**Redesign Project**

**Suicide Prediction and visualization**



**STAT 515**

**Visualization for Analytics**

**Professor: Scott Bruce**

**Submitted By**

**Srashti Agrawal**

**Introduction:**

Suicide is the act of taking one’s own life intentionally. It is considered as a major mental health problem in the world. Suicide has many effects that go beyond the person who acts to take their lives, it could have an effect on their families, friends and the communities.

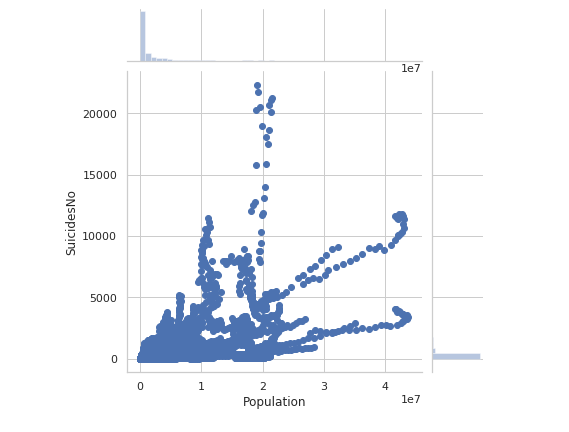
In our project, we present the suicide rates in the world with respect to seven categories, including sex (males and females), year, age group, count of suicides, population, suicide rates and generation. We manipulate different tools (graph and model) to show if there is any relationship among different categories. The data we used in this project are taken from United Nations Development Program (2018), World Bank (2018), Suicide in the Twenty-First Century (2017) and World health Organization (2018) Suicide prevention.

The purpose of this project is to improve the current visualizations by incorporating more variables in an interactive plot, which provides a rather comprehensive description of the dataset, and to make a prediction based on some transformed variables by employing a statistical model. Thus, it may be beneficial to study the suicide issue and therefore to decrease suicide rates in the world.

**Visualizations:**

**Comparison 1:**

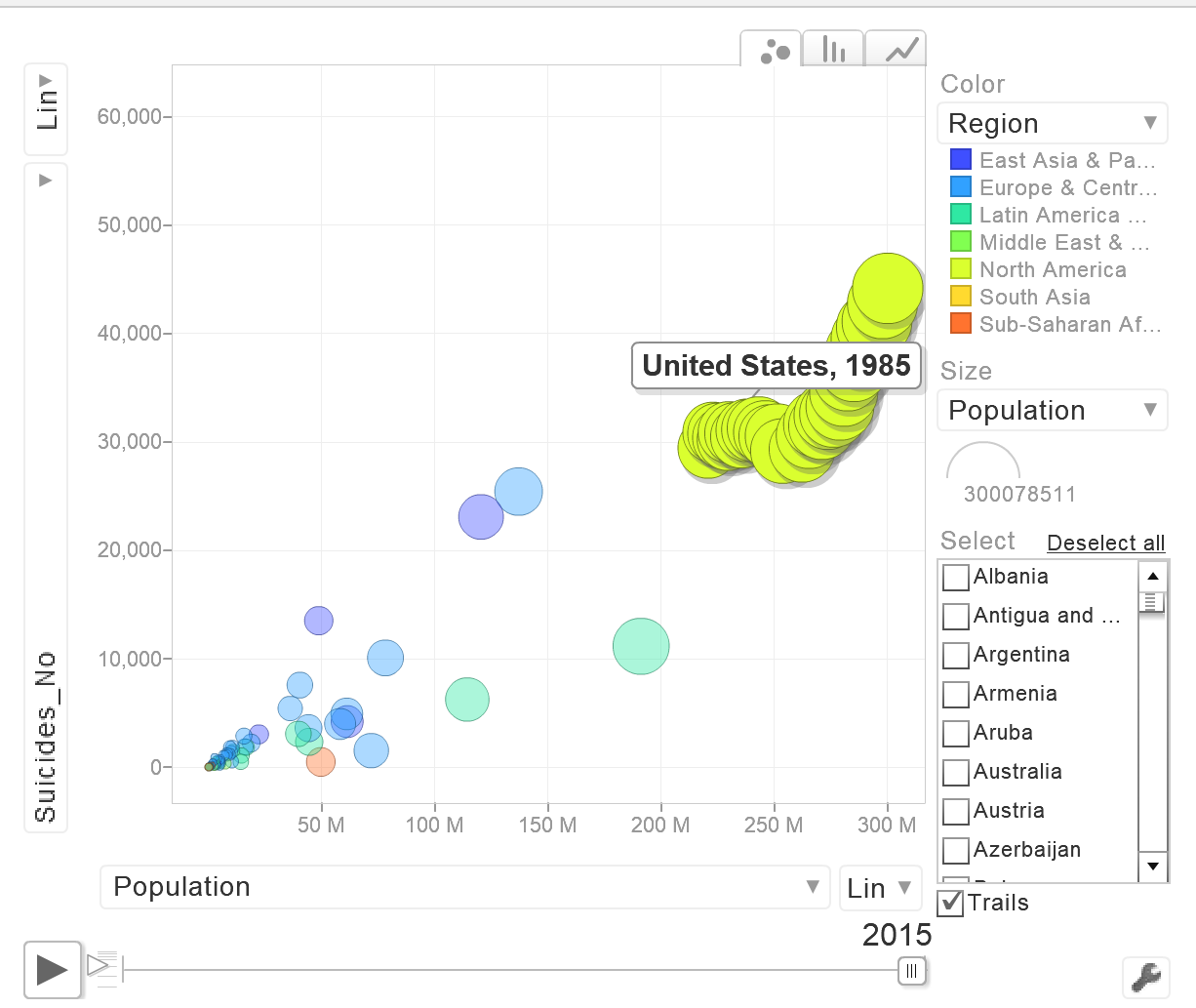
In this part we show the relation between suicide number and population in figure 1. As it is seen, there is only one color used to represent the relation between population and suicides number. The possible region variable is not presented in current plot, but it is very natural to connect population sizes with regions. Also, using one color made the plot too busy.



**(Figure 1. Suicide versus Population)**

In our visualization, we used different colors to represent different regions, and the size of each bubble represents the population of each country and how it changes by time. In addition, we used an interactive chart that allows us to change the axes in order to reveal different relations. We can also use the Trail feature, which allows us to track a specific country and see its changes across different years.

As displayed in figure 2, the visualization now is clearer and has meaning by having different colors and sizes of the bubbles. And we can track the change of the suicide numbers among the population from year 1985 to 2015. For example, if we want to track the change in United State, we can simply click on the bubble, check the trail option and hit play.

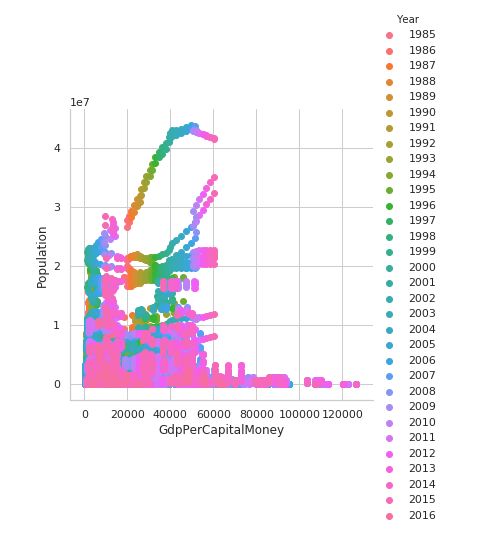


**(Figure 2. Suicide number versus Population)**

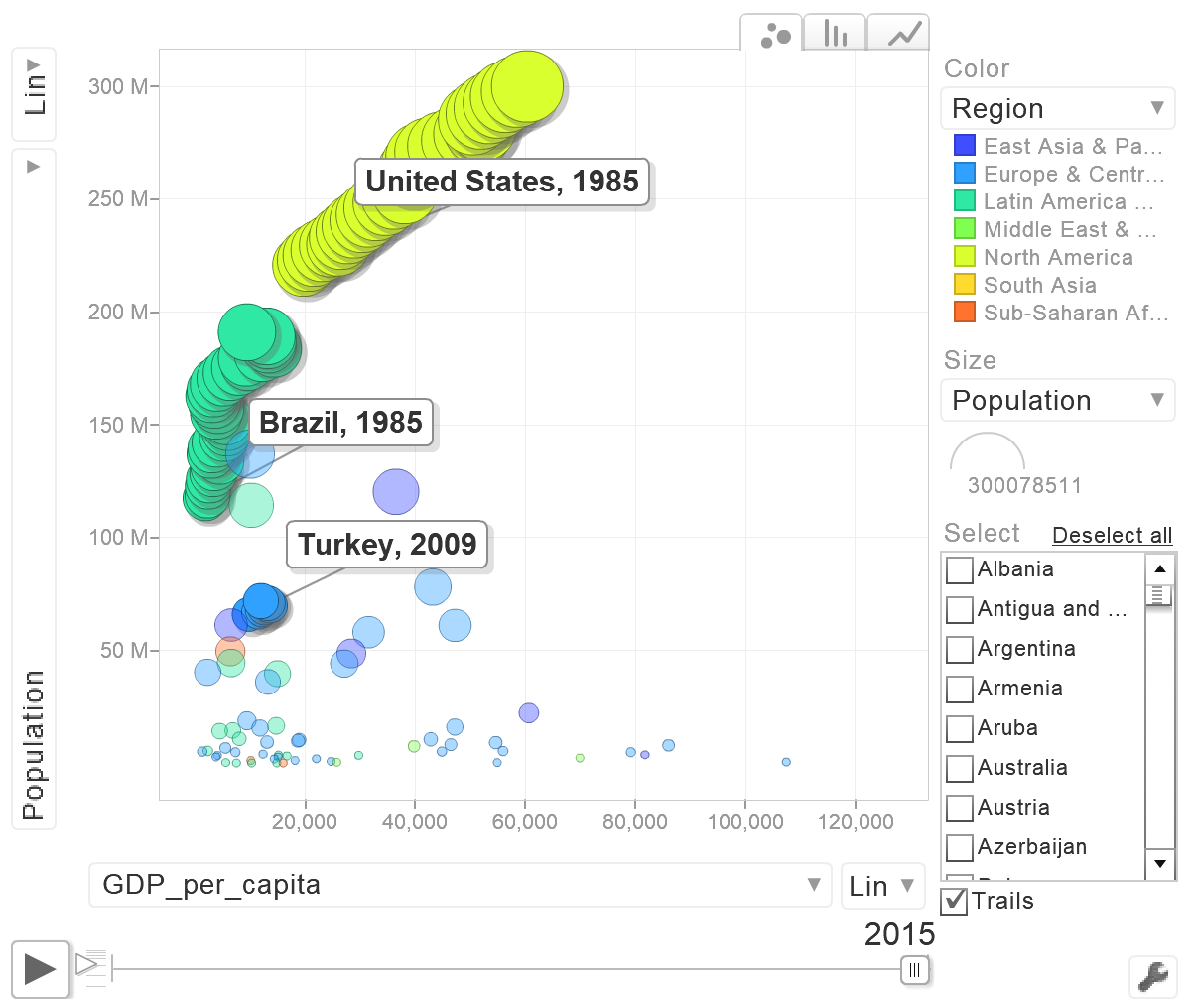
**Comparison 2:**

As it is shown in figure 3, using different colors does not guarantee a better visualization. Having different shades of the same color (representing the years) makes the plot looks busy and unreadable. Some of the bubbles are hidden, and some of them look like merging with others. At some point, we cannot even tell which year each bubble belongs to.

In our visualization (figure 4), we used different colors to represent countries rather than years, because the number of years, 30, makes color coding impossible. If we tried so, the plot would be messy and the year information is difficult to be extracted, as in Figure 3.

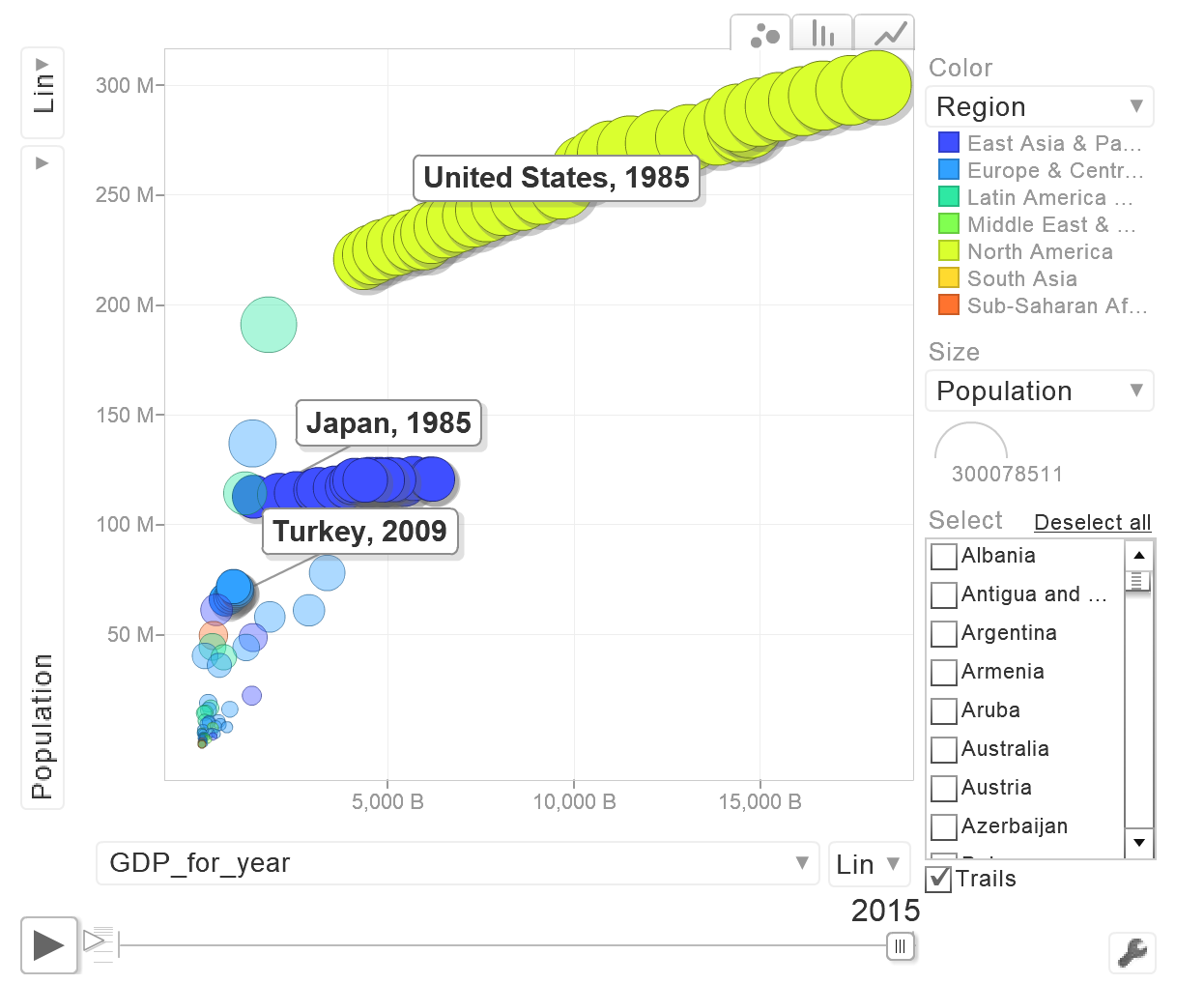


**(Figure 3. Population Versus GDP per capita)**



**(Figure 4. Population versus GDP per capita)**

We can also compare GDP per year with respect to population which represents the expense they spend in a particular year. We can change the axis depending on our requirements which can also facilitate further analysis (figure 5).



**(Figure 5. Population versus GDP per year)**

**Modeling:**

The random forest model is applied to predict the suicide proportion of 100 thousand population based on the five variables transformed from the original dataset. They are continent where a country locates, year, gender, age group and GDP per capita ($). The dataset is split into a training set and a test set. The split ratio is 70% vs. 30%. When fitting the random forest model, all the five predictors were set to be considered for each split of the tree, and the number of trees were set to be 20. After applying the model to the test set, the MSE for the test set is around 110. As shown Figure 6, the overall distribution of predicted proportion is similar to that of the real data. More real data are in the range between zero and five, compared the predicted values. The numbers from prediction and real data in other ranges are similar based on Figure 7.

|  |  |
| --- | --- |
| https://lh3.googleusercontent.com/BA5jTwIMX1rZTyhLjVbsCP02GZTo99ZlRRQ9Za2gNRBl6AGEHLX4JBlbQB-oTYIAidticJJjqrNdhKgdd3u-j_g4hLErypNfu8UuLNTO5nGqra1fYEPUkskhnZ6QDYrPvoO4eHRM | https://lh6.googleusercontent.com/Za35xT2wm0h7MvcC0rxAk24hHXSrYYnsaME-lzyL_5oNSEZyFN3t4MVlzxtBB5nL33ZFCrkfDAd_6Yie8LeaOHxRv489hH7wUwWjbKwgzhvr-AjfYG2fAY5L_RsaJIJKqVs1cClG |
| **Figure 6. Box Plot** | **Figure 7. Bar Plot** |

The results of importance function indicate that sex (74.8%: %incMSE) and continent (50.8%: %incMSE) are two most important variables across all trees to predict the suicide proportion in the random forest model.

**Data Limitation/ Bias:**

As we look in our dataset, we will find that there is some missing countries. Some of the most populous countries in Asia and Africa are not included, such as China and India. That means the dataset we have is biased. Although it lacks the prominent countries of Asia and Africa, most of the countries of North and South America and Europe are covered.

We should be aware with our assumptions while working on this dataset. That does not mean that the dataset is incorrect or we cannot use it. It means that our assumption will apply only on the countries we have data for.

Our visualization will still work if we get more data of the other countries. We just need to add the countries into the dataset and they will be included in our chart.

**Conclusion:**

We can conclude from this dataset, that some of the countries grow in economics over years and the suicide rates get affected. However, higher standards of living do not necessarily indicate lower suicide rates. Take United States for example, as the GDP increased over years the rates of suicide were also increased. Whereas, the GDP of Russia was stable over the years but the rates of suicide decreased.

Our Visualization can also employ log-transformation to the highly skewed data to make it less skewed. Our chart is also able to show trails, to highlight the countries depending on regions, to play or pause the years based on our needs, and to build connections among other aspects of countries, such as population, GDP etc.

**Future work:**

For future work, hopefully we can also include more countries in our dataset to make better comparisons and analysis. Also we can add gender to our drop list. When do modeling, validation or cross-validation method can also be employed to tune hyper parameters of the model.

**References:**

Dataset is Taken from: <https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016>

* United Nations Development Program. (2018). Human development index (HDI). Retrieved from <http://hdr.undp.org/en/indicators/137506>
* World Bank. (2018). World development indicators: GDP (current US$) by country:1985 to 2016. Retrieved from [http://databank.worldbank.org/data/source/world-development-indicators#](http://databank.worldbank.org/data/source/world-development-indicators)
* [Szamil]. (2017). Suicide in the Twenty-First Century [dataset]. Retrieved from <https://www.kaggle.com/szamil/suicide-in-the-twenty-first-century/notebook>
* World Health Organization. (2018). Suicide prevention. Retrieved from <http://www.who.int/mental_health/suicide-prevention/en/>