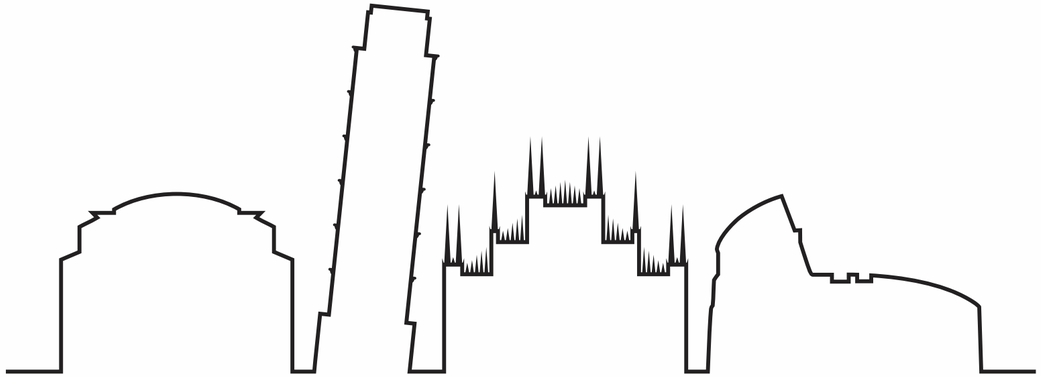


The quality of life in Italian Municipalities

STATISTICAL LEARNING

A project by
Andrea Pio Cutrera

Municipalities in Italy

- Italy is a very heterogeneous country
- Fundamental political units have some similarities and much differences
- Some areas share economic and social characteristics
- other differs in natural and cultural resources

Quality of life

Is it possible to model this measure for the municipalities of Italy as a function of some features?

Data Understanding

Data sources:

- CIPU office of the “Department for the Programming and Coordination of Political Economy” for the Urban Index set of variables of 2015 (2011 data)
- Il Sole 24 Ore - for the quality of life score of municipalities in Italy in 2011

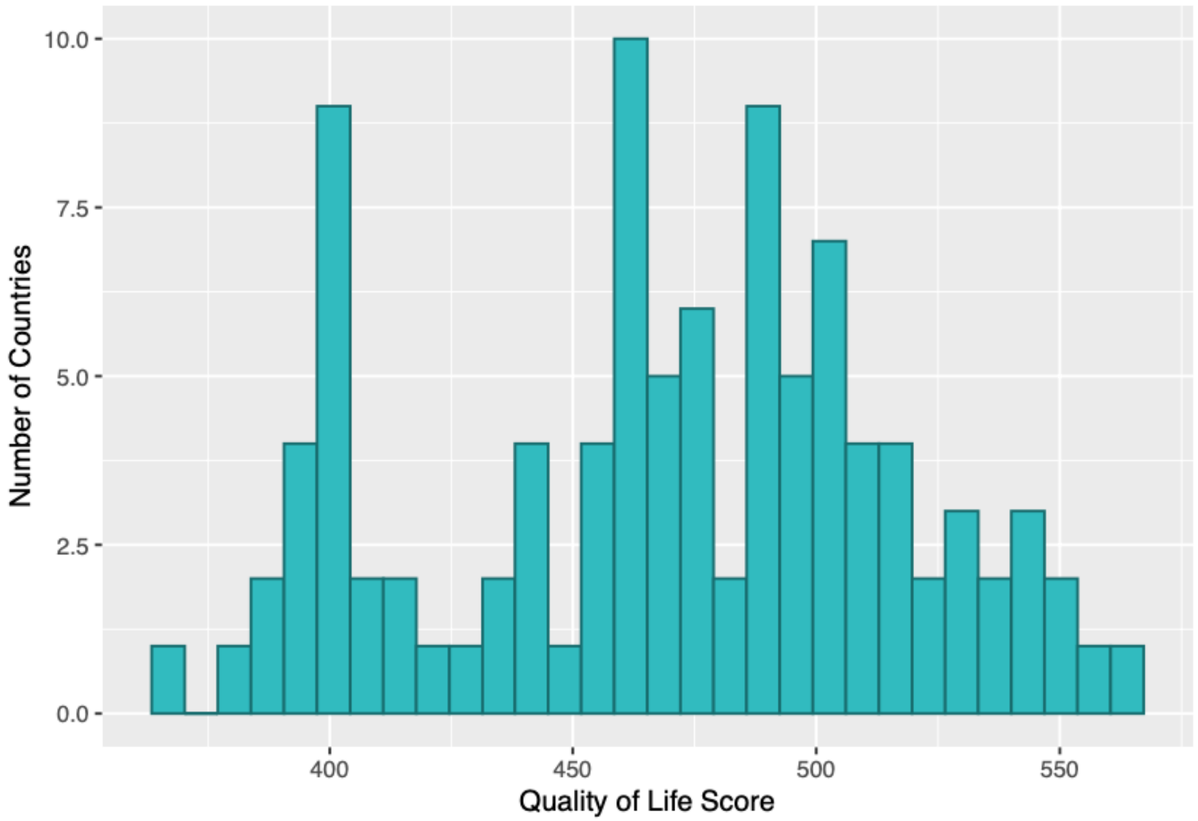
35 variables

INDEPENDENT VARIABLES

- decennal change in resident population (1991- 2001 and 2001-2011
- human densisty
- private mobility
- oldness
- geen areas per capita
- incidence of young couples with children
- presence of university
- commuting for work or study
- economic dynamism
- compound index of tourism reception capability
- annual number of visitors in cultural places
- entertainment
- digital divide
- accessibility to train stations
- hospital seats for 10.000 inhabitants
- libraries for 10,000 inhabitants
- young people with risk of quitting secondary school
- unemployment
- economical hardship of families
- old people alone
- suicides
- crowded houses
- services at home
- share of foreign population
- Gini index of inequality
- ratio between occupied italian foreign citizens
- ratio between the share of occupied males and females
- agricultural land
- waste produced per capita
- slow mobility
- drinking water influed in the municipal net
- waste differentiation
- density of photovoltaic systems
- car classified as E5 and E6
- centres of excellence

Response Variable

Quality of life



- Source: Il Sole 24 Ore
- Year: 2011
- Observations: 100 provinces
- Computed by taking into account many indexes measuring economy, health, crime, opportunities for free time, income and many more dimensions

Supervised Learning

STEPWISE MODEL SELECTION

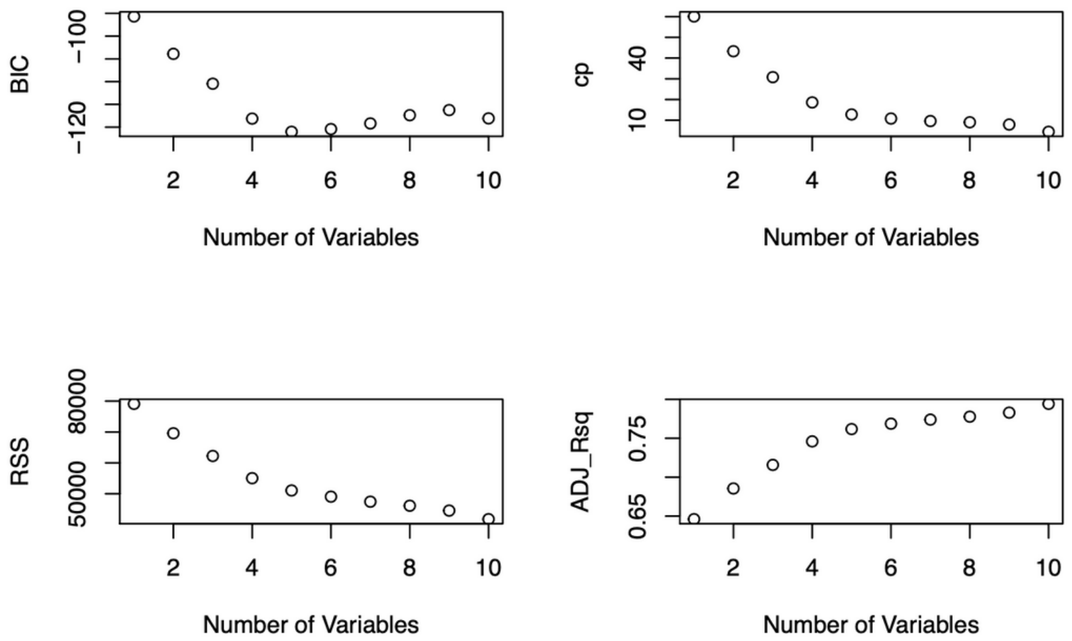
Find the most important features to add to the null model

LINEAR REGRESSION

Create a model with the variables selected as a function which maps features to the response variable

K-NEAREST NEIGHBOURS

Create a model able to classify a municipality to the category of quality of life




```
lm(formula = life_quality ~ disoccupazione + auto_e5_e6 + occupazione_m_f +
    popolazione_straniera + visitatori_luoghi_cultura, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-49.283 -13.903  -1.577  12.546  62.349

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                5.811e+02  3.058e+01  19.004  < 2e-16 ***
disoccupazione            -2.955e+00  8.692e-01  -3.399 0.000993 ***
auto_e5_e6                 1.239e+00  3.529e-01   3.512 0.000686 ***
occupazione_m_f           -7.955e+01  2.311e+01  -3.443 0.000862 ***
popolazione_straniera      2.757e+00  8.022e-01   3.436 0.000880 ***
visitatori_luoghi_cultura  6.249e-06  2.311e-06   2.704 0.008140 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.31 on 94 degrees of freedom
Multiple R-squared:  0.7738, Adjusted R-squared:  0.7618
F-statistic: 64.32 on 5 and 94 DF,  p-value: < 2.2e-16
```

- BIC information criterion selects the model with 5 variables
- All the variables are strongly statistically significative

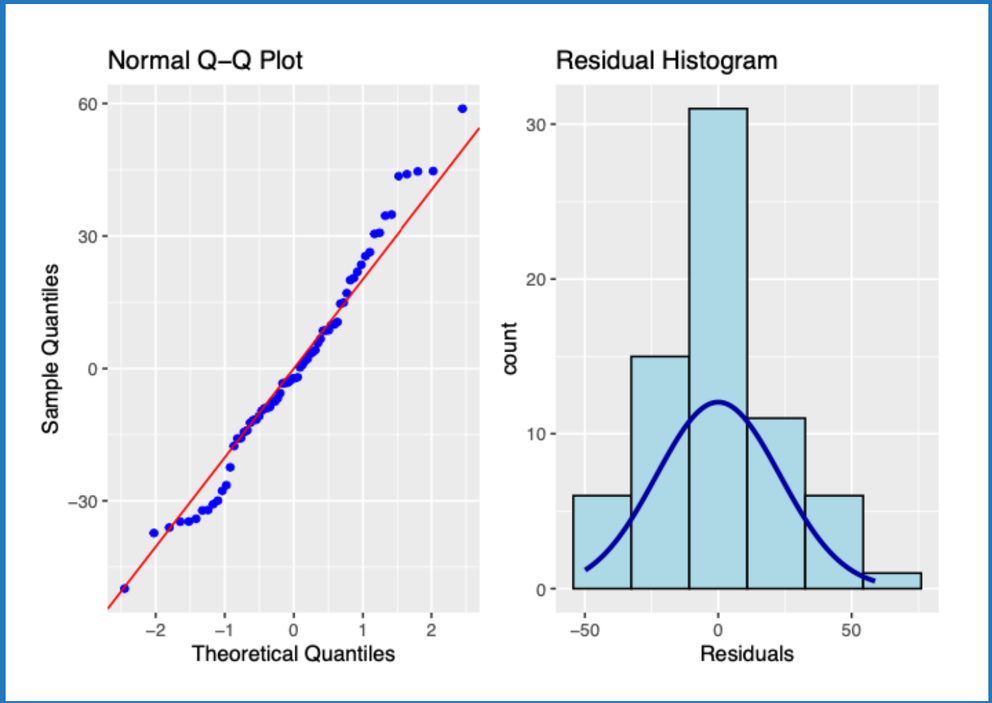
OLS model with 5 variables selected

Residuals are normally distributed

Square root of mean squared error

Definitely decreased with the forward search with respect to the full model

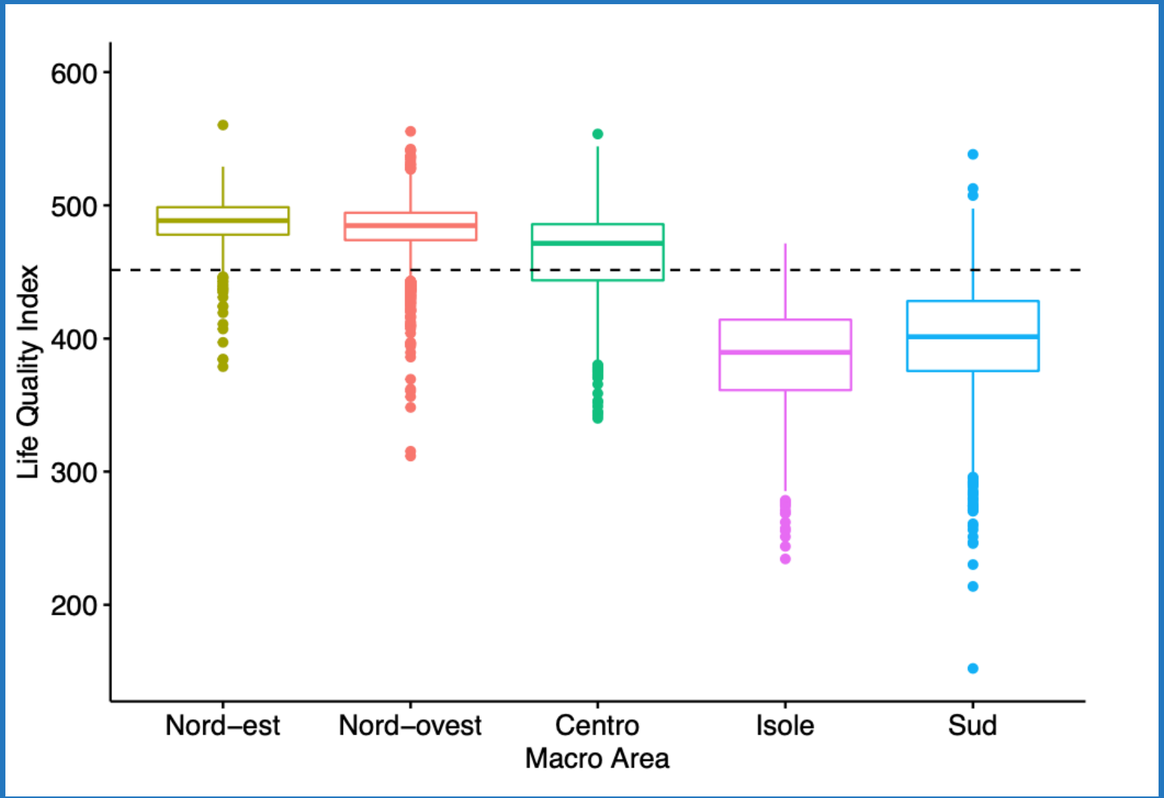
from 30 to 22

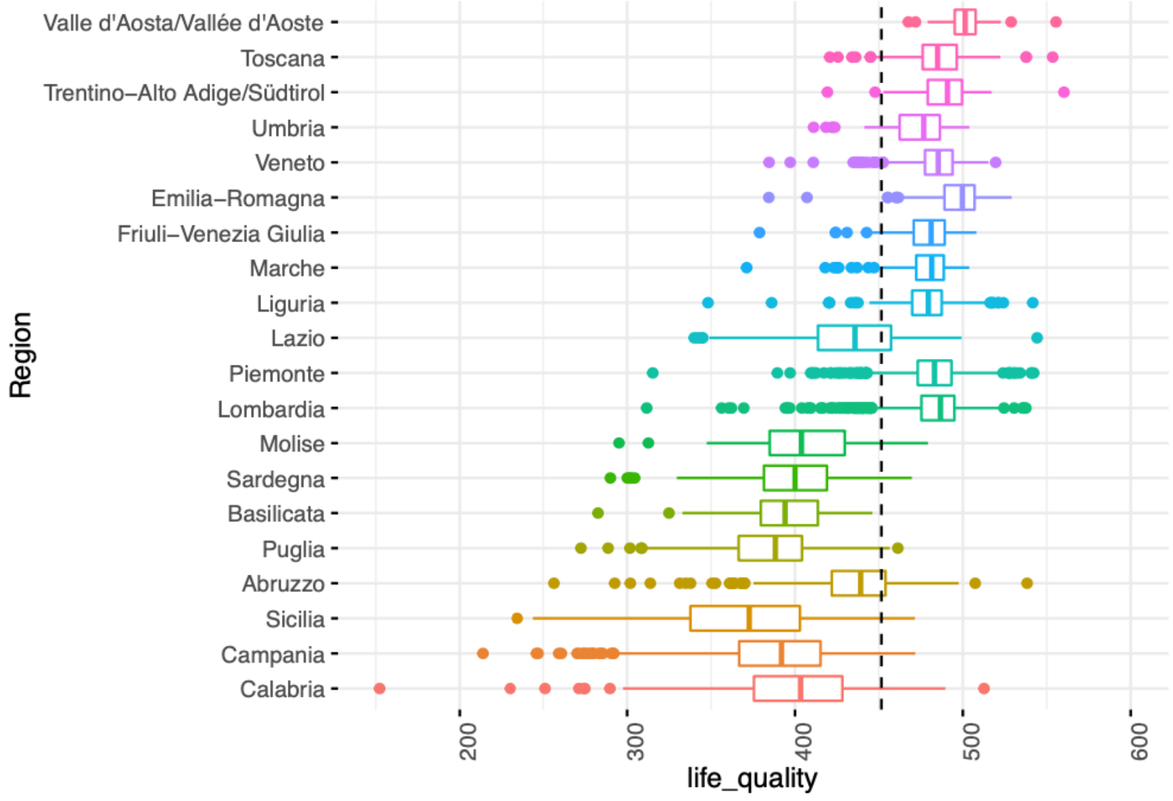


Inference on the remaining municipalities

Model created on the provincial capital municipalities is used to make prediction for the response variable on the remaining observations

GEOGRAPHICAL AREAS





REGIONAL AREAS

Quality of life

Category

low: from 150 to 400

1361 municipalities

medium: from 401 to 480

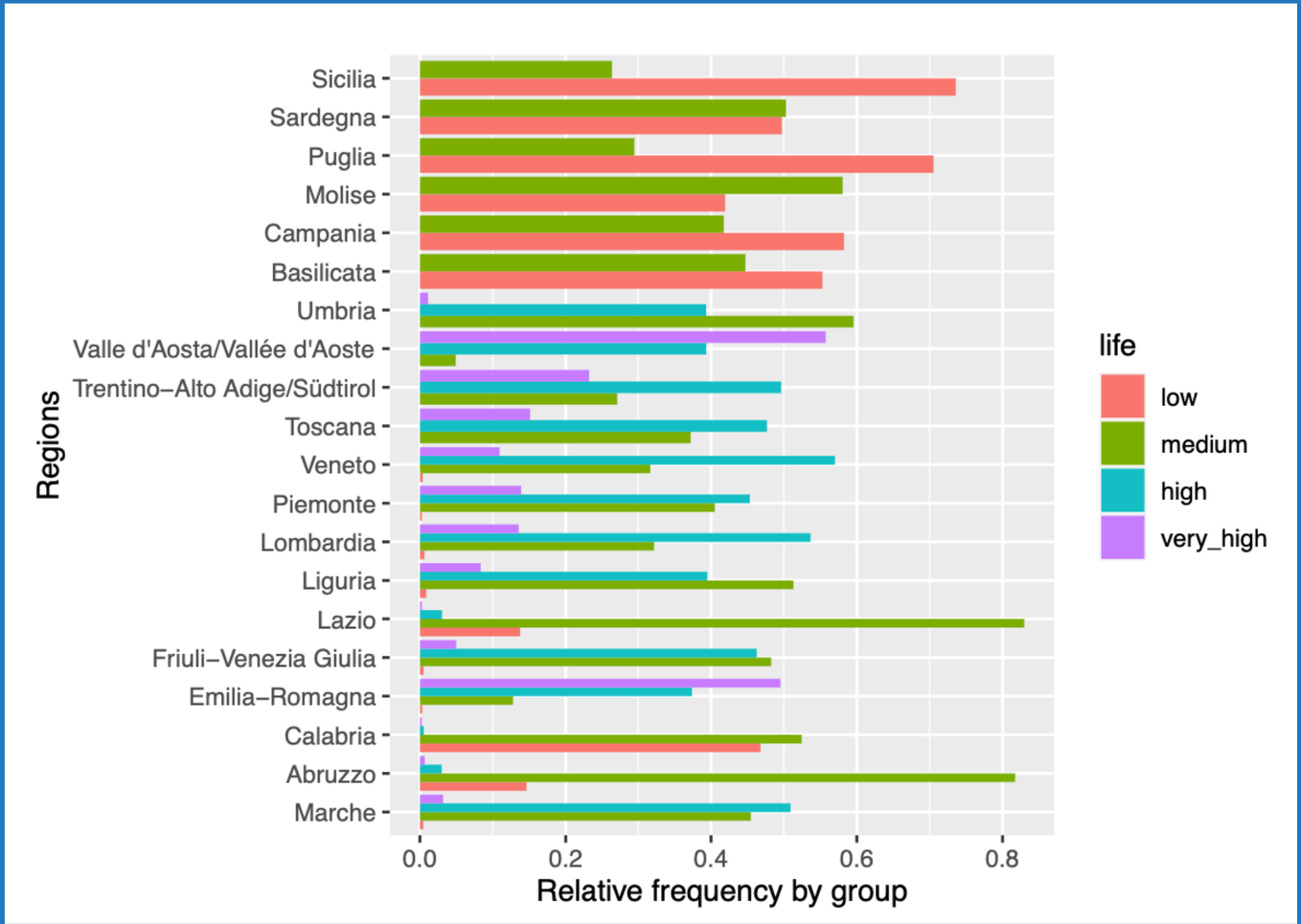
3088 municipalities

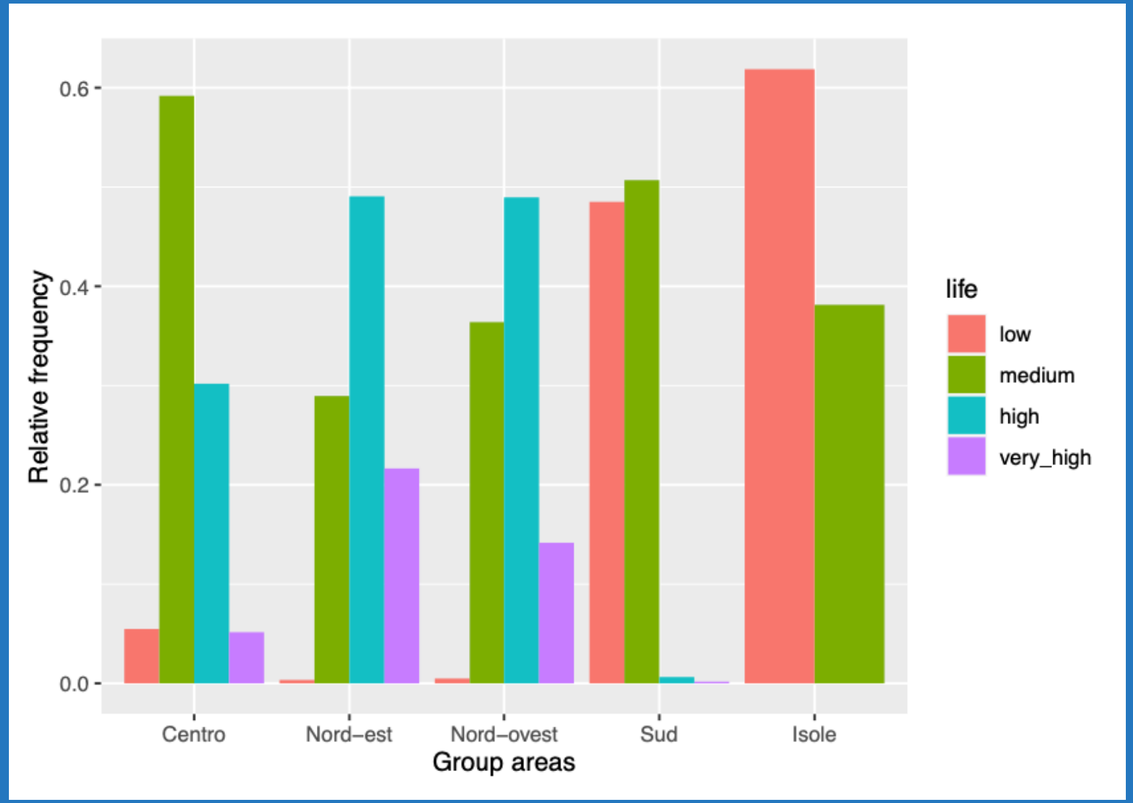
high: from 481 to 500

2266 municipalities

very high: from 501 to 570

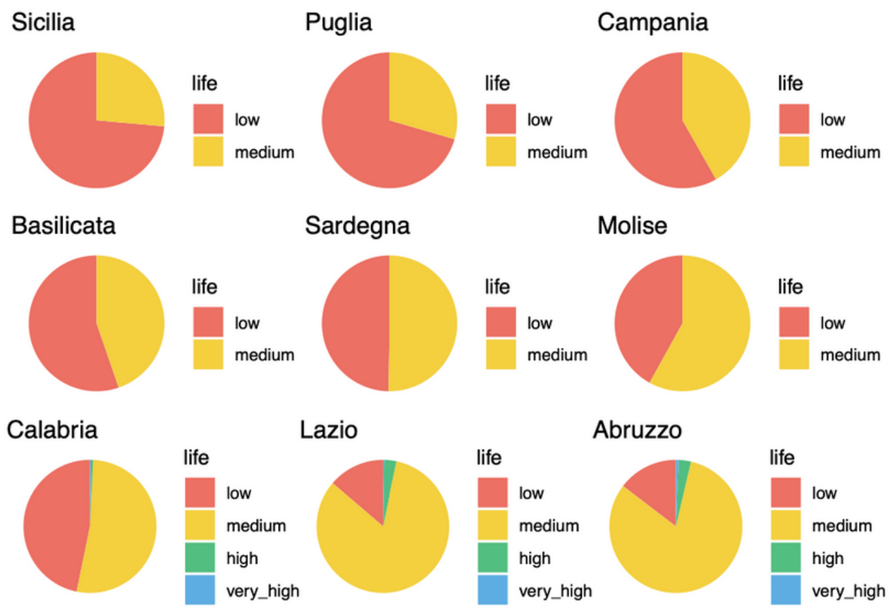
709 municipalities

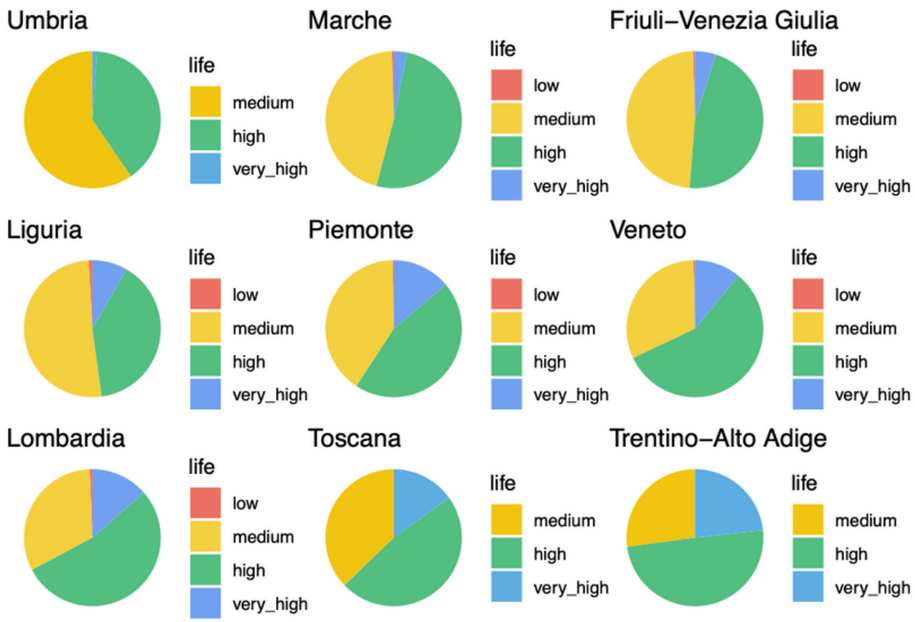


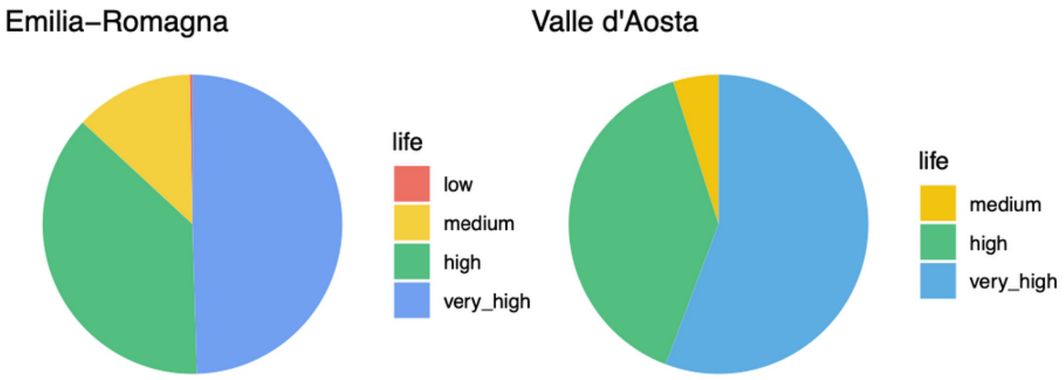


RELATIVE FREQUENCIES OF CATEGORIES AMONG GEOGRAPHICAL AND REGIONAL AREAS

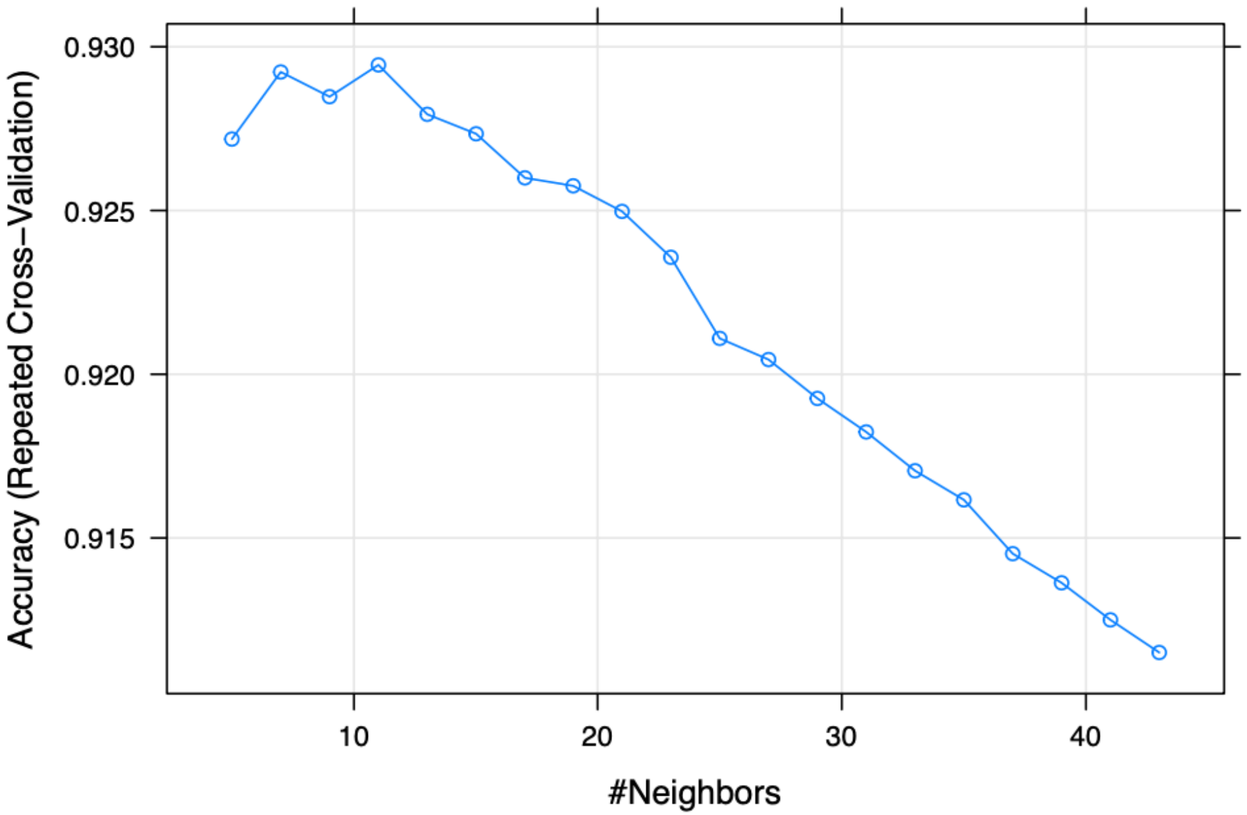
Pie charts for relative frequencies







K-Nearest Neighbor



CLASSIFICATION ALGORITHM

- REPEATED CROSS VALIDATION ON THE ENTIRE DATASET REVEALS THAT 11 IS THE BEST HYPERPARAMETER WHICH MAXIMIZES THE ACCURACY

	test_labels			
knn.model	low	medium	high	very_high
low	386	12	0	0
medium	15	891	14	0
high	0	52	644	41
very_high	0	0	7	166

11-NN

CLASSIFICATION ALGORITHM

- ALMOST **94% OF ACCURACY** ON THE TEST SET

Unsupervised Learning

PRINCIPAL COMPONENT ANALYSIS

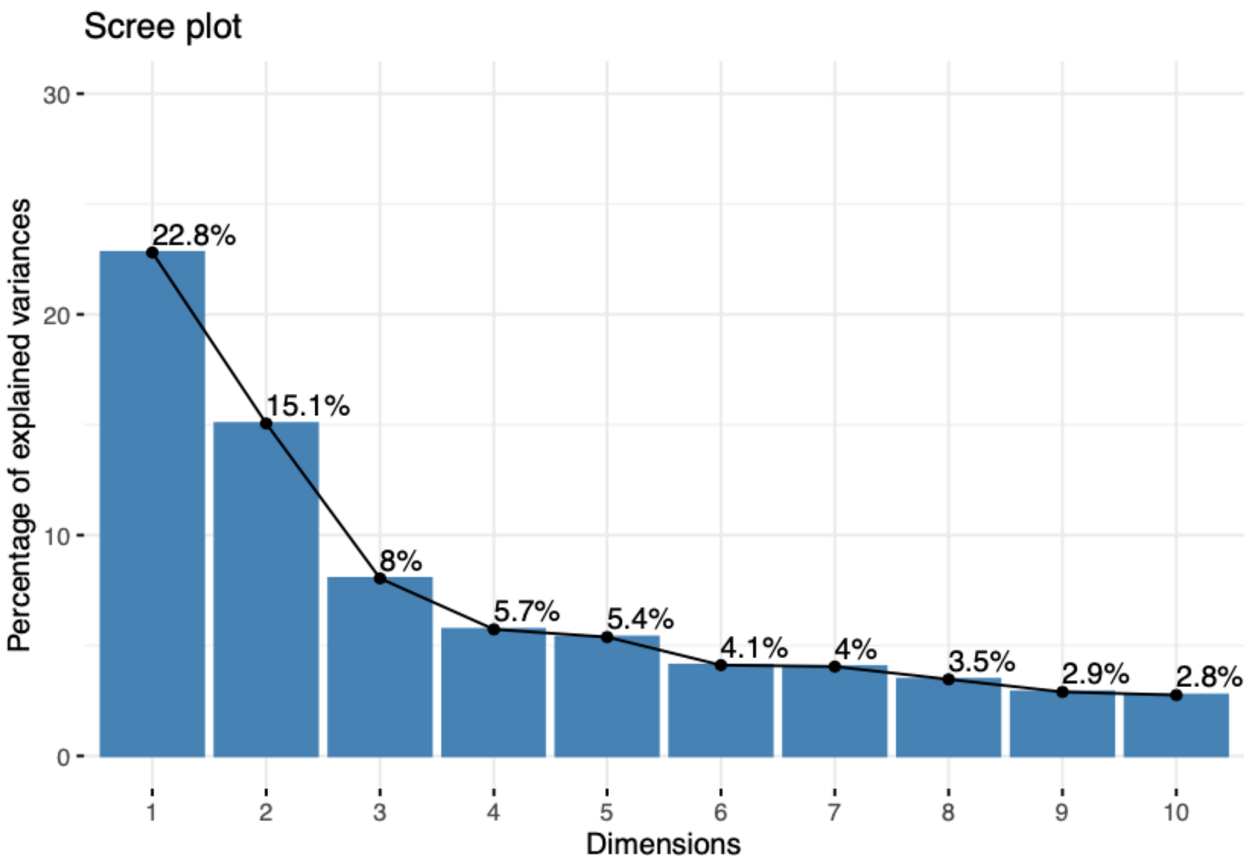
To reduce dimensionality of our data matrix

K-MEANS CLUSTERING

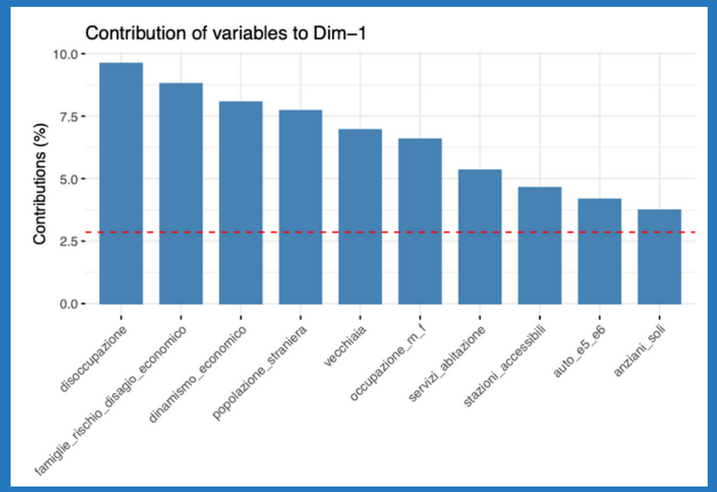
Divide our observations into groups through a similarity principle

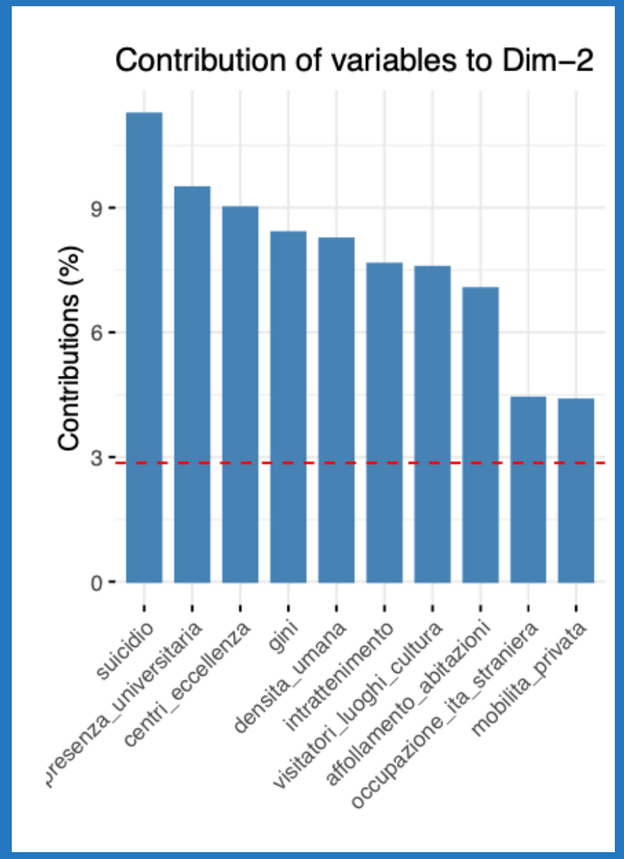
PCA

Reducing the dimensionality of our data is not an easy task. At least 5 dimensions should be retained to have a cumulative variance explained of 57%



ANALYSIS OF PRINCIPAL DIMENSIONS





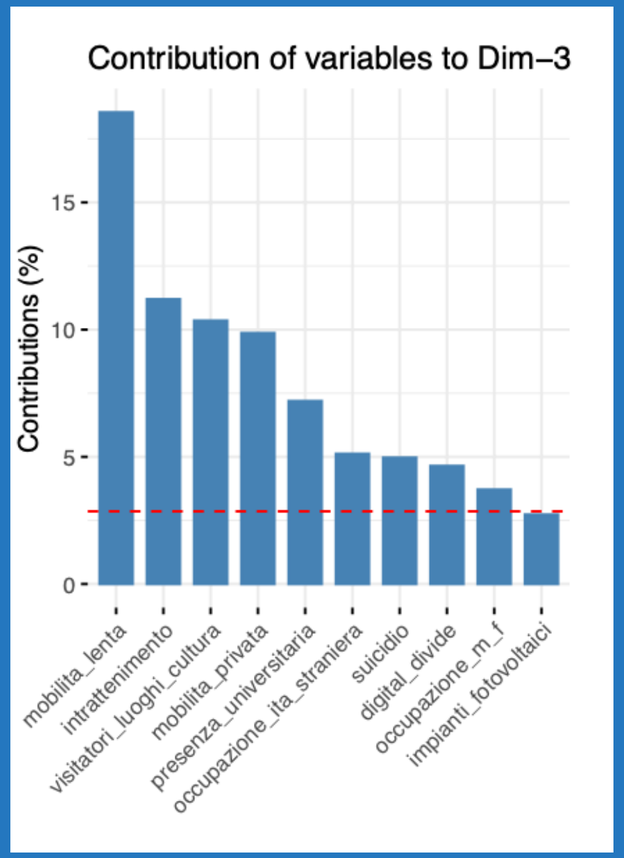
ECONOMIC DIMENSION

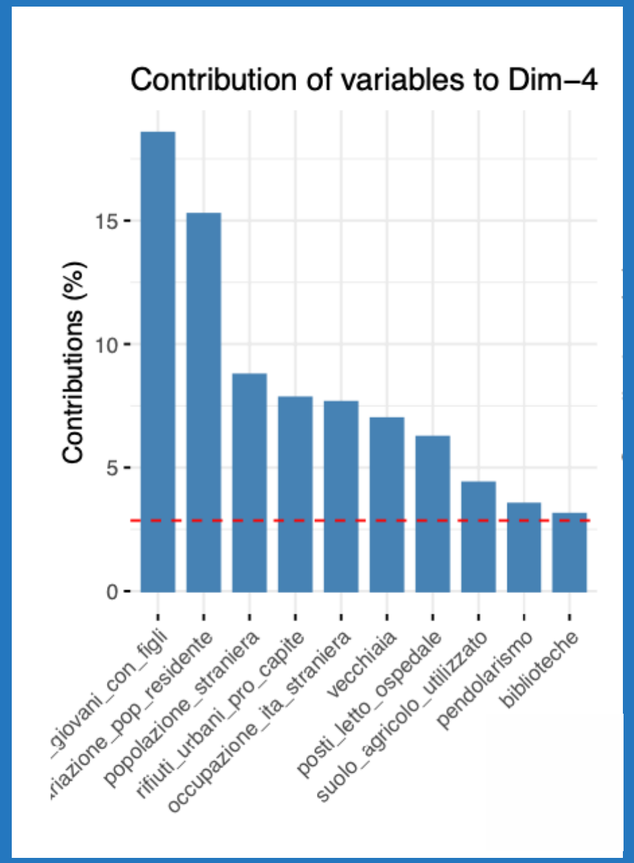
- unemployment
- families in risk of hardship
- economic dynamism

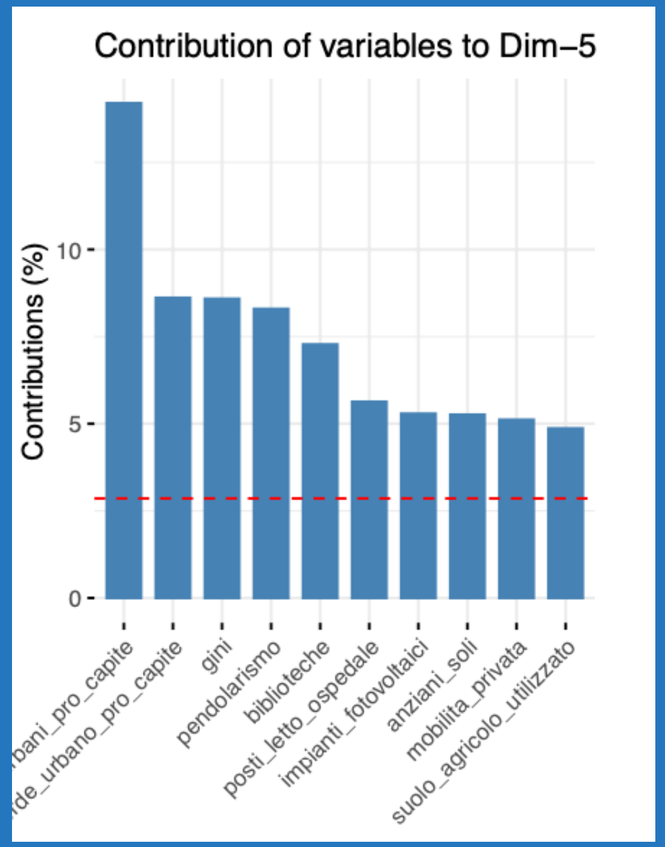
SOCIAL-CULTURAL DIMENSION

- suicides
- presence of universities
- centres of excellence
- gini index of inequality

ANALYSIS OF PRINCIPAL DIMENSIONS







TRASPORTATION DIMENSION

- slow mobility

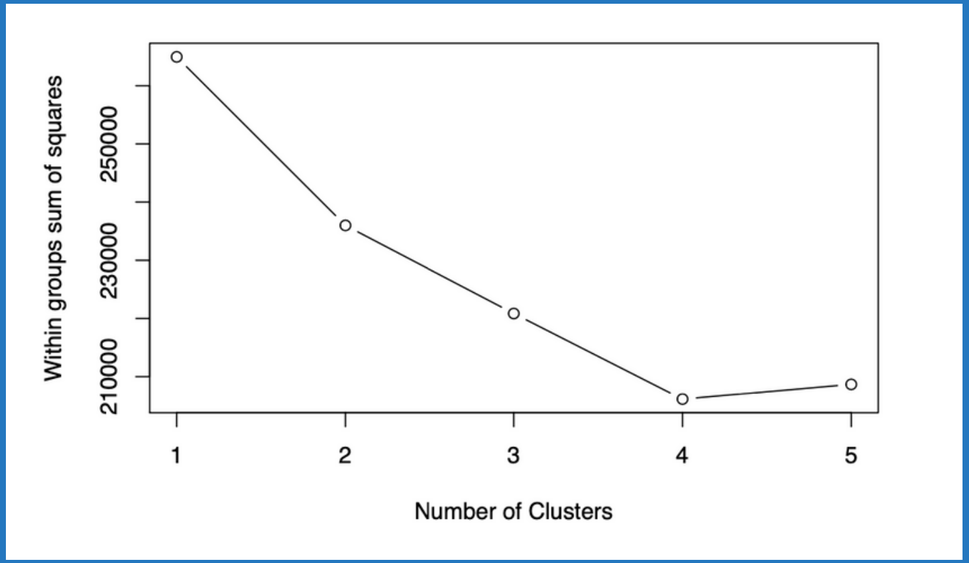
DEMOGRAPHIC DIMENSION

- number of young couples with children
- change in resident population
- foreign population

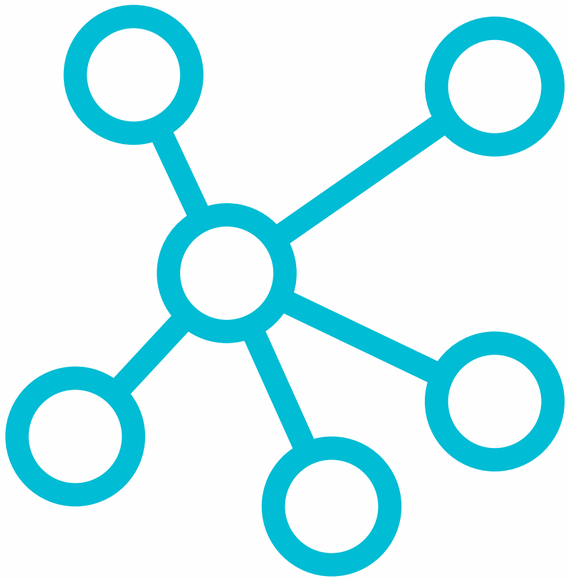
GREEN DIMENSION

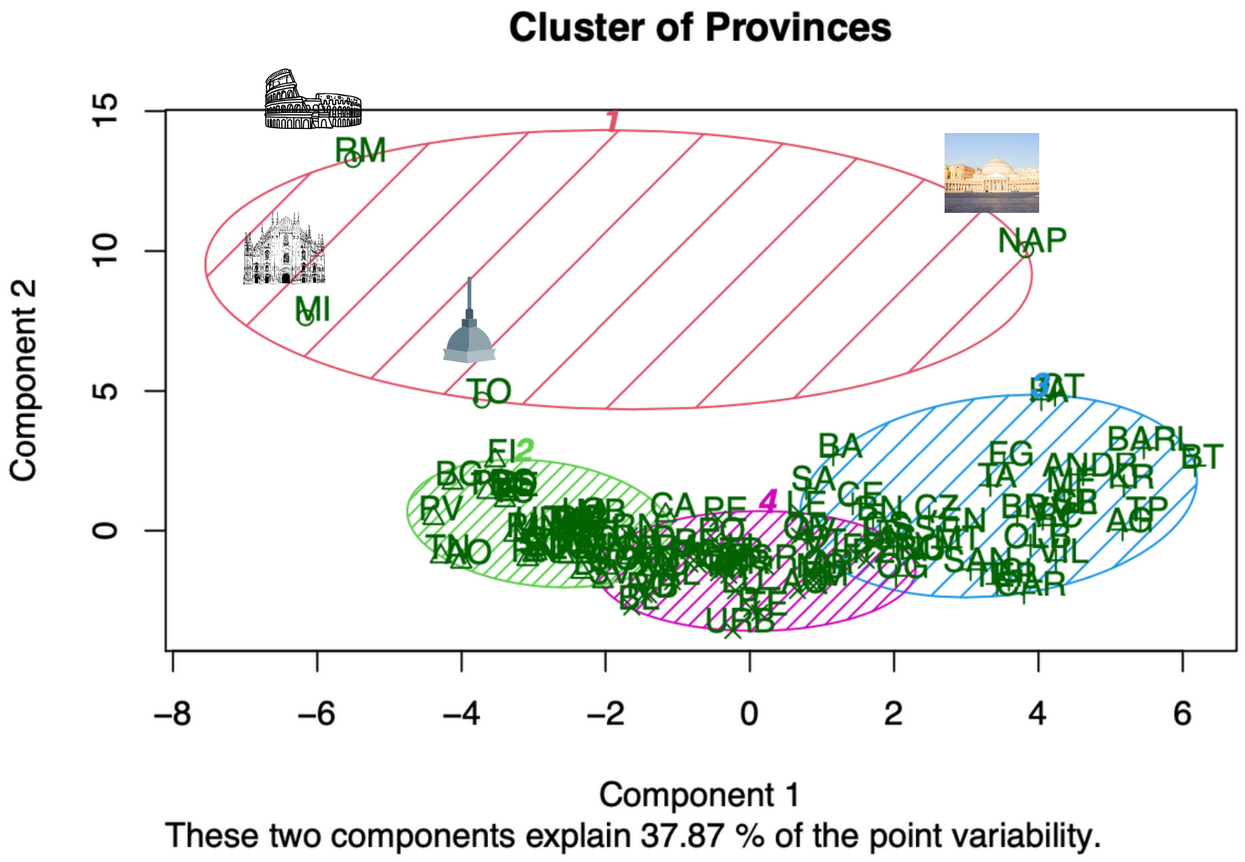
- urban waste per capita
- green areas per capita

K-Means Clustering

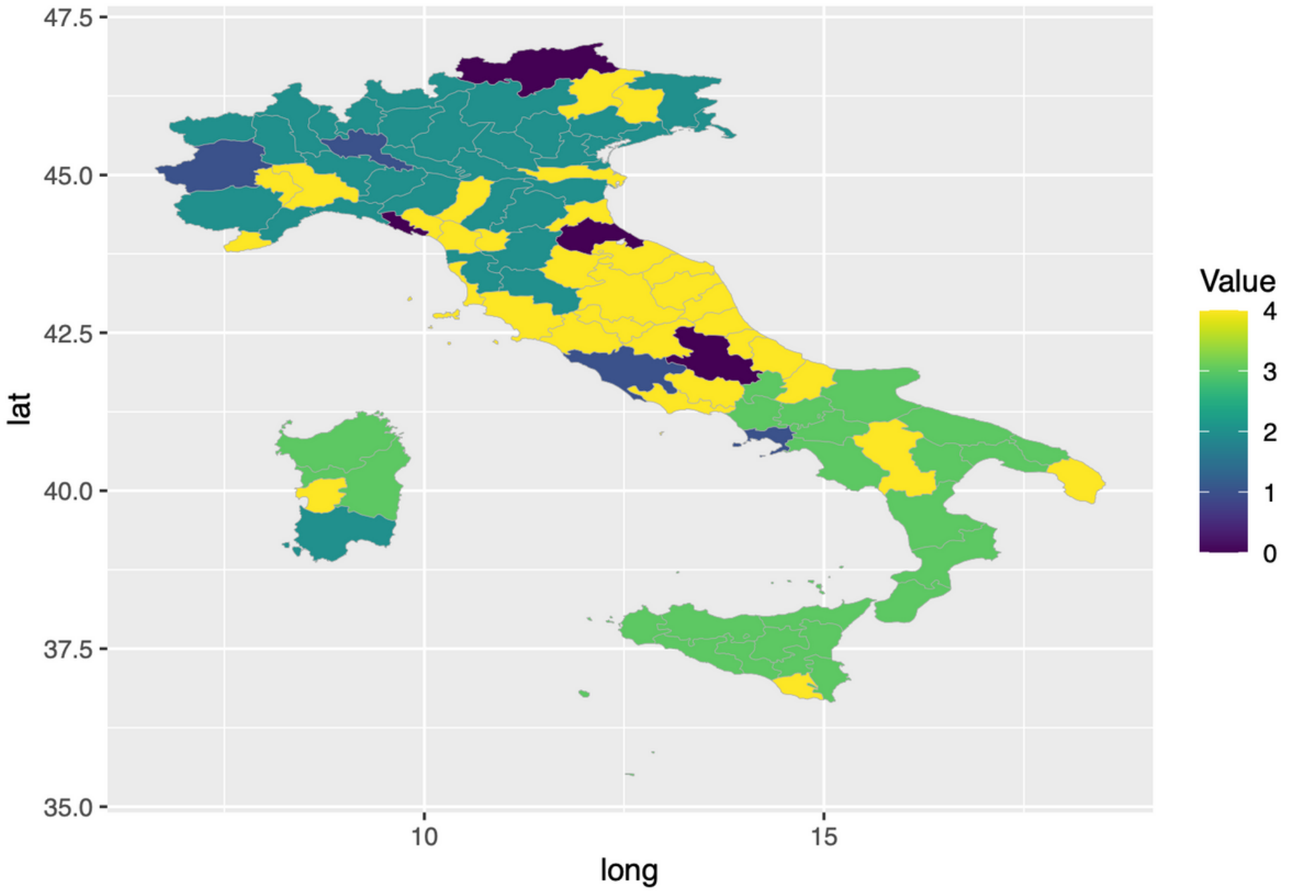


4-Means clustering is the algorithm which minimize the within -cluster variation





These two components explain 37.87 % of the point variability.



Conclusions

AND MAIN FINDINGS

QUALITY OF LIFE IN MUNICIPALITIES IS AFFECTED:

- **Negatively** by unemployment
- **Positvely** by the number of cars classifieds E5-E6
- **Negatively** by the gender gap in occupation
- **Positively** by foreign population
- **Positively** by visitors of cultural places

HAVING ALL THESE INFORMATION FOR A MUNICIPALITY ALLOWS US TO **PREDICT** THE **SCORE IN QUALITY OF LIFE**

by means of the **OLS model** created

ALSO THE CATEGORY CAN BE PREDICTED

by means of **11-NN model** created

Conclusions

AND MAIN FINDINGS

PCA REVEALS THAT OUR DATA ARE SPREAD ALONG THE FOLLOWING MAIN DIMENSIONS:

- **Economic** dimension
- **Social-Cultural** dimension
- **Transportation** dimension
- **Demographic** dimension
- **Green** dimension

K-MEANS REVEALS INSTEAD THE PRESENCE OF 4 CLUSTERS OF PROVINCIAL CAPITALS MUNICIPALITIES:

- **Milan**, **Rome**, **Naples** and **Turin**: the greatest cities
- Mostly cities of the **North**
- Mostly cities of the **Centre**
- Mostly cities of the **South**

Thank you for your time

- Andrea Pio Cutrera
- 965591
- andrea.cutrera@studenti.unimi.it