Statistical Learning and Data Analysis

FATER CHALLENGE

Prof ROBERTA SICILIANO

Group:

Seyed Sadegh Elmi Mousavi (D03000009)

Seyedeh sara Hashemi (D03000036)

Sahel Memariani (D03000025)

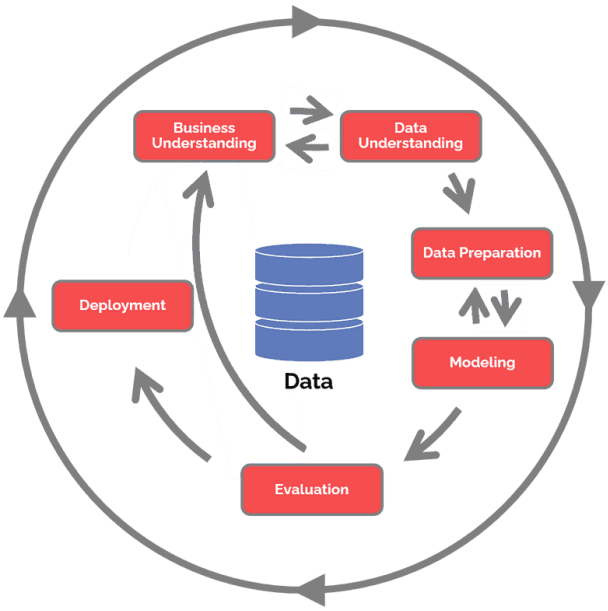
Fatemeh Rahimi (D03000027)

First of all, the cross-industry standard process for data mining (CRISP-DM) methodology is introduced in this report in order to systemize the study and make it clearer to perform based on this process in the next stages.

CRISP-DM methodology

CRISP-DM method is one of the most powerful and common methods, which is based on problem solution strategy. This technique presents a process model for reviewing the lifecycle of each data mining project, and it has six stages, including business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

Understanding the business problem is the first phase in which the business environment is got to be known. In the first place, the problem is defined according to the requirements of the organization so that the data mining methods can handle it. The data understanding is the second phase. The first action in this phase is data gathering. Once the data is gathered, initial processing is carried out on these datasets so that the data are fully identified. The third phase is the data preparation, where the preprocessing and preparation of the data is done in order to apply the data mining algorithm in this phase. Data preprocessing contains four stages, including data cleansing, data reduction, data integration, and data transformation. The lost values are removed or replaced with suitable values in the first stage, and the noises are removed. The repeated rows are removed or sampled in the second stage. Datasets of the locations and different resources are integrated into the third phase. Data normalization is carried out in the fourth stage, and the type of data is transformed. The fourth phase is modeling, which is the major goal of the data mining process, and suitable techniques regarding the subject are employed to analyze data and extract knowledge out of it. Some techniques require special data forms. Hence, sometimes it is needed to return the data preparation phase. The fifth phase is devoted to the evaluation of the results. Once the results of the previous phase are obtained, they are evaluated so that their efficiency is measured. In this phase, the obtained results are evaluated to make sure that these results are consistent with the defined goals of the project. The sixth phase is the implementation of the model, and the outcomes of the actions in the previous phases are obtained. This phase is focused on the application of the obtained knowledge in the business processes so that the business problems are resolved.



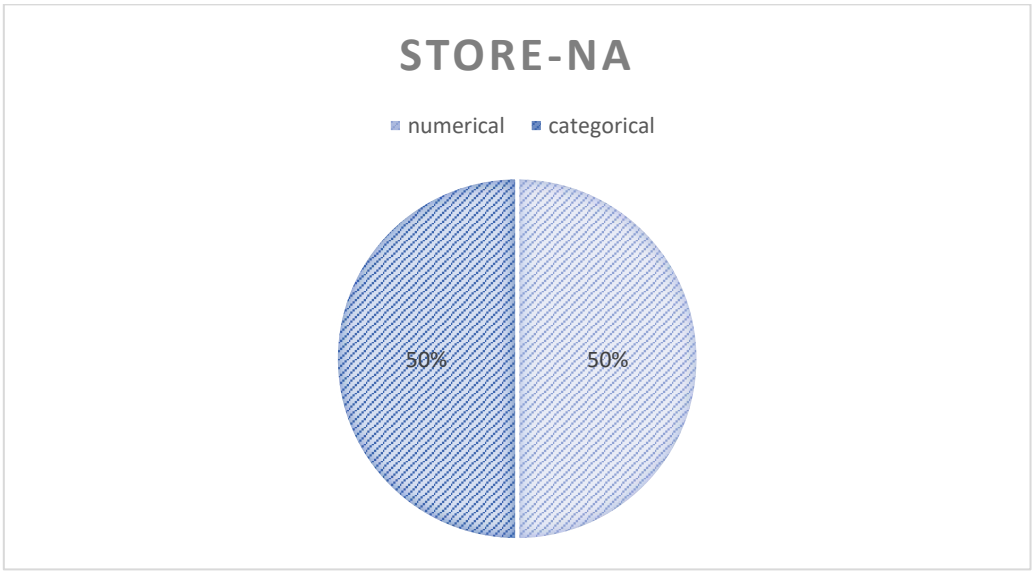
The CRISP-DM method

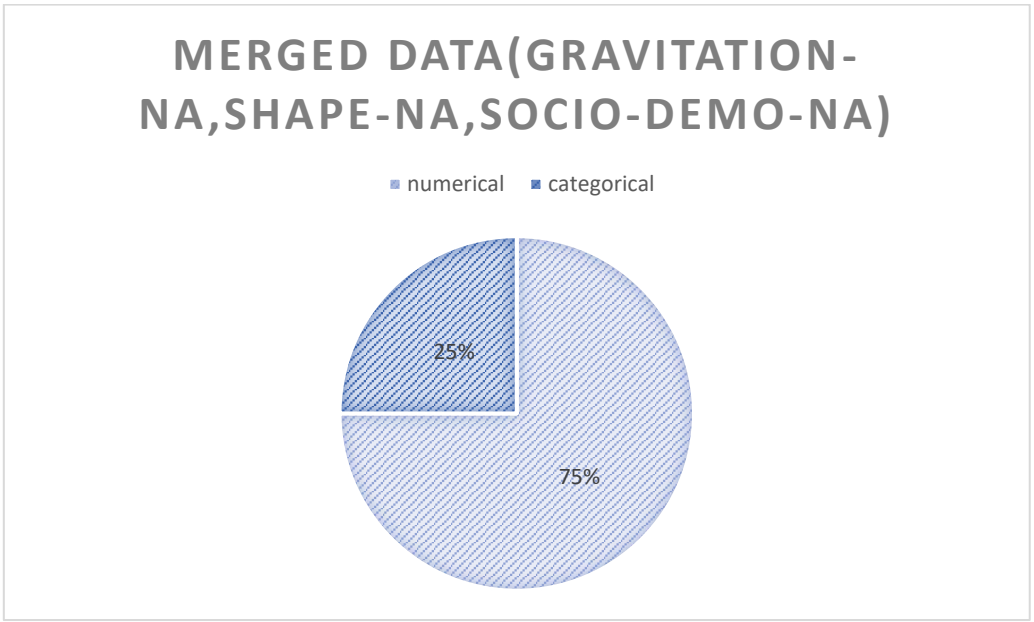
Business understanding

This challenge asks us about estimating store potential revenue, we try to identify more critical stores to improve their performances also more potential stores to boost sales. In this case study we are processing Naples store (entire province) for selling Diapers. we have four main datasets which we explain more about them in below.

Data understanding

In the process of scrutinizing datasets, our focus gravitated toward the pivotal "microcode" column, which surfaced as a common thread weaving through three distinct Excel files: "gravitation_NA," "shapes_NA," and "socio_demo_NA." Due to the inherent significance of this shared attribute, we created an mdf dataset. So now we have two main dataset which we will work on them mdf dataset and store-NA dataset.
Here we created a pie chart to have an overview of the datatypes how much of them is numerical and how much of them is categorical.





Store and merge data pie charts

Summary of raw data

Field name	Length	Class	Mode
datatype	806160	character	character
geometry	10077	character	character
District	10077	character	character
Province	10077	character	character
region	10077	character	character
Insegna	1270	character	character
TipologiaPdV	1270	character	character
Parking	1270	character	character
Indirizzo	1270	character	character
Comune	1270	character	character

Summary of raw data

Field name	Min	1st Qu	Median	Mean	3rd Qu	Max
x	33941450	34289480	34644932	34638766	34985824	35333854
microcode	6.300e+11	6.304e+11	6.305e+11	6.305e+11	6.306e+11	6.309e+11
daytype	1.0	1.0	1.5	1.5	2.0	2.0
fasciaoraria	2	3	4	4	5	6
Media-annuale	1.00	11.00	29.00	66.25	73.00	4903.00

Gravitation_NA

Field name	Min	1st Qu	Median	Mean	3rd Qu	Max
Microcode	6.300e+11	6.304e+11	6.305e+11	6.305e+11	6.306e+11	6.309e+11

Shapes_NA

Field name	Min	1st Qu	Median	Mean	3rd Qu	Max
Microcode	6.300e+11	6.304e+11	6.305e+11	6.305e+11	6.306e+11	6.309e+11
Population	0.0	68.0	186.0	307.0	428.0	4197.0
Population-m	0.0	33.0	90.0	149.4	206.0	2089.0
Population-f	0.0	35.0	96.0	158.3	221.0	2108.0
Population-age-0-4	0.00	3.00	8.00	14.19	19.00	348.00
Population-age-5-14	0.0	7.0	20.0	33.7	45.0	642.0
Population-age-15-34	0.00	16.00	45.00	77.22	103.00	1163.00
Population-age-35-44	0.00	9.00	26.00	43.05	58.00	831.00
Population-age-45-54	0.00	10.00	28.00	47.15	65.00	670.00
Population-age-55-64	0.00	8.00	23.00	38.88	54.00	491.00
Population-age-65-up	0.00	10.00	31.00	53.55	78.00	791.00

Socio_demo_NA

Field name	Min	1st Qu	Median	Mean	3rd Qu	Max
Cod3HD	198	11466	19384	20225	24803	51371
MQVEND	100.0	150.0	250.0	417.1	488.8	11978.0
Lat	40.55	40.83	40.86	40.85	40.91	41.00
Long	13.86	14.21	14.27	14.29	14.38	14.58
Potenziale	0.00100	0.00200	0.00500	0.01431	0.01300	0.83000

Stores_NA

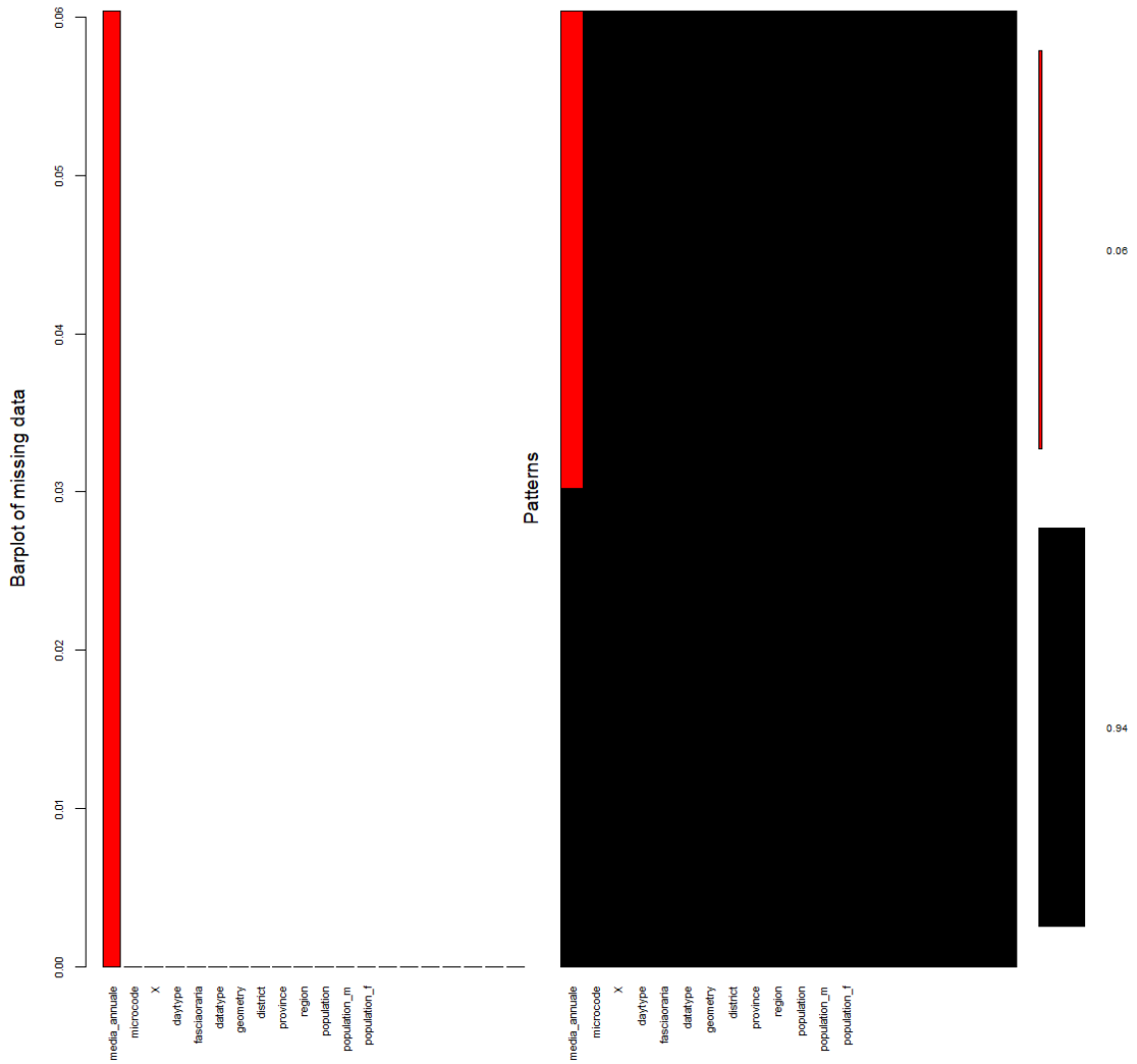
Data preprocessing

Duplicate

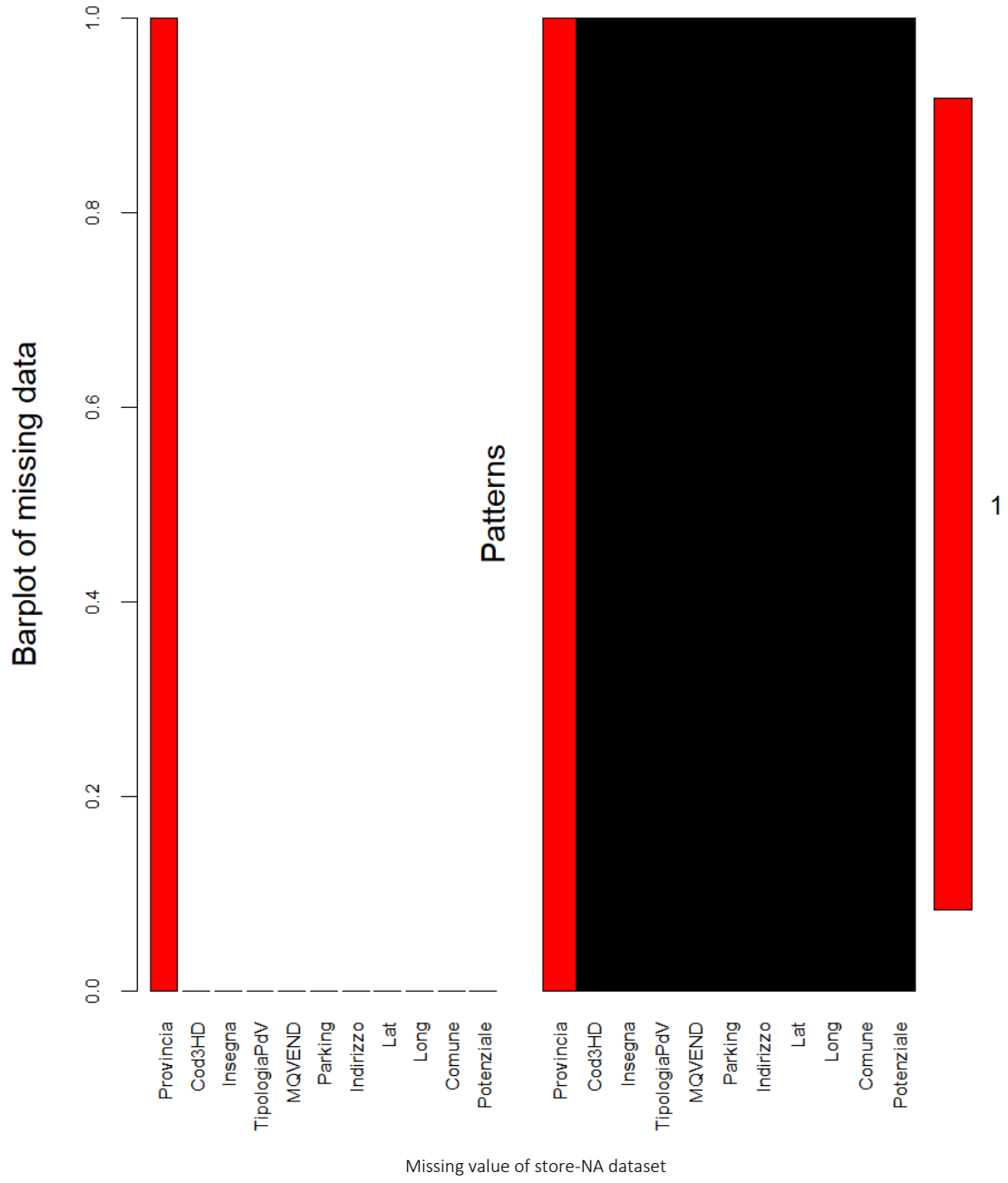
In the next phase of our project, we meticulously checked all columns of the dataset for duplicate rows. There were no duplicate rows found as a result of this rigorous analysis.

Missing values (VIM)

The provided R code uses functions from the VIM (Visualization and Imputation of Missing Data) package to create a visualization of missing data patterns in a data frame.
resulting visualization show a bar plot of missing data for each variable in the dataset, with black and red bars representing non-missing and missing data, respectively. Additionally, it will display patterns of missing data across variables.

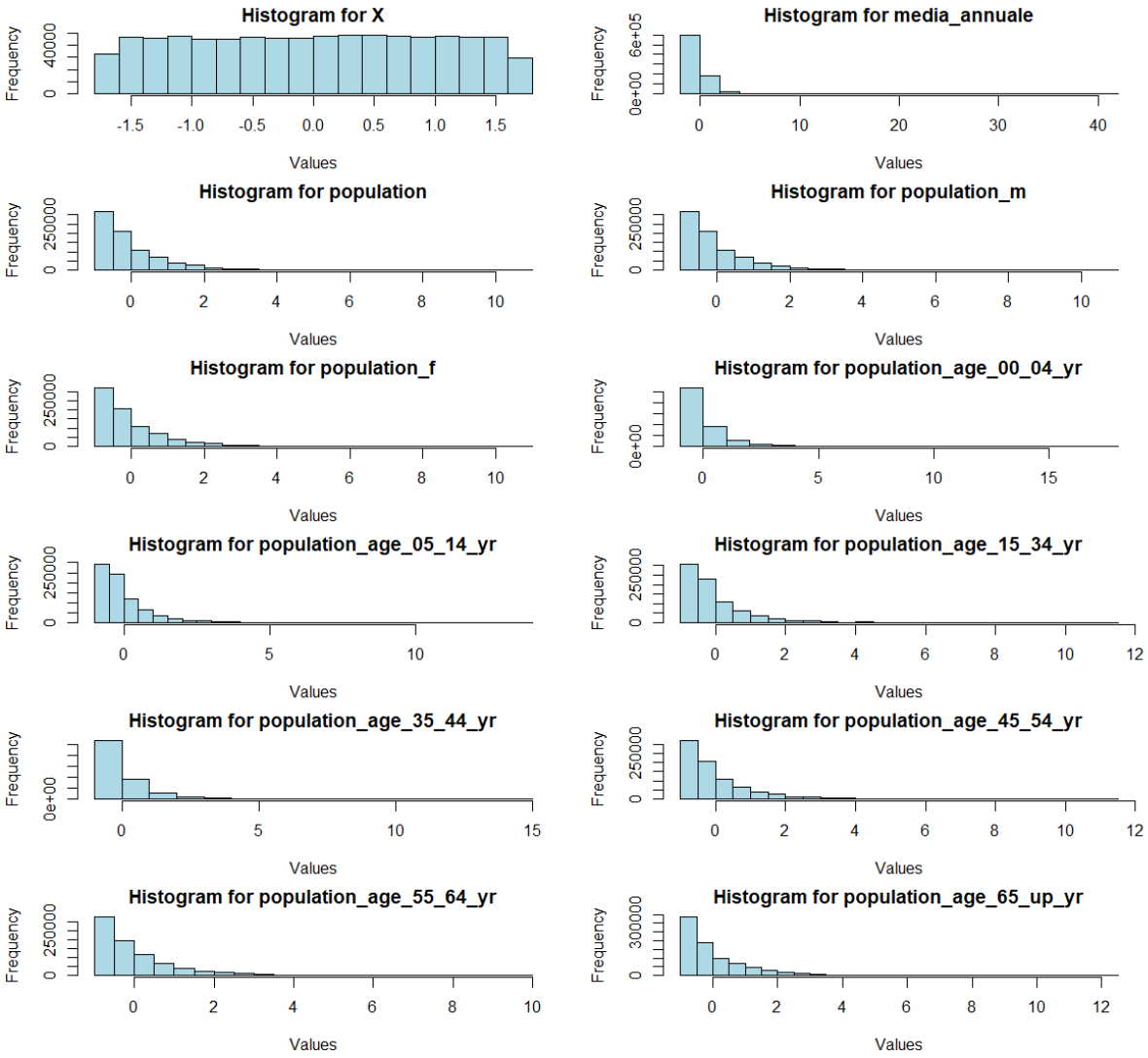


Missing value in mdf dataset

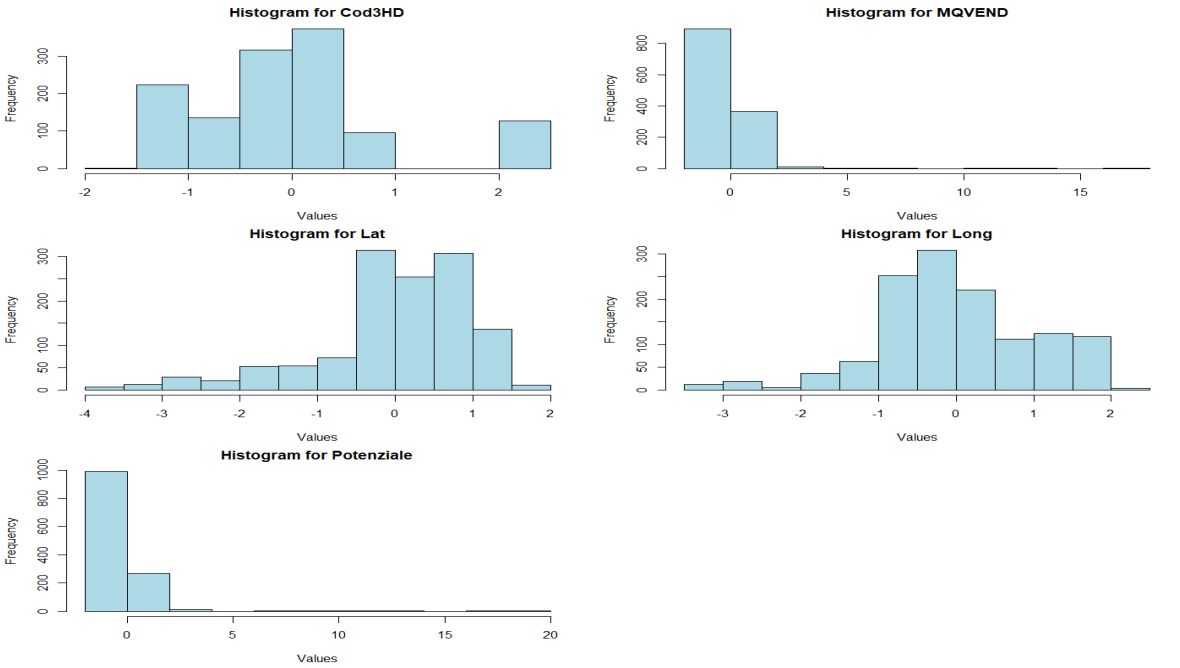


Impute the missing values with mean

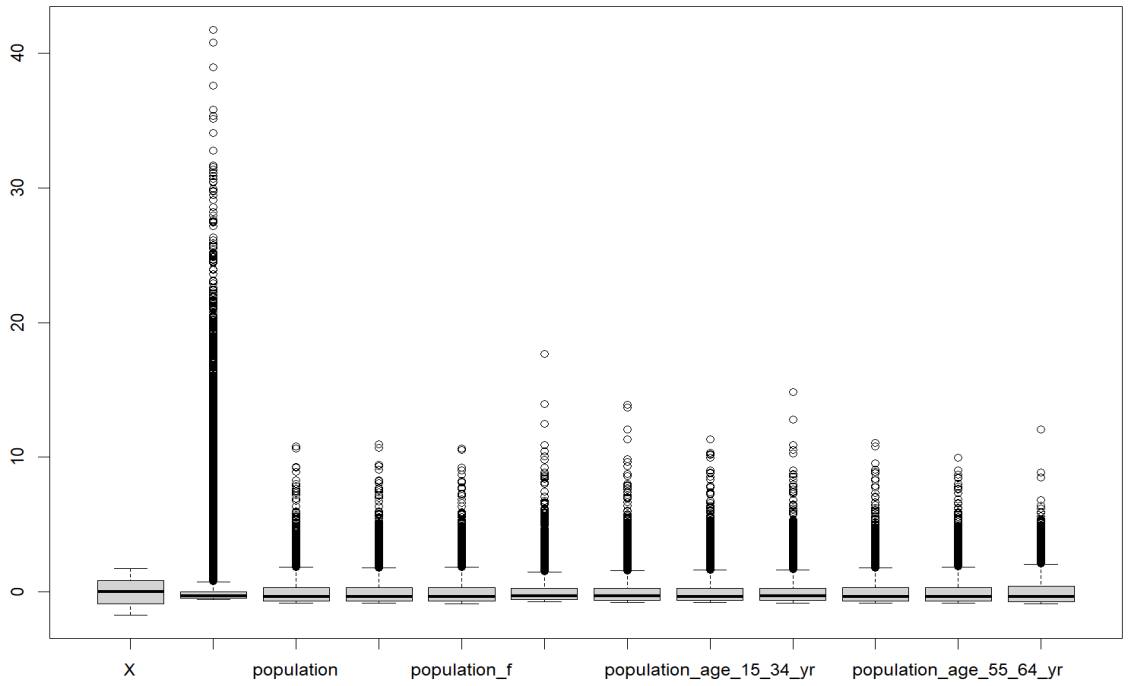
In this case, the missing values in the media_annuale column in mdf dataset and provincia column in store_na dataset are replaced with the mean of the non-missing values in the same column. Mean imputation is a simple method but should be used cautiously as it assumes that missing values are missing completely at random and that the mean is a representative value for imputation.
This can be a common strategy for handling missing data, especially when the missing values are assumed to be missing at random. After imputing missing values, we separate the categorical columns and make a Histogram plots for numerical data.



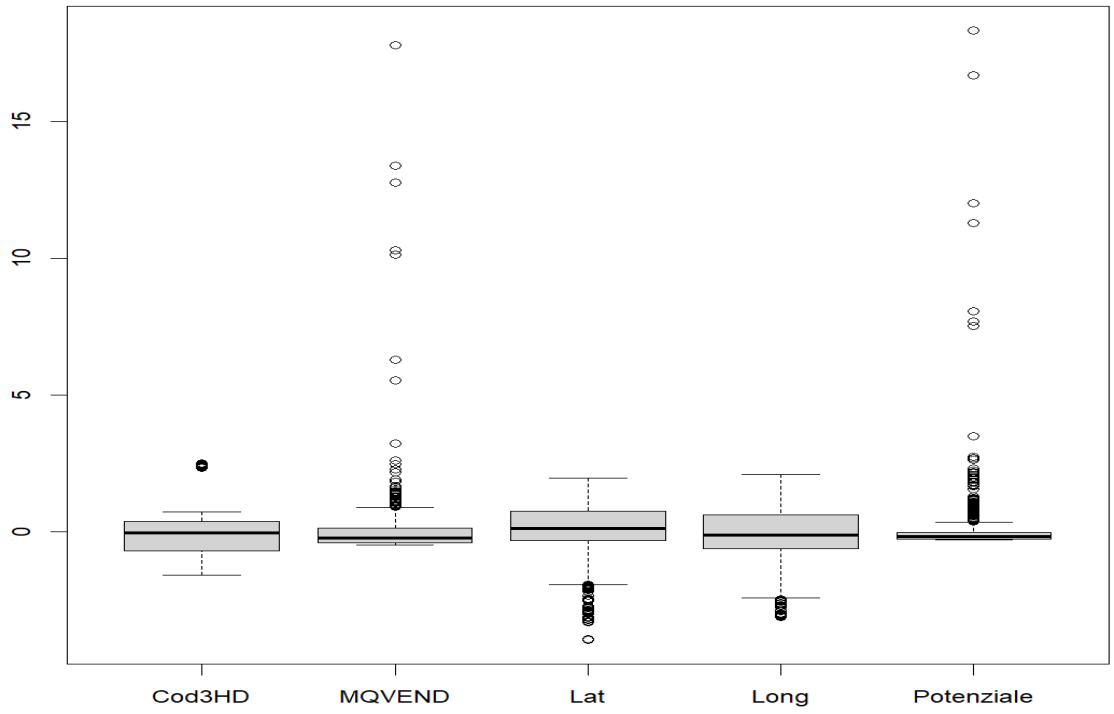
Histograms for all mdf features before cleaning



Histograms for store before cleanin

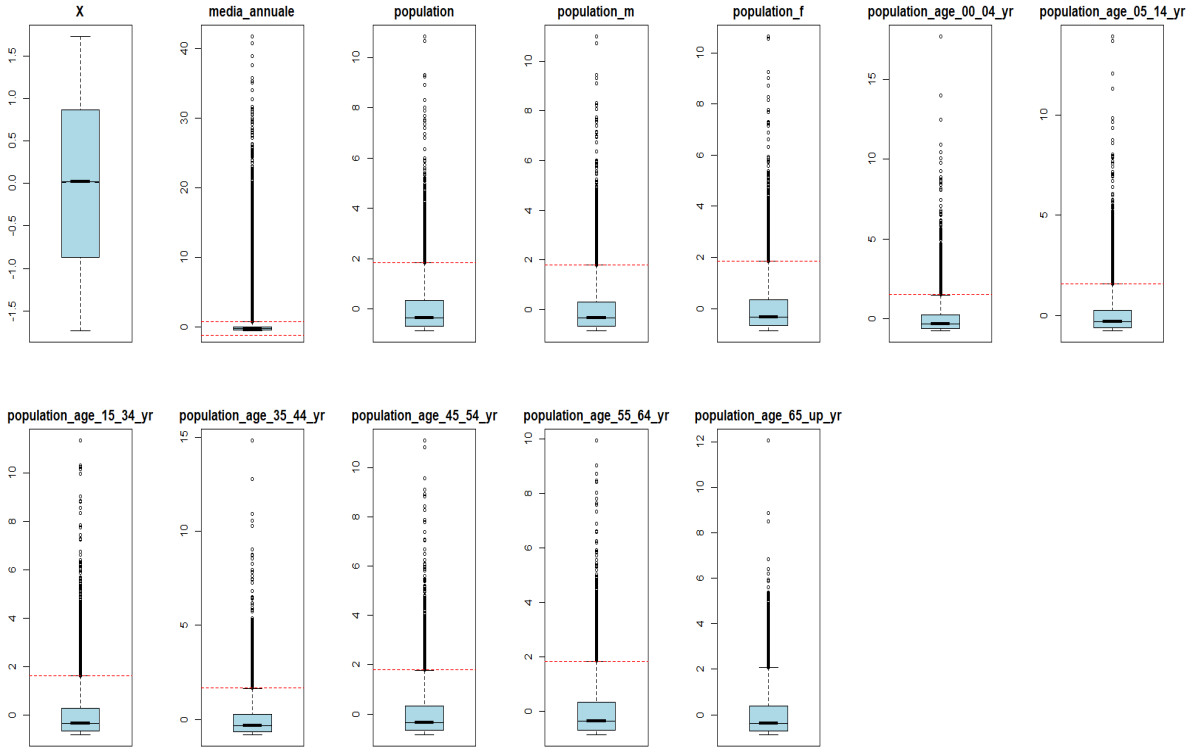


Boxplot(mdf)

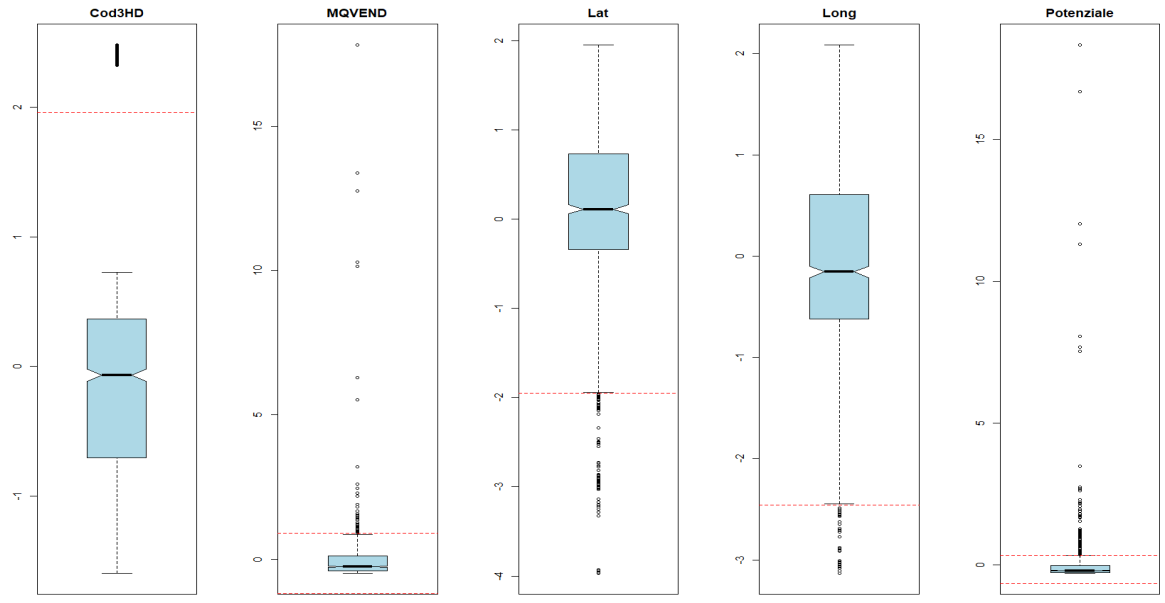


Boxplot store

In this step, we standardize the variables in our mmdf data frame, then convert the standardized matrix into a data frame (mmdf). Then, boxplots are generated to visualize the distributional characteristics of each standardized variable. In our statistical analysis, standardization is often performed as a preprocessing step to ensure that our variables are on a common scale, making it easier to compare and interpret their relative importance.
Media-annuale is the column with the most outliers in the box plot for each feature.



Boxplot with red line for each mdf columns

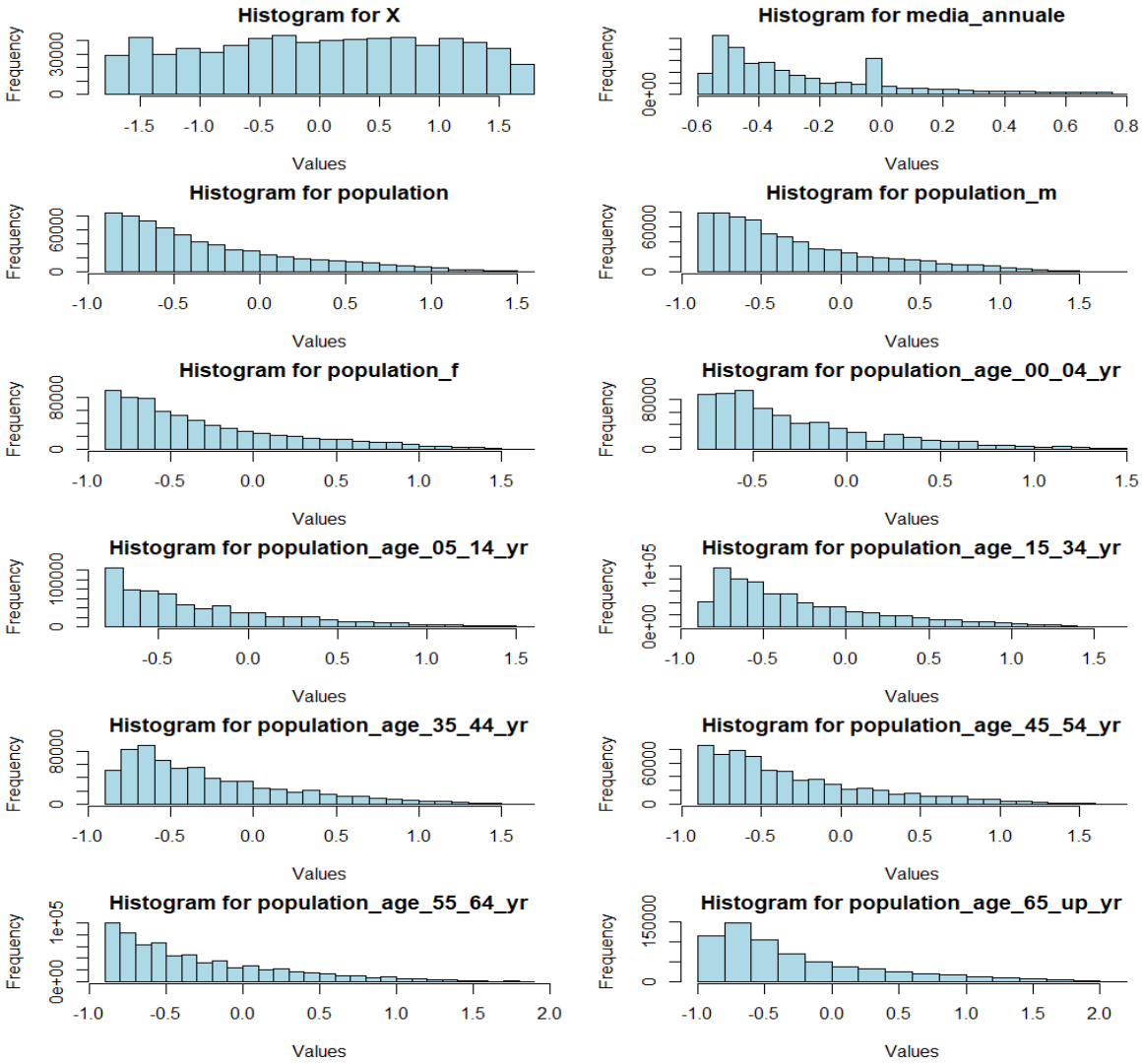


Boxplot with red line for each store columns

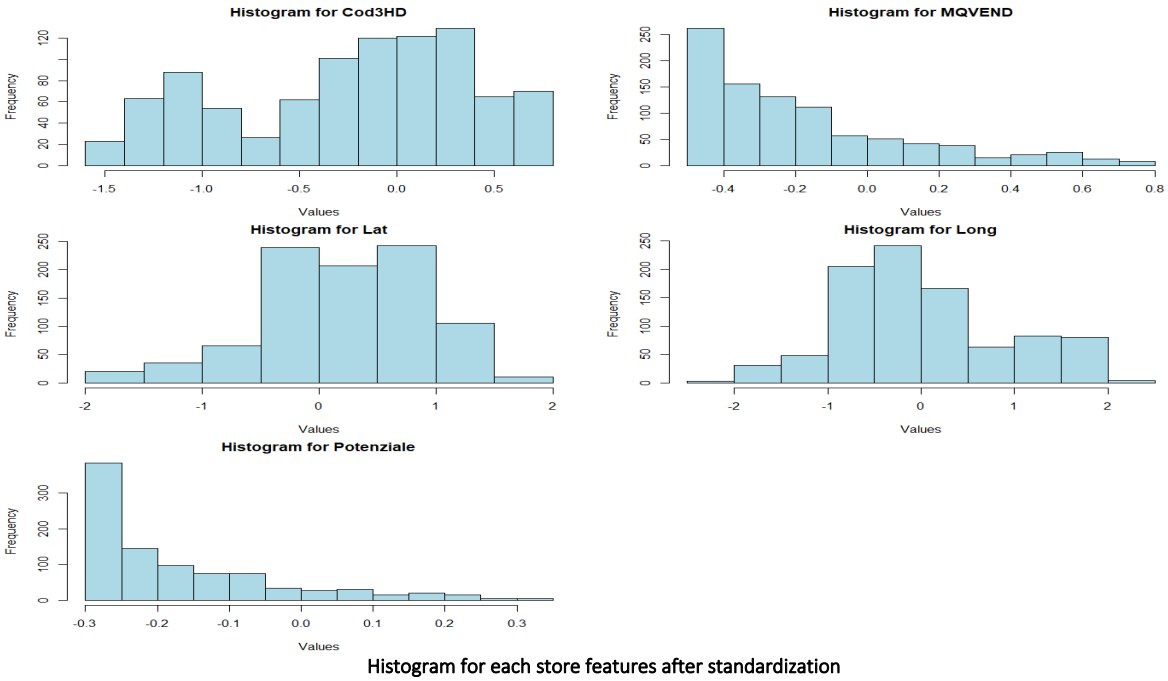
With using IQR methode we try to remove the outliers.Then we make a box plot for each feature in our cleaned dataset and now we can figure out the difference before and after cleaning . the highest and the lowest limit are created to detect outliers for each variable. the points that are upper than U (red line) or lower than L (red line) are outlier.

$$U = Q_3 + 1.5 * (Q_3 - Q_1)\ (1)$$

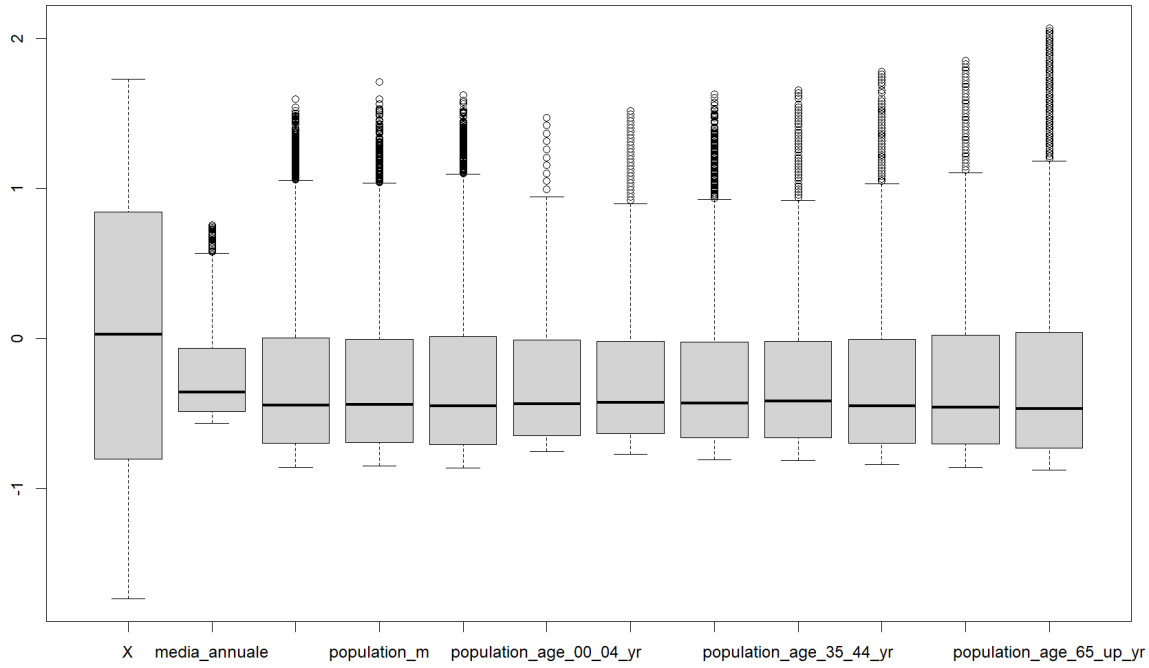
$$\mathrm{L} = Q_1 - 1.5 * (Q_3 - Q_1)\ (2)$$



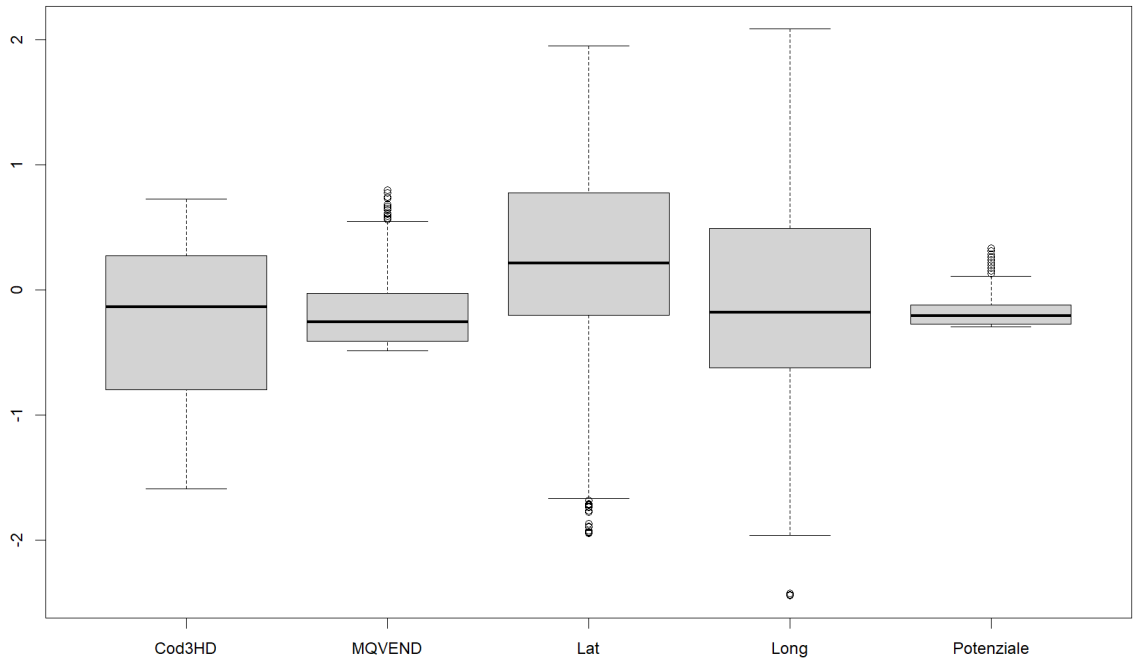
Histogram for each mdf features after standardization



Histogram for each store features after standardization



Boxplot for mdf cleaned data



Boxplot for store cleaned data

Correlation analysis for cleaned-mdf

Correlation analysis is related to finding a relationship between variables and then determining the extent and action of that relationship. Pearson and Spearman correlation coefficients are in the range of +1 to -1. Hence, if the correlation coefficient is near or equal to 1, there is a strong and co-directional relationship between the two variables and they have the same information.

In this condition, if one variable increases, the other one will increase too, and the direction of the variation is the same for the two variables. Besides, when one variable decreases, the other variable will decrease too. In this condition, the two variables have proportional relationships. On the other side, if the correlation coefficient is near or equal to -1, the two variables have a strong but opposite-direction relationship.

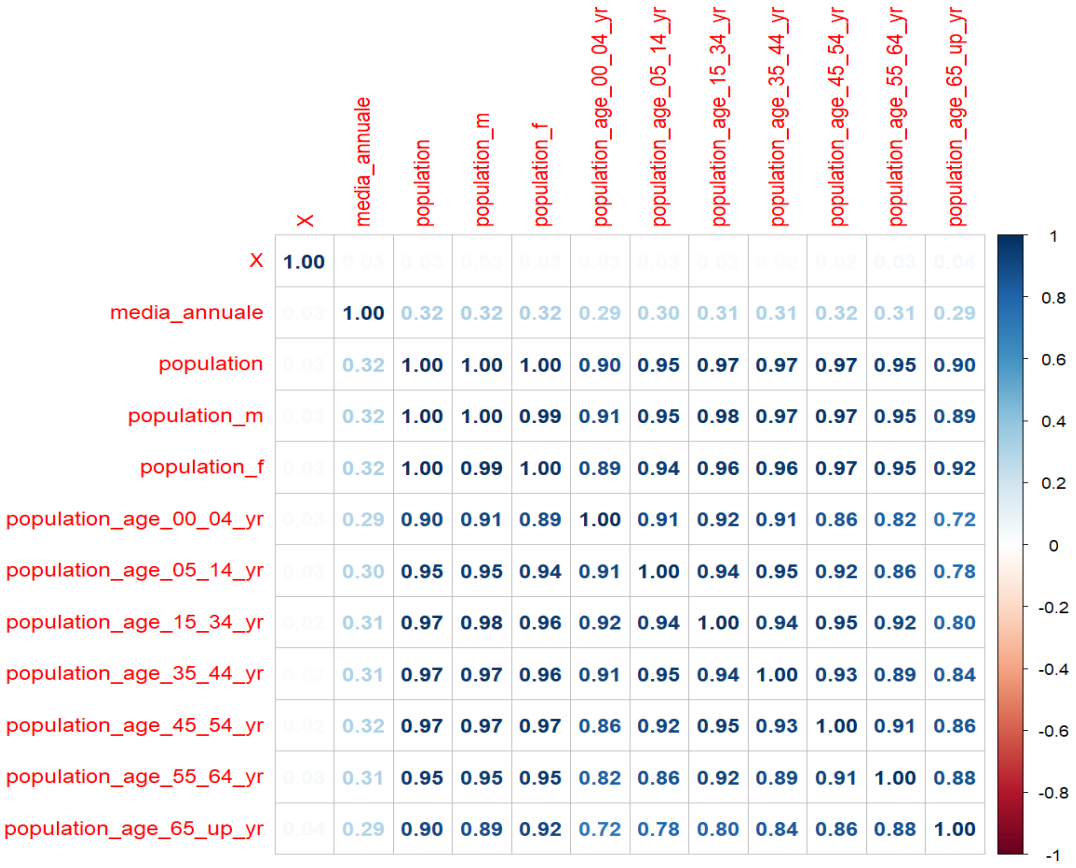
Therefore, if one variable increases, the other one will decrease. However, the prediction is still possible in this condition. This is called the "inverse relationship." In this study, Pearson linear correlation coefficient has been used to measure the linearity of the relationship between two variables, which is expressed in equation (3). $Cov(x,y)$ in equation (3) is the covariance between X and y, and standard deviation of the variables are indicated with σ_x and σ_y, and ρ

$$\rho(X,Y) = r_{xy} = \frac{\mathrm{Cov}\,(x,y)}{\sigma_x \sigma_y} \quad (3)$$

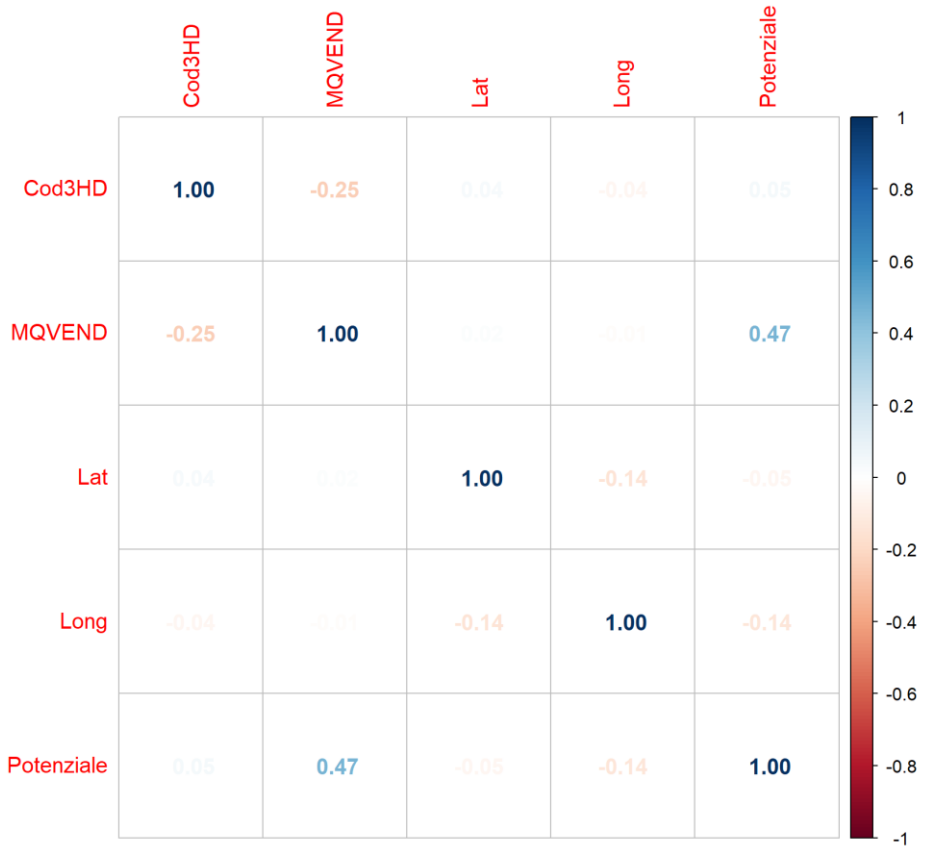
A statistical test with sample statistic with t-student distribution and n-2 degree of freedom has been used based on equations (4) and (5) (n is the number of observations). The chart below demonstrates the results of the correlation matrix.

$$\begin{cases} H_0 & \rho = 0 \\ H_1 & \rho \neq 0 \end{cases} \quad (4)$$

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \sim t(n-2) \quad (5)$$



Correlation matrix for mdf



Correlation matrix for store

The heatmap color gradient indicates the strength of correlations:
Blue: Positive correlations (values close to 1).
White: Little to no correlation (values around 0).
Red: Negative correlations (values approaching -1).
Most cells appear blue, suggesting positive associations.

Correlation matrix for mdf

We take the p-value = 0.05, and the analyze for the correlation matrix of mdf is like below:
Total Population: Positively correlated with media_annuale (blue cell).
As the total population increases, media_annuale tends to increase as well.

Male Population (population_m): Positively correlated with media_annuale (blue cell).
When the male population rises, media_annuale tends to rise too.

Female Population (population_f): Positively correlated with media_annuale (blue cell).
As the female population grows, media_annuale tends to grow as well.

Population by Age Brackets:
Generaly positively correlated with media_annuale (blue cells).

The specific age brackets (e.g., 00-04 years, 05-14 years, etc.) show positive associations with media_annuale.

Correlation matrix for store

We take the p-value = 0.05, and the analyze for the correlation matrix of store is like below:
Cod3HD:
Negatively correlated with MQVEND (correlation coefficient: -0.25).

As Cod3HD decreases, MQVEND tends to increase (and vice versa).

MQVEND:
Positively correlated with Potenziale (correlation coefficient: 0.47).

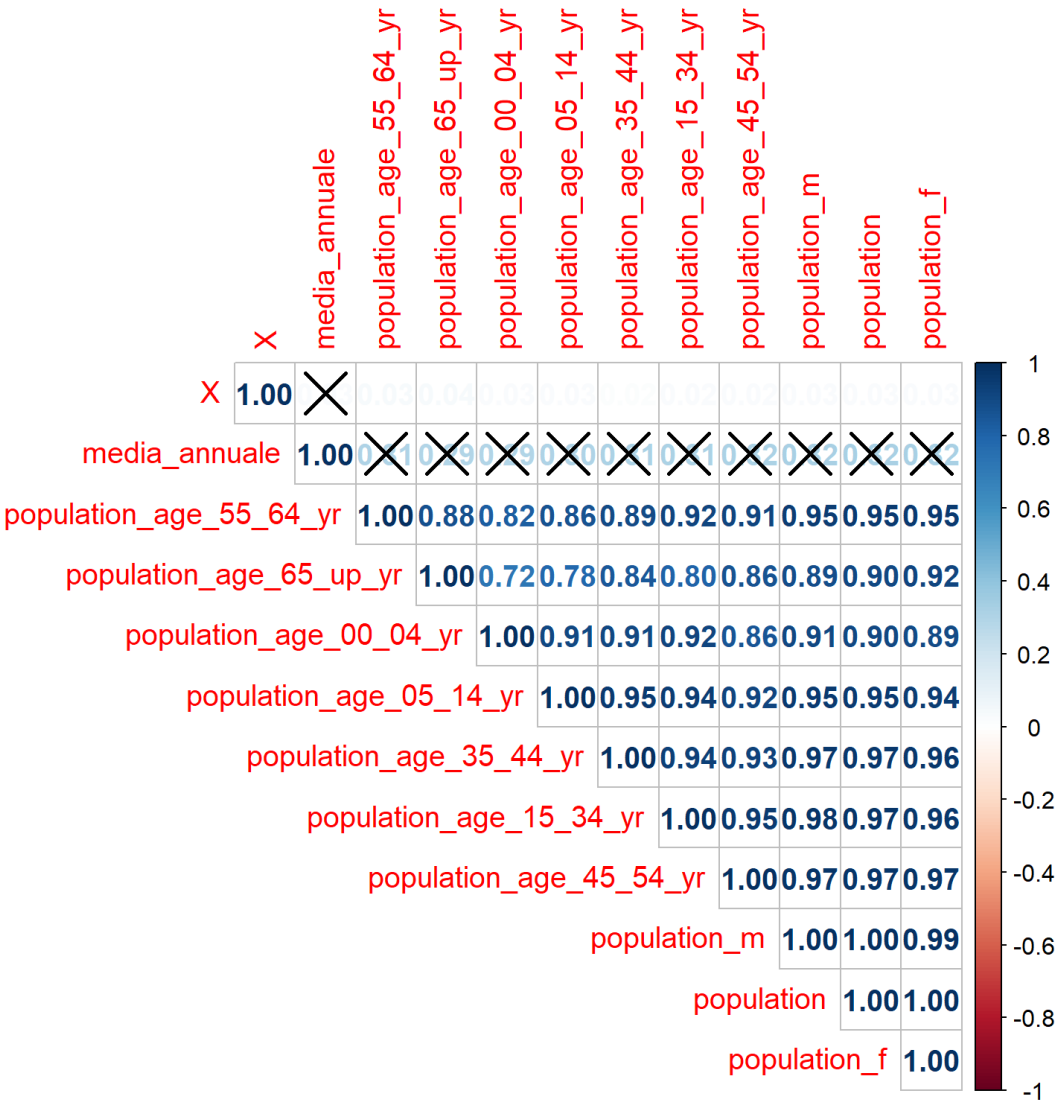
When MQVEND increases, Potenziale tends to increase as well.

Lat and Long:
Negatively correlated (correlation coefficient**:** -0.14).

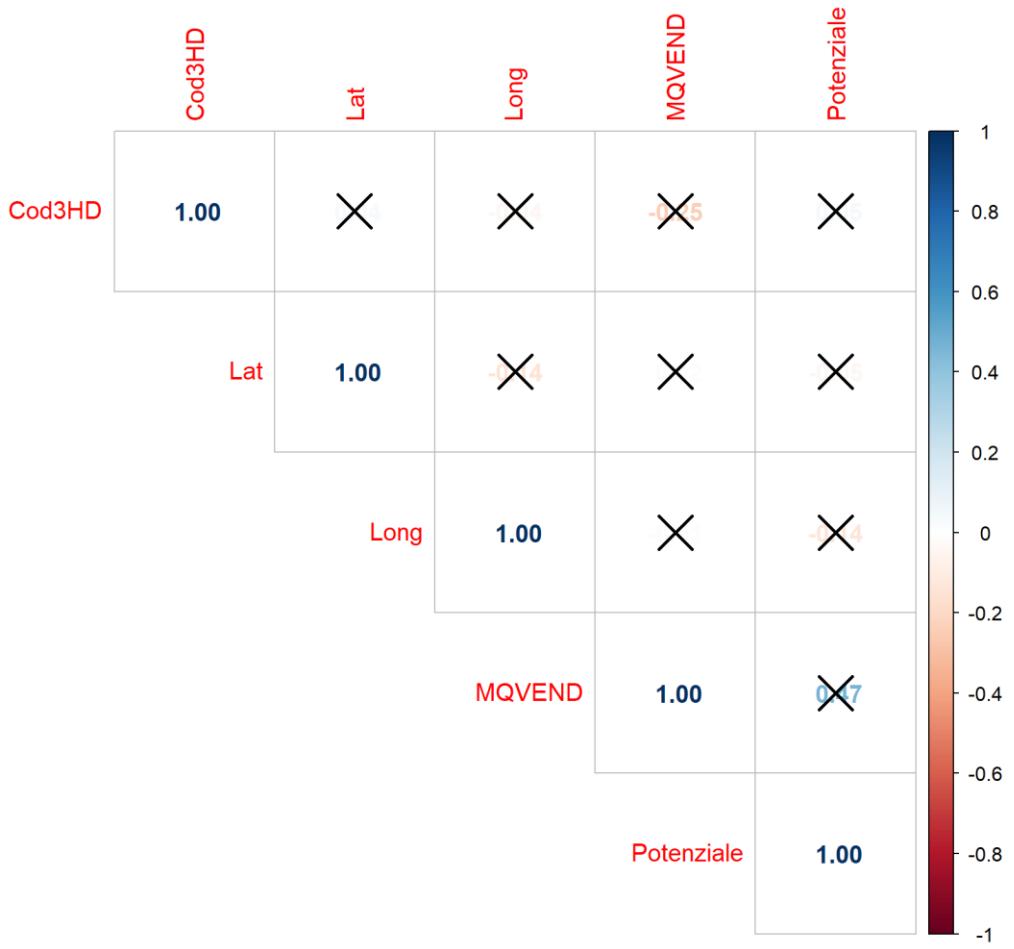
As latitude (Lat) increases, longitude (Long) tends to decrease (and vice versa).

Other Pairs:
The diagonal cells (self-correlations) are all red with a value of 1.00, indicating perfect positive self-correlation for each variable.

Other pairs do not exhibit significant correlations (values close to zero).



Correlation matrix of mdf after statistical test



Correlation matrix of store after statistical test

Hypothesis testing with the matrix of p-values of the correlation involves assessing the significance of correlations between variables in a dataset. The correlation matrix provides a measure of the strength and direction of linear relationships between pairs of variables. However, to determine whether these correlations are statistically significant, hypothesis testing is commonly performed.

Before going to the modeling step, we need to explore the data. After investigating the mdf-cleaned dataset, a crucial step involves group by the data based on the microcode feature to obtain unique rows. This process helps identify distinct sets of data with similar microcode values. To accomplish this, we create a new dataframe, referred to as dfg1. This action effectively organizes the data into groups, with each group representing unique values of the microcode column. Following figure shows the summary of dfg1.

```
       microcode       count_X      mode_daytype mode_fasciaoraria mode_datatype mean_media_annuale           district      population   population_m
630010000001:   1   Min.   :10.00   1:8477       2:8831            F1:9000       Min.   :  6.00     NAPOLI         :3802   Min.   :  0   Min.   :  0
630010000002:   1   1st Qu.:80.00   2: 523       6:  84                          1st Qu.: 20.00     NOLA           : 363   1st Qu.: 58   1st Qu.: 28
630010000003:   1   Median :80.00                3:  30                          Median : 32.00     AFRAGOLA       : 357   Median :154   Median : 74
630010000004:   1   Mean   :74.35                5:  19                          Mean   : 38.42     QUARTO         : 293   Mean   :213   Mean   :103
630010000005:   1   3rd Qu.:80.00                4:  36                          3rd Qu.: 56.00     ACERRA         : 264   3rd Qu.:326   3rd Qu.:157
630010000006:   1   Max.   :80.00                                                Max.   :115.00     TORRE DEL GRECO: 263   Max.   :881   Max.   :451
(Other)     :8994                                                                                   (Other)        :3658
 population_f   population_age_00_04_yr population_age_05_14_yr population_age_15_34_yr population_age_35_44_yr population_age_45_54_yr population_age_55_64_yr
Min.   :  0.0   Min.   : 0.00           Min.   :  0.00          Min.   :  0.00          Min.   :  0.0           Min.   :  0.00          Min.   :  0.00
1st Qu.: 30.0   1st Qu.: 2.00           1st Qu.:  6.00          1st Qu.: 14.00          1st Qu.:  8.0           1st Qu.:  8.00          1st Qu.:  7.00
Median : 79.0   Median : 7.00           Median : 16.00          Median : 37.00          Median : 21.5           Median : 23.00          Median : 19.00
Mean   :110.1   Mean   : 9.54           Mean   : 22.64          Mean   : 52.43          Mean   : 29.4           Mean   : 32.63          Mean   : 27.29
3rd Qu.:168.0   3rd Qu.:14.00           3rd Qu.: 34.00          3rd Qu.: 79.00          3rd Qu.: 45.0           3rd Qu.: 50.00          3rd Qu.: 42.00
Max.   :456.0   Max.   :42.00           Max.   :100.00          Max.   :233.00          Max.   :131.0           Max.   :147.00          Max.   :123.00

population_age_65_up_yr
Min.   :  0.0
1st Qu.:  9.0
Median : 26.0
Mean   : 39.1
3rd Qu.: 58.0
Max.   :180.0
```

Modeling

First, we create a new data frame (dfg1) and we put mdf-cleaned on it. Then groups the data frame dfg1 by microcode column, this means that subsequent operations will be applied within each group defined by unique values of the microcode column.

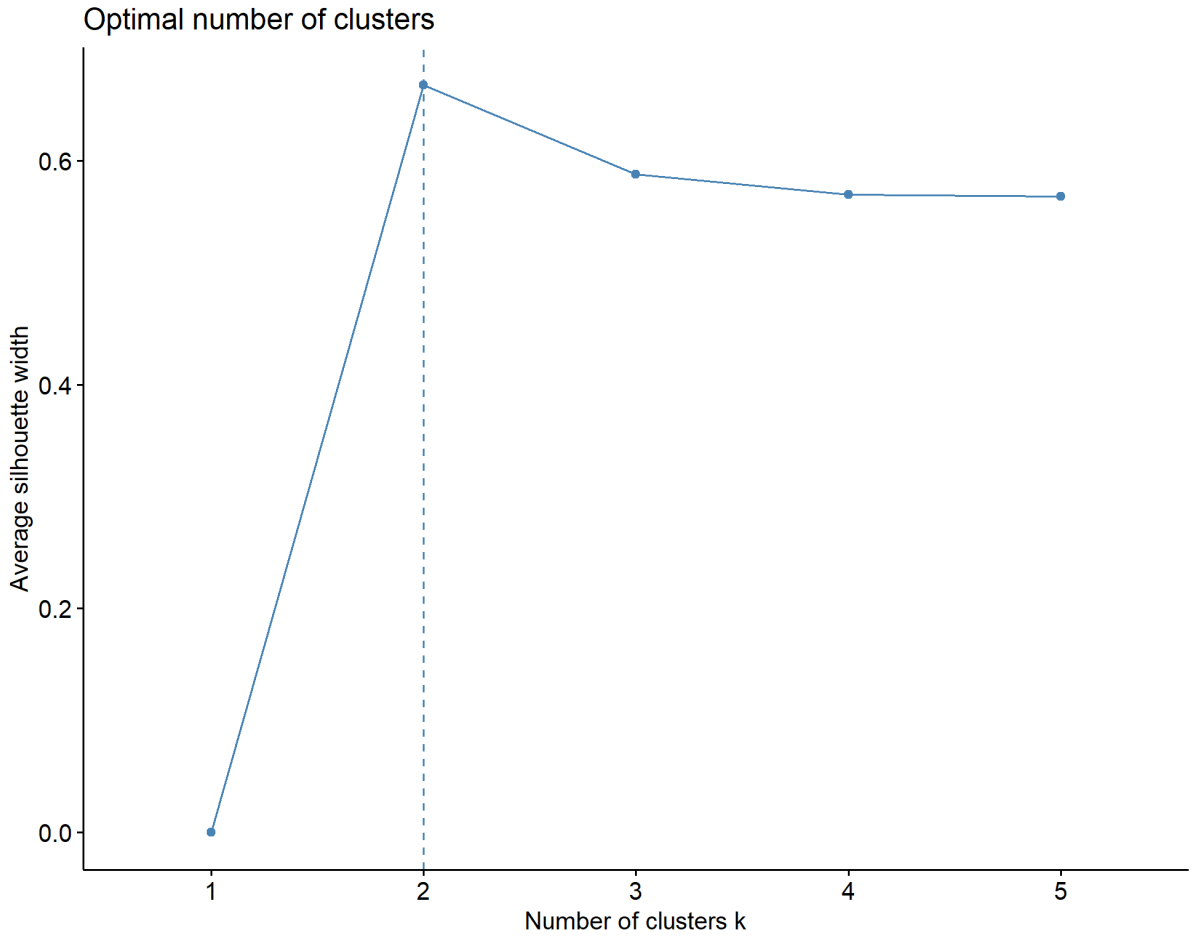
K-means is an unsupervised learning algorithm, meaning it doesn't require labeled training data. It organizes data into clusters based on inherent patterns or similarities. We used the K-means algorithm. The number of the optimal clusters can be understood using following Fig. Silhouette method has been used in this Fig to determine the number of the optimal cluster. This index focuses on the quality of the clustering. In fact, this index defines how the data disperse in clusters, or in other words, it measures how much a point belongs to its cluster in comparison with the adjacent cluster, and it is dependent on the within-correlation and separability of the clusters. This index is in the range of -1 to +1. The more this index gets (nearer to +1), the better the quality of the clustering is. If k clusters that have been created bb the algorithm are indicated with $(C_1, C_2, \dots, C_k)$ and x_i is a point among the clustered data, the Silhouette index for x_i can be obtained by using equation as follows:

$$s(i) = \frac{b(i) - a(i)}{\max(b(i), a(i))}$$

Where $a(i)$ is the average distance between one point of the cluster and other points of the cluster, which can be obtained using the first equation, and $b(i)$ is the minimum average distance between one point and other clusters that can be derived using the second equation.

$$a(i) = \frac{1}{n_i} \sum_{l=1}^{n_i} d(x_i, x_l) \qquad b(i) = \min \frac{1}{n_i} \sum_{y_m \in C_l} d\big((x_i, y_m)\big) \ \forall\, l = 1, \dots, k$$

In this method, the clustering algorithm has been implemented for different values of k ($k = 1, \dots, 5$), and the Silhouette index is calculated for each member of the clusters in every operation. Afterward, the average value of the obtained Silhouette indices is calculated. The optimal value for k is the value that yields the maximum value of the Silhouette index average. In this Fig., this optimal value happens in $k = 3$ (without considering k=2).



Number of optimal clusters based on Silhouette

We have clustered the stores in three clusters according to potenziale (potential for sale) and MQVEND (store size) (Following Fig).

```
> k1$centers
    MQVEND  Potenziale
1 394.1269 0.009108209
2 172.3615 0.004449541
3 709.6455 0.011454545
> k1$size
[1] 268 545 110
```

Result of clustering

According to this figure we figure out that in second cluster we have more potential for sale. The distribution of stores in clusters shows that the second cluster has the best potential although the stores are not big stores, after that the first cluster has better potential and the last one is the third cluster. So in continue we prioritized the clusters in this way => (2,1,3).

Following fig shows the cluster plot.

After clustering, we add the number of clusters to the stored dataset as a new feature and create the new dataset. Then, we merge the new dataset with dfg1 from the last part based on Comune and "district" features. After removing NA rows, we have the cmer dataset. The objective of this section is to establish a connection between the store and mdf datasets.

We can group by the new data set based on the Cod3HD (Store id) feature and create a cmerg dataframe. This action effectively organizes the data into groups, with each group representing unique values of each store. Following figure shows the summary of cmerg.

```
     Cod3HD                  Comune                 Insegna     TipologiaPdV     MQVEND       Parking      Potenziale       cluster_number      mode_microcode mode_daytype mode_fasciaoraria
Min.   :  198   NAPOLI           :313   N. D.           :153   LIS:564      Min.   :100.0   False:619   Min.   :0.001000   1:545          630495100011:313    1:923        2:923
1st Qu.:10188   GIUGLIANO IN CAMPANIA: 48   DECO' SUPERMERCATI: 73   SUP:193      1st Qu.:150.0   True :304   1st Qu.:0.002000   2:268          630340000001: 48
Median :18498   MARANO DI NAPOLI : 31   MD              : 58   DIS: 71      Median :250.0               Median :0.005000   3:110          630410000002: 31
Mean   :17215   CASORIA          : 27   CONAD CITY      : 53   SSD: 95      Mean   :300.8               Mean   :0.006637                  630230000001: 27
3rd Qu.:23616   PORTICI          : 23   CONAD           : 50                3rd Qu.:400.0               3rd Qu.:0.009000                  630590000001: 23
Max.   :29351   POZZUOLI         : 20   PROSHOP         : 38                Max.   :935.0               Max.   :0.029000                  630600000001: 20
                (Other)          :461   (Other)         :498                                                                              (Other)     :461
mode_datatype mean_media_annuale mean_population mean_population_m mean_population_f mean_population_age_00_04_yr mean_population_age_05_14_yr mean_population_age_15_34_yr
F1:923        Min.   : 6.00      Min.   : 91.0   Min.   : 44.0    Min.   : 47.0    Min.   : 4.0                 Min.   : 9.00              Min.   : 22.00
              1st Qu.:37.00      1st Qu.:187.0   1st Qu.: 89.0    1st Qu.: 97.0    1st Qu.: 8.0                 1st Qu.:19.00              1st Qu.: 44.00
              Median :37.00      Median :213.0   Median :105.0    Median :105.0    Median :11.0                 Median :25.00              Median : 58.00
              Mean   :42.95      Mean   :252.8   Mean   :122.5    Mean   :130.1    Mean   :11.5                 Mean   :27.08              Mean   : 63.31
              3rd Qu.:49.00      3rd Qu.:311.0   3rd Qu.:152.0    3rd Qu.:161.0    3rd Qu.:14.0                 3rd Qu.:34.00              3rd Qu.: 80.00
              Max.   :76.00      Max.   :658.0   Max.   :327.0    Max.   :331.0    Max.   :31.0                 Max.   :62.00              Max.   :172.00

mean_population_age_35_44_yr mean_population_age_45_54_yr mean_population_age_55_64_yr mean_population_age_65_up_yr
Min.   :13.00                Min.   : 14.00               Min.   :12.00                Min.   : 13.00
1st Qu.:25.00                1st Qu.: 29.00               1st Qu.:25.00                1st Qu.: 36.00
Median :31.00                Median : 32.00               Median :26.00                Median : 36.00
Mean   :35.08                Mean   : 39.01               Mean   :31.91                Mean   : 44.72
3rd Qu.:44.00                3rd Qu.: 49.00               3rd Qu.:38.00                3rd Qu.: 53.00
Max.   :80.00                Max.   :100.00               Max.   :84.00                Max.   :135.00
```

summary of cmerg

Then create three new data frames from each cluster that we discussed on them separately.

Group by based on Cod3HD

we group by the dataset by feature Cod3HD (store id) to figure out ….

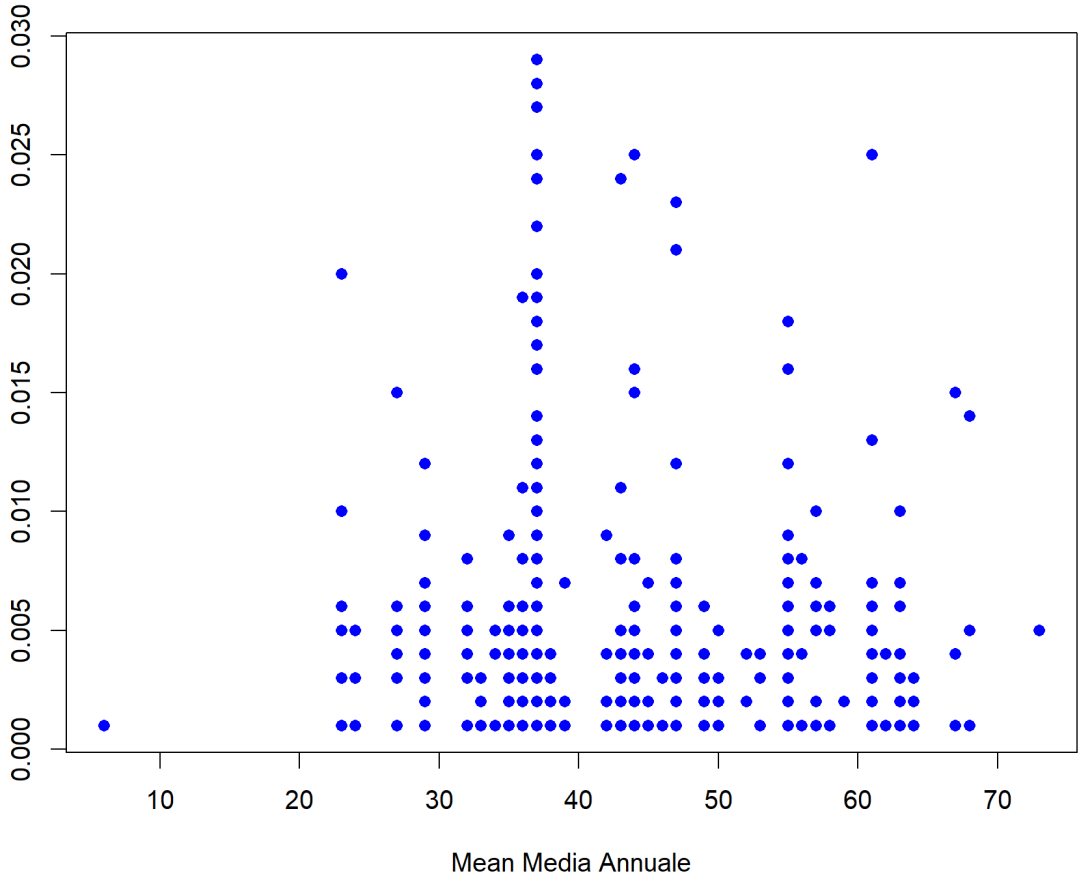
Then create three new data frames from each cluster that we discussed on them separately.

Summary C MERG 1

```
> summary(cmerg1)
    Cod3HD                  Comune           Insegna    TipologiaPdV     MQVEND       Parking      Potenziale      cluster_number      mode_microcode mode_daytype mode_fasciaoraria
Min.   :  198   NAPOLI             :180   N. D.      :123   LIS:449      Min.   :100.0   False:435   Min.   :0.00100   1:545          630495100011:180    1:545         2:545
1st Qu.:15062   GIUGLIANO IN CAMPANIA: 30   CONAD CITY : 36   SUP:  0      1st Qu.:120.0   True :110   1st Qu.:0.00100   2:  0          630340000001: 30
Median :20076   MARANO DI NAPOLI   : 21   PROSHOP    : 35   DIS:  7      Median :165.0               Median :0.00300   3:  0          630410000002: 21
Mean   :18886   CASORIA            : 16   DECO' MARKET: 26   SSD: 89      Mean   :172.4               Mean   :0.00445                  630230000001: 16
3rd Qu.:24733   PORTICI            : 16   MARKETPIU'  : 21                3rd Qu.:220.0               3rd Qu.:0.00500                  630590000001: 16
Max.   :29350   QUARTO             : 13   DESPAR     : 21                Max.   :280.0               Max.   :0.02900                  630630000001: 13
                (Other)            :269   (Other)    :283                                                                             (Other)     :269
mode_datatype mean_media_annuale mean_population mean_population_m mean_population_f mean_population_age_00_04_yr mean_population_age_05_14_yr mean_population_age_15_34_yr
F1:545        Min.   : 6.00      Min.   : 91.0   Min.   : 44.0     Min.   : 47.0     Min.   : 4.0               Min.   : 9.00              Min.   : 22.00
              1st Qu.:37.00      1st Qu.:187.0   1st Qu.: 89.0     1st Qu.: 97.0     1st Qu.: 8.0               1st Qu.:19.00              1st Qu.: 44.00
              Median :37.00      Median :205.0   Median :100.0     Median :105.0     Median :10.0               Median :24.00              Median : 58.00
              Mean   :42.45      Mean   :249.2   Mean   :120.7     Mean   :128.2     Mean   :11.3               Mean   :26.66              Mean   : 62.17
              3rd Qu.:47.00      3rd Qu.:311.0   3rd Qu.:152.0     3rd Qu.:159.0     3rd Qu.:14.0               3rd Qu.:33.00              3rd Qu.: 76.00
              Max.   :73.00      Max.   :658.0   Max.   :327.0     Max.   :331.0     Max.   :31.0               Max.   :62.00              Max.   :172.00

mean_population_age_35_44_yr mean_population_age_45_54_yr mean_population_age_55_64_yr mean_population_age_65_up_yr
Min.   :13.00                Min.   : 14.00               Min.   :12.00                Min.   : 13.00
1st Qu.:25.00                1st Qu.: 29.00               1st Qu.:25.00                1st Qu.: 34.00
Median :28.00                Median : 31.00               Median :26.00                Median : 36.00
Mean   :34.56                Mean   : 38.38               Mean   :31.47                Mean   : 44.41
3rd Qu.:44.00                3rd Qu.: 48.00               3rd Qu.:38.00                3rd Qu.: 53.00
Max.   :80.00                Max.   :100.00               Max.   :84.00                Max.   :135.00
```

Scatterplot for cmerg1



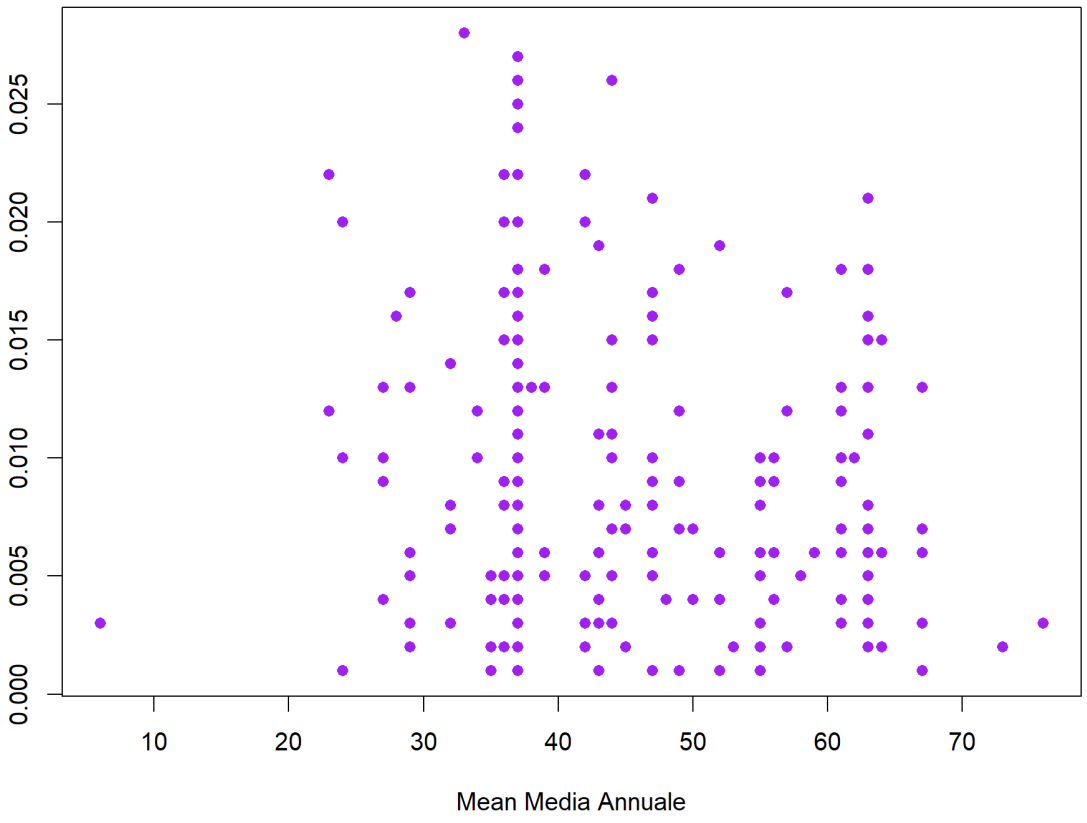
It shows a relationship between two variables: "Mean Media Annuale" on the x-axis and potenziale value on the y-axis. The data points are widely spread across the plot, which indicates variability in the relationship between the two variables. As "Mean Media Annuale" increases, the spread of "Potenziale" values also appears to increase. This suggests there might be greater variability in "Potenziale" at higher levels of "Mean Media Annuale".

Summary C MERG 2

```
> summary(cmerg2)
    Cod3HD                    Comune                  Insegna    TipologiaPdV     MQVEND      Parking     Potenziale      cluster_number      mode_microcode mode_daytype mode_fasciaoraria
Min.   : 1924   NAPOLI              : 94   DECO' SUPERMERCATI: 51   LIS:115      Min.   :285.0   False:151   Min.   :0.001000   1:  0          630495100011: 94    1:268         2:268
1st Qu.: 7276   GIUGLIANO IN CAMPANIA: 15   CONAD             : 36   SUP:122      1st Qu.:320.0   True :117   1st Qu.:0.004000   2:268          630340000001: 15
Median :15949   MARANO DI NAPOLI    :  8   MD                : 25   DIS: 26      Median :400.0               Median :0.008000   3:  0          630410000002:  8
Mean   :14825   CASALNUOVO DI NAPOLI :  7   N. D.             : 23   SSD:  5      Mean   :394.1               Mean   :0.009108                  630170000001:  7
3rd Qu.:21628   POZZUOLI            :  6   CONAD CITY        : 17                3rd Qu.:450.0               3rd Qu.:0.013000                  630600000001:  6
Max.   :29351   CASORIA             :  6   SIGMA             : 15                Max.   :550.0               Max.   :0.028000                  630230000001:  6
                (Other)             :132   (Other)           :101                                                                              (Other)     :132
mode_datatype mean_media_annuale mean_population mean_population_m mean_population_f mean_population_age_00_04_yr mean_population_age_05_14_yr mean_population_age_15_34_yr
F1:268        Min.   : 6.00      Min.   : 91.0   Min.   : 44.0     Min.   : 47.0     Min.   : 4.00               Min.   : 9.00               Min.   : 22.00
              1st Qu.:37.00      1st Qu.:187.0   1st Qu.: 89.0     1st Qu.: 97.0     1st Qu.: 8.00               1st Qu.:19.00               1st Qu.: 44.00
              Median :37.00      Median :210.5   Median :105.0     Median :105.0     Median :10.00               Median :24.50               Median : 58.00
              Mean   :43.46      Mean   :256.7   Mean   :124.6     Mean   :131.8     Mean   :11.77               Mean   :27.53               Mean   : 64.68
              3rd Qu.:52.00      3rd Qu.:319.0   3rd Qu.:159.0     3rd Qu.:162.0     3rd Qu.:15.00               3rd Qu.:34.75               3rd Qu.: 80.00
              Max.   :76.00      Max.   :561.0   Max.   :275.0     Max.   :286.0     Max.   :31.00               Max.   :59.00               Max.   :153.00

mean_population_age_35_44_yr mean_population_age_45_54_yr mean_population_age_55_64_yr mean_population_age_65_up_yr
Min.   :13.00                Min.   :14.00                Min.   :12.00                Min.   : 13.0
1st Qu.:25.00                1st Qu.:29.00                1st Qu.:25.00                1st Qu.: 36.0
Median :31.00                Median :31.50                Median :26.00                Median : 36.0
Mean   :35.71                Mean   :39.75                Mean   :32.38                Mean   : 44.7
3rd Qu.:47.00                3rd Qu.:50.00                3rd Qu.:40.00                3rd Qu.: 53.0
Max.   :73.00                Max.   :88.00                Max.   :74.00                Max.   :101.0
```

Scatterplot for cmerg2

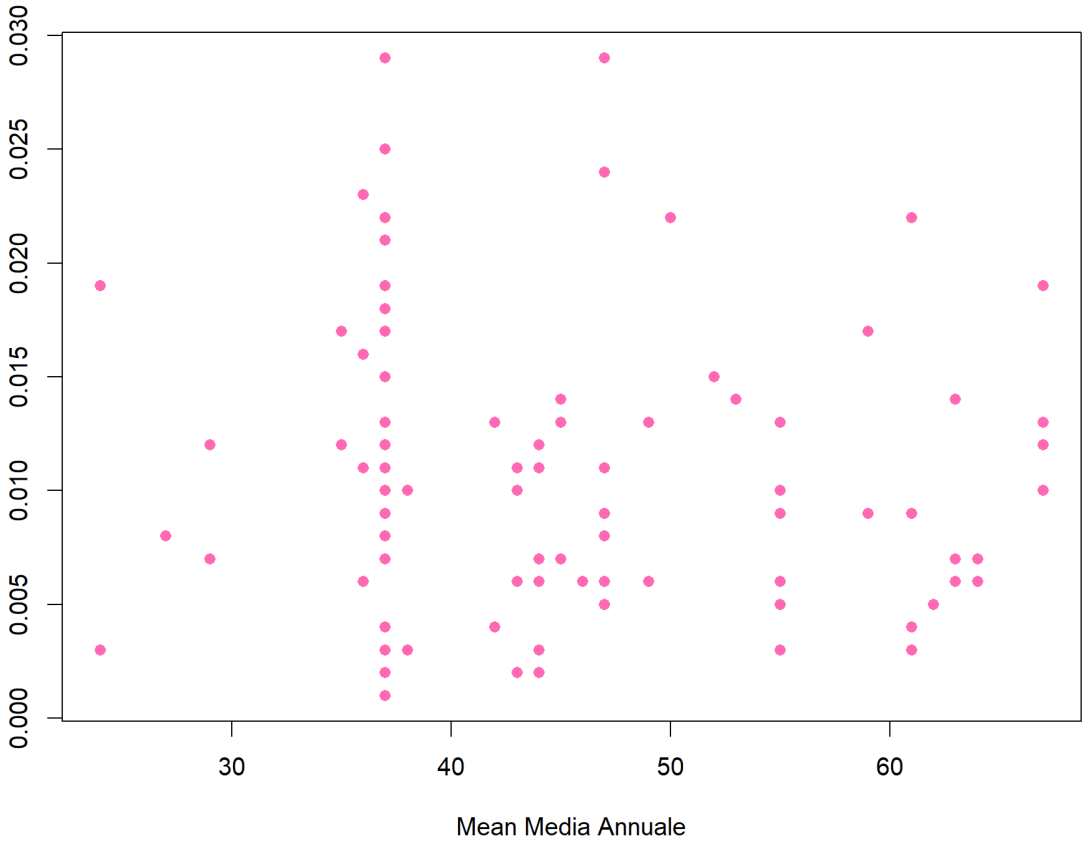


Summary C MERG 3

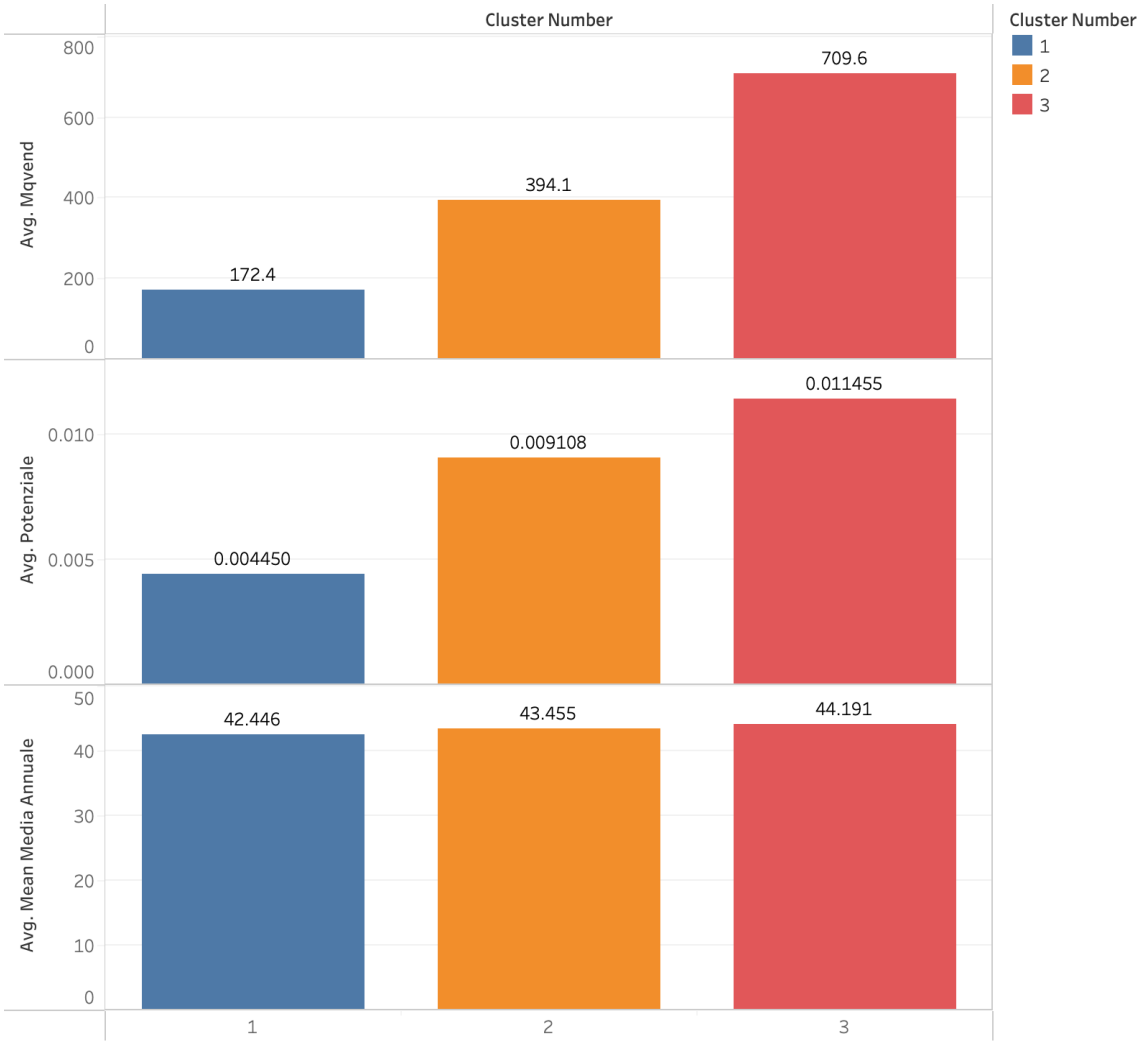
```
> summary(cmerg3)
     Cod3HD                  Comune                Insegna   TipologiaPdV      MQVEND       Parking     Potenziale      cluster_number      mode_microcode mode_daytype mode_fasciaoraria
 Min.   : 1925   NAPOLI           :39   MD              :28   LIS: 0       Min.   :555.0   False:33   Min.   :0.00100   1:  0          630495100011:39     1:110        2:110
 1st Qu.: 7985   POZZUOLI         : 5   CONAD           :11   SUP:71       1st Qu.:600.0   True :77   1st Qu.:0.00600   2:  0          630600000001: 5
 Median :15818   CASORIA          : 5   SIGMA           : 8   DIS:38       Median :700.0              Median :0.01000   3:110          630230000001: 5
 Mean   :14757   TORRE DEL GRECO  : 5   DECO' SUPERMERCATI: 8   SSD: 1       Mean   :709.6              Mean   :0.01145                  630840000002: 5
 3rd Qu.:21367   MUGNANO DI NAPOLI: 4   SISA            : 7                3rd Qu.:800.0              3rd Qu.:0.01500                  630480000001: 4
 Max.   :27010   MARIGLIANO       : 3   N. D.           : 7                Max.   :935.0              Max.   :0.02900                  630430000001: 3
                 (Other)          :49   (Other)         :41                                                                            (Other)     :49
 mode_datatype mean_media_annuale mean_population mean_population_m mean_population_f mean_population_age_00_04_yr mean_population_age_05_14_yr mean_population_age_15_34_yr
 F1:110        Min.   :24.00      Min.   : 91.0   Min.   : 44.0     Min.   : 47.0     Min.   : 4.00               Min.   : 9.00               Min.   : 22.00
               1st Qu.:37.00      1st Qu.:187.0   1st Qu.: 89.0     1st Qu.: 97.0     1st Qu.: 8.00               1st Qu.:19.00               1st Qu.: 44.00
               Median :42.00      Median :254.5   Median :123.5     Median :131.0     Median :11.00               Median :28.00               Median : 64.00
               Mean   :44.19      Mean   :261.4   Mean   :126.4     Mean   :134.8     Mean   :11.79               Mean   :28.07               Mean   : 65.59
               3rd Qu.:49.75      3rd Qu.:324.2   3rd Qu.:159.0     3rd Qu.:165.8     3rd Qu.:15.00               3rd Qu.:35.50               3rd Qu.: 81.50
               Max.   :67.00      Max.   :439.0   Max.   :209.0     Max.   :230.0     Max.   :20.00               Max.   :49.00               Max.   :115.00

 mean_population_age_35_44_yr mean_population_age_45_54_yr mean_population_age_55_64_yr mean_population_age_65_up_yr
 Min.   :13.00                Min.   :14.00                Min.   :12.00                Min.   :13.00
 1st Qu.:25.00                1st Qu.:29.00                1st Qu.:25.00                1st Qu.:36.00
 Median :35.00                Median :37.00                Median :30.50                Median :39.00
 Mean   :36.15                Mean   :40.32                Mean   :32.98                Mean   :46.29
 3rd Qu.:47.00                3rd Qu.:50.00                3rd Qu.:40.00                3rd Qu.:54.00
 Max.   :61.00                Max.   :68.00                Max.   :57.00                Max.   :83.00
```

Scatterplot for cmerg3

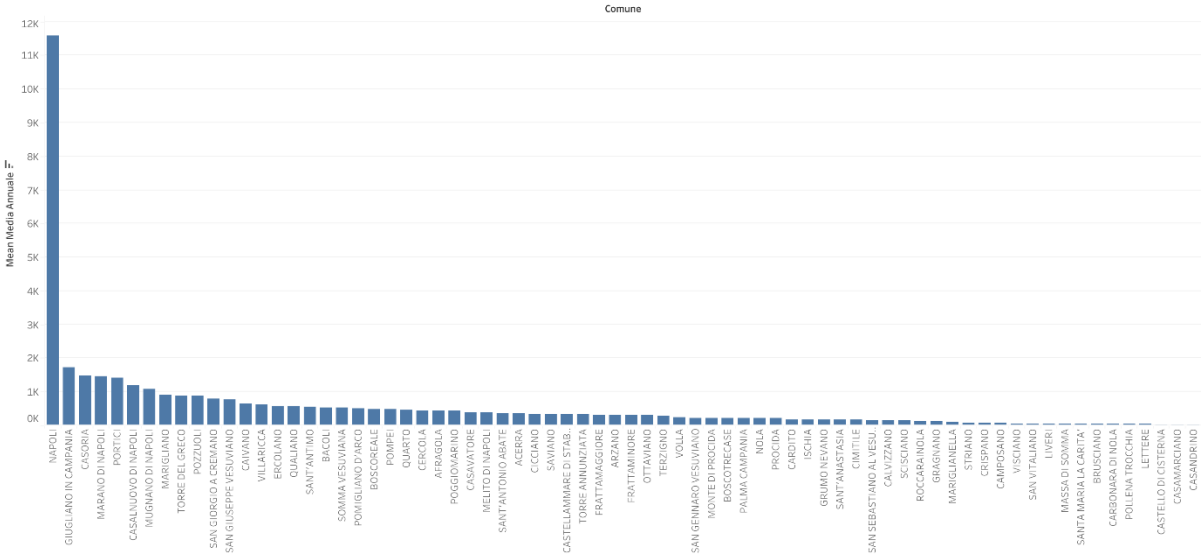


In the following figures, we analyze three datasets together. The first figure compares three datasets in terms of Mqvend, Potenziale, and Media Annuale. The chart demonstrates that although the average of media annual is very close for the three clusters, potential and store size vary significantly. For instance, in the first cluster with smaller stores, among other clusters, lower competition could influence more people. Therefore, it has more potential for expanding sales and marketing strategies.

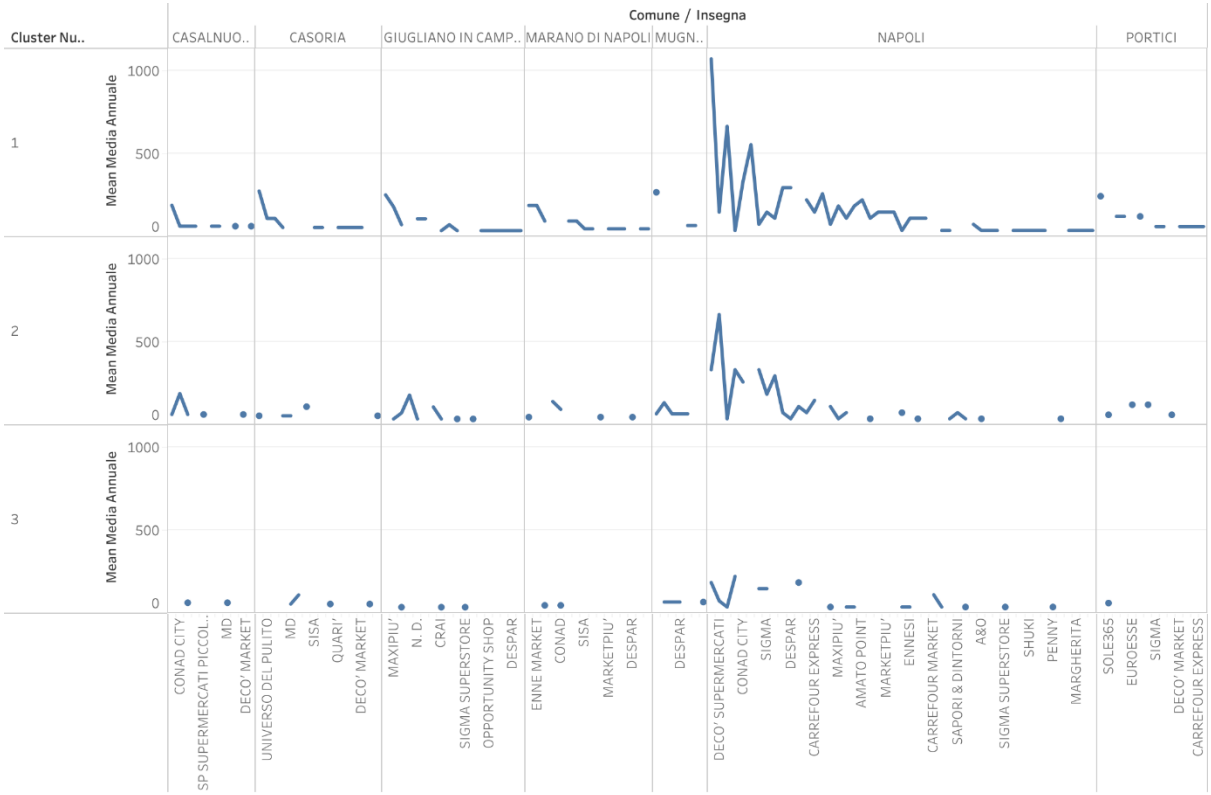


comparison of three datasets in terms of Mqvend, Potenziale, and Media Annuale

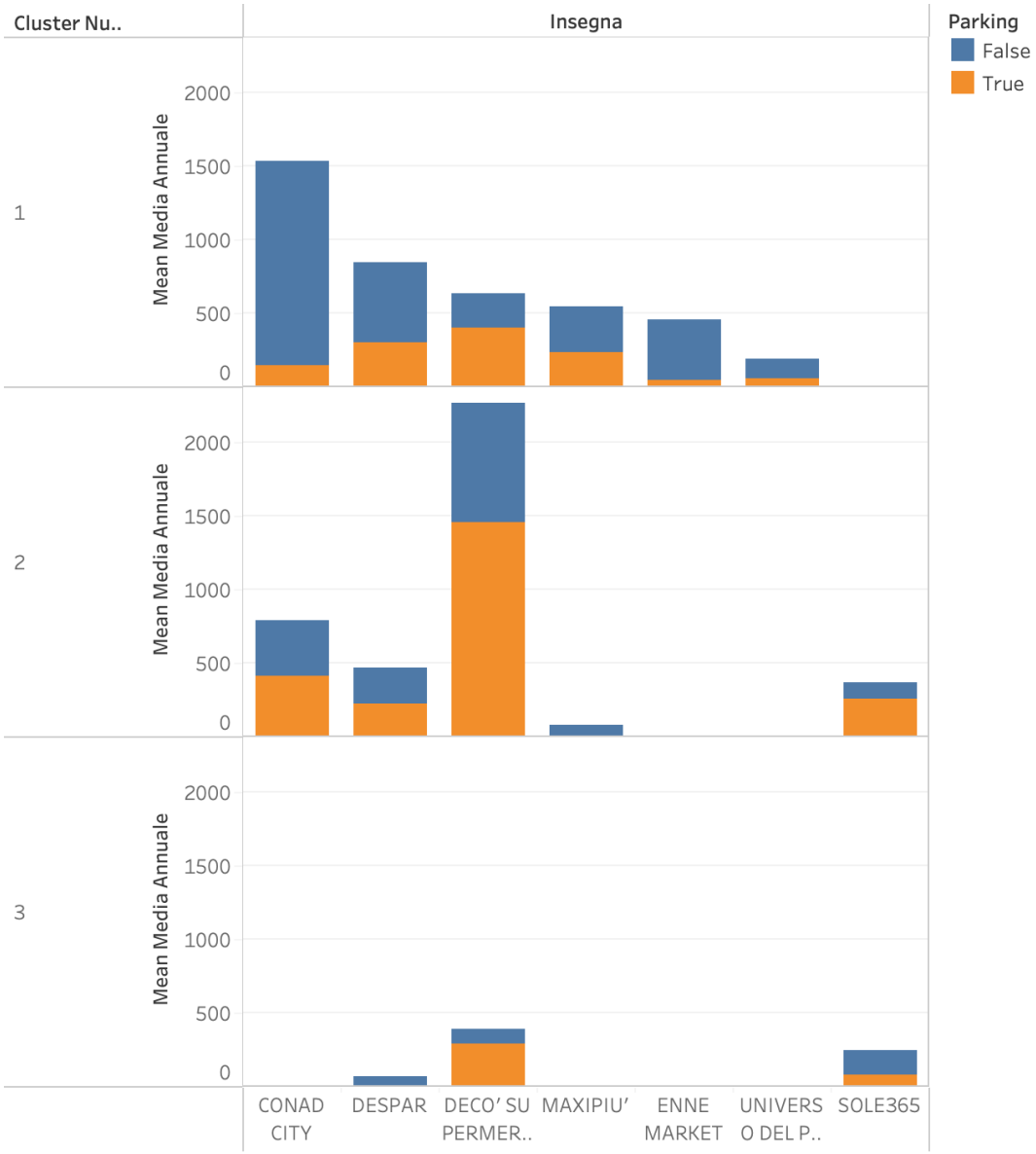
This following chart displays cities that attract more attention. Therefore, we will use the top 6 of them for further analysis.



The third chart illustrates the more active stores attracting people in each city, as chosen by the last chart, according to separate clusters. Therefore, we will use the top store from each city for further analysis.



The last chart compares stores availability of parking or not in each cluster.



Comparison of three datasets based on parking facilities categorized by store type

The aim of this part is using a random forest algorithm for predicting the potential of each store related to its cluster and identifying and prioritizing key features using the mean decrease in Gini coefficient criterion simultaneously. So, each data set is divided into two sets: training set (80% data) and test set (20% data) with considering numerical features. A random forest algorithm with suitable parameters is then applied to them for each dataset to ensure the highest prediction accuracy for both sets. After implementing the algorithm, the importance of the variables is determined based on the mean decrease in the Gini coefficient. The variables are prioritized by sorting them in a descending order based on mean decrease in the Gini coefficient.

Random forest

The random forest algorithm is a combination of multiple decision trees, constructed from several bootstrap samples of data. A number of input variables are randomly involved in the construction of each tree. The bootstrap method is indeed sampling with replacement. As a result of repeating the sampling operation, a number of datasets are generated from the training set, which can be calculated for each set of a decision tree. It would be a random forest classification if the expected value is discrete and a random forest regression if it is continuous.

Gini coefficient criterion

A random forest algorithm is used because the random forest classifier with the importance of the associated Gini coefficient feature allows direct elimination of irrelevant features in comparison with other classification methods. The Gini coefficient is a type of error that can be measured using the following equation:

$$Gini - index = \sum_{k=1}^{K} \widehat{p_{mk}}(1 - \widehat{p_{mk}})$$

Where $\widehat{p_m}$ indicates the probability of classification accuracy. The higher the mean decrease in the Gini coefficient, the higher the importance of the variable.

Prediction result for each data set and plot of importance features

Or any model for predicting…

To figure out how many ntree we should have and how depth we should go, we use cross-validation.

Cross-validation

Cross-Validation can be used to estimate the test error associated with a given statistical learning method in order to evaluate its performance, or to select the appropriate level of flexibility.

We use V-FOLD cross validation which we explain in below.

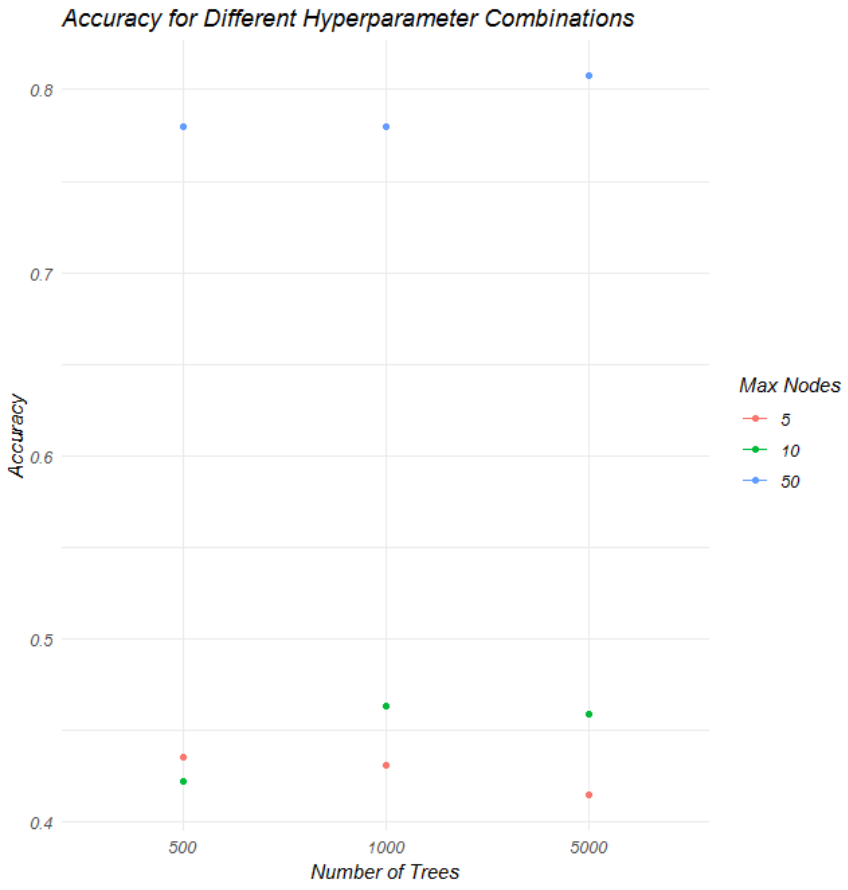
This approach involves randomly V-fold CV dividing the set of observations into V groups, or folds, of approximately equal size.
• The first fold is treated as a validation set, and the method is fit on the remaining V - 1 folds. The error, say MSE_i, is computed on the observations in the held-out fold.
• This procedure is repeated V times; each time, a different group of observations is treated as a validation set.
• The V -fold CV error is

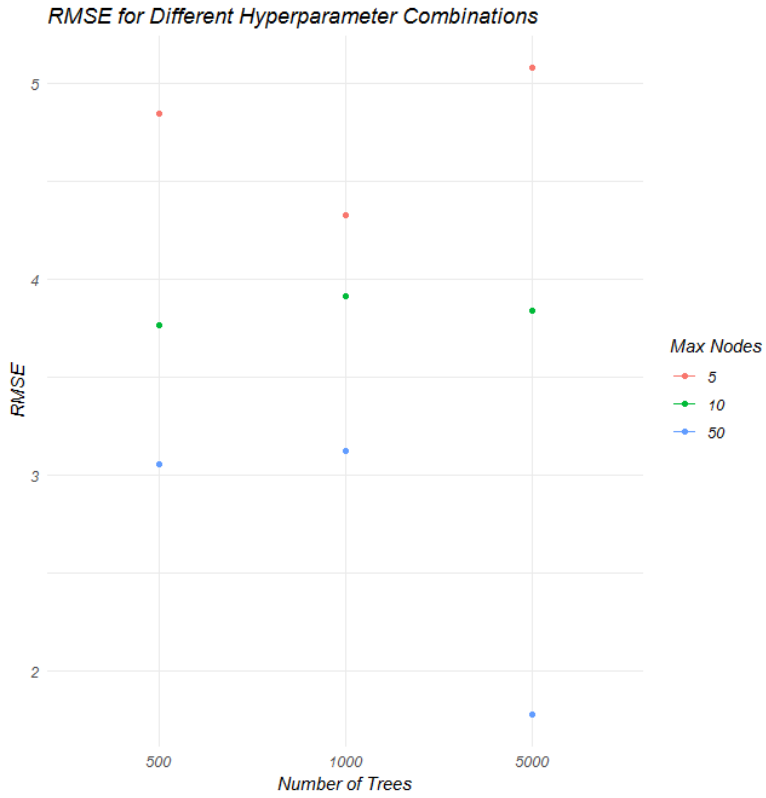
$$CV_{(V)} = \frac{1}{v}\sum_{i=1}^{v} MSE_i$$

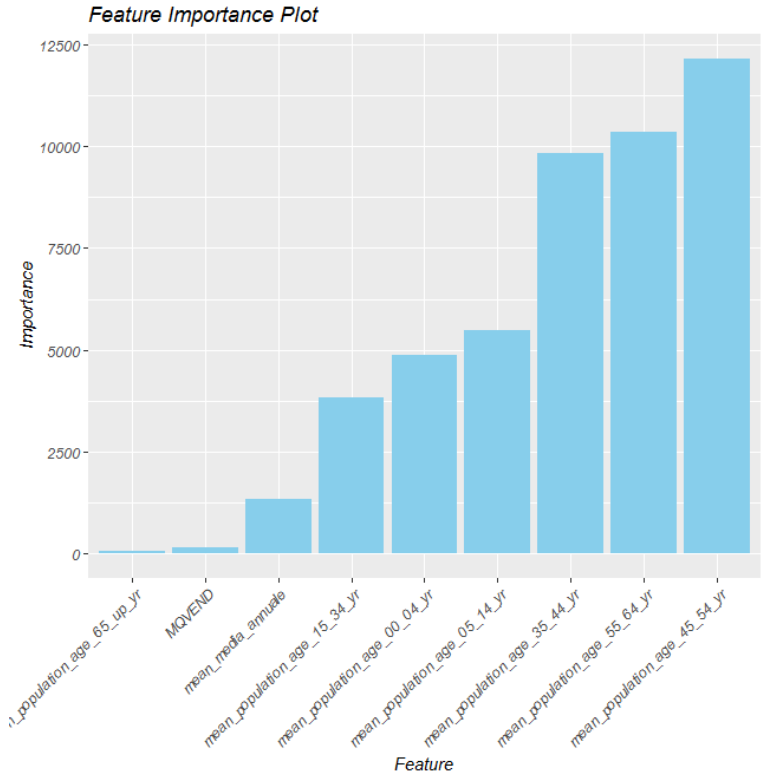
For rfcmerg1:

```
Call:
 randomForest(formula = mean_media_annuale ~ ., data = train_data,      ntree = ntree, maxnodes = maxnodes)
               Type of random forest: regression
                     Number of trees: 5000
No. of variables tried at each split: 3

          Mean of squared residuals: 6.219793
                    % Var explained: 94.43
> print(paste("Max Nodes:", maxnodes))
[1] "Max Nodes: 50"
>
```





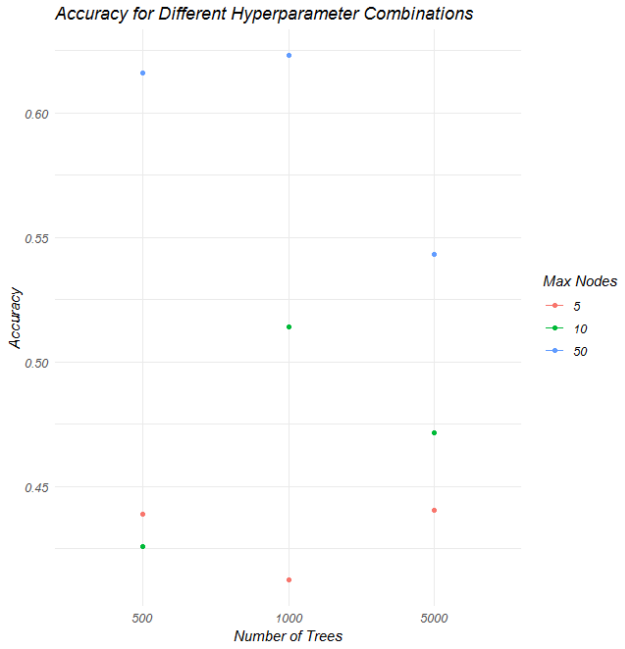


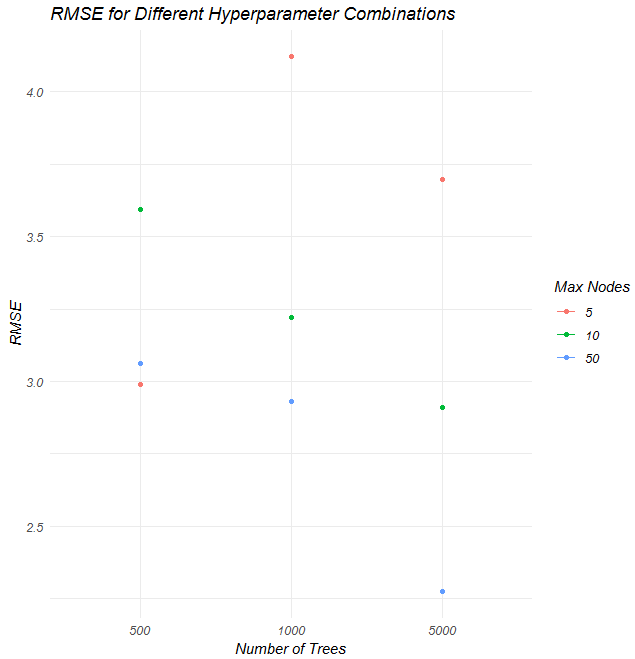
As the bar chart shows the most important feature is mean-population-age-45-54-yr,it means this feature can paly an important role in the dataset. the most influential feature is belong to this age. The height of each bar corresponds to the importance score of each feature, which quantifies the contribution of the feature to the model's predictive power. The scores are presumably calculated based on the mean decrease in impurity (such as Gini impurity) that results from splits over a particular feature, a common metric in tree-based models

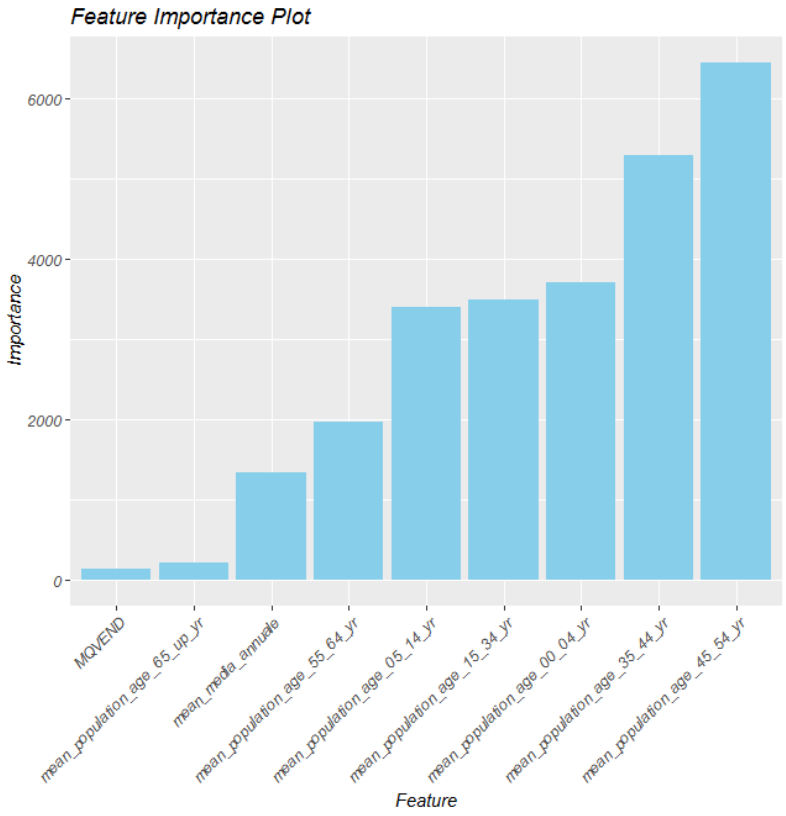
For rfcmerg2:

```
Call:
 randomForest(formula = mean_media_annuale ~ ., data = train_data,      ntree = ntree, maxnodes = maxnodes)
               Type of random forest: regression
                     Number of trees: 1000
No. of variables tried at each split: 3

          Mean of squared residuals: 6.545076
                    % Var explained: 94.67
> print(paste("Max Nodes:", maxnodes))
[1] "Max Nodes: 50"
```



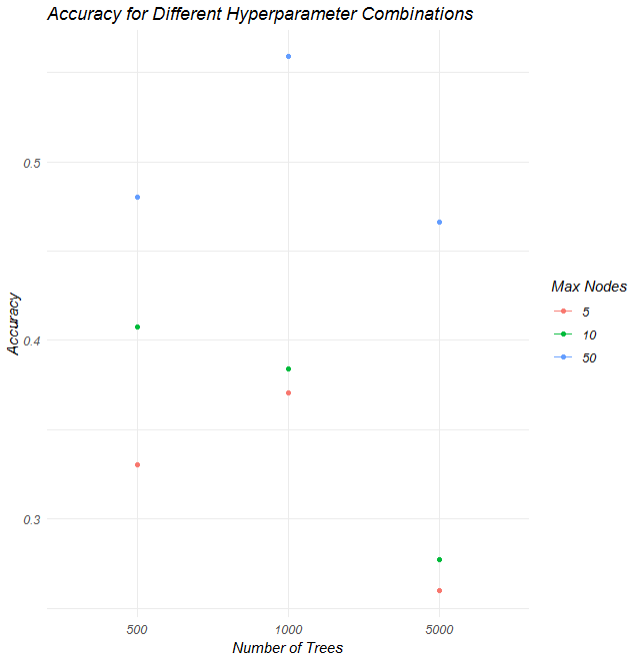


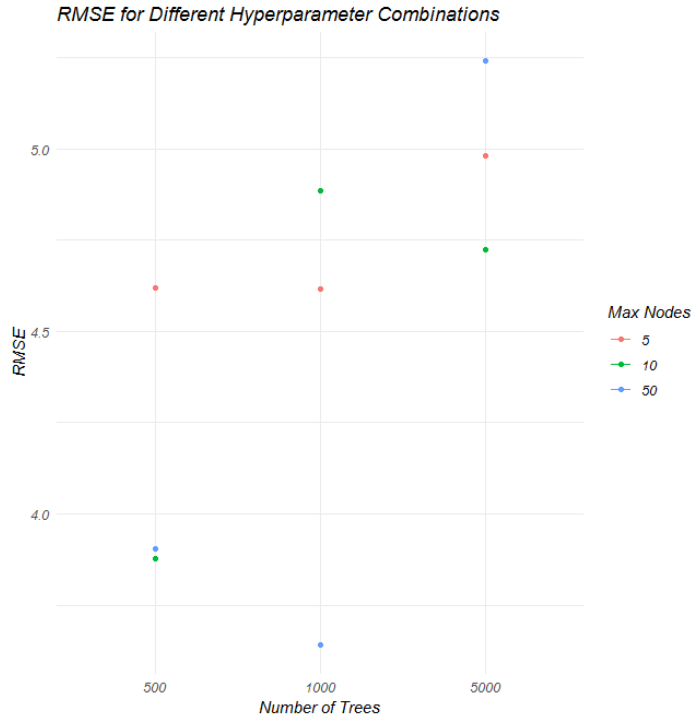


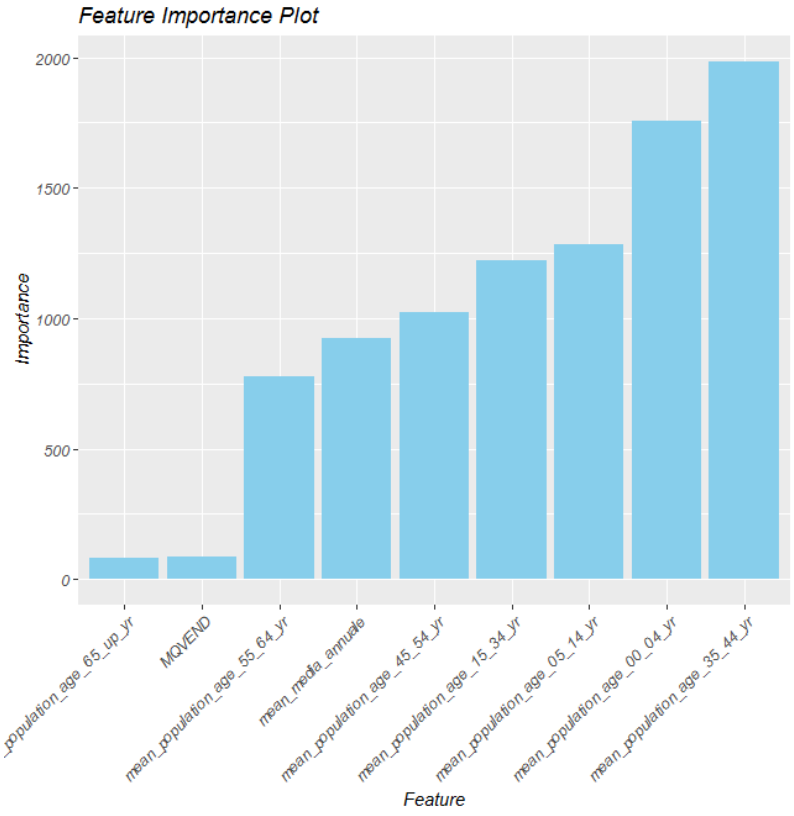
For rfcmerg3:

```
Call:
 randomForest(formula = mean_media_annuale ~ ., data = train_data,      ntree = ntree, maxnodes = maxnodes)
               Type of random forest: regression
                     Number of trees: 1000
No. of variables tried at each split: 3

          Mean of squared residuals: 17.24651
                    % Var explained: 83.72
> print(paste("Max Nodes:", maxnodes))
[1] "Max Nodes: 50"
```




RMSE for Different Hyperparameter Combinations
RMSE
5.0
4.5
4.0
500
1000
5000
Number of Trees
Max Nodes
5
10
50


Feature Importance Plot
Importance
2000
1500
1000
500
0
mean_population_age_65_up_yr
MQVEND
mean_population_age_55_64_yr
mean_media_annuale
mean_population_age_45_54_yr
mean_population_age_15_34_yr
mean_population_age_05_14_yr
mean_population_age_00_04_yr
mean_population_age_35_44_yr
Feature