The housing cost among Italy's regions and its relationship with other factors

Research questions covered

- Unsupervised Learning Model:

“ What factors influence the similarities among the 18 regions of Italy and how could they be classified? “

- Supervised Learning Model:

“ What are the key factors valuable in estimating Italy's overall cost of housing? “

Dataset introduction

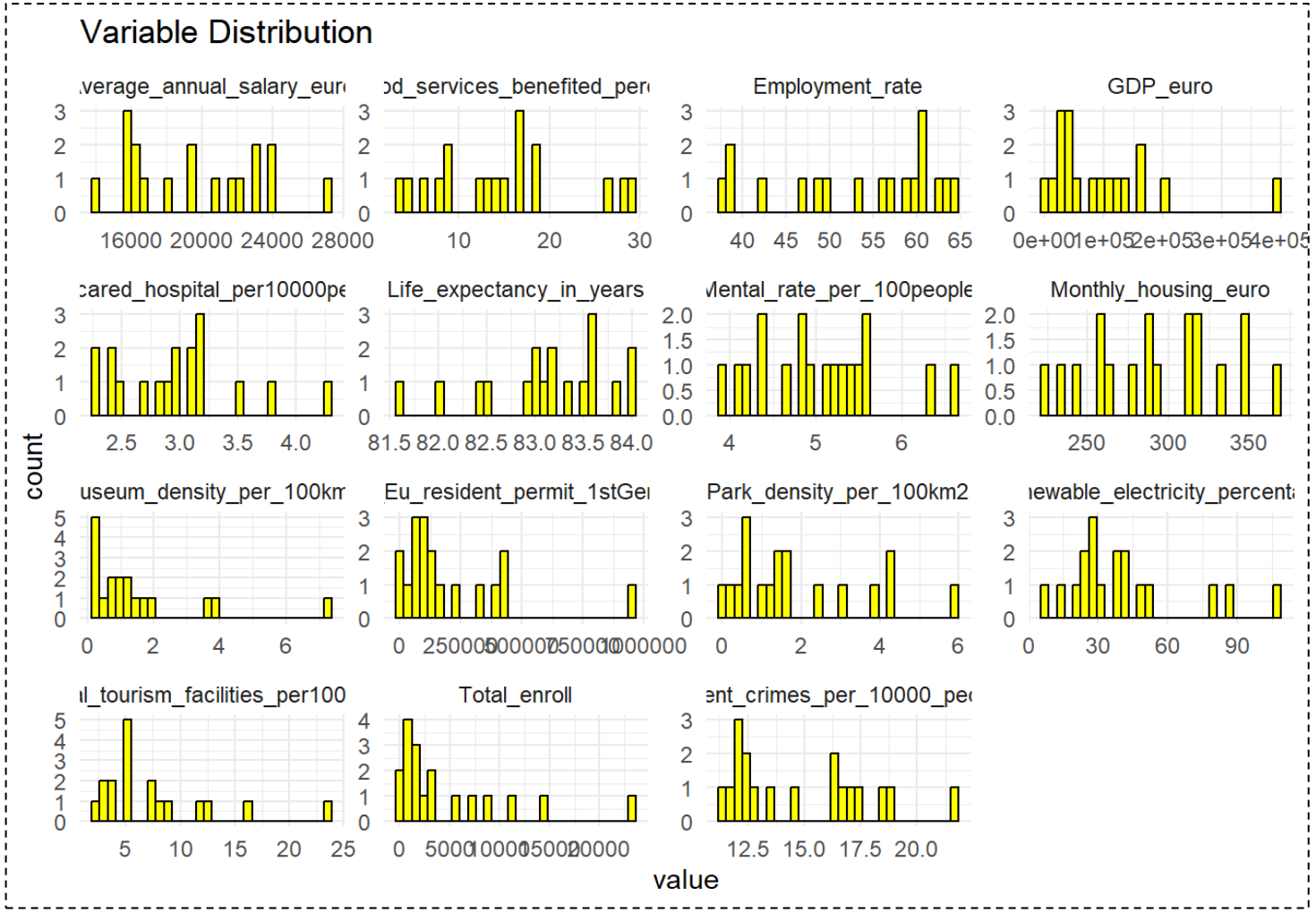
- The "merged_final" dataset is the result of merging 7 different smaller datasets, including:
  - employ_rate.csv
  - enroll.csv
  - gdp.csv
  - housing.csv
  - mental.csv
  - Residence permits of non-EU citizens.csv
  - raw.csv where the project gets information about "Average annual salary", "Violent crimes reported", "Museum density", "Park density", "Life expectancy", "Childhood services benefited", "Rural tourism facilities", "Renewable electricity rate", "High-care hospital rate"

```
merged_final<-read.csv("merged_final.csv")
head(merged_final)
```

```
##              Territory Monthly_housing_euro GDP_euro Employment_rate
## 1              Abruzzo                  264  32898.4        53.08077
## 2           Basilicata                  259  12901.9        46.74003
## 3             Calabria                  234  33443.9        38.91868
## 4             Campania                  290 109504.1        38.73587
## 5       Emilia-Romagna                  346 163994.2        64.71632
## 6 Friuli-Venezia Giulia                  311  38735.4        60.76439
##   Mental_rate_per_100people Total_enroll Non_Eu_resident_permit_1stGenuary
## 1                       5.1         1225                             58025
## 2                       5.3           32                             11976
## 3                       6.6         1118                             52425
## 4                       4.4         3175                            176897
## 5                       5.2        10977                            420312
## 6                       4.1         1959                             83895
##   Average_annual_salary_euro Violent_crimes_per_10000_people
## 1                    18001.2                            12.7
## 2                    16543.2                            12.2
## 3                    13978.4                            12.1
## 4                    15839.8                            21.7
## 5                    23756.6                            18.3
## 6                    22873.8                            12.6
##   Museum_density_per_100km2 Park_density_per_100km2 Life_expectancy_in_years
## 1                      0.14                     0.7                     83.3
## 2                      0.23                     4.2                     82.5
## 3                      0.28                     0.5                     82.4
## 4                      3.61                     1.7                     81.6
## 5                      1.15                     0.7                     83.6
## 6                      1.41                     6.0                     83.5
##   Childhood_services_benefited_percentage Rural_tourism_facilities_per100km2
## 1                                     9.1                                5.1
## 2                                     7.7                                2.0
## 3                                     3.1                                3.8
## 4                                     4.0                                5.4
## 5                                    28.7                                5.3
## 6                                    28.4                                8.6
##   Renewable_electricity_percentage Highcared_hospital_per10000people
## 1                             48.2                               3.0
## 2                            109.1                               2.8
## 3                             80.3                               2.4
## 4                             29.0                               2.4
## 5                             20.5                               3.2
## 6                             29.6                               2.5
```

Preprocessing Data

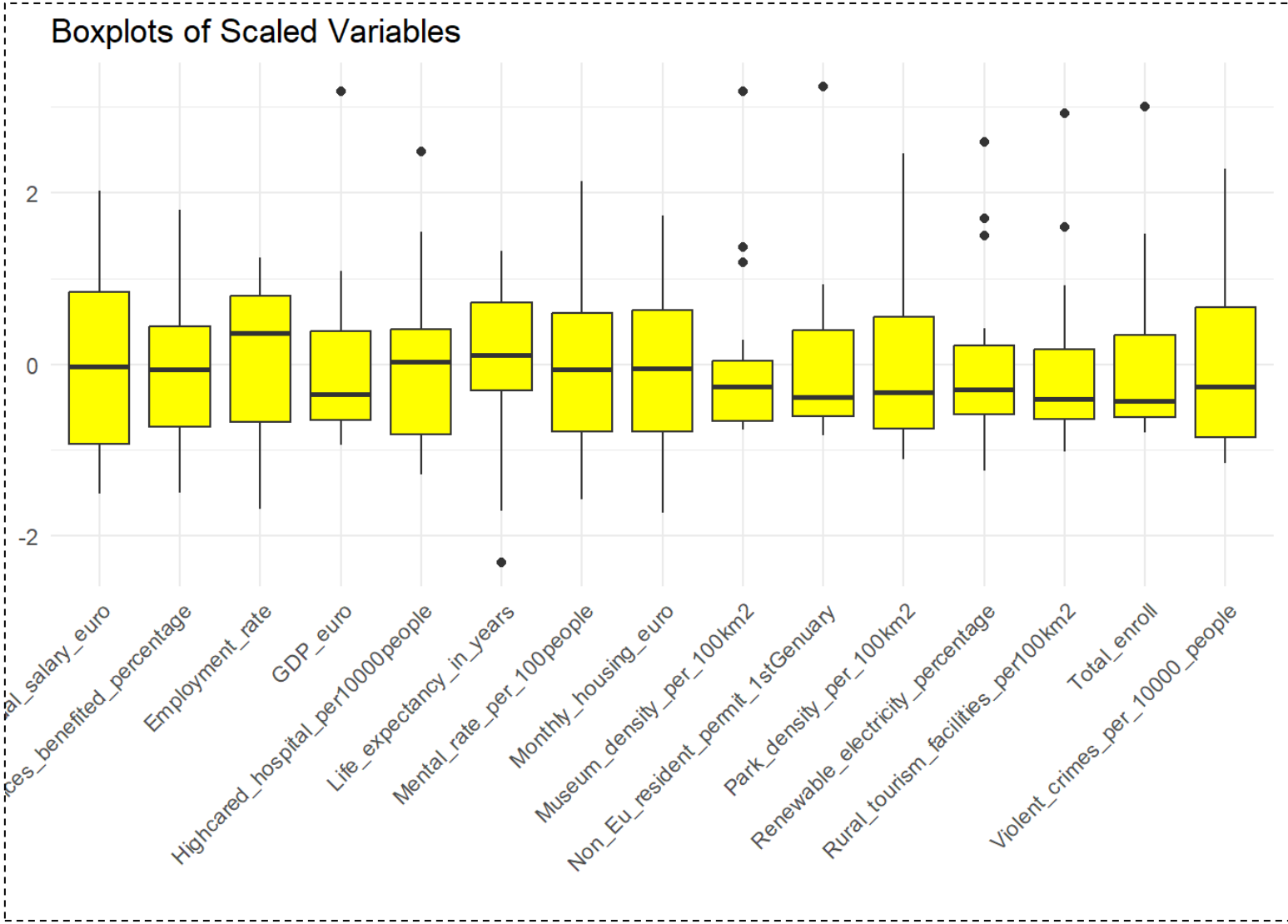
Variables distribution



The variables seem to be distributed differently.

Outliers checking

- According to the plot, there are not many outliers highlighted.
- However, to return best result, the project will choose the Supervised models which are not sensitive with outliers, namely Ridge Regression and Random forest

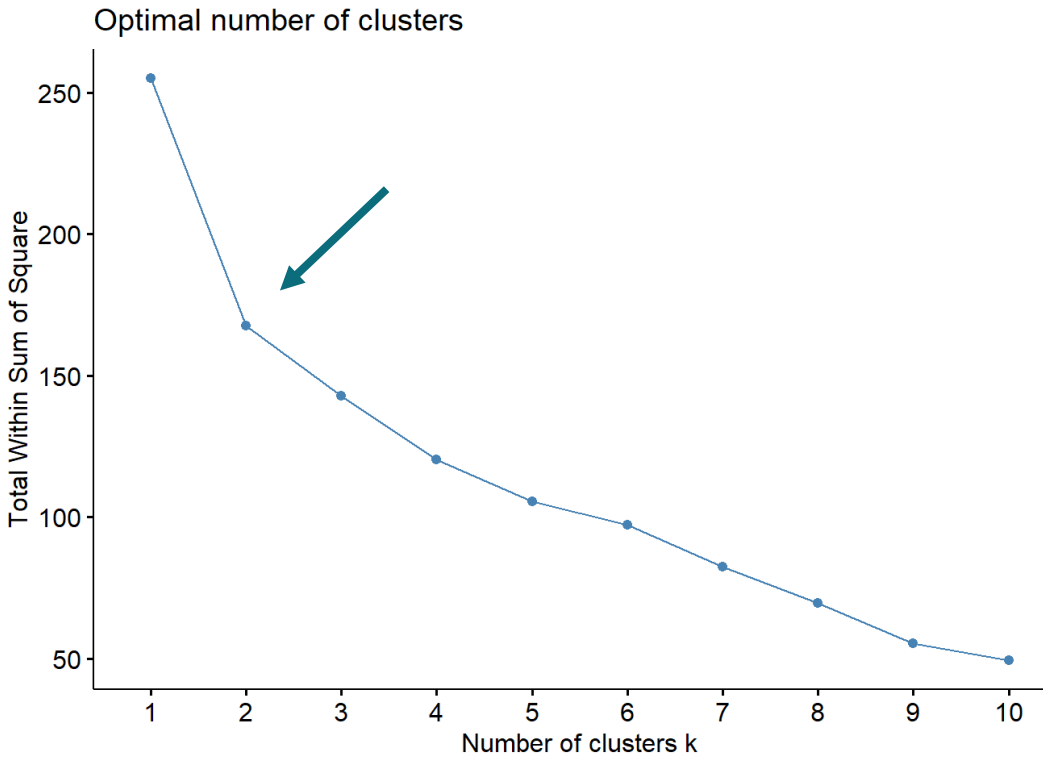


Unsupervised Learning

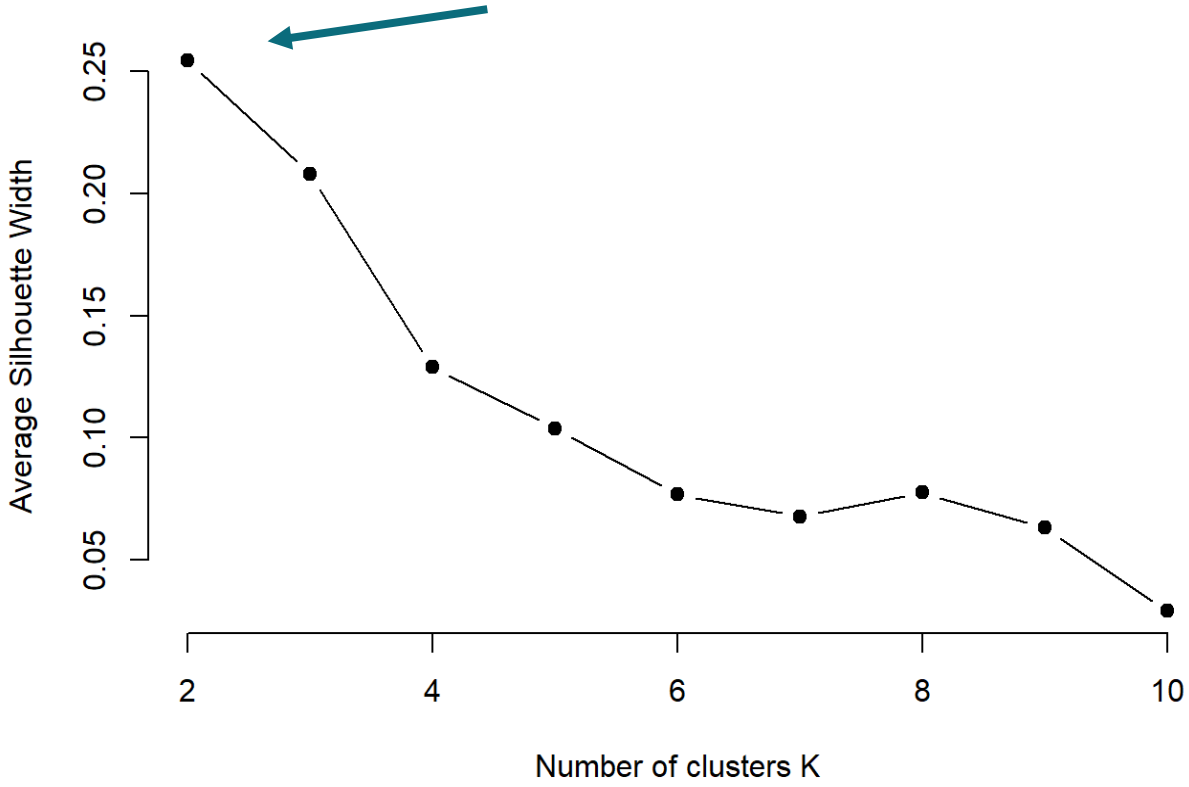
K-medoids clustering & Hierarchical clustering

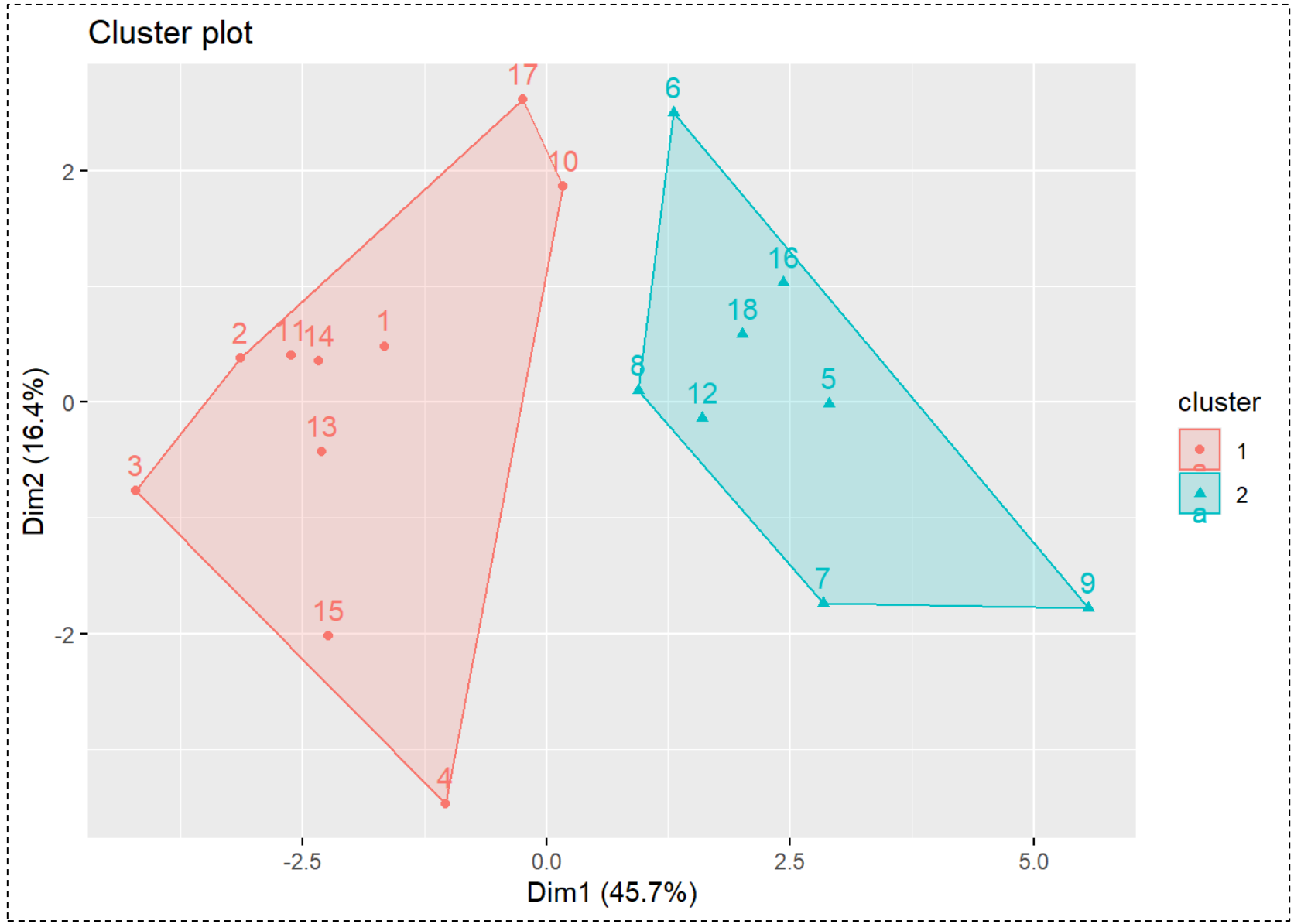
Optimal k-value determination **K=2**

Elbow method



Silhouette method



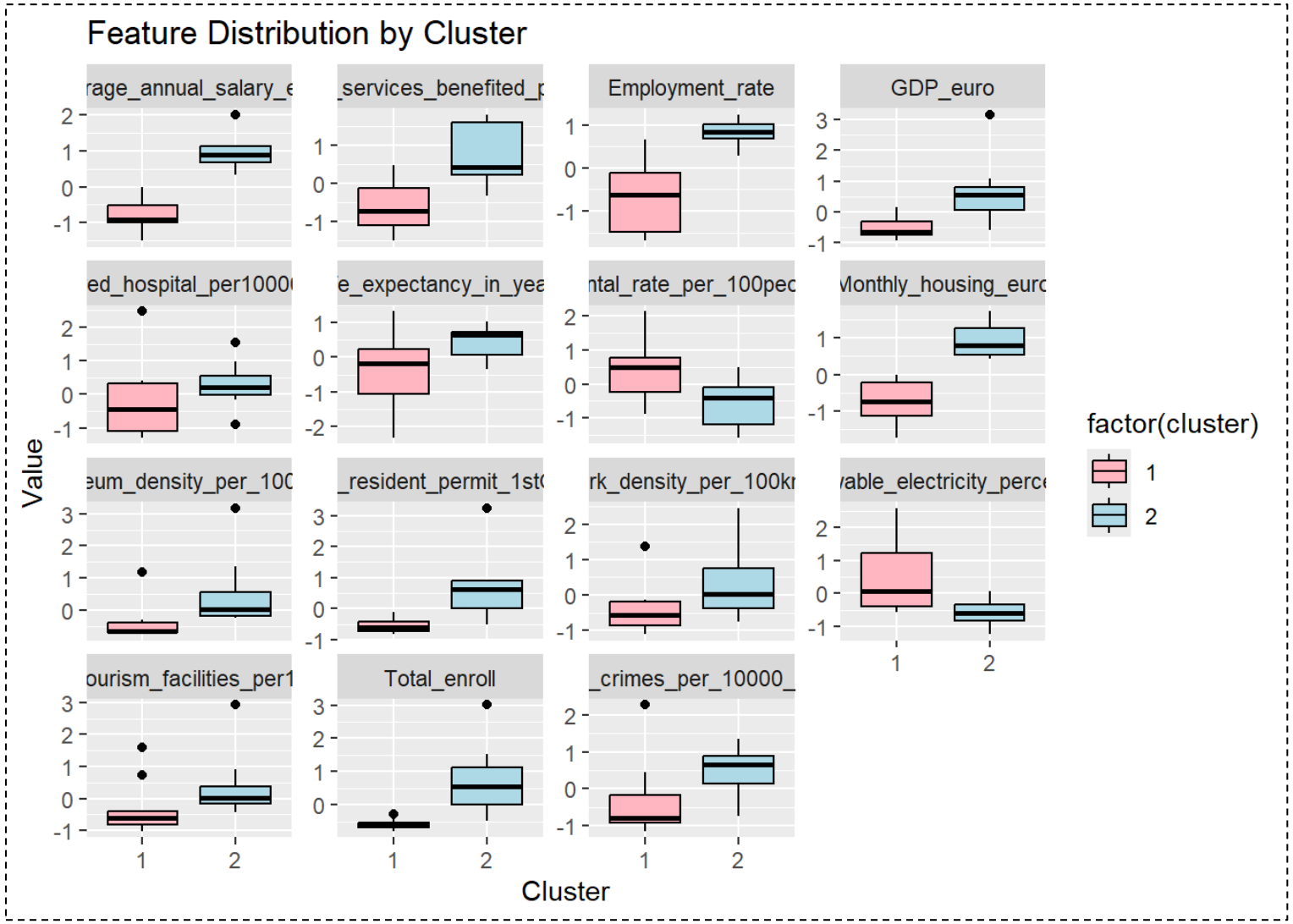

Cluster plot
Dim2 (16.4%)
Dim1 (45.7%)
2
0
-2
-2.5
0.0
2.5
5.0
17
10
6
16
18
2
11
14
1
8
12
5
13
3
7
9
15
4
cluster
1
2

- This figure shows the information of K-medoids centroids of two clusters
- *Abruzzo and Piemonte* are the medoids of cluster 1 and cluster 2, respectively

```
#view results
kmed
```

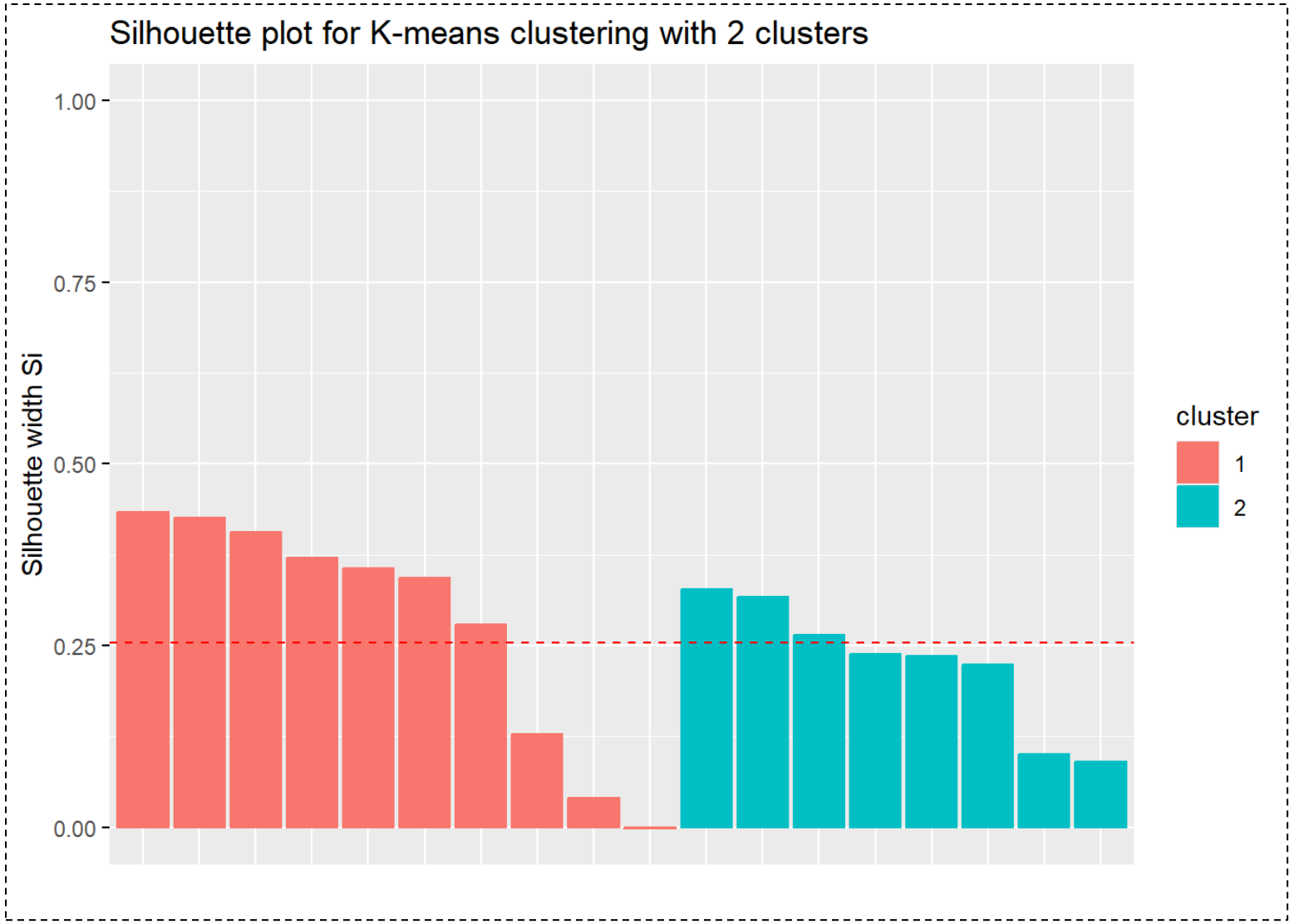
```
## Medoids:
##      ID Monthly_housing_euro   GDP_euro Employment_rate
## [1,]  1           -0.6968772 -0.6631324     -0.03603692
## [2,] 12            1.3014617  0.4327720      0.76990225
##      Mental_rate_per_100people Total_enroll Non_Eu_resident_permit_1stGenuary
## [1,]                0.07616598   -0.5991780                         -0.6158775
## [2,]               -0.47222905    0.6694967                          0.1326106
##      Average_annual_salary_euro Violent_crimes_per_10000_people
## [1,]                 -0.4252301                      -0.7229501
## [2,]                  1.1215582                       0.4419056
##      Museum_density_per_100km2 Park_density_per_100km2 Life_expectancy_in_years
## [1,]                -0.7591887              -0.7485785                0.2603998
## [2,]                -0.2338115               1.1245463               -0.3443998
##      Childhood_services_benefited_percentage Rural_tourism_facilities_per100km2
## [1,]                              -0.72464346                        -0.4510742
## [2,]                              -0.01575312                        -0.4326212
##      Renewable_electricity_percentage Highcared_hospital_per10000people
## [1,]                        0.26913696                         0.0313492
## [2,]                        0.05569081                         0.2194444
## Clustering vector:
##  [1] 1 1 1 1 2 2 2 2 2 1 1 2 1 1 1 2 1 2
## Objective function:
##    build     swap
## 3.150401 3.150401
##
## Available components:
##  [1] "medoids"    "id.med"     "clustering" "objective"  "isolation"
##  [6] "clusinfo"   "silinfo"    "diss"       "call"       "data"
```

Variables distribution by clusters



- The regions which are put in cluster 2 could be considered as *more developed* than the others in cluster 1.
- Especially, the expenditures on rent in those regions in the second cluster seem to be more costly than those in the first one.

Clusters checking

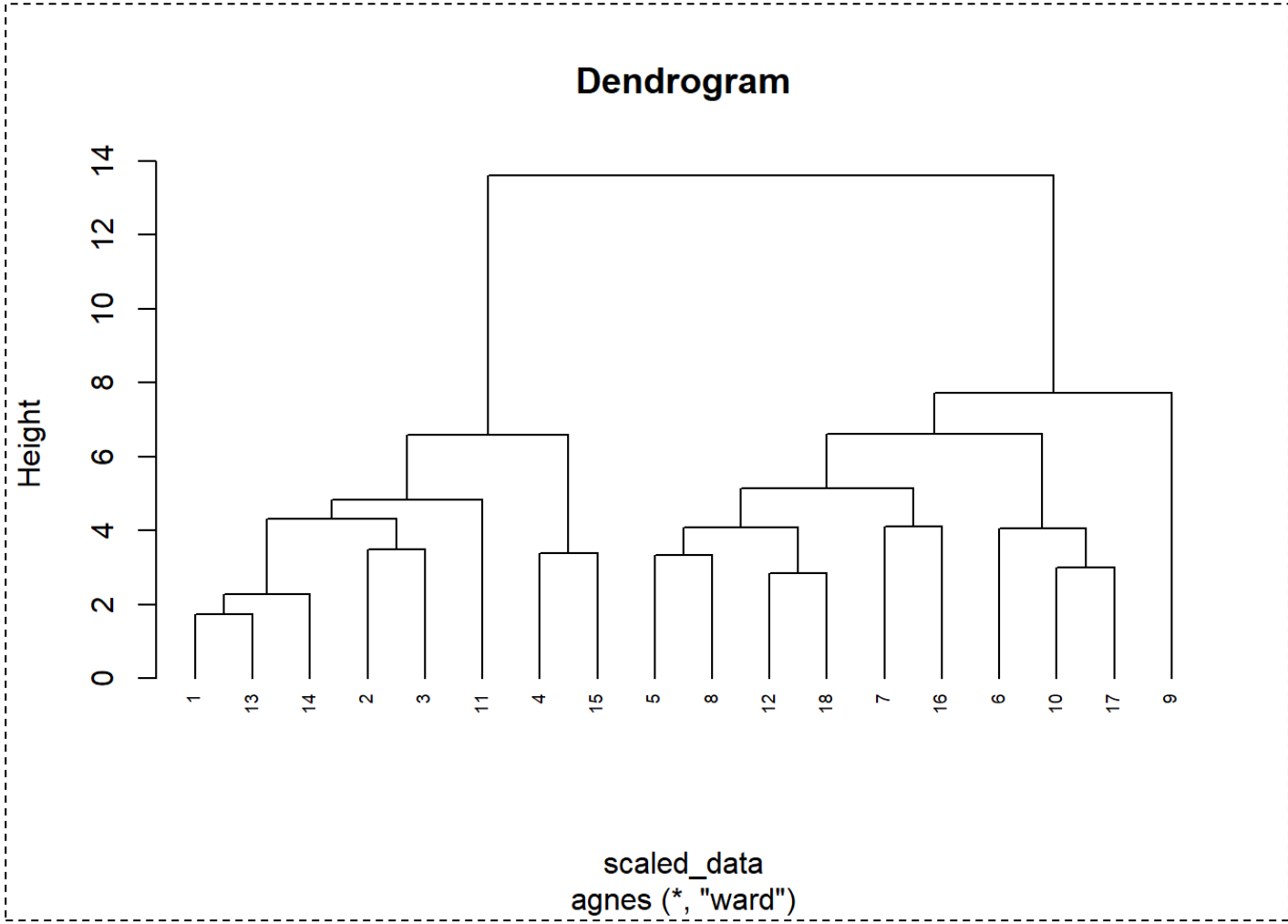


- Most of the data are reasonably well-clustered and placed properly

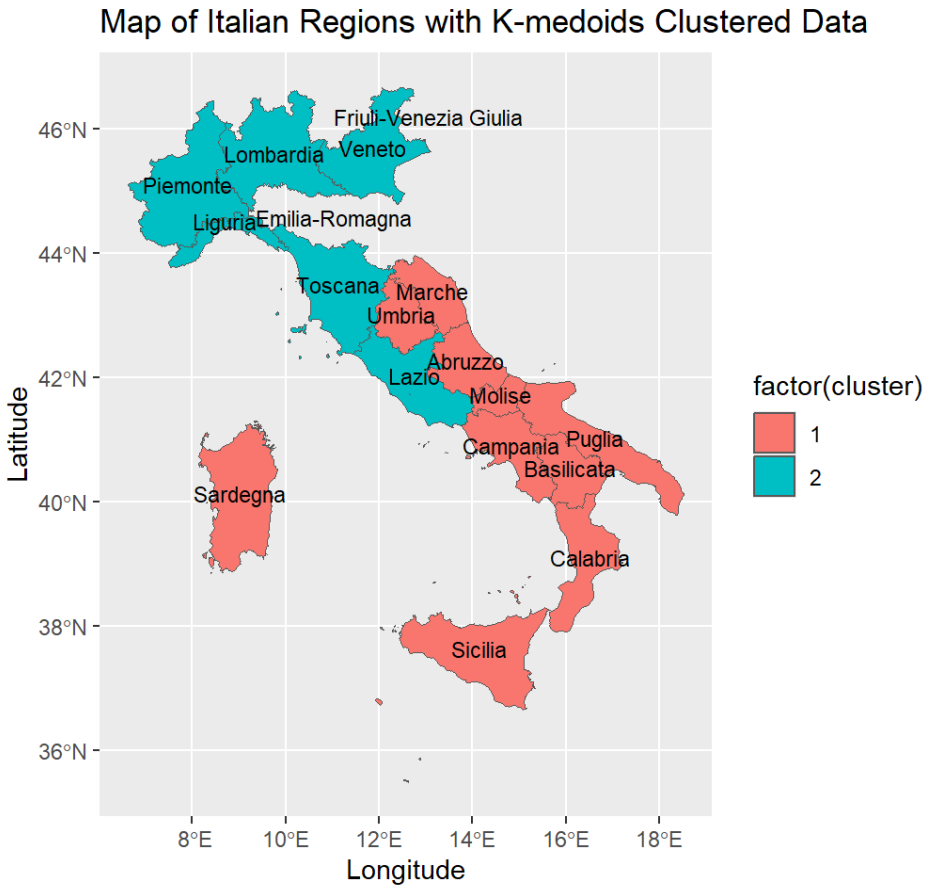
Define Linkage methods

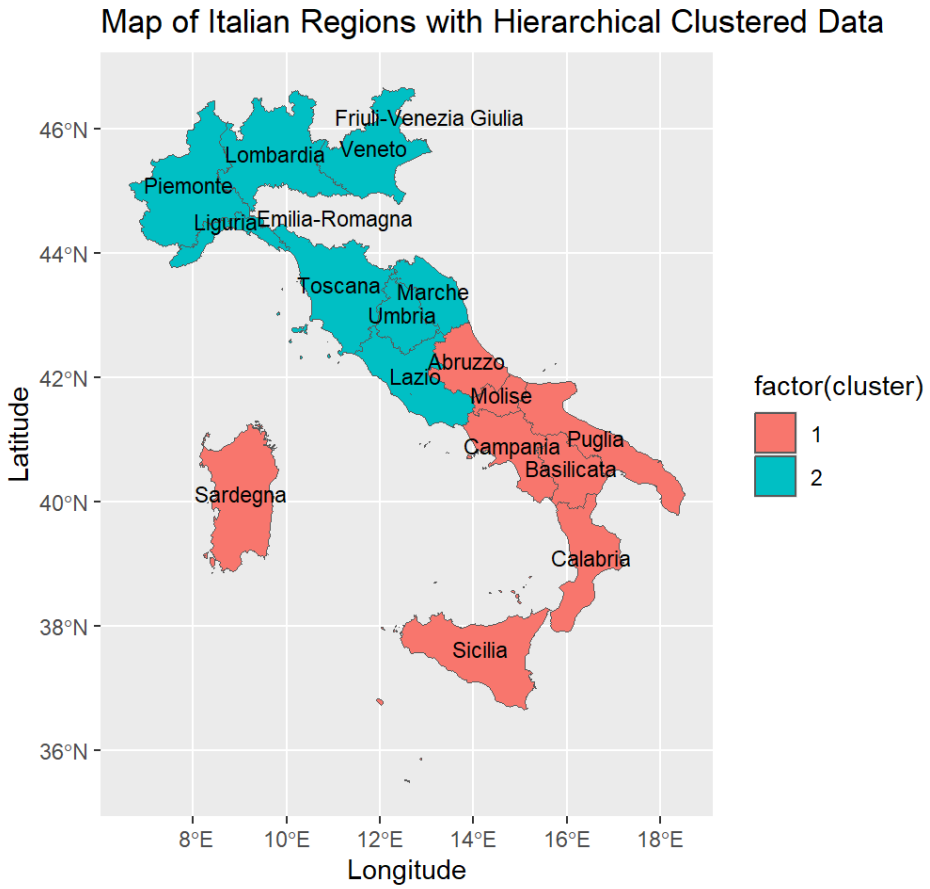
Average	0.5337248
Single	0.3675819
Complete	0.6231460
Ward	0.7437755

- *The Ward method* gives the highest value of agglomerative coefficients



Clustering on map





- The two maps can tell that almost every regions in cluster 2, which are labeled as developed regions, are from the *North of Italy*

Supervised Learning

Ridge Regression & Random Forest

Checking multicollinearity

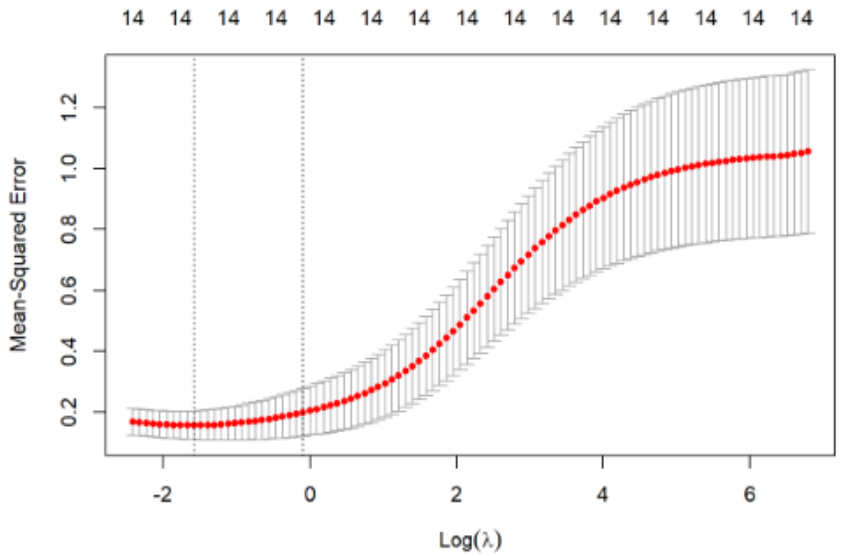
```
# Print the VIF values
print(vif_values)
```

```
##                                GDP_euro                         Employment_rate
##                              157.443994                              48.687427
##                   Mental_rate_per_100people                            Total_enroll
##                                2.847983                              77.859807
##          Non_Eu_resident_permit_1stGenuary                  Average_annual_salary_euro
##                              100.249850                              96.367767
##           Violent_crimes_per_10000_people                  Museum_density_per_100km2
##                                5.252899                               3.606869
##                    Park_density_per_100km2                  Life_expectancy_in_years
##                                8.491575                              11.986922
## Childhood_services_benefited_percentage      Rural_tourism_facilities_per100km2
##                                4.988962                               3.778604
##           Renewable_electricity_percentage      Highcared_hospital_per10000people
##                                9.252784                               4.138730
```

- Variables with higher VIF values include *GDP, employment rate, number of residence permits issued, total enrollment, average annual salary and life expectancy* while others are returned with lower VIF

Ridge Regression: Determining best lambda

```
# Plot cross-validated mean squared error (CV-MSE) vs. Log(lambda)
plot(cv_model)
```




```
# Find the optimal lambda value that minimizes test MSE
best_lambda <- cv_model$lambda.min
print(paste("Optimal lambda value:", best_lambda))
```

```
## [1] "Optimal lambda value: 0.206093061166557"
```

0.2061 is the result of the best lambda obtained

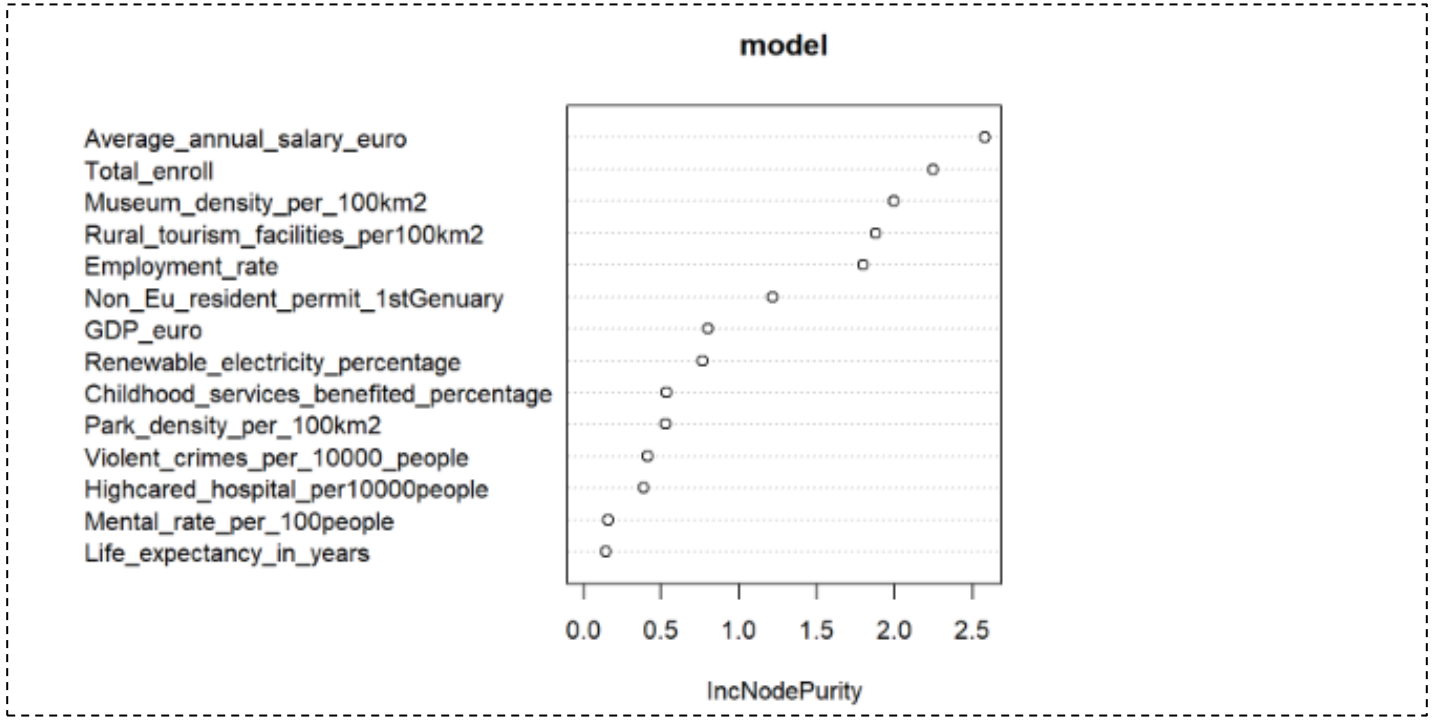
Ridge Regression: Predictors coefficient

```r
# Find coefficients of the best model
best_model <- glmnet(x, y, alpha = 0, lambda = best_lambda)
coef(best_model)
```

```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                                                s0
## (Intercept)                          -6.548140e-16
## GDP_euro                              8.549355e-02
## Employment_rate                       2.328712e-01
## Mental_rate_per_100people             4.013725e-02
## Total_enroll                          9.720672e-02
## Average_annual_salary_euro            2.868945e-01
## Violent_crimes_per_10000_people       2.264344e-01
## Museum_density_per_100km2            -6.033744e-02
## Park_density_per_100km2               1.635823e-01
## Non_Eu_resident_permit_1stGenuary     8.136465e-02
## Life_expectancy_in_years             -9.969069e-02
## Childhood_services_benefited_percentage  2.690649e-02
## Rural_tourism_facilities_per100km2    2.376228e-02
## Renewable_electricity_percentage     -1.922333e-01
## Highcared_hospital_per10000people    -1.462110e-02
```

"*Average_annual_salary_euro*" has the strongest impact on the increase in the target variable.

Random Forest: Predictors importance



- Random Forest method also tells that "*Average_annual_salary_euro*" is the most important predictor variable

Evaluate Supervised methods' performance

Ridge Regression

```r
# Calculate R-squared of the model on training data
y_predicted <- predict(ridge_model, s = best_lambda, newx = x)

# Calculate Sum of Squares Total (SST) and Sum of Squares Error (SSE)
sst <- sum((y - mean(y))^2)
sse <- sum((y_predicted - y)^2)

# Calculate R-squared
rsq <- 1 - sse / sst
print(paste("R-squared:", rsq))
```

```
[1] "R-squared: 0.958995497256194"
```

Random Forest

```
##
## Call:
##  randomForest(formula = Monthly_housing_euro ~ ., data = scaled_data)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##          Mean of squared residuals: 0.2456979
##                    % Var explained: 73.98
```

- The R-squared result obtained from Ridge Regression model is 0.95899549, which means that the model is able to explain almost 96% of the variation in the response values of the training data.
- Meanwhile, approximately 73.98% of the variance in monthly housing costs is explained by the predictors according to Random Forest model, which is slightly lower

Supervised models' observation predictions

It can be seen that the two results are quite similar.

```r
# Create a new observation based on the information provided
new_observation <- data.frame(GDP_euro = -0.35,
                              Employment_rate = 0.6,
                              Mental_rate_per_100people = -0.7,
                              Total_enroll = 2,
                              Non_Eu_resident_permit_1stGenuary = 2.9,
                              Average_annual_salary_euro = 1,
                              Violent_crimes_per_10000_people = -0.5,
                              Museum_density_per_100km2 = -0.3,
                              Park_density_per_100km2 = 0.4,
                              Life_expectancy_in_years = 0.7,
                              Childhood_services_benefited_percentage = -0.7,
                              Rural_tourism_facilities_per100km2 = -0.6,
                              Renewable_electricity_percentage = 1.2,
                              Highcared_hospital_per10000people = -0.8)
```

Estimated value 1 (Ridge Regression)	Estimated value 2 (Random Forest)
335.5724	305.5346

Conclusions

- Unsupervised learning models namely *K-medoids method and the Ward's method* have classified 18 regions in Italy into two different clusters with specific characteristics:
  - Cluster 1 represents the regions located in the South of the country, which are less developed and have lower housing costs compared to those in cluster 2, which are the regions from the North of the country.

- Both Supervised Learning models used throughout this project, which are Ridge Regression and Random Forest, have concluded that the annual salary per capita is the key feature in deciding the overall housing costs in Italy
  - Both models seem to return good estimation for the target variable based on the percentage covered in the data variance

Thank You Very Much