**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**

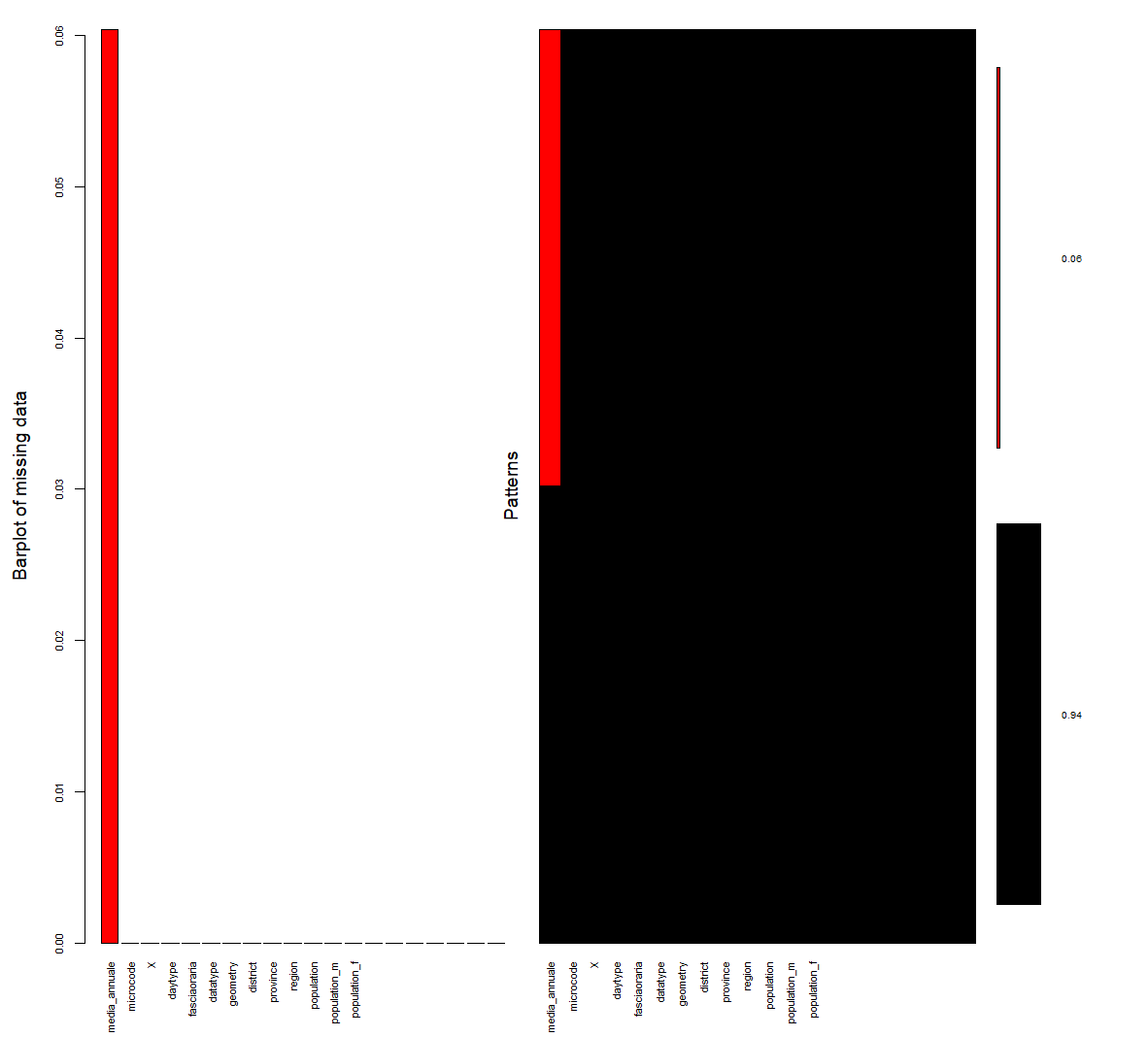
**Duplicate**

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.

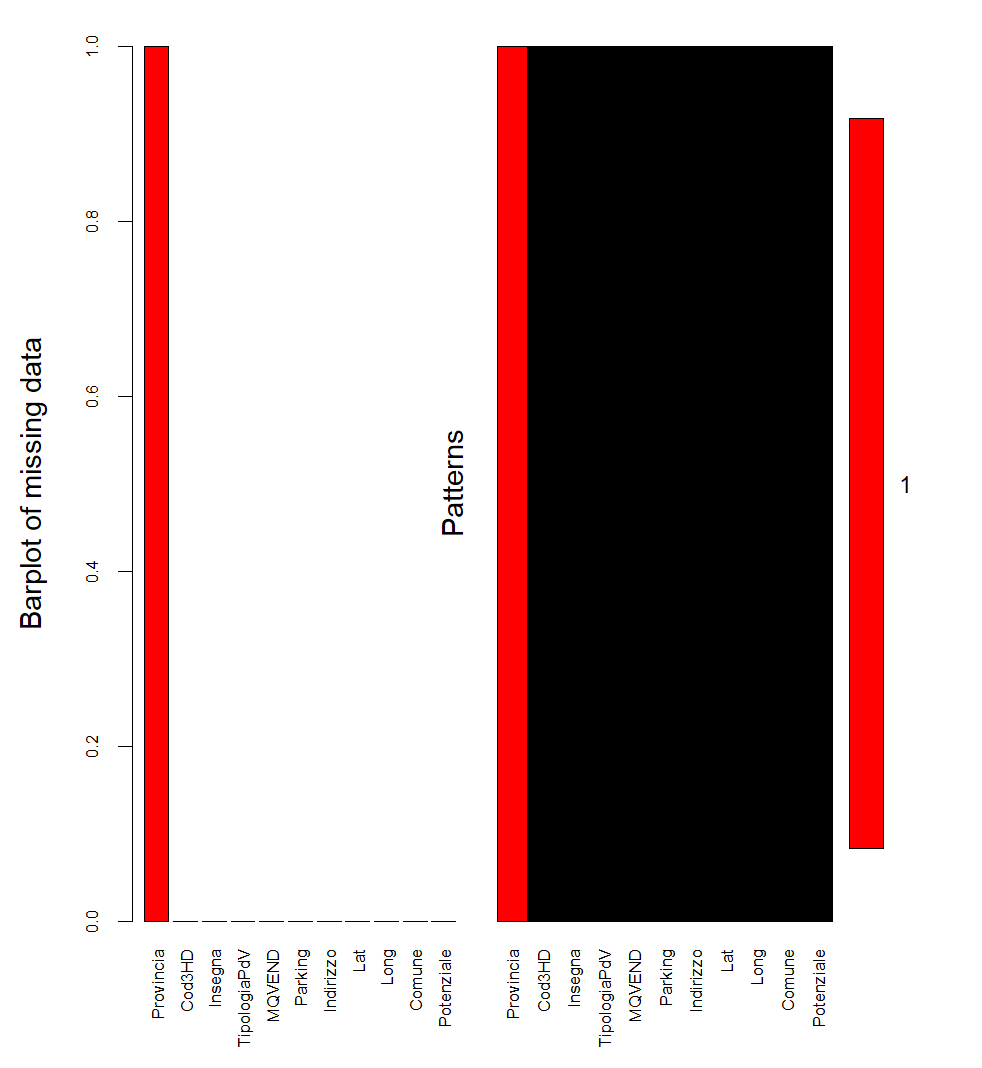
**Missing values (VIM)**

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

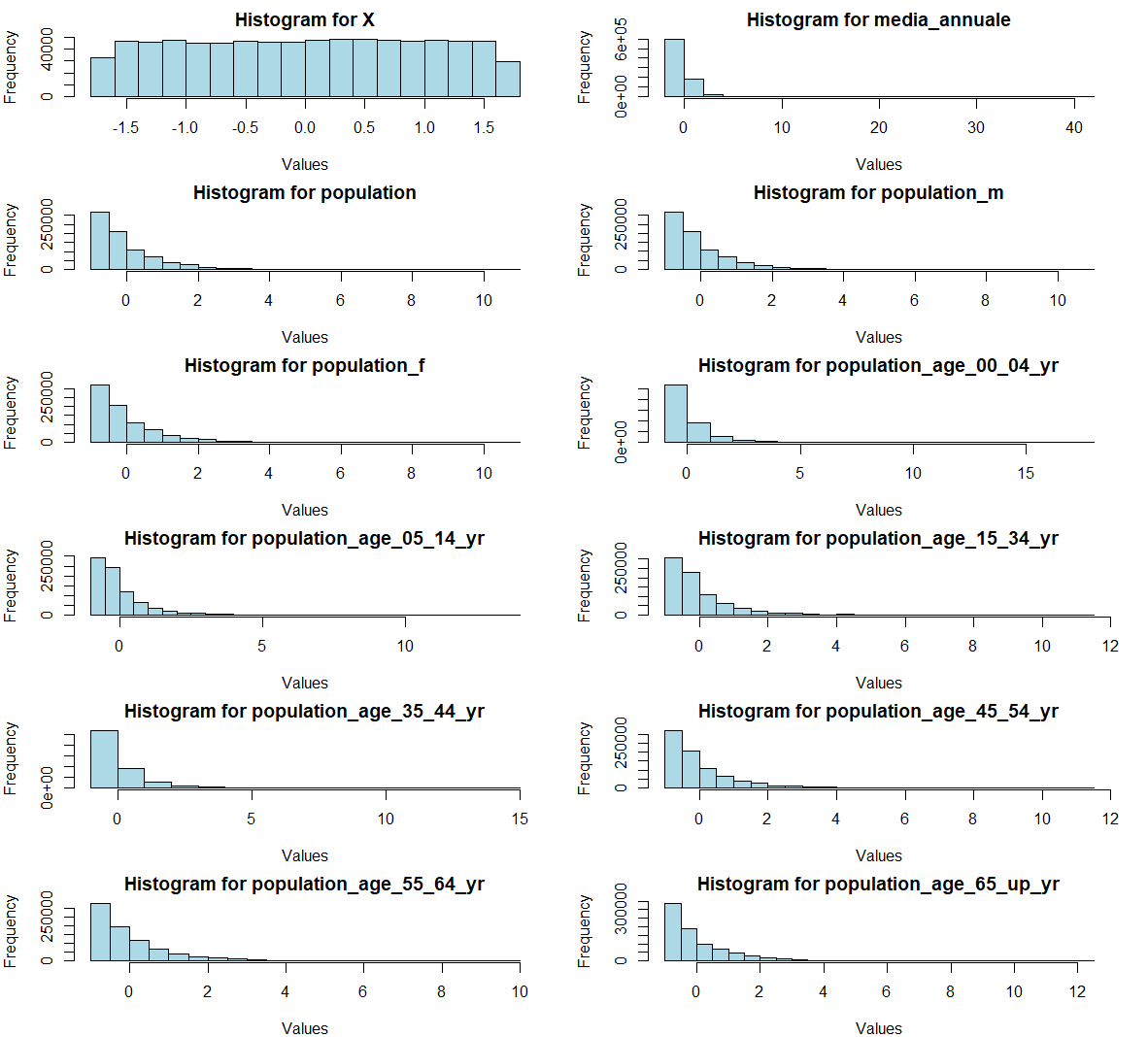
****

Missing value of store-NA dataset

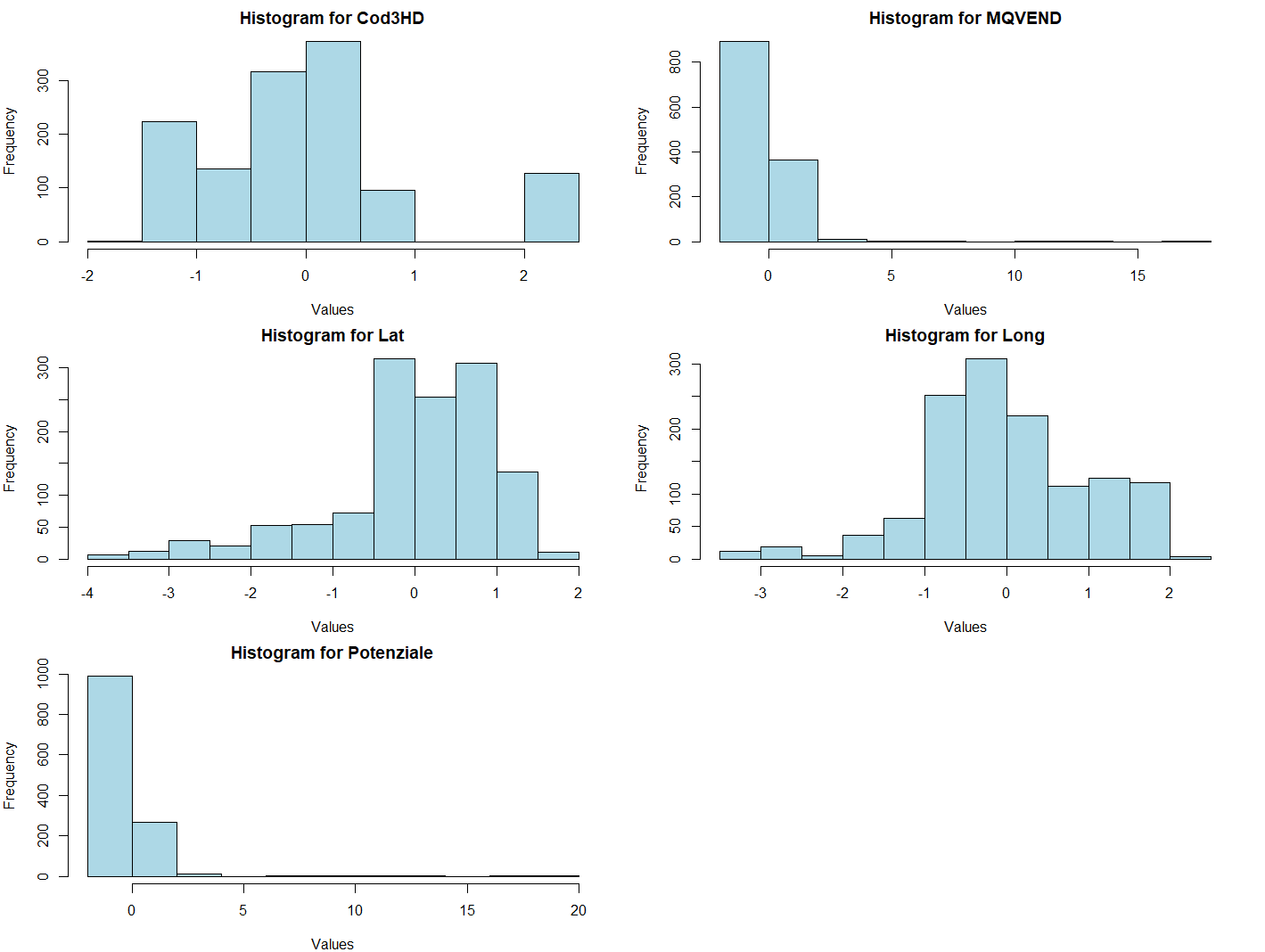
**Impute the missing values with mean**

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.

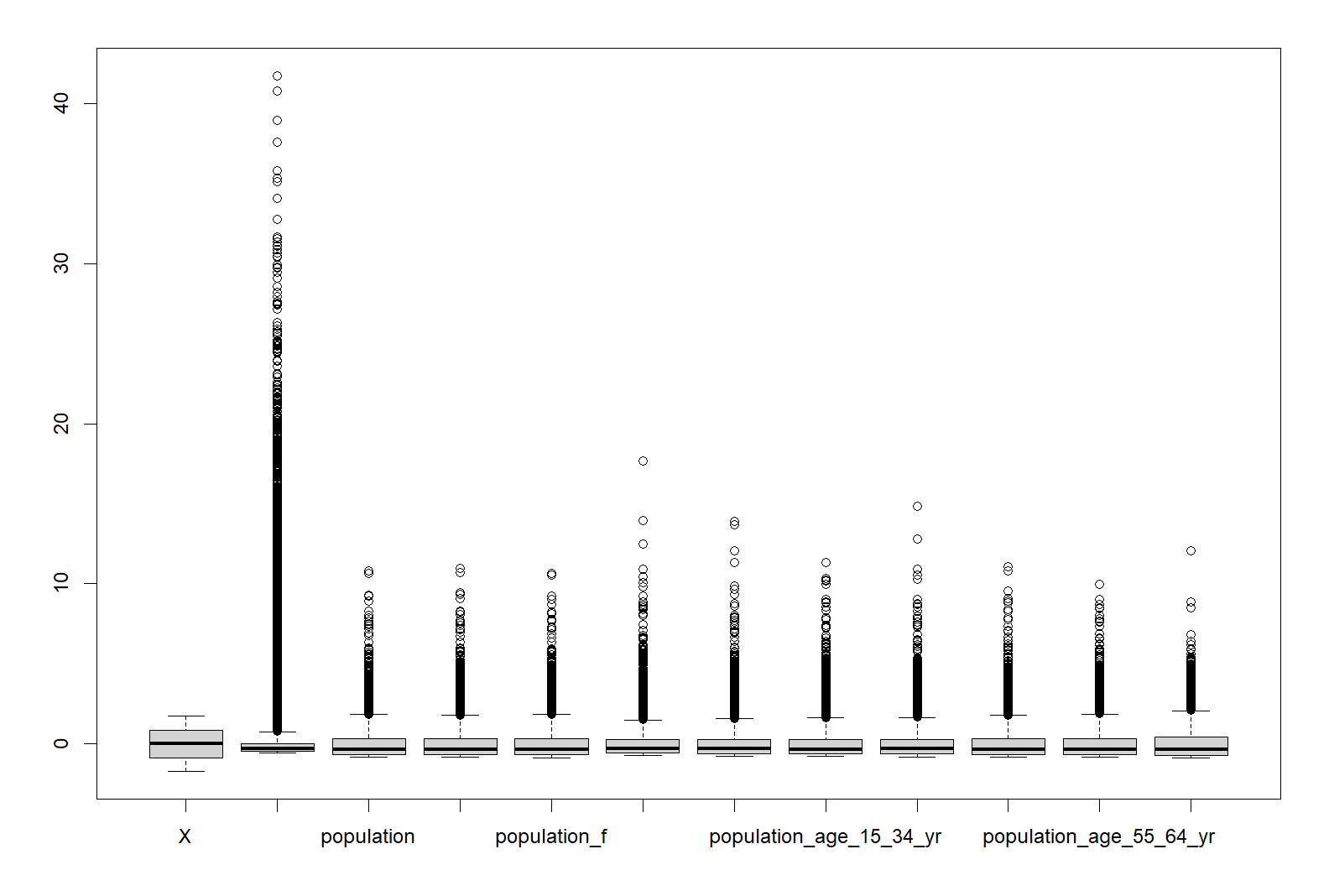
This can be a common strategy for handling missing data, especially when the missing values are assumed to be missing at random. After imputing missing values, we separate the categorical columns and make a Histogram plots for numerical data.

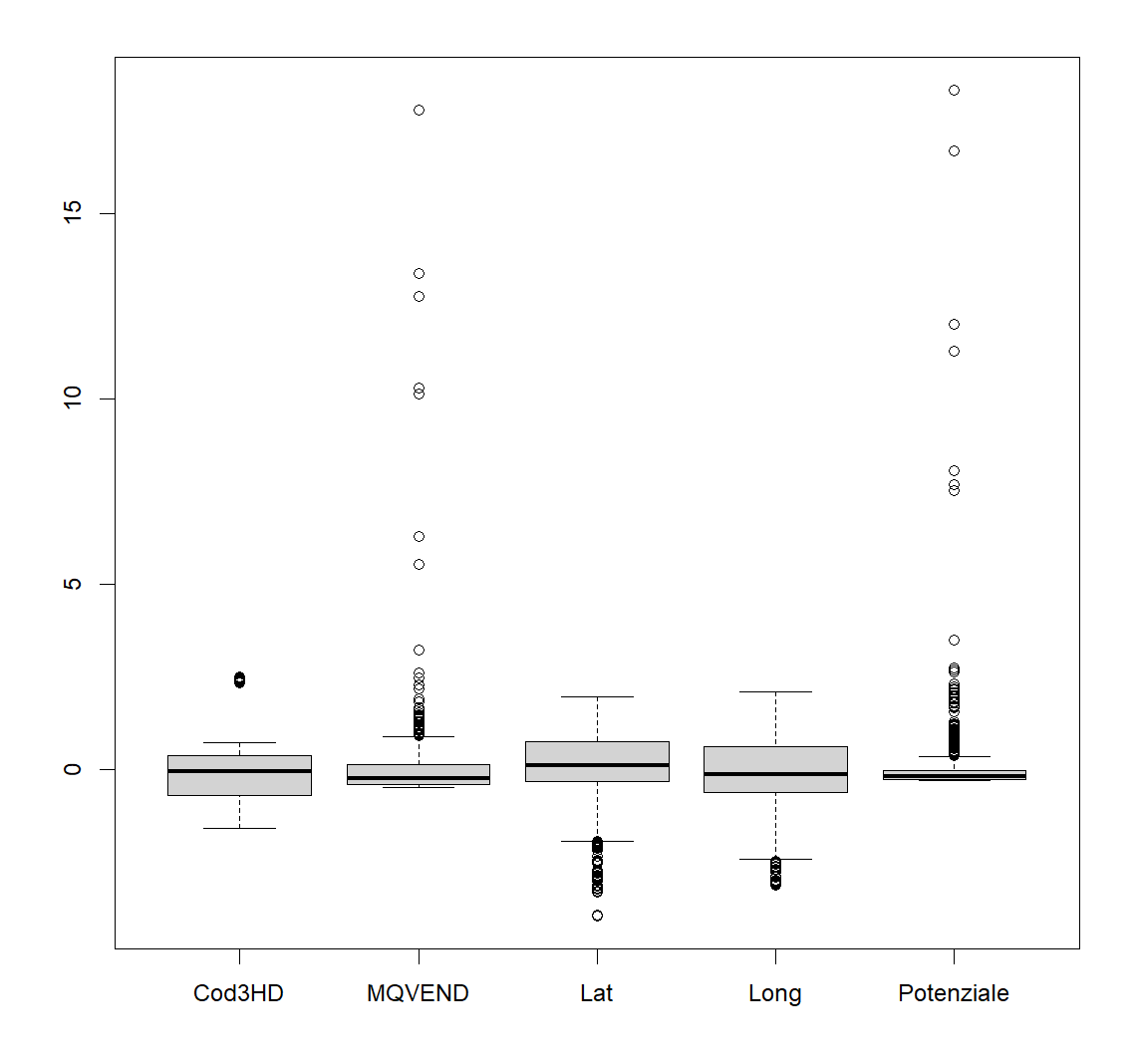


Histograms for all mdf features before cleaning



Histograms for store before cleanin

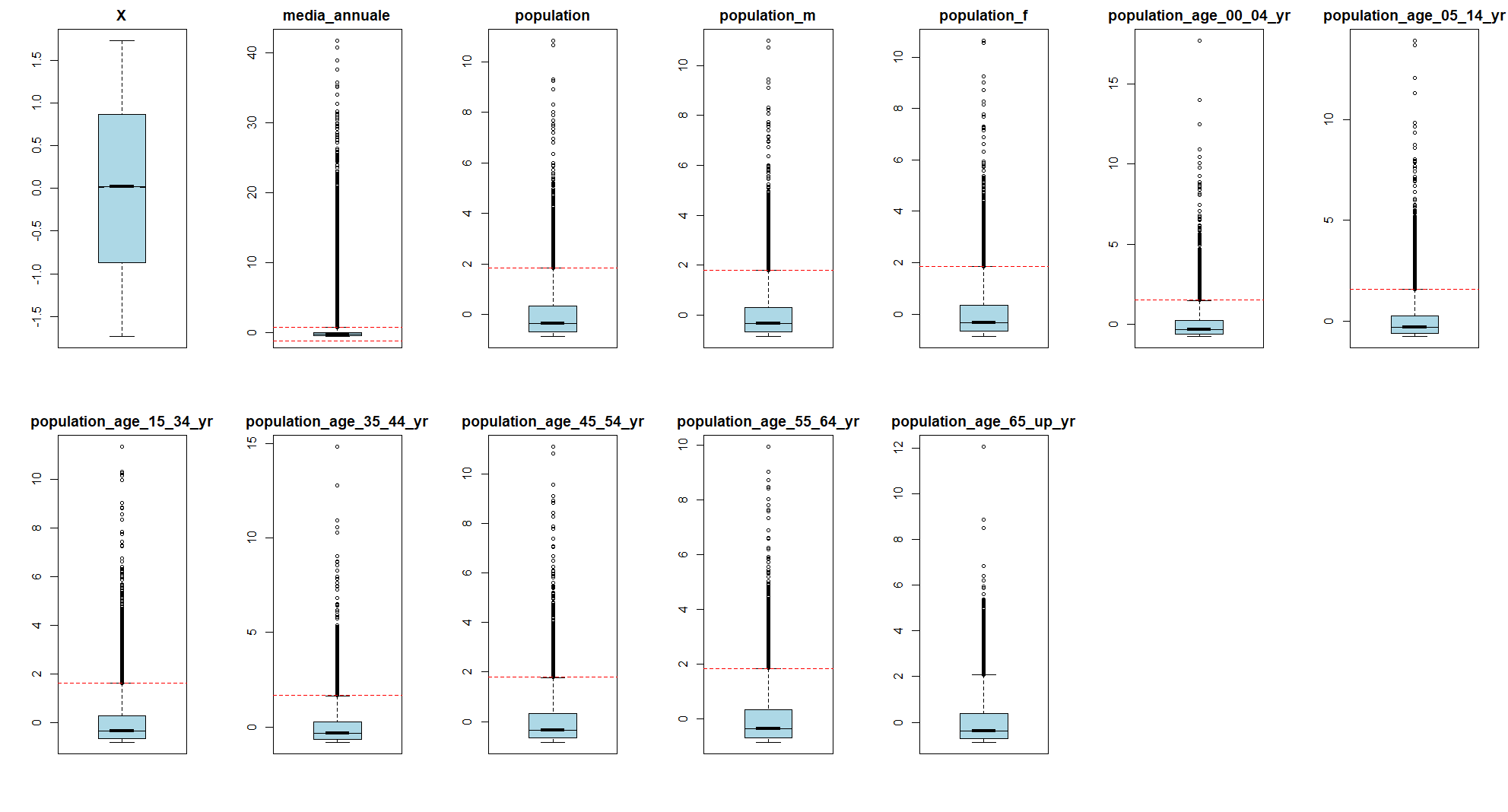
Boxplot(mdf)



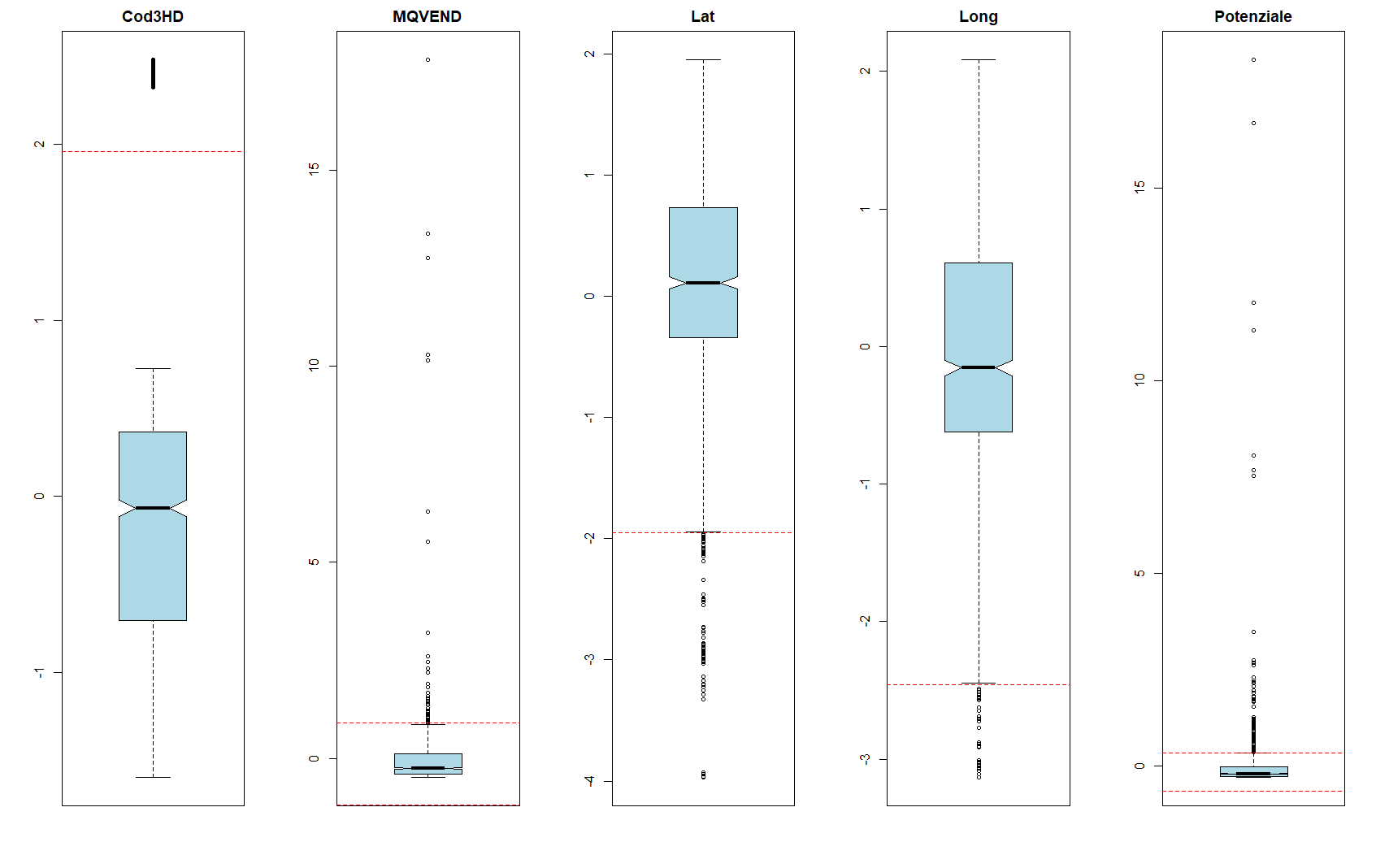
Boxplot store

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.

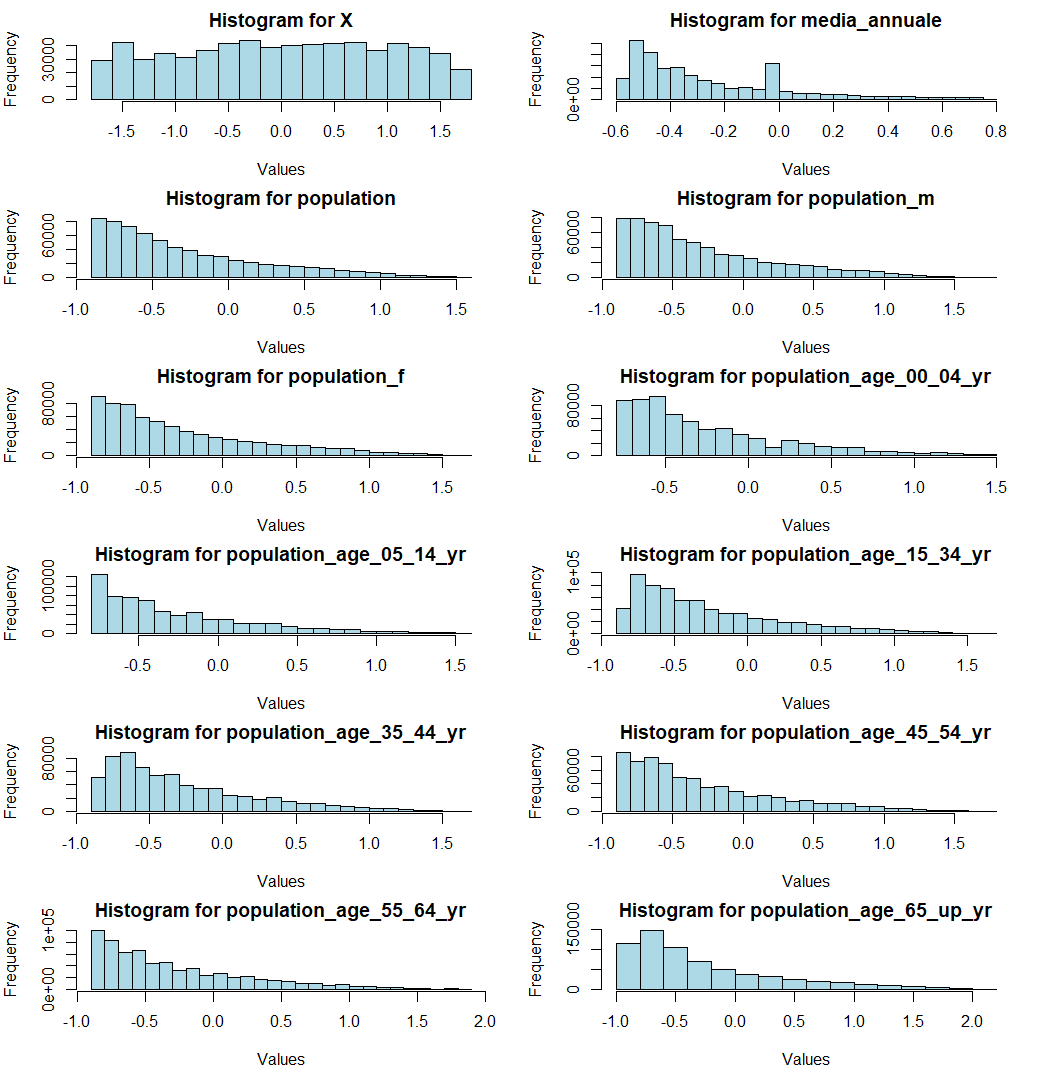


Boxplot with red line for each mdf columns

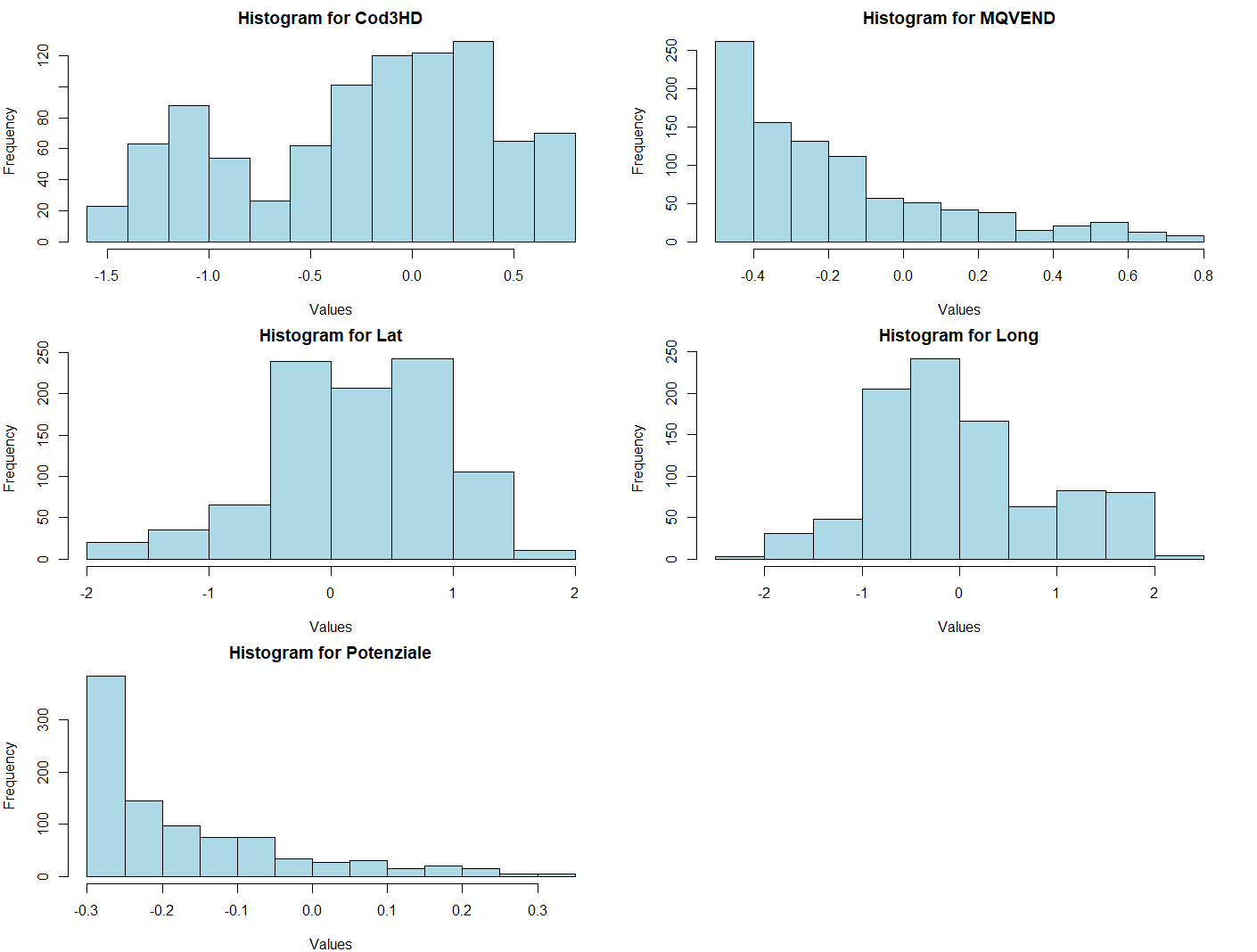
Boxplot with red line for each store columns

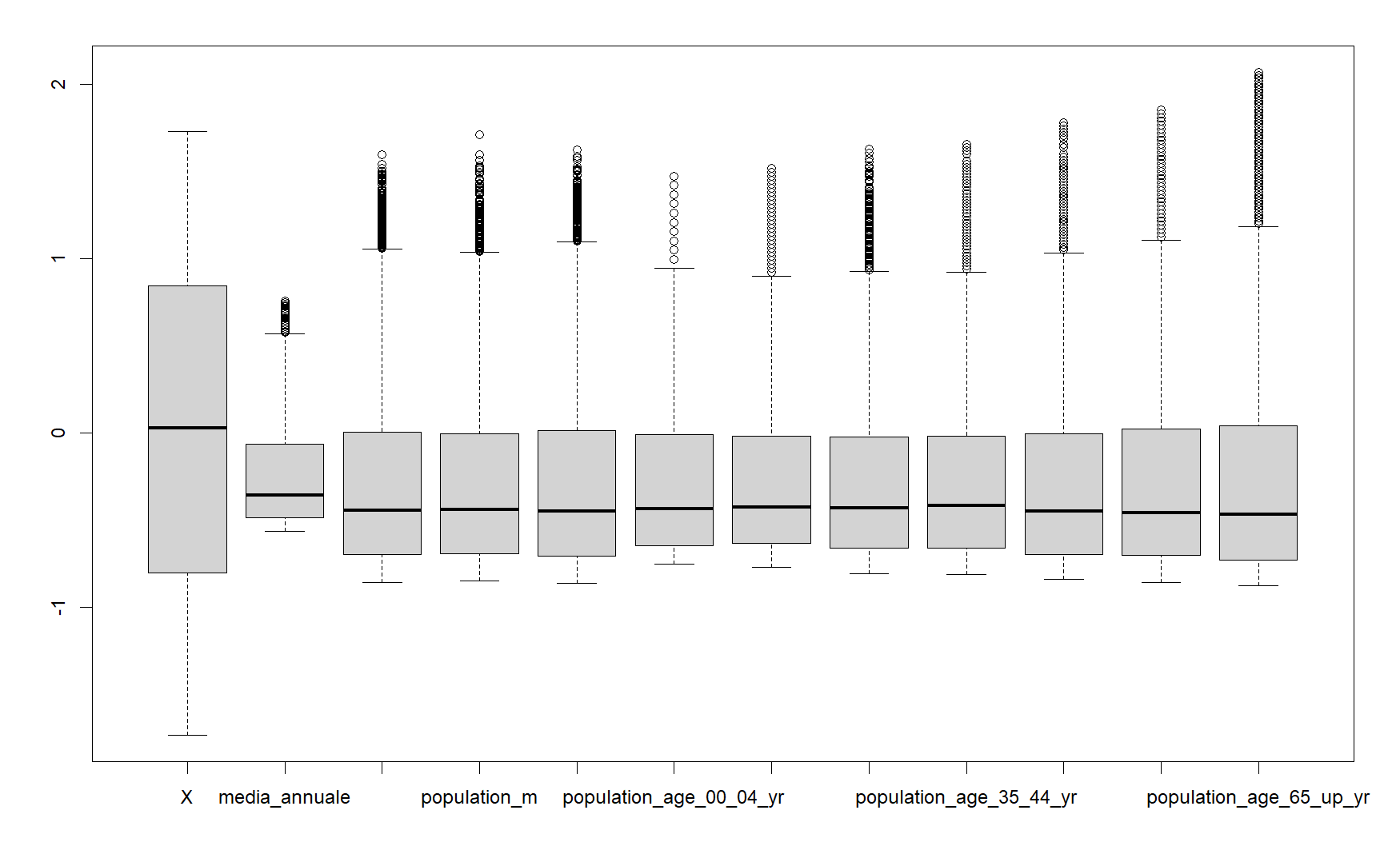
With using IQR methode we try to remove the outliers.Then we make a box plot for each feature in our cleaned dataset and now we can figure out the difference before and after cleaning . the highest and the lowest limit are created to detect outliers for each variable. the points that are upper than U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

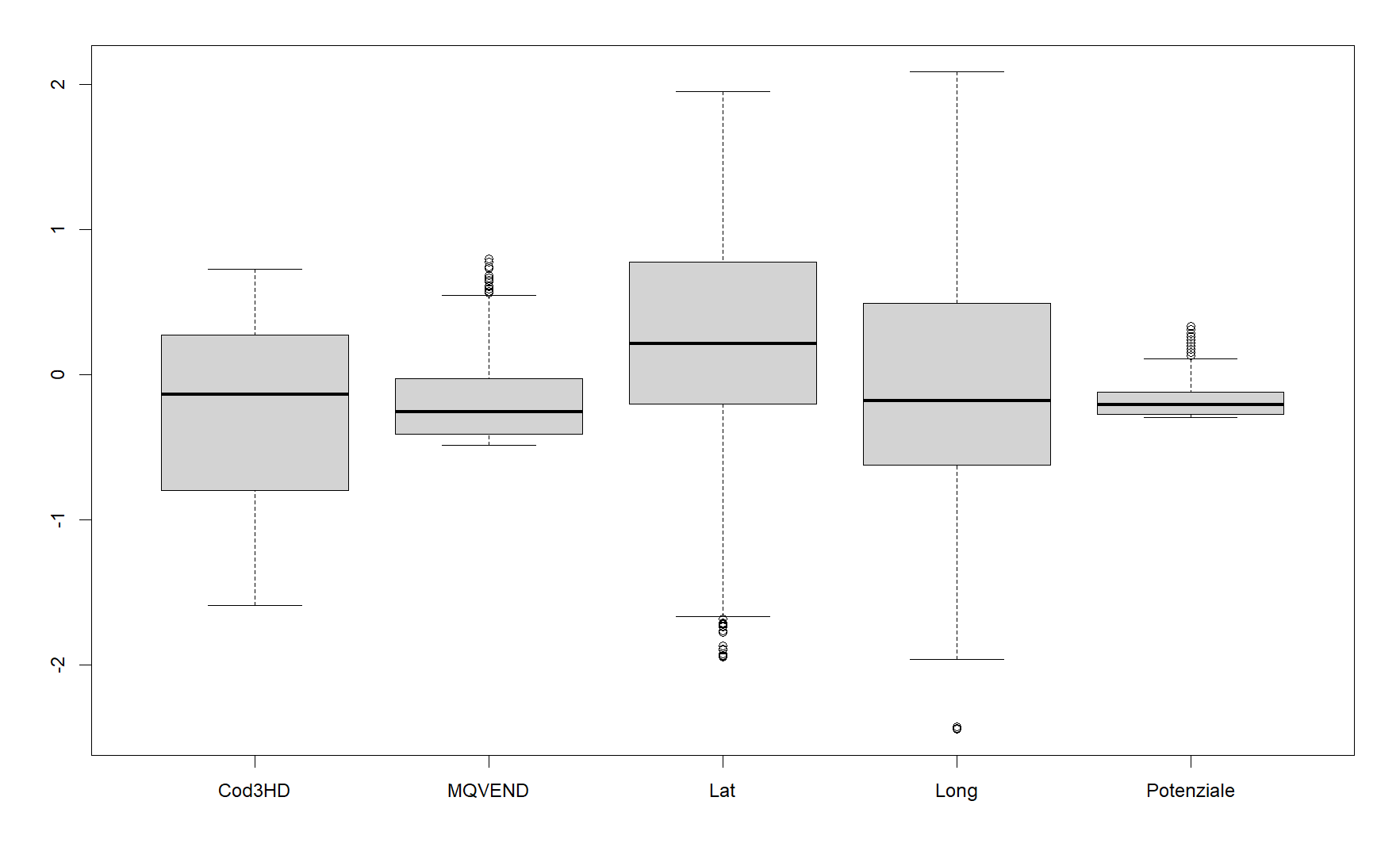


Histogram for each mdf features after standardization



**Histogram for each store features after standardization**

Boxplot for mdf cleaned data



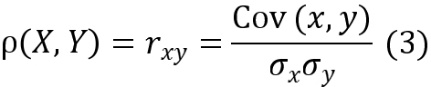
Boxplot for store cleaned data

**Correlation analysis for cleaned-mdf**

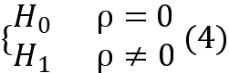
Correlation analysis is related to finding a relationship between variables and then determining the extent and action of that relationship. Pearson and Spearman correlation coefficients are in the range of +1 to -1. Hence, if the correlation coefficient is near or equal to 1, there is a strong and co-directional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

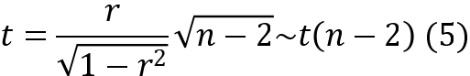
In this condition, if one variable increases, the other one will increase too, and the direction of the variation is the same for the two variables. Besides, when one variable decreases, the other variable will decrease too. In this condition, the two variables have proportional relationships. On the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

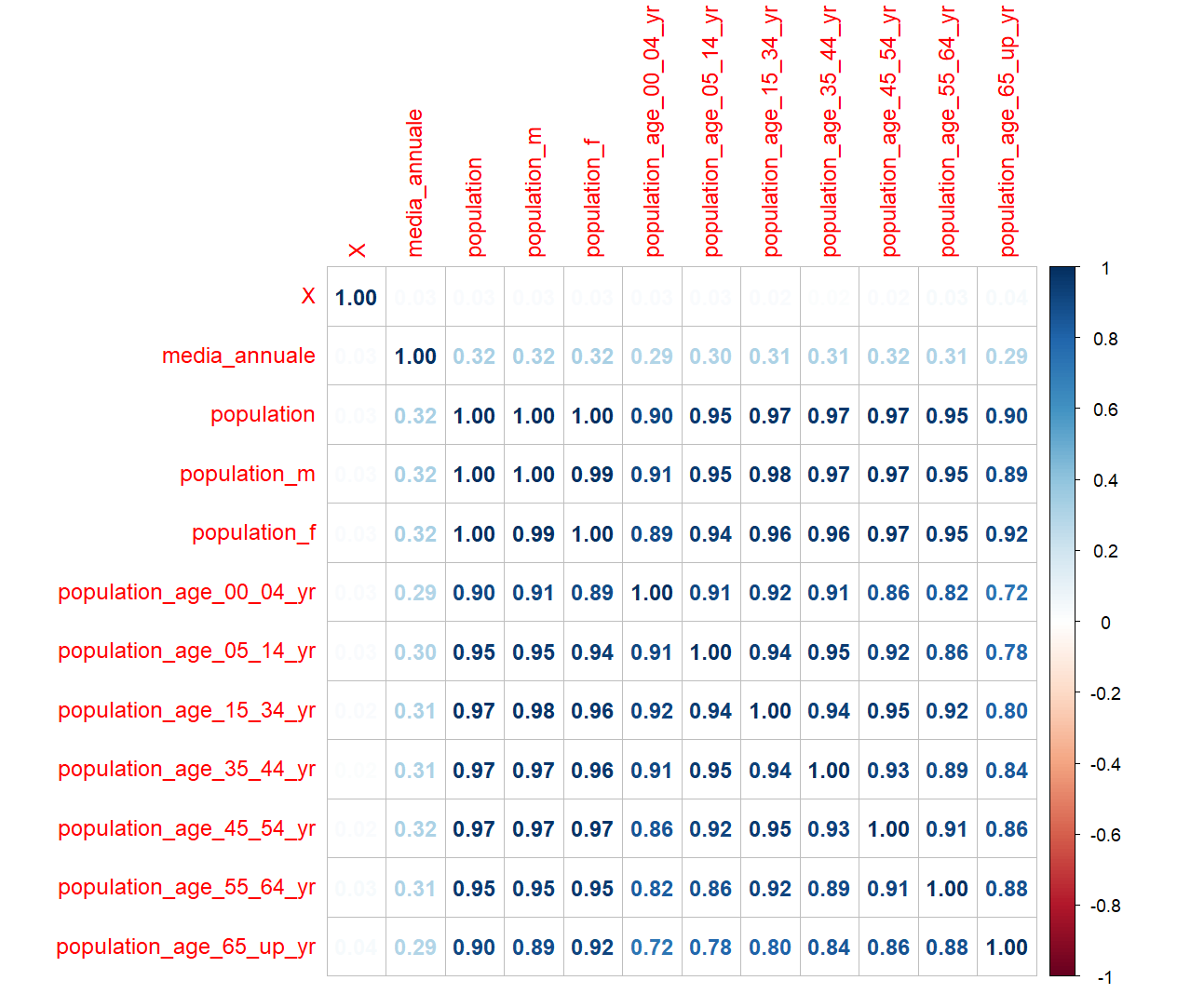
Therefore, if one variable increases, the other one will decrease. However, the prediction is still possible in this condition. This is called the "inverse relationship." In this study, Pearson linear correlation coefficient has been used to measure the linearity of the relationship between two variables,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



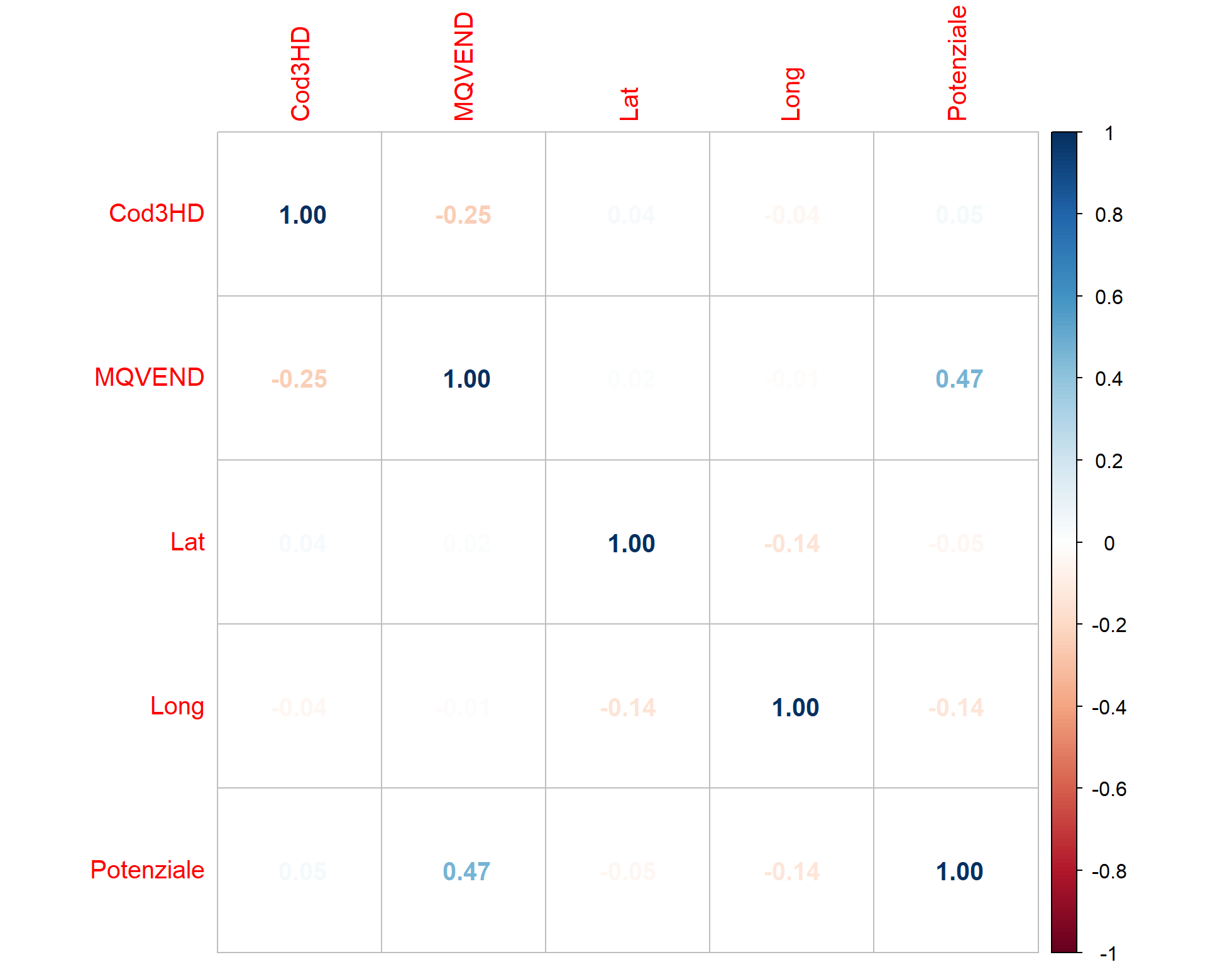
A statistical test with sample statistic with t-student distribution and n-2 degree of freedom has been used based on equations (4) and (5) (n is the number of observations). The chart below demonstrates the results of the correlation matrix.







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:

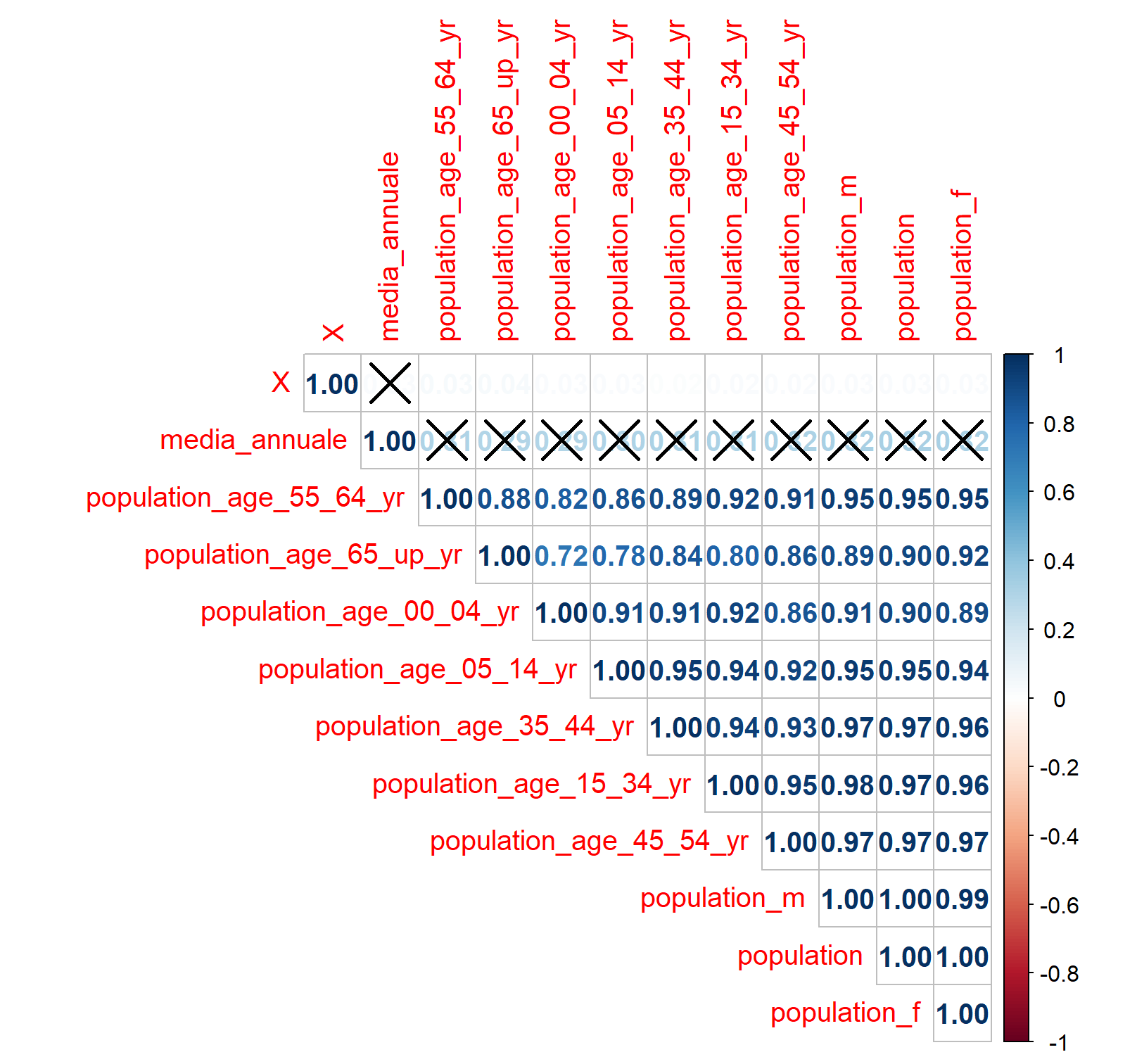
Negatively correlated (correlation coefficient**:**-0.14).

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

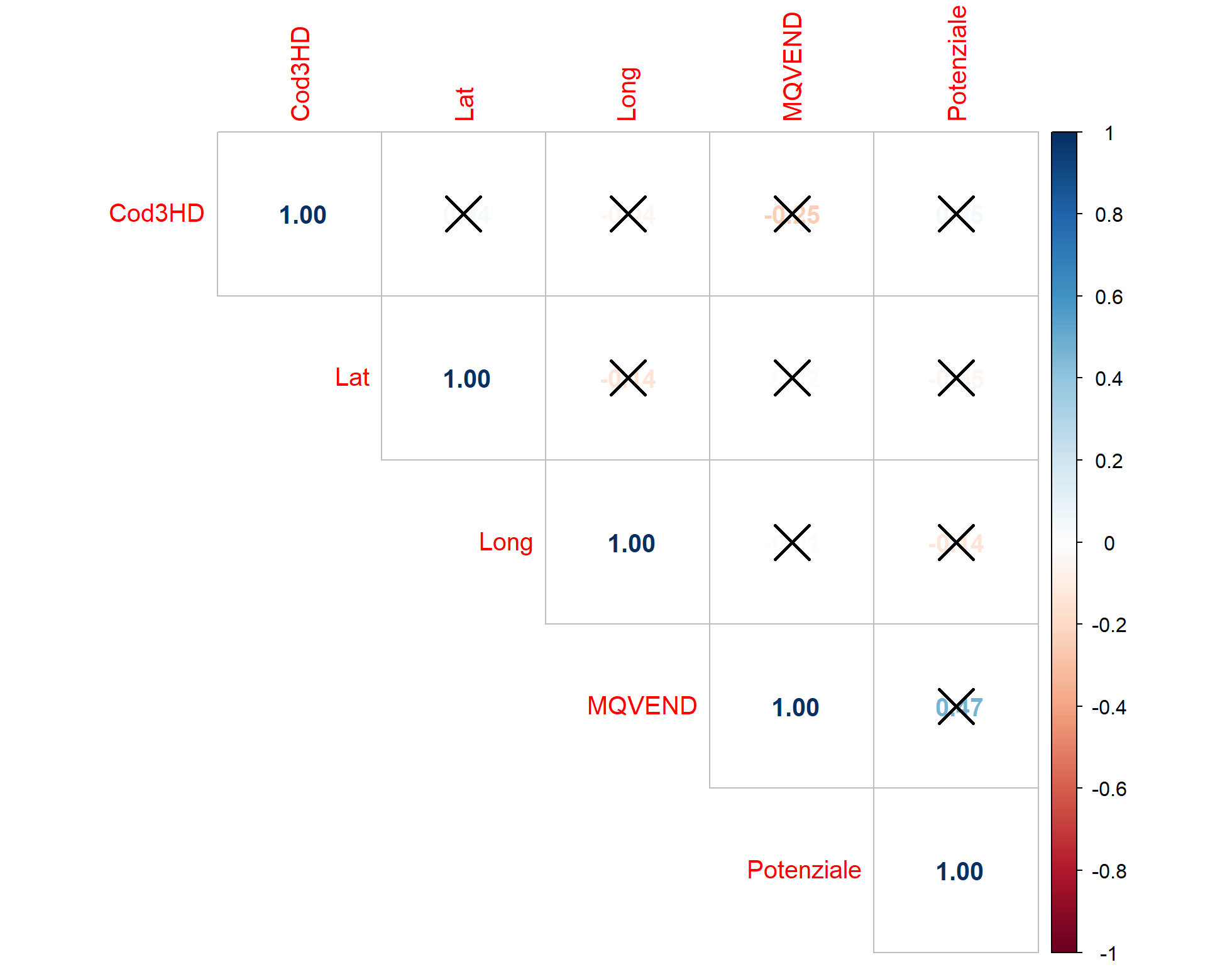
Other Pairs:

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).



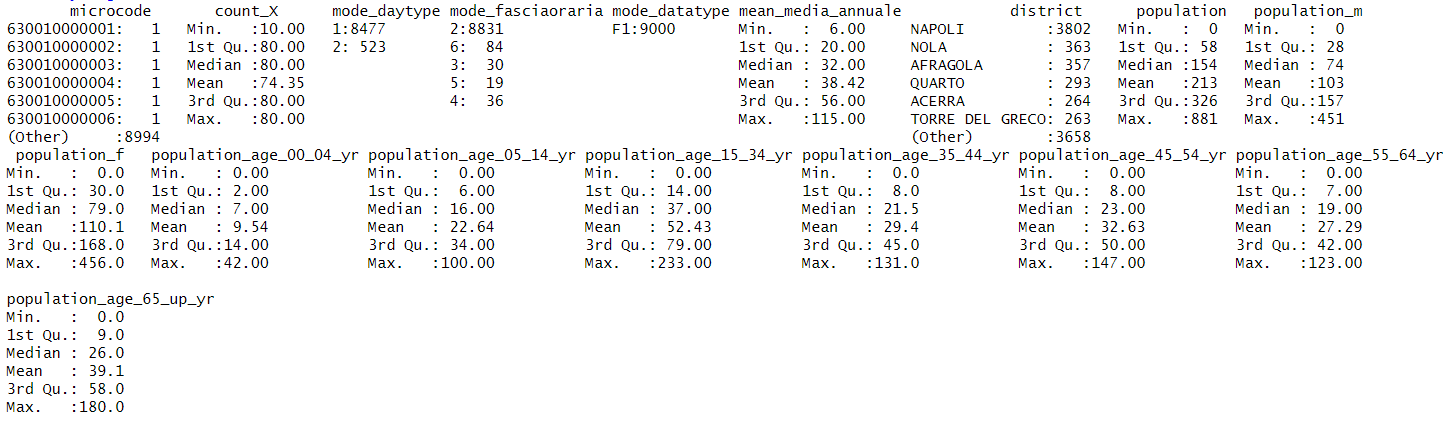
Correlation matrix of mdf after statistical test



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.



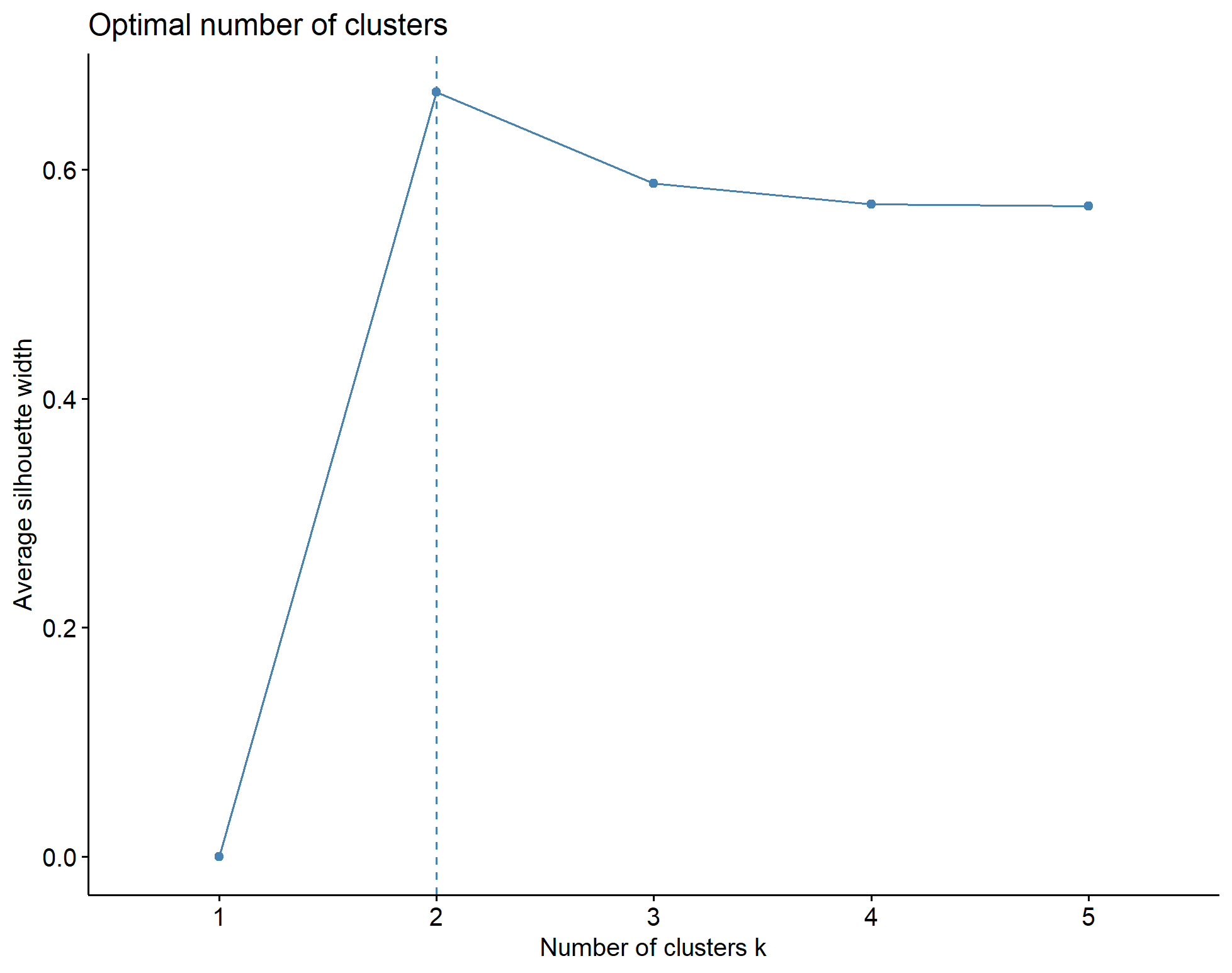
**Modeling**

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 and is a point among the clustered data, the Silhouette index for can be obtained by using equation as follows:

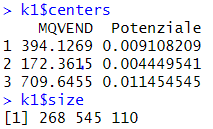
Where is the average distance between one point of the cluster and other points of the cluster, which can be obtained using the first equation, and is the minimum average distance between one point and other clusters that can be derived using the second equation.

In this method, the clustering algorithm has been implemented for different values of k (), 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 is the value that yields the maximum value of the Silhouette index average. In this Fig., this optimal value happens in (without considering k=2).



Number of optimal clusters based on Silhouette

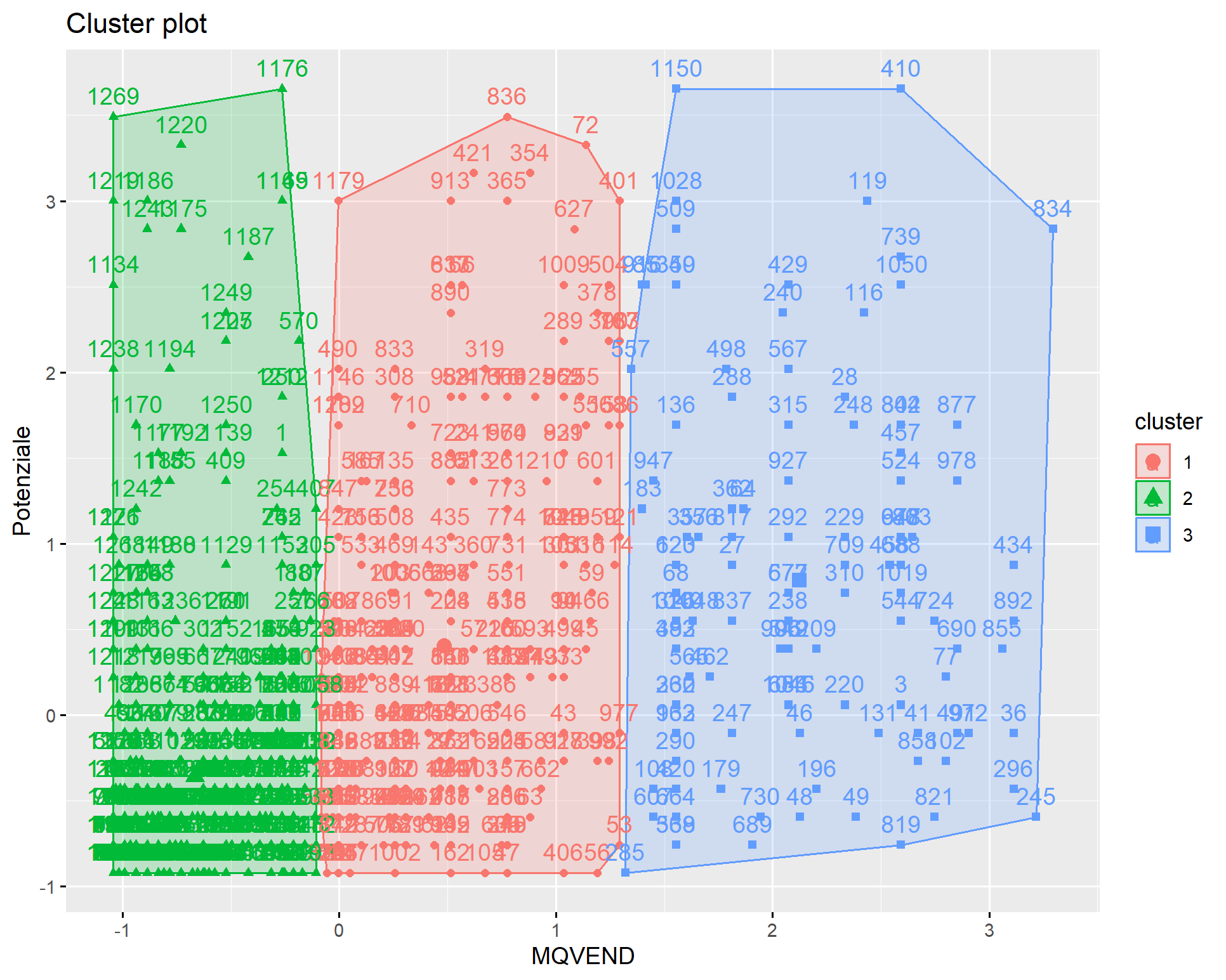
We have clustered the stores in three clusters according to potenziale (potential for sale) and MQVEND (store size) (Following Fig).



Result of clustering

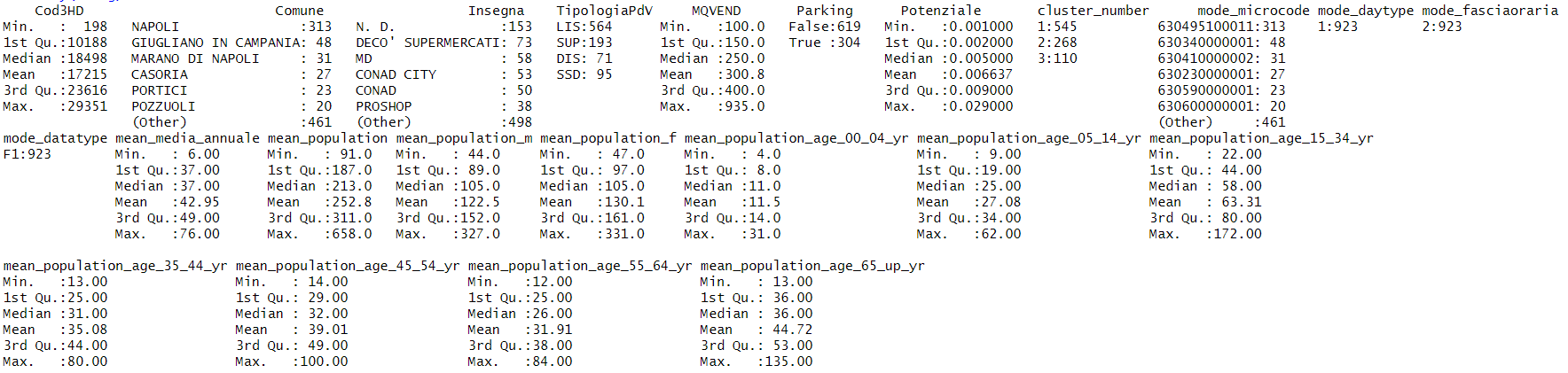
According to this figure we figure out that in second cluster we have more potential for sale. The distribution of stores in clusters shows that the second cluster has the best potential although the stores are not big stores, after that the first cluster has better potential and the last one is the 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.



summary of cmerg

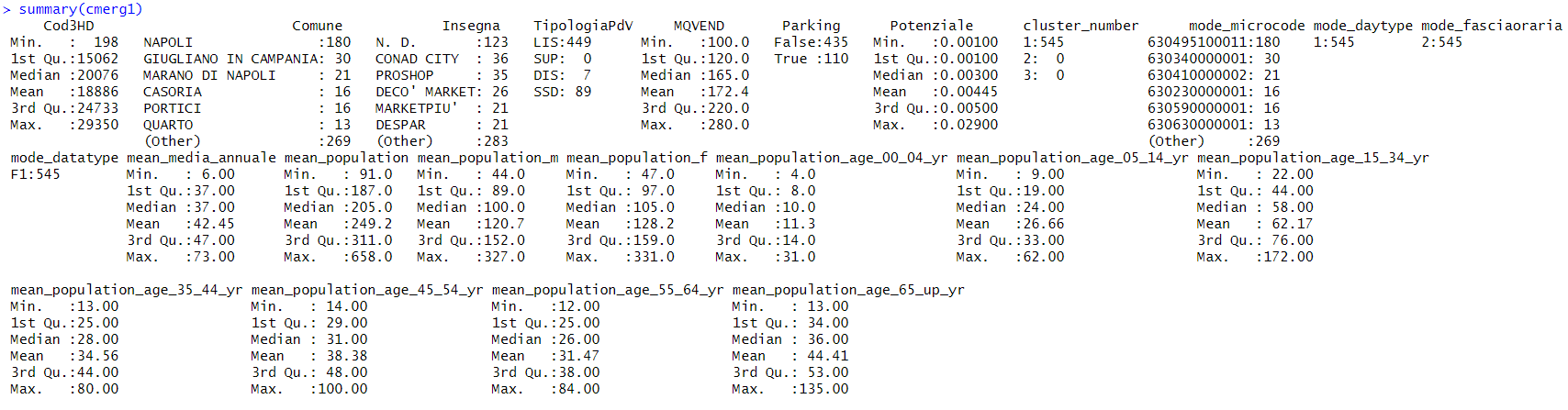
Then create three new data frames from each cluster that we discussed on them separately.

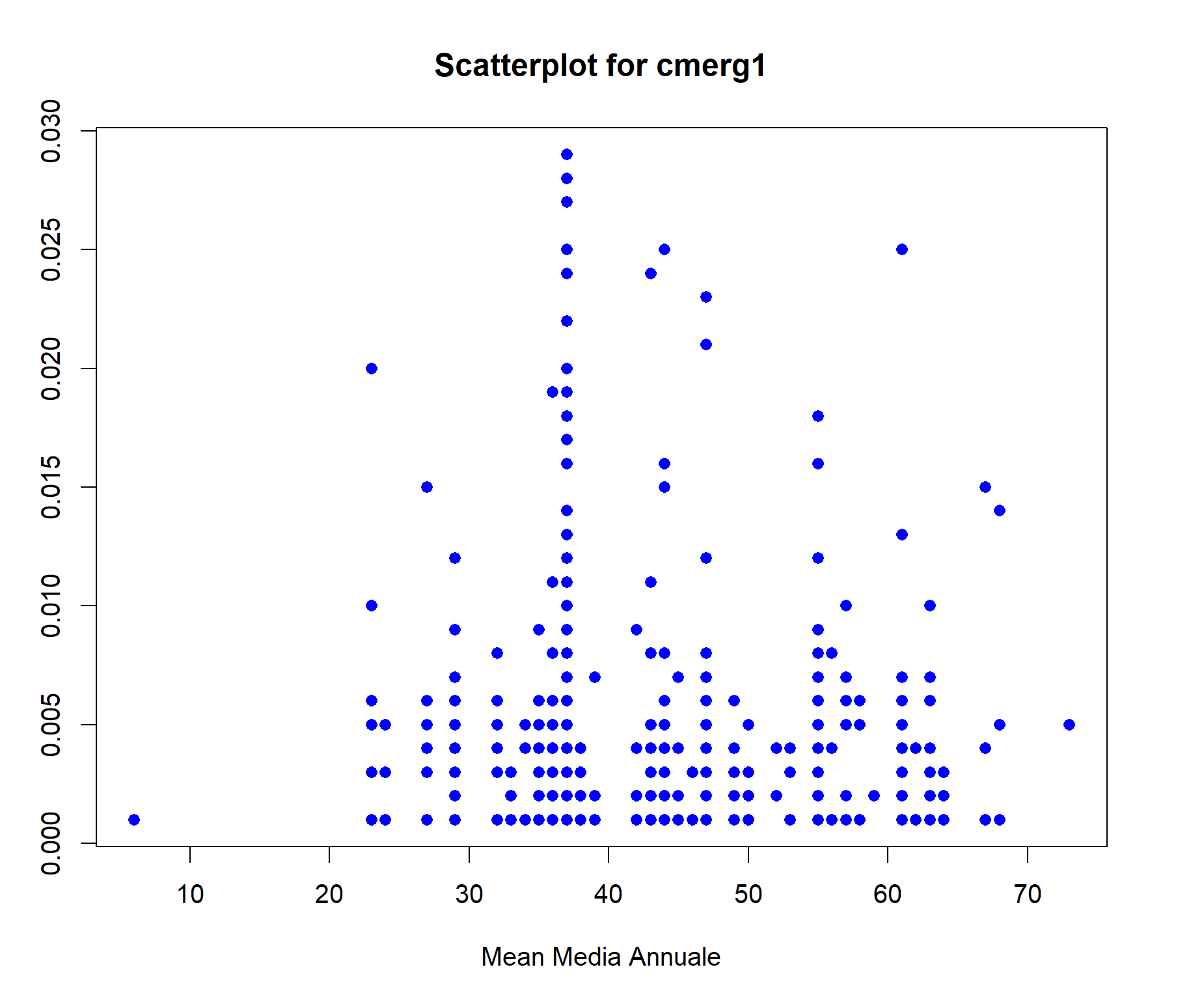
**Group by based on Cod3HD**

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.

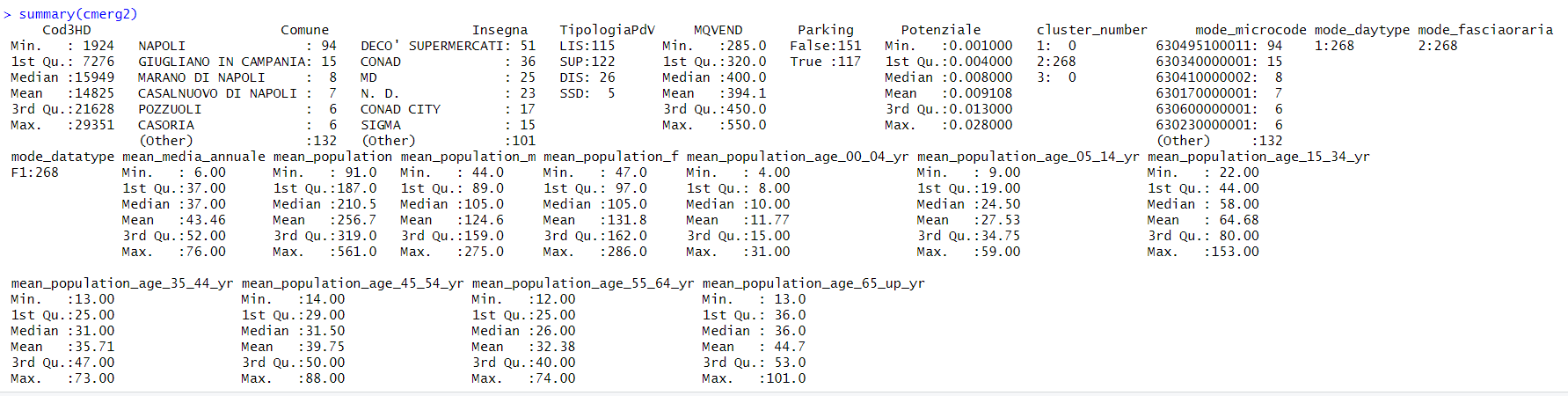
**Summary C MERG 1**

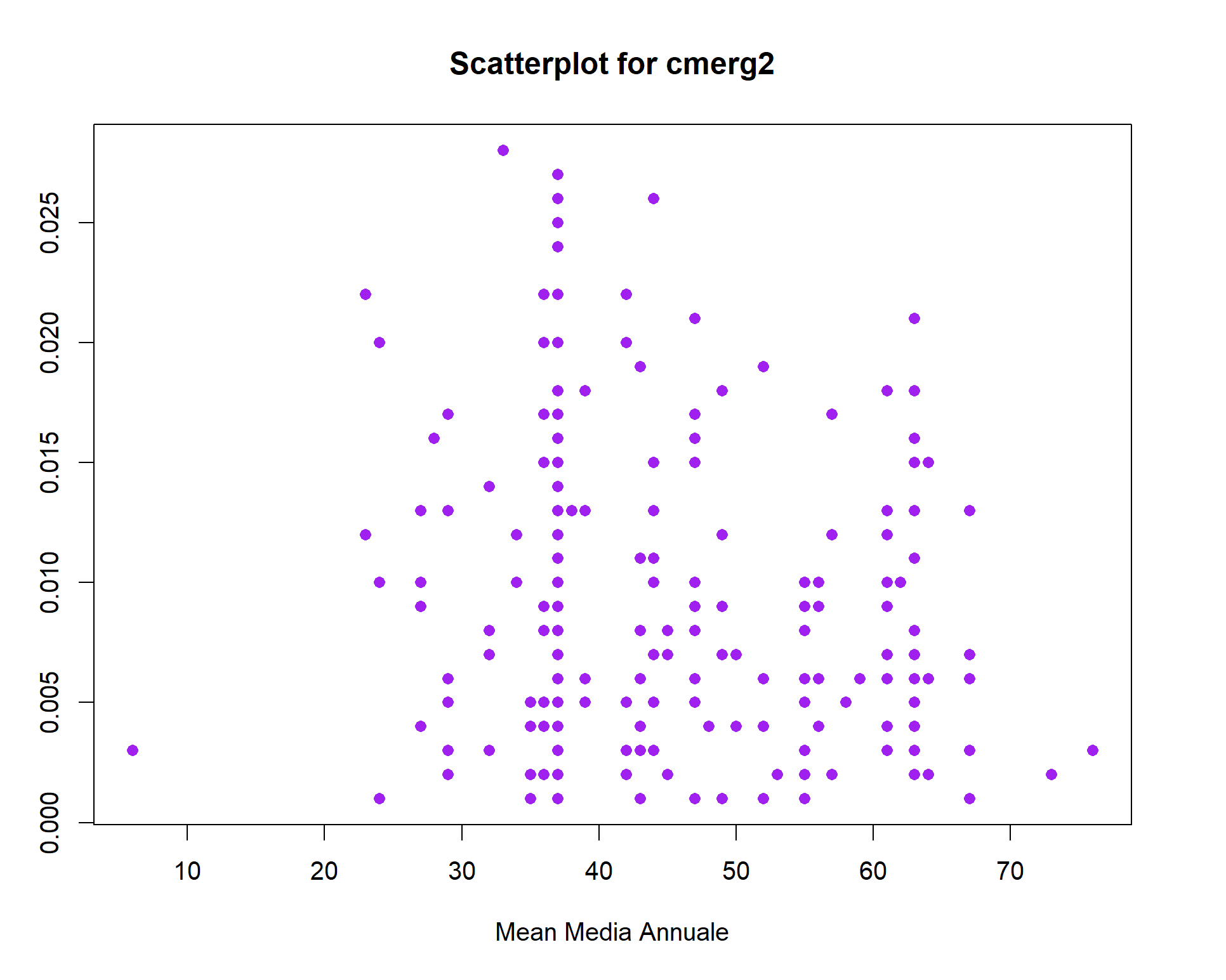




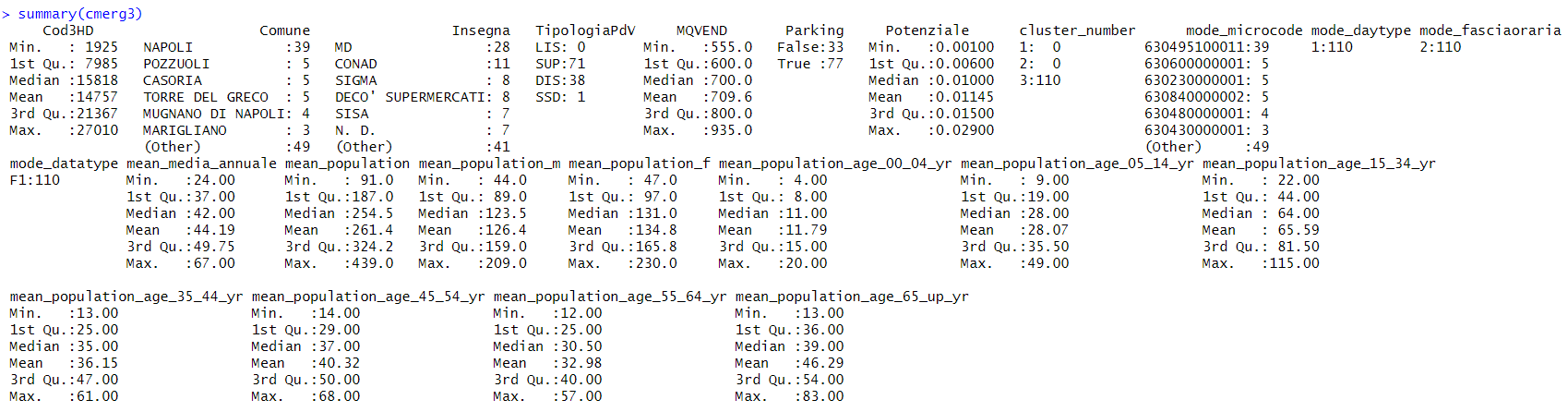
It shows a relationship between two variables: "Mean Media Annuale" on the x-axis and potenziale value on the y-axis. The data points are widely spread across the plot, which indicates variability in the relationship between the two variables. As "Mean Media Annuale" increases, the spread of "Potenziale"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

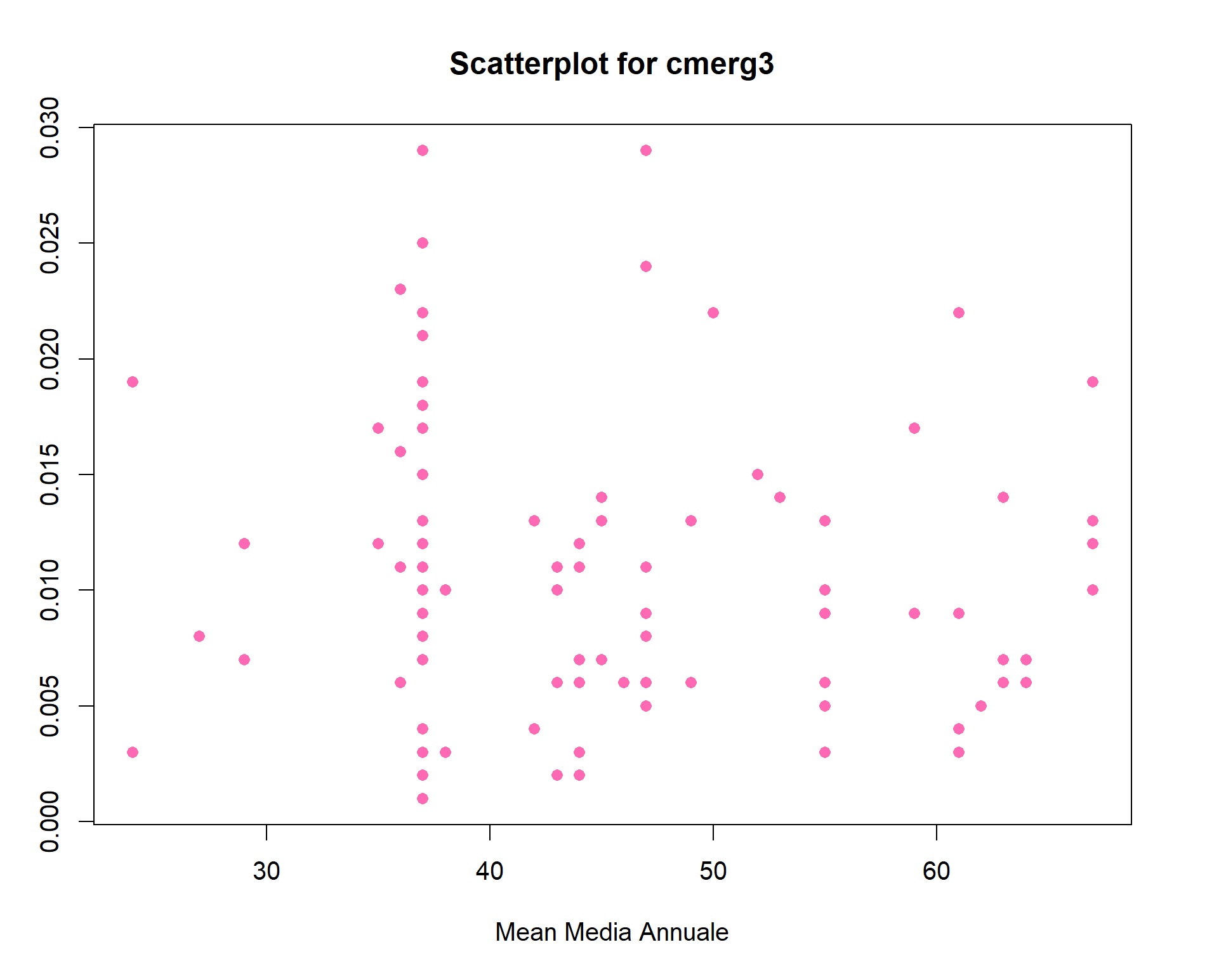
**Summary C MERG 2**





**Summary C MERG 3**



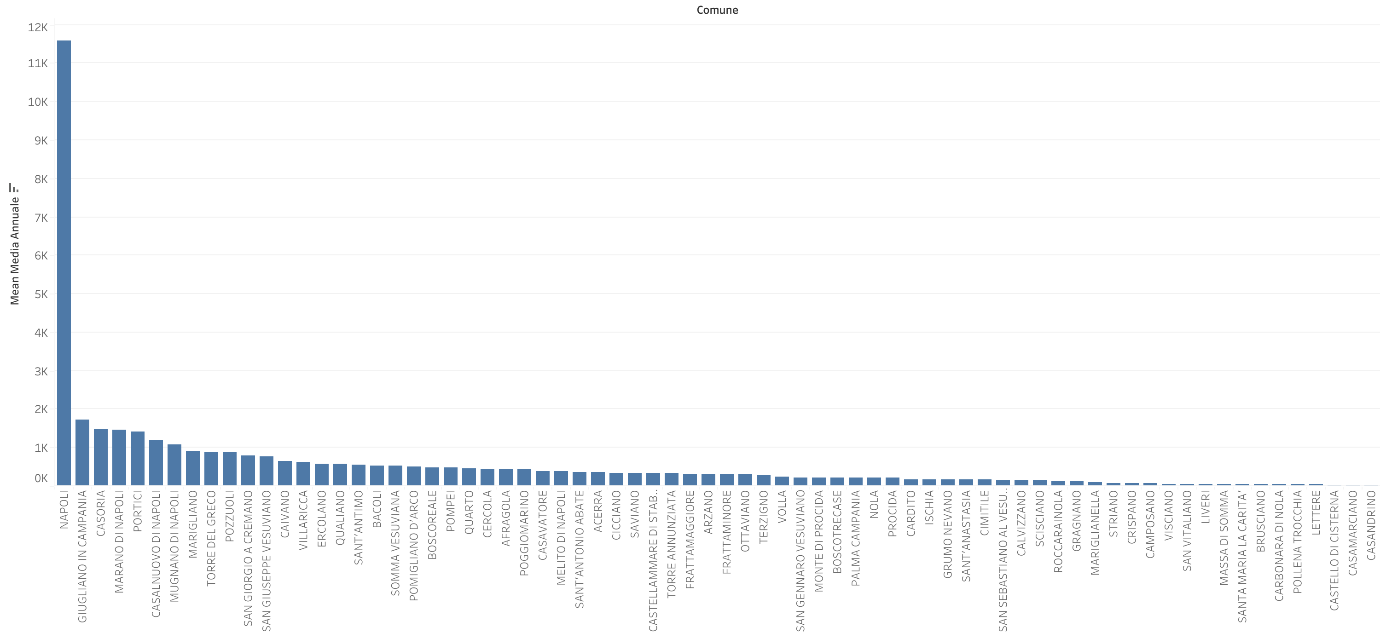


In the following figures, we analyze three datasets together. The first figure compares three datasets in terms of Mqvend, Potenziale, and Media Annuale. The chart demonstrates that although the average of media annual is very close for the three clusters, potential and store size vary significantly. 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.

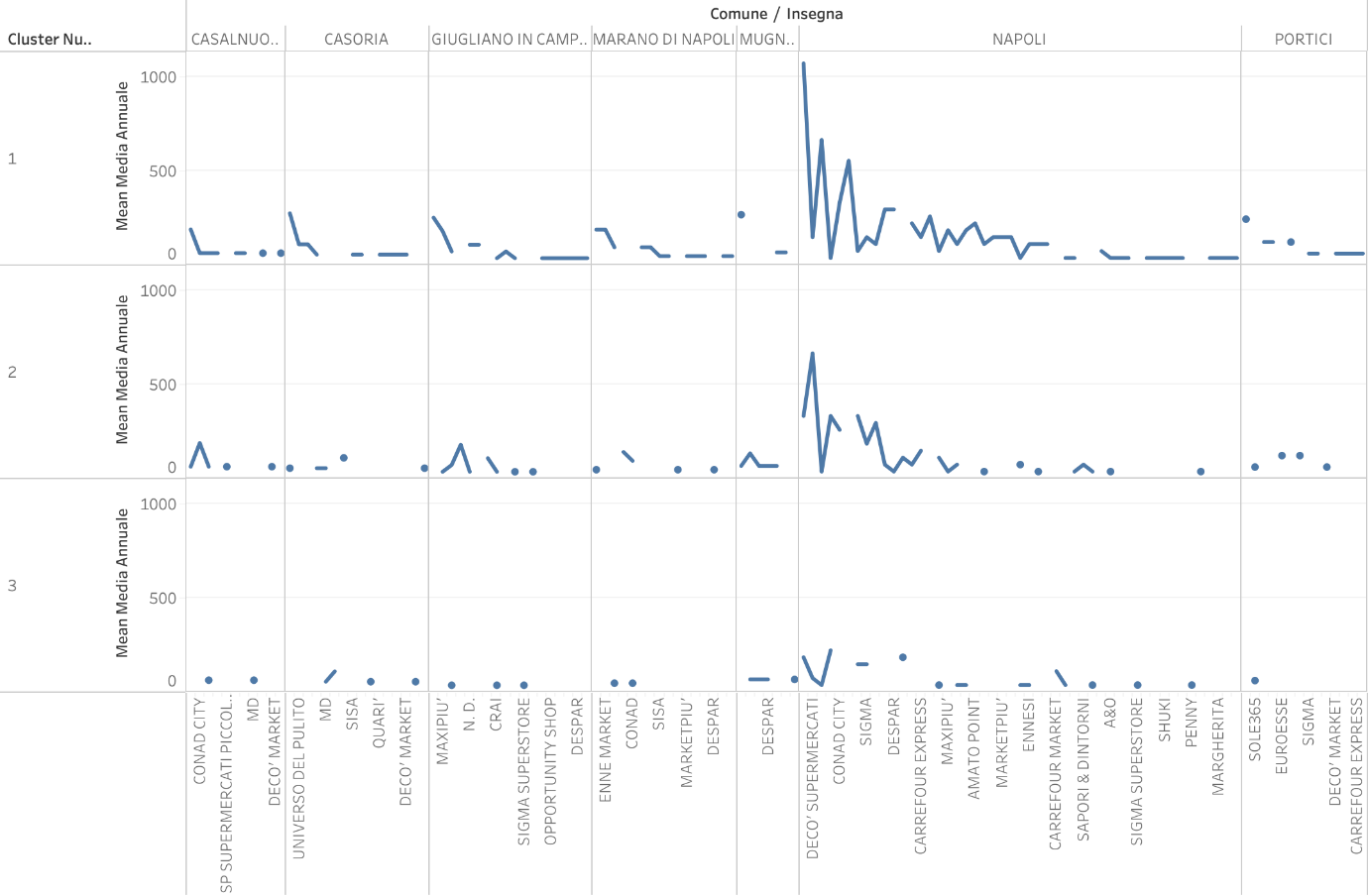


comparison of three datasets in terms of Mqvend, Potenziale, and Media Annuale

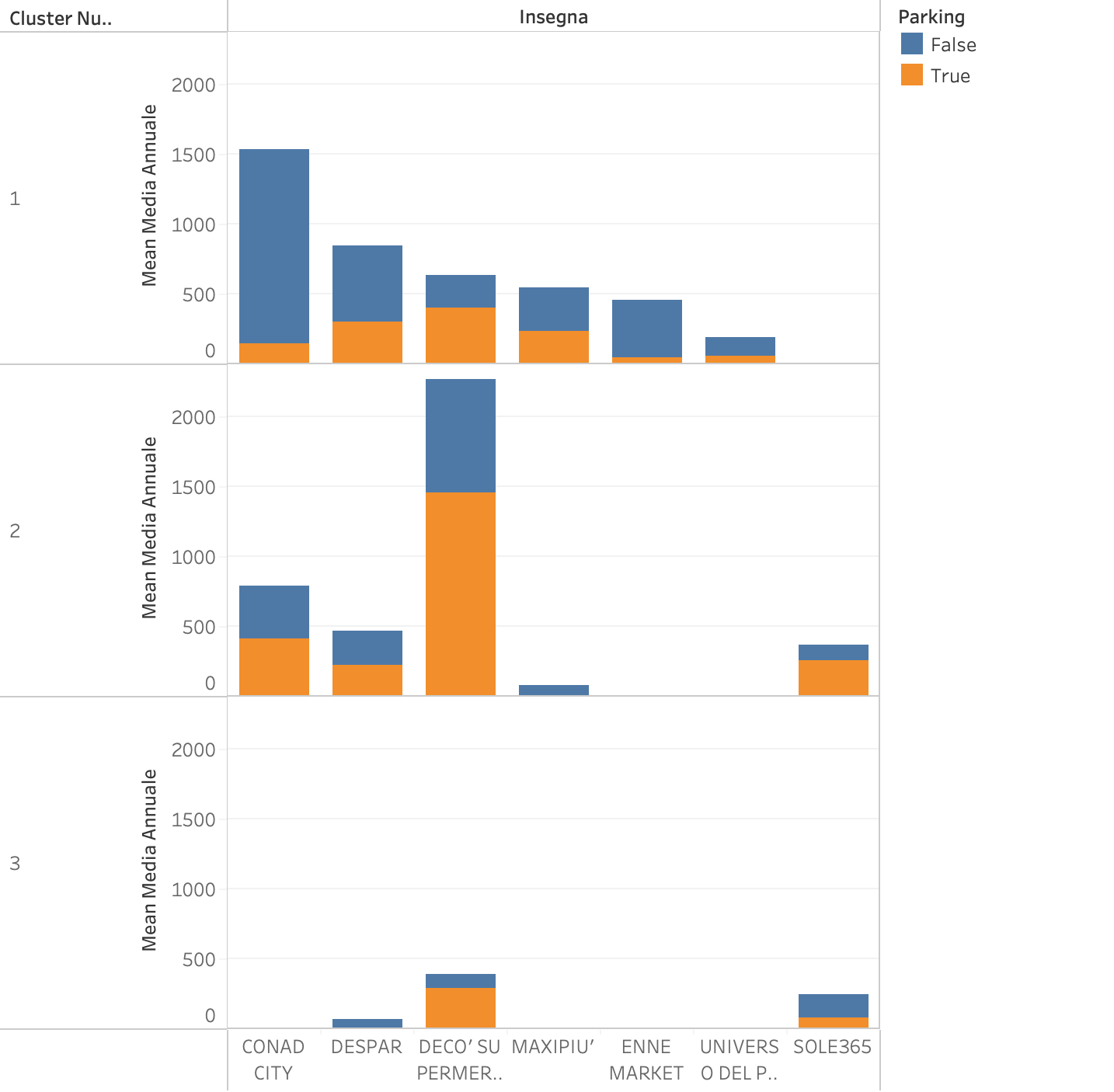
This following chart displays cities that attract more attention. Therefore, we will use the top 6 of them for further analysis.



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:

Where 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.

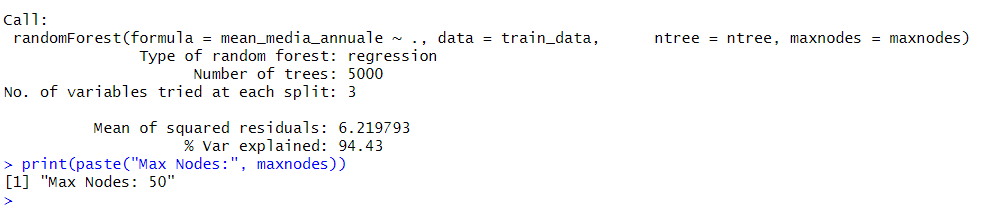
The error, say , is computed on the observations in the held-out fold.

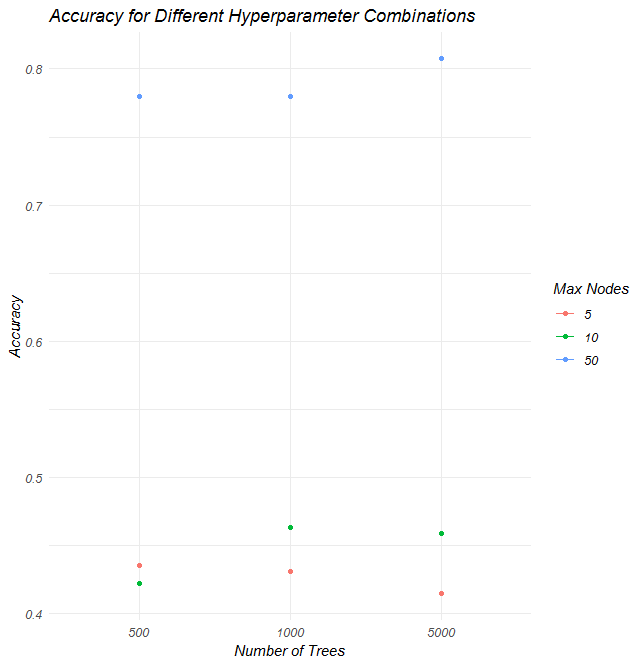
• This procedure is repeated 𝑉 times; each time, a different group of observations is treated as a

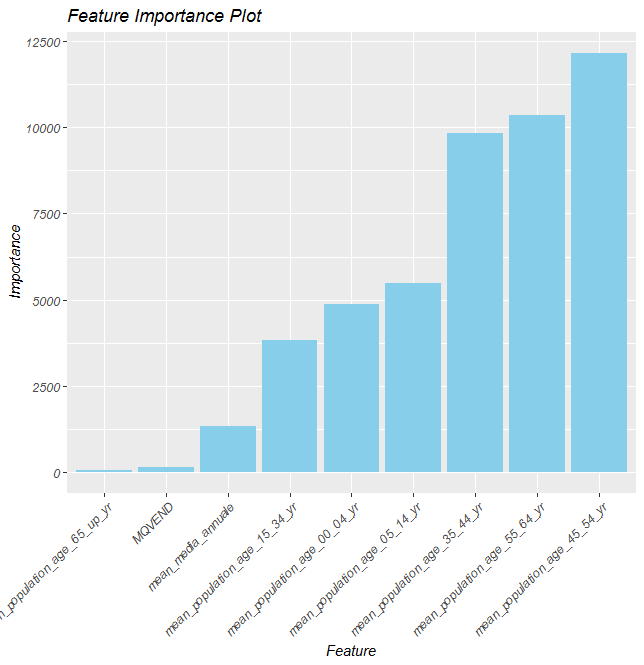
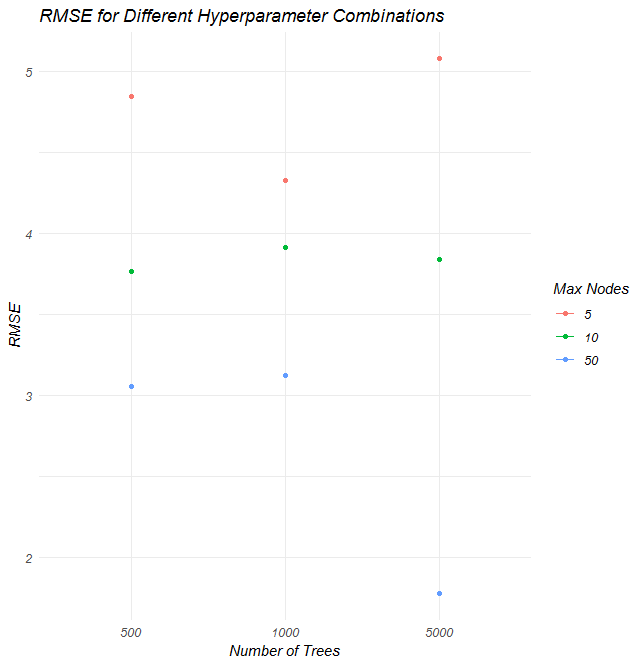
validation set.

• The 𝑉 -fold CV error is

For rfcmerg1:



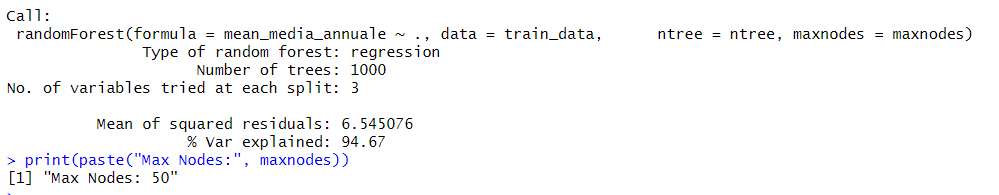


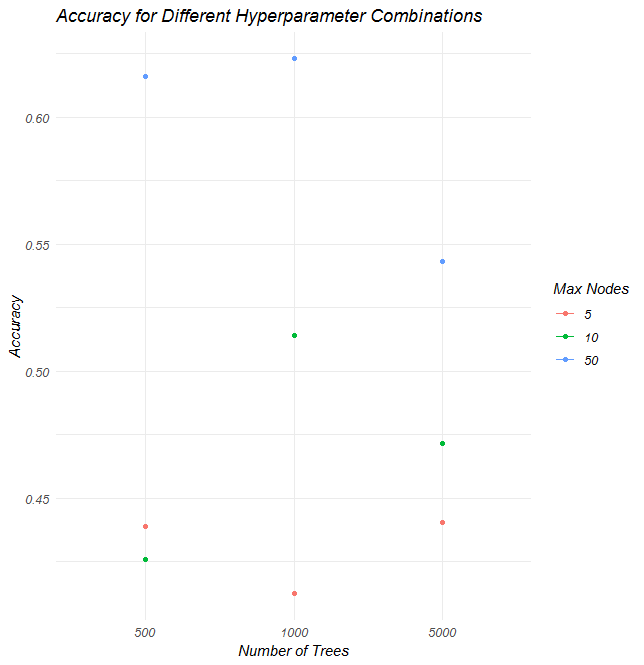


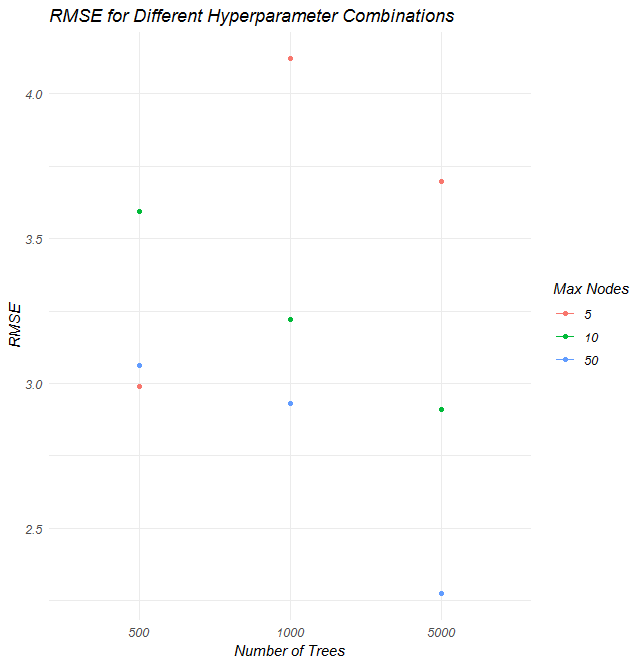
As the bar chart shows the most important feature is mean-population-age-45-54-yr,it means this feature can paly an important role in the dataset. the most influential feature is belong to this age.

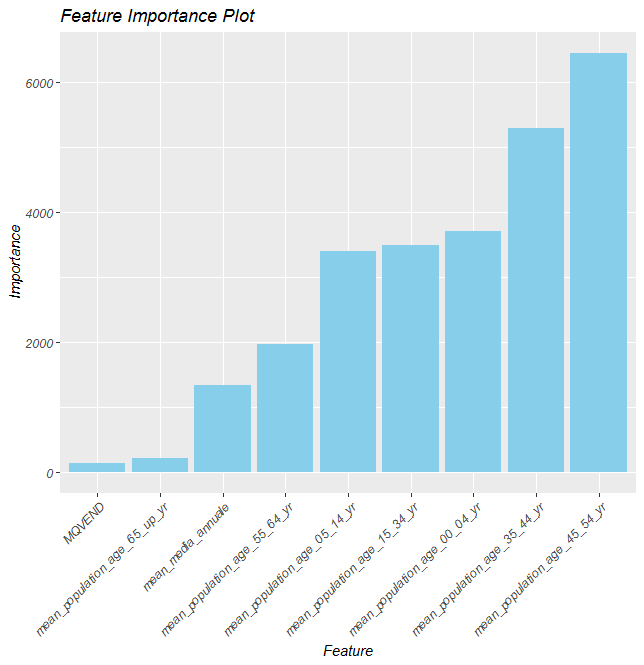
The height of each bar corresponds to the importance score of each feature, which quantifies the contribution of the feature to the model's predictive power. The scores are presumably calculated based on the mean decrease in impurity (such as Gini impurity) that results from splits over a particular feature, a common metric in tree-based models

For rfcmerg2:









For rfcmerg3:

