MSDS6306CaseStudy

Priyanka Tilak

March 19, 2017

## Introduction to the Project

This document describes the Gross Domestic Product data and Educational data for 190 countries across the world. We are analyzing GDP of the countries, different income groups of these countries and their contribution in million US dollars of Gross domestic product in each country.

We have downloaded 2 data files, 1 for GDP data and Countries educational data from the following web sources:

GDP of 190 countries: <https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv>

Educational statistics of Different countries: <https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS_Country.csv>

#### Information about raw data files:

* Both the raw data files are in Comma separated format (csv)
* GDP file has NAs, missing heading, missing rows, content which is not very informative
* Education information data files has NAs
* Filtering required to perform analysis on these files

#### Objectives:

* Merge the data based on the country short code. How many of the IDs match?
* Sort the data frame in ascending order by GDP (so United States is last). What is the 13th country in the resulting data frame?
* What are the average GDP rankings for the "High income: OECD" and "High income: nonOECD" groups?
* Show the distribution of GDP value for all the countries and color plots by income group. Use ggplot2 to create your plot.
* Provide summary statistics of GDP by income groups.
* 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?

To begin with we will first set up a working directory.

#Set the present directory as working directory  
setwd(".")  
#view present working directory  
getwd()

## [1] "C:/Personal/SMU/SAS/R/R\_REPO/MSDS6306CaseStudy1"

Setting up the working directory and downloading data files from web resources

#Downloading GDP file from the WEB URL  
download.file("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv",destfile = ".\\Data\\GDP.csv", mode = "w")  
  
#Downloading Edstats\_country file from the WEB URL  
download.file("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS\_Country.csv",destfile = ".\\Data\\Edstats\_Country.csv", mode= "w")  
  
#Libraries Required  
library(downloader)  
library(plyr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)  
library(sqldf)

## Loading required package: gsubfn

## Loading required package: proto

## Loading required package: RSQLite

#### Problems with Raw Data files

Raw data files required cleaning in order to move further for analysis and merging

* GDP data file has empty rows
* GDP data file has empty column
* GDP data file needs to have fix for column headings
* GDP data file has data for 190 countries and other geographical regions which are not country
* GDP for countries has NAs and missing value
* Education data for country has many columns which are not required for the analysis or objectives
* Education data for countries has NA's and missing value

## Data Wrangling

To clean the raw data files in order to use it for merging and for analysis, Data Wrangling step is performed

#read the Downloaded files in objects  
GDP<-read.csv(".\\Data\\GDP.csv",sep=",",header=F, skip= 5)  
Edstats\_Country<-read.csv("C:\\Personal\\SMU\\SAS\\R\\R\_REPO\\MSDS6306CaseStudy1\\Data\\Edstats\_Country.csv",sep=",",header=T)  
  
library(plyr)  
#GDP data file cleanup\n  
#giving new column names and ignoring empty rows for GDP table  
colnames(GDP) <- c("CountryCode","Rank","Empty","CountryName","GDPA" )  
keep\_columns<-c("CountryCode","Rank","CountryName","GDPA")  
GDP<-GDP[keep\_columns]  
  
# Calculating missing values and NAs before cleaning up GDP data  
sum(is.na(GDP$CountryCode) | GDP$CountryCode=="")

## [1] 98

sum(is.na(GDP$Rank) | GDP$Rank=="")

## [1] 132

sum(is.na(GDP$GDPA) | GDP$GDPA =="")

## [1] 99

#Cleaning GDP file data  
GDP<-GDP[!(is.na(GDP$CountryCode) | GDP$CountryCode=="" | is.na(GDP$GDPA) | GDP$GDPA=="" | is.na(GDP$Rank) | GDP$Rank==""),]  
#view first 5 rows of the dataframe  
head(GDP)

## CountryCode Rank CountryName GDPA  
## 1 USA 1 United States 16,244,600   
## 2 CHN 2 China 8,227,103   
## 3 JPN 3 Japan 5,959,718   
## 4 DEU 4 Germany 3,428,131   
## 5 FRA 5 France 2,612,878   
## 6 GBR 6 United Kingdom 2,471,784

#Cleaning Edstats\_Country data file  
colnames(Edstats\_Country)[1] <- "CountryCode"  
colnames(Edstats\_Country)[2] <- "LongName"  
colnames(Edstats\_Country)[3] <- "IncomeGroup"  
  
# Calculating missing values and NAs before cleaning up Edstats\_Country data  
sum(is.na(Edstats\_Country$CountryCode) | Edstats\_Country$CountryCode=="")

## [1] 0

sum(is.na(Edstats\_Country$Income.Group) | Edstats\_Country$Income.Group=="")

## [1] 0

library(sqldf)  
Edstats\_Country<-sqldf("select Edstats\_Country.CountryCode,Edstats\_Country.LongName, Edstats\_Country.IncomeGroup from Edstats\_Country ")

## Loading required package: tcltk

#view first 5 rows of the dataframe  
head(Edstats\_Country)

## CountryCode LongName IncomeGroup  
## 1 ABW Aruba High income: nonOECD  
## 2 ADO Principality of Andorra High income: nonOECD  
## 3 AFG Islamic State of Afghanistan Low income  
## 4 AGO People's Republic of Angola Lower middle income  
## 5 ALB Republic of Albania Upper middle income  
## 6 ARE United Arab Emirates High income: nonOECD

## Analysis

To perform the analysis on GDP of the countries and their income groups, both cleaned data files "GDP"" and Edstats\_Country need to be merged

### Objective 1: Merge the data based on the country short code. How many of the IDs match?

Following is the Merged Table and its structure, which is merged table of GDP and Edstats\_Country

#merging 2 tables by CountryCode in to new data frame Merged\_Table  
Merged\_Table<-merge(GDP,Edstats\_Country, by="CountryCode")  
#viewving first five records of the merged data frame "Merged\_Table"   
head(Merged\_Table)

## CountryCode Rank CountryName GDPA  
## 1 ABW 161 Aruba 2,584   
## 2 AFG 105 Afghanistan 20,497   
## 3 AGO 60 Angola 114,147   
## 4 ALB 125 Albania 12,648   
## 5 ARE 32 United Arab Emirates 348,595   
## 6 ARG 26 Argentina 475,502   
## LongName IncomeGroup  
## 1 Aruba High income: nonOECD  
## 2 Islamic State of Afghanistan Low income  
## 3 People's Republic of Angola Lower middle income  
## 4 Republic of Albania Upper middle income  
## 5 United Arab Emirates High income: nonOECD  
## 6 Argentine Republic Upper middle income

#removing blank and "," values in GDPA and Rank column and making it as numeric column  
Merged\_Table$GDPA <- as.numeric(gsub(",","", Merged\_Table$GDPA))  
Merged\_Table$Rank <- as.numeric(gsub(",","", Merged\_Table$Rank))  
#structure of the Merged\_Table  
str(Merged\_Table)

## 'data.frame': 189 obs. of 6 variables:  
## $ CountryCode: Factor w/ 229 levels "","ABW","ADO",..: 2 4 5 6 7 8 9 11 12 13 ...  
## $ Rank : num 161 105 60 125 32 26 133 172 12 27 ...  
## $ CountryName: Factor w/ 229 levels ""," East Asia & Pacific",..: 20 11 16 12 216 18 19 17 21 22 ...  
## $ GDPA : num 2584 20497 114147 12648 348595 ...  
## $ LongName : Factor w/ 234 levels "American Samoa",..: 5 57 99 108 226 4 109 2 15 110 ...  
## $ IncomeGroup: Factor w/ 6 levels "","High income: nonOECD",..: 2 4 5 6 2 6 5 6 3 3 ...

### Objective 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?

Following is the result of merging and sorting to get United State as a last country in the list, and 13th country in the list

sortedMergeTable <- Merged\_Table[ order(Merged\_Table$GDPA), ]  
#sorted merged table, "sortedMergeTable""  
head(sortedMergeTable)

## CountryCode Rank CountryName GDPA  
## 173 TUV 190 Tuvalu 40  
## 92 KIR 189 Kiribati 175  
## 113 MHL 188 Marshall Islands 182  
## 137 PLW 187 Palau 228  
## 155 STP 186 São Tomé and Principe 263  
## 59 FSM 185 Micronesia, Fed. Sts. 326  
## LongName IncomeGroup  
## 173 Tuvalu Lower middle income  
## 92 Republic of Kiribati Lower middle income  
## 113 Republic of the Marshall Islands Lower middle income  
## 137 Republic of Palau Upper middle income  
## 155 Democratic Republic of São Tomé and Principe Lower middle income  
## 59 Federated States of Micronesia Lower middle income

#to create output file  
sortedMergeTable<-sortedMergeTable[(!(is.na(sortedMergeTable$IncomeGroup) |sortedMergeTable$IncomeGroup =="")),]  
write.csv(sortedMergeTable, file = ".\\Data\\sortedMergeTable.csv")  
head(sortedMergeTable)

## CountryCode Rank CountryName GDPA  
## 173 TUV 190 Tuvalu 40  
## 92 KIR 189 Kiribati 175  
## 113 MHL 188 Marshall Islands 182  
## 137 PLW 187 Palau 228  
## 155 STP 186 São Tomé and Principe 263  
## 59 FSM 185 Micronesia, Fed. Sts. 326  
## LongName IncomeGroup  
## 173 Tuvalu Lower middle income  
## 92 Republic of Kiribati Lower middle income  
## 113 Republic of the Marshall Islands Lower middle income  
## 137 Republic of Palau Upper middle income  
## 155 Democratic Republic of São Tomé and Principe Lower middle income  
## 59 Federated States of Micronesia Lower middle income

#to view 13th country  
sortedMergeTable[13,3]

## [1] St. Kitts and Nevis  
## 229 Levels: East Asia & Pacific Euro area ... Zimbabwe

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

Average Ranking for the different Income group and their respective mean in column 'X'

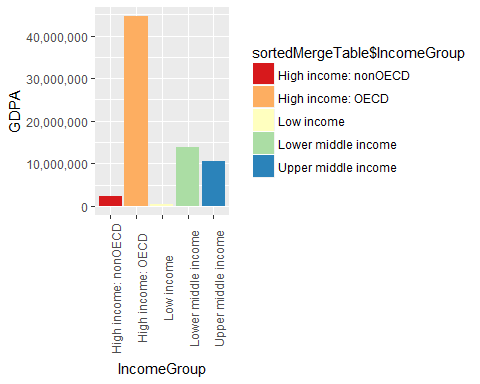
library(dplyr)  
AvgRanking<-aggregate(sortedMergeTable$Rank,list(sortedMergeTable$IncomeGroup),mean)  
print(AvgRanking)

## Group.1 x  
## 1 High income: nonOECD 91.91304  
## 2 High income: OECD 32.96667  
## 3 Low income 133.72973  
## 4 Lower middle income 107.70370  
## 5 Upper middle income 92.13333

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

Bar chart representation of GDP values of the countries and their Income group Contributions

library(ggplot2)  
library(scales)  
ggplot(sortedMergeTable,aes(x=IncomeGroup, y=GDPA,group\_by(sortedMergeTable$IncomeGroup)))+ geom\_bar(aes(fill=sortedMergeTable$IncomeGroup), stat = "identity") + scale\_fill\_brewer(palette = "Spectral") + theme( axis.text.x=element\_text(angle=90)) +scale\_y\_continuous(labels = comma)



### Objective 5: Provide summary statistics of GDP by income groups

Summary Statistics of the complete merged data set for GDP

library('dplyr')  
#Provide summary statistics of GDP by income groups.  
summary(sortedMergeTable$GDPA, by=sortedMergeTable$IncomeGroup)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 40 6972 28240 379600 205800 16240000

Summary Statistics of the complete merged data set for GDP for Lower Income group

#Statistics for Lower Income group  
lowerIncome<-sortedMergeTable[sortedMergeTable$IncomeGroup=="Low income",]  
print(summary(lowerIncome$GDPA))

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

Summary Statistics of the complete merged data set for GDP for Lower Middle Income group

#Statistics for Lower Middle Income group  
lowerMiddleIncome<-sortedMergeTable[sortedMergeTable$IncomeGroup=="Lower middle income",]  
print(summary(lowerMiddleIncome$GDPA))

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

Summary Statistics of the complete merged data set for GDP for Upper Middle Income group

#Statistics for Upper Middle Income group  
upperMiddleIncome<-sortedMergeTable[sortedMergeTable$IncomeGroup=="Upper middle income",]  
print(summary(upperMiddleIncome$GDPA))

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

Summary Statistics of the complete merged data set for GDP for High Income group

#Statistics for High Income group  
highIncomeOECD<-sortedMergeTable[sortedMergeTable$IncomeGroup=="High income: OECD",]  
print(summary(highIncomeOECD$GDPA))

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

Summary Statistics of the complete merged data set for GDP for High Income group

#Statistics for High Income Income group  
highIncomeNonOECD<-sortedMergeTable[sortedMergeTable$IncomeGroup=="High income: nonOECD",]  
summary(highIncomeNonOECD$GDPA)

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

#### Objective 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?

quantileGrps <-quantile(sortedMergeTable$Rank, probs=seq(0,1,0.2),na.rm=TRUE)  
print(quantileGrps)

## 0% 20% 40% 60% 80% 100%   
## 1.0 38.6 76.2 113.8 152.4 190.0

gdpQuantile <- cut(sortedMergeTable$Rank, breaks = quantileGrps)  
#print(gdpQuantile)  
sortedMergeTable$RankGroup <- gdpQuantile  
tableIncGrpVsRnkGrp <- table(sortedMergeTable$IncomeGroup, sortedMergeTable$RankGroup)  
tableIncGrpVsRnkGrp

##   
## (1,38.6] (38.6,76.2] (76.2,114] (114,152] (152,190]  
## 0 0 0 0 0  
## 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

Countries count which are lower middle income but among the 38 nations with the highest GDP

lower\_mid\_inc\_count <-sqldf("select count(\*) from sortedMergeTable where IncomeGroup=='Lower middle income' and Rank <39")  
print("Countries count which are lower middle income but among the 38 nations with the highest GDP")

## [1] "Countries count which are lower middle income but among the 38 nations with the highest GDP"

print(lower\_mid\_inc\_count)

## count(\*)  
## 1 5

## Conclusion

* According to the plot, the High Income OECD group is leading the total GDP as compared to all the other groups, followed surprisingly by the Lower middle income group. This indicates that the Lower Middle income group contribute significantly to the overall GDP of the world.
* Looking at the statistical summary across various groups, we can see that the median income for High Income OECD group, comprising of only 23 countries, is atleast 10 times the median income of the other groups.
* According to the plot, GDP of countries in the Lower middle income and Upper middle income are very comparable.

## References

* <https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf>
* <https://learnr.wordpress.com/2009/04/15/ggplot2-qualitative-colour-palettes/>
* <https://www.r-bloggers.com>
* <https://stackoverflow.com>
* <http://www.statmethods.net>
* <https://www.r-project.org>