# Case Study 1

*Hieu Nguyen*

*March 23, 2017*

## Introduction

The following documents describes the analysis performed on the data set of GDP and EdStats. The two data sets were cleaned and then merged on the matching country code. There are 4 most useful colums which are CountryCode, Rank, Country.Name, GDP.Value.

*Data set can be found here:*

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

*Packages used:*

install the plyr, ggplot2 and Hmisc packages and load in to R. library(plyr) library(ggplot2) library(Hmisc)

*Download the files and format data*

```
file1 <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv"
file2 <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS_Country.csv"
download.file(file1, destfile = "GDP.csv")
download.file(file2, destfile = "EDU.csv")
gdpdata <- read.csv("GDP.csv", skip = 4, nrows = 190)
edudata <- read.csv("EDU.csv")
gdpdata <- gdpdata[, c(1,2,4,5)]
colnames(gdpdata) <- c("CountryCode", "Rank", "Country.Name", "GDP.Value")
gdpdata$GDP.Value <- as.numeric(gsub(",", "", gdpdata$GDP.Value))
```

## Analysis

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

```
CombineData <- merge(gdpdata, edudata, by.x = "CountryCode", by.y = "CountryCode")
dim(CombineData)[1]
```

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

### 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?

```
arrange(CombineData, desc(Rank))[13,3]
```

```
## [1] St. Kitts and Nevis
## 190 Levels: Afghanistan Albania Algeria Angola ... Zimbabwe
```

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

```
### average GDP rankings for the "High income: OECD"
mean(subset(CombineData, Income.Group %in% "High income: OECD", select = c(Rank))$Rank)
```

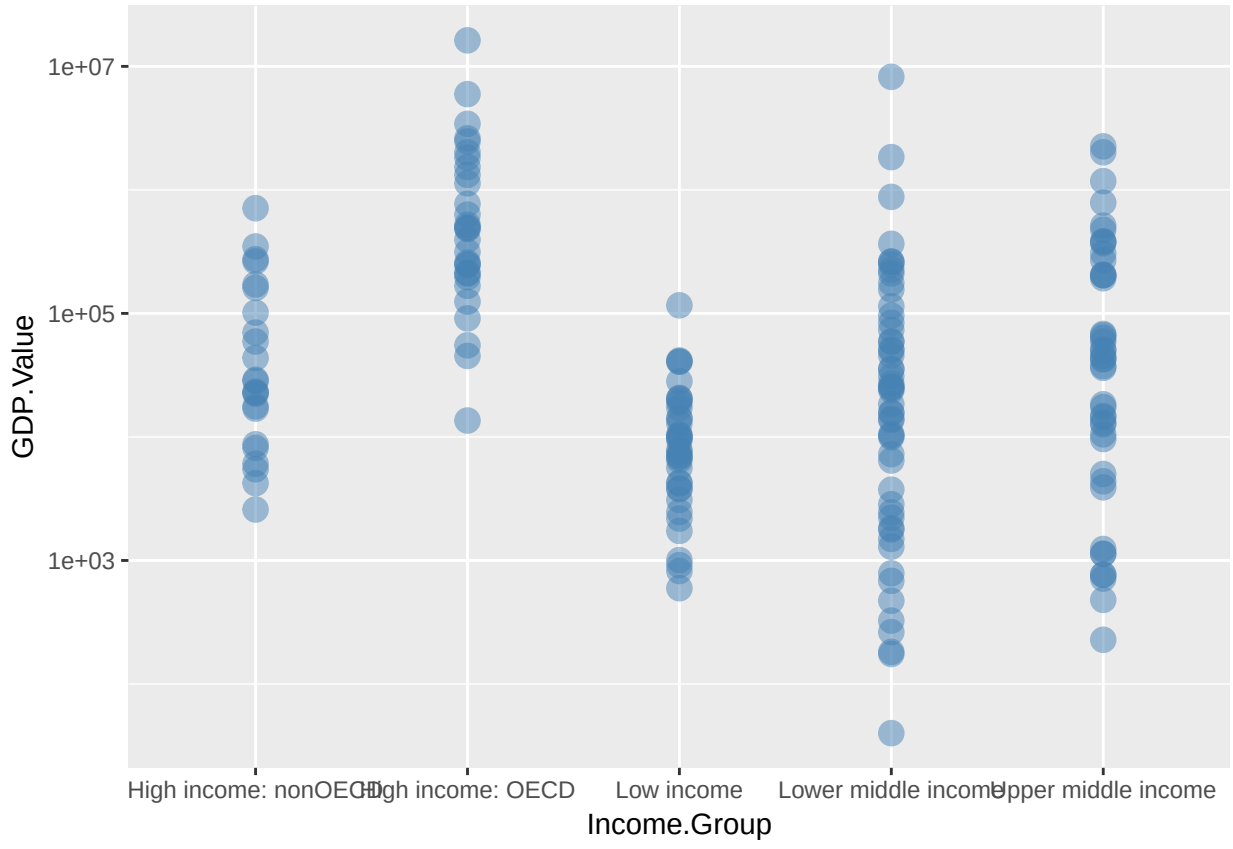
```
## [1] 32.96667
```

```
### average GDP rankings for the "High income: nonOECD"
mean(subset(CombineData, Income.Group %in% "High income: nonOECD", select = c(Rank))$Rank)
```

```
## [1] 91.91304
```

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

```
ggplot(CombineData,aes(x =Income.Group, y = GDP.Value)) + scale_y_log10() + geom_point(color = "steelblu
```




### 5) Provide summary statistics of GDP by income groups.

```
tapply(CombineData$GDP.Value, CombineData$Income.Group, summary)
```

```
## [[1]]
## NULL
```

```
##
## $`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
```

### 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?

```
CombineData$Rank.Groups = cut2(CombineData$Rank, g = 5)
table(CombineData$Income.Group, CombineData$Rank.Groups)
```

```
##
##                          [  1, 39) [ 39, 77) [ 77,115) [115,154) [154,190]
##                                  0         0         0         0         0
##   High income: nonOECD           4         5         8         5         1
##   High income: OECD             18        10         1         1         0
##   Low income                     0         1         9        16        11
##   Lower middle income            5        13        12         8        16
##   Upper middle income           11         9         8         8         9
```

## Conclusion

There were 189 matching countries between the two data sets. The US had the highest GDP value, Tuvula had the lowest GDP value. The average GDP ranking for the "High income: nonOECD" group was 91.9, while the average GDP ranking for the "High income: OECD"was 32.97. This shows big diferrent between these two groups. From the plot on question four, I found that there were normal GDP distributions for all income groups.

## References


[1]https://rpubs.com/Araya1982/191841

[2]https://cran.r-project.org/web/packages/plyr/plyr.pdf

[3]https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf
[4]https://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf