The quality of life in Italian Municipalities

Statistical Learning Project

Andrea Pio Cutrera - 965591

7/3/2021

Abstract

Italy is a very heterogeneous country in which the fundamental political units (i.e., municipalities) have some similarities and much differences. This is what makes Italy a unique country in the world, full of different richness, but also many weakness; some areas share economical and social characteristics, other differs in natural and cultural resources. In this paper, Quality of life, which is an indicator measured in one dimension, is modelled to find the best explanatory variables which are able to describe it for the municipalities that are provincial capital. A classification algorithm is implemented for the categorization between low, medium, high and very high quality of life. Then in the last part it is made an effort to understand the main dimensions along which data are spread the most, trying also to find clusters between municipalities and provinces.

1. Introduction

The research question of this paper started from the curiosity of whether there exists a model which describes the **quality of life of italian Municipalities**. Every year the italian journal "Il Sole 24 Ore" makes an aggregate ranking on the quality of life for all the provinces of Italy based on a bunch of indicators. But, which are the main features that explain a certain level of quality of life?

I started collecting data from many sources like the Urban Index from *"Department for Programming and Coordination of Political Economy"* (in particular the office called *CIPU* - Comitato Interministeriale per le Politiche Urbane) of the Presidency of the Council of Minister and the *"Il Sole 24 Ore"* for the response variable. The first source had more than 100 variables and more than 8,000 observations (Municipalities), the second one, from which I took the variable used as dependent one, has the ranking and the associated score of quality of life for slightly more than 100 observations (Provinces). I made a subset for the municipalities' dataset choosing 35 variables among the 100+ which I considered as the most relevant for the aim of describing the quality of life; then I merged the two sources on the municipalities which are capitals of each province and I got a dataframe with at least **35 variables** (features) and **100 municipalities** (provincial capital) as observations.

For the *Supervised Learning* Part it has been used a feature selection model (**forward search**) in order to find **standard linear regression model** (i.e., *OLS* - Ordiranry Least Square) with not too much regressors, choosing the number of regressors with the Bayes Information Criterion (i.e., *BIC*), the one which penalizes models with many independent variables. With the model generated and evaluated on the provincial capital municipalities I decided to make the **predictions on the remaining municipalities** (the observations for which the response variable is not available) inferring the response of quality of life. Then a classification model with k-Nearest Neighbour is implemented for the categorization of the predicted response (4 levels: low, medium, high, very high quality of life).

For the *Unsupervised Learning* part instead it is applied a **Principal Component Analysis** technique to project the data onto a lower dimansional space, which can be reduced to 5 (dimensions) but loosing a bit of information. This is maybe because our 35 variables measure pretty different things, evaluating each

municipality from many perspectives. Nontheless the five principal dimensions are described and interpreted. Then it is applied the **k-means** algorithm to find relevant clusters between observations with some notion of distance.

An expected result is to find the usual picture of a fragmented country, especially by latitude: a rich and prosper *north* against a poor and not developed *south*. Another expectation is the possibility to find a model which describes the quality of life score as positively affected by economic performance, availability of services and cultural richness.

2. Data Understanding

The two data sources, as above mentioned are the *CIPU* office[1] of the "Department for the Programming and Coordination of Political Economy" for the **Urban Index** and the journal *Il Sole 24 Ore*[2] for the **quality of life score**.

The Urban index is a project in collaboration with the the Department of Architecture and Urban Studies of University "Politecnico di Milano" which in 2015 collected data of 2011 at municipality level on a set of heterogeneous indicators.

All these indicators are the ones which have been choosen for the developing of useful tools for definition and evaluation of public policies in urban areas and also to define the contents of the National Urban Agenda[3].

2.1. Independent variables

The set of features of our observations taken from the *Urban Index* is very big and comprehend even detailed measures of some characteristics of the municipalites (*7000+ observations* without missing values), and for this reason it has been choosen to make a subset with the variables which could be more apt in describing the wellness in town. The selected variables are 35, and they are named in italian. For the sake of understanding, you can refer to the following dictionary in which they are named (in english) and briefly described:

- *variazione_pop_residente*: decennal change in resident population (difference between the mean 1991-2001 and the mean 2001-2011);
- *densita_umana*: human densisty (ratio between resident population + 1/3 of hotel seats over the total surface of land in square kilometers);
- *mobilita_privata*: private mobility (share of people which daily uses a owned engined-mean of transportation);
- *vecchiaia*: oldness measured as the ratio between people 65+ and 0-14;
- *verde_urbano_pro_capite*: geen areas per capita (not agricultural);
- *coppie_giovani_con_figli*: incidence of young couples with children (less than 35 years old);
- *presenza_universitaria*: presence of university is an index made by the sum of universities (counted as 1) and other branches (counted as 0.5);
- *pendolarismo*: commuting for work or study as the ratio between the sum of daily ingoing and outgoing people and total population;

[1] CIPU office website: http://presidenza.governo.it/AmministrazioneTrasparente/Organizzazione/ArticolazioneUffici/Dipartimenti/DISET.html

[2] Il Sole 24 Ore website:https://st.ilsole24ore.com/includes2007/speciali/qualita-della-vita/scheda_finale.shtml?refresh_ce=1

[3] Urban Index: Governo Italiano Presidenza del Consiglio dei Ministri, Segreteria tecnica del Comitato Interministeriale per le Politiche Urbane (CIPU) https://www.urbanindex.it

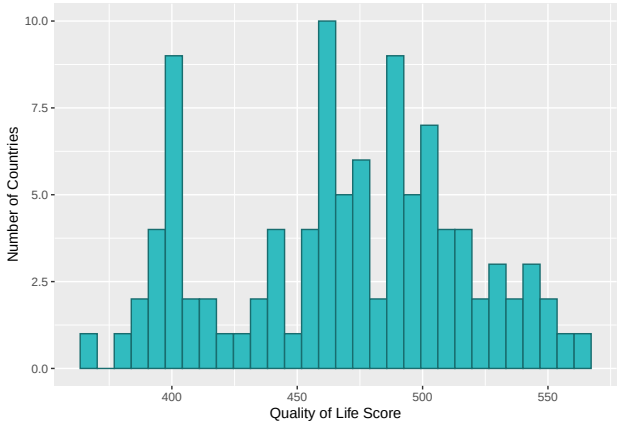
- *dinamismo_economico*: economic dynamism is the arithmetic mean of standardized values of 4 indexes:
  - *Agriculture*: employed people / total population * 100;
  - *Manufacturing*: employed people / total population * 100;
  - *Trade*: employed people / total population * 100;
  - *Services*: employed people / total population * 100;
- *tasso_funzione_ricettiva*: compound index of tourism reception capability; it is the ratio between the hotel seats * 10,000 and the president population * square kilometers of land;
- *visitatori_luoghi_cultura*: annual number of visitors in cultural places of the State;
- *intrattenimento*: entertainment measured with the absolute number of places like amusement parks, aquariums, multiplex;
- *digital_divide*: share of population excluded by the wide-band internet connection;
- *stazioni_accessibili*: composite index which measures how much accessible are the train stations;
- *posti_letto_ospedale*: number of hospital seats for 10.000 inhabitants;
- *biblioteche*: number of libraries for 10.000 inhabitants;
- *abbandono_scolastico_2grado*: young people with risk of quitting secondary school (as a share of total young students);
- *disoccupazione*: ratio between people over 15 looking for a job and total active people in the same age range;
- *famiglie_rischio_disagio_economico*: share of families with children with no active members in economical activity;
- *anziani_soli*: share of families with one member in the age range 65+;
- *suicidio*: mean of suicides in 2010-2011-2012;
- *affollamento_abitazioni*: share of houses with less than 40 squared metres and 4 inhabitants, houses between 40-59 squared metres and 5 inhabitants, houses between 60-79 squared metres and 6 inhabitants out of the total occupied houses;
- *servizi_abitazione*: arithmetic mean of the ratios between the 5 essential services in a house (drinking water, internal toilette, bathtub or shower, heating system and warm water);
- *popolazione_straniera*: share of foreign population;
- *gini*: is an index which computes the heterogeneity in between groups of the *IRPEF* tax (income tax), since only the mean for each subgroup is available for that data (proxy of inequality);
- *occupazione_ita_straniera*: ratio between the share of occupied italian citizens and the share of occupied foreign citizens in each age group of 15+ years old;
- *occupazione_m_f*: ratio between the share of occupied males and the share of females in each age group of 15+ years old;
- *suolo_agricolo_utilizzato*: percentage of land used for agriculture;
- *rifiuti_urbani_pro_capite*: waste produced per capita;
- *mobilita_lenta*: index measuring how much people moves on foot or by bike;
- *acqua_potabile*: drinking water influed in the municipal net in cubic metres per year per capita;

- *raccolta_differenziata*: percentage of waste differentiation (for recycling);
- *impianti_fotovoltaici*: density of photovoltaic systems;
- *auto_e5_e6*: percentage of car classified as E5 and E6;
- *centri_eccellenza*: number of technological districts, centres of excellence and scientific parks;

2.2. Response variable

The response taken from the *Il Sole 24 Ore* is a numeric variable which is available for 100 provinces and it is computed by taking into account many indexes measuring economy, health, crime, opportunities for free time, income and many more dimensions for 2011[4]. It has a distribution that can be seen in the following summary statistics and better with a histogram:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   369.0   439.0   475.5   469.6   505.2   566.0
```



With the histogram it seems to be normally distributed, but a check in the Shapiro-test (for normality) does not confirm it; it needs some *transformation*, and the *quadratic one* is the only able to make it *"more" normal*, allowing us to reject no more the hypotheis of normality but at 90% of confidence level.

```
##
##  Shapiro-Wilk normality test
##
## data:  (dataset$life_quality)^2
## W = 0.97205, p-value = 0.03179
```

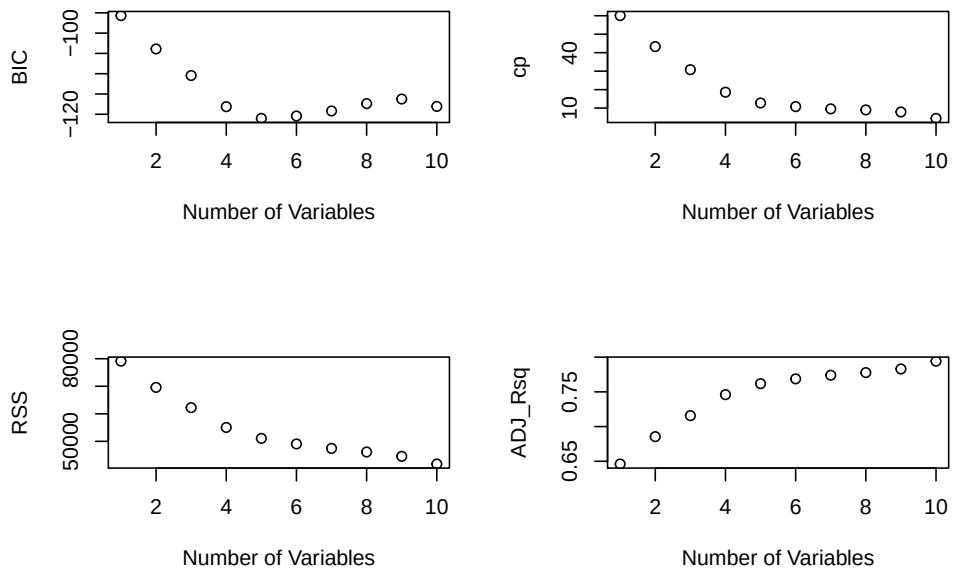
3. Supervised Learning Methods

3.1. Linear Regression and Stepwise Model Selection

For the Supervised Learning part, as described above, it has been used the **forward stepwise selection** technique to find the most important features to be added to the null model and according to some method of evaluation (i.e., *BIC - Bayes Information Briterion*) it is choosen the model with 5 regressors. BIC criterion

[4]Il Sole 24 Ore website:https://st.ilsole24ore.com/includes2007/speciali/qualita-della-vita/scheda_finale.shtml?refresh_ce=1

penalizes models with many features because of the *parsimonious modelling* reason, as the problem-solving principle called **Occam's Razor** suggests: *"entities should not be multiplied without necessity"*, sometimes inaccurately paraphrased as "the simplest explanation is usually the best one"[5].



Forward search selects the following regressors:

- *disoccupazione* (**unemployment**);
- *auto_e5_e6* (**car classified as E5-E6**);
- *occupazione_m_f* (**employment male/female**);
- *popolazione_straniera* (**foreign population**);
- *visitatori_luoghi_cultura* (**visitors of cultural places**);

and the *OLS model* has **all the coefficients statistically significative**, and an adjusted R^2 of 0.7618:

```
##
## Call:
## 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 ***
```

[5]Wikipedia: Occam's Razor: https://en.wikipedia.org/wiki/Occam%27s_razor

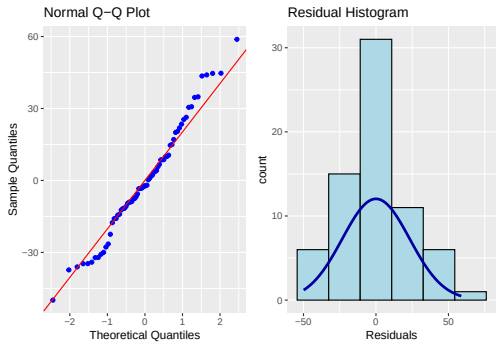
```
## occupazione_m_f           -7.955e+01  2.311e+01  -3.443 0.000862 ***
## popolazione_straniera      2.757e+00  8.022e-01   3.436 0.000880 ***
## visitatori_luoghi_cultura  6.249e-06  2.311e-06   2.704 0.008140 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual 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
```

At least training both models (the full model and the one with the 5 variables selected) on the training set and testing it with the test set, the **square root of mean squared error** is definetely **decresed** with the *forward search*.

```
##   root_mse_full_model root_mse_fwd_model
## 1            30.47695           22.06351
```

The residual diagnostics are at least confirming that errors of the model with 5 regressors are normally distributed, as it can be easily seen from the following output.

```
## ------------------------------------------------
##        Test             Statistic       pvalue
## ------------------------------------------------
## Shapiro-Wilk              0.9823         0.4275
## Kolmogorov-Smirnov        0.0676         0.8851
## Cramer-von Mises          5.6582         0.0000
## Anderson-Darling          0.3838         0.3864
## ------------------------------------------------
```

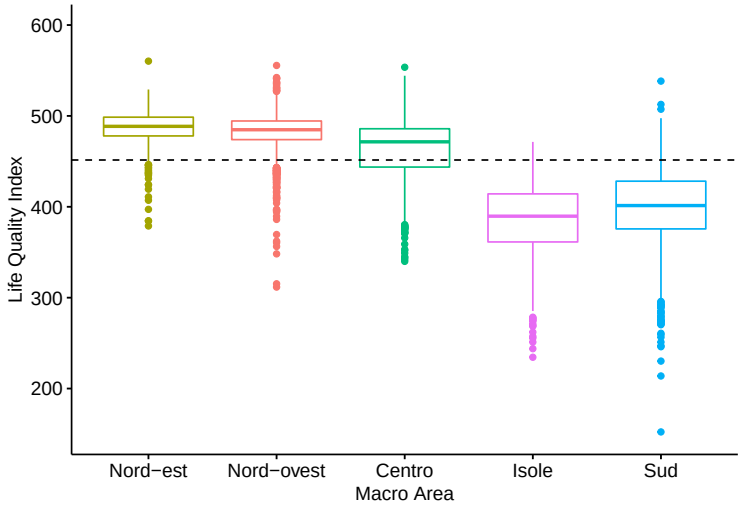


Once created and evaluated our simple regression model on the municipalities which are provincial capital, it has been used to make inference **predicting the response variable** for all the remaining observations (the entire set of municipality for which the same regressors are available). Below we can see the descriptive statistics of the response variable inferred on the bigger dataset. The quality of life now ranges from a minimum of 152 to a maximum of 560, with a mean of 451.

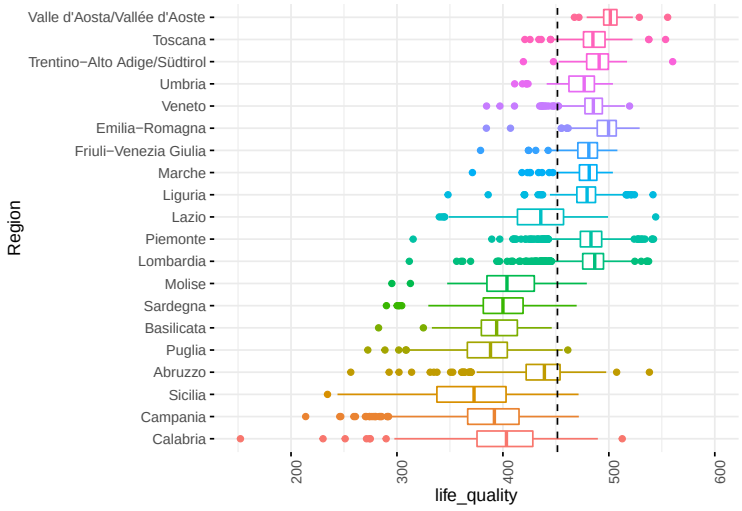
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   152.3   417.3   472.0   451.5   489.1   560.3
```

We can also make a comparison of the distributions of the quality of life scores between the **5 levels of geographical areas** (North-East, North-West, Centre, Islands and South). From there It can be easily

found that the first three groups, North-East, North-West and Centre have their distributions centered (with their median) **above the mean**, contrary to the two remaining groups of the South and Islands which are **well below**.



If we want to come more in depth of the decomposition for the geographical areas, we can check at **regional level** the 20 ditributions of the quality of life score of municipalities. Even here it is possible to check what it has been found before: all the regions of the south and the islands have their distributions centered with their median below the mean, and the center-north ones are above, *except* for the *Lazio*.



3.2. Classification of quality of life level

Before starting to work on the *classification task*, we need to construct a categorical variable. To accomplish this task we discretize our observations in 4 categories according to their level of quality of life:

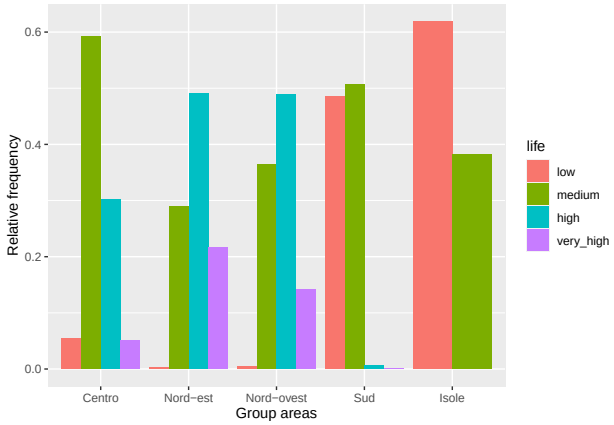
- `low`: *from 150 to 400*;
- `medium`: *from 401 to 480*;
- `high`: *from 481 to 500*;
- `very high`: *from 501 to 570*;

and they are distributed in the following way:

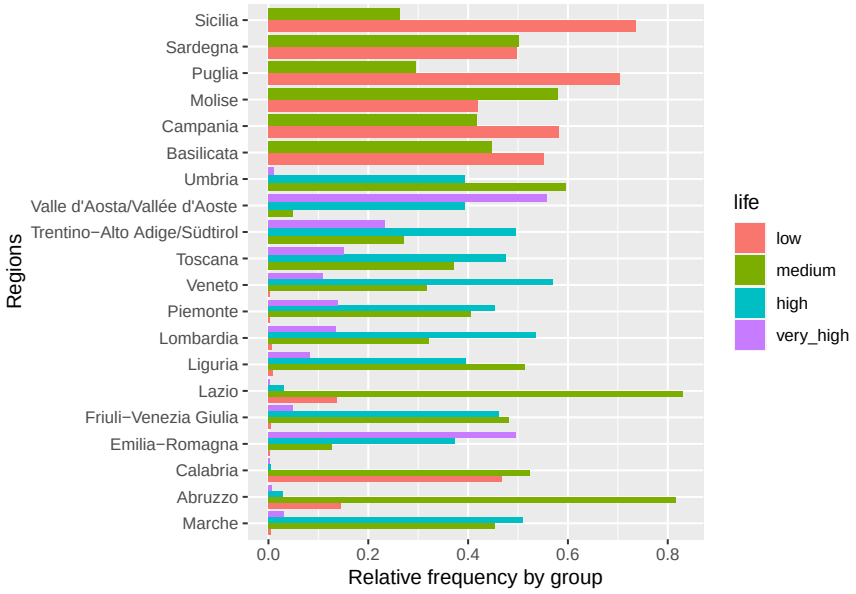
```
##
##       low    medium      high very_high
##      1361      3088      2266       709
```

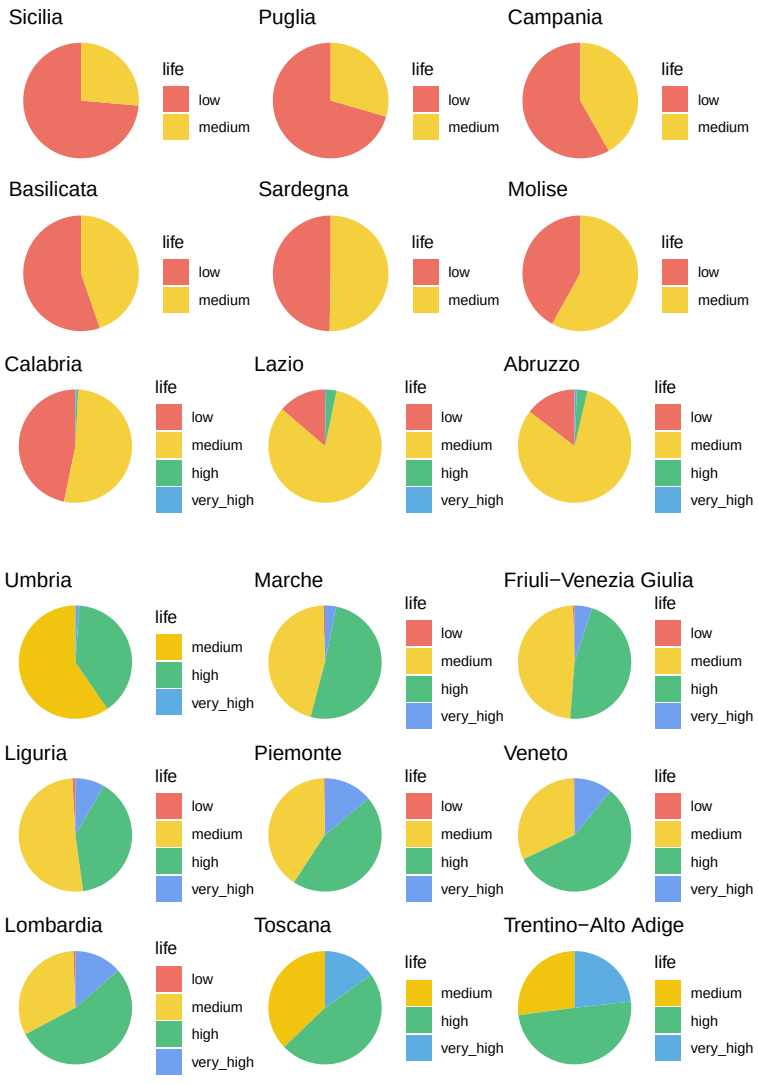
Another pleasant way to understand the **relative frequencies** of the 4 levels of quality of life between the geographical areas and regional areas is to use the bar chart. From there it can be seen that:

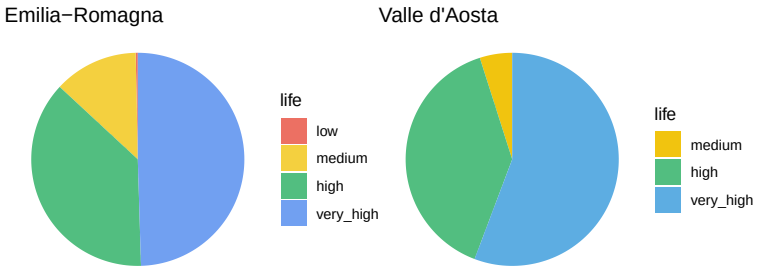
- **islands** have only low and medium quality of life municipalities (more than 60% low and less than 40% medium)
- **south** has a very similar situation with almost the same number of low and medium municipalities, and just a few in high and very high
- **centre** has the majority of medium quality of life municipalities (more than 60%) and more than 25% of high quality ones
- **north-east** and **north-west** share quite the same situation with a majority of high quality municipalities, but north-east has a slightly better situation



Even at regional level it can be found the discussion already done for the geographical areas. And plotting the same information with pie charts allows us to discriminate in two "clusters" the regions in which the quality of life is at least very low, from the regions in which is high. The discriminant is ultimately the geography: *south versus north*.


Sicilia
Sardegna
Puglia
Molise
Campania
Basilicata
Umbria
Valle d'Aosta/Vallée d'Aoste
Trentino−Alto Adige/Südtirol
Toscana
Veneto
Piemonte
Lombardia
Liguria
Lazio
Friuli−Venezia Giulia
Emilia−Romagna
Calabria
Abruzzo
Marche
Regions
0.0
0.2
0.4
0.6
0.8
Relative frequency by group
life
low
medium
high
very_high


Sicilia
Puglia
Campania
Basilicata
Sardegna
Molise
Calabria
Lazio
Abruzzo
Umbria
Marche
Friuli−Venezia Giulia
Liguria
Piemonte
Veneto
Lombardia
Toscana
Trentino−Alto Adige
life
low
medium
high
very_high

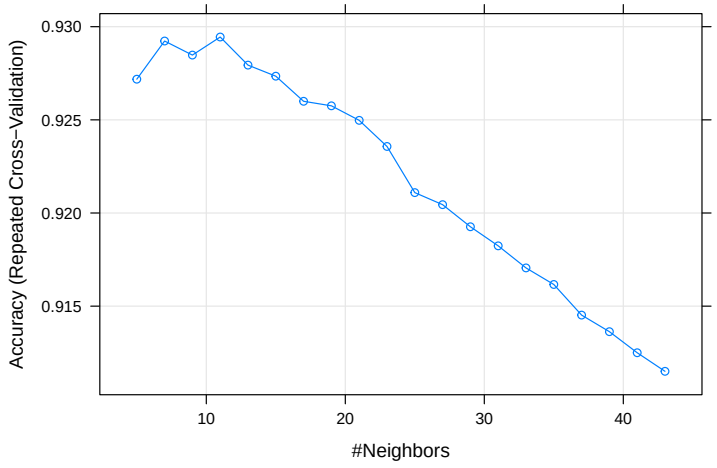


3.3. Classification with k-Nearest Neighbour algorithm

One simple and powerful way to deal with classification of observations with numerical features is the **k-Nearest Neighbour** algorithm. It uses some notion of distance (usually the Euclidean one) to classify each point with the majority label of the `k` closest points. In this way we can construct a model able to classify whichever municipality, according to a *similarity (by proximity) principle*, as a municipality in which there is a low, medium, high or very high level of quality of life.

For the sake of model creation it has been used only a subset of features (the ones selected by the forward search) in order to avoid the *curse of dimensionality* for the k-NN algorithm, which is an exponential dependence on the number of features of the training set size; in practice it means that it becomes difficult for a learning algorithm when observations lie in a high-dimensional space[6].

For the tuning of the best hyperparameter k, a **repeated cross validation** with 5 repetitions is made on the entire dataset (not only with the training set), and as it can be seen in the graph below, the **cross validated accuracy** is *maximized* with the hyperparameter $k = 11$.



[6]Machine Learning — Statistical Methods for Machine Learning: Risk Analysys for Nearest-Neighbor, Nicolò Cesa-Bianchi http://cesa-bianchi.di.unimi.it/MSA/Notes/nnRisk.pdf

```
##    k
## 4 11
```

Training the `11-NN` classification model on the training set and checking how the model performs with test data confirms the **accuracy of almost 94%**.

```
## [1] 93.67145
```

```
##            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
```

4. Unsupervised Learning Methods

Switching our mind to *unsupervised learning techniques*, no more considering the response variable of the quality of life, we can try to see whether it is possible to **reduce the dimensionality** of our data projecting the data matrix onto a lower dimensional space (without loosing too much information) and ultimately if there exists some way to **cluster** our observations (for all the municipalities or just the ones which are provincial capital).

4.1. Principal Component Analysis

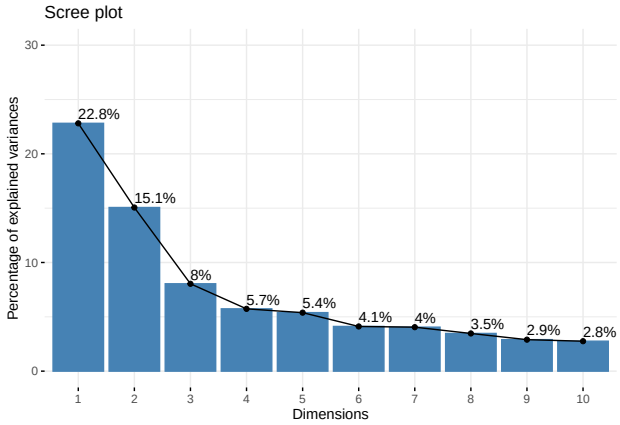
Principal Component Analysis is a technique which allows us to project data (scaled) onto a lower dimensional space, without loosing too much information. It is based on the **singular value decomposition principle** (which is a generalization of the *spectral theorem*) which decomposes our data matrix into: $X = U\Sigma V^T$.

1. X is the matrix of our data scaled
2. U is the orthonormal matrix composed by the eigenvectors of XX^T
3. V^T is the orthonormal matrix composed by the eigenvectors of X^TX
4. Σ is the diagonal matrix of the singular values associated to the principal components.

With **PCA** we retain the **principal components** associated with the **highest singular values**. For the data matrix we are analyzing unfortunately we are not able to project data on few dimensions, maybe because of heterogeneity of features. The independent variables measure pretty different things for our datapoints, and this makes our dimensionality reduction not an easy task.

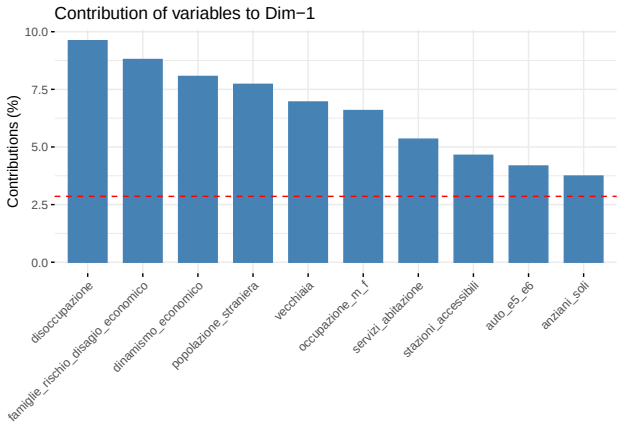
Using the subset of observations for the provincial capital municipalities we got a principal component analysis which can be also described by a scree plot.

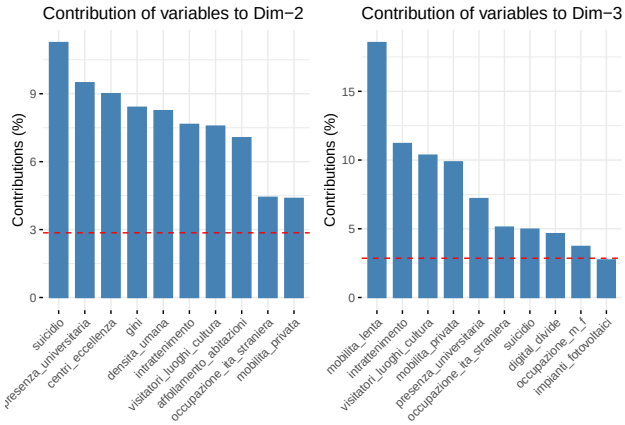
```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1   7.9822336        22.806382                    22.80638
## Dim.2   5.2705451        15.058700                    37.86508
## Dim.3   2.8139845         8.039956                    45.90504
## Dim.4   2.0065892         5.733112                    51.63815
## Dim.5   1.8844246         5.384070                    57.02222
## Dim.6   1.4393775         4.112507                    61.13473
## Dim.7   1.4164008         4.046859                    65.18159
## Dim.8   1.2140937         3.468839                    68.65043
## Dim.9   1.0144261         2.898360                    71.54879
## Dim.10  0.9644233         2.755495                    74.30428
```

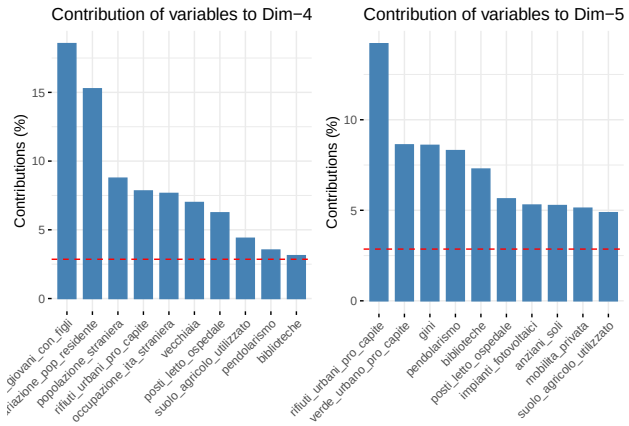


In order to understand what each of the 5 dimensions found represent, we can check the **loadings** (i.e., *contributions*) of the variables with a simple bar chart. In that way we are able to recognize what each dimension stands for.

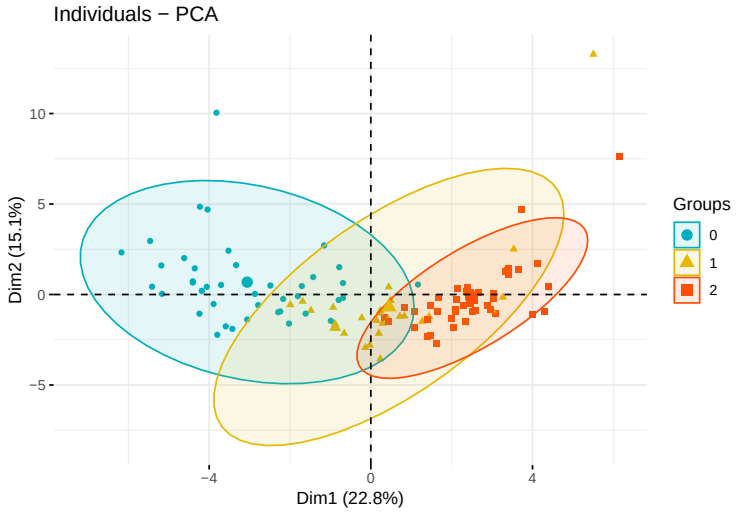
1. *dimension 1*: **Economical dimension**
2. *dimension 2*: **Social Cultural dimension**
3. *dimension 3*: **Transportation dimension**
4. *dimension 4*: **Demographic dimension**
5. *dimension 5*: **Green dimension**







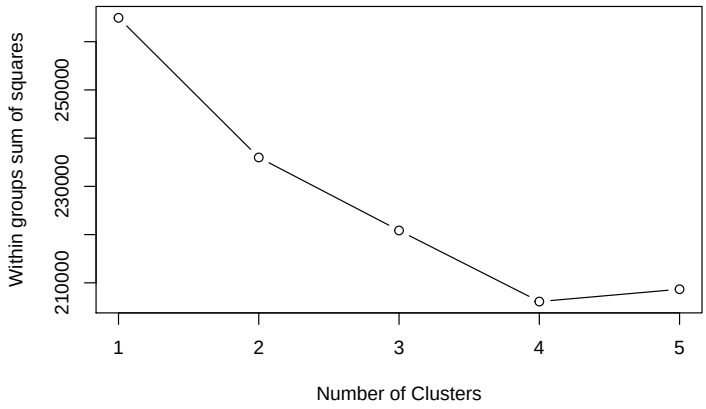
One last check could be to visualize the positioning of our observations in the 2 main dimensions grouped by the geographical areas. The south group seems distant from the center and north ones, which are instead pretty close to each other.

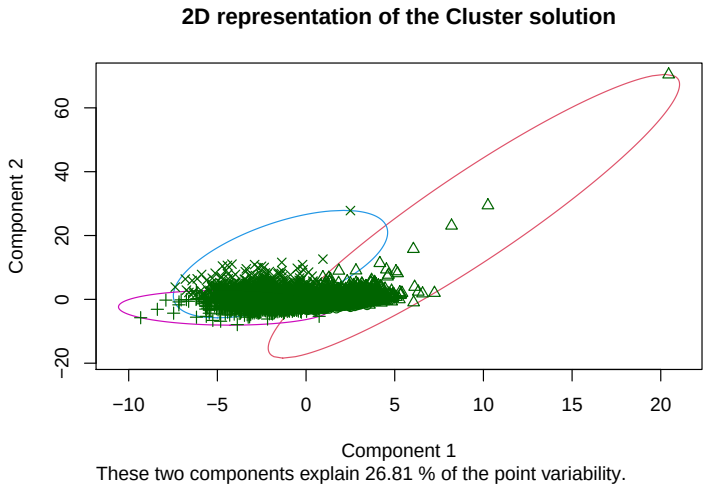


4.2. Clustering with k-means algorithm

Clustering methods are many, but the one which best fits numerical features is the **k-Means** algorithm. This algorithm makes a partition into the K pre-specified clusters *minimizing the within variation of each cluster*, and at least assigning each datapoint to the cluster K with the *closest centroid*.

Making an attempt with 5 clusters, as the ones used for geographic classification (e.g. North-East, North-West, Centre, South and Islands) we can see that the within-cluster variation is *minimized* for the number of cluster $K = 4$.

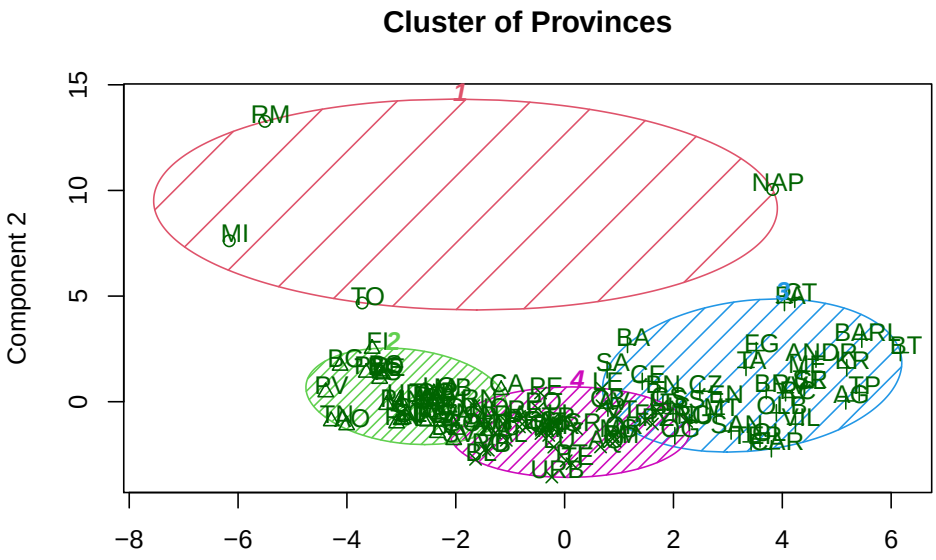




Plotting the results of the clustering for the municipalities, we can see nothing, so, in order to make something more interpretable we do the same procedure on the subset of observations with the municipalities which are provincial capitals, just to better understand how the 4 clusters are composed.

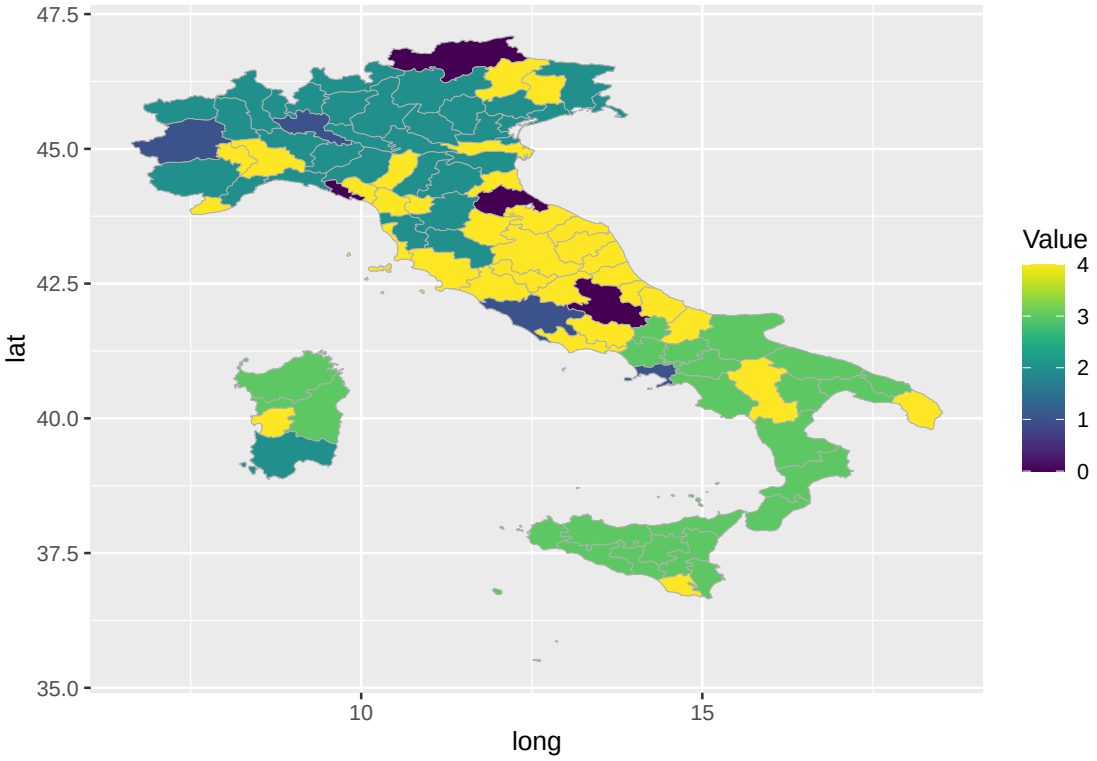
Provincial capitals municipalities are in a sense clustered by the *4-Means* algorithm in the following way:

1. **Milan**, **Naples**, **Rome** and **Turin**, probably the **biggest** among **italian cities**;
2. mostly **nothern cities**;
3. mostly **southern cities**;
4. mostly cities of the **centre**.



These two components explain 37.87 % of the point variability.

For a better visualization of the cluster just plotted above, I decided to create a geographical map of Italy because it better delivers the message: Italy is clearly divided in 3 groups that correspond to the geographical positioning of cities. Therefore similarities and differences correspond not only to the latitude but also the longitude. The result plotted is not completely in accordance with the expectations set at the beginning of the analysis.



5. Conslusions

The main findings of this paper can be summarized in the following points:

1. A **descriptive model** with *linear regression* for the response of quality of life score is found with statistically significative coefficients:

- Quality of life of Italian municipalities is **negatively** related with the **unemployment rate**. The higher the unemployment, the lower is the quality of life experienced by citizens;
- The cars driven by italians influence their lives: the **newer and less fuel-consuming** the **cars** are, the higher is the quality of life;
- When **women are not actively engaged in the labour market** as much as men, italian municipalities experience a life with **lower quality**;
- Municipalities where there is a **cosmopolitan and inclusive attitude** towards **foreign citizens**, all the community can benefit from it, at least **increasing the life quality**;
- **Tourism positively** affects the conditions in which citizens live: having tourists that visit the historical buildings and museums, is a factor of **better everyday life**.

The model also seems to predict well when tested on unaccessed data.

2. A **classification model** with *11-Nearest Neighbour* is found with an accuracy of almost 94%. It means that asking to the model which is one of the 4 category level (low, medium, high, very high quality of life) attached to a municipality for which we have the 5 information (i.e., unemployment, E5-E6 cars, ratio between males and females employment, foreign population and visitors of cultural places) available, approximately in the 94% of the times it will classify well.

3. *Principal component analysis* reveals 5 main dimensions along which our data are spread the most, and they are the following:

- The first and *main component* is the **Economic dimension** which turns out to have the following variables as the main *contributors*: unemployment, risk of families living in disconfort and economic dynamism.

As expected the most important dimension along which data are spread is the economic one.

- The second principal component has been called **Social Cultural dimension**, because it identifies the variables which measure social factors such as suicide, gini coefficient of inequality and cultural factors driven by presence of universities, centres of excellence, presence of tourists visiting cultural places, museums and places of entertainment.
- The third dimension is the **Transportation** one, because the main contribution comes from the measure of slow mobility.
- The **Demographic dimension** is the fouth component and it is prevalently guided by the number of young couples with children, the variation of resident population and share of foreign residents.
- The last component found is the **Green dimension**, and it has the most important contribution from the urban waste and green areas per capita.

4. Clustering method of *K-means* reveals the possibility to make a partition of our observations (provincial capital municipalities) into 4 groups, according to the first 2 dimensions just identified: *Economic* and *Social-Cultural* dimension.

- **Milan**, **Naples**, **Rome** and **Turin**, which are the **biggest italian cities**, can be identified in the upper part of the graph because of much difference in the Social-Cultural dimension with respect to the other municipalities. These cities are clearly differentiated from the other 3 clusters in particular for the suicides, the inequality, the presence of universities, centres of excellence and museums. We are clearly discussing about the most important cities where the major part of Italians live, and they seem to be the ones where there is much inequality.

In the lower part of the Social-Cultural dimension, the remaining 3 clusters are mainly distinguished by the first principal component: the Economic one.

- Lower coordinates for the first component identifies mostly **nothern cities**; they are municipalities in which there is at least a lower unemployment rate and a lower risk for families to live in hardship.
- Middle coordinates for the economic component then identifies mostly cities of the **centre**.
- **Southern cities** instead comes with higher coordinates for economic dimension.