John Kamau

Kamau Kamau

10/02/2020

The Gapminder Project

Gapminder is an organzation based in Sweden that seeks to fight devastating ignorance using statistical tools. They have colelcted alot of data on most countries in the world on ndicators that show performance of the countries in welfare, income etc. It promotes the achievement of the sustainable goals and millenium goals from the united Nations through better understanding of world data. They have a number of interesting tools that I would recommend that you check out on their website and play around with the data.

```
library(knitr)
library(tidyverse)
library(gapminder)
library(rworldmap) ## plotting the data on World Map
library(countrycode) ## Converting the country name to Country code
library(dplyr) ## For manipulating, transforming, filtering, summarizing the data
library(Hmisc)
library(printr)
library(RColorBrewer)
df <- gapminder
```

Data Description

The data used comes from the gapminder organization mentioned above. It has a total of six columns and 1704 rows. The columns are Country, Continent, year, Life expectancy, population and GDP (gross domestic product) per capita which basically means the GPD / population. Its a better measure of economic performance than just GDP because it takes into account the country size based on its population and also indicates income distribution in the country. The years range from `min(gapminder$year)` to `max(gapminder$year)`

```
sum(complete.cases(gapminder)) ## No missing values found
```

```
## [1] 1704
```

```
#describe(gapminder_unfiltered) %>% t()
head(gapminder) %>% kable()
```

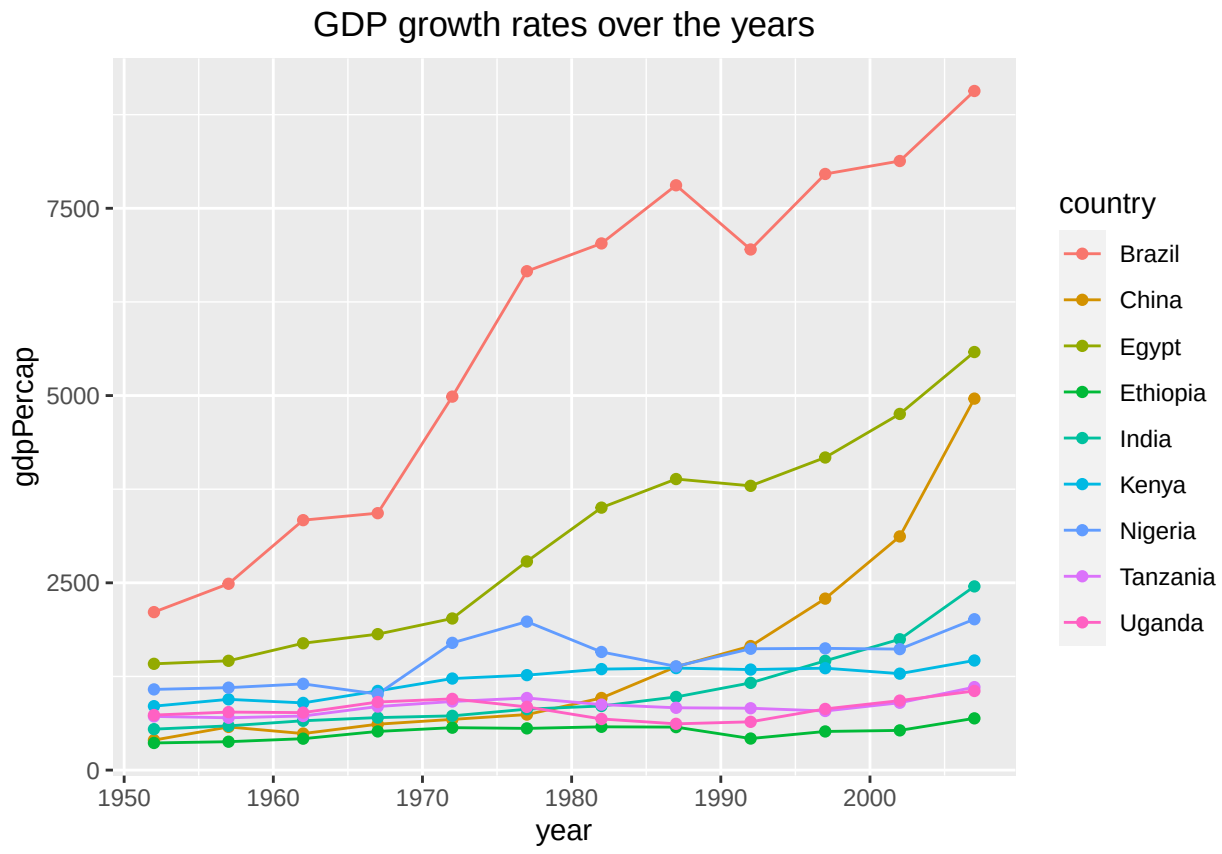
country	continent	year	lifeExp	pop	gdpPercap
Afghanistan	Asia	1952	28.801	8425333	779.4453
Afghanistan	Asia	1957	30.332	9240934	820.8530

country	continent	year	lifeExp	pop	gdpPercap
Afghanistan	Asia	1962	31.997	10267083	853.1007
Afghanistan	Asia	1967	34.020	11537966	836.1971
Afghanistan	Asia	1972	36.088	13079460	739.9811
Afghanistan	Asia	1977	38.438	14880372	786.1134

Exploration of the Data

We can look at the general trend in the GDP of some countries within the gapminder dataset to see how the economies have been performing over the years. The general trend is positive,its safe to say that gdp has been growing in most countries the only difference is the growth rates. Evidently, Brazil, China and Egypt are the most dramatic for the current dataset slice with a few countries from Africa,Asia and USA with china showing the fastest growth rate while that of Brazil is somewhat gradual.

```
library(plotly)
countries <- c("Kenya","Uganda","Tanzania","Nigeria","Ethiopia","Egypt","SouthAfrica","USA","Brazil","Bi
#describe(df)
Graph <- df %>% filter(country %in% countries) %>% ggplot(aes(x = year,y =gdpPercap,color = country))+g
Graph
```



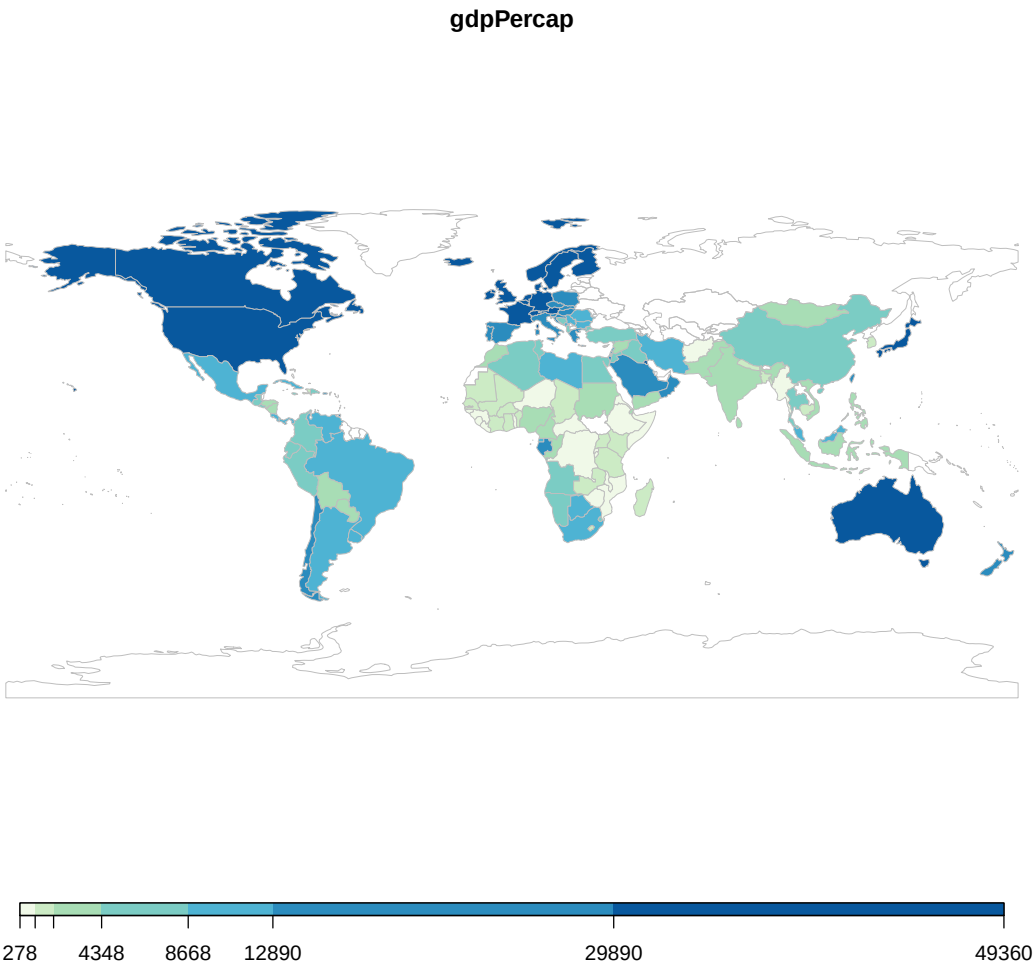
GDP Distribution 2007

Wee can look at the distribution of incomes all over the world in a single year (in this case 2007) using a worldmap. Worldmaps are a great tool to visualize global data that has the countries in it. As expected, european countries seem to have the largest GDP with Africa having the least.

```
df$countrycode <- countrycode(df$country, 'country.name', 'iso3c')

sPDF <- joinCountryData2Map(df %>% filter(year == 2007)
                            ,joinCode = "ISO3"
                            ,nameJoinColumn = "countrycode"
                            ,mapResolution = "coarse"
                           , verbose = T)
```

```
## 141 codes from your data successfully matched countries in the map
## 1 codes from your data failed to match with a country code in the map
##      failedCodes
## [1,] "REU"
## 103 codes from the map weren't represented in your data
```

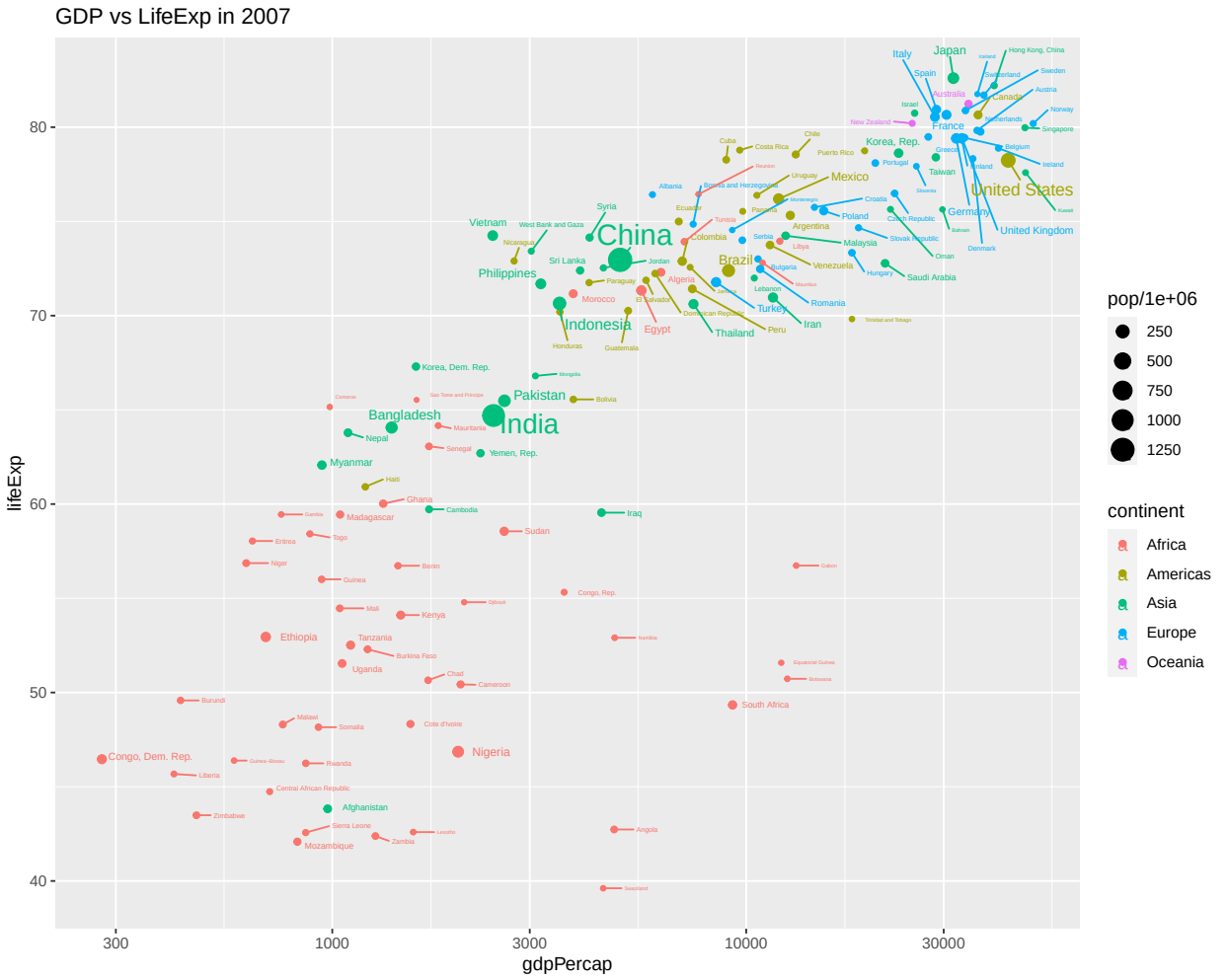
```
 colourPalette <- brewer.pal(7,'GnBu')

mapParams <- mapCountryData(sPDF,
                            nameColumnToPlot="gdpPercap",
                            addLegend=F,
                            colourPalette=colourPalette )

do.call(addMapLegend
        ,c(mapParams
           ,legendLabels="all"
           ,legendWidth=0.5
           ,legendIntervals="data"
           ,legendMar = 2))
```



Scatter Plot

we can try and visualize most of this data using a scatter plot where the x axis represents the GDP and the y axis the life expectancy. It is clear that GDP and lifeExp have a positive relationship as shown below. We have selected a single year (2007) for this analysis, a better method would be to use GGanimate to see what has been happening to the GDP and Life expectancy over the years. It is clear that African countries (in Red) are doing very poorly particularly to life expectancy as some have high GDP but low life expectancy like South Africa.

```
library(ggrepel)
gapminder %>% filter(year == 2007) %>% ggplot(aes(x = gdpPercap,y = lifeExp,size = pop/1000000,color = c
```




```
#+theme(legend.position = "none")
```

Countries with highest GDP 2007

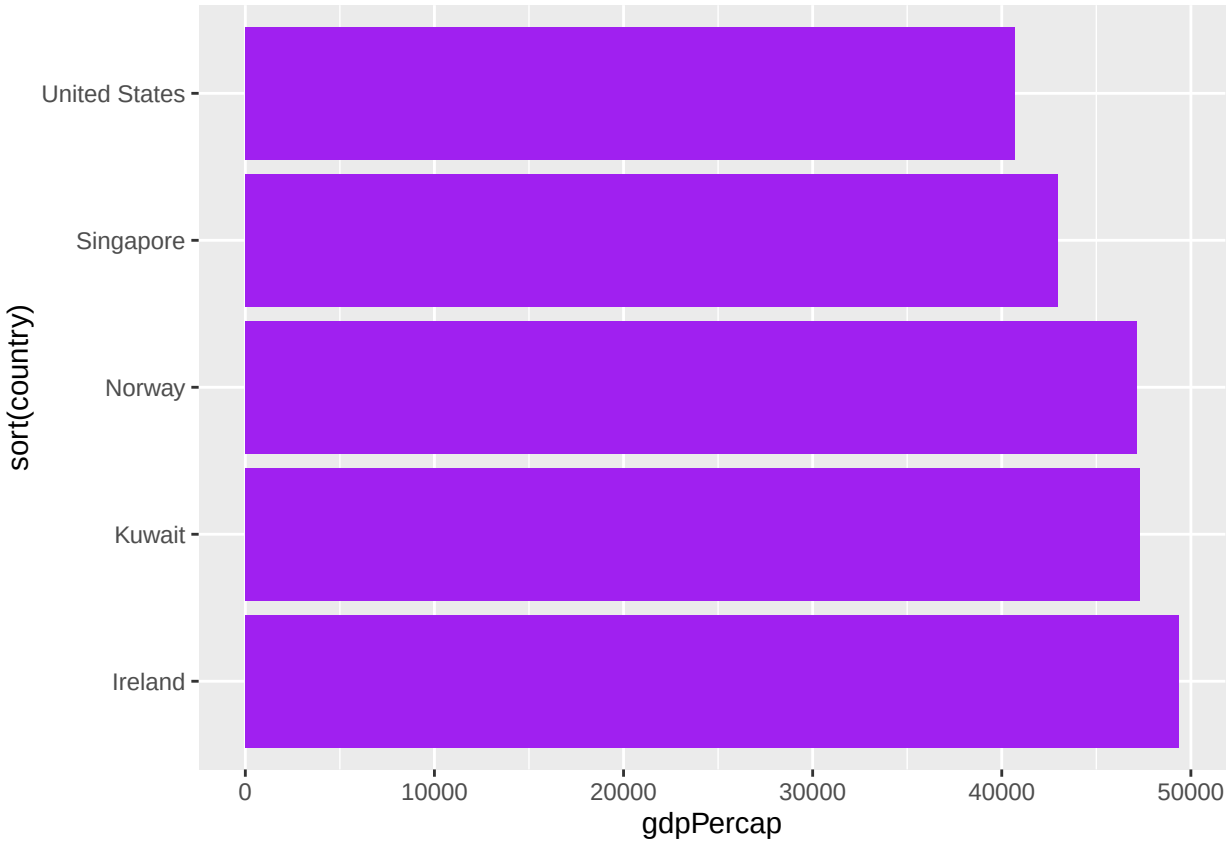
Worldwide

Here are some countries with the highest GDP per capita. Please note that per capita means that the GDP is divided over the countries population so a country like USA that has very high GDP is seen to have lower GDP per capita due to its high population.

```
gap2007 <- df %>% filter(year == 2007)
summary(gap2007$gdpPercap)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
277.5519	1624.842	6124.371	11680.07	18008.84	49357.19

```
count <- gap2007 %>% arrange(-gdpPercap) %>% select(country, gdpPercap) %>% filter(gdpPercap > 40000)
count$country <- factor(count$country)
count%>% ggplot(aes(x = sort(country),y = gdpPercap))+geom_bar(stat = "identity",fill = 'purple')+coord_
```



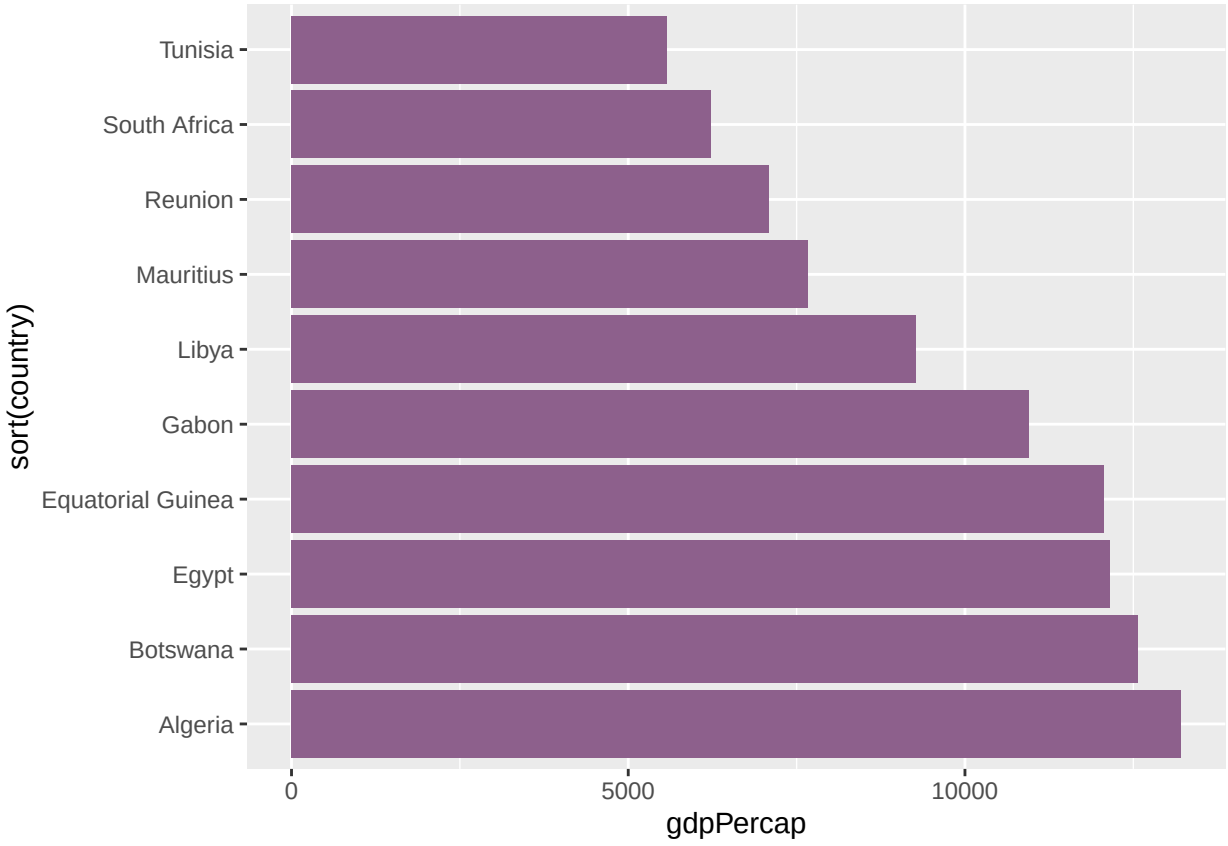
Africa

The same thing goes for African Countries, Algeria is seen to have the highest GDP per cap

```
gap2007<-gap2007 %>% filter(continent =="Africa") %>%  filter(year == 2007)
summary(gap2007$gdpPercap)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
277.5519	862.9515	1452.267	3089.033	3993.502	13206.48

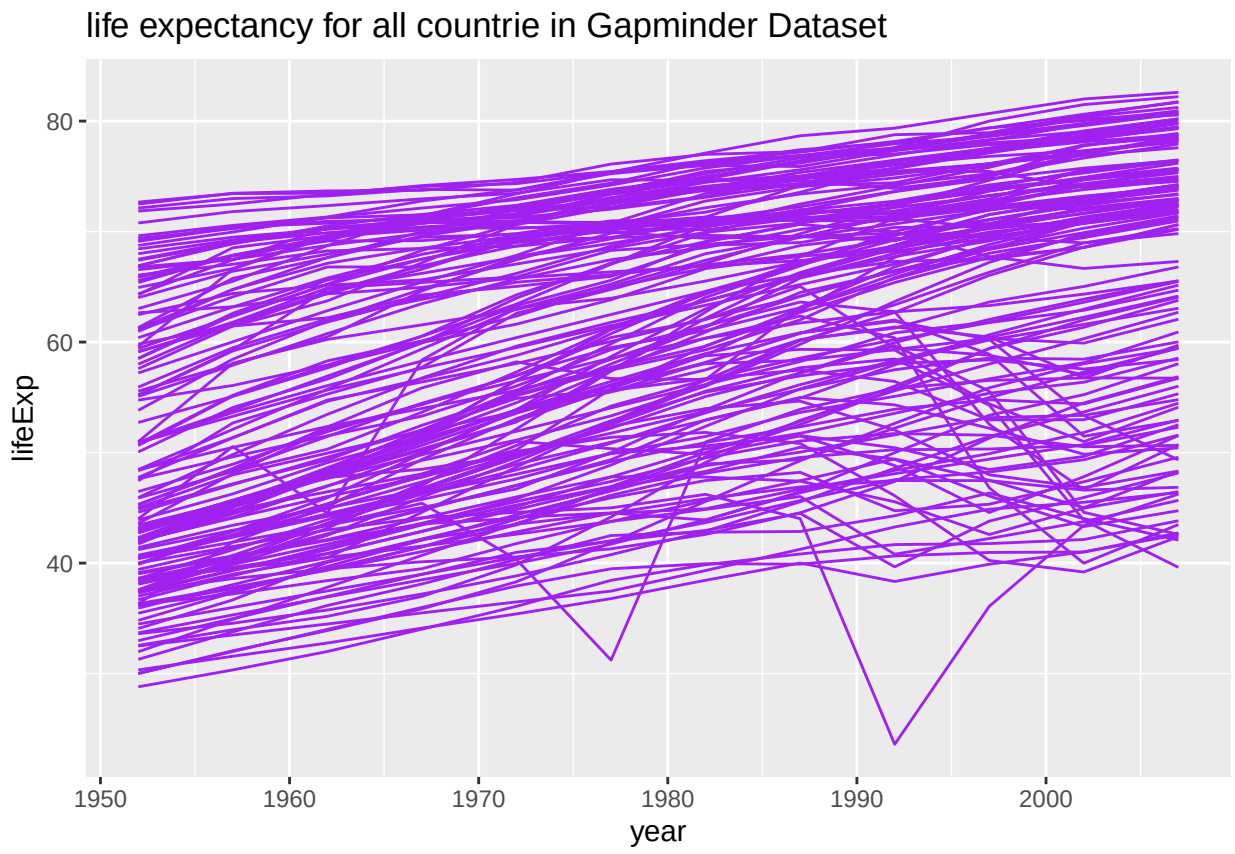
```
count <- gap2007 %>% arrange(-gdpPercap) %>% select(country, gdpPercap) %>% filter(gdpPercap>5000)
count$country <- factor(count$country)
count%>% ggplot(aes(x = sort(country),y = gdpPercap))+geom_bar(stat = "identity",fill = "#8D608C")+coor
```



Does the Life expectancy really grow each year?

although the graph below seems to indicate that for most countries, the Lifeexpectancy grows each year, We cannot just trust our eyes and conclude that most countries have seen a rise in in life expectancy. This is because there are some countries that have declining life expectancy and we can therefore not tell just by looking at the graph below that for sure the average behavior is a rise in life expectancy over all.

```
df %>% ggplot(aes(x = year,y =lifeExp,group = country))+geom_line(color ="purple")+labs(title = "life e
```



Regression

A more robust way to find the behavior of life expectancy over time is to use regression with the following equation:

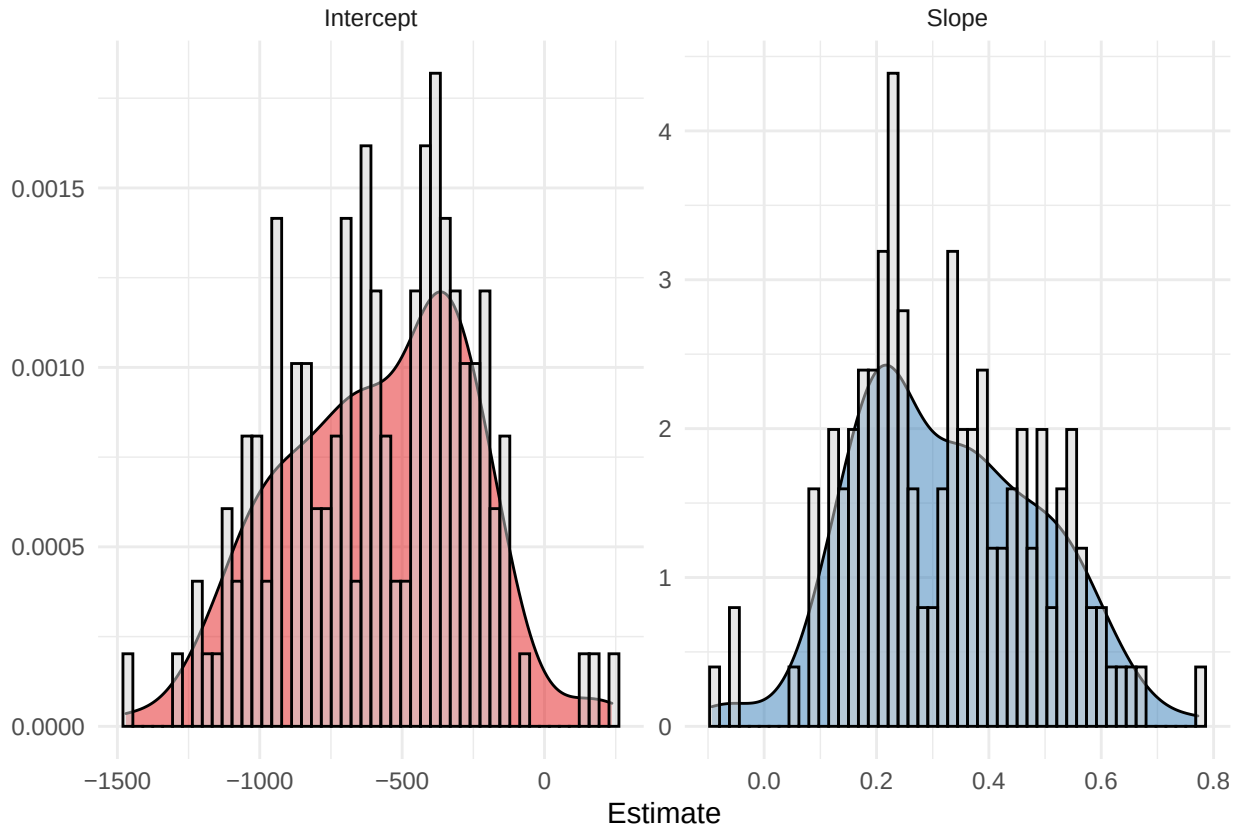
$lifeExp = B1 + B2 * year$

here we are looking at the relationship between changes over time and life expectancy. The analysis for all the countries is a bit involved and requires using Nestng function which splits the data in our case by country, we do a regression for each country and extract the coefficient for the year, we then plot a histogram for all these coefficients and see if they are positive for the most part.

It is clear from the graphs below that whereas the intercep is mostly negative, the slope is positive for a majority of the countries. (the part beyond zero is the largest, actually I would estimate over 99% of the countries have had a positive growth of lifeExp over the years)

```
library(tidyverse)
nested <- df %>% group_by(country) %>% nest()


reg <- function(df){
  m = lm(lifeExp~year,data = df)
  return(m)
}
Models  = nested %>% mutate(model = map(data,reg),
                        coef = map(model,broom::tidy))
```

```
Models.coef  <- Models %>% unnest(coef)
#Models.coef
Models.coef %>%
  mutate(term = fct_recode(term,
                           Intercept = "(Intercept)",
                           Slope = "year")) %>%
  ggplot(aes(estimate, fill = term)) +
  geom_density(show.legend = FALSE, alpha = 0.5) +
  geom_histogram(col = "black", fill = "lightgrey",
                 alpha = 0.5,bins =50,
                 aes(y = ..density..)) +
  facet_wrap(~term, scales = "free") +
  scale_fill_brewer(palette = "Set1") +
  theme_minimal() +
  labs(y = NULL, x = "Estimate")
```



Conclusions and Further Analysis

The Gapminder is an interesting dataset especually because its collected for such a long time and includes so many countries with each indicator showing an important part of the countries overall performance both economically and welfare. We can do alot of analysis: cross sectional where we look at comparisons over a specific year, or time series where we see analysis over the years.

One can consider doing more analysis, in particular using the GGanimate library to plot an animated scatter

plot (or bubble plot as popularized by the gapminder organixation) to see how GDP vs life Exp performs over the years.