CSE564 Data Visualization

Final Project

Data Visualization on WHO Suicide Data

Motivation

Every year close to 800 000 people take their own life and there are many more people who attempt suicide. One person is died of suicide in every 40 seconds, the fact is frustrating. Suicide as a global phenomenon which occurs throughout the lifespan, is a tragedy that affects families, communities and entire countries and has long-lasting effects on the people left behind.

In my project, I try to use the data from WHO to visualize the suicide information about different countries, different genders, and different age groups. In addition, I try to find the correlation the correlation between the suicide numbers and other features, thus providing inspirations for relating researchers to conduct deeper work.

Data and Preparation

I develop the project with Python and JavaScript. I use D3 in JS as a technique to visualize and pandas in Python to do the data preparation. Between them, I use flask to transport the data from the frontend to the backend, thus realizing the update of different charts.

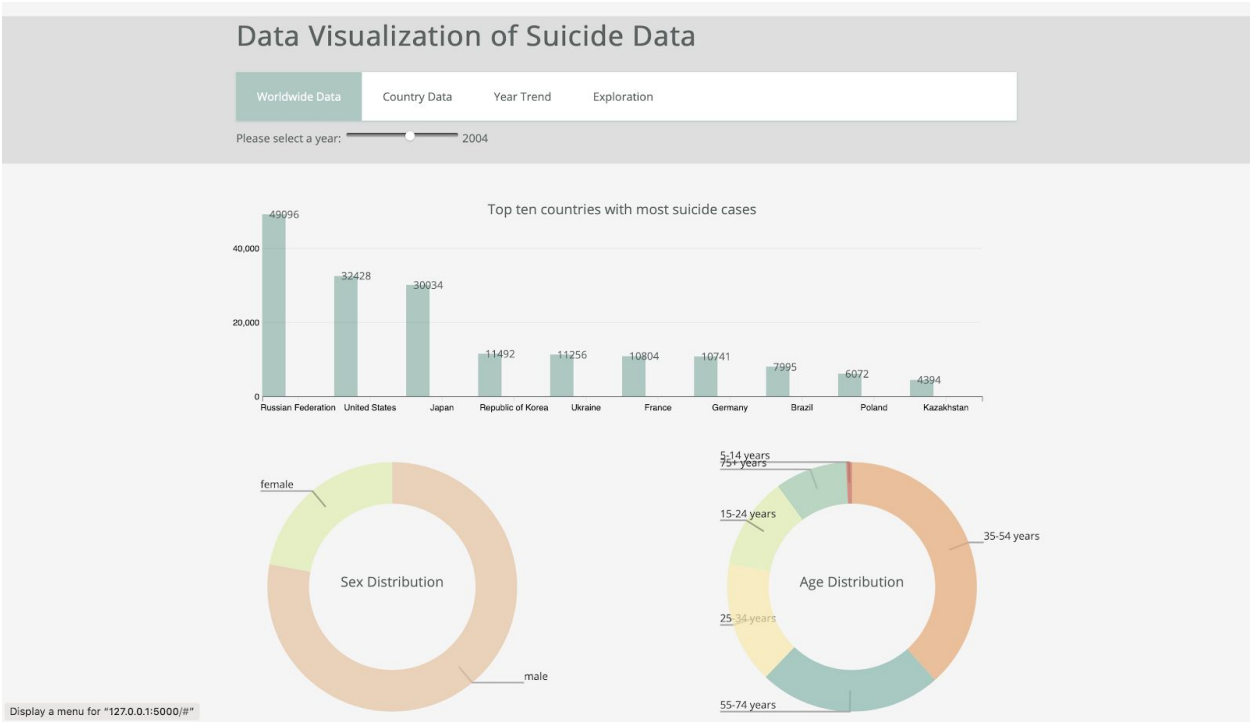
The data is about the suicide cases from 1987 to 2016. It was built to find signals correlated to increased suicide rates among different cohorts globally, across the socio-economic spectrum. It contains 27,821 rows and 11 columns in the data. The features show: `country`, `year`, `sex`, `age`, `suicides_no`, `population`, `suicides/100k pop`, `HDI for year`, `gdp_for_year ($)`, `gdp_per_capita ($)`, `generation` . Note that the `suicides_no` represents the number of suicide cases with respect to the different features.

Layout

There are 4 main parts in the whole webpage, respectively, Worldwide, Countrywide, Year Trend and the Exploration. Each of them can be evoked by clicking the corresponding menu. The design of the layout shows as follows:

Now, the detailed information about each part shows in the following report:

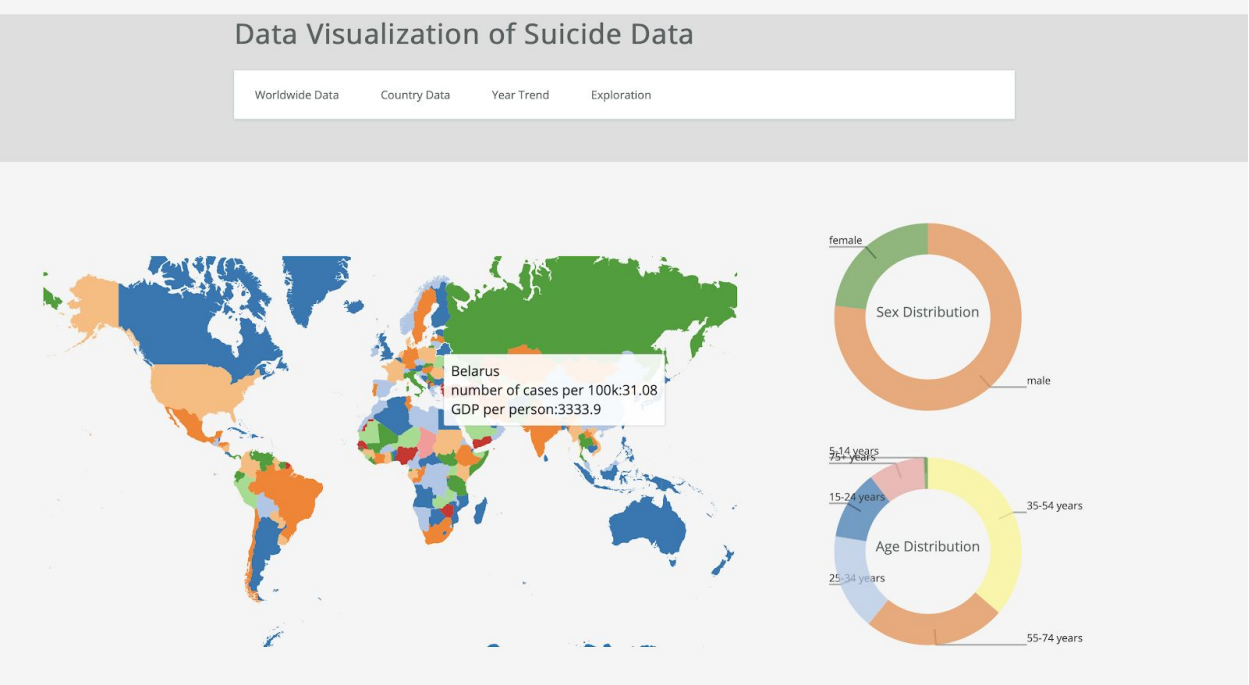
Worldwide



This part is designed to show the data worldwidely, users can choose a specific year they want to see by moving the mouse on the slider at the top. In the bar chart, I aggregate the data and find the top ten countries with the most suicide cases. Under the bar chart, there are two donut charts, each of them show the fraction of different groups. In this part, the data is directed from the source, which means that the chart can be updated from the change of the data source.

As the data set contains information of 101 countries, I found that the information of the countries with low cases is not accurate (which is suggested by the huge difference between two successive years). Thus, I decided to only show the countries with more cases. This part is the total number, the cases considering the population of the region will show in the country data.

Countrywide



In this part, I show a world map with tooltips, which provide information about the GDP information and number of cases considering the population. Users can move the mouse to the corresponding area they want to see and after clicking the area, the fraction of the gender and age group of this country will be shown as two donut charts beside the map. These charts can also be up-to-date.

I use the `toposon.json` to draw the map of the world, and note that the tooltips will only show when the data of the country exists in the data source. When the data does not exist, the tooltips will be hidden.

Year Trend

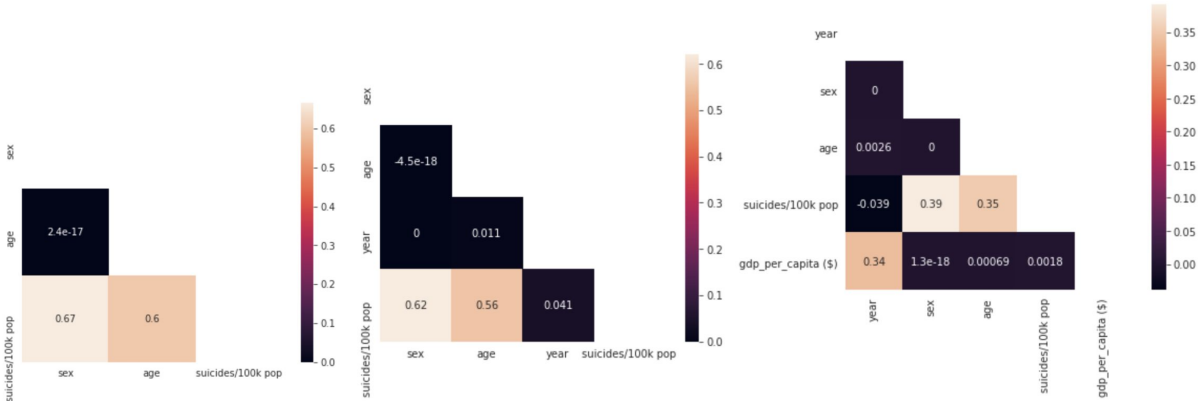
In this part, I show the trend of the number of cases of the countries with most cases through the whole time period. Users can use the menu to select the country and a tooltip containing the information of both axes will be shown when the mouse moves to the corresponding year.

Exploration

To fully observe the correlation between each feature and the `suicide_no` , firstly, I aggregate the data with respect to sex, age, gdp and generations. (For generations, I tried two methods to transform the name to the term that can be put into calculation. 1: sample an integer from a large random set to represent the name of the generations. 2: arrange it into successive numbers considering the time sequence of generations. After comparison between these two, I found there is little correlation.). The pearson correlation shows as follows:

```
sex : 1.0
age : 0.9415702060072109
gdp_per_capita ($) : -0.024948936887860214
generation : -0.6279750370920653
```

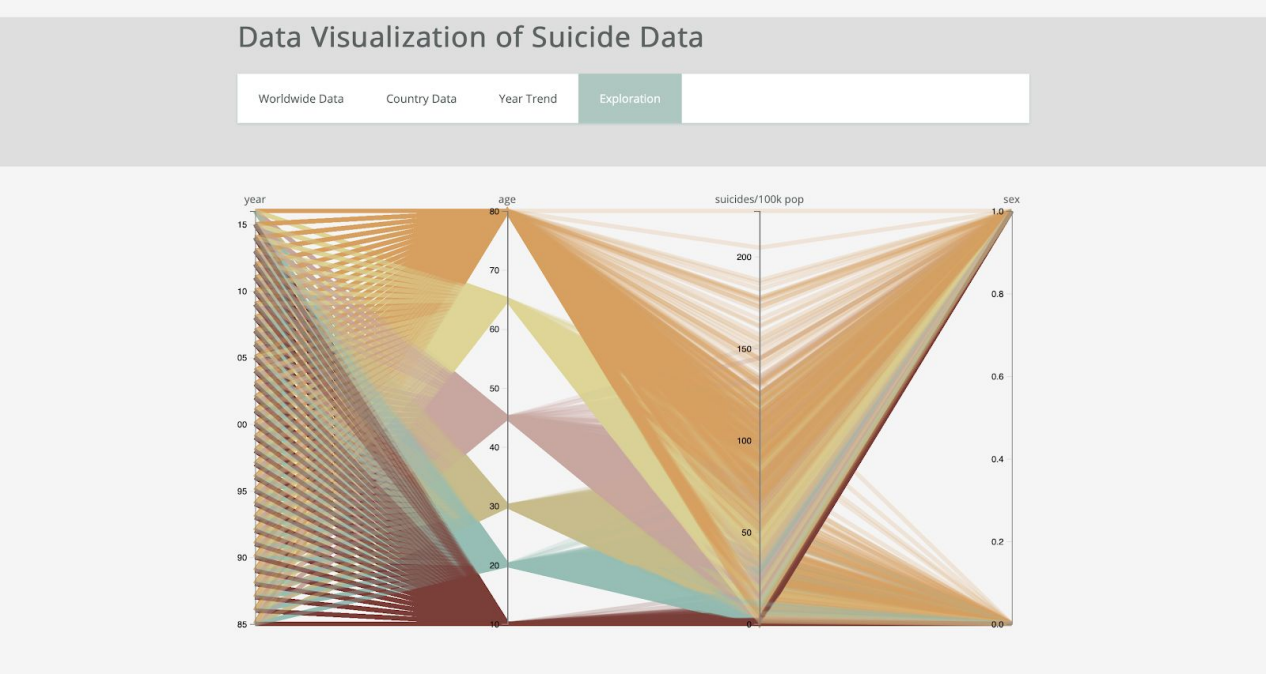
The above statistics show the correlation considering only one feature, in addition, I tried to add more features and find the correlation, the heatmap about the pearson correlation shows as follows:



We see a strong correlation between the number of suicide cases between age and gender. To my surprise, the GDP per person seems not to be a critical term. However, the exploration is still inadequate, and I will keep digging the correlation with other useful techniques.

To show the result of the correlation, I choose to draw parallel coordinates. I put the suicide rate in the middle of the gender and age, thus we can see the correlation clearly. Then, I

assign the different colors to the line by the age groups. Also, I implement the highlight function to enable users to focus on a specific age group, but there seems to be some bugs, it works on chrome but not on safari. I will try to fix these bugs.



Conclusion

In this project, I developed a webpage that allows user to view the suicide data from WHO from different perspective: worldwide, country wide and trending. In addition, I try to explore the correlation between the features and the suicide numbers and show it by a parallel cordinates.