Exploratory Analysis Of the Gapminder Data.

Emmanuel Victor Barbosa Sampaio
Department of Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Brazil
emmanuelsampaio@alu.ufc.br

Carlos Alfredo Cordeiro de Vasconcelos Filho
Department of Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Brazil
carlosalfredo@alu.ufc.br

***Abstract*—In this article, we will analyze a data set containing information on Per Capita GDP, population size and life expectancy for different countries in different years. To perform the analysis, we used univariate, bivariate and multivariate analysis techniques. Based on the result of the analysis of these variables, we want to find similarities and differences of the continents.**

***Index Terms*—Gapminder, Data Analysis**

I. Introduction

Human development around the world happens in a different pace and way on each different regions. In a economical perspective, there are some information that can be used to measure the development of countries. Based on these information it is possible to identify the difference between these countries and understand their development, this is the context that is insert the database from Foundation Gapminder. Gapminder is a project, created in Sweden, that has as focus explore and develop new ways of explaining important global trends and proportions to make them easier to understand.

Given the context, this work has as goal verify, based on the analysis of the data from Gapminder datasets, the difference between the continents development. In order to reach this objective, we had access to datasets that contain data about life expectation, Per Capita GDP and population size for different countries in different years. To perform the analysis, we verify the distribution, relationship and also the combination of these available variables for each continent.

The outline of this paper is organized as follows: Section II explains the concepts that were used in the analysis, Section III shows the analysis of the data set variables, Section IV presents the analysis of the variables in order to understand the continent differences. We conclude this work in Section V with a summary of the results of this paper.

II. Analysis Background

Before introduce the mathematical background that going to be used in the analysis, we need to divide the analyzed data into quantitative variables and qualitative variables, because the analysis going to have a deep focus on the quantitative ones. Based on [2], the qualitative variables can assume values that fit into categories and the quantitative variables can assume continues range of values. In our case, we assume the country name, the year of the observation and the continent as qualitative variables. The quantitative variable are the per capita GDP, the life expectation and the population size.

For the quantitative variables we can create a description of their values based on their distribution, the process of analyze the distribution is called univariate analysis. Using the univariate analysis it is possible to identify how the quantitative variables are distributed on each continent, and these can create a initial insight about the importance of a variable in order to differentiate the continents.

As just one variable maybe not be enough to reach our goal, we can use two variables, in a approach called bivariate analysis, here we verify the relationship between the variables. To finish, we can also perform a multivariate analysis, that uses more than just two variables, in this approach we going to use a dimensionality reduction techniques in order to summarize the variables in lower dimension. The next subsections introduce the mathematical concepts that were used in this paper in order to perform univariate, bivariate and multivariate analysis of the available variables.

A. Univariate analysis

The univariate analysis performed in this work has a focus on verify how the data is distributed related to the mean, that is how the data is distributed related to the average of the values of the variable. In this sense we going to calculate two metrics, the skewness and standard deviation.

The skewness is a metric that evaluate the symmetry of the data distribution related to the mean. A positive skewness indicates that the distribution is concentrated in values below the mean and a negative indicates that the majority of the data is greater than the mean. We can calculate the skewness for a variable $x \in \mathbb{R}^n$ that has a mean μ_x using Equation 1:

$$\text{skewness} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)^3 (n-1)^{1/2}}{(\sum_{i}^{n}(x_i - \mu_x)^2)^{3/2}} \tag{1}$$

The standard deviation, measures how spread is the data related to the mean of the variable, low values of standard deviation tells that the data is close to the mean, high value of standard deviation tells the opposite. To calculate the standard deviation of a variable x with mean μ_x the following equation is used:

$$\sigma = \sqrt{\frac{\sum_{i}^{n}(x_i - \mu_x)^2}{(n-1)}} \tag{2}$$

In order to visualize the univariate analysis we going to use two different visualizations: histograms and box plots. The

histograms shows the amount of values of the variable in a determine group of ranges inside the numerical interval of values of the variable. The box plot gives us a way to split the data into intervals and verify how they are distributed related to the quartiles, which are values that divide these intervals. In this visualization, 25% of the data is lower than the first quartile, 50% of the data is lower than the second quartile, also called median, and 75% of the data is lower than the third quartile.

B. Scaling, Centering and Skewness reduction

In order to improve the numerical stability of the data, better visualize the linear relation between predictors as well as making the use of the Principal Component Analysis more efficient when the scale of the predictors is different, we performed some manipulations with the data that are: Centering, scaling and skewness reduction.

Centering a predictor variable consists in subtracting the mean of the predictor from all the values. As a result of this manipulation the mean of the transformed data is zero. Scaling a predictor variable consists in dividing each value of the predictor by its standard deviation. This makes so that the transformed data's standard deviation becomes one.

To reduce skewness the data must be replaced by a logarithm or exponential transformation. In this paper, the transformation used to reduce skewness will be from the Box and Cox family [4], that is represented by the following equation:

$$x* = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x) & \text{if } \lambda = 0 \end{cases} \tag{3}$$

From scikit-learn documentation and from [4], λ is a transformation parameter that can be defined manually or calculated using a maximum likehood estimator in order to find optimal parameter for stabilizing variance and minimizing skewness.

C. Bivariate Analysis

In this approach, we verify the relation between two variables, with this we can understand if they have some relationship. In our work domain, we perform these analysis using a graphical and a statistical approach.

The graphical approach is the scatter plot, which is a visualization of the data based on the values of two variables. With this graphical approach we can create a initial insight on a empirical relationship between the variables, in a sense that we create a understating if the variables increasing or decreasing in the same time. Also, using the scatter plot we can use colors based on the categorical variables in order to identify the location of the classes based on two variables.

The statistical approach is the calculus of the correlation coefficient [3], this coefficient verify the linear relationship between two variables. To calculate the coefficient between two variables x, y we going to use Pearson Correlation Coefficient, that is showed bellow:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}} \tag{4}$$

This coefficient can assume values between -1 and 1, where negative values indicate that when one variable increase the other decrease and a positive values indicates that one variable increase the other increase. The magnitude of the coefficient indicates how linear is the relation, which means that a 0 value indicates that there is no linear relation between the variables.

D. Principal Component Analysis

The Principal Component Analysis or PCA is a method that consist in creating linear combinations of predictors which aim to capture the most possible variance. This allow the use of a smaller set of predictors that still contain the the majority of the information of the original variables.

Based on [4] and [1] lets introduce the calculus of the principal components. Suppose a data set that will be analysed by PCA is formed by i observations described by j quantitative variables. It going to be represented by a matrix $X \in \mathbb{R}^{i \times j}$. This matrix has a rank L, that corresponds to the minimum value between i and j. In order to perform the PCA, the columns of X, that represent each variable data, will have their values centered and scaled to avoid distortions by the distribution and scale of the measures. After centering and scaling the columns of X, the matrix will be decomposed using singular value decomposition:

$$X = P\Delta Q^T \tag{5}$$

Where P is the $i \times L$ matrix of left singular vectors, Q is the $j \times L$ matrix of right singular vectors, and Δ is the diagonal matrix of singular values. Furthermore, we will call F the matrix of factor scores and define it by the following equation:

$$F = P\Delta \tag{6}$$

Multiplying the equation above for Q^TQ we obtain:

$$FQ^TQ = P\Delta Q^TQ = XQ \tag{7}$$

Because Q is a orthogonal matrix, we can rewrite the equation as:

$$F = XQ \tag{8}$$

Where F is the new dataset after the application of PCA and Q is a $J \times L$ matrix where each individual column represents the coefficients of a component.

Because of the definition of right singular vectors, Q is a matrix formed by putting all the J eigenvectors of X^TX in it's lines. To find witch of the columns of Q represents the most adequate coefficients of the first component we must calculate the inertia of all the columns of F.

$$\gamma_j^2 = \sum_{i=1}^{I}(f_{i,j})^2 \tag{9}$$

Where γ_j is the inertia of the column j. After that, we must find the column j_M with the most inertia. Having found j_M, the equation that defines the first component is:

$$PC_1 = q_{1,j_M} J_1 + q_{2,j_M} J_2 + q_{3,j_M} J_3 + \cdots + q_{J,j_M} J_J \quad (10)$$

Where: J_i is the i-th variable and $q_{a,b}$ is the element of Q in the row a and column b.

To obtain the subsequent components suffices to select the next column with the most inertia.

III. Introduction to the Data Sets

In this work, we going to analyze a data set that contains 1704 observations of 142 different countries from 5 different continents in different years. In order to visualize the observed continents we plotted two bar plots, they are presented in Fig.1. In the left side, the number of observations per continent, which means the total of times that a country for that continent appears in the dataset. In the right side there is the number of countries observed per continent. Based on the bar plots, we can identify that there are few observations from Oceania, looking deeply they are Australia and New Zealand, and huge amount of observations from Africa.

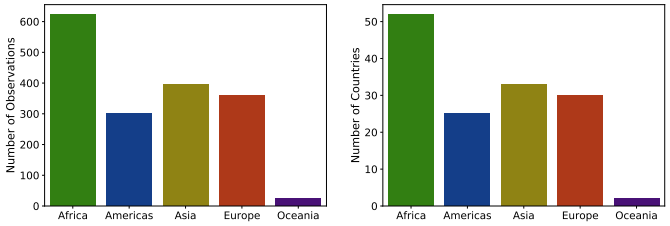

Fig. 1. The bar plot showing the number of observations and countries per continent.

A. Analysis Of The Quantitative Variables

We started the analysis by looking to the Per Capita GDP. This variable has a positive skewness, equal to 3.85. This indicates that the majority of the data is situated before the mean, which has a value of 7215.32 dollars. In this that set, few observations carries high values of Per Capita GDP.

For the population size, the skewness has a positive value, 2.1, which indicates that we have few observations with values greater than the mean, which is 29.60 million people. The standard deviation has value of 106 million, this show that the population is distributed between the observation and due to the skewness and the standard deviation, there are few countries we high values of population size.

To visualize both population size and Per Capita GDP distribution it was used histograms, they are present in Fig.2. But, the skewness and the scale of the variables difficulties the interpretation, since the count for high values of these variables are almost not possible see as well as a detail of the distribution on the low values. Based on this, in order to improve the visualization we used a logarithm of base 10, because it going to reduce the scale and distance between the points, do to the fact that logarithm is a strictly increasing function does not affect the order of the values. The histograms in logarithm scale are present in Fig.3.

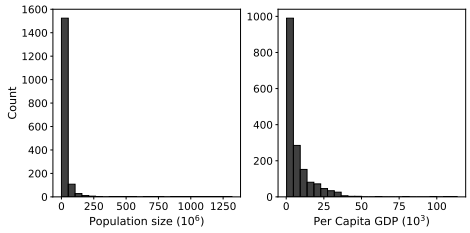

Fig. 2. Per capita GDP and population size histograms from the sample .

In the logarithm scale histograms it is possible to see, for the population size distribution, that few countries have population in a scale greater 10^8, the majority have values around 10^7 and lower than this scale. it also indicates that some countries have observed population size lower than 10^4. For the Per capita GDP it is possible to verify the number of observations with low values seams to be greater than the observation with high values, and also there are few observations with a scale of 10^5 and $10^{2.5}$.

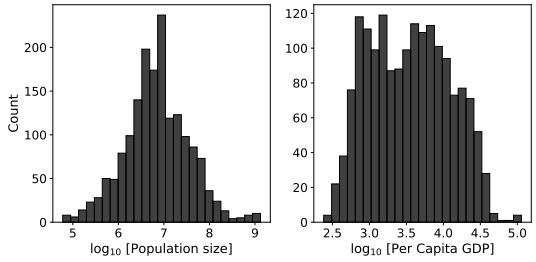

Fig. 3. The Per Capita GDP and Population Size using logarithm scale.

The last analyzed variable is the Life Expectation, here we can verify a negative skewness of -0.25 and a mean of 59 years, that indicates that the majority of the observations have a values of life expectation greater than 59 and few observations that have low values of life expectation These information are possible to visualize in the histogram in Fig.4.

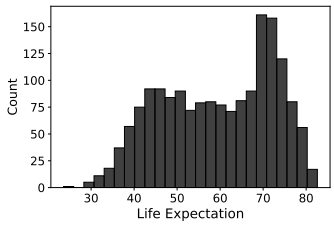

Fig. 4. Life Expectation Distribution. We can verify the negative skewness, that indicates the majority of the observations have values greater than the mean.

B. Quantitative Variable Distribution per Year

This data set contains information from 12 different years, which were collected every five years, starting in 1952 and with 2007 as the last year observed. Each country of the data

set was observed once a year. In this context 5 shows the distribution per year of Per Capita GDP and Population size in different years.

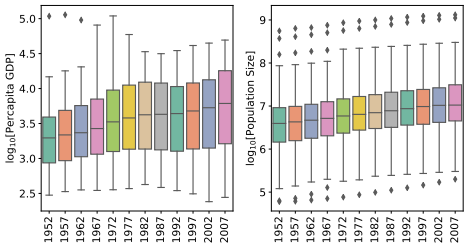

Fig. 5. The box plot of the logarithm in based 10 of Per capita GDP and Population size for different years.

For the Per Capita GDP, that with the advance in time that more countries are having values greater than the mean, but is also possible to visualize in the last observed years there still existing countries with Per Capita GDP as lower as the lower values of these variable in the first measured years. It is also possible to see that occurs a reduction of the maximum value of Per Capita GDP observed on the years and a reduction of the minimum value observed between 1982 and 2002.

For the population we can mentioned the existence of countries with huge population size, these countries are China and India, they are population in billion of people scale. The lower outlier in the years is São Tome and Principe a small island in Africa. The other countries are situated between 10^6 and 10^7 as we can also see from the histogram in Fig.3.

We also verified how the life expectation change in time. Fig.6 shows that the median of the life expectation is increasing in time, it reach a value near to 72 years in 2007. In 1952, the median was 45 years. It is important to say that despite the consistent increase in the median of life expectation, the first quartil actually decreased from 1992 to 2002, witch shows that the life expectation got worse in the worst off countries in that time frame. Another curiosity that we can observe is an outlier in the lower direction in 1992. This outlier has a life expectation of 23.6 and is related to Rhuand. In this country, 1992 was the second year of a civil war that was marked by genocide of determined population group, the war happened between 1990 and 1994.

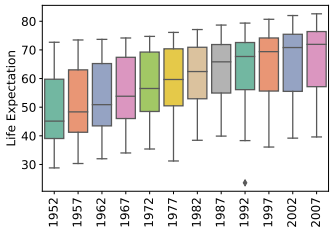

Fig. 6. The box plot of the life expectation for different years.

IV. Continent Contrast

To show the differences between the distributions for each continent, we created plots that color the histograms based on the continent analyzed. The histograms are diagonal from Fig.7. In these histograms, to improve visualization, we use a Gaussian Kernel Density Estimation function, which is the curve over the histogram. These estimates seek to model the probability distribution that generated the variable data.

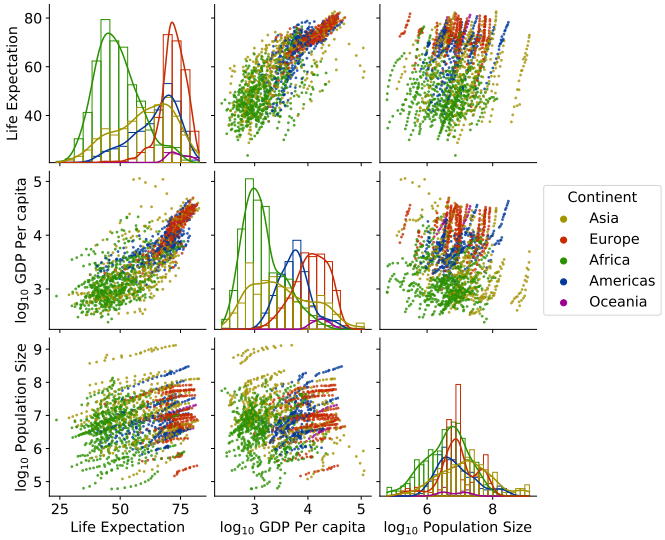

Fig. 7. The relation between the quantitative variables. The colors indicate the continent of the observation. In the diagonal there is histograms showing their distribution of the quantitative variable for each continent.

We start the analysis by looking to histogram related to the logarithm of population size, that is present in the lower corner of 7. It is possible to see that the continents are to be concentrate around the scale of 10^7, but it is possible to see that Africa and Europe has a lower spreading, compare to Asia and America. It is also possible to see calculating the standard deviation of the original variable. For America the value of standard deviation is 56 million, for Europe is 20 million, for Asia is 200 million and for Africa is 15 million. This high value of spreading for Asia observations a reflex of this continent has low number countries as we see from 1, countries with low population and two countries that have the greatest value of population in the whole sample, Asia and India.

Now, for the life expectation, is possible to verify that in the histogram, there is a distinguish distribution between Africa and Europe, where Africa assume low values and Europe Assume high values. In the other continents, America has observations near to the European ones, and Asia are more spread, with values close to African ones and other close to the European ones. The Ocenia countries, that are Australia and New Zealand are near to the Europe data. In a numerical perspective for the life expectation, African has the lowest mean, 48 years, and a positive skewness and a standard deviation equal to 9.1, that shows the majority of the African countries have values near and lower to the mean. In other hand Europe and America have negative skewness, where Europe has a mean equal to 71.9 and America has a mean equal to 64.65 and a standard deviation of 5.4 for Europe

and 9.3 for America. Asia has the biggest standard deviation, with a value of 11.8, with a mean close to 60 and a negative skewness of -0.4. To finish, Oceania has as mean 74.3, and a standard deviation of 3.8 and a positive skewness. These values show that the life expectation has some similarities around the world for example Oceania and Europe, also some observations from America are near to Europe. In other hand, Africa has the lowest values of life expectation while Asia is the continent that has more spread of these value, having observations near to the lowest values of the dataset and others near to the high values.

For the Per Capita GDP, it is possible to verify that Africa again assume the lowest values, Europe is situated in the highest values side. America is distributed near to Europe and Asia. And Asia has the greatest spreading. Oceania again is situated near to the European countries. In a numerical perspective, Africa has a mean equal to 2193.75 dollars a positive skewness equal to 3.5, and a standard deviation of 2827.92, while Europe has a mean equal to 14469.47 dollars a skewness equal to 0.85 and standard deviation of 9355.21 dollars. America has as mean 7136.11 dollars, a positive skewness equal to 2.8 with a standard deviation of 6396.76. To finish, Asia has a mean 7902.15043 a positive skewness of 4.1 and a standard deviation of 14045.37. Given the numerical results and graphical, we can verify that as life expectation, Per Capita GDP from Africa and Europe are opposite to each other, Oceania has values near to Europe, and America has values near to Europe and Asia. And this last has the biggest spreading.

The result from the last two paragraphs, shows the importance of life expectation and Per Capita GDP to differentiate the continents. To complement this analysis we verify the relationship between the variables plotting scatter plots for each pair of variables and also calculating the correlation coefficient 4 for each pair. From the scatters of Fig.7, it is possible to verify that high values of life expectation are associated with high values of Per Capita GDP and the population size is not so related to the the other variables. Fig.8, shows that there exists a positive correlation between life expectation and Per Capita GDP, that is increased for the logarithm scale.

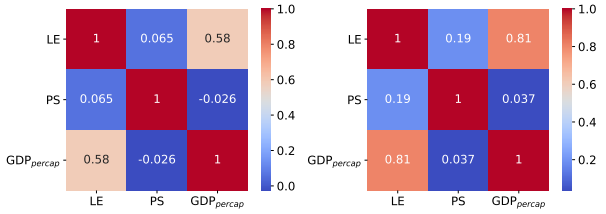

Fig. 8. Correlation Matrix. In the left side, considering the raw data, in the right side considering the transformation of Per Capita GDP (GDPpercap) and population size (PS). The colors are associated with the value of the Pearson Coefficient.

A. *Comparing The continents contrast in different years*

Here we going to verify he distribution of the quantitative variables in the first and the last year of the data set, and going to use color in order to show the data from different countries.

For both plots, the values below the first quartile are predominant from African countries and above the third quartile are predominant from European countries for both years. In this context, from the 52 countries of Africa that were evaluated in 2007, 30 have a logarithm of their per capita GDP lower than the first quartile that was 1624.84 dollars and 35 have a life expectation lower than the 57.16 that was the first quartile of the data. For Europe, from 30 evaluated countries, 20 have Per Capita GDP greater than third quartile, 9 have values between median and third quartile. This values shows that the majority of the countries of these two mentioned continents have a clear distinction on Per Capita GDP and life expectation.

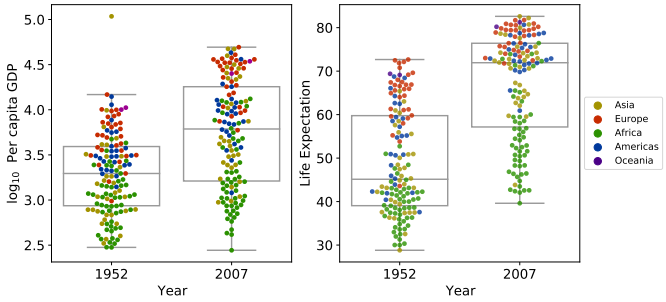

Fig. 9. The life expectation and the logarithm of Per Capita GDP distributed for the first and last year of the data set. The colors are based on the continent.

To complete the analysis we verified the scatter plot where the marker size is regulated by the population size of the country, and the color by the continent of the country. From the generated scatter Fig.10, it is possible to verify the advance of Per Capita GDP and life expectation in time and also the increasing of the population, wich shows a inclusion of more people living in a country with a better Per Capita GDP and better life expectation.

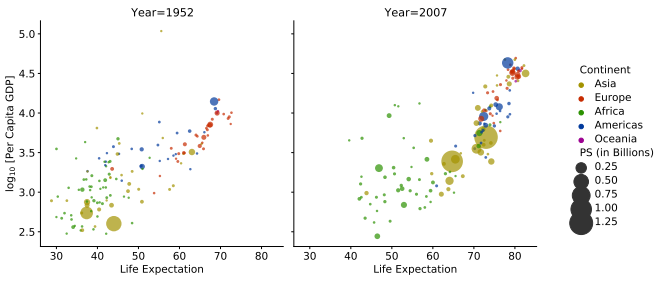

Fig. 10. The scatter plots for Per Capita GDP and life expectation, where the colors are adjust by the continent and the marker size by the population size.

B. *Applying Principal Component Analysis*

The last step of the analysis is the multivariate analysis. Here, as already mentioned, we used the PCA in order to

reduce the data dimension and try to perform a visualization. In our context, we have 3 variables, in different scales, that were not centered and have skewness so we perform the preprocess techniques introduced in the Subsection II-B.

After the preprocess, we perform the calculus of the principal components, the inertia associate with each components and the coefficients that multiply the variables values in the linear combination for each component combination are present in the Fig.11. Based on this figure, the component associated with the combination of Per Capita GDP and life expectation has the greatest inertia. The second component with a great inertia has is basically the population size. Since these two components correspond to almost 90% of all inertia, we can use them to perform the visualizations, since they are the components that summarize better the data, the scatter from these two components is present in Fig.12.

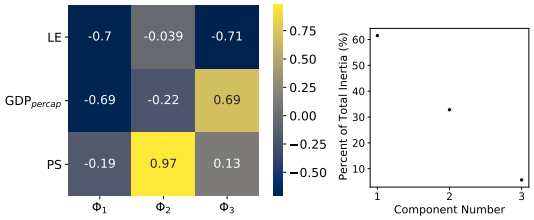

Fig. 11. In the left side a plot showing the coefficients associated to each principal component. In the right side, the inertia associated to each principal component.

In Fig.12 it is possible to verify in the related to the PC_1 that there is a division between the observations in one side there are many African countries few Asians and few Americans countries. in the other side European, Oceania and few Asians and American countries. This can be related to the fact that this component is a combination of life expectation and Per Capita GDP, and as we saw in 7, there is a distinction between the distribution of these two variables in the continents. For PC_2 it is possible to verify that the majority of the observations are situated around the same values, despite few observations that are in the top, and are related to the Asian countries that in Fig.10, are showing to be the outliers in population size, China and India.

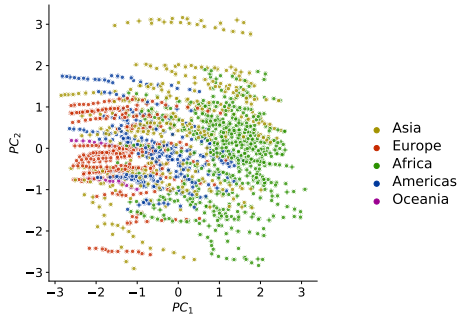

Fig. 12. Scatter plot between the two principal components.

To verify the difference in time, we plotted the same scatter of Fig.12 but now filtering the data from 1952 and 2007. The result is presented in Fig.13 it is possible to verify, that the observed countries, change their position over PC_1, which indicates the advance of life expectation and Per Capita GDP in time. But this advance was not equality distributed between regions since European countries has more data in one side of the graph compare to the observation of Africa for example. It is also possible to see that Asian countries modify their position, 1952 they are situated in one side, and in 2007 many of the are moving to the other side of PC_1 axis.

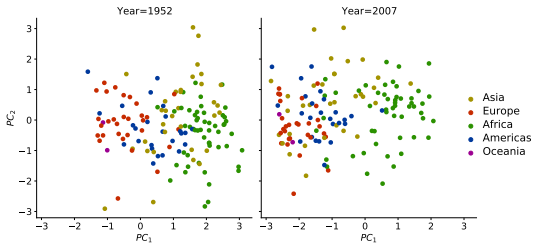

Fig. 13. Scatter plot between the two principal components in different years. It is possible to verify the distribution modification in the position in the PC_1 axis. It is possible to verify that over this exists there exists two mine clusters of countries one composed in by more European and the other by more African countries.

V. Conclusions

It is possible to conclude, that there exist differences between the continents on the available variables, and this difference can was better verified when we look to the Per Capita GDP and life expectation and the combination of this variables. In this context, there are continents in which the majority of their countries have big values of life expectation and high Per Capita GDP, such as Europe. And other that the majority of their countries have low life expectation and low Per Capita GDP, such as Africa. In another hand, that are countries with a diversity in the observation such as Asia, that at the same time have countries with high Per Capita GDP and High Life Expectation and others with low per Capita GDP and low expectation. It also possible to conclude, that the development of the Per Capita GDP and life expectation in time occurs in a different manner for the continents, since, when we compare the distribution of the first analyzed and the last analyzed year we have verified that some countries does not change to much their original values of these metrics, compare to others that have modify.

References

[1] Hervé Abdi and Lynne J. Williams. “Principal component analysis”. In: *WIREs Computational Statistics* 2.4 (2010), pp. 433–459.

[2] Gareth James et al. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014.

[3] Joseph K. Blitzstein and Jessica Hwang. *Introduction to Probability Second Edition*. 2019.

[4] Max Kuhn and Kjell Johnson. *Applied predictive modeling*. New York, NY: Springer, 2013.