EDA_dgonzales2

Danielle Gonzales

2022-09-19

NOTES_DELETE AFTER fig height and width to make plots smaller this is “italicized” and this is ““bold””

Exploratory Data Analysis

Depression

Introduction

Pathway- /Users/daniellegonzales/Desktop/math130/Data/depress_081217.txt

I will be exploring the Depression data set specifically looking at two three variables, age, income, and sex. My research question reads; ‘is there a correlation between the age, sex, and income of inviduals with depression in this data set?’ I hypothesize that younger people with lower income will make up a larger portion of those with depression in this data set. I also anticipate to see more indiviuals with the sex female to have depression in this data set.

```
Depression <- read.table("/Users/daniellegonzales/Desktop/math130/Data/depress_081217.txt", header=TRUE
str(Depression)
```

```
## 'data.frame':    294 obs. of  37 variables:
##  $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ sex     : int  1 0 1 1 1 0 1 0 1 0 ...
##  $ age     : int  68 58 45 50 33 24 58 22 47 30 ...
##  $ marital : chr  "Widowed" "Divorced" "Married" "Divorced" ...
##  $ educat  : chr  "Some HS" "Some college" "HS Grad" "HS Grad" ...
##  $ employ  : chr  "Retired" "FT" "FT" "Unemp" ...
##  $ income  : int  4 15 28 9 35 11 11 9 23 35 ...
##  $ relig   : int  1 1 1 1 1 1 1 1 2 4 ...
##  $ c1      : int  0 0 0 0 0 0 2 0 0 0 ...
##  $ c2      : int  0 0 0 0 0 0 1 1 1 0 ...
##  $ c3      : int  0 1 0 0 0 0 1 2 1 0 ...
##  $ c4      : int  0 0 0 0 0 0 2 0 0 0 ...
##  $ c5      : int  0 0 1 1 0 0 1 2 0 0 ...
##  $ c6      : int  0 0 0 1 0 0 0 1 3 0 ...
##  $ c7      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ c8      : int  0 0 0 3 3 0 2 0 0 0 ...
##  $ c9      : int  0 0 0 0 3 1 2 0 0 0 ...
##  $ c10     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ c11     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ c12     : int  0 1 0 0 0 1 0 0 3 0 ...
```

```
##  $ c13     : int  0 0 0 0 0 2 0 0 0 0 ...
##  $ c14     : int  0 0 1 0 0 0 0 0 3 0 ...
##  $ c15     : int  0 1 1 0 0 0 3 0 2 0 ...
##  $ c16     : int  0 0 1 0 0 2 0 1 3 0 ...
##  $ c17     : int  0 1 0 0 0 1 0 1 0 0 ...
##  $ c18     : int  0 0 0 0 0 0 0 0 1 0 0 ...
##  $ c19     : int  0 0 0 0 0 0 0 0 1 0 0 ...
##  $ c20     : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ cesd    : int  0 4 4 5 6 7 15 10 16 0 ...
##  $ cases   : int  0 0 0 0 0 0 0 0 0 1 0 ...
##  $ drink   : int  0 1 1 0 1 1 0 0 1 1 ...
##  $ health  : int  2 1 2 1 1 1 3 1 4 1 ...
##  $ regdoc  : int  1 1 1 1 1 1 1 0 1 1 ...
##  $ treat   : int  1 1 1 0 1 1 1 0 1 0 ...
##  $ beddays : int  0 0 0 0 1 0 0 0 1 0 ...
##  $ acuteill: int  0 0 0 0 1 1 1 1 0 0 ...
##  $ chronill: int  1 1 0 1 0 1 1 0 1 0 ...
```

This data set explores depression and its relation to other variables in an affected person's life.

```
head(Depression)
```

```
##   id sex age   marital       educat  employ income relig c1 c2 c3 c4 c5 c6 c7
## 1  1   1  68   Widowed      Some HS Retired      4     1  0  0  0  0  0  0  0
## 2  2   0  58  Divorced Some college      FT     15     1  0  0  1  0  0  0  0
## 3  3   1  45   Married      HS Grad      FT     28     1  0  0  0  0  1  0  0
## 4  4   1  50  Divorced      HS Grad   Unemp      9     1  0  0  0  0  1  1  0
## 5  5   1  33 Separated      HS Grad      FT     35     1  0  0  0  0  0  0  0
## 6  6   0  24   Married      HS Grad      FT     11     1  0  0  0  0  0  0  0
##   c8 c9 c10 c11 c12 c13 c14 c15 c16 c17 c18 c19 c20 cesd cases drink health
## 1  0  0   0   0   0   0   0   0   0   0   0   0   0    0     0     0      2
## 2  0  0   0   0   1   0   0   1   0   1   0   0   0    4     0     1      1
## 3  0  0   0   0   0   0   1   1   1   0   0   0   0    4     0     1      2
## 4  3  0   0   0   0   0   0   0   0   0   0   0   0    5     0     0      1
## 5  3  3   0   0   0   0   0   0   0   0   0   0   0    6     0     1      1
## 6  0  1   0   0   1   2   0   0   2   1   0   0   0    7     0     1      1
##   regdoc treat beddays acuteill chronill
## 1      1     1       0        0        1
## 2      1     1       0        0        1
## 3      1     1       0        0        0
## 4      1     0       0        0        1
## 5      1     1       1        1        0
## 6      1     1       0        1        1
```

Univariate Exploration

Age and Depression

```
summary(Depression$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   28.00   42.50   44.41   59.00   89.00
```

```
mean(Depression$age)
```

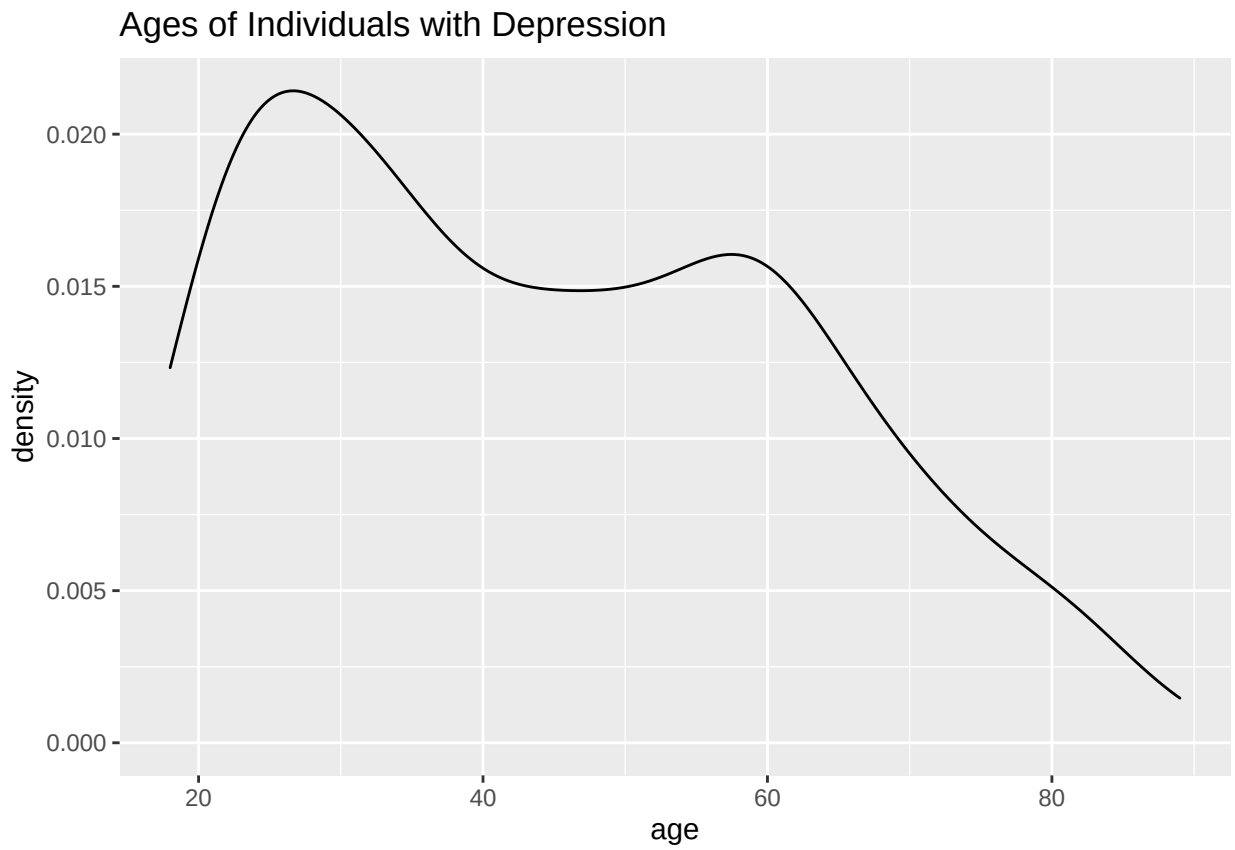
```
## [1] 44.41497
```

```
median(Depression$age)
```

```
## [1] 42.5
```

The data set displays the ages of those with depression involved in this survey, the youngest individual with depression at 18 years old and the oldest at 89 years old. Both the summary function and the mean function displayed the same mean age (44.41 years) of all individuals in the data set. Additionally, both the summary function and the median function provided the same median age of all individuals in the data set (42.5 years).

```
ggplot(data = Depression, aes(age)) +
     geom_density() +
     ggtitle("Ages of Individuals with Depression")
```



I used a bar chart to visually show the ages of people with depression in this survey and the amount of people in this age group. Interestingly, the majority of depression cases are among younger people, but there is a noticeable spike at the age of 60.

Income and Depression

```
summary(Depression$income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    9.00   15.00   20.57   28.00   65.00
```

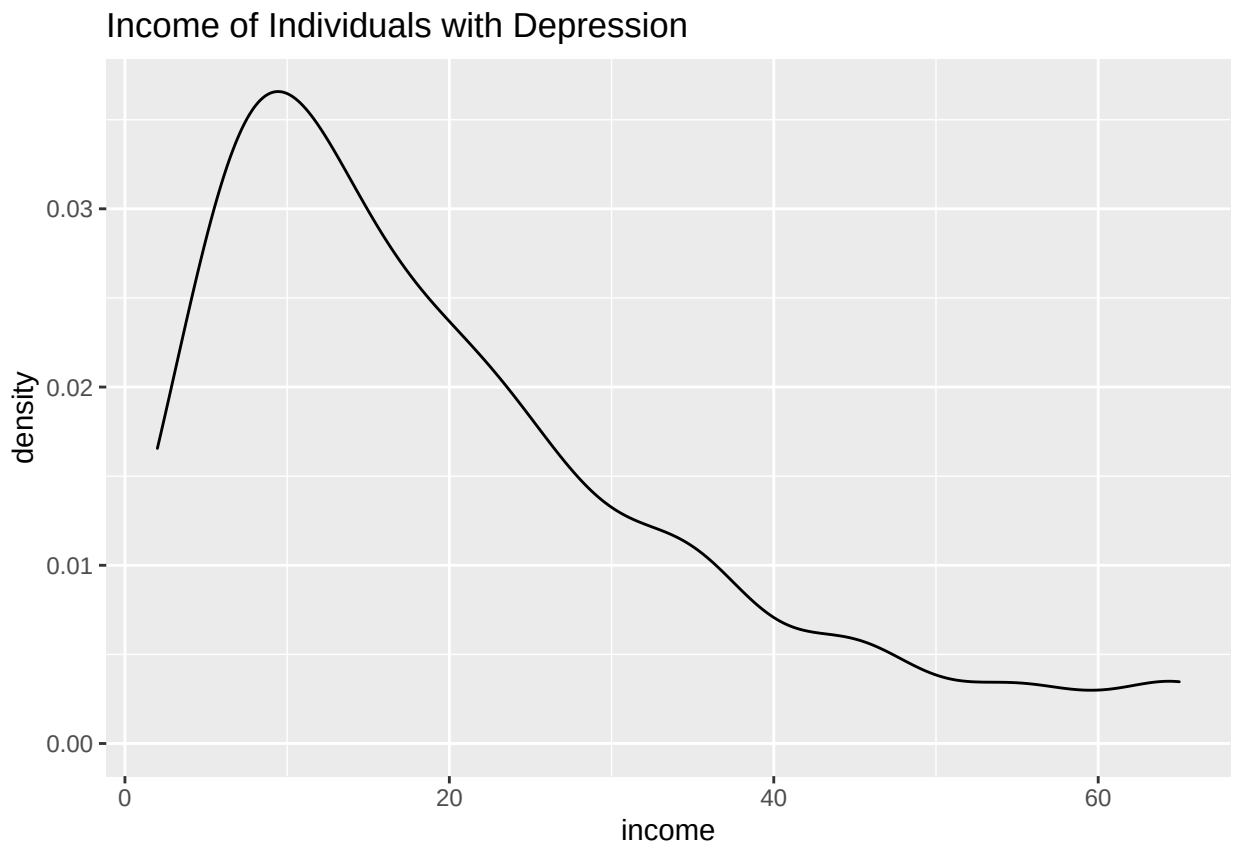
```
mean(Depression$income)
```

```
## [1] 20.57483
```

```
median(Depression$income)
```

```
## [1] 15
```

```
ggplot(data = Depression, aes(income)) +
     geom_density() +
     ggtitle("Income of Individuals with Depression")
```



Based on the overall layout of this density plot, there is a higher amount of individuals suffering from depression that have lower income.

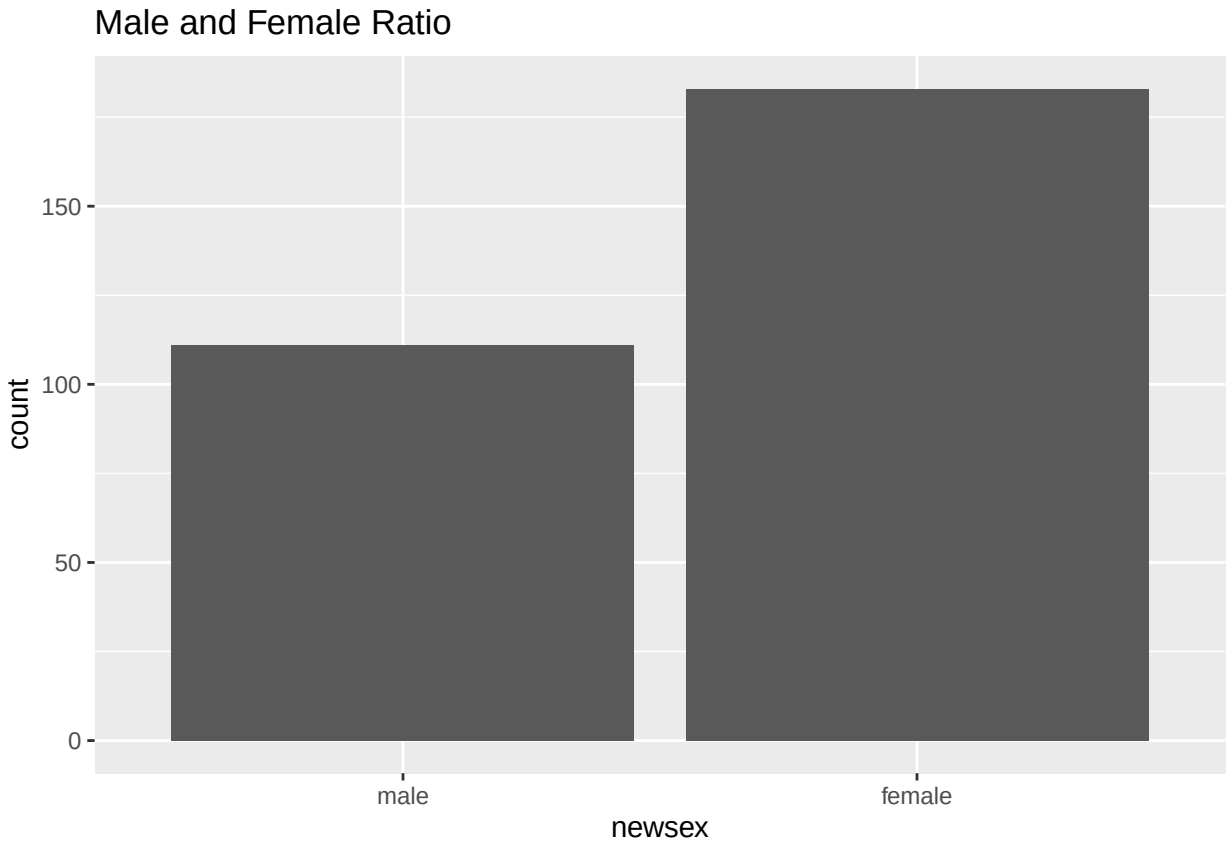
Sex and Depression

First, I renamed the variables to be clearer for the reader. In this data set, 1=male and 0=female. Now, the variable "sex" will be under the name "newsex" in my code.

```
Depression$newsex <- factor(Depression$sex, labels=c( "male", "female"))
table(Depression$sex, Depression$newsex, useNA="always")
```

```
##
##         male female <NA>
##   0      111      0    0
##   1        0    183    0
##   <NA>     0      0    0
```

```
ggplot(data = Depression, aes(newsex)) +
    geom_bar() +
    labs(title = "Male and Female Ratio")
```



Similar to my prediction, there appears to be a greater amount of female individuals with depression based on this plot.

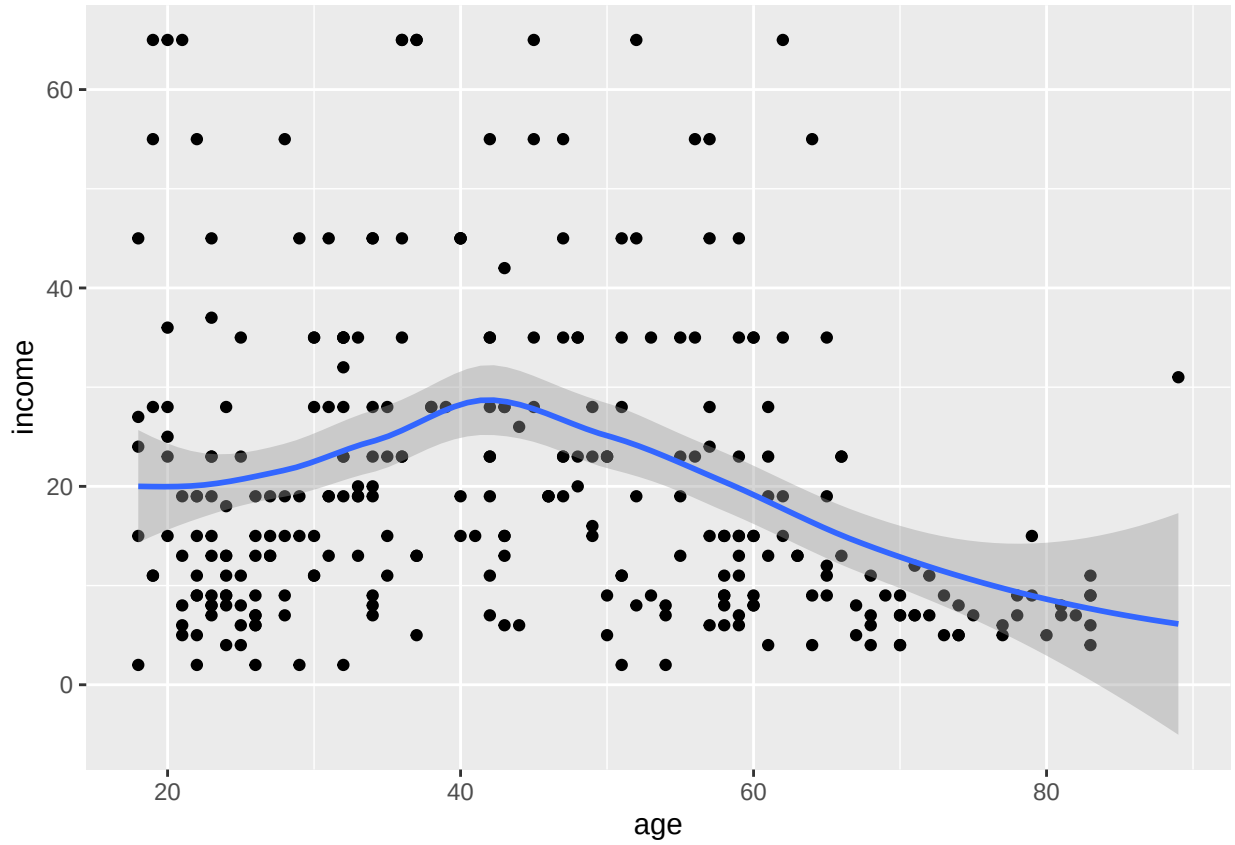
Bivariate Exploration

First, I will compare ages with income of those with depression in this study. I expect to see an increased amount of people with depression with lower income.

Age, Income, and Depression

```
ggplot(Depression, aes(x=age, y=income)) + geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

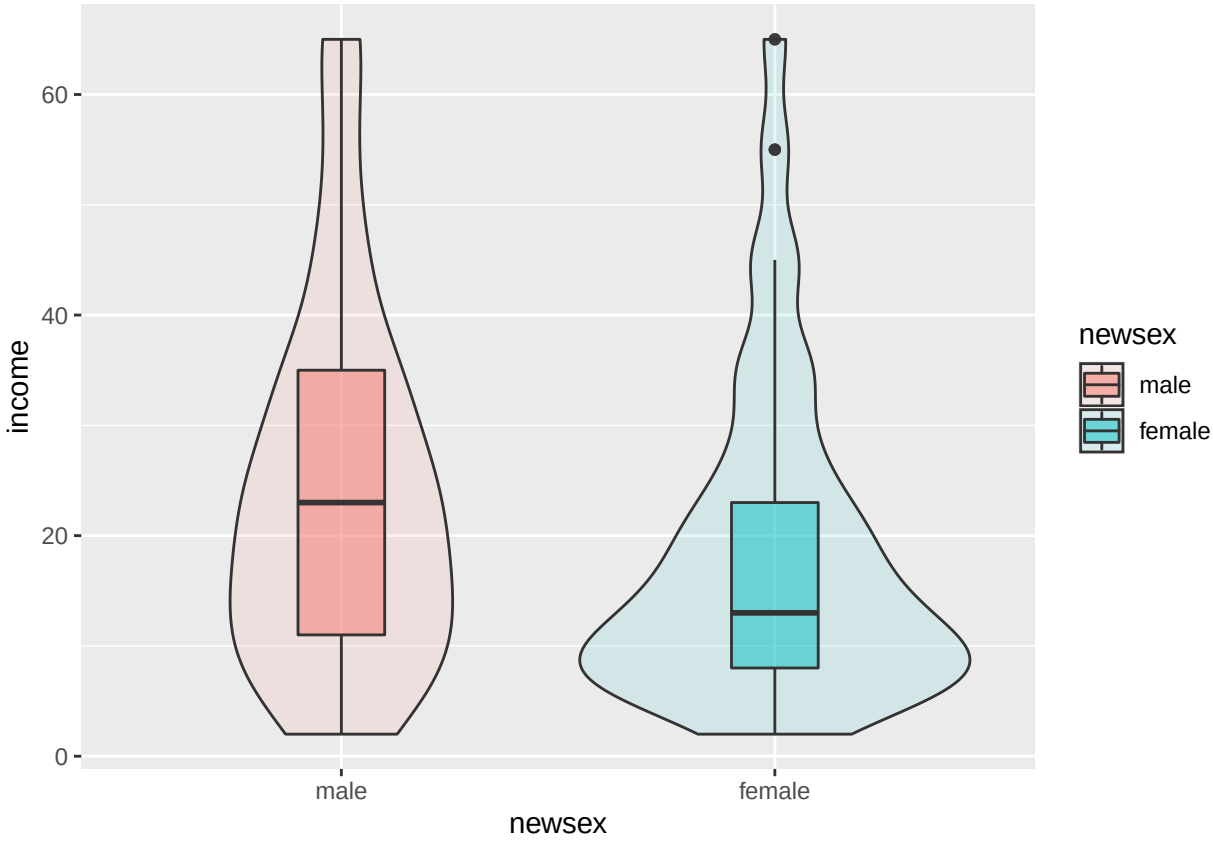


Because both age and income are continuous variables, I used a scatterplot and added a line of best fit to make the data more readable.

Based on this plot, much of the data is skewed to the left, meaning the majority of those with depression are younger and with lower income. There also appears to be a trend of middle aged individuals with middle income and depression.

Income, Sex, and Depression

```
ggplot(Depression, aes(x=newsex, y=income, fill=newsex)) +
        geom_violin(alpha=.1) +
        geom_boxplot(alpha=.5, width=.2)
```

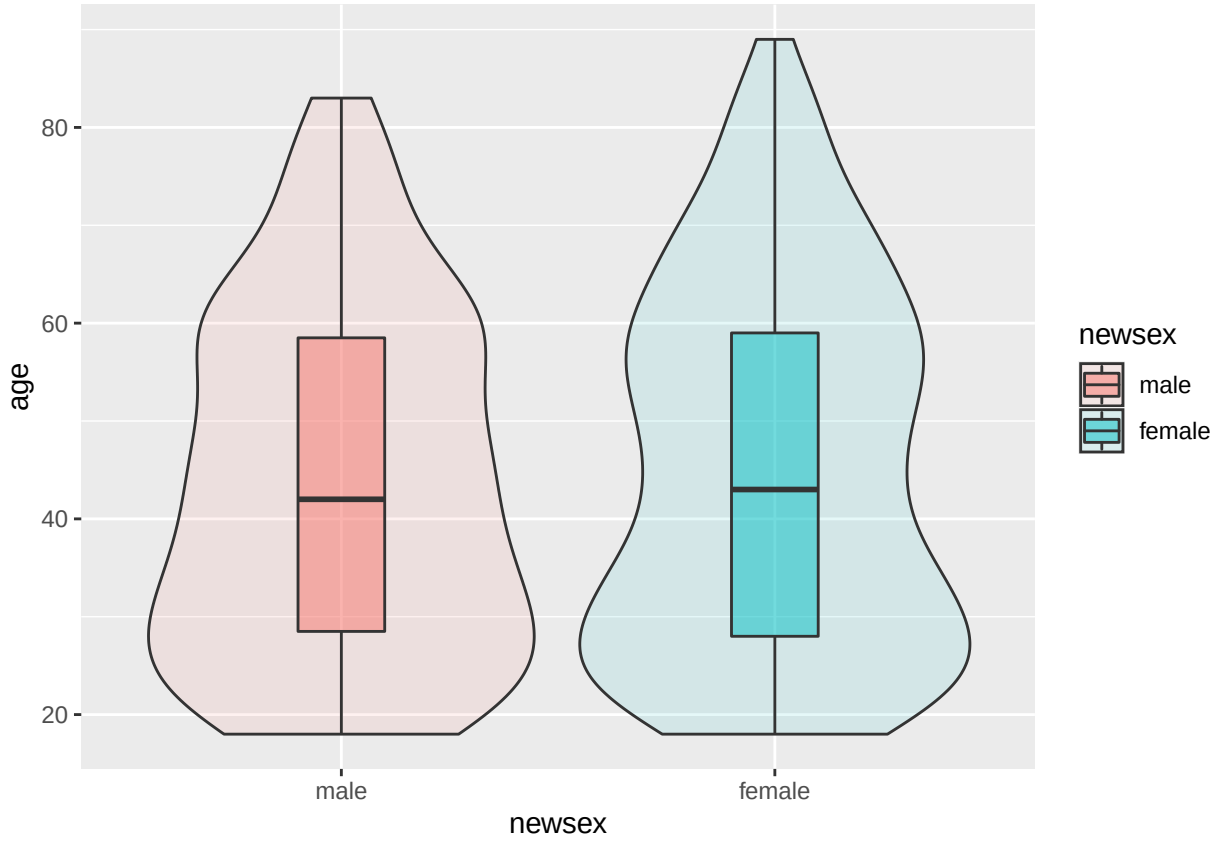


When comparing a continuous and categorical variable I used a box plot.

From this plot, it appears that the majority of female individuals with depression have low income, and male individuals have low to middle income.

Age, Sex, and Depression

```
ggplot(Depression, aes(x=newsex, y=age, fill=newsex)) +
        geom_violin(alpha=.1) +
        geom_boxplot(alpha=.5, width=.2)
```



In both male and female individuals, the majority of the participants were from the ages 30-60. Interestingly both boxes are almost exactly the same.

Conclusion

Overall, the plots and data analysis supports my hypothesis that there is a correlation between females and depression, young people and depression, and low income and depression. One area that could be studied further was the interesting spike in depression around the age of 60, so a study on depression and the elderly could have interesting insight.