Case Study: GDP

Clinton Tippett

June 12, 2016

## Introduction

GDP by country. Based on two datasets, I will present findings around multiple country's GDP and the factors that distinguish them. I will explore income groups, GDP rank, and how each of the variables relate to each other.

## Loading the data...

In order to conduct data analysis, we must have data! In the next chunk of code, I am going to load the data into R.

install.packages("downloader", repos = 'http://cran.us.r-project.org')

##   
## The downloaded binary packages are in  
## /var/folders/z5/9mhthf2950d9twsm48csc45w0000gn/T//RtmpnTaaaE/downloaded\_packages

library(downloader)  
  
download("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv", destfile="gdp.csv")  
gdp <- read.csv("gdp.csv", stringsAsFactors=FALSE)  
  
download("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS\_Country.csv", destfile="edu.csv")  
edu <- read.csv("edu.csv", stringsAsFactors=FALSE)

Now, the two datasets that will be used are loaded into R. This was accomplished using the 'download' function in R to import the data and then read, using read.csv() function, in the files and store them as dataframes.

## Cleaning the Data

Once the data was loaded, I could see (by calling the dataframe) that there was a lot of missing or unused data. Before I merge the files, this must be cleaned up.

gdp1 <- gdp[5:235,]  
  
names(gdp1) <- c("CountryCode", "Rank", "Un3", "Country", "Dollars", "un6", "un7", "un8", "un9", "un10")  
gdp1$Dollars <- gsub(",","", gdp1$Dollars)  
gdp1$Dollars <- as.numeric(gdp1$Dollars)

## Warning: NAs introduced by coercion

gdp1$Rank <- as.numeric(gdp1$Rank)  
str(gdp1)

## 'data.frame': 231 obs. of 10 variables:  
## $ CountryCode: chr "USA" "CHN" "JPN" "DEU" ...  
## $ Rank : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ Un3 : logi NA NA NA NA NA NA ...  
## $ Country : chr "United States" "China" "Japan" "Germany" ...  
## $ Dollars : num 16244600 8227103 5959718 3428131 2612878 ...  
## $ un6 : chr "" "" "" "" ...  
## $ un7 : logi NA NA NA NA NA NA ...  
## $ un8 : logi NA NA NA NA NA NA ...  
## $ un9 : logi NA NA NA NA NA NA ...  
## $ un10 : logi NA NA NA NA NA NA ...

gdp2 <- within(gdp1, rm("Un3", "un6", "un7", "un8", "un9", "un10"))  
  
edu1 <- within(edu, rm("Special.Notes", "Long.Name", "Region", "Lending.category", "Other.groups", "Currency.Unit", "Latest.population.census", "Latest.household.survey", "National.accounts.base.year", "National.accounts.reference.year", "System.of.National.Accounts", "SNA.price.valuation", "Alternative.conversion.factor", "PPP.survey.year", "Table.Name", "WB.2.code", "X2.alpha.code", "Balance.of.Payments.Manual.in.use", "External.debt.Reporting.status", "System.of.trade", "Government.Accounting.concept", "IMF.data.dissemination.standard", "Source.of.most.recent.Income.and.expenditure.data", "Vital.registration.complete", "Latest.agricultural.census", "Latest.industrial.data", "Latest.trade.data", "Latest.water.withdrawal.data"))  
  
  
comb <- merge(gdp2, edu1, by="CountryCode", all=TRUE)

Here, I selected only the rows of data that I needed, as a lot of the rows were blank. Next, I gave the variables names and used the label of "un3, un4, etc" to label the columns that I wanted to remove. I repeated this process for the edu dataset. Using the within() function, I can select which variables I want to use for analysis. Also, it allows variables to easily be put back in place if needed.

## Question 1: By using the str() function below, I can see that 231 of the 'CountryCode' IDs match.

str(gdp2)

## 'data.frame': 231 obs. of 4 variables:  
## $ CountryCode: chr "USA" "CHN" "JPN" "DEU" ...  
## $ Rank : num 1 2 3 4 5 6 7 8 9 10 ...  
## $ Country : chr "United States" "China" "Japan" "Germany" ...  
## $ Dollars : num 16244600 8227103 5959718 3428131 2612878 ...

str(edu1)

## 'data.frame': 234 obs. of 3 variables:  
## $ CountryCode : chr "ABW" "ADO" "AFG" "AGO" ...  
## $ Income.Group: chr "High income: nonOECD" "High income: nonOECD" "Low income" "Lower middle income" ...  
## $ Short.Name : chr "Aruba" "Andorra" "Afghanistan" "Angola" ...

## Question 2: To rearrange each country in ascending order by GDP I used the order function. Based on the new data frame, the 13th country on the list is:KNA, St. Kitts and Nevis with a GDP of 767.

combOr <- comb[order(comb$Dollars),]  
head(combOr, 13)

## CountryCode Rank Country Dollars  
## 221 TUV 190 Tuvalu 40  
## 112 KIR 189 Kiribati 175  
## 143 MHL 188 Marshall Islands 182  
## 175 PLW 187 Palau 228  
## 201 STP 186 S\xe3o Tom\xe9 and Principe 263  
## 74 FSM 185 Micronesia, Fed. Sts. 326  
## 217 TON 184 Tonga 472  
## 55 DMA 183 Dominica 480  
## 45 COM 182 Comoros 596  
## 236 WSM 181 Samoa 684  
## 229 VCT 180 St. Vincent and the Grenadines 713  
## 84 GRD 178 Grenada 767  
## 113 KNA 178 St. Kitts and Nevis 767  
## Income.Group Short.Name  
## 221 Lower middle income Tuvalu  
## 112 Lower middle income Kiribati  
## 143 Lower middle income Marshall Islands  
## 175 Upper middle income Palau  
## 201 Lower middle income S\xe3o Tom\xe9 and Principe  
## 74 Lower middle income Micronesia  
## 217 Lower middle income Tonga  
## 55 Upper middle income Dominica  
## 45 Low income Comoros  
## 236 Lower middle income Samoa  
## 229 Upper middle income St. Vincent and the Grenadines  
## 84 Upper middle income Grenada  
## 113 Upper middle income St. Kitts and Nevis

str(comb)

## 'data.frame': 241 obs. of 6 variables:  
## $ CountryCode : chr "" "" "" "ABW" ...  
## $ Rank : num NA NA NA 161 NA 105 60 125 32 26 ...  
## $ Country : chr "" "" "" "Aruba" ...  
## $ Dollars : num NA NA NA 2584 NA ...  
## $ Income.Group: chr NA NA NA "High income: nonOECD" ...  
## $ Short.Name : chr NA NA NA "Aruba" ...

## Question 3: The average GDP rankings were also calculated based on if the country fell into one of two groups: High income: OECD and High income: nonOECD. Since NA values were present, the code had to be specified to remove these values when calculating the mean.

mean(comb$Rank[comb$Income.Group=="High income: OECD"], na.rm=TRUE)

## [1] 32.96667

mean(comb$Rank[comb$Income.Group=="High income: nonOECD"], na.rm=TRUE)

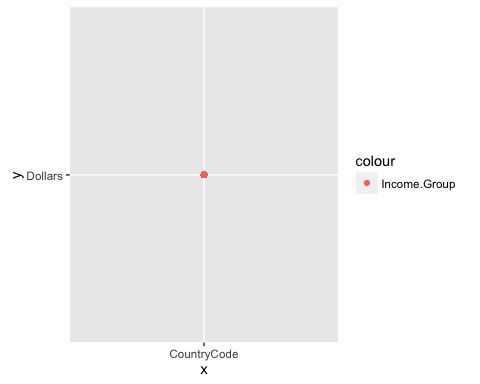
## [1] 91.91304

## Question 4: I cannot seem to get this to generate. I've looked at several examples and my syntax seems correct...

install.packages("ggplot2", repos = 'http://cran.us.r-project.org')

##   
## The downloaded binary packages are in  
## /var/folders/z5/9mhthf2950d9twsm48csc45w0000gn/T//RtmpnTaaaE/downloaded\_packages

library(ggplot2)  
qplot("CountryCode", "Dollars", data=comb, color="Income.Group")



## Question 5: Using the cut method, I divided the GDP rankings into five groups. When plotted into a table against the different income groups we can see that there are 5 countries that fall into the 'Lower middle' income group, but are in the quantile with the highest GDP.

rankcut <- cut(comb$Rank, breaks=5)  
table(rankcut, comb$Income.Group)

##   
## rankcut High income: nonOECD High income: OECD Low income  
## (0.811,38.8] 0 4 18 0  
## (38.8,76.6] 0 5 10 1  
## (76.6,114] 0 8 1 9  
## (114,152] 0 4 1 16  
## (152,190] 0 2 0 11  
##   
## rankcut Lower middle income Upper middle income  
## (0.811,38.8] 5 11  
## (38.8,76.6] 13 9  
## (76.6,114] 12 8  
## (114,152] 8 8  
## (152,190] 16 9

## Conclusion: The GDP ranking can be broken down by several factors. The main factors taken into consideration in this project was that of the "Income Group". Surprisingly, the countries that ranked highest in GDP were not typically in the highest income group. In fact, over 40% of those countries were grouped in middle income or lower. Additionally, there is a sizable difference in GDP amounts between the top 5 and the rest of the world. This gap can help explain the variance shown in the rank quantile.