GDP2017 Analysis

Nitish Goyal

23 September 2018

Executive Summary

This document presents an analysis of data concerning Gross Domestic Product(GDP) of various economies in 2017. The analysis is based on GDP data of 200 countries and aggregate GDPs by geographic regions and income groups.

After cleaning the data, the data was explored by calculating summary statitics, and by creating visulaizations of the data. This exploaratory data analysis helped in identifying characteristics of GDP distribution across the globe.

After performing the analysis, the author presents the following conclusions:

- The GDP distribution across the globe is very uneven
- GDP distribution can be approximated well by a log-normal distribution
- 75% of the world's GDP is accounted by the GDP of top 7 percentile economies

Data Loading and Cleaning

First, the GDP dataset was imported from the URL address of the GDP.csv file. Strings were not imported as factors because they will not be used as categorical variables. Instead they will be used to describe the country name or code.

After importing that data, a quick look at it showed that there were redundant columns as well as rows in the dataset. The following steps were taken to make the dataset more convenient to use:

- Columns 3, 6, 7 and 8 were removed as they consisted no relevant information
- Column names, which were provided in the intial rows were incorporated as column names of the data frame
- Rows with no GDP data were removed as either they were empty or contained unimportant comments or countries without GDP data
- First two rows were removed as the column names were incorporated properly
- The commas were removed from GDP numbers and then GDP and ranking column observations were converted into numeric data types

Finally, the dataset was split into two parts; one containing the GDP data only for countries, the other conataining aggregate GDP data for various regions, income groups and the world.

Exploratory Data Analysis

Summary Statistics

Summary statistics including mean, median, standard deviation (sd), interquartile range(iqr), range and maximum were calculated for GDP data, and the results taken from 200 observations are shown below:

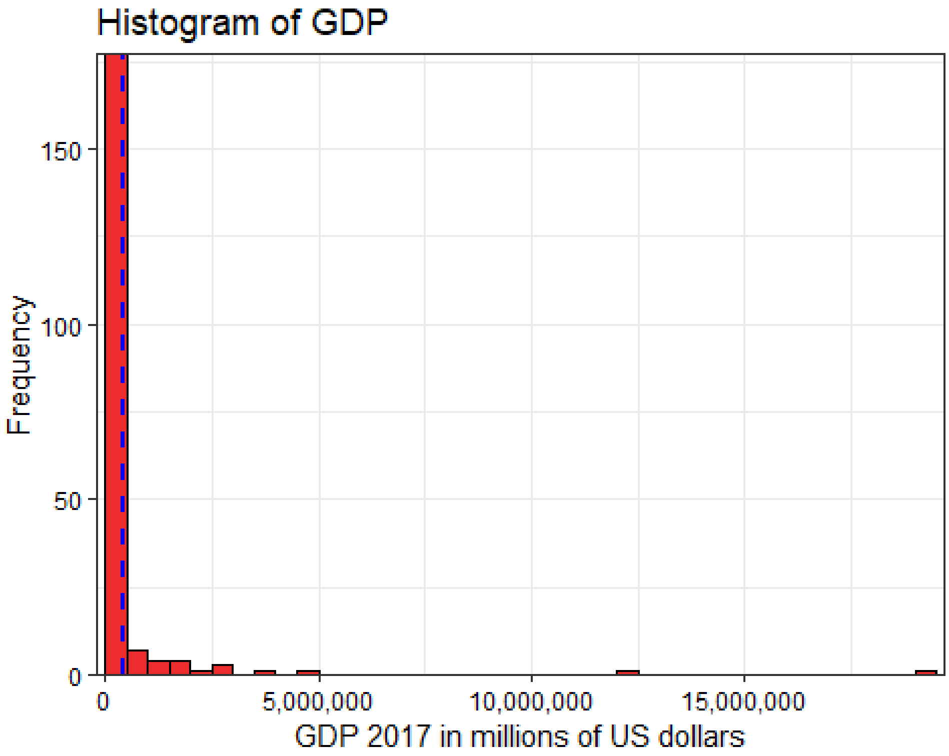
GDP summary

Mean	Median	sd	iqr	Range	Maximum
397286	25906	1706737	195380	19390564	19390604

Significantly large standard deviation implies that the there is a lot of variation in the GDP data. Also, the mean is more than 10 times greater than median, which implies that there are a few countries with GDPs much larger than the median GDP.

Data Visualization

Summary statistics provided some sense of the distribution of the data, but to understand the data better data visualizations of various kinds were used. The first step was to look at a histogram of the GDP data for all 200 economies:



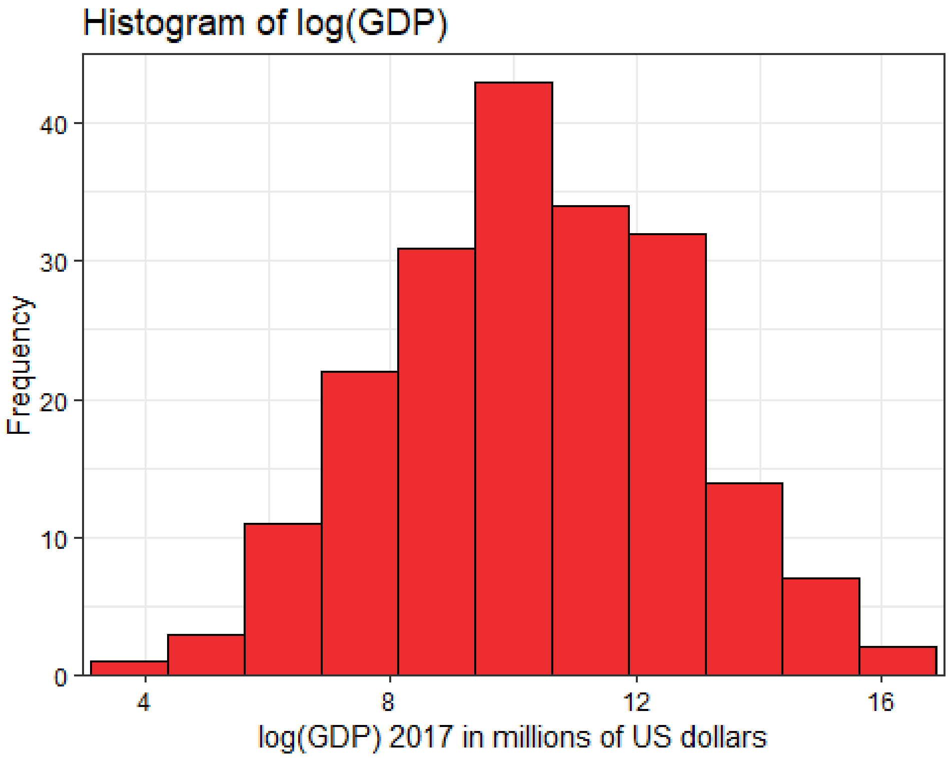
This figure shows that the GDP distribution is right-skewed. In fact, GDPs of most economies lie below the 5,000,000 million dollar mark. As noted, the data is extremely skewed and since the range is almost as large as the maximum observation and many times the mean, a boxplot of all the observations will not be very helpful in gaining insights about GDP distribution.

To visualize the data better, the following two options were explored:

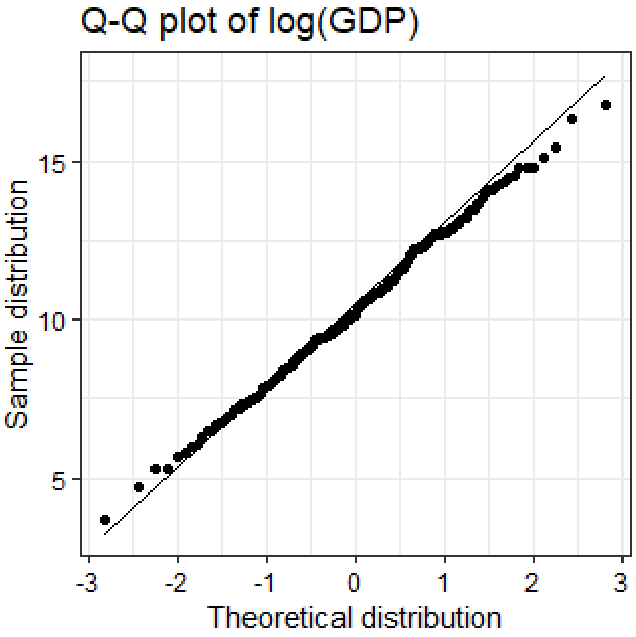
- Visualizations based on log(GDP)
- Visualizations including countries up to a certain percentile(based on GDP)

Visualizations based on log(GDP)

Firstly, a new data set containing a new column log(GDP) was created, and then a histogram of this new column was made, as shown below:



The figure above seems to be similar to a normal distribution. This would imlpy that GDP data is log-normally distributed. To check this claim, a Q-Q plot was created with log(GDP) data using the default distribution (normal), as shown below:



As the points lie mostly on a straight line in the Q-Q plot, it confirms that the GDP data is approximately log-normally distributed.

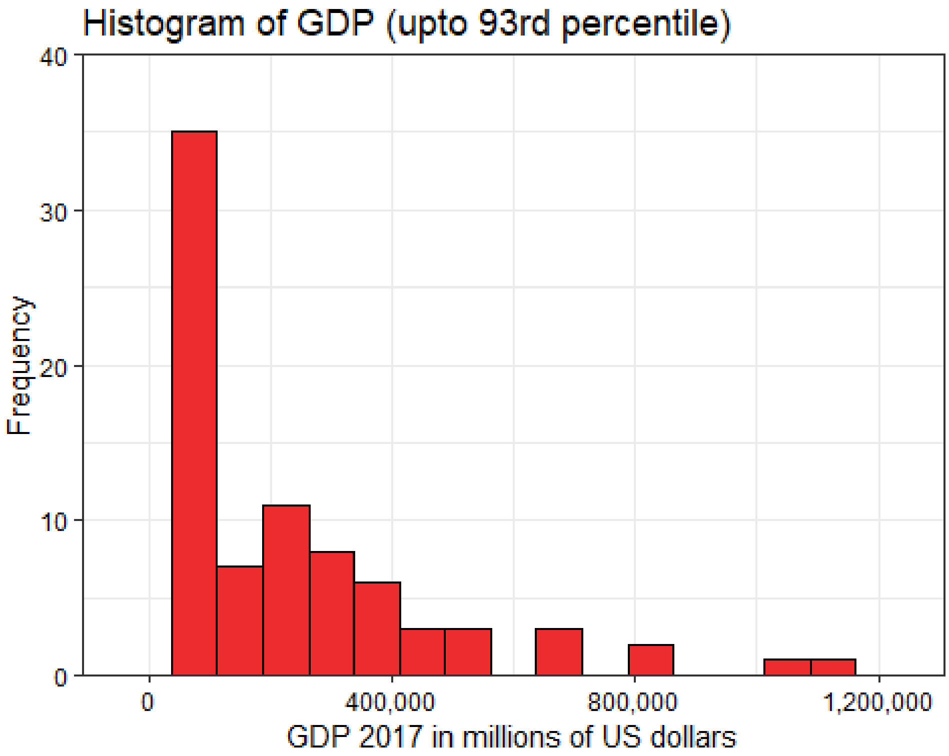
Visualizations including a fraction of countries

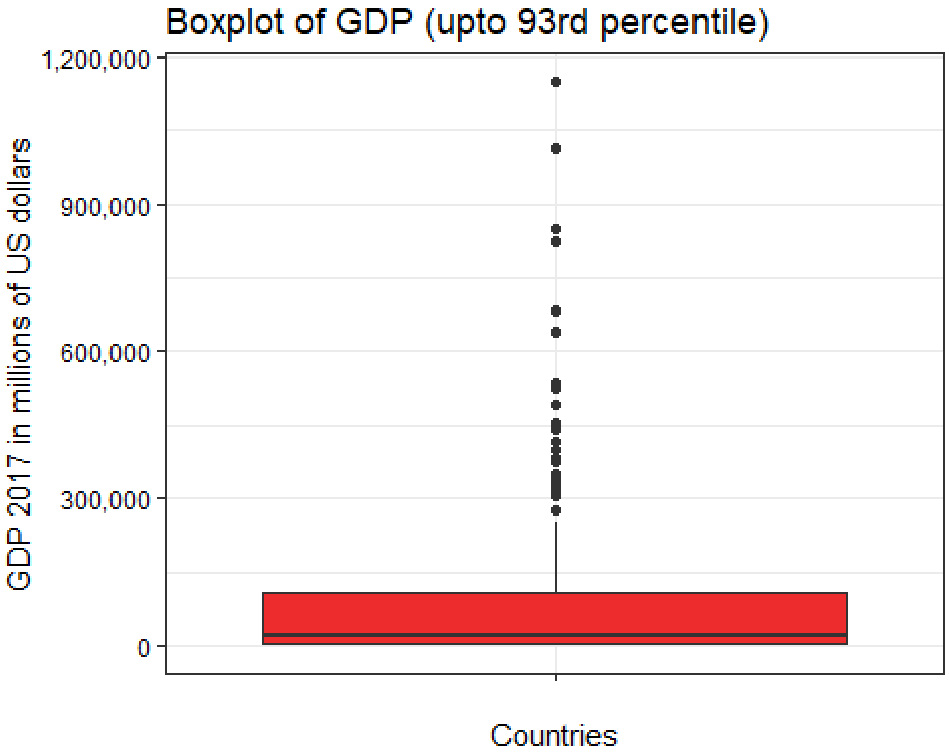
Although the histogram of log(GDP) was helpful in visualizing the data, it is still useful to plot the data as it is because taking the log suppresses the extent of difference in observations. However, as was seen earlier, a histogram of the GDP data as it is, is not very informative about the vast majority of the countries. One way to go around this is to plot the data for a fraction of countries, excluding the outliers, which in this case would be the top few countries. To get a better understanding of how many top countries to exclude a table of perentiles based on GDP was created, as shown below:

GDP by percentile

	x
0%	40
25%	6300
50%	25906
75%	201679
100%	19390604

This table shows that the GDP of the 75th percentile country is almost 100 times lower than the 100th percentile country. This table also confirms the skewness in the data. However, instead of dropping the top quartile of the countries (50), it was decided to drop the top 7% of countries (14), so as to minimize the loss of data due to this truncation, while still making the visualizations more appealing. The histogram and boxplot of countries upto 93rd percentile are shown below:





Although the above two visualizations exclude top 7 percentile of countries, they provide a much better picture about the remaining 186 countries. The reason for this is simple, the GDP of the top 14 countries is so high, that not only does it make the initial histogram hard to read, it also makes the mean of the entire dataset shift significantly. The histogram of 186 countries still shows that this distribution is also right skewed. Another point to be noted is that the sum of GDPs of all countries up to the 93rd percentile account only for 25% of the world's GDP.

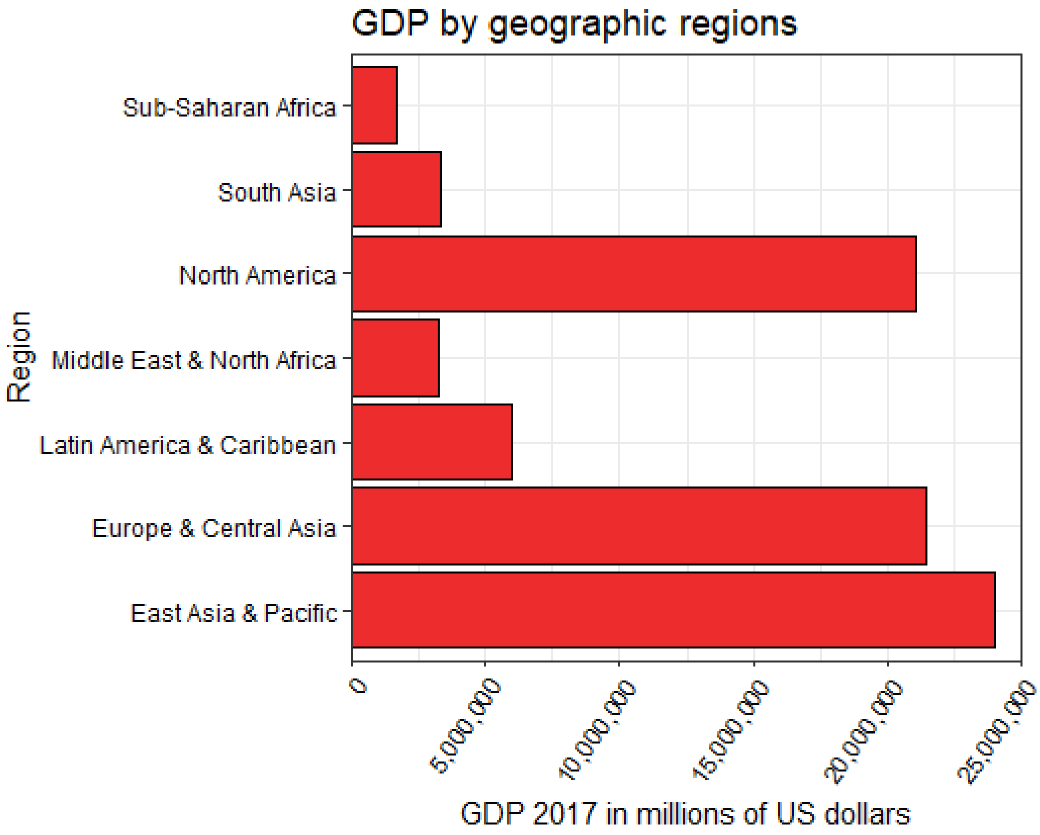
The boxlot with data of 186 countries should be used carefully, as it is not representative of the summary statistics of the overall data (for which the table in the begining should be referred to). However the boxplot does provide some useful information. It shows that there are many outliers even in this truncated dataset. The maximum GDP for this truncated dataset is a little less than 300,000 millions of USD and the GDP of the outliers goes up as high as 12,000,000 millions of USD.

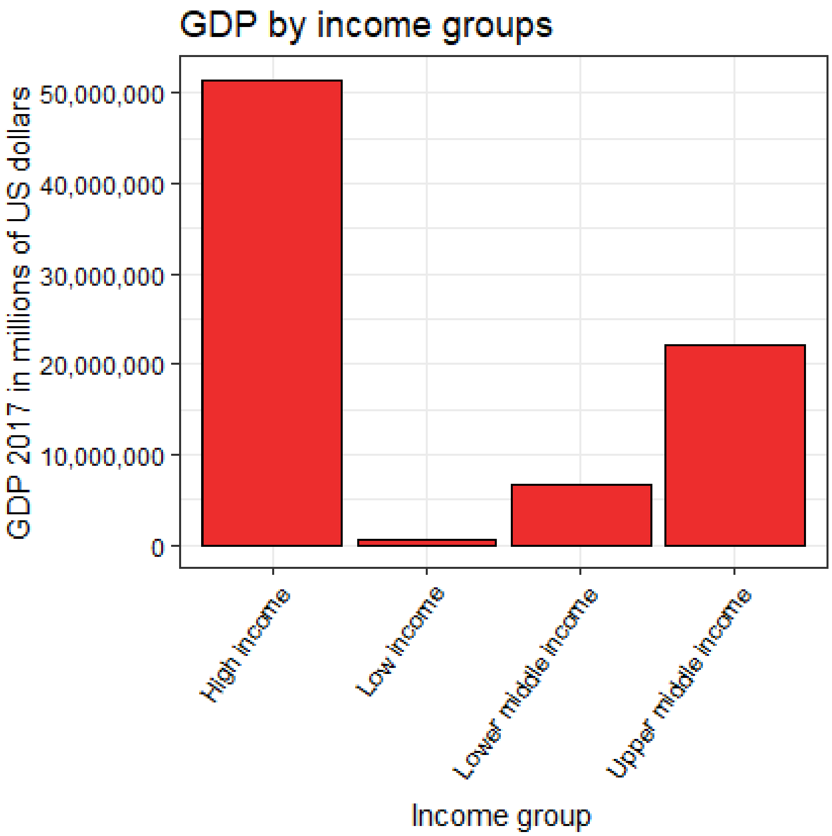
Visualizations based on aggregate data

Apart from the visualizations based on GDPs of individual countries, a few more plots were created based on GDP data of the following groups:

- GDP by geographic regions
- GDP by income groups

These two plots are shown below:





The visualizations based on groups show the distribution amongst the groups. For instance the visualization based on income groups shows that the GDP of High income countries is not only more than the GDP of the rest of the income groups, but it is almost double of that. However these visualizations do not explain the reasons for disparities in GDP by region or income groups. As an example, the GDP of Middle East and North Africa is relatively low, but if it is low because of smaller size of countries, lesser population or less advancement of technology and industry or any other reason, can not be concluded.

Conclusion

This analysis has explored the GDP data for 200 countries from the year 2017. The summary statistics and data visulaizations reveled that the GDPs are very unevenly distributed through out the world. Countries within the top 7 percentile (based on GDP) represent 75% of the world GDP. The distribution of GDP data is right skewed as shown by the histogram. ALso, it was verified that the GDP distribution can be approximated well following a log-normal distribution. Distribution of GDP by income groups and geographic regions also provided some insights.