McKibben DSC520 Final Project Week 10

Makayla McKibben

2024-08-12

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.

```
# 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
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)
```

```r
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...."
```

```r
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),]
```

```r
# 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"
```

```r
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"
```

```r
# 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
```

```r
# 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"
```

```r
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")
```

```r
# 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
```

```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)
# The following would qualify as something I don't know how to do that
# could be helpful to know.
```

```r
# 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.

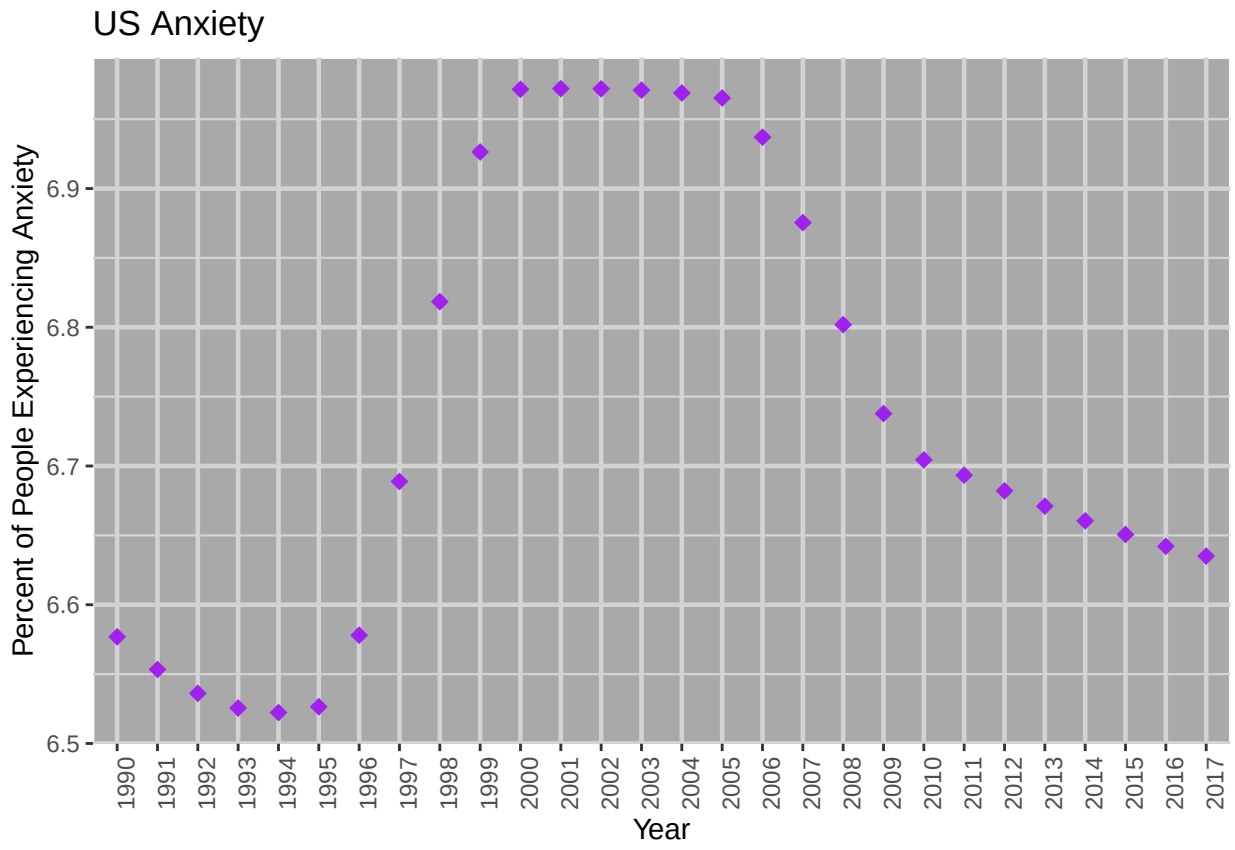
How you addressed this problem statement

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.

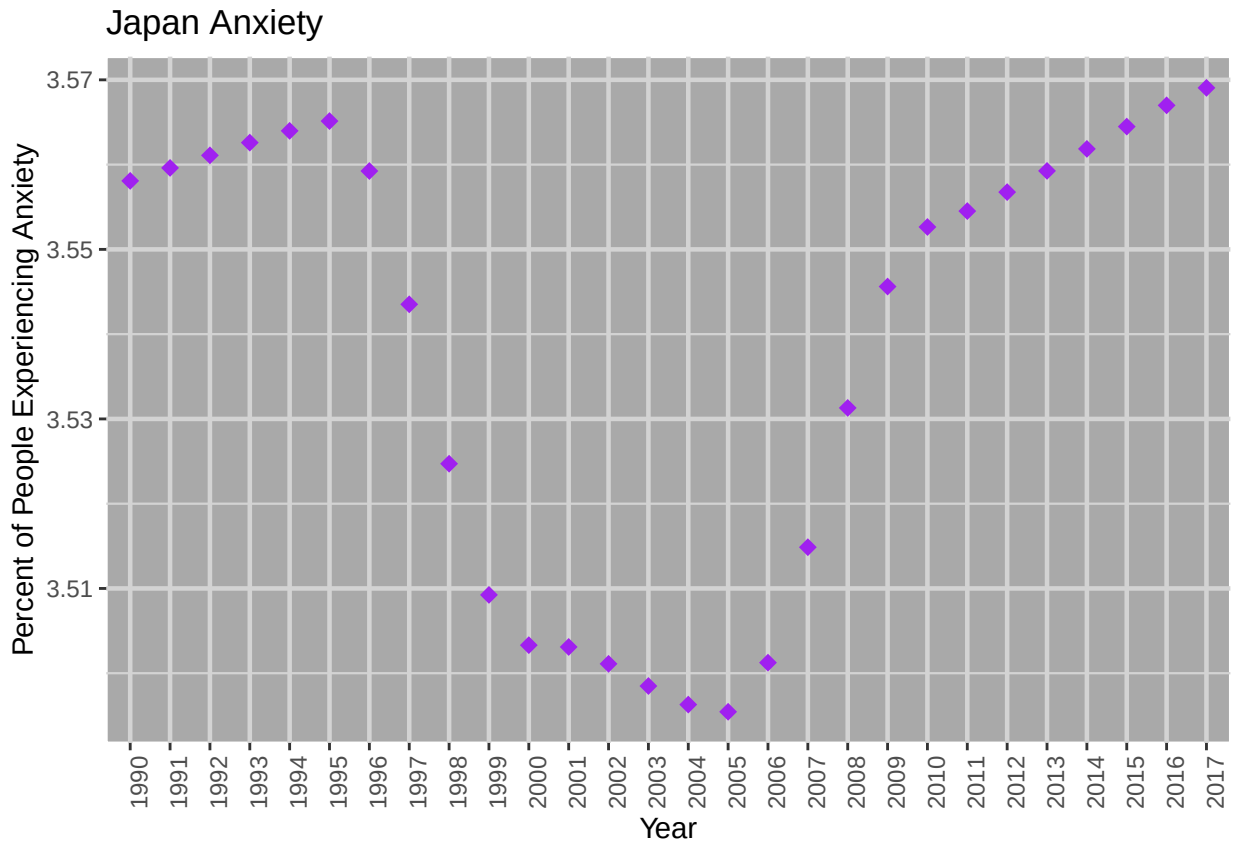
Analysis

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.

```
# 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))
```

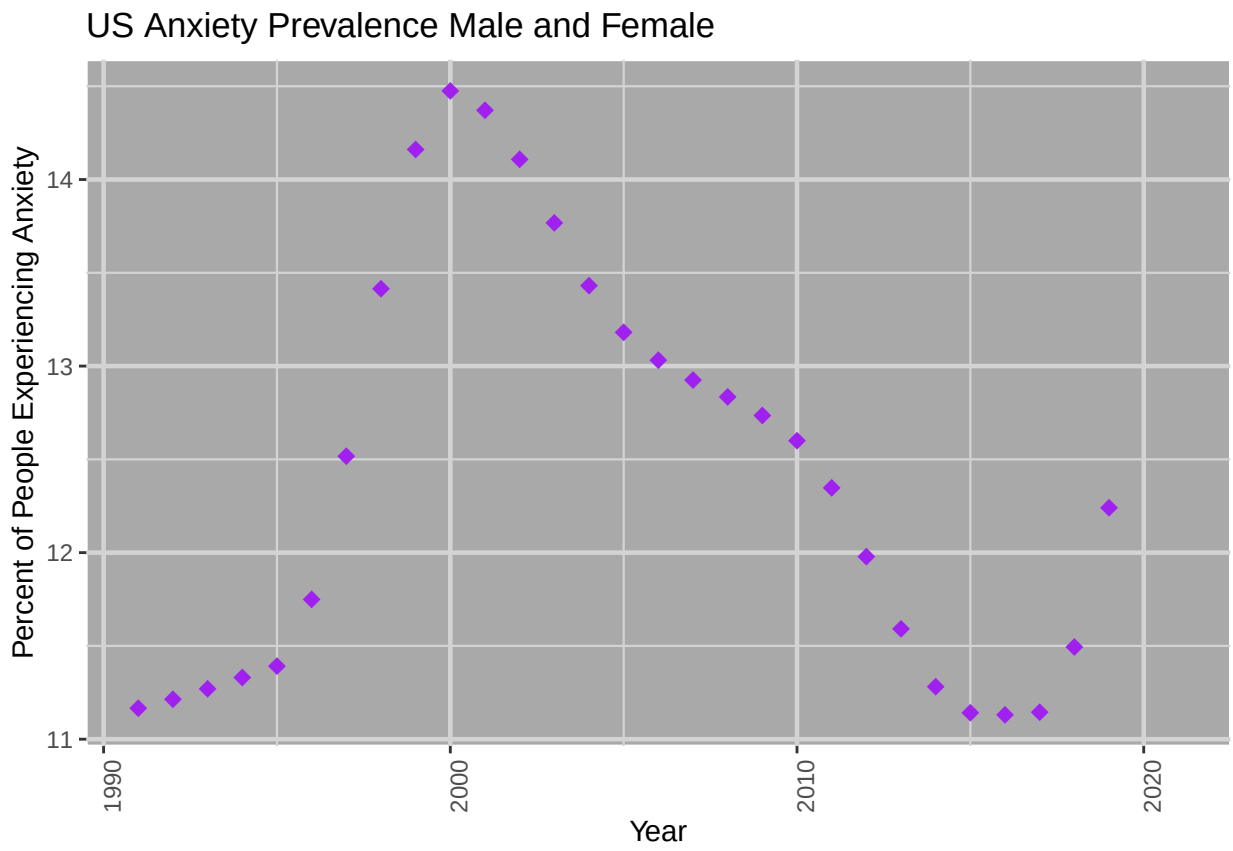


```
# Plot % of people experiencing anxiety Japan
global_plot_japan <- ggplot(global_mental_japan, aes(Year, anxiety))
global_plot_japan + geom_point(shape = 18, size = 2.8, color = "purple") +
  theme(panel.grid = element_line(color = "lightgrey", linewidth = 0.8, linetype = 1),
  panel.background = 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))
```



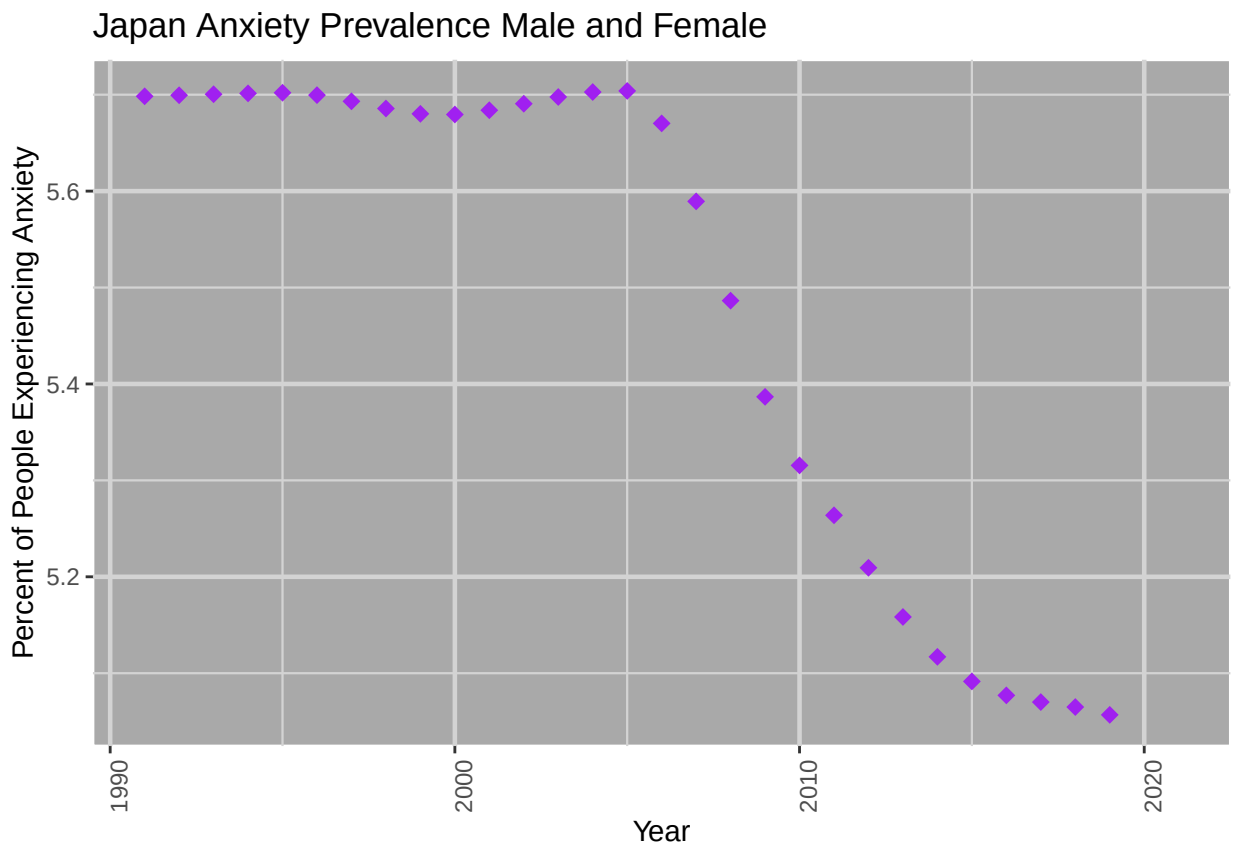

```
# Plot % of people experiencing anxiety US
prevalence_plot_us <- ggplot(prevalence_us, aes(year, male + female))
prevalence_plot_us + geom_point(shape = 18, size = 2.8, color = "purple") +
  theme(panel.grid = element_line(color = "lightgrey", linewidth = 0.8, linetype = 1),
   panel.background = 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()`).
```




```
# 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))
```

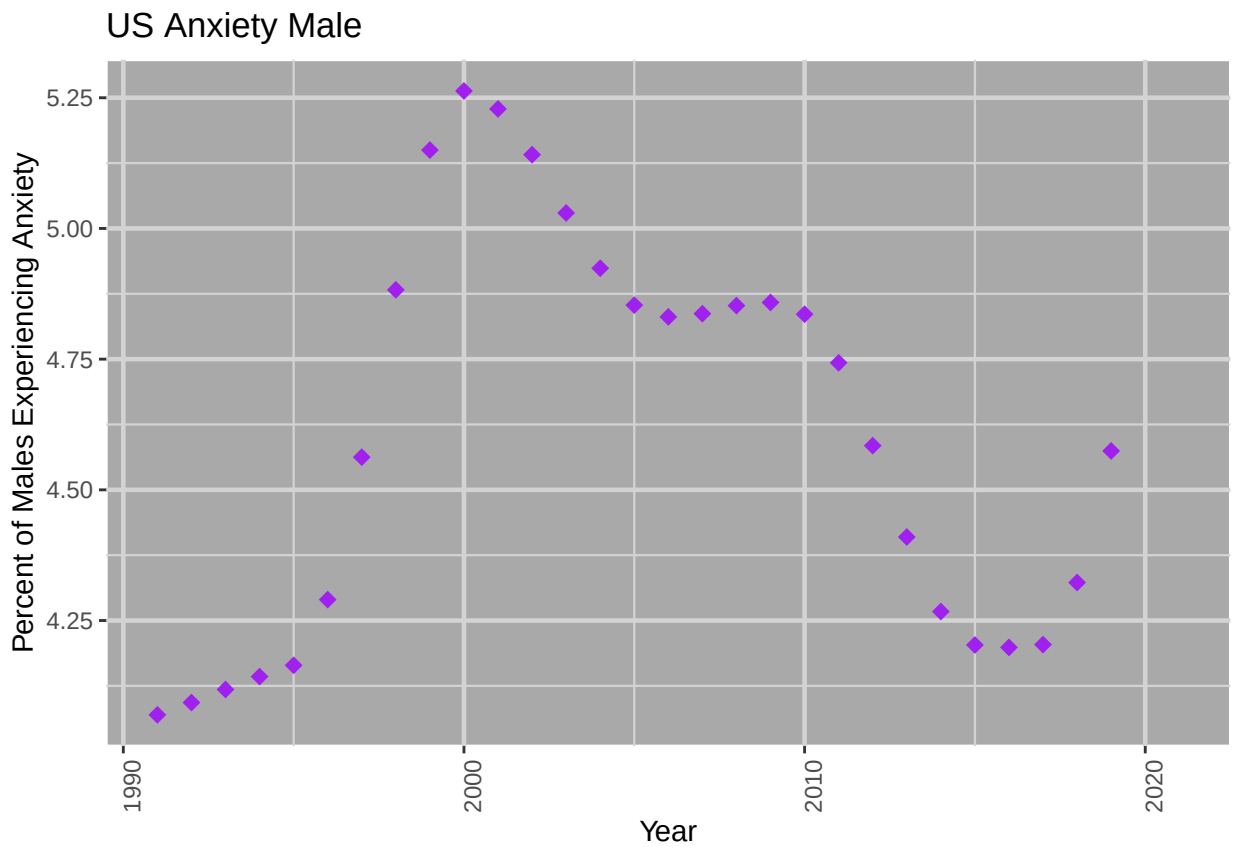
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



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.

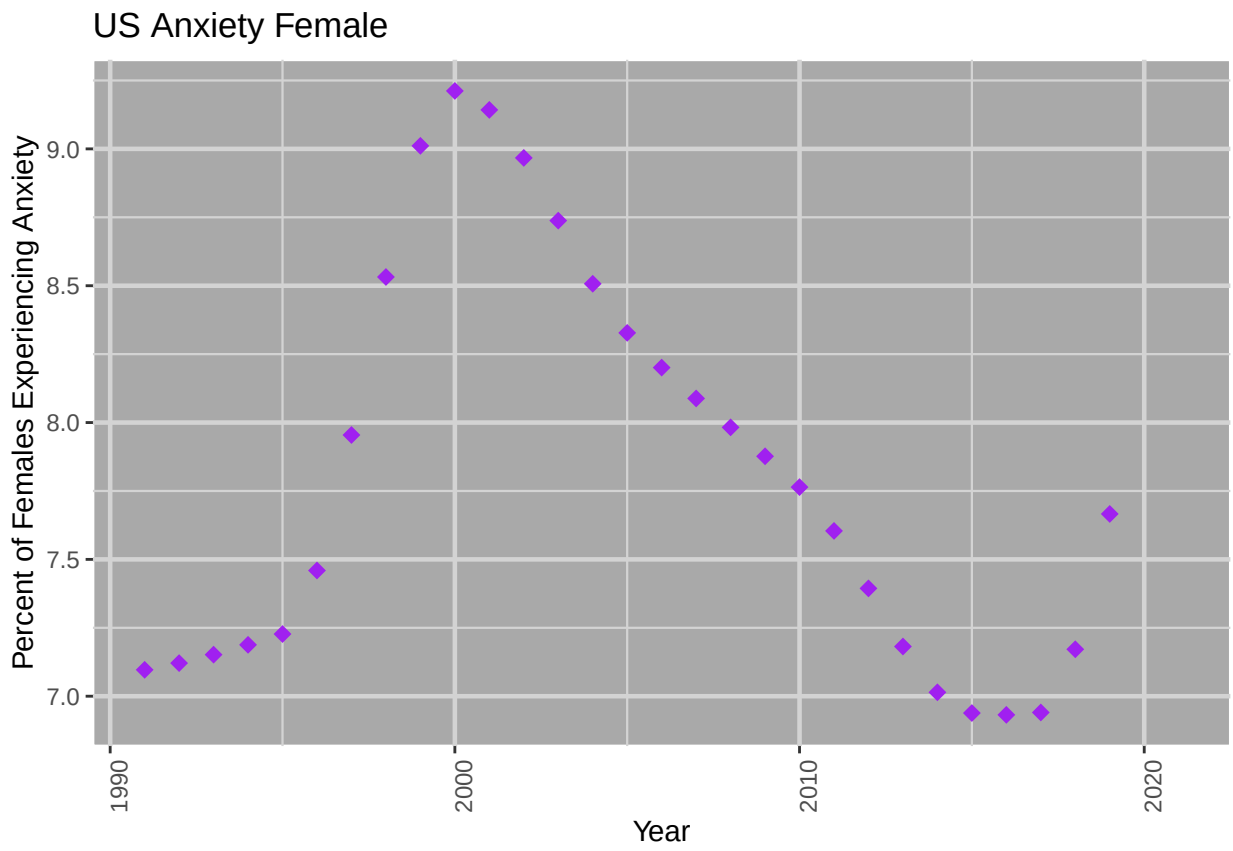
```
    # Plot % of people experiencing anxiety male
prevalence_plot_male_us <- ggplot(prevalence_us, aes(year, male))
prevalence_plot_male_us + geom_point(shape = 18, size = 2.8, color = "purple") +
  theme(panel.grid = element_line(color = "lightgrey", linewidth = 0.8, linetype = 1),
   panel.background = 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()`).
```



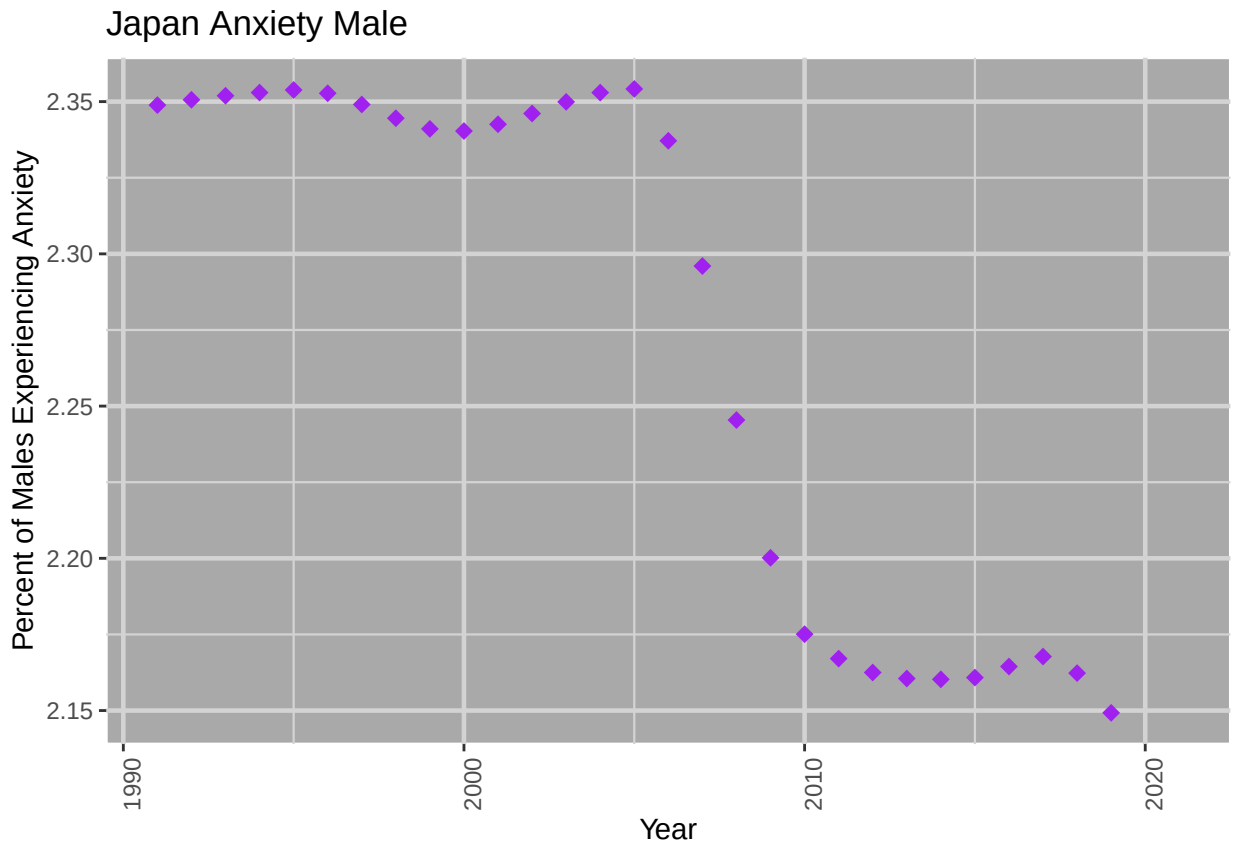

```
# Plot % of people experiencing anxiety female
prevalence_plot_female_us <- ggplot(prevalence_us, aes(year, female))
prevalence_plot_female_us + geom_point(shape = 18, size = 2.8, color = "purple") +
  theme(panel.grid = element_line(color = "lightgrey", linewidth = 0.8, linetype = 1),
 panel.background = 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()`).
```



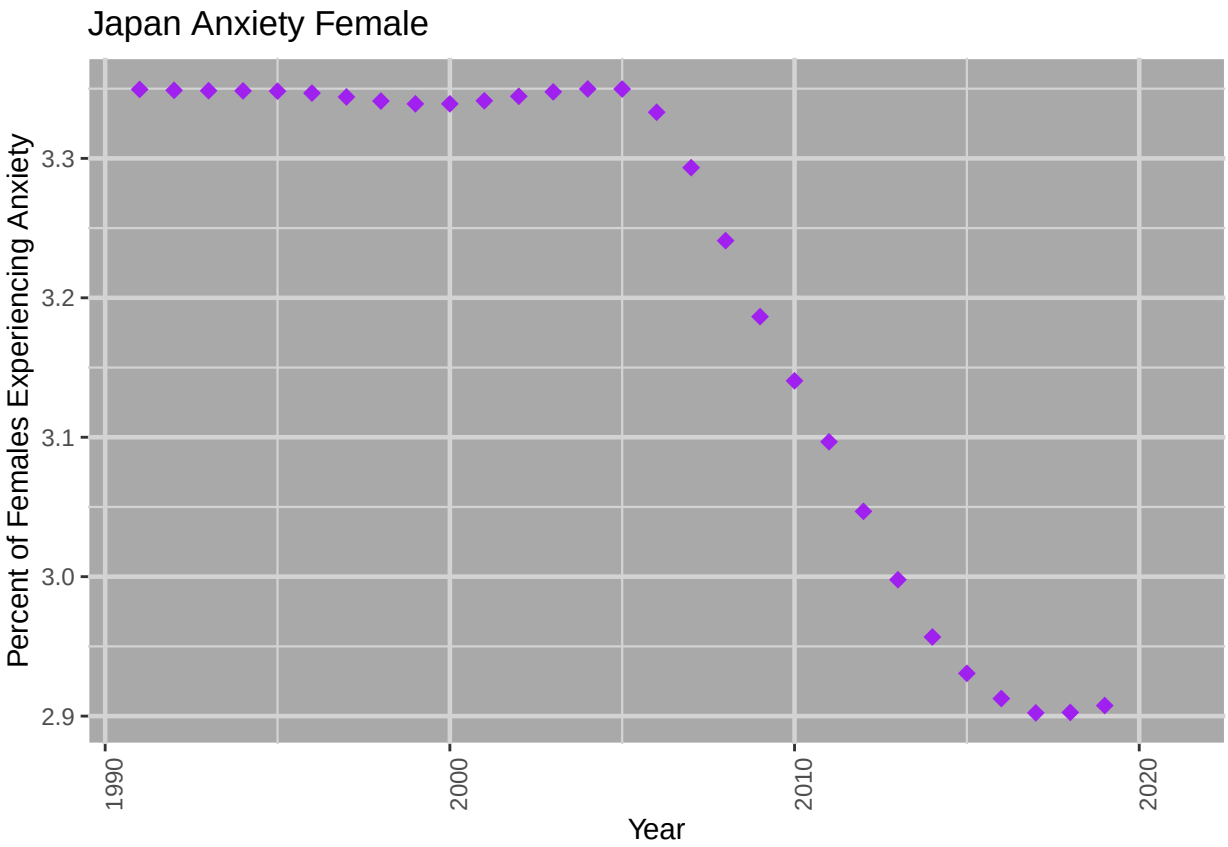

```
# Plot % of people experiencing anxiety male Japan
prevalence_plot_male_japan <- ggplot(prevalence_japan, aes(year, male))
prevalence_plot_male_japan + geom_point(shape = 18, size = 2.8, color = "purple") +
  theme(panel.grid = element_line(color = "lightgrey", linewidth = 0.8, linetype = 1),
   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()`).
```

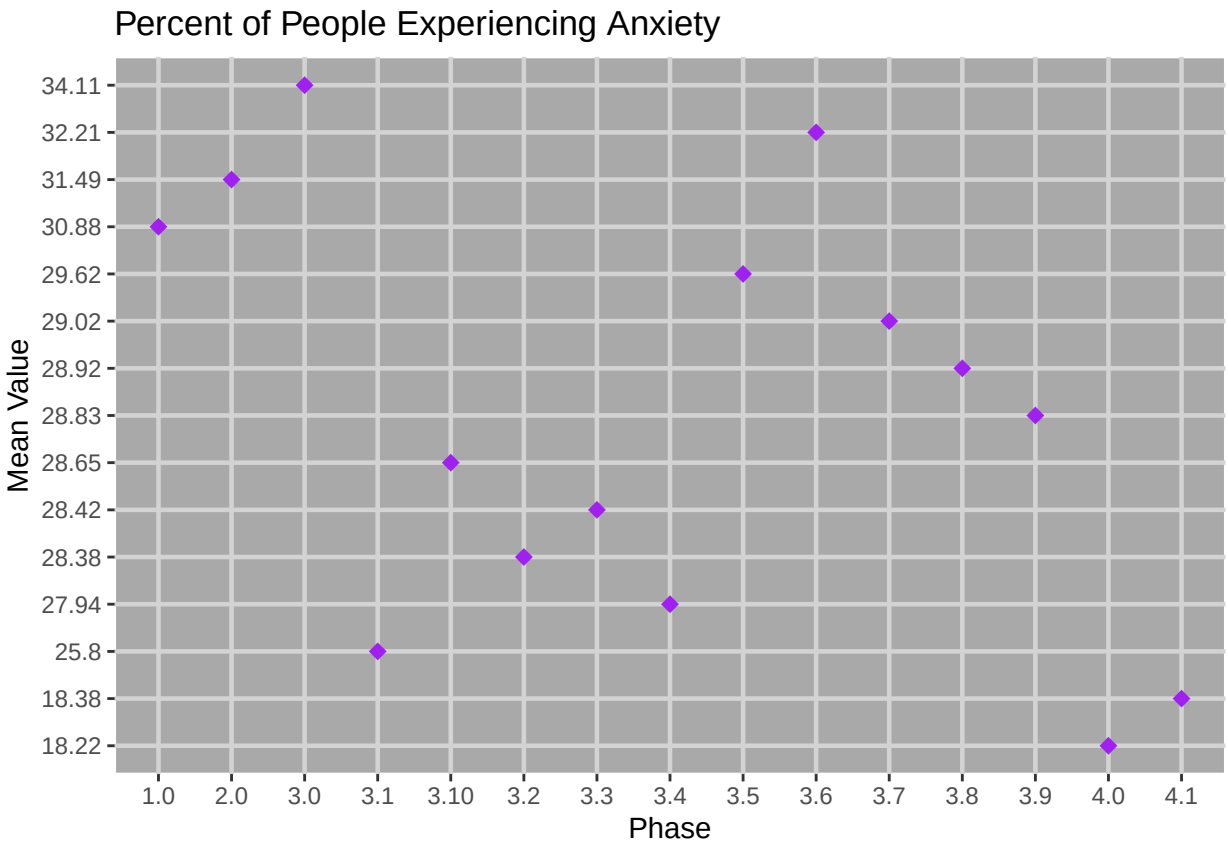


The dichotomy between men's and women's reporting of anxiety is further supported by the "Mental Health Dataset" from Kaggle (Kaggle, 2024). Analysis of the percentage of men and women receiving treatment for their anxiety disagrees with the WHO estimate of 1 in 4 receiving treatment (WHO, 2023). According 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).

```
# 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

```r
# 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.