Milestone 1 Project

For this project, I have decided to make an interactive dashboard. The objective of this project is to look at trends of mental health datasets from Kaggle. The data sets are “Mental Health” (1)
https://www.kaggle.com/datasets/imtkaggleteam/mental-health/data?select=5-+anxiety-disorders-treatment-gap.csv
“Mental Health Dataset” (2)
https://www.kaggle.com/datasets/bhavikjikadara/mental-health-dataset
“Student Mental Health” (3)
https://www.kaggle.com/datasets/shariful07/student-mental-health/data

For the tech stack, I intend to use matplotlib, seaborn, numpy, pandas, os, and scikit learn, statsmodels. Perhaps others, but unknown yet.

For each dataset, I am following proper usage.

Data Collection

For the first dataset, it was split into 7 separate csv’s.
CSV 1 data_1_1-mental-illnesses-prevalence: Dimension n x 8

Features:
Entity
Code
Year
Schizophrenia disorders (share of population) - Sex: Both - Age: Age-standardized
Depressive disorders (share of population) - Sex: Both - Age: Age-standardized
Anxiety disorders (share of population) - Sex: Both - Age: Age-standardized
Bipolar disorders (share of population) - Sex: Both - Age: Age-standardized
Eating disorders (share of population) - Sex: Both - Age: Age-standardized

Where each disorder is reported as a percent of the population
Entity is a location in the world
Code is the code associated with that location
Year is the year the data is from

CSV 2 data_1_2-burden-disease-from-each-mental-illness: Dimension n x 8
Features:
Entity

Code
Year
DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Depressive disorders
DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Schizophrenia
DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Bipolar disorder
DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Eating disorders
DALYs (rate) - Sex: Both - Age: Age-standardized - Cause: Anxiety disorders

Disability-adjusted life years are a measure of overall disease burden, expressed as the number of years lost due to ill-health, disability, or early death.
This means that the burden of the disease on the population is equivalent to <DALYs rate> healthy years of life lost due to disability or premature death.
https://www.pharmdinfo.com/pharmacoepidemiology-and-pharmacoeconomics-f65/understanding-daly-calculation-a-step-by-step-example-with-model-data-t4026.html

Entity is the location of where the data was sourced from.
Code is the relevant code for the entity.
Year is the year the data was taken.

CSV 3
data_1_3-adult-population-covered-in-primary-data-on-the-prevalence-of-major-depression: Dimension n x 4
Features:
Entity
Code
Year
Major Depression

Entity is the location the data is sourced from
Code is the relevant area code
Year is the year the data was taken
Major Depression is a score 100 - prevalent, 0 not prevalent at all

CSV 4
data_1_4-adult-population-covered-in-primary-data-on-the-prevalence-of-mental-illnesses: Dimension n x 9
Features:

Entity
Code

Year
Major depression
Bipolar disorder
Eating disorders
Dysthymia
Schizophrenia
Anxiety disorders

Entity is the location
Code is the relevant entity location code
Year is the year the data was taken
Each disorder is assigned a score, 0 for not prevalent, 100 for prevalent

CSV 5 data_1_5-anxiety-disorders-treatment-gap: Dimension n x 6
Features:
Entity
Code
Year
Potentially adequate treatment, conditional
Other treatments, conditional
Untreated, conditional

Entity is the location the data is taken from
Code is the relevant location code
Year is the year the data was taken
Potentially adequate treatment, conditional is the percentage of people who could get adequately treated but are not
Other treatments, conditional, is the percentage of people who could find some kind of treatment, but are not
Untreated, conditional is the percentage of people living with a mental illness that is not treated

For this CSV, the years are all different, but offer important insights to those who get treated across countries. I think this is still usable.

CSV 6 data_1_6-depressive-symptoms-across-us-population: Dimension n x 7
Features:
Entity

Code
Year
Nearly every day
More than half the days
Several days
Not at all

Entity is the type of symptom
Code is the relevant code for that symptom
Year is the year the data was taken
Nearly every day, More than half the days, Several days , Not at all are percentages of people that answered that.

CSV 7
data_1_7-number-of-countries-with-primary-data-on-prevalence-of-mental-illnesses-in-the-global-burden-of-disease-study: Dimension n x 4
Features:
Entity
Code
Year
Number of countries with primary data on prevalence of mental disorders

Entity is the type of disorder
Code is the relevant code for that disorder
Year is the year the data was taken
Number of countries with primary data on prevalence of mental disorders is exactly the name

For dataset 2, there was 1 csv. The self_employed has some nulls.

CSV data_2_Mental Health Dataset: Dimension n x 17
Features:
Timestamp
Gender
Country
Occupation
self_employed

Family_history
Treatment
Days_Indoors
Growing_Stress
Changes_Habits
Mental_Health_History
Mood_Swings
Coping_Struggles
Work_Interest
Social_Weakness
Mental_health_interview
care_options

Timestamp when the data was taken
Gender, if the person is male or female
Country, the country from which the data was taken
Occupation, What occupation the person falls under
self_employed , if the person is self employed or not
Family_history, if the person has any family history of mental illness
Treatment, if the person is seeking treatment for their mental health
Days_Indoors, the number of days spent indoors
Growing_Stress, if whether or not stress increased with their mental illness
Changes_Habits, if habits changed as a result of their mental illness
Mental_Health_History, if there is a history working with their mental illness
Mood_Swings, Mood swings associated with their mental illness
Coping_Struggles, if person is struggling to cope
Work_Interest, if person is interested in their work
Social_Weakness, if person experiences social weakness
Mental_health_interview, if they conducted a mental health interview
Care_options, yes if seeking appropriate care, no if remaining untreated, not sure for not sure

For dataset 3, there was 1 csv.

CSV data_3_Student Mental health: Dimension n x 11
Features:
Timestamp
Choose your gender
Age

What is your course?
Your current year of Study
What is your CGPA?
Marital status
Do you have Depression?
Do you have Anxiety?
Do you have Panic attack?
Did you seek any specialist for a treatment?

Timestamp, the timestamp that the data was taken
Choose your gender, the gender of the participant
Age, age of participant
What is your course?, Area of study for student
Your current year of Study, the undergrad year of the student
What is your CGPA?, the GPA of the student
Marital status, whether or not the participant is married
Do you have Depression?, whether or not, self reported, the student has depression
Do you have Anxiety?, whether or not, self reported, the student has anxiety
Do you have Panic attack?, whether or not, self reported, the student has panic attacks
Did you seek any specialist for a treatment? whether or not, self reported, the student has seeked treatment

Data Preprocessing

https://www.statology.org/top-5-statistical-techniques-detect-handle-outliers-data/
Z-score was calculated via this article.
analysis.py contains the code run to calculate outliers and generate figures.
multi.py contains the code to generate multicollinearity corr matrices.

Dataset 1:
CSV 1:

1) There was no missing data or nulls.
2) There were outliers. A list of outliers with their relevant Z score were made. There were no outliers for eating_disorders. The lists of outliers for each column are as follows: data_1_1_anxiety_outliers, data_1_1_bipolar_outliers, data_1_1_depressive_outliers, data_1_1_eating_outliers, data_1_1_schizophrenia_outliers
3) There was no need to scale features.

CSV 2:

1) There was no missing data or nulls.
2) There were outliers. Several lists of outliers was made with data index and Z scores. The lists for each columns are as follows: data_1_2_anxiety_outliers, data_1_2_bipolar_outliers, data_1_2_depressive_outliers, data_1_2_eating_outliers, data_1_2_schizophrenia_outliers
3) There was no need to scale features.

CSV 3:

1) There was no missing data or nulls.
2) There were no outliers. This is shown in the blank csv generated by checking the z score of every point. The output csv is data_1_3_depression_score_outliers.
3) There was no need to scale features.

CSV 4:

1) There was no missing data or nulls.
2) There were no outliers. This is shown by the blank csv's. The output csv's are as follows: data_1_4_anxiety_score_outliers, data_1_4_bipolar_score_outliers, data_1_4_depression_score_outliers, data_1_4_dysthymia_score_outliers, data_1_4_eating_score_outliers, data_1_4_schizophrenia_score_outliers.
3) There was no need to scale features.

CSV 5:

1) There was no missing data or nulls.
2) There were no outliers. This is shown by the blank output csv's. The output csv;s are as follows: data_1_5_other_outliers, data_1_5_potentially_adequate_outliers, data_1_5_untreated_outliers
3) There was no need to scale features.

CSV 6:

1) There was no missing data or nulls.
2) There were no outliers. Outliers wouldn't make sense here, as they are all different mental illnesses.
3) There was no need to scale features.

CSV 7:

1) There was no missing data or nulls.
2) There were no outliers. Outliers wouldn't make sense here, as they are all different countries.
3) There was no need to scale features.

Dataset 2:

1) Self_employed has nulls. They will be dropped.
2) There were no outliers
3) There is no need to scale features for analysis

Dataset 3:

1) There was no missing data or nulls.
2) There were no outliers
3) There is no need to scale features for analysis

Exploratory Data Analysis

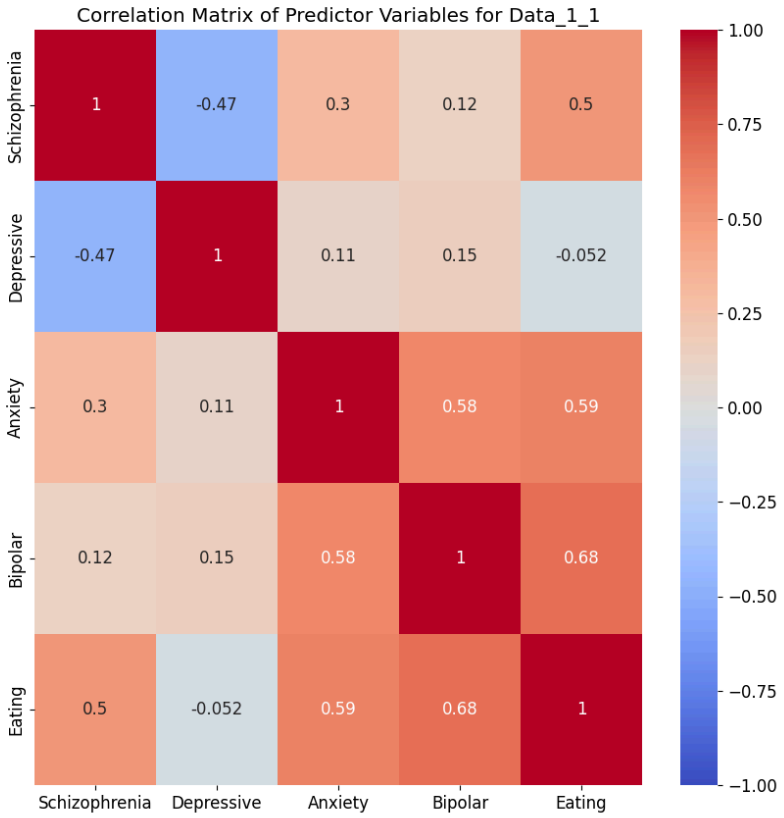
Dataset 1:

CSV 1:

1)
    a) Schizophrenia
        i) Mean: 0.27
        ii) Median: 0.27
        iii) Standard Dev: 0.04
        iv) Skew: -0.52
    b) Depressive
        i) Mean: 3.77
        ii) Median: 3.64
        iii) Standard Dev: 0.93
        iv) Skew: 0.42
    c) Anxiety
        i) Mean: 4.10
        ii) Median: 3.94
        iii) Standard Dev: 1.05
        iv) Skew: 0.46
    d) Bipolar
        i) Mean: 0.64
        ii) Median: 0.58
        iii) Standard Dev: 0.23
        iv) Skew: 0.74

e) Eating
    i) Mean: 0.20
    ii) Median: 0.14
    iii) Standard Dev: 0.14
    iv) Skew: 1.12

2) See Data_1_1 Figures
3) Given that the diseases under consideration (Schizophrenia, Depression, Anxiety, Bipolar, Eating Disorder) are known to have distinct etiologies and underlying mechanisms, we do not expect significant multicollinearity among these variables. This expectation is confirmed by the correlation matrix below, which shows low correlations between all disease pairs, and the VIF values listed below, all of which are well below the commonly used threshold of 5, indicating minimal multicollinearity.



4) VIF TABLE

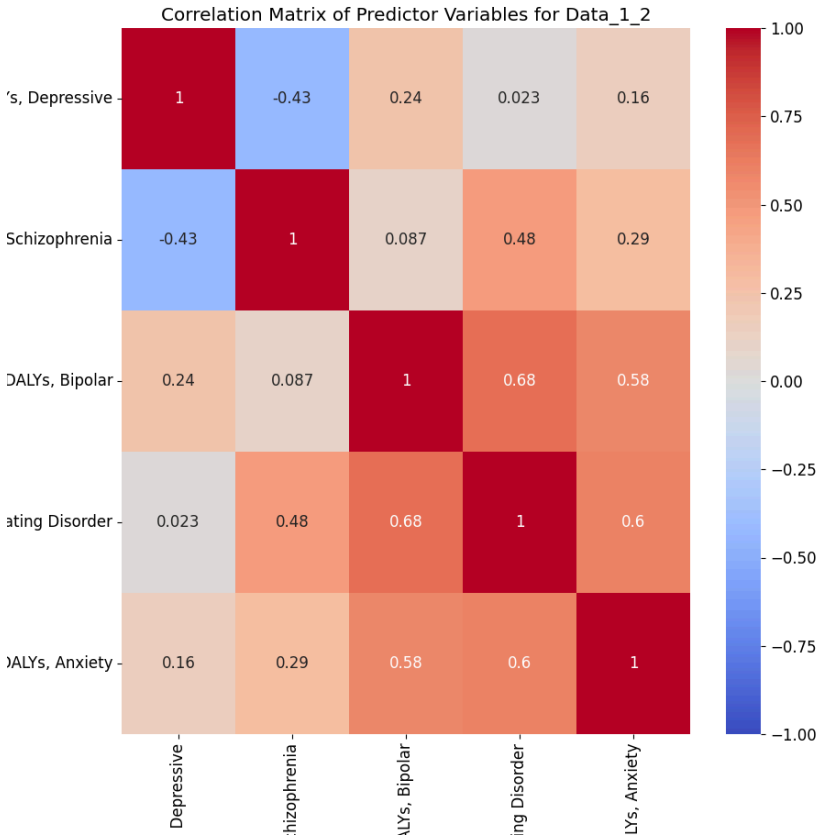
Feature	VIF Value

Const	153.37
Schizophrenia	2.03
Depressive	1.44
Anxiety	1.78
Bipolar	2.32
Eating Disorder	2.9

CSV 2:

1)
    a) Schizophrenia
        i) Mean: 171.09
        ii) Median: 175.12
        iii) Standard Dev: 26.23
        iv) Skew: -0.46
    b) Depressive
        i) Mean: 652.22
        ii) Median: 640.10
        iii) Standard Dev: 183.64
        iv) Skew: 0.19
    c) Anxiety
        i) Mean: 137.93
        ii) Median: 124.23
        iii) Standard Dev: 51.20
        iv) Skew: 0.80
    d) Bipolar
        i) Mean: 42.39
        ii) Median: 31.43
        iii) Standard Dev: 29.39
        iv) Skew: 1.11
    e) Eating
        i) Mean: 392.94
        ii) Median: 376.32
        iii) Standard Dev: 100.82
        iv) Skew: 0.49
2) See Data_1_2 Figures

3) Given that the diseases under consideration (Schizophrenia, Depression, Anxiety, Bipolar, Eating Disorder) are known to have distinct etiologies and underlying mechanisms, we do not expect significant multicollinearity among these variables. This expectation is confirmed by the correlation matrix below, which shows low correlations between all disease pairs, and the VIF values listed below, all of which are well below the commonly used threshold of 5, indicating minimal multicollinearity.



4)

5) VIF TABLE

Feature	VIF Value
Const	127.5
Depressive	1.42
Schizophrenia	1.94
Bipolar	2.41
Eating Disorder	2.9

Anxiety	1.8

CSV 3:

1)
    a) Depression Score
        i) Mean: 28.99
        ii) Median: 15.25
        iii) Standard Dev: 33.23
2) See Data_1_3 Figures
3) This is a duplicate. No need to run tests. See CSV4.

CSV3 is contained in CSV4, so this one is completely useless and duplicate data.

CSV 4:

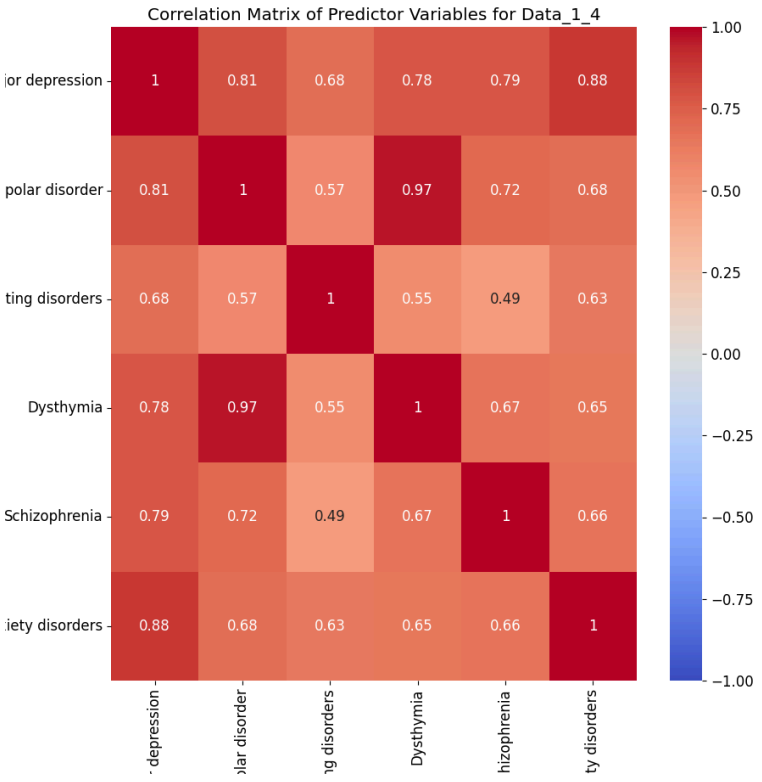
1)
    a) Depression Score
        i) Mean: 28.99
        ii) Median: 15.25
        iii) Standard Dev: 33.23
        iv) Skew: 0.19
    b) Bipolar Score
        i) Mean: 15.93
        ii) Median: 2.75
        iii) Standard Dev: 28.35
        iv) Skew: 1.39
    c) Eating Score
        i) Mean: 14.57
        ii) Median: 0.0
        iii) Standard Dev: 25.46
        iv) Skew: 1.71
    d) Dysthymia Score
        i) Mean: 17.37
        ii) Median: 0.85
        iii) Standard Dev: 29.17
        iv) Skew: 1.69
    e) Schizophrenia Score
        i) Mean: 15.13
        ii) Median: 0.55
        iii) Standard Dev: 28.35
        iv) Skew: 1.54

f) Anxiety Score
   i) Mean: 34.18
   ii) Median: 23.9
   iii) Standard Dev: 35.83
   iv) Skew: 0.86

2) See Data_1_4 Figures

3) Given that the diseases under consideration (Schizophrenia, Depression, Anxiety, Bipolar, Eating Disorder) are known to have distinct etiologies and underlying mechanisms, we do not expect significant multicollinearity among these variables. This expectation is not confirmed by the correlation matrix below, which shows high correlations between some disease pairs, and the VIF values listed below, two of which are well above the commonly used threshold of 5, indicating high multicollinearity. I guess they are collinear??

4)

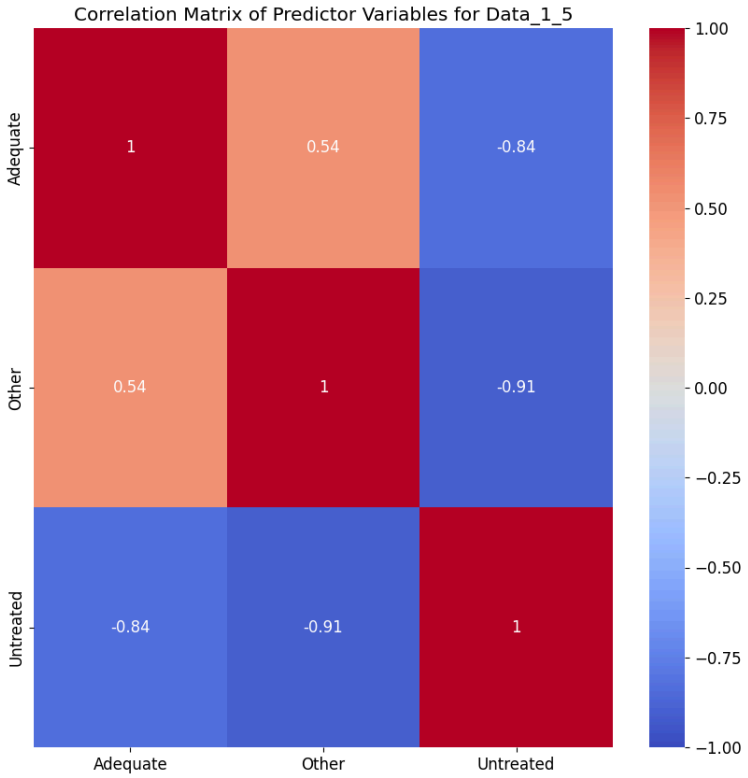


Feature	VIF Value
const	2.07
Major Depression	9.32
Bipolar Disorder	18.59

Eating Disorders	1.91
Dysthymia	15.98
Schizophrenia	2.94
Anxiety Disorders	4.71

CSV 5:

1)
    a) Potentially Adequate Treatment
        i) Mean: 8.65
        ii) Median: 9.8
        iii) Standard Dev: 5.07
        iv) Skew: -0.67
    b) Other Treatment
        i) Mean: 16.37
        ii) Median: 15.25
        iii) Standard Dev: 6.57
        iv) Skew: 0.50
    c) Untreated
        i) Mean: 74.98
        ii) Median: 76.40
        iii) Standard Dev: 10.23
        iv) Skew: -0.41
2) See Data_1_5 Figures
3) These are percent values that must add to 100. It would make sense that they are related to one another, but not in any meaningful way, so this correlation matrix will be garbage, as well as the VIF values.



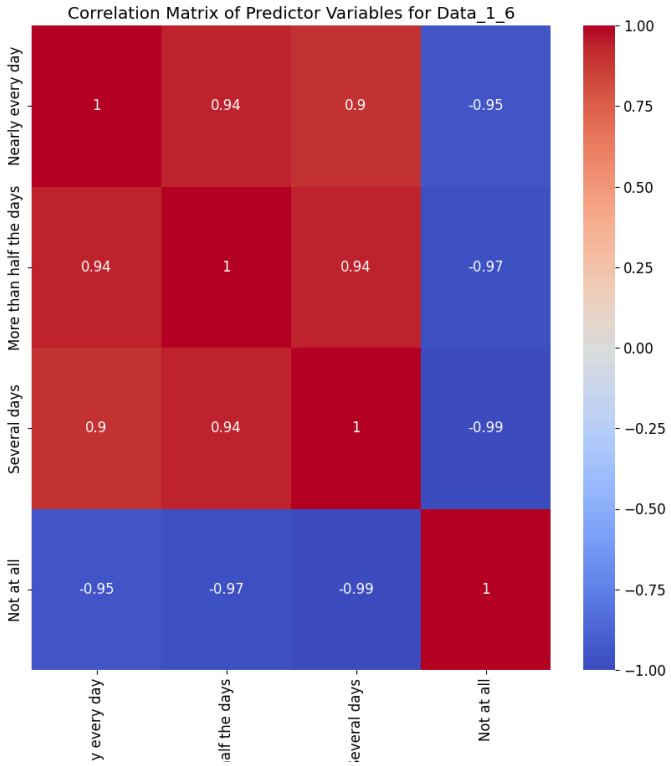
4)

Feature	VIF Value
const	0.0
Adequate	inf
Other	inf
Untreated	inf

5)

CSV 6:

1) No mean, median, or standard dev makes sense here.
2) See Data_1_6 Figures
3) These are percent values that must add to 100. It would make sense that they are related to one another, but not in any meaningful way, so this correlation matrix will be garbage, as well as the VIF values.



4)

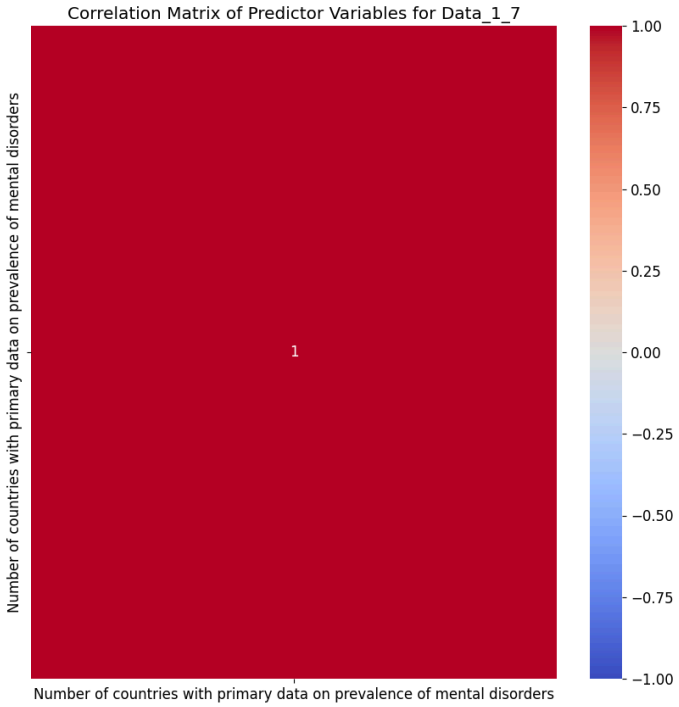
5)

Feature	VIF Value
const	6.24*10^6
Nearly Every Day	4.2*10^3
More than half the days	2.57*10^3
Several Days	4.09*10^4
Not at all	9.71*10^4

6)

CSV 7:

1) No mean, median, or standard dev makes sense here.
2) See Data_1_7 Figures
3) Multicollinearity and Skewed don't make sense here, there is only one column, and it would be 1. VIF Value divided by 0, so it was inf.

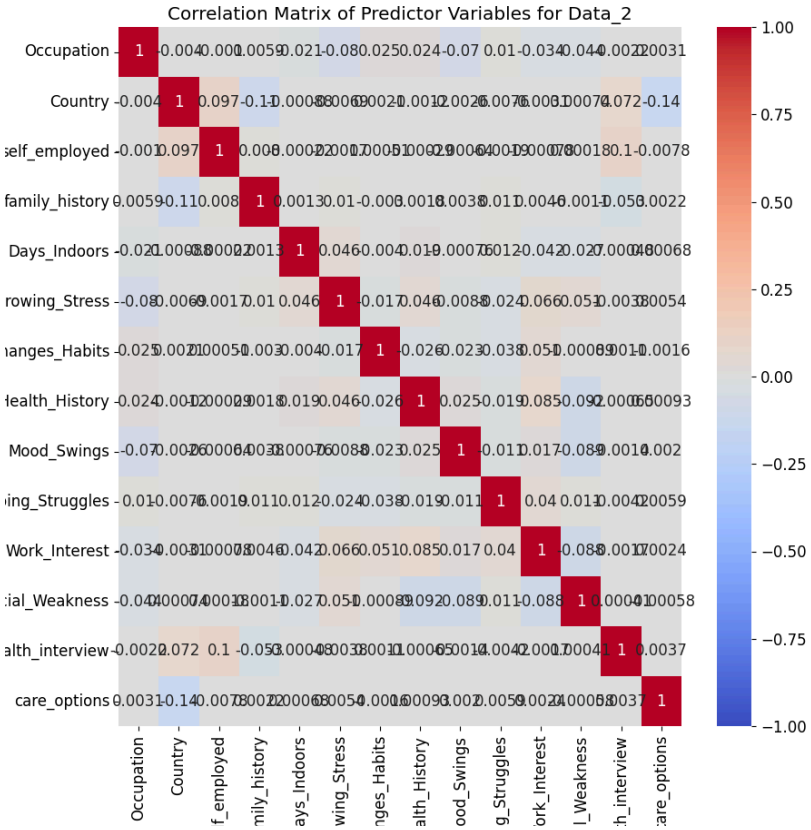


4)

Dataset 2:

CSV 1:

1) No mean, median, or standard dev makes sense here. Qualitative Data.
2) See Data_2 Figures
3) To run tests, the qualitative data was converted to numerical categories. It doesn't seem these variables are highly correlated, so multicollinearity won't be a problem, according to the correlation matrix. (see below). The VIF values are far below 5, also indicating multicollinearity won't be a problem, as seen below.



4)

Feature	VIF Value
Occupation	3.24
Country	1.28
self_employed	1.11
Days_indoors	2.55
Growing_Stress	2.34
Changes_Habits	2.59
Mental_Health_History	2.17
Mood_swings	2.00
Coping_struggles	2.12
Work_interest	2.06
Social Weakness	1.99

mental_health_interview	1.05
care_options	1.96

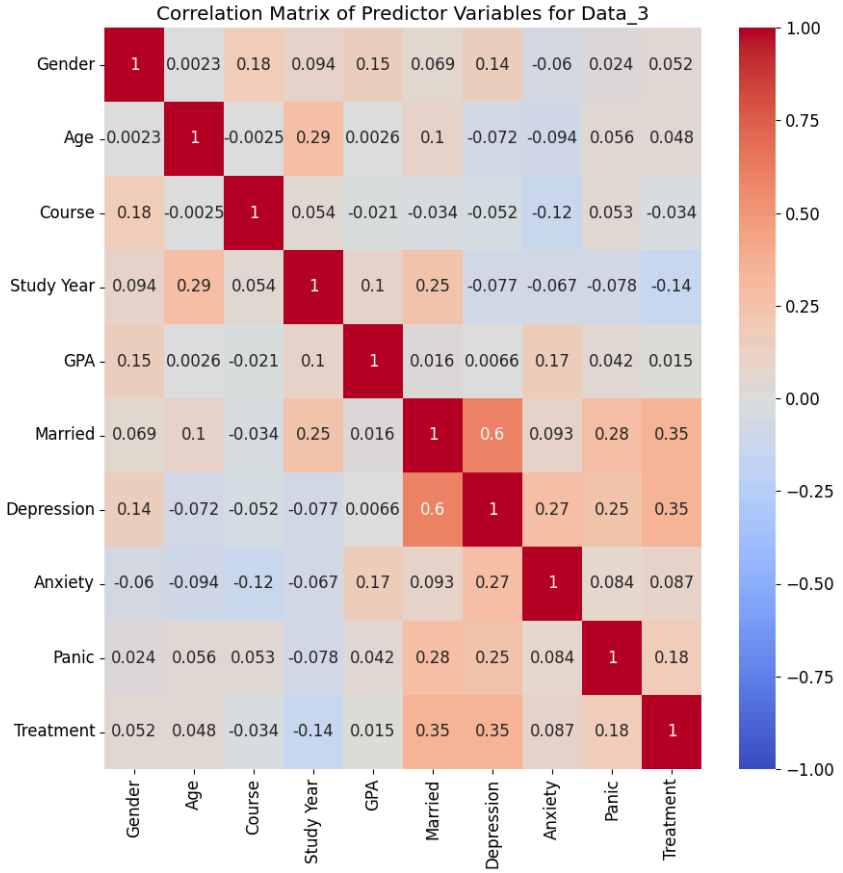
5)

Dataset 3:

CSV 1:

1) No mean, median, or standard dev makes sense here. Qualitative Data.
2) See Data_3 Figures
3) To run tests, the qualitative data was converted to numerical categories. It doesn't seem these variables overall are highly correlated, except for GPA and Gender, and Age, so multicollinearity might be a problem, according to the correlation matrix. (see below). Three VIF values are above 5, indicating multicollinearity might be a problem, as seen below.

4)

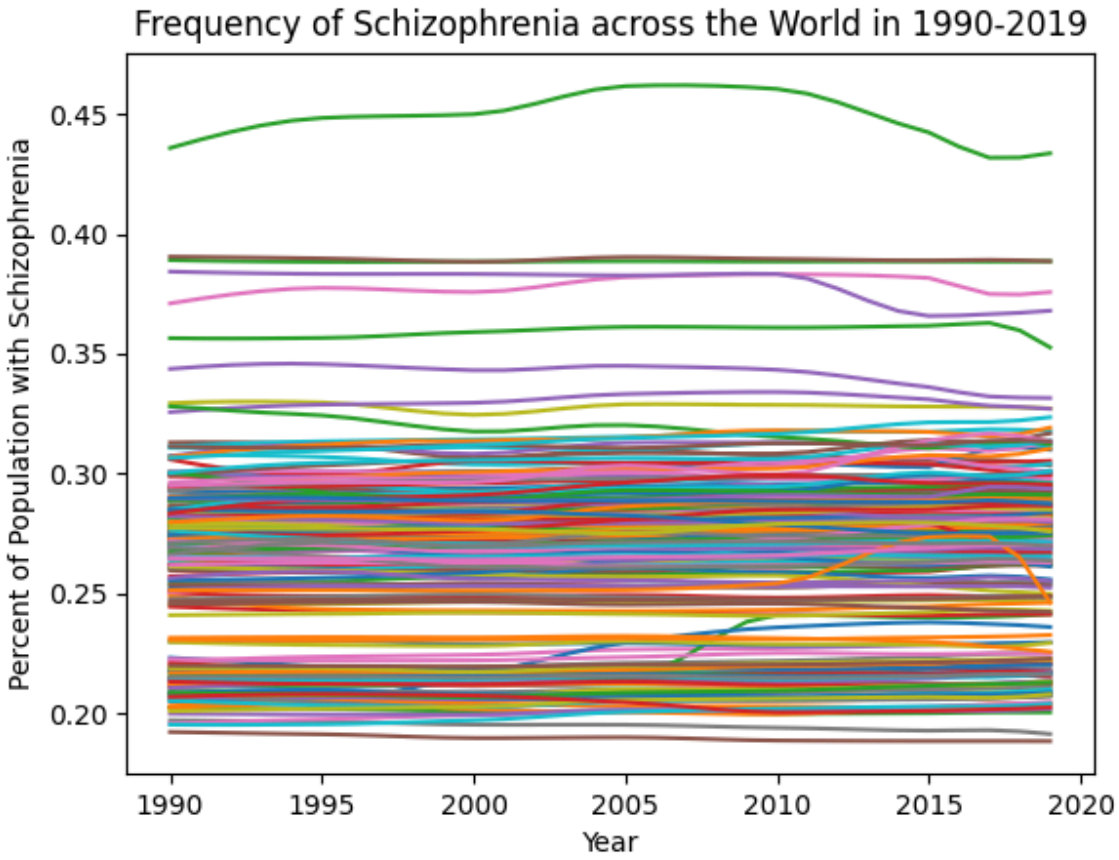


Feature	Value
Gender	5.98
Age	18.53
Course	2.19
Study Year	1.75
GPA	20.53
Married	1.85
Depression	2.49
Anxiety	1.61
Panic Attacks	1.85
Treatment	1.24

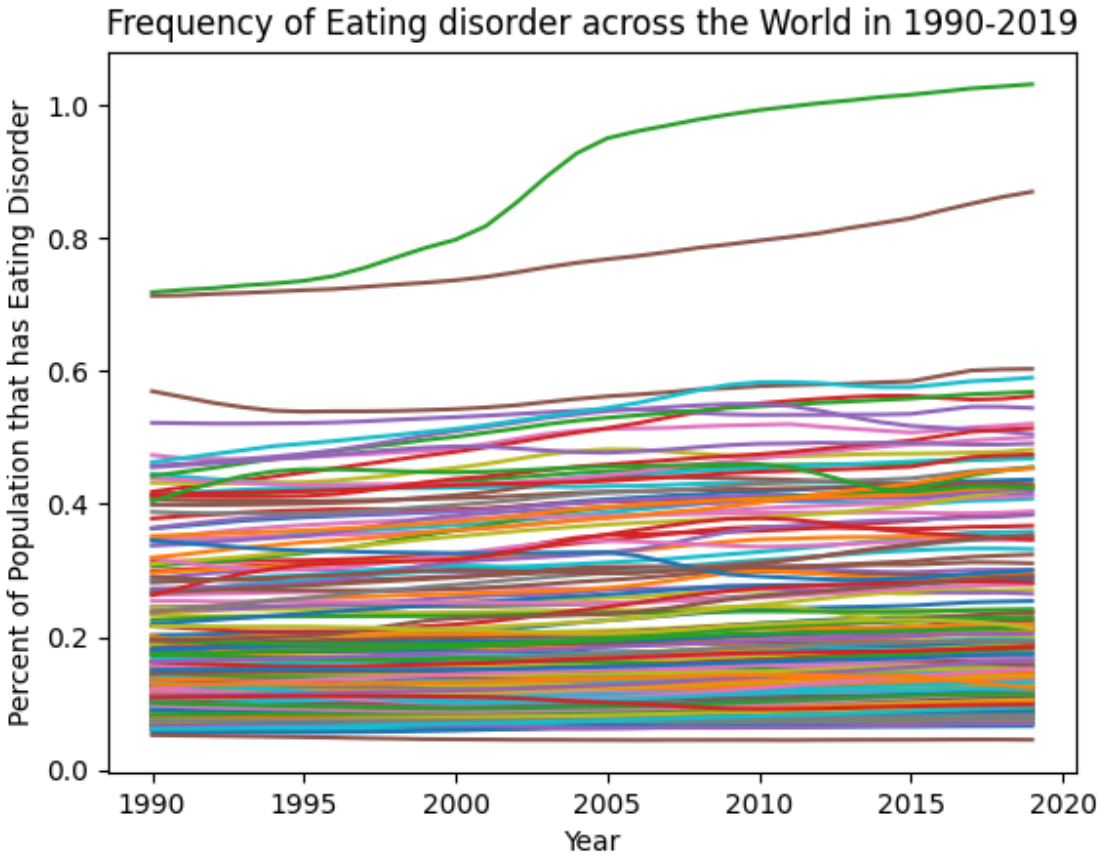
Timeline

Looking at when each milestone is due, on the time interval Feb 21 - Mar 21, week 1 I will complete ‘Feature Engineering’, week 2 I will complete ‘Feature Selection’, week3 I will complete ‘Data Modeling ’. On the time interval Mar 24 - Apr 23, week 1 I will complete ‘Evaluation and Interpretation’ and week 2 and onward I will complete ‘Tool’.

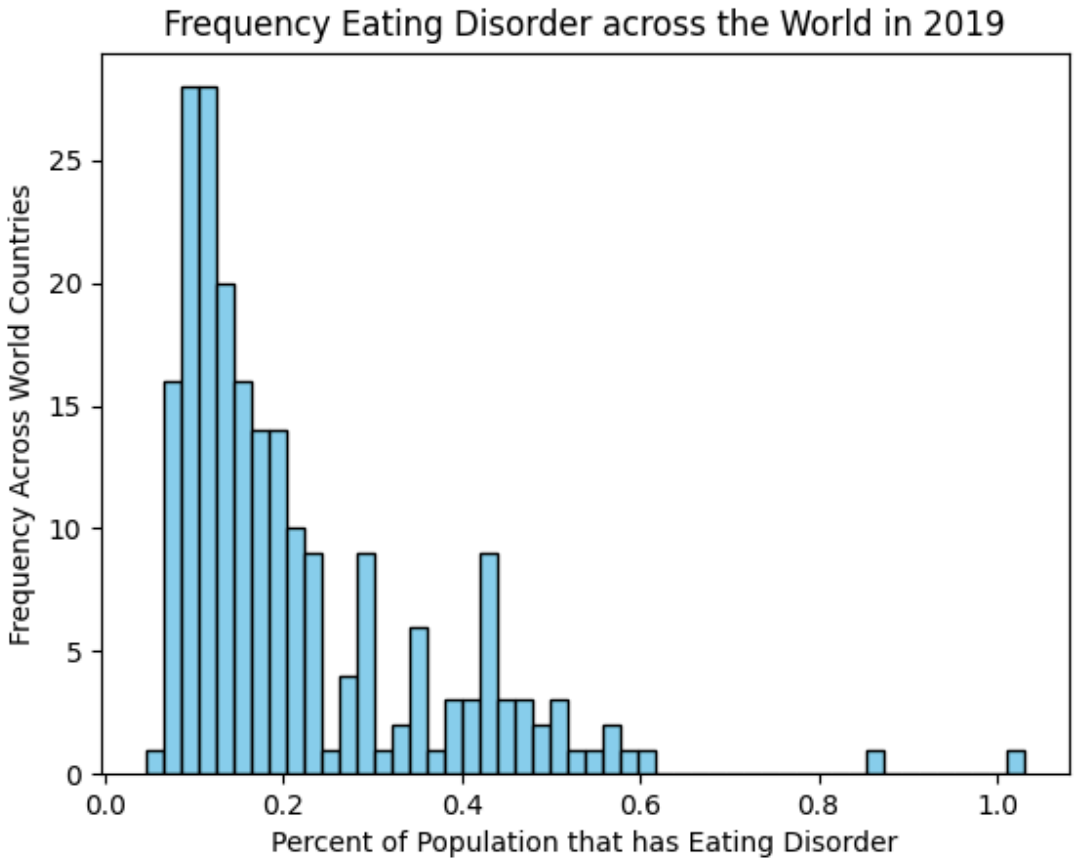
Data_1_1 Figures



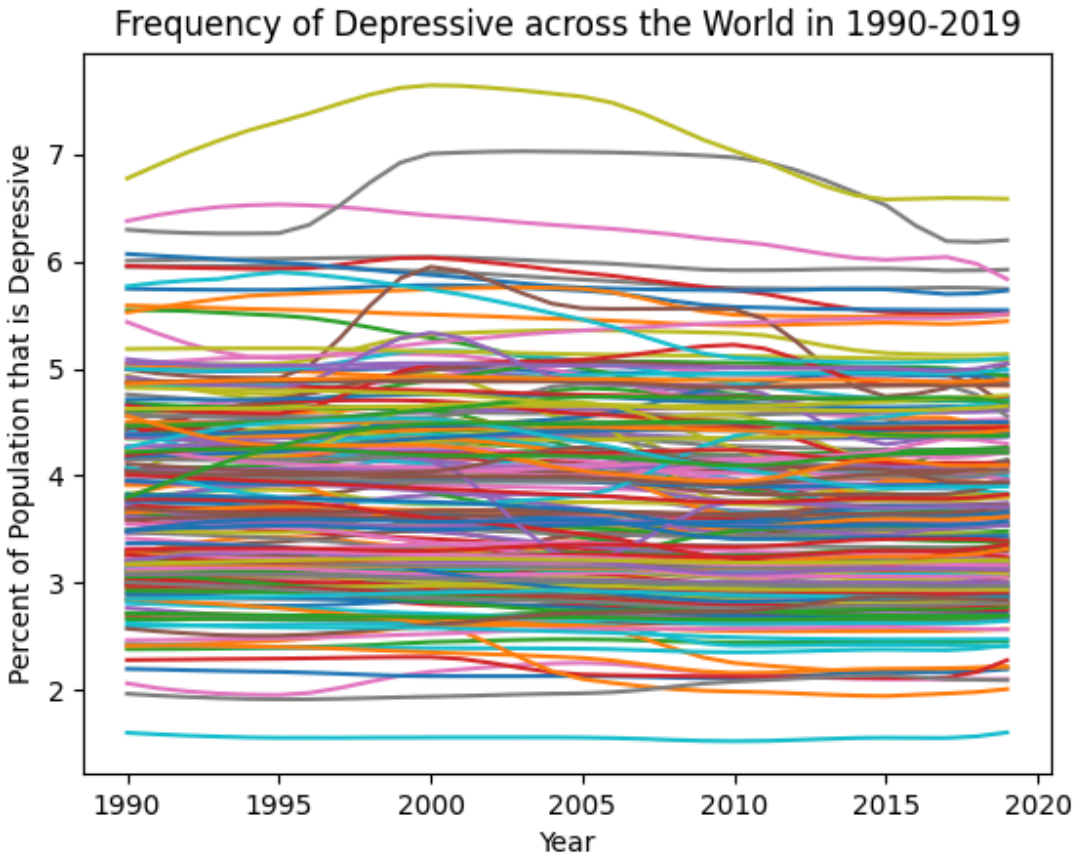
It seems that the United States in particular has a high rate of schizophrenia! (When compared to all other countries.
Overall the majority of countries seem to be in the 0.2-0.35 region.



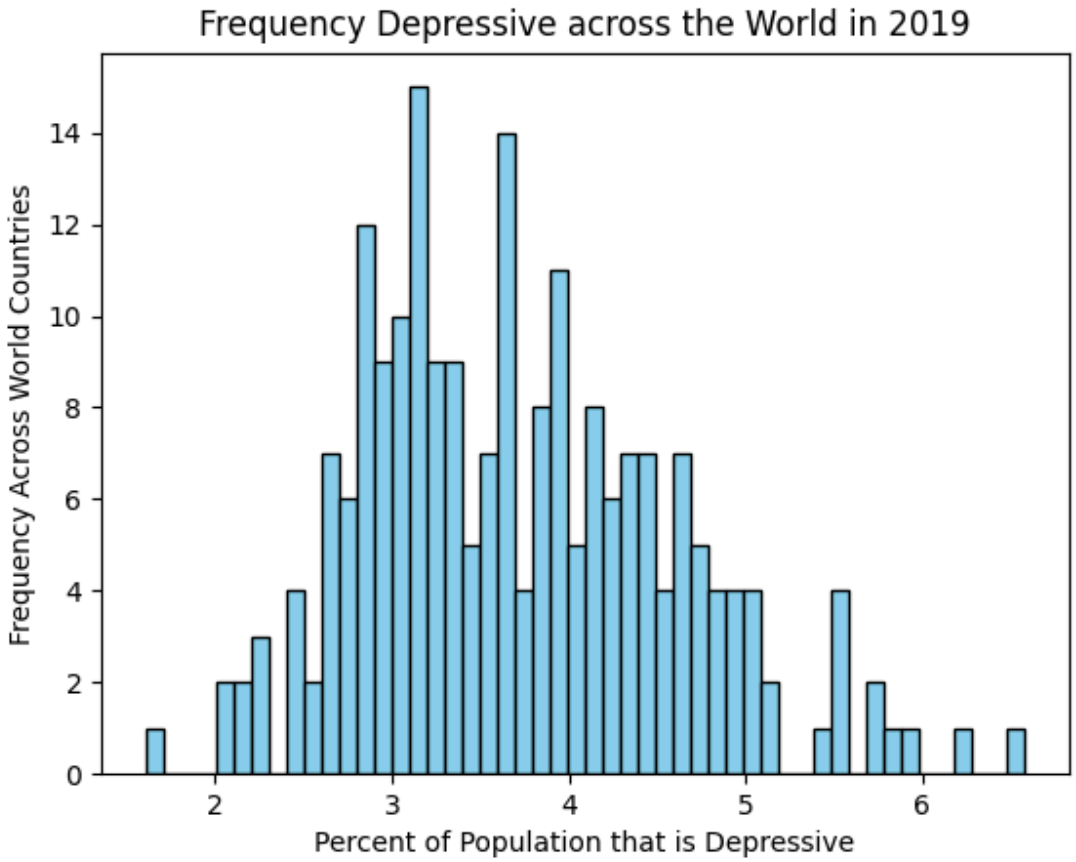
It seems Australia has the highest percent overall, with Monaco in second. Majority falls in the 0-0.6% region.



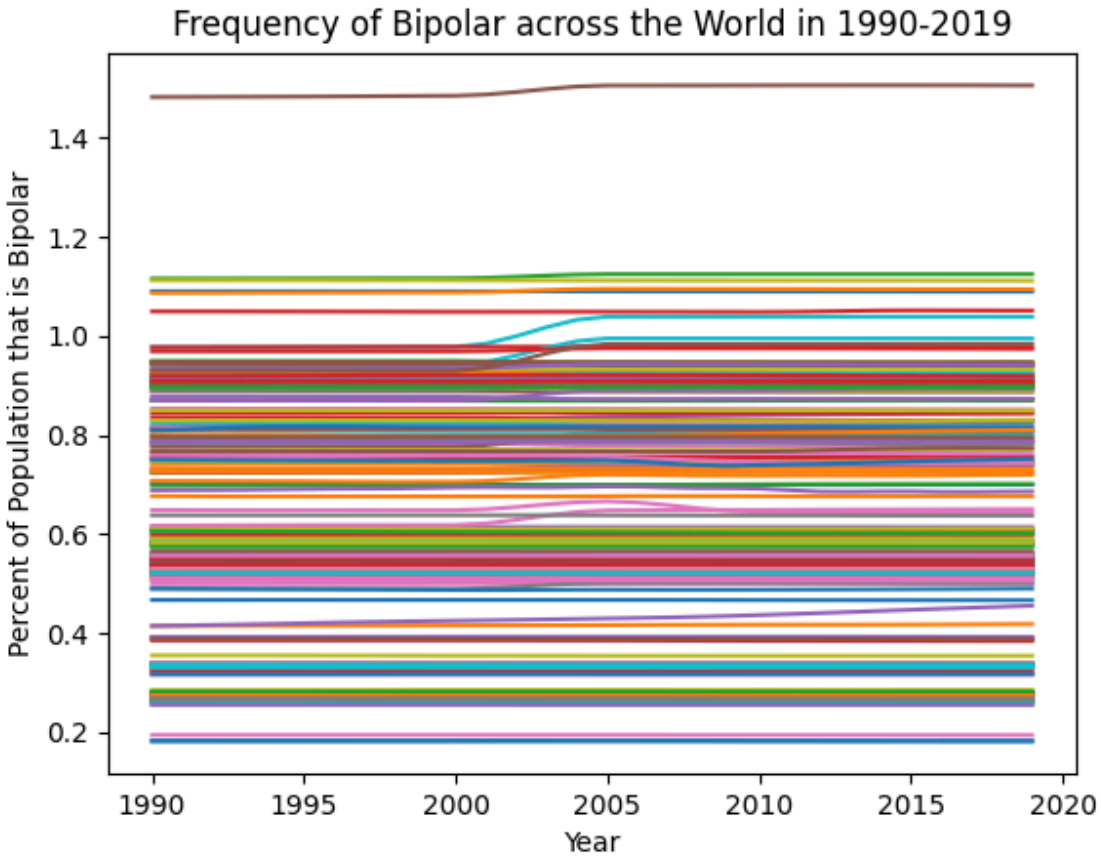
It seems most countries of the world have a lower percent with eating disorders, in the range 0 - 0.3



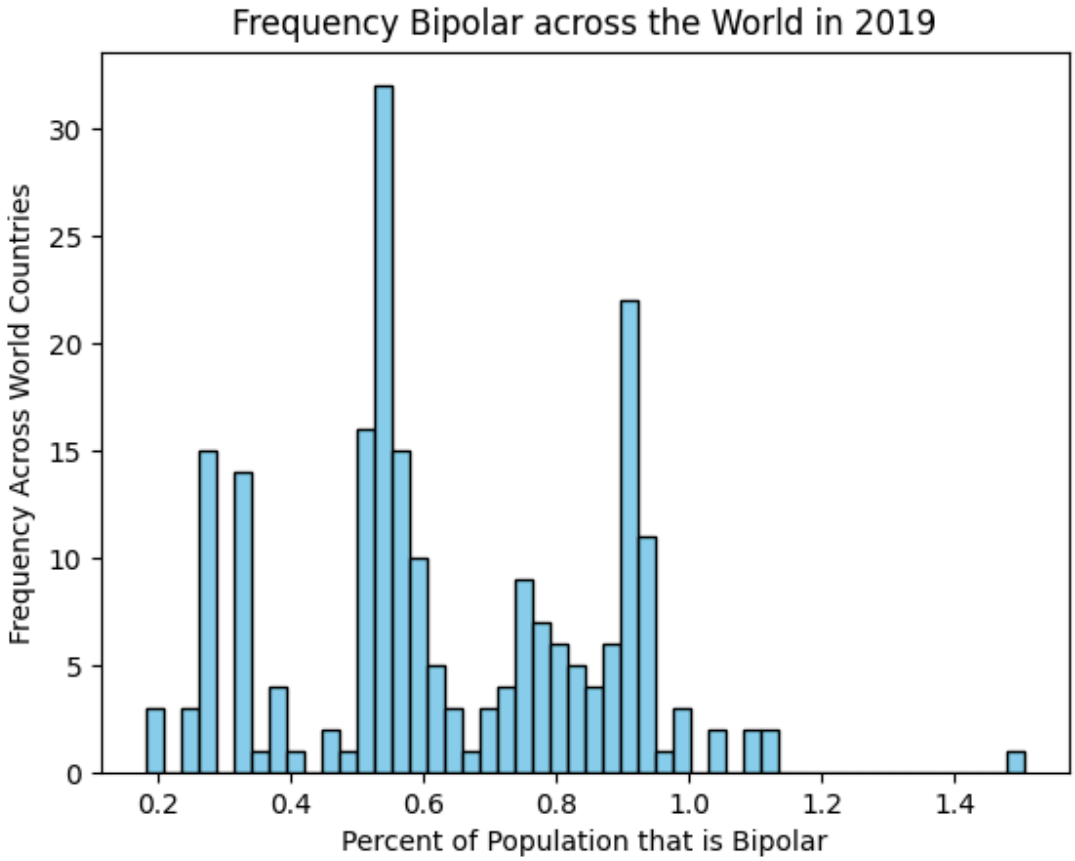
It seems like the majority of the world has 2% - 5% of the population that suffers from depression. The most depressive country is Uganda.



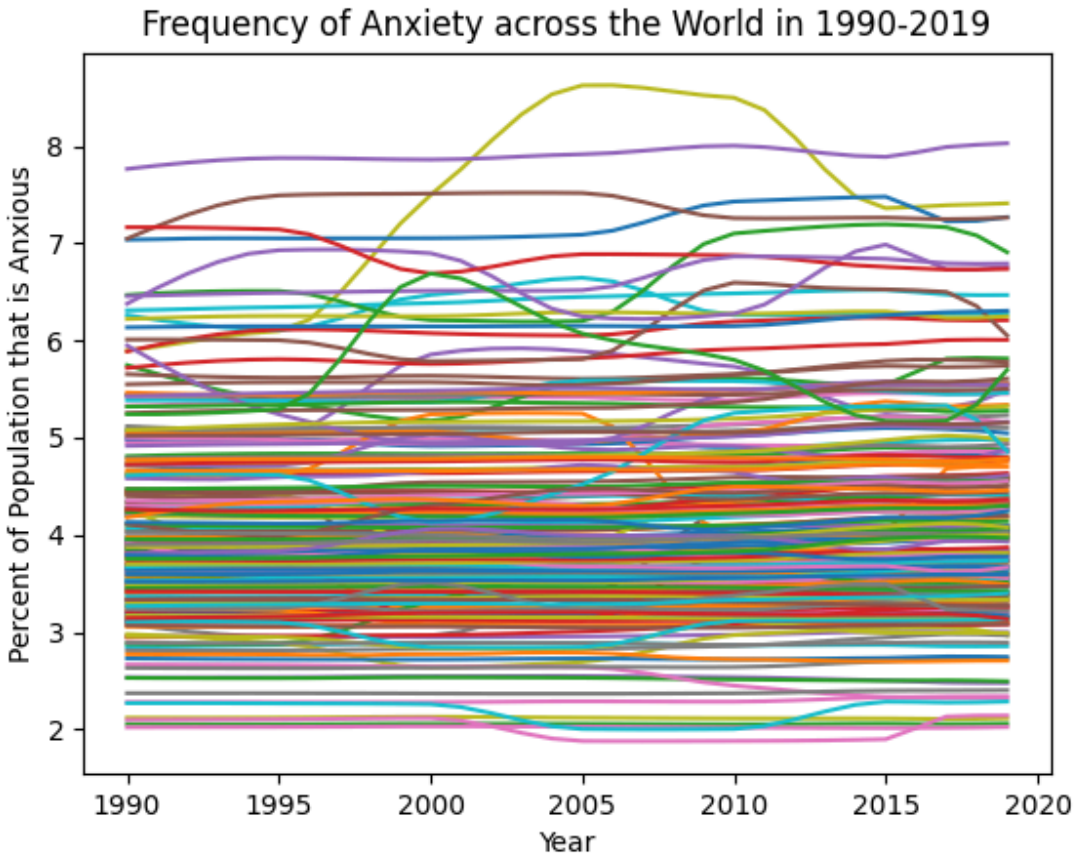
It seems like most of the world population falls in the 2.5-5% depressive range.



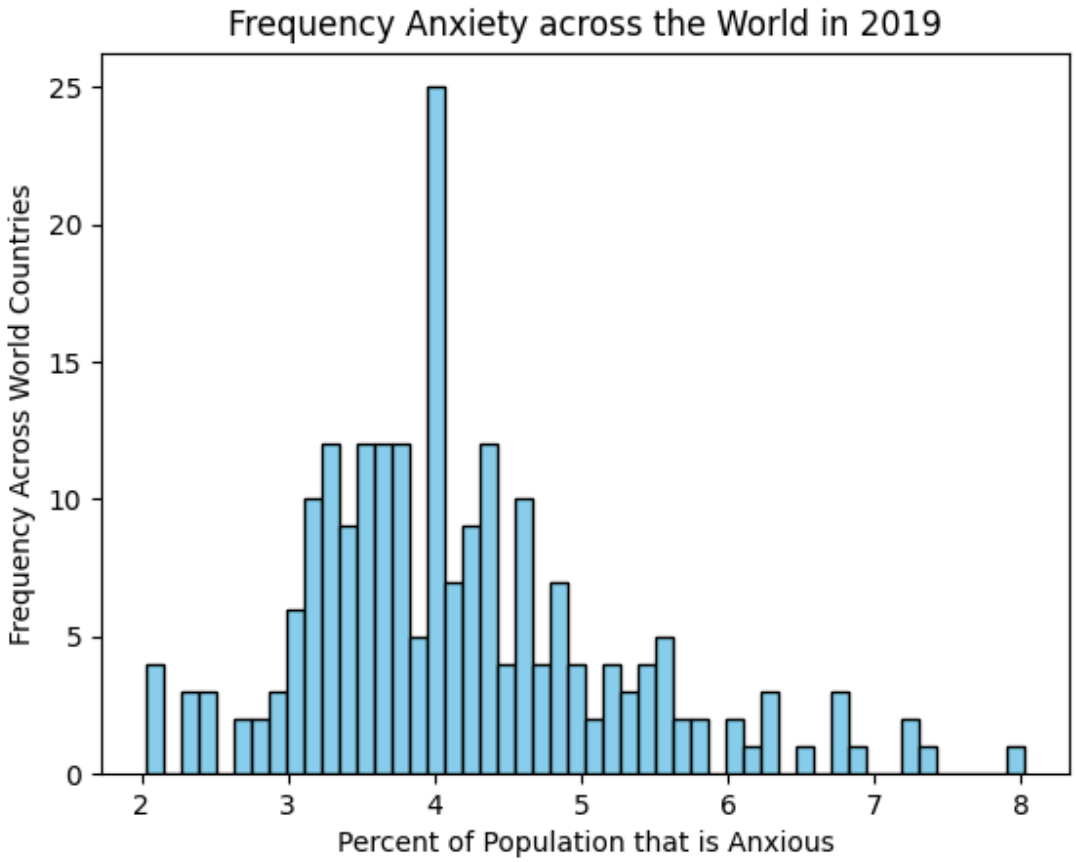
It seems like New Zealand is the most bipolar of any country! Overall the world's population falls between 0 - 1.2% bipolar.



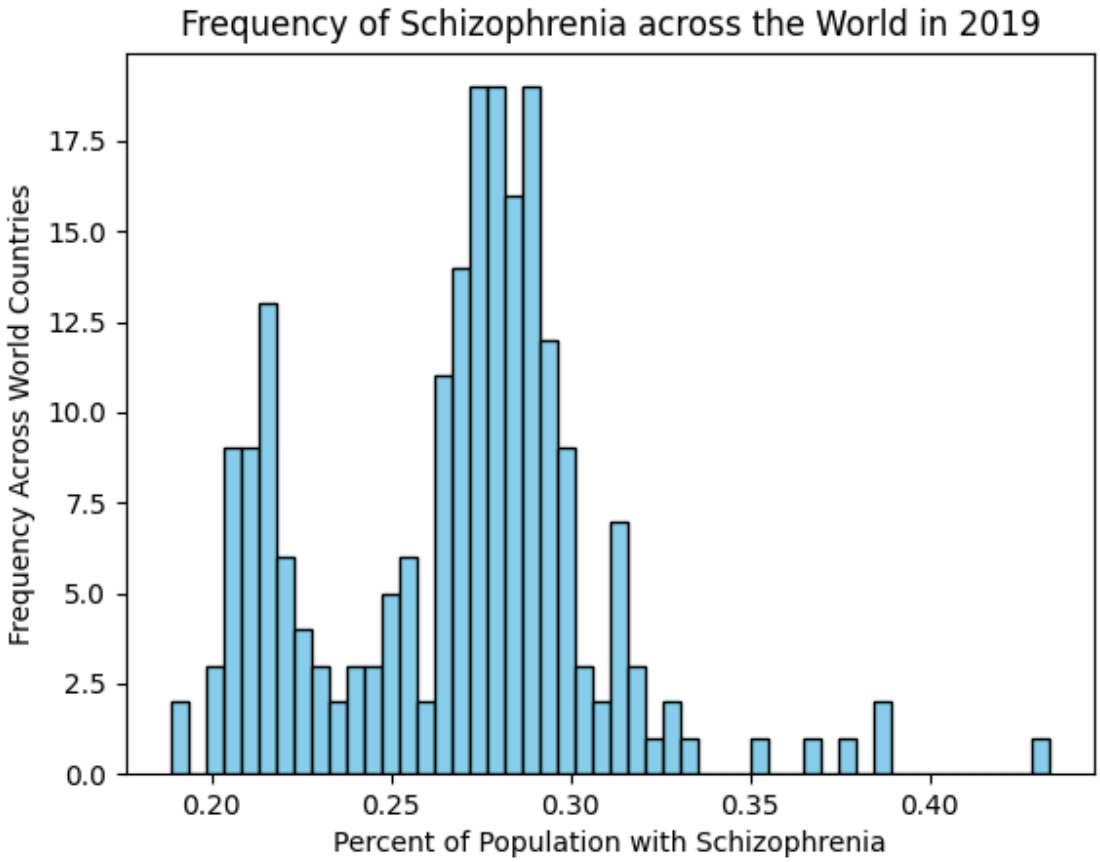
It seems like there are three ranges, 0.2-0.4%, 0.5-0.65%, 0.7-1% for bipolar distribution.



It seems Brazil was the most anxious in 2008, and reached the overall peak. Otherwise, the world is 2-8% anxious.

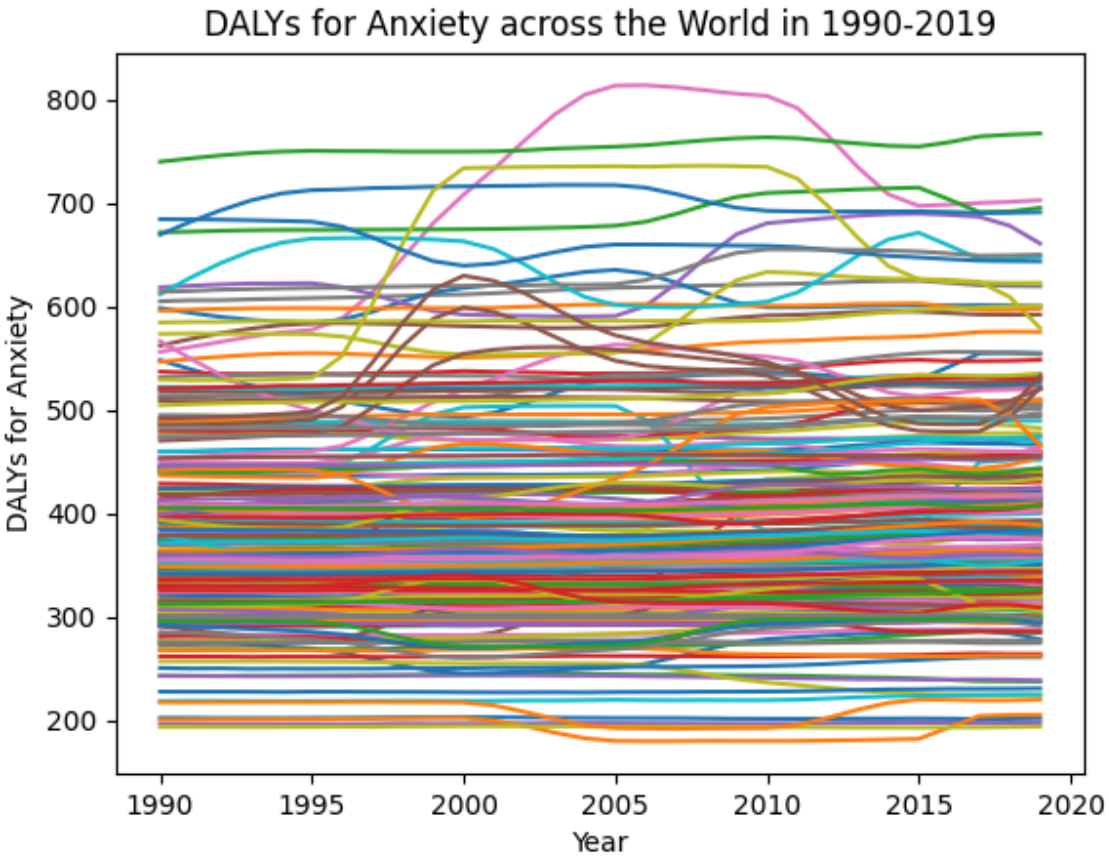


It seems that the highest amount of countries are 4% anxious, with the majority of countries falling in the 3-5% range.

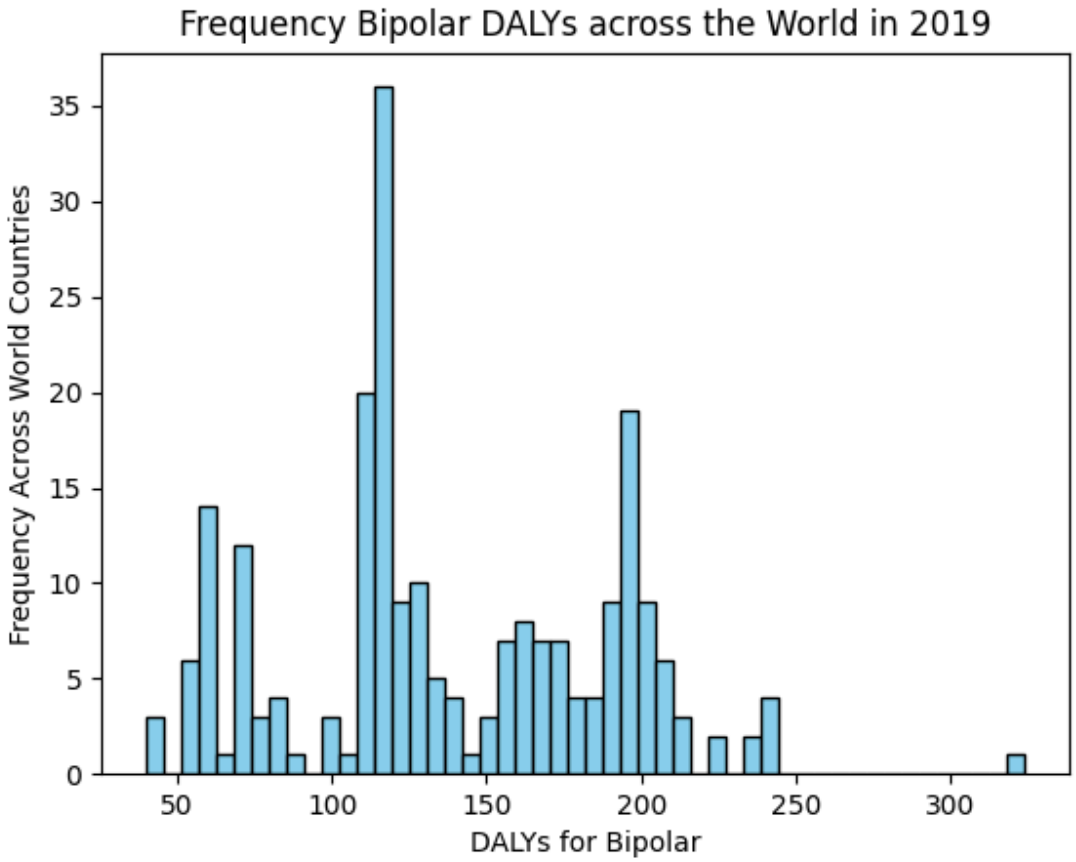


It seems like the majority of the world is 0.25-0.35 % schizophrenic.

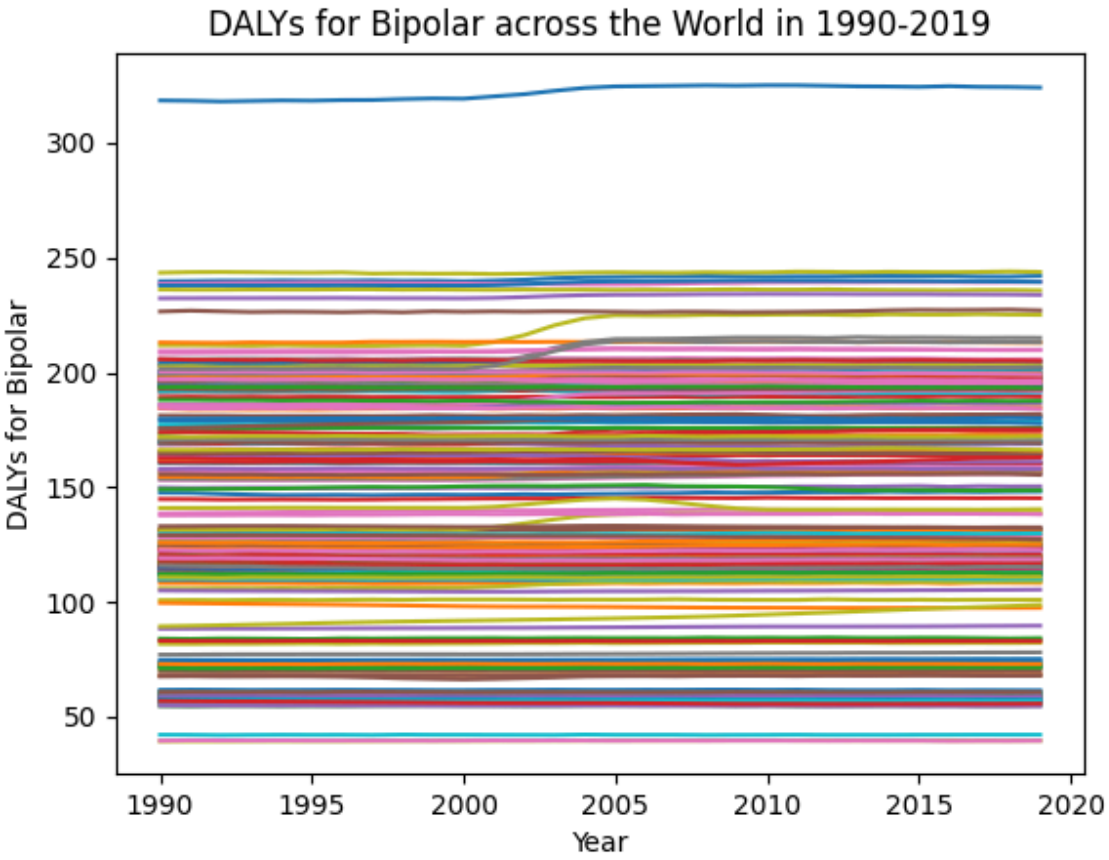
Data 1_2 Figures



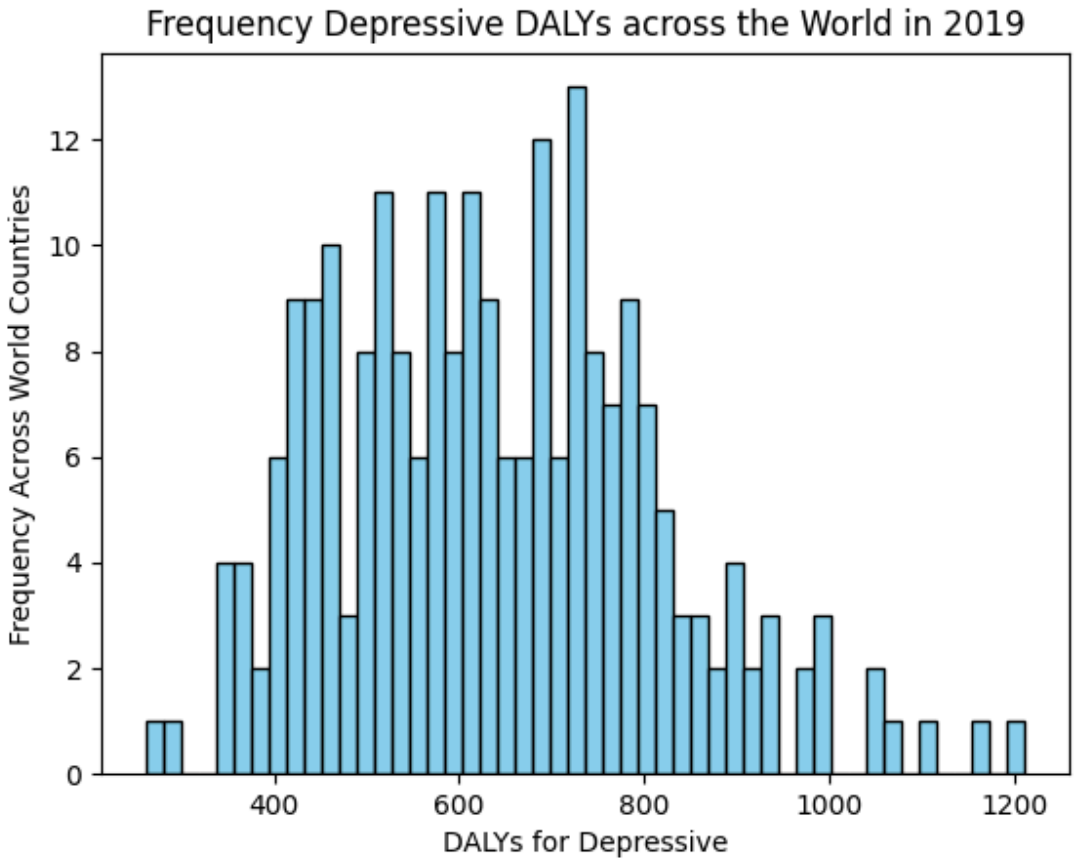
Looks like the majority of the world falls in the DALY range of 200-600.
The United States was most anxious.



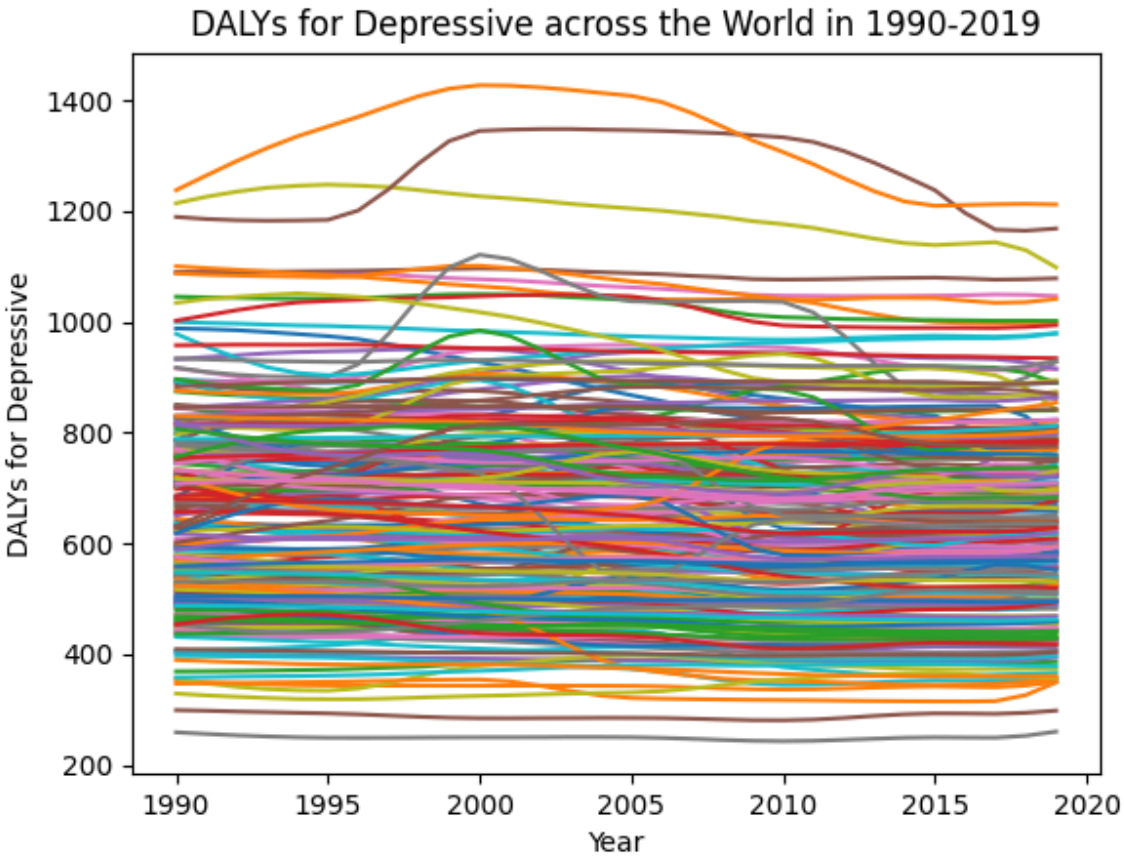
Looks like most of the world falls in the 100-230 range for bipolar.



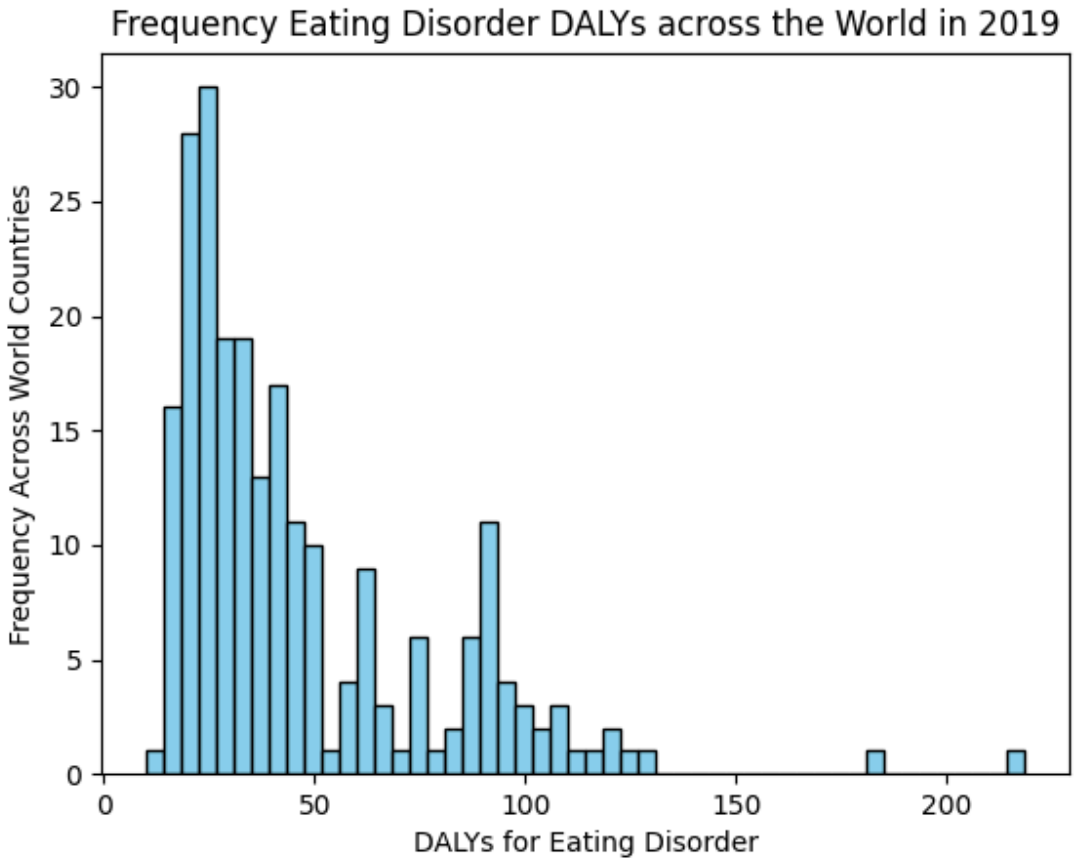
It looks like the United States was the most bipolar. Majority of the world has DALYs in 50 - 250 range



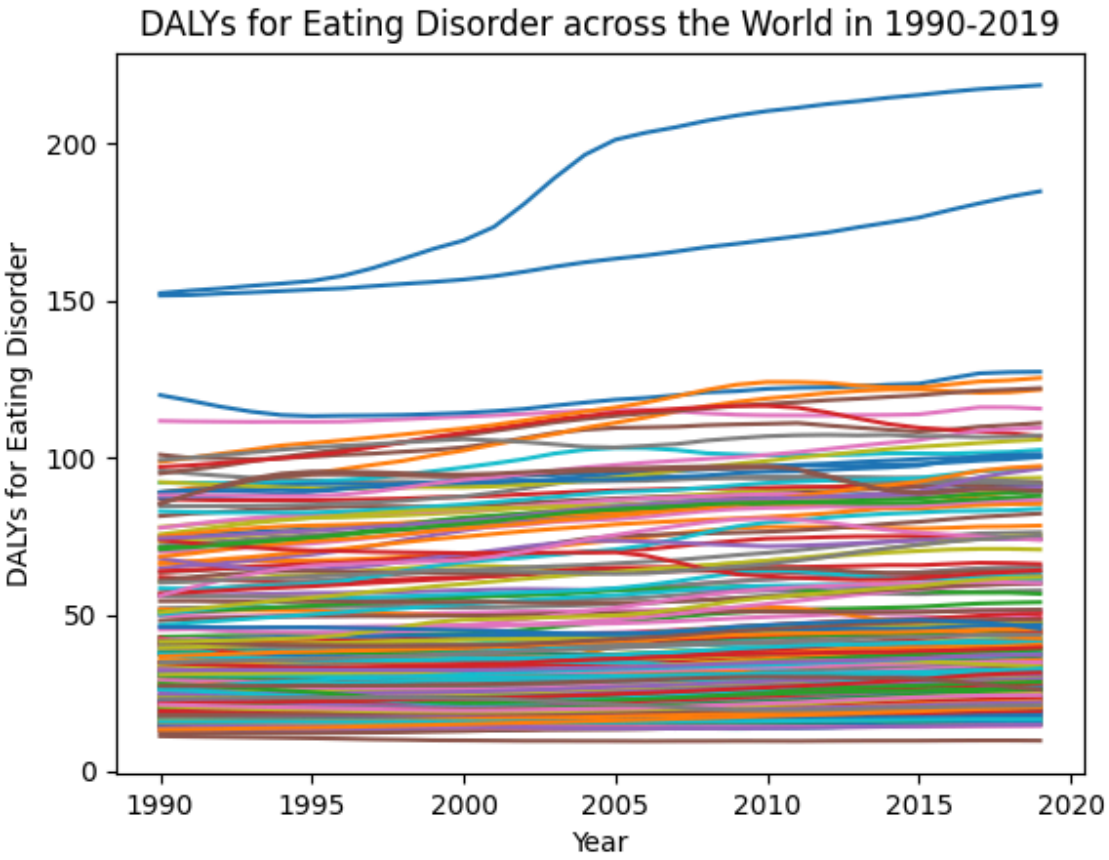
The majority of the world has DALYs in the 400-1000 range for Depression.



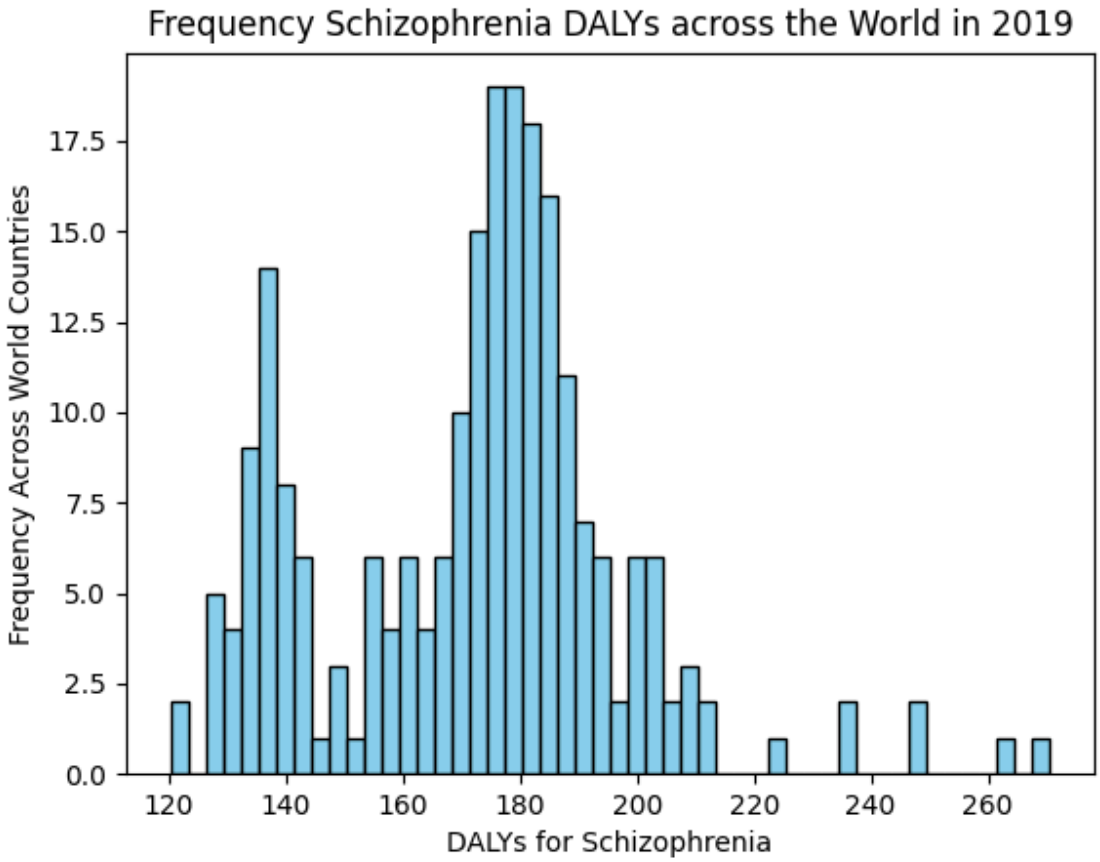
It looks like Uganda has the highest DALYs for depression.



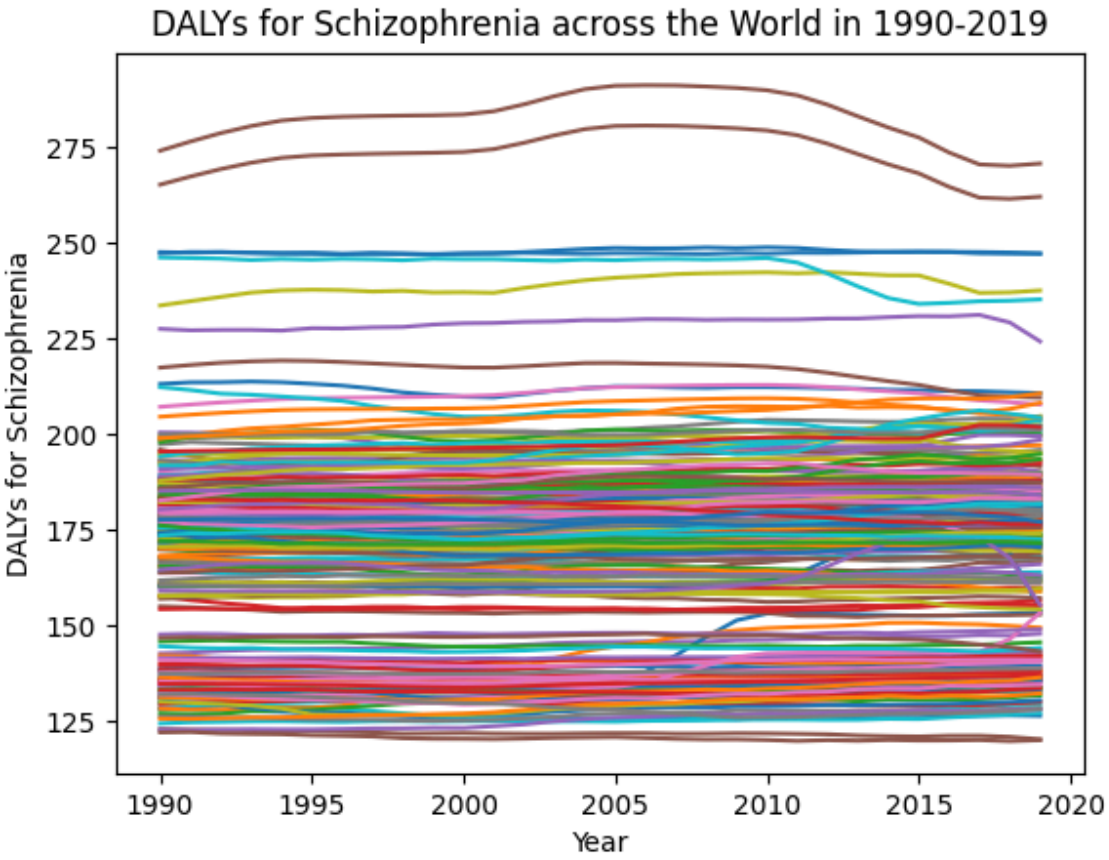
The majority of the world falls in the 0 - 100 range for DALYs for the world.



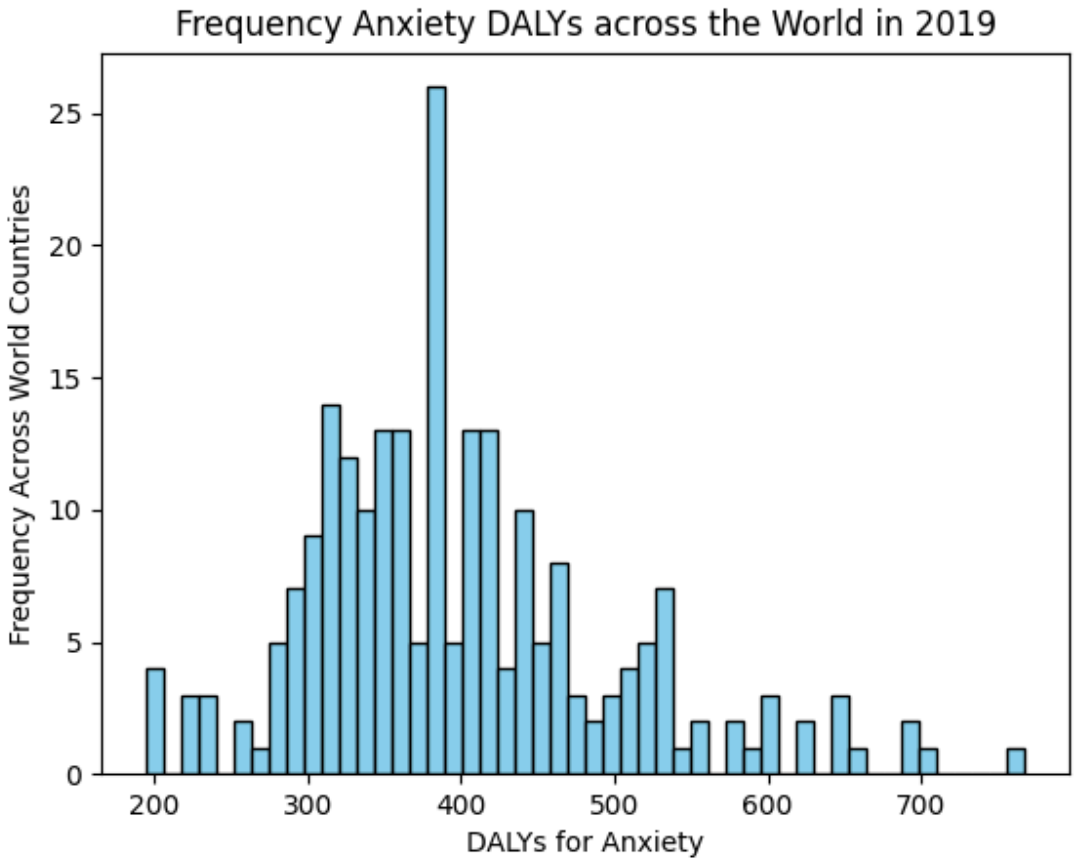
Looks like Zimbabwe is the worst for DALYs for eating disorders.



Looks like the majority of the world for Schizophrenia falls between 120-210 DALYs.

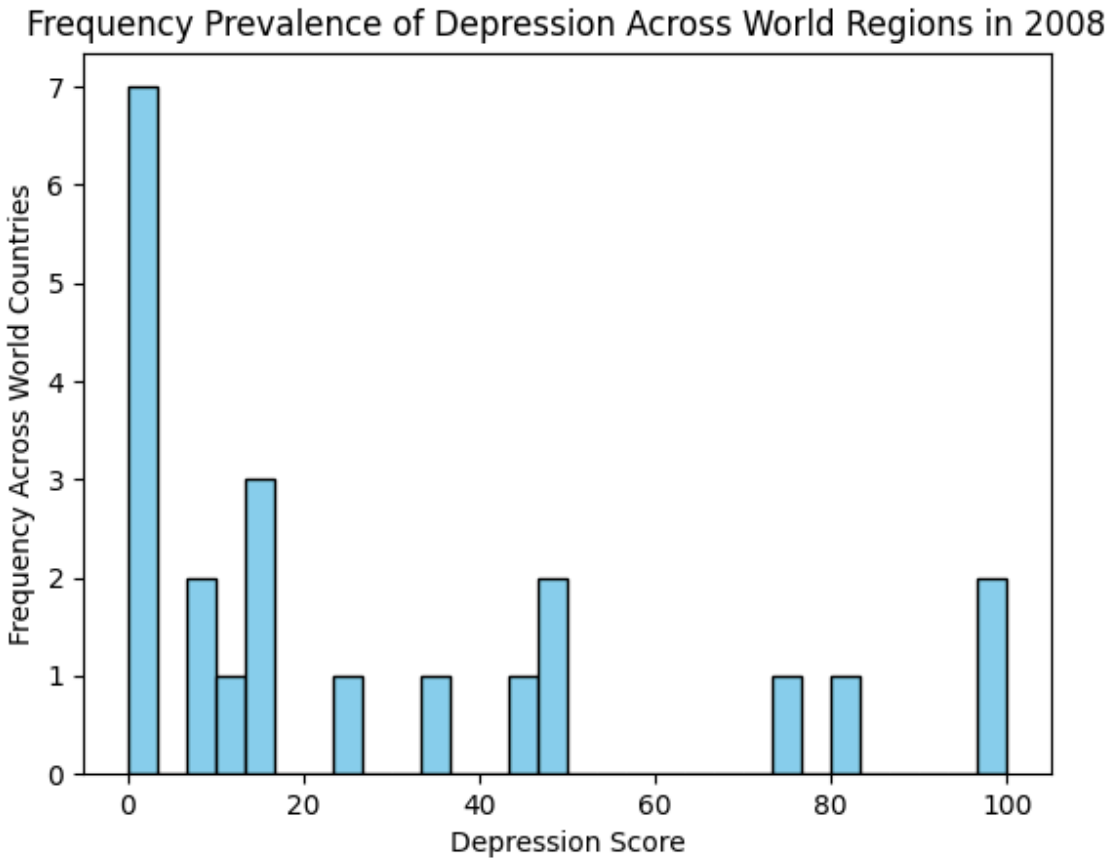


Looks like the United States has the highest schizophrenia DALYs.



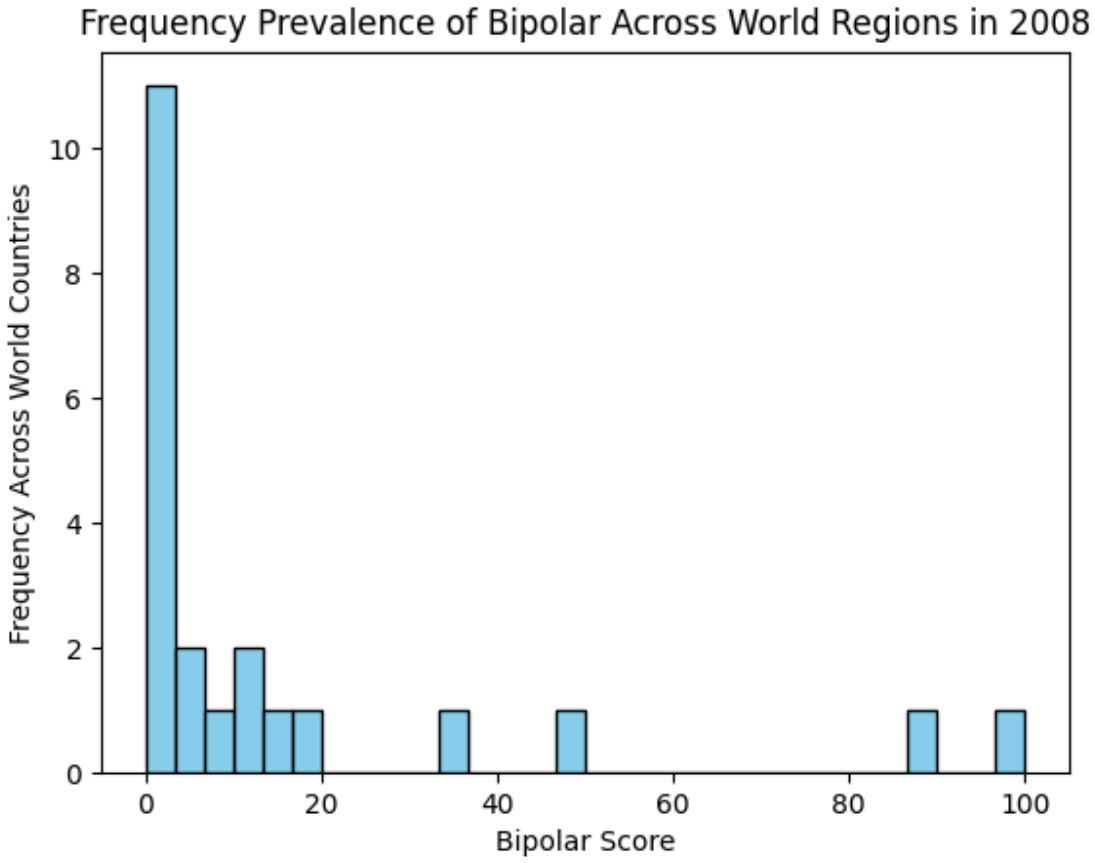
Looks like the majority of the world is 300 - 550 DALYs for Anxiety.

Data 1_3 Figures

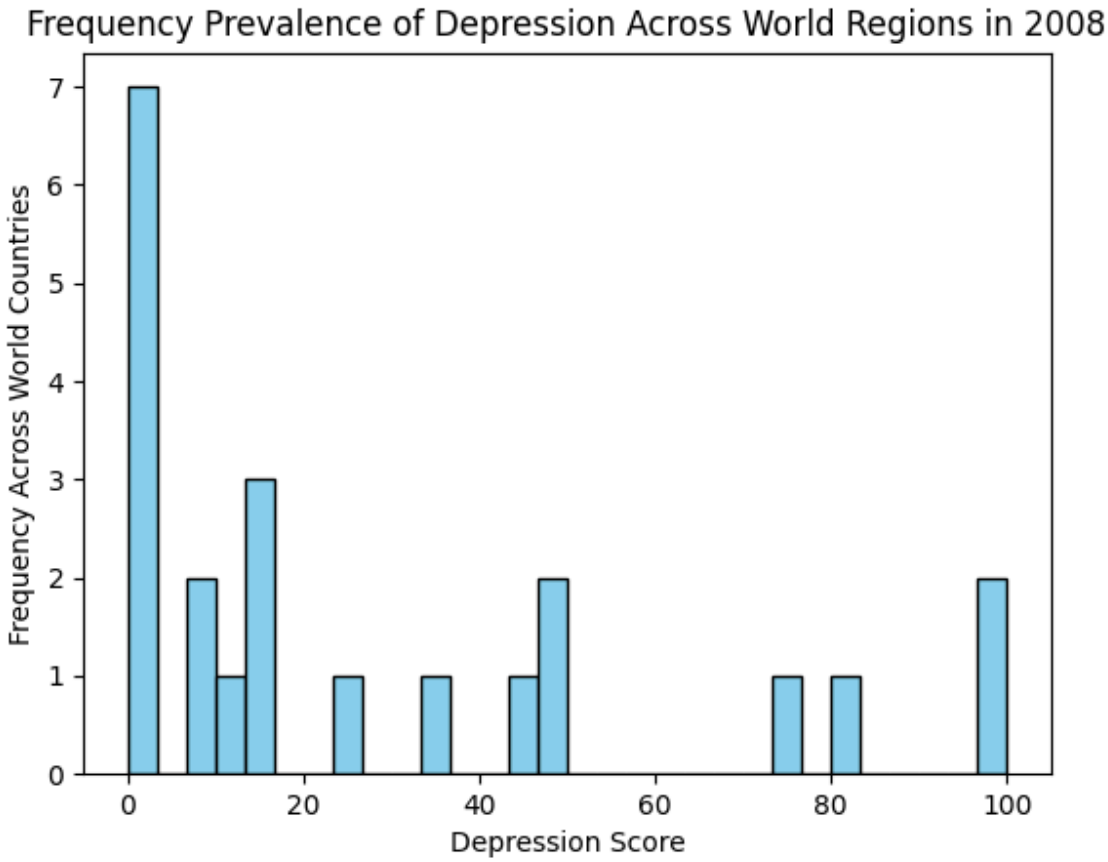


Looks like the world has a majority 0 - 50 depression score.

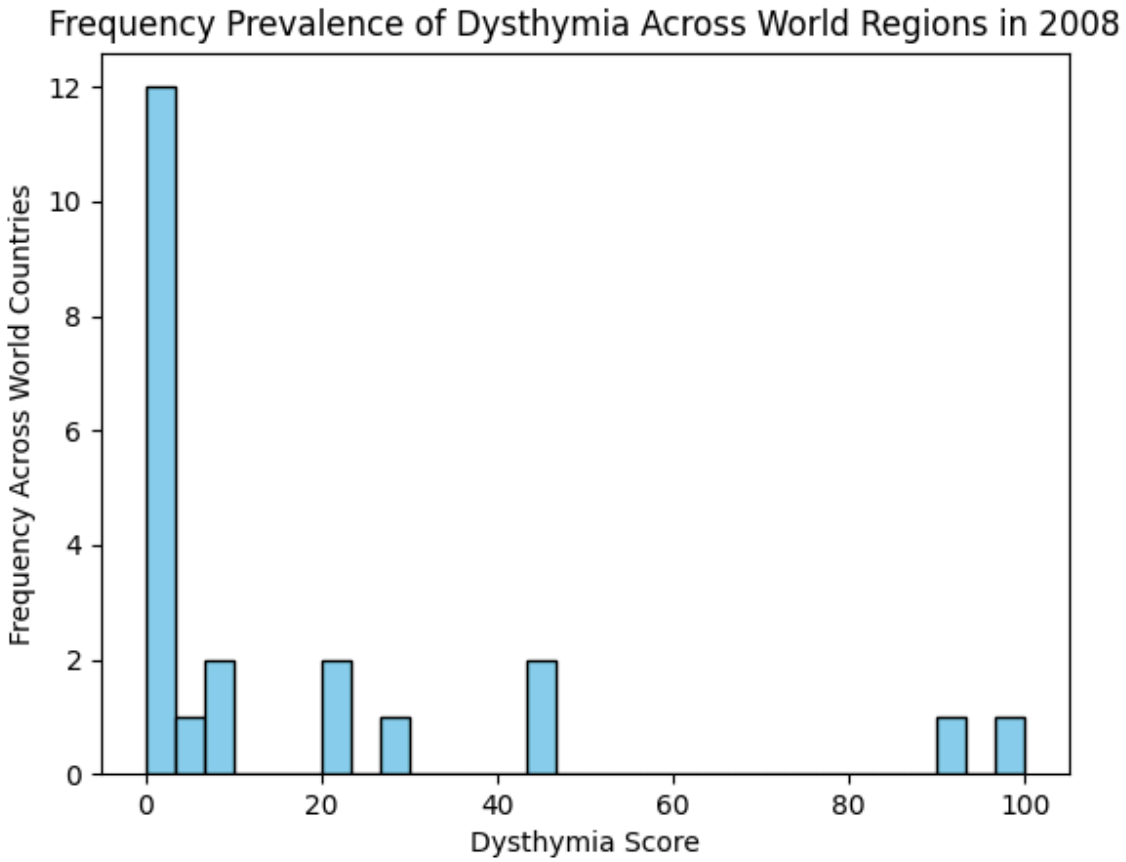
Data 1_4 Figures



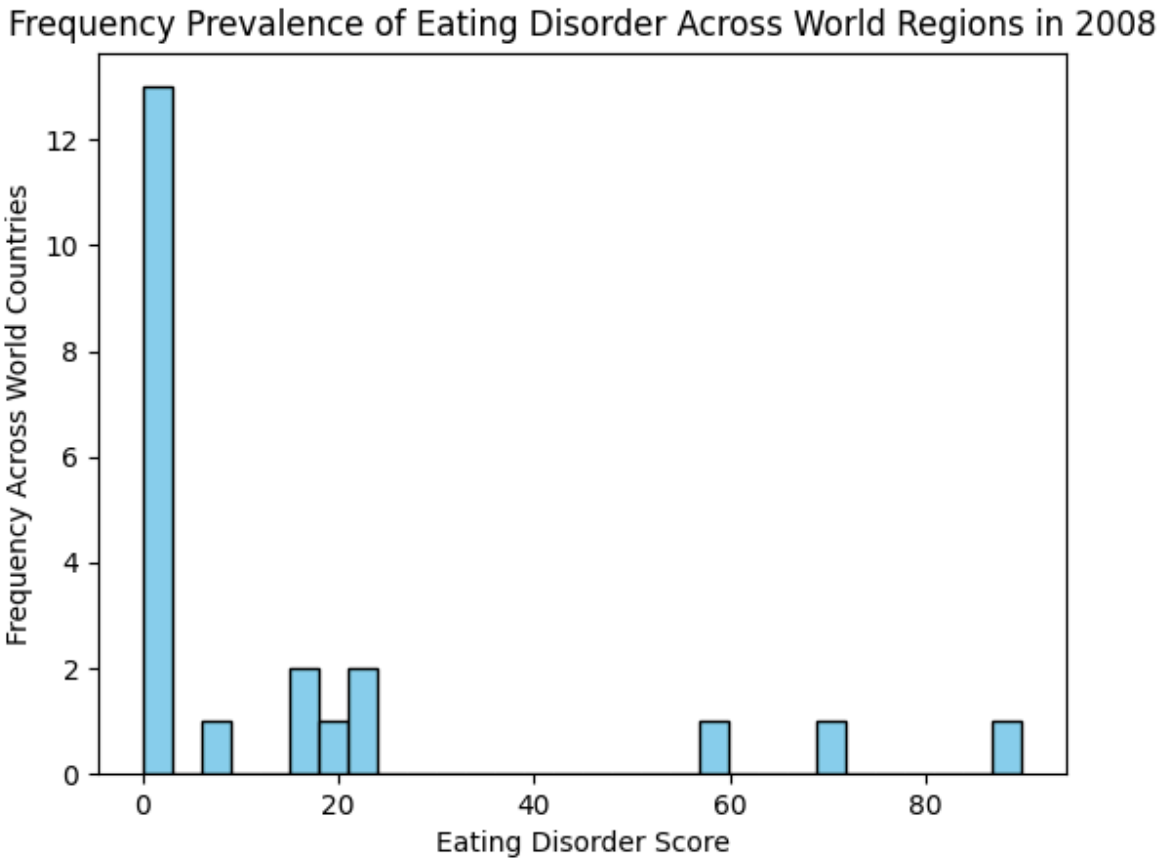
Looks like the world has a majority 0 - 20 Bipolar score.



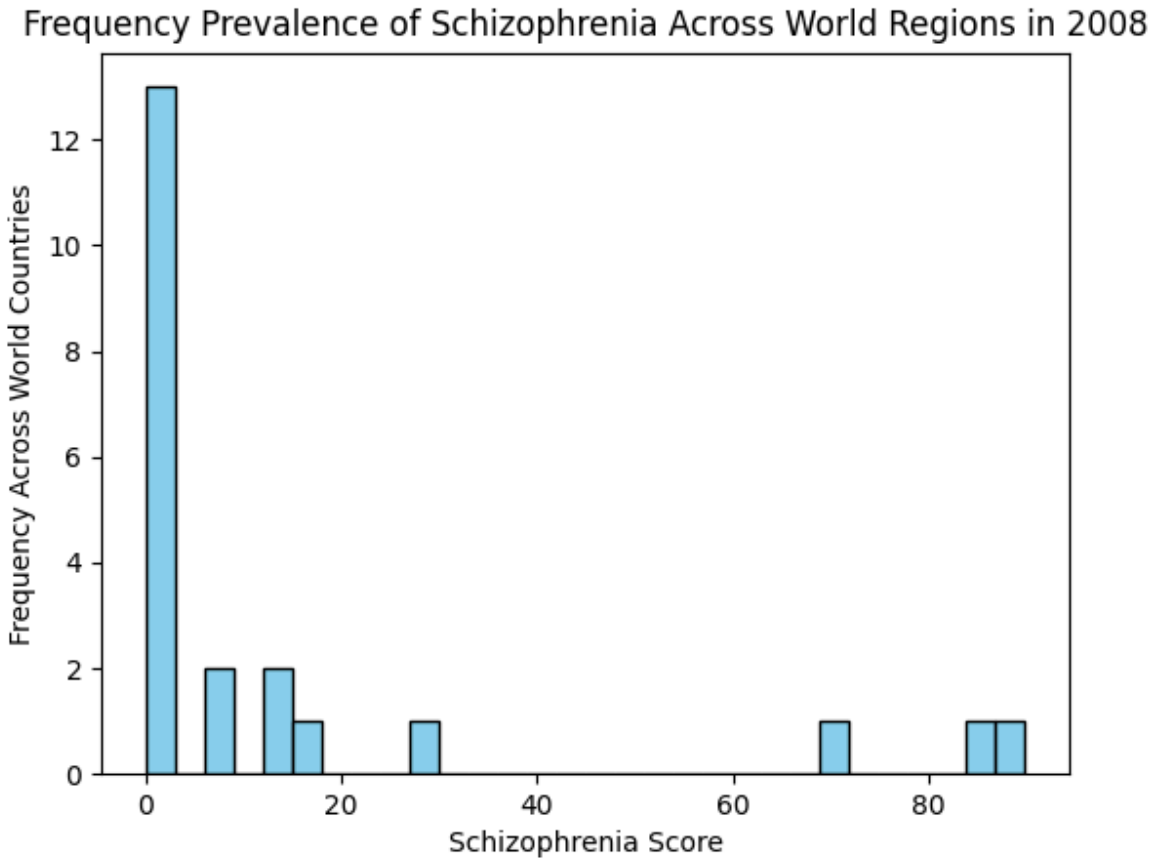
Looks like the Depression score majority was 0 - 50 for the world.



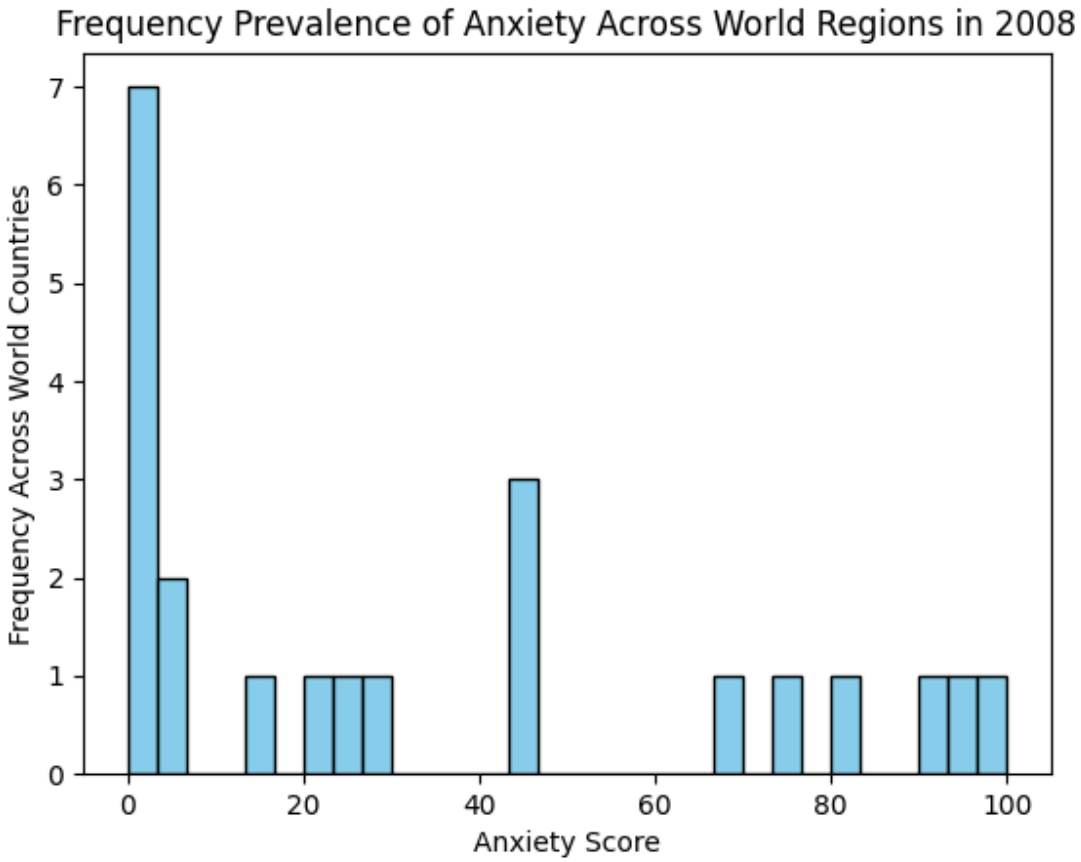
Looks like the Dysthymia score was a majority in the range 0 - 40.



Looks like the world overall falls in the 0 -20 range.

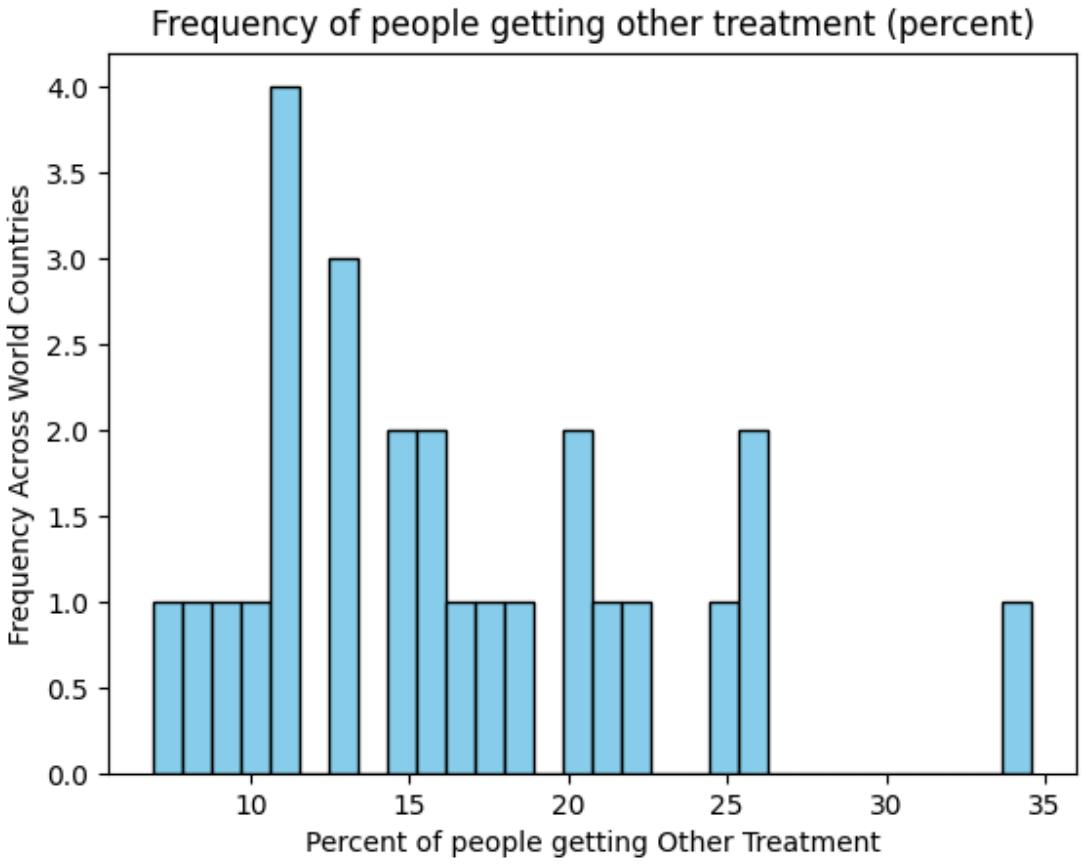


Looks like the world overall falls in the 0 - 20 range.

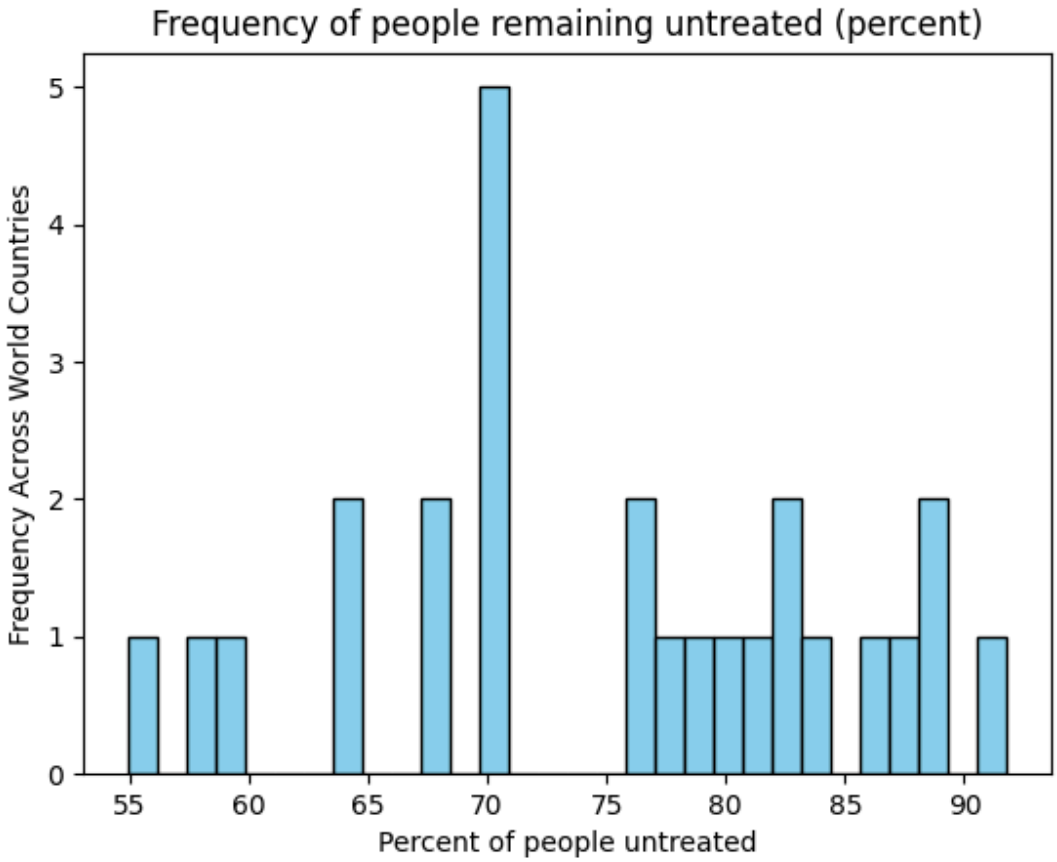


Looks like the majority of the world falls in the 0 - 50 range.

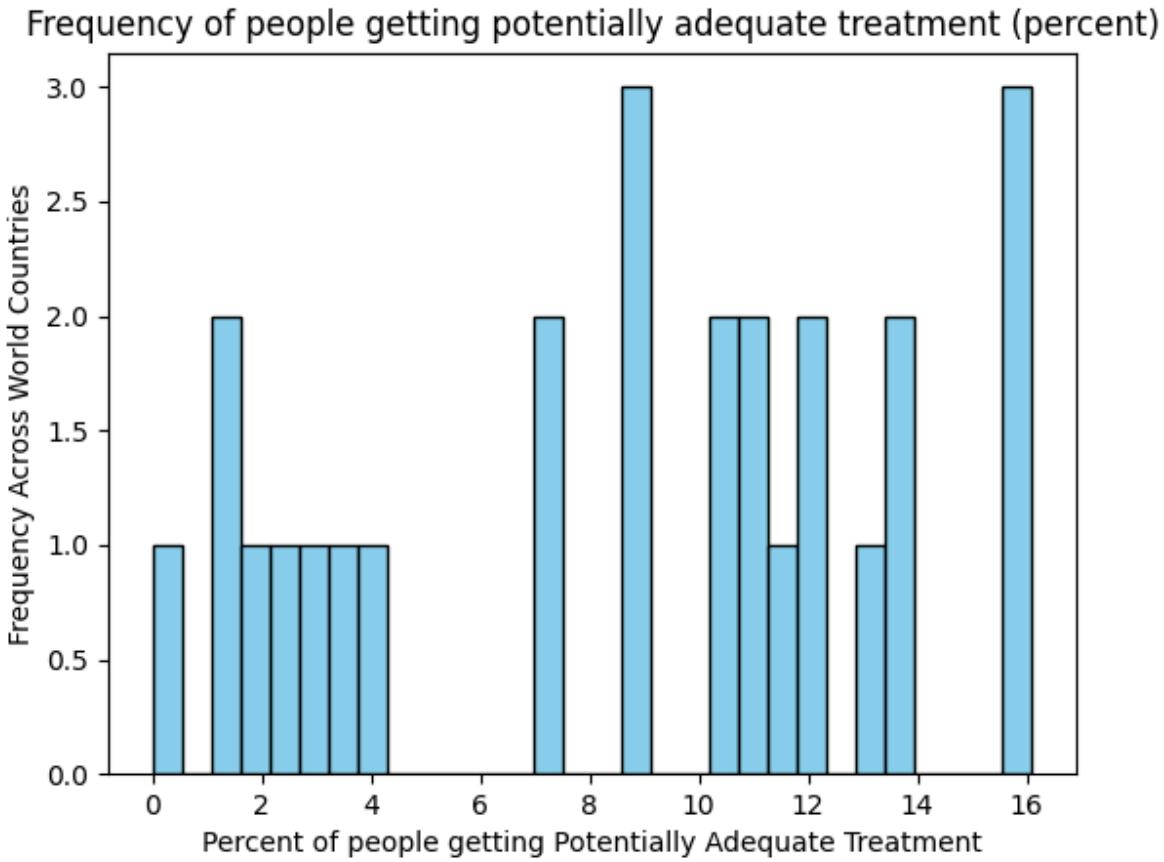
Data 1_5 Figures



It seems between 10 - 25% is the most common percent finding other treatment.

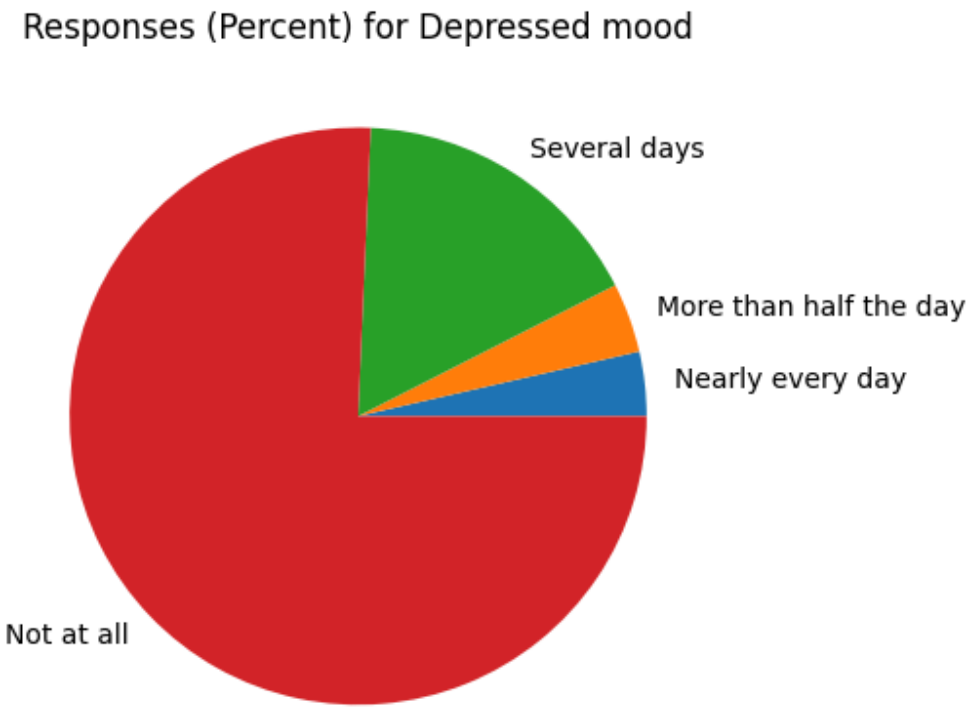


The percent of people remaining untreated is very high, falling in the 65 - 90 % range.

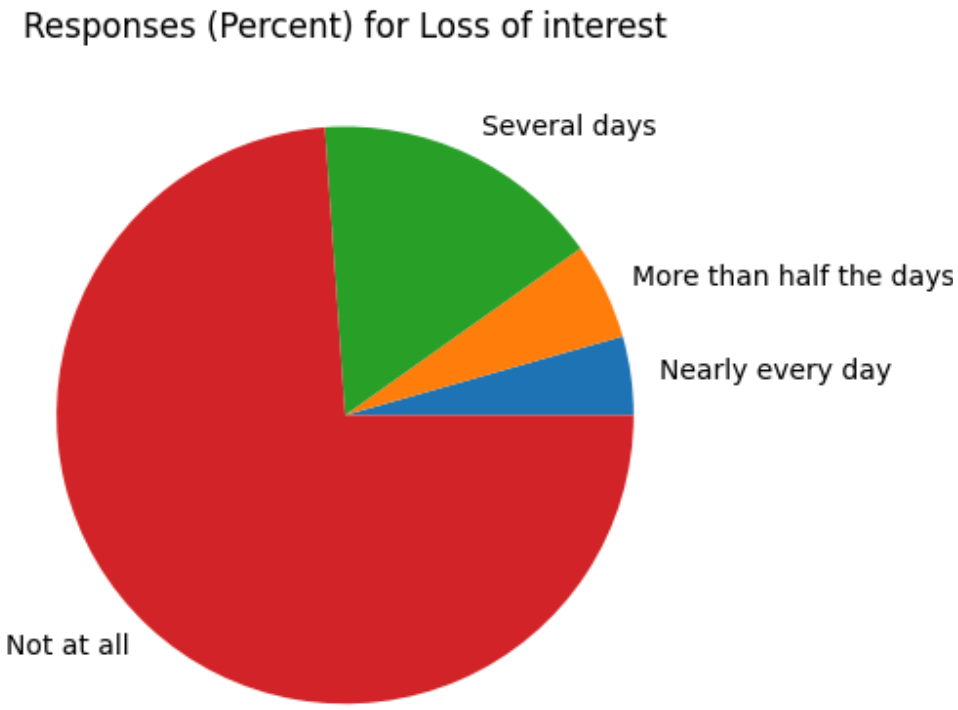


Overall the lowest chunk of people find adequate treatment. Majority falling in the 10 - 15% range

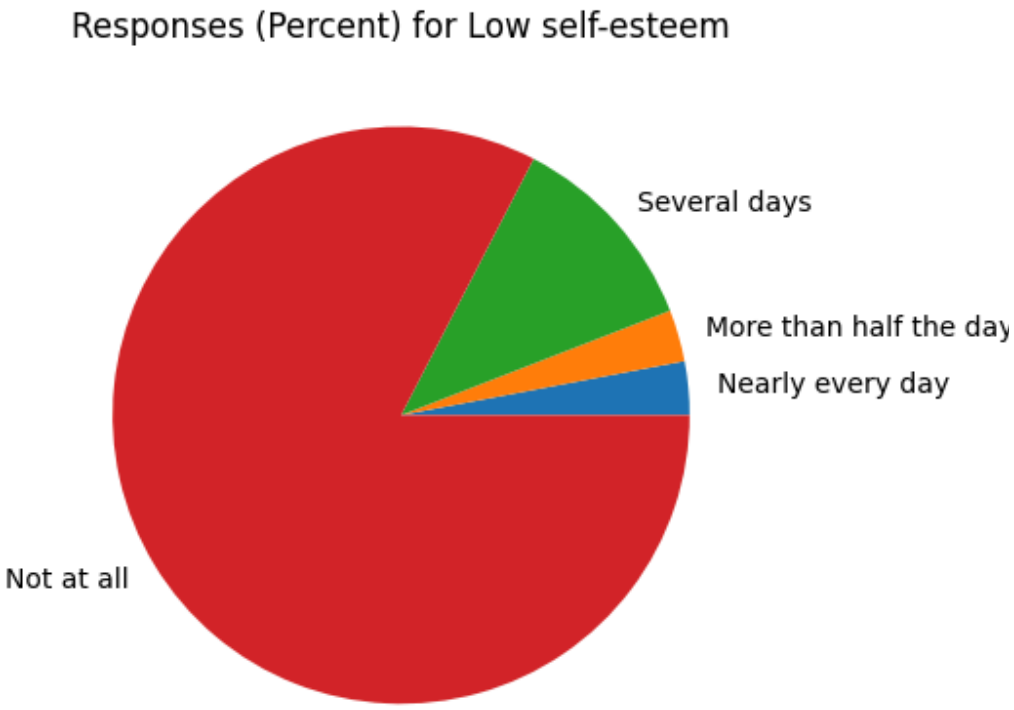
Data 1_6 Figures



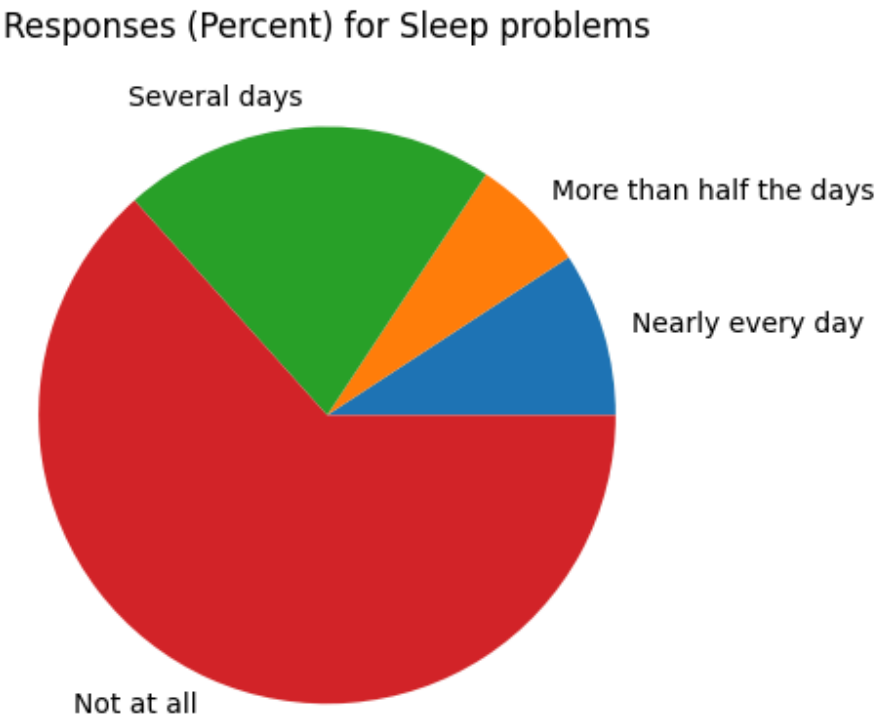
Majority seem not depressed, but a decent chunk feel depressed for several days.



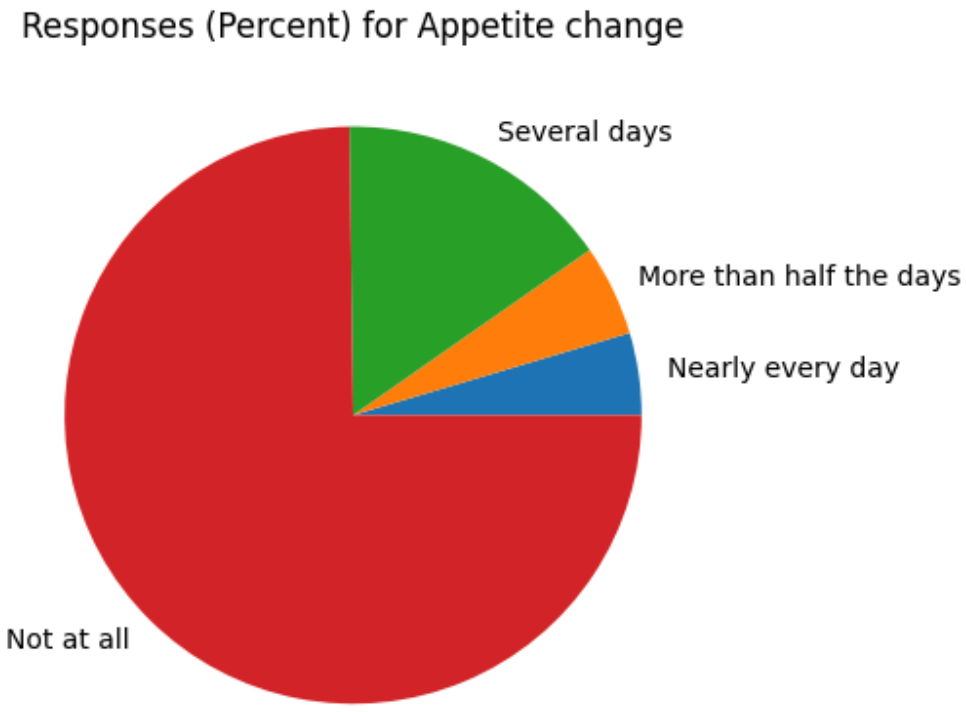
It seems like the majority of people are not feeling this symptom, but a decent chunk felt it several days.



It seems like the majority of people are not feeling this symptom, but a decent chunk felt it several days.

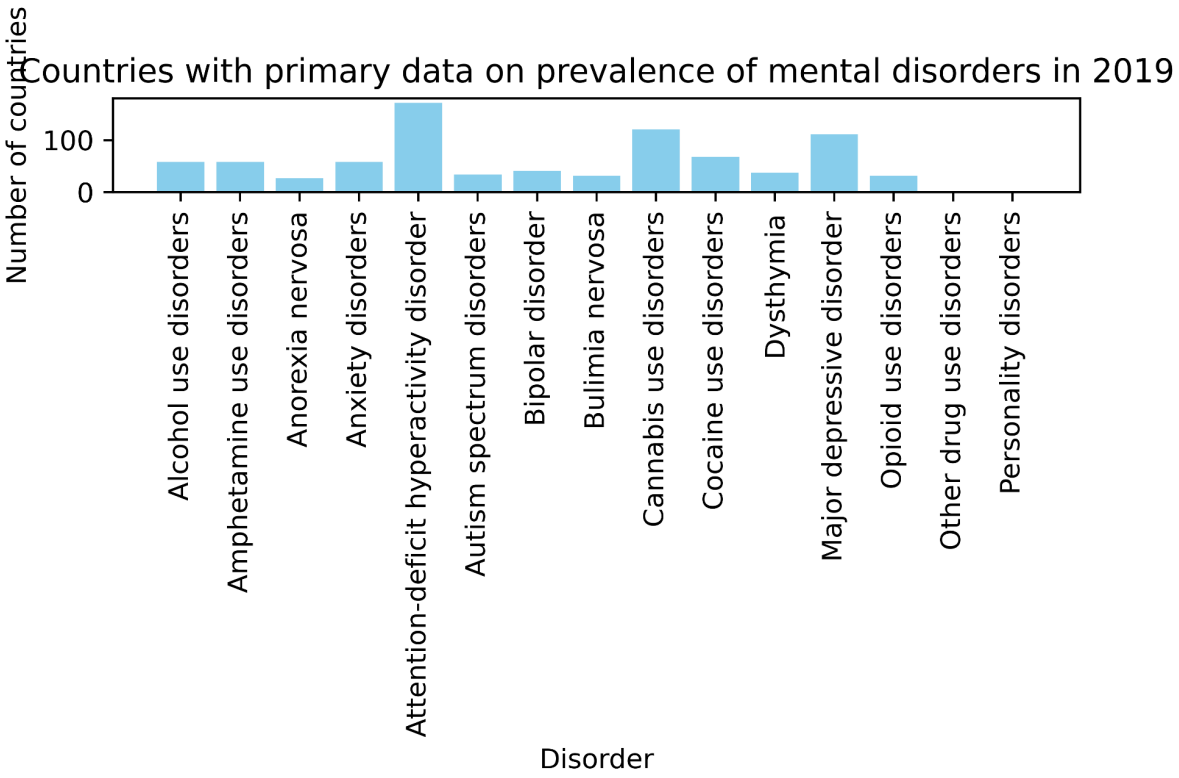


It seems like the majority of people are not feeling this symptom, but a decent chunk felt it several days.



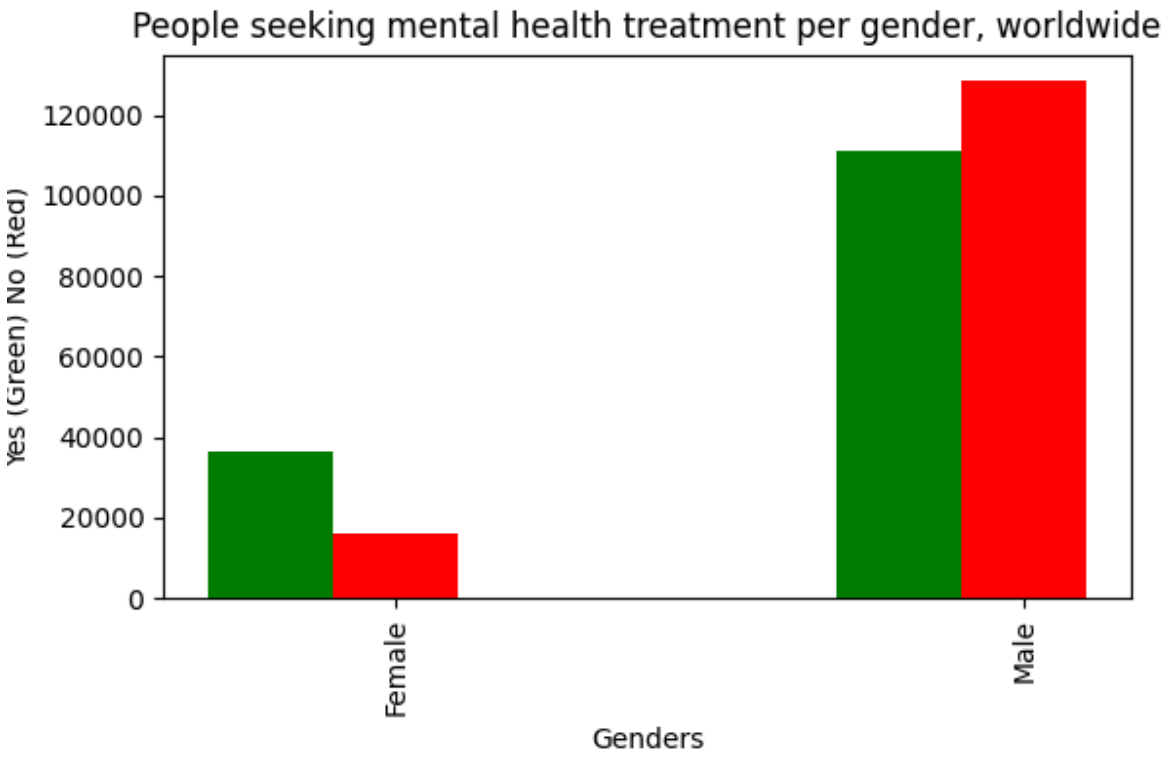
It seems like the majority of people are not feeling this symptom, but a decent chunk felt it several days.

Data 1_7 Figures

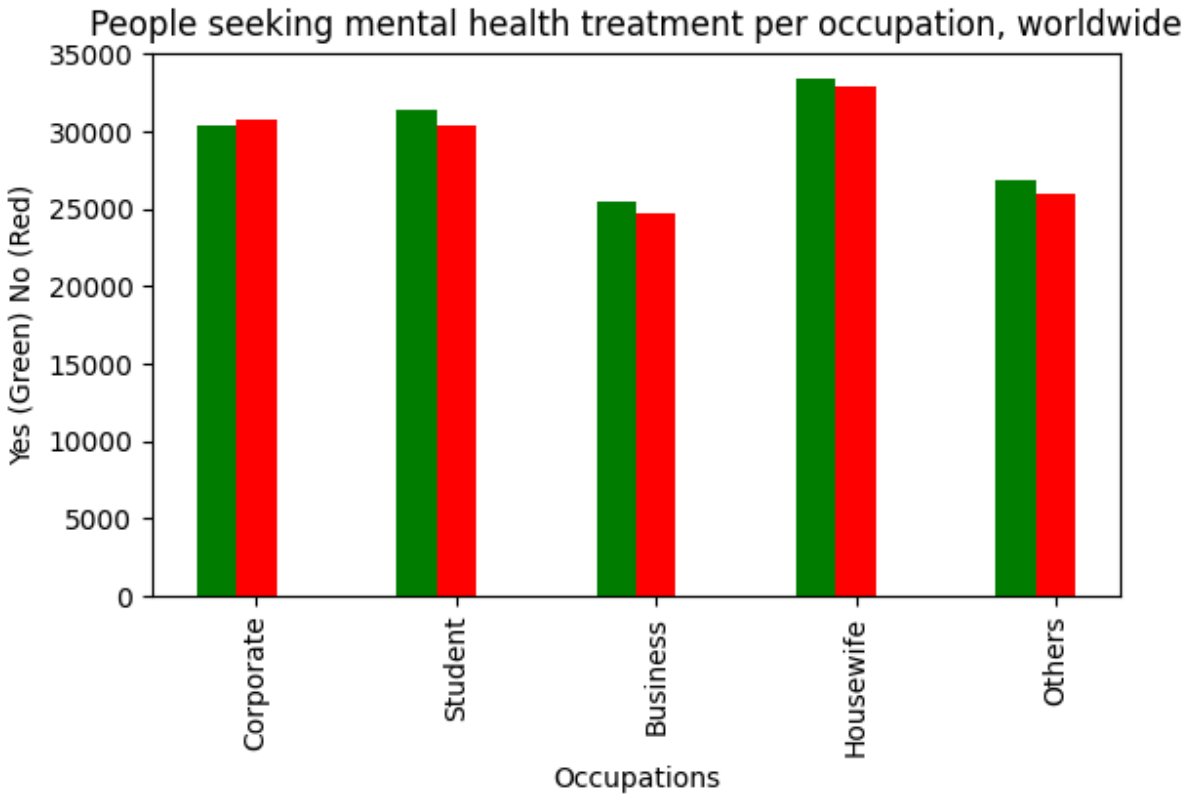


It seems like ADHD was the most prevalent disorder, with depression and cannabis use behind it.

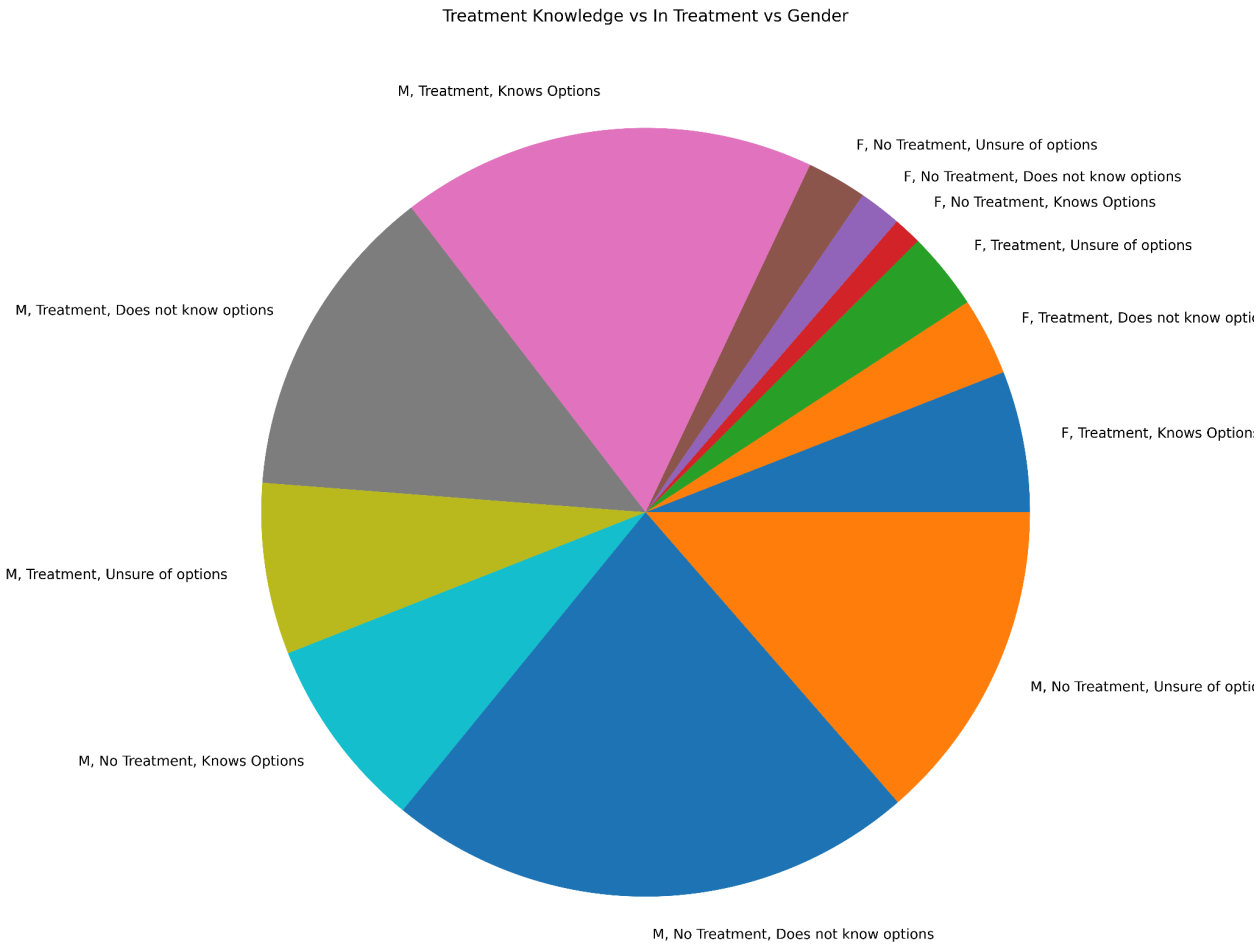
Data 2 Figures



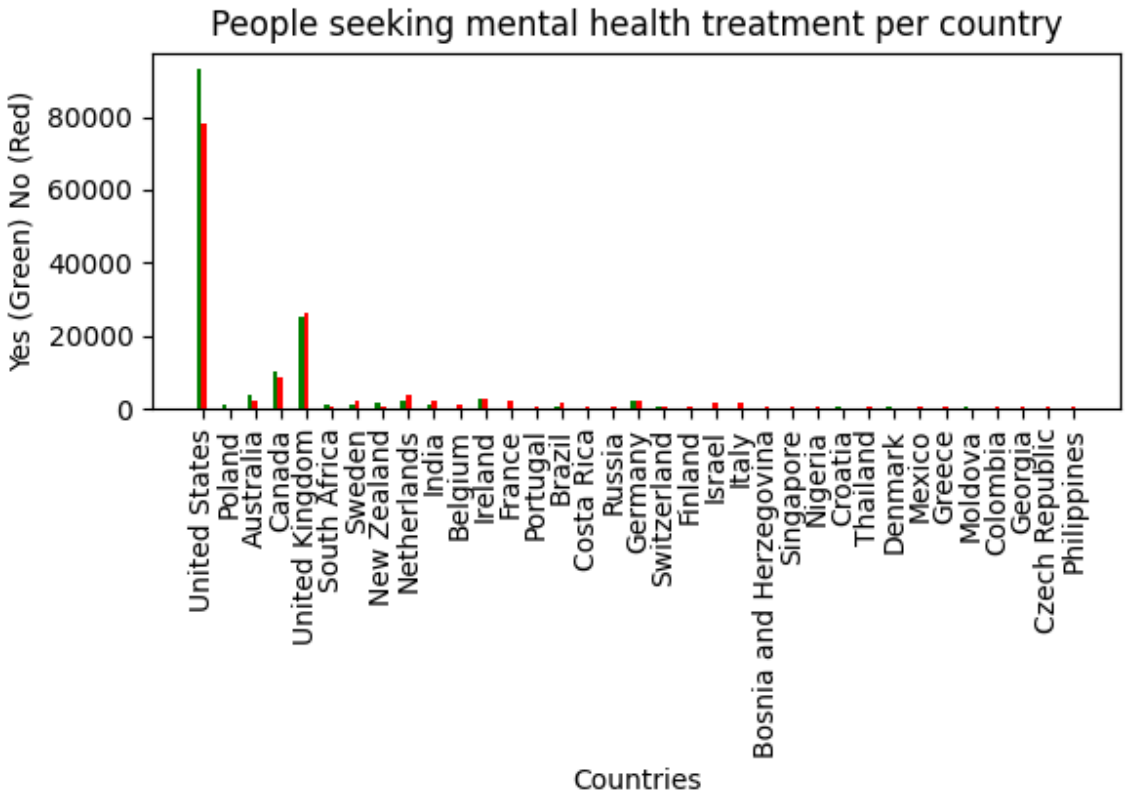
Males don’t seek treatment when compared to females.



Corporate was the only one ignoring mental health treatment, when compared to all the others.

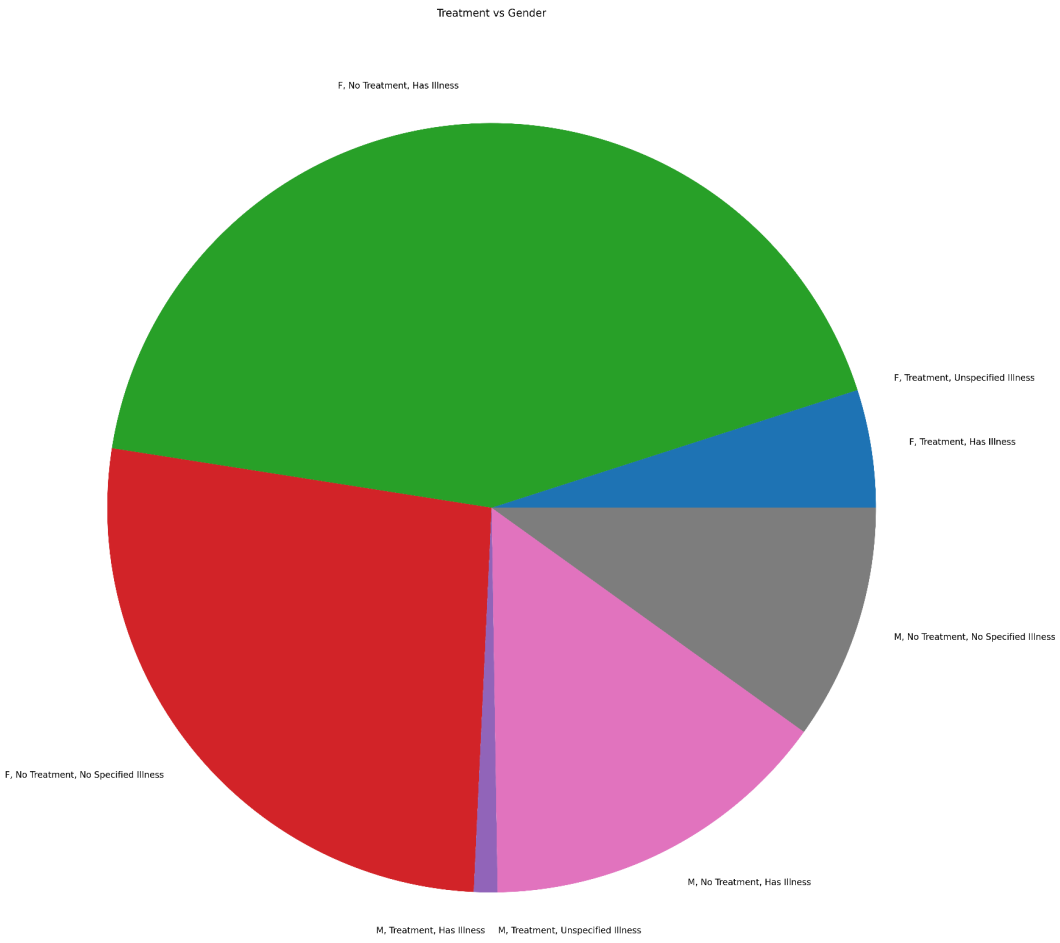


It seems the majority of men do not seek treatment and do not know options.
It seems the majority of women do seek treatment and know options.

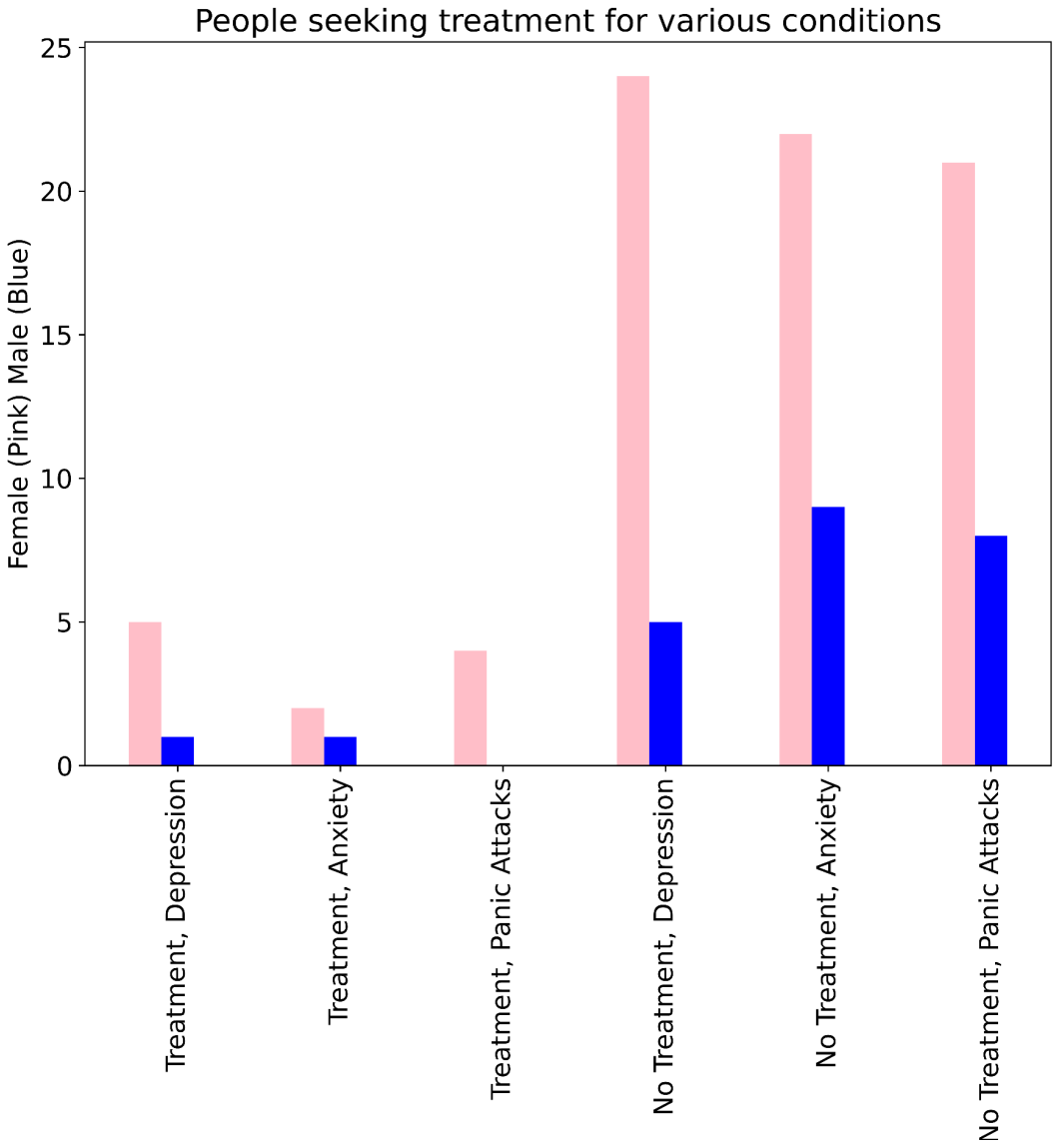


More people in the United States find treatment than don’t. The UK is the opposite, with less people finding treatment than those who do.

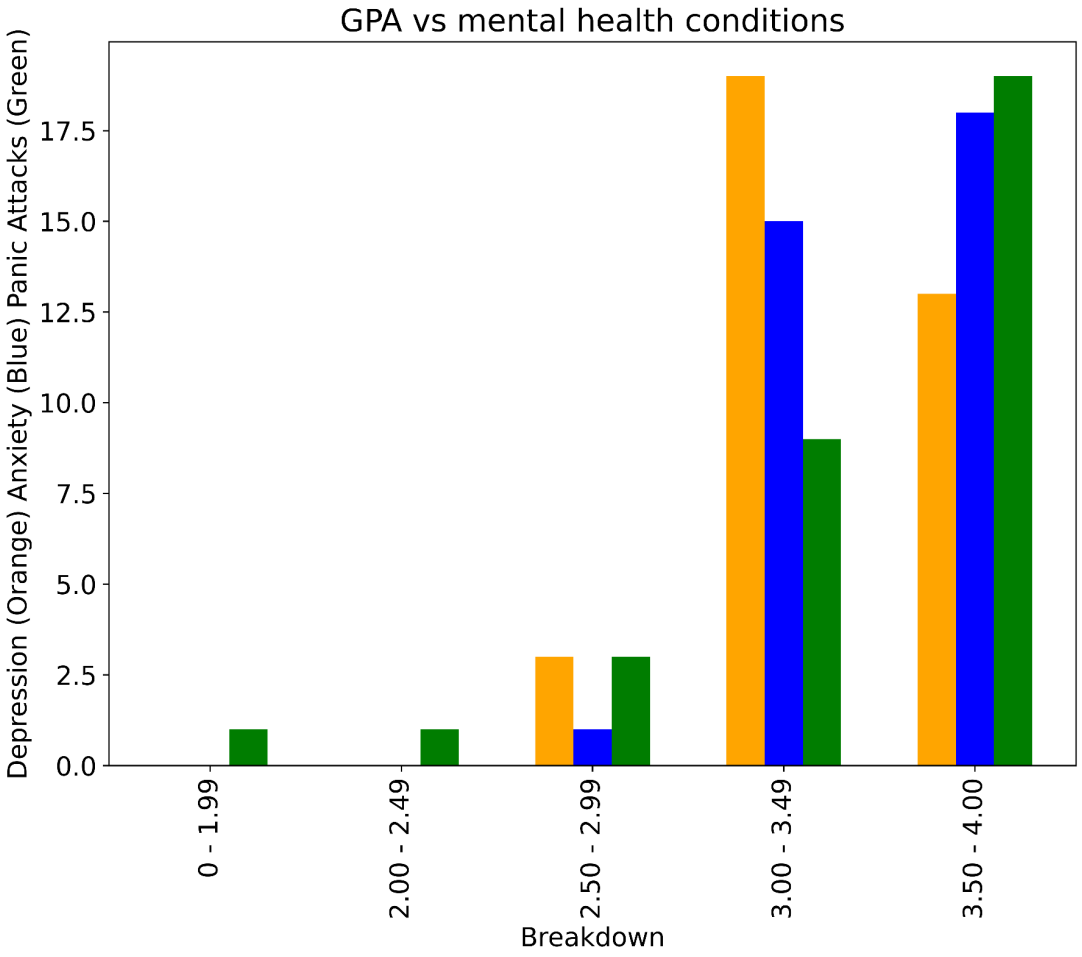
Data 3 Figures



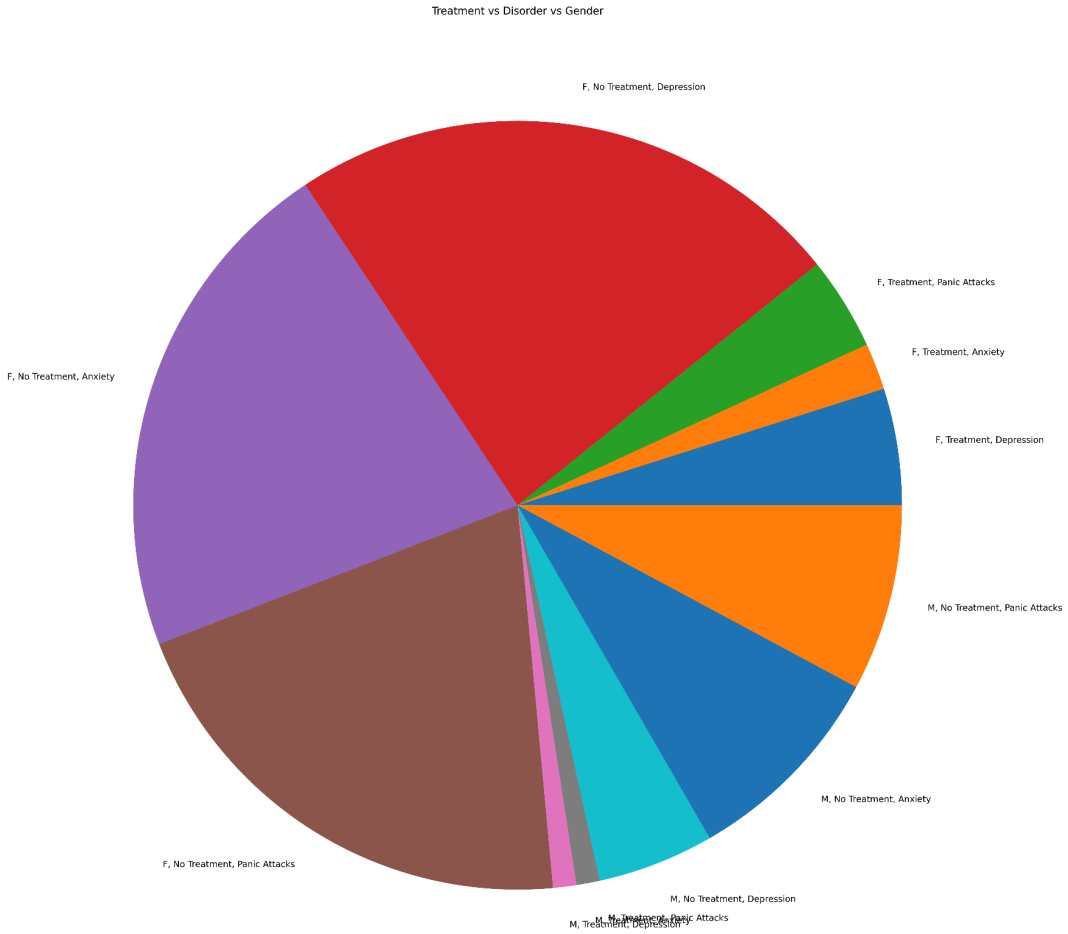
It seems women largely do not take care of their mental health as a student.



It seems like women vastly overshadow men when it comes to student distribution of mental health treatment, with more women not seeking treatment.



GPA was a significant indicator of mental health issues, with high GPA indicating high disorder chance.



It seems like females do not find treatment for any of these three illnesses on a frequent basis given this data, when compared to the males.