EDA Project on Depression

Mikaela Wiley

2/22/2022

Introduction

I will be using the data set "Depression" which is from a Los Angeles County study on depression. In this data set there are 294 observations and 37 variables. Of those variables I will be using CESD, HEALTH, MARITAL and INCOME. The question that I'm asking is how does marital status, health, and income affect depression? Does one have more of an impact than the other? My hypothesis is that poor health will have the biggest impact on depression.

The CESD variable is the scores on the CESD test to see if the people being surveyed have depression. A score of 16 or higher is considered depressed.

The HEALTH variable asks for health status and has four options from 1-4: 1- Excellent health, 2-Good health, 3- Fair health, 4- Poor health.

The MARITAL variable asks for marital status and has five options: Never married, married, separated, divorced, and widowed.

The INCOME variable asks for yearly income. 0-65 in thousands of dollars.

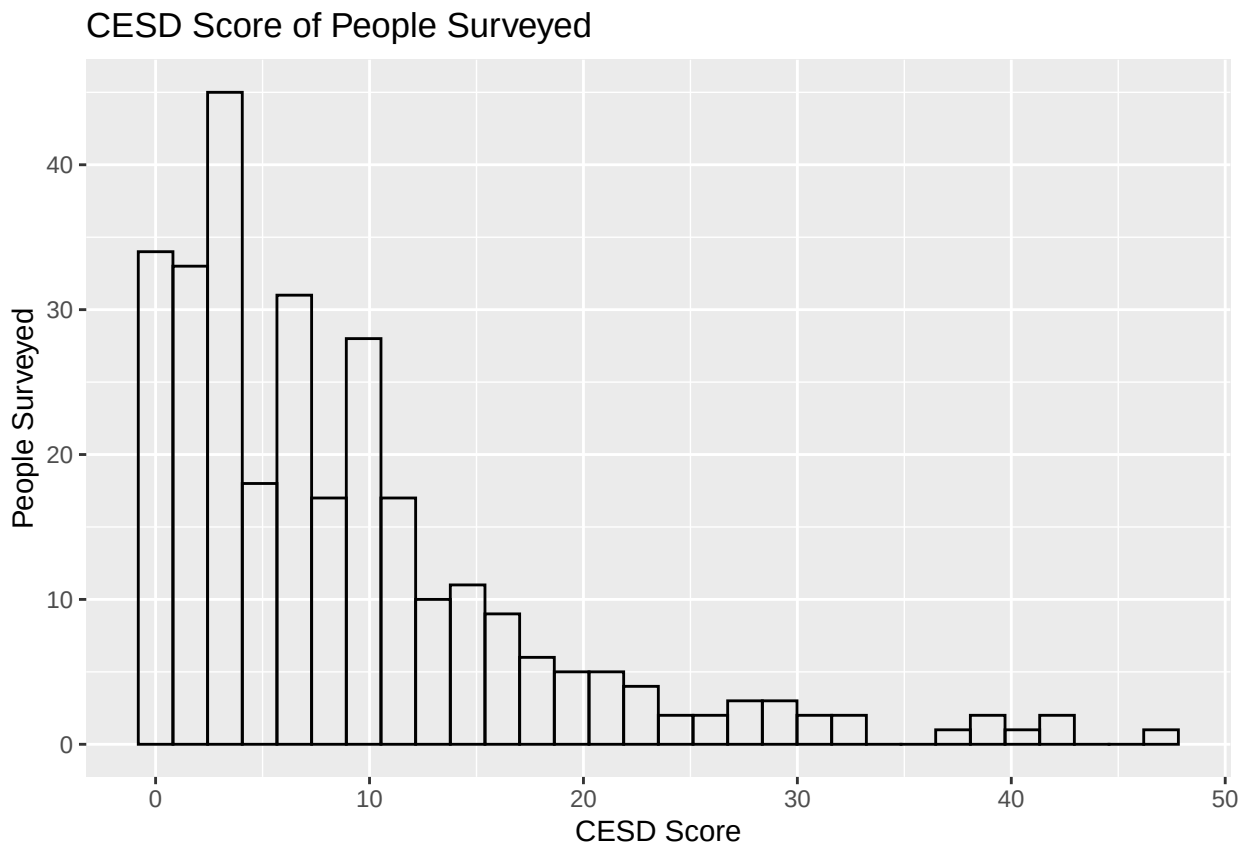
Univariate exploration

Variable:CESD

```
summary(depression$cesd)
```

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

```
ggplot(depression, aes(x=cesd)) + geom_histogram(color="black", fill=NA, bins=30) + ggtitle("CESD Score
```




```
factor(depression$cesd >=16, labels = c("not depressed", "depressed")) %>% table()
```

```
## .
## not depressed     depressed
##           244            50
```

From this data we can see that the majority of people surveyed (83%) are not depressed according to the Center for Epidemiological Studies Depression(CESD) which says someone must score a 16 or higher to be diagnosed.This is right skewed which tells us higher CESD scores are less frequent.

Variable: HEALTH

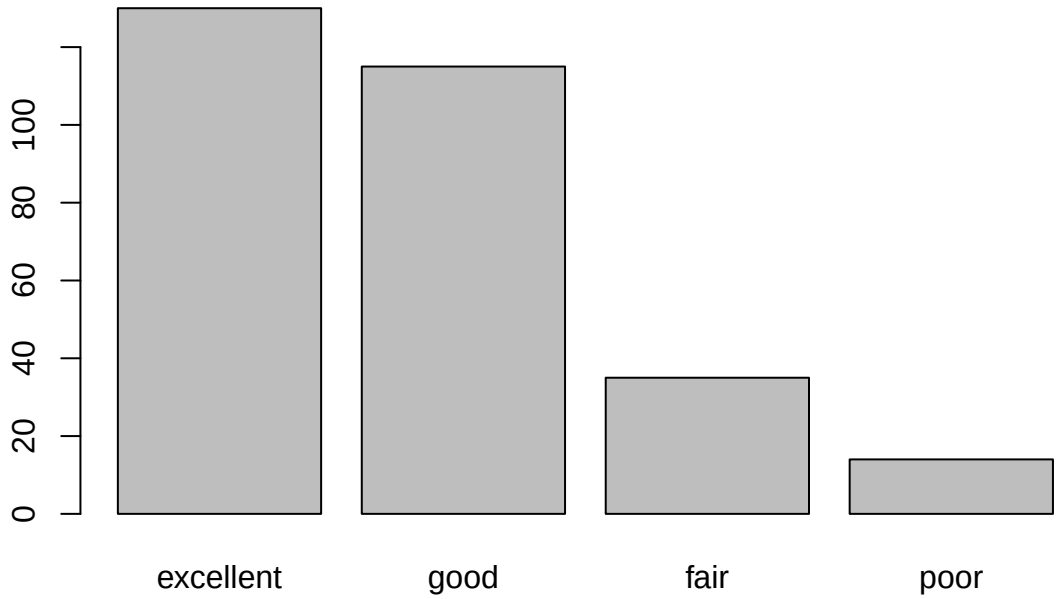
```
summary(depression$health)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   1.772   2.000   4.000
```

```
factor(depression$health, labels = c("excellent","good", "fair", "poor")) %>% table()
```

```
## .
## excellent      good      fair      poor
##       130       115        35        14
```

```
factor(depression$health, labels = c("excellent","good", "fair", "poor")) %>% table() %>% barplot()
```

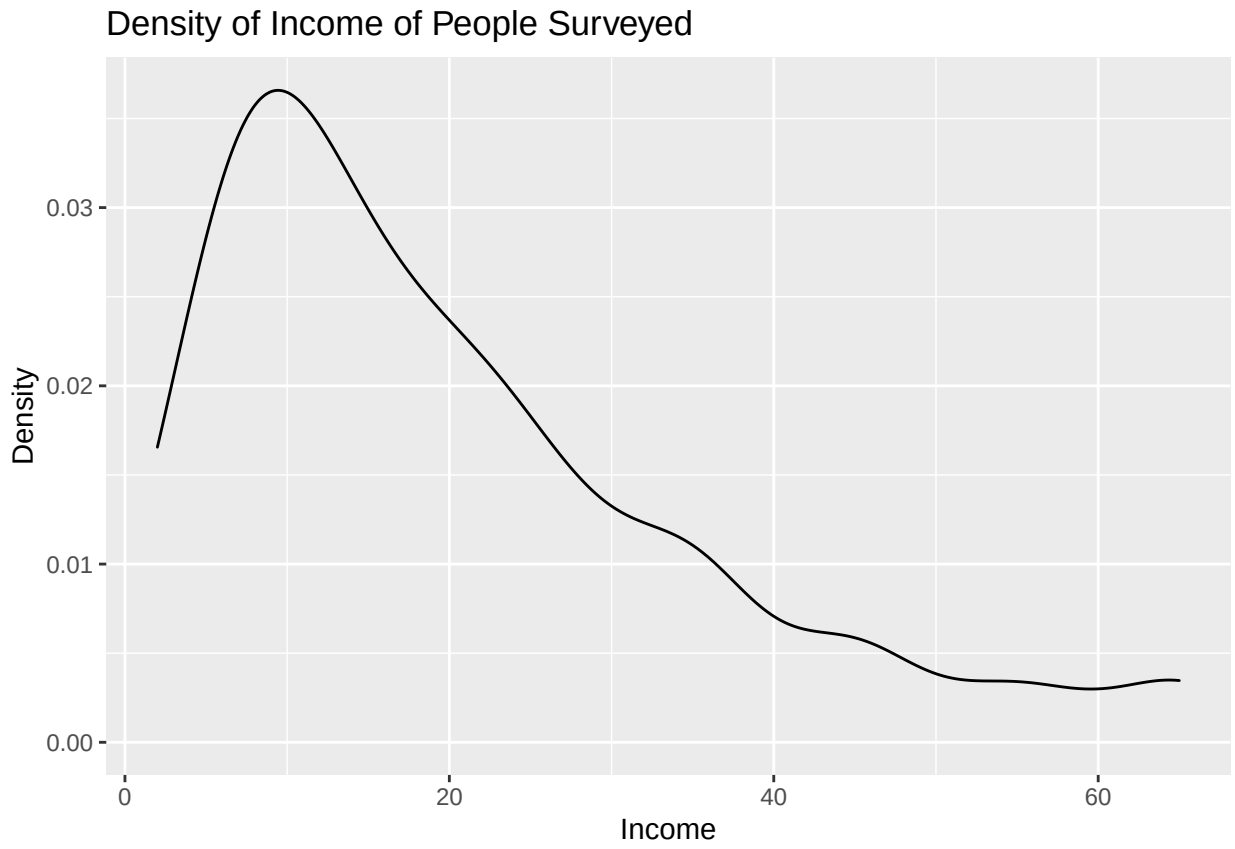



```
depression$health <- factor(depression$health,  labels=c("excellent", "good", "fair", "poor"))
```

We can see here that most of the sample is in excellent or good health, with only 16.7% of the participants having fair or poor health.

Variable: INCOME

```
ggplot(depression, aes(x=income)) + geom_density() +ggtitle("Density of Income of People Surveyed") + xl
```




```
summary(depression$income)
```

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

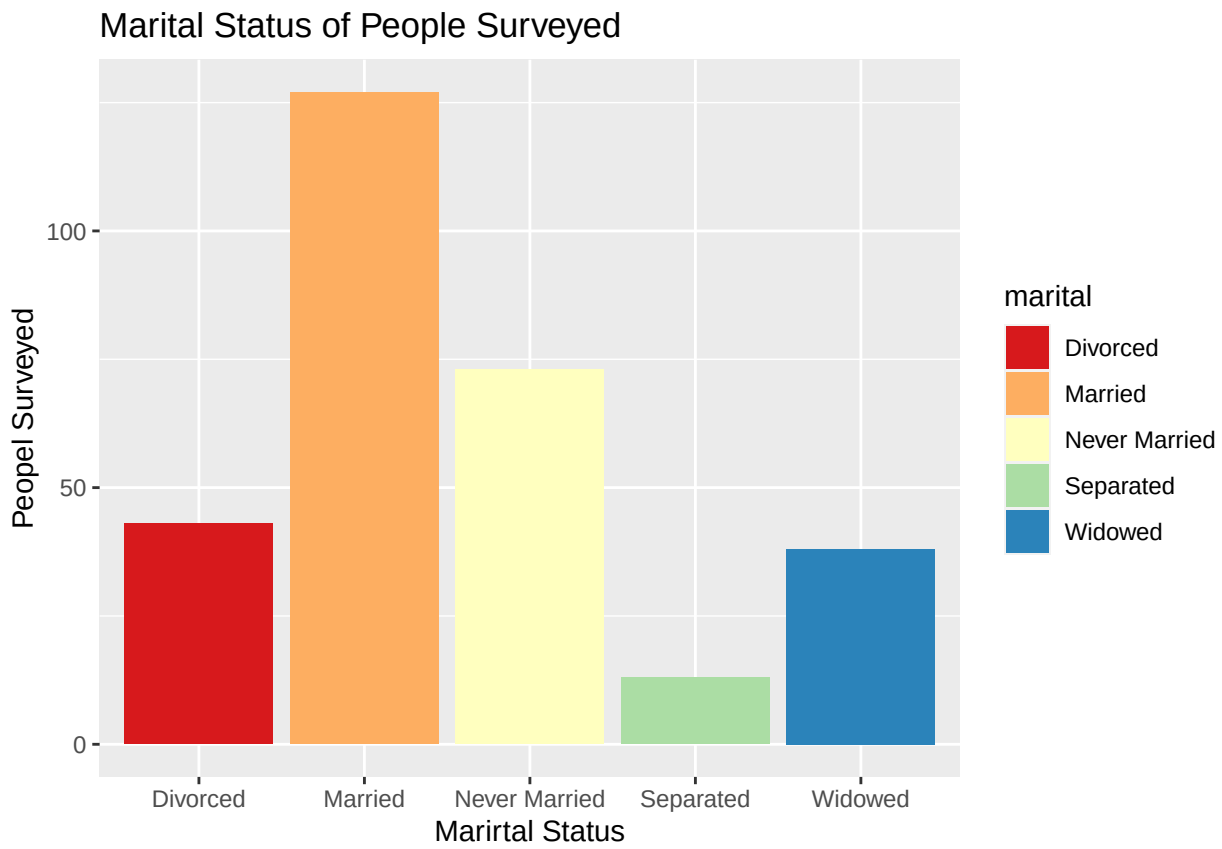
From this we can see that the median is 15,00 a year with a high of 65,00 and a low of 2,000. This is right skewed so out of the people surveyed not many of them make over 30,000 a year.

Variable:MARITAL

```
table(depression$marital)
```

```
##
##      Divorced       Married Never Married     Separated       Widowed
##            43           127            73            13            38
```

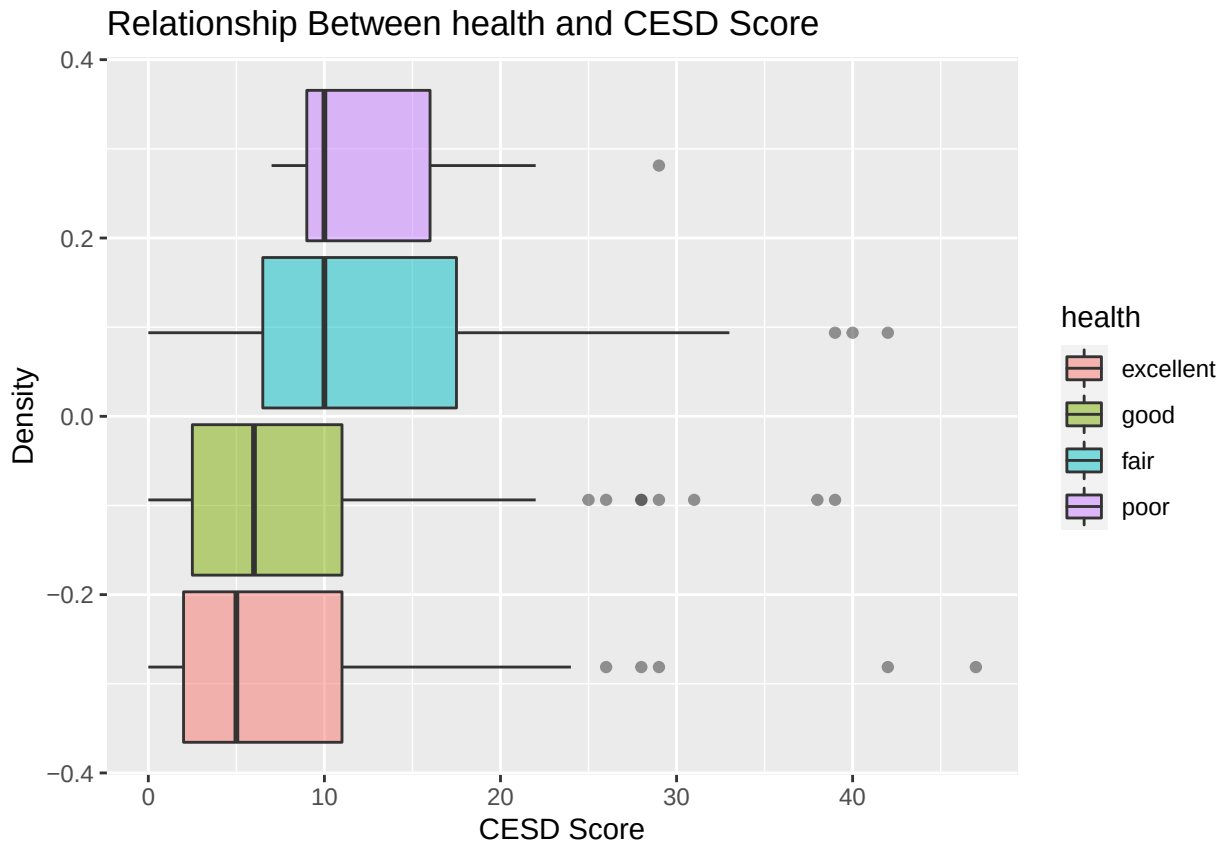
```
ggplot(depression, aes(x=marital, fill=marital)) + geom_bar() + scale_fill_brewer(palette="Spectral") +
```



The biggest group in this is married with 43% and the smallest by far is separated with 4%.

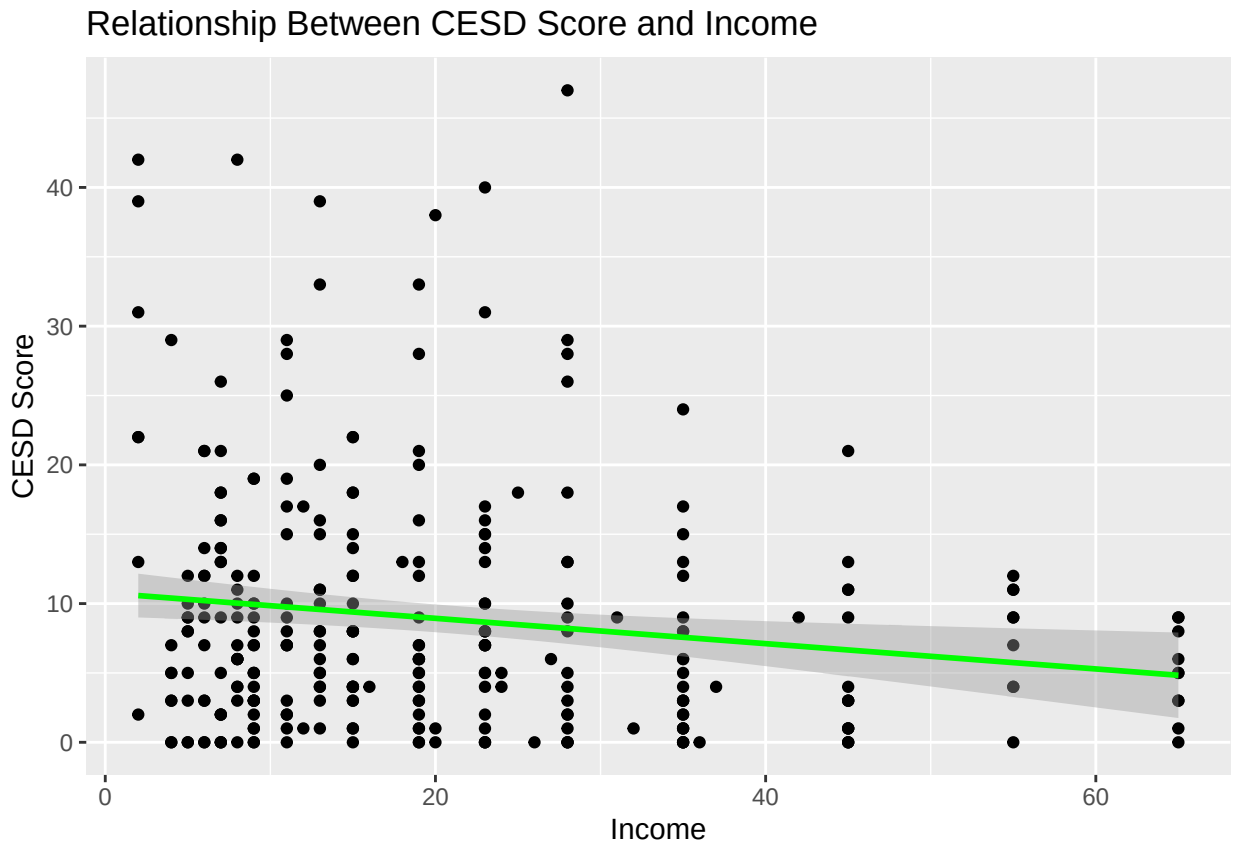
Bivariate Exploration

```
ggplot(depression, aes(cesd, fill=health)) + geom_boxplot(alpha=.5) +ggtitle("Relationship Between heal
```



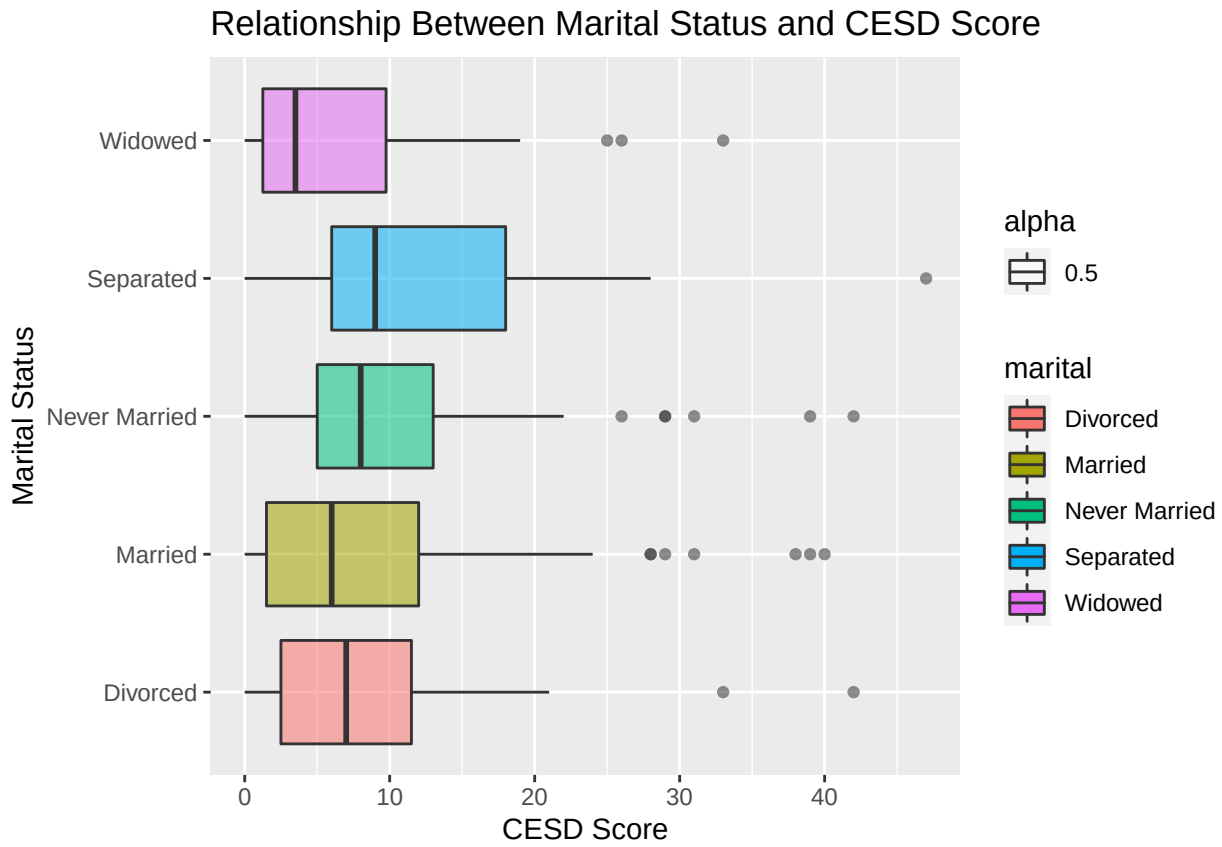
This plot shows that the group poor health and fair health's median are around the same, and higher than excellent and good health. The interesting thing is the poor health group only has one high score around 30. The other three have higher scores and more outliers. In fact the group excellent health has both the highest and the lowest scores. In general people with excellent health and good health have around the same frequency of scores. It is important to note that people with excellent and good health make up 83% of people surveyed.

```
ggplot(depression, aes(y=cesd, x=income)) + geom_point() + geom_smooth(method="lm", color="green", formu
```



In this graph you can see that almost all the CESD scores above 16 make below 40,000 thousand a year. The highest scores all make less than 30,000 thousand a year.

```
ggplot(depression, aes(cesd, marital, fill=marital, alpha=.5)) + geom_boxplot() +ggtitle("Relationship
```



In this graph we can see that separated people have the highest median score for the CESD test, and this group has the highest score. Widowed people had the lowest median score, but still have some outliers. In general the means of each group are pretty similar and all under 10.

Conclusion

Depression is a lot more complex than just the variables I explored, but out of the variables I explored there is one variable that seems to have the biggest impact on depression. My original hypothesis was that poor health would have the biggest impact on depression and that was kind of inconclusive. In the poor health group there were no very low scores like in the rest of the groups, but they also didn't have any of the highest scores. Poor health had a higher median than excellent or good health and around the same as fair health. People with excellent health varied a lot from 0 to around 50. So that group had more diverse and higher scores,but a lower median. As far as marital status, each one seems to be similar. All of the medians are under 10, with separated having the highest median. Each group has some high outliers, though Married and Never married have the most outliers and separated has the highest outlier. It's also important to note the biggest group of people survived were married (43%). The variable that seemingly has the biggest impact on depression is income. In fact there is no cases of depression for people who make over 45,000 a year. On the other hand all of the worst cases of depression are clustered at under 30,000 a year. This could be because not as many people surveyed made an income the higher end.

I found these results to be very surprising. I do think looking into different variables and surveying more people would make it more accurate. At the same time I believe no matter how many people are surveyed money will play a part in depression level.