Suicide Analysis

Mayank Mohan Yadav

10/12/2020

#The World Health Organization reports that 800,000 people die by suicide each year worldwide, while suicide is the 10th leading cause of death in the United States. Suicide is a major global health problem. In addition to the enormous toll that suicide takes on individuals and families, a high suicide rate can be detrimental to the long-run growth of a society, particularly if mostly young people are affected. Let's take a look.

#The data is provided by WHO and The World Bank. #Raw Data:

```
## Parsed with column specification:
## cols(
##   country = col_character(),
##   year = col_double(),
##   sex = col_character(),
##   age = col_character(),
##   suicides_no = col_double(),
##   population = col_double(),
##   ‘suicides/100k pop‘ = col_double(),
##   ‘country-year‘ = col_character(),
##   ‘HDI for year‘ = col_double(),
##   ‘gdp_for_year ($)‘ = col_number(),
##   ‘gdp_per_capita ($)‘ = col_double(),
##   generation = col_character()
## )
```

```
## # A tibble: 3 x 12
##   country  year sex   age   suicides_no population ‘suicides/100k ~
##   <chr>   <dbl> <chr> <chr>       <dbl>      <dbl>            <dbl>
## 1 Albania  1987 male  15-2~          21     312900             6.71
## 2 Albania  1987 male  35-5~          16     308000             5.19
## 3 Albania  1987 fema~ 15-2~          14     289700             4.83
## # ... with 5 more variables: ‘country-year‘ <chr>, ‘HDI for year‘ <dbl>,
## #   ‘gdp_for_year ($)‘ <dbl>, ‘gdp_per_capita ($)‘ <dbl>, generation <chr>
```

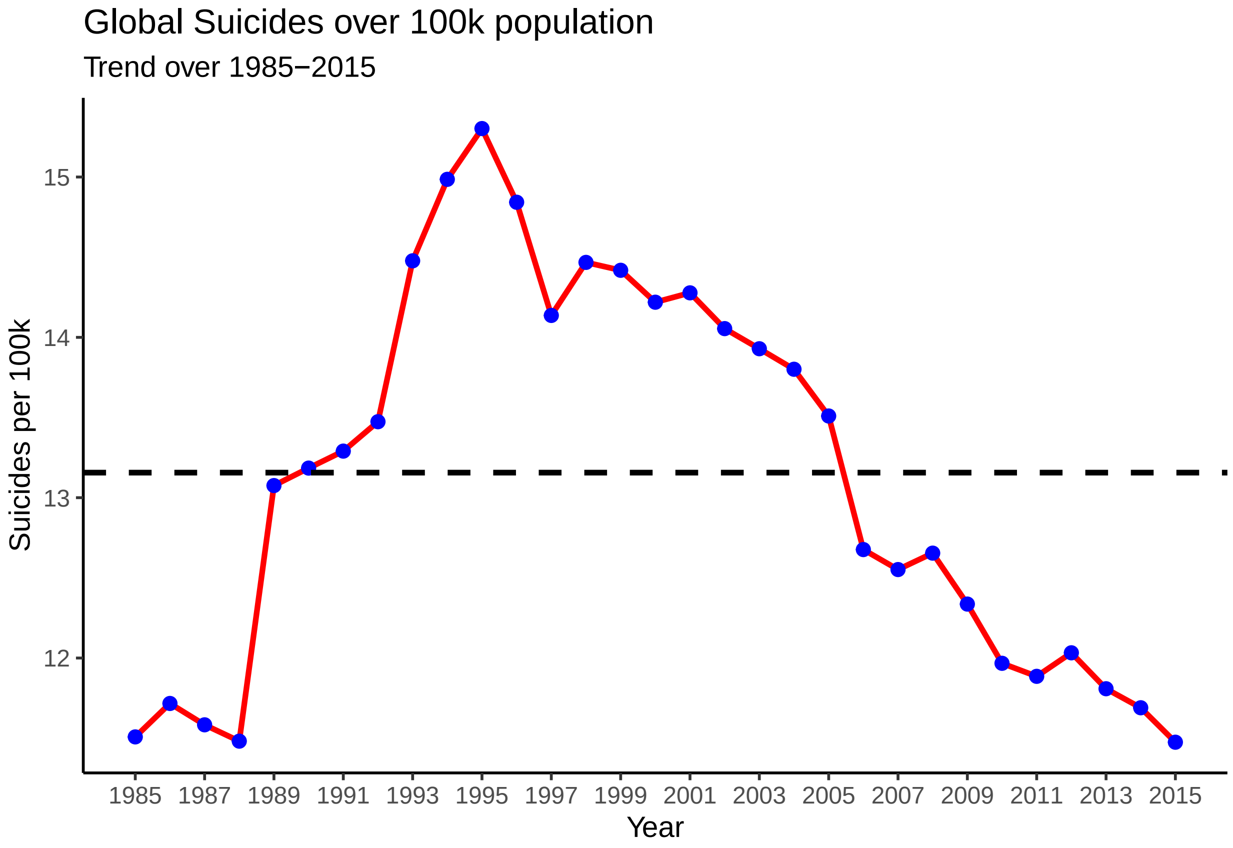
```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

#Cleaned data: removed missing rows and columns, tidied up column names, created new column continents using countrycode function, and removed unusable columns

```
##    country year    sex   age suicides_no population suicide_rate gdp_for_year
## 1  Albania 1987   Male 15-24          21     312900         6.71   2156624900
## 2  Albania 1987   Male 35-54          16     308000         5.19   2156624900
## 3  Albania 1987 Female 15-24          14     289700         4.83   2156624900
## 4  Albania 1987   Male   75+           1      21800         4.59   2156624900
## 5  Albania 1987   Male 25-34           9     274300         3.28   2156624900
## 6  Albania 1987 Female   75+           1      35600         2.81   2156624900
## 7  Albania 1987 Female 35-54           6     278800         2.15   2156624900
## 8  Albania 1987 Female 25-34           4     257200         1.56   2156624900
## 9  Albania 1987   Male 55-74           1     137500         0.73   2156624900
## 10 Albania 1987 Female  5-14           0     311000         0.00   2156624900
##    gdp_per_capita      generation continent
## 1             796    Generation X    Europe
## 2             796          Silent    Europe
## 3             796    Generation X    Europe
## 4             796 G.I. Generation    Europe
## 5             796         Boomers    Europe
## 6             796 G.I. Generation    Europe
## 7             796          Silent    Europe
## 8             796         Boomers    Europe
## 9             796 G.I. Generation    Europe
## 10            796    Generation X    Europe
```

#The overall trend is decreasing which is a good thing but given the limited data collection techniques and tools (lets say pre 1990) cannot really be too trusting of this data. The world saw a peak in the rate of suicides in 1995

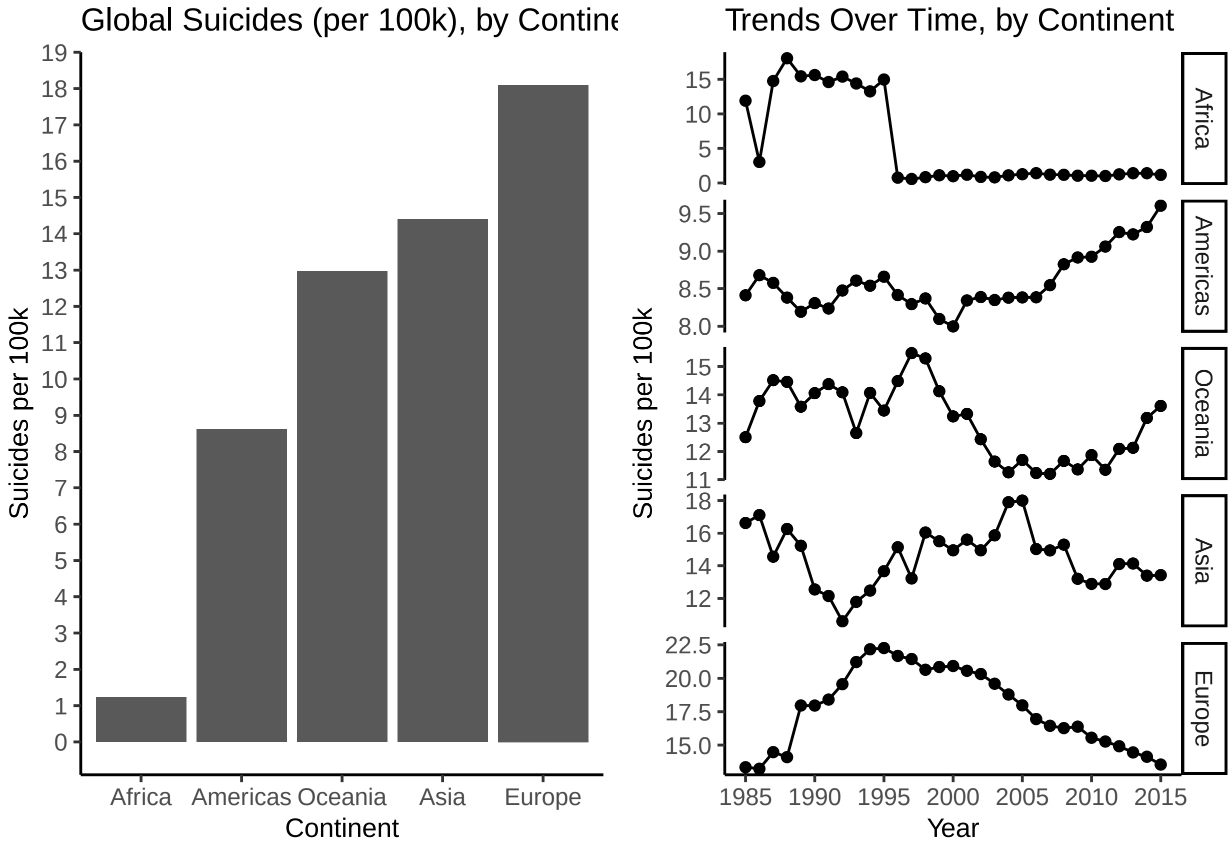
```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```



#Countries in the continent of Africa report a suicide rate of 0 after 1995 which may actually not be the case and it might be that the governments have not properly collected and reported the data to the World Bank and WHO. Trends of continents Americas and Oceania are on the rise which is concerning but even more concerning is the suicide rate of European countries through the years. They are literally on a different scale.

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

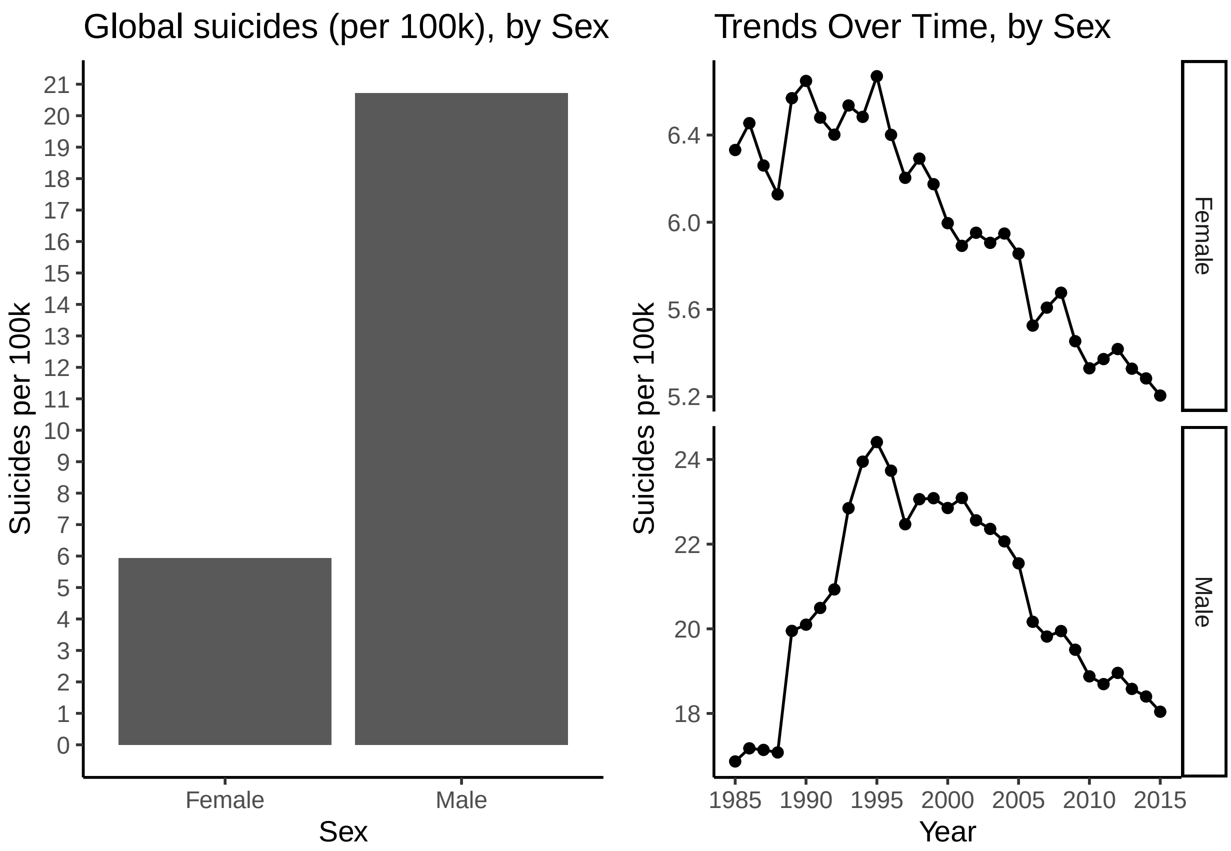
```
## ‘summarise()‘ regrouping output by ’year’ (override with ‘.groups‘ argument)
```



#The gender stereotype of men being ‘tough’ and ‘strong’ does not allow for failure. We see a very clear over-representation of men here. Even with these decreasing trends by the end of 2015 global average of suicides for men is 18/100k population compared to 5.2 for women. A staggering difference.

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

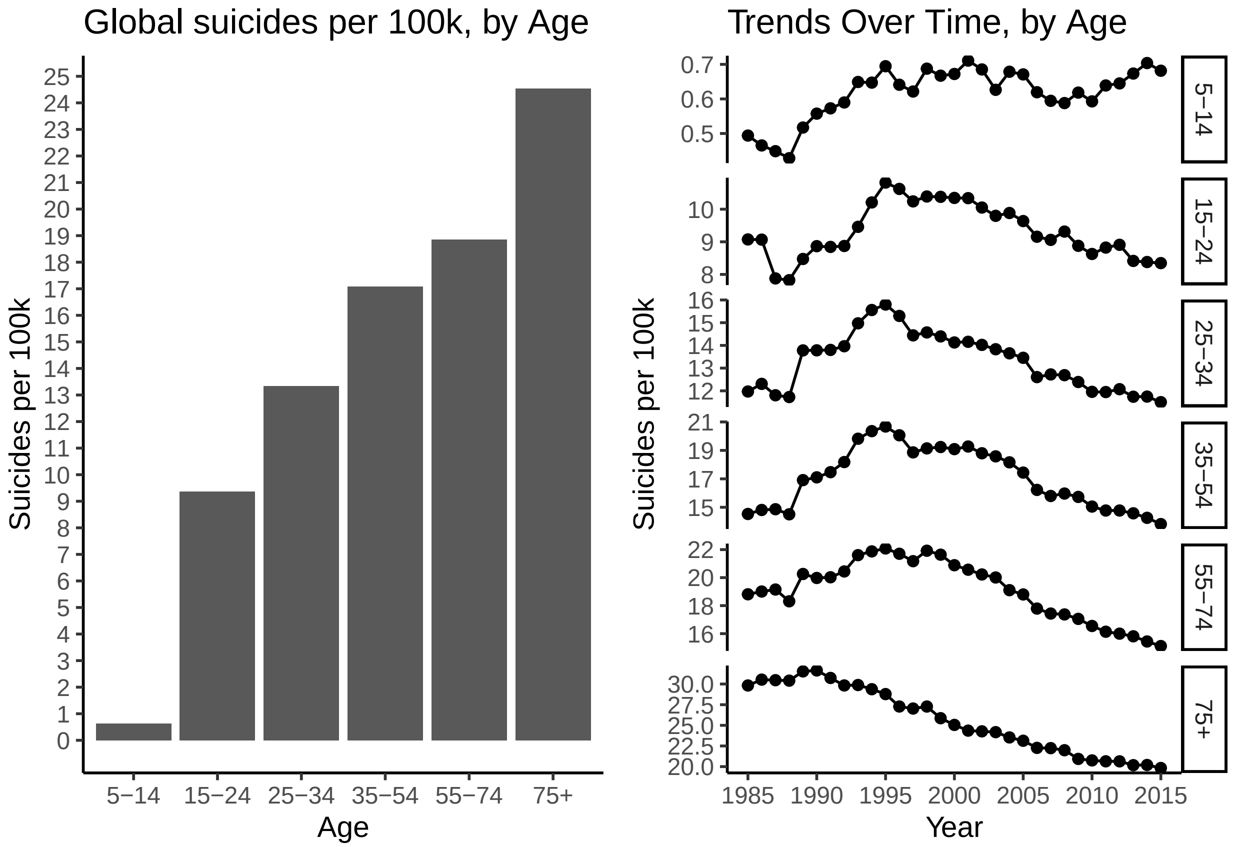
```
## ‘summarise()‘ regrouping output by ’year’ (override with ‘.groups‘ argument)
```



#Through the years we are seeing a decrease in suicidal patterns across all age groups but one thing can be established-as age of a person increases the suicidal tendencies increase

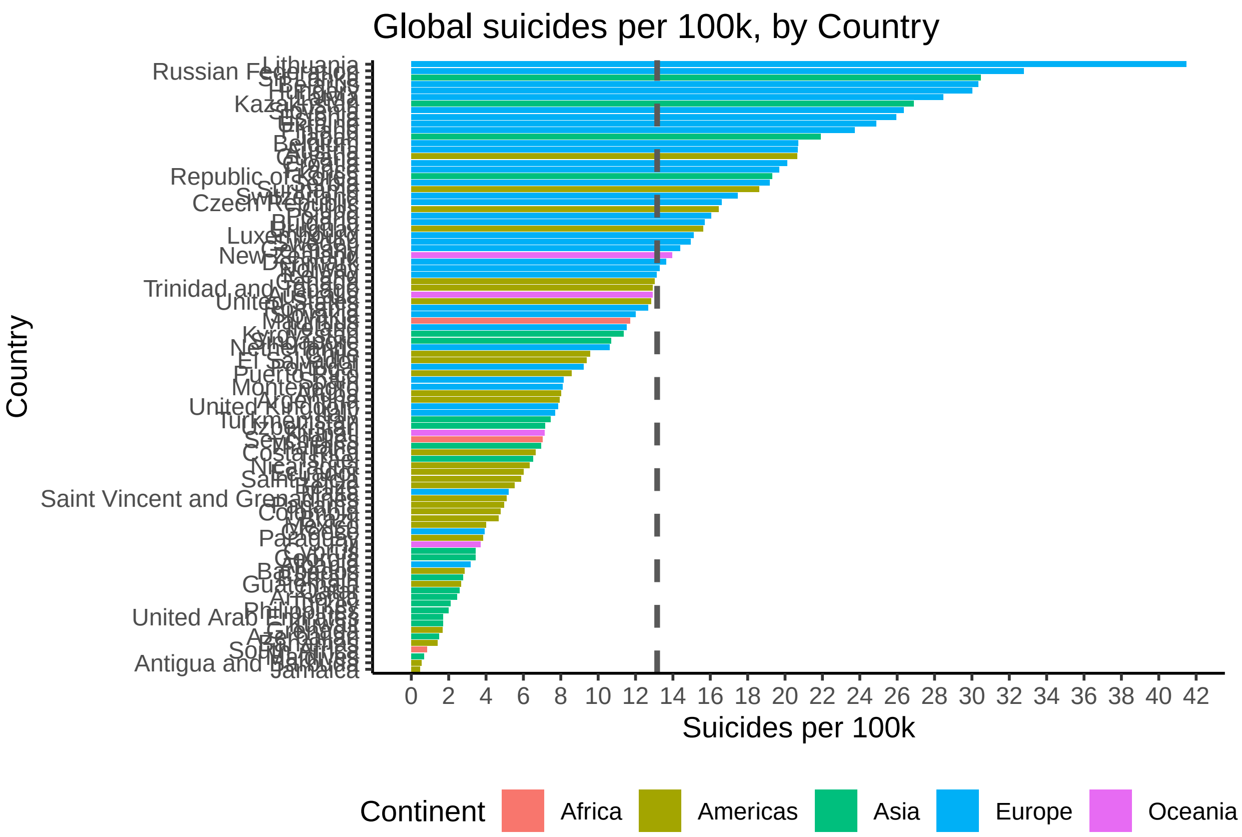
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## `summarise()` regrouping output by 'year' (override with `.groups` argument)
```



#The suicide rate of Lithuania is concerning, but even more concerning is that a large number of European countries make up half of the chart, that too on the higher end. Suicide in Lithuania is such a big concern that it has it's own Wiki page. Sociologists explain it as Lithuania being exposed to new and unfamiliar social environment after the collapse of USSR that started in 1988. Suicide rates of many European countries peak in in the early 1900's and show a steady decline post 1995. The social and economic factors of the post USSR dissolution era must have played a major role in European suicides.

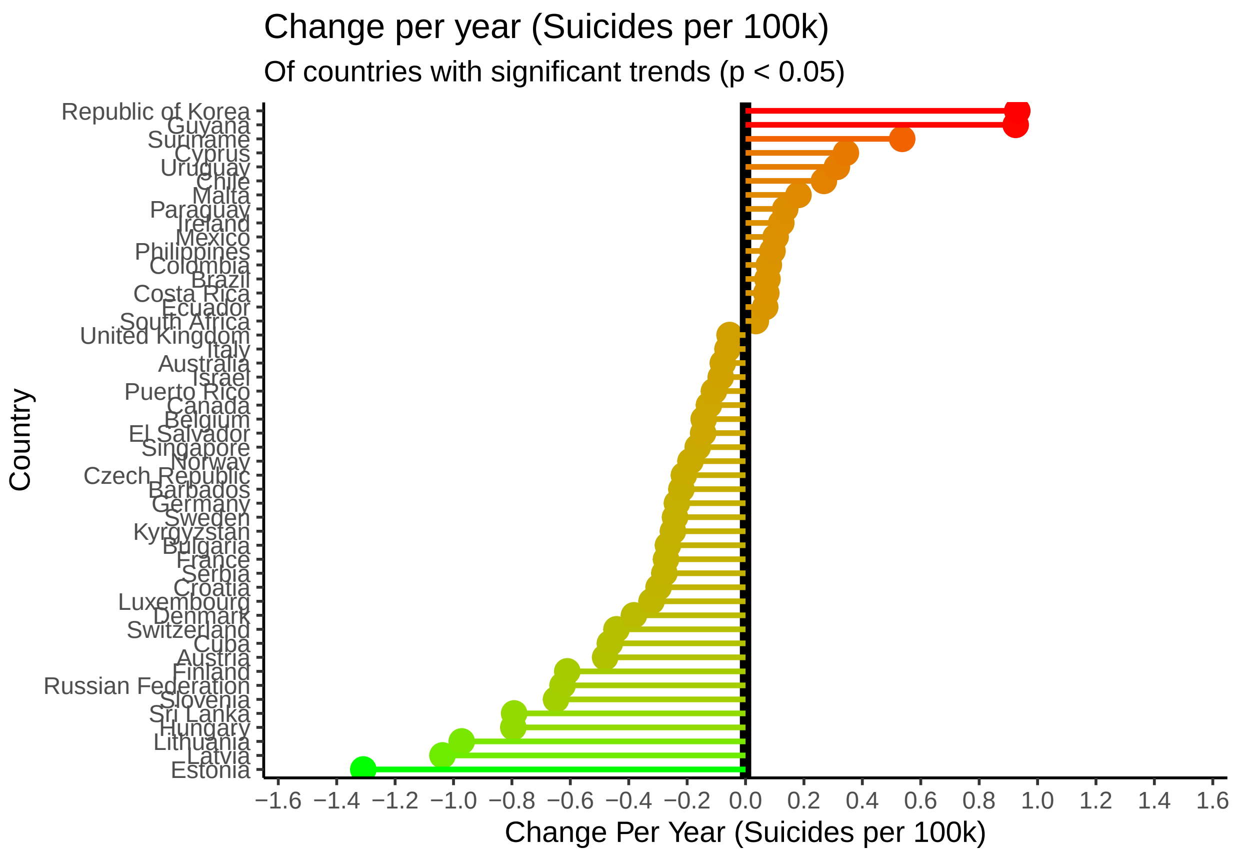
```
## 'summarise()' regrouping output by 'country' (override with '.groups' argument)
```



#I fit a linear model to explain the rate of change in suicide rate per year. #The following is the **rate of change**, not the absolute value of change; hence we don't see Lithuania at the top, instead Lithuania has a decreasing rate of change which is a good thing. On the other hand 16 countries have an increasing rate of change, with Republic of Korea and Guyana observing 1 more suicide every year!

```
## ‘summarise()‘ regrouping output by ’country’ (override with ‘.groups‘ argument)
```

```
## Warning: All elements of ‘...‘ must be named.
## Did you want ‘data = c(year, suicides, population, suicideper100k, gdp_per_capita)‘?
```



#Before we establish any relationship between GDP and suicide rates, we need to establish a correlation between the GDP of a country with years. If in fact the GDP of countries increase with time, only then would there be any meaningful interpretation of GDP and suicide rates. On observing the p-value we note that almost all countries have a very strong positive correlation for year and GDP(per capita)

```
## ‘summarise()‘ regrouping output by ’country’ (override with ‘.groups‘ argument)
```

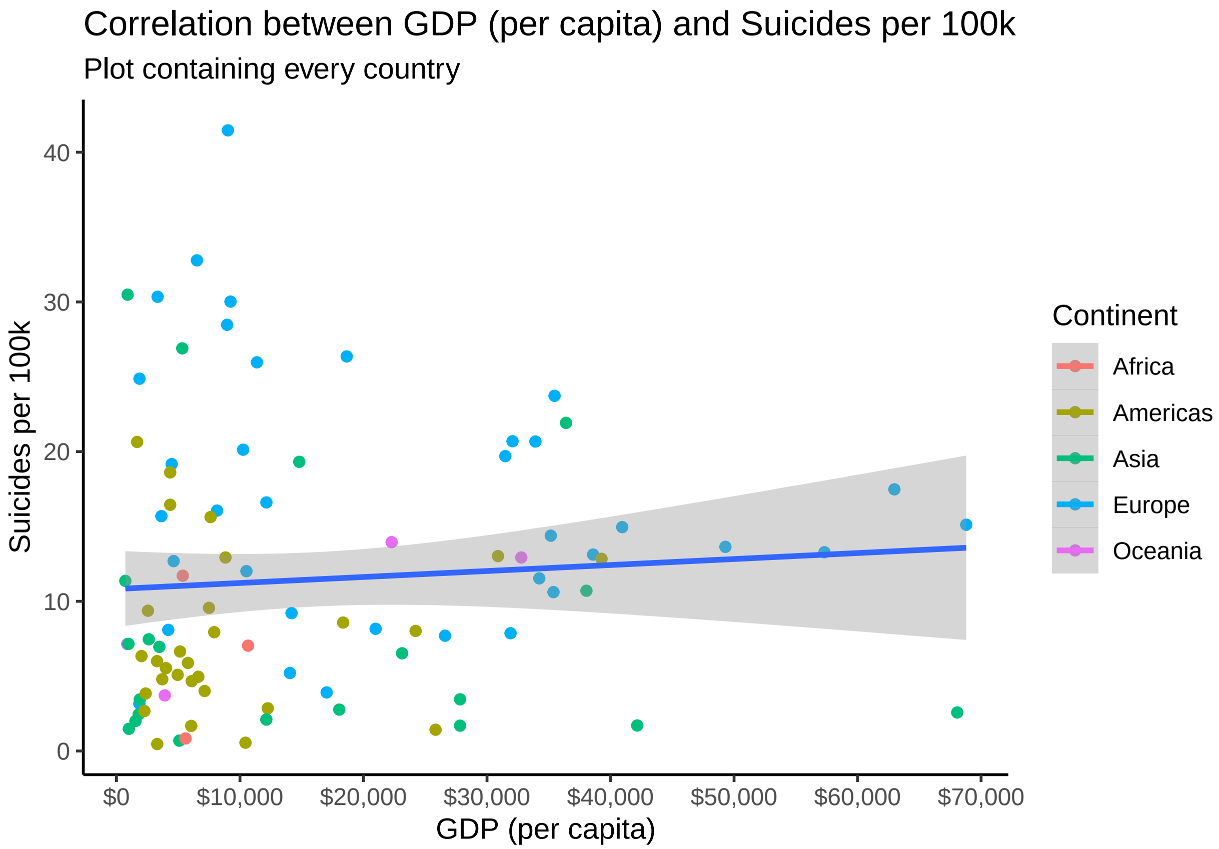
```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
## # A tibble: 93 x 2
##    country             year_gdp_correlation
##    <fct>                              <dbl>
##  1 Albania                            0.882
##  2 Antigua and Barbuda                0.944
##  3 Argentina                          0.738
##  4 Armenia                            0.915
##  5 Aruba                              0.914
##  6 Australia                          0.905
##  7 Austria                            0.943
##  8 Azerbaijan                         0.427
##  9 Bahamas                            0.843
## 10 Bahrain                            0.928
## # ... with 83 more rows
```

#I see a weak positive linear relationship between GDP and suicide_rate; but the regression line does not represent the data which is to be expected because datapoints are scattered all over the place; the data is very skewed towards European suicide rates; nevertheless countries with high GDP might actually be associated with higher suicide_rate.

```
## ‘summarise()‘ regrouping output by ’country’ (override with ‘.groups‘ argument)
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```

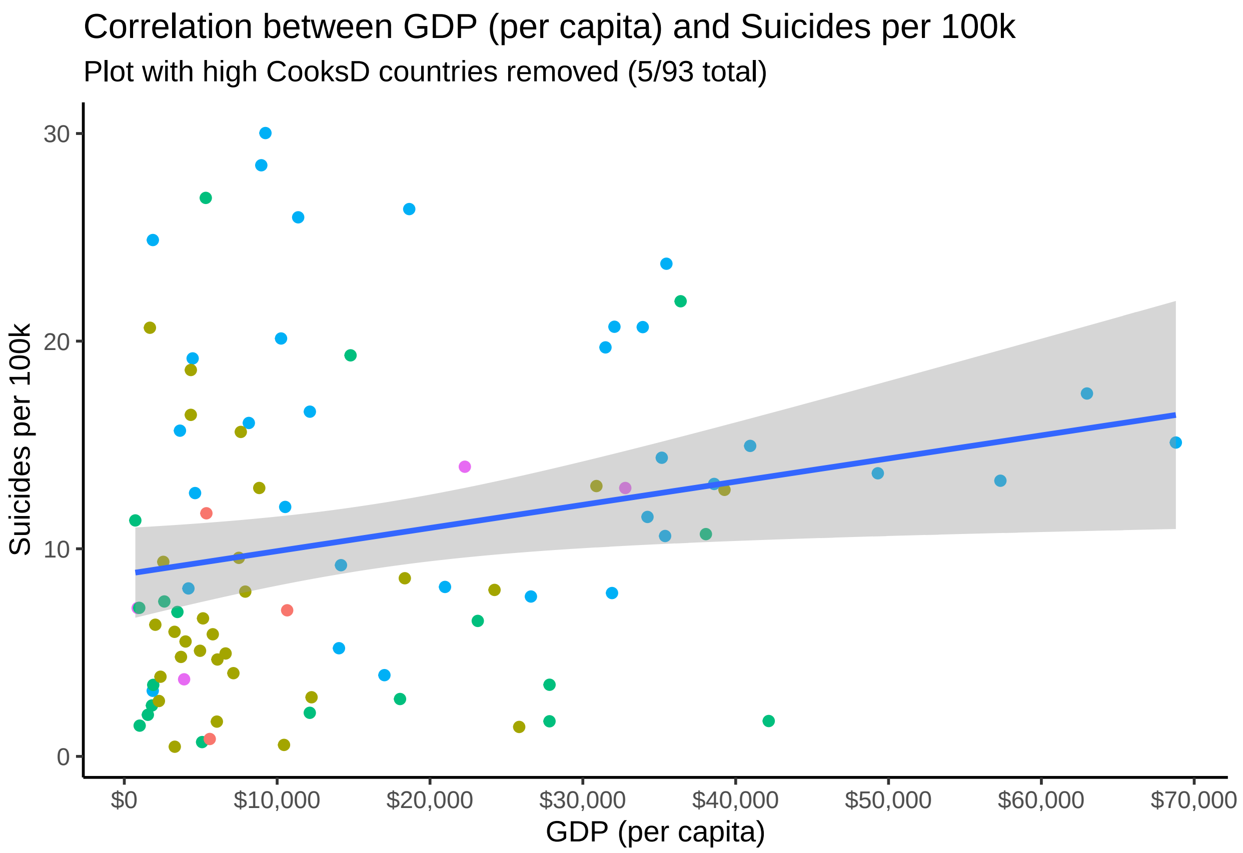


#GDP is not really capturing the variation in suicide_rate; suicide_rate is not VERY depedent on GDP and varies across countries hence cannot say that if a country has higher GDP it will definitely observe fewer suicides

#What we really want is to capture the trend of global suicide rates without the data being so skewed; I am calculating and removing high values of Cooks distance (will help make data less skewed) for suicide rates so that abnormally high values if countries like Lithuania are removed #Although this is an improved attempt at fitting a regression line, it does not make a whole lot of difference like i thought it would, but we do see a better fitted regression line

#One thing that is for certain though is that with GDP increase we do see a positive linear increase in suicide_rate; i personally found it counter-intuitive

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



#https://rstudio-pubs-static.s3.amazonaws.com/484548_f68efb9b21244a4099f34114a2ca7218.html#root-mean-squared-error-rmse #rmse to evaluate the model performance

#https://tidyr.tidyverse.org/reference/gather.html #gather to create key-value pairs for comparision between both models

#RMSE of random forest is better than linear model by double which is understandable given that linear models do not fit well with dataframes having many dimensions as is the case here.

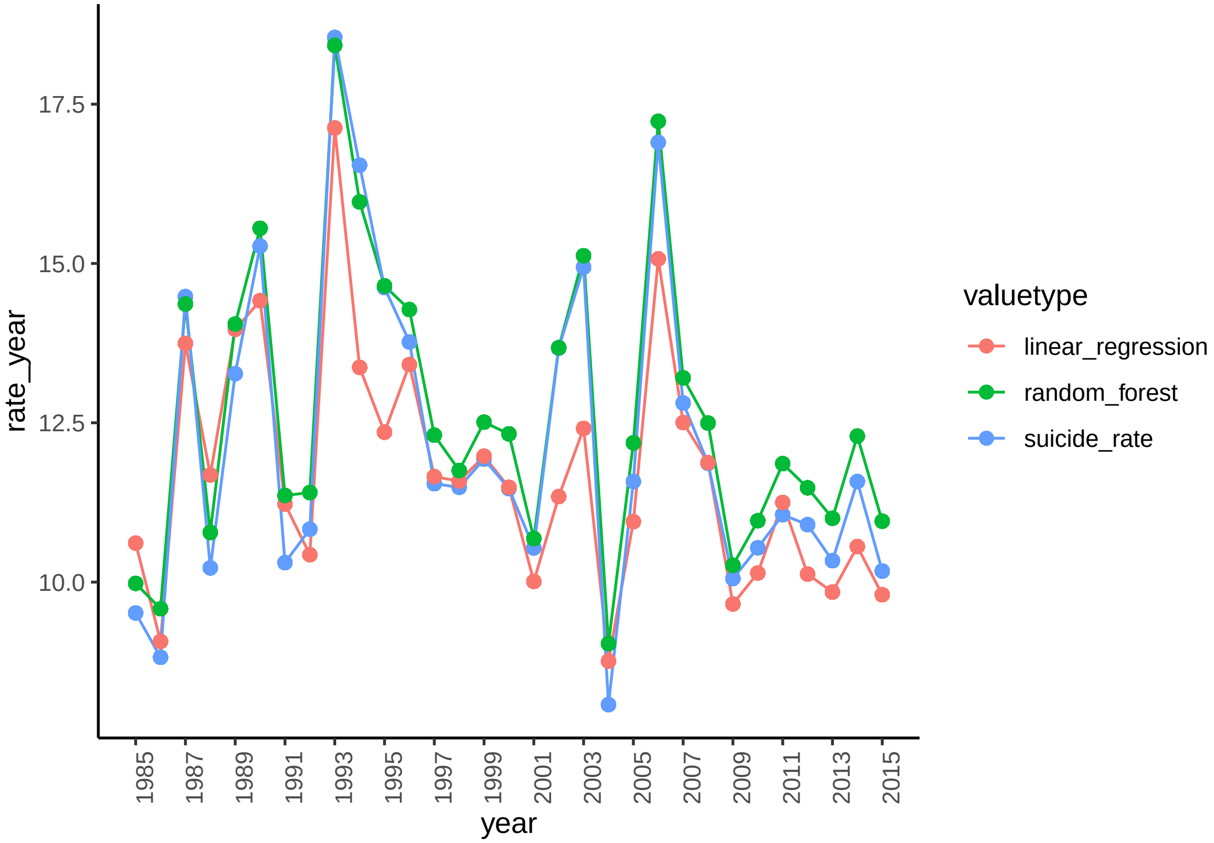
```
## Warning in predict.lm(lm1, newdata = test): prediction from a rank-deficient fit
## may be misleading
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
## # A tibble: 2 x 2
##   model              rmse
##   <chr>             <dbl>
## 1 linear_regression 12.2
## 2 random_forest      6.80
```

#Random forest (predicted values grouped by year) are very closely aligned with the test data predicted values grouped by year and hence is better(and possibly the best)than a linear model

#I chose RF because i did not have to do any feature scaling (no standardization or normalization as such except the log transformation); it automatically handles missing data and is less impacted by noise; even if we treat abnormal values of suicide_rate for European countries as outliers the prediction is very good



#This RF model can also be made sensitive to dimensions like unemployement rate or number of jobs created per year/month in a country and we could potentially very accurately predict what effect changes in such dimensions would have on suicide rates.