Armand Post

MSDS 6306 – Section 403

06/20/2016

Unit 6 Case Study

**Introduction**

The purpose of this case study is to assess GDP data available from [www.worldbank.org](http://www.worldbank.org). The deliverable is a github repository (found at <https://github.com/armandpost/CaseStudyWeek6>) containing the following:

* Readme file (contains a brief explanation of the github repository).
* Data dictionary
* Original datasets in .csv format
* R-Studio files (.rmd, .md, .rProj, .html)

**Methods**

R-Studio was used to pull the data in, scrub the data, and produce graphs and tables. Two datasets were merged together with the purpose of attaching income groupings to the GDP data. These income groupings include the below:

* High Income: OECD
* High Income: Non OECD
* Upper Middle Income
* Lower Middle Income
* Low Income

Once the data merge was complete, countries with missing GDP values were removed. There were 24 countries that were removed which are listed below:

American Samoa, Andorra, Cayman Islands, Channel Islands, Curaçao, Djibouti, Faeroe Islands, French Polynesia, Greenland, Guam, Isle of Man, Korea, Dem. Rep., Libya, Liechtenstein, Myanmar, New Caledonia, Northern Mariana Islands, San Marino, Sint Maarten (Dutch part), Somalia, St. Martin (French part), Turks and Caicos Islands, Virgin Islands (U.S.), West Bank and Gaza

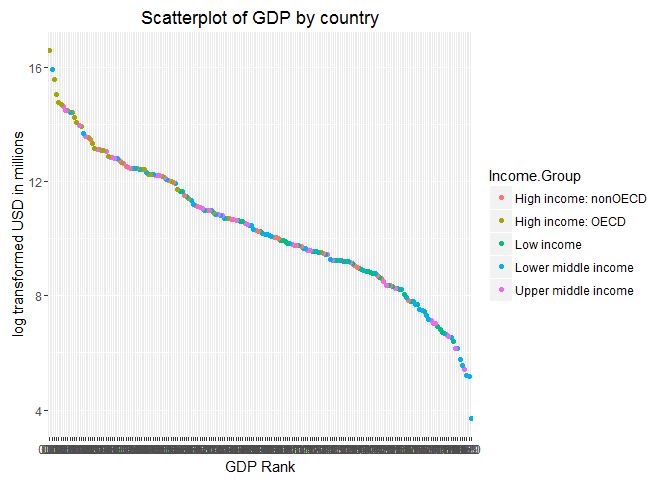
There were also a number of non-country specific records which were removed so that only countries would be plotted. These include:

World, low income, middle income, lower middle income, upper middle income, low & middle income, high income, East Asia & Pacific, Europe & Central Asia, Latin America & Caribbean, Middle East & North Africa, South Asia, Sub-Saharan Africa, and Euro area

Additionally, one country contained no income group classification so was excluded from the analysis bringing the total country count to 189.

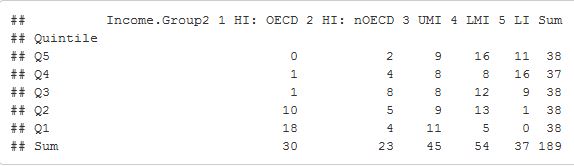
**Results**

Once all of the data was merged and scrubbed, an assessment was performed by creating a graph of every country’s GDP color coded by income group. Initially the graph was created without a log transformation, however since the United States, China, and Japan all have GDPs greater than $5 trillion while the rest of the countries are below $5 trillion, the graph was difficult to interpret due to the scale. A log transformation was created in order to better visualize.



From this plot it can be inferred that OECD countries are more likely to have an overall GDP higher than non OECD countries. The middle income countries appear to be spread around the middle while low income countries appear mostly in the lower GDP Ranks. Since this graph has a large number of data points, some tables were created to further assess.

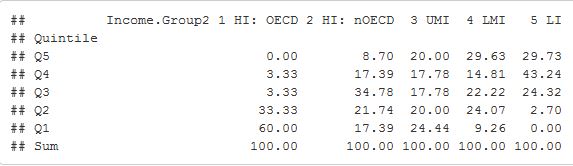
The GDP Ranks were split into quintiles and crosstabs were created showing quintile ranking vs income group. The table below shows counts of countries in each crosstab segmentation.



We can see from this table that a number of exceptions occur for lower middle income countries. These are nations with high populations which drive their total GDP amongst the high income countries. For instance, five lower middle income countries appear in the highest quintile. These include:

* China
* India
* Indonesia
* Thailand
* Egypt

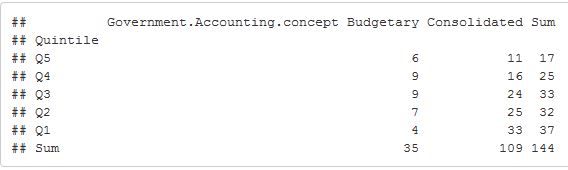
An additional crosstab was created which show the percentage distributions in quintiles for each income group.

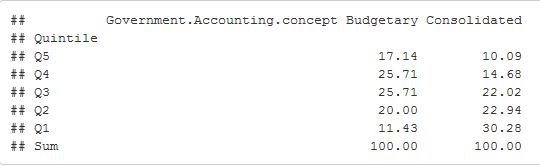


This crosstab shows:

* High income OECD countries appear overwhelmingly in the two upper quintiles.
* High income non OECD countries have a somewhat normal distribution amongst the quintiles.
* Upper middle income countries have an evenly spread distribution amongst the quintiles.
* Lower middle income countries skew more towards the middle and lower quintiles, with some exceptions in the higher ranges.
* Lower income countries are heavily skewed towards the lower quintiles.

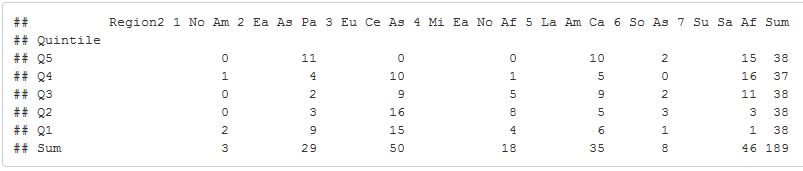
Additional analysis was performed to see if government accounting concepts and / or region are predictive of total GDP. The next tables show counts and column percentages for government accounting concepts.

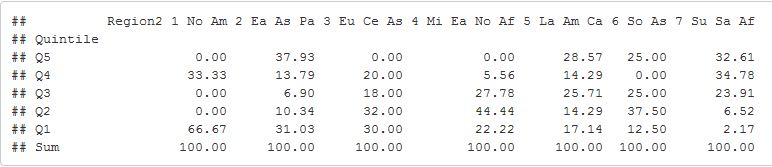
****

****

From the above, 45 countries did not have a record for government accounting concept. Consolidated accounting concepts were used 76% of the time vs 24% for budgetary. While the sample size is somewhat small for budgetary, consolidated accounting concepts are more often used in higher GDP ranked countries which may indicate it leads to higher economic growth.

Finally, geographic region was assessed to see if there are regional clusters of high GDP countries.





From the above, all 189 countries had a region category assigned. The data becomes somewhat thin, but some trends can be observed.

* North America only has three countries, but two appear in the upper quintile.
* East Asia and the South Pacific has a U shaped curve with most countries in the highest and lowest quintiles.
* Europe and Central Asia is skewed towards the two upper quintiles.
* The Middle East and North Africa is skewed towards the two upper quintiles, but has some countries in the lower two.
* Latin America and the Caribbean has the closest to a normal distribution, but is more skewed towards the lower quintiles.
* South Asia has a small sample size, but skews more towards the upper quintiles.
* Sub-Saharan Africa is heavily skewed towards the lower quintiles.

**Conclusion**

While some mixed results were observed, a number of conclusions can be drawn. First of which is that being classified as an OECD nation is indicative of high GDP. Also, high GDP nations tend to have a consolidated government accounting concept. There are also regional clusters of high GDP in North America, East Asia & the Pacific, Europe & Central Asia, and the Middle East & North Africa. Conversely, low GDP nations appear clustered in Latin America & the Caribbean and Sub-Saharan Africa.

While conclusions can be drawn from looking at total GDP, an assessment of GDP per capita may be more telling due to large population differences between countries. However, population data was not present in the datasets used. An attempt was made to merge a dataset with GDP per capita obtained from www.worldbank.org with the original dataset. The merge was successful, however due to my limitations with R, I was unable to convert the GDP per capita field to a format acceptable to the ddply function in the dplyr library.

**Appendix**

The appendix section includes .Rmd code for assessing the data after it was cleaned.

### Question 1. Show row counts. Note that two records were in GDP, but not Education (Andorra, South Sudan). Additonally, 21 records were in Education, but not GDP

```{r Rows}

nrow(GDPRecords)

nrow(EduRecords)

nrow(FullData2)

```

### Remove the records that didn't match and non-country specific data. Also add in a GDP rank column with leading zeros for sorting and sort by descending order. Once removing records with no GDP rank, the record count is reduced to 190 from 235.

```{r Order}

FullData3 <- subset(FullData2, GDPRank != "")

FullData3$GDPRank2 <- str\_pad(FullData3$GDPRank, width=3, side="left", pad="0")

FullData3 <- FullData3[order(FullData3$GDPRank2, decreasing = TRUE),]

nrow(FullData3)

head(FullData3)

```

### For Question 2, once the data is in descending order, show the 13th from lowest GDP country.

#### From the below, we can see that Grenada and St. Kitts and Nevis are tied for 13th from last.

```{r Show 13}

head(FullData3,13)

```

### For Question 3, What are the average GDP rankings for the "High income: OECD" and "High income: non OECD" groups?

```{r Average GDP}

## Convert GDP to numeric

FullData3$GDP = gsub(pattern = ",", replacement = "", FullData3$GDP)

FullData3[, c(4)] <- sapply(FullData3[, c(4)], as.numeric)

ddply(FullData3, .(Income.Group), summarize, GDPMean = mean(GDP))

head(FullData3)

```

### High income: nonOECD countries have a mean GDP of $104,349 USD in millions while High income: OECD countries have a mean GDP of $1,483,917 USD in millions.

### Question 4. Plot the GDP data for all countries

```{r Plot}

scale\_y\_continuous(labels=function(n){format(n, scientific = FALSE)})

Graph1 <- ggplot(FullData3, aes(x = GDPRank2, y = GDP, color = Income.Group))

Graph1 + geom\_point() + scale\_y\_continuous(labels = comma) + geom\_point() + xlab("GDP Rank") + ylab("USD in millions") + ggtitle("Scatterplot of GDP by country")

```

### Question 4. Since the data has a few high GDP countries and many low GDP countries, the above graph is difficult to interpret due to the scale being off. The following step performs a Log Transformation of GDP and replots.

```{r Log Plot}

scale\_y\_continuous(labels=function(n){format(n, scientific = FALSE)})

Graph1 <- ggplot(FullData3, aes(x = GDPRank2, y = log(GDP), color = Income.Group))

Graph1 + geom\_point() + scale\_y\_continuous(labels = comma) + geom\_point() + xlab("GDP Rank") + ylab("log transformed USD in millions") + ggtitle("Scatterplot of GDP by country")

```

### Once the data is log transformed, it takes on an S-Shape. We can also see high income: OECD countries make up most of the high GDP Rank countries, while high income: nonOECD countries tend to lag behind their OECD counterparts. There are some exceptions, most notably China which has a very high GDP rank despite being classified as a lower middle income nation. Lower middle income countries typically appear around the middle and low end of GDP Rank.

### Cut the GDP rankings into quintile Groups.

```{r Quintiles}

FullData3$Quintile <- with(FullData3, factor(findInterval(GDP, c(-Inf, quantile(GDP, probs=c(0.20, 0.40, 0.60, 0.80)), Inf)), labels=c("Q5","Q4","Q3","Q2","Q1")))

```

### Make a crosstab of Quintiles vs Income Groups

```{r Crosstab Income}

# Load function

source("http://pcwww.liv.ac.uk/~william/R/crosstab.r")

# create new variable for income group so it can be shortened for display purposes

FullData3$Income.Group2 <- NA

# shorten income group variable names to fit and create order

FullData3$Income.Group2[FullData3$Income.Group == "High income: OECD"] <- "1 HI: OECD"

FullData3$Income.Group2[FullData3$Income.Group == "High income: nonOECD"] <- "2 HI: nOECD"

FullData3$Income.Group2[FullData3$Income.Group == "Upper middle income"] <- "3 UMI"

FullData3$Income.Group2[FullData3$Income.Group == "Lower middle income"] <- "4 LMI"

FullData3$Income.Group2[FullData3$Income.Group == "Low income"] <- "5 LI"

# Frequency count

crosstab(FullData3, row.vars = "Quintile", col.vars = "Income.Group2", type = "f")

# Total Percentages

crosstab(FullData3, row.vars = "Quintile", col.vars = "Income.Group2", type = "t", percentages = FALSE)

# Column percentages

crosstab(FullData3, row.vars = "Quintile", col.vars = "Income.Group2", type = "c")

```

### Question 5. From the above, there are 5 countries that are lower middle income and in the first quintile. These countries are in order of GDP Rank: China, India, Indonesia, Thailand, and Egypt. We can also see from the counts that one country did not have an income group classifcation and was excluded from both the above plot and tables.

### Also of note are the distributions for the income group categories. High income: OECD countries tend to fall in the two upper quintiles while high income: non OECD have a somewhat normal distribution amongst quintiles. Upper middle income countries have the most evenly spread distributions amongst quintiles. Lower middle income and lower income are skewed towards the lower quartiles.

## The code below contains additional analysis beyond the five questions. Government accounting concepts and geographical region are assessed to see if they may be indicitive of a high income country.

### Make a crosstab of Quintiles vs government accounting concepts.

```{r Crosstab Accounting}

# Load function

source("http://pcwww.liv.ac.uk/~william/R/crosstab.r")

# Frequency count

crosstab(FullData3, row.vars = "Quintile", col.vars = "Government.Accounting.concept", type = "f")

# Total Percentages

crosstab(FullData3, row.vars = "Quintile", col.vars = "Government.Accounting.concept", type = "t", percentages = FALSE)

# Column percentages

crosstab(FullData3, row.vars = "Quintile", col.vars = "Government.Accounting.concept", type = "c")

```

### From the above, 45 countries did not have a record for government accounting concept. Consolidated accounting concepts were used 76% of the time vs 24% for budgetary. While the sample size is somewhat small for budgetary, consolidated accounting concepts are more often used in higher GDP ranked countries which may indicate it leads to higher economic growth.

### Make crosstabs of Quintiles vs Region

```{r Crosstab Region}

# Load function

source("http://pcwww.liv.ac.uk/~william/R/crosstab.r")

# create new variable for income group so it can be shortened for display purposes

FullData3$Region2 <- NA

# shorten income group variable names to fit and create order

FullData3$Region2[FullData3$Region == "North America"] <- "1 No Am"

FullData3$Region2[FullData3$Region == "East Asia & Pacific"] <- "2 Ea As Pa"

FullData3$Region2[FullData3$Region == "Europe & Central Asia"] <- "3 Eu Ce As"

FullData3$Region2[FullData3$Region == "Middle East & North Africa"] <- "4 Mi Ea No Af"

FullData3$Region2[FullData3$Region == "Latin America & Caribbean"] <- "5 La Am Ca"

FullData3$Region2[FullData3$Region == "South Asia"] <- "6 So As"

FullData3$Region2[FullData3$Region == "Sub-Saharan Africa"] <- "7 Su Sa Af"

# Frequency count

crosstab(FullData3, row.vars = "Quintile", col.vars = "Region2", type = "f")

# Total Percentages

crosstab(FullData3, row.vars = "Quintile", col.vars = "Region2", type = "t", percentages = FALSE)

# Column percentages

crosstab(FullData3, row.vars = "Quintile", col.vars = "Region2", type = "c")

```

## From the above, all 189 countries had a region category assigned. The data becomes somewhat thin, but some trends can be observed. North America only has three countries, but two appear in the upper quintile. East Asia and the South Pacific has a U shaped curve with most countries in the highest and lowest quintiles. Europe and Central Asia is skewed towards the two upper quintiles. The Middle East and North Africa is skewed towards the two upper quintiles, but has some countries in the lower two. Latin America and the Caribbean has the closest to a normal distribution, but is more skewed towards the lower quintiles. South Asia has a small sample size, but skews more towards the upper quintiles. Sub-Saharan Africa is heavily skewed towards the lower quintiles.