, which yielded similar results for our dataset.

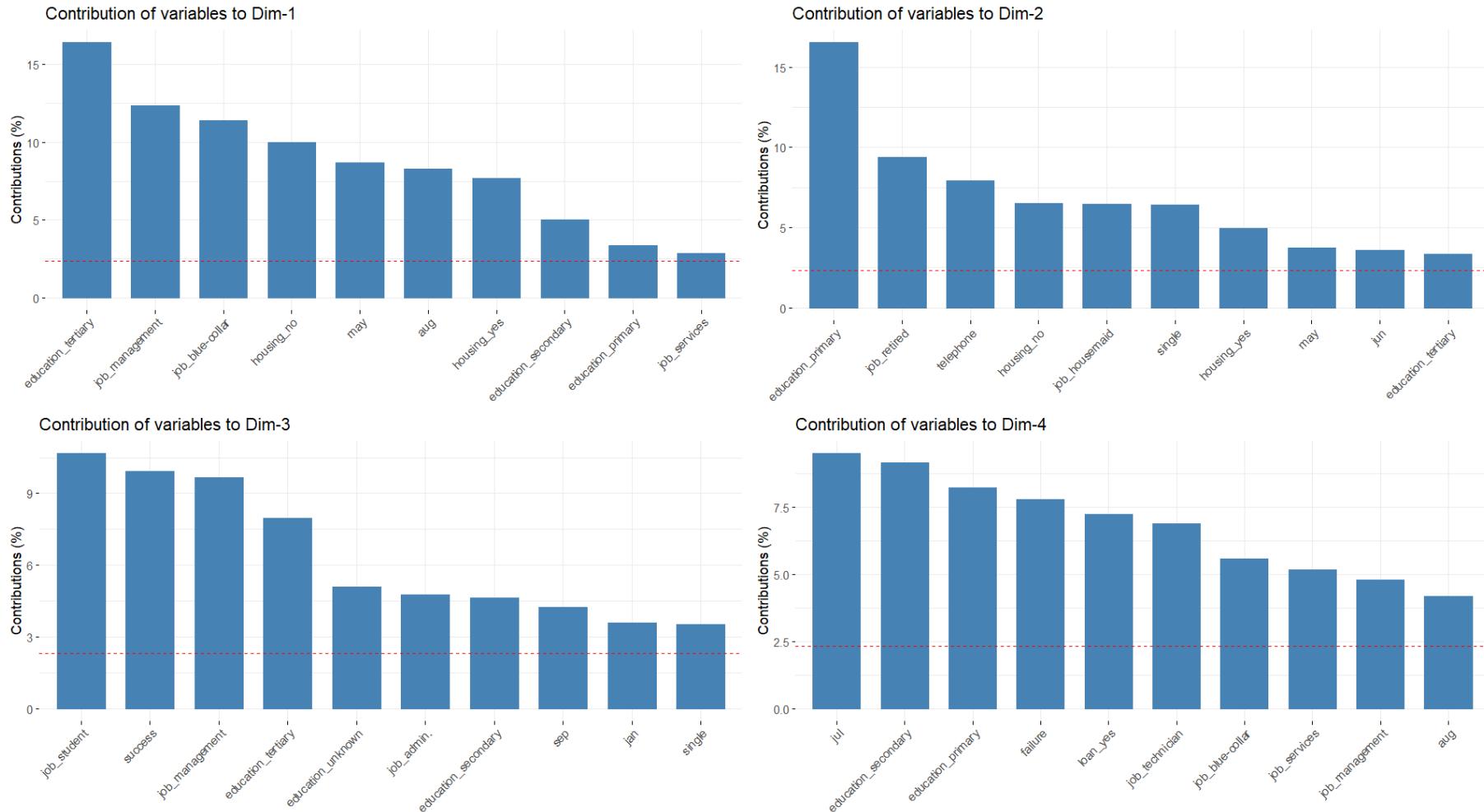


So, as both methods map the variables to the same relative locations, only changing the scale and the variance each dimension represents, we'll continue the analysis with the Burt method's results.

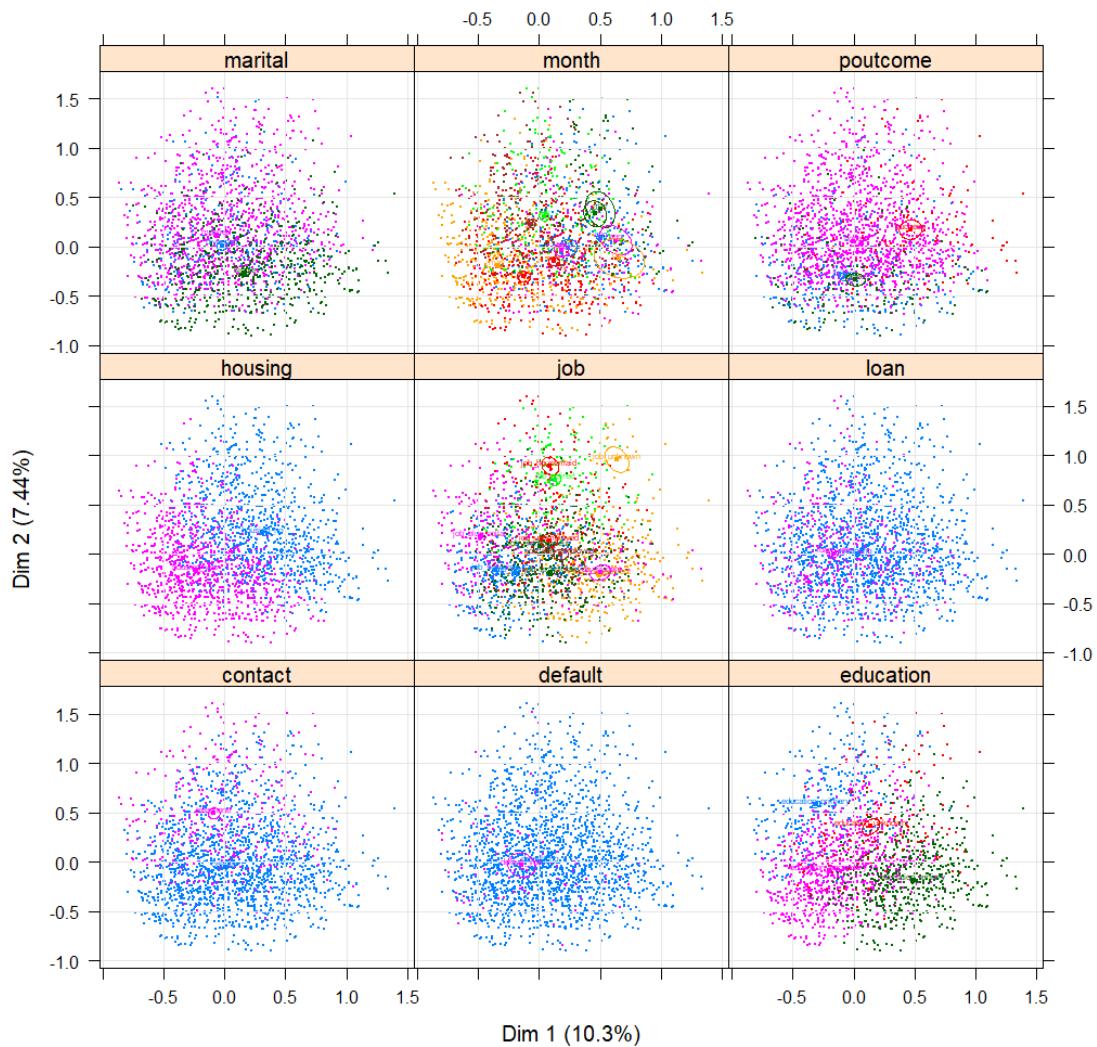


Variable categories - MCA



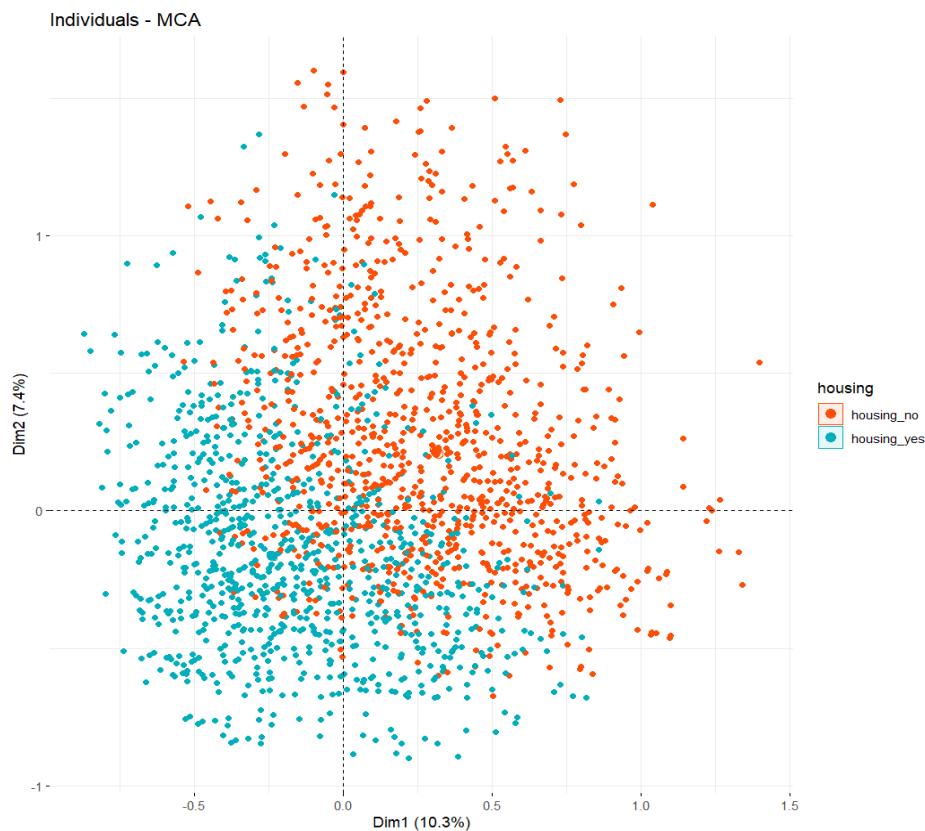
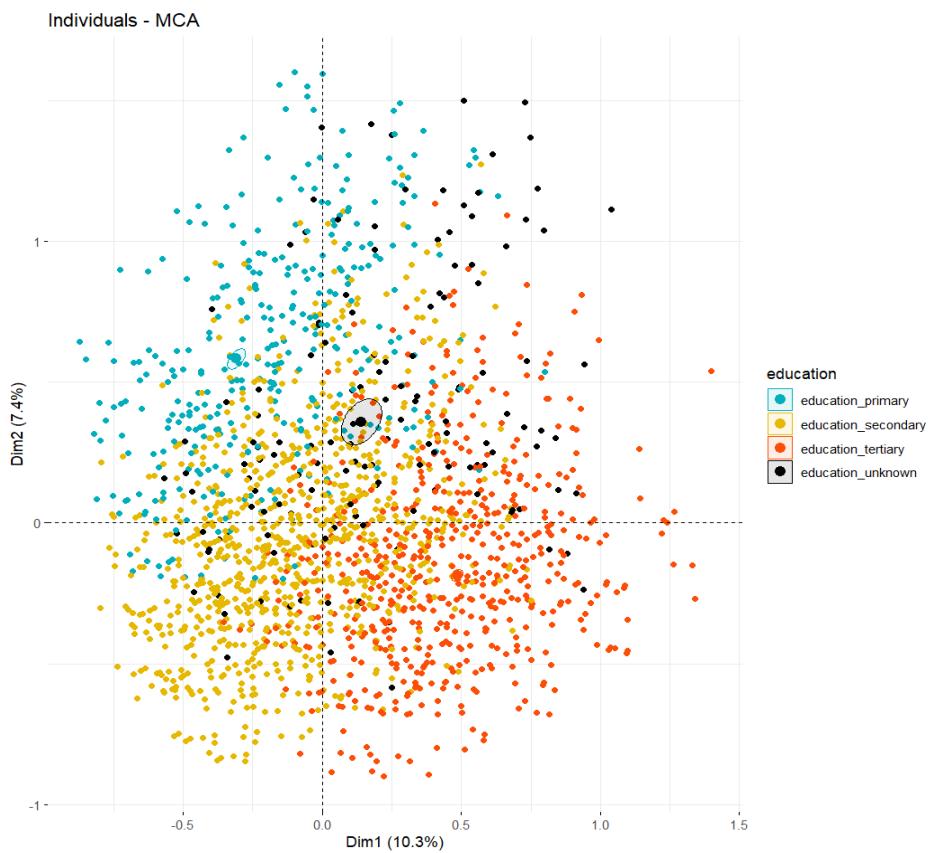


For the first two plots, they represent the space after the MCA transformations for the first and second dimensions or third and fourth dimensions respectively. The position of each value shows its correlation with that dimension and the colour shows its contribution for those two dimensions. With this plot we can see which variables may have a bigger leverage (with the colour) and if their modalities are very similar (close to one another) and thus not very informative or more distinct (further away) meaning they may be useful categories. Because the contributions may not be fully clear, we also have a plot of the main contributors to each of the first four dimensions. Using this information, we can gather that the most relevant categories will end up being the education level, the job, if they have a housing loan, the previous outcome and maybe also the month, the marital status and the contact method. If we plot the individuals coloured according to the different categories, we can check which of the factors (most likely the ones with higher contributions) may be the most useful later on.



The most notable ones are housing and education that show a clear distinction between the individuals of different modalities, then marital and poutcome to a lesser degree, and

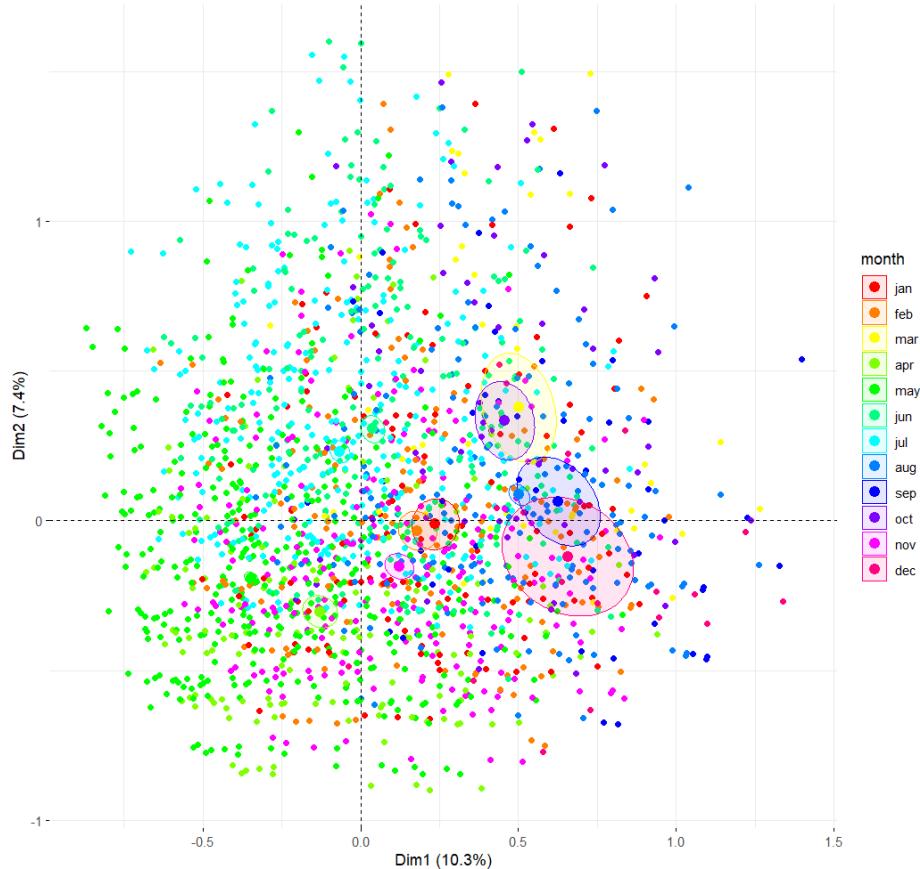
finally, for the plots of job and month, the data is so cluttered that that we'll need a better view.



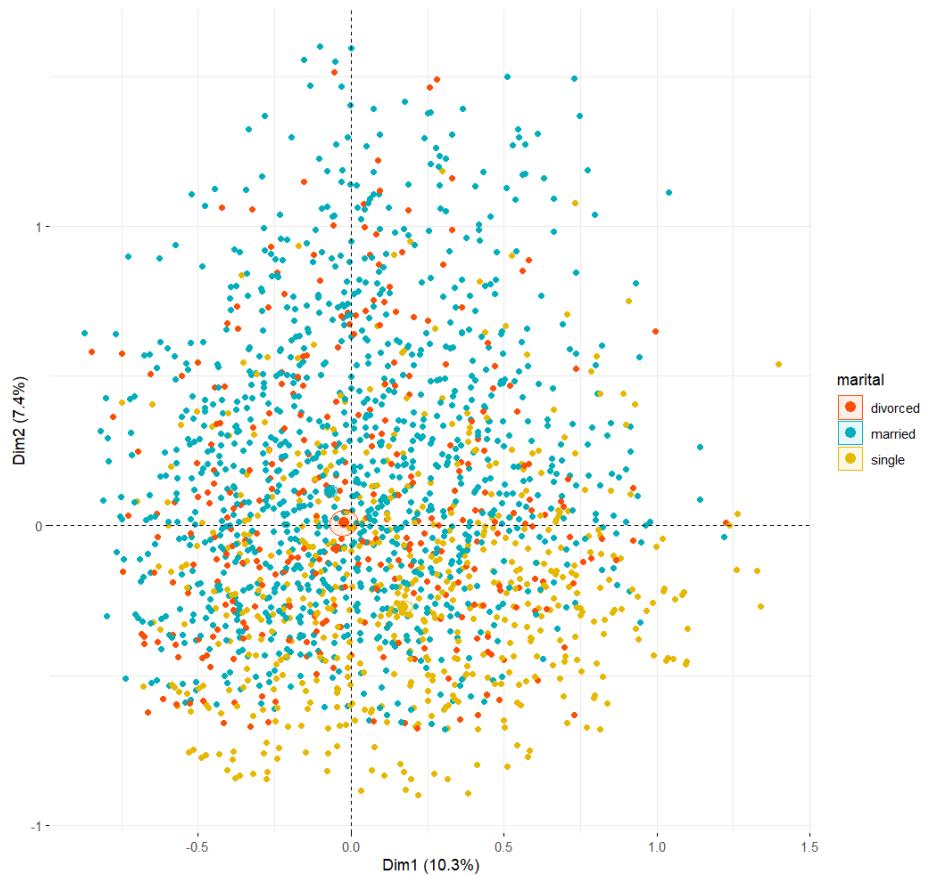
Individuals - MCA



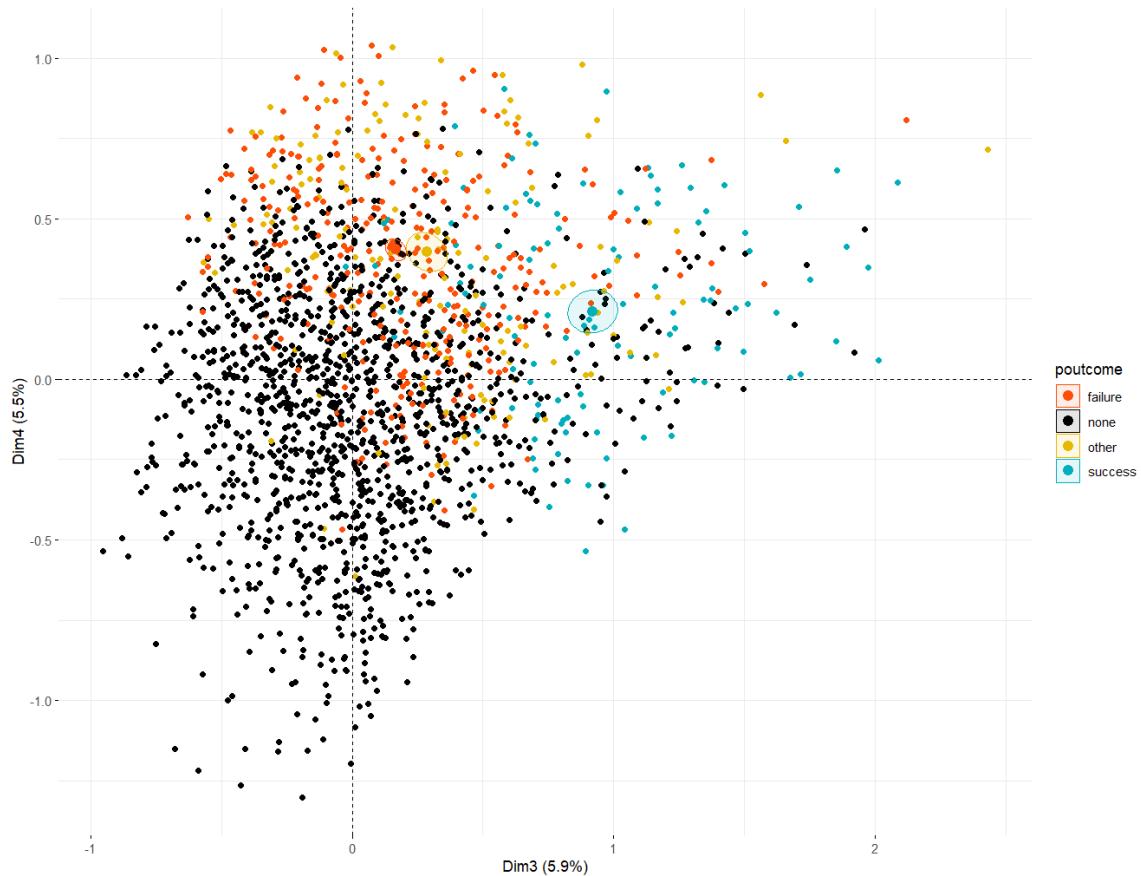
Individuals - MCA



Individuals - MCA



Individuals - MCA



For education, we see the clear difference according to the level with education increasing from the second quadrant towards the fourth. In housing loans, we see also the difference with no loans on the first quadrant and instead on the third. For jobs we see some of them are more remarkable: blue-collar, housemaid, retired, unknown, student, admin and services. So, for this factor, if need be, it could be considered to join the rest of modalities. Regarding months, we have from April to July (spring to summer) on the left side, while on the right, the rest of them are mixed. For the marital status we see again only single people are somewhat separate from the rest, benign more prominent on the third and fourth quadrants. Finally, the outcome of the previous campaign, projected on the third and fourth axis as that's where it has relatively the biggest relevance, shows three distinct (but still overlapping) groups. For the not contacted slightly towards the third quadrant, for the successful on the first and the other two together between the first and second quadrant.

To end the MCA, we have to see how these new dimensions relate to the response variable. If we plot the samples along the four first dimensions, we see that the modalities for the variables are overlapping significantly, so the factors alone probably won't solve the problem alone. We also see that the two centroids are only separated in the first and third dimensions. This means that the people more likely to subscribe are of higher education, without house loans, with jobs either unknown or still students instead of blue-collar or services, not by the end of spring and that already said yes to the previous campaign.

