Country GDP and Educational Data Analysis (Group Project)

Group B

Lisa Street

Annahid Lee

Brian Coari

October 23, 2017

Introduction

The goal of this case study is to analyze the Income Groups of the educational data for countries around the world and compare them to the Gross Domestic Products and rankings of Gross Domestic Product for those countries.

For this study we have access to Education Data and GDP data from 2012, though some of the GDP data might come from 2010 or 2011 if the 2012 data for that country were unavailable. It is unclear which countries had data for which years, so if this is determined to be critical information we might need to go back to the data's source for clarification.

Installing Necessary Libraries

Install necessary packages if not installed:

```r
if (!is.element("dplyr", installed.packages()[,1]))
  install.packages("dplyr",repos="http://cran.rstudio.com")
if (!is.element("ggplot2", installed.packages()[,1]))
  install.packages("ggplot2",repos="http://cran.rstudio.com")
if (!is.element("stringr", installed.packages()[,1]))
  install.packages("stringr",repos="http://cran.rstudio.com")
if (!is.element("data.table", installed.packages()[,1]))
  install.packages("data.table",repos="http://cran.rstudio.com")
if (!is.element("readr", installed.packages()[,1]))
  install.packages("readr",repos="http://cran.rstudio.com")
```

Importing Data

For this analysis we have access to two data sets:

1. EDSTATS_Country.csv
    - Educational data for countries

    - Fields Applicable to this analysis:
        - Country Code: Short code of the country, primary key and foreign key to GDP.csv
        - Income Group: Primary Income Group of the country, broken down to these discrete values:
            - Low income
            - Lower middle income
            - Upper middle income
            - High income: OECD
            - High income: nonOECD

2. GDP.csv
    - GDP Information by country.
    - Raw data, processed in the Data Cleanup phase to remove blank or irrelevant rows
    - Fields Applicable to this analysis:
        - Country Code: Short code of the country, primary key and foreign key to EDSTATS_Country.csv
        - Ranking: from 1-n of the country's GDP.
        - Long Name: Long name of the country
        - GDP Ranking: GDP value for the country in a year. The data is supposed to be mostly for 2012, but the CSV file references that some of the data might come from 2010 or 2011. It is unclear which rows come from which year, which is noted in the analysis
        - Note: Rankings include only those economies with confirmed GDP estimates. Figures in italics are for 2011 or 2010.
    a. Includes Former Spanish Sahara. b. Excludes South Sudan c. Covers mainland Tanzania only. d. Data are for the area controlled by the government of the Republic of Cyprus. e. Excludes Abkhazia and South Ossetia. f. Excludes Transnistria.

```
# Download data from URL
GDPurl <-  "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv"
download.file(GDPurl, "Data\\GDP.csv", quiet=TRUE)

EduURL <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS_Country.csv"
download.file(EduURL, "Data\\EDSTATS_Country.csv", quiet=TRUE)

#read CSV input into data frames
EDSTATS_Country <- read.csv("Data\\EDSTATS_Country.csv",header=TRUE, sep=",", stringsAsFactors=FALSE)
GDP <- read.csv("Data\\GDP.csv",header=TRUE, sep=",", stringsAsFactors=FALSE)
```

Cleaning Data

In order to ensure we get correct results on our analysis some data cleanup is necessary since the EDP.CSV dataset contains many blank rows or rows that are irrelevant to this analysis. It might be possible to perform these exclusions at every calculation but the risk of a mistake is higher and the code would become much more complicated, so we will clean the code prior to analysis.

All of our cleanup is to GDP.CSV since EDSTATS_Country.csv is seemingly clean.

Cleanup is as follows:

1. Set column names to 'CountryCode','Ranking','Long

Name’,‘GDPInMillions2010_OR_2011_OR_2012’,‘note’ in order to be more readable (Purposes of these fields noted in the “Importing Data” section)

2. Deleted blank lines for country code and ranking since they did not contain meaningful data for this analysis and might throw off mean analysis.
3. Convert Ranking and GDP fields to numeric, removing any commas to ensure a good conversion

```
#Data Cleanup#
#set row names

#Remove blank row
GDP <- GDP[,-3]

names(GDP) <- c('CountryCode','Ranking','Long Name','GDPInMillions2010_OR_2011_OR_2012
','note')

#IN GDP.CSV: Deleted blank lines for country code and ranking since they did not conta
in meaningful data for this analysis and might throw off mean analysis.
GDP <- GDP[GDP$Ranking!="", ]
GDP <- GDP[GDP$CountryCode!="", ]

#convert Ranking to numeric
GDP$Ranking <- as.numeric(GDP$Ranking)

#convert GPD to numeric and remove commas
GDP$GDPInMillions2010_OR_2011_OR_2012 <- as.numeric(gsub(",","",GDP$GDPInMillions2010_
OR_2011_OR_2012))
```

Analyzing Data

It is time for our data analysis in order to answer questions on these data.

First we will merge the data based on the country shortcode and see how many of the IDs match between our two data sets:

```
#Merge EDSTATS_Country and GDP by Country Code
merged_data <- merge(EDSTATS_Country,GDP,by="CountryCode")
print(dim(merged_data))
```

```
## [1] 189  39
```

From these results we can see that we have 189 matching country codes between the data sets.

Now we will sort our merged data frame in ascending order by GDP (so United States is last). After the list is sorted we will find the 13th country in the resulting data frame:

```
#order the merged data by GDP
order.GDP <- order(merged_data$GDPInMillions2010_OR_2011_OR_2012)

#Build a new data frame with the ordered indexes
merged_data_sorted_GDP_asc <- merged_data[order.GDP, ]

#Print the long name of the 13th country in the ordered list
print(merged_data_sorted_GDP_asc[13,"Long.Name"])
```

```
## [1] "St. Kitts and Nevis"
```

The 13th country in the merged data after sorted by GDP is St. Kitts and Nevis.

Next we will find the average GDP rankings in our merged data for the “High income: OECD” and “High income: nonOECD” income groups:

```
#Form the required data frames based on Income Group
HighIncomeOECD <- merged_data[merged_data$Income.Group == 'High income: OECD',]
HighIncomeNonOECD <- merged_data[merged_data$Income.Group == 'High income: nonOECD',]

#Print means
print(paste("High Income OECD mean Ranking: ",mean(HighIncomeOECD$Ranking)))
```

```
## [1] "High Income OECD mean Ranking:  32.9666666666667"
```

```
print(paste("High Income Non OECD mean Ranking: ",mean(HighIncomeNonOECD$Ranking)))
```

```
## [1] "High Income Non OECD mean Ranking:  91.9130434782609"
```

From the means of the filtered merge data we can see that the average GDP rankings of the High Income OECD is 32.97, and the average GDP rankings of the High Income Non OECD is 91.91.

Our next analysis will have is plot the GDP for all of the countries, color-coding the points by Income Group. We will show two different plots for our analysis, one with the x-axis grouped by country and the second with the x-axis grouped by Income Group.

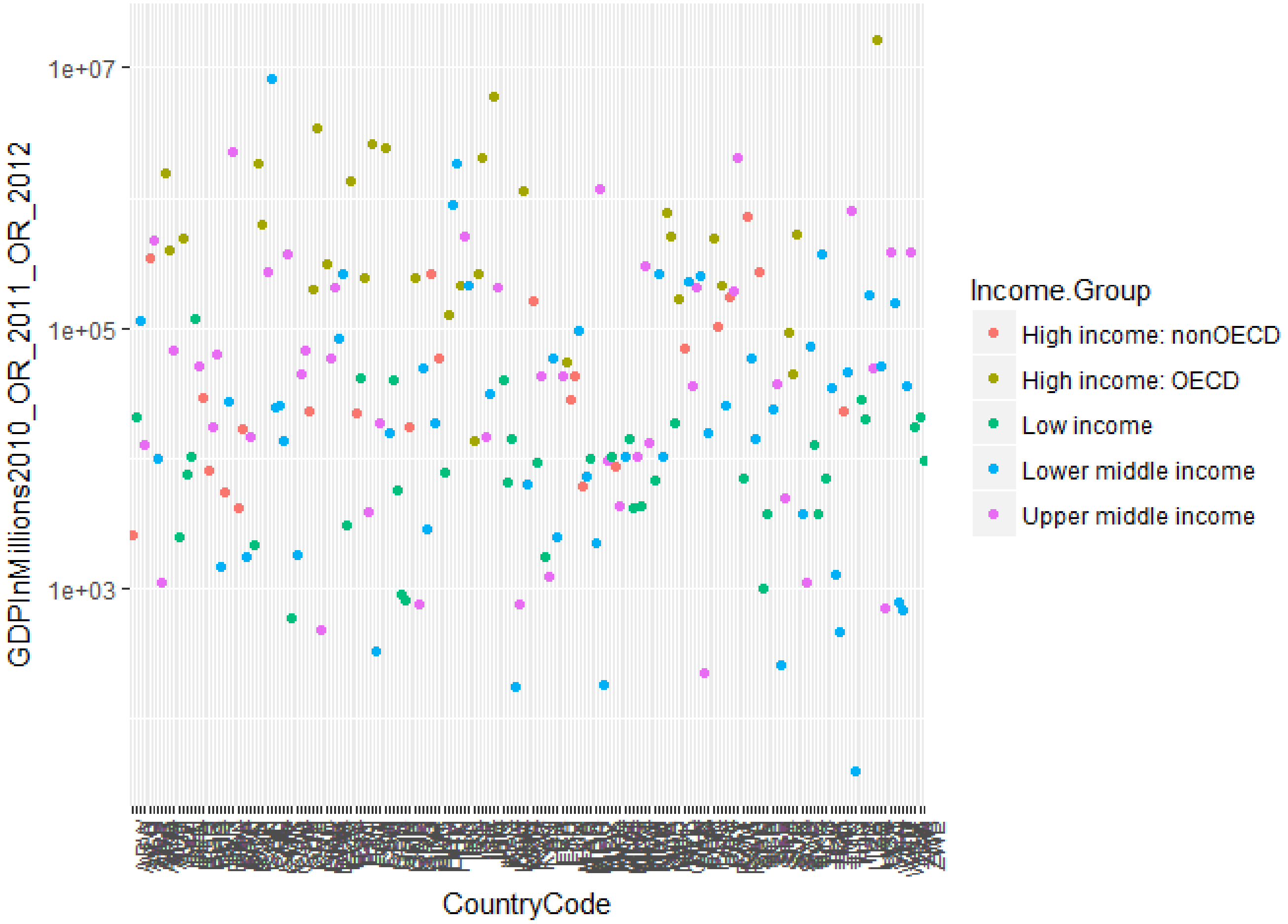
This plot shows our countries from the GDP data on the x-axis and the GDP for that county on the y-axis with the color of each point representing its income group.

```
#Include plotting library
library(ggplot2)
```

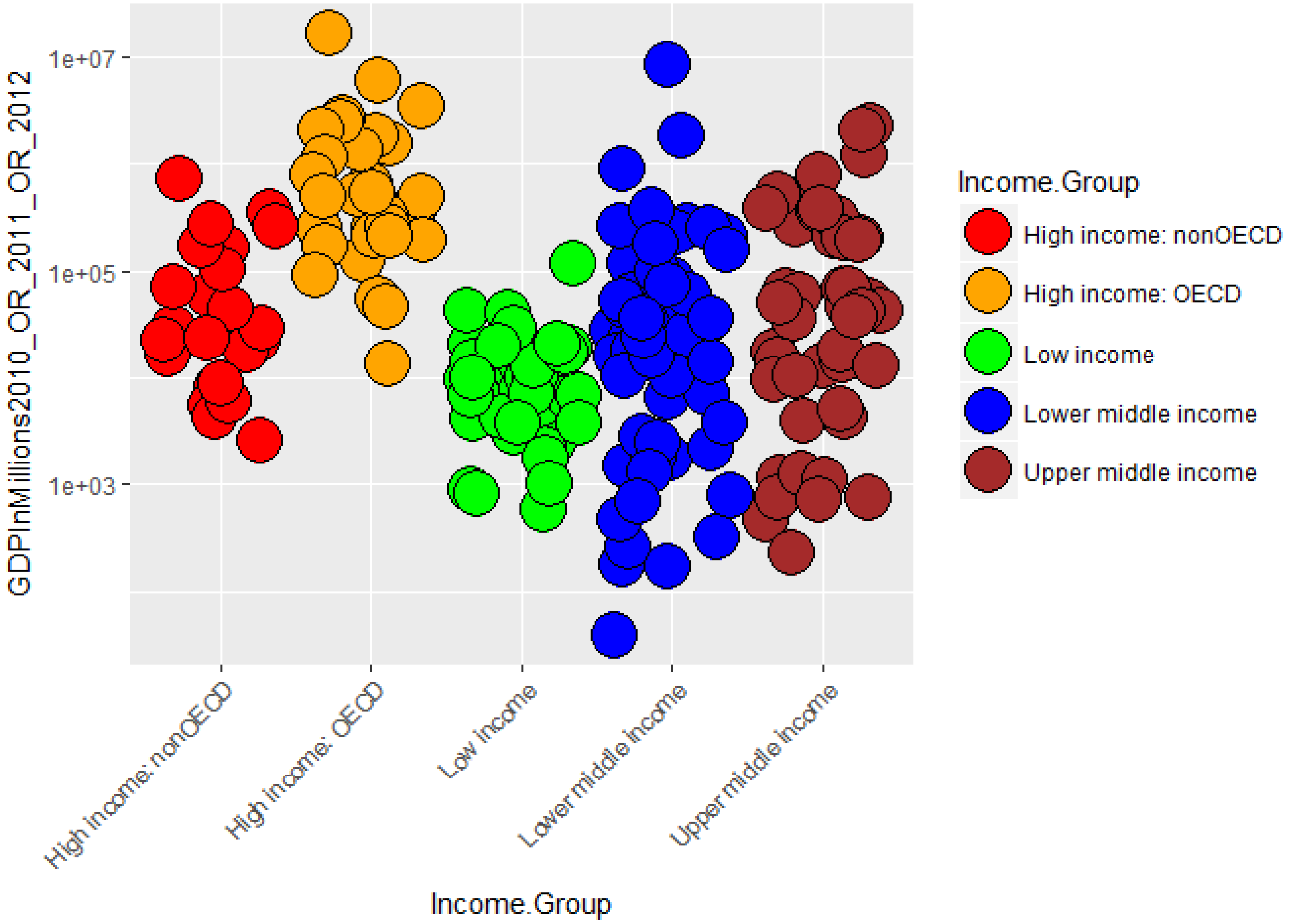
```
## Warning: package 'ggplot2' was built under R version 3.4.2
```

```
#Form the basic plot options
GdpPlot <- ggplot(merged_data, aes(x=CountryCode, y=GDPInMillions2010_OR_2011_OR_2012,
colour=Income.Group))

#Print the plot with logarithmic scaling for higher readability, each Income Group in
a different color
print(GdpPlot + geom_point() +  scale_y_log10() + theme(axis.text.x = element_text(ang
le = 90, hjust = 1)))
```



```
#print a plot grouped by Income.Group since it looks a little cleaner and might provid
e insights that are easier to interpret
print(ggplot(merged_data, aes(y = GDPInMillions2010_OR_2011_OR_2012, x=Income.Group,fi
ll=Income.Group)) + scale_y_log10()+ geom_point(pch = 21, size = 8, stat = "identity",
position=position_jitter())+ scale_fill_manual(values = c("red", "orange", "green", "b
lue","brown"), na.value = "grey50" ) + theme(axis.text.x = element_text(angle = 45, hj
ust = 1)))
```



This plot shows ourr countries from the GDP data on the x-axis and the GDP for that county on the y-axis with the color of each point representing its income group, but this time the countries are grouped by their income group from the educational data.

A design decision was made to show the y-axis in logarithmic scale since some countries had a vastly higher GDP than most countries and the detail of the spread of the lower countries was lost, making the graphs less impactful since our goal is to see if a country's income group is somehow correlated to its GDP. Logarithmic scale makes these outliers less impactful in the skewing of the plots.

In the first plot it is difficult to see any correlation between the income groups and the GDP since the data is so scattered. The second plot was created in order to help groups the data for easier comparison.

The plots show a definite trend in certain income groups to have a higher average GDP. The means for the higher income groups are higher than the upper middle income group, which is higher than the lower middle income group, which is higher than the low income group. In addition, the mean GDP for the high income countries participating in the Organisation for Economic Co-operation and Development (OECD) seem to have a higher mean GDP than the high income countries not participating in the OECD. Lastly of note is that the spread, or deviation, in the lower middle and upper middle income groups is much larger than the spread in the other groups, with the lower middle income group havig the most spread. It's difficult to derive many conclusions as to why that is, but suffice it to say that these income groups will not correlate as well with GDP.

More analysis would be necessary to draw any conclusions but the results are interesting and might warrant further investigation.

Lastly, we will cut the merged GDP rankings into 5 separate quantile groups and make a table of the rankings versus the income groups. We will use this table to tell us how many countries are Lower middle income but among the 38 nations with highest GDP:

```
#Determine the quantiles
merged_data_quantiles <- quantile(merged_data$Ranking, probs = seq(0, 1, 0.2))

#Assign the ranks into quantile groups
merged_data$Ranking_quantiles <- cut(merged_data$Ranking, breaks = merged_data_quantil
es)

#print the table of income group vs ranking quantiles
print(table(merged_data$Income.Group,merged_data$Ranking_quantiles))
```

```
## 
##                         (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
```

```
tblData <- merged_data[merged_data$Income.Group == "Lower middle income" & merged_data
$Ranking <= 38,]

print(tblData["Short.Name"])
```

```
##     Short.Name
## 34       China
## 51       Egypt
## 77   Indonesia
## 78       India
## 165   Thailand
```

This Table shows the five income groups tracked in our educational data on the x-axis versus the ranking group of the country’s GDP. Each individual cell show the number of countries in that income group and ranking group. From this table we can see that The intersection of the top ranking group “(1,38.6]” and the “Lower middle income” group is 5, therefore there are 5 countries with Lower Middle income among the 38 nations with highest GDP in our dataset. Finding those countries in the merged data allows us to see these countries are China, Egypt, Indonesia, India and Thailand.

Conclusion

Through the merging of the datasets and our analysis we were able to determine that here are 189 matching country codes between the two datasets. We were able to see that St. Kitts and Nevis has the 13th-highest GDP in our dataset, and that the average ranking for the countries assigned an income group of ‘High Income: OECD’ is 32.97 and for the income group of ‘High Income: nonOECD’ is 91.91.

Plotting and tables shows that a higher income group seems to correlate with a higher GDP. Our quantile comparison shows 5 countries that have a Lower Middle income but fall into one of the 38 counties with the highest GDP, and they are China, Egypt, Indonesia, India and Thailand.

From our analysis we can see a probable correlation between GDP and income groups, as well as strong evidence that countries that High Income countries that are part of the Organisation for Economic Co-operation and Development (OECD) have, on average, a higher ranking when it comes to GDP than High Income countries that are not part of the OECD.

Since the OECD and income groups were not randomly assigned, we cannot use these data or analyses to determine causality, only a correlation. Furthermore since these data were not randomaly attained we cannot draw any conclusions about populations as a whole, only the 189 countries in both data sets involved in these analyses.