**Group Member Information:**

Assignment will be completed in a group of 4-5 people. These will be the same people you work with for the final project. This assignment will help get you started on the final project.

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

**Question**

Clearly state the specific data science question you're interested in answering. (5 pts)

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 reasons that could ultimately trigger 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. Finally, we would like to extract the principal and decisive components / features, so that future work can concentrate on fewer and more concise data.

**Hypothesis**

Write down your group’s hypothesis to your question. Provide justification how you came to this hypothesis. (What background information or instinct led you to that hypothesis?). (10 pts)

Hypothesis 1: **Loneliness could increase the probability of suicide**. Given the intuition that humans are social creatures, loneliness might be one of the most vital factors that impacts people’s living attitude. Segregated apart from the majority, one could suffer more from the negative emotions and therefore tends to have higher probability of suicide behavior.

Hypothesis 2: **Work pressure could increase the probability of suicide**. Work pressure is very common in certain developed countries such as Korea and Japan, where employees are required to work long hours under great pressure, interestingly, we also found out that these countries share very high suicide rates, suggesting that too much work pressure could be related to suicide incidents.

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.

**Background Information**

Include a few paragraphs of background research and information on your topic. This should include at least 2 citations to work from others. Including hyperlinks to reputable sources is fine. (10 pts)

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 [1].

With revolutionary cyber and transport technology, it is more convenient for people to communicate with others, governments around the world invest more in welfare to help this situation thanks to the economy prosperity, but many people tend to suffer even more from loneliness [2]. Therefore, we would like to explore the latent relationship of this phenomenon with other factors, such as loneliness, work pressure and social welfare.

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. For instance, homesick and unfamiliar environments could bring loneliness to many people, without proper coping skills, it is likely that they tend to have a higher suicide rate [3].

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 [4], we found that this relation is not significant as it appears that both high and low suicide rates can exist in either developing or developed countries. Therefore, we need to do more research on different social welfare policies in relevant countries, before we could prove or disprove our hypothesis.

In addition, 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 [5]. 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 long-term work stress can be a serious risk factor causing commit suicide among working populations.

**Data**

Include a description of the perfect dataset you would need/want to answer this question. How many observations would you need? What variables would you collect? Explain the perfect dataset that you would want to answer this question.

Then, look online for available datasets. Find a dataset that could be used to answer this question. Describe how many observations are included and what variables have been collected. Discuss the dataset’s limitations and how it differs from your ideal dataset. Alternatively, you can collect your own data. You must explain what information you will collect and how you are going to collect these data. These do not have to be collected by the time this proposal is submitted, but they must be collected by the time the final project is submitted. (5 pts)

For the perfect dataset 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 as many as possible, 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”, “work ethic”, “natural disaster” 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 was found: Suicide Rates Overview 1985 to 2016 [6].

The dataset collects 27.8k observation 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 loneliness 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 are also expecting to find other databases of smaller sizes to support our hypothesis involving qualitative variables and other details related to suicide incidents. However, such databases are either not accessible or come with very limited usability. If not available, it’s also possible to collect data ourselves, although the collection could be extremely difficult and possibly causing pain for the relatives, and collected data might not be very representative or even biased due to limited mobility and group size.

**Ethical Considerations**

Read the data science ethics checklist from lecture. Then, discuss what ethical considerations must be made when answering your specific data science question. Brainstorm and explain how you would address these considerations for each of the following categories in your specific project: Team Bias, Sampling Bias, Data Bias, Consent, Data Privacy / Ownership, Algorithmic Bias / Discrimination, Transparency, Unintended Consequences, Continued Monitoring / Accountability. Feel free to write about additional ethical considerations you would make that aren't included on the checklist. Note that data privacy is NOT the only ethical consideration for a data science project. It is a piece, but there is a lot more that has to be considered. (10 pts)

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.

[1] <https://en.wikipedia.org/wiki/List_of_countries_by_suicide_rate>

[2] <https://en.wikipedia.org/wiki/Suicide_in_the_United_States>

[3] <https://en.wikipedia.org/wiki/Loneliness>

[4] <http://apps.who.int/gho/data/node.main.MHSUICIDEASDR?lang=en>

[5] <https://blogs.cdc.gov/niosh-science-blog/2018/09/13/suicide-prevention>

[6] <https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016>