

# Final Project

Mariana MacDonald

2022-05-22

**Data importing and cleaning steps are explained in the text and follow a logical process. Outline your data preparation and cleansing steps.**

- familiarized with the data sets;
- checked for NAs, errors or missing values;
- changed the names of the variables when needed to make it standard and easier to read and use;
- extracted only relevant variables from the data sets for my research;
- most of the data I am using, is already clean

**With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.**

DATA1 - Mental health Depression disorder Data

```
setwd("/Users/arianamacdonald/Documents/DATA SCIENCE/DSC 520/Statistics R/Week 2/dsc520")
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##   filter, lag
```

```
## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
```

```
DATA1 <- read_excel("DATA 1 - Final Project - Mental health Depression disorder Data.xlsx")
library(tidyverse)
```

```
## -- Attaching packages ----- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5      v purrr   0.3.4
## v tibble  3.1.7      v stringr 1.4.0
## v tidyr   1.2.0      v forcats 0.5.1
## v readr   2.1.2
```

```

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag() masks stats::lag()

library(magrittr)

##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##
## set_names

## The following object is masked from 'package:tidyr':
##
## extract

library(tidyr)
library(purrr)
names(DATA1)

## [1] "Entity" "Code"
## [3] "Year" "Schizophrenia (%)"
## [5] "Bipolar disorder (%)" "Eating disorders (%)"
## [7] "Anxiety disorders (%)" "Drug use disorders (%)"
## [9] "Depression (%)" "Alcohol use disorders (%)"

colnames(DATA1) <- c("entity", "code", "year", "schizophrenia", "bipolar_disorder",
                    "eating_disorders", "anxiety_disorders", "drug_use_disorders", "depression",
                    "alcohol_use_disorders")

newdata1 <- subset(DATA1, code == "USA", select=c(code, year, depression))
head(newdata1)

## # A tibble: 6 x 3
## code year depression
## <chr> <dbl> <dbl>
## 1 USA 1990 4.68
## 2 USA 1991 4.66
## 3 USA 1992 4.65
## 4 USA 1993 4.65
## 5 USA 1994 4.65
## 6 USA 1995 4.65

newdata2<- subset(DATA1, select=c(code, depression))
head(newdata2)

## # A tibble: 6 x 2
## code depression
## <chr> <dbl>
## 1 AFG 4.07

```

```
## 2 AFG      4.08
## 3 AFG      4.09
## 4 AFG      4.10
## 5 AFG      4.10
## 6 AFG      4.10
```

## DATA 2 - Glassdoor Gender Pay Gap

```
setwd("/Users/marianamacdonald/Documents/DATA SCIENCE/DSC 520/Statistics R/Week 2/dsc520")
DATA2 <- read.csv("DATA 2 - Final Project - Glassdoor Gender Pay Gap.csv", header = T,
                  stringsAsFactors = T)
names(DATA2)
```

```
## [1] "JobTitle" "Gender" "Age" "PerfEval" "Education" "Dept"
## [7] "Seniority" "BasePay" "Bonus"
```

```
head(DATA2)
```

```
##      JobTitle Gender Age PerfEval Education      Dept Seniority
## 1  Graphic Designer Female   18      5   College Operations      2
## 2  Software Engineer   Male   21      5   College Management      5
## 3 Warehouse Associate Female   19      4      PhD Administration      5
## 4  Software Engineer   Male   20      5   Masters      Sales      4
## 5  Graphic Designer   Male   26      5   Masters Engineering      5
## 6              IT Female   20      5      PhD      Operations      4
##      BasePay Bonus
## 1    42363  9938
## 2   108476 11128
## 3    90208  9268
## 4   108080 10154
## 5    99464  9319
## 6    70890 10126
```

## DATA 3 - The Depression Dataset -

From this data, I intend to use only a few variables, I have removed the melanch and inpatient, which had NA values and I am not interested on them. Also, from the entire dataset, I am only able to use a few rows. The others are missing basically all the information. Conditions 7,8,9 have been removed.

```
library(tidyr)

setwd("/Users/marianamacdonald/Documents/DATA SCIENCE/DSC 520/Statistics R/Week 2/dsc520")
DATA3 <- read.csv("DATA 3 - Final Project - scores.csv", header = T)
names(DATA3)
```

```
## [1] "number" "days" "gender" "age" "afftype" "melanch"
## [7] "inpatient" "edu" "marriage" "work" "madr1" "madr2"
```

```
newdata3 <- DATA3 %>% drop_na(afftype, melanch)
newdata3
```

##	number	days	gender	age	afftype	melanch	inpatient	edu	marriage	work
## 1	condition_1	11	2	35-39	2	2	2	6-10	1	2
## 2	condition_2	18	2	40-44	1	2	2	6-10	2	2
## 3	condition_3	13	1	45-49	2	2	2	6-10	2	2
## 4	condition_4	13	2	25-29	2	2	2	11-15	1	1
## 5	condition_5	13	2	50-54	2	2	2	11-15	2	2
## 6	condition_6	7	1	35-39	2	2	2	6-10	1	2
## 7	condition_10	9	2	45-49	2	2	2	6-10	1	2
## 8	condition_11	14	1	45-49	2	2	2	6-10	1	2
## 9	condition_12	12	2	40-44	1	2	2	6-10	2	2
## 10	condition_13	14	2	35-39	1	2	2	11-15	2	2
## 11	condition_14	14	1	60-64	1	2	2	6-10	2	2
## 12	condition_15	13	2	55-59	2	2	2	11-15	1	1
## 13	condition_16	16	1	45-49	2	2	2	11-15	1	2
## 14	condition_17	13	1	50-54	1	2	2	6-10	1	2
## 15	condition_18	13	2	40-44	3	2	2	11-15	2	2
## 16	condition_19	13	2	50-54	2	2	1	16-20	2	2
## 17	condition_20	13	1	30-34	2	1	1	6-10	1	2
## 18	condition_21	13	2	35-39	2	2	1	6-10	2	2
## 19	condition_22	14	1	65-69	2	2	1		2	2
## 20	condition_23	16	1	30-34	2	2	1	16-20	2	2
##	madrsl	madrsl								
## 1	19	19								
## 2	24	11								
## 3	24	25								
## 4	20	16								
## 5	26	26								
## 6	18	15								
## 7	28	21								
## 8	24	24								
## 9	25	21								
## 10	18	13								
## 11	28	19								
## 12	14	18								
## 13	13	17								
## 14	17	15								
## 15	18	15								
## 16	26	21								
## 17	27	25								
## 18	26	21								
## 19	29	28								
## 20	29	23								

Description of variables number (patient identifier), days (number of days of measurements), gender (1 or 2 for female or male), age (age in age groups), afftype (1: bipolar II, 2: unipolar depressive, 3: bipolar I), melanch (1: melancholia, 2: no melancholia), inpatient (1: inpatient, 2: outpatient), edu (education grouped in years), marriage (1: married or cohabiting, 2: single), work (1: working or studying, 2: unemployed/sick leave/pension), madrs1 (MADRS score when measurement started), madrs2 (MADRS when measurement stopped).

**What do you not know how to do right now that you need to learn to import and cleanup your dataset?**

I have learned how to import csv, excel and arff dataset so I believe I have learned what I need for this project. What is pending is machine learning.

**Discuss how you plan to uncover new information in the data that is not self-evident.**

At this moment, I am not sure if the predictors I am selecting will have relationship to the questions I want to answer, so I might need to use other variables to get to my solution. I might use correlation, regression, ANOVA, histograms and/or graphs to uncover new information.

**What are different ways you could look at this data to answer the questions you want to answer?**

DATA1 Instead of only considering depression, I can sum the % of all the mental disorders and create a new variable (called Sum)

```
library(readxl)
getwd()
```

```
## [1] "/Users/marianamacdonald/Documents/DATA SCIENCE/DSC 520/Statistics R/Week 2/dsc520"
```

```
disorders_df <- read_excel("DATA 1 - Final Project - Mental health Depression disorder Data.xlsx")
head(disorders_df)
```

```
## # A tibble: 6 x 10
##   Entity      Code  Year 'Schizophrenia (%)' 'Bipolar disord~' 'Eating disord~'
##   <chr>      <chr> <dbl>             <dbl>             <dbl>             <dbl>
## 1 Afghanistan AFG   1990             0.161             0.698             0.102
## 2 Afghanistan AFG   1991             0.160             0.698             0.0993
## 3 Afghanistan AFG   1992             0.160             0.698             0.0967
## 4 Afghanistan AFG   1993             0.160             0.698             0.0943
## 5 Afghanistan AFG   1994             0.160             0.698             0.0924
## 6 Afghanistan AFG   1995             0.160             0.699             0.0910
## # ... with 4 more variables: 'Anxiety disorders (%)' <dbl>,
## #   'Drug use disorders (%)' <dbl>, 'Depression (%)' <dbl>,
## #   'Alcohol use disorders (%)' <dbl>
```

```
colnames(disorders_df)
```

```
## [1] "Entity"      "Code"
## [3] "Year"        "Schizophrenia (%)"
## [5] "Bipolar disorder (%)" "Eating disorders (%)"
## [7] "Anxiety disorders (%)" "Drug use disorders (%)"
## [9] "Depression (%)" "Alcohol use disorders (%)"
```

```
disorders_df$Sum <- rowSums(disorders_df[c('Schizophrenia (%)', 'Bipolar disorder (%)',
'Eating disorders (%)', 'Anxiety disorders (%)',
'Drug use disorders (%)', 'Depression (%)',
'Alcohol use disorders (%)')], na.rm = TRUE)
head(disorders_df)
```

```
## # A tibble: 6 x 11
##   Entity      Code   Year 'Schizophrenia (%)' 'Bipolar disord~' 'Eating disord~'
##   <chr>      <chr> <dbl>          <dbl>          <dbl>          <dbl>
## 1 Afghanistan AFG    1990            0.161            0.698            0.102
## 2 Afghanistan AFG    1991            0.160            0.698            0.0993
## 3 Afghanistan AFG    1992            0.160            0.698            0.0967
## 4 Afghanistan AFG    1993            0.160            0.698            0.0943
## 5 Afghanistan AFG    1994            0.160            0.698            0.0924
## 6 Afghanistan AFG    1995            0.160            0.699            0.0910
## # ... with 5 more variables: 'Anxiety disorders (%)' <dbl>,
## #   'Drug use disorders (%)' <dbl>, 'Depression (%)' <dbl>,
## #   'Alcohol use disorders (%)' <dbl>, Sum <dbl>
```

DATA 2 I can separate the data into male and female base pay, and look at the summary to find the mean and compare. (Male USD 98,458 x Female USD 89,943)

```
malepay <- subset(DATA2, Gender == "Male", select=c(Gender, BasePay))
head(malepay)
```

```
##   Gender BasePay
## 2   Male 108476
## 4   Male 108080
## 5   Male  99464
## 8   Male  97523
## 11  Male 102261
## 19  Male  90386
```

```
summary(malepay)
```

```
##      Gender      BasePay
## Female: 0   Min.   : 36642
## Male  :532 1st Qu.: 81452
##           Median : 98223
##           Mean   : 98458
##           3rd Qu.:115606
##           Max.   :179726
```

```
femalepay <- subset(DATA2, Gender == "Female", select=c(Gender, BasePay))
head(femalepay)
```

```
##   Gender BasePay
## 1 Female  42363
## 3 Female  90208
## 6 Female  70890
## 7 Female  67585
## 9 Female 112976
## 10 Female 106524
```

```
summary(femalepay)
```

```
##      Gender      BasePay
## Female:468   Min.    : 34208
## Male  :    0   1st Qu.: 73186
##                      Median : 89914
##                      Mean    : 89943
##                      3rd Qu.:106923
##                      Max.    :160614
```

Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.

I will not be joining data frames. They are very different and I won't benefit from joining them. I might create new variables.

How could you summarize your data to answer key questions?

```
summary(newdata1)
```

```
##      code      year      depression
## Length:28      Min.    :1990      Min.    :4.649
## Class :character 1st Qu.:1997      1st Qu.:4.686
## Mode  :character Median :2004      Median :4.766
##                      Mean    :2004      Mean    :4.745
##                      3rd Qu.:2010      3rd Qu.:4.783
##                      Max.    :2017      Max.    :4.836
```

```
summary(newdata2)
```

```
##      code      depression
## Length:6468      Min.    :2.140
## Class :character 1st Qu.:3.006
## Mode  :character Median :3.500
##                      Mean    :3.498
##                      3rd Qu.:3.912
##                      Max.    :6.603
```

```
summary(DATA2)
```

```
##      JobTitle      Gender      Age      PerfEval
## Marketing Associate:118 Female:468   Min.    :18.00   Min.    :1.000
## Software Engineer  :109 Male   :532   1st Qu.:29.00   1st Qu.:2.000
## Data Scientist     :107              Median :41.00   Median :3.000
## Financial Analyst   :107              Mean    :41.39   Mean    :3.037
## Graphic Designer    : 98              3rd Qu.:54.25   3rd Qu.:4.000
## IT                  : 96              Max.    :65.00   Max.    :5.000
## (Other)             :365
```

```
##      Education      Dept      Seniority      BasePay
## College :241 Administration:193 Min. :1.000 Min. : 34208
## High School:265 Engineering :192 1st Qu.:2.000 1st Qu.: 76850
## Masters :256 Management :198 Median :3.000 Median : 93328
## PhD :238 Operations :210 Mean :2.971 Mean : 94473
## Sales :207 3rd Qu.:4.000 3rd Qu.:111558
## Max. :5.000 Max. :179726
##
##      Bonus
## Min. : 1703
## 1st Qu.: 4850
## Median : 6507
## Mean : 6467
## 3rd Qu.: 8026
## Max. :11293
##
```

```
summary(newdata3)
```

```
##      number      days      gender      age
## Length:20 Min. : 7.0 Min. :1.00 Length:20
## Class :character 1st Qu.:13.0 1st Qu.:1.00 Class :character
## Mode :character Median :13.0 Median :2.00 Mode :character
## Mean :13.1 Mean :1.55
## 3rd Qu.:14.0 3rd Qu.:2.00
## Max. :18.0 Max. :2.00
##      afftype      melanch      inpatient      edu      marriage
## Min. :1.00 Min. :1.00 Min. :1.00 Length:20 Min. :1.00
## 1st Qu.:1.75 1st Qu.:2.00 1st Qu.:1.75 Class :character 1st Qu.:1.00
## Median :2.00 Median :2.00 Median :2.00 Mode :character Median :2.00
## Mean :1.80 Mean :1.95 Mean :1.75 Mean :1.55
## 3rd Qu.:2.00 3rd Qu.:2.00 3rd Qu.:2.00 3rd Qu.:2.00
## Max. :3.00 Max. :2.00 Max. :2.00 Max. :2.00
##      work      madrs1      madrs2
## Min. :1.0 Min. :13.00 Min. :11.00
## 1st Qu.:2.0 1st Qu.:18.00 1st Qu.:15.75
## Median :2.0 Median :24.00 Median :20.00
## Mean :1.9 Mean :22.65 Mean :19.65
## 3rd Qu.:2.0 3rd Qu.:26.25 3rd Qu.:23.25
## Max. :2.0 Max. :29.00 Max. :28.00
```

What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).

DATA 1

Depression in the USA during the years

```
library(ggplot2)
theme_set(theme_minimal())
ggplot(newdata1, aes(x=year, y=depression)) + geom_point() +
  ggtitle("Depression in the USA")+xlab("Country") + ylab("Depression %")
```



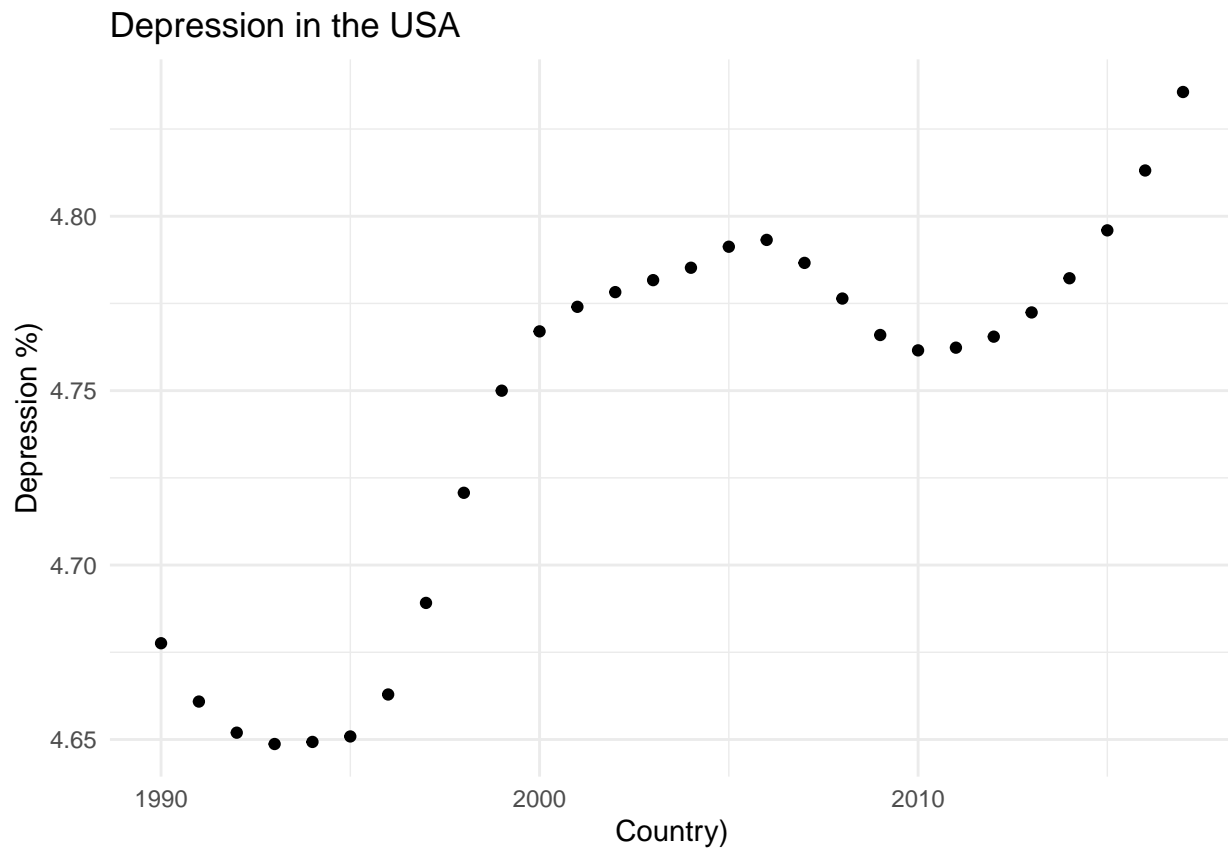


Table of USA, Year and Depression %

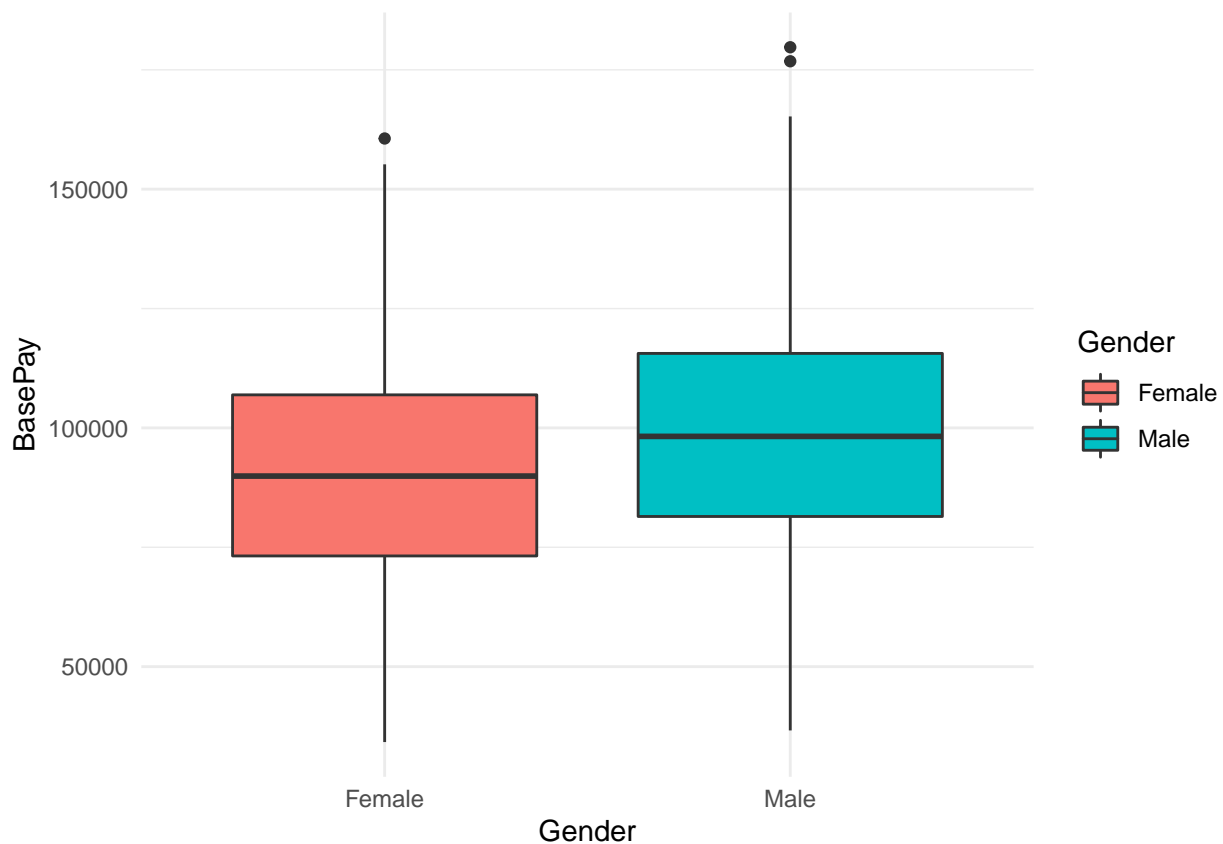
code	year	depression
USA	1990	4.678
USA	1991	4.661
USA	1992	4.652
USA	1993	4.649
USA	1994	4.649
USA	1995	4.651
USA	1996	4.663
USA	1997	4.689
USA	1998	4.721
USA	1999	4.75
USA	2000	4.767
USA	2001	4.774
USA	2002	4.778
USA	2003	4.782
USA	2004	4.785
USA	2005	4.791
USA	2006	4.793
USA	2007	4.787
USA	2008	4.776
USA	2009	4.766
USA	2010	4.762
USA	2011	4.762
USA	2012	4.765

code	year	depression
USA	2013	4.772
USA	2014	4.782
USA	2015	4.796
USA	2016	4.813
USA	2017	4.836

DATA 2

gap pay between male and female

```
qplot(Gender, BasePay, geom = "boxplot", data = DATA2, na.rm=TRUE, fill=Gender)
```



Education x gender

```
library(dplyr)
DATA2 %>% group_by(Gender, Education) %>% summarize(count = n())
```

```
## 'summarise()' has grouped output by 'Gender'. You can override using the
## '.groups' argument.
```

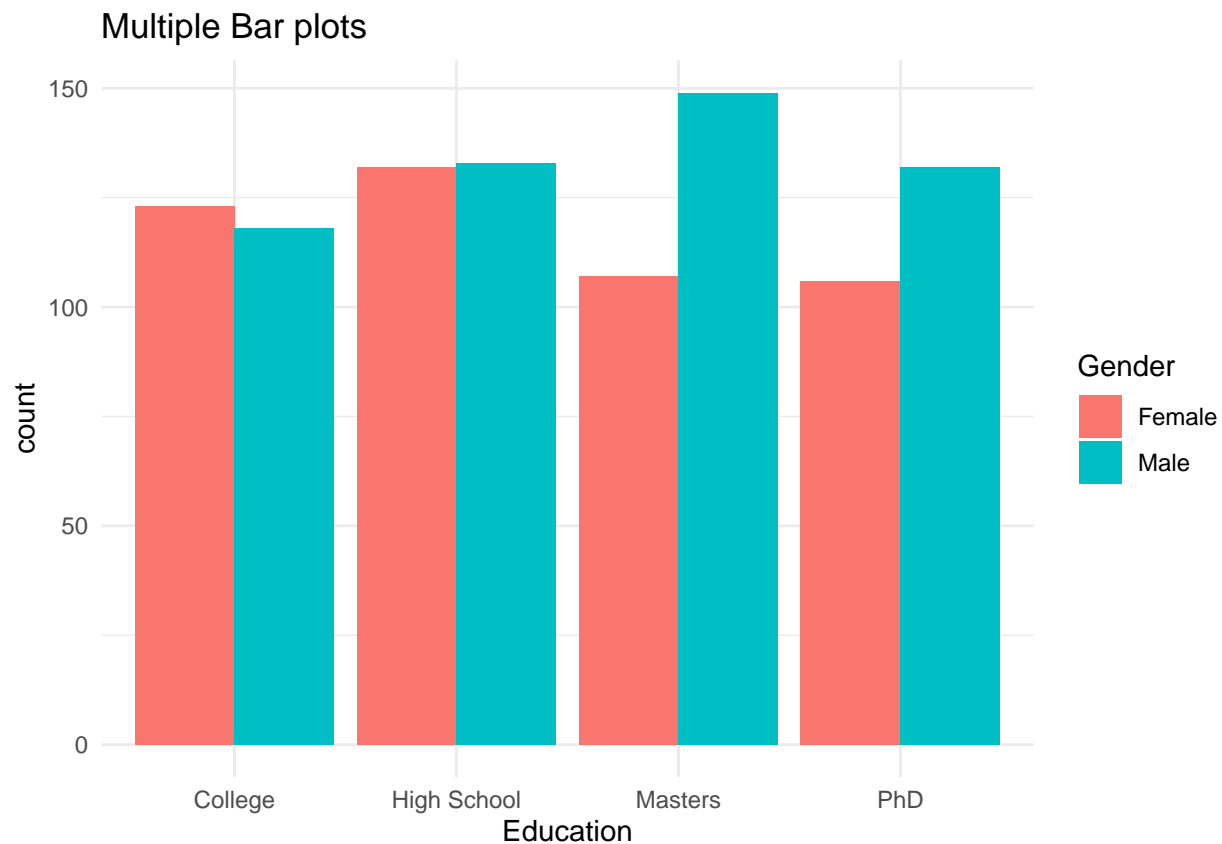
```
## # A tibble: 8 x 3
## # Groups:   Gender [2]
##   Gender Education count
##   <fct>   <fct>     <int>
## 1 Female    High School    12
## 2 Female    Bachelor's     10
## 3 Female    Master's         5
## 4 Male      High School    15
## 5 Male      Bachelor's     12
## 6 Male      Master's         8
```

```
## 1 Female College      123
## 2 Female High School  132
## 3 Female Masters      107
## 4 Female PhD          106
## 5 Male   College      118
## 6 Male   High School  133
## 7 Male   Masters      149
## 8 Male   PhD          132
```

```
library(dplyr)
new_glassdoor_df <- DATA2 %>% group_by(Gender, Education) %>% summarize(count = n())
```

```
## 'summarise()' has grouped output by 'Gender'. You can override using the
## '.groups' argument.
```

```
library(ggplot2)
ggplot(new_glassdoor_df, aes(Education, count, fill = Gender)) +
  geom_bar(stat="identity", position = 'dodge') +
  labs(title="Multiple Bar plots")
```



### DATA 3

#### Depression by gender

Regrettably, this is not a very good data to analyze this correlation. Based on many studies that I will discuss at the next step, about twice as many women as men experience depression (Staff, n.d.)

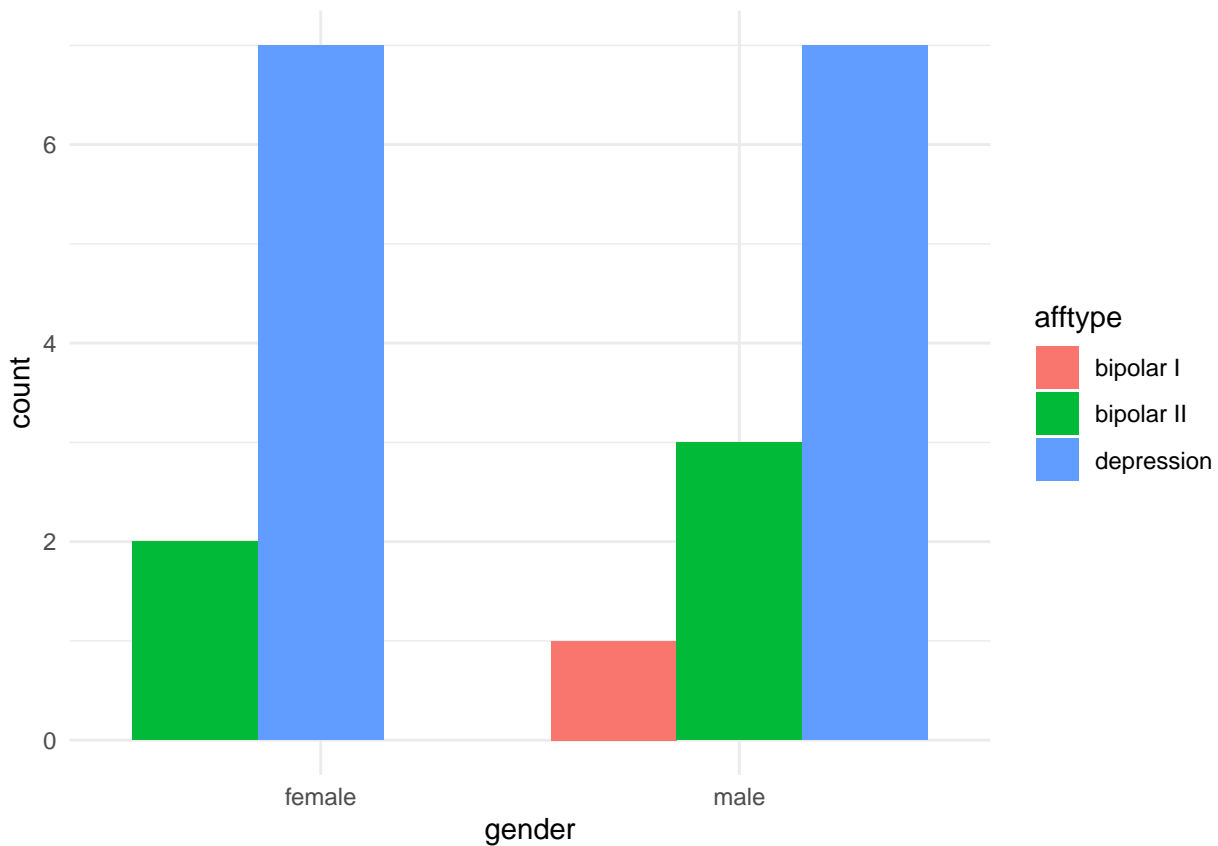
“The prevalence of major depression is higher in women than in men;6,7 in 2010 its global annual prevalence was 5.5% and 3.2%, respectively, representing a 1.7-fold greater incidence in women.” (Paul R. Albert, n.d.)

```
library(ggplot2)

newdata3[newdata3$gender==1, 'gender'] <- "female"
newdata3[newdata3$gender==2, 'gender'] <- "male"

newdata3[newdata3$afftype==1, 'afftype'] <- "bipolar II"
newdata3[newdata3$afftype==2, 'afftype'] <- "depression"
newdata3[newdata3$afftype==3, 'afftype'] <- "bipolar I"

ggplot(newdata3,
       aes(x = gender,
           fill = afftype)) +
  geom_bar(position = position_dodge(preserve = "single"))
```



Marriage x Depression

```
library(ggplot2)

newdata3[newdata3$marriage==1, 'marriage'] <- "married"
newdata3[newdata3$marriage==2, 'marriage'] <- "single"

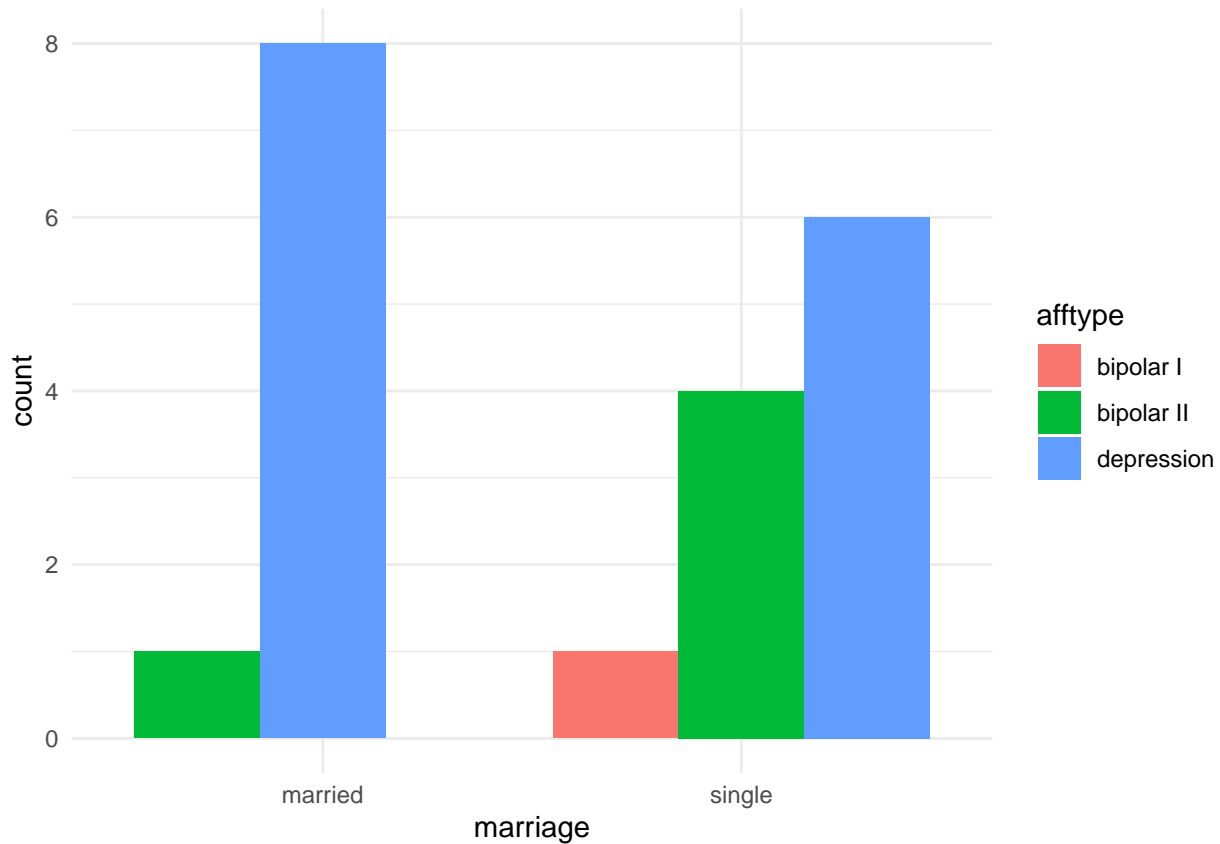
newdata3[newdata3$afftype==1, 'afftype'] <- "bipolar II"
newdata3[newdata3$afftype==2, 'afftype'] <- "depression"
newdata3[newdata3$afftype==3, 'afftype'] <- "bipolar I"

ggplot(newdata3,
```

```

aes(x = marriage,
    fill = afftype)) +
geom_bar(position = position_dodge(preserve = "single"))

```



Work x Depression

```

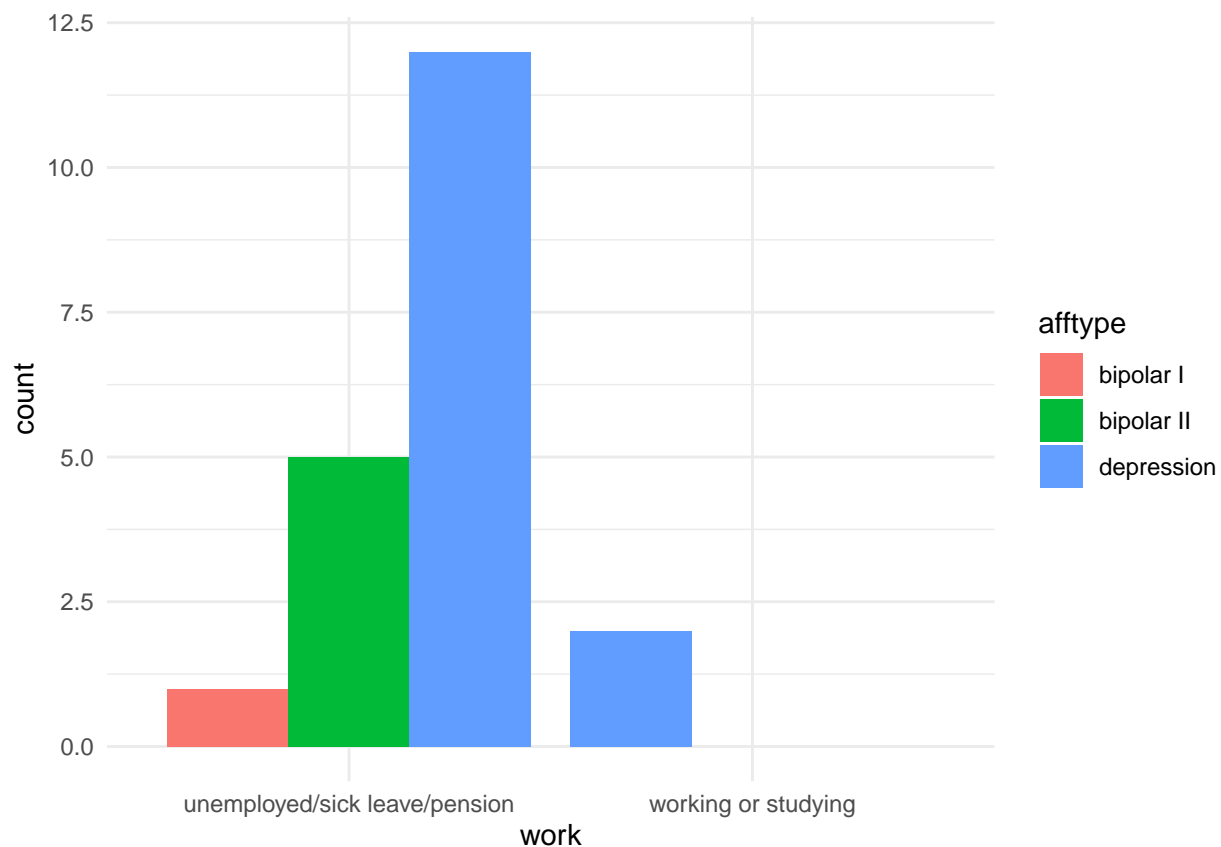
library(ggplot2)

newdata3[newdata3$work==1, 'work'] <- "working or studying"
newdata3[newdata3$work==2, 'work'] <- "unemployed/sick leave/pension"

newdata3[newdata3$afftype==1, 'afftype'] <- "bipolar II"
newdata3[newdata3$afftype==2, 'afftype'] <- "depression"
newdata3[newdata3$afftype==3, 'afftype'] <- "bipolar I"

ggplot(newdata3,
    aes(x = work,
        fill = afftype)) +
geom_bar(position = position_dodge(preserve = "single"))

```



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

Machine learning.

**Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.**

I might. I will have to wait until next week to see if it's applicable. I am interested in learning and applying Nearest Neighbors Classification, K-Means Clustering

## Future Steps

I won't be adding new data, but I will be adding new research information to discuss about depression and gender. While women are more depressed than men, there are questions that came up, such as, how often are men questioned about depression symptoms, are the symptoms the same? Do men go to primary care or doctor visits in general as often as women?

## #References

Paul R. Albert, PhD. n.d. "Why Is Depression More Prevalent in Women?"

Staff, By Mayo Clinic. n.d. "Depression in Women: Understanding the Gender Gap."