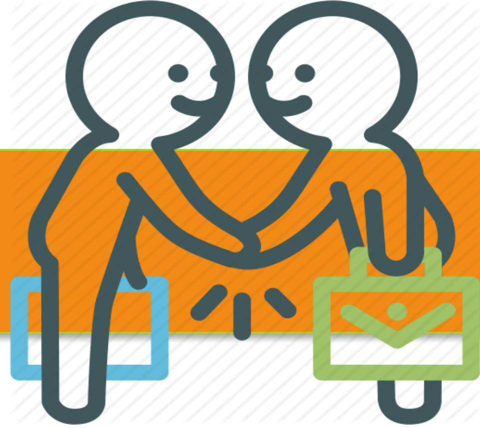
Forecasting Suicides using Historical Data

ALY6015- Second Quarter, Term A, Dr. Matthew Goodwin
Analysis Report of Final Capstone Project

Assignment Completion Date: 02/16/19

Manvinder Kaur [NUID: 001402192]
Ai Uyen Thi Huynh [NUID: 001406618]

Silicon Valley, Northeastern University, CA.



Introducing Dataset:
Suicide Rates Over 1985 to 2016

The Dataset contains number of suicides in different countries over 1985 to 2016. It covers data for 101 countries in the form of categorical fields and continuous numerical fields

<u>Categorical Data :</u>

- Age
- Gender
- Generation

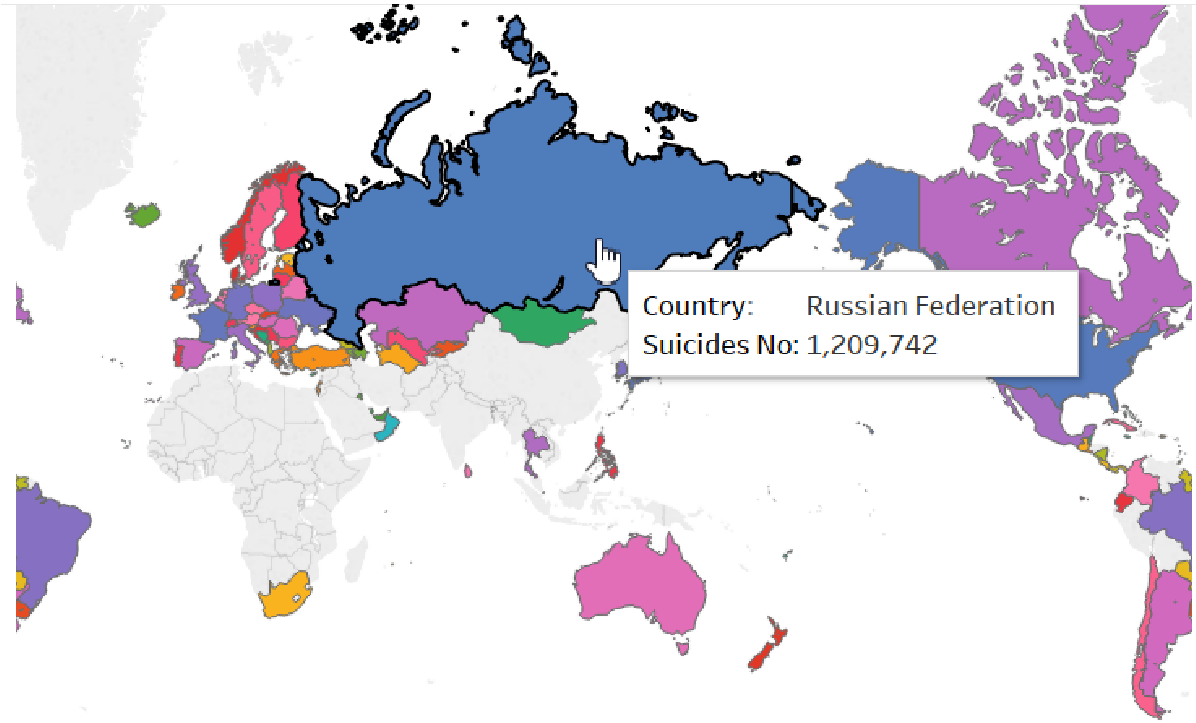
<u>Numerical Data :</u>

- Number of suicide
- Suicide Rate
- Population
- GDP for year
- GDP per capita

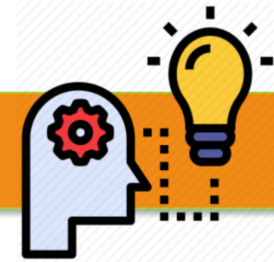
Research Questions

- Forecasting the number of suicides in year 2019.
- Country with maximum number of suicides in 2019.

– *Currently its Russia*



Visualization done in Tableau

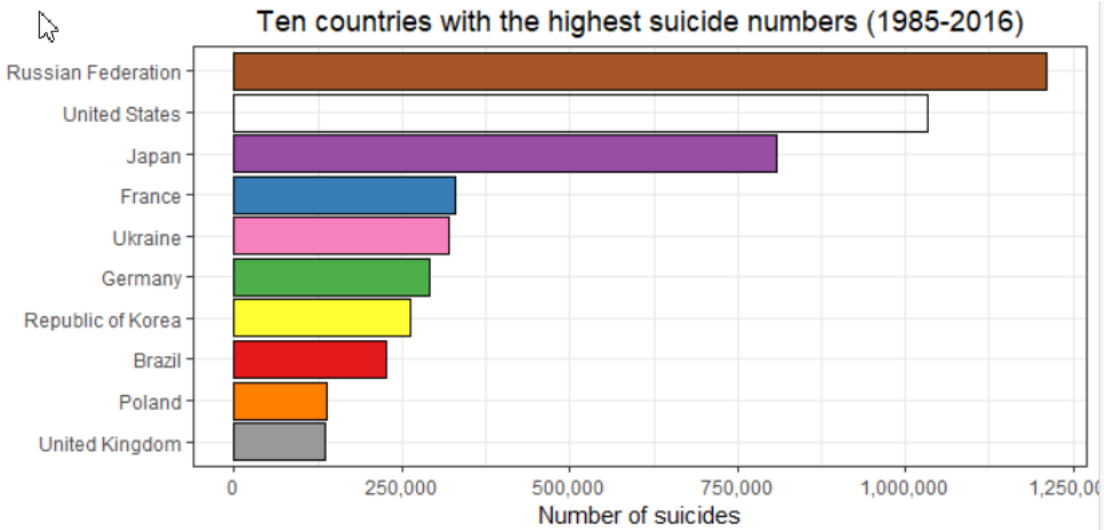


Model Implemented

Methods/Models tried and implemented

- ✓: **We** developed the R-code to find top 10 countries with highest suicide over the period of time.
- ✓: We **answered** questions like “ The female commit more suicide or the males and visualize it for all the age groups“ using the graphical representations.
- ✓: **We** conducted an experiment to finding the suicide rate and number of suicides among different age groups.
- ✓: We **worked** on finding the correlation of suicide with all the “dimensions“
- ✓: “We answered question by preforming “**Time series using ARIMA prediction model**“ and preforming “**Linear regression model**” for better results.

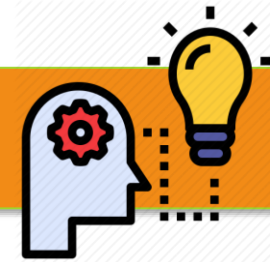
Top Ten countries with highest suicide number



Inference:

We can see from this graph that Russia has the highest number of deaths by suicide, following by United States and Japan – the top three countries dominate the list, accounting for 64% of the total number of the top 10.

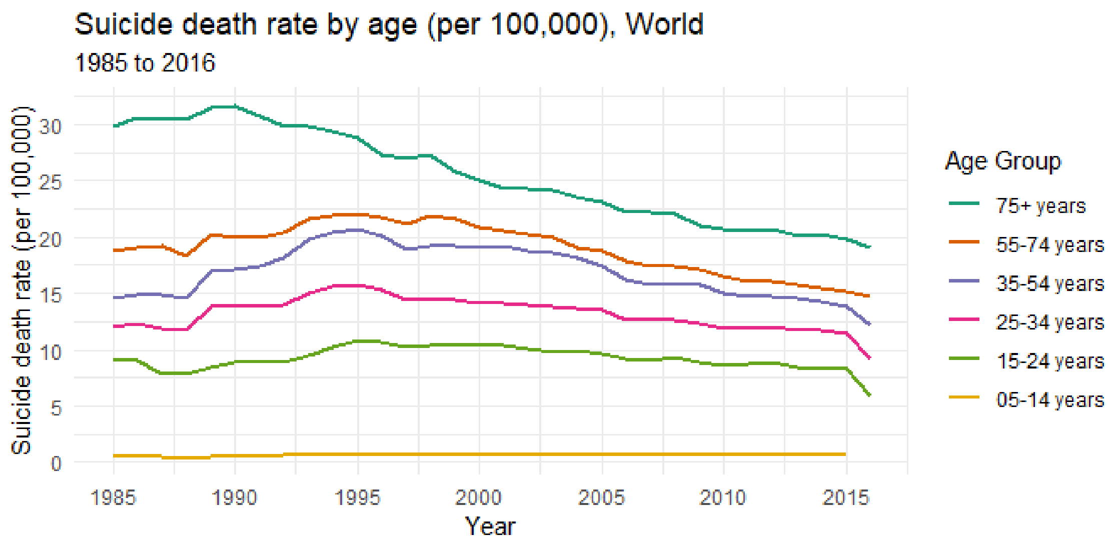
```
# plot the 10 countries with the highest suicide numbers
library(dplyr)
country_group <- group_by(mydata, country)
mydata_by_country <- summarize(country_group,
                               sum_suicide = sum(suicides_no))
mydata_by_country <- arrange(mydata_by_country, desc(sum_suicide))
top_10 <- head(mydata_by_country, 10)
ggplot(data=top_10, aes(x=reorder(country, sum_suicide), y=sum_suicide)) +
  geom_bar(colour="black", stat = "identity", aes(fill=country)) +
  coord_flip() + guides(fill=FALSE) +
  scale_fill_brewer(palette="Set1") +
  scale_y_continuous(labels = scales::comma) +
  ggtitle("Ten countries with the highest suicide numbers (1985-2016)") +
  theme_bw() + theme(plot.title = element_text(hjust=0.5)) +
  xlab('') + ylab("Number of suicides")
```



Top Ten countries with highest suicide number

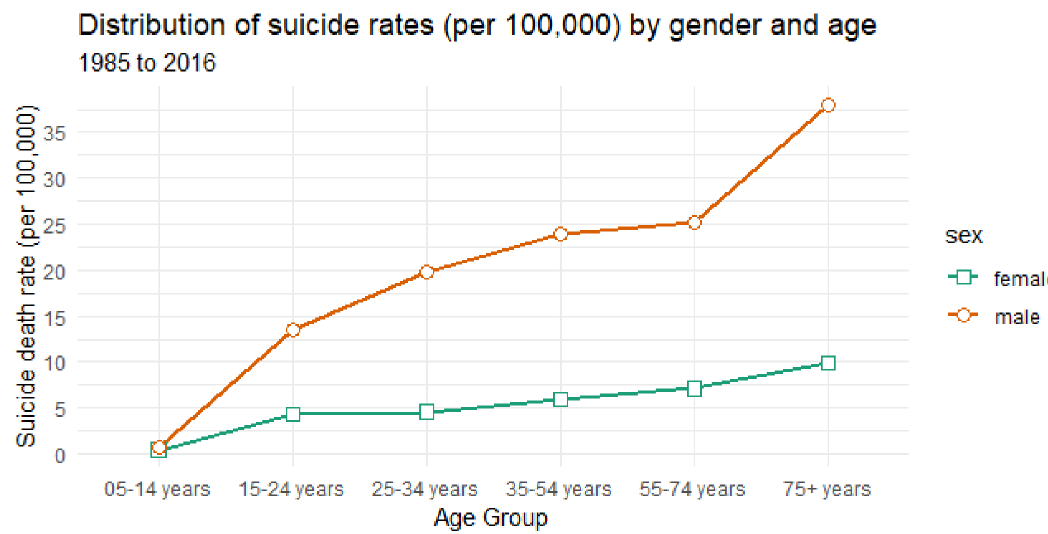
Inference:

There is an overall decline in suicide rates (measured per 100,000 people) across various age categories. The oldest age group had decreasing rates since 1990, while the younger ones saw the decreasing trend 10 years later, since 2000.




```
# Plot the suicide death number by age
age_group <- group_by(mydata, age, year)
mydata_by_age <- summarize(age_group,
                               sum_suicide = sum(suicides_no))
ggplot(aes(x=year, y=sum_suicide/1000, fill=forcats::fct_rev(age)),
       data = mydata_by_age) +
  geom_area(colour="black", size=.2, alpha=.8) +
        theme_bw() +
        scale_x_continuous(breaks=seq(1985,2015,5)) +
        scale_y_continuous(breaks=seq(0,300,50)) +
  labs(title = "Suicide deaths number by age, World",
       subtitle = "1985 to 2016",
       x = "Year",
       y = "Number of suicide deaths in Thousands",
       fill = "Age Group") +
  scale_fill_brewer(palette = "Set2") +
  theme_minimal()
```

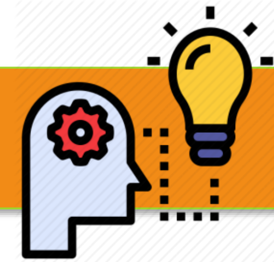
Top Ten countries with highest suicide number



Inference:

The gap between male and female rate for suicide is also different and not constant by age. There is almost no difference in the youngest group (05-14 years old), then the difference is getting bigger when the age increases, and it is biggest in the eldest group (> 75 years old) and middle-age group (55-74 years old

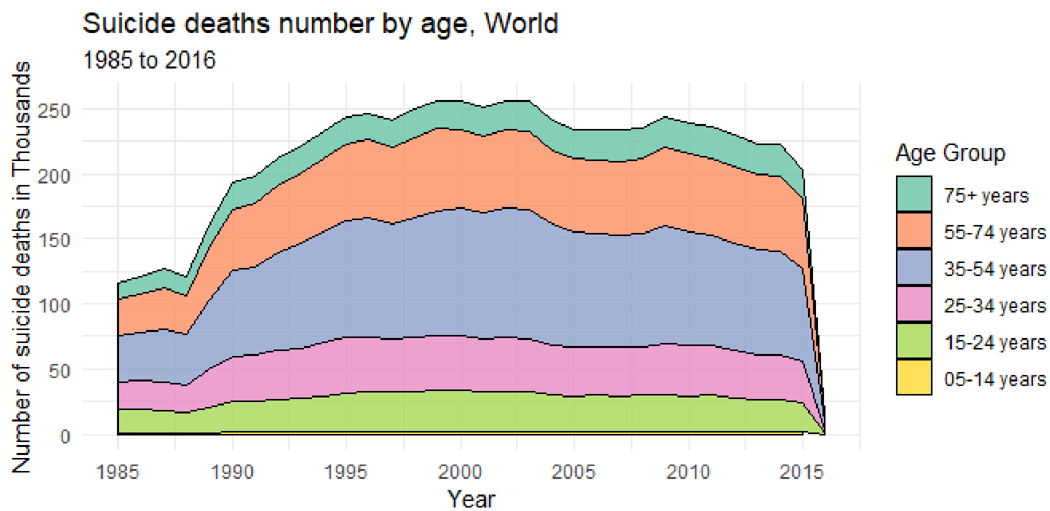
```
# Distribution of suicide rates (per 100,000) by gender and age
mydata %>% group_by(sex, age) %>% summarize(rate=mean(suicides_per100k)) %>%
  ggplot(aes(age, rate, group=sex, color=sex, shape=sex))+
  geom_line(size=.8) +
  geom_point(size=3, fill="white") +
  scale_shape_manual(values=c(22,21)) +
  scale_y_continuous(breaks=seq(0,40,5)) +
  labs(title = "Distribution of suicide rates (per 100,000) by gender and age",
       subtitle = "1985 to 2016",
       x = "Age Group", y = "Suicide death rate (per 100,000)") +
    theme_minimal() +
    scale_color_brewer(palette = "Dark2")
```



Top Ten countries with highest suicide number

Inference:

The 35-54 year old group takes up the largest share of number of suicide deaths, roughly 36.3% of deaths. The younger groups, aged 15-34, are at lower risks of taking their own lives, with 28.6%. As age increases (55 and above), the number of suicides decrease.

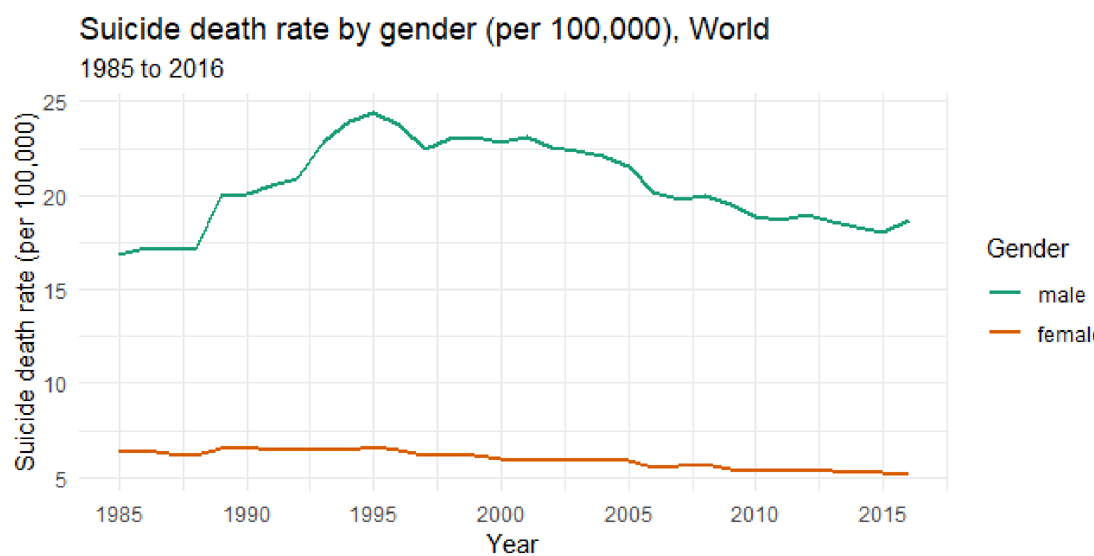



```
# Plot the suicide death rate by age (per 100,000)
mydata %>% group_by(year, age) %>% summarize(s = sum(suicides_no),
                                              p = sum(population)) %>%
  ggplot(aes(year, (s/p)*100000, color=forcats::fct_rev(age))) +
  geom_line(size=1) +
  scale_x_continuous(breaks=seq(1985,2015,5)) +
  scale_y_continuous(breaks=seq(0,40,5)) +
  labs(title = "Suicide death rate by age (per 100,000), World",
       subtitle = "1985 to 2016",
       x = "Year", y = "Suicide death rate (per 100,000)", color="Age Group") +
  theme_minimal() +
  scale_color_brewer(palette = "Dark2")
```

Top Ten countries with highest suicide number

Inference:

There is a big gap in suicide rate between male and female. We can see that male suicide rate is more than 3 times higher than rate for female. Both rates for male and female are seeing the declining trends over years.

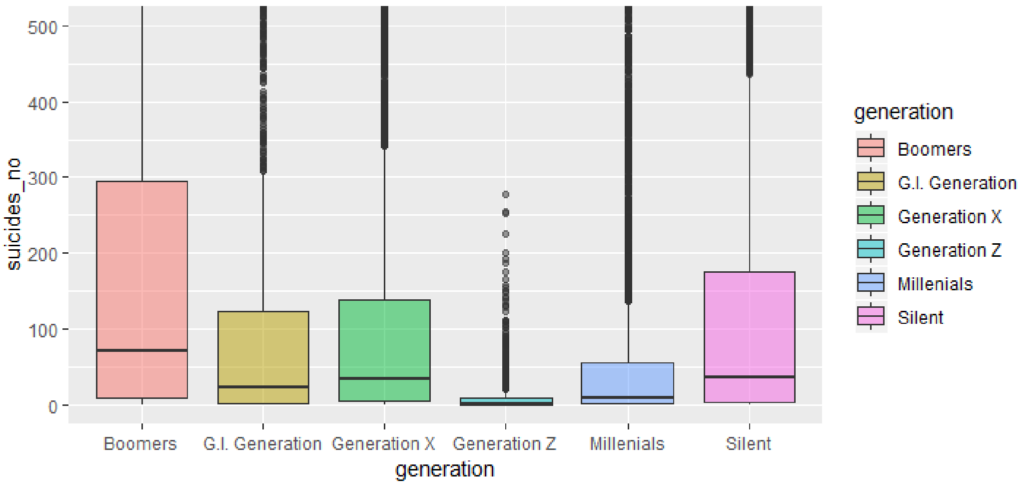



```
# Plot suicide rates (per 100,000) by gender
mydata %>% group_by(year, sex) %>% summarize(s = sum(suicides_no),
                                             p = sum(population)) %>%
  ggplot(aes(year, (s/p)*100000, color=forcats::fct_rev(sex))) +
  geom_line(size=1) +
  scale_x_continuous(breaks=seq(1985,2015,5)) +
  scale_y_continuous(breaks=seq(0,40,5)) +
  labs(title = "Suicide death rate by gender (per 100,000), World",
       subtitle = "1985 to 2016",
       x = "Year", y = "Suicide death rate (per 100,000)", color="Gender") +
  theme_minimal() +
  scale_color_brewer(palette = "Dark2")
```

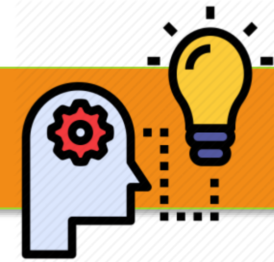
Top Ten countries with highest suicide number

Inference:

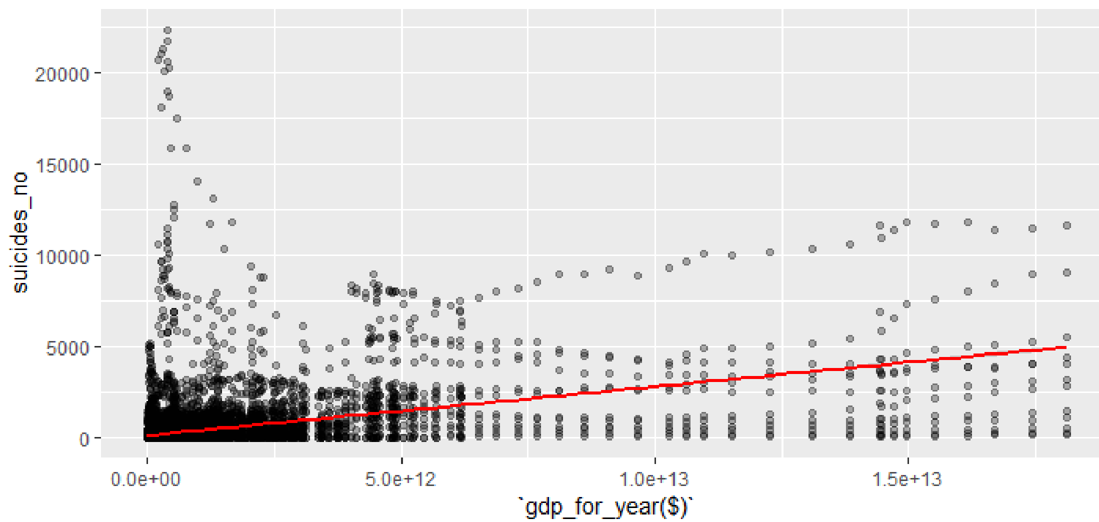
Looking at the box plot, we see that baby boomer generation has the highest rate of suicide (34%) compare to 26% of silent generation and 23% of generation X. Generation Z holds the lowest share, with only 0.24%.




```
#Number of suicides categorised by different Generation
mydata %>%
  ggplot(aes(x = generation , y = suicides_no , fill = generation))+
  geom_boxplot(alpha = .50)+
  coord_cartesian(ylim = c(0,500))
```



Top Ten countries with highest suicide number



Inference:

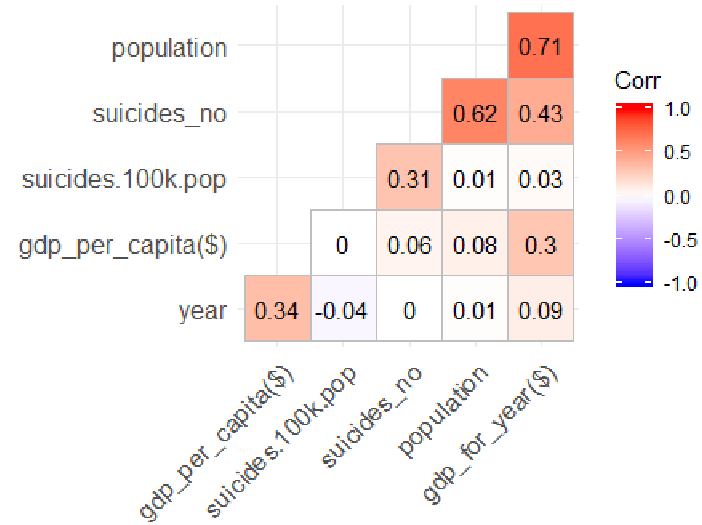
Looking at the graph we see a positive correlation between number of suicides and GDP for year, effect size on the correlation is -0.4335046. It's considering to be small effect size. Therefore, we can firmly say that there is slightly positive correlation and yet significant between suicides_no and gdp_for_year.

```
#Positive correlation between suicides_no and gdp_for_year
#install.packages("ggplot2")
library(ggplot2)
mydata %>%
  ggplot(aes(x = `gdp_for_year($)` , y = suicides_no))+
  geom_jitter(alpha = .30)+
  geom_smooth(method = 'lm' ,color = "red")
```

Top Ten countries with highest suicide number

Inference:

We want to understand if there is a correlation between suicide and GDP of a country. However, we found a weak positive association between suicide number and GDP for year, 0.43, which is very close to 0, so we cannot say anything about this relation. Other variables, such as population or year, do not reasonably associated with the decrease or increase in rate of suicide.



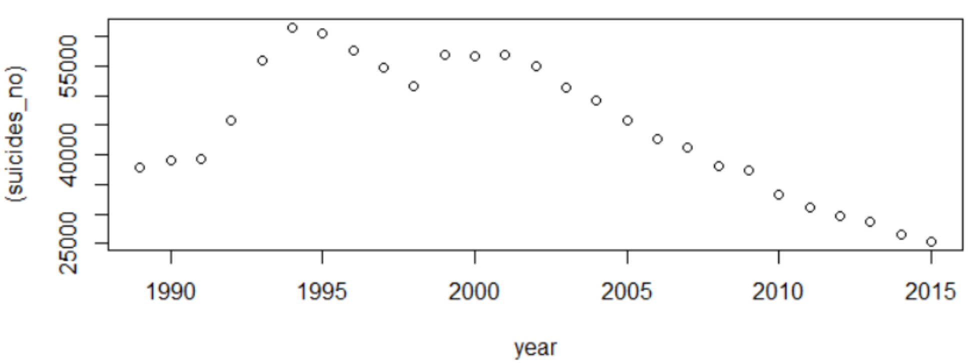

```
######## Correlation
library(dplyr)
# select numeric variables
df <- dplyr::select_if(mydata, is.numeric)
# calulate the correlations
r <- cor(df, use="complete.obs")
round(r,2)
#visualize correlation
library(ggcorrplot)
ggcorrplot(r, hc.order = TRUE, type = "lower", lab = TRUE)
#linear regression
mydata_lm <- lm(suicides_per100k ~ gdp_for_year+ GDP_per_capita, data = mydata)
mydata_lm
library(visreg)
visreg(mydata_lm, "gdp_for_year", gg = TRUE)
visreg(mydata_lm, "GDP_per_capita", gg = TRUE)
```

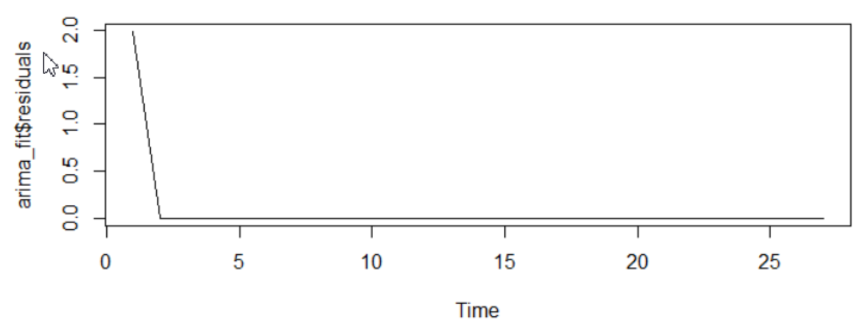
Results- Time series ARIMA

R-code and console out-put: The model is run on the Country with highest suicides i.e. Russian Federation

```r
#Filtering the data required for runing the model
mydata <- mydata %>% filter(mydata$country == "Russian Federation")
tsdata <- subset(mydata,select=c('year','suicides_no'))
tsdata <- as.ts(tsdata)
class(tsdata)
byyear <- aggregate((suicides_no)~year,
                   data=tsdata,FUN=sum)
head(byyear)

#cor(x, y = NULL, use = "everything", method ="pearson")
cor(byyear, method ="pearson")

plot(byyear)
# Smoothing the data, considering the trend from year 2000 to year 2015
adenoTS = ts(byyear)
arima_fit = auto.arima(adenoTS[,1])

arima_fit = auto.arima(adenoTS[,1], trace = TRUE)
plot.ts(adenoTS[,2])
plot.ts(arima_fit$residuals)


#validate the model
Box.test(adenoTS[,2], lag = 5, type = "Ljung-Box")
Box.test(adenoTS[,2], lag = 10, type = "Ljung-Box")
Box.test(adenoTS[,2], lag = 15, type = "Ljung-Box")
```

```
> cor(byyear, method ="pearson")
                   year (suicides_no)
year          1.0000000    -0.6340826
(suicides_no) -0.6340826    1.0000000
> plot(byyear)
> # Smoothing the data, considering the trend from year 2000 to year 2015
> adenoTS = ts(byyear)
> arima_fit = auto.arima(adenoTS[,1])
> arima_fit = auto.arima(adenoTS[,1], trace = TRUE)
> plot.ts(adenoTS[,2])
> plot.ts(arima_fit$residuals)
> #validate the model
> Box.test(adenoTS[,2], lag = 5, type = "Ljung-Box")

        Box-Ljung test

data:  adenoTS[, 2]
X-squared = 59.084, df = 5, p-value = 1.879e-11

> Box.test(adenoTS[,2], lag = 10, type = "Ljung-Box")

        Box-Ljung test

data:  adenoTS[, 2]
X-squared = 65.347, df = 10, p-value = 3.479e-10
```



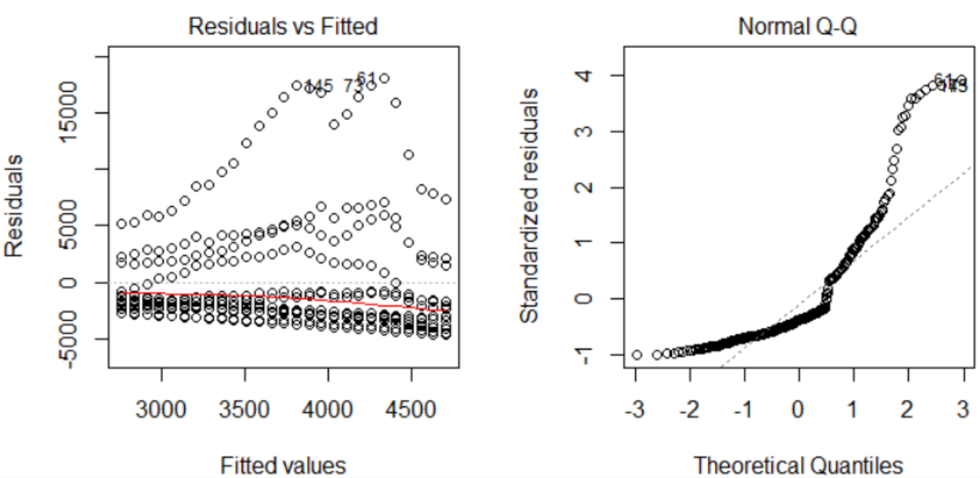


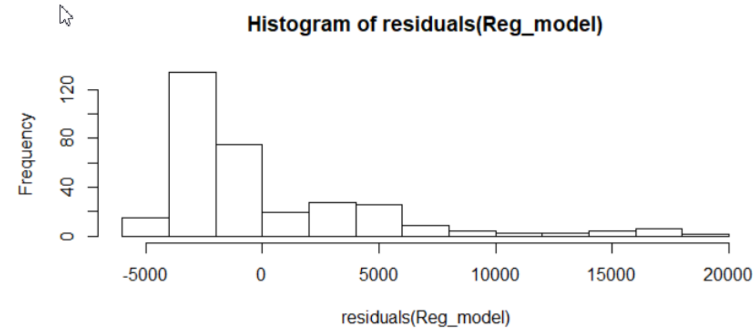
Results- Linear Regression

R-code and console out-put

```
# R Linear Regression
#X <- subset(mydata,select=c('suicides_no'))
Y <- subset(mydata,select=c('year','suicides_no'))
cor(Y)
plot(Y)

# Fit our regression model
Reg_model <- lm(suicides_no ~ year, # regression formula
              data=mydata) # data set
# Summarize and print the results
summary(Reg_model) # show regression coefficients table

confint(Reg_model)

hist(residuals(Reg_model))

par(mar = c(4, 4, 2, 2), mfrow = c(1, 2)) #optional
plot(Reg_model, which = c(1, 2)) # "which" argument optional
```

```
> # R Linear Regression
> #X <- subset(mydata,select=c('suicides_no'))
> Y <- subset(mydata,select=c('year','suicides_no'))
> cor(Y)
                  year suicides_no
year         1.0000000  -0.1268606
suicides_no -0.1268606   1.0000000
> summary(Reg_model) # show regression coefficients table

Call:
lm(formula = suicides_no ~ year, data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
 -4661  -2918  -1731   1926  18004

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 153938.98   65450.15   2.352   0.0193 *
year           -75.03      32.69  -2.295   0.0224 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4583 on 322 degrees of freedom
Multiple R-squared:  0.01609,	Adjusted R-squared:  0.01304
F-statistic: 5.267 on 1 and 322 DF,  p-value: 0.02238

> confint(Reg_model)
                 2.5 %       97.5 %
(Intercept) 25175.0705 282702.88010
year         -139.3447    -10.71043
```





Results

Time series model:
ARIMA Model include
“Running the model,
validating and
predicting the model”

Not ideal: The data set don’t have enough frequency data

Regression model:
Linear regression Model include “Running the model, validating and predicting the model”

Ideal: The data set has the linear decreasing trend

The suicide number has started to decrease from year 2000 and it is projected that the number will further fall. As the correlation coefficient in negative for number of suicides.

Reference Page

- Kan Nishida (2016). Filtering Data with dplyr. Medium blog. Retrieved from https://blog.exploratory.io/filter-data-with-dplyr-76cf5f1a258e
- Lindsay Lee, Max Roser and Esteban Ortiz-Ospina (2016). Suicide. Our World in Data. Retrieved from https://ourworldindata.org/suicide
- RDocumentation. Correlation, Variance and Covariance (Matrices). Retrieved from https://www.rdocumentation.org/packages/stats/versions/3.5.2/topics/cor
- World Bank Dataset. Suicide Rates Overview 1985 to 2016. Kaggle. Retrieved from https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

The End