Group: 19

Members: Kelvin Lao, Bobo Tam, Xin Yang

Github Link: https://github.com/bobotamm/data_science_pipeline.git

Introduction and motivation

Since mental health has emerged as an important health problem that would require attention, it is important to study how suicide rates develop across the globe. By understanding the underlying fundamentals, researchers might be able to develop solutions that could improve mental hygiene.

In our project, we would like to explore how the suicide rates changed over the course of the past few decades by breaking them down into different genders, age groups, and generation groups. In addition, we would like to examine whether the financial prosperity of a country (GDP) plays a role in affecting the rates of suicide. Based on our result, we would be able to provide a reference for predicting future suicide rates, and along with further research, we might be able to know the possible correlated variables contributing to those changes.

Dataset

Our dataset (Suicide Rates Overview 1985 to 2016 from Kaggle) is compiled from four other datasets linked by time and place and was built to find signals correlated to changes in suicide rates among different cohorts globally, across the socio-economic spectrum.

The entire dataset contains 28,000 samples and information on global suicide rates from 1985-2016. We used the country, year, sex, age, suicides/100k pop, generation, and gdp_for_year from the dataset in our project. Specifically, suicide/100k pop and year would be used to explore how the suicide rates changed over the past few decades. Gender, generation, age, and gdp_for_year would be used when we further break down into different groups to study the correlation between those groups and the changes in the suicide rate.

Data Cleaning

The dataset we used is already preprocessed and many of the items are clean. There are several different data types and formatting that we changed in order to better perform exploratory data analysis. We performed the following data cleaning (some are done during the process of data manipulation and analysis):

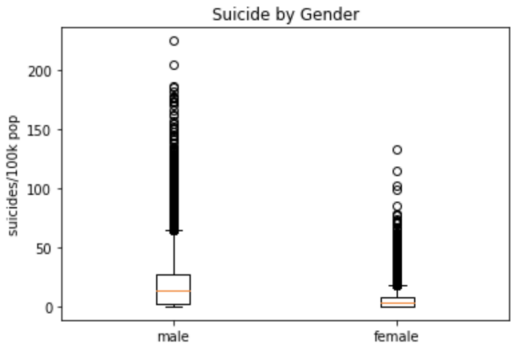
- Converted GDP from string to integer for easier data comparison and plot in the graph.
- Renamed '($)' in the GDP and gdp_per_capita to easily access the columns.
- Remove HDI because most of the data is missing and we did not use it.

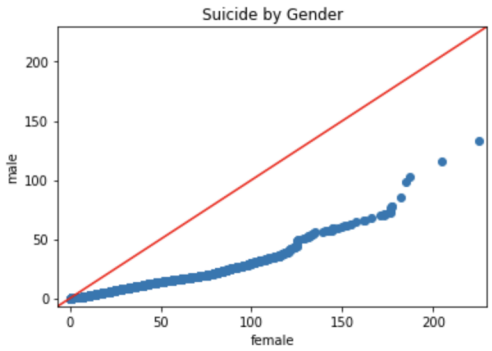
- Created a data frame with a new column named “suicide rate” for mean suicides/100k pop
- Created a data frame with a new column that has mean gdp_per_capita. Converted age from string to integer so that it’s easier to compare age groups.
    - 5-14 (1)
    - 15-24 (2)
    - 25-34 (3)
    - 35-54 (4)
    - 55-74 (5)
    - 75+ (6)
- Figure out the year period for each generation. Change generation to year range of birth year to better conceptualize the data.
    - G.I Generation: 1901-1927 (1)
    - Silent generation: 1928-1945 (2)
    - Boomers Generation: 1946-1964 (3)
    - Generation X: 1965-1980 (4)
    - Millennials Generation: 1981-1996 (5)
    - Generation Z: 1997-2016 (6)

EDA

Suicide by Gender

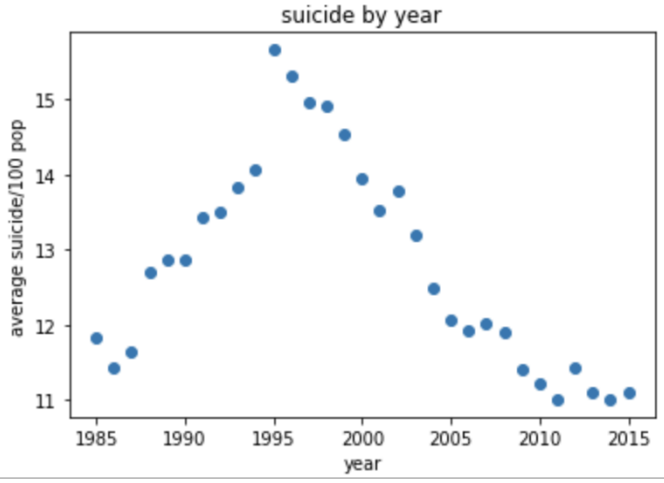
Box Plot and QQ plot of suicide by gender globally from 1985-2016. From this, it is clear that male suicides are relatively higher than female suicides. Using the mean function, the male suicide/100k population is 20.24 while the male suicide/100k population is 5.39. We will use this information to determine which gender group has a higher suicide rate and further study whether or not this difference has any effect on the suicide rate trend.



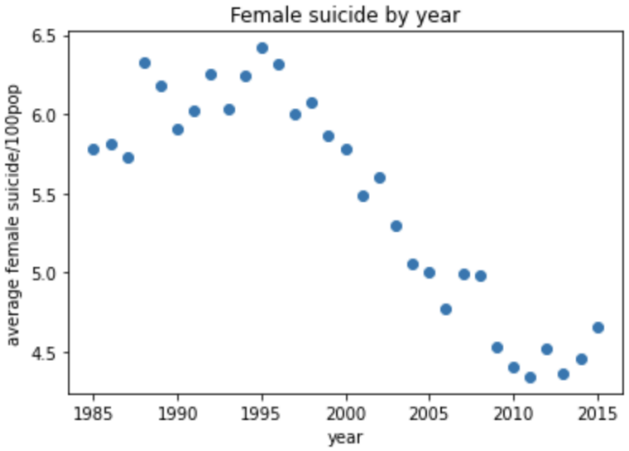


Scatter plots of suicide/100k by year, female suicide/100k by year, male suicide/100k by year, and U.S. suicide/100k by year. Overall, it is surprising that the male suicide rate scatters plot is very similar to the overall suicide rate. However, the female suicide rate scatters plot has a

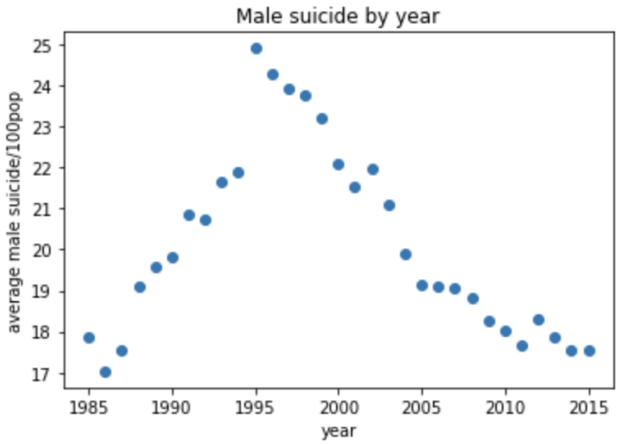
different trend. The U.S. suicide rate scatters plot is nearly opposite from the overall scatter plot. We want to use this information to explore how the suicide rate changed throughout the past few decades. By breaking it down into gender groups, we can further study how the suicide rates differ between the different two genders.



Since 1985, the suicide rate had an increasing trend up until 1995 and peaked there. After 1995, the rate has been decreasing.



The female suicide rate generally had been decreasing since 1995.

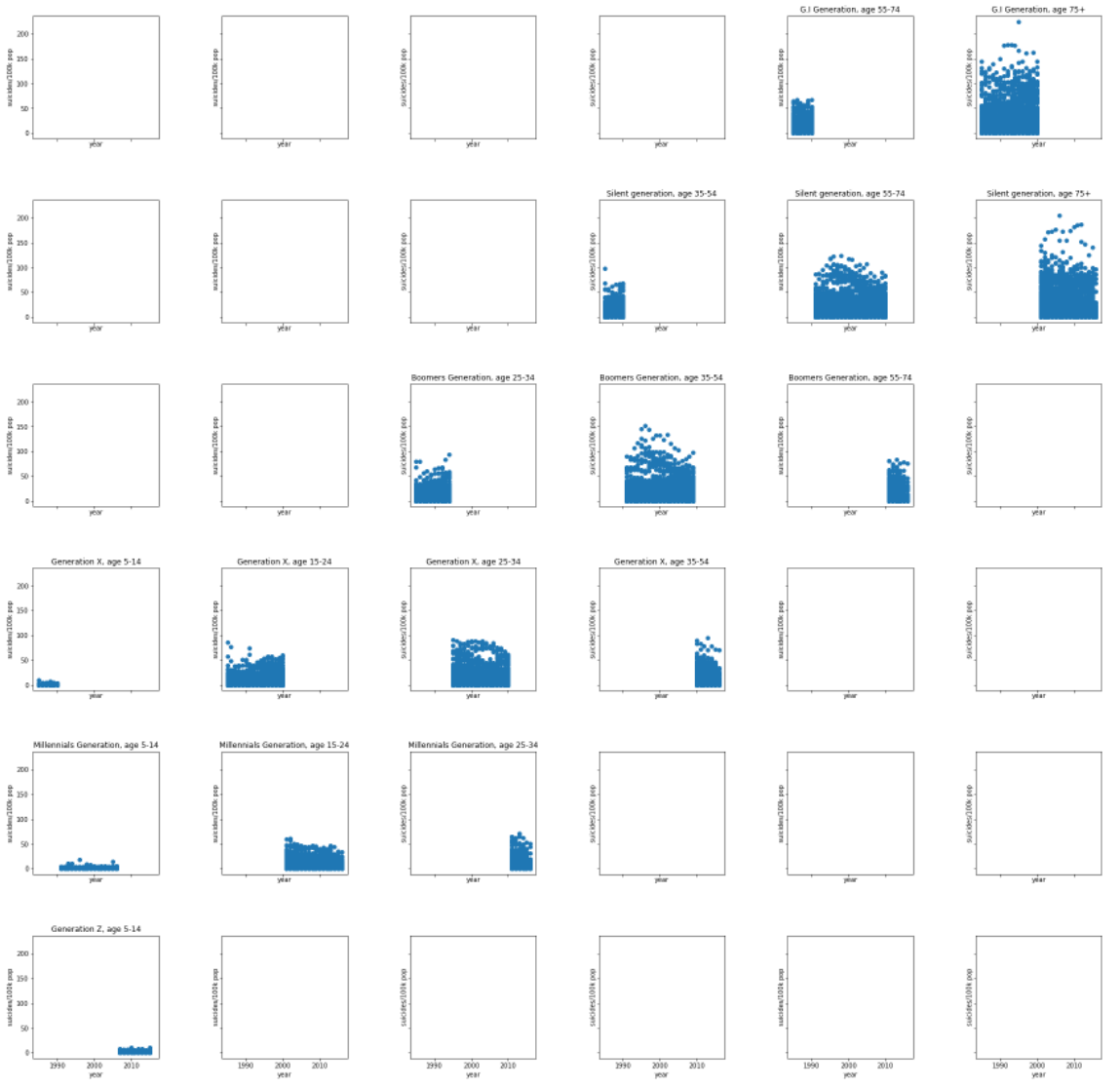


From 1985-1995, the male suicide rate had been increasing. It peaked in 1995 and since then, the rate has been decreasing

From these plots, we see that gender and suicide rates have a high correlation. Thus, we will be including gender as one of the features in our modeling section.

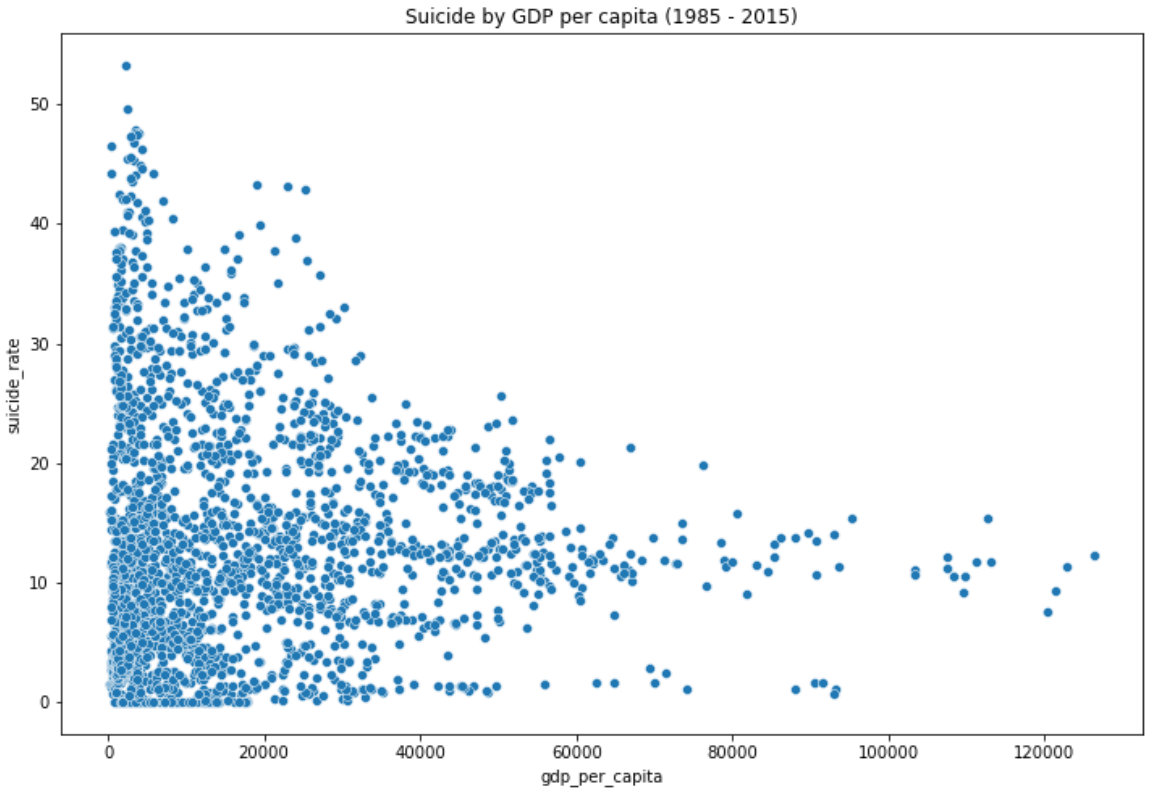
Suicide by Age Group

Scatter plot subplots for age and generation. The subplots are organized where the oldest generation is on top and the youngest is on the bottom and then the youngest age is on the left and the oldest age is on the right. The subplots show how generally, the older ages tend to have more suicides per 100k population each year. Older generations also seem to have a larger number of suicides per year but this may be a result of the age of the older generations being higher. There are a relatively small amounts of suicides for people in the range of 5-14 years old.

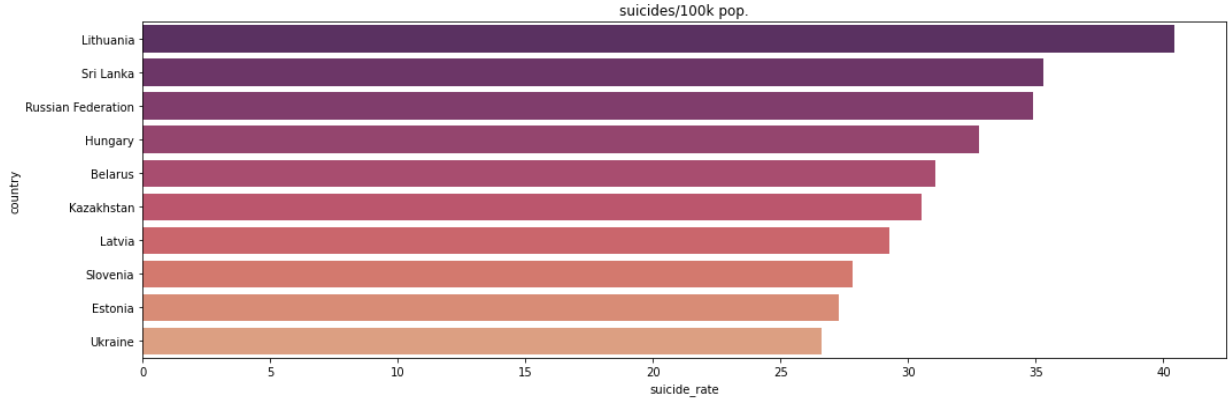


Suicide by GDP

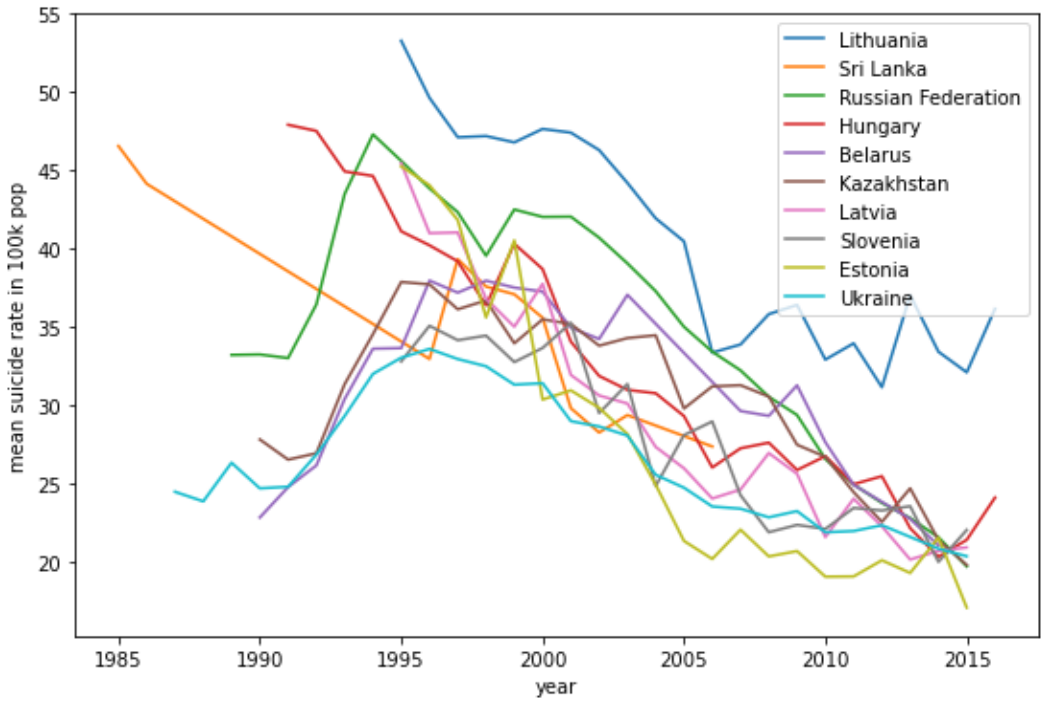
Scatterplot between the suicide rate and GDP per capita. Since both GDP and suicide numbers might depend on the population, we evaluate the GDP per capita and suicide number per 100k population (suicide rate) that can be more representative of the population. Based on the scatterplot, there appears to be a negative correlation between the two variables in the sense that as the number of GDP per capita ($) increases, the rate of suicide also increases. Compared to countries with low GDP per capita, countries with high GDP capita are experiencing lower suicide rates. However, there are also countries with both low GDP per capita and low suicide rates suggesting that GDP is not the sole factor that affects the suicide rate and there are other societal influences.

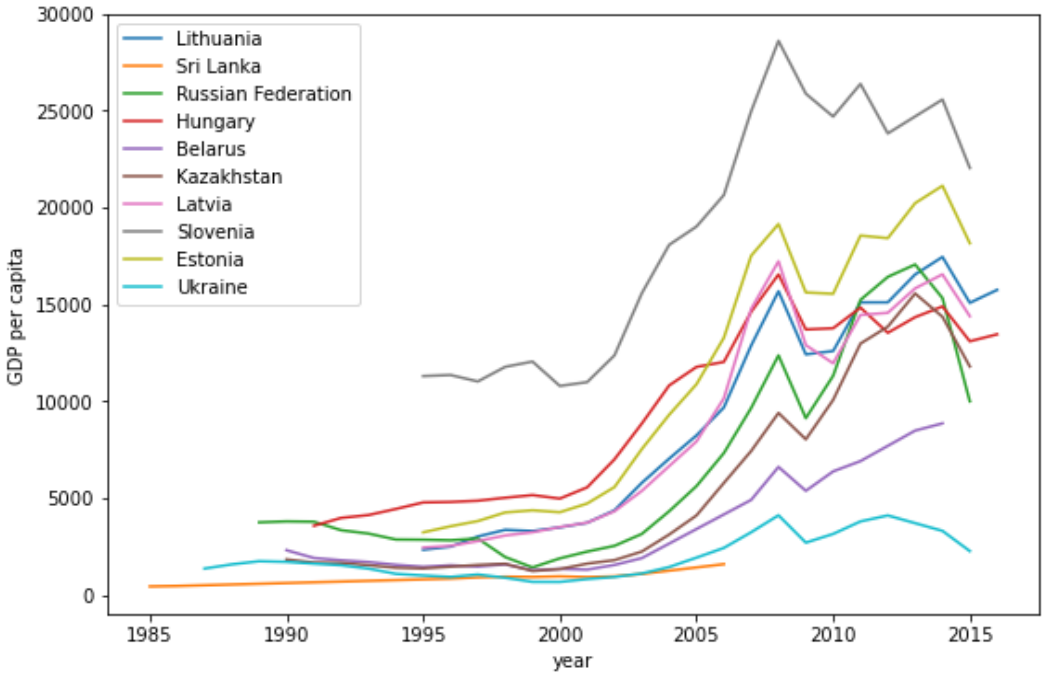


The bar chart shows the top 10 countries with the highest mean suicide number per 100k population. They are 'Lithuania','Sri Lanka', 'Russian Federation','Hungary', 'Belarus', 'Kazakhstan', 'Latvia', 'Slovenia', 'Estonia', and 'Ukraine'. The highest mean suicide number per 100k population is around 40.

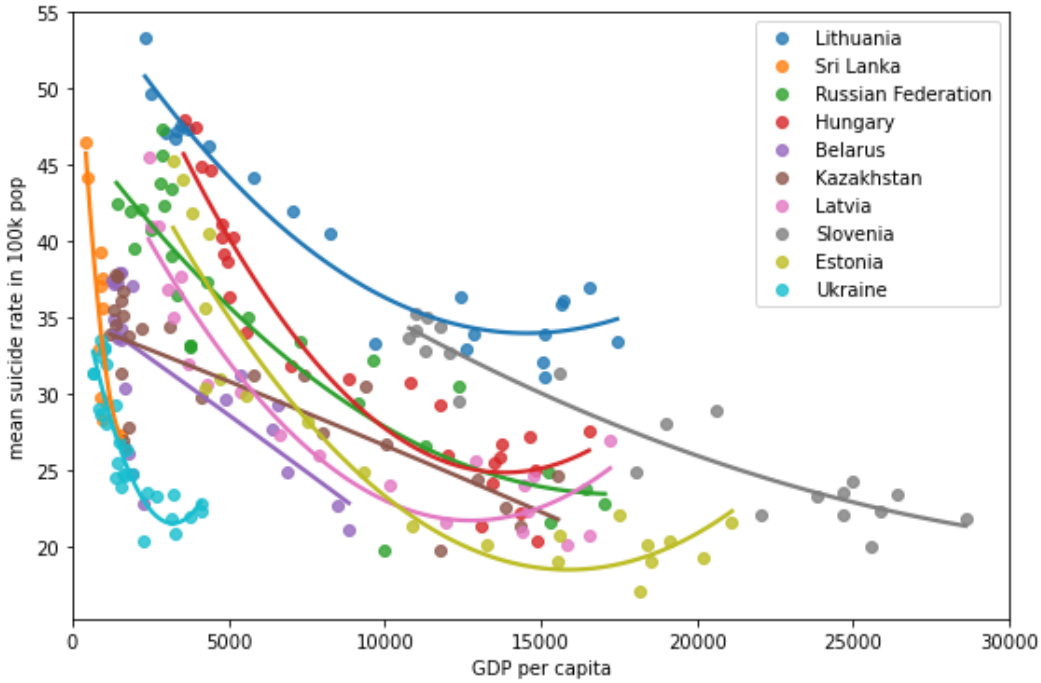


Based on the top 10 countries with the highest mean suicide number per 100k population, we evaluated how each country's GDP per capita and rate of suicide changed over the years between 1985 and 2016. The GDP per capita for all 10 countries gradually increased in the span of 31 years with a slight dip in 2008 when the great recession occurred. The rate of suicide follows an opposite trend as it gradually decreased over the years.





Regression plot between suicide rate and GDP per capita. To see clearly what happened between the GDP per capita and mean suicide number per 100k population for the 10 countries, we performed a regression plot on each country. Unsurprisingly, for each country, there is a negative correlation between the GDP per capita and the rate of suicide. This is indicative that as the GDP per capita in a country increases, the rate of suicide will likely decline.



Data Modeling

After completing the EDA we tried to include as many features as possible in our ML model to give us the most accurate model to predict suicides per 100k population.

Features: gender, generation group, age group, year of suicide, and GDP for the year

*** Generation group refers to the birth cohort born around the same time. For example, anyone born between 1981 and 1996 (ages 23 to 38 in 2019) is considered a Millennial.

Targets: Suicide/100k pop

Sample Size: 28,000

Feature Engineering: We cleaned all the data by changing the age range, generation, and gender which were already in bins to integers to use in our ML model. We originally tried no preprocessing on all our data but none of them ended up very accurate. Initially, we attempt to perform a linear regression model on the features and targets. In order to improve the feature of the model, we attempt to min-max scale all the features since linear regression is a model with smooth functions of input features and is sensitive to the scale of the input. Using min-max scaling will help reduce outliers and improve the accuracy of the linear models.

Relevance to Project: Our goals are to explore how the suicide rates changed over the course of the past few decades and we broke down into age groups, sex, generation groups in our EDA. We also performed EDA on the relationship between GDP and suicide rate. Based on the observation of the EDA, sex, age, generation, GDP all exhibit some correlation and pattern associated with the rate of suicide. Thus, in our machine learning models, we attempt to use these as features to perform regression on the rate of suicide.

Machine Learning Models

The goal of our data modeling is to predict suicide rates. Given that suicide rates is continuous data, we would need to perform regression instead of classification. Therefore, all of our machine learning models are regression models. The first three are linear regression and polynomial regression while the last two are ensemble regression with tree structure models. Linear regression and polynomial regression are approaches for modeling the relationship between the target value and exploratory variables, which would best approximate the point on the graph. The ensemble regression, however, takes on a different approach by reducing the error residual each time and combining models to improve prediction accuracy. We compared the different approaches and see how they best fit in our data modeling,

Linear Regression

- Parameters: Fit_intercept, n_jobs, normalize
- Scoring metric: Mean of CV with 3 cross-validation

- Model selection: Grid Search Cross-validation
- Best parameters: {fit_intercept: True, n_jobs: 3, normalize: True}
- Performance: 0.288
    - Data does not fit the model well

Polynomial Regression (Linear Regression)

- Parameters: Fit_intercept, normalize
- Scoring metric: Mean of CV with 3 cross-validation
- Model selection: Grid Search Cross-validation
- Best parameters: {fit_intercept': True, 'n_jobs': 0, 'normalize': True}
- Performance: 0.2987
    - Data does not fit the model well

Polynomial Regression (SGD Regression)

- Parameters: Fit_intercept, loss, penalty
- Scoring metric: Mean of CV with 3 cross-validation
- Model selection: Grid Search Cross-validation
- Best parameters: {'fit_intercept': True, 'loss': 'epsilon_insensitive', 'penalty': 'l1'}
- Performance: 0.2724
    - Data does not fit the model well

Gradient Boosting Regression

- Parameters: Learning Rate, Maximum Depth, Number of Estimators
- Scoring metric: 5 cross-validation and r2 scoring
- Model selection: Grid Search Cross-validation
- Best parameters: {'learning_rate': 0.3, 'max_depth': 5, 'n_estimators': 150}
- Performance: Training Data: 0.8619. Testing Data: 0.6504
    - Data fit the modele moderately well

Random Forest Regression

- Parameters: Min Samples Leaf, Min Samples Split, number of estimators
- Scoring metric: 5 cross-validation and r2 scoring
- Model selection: Grid Search Cross-validation
- Best parameters: {'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 150}
- Performance: Training Data: 0.9584. Testing Data: 0.6911
    - Data fit the model well

Summary of findings/Potential implications

Exploratory Data Analysis:

Gender differences: From the EDA, we found that male suicide rates are relatively higher than female suicide rates. With that, we researched the reason for that difference. According to BBC, this is due to the more violent methods of suicide by males. These methods made them more likely to die before anyone can intervene, which increase the number of deaths by suicide and increase the male suicide rate.

Age group: The EDA for age and generation shows a trend of older generations and older age ranges having higher suicide rates. This trend is in line with a study from the US National Library of Medicine where it is stated that "Suicide rates tend to rise as a function of age for both men and women to a peak in old, old age". According to the Suicide Prevention Resource Center, potential causes may be because alder adults plan more carefully and use more deadly methods, they're are less likely to be discovered and rescued, and the physical frailty of older adults means they are less likely to recover from an attempt.

GDP: Based on the observation found in EDA, GDP per capita is negatively correlated with suicide rate. According to overall scatterplot between suicide rate and GDP per capita, countries with higher GDP tend to experience lower suicide rate. While on the lower end of the GDP per capita, there seems to be a mixed signals: some countries with lower GDP per capita also have lower suicide rates. This indicates that GDP is only partially associated with suicide rate and there are many factors within a country that contribute to its development. Looking closely at the top 10 countries with the highest mean suicide rate, we can see that over the past few decades the rate of suicide generally decreases as the GDP per capita increases and the regression plot has also confirmed the negative correlation between the two variables.

Machine Learning Models:

With all our models, we found that our data fits the best using the ensemble learning algorithm (Gradient Boosting Regression and Random Forest). This may be due to the fact that the data have a non-linear shape which means the linear model could not capture the non-linear features, which result in lower accuracy. However, with Ensemble Regression using trees, the models were able to handle non-linear features. The high performance score in the ensemble regression indicates that our features (gender, generation group, age group, year of suicide, and GDP for the year) can be useful in prediction of suicide rate. However, the factors that might contribute to the suicide rate are not limited to these variables. This indicates that suicide is complex human public health issues involves factors beyond our dataset. In the future, we can include more factors from other datasets to further analysis the suicide rate.

With the Gradient Boosting Regression and Random Forest, we were able to generate machine learning models that could be meaningful for public health scientist and policy maker in finding

solution for suicide prevention. Using the models perhaps would help them determine what kind of issues should be aware of and provide some directions they could follow to examine the factors that lead to higher suicide rate. This could be used as an initial step in finding the underlying reasons that contribute to increase or decrease in suicide rate.

Suicide Rate Over Time:

Since 1995, the suicide rate has been decreasing. This may be reflected by many different factors. For example, economic growth has been steadily increasing over the past few decades. This provides more resources and employment opportunities for the population and people would be more likely to live a comfortable life. According to Tony Blakely's research on unemployment and suicide rate, unemployment is associated with two to three times of increase in relative risk of death by suicide compared with employment.

Secondly, as the field of psychology and medical science advances, people are becoming more open about mental health problems. According to a survey done by American Psychological Association, more and more American adults believe that having mental health problem is nothing to be ashame of and believe that people should talk about suicide more openly and recognize it as an important issue that should be brought attention to.

In addition, according to the study from the US National Library of Medicine, the sucide rate for older adults has dropped by 35% since 1986. This may partially be because of the advancements in medical technology allowing for later life to last longer for peers which reduces suicide rates due to the construct of social connectedness.

Work Cited

Survey: Americans becoming more open about Mental Health. (2019). *PsycEXTRA Dataset*. https://doi.org/10.1037/e504922019-001

Blakely, T. A., Collings, S. C., & Atkinson, J. (2003). Unemployment and suicide. Evidence for a causal association?. *Journal of epidemiology and community health*, 57(8), 594–600. https://doi.org/10.1136/jech.57.8.594

Schumacher, H. (2019). Why more men than women die by suicide. *BBC Future*.

https://www.bbc.com/future/article/20190313-why-more-men-kill-themselves-than-women

Conwell, Y., Van Orden, K., & Caine, E. D. (2011, June). Suicide in older adults. The Psychiatric clinics of North America. Retrieved December 7, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3107573/