Data Mining Project

Submission date : January 12, 2018

YAO ZELIANG* & *ZHAO HE* & *PAULINE LOR

Project goal

The goal of this data mining project is to work on a real data set. The main idea of the project is to implement one or more methods studied during the courses on a proposed data set. Our group will be working on the “Pays” data set.

Report content

0.Introduction

1.The purpose of the analysis

2.Data description

3.The data analysis

4.PCA analysis

0.Introduction

The goal of the project is to analyse the datasets and make some comments on the results.

1.The purpose of the analysis

The purpose of this analysis is to apply the most appropriate method to analyse, reduce and process the dataset.

2.Data description

The individuals are the countries (European countries) on which we are making the analysis. Each country is represented in one row.

Each variable will be represented in one column.

The description of the variables:

- **esp life F**: average number of years lived by a girl born in 2001 if the female mortality by age remained the same as in 2001

- **Mort_inf**: number of children <1 year old dead in 2001 / number of children born alive in 2001

- **Activ F**: number of women in employment / number of women of working age

-**% chom**: (number of unemployed / number of workers aged over 15) * 100

- **Pnb / hb**: annual gross national product per capita (expressed in $)

-**% education**: education expenditure (public or private) as% of Pnb

-**% health**: health expenditure (public or private) as% of Pnb

Once we import the data in R studio, we can turn the original txt.file to the dataset below:

```
> data<-read.table("D:/BigData/Datamining/Binome7/pays.txt",encoding = "UTF-8")
```

	pays	esp_vie_F	Mort_inf	Activ_inf	%chom	Pnb/hb	%education	%sante
1	Allamagne	74.8	4.4	48.8	8.2	26768	4.3	10.6
2	Autriche	75.4	4.8	49.0	4.1	29075	4.9	8.0
3	Belgique	75.1	5.0	42.3	7.3	27952	5.8	8.8
4	Chypre	75.3	5.6	50.9	3.8	12724	5.8	6.0
5	Danemark	74.5	5.3	73.8	4.5	30096	8.1	8.4
6	Espagne	75.6	3.9	40.3	11.4	22538	5.6	7.7
7	Estonie	64.9	8.4	52.2	6.8	10201	6.8	5.5
8	Finlande	74.6	3.8	56.8	9.1	27215	5.6	6.6
9	France	75.5	4.6	47.8	8.7	27560	5.6	9.4
10	Greece	75.4	6.1	37.6	10.3	17670	2.3	9.2
11	Hongrie	68.4	9.2	45.6	8.4	12733	5.2	5.7
12	Irlande	73.0	5.9	47.5	4.4	32549	4.5	7.2
13	Italie	76.7	4.5	36.0	9.1	26946	4.6	8.0
14	Lettonie	64.5	10.4	49.7	8.5	7809	6.2	4.8
15	Lituanie	65.9	8.6	54.6	10.9	8359	5.2	5.7
16	Luxembourg	75.2	5.8	42.5	2.4	50410	4.0	6.1
17	Malte	76.2	6.0	22.9	7.4	9875	5.7	8.9
18	Norvege	76.2	3.8	69.2	3.9	37070	7.4	8.5
19	Pays-Bas	75.5	5.1	52.9	2.7	29614	5.2	8.2
20	Pologne	70.3	8.1	49.5	18.1	9852	5.1	4.2
21	Portugal	72.4	5.5	54.1	5.0	18500	5.5	8.2
22	Royaume-Uni	75.0	5.6	53.0	5.1	26756	4.7	7.3
23	Slovaquie	69.7	8.6	52.9	17.4	12314	4.3	6.4
24	Slovénie	72.3	4.9	51.3	11.3	17762	5.2	8.2
25	Suède	77.5	3.4	76.2	4.9	26849	7.3	7.9
26	Suisse	77.2	4.9	58.8	3.1	30058	5.1	10.7
27	Tch<U+FFFD>quie	72.2	4.1	51.3	9.8	15011	4.6	7.4

Type of variables of dataset: str(data)

```
> str(data)
'data.frame':   27 obs. of  8 variables:
 $ pays      : chr  "Allamagne" "Autriche" "Belgique" "Chypre" ...
 $ esp_vie_F : num  74.8 75.4 75.1 75.3 74.5 75.6 64.9 74.6 75.5 75.4 ...
 $ Mort_inf  : num  4.4 4.8 5 5.6 5.3 3.9 8.4 3.8 4.6 6.1 ...
 $ Activ_inf : num  48.8 49 42.3 50.9 73.8 40.3 52.2 56.8 47.8 37.6 ...
 $ %chom     : num  8.2 4.1 7.3 3.8 4.5 11.4 6.8 9.1 8.7 10.3 ...
 $ Pnb/hb    : int  26768 29075 27952 12724 30096 22538 10201 27215 27560 17670 ...
 $ %education: num  4.3 4.9 5.8 5.8 8.1 5.6 6.8 5.6 5.6 2.3 ...
 $ %sante    : num  10.6 8 8.8 6 8.4 7.7 5.5 6.6 9.4 9.2 ...
```

From the result, we can see that 'pays ' has the type "Character", "Pnb/hb" has the type "int"and all the other variable's type are "numeric".

Type of variables of dataset: attributes(data)

```
> attributes(data)
$names
[1] "pays"       "esp_vie_F"  "Mort_inf"   "Activ_inf"  "%chom"      "Pnb/hb"
[7] "%education" "%sante"

$row.names
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27

$class
[1] "data.frame"
```

Function attributes() access our data's attributes, we can get the names of variables and rows, and the type of the data class.

Univariate analysis (position criteria, dispersion criteria) : summary(data)

```
> summary(data)
     pays             esp_vie_F        Mort_inf         Activ_inf         %chom
 Length:27          Min.   :64.50   Min.   : 3.400   Min.   :22.90   Min.   : 2.400
 Class :character   1st Qu.:72.25   1st Qu.: 4.550   1st Qu.:46.55   1st Qu.: 4.450
 Mode  :character   Median :75.00   Median : 5.300   Median :50.90   Median : 7.400
                    Mean   :73.31   Mean   : 5.789   Mean   :50.65   Mean   : 7.652
                    3rd Qu.:75.50   3rd Qu.: 6.050   3rd Qu.:53.55   3rd Qu.: 9.450
                    Max.   :77.50   Max.   :10.400   Max.   :76.20   Max.   :18.100
     Pnb/hb        %education        %sante
 Min.   : 7809   Min.   :2.300   Min.   : 4.200
 1st Qu.:12728   1st Qu.:4.650   1st Qu.: 6.250
 Median :26756   Median :5.200   Median : 7.900
 Mean   :22380   Mean   :5.356   Mean   : 7.541
 3rd Qu.:28514   3rd Qu.:5.750   3rd Qu.: 8.450
 Max.   :50410   Max.   :8.100   Max.   :10.700
```

The summary function gives all the statistical results of the data according to each variable, which are : the minimum value, 1st quartile, median, mean, 3rd quartile, the maximum value.

Variance list (dispersion criteria): apply(data,2,var)

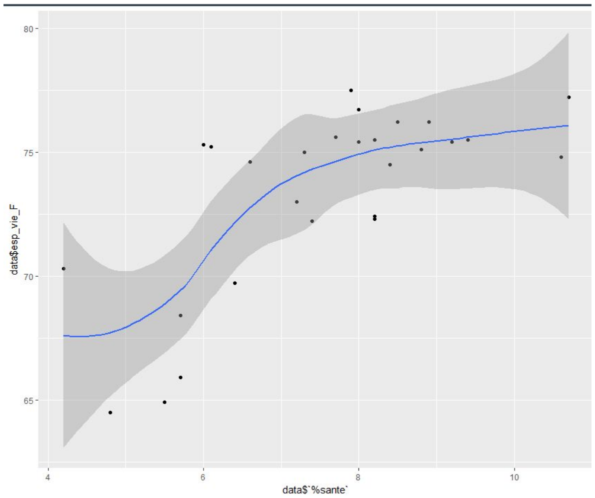
```
> apply(data,2,var)
        pays    esp_vie_F     Mort_inf    Activ_inf        %chom       Pnb/hb   %education       %sante
          NA 1.369148e+01 3.465641e+00 1.205272e+02 1.605721e+01 1.058494e+08 1.351795e+00 2.652507e+00
```

Bivariate analysis (covariance matrix) : cov(data)

```
              esp_vie_F    Mort_inf    Activ_inf     %chom       Pnb/hb    %education      %sante
esp_vie_F   1.000000000 -0.86208619 -0.003531218 -0.4116394  0.6511476638 -0.0628593657  0.701675316
Mort_inf   -0.862086189  1.00000000 -0.176850877  0.3782614 -0.6166925865 -0.0862421524 -0.674585139
Activ_inf  -0.003531218 -0.17685088  1.000000000 -0.2425572  0.2178180313  0.5999525892 -0.009578663
%chom      -0.411639449  0.37826143 -0.242557163  1.0000000 -0.5932732219 -0.2757933732 -0.340795849
Pnb/hb      0.651147664 -0.61669259  0.217818031 -0.5932732  1.0000000000  0.0006702915  0.423179685
%education -0.062859366 -0.08624215  0.599952589 -0.2757934  0.0006702915  1.0000000000 -0.086752869
%sante      0.701675316 -0.67458514 -0.009578663 -0.3407958  0.4231796846 -0.0867528687  1.000000000
```

The top positively relative pairs is **%sante / esp_vie-F: 0.702** and the most negatively relative pair is **esp_vie_F/Mort_inf : -0.862**

Scatter plot (ggplot2)



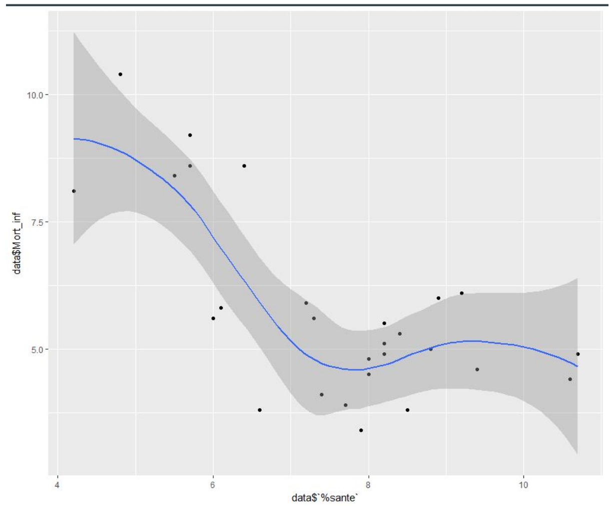


Figure 1

Figure 2

From these results of figure 1, we can see that there is a strong correlation between health expenditure and life expectancy. We can deduce that higher health expenditure is , longer girl life expectancy is.

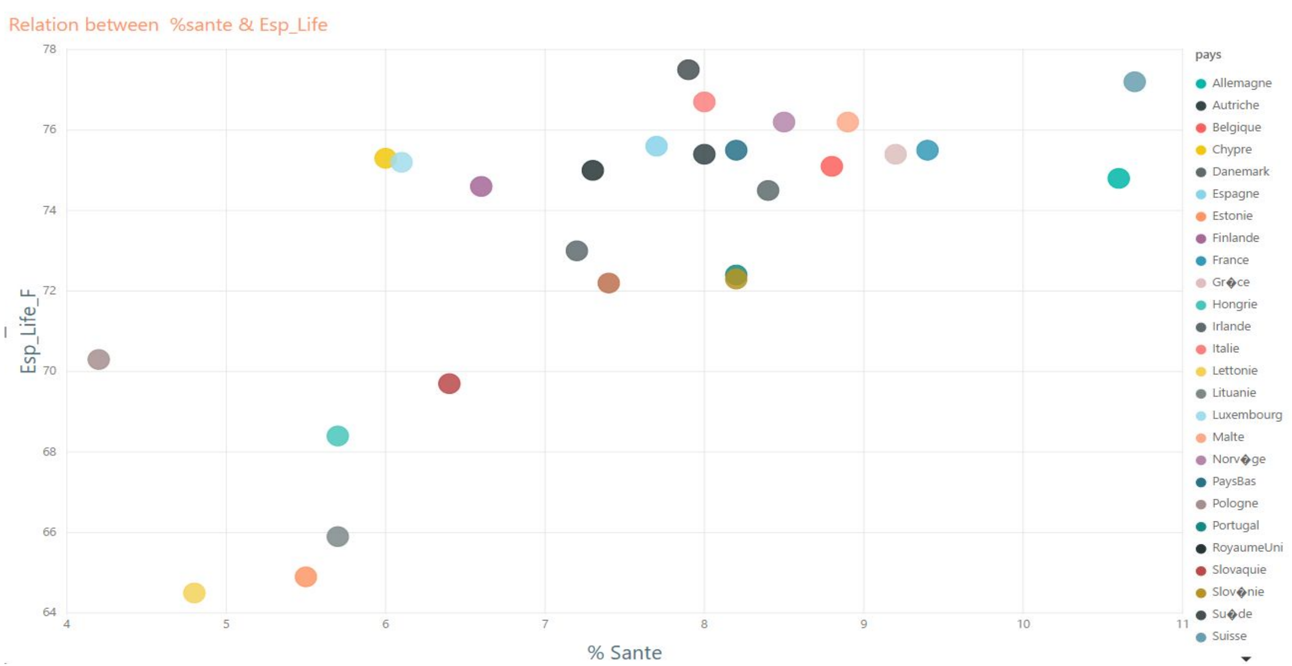
The interpretation is : If more money are spent on the health expenditure, girls born in 2001 will live longer.

On the other hand in figure 2, the correlation is very low between girl life expectancy and childhood death. We can deduce that these two variables have an inverse relationship.

The interpretation is : When more money is spent on childhood cares, girls born in 2001 will have a longer life expectancy.

Some Visualizations :

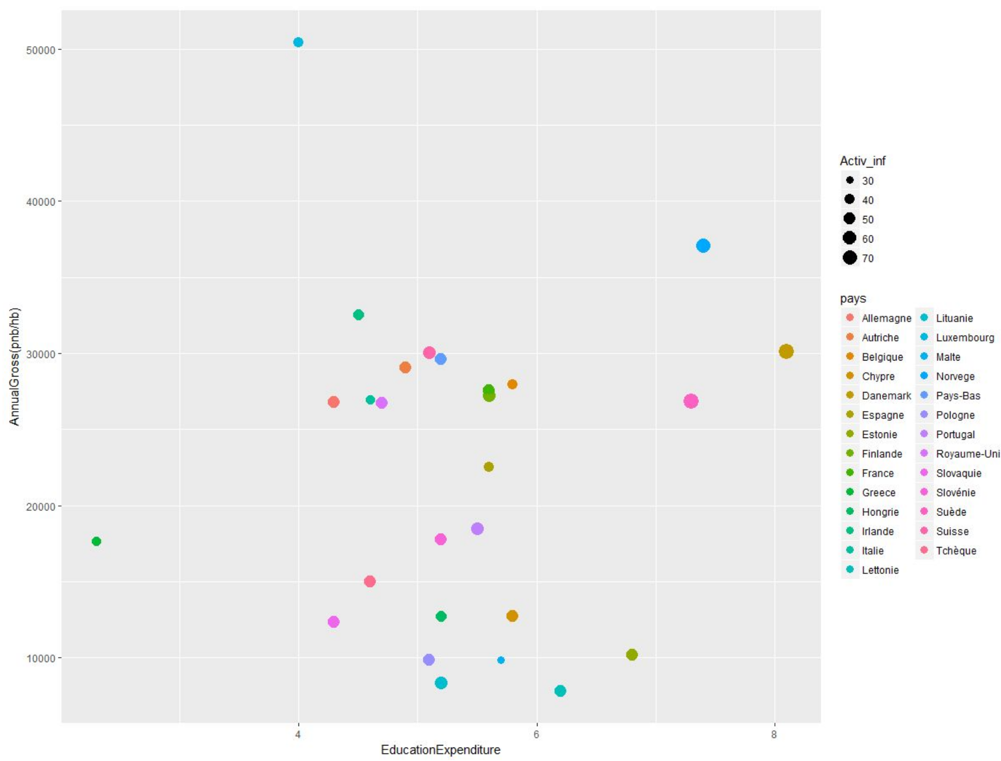
Conclusion: As it can be seen from the TreeMap here (**2 dimension graph**), ***Slovaquie*** has the highest unemployment rate and ***Luxembourg*** has the lowest one. We also analyze the other indexes(Esp_Life_F,Mort_inf...).



From this scatter plot, we can visualize for each country, the women life expectancy depending on the health expenditure per country.

We can see that the more money countries spend on health expenditure, the longer their life expectancy is. Here, Switzerland is the first one in this domain.

```
> qplot(data$`%education`,data$`Pnb/hb`,colour=pays,size=Activ_inf)+geom_point(size
  =3)+labs(x="EducationExpenditure",y="AnnualGross(pnb/hb)")
```



From this scatter plot, we have a **four- dimension** graph. We can view annual gross national product per capita in function of education expenditure for each country as well as the rate of number of active women depending on the size of the points.

1st description and interpretation: the more money is spent on Education expenditure, higher is annual gross national product per capita. ***Norway*** performs the best here, but we can notice that ***Luxembourg*** spends less on Education expenditure but gets the highest annual gross. We can interpret this phenomenon with the fact that Luxembourg citizens are already high educated and the country does not need to spend more on education.

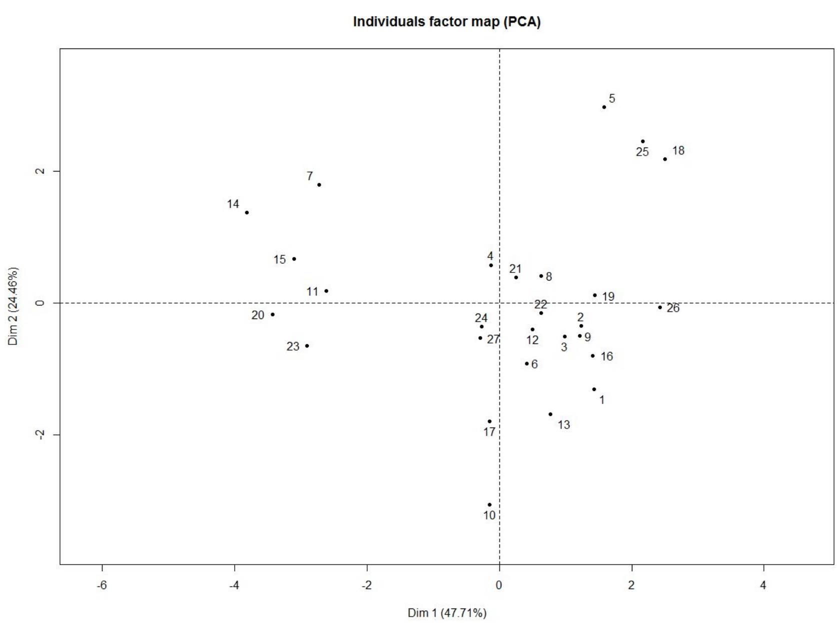
2nd description and interpretation : from the size of the points, we can isolate three countries with the biggest number of active women : ***Denmark, Norway and Switzerland.***

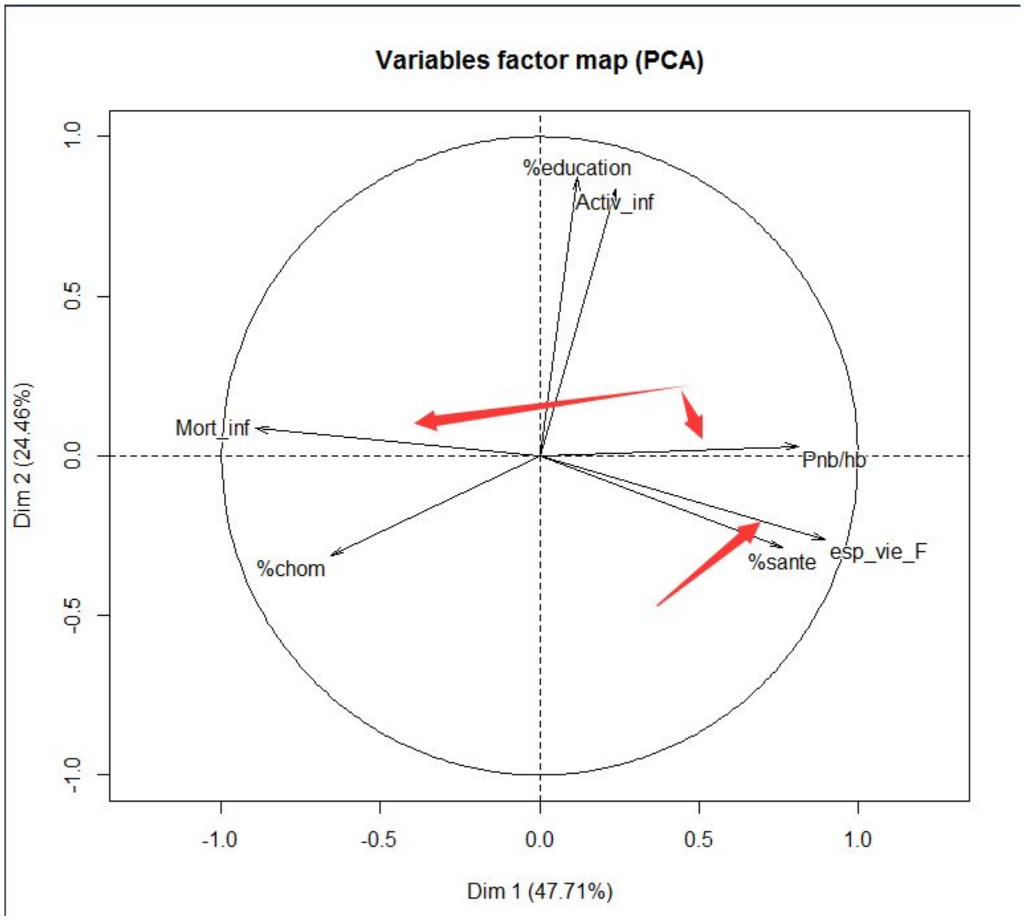
These three countries also spend the more money on Education expenditure.

4.The PCA analysis

PCA is a classic unsupervised method in data mining for linear dimension reduction and principal components analysis. In our project, we can use PCA to find out the most important axes which contain most of the information in the data set in order to represent the most significant features. (Here we use package ***FactoMineR*** for analysis).

```
> library("FactoMineR")
> result<-PCA(data[,-1])
> result$eig
       eigenvalue percentage of variance cumulative percentage of variance
comp 1  3.3396243              47.708919                          47.70892
comp 2  1.7119496              24.456423                          72.16534
comp 3  0.7392525              10.560750                          82.72609
comp 4  0.5187857               7.411225                          90.13732
comp 5  0.3762541               5.375059                          95.51237
comp 6  0.2019273               2.884676                          98.39705
comp 7  0.1122065               1.602950                         100.00000
```





From the variables factor map, esp_vie seems to be very highly correlated to %sante, from the correlation matrix we got before we know that their correlation value is 0.702 (top positively relative pairs). The same way for Mort_inf and Pnb/hb (top negatively relative pairs).

As a result, we consider that the implementation of PCA by 2 axes in this case is accurate. The graph matches perfectly with our analysis.