MSDS 6306 Case Study 1

Author: Simrat Ahluwalia

March 19, 2017

# Case Study Introduction

Case study about GDP of the countries, their ranking and the income group they fall in. The make file is the mother file to call other sources to execute the complete analysis done on this case study.

### GDP (Gross Domestic Product)

**GDP** is the combination of three individual terms Gross, Domestic and Product commonly know as Gross Domestic Product (GDP) Individually these three terms are summarized as:

**GROSS**: The depreciation in the capital assets of the country, occurred during the year is inclusive.

**DOMESTIC**: Domestic implies, produced within the geographical boundaries.

**PRODUCT**: The final goods and services. Final implies that intermediate goods are not taken into account.

In other words , in a given period of time GDP is the total monetary value of the final goods and services which are produced within the geographical boundaries of any country.

**Data files used for the analysis were downloaded from:**

<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv> <https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS_Country.csv>

### \* The purpose of this study is to clean, merge and analyze these data files to answer following questions:

* Merge the data based on the country short code and find how many of the IDs match?
* Sort the merged file in ascending order by GDP so that United States comes at the last and to find the 13th country from this merged file , since the file is sorted in ascending order by GDP value (USD) , 13th country will be considered from the top of the sorted data file .
* The average GDP rankings for the countries which comes under "High income: OECD" and "High income: nonOECD" groups.
* Display the distribution of GDP value for all the countries by income group using tools such as ggplot.
* 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 highestGDP?

After doing the complete cleanup and analysis of the case study, it has been analyzed that there are **189** countries with complete information about the relevant data which was considered for this analysis. For example their **Country code, GDP ranking, GDP in millions of US dollars, Country name and the Income group** the country falls in out of the **5** groups category.

**> 13th country from the top of the merged data file in ascending order is -> St. Kitts and Nevis**

**> Average GDP rankings for the High income: OECD group = 32.96667**

**> Average GDP rankings for the High income: nonOECD group = 91.91304**

**> There are 5 countries which are in Lower middle income group but among the 38 nations with highest GDP's**

**> Further data between income group of countries and their GDP value has been plotted for visual reference.**

* This visual plot suggests that the countries which are in Highest Income:OECD group have higest GDP values

## Packages included for Data Analysis

require(ggplot2)  
require(RCurl)  
require(plyr)  
require(scales)  
library(RCurl)  
library(dplyr)  
library(reshape2)  
library(plyr)  
library(tcltk)  
library(sqldf)

##### *Load\_data.R script will retrieve the Raw data files from the internet and upload them in the projects Data folder for further analysis.*

*source("load\_data.R")*

# Load Data

First of all both the data files will be downloaded from internet and will be stored in a folder called 'Data' witin the project.

#Setting the working directory as current directory  
setwd('.')  
#Displaying the current working directory  
getwd()

## [1] "C:/Users/sahluwalia/Desktop/Simrat/1-SAP-Proj-Mngt-Data-Sc/Data-Science/SMU-Docs/2.Mahesh-MSDS-6306-Doing-Data-Science/Case-Study-1/GitHub\_Project/MSDS-6306-Case\_Study\_1"

# Saving data files URL's in variables.  
gdpURL <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv"  
CountryURL <-"https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS\_Country.csv"  
  
# Downloading GDP data file and Country Stats data csv files.  
download.file(gdpURL, destfile = "./Raw\_Data/gdpData.csv" )  
download.file(CountryURL, destfile = "./Raw\_Data/CountryData.csv")  
  
# Reading getdata-2Fdata-2FGDP.csv into gdp data frame  
gdp <- read.csv(".\\Raw\_Data\\gdpData.csv")  
#Reading getdata\_data\_EDSTATS\_Country.csv into country data frame  
country <- read.csv(".\\Raw\_Data\\CountryData.csv")

# 

# Clean and merge data

Clean up file will do the merge and data cleanup of both the data files. Clean up includes removing the records with null values from columns which are most relevent for the analysis of this case study, such as CountrtyCode, Ranking, Economy\_Name, USD\_in\_Millions, Income\_Group.

|  |
| --- |
| Data files contained many blank columns, those extra columns had no relevance thus some of them were removed from the data sets. |
| Following blank columns were removed from gdpData data file -> : X.1, X.4, X.5, X.6, X.7 and X.8 |

|  |
| --- |
| Few columns in gdp data file were renamed so the column name can make more sense while reading the data: |
| column 'X' was renamed to Countrycode |
| Column 'Gross.domestic.product2012' was renamed to Ranking |
| Column 'X.2' was renamed to Economy\_name |
| Column 'X.3' was renamed to USD\_in\_Millions |

|  |
| --- |
| Following column was renamed in CountryData data file: |
| Column 'Income.Group' was renamed to Income\_group |
| gdpData and CountryData files were merged into one file called mergedata1 and |

|  |
| --- |
| Following columns were used from the merged file ( mergedata1) for further data analysis |
| CountryCode |
| Ranking |
| Economy\_Name |
| USD\_in\_Millions ( Country GDP ) |
| Income\_Group |

*source("clean\_and\_merge\_data.R")*

#------------------Clean Up and merge raw data files ----------------------  
  
# Deleting X.1,X.4,X.5,X.6,X.7 and X.8 blank columns from gdp dataframe.  
gdp <- subset(gdp,select =-c(X.1,X.4,X.5,X.6,X.7,X.8))  
  
# Replacing following column name to give them more meaningful names  
  
# Replacing the blank column name with 'CountryCode'  
colnames(gdp)[1] <- 'CountryCode'  
# Replacing the 2nd column name with 'Ranking'  
colnames(gdp)[2] <- 'Ranking'  
# Replacing Economy with Economy\_name  
colnames(gdp)[3] <- 'Economy\_Name'  
# Replacing US.dollars. with USD\_in\_Millions  
colnames(gdp)[4] <- 'USD\_in\_Millions'  
  
# Income.Group as Income\_Group  
colnames(country)[3] <- 'Income\_Group'  
  
# Merged data frames 'gdp' and 'country' on columan 'CountryCode' [ key (by) = CountryCode ]  
  
# 1 -> Merged data frame of above 2 data frames, key (by) = CountryCode   
mergedata1 <- merge(x=gdp, y=country, by = "CountryCode", all = TRUE, incomparables = TRUE )  
  
# Removing records having 'NA' or spaces from columns CountryCode and Ranking   
mergedata1 <- mergedata1[!(is.na(mergedata1$CountryCode)| mergedata1$CountryCode=="" | is.na(mergedata1$Ranking) | mergedata1$Ranking==""),]  
cat("\n")

### Data was merged based on the country shortcode

### Q 1 > Merge the data based on the country shortcode. How many of the IDs match?

cat(" Data was merged based on the country shortcode and total of ", as.vector(dim(mergedata1))[1] ,"IDs were matched \n")

## Data was merged based on the country shortcode and total of 190 IDs were matched

# no\_Income\_Group <- sqldf("select CountryCode,Ranking,Economy\_Name,USD\_in\_Millions,Income\_Group from mergedata1 where Income\_Group is NULL")  
  
# To find if there are any countries with no Income Group associated to them.  
cat(" Following Country(s) does not have any Income Groups associated with them \n\n")

## Following Country(s) does not have any Income Groups associated with them

sqldf("select CountryCode,Ranking,Economy\_Name,USD\_in\_Millions,Income\_Group from mergedata1 where Income\_Group is NULL")

## Warning: Quoted identifiers should have class SQL, use DBI::SQL() if the  
## caller performs the quoting.

## CountryCode Ranking Economy\_Name USD\_in\_Millions Income\_Group  
## 1 SSD 131 South Sudan 10,220 <NA>

### Ans 1 > After cleaning up the merged data file, the number of Id's which matched were 189

# Removing records having 'NA' or spaces from column Income Group  
mergedata1 <- mergedata1[!(is.na(mergedata1$Income\_Group)| mergedata1$Income\_Group==""),]  
cat(" After cleaning the data with missing values in Country Code, Ranking, Economy Name, GDP(in USD) and Income Group , -> ", as.vector(dim(mergedata1))[1] ,"IDs were matched \n")

## After cleaning the data with missing values in Country Code, Ranking, Economy Name, GDP(in USD) and Income Group , -> 189 IDs were matched

#convert Ranking and US.dollars. from factor to numeric  
mergedata1$Ranking <- as.numeric(gsub(",", "", mergedata1$Ranking))  
mergedata1$USD\_in\_Millions <- as.numeric(gsub(",", "", mergedata1$USD\_in\_Millions))  
  
# dimensions of the cleaned data (mergedata1)  
dim(mergedata1)

## [1] 189 34

cat("\n")

cat(" Data was merged based on the country shortcode and total of ", as.vector(dim(mergedata1))[1] ,"IDs were matched \n")

## Data was merged based on the country shortcode and total of 189 IDs were matched

## Data analysis

### Q 2 [a] > Sort the data frame in ascending order by GDP (so United States is last).

### [b] What is the 13th country in the resulting data frame?

#--- Q2   
# Sort the data frame in ascending order by GDP (so United States is last).  
# Find the 13th country (descending order)  
  
mergedata1 <- mergedata1[order(mergedata1$USD\_in\_Millions, decreasing = FALSE, na.last=NA),]  
  
# After data cleanup and sorting on USD\_in\_Millions (countries GDP in millions),   
# here is the snapshot of the selective cleaned data.  
# Top 10 records of the cleaned data in decreasing order of their GDP and Rank  
# Their Country code, Rank, Name, GDP In Millions and their Income Group (in decreasing order of their GDP and Rank)

### Ans 2 [a]> Merged data got sorted in ascending order of GDP (United States is last)

### Below you can see the sorted data from the top and bottom of the data file. .

headofmergedata1 <- head(mergedata1[,c("CountryCode","Ranking","Economy\_Name","USD\_in\_Millions","Income\_Group")],10)  
print(headofmergedata1)

## CountryCode Ranking Economy\_Name USD\_in\_Millions  
## 310 TUV 190 Tuvalu 40  
## 208 KIR 189 Kiribati 175  
## 236 MHL 188 Marshall Islands 182  
## 265 PLW 187 Palau 228  
## 290 STP 186 São Tomé and Principe 263  
## 171 FSM 185 Micronesia, Fed. Sts. 326  
## 306 TON 184 Tonga 472  
## 154 DMA 183 Dominica 480  
## 144 COM 182 Comoros 596  
## 325 WSM 181 Samoa 684  
## Income\_Group  
## 310 Lower middle income  
## 208 Lower middle income  
## 236 Lower middle income  
## 265 Upper middle income  
## 290 Lower middle income  
## 171 Lower middle income  
## 306 Lower middle income  
## 154 Upper middle income  
## 144 Low income  
## 325 Lower middle income

cat("\n")

# Bottom 10 records with their Country code, Rank, Name, GDP In Millions and their Income Group   
# File is sorted in the decreasing order of their GDP and ranking thats why highest GPS and Rank is at the bottom.  
tailofmergedata1 <- tail(mergedata1[,c("CountryCode","Ranking","Economy\_Name","USD\_in\_Millions","Income\_Group")],10)  
print(tailofmergedata1)

## CountryCode Ranking Economy\_Name USD\_in\_Millions  
## 194 IND 10 India 1841710  
## 200 ITA 9 Italy 2014670  
## 275 RUS 8 Russian Federation 2014775  
## 129 BRA 7 Brazil 2252664  
## 173 GBR 6 United Kingdom 2471784  
## 169 FRA 5 France 2612878  
## 152 DEU 4 Germany 3428131  
## 203 JPN 3 Japan 5959718  
## 139 CHN 2 China 8227103  
## 316 USA 1 United States 16244600  
## Income\_Group  
## 194 Lower middle income  
## 200 High income: OECD  
## 275 Upper middle income  
## 129 Upper middle income  
## 173 High income: OECD  
## 169 High income: OECD  
## 152 High income: OECD  
## 203 High income: OECD  
## 139 Lower middle income  
## 316 High income: OECD

# Writing output to a CSV file  
write.csv(mergedata1, file = ".\\Clean\_data\\mergedata1.csv")  
  
#--How many total Ranking Countries ?  
cat (" Total number of Ranking countries are -> ", nrow(mergedata1),"\n")

## Total number of Ranking countries are -> 189

### Ans 2 [b] > 13th country from the top of the merged data file in ascending order is -> St. Kitts and Nevis

#--Which Country is the 13th Country from top decreasing order  
cat(" 13th Country in the cleaned data file with decreasing order in GDP and Ranking is -> ",as.vector(mergedata1$Economy\_Name[13])," \n")

## 13th Country in the cleaned data file with decreasing order in GDP and Ranking is -> St. Kitts and Nevis

cat("\n")

#--- Q3   
# Average GDP rankings for the "High income: OECD" and "High income: nonOECD" groups  
# Sub setting highIncomeOECD group from mergedata1   
highIncomeOECD <- subset(mergedata1,mergedata1$Income\_Group=="High income: OECD")  
  
# Storing Mean of highIncomeOECD$Ranking  
highIncomeOECDRanking <-mean(highIncomeOECD$Ranking)  
  
# Sub setting highIncomeNonOECD group from mergedata1  
highIncomeNonOECD <- subset(mergedata1,mergedata1$Income\_Group=="High income: nonOECD")

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

### Ans 3 > Average GDP rankings for the "High income: OECD" and "High income: nonOECD" groups are following:

#### Average GDP rankings for the High income: OECD group = 32.96667

#### Average GDP rankings for the High income: nonOECD group = 91.91304

# Storing Mean of highIncomeNonOECD$Ranking  
highIncomeNonOECDRanking<-mean(highIncomeNonOECD$Ranking)  
cat(" Average GDP rankings for the High income: OECD group = ",highIncomeOECDRanking,"\n")

## Average GDP rankings for the High income: OECD group = 32.96667

cat(" Average GDP rankings for the High income: nonOECD group = ",highIncomeNonOECDRanking,"\n\n")

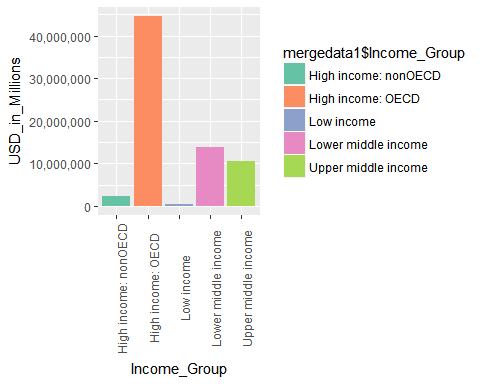
## Average GDP rankings for the High income: nonOECD group = 91.91304

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

### Ans 4 > Below is the distribution of GDP value for all the countries and color plots by income group.

#### HighIncome:OECD group has the highest distribution of GDP values in USD millions.

#--- Q4  
# Plot the GDP for all of the countries. Used ggplot2 to color plot by Income Group.  
# There was one record with Income\_Groups column has "NA" in it   
# below syntax is to remove the "NA" and create groupwise bar chart [different color represent different group]  
  
ggplot(mergedata1,aes(x=Income\_Group, y=USD\_in\_Millions,group\_by(mergedata1$Income\_Groups)))+ geom\_bar(aes(fill=mergedata1$Income\_Group), stat = "identity") + scale\_fill\_brewer(palette="Set2") + theme( axis.text.x=element\_text(angle=90)) +scale\_y\_continuous(labels = comma)



### Q 5 > Provide summary statistics of GDP by income groups.

### Ans 5 > There are 5 income groups , below are the summary statictics of each group

#--- Q5  
# Summary statistics of GDP by Income Group  
# Subsetting the Income Groups to generate summary statistics of each Group  
low\_income <- subset(mergedata1,mergedata1$Income\_Group=="Low income")  
lower\_middle\_income <- subset(mergedata1,mergedata1$Income\_Group=="Lower middle income")  
upper\_middle\_income <- subset(mergedata1,mergedata1$Income\_Group=="Upper middle income")

### Let's start with the aggregated summary statistics of all the five Income Groups:

# Aggregated summary statistics of all the five Income Groups.  
cat(" Aggregate Group by GDP in USD millions \n")

## Aggregate Group by GDP in USD millions

aggregate(mergedata1$USD\_in\_Millions, by=list(mergedata1$Income\_Group),FUN=sum)

## Group.1 x  
## 1 High income: nonOECD 2400046  
## 2 High income: OECD 44517514  
## 3 Low income 533199  
## 4 Lower middle income 13859828  
## 5 Upper middle income 10433153

highIncomeOECD\_Summary <- summary(highIncomeOECD$USD\_in\_Millions, digits = 10)  
highIncomeNonOECD\_summary <- summary(highIncomeNonOECD$USD\_in\_Millions, digits = 10)  
low\_income\_Summary <- summary(low\_income$USD\_in\_Millions, digits = 10)  
lower\_middle\_income\_Summary <- summary(lower\_middle\_income$USD\_in\_Millions, digits = 10)  
upper\_middle\_income\_Summary <-summary(upper\_middle\_income$USD\_in\_Millions, digits = 10)

### Summary statistics for highIncomeOECD income group are following :

# Summary statistics for highIncomeOECD income group.  
cat("Summary statistics for High Income OECD group\n\n"," Lowest GDP in Millions (USD) in high income OECD groups is ->",as.vector(highIncomeOECD\_Summary)[1]," M\n"," Mean GDP in Millions (USD) in high income OECD groups is ->",as.vector(highIncomeOECD\_Summary)[4]," M\n"," Median GDP in Millions (USD) in high income OECD groups is ->",as.vector(highIncomeOECD\_Summary)[3]," M\n"," Highest GDP in Millions (USD) in high income OECD groups is ->",as.vector(highIncomeOECD\_Summary)[6],"M\n")

## Summary statistics for High Income OECD group  
##   
## Lowest GDP in Millions (USD) in high income OECD groups is -> 13579 M  
## Mean GDP in Millions (USD) in high income OECD groups is -> 1483917 M  
## Median GDP in Millions (USD) in high income OECD groups is -> 486528.5 M  
## Highest GDP in Millions (USD) in high income OECD groups is -> 16244600 M

### Summary statistics for highIncomeNonOECD income group are following:

# Summary statistics for highIncomeNonOECD income group.  
cat("Summary statistics for High Income Non OECD group\n\n"," Lowest GDP in Millions (USD) in high income Non OECD groups is ->",as.vector(highIncomeNonOECD\_summary)[1]," M\n"," Mean GDP in Millions (USD) in high income Non OECD groups is ->",as.vector(highIncomeNonOECD\_summary)[4]," M\n"," Median GDP in Millions (USD) in high income Non OECD groups is ->",as.vector(highIncomeNonOECD\_summary)[3]," M\n"," Highest GDP in Millions (USD) in high income Non OECD groups is ->",as.vector(highIncomeNonOECD\_summary)[6]," M\n")

## Summary statistics for High Income Non OECD group  
##   
## Lowest GDP in Millions (USD) in high income Non OECD groups is -> 2584 M  
## Mean GDP in Millions (USD) in high income Non OECD groups is -> 104349.8 M  
## Median GDP in Millions (USD) in high income Non OECD groups is -> 28373 M  
## Highest GDP in Millions (USD) in high income Non OECD groups is -> 711050 M

### Summary statistics for low income group are following:

# Summary statistics for low income group.  
cat("Summary statistics for Low Income group \n\n"," Lowest GDP in Millions (USD) in low income groups is ->",as.vector(low\_income\_Summary)[1]," M\n"," Mean GDP in Millions (USD) in low income groups is ->",as.vector(low\_income\_Summary)[4]," M\n"," Median GDP in Millions (USD) in low income groups is ->",as.vector(low\_income\_Summary)[3]," M\n"," Highest GDP in Millions (USD) in low income groups is ->",as.vector(low\_income\_Summary)[6]," M\n")

## Summary statistics for Low Income group   
##   
## Lowest GDP in Millions (USD) in low income groups is -> 596 M  
## Mean GDP in Millions (USD) in low income groups is -> 14410.78 M  
## Median GDP in Millions (USD) in low income groups is -> 7843 M  
## Highest GDP in Millions (USD) in low income groups is -> 116355 M

### Summary statistics for lower middle income group are following:

# Summary statistics for lower middle income group.  
cat("Summary statistics for Lower Middle Income group \n\n"," Lowest GDP in Millions (USD) in lower middle income groups is ->",as.vector(lower\_middle\_income\_Summary)[1]," M\n"," Mean GDP in Millions (USD) in lower middle income groups is ->",as.vector(lower\_middle\_income\_Summary)[4]," M\n"," Median GDP in Millions (USD) in lower middle income groups is ->",as.vector(lower\_middle\_income\_Summary)[3]," M\n"," Highest GDP in Millions (USD) in lower middle income groups is ->",as.vector(lower\_middle\_income\_Summary)[6]," M\n")

## Summary statistics for Lower Middle Income group   
##   
## Lowest GDP in Millions (USD) in lower middle income groups is -> 40 M  
## Mean GDP in Millions (USD) in lower middle income groups is -> 256663.5 M  
## Median GDP in Millions (USD) in lower middle income groups is -> 24272 M  
## Highest GDP in Millions (USD) in lower middle income groups is -> 8227103 M

### Summary statistics for upper middle income group are following:

# Summary statistics for upper middle income group.  
cat("Summary statistics for Upper Middle Income group\n\n"," Lowest GDP in Millions (USD) in upper middle income groups is ->",as.vector(upper\_middle\_income\_Summary)[1]," M\n"," Mean GDP in Millions (USD) in upper middle income groups is ->",as.vector(upper\_middle\_income\_Summary)[4]," M\n"," Median GDP in Millions (USD) in upper middle income groups is ->",as.vector(upper\_middle\_income\_Summary)[3]," M\n"," Highest GDP in Millions (USD) in upper middle income groups is ->",as.vector(upper\_middle\_income\_Summary)[6]," M\n")

## Summary statistics for Upper Middle Income group  
##   
## Lowest GDP in Millions (USD) in upper middle income groups is -> 228 M  
## Mean GDP in Millions (USD) in upper middle income groups is -> 231847.8 M  
## Median GDP in Millions (USD) in upper middle income groups is -> 42945 M  
## Highest GDP in Millions (USD) in upper middle income groups is -> 2252664 M

### Q 6 [a]> Cut the GDP ranking into 5 separate quantile groups. Make a table versus Income.Group.

### [b] How many countries are Lower middle income but among the 38 nations with highest GDP?

### Ans 6 [a] > GDP Ranking was cut into 5 separate quantile and these quantile groups were assigned to each country they fall in.

#### A new data column "RankingGroup" is introduced in the cleaned data file,

#### which can eventually help us to generate statistics of which country income group falls in which quantile and can give us income group vs quantile range statistics.

#--- Q6.   
# a] Cut GDP ranking into 5 separate quantile groups. Table versus Income.Group.  
  
gdpQuantile <- quantile(mergedata1$Ranking, probs = seq(0,1,0.2),na.rm=TRUE)  
gdpQuantile

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

cutgdpQuantile <- cut(mergedata1$Ranking, breaks = gdpQuantile)  
mergedata1$RankingGroup <- cutgdpQuantile  
  
merge\_RankingGroup <- table(mergedata1$Income\_Group, mergedata1$RankingGroup )  
merge\_RankingGroup

##   
## (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

### Ans 6 [b] > Total number of countries with Lower middle income but among the 38 nations with highest GDP are -> 5

# b] Find Countries that are in Lower middle income but among the first 38 nations with highest GDP  
# Number under column name = TRUE are the number of countries which are in Lower middle income group  
  
table(mergedata1$Income\_Group, mergedata1$Income\_Group=="Lower middle income" & mergedata1$Ranking <39)

##   
## FALSE TRUE  
## 0 0  
## High income: nonOECD 23 0  
## High income: OECD 30 0  
## Low income 37 0  
## Lower middle income 49 5  
## Upper middle income 45 0

count\_lower\_\_middle\_income\_group\_countries<-table(mergedata1$Income\_Group, mergedata1$Income\_Group=="Lower middle income" & mergedata1$Ranking <39)  
  
cat(" Total number of countries in Lower middle income group are -> ", count\_lower\_\_middle\_income\_group\_countries[5,2],"\n")

## Total number of countries in Lower middle income group are -> 5

### Countries with Lower middle income but among the 38 nations with highest GDP are following:

cat(" Following are the countries which are in the lower middle income group but among the 38 nations with highest GDP :")

## Following are the countries which are in the lower middle income group but among the 38 nations with highest GDP :

lower\_middle\_income\_group\_countries <-sqldf("select CountryCode,Ranking,Economy\_Name,USD\_in\_Millions,Income\_Group from mergedata1 where Income\_Group=='Lower middle income' and Ranking <39")  
print(lower\_middle\_income\_group\_countries)

## CountryCode Ranking Economy\_Name USD\_in\_Millions Income\_Group  
## 1 EGY 38 Egypt, Arab Rep. 262832 Lower middle income  
## 2 THA 31 Thailand 365966 Lower middle income  
## 3 IDN 16 Indonesia 878043 Lower middle income  
## 4 IND 10 India 1841710 Lower middle income  
## 5 CHN 2 China 8227103 Lower middle income

## Conclusion:

#### Countries in this case study are basically divided into five different income categories based upon their total monetary value of the final goods and services which are produce within the geographical boundaries of their country.

* High Income OECD group has least number of countries (23) present in that group and they have the highest combined GDP as a group.
* Countries which are in Low Income group, their combined GDP is only 12% in comparison to the High IncomeOECD group
* Countries in the Lower middle income and Upper middle income groups are very close in terms of GDP based on the number of countries and GDP value in total in their respective group.
* Looking at the bar graph it looks like a right tailed distribution and does not suggest a normal distribution because of the extreme outliers in the High Income OECD group.

## References:

* <https://www.r-bloggers.com/>
* <https://cran.r-project.org/web/packages/available_packages_by_name.html>
* <https://www.r-bloggers.com/make-r-speak-sql-with-sqldf/>
* <http://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html>
* <https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet>