An overview of Supervised/Unsupervised Analysis

Giuseppe Scaffidi Caruso, MAT 989728

Juy 2022

Sommario

The aim of the following analysis will be to use certain statistical model such as **Supervised Unsupervised learning** and the associated models with it in order to come up clear statistics from different data sets. The project will apply Supervised techniques such as Logistic Regression, Tree Predictor and Unsupervised Techniques such as K-Mean and Principal Component Analysis on *Heart Disease* and *Mall Customer* data set respectively.

Indice

Supervised Learning

As the theoretical parts suggest the Supervised Learning is a ML method where the algorithm, exploiting a label data set, is able to train on a **Training Set** and test this trained model on a **Test Set**. The peculiarity of this model is related to the label data set in fact the latter have to be encoded by human annotator, this implicates a long, expensive process. Supervised machine learning is often used for:

- Classifying different file types such as images, documents, or written words.
- Forecasting future trends and outcomes through learning patterns in training data.

1 Heart Failure prediction

1.1 Project Overview

As the title suggest, the data set collect information about 299 patients with Hearth Failure and a cardiovascular Disease. The main effect given by this problem is the progressive inability of the heart to pump blood as the time goes by. The data set contains 13 feature about human functions, their aim is to predict one dependent variable which is the **DEATH EVENT**

1.2 Data set Explanation

The data set presented for this project collect 299 patient's observation collected between April-December 2015. The data structure is composed by 105 women observation and 195 man observation whose age range are between 40 and 95 years old. The column are 13 that indicates the same number of variables in which 12 of them are related to independent variable (*predictors* and 1 of them is the dependent one (*criterion*). In the table above there will be the table of features.

Feature	**Explanation**
Age	Age of the patient
Anaemia	Decrease of red blood cells
High blood pressure	If a patient has hypertension
Creatinine phosphokinase	Level of CPK enzyme in the blood
Diabets	If the patients has diabets
Ejection Fraction	Percentage of blood leaving
Sex	Woman or Man
Platelets	Plateletes in the blood
Serum creatinine	Level of creatinine in the blood
Serum sodium	Level of sodium in the blood
Smoking	If the patient smokes
Time	Observation Period (or Follow-up period)
Death Event	If the patient die during the follow up period

Tabella 1: Data set Features

Taking a look to the data set's summary it's possible to understand the main characteristic for all features.

```
── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable            n_missing complete_rate      mean        sd     p0      p25      p50      p75     p100 hist
1 age                              0             1     60.8      11.9     40       51       60       70       95 ▆▇▃▂▁
2 anaemia                          0             1      0.431     0.496    0        0        0        1        1 ▆▁▁▁▇
3 creatinine_phosphokinase         0             1    582.      970.      23      116.     250      582     7861 ▇▁▁▁▁
4 diabetes                         0             1      0.418     0.494    0        0        0        1        1 ▆▁▁▁▇
5 ejection_fraction                0             1     38.1      11.8     14       30       38       45       80 ▂▇▅▁▁
6 high_blood_pressure              0             1      0.351     0.478    0        0        0        1        1 ▆▁▁▁▇
7 platelets                        0             1 263358.    97804.    25100   212500   262000   303500   850000 ▂▇▁▁▁
8 serum_creatinine                 0             1      1.39      1.03     0.5      0.9      1.1      1.4      9.4 ▇▁▁▁▁
9 serum_sodium                     0             1    137.        4.41   113      134      137      140      148 ▁▁▂▇▃
10 sex                             0             1      0.649     0.478    0        0        1        1        1 ▅▁▁▁▇
11 smoking                         0             1      0.321     0.468    0        0        0        1        1 ▆▁▁▁▇
12 time                            0             1    130.       77.6      4       73      115      203      285 ▇▆▆▆▃
13 DEATH_EVENT                     0             1      0.321     0.468    0        0        0        1        1 ▆▁▁▁▇
```

Figura 1: Summary Heart Failure

As we can see from the plotted statistics the mean of CPK (*creatinine phosphokinase* is 581.8 g while the max amount is 7861 g, this means there are certain outlier inside the variable. Clinically speaking the values of certain substances have to be in a certain range in order the body works well. For example:

- The platelets count range between 150.000 and 450.000
- Serum Creatinine range between 0.7 and 1.2 mg/dL

It can be noticed how the range quantities listed above have some peak in the dataset with a max platelets amount of 850.000[1] and a max Serum Creatinine of 9.4g. It is also relevant take a look at the observation period (Time) which gives us information about the Death time period of each patient, the mean death time is at 130.3 days. In order to make a strongest idea about that, below it has been shown the same concept with a plotted graph:

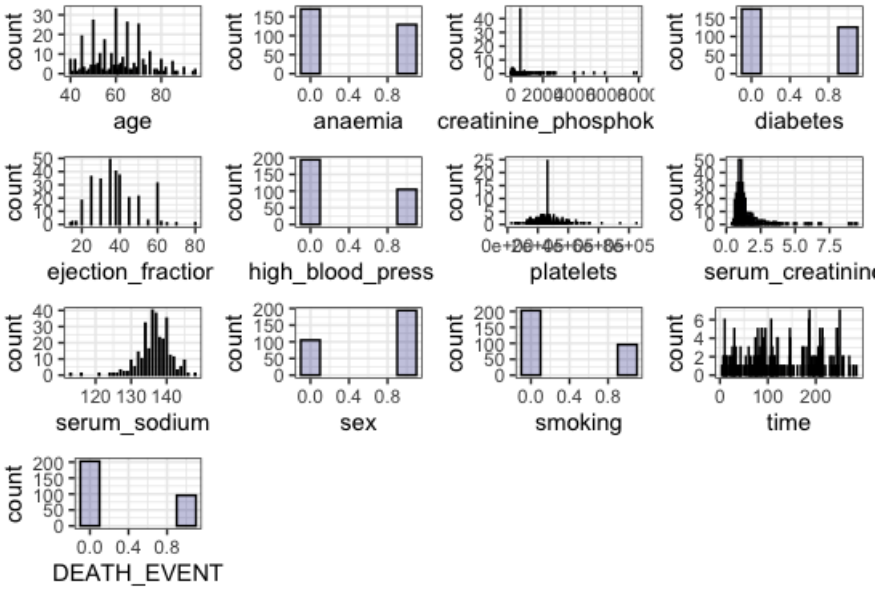

Figura 2: Plot of Summary

[1]From the plotted graph the output can be wrong but it's only a graph adjustment due to too many values.

From now on the task will be to build a framework in which it's possible to pick the correlation of the independent variable with the dependent variable (Death Event). Some features won't have differences in the plot while other will have slightly differences such as serum creatinine, serum sodium and time. We can confirm if the features are normal distributed with the

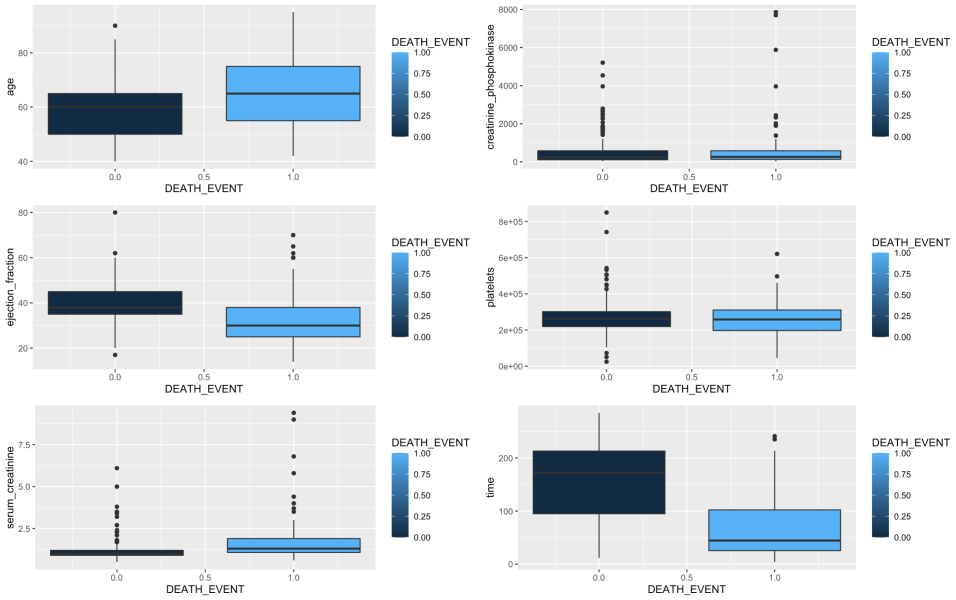

Figura 3: Death Event ~ features

Kolmogorov-Smirnov test. The Kolmogorov-Smirnov test is (The Kolmogorov-Smirnov test is used to test the null hypothesis that a set of data comes from a Normal distribution in particular the Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution functions of two samples."). From the Figure 3 recalling a theoretical and graphical demonstration of this normal distribution test and plot we obtain:

	p.value
anaemia	0
creatinine_phosphokinase	0
diabetes	0
ejection_fraction	0
high_blood_pressure	0
platelets	0
serum_creatinine	0

Figura 4: Kolmogorov-Smirnov test

None of these value tested are normal distributed cause their values are less than 0.05 (all of them ¡ 0.05). The non normality assumption earned from the test admit another assumption, based on the fact that some outlier brake the normality and does not favor it. For these reason it's possible to plot a graph where these outlier can be noticed

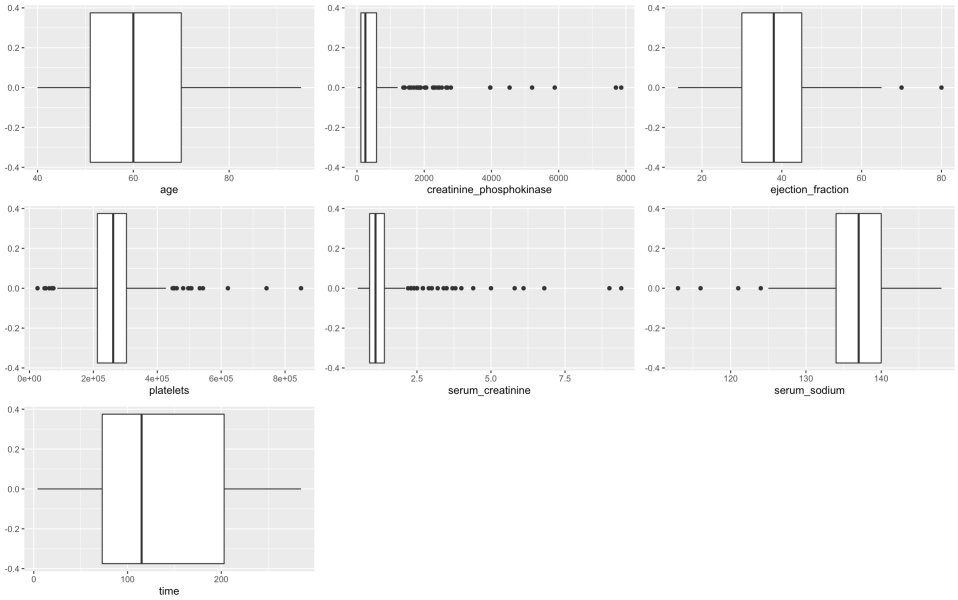

Figura 5: Box Plot for Outliers

From the Figure 5 it's clear how the only values don't contain outliers are **time** and **age** the rest of them have outliers. The particular case of outlier is given by the serum creatinine and CPK (Creatinine phosphokinase) values.

1.3 Check for Collinearity

The next step will be to take a look for collinearity, analyzing the correlations between independent variables. In the heart failure data set the column does not require any further transformation. To be able to take a look at multicollinearity properly, we'll use a graph to plot the correlation matrix.

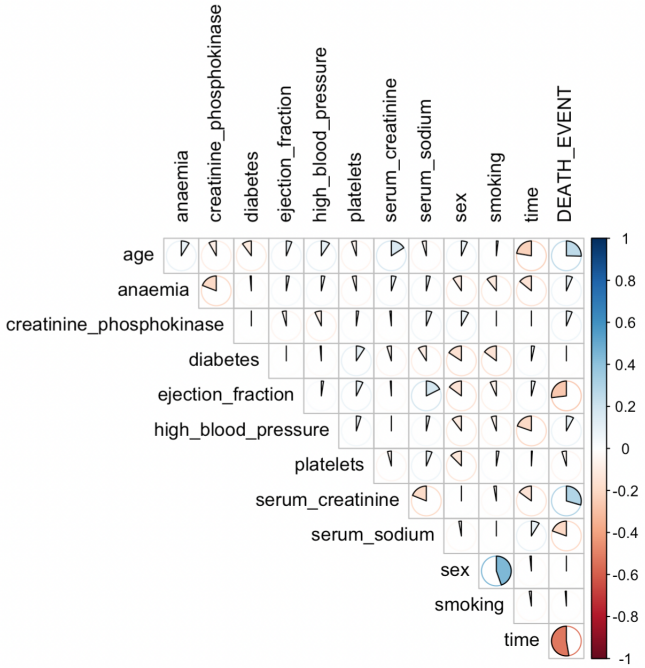

Figura 6: Correlation Matrix

1.4 Data Manipulation

This parts is related to data manipulation procedure in which are performed different operations in order to clean and set the raw data set. It has been performed an adjust operation of platelets where the entire columns has been re sized dividing it for 1000 then it has also been performed a scale operation function. The scaling is a way to compare data that is not measured in the same way.

1.4.1 Chi Square Test

The aim is to investigate if the categorical variables have correlation. To perform this operation the **Chi Square** is used. In this case it allows us to check if categorical variable such as **high blood pressure** and **Death Event** has significant correlation between them.
This test allow us to understand whether both variable in our Table of Chi Square are independent.

	0	1
0	137	57
1	66	39

The P-value > 0.05 allow us to understand we do not reject the null hyphotesis (no relationship exist on the categorical variables meaning they are independent).

2 Logistic Regression

Logistic regression is used to predict the class (or category) of individuals based on one or multiple predictor variables (x). It is used to model a binary outcome, that is a variable, which can have only two possible values: 0 or 1, yes or no, diseased or non-diseased.In the project specific case the algorithm will predict the Death Event which is a numerical variable 0, 1. The dependent variable (Y) is given by the DEATH EVENT (which is binary 0= no , 1= yes)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1848  -0.5706  -0.2401   0.4466   2.6668

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               1.031e+01  5.660e+00   1.822 0.068433 .
age                       4.742e-02  1.580e-02   3.001 0.002690 **
anaemia                  -7.470e-03  3.605e-01  -0.021 0.983467
creatinine_phosphokinase  2.222e-04  1.779e-04   1.249 0.211684
diabetes1                 1.451e-01  3.512e-01   0.413 0.679380
ejection_fraction        -7.666e-02  1.633e-02  -4.695 2.67e-06 ***
high_blood_pressure1     -1.027e-01  3.587e-01  -0.286 0.774688
platelets                -1.200e-06  1.889e-06  -0.635 0.525404
serum_creatinine          6.661e-01  1.815e-01   3.670 0.000242 ***
serum_sodium             -6.698e-02  3.974e-02  -1.686 0.091855 .
sex1                     -5.337e-01  4.139e-01  -1.289 0.197299
smoking1                 -1.349e-02  4.126e-01  -0.033 0.973915
time                     -2.104e-02  3.014e-03  -6.981 2.92e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 375.35  on 298  degrees of freedom
Residual deviance: 219.55  on 286  degrees of freedom
AIC: 245.55
```

Figura 7: Summary Logistic Regression

What we can do about this model is that we're using all the predictors we have but we could also use some other techniques, named a sort of MODEL SELECTION technique in order to select the right number of feature. We're gonna use the backwards STEPWISE regression through the package MASS. The backward stepwise start with the satured model first and then in each step will remove one variable up it reaches a model where all the variable are significant (it works like LASSO and RIDGE penalizing the non significant variable) What the model do ? It will compare the AIC of each model and select the lowest AIC. LOWER value indicates an improved fit for the model based on the data being used.

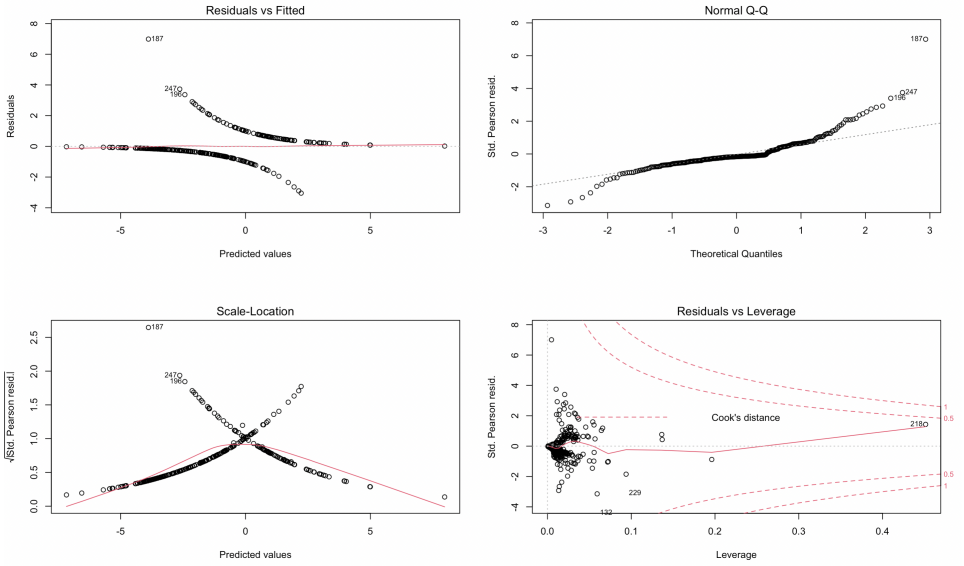

Figura 8: Logistic Regression with StepWise

2.0.1 Standardized and Studentized Residuals

The only difference between the standardized residuals and the studentized residuals is that standardized residuals use the mean square error for the model based on all observations, MSE, while studentized residuals use the mean square error based on the estimated model with the ith observation deleted, MSE(i). To simplify the **Standardize residuals**. Considering residual the difference between an observed value and a predicted value in a regression or other relevant statistical tool. A standardized residual is the residuals divided by an overall standard deviation of the raw residuals. It' important to take a look a at both measure to understand the residuals across different predictor values. Residuals with $<$ *absolute value* of 3 is deemed to be with no significant outlier.

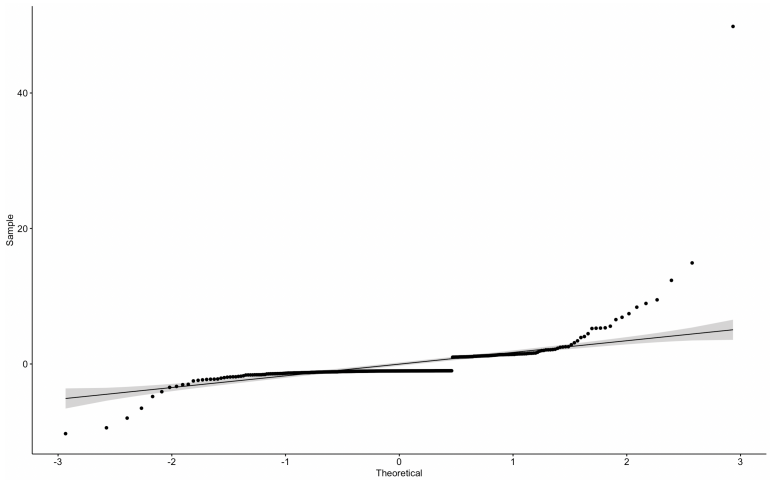

Figura 9: Logistic Step Residuals

In the option of logistic regression it has been selected *Backwards* this because the model start with a fully saturate model. Applying the Stepwise.AIC the best model suggest at the end will be:

```
DEATH.EVENT ~ age + ejection.fraction + serum.creatinine + serum.sodium + time,
family = "binomial")
```

The problem of step AIC is that leads to inflated R square [2] and overfitting. Most of the time is suggested to use classic method in which it will be defined the training and test. ALTERNATIVE WAY to step AIC is the bootstrap resampling. We use the bootstrap re-sampling with replacement method to assess consistency predictors selected with stepwwise.

2.0.2 Boot strapping

In order to have a better measurement of our algorithm and avoid time consuming computation, a proper method is the boot strapping. Going to simply explain it, the computation is composed by 4 steps:

- Make a boot strapped data set
- Calculate the parameter we're interested in
- Keep track of that calculation
- repeat that steps as much as we want

In the specific case of this project, the boot strapping algorithm will take a sample of the data set that it's using, 299 participant in this case and it'll do this a certain amount of times and it'll run this step AIC over and over again on each sample and it's sampling replacement and then what it's going to provide (the AIM) is a diagnostic on the consistency of the variables because there could be variable that are significant in one sample or in another sample. so to summarize the final result from the process will be for each variable in the saturated model how often (in % of times) the predictor is selected. *Notice that*:The boot strapped will be repeated 50 number of times The final model chosen by the boot strap selection is:

```
glm(DEATH_EVENT ~ age+ejection_fraction + serum_creatinine + serum_sodium +
                  time,
                data= heart, family = "binomial")
```

[2]Recalling that is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

2.1 Result

The model contained independent variable as *age*,*ejection fraction*, *serum creatinine*, *serum sodium*, *time* has a plotted summary of the type:

```
glm(formula = DEATH_EVENT ~ age + ejection_fraction + serum_creatinine +
    serum_sodium + time, family = "binomial", data = heart)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1590  -0.5888  -0.2281   0.5144   2.7959

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        9.493034   5.405768   1.756  0.07907 .
age                0.042466   0.015030   2.825  0.00472 **
ejection_fraction -0.073430   0.015785  -4.652 3.29e-06 ***
serum_creatinine   0.685990   0.174044   3.941 8.10e-05 ***
serum_sodium      -0.064557   0.038377  -1.682  0.09254 .
time              -0.020895   0.002916  -7.166 7.74e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 375.35  on 298  degrees of freedom
Residual deviance: 223.49  on 293  degrees of freedom
AIC: 235.49

Number of Fisher Scoring iterations: 6
```

Figura 10: Final Model

From the summary we can notice how the result plotted are the log probability(the estimates in log probability) that are interpreting as for example: age for every one year increase of 0.04 log probability but what I want it's a probability ratio. This problem is solved get the *Return Exponentiated coefficients* and get a *probability Ratio*.
The interpretation of the result is self explained providing the right interpretation under the form of probability. Interpreting the coefficient age= 4.338089e+00.
This means that for every additional year older the probability of death increases by 4.3 percent. By reasoning with the age (independent variable) it can be stated how one year old may just be 4.33 % increase footnoteIf the number in plotted graph was negative you'd be interpreting as decrease.
Therefore, for every 10 year older the odds of death increases by 53% while controlling for all other predictors in the model.

2.2 Check for Multicollinearity

The aim is to check if the independent variable(chosen by the model selection) are not too correlated, in the specific, they are not linear dependent on one other. If we have multicollinearity basically what we'll have is that, in the model there will be an increase in the standard error and a decrease in the reliability of the coefficient estimates.
In order to detect multicollinearity we'll take a look at the **Variables inflation Factor** (VIF[3]) if the VIF is > than 5 this result would suggest high correlation.

age	ejection fraction	serum creatinine	serum sodium	time
1.053111	1.133484	1.079122	1.028355	1.096862

2.3 Test the Model

Firstly the test of the algorithm developed will be done through a generic invented data frame containing only one row of features but with different sensible parameters in order to understand hot the algorithm deployed on that one behaves. The 2 rows of different data-frame are composed like this:

age	ejection fraction	serum creatinine	serum sodium	time
61	38	1.4	136	130

Tabella 2: Trial DF test

age	ejection fraction	serum creatinine	serum sodium	time
95	38	1.4	136	4

Tabella 3: Trial DF 1 test

It can be noticed how the changed features are age and time. Applying the function prediction with the model developed before the probability on the 2 result change sharply. The probability scored on the **Trial DF** is 22.398% of **death event** while on the **Trial DF 1** computed, by increasing the number of years old and decreasing the time the probability of **death event** is equal to 94.44%.

2.4 Train and Test split

In the previous section were built 2 sample Data Frame in order to test the prediction of the model. In this section 2.4 the aim will be to build a stratified random sampling based on the DEATH EVENT to generate train data from 70% of the full data set and test data from 30% of the full data set.
This other approach allow us to understand what happens when the power deployed by the logistic algorithm is applied on the

[3]Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable

2.4.1 Probability and Confusion Matrix

After the model is built, we will use the same data set as a testing data set, and try to predict the probability of DEATH EVENT. After the probabilities are generated, it need to classify the binary events (death or alive) based on the probability 0.5. The (50%) probability is commonly used as the benchmark of the binary events, which means any probability above 50% will be considered as death, vice-versa. From the CONFUSION MATRIX we want to see how the model predict. From the table we can see how the 52 prediction are right and also 23 prediction are right, so we have 75 right classification out of 89. From this score we are able to compute the accuracy of our model, 0.84%.

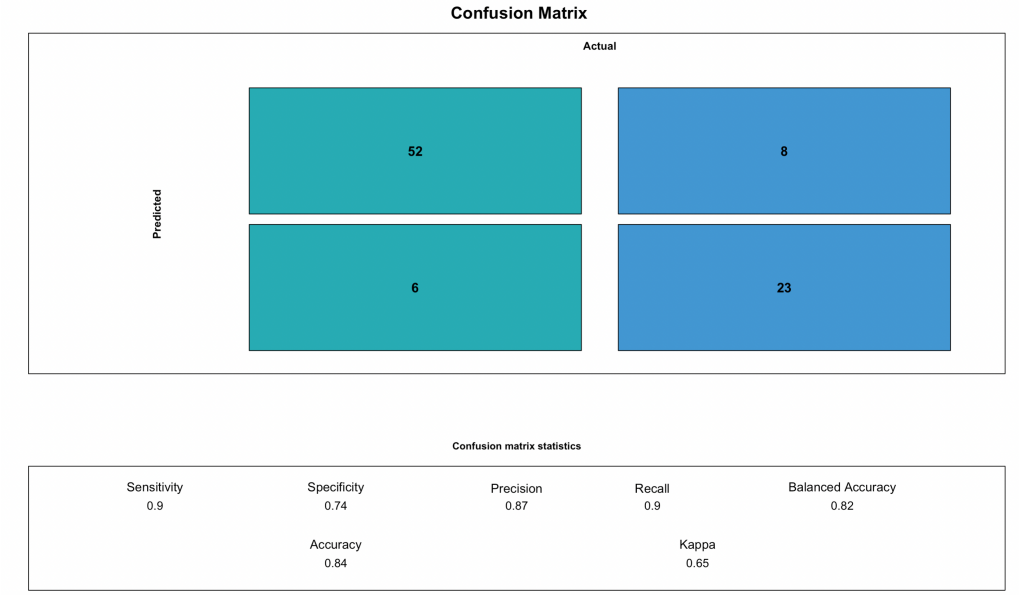

Figura 11: Confusion Matrix

Executing the AUC[4] it's possible to measure the general performance of a binary classifier. In this specific case the **AUC** value is 0.8299. Last but not least through a regression plot it's possible to see how AUC values and the accuracy of this model indicates that the parameters chosen from the stepwise logistic regression model are good in order to predict mortality by heart disease.

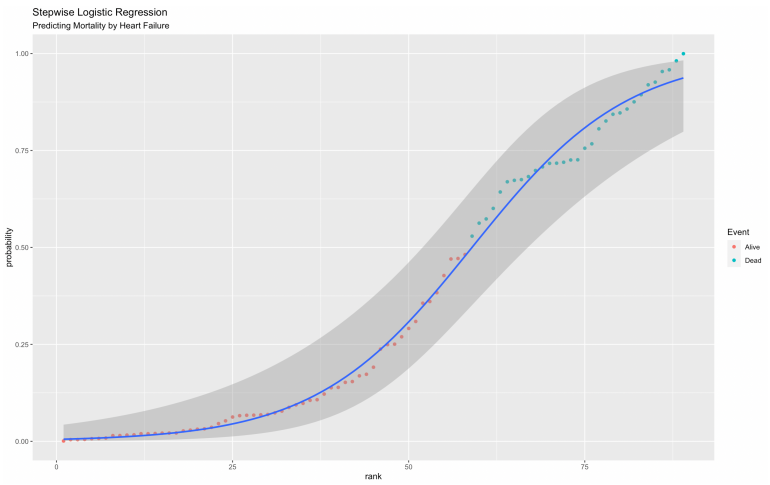

Figura 12: Regression Plot

[4]Area Under the Roc

3 Tree Predictor

The next approach will be focus on one of the most used algorithms in statistical learning, the method is gonna be applied is the Decision Tree Analysis. This one is a non-parametric supervised learning algorithm used for both classification and regression task. In the next section has been provided the explanation process.

3.1 Building the Model

Exploiting the R tool to run the tree predictors, for the composition of the algorithm have been used all the independent features belong to the data set, always paying attention on the train and test split in order to do not take into account the part of data related to the test.

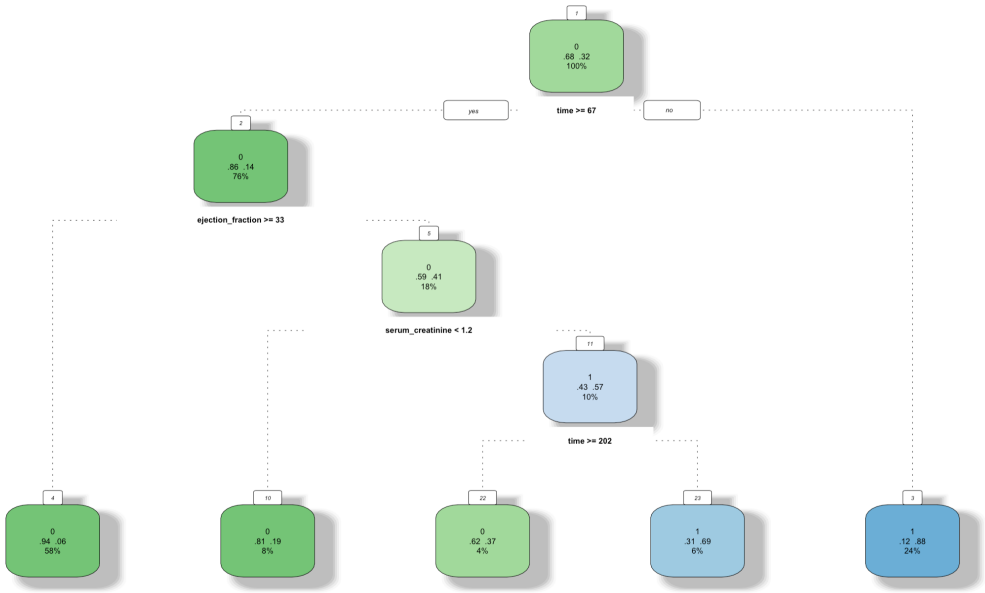

Figura 13: Tree Predictor Structure

It's clear from the graph how, among 12 features, the tree take evaluate as relevant only 3 of them, respectively:

- Time
- Serum Creatinine
- Ejection Fraction

How the features kept from the graph can be evaluated? This effect is given mainly due to highly Independence in the categorical variable followed also by numerical variable with the same mean. The result we stated above can be showed through the computation of the **Analysis of Variance** through the **homogeneity of variance** test. The assumption will be that within the samples under observation they should look like relatively similar.

3.2 Result

The accuracy scored from the tree model is 85,33% In the tree predictor it has been scored the above result with a selected set of variables output by the algorithm while if in the model is implemented the random forest which essentially is a set of trees where the only difference, except the first one already said, is related to the impossibility of directly seen the tree connections through a graph.

4 Random forest

Starting to compute the **Random Forest**

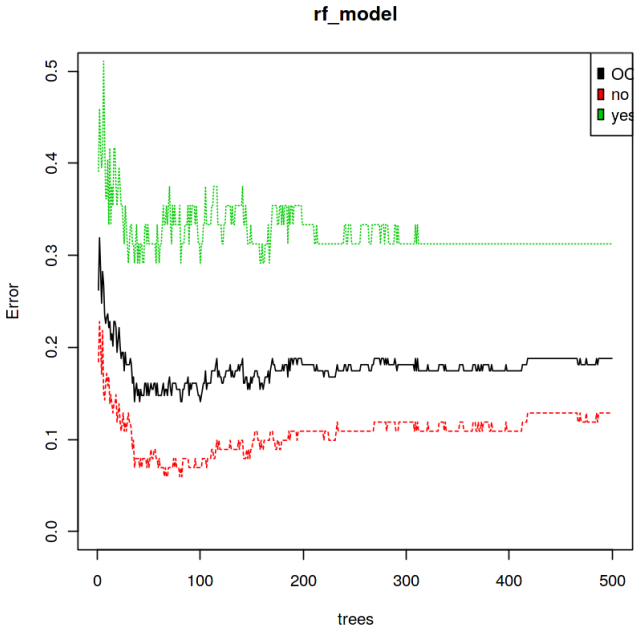

Figura 14: Random Forest Error

it can be seen how as the number of trees increase the error tends to decrease until the line reach the steady state after a few number of trees which seems to be about 50 trees. The situation is stable starting from 100 trees more and it can be seen how the **OOB estimate of error** is 14.77%.
What matter in the evaluation process is to understand once more what are the proportion of the variables mostly used in this Random Forest application. (scale colour gradient add)

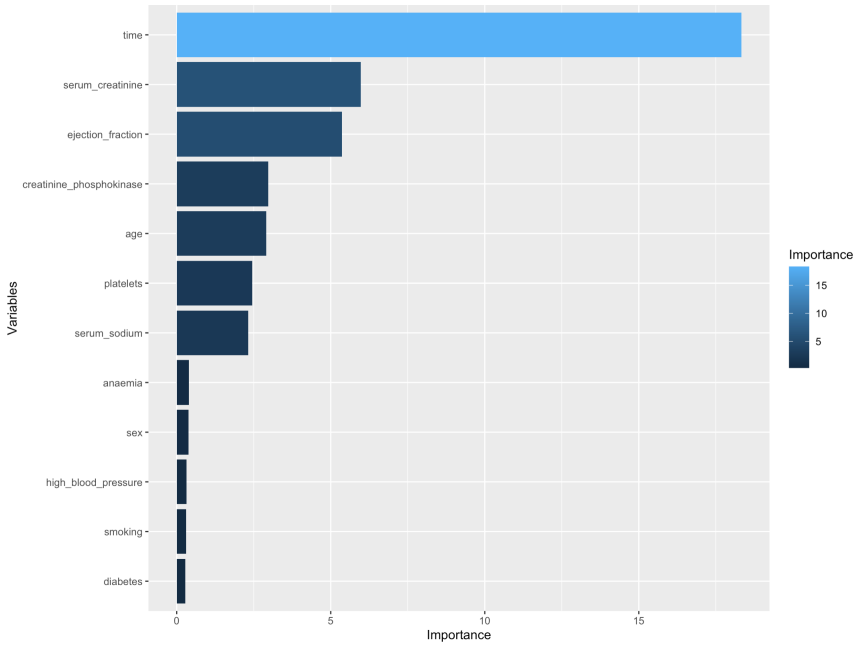

Figura 15: Variable Importance

Also this time as the output of the tree predictors the top 3 important variable are *Time*, *Serum Creatinine*, *Ejection Fraction*.

Unsupervised Learning

Unsupervised learning is a process in which an artificial neural network try to catch for similarities among a certain specific data set. During unsupervised learning, a computer attempts to recognize patterns and structures within the input data on its own. The analysis is not based on a specific prediction of the dependent variable but is bounded on the catch and highlights of cluster and pattern.

4.1 Project Overview

The aim of the analysis is to deep into customers segmentation going to analyse how the features of different clients can affect certain marketing strategies. Clustering the relative customers is a useful practice which could grow the cash-flow or more in general the relative statistics that govern aspect such as: Customer journey, advertising campaign. Going to analyze the feature related to data set we can understand how the latter is composed by 2240 observation from 29 variables. From 29 column , 3 of them are *Character* and 26 are *numeric*.

4.2 Data set Explanation

The data set we're going to analyze is composed by:
Data Summary

Name	marketing
Number of rows	2240
Number of columns	29
_____	
Column type frequency:	
character	3
numeric	26
_____	
Group variables	None

Figura 16: Data Summary

Categorical Variables

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Education	0	1	3	10	0	5	0
Marital_Status	0	1	4	8	0	8	0
Dt_Customer	0	1	10	10	0	663	0

Figura 17: Categorical Variables

Numerical Variables

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
ID	0	1.00	5592.16	3246.66	0	2828.25	5458.5	8427.75	11191	
Year_Birth	0	1.00	1968.81	11.98	1893	1959.00	1970.0	1977.00	1996	
Income	24	0.99	52247.25	25173.08	1730	35303.00	51381.5	68522.00	666666	
Kidhome	0	1.00	0.44	0.54	0	0.00	0.0	1.00	2	
Teenhome	0	1.00	0.51	0.54	0	0.00	0.0	1.00	2	
Recency	0	1.00	49.11	28.96	0	24.00	49.0	74.00	99	
MntWines	0	1.00	303.94	336.60	0	23.75	173.5	504.25	1493	
MntFruits	0	1.00	26.30	39.77	0	1.00	8.0	33.00	199	
MntMeatProducts	0	1.00	166.95	225.72	0	16.00	67.0	232.00	1725	
MntFishProducts	0	1.00	37.53	54.63	0	3.00	12.0	50.00	259	
MntSweetProducts	0	1.00	27.06	41.28	0	1.00	8.0	33.00	263	
MntGoldProds	0	1.00	44.02	52.17	0	9.00	24.0	56.00	362	
NumDealsPurchases	0	1.00	2.33	1.93	0	1.00	2.0	3.00	15	
NumWebPurchases	0	1.00	4.08	2.78	0	2.00	4.0	6.00	27	
NumCatalogPurchases	0	1.00	2.66	2.92	0	0.00	2.0	4.00	28	
NumStorePurchases	0	1.00	5.79	3.25	0	3.00	5.0	8.00	13	
NumWebVisitsMonth	0	1.00	5.32	2.43	0	3.00	6.0	7.00	20	
AcceptedCmp3	0	1.00	0.07	0.26	0	0.00	0.0	0.00	1	
AcceptedCmp4	0	1.00	0.07	0.26	0	0.00	0.0	0.00	1	
AcceptedCmp5	0	1.00	0.07	0.26	0	0.00	0.0	0.00	1	
AcceptedCmp1	0	1.00	0.06	0.25	0	0.00	0.0	0.00	1	
AcceptedCmp2	0	1.00	0.01	0.11	0	0.00	0.0	0.00	1	
Complain	0	1.00	0.01	0.10	0	0.00	0.0	0.00	1	
Z_CostContact	0	1.00	3.00	0.00	3	3.00	3.0	3.00	3	
Z_Revenue	0	1.00	11.00	0.00	11	11.00	11.0	11.00	11	
Response	0	1.00	0.15	0.36	0	0.00	0.0	0.00	1	

Figura 18: Numerical Variables

4.3 Data Exploration

4.3.1 Advertisement Campaign Focus

Deepen into the Accepted variables class we can understand the type of campaign accepted or rejected by each customers. This means that there will be 6 **Marketing Campaign** types where 1 indicates the value has been accepted and 0 indicates the campaign has been rejected (from customer point of view).

- Marketing Campaign 1 (*AcceptedCmp1*)

Accepted	Rejected
1	0

- Marketing Campaign 2 (*AcceptedCmp2*)

Accepted	Rejected
1	0

- Marketing Campaign 3 (*AcceptedCmp3*)

Accepted	Rejected
1	0

- Marketing Campaign 4 (*AcceptedCmp4*)

Accepted	Rejected
1	0

- Marketing Campaign 5 (*AcceptedCmp5*)

Accepted	Rejected
1	0

4.3.2 Amount Spent on different product

Deep into the analysis of all the goods purchased by the customers, understanding the type and the price products.

- **Mntwines**: Amount Spent On Wine In Last 2 Years
- **MntFruits**: Amount Spent On Fruits In Last 2 Years
- **Mntmeatproducts**: Amount Spent On Meat In Last 2 Years
- **Mntfishproducts**: Amount Spent On Fish In Last 2 Years
- **Mntsweetproducts**: Amount Spent On Sweets In Last 2 Years
- **Mntgoldprods**: Amount Spent On Gold In Last 2 Years

4.3.3 Platform usage

- **NumDealsPurchases**: Number Of Purchases Made Through The Company's Website
- **NumWebPurchases**: Number Of Purchases Made Using A Catalogue
- **NumCatalogPurchases**: Number Of Purchases Made Directly In Stores
- **NumStorePurchases**: Amount Spent On Fish In Last 2 Years
- **NumWebVisitsMonth**: Number Of Visits To Company's Website In The Last Month

4.3.4 Customers Data

- **Id**: Customer's Unique Identifier
- **Year birth**: Customer's Birth Year
- **Education**: Customer's Education Level
- **Marital status**: Customer's Marital Status
- **Income**: Customer's Yearly Household Income
- **Kidhome**: Number Of Children In Customer's Household
- **Teenhome**: Number Of Teenagers In Customer's Household
- **Dt customer**: Date Of Customer's Enrollment With The Company
- **Recency**: Number Of Days Since Customer's Last Purchase
- **Complain**: 1 If The Customer Complained In The Last 2 Years, 0 Otherwise

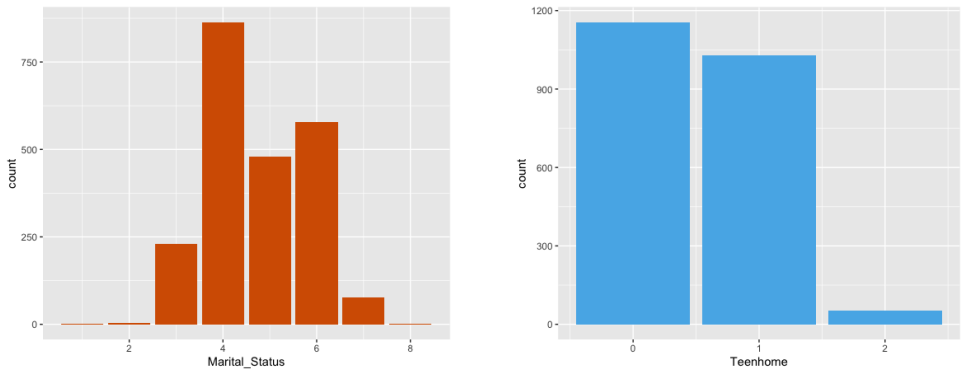

Figura 19: Teen home and Marital Status statistics

From the analysis of the Income variable it can be seen an outlier, its value is equal to 66666. The outlier shown will be changed as 0 because the main assumption will be,every person in the gap will have income 0 and only a few person has value equal to 6666.
Marital status deep: 1= absurd , 2= alone, 3= divorced, 4= married, 5= single, 6= together, 7=Widow, 8=yolo
Education deep: 1= 2n cycle, 2=Basic, 3= graduation, 4= master, 5= PhD

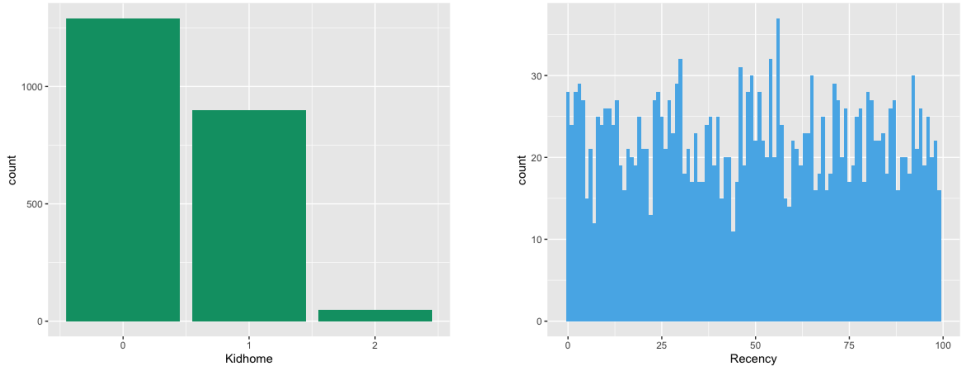

Figura 20: Kid home and Recency Statistics

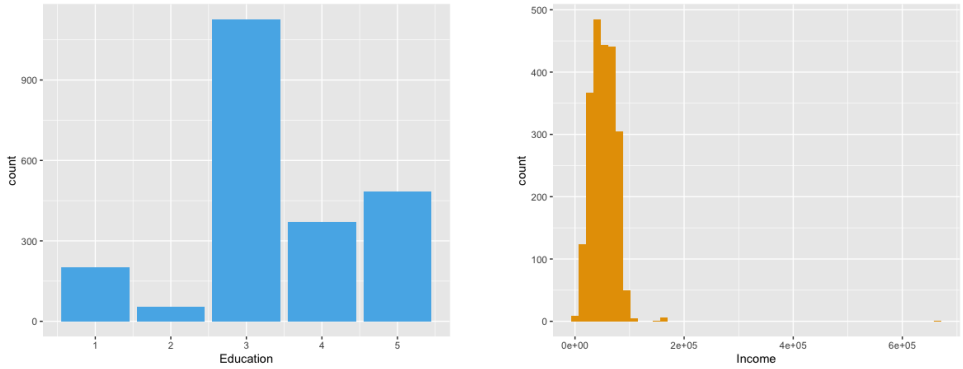

Figura 21: Education and Income Variable

4.3.5 Converting the Variable

The variable *Dt Customer* will be changed in factor due to its form. Originally this one is expressed as character and it has been followed by a transformation process. This will then be numerically encoded into 1,2 and 3. The variable *Year Birth* contains some outlier such as the data 1893,1899 and 1900 seem unrealistic. For this reason they will be removed leaving only the birth date starting from 1940. The next step will be to encode the remaining categorical

```
##
## 1893 1899 1900 1940 1941 1943 1944 1945 1946 1947 1948 1949 1950 1951 1952 1953
##    1    1    1    1    1    7    7    8   16   16   21   30   29   43   52   35
## 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969
##   50   49   55   43   53   51   49   36   44   45   42   74   50   44   51   71
## 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985
##   77   87   79   74   69   83   89   52   77   53   39   39   45   42   38   32
## 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996
##   42   27   29   30   18   15   13    5    3    5    2
```

Figura 22: Caption

variable into numeric, the converted variable will be:

- **Education**
- **Marital Status**
- **Education**

5 Principal Components Analysis

Principal Components analysis is a way to print screen our data. It's a method to find the direction of our data set that describe most of the variability, this explained is the principal component 1 then there will be principal component 2 and so on. Always remember how the direction of the principal component is orthogonal and perpendicular with respect to the former .
First problem to solve is the column variance equal to zero. You can check which column of a data frame is constant this way. Once the 1st problem is solved there will be **NA removing** process.
The **PRINCIPAL COMPONENT ANALYSIS** will be ran with the *Prcomp* command.

5.1 Scree Plot

The scree plot indicates how much variation each principal component captures from the data. From the graphs it's clear how the first 8 component allow us to explain most of the variance (the variance is cumulative). Ideally the variance explained should not be less than 60%.

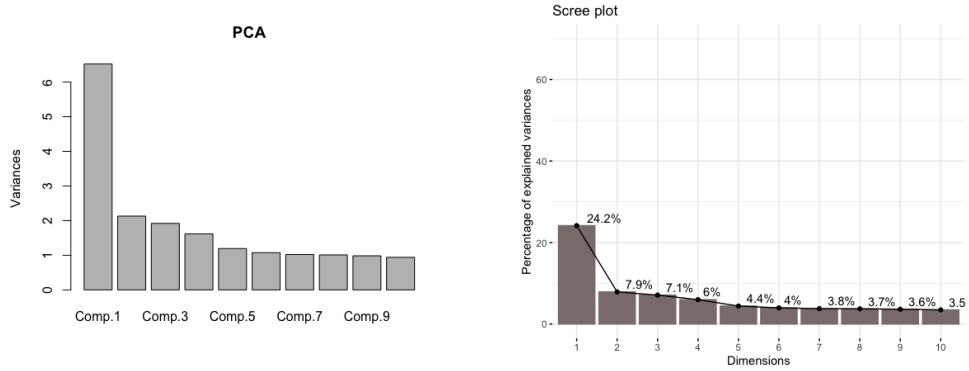

Figura 23: Different view of Scree Plot

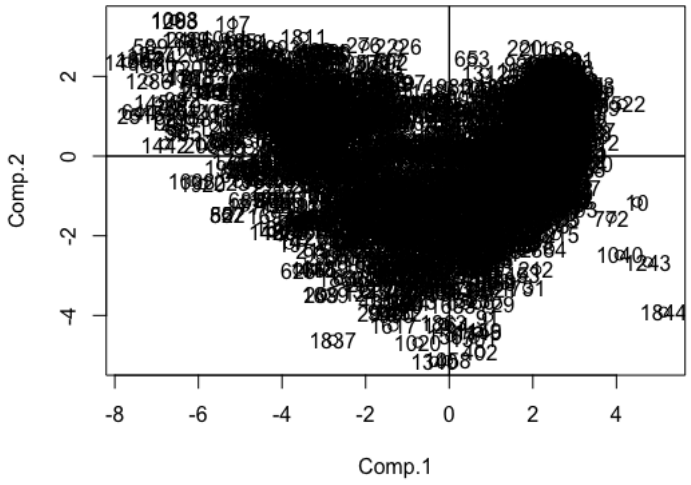

Figura 24: Score plot

Thus score plots allow us to rapidly locate similar observations, clusters, outliers and time-based patterns.

Points close the average appear at the origin of the score plot. An observation that is at the mean value for all k-variables will have a score vector

$$Zi = [0, 0, ..., 0] \tag{1}$$

5.2 Loading Plot

The loading plot is a plot of the direction vectors that define the model. They show how the original variables contribute to creating the principal component.

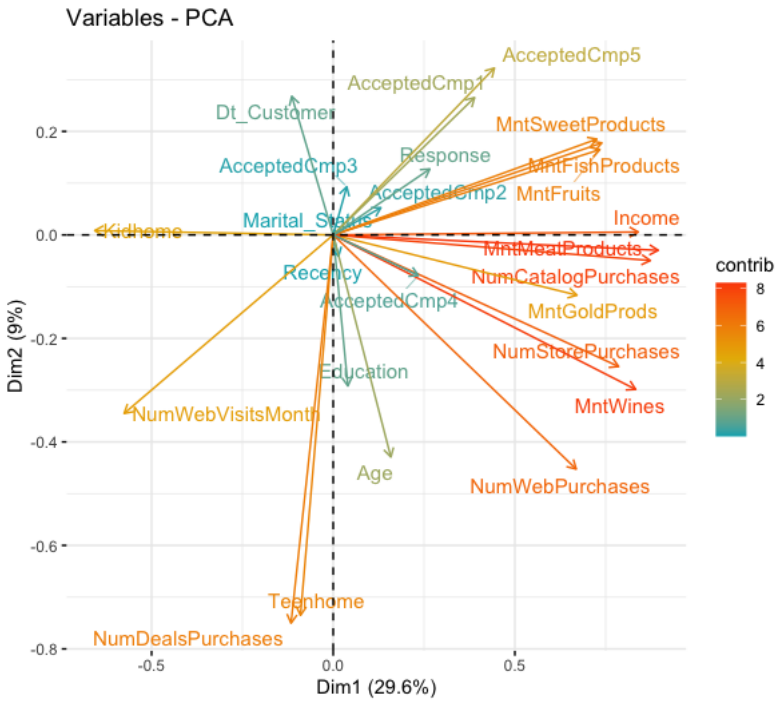

Figura 25: Loading Plot

Recall to guide lines:

- **Direction**:The more parallel to a PC axis is a vector, the more it contributes only to that PC (in this case PC1 and PC2).
- **Length**:The longer the vector, the more variability of this variable is represented by the first two components. Short vectors are thus better represented in other dimension.
- **Angle**:The angle between vectors of the variables show their correlation in this space. To be clear: small angles represent high positive correlation, right angles represent lack of correlation, opposite angles represent high negative correlation.

What stated for the direction it also provided by the heat map on the right hand side. Just to make an example it can be noticed how variable such as *Num Catalog purchase* , *Num Store Purchases* and *Mnt Wines* have long length and they are parallel to 2nd dimension. It can be seen how the variable *Income* form an angle of 180° with *Kid home* this means that they are inverse correlated.

5.3 Bi-Plot

The same information plotted above can be also plotted with the **Bi-Plot** even if the union of the graphs above is not clear.

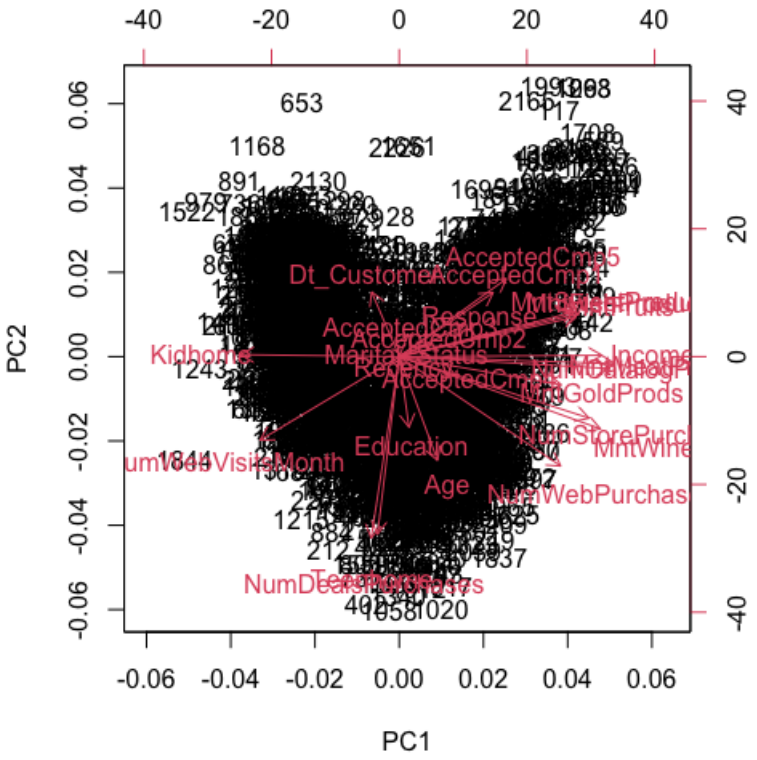

Figura 26: Bi-Plot

5.4 Contribution of the Variables

We're going to analyze the contribution of each variables to each dimension understand how this contribution become consistently low every time we increase the dimension.

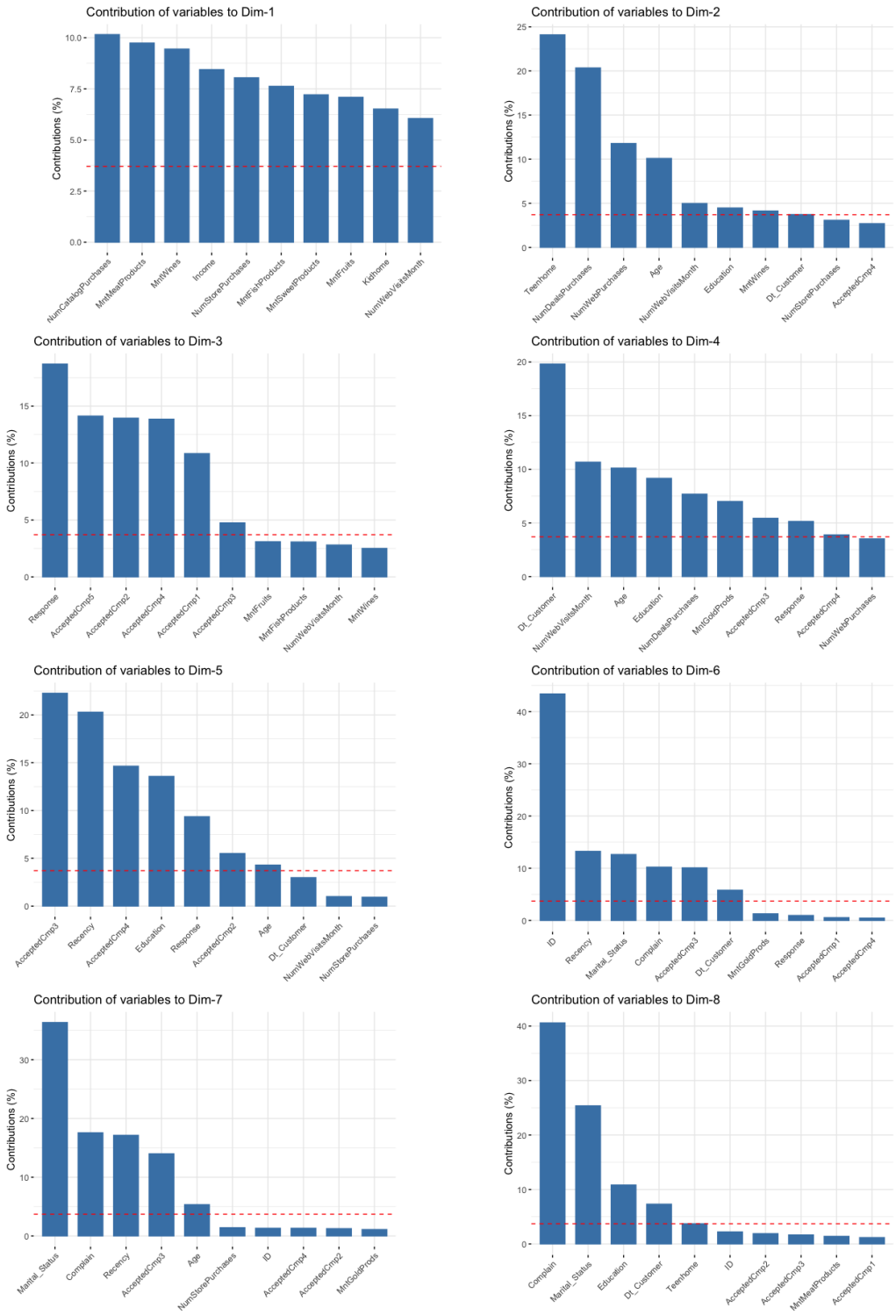

Figura 27: Contribution of variables to each Dimension

6 K-Mean Clustering

The k-means clustering method is an unsupervised machine learning technique used to identify clusters of data objects in a data set. The hyphotesis behind the model in order to choose the parameter relying on the K estimation which is based on the **Elbow approach**. The empirical approach is ran as follow: Calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS becomes first starts to diminish.In the plot of WSS-versus-k, this is visible as an elbow.
In this specific application of K-Mean due to the fact the result from the scale Marketing data set led to inconclusive result due to miss classification of Cluster I choose the variable **process_pca** to run the the K mean function. The *process_pca* variable allow to convert numerical data in one or more principal components and in this case the number of principal components equal to 8. Even if the difference between k-mean and PCA is clear there is a a way in which can be applied the data set with the result of PCA directly into the K mean formula. In short, using PCA before K-means clustering reduces dimensions and decrease computation cost. On the other hand, its performance depends on the distribution of a data set and the correlation of features.So if we need to cluster data based on many features (like in this specific case), using PCA before clustering is very reasonable. ! *]40

6.1 K-Estimation

From the Figure 28 plotted below it can be seen how the number of clusters the model has going to implement is equal to 5.

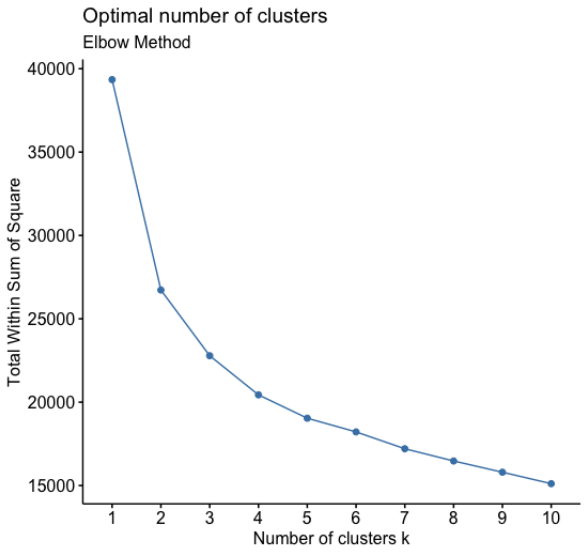

Figura 28: Elbow K estimation

6.2 Plotted cluster

In the plotted graph below there are 2 way in which the K-means clustering can be done. The first aim will be to apply the K-mean clustering using as input the 8 Principal components earned from PCA. In the second method the aim will be to give a comparison applying K-mean with the entire input data set just scaling the feature. Just to give a comparison of how is the graph without this assumption on the input data based on PCA. Using scaled data only without (PCA assumption) we can see how the percentage given by the PC1 is half 24% compared to the model above in which the percentage is equal to 42%.

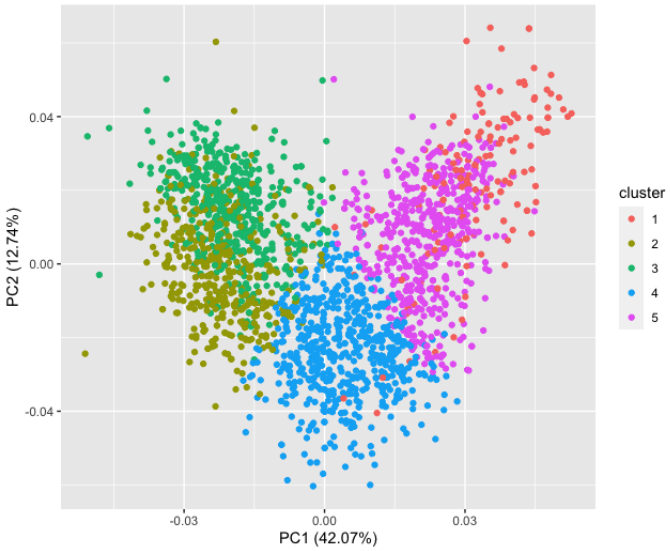

Figura 29: K means Clustering with 8 PC

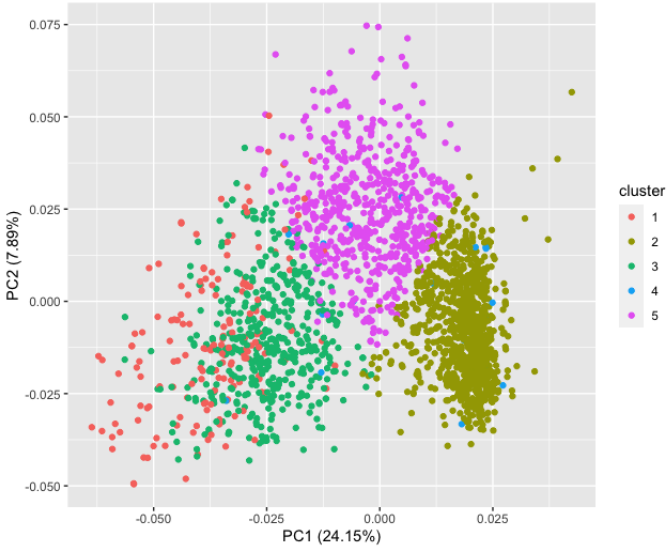

Figura 30: K means with scaled data

6.3 Inspection of the Cluster

From here it's good to go deep into the data which represent the cluster and understand what is the logic of clustering applied by the model. Each plotted graph on usually on the x axis will have the 5 clusters analyzed.

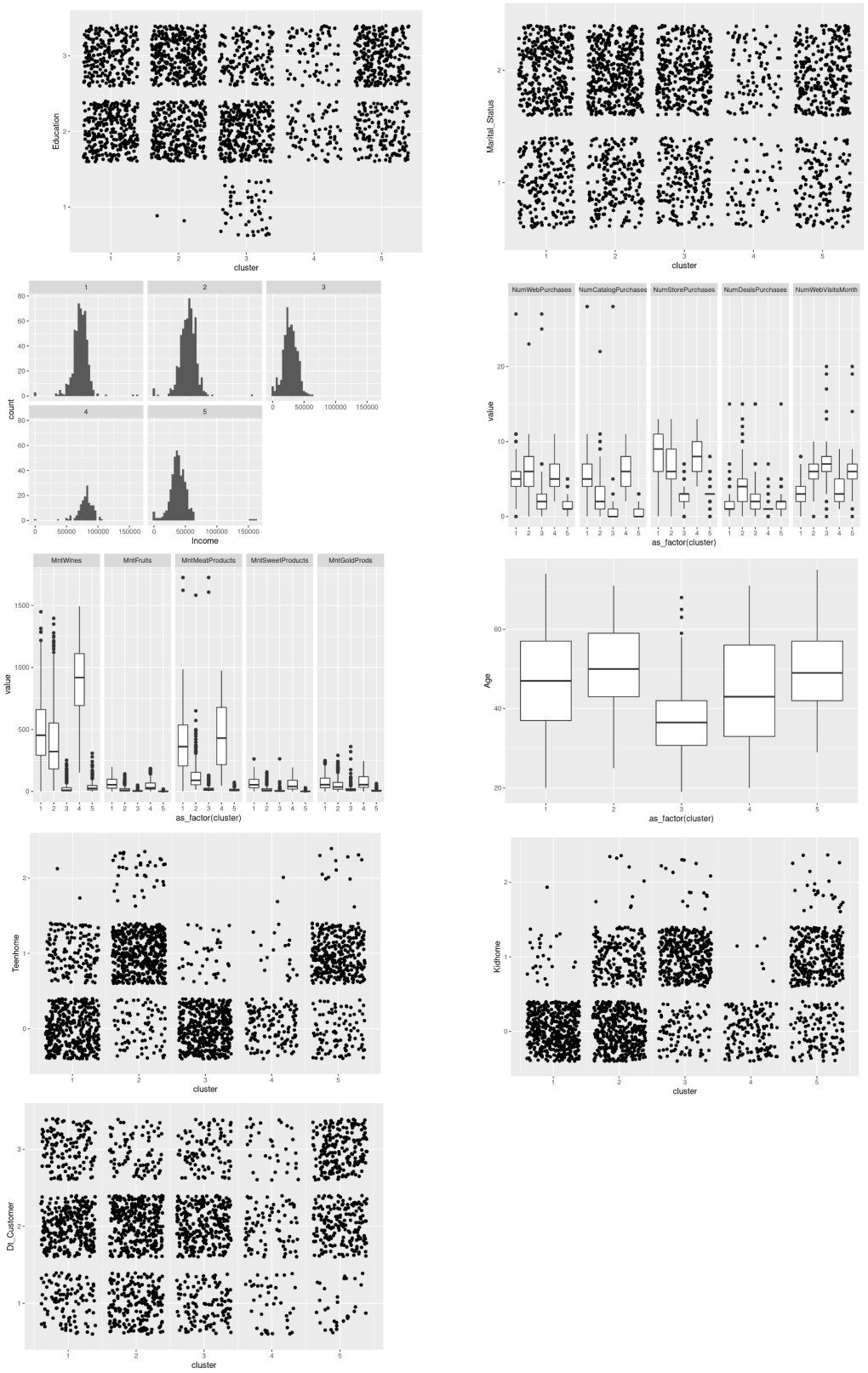

Figura 31: Analysis of variables for each cluster

6.4 Resume of all the clusters

From figure 31 it's possible to compile the clusters by looking at the difference between each of them.

6.4.1 Cluster 1

This is the clusters of people with highest level of education among all the clusters analyzed. The cluster is **High income** group. In this type of result we're clustering individual with high income and not to many young dependents. From the analysis of each segment we can understand how the cluster 1 prefer to purchase via store and catalogs and they are low web addicted. Regardless the type of platform they like to buy all goods types.

Variable	Values
Education	Most of the individuals with specialization and graduation. [None with basic education]
Income	High
Web Purchases	Medium
Catalog Purchases	High
Store Purchases	High
Deal Purchases	Low
Web visits	Low
Amount on Wines	Medium
Amount on Fruits	High
Amount on Meats	High
Amount on Sweets	High
Amount on Gold	High
Advertisement Campaigns	High response to 5th and last campaign
Age	Mid 40s
TeenHome	Some have 1 teen at home.
KidHome	Most of them don't have a kid at home

6.4.2 Cluster 2

In this cluster the people with the graduation and specialization are almost the same of the cluster 1 but now this time we have to add also people with basic education. This cluster prefer to buy online and they represent the **mid income** group with more teens than kids at home.

Variable	Values
Education	Most of the individuals with specialization and graduation + people with basic education
Income	Mid value
Web Purchases	High
Catalog Purchases	Low
Store Purchases	medium
Deal Purchases	High
Web visits	High
Amount on Wines	Medium
Amount on Fruits	Low
Amount on Meats	Medium
Amount on Sweets	Low
Amount on Gold	Medium
Advertisement Campaigns	High response to 4th and last campaign
Age	Early 50s
TeenHome	Most have at least 1 teen at home
KidHome	Some have kids at home

6.4.3 Cluster 3

This cluster represent the **low income** one with low purchase on the entire kinds of products and high visits on web stores. The interesting data is that despite they have low income they are also the youngest among the cluster analyzed where they don't have any teens at home, but they do have at least a kid at home.

Variable	Values
Education	specialization low with respect to graduation. Individuals with basic education are included only in this category
Income	Low value
Web Purchases	Low
Catalog Purchases	Low
Store Purchases	Low
Deal Purchases	Medium
Web visits	High
Amount on Wines	Low
Amount on Fruits	Low
Amount on Meats	Low
Amount on Sweets	Low
Amount on Gold	Low
Advertisement Campaigns	response only to 3th and last campaign
Age	Below 40
TeenHome	Most have don't have teen at home
KidHome	Most have at least a kid at home

6.4.4 Cluster 4

This cluster represent the **highest income on average** one , they are the early 40's population which don't have both kids and teen at home. Their level of purchase of Wines, Meet and Gold respectively, is high and they give a full response to all the marketing campaigns proposed.

Variable	Values
Education	equal number with specialization and graduation
Income	highest average income among all the group
Web Purchases	Medium
Catalog Purchases	High
Store Purchases	High
Deal Purchases	Low
Web visits	Low
Amount on Wines	High
Amount on Fruits	Medium
Amount on Meats	Medium
Amount on Sweets	Medium
Amount on Gold	High
Advertisement Campaigns	response to all the campaign,from 1th up to last campaign
Age	Early 40s
TeenHome	Most have don't have teen at home
KidHome	Most dont have a kid at home

6.4.5 Cluster 5

This cluster represent the **average income group** , they are the late 40's population which have at least one kids and 1 teen at home. Their level of purchase of Wines, Meet and Gold respectively, is low and they give a respond to certain specific marketing campaigns proposed.

Variable	Values
Education	specialization relatively lower compared to graduation.
Income	average income group
Web Purchases	Low
Catalog Purchases	Low
Store Purchases	Low
Deal Purchases	Low
Web visits	High
Amount on Wines	Low
Amount on Fruits	Low
Amount on Meats	Low
Amount on Sweets	Low
Amount on Gold	Low
Advertisement Campaigns	response only to campaign 3rd, 4th and last campaign
Age	Late 40s
TeenHome	Most have at least 1 teen at home
KidHome	Most have a kid at home.

6.5 Conclusion

We can notice how the advertisement campaign earn power at the final stage with moderate and low income people while they work very well with high income people. From this clusters it's possible to understand how the people, depending on their income, prefer to visit and purchase online or in store. In particular high income prefer in store whereas low/medium income prefer to buy online. It can be also seen how people with high income consume more certain types of goods with respect to low income one.