**World Population**

**A detailed data analysis on the World Population and Countries**

**By**

**Ashutosh Burde**



**Information Science**

**School of Computing and Information**

**University of Pittsburgh**

**2018-2019**

**About my Project:**

My project is a detailed data analysis on the world population and countries. It helps to display the divide that is present between the various countries of the world. The analysis deals with properties of countries that ranges from GDP\* and urban percentage to maternal and child mortality rates. I have done most of the analysis using Python and Jupyter notebook

**DataSet:**

I used raw csv data from the files that I have attached with this paper. They are from CIAWorldFactBook\* and Google\*.

**Approach:**

My approach can be broken down into

Importing the data

Cleaning the data

Performing data analysis

Gathering the findings and results

Phase One……

Started with importing the raw data from the csv files.

Imported the respective libraries and stored the data in a data frame.

Then printed the data to just verify the import.

Phase Two ….

Cleaned the data by filling the empty spaces in the data frame with 0’s and dropping the data with N/A in the dataframe.

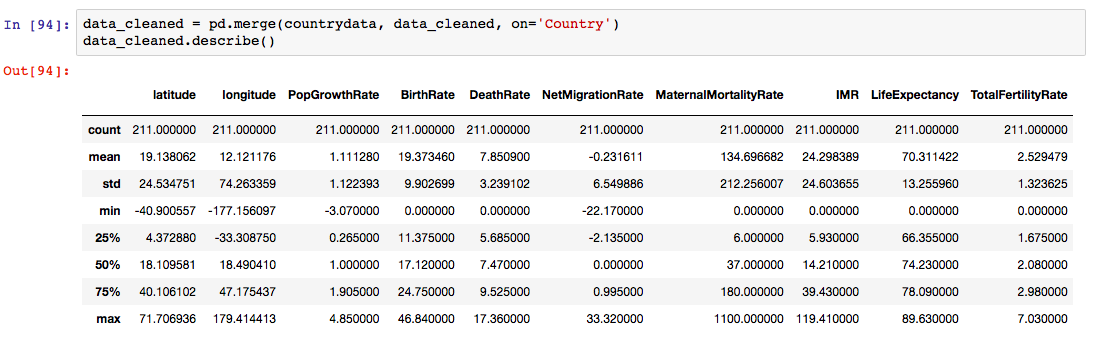
Then for the percentages; replaced the % sign with blank and converted the data type to float, so that it can be used for analysis in the future.

Replaced the $ and , sign in the GDP\_PPP\_percapita to blanks and stripped the leading and trailing blank spaces in that column of dataframe

Converted the datatype from string to numeric so that it can be used for data analysis in the future

Merged the 2 data frames on the Country column

Described the data to check if the data is cleaned by checking the rowcount of all columns

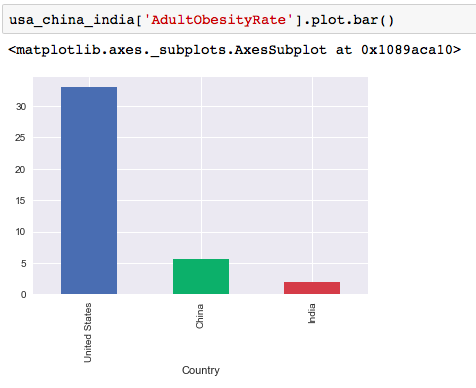


Phase Three…..

Checked the data types of all the columns to determine if they can be used for data analysis.

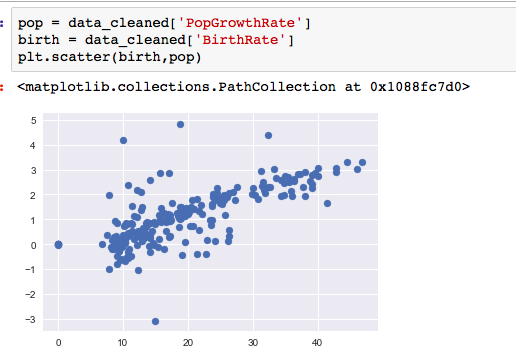
Changed the index of the data frame to countries and selected 3 countries to display bar graphs.

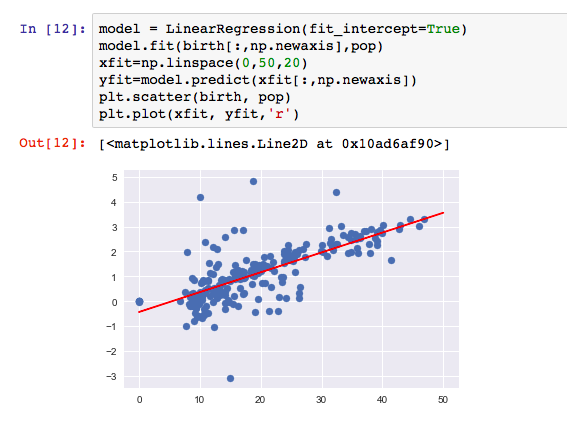
One of the properties compared was Adult Obesity Rate



Phase Four….

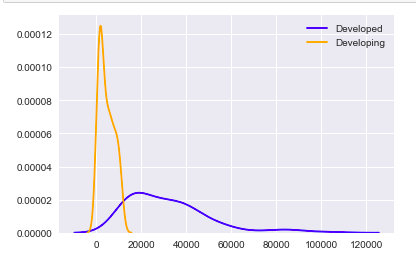
Showing a linear regression between the Population Growth Rate and Birth Rate





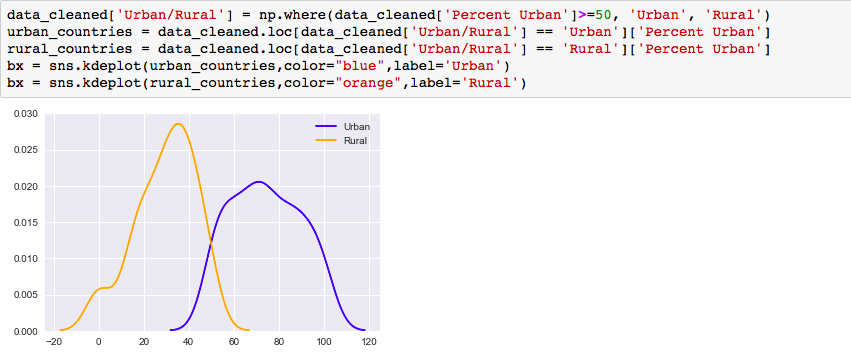
Phase Five …

Performed Naïve Based Classification to classify countries into Developed and Developing countries based on GDP\*



Performed predictions of countries given only their GDP\* and classified them as Developed or Developing.

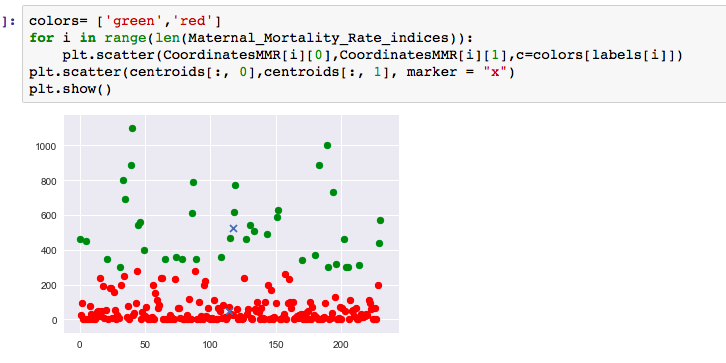
Performed Naïve Based Classification to classify countries into Urban/Rural based on Percent Urban



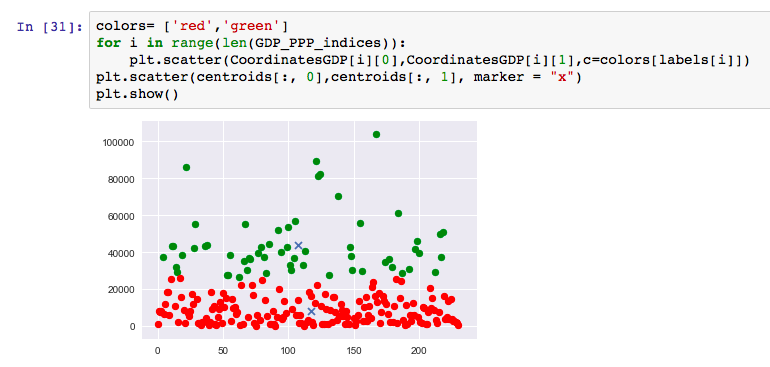
Performed predictions of countries given only their Percent Urban and classified them as Urban or Rural

Phase Six …..

Performed K means clustering on Maternal Mortality Rates



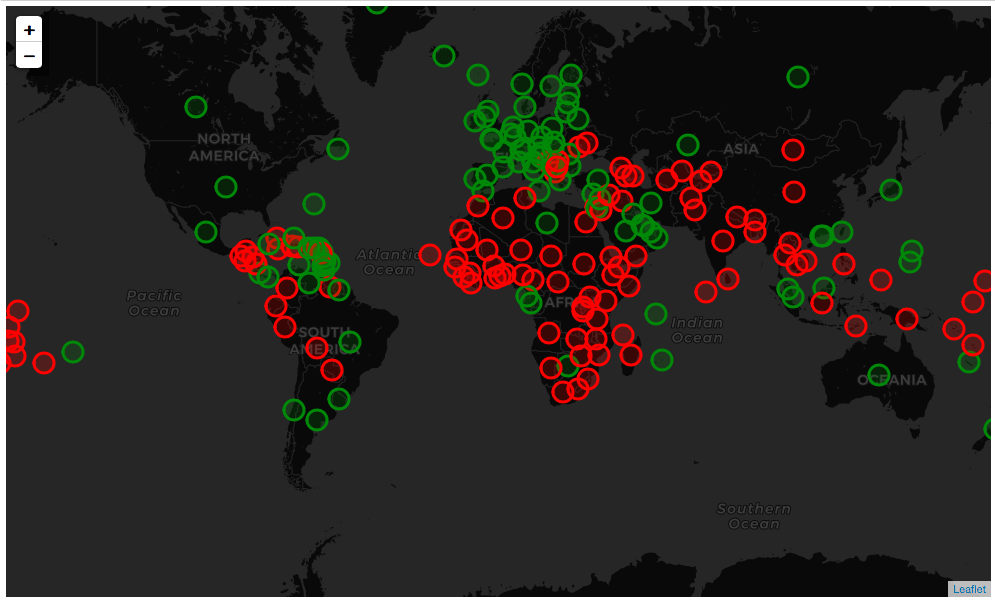
Performed K means clustering on GDP\_PPP\_percapita



Phase Seven ……

Installed and imported the folium\* library package

Classification based on developed and developing countries. With the green and red colors respectively on the folium map.



References and Glossary:

\*GDP: Gross Domestic Product

<http://gsociology.icaap.org/dataupload.html>

<https://blog.prototypr.io/interactive-maps-with-python-part-1-aa1563dbe5a9>

<https://developers.google.com/public-data/docs/canonical/countries_csv>