Global Suicide Statistics

An analysis of critical variables

Team BSJ - Shannon Houser, Jack McNeilly, Brian Linder

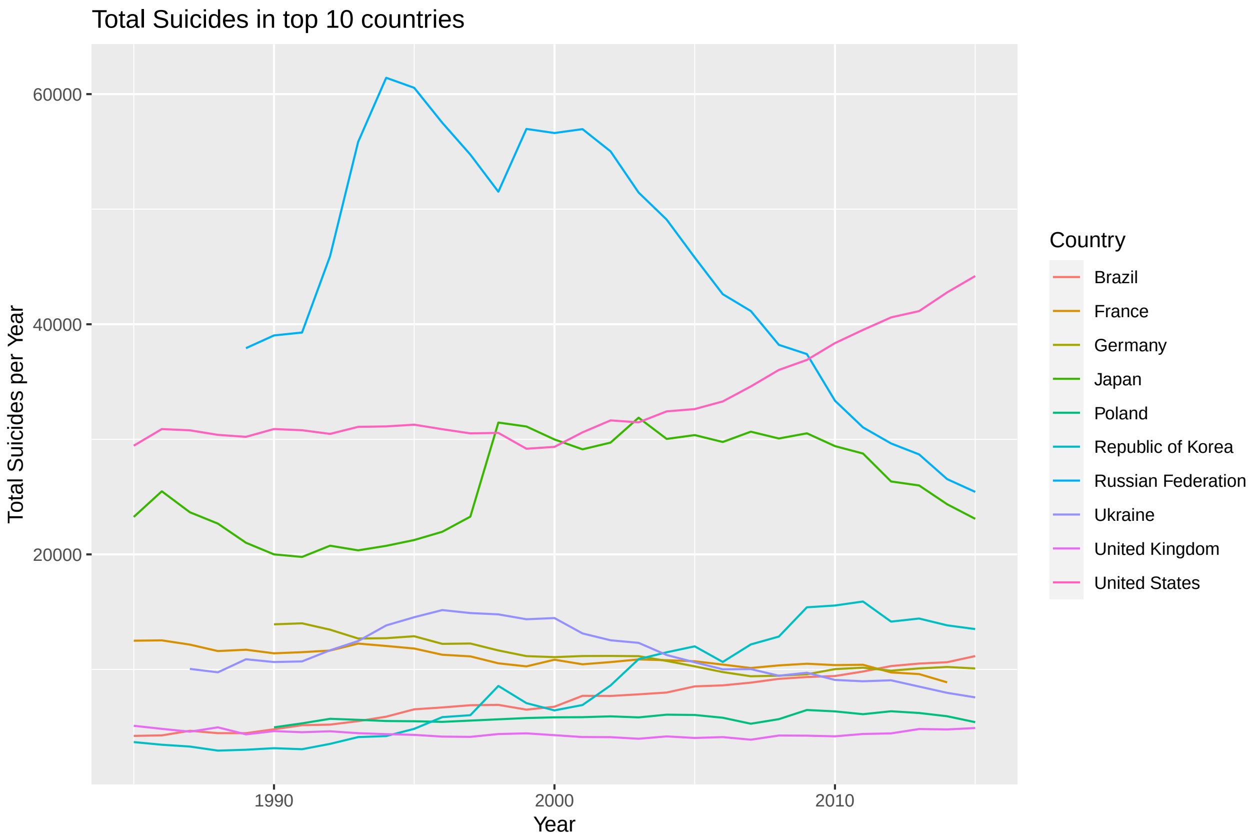
Introduction

As the Duke community grieves the loss of two of our classmates to suicide in the last 2 weeks, our team plans to analyze general global suicide rates from 1985-2016 in order to see if there are any prevalent factors that might contribute to people taking their own lives. The dataset we have selected compiles data from four distinct datasets that includes information on suicides from over 100 different countries throughout the world.The data compares socio-economic info with suicide rates by country and year. The data is sourced from the World Health Organization, the World Bank and, the United Nations Development Program.
Our goal is to examine these different socio-economic, location, and gender factors to gain insight regarding how the variables of the dataset impact increased suicide rates. Each observation corresponds to the number of suicides that occurred in a certain country and within a certain age and gender group. The variables include country, year, sex, age group, count of suicides, population, suicide rate, country-year composite key, HDI for year, gdp for year, gdp per capita, and generation.

Data Analysis Plan

Summary Statistics and Visualizations

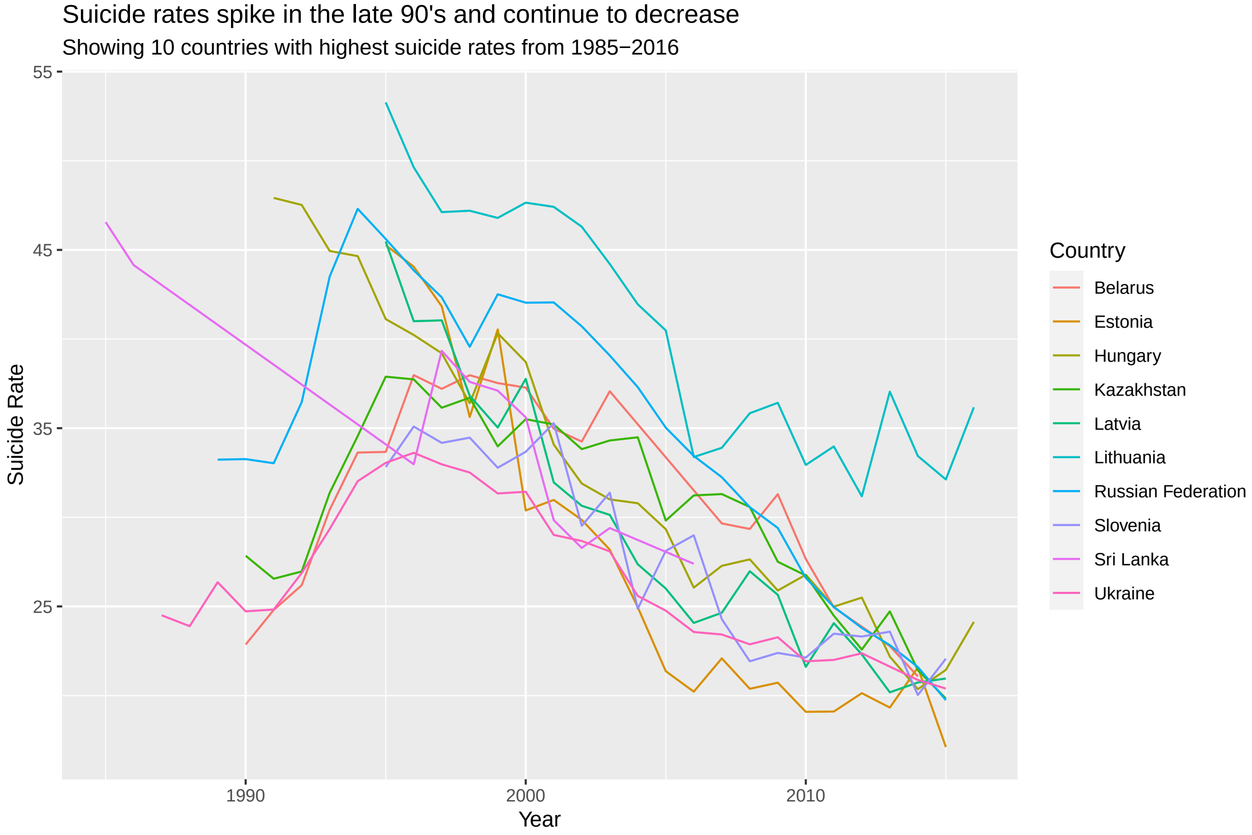
```
# A tibble: 10 x 2
   country            total_suicides
   <chr>                       <dbl>
 1 Russian Federation        1209742
 2 United States             1034013
 3 Japan                      806902
 4 France                     329127
 5 Ukraine                    319950
 6 Germany                    291262
 7 Republic of Korea          261730
 8 Brazil                     226613
 9 Poland                     139098
10 United Kingdom             136805
```



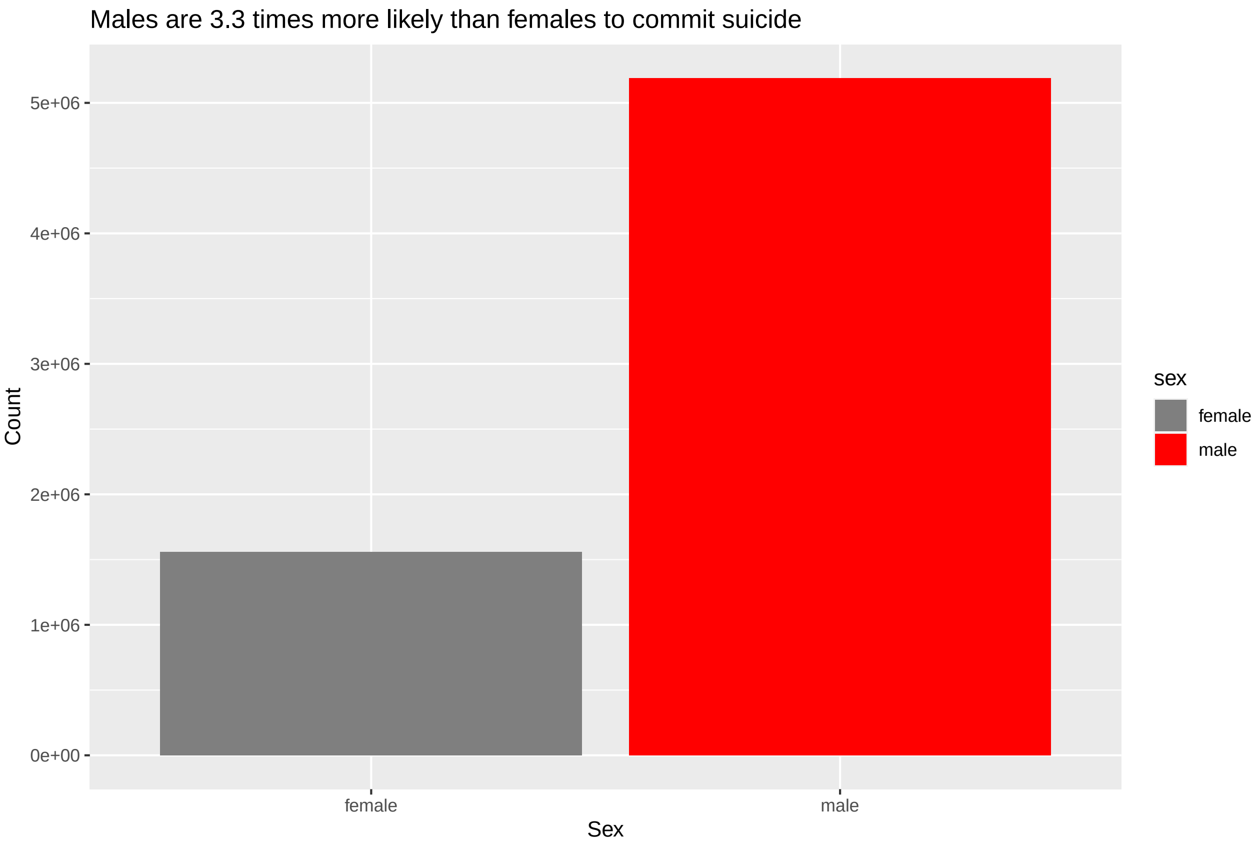
The top ten countries in terms of total suicides seem to be those that are most populated. We then decided to graph how the total number of suicides in these countries has changed over the years. We decided that this is not really helpful information and went on to explore further.

```
# A tibble: 10 x 2
   country            rate_suicide
   <chr>                     <dbl>
 1 Lithuania                  40.4
 2 Sri Lanka                  35.3
 3 Russian Federation         34.9
 4 Hungary                    32.8
 5 Belarus                    31.1
 6 Kazakhstan                 30.5
 7 Latvia                     29.3
 8 Slovenia                   27.8
 9 Estonia                    27.3
10 Ukraine                    26.6
```

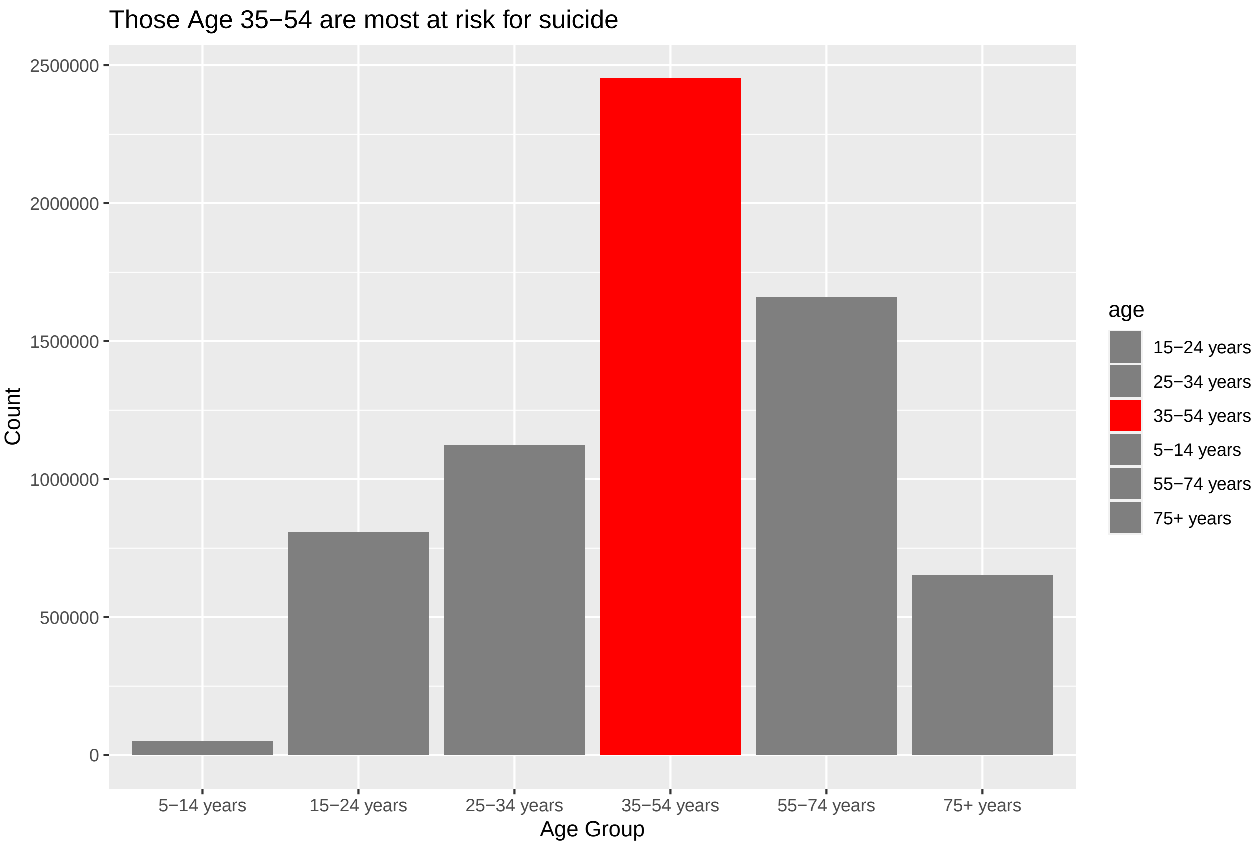
When we examined the suicide rate (per 100k people), we found a strong correlation between geopolitical circumstances as 9/10 of the top 10 countries for suicide rate were part of the ex- Soviet Union. They are all Eastern European countries that may share history, religions, wars, etc. that we are unable to currently predict.



Graphing the average suicide rates over time in the ten countries with the highest suicide rates, we can see that the suicide rates peaked in the late 1990's and have continued to decrease since.



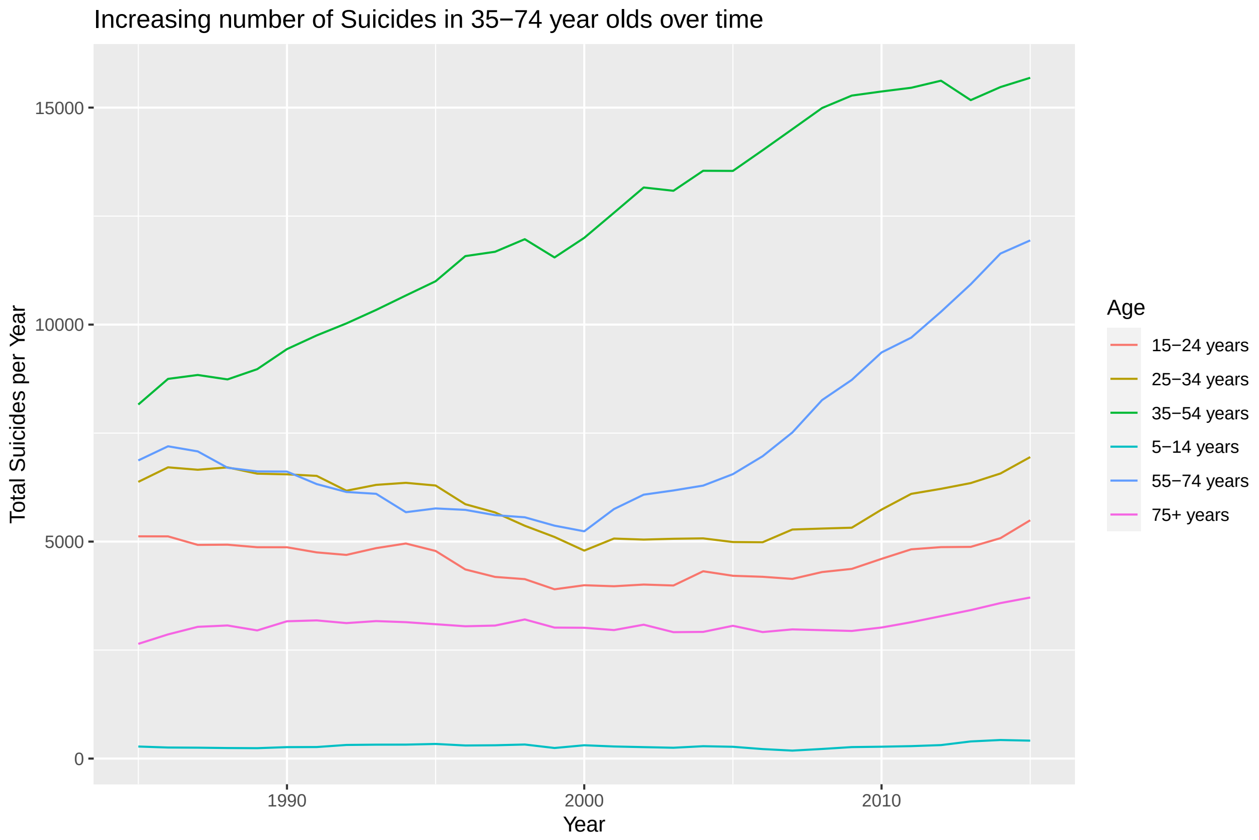
From this data visualization, it is obvious that sex is probably a very important variable when predicting suicide. Men are more than 3.3 times more likely than women to commit suicide.



This visualization shows us that age may also have a strong influence on the likeliness of someone to commit suicide. From the last two visualizations, we may suggest the possibility that middle-aged men are most at risk for suicide. We may also want to examine this relationship more and see brainstorm what life factors make middle-aged men more likely than any other group to commit suicide.

```
# A tibble: 6 x 2
  generation      tot_gen
  <chr>             <dbl>
1 Boomers         2284498
2 Silent          1781744
3 Generation X    1532804
4 Millenials       623459
5 G.I. Generation  510009
6 Generation Z      15906
```

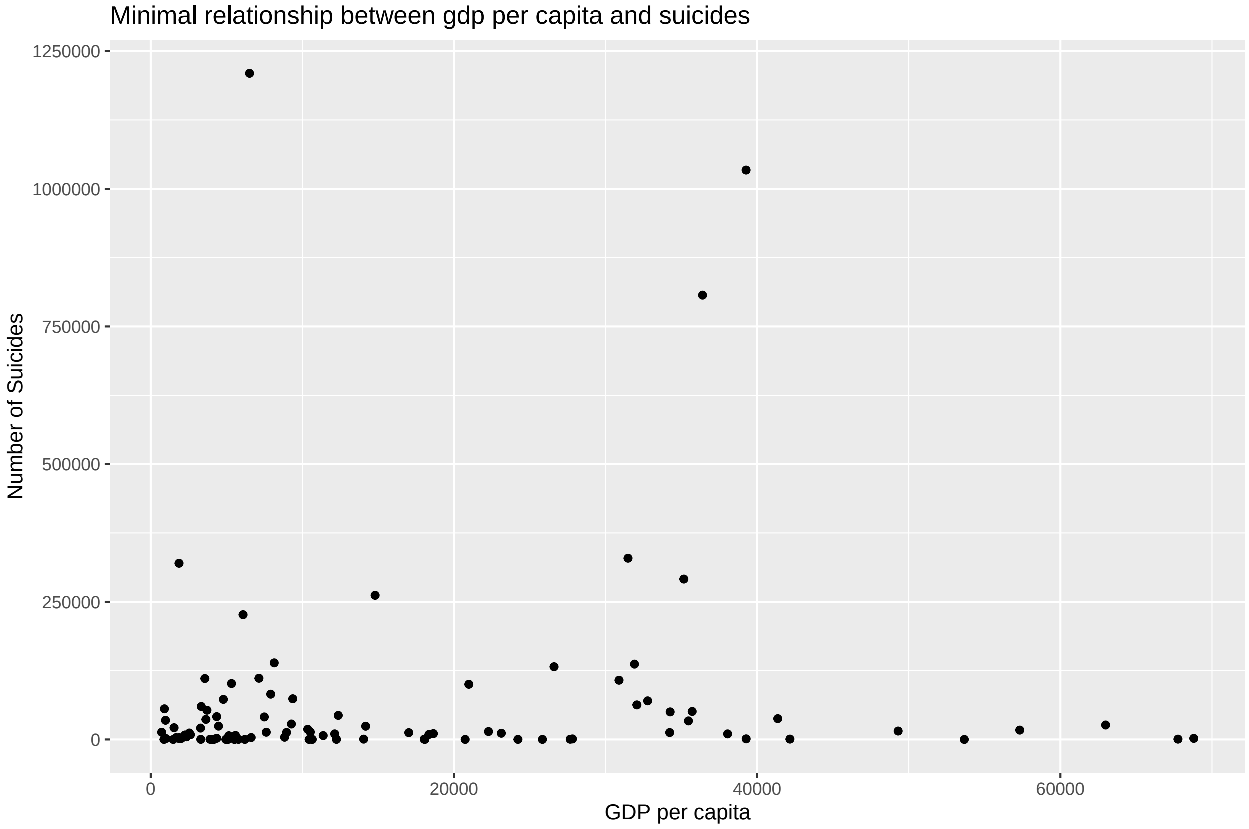
This shows the total number of suicides per generation. This has a lot to do with age group and thus is redundant; however, it may help us to better understand what kinds of life circumstances outside of the data these people may have faced to lead them to commit suicide.



From this visualization we can see that over time, the number of global suicides for those between the ages of 35- 74 have increased the most drastically. The other age groups seem to be roughly stable; however, it does appear that all other age groups are increasing at the very end of the graph. It would be interesting to see if this sad trend continued past the last year of this study's data collection, 2016.

```
# A tibble: 101 x 2
   country            avg_gdp
   <chr>                <dbl>
 1 United States      1.05e13
 2 Japan              4.34e12
 3 Germany            2.74e12
 4 United Kingdom     1.82e12
 5 France             1.78e12
 6 Italy              1.48e12
 7 Brazil             1.02e12
 8 Canada             9.13e11
 9 Russian Federation 8.84e11
10 Spain              8.57e11
# ... with 91 more rows
```

```
# A tibble: 101 x 2
   country              avg_gdp_capita
   <chr>                         <dbl>
 1 Luxembourg                   68798.
 2 Qatar                        67756.
 3 Switzerland                  62982.
 4 Norway                       57320.
 5 San Marino                     53664.
 6 Denmark                        49300.
 7 United Arab Emirates           42162
 8 Sweden                         41358.
 9 Iceland                        39275.
10 United States                  39270.
# ... with 91 more rows
```



Planning

The response variables we will test are the numbers of suicides and the number of suicides per 100k people. The explanatory variables we will examine are different age groups, sex, years, countries, and the socioeconomic status of each country, including their Human Development Index (HDI),growth domestic product (GDP), and GDP per capita.

In addition to observing each individual explanatory variable's impact on the response variables, we will examine how the following combination of explanatory variables and the corresponding result on the response variables:

-Age groups faceted by sex -HDI with age -HDI with sex -HDI with sex and age -GDP with age -GDP with sex -GDP with sex and age -the above combinations faceted by time period (years) -the above combinations for Each individual country -the above combinations for World regions including continents and sub-regions of each continent

In our analysis of the dataset we plan to use statistical methods and tools in R including, linear modeling, regression modeling, a combination of visualization techniques, and null hypothesis testing.

In our analysis of the dataset we plan to use statistical methods and tools in R including, linear modeling, regression modeling, a combination of visualization techniques, and null hypothesis testing.

From our preliminary analysis, we believe that trying to find predictors of suicide using modeling techniques would be a good place to start. We believe that such variables as sex and age may have large impacts on the response variables of total suicide numbers and mean suicide rates. We also believe that gdp and gdp per capita may not play as large of a role as people may think. Instead, perhaps geopolitical factors that are outside of our datasets scope play a large part in suicide determinants.

Furthermore, we plan to explore how these factors have changed over time and if the changes are statistically significant. For example, we will explore whether the total number of suicides in the US has significantly changed between 1985 and 2016, and compare these changes with comparable nations within the data.

Glimpse of Data

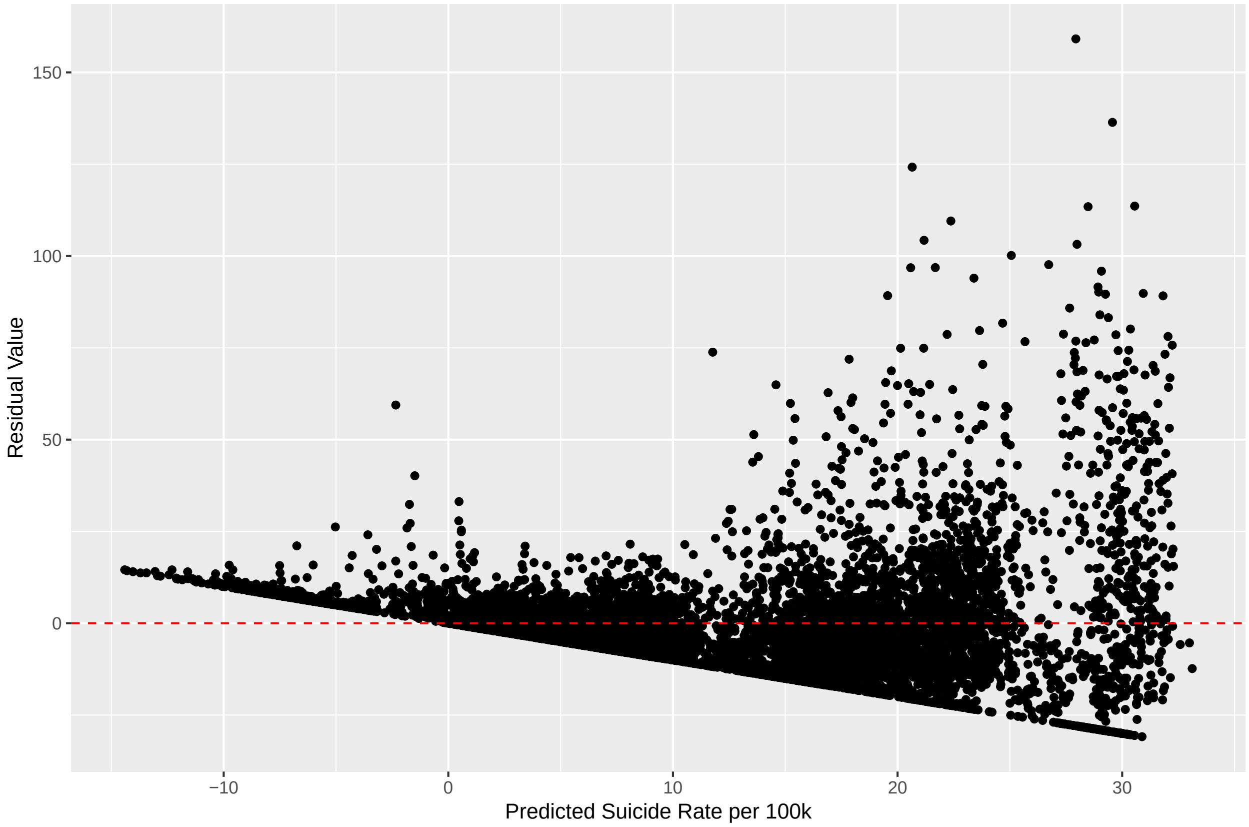
```
Observations: 27,820
Variables: 12
$ country              <chr> "Albania", "Albania", "Albania", "Albania", "A...
$ year                 <dbl> 1987, 1987, 1987, 1987, 1987, 1987, 1987, 1987...
$ sex                  <chr> "male", "male", "female", "male", "male", "fem...
$ age                  <chr> "15-24 years", "35-54 years", "15-24 years", "...
$ suicides_no          <dbl> 21, 16, 14, 1, 9, 1, 6, 4, 1, 0, 0, 0, 2, 17, ...
$ population           <dbl> 312900, 308000, 289700, 21800, 274300, 35600, ...
$ `suicides/100k pop`  <dbl> 6.71, 5.19, 4.83, 4.59, 3.28, 2.81, 2.15, 1.56...
$ `country-year`       <chr> "Albania1987", "Albania1987", "Albania1987", "...
$ `HDI for year`       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
$ `gdp_for_year ($)`   <dbl> 2156624900, 2156624900, 2156624900, 2156624900...
$ `gdp_per_capita ($)` <dbl> 796, 796, 796, 796, 796, 796, 796, 796, 796, 7...
$ generation           <chr> "Generation X", "Silent", "Generation X", "G.I...
```

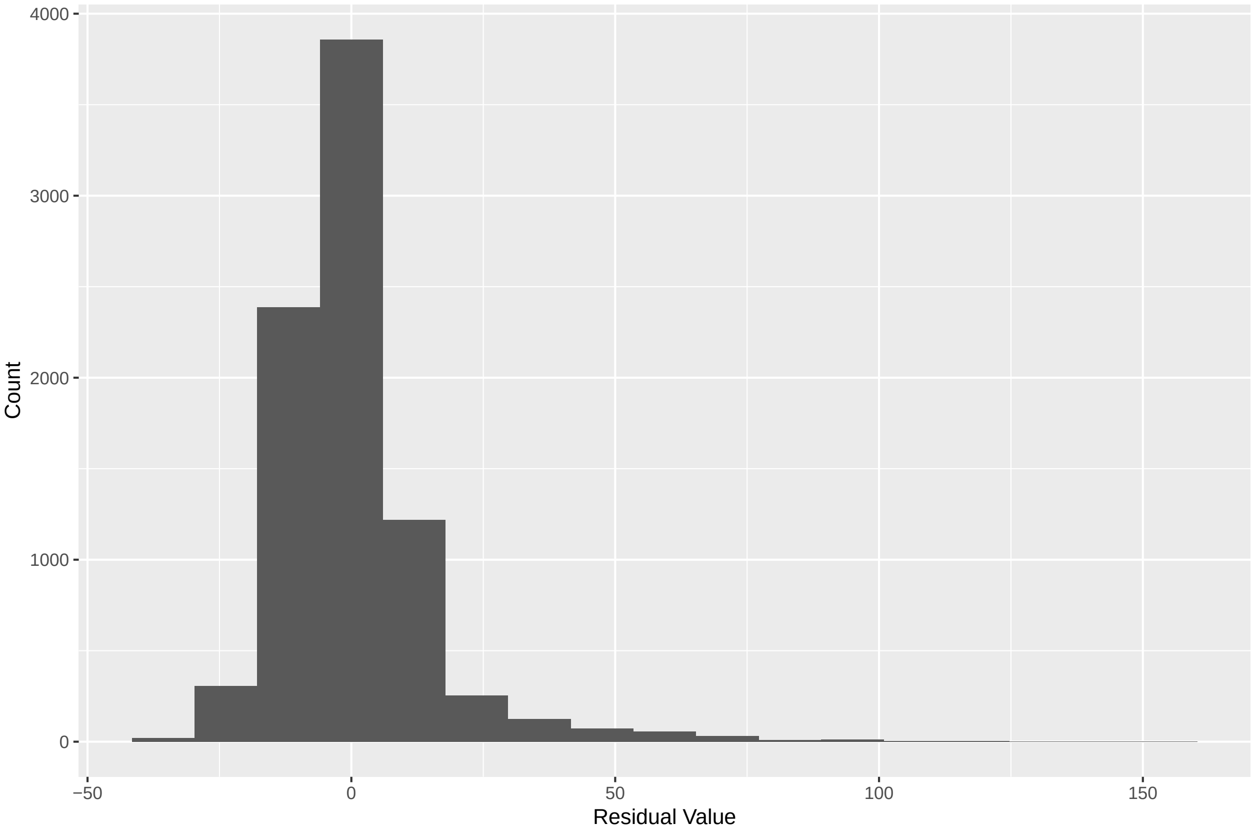
Try at LM

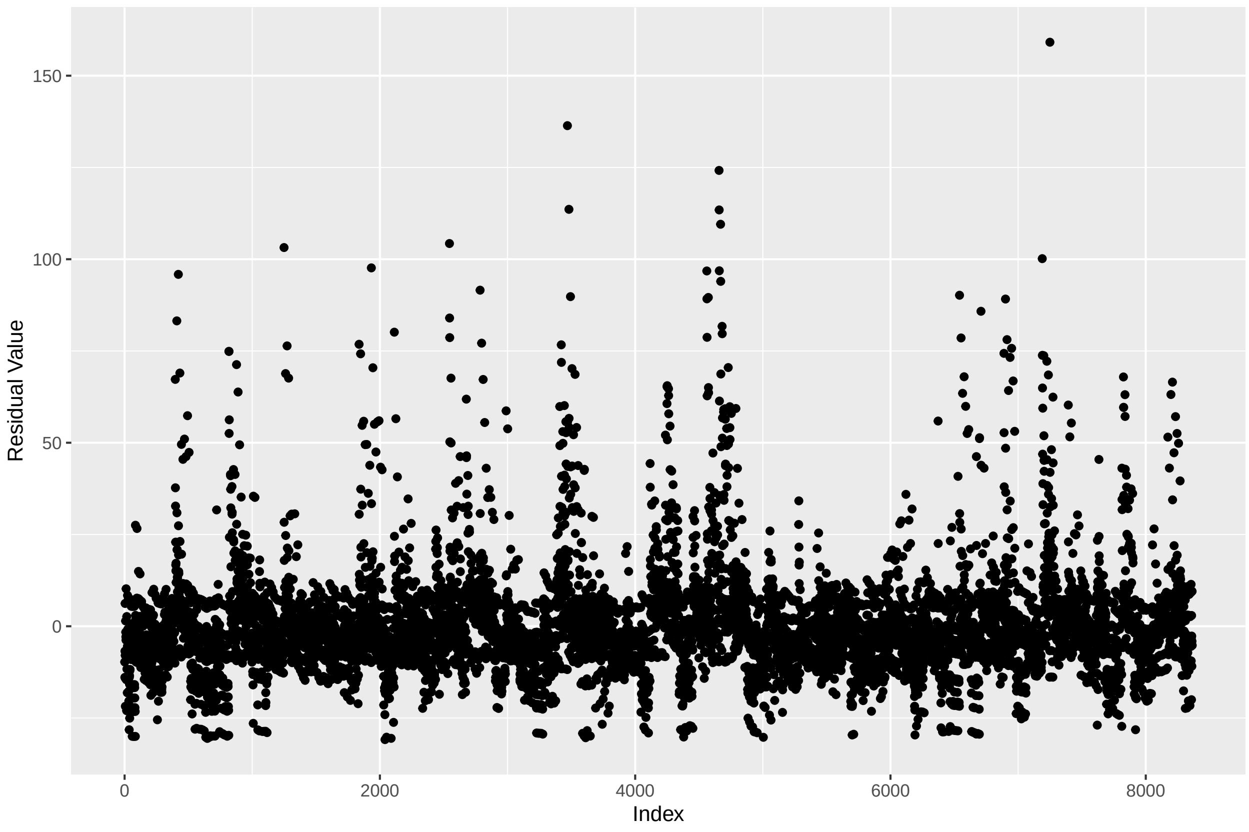
```
Start:  AIC=44538.69
suicides_100k_pop ~ factor(sex) + factor(age) + hdi_for_year + 
    gdp_per_capita

                 Df Sum of Sq     RSS   AIC
<none>                        1714421 44539
- gdp_per_capita  1     22117 1736538 44644
- hdi_for_year    1     36012 1750433 44711
- factor(age)     5    354151 2068572 46099
- factor(sex)     1    416271 2130692 46355
```

```
# A tibble: 9 x 5
  term                     estimate std.error statistic   p.value
  <chr>                       <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)            -23.1      1.92         -12.0  3.88e- 33
2 factor(sex)male         14.1      0.313         45.0  0.       
3 factor(age)25-34 years   2.82     0.543          5.19 2.16e-  7
4 factor(age)35-54 years   5.46     0.543         10.1  1.01e- 23
5 factor(age)5-14 years   -8.02     0.543        -14.8  7.37e- 49
6 factor(age)55-74 years   6.49     0.543         12.0  1.02e- 32
7 factor(age)75+ years    13.3      0.543         24.5  4.60e-128
8 hdi_for_year            34.9      2.64          13.2  1.17e- 39
9 gdp_per_capita          -0.000113 0.0000109    -10.4  4.25e- 25
```

```
[1] 0.3192574
```


Residual Value
150
100
50
0
−10
0
10
20
30
Predicted Suicide Rate per 100k


Count
4000
3000
2000
1000
0
−50
0
50
100
150
Residual Value


Residual Value
150
100
50
0
0
2000
4000
6000
8000
Index

We wanted to fit a linear model to decide which factors were the best when predicting suicide rates across the world. We created a linear model and used the step function to see which were the best variables for predicting suicide rates. However, our first attempt at a linear model was unsuccessful. We saw that the data was largely skewed to the right and thought it best to do a data transformation and predict the log of the suicide rate instead.

Log Linear Model

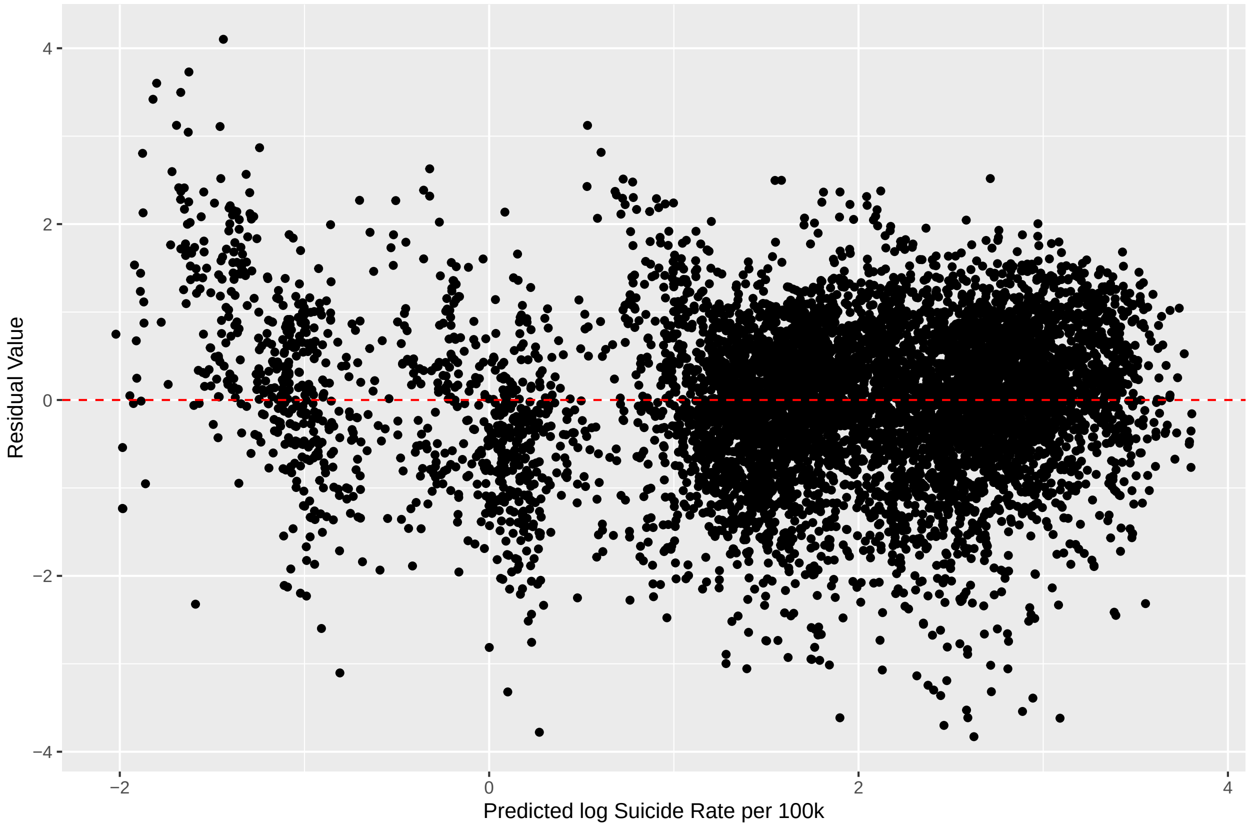
```
Start:  AIC=-1998.11
log(suicides_100k_pop) ~ factor(sex) + factor(age) + year + hdi_for_year + 
    log(gdp_per_capita)

                      Df Sum of Sq     RSS     AIC
<none>                              5446.8 -1998.1
- year                 1     120.8  5567.6 -1842.0
- log(gdp_per_capita)  1     210.3  5657.1 -1727.1
- hdi_for_year         1     408.2  5855.1 -1479.2
- factor(sex)          1    2479.8  7926.6   703.9
- factor(age)          5    6746.9 12193.7  3799.9
```

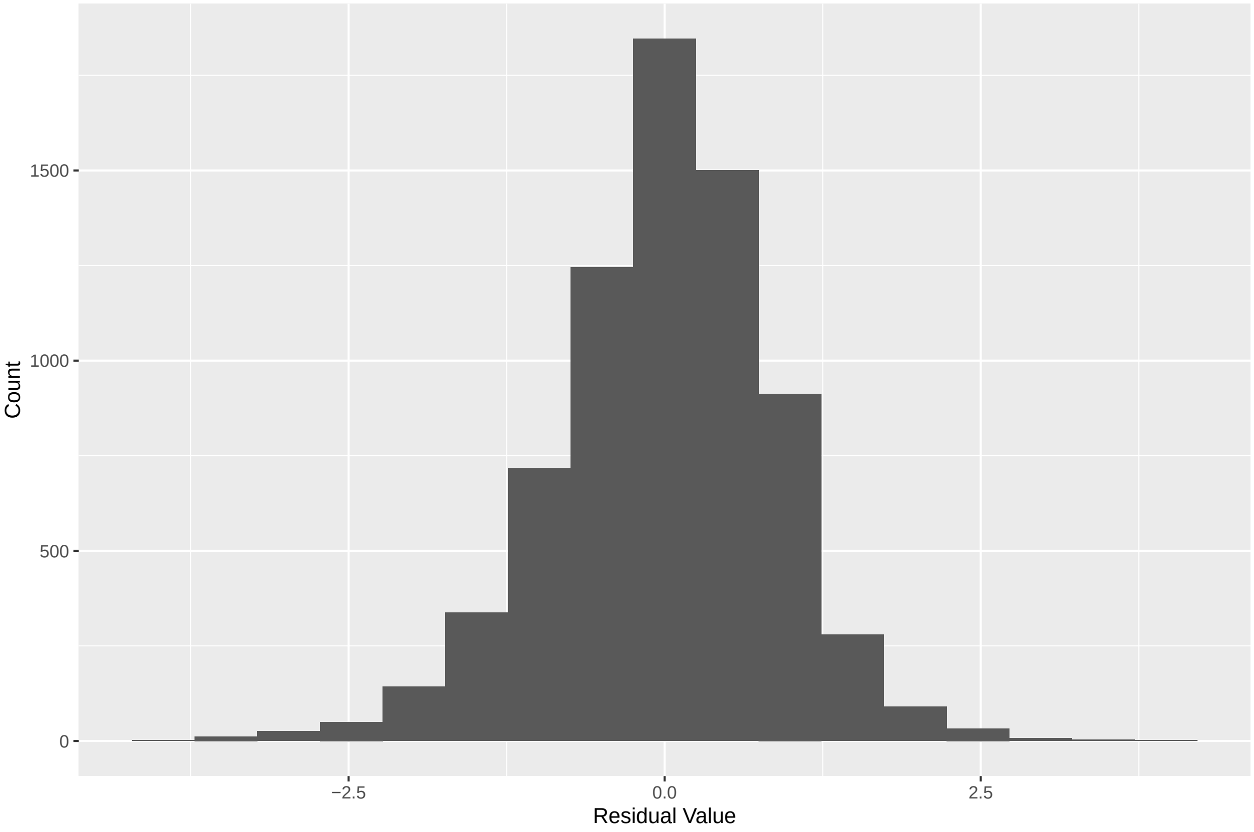
```
# A tibble: 10 x 5
   term                     estimate  std.error statistic   p.value
   <chr>                       <dbl>      <dbl>     <dbl>     <dbl>
 1 (Intercept)             0.144     0.0878          1.64 1.01e-  1
 2 factor(sex)male         1.17      0.0205         57.2  0.       
 3 factor(age)25-34 years  0.192     0.0345          5.58 2.52e-  8
 4 factor(age)35-54 years  0.339     0.0343          9.90 5.71e- 23
 5 factor(age)5-14 years  -2.41      0.0372        -64.6  0.       
 6 factor(age)55-74 years  0.467     0.0346         13.5  7.32e- 41
 7 factor(age)75+ years    0.802     0.0355         22.6  4.61e-109
 8 year                   -0.0000431 0.00000341    -12.6  3.25e- 36
 9 hdi_for_year            6.28      0.271          23.2  3.97e-115
10 log(gdp_per_capita)    -0.348     0.0209        -16.7  2.98e- 61
```

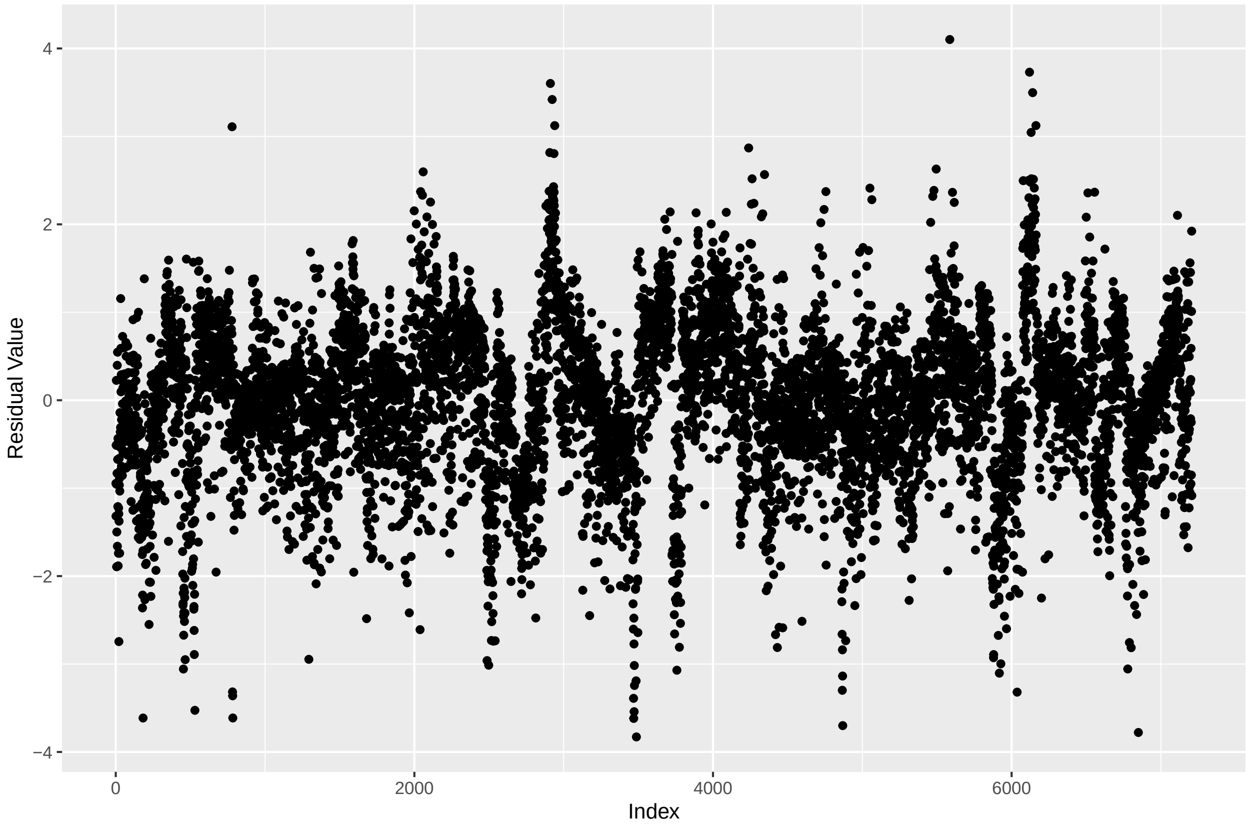
```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic p.value    df logLik    AIC    BIC
      <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl>  <dbl>  <dbl>
1     0.640         0.640 0.870     1422.       0    10 -9217. 18456. 18532.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

$$log(\widehat{suicides}/100k) = sex + agegroup + year + hdi + log(gdp/capita)$$

This model is a much better predictor. The adjusted Rˆ2 value is .6396, which means that the model accurately predicts about 64% of the variation in the log suicide rate globally. Note that this model was created by data that had to filter out the 0s in order to create a better log model. In this way, countries with suicide rates of 0 in a particular year are not able to be accurately predicted using this model. We can interpret all of the coefficients in the model. For example, men will have a log suicide rate that is 1.17 higher than their female counterparts, on average. The model also suggests, that on average the log rate of suicides has decreases slightly over time because the coefficient of the year is negative. The model as is, is based on females age 15-24, however, we can use the coefficients given to change the model to predict the suicide rate of indivisuals with very different characteristics.


4
2
0
−2
−4
Residual Value
−2
0
2
4
Predicted log Suicide Rate per 100k

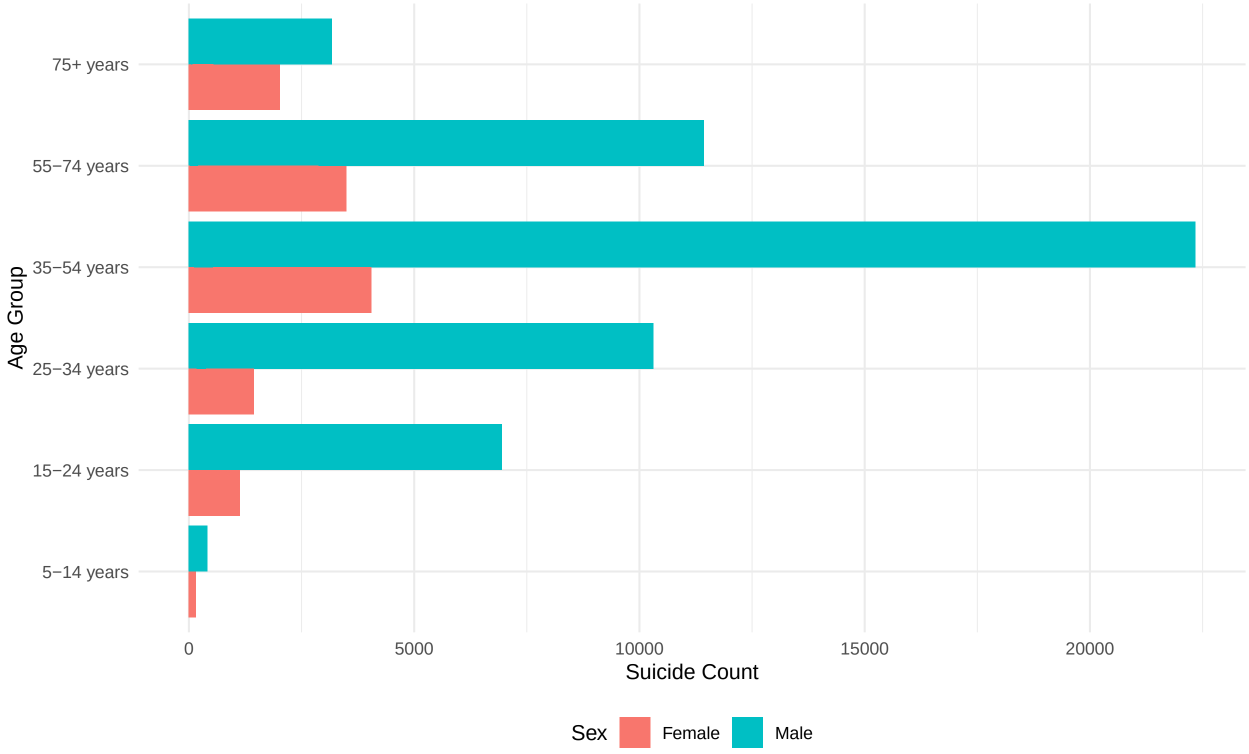

1500
1000
500
0
Count
−2.5
0.0
2.5
Residual Value

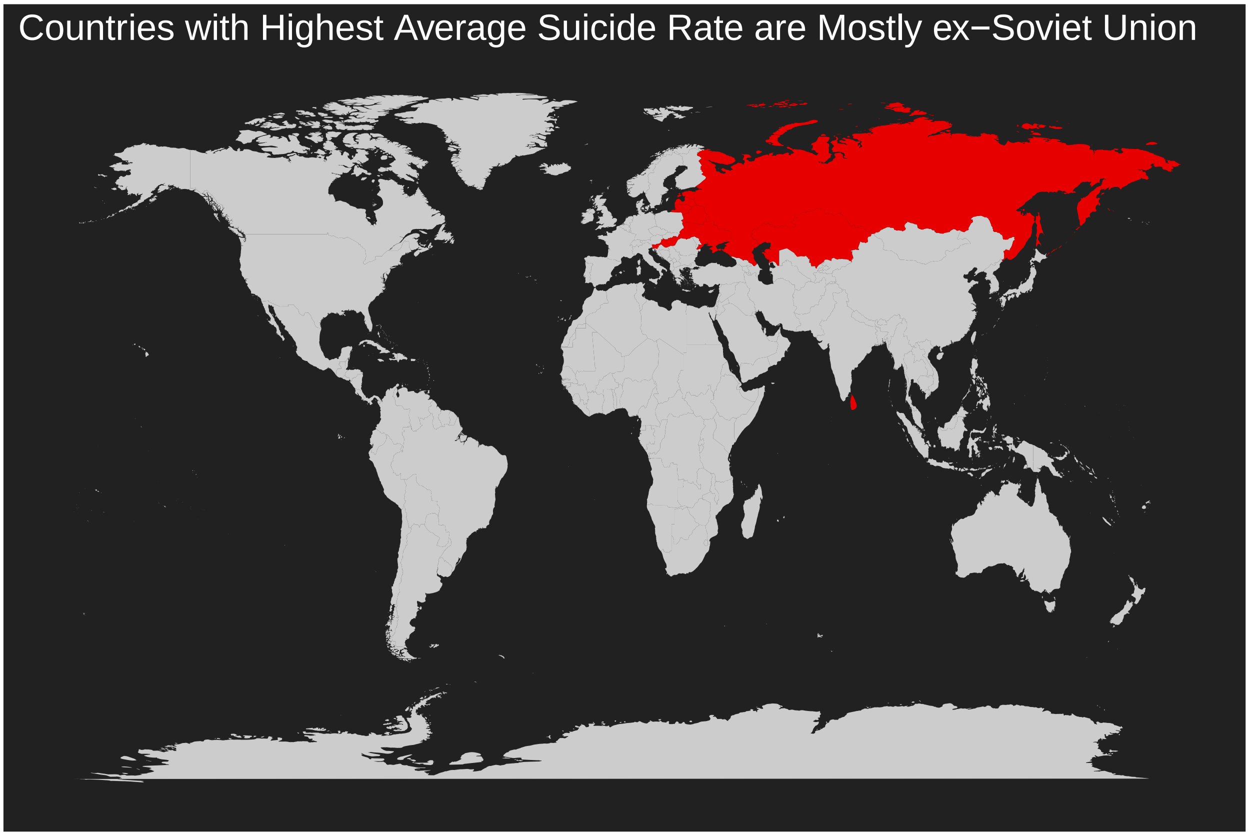


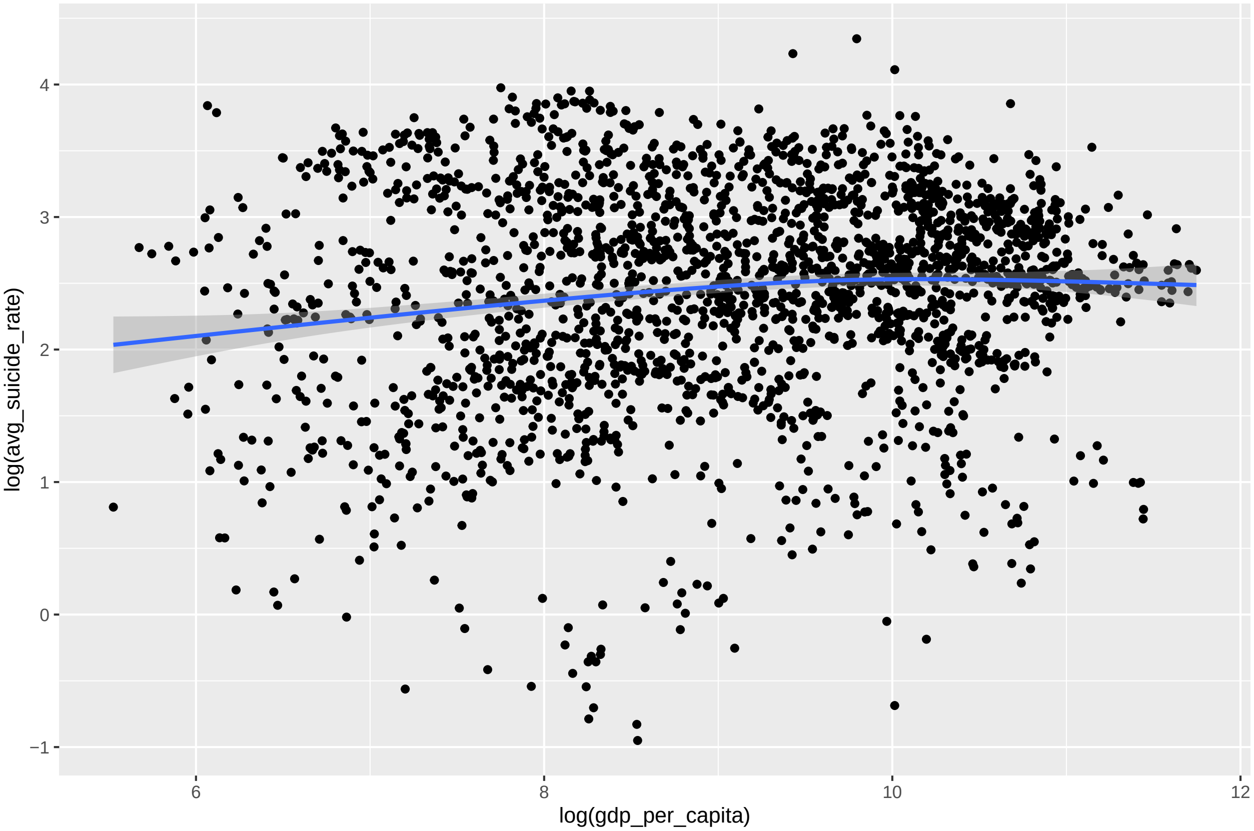
When checking linearity with the log model, it is obvious that the residuals are not evenly spread out around 0 at every portion of the line. The residuals do make an approximately normal distribution. The residuals also do not look completely independent of one another; however, this was not unexpected as the data is time series and therefore the suicide rates are not completely independent.

Create better Graphic

Males Age 35−54 Commit Suicide the Most and Males are 3.3 times more likely than Females to Commit Suicide at any Age




Countries with Highest Average Suicide Rate are Mostly ex−Soviet Union



References

Coor_flip() https://ggplot2.tidyverse.org/reference/coord__flip.html

Converting to date https://stackoverflow.com/questions/30255833/convert-four-digit-year-values-to-a-date-type

Map https://www.r-bloggers.com/how-to-make-a-global-map-in-r-step-by-step/