Multi-Dimensional Study of Suicide Trend from 1985 to 2016

Lang Qin, Jasmine Xie, Wenjin Lyu

Research questions:

1. How does geographical location correlate to the suicide rate?

   Result: Russian federation, Japan and United States have a considerably high suicide rate.

2. Does suicide happen more in young people, middle-aged people, or old people?

   Result: We found that suicide rate happend more in middle-aged people with a range from 35 to 54 years old from the graph "The distribution of suicides by age group"

3. How does suicide number grow in United States? In future?

   Result: We predicted the suicide number is 38511 in United States in 2016 (where dataset ends), and actual data we found in public sources is 44965 in year 2016. And we predicted that 41431 in 2018, while the actual number is 48344.

Motivation and background:

Suicide is a reaction of a mixture of extremely intense multiple negative feelings, which involves voluntarily taking one's own life. In analyzing this dataset, we hope to uncover the trends that could raise the importance of mental health and better prevent suicide from happening in the future. By having insight into how the suicide rate change in different ages, genders over years, and how a country's GDP and location affect suicide rate, we hope every country or region can develop better suicide prevention for specific group of people with different ages and genders. With deep insight of reviewing the data, we also can break the assumptions and misunderstandings while we first came up with this topic. Hopefully through this data analyses, every region can decrease the suicide rate by a large amount and start more thorough suicide prevention program.

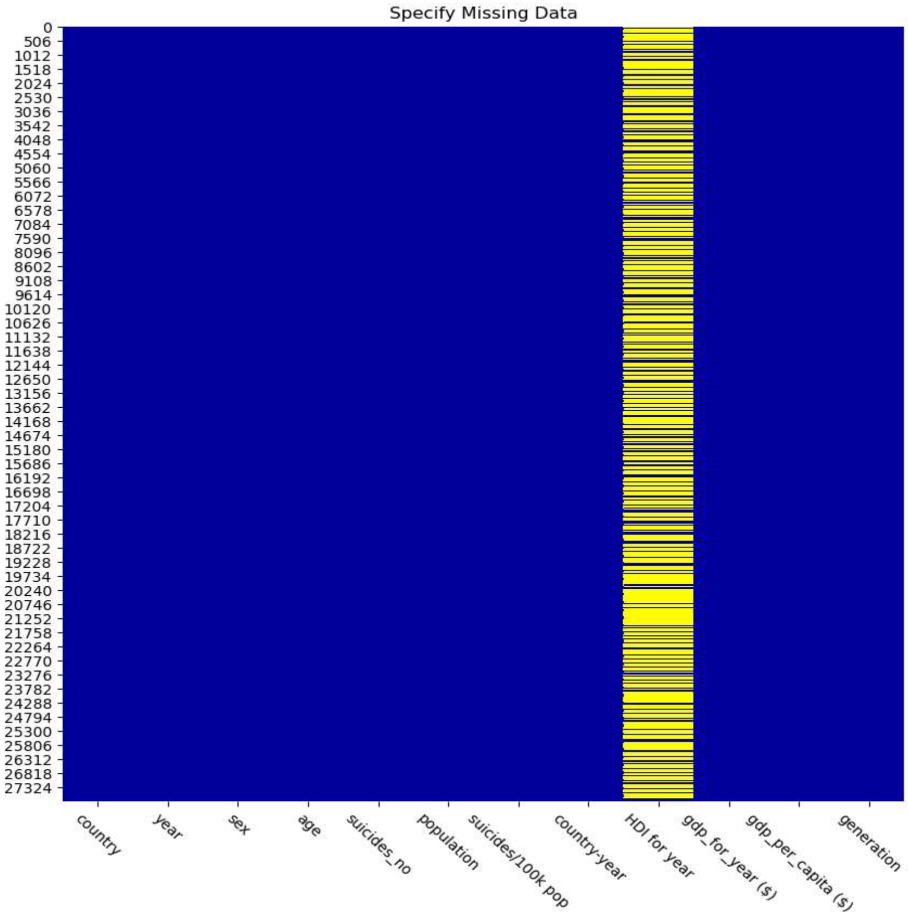
Dataset:

We are going to use a dataset which contains suicide rates overview from 1985 to 2016. The dataset includes columns of country name, year, sex, age, suicides number, population, suicides per 100k population, country at year, HDI for the year, and gdp for the year. This compiled dataset pulled from four other datasets linked by time and place was built to find signals correlated to increased suicide rates among different cohorts globally, across the socio-economic spectrum. We will use this dataset to analyze the suicide rate from multiple dimensions or factors.

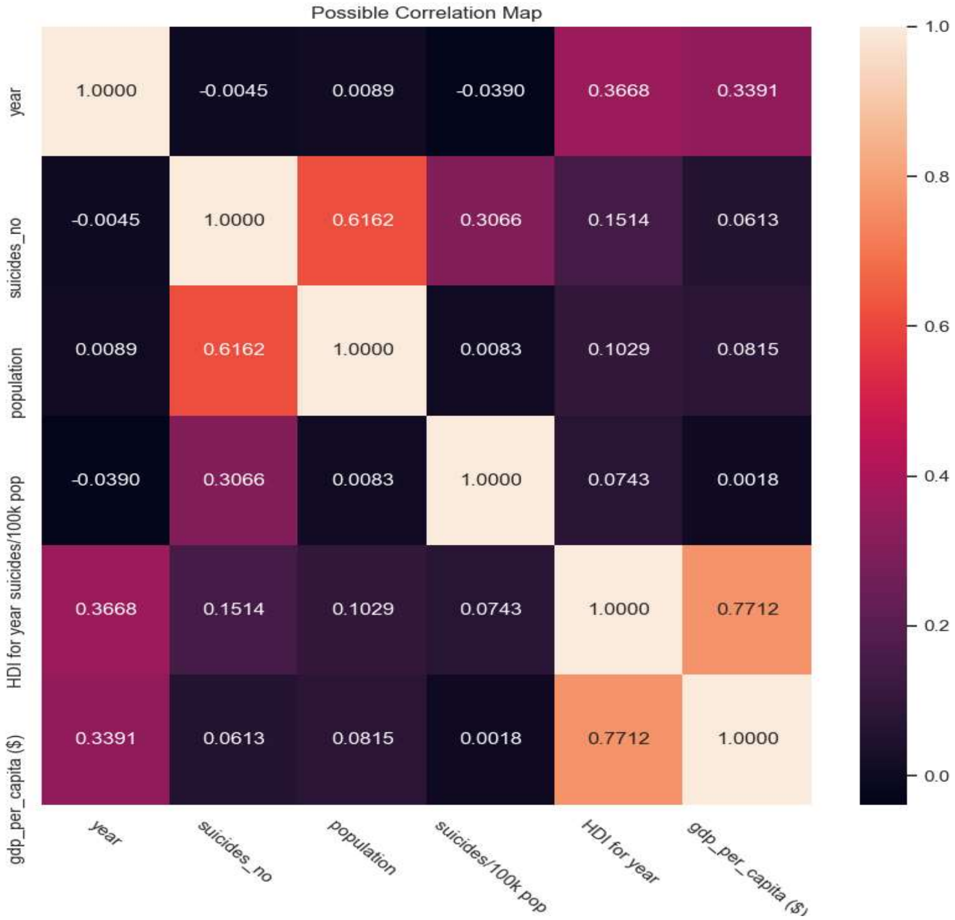
URL: https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

Methodology:

First we did some Initial computations, we tried to find the correlation between columns. In order to see how much data is missing. We used seaborn to draw a graph, and save the graph with the title as “Specify Missing Data”. We found that there is a large amount of missing data in the column “HDI for year”.



Next, we try to find the correlation of each column in the dataset by corr method and see what the relationship between these columns is. We used a heat map to demonstrate, with the title “Possible Correlation Map”



With reviewing the “Specify Missing data”, HDI for year data is the only column with missing values. Moreover, from the “Possible Correlation Map”, the correlation between HDI for year and gdp_per_capita is 0.7712, which shows the high correlation. If we use both columns, the information is redundant for us to analyze. Therefore, we choose to remove the HDI for year column instead of making predictions for those lost values. Since original dataset is stratified by age groups and gender, while we want to see the growth of suidcide number based on countries over year trend, we aggregates the suicide numbers of different groups as total number for each country.

After seeing the correlation of each column and cleaning the dataset, we start on graphing and visualization to answer our questions.

1. In the first question, we want to see the correlation based on the general geographical locations. Besides the original dataset we have, we also use geographical coordinates data from geopandas in order to visualize our data. While merging the data, we find that our original dataset and geospatial data set have different country names for the same country, so we changed the different country names to a standard names after looping over the whole list. Through the world map, we can see the color changed in different countries and whether they are consistent with their neighborhood, so it can tell the correlation among the geographic regions instead of simply countries.
2. For the second question, we were only focusing on the suicide rate over the different age groups. However, our data is stratified by countries and genders, so we then modify our dataset by aggregating in ages and years. Therefore, we can see how the different age groups are growing over the years, and which group is the largest number of suicide at each year.
3. We intended to make predictions by using machine learning. Before training the model with machine learning, we first want to have better understanding of how ages, GDP per capita, and gender correlate to the suicide in United States, because we found these three elements might corelate differently with the suicide number in different countries. We

splitted the United States data solely, and we visualized those three elements. And then, we began to build our machine learning model. In this part, we also create two machine learning models with two different tools. Basic one is using the linear regression model that we learned in the class, and the Advanced one is using the K neighborhood regressor looping over each United States data we have collected and find the least error k to predict 2016 suicide number of United States.

Challenge Goals:

1. Machine learning:

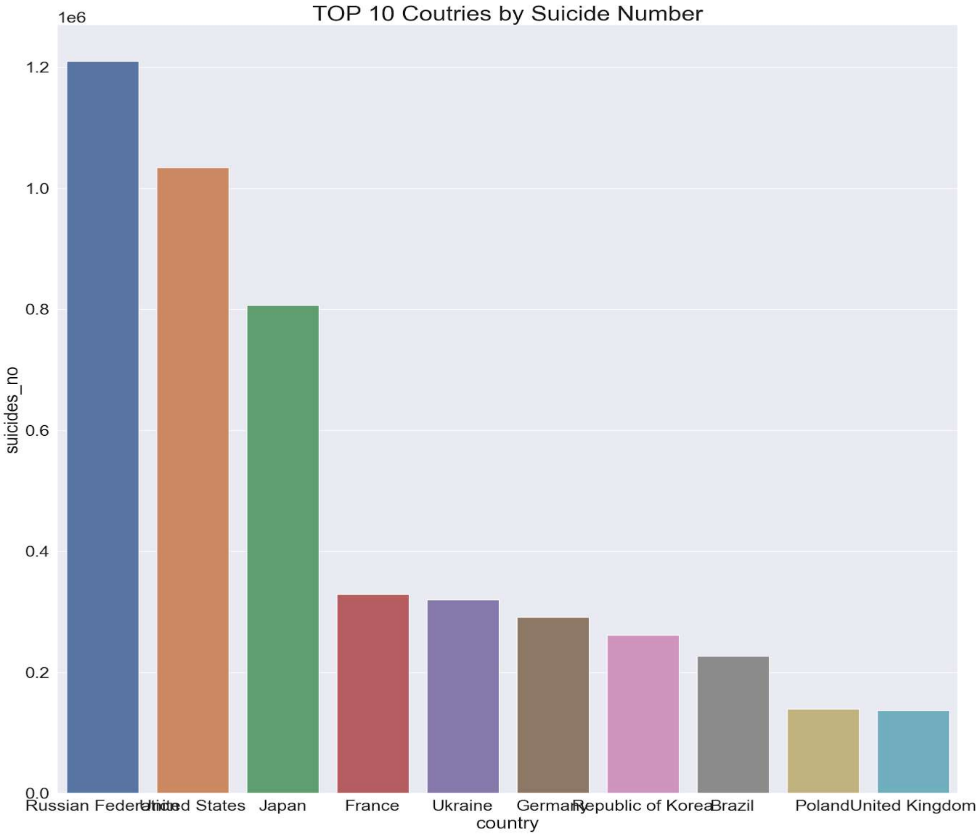
   We will use machine learning to predict the trend of future suicide rate. It will look at various models, for example the trends for each of the year, and how the trend changes over time in order to predict the future. To accomplish this goal, we use two ways to train the machine learning models, linear regression and k neighborhood regression. The second one is the new method we learned to train the model by computing the lowest root mean square error with trying out different k values.

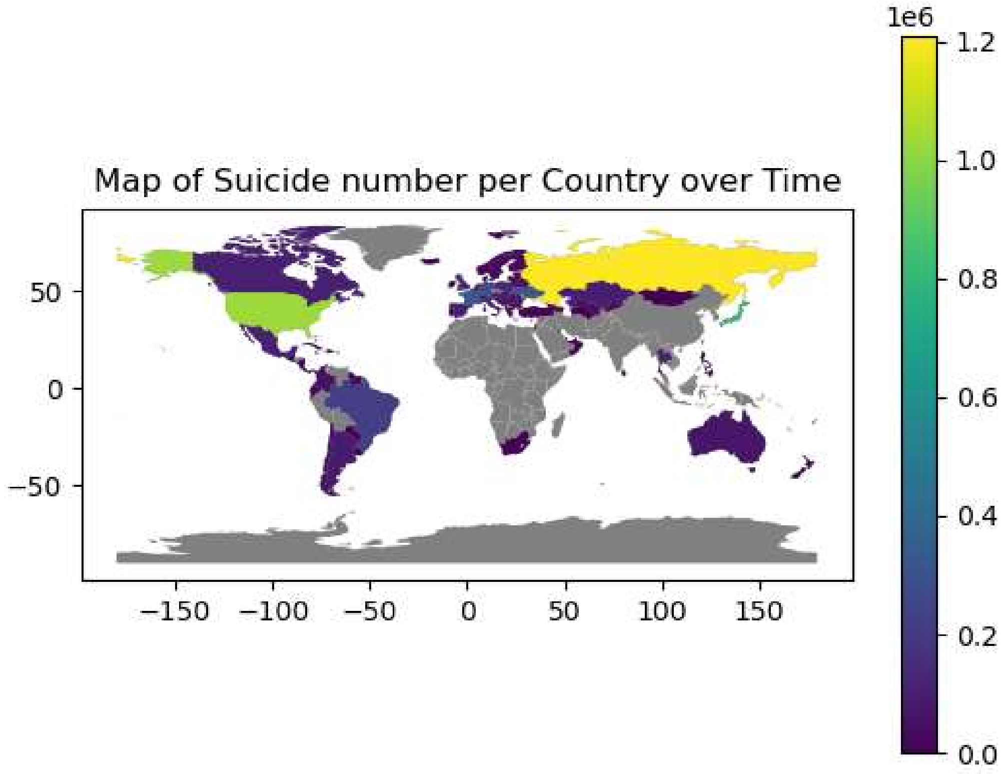
2. Pre-process data (Messy data):

   The dataset is not perfect as we expected. It was with missing values, lack of the geospatial information we want to use, and it is well stratified but most of the time, we have to aggregate columns over and over. We think we accomplished this goal, while we found a way to clean the data, merged original data with geographic coordinates, and created multiple data frames for later visualizations.

Result:

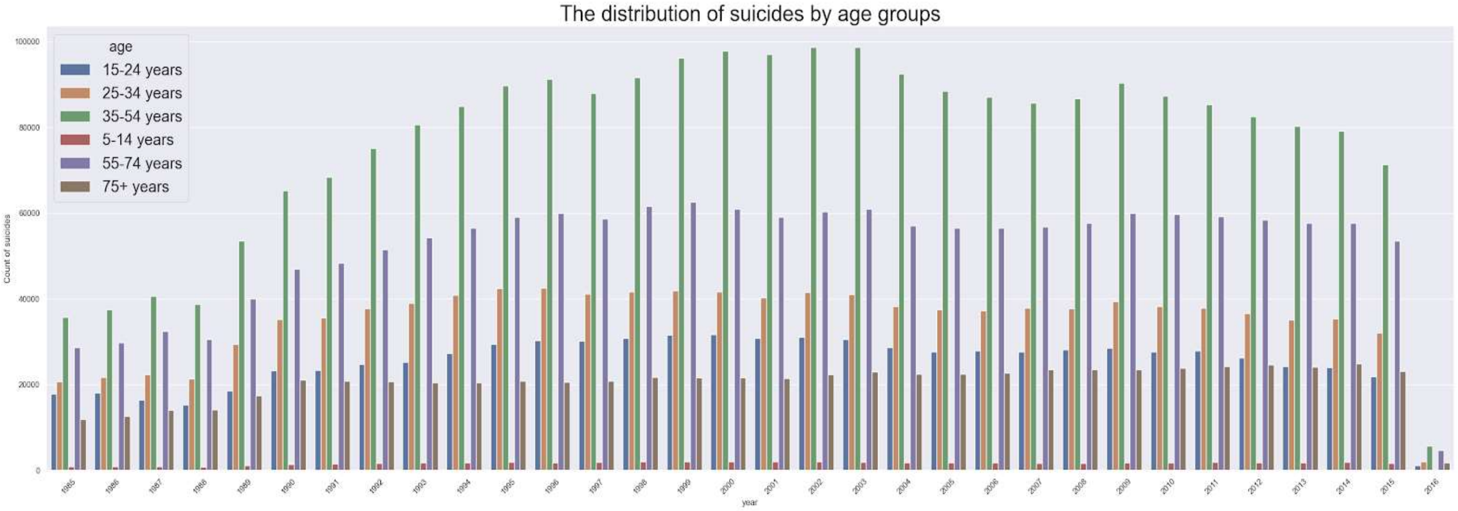
1st question:





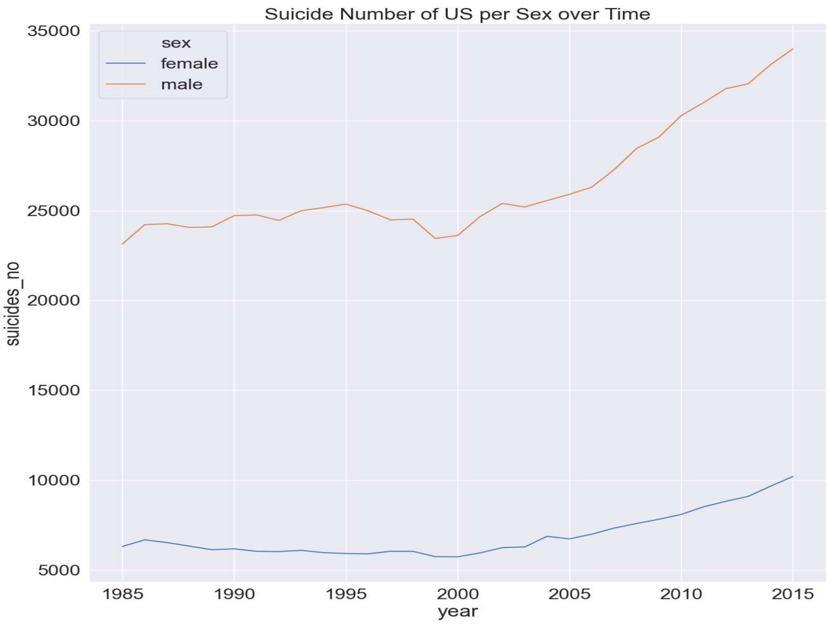
For the first research question of our project, we found that in Europe and part of Asian continent have a very high suicide rate, which is around Russia area. We also found that North America continent has a high suicide rate, which is in the area of the United States. The other areas either don’t have data or have a comparably equal suicide rate. Our initial hypothesis was that in undeveloped geographical areas, there will be higher suicide rate, and the developed countries will have lower suicide rate. The “Map of suicide number per Country over Time” showed us that the dataset is not in favor of our hypothesis, because if we look at some specific countries like Japan, United States, and Russia, we noticed that the suicide rates in these developed and developing are actually higher than other areas. From this map, we can see that the suicide rate is more related to the developement state of a country. And we can see living in a developed country is possibly with more pressure from life.

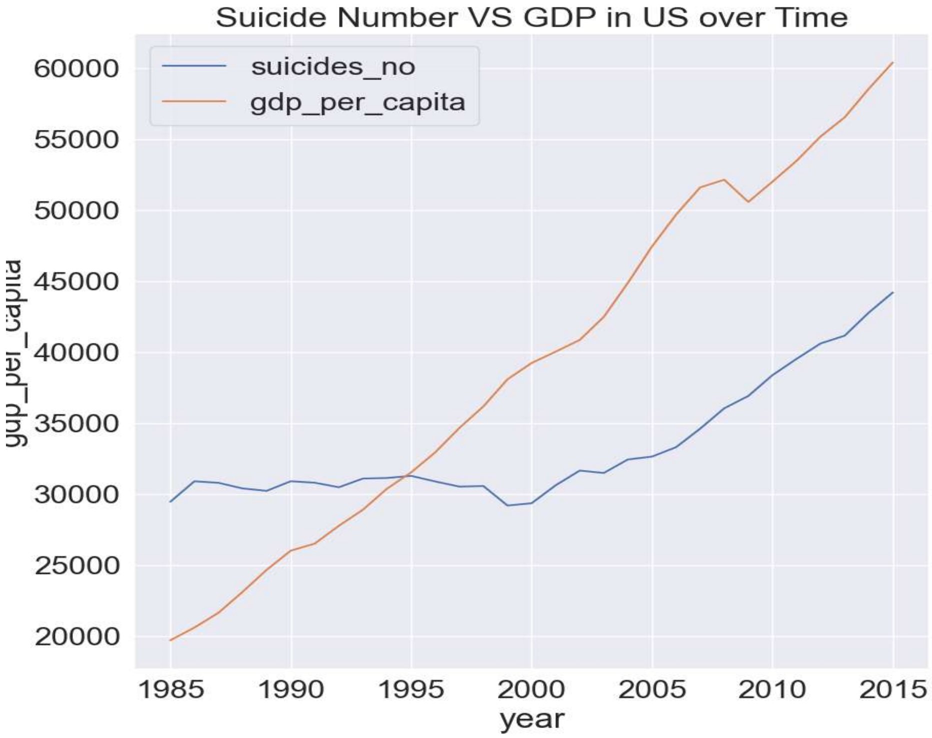
2nd question:



For our second research question, our initial thought was that suicide rate is higher among teenagers. Our graph "The distribution of suicides by age groups" shows that 35-54 years old people have the highest suicide rate. Then it is followed by the 55-74 years old age group. The rank is shown exactly the same each year after we aggregated the data from all countries and both genders. This result indicates the middle aged and the elder are more frequently committing suicide, while the 5 - 14 year have the lowest suicide number. We can try to interpret this from a psychological perspective, that suicide committing induced by stress. The middle ages and older people usually experience more stress rather than teenagers and children. Hardship and complexity are more pushing toward them.

3rd question:

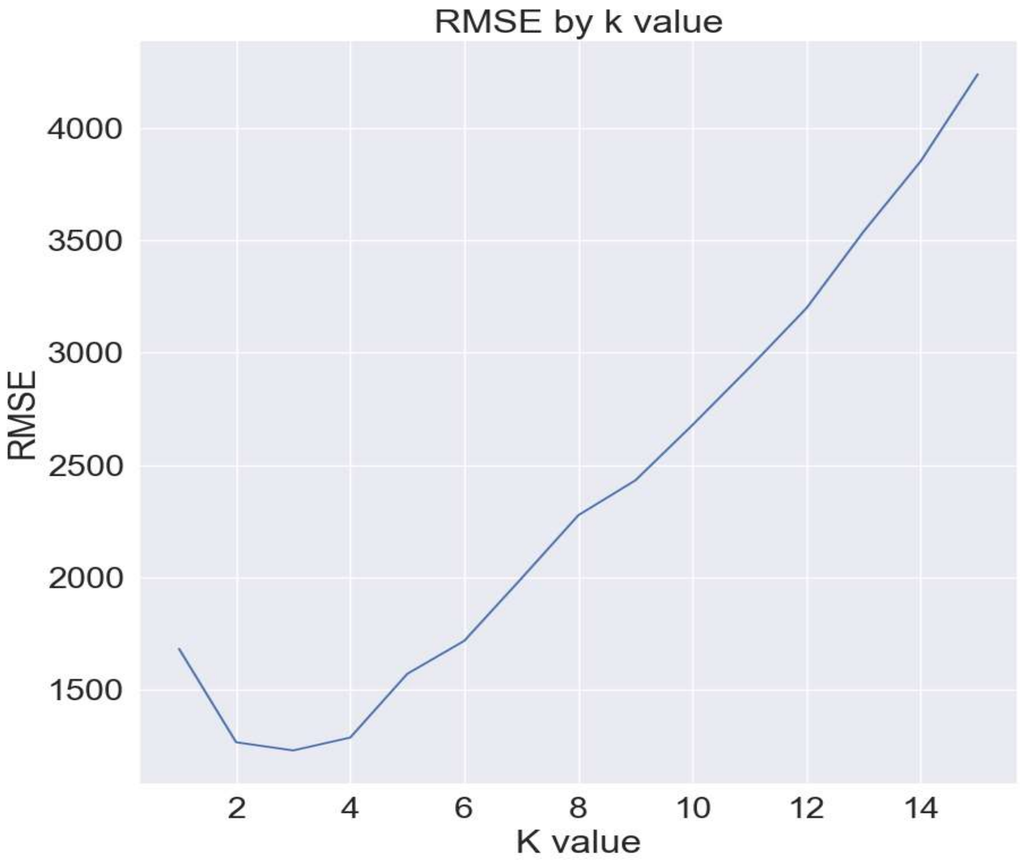




We first investigated data of the United States into further depth. Compared the suicide number for male and female of United States, the suicide number of the male is distinctly more than the number of the female. There is an approximate difference of 20000 between the male and female with growth over 1985 to 2015, which suggests that the male may experience more life stress or may be potentially more impulsive than the female in the United States.

In the second graph, *Suicide Number VS GDP over Time*, it also illustrates the correlation between GDP and the suicide number that we had suggested before. Specially through year 2000 to 2015, suicide number growed highly positive with the growth of GDP per capita.

Then we began to make our machine learning. We used the linear regression model to fit all features and produce the predictions. The flaw of this model is we need many features to give a prediction and it can be overfitting sometimes. Hence, we want to have an advanced model to predict the value. We learned about the K Neighborhood regressor. By trying out different k values, we find the best root mean square error is when k = 3.



With this k value, our model predict the suicide number of year 2016 is 38511, and actual data we found in public sources is 44965 in year 2016. And we predicted that 41431 in 2018, while the actual number is 48344.

Work Plan:

1. Check and Clean Data (Predicted: 1 Hours)
    a. Read in csv file and build dataframe.

b. Check distribution of data.

c. Fill or remove missing values.

2. Make some initial Computations (Predicted: 1 Hours)

    a. Check correlation among columns with pair plots.

    b. Plot growth of suiside numbers for some countries of interest.

    c. Plot growth of GDP for some countries of interest.

3. Train models to predict suicide numbers in future (Predicted: 6 Hours)

    a. Split into a train and test set.

    b. Train model with possible features with strong correlation.

    c. Repeat with different hyperparameters, check both train and test accuracy scores.

4. Tests (Predicted: 3 Hours)

    a. Do some sanity checks: Are there some outliers? Do we need to change the train and test set in case that we are overfitting? Is the result reasonable for our hypothesis?

5. Write report (Predicted: 6 Hours)

    a. Analyze the result and try to answer the question we have.

    b. If there are some questions we can not answer, list the reasons.

    c. Draw a conclusion.

To work coordinately:

Jasmine Xie:

- Checking and cleaning data.

- Checking Lang Qin’s initial computations.
- Writing the final report.

Lang Qin:

- Making some initial Computations.
- Checking Wenjing Lyu’s train and test set.
- Checking final report

Wenjing Lyu:

- Splitting into a train and test set.
- Checking Jasmine’s data cleaning
- Checking final report

The rest of training models, predicting, and tests will be done with pair programming. The three of us will be taking turns to be the driver and the navigator.

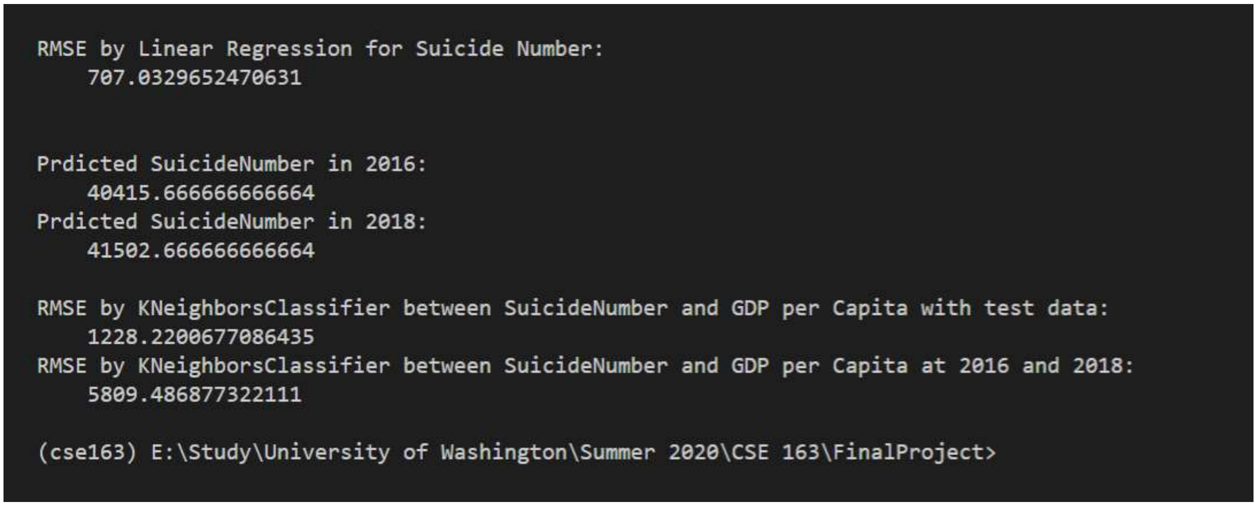
Work Plan Evaluation:

Most of our plan is pretty accurate. We had a good outline for what to do first and what comes next. However, there is one part that we did not expect which is the checking and cleaning data part. We spent more hours than expected, because we did not pay attention to the names in shape file and csv file before starting. Some of the country names in one file do not match the ones in the other file. For example, we have “United States” in one file and “United States of America” in another file, which costs some challenge for us to merge the data. The next thing that we did not have and right estimate is the machine learning part. We were not thinking ahead about what to use and how to do the machine learning part ahead of time, so that part takes longer than we expected. The estimate on the collaboration part is actually really accurate, the three of us spent a lot of time on zoom doing pair programming and checking each other's work. The estimate of visualization and plotting is also very accurate. We had a decent idea on how much time and effort does it take to do plotting. The report writing part is better than we expected, we planned to write the report after doing the coding part, however, we did the coding with the report at pretty much the same time. The overall work did take longer work than we expected, it is also harder than we imagined.

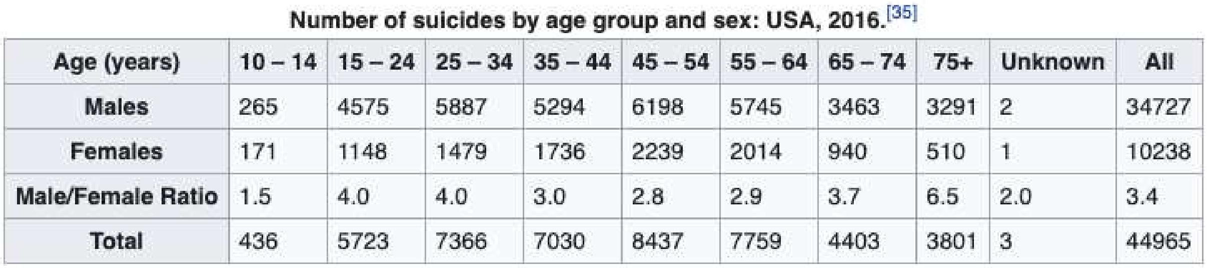
Testing:

To test our machine learning, we used our predicted values to compare the real world data. For example, our dataset is only from 1985 to 2016(year 2016 has missing suicide number), so we predicteed the suicide number for year 2016 and 2018. Then we finds the the real suicide number for year 2016 and 2018 from https://afsp.org/suicide-statistics/ and https://en.wikipedia.org/wiki/Suicide_in_the_United_States. The results are not exactly the same as the real world results; however, we are in a reasonable range of suicide rate for both 2016 and 2018.

We also used mean square error to check our predictions. The graph below shows our two model's root mean squared error.

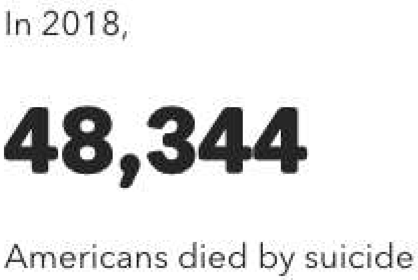


```
RMSE by Linear Regression for Suicide Number:
    707.0329652470631


Prdicted SuicideNumber in 2016:
    40415.666666666664
Prdicted SuicideNumber in 2018:
    41502.666666666664

RMSE by KNeighborsClassifier between SuicideNumber and GDP per Capita with test data:
    1228.2200677086435
RMSE by KNeighborsClassifier between SuicideNumber and GDP per Capita at 2016 and 2018:
    5809.486877322111

(cse163) E:\Study\University of Washington\Summer 2020\CSE 163\FinalProject>
```



Number of suicides by age group and sex: USA, 2016.[35]

Age (years)	10 – 14	15 – 24	25 – 34	35 – 44	45 – 54	55 – 64	65 – 74	75+	Unknown	All
Males	265	4575	5887	5294	6198	5745	3463	3291	2	34727
Females	171	1148	1479	1736	2239	2014	940	510	1	10238
Male/Female Ratio	1.5	4.0	4.0	3.0	2.8	2.9	3.7	6.5	2.0	3.4
Total	436	5723	7366	7030	8437	7759	4403	3801	3	44965



In 2018,

48,344

Americans died by suicide

Collaboration:

This project is done coordinately by Lang Qin, Wenjin Lyu, and Jasmine Xie. We received no additional assistance from people besides the course staff and our group mates. We used several online resources. We looked at some examples from https://seaborn.pydata.org/tutorial.html to help us on the visualization and plotting part. We also searched at some examples from https://stackoverflow.com/questions/32751229/pandas-sum-by-groupby-but-exclude-certain-columns to help us with aggregating columns and merging different dataset.