COGS9: Introduction to Data Science
Final Project
Due date: Thursday, March 19, 2020 23:59:59
Grading: 10% of overall course grade. 40 points total.
Completed as a group. One submission per group on Gradescope.

Group Member Information:

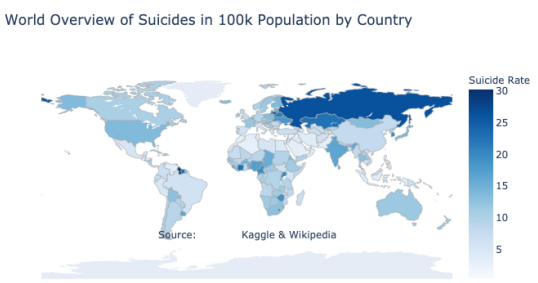
First Name	Last Name	PID
Huimin	**Zeng**	**U08498860**
Jie	**Zhang**	**U08474038**
Qingyang	**Chen**	**U08602061**
Tian	**Lan**	**U08397856**
Yinan	**Guo**	**U08608213**
Zhenrui	**Yue**	**U08438887**

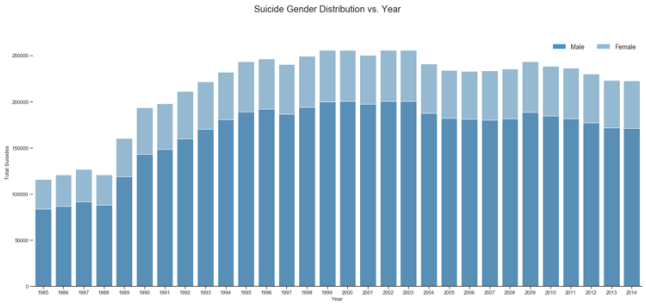
Question

It is estimated that around 800,000 people committed suicide every year, suicide is becoming one of the major causes of death in the world, outranked homicide (ca. 400,000), Parkinson's disease (ca. 340,000) and many other metal or physical illnesses. This is increasingly leading to misfortune and wealth loss to the whole society.

In this sense, we want to use data science to explore the relevant datasets and analyze questions like **what are the possible factors that could be related to these tragedies and what kind of measures could be taken to decrease the probability of suicide**, so that a better suicide prevention system and effective individual interference could be achieved, future losses could be significantly reduced based on our analysis.

Since we would like to extract the principal and decisive variables, we visualized the suicide rate in different countries with choropleth first and investigated all possible factors that could be related to the suicide incidents, then we analyzed the data statistically and found out the variables with statistical significance. A few suggestions were made based on the data analysis for future suicide prevention.





Hypothesis

Hypothesis 1: Poor working conditions could increase the probability of suicide. Work pressure is very common in certain developed countries such as Korea, where employees are required to work very long hours with relatively low productivity, interestingly, we also found out that these countries share very high suicide rates, suggesting that working conditions could be related to suicide ratio.

Hypothesis 2: Economic development could reduce the probability of suicide. A booming economy not only provides better jobs, increase people's salary and life quality but also make them feel secure and give hope of a promising future life. It is obvious that a person with increasing financial success and positive expectations of the future would less likely to have thoughts of suicide.

Hypothesis 3: Social Welfare could reduce the probability of suicide. Governmental investment might play a role in the prevention of this phenomenon as well. With more financial and supportive help from the society, it is more likely that individuals with more healthcare and psychological treatment available would have brighter mindset and fight against their suicide tendencies.

Hypothesis 4: Warm and sunny climate could reduce the probability of suicide. In general, warm weather and sunshine are preferred by most human. Researches show that sunny and warm climate are beneficial to people's mental health, while most countries in the tropics have relatively low suicide ratio. Hence, we made a hypothesis that warm climate and sunny days could reduce the probability of suicide.

Background Information

The world has been witnessing the rapid development in both technology and economics these years. It seems to be reasonable to assert that people are living a better life. However, according to WHO, suicide is a globally observed phenomenon and the suicide rates in many specific areas are increasing. According to the data from World Health Organization (WHO), it is

to observe that the suicide rate is surprisingly high in some developed countries, for instance Finland with 0.0138% compared to only 0.0055% in Italy.

It is assumed that the suicidal tendency is, to some extent, driven by the severe loneliness that a person is facing. That is, such people might struggle to escape from it as far as they can, however, in an extreme and tragic way. This is very common in both younger and older generations, the National Bureau of Statistics of England found that one of the top ten reasons young people attempt suicide is that they feel lonely. However, due to lack of accessible data and difficulties in quantifying loneliness, we decided to explore further possible factors related to suicide rate of different countries.

The National Survey of Midlife Development in the United States (MIDUS) II study (2004 - 2009) shows that about 11% of workers reported that they had suicidal thoughts, where 3% even reported severe suicidal thoughts. The result clearly showed that work stress and long-time working hours (> 40 hours per week) were significantly positively correlated with moderate to severe suicidal thoughts. This indicates that poor working conditions can be a serious risk factor causing commit suicide among working populations.

As for social welfare, it is natural to assume there is a connection between a country's suicide rate and welfare standard. When we explored the data provided by WHO concerning suicides per 100,000 people in 2016, we found that this relation is quite significant as it appears that both high and low suicide rates can exist in either developing or developed countries, but with increasing spending of social welfare, the suicide rate tends to decrease. Therefore, we need to do certain research on social welfare policies and budgets as percentage in GDP in different countries, before we could prove our hypothesis.

Last but not least, when we explored our data set of suicide rate among countries, we also found that countries with high latitude location (e.g. Russia, Norway and Finland) suffer a relative higher suicide possibility. These countries have relatively low average temperature and fewer sunny days per year compared to those with lower suicide rates such as Mexico and Malaysia. That could also be one of the reasons that change the suicide rate in different locations, thus, we will also consider this feature in the analysis.

Data

For datasets we need to analyze for this question, the observations should be in a reasonable recent period (10-15 years), covering most of the major economies (OECD countries), with all different suicide attempter and committer groups and as many observations as possible (20,000 observations would be a minimum). Also, details of individual cases should be a big plus, including but not limited to age, education, living area, income status, marital status etc. Since we are also looking for other possible reasons that could trigger suicide, the variables should include most of the relevant information of suicides (ideally, include all details of suicide incidents without violating individual privacy). In practice, we would also collect certain

qualitative information and quantify these for the convenience of analysis (e.g. loneliness, stress, emotional stability etc.). General information of investigated countries / areas or even time periods that reflects the living environment, cultural background or even historical events could also help clarify cases, such as “social welfare”, “annual average sunny days”, “absolute latitude value” etc.

After searching for various datasets related to suicide incidents around the world, a suicide dataset with the name “Suicide Rates Overview 1985 to 2016” from Kaggle that could be used to analyze this question. The dataset collects 27.8k observations range from 1985 to 2016, and adopts features like country, year, sex, age, suicides number, population, suicides / 100k population, HDI for year, GDP for year, GDP per capita and generation as variables. The limitation of this dataset is that it doesn’t provide any qualitative variables (not contained in the dataset) and details for cases to support our hypothesis such as the relationship between working conditions and suicide. The variables in the dataset can only reflect quantitative information like age and average income as well as their correlations to suicide rates.

Hence, we also utilized other data of smaller sizes from Wikipedia, OECD and World Bank to support our hypothesis involving other geographical and economic variables as well as other details related to suicide incidents. A list of helpful databases is listed below:

This dataset from Kaggle worked as a “base” in our research:
Suicide Rates Overview 1985 to 2016

Followings were used to estimate the working hours, conditions and stress among countries:
Wikipedia: Working Time (Annual Working Hours)
Wikipedia: List of countries by GDP per hour worked (Productivity)

These datasets were used to quantify the economic development of a country:
Wikipedia: List of countries by GDP (PPP) per capita
OECD Data: Average wages by country

This dataset helped us to evaluate the effects of social welfare among countries:
Wikipedia: List of countries by social welfare spending

Following datasets helped us to determine the relationship among suicide rate and sunshine duration as well as latitude position:
Wikipedia: List of capital of countries by latitude.
Wikipedia: List of countries by sunshine duration

Ethical Considerations

Team Bias: During data preparing and preprocessing, certain data entries might be manually altered or even filtered based on subjective criteria suggested by the whole team. However, the

criteria might not be entirely fair and objective, this could have a negative impact on the accountability of data.

Sampling Bias: We should pay more attention to the collection of suicide data for people in a certain age group, job position, their city or their country, especially the ones we are more familiar with as students, which could eventually lead to an unbalanced dataset and a biased conclusion.

Data Bias: If we utilize available datasets from the Internet, we would have to pay attention to the provided data and test it on fairness and accountability. Based on a biased dataset, it is most likely that the model and conclusion are not solid, thus, it's crucial to select unbiased dataset and base our analysis on it.

Consent: When we collect data to analyze different causes of death and reasons of suicide, this could negatively affect the emotions of the deceased's families and lead them to grieve. Therefore, it is necessary to acquire informed consents on the mentioned individuals from their relatives

Data Privacy / Ownership: Personal information and other privacy data could be leaked when gathering the data and using them to generate textual or visual analysis in the project, such as examples in the report describing the suicide of an individual.

Algorithmic Bias / Discrimination: Preparing and processing data: the bias might occur when we quantify the qualitative variables to fit a statistical model, for instance loneliness. it is likely that our evaluation on the level of loneliness could be biased due to our personal experience and knowledge.

Transparency: Lack of transparency of analysis methodology, statistical model selection and evaluation criteria. The lack of transparency could accumulate during the whole process from data processing, model training to conclusion drawing, which could significantly impair the transparency of the entire project.

Unintended Consequences: When the analysis is published, the relevant districts, companies and minority groups with higher rate of suicide could bear public prejudice or malicious comments. Social stigma of the mentioned individuals in the report Reader's emotions might be negatively affected

Continued Monitoring / Accountability: The accountability of the data science project is essential to the project since the model and conclusion are based on accountable data. It's absolutely important to guarantee the accountability and traceability of the data processing pipeline and avoid intentional / unintentional data manipulation.

Analysis Proposal

Data Collection:

- We searched “Suicide Rate” on Kaggle, find out data of suicide rate overview from 1985 to 2016, and then download the available data in tabular format.
- Collected data from Wikipedia, search for unemployment rate, working hours, GDP per capita, social welfare spending, minimum annual leave, etc. Explored the web API from OECD Data and World Bank for further data, recorded all data in tabular format (xlsx, csv).

Data Wrangling:

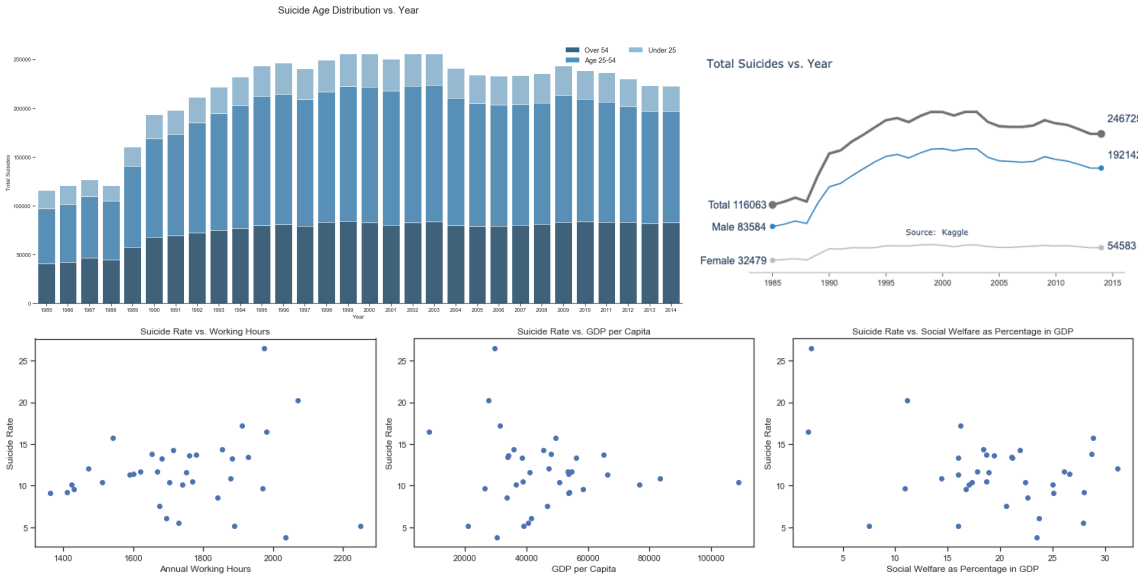
- Filter and delete invalid & missing entries.
- Select suicide rate / total incidents with every variable (e.g. GDP per capita, age, etc.) into subset datasets.
- Combine data groups in same category for exploratory analysis (combine age and gender groups for overall evaluation).

Descriptive & Exploratory Data Analysis:

- Look for outlier data: According to overall data distribution, find and remove outlier data (e.g. extremely high suicide rate of one region in one year compared with other years).
- Generate plots with different variables as axes to explore relationships between suicide rate and possible factors, we utilized different kinds of plots to capture the inner relations of different variables.

Data Visualization:

- Histogram: Show distribution / proportion of total suicide incidents.
- Line chart: Show trend of the number of suicides over time, as well as the number of female suicides and the number of male suicides.
- Scatterplot: Use scatter plots to analyze the correlation between suicide and other variables (Annual Working Hours, Unemployment Rate, Social Welfare, Latitude, Sunny Days, etc.) to show the relationship between two variables.

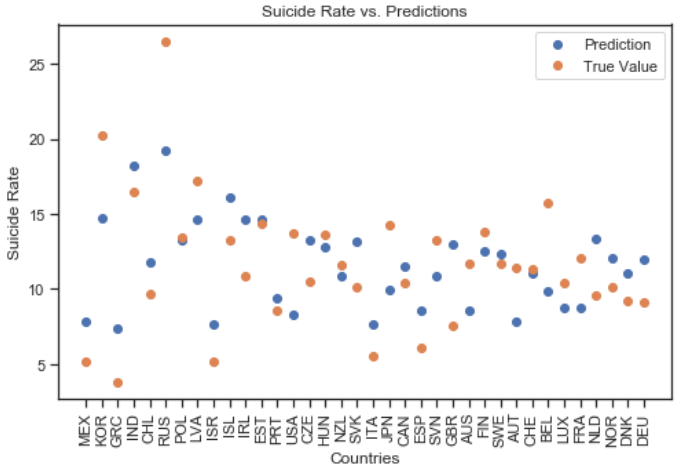


Statistical Analysis: (Inference)

- Correlations: We used the scatter plot to analyze the correlations between two different variables from which we find that there are different correlations between suicide rate and other variables with different statistical significance, for example, we found a negative correlation between ratio of social welfare ratio in GDP and suicide rate with very high confidence (Pearson correlation with low p-value <= 0.05), suggesting that an increase of social welfare could effectively reduce suicide incidents.

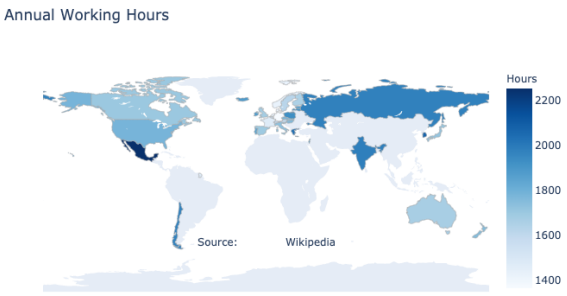
Predictive Analysis:

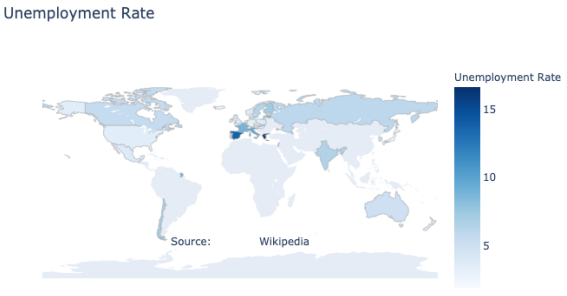
- Supervised Learning (Linear Regression): we used part (80%) of the data as the training set, fit all relevant variables into a linear regression model and generate suicide rate predictions for both training and test data. Results show that our predictions are not far away from the ground truth, therefore we can use the model to predict the suicide rate with given data.

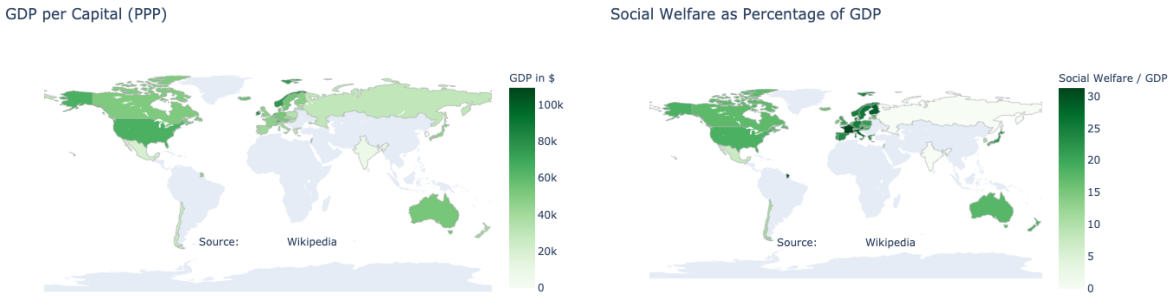


Geospatial Analysis:

- Choropleth: visualized clear regional differences in suicide rates with all data collected, certain variables such as social welfare as percentage of GDP and GDP per Capita were also visualized as choropleth. We compared suicide rate choropleth with others and came up with different ideas of the relationship between these factors and suicide rate. Following are a few examples of geospatial analysis with OECD data.







Discussion

How would you interpret the results of your proposed analysis? What are the limitations, pitfalls, and potential confounds of your methods, or biases in your data sources (e.g., how does the selection of the sources of your crowds affect your outcomes?)? How would you set out to address them? In addition, outline how you would address any societal and/or ethical implications of your proposed project discussed in your Ethical Considerations section. (10 pts)

After using the mentioned methods to analyze the data, we finally came up with several distinctive conclusion.

- First of all, we found that suicide rate has a correlation with working conditions in different OECD countries. We computed Pearson the correlation between Working Hours + Productivity and Suicide Rate data, the p-value 0.020 confirmed that working conditions are related to suicide rates in different countries, specifically, longer working hours would increase, while higher productivity decrease the probability of suicide.
- Then, we assumed that economic development and Suicide Rate might share strong connections as well, so we evaluated unemployment rate, GDP per capita and average annual wage in OECD countries, but the result showed that both these factors don't have statistical significance (p-value 0.365).
- Third, we explored the relationship between social welfare and suicide rate, there's an obvious negative correlation between Social Welfare as Percentage of GDP and Suicide Rate as we viewed the data with p-value of 0.021, suggesting that more investments in social welfare could actually reduce the suicide incidents.
- Afterwards, we tested the assumption that there's a strong correlation between Climate and Suicide Rate. When we combined the two factors Latitude & Sunny Days together and computed the Pearson correlation coefficient, p-value of 0.042 indicates that Climate would influence the decision of committing suicide, as warmer and sunnier weather (lower latitude and more sunny days) would effectively decrease suicide rate of a country.
- In addition, as we visualized the distribution of Suicide Age and Gender, we can see that suicide incidents are more frequent in male than in female, middle-aged and elderly people are also more likely to commit suicide, which could be traced back to factors like work pressure and loneness.

However, there are still some limitations and pitfalls we need to pay attention to. To begin with, we may not have collected enough data from all countries with all possible details. In other words, we might face some trouble caused by bias if our data are not representative. Therefore, more representative data should be collected in order to let us make more precise predictions and conclusions. Also, our data may not be precise enough. For example, many of our data did not show the exact time or exact age of people who commit suicide. Hence, we need to classify the data into data groups so that the data could be compared with meaningful control groups.

Considering the ethical part, we also need to improve on teamwork and analysis details. For the team bias consideration, different people may consider different perspectives during the data wrangling and interpretation, this may cause controversial opinions affecting the fairness of the result. Therefore, we need to put away our distinctive thoughts and focus on the objectivity of data. What's more, sampling bias could also happen because most of our data was collected online from OECD countries which indicates that we might ignore a big hidden part of the statistics in the rest of the world, and there were no effective methods to evaluate the data regarding its fairness and accountability. Hence, more representative and comprehensive data should be collected to for the analysis, we should also improve our strategies dealing with data to guarantee unbiased and representative conclusions. Furthermore, we weren't too careful with the protection of data privacy as the data don't contain specific suicide case details, thus, we need to remedy the methods with data analysis and privacy protection.

Group Participation

Include one paragraph briefly outlining the contribution of each group member throughout the quarter while working on this project. Each of you must also fill out the survey (link provided toward the end of the quarter) about individual and group participation. (3 pts)

Huimin Zeng: Wrote some python code and created various machine learning models for suicide rate prediction, he also wrote and reviewed some of the report.

Jie Zhang: Searched accessible and reliable datasets for our research. For the final report he wrote Background Information and Data.

Qingyang Chen: Chose appropriate data processing methods, used various plots to analyze data, and clarified the moral issues behind data analysis.

Tian Lan: Analyzed the proposed data and interpreted the result. Discussed the limitation and pitfalls and explored the solution of ethical considerations.

Yinan Guo: Decided the methods of data analysis. In charge of data analysis and ethical consideration part in final report.

Zhenrui Yue: wrote most of the python code to analyze data, generate visualizations (including choropleth) and used regression models for suicide rate prediction, he also reviewed the final report.

Reference

[1] https://en.wikipedia.org/wiki/Suicide_in_the_United_States
[2] https://en.wikipedia.org/wiki/Loneliness
[3] http://apps.who.int/gho/data/node.main.MHSUICIDEASDR?lang=en
[4] https://blogs.cdc.gov/niosh-science-blog/2018/09/13/suicide-prevention
[5] https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016
[6] https://en.wikipedia.org/wiki/List_of_countries_by_suicide_rate
[7] https://en.wikipedia.org/wiki/Working_time
[8] https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(PPP)_per_hour_worked
[9] https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(PPP)_per_capita
[10] https://data.oecd.org/earnwage/average-wages
[11] https://en.wikipedia.org/wiki/List_of_countries_by_social_welfare_spending
[12] https://en.wikipedia.org/wiki/List_of_national_capitals_by_latitude
[13] https://en.wikipedia.org/wiki/List_of_cities_by_sunshine_duration