5730 Project

Ryan Ulring

4/24/2021

Question

The question I would like to ask for this project, is how does the distribution of GDP per capita (US dollars, inflation-adjusted) vary by year for small, medium, and large countries?

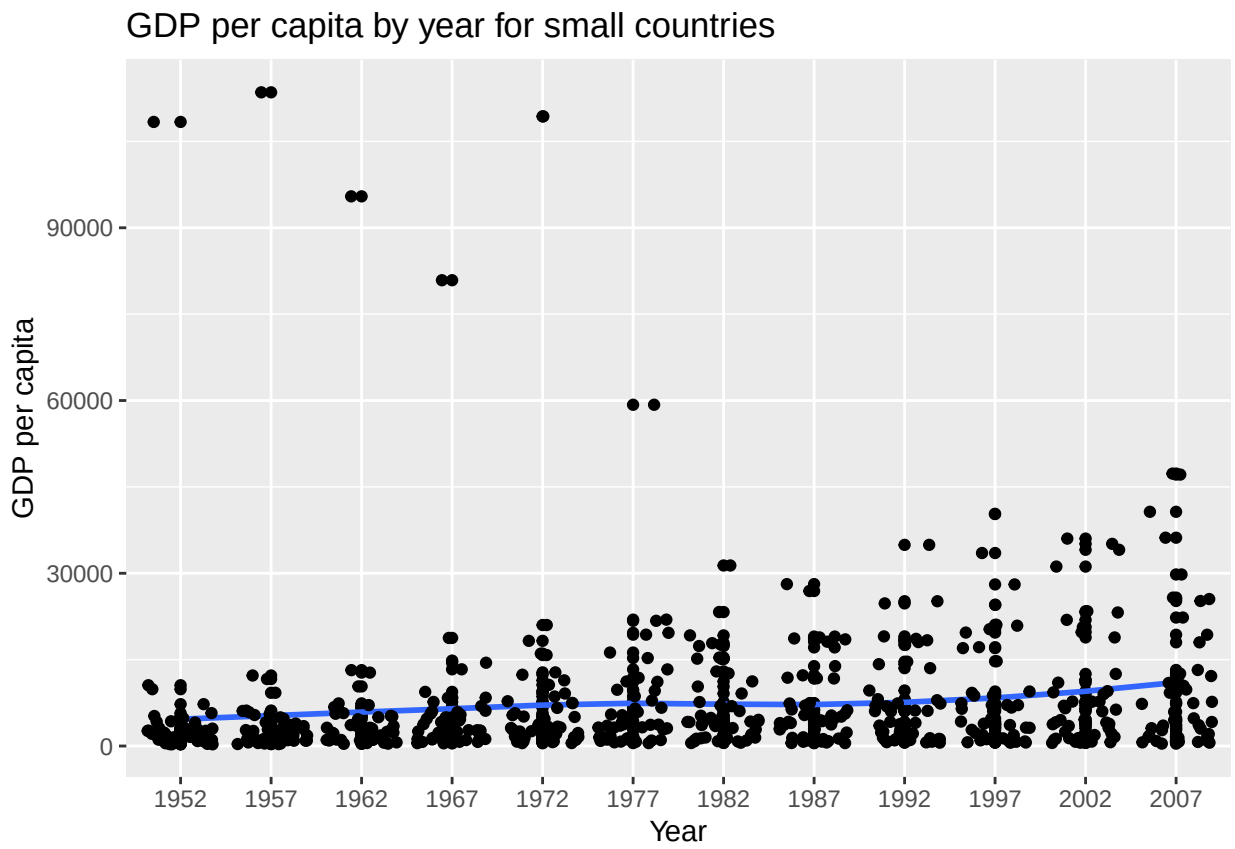
Analysis

First I will need to separate the datasets into small, medium, and large countries. I will do this by separating the countries according to their average population during the time period of this data. The "small" countries will be the countries with the lower third average population, the "medium" countries will be the countries with the middle third average population, and the "large" countries will be the countries with the upper third average population.

```
library(dplyr)
library(ggplot2)
library(gapminder)
data <- gapminder

# find the avg populations by country
avg.pop.by.country <- data %>% group_by(country) %>% summarise(avg.pop = mean(pop))

# find the small, medium, and large countries
small.countries <- arrange(avg.pop.by.country, avg.pop)[1:round(142 / 3),] %>% select(country)
medium.countries <- arrange(avg.pop.by.country, avg.pop)[(round(142 / 3) + 1):round(142 / 3 * 2),] %>% s
large.countries <- arrange(avg.pop.by.country, avg.pop)[(round(142 / 3 * 2) + 1):142,] %>% select(counti

# perform a right join to get the data for the small, medium, and large countries
small.data <- data %>% right_join(small.countries)
medium.data <- data %>% right_join(medium.countries)
large.data <- data %>% right_join(large.countries)
```

Next, I will analyze the GDP per capita by year for small countries by plotting the data.

```
# plot the gdp per capita by a factor of the year, show the trend with
# geom_smooth(), and get rid of some overplotting with geom_jitter()
ggplot(data = small.data, aes(x = factor(year), y = gdpPercap)) + geom_point() + geom_smooth(aes(group =
```



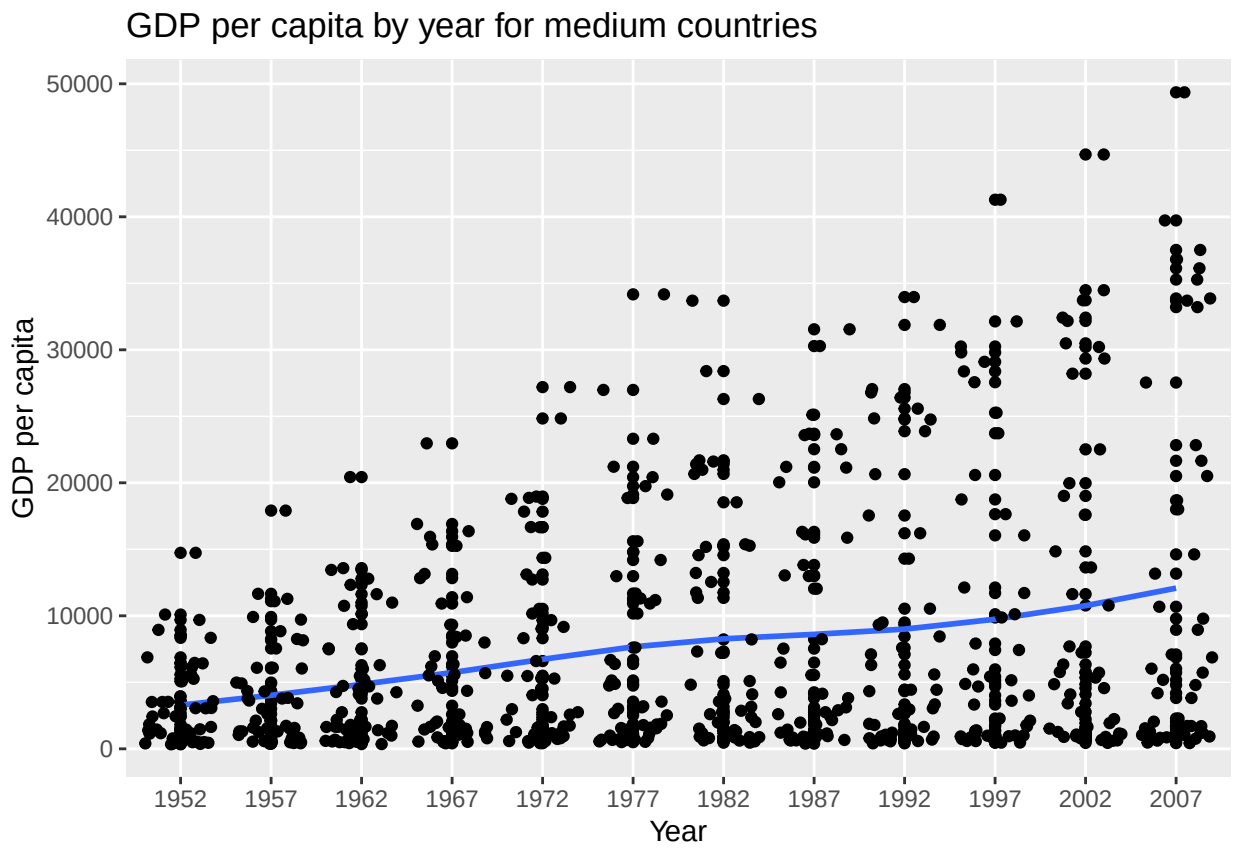
Based on this plot, it appears that the GDP per capita is increasing as the year increases. In addition to this, the variance of GDP per capita seems to increase as the year increases as well. Another thing to note about this plot is that there seems to be some large outliers in the data, which could mean that my conclusions aren't valid. From the plot, it looks like the largest outliers are above 50000, so I will take a look at these outliers to see if I should be concerned about my conclusions.

```
# look at the data for small countries with a GDP per capita greater than 50000
small.data[small.data$gdpPercap > 50000,]
```

```
## # A tibble: 6 x 6
##   country continent  year lifeExp     pop gdpPercap
##   <fct>   <fct>     <int>   <dbl>   <int>     <dbl>
## 1 Kuwait  Asia       1952    55.6  160000   108382.
## 2 Kuwait  Asia       1957    58.0  212846   113523.
## 3 Kuwait  Asia       1962    60.5  358266    95458.
## 4 Kuwait  Asia       1967    64.6  575003    80895.
## 5 Kuwait  Asia       1972    67.7  841934   109348.
## 6 Kuwait  Asia       1977    69.3 1140357    59265.
```

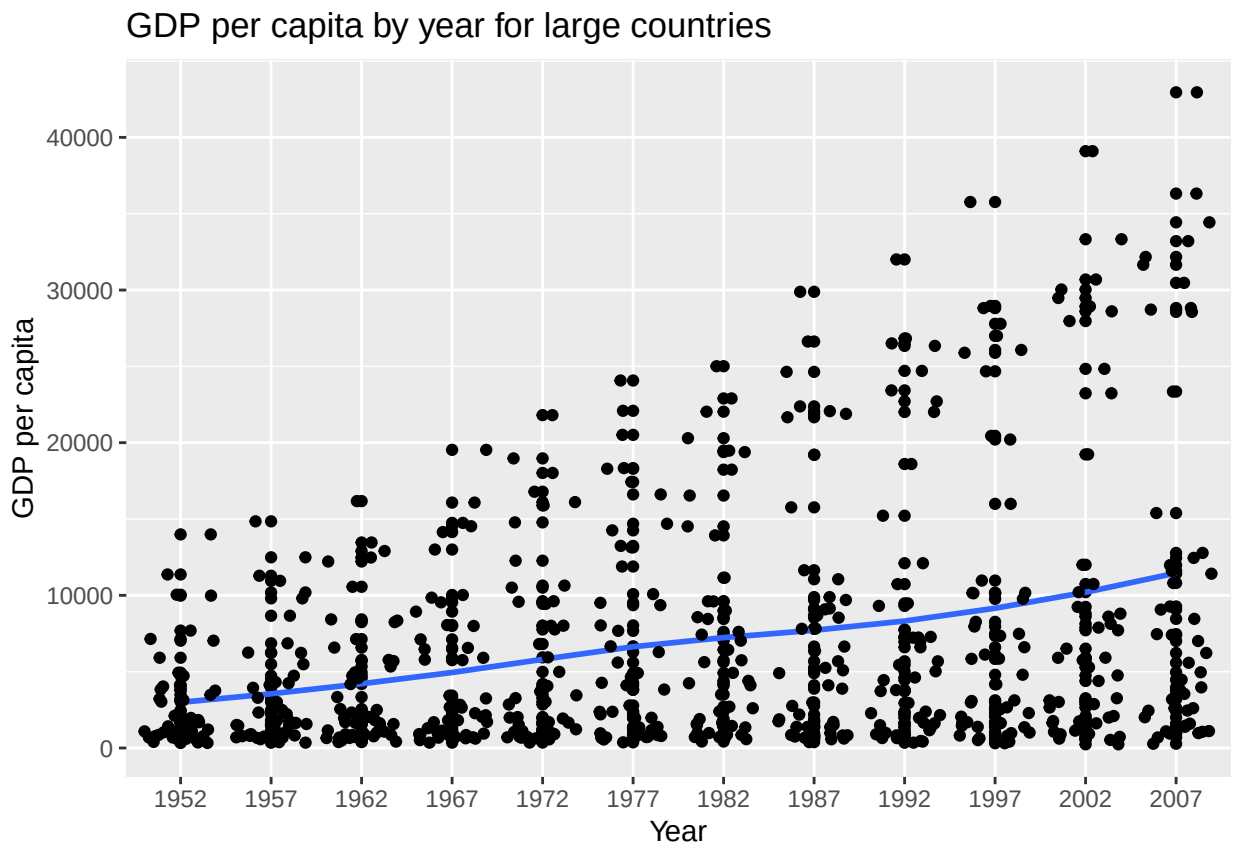
Looking at the data for small countries with a GDP greater than 50000, it appears that all six outliers I found were from one country, Kuwait, so I am not concerned about the conclusions I made for this data. Next, I will analyze the GDP per capita by year for medium countries by plotting the data.

```
# plot the gdp per capita by a factor of the year, show the trend with
# geom_smooth(), and get rid of some overplotting with geom_jitter()
ggplot(data = medium.data, aes(x = factor(year), y = gdpPercap)) + geom_point() + geom_smooth(aes(group
```



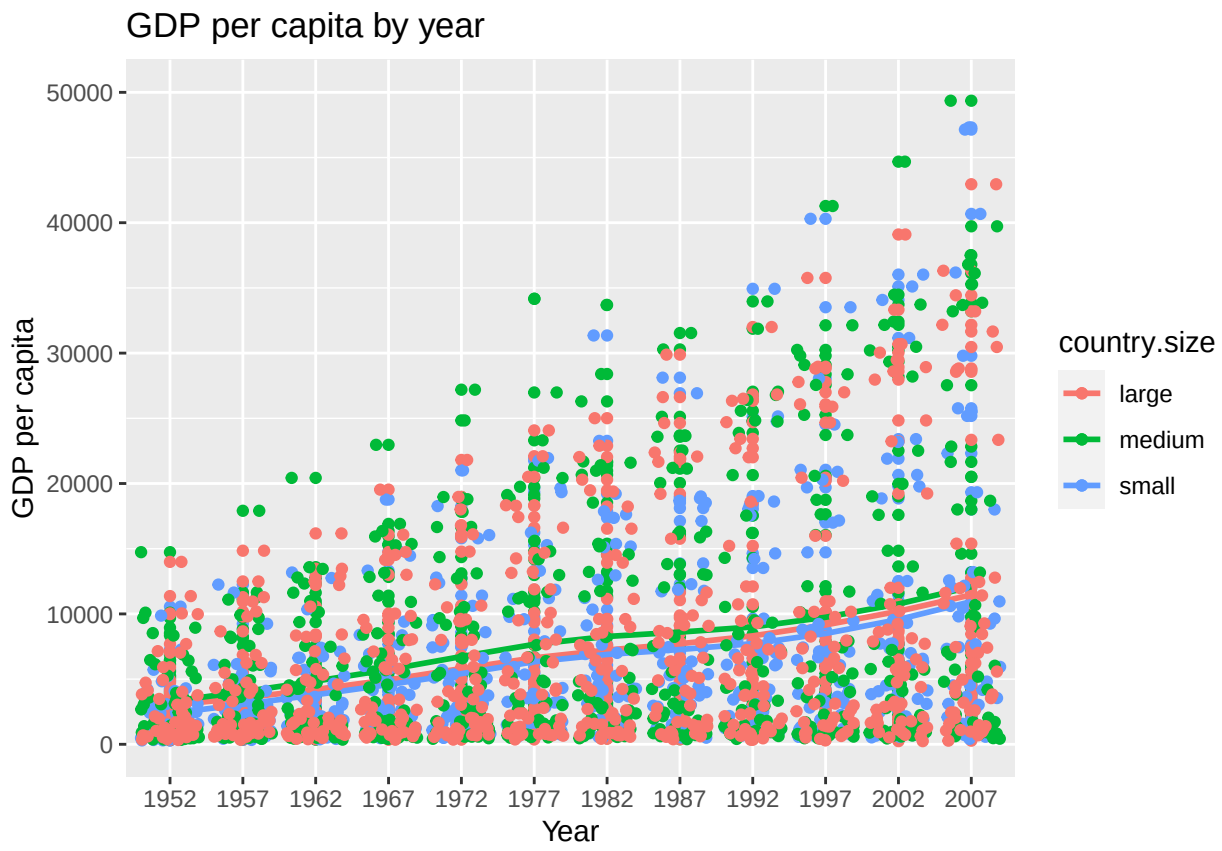
Similarly to the small countries plot, this plot also shows that the GDP per capita appears to be increasing as the year increases. Also, the variance of the GDP per capita appears to increase as the year increases as well. Unlike the small countries plot, there are no large outliears that I find concerning, so I will continue on to the next part of my analysis by plotting the GDP per capita by year for medium countries.

```
# plot the gdp per capita by a factor of the year, show the trend with
# geom_smooth(), and get rid of some overplotting with geom_jitter()
ggplot(data = large.data, aes(x = factor(year), y = gdpPercap)) + geom_point() + geom_smooth(aes(group =
```



Like the other plots, this plot shows that the GDP per capita appears to be increasing as the year increases. The variance of GDP per capita appears to increase as the year increases as well. There are no concerning outliers here, so next I will compare the GDP per capita by year for small, medium, and large countries by plotting the data together, and using color to differentiate between small, medium, and large countries. Also, I will limit the y-values for this plot to be from 0 to 50000, so that we aren't looking at the outliers from the small countries plot.

```
# create a new data set with a variable to account for the country size
# by using the small, medium, and large data sets
small.data$country.size = rep("small", nrow(small.data))
medium.data$country.size = rep("medium", nrow(medium.data))
large.data$country.size = rep("large", nrow(large.data))
newData <- small.data %>% full_join(medium.data) %>% full_join(large.data)

# plot the data (except for points with a GDP per capita of over 50000),
# show the trends for each country size with geom_smooth(), and get rid of some
# overplotting with geom_jitter()
ggplot(data = newData, mapping = aes(x = factor(year), y = gdpPercap, color = country.size)) + geom_poi
```



After looking at the plot, it appears that medium sized countries tend to have the highest GDP per capita, then large countries, and then small countries. Also, the variance of the GDP per capita by year seems to be very similar for each country size. Knowing this information, if someone wanted to move to a country with a high GDP per capita, it may be best to move somewhere that doesn't have a very small or very large population. Also, if someone wanted to move to a country with a small GDP per capita, it might be best to move somewhere with a small population.