# Data Visualization to Explore the Countries Dataset for Pattern Creation

https://doi.org/10.3991/ijoe.v17i13.20167

Shakir Khan
Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia
sgkhan@imamu.edu.sa

**Abstract—**Data visualization is graph representation of data. It produces interactive graphs that explain the relationships among the data to viewers of the graph. The aim of data visualization is to communicate data value clearly and effectively through graphs [1]. Here we take the advantage of data visualization to explore the countries dataset to provide a holistic and interpretive view about the world. In addition to examine some hypotheses about gross domestic product (GDP), literacy and more of the countries effects on different factors showing on the dataset such as the literacy and the migration.

**Keywords**—data visualization, countries data, world dataset, regression analysis

## 1 Introduction

Data visualization is an easy way to understand data for humans. It is an attractive way to interpret huge and overlapping numbers and make it easy to identify patterns and trends. It is the way of human brain works. This research aims to represent information about countries and when we talk about countries there will be a hundreds number of countries in different regions with multiple data normal people cannot understand it by just reading especially when they want to make some compressions between different countries. Countries data such as population, distribution, their contributions, birthrate, death rate, and so on, huge data is needed to represent and to analysis the relationships between different aspects such as the relationship between gross domestic product (GDP) and literacy and many more. This paper will focus to visualize the dataset about countries around the world in addition to examine different hypotheses for different relationships between those attributes. This paper is organized as follows section 2 gives a brief overview of the literatures about analyzing data of countries. The methodology in section 3 is divided into three stage data collection where the dataset downloaded from kaggle [2] then data preprocessing and data visualization by using python programming language. Furthermore, we assume that there are relationships between many attributes and those are tested depending on visualizing them. Section 4 includes a discussion of the findings of the visualization. Finally, the conclusion gives a brief summary of the key results.

## 2 Literature review

A considerable amount of literature has been studied data about countries. Those study analyzing countries data which explore many factors in different areas such as health, economy and so on. They also test hypotheses about the effect of many factors on countries growing and more. Here we will assume different factors in our dataset affect or affected by GDP which is gross domestic product calculated in dollar per capita. Many literatures have been study different factors such as population, area, transparency ranking, youth unemployment rate, transparency score, education period and type of government. Recent study by Ilter(2017)[3] exams the effect of eleven social and economic factors on GDP using regression analysis and found that population, GDP, transparency score and compulsory education effect on GDP. Zhang, et al. (2016) [4] study the relationship between city size and GDP and found there is no relationship between them. Jain, et al.(2015)[5] studied the effecting of many factors which are Foreign Direct Investment FDI ,Net FII Debt and Equity, Net FII debt, Import and Export and found FDI, Net FII equity and Import effect on GDP. There are many research studies on those factors and much more in detail. Hnusuwan et al. (2020). [21] Study the effect of geospatial data in the health sector which allows them to map the potential outbreak of dengue fever. Here we try to depend on visualizing data to understand different aspects about countries and regions and give a comprehensive view on data in a manner that can be used to explain the differences and the relationships between other factors as well as the similarities. My research has limitation that I did not document the visualization projects because most of them did not in research paper instead of that it published in websites such as kaggle and guthip. El Mouden et al. (2020) [22] pointed to the idea of data modeling by graphs, where the input and the output are graph schemas were no approaches were proposed to deal with that before. However, this research attempts to show different visualization methods to explore the countries data in addition to using visualization to explain any relationships between the attributes in the dataset.

## 3 Methodology

The methodology deals with different stages of the model it consists of data collection, data preprocessing and data visualization. We use Python programming languages and Google Colab[6] platform which allows to write and execute Python in the browser, the details about the stages describe below:

### 3.1 Data collection

The dataset requirement for this research is fulfilled through Kaggle[1] it is include data about 227 countries and 20 attributes. Those attributes are Country, Region, Population, Area, Population Density, Coastline, Net migration, Infant mortality, GDP, Literacy, Phones, Arable ratio, Crops ratio, Other ratio, Climate, Birthrate, Death rate, Agriculture, Industry and Service. The region represent where the countries are such as

Northern Africa, Northern America etc. the population refer to the number of population in the country, Area at square meter of the country, Population Density include is the number of people per square meter, Coastline which is defined as the area where land meets the sea or ocean it calculated here in coast/area ratio[7], Net migration which is refer to the difference between the number of immigrants who are coming into the country and the number of emigrants who left the country throughout the year[8], Infant mortality represents Infant mortality per 1000 births, GDP which is the short of gross domestic product calculated in dollar per capita. Literacy ratio of educated people in the country, Phones refer to the numbers of phones per 1000 person, Arable land ratio in the country , Crops ratio that country produce, Other which is refer to other products the country produce ratio, Climate is describe the country climate which ranges from 1 to 4 and they are as follow : 1 means dry tropical or tundra and ice, 2 means wet tropical, 3 means temperate humid subtropical and temperate continental and 4 means dry hot summers and wet winters[9] , Birthrate represent the ratio of the birth in the country, as well as death rate represent the ratio of the death in the country, Agriculture, Industry and Service those represent their value added which is determined by the International Standard Industrial Classification (ISIC). Value added is the net output of a sector after adding up all outputs and subtracting intermediate inputs. It is calculated without making deductions for depreciation of fabricated assets or depletion and degradation of natural resources.

Data collected by Fernando Lasso [10]. He collected the data from World Facebook [9] which is public domain. The details about the source of the attributes mentioned in webpage [11] pointed to that anyone can freely use the dataset.

### 3.2 Data preprocessing

In this stage the dataset is prepared for data visualization. Preprocessing methods include cleaning, variable transformation and so on. Cleaning the data includes filling missing data values and replacing the outliers with mean. Feature engineering also applied on the dataset. Trying to explore the data, we visualize the attribute in order to understand the data in an easy way. After visualizing the data, we will test six hypotheses and try to test the relation between GDP of the countries and other factors. We test many factors that we assume they are affected or affected by the GDP. Those factors are Population Density, migration, Infant mortality, Literacy, Birthrate and Death rate. Hypothesizes that we are try to test are:

- *H1:* GDP effects on birthrate.
- H2: GDP effects on death rate.
- H3: GDP effects on Infant mortality.
- H4: GDP effects on migration.
- H5: Literacy effects on GDP.
- H6: Population Density effects on GDP.

Through visualizing the correlations between those factors we will be able to understand the if there are any correlation between them.

### 3.3 Data visualization

Here we use the python programming language because it is rich in libraries for visualizing data which present the data in the best way. The libraries used are NumPy [12], Panda [13], Matplotlib [14] and Seaborn [15].

As an exploring process we visualize the attributes of the dataset which will give us a comprehensive look on the data. Figure 1 shows the percentage of distribution of the countries in the regions. The regions according to the dataset are Asia, Baltics, C.W. Of India States, Eastern Europe, Latin America and Caribou, Near East, Northern Africa, Northern America, Oceania Sub-Saharan Africa, Western Europe. The figure shows the Sub-Saharan Africa has the largest number of countries in the world. we use pie charts which make the comparison understandable.

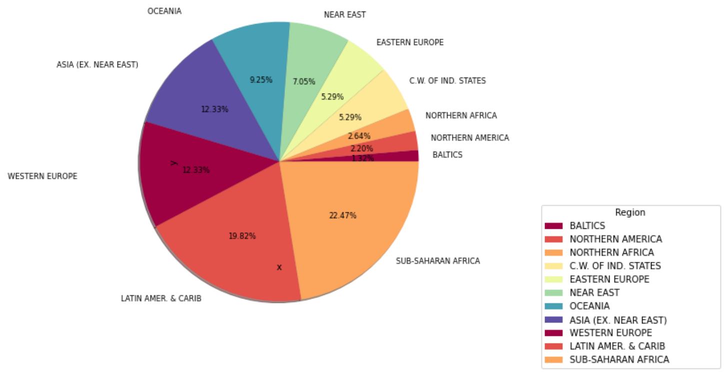


**Fig. 1.** Distribution of the Countries in the Regions

Figure 2 shows the top ten countries in population density which shows, Monaco has largest population density in the world. On other hand figure 3 shows the ten countries have less population density in the world shows the Green Land has less population density in the world.

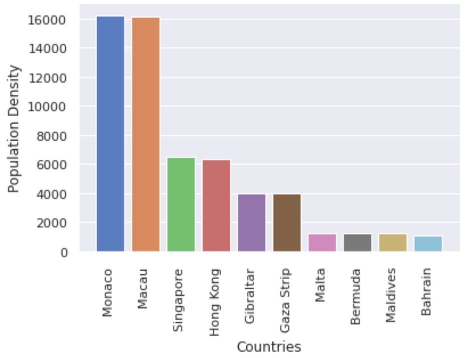


**Fig. 2.** Top Ten Countries in Population Density

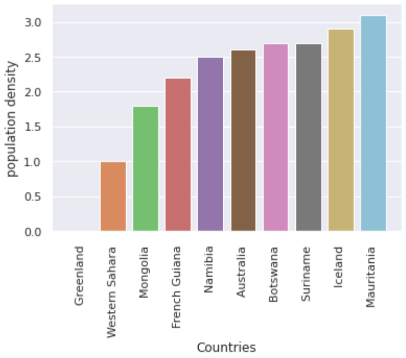


**Fig. 3.** Less Ten Countries in Population Density

For Areas the top ten largest countries and top ten smallest countries area showed in figure 4 and 5. Those showed Russia has the largest area in the world and Monaco is the smallest country in the world. Try to understand if there any relation between the population density and the area figure 6 represents the relation and show there is no relation between them.

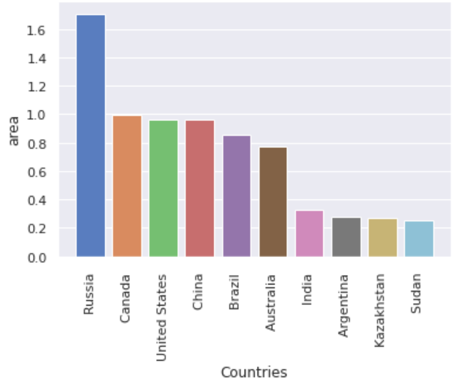


**Fig. 4.** Top Ten Countries in Area

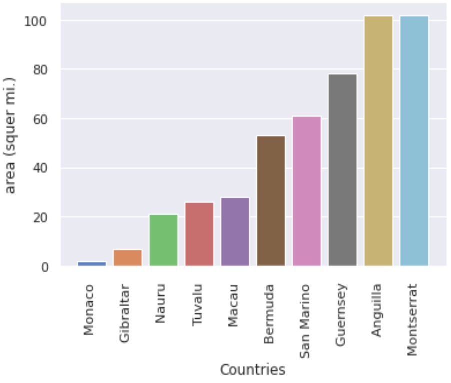


**Fig. 5.** Smallest Ten Countries in Area

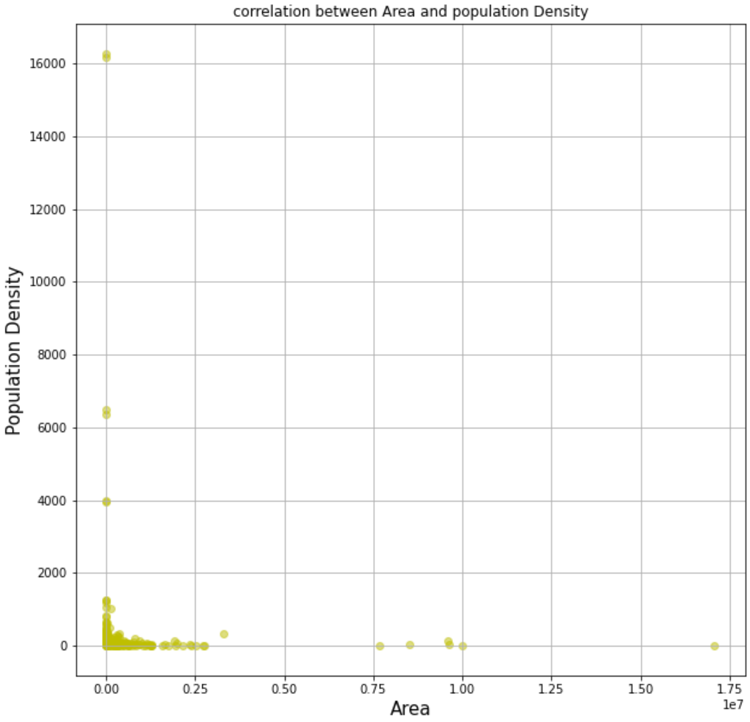


**Fig. 6.** The Relation between Population Density and the Area

Figure 7 and 8 try to show the birth and death rate in the regions of the world.in addition to show the distribution of the birth rate and death rate across the world in figure 9 and 10 we use distplot which produces a plot of cumulative distribution function [16].

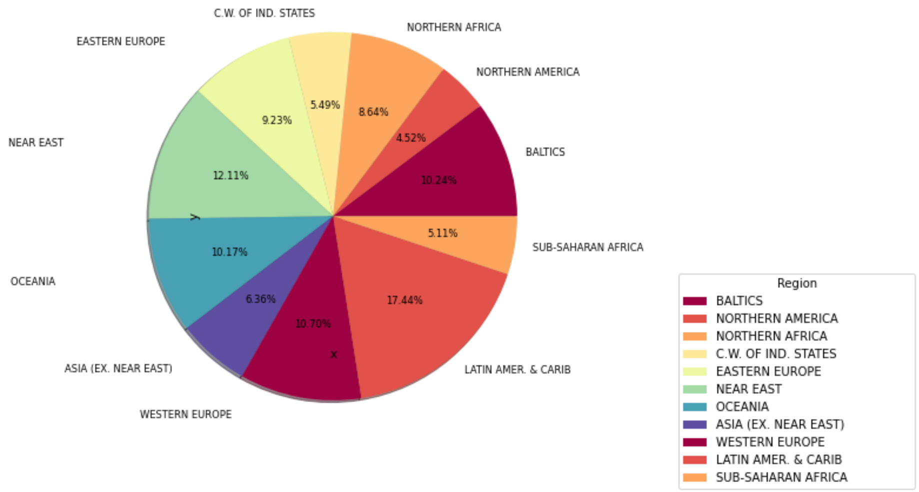


**Fig. 7.** Birth Rate in the Region

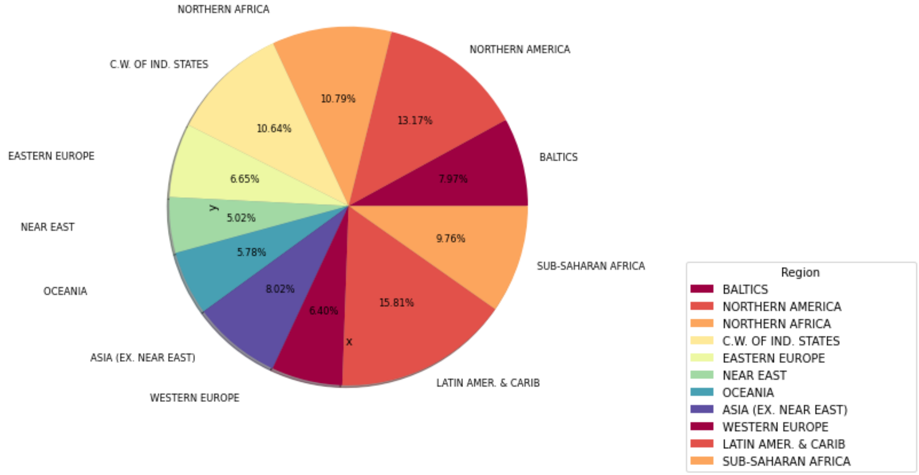


**Fig. 8.** Death Rate in the Region

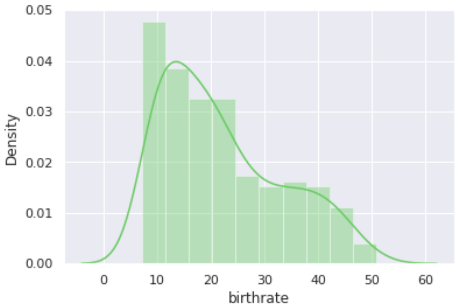


**Fig. 9.** Density of the Birth Rate in the World

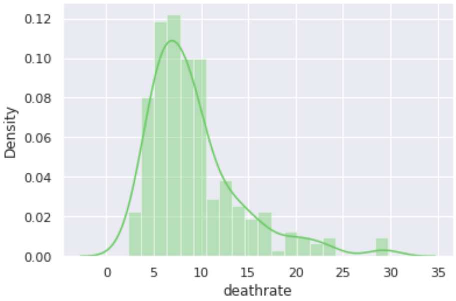


**Fig. 10.** Density of the Death Rate in the World

Then figure 11 and 12 show the top ten countries in birth rate and death rate. On the other hand, figure 13 and 14 show the countries that have less birth rate and death rate. Those figures indicate that Niger has the biggest rate of Birth in the world and one of the top ten of the death rate and Swaziland has the biggest rate of death in the world. In addition to indicating that Hong Kong has the lowest rate of birth in the world and Mariana Island has the lowest Rate of death, we notice in the lowest ten countries in death rate there are five Arab countries in the list which is an interesting result.

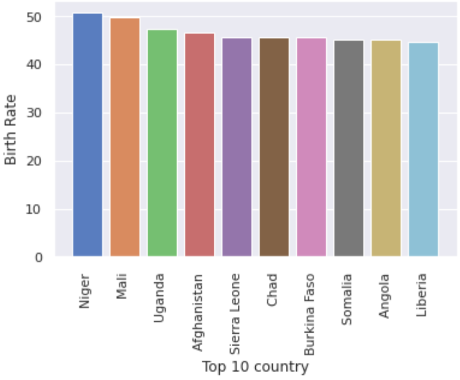


**Fig. 11.**Top Ten Countries in Birth Rate

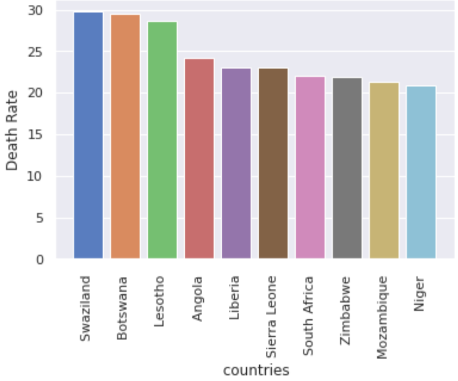
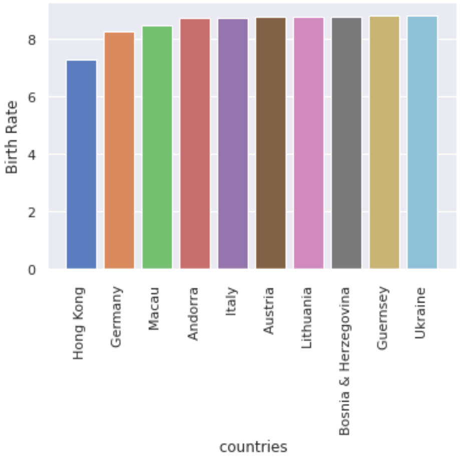


**Fig. 12.**Top Ten Countries in Death Rate



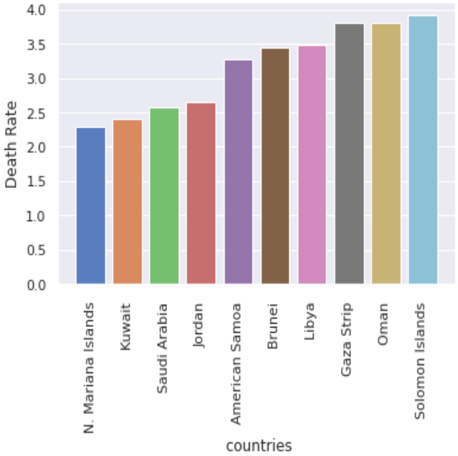


**Fig. 13.** Lowest Ten Countries in Birth Rate

**Fig. 14.**Lowest Ten Countries in Death Rate

Figure 15 shows the relation between the birth rate and death rate which pointed to there are small indicate to correlation between them we should not consider it.

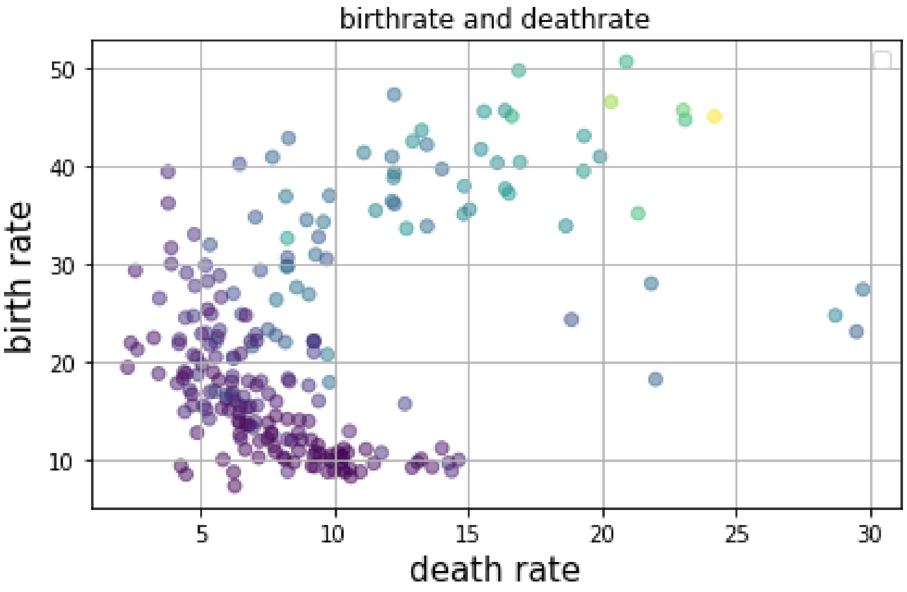


**Fig. 15.** Relation between Birth Rate and Death Rate where the color represent the infant mortality

We also consider the infant mortality rate we show the distribution rate by displot in figure 16 and the relation with birth rate in figure 17, shows there are positive correlations between birth rate and infant mortality.

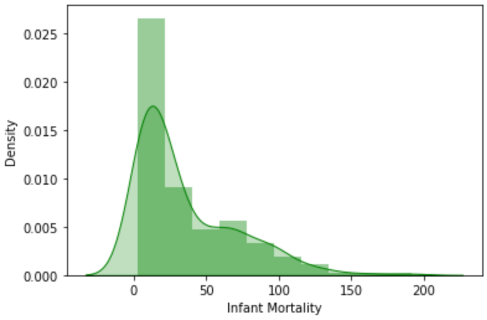


**Fig. 16.**Infant Mortality Rate Distribution in the world

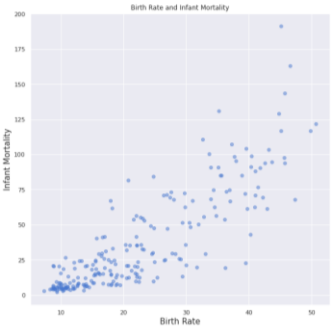


**Fig. 17.** Relation Infant Mortality Rate and Birth Rate

The literacy rate of the countries also interested information for a lot of people, in figure 18 show the distribution of the literacy rate across the world which showed the high number of countries in the high rate of literacy. We can see the distribution of the rate by the regions in figure 19.

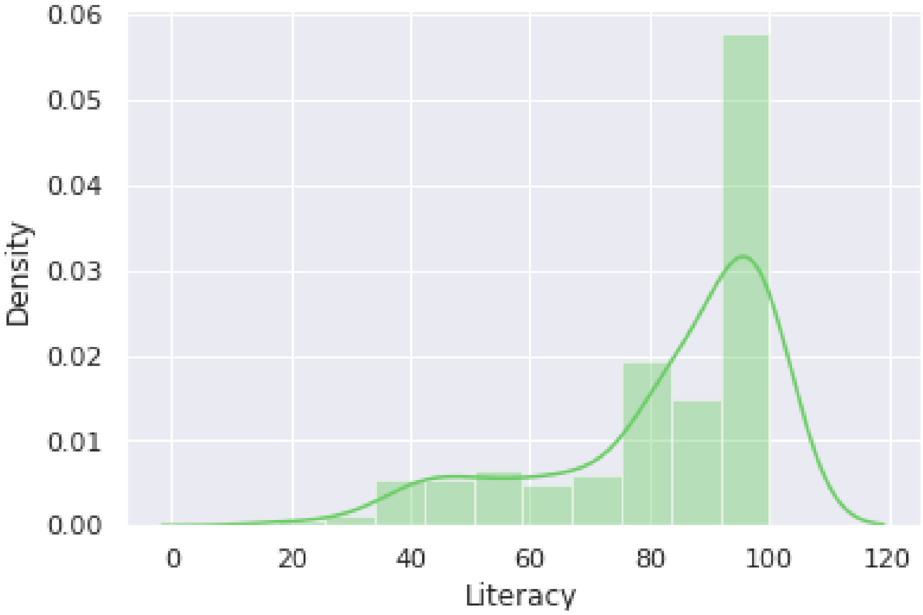


**Fig. 18.** Literacy Rate Distribution in the world

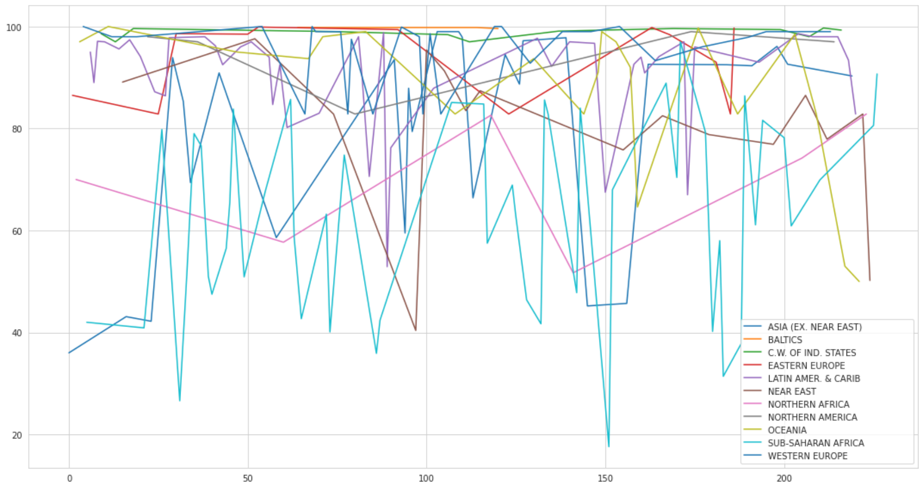


**Fig. 19.**Literacy Rate Distribution in the Regions

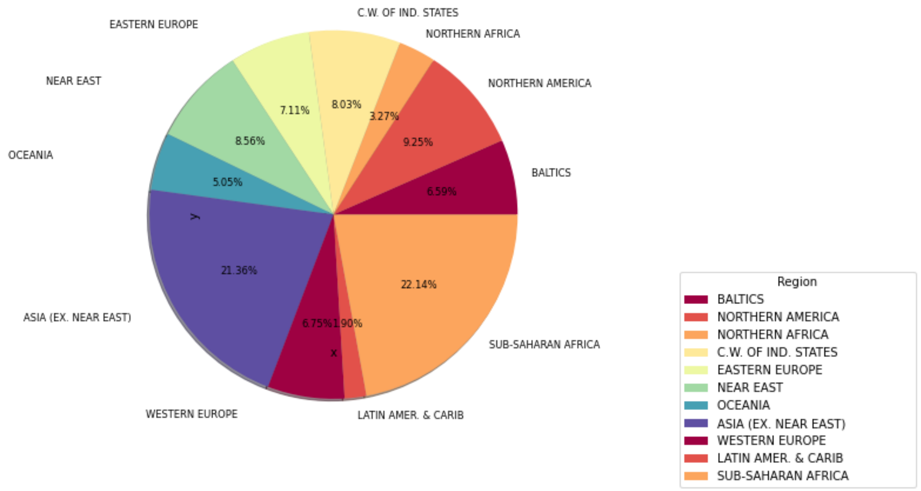


**Fig. 20.** GDP by the Regions

GDP is an important attribute. We showed the total GDP by the region in figure 20 and 21. Pie Chart indicates that Western Europe has the largest share of the GDP in the world by 22.3%. We Use Boxplot to provide details more than a simple pie chart. GDP distribution in the world shows in figure 22 by dispolt. The top ten countries in GDP per capita shows in figure 23. The lowest 10 countries in GDP also shows in figure 24. It shows Luxembourg which has the highest GDP in the world by more than 5000$ per capita. And East Timor, Sierra Leone and Somalia have the lowest GDP in the world by 500$ per capita.

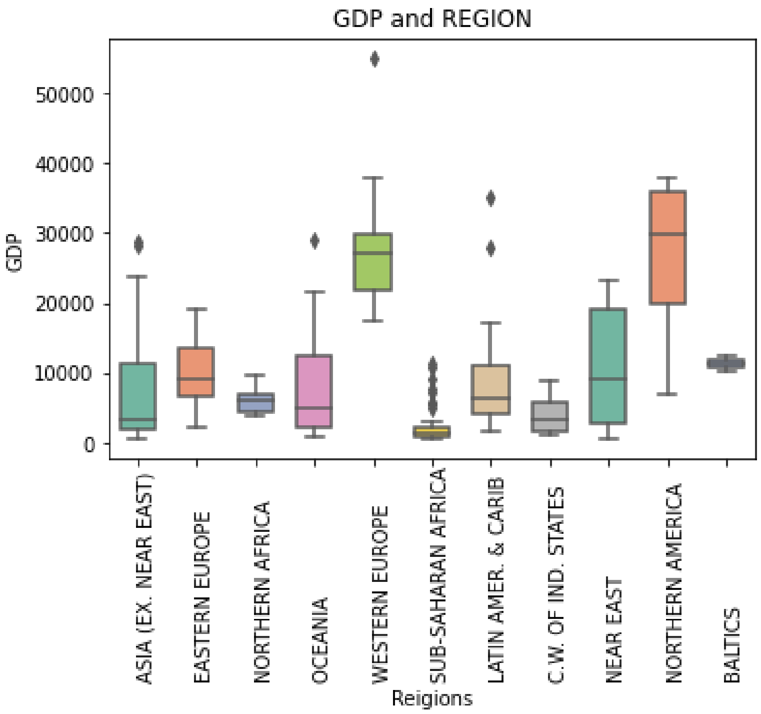


**Fig. 21.** Boxplot of the GDP by Regions

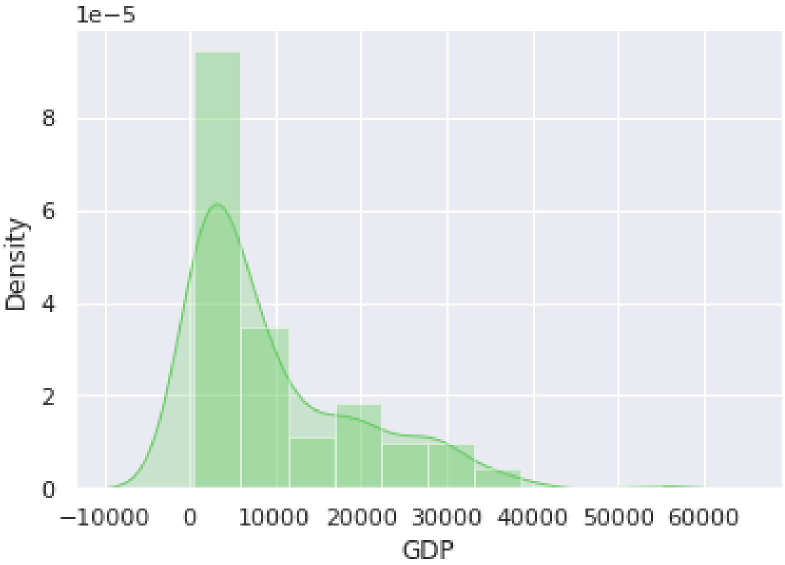
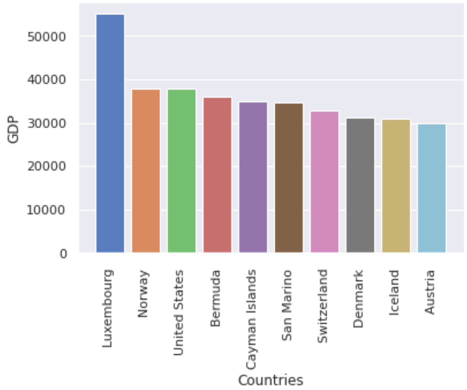


**Fig. 22.**Distribution of the GDP in the world



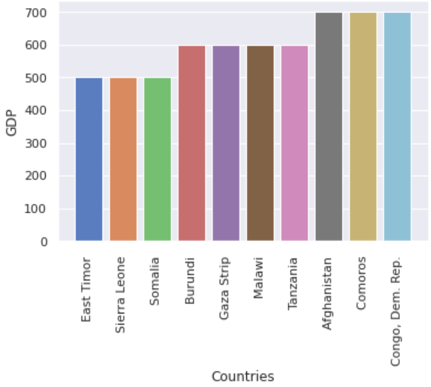


**Fig. 23.**Top Ten Countries in GDP in the world

**Fig. 24.**Lowest Ten Countries in GDP in the world

Heatmap is fundamental and common visualization method for correlation [16] figure 25 show the heatmap of all attributes in the dataset. Then to test the hypothesis above of the GDP effect, we used the scatterplot which is used to show how much one variable is affected by another [18]. Figure 26, 27, 28 and 29shows the scatter plot of the GDP effects on Birth rate, Death rate, Infant mortality and migration. Where figure 30 and 31 show the effect of Literacy and Population density on GDP. The result shows there are no correlations between the factors and the GDP, which lead us to reject all hypotheses. We notice in heatmap the only attributes that have correlation with GDP is the phones which refer to the numbers of phones per 1000 people.

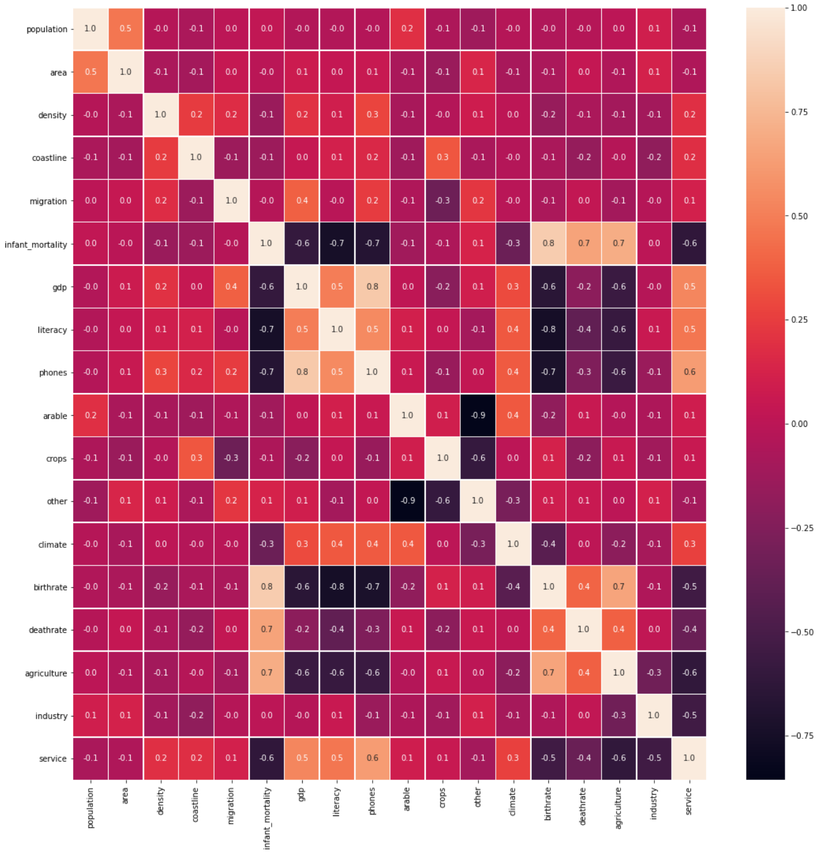
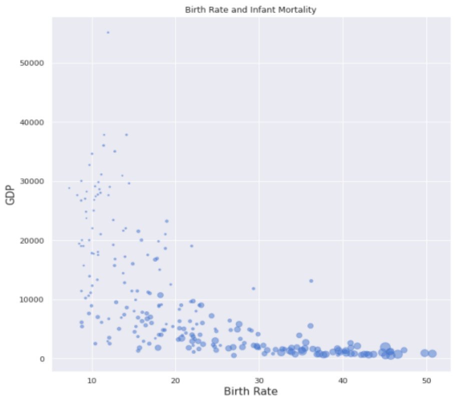


**Fig. 25.** Heatmap of Correlations between Attributes in the dataset



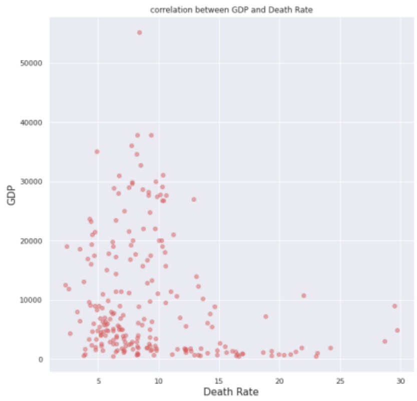


**Fig. 26.**Correlation between Birth rate and GDP

**Fig. 27.**Correlation between Death rate and GDP

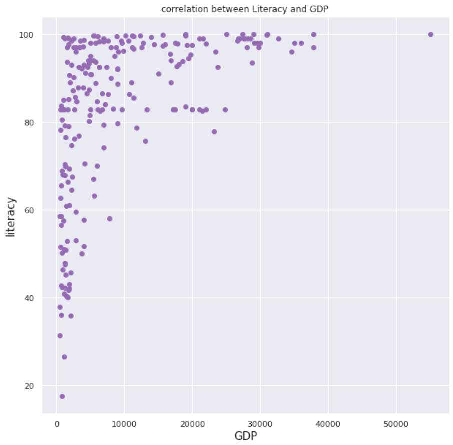


**Fig. 28.**correlation between Infant mortality and GDP

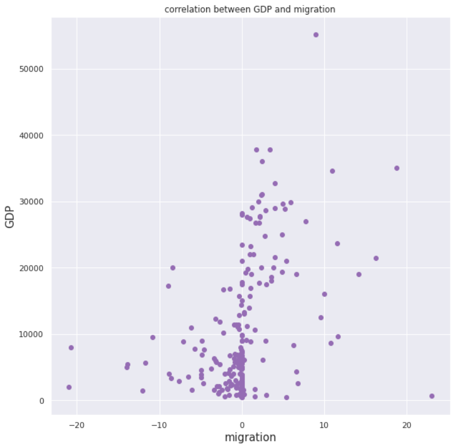


**Fig. 29.**correlation between migration and GDP

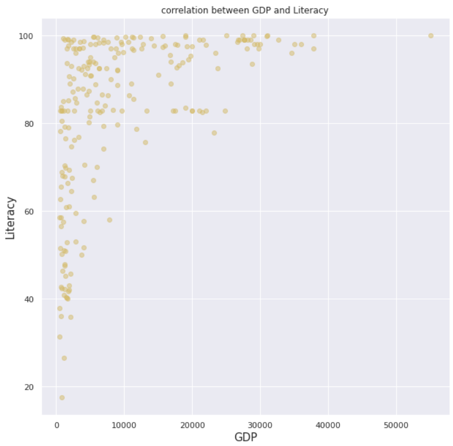


**Fig. 30.**correlation between GDP and Literacy

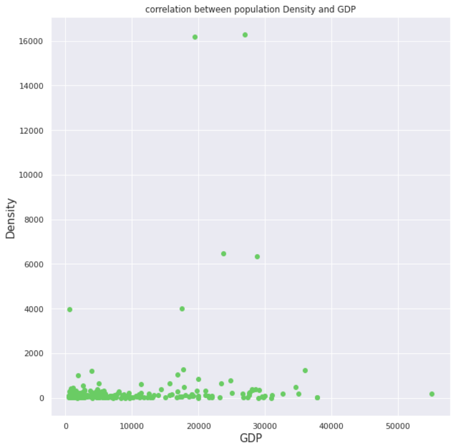


**Fig. 31.**correlation between GDP and Population density

# 4 Discussion

There is a need for extracting information from the massive amount of data [19] Using different visualization methods can help to explore the data and make it easy to see the useful information in graph form which allows us to have a comprehensive view on the data. Furthermore, the data mining technologies which can help to extract useful information and increase the quality and efficacy of pattern detection [20]. This research study countries dataset. Visualization of the dataset give us insights about the countries we represent the distribution of countries in the regions. The population density show Monaco has the most rate and Greenland the less. Monaco also has the smallest area in the world although the representation show there are no relationship between the area and the papulation density. Birth rate show most of top 10 are poor countries where the death rate different and there are no relations between them nether with infant mortality. The less rate of birth was in Hong Kong where the top 10 less rate of death has five Arabian countries which is Kuwait, Saudi Arabia, Jorden, Oman and Libya. We notice literacy is high in most countries and different in regions. GDP is the important factor we found the Western Europe and northern America have high value at most of their countries and show Luxemburg have the highest value of GDP in the world. Furthermore, we examine some factors by assuming they are affected on GDP or affected by those factors are Birth rate, Death rate, Infant mortality, migration, literacy and Population Density and find there are no relationship between those factors and GDP. We found the only attribute affected by GDP is the phone number which refers to the number of phones per 1000 people. That led us to reject all those hypothesis.

## 5 Conclusion

Graph is more than a thousand words and it is the way of human brain works. Python provide many interactive libraries which help to represent huge data as countries data such as NumPy [12], Panda [13], Matplotlib [14] and Seaborn [15]. This paper aims to use visualization as exploring method of the countries dataset in addition to show the relations in the data set through visualization. Visualization the data allowed us to explain the data and represented it in understandable way in addition to test effect of some factors on each other. As many research do, we try to test the effect of GDP on other factors and if it effected by them. The most obvious finding of this study is that no relations between any attributes in the dataset and GDP unless the phone number which refers to the number of phones per 1000 people. The key strength of this study is there are considerable amount of visualization countries dataset have been done in different websites but based on authors' best knowledge there are no one documented in science.

## 6 References

[1] Friedman, V. (2008). Data visualization and infographics. Graphics, Monday Inspiration, 14, 2008.

[2] Fernandol (2017), from https://www.kaggle.com/fernandol/countries-of-the-world

[3] Ilter, C. (2017). What economic and social factors affect GDP per capita? A study on 40 countries. Journal of Global Strategic Management, 11(2), 51-62. https://doi.org/10.20460/jgsm.2018.252

[4] Zhang, W., Yang, D., &Huo, J. (2016). Studies of the relationship between city size and urban benefits in China based on a panel data model. Sustainability, 8(6), 554. https://doi.org/10.3390/su8060554

[5] Jain, D., Nair, K., & Jain, V. (2015). Factors affecting GDP (manufacturing, services, industry): An Indian perspective. *Annual Research Journal of SCMS Pune*, *3*, 38-56.

[6] Google Colab, (n.d) from https://colab.research.google.com/

[7] Coast (2020,October 10) In *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Coast

[8] Net Migration Rate (2020, October 10) In *Wikipedia*. Retrieved from https://en.wikipedia.org/wiki/Net_migration_rate

[9] World Factbook (n.d) October ,2020 from https://www.cia.gov/library/publications/the-world-factbook/index.html

[10] Fernandol (n.d) October ,2020 from https://www.kaggle.com/fernandol

[11] Data Set (2017) October ,2020 from https://gsociology.icaap.org/dataupload.html

[12] NumPy Library (n.d) October ,2020 from http://www.numpy.org/

[13] Panda Library (n.d) October ,2020 from http://pandas.pydata.org/

[14] Matplotlab Library (n.d) October ,2020 from https://matplotlib.org/

[15] Seaborn Library (n.d) October ,2020 from https://seaborn.pydata.org/

[16] Cox, N. (2017). DISTPLOT: Stata module to generate distribution function plot.

[17] Gu, Z., Eils, R., &Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics, 32(18), 2847-2849. https://doi.org/10.1093/bioinformatics/btw313

[18] Keim, D. A., Hao, M. C., Dayal, U., Janetzko, H., &Bak, P. (2010). Generalized scatter plots. Information Visualization, 9(4), 301-311. https://doi.org/10.1057/ivs.2009.34

[19] AlAjmi, M. F., Khan, S., & Sharma, A. (2013). Studying data mining and data warehousing with different e-learning system. IJACSA) International Journal of Advanced Computer Science and Applications, 4(1). https://doi.org/10.14569/ijacsa.2013.040122

[20] Khan, S.(2016) How Data Mining Can Help Curb City Crime, International Journal of Control Theory and Applications (IJCTA) 9 (23), 483-488.

[21] Hnusuwan, B., Kajornkasirat, S., & Puttinaovarat, S. (2020). Dengue Risk Mapping from Geospatial Data Using GIS and Data Mining Techniques, International Journal of Online and Biomedical Engineering (iJOE), Vol. 16, No. 11, 2020. https://doi.org/10.3991/ijoe.v16i11.16455

[22] El Mouden, Z. A., & Jakimi, A. (2020). A New Algorithm for Storing and Migrating Data Modelled by Graphs, International Journal of Online and Biomedical Engineering (iJOE), Vol. 16, No. 11, 2020. https://doi.org/10.3991/ijoe.v16i11.15545

## 7 Author

**Dr. Shakir Khan** received his BSc, MSc and PhD in computer science in 1999, 2005 and 2011 respectively. He is member of the International Association of Online Engineering (IAOE) and IEEE; He is currently working as Associate Professor at College of Computer and Information Sciences in Imam Mohammad Ibn Saud Islamic University, Riyadh (Saudi Arabia). His research interest is Big Data, Data Science, Data Mining, Machine Learning, Internet of Things (IoT), and eLearning, Artificial Intelligence, Emerging Technology, Open Source Software, Library Automation and Mobile / Web Application. He published many research papers in international journals and conferences in his research domain. He has around 15 years of teaching, research and IT experience in India and Saudi Arabia. Dr. Khan is teaching bachelor and master degree courses in the college of computer at Imam University. He is reviewer for many international journals.

Article submitted 2020-11-30. Resubmitted 2021-05-11. Final acceptance 2021-07-08. Final version published as submitted by the author.