Welcome!

Let's look at italian regions in 2019!

AIM OF THE ANALYSIS

The analysis is carried out among the italian regions in 2019.
The dataset has been made starting from tables reported in Istat.
The aims of the analysis are:

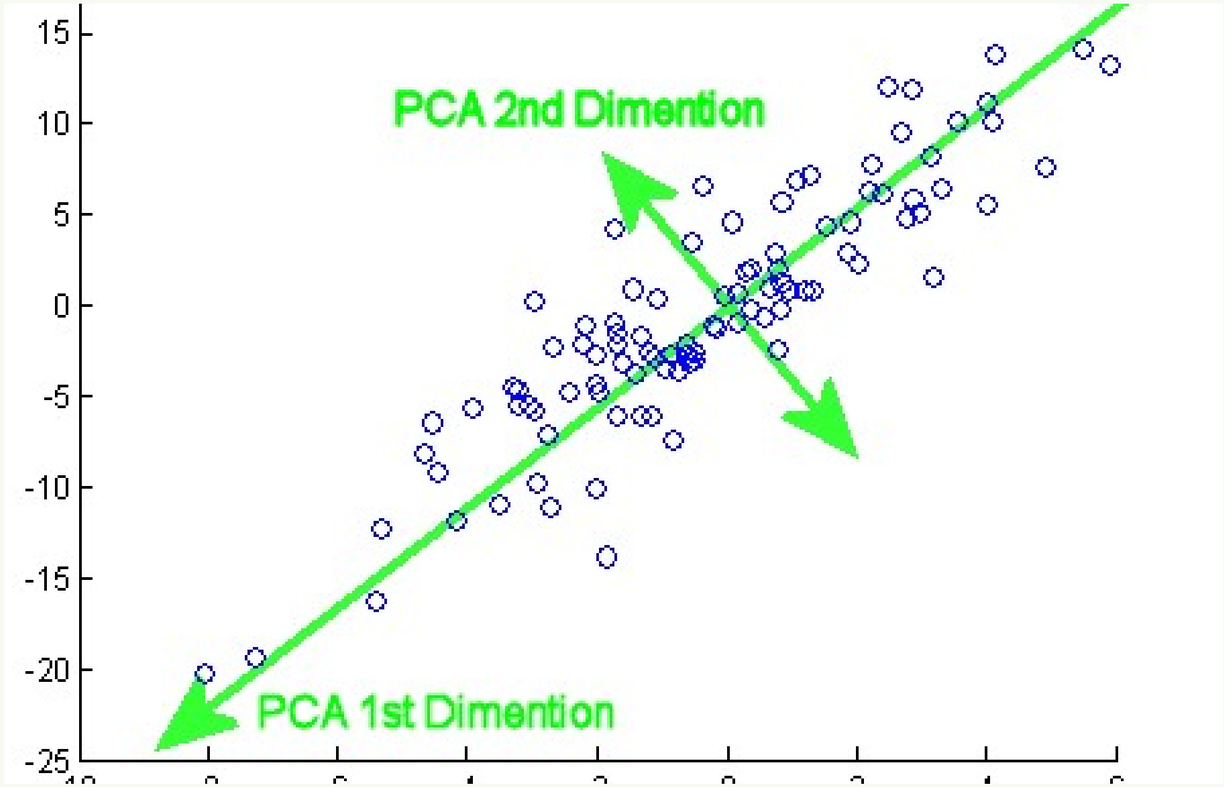
— **Obtain a low-dimensional representation of data.**
— Detect the presence of possible clusters.

Contents of variables

- **wine** = percentage, number of people every 100 inhabitants that drink almost one half liter of wine each day.
- **beer** = percentage, number of people every 100 inhabitants that drink almost one half beer of wine each day.
- **ape** = percentage, number of people every 100 inhabitants that drink at least one alcoholic appetizer each day.
- **liquor** = percentage, number of people every 100 inhabitants that drink liquor each day.
- **bitter** = percentage, number of people every 100 inhabitants that drink bitters each day.
- **cigarettes11.20** = percentage, number of people every 100 inhabitants that smoke between 11 and 20 cigarettes each day.
- **cigarettesmore20** = percentage, number of people every 100 inhabitants that smoke more than 20 cigarettes each day.
- **excigarettes** = percentage, number of people every 100 inhabitants that stopped smoking.
- **minsatisfaction** = percentage, number of people every 100 inhabitants that evaluated their quality of life with a 0 rating on a scale from 0 to 10. This data is gotten from an interview conducted by Istat in 2019.
- **maxsatisfaction** = percentage, number of people every 100 inhabitants that evaluated their quality of life with a 10 rating on a scale from 0 to 10. This data is gotten from an interview conducted by Istat in 2019.
- **confidence** = percentage, number of people every 100 inhabitants that considers other persons of the region trustworthy.
- **education** = percentage, number of persons every 100 inhabitants that completed high school.
- **income** = average income of a person in the region.

PCA analysis

Looking for a low dimensional representation of data.

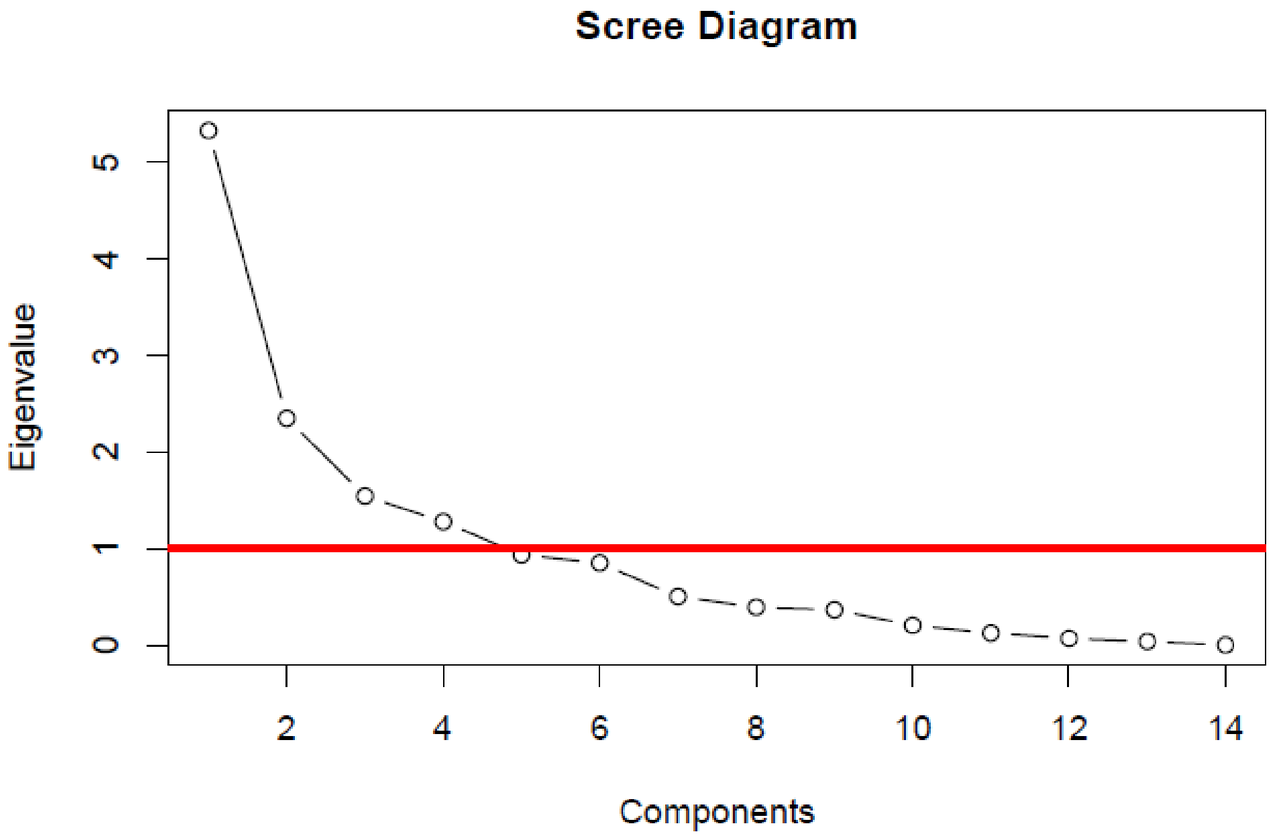


Eigenvalues and variances

```
##        eigenval   %var %cumvar
##  [1,]     5.326 38.045  38.045
##  [2,]     2.348 16.771  54.816
##  [3,]     1.543 11.025  65.841
##  [4,]     1.280  9.144  74.985
##  [5,]     0.936  6.687  81.672
##  [6,]     0.851  6.080  87.752
##  [7,]     0.504  3.599  91.351
##  [8,]     0.394  2.812  94.162
##  [9,]     0.368  2.626  96.788
## [10,]     0.207  1.480  98.268
## [11,]     0.127  0.910  99.179
## [12,]     0.071  0.509  99.688
## [13,]     0.040  0.284  99.972
## [14,]     0.004  0.028 100.000
```

The datset contains observations of regions and autonomous provinces. I started computing the eigenvalues, which suggest how much more each eigenvalue is able to explain the variance of the data respect to one variable of the starting dataset. I reported the list with the associated variance and cumulative variance to get to amount of variablity considered.

Eigenvalue selection



The scree diagram is presented with the threshold fixed at 1 level, since each new variable should explain more than a variable of the starting dataset to obtain a reduction of the dimensionality. The result seems to suggest 4 eigenvalues, however I considered 3 of them to get a visualization of the data.

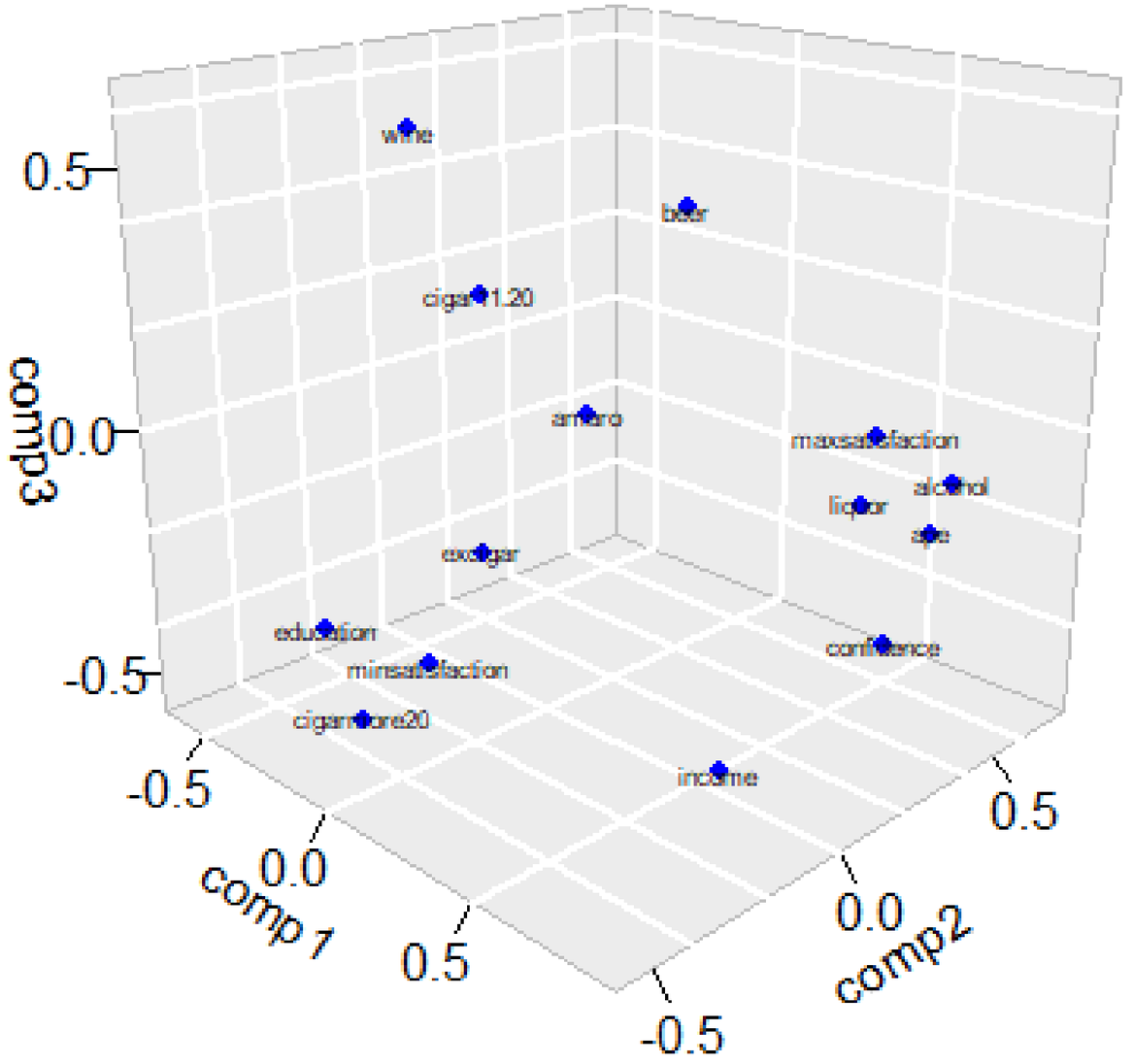
Components and communality

	comp1	comp2	comp3	communality
wine	0.2393	-0.4846	0.6584	0.7255922
beer	0.0710	0.3550	0.3752	0.2718410
ape	0.9319	0.2376	-0.0572	0.9281632
liquor	0.9336	0.0219	0.0547	0.8750807
amaro	-0.4924	0.5198	-0.1932	0.5499760
alcohol	0.9029	0.3351	0.0014	0.9275224
cigar11.20	-0.6627	0.3247	0.0859	0.5519802
cigarmore20	-0.0484	-0.4599	-0.4639	0.4290538
excigar	0.5694	-0.5877	0.0589	0.6730769
minsatisfaction	-0.3925	-0.0512	-0.5899	0.5046597
maxsatisfaction	0.3103	0.7612	-0.1381	0.6947831
confidence	0.8178	0.2370	-0.2986	0.8141278
education	-0.2368	-0.4241	-0.3420	0.3528990
income	0.8516	-0.2592	-0.3556	0.9188586

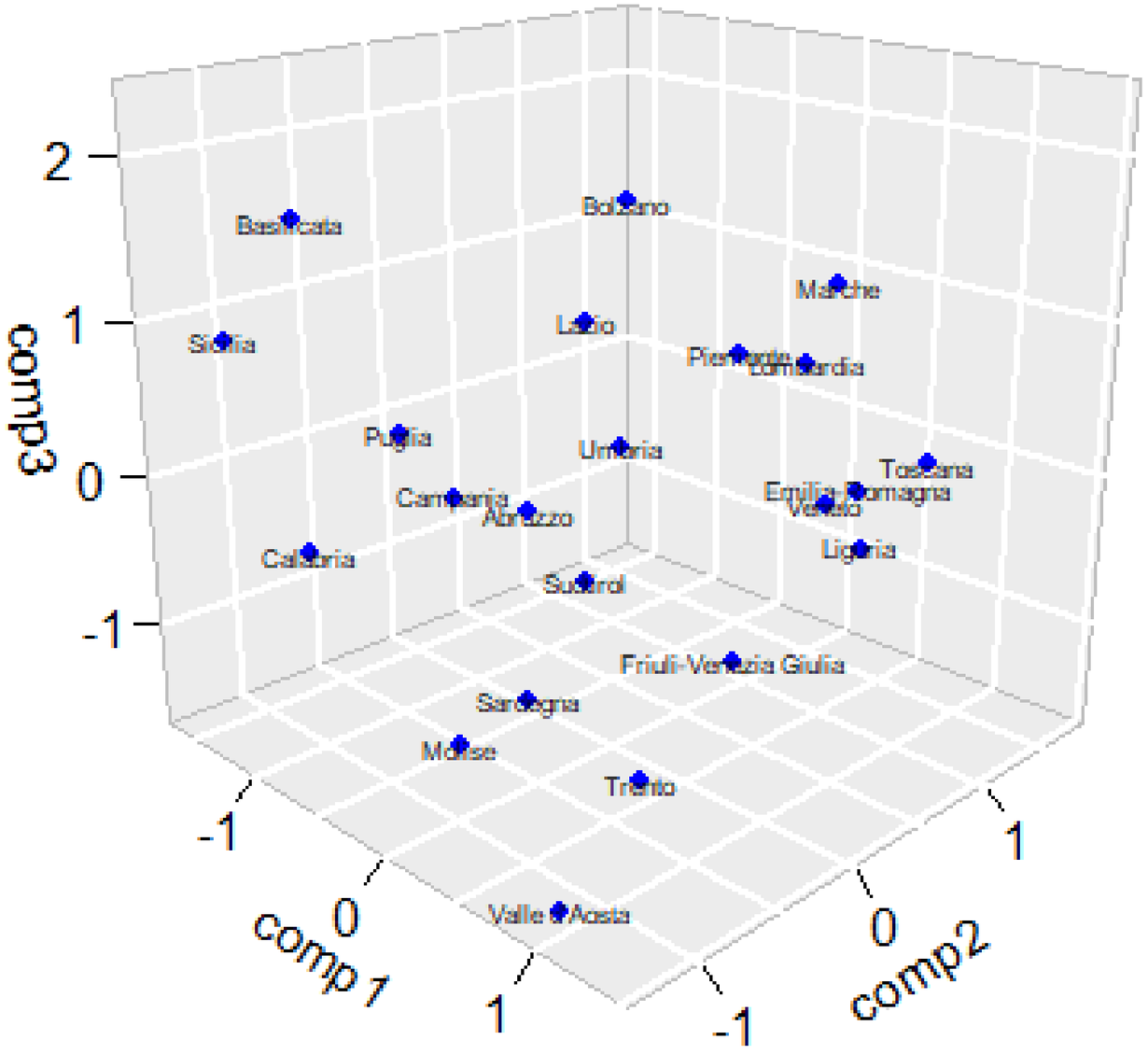
The model with three components is reported here with the associated communality. There are good findings. The first component brings clearly the biggest information; moreover, almost all the variables seem to be explained in a consistent way, given the widespread value of communality greater than 0.5.

In the following slide, the loading plot is reported to understand how the variables are spread according to the three components.

Loading plot - 3 components



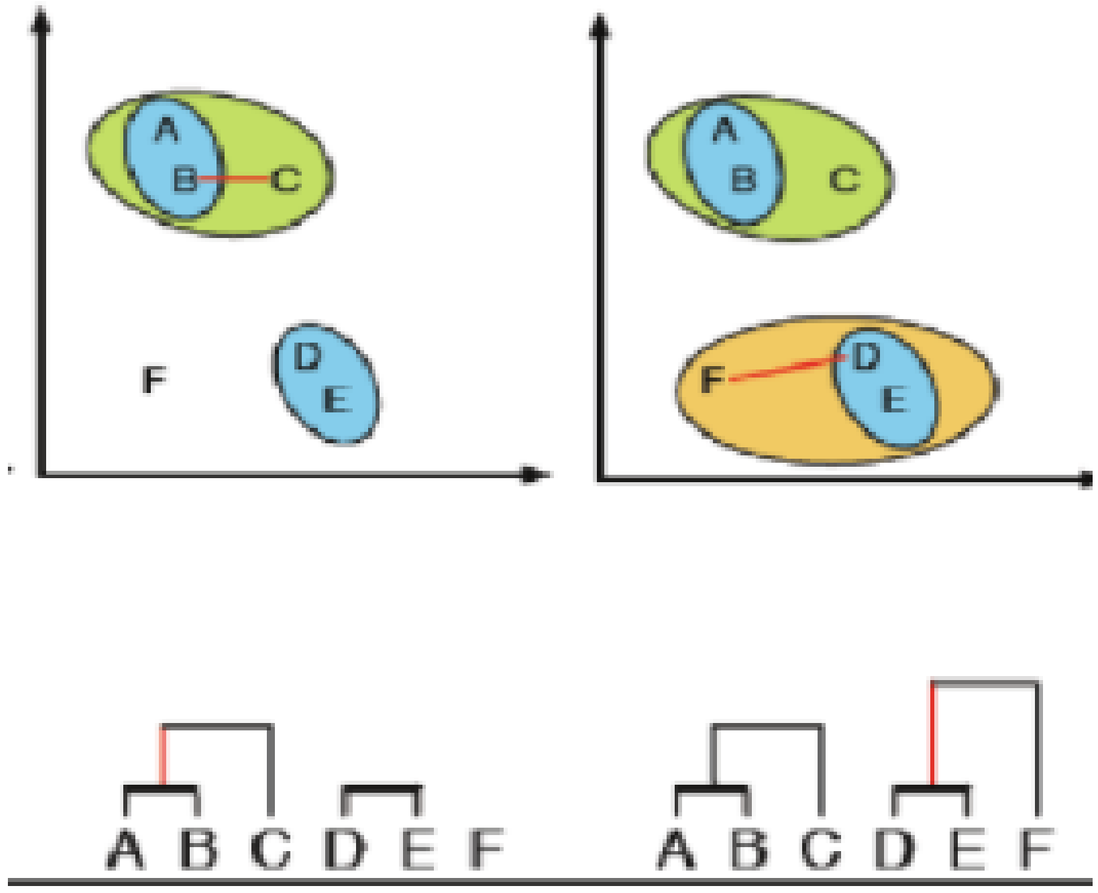
Score plot - 3 components



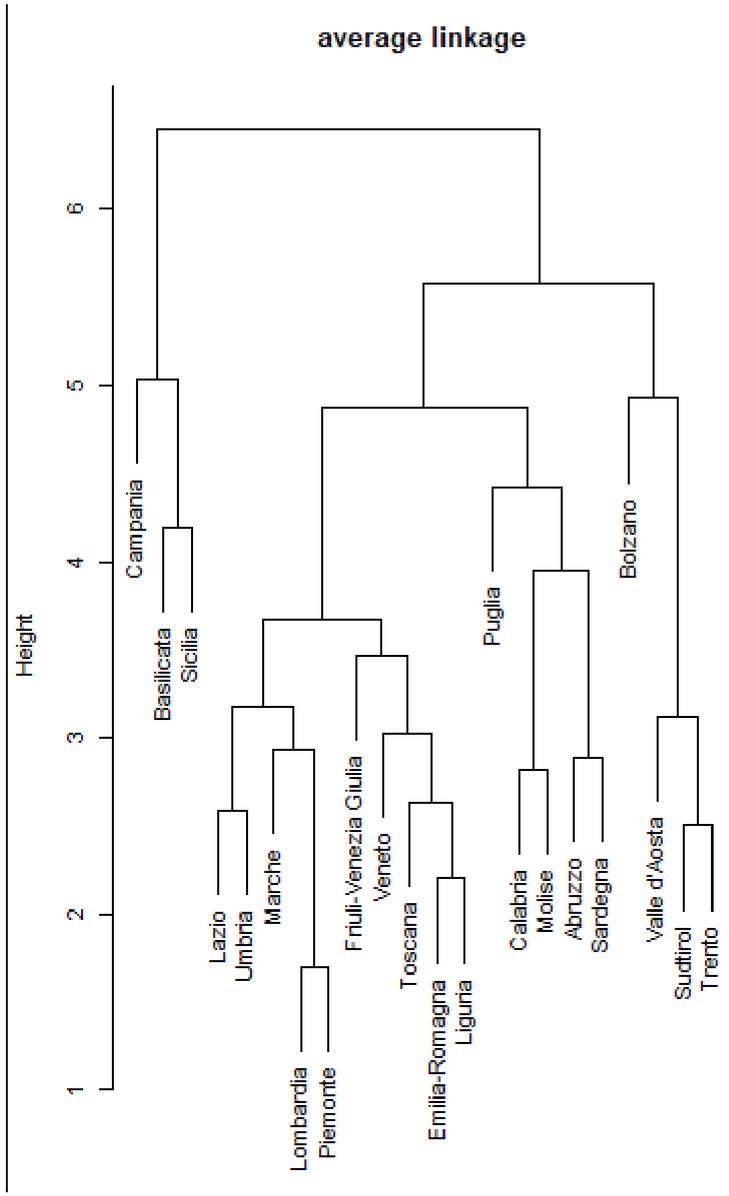
Hierarchical clustering

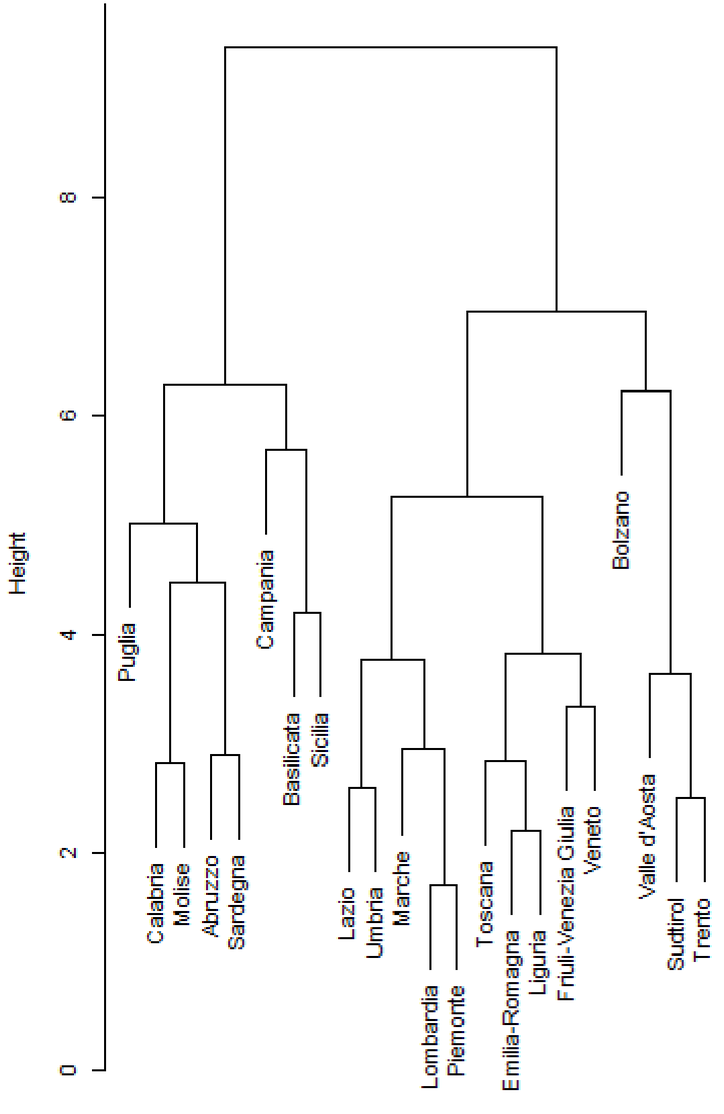
Detecting clusters among the data.

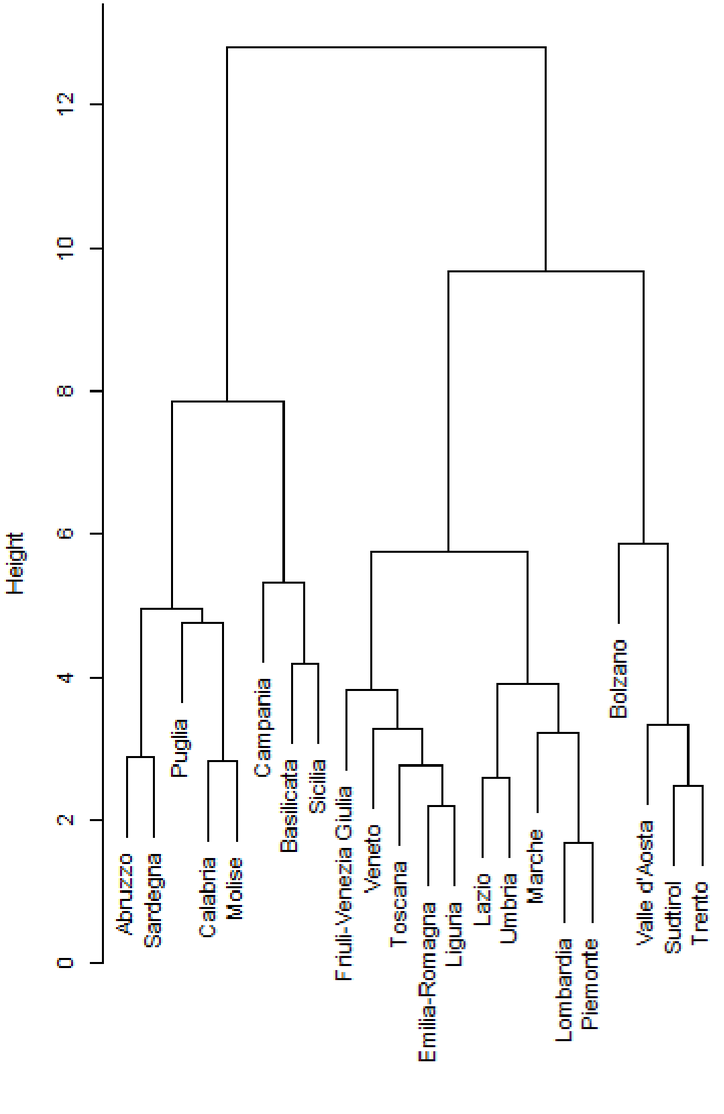
In the next slide, dendrogrmas obtain with different linkage methods are reported.



Dendrograms







Method selection

To select the best number of clusters, I used the elbow method, which has given 3 as best result. After that, I compared the three methods with 3 clusters and I chose the ward method to go on with the analysis, because it has the same result of the complete method and it is more able to face the outliers. This is the clusters I found:

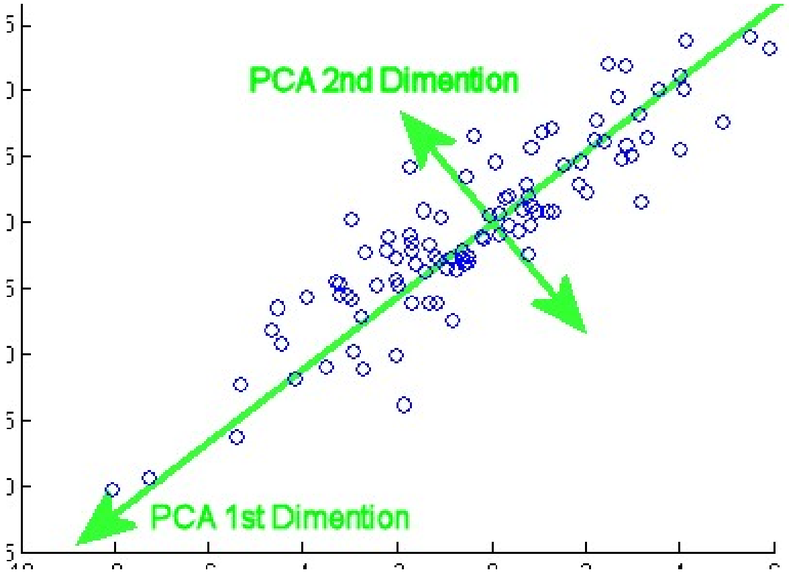
```
##                 Abruzzo              Basilicata                 Bolzano
##                       1                       1                       2
##                Calabria                Campania          Emilia-Romagna
##                       1                       1                       3
## Friuli-Venezia Giulia                   Lazio                 Liguria
##                       3                       3                       3
##                Lombardia                 Marche                  Molise
##                       3                       3                       1
##                Piemonte                  Puglia                Sardegna
##                       3                       1                       1
##                 Sicilia                Sudtirol                 Toscana
##                       1                       2                       3
##                  Trento                  Umbria            Valle d'Aosta
##                       2                       3                       2
##                  Veneto
##                       3
```

Features of the clusters

```
##   Group.1       wine       beer        ape     liquor      amaro    alcohol
## 1       1 -0.3123305  0.4518803 -0.9566241 -1.0010347  0.7435613 -0.7982472
## 2       2 -0.4196941  0.1388650  1.4823827  1.1199941 -0.2334437  1.5643127
## 3       3  0.4177421 -0.4170503  0.1723462  0.3528301 -0.5014716  0.0128727
##   cigar11.20 cigarmore20    excigar minsatisfaction maxsatisfaction
## 1  0.7182588  -0.1004421 -0.7214088      0.27721374      0.02042701
## 2 -0.1582651  -0.6863541 -0.3482663     -0.70796124      1.34818269
## 3 -0.5113010   0.3548953  0.7164335      0.06141351     -0.55561468
##      confidence  education     income
## 1 -0.8473830081 -0.1344177 -1.1235455
## 2  1.6929854256 -0.6905241  0.8349968
## 3  0.0007122362  0.3837438  0.5648376
```

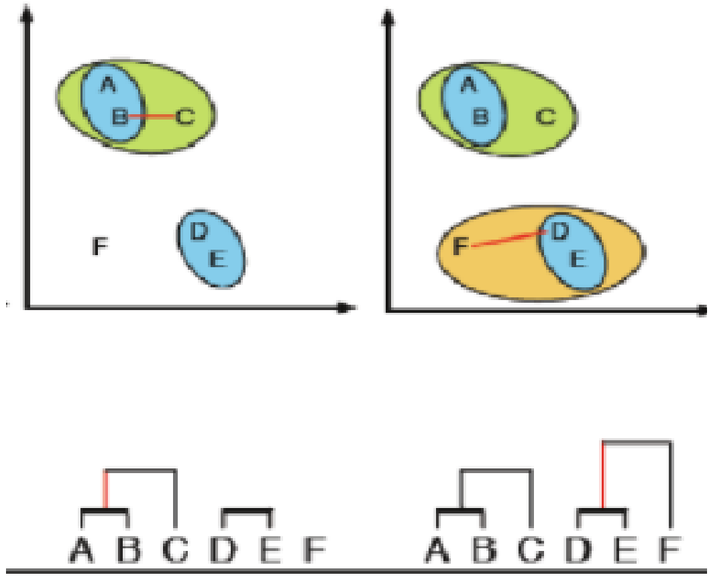
Variables are normalized to take into account different units of measure.
Cluster 2 seems to be the best one: higher income, best satisfaction and confidence and worst levels of smoke. Cluster 1, including the south-centre regions of Italy, has the worst values and lower than the mean for several variables (income, satisfaction, cigar11.20), while cluster 3 has values at an half way between the previous two clusters.

Conclusions



PCA

Low dimensional representation of data made with 3 components.



CLUSTERING

3 clusters detected: autonomous provinces, with best living conditions, north italy, at an half way, and south-center of italy, the worst situation.


Thank you for
the attention!