Suicide Data by Country 1985-2016

Analysis and Prediction

Authors: Steven Kinnunen and Asel Kushkeyeva

Date: December 24, 2020.

Table of Content

Introduction

This analysis seeks to understand some of the complicated relationships that influence suicide rate. Our analysis is based on a dataset covering several countries from 1985-2016. Looking at Lee's CBC article on suicide rate and unemployment rate in Alberta (Lee, 2019), we noted that there appears to be a link between suicide rate and unemployment rate. To that end, the OECD data will help determine whether we can predict suicide rate and notice any meaningful trends based on our data.

There are four main issues that we wanted to address with our available data. First, which countries / age group(s) / gender(s) can be classified as a 'high', 'moderate', or 'low' suicide risk? Secondly, Is it possible to predict suicide rates by GDP per capita? Additionally, Is it possible to predict the suicide rates using unemployment data along with other socio-economic factors? Finally, are we able to predict a change in the suicide rates after the global recession of 2008 ?

Data Description

The suicides dataset can be found on Kaggle (https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016). As previously mentioned, our dataset covers several countries from 1985-2016. There are 101 countries in total with data on various socio-economic factors and suicide rates with a total of 27820 rows. The features included in the dataset are: country, year, sex, age, country-year, generation, population, suicide_no, suicides/100k pop (suicide rate), HDI for year, gdp_for_year($) and gdp_per_capita.
In order to better analyze the dataset, we created some new columns based on existing data: Risk ('low', 'moderate', and 'high'), continent, binary columns for each gender, each age group as well as each continent, and binary column for Risk ('low' and 'high').

We also used an unemployment dataset that can be retrieved from the OECD website (https://data.oecd.org/unemp/unemployment-rate.htm) and detailed instructions for retrieval can be found in appendix I. This dataset similarly contains data from 1985-2016. However, there are only 1052 rows. Columns include: Location (abbreviation for country or organization), Indicator, Subject, Measure, Frequency, Time (Year), Value (Unemployment Rate) and Flag Codes.

Exploratory Data Analysis **[Please refer to SUICIDES1 jupyter notebook]**

The Suicide dataset contains information spanning over 30 years from 1985 with number of observations fluctuating around 600 in the first five years, increasing steadily until about 1000 data points per year and dropping to 700 in 2015 (Figure 1, this and the following figures of this section were created with the help of seaborn package of Waskom et al.,

2020). The year of 2016 is the most underrepresented with less than 200 observations.

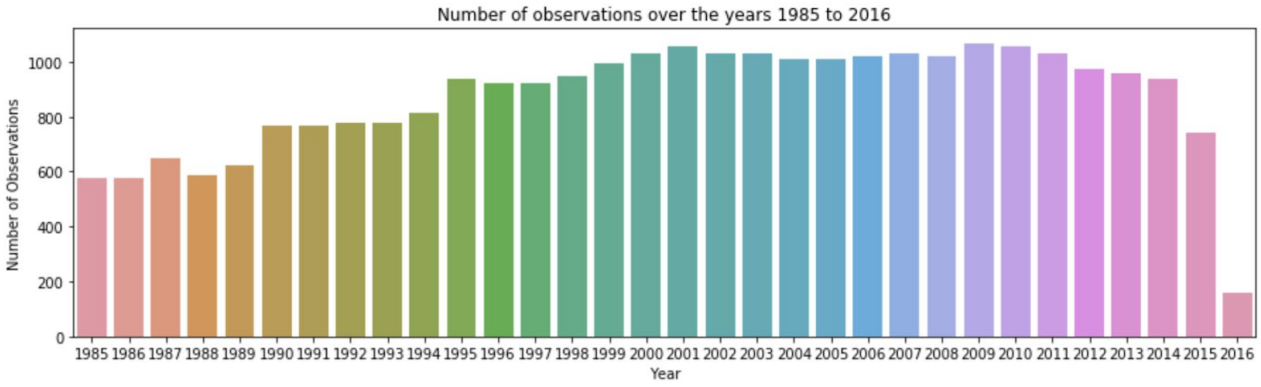


Figure 1. Number of observations over the years 1985 to 2016

Exploring the data further, we excluded the 'HDI for year' variable from our analysis since it had about 20,000 missing values (Image 1). Conveniently, the age groups were represented equally, about 4,600 observations in each group (Figure 2).

```
1 suicide.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27820 entries, 0 to 27819
Data columns (total 12 columns):
country                 27820 non-null object
year                    27820 non-null int64
sex                     27820 non-null object
age                     27820 non-null object
suicides_no             27820 non-null int64
population              27820 non-null int64
suicides/100k pop       27820 non-null float64
country-year            27820 non-null object
HDI for year             8364 non-null float64
 gdp_for_year ($)       27820 non-null object
gdp_per_capita ($)      27820 non-null int64
generation              27820 non-null object
dtypes: float64(2), int64(4), object(6)
memory usage: 2.5+ MB
```

Image 1. Missing values

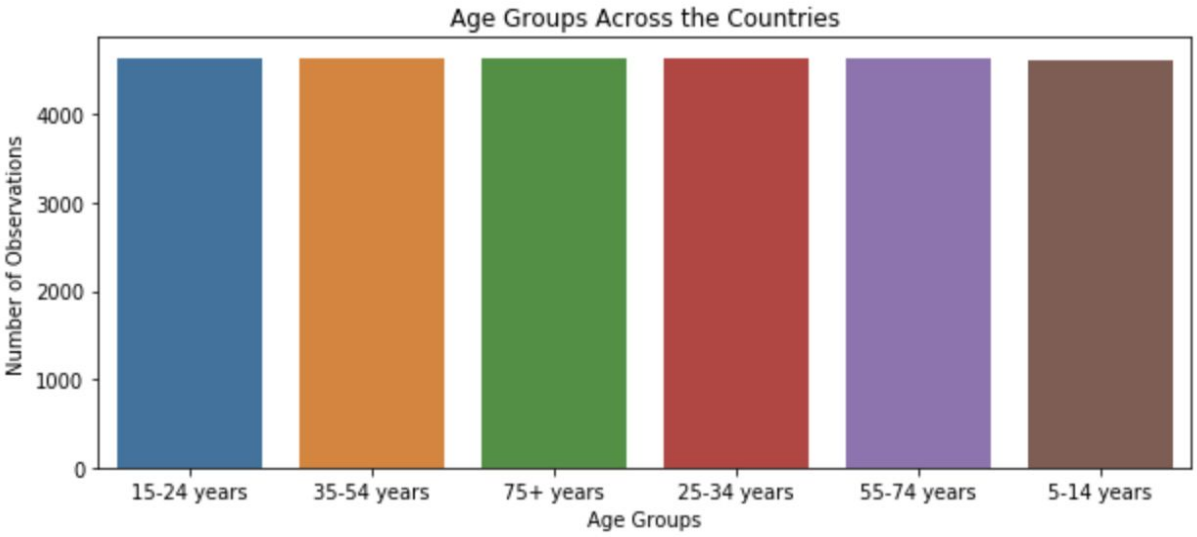


Figure 2. Age groups across the countries

One of the notable facts of the Suicide dataset is that by age group and generation, with the exception of the youngest age group of 5 to 14, males commit more suicides than

females with the highest numbers for over 75-years, followed by 35 to 54 and 54 to 75 years old groups.

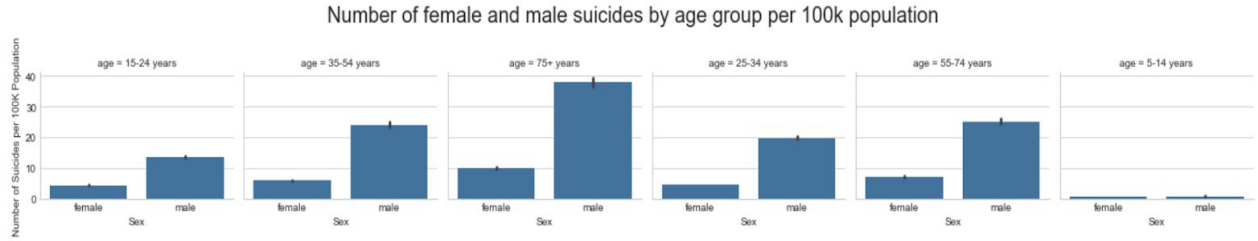


Figure 3. Number of female and male suicides by age group per 100,000 population

As mentioned earlier, we came across some information that suggested a strong correlation between suicide rate and the unemployment rate. As you can see from the boxplot (Figure 4), most values for unemployment rate fall to about 5-10%, with several outliers after about 17%. On the right of Figure 4, you can see the distribution of suicides per 100k by unemployment rate. At first glance, it would not appear that this relationship is a very linear one and there are quite a few outliers. Nonetheless, we created a linear model to see how well, or not, it could predict suicide rate by unemployment.

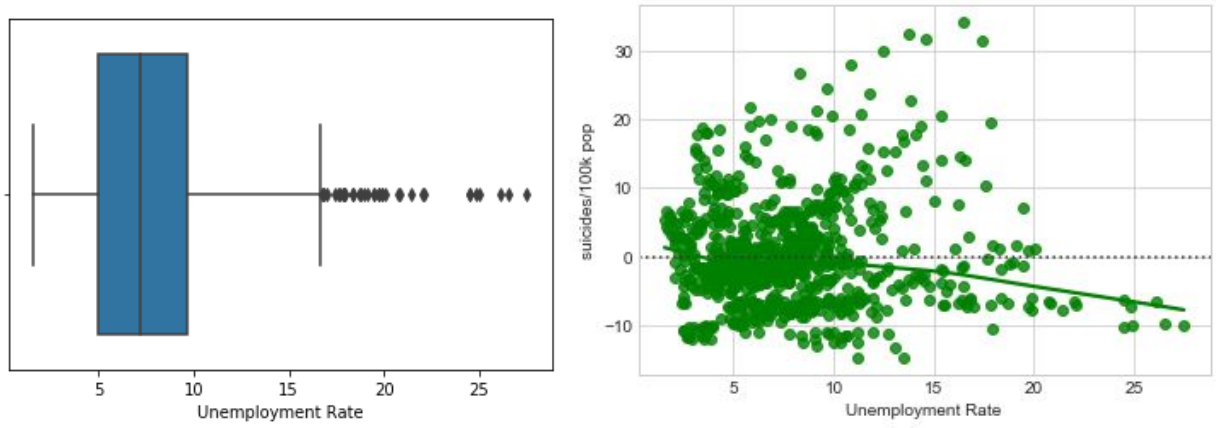


Figure 4. Missing values for unemployment data and its distribution by suicides per 100,000 population

As shown in Figure 5, we hoped to find a strong negative correlation between GDP per capita and suicides (the top left scatterplot in blue). However, as the boxplots in the same figure illustrate, both variables had outliers, and their distributions were right-skewed. After removing the outliers, the dataset reduced from 27,820 to 24,785 observations, and the

updated scatterplot in green shows no correlation between the variables (top right scatter plot).

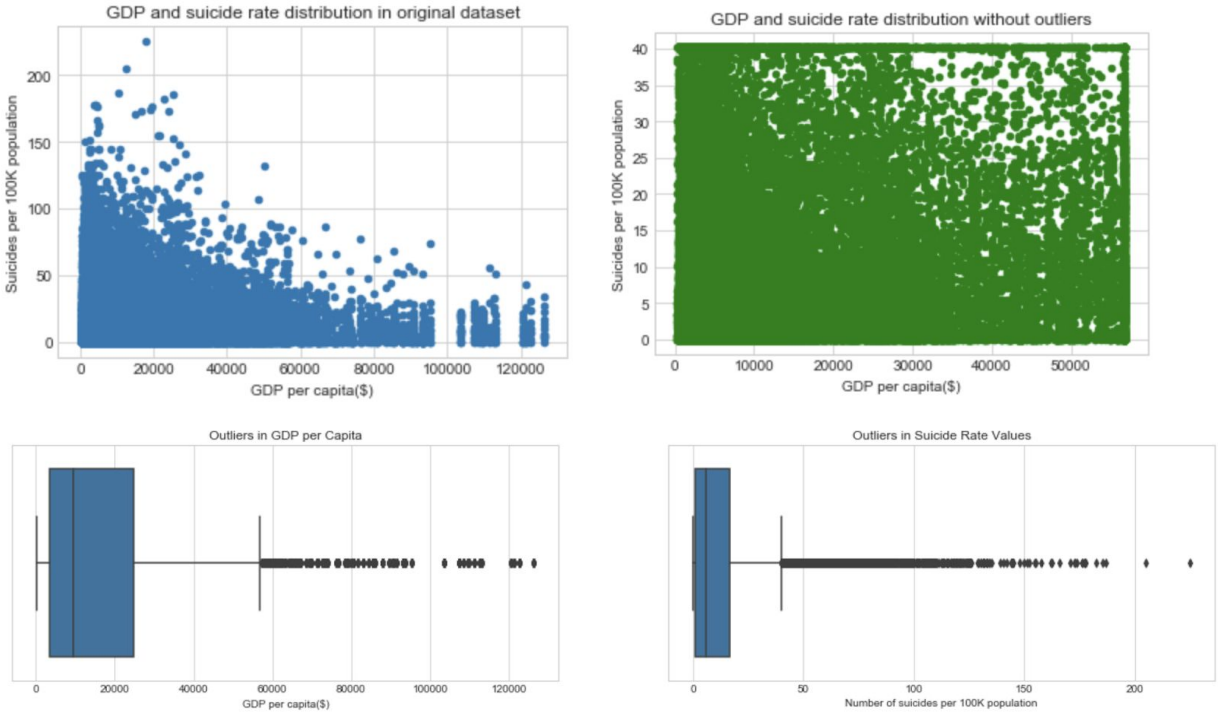


Figure 5. GDP per capita and suicides distributions with and without outliers

Research Questions

We proposed four questions to explore with the datasets:

1. What countries / age group / gender can be classified as ‘high’, “moderate, or ‘low’ suicide risk?

2. Is it possible to predict suicide rates by GDP per capita?*

*The question was changed from “Is it possible to predict GDP per capita looking at suicide rates in some countries, age groups, and/or sexes?”

3. Is it possible to predict the suicide rates using unemployment data along with other socio-economic factors?

4. Are we able to predict a change in the suicide rates after the Great Recession of 2008?

Since all the questions focused on suicides rate, our main motivation was to help policy makers determine which age groups, or gender to direct their attention and resources for mental health support; and for international organizations to prioritize the geographic locations. We were also trying to identify which major events had an impact on suicide rate and risk.

Research Question 1: What countries / age group / gender can be classified as ‘high’, or ‘low’ suicide risk? [Please refer to “Suicides 1985-2016 Report” jupyter notebook]

Decision Tree: 3 Levels of Risk

In order to predict risk and determine which features act as better predictors, we turned to decision trees and logistic regression. It is important to note that these decision tree models were used without removing outliers since decision trees are robust when dealing with outliers.

Our first decision tree (Figure 6) operates on three levels of risk: ‘low’, ‘moderate’ and ‘high’. These values are set for 10 or lower suicides per 100 000 people, between 10 and 20 suicides per 100 00 people and greater than 20 suicides per 100 000 people respectively. These values were based off of the scale used in the global suicide map on the Ourworldindata website (Ritchie et al, 2015). It is important to note that there is a class imbalance issue at play: there are 17 358 ‘low’ values whereas there are only 5747 ‘high’ values and 4715 ‘high’ values. In order to circumvent this issue, we undersampled from the majority class so that there were only 4715 ‘low’ values with the number of ‘high’ and ‘moderate’ values unchanged (Brownlee, 2020). The decision tree was created with a maximum depth of 5.

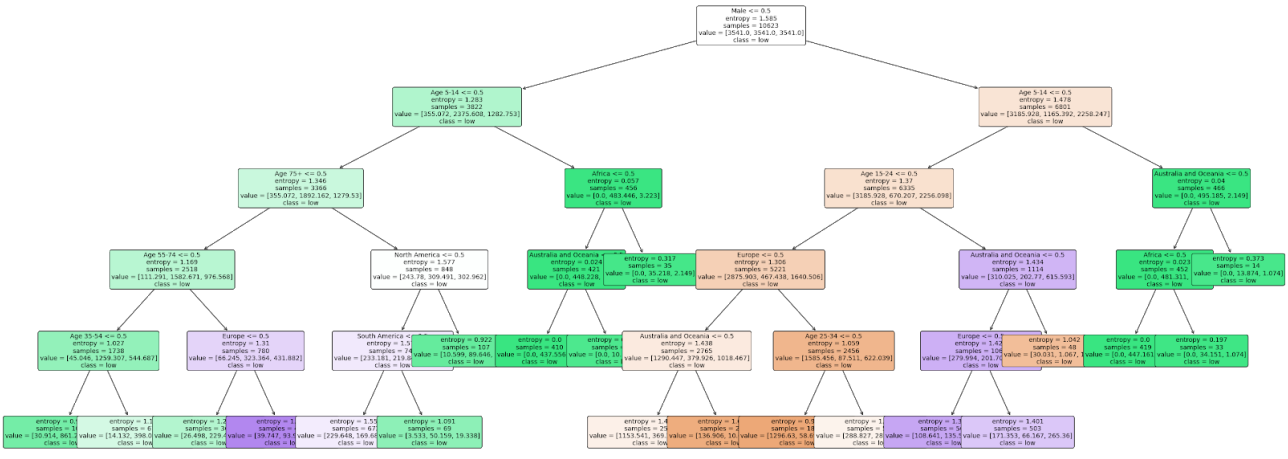


Figure 6. Decision tree on three levels of suicide risk

	precision	recall	f1-score	support
high	0.62	0.82	0.71	1738
low	0.74	0.72	0.73	1397
moderate	0.50	0.31	0.38	1419
accuracy			0.63	4554
macro avg	0.62	0.62	0.61	4554
weighted avg	0.62	0.63	0.61	4554

Image 2. Decision tree classification report

Above in Image 2, you can see the classification report for this decision tree model. Unfortunately, the overall accuracy is not very good at 63%. However, given the nature of this problem it is important to note that it is best to minimize the number of false negatives for ‘high’ values while also seeking to capture as many true positives as possible. This is the case because ‘high’ values correspond to higher suicide rates and therefore more resources that would likely need to be used by policy makers to help those more at risk. Therefore, the recall rate would be most important for ‘high’ values and it is fairly good at 82% (Shung, 2020). This is coupled with a precision of 62% and f1-score of 71%.

However, the precision, recall and f1-score for ‘moderate’ risk are not very good at 50%, 31% and 38% respectively. While it is important to capture as many true high risk cases as possible, moderate risk cases may also need increased resources over low risk cases. Therefore, these shortcomings must be considered before using this model.

Since we are concerned predominately with correctly identifying ‘high’ risk cases, this model may still be useful given the high recall for those ‘high’ values. However, the bad precision, recall and f1-score for moderate values should prompt policy makers to strongly consider whether to use this model.

Decision Tree: 2 Levels of Risk

Our second decision tree model (Figure 7) only operates on two levels of risk: ‘low’ and ‘high’. ‘Low’ values correspond to suicide rates less than 6, whereas ‘high’ values are any values that are 6 or more per 100 000 people. This means that ‘high’ risk values, the value most likely to be of interest to policymakers, will be less specific than in the previous model where 3 levels of risk were used. We used the median value (5.99) of suicide rate to determine the threshold for ‘low’ and ‘high’ risk values. We used a maximum tree depth of 5 as well as setting the minimum samples for a leaf node to be 500 to attempt to prune the tree. Before, higher values for the minimum number of samples for a leaf node were

attempted (750 and 1000 in particular), however this dropped the overall accuracy of the model by a few percentage points.

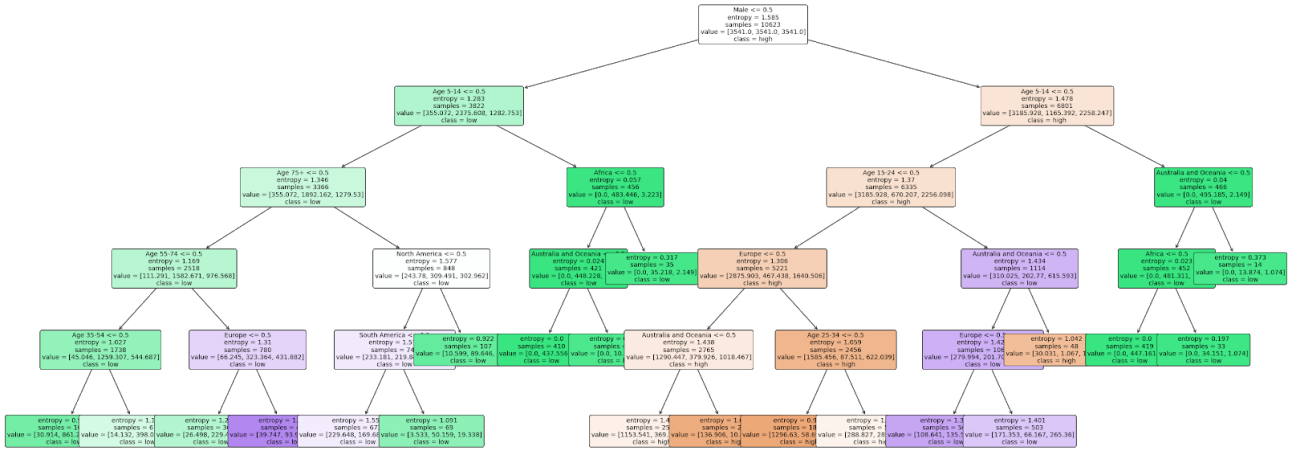


Figure 7. Decision tree with two levels of suicide risk.

	precision	recall	f1-score	support
high	0.79	0.83	0.81	4150
low	0.82	0.78	0.80	4196
accuracy			0.81	8346
macro avg	0.81	0.81	0.81	8346
weighted avg	0.81	0.81	0.81	8346

Image 3. Decision tree classification report

Just above (Image 3) you can see the classification report for the second decision tree model. There is an overall accuracy of 81%. Consider once again that 'high' values are more likely to be of interest for policy makers than low values. This model, while being less specific, produces better precision, recall and f1-scores for 'high' values at 79%, 83% and 81%. The recall is only 1% better than the previous model (using 3 risk levels), so the specificity of the 'high' risk values should be weighted against the overall performance of each model.

It is important to note that the two most important features appear to be Male (1 represents 'yes', 0 'no') and Age 5-14 (1 represents 'yes', 0 represents 'no'). Looking at the decision tree, one can see that Males are more likely to be classified as high risk. Conversely, if someone is male but also within the 5-14 age group then a 'low' class label is more likely. This is consistent with our exploratory data analysis where it appeared that males committed more suicides overall and at higher rates than females. Additionally, everyone within the 5-14 age group appeared to commit less suicides overall and at much lower rates.

Logistic Regression: 2 Levels of Risk

	precision	recall	f1-score	support
0	0.76	0.79	0.78	4139
1	0.79	0.75	0.77	4207
accuracy			0.77	8346
macro avg	0.77	0.77	0.77	8346
weighted avg	0.77	0.77	0.77	8346

Image 4. RQ1: Logistic regression classification report

The classification report above (Image 4) demonstrates the results of the logistic regression model. We calculated importance for the features and determined that male, female, age 5-14, age 55-74, age 75+, North America and South America were all important enough to retain. Once again, we used two levels of risk at the threshold of 6: 'low' values were below 6 per 100 000 people whereas anything 6 or higher was counted as 'high'. It is important to note that 0 represents 'low' whereas 1 represents 'high'.

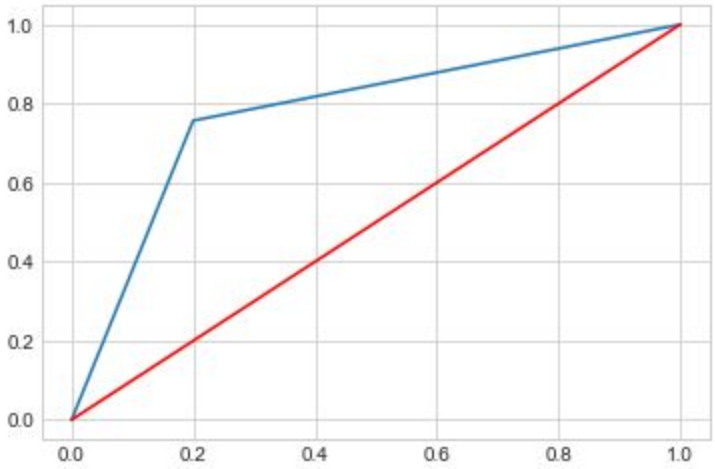


Figure 8. RQ1: ROC curve

The overall accuracy of the model is 77%. Keeping in mind the relative importance of 'high' values, we can see that there are precision, recall and f1-scores of 79%, 75% and 77%. While these values are not bad, all these values are outperformed by the decision tree model for 2 levels of risk. The recall in particular is 8% lower in our logistic regression model. Therefore, if policy makers are to evaluate risk at 2 levels based on the thresholds we have established, we would strongly encourage using the decision tree model over the logistic regression model since that 8% represents a large difference in the number of

false negatives for ‘high’ values.

Research Question 2: Is it possible to predict suicides rate by GDP per capita? [Please refer to SUICIDES1 jupyter notebook]

As discussed in the EDA section above, the original Suicide data contained a little over 3,000 outliers in GDP per capita and suicides rates. Considering that the higher values of GDP represent wealthier economies, whereas the higher numbers of suicides might be a result of major events, in other words, there is a lower chance of them being a typo or other unintended error, we settled on two methods of handling the outliers: (1) removing them, and (2) replacing the outliers with the maximum values, which was done in an attempt to preserve those high values. We created two datasets reflecting the method of handling outliers above. However, the logistic regression and K-Nearest Neighbor (KNN) models performed similarly with both datasets, so further in the analysis, we chose to go ahead with the dataset where the outliers were replaced by maximum values.

Since balanced datasets allow for more accurate predictions, we used the 50th percentile of the target variables to create balanced categories of 0 being low and 1 being high (Figure 9).

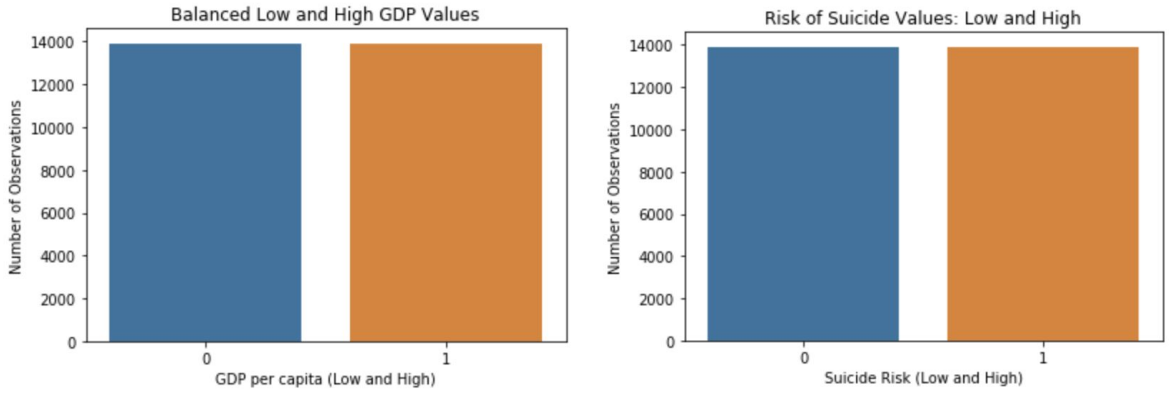


Figure 9. Balanced categories of GDP per capita and risk of suicide

Logistic Regression

First, we created binary suicide risk variable splitting the data by the 50th percentile - 0 (‘low’) being the values below 6 and 1 (‘high’) for values above 6. Then, running the model with only one predictor, GDP per capita, we have achieved a 0.54 accuracy score by setting max_iteration to default of 100 and solver to liblinear (Figure 10). The Receiver Operator Characteristic (ROC) curve gives us the trade-off between true positive and false positive rates. Here, as expected, we see that area under the curve is not large.

```
[[2298 1909]
 [1925 2214]]
              precision    recall  f1-score   support

           0       0.54      0.55      0.55      4207
           1       0.54      0.53      0.54      4139

    accuracy                           0.54      8346
   macro avg       0.54      0.54      0.54      8346
weighted avg       0.54      0.54      0.54      8346
```

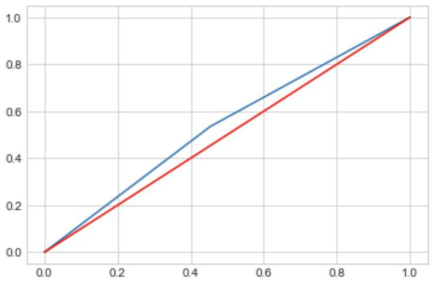


Figure 10. RQ2: Confusion matrix, classification report, and the ROC curve

Since only one predictor was used for this model, the feature selection method was not applicable.

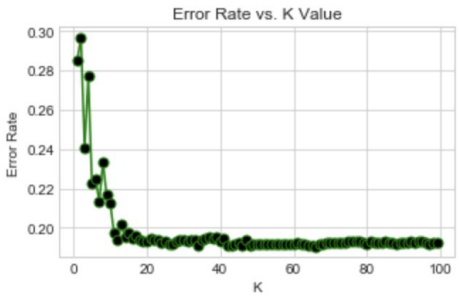
K-Nearest Neighbor

Since KNN works best with fewer features, we hoped to see better results than of the previous model. Indeed, using the same parameters as in the logistic regression classifier, KNN achieved an accuracy score of 0.64 with k=1 (Image 5).

```
WITH K=1

              precision    recall  f1-score   support

           0       0.77      0.74      0.75      4134
           1       0.75      0.79      0.77      4212

    accuracy                           0.76      8346
   macro avg       0.76      0.76      0.76      8346
weighted avg       0.76      0.76      0.76      8346
```

Image 5. RQ2: K-Nearest Neighbor classification report with K = 1

Then we attempted to find the best K, limiting the value of K to 100 as the error rate fluctuated slightly around 0.26 and remained constant beyond the threshold. The figure 11 shows that the lowest dip in the graph is around 15, which is the best K for this model. Running the KNN classifier with K=15, improved the accuracy score by 0.06 points.



WITH K=50

	precision	recall	f1-score	support
0	0.83	0.78	0.80	4150
1	0.80	0.84	0.82	4196
accuracy			0.81	8346
macro avg	0.81	0.81	0.81	8346
weighted avg	0.81	0.81	0.81	8346

Figure 11. RQ 2: Finding best K value and classification report with K = 50

Research Question 3: Is it possible to predict the suicide rates using unemployment data along with other socio-economic factors? [Please refer to "Suicides 1985-2016 Report" jupyter notebook]

We used linear regression on the merged suicides and unemployment dataset in order to determine whether there was a direct link to suicide rate and unemployment rate in our data. Using only unemployment as a predictor we got a root mean squared error of 7.64. This corresponds to 7.64 deaths per 100 000 people. Returning to the question of risk for a moment, this could drastically move someone from one risk category such as 'low' to another risk category such as 'moderate' or 'high' depending on which thresholds are being used.

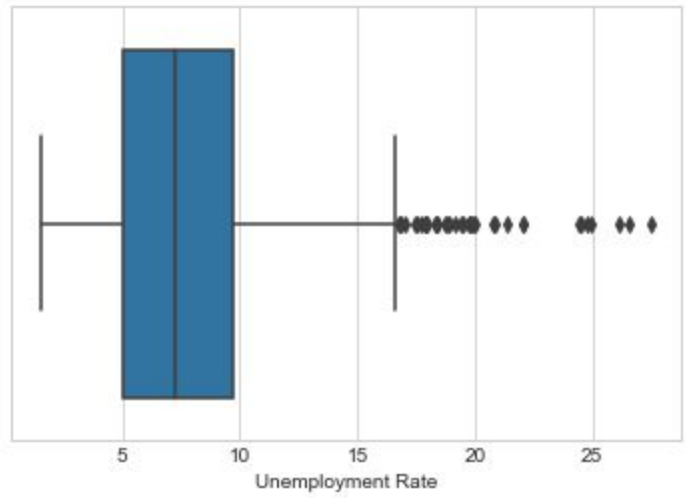


Figure 12. Unemployment rate boxplot

Looking at the boxplot (Figure 12) for unemployment rate we can see a number of outliers in the data starting at about 17% onwards. Moreover, we can look at the two plots for unemployment and suicide rate below (Figure 13) and can see clearly that the

assumptions of linearity and homoscedasticity are violated. Recall that homoscedasticity is constant variance of residuals at every level of x (Zach, 2020). The variance appears to change drastically for different values of x.

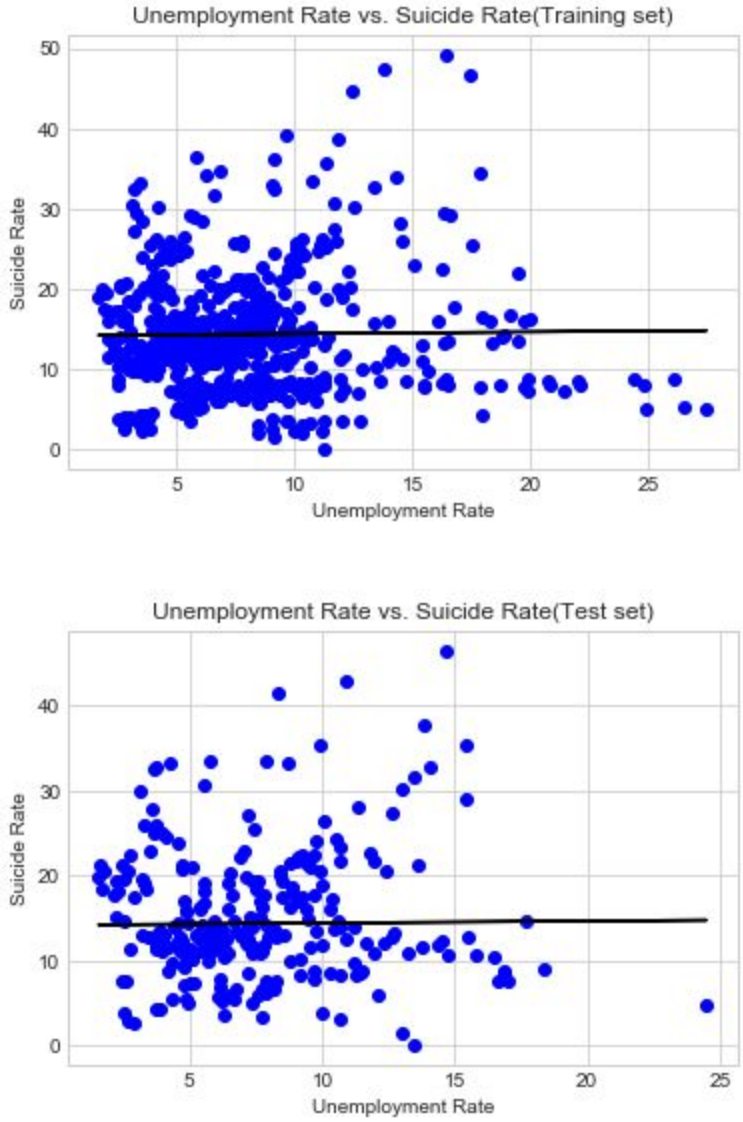


Figure 13. Unemployment rate versus suicide rate in training and test sets

This linear regression model was particular to the merged suicides and unemployment dataset. However, there are other socio-economic factors that may be useful predictors found in the suicides dataset. Using a regression tree and the suicides dataset, we constructed a model to predict on gender, age group and continent. Using a 70:30 train-test split and a maximum tree depth of 5, we get a model that produces a root mean square error of 14.361. This appears to be even worse than the simple linear regression model used on unemployment. However, the two datasets are drastically different as the merged dataset contains less than 1000 rows, whereas the suicides dataset contains 27820 rows. Furthermore, there are more countries in the suicides dataset (over 100) as

compared to the merged dataset (a little over 30). These may account for some of the observed differences.

Research Question 4: Are we able to predict a change in the suicide rates after the Great Recession of 2008?

Classification Models **[Please refer to SUICIDES1 jupyter notebook]**

The financial crisis of 2008 had a major impact on people's well being both economically and mentally. The younger generations entered the broken job market, whereas the older individuals lost a significant portion of their retirement savings. Since our goal is to predict psychological consequences of the Great Recession and they may take a few years to manifest, to separate two periods - pre- and post-recession - we included the year of 2008 into the pre-recession years. To be consistent with the rest of the analysis, we introduced the year_binary variable to explore the fourth research question. Here, 0 represents years before and including 2008 and 1 for years after 2008 (Figure 14). Althouth, the year_binary variable is not as balanced as the binary values of GDP per capita and suicide risk rate, the following logistic regression and KNN classifiers performed well.

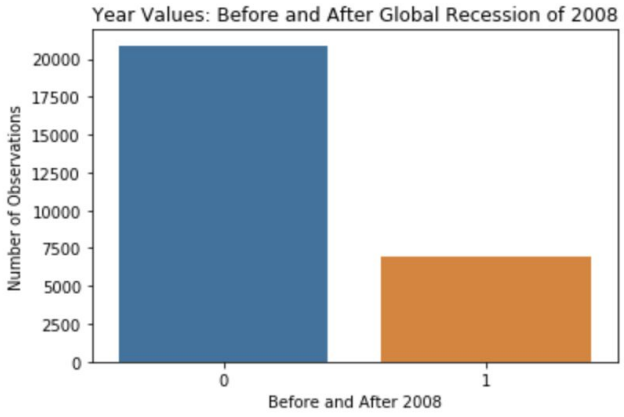


Figure 14. Binary year variable

Logistic regression

To determine what predictors to use for the analysis, we performed a correlation analysis that revealed a few highly-correlated pairs of features such as Male and Female, suicides/100k pop and Risk_binary, and suicides_no and population (Figure 15). After removing the correlated predictors and adding binary GDP and year, we arrived at 15 features: "Male","Age 5-14", "Age

15-24", "Age 25-34", "Age 35-54", "Age 55-74", "Age 75+", "North America", "South America", "Europe", "Asia", "Australia and Oceania", "Africa", "gdp_binary", "year_binary".

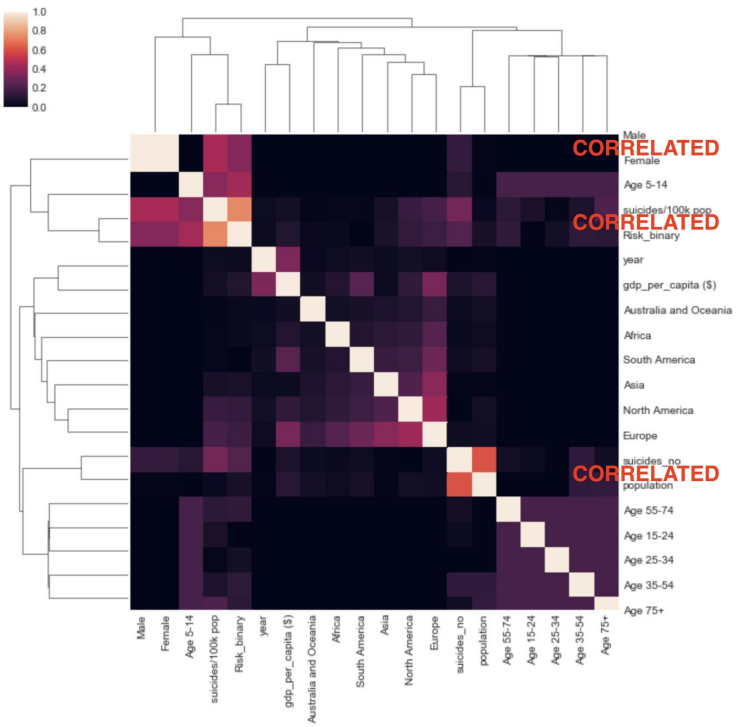


Figure 15. Correlation cluster map

The logistic regression's confusion matrix (Figure 16) shows that, for years before the financial crisis, *3,194* of *4,162* data points were predicted correctly, and, for the post-recession years, the numbers are slightly higher (*3,506* of *4,184* observations). The area under the ROC curve (on the right) is large as the accuracy score is 0.80.

```
[[3194  968]
 [ 678 3506]]
              precision    recall  f1-score   support

           0       0.82      0.77      0.80      4162
           1       0.78      0.84      0.81      4184

    accuracy                           0.80      8346
   macro avg       0.80      0.80      0.80      8346
weighted avg       0.80      0.80      0.80      8346
```

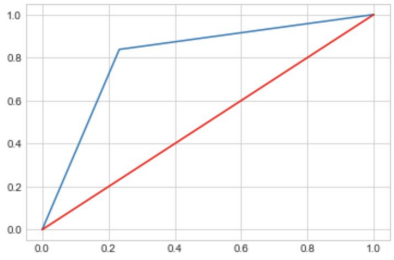


Figure 16. RQ4: Logistic regression classification report and ROC curve

Next, we attempted to improve the model by selecting the best features, for which we utilized SelectFromModel function from scikit-learn (Pedregosa et al., 2011). The function "chose" six predictors that were used to re-run the classifier, which, unfortunately, deteriorated by a few points (Image 6).

```
[[3526  640]
 [1361 2819]]
              precision    recall  f1-score   support

           0       0.72      0.85      0.78      4166
           1       0.81      0.67      0.74      4180

    accuracy                           0.76      8346
   macro avg       0.77      0.76      0.76      8346
weighted avg       0.77      0.76      0.76      8346
```

Image 6. RQ4: Logistic regression confusion matrix and classification report after feature selection analysis

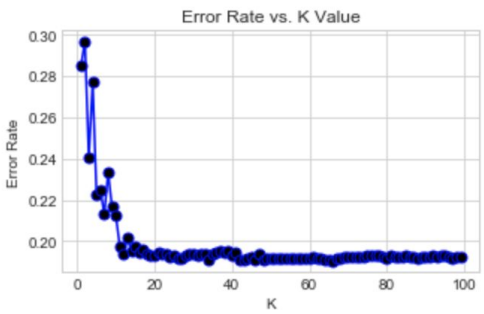
KNN

By employing the same 15 predictors as for logistic regression, as illustrated in the Image 7 below, KNN classifier hit 0.73 accuracy score (k=1).

```
WITH K=1


              precision    recall  f1-score   support

           0       0.74      0.71      0.73      4163
           1       0.73      0.75      0.74      4183

    accuracy                           0.73      8346
   macro avg       0.73      0.73      0.73      8346
weighted avg       0.73      0.73      0.73      8346
```

Image 7. RQ4: KNN Classification report with K = 1

After plotting the error rate and K values (Figure 17), we chose K to be 20, which increased the accuracy score to 0.80. With F1-score being 0.80 and 0.81 for years before and after the recession, respectively, the KNN's results are identical to the logistic regression model.




```
WITH K=20


              precision    recall  f1-score   support

           0       0.81      0.78      0.80      4163
           1       0.79      0.82      0.81      4183

    accuracy                           0.80      8346
   macro avg       0.80      0.80      0.80      8346
weighted avg       0.80      0.80      0.80      8346
```

Figure 17. RQ4: Finding best K value and classification report with K = 20

Regression Results

[Please refer to "Suicides 1985-2016 Report" jupyter notebook]

Finally, we turn to regression trees in order to predict suicide rate. As previously noted, we created a regression tree that produced a RMSE of 14.361. However, this did not include the 'Recession' feature, which determines whether an observation took place before (including the year 2008) the year 2008 or after.

The second model uses all features of the previous model (male, female, all age groups and continents) as well as this recession variable. We created train-test sets at a 70:30 ratio and only managed to change the RMSE to 14.349. This produces a difference of 0.013 between the two RMSE values. Therefore, it seems that there is very little difference that 'Recession' has on the outcome of the regression tree algorithm. Furthermore, given the large value for RMSE, these models are not very effective at predicting suicide rates.

Results

RQ1: What countries / age group / gender can be classified as a 'high', 'moderate', or 'low' suicide risk?		RQ2: Is it possible to predict suicide rates by GDP per capita?	
Model	Accuracy Score	Model	Accuracy Score
Decision Tree	0.81	Logistic Regression	0.54
Logistic Regression	0.77	KNN	0.81

Table 1. Research questions 1 and 2 results

RQ3: Is it possible to predict the suicide rates using unemployment data along with other socio-economic factors?		RQ4: Are we able to predict a change in the suicide rates after the global recession of 2008?	
Model	Accuracy Score	Model	Accuracy Score
Linear Regression	RMSE: 7.638	Logistic Regression	0.80
Regression Tree	RMSE: 14.361	KNN	0.80

Table 2. Research questions 3 and 4 results

In order to predict risk level and determine important features that determine risk (RQ1), we created 2 decision tree models and a logistic regression model. The first decision tree model did not perform well overall with an accuracy of 63%. However, the recall for high values was good at 82% and this model retained a specificity that the second decision tree lacked for high risk values as it retained 3 levels of risk. Conversely, the second decision tree had a much better accuracy at 81%. However, the recall for high values was much lower (which is an important consideration for policy makers) and the model lacked the specificity of the first decision tree for high risk values. Logistic regression did not perform as well with an accuracy of 77% and lower metrics in precision, recall and f1-score compared to the second decision tree. Overall, the second decision tree determined that variables male and age 5-14 were particularly important in determining binary risk level. Feature selection for logistic regression determined that features male, female, age 5-14, age 55-74, age 75+, North America and South America were important features for classifying a binary risk level.

We ran two models to predict suicide rates by GDP per capita (RQ2): Logistic regression and K-Nearest Neighbor. With only one predictor, the first model did not perform well (accuracy score of 0.54), whereas KNN predicted 81% of the observations in the data (Table 1).

In order to predict suicide rate, we used the OECD dataset to use unemployment rate (RQ3). However, the linear regression model suffered from violating the assumptions of linearity and homoscedasticity. It also produced a fairly large RMSE of 7.64. The regression tree was applied to a much larger dataset for other socio-economic factors and produced an even larger RMSE of 14.36. Neither model predicts suicide rate very well.

We tackled the fourth question both with classification and regression models. The logistic regression and KNN classifiers' accuracy scores were equally high, 80% of the data was predicted correctly (Table 2). However, the regression tree models, in addition to not

predicting suicide rate very well, also did not have much of a difference between RMSE when accounting for the difference between including the ‘Recession’ feature or excluding it.

Discussion

Our research focused on classifying and predicting suicide risk globally. Although we operated with a limited number of predictors, our main goal was to formulate the research questions to cover maximum information available. In an attempt to review our questions from various angles, we used numerous models such as decision tree, logistic regression, KNN, and linear regression. Overall, across all four questions KNN and decision tree (using binary risk values) models produced the best results.

However, the way we handled outliers might have influenced the outcomes. As mentioned earlier, in the Suicide dataset the GDP outliers suggest wealthier economies and the suicide rate outliers could be the result of major events in the countries or any other reasons not captured by this dataset. Secondly, grouping the countries into continents might have prevented from performing more in-depth analysis and prediction. For future research, we might attempt to categorize the geographic areas differently such as the developing versus developed world.

Lastly, since the unemployment rate dataset covered only one third of the countries presented in the Suicide data, we believe that collecting more data on unemployment could have enriched our research. According to Suicide article in Our World in Data “With timely, evidence-based interventions, suicides can be prevented” (Ritchie, Roser, & Ortiz-Ospina, 2015). We hope that our analysis could contribute to the ways of preventing them globally.

Self Assessment

Steven Kinnunen

On the suicide dataset I performed some data preprocessing. I also performed some data preprocessing on the Unemployment dataset and created the merged dataset. I mainly focused on research questions 1 and 3, although I contributed the regression trees to research question 4. I contributed to the results section of the report as well.

Asel Kushkeyeva

On the Suicide dataset, I performed Exploratory Data Analysis, identified main variables, transforming and cleaning the data as was required for various models. I mainly focused on research question 2 and 4, trained the classifiers and contributed to the Results and Discussion sections of the report.

References

Brownlee, J. (2020, January 19). Undersampling Algorithms for Imbalanced Classification. https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification/

Harris, C., Millman, K. J., van der Walt, S., Gommers, R., Virtanen, P., Cournapeau, D., … Travis, E. (2020). Array programming with {NumPy}. *Nature*. 585. doi: 10.1038/s41586-020-2649-2

Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9, 90-95. doi:10.1109/MCSE.2007.55

Lee, J. (2019, September 27). Advocates raise alarm about suicide and the economy. https://www.cbc.ca/news/canada/calgary/suicide-unemployment-increase-university-calgary-1.5300009

Lematre, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research, 18*(17), 1-5. doi:http://jmlr.org/papers/v18/16-365.html

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference,* 51-56

OECD. (n.d.). Unemployment rate. https://data.oecd.org/unemp/unemployment-rate.htm

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research.* 12, 2825-2830

Ritchie, H., Roser, M., & Ortiz-Ospina, E. (2015). Suicide. https://ourworldindata.org/suicide

Rusty. (2018). Suicide Rates Overview 1985 to 2016. https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016/notebooks

Shung, K. (2020, April 10). Accuracy, Precision, Recall or F1? https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

Waskom, M. and the seaborn development team. (2020, September). mwaskom/seaborn. doi: 10.5281/zenodo.592845

Zach. (2020, January 08). The Four Assumptions of Linear Regression. https://www.statology.org/linear-regression-assumptions/

Appendix

Appendix I : Retrieving OECD Unemployment Dataset

In order to retrieve the unemployment dataset, please visit (https://data.oecd.org/unemp/unemployment-rate.htm). Next, below the graph be sure to click on 'yearly' underneath 'Time'. Also underneath 'Time', select the years 1985-2016 on the slider. Ensure that the drop down menu under 'Perspectives' shows as 'Total'. Ensure that there are no specific countries selected in the dropdown menu under 'Countries', the dropdown menu should say 'Highlighted Countries (0)'. Finally, download the dataset using the download dropdown menu above the graph, select 'Selected data only (.csv)' and you will now have a file called something similar to 'DP_LIVE_23122020174926163.csv'. For simplicity, we renamed the file to 'OECD_unemployment1985-2016.csv' for our Jupyter notebook.