Case Study 1

Group from MSDS 6306 - 404: Jethin Abraham, Daniel Higdon, Mustafa Sakarwala, and Keyue Wang

3/8/2017

# Introduction

The puropose of our project is to demonstrate, and understand, how the annual GDP (Gross Domestic Product) of countries around the globe compares to the their income grouping according to The World Bank. Throughout the project we will utilize two World Bank data sets:

1. GDP of 190 countries in 2012
2. Educational data on 235 various countries

The project will take readers through the analysis step-by-step, from downloading of the data to the final analysis. We will conclude with our findings of the analysis.

Before getting started, you should make sure you have installed and loaded the ggplot2, and dplyr packages into your R workspace. We will use functions from both packages through the project.

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)  
#Set your working directory  
setwd("C:/Users/Jethin/Documents/Git/MSDS6306\_CaseStudy1/source")

## Step 1: Downloading the data

We will be downloading our data from the provided URLs and read them into the variables GDP and econ. After analyzing the data, we also used this step to skip the first 3 rows of the GDP data because they are irrelevant to our analysis. Next we remove the sub-totals and totals at the bottom of the data set, as they will not be used in this analysis. In addition, we replaced empty values with 'NA' for both the GDP and econ data.

#download and read the data  
  
GDP <- read.csv('https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FGDP.csv ' ,skip = 3 ,nrows=236,na.strings=c("","NA"))  
econ <- read.csv('https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FEDSTATS\_Country.csv',na.strings=c("","NA"))

## Step 2: Cleaning the data

Now that we have our data, a quick view of it shows that the GDP data set is not clean. First, we rename each field to something that makes more sense ("CountryCode", "Rank", "Country", and "USD"). We will then sum the number of 'NA' that we find in our CountryCode column into the variable NA\_Count. We see that there are a total of 8 counts of 'NA' in the CountryCode column. Finally, we have a remaining problem. R has saved the Rank as USD columns as factors, not numbers. We are going to do math on these columns later in the analysis, so we'll go ahead and change them to numeric in the last two lines of code.

#Cleaning the data, deleting unnecessary rows and columns  
names(GDP) <- c('CountryCode','Rank','x','Country','USD')   
  
#Subsetting columns of our interest   
Clean\_GDP\_NA <- subset(GDP,select = c('CountryCode','Rank','Country','USD'))  
  
#Creating a function to check for NA is rows  
row.has.na <- apply(Clean\_GDP\_NA, 1, function(x){any(is.na(x))})  
  
#Removing the rows that are NULL and getitng a clean set of data   
Clean\_GDP <- Clean\_GDP\_NA[!row.has.na,]  
  
#Count the number of missing values  
NA\_Count <- sum(is.na(Clean\_GDP\_NA$CountryCode))   
NA\_Count

## [1] 8

#making Rank and GDP numeric  
Clean\_GDP$Rank<-as.numeric(as.character((Clean\_GDP$Rank)))  
Clean\_GDP$USD<-as.numeric(gsub(",","",Clean\_GDP$USD))

## Step 3: Merge the data sets (Question 1)

After merging the data sets you'll notice a warning message that NAs have been duplicated. This is again caused by R saving the Rank and USD fields as factors rather than the numeric data type. We'll fix it with the second and third line of code, similar to how we did in step two.

#Merging the two datasets  
Comb\_data<-merge(Clean\_GDP,econ,by="CountryCode")  
#Making Rank and USD numeric  
Comb\_data$Rank<-as.numeric(as.character(Comb\_data$Rank))  
Comb\_data$USD<-as.numeric(gsub(",","",Comb\_data$USD))  
  
#Q1  
summary(Clean\_GDP)

## CountryCode Rank Country   
## ABW : 1 Min. : 1.00 Afghanistan : 1   
## AFG : 1 1st Qu.: 48.25 Albania : 1   
## AGO : 1 Median : 95.50 Algeria : 1   
## ALB : 1 Mean : 95.49 Angola : 1   
## ARE : 1 3rd Qu.:142.75 Antigua and Barbuda: 1   
## ARG : 1 Max. :190.00 Argentina : 1   
## (Other):184 (Other) :184   
## USD   
## Min. : 40   
## 1st Qu.: 7005   
## Median : 27638   
## Mean : 377652   
## 3rd Qu.: 205289   
## Max. :16244600   
##

summary(Comb\_data)

## CountryCode Rank Country   
## ABW : 1 Min. : 1.00 Afghanistan : 1   
## AFG : 1 1st Qu.: 48.00 Albania : 1   
## AGO : 1 Median : 95.00 Algeria : 1   
## ALB : 1 Mean : 95.31 Angola : 1   
## ARE : 1 3rd Qu.:143.00 Antigua and Barbuda: 1   
## ARG : 1 Max. :190.00 Argentina : 1   
## (Other):183 (Other) :183   
## USD Long.Name   
## Min. : 40 Antigua and Barbuda : 1   
## 1st Qu.: 6972 Arab Republic of Egypt: 1   
## Median : 28242 Argentine Republic : 1   
## Mean : 379597 Aruba : 1   
## 3rd Qu.: 205789 Barbados : 1   
## Max. :16244600 Belize : 1   
## (Other) :183   
## Income.Group Region   
## High income: nonOECD:23 East Asia & Pacific :29   
## High income: OECD :30 Europe & Central Asia :50   
## Low income :37 Latin America & Caribbean :35   
## Lower middle income :54 Middle East & North Africa:18   
## Upper middle income :45 North America : 3   
## South Asia : 8   
## Sub-Saharan Africa :46   
## Lending.category Other.groups Currency.Unit  
## Blend:16 Euro area: 16 Euro : 18   
## IBRD :62 HIPC : 39 CFA franc : 14   
## IDA :60 NA's :134 U.S. dollar : 8   
## NA's :51 East Caribbean dollar: 6   
## Australian dollar : 3   
## Afghan afghani : 1   
## (Other) :139   
## Latest.population.census Latest.household.survey  
## 2001 :45 MICS, 2006:21   
## 2000 :30 DHS, 2005 : 7   
## 2002 :21 DHS, 2007 : 6   
## 2006 :14 MICS, 2005: 6   
## 2004 :12 DHS, 2008 : 5   
## (Other):66 (Other) :71   
## NA's : 1 NA's :73   
## Special.Notes  
## Fiscal year end: June 30; reporting period for national accounts data: FY. : 7   
## Fiscal year end: March 31; reporting period for national accounts data: CY.: 7   
## Fiscal year end: June 30; reporting period for national accounts data: CY. : 6   
## Fiscal year end: March 20; reporting period for national accounts data: FY.: 2   
## Fiscal year end: March 31; reporting period for national accounts data: FY.: 2   
## (Other) : 42   
## NA's :123   
## National.accounts.base.year National.accounts.reference.year  
## 2000 :42 Min. :1987   
## 1990 :15 1st Qu.:1996   
## 1995 : 9 Median :2000   
## 1994 : 7 Mean :1999   
## 1991 : 5 3rd Qu.:2000   
## (Other):79 Max. :2007   
## NA's :32 NA's :155   
## System.of.National.Accounts SNA.price.valuation  
## Min. :1993 VAB :150   
## 1st Qu.:1993 VAP : 35   
## Median :1993 NA's: 4   
## Mean :1993   
## 3rd Qu.:1993   
## Max. :1993   
## NA's :106   
## Alternative.conversion.factor PPP.survey.year  
## 1990-95: 8 Min. :2005   
## 1987-95: 5 1st Qu.:2005   
## 1993 : 3 Median :2005   
## 1991 : 2 Mean :2005   
## 1992-95: 2 3rd Qu.:2005   
## (Other): 25 Max. :2005   
## NA's :144 NA's :45   
## Balance.of.Payments.Manual.in.use External.debt.Reporting.status  
## BPM4: 9 Actual :92   
## BPM5:158 Estimate :11   
## NA's: 22 Preliminary:22   
## NA's :64   
##   
##   
##   
## System.of.trade Government.Accounting.concept  
## General:104 Budgetary : 35   
## Special: 69 Consolidated:109   
## NA's : 16 NA's : 45   
##   
##   
##   
##   
## IMF.data.dissemination.standard  
## GDDS:92   
## SDDS:68   
## NA's:29   
##   
##   
##   
##   
## Source.of.most.recent.Income.and.expenditure.data  
## IHS, 2007 : 10   
## IHS, 2000 : 9   
## ES/BS, 2005: 6   
## IHS, 2006 : 6   
## ES/BS, 2004: 5   
## (Other) :104   
## NA's : 49   
## Vital.registration.complete Latest.agricultural.census  
## Yes :91 2001 :13   
## NA's:98 2000 :12   
## 1999-2000:11   
## 2002 : 8   
## 2003 : 7   
## (Other) :80   
## NA's :58   
## Latest.industrial.data Latest.trade.data Latest.water.withdrawal.data  
## Min. :1995 Min. :1975 Min. :1990   
## 1st Qu.:2002 1st Qu.:2007 1st Qu.:2000   
## Median :2004 Median :2008 Median :2000   
## Mean :2003 Mean :2007 Mean :2001   
## 3rd Qu.:2005 3rd Qu.:2008 3rd Qu.:2000   
## Max. :2006 Max. :2008 Max. :2006   
## NA's :94 NA's :10 NA's :42   
## X2.alpha.code WB.2.code Table.Name   
## AE : 1 AE : 1 Afghanistan : 1   
## AF : 1 AF : 1 Albania : 1   
## AG : 1 AG : 1 Algeria : 1   
## AL : 1 AL : 1 Angola : 1   
## AM : 1 AM : 1 Antigua and Barbuda: 1   
## (Other):182 (Other):183 Argentina : 1   
## NA's : 2 NA's : 1 (Other) :183   
## Short.Name   
## Afghanistan : 1   
## Albania : 1   
## Algeria : 1   
## Angola : 1   
## Antigua and Barbuda: 1   
## Argentina : 1   
## (Other) :183

#189 of 190 IDs match

Looking at the output of the summary() functions on each data set, we see (from the length) that the Clean\_GDP data set had 190 records, but the merged data set (Comb\_data) has only 189. This tells us that only 189 of the IDs matched.

## Step 4: Sort the data in ascending order (Question 2)

After sorting the data we can see that St. Kitts and Nevis has the 13th smallest GDP. Note that St. Kitts and Nevis has the same GDP as Grenada, but due to sorting rules, St. Kitts and Nevis shows up as the 13th record returned.

#sorting  
Comb\_data<-Comb\_data[order(Comb\_data$USD,decreasing=FALSE),]  
#Q2  
Comb\_data[13,3]

## [1] St. Kitts and Nevis  
## 228 Levels: East Asia & Pacific Euro area ... Zimbabwe

## Step 5: Determine the average GDP rankings for the "High income: OECD" and "High income nonOECD" groups (Question 3)

What if we wanted to know the average GDP rank for "High income: OECD" countries and the "High income: nonOECD" countries? We could use the subset function to limit our data for only countries falling into those categories, and then take the mean of the rank column:

#Q3  
Hi\_OECD<-subset(Comb\_data,Income.Group=="High income: OECD",select=c(Rank,CountryCode,USD,Income.Group))  
mean(Hi\_OECD$Rank)

## [1] 32.96667

Hi\_nonOECD<-subset(Comb\_data,Income.Group=="High income: nonOECD",select=c(Rank,CountryCode,USD,Income.Group))  
mean(Hi\_nonOECD$Rank)

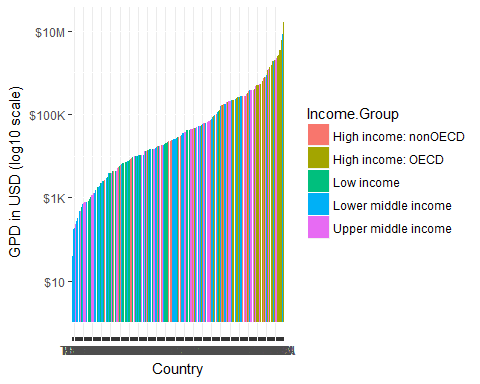
## [1] 91.91304

After doing so, we can see that the High income: OECD countries have a higher average GDP ranking of 32.9667 vs. 91.91304.

## Step 6: Plot, and color code based on income group, the GDP for all of the countries (Question 4)

Charts can often help us understand a large amount of data, very quickly. To accomplish that, we'll plot the GDP in USD on the y-axis, and provide a column for each country on the x-axis. To make the chart even more helpful, we will color code the columns by income group, and put the y-axis on a log10 scale so that the outliers do not overwhelm the other data points:

#Q4  
Comb\_data$CountryCode<-factor(Comb\_data$CountryCode,levels=Comb\_data[order(Comb\_data$USD,decreasing=FALSE),"CountryCode"])  
qplot(CountryCode,USD,data=Comb\_data,fill=Income.Group,geom="col")+scale\_y\_log10(labels=c("$0","$10","$1K","$100K","$10M"))+labs(x="Country",y="GPD in USD (log10 scale)")



## Step 7: Break the country ranks (by GDP) in to quantiles (Question 5)

Finally, we want to know how many countries are considered to be in the Lower middle income group, but in the top quantile of countries in terms of GDP. To accomplish this we'll create a new field called "quantile" and use the cut function to divide our data into quantiles. Then we'll subset the data to only those countires where "Income.Group" is equal to "Lower middle income" and their GDP rank is in the first quantile:

#Q5  
Comb\_data$quantile<-with(Comb\_data,cut(Rank,breaks=quantile(Rank, probs=seq(0,1,by=0.2),na.rm=TRUE),include.lowest=TRUE))  
IncvsGDP<-Comb\_data[(Comb\_data$Income.Group=="Lower middle income") & (Comb\_data$quantile=="[1,38.6]"),c("Rank","CountryCode","Income.Group","quantile")]  
str(IncvsGDP)

## 'data.frame': 5 obs. of 4 variables:  
## $ Rank : num 38 31 16 10 2  
## $ CountryCode : Factor w/ 189 levels "TUV","KIR","MHL",..: 152 159 174 180 188  
## $ Income.Group: Factor w/ 5 levels "High income: nonOECD",..: 4 4 4 4 4  
## $ quantile : Factor w/ 5 levels "[1,38.6]","(38.6,76.2]",..: 1 1 1 1 1

Using the str() function to look at our final data set (which we called IncvsGDP) we can see that four countries fall both into the "Lower middle income" group, and have a GDP ranked in the top quantile.

## Conclusion

Throughout this project we have downloaded data from the internet, saved it onto our machines locally, as well as in our workspace, and then we cleaned and merged the data. After accomplishing this we were able to show the following:

* 189 of the 190 countries in the GDP data set matched with the educational data set
* St. Kitts and Nevis has the 13th ranked GDP in the data set
* The "High income: OECD" group had an average rank of 33.0, which was higher than the "High income: nonOECD" group with an average rank of 91.9
* The United States GDP was significantly higher than the other countries, and therefore is better displayed in a plot on a log scale
* 5 countries (Egypt, Thailand, Indonesia, India, and China) were both in the top quantile of GDPs but also classified as "Lower middle income"