Predicting Suicide Rates

Capstone Project

Rahma Ali

01/10/2019

This project is submitted in partial fulfillment of the requirements for obtaining HarvardX Professional Certificate of Data Science from, offered via EdX.

1. Introduction and Project Motivation

This project aims at creating a machine learning algorithm for suicide rate prediction. It uses country level suicide rates data from the World Health Organization and other country level indicators from The World Bank (WB) and The United Nations Development Program (UNDP), starting in 1985 to 2016. The data contains information on suicide rates per 100k population with respect to populaion age, sex, generation, country Gross Domestinc Product (GDP), population size, and Human Development Index (HDI) country score. The combined dataet is available online though https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016.

At first, the data is downloaded from the web. I added the combined csv data file (as downloaded from kaggle link above) to my github repository for ease of download and code reproducibility.

```
# Download data file
################################
dl <- tempfile()
download.file("https://raw.githubusercontent.com/rali314/Suicide_Rates/master/master.csv",
        dl)
suicide <- read.csv(dl, col.names = c("country", "year", "sex", "age", "suicides_no",
        "population", "suicide_rate", "country.year", "HDI.for.year", "gdp_for_year",
        "gdp_per_capita", "generation"))
```

2. Data Exploration

Now, we take an overall look on the suicide data.

```
# Explore the data
################################
glimpse(suicide)
```

```
## Observations: 27,820
## Variables: 12
## $ country        <fct> Albania, Albania, Albania, Albania, Albania, Al...
## $ year           <int> 1987, 1987, 1987, 1987, 1987, 1987, 1987, 1987,...
## $ sex            <fct> male, male, female, male, male, female, female,...
## $ age            <fct> 15-24 years, 35-54 years, 15-24 years, 75+ year...
## $ suicides_no    <int> 21, 16, 14, 1, 9, 1, 6, 4, 1, 0, 0, 0, 2, 17, 1...
## $ population     <int> 312900, 308000, 289700, 21800, 274300, 35600, 2...
## $ suicide_rate   <dbl> 6.71, 5.19, 4.83, 4.59, 3.28, 2.81, 2.15, 1.56,...
## $ country.year   <fct> Albania1987, Albania1987, Albania1987, Albania1...
## $ HDI.for.year   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
## $ gdp_for_year   <fct> "2,156,624,900", "2,156,624,900", "2,156,624,90...
```

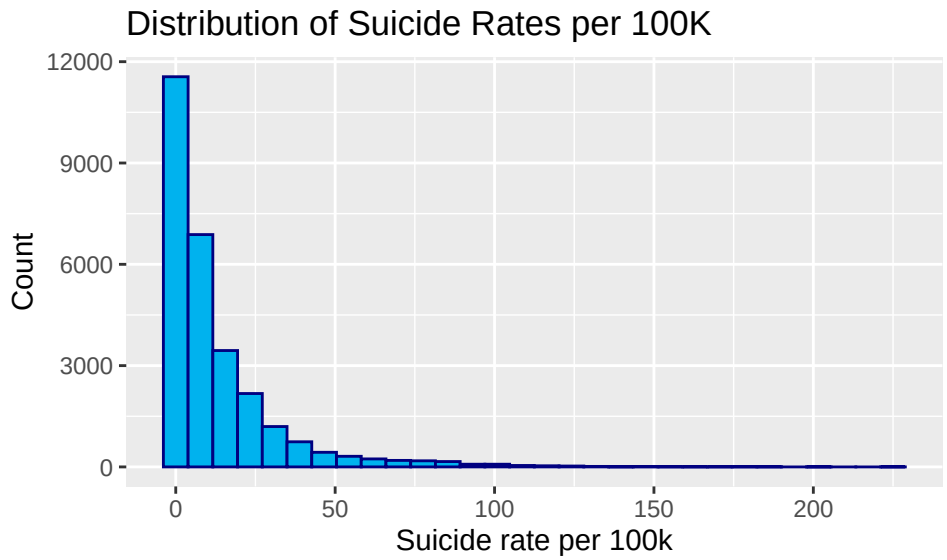
```
## $ gdp_per_capita <int> 796, 796, 796, 796, 796, 796, 796, 796, 796, 79...
## $ generation     <fct> Generation X, Silent, Generation X, G.I. Genera...
```

The `suicide` dataset in hand includes 12 varibales: 1 target variable and 11 features. The target variable is `suicide_rate`, which is the suicide rate per 100k population. The features, respictively, are: country name, year, sex, age, population size, HDI score, GDP and GDP per capita and generation.

The Target Variable: `suicide_rate`

The distribution of the `suicide_rate` variable seems to be extremely positivly skewed to the right, with a spike at the first bin closest to the value 0. This shape suggests that the majority of the observations have a very small value of suicide rate with a small number of observations with very high values, causing a very long positive tail to the shape of the distribution. This severe skewness might suggest the use of a transformation. This topic will be discussed further later on in the report.

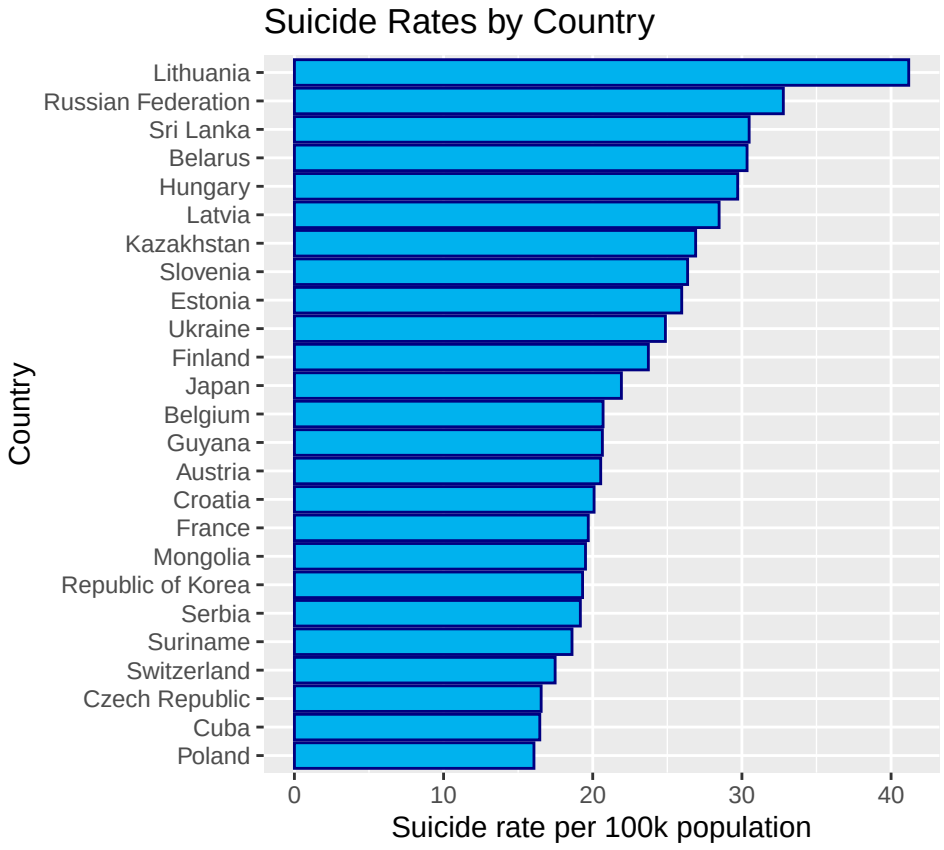
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Exploring Suicide rate Variability by Country

Suicide rate varies from one country to another. This is confirmed by the following bar chart. For easier interpretation, the figure shows the top 25 countries in terms of suicide rates. Lithuania is universally the top country in terms of suicide at a little over 40 suicides per 100k population. This rate is extremely high especially that Lithuania is not a big country like Russia, for example, which comes second after Lituania in the ranking. In fact, the average population size in Lithuania between 1985 and 2016 is 40.415572519084 compared to 34.8923765432099. This disproportion in the case of Lithuania induces further exploration of the country populatio size.

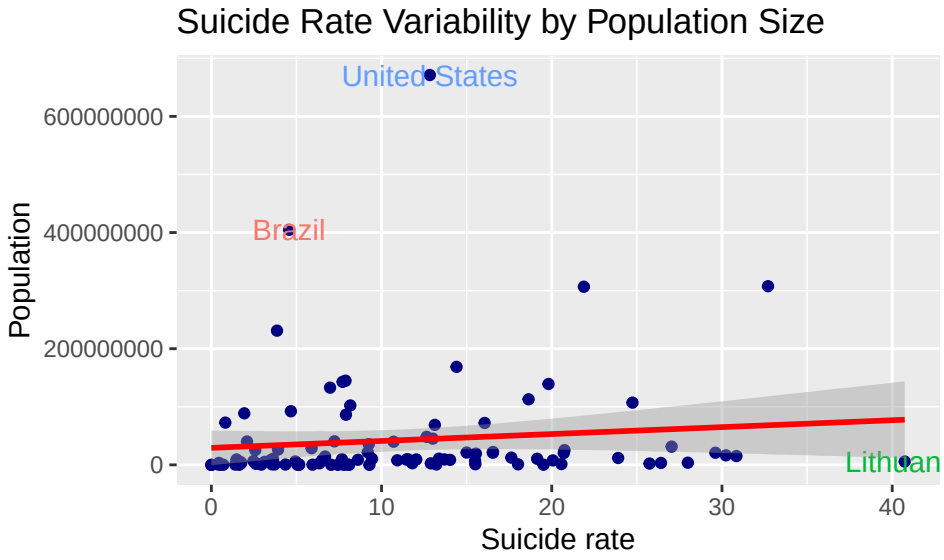
```
## Selecting by country_suicide_rate
```



Suicide Rate Variability by Population Size

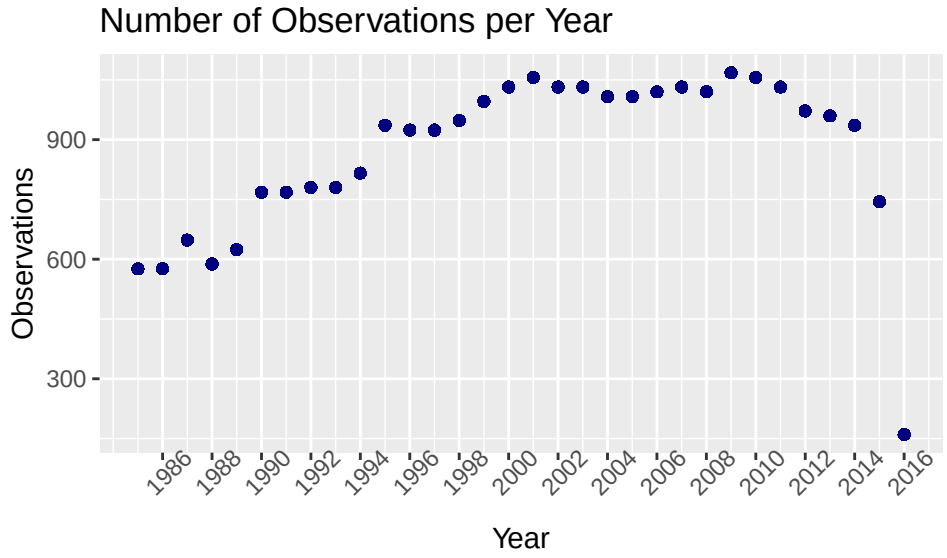
There seems to be some sort of positive linear association between the country population size and its corresponding suicide rate. Lithuania and the United States can be considered as two outliers in 2 opposite directions in the data. Presence of such extreme values affects the prediction process.

```
## Warning: Width not defined. Set with `position_dodge(width = ?)`
```

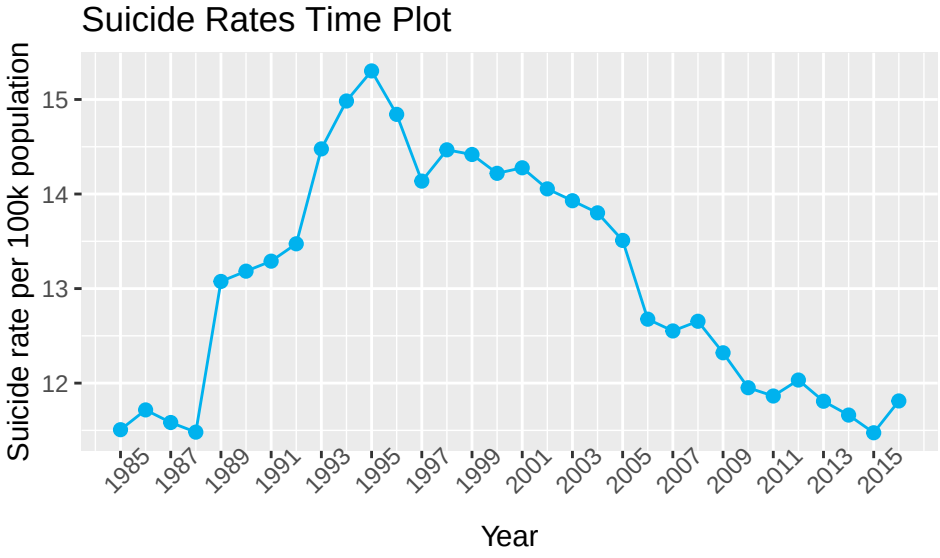


It seems that not all years included in the dataset have the same number of observations. The following plot

shows that the year 2016 has the least number of observations. For this reason, year 2016 will be excluded later in the analysis

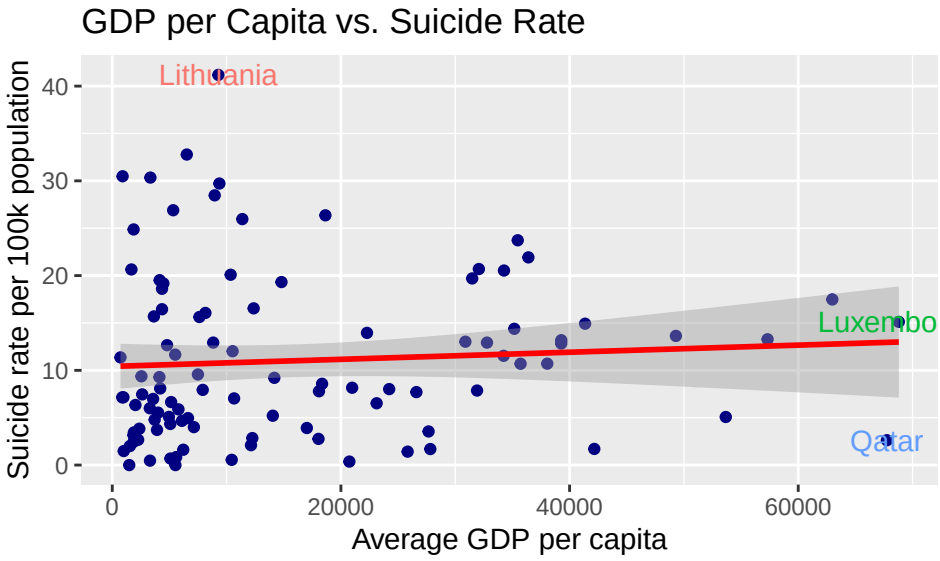


Suicide rates vary from one year to another. The following timeplot shows that before 1995, there was a global ascending trend of suicide. The opposite is true after 1995, as we see suicide rates decrease in a downward trend.



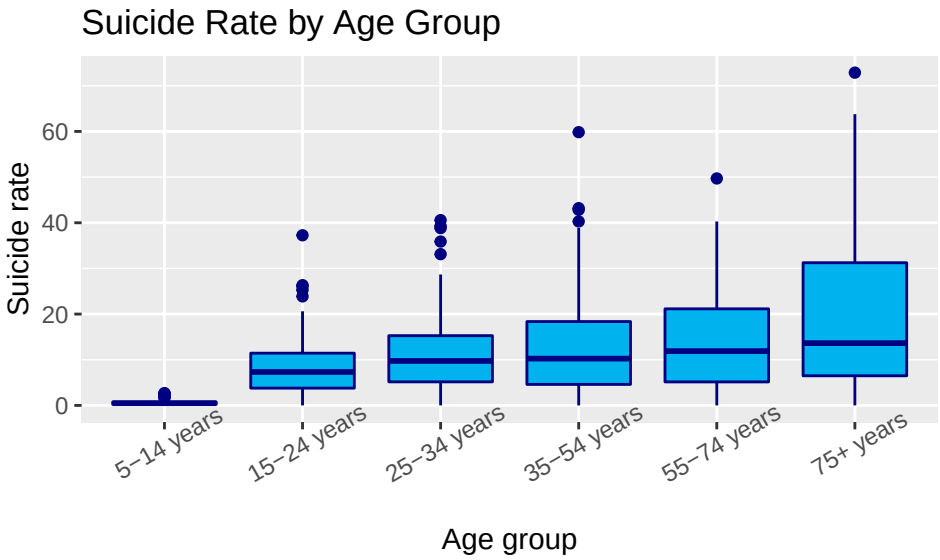
Effect of Nation Wealth on Suicide Rates

The next plot shows suicide rates plotted against per capita GDP of the countries in the dataset. It seems that there is a positive linear correlation between suicide rates per 100k population and the country's GDP per capita. There exists some outlier values in the case of Lithuania, where we have relatively small per capita GDP and very high suicide rate (the highest as seen earlier).



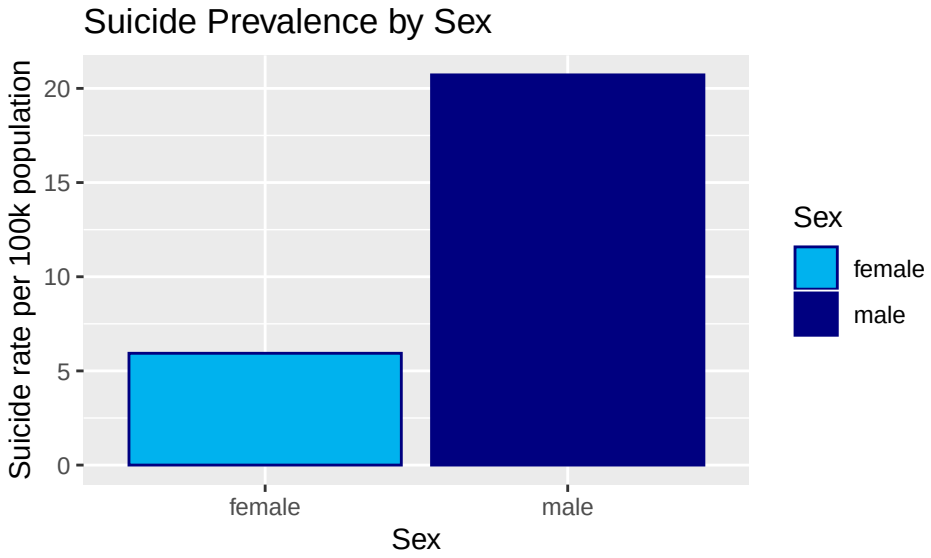
Population Characteristics Variability

Now after taking an overview on suicide variability by several country and nation level variables, we move to demographic variables that characterise the populations of these countries. First, we look at suicide variability by age goup. The following plot shows that suicide rates distribution varies from one age group to another. The highest suicide rates are found in individuals aged 75+ and the lowest suicide rates are found in individuals aged between 5-14.

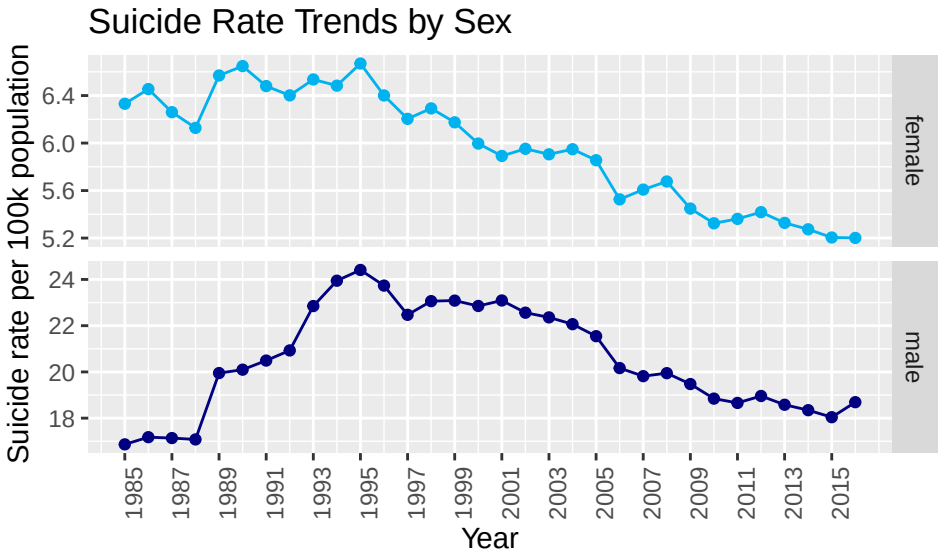


When looking at suicide rates by sex, the data shows that suicide is more prevalent among males than females, with universal suicide rate of almost 21 versus about 6 suicides per 100k male population.

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



Suicide rate trends across time varies by sex as well. The following plot shows that the suicide rates for females exhibit a universal descending trend across time while the trend for males fluctuates around 1995 which, resembles the trend we saw before.



3. Method and Analysis

From the exploratory analysis performed above, two main observations emerge:

- Lithuania was found to have an extreme value for suicide rates per 100k population.
- The year 2016 included the least numnber of observations.
- The target variable, `suicide_rate`, is severly positively skewed

Due to these factors, predicting suicide rates can be highly jeopardized if the data is used as is. Several model attempts are made and performance of each attempt is measured.

Specifying the Model

Since the target variable, `suicide_rate` is a continuous variable, multiple linear regression algorithm is considered. A regression model is fitted that takes into account country, population size, per capita CGP, year, sex and age group effects on suicide rates per 100k population. The following model is consideres:

$$suicide\_rate_{c,p,g,y,s,a} = \mu + \beta_c + \beta_p + \beta_g + \beta_y + \beta_s + \beta_a + \epsilon_{c,p,g,y,s,a}$$

where β_c is the country effect, β_p is the population effect, β_g is the per capita GDP effect, β_y is the year effect, β_s is the sex effect, β_a is the age group effect and ϵ is the model error term.

Prepping the Data for the Model

Since the target variable is far from normality, a transformation is considered in order to scale the distribution to symmetry. Many transformations are discussed in the Staitstical literature for the purpose of transforming a skewed variable to a symmetric one, however, not all of them are suitable for the `suicide_rate` variable. The natural *log* transformation produces infinte values since `suicide_rates` include zero values. For this reason, the value 1 is added to the variable prior to appling the natural *log* transformation.

```
# Variable transformation
suicide_log <- suicide %>%
  mutate(suicide_rate_log=log(1+suicide_rate))
```

Create Training and Testing sets

Training and testing datasets are created. Testing set is 20% of the entire dataset.

```
# Split to training and testing datasets
set.seed(1, sample.kind="Rounding")
```

```
## Warning in set.seed(1, sample.kind = "Rounding"): non-uniform 'Rounding'
## sampler used
```

```
test_index <- createDataPartition(y = suicide_log$suicide_rate_log, times = 1,
            p = 0.2, list = FALSE)
train_log <- suicide_log[-test_index,]
test_log <- suicide_log[test_index,]
```

Train the Model

A linear regression model is fitted using the train data using the `lm` function.

```
# Train the model
fit <- train_log %>%
  lm(suicide_rate_log ~ population + country + as.factor(year) + sex + age + gdp_per_capita,
     data=.)
```

Test the Model

After this, the model is applied to the test data. This is done through the `predict` function. To test the performance of the model, the Root Mean Squared Errors (RMSE) is considered. After generating the model predictions, the RMSE is calculated by comparing the model predictions against the true value of the suicide rates.

The value of RMSE is 0.6912835.

2 other models are attempted with the following changes in each attempt:

- Eliminate the year 2016
- Eliminate Lithuania

Training and testing data sets are generated again to accommodate the above changes for each case and the models are refitted. Comparison between models perfomance is done through RMSE values.

4. Results and Conclusions

The suicide rates per 100k populations prediction algorithm includes several country-level variables in addition to population demographic characteristics. The model used for prediction is a multiple linear regression model fitted to the training data and tested on the testing data. 3 models are fitted based on 3 different cases. Assessment of the perfomance is based on the value of RMSE

The following table summarizes the performance of each model:

```
rmse_results <- tibble(method = c("Model1: log trans",
                "Model2: log trans, 2016 trim",
                "Model3: log transform, 2016 and Lithuania trim"),
                RMSE = c(rmse, rmse_trim, rmse_trim2))
rmse_results
```

```
## # A tibble: 3 x 2
##   method                                          RMSE
##   <chr>                                          <dbl>
## 1 Model1: log trans                              0.691
## 2 Model2: log trans, 2016 trim                   0.685
## 3 Model3: log transform, 2016 and Lithuania trim 0.697
```

Model 2 yielded the least value of RMSE at 0.6854295 and is considered the best model for predicting suicide rates per 100k population.

Additional modeling attempts were made. They are based on the concept of regularization; to penalize the least squares estimates using the parameter `lambda` to optimize for `lambda` that minimizes the RMSE. The attempt yielded value of zero for `lambda`: suggesting that no penalty for the least squares estimates of the model would further enhance the model performance. The attempt is not discussed in the report but it is included in the R script file.

There are other variables that exist in the dataset but not included in the analysis and the modeling. These variables are:

- `gdp_for_year`, which is the GDP of the country at a given year. It is eliminated as it is highly correlated with `gdp_per_capita` to eliminate multicollinearity in the model. `gdp_per_capita` was selected over `gdp_for_year` as it is a better measure for the GDP and wealth of the nations that takes into account population size.
- `generation`, which is a categorical variable for the generation of the population. It is left out as it is highly correlated with `age`. `age` is selected over `generation` to include in the analysis as it more easily and intuitively understood.
- `HDI.for.year`, as over two thirds of the variable is missing.

5. Limitations and Future Work

In this project, a transformation for the target variable is introduced which is the log transformation after adding +1 to the variable. This particular transfomration and the value 1 were are selected based on convenience. Other transformations can be explored and optimized for the purpose of predicting suicide rates.

In one of the model fitting attemps, records for Lithuania are eliminated. This is because Lithuania was identified as an outlier amongst the other countries in the data. This approach may not be the best to handle such case, especially if the suicide rate values reported for Lithuania are accurate. Further work can go into exploring methods to better understand and penalize the leverage that Lithuania imposes on the data