McKibben DSC520 Final Project 10.3

Makayla McKibben

2024-08-11

Week 8 Ex. 8.3

Introduction

Global awareness and tolerance of mental health conditions have been increasing rapidly in the last few years. However, a recent comprehensive analysis of the prevalence and most effective treatments of anxiety is not something I've come across. Mental illness and, specifically, anxiety is a topic that affects a great many people. Whether you are afflicted or have a spouse, relative, or friend who deals with mental illness, the impact of anxiety on a person's quality of life can touch nearly everyone. In order to determine the prevalence of anxiety and the effectiveness of various treatments, we will look at several datasets that provide recent, relevant information. R will be critical in analyzing this much data from these expansive datasets. While the World Health Organization published a study with data from 2019 in 2023, and the NIH published a study using data from 2001 - 2004 in 2010, I'd like to see how a more recent analysis will compare with their findings. This data and our inferences can be checked against the 2019 WHO analysis to see the progression of anxiety's pervasiveness and treatment options in the last five years and against the NIH study to see the change in the last 20 years.

Research questions

What are the most prevalent symptoms of anxiety? How many people experience symptoms of anxiety? Are specific demographics affected more heavily by anxiety? In the last five years, has the number of people affected by anxiety changed from the WHO study's predictions? In the last 20 years, has the number of people affected by anxiety changed from the NIH study? What medications are available to treat anxiety? Which medications appear to be most effective at treating anxiety? How many cases of anxiety are managed by the most effective medications?

Approach

The first part of the analysis will deal with datasets that involve the symptoms that indicate an anxiety disorder. We can establish trends from the number of people experiencing symptoms versus those diagnosed and possibly create theories about its pervasiveness; this will dovetail with the second part of our analysis. The second step will be to work with different global datasets to determine the trends of the prevalence of anxiety disorders and compare the data results between the first two analysis phases to see if our prediction about the number of people affected matches up. The last step will look at a dataset that has ratings of the effectiveness of psychiatric medications, which we will narrow down to anxiety. We can take the trends of prevalence and compare them to the effectiveness of various medications and make a supposition about whether or not certain medications seem more effective at treating anxiety. We can examine all of our datasets to see if specific demographics are particularly susceptible to experiencing anxiety and if certain medications are more effective for different demographic groups as well. We can use the data from the first two parts of our analysis and cross-correlate with the number of people prescribed effective medications to make a conjecture about how many cases of anxiety are well managed.

How your approach addresses (fully or partially) the problem.

The datasets I intend to use are more recent than those from the 2019 WHO study which used data from 2018 and 2019 and the 20-year-old data from the NIH study. I have selected six datasets, two from 2022 and the remaining four updated this year. Once I've analyzed all of this more recent data, I can compare this new information with the WHO and NIH studies and see if there have been unaccounted-for changes.

Data (Minimum of 3 Datasets - but no requirement on number of fields or rows)

Global Mental Health Disorders Kaggle https://www.kaggle.com/datasets/thedevastator/global-mental-health-disorders "This dataset contains valuable information about the prevalence of mental health disorders including schizophrenia, bipolar disorder, eating disorders, anxiety disorders, drug use disorders, depression, and alcohol use disorders from various countries across the globe." 108554 rows, 11 columns Updated 2022 Missing values left blank

Gender Mental Disorder Prevalence Kaggle https://www.kaggle.com/datasets/thedevastator/gender-mental-disorder-prevalence-2019 "This dataset provides the gender-based prevalence of mental health disorders around the world in 2019." 56396 rows, 8 columns Updated 2022 Reisha Hermana Missing values left blank

Mental Health Data (Anxiety) Kaggle https://www.kaggle.com/datasets/michellevp/predicting-anxiety-in-mental-health-data "This dataset appears to contain a variety of features related to text analysis, sentiment analysis, and psychological indicators, likely derived from posts or text data. Additionally, there are features related to psychological aspects such as economic stress, isolation, substance use, and domestic stress. The dataset seems to cover a wide range of linguistic, psychological, and behavioral attributes, potentially suitable for analyzing mental health-related topics in online communities or text data." 1968 rows, 350 columns Updated 2024 Collected 2018-2019 Missing values blank or zero

Indicators of Anxiety or Depression Kaggle https://www.kaggle.com/datasets/melissamonfared/indicators-of-anxiety-or-depression "This dataset contains information on the indicators of anxiety or depression based on the reported frequency of symptoms during the last 7 days. The data is collected through the Household Pulse Survey, launched by the U.S. Census Bureau in collaboration with five federal agencies." 16093 rows, 14 columns Updated 2024 Collected 2020-2024 No missing values

Mental Health Dataset Kaggle https://www.kaggle.com/datasets/divaniazzahra/mental-health-dataset "This dataset records a global survey conducted to track trends in mental health. The data covers a range of variables such as levels of stress, depression, anxiety, subjective well-being, and use of mental health services. The survey involved respondents from various demographic backgrounds, including gender, employment status, and geographic region." 292365 rows, 17 columns Updated 2024 Collected 2014-2016 No missing values

WebMD Reviews for Psychiatric Drugs Kaggle https://www.kaggle.com/datasets/sepidehparhami/psychiatric-drug-webmd-reviews "This dataset consists of unstructured text reviews, categorical ratings, and demographics from patients and caregivers of patients on various psychiatric drugs. The current version of the dataset contains over 61,000 reviews for hundreds of medications used to treat psychiatric disorders." 61321 rows, 13 columns Updated 2024 Collected 2007-2024 Missing data cell is left blank

Anxiety disorders World Health Organization https://www.who.int/news-room/fact-sheets/detail/anxiety-disorders

Any Anxiety Disorder NIH https://www.nimh.nih.gov/health/statistics/any-anxiety-disorder

Required Packages

At least: purrr tidyverse Metrics ggplot2

Plots and Table Needs

We will need quite a few plots and tables. At least: Number of people experiencing anxiety symptoms globally % globally by country % by country by demographic % by demographic Number of people diagnosed with anxiety globally % globally by country % by country by demographic % by demographic The most prevalent anxiety symptoms globally % globally by country % by country by demographic % by demographic Most prevalent medications globally % globally by country % by country by demographic % by demographic Ratings of medications Number of people taking specified medication The number of cases we predict will be well-managed

Questions for future steps

Do we anticipate the trends in prevalence we've found to continue over the next five years? The following 10, 20? Do we anticipate the number of people with anxiety that is well-managed to change? How? Are there newer medications that we do not have data on? Do we expect the representation of those diagnosed in specific demographics to change? How?

What do you not know how to do right now that you need to learn to answer your research questions?

I think I have all the resources and information I need to complete this project; it will just take time.

Week 9 Ex. 9.3

Code for Importing and Cleaning the Data

```r
# Import necessary packages
#install.packages("tidyverse")
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.1

## Warning: package 'ggplot2' was built under R version 4.4.1

## Warning: package 'tibble' was built under R version 4.4.1

## Warning: package 'tidyr' was built under R version 4.4.1
## Warning: package 'readr' was built under R version 4.4.1

## Warning: package 'purrr' was built under R version 4.4.1

## Warning: package 'dplyr' was built under R version 4.4.1

## Warning: package 'forcats' was built under R version 4.4.1

## Warning: package 'lubridate' was built under R version 4.4.1

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
# Import all data files
healthanxiety <-
  read.csv(file = 'healthanxiety_dataset.csv',
           header = TRUE, sep =",", stringsAsFactors = FALSE)

indicators_anxiety <-
  read.csv(file = 'Indicators_of_Anxiety_or_Depression.csv',
           header = TRUE, sep =",", stringsAsFactors = FALSE)

global_mental <-
  read.csv(file = 'Mental Health Data Global.csv',
           header = TRUE, sep =",", stringsAsFactors = FALSE)

mental <-
  read.csv(file = 'Mental Health Dataset.csv',
           header = TRUE, sep =",", stringsAsFactors = FALSE)

prevalence <-
  read.csv(file = 'prevalence-of-anxiety-disorders-males-vs-females.csv',
           header = TRUE, sep =",", stringsAsFactors = FALSE)

meds <-
  read.csv(file = 'psychiatric_drug_webmd_reviews.csv',
           header = TRUE, sep =",", stringsAsFactors = FALSE)


# Check out the data
# First dataset
#head(healthanxiety)

# Second dataset
head(indicators_anxiety)
```

```
##                          Indicator             Group         State      Subgroup
## 1 Symptoms of Depressive Disorder National Estimate United States United States
## 2 Symptoms of Depressive Disorder            By Age United States 18 - 29 years
## 3 Symptoms of Depressive Disorder            By Age United States 30 - 39 years
## 4 Symptoms of Depressive Disorder            By Age United States 40 - 49 years
## 5 Symptoms of Depressive Disorder            By Age United States 50 - 59 years
## 6 Symptoms of Depressive Disorder            By Age United States 60 - 69 years
##   Phase Time.Period    Time.Period.Label Time.Period.Start.Date
## 1   1.0           1 Apr 23 - May 5, 2020             04/23/2020
## 2   1.0           1 Apr 23 - May 5, 2020             04/23/2020
## 3   1.0           1 Apr 23 - May 5, 2020             04/23/2020
## 4   1.0           1 Apr 23 - May 5, 2020             04/23/2020
## 5   1.0           1 Apr 23 - May 5, 2020             04/23/2020
## 6   1.0           1 Apr 23 - May 5, 2020             04/23/2020
##   Time.Period.End.Date Value Low.CI High.CI Confidence.Interval Quartile.Range
## 1           05/05/2020  23.5   22.7    24.3         22.7 - 24.3
## 2           05/05/2020  32.7   30.2    35.2         30.2 - 35.2
## 3           05/05/2020  25.7   24.1    27.3         24.1 - 27.3
## 4           05/05/2020  24.8   23.3    26.2         23.3 - 26.2
## 5           05/05/2020  23.2   21.5    25.0         21.5 - 25.0
## 6           05/05/2020  18.4   17.0    19.7         17.0 - 19.7
```

```r
# Turn data into a df
indicators_anxiety <- as.data.frame(indicators_anxiety)
# Find unique indicators
unique(indicators_anxiety$Indicator)
```

```
## [1] "Symptoms of Depressive Disorder"
## [2] "Symptoms of Anxiety Disorder"
## [3] "Symptoms of Anxiety Disorder or Depressive Disorder"
```

```r
# Filter for only anxiety not depression
indicators_anxiety <- filter(indicators_anxiety,
                             Indicator == 'Symptoms of Anxiety Disorder')
# Validate that it worked
head(indicators_anxiety, 8)
```

```
##                       Indicator             Group         State
## 1 Symptoms of Anxiety Disorder National Estimate United States
## 2 Symptoms of Anxiety Disorder            By Age United States
## 3 Symptoms of Anxiety Disorder            By Age United States
## 4 Symptoms of Anxiety Disorder            By Age United States
## 5 Symptoms of Anxiety Disorder            By Age United States
## 6 Symptoms of Anxiety Disorder            By Age United States
## 7 Symptoms of Anxiety Disorder            By Age United States
## 8 Symptoms of Anxiety Disorder            By Age United States
##             Subgroup Phase Time.Period    Time.Period.Label
## 1      United States   1.0           1 Apr 23 - May 5, 2020
## 2      18 - 29 years   1.0           1 Apr 23 - May 5, 2020
## 3      30 - 39 years   1.0           1 Apr 23 - May 5, 2020
## 4      40 - 49 years   1.0           1 Apr 23 - May 5, 2020
## 5      50 - 59 years   1.0           1 Apr 23 - May 5, 2020
## 6      60 - 69 years   1.0           1 Apr 23 - May 5, 2020
## 7      70 - 79 years   1.0           1 Apr 23 - May 5, 2020
## 8 80 years and above   1.0           1 Apr 23 - May 5, 2020
##   Time.Period.Start.Date Time.Period.End.Date Value Low.CI High.CI
## 1             04/23/2020           05/05/2020  30.8   30.0    31.7
## 2             04/23/2020           05/05/2020  40.2   38.1    42.4
## 3             04/23/2020           05/05/2020  34.4   32.6    36.1
## 4             04/23/2020           05/05/2020  34.1   32.1    36.2
## 5             04/23/2020           05/05/2020  31.0   29.0    33.1
## 6             04/23/2020           05/05/2020  24.9   23.6    26.3
## 7             04/23/2020           05/05/2020  16.4   14.8    18.1
## 8             04/23/2020           05/05/2020  14.6   11.5    18.2
##   Confidence.Interval Quartile.Range
## 1         30.0 - 31.7
## 2         38.1 - 42.4
## 3         32.6 - 36.1
## 4         32.1 - 36.2
## 5         29.0 - 33.1
## 6         23.6 - 26.3
## 7         14.8 - 18.1
## 8         11.5 - 18.2
```

```r
# Third dataset
head(global_mental)
```

```
##   index      Entity Code Year Schizophrenia.... Bipolar.disorder....
## 1     0 Afghanistan  AFG 1990           0.16056             0.697779
## 2     1 Afghanistan  AFG 1991          0.160312             0.697961
## 3     2 Afghanistan  AFG 1992          0.160135             0.698107
## 4     3 Afghanistan  AFG 1993          0.160037             0.698257
## 5     4 Afghanistan  AFG 1994          0.160022             0.698469
## 6     5 Afghanistan  AFG 1995          0.160076             0.698695
##   Eating.disorders.... Anxiety.disorders.... Drug.use.disorders....
## 1             0.101855              4.828830               1.677082
## 2             0.099313              4.829740               1.684746
## 3             0.096692              4.831108               1.694334
## 4             0.094336              4.830864               1.705320
## 5             0.092439              4.829423               1.716069
## 6              0.09098              4.828337               1.728112
##   Depression.... Alcohol.use.disorders....
## 1       4.071831                  0.672404
## 2       4.079531                  0.671768
## 3       4.088358                  0.670644
## 4       4.096190                  0.669738
## 5       4.099582                  0.669260
## 6       4.104207                  0.668746
```

```r
# Turn data into a df
global_mental <- as.data.frame(global_mental)
# Remove columns that we don't need
colnames(global_mental)
```

```
##  [1] "index"                     "Entity"
##  [3] "Code"                      "Year"
##  [5] "Schizophrenia...."         "Bipolar.disorder...."
##  [7] "Eating.disorders...."      "Anxiety.disorders...."
##  [9] "Drug.use.disorders...."    "Depression...."
## [11] "Alcohol.use.disorders...."
```

```
global_mental <- subset(global_mental, select = -c(Schizophrenia....,
                                                   Bipolar.disorder....,
                                                   Eating.disorders....,
                                                   Drug.use.disorders....,
                                                   Depression....,
                                                   Alcohol.use.disorders....))
# Remove rows missing data
global_mental <- global_mental[complete.cases(global_mental),]
# Rename
global_mental <- global_mental %>%
  rename(anxiety = Anxiety.disorders....)
# Check that the data has been trimmed down
head(global_mental, 8)
```

```
##   index      Entity Code Year  anxiety
## 1     0 Afghanistan  AFG 1990 4.828830
## 2     1 Afghanistan  AFG 1991 4.829740
## 3     2 Afghanistan  AFG 1992 4.831108
## 4     3 Afghanistan  AFG 1993 4.830864
## 5     4 Afghanistan  AFG 1994 4.829423
## 6     5 Afghanistan  AFG 1995 4.828337
## 7     6 Afghanistan  AFG 1996 4.828083
## 8     7 Afghanistan  AFG 1997 4.827726
```

```
# Fourth dataset
head(mental)
```

```
##             Timestamp Gender       Country Occupation self_employed
## 1 2014-08-27 11:29:31 Female United States  Corporate
## 2 2014-08-27 11:31:50 Female United States  Corporate
## 3 2014-08-27 11:32:39 Female United States  Corporate
## 4 2014-08-27 11:37:59 Female United States  Corporate            No
## 5 2014-08-27 11:43:36 Female United States  Corporate            No
## 6 2014-08-27 11:49:51 Female        Poland  Corporate            No
##   family_history treatment Days_Indoors Growing_Stress Changes_Habits
## 1             No       Yes    1-14 days            Yes             No
## 2            Yes       Yes    1-14 days            Yes             No
## 3            Yes       Yes    1-14 days            Yes             No
## 4            Yes       Yes    1-14 days            Yes             No
## 5            Yes       Yes    1-14 days            Yes             No
## 6             No       Yes    1-14 days            Yes             No
##   Mental_Health_History Mood_Swings Coping_Struggles Work_Interest
## 1                   Yes      Medium               No            No
## 2                   Yes      Medium               No            No
## 3                   Yes      Medium               No            No
## 4                   Yes      Medium               No            No
## 5                   Yes      Medium               No            No
## 6                   Yes      Medium               No            No
##   Social_Weakness mental_health_interview care_options
## 1             Yes                      No     Not sure
## 2             Yes                      No           No
## 3             Yes                      No          Yes
## 4             Yes                   Maybe          Yes
## 5             Yes                      No          Yes
## 6             Yes                   Maybe     Not sure
```

```
# Turn data into dataframe
mental <- as.data.frame(mental)
# Remove columns that we don't need
colnames(mental)
```

```
##  [1] "Timestamp"               "Gender"
##  [3] "Country"                 "Occupation"
##  [5] "self_employed"           "family_history"
##  [7] "treatment"               "Days_Indoors"
##  [9] "Growing_Stress"          "Changes_Habits"
## [11] "Mental_Health_History"   "Mood_Swings"
## [13] "Coping_Struggles"        "Work_Interest"
## [15] "Social_Weakness"         "mental_health_interview"
## [17] "care_options"
```

```
mental <- subset(mental, select = -c(Occupation,
                                     self_employed,
                                     family_history,
                                     Days_Indoors,
                                     Growing_Stress,
                                     Changes_Habits,
                                     Mental_Health_History,
                                     Mood_Swings,
                                     Coping_Struggles,
                                     Work_Interest,
                                     Social_Weakness,
                                     mental_health_interview,
                                     care_options))
# Remove rows missing data
mental <- mental[complete.cases(mental),]
# Check that the data has been trimmed
head(mental, 8)
```

```
##             Timestamp Gender       Country treatment
## 1 2014-08-27 11:29:31 Female United States       Yes
## 2 2014-08-27 11:31:50 Female United States       Yes
## 3 2014-08-27 11:32:39 Female United States       Yes
## 4 2014-08-27 11:37:59 Female United States       Yes
## 5 2014-08-27 11:43:36 Female United States       Yes
## 6 2014-08-27 11:49:51 Female        Poland       Yes
## 7 2014-08-27 11:51:34 Female     Australia       Yes
## 8 2014-08-27 11:52:41 Female United States        No
```

```
# Fifth dataset
head(prevalence)
```

```
##   index      Entity     Code Year
## 1     0    Abkhazia OWID_ABK 2015
## 2     1 Afghanistan      AFG 1990
## 3     2 Afghanistan      AFG 1991
## 4     3 Afghanistan      AFG 1992
## 5     4 Afghanistan      AFG 1993
## 6     5 Afghanistan      AFG 1994
##   Prevalence...Anxiety.disorders...Sex..Male...Age..Age.standardized..Percent.
## 1                                                                           NA
## 2                                                                     3.556843
## 3                                                                     3.548885
## 4                                                                     3.542779
## 5                                                                     3.538304
## 6                                                                     3.535309
##   Prevalence...Anxiety.disorders...Sex..Female...Age..Age.standardized..Percent.
## 1                                                                             NA
## 2                                                                       5.971172
## 3                                                                       5.980482
## 4                                                                       5.988175
## 5                                                                       5.993858
## 6                                                                       5.997363
##   Population..historical.estimates. Continent
## 1                                NA      Asia
## 2                          12412311
## 3                          13299016
## 4                          14485543
## 5                          15816601
## 6                          17075728
```

```
# Turn data into dataframe
prevalence <- as.data.frame(prevalence)
# Remove columns that we don't need
colnames(prevalence)
```

```
## [1] "index"
## [2] "Entity"
## [3] "Code"
## [4] "Year"
## [5] "Prevalence...Anxiety.disorders...Sex..Male...Age..Age.standardized..Percent."
## [6] "Prevalence...Anxiety.disorders...Sex..Female...Age..Age.standardized..Percent."
## [7] "Population..historical.estimates."
## [8] "Continent"
```

```
prevalence <- subset(prevalence, select = -c(Continent,
                                     Population..historical.estimates.))
# Rename columns
prevalence <- prevalence %>%
  rename(index = index, entity = Entity, code = Code, year = Year, male = Prevalence...Anxiety.disorders
female =
```

```r
Prevalence...Anxiety.disorders...Sex..Female...Age..Age.standardized..Percent.)
# Check changes
head(prevalence, 8)
```

```
##   index      entity     code year     male   female
## 1     0    Abkhazia OWID_ABK 2015       NA       NA
## 2     1 Afghanistan      AFG 1990 3.556843 5.971172
## 3     2 Afghanistan      AFG 1991 3.548885 5.980482
## 4     3 Afghanistan      AFG 1992 3.542779 5.988175
## 5     4 Afghanistan      AFG 1993 3.538304 5.993858
## 6     5 Afghanistan      AFG 1994 3.535309 5.997363
## 7     6 Afghanistan      AFG 1995 3.533797 5.998540
## 8     7 Afghanistan      AFG 1996 3.535415 5.996443
```

```r
# Sixth dataset
head(meds)
```

```
##   X       drug_name      date   age gender           time_on_drug reviewer_type
## 1 0 Sertraline Oral 5/12/2024 45-54 Female 1 to less than 2 years       Patient
## 2 1 Sertraline Oral 4/21/2024 35-44 Female      less than 1 month       Patient
## 3 2 Sertraline Oral 4/16/2024 25-34 Female 2 to less than 5 years       Patient
## 4 3 Sertraline Oral 4/11/2024 45-54   Male      less than 1 month       Patient
## 5 4 Sertraline Oral  4/8/2024 13-18 Female                              Patient
## 6 5 Sertraline Oral 3/29/2024 45-54 Female      less than 1 month       Patient
##                       condition rating_overall rating_effectiveness
## 1 Posttraumatic Stress Syndrome            5.0                    5
## 2                    Depression            1.0                    1
## 3  Repeated Episodes of Anxiety            4.3                    4
## 4                Panic Disorder            1.7                    1
## 5     Major Depressive Disorder            3.0                    2
## 6 Posttraumatic Stress Syndrome            2.3                    1
##   rating_ease_of_use rating_satisfaction
## 1                  5                   5
## 2                  1                   1
## 3                  4                   5
## 4                  3                   1
## 5                  4                   3
## 6                  5                   1
##
## 1
## 2
## 3
## 4 Of course, take this with a pinch of salt because everyone’s chemistry is different, and I am DEFI
## 5
## 6
```

```r
# Turn data into dataframe
meds <- as.data.frame(meds)
# Remove columns that we don't need
colnames(meds)
```

```
##  [1] "X"                    "drug_name"            "date"
##  [4] "age"                  "gender"               "time_on_drug"
##  [7] "reviewer_type"        "condition"            "rating_overall"
## [10] "rating_effectiveness" "rating_ease_of_use"   "rating_satisfaction"
## [13] "text"
```

```r
meds <- subset(meds, select = -c(time_on_drug, text))
# Find conditions that are anxiety related
unique(meds$condition)
```

```
##   [1] "Posttraumatic Stress Syndrome"
##   [2] "Depression"
##   [3] "Repeated Episodes of Anxiety"
##   [4] "Panic Disorder"
##   [5] "Major Depressive Disorder"
##   [6] "Bipolar Depression"
##   [7] "Depressed Mood Disorder Occurring Every Year at the Same Time"
##   [8] "Anxiousness associated with Depression"
##   [9] "Other"
##  [10] "Obsessive Compulsive Disorder"
##  [11] "Extreme Apprehension or Fear of Social Interaction"
##  [12] "Premenstrual Disorder with a State of Unhappiness"
##  [13] "Premature Ejection of Semen"
##  [14] "Problem Behavior"
##  [15] "Schizophrenia"
##  [16] "Osteoporosis"
##  [17] "Aggressive Behavior"
##  [18] "Binge Eating Disorder"
##  [19] "Acute Repetitive Seizures"
##  [20] "Anxious"
##  [21] "Muscle Spasm"
##  [22] "Neuropathic Pain"
##  [23] "Attention Deficit Disorder with Hyperactivity"
##  [24] "Stop Smoking"
##  [25] "Manic"
##  [26] "Convulsive Seizures"
##  [27] "Facial Nerve Pain"
##  [28] "Bipolar I Disorder with Most Recent Episode Mixed"
##  [29] "Nerve Pain"
##  [30] "Bipolar Disorder in Remission"
##  [31] "Epileptic Seizure"
##  [32] "Tonic"
##  [33] "Prevention of Seizures following Head Injury or Surgery"
##  [34] "Inducing of a Relaxed Easy State"
##  [35] "Sleep Disturbance with Extreme Anxiety"
##  [36] "Psychosis caused by Sudden Alcohol Withdrawal"
##  [37] "Seizure with Loss of Normal Tone or Strength"
##  [38] "Chronic Trouble Sleeping"
##  [39] "A Feeling of Restlessness with Inability to Sit Still"
##  [40] "Tension Headache"
##  [41] "Calming of Pediatric Patient by Administration of Sedative"
##  [42] "Symptoms from Stopping Treatment with Opioid Drugs"
##  [43] "Symptoms from Alcohol Withdrawal"
##  [44] "Brief Muscle Spasms in an Infant"
##  [45] "Mania associated with Bipolar Disorder"
##  [46] "Mental Disorder with Loss of Normal Personality"
##  [47] "Nausea and Vomiting"
##  [48] "Hiccups that are Hard to Cure"
##  [49] "Delirium"
##  [50] "Tourette"
##  [51] "Generalized Attack of Muscular Weakness"
##  [52] "Simple Partial Seizures"
##  [53] "Rapid Cycle Manic"
##  [54] "Additional Medications to Treat Depression"
##  [55] "Multiple Seizure Types"
##  [56] "Apprehension"
##  [57] "Combative and Explosive Behavior"
##  [58] "Schizophrenia With Mood Changes"
##  [59] "Psychosis caused by a Disease"
##  [60] "Pervasive Developmental Disorder"
##  [61] "A Rare Developmental Disorder of Infants"
##  [62] "Paroxysmal Choreoathetosis"
##  [63] "Nerve Pain of Tongue and Throat from 9th Cranial Nerve"
##  [64] "Extreme Discomfort in Calves when Sitting or Lying Down"
##  [65] ""
##  [66] "Nicotine Addiction"
##  [67] "Depression associated with Bipolar Disorder"
##  [68] "Nausea and Vomiting caused by Cancer Drugs"
##  [69] "Anxiety associated with an Operation"
##  [70] "Feeling Restless"
##  [71] "Additional Medication for Calming"
##  [72] "Additional Agent to Induce General Anesthesia"
##  [73] "Sedation with Ability to Respond to Stimulation or Speech"
##  [74] "Induce Temporary Amnesia"
##  [75] "Psychosis associated with Alzheimer"
##  [76] "Prevent Nausea and Vomiting from Cancer Chemotherapy"
##  [77] "Depression following Delivery of Baby"
##  [78] "Overweight"
##  [79] "Bulimia"
##  [80] "Anorexia Nervosa"
##  [81] "Muscle Weakness associated with Sleeping Disease"
##  [82] "Migraine Prevention"
##  [83] "Disorder characterized by Stiff"
##  [84] "Agitation associated with Schizophrenia"
##  [85] "Psychotic Depressive Illness"
##  [86] "Severe Anxiety"
##  [87] "Itching"
##  [88] "Severe Itching"
##  [89] "Life Threatening Allergic Reaction"
##  [90] "Additional Medications to Treat Pain"
##  [91] "Ventricular Premature Beats"
##  [92] "Heart Ventricle Rhythm Problem"
##  [93] "High Blood Pressure"
##  [94] "Feeling of Apprehension Before an Operation"
##  [95] "Involuntary Quivering"
##  [96] "Psychosis caused by a Drug"
##  [97] "Brief Episode of Schizophrenia with Rapid Onset"
##  [98] "Chronic Type of Schizophrenia"
##  [99] "Over Excitement"
## [100] "Allergic Conjunctivitis"
## [101] "Inflammation of the Nose due to an Allergy"
## [102] "Inflammation of Skin caused by an Allergy"
## [103] "Sneezing"
## [104] "Hives"
## [105] "Feel Like Throwing Up"
## [106] "Welt from Pressure on Skin"
## [107] "Cluster Headache Prevention"
## [108] "Seizures with Breaks in Consciousness"
## [109] "Petit Mal Seizures"
## [110] "Petit Mal Epilepsy with Multiple Seizure Types"
## [111] "Seizures with Irregular Muscle Contractions"
## [112] "Runny Nose"
## [113] "Additional Medication for Myoclonic Epilepsy"
## [114] "adjunct therapy for obsessive compulsive disorder"
## [115] "Frequent Headaches"
## [116] "Mitral Valve Prolapse Syndrome"
## [117] "Myocardial Reinfarction Prevention"
## [118] "Problems with Bladder Control"
## [119] "Bedwetting"
## [120] "Stuffy Nose"
## [121] "Essential Tremor"
## [122] "Irritable Colon"
## [123] "Nerve Pain after Herpes"
## [124] "Ulcer from Stomach Acid"
## [125] "Behaving with Excessive Cheerfulness and Activity"
## [126] "Recurring Sleep Episodes During the Day"
## [127] "Chronic Muscle or Bone Pain"
## [128] "Diabetic Complication causing Injury to some Body Nerves"
```

```r
# The following would qualify as something I don't know how to do
# that could be helpful to know.
# I don't know if there is a way to iterate through the unique items in the
# condition column and find conditions that are anxiety related using R.
# I would imagine it could be done through a loop and doing a partial match
# to condition which contains the letters "anx". I looked through the printed
# conditions myself and then coded the following lines.

# Remove rows that aren't anxiety related
meds <- subset(meds, condition == "Anxious" |
                 condition == "Severe Anxiety" |
                 condition == "Repeated Episodes of Anxiety")
# Remove rows that have missing values
meds <- meds[complete.cases(meds),]
# Check changes
head(meds, 8)
```

```
##     X       drug_name      date   age gender reviewer_type
## 3   2 Sertraline Oral 4/16/2024 25-34 Female       Patient
## 8   7 Sertraline Oral 3/26/2024  7-12   Male     Caregiver
## 21 20 Sertraline Oral 9/17/2023 25-34 Female       Patient
## 25 24 Sertraline Oral  9/5/2023 45-54 Female       Patient
## 29 28 Sertraline Oral  8/2/2023 55-64 Female       Patient
## 37 36 Sertraline Oral 3/20/2023 55-64   Male       Patient
## 56 55 Sertraline Oral 9/15/2022 35-44 Female       Patient
## 67 66 Sertraline Oral 7/13/2022 19-24 Female       Patient
##                         condition rating_overall rating_effectiveness
## 3  Repeated Episodes of Anxiety            4.3                    4
## 8  Repeated Episodes of Anxiety            5.0                    5
## 21 Repeated Episodes of Anxiety            5.0                    5
## 25 Repeated Episodes of Anxiety            1.0                    1
## 29 Repeated Episodes of Anxiety            1.3                    2
## 37 Repeated Episodes of Anxiety            1.3                    1
## 56 Repeated Episodes of Anxiety            4.3                    4
## 67 Repeated Episodes of Anxiety            1.7                    2
##    rating_ease_of_use rating_satisfaction
## 3                   4                   5
## 8                   5                   5
## 21                  5                   5
## 25                  1                   1
## 29                  1                   1
## 37                  2                   1
## 56                  5                   4
## 67                  2                   1
```

```
# Summarize all datasets
head(indicators_anxiety, 8)
```

```
##                        Indicator             Group         State
## 1 Symptoms of Anxiety Disorder National Estimate United States
## 2 Symptoms of Anxiety Disorder            By Age United States
## 3 Symptoms of Anxiety Disorder            By Age United States
## 4 Symptoms of Anxiety Disorder            By Age United States
## 5 Symptoms of Anxiety Disorder            By Age United States
## 6 Symptoms of Anxiety Disorder            By Age United States
## 7 Symptoms of Anxiety Disorder            By Age United States
## 8 Symptoms of Anxiety Disorder            By Age United States
##              Subgroup Phase Time.Period    Time.Period.Label
## 1      United States   1.0           1 Apr 23 - May 5, 2020
## 2      18 - 29 years   1.0           1 Apr 23 - May 5, 2020
## 3      30 - 39 years   1.0           1 Apr 23 - May 5, 2020
## 4      40 - 49 years   1.0           1 Apr 23 - May 5, 2020
## 5      50 - 59 years   1.0           1 Apr 23 - May 5, 2020
## 6      60 - 69 years   1.0           1 Apr 23 - May 5, 2020
## 7      70 - 79 years   1.0           1 Apr 23 - May 5, 2020
## 8 80 years and above   1.0           1 Apr 23 - May 5, 2020
##   Time.Period.Start.Date Time.Period.End.Date Value Low.CI High.CI
## 1             04/23/2020           05/05/2020  30.8   30.0    31.7
## 2             04/23/2020           05/05/2020  40.2   38.1    42.4
## 3             04/23/2020           05/05/2020  34.4   32.6    36.1
## 4             04/23/2020           05/05/2020  34.1   32.1    36.2
## 5             04/23/2020           05/05/2020  31.0   29.0    33.1
## 6             04/23/2020           05/05/2020  24.9   23.6    26.3
## 7             04/23/2020           05/05/2020  16.4   14.8    18.1
## 8             04/23/2020           05/05/2020  14.6   11.5    18.2
##   Confidence.Interval Quartile.Range
## 1         30.0 - 31.7
## 2         38.1 - 42.4
## 3         32.6 - 36.1
## 4         32.1 - 36.2
## 5         29.0 - 33.1
## 6         23.6 - 26.3
## 7         14.8 - 18.1
## 8         11.5 - 18.2
```

```
head(global_mental, 8)
```

```
##   index      Entity Code Year  anxiety
## 1     0 Afghanistan  AFG 1990 4.828830
## 2     1 Afghanistan  AFG 1991 4.829740
## 3     2 Afghanistan  AFG 1992 4.831108
## 4     3 Afghanistan  AFG 1993 4.830864
## 5     4 Afghanistan  AFG 1994 4.829423
## 6     5 Afghanistan  AFG 1995 4.828337
## 7     6 Afghanistan  AFG 1996 4.828083
## 8     7 Afghanistan  AFG 1997 4.827726
```

```
head(mental, 8)
```

```
##              Timestamp Gender       Country treatment
## 1 2014-08-27 11:29:31 Female United States       Yes
## 2 2014-08-27 11:31:50 Female United States       Yes
## 3 2014-08-27 11:32:39 Female United States       Yes
## 4 2014-08-27 11:37:59 Female United States       Yes
## 5 2014-08-27 11:43:36 Female United States       Yes
## 6 2014-08-27 11:49:51 Female        Poland       Yes
## 7 2014-08-27 11:51:34 Female     Australia       Yes
## 8 2014-08-27 11:52:41 Female United States        No
```

```
head(prevalence, 8)
```

```
##   index      entity     code year     male   female
## 1     0    Abkhazia OWID_ABK 2015       NA       NA
## 2     1 Afghanistan      AFG 1990 3.556843 5.971172
## 3     2 Afghanistan      AFG 1991 3.548885 5.980482
## 4     3 Afghanistan      AFG 1992 3.542779 5.988175
## 5     4 Afghanistan      AFG 1993 3.538304 5.993858
## 6     5 Afghanistan      AFG 1994 3.535309 5.997363
## 7     6 Afghanistan      AFG 1995 3.533797 5.998540
## 8     7 Afghanistan      AFG 1996 3.535415 5.996443
```

```
head(meds, 8)
```

```
##     X       drug_name      date   age gender reviewer_type
## 3   2 Sertraline Oral 4/16/2024 25-34 Female       Patient
## 8   7 Sertraline Oral 3/26/2024  7-12   Male     Caregiver
## 21 20 Sertraline Oral 9/17/2023 25-34 Female       Patient
## 25 24 Sertraline Oral  9/5/2023 45-54 Female       Patient
## 29 28 Sertraline Oral  8/2/2023 55-64 Female       Patient
## 37 36 Sertraline Oral 3/20/2023 55-64   Male       Patient
## 56 55 Sertraline Oral 9/15/2022 35-44 Female       Patient
## 67 66 Sertraline Oral 7/13/2022 19-24 Female       Patient
##                       condition rating_overall rating_effectiveness
## 3  Repeated Episodes of Anxiety            4.3                    4
## 8  Repeated Episodes of Anxiety            5.0                    5
## 21 Repeated Episodes of Anxiety            5.0                    5
## 25 Repeated Episodes of Anxiety            1.0                    1
## 29 Repeated Episodes of Anxiety            1.3                    2
## 37 Repeated Episodes of Anxiety            1.3                    1
## 56 Repeated Episodes of Anxiety            4.3                    4
## 67 Repeated Episodes of Anxiety            1.7                    2
##    rating_ease_of_use rating_satisfaction
## 3                    4                   5
## 8                    5                   5
## 21                   5                   5
## 25                   1                   1
## 29                   1                   1
## 37                   2                   1
## 56                   5                   4
## 67                   2                   1
```

Data Importing and Cleaning

The first step to take is going to be importing all of the data from the csv files. After that it will be to parse the data in order to condense the datasets down to the appropriate columns. Once the data has been parsed we will remove rows with missing data to create datasets that have all the necessary columns. The code in the above section prints the kept columns and first eight rows of each dataset.

Exceptions

Unfortunately, the dataset that has the symptoms isn't very useable. I don't believe this first dataset is actually useful for me. There's lots of data but not useful for this particalur analysis.

Data Discovery

The data from the indicators dataset seems very promising. It breaks down the data by date, age group, ethnicity, gender, education level, and state. We can group the data by these categories throughout the timeframes that all the categories span. We should then make plots over time for all of these values. We can then look to the global_mental dataset and find the prevalence per country. In that prevalence dataset we can average the prevalence between male and female and compare that to the results from the indicators and global_mental datasets. We can then compare the prevalence by male or female from the indicator groups, and compare that to the mental and prevalence datasets sorted by gender. We should average all the results for female and male from all of the datasets. We can then look at the treatment dataset from WebMD and see how many people are receiving treatment and what medications they are taking. This can also be broken down by gender for more information. To answer our research questions we look at the prevalence of anxiety per location, gender, race, and age. Then we can look at the effectiveness of medication for each group.

Variables and Data Preparation

I think that the datasets have all the appropriate variables aside from the cumulative female and male results. We can compare all of our results to the NIH and WHO findings. And all of the results can be used in comparisons plots for all of the demographic categories.

Machine Learning

I'm not sure how to incorporate machine learning currently. I've removed some categories from the datasets that could potentially be useful for making models that predict anxiety from symptoms, demographics, and comorbidities. If I could make this model I think machine learning would be useful in order to look at more datasets.

For the Future

I think making models and using machine learning to look at new data would be something I'd really like to do. There are some ethical and privacy concerns with doing this though.

Things to Learn

The first dataset has a text field that I'm unsure how to parse for symptoms in R. The way I would approach this would be to parse the whole text field down into individual words. I'd then compare the words individually to a list of symptoms I defined. Alternatively, I'd probably need AI to parse the text field to find all the symptoms. That may be the better way to search through the text field because doing it myself does not account for if someone says these are symptoms they're not experiencing. The string comparison method only looks for the presence of a word not the context. I'd also like to know how to print nice tables of data both in R and to files from R.

Week 10 Ex. 10.3

For this final analysis I've edited the data and made plots and a table. The actual analysis can be found below.

```r
# Import necessary packages
# install.packages("tidyverse")
library(tidyverse)
library(dplyr)

#Import Dataset 2
indicators_anxiety <-
  read.csv(file = 'Indicators_of_Anxiety_or_Depression.csv', header = TRUE,
           sep =",", stringsAsFactors = FALSE)
# Turn data into a df
indicators_anxiety <- as.data.frame(indicators_anxiety)
# Find unique indicators
unique(indicators_anxiety$Indicator)
```

```
## [1] "Symptoms of Depressive Disorder"
## [2] "Symptoms of Anxiety Disorder"
## [3] "Symptoms of Anxiety Disorder or Depressive Disorder"
```

```r
# Filter for only anxiety not depression
indicators_anxiety <- filter(indicators_anxiety, Indicator ==
                               'Symptoms of Anxiety Disorder')
# Remove missing data
indicators_anxiety <- na.omit(indicators_anxiety)
# Check the phases of the study
unique(indicators_anxiety$Phase)
```

```
##  [1] "1.0"                      "2.0"                      "3.0 (Oct 28 - Dec 21)"
##  [4] "3.0 (Jan 6 - Mar 29)"     "3.1"                      "3.2"
##  [7] "3.3"                      "3.4"                      "3.5"
## [10] "3.6"                      "3.7"                      "3.8"
## [13] "3.9"                      "3.10"                     "4.0"
## [16] "4.1"
```

```r
# Change phases to simple numbers
indicators_anxiety <- indicators_anxiety %>%
  mutate(Phase = recode(Phase, '-1' = '0', '3.0 (Oct 28 - Dec 21)'
                        = '3.0', '3.0 (Jan 6 - Mar 29)' = '3.0'))
indicators_phase_1 = subset(indicators_anxiety, Phase == '1.0')
indicators_val_1 <- mean(indicators_phase_1$Value)
indicators_phase_2 = subset(indicators_anxiety, Phase == '2.0')
indicators_val_2 <- mean(indicators_phase_2$Value)
indicators_phase_3 = subset(indicators_anxiety, Phase == '3.0')
indicators_val_3 <- mean(indicators_phase_3$Value)
indicators_phase_3_1 = subset(indicators_anxiety, Phase == '3.1')
indicators_val_3_1 <- mean(indicators_phase_3_1$Value)
indicators_phase_3_2 = subset(indicators_anxiety, Phase == '3.2')
indicators_val_3_2 <- mean(indicators_phase_3_2$Value)
indicators_phase_3_3 = subset(indicators_anxiety, Phase == '3.3')
indicators_val_3_3 <- mean(indicators_phase_3_3$Value)
indicators_phase_3_4 = subset(indicators_anxiety, Phase == '3.4')
indicators_val_3_4 <- mean(indicators_phase_3_4$Value)
indicators_phase_3_5 = subset(indicators_anxiety, Phase == '3.5')
indicators_val_3_5 <- mean(indicators_phase_3_5$Value)
indicators_phase_3_6 = subset(indicators_anxiety, Phase == '3.6')
indicators_val_3_6 <- mean(indicators_phase_3_6$Value)
indicators_phase_3_7 = subset(indicators_anxiety, Phase == '3.7')
indicators_val_3_7 <- mean(indicators_phase_3_7$Value)
indicators_phase_3_8 = subset(indicators_anxiety, Phase == '3.8')
indicators_val_3_8 <- mean(indicators_phase_3_8$Value)
indicators_phase_3_9 = subset(indicators_anxiety, Phase == '3.9')
indicators_val_3_9 <- mean(indicators_phase_3_9$Value)
indicators_phase_3_10 = subset(indicators_anxiety, Phase == '3.10')
indicators_val_3_10 <- mean(indicators_phase_3_10$Value)
indicators_phase_4 = subset(indicators_anxiety, Phase == '4.0')
indicators_val_4 <- mean(indicators_phase_4$Value)
indicators_phase_4_1 = subset(indicators_anxiety, Phase == '4.1')
indicators_val_4_1 <- mean(indicators_phase_4_1$Value)
indicators_mean_val <- c(indicators_val_1, indicators_val_2,
                         indicators_val_3, indicators_val_3_1,
                         indicators_val_3_2, indicators_val_3_3,
                         indicators_val_3_4, indicators_val_3_5,
                         indicators_val_3_6, indicators_val_3_7,
                         indicators_val_3_8, indicators_val_3_9,
                         indicators_val_3_10, indicators_val_4,
                         indicators_val_4_1)
indicators_mean_val <- round(indicators_mean_val, 2)
phases <- unique(indicators_anxiety$Phase)
indicators <- cbind(phases, indicators_mean_val)
indicators <- as.data.frame(indicators)


# According to the phases of the study there is a general downward trend as the phases progress.
# In 2024, which are phases 4.0 and 4.1 there is the smallest average value from the dataset.

# Import dataset 3
global_mental <-
  read.csv(file = 'Mental Health Data Global.csv', header = TRUE, sep =",",
           stringsAsFactors = FALSE)
# Turn data into a df
global_mental <- as.data.frame(global_mental)
# Remove columns that we don't need
colnames(global_mental)
```

```
##  [1] "index"                     "Entity"
##  [3] "Code"                      "Year"
##  [5] "Schizophrenia...."         "Bipolar.disorder...."
##  [7] "Eating.disorders...."      "Anxiety.disorders...."
##  [9] "Drug.use.disorders...."    "Depression...."
## [11] "Alcohol.use.disorders...."
```

```
global_mental <- subset(global_mental, select = -c(Schizophrenia....,
                                                   Bipolar.disorder....,
                                                   Eating.disorders....,
                                                   Drug.use.disorders....,
                                                   Depression....,
                                                   Alcohol.use.disorders....))
# Remove rows missing data
global_mental <- global_mental[complete.cases(global_mental),]
# Rename
global_mental <- global_mental %>%
  rename(anxiety = Anxiety.disorders....)
# Check unique locations
# unique(global_mental$Entity)
# Subset to US for this analysis
global_mental_us <- subset(global_mental, Entity == "United States")
# Subset to US for this analysis
global_mental_japan <- subset(global_mental, Entity == "Japan")

# According to the global_mental dataset there is a downward trend
# of people experiencing anxiety from 2007 on.
# This seems odd given the recession starting in that timeframe.
# There may not be a strong correlation between financial
# stressors and anxiety.
# Import dataset 4
mental <-
  read.csv(file = 'Mental Health Dataset.csv', header = TRUE, sep =",", stringsAsFactors = FALSE)
# Turn data into dataframe
mental <- as.data.frame(mental)
# Remove columns that we don't need
colnames(mental)
```

```
##  [1] "Timestamp"               "Gender"
##  [3] "Country"                 "Occupation"
##  [5] "self_employed"           "family_history"
##  [7] "treatment"               "Days_Indoors"
##  [9] "Growing_Stress"          "Changes_Habits"
## [11] "Mental_Health_History"   "Mood_Swings"
## [13] "Coping_Struggles"        "Work_Interest"
## [15] "Social_Weakness"         "mental_health_interview"
## [17] "care_options"
```

```
mental <- subset(mental, select = -c(Occupation, self_employed,
                                     family_history, Days_Indoors,
                                     Growing_Stress, Changes_Habits,
                                     Mental_Health_History, Mood_Swings,
                                     Coping_Struggles, Work_Interest,
                                     Social_Weakness,
                                     mental_health_interview,
                                     care_options))
# Limit to US
mental <- subset(mental, Country == "United States")
# Remove rows missing data
mental <- mental[complete.cases(mental),]
# Female to male anxiety percentage
mental$Timestamp <- substr(mental$Timestamp, 1, 4)
unique(mental$Timestamp)
```

```
## [1] "2014" "2015" "2016"
```

```
# Treatment vs. No Treatment 2014
mental_total_no_treatment_male <- nrow(mental[mental$Timestamp == '2014' &
                                                mental$Gender == 'Male' &
                                                mental$treatment == 'No',])
mental_total_no_treatment_fem <- nrow(mental[mental$Timestamp == '2014' &
                                               mental$Gender == 'Female' &
                                               mental$treatment == 'No',])
mental_total_treatment_male <- nrow(mental[mental$Timestamp == '2014' &
                                             mental$Gender == 'Male' &
                                             mental$treatment == 'Yes',])
mental_total_treatment_fem <- nrow(mental[mental$Timestamp == '2014' &
                                            mental$Gender == 'Female' &
                                            mental$treatment == 'Yes',])
mental_total_female <- mental_total_treatment_fem +
  mental_total_no_treatment_fem
mental_total_male <- mental_total_treatment_male +
  mental_total_no_treatment_male
percent_fem_treatment <- (mental_total_treatment_fem/(mental_total_treatment_fem +
                                                        mental_total_no_treatment_fem))*100
percent_male_treatment <- (mental_total_treatment_male/(mental_total_treatment_male +
                                                          mental_total_no_treatment_male))*100
total_people <- mental_total_treatment_male + mental_total_no_treatment_male +
  mental_total_treatment_fem + mental_total_no_treatment_fem
total_treatment <- (mental_total_treatment_male + mental_total_treatment_fem)
total_no_treatment <- total_people - total_treatment
percent_total_treatment <- (total_treatment/total_people)*100
# Make a dataframe of the data
male <- c(format(round(mental_total_male, 0)),
          format(round(mental_total_no_treatment_male, 0)),
          format(round(mental_total_treatment_male, 0)),
          format(round(percent_male_treatment, 2)))
female <- c(format(round(mental_total_female, 0)),
            format(round(mental_total_no_treatment_fem,0)),
            format(round(mental_total_treatment_fem, 0)),
            format(round(percent_fem_treatment, 2)))
total <- c(format(round(total_people, 0)), format(round(total_no_treatment, 0)),
           format(round(total_treatment, 0)), format(round(percent_total_treatment, 2)))
tab_tot_fem_male <- cbind(total, male, female)
colnames(tab_tot_fem_male) <- c('Total', 'Male', 'Female')
rownames(tab_tot_fem_male) <- c('Total People', 'No Treatment', 'Treatment',
                                'Percentage Receiving Treatment')

####MUCH SMALLER DATASET NOT USED
## Female to male anxiety percentage 2015
# mental_total_15 <- nrow(mental[mental$Timestamp == '2015' &
# mental$treatment == 'Yes',])
# mental_total_fem_15 <- nrow(mental[mental$Timestamp == '2015' &
# mental$treatment == 'Yes' & mental$Gender == 'Female',])
# mental_total_male_15 <- nrow(mental[mental$Timestamp == '2015' &
# mental$treatment == 'Yes' & mental$Gender == 'Male',])
# mental_total_15
# mental_total_fem_15
# mental_total_male_15
# percent_tot_fem_15 <- (mental_total_fem_15/mental_total_15)*100
# percent_tot_fem_15
# percent_tot_male_15 <- (mental_total_male_15/mental_total_15)*100
# percent_tot_male_15

####MUCH SMALLER DATASET NOT USED
## Female to male anxiety percentage 2016
# mental_total_16 <- nrow(mental[mental$Timestamp == '2016' &
# mental$treatment == 'Yes',])
# mental_total_fem_16 <- nrow(mental[mental$Timestamp == '2016' &
# mental$treatment == 'Yes' & mental$Gender == 'Female',])
# mental_total_male_16 <- nrow(mental[mental$Timestamp == '2016' &
# mental$treatment == 'Yes' & mental$Gender == 'Male',])
# mental_total_16
# mental_total_fem_16
# mental_total_male_16
# percent_tot_fem_16 <- (mental_total_fem_16/mental_total_16)*100
# percent_tot_fem_16
# percent_tot_male_16 <- (mental_total_male_16/mental_total_16)*100
# percent_tot_male_16

# Import datatset 5
prevalence <-
  read.csv(file = 'prevalence-of-anxiety-disorders-males-vs-females.csv', header = TRUE,
           sep =",", stringsAsFactors = FALSE)
# Turn data into dataframe
prevalence <- as.data.frame(prevalence)
# Remove columns that we don't need
colnames(prevalence)
```

```
## [1] "index"
## [2] "Entity"
## [3] "Code"
## [4] "Year"
## [5] "Prevalence...Anxiety.disorders...Sex..Male...Age..Age.standardized..Percent."
## [6] "Prevalence...Anxiety.disorders...Sex..Female...Age..Age.standardized..Percent."
## [7] "Population..historical.estimates."
## [8] "Continent"
```

```
prevalence <- subset(prevalence, select = -c(Continent, Population..historical.estimates.))
# Rename columns
prevalence <- prevalence %>%
  rename(index = index, entity = Entity, code = Code, year = Year,
         male = Prevalence...Anxiety.disorders...Sex..Male...Age..Age.standardized..Percent.,
     female = Prevalence...Anxiety.disorders...Sex..Female...Age..Age.standardized..Percent.)
# Make changes
# unique(prevalence$entity)
prevalence <- filter(prevalence, year > 1990)
# Subset to Japan
prevalence_japan <-subset(prevalence, entity == "Japan")
# Subset to just the US
prevalence_us <- subset(prevalence, entity == "United States")


# Import dataset 6
meds <-
  read.csv(file = 'psychiatric_drug_webmd_reviews.csv', header = TRUE, sep =",",
           stringsAsFactors = FALSE)
# Sixth dataset
head(meds)
```

```
##   X       drug_name      date   age gender           time_on_drug reviewer_type
## 1 0 Sertraline Oral 5/12/2024 45-54 Female 1 to less than 2 years       Patient
## 2 1 Sertraline Oral 4/21/2024 35-44 Female      less than 1 month       Patient
## 3 2 Sertraline Oral 4/16/2024 25-34 Female 2 to less than 5 years       Patient
## 4 3 Sertraline Oral 4/11/2024 45-54   Male      less than 1 month       Patient
## 5 4 Sertraline Oral  4/8/2024 13-18 Female                              Patient
## 6 5 Sertraline Oral 3/29/2024 45-54 Female      less than 1 month       Patient
##                       condition rating_overall rating_effectiveness
## 1 Posttraumatic Stress Syndrome            5.0                    5
## 2                    Depression            1.0                    1
## 3  Repeated Episodes of Anxiety            4.3                    4
## 4                Panic Disorder            1.7                    1
## 5     Major Depressive Disorder            3.0                    2
## 6 Posttraumatic Stress Syndrome            2.3                    1
##   rating_ease_of_use rating_satisfaction
## 1                  5                   5
## 2                  1                   1
## 3                  4                   5
## 4                  3                   1
## 5                  4                   3
## 6                  5                   1
##
## 1
## 2
## 3
## 4 Of course, take this with a pinch of salt because everyone's chemistry is different, and I am DEFI
## 5
## 6
```

```
# Turn data into dataframe
meds <- as.data.frame(meds)
# Remove columns that we don't need
colnames(meds)
```

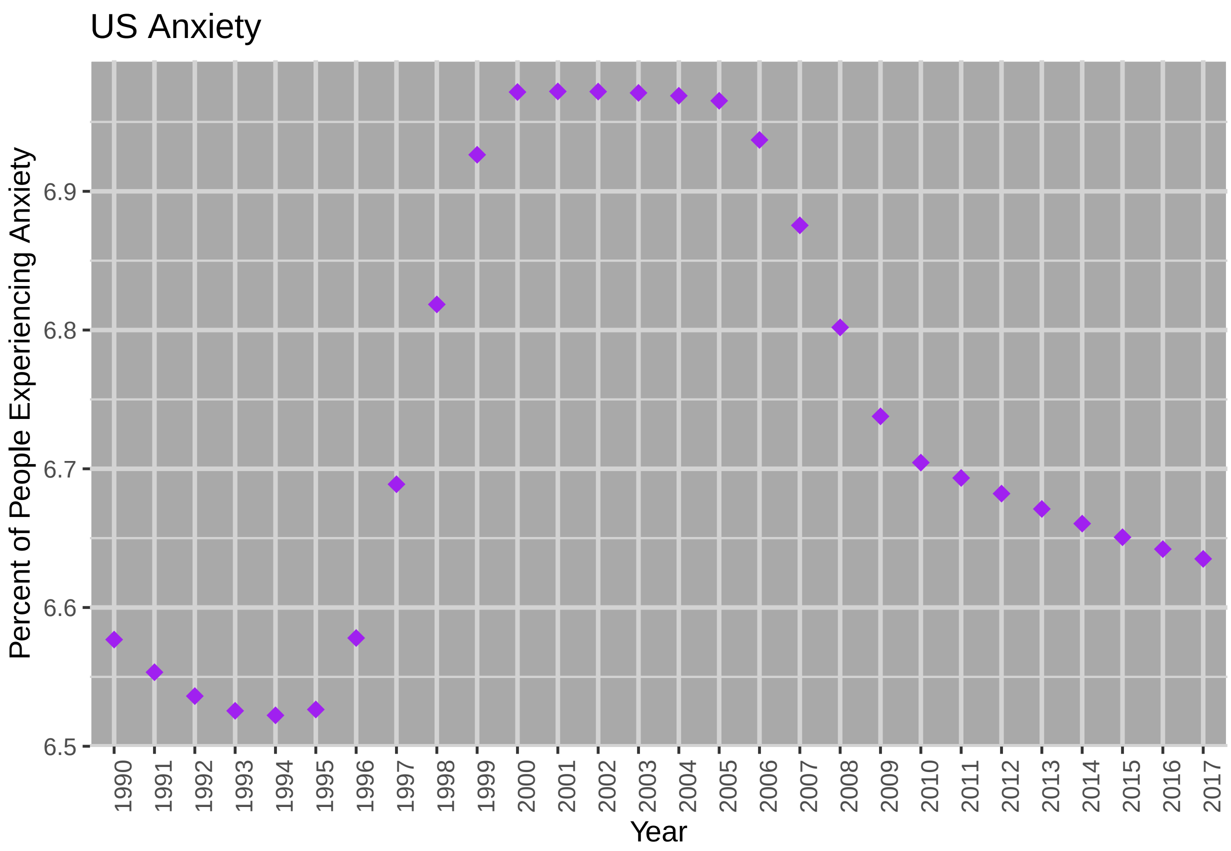
```
##  [1] "X"                    "drug_name"            "date"
##  [4] "age"                  "gender"               "time_on_drug"
##  [7] "reviewer_type"        "condition"            "rating_overall"
## [10] "rating_effectiveness" "rating_ease_of_use"   "rating_satisfaction"
## [13] "text"
```

```
meds <- subset(meds, select = -c(time_on_drug, text))
# Find conditions that are anxiety related
# unique(meds$condition)
# The following would qualify as something I don't know how to do that
# could be helpful to know.
# I don't know if there is a way to iterate through the unique items in
# the condition column and find conditions that are anxiety related using R.
# I would imagine it could be done through a loop and doing a partial match
# to condition which contains the letters "anx".
# I looked through the printed conditions myself and then coded the following lines
# unique(meds$drug_name)
# Remove rows that aren't anxiety related
meds <- subset(meds, condition == "Anxious" | condition == "Severe Anxiety" |
                 condition == "Repeated Episodes of Anxiety")
# Remove rows that have missing values
meds <- meds[complete.cases(meds),]
```

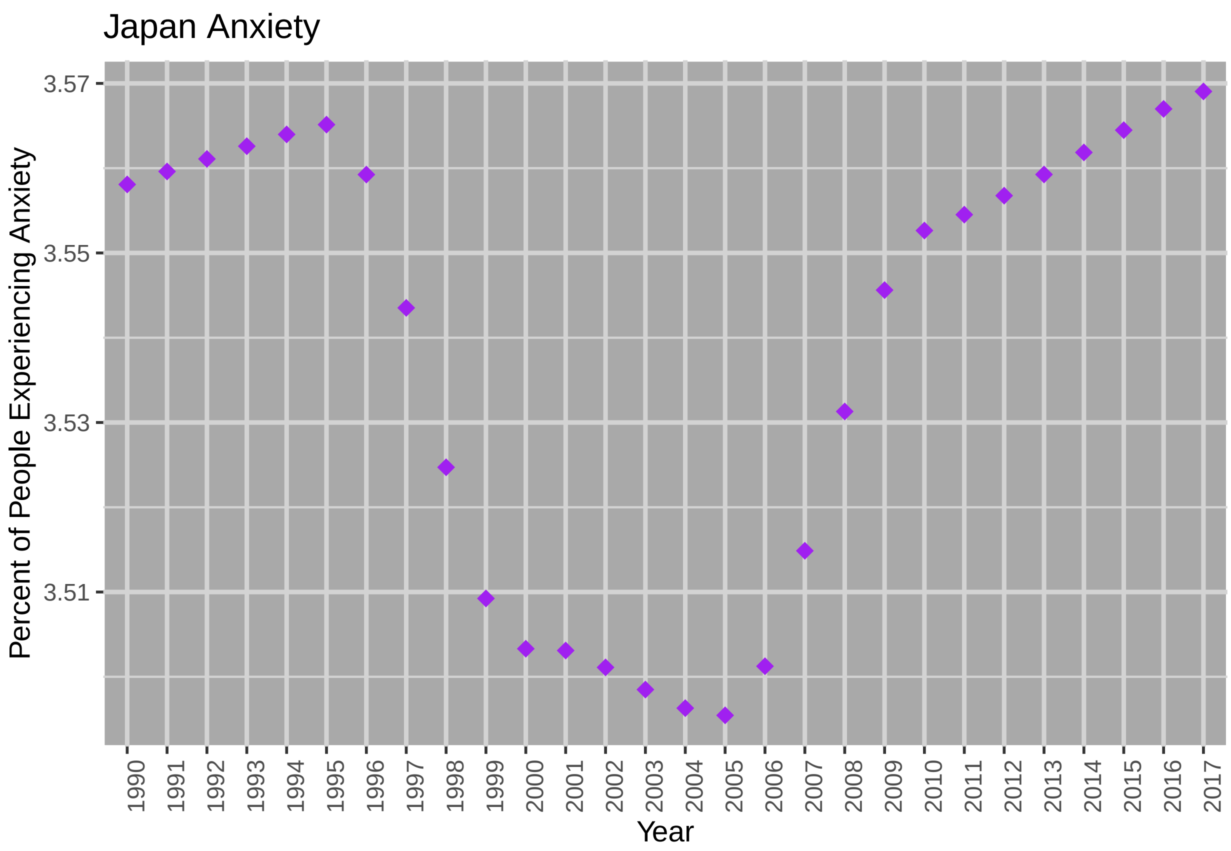
Introduction

The world is experiencing a change in how it views mental health's importance. Some companies now offer mental health days, sick days, and paid time off. The addition of mental health days is an implied acknowledgment that mental health is impactful in the day-to-day lives of the average person. The stigma of having a mental illness is changing, particularly among the younger generations. Many people in the millennials and Gen Z age groups are much more open about having mental health problems, seeking therapy, and potentially medications or other treatment options. According to the World Health Organization, the estimated prevalence of anxiety in the world population is 4% (WHO, 2023). This estimate makes anxiety the most common of all mental health disorders (WHO, 2023). The estimated 4% is likely lower than the actual number as the stigma reduces the number of people seeking help. The WHO also notes that only 1 in 4 people with anxiety receive treatment (WHO, 2023). I believe it is also culturally dependent on how much of a role the stigma plays in people reporting their illness and seeking help. A brief exploration of the cultural effect of stigma is conducted by analyzing the chosen datasets. I believe culture plays a significant role, as do other demographics of patients. The NIH has a published study from 2010 that contradicts the numbers from the WHO (NIH, 2010). The NIH study estimated that 31.1% of U.S. adults experienced an anxiety disorder at some point in their lives and that 19.1% of U.S. adults experienced an anxiety disorder in the past year (NIH, 2010). The NIH study also acknowledges that there is a difference in the prevalence of anxiety disorders between male and female patients, with females affected at a rate of 23.4% and males at 14.3% (NIH, 2010). ## The problem statement you addressed. The original intent of this analysis was to compare the findings from the chosen datasets with those of the WHO and NIH studies. While working through comparing the Kaggle datasets with the WHO and NIH, an additional idea presented itself. The basic idea was to explore the ability of machine learning to make predictions as an excellent opportunity for improving mental health screenings. This report compares findings from the chosen datasets with the results from the WHO and NIH studies. Additionally, this report explores the potential of using machine learning to ascertain if someone may be experiencing an anxiety disorder, if they need treatment, and what treatment may be best for them. ## How you addressed this problem statement In order to ascertain if there has been a change in the prevalence or diagnosis of anxiety disorders since the publishing of the WHO and NIH studies, the datasets were edited, condensed, or expanded to retain the relevant information, and the data was plotted for a more straightforward interpretation. ## Analysis The dataset "Mental Health Data Global" from Kaggle spans 1990 - 2017 (Kaggle, 2024). The percentage of people experiencing anxiety in the U.S. in this Global dataset ranges from 6.5% - 7%. The percentage of people experiencing anxiety in Japan in this dataset ranges from 3.5% - 3.6%. The U.S. range is higher than the WHO estimate of 4%, while Japan is lower than the WHO estimate (WHO, 2023). The finding that the U.S. has a higher percentage of people experiencing anxiety than Japan is also present in the Kaggle dataset, "Prevalence of Anxiety Disorders Males vs. Females" (Kaggle, 2024). The prevalence dataset has the U.S. range of people experiencing anxiety at 11% - 14.5% and Japan coming in much lower at 5% - 5.7%. The data from the Prevalence dataset shows that both the people of Japan and the U.S. are experiencing anxiety at a rate more significant than the published 4% from the WHO (WHO, 2023). However, the Global and Prevalence datasets come in lower than the NIH estimate of 19.1% (NIH, 2010). I believe that comparing the U.S. and Japan in both these datasets supports the idea that there is a significant impact of culture on the number of people seeking help or treatment for their anxiety.

```r
# Plot % of people experiencing anxiety US
global_plot_us <- ggplot(global_mental_us, aes(Year, anxiety))
global_plot_us + geom_point(shape = 18, size = 2.8, color = "purple") +
  theme(panel.grid = element_line(color = "lightgrey", linewidth = 0.8, linetype = 1),
  panel.background = element_rect(color = "white", fill = "darkgrey")) +
  labs(title = "US Anxiety", x ="Year",
       y = "Percent of People Experiencing Anxiety") +
  theme(axis.text.x = element_text(angle = 90))
```

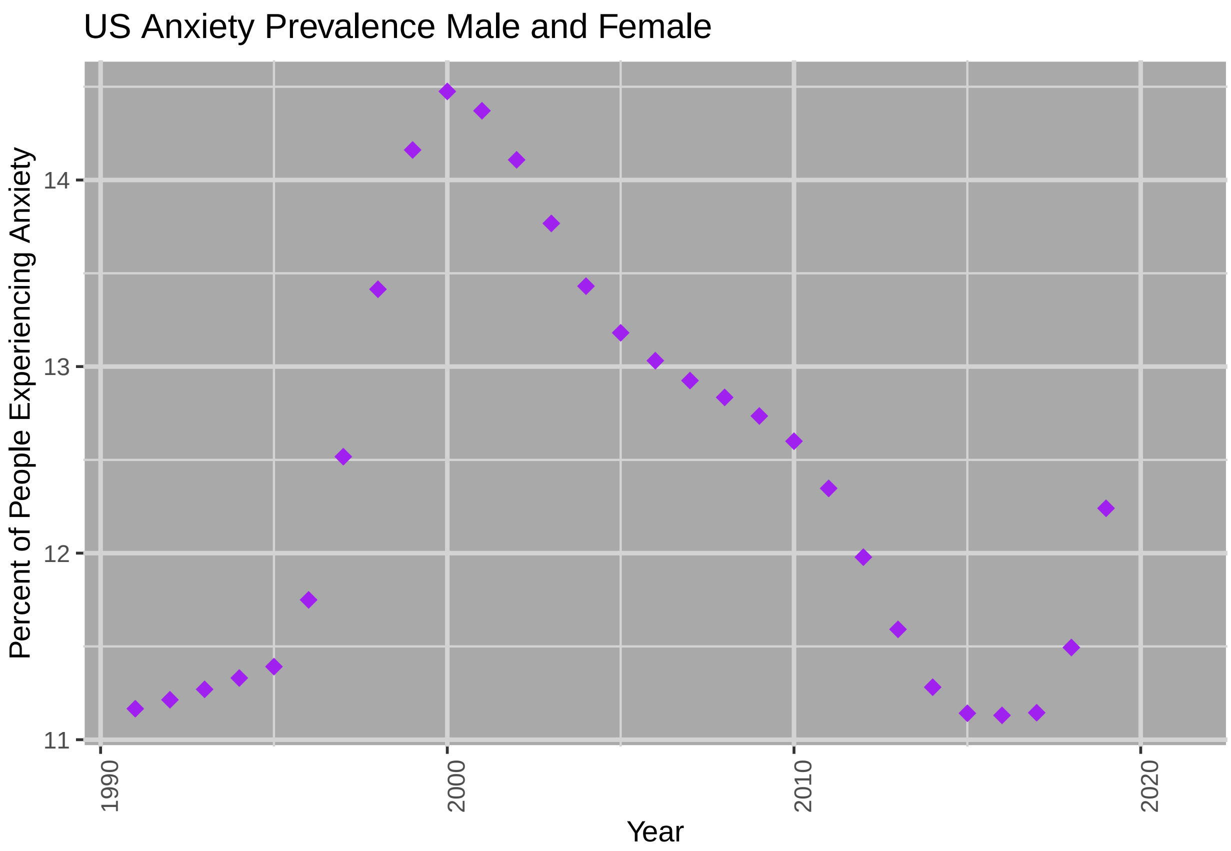


```
# Plot % of people experiencing anxiety Japan
global_plot_japan <- ggplot(global_mental_japan, aes(Year, anxiety))
global_plot_japan + geom_point(shape = 18, size = 2.8, color = "purple") +
  theme(panel.grid = element_line(color = "lightgrey", linewidth = 0.8, linetype = 1),
  panel.background = element_rect(color = "white", fill = "darkgrey")) +
  labs(title = "Japan Anxiety", x ="Year",
       y = "Percent of People Experiencing Anxiety") +
  theme(axis.text.x = element_text(angle = 90))
```



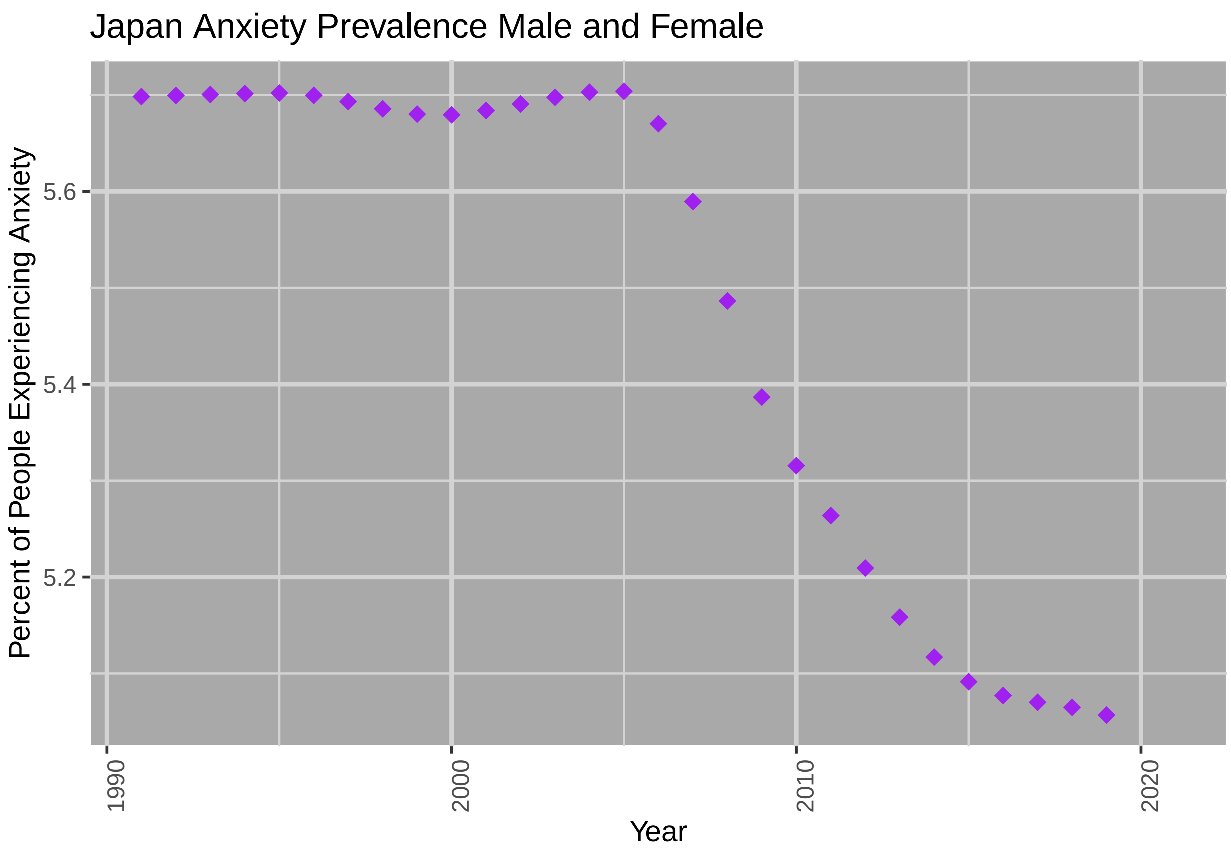
```
# Plot % of people experiencing anxiety US
prevalence_plot_us <- ggplot(prevalence_us, aes(year, male + female))
prevalence_plot_us + geom_point(shape = 18, size = 2.8, color = "purple") +
  theme(panel.grid = element_line(color = "lightgrey", linewidth = 0.8, linetype = 1),
   panel.background = element_rect(color = "white", fill = "darkgrey")) +
  labs(title = "US Anxiety Prevalence Male and Female", x ="Year",
       y = "Percent of People Experiencing Anxiety") +
  theme(axis.text.x = element_text(angle = 90))
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



```r
# Plot % of people experiencing anxiety Japan
prevalence_plot_japan <- ggplot(prevalence_japan, aes(year, male + female))
prevalence_plot_japan + geom_point(shape = 18, size = 2.8, color = "purple") +
  theme(panel.grid = element_line(color = "lightgrey", linewidth = 0.8, linetype = 1),
   panel.background = element_rect(color = "white", fill = "darkgrey")) +
  labs(title = "Japan Anxiety Prevalence Male and Female", x ="Year",
       y = "Percent of People Experiencing Anxiety") +
  theme(axis.text.x = element_text(angle = 90))
```

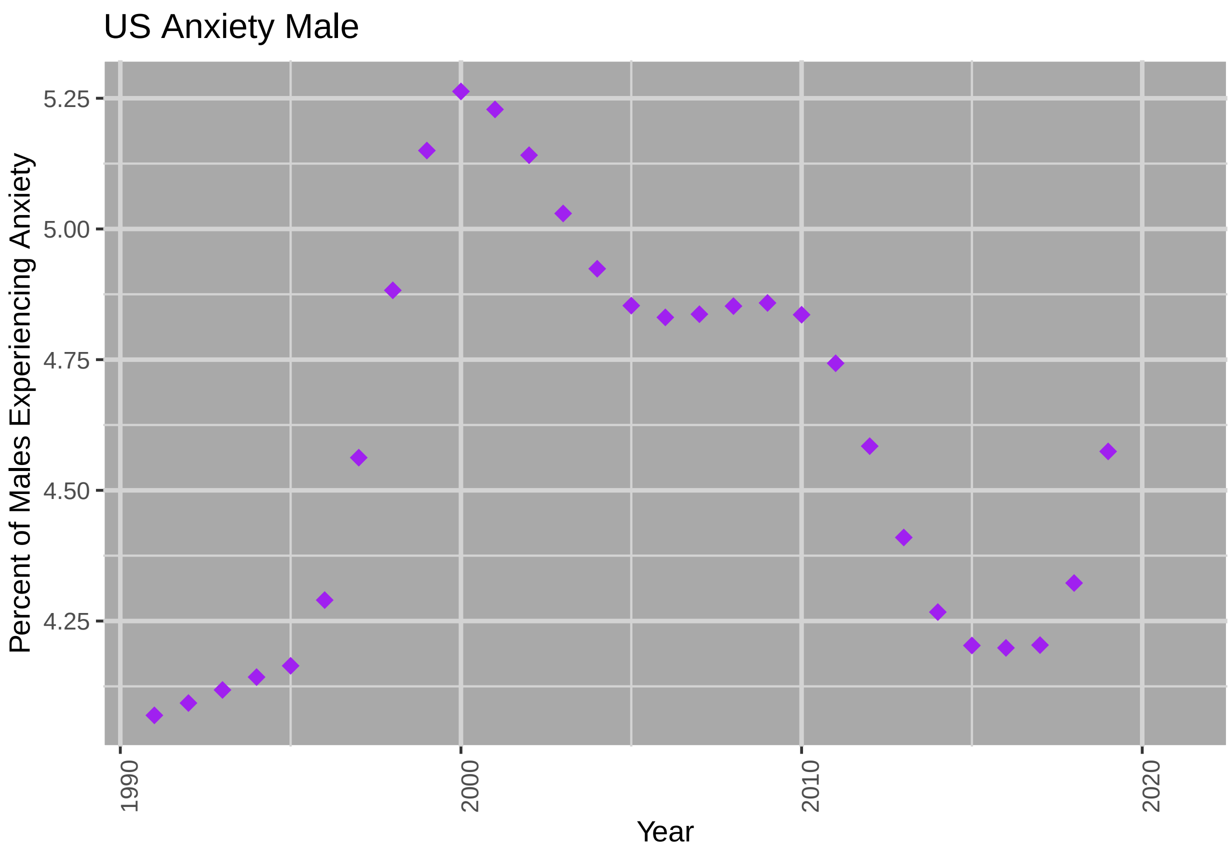
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



Seeing the effect of gender on the number of people experiencing or seeking help for anxiety from the Prevalence dataset, has results that fall in line with the WHO study (WHO, 2023). In both the U.S. and Japan, women experience or report anxiety at a higher rate than their male counterparts. In the U.S., the range for men is from 4% to 5.25%, and for women it is from 6.9% to 9.5%. In Japan, men report experiencing anxiety at a rate of 2.15 % to 2.35% and women from 2.9% to 3.35%. The discrepancy between men and women could be an example of how the stigma of having a mental illness is affected by the idea of masculinity or femininity. The idea that for a man to be masculine, they must not worry or show weakness is a troubling social norm. According to the CDC, men commit suicide at a rate nearly four times that of women in the U.S. (AFSP, 2024). This suicide rate is a troubling statistic that I believe is supportive of the idea that perceived masculinity is diminished if a man has a mental illness. I believe there is a strong correlation between perceived masculinity and men reporting their mental health struggles, which could help explain why the data shows men reporting anxiety at a rate lower than women.

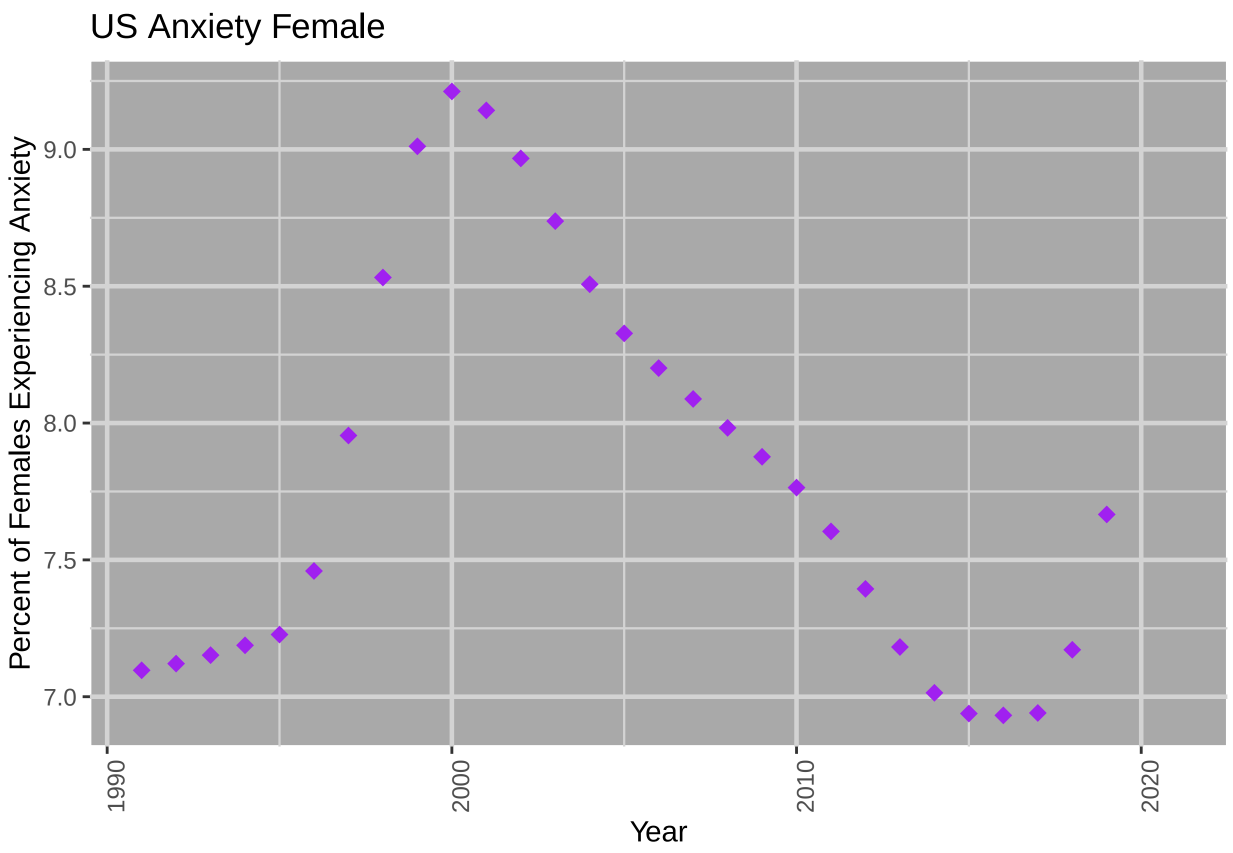
```
    # Plot % of people experiencing anxiety male
prevalence_plot_male_us <- ggplot(prevalence_us, aes(year, male))
prevalence_plot_male_us + geom_point(shape = 18, size = 2.8, color = "purple") +
  theme(panel.grid = element_line(color = "lightgrey", linewidth = 0.8, linetype = 1),
   panel.background = element_rect(color = "white", fill = "darkgrey")) +
  labs(title = "US Anxiety Male", x ="Year",
       y = "Percent of Males Experiencing Anxiety") +
  theme(axis.text.x = element_text(angle = 90))
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



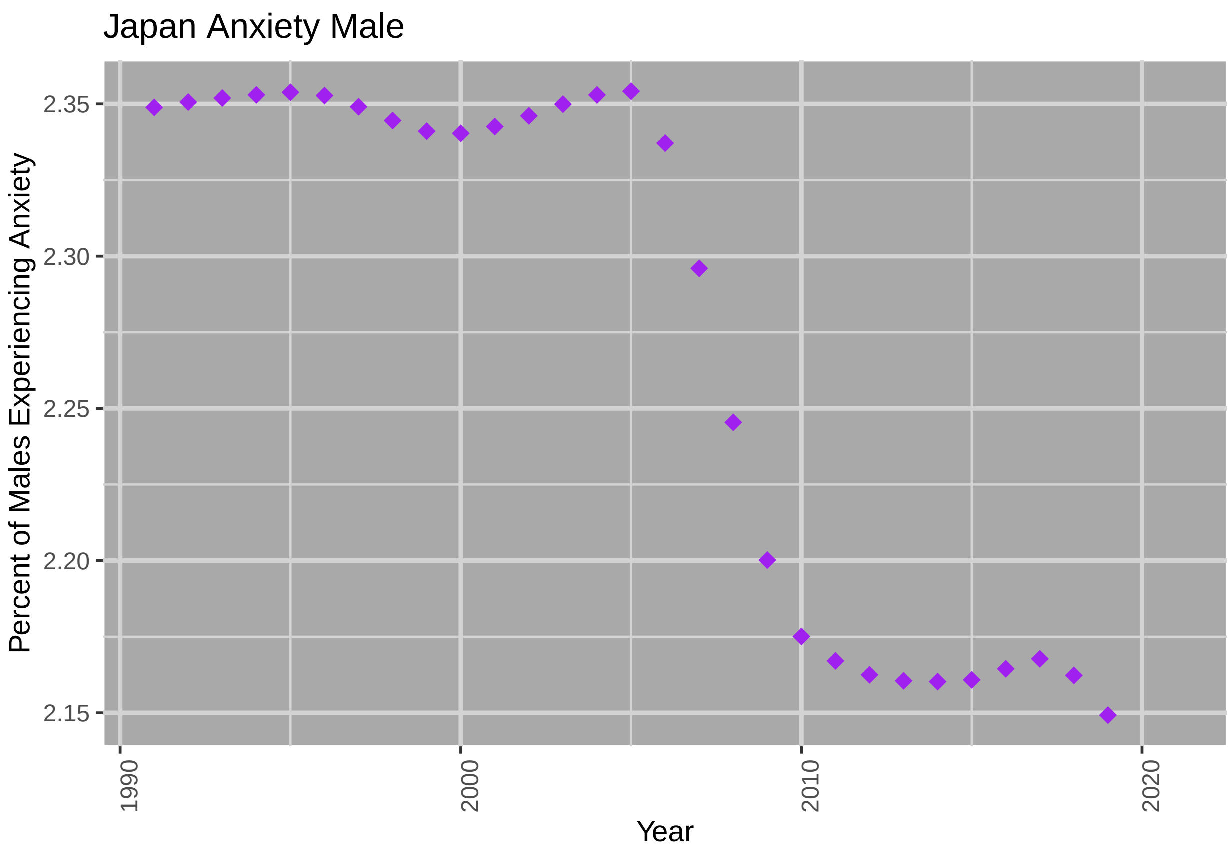
```
# Plot % of people experiencing anxiety female
prevalence_plot_female_us <- ggplot(prevalence_us, aes(year, female))
prevalence_plot_female_us + geom_point(shape = 18, size = 2.8, color = "purple") +
  theme(panel.grid = element_line(color = "lightgrey", linewidth = 0.8, linetype = 1),
 panel.background = element_rect(color = "white", fill = "darkgrey")) +
  labs(title = "US Anxiety Female", x ="Year",
       y = "Percent of Females Experiencing Anxiety") +
  theme(axis.text.x = element_text(angle = 90))
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



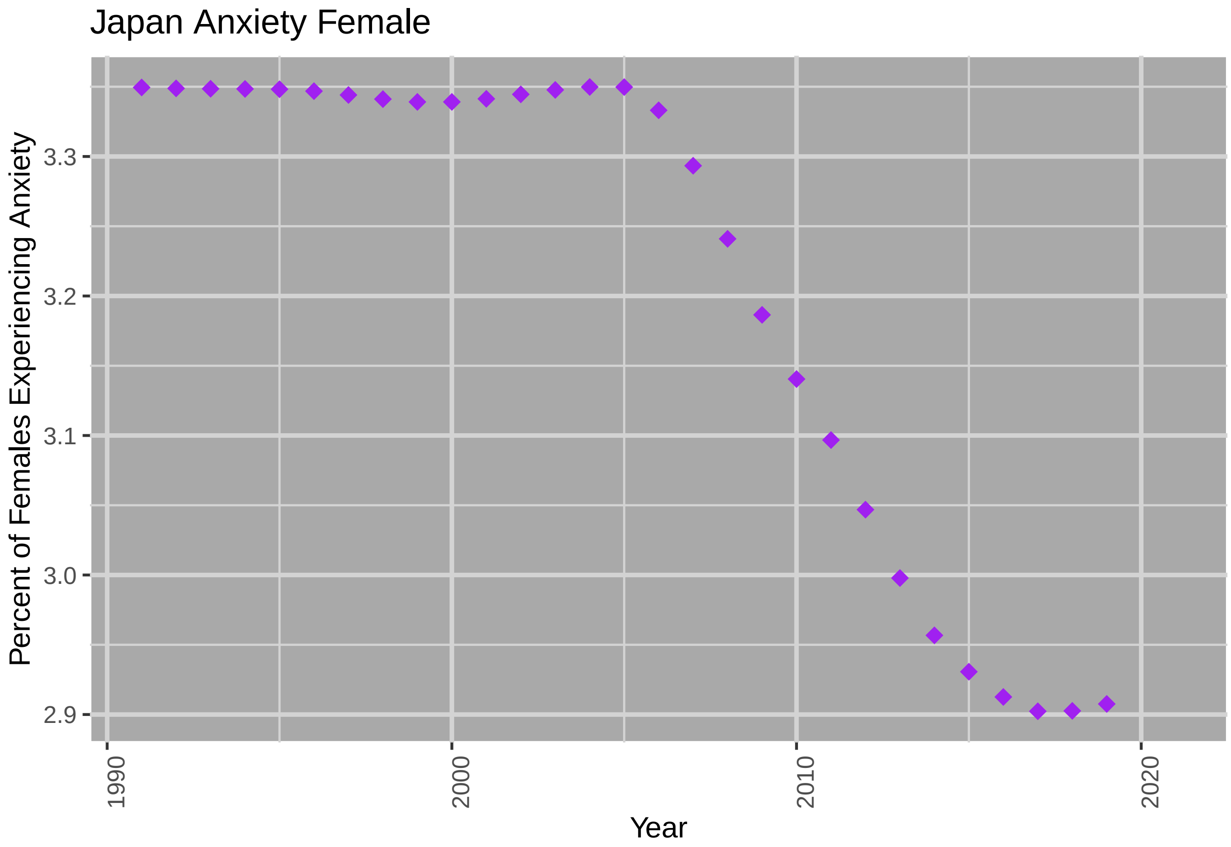
```
# Plot % of people experiencing anxiety male Japan
prevalence_plot_male_japan <- ggplot(prevalence_japan, aes(year, male))
prevalence_plot_male_japan + geom_point(shape = 18, size = 2.8, color = "purple") +
  theme(panel.grid = element_line(color = "lightgrey", linewidth = 0.8, linetype = 1),
   panel.background = element_rect(color = "white", fill = "darkgrey")) +
  labs(title = "Japan Anxiety Male", x ="Year",
       y = "Percent of Males Experiencing Anxiety") +
  theme(axis.text.x = element_text(angle = 90))
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



```
# Plot % of people experiencing anxiety female Japan
prevalence_plot_female_japan <- ggplot(prevalence_japan, aes(year, female))
prevalence_plot_female_japan + geom_point(shape = 18, size = 2.8, color = "purple") +
  theme(panel.grid = element_line(color = "lightgrey", linewidth = 0.8, linetype = 1),
 panel.background = element_rect(color = "white", fill = "darkgrey")) +
  labs(title = "Japan Anxiety Female", x ="Year",
       y = "Percent of Females Experiencing Anxiety") +
  theme(axis.text.x = element_text(angle = 90))
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```

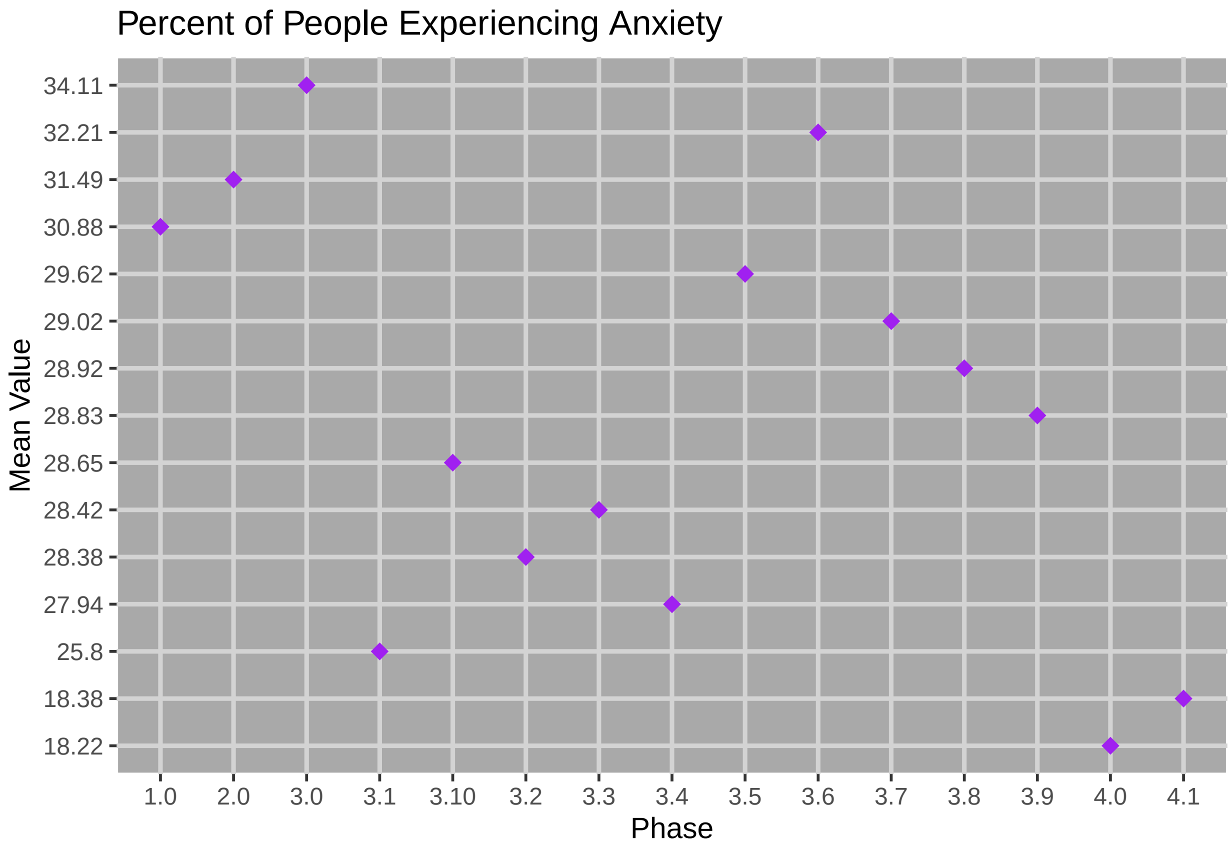


The dichotomy between men's and women's reporting of anxiety is further supported by the "Mental Health Dataset" from Kaggle (Kaggle, 2024). Analysis of the percentage of men and women receiving treatment for their anxiety disagrees with the WHO estimate of 1 in 4 receiving treatment (WHO, 2023). According to the "Mental Health Dataset," women receive treatment for their anxiety at a rate of 70%, while men are at 49%. I do believe this has to do with self-perception, outside perception, and societal expectations. This report will not delve deeper into the psychological reasons for the discrepancy between men's and women's willingness to seek help and accept treatment, but analysis of the data warrants some discussion and theorizing about the causes of statistical differences. The "Indicators of Anxiety or Depression" dataset collected data through the U.S. Census Bureau, asking if people have been experiencing symptoms of anxiety in the last seven days. This dataset has a range of people experiencing anxiety at a rate of 18.2% - 34.1% over all phases of their study. While experiencing symptoms of anxiety over seven days does not indicate a person should receive an anxiety disorder diagnosis, the number of people experiencing these symptoms matches more closely with the number of people with reported anxiety disorder symptoms in the NIH dataset, 19.1% in the last year and 31.1% throughout their lifetime (NIH, 2010).

```r
# Table of treatment
tab_tot_fem_male <- as.data.frame(tab_tot_fem_male)
knitr::kable(tab_tot_fem_male, format = "markdown")
```

	Total	Male	Female
Total People	165370	130650	34720
No Treatment	77496	67080	10416
Treatment	87874	63570	24304
Percentage Receiving Treatment	53.14	48.66	70

```
# Plot % of people experiencing
indicators_plot <- ggplot(indicators, aes(phases, indicators_mean_val))
indicators_plot + geom_point(shape = 18, size = 2.8, color = "purple") +
  theme(panel.grid = element_line(color = "lightgrey", linewidth = 0.8, linetype = 1),
panel.background = element_rect(color = "white", fill = "darkgrey")) +
  labs(title = "Percent of People Experiencing Anxiety", x ="Phase", y = "Mean Value")
```



## Implications The most important takeaways from this analysis were that the 4% of people reporting an anxiety disorder from the WHO study were biased low compared to all of the datasets analyzed in this study and that culture, gender, and other demographic information influences the number of people reporting anxiety disorders. With the knowledge that the WHO study was biased low, I believe there should be a more significant effort to reduce stigma, improve access to mental health resources, and find more accurate ways to track mental health data. ## Limitations, Future Work Two additional datasets were initially selected to be included in this analysis. The "Health Anxiety Dataset" and "Psychiatric Drug WebMD Reviews" datasets from Kaggle. The "Health Anxiety Dataset" included 350 columns of data; I believe most of them were predictor variables, but their descriptions were insufficient for me to figure out what they were. The "Psychiatric Drug WebMD Reviews" dataset included a text field that I did not know how to parse to draw conclusions from. Additionally, the WebMD dataset has well over 100 medications to evaluate but analyzing any of them would not have been useful without knowing what symptoms they were treating which I couldn't parse from the text field. The combination of information from these two datasets would have been instrumental in creating a model to predict if someone A) Has anxiety, B) Needs medication, and C) What the best medication for them would be. Instead of analyzing these datasets manually, what would be the implications of having a machine learning program look at the data, the two excluded datasets, and the other datasets that were not relevant to this analysis? While working on this project, I found other irrelevant datasets, but they would be helpful to feed the ML program. The additional datasets I'm thinking of for this involve text responses and social media posts from people with anxiety. We could potentially

use machine learning to set the foundation for screening methods for mental health treatment. There are some substantial potential problems with using ML as a screening tool. What happens if the ML program makes the wrong decision is the most glaring to me. Does the person in question not receive care? I would hope that it could look for cases where it is definite that the person is exhibiting anxious behavior and flag them for review with the person's doctor but not exclude a person from receiving care if the ML program does not flag them. Flagging people based on personal data leads us to another dilemma: How much of this person's private health information would the ML program need access to for the program to be successful? How much of their digital information would the program need to look through, think social media, browser history, subscriptions, etc.? What is considered too invasive? These considerations need to be reviewed, and even with review, we may not reach a unanimous conclusion as the questions are subjective. Person A may be okay with the program tracking their browser history, while person B says absolutely not. We could utilize ML or DL to look at a holistic picture of a person (demographics, economic status, stress factors, family history, previous medical history, etc.) and what medication worked best for them. We can then predict which medication would work best for someone with a similar profile. There are also services available that can theorize about medication effectiveness based on a person's DNA (I personally have used GeneSight). They are relatively inexpensive but would be an excellent asset for a ML program of this type. ## Concluding Remarks I believe that there should be a more significant investment of money, time, and effort into accurately tracking mental health data and providing access to those in need. We need to take steps that remove the stigma of seeking help for mental illness for all and, in particular, provide support and reduce the social impact of men seeking help. The proposed implementation of ML models for making mental health predictions has potentially significant problems, both ethical and practical problems, that require input from the community at large. If we reach a consensus on algorithm parameters, then some ethical issues could be alleviated with a voluntary waiver. While the subject of mental illness is important to me personally, it impacts nearly everyone in some way. Using technology in sensitive matters can be problematic and is made even more complicated when the matter has subjective components. I do not believe that means we shouldn't try it; we just have to make intelligent choices about when and how we use it.