Exploratory Data Analysis Project

Marisol Cisneros

2022-09-21

Introduction:

The data that I will be analyzing in this project is the depression data set. It contains 294 observations and 37 variables. The information from the data set comes from a study of depression done by Los Angeles County. The variables that I will be using will be sex, age, and education. My goal in this is to see what variable is more prone to depression.

```
depression <- read.table("C:/Users/sunli/OneDrive/Desktop/EDA_mcisneros1/depress_081217 (1).txt"
, header=TRUE, sep="\t")
library(ggplot2)
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
```

```
library(RColorBrewer)
```

Univariate Exploration:

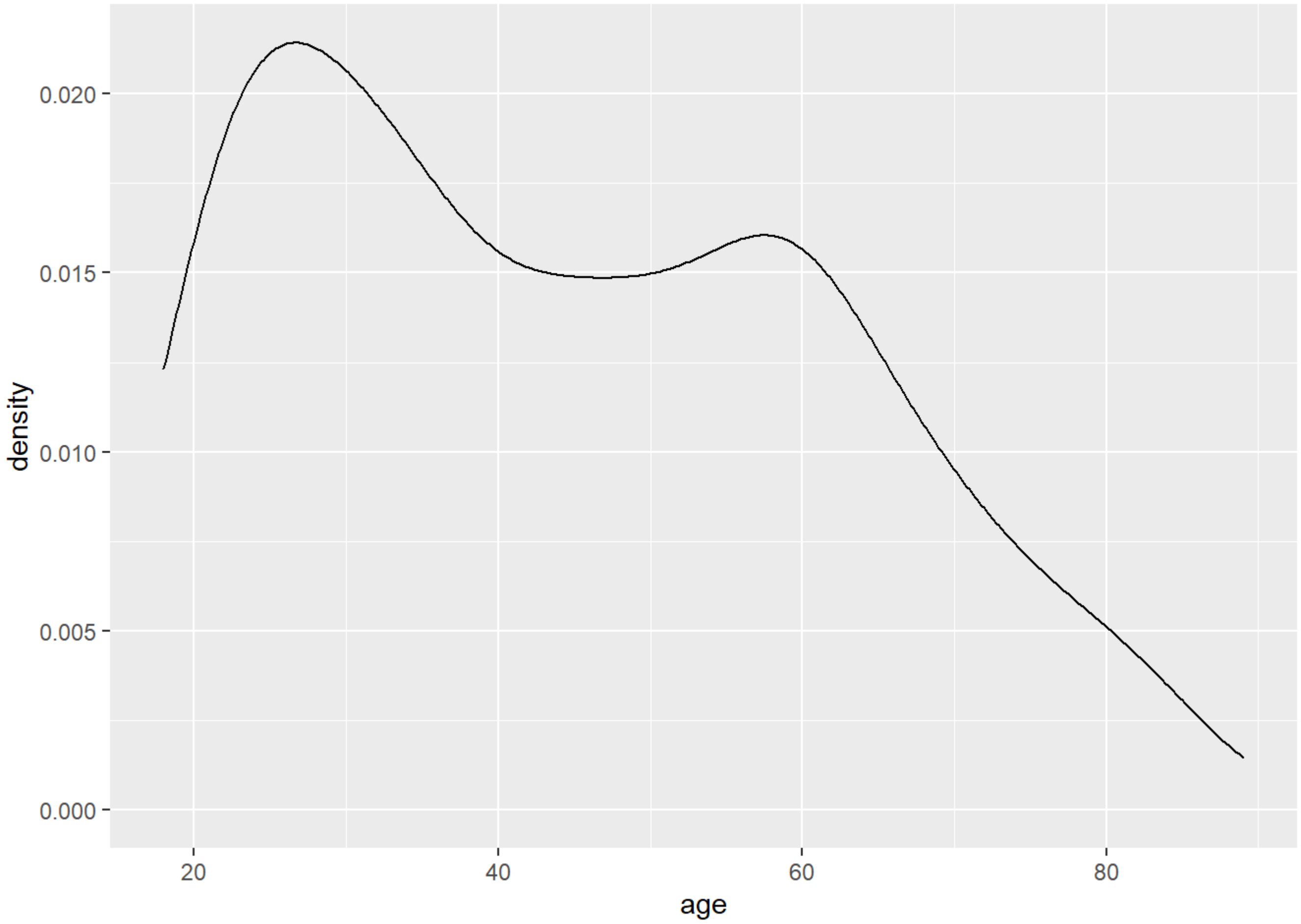
In this function, I will compare which sex is much more to depressed than the other one.

```
table(depression$sex)
```

```
##
##   0   1
## 111 183
```

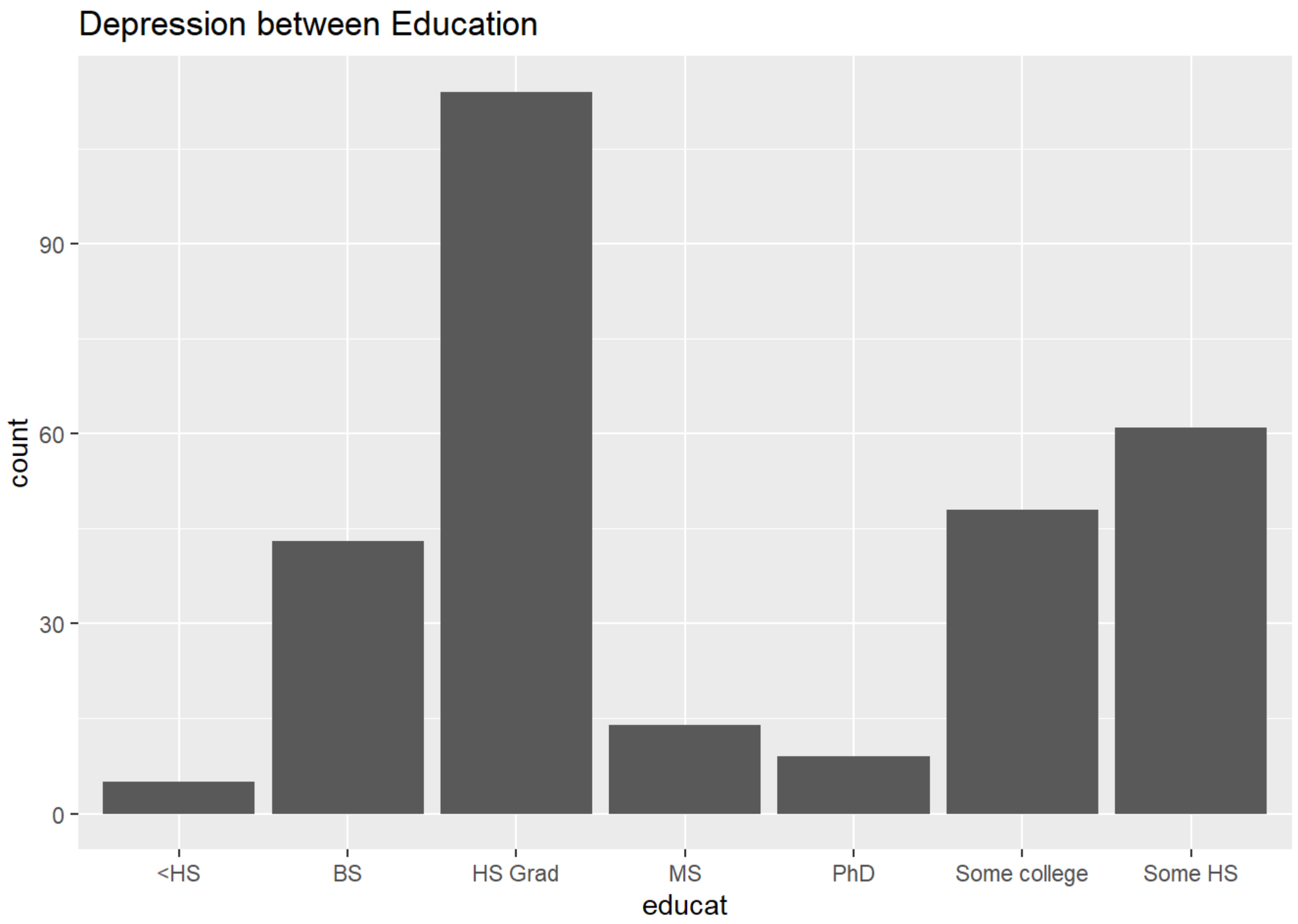
Based on the table made, males were the ones who experienced more depression out of the females who participated.

```
ggplot(depression, aes(x=age)) + geom_density()
```



Here we can see that the density plot shows the depression rates between the ages of 20 to 80. Based on the plot, the ages around 30 were more prone to depression than the other ones. After the age of 35, it would start to decrease as it reached 80.

```
ggplot(depression, aes(x=educat)) + geom_bar()+ggtitle("Depression between Education")
```



From the box plot, we are able to see the differences between the educations and how they relate to depression. From what we can see, the participants who are high school graduates were more prone to depression than the rest of the participants.

Biavariate Exploration:

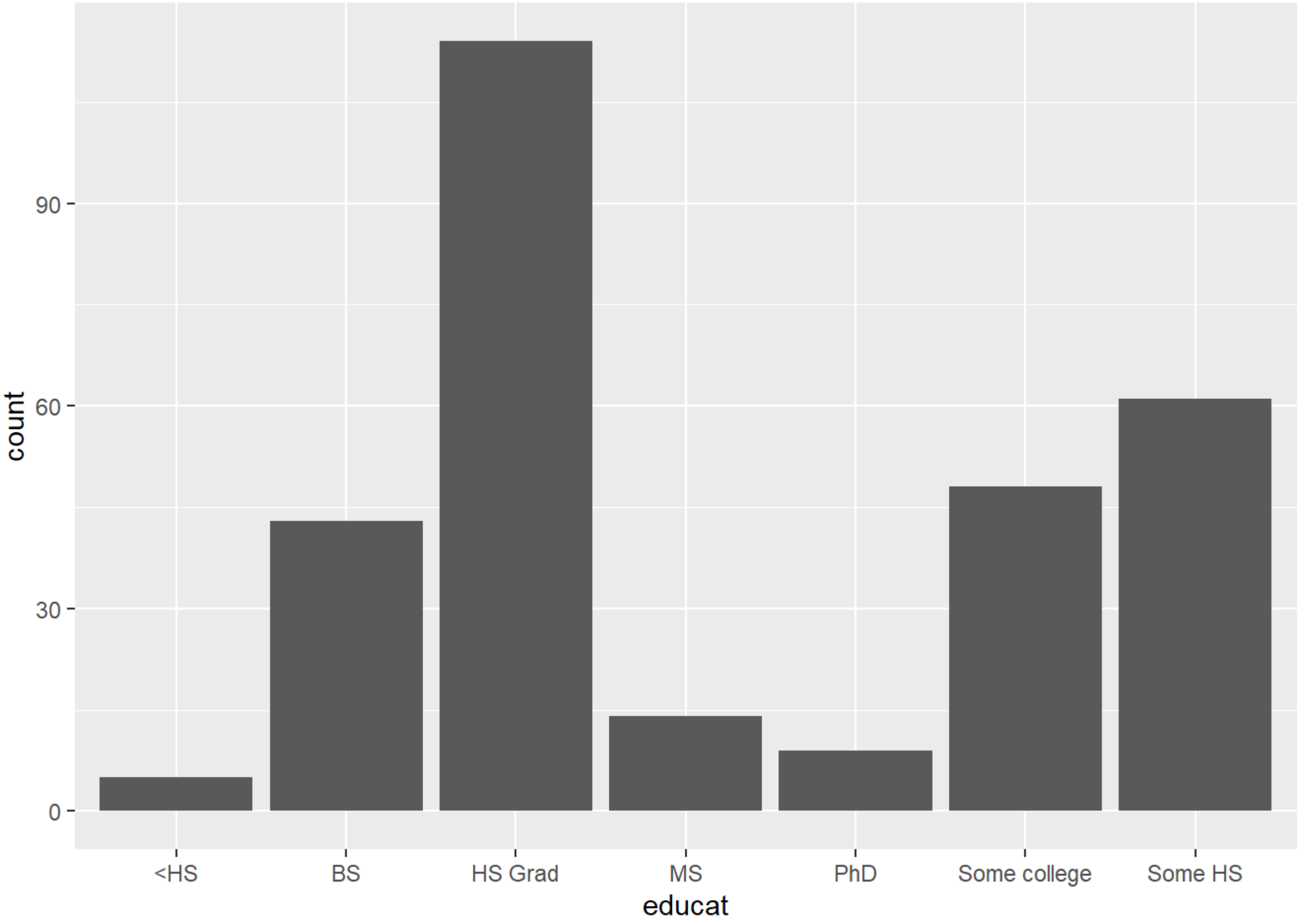
In this section, I will be focusing on the variables of education and sex to see if there are is any connection between and if they are more likely to be depressed that any other variable.

```
table(depression$educat, depression$sex)
```

```
##
##                  0  1
##   <HS            4  1
##   BS            17 26
##   HS Grad       39 75
##   MS             8  6
##   PhD            6  3
##   Some college 18 30
##   Some HS       19 42
```

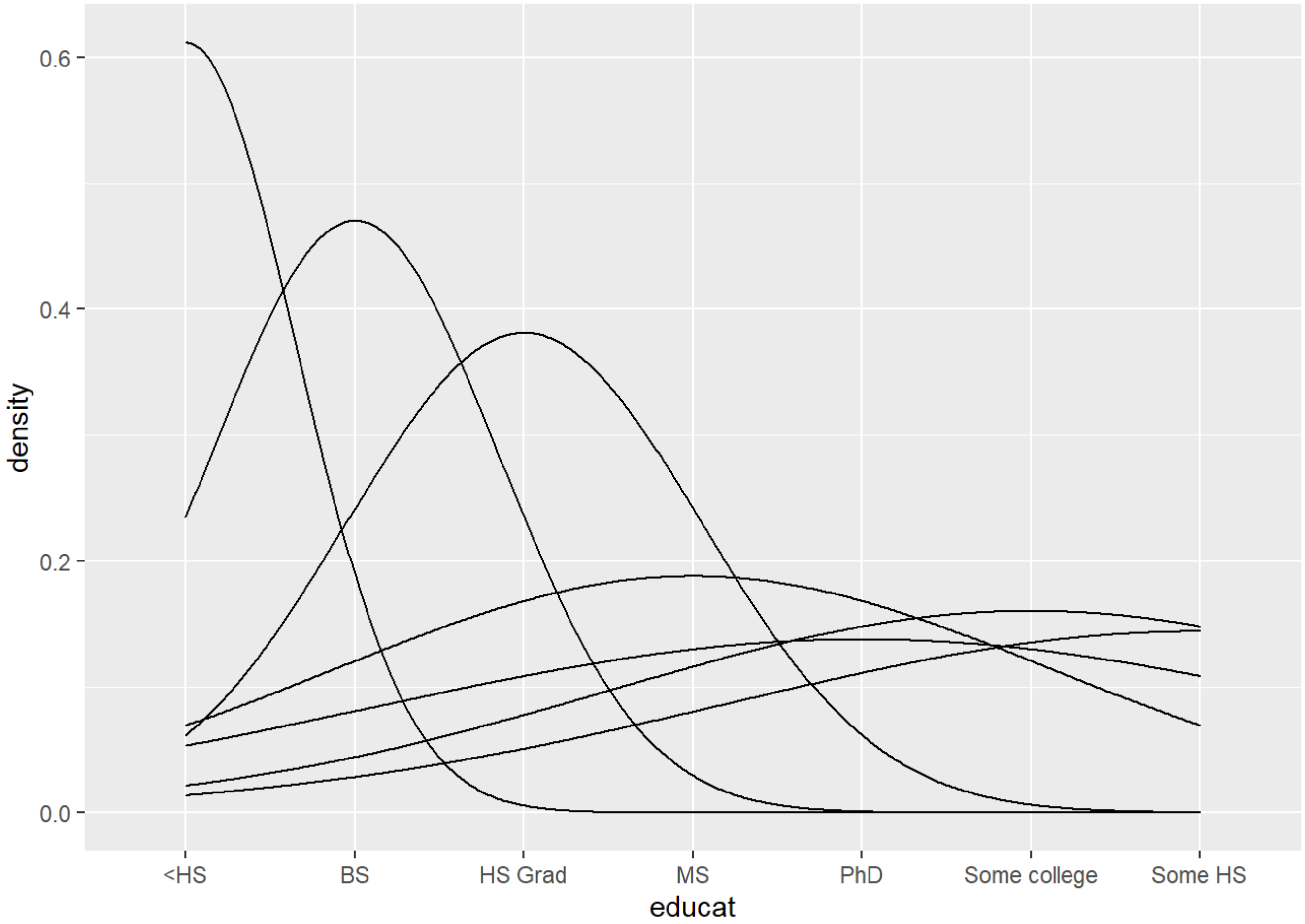
Based on what is shown in the table, it shows the number of participants and their education levels. From what can be seen, high school graduates have a higher number of depression but it can be mostly seen in male high school graduates which have a number of 75.

```
ggplot(depression, aes(x=educat, fill=sex)) + geom_bar()
```



In this grouped bar chart, it shows the education of male and female and it varries between education. The highest of them is from high school graduates which show that they were more prone to depression than the rest. Next from it is some high school education that showed depression.

```
ggplot(depression, aes(x=educat, fill=sex)) + geom_density(alpha=.3)
```



This overlaid density curve shows the educations and how high of a curve it goes. The highest starts as less than a high school education and gores from a BS to a high school graduate.

Conclusion

In conclusion, we can see that the three different chosen variables all had different outcomes. In the univariate, it can be seen that males were more prone to depression between the age of 30. The biavariate variable shows that males that were high school graduates show more depression than females. Based on the results of the project, it can be seen that males that had some high school education are more likely to develop depression than the rest of the variables.