Zulato Davide mat. 876101

July 8, 2022

ECONOMICS FOR DATA SCIENCE M

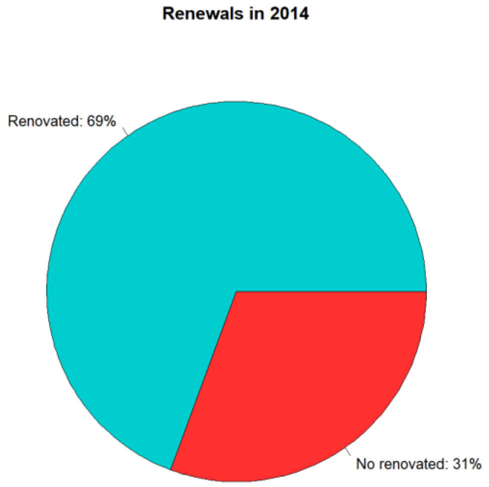
In this work I consider the datasets data1, an13, and in13 containing information on users of a museum card. Each user bought a museum card, which allow them to enter in any museum for free.

1) ISSUE 1

Describe the most interesting variables by plotting distributions, correlations, co-occurrence

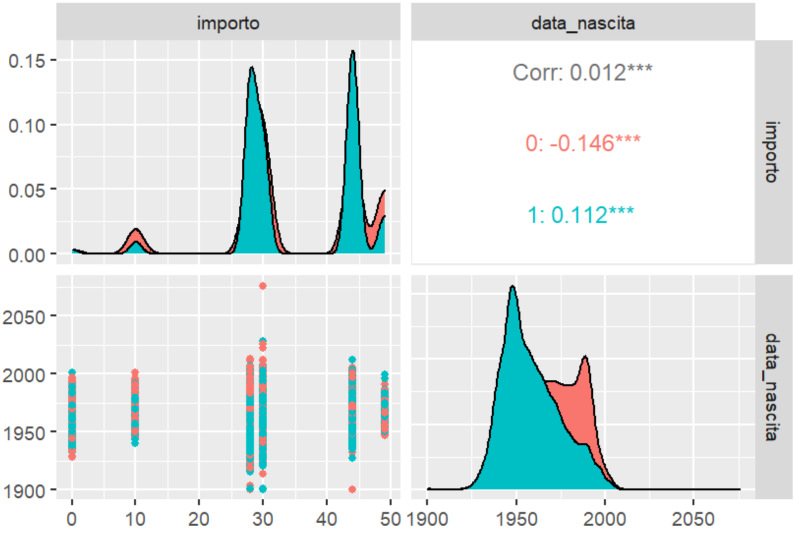
The first interesting variable I should care is the churning rate, in the variable "Si2014" (label 0 for churn) 31% of the users of the museum card didn't renew the card while 69% did.

Figure 1



The next plot (*Fig.2*) provides a correlation plot and the univariate distribution by churn (0=churn, 1=renewal) for the variables price paid and date of birth for the 80,140 clients in the dataset. The correlation coefficient is statistically significant (at 1% level) but from this ρ (0.012) I cannot assess a substantial significance of the effect of being a churner or not.

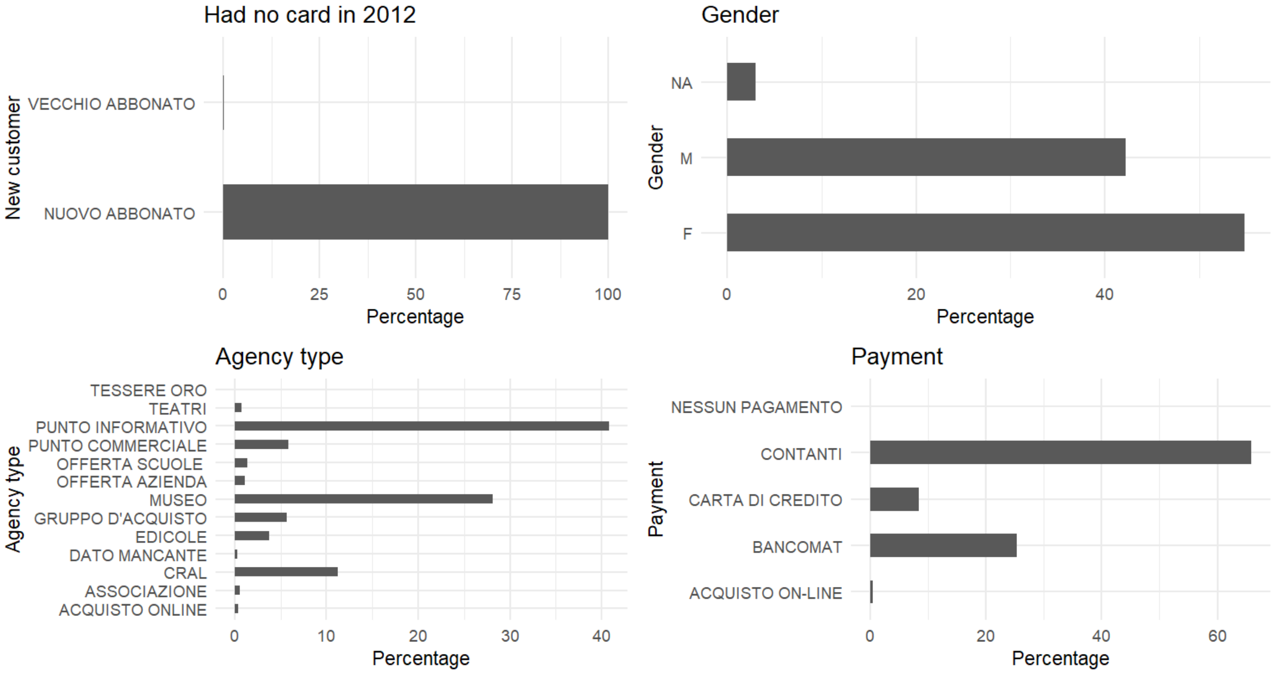
Figure 2



0=CHURN 1=RENEWAL

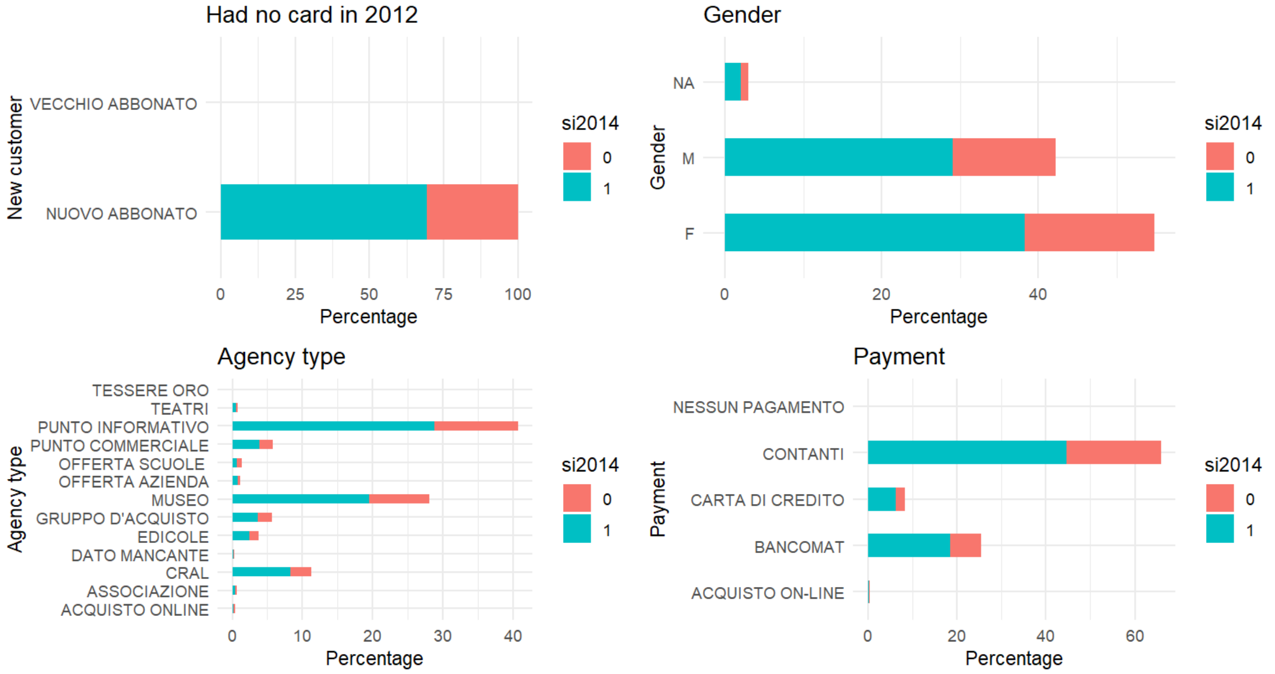
From the plot of the distribution, we can see that there are some outliers for the age (/birth date); I'll deal with it later. The correlation between price paid and birth date is statistically significant but weak, it is interesting to observe the change of sign for the churners and not churners (positive for renewals, negative for churners)

Figure 3



This bar plot provides a distribution for some interesting categorical variables, about 35% of the clients has not discount and 45% has a discount coming from the renewal of the membership. The "NUOVO ABBONATO" in the new customer variable are those who had no card in 2012, only 41 people in this dataset had a card in 2012. Most of the customers paid in cash (more than 60%).

Figure 4



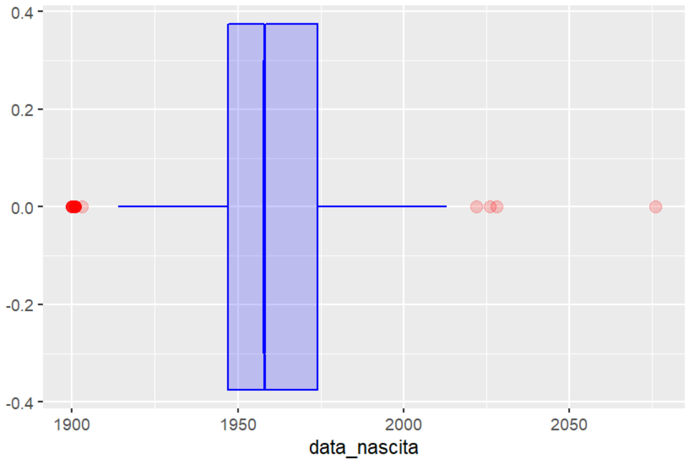
It is interesting to observe a higher proportion of churners for the Organization type “OFFERTA SCUOLE”, I’ll introduce a dummy variable = 1 if the individual belongs to this organization.

2) ISSUE 2

Do you spot some problem with the variables? Are there any specific problems you should take care about?

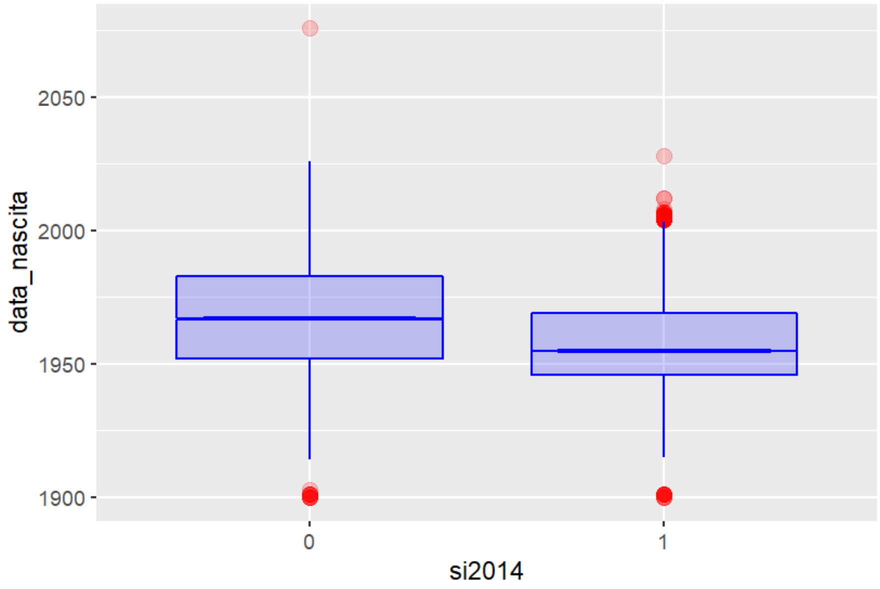
The next figure displays the boxplot of the birth date for the 80,140 clients: the outliers (>1.5 IQR) are marked in red

Figure 5



Looking at the boxplots I noticed some outliers to be removed or imputed, the next figure (Fig.6) shows the boxplot of birth date by Renewals and churners, the churners are in median younger than the retained

Figure 6



Remember that 0 indicates churn and 1 indicates renewal. Since the

in 2014, according to ISTAT, there were 18 people over 110 y.o. (18 all-female). I assume that people born before 1901 are mistakes in the data entry, the same for the people born after. I impute the mean of the response variable (si2014, churn=0, renewal=1) because as we have seen in Fig.2 the distribution of the age changes over the variable churn.

For the NAs in the gender (see them in the bar plot on top right in figure 3/4) I assigned randomly a gender to the customers to don't have missing values in ISSUE6.

3) ISSUE 3

Analyse the pattern of missing values. Is there any variable you should drop from the analysis?

Looking at the pattern of the missing values (*Fig.2*) I would drop the variable "Job" (*professione*) because it is a column of NAs, from the following graphical representation we can see, for each dataset, the distribution of missing values for every variable in each dataset (data1, an13, in13 and the merged dataset[1]).

[1] It includes all rows in dataset data1 such that every row is a client id and we get information also from an13 and in13

This plot (*Fig.7*) represents the distribution of NAs in the dataset “data1”

Figure 7 Data1

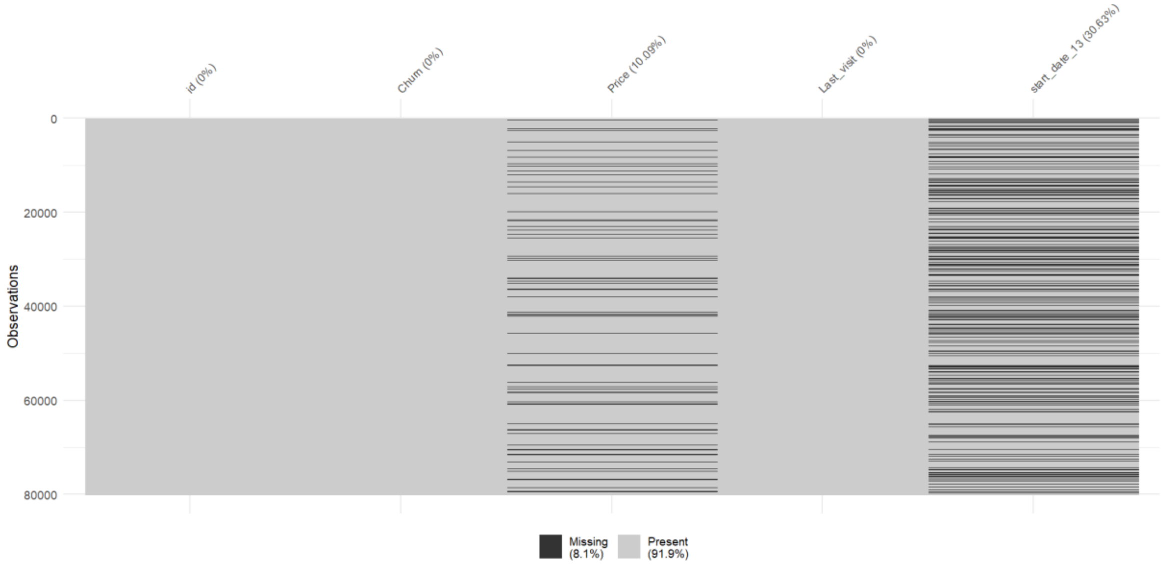
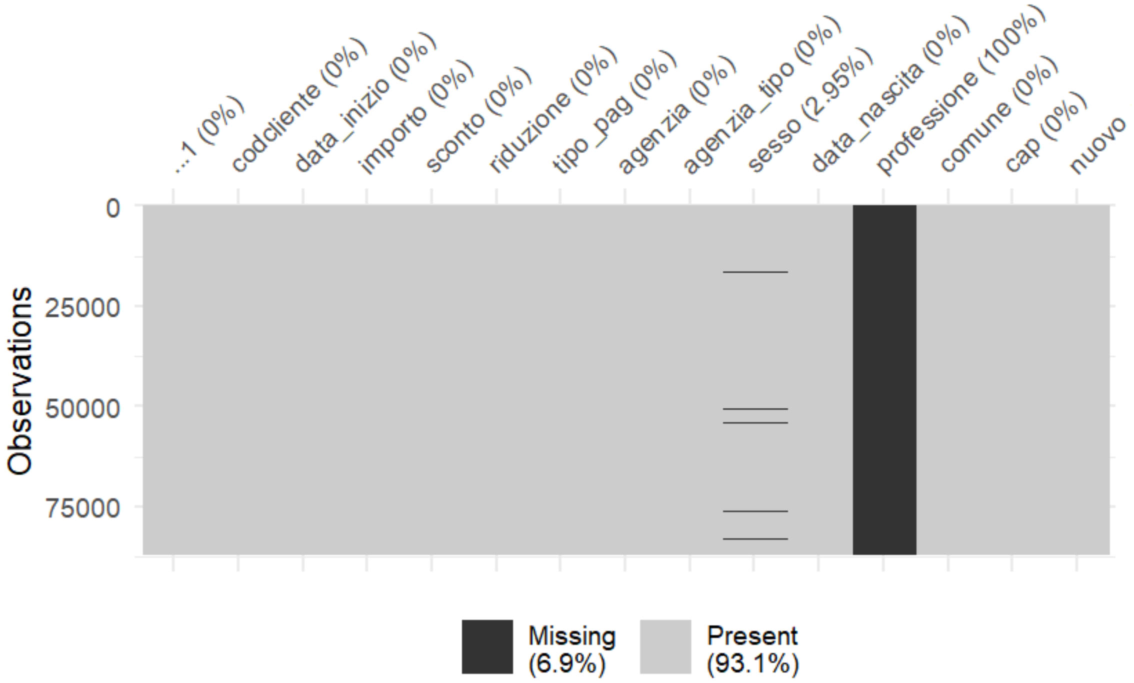


Figure 8 an13

The following plot represents the distribution of the NAs in the dataset an13 across the variables, we can delete the variable “*professione*”

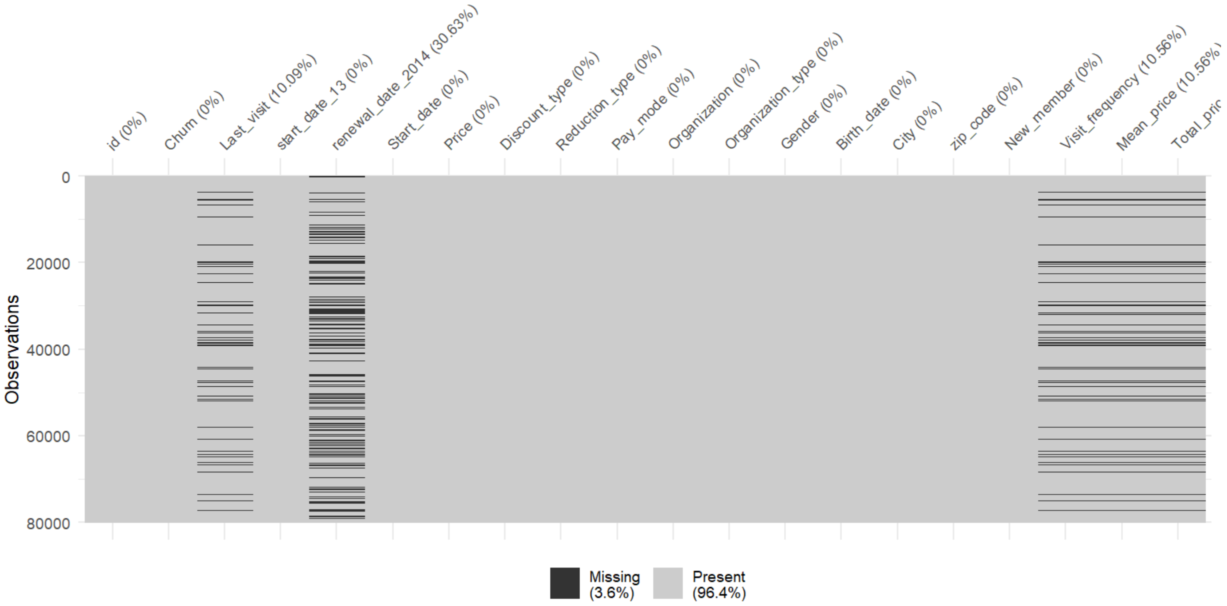
Figure 9



There are no NAs in the dataset in13

Next plot concerns the merged Dataset, the NAs in gender has been randomly replaced and those in birth date have been replaced with the mean between churners and renewals

Figure 10 Merged Dataset before Processing



At the End, for the next ISSUES I created a merged dataset with also the new variable created in ISSUE 5 (Degree), the frequency of visits (counting the number of visits by client) and the total price spent, summing up the Prices and the mean price (later deleted for avoid multicollinearity). In ISSUE 5 I'll also create the Degree variable as output of the Network.

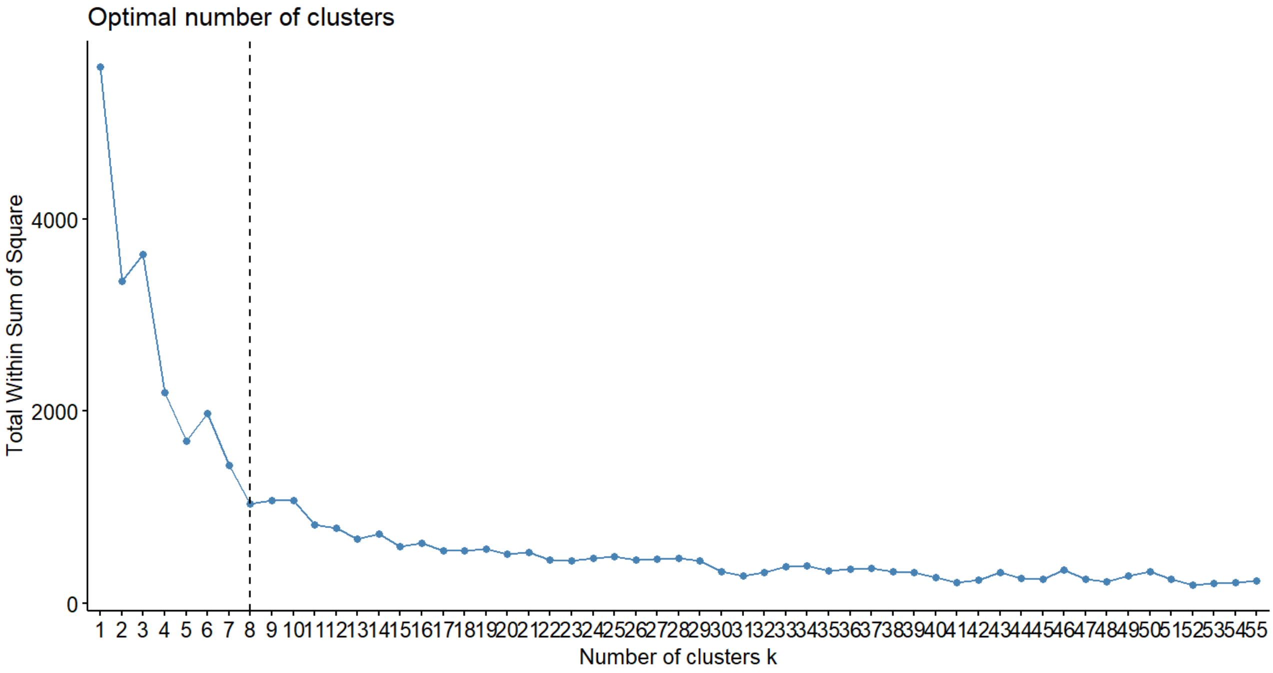
In conclusion I deleted the profession column, replaced the missing Genders randomly, deleted Renewal date because it has a similar distribution to Churn (Si2014) and for the birth dates I replaced the impossible values with the group mean for retained and churners

4) ISSUE 4

Can you cluster the observations? Is there a cluster with most churners?

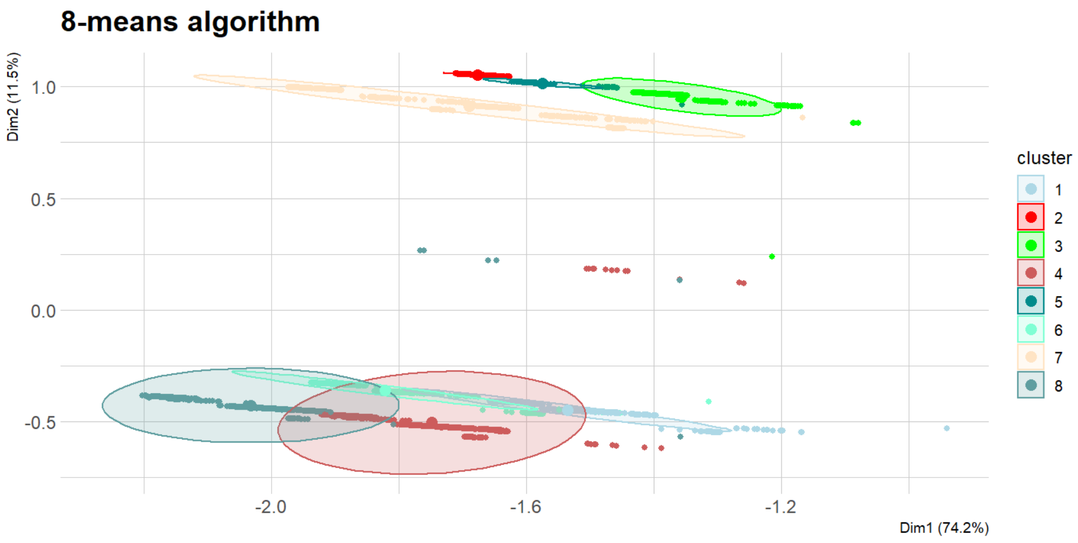
For this issue I used first the classical two-step approach with dimensional reduction and k-means and then the kohonen map to obtain an easy-to use map and trying to interpret the results. We are in an unsupervised environment; it means that I must choose the number of centroids and the dimension of the Kohonen map. Looking at the marginal distribution of the covariates I cannot assume normality, so I expect better and mor interpretable results with the Kohonen map. For Example, as noticed in Fig.2 the distribution of the price variable is not symmetric and it is far from normality.

Figure 11



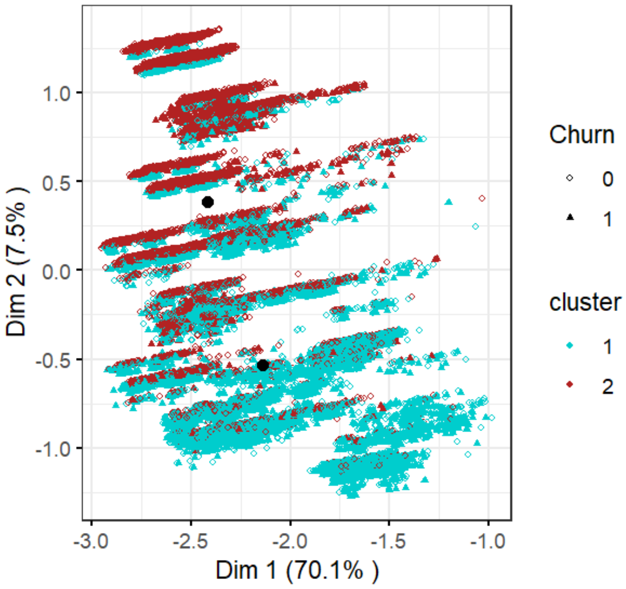
Using the criterion of minimum Total Within Sum of Square the optimal number of clusters would be too high (around 50 centroids) for a nice visualization; so, I looked for some elbows like the one with k=8. Also, k=2 and k=5 looks feasible. The total sum of squares with 8 clusters is 5169.051, the Betweenness is 4628.666.

Figure 12



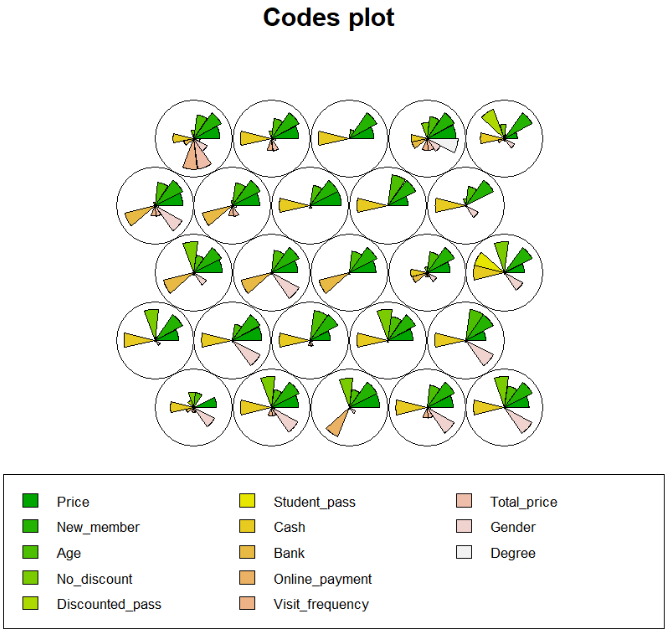
With 8 Centroids I am not able to see a clear separation of the Clusters; in order to cluster the observations, categorical variables were converted in dummies. Once scaled the numeric variables and applied a PCA I tried also with a 2-means algorithm to identify a cluster with more churners (with 8/9 clusters an interpretable visualization was not viable). The figure shows the two centroids (black dots) and the two clusters, one in red and one in Cyan. Also, from this representation I cannot see a clear separation and identify a cluster with most churners, the churners in the figure are identified with a circle and the retained as a triangle.

Figure 13



Taking account af the problems related to the variables (Dummies and violation of normality assumption) I proceed with an SOM map aiming to get interpretable results.

After Some Trials I decided to draw a 5x5[2] (the dimension of the map is up to the User) Kohonen map with hexagonal shape.

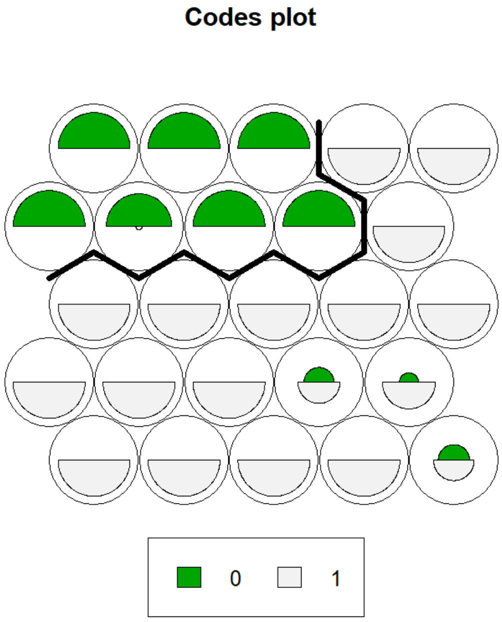


The Red black line in the next figure (*Fig.14*) indicates the separation of the two classes; I can do that because the ground truth is provided by the variable Churn (si2014).

[2] For example, with a bigger map I get too many empty cells

Another kind of visualization for the binary classification task.

Figure 14 0 = Churn 1= No-Churn



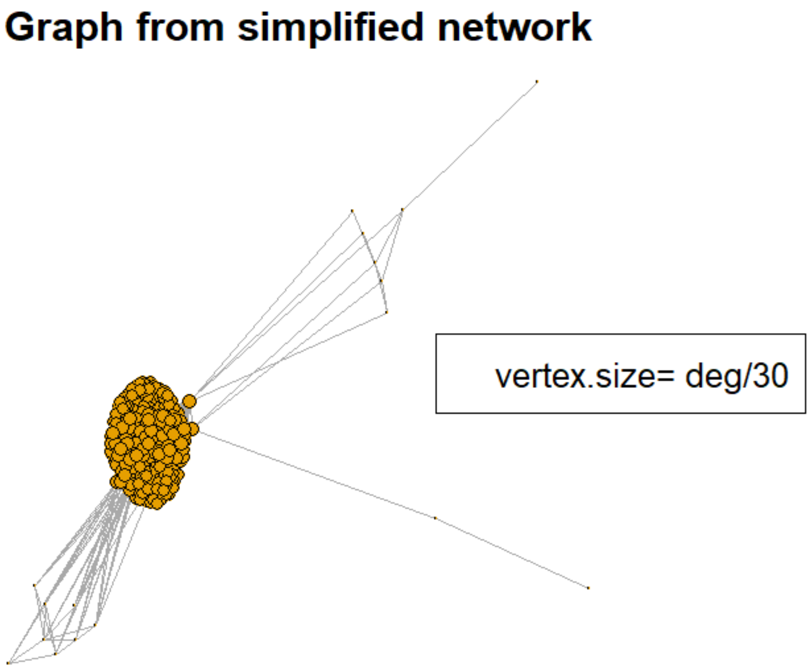
From those graphical representation it is possible to visualize the cells with most Churners.

5) ISSUE 5

Consider as connected, customers who visited the same museum at the same time more than twice. Draw a customers' network and compute measure of centralities for each node.

In this section we are dealing with a one-mode social network defined by the set of units and the relationships only between them (relationship is being at the same time at the same museum more than twice). To create a graph, I had to convert the in13 dataset into a matrix from-to (undirected). I obtained 48170 nodes from a link matrix 82789 obs. of 3 variables (from, to and attached the number of times that the pair of units went together in a museum). The Graph of the complete Network is shown at the end of the paper; it does not provide a nice visualization because of high number of nodes and edges, so I reduced the size of the network and represented a network with the most frequent individuals. The next figure shows a Graph of the network with self-loops omitted and 300 nodes. The full version of the network is presented in the section 10 of this work.

Figure 15



The vertices have been Placed on the plane using the force-directed layout algorithm by Fruchterman and Reingold[3]. The size of the vertices is proportional to the degree of the vertex and is $\frac{degree}{30}$. Higher is the degree, higher will be the size of the node (degree of vertex v, *deg (v)* is the number of lines with v as end-vertex, in this undirected context I don't distinguish between indegree and outdegree). The dense core in the centre of the figure contains the bigger vertices that are the one with the greatest number of connections. Since the graph is undirected (because they just went together to the museum, edges have no orientation) I cannot compute some measures of importance such like the authority score and reciprocity. There are many ways to draw the graph, I chose this one because of its better visualization, it is clear especially for the low-degree vertices.

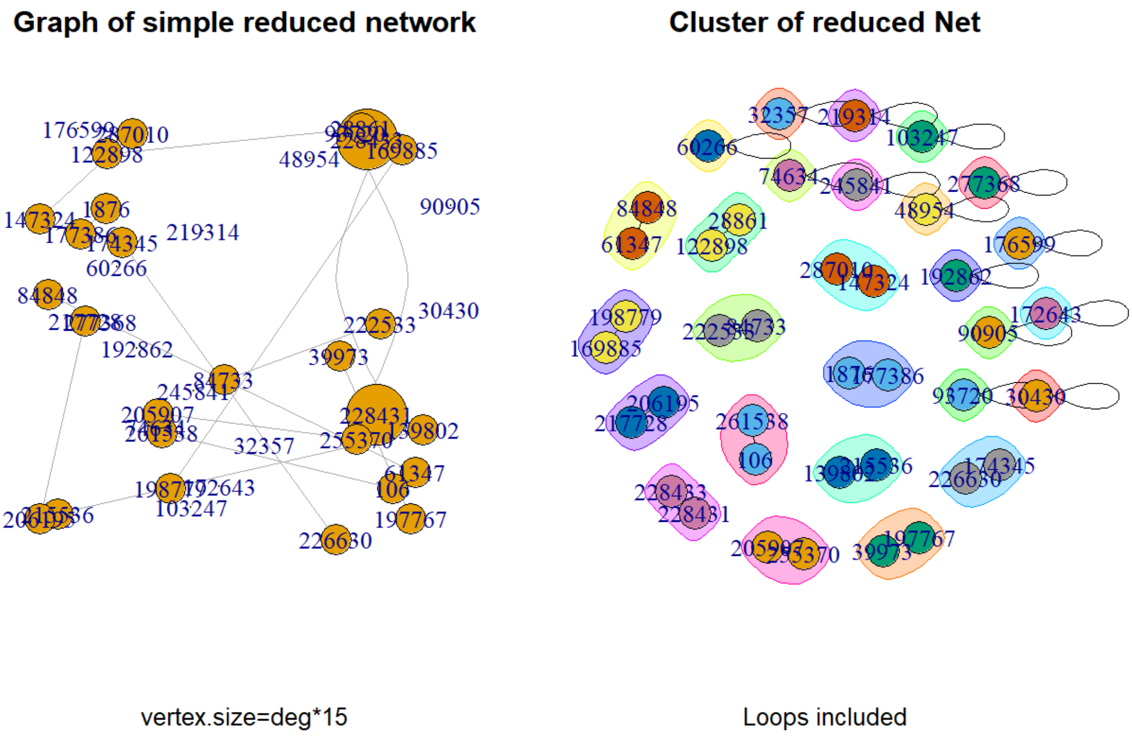
Here a little Set of related pairs of units, those who visited a museum more than 21 (max is 31) times together (I can observe it from the groups of people in the same museum, at the same time and date)

$$R = \{(39973: 197767), (61347: 84848), (84733: 222533), (139802: 215536), (147324: 287010\ ),$$
$$(122898: 28861), (147324: 287010\ ), (174345: 226630), (177386: 1876), (198779: 169885),$$
$$(217728: 206195\ ), (228431: 228433\ ), (228431: 228433), (255370: 205907), (261538: 106)\}$$

[3] uses an analogy of physical springs as edges that attract connected vertices toward each other and a competing repulsive force that pushes all vertices away from one another, whether they are connected or not. (https://www.sciencedirect.com/book/9780128177563/analyzing-social-media-networks-with-nodexl)

From this set of pair of units with 40 nodes (I omitted the loops in the previous list “R = {…}”) I made the following graph with a random layout; it is an undirected graph, and the dimension of the nodes is 15 times the degree of the vertex.

Figure 16



Graph of simple reduced network

Cluster of reduced Net

vertex.size=deg*15

Loops included

The aim of this representation is to provide a graph with labelled nodes (impossible to visualize for the full links matrix) in order to understand which client has more connections.

On the left is presented a community detection based on edge betweenness (Newman-Girvan) plot. High-betweenness edges are removed sequentially (recalculating at each step) and the best partitioning of the network is selected; I can see for instance that clients 228431 and 228433 have degree=2 while 39973 and 222533 degree=1 and consequently are bigger in the graph. On the right side there are clusters of clients, the loops (later removed) indicate a client who bought two tickets.

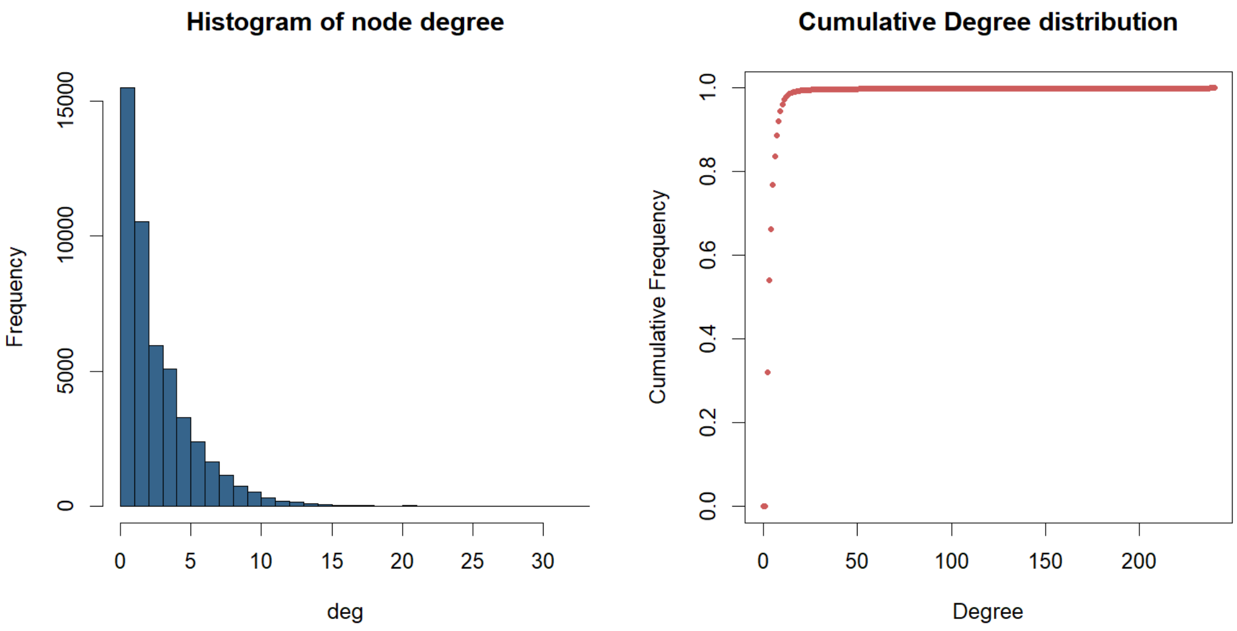
Here a synthesis of the number of adjacent edges for each of the 40 nodes:

Degree	Id (Client Code Most recurrent clients)
0	30430;32357; 48954;60266; 74634; 90905;93720; 103247; 172643; 176599; 192862; 219314; 245841; 277368
1	222533;28861;215536;287010;226630;1876;169885 206195;205907;106; 217728; 255370;261538;197767;84848;122898;139802;147324; 174345; 177386;198779; 39973; 61347; 84733
2	228431;228433

Note that on the left size of the figure (Fig.) those with degree=0 are omitted in the graph because the self-loops are omitted.

Now the distribution of the degree for the complete net

Figure 17



The following table summarizes the centrality measures for the whole network. The degree for each node is kept as new column in the dataset.

Network and node descriptives

Measure of centrality	value
Density	7.136058e-05
Global Transitivity	0.4048376
Diameter	46
Centralization	0.004911097
Eigen centrality (value)	75.6639
Reciprocity	NA (only for directed)
Betweenness (centralization)	0.01462364

Note that centrality measures could be computed also for al little subset of the nodes if the network is big. Density is the proportion of present edges from all possible edges in the network, from this kind of data I expected a low value because most of the customers of the museum are unrelated.

For the whole network I also create a dataset with centrality measures with client ids on the row and centrality measures by column. The Degree is set=0 when there are no connections (NAs in the merged dataset)

6) ISSUE 6

Is there a causal impact of gender on the probability of churning? identify a suitable model to create a counterfactual group with observational data

I will use propensity score matching to create counterfactual group to assess potential churning. The statistical quantity of interest is the causal effect of the exposure (gender) on probability of churning. In that way I'll estimate the causal effect of the binary exposure on the outcome (churning) while controlling for measured variables. In this context I am not able to measure the pre-treatment variables (since I assume that customer's gender pre-existed the other measured variables), but I can assume that those variables are not affected by the treatment. Treatment refers to the exposure of being a male (Gender(*sesso*)=1)

The general steps for matching analysis are:

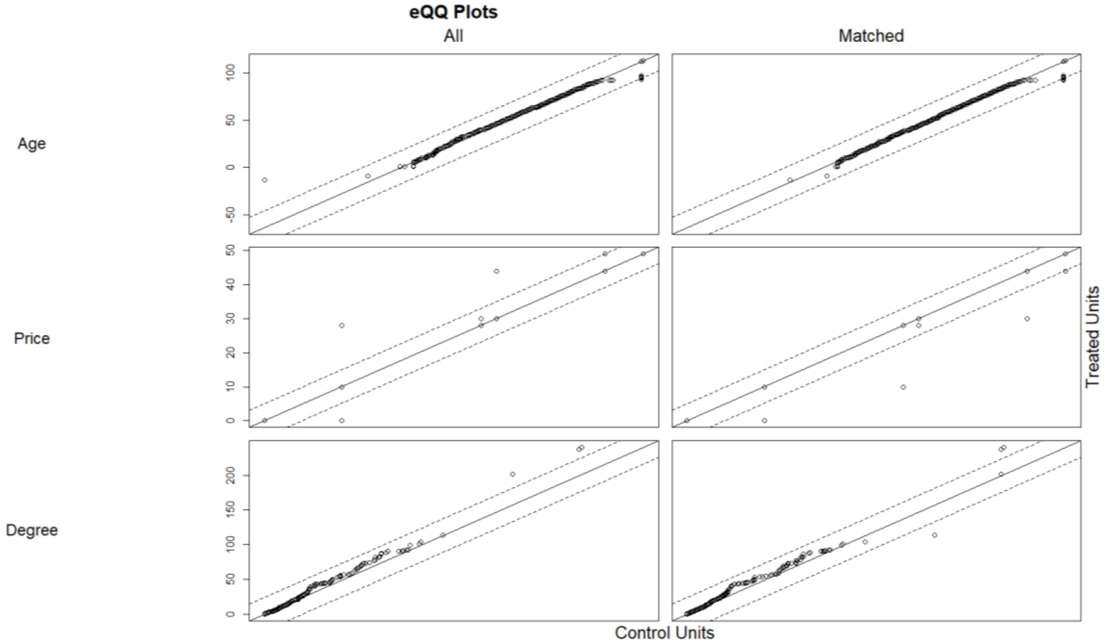
- Planning
- Matching
- Assessing quality of matches
- Estimating the treatment effect and its uncertainty

The planning ideally is prior to data collection, in this context the first decision is the formula, this object relates the exposure to the covariates used in estimating the propensity score and for which balance is to be assessed. The covariates selected to balance and ensuring the resulting treatment is free of confounding are Churn, Age, Price, Degree (The new variable from the previous ISSUE) and New Member.

The chosen method for matching is the nearest neighbour propensity score matching. (1:1 NN PS matching without replacement).

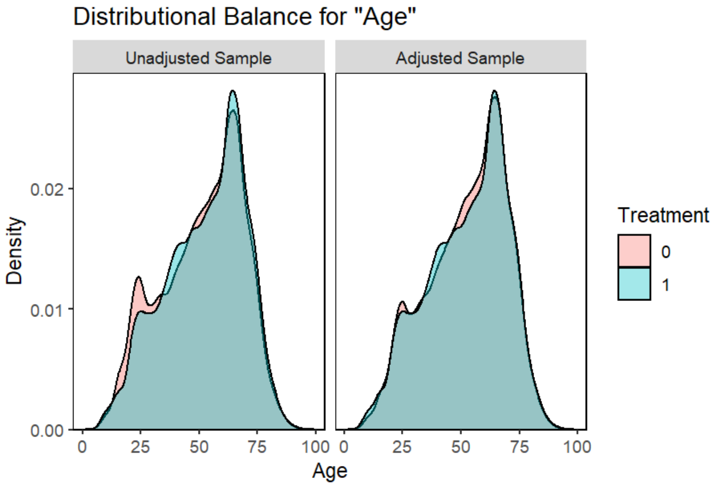
The QQplot between matched treated and control shows that the matching works (I am looking for the dots laying on the diagonal)

Figure 18



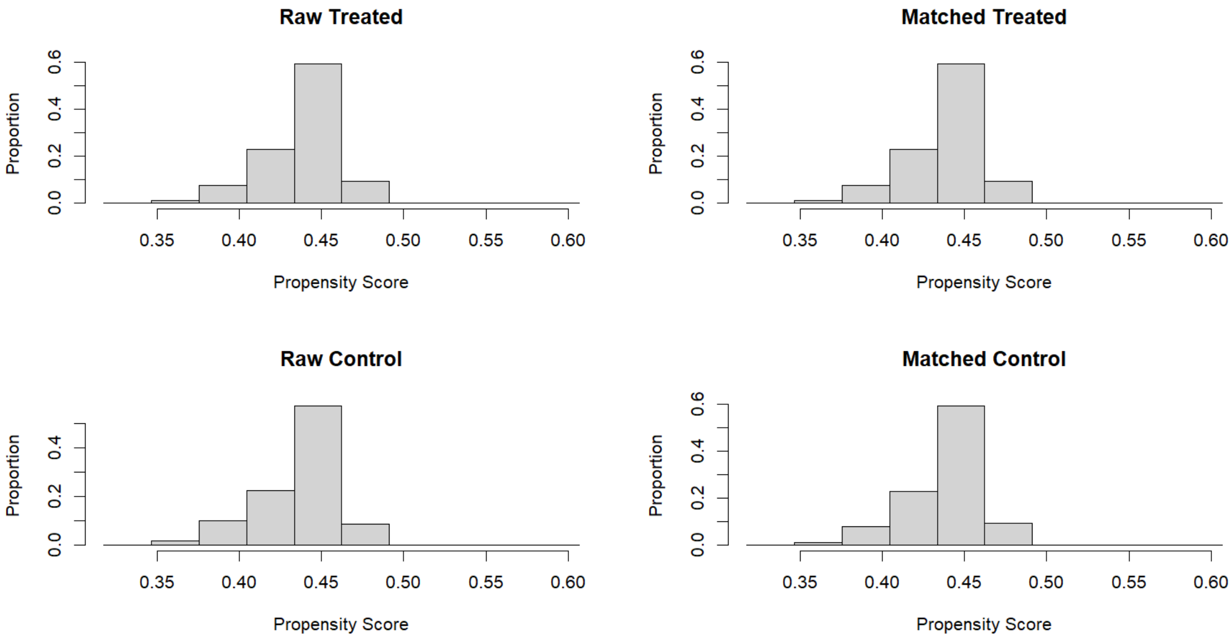
The next plot, for the variable "Age" (result of subtracting the dataset year to the variable "data_nascita") provides a graphical representation of the distributional balance; I am satisfied to see an overlap and the common support.

Figure 19



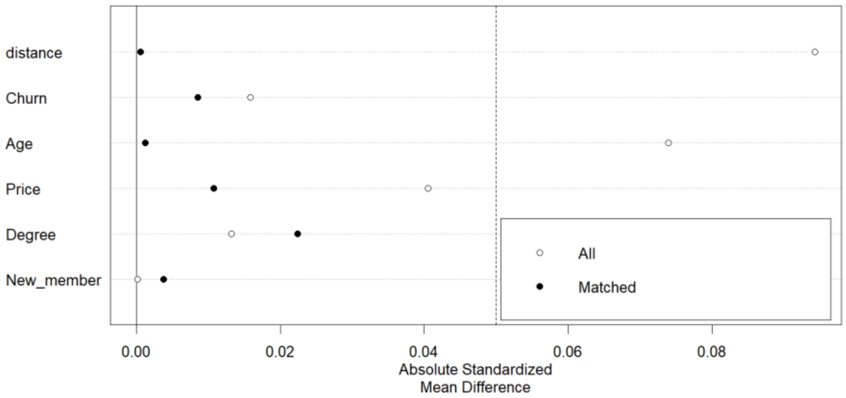
The next Figure shows the distributions of the propensity score for treated (Gender=1=Male) and Control in Raw and matched data, they all look similar.

Figure 20



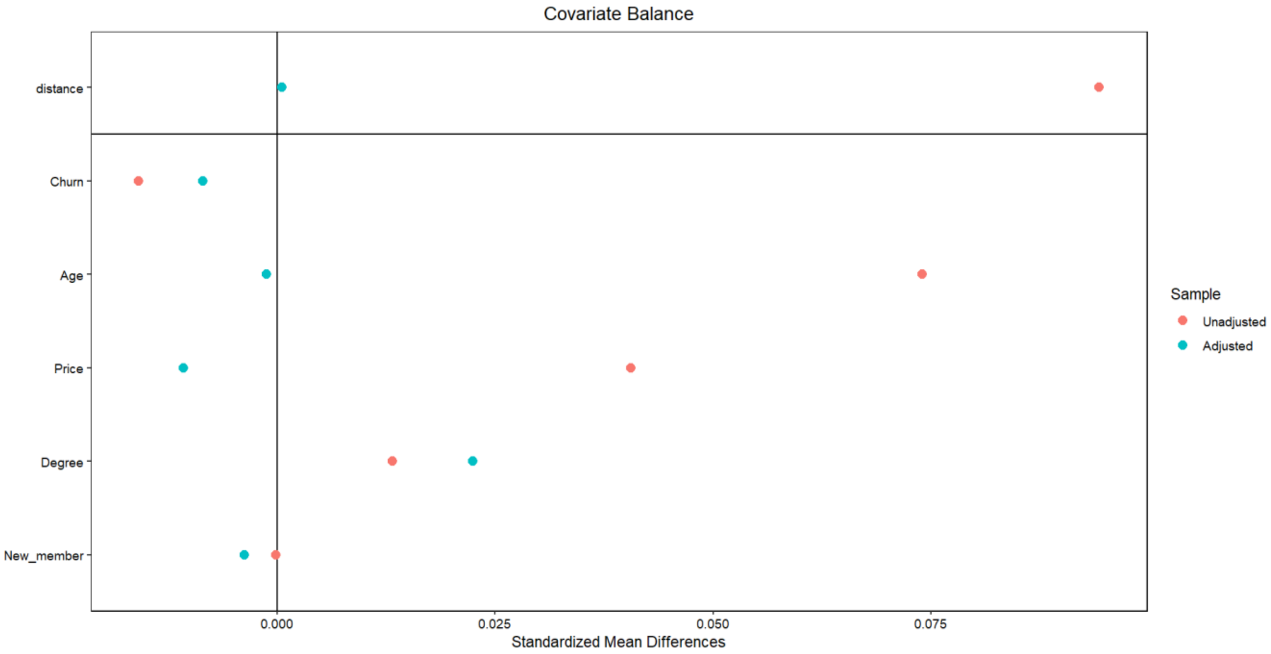
The next plot shows the standardized mean difference (SMD),close to zero which indicate good balance.

Figure 18



With the next plot (Love plot) we observe the standardized mean difference (SMD)[4] it is close to zero which indicate good balance. The vertical straight line represents the zero, all values are closer than 0.1, indicating a good balance.

Figure 21



For **Estimating the treatment effect** after 1:1 matching without replacement I run a simple regression of the churn on the gender in the matched sample (after having assessed the balance) including (and not) the matching weights. The coefficient on the treat (exposure of being a male) is the estimated Average treatment effect on the treated (ATT i.e., average effect of being a male for units like those that were male(treated)). ATT = -0.003939 (Coefficient on the exposure not significant, p-value=0.259)

[4] SMD is the difference in the means of each covariate between treatment groups standardized by a standardization factor so that it is on the same scale for all covariates. The standardization factor should be the same before and after matching to ensure changes in the mean difference are not confounded by changes in the standard deviation of the covariate. SMDs close to zero indicate good balance.

Applying a t-test to the difference in means I conclude that being a Male has an effect on the probability of churning. Note that a statistical significance given by the t-test doesn't imply a substantial influence of the gender on the Churning: even if significant the difference between the means is very little. The mean in group 0 is 0.6936778 while the mean in group 1 is 0.6897388 Welch Two Sample t-test the t is 1.1289 with 70067 degrees of freedom and p-value = 0.2589 that gives evidence against the null hypothesis of zero difference. (Alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0) The 95% C.I. is: (-0.002899634; 0.010777471).

NOTE: The "Full" Matching technique (matches every treated unit to at least one control and every control to at least one treated unit with probit link for the propensity score model.). has not been treated because of the size of the dataset

7) ISSUE 7

Which models could you use to predict churners? Run at least three prediction models and show the ROC curves for them. Compute the predicted probability on the test-set and show its distribution.

For this prediction task I'll run the following models:

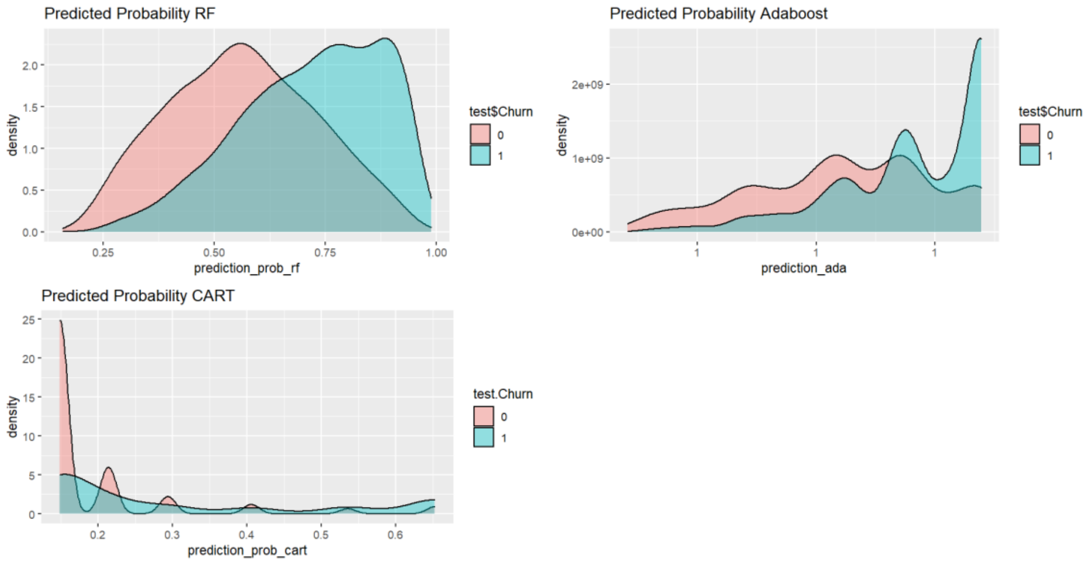
- Random Forest
- Classification tree (CART model)
- Adaboost

Before running the models, the dataset (the merged dataset, in which every row is a customer and the columns are the variables from the datasets data1, an13 and in13) is divided randomly in training and test set; I checked that the target distribution is kept over the train set easy and not perfect task to check whether in the creation we are doing right. check for distribution with a Kolmogorov-Smirnov Test (for numeric features[5]), looking for high p-values to not reject H0, otherwise I would have to try another split. The distribution of the target variable (0.31 churners and 0.69 non-churners) is kept both in training and test set.

In ISSUE 5 I created a dataset with some interesting centrality measures for the nodes (clients), I merged the two datasets by client code. The degree of the individual is a proxy of its popularity, individual who is likely to quickly connect with the wider network (or he's simply a very active user of the card). The variable considered are Churn as response: id + Price + City + New member + Age + No discount + Discounted pass + Student pass + Cash + Bank + Online payment + Visit frequency + Total price + Gender + Degree.

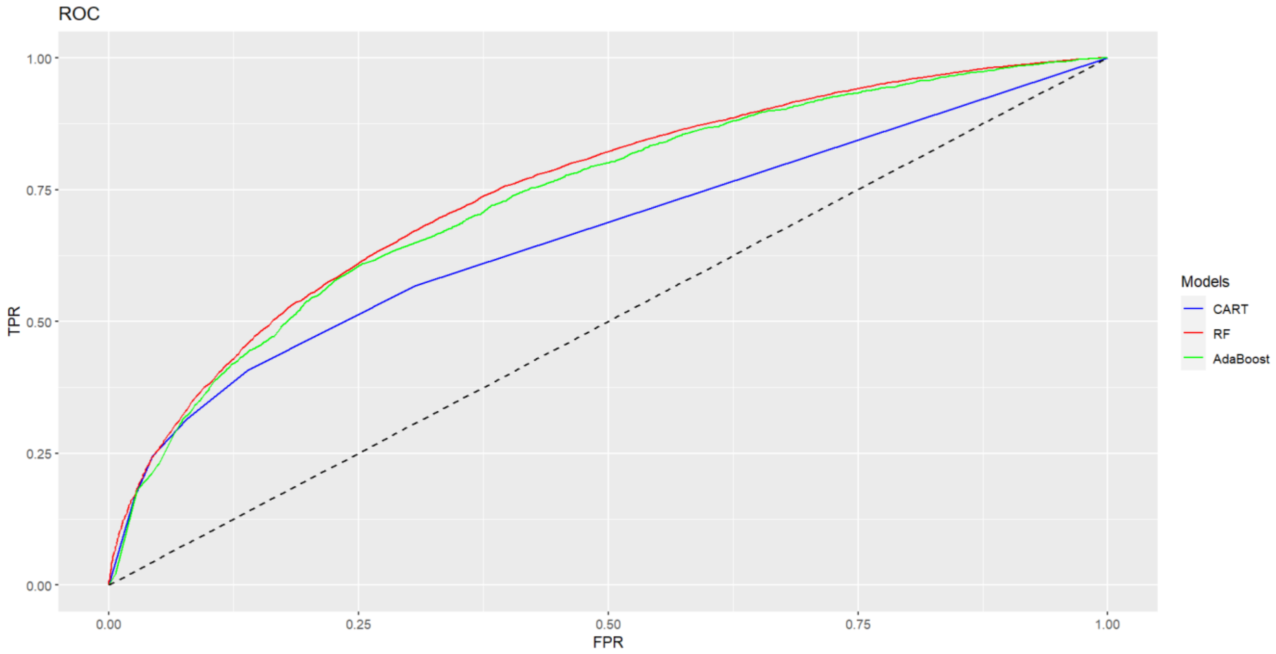
[5] p-value = 0.9362 for the variable Price, good

Figure 22



The previous plot (*Fig.22*) shows the distributions of the predicted probability on the test-set for the 3 models considered grouped by churning (0=Churn, 1=Retain).

Figure 23



This figure Shows the ROC curves for the 3 models, The Random Forest in red looks has the highest AUC followed by adaboost (73.56%). All models perform better than random guess. The next table provides a syntesis of the performances for the best and the worst model.

Statistics CART and RF

MODEL	AUC	Accuracy	Sensitivity	Specificity
CART	66.24%	0.4626	0.9561	0.2449
Random Forest	74.89%	0.7312	0.3658	0.8923

In this context the sensitivity is the Frequency of being correct on positive (=1 i.e., Retained) and specificity is the frequency of being correct on negative (=0, i.e. churners). I would prefer a model with higher specificity (True Negative Rate) because it is better in detecting churners.

8) ISSUE 8

Consider a marketing campaign addressing directly single customers. We know that each contact costs 2 euro. We can also compute the consumer value for each single customer. We can reasonably assume that a churner, contacted for the campaign has a probability of 10% of not churning. Non-churners contacted are simply a cost of 2 euros. With this additional information, generate a profit curve of each prediction model. Discuss your results.

There are two critical conditions underlying the profit calculation: First the class priors The proportion of positive and negative instances in the target population; secondly the costs and benefits the expected profit is specifically sensitive to the relative levels of costs and benefits for the different cells of the cost-benefit matrix.

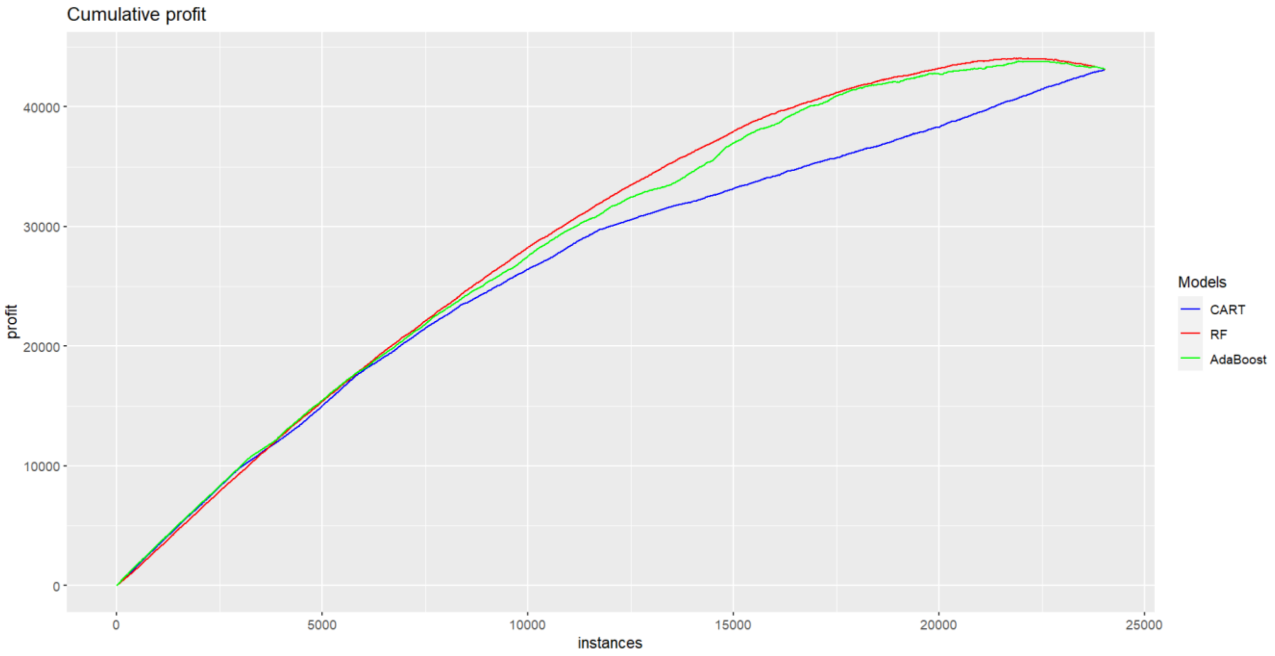
As suggested, I'll take the test set, consider the profit generated by the campaign for non-churners and the expect profit for the churners and order them according to their predicted probability of churn.

if I act on a positive, I get Gamma (already cost considered). if I act on negative, I get -cost. Gamma could also be client-specific (for instance the highest one is 6.38 €) and could be seen as the benefit of contacting the right client.

$$\gamma = 0 \times 0.9 + 0.1 \times (\text{Price} + \text{Visit\_frequency} \times 0.2 - 2)$$

0.2 is the contribution for every visit from the city of Turin. I recall that with positives I mean retained (Churn=1) and with negatives I mean Churners (Churn=0) 0.1 is the probability of the contacted churner of non-churning, 0.9 is its probability of churning after the call. In other words, with 0.1 we weight the benefit of retaining a client with the call (Price paid (*importo*) plus the 20 cents from Tourin for each visit, minus the cost of the contact).

Figure 24



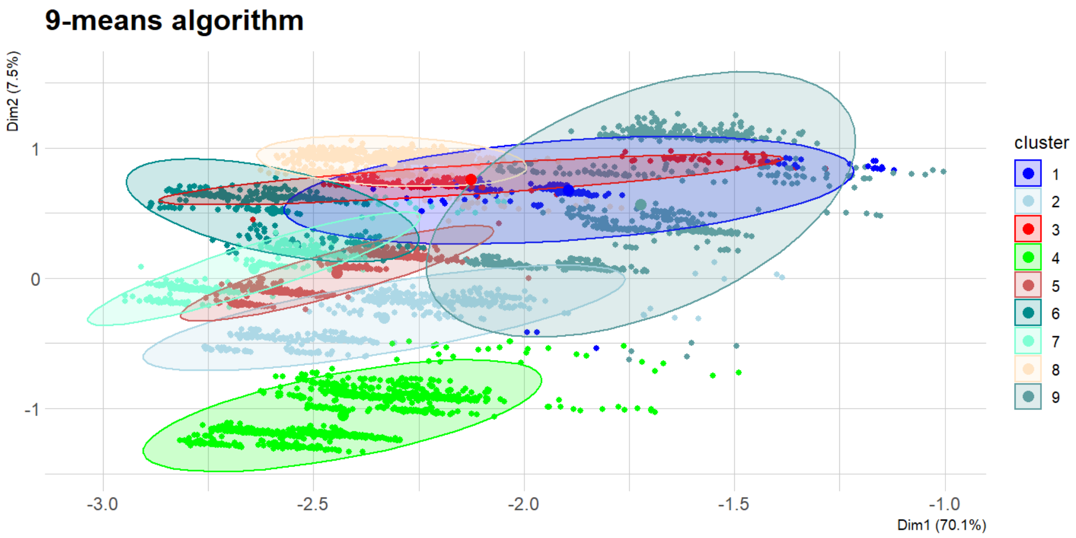
The models that better succeed in generating the profit curve are the Random Forest and the Adaboost. The classes are imbalanced (around 30% churners and 70% non-churners). Since the Random Forest has a better Sensitivity (works better in predicting churers) than the other models this result (RF as best model for this task) is not unexpected, if contacting a churner who won't be retained has an high cost, identifying the churner is crucial.

9) REFERENCES/SOURCES

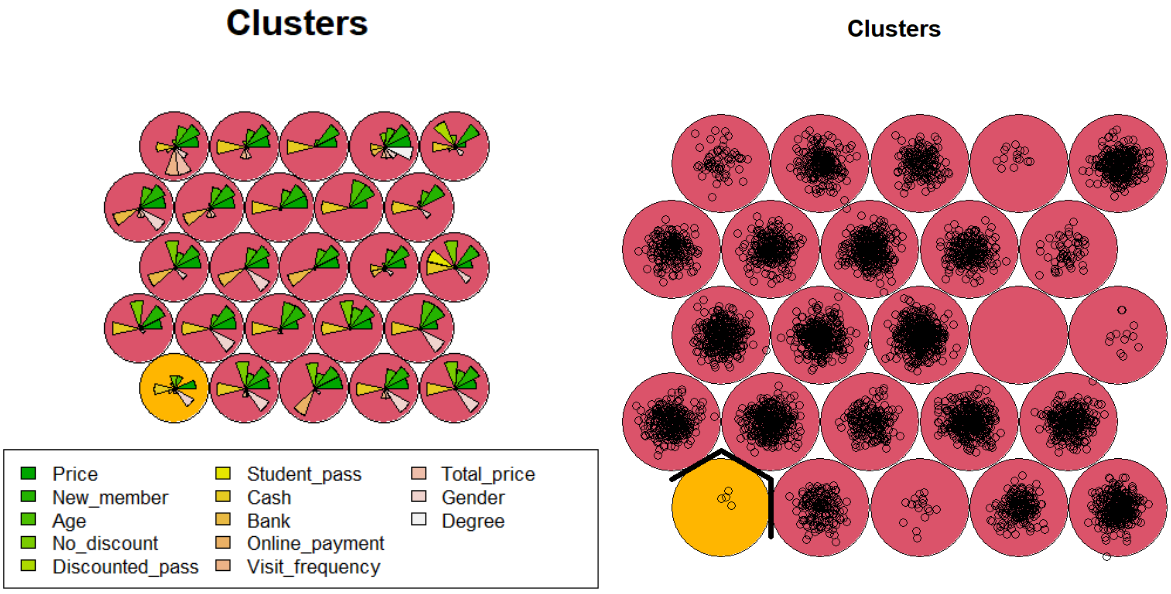
- https://kateto.net/networks-r-igraph library igraph
- ISTAT rapport annual 2014 https://www.istat.it/it/files/2014/05/Rapporto-annuale-2014.pdf
- Kuhn, Silge (2021+). Tidy Modeling with R. In progress https://www.tmwr.org/
- Lecture notes & codes 2021/2022
- Wickham, Grolemund (2017). R for Data Science. O'REILLY

10) ADDITIONAL FIGURES

9-means with less pre-processing (ISSUE 4)

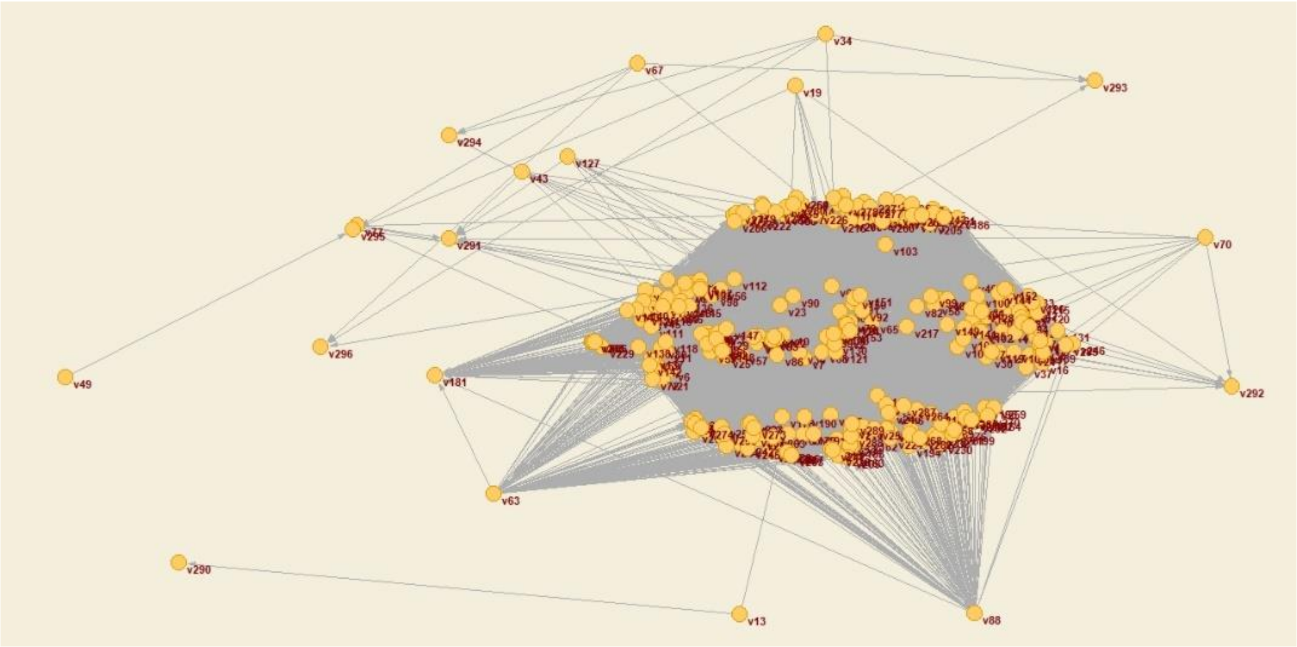


Alternative clustering with Kohonem map (ISSUE 4)



The results of the clustering in the supervised exercise with the Kohonen map doesn't seems to be stable across different specifications, the presented result was the best in terms of interpretability.

Complete Network with Pajek (ISSUE 5)



The arrows refer to the relation from->to, in the issue I have considered the network as undirected