

MATH130 Exploratory Data Analysis

Nile Reed

2-21-22

Data Set: Depression. This data set is from a study on Depression in Adults in Los Angeles County. The study can be found in “Practical Multivariate Analysis, 5th edition by Afifi, May and Clark”. It includes 294 observations of 37 variables. The variables I will be comparing are CESD scores (Measure of depression, higher CESD scores equate to higher levels of depression), Chronic illness, and Education level. I want to see if chronic illness is associated with higher levels of depression, and if education level has any correlation to depression.

```
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(ggplot2)
```

```
library(forcats)
```

```
depression <- read.table("/Users/nilereed/Desktop/MATH130/DATA/depress_081217.txt", header = TRUE, sep=  
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 1 0 0 ...
## $ c19     : int  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 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 ...
```

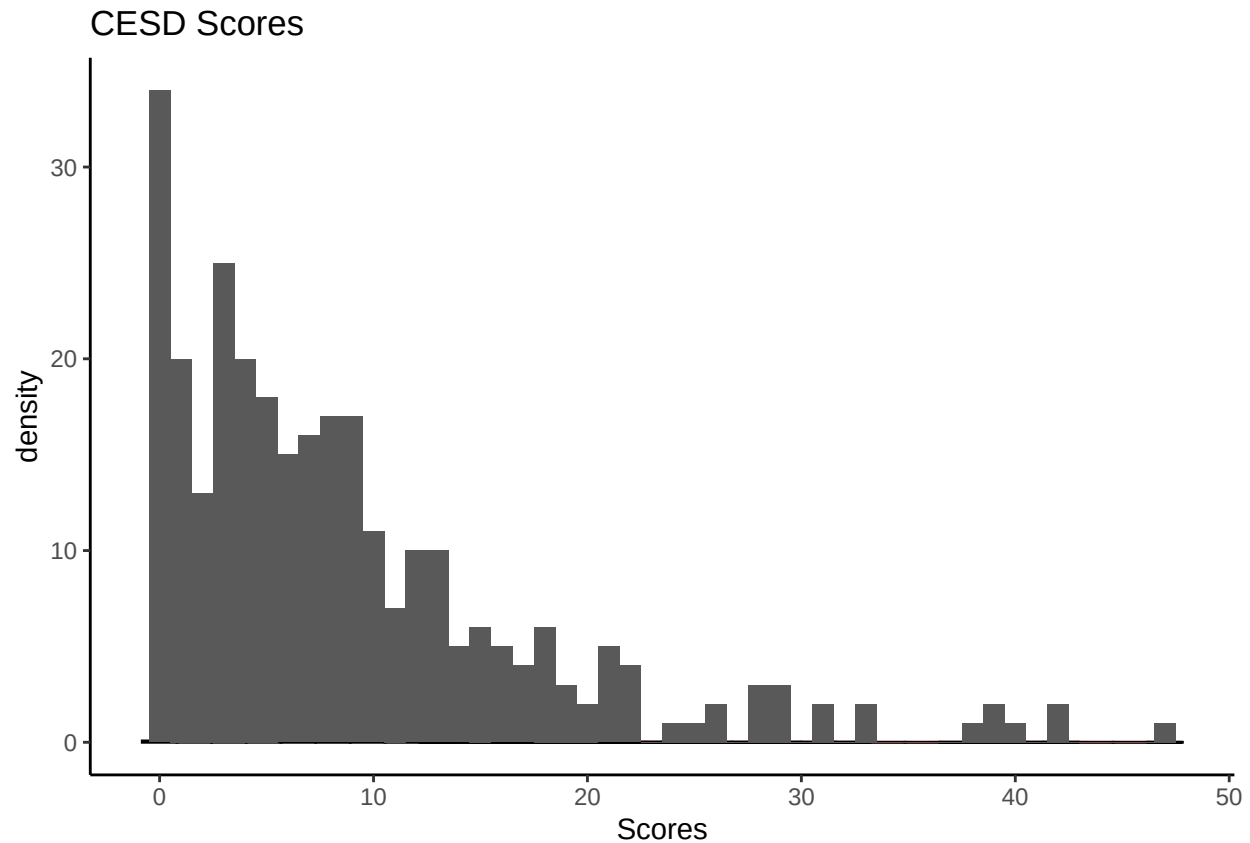
Depression

```
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_density(col="red") + geom_histogram(aes(y=..density..), colour=
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



CESD scores range from 0 to 47. The average score is 8.9, and the median is 7.0. The graph is skewed to the left showing that most of the participants have relatively lower scores.

Chronic Illness

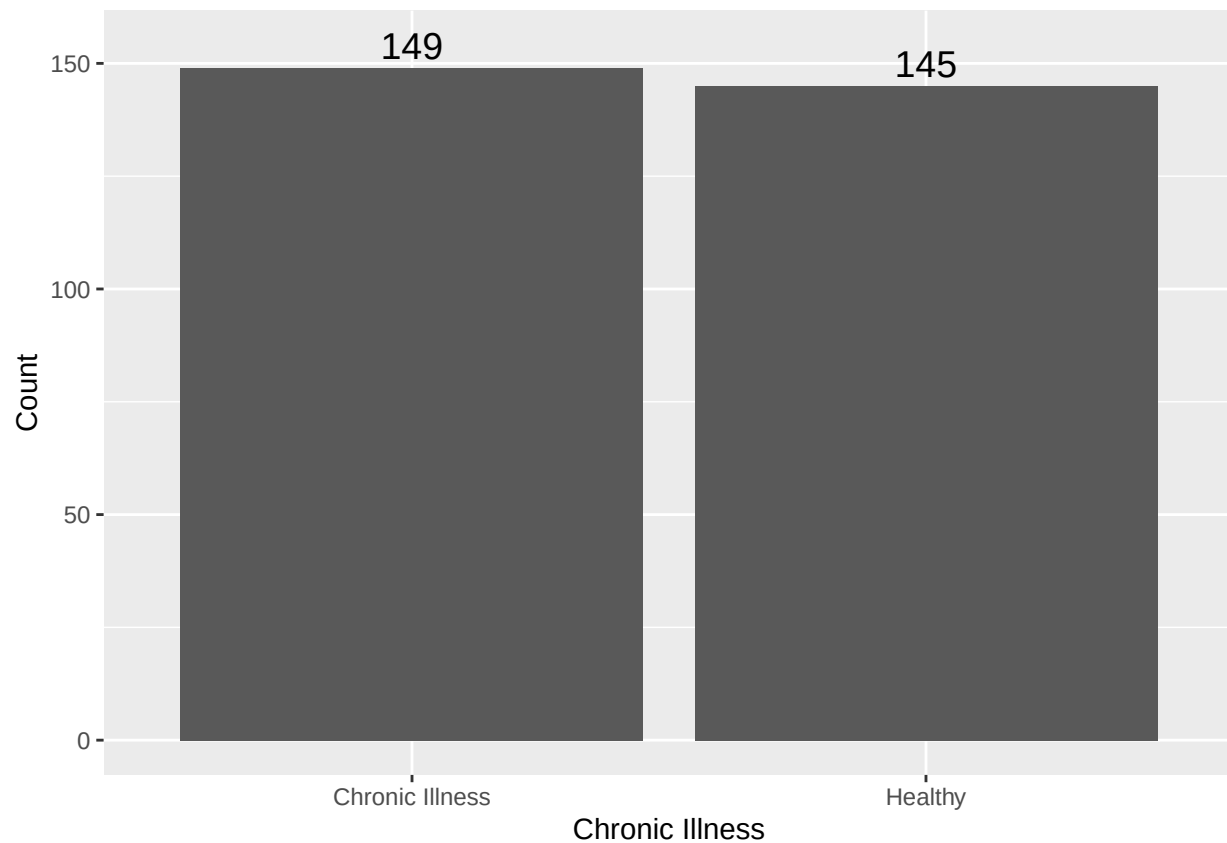
```
depression$chronicillness <- ifelse(depression$chronill==1, "Chronic Illness", "Healthy")
table(depression$chronicillness)
```

```
##
## Chronic Illness      Healthy
##           149           145
```

```
mean(depression$chronill)
```

```
## [1] 0.5068027
```

```
ggplot(depression, aes(x=chronicillness)) + geom_bar() + ylab("Count") + xlab ("Chronic Illness") + geom
```

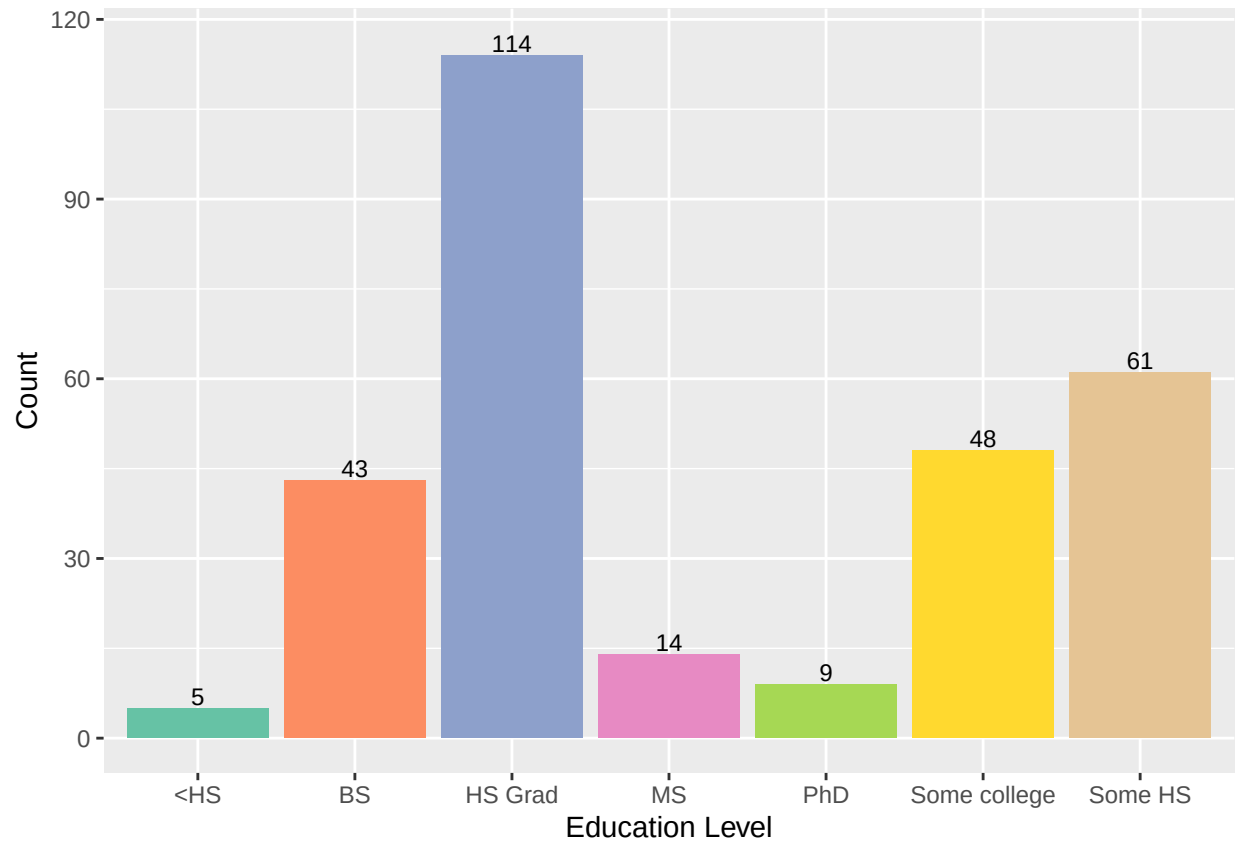


Just over half of the subjects interviewed had experienced chronic illness in the last year.

```
table(depression$educat)
```

```
##
##      <HS      BS      HS Grad      MS      PhD Some college
##      5       43      114      14       9       48
##      Some HS
##      61
```

```
ggplot(depression, aes(x=educat, fill=educat)) + geom_bar() + scale_fill_brewer(palette="Set2", guide='')
```



The most common education level was High School graduate.

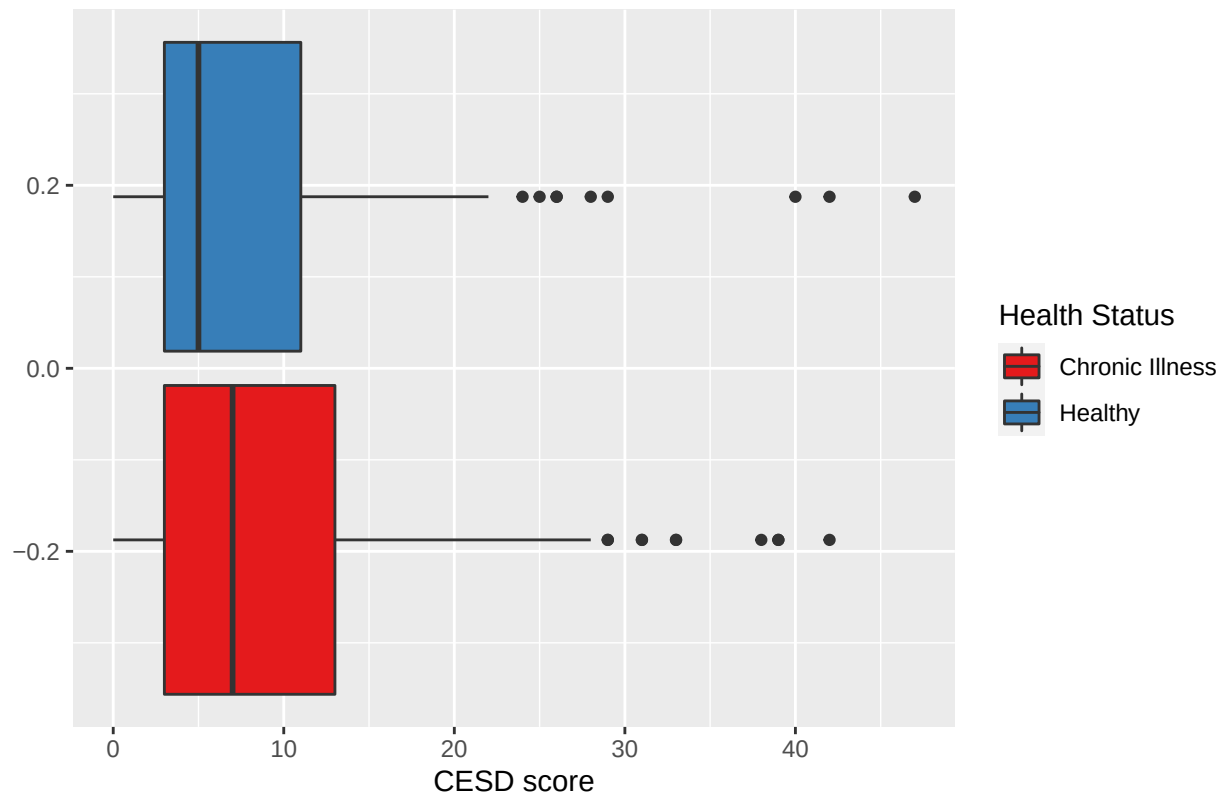
Chronic Illness and Depression

```
table(depression$chronicillness, depression$cesd)
```

```
##
##           0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19
## Chronic Illness 15  8  8 10  7  8 10  9 10 10  5  4  5  6  2  1  4  3  3  3
## Healthy       19 12  5 15 13 10  5  7  7  7  6  3  5  4  3  5  1  1  3  0
##
##           20 21 22 24 25 26 28 29 31 33 38 39 40 42 47
## Chronic Illness  1  3  2  0  0  0  2  2  2  2  1  2  0  1  0
## Healthy         1  2  2  1  1  2  1  1  0  0  0  0  1  1  1
```

```
ggplot(depression, aes(x=cesd, fill=chronicillness)) + geom_boxplot() + scale_fill_brewer(palette = "Se
```

Chronic Illness and Depression Levels Box Plot



```
ggplot(depression, aes(x=cесd, fill=chronicillness)) + geom_density(alpha=.4) + scale_fill_brewer(palette="Set1")
```

Chronic Illness and Depression Levels Density Plot



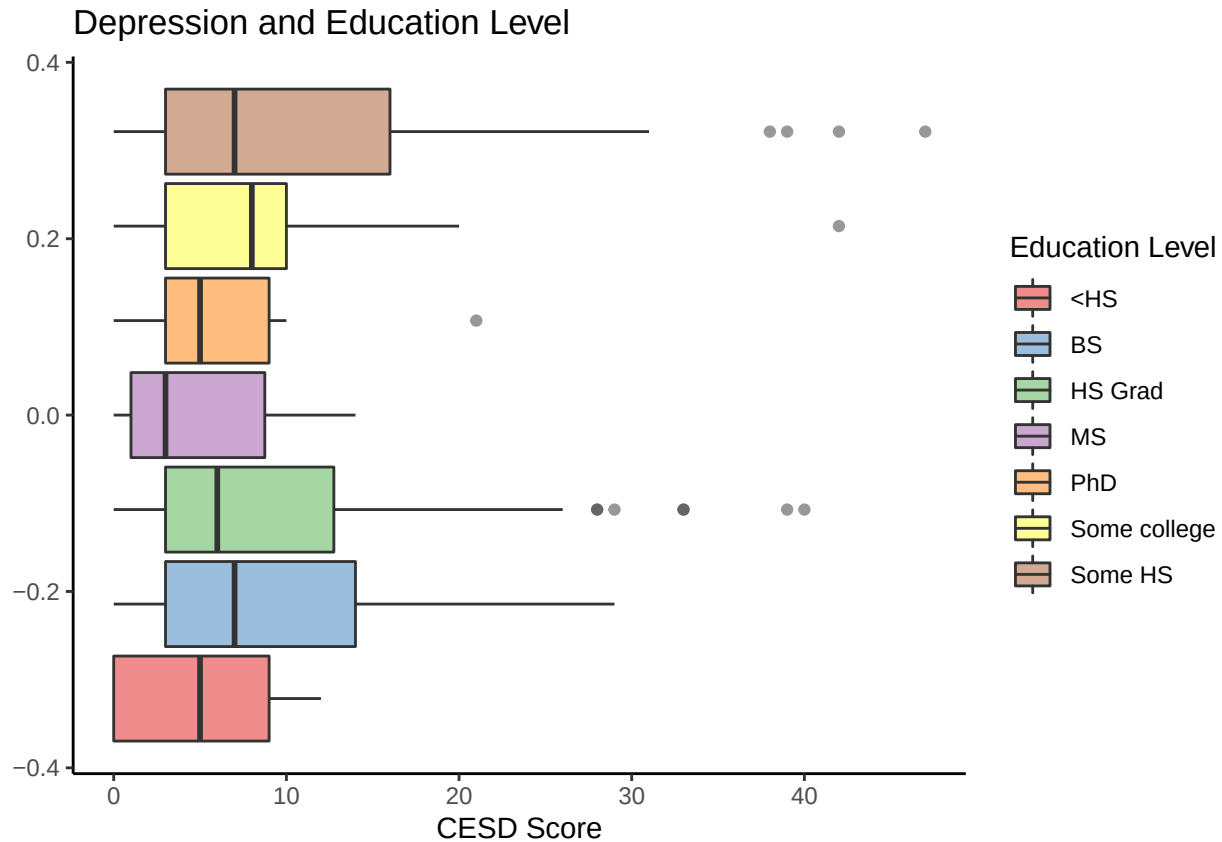
These graphs show that individuals with chronic illness tend to have higher CESD scores, indicating higher levels of depression.

Depression and Education Level

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

```
##
##           0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##   <HS      2  0  0  0  0  1  0  0  0  1  0  0  1  0  0  0  0  0  0  0  0
##   BS       2  3  2  5  4  2  2  4  1  2  0  0  2  2  2  2  0  1  1  0  0
##   HS Grad  10 10  5  7 11  8  7  8  5  1  6  3  4  5  1  2  3  2  1  3  0
##   MS       3  2  1  3  0  0  0  0  1  1  1  0  1  0  1  0  0  0  0  0  0
##   PhD      2  0  0  1  1  1  0  0  0  2  1  0  0  0  0  0  0  0  0  0  0
##   Some college 4  4  2  3  3  2  3  1  6  6  3  4  1  1  0  1  1  0  1  0  1
##   Some HS   11  1  3  6  1  4  3  3  4  4  0  0  1  2  1  1  1  1  3  0  1
##
##           21 22 24 25 26 28 29 31 33 38 39 40 42 47
##   <HS      0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   BS       1  3  0  0  0  1  1  0  0  0  0  0  0  0
##   HS Grad  2  1  1  0  1  2  1  0  2  0  1  1  0  0
##   MS       0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   PhD      1  0  0  0  0  0  0  0  0  0  0  0  0  0
##   Some college 0  0  0  0  0  0  0  0  0  0  0  0  1  0
##   Some HS   1  0  0  1  1  0  1  2  0  1  1  0  1  1
```

```
ggplot(depression, aes(x=cesd, fill=educat)) + geom_boxplot(alpha=.5) + theme_classic() + scale_fill_brewer()
```



Education levels of “Some High School” and “Bachelor of Science” appear to have the highest CESD scores. With “Master of Science” having the lowest CESD scores.

Conclusion: CESD scores are used to measure depression level. They range from 0 to 47, with higher scores indicating higher levels of depression. Individuals who had experienced some form of chronic illness in the last year tended to have higher CESD scores, indicating that chronic illness is linked to higher levels of depression. Looking at education, it didn’t show a very clear pattern as to whether higher education levels or lower education levels were linked to higher CESD scores. “Some High School” had the highest upper quartile value, and “Some College” had the highest median, but all education levels were in pretty similar ranges.