MSDS 6306 401 Case Study 1: Exploration of Countries' GDPs Vs Income Groups

Yao Yao

March 9, 2017

Contents

Introduction:

The following is a data exploration of GDPs and income groups for countries around the world. GDP data of countries is taken from http://data.worldbank.org/data-catalog/GDP-ranking-table, which was last updated on 01-Feb-2017. Income group of countries is taken from http://data.worldbank.org/data-catalog/ed-stats, which was last updated on 22-Feb-2017.

This case study is an exercise of gathering, cleaning, and analyzing data using R markdown to source .R files from various directories, and thus creating the paper file.

Both data sets are stored in .csv format where headers are imported directly from the original data set, where the columns that are not used to answer the questions in the analysis are then eliminated.

Column name description:

1. CountryCode – The 3 letter country shortcode
2. Ranking – Country ranking by GDP with 1 being the highest
3. Economy – Country name
4. US Dollars (millions) – Gross Domestic Product of a certain country, in U.S. Dollars
5. Income.Group – The income group of a country

Problems with the data:

1. The countries that have missing values for those columns listed above are not included in the analysis
2. The download file is updated regularly and may create different results later on

Folder Description:

The folder "DDS-Case-Study-1" contains the following:

Data Directory:

1. main_helper.R – contains .R code that sources everything together prior to analysis
2. gather.R – contains .R code that gathers data sets from websites and stores into EducationalWeb.csv and GDPWeb.csv
3. clean_GDP.R – contains .R code that cleans GDPraw.csv
4. clean_education.R – contains .R code that cleans Educationraw.csv
5. merge.R – contains .R code that merges GDPdata.csv and Educationraw.csv into MergeData1 and creates subsets for later analysis
6. Contains all the .csv files to validate automated processing and analysis

Analysis Directory:

1. main.R – contains .R code that has working functions to do further analysis

Directions to run the code:

main_helper.R downloads the required packages and datasets; it also cleans and merges the datasets for analysis

```
source(".\\Data\\main_helper.R")
```

main.R has functions that are used for the analysis

```
source(".\\Analysis\\main.R")
```

Analysis:

The reason for the analysis is for us to explore and understand the relationship between country GDP and income group for the different nations around the world to come up with meaningful conclusions.

0) Include code to count the number of missing values for each variable used in the analysis:

Finds the number of missing values per column in the merged data set

```
NumberMissingValues(MergeData1)
```

```
## [1] "Total Number of Rows in Merged Data: 235"
## [1] "Number of N/A Rankings: 45"
## [1] "Number of N/A Economies: 45"
## [1] "Number of N/A GDPs: 45"
## [1] "Number of N/A Income Groups: 1"
## [1] "Total Number of Unmatched Rows: 46"
```

For each of the variables utilized in the analysis, there are 45 N/A values for rankings, economies, and GDP. There is one country with a N/A value for income groups and is also eliminated, which results in a total of 46 unmatched rows

1) Merge the data based on the country shortcode. How many of the IDs match?

Finds the total number of matched values in the data set for further analysis

```
NumberMatchedValues(MergeData2)
```

```
## [1] "Number of Rows in Merged Data with complete values: 189"
```

After merging the data by country shortcode and eliminating pertinent rows with N/A values, 189 countries has GDP, educational, and ranking values matching with income group.

2) Sort the data frame in ascending order by GDP (so United States is last). What is the 13th country in the resulting data frame?

Finds the country based on ascending GDP

```
FindCountryByAscendingGDP(NegGDP, 13)
```

```
## [1] "The 13th country in ascending order by GDP is: St. Kitts and Nevis"
```

Shows the subset of the country ranks to show any discrepancies

```
FindSubsetByAscendingGDP(NegGDP, 12, 13)
```

	Ranking	Economy	US Dollars (millions)	Income.Group
69	178	Grenada	767	Upper middle income
93	178	St. Kitts and Nevis	767	Upper middle income

From ascending GDP, country #13 is St. Kitts and Nevis in the resulting data frame NegGDP. Technically, St. Kitts and Grenada are tied at 12th in ascending GDP and further ascending alphabetical sorting makes St. Kitts appear at 13th and Grenada at 12th place.

3) What are the average GDP rankings for the "High income: OECD" and "High income: nonOECD" groups?

Find the mean GDP rank from income group of High Income OECD countries

```
AverageGDPByGroup(MergeData2, "High income: OECD")
```

```
## [1] "The average GDP ranking of High income: OECD countries is: 32.97"
```

Find the mean GDP rank from income group of High Income nonOECD countries

```
AverageGDPByGroup(MergeData2, "High income: nonOECD")
```

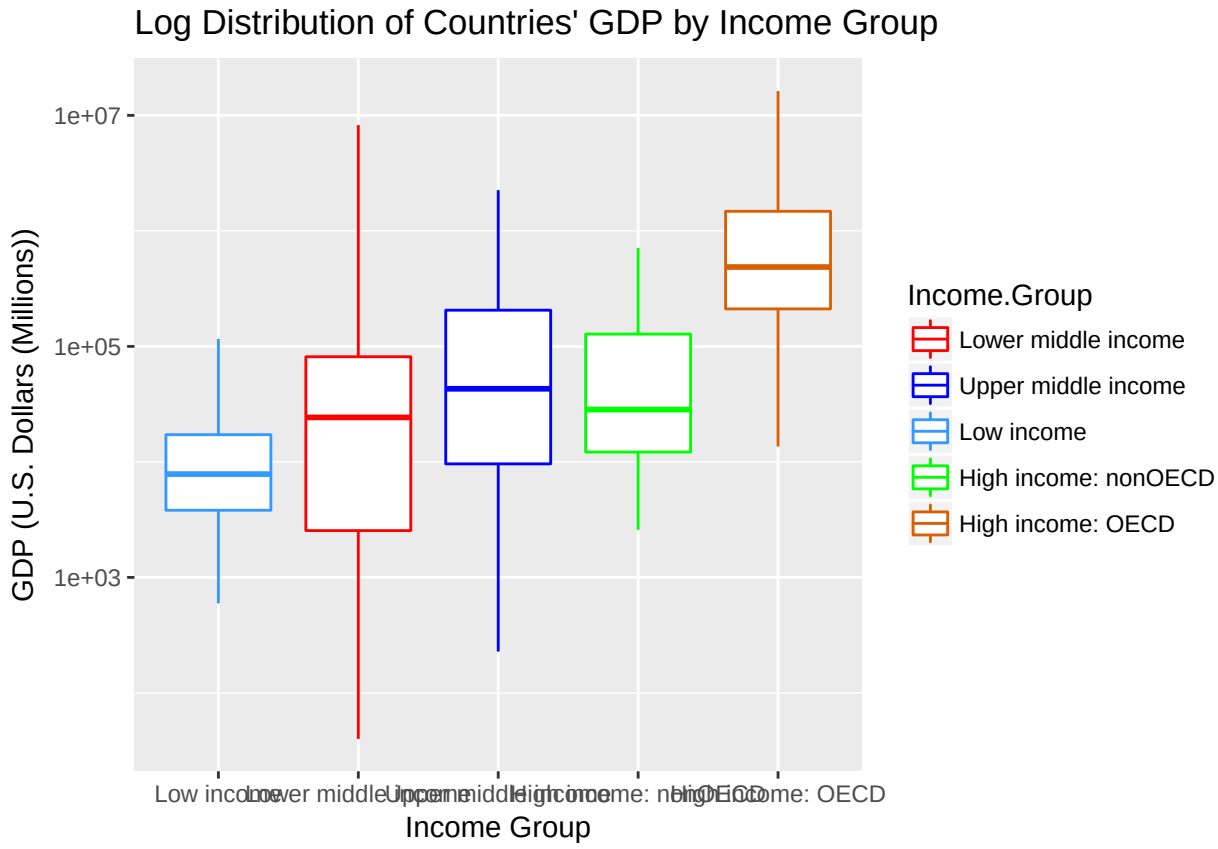
```
## [1] "The average GDP ranking of High income: nonOECD countries is: 91.91"
```

By income group, the average GDP rankings for High income: OECD countries is 32.97 and for High income: nonOECD countries is 91.91. High income OECD countries have higher GDP than that of High income nonOECD countries. Higher GDP ranking suggests that high income countries that are open to free world trade and development are more prosperous.

4) Show the distribution of GDP value for all the countries and color plots by income group. Use ggplot2 to create your plot.

Plots the distribution of GDP values for all countries and color plots by income group

```
GraphBoxPlotsByGroup(NegGDP, "#3399FF", "#FF0000","#0000FF", "#00FF00", "#D95F02")
```



Graphically by boxplot log distribution, it was expected that the median GDP of countries grouped by income group rose from low income to lower middle income to upper middle income. For high income countries, there is a discrepancy between OECD and nonOECD countries. If the country is high income but does not allow free global trade and development, they have an median GDP lower than that of upper middle income countries and about equivalent to that of lower middle income countries. Otherwise, if the high income country is an OECD member, they countinue the trend of GDP prosperity. In addition, the giant range of counties that fall into the lower middle income category suggests that the distinction of countries by income groups is not solely based on GDP qualities alone.

5) Provide summary statistics of GDP by income groups.

GDP summary statistics of countries based on income groups

```
SummaryStats(NegGDP)
```

- **High income: nonOECD**:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2584	12840	28370	104300	131200	711000

- **High income: OECD**:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
13580	211100	486500	1484000	1480000	16240000

- **Low income**:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
596	3814	7843	14410	17200	116400

- **Lower middle income**:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
40	2549	24270	256700	81450	8227000

- **Upper middle income**:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
228	9613	42940	231800	205800	2253000

From the boxplot log distribution of countries' GDP separated by income groups, the quantile distributions were plotted by range, interquantile range, and medians. The summary statistics show that the mean GDP per income group is very different than that of the median, with the mean being 0.8x, 9.5x, 4.3x, 2.6x, and 2x greater than that of the median for their respective income groups by ascending classification.

The range overlap in country GDP further suggests that countries separated by income group was not solely based on GDP. The order of mean GDP by income group is low income, high income: nonOECD, upper middle income, lower middle income, and high income: OECD, which means that there are more factors that dictate how a country is classifed into income groups than GDP alone.

6) Cut the GDP ranking into 5 separate quantile groups. Make a table versus Income.Group. How many countries are Lower middle income but among the 38 nations with highest GDP?

Creates separate quantile groups for countries based on income groups

```
GDPRankingsPerQuantByIncome(NegGDP, Quantiles, 0.2)
```

	(1,38.6]	(38.6,76.2]	(76.2,114]	(114,152]	(152,190]
High income: nonOECD	4	5	8	4	2
High income: OECD	17	10	1	1	0
Low income	0	1	9	16	11
Lower middle income	5	13	11	9	16
Upper middle income	11	9	8	8	9

Finds the countries by name based on top GDP ranking and income group

```
CountriesByGDPRankIncomeGroup(NegGDP, 38, "Lower middle income")
```

	Ranking	Economy	US Dollars (millions)	Income.Group
34	2	China	8227103	Lower middle income
78	10	India	1841710	Lower middle income
77	16	Indonesia	878043	Lower middle income
165	31	Thailand	365966	Lower middle income
51	38	Egypt, Arab Rep.	262832	Lower middle income

It was expected that low income has more countries that fall inside the higher quantile GDP rankings while that of the higher income: OECD has more that fall inside the lower GDP quantile rankings. Lower middle income has a concentration of counties that fall inside ther higher GDP rankings with some of its countries in the lower quantile GDP rankings. Upper middle income countries has an even distribution of countries in each quantile category while that of high income: nonOECD countries have countries falling in the middle GDP quantile rankings.

GDP ranking 1 to 38 is the top 20% quantile of all the nations. There are 5 lower middle income countries among the 38 nations with the highest GDP.

Conclusion:

0 and 1) As the online data set updates to include more GDP and income groups, more of the world's countries would be included to do a full-world analysis. For now, the analysis is for 189 of the 235 available countries, with 46 countries with missing data.

2) If there is a tie in GDP rankings at #12 for Grenada and St. Kitts, further alphabetical sorting is used to distinguish St. Kitts as the 13th country in ascending GDP ranking.

3) The rankings gap between the average GDP ranking of high income, OECD countries (32.97) and that of high income, nonOECD countries (91.91) is quite significant, given that the range of GDP rankings is from 1 to 189. High income OECD countries that are open to free trade and development have a higher average GDP ranking than those nonOECD countries that do not.

4) When boxplot distributions are plotted for GDPs by income group, there is some upwards trend when comparing median GDPs for low income to lower middle income to upper middle income to high income OECD countries. As stated for number 3, high income nonOECD countries cripple their GDP by not having open trade to all countries for development and its median GDP fall close to that for lower middle income. There are non-GDP factors when categorizing certain countries by income group because of the wide GDP range that the lower middle income group constitutes.

5) The summary statistics show that the mean GDP per income group is very different than that of the median, with the mean being 0.8x, 9.5x, 4.3x, 2.6x, and 2x greater than that of the median for their respective income groups by ascending classification. As stated in number 4, there is quite a bit of GDP overlap when classifying certain countries to income groups and classification of income group is not solely based on GDP.

6) There are 5 lower middle income countries among the 38 nations with the highest GDP, which constitutes the top 20% quantile of all the nations analyzed. As stated in number 4) there are factors outside of GDP that qualify certain countries to certain income classifications.

- The world data sets are observational and no causal effect could be inferenced. The country data sampled are not randomized for population inference and does not reflect data from all the nations in the world.
- Writing functions in R makes the work reproducible for future analysis and R markdown is good for documenting all the steps.

Further Work:

Future work would be to analyze country GDP per capita or per land size to see if the GDP distributions per income group would change based on those incremental factors. It would also be good to know what constitutes a country to be categorized to a certain income group and see if any of the other columns imported from world data sets could indicate more trends based on column data from other factors.