JWheelerCaseStudyUnit6

Jessica Wheeler

Due June 20, 2016

## Introduction

In this case study, we examine 2 datasets from the World Bank: one titled "GDP" containing the GDP ranking for countries and more, and one entited "Education" containing income groups and more information on each country. We will merge these datasets, explore the merged dataset and examine how GDP ranking relates to income group.

## Loading, Cleaning, and Merging the datasets

#This makefile sources r code to load, clean, then merge the two datasets:  
source("Makefile.R")

## Loading required package: repmis

## Loading required package: 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

## Loading required package: ggplot2

#Once sourced, we have an object called merged which contains columns 1-6 of the merged dataset with information on countrycode, GDPranking, country, millionsUSdollars, and income.group.  
  
#Check NA's:  
sum(is.na(merged))

## [1] 137

# Questions

**1. Match the data based on country short code. How many of the IDs match?**  
\* In this step, first we check how many country shortcodes are in each individual dataset.  
\* In order to accomplish this, I've created a vector for the column countrycode in each dataset.

#Checking the country shortcodes for gdpdata:  
gdpcc <- gdpdata[, 1]  
gdpcc

## [1] "USA" "CHN" "JPN" "DEU" "FRA" "GBR" "BRA" "RUS" "ITA" "IND" "CAN"  
## [12] "AUS" "ESP" "MEX" "KOR" "IDN" "TUR" "NLD" "SAU" "CHE" "SWE" "IRN"  
## [23] "NOR" "POL" "BEL" "ARG" "AUT" "ZAF" "VEN" "COL" "THA" "ARE" "DNK"  
## [34] "MYS" "SGP" "CHL" "HKG" "EGY" "NGA" "ISR" "PHL" "GRC" "FIN" "PAK"  
## [45] "PRT" "IRL" "IRQ" "DZA" "PER" "KAZ" "CZE" "ROM" "UKR" "QAT" "NZL"  
## [56] "KWT" "VNM" "HUN" "BGD" "AGO" "PRI" "MAR" "SVK" "ECU" "SYR" "OMN"  
## [67] "CUB" "AZE" "BLR" "LKA" "HRV" "DOM" "SDN" "LUX" "UZB" "BGR" "GTM"  
## [78] "URY" "TUN" "SVN" "CRI" "MAC" "LBN" "LTU" "ETH" "GHA" "KEN" "SRB"  
## [89] "PAN" "YEM" "TKM" "JOR" "BHR" "LVA" "TZA" "BOL" "PRY" "CMR" "CIV"  
## [100] "SLV" "TTO" "CYP" "EST" "ZMB" "AFG" "UGA" "NPL" "HND" "GAB" "GNQ"  
## [111] "BIH" "ZAR" "BRN" "GEO" "PNG" "JAM" "BWA" "MOZ" "SEN" "KHM" "COG"  
## [122] "ISL" "NAM" "TCD" "ALB" "NIC" "MUS" "BFA" "MLI" "MNG" "SSD" "MDG"  
## [133] "ARM" "ZWE" "MKD" "LAO" "MLT" "BHS" "HTI" "BEN" "MDA" "RWA" "TJK"  
## [144] "NER" "KGZ" "KSV" "MCO" "GIN" "BMU" "SUR" "MNE" "MWI" "BRB" "MRT"  
## [155] "FJI" "TGO" "SLE" "SWZ" "ERI" "GUY" "ABW" "BDI" "LSO" "MDV" "CAF"  
## [166] "CPV" "BTN" "LBR" "BLZ" "TMP" "LCA" "ATG" "SYC" "SLB" "GMB" "GNB"  
## [177] "VUT" "GRD" "KNA" "VCT" "WSM" "COM" "DMA" "TON" "FSM" "STP" "PLW"  
## [188] "MHL" "KIR" "TUV"

#Checking the country shortcodes for educationdata:  
educationcc <- educationdata[, 1]  
educationcc

## [1] "ABW" "ADO" "AFG" "AGO" "ALB" "ARE" "ARG" "ARM" "ASM" "ATG" "AUS"  
## [12] "AUT" "AZE" "BDI" "BEL" "BEN" "BFA" "BGD" "BGR" "BHR" "BHS" "BIH"  
## [23] "BLR" "BLZ" "BMU" "BOL" "BRA" "BRB" "BRN" "BTN" "BWA" "CAF" "CAN"  
## [34] "CHE" "CHI" "CHL" "CHN" "CIV" "CMR" "COG" "COL" "COM" "CPV" "CRI"  
## [45] "CUB" "CYM" "CYP" "CZE" "DEU" "DJI" "DMA" "DNK" "DOM" "DZA" "EAP"  
## [56] "EAS" "ECA" "ECS" "ECU" "EGY" "EMU" "ERI" "ESP" "EST" "ETH" "FIN"  
## [67] "FJI" "FRA" "FRO" "FSM" "GAB" "GBR" "GEO" "GHA" "GIN" "GMB" "GNB"  
## [78] "GNQ" "GRC" "GRD" "GRL" "GTM" "GUM" "GUY" "HIC" "HKG" "HND" "HPC"  
## [89] "HRV" "HTI" "HUN" "IDN" "IMY" "IND" "IRL" "IRN" "IRQ" "ISL" "ISR"  
## [100] "ITA" "JAM" "JOR" "JPN" "KAZ" "KEN" "KGZ" "KHM" "KIR" "KNA" "KOR"  
## [111] "KSV" "KWT" "LAC" "LAO" "LBN" "LBR" "LBY" "LCA" "LCN" "LDC" "LIC"  
## [122] "LIE" "LKA" "LMC" "LMY" "LSO" "LTU" "LUX" "LVA" "MAC" "MAR" "MCO"  
## [133] "MDA" "MDG" "MDV" "MEA" "MEX" "MHL" "MIC" "MKD" "MLI" "MLT" "MMR"  
## [144] "MNA" "MNE" "MNG" "MNP" "MOZ" "MRT" "MUS" "MWI" "MYS" "NAC" "NAM"  
## [155] "NCL" "NER" "NGA" "NIC" "NLD" "NOC" "NOR" "NPL" "NZL" "OEC" "OMN"  
## [166] "PAK" "PAN" "PER" "PHL" "PLW" "PNG" "POL" "PRI" "PRK" "PRT" "PRY"  
## [177] "PYF" "QAT" "ROM" "RUS" "RWA" "SAS" "SAU" "SDN" "SEN" "SGP" "SLB"  
## [188] "SLE" "SLV" "SMR" "SOM" "SRB" "SSA" "SSF" "STP" "SUR" "SVK" "SVN"  
## [199] "SWE" "SWZ" "SYC" "SYR" "TCA" "TCD" "TGO" "THA" "TJK" "TKM" "TMP"  
## [210] "TON" "TTO" "TUN" "TUR" "TUV" "TZA" "UGA" "UKR" "UMC" "URY" "USA"  
## [221] "UZB" "VCT" "VEN" "VIR" "VNM" "VUT" "WBG" "WLD" "WSM" "YEM" "ZAF"  
## [232] "ZAR" "ZMB" "ZWE"

We see that there are 234 levels of country shortcodes for educationdata and 190 in gdpdata. When merged, there are now 235 total countrycodes so that means only **1 countrycode was added after merging.**

**2. Sort the data frame in ascending order by GDP rank (so the United States is first). What is the 13th country in the resulting data frame?**

sortedrank <- merged[order(desc(merged$millionsUSdollars),merged$GDPranking),]  
head(sortedrank, n=20)

## countrycode GDPranking country millionsUSdollars  
## 1 USA 1 United States 16244600  
## 2 CHN 2 China 8227103  
## 3 JPN 3 Japan 5959718  
## 4 DEU 4 Germany 3428131  
## 5 FRA 5 France 2612878  
## 6 GBR 6 United Kingdom 2471784  
## 7 BRA 7 Brazil 2252664  
## 8 RUS 8 Russian Federation 2014775  
## 9 ITA 9 Italy 2014670  
## 10 IND 10 India 1841710  
## 11 CAN 11 Canada 1821424  
## 12 AUS 12 Australia 1532408  
## 13 ESP 13 Spain 1322965  
## 14 MEX 14 Mexico 1178126  
## 15 KOR 15 Korea, Rep. 1129598  
## 16 IDN 16 Indonesia 878043  
## 17 TUR 17 Turkey 789257  
## 18 NLD 18 Netherlands 770555  
## 19 SAU 19 Saudi Arabia 711050  
## 20 CHE 20 Switzerland 631173  
## long.name  
## 1 United States of America  
## 2 People's Republic of China  
## 3 Japan  
## 4 Federal Republic of Germany  
## 5 French Republic  
## 6 United Kingdom of Great Britain and Northern Ireland  
## 7 Federative Republic of Brazil  
## 8 Russian Federation  
## 9 Italian Republic  
## 10 Republic of India  
## 11 Canada  
## 12 Commonwealth of Australia  
## 13 Kingdom of Spain  
## 14 United Mexican States  
## 15 Republic of Korea  
## 16 Republic of Indonesia  
## 17 Republic of Turkey  
## 18 Kingdom of the Netherlands  
## 19 Kingdom of Saudi Arabia  
## 20 Switzerland  
## income.group  
## 1 High income: OECD  
## 2 Lower middle income  
## 3 High income: OECD  
## 4 High income: OECD  
## 5 High income: OECD  
## 6 High income: OECD  
## 7 Upper middle income  
## 8 Upper middle income  
## 9 High income: OECD  
## 10 Lower middle income  
## 11 High income: OECD  
## 12 High income: OECD  
## 13 High income: OECD  
## 14 Upper middle income  
## 15 High income: OECD  
## 16 Lower middle income  
## 17 Upper middle income  
## 18 High income: OECD  
## 19 High income: nonOECD  
## 20 High income: OECD

The 13th country in the dataframe is **Spain.**

**3. What are the average GDP rankings for the "High income:OECD" and "High income: Non-OECD" groups?**

* As I came across this question, I had another question: What is OECD? According to [usoecd](http://usoecd.usmission.gov/mission/overview.htmlo), it is the Organization for Economic Cooperation and Development, which is a "unique forum where governments of 34 democracies with market economies work with each other, as well as with more than 70 non-member economies to promote economic growth, prosperity, and sustainable development."

#Let's create a subset of GDP ranking and country name:  
countryrank <- merged  
  
#Let's convert GDPranking into a numeric (for some reason, when I did this before merging the datasets, the ranking became out of order after merging, so we will do this now):  
countryrank$GDPranking <- as.numeric(countryrank$GDPranking)  
  
#Check of NA's and get rid of them:  
sum(is.na(countryrank))

## [1] 137

countryrank <-na.omit(countryrank)  
  
#We could use the tapply function to find the sum of each income group:  
tapply(countryrank$GDPranking, countryrank$income.group, mean)

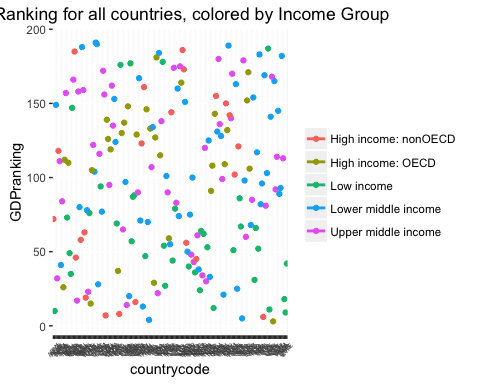
## High income: nonOECD High income: OECD Low income   
## 93.73913 110.06667 66.97297   
## Lower middle income Upper middle income   
## 105.03704 106.13333

As can be seen in the output above, the average GDP ranking for the High Income: OECD group is 110.066 while the average GDP ranking for the High Income: nonOECD group is 93.739.

* The output above also shows the average GDP ranking for the rest of the income groups. From this we can see that the Low income group has the lowest average GDP ranking, followed by the High Income: nonOECD group. The highest average GDP rankings are in the groups Lower middle income, upper middle income, and High income: OECD. Thus we can conclude that being in the high income group does not necessarily indicate a high average GDP ranking.

**4. Plot the GDP for all of the countries. Use ggplot2 to color the plot by Income Group.**

#Creating the basic plot:  
plot <- ggplot(countryrank, aes(countrycode, GDPranking, color=factor(income.group))) +geom\_point() +stat\_smooth(se = F)  
  
#Title for the plot:  
p <- plot+ggtitle('GDP Ranking for all countries, colored by Income Group')  
#Turn off the legend title  
p+theme(legend.title=element\_blank()) +theme(axis.text.x=element\_text(angle=50, size=5, vjust=0.5))



The plot above appears to look like a random cloud. Although the income groups are distinguished by color, each color can be seen on each quadrant of the plot. Therefore, I think it is safe to say that there is no linear relationship between income group and GDP ranking. In other words, a high GDP ranking does not necessarily indicate that a country is in a high income group.

**5. Cut the GDP ranking into 5 separate quantile groups. Make a table vs. Income Group. How many countries are Lower middle income but among the 38 nations with highest GDP?**

#GDP ranking goes from 1 to 191, we will divide the rankings into these 5 groups:  
Q1 <- subset(countryrank, GDPranking <=38)  
Q2 <- subset(countryrank, GDPranking >38 | GDPranking <=76)  
Q3 <- subset(countryrank, GDPranking >76 | GDPranking <=114)  
Q4 <- subset(countryrank, GDPranking >114 | GDPranking <=152)  
Q5 <- subset(countryrank, GDPranking >=152)  
  
#Using the code above, we can use the sum function to answer the question of how many countries are in the lower middle income group and are in Q1, among the 38 nations with the highest GDP:  
sum(Q1$income.group == "Lower middle income")

## [1] 9

#Let's create a new object with only the columns we want to make a data table- which are GDP ranking and income.group:  
quantile <- countryrank[, c(2,6)]  
  
#To make a table of the subsets Q1-Q5 vs. income group:  
#First we'll convert GDPranking into a numeric vector and sort the data while assigning it to a new object:  
quantile$GDPranking <- as.numeric(quantile$GDPranking)  
sortedq <- quantile[order(countryrank$GDPranking),]  
  
  
#Then, let's cut the data into respective quantile groups according to GDP rank and name those Q1 through Q5:  
quantileranks <- cut(sortedq$GDPranking, breaks=c(-0.01, 37.99, 75.99, 113.99, 152.99, 999.99), labels=c("Q1", "Q2", "Q3", "Q4", "Q5"))  
  
#The object quantileranks shows the sets of Q1-Q5 of the dataset.  
#I'll add a new column (Q) to the data frame sortedq to contain the respective quantile groups:  
sortedq$Q <- quantileranks  
#Now this data set contains 3 columns: GDP ranking, income group, and Q (quantile).  
  
#I'll subset the data and choose only the columns we want: Income group and Q:  
datatable <- sortedq[, 2:3]  
datatable

## income.group Q  
## 1 High income: OECD Q1  
## 10 Lower middle income Q1  
## 100 Lower middle income Q1  
## 101 High income: nonOECD Q1  
## 102 High income: nonOECD Q1  
## 103 High income: nonOECD Q1  
## 104 Low income Q1  
## 105 Low income Q1  
## 106 Low income Q1  
## 107 Low income Q1  
## 108 Lower middle income Q1  
## 109 Upper middle income Q1  
## 11 High income: OECD Q1  
## 110 High income: nonOECD Q1  
## 111 Upper middle income Q1  
## 112 Low income Q1  
## 113 High income: nonOECD Q1  
## 114 Lower middle income Q1  
## 115 Lower middle income Q1  
## 116 Upper middle income Q1  
## 117 Upper middle income Q1  
## 118 Low income Q1  
## 119 Lower middle income Q1  
## 12 High income: OECD Q1  
## 120 Low income Q1  
## 121 Lower middle income Q1  
## 122 High income: OECD Q1  
## 123 Upper middle income Q1  
## 124 Low income Q1  
## 125 Upper middle income Q1  
## 126 Lower middle income Q1  
## 127 Upper middle income Q1  
## 128 Low income Q1  
## 129 Low income Q1  
## 13 High income: OECD Q1  
## 130 Lower middle income Q2  
## 131 Low income Q2  
## 132 Lower middle income Q2  
## 133 Low income Q2  
## 134 Upper middle income Q2  
## 135 Low income Q2  
## 136 High income: nonOECD Q2  
## 137 High income: nonOECD Q2  
## 138 Low income Q2  
## 14 Upper middle income Q2  
## 139 Low income Q2  
## 140 Lower middle income Q2  
## 141 Low income Q2  
## 142 Low income Q2  
## 143 Low income Q2  
## 144 Low income Q2  
## 145 Lower middle income Q2  
## 146 High income: nonOECD Q2  
## 147 Low income Q2  
## 148 High income: nonOECD Q2  
## 15 High income: OECD Q2  
## 149 Upper middle income Q2  
## 150 Upper middle income Q2  
## 151 Low income Q2  
## 152 High income: nonOECD Q2  
## 153 Low income Q2  
## 154 Upper middle income Q2  
## 155 Low income Q2  
## 156 Low income Q2  
## 157 Lower middle income Q2  
## 158 Low income Q2  
## 16 Lower middle income Q2  
## 159 Lower middle income Q2  
## 160 High income: nonOECD Q2  
## 161 Low income Q2  
## 162 Lower middle income Q2  
## 163 Lower middle income Q2  
## 164 Low income Q3  
## 165 Lower middle income Q3  
## 166 Lower middle income Q3  
## 167 Low income Q3  
## 168 Lower middle income Q3  
## 17 Upper middle income Q3  
## 169 Lower middle income Q3  
## 170 Upper middle income Q3  
## 171 Upper middle income Q3  
## 172 Upper middle income Q3  
## 173 Low income Q3  
## 174 Low income Q3  
## 175 Low income Q3  
## 176 Lower middle income Q3  
## 177 Upper middle income Q3  
## 178 Upper middle income Q3  
## 18 High income: OECD Q3  
## 179 Upper middle income Q3  
## 180 Lower middle income Q3  
## 181 Low income Q3  
## 182 Upper middle income Q3  
## 183 Lower middle income Q3  
## 184 Lower middle income Q3  
## 185 Lower middle income Q3  
## 186 Upper middle income Q3  
## 187 Lower middle income Q3  
## 188 Lower middle income Q3  
## 19 High income: nonOECD Q3  
## 189 Lower middle income Q3  
## 2 Lower middle income Q3  
## 20 High income: OECD Q3  
## 21 High income: OECD Q3  
## 22 Upper middle income Q3  
## 23 High income: OECD Q3  
## 24 High income: OECD Q3  
## 25 High income: OECD Q3  
## 26 Upper middle income Q3  
## 27 High income: OECD Q3  
## 28 Upper middle income Q3  
## 29 Upper middle income Q4  
## 3 High income: OECD Q4  
## 30 Upper middle income Q4  
## 31 Lower middle income Q4  
## 32 High income: nonOECD Q4  
## 33 High income: OECD Q4  
## 34 Upper middle income Q4  
## 35 High income: nonOECD Q4  
## 36 Upper middle income Q4  
## 37 High income: nonOECD Q4  
## 38 Lower middle income Q4  
## 39 Lower middle income Q4  
## 4 High income: OECD Q4  
## 40 High income: OECD Q4  
## 41 Lower middle income Q4  
## 42 High income: OECD Q4  
## 43 High income: OECD Q4  
## 44 Lower middle income Q4  
## 45 High income: OECD Q4  
## 46 High income: OECD Q4  
## 47 Lower middle income Q4  
## 48 Upper middle income Q4  
## 49 Upper middle income Q4  
## 5 High income: OECD Q4  
## 50 Upper middle income Q4  
## 51 High income: OECD Q4  
## 52 Upper middle income Q4  
## 53 Lower middle income Q4  
## 54 High income: nonOECD Q4  
## 55 High income: OECD Q4  
## 56 High income: nonOECD Q4  
## 57 Lower middle income Q4  
## 58 High income: OECD Q4  
## 59 Low income Q4  
## 6 High income: OECD Q4  
## 60 Lower middle income Q4  
## 61 High income: nonOECD Q4  
## 62 Lower middle income Q4  
## 63 High income: OECD Q4  
## 64 Lower middle income Q5  
## 65 Lower middle income Q5  
## 66 High income: nonOECD Q5  
## 67 Upper middle income Q5  
## 68 Upper middle income Q5  
## 69 Upper middle income Q5  
## 7 Upper middle income Q5  
## 70 Lower middle income Q5  
## 71 High income: nonOECD Q5  
## 72 Upper middle income Q5  
## 73 Lower middle income Q5  
## 74 High income: OECD Q5  
## 75 Lower middle income Q5  
## 76 Upper middle income Q5  
## 77 Lower middle income Q5  
## 78 Upper middle income Q5  
## 79 Lower middle income Q5  
## 8 Upper middle income Q5  
## 80 High income: OECD Q5  
## 81 Upper middle income Q5  
## 82 High income: nonOECD Q5  
## 83 Upper middle income Q5  
## 84 Upper middle income Q5  
## 85 Low income Q5  
## 86 Low income Q5  
## 87 Low income Q5  
## 88 Upper middle income Q5  
## 89 Upper middle income Q5  
## 9 High income: OECD Q5  
## 90 Lower middle income Q5  
## 91 Lower middle income Q5  
## 92 Lower middle income Q5  
## 93 High income: nonOECD Q5  
## 94 High income: nonOECD Q5  
## 95 Low income Q5  
## 96 Lower middle income Q5  
## 97 Lower middle income Q5  
## 98 Lower middle income Q5  
## 99 Lower middle income Q5

As can be seen in the output above, there are 9 out of the 38 nations with the highest GDP ranking that classify as Lower middle income.

# Conclusion

In merging the two datasets above, we could examine the relationship between the GDP ranking of a country as it relates to the country's income group category.

* First we observed the merged dataset and saw that only one country shortcode was created which means all the others in each dataset were "matched."
* Then, we saw that the average GDP ranking for the High Income: OECD group is a bit higher than that of the High Income: nonOECD group.
* The scatterplot showed that GDP ranking does not indicate income group.
* Finally, there were a total of 9 out of the 38 countries that have the highest GDP ranking that are classified as Lower middle income. This shows that a country does not necessarily need to be classified as high income to have a high GDP ranking.