CaseStudy1_Pefthimion

Introduction

With the given data files we have information on the annual Gross Domestic Product (GDP) for 190 countries in 2012. Each of them are also assigned a GDP ranking (from # 1 to 190). We are also provided with a further categorization of 5 different income groups that each ranked country falls into.

The first file provides us with the annual GDP (in terms of millions of US Dollars) as well as each country's GDP ranking. The second data file has a lot of information, mostly about the educational system of each country, but we will be primarily looking at each country's income group.

We will be analyzing how a country's ranking and GDP relates to it's income group.

Downloading

First, we will being by downloading the 2 files where we pulled our data from. This data was pulled from the internet as .csv files and was provided to us from you, our client.

```
source("Download_Files_Case_Study_1.R")
```

```
##
## 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
```

Tidying

Next we will begin by tidying up the data set. Many countries in the data set do not have any information on their GDP or rankings. We will not be using that information for our analysis. Also, there were regions listed, which are not countries. Those too will not be used in our analysis. The data set with educational information contains a lot of information that we will not be using. These columns will also be removed in order to give us a cleaner data set.

```
# Cleans the data of blanks and NAs for both data sets
source('Tidying_Case_Study_1.R')
```

Merging and Question 1

We are being asked to merge all of the data based on the country shortcode (CountryCode) and need to find if all of the IDs match.

We will now merge the 2 data sets together so we have one file with the CountryCode, the GDP of each country, their rankings, and each country's income group. Before the merge happens, we have to change the type of some of the data. We will be making the GDP numbers, the CountryCode, and Income Groups characters. This will facilitate the merge. The data will now be easier to read in one place.

```
# merges the GDPdata2 and Edu2 into Merge1
source('Merge_Case_Study_1.R')
```

Now we will try to find how many countries matched. To find out how many match we will remove any blanks or NAs in the Rankings. This is because every country from one of the datasets had a ranking. This will subset out any that don't and therefore do not match. It will also potentialy help us further tidy up the data.

```
# Eliminate any black spaces and NAs from Merge1
Merge1 <- subset(Merge1, Merge1$Ranking !="")
Merge1 <- Merge1[!is.na(Merge1$Ranking),]

# Now we wil look by income group, the other variable not in both of the original data s
ets. It will tell us if anything does not match.
Merge1$Income.Group
```

```
##   [1] High income: nonOECD Low income           Lower middle income
##   [4] Upper middle income  High income: nonOECD Upper middle income
##   [7] Lower middle income  Upper middle income  High income: OECD
##  [10] High income: OECD    Upper middle income  Low income
##  [13] High income: OECD    Low income           Low income
##  [16] Low income           Upper middle income  High income: nonOECD
##  [19] High income: nonOECD Upper middle income  Upper middle income
##  [22] Lower middle income  High income: nonOECD Lower middle income
##  [25] Upper middle income  High income: nonOECD High income: nonOECD
##  [28] Lower middle income  Upper middle income  Low income
##  [31] High income: OECD    High income: OECD    Upper middle income
##  [34] Lower middle income  Lower middle income  Lower middle income
##  [37] Lower middle income  Upper middle income  Low income
##  [40] Lower middle income  Upper middle income  Upper middle income
##  [43] High income: nonOECD High income: OECD    High income: OECD
##  [46] Upper middle income  High income: OECD    Upper middle income
##  [49] Upper middle income  Lower middle income  Lower middle income
##  [52] Low income           High income: OECD    High income: nonOECD
##  [55] Low income           High income: OECD    Upper middle income
##  [58] High income: OECD    Lower middle income  Upper middle income
##  [61] High income: OECD    Lower middle income  Low income
##  [64] Low income           Low income           Low income
##  [67] High income: nonOECD High income: OECD    Upper middle income
##  [70] Lower middle income  Lower middle income  High income: nonOECD
##  [73] Lower middle income  High income: nonOECD Low income
##  [76] High income: OECD    Lower middle income  Lower middle income
##  [79] High income: OECD    Upper middle income  Lower middle income
##  [82] High income: OECD    High income: OECD    High income: OECD
##  [85] Upper middle income  Lower middle income  High income: OECD
##  [88] Upper middle income  Low income           Low income
##  [91] Low income           Lower middle income  Upper middle income
##  [94] High income: OECD    Lower middle income  High income: nonOECD
##  [97] Low income           Upper middle income  Low income
## [100] Upper middle income  Lower middle income  Lower middle income
## [103] Upper middle income  High income: OECD    High income: nonOECD
## [106] High income: nonOECD Lower middle income  High income: nonOECD
## [109] Lower middle income  Low income           Lower middle income
## [112] Upper middle income  Lower middle income  Upper middle income
## [115] Low income           High income: nonOECD Upper middle income
## [118] Lower middle income  Low income           Low income
## [121] Upper middle income  Low income           Upper middle income
## [124] Upper middle income  Low income           Lower middle income
## [127] Lower middle income  High income: OECD    High income: OECD
## [130] Low income           High income: OECD    High income: nonOECD
## [133] Lower middle income  Upper middle income  Upper middle income
## [136] Lower middle income  Upper middle income  Lower middle income
## [139] High income: OECD    High income: nonOECD High income: OECD
## [142] Lower middle income  High income: nonOECD Upper middle income
## [145] Upper middle income  Low income           High income: nonOECD
## [148] Lower middle income  Lower middle income  High income: nonOECD
## [151] Low income           Low income           Lower middle income
## [154] Upper middle income  <NA>                 Lower middle income
## [157] Upper middle income  High income: OECD    High income: OECD
```

```
## [160] High income: OECD    Lower middle income  Upper middle income
## [163] Lower middle income  Low income           Low income
## [166] Lower middle income  Low income           Lower middle income
## [169] Lower middle income  Lower middle income  High income: nonOECD
## [172] Lower middle income  Upper middle income  Lower middle income
## [175] Low income           Low income           Lower middle income
## [178] Upper middle income  High income: OECD    Lower middle income
## [181] Upper middle income  Upper middle income  Lower middle income
## [184] Lower middle income  Lower middle income  Lower middle income
## [187] Upper middle income  Low income           Low income
## [190] Low income
## 6 Levels:  High income: nonOECD High income: OECD ... Upper middle income
```

```
Merge1$Income.Group[155]
```

```
## [1] <NA>
## 6 Levels:  High income: nonOECD High income: OECD ... Upper middle income
```

```
Merge1[155,]
```

```
##     CountryCode Ranking  Table.Name   GDP Income.Group
## 173         SSD     131 South Sudan 10220         <NA>
```

We have 190 IDs listed and have found one that does not match. Therefore, we have found that 189 IDs match. South Sudan does not have an income group. Therefore, it must have existed in one data set, but not the other. In order to keep our data clean, we will remove South Sudan from our further analysis because we will be dealing with income groups and South Sudan does not have data for one in the information that we have been given.

```
# We will now remove South Sudan from our data set, as we cannot have any NAs in our inc
ome group data
Merge1 <- Merge1[!is.na(Merge1$Income.Group),]
```

Question 2

We are asked to find the country with the 13th smallest GDP. To do this we have to put our GDP in ascending order so the country with the smallest GDP is at the top. This is why we had to convert GDP into numeric from a factor before.

```
# Find country with 13th lowest GDP
Merge1.Ascending <- Merge1[order(Merge1$GDP),]
Merge1.Ascending[13,]
```

```
##     CountryCode Ranking          Table.Name GDP        Income.Group
## 102         KNA     178 St. Kitts and Nevis 767 Upper middle income
```

The country with the 13th smallest GDP of our ranked countries is St. Kitts and Nevis (KNA). In 2012, St. Kitts and Nevis had an annual GDP of $767,000,000. Remeber, our data's GDP is in terms of millions of US dollars.

Question 3

We need to look at the GDP rankings for countries that are in the “High Income: OECD” income group and countries that are in the “High Income: nonOECD” group.

To do so first we will subset our data into each of our 5 income groups (low income, lower middle income, upper middle income, high income:OECD, and high income: nonOECD).

```
# Makes a seperate subset for each income group
Merge1.LowIncome <- subset(Merge1, Merge1$Income.Group == "Low income")
Merge1.LowerMiddleIncome <- subset(Merge1, Merge1$Income.Group == "Lower middle income")
Merge1.UpperMiddleIncome <- subset(Merge1, Merge1$Income.Group == "Upper middle income")
Merge1.HighIncomeNonOECD <- subset(Merge1, Merge1$Income.Group == "High income:
nonOECD")
Merge1.HighIncomeOECD <- subset(Merge1, Merge1$Income.Group == "High income: OECD")
```

This is why we removed South Sudan from our data, it does not have listed any income type.

Then, we need to find the average. This will be a 2 step process. First, we need to convert our rankings datatype from factor to numeric. Then we will take the average ranking of each of the 2 specified income types.

```
# Makes the 2 subsets Ranking column numeric, then finds the mean for the income group
Merge1.HighIncomeNonOECD$Ranking <- as.numeric(Merge1.HighIncomeNonOECD$Ranking)
mean(Merge1.HighIncomeNonOECD$Ranking)
```

```
## [1] 93.73913
```

```
Merge1.HighIncomeOECD$Ranking <- as.numeric(Merge1.HighIncomeOECD$Ranking)
mean(Merge1.HighIncomeOECD$Ranking)
```

```
## [1] 110.0667
```

The average ranking for countries in the high income:NonOECD income group is 93.73913. The average ranking for countries in the high income:OECD income group is 110.0667.

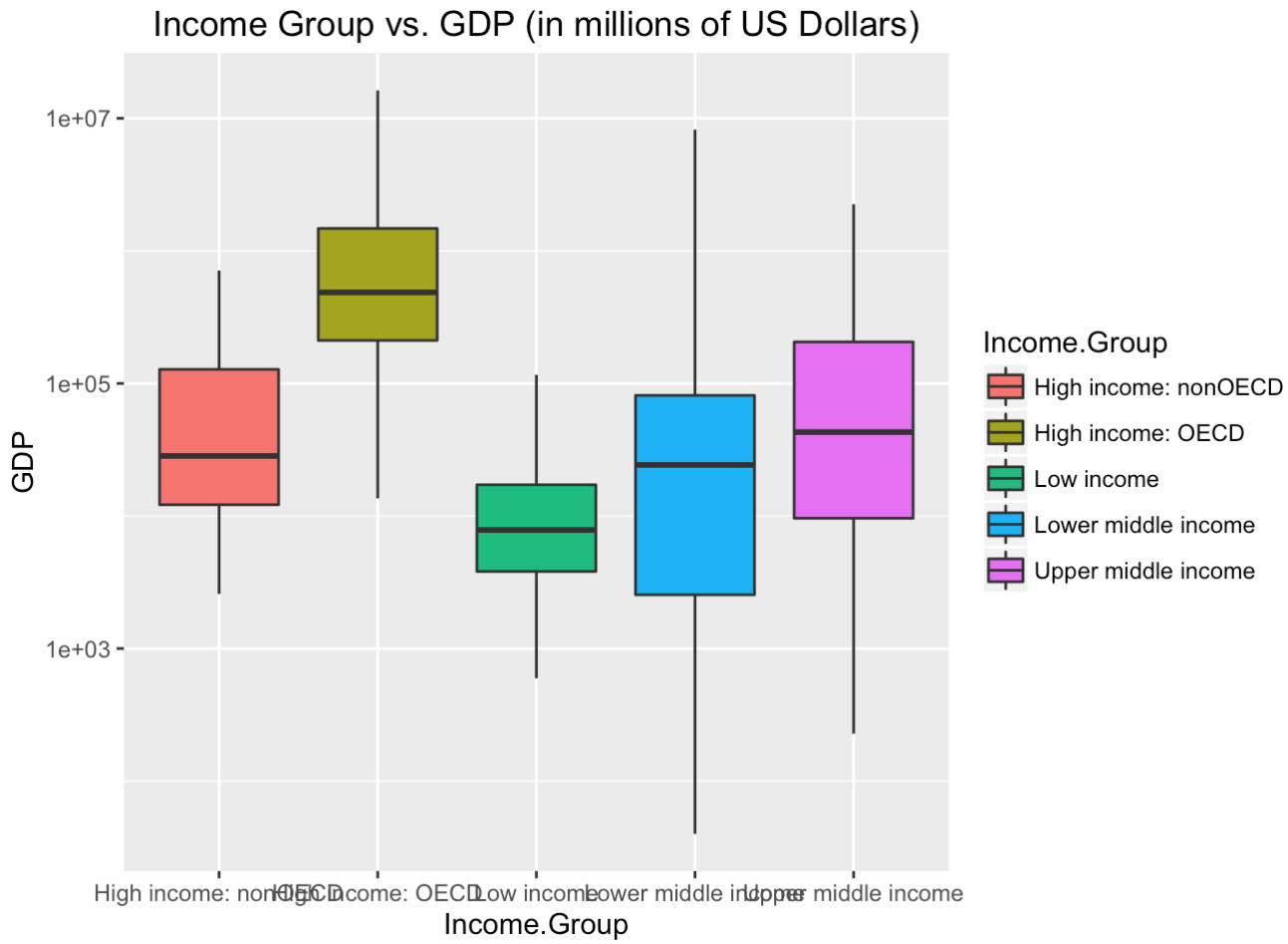
Question 4

We will be plotting the GDP for all of the countries. We will create a graph of boxplots broken up by income group and plot income group versus GDP, which will be in terms of millions of US Dollars.

We need to log transform the data because of the outliers in the data.

```
# Create boxplot where of Income Group vs log(GDP)
Plot4Box <- ggplot(Merge1, aes(x=Income.Group, y=GDP, fill = Income.Group)) + geom_boxpl
ot() +scale_y_log10() + ggtitle('Income Group vs. GDP (in millions of US Dollars)')

# If this were local
Plot4Box
```




```
# source("https://github.com/pefthimion/Case-Study-1/blob/master/Analysis/Boxplot_Income_Group_vs_GDP.pdf")

# Boxplot is also in Analysis directory
```

The high income:OECD group has both the country with the highest GDP as well as being the income group with the highest average GDP. Then, the income group with the second highest average is the upper middle class. It, suprisingly, has a higher maximum value and a higher mean than high income:nonOECD countries. However, the upper middle class countries have much higher variation.

Lower middle income has the most variation as well as the highest range. The GDP of a lower middle class country could be higher or lower than most of the countries in any other income group. Part of this could be due to the fact that there are more countries considered lower middle income than any of the other income groups.

Low income countries have the lowest average GDP and also have the smallest amount of variation. One notable is that there are countries that are considered lower middle income that have a GDP well below that of the average low income country.

Question 5

To calculate the 5 seperate quantile groups we find the length of our rankings and then divide it into 5 equal quantiles.

```
# Find range of each quantile
length(Merge1$Ranking)
```

```
## [1] 189
```

```
189*.2
```

```
## [1] 37.8
```

```
189*.4
```

```
## [1] 75.6
```

```
189*.6
```

```
## [1] 113.4
```

```
189*.8
```

```
## [1] 151.2
```

This means that our quantiles for the rankings will look like the following:

Q1: 1-37 Q2: 38-75 Q3: 76-113 Q4: 114-151 Q5: 152 - 189

First, we will make sure that our rankings data is the correct file type (numeric), then we will order the data by ranking.

```
# Make the lower middle income subset's Ranking numeric
Merge1.LowerMiddleIncome
```

```
##     CountryCode Ranking                  Table.Name     GDP
## 4           AGO      60                      Angola  114147
## 8           ARM     133                     Armenia    9951
## 24          BLZ     169                      Belize    1493
## 26          BOL      96                     Bolivia   27035
## 30          BTN     167                      Bhutan    1780
## 37          CHN       2                       China 8227103
## 38          CIV      99             C\xf4te d'Ivoire   24680
## 39          CMR      98                    Cameroon   25322
## 40          COG     121                 Congo, Rep.   13678
## 43          CPV     166                  Cape Verde    1827
## 55          ECU      64                     Ecuador   84040
## 56          EGY      38            Egypt, Arab Rep.  262832
## 65          FSM     185       Micronesia, Fed. Sts.     326
## 68          GEO     114                     Georgia   15747
## 77          GTM      77                   Guatemala   50234
## 79          GUY     160                      Guyana    2851
## 81          HND     108                    Honduras   18434
## 85          IDN      16                   Indonesia  878043
## 87          IND      10                       India 1841710
## 90          IRQ      47                        Iraq  210280
## 95          JOR      92                      Jordan   31015
## 101         KIR     189                    Kiribati     175
## 104         KSV     146                      Kosovo    6445
## 112         LKA      70                   Sri Lanka   59423
## 113         LSO     163                     Lesotho    2448
## 118         MAR      62                     Morocco   95982
## 120         MDA     141                     Moldova    7253
## 122         MDV     164                    Maldives    2222
## 124         MHL     188            Marshall Islands     182
## 130         MNG     130                    Mongolia   10271
## 140         NGA      39                     Nigeria  262597
## 141         NIC     126                   Nicaragua   10507
## 147         PAK      44                    Pakistan  225143
## 150         PHL      41                 Philippines  250182
## 152         PNG     115            Papua New Guinea   15654
## 157         PRY      97                    Paraguay   25502
## 164         SDN      73                       Sudan   58769
## 165         SEN     119                     Senegal   14046
## 169         SLV     100                  El Salvador   23864
## 174         STP     186 S\xe3o Tom\xe9 and Principe     263
## 179         SWZ     158                   Swaziland    3744
## 181         SYR      65        Syrian Arab Republic   73672
## 185         THA      31                    Thailand  365966
## 187         TKM      91                Turkmenistan   35164
## 188         TMP     170                 Timor-Leste    1293
## 189         TON     184                       Tonga     472
## 191         TUN      79                     Tunisia   45662
## 193         TUV     190                      Tuvalu      40
## 196         UKR      53                     Ukraine  176309
## 199         UZB      75                   Uzbekistan   51113
## 203         VNM      57                     Vietnam  155820
## 204         VUT     177                     Vanuatu     787
```

```
## 206         WSM     181                  Samoa     684
## 207         YEM      90            Yemen, Rep.   35646
##            Income.Group
## 4   Lower middle income
## 8   Lower middle income
## 24  Lower middle income
## 26  Lower middle income
## 30  Lower middle income
## 37  Lower middle income
## 38  Lower middle income
## 39  Lower middle income
## 40  Lower middle income
## 43  Lower middle income
## 55  Lower middle income
## 56  Lower middle income
## 65  Lower middle income
## 68  Lower middle income
## 77  Lower middle income
## 79  Lower middle income
## 81  Lower middle income
## 85  Lower middle income
## 87  Lower middle income
## 90  Lower middle income
## 95  Lower middle income
## 101 Lower middle income
## 104 Lower middle income
## 112 Lower middle income
## 113 Lower middle income
## 118 Lower middle income
## 120 Lower middle income
## 122 Lower middle income
## 124 Lower middle income
## 130 Lower middle income
## 140 Lower middle income
## 141 Lower middle income
## 147 Lower middle income
## 150 Lower middle income
## 152 Lower middle income
## 157 Lower middle income
## 164 Lower middle income
## 165 Lower middle income
## 169 Lower middle income
## 174 Lower middle income
## 179 Lower middle income
## 181 Lower middle income
## 185 Lower middle income
## 187 Lower middle income
## 188 Lower middle income
## 189 Lower middle income
## 191 Lower middle income
## 193 Lower middle income
## 196 Lower middle income
## 199 Lower middle income
## 203 Lower middle income
```

```
## 204 Lower middle income
## 206 Lower middle income
## 207 Lower middle income
```

```
Merge1.LowerMiddleIncome$Ranking <- as.numeric(Merge1.LowerMiddleIncome$Ranking)
str(Merge1.LowerMiddleIncome)
```

```
## 'data.frame':    54 obs. of  5 variables:
##  $ CountryCode : Factor w/ 229 levels "","ABW","ADO",..: 5 9 25 27 31 38 39 40 41 44
 ...
##  $ Ranking     : num  149 41 80 188 78 104 191 190 28 77 ...
##  $ Table.Name  : Factor w/ 230 levels "","  East Asia & Pacific",..: 16 19 30 34 33 5
2 42 44 56 46 ...
##  $ GDP         : num  114147 9951 1493 27035 1780 ...
##  $ Income.Group: Factor w/ 6 levels "","High income: nonOECD",..: 5 5 5 5 5 5 5 5 5 5
 ...
```

```
# Order lower middle income by Ranking
Merge1.LowerMiddleIncome.OrderRanking <- Merge1.LowerMiddleIncome[order(Merge1.LowerMidd
leIncome$Ranking),]
```

Next we will subset the data to find the countries with a top 38 ranking and finally, show all the countries with a top 38 ranking that are considered lower middle income.

```
# Find top 38 ranked countries
Merge1.LowerMiddleIncome.OrderRankingTop38 <- subset(Merge1.LowerMiddleIncome.OrderRanki
ng, Merge1.LowerMiddleIncome.OrderRanking$Ranking <= 38)

# Find how many countries in top 38 are lower middle income
Merge1.LowerMiddleIncome.OrderRankingTop38
```

```
##     CountryCode Ranking       Table.Name     GDP        Income.Group
## 87          IND       4            India 1841710 Lower middle income
## 169         SLV       5      El Salvador   23864 Lower middle income
## 81          HND      13         Honduras   18434 Lower middle income
## 68          GEO      20          Georgia   15747 Lower middle income
## 152         PNG      21 Papua New Guinea   15654 Lower middle income
## 165         SEN      25          Senegal   14046 Lower middle income
## 40          COG      28      Congo, Rep.   13678 Lower middle income
## 141         NIC      33        Nicaragua   10507 Lower middle income
## 130         MNG      38         Mongolia   10271 Lower middle income
```

```
dim(Merge1.LowerMiddleIncome.OrderRankingTop38)
```

```
## [1] 9 5
```

As shown above there are 9 countries that are considered lower middle income, but have a GDP that is one of the top 38 in the world. This would make them inside the top quantile (except one that would be at the top of quantile 2). This reinforces the information from box plot shown above in question 4 where we saw that there are countries

that are lower middle class (the blue shaded box)

The countries that are in the lower middle income group that have a top 38 ranking are India, El Salvador, Honduras, Georgia, Papau New Guinea, Senegal, Republic of Congo, Nicaragua, and Mongolia.

Now we will check our answer by creating a table.

First, we will create a set quantiles based on the quantiles we created above.

```
# Create a new column that lists each country's quantile group
Quantiles <-NULL
Quantiles[1:37] <- 1
Quantiles[38:75] <- 2
Quantiles[76:113] <- 3
Quantiles[114:151] <- 4
Quantiles[152:189] <- 5
Quantiles
```

```
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [36] 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [71] 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
## [106] 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
## [141] 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
## [176] 5 5 5 5 5 5 5 5 5 5 5 5 5 5
```

Now we will put them into a table where it is income groups vs. quantiles.

```
# Create a table of income group vs quantile
table(Merge1$Income.Group, Quantiles)
```

```
##                         Quantiles
##                           1  2  3  4  5
##                           0  0  0  0  0
##   High income: nonOECD    7  5  4  6  1
##   High income: OECD       5  8  8  5  4
##   Low income              6  8  6  8  9
##   Lower middle income     9  8 12  9 16
##   Upper middle income    10  9  8 10  8
```

As the table above shows, there are 9 countries that have lower middle incomes that are in the top quantile.

Conclusion

From the given datasets we were given information about country's GDP (in terms of millions of US dollars) and the countries GDP ranking for the year 2012. We concluded that out of all of the countries, 189 of them matched. In terms of ascending order of GDP (with the USA being at the bottom), St. Kitt's and Nevis (KNA) had the 13th lowest GDP with a GDP of $767,000,000. By analyzing the countries income groups we found, on average, high income non:OECD countries have better GDP ranking than high income OECD countries. We also concluded that there are 9 countries that are in the lower middle income group that are among the 38 nations with the highest GDP.