Analysis supporting Case Study 1

Analysis

Question 1:

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

Answer:

The total number of records returned in the joined data set is 189. This number matches the total number of records in the gdp data frame and we thus have a perfect match using the country codes. A 100% join match indicates that our method for removing NA’s by omitting extraneous columns was successful.

# gdp\_educ <- merge(gdp, educNoNA)

#head(gdp\_educ)

#gdp\_educ

#educNoNA

gdp\_educraw <- merge(gdp, educ)

head(gdp\_educraw)

## CountryCode Rank CountryFullName Economy Income.Group

## 1 ABW 161 Aruba 2584 High income: nonOECD

## 2 AFG 105 Afghanistan 20497 Low income

## 3 AGO 60 Angola 114147 Lower middle income

## 4 ALB 125 Albania 12648 Upper middle income

## 5 ARE 32 United Arab Emirates 348595 High income: nonOECD

## 6 ARG 26 Argentina 475502 Upper middle income

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?

Answer: St. Kitts and Nevis

To obtain this answer, it is necessary to convert the Economy variable to a numeric data type and then sort ascending so that the highest GDP is last in the list. Once sorted, the 13th record is St. Kitts and Nevis.

order\_gdp <-order(gdp\_educraw$Economy)

head (gdp\_educraw[order\_gdp,])

## CountryCode Rank CountryFullName Economy Income.Group

## 173 TUV 190 Tuvalu 40 Lower middle income

## 92 KIR 189 Kiribati 175 Lower middle income

## 113 MHL 188 Marshall Islands 182 Lower middle income

## 137 PLW 187 Palau 228 Upper middle income

## 155 STP 186 São Tomé and Principe 263 Lower middle income

## 59 FSM 185 Micronesia, Fed. Sts. 326 Lower middle income

# gdp\_educraw [13,] This code returns the original position of the record, not the newly sorted position

Question 3:

What are the average GDP rankings for the “High income: OECD” and “High income: nonOECD” groups?

Answer:

Average Rank of High Income OECD group: 32.9667

Average Rank of High Income nonOECD group: 91.91304

Note: To obtain this answer, the data was subsetted into 2 data frames called HighIncome\_OECD and HighIncome\_nonOECD. Data containing only records with Income.Group equal to HighIncome: OECD was subset or queried and stored in the HighIncome\_OECD data frame. Data containing only records with Income.Group equal to HighIncome: nonOECD was subset or queried and stored in the HighIncome\_nonOECD data frame. Subsequently, the mean was taken of both data frames to obtain the answers 32.9667 and 91.91304.

The inferences that could be made about this analysis is that the HighIncome: OECD have a much better GDP ranking than the HighIncome: nonOECD group. In fact, the average rank of HighIncome: OECD is nearly 3 times as better as HighIncome: nonOECD.

HighIncome\_OECD <- subset(x=gdp\_educraw, Income.Group == 'High income: OECD')

mean(HighIncome\_OECD$Rank)

## [1] 32.96667

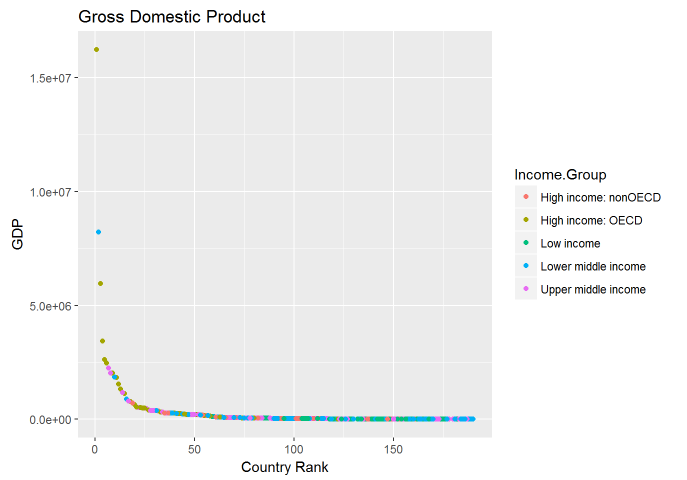
HighIncome\_nonOECD <- subset(x=gdp\_educraw, Income.Group == 'High income: nonOECD')

mean(HighIncome\_nonOECD$Rank)

## [1] 91.91304

Question 4:

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



Answer:

Looking at the graph below, it is apparent that there are 4 distinct outliers with very high GDP’s. The GDP numbers drop sharply until around the 25th ranked country. The numbers continue to drop steadily until around the 50th ranked country and then the numbers stay relatively flat-lined until the last country.

The logical deduction that can be made from this graph is that there are only a few very wealthy nations in the world and about 20 fairly wealthy nations, however, the vast majority of nations have an extremely low GDP compared to the wealthy nations.

#Istall and load ggplot2

library("ggplot2")

## Warning: package 'ggplot2' was built under R version 3.3.3

ggplot(gdp\_educraw, aes(x = Rank, y = Economy, colour = Income.Group)) + geom\_point() + ggtitle("Gross Domestic Product")+ labs(x="Country Rank",y="GDP")

Question 5:

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?

Answer:

5 Countries are in the lower middle income group and also in the highest ranked GDP.

The logical conclusion that can be made regarding this analysis is that even though a country has a high GDP, it does not guarantee that people dwelling in these nations have a high income.

# Sort the data in ascending order on Rank

order\_rank <-order(gdp\_educraw$Rank)

order\_rank

## [1] 178 34 87 45 58 61 25 145 84 78 31 9 53 112 94 77 172

## [18] 128 147 32 159 80 129 139 13 6 10 186 181 38 165 5 47 123

## [35] 150 33 72 51 126 83 136 68 56 133 141 79 81 49 135 88 44

## [52] 144 176 143 131 96 182 76 16 3 140 107 157 50 162 132 42 11

## [69] 21 101 74 48 148 104 179 17 70 177 171 158 41 106 98 103 55

## [86] 63 89 154 134 185 167 86 18 105 174 24 142 36 35 153 170 43

## [103] 54 188 2 175 130 73 60 67 20 187 27 62 138 85 29 119 149

## [120] 91 37 82 124 163 4 127 121 15 115 118 110 7 189 114 97 116

## [137] 19 75 14 109 146 166 125 90 95 108 64 23 156 117 122 26 120

## [154] 57 164 152 160 52 71 1 12 102 111 30 40 28 99 22 168 100

## [171] 8 161 151 65 66 183 69 93 180 184 39 46 169 59 155 137 113

## [188] 92 173

head(gdp\_educraw[order\_rank,])

## CountryCode Rank CountryFullName Economy Income.Group

## 178 USA 1 United States 16244600 High income: OECD

## 34 CHN 2 China 8227103 Lower middle income

## 87 JPN 3 Japan 5959718 High income: OECD

## 45 DEU 4 Germany 3428131 High income: OECD

## 58 FRA 5 France 2612878 High income: OECD

## 61 GBR 6 United Kingdom 2471784 High income: OECD

# Cut the data into 5 equal quantile groups

gdp\_rank <- cut(gdp\_educraw$Rank, 5, include.lowest = TRUE, labels = c("Rank\_Highest", "Rank\_MedHigh", "Rank\_Medium", "Rank\_MedLow", "Rank\_Lowest"))

# Append gdp\_rank list as a variable called rank\_group to gdp\_educraw data frame

gdp\_educraw$rank\_group <- gdp\_rank

# Create table with variables Income.Group and rank\_group

IncomeGroup\_RankGroup <- gdp\_educraw[, c("Income.Group", "rank\_group")]

# Convert Income.Group and rank\_group to factors for subsetting

IncomeGroup\_RankGroup$Income.Group <- as.factor(IncomeGroup\_RankGroup$Income.Group)

# Subset records where Income.Group is lower middle and rank\_group is highest

Low\_High <- subset(IncomeGroup\_RankGroup, Income.Group == "Lower middle income" & rank\_group == "Rank\_Highest",select=Income.Group:rank\_group)

Low\_High

## Income.Group rank\_group

## 34 Lower middle income Rank\_Highest

## 51 Lower middle income Rank\_Highest

## 77 Lower middle income Rank\_Highest

## 78 Lower middle income Rank\_Highest

## 165 Lower middle income Rank\_Highest