Suicide Rate across the world

Team member: 1 person

Yang Lu

Suicide rate across the world from 1985 to 2016

##1.Introduction

Suicide, it is a disaster that happen on people all over the world. Some people may have very bad memories about the suicide because some of the famous people died because of that. WIth the development of technology, we want to focus more on the mental health and reduce the problems of suicide in our society.

In order to investigate on the suicide and help reduce this kind of tragedy from happening, we need to know the statistics for that and know how many people in each country are suffering from that every year. Because one’s suicide is related with their entire family, mitigating suicide rate can have a significant impact in our world.

Data science helped researchers to find out the correlation between suicide rate and its country to see whether the wealth of a country and how they can affect a country’s suicide rate and what are other factors that may affect it.We are also curious on finding what are the countries with the lowest suicide rate and the highest suicide as well to see what are the characteristics of such countries and help us better understand the phenomenon.

Here we explore a dataset regarding the suicide rate from all over the world from 1985 to 2016, while also providing a walk through tutorial of the entire data science pipeline.

Dataset URL: https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016/version/1

##Motivation I am motivated to do this topic because citizens’ health situation is the priority for every country in nowadays society. In addition to that, the development of technology should serve for the people in our world. As one of the biggest problem our society is facing that cause a huge amount of deaths, it is important for data science as this is useful for finding out the reasons behind the phenomenon and solve the problem in the future and this is the foundation data science stands for which is to help the society and solve the problems.

2.Getting Started with the Data

First, we are calling numerous libraries that are necessary for our code because these libraries consist of many useful R functions that will be used in the code.

Required Libraries

```
# Calling Libraries
library(tidyverse)
library(rvest)
library(dplyr)
library(ggplot2)
library(broom)
library(magrittr)
library(tidyr)
library(stringr)
```

Importing the Data

After downloading a CSV file from Kaggle, use the read_csv() function with the file path as the parameter to read the data from a CSV file into a data frame in R. When loading a CSV file, you can parse it with column specifications using the cols() or cols_only() function. cols() includes all columns in the data, whereas cols_only() only include the columns specified.

For more information on the read_csv() function:
- https://readr.tidyverse.org/reference/read_delim.html for more information on the cols() and cols_only() function: - https://www.rdocumentation.org/packages/readr/versions/1.3.1/topics/cols Find datasets through links provided here: - http://www.hcbravo.org/IntroDataSci/resources/ - https://www.kaggle.com/datasets

```
csv_file <- "master.csv"
df <- read_csv(csv_file)
```

```
## 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()
## )
```

```
# Parsed with column specification
cols(
  country = col_character(),
  year = col_double(),
  sex = col_character(),
  age = col_character(),
  suicides_no = col_double(),
  population = col_double(),
  suicidesper100pop = col_integer(),
  country_year = col_character(),
  HDI = col_character(),
  gdp_for_year = col_double()
)
```

```
## cols(
##   country = col_character(),
##   year = col_double(),
##   sex = col_character(),
##   age = col_character(),
##   suicides_no = col_double(),
##   population = col_double(),
##   suicidesper100pop = col_integer(),
##   country_year = col_character(),
##   HDI = col_character(),
##   gdp_for_year = col_double()
## )
```

```
df
```

```
## # A tibble: 27,820 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
##  4 Albania  1987 male  75+ ~           1      21800             4.59
##  5 Albania  1987 male  25-3~           9     274300             3.28
##  6 Albania  1987 fema~ 75+ ~           1      35600             2.81
##  7 Albania  1987 fema~ 35-5~           6     278800             2.15
##  8 Albania  1987 fema~ 25-3~           4     257200             1.56
##  9 Albania  1987 male  55-7~           1     137500             0.73
## 10 Albania  1987 fema~ 5-14~           0     311000             0
## # ... with 27,810 more rows, and 5 more variables: `country-year` <chr>, `HDI
## #   for year` <dbl>, `gdp_for_year ($)` <dbl>, `gdp_per_capita ($)` <dbl>,
## #   generation <chr>
```

The set_colnames function allows for renaming column names.

```
# Change column names
df <- df %>%
  set_colnames(c("Country", "year", "sex", "age range", "suicide_number", "total_population", "suicide_rate"
, "country_year","HDI","gdp_year","gdp_capita","generation"))
df
```

```
## # A tibble: 27,820 x 12
##    Country  year sex   `age range` suicide_number total_population suicide_rate
##    <chr>   <dbl> <chr> <chr>                <dbl>            <dbl>        <dbl>
##  1 Albania  1987 male  15-24 years             21           312900         6.71
##  2 Albania  1987 male  35-54 years             16           308000         5.19
##  3 Albania  1987 fema~ 15-24 years             14           289700         4.83
##  4 Albania  1987 male  75+ years                1            21800         4.59
##  5 Albania  1987 male  25-34 years              9           274300         3.28
##  6 Albania  1987 fema~ 75+ years                1            35600         2.81
##  7 Albania  1987 fema~ 35-54 years              6           278800         2.15
##  8 Albania  1987 fema~ 25-34 years              4           257200         1.56
##  9 Albania  1987 male  55-74 years              1           137500         0.73
## 10 Albania  1987 fema~ 5-14 years               0           311000         0
## # ... with 27,810 more rows, and 5 more variables: country_year <chr>,
## #   HDI <dbl>, gdp_year <dbl>, gdp_capita <dbl>, generation <chr>
```

Exploratory Data Analysis

Operations, such as select, filter, slice, arrange, group_by, and summarise, are used to help perform almost any analysis on data frames.

Using the various operations, we can determine which state has had the largest and smallest total average honey production between 1998 and 2012, and on average, how many colonies it takes to create 1 pound of honey for each state.

Documentation of operations used: - group_by(): https://www.rdocumentation.org/packages/dplyr/versions/0.7.8/topics/group_by - summarise(): https://www.rdocumentation.org/packages/dplyr/versions/0.7.8/topics/summarise - filter(): https://www.rdocumentation.org/packages/dplyr/versions/0.7.8/topics/filter - mutate(): https://www.rdocumentation.org/packages/dplyr/versions/0.5.0/topics/mutate

URL for the problem of Lithuania: https://en.wikipedia.org/wiki/Suicide_in_Lithuania

We can see economic problem and social transition has been a great factor which cause LIthuania to be so depressed. From this chart, we can clearly see that a country's economic situation is positively related to its suicide rate.

```
# Lithuania has the highest suicide rate among all countries
df %>%
  group_by(Country) %>%
  summarise(avg_suiciderate = mean(suicide_rate)) %>%
  filter(avg_suiciderate == max(avg_suiciderate))
```

```
## # A tibble: 1 x 2
##   Country   avg_suiciderate
##   <chr>               <dbl>
## 1 Lithuania            40.4
```

```
# Dominica and Saint Kitts and Nevis has the lowest suicide rates
df %>%
  group_by(Country) %>%
 summarise(avg_suiciderate = mean(suicide_rate)) %>%
   filter(avg_suiciderate == min(avg_suiciderate))
```

```
## # A tibble: 2 x 2
##   Country               avg_suiciderate
##   <chr>                           <dbl>
## 1 Dominica                            0
## 2 Saint Kitts and Nevis               0
```

```
# On average, the amount of gpd shared by one person in the country and its suicide data
df %>%
  group_by(Country) %>%
  mutate(mean_gdp = gdp_year/total_population)
```

```
## # A tibble: 27,820 x 13
## # Groups:   Country [101]
##    Country  year sex   `age range` suicide_number total_population suicide_rate
##    <chr>   <dbl> <chr> <chr>                <dbl>            <dbl>        <dbl>
##  1 Albania  1987 male  15-24 years             21           312900         6.71
##  2 Albania  1987 male  35-54 years             16           308000         5.19
##  3 Albania  1987 fema~ 15-24 years             14           289700         4.83
##  4 Albania  1987 male  75+ years                1            21800         4.59
##  5 Albania  1987 male  25-34 years              9           274300         3.28
##  6 Albania  1987 fema~ 75+ years                1            35600         2.81
##  7 Albania  1987 fema~ 35-54 years              6           278800         2.15
##  8 Albania  1987 fema~ 25-34 years              4           257200         1.56
##  9 Albania  1987 male  55-74 years              1           137500         0.73
## 10 Albania  1987 fema~ 5-14 years               0           311000         0
## # ... with 27,810 more rows, and 6 more variables: country_year <chr>,
## #   HDI <dbl>, gdp_year <dbl>, gdp_capita <dbl>, generation <chr>,
## #   mean_gdp <dbl>
```

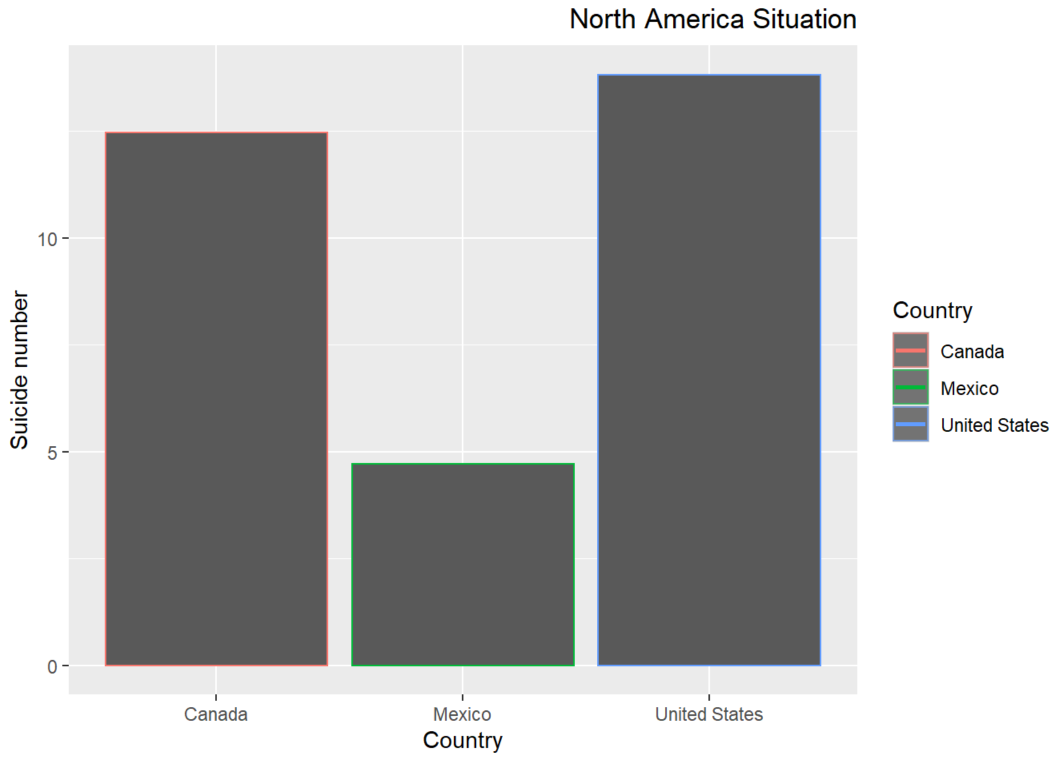
The plot below shows the suicide rate in each continent and it uses plots to help readers better understand the graph.

geom_point() is used to create the scatterplot and geom_smooth() is used to create the trend line. labs() and ggtitle() is used to customize the names of the axises and title of the plot. In addition, theme(plot.title = element_text(hjust = 0.5)) helps center the title of the plot.

Here is a ggplot2 reference sheet: https://ggplot2.tidyverse.org/reference/

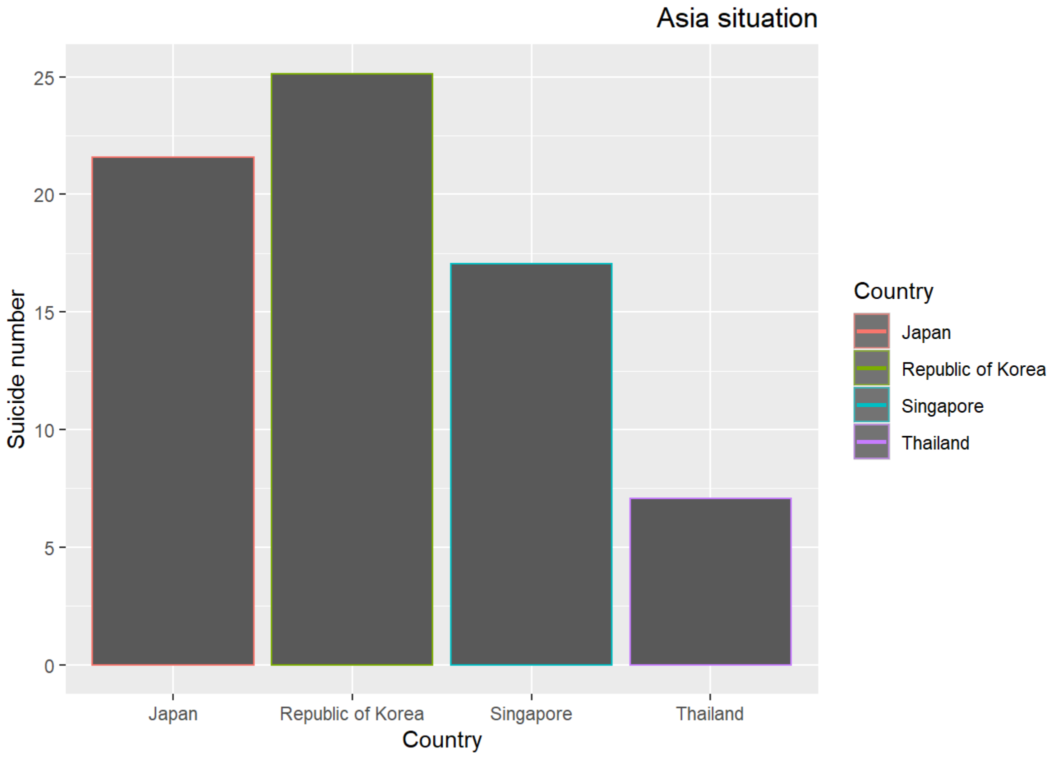
Below are some countries from North America

```
df %>%
  group_by(Country) %>%
 summarise(avg_suiciderate = mean(suicide_rate))      %>%
  filter(Country %in% c("United States", "Canada","Mexico" )) %>%
  ggplot(aes(x = Country, y = avg_suiciderate, color = Country)) + geom_col() +
  geom_smooth(method = loess) + labs(x = "Country", y = "Suicide number") +
  ggtitle("North America Situation") +
  theme(plot.title = element_text(hjust = 1.0))
```



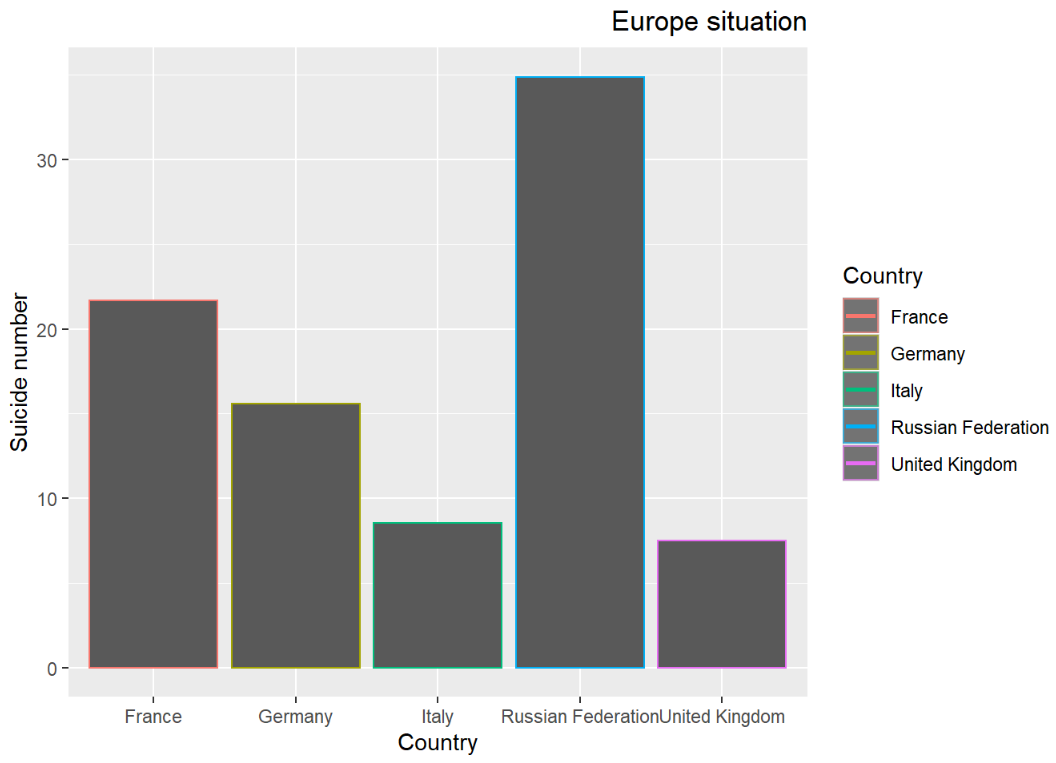
Below are countries from Asia

```
df %>%
  group_by(Country) %>%
 summarise(avg_suiciderate = mean(suicide_rate))      %>%
  filter(Country %in% c("Republic of Korea", "Japan","Thailand","Singapore" )) %>%
  ggplot(aes(x = Country, y = avg_suiciderate, color = Country)) + geom_col() +
  geom_smooth(method = loess) + labs(x = "Country", y = "Suicide number") +
  ggtitle("Asia situation") +
  theme(plot.title = element_text(hjust = 1.0))
```

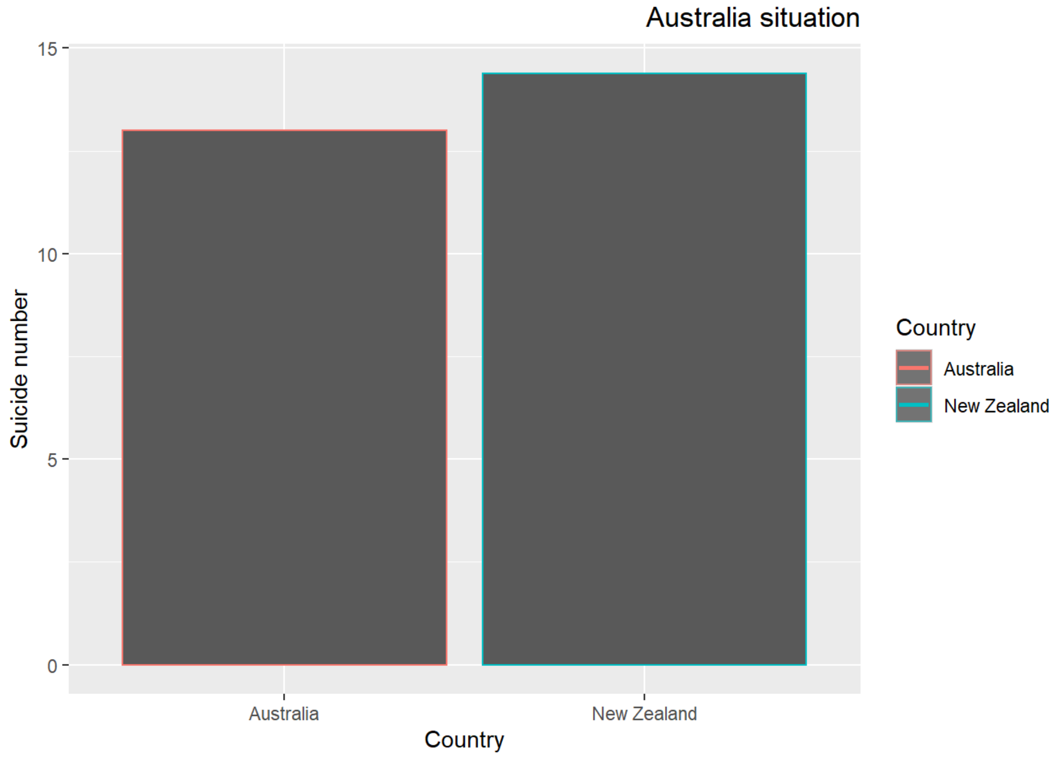


Below are countries from Europe

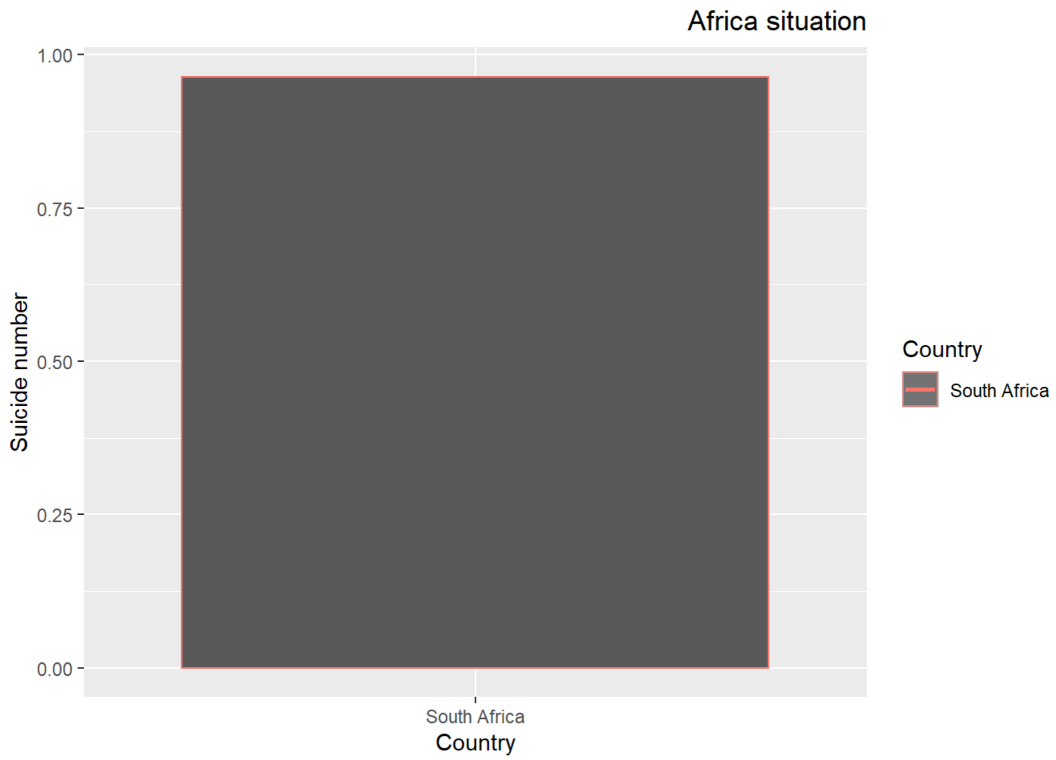
```
df %>%
  group_by(Country) %>%
 summarise(avg_suiciderate = mean(suicide_rate))      %>%
  filter(Country %in% c("Russian Federation", "France","Germany","United Kingdom","Italy" )) %>%
  ggplot(aes(x = Country, y = avg_suiciderate, color = Country)) + geom_col() +
  geom_smooth(method = loess) + labs(x = "Country", y = "Suicide number") +
  ggtitle("Europe situation") +
  theme(plot.title = element_text(hjust = 1.0))
```



```
df %>%
  group_by(Country) %>%
 summarise(avg_suiciderate = mean(suicide_rate))      %>%
  filter(Country %in% c("Australia", "New Zealand")) %>%
  ggplot(aes(x = Country, y = avg_suiciderate, color = Country)) + geom_col() +
  geom_smooth(method = loess) + labs(x = "Country", y = "Suicide number") +
  ggtitle("Australia situation") +
  theme(plot.title = element_text(hjust = 1.0))
```



```
df %>%
  group_by(Country) %>%
 summarise(avg_suiciderate = mean(suicide_rate))      %>%
  filter(Country %in% c("South Africa")) %>%
  ggplot(aes(x = Country, y = avg_suiciderate, color = Country)) + geom_col() +
  geom_smooth(method = loess) + labs(x = "Country", y = "Suicide number") +
  ggtitle("Africa situation") +
  theme(plot.title = element_text(hjust = 1.0))
```



Linear Regression

Linear regression is a very useful technique for data analysis. It allows for constructing confidence intervals, utilizing hypothesis testing for relationships between variables, and providing continuous outcomes of interest.

From http://r-statistics.co/Linear-Regression.html , “linear regression is used to predict the value of an outcome variable Y based on one or more predictor variables X. The aim is to establish a linear relationship (a mathematical formula) between the predictor variable(s) and the response variable, so that, we can use this formula to estimate the value of the response Y, when only the predictors (Xs) values are known.”

For more information on Linear Regression and Linear Models: - https://www.statisticssolutions.com/what-is-linear-regression/ - http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm - https://data.princeton.edu/r/linearmodels - This link helps explains how to fit a model, examine a fit, extract results, and much
more.

```
# Fit a linear regression model for Suicide rate vs. Year
df_fit_suicide <- lm(suicide_rate~year, data = df)
df_fit_suicide
```

```
##
## Call:
## lm(formula = suicide_rate ~ year, data = df)
##
## Coefficients:
## (Intercept)         year
##    187.7264      -0.0874
```

```
df_fit_suicide_stats <- df_fit_suicide %>%
  tidy()
df_fit_suicide_stats
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) 188.        26.8        6.99 2.75e-12
## 2 year         -0.0874     0.0134    -6.52 7.35e-11
```

```
cat("On average, the number of people commit suicide amoung 100k people decreased by", df_fit_suicide_stats$
estimate[2],
    "pounds per year from 1985 to 2016.")
```

```
## On average, the number of people commit suicide amoung 100k people decreased by -0.08740015 pounds per ye
ar from 1985 to 2016.
```

If there was a null hypothesis of no relationship between suicide rate and year, I would reject that null hypothesis because there is a relationship between the two variables: Over time, the total suicide rate decreases by around -0.874 per year.

Now, we augment the df_fit_suicide, in which columns, such as predictions, residuals, etc. are added.

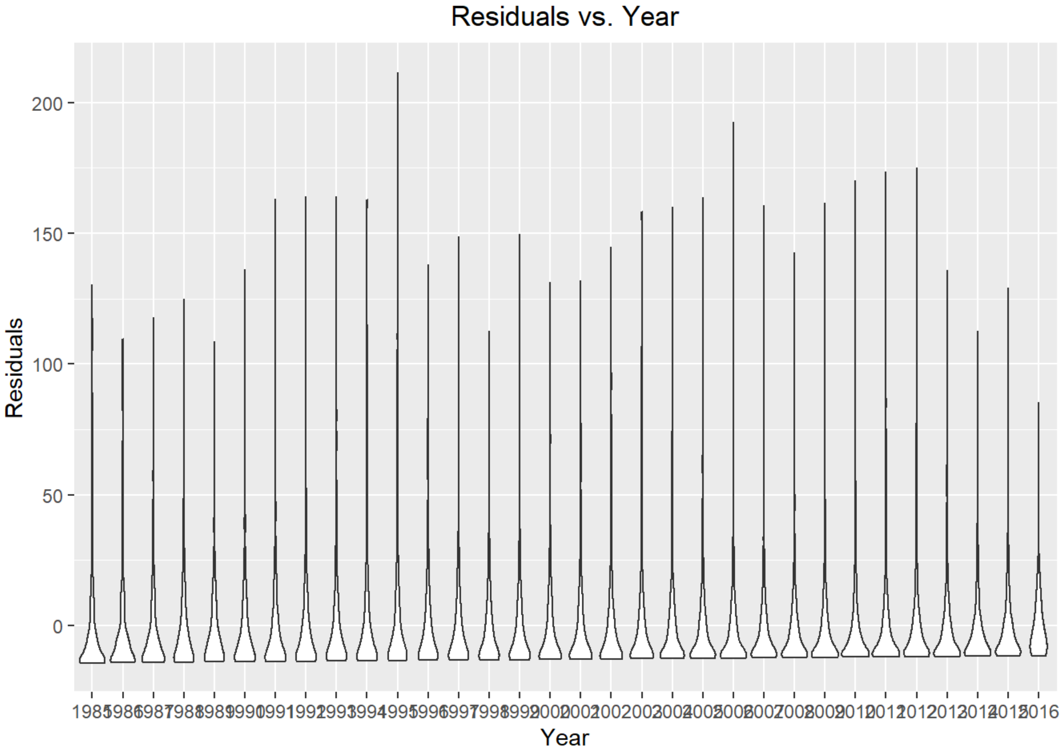
Residuals is the difference between the observed value of the dependent variable and the predicted value. Fitted values are also known as predicted values. One way to check if your linear regression model is appropriate is to plot a graph of Residuals vs. Fitted Values. This graph will check for the linearity assumption. If the regression model is appropriate, the mean of residuals will be approximately 0.

The augment() function Documentation: - https://www.rdocumentation.org/packages/broom/versions/0.4.3/topics/augment Residuals: - http://www.r-tutor.com/elementary-statistics/simple-linear-regression/residual-plot

```
aug_df_suicide <- df_fit_suicide %>%
  augment()
aug_df_suicide
```

```
## # A tibble: 27,820 x 9
##    suicide_rate  year .fitted .se.fit .resid     .hat .sigma  .cooksd .std.resid
##           <dbl> <dbl>   <dbl>   <dbl>  <dbl>    <dbl>  <dbl>    <dbl>      <dbl>
##  1         6.71  1987    14.1   0.222  -7.35 0.000138   18.9 1.04e-5     -0.388
##  2         5.19  1987    14.1   0.222  -8.87 0.000138   18.9 1.51e-5     -0.468
##  3         4.83  1987    14.1   0.222  -9.23 0.000138   18.9 1.64e-5     -0.487
##  4         4.59  1987    14.1   0.222  -9.47 0.000138   18.9 1.72e-5     -0.500
##  5         3.28  1987    14.1   0.222 -10.8  0.000138   18.9 2.23e-5     -0.569
##  6         2.81  1987    14.1   0.222 -11.3  0.000138   18.9 2.43e-5     -0.594
##  7         2.15  1987    14.1   0.222 -11.9  0.000138   18.9 2.72e-5     -0.629
##  8         1.56  1987    14.1   0.222 -12.5  0.000138   18.9 3.00e-5     -0.660
##  9         0.73  1987    14.1   0.222 -13.3  0.000138   18.9 3.41e-5     -0.704
## 10         0     1987    14.1   0.222 -14.1  0.000138   18.9 3.80e-5     -0.742
## # ... with 27,810 more rows
```

```
# A violin plot of model Residuals vs. Year
aug_df_suicide %>%
  ggplot(aes(x = factor(year), y = .resid)) + geom_violin() +
  labs(x = "Year", y = "Residuals") + ggtitle("Residuals vs. Year") +
  theme(plot.title = element_text(hjust = 0.5))
```



Conclusion

The suicide rate is positively related with a country's financial situation. We can see Lithuania is one of the poorest country in the world and it has a highest suicide rate in the world. However, once the economy level of a country reaches a certain level, we can see that the suicide problem is not only related to the economy situation of the country.

In my experienment I find out that Australia is among the countries with the lowest suicide rate, I conclude that it is not only related with financial situation but also with the living standard of the country. URL: https://degreesplusaustralia.com.au/why-australia/living-standards/ WE can see from this article that Australia's people's costs of livving is low and they have a very high standard of living, they do not need to work for many hours and they can afford their family easily.

This means they do not have many things to be bothered and that is the direct reason why Australians have the lowest suicide rate in the world. I think I find out the solutions to solve the problems: it is for government to provide more social warefares to citizens and help reduce their daily pressure. As technology develops, the introduction of data science can replace many hard works that are normally done by traditional workers. Once the government can maintain operation and gain benefits with the benefit of having data science. I believe most of people's pressure can be reduced and the problem of suicide will decrease significantly.