ANALYSIS OF INEQUALITY FACTORS BETWEEN COUNTRIES

Lorena Recio Cabeza

8 de junio de 2018

INTRODUCTION

Due to we live in a world with a lot of inequality among countries that form it, it's interesting to see what factors are causing this inequality. They can be political, economic, social or health factors. The main goal of this projects is to analyse these factors and determine which ones are the most influential.

The richness of a country can be measured by the "Gross National Income" since it corresponds to the sum of the remuneration of all the national production factors. Therefore, for the entire project it will be the dependent variable that I will use ("GNIPPP" or "Gross National Income per capital" in *"WHO"* dataset).

The main tools chosen for this task are **R**, **Python** and **Tableau Public**. Some programming and statistical knowledge are needed in order to completely understand these pages, and a personal computer capable of running the programs listed above.

Python will be developed in Jupyter notebook and R in R Studio. To be able to work from Windows on Jupyter notebook, you can download the graphical user interface **Anaconda Navigator** that allows you to launch applications like Jupyter.

R and R Studio you can download it for free in the next web pages:

- **R**: Link to R website
- **R Studio**: Link to download R Studio

Tableau Public is used for data visualizations, so I put the link of the dashboard and you only need a internet navigator.

The packages that I've used in the project and you need to install are the following, depending on whether I work with R or Python.

Python packages:

- `pandas`: It's a package providing fast, flexible, and expressive data structures designed to make working with relational or labeled data both easy and intuitive.
- `numpy`: This package is used for the treatment of vectors and matrix
- `matplotly`: It's used for generating plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc.
- `shutil`: It's a specialized module to perform operations with files and folders (move, copy, paste, etc.).

R packages:

- `ggplot`: it/'s used for creating graphics or data visualizations.
- `dplyr`: this package is used to manipulate the data files in R in a simple way.

DATASET DESCRIPTION

I have worked with *"WHO"* Public Dataset which you can download in the next website **"Data world - WHO dataset"** or load it from DATASETS directory.

"WHO" dataset is in semicolon CSV format and it contains 202 countries in rows and 358 different features of this countries.

I've only chosen features which have more than 50% of data, those that not fulfil this condition I've dropped them because they won't give an accurate information. In the analysis, I've focused on the variables with more than 50% of correlation with "GNIPPP", whether positive or negative.

METHODOLOGY

In the notebooks are explain all the steps that I've followed to be able to reproduce the project. Here I explain the main process and the location where you can find the files.

Firstly I've cleaned the dataset seeing what to do with NA values, and select the most interesting features to analyse, I've done this first part in a Jupyter notebook with Python. You can find this notebook as Data_Cleaning.ipynb save in DATA CLEANING directory.

Once cleaned I've worked with *"whocomplete"* dataset for the first part of the analysis and *"whodfR"* for the second part (both of them are save in DATASETS directory).

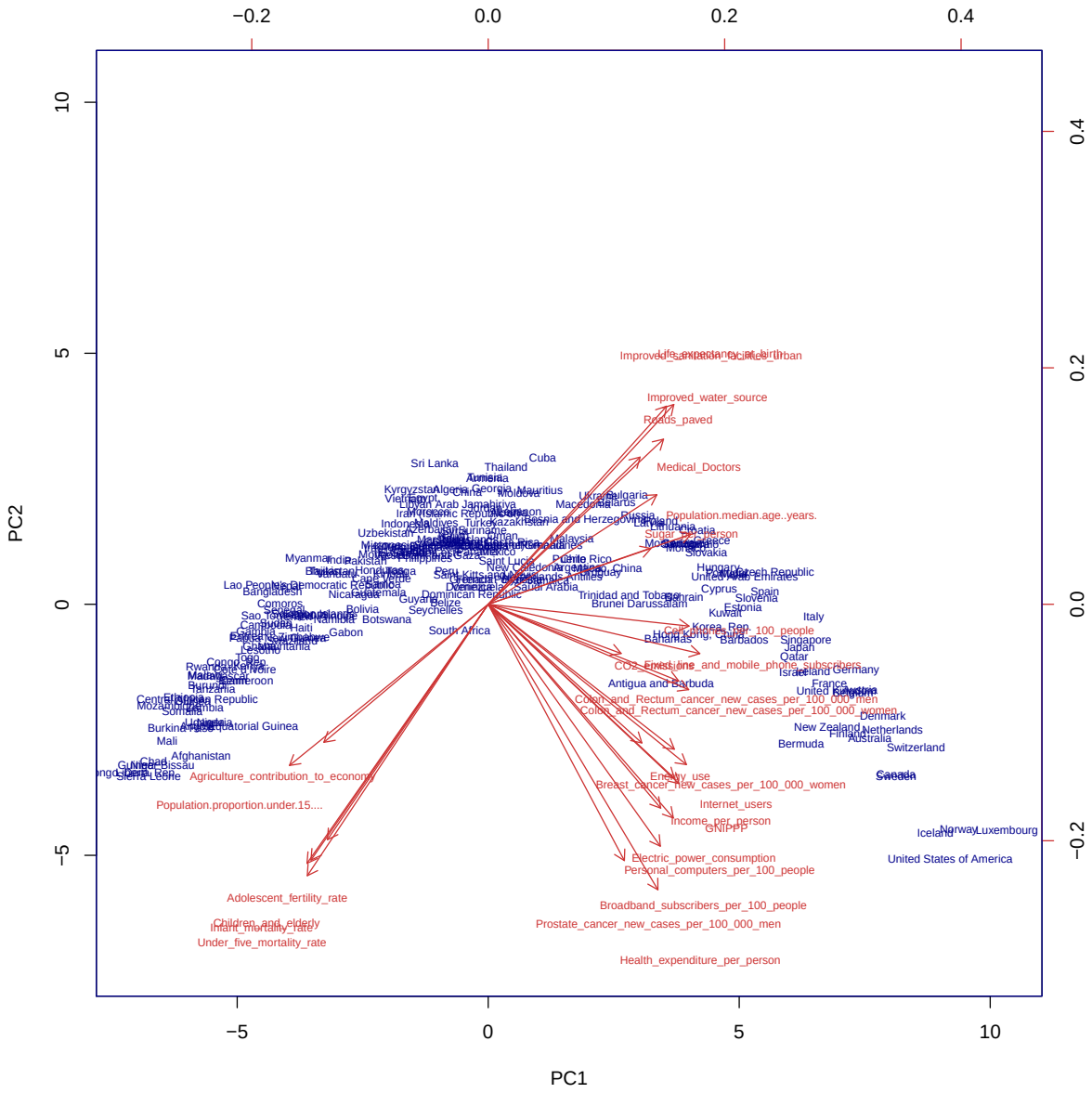
The first part of the data analysis it's done with Python. I've tried to see the correlation between the different features with the correlation matrix and, as I've said above, select which ones that have more than 50% of correlation with "GNIPPP".
The second part it has been worked with R Studio. Here I've used a main components analysis (PCA) to reduce the dimensionality but retaining the original information. Then I've performed some different linear regression models by OLS (Ordinary Least Squares) to see what features explain to a greater extend the changes in "GNIPPP".

The notebooks are save in DATA ANALYSIS directory as 1.Data_analysis.ipynb (Jupiter notebook), 2.Data Analysis.R (R project) and 2.Analysis.R (R script). To work with R, firstly you have to open the R project file, and inside of R project, you can open R script.

RESEARCH FINDING

The main research findings are two:

1. In the PCA analysis we can see in the `biplot()` graph, that the main component, which have the highest weight of variance, have more positive correlation with "GNIPPP", "Population median age","Fixed line and mobile phone subscribers", "Internet users", "Cell phones per 100 people" and "Colon and rectum cancer in men and woman". So their positive values could be assimilated with those countries that stand out for their numbers of users using tecnologies, a high GNIPPP and median age population, and their have many cases of Colon and rectum cancer. Here we can see in the graph:



2. The linear multiple regression model called *"modelSingnificant"* show that the most influent features in the inequality between countries are health expenditure, income per person, populations of median age, personal computer, broadband subscribers, agriculture and electric power consumption.
   The result of the linear multiple regression model, as we can see, show that all the variables are significant and the whole model with a p.value: <2.2e-16, also it's significant. Furthermore, the adjust R-squared is 0.8484, so it's the best model.

```
##
## Call:
## lm(formula = GNIPPP ~ Health_expenditure_per_person + Income_per_person +
##     Population.median.age..years. + Children_and_elderly + Personal_computers_per_100_people +
##     Electric_power_consumption + Broadband_subscribers_per_100_people +
##     Agriculture_contribution_to_economy, data = who)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -28955.1  -1346.1   -218.2   1226.2  26611.8
##
## Coefficients:
##                                     Estimate Std. Error t value
## (Intercept)                       -8.542e+03  4.176e+03  -2.045
## Health_expenditure_per_person      3.154e+00  5.726e-01   5.507
## Income_per_person                  3.289e-01  4.962e-02   6.628
## Population.median.age..years.      3.078e+02  8.755e+01   3.515
## Children_and_elderly               8.085e+01  3.583e+01   2.256
## Personal_computers_per_100_people  1.582e+02  4.150e+01   3.811
## Electric_power_consumption         3.182e-01  1.332e-01   2.389
## Broadband_subscribers_per_100_people -4.549e+02  1.089e+02  -4.178
## Agriculture_contribution_to_economy  -1.042e+02  3.319e+01  -3.141
##                                     Pr(>|t|)
## (Intercept)                         0.042173 *
## Health_expenditure_per_person       1.15e-07 ***
## Income_per_person                   3.32e-10 ***
## Population.median.age..years.       0.000547 ***
## Children_and_elderly                0.025175 *
## Personal_computers_per_100_people   0.000186 ***
## Electric_power_consumption          0.017861 *
## Broadband_subscribers_per_100_people 4.46e-05 ***
## Agriculture_contribution_to_economy  0.001951 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4653 on 193 degrees of freedom
## Multiple R-squared:  0.8544, Adjusted R-squared:  0.8484
## F-statistic: 141.6 on 8 and 193 DF,  p-value: < 2.2e-16
```

DESCRIPTION OF FRONTEND

In the first Dashboard, I've showed a world map that change the color depends on the "GNIPPP". The countries that have a high "GNIPPP" are in red and, in the other hand, the countries that have the lowest "GNIPPP" are in green. In the following link you can see the graph:

DASHBOARD 1: GNIPPP BY CONTINENT

The second Dashboard show the difference between the median of health expenditure and income per person in the different continents. Both caracteristics have the same structure; in Europe and in North America is where they are higher. Also in this dashboard, we can see the positive relation between health expenditure and incomes per person features. The countries in this graph are in different colors and if you put your mousse in the circles, you can see what country is each other. The following link show the dashboard:

DASHBOARD 2: COMPARATIONS HEALTH EXPENDITURE AND INCOME BY CONTINENT

BIBLIOGRAPHY

- https://stackoverflow.com/
- https://cran.r-project.org/web/packages
- http://rpubs.com/
- https://es.wikipedia.org/
- https://rstudio-pubs-static.s3.amazonaws.com/287787_1c53df3fcf6b432dbc775a91cb2090ce.html#pcr:_pca_aplicado_a_regresi%C3%B3n_lineal