Predictive models for Torino's museums subscription churn rate

Leonardo Patricelli

Introduction to the dataset

In this paper I am going to show you some models in order to predict what kind of customers tend to renew the subscription of Piemonte's museums and who do not; To do this I used the dataset with the informations of 2014's customers. First of all let's have a look at the most interesting variables :

```
##   codcliente spesa_tot n_visite sesso     comune   cap eta_abbonati si2014
## 1          6      8.60        2     0     TORINO 10122           45      1
## 2         13     33.50        9     0     TORINO 10147           44      1
## 3         16     25.75        6     1 SAVIGLIANO 12038           29      1
## 4         23      9.25        2     0   VOLPIANO 10088           18      1
## 5         27      5.00        1     1  CARIGNANO 10041           52      1
## 6         28     16.50        5     0     TORINO 10138           80      1
```

Now a quick introduction to the variables:

- **codliente** is the customer's ID
- **spesa_tot** is the total amount of money spent in museum tickets
- **n_visite** is the number of museums visited by the customer
- **sesso** is the gender of the customer
- **comune** is the city in which the customer lives
- **cap** is the zip code of the customer's city
- **eta_abbonati** is th ecustomer's age
- **si2014** is a binary variable, 0 is for the customers who did not renew the card in 2014, 1 otherwise.

Variable "Residenza" and "mesi_tot"

In order to improve the dataset I created two more variables: **Residenza** and **mesi_tot**. **Residenza** is a factor variable with 4 levels that express the distance from Torino; in fact it was shown by following analysis that people who lives close to Torino renew more often the subscription rather than who lives further.

```
##   codcliente     comune   cap Residenza
## 1          6     TORINO 10122         3
## 2         13     TORINO 10147         3
## 3         16 SAVIGLIANO 12038         1
## 4         23   VOLPIANO 10088         2
## 5         27  CARIGNANO 10041         2
## 6         28     TORINO 10138         3
```

The levels are assigned according to the distance from Torino ;

- **3** if the customer lives in Torino

- **2** if the customer lives in Torino's province

- **1** if the customer lives outside Torino's province but in Piemonte

- **0** if the customer lives outside Piemonte.

Mesi_tot is another variable created by me, it includes values from 1 to 12, and shows in how many different months per year the customer went to visit any museum.

```
##   codcliente spesa_tot n_visite mesi_tot
## 1          6      8.60        2        2
## 2         13     33.50        9        4
## 3         16     25.75        6        4
## 4         23      9.25        2        2
## 5         27      5.00        1        1
## 6         28     16.50        5        3
```

The logic behind the variable is that casual customers (and especially tourists) visit museum in less than a month, because usually their permanence last for less than a month. Secondly, in this way is possible to identify people who go in museums during the year, and not only in a period. Here we can see the frequency of each level of the variable.

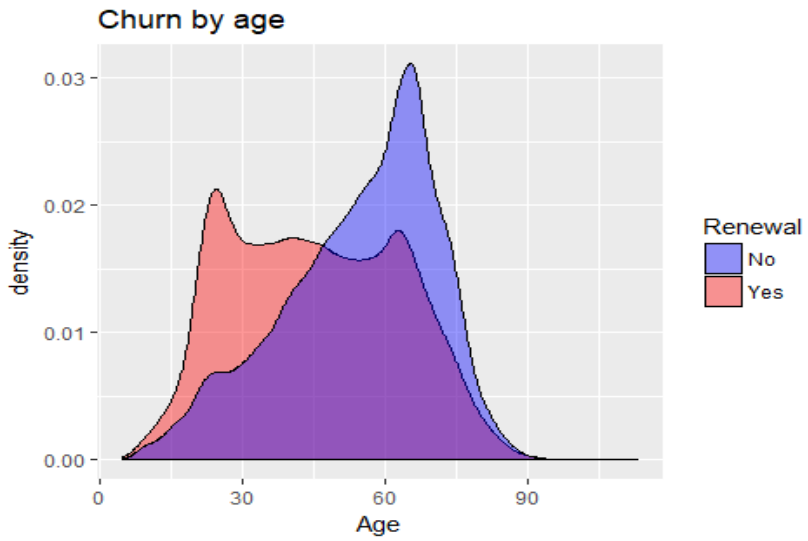
```
##    mesi_tot  freq
## 1         1 11221
## 2         2 12734
## 3         3 12326
## 4         4 10934
## 5         5  8749
## 6         6  6350
## 7         7  4182
## 8         8  2558
## 9         9  1455
## 10       10   751
## 11       11   308
## 12       12    98
```

Analyzing the dataset

Customer's age

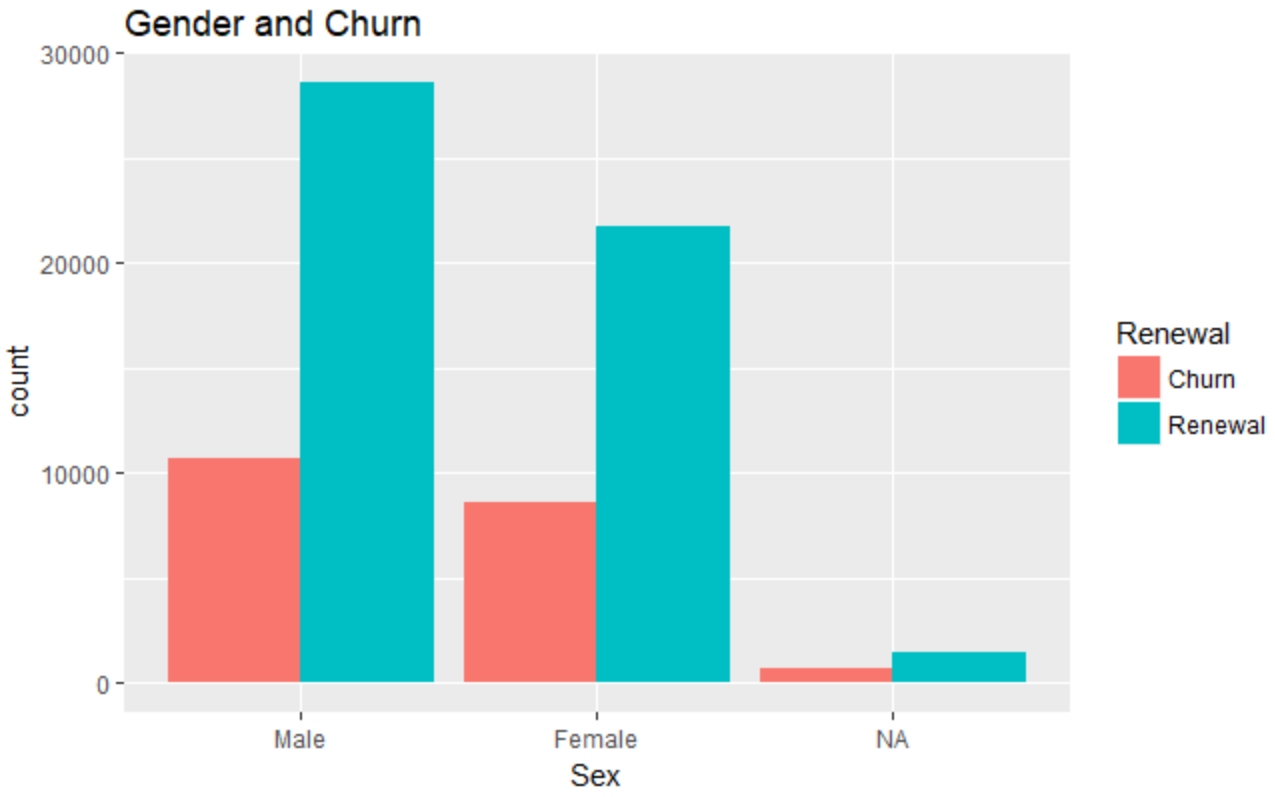
From the graph it is easy to see that there is no difference in subscriber's age distribution; in fact female and male distribution curves are almost overlapping. So, we can assume that the gender is not important for subscribing.

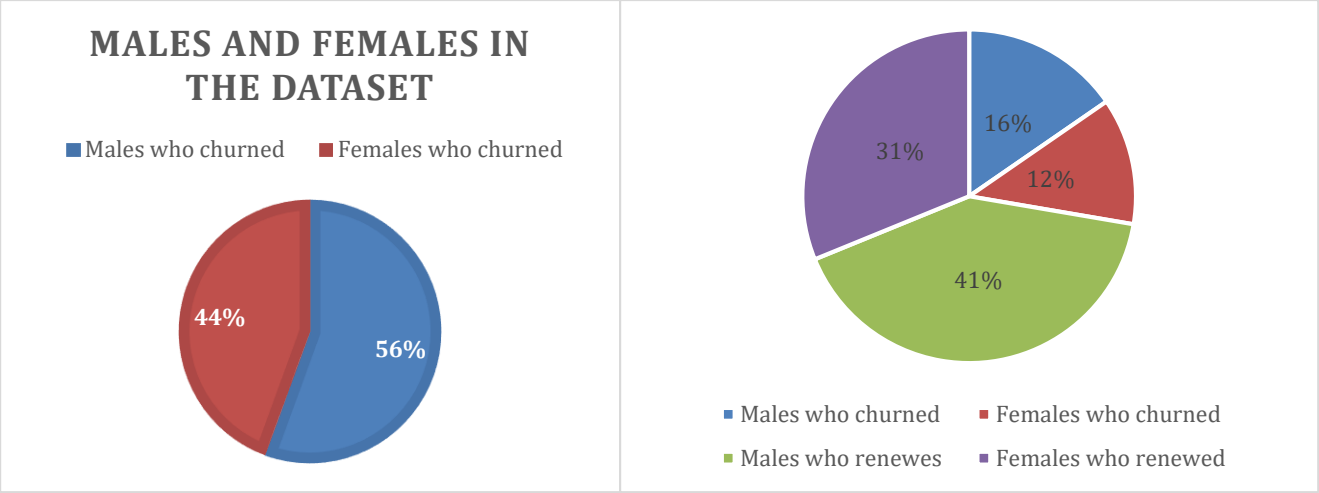
Renewals per age



The graph shows the age of who Renew the subscription and who does not. As you can see the people tend to not renew the signature according to their age : higher the age, higher the chance to not renew. Instead, the age distribution between who renew is more or less constant until the sixties. So, we can conclude that the age can be an important variable to include in our prediction analysys.

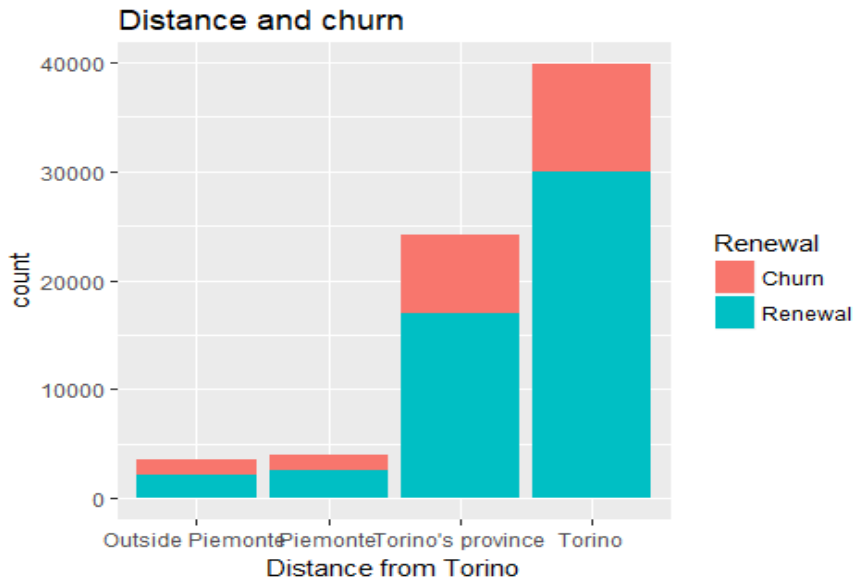
Renewals by Gender





Here we can see the gender of the people who renew the subscription and of who did not. In matter of churning, the proportions of female and male is the same: in fact the 72,78% of males renew the subscription and so do the 72,7% of Female. According to these results, it seems that the gender is not an important variable to predict renewals.

Distance and churn

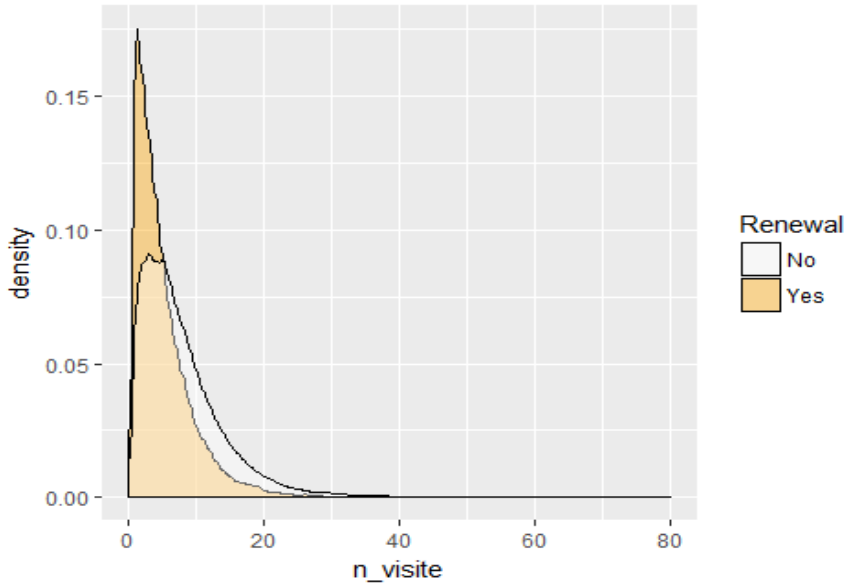


This graph shows the number of Renewals per region. It is clear that most of Renewals come from people who lives in Torino and in its province. Infact there are few subscriptions from people who lives far away from Torino, and about half of them churn. This variable seems to be important for the prediction model.

Spending and churn

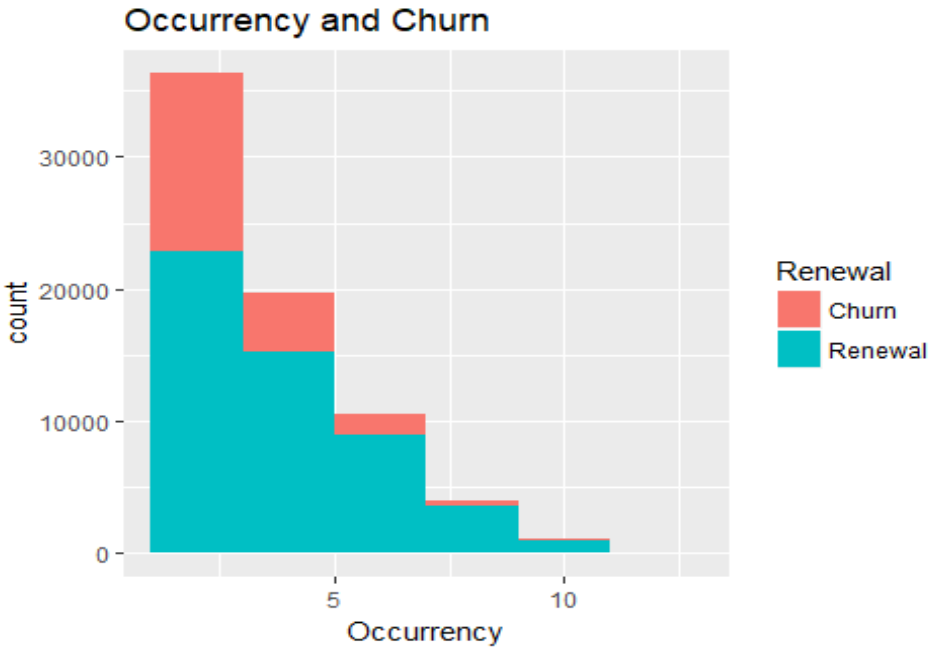
Here we can have a look at the total spending for customers who renewed and who did not. The average Customer who renewed the subscription spend more compared to the other ones. We can conclude that this affect in a positive way the probability of renewal (more you spend higher the probability).

Number of visits and churn



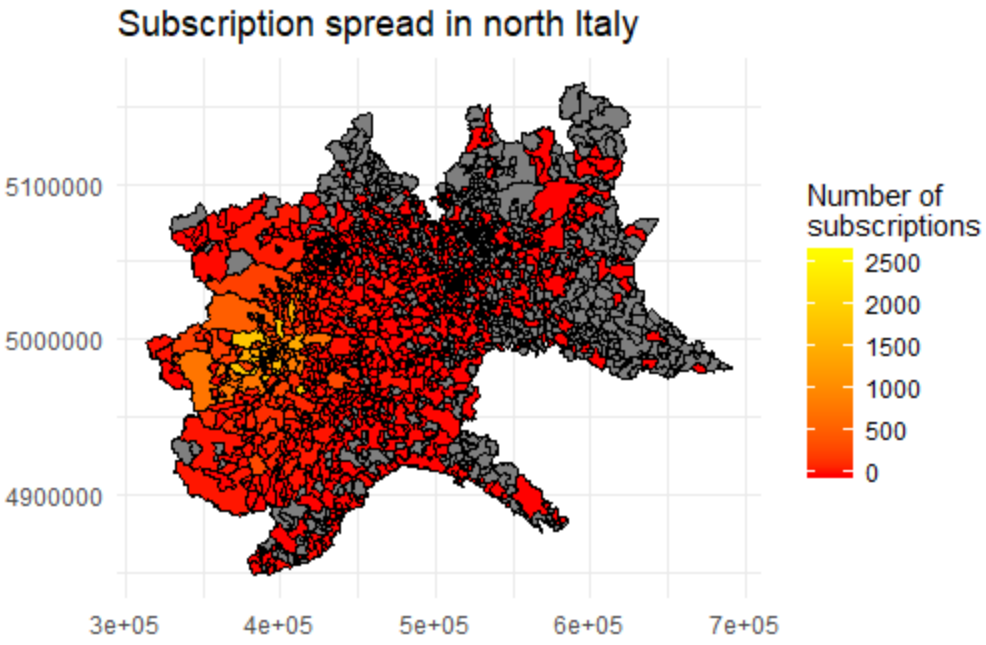
Here we are analyzing the variable relative to the number of visits. This variable is related to the previous one : in fact higher the number of visits higher the total spending. Because of this we have results similar to the previous ones. In conclusion, higher the visits higher the renewal probability.

"Mesi tot" and Renewal

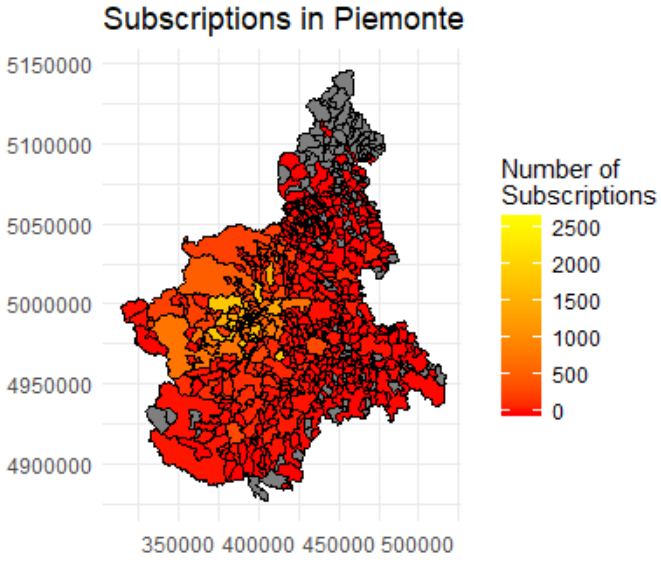


Here is visualized the "mesi_tot" variable among the subscribers. It is clear that people who use the subsciption only in a short time (one or two month) tend to churn; this could be possible because most of people are turists, and will go away from Torino after a while.

Subscriptions distribution in Nord-Italy

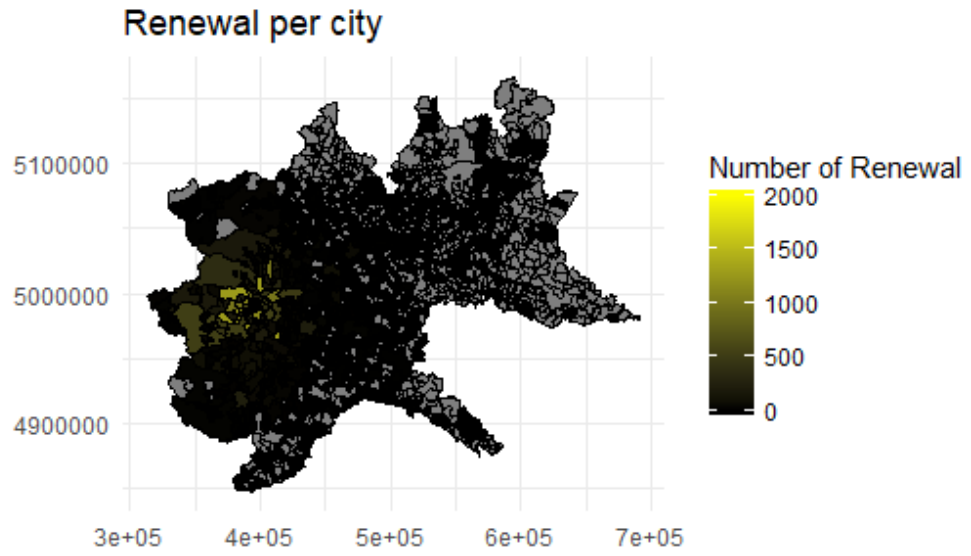


The map show what is the distributon of the subscriptions in north-Italy. The borders in the shapefile define the italian cities (in fact each region defined by the borders has an own zipcode) At first glance we can see that we have colours close to yellow in proximity of Torino, because in all the other cities outside Piemonte the color is red. Gray color is for cities which have no subscriptions at all (missing values). So it is better to do a focus to Piemonte status.

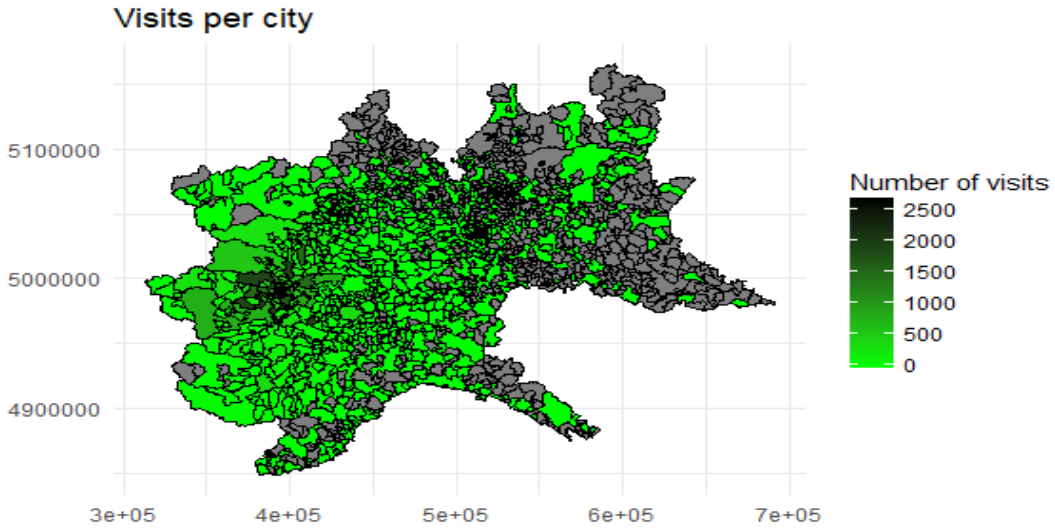


Zooming on Piemonte it is clear that in Torino's Province there is the highest number of subscriptions. This fact support our decision to build a variable that takes into account the distance from Torino. Also, Analizing the number of visits per city it is clear that people who lives in cities close to torino tend to do more visits.

Renewals Distribution in North Italy



The picture shows subscription's renewal distribution in North Italy. It was to be expected by previous analysis that most of the renewals are in Piemonte's region, so we should focus on this area. As we can see here, the Torino's province is the area with the highest number of renewals; this result supports the decision to include a variable to consider Torino's distance.

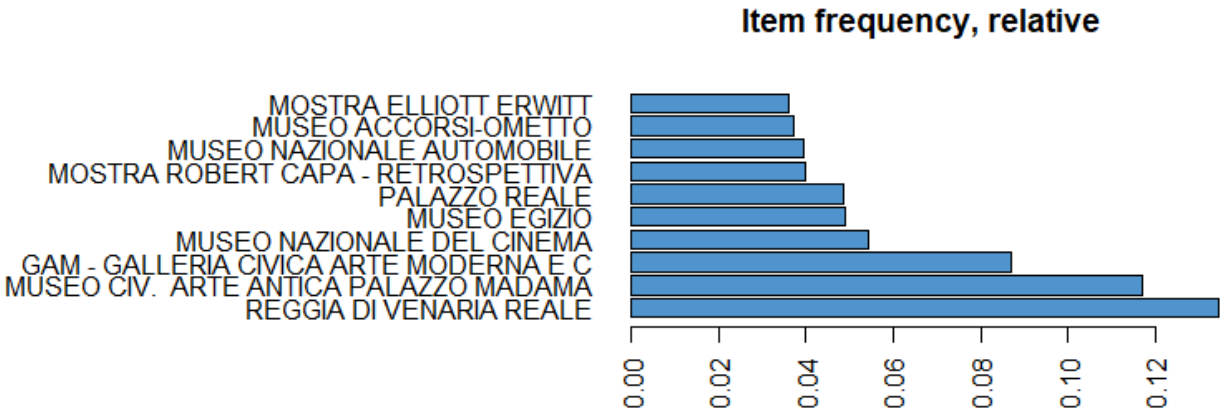


Also analyzing the visit per city, we can easily see that Torino's zone is the one with the highest density.

Market Basket Analysis

Now we are going to proceed with Market Basket Analysis in order to identify what are the most visited museums and to understand if there are any association rules between museums.

The graph below shows the most visited places in the dataset.



Now let's proceed analyzing the items.

```
##      items                                          support count        lift
## [1]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA}  0.010444382  4369   2.4824341
## [2]  {MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       PALAZZO REALE}                          0.004384298  1834   0.7725365
## [3]  {MOSTRA ELLIOTT ERWITT,
##       MOSTRA ROBERT CAPA - RETROSPETTIVA}     0.004300628  1799   3.0116604
## [4]  {MOSTRA BORN SOMEWHERE,
##       MUSEO REGIONALE DI SCIENZE NATURALI}    0.004150022  1736  31.6763349
## [5]  {MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       PALAZZO REALE}                          0.003803390  1591   1.9745540
## [6]  {MUSEO ANTROP. CRIMINALE CESARE LOMBROSO,
##       MUSEO DELLA FRUTTA}                     0.002806524  1174 168.4521975
## [7]  {MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO EGIZIO}                           0.002440768  1021   0.4264981
## [8]  {MUSEO EGIZIO,
##       MUSEO NAZIONALE DEL CINEMA}             0.002199321   920   0.8315469
## [9]  {MUSEO EGIZIO,
##       PALAZZO REALE}                          0.002034372   851   0.8597911
## [10] {MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA} 0.001931577   808   0.4146084
```

Here we have ordered our transactions by support, so we have the items with the highest frequency of occurrence. Especially {MOSTRA BORN SOMEWHERE,MUSEO REGIONALE DI SCIENZE NATURALI} and {MUSEO ANTROP. CRIMINALE CESARE LOMBROSO, MUSEO DELLA FRUTTA} have a very high lift.

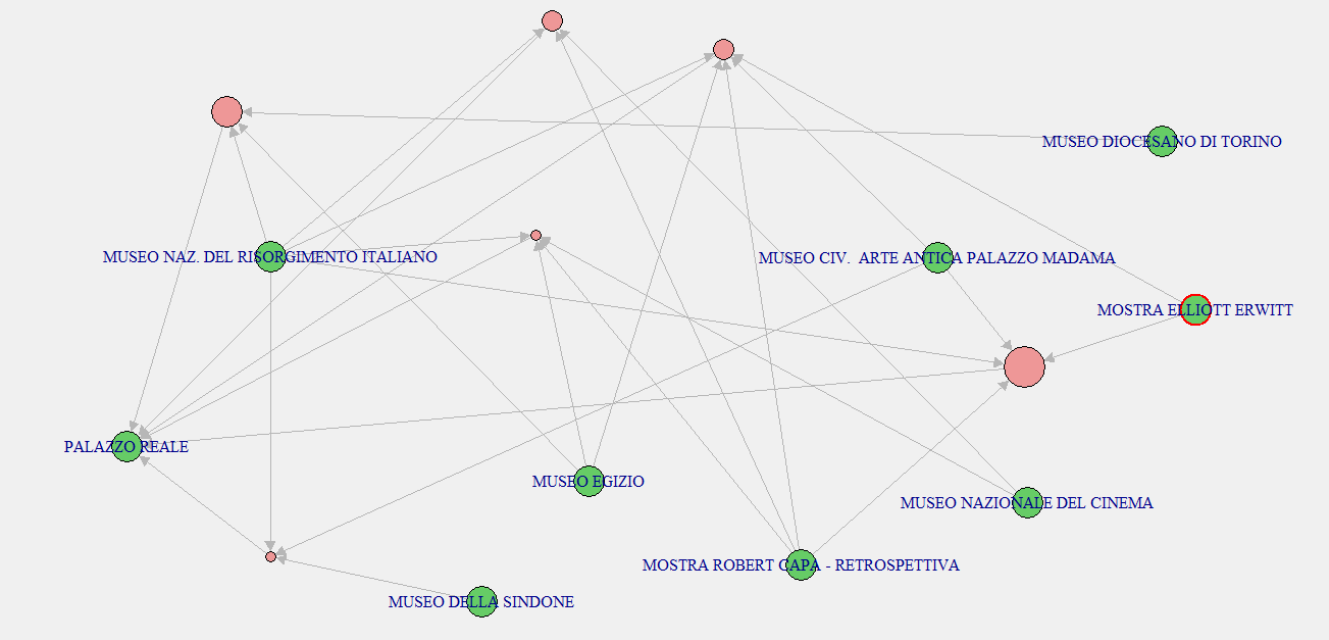
What patterns lead customers to visit a specific museum?

Now let's proceed to search rules for the most visited museums. Our aim is to find out what customers have purchased before visiting the museum under the column "rhs". This will help us to understand the patterns that led to the visit of the museum under the colum "rhs". I added the criterium "confidence > 0.7". Confidence is probability of purchase B, given purchase A happened (basically the conditional probability P(B|A)). For recommending a good rule we prefer higher confidence. Here I propose you only the best "10 rules" per museum that I found out.

Palazzo Reale

```
##      lhs                                          rhs                  support confidence     lift
count
## [1]  {MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO DELLA SINDONE,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO}   => {PALAZZO REALE} 1.195283e-05  1.0000000 20.64408
5
## [2]  {MUSEO DIOCESANO DI TORINO,
##       MUSEO EGIZIO,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO}   => {PALAZZO REALE} 1.673396e-05  1.0000000 20.64408
7
## [3]  {MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA}             => {PALAZZO REALE} 1.434340e-05  1.0000000 20.64408
6
## [4]  {MOSTRA ELLIOTT ERWITT,
##       MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO}   => {PALAZZO REALE} 1.912453e-05  1.0000000 20.64408
8
## [5]  {MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO EGIZIO,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA}             => {PALAZZO REALE} 1.195283e-05  1.0000000 20.64408
5
## [6]  {MOSTRA ELLIOTT ERWITT,
##       MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO EGIZIO,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO}   => {PALAZZO REALE} 1.434340e-05  1.0000000 20.64408
6
## [7]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA}             => {PALAZZO REALE} 1.912453e-05  0.8888889 18.35029
8
## [8]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO EGIZIO,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA}             => {PALAZZO REALE} 1.912453e-05  0.8888889 18.35029
8
## [9]  {MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO EGIZIO,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO}   => {PALAZZO REALE} 1.434340e-05  0.8571429 17.69493
6
## [10] {MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO EGIZIO,
##       MUSEO NAZIONALE DEL CINEMA}             => {PALAZZO REALE} 2.390566e-05  0.8333333 17.20340
10
```

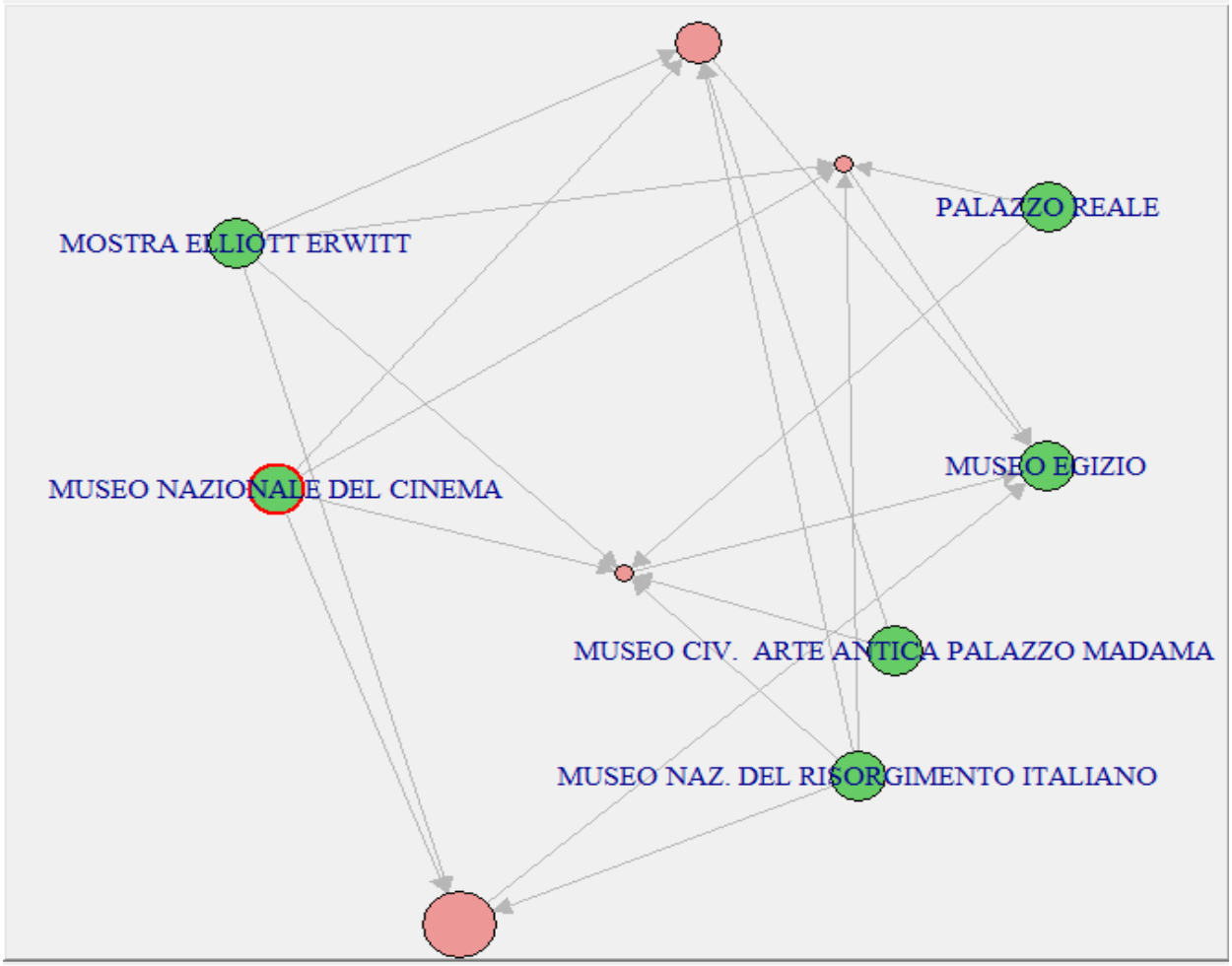
Here the visualization of the 6 rules with the highest confidence.



Museo Egizio

```
##      lhs                                    rhs                support confidence     lift count
## [1]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA}        => {MUSEO EGIZIO} 2.390566e-05  1.0000000 20.47232    10
## [2]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA,
##       PALAZZO REALE}                     => {MUSEO EGIZIO} 1.912453e-05  1.0000000 20.47232     8
## [3]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA}        => {MUSEO EGIZIO} 2.151509e-05  1.0000000 20.47232     9
## [4]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA,
##       PALAZZO REALE}                     => {MUSEO EGIZIO} 1.912453e-05  1.0000000 20.47232     8
## [5]  {MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA}        => {MUSEO EGIZIO} 1.195283e-05  0.8333333 17.06027     5
## [6]  {MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA,
##       PALAZZO REALE}                     => {MUSEO EGIZIO} 1.195283e-05  0.8333333 17.06027     5
## [7]  {MUSEO DIOCESANO DI TORINO,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       PALAZZO REALE}                     => {MUSEO EGIZIO} 1.673396e-05  0.7777778 15.92292     7
## [8]  {MOSTRA ELLIOTT ERWITT,
##       MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       PALAZZO REALE}                     => {MUSEO EGIZIO} 1.434340e-05  0.7500000 15.35424     6
## [9]  {MOSTRA ELLIOTT ERWITT,
##       MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO}   => {MUSEO EGIZIO} 1.434340e-05  0.7500000 15.35424     6
## [10] {MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       PALAZZO REALE}                     => {MUSEO EGIZIO} 1.434340e-05  0.7500000 15.35424     6
```

Here the visualization of the 4 rules with the highest confidence.



Now let's see another type of association rules; this time I will find out what products were purchased after/along with product X, so we will have our "given item" under the "lhs" column. The criterium "confidence > 0.7" is still valid. I will add the visualization of the rules with confidence = 1 under the output.

I will not show the plot of the rules because they will result too much chaotic.

Mostra Elliot Erwitt

```
##      lhs                                        rhs                                          support confidence      lift
count
## [1]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO DIOCESANO DI TORINO,
##       PALAZZO REALE}                          => {MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA} 1.195283e-05          1  8.535391
5
## [2]  {MOSTRA ELLIOTT ERWITT,
##       MOSTRA WILDLIFE PHOTOGRAPHER OF THE YEAR,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA}   => {MUSEO REGIONALE DI SCIENZE NATURALI}    2.390566e-05          1 37.770745
10
## [3]  {MOSTRA BORN SOMEWHERE,
##       MOSTRA ELLIOTT ERWITT,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA}   => {MUSEO REGIONALE DI SCIENZE NATURALI}    1.912453e-05          1 37.770745
8
## [4]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA}               => {MUSEO EGIZIO}                           2.390566e-05          1 20.472324
10
## [5]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO ACCORSI-OMETTO,
##       MUSEO NAZIONALE DEL CINEMA}               => {MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA} 1.434340e-05          1  8.535391
6
## [6]  {MOSTRA ELLIOTT ERWITT,
##       MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       PALAZZO REALE}                          => {MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA} 1.912453e-05          1  8.535391
8
## [7]  {MOSTRA ELLIOTT ERWITT,
##       MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO}     => {PALAZZO REALE}                          1.912453e-05          1 20.644080
8
## [8]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA,
##       PALAZZO REALE}                          => {MUSEO EGIZIO}                           1.912453e-05          1 20.472324
8
## [9]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA}               => {MUSEO EGIZIO}                           2.151509e-05          1 20.472324
9
## [10] {MOSTRA ELLIOTT ERWITT,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA,
##       PALAZZO REALE}                          => {MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA} 1.912453e-05          1  8.535391
8
```

Mostra Robert Capa - Retrospettiva

```
##      lhs                                       rhs                                       support confidence      lift
count
## [1]  {MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO NAZIONALE DELLA MONTAGNA}        => {MOSTRA ELLIOTT ERWITT}                 1.195283e-05          1 27.846558
5
## [2]  {MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA}            => {PALAZZO REALE}                         1.434340e-05          1 20.644080
6
## [3]  {MOSTRA ELLIOTT ERWITT,
##       MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       PALAZZO REALE}                         => {MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA} 1.912453e-05          1  8.535391
8
## [4]  {MOSTRA ELLIOTT ERWITT,
##       MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO}  => {PALAZZO REALE}                         1.912453e-05          1 20.644080
8
## [5]  {MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       PALAZZO REALE}                         => {MOSTRA ELLIOTT ERWITT}                 1.912453e-05          1 27.846558
8
## [6]  {MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO EGIZIO,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA}            => {PALAZZO REALE}                         1.195283e-05          1 20.644080
5
## [7]  {MOSTRA ELLIOTT ERWITT,
##       MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO EGIZIO,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       PALAZZO REALE}                         => {MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA} 1.434340e-05          1  8.535391
6
## [8]  {MOSTRA ELLIOTT ERWITT,
##       MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO EGIZIO,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO}  => {PALAZZO REALE}                         1.434340e-05          1 20.644080
6
## [9]  {MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO EGIZIO,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       PALAZZO REALE}                         => {MOSTRA ELLIOTT ERWITT}                 1.434340e-05          1 27.846558
6
## [10] {MOSTRA ELLIOTT ERWITT,
##       MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO EGIZIO,
##       PALAZZO REALE}                         => {MUSEO NAZ. DEL RISORGIMENTO ITALIANO}  1.434340e-05          1 62.276463
6
```

Palazzo Reale

```
##      lhs                                       rhs                                        support confidence       lift
count
## [1]  {MUSEO DELLA FRUTTA,
##       PALAZZO REALE}                        => {MUSEO ANTROP. CRIMINALE CESARE LOMBROSO} 4.303019e-05          1 210.206533
18
## [2]  {MUSEO DELLA FRUTTA,
##       MUSEO DI ANATOMIA UMANA LUIGI ROLANDO,
##       PALAZZO REALE}                        => {MUSEO ANTROP. CRIMINALE CESARE LOMBROSO} 2.868679e-05          1 210.206533
12
## [3]  {MUSEO ANTROP. CRIMINALE CESARE LOMBROSO,
##       MUSEO DI ANATOMIA UMANA LUIGI ROLANDO,
##       PALAZZO REALE}                        => {MUSEO DELLA FRUTTA}                    2.868679e-05          1 285.536519
12
## [4]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO DIOCESANO DI TORINO,
##       PALAZZO REALE}                        => {MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA} 1.195283e-05          1   8.535391
5
## [5]  {MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO DELLA FRUTTA,
##       PALAZZO REALE}                        => {MUSEO ANTROP. CRIMINALE CESARE LOMBROSO} 1.673396e-05          1 210.206533
7
## [6]  {MUSEO ANTROP. CRIMINALE CESARE LOMBROSO,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       PALAZZO REALE}                        => {MUSEO DELLA FRUTTA}                    1.673396e-05          1 285.536519
7
## [7]  {MOSTRA ELLIOTT ERWITT,
##       MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       PALAZZO REALE}                        => {MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA} 1.912453e-05          1   8.535391
8
## [8]  {MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       PALAZZO REALE}                        => {MOSTRA ELLIOTT ERWITT}                 1.912453e-05          1  27.846558
8
## [9]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA,
##       PALAZZO REALE}                        => {MUSEO EGIZIO}                          1.912453e-05          1  20.472324
8
## [10] {MOSTRA ELLIOTT ERWITT,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA,
##       PALAZZO REALE}                        => {MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA} 1.912453e-05          1   8.535391
8
```

Museo Egizio

```
##      lhs                                          rhs                                        support       confidence    lift   count
## [1]  {MOSTRA BORN SOMEWHERE,
##       MUSEO EGIZIO}                           => {MUSEO REGIONALE DI SCIENZE NATURALI}    3.824905e-05        1.0 37.770745    16
## [2]  {MUSEO DIOCESANO DI TORINO,
##       MUSEO EGIZIO,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO}   => {PALAZZO REALE}                          1.673396e-05        1.0 20.644080     7
## [3]  {MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO EGIZIO,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA}             => {PALAZZO REALE}                          1.195283e-05        1.0 20.644080     5
## [4]  {MOSTRA ELLIOTT ERWITT,
##       MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO EGIZIO,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       PALAZZO REALE}                          => {MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA} 1.434340e-05        1.0  8.535391     6
## [5]  {MOSTRA ELLIOTT ERWITT,
##       MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO EGIZIO,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO}   => {PALAZZO REALE}                          1.434340e-05        1.0 20.644080     6
## [6]  {MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO EGIZIO,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       PALAZZO REALE}                          => {MOSTRA ELLIOTT ERWITT}                  1.434340e-05        1.0 27.846558     6
## [7]  {MOSTRA ELLIOTT ERWITT,
##       MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO EGIZIO,
##       PALAZZO REALE}                          => {MUSEO NAZ. DEL RISORGIMENTO ITALIANO}   1.434340e-05        1.0 62.276463     6
## [8]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO EGIZIO,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA,
##       PALAZZO REALE}                          => {MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA} 1.912453e-05        1.0  8.535391     8
## [9]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO EGIZIO,
##       MUSEO NAZIONALE DEL CINEMA,
##       PALAZZO REALE}                          => {MUSEO NAZ. DEL RISORGIMENTO ITALIANO}   1.912453e-05        1.0 62.276463     8
## [10] {MOSTRA ELLIOTT ERWITT,
##       MUSEO EGIZIO,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA}             => {MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA} 2.151509e-05        0.9  7.681852     9
```

Museo nazionale del cinema

```
##      lhs                                          rhs                                          support confidence      lift count
## [1]  {MOSTRA WILDLIFE PHOTOGRAPHER OF THE YEAR,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO NAZIONALE DEL CINEMA}              => {MUSEO REGIONALE DI SCIENZE NATURALI}    1.195283e-05          1 37.770745     5
## [2]  {MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA}              => {PALAZZO REALE}                          1.434340e-05          1 20.644080     6
## [3]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA}              => {MUSEO EGIZIO}                           2.390566e-05          1 20.472324    10
## [4]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO ACCORSI-OMETTO,
##       MUSEO NAZIONALE DEL CINEMA}              => {MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA} 1.434340e-05          1  8.535391     6
## [5]  {MOSTRA ROBERT CAPA - RETROSPETTIVA,
##       MUSEO EGIZIO,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA}              => {PALAZZO REALE}                          1.195283e-05          1 20.644080     5
## [6]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA,
##       PALAZZO REALE}                           => {MUSEO EGIZIO}                           1.912453e-05          1 20.472324     8
## [7]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA}              => {MUSEO EGIZIO}                           2.151509e-05          1 20.472324     9
## [8]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA,
##       PALAZZO REALE}                           => {MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA} 1.912453e-05          1  8.535391     8
## [9]  {MOSTRA ELLIOTT ERWITT,
##       MUSEO EGIZIO,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA,
##       PALAZZO REALE}                           => {MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA} 1.912453e-05          1  8.535391     8
## [10] {MOSTRA ELLIOTT ERWITT,
##       MUSEO CIV.  ARTE ANTICA PALAZZO MADAMA,
##       MUSEO NAZ. DEL RISORGIMENTO ITALIANO,
##       MUSEO NAZIONALE DEL CINEMA,
##       PALAZZO REALE}                           => {MUSEO EGIZIO}                           1.912453e-05          1 20.472324     8
```

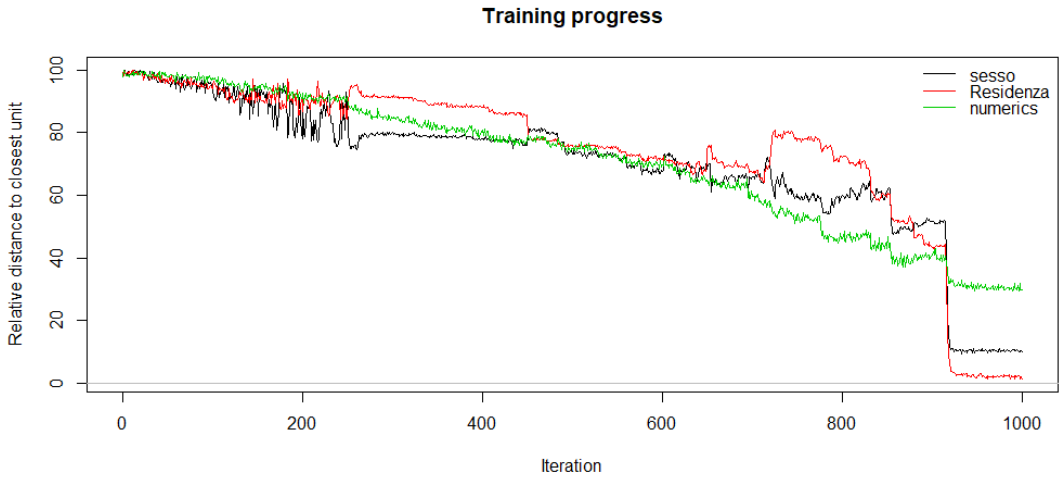
Analyzing the rules and the position of each museum most of the times we can see that all the museums in the itemset are pretty close to each others, especially the ones in the city centre such as “museo egizio”, “palazzo madama”, “palazzo reale” and “museo nazionale del risorgimento italiano”. This can explain most of the rules analyzed.

Customers Segmentation via SOM

Self-Organising Maps (SOMs) are an unsupervised data visualisation technique that can be used to visualise high-dimensional data sets in lower (typically 2) dimensional representations and to create clusters among the observations. In this chapter we will create clusters with SOM.
I must premise that I used the "Tanimoto Distance" to create the categories.

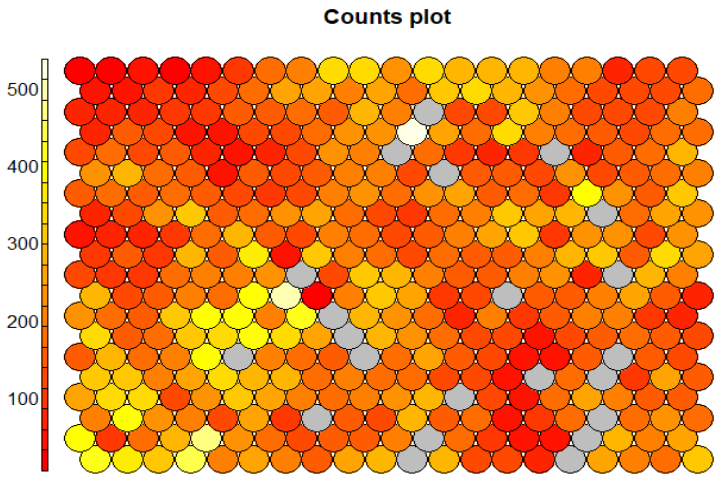
Phase 1: Training Progress

First of all we must train our SOM model: As the SOM training iterations progress, the distance from each node's weights to the samples represented by that node is reduced. Ideally, this distance should reach a minimum plateau. This plot option shows the progress over time. If the curve is continually decreasing, more iterations are required.



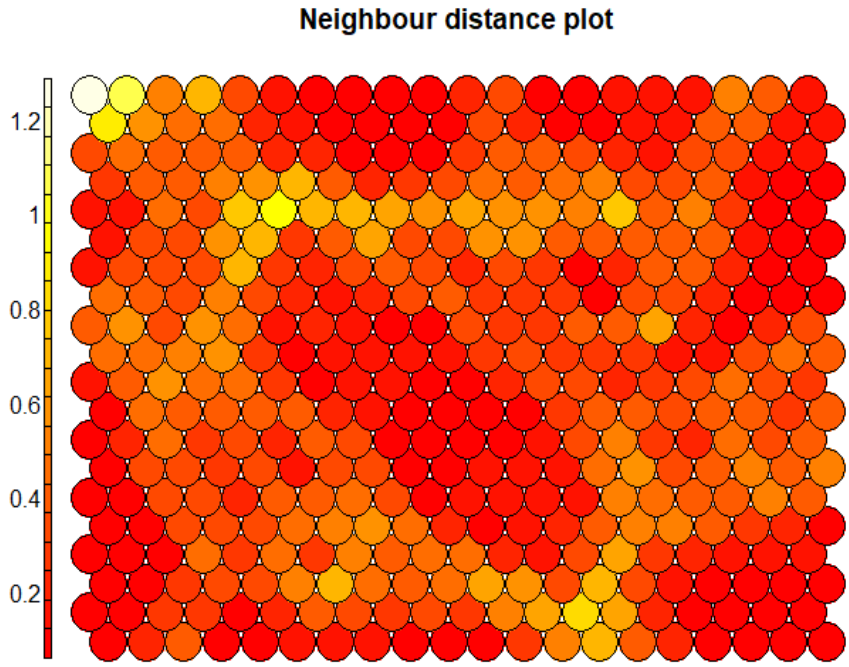
Phase 2: Counts plot

The Kohonen packages in R allows us to visualize the count of how many samples are mapped to each node on the map. This metric can be used as a measure of map quality – ideally the sample distribution is relatively uniform. Our distribution seems quite good.



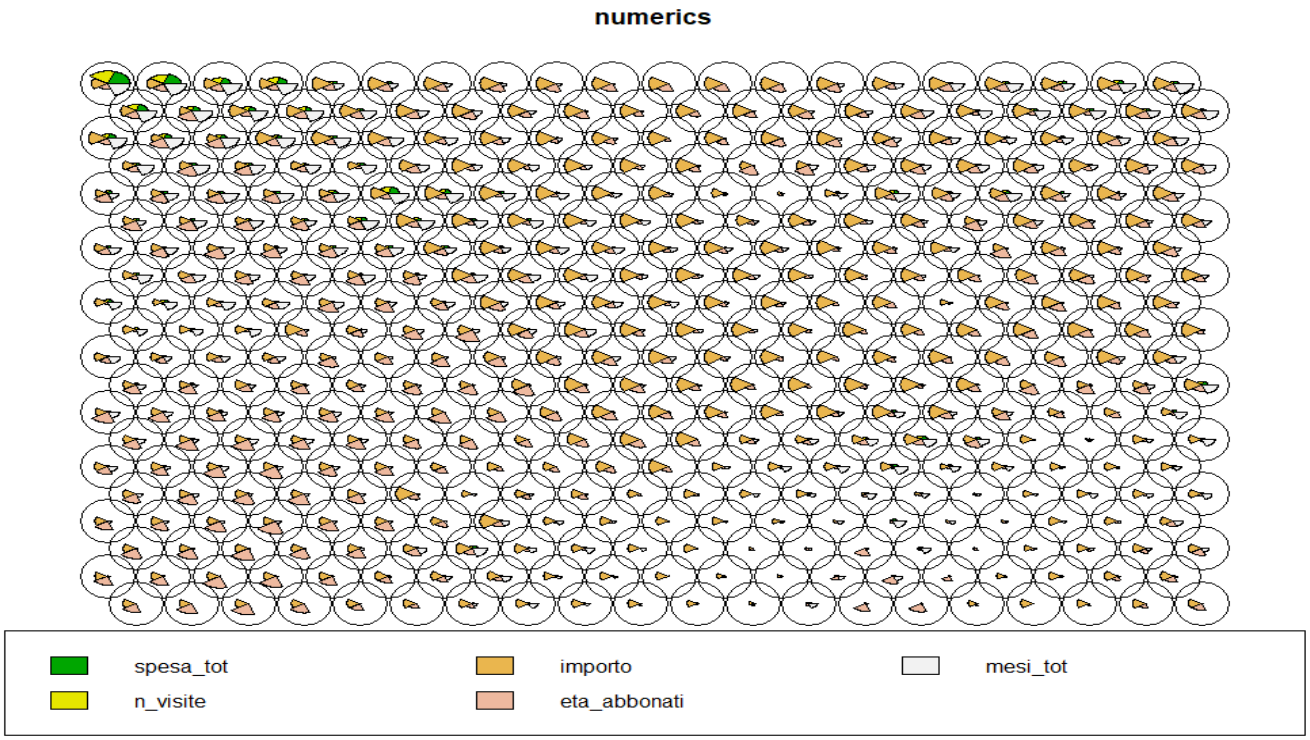
Phase 3: Neighbour distance plot

Often referred to as the "U-Matrix", this visualization is of the distance between each node and its neighbours. Areas of low neighbour distance indicate groups of nodes that are similar. Areas with large distances indicate the nodes are much more dissimilar – and indicate natural boundaries between node clusters.



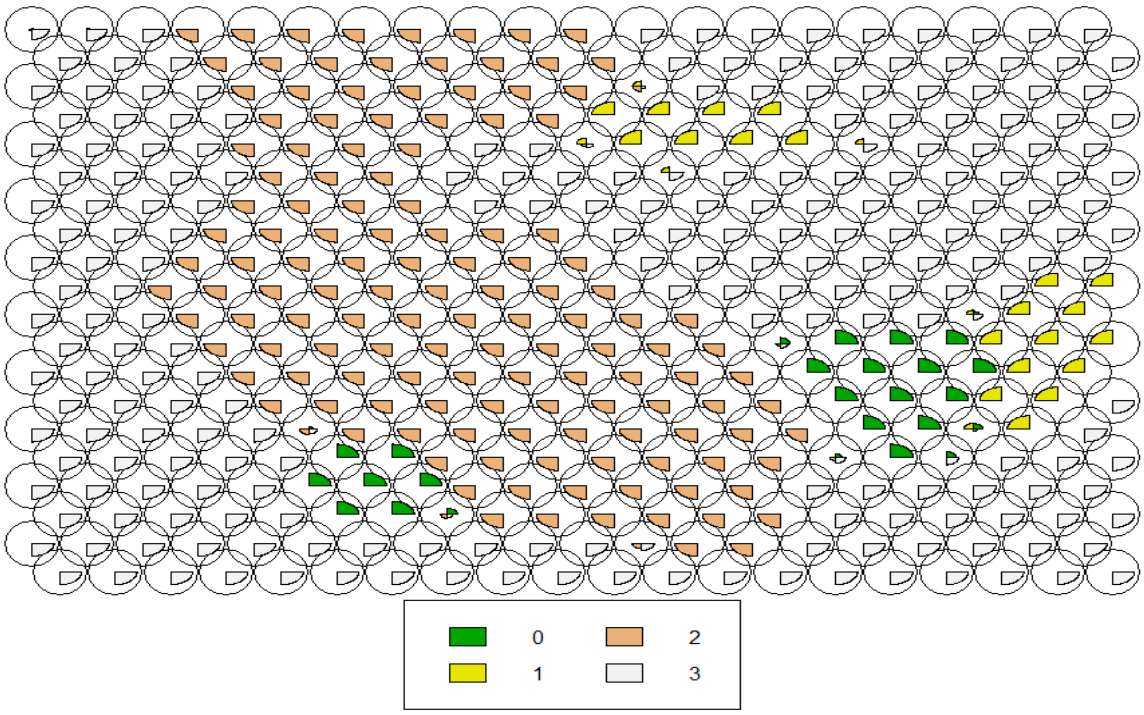
Phase 4: Codes / Weight vectors

The node weight vectors, or "codes", are made up of normalised values of the original variables used to generate the SOM. Each node's weight vector is representative / similar of the samples mapped to that node. By visualising the weight vectors across the map, we can see patterns in the distribution of samples and variables. The default visualization of the weight vectors is a "fan diagram", where individual fan representations of the magnitude of each variable in the weight vector is shown for each node.

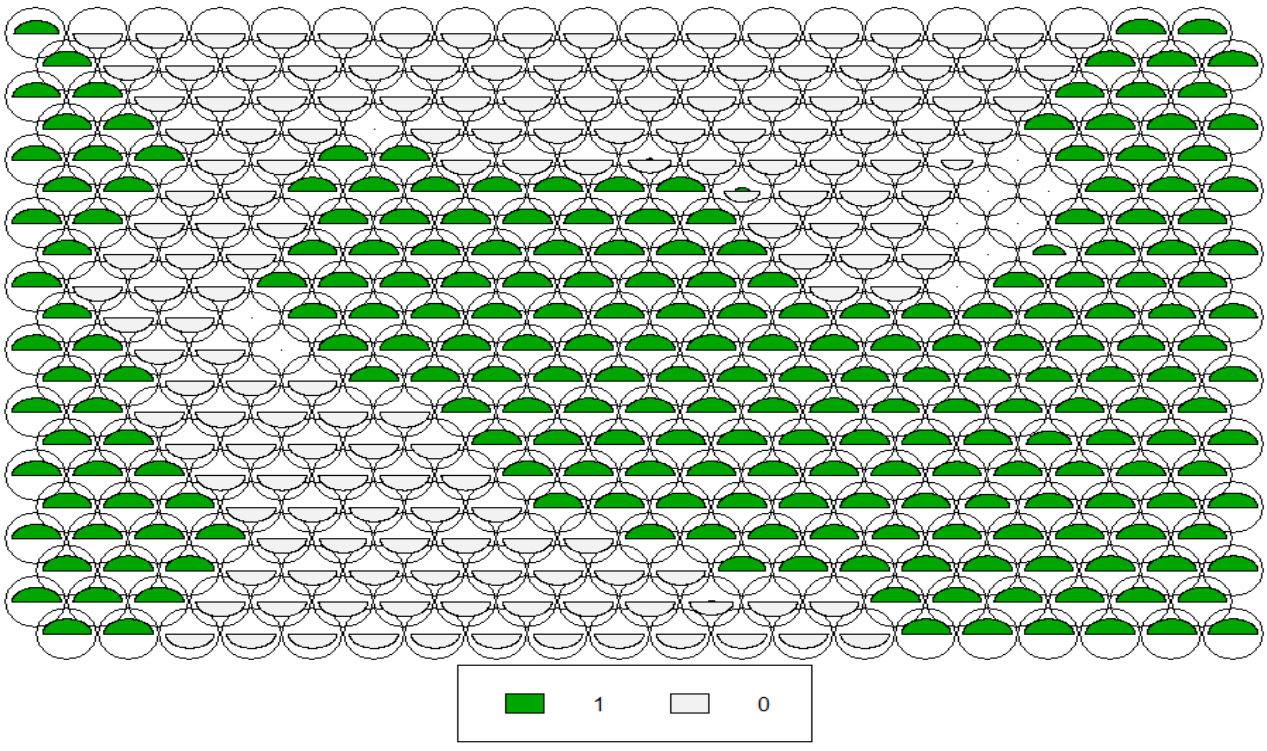


Thanks to the codes plot it is possible to find areas with similar conjunct distribution of variables across the nodes. At the top-left of the map it is possible to see customers with high values in the variables, who are the ones with the lowest probability to churn. Also going from the right toward the center of the SOM map are listed groups with an increasing value of "importo".
The people with the lowest values in the variables, who are the ones with higher probability to churn, are situated at the bottom right of the map.
With SOM is also possible to see how each node vary with each single variable, so undirectly this let us know which variable is the most efficient to create clusters.

Residenza



sesso



Once built the SOM model we must find out what is the optimal number of cluster; to do this I used two methods: the Ward's method (figure 1) and the Silhouette method (figure 2).

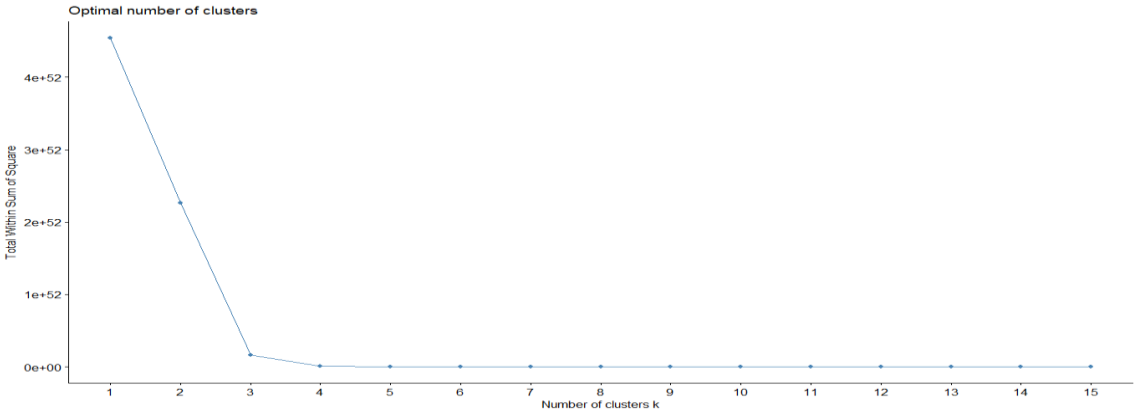


Figure 1

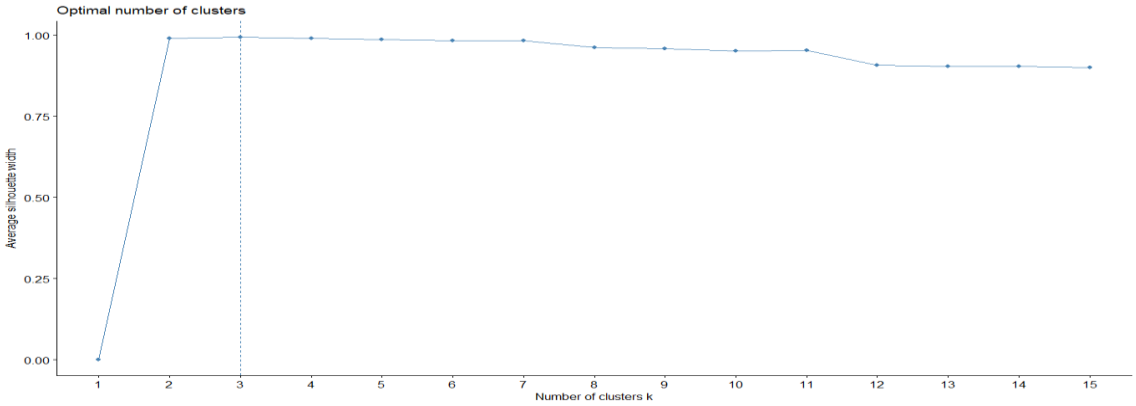
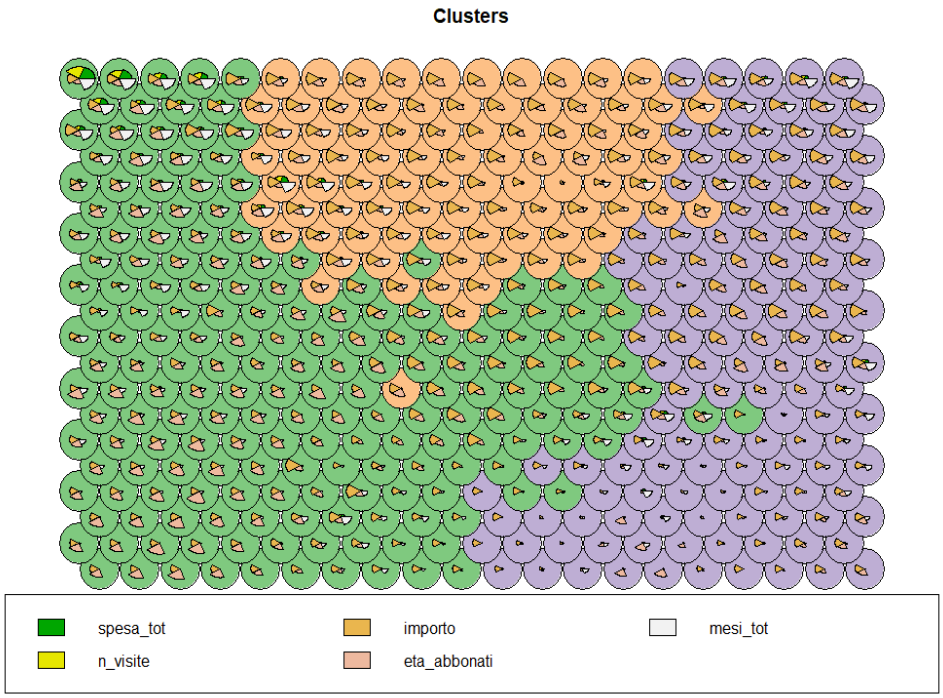


Figure 2

Both methods agree with creating 3 clusters.

Once created these clusters I will use them as a categorical variable for the classification.



Econometrical approach: The Logit Model

Here I will Provide evidence on which variables have the highest impact on the probability to churn, and in order to do this I will use a logit model using as dependent variable "si2014", that is the variable for the Renewal. Let's have a look to the variables used to build the model and to their coefficients. Below there are the coefficients and the variables used:

```
##
## Call:
## glm(formula = si2014 ~ ., family = "binomial", data = train.4)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8759  -1.0558   0.5823   0.8148   1.7489
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.5324320  0.0963112 -15.911  < 2e-16 ***
## spesa_tot     0.0112480  0.0031047   3.623 0.000291 ***
## n_visite     -0.0398362  0.0139522  -2.855 0.004301 **
## importo      -0.0084001  0.0013833  -6.072 1.26e-09 ***
## sesso0       -0.1125778  0.0450700  -2.498 0.012495 *
## eta_abbonati  0.0300023  0.0008936  33.574  < 2e-16 ***
## mesi_tot      0.2611983  0.0125873  20.751  < 2e-16 ***
## Residenza1    0.1708554  0.0772871   2.211 0.027060 *
## Residenza2    0.2923936  0.0604290   4.839 1.31e-06 ***
## Residenza3    0.4316778  0.0598039   7.218 5.27e-13 ***
## cluster2      0.0429661  0.0499659   0.860 0.389840
## cluster3      0.0934486  0.0627562   1.489 0.136468
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 36923  on 31315  degrees of freedom
## Residual deviance: 33168  on 31304  degrees of freedom
## AIC: 33192
##
## Number of Fisher Scoring iterations: 5
```

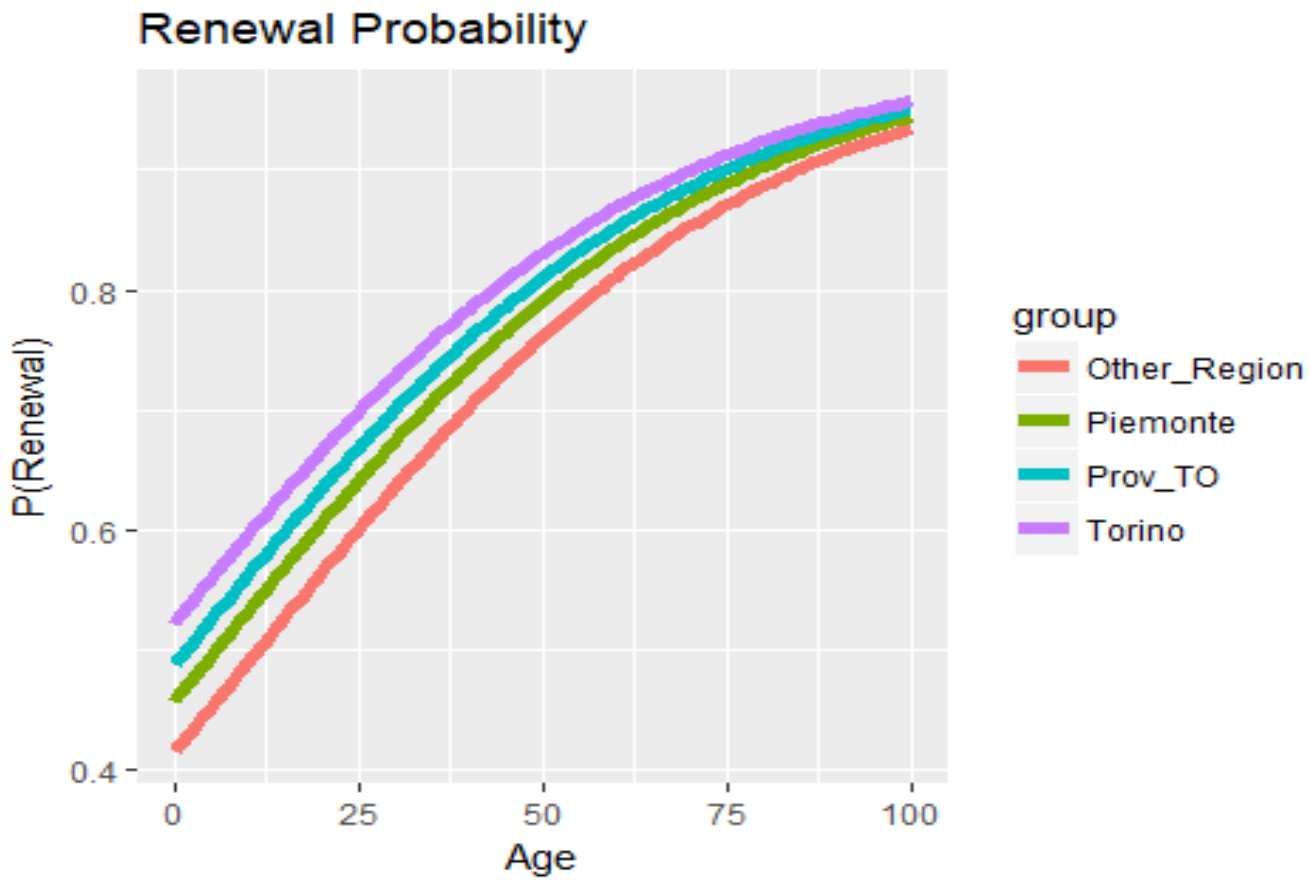
All the coefficients are significative except the clusters created with the SOM; **according to the logit model**, these variables are not really useful to explain the churn.
"Gender" also is not much significative, but we could imagine it since descriptive statistics.

Here Below instead the exponentials of the coefficients.

```
##  (Intercept)    spesa_tot     n_visite      importo       sesso0
##    0.2160097    1.0113114    0.9609469    0.9916351    0.8935278
## eta_abbonati     mesi_tot   Residenza1   Residenza2   Residenza3
##    1.0304569    1.2984851    1.1863192    1.3396303    1.5398389
##     cluster2     cluster3
##    1.0439025    1.0979541
```

The coefficients it selves have no much sense; they became useful when insert in an exponential function, because we can interpret them as an odds-ratio.
It is clear that the most important variable seems being "Residenza", the variable that represent the area in which the customers lives: In fact, who lives in Torino has and Odd to renew 53% higher than who lives outside Piemonte.

In fact as we get closer and closer to Torino we notice that the odds-ratio increases, this influence the probability of renewing.
Also “mesi_tot” seems a good variable to explain the churn: in fact, increasing by one variable the odd of renewal increase by 29,84%. Same for the age of the customers: increasing by one the age the odd of renewal increase by 3 %.



Renewals Predictions

Right now I will show you some models to predict churning; it is a different matter from finding causal relations, as it has been done in the previous chapter.
I used three different methods to do this: Support vector machine (SVM), Random Forest and logit model (that I used before). I used the same variables for all the three models.
Firstly, all the methods were trained on part of the dataset (train dataset) and then tested on the rest of data not used for training (test dataset). In particular I used the cross-validation method for Rainbow Forest and SVM in order to choose the best parameters for the algorithms; before continuing with the analysis I must premise that due to the processor of my PC (an intel i5) I have trained the models on 10'000 obserations. It was not possible for me using more observations.
Let's start comparing the confusion matrix (of the test dataset) for each model and their ROC curves. I used a threshold of 0.5 to do the classification.
The *confusion matrix* (or *error matrix*) is one way to summarize the performance of a classifier for binary classification tasks. This square matrix consists of columns and rows that list the number of instances as absolute or relative "actual class" vs. "predicted class" ratios.
Let P be the label of class 1 and N be the label of a second class or the label of all classes that are *not class 1* in a multi-class setting.

		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Logit model

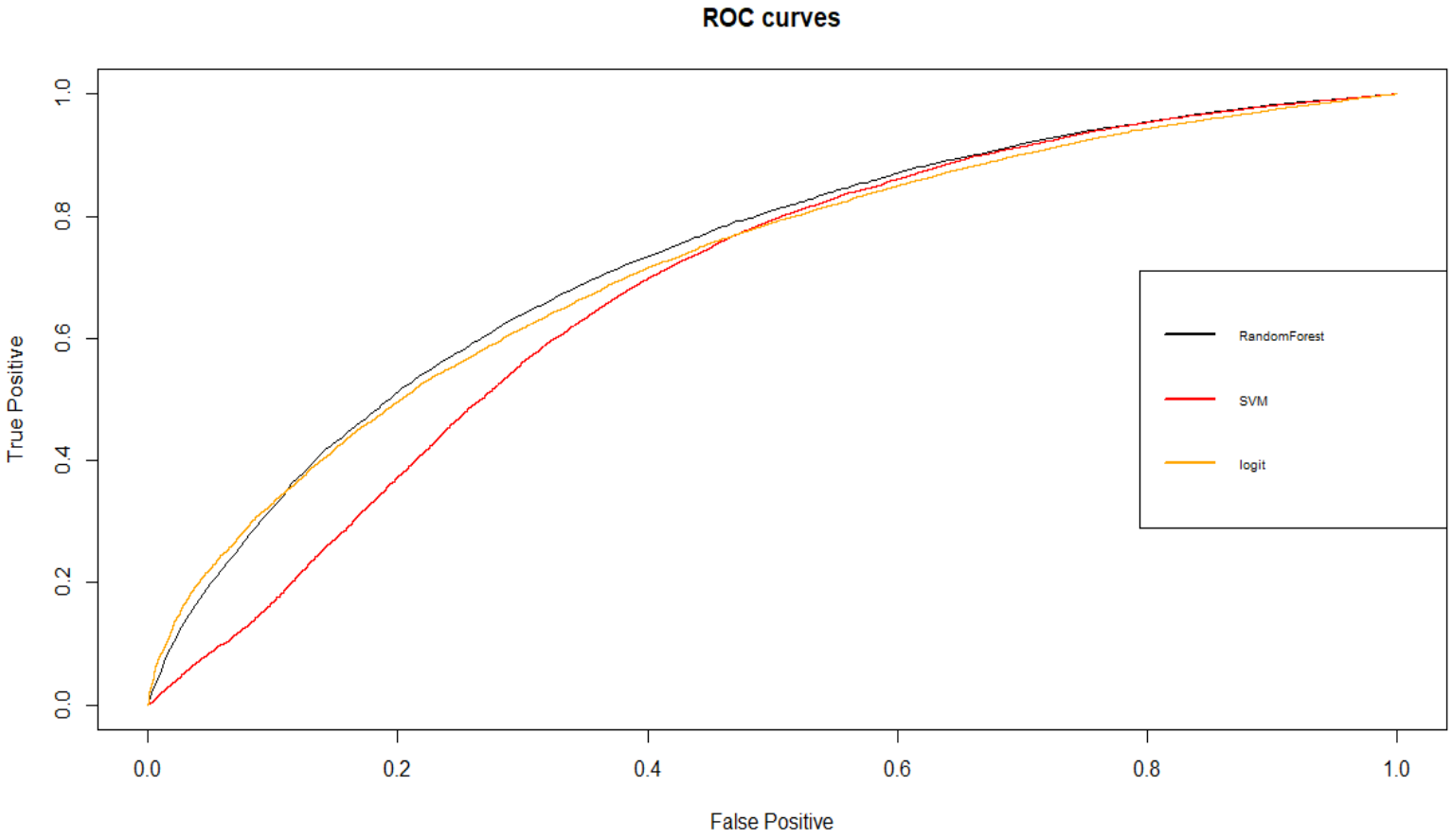
	False	True
False	4657	5956
True	4821	22840
Accuracy model : 0.718425		

SVM model

	False	True
False	2828	1975
True	7785	25686
Accuracy model : 0.7449966		

Random Forest Model

	False	True
False	3081	2152
True	7532	25509
Accuracy model : 0.7462823		

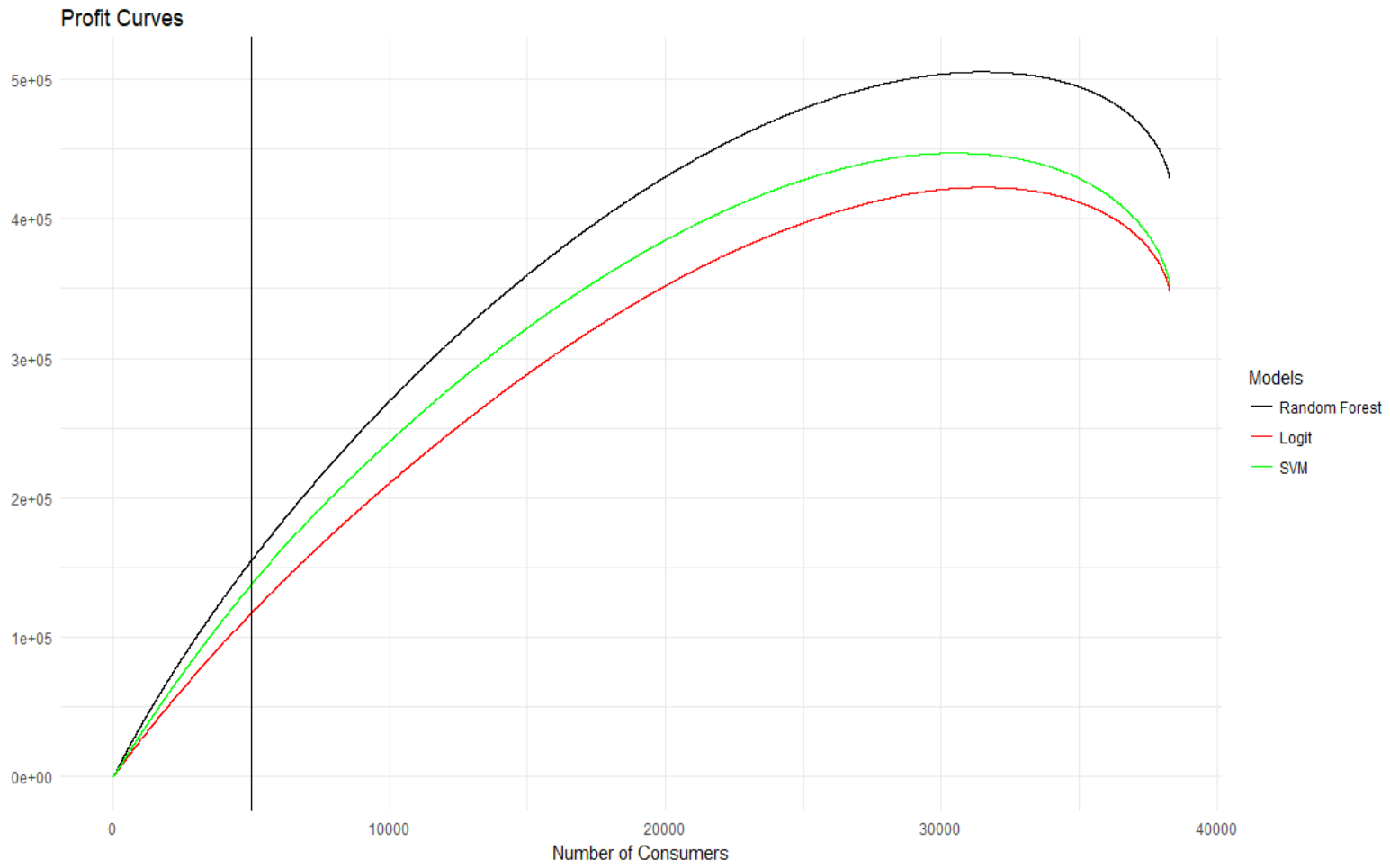


It seems that Random forest and logit models work better than SVM, but it is quite difficult to choose between these two the best one.
In order to do this, we must control our constraints.
We should contact customers with higher probability to renew and ask them if they want to renew their subscription.
Our budget is 5000 euro, and Each phone call has a variable cost of 1 euro (line and operator).
In this context we should choose the model that maximize our profits.
To calculate the expected profit by each person I decided to use the expected value of a binomial distribution with "p" as probability to renew.

$$E(X_i) = p_i(a_i - b_i) + (1 - p_i)b_i$$

$p_i = probability\ to\ renew\ for\ person\ "i"$
$a_i = income\ related\ to\ person\ "i"$
$b_i = costs\ related\ to\ person\ "i"$
$X_i = person\ "i"$

In order to calculate the profit I calculated it as $\sum E(X_i)$, the sum of all the expected values of each customer, using as p_i the probability assigned to the customer by the model.
To calculate $\sum E(X_i)$ I previously ordered the $E(X_i)$ from the highest to the lowest.
According to these assumptions we can contact only 5000 people, so we should contact the 5000 people with the highest probability to renew in each model in order to maximize the profit.
I insert a vertical line on the 5000th customer. We will choose the model with the highest profit on the vertical line.



Model	Expected Profit at the 5000th customer
Random Forest Model	155'074,1
Logit Model	117'327,3
SVM Model	137'820

It seems that the Random Forest model it is the best choice for us.