World Development Indicators

Kedar Deodhar, Prateek Rajput

18/12/2020

Abstract

This is an exploratory data analysis of the World Development Indicators dataset. [1] The dataset has a variety of data from different domains such as economic policies of a country, poverty, education, infrastructure and technology, healthcare. Our goal was to evaluate performance of countries the best we could, see if there are certain countries that have significantly been able to change their results and try to find out if any identifiable reasons for that exist. As a goal, we tried to see if there are some controllable factors like government expenditure which generally result in human prosperity or better life quality. We tried to find out which specific policy decisions, societal norms or perception of people is the difference between developed countries and developing countries. In order to check the overall importance of an indicator in the quality of living in a country, we analyzed developed nations with a focus on US and Canada and their comparison. To find the differentiating factors between developing and developed countries, we used India as an example of a developing nation. In the end, we did find some indicators that are related to indicators in seemingly other domains. The results could potentially serve as a guide to the actions that can be taken by governments for betterment of quality of life.

Table of content

Introduction

The reason behind the choice of this dataset is the variety of data it provided. The fact that the dataset had covered all the aspects of human life, especially Economy, Health and Education made it worth exploring. To get better insights on our findings, we also added the data we collected from Happiness Index Report [2]. The data from the World Development Indicators dataset helped us in understanding the objective, measurable elements of human development, but we also wanted to see how well how well some subjective, self-identified factors such as self-described happiness, perception of freedom, corruption, generosity depend on the objective measures of well being. The project was a great opportunity for us to dive deep into all the aspects of data science. Data manipulation, Statistical Analyses, Machine Learning and Data Visualization. Statistical Analysis on the data was a crucial part for us since both of us are from Computer science background. We tried to be as thorough as we could in the statistical analysis part by analyzing our Hypothesis Test results deeply, checking for significance or magnitude of the results by running post-hoc tests. Visualization was another challenge for us because our project is data driven which makes representation of data really important. We tried to find best possible diagrams and tools that make it easy to interpret the information. We also familiarized ourselves with the concept of Time Series Analysis. Although we had learnt these things to one extent or another throughout the Big Data program, the project enhanced our knowledge about them and helped us learn how to put our knowledge into practice. Applying these methods to a real world problem, the WDI dataset, gave us a chance to learn and overcome the problems we faced while implementing them while simultaneously getting to know a lot about the problems faced by many countries and their possible solutions. There was some work that we did which ended up being insignificant. This echos either models that failed, tests that did not give expected or significant results, or results that we could not be sure about. We have added all of this in the Appendix part.

Prior Research

We looked for related work done by others on this dataset. There was not much that was done to comprehensively cover the entire dataset, but we found some interesting studies done on specific parts of the dataset.

This study [3] by Krishna Ravikumar is a comparison between India and China’s development indicators on the same database. The focus was on some economic indicators such as GDP, Trade and import export as well as life expectancy and literacy rate. The conclusion was that China is ahead of India in most indicators. India’s lag from China was between 10 to 25 years depending on the indicator. The stated result was that India is that many years behind China in those indicators. This seemed like a too simplistic way to draw a conclusion and it was based on minimal data. However, the approach taken was novel, the visualization was good, and the overall idea seemed interesting.

This study [4] by itssangeeta tries to explore global wealth inequality and growth over the years. It has some great observations about how far countries have come from 1960. It provides insightful information about countries that have been consistently poor or rich, countries that are emerging, and countries that stand out in terms of accumulation of wealth, average income or wealth distribution. The comparison it attempts to do is interesting, but it sometimes fails to recognize missing data at some points, specially in the earlier years, which sometimes makes the outcome of the graph look a bit contextless. Yet, there were some interesting observations and based on that they went deep into interesting specifics.

From these projects, we picked up how to roughly observe outliers or significant deviations from ordinary and then try to find out more about them. These projects don’t go into any further statistical analyses or Machine learning. We wanted to do that but there was something to be gained from these projects about what things can be tried to find interesting trends and how to focus on interesting specifics which we could then analyse further.

Experiment and results

Health

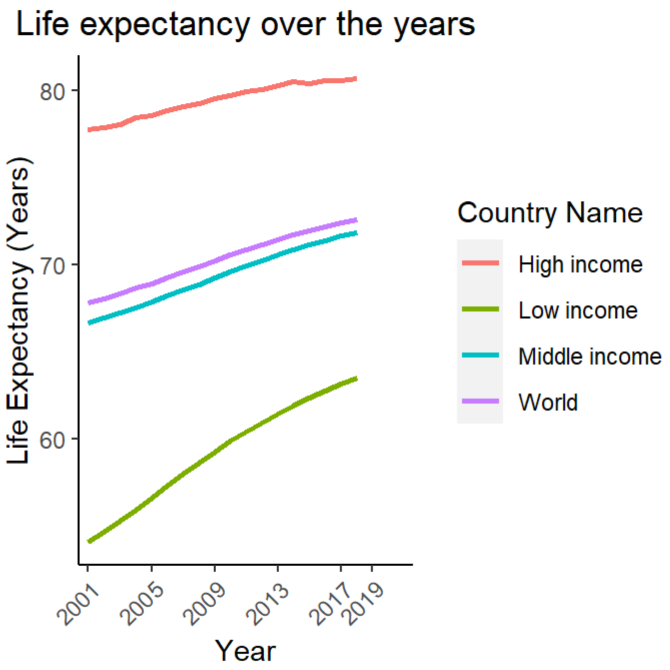
After importing the data, the first domain that we chose to analyse was Health since quality and availability of healthcare in a country is one of the best indicators its overall well-being.

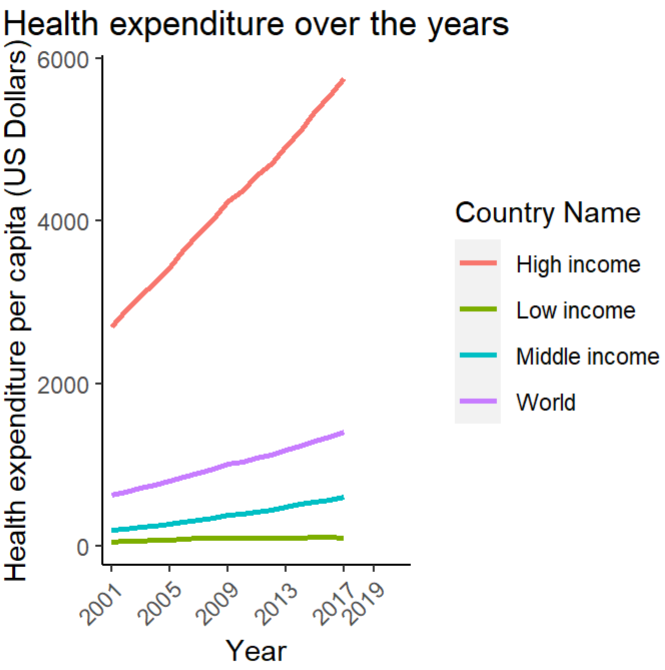
The first thing we checked was the life expectancy of countries in 2018. Then, we checked which countries spent more on healthcare. We measured the spending per capita in USD.

```
## # A tibble: 264 x 1
##    `Country Name`
##    <chr>
##  1 Hong Kong SAR, China
##  2 Japan
##  3 Macao SAR, China
##  4 Switzerland
##  5 Spain
##  6 Italy
##  7 Singapore
##  8 Liechtenstein
##  9 Channel Islands
## 10 Iceland
## # ... with 254 more rows
```

```
## # A tibble: 217 x 1
##    `Country Name`
##    <chr>
##  1 Tuvalu
##  2 United States
##  3 Marshall Islands
##  4 Sierra Leone
##  5 Micronesia, Fed. Sts.
##  6 Switzerland
##  7 Palau
##  8 Afghanistan
##  9 Cuba
## 10 France
## # ... with 207 more rows
```

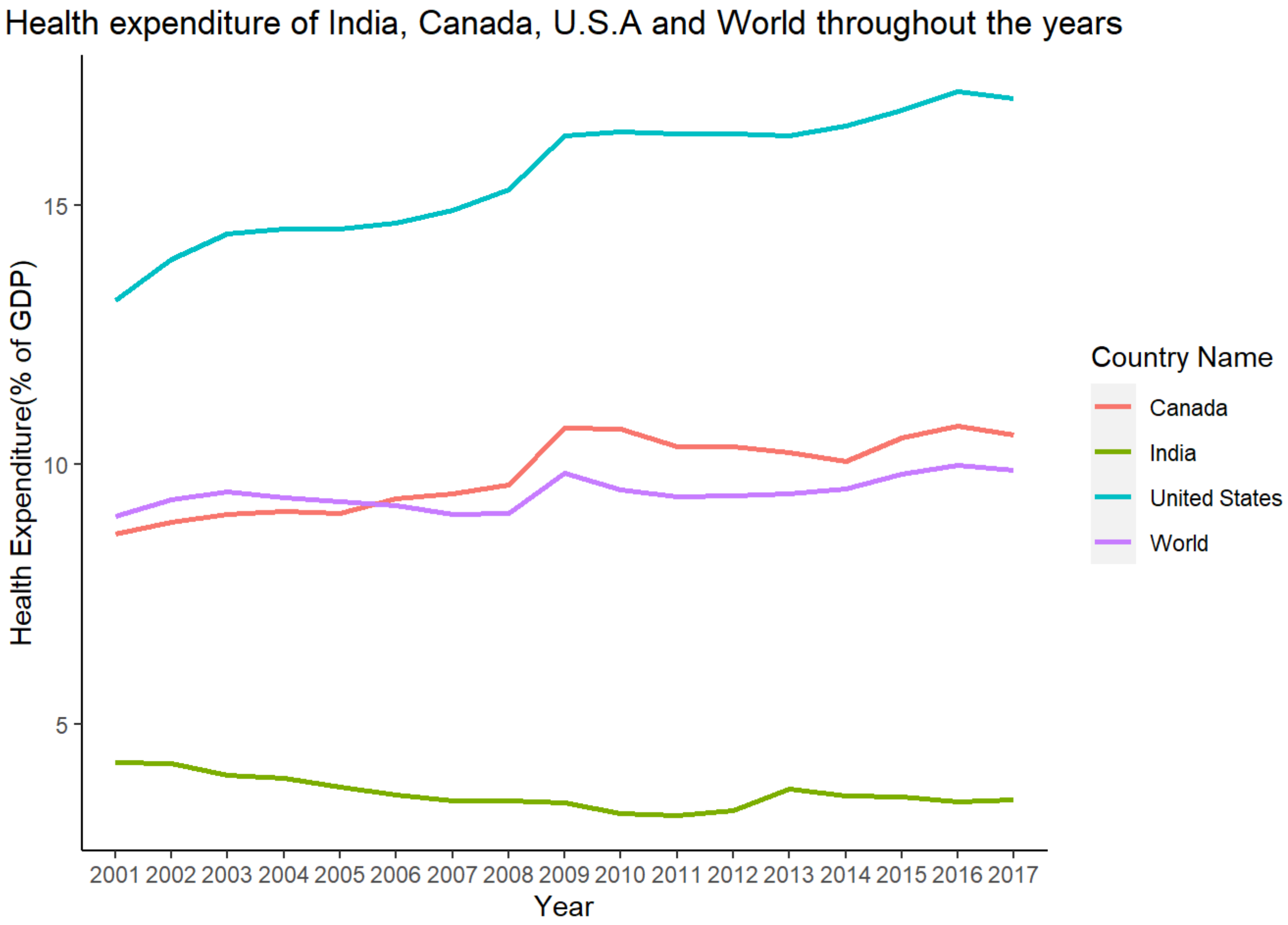
Then, we decided to see how both these things have progressed in different country categories. These categories were based on the economic condition of a country.





We saw that the Life expectancy had gone up for the entire world progressively. Same can be said for spending as well. The average of high-income countries was $2700 in 1964 which went up to $5750 in 2017. For lower income countries, it went from $54 to $104.

Next, we compared the health spending per capita of U.S., Canada and India. For a fair comparison, we compared their expenditure as a percent of their GDP.



We saw that Canada’s expenditure was the closest to the world. United States spent a lot higher than Canada whereas India’s spending seemed significantly lower than the world average.

Does more expenditure give better outcomes?

Looking at these graphs, it looks like countries with higher spending have better Health coverage. We decided to verify this by doing a pearson correlation test between life expectancy and health expenditure.

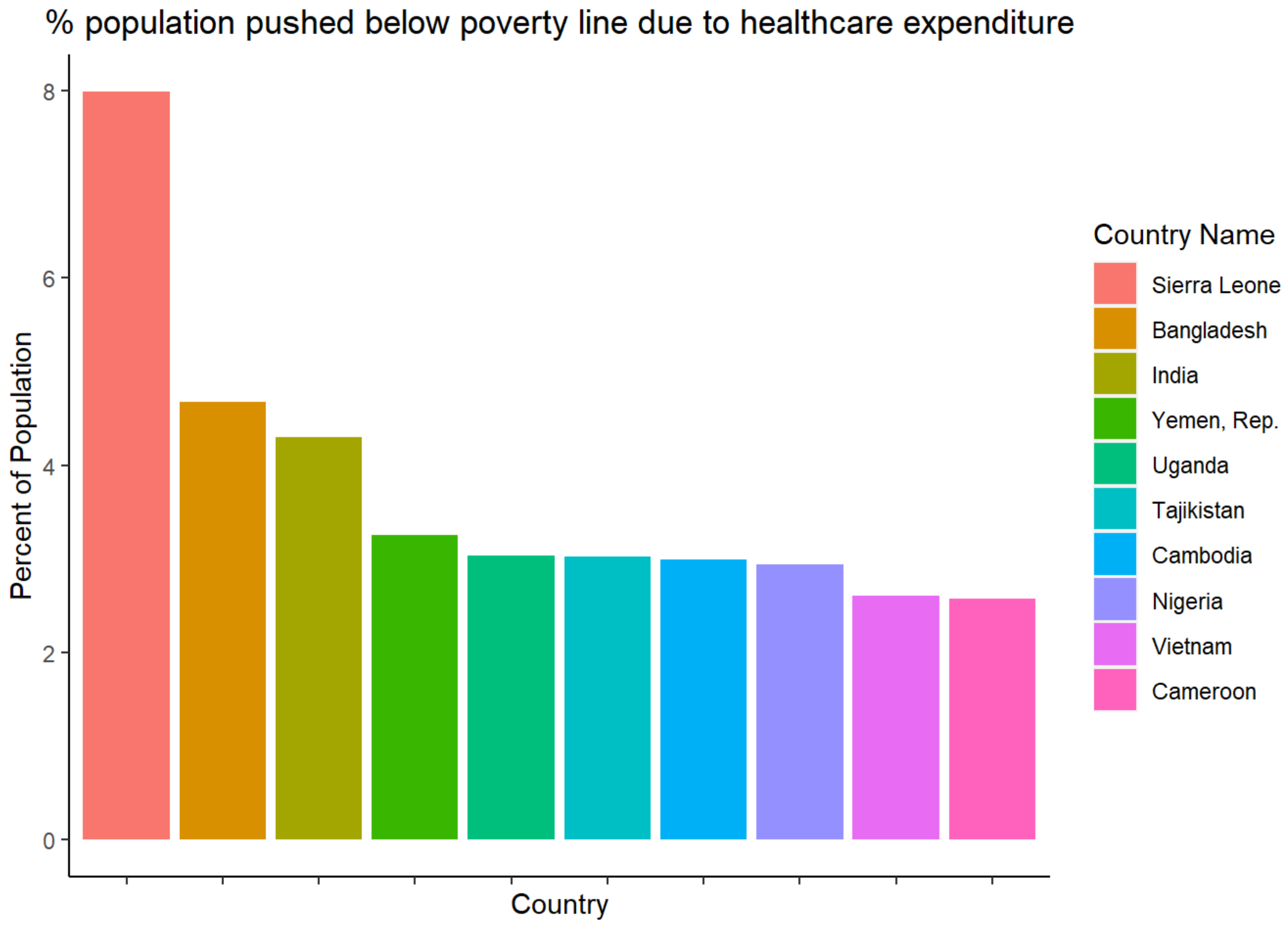
```
##
##  Pearson's product-moment correlation
##
## data:  data_proper_all$SP.DYN.LE00.IN and data_proper_all$SH.XPD.CHEX.PP.CD
## t = 51.394, df = 3999, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6116550 0.6489926
## sample estimates:
##       cor
## 0.6306887
```

The correlation coefficient was 0.63, which makes it a decent correlation.

After this, we checked the correlation between Per capita expenditure and UHC index, which seemed like an even better indicator of a country’s Health coverage and quality. UHC (Universal Health Coverage) index is given by the World Health Organization and it also considers factors such as access to healthcare, quality, child or new-born mortality rate, spread and prevalence of infectious disease etc.

```
##
##  Pearson's product-moment correlation
##
## data:  data_proper_all$SH.XPD.CHEX.PP.CD and data_proper_all$SH.UHC.SRVS.CV.XD
## t = 19.306, df = 440, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6233042 0.7247023
## sample estimates:
##       cor
## 0.6772056
```

The correlation was 0.677 which is an even better correlation. We could firmly say that spending more on healthcare results in better outcomes objectively.



Does India spend significantly less than other lower middle income countries?

India was 3rd in this list. From this and other the data we saw, it looked like India’s poor performance in healthcare was due to low spending on healthcare, which was substantially less than the world average. In fact, it seemed even less than the spending of countries of Lower middle income category. We wanted to statistically verify this hypothesis so we used t-test to see if the difference was actually statistically significant.

```
## # A tibble: 1 x 8
##   .y.              group1 group2              n1    n2 statistic    df        p
## * <chr>            <chr>  <chr>            <int> <int>     <dbl> <dbl>    <dbl>
## 1 SH.XPD.CHEX.GD.~ India  Lower middle inc~    18    18     -5.09  20.3  5.28e-5
```

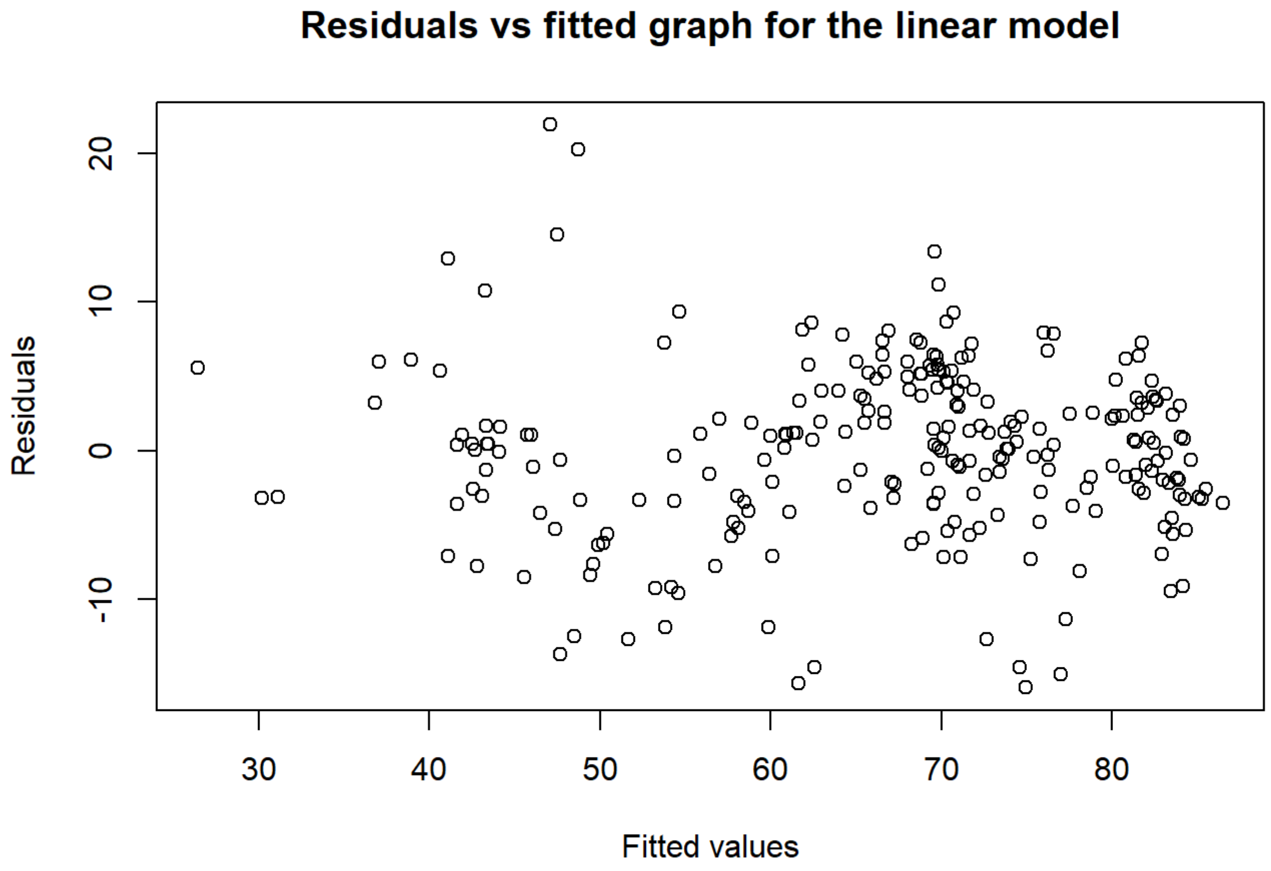
```
## # A tibble: 1 x 7
##   .y.                group1 group2              effsize    n1    n2 magnitude
## * <chr>              <chr>  <chr>                 <dbl> <int> <int> <ord>
## 1 SH.XPD.CHEX.GD.ZS India  Lower middle income   -1.70    18    18 large
```

It turned out that the difference between India’s health expenditure as a percent of GDP and the average of spending of lower middle income countries was statistically significant with India’s spending being low. We used cohens_d method to get the effect size, which was large.

Predictors of Universal Health Coverage Index

Then we wanted to see what factors impact the UHC index the most, so we ran regression tests between UHC coverage and various indicators in the healthcare domain. In the end, the most effective combination of predictors was Government health expenditure and physicians per 1000.

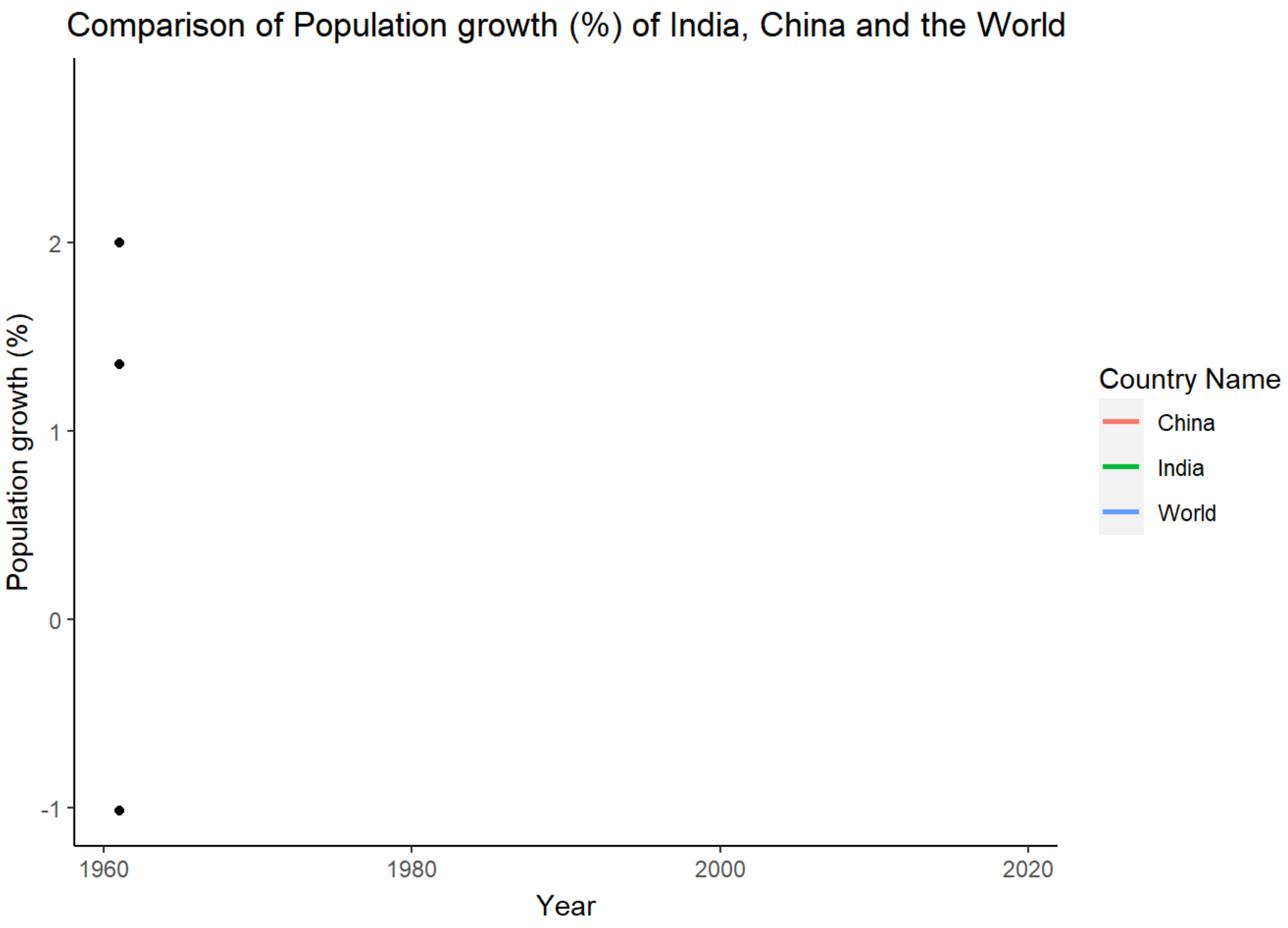
```
## 
## Call:
## lm(formula = SH.UHC.SRVS.CV.XD ~ SH.MED.PHYS.ZS + SP.DYN.LE00.IN, 
##     data = data_proper_all)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -15.940  -3.286   0.399   3.589  21.928 
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    -59.27409    5.01974 -11.808  < 2e-16 ***
## SH.MED.PHYS.ZS   1.21770    0.34221   3.558 0.000446 ***
## SP.DYN.LE00.IN   1.68230    0.07464  22.540  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5.804 on 250 degrees of freedom
##   (16115 observations deleted due to missingness)
## Multiple R-squared:  0.8415, Adjusted R-squared:  0.8403 
## F-statistic: 663.9 on 2 and 250 DF,  p-value: < 2.2e-16
```



Running a Linear model with Universal Health Coverage Index as the dependant variable and Life expectancy and Doctors per 1000 people as predictors, gave us an adjusted R-squared of 0.84 with 663.9 F-statistic on 2 and 250 df (p < 0.0001).

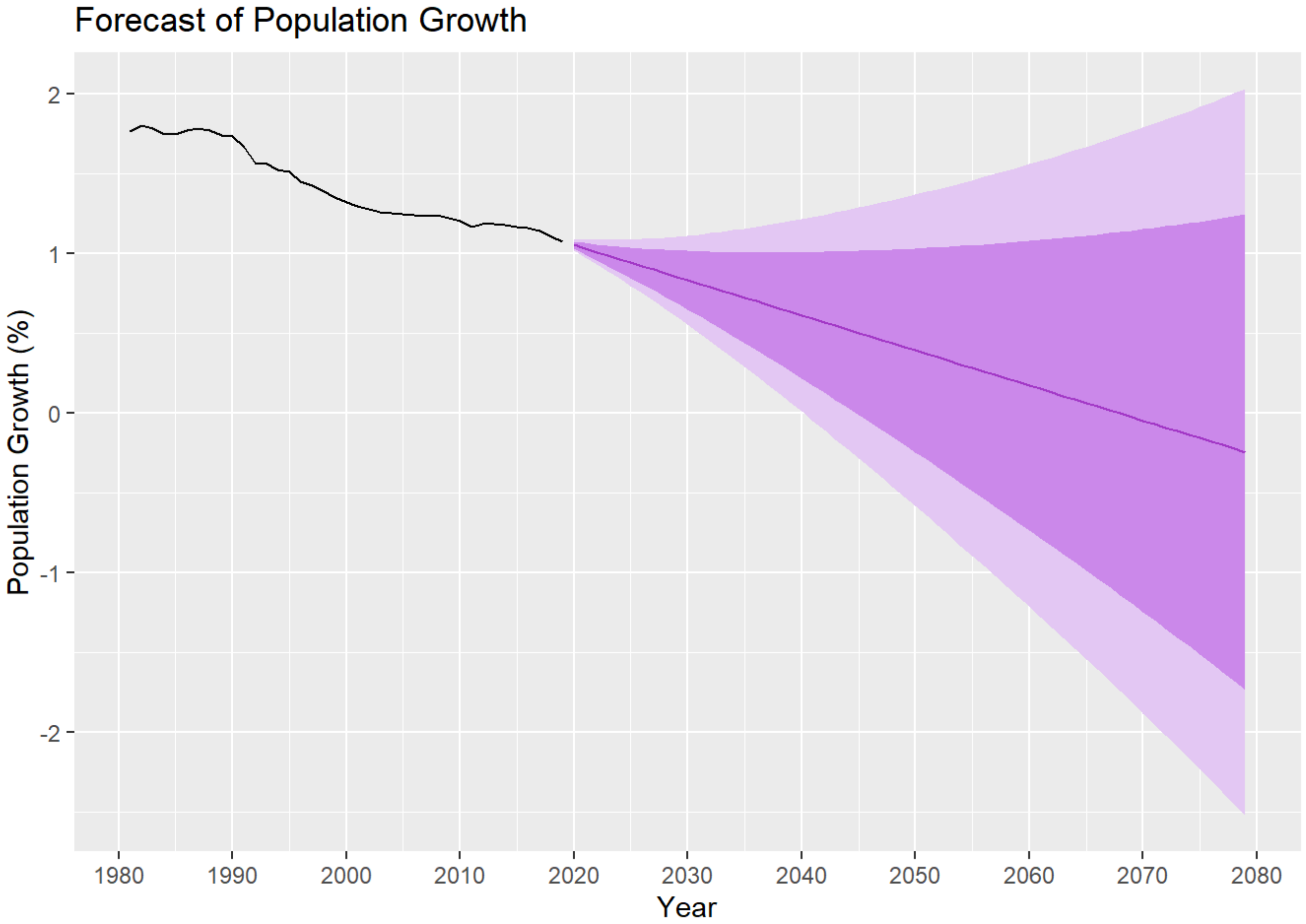
The value of adjusted r-squared was 0.84 which means that 84% of variability in UHC index is explained by our selected predictors.

Our focus then was the population growth problem. We started by seeing the population growth of world’s most populated countries India and China and compared their growth to the world.



We saw that the population growth rate was decreasing for both these countries. In fact, both these countries have a lower population growth rate than the world average currently. An interesting discovery is that the sudden dip in China’s population growth occurred in 1979. The reason behind that was the One child policy introduced by China. Though they have lifted it in 2015, its immediate results can be clearly seen.

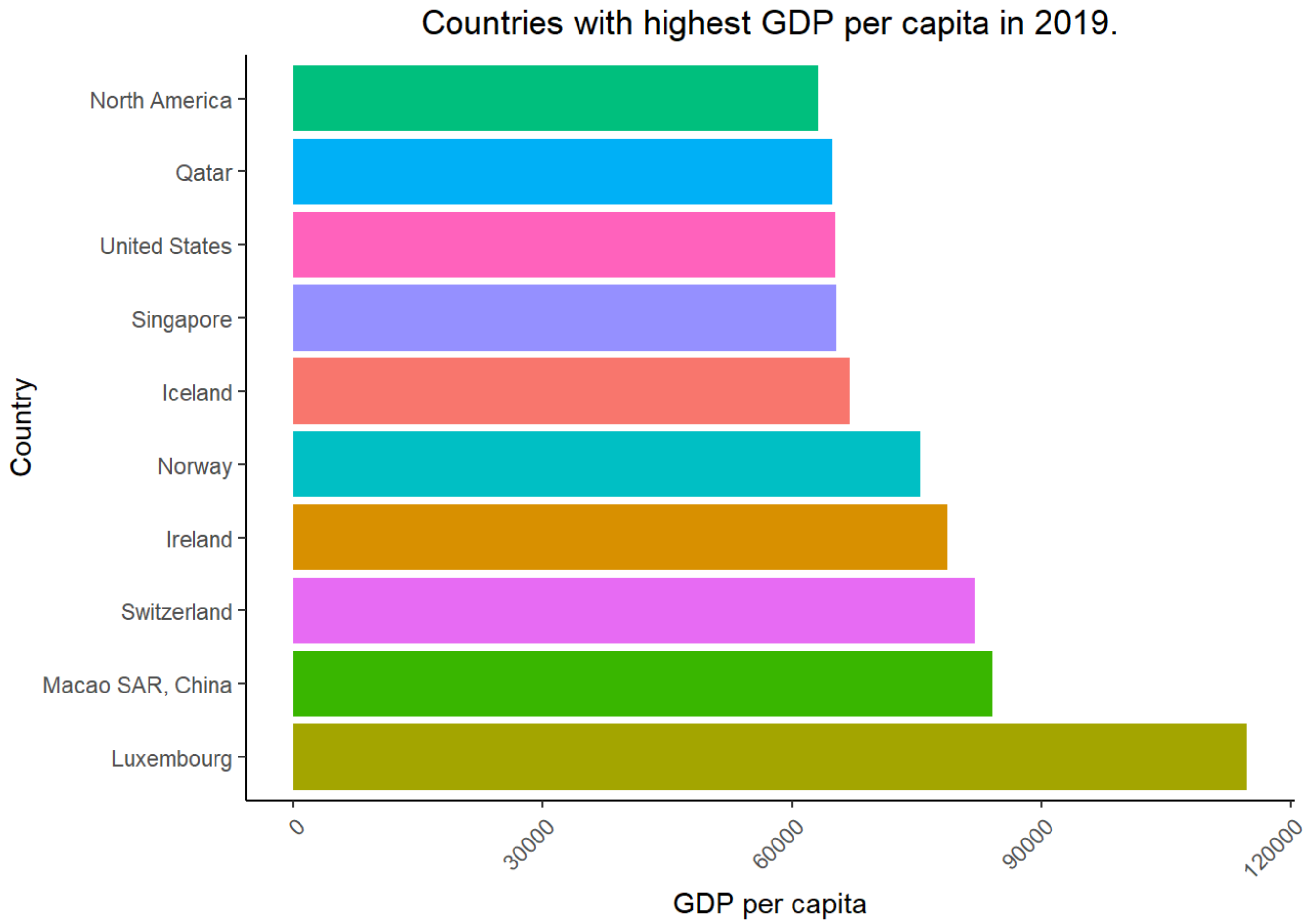
Since the population growth for these countries and the world is gradually decreasing, it seemed that the population may stabilize someday in the future. We wanted to see when the population growth will get to 0. We used forecast function by r, which can be used for forecasting time series, to find the approximate year when population growth may get to 0.

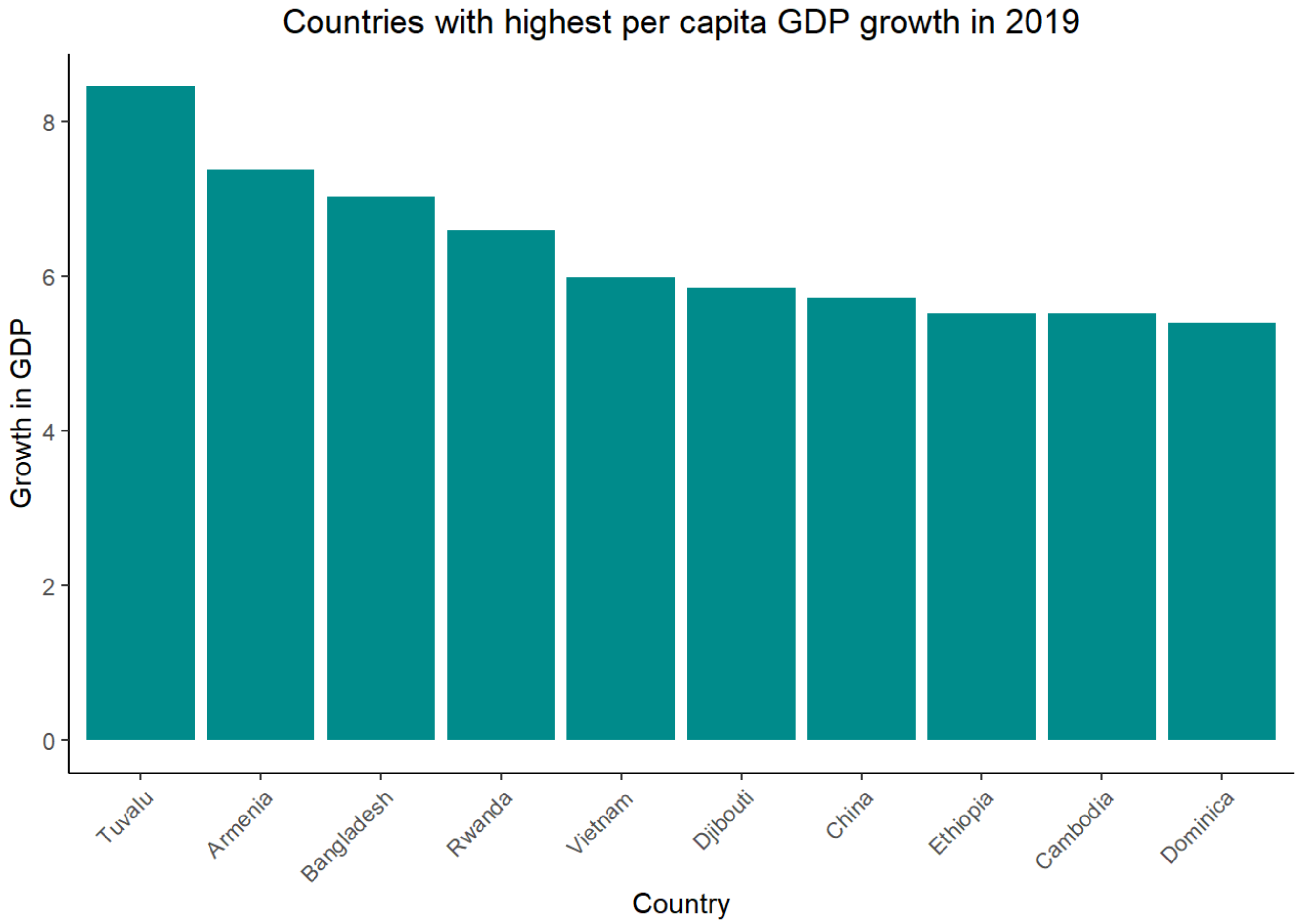


The results suggest that the population growth will get to 0 % around 2070. The results are close to a study done in 2014 which showed it will happen around 2075.

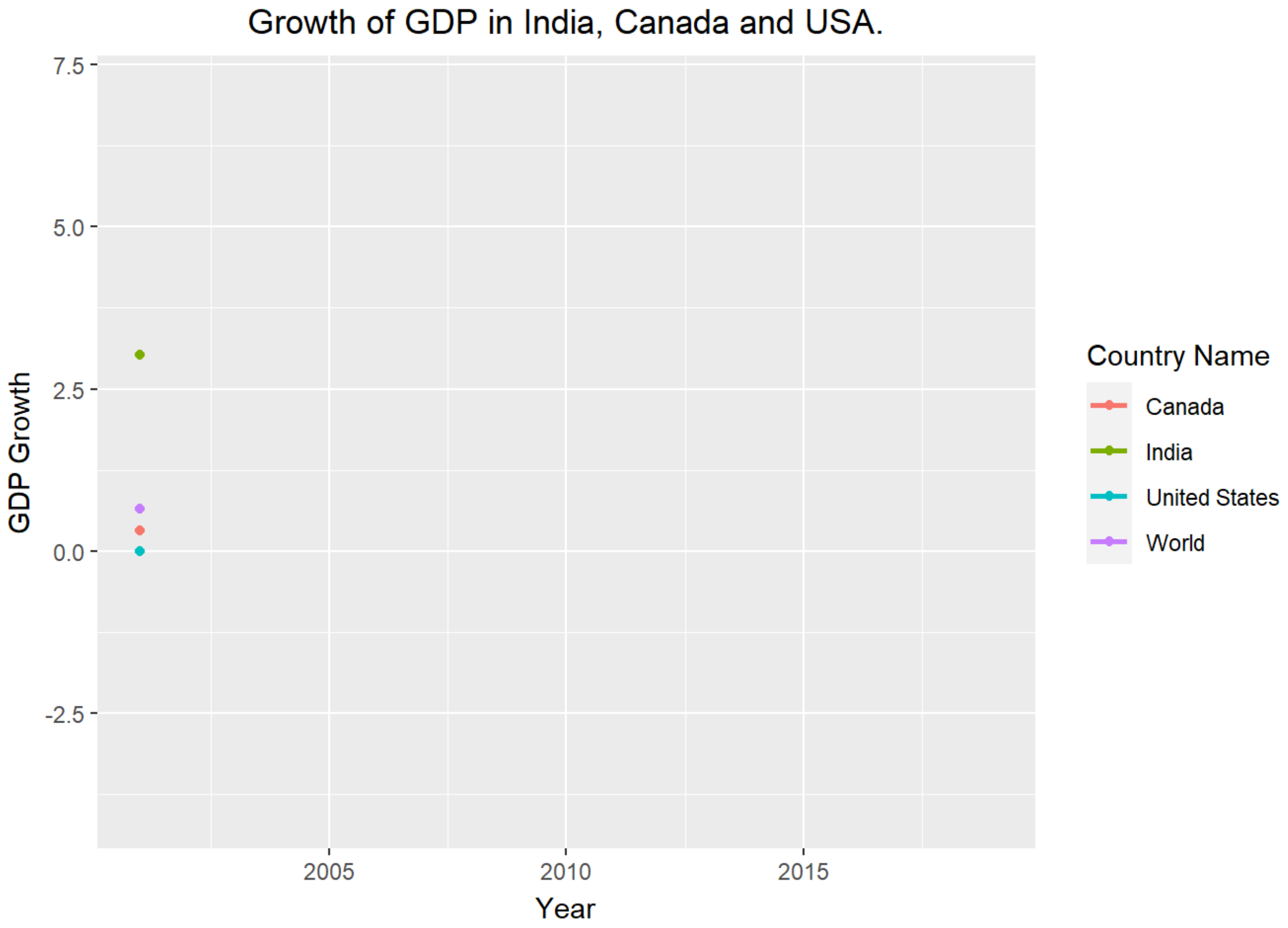
Economy

Our next focus was economy. We started by seeing the countries with highest GDP. European and North American countries topped this list. We then saw which countries have highest growth rate of GDP and as expected, this list was topped by countries that are developing but have good economies.



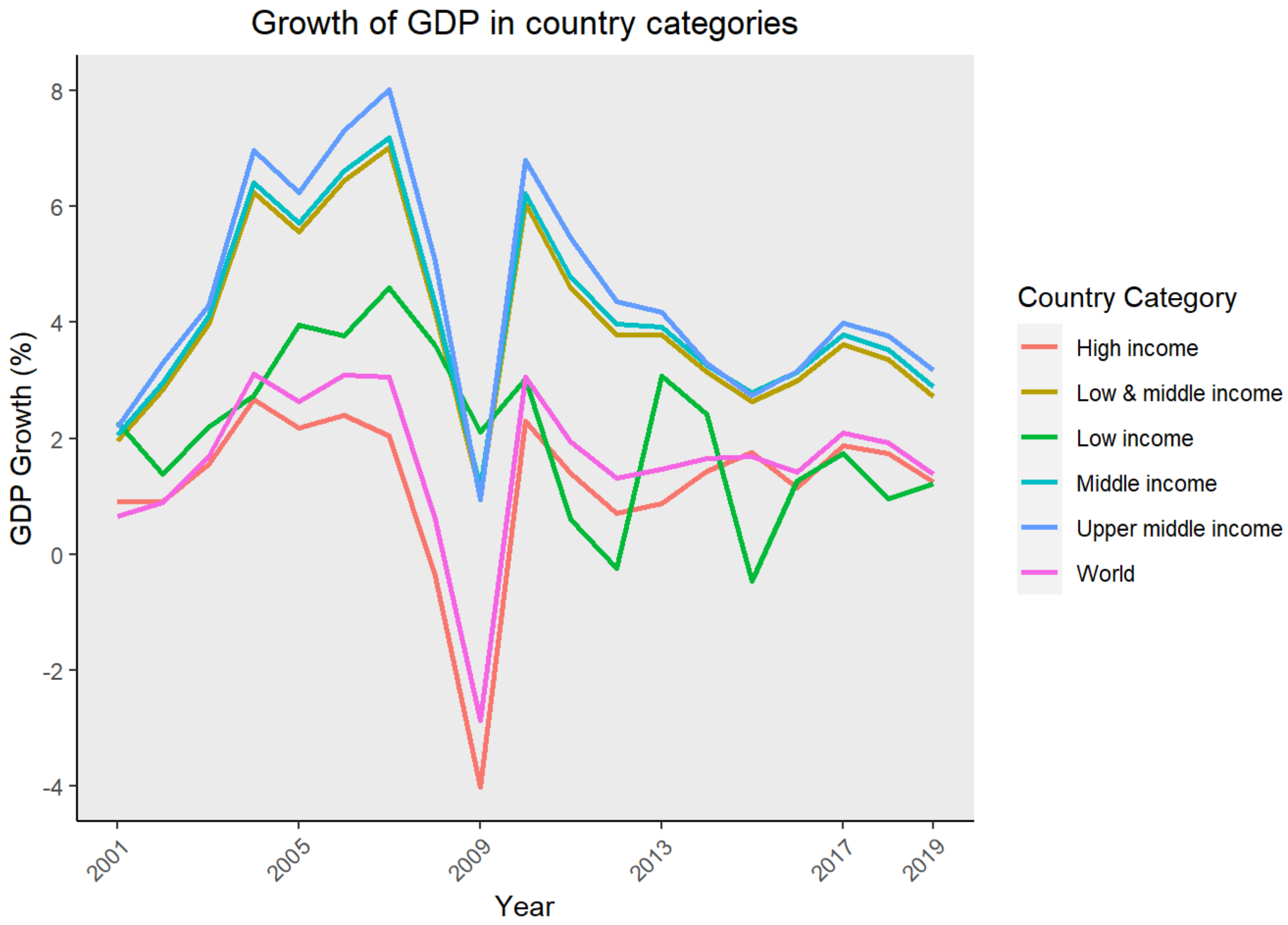


The next thing we checked was the growth rate of USA and Canada, which are first world countries and India which is a developing nation to look for conclusive trends.



Out of these countries, India has consistently had highest GDP growth rate. US and Canada have had similar growth rates. Around 2008, all the countries had lowest growth rate in years. US, Canada and world average had gone below zero percent. This can be attributed to the Economic recession which hurt almost the entire world economically.

We wanted to see if similar results are reflected by other countries with similar economic categories, so we plotted various country categories.



The results we found were interesting. One of the things we noticed was how Lower middle income, middle income and upper middle income countries' economies almost go hand in hand. These categories are above the rest or world average which suggests that they are economies that are getting constantly better.High income countries have economies who have relatively stable growth rate just above zero. The world economy, however, is heavily influenced by the higher income countries. The line graphs of these two categories go together through the years.

Does US have a higher per capita GDP than Canada?

We compared the adjusted GDP per capita of Canada and USA which are believed to be equally wealthy. We ran a hypothesis test with the null hypothesis that USA and Canada have the same GDP per capita over the years.

```
## # A tibble: 1 x 12
##   estimate .y.   group1 group2    n1    n2 statistic       p conf.low conf.high
## *    <dbl> <chr> <chr>  <chr>  <int> <int>     <dbl>   <dbl>    <dbl>     <dbl>
## 1  -26649. SL.G~ Canada Unite~    28    28        64 3.00e-9  -33099.   -19293.
## # ... with 2 more variables: method <chr>, alternative <chr>
```

```
## # A tibble: 1 x 7
##   .y.               group1 group2        effsize    n1    n2 magnitude
## * <chr>             <chr>  <chr>           <dbl> <int> <int> <ord>
## 1 SL.GDP.PCAP.EM.KD Canada United States   -2.20    28    28 large
```

The result showed that the GDP of USA was statistically higher than that of Canada. Since the data was not normally distributed, we ran Wilcoxon test which showed that the GDP Per Capita of US was high with a W-statistic of 64 (p<0.0001). Running the cohens_d test revealed that the magnitude of this difference was large.

Income inequality and GDP per capita

One of the most interesting facts we found was the correlation between GDP and percent of income share held by the wealthiest and least wealthiest. Here are the results of correlation tests between GDP per capita and income share held by the bottom 10% :

```
## 
##  Pearson's product-moment correlation
## 
## data:  data_proper$SI.DST.FRST.10 and data_proper$NY.GDP.PCAP.CD
## t = 13.75, df = 1664, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2756254 0.3618957
## sample estimates:
##       cor 
## 0.3194222
```

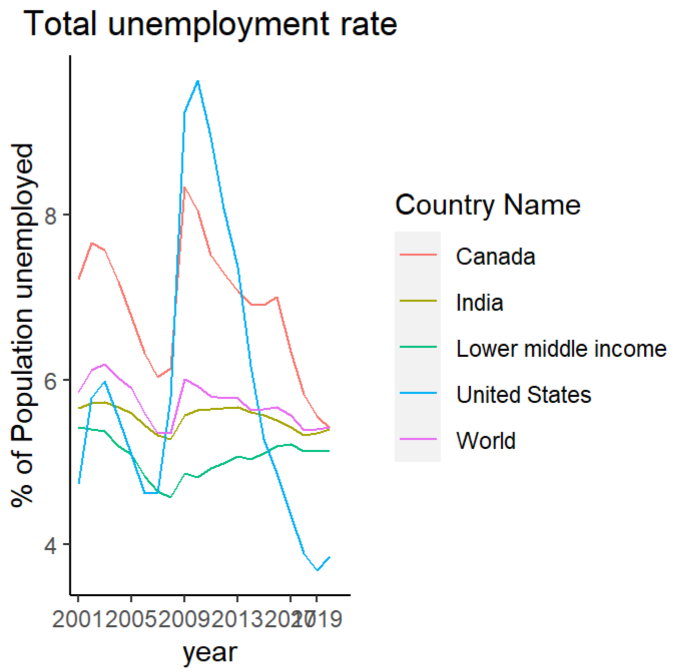
And here is the result of pearson correlation test between GDP per capita and income share held by the top 10% :

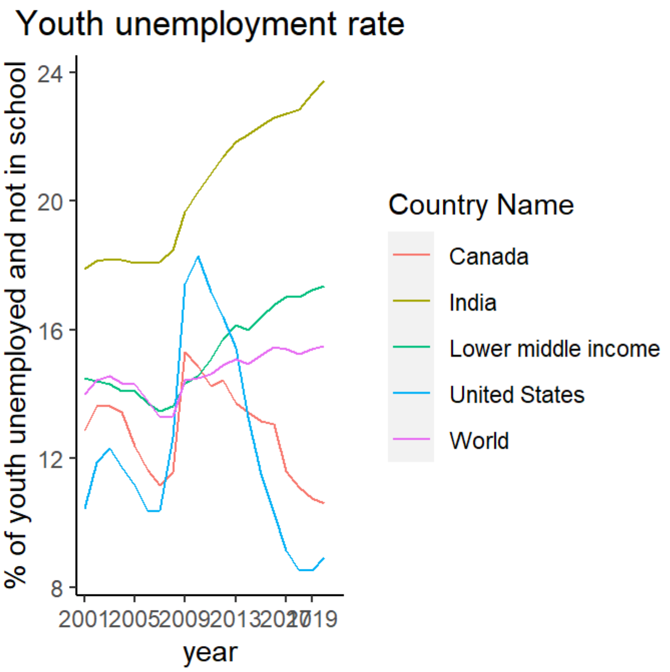
```
## 
##  Pearson's product-moment correlation
## 
## data:  data_proper$SI.DST.10TH.10 and data_proper$NY.GDP.PCAP.CD
## t = -19.352, df = 1665, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4669204 -0.3884979
## sample estimates:
##        cor 
## -0.4285159
```

The GDP per capita was directly proportional to the income share held by the bottom 10% of the country (r=0.32), but it was inversely proportional to the income share held by the richest 10% (r= -0.43). This was surprising and it clearly displayed how income inequality was a problem for the overall well-being of a country.

Unemployment problem in India

When we checked the unemployment rate of India, US and Canada, India’s unemployment rate turned out to be the lowest. It was surprising since US and Canada have much better economies than India. It turns out that the real problem in India is underemployment, which isn’t measured well in a country like India. We decided to check Youth unemployment rates (People between 16-25 who are not in school nor working). That gave us a much better picture of the performance of these countries.

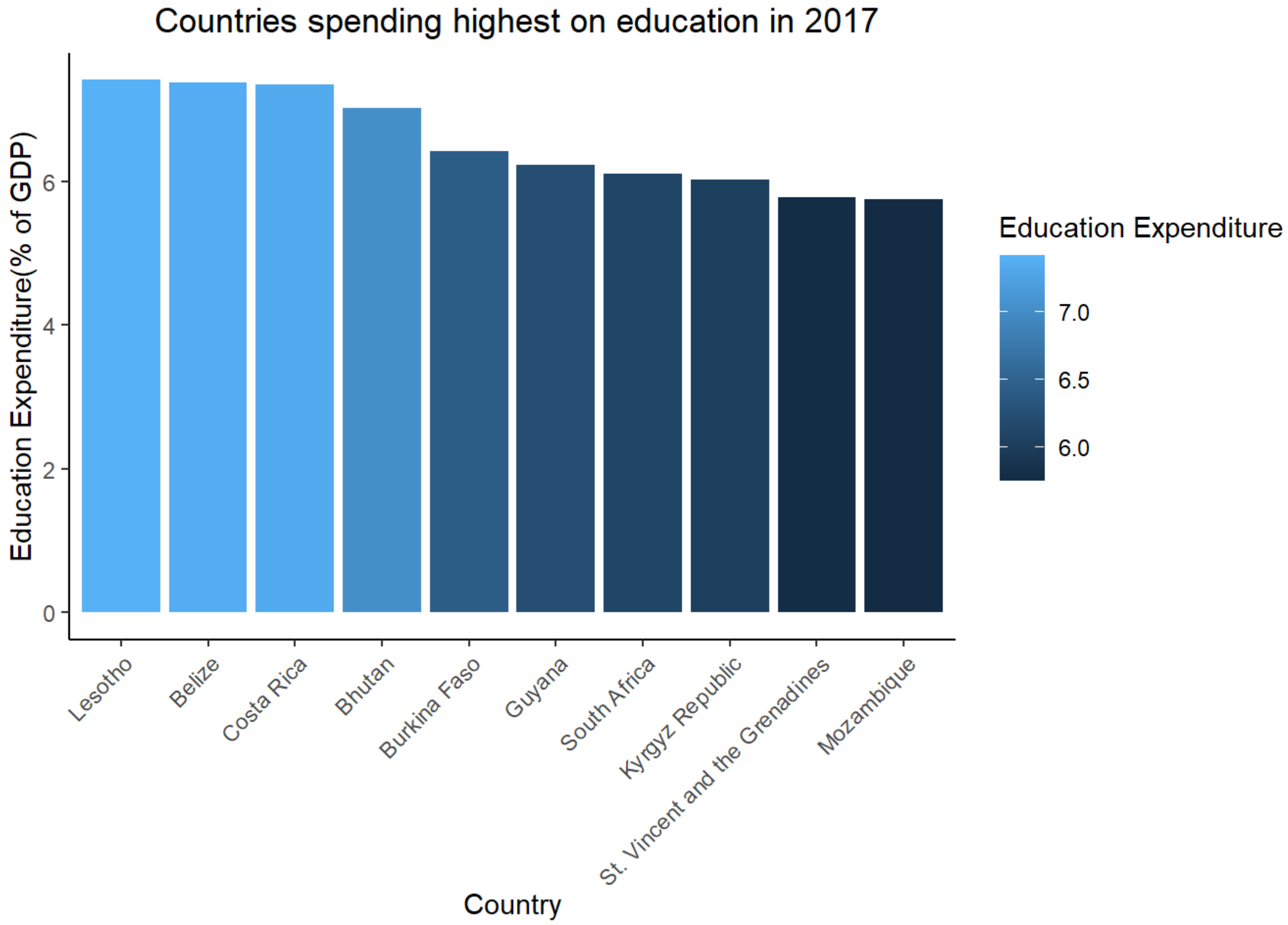




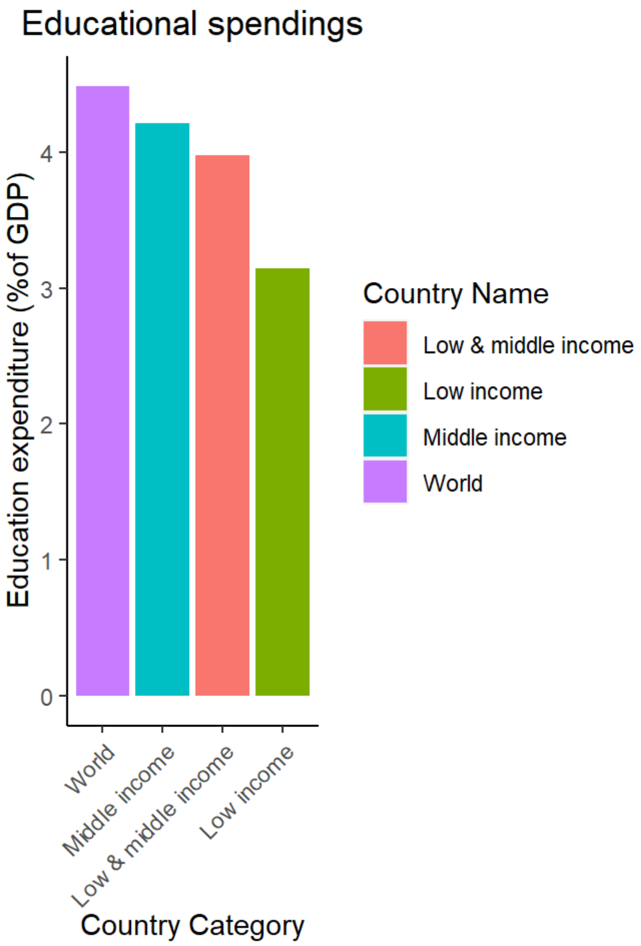
In both of the graphs, US and Canada have a drop at around 2008 near the recession years, but India has a much higher rate of youth unemployment than US and Canada. The fact that it is much higher than even the lower middle income countries and is continuously increasing in past few years is concerning.

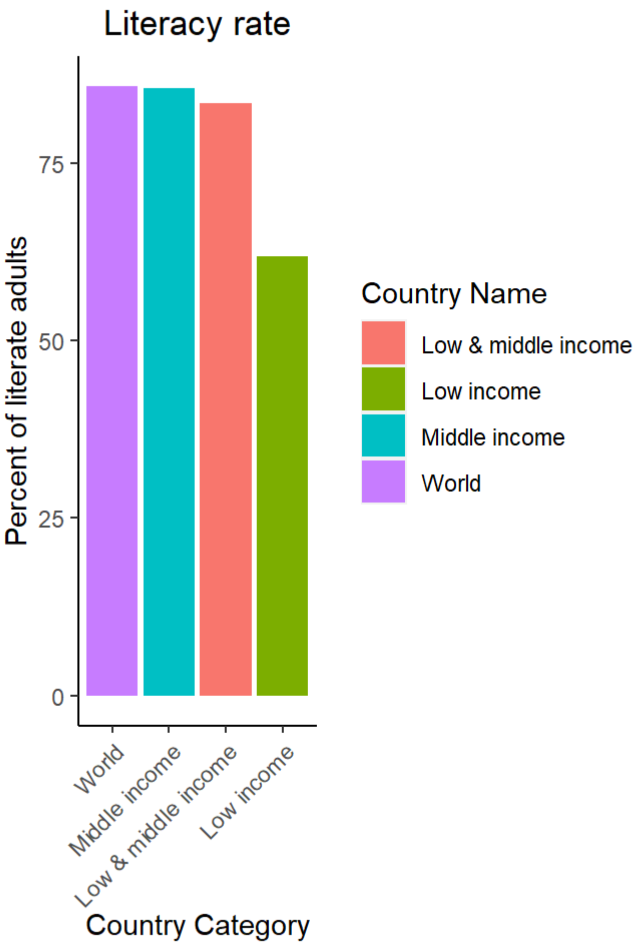
Education

The very first thing that we wanted to check in education was to see that which were the countries that spent the most on education. For most of the parts, we wanted to analyze the indicators for the latest year in which the largest amount of data was available. So, the spending was checked for the year 2017, which turned out to be that the most of the countries among the top 10 countries were African countries such as Lesotho(7.42%), Burkina Faso(6.42%), Guyana(6.23%), South Africa(6.11%) and Mozambique(5.75%).



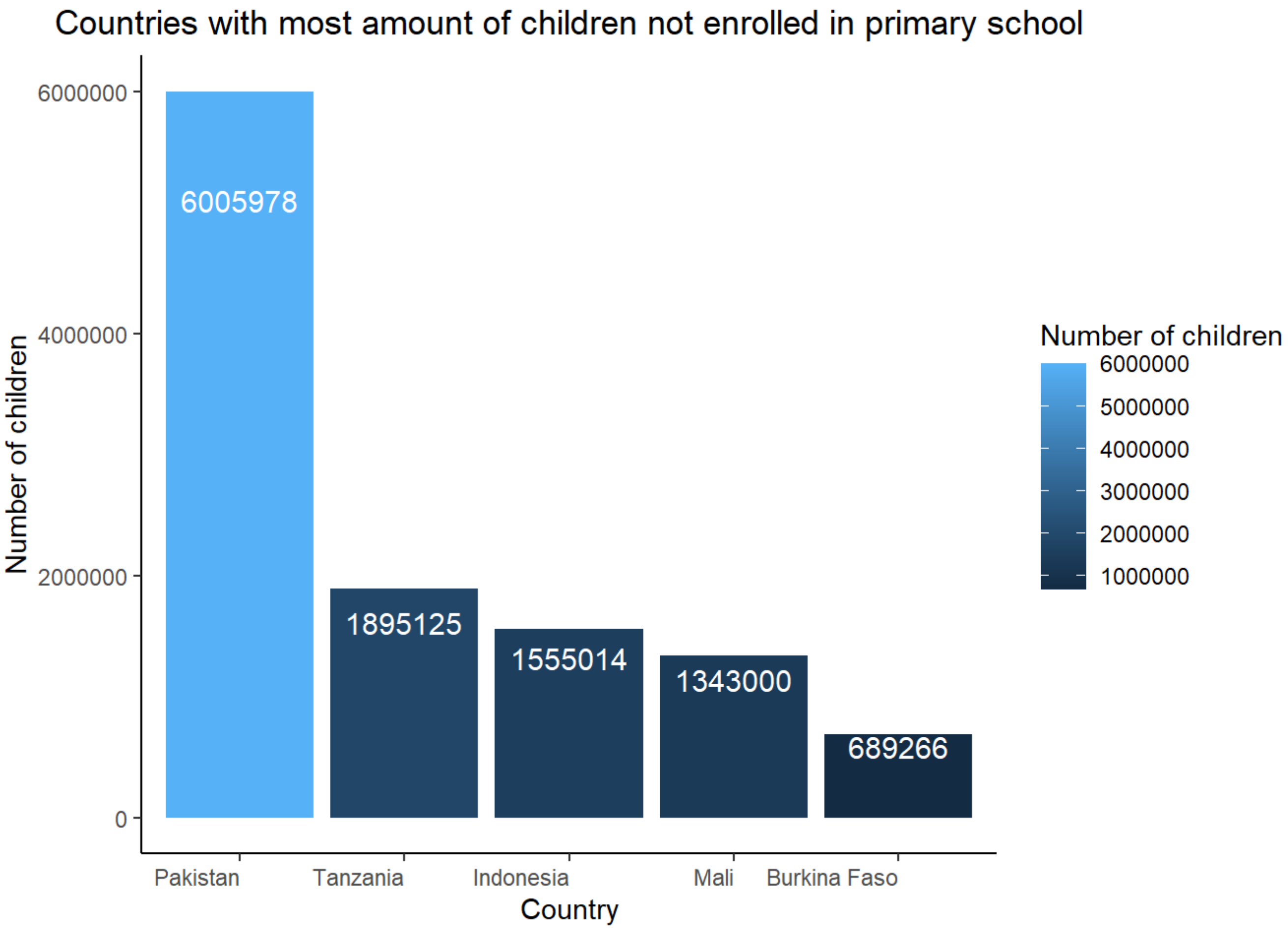
Then we checked the educational expenditure for the country categories according to the income and also the average of the world and compared it to the literacy rates for the same. The country categories were Low income, Low & middle income, Middle income. The results indicated that the more a country spends on education, the more the literacy rate of the country is. Also, the better the economy of a country is, the more it spends on education. Therefore, middle income countries spend more on education than low & middle income countries, and as a result, the literacy rate of middle income countries is better than that of low income countries as well.





Education problem in Pakistan

One really interesting indicator in education was the one that displayed the amount of primary age children that should have been enrolled in primary or secondary school, but are not. After delving into this indicator, one shocking fact was found that in the year 2018, there were more than six million primary school age children in Pakistan who were not enrolled in school. This number is three times greater than that of Tanzania, which is the 2nd country for this indicator.



The last thing that we checked in education sector was the educational attainment for different levels and seeing the top 10 countries in each country. It turns out that in Germany, 100% of adults have attained primary level education. Germany also comes in top 10 countries in other level of education (upper secondary, bachelors, masters and doctoral). Uzbekistan, Kazakhstan, Australia, Maldova and United States of America have 99% of adults that completed primary education. Kazakhstan also tops in upper secondary level education.United Arab Emirates has most number of bachelors graduated while Switzerland has most number of people that have masters and doctorate degrees.

Countries with highest people with Primary education attainment

```
## # A tibble: 10 x 2
##    `Country Name`     `Percent of people with primary education`
##    <chr>                                                   <dbl>
##  1 Germany                                                 100
##  2 Uzbekistan                                              100.
##  3 Kazakhstan                                               99.9
##  4 Australia                                                99.7
##  5 Moldova                                                  99.4
##  6 United States                                            99.0
##  7 Malta                                                    98.9
##  8 Netherlands                                              98.5
##  9 San Marino                                               96.6
## 10 West Bank and Gaza                                       94.6
```

Countries with highest people with Secondary education attainment

```
## # A tibble: 10 x 2
##    `Country Name` `Percent of people with secondary education`
##    <chr>                                                 <dbl>
##  1 Kazakhstan                                             97.4
##  2 Uzbekistan                                             96.1
##  3 Latvia                                                 90.3
##  4 United States                                          89.8
##  5 Estonia                                                87.6
##  6 Switzerland                                            85.7
##  7 Germany                                                83.5
##  8 Denmark                                                79.3
##  9 Australia                                              78.3
## 10 Moldova                                                75.2
```

Countries with highest people with Bachelor’s degree

```
## # A tibble: 10 x 2
##    `Country Name`       `Percent of people with Bachelors degree`
##    <chr>                                                    <dbl>
##  1 United Arab Emirates                                      47.3
##  2 United States                                             35.0
##  3 Kazakhstan                                                34.1
##  4 Denmark                                                   32.4
##  5 Australia                                                 31.7
##  6 Singapore                                                 31.6
##  7 Netherlands                                               31.1
##  8 Latvia                                                    30.0
##  9 Germany                                                   25.1
## 10 Spain                                                     21.9
```

Countries with highest people with Master’s degree

```
## # A tibble: 10 x 2
##    `Country Name` `Percent of people with Masters degree`
##    <chr>                                            <dbl>
##  1 Switzerland                                       20.0
##  2 Latvia                                            14.6
##  3 Portugal                                          13.6
##  4 United States                                     13.1
##  5 Spain                                             12.9
##  6 Denmark                                           12.7
##  7 Peru                                              11.9
##  8 Netherlands                                       11.7
##  9 Germany                                           11.2
## 10 San Marino                                        11.1
```

Countries with highest people with PhD

```
## # A tibble: 10 x 2
##    `Country Name`       `Percent of people with a PhD`
##    <chr>                                         <dbl>
##  1 Switzerland                                   2.93
##  2 United States                                 2.03
##  3 Germany                                       1.25
##  4 Australia                                     1.14
##  5 Latvia                                        1.11
##  6 United Arab Emirates                          0.905
##  7 Denmark                                       0.779
##  8 Spain                                         0.660
##  9 Netherlands                                   0.627
## 10 Portugal                                      0.589
```

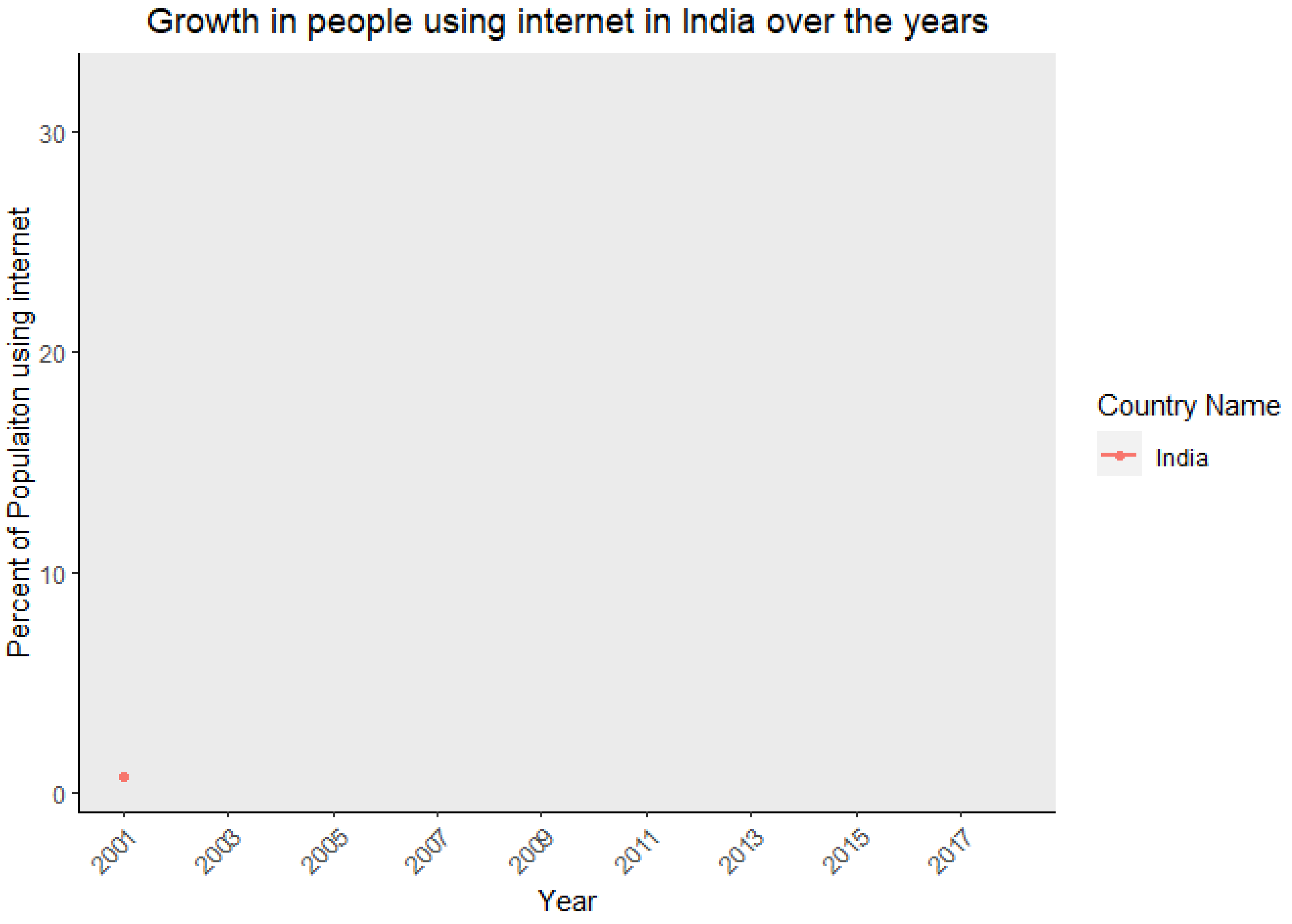
Then we wanted to see the correlation between the percent of youth educated and percent of youth unemployed. The value for the Pearson Correlation Coefficient is 0.36 indicating that there is a moderate correlation between the aforementioned indicators.

```
## 
##  Pearson's product-moment correlation
## 
## data:  scale(data_proper$SL.UEM.1524.ZS) and scale(data_proper$SE.ADT.1524.LT.ZS)
## t = 16.571, df = 1844, p-value < 0.00000000000000022
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3196377 0.3990789
## sample estimates:
##       cor 
## 0.3600107
```

Infrastructure

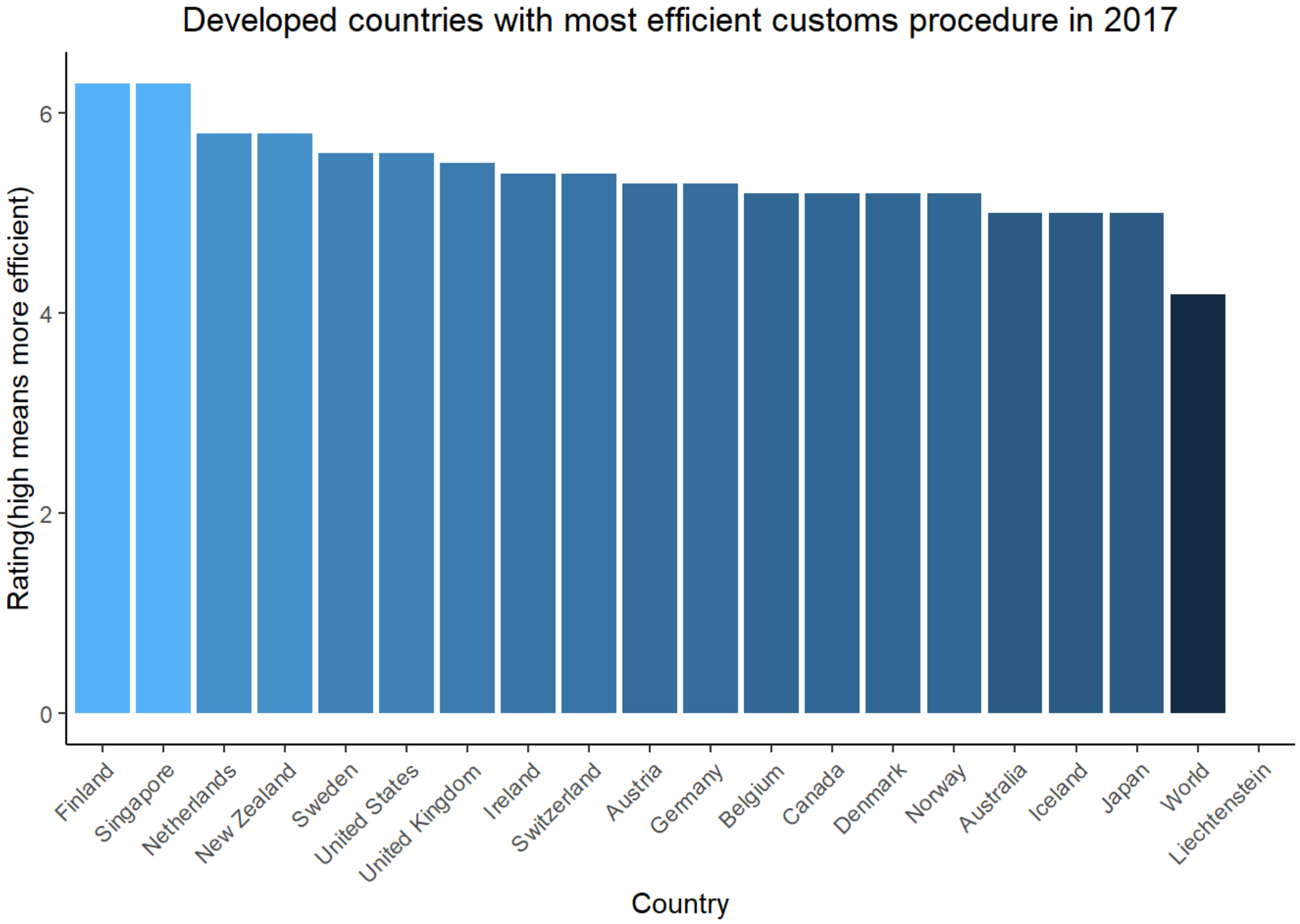
Cellphone users in India

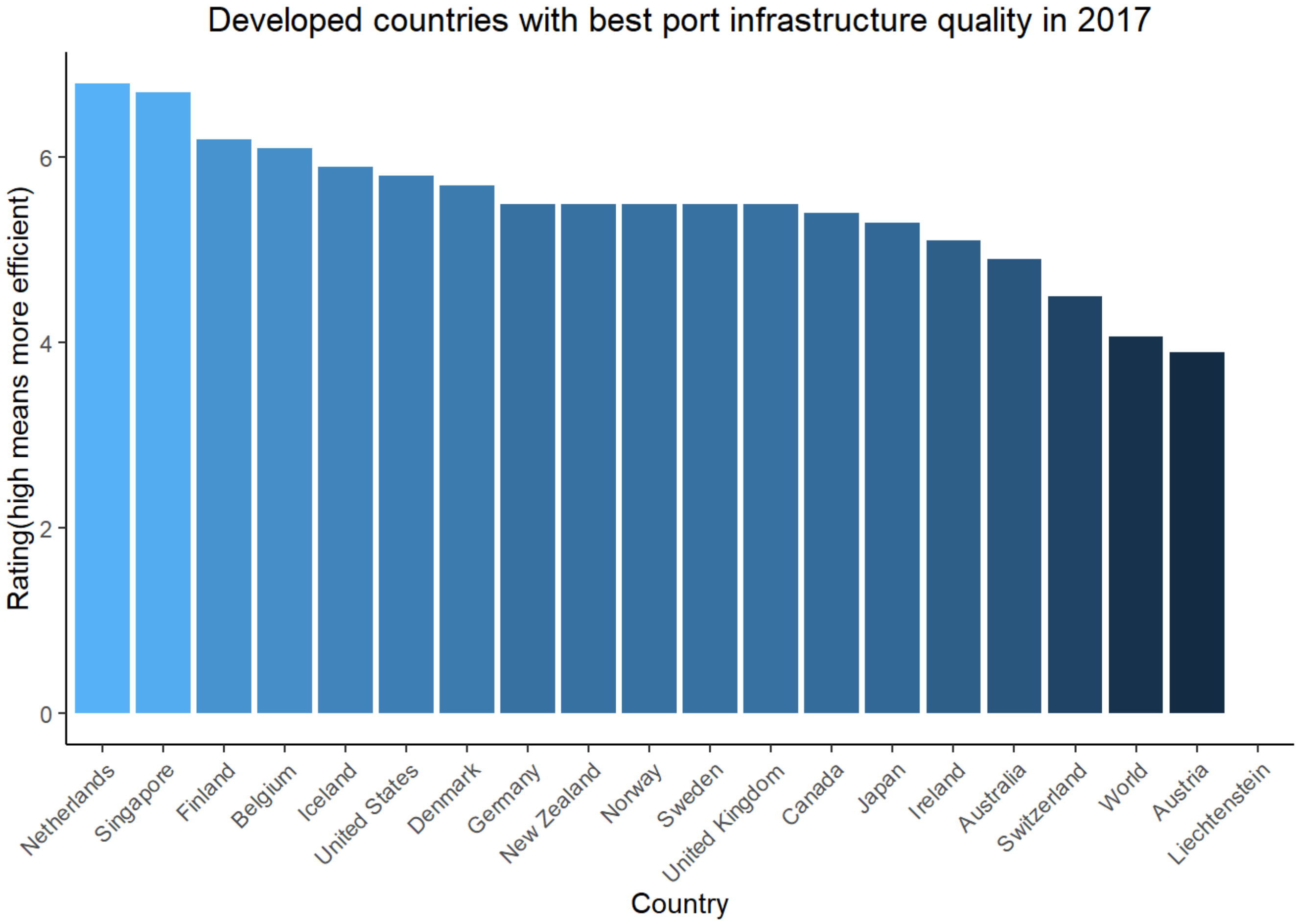
While seeing the indicators available in infrastructure sector, we were particularly intrigued by the percent of population using internet. So, we checked it for the years after 2000 in India. And an interesting thing was found that in the early 2000s, almost 0% of the population in India was using internet. The number grew gradually and reached over 20% in 2014, then dropped a bit in the following year. But after 2015, a huge spike can be seen, a 15% increase was experienced. The reason behind the growth was the introduction of a telecommunication company named Jio. The very first year of the company, a usage of 2gb of data was given to everyone who bought the sim at no cost for a whole year.



Europe’s supremacy in infrastructure

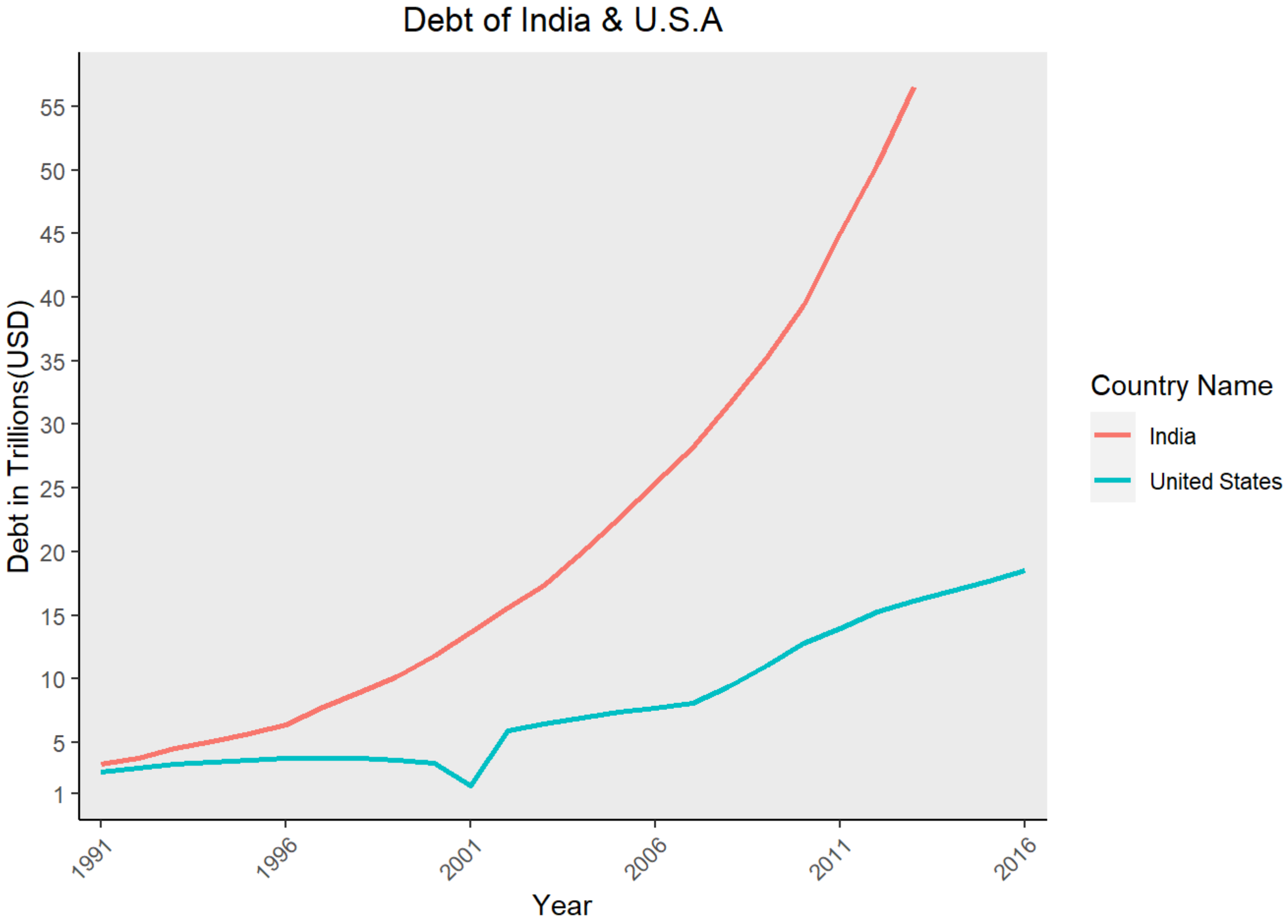
Another thing that we found in Infrastructure was that the European developed countries outperformed other developed countries of the world in the infrastructure sector. In efficiency of customs procedure, countries like Finland(6.3), Netherlands(6.3) Sweden(5.6), United Kingdom(5.5), Ireland(5.4), Switzerland(5.4) and Austria(5.3) have the highest ratings.This indicator is a rating given by the business executives of the country according to their perspective on their country’s efficiency on customs procedure. For quality of port infrastructure too, European developed countries were among the top 10. With having rating above 5 out of 7.

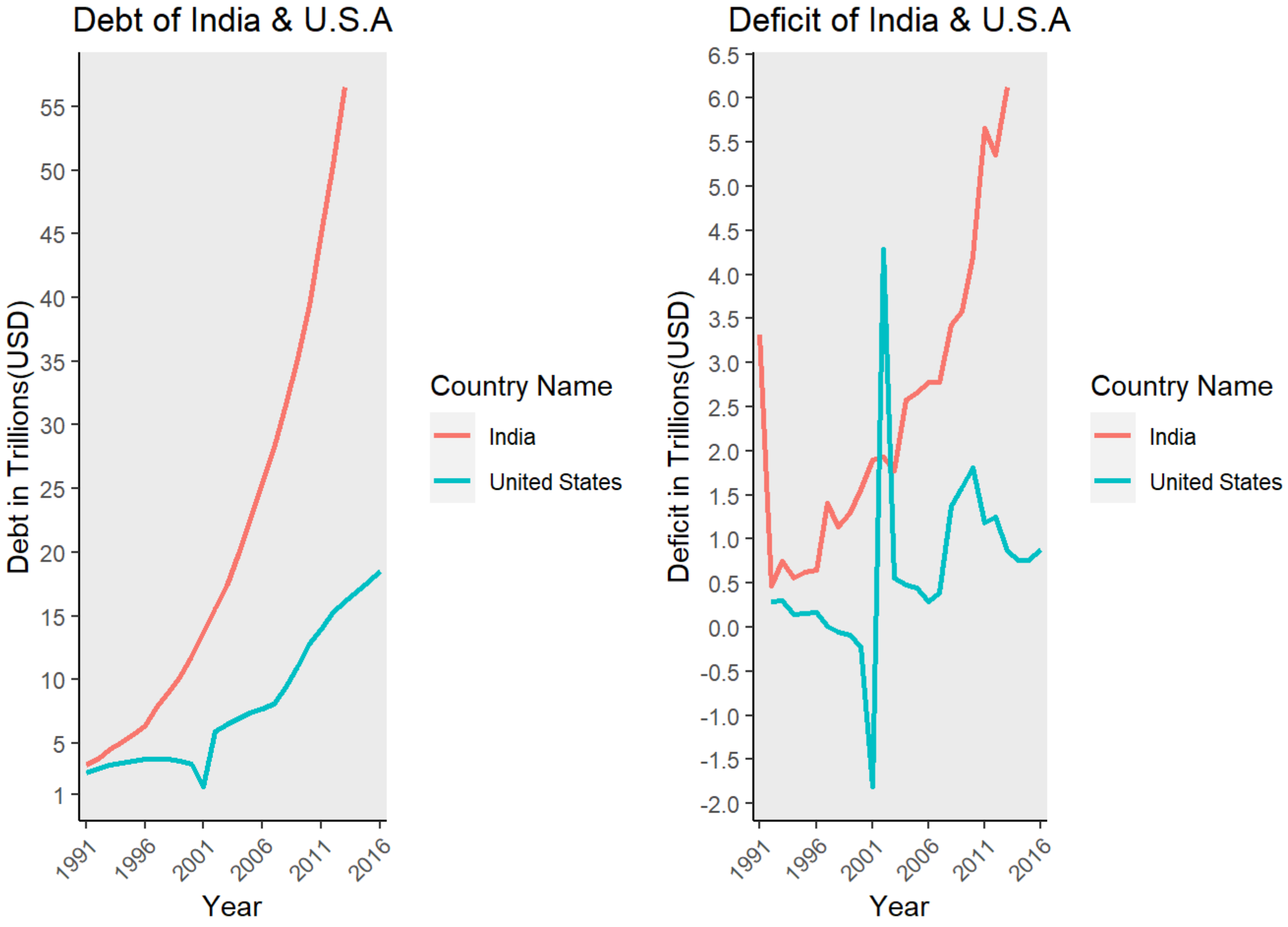




Public and Private Sector

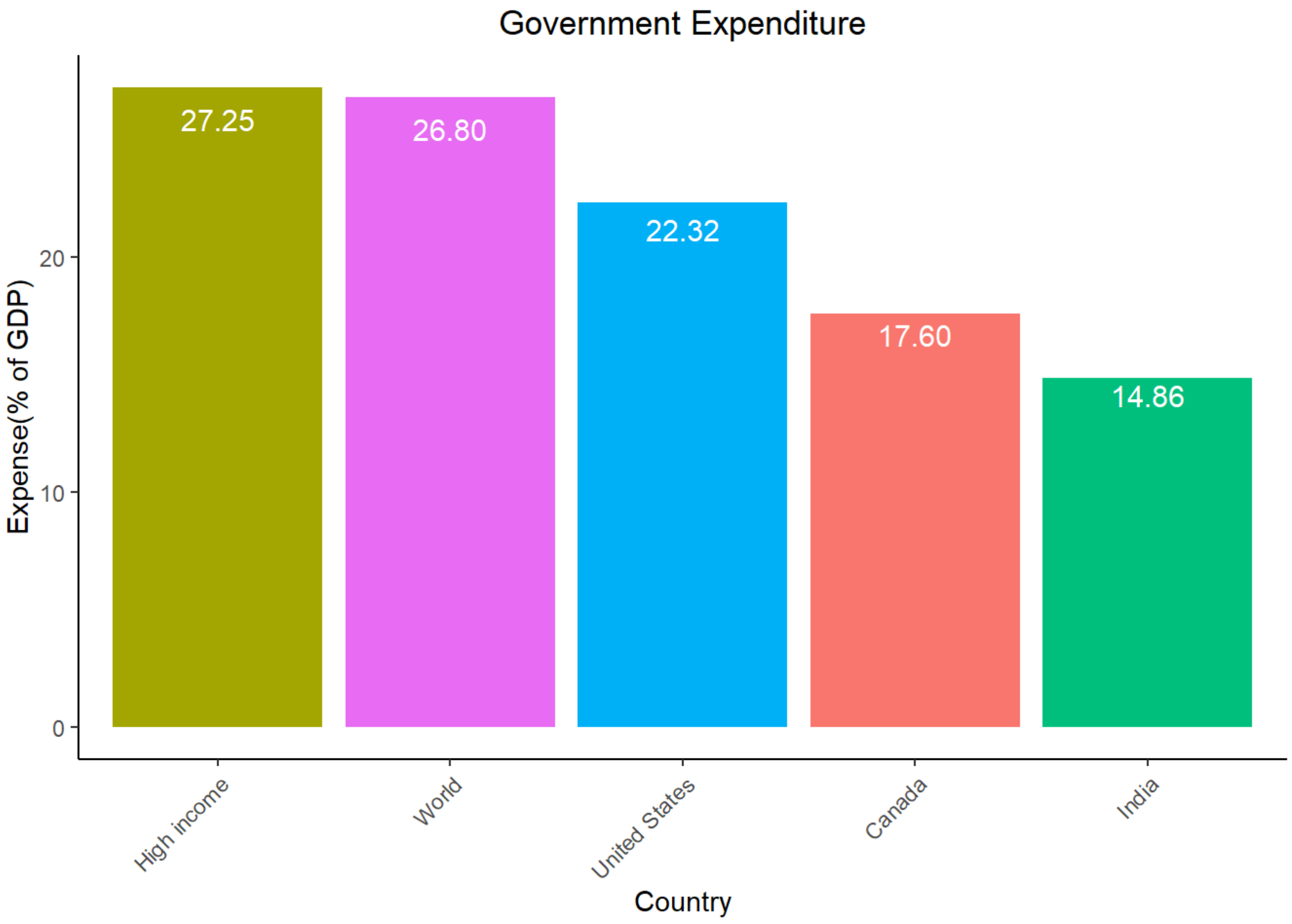
The first thing that we analyzed in this domain was the debts of countries, but we were specifically interested in India. Then we compared the debt of India with United States of America. We saw the progression of debt of both of the countries over the years. In the first graph, it can be seen that the debt of India has been rising on a very fast pace each year. It started with a bit above 1 trillion USD in 1991 and by the year 2015, it reached above 55 trillion USD. The debt of USA has also been rising over the years, but it is not as huge amount as India and also the progression is not that large. After seeing the debt, we then calculated and added the deficit for each year. We noticed a huge spike in the deficit of USA in the year 2001. On researching about it, we came to know that there was a huge tax cut in USA in 2001 and also the military expenditure of USA was very high.


Debt of India & U.S.A
Debt in Trillions(USD)
55
50
45
40
35
30
25
20
15
10
5
1
1991
1996
2001
2006
2011
2016
Year
Country Name
India
United States

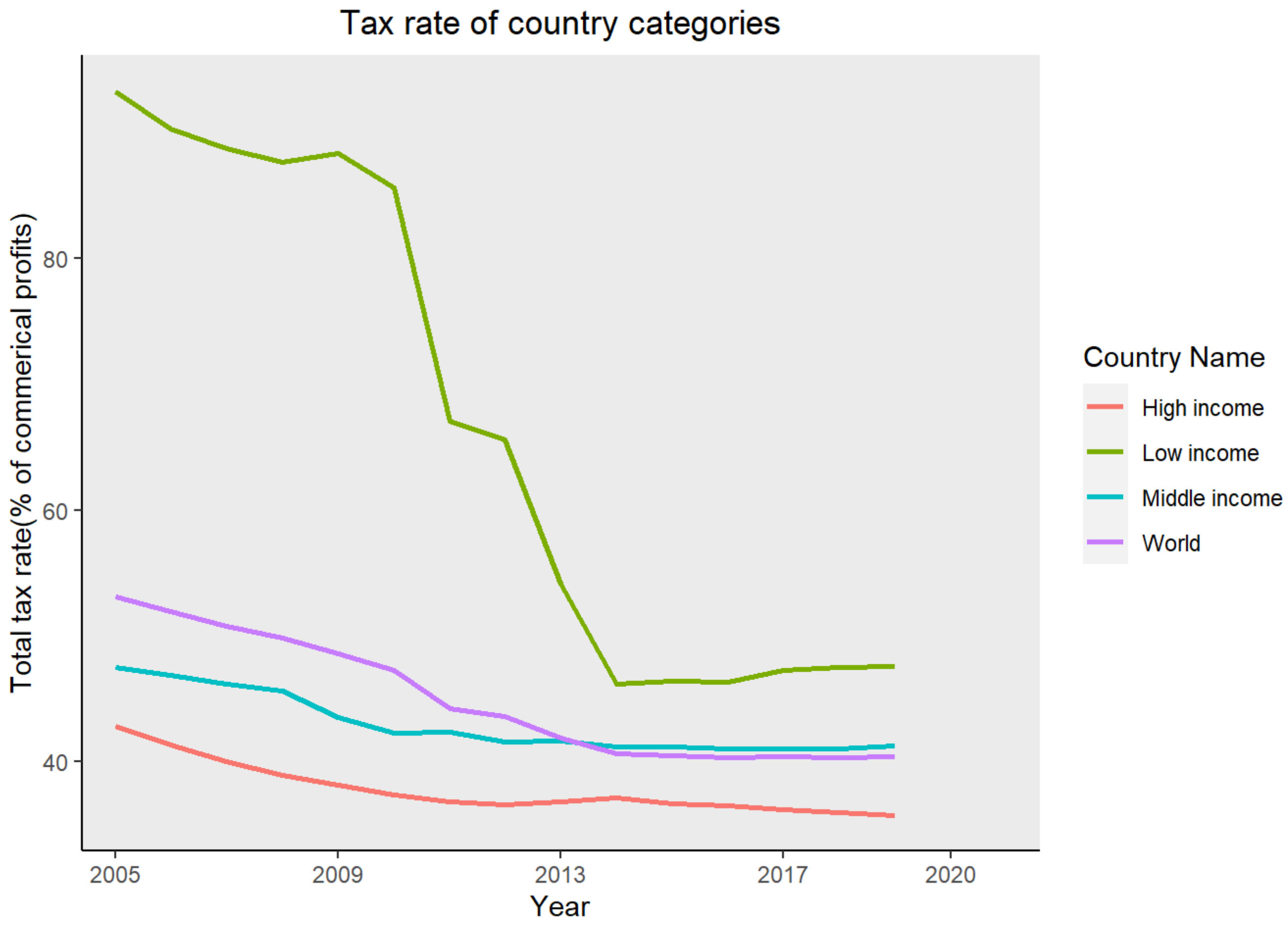

Debt of India & U.S.A
Debt in Trillions(USD)
55
50
45
40
35
30
25
20
15
10
5
1
1991
1996
2001
2006
2011
2016
Year
Country Name
India
United States
Deficit of India & U.S.A
Deficit in Trillions(USD)
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
1991
1996
2001
2006
2011
2016
Year
Country Name
India
United States

The next thing was to check the expenditure of government.For this, we checked it for the high income category to get a general idea on how much a country with a good economy would spend. We also considered World average, United States of America, Canada and India for this one, to get a diverse outcome.

In 2017,the high income countries spend more than that of the world average and as expected, USA's expenditure was also more than Canada and India.

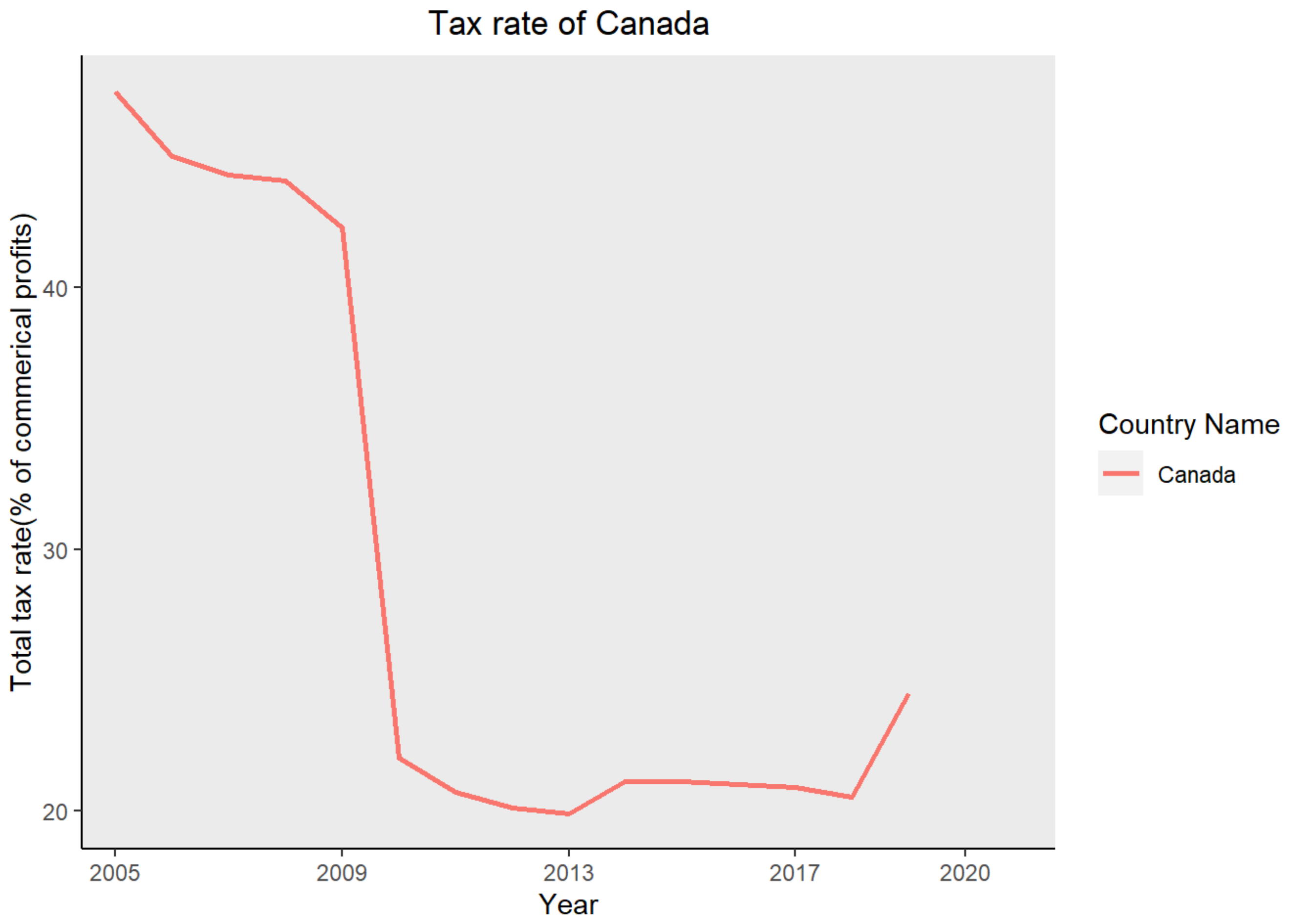


While exploring the indicators for this sector, we came around the total tax rate of a country and decided to analyze the same for the different country categories and compared it with the World average too. The tax rate of the countries with better economy have a fairly stable and low tax rate over the years but it's not the same with low income countries. As wealthy a country gets, the tax rate also decreases accordingly.



Drop in Canada’s corporate tax rates

One interesting high income country was Canada because we noticed a huge drop in the tax rate in the year 2009. So after researching why it happened, we came to know that due to recession, to keep up with the world market, Canada had to lower the tax rate to almost half(42% - 22%). Since then, it has been a bit stable and is increasing in the recent years.



Do better economies have less bribery incidences?

We came across an interesting indicator that recorded that how much percent of firms were offered bribe at least once in a country. So, we then delved into the indicator and noticed that better economies had less bribery incidences. Therefore, to confirm it, we ran a Pearson correlation test between bribery incidences and GDP of a country. The value for the coefficient was -0.45 suggesting that better economies do experience less bribery incidences.

```
## 
##  Pearson's product-moment correlation
## 
## data:  data_proper_all$NY.GDP.PCAP.PP.CD and data_proper_all$IC.FRM.BRIB.ZS
## t = -9.3466, df = 340, p-value < 0.00000000000000022
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5326350 -0.3635021
## sample estimates:
##        cor 
## -0.4521233
```

Predictors of GDP

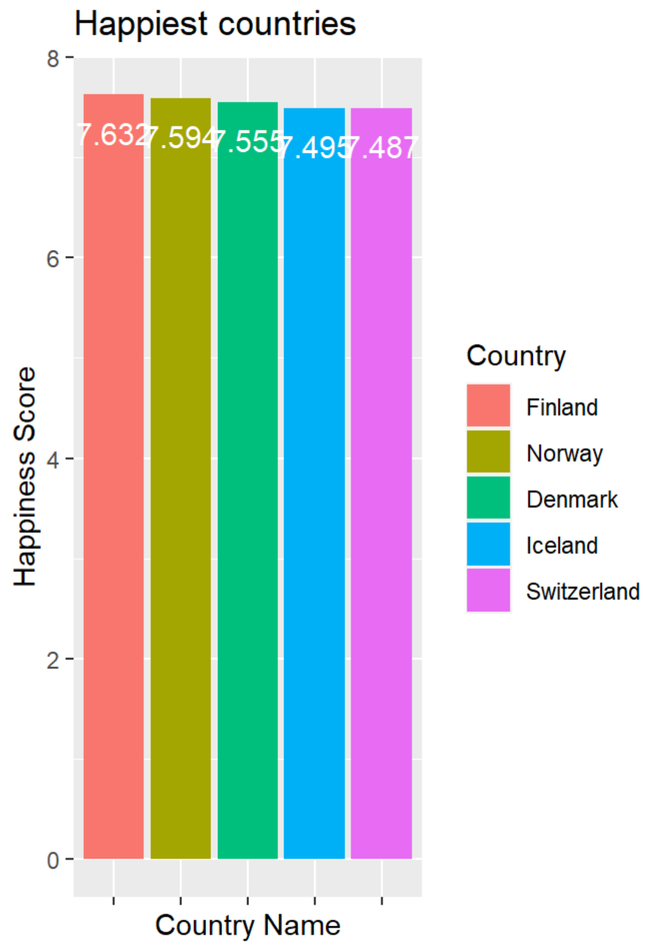
After observing all the data and finding interesting models and correlations, we wanted to see if we can find some definitive facts about how human well being can be achieved. Firstly, we used GDP per capita as the metric of human well being and checked which specific factors affect it the most. We ran multiple models with a combination of different predictors but the one with highest value of adjusted r-squared was the model with a combination of UHC index, Bachelor’s education attainment rate and unemployment rate. Out of these three, UHC index and Educational attainment were the most significant ones. This makes sense as Health and Education are considered really important and basis for the progress of a country.

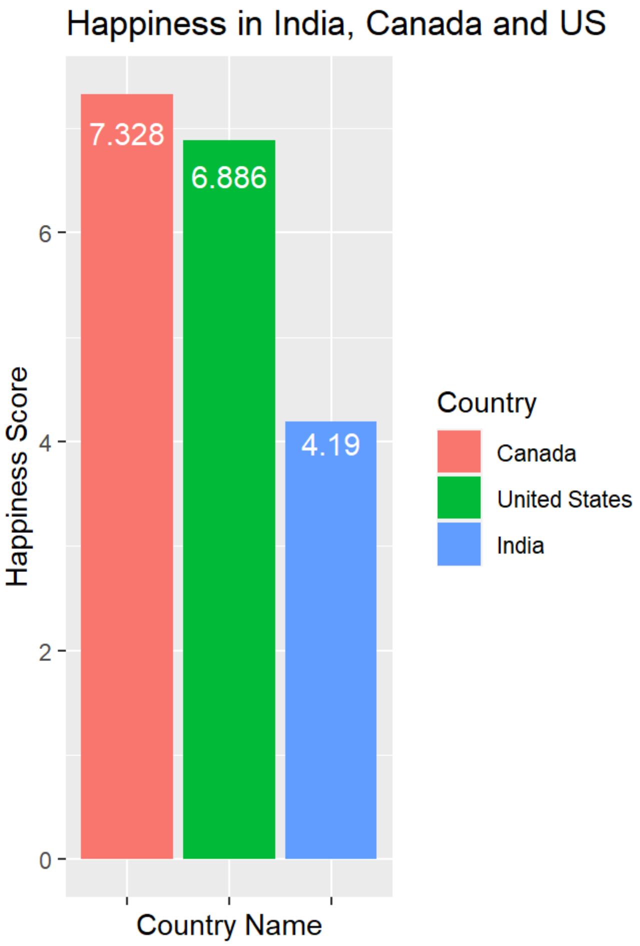
```
## 
## Call:
## lm(formula = NY.GDP.PCAP.CD ~ SH.UHC.SRVS.CV.XD + SE.TER.CUAT.BA.ZS + 
##     SL.UEM.TOTL.ZS, data = data_proper_all)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -19849 -11214  -2991  10827  36569 
## 
## Coefficients:
##                   Estimate Std. Error t value   Pr(>|t|)    
## (Intercept)       -48341.0    11532.2  -4.192 0.00007015 ***
## SH.UHC.SRVS.CV.XD    850.2      178.2   4.770 0.00000803 ***
## SE.TER.CUAT.BA.ZS    805.9      204.3   3.945   0.000169 ***
## SL.UEM.TOTL.ZS      -787.9      304.9  -2.585   0.011545 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 14140 on 81 degrees of freedom
##   (16283 observations deleted due to missingness)
## Multiple R-squared:  0.5803, Adjusted R-squared:  0.5648 
## F-statistic: 37.34 on 3 and 81 DF,  p-value: 0.000000000000002975
```

The adjusted r-squared for this was 0.56. UHC index, Bachelor’s attainment rate and total unemployment rate can explain 56% of variability in GDP per capita.

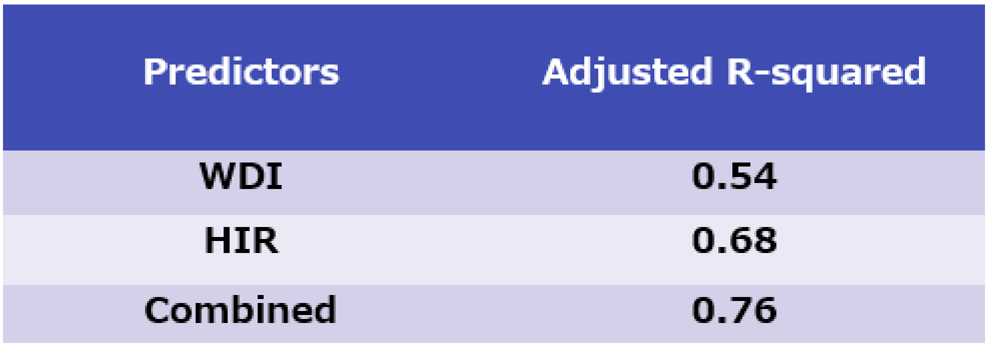
What Predicts Happiness?

Though GDP is certainly a really important factor of well being of a country, we decided to go a step further and echo the Happiness index report dataset. Happiness Index Report consists of subjective data where questions are asked to people to get their perceived answers. The dataset has Happiness Score, Freedom to make life choices, Generosity, Social support, perception of corruption. Here is a brief look at the dataset:





We considered the average self-perceived happiness score of a country as a metric. Since all of the indicators in the HIR were self-perceived, we categorized them as Subjective. In contrast, the indicators in the WDI were Objective, since they were measurable numbers. Then, we took the Happiness Score as our dependant variable and again ran various models once with Subjective indicators, once with Objective indicators and once with a combination of both. The results are as follows:



Predictors	Adjusted R-squared
WDI	0.54
HIR	0.68
Combined	0.76

‘Predictors of happiness’

Details of these models:

1. Objective Predictors (From World Development indicators dataset)

```
##
## Call:
## lm(formula = Score ~ NY.GDP.PCAP.CD + SL.UEM.TOTL.ZS + IC.TAX.TOTL.CP.ZS,
##     data = data_whole)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2718 -0.4542  0.1280  0.5687  1.7054
##
## Coefficients:
##                      Estimate   Std. Error t value            Pr(>|t|)
## (Intercept)        4.622817186  0.233135357  19.829 <0.0000000000000002 ***
## NY.GDP.PCAP.CD     0.000038446  0.000003153  12.193 <0.0000000000000002 ***
## SL.UEM.TOTL.ZS    -0.010792205  0.013291297  -0.812               0.418
## IC.TAX.TOTL.CP.ZS  0.006554209  0.004490923   1.459               0.147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7764 on 129 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.5502, Adjusted R-squared:  0.5398
## F-statistic: 52.61 on 3 and 129 DF,  p-value: < 0.00000000000000022
```

2. Subjective Predictors (From Happiness Index dataset)

```
##
## Call:
## lm(formula = Score ~ as.numeric(Perceptions.of.corruption) +
##     Freedom.to.make.life.choices + Generosity + Social.support,
##     data = data_whole)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1226 -0.4032  0.1388  0.4465  1.3543
##
## Coefficients:
##                                       Estimate Std. Error t value
## (Intercept)                             1.5696     0.2574   6.098
## as.numeric(Perceptions.of.corruption)   1.9610     0.6622   2.961
## Freedom.to.make.life.choices            1.5314     0.4265   3.591
## Generosity                             -0.1737     0.6066  -0.286
## Social.support                          2.4007     0.2066  11.622
##                                                   Pr(>|t|)
## (Intercept)                                   0.0000000113 ***
## as.numeric(Perceptions.of.corruption)             0.003639 **
## Freedom.to.make.life.choices                      0.000465 ***
## Generosity                                        0.775055
## Social.support                         < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6502 on 131 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6854, Adjusted R-squared:  0.6758
## F-statistic: 71.36 on 4 and 131 DF,  p-value: < 0.00000000000000022
```

3. Combination of objective and subjective predictors

```
## 
## Call:
## lm(formula = Score ~ as.numeric(Perceptions.of.corruption) + 
##     NY.GDP.PCAP.CD + SL.UEM.TOTL.ZS + IC.TAX.TOTL.CP.ZS + Freedom.to.make.life.choices + 
##     Generosity + Social.support, data = data_whole)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.50618 -0.32125  0.08729  0.32087  1.36976 
## 
## Coefficients:
##                                           Estimate   Std. Error t value
## (Intercept)                             1.965944390  0.340210465   5.779
## as.numeric(Perceptions.of.corruption) -0.461545245  0.693563951  -0.665
## NY.GDP.PCAP.CD                          0.000022363  0.000003339   6.698
## SL.UEM.TOTL.ZS                         -0.010095383  0.011064325  -0.912
## IC.TAX.TOTL.CP.ZS                       0.009141573  0.003372113   2.711
## Freedom.to.make.life.choices            1.396669982  0.398969367   3.501
## Generosity                             -0.367679586  0.565512346  -0.650
## Social.support                          1.849036945  0.209955668   8.807
##                                                   Pr(>|t|)    
## (Intercept)                           0.00000005730560816 ***
## as.numeric(Perceptions.of.corruption)            0.506987    
## NY.GDP.PCAP.CD                        0.00000000065641500 ***
## SL.UEM.TOTL.ZS                                   0.363315    
## IC.TAX.TOTL.CP.ZS                                0.007661 ** 
## Freedom.to.make.life.choices                     0.000646 ***
## Generosity                                       0.516785    
## Social.support                        0.00000000000000945 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.5577 on 124 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.7746, Adjusted R-squared:  0.7619 
## F-statistic: 60.87 on 7 and 124 DF,  p-value: < 0.00000000000000022
```

We can see that when the indicators from World development indicators and Happiness index report are combined, the results get even better. This model explains 76% variability in Happiness score. We can safely say that Happiness is indeed a combination of perceived factors and objective indicators.

Limitations and future work

One of the shortcomings of our project is the lack of data. Many countries don’t provide some data at all, some data wasn’t available in the earlier years, some data is collected only once in a few years and these factors make it hard to implement deep analysis, especially machine learning. Because of the volume of the missing data, imputing missing data wasn’t a possibility either. The other issue is that the data is static. Some of the indicators mentioned in the data have values that change or are updated regularly, but the dataset as a whole is updated quarterly. Although most of the data won’t be affected by dynamic updating, it certainly could be a way to get much better outputs and tracking the progress of countries would be much easier.

Due to the size of the dataset, we were almost certain that we wouldn't be able to cover all of it. There were some domains that we couldn't cover and even within the domains we did, there could be some indicators we missed that could have given revealing results or enhanced our statistical analyses. There is a lot more that can be done with this dataset. People with expertise in specific domains such as Economy, health or gender studies can analyse that part of data substantially better than we did and the results they find would be much more accurate. Not only would they be able to explain the data well, but they would also be able to do a lot in terms of practical suggestions of solutions in the domains.

We intended to show rest of our work (Graphs, tests that didn't give good results, tests where we later realized that the data isn't sufficient) at the end of the report in the appendix, but there is a lot of content in this section, so we decided to not show it in the report. We have echod the code in the code file at the end.

References

1. https://datacatalog.worldbank.org/dataset/world-development-indicators (https://datacatalog.worldbank.org/dataset/world-development-indicators)
2. https://worldhappiness.report/ (https://worldhappiness.report/)
3. https://www.kaggle.com/kmravikumar/how-far-is-china-ahead-of-india (https://www.kaggle.com/kmravikumar/how-far-is-china-ahead-of-india)
4. https://www.kaggle.com/smondal93/exploring-global-inequality-and-growth (https://www.kaggle.com/smondal93/exploring-global-inequality-and-growth)