# Mental health and substance use disorders -- Project final report


Yu Han
10/27/2020

# Introduction

Mental health and substance use disorders affect our life in all directions and are very common illnesses in the world. The prevalence of mental health and substance disorders includes depression, anxiety, bipolar, eating disorders, alcohol or drug use disorders, and schizophrenia. They involve changes in thinking, mood, and/or behavior. In 2017, there were more than one in ten people globally suffering from mental health disorder (10.7%). Mental and substance use disorders account for around 5 percent of the global disease burden in 2017, but this reaches up to 10 percent in several countries. The data showed mental health and substance use disorders are common everywhere but with different rates for different countries and with different rates by gender and age.

The goal of this project is to show people how mental health and substance use disorders look like in the world and different countries, how it behaves differently by ages and genders, and use unsupervised machine learning to detect outlier countries with gender rates data.

Hopefully, this project will make society organizations, the government, and general people like us pay attention to mental health and substance disorders and have an idea directly from this project to help people who are suffering.

# Datasets & Data wrangling

The datasets used for this project were produced by the Institute for Health Metrics and Evaluation and reported in their flagship Global Burden of Disease study. The original blog was published on Our World in Data (link). 8 datasets were downloaded from Our World in Data chart sources. All datasets included Entities (231 uniques), Country code, and Year (1990 to 2017) columns. Additional columns belonging to each dataset show below, and you can download them by clicking each title. All datasets are very clean. Only No. 7 and 8 needed to deal with null values. These two datasets include null values from 1800 to 1989, so I just dropped the null values since we don't need them. I also changed the long column names to brief ones. I then combined No 1, 3, 4, 6, and 8 for further exploration, and the rest of them explored individually. A brief explanation for each dataset shows below.

**1. Death rates from mental health and substance use disorders**

Data were age-standardized and measured per 100,000 individuals. Data do not include deaths resultant from suicide.

**2. GDP per capita**

Measured in constant international ($) from 1990 to 2017 for most countries in the world.

**3. Mental and substance use disorders as a share of total disease burden**

Disease burden is measured in DALYs (Disability-Adjusted Life Years). DALYs measure the total burden of disease, both from years of life lost and years lived with a disability. One DALY equals one lost year of a healthy life.

**4. Prevalence by mental and substance use disorder**

Share of the total population with a given mental health or substance use disorder that includes depression, anxiety, bipolar, eating disorders, alcohol use disorders, drug use disorders, and schizophrenia.

**5. Prevalence of mental and substance use disorders across age groups**

Share of the population by age groups suffering from any mental health or substance use disorders; this includes depression, anxiety, bipolar, eating disorders, alcohol or drug use disorders, and schizophrenia. The features include 10-14 years old, 15-19 years old, 20-24 years old, 25-29 years old, 30-34 years old, all ages, 5-14 years old, 15-49 years old, 50-69 years old, 70+ years old, and age-standardized.

**6. Share of population with mental health and substance use disorders**

Share of population with any mental health or substance use disorder; this includes depression, anxiety, bipolar, eating disorders, alcohol or drug use disorders, and schizophrenia.

**7. Share of population with mental or substance disorders, male vs. female**

Share of males vs. females with any mental health or substance use disorder; this includes depression, anxiety, bipolar, eating disorders, schizophrenia, alcohol, and drug use disorders, and neurodevelopmental disorders.

**8. Suicide death rates vs. prevalence of mental & substance use disorders**

Age-standardized suicide death rates, measured per 100,000 individuals versus rates of mental and substance use disorders per 100,000 individuals. This includes depression, anxiety, bipolar, eating disorders, alcohol or drug use disorders, and schizophrenia.

# Exploratory Data Analysis (EDA)

This part was explored separately by three different datasets.

**1. Combined dataset:**

The dataset includes a few prevalence mental health disorders rate, its share of total disease burden, the direct death from mental and substance use disorders, and suicide rate presented in 231 entities from 1997 to 2017. Here are the findings I explored.

- I first checked the outlier countries and noticed that all outliers are upper outliers and outlier countries are either big countries or rich countries.

- Then I checked the correlation coefficient between death and suicide rate with mental health disorders and visualized the relationship with the regression plots. The correlation coefficient showed that direct death from mental and substance use disorders has a higher positive correlation with eating disorders, but I barely see it from the regression plot between two variables. The correlation coefficient also showed the suicide rate has a positive correlation with depressive disorders and alcohol use disorders, but we don't see relationships between suicide rate and depressive disorders from the plot. After removing the obvious outlier points over 50 by the y-axis, we do see a positive linear correlation between suicide rate and alcohol use disorders (showed in Fig.1).

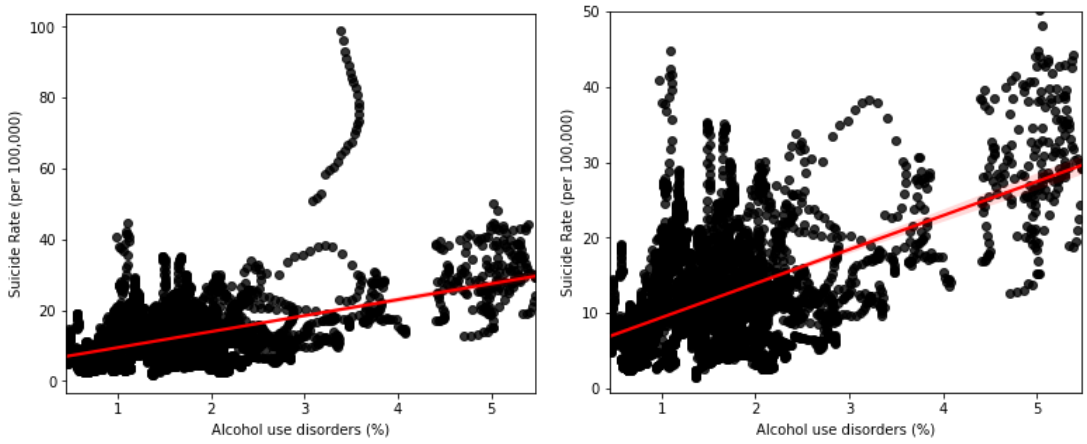


*Fig.1 These two regression plots showed the correlation between suicide rate and alcohol use disorders. (Left) Plotted with the whole data. As you can see the points on the top are extremely different and impact the visualization. (Right) Plotted it again by changing the y-axis. We can see there is a positive correlation relationship between suicide rate and alcohol use disorders.*

- I check how mental and substance use disorders increase in decades. I notice that the top 3 countries with the biggest increase for each decade are different. And for most features, the biggest increase has decreased decades in decades. Table 1 showed the top 3 countries with the biggest increase for mental and substance use disorder as an example. Then I checked the top 3 countries with the biggest increase for each decade (shown in table 2). Very interesting, only Japan is increasing with increasing change in decades. Australia and Italy have increased in population from 2000 to 2009. South Korea and Libya have decreased from 2000 to 2009 and jump back increased from 2010 to 2017.

| 1990 ~ 1999 | | 2000 ~ 2009 | | 2010 ~ 2017 | |
|---|---|---|---|---|---|
| Entity | | Entity | | Entity | |
| Afghanistan | 0.849154 | Australia | 0.372023 | Libya | 0.234751 |
| Brazil | 0.797578 | Somalia | 0.368775 | Japan | 0.200297 |
| Netherlands | 0.625952 | Italy | 0.250366 | South Korea | 0.168606 |

*Table 1. Mental and substance use disorders (%)-- top 3 countries with the biggest increase*

| Entity | 1990 ~ 1999 | 2000 ~ 2009 | 2010 ~ 2017 |
|---|---|---|---|
| Afghanistan | 0.849154 | -1.085967 | -0.372593 |
| Brazil | 0.797578 | -0.173188 | -0.404024 |
| Netherlands | 0.625952 | -0.588988 | 0.075779 |
| Somalia | 0.481309 | 0.368775 | -0.609908 |
| Australia | 0.179127 | 0.372023 | -0.237284 |
| South Korea | 0.157959 | 0.030811 | 0.168606 |
| Libya | 0.101766 | 0.039928 | 0.234751 |
| Japan | -0.052152 | 0.141358 | 0.200297 |
| Italy | -0.600999 | 0.250366 | 0.025461 |

*Table 2. Mental and substance use disorders (%)-- top 3 countries with the biggest increase for each decade.*

- Next, I checked the relationship between mental or substance use disorders rate and GDP per capita. The correlation coefficient shows GDP per capita has stronger positive correlations with mental health and substance use disorders as a share of total disease burden (r=0.73) and eating disorders (r=0.75). The regression figure shows below in fig.2.

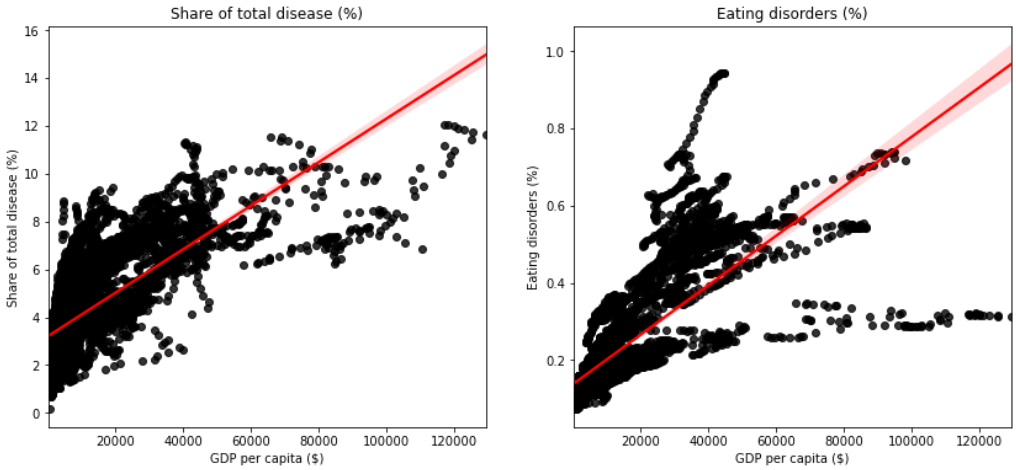


*Fig.2 (Left) Regression relationship between mental health and substance use disorders as a share of total disease burden and GDP per capita. (Right) Regression relationship between eating disorders and GDP per capita.*

- From the above, we know that the economy is a factor for mental and substance use disorders. So I want to see how mental health and substance use disorders differ between rich and poor countries. I filtered the top 3 richest and poorest countries from

the GDP per capita dataset (based on 2017). Then I plotted the eating disorders (shown in Fig. 3) and Mental health and substance use disorders (shown in Fig. 4) trend for each country from 1990 to 2017.

- For Eating disorder:
  - All rich countries have much higher rates than poor countries.
  - All rich countries show an increased rate year by year.
  - All poor countries show lower rates and slight changes over the years.

- Mental health and substance use disorders
  - The rich countries tend to have more people rates suffering from mental and substance use disorders, and poor countries have less.
  - Singapore(rich) has less rates than Burundi(poor).
  - The rates tend to decrease through the years.

In summary, rich countries tend to have more people rates suffering from mental health and substance use disorders than poor countries.

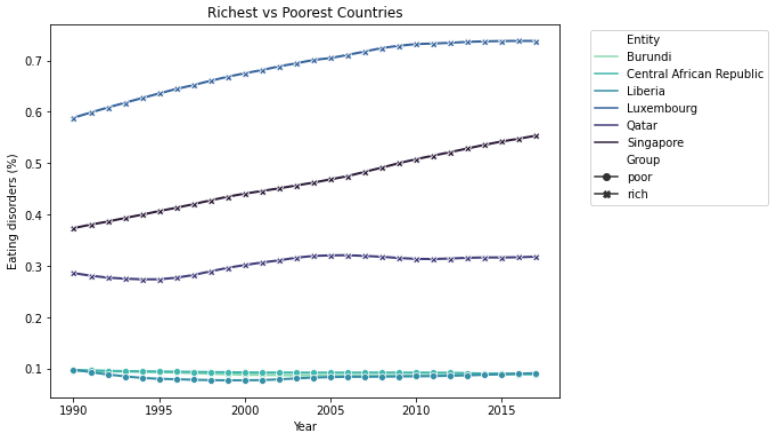


*Fig.3 Eating disorder trends by years to compare richest and poorest countries in the world.*

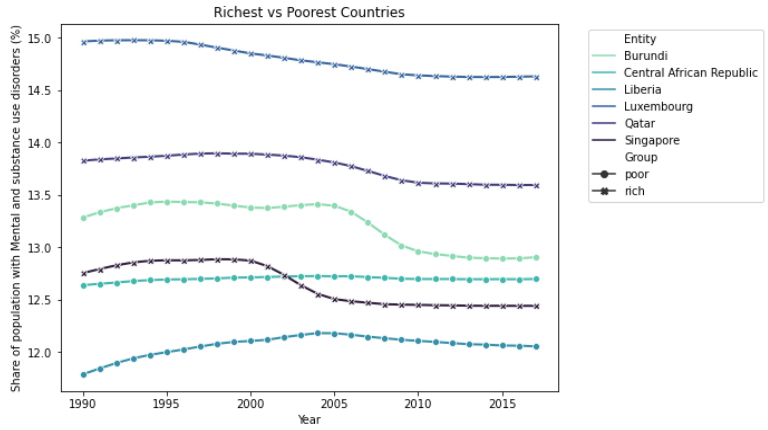


*Fig.4 Mental health and substance use disorders trends by years to compare the richest and poorest countries in the world.*

**2. Explore the share of population by age groups suffering from any mental health or substance use disorders data.**

This dataset includes a few age groups that cover most age ranges. On average, it shows younger people who are under 14 years old have the lowest rate, and people who are between 15-49 years old have the highest rate. Dut to 15-19 years old, 20-24 years old, 25-29 years old, 30-34 years old age groups have slightly different on average, 5-14 years old, 15-49 years old, 50-69 years old, 70+ years old, and age-standardized columns in the whole world level will be chosen for the future study, and also because those groups cover most age range. Here are some findings I explored.

- I visualized 5-14 years old, 15-49 years old, 50-69 years old, 70+ years old, and age-standardized age groups in the whole world level (shown in fig. 5). Mental health and substance use disorders rate is decreasing for most age groups in the whole world level through 2000 to now. The 50-69 years old age group shows the highest percentage, the 5-14 years old age group shows the lowest percentage. The 15-19 years old age group shows the second-highest percentage, and it is going up in recent years.

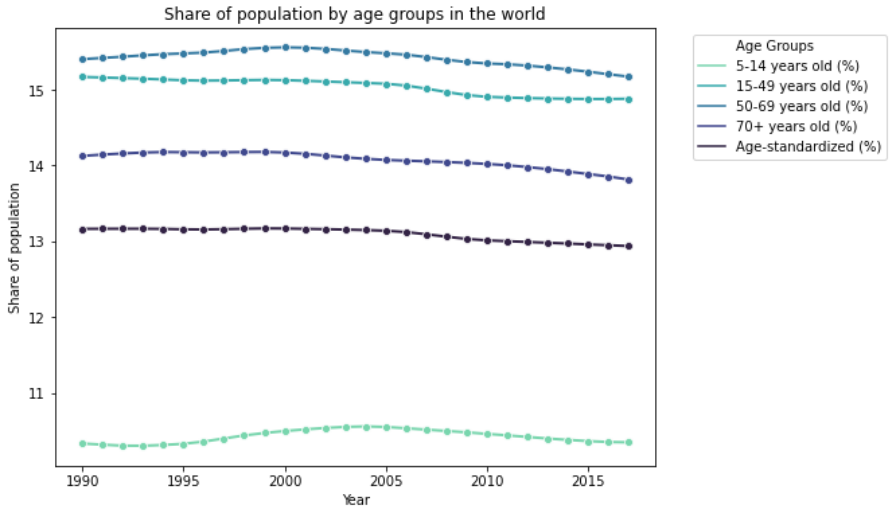


*Fig. 5 The trend by years for 5-14 years old, 15-49 years old, 50-69 years old, 70+ years old, and age-standardized features in the whole world level.*

- From fig.5, we can observe that from 2000 to 2005, there is an increased jump for 5-14 years old, 15-49 years old and 50-69 years old age groups. Let's take a closer look at how it changes over the years. I checked how each age group changed in decades. Figure 6 clearly showed that the mental health and substance use disorders rate is decreasing for most age groups in the whole world level through 2000 to now.
    - 1990~1999: only 15-49 years old age group mental health and substance use disorders decreased, 50-69 years old age group has maximum increase.

- 2000~2009: all age groups show a decrease, 15-49 years old and 50-69 years old age groups have a maximum decrease, 5-14 years old shows a minimum decrease.
- 2010~2017: all age groups show a decrease, 50-69 years old, and 70+ years old age groups have a maximum decrease, 5-14 shows a minimum decrease.
- 1990~2017: only 15-49 years old age group increased, 70+ years old age group has maximum decrease.

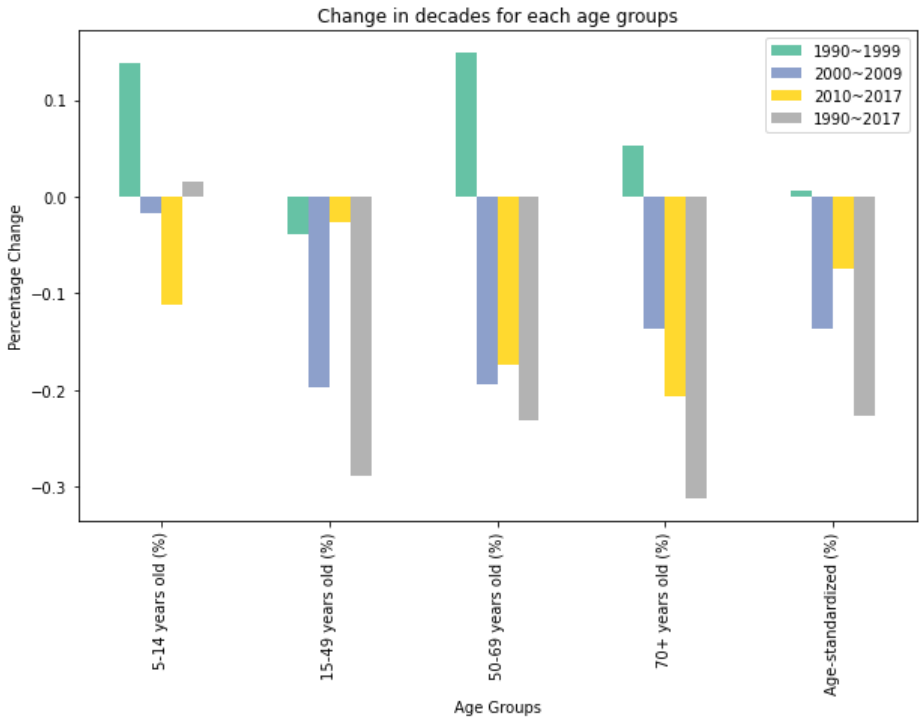


*Fig.6 Mental health and substance use disorders rate change in decades for different age groups.*

- I used Simpson's paradox method to compare relationships between different age groups and GDP per capita in all countries. When looking at the data individually (Fig. 7), there is a negative correlation between 70+ years old and GDP per capita, a positive correlation between 5-14 years old and 15-49 years old and GDP per capita, no correlation between 50-69 years old and GDP per capita, but when aggregating the data (Fig. 8), the correlation is only slightly positive! This is due to the presence of another cause, age, on the chance of developing a disease. To determine the effect of GDP per capita on the probability of disease, we need to control for the age of patients.

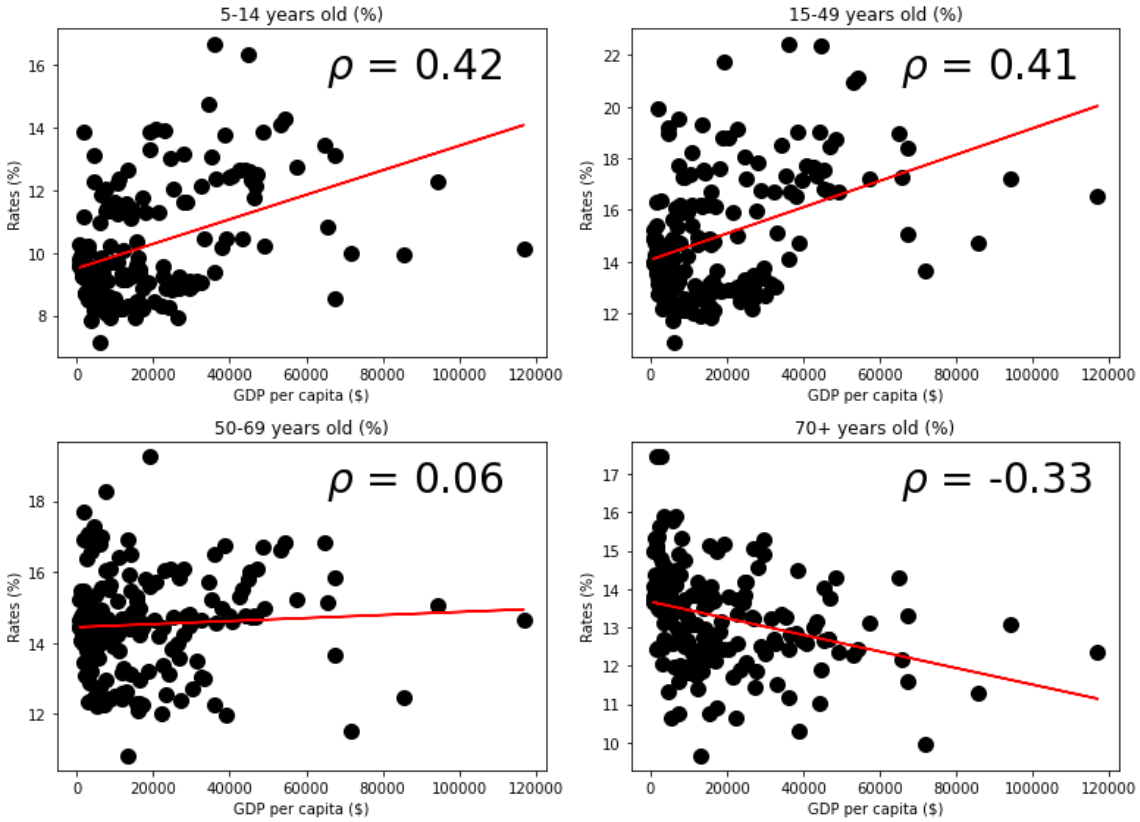


*Fig. 7 The relationship between mental health and substance use disorders rate for each group and GDP per capita.*

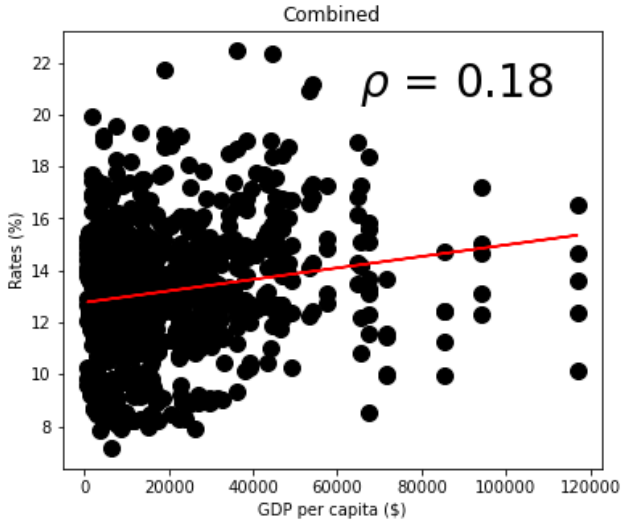


*Fig. 8 The relationship between mental health and substance use disorders rate for all groups and GDP per capita.*

### 3. Share of population with mental or substance disorders, male vs. female

Then I checked the share of mental health and substance use disorders by sex data. I compared the average rate for males and females to check if there's a significant difference between them. The average rate for the male is 12.610292 and the average rate for females is 13.358836. It seems there is a significant difference between them with (13.358836 - 12.610292) 0.748544 observed. We can also see differences from boxplots (Fig. 9).

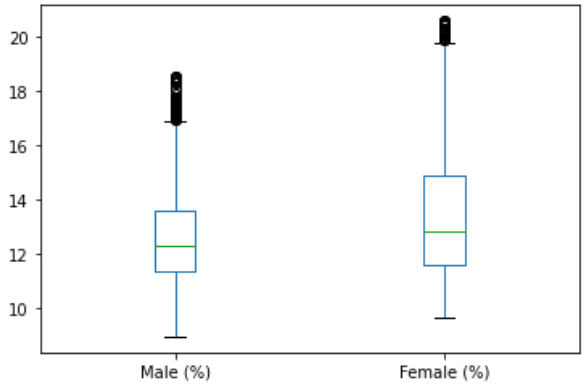


*Fig. 9 The boxplot for male and female*

Then I applied statistical tests to check if it's a significant difference between male and female. 4 steps were taken to apply the statistics test:

- Hypothesis formulation
- Getting the distribution of the data
- Statistic test (t-test and permutation test)
- Conclusion

Here, I will explain how each step approached and the main fundings.

1) **Hypothesis formulation**

   $H_{null}$: the observed difference in the mean percentage of the male and female is due to chance (and thus not due to the gender).

   $H_{alternative}$: the observed difference in the mean percentage of male and female is not due to chance (and is actually due to gender)

   I am also going to pick a significance level of 0.05.

2) **Getting the distribution of the data**

To check whether male and female data are normally distributed, I visualized the distribution of the data (shown in Fig. 10) and calculated the p-value with stats.normaltest() method. The distribution of the data visually with a histogram is not symmetric, and the p-values for both tests are less than 0.05 (reject the null hypothesis of the normaltest() which is the data are normally distributed), so the conclusion is that the data are not normally distributed. I chose t-test and permutation for this non-normal distributed data for the statistic test.

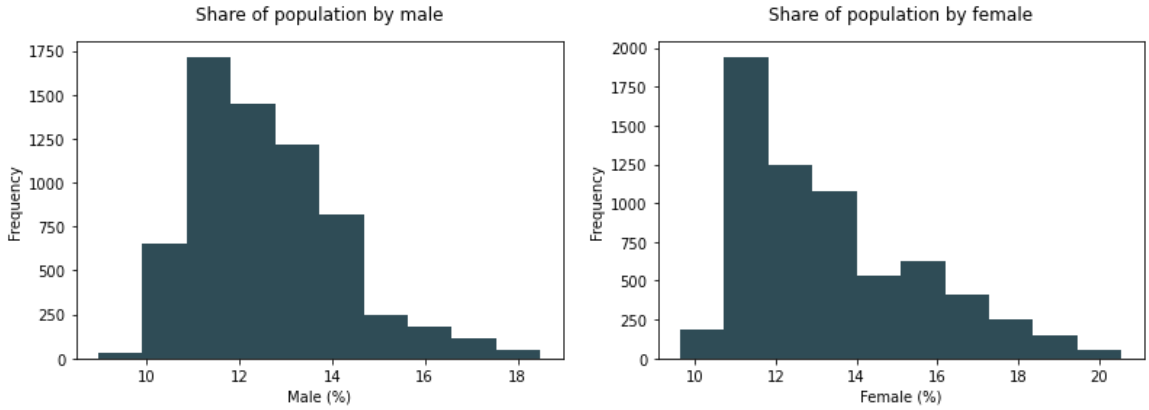


*Fig. 10 The distribution for male (Left) and female (Right).*

## 3) Statistic test

3.1) t-test

I used a t-test method because of the Central Limit Theorem (CLT) and the size of our sample is large (6468 variables). Typically, I used the t-test function from scipy.stats to calculate the value of the test statistic and its probability (the p-value). The result of the statistic is -22.1844, and the result of the p-value is 7.611770095217847e-107.

3.2) Permutation test

Since the data isn't normally distributed, we can also use a non-parametric test here -- permutation test. I used the permutation_test function from the mlxtend library. The p-value from the permutation test is 0.0.

## 4) Conclusion

The p-value from the t-test and permutation test is less than the significance level of 0.05. We can reject our null hypothesis. So our observed data is statistically significant that the share of the population with mental or substance use disorders between male and female are significantly different. And from the average rates, we can see that the female tends to have higher rates than male.

# Model selection

I used the share of the population with mental or substance disorders (male vs. female) data to identify the country outliers using unsupervised clustering machine learning. The data is an unlabeled anomaly, so we can only model this data with an unsupervised method. I filtered the data from 2017, then applied the following different models and evaluated the models with visualization. All functions used were from scikit-learn library. The One-Class SVM was chosen because it best identified the outliers.

- Isolation Forest (Fig.11)
- One-Class SVM (Fig.12)
- DBSCAN (Fig.13)
- Local Outlier Factor (LOF)(Fig.14)
- Elliptic Envelope (Fig.15)

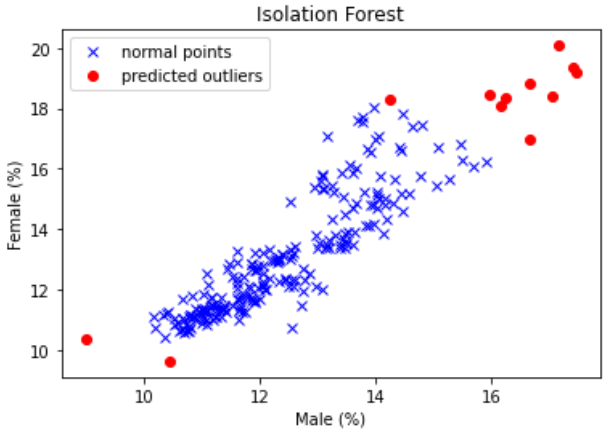


*Fig.11 Isolation Forest*

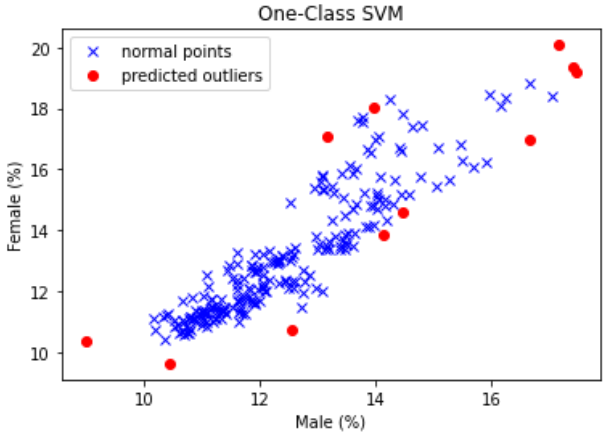

*Fig.12 One-Class SVM*

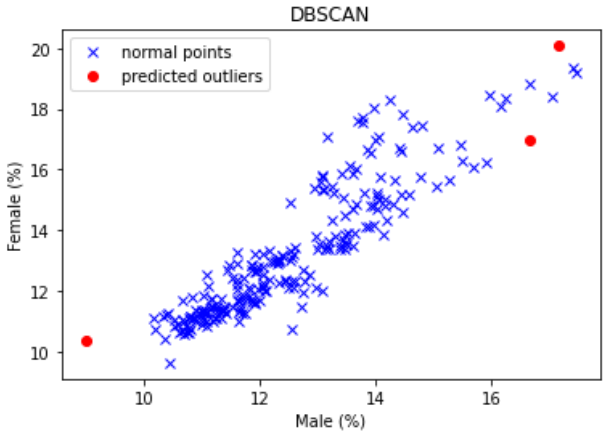


*Fig.13 DBSCAN*

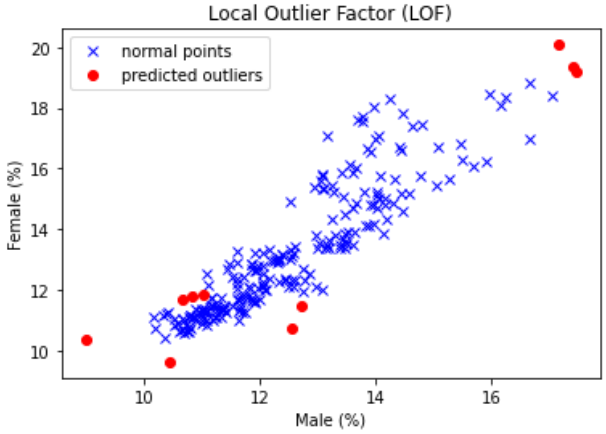


*Fig.14 Local Outlier Factor (LOF)*

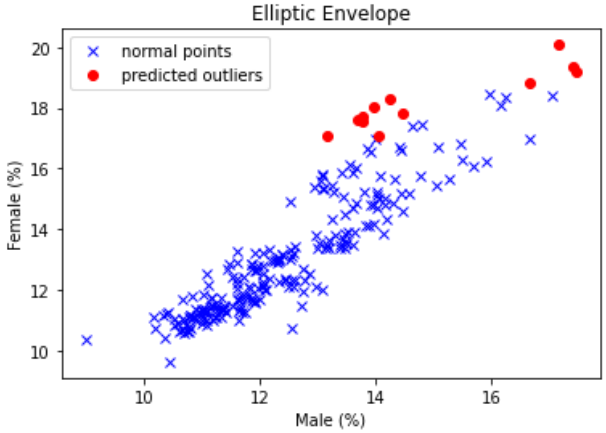


*Fig. 15 Elliptic Envelope*

## Conclusion

In summary, mental health and substance use disorders have a high share of the population in the whole world. Alcohol use disorders are a risk factor for suicide rates. GDP per capita has a positive linear relationship with mental health and substance use disorders as a share of total disease burden and eating disorders. So the rich countries have more people rate suffering from mental health or substance use disorders than poor countries. Mental health and substance use disorders have different rates in different age groups, the 50-69 years old age group shows the highest percentage and the 5-14 years old age group shows the lowest percentage. Mental and substance disorders between males and females are significantly different. And from the mean percentage, we can see that females tend to have higher rates than males.

The One-Class SVM is the best method to detect country outliers in this case.

## Future work

In the future, I would like to identify the country outliers for each year and check if there're commons for each year. I also would like to check what makes those countries be outliers.

We saw strong positive correlations between mental health and substance use disorders as a share of total disease burden and GDP per capita (r=0.73), eating disorders, and GDP per capita (r=0.75). It'll be interesting to apply the linear regression to predict the share of disease burden and eating disorders based on GDP per capita.

There is an increased jump for 5-14 years old, 15-49 years old and 50-69 years old age groups around 2000 to 2005. I would like to gather more data and read more about mental health and substance use disorders and events that happened at that time and combine with data science methods to look at the insights about that.