Final Project, Depression Data Set

Lily Bergeron

2023-09-21

Introduction:

For the final project, I selected the Depression data set. It shows a study of depression levels among 294 male and female adult residents in L.A.. Within this data set, there are 37 variables. I initially studied three variables, but drew my final conclusion from the data set based on only two. I studied "age," "cesd," and "sex," but only used "cesd" and "sex" in my conclusion and for my bivariate study.

```
depress <- read.delim("C:\\Users\\lilyb\\OneDrive\\Desktop\\Math130\\data\\depress_081217.txt",
    header = TRUE, sep = "\t")
library(ggplot2)
```

Univariate Exploration:

```
summary(depress$age)
```

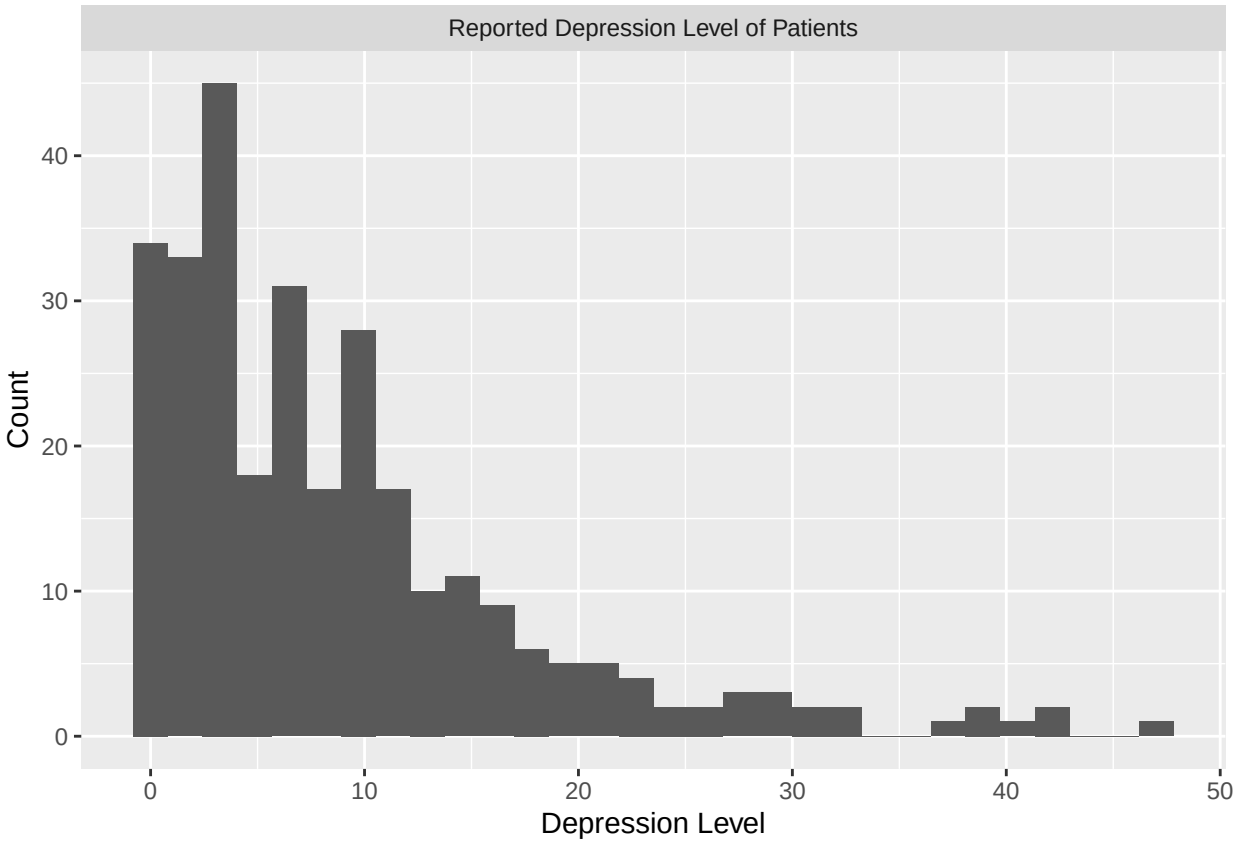
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   28.00   42.50   44.41   59.00   89.00
```

```
summary(depress$cesd)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.000   7.000   8.884  12.000  47.000
```

The first table shows a summary of age within the depression data set. The second table is a summary of depression levels within the same data set. For both, the mean is greater than that of the median. Therefore, the graphs will be skewed to the right. The youngest person within the study is 18 years old, and the oldest is 89 years old. The average age is 44.41, the 25th percentile is 28 or younger, and the 75th percentile is 59 or younger. The lowest depression level of participants is 0, while the highest is 47. The average among patients is 8.884, the 25th percentile had a score of 3 or lower, and the 75th percentile had a score of 12 or less.

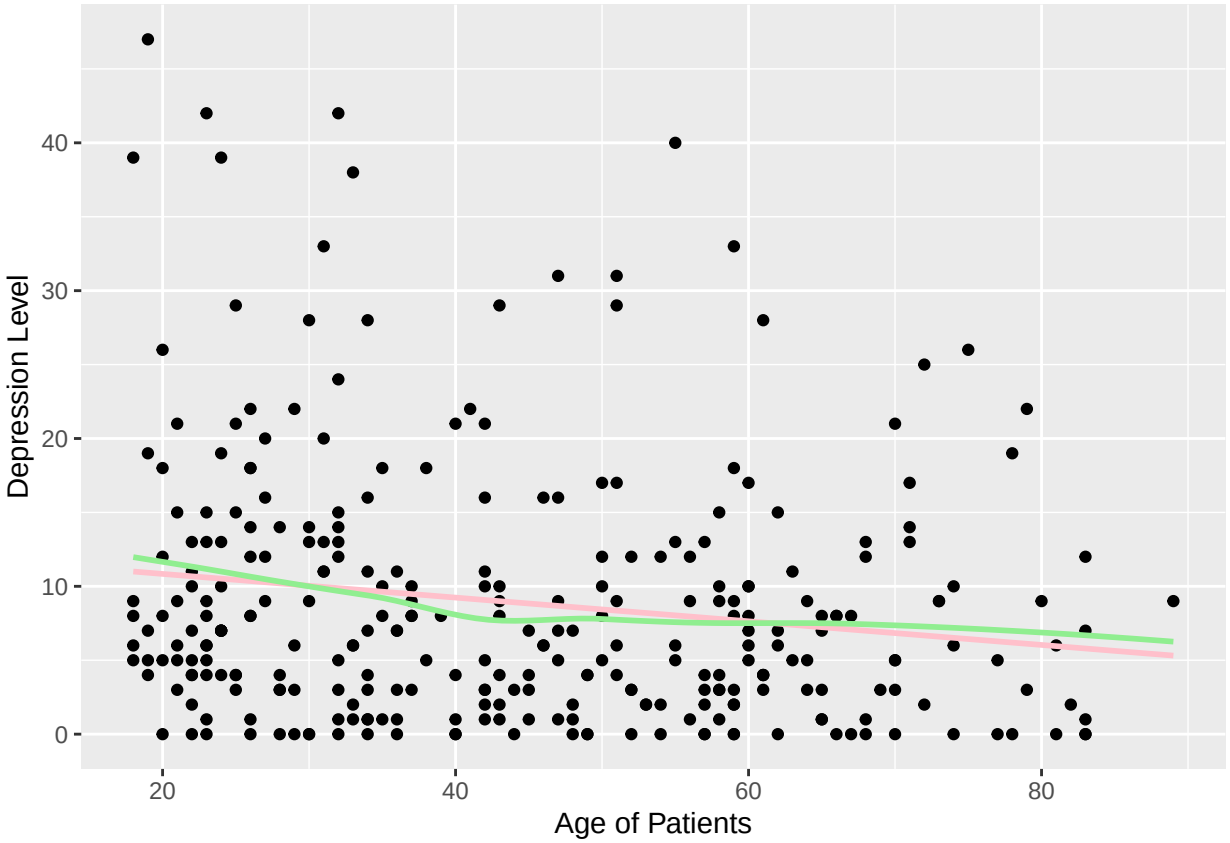
```
ggplot(depress, aes(x = cesd)) + geom_histogram() + ylab("Count") + xlab("Depression Level") +
    facet_wrap("scales=\"Reported Depression Level of Patients\"")
```




```
ggplot(depress, aes(x = age)) + geom_histogram() + ylab("Count") + xlab("Age") +
    facet_wrap("scales=\"Age of Patients Within Study\"")
```

The first graph displays the frequency distribution of depression levels of patients, while the second illustrates the frequency distribution of the age of patients within the study. Both are skewed to the right, as predicted by the generated tables; however, age had a much smaller skew than depression levels. For this reason, I will expand further upon the age variable in order to see if there is a significant relationship.

```
ggplot(depress, aes(x = age, y = cesd)) + geom_point() + geom_smooth(se = FALSE,
    method = "lm", color = "pink") + geom_smooth(se = FALSE, color = "lightgreen") +
    xlab("Age of Patients") + ylab("Depression Level")
```

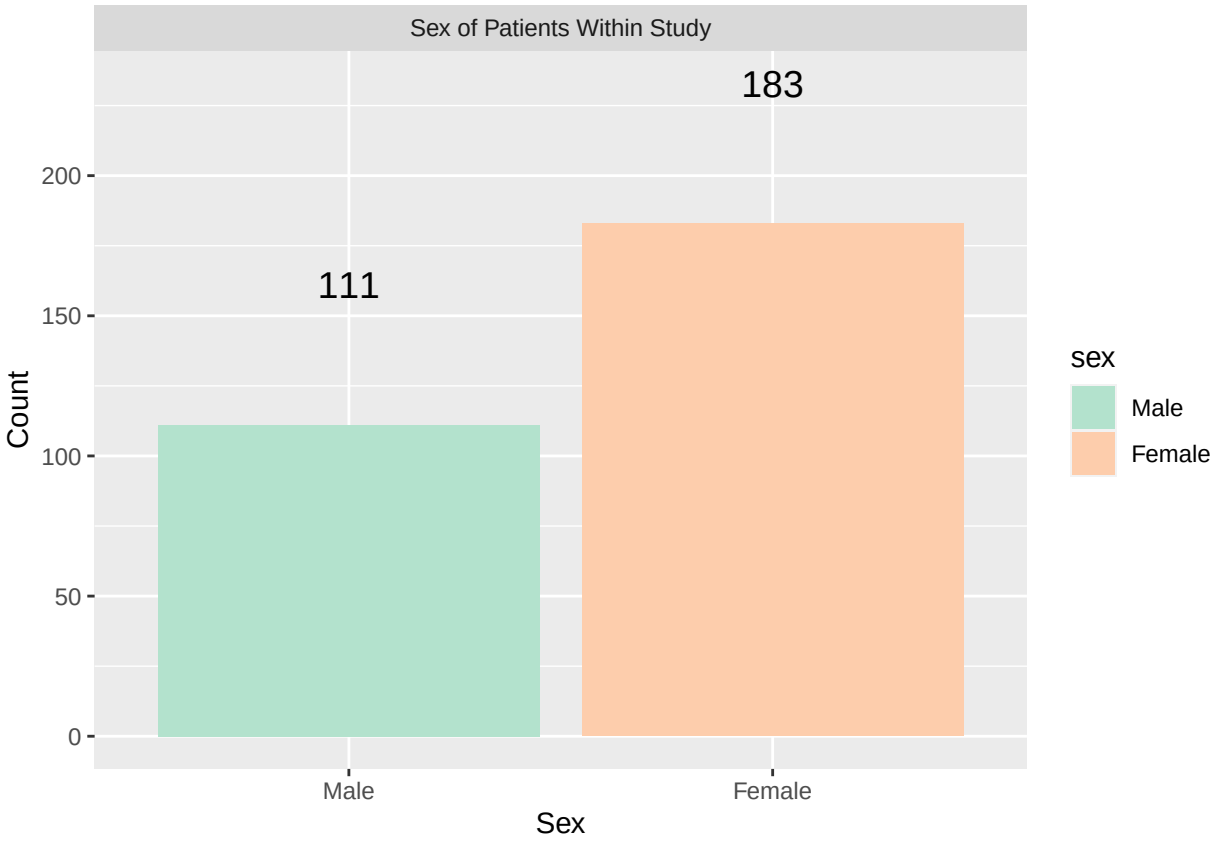


The above scatterplot, with a line of best fit and lowess line, shows that there is not a significant relationship between age and depression levels in patients. Both lines are nearly overlapping and with a very small negative relationship. Because there is no significant relationship, I will now look further into the sex variable as opposed to the age variable. Depression levels of patients will continue to be studied.

```
depress$sex <- factor(depress$sex, labels = c("Male", "Female"))
table(depress$sex)
```

```
##
##   Male Female
##    111    183
```

```
ggplot(depress, aes(sex, fill = sex)) + geom_bar() + ylab("Count") + xlab("Sex") +
    facet_wrap("scales=\"Sex of Patients Within Study\"") + geom_text(aes(y = ..count.. +
    50, label = ..count..), stat = "count", size = 5) + scale_fill_brewer(palette = "Pastel2")
```

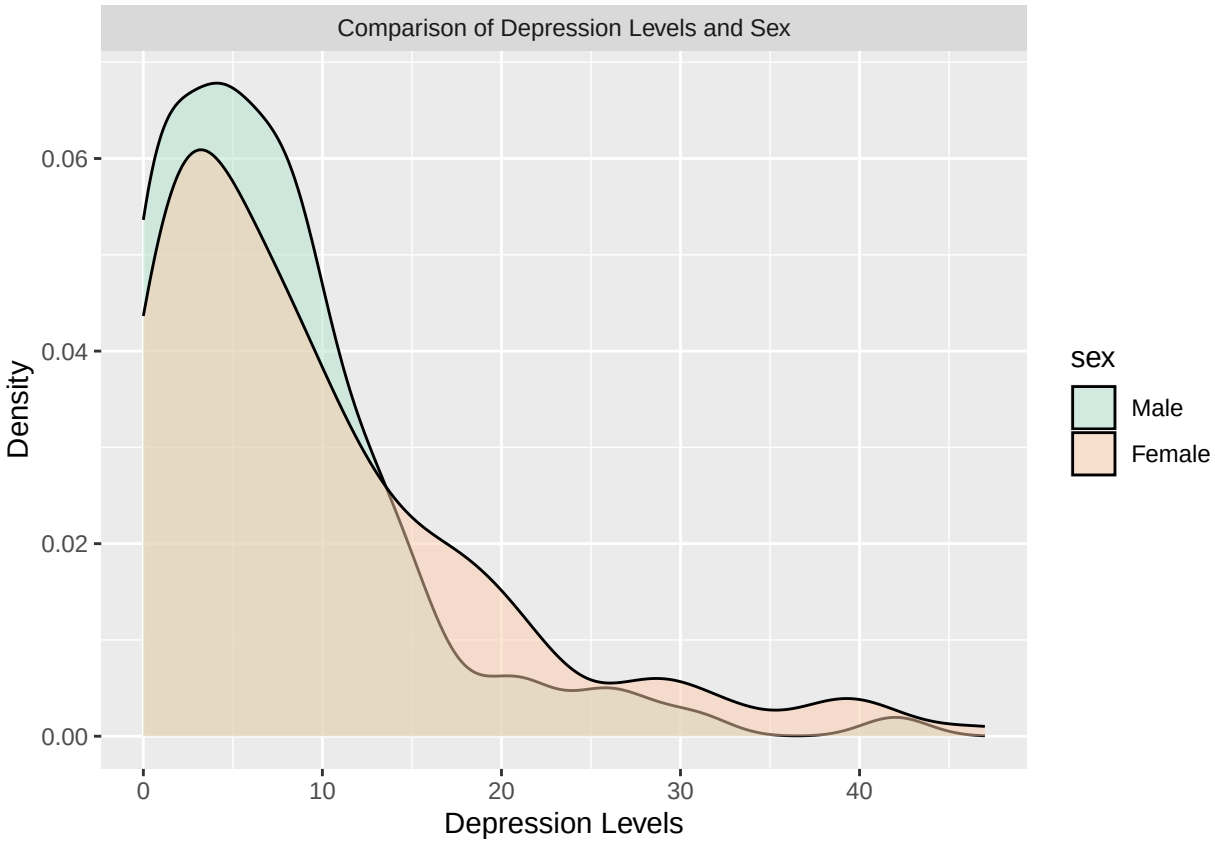


For this study, there are 111 Male patients and 183 Female patients. There are a total of 294 individuals in the study.

Bivariate Exploration:

I decided to explore more about the relationship between sex and depression levels, as age seemed to have little to no impact on depression levels when compared to sex.

```
ggplot(depress, aes(x = cesd, fill = sex)) + geom_density(alpha = 0.5) + scale_fill_discrete(name = "se
    scale_fill_brewer(palette = "Pastel2") + ylab("Density") + xlab("Depression Levels") +
    facet_wrap("scales=\"Comparison of Depression Levels and Sex\"")
```



The density graph above shows a relationship between depression levels in men versus women, with female score represented in orange and male scores in green. Overall, the graph illustrates that females have higher depression scores than men. Not only are high female scores more abundant, but there are a significantly higher amount of low male scores.

Conclusion: Initially, depression levels were compared to the variable of age. When it was noted that there was not a very significant relationship, the variable of sex replaced that of age. It was discovered from the data that depression levels are different between men and women, with women overall being more depressed than men.