**March 15, 2020**

# SUICIDE MORTALITY RATE AND SOCIO-ECONOMIC FACTORS IN THE UNITED STATES

**PRESENTED BY:**
Saruul Khasar

## TABLE OF CONTENTS

## I. SUMMARY

1. **How does the U.S. suicide rate trend look like? Is it really increasing?**

Result: Past 18 years of the U.S. suicide rate data shows us a fast-increasing trend. Comparing the most recent data with a trend line drawn by past data (from 1997 to 2007) shows us that the recent data is diverging far from the past trend line. It implies the suicide rate in the U.S. is increasing at a concerning rate.

2. **Do suicide rates differ among 50 states in the U.S.?**

Result: Yes, it does. Midwest and eastern part of the West U.S. has the highest suicide rates. Wyoming, New Mexico, and Montana have the highest suicide rate in the nation; whereas, Rhoda Island, New York and New Jersey has the lowest suicide rates in the nation.

3. **In states with highest and lowest suicide rate, what pattern do we see in the socioeconomic factors to suicide rate? Does it follow other research outcomes?**

Result: We see an interesting pattern in socioeconomic factors to suicide rates. States with highest suicide rates do not necessarily have the worst socioeconomic factors. For example, unemployment in these states are higher than the national average; however, in the states with lowest suicide rates we see an unemployment even higher than the national average. GDP growth was higher in the states with highest suicide rates as well as the education attainment.

4. **Do socioeconomic factors really look like a leading cause to this increasing suicide rate? And what can each state do about it?**

Result: Given the data we have, we don't see any worse situation in the states with high suicide rates. These states with high suicide rate have favorable employment situation, favorable economic and favorable education attainment rate. There are other factors such as isolation may cause the suicide rate more than these socioeconomic factors. The Economist article on increasing suicide rate in the U.S. mentioned that studies conducted by Ohio State University and West Virginia University found that isolation may be an important factor to suicide rates. However, another criticism to my analysis is that state level data is very limited (only past 5 years) in this analysis. If we had a data spanning for 10 years, then things might look different.

## II. MOTIVATION AND BACKGROUND

On January 30, 2020, The Economist magazine published an article titled "America's suicide rate has increased for 13 years in a row" and the article says those living in rural and less-populated areas have been hit especially hard. Another article published by the American Psychological Association says that the U.S. has the highest suicide rate among any wealthy nations.

Suicide rate is an indication of existing social and mental problems in the society. Therefore, countries work toward decreasing this rate. Understanding what causes high suicide rate is an important first step to work toward alleviating the factors.

When I read this article, I was curious whether the suicide rate is high in every state in the U.S. and whether the overall rate is really increasing as the article suggests. I was curious what could be causes to this increasing suicide rates especially in those states with the highest suicide rates.

Studies say the suicide rate is strongly associated with socio-economic factors such as unemployment, low income, education and inequality. There are other factors such as isolation, culture, history etc. However, I thought these socio-economic factors is an interesting starting point to really pinpoint what might be the causes of these increasing suicide rate.

## III. DATASET

State level public dataset was limited; therefore, I am going to use what is available publicly which is 5 years of data from 2014 to 2018.

| | Indicator | Definition | Source | Dates | Data format |
|---|---|---|---|---|---|
| 1. | Suicide rate | The number of deaths per 100,000 total population. Adjusted by age-distribution and population size. | Centers for Disease Control and Prevention http://bit.ly/3axglEU | 1999-2018 (national level)<br><br>2014-2018 (state level) | .csv |
| 2. | Real GDP growth | Real GDP by state: All industry total (Percent change from preceding period) | Bureau of Economic Analysis http://bit.ly/3aCDlme | 2014-2018 | .csv |
| 3. | Unemployment rate | Unemployment as a percentage of the labor force. | U.S. Bureau of Labor Statistics http://bit.ly/38DrX7X | 2014-2018 | html and .csv |
| 4. | Education | Percentage of population 25 years and over with bachelor's or above education | United States Census Bureau http://bit.ly/2W2tQsa | 2014-2018 | html |

## IV. METHODOLOGY

### Part I. Store datasets into Python

In this part of the project, I will import tabular and html data from different sources and save the file as pandas dataframe. I will clean the data by extracting important columns in each dataset and also adding columns or rows that are missing in the dataset. Following are the list of datasets that I will import into Python:

i) Unemployment (from webscraping and .csv file)
ii) Education attainment (from .csv file)
iii) Suicide rate (from .csv file)
iv) GDP growth (from .csv file)
v) US map (from .json file)

### Part II. Preliminary data analysis

This part has three sections:

i) Time trend analysis

I will write a function to find plot time trend on aggregate suicide rates data for the past 18 years (I only found 18 years of data). This function will train the data on 1999 to 2007 dataset and draw the data in terms of all-time horizon. The reason for choosing this time period for my training set is because it was a normal time (no financial crisis) economically in the US. This trend analysis will allow us to see whether the recent data follows the trend in the past.

ii) Top 3 and bottom 3 states

I will write a function to plot 3 states with highest suicide rate and 3 states with lowest suicide rate with bar chart. This function will sort the suicide rate data and extract top 3 and bottom states to use for the bar chart.

iii) States by suicide rate

I will write a function to draw U.S. map colored by the suicide rates in each state. This function will merge U.S. map data with suicide rate data and draw a map color differentiating each state by average suicide rate of the past 5 years (from 2014 to 2018).

**Part III. Socioeconomic factor analysis**

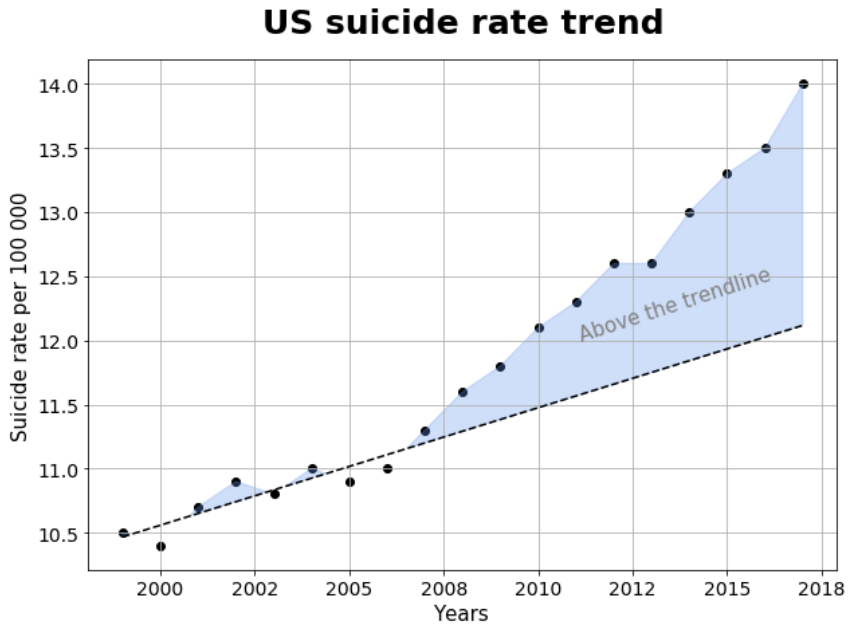
In this part I will write a function to plot data using GridSpec. GridSpec will allow me to show U.S. map geospatial data along with other trend datasets. This function will take several arguments, including all dataframes and an argument for top or bottom states, and number of states to draw in this visualization. For instance, if I give an argument of top and number 10 to this function along with all dataframes, this function will draw U.S. map on top highlighting the top 10 states with highest suicide rates, and on the bottom of this map we will see 4 trend diagrams. These 4 diagrams will have a trend line for average suicide rates in these states along with the national average, a trend line for average unemployment rate in these states along with the national average, a trend line for average education attainment in these states along with the national average, and finally GDP growth rate trend along with the national average. After writing this function, I will plot following two scenarios:

i) Top 10 states and corresponding socioeconomic factors
ii) Bottom 10 states and corresponding socioeconomic factors

## V. RESULTS

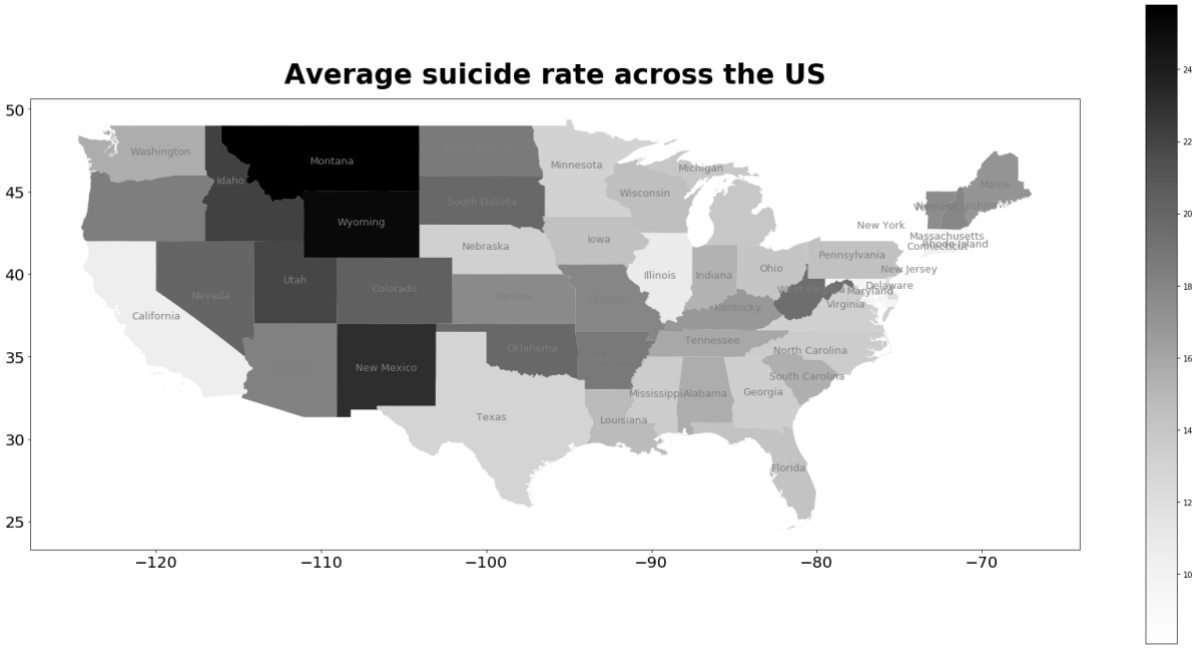
### 1. How does the U.S. suicide rate trend look like? Is it really increasing?

Plot below shows the suicide rate trend drawn by a data from 1997 to 2007 along with actual datasets. Data in recent years is well above this trend and even diverging faster from the past trend. This result implies the suicide rate in the U.S. is increasing at a concerning rate.

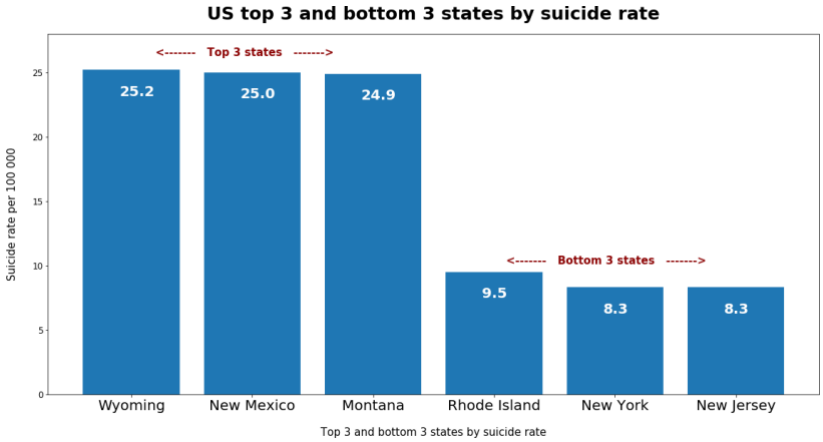



### 2. Do suicide rates differ among 50 states in the U.S.?

As the plot below shows Midwest and eastern part of the West U.S. has the highest suicide rates. Eastern part of the U.S. and West Coast has lower suicide rates.

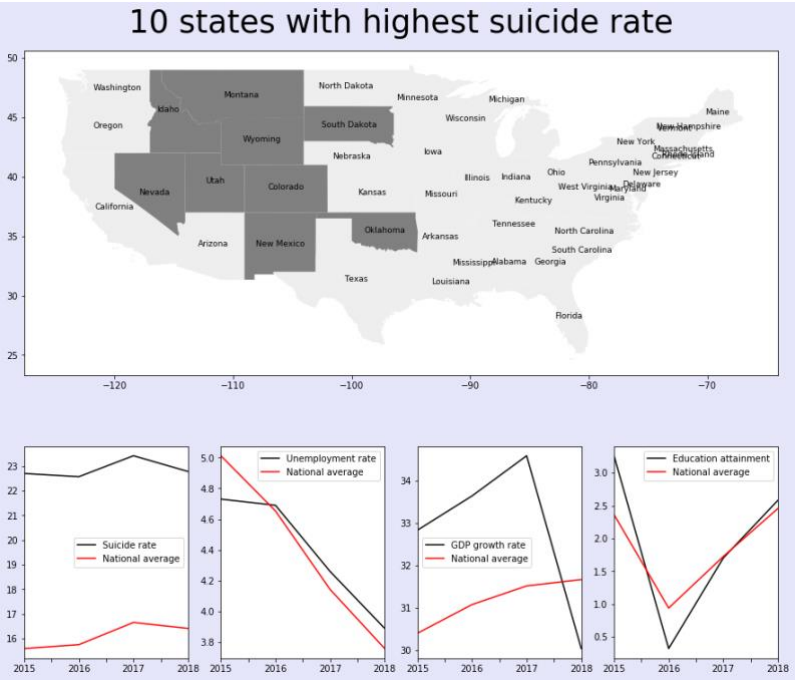


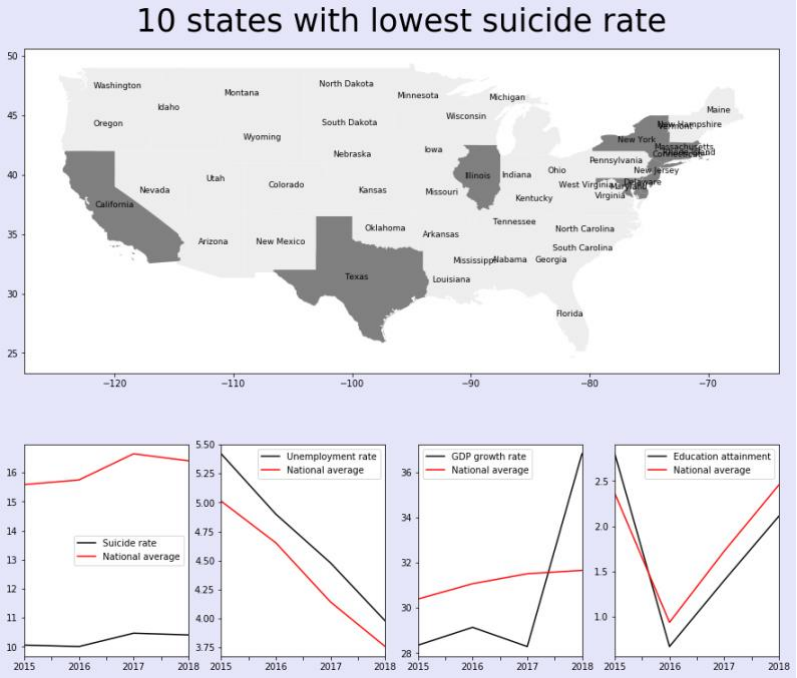
As the plot below shows, Wyoming, New Mexico, and Montana have the highest suicide rate in the nation; whereas, Rhoda Island, New York and New Jersey has the lowest suicide rates in the nation.




3. **In states with highest and lowest suicide rate, what pattern do we see in the socioeconomic factors to suicide rate? Does it follow other research outcomes?**

As the two plots below show, we see an interesting pattern in socioeconomic factors to suicide rates. States with highest suicide rates do not necessarily have the worst socioeconomic factors. For example, unemployment in these states are higher than the national average; however, in the states with lowest suicide rates we see an unemployment even higher than the national average. GDP growth was higher in the states with highest suicide rates as well as the education attainment.






## VI. REPRODUCING THE RESULTS

**Running project.py file:**

**Step 1.** Copy and paste ***saruul folder*** on to your desktop
**Step 2.** Open project.py file
**Step 3.** Change directory to ***data folder*** on row 488. Change the highlighted part to current directory to the data folder. Save the project.py file.

dir_to_data_folder = "/Users/saruul/Desktop/saruul/data"

**Step 3.** Open your terminal and go into the ***saruul folder***
**Step 4.** Open ipython on your terminal and ***run project.py***
All the images will be saved in your ***saruul folder*** with .png format.

**Running test.py file:**

**Step 1.** Open ***test.py*** file
**Step 2.** Change directory to data folder on row 12. Chane the highlighted part to current directory to the data folder. Save the test.py file.

dir_to_data_folder = "/Users/saruul/Desktop/saruul/data"

**Step 3.** Run test.py file from the terminal using ipython

## VII. TESTING

Since it's hard to test the accuracy of visualizations, I have decided to test whether I imported the right datasets. In order to do that, I had to filter my dataframes and confirm if it equals the dataset that is on the official websites. I filtered out my dataset in terms of random states and transformed the dataframe to numpy array and used np.array_equal() function to transform it into a numpy array. Then, I typed out the dataset I see from each website that I used to download or scrape my dataset and used np.array_equal function to see if it equals the filtered dataset I have in my program. I used this testing method on functions that I can test.