Unsupervised learning exam

Matteo Ciarrocchi

Abstract

This analysis is focused on the habits of the population all over one country: Italy. It is carried out at a regional level in 2019, just to look at a period which wasn't affected by extraordinary events like Covid pandemic. The analysis aims to present a low-dimensional representation and to detect possible clusters among the considered variables. PCA satisfied the need to visualize data, while hierarchical clustering selected our groups and which of them have the best living conditions and healthy habits, finding as the best the cluster including the autonomous provinces.

Problem understanding and data description

The analysis I prepared takes as units of observation the regions and aims to find relations between two macro-sets of variables: one includes regressors which represent typically habits, sometime defined as unhealthy, like smoke and drink, and they are grouped based on the frequency with which they are consumed, while the other contains variables which are related mostly to economic and instruction indicators, like income and education. The hope is to detect possible clusters among the regions and to understand how some variables move together according to the considered observations. All the data are taken from tables published by *Istat*, cleaned of irrelevant information through an accurate data manipulation. All the variables which compose the resulting dataset are presented more technically and deeply below:

- **Region** = set of regions and autonomous provinces.
- **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.

I analyzed the dataset for two different aims with two different methods: *principal component analysis* and *clustering*.

Principal Component Analysis (PCA)

I carried out an analysis based on PCA because I wanted to find out a low-dimensional visualization of the dataset with all the needed information. I tried to reduce the dimensionality of the dataset because it seemed a reasonable choice, since there are variables with similar contents which differ only for the considered frequency, like the cigarettes-related variables. First of all, I computed eigenvalues and eigenvectors of the dataset starting from the correlation matrix, as reported the following results:

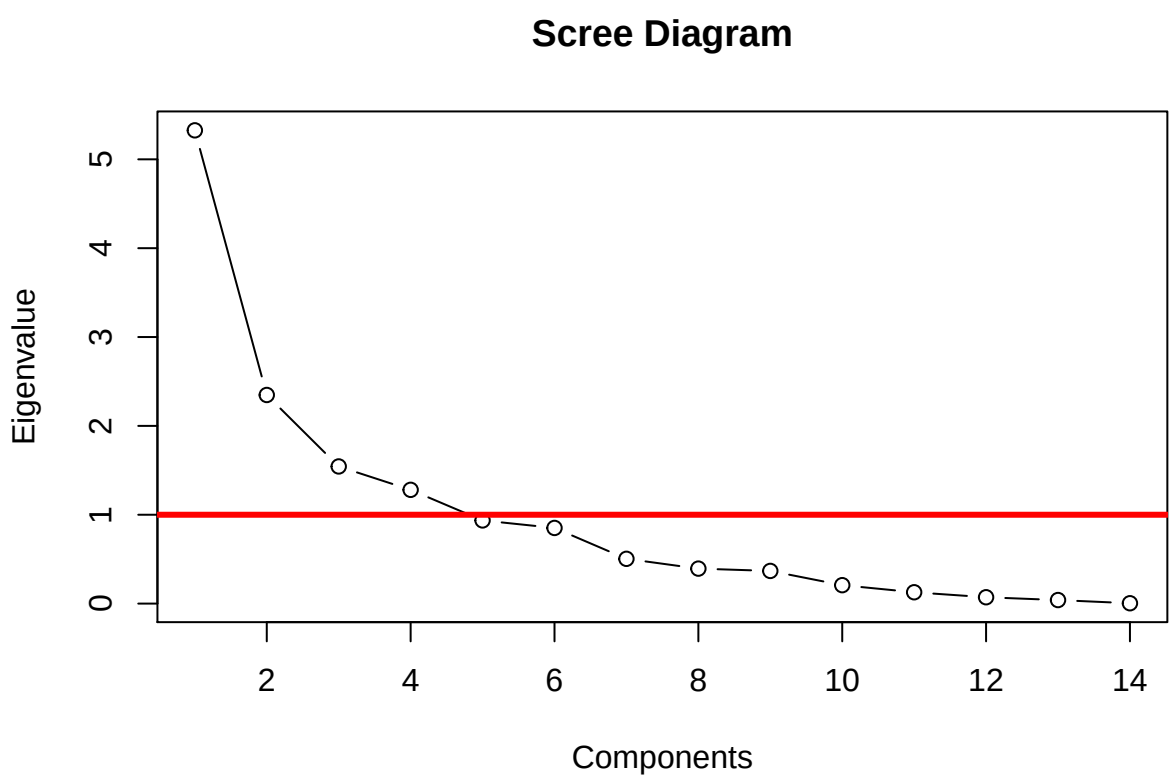
```
## eigen() decomposition
## $values
##  [1] 5.326283981 2.347936058 1.543459949 1.280183569 0.936149858 0.851247653
##  [7] 0.503814327 0.393647639 0.367612750 0.207217094 0.127458813 0.071300634
## [13] 0.039796128 0.003891548
##
## $vectors
##              [,1]        [,2]        [,3]         [,4]         [,5]
##  [1,] -0.10368443  0.31627479 -0.52992493  0.123207050  0.004102244
##  [2,] -0.03075419 -0.23170340 -0.30198502  0.684501464  0.091687884
##  [3,] -0.40377163 -0.15506126  0.04604466 -0.073871440 -0.054157329
##  [4,] -0.40453978 -0.01427411 -0.04400139  0.126201648 -0.191779612
##  [5,]  0.21335293 -0.33921473  0.15550483  0.375441402 -0.294527884
##  [6,] -0.39122383 -0.21872100 -0.00115683  0.003971333  0.089179360
##  [7,]  0.28714502 -0.21188141 -0.06916037 -0.325664179  0.131569455
##  [8,]  0.02098579  0.30015835  0.37341240  0.255829390  0.277129245
##  [9,] -0.24672090  0.38356695 -0.04744601  0.226497114  0.217840576
## [10,]  0.17006066  0.03339318  0.47479116  0.233278223  0.415555943
## [11,] -0.13446525 -0.49674186  0.11112607  0.129802979  0.059609591
## [12,] -0.35436726 -0.15467536  0.24034559 -0.163699429 -0.059688751
## [13,]  0.10261495  0.27678318  0.27529728  0.171020695 -0.731936465
## [14,] -0.36901601  0.16916615  0.28625851 -0.061660394 -0.008062108
##               [,6]        [,7]        [,8]        [,9]        [,10]        [,11]
##  [1,] -0.336894560 -0.38600351  0.36988845  0.22492677 -0.05124995 -0.08396888
##  [2,]  0.193933352 -0.22040706 -0.34065856 -0.12137854 -0.16370833 -0.14569701
##  [3,]  0.070913689 -0.10906857  0.23595079 -0.08939289 -0.03166466  0.43191712
##  [4,]  0.007130684 -0.18305339  0.14873283 -0.01125012  0.10180255 -0.11124410
##  [5,] -0.249697874  0.03610734  0.31577078 -0.28901811  0.54626485  0.01271358
##  [6,]  0.079569152 -0.13026330 -0.05780599 -0.12530854 -0.07342990  0.54154372
##  [7,]  0.027776327 -0.64683017 -0.29994018  0.16184572  0.42123745  0.12750911
##  [8,] -0.656907214 -0.09935875 -0.26725016 -0.08689981 -0.08800447  0.26210107
##  [9,]  0.303663597  0.24391131 -0.26536557  0.13643654  0.56826333  0.06970228
## [10,]  0.394495444 -0.30678408  0.43027234  0.09138445 -0.16186913 -0.07931990
## [11,] -0.218835994  0.20488951 -0.06080740  0.76534755 -0.04816202 -0.04956856
## [12,] -0.101120464 -0.21072792 -0.29769077 -0.28523062 -0.10624286 -0.51607691
## [13,]  0.186991013 -0.22417656 -0.20299399  0.26910834 -0.15271496  0.17841825
## [14,] -0.043760800 -0.14224461  0.12579156  0.16455853  0.29012031 -0.28745718
##              [,12]        [,13]        [,14]
##  [1,] -0.36088310 -0.055798897 -0.016399199
##  [2,]  0.12063031  0.298482492  0.092638591
##  [3,] -0.03009678  0.161557933  0.712224654
##  [4,]  0.64161691 -0.532023324 -0.090740991
```

```
##  [5,] -0.19910890 -0.022973635 -0.043528481
##  [6,] -0.23474917  0.001419023 -0.628506710
##  [7,]  0.09625214 -0.037748560  0.069247664
##  [8,]  0.14625215 -0.051095538  0.083443204
##  [9,] -0.24318205 -0.220342890  0.121615653
## [10,] -0.10561292 -0.174351638 -0.012560890
## [11,] -0.08350717 -0.097713241  0.042358517
## [12,] -0.44034172 -0.232078163  0.111270780
## [13,] -0.13368272 -0.013192244 -0.006849445
## [14,]  0.16790174  0.671973007 -0.194441158
```

It is worth of notice to talk about eigenvalues. They are ordered in a descending way, and we can interpret each of them as how much more each eigenvalue is able to explain the variance of the data respect to one variable of the starting dataset. Following this assumption, we can say that the first eigenvalue explain as much variance as 5.32 variables among the initial regressors, and so on. Data about variance and cumulative variance explained by eigenvalues are reported:

```
##       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
```

Considering the statistical implications of eigenvalues, it is needed to decide a level at which we should stop considering eigenvalues for the low-dimensional representation of data. It seems reasonable to put a threshold at level corresponding to 1, since we want components working as new variables and able to explain our data as good as possible, without losing a lot of information. The story is graphically told by the following scree plot:

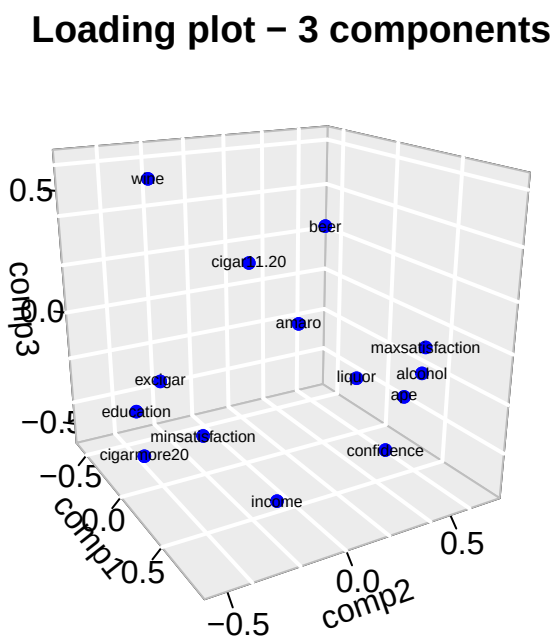


As analytically detected, the scree diagram shows the ideal number of eigenvalues to use and, according to the imposed threshold, It would be 4. However, I ran the PCA to obtain a low-dimensional visualization of the data, therefore I continued the analysis only with 3 components. Here I reported the relative loadings and the communality:

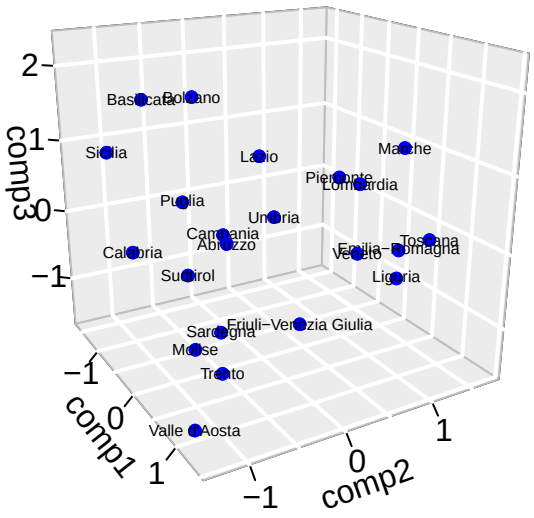
```
##                    comp1   comp2   comp3 communality
## wine              0.2393 -0.4846  0.6584   0.7255922
## beer              0.0710  0.3550  0.3752   0.2718410
## ape               0.9319  0.2376 -0.0572   0.9281632
## liquor            0.9336  0.0219  0.0547   0.8750807
## amaro            -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
```

Clearly, the first component carries out the bigger part of information about the data: there are big values for variables *ape*, *liquor*, *alcohol*, *confidence* and *income* (respectively 0.93, 0.93, 0.9, 0.81, 0.85). The second component carries out good amount of information about *amaro* and *maxsatisfaction* (0.51 and 0.76), while the last component is important for *wine* and *minsatisfaction* (0.65 and 0.59). Communality is computed as the sum of the squares of components for each row and it has a really important statistical meaning: it

detects how much variance is explained by the components for each variable with respect to the starting dataset. For example, considering *wine* variable, this new model composed only by 3 components has the capacity to explain the 72% of the variability of the *wine* variable of the starting dataset. Looking at the other variables and assuming a threshold value equal to 0.5 as guide judgment, I can state that this is a good model to explain almost all our starting variables, except for *beer*, *cigarmore20* and *education*: the communality selected them as non-relevant variables to explain the spread of the data. The *loading plot* is here reported to understand how the starting variables are correlated with the three components.



Just to give an example, high income regions have high levels of component 1, small levels of components 2 and 3 (smaller than 0). To finish with the PCA analysis, I presented another 3D plot, the *score plot*, to check how the states behave according to the components:



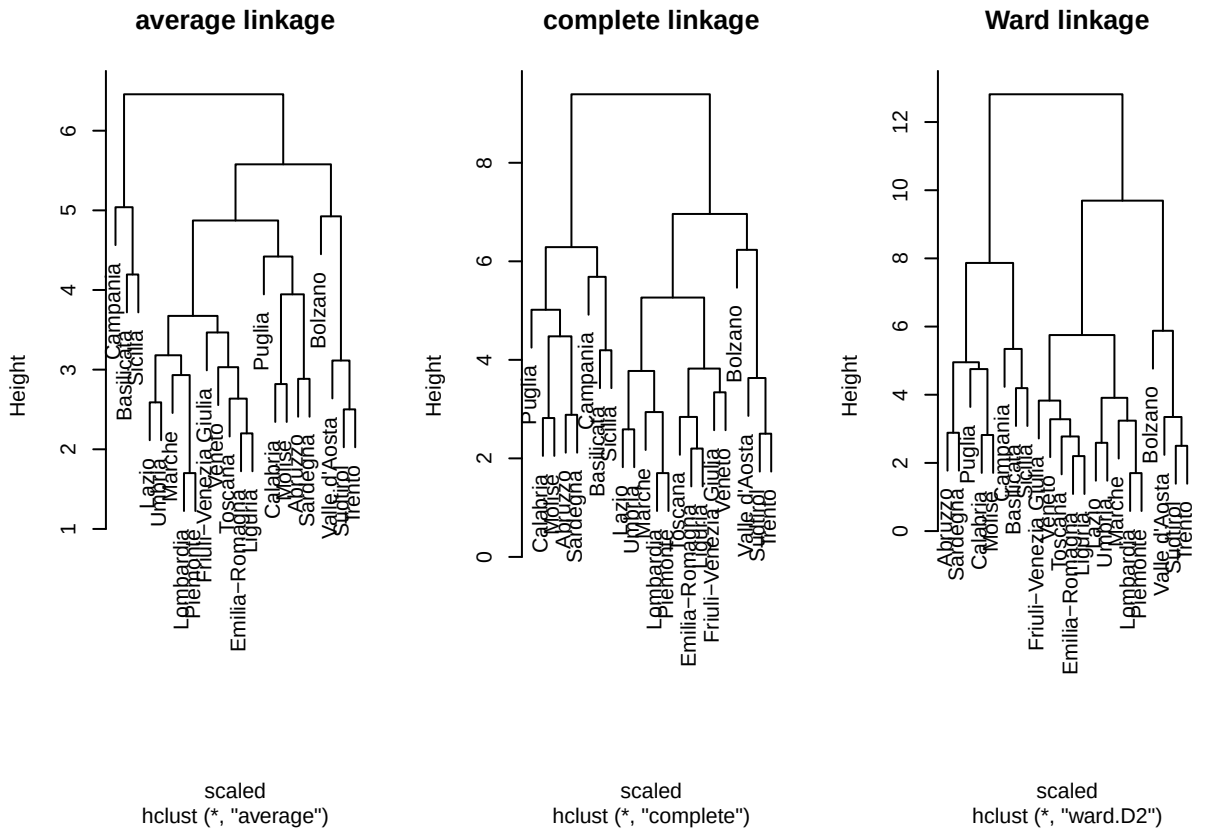
This is the low-dimensional representation of regional data which I was looking for.

Clustering

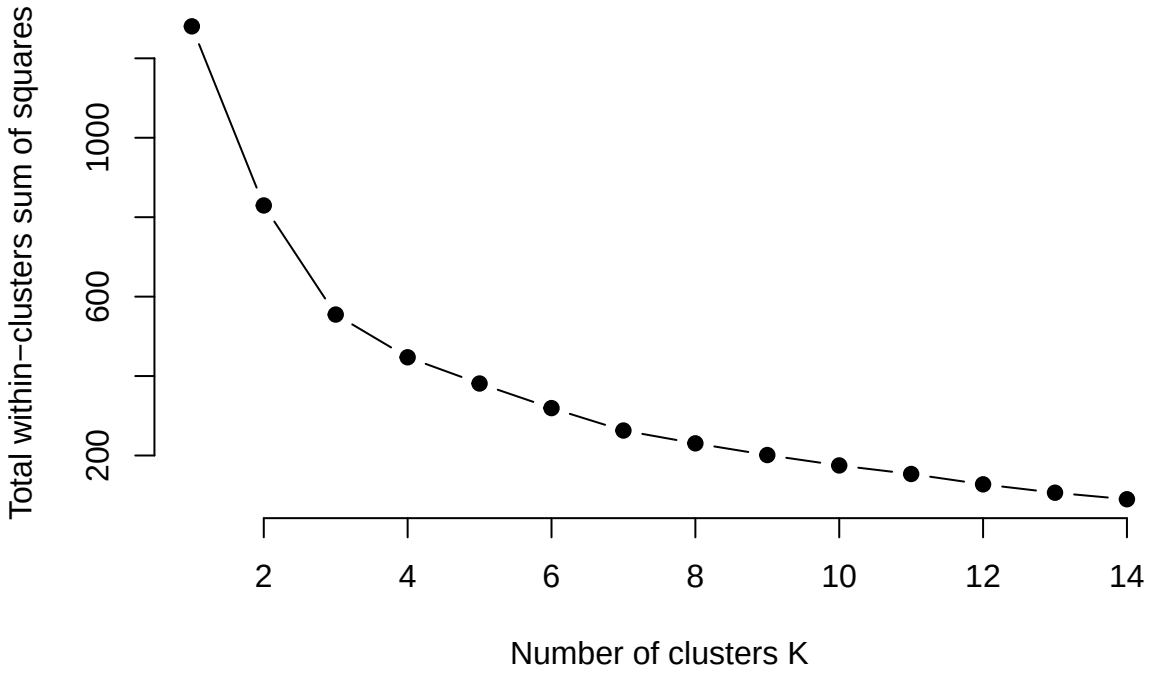
Clustering is used to find possible subgroups among data which share some features. In order to have an analysis as deep as possible and to obtain consistent results, I used several **hierarchical clustering** methods.

Hierarchical clustering

This kind of clustering could be run according to different methods. First of all, I scaled the data to cope with the different units of measures of the variables and, then, I decided to report here three dendrograms, each one obtained with a different linkage method: *average linkage*, *complete linkage* and *Ward linkage*.



All the three methods report a remarkable result: a potential cluster is composed by the autonomous provinces located in the north Italy, *Bolzano*, *Valle d'Aosta*, *sudtirol* and *Trento*. From the graph of complete linkage we can clearly see the main outliers of our dataset: *Bolzano* and *Campania*. At a first glance, the right number of clusters could be 3, considering the usual division of Italy in north, centre and Sud, or 4, adding the autonomous provinces. To select the optimal number of clusters in an analytical way, the graph of the elbow method is reported below.



The optimal choice seems to be K = 3. In fact, considering a further cluster, the decrease of the total within-clusters sum of squares would not be significant: the elbow of the graph supports this choice too. Having selected the number of clusters, the next step is to check whether the result differs on the basis of the linkage method. In order to do it, three cross-tabs are presented.

```
##       cut_compl
## cut_av  1  2  3
##      1  5  0 10
##      2  3  0  0
##      3  0  4  0
```

```
##       cut_ward
## cut_av  1  2  3
##      1  5  0 10
##      2  3  0  0
##      3  0  4  0
```

```
##          cut_ward
## cut_compl  1  2  3
##         1  8  0  0
##         2  0  4  0
##         3  0  0 10
```

From the tables, we understand that ward method and complete method have grouped regions in the same way. The difference is between these two methods and the average one: cluster 1 of average method is

composed of 15 regions, 10 of which are in the cluster 3 of complete and ward linkage. Cluster 2 is composed of 3 regions, present in cluster 1 of the other two methods, while cluster 3 of average linkage is composed of 4 regions which are in cluster 2 of the other ones. I decided to not consider the clusters detected by the average linkage because the size of the first cluster is too big and, since complete and ward method have the same result, I decided to go on with the ward method because it is more able to manage with outliers, as also shown in the dendrogram. Here the allocation of regions among the clusters is reported.

```
##               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
```

Finally, I described features of clusters using the starting variables. I computed normalized measures since the units of measure of the initial dataset were different, including percentage measures and integer numbers. The result is reported below.

```
##   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
```

Cluster 2 includes the autonomous provinces as a set of regions where there is a better life. Here, *income* value is the higher among the three clusters and higher than the mean and the same holds for *confidence* and *maxsatisfaction* (respectively value 0.8, 1.69 and 1.34, remembering that extreme values for a normal distribution are -3 and 3), while the *misatisfaction* value is lower than the mean (-0.7). Values related to the cigarette variables are all negative, maybe suggesting a better well-being of population, while data about drinkage are various: daily consumption of *wine* lower than the mean, as for *amaro*, while *ape*, *liquor* and *alcohol* take high values, suggesting drinking habits related to the territory and, for example, to the adverse weather conditions. **Cluster 1** is mainly composed by provinces located in the south-centre of Italy; for example, Campania, Sicilia and Calabria are present. In this cluster, the *income* takes value significantly low (-1.12), detecting for a widespread poverty, the *confidence* is low (-0.84) and a lot of people smoke from 11 to 20 cigarettes each day, suggesting a stressful and less healthy life. **Cluster 3** presents regions typically located in the north-centre of Italy, including Lombardia, Piemonte and Liguria. Here, the state of life seems

to be at an half way between the other two clusters: the *income* is slightly better than the mean (0.56), there isn't a *maxsatisfaction* high (-0.55), the life seems a little bit stressful (*excigar* = 0.71 and *cigarmore20* = 0.35), while data about drinking habits are around the mean.

Conclusion

I ran this analysis for two reasons. First, I wanted a low-dimensional visualization of my units of observation based on the variables in which I was interested, paying attention to which of them were particularly useful to explain the spread of the data. Second, I was interested to understand how and if regions could be grouped based on habits judged as unhealthy, like smoke and daily excessive drinking, variables about satisfaction and confidence and two more control variables (*education* and *income*). The results are satisfactory: the 3-dimensional representation of data is able to explain the 65% of the variance of the starting dataset and, looking at the communality, the variance of almost all the variables is taken into account. With regard to the clustering, I used a ward linkage method and selected the optimal number of clusters with the elbow method and the results have been good and consistent with the usual division of Italy between north and south. In fact, two out of three found clusters resemble this division, while the third one contain the autonomous provinces included in the analysis. The variables toke values which suggest that north-center Italy and autonomous provinces (located in the north Italy too) seem to live more quiet, healthy and with higher incomes.

Theoretical background

PCA

Used for data visualization or data pre-processing, the PCA is a technique which gives a low dimensional representation of the dataset, exploiting the covariance matrix of the starting dataset. The aim is to find a sequence of combination of variables that have maximal variance and are uncorrelated: the *components*. Clearly, the first component aims to maximize the explained variance, while the next ones need to be uncorrelated with the previous ones. PCA theory is developed based on a linear algebra technique: *singular value decomposition*, with which we can find eigenvalues and eigenvectors, reported in this paper. In this setting, eigenvalues represent how much variance the corresponding eigenvectors are able to explain, while these latter are vectors useful to describe how our units are spread. The scree diagram is present in the analysis and it is needed to choose the optimal number of components to use for the reduction of the data. This optimal choice is not mandatory, but it usually has value 1 as threshold because lower values work worse than whichever starting variable. Since PCA is made to obtain a low-dimensional reduction and a 4D visualization is clearly not possible, I chose 3 components for my model. The first component has the aim to explain as much variance as possible and, for this reason, values of the first component are the highest. The second component has to be uncorrelated with the previous one and, therefore, it is useful to explain the variance of those variables not considered by the first component and so on with the other components. Communality represents the total variance explained by the considered components with respect to the starting variables. Loading plot and score plot are also reported in the analysis: the first one explains how the initial variables move respect to the components, while the second one explains how regions are distributed respect to the components.

Clustering

The aim of the clustering is different respect to PCA: here, we want to detect if subgroups or clusters are present in the dataset, seeking a data partition such that observations within the same group are *similar*. There are different types of clustering and I made a hierarchical clustering analysis, useful because it could be run without choosing the number of clusters a priori, but there are also differences within each kind of

analysis. In fact, the results differ based on the type of distance used (I used only the Euclidean distance) and the type of linkage adopted, with regard to the kind of considered dissimilarity. I represented three dendrograms with three different linkage methods, just to compare graphically the results and to understand which type it was better to use.I presented the plot related to the elbow method too. It minimizes the total within-clusters sum of squares and it is needed to select the right number of clusters, which is located on the graph's elbow. After this, a comparison through cross-tables among the three methods is presented and it is useful to understand what differ among the methods and to detect which are the regions that changed their position. At the end, I reported selecte clusters with their regions and, then, the behaviour of the clusters respect to the standardized variables, since the initial units of measure were different, distributed between integer numbers and percentages.

Appendix

```
library(haven)
library(plot3D)
library(readxl)
final_dataset <- read_excel("C:\\Users\\Utente\\OneDrive\\Desktop\\SL exam\\unsupervised_dataset.xlsx")
unsupervised_dataset <- final_dataset
unsupervised_dataset <- unsupervised_dataset[,-c(1)]
rownames(unsupervised_dataset) <- final_dataset$Territorio
rho <- cor(unsupervised_dataset)
eigen(rho) ### eigenvalues and eigenvactors

### Variance and cumulative variance
autoval <- eigen(rho)$values
autovec <- eigen(rho)$vectors
pvarsp = autoval/p
pvarspcum = cumsum(pvarsp)
tab<-round(cbind(autoval,pvarsp*100,pvarspcum*100),3)
colnames(tab)<-c("eigenval","%var","%cumvar")
tab

### Scree diagram
plot(autoval, type="b", main="Scree Diagram", xlab="Components", ylab="Eigenvalue")
abline(h=1, lwd=3, col="red")
```

```
### Componnets and communality
comp<-round(cbind(
  -eigen(rho)$vectors[,1]*sqrt(autoval[1]),
  -eigen(rho)$vectors[,2]*sqrt(autoval[2]),
  -eigen(rho)$vectors[,3]*sqrt(autoval[3])
),4)
rownames(comp)<-colnames(unsupervised_dataset)
colnames(comp)<-c("comp1","comp2","comp3")
communality<-comp[,1]^2+comp[,2]^2+comp[,3]^2
comp<-cbind(comp,communality)
comp

#### 3D graphical representation of loadings
scatter3D(comp[,1], comp[,2], comp[,3],
          xlab = "comp1", ylab = "comp2", zlab = "comp3",
```

```
          main = "Loading plot - 3 components",
          bty = "g", ticktype = "detailed", d = 2,
          theta = 60, phi = 20, col = 'blue',
          pch = 19, cex = 0.8)
text3D(comp[,1],comp[,2],comp[,3],labels = rownames(comp),add = TRUE,
       cex = 0.5,
       adj = 0.5, font = 1)
```

```
#### 3D score plot
score <- unsupervised.scale%*%autovec[,1:3]

#score plot
scorez<-round(cbind
              (-score[,1]/sqrt(autoval[1]),
                score[,2]/sqrt(autoval[2]),
                score[,3]/sqrt(autoval[3])),2)
x <- scorez[,1]
y <- scorez[,2]
z <- scorez[,3]
#plots
scatter3D(x, y, z,
          xlab = "comp1", ylab = "comp2", zlab = "comp3",
          main = "Scores plot - 3 components",
          bty = "g", ticktype = "detailed", d = 2,
          theta = 60, phi = 20, col = 'blue',
          pch = 19, cex = 0.8)
text3D(x,y,z,labels = rownames(unsupervised_dataset),add = TRUE,
       cex = 0.5,
       adj = 0.5, font = 1)
```

```
################################ Clustering

### Dendrograms of the three methods
scaled <- dist(scale(unsupervised_dataset))
opar <- par(mfrow =  c(1,3) )
h1<-hclust(scaled, method="average")
plot(h1, main="average linkage")
h2<-hclust(scaled, method="complete")
plot(h2, main="complete linkage")
h3<-hclust(scaled, method="ward.D2")
plot(h3, main="Ward linkage")
```

```
####compute and plot wss ---> elbow method
wss <- sapply(1:14,
              function(k){kmeans(scaled, k, nstart=21,iter.max = 14 )$tot.withinss})
wss
plot(1:14, wss,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares")

### Table-comparisons
cut_av <- cutree(h1, k = 3)
```

```r
cut_compl <- cutree(h2, k = 3)
cut_ward <- cutree(h3, k = 3)

#Confront the different clustering methods
table(cut_av,cut_compl)
table(cut_av,cut_ward)
table(cut_compl,cut_ward)

#### Clusters using Ward method
cut_ward <- cutree(h3, k = 3)
cut_ward

###### Variables in relation with clusters
medie_ward_scaled<-aggregate(scale(unsupervised_dataset), list(cut_ward), mean)
medie_ward_scaled
```