GDP and Income groupings of nations

Shravan Kuchkula

3/14/2017

Introduction

In the *World Development Indicators* database, all 189 World Bank member countries, plus 28 other economies with populations of more than 30,000, are classified based on income groups, so that data users can aggregate, group, and compare statistical data of interest, and for the presentation of key statistics. From this database, two data sets: `EdStats` and `GDP rank table` were obtained for the year of 2012 to analyse GDP based rankings and income group classifications.

The `EdStats` data set (*csv format*) for the year 2012 contains two key variables:

- **CountryCode** - A unique three letter code to identify a Country/Economy.
- **Income Group** - One of five income groups: low, lower-middle, upper-middle, high OECD and high non-OECD

The `GDP rank table` data set (*csv format*) contains a ranking table with no headers. The header to the columns are added as listed:

- **CountryCode** - A unique three letter code to identify an Country/Economy.
- **Rank** - Ranking based on GDP.
- **Country** - Country name.
- **GDP** - Gross Domestic Product in millions of US dollars.

Both these data sets are merged based on the matching country code to facilate with the analysis.

Cleaning the data sets

The `GDP rank table` data set has the following problems:

1. Columns names are not mapped correctly to the columns.
2. Actual data starts at row 6.
3. Third column is empty.
4. Sixth column is sparse and contains reference to footnotes.
5. Bottom part of the data set contains additional information.
6. The GDP column contains ".." for some missing values.
7. The CountryCode column contains missing values.

Problems 1-to-6 are fixed while importing the data into R. The script that is used to do this is `GatherData1.R`. As there are missing values in CountryCode which we will use later while merging with EdStats dataset, the rows containing missing CountryCode are removed from the dataset in this script `CleanData1.R`

The `EdStats` data set does not have any problems, it is imported into R using this script: `GatherData2.R`

File and Directory Organization

- `Makefile.txt` - Downloads the two data sets, cleans and merges them.
- `CleanData1.R` - Cleans the GDP Rank table data set.
- `GatherData1.R` - Downloads the GDP Rank table data set.
- `GatherData2.R` - Downloads the EducStats data set.
- `MergeData.R` - Merges GDP and EducStats based on CountryCode.

The project structure is below:

```
GDPEduc
|_
  Analysis
  |_
    Data
    |_
      Makefile.txt
      CleanData1.R
      GatherData1.R
      GatherData2.R
      MergeData.R
      EDUC.csv
      GDP.csv
      MergedData.csv
  |_
    Analysis.R
    Questions.R
    Main.R
|_
  Paper
    |_
      Report.html
      Report.pdf
      Report.Rmd
```

Instructions to run the code

When you download this project from Github, you will be in project's root directory, which in this case is: GDPEduc. You have 2 methods to reproduce the analysis done in this project.

Method 1: From the GDPEduc/Paper directory, run the Report.Rmd file. When running the Report.Rmd from the Paper directory, the working directory is automatically changed to GDPEduc/Paper. Thus, we need to explicitly set the working directory to Analysis/Data, since we want the .csv files to be located over there.

```
setwd('../Analysis/Data')
source('Makefile.txt')
```

```
## *****************************
## CleanData1.R
## *****************************
```

```
setwd('../../Paper')
```

Next, load all the analysis R scripts that are needed for this project

To run the analysis, change the directory to Analysis and run the Analysis.R which contains functions to aid in the analysis.

```
setwd('../Analysis')
source('libraries.R')
source('Analysis.R')
setwd('../Paper')
```

Method 2: Second way to reporduce this project is by running the Makefile.txt from the R command prompt .

I will be using Method 1 for answering the questions in this case study.

Analysis

Question 1:

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

The data sets `GDP rank table` and `EdStats` are merged based on CountryCode. Invoke the function idMatches in Analysis.R

```
num_id_matches <- idMatches()
```

```
paste("The number of IDs matched by merging GDP and EdStats datasets are ",
        num_id_matches)
```

```
## [1] "The number of IDs matched by merging GDP and EdStats datasets are  224"
```

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?

Before we sort the data frame, we need to fix the following problems with the data set: 1. Remove NA's from GDP and Ranking columns. 2. Format the GDP data by removing commas "," 3. Convert GDP data to numeric.

The `gdpRank()` function in the `Analysis.R` script displays the n-th smallest economy.

```
paste("The 13th smallest GDP country is: ", gdpRank(13))
```

```
## [1] "The 13th smallest GDP country is:  St. Kitts and Nevis"
```

Question 3:

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

The `groupRankAverages()` function in the `Analysis.R` script displays the average GDP Ranking of a given income group.

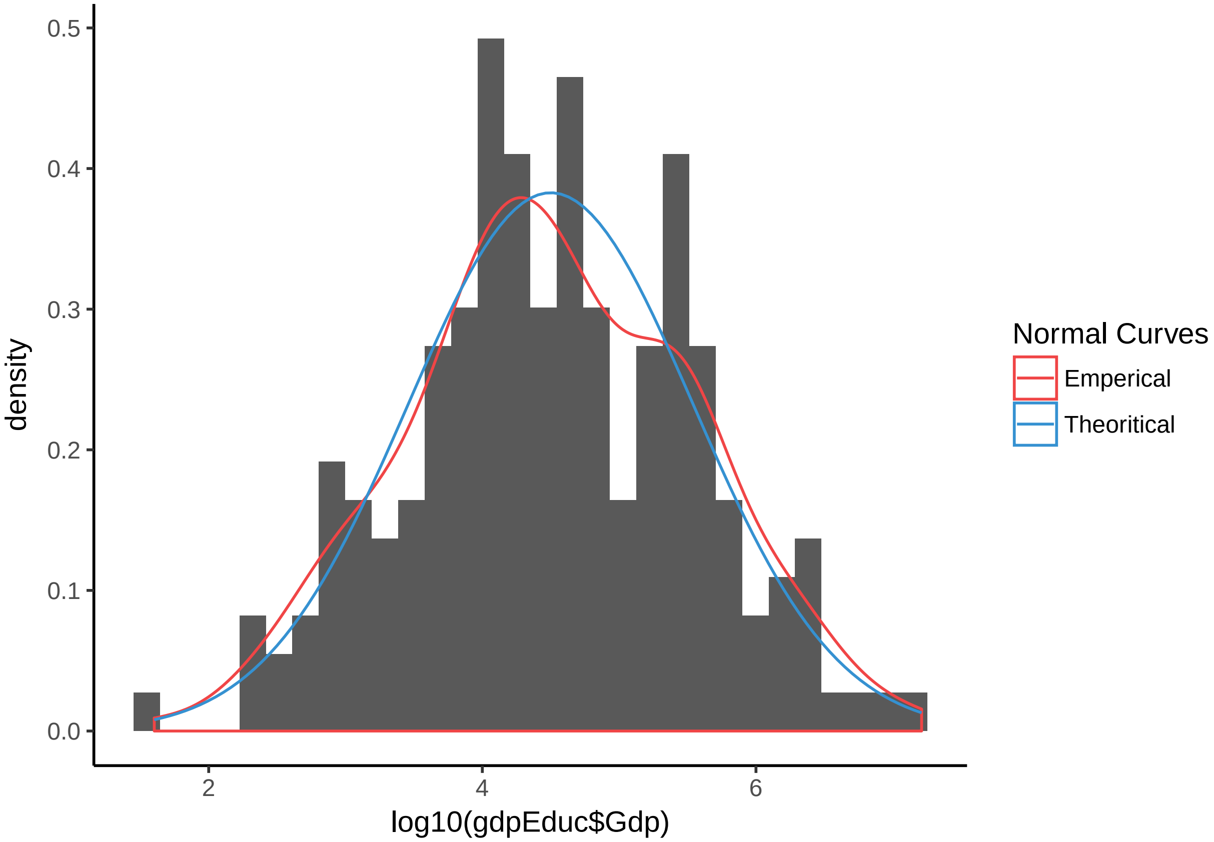
```
grpAvgs <- groupRankAverages()
grpAvgs[grepl(".*OECD$", grpAvgs$`Income Group`),]
```

```
## # A tibble: 2 × 2
##          `Income Group`      avg
##                   <chr>    <dbl>
## 1 High income: nonOECD 91.91304
## 2    High income: OECD 32.96667
```

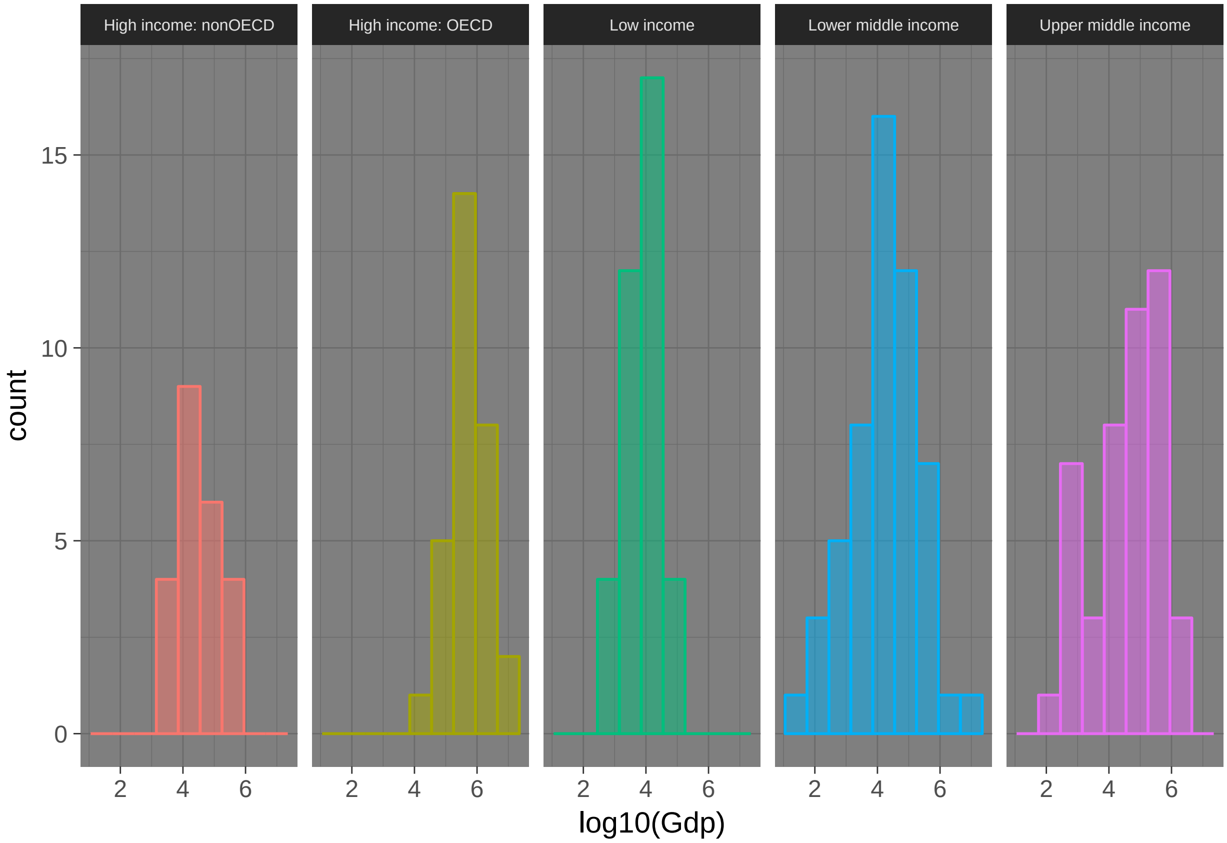
Question 4:

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

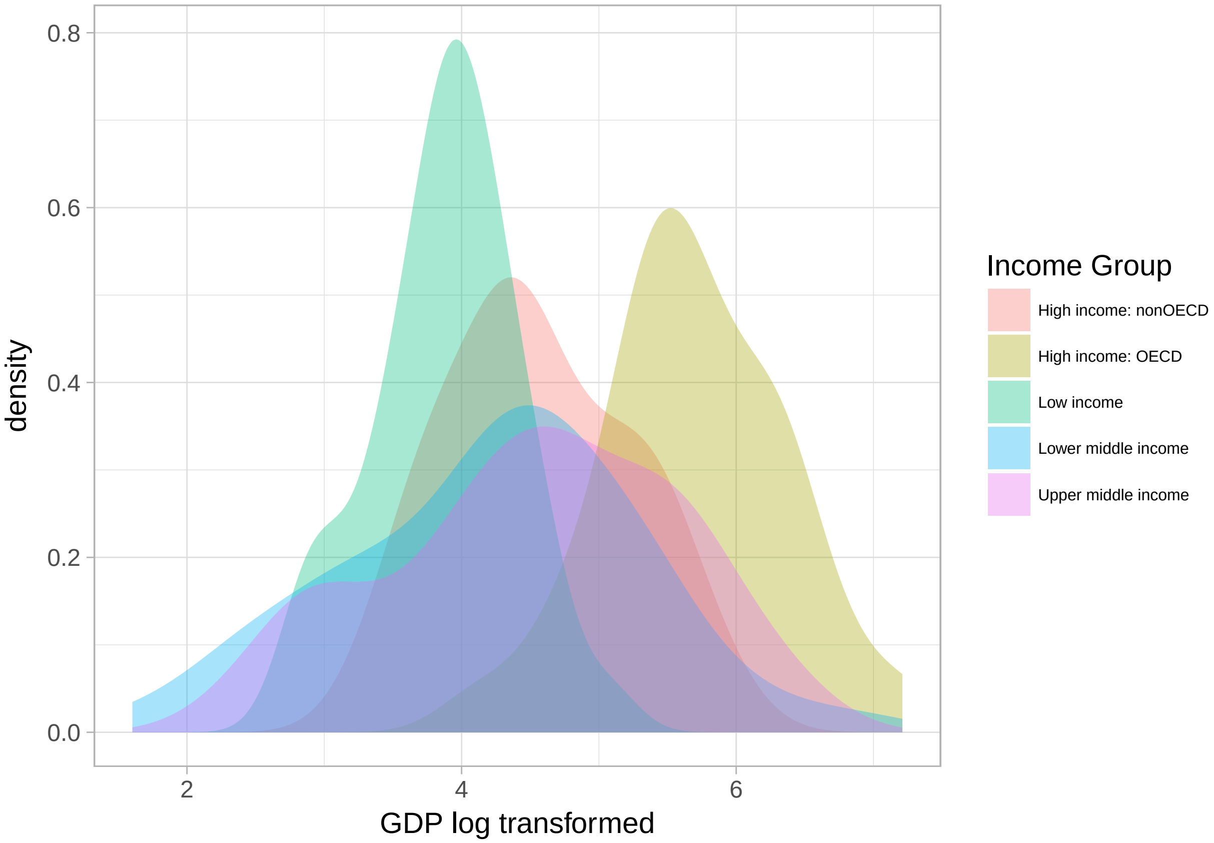
```
# Draw a density curve
cols <- c("Emperical"="#f04546","Theoritical"="#3591d1")
  fun_args <- list(mean = mean(log10(gdpEduc$Gdp)),
                   sd = sd(log10(gdpEduc$Gdp)))
  ggplot(gdpEduc, aes(x = log10(gdpEduc$Gdp))) +
    geom_histogram(aes(y = ..density..)) +
    geom_density(aes(col = "Emperical")) +
    stat_function(fun = dnorm, args = fun_args, aes(col = "Theoritical")) +
    theme_classic() +
    scale_color_manual(name = "Normal Curves", values = cols)
```



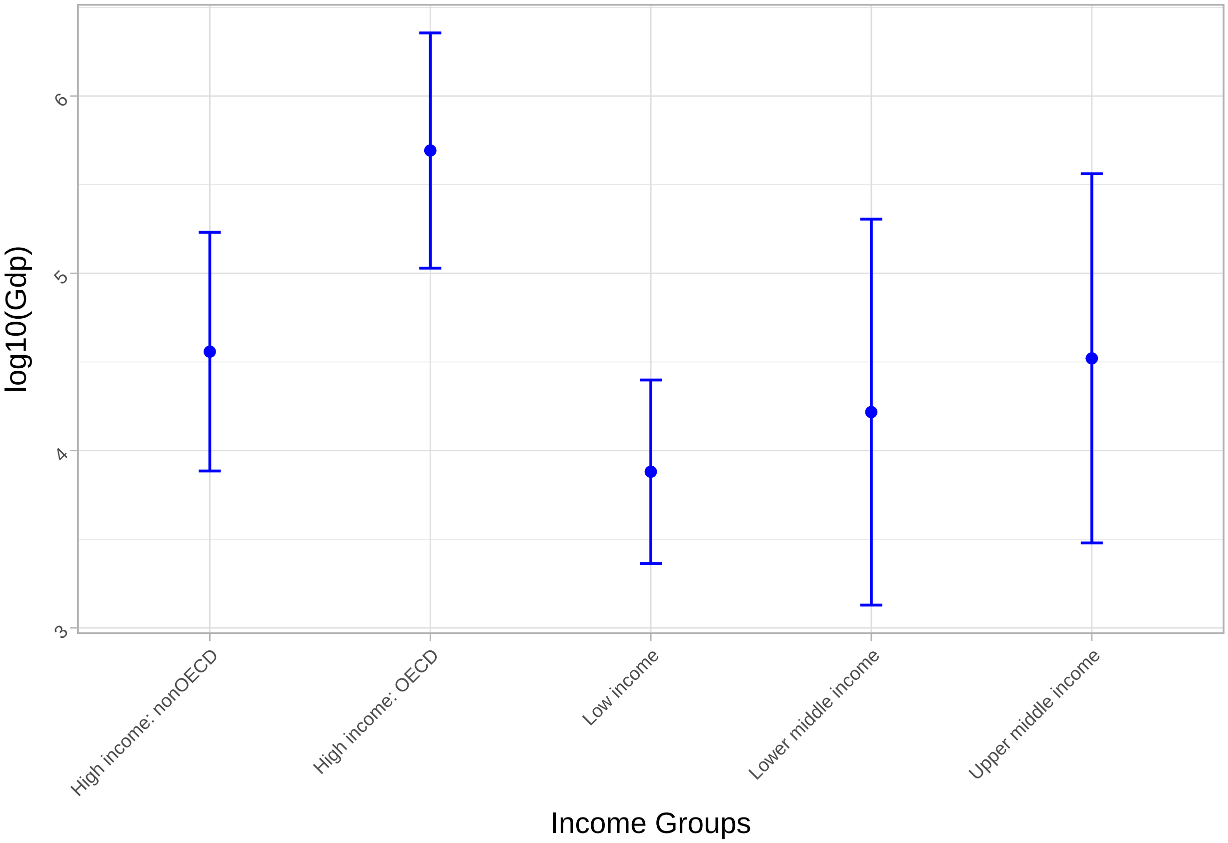
```
  # Draw the histogram of count vs log10(GDP) faceted by Income Groups
  ggplot(gdpEduc, aes(x = log10(Gdp), col = factor(`Income Group`),
                   fill = factor(`Income Group`))) +
    geom_histogram(binwidth = 0.7, alpha = 0.5) +
    facet_grid(.~ factor(`Income Group`)) +
    theme_dark() +
    theme(legend.position = "none", strip.text = element_text(size = rel(0.5))) +
    labs(x = "log10(Gdp)")
```



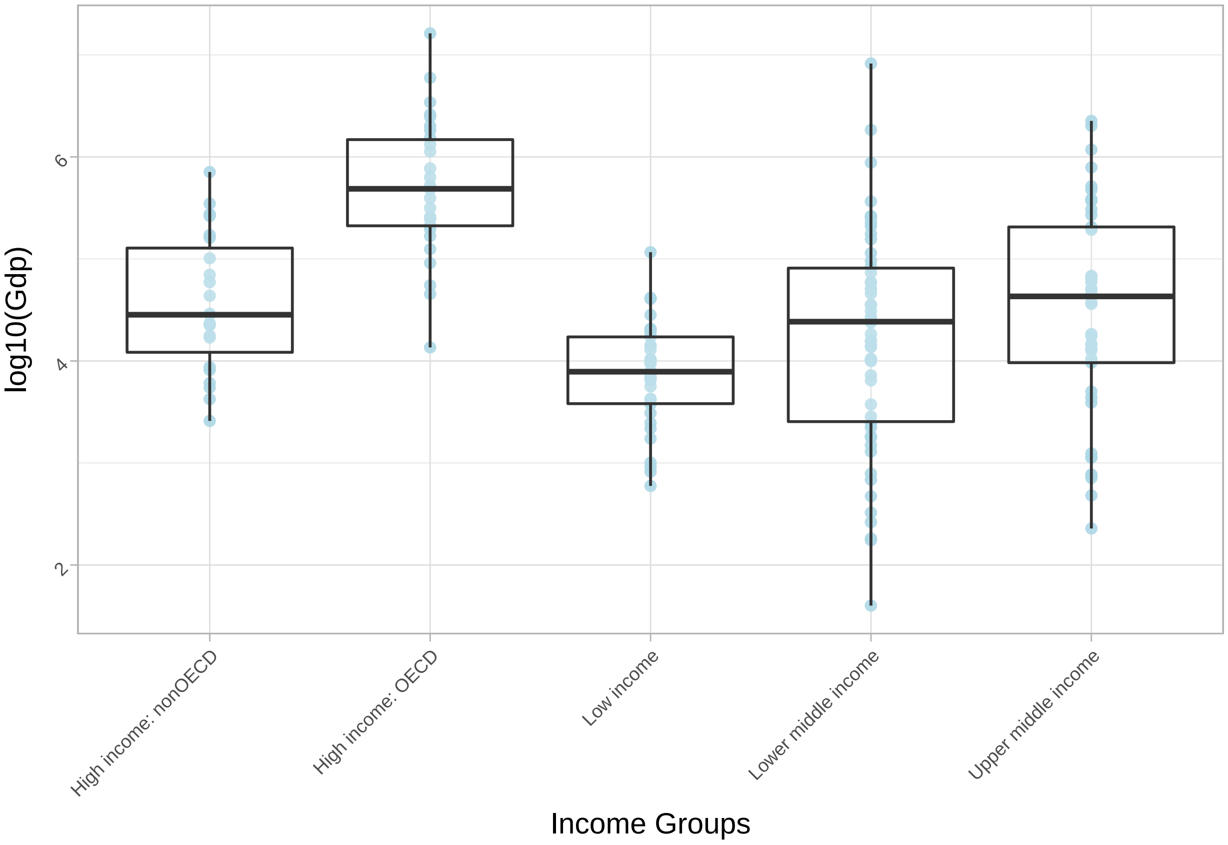
```
# Overlay multiple density plots
ggplot(gdpEduc, aes(x = log10(Gdp), fill = factor(`Income Group`))) +
  geom_density(col = NA, alpha = 0.35) +
  theme_light() +
  theme(legend.text = element_text(size=rel(0.5))) +
  labs(x = "GDP log transformed") +
  scale_fill_discrete(name = "Income Group")
```



```
# Display mean and 1 sd
ggplot(gdpEduc, aes(x = factor(`Income Group`), y = log10(Gdp))) +
  stat_summary(geom = "point", fun.y = mean, col = "blue") +
  stat_summary(geom = "errorbar", fun.data = mean_sdl,
               fun.args = list(mult = 1), col = "blue", width = 0.1) +
  theme_light() +
  theme(axis.text = element_text(angle = 45, hjust = c(1), size = rel(0.6))) +
  labs(x = "Income Groups")
```



```
# Box plots with data overlayed to get an idea of the distribution of GDP
ggplot(gdpEduc, aes(x = factor(`Income Group`), y = log10(Gdp))) +
geom_point(colour="lightblue", alpha=0.9, position="identity") +
  geom_boxplot(outlier.size=0, alpha=0.2) +
  theme_light() +
  theme(axis.text = element_text(angle = 45, hjust = c(1), size = rel(0.6))) +
  labs(x = "Income Groups")
```



Question 5:

Provide summary statistics of GDP by income groups.

```
# Descriptive statistics using psych package.
  descStatsGDP <- describeBy(gdpEduc$Gdp, gdpEduc$`Income Group`, mat=TRUE)
  descStatsGDP %>%
    select(-item, -vars, -mad) %>%
    print(row.names = FALSE)
```

```
##                 group1  n        mean          sd   median   trimmed    min
##  High income: nonOECD 23  104349.83  165334.45  28373.0  70189.05  2584
##     High income: OECD 30 1483917.13 3070463.52 486528.5 782126.21 13579
##            Low income 37   14410.78   20473.09   7843.0  10715.90   596
##   Lower middle income 54  256663.48 1139619.92  24272.0  51890.64    40
##   Upper middle income 45  231847.84  476872.04  42945.0 113409.27   228
##       max    range     skew  kurtosis         se
##    711050   708466 2.284047  5.230623  34474.616
##  16244600 16231021 3.776308 14.976956 560587.378
##    116355   115759 3.446657 13.898099   3365.755
##   8227103  8227063 6.381095 41.381778 155082.628
##   2252664  2252436 3.013289  8.925480  71087.887
```

Question 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?

```
# QuantileCut to divide the GDP rankings into 5 quantile groups
  gdpEducTemp <- gdpEduc
  gdpEducTemp %>%
    mutate(quantiles = quantileCut(gdpEduc$Ranking,5,
                                   labels=c("Q1","Q2","Q3","Q4","Q5"))) %>%
    filter(as.character(quantiles) == "Q1", `Income Group` == "Lower middle income") %>%
    select(CountryCode, Economy, Ranking, `Income Group`, quantiles)
```

```
## # A tibble: 5 × 5
##   CountryCode          Economy Ranking      `Income Group` quantiles
##         <chr>            <chr>   <dbl>               <chr>    <fctr>
## 1         CHN            China       2 Lower middle income        Q1
## 2         IND            India      10 Lower middle income        Q1
## 3         IDN        Indonesia      16 Lower middle income        Q1
## 4         THA         Thailand      31 Lower middle income        Q1
## 5         EGY Egypt, Arab Rep.      38 Lower middle income        Q1
```