Case Study I

Damon Resnick

October 21, 2016

Introduction

```
    Gross Domestic product and Education data was downloaded from the World Bank website. The data was
```

.
1) Merge the data based on the country shortcode. How many of the IDs match?
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?
3) What are the average GDP rankings for the "High income: OECD" and "High income:nonOECD" groups?
4) Plot the GDP for all of the countries. Use ggplot2 to color your plot by Income Group.
5) 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?
.

The answers are presented within the file below as it becomes convenient to answer them, and they are also presented and discussed in more detail at the end of this file in the summary and conclusion.

.

The code below should be able to be run as R Markdown and present the results in a useful way.

This code was designed for use in RStudio on a Windows 10 machine. It should work on other platforms with no changes, however this may not be the case.

.

Load packages I may need. Set working directory.

```
library(stats)
library(plyr)
library(ggplot2)
library(repmis)

## setwd("C:/Users/hp/Desktop/SMU/Doing Data Science/Homework/Case Studies")

setwd(".")
```

.

Upload the data from the web and load data into two raw data sets in R.

This method is quick, but the data is only saved in active memory not on the hard drive.

This also makes the columns numeric and character by default.

```
gdpraw <- source_data("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv",skip=5,header=FA
```

```
## Downloading data from: https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv
```

```
## SHA-1 hash of the downloaded data file is:
## 18dd2f9ca509a8ace7d8de3831a8f842124c533d
```

```
fedraw <- source_data("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS_Country.csv",head
```

```
## Downloading data from: https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS_Country.csv
```

```
## SHA-1 hash of the downloaded data file is:
## 20be6ae8245b5a565a815c18a615a83c34745e5e
```

.

Another way to load the data. Saves files on the hard drive and then loads into R.

This loads the data with slightly different variable names so if you use this make sure to rename them correctly.

```
## GDPFileUrl<-"https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv"
## FEDFileUrl<-"https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS_Country.csv"

## download.file(GDPFileUrl,destfile="Data/FGDP_Rank_raw.csv")
## download.file(FEDFileUrl,destfile="Data/FEDSTATS_Country_raw.csv")

## gdpraw <- read.csv("./Data/FGDP_Rank_raw.csv",skip=5,header=FALSE)
## fedraw <- read.csv("./Data/FEDSTATS_Country_raw.csv",header=TRUE)
```

.

Looking at the data

```
## head(gdpraw)
## head(fedraw)
## str(gdpraw)
## str(fedraw)
```

Making data file objects so I won't have to load the data later if I need to go back and look at the raw data.

```
gdpa <- gdpraw
feda <- fedraw
```

.

Remove columns I don't need, rename header, and Remove rows without country codes

```
## get rid of all the other columns
gdpa <- gdpa[,c(1,2,4,5)]
head(gdpa)
```

```
##    V1 V2             V4           V5
## 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
```

```
## rename header
names(feda)
```

```
##  [1] "CountryCode"
##  [2] "Long Name"
##  [3] "Income Group"
##  [4] "Region"
##  [5] "Lending category"
##  [6] "Other groups"
##  [7] "Currency Unit"
##  [8] "Latest population census"
##  [9] "Latest household survey"
## [10] "Special Notes"
## [11] "National accounts base year"
## [12] "National accounts reference year"
## [13] "System of National Accounts"
## [14] "SNA price valuation"
## [15] "Alternative conversion factor"
## [16] "PPP survey year"
## [17] "Balance of Payments Manual in use"
## [18] "External debt Reporting status"
## [19] "System of trade"
## [20] "Government Accounting concept"
## [21] "IMF data dissemination standard"
## [22] "Source of most recent Income and expenditure data"
## [23] "Vital registration complete"
## [24] "Latest agricultural census"
## [25] "Latest industrial data"
## [26] "Latest trade data"
## [27] "Latest water withdrawal data"
## [28] "2-alpha code"
## [29] "WB-2 code"
## [30] "Table Name"
## [31] "Short Name"
```

```
names(gdpa) <- c("CountryCode", "Rank", "Long Name", "GDP")
names(gdpa)
```

```
## [1] "CountryCode" "Rank"        "Long Name"   "GDP"
```

```
##get rid of rows without country codes
gdpa <- gdpa[1:215,]
```

.

Merge files gdpa and feda using the CountryCode as a common column then count the NAs in each column.

```
mergedfile <- merge(gdpa, feda, by = "CountryCode", all=TRUE)

## Count the number of NAs in each column
mergefile.na <- colSums(is.na(mergedfile))
mergefile.na
```

```
##                                       CountryCode
##                                                 0
##                                              Rank
##                                                24
##                                       Long Name.x
##                                                24
##                                               GDP
##                                                24
##                                       Long Name.y
##                                                 5
##                                      Income Group
##                                                 5
##                                            Region
##                                                 5
##                                  Lending category
##                                                 5
##                                      Other groups
##                                                 5
##                                     Currency Unit
##                                                 5
##                          Latest population census
##                                                 5
##                           Latest household survey
##                                                 5
##                                     Special Notes
##                                                 5
##                       National accounts base year
##                                                 5
##                  National accounts reference year
##                                               202
##                       System of National Accounts
##                                               154
##                               SNA price valuation
##                                                 5
##                     Alternative conversion factor
##                                                 5
##                                   PPP survey year
##                                                94
##                  Balance of Payments Manual in use
##                                                  5
##                     External debt Reporting status
##                                                  5
##                                    System of trade
##                                                  5
##                      Government Accounting concept
##                                                  5
##                    IMF data dissemination standard
##                                                  5
## Source of most recent Income and expenditure data
##                                                  5
##                        Vital registration complete
##                                                  5
##                         Latest agricultural census
##                                                  5
##                             Latest industrial data
##                                                144
##                                  Latest trade data
##                                                 51
##                       Latest water withdrawal data
##                                                 87
##                                       2-alpha code
##                                                  6
##                                          WB-2 code
##                                                  6
##                                         Table Name
##                                                  5
##                                         Short Name
##                                                  5
```

.

Answer to question 1.

```
## Answer to question 1:
length(intersect(gdpa$CountryCode,feda$CountryCode))
```

```
## [1] 210
```

```
## Just a way of seeing the number that did not intersect
length(mergedfile$CountryCode) - length(intersect(gdpa$CountryCode,feda$CountryCode))
```

```
## [1] 29
```

.

Make blanks NAs and make sure the GDP and Rank columns are numeric so that we can answer question 2.

We also want to keep track of the NAs before and after.

```
count(is.na(mergedfile$GDP))
```

```
##       x freq
## 1 FALSE  215
## 2  TRUE   24
```

```
mergedfile$GDP <- as.numeric(gsub("[^[:digit:]]","", mergedfile$GDP))
count(is.na(mergedfile$GDP))
```

```
##       x freq
## 1 FALSE  190
## 2  TRUE   49
```

```
str(mergedfile$GDP)
```

```
##  num [1:239] NA 2584 NA 20497 114147 ...
```

```
count(is.na(mergedfile$Rank))
```

```
##       x freq
## 1 FALSE  215
## 2  TRUE   24
```

```
mergedfile$Rank <- as.numeric(gsub("[^[:digit:]]","", mergedfile$Rank))
count(is.na(mergedfile$Rank))
```

```
##       x freq
## 1 FALSE  190
## 2  TRUE   49
```

```
str(mergedfile$Rank)
```

```
##  num [1:239] NA 161 NA 105 60 125 32 26 133 NA ...
```

.

Answer to question 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?

Making a merged file that is sorted by Rank in ascending order and putting any NAs last.

```
## Use this just to check
mergedsort <- sort(mergedfile$GDP, decreasing = FALSE ,na.last = TRUE)

## Answer to question 2: sorting the merged file by GDP in ascending order
mergedsort2 <- mergedfile[order(mergedfile$GDP, decreasing = FALSE ,na.last = TRUE),]
mergedsort2[c(12,13,14),c(2,3,4)]
```

```
##     Rank          Long Name.x GDP
## 82   178              Grenada 767
## 111  178 St. Kitts and Nevis 767
## 231  177              Vanuatu 787
```

```
mergedsort2[c(13),c(2,3,4)]
```

```
##     Rank          Long Name.x GDP
## 111  178 St. Kitts and Nevis 767
```

As you can see there are 2 12th to last GDPs so there is no 13! But alphabetically St. Kitts and Nevis is 13th.

.

Remove all rows but the ones with Ranks, and keep only the first 6 columns.

Also make two objects that represent only the "High income: OECD" and "High income: nonOECD"

```
mergedsort3 <- mergedsort2[1:190, c(1:6)]
highincomeOECD <- mergedsort3[mergedsort3$`Income Group` == "High income: OECD",]
highincomenonOECD <- mergedsort3[mergedsort3$`Income Group` == "High income: nonOECD",]

## Gets rid of NA rows
highincomeOECD <- highincomeOECD[complete.cases(highincomeOECD$GDP),]
highincomenonOECD <- highincomenonOECD[complete.cases(highincomenonOECD$GDP),]
```

.

Answers to question 3: The averages of the Ranks for the "High income: OECD" and "High income: nonOECD".

```
## Question 3 Answers:
mean(highincomeOECD$Rank)
```

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

```
mean(highincomenonOECD$Rank)
```

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

```
## Average of GDPs as well. Why not?
mean(highincomeOECD$GDP)
```

```
## [1] 1483917
```

```
mean(highincomenonOECD$GDP)
```

```
## [1] 104349.8
```

.

Cleaning up everything else now. Making an object with only countries with Ranks, and getting rid of all the columns I don't seem to need.

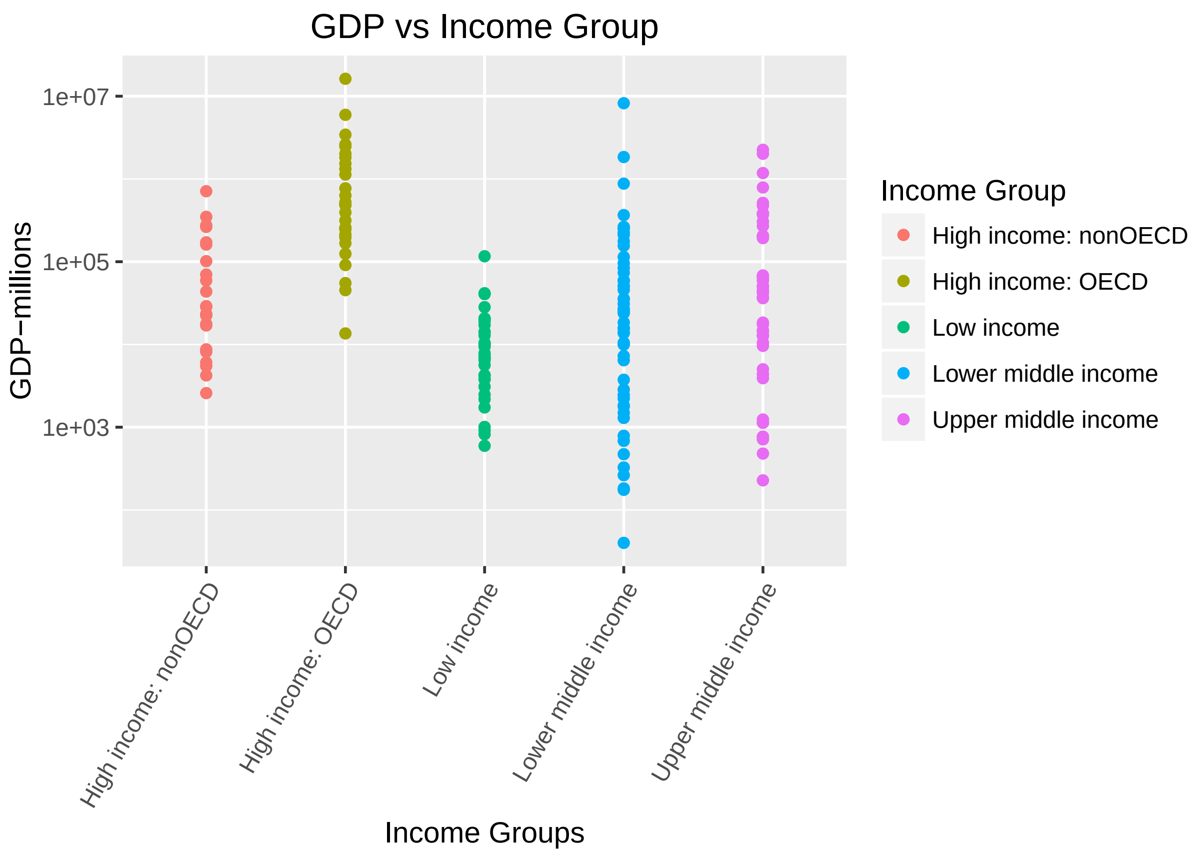
```
GDPall <- mergedsort2[1:190, c(1:6)]

## Clean up GDPall and get rid of all the factors
GDPall$`Income Group`<-GDPall$`Income Group` <- as.character(GDPall$`Income Group`)
GDPall$`Long Name.x` <- as.character(GDPall$`Long Name.x`)
GDPall$`Long Name.y` <- as.character(GDPall$`Long Name.y`)
GDPall$CountryCode <- as.character(GDPall$CountryCode)
```

Answer to question 4: Make a plot of GDP vs. Income Group colored by Income Group.

```
plot1a <- ggplot(GDPall) + geom_point(aes(y=GDP,x=`Income Group`,colour=`Income Group`)) + scale_y_log10
plot1a + labs(title="GDP vs Income Group", x="Income Groups",y="GDP-millions",colour="Income Group") +
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



.

One row has an NA in the Income Group column so we will replace it with “NA-South Sudan” and plot again.

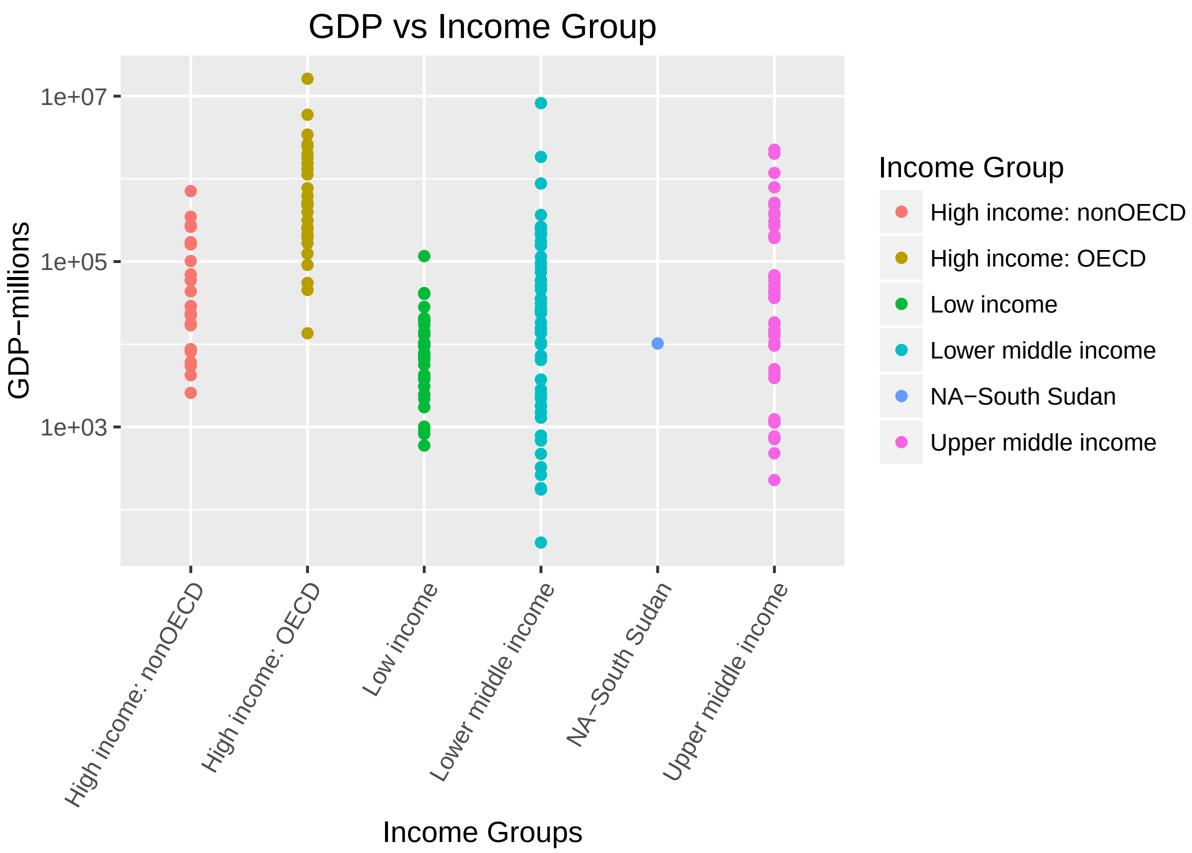
```
count(is.na(GDPall$`Income Group`))
```

```
##       x freq
## 1 FALSE  189
## 2  TRUE    1
```

```
GDPall$`Income Group`[is.na(GDPall$`Income Group`)] <- "NA-South Sudan"
count(is.na(GDPall$`Income Group`))
```

```
##       x freq
## 1 FALSE  190
```

```
plot1a <- ggplot(GDPall) + geom_point(aes(y=GDP,x=`Income Group`,colour=`Income Group`)) + scale_y_log1
plot1a + labs(title="GDP vs Income Group", x="Income Groups",y="GDP-millions",colour="Income Group") +
```



.

Making objects for the other income groups. Why not?

```
lowincome <- mergedsort3[mergedsort3$`Income Group` == "Low income",]
lowermiddleincome <- mergedsort3[mergedsort3$`Income Group` == "Lower middle income",]
uppermiddleincome <- mergedsort3[mergedsort3$`Income Group` == "Upper middle income",]

## Gets rid of NA rows I don't think there are any left now.
lowincome <- lowincome[complete.cases(lowincome$GDP),]
lowermiddleincome <- lowermiddleincome[complete.cases(lowermiddleincome$GDP),]
uppermiddleincome <- uppermiddleincome[complete.cases(uppermiddleincome$GDP),]
```

.

Answers to question 5

Use quantile function to break up the data into 5 Rank Quantiles

```
data.quant<-quantile(GDPall$Rank,seq(0, 1, 0.2))
data.quant
```

```
##    0%   20%   40%   60%   80%  100%
##   1.0  38.8  76.6 114.4 152.2 190.0
## This gave us 38.8 as the cuttoff, so below are two ways to get the answer
lowermiddleincome[which(lowermiddleincome$Rank<38.8), c(1,2,3,4,6)]
```

```
##     CountryCode Rank      Long Name.x     GDP        Income Group
## 62          EGY   38 Egypt, Arab Rep.  262832 Lower middle income
## 211         THA   31         Thailand  365966 Lower middle income
## 94          IDN   16        Indonesia  878043 Lower middle income
## 96          IND   10            India 1841710 Lower middle income
## 38          CHN    2            China 8227103 Lower middle income
```

```
GDPall[which(GDPall$Rank < 38.8 & GDPall$`Income Group` == "Lower middle income"), c(1,2,3,4,6)]
```

```
##     CountryCode Rank      Long Name.x     GDP        Income Group
## 62          EGY   38 Egypt, Arab Rep.  262832 Lower middle income
## 211         THA   31         Thailand  365966 Lower middle income
## 94          IDN   16        Indonesia  878043 Lower middle income
## 96          IND   10            India 1841710 Lower middle income
## 38          CHN    2            China 8227103 Lower middle income
```

```
nrow(GDPall[which(GDPall$Rank < 38.8 & GDPall$`Income Group` == "Lower middle income"),])
```

```
## [1] 5
```

```
## Another way to do it is to make a table, first figure out the quantile cutoff points
brk<-with(GDPall, quantile(GDPall$GDP, probs = c(0, 0.20, 0.4, 0.6, 0.8, 1.0)))
data.quant2 <- within(GDPall, quantile <- cut(GDPall$GDP, breaks = brk, labels = 1:5, include.lowest =

## Checking
nrow(data.quant2[which(data.quant2$quantile == 5 & data.quant2$`Income Group` == "Lower middle income")
```

```
## [1] 5
```

```
## Table answers question 5.
table(data.quant2$`Income Group`, data.quant2$quantile)
```

```
##
##                          1  2  3  4  5
##   High income: nonOECD   2  4  8  5  4
##   High income: OECD      0  1  1 10 18
##   Low income            11 16  9  1  0
##   Lower middle income   16  8 12 13  5
##   NA-South Sudan         0  1  0  0  0
##   Upper middle income    9  8  8  9 11
```

```
## You can see that there are only 5 total countries in the Lower middle income group that are also in
## Those countries are again
GDPall[which(GDPall$Rank < 38.8 & GDPall$`Income Group` == "Lower middle income"), c(1,2,3,4,6)]
```

```
##     CountryCode Rank      Long Name.x     GDP        Income Group
## 62          EGY   38 Egypt, Arab Rep.  262832 Lower middle income
## 211         THA   31         Thailand  365966 Lower middle income
## 94          IDN   16        Indonesia  878043 Lower middle income
## 96          IND   10            India 1841710 Lower middle income
## 38          CHN    2            China 8227103 Lower middle income
```

.
### Summary of all the Answers

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

```
## Answer to question 1:
length(intersect(gdpa$CountryCode,feda$CountryCode))
```

```
## [1] 210
```

Answer: 210 match using the intersect function.
.

Answer to question 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?

```
mergedsort2[c(12,13,14),c(2,3,4)]
```

```
##     Rank         Long Name.x GDP
## 82   178             Grenada 767
## 111  178 St. Kitts and Nevis 767
## 231  177             Vanuatu 787
```

```
mergedsort2[c(13),c(2,3,4)]
```

```
##     Rank         Long Name.x GDP
## 111  178 St. Kitts and Nevis 767
```

Answer: As you can see there are 2 12th to last GDPs so there is no 13! But alphabetically St. Kitts and Nevis is 13th.
.

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

```
mean(highincomeOECD$Rank)
```

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

```
mean(highincomenonOECD$Rank)
```

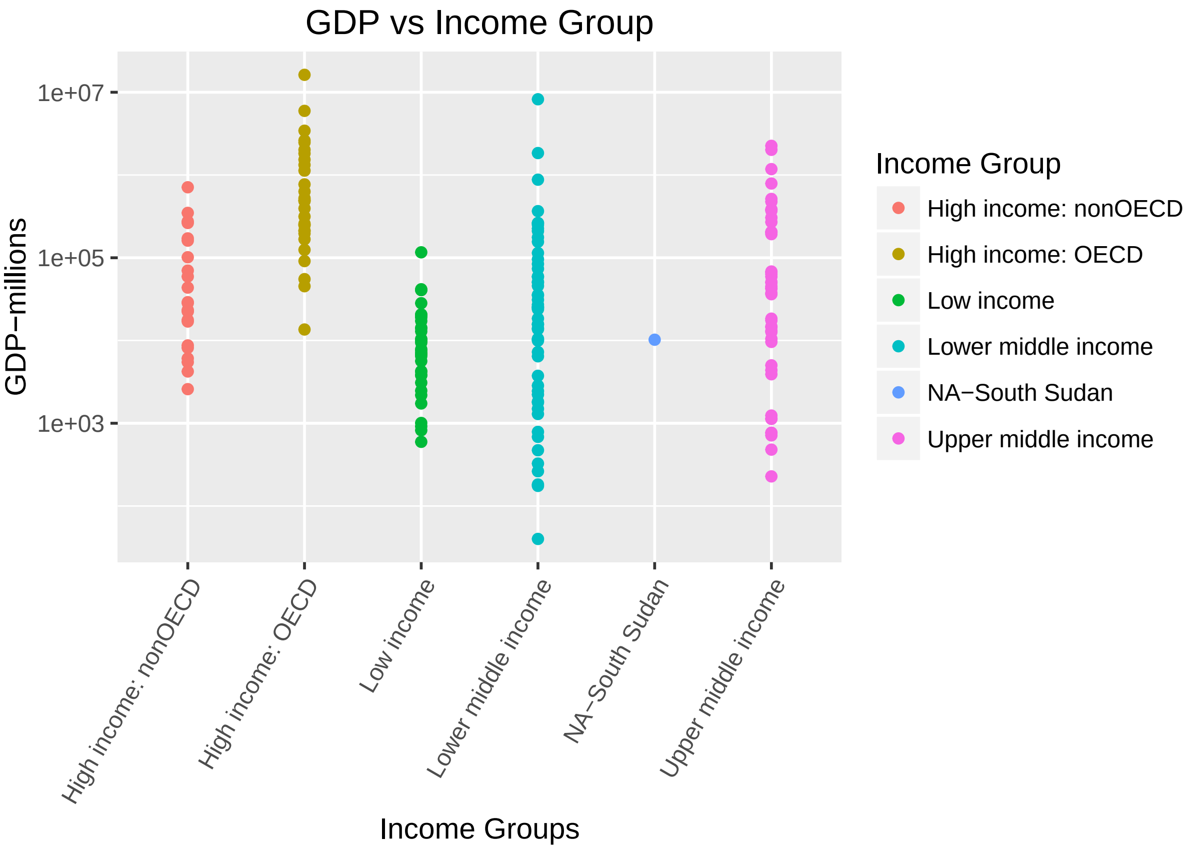
```
## [1] 91.91304
```

Answer: The “High income: OECD” has an average Rank of 32.97, and the “High income:nonOECD” has an average Rank of 91.91.

.

Answers to question 4) Plot the GDP for all of the countries. Use ggplot2 to color your plot by Income Group.

```
plot1a <- ggplot(GDPall) + geom_point(aes(y=GDP,x=`Income Group`,colour=`Income Group`)) + scale_y_log1
plot1a + labs(title="GDP vs Income Group", x="Income Groups",y="GDP-millions",colour="Income Group") + 
```



.

.

Answer: The plot shows all five income groups as well as South Sudan which was not assigned an Income Group.

.

Answers to question 5) 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?

```
table(data.quant2$`Income Group`, data.quant2$quantile)
```

```
##
##                         1  2  3  4  5
##   High income: nonOECD  2  4  8  5  4
##   High income: OECD     0  1  1 10 18
##   Low income           11 16  9  1  0
##   Lower middle income  16  8 12 13  5
##   NA-South Sudan        0  1  0  0  0
##   Upper middle income   9  8  8  9 11
```

```
GDPall[which(GDPall$Rank < 38.8 & GDPall$`Income Group` == "Lower middle income"), c(1,2,3,4,6)]
```

```
##     CountryCode Rank      Long Name.x     GDP        Income Group
## 62          EGY   38 Egypt, Arab Rep.  262832 Lower middle income
## 211         THA   31         Thailand  365966 Lower middle income
## 94          IDN   16        Indonesia  878043 Lower middle income
## 96          IND   10            India 1841710 Lower middle income
## 38          CHN    2            China 8227103 Lower middle income
```

Answer: The table shows that there are only five total countries in the upper 38, or 5th quantile, of GDP assigned to the Lower middle income group.

.

Summary and Conclusions

```
    Gross Domestic product and Education data was downloaded from the World Bank website. The data was 
```

.

I would like to thank the instructor and others for help creating this RMD file.