

MULTIVARIATE STATISTICAL ANALYSIS

REPORT FOR MULTIVARIATE STATISTICAL ANALYSIS - PROBLEM SET 3

 **Bruno Casella**

Department of Computer Science
University of Turin
casella@di.unito.it

May 27, 2022

ABSTRACT

This report describes the results and the implementation of Problem Set 3 of Multivariate Statistical Analysis. All the R code is available at the following GitHub link: <https://github.com/CasellaJr/Problem-Set-3>. The exercise has been solved in collaboration with the PhD student Lorenzo Paletto.

Exercise 1

The data set **swiss** (part of R library datasets) gives measures of socio-economic data on 47 Swiss provinces about 1888. Switzerland was entering a period known as the "demographic transition"; i.e., its fertility was beginning to fall from high level generally found in underdeveloped countries to the lower level it has today. Together with a standardized fertility measure, variable Fertility, the data contains the following 5 indicators:

1. Agriculture % of males involved in agriculture as occupation
2. Examination % draftees receiving highest mark on army examination
3. Education % education beyond primary school for draftees
4. Catholic % "catholic" (as opposed to "protestant")
5. Infant.Mortality % live births who live less than 1 year.

The task is to cluster the observations by considering the 5 indicators only.

Before starting with the first point of the exercise, we performed an Exploratory Data Analysis. Let's see a brief summary of the statistics of the dataset.

Table 1

Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Min. :35.00	Min. : 1.20	Min. : 3.00	Min. : 1.00	Min. : 2.150	Min. :10.80
1st Qu.:64.70	1st Qu.:35.90	1st Qu.:12.00	1st Qu.: 6.00	1st Qu.: 5.195	1st Qu.:18.15
Median :70.40	Median :54.10	Median :16.00	Median : 8.00	Median : 15.140	Median :20.00
Mean :70.14	Mean :50.66	Mean :16.49	Mean :10.98	Mean : 41.144	Mean :19.94
3rd Qu.:78.45	3rd Qu.:67.65	3rd Qu.:22.00	3rd Qu.:12.00	3rd Qu.: 93.125	3rd Qu.:21.70
Max. :92.50	Max. :89.70	Max. :37.00	Max. :53.00	Max. :100.000	Max. :26.60

Let's see the first rows of the dataframe:

Table 2

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Courtelary	80.2	17.0	15	12	9.96	22.2
Delemont	83.1	45.1	6	9	84.84	22.2
Franches-Mnt	92.5	39.7	5	5	93.40	20.2
Moutier	85.8	36.5	12	7	33.77	20.3
Neuveville	76.9	43.5	17	15	5.16	20.6
Porrentruy	76.1	35.3	9	7	90.57	26.6

We standardized the data to make variables comparable.

Table 3

	Agriculture	Examination	Education	Catholic	Infant.Mortality
Courtelary	-1.4820682	-0.18668632	0.1062125	-0.7477267	0.77503669
Delemont	-0.2447942	-1.31480509	-0.2057867	1.0477479	0.77503669
Franches-Mnt	-0.4825622	-1.44015162	-0.6217858	1.2529998	0.08838778
Moutier	-0.6234617	-0.56272591	-0.4137863	-0.1768099	0.12272023
Neuveville	-0.3152440	0.06400674	0.4182118	-0.8628212	0.22571757
Porrentruy	-0.6762990	-0.93876550	-0.4137863	1.1851420	2.28566429

Before starting with the cluster analysis, it is useful to check a graphical representation of the data in order to see if the data present a grouping propensity or not. In general before choosing a clustering approach, we need to decide whether the dataset contains meaningful clusters or not, and if yes, how many clusters are there. This process is defined as clustering tendency (i.e. clusterability). So, for this reason, I am going to show some steps of clustering tendency.

First of all, let's see a graphical representation of the data:

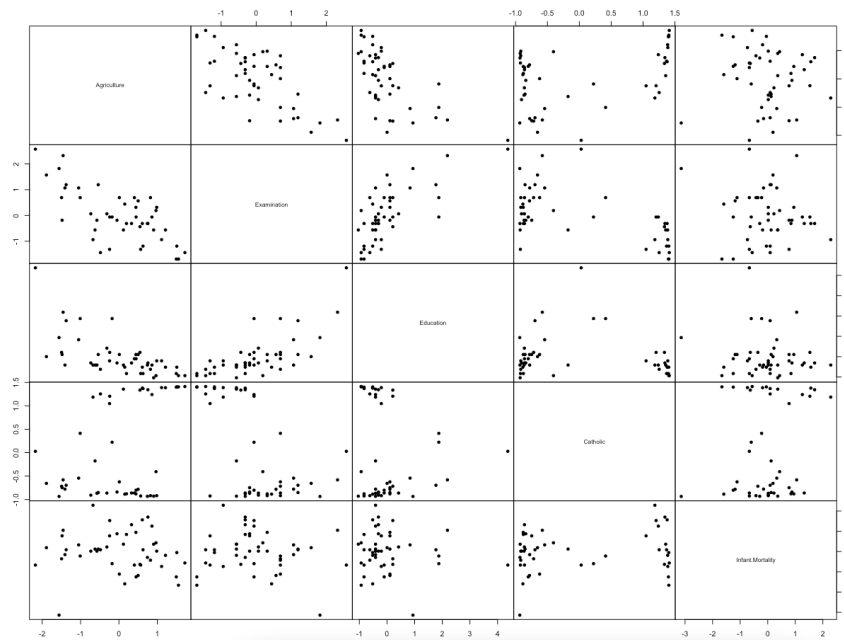


Figure 1: Graphical representation of data

We can see, from the graphical pairs representation that the data presents a grouping propensity, conversely to a random dataset generated from the swiss dataset, presented below:

```
random_df <- apply(df, 2,
  function(x){runif(length(x), min(x), (max(x)))})
random_df <- as.data.frame(random_df)
scaled.random_df = scale(random_df)
pairs(scaled.random_df, gap = 0, pch = 16)
```

Let's see a graphical representation of the data:

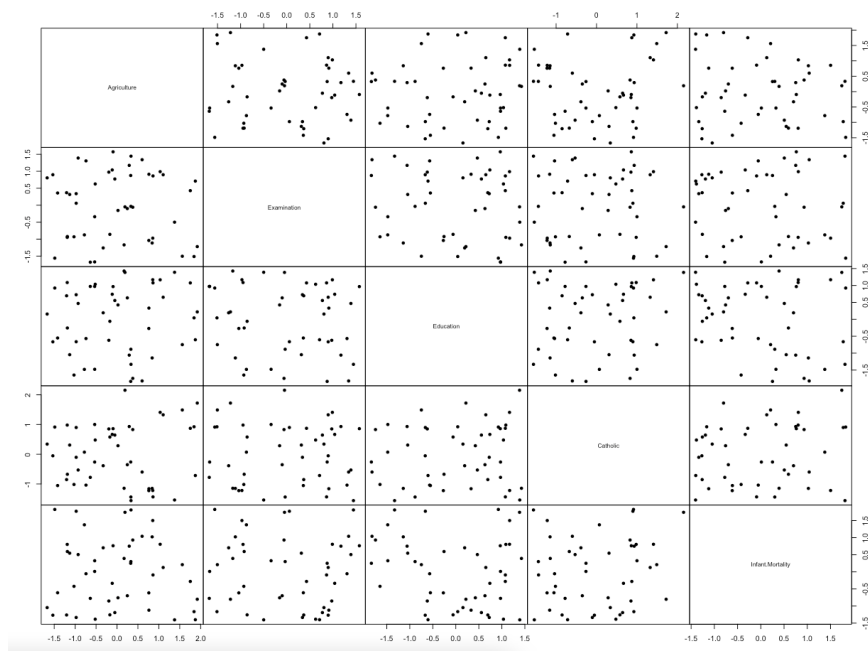


Figure 2: Graphical representation of random data generated from Swiss

Let's check in the PC space:

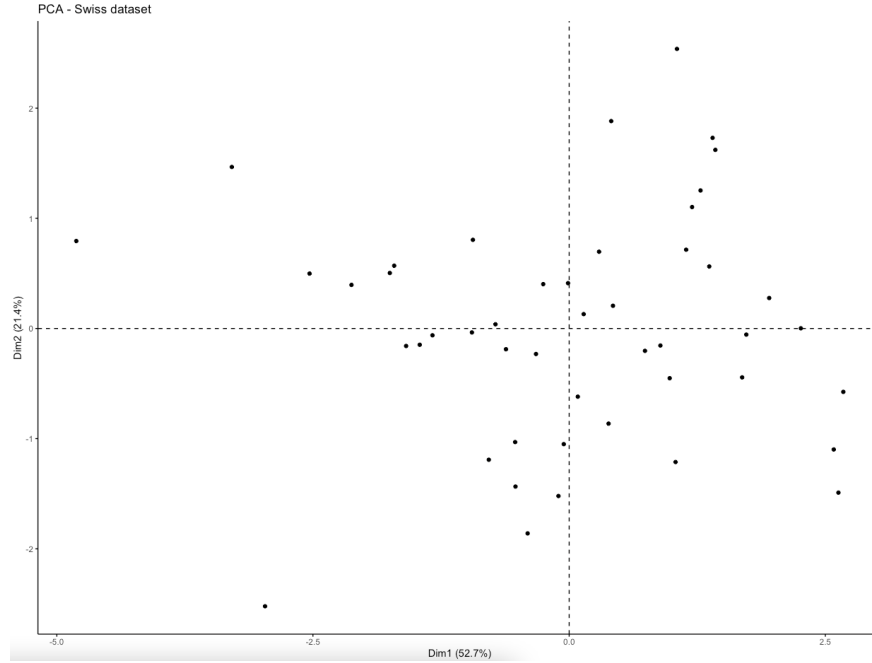


Figure 3: Graphical representation of swiss data in the PC space

and now check the random data:

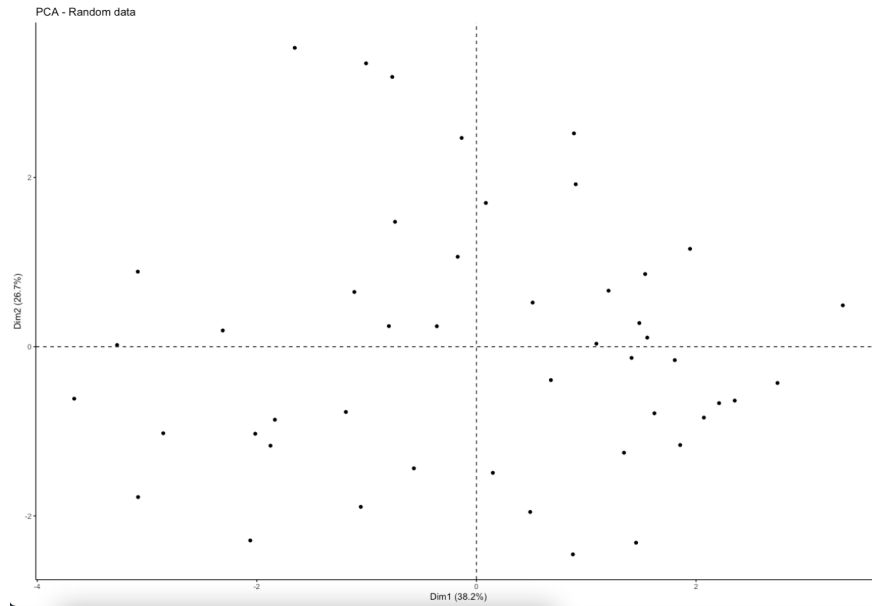


Figure 4: Graphical representation of random data in the PC space

It is not clear how many clusters contain the real data in the PC space; Instead, for sure it can be seen, in the PC space too, that the standardized randomly generated uniform data are more "isolated".

Now, for evaluating the clustering tendency, we can use a statistical method called Hopkins statistic, that takes values in $[0,1]$. The value H of this method close to 0 indicates clustered data; H close to 0.5 indicates uniformly distributed data (no meaningful clusters). For the swiss dataset, $H = 0.3165011$ while for the random dataset generated from swiss $H = 0.5237217$. So, the swiss dataset is clusterable because his H value (0.3165011) is close enough to 0. However, the random dataset is not clusterable ($H = 0.5237217$).

There is also a visual method for confirming the cluster tendency of a dataset, and it is the visual assessment of cluster tendency (VAT). It computes the dissimilarity matrix (DM) between the units in the dataset using the Euclidean distance, and it reorders it (the DM) so that similar units are close to one another. This process creates an ordered dissimilarity matrix (ODM) that is displayed as an ordered dissimilarity image (ODI), which is the visual output of the VAT algorithm. For the visual assessment of clustering tendency, we start by computing the dissimilarity matrix between observations using the function `dist()` and then using the function `fviz_dist()` to display the dissimilarity matrix.

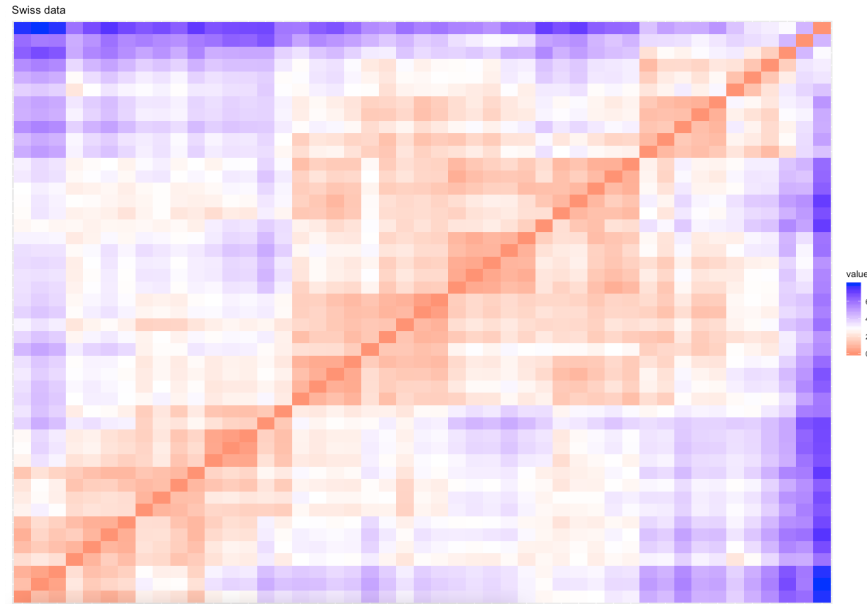


Figure 5: Visual assessment of cluster tendency of swiss

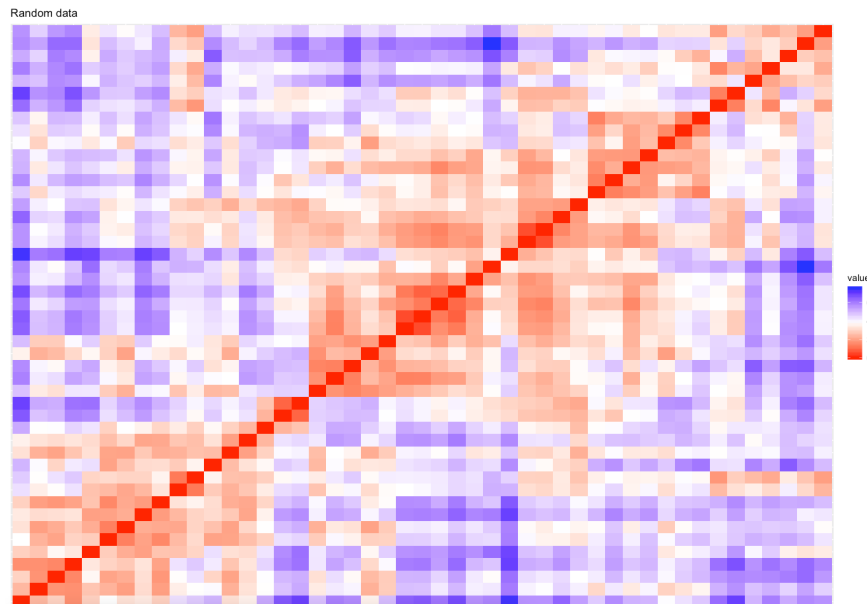


Figure 6: Visual assessment of cluster tendency of random data generated from swiss

The color level is proportional to the value of the dissimilarity between observations: red denotes high similarity (i.e. low dissimilarity); blue denotes low similarity (i.e. high dissimilarity). The dissimilarity matrix image seems to confirm that there is a cluster structure in the standardized swiss dataset but not in the random one. The VAT algorithm detects

the clustering tendency in a visual form by counting the number of a square shaped red blocks along the diagonal in a VAT image.

Now we can start with the first point of the exercise.

1) Obtain a 3 clusters solution via hierarchical clustering with Ward's method. Plot the observations in the space of the first two principal components with colors determined by cluster memberships. Add the projected cluster centroids. Give an interpretation to the clusters. What is the cluster assigned to province "V. De Geneve"?

The first clustering algorithm applied is the agglomerative hierarchical clustering, that is a “bottom-up” approach: each observation starts in its own cluster (leaf), and pairs of cluster are merged as one moves up the hierarchy. This process goes on until there is just one single big cluster (root). In this case the leaves are 47 (because the dataset contains 47 observations), and for every step, the algorithm merges pairs of cluster with the lowest dissimilarity according to a different given linkage method, and then it moves up the hierarchy, until there is just one single big cluster, with all the observations within, by building a sort of tree diagram called dendrogram. We use the Ward (minimum deviance) linkage method. Ward's linkage methods assume that a cluster is represented by its centroid (cluster center). So it doesn't use a unit, but a point, and for this reason it is feasible for numeric dataset only; Ward's method minimizes iteratively the sum of the squared Euclidean distances of units within a cluster for every cluster (WSS – Within deviance), and maximizes the sum of the squared average distances of units between centroids (BSS – Between deviance). So, the Ward's method accepts as input only a Euclidean distance matrix. There are two Ward's methods: the difference between the methods Ward.D and Ward.D2 is the distance matrix to be given as an input to hclust(). In Ward.D it is squared, in Ward.D2 it is not squared. Below the dendrograms obtained with both methods: Ward.D:

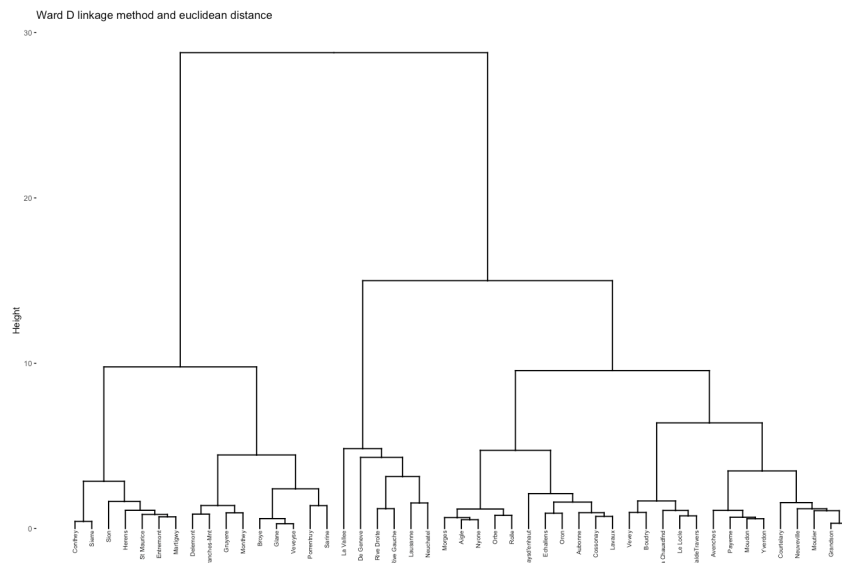


Figure 7: Dendrogram obtained with the Ward's method

and Ward.D2:

Here is the plot of the observations in the space of the first two principal components with colors determined by cluster memberships.

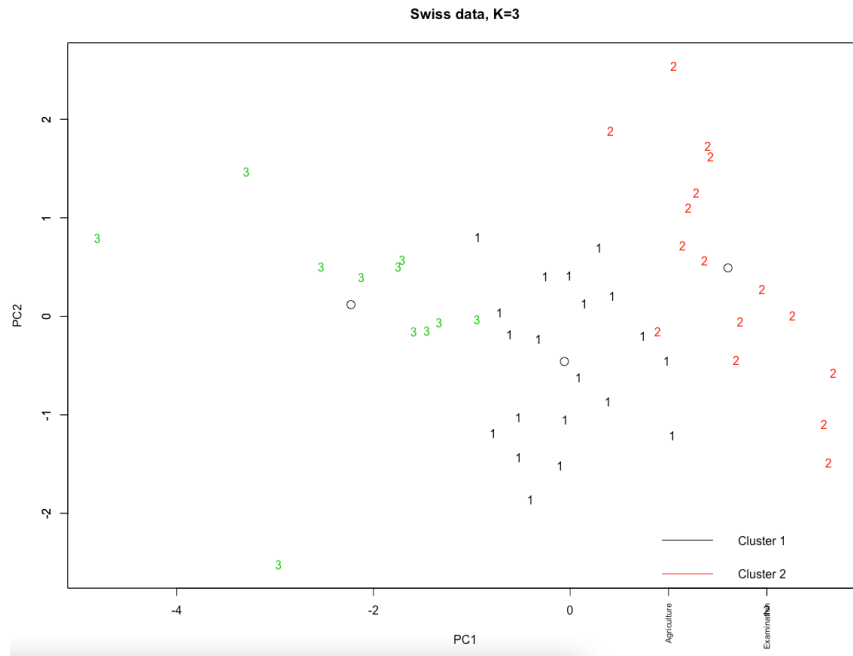


Figure 11: Plot in the space of the first 2 PCs

The first two components account for about 73% of the data variability:

Importance of components:					
	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.6228	1.0355	0.9033	0.55928	0.40675
Proportion of Variance	0.5267	0.2145	0.1632	0.06256	0.03309
Cumulative Proportion	0.5267	0.7411	0.9043	0.96691	1.00000

2) Find a 3 clusters solution with k-means using the centroids found at previous point as initial choice of seeds. Plot the observations in the space of the first two principal components as done in the previous point and compare the k-means cluster solution with the hierarchical one.

The plot in the space of the first 2 PCs is shown below:

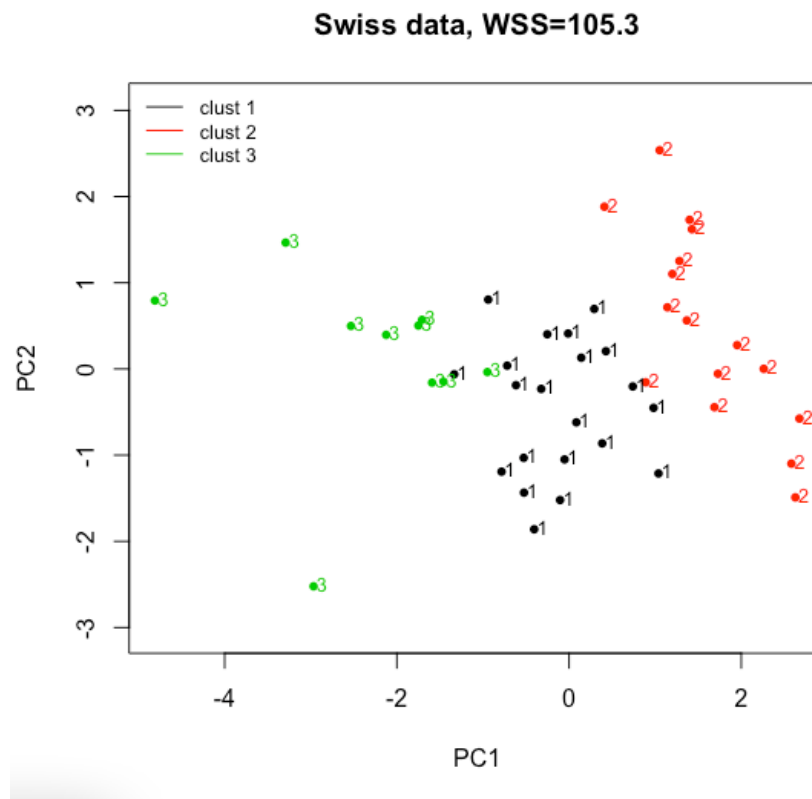


Figure 12: Plot in the space of the first 2 PCs when using KMeans

Basically the 2 clustering methods (Ward.D2 and K-Means) present exactly the same results except for the observation in the middle of the picture (which is Boudry): with Ward it is clustered as Cluster 3, while with K-Means is clustered as Cluster 1. All the others are equal.

3) Compute the average Fertility in the three clusters found via k-means. What is the cluster with the highest fertility? Can you make sense of this finding?

These are the mean values of Fertility for each cluster:

Cluster 1: 68.89

Cluster 2: 80.55

Cluster 3: 56.12

So, the cluster with the highest fertility is cluster 2, with 80.55. Let's try to understand why.

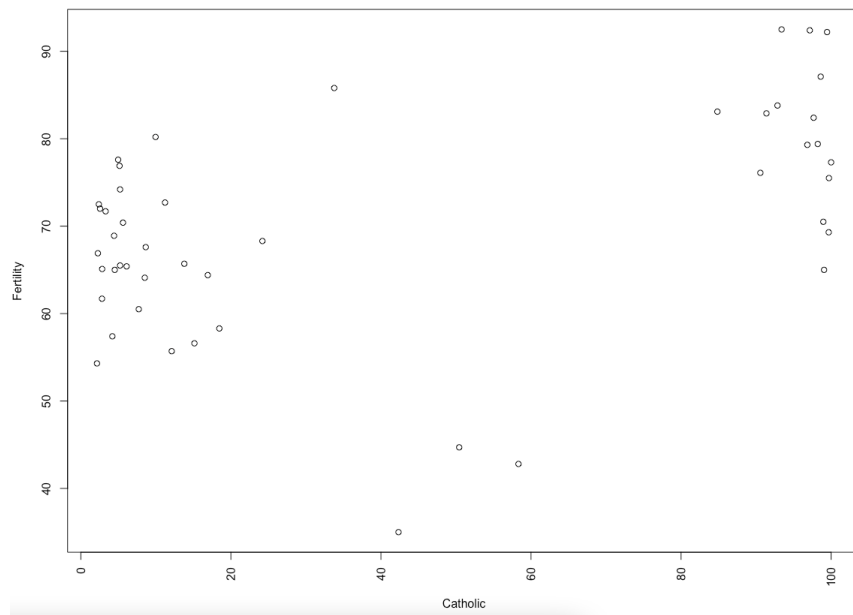


Figure 13: Scatterplot Fertility vs Catholic

From this picture we can see a strong correlation: the higher degree of catholic the higher fertility. Looking at the mean values we can see that cluster 2 contains the highest percentage of catholic people:

Cluster 1: 7.67

Cluster 2: 96.15

Cluster 3: 23.44

For this reason in cluster 2 we have the highest fertility.

4) Consider now model-based clustering with Gaussian mixtures. Use R library Mclust to select the best model with 3 groups by BIC method. What model is chosen? Discuss.

This is the plot of the model-based clustering with Gaussian mixtures.

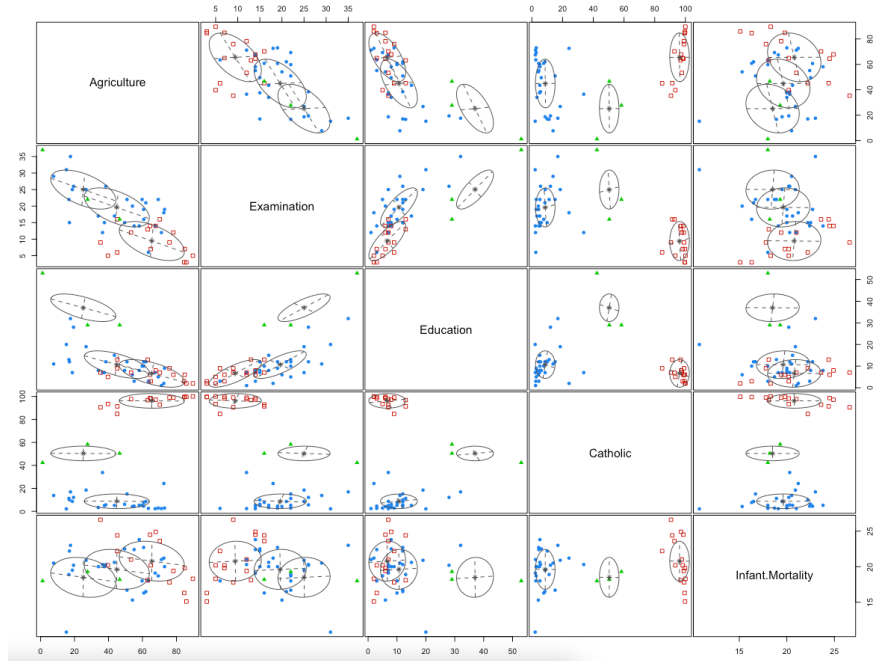


Figure 14: Plot of model-based Clustering with Gaussian mixtures

The best model is EEE: ellipsoidal, equal volume, shape and orientation. This model fits a separate mean vector for each class, but the same ellipsoidal covariance matrix, which is essentially equivalent to linear discriminant analysis.

5) Compare the k-means cluster solution with the model based one. Total freedom in this can you make sense of the differences?

With k-Means:

```
km_label3      1  2  3
1 21  0  0
2  0 16  0
3  7  0  3
```

With model-based solution:

```
hc_labels      1  2  3
1 20  0  0
2  0 16  0
3  8  0  3
```

```
fit$classification
Courtelary      Delemont Franches-Mnt      Moutier      Neuveville      Porrentruy      Broye      Glane
1              2          2              1              1              2          2
Sarine          Veveyse      Aigle      Aubonne      Avenches      Cossonay      Echallens      Grandson
2              2          1              1              1              1          1
La Vallee       Lavaux      Morges      Moudon      Nyone      Orbe      Oron      Payerne
1              1          1              1              1              1          1
Rolle          Vevey      Yverdon      Conthey      Entremont      Herens      Martigwy      Monthey
1              1          1          2              2              2          2
Sierre          Sion      Boudry      La Chauxdfnd      Le Locle      Neuchatel      Val de Ruz      ValdeTravers
2              2          1              1              1              1          1
Rive Droite    Rive Gauche
3              3
```

Basically, with model-based clustering, cluster 3 contains only 3 observations: Rive Droite, Rive Gauche and V. De Geneve, that are all the provinces belonging to Canton Geneve.