# EDA_tharry

Trinity Lloyd-Harry

Due 3/1

## Reading in Data Set

```
depress <- read.table("C:/Users/trini/OneDrive/Documents/MATH 130/Data/Depress.txt", header=TRUE, sep="
head(depress)
```

```
##   ID SEX AGE MARITAL EDUCAT EMPLOY INCOME RELIG C1 C2 C3 C4 C5 C6 C7 C8 C9 C10
## 1  1   2  68       5      2      4      4     1  0  0  0  0  0  0  0  0  0   0
## 2  2   1  58       3      4      1     15     1  0  0  1  0  0  0  0  0  0   0
## 3  3   2  45       2      3      1     28     1  0  0  0  0  1  0  0  0  0   0
## 4  4   2  50       3      3      3      9     1  0  0  0  0  1  1  0  3  0   0
## 5  5   2  33       4      3      1     35     1  0  0  0  0  0  0  0  3  3   0
## 6  6   1  24       2      3      1     11     1  0  0  0  0  0  0  0  0  1   0
##   C11 C12 C13 C14 C15 C16 C17 C18 C19 C20 CESD CASES DRINK HEALTH REGDOC TREAT
## 1   0   0   0   0   0   0   0   0   0   0    0     0     2      2      1     1
## 2   0   1   0   0   1   0   1   0   0   0    4     0     1      1      1     1
## 3   0   0   0   1   1   1   0   0   0   0    4     0     1      2      1     1
## 4   0   0   0   0   0   0   0   0   0   0    5     0     2      1      1     2
## 5   0   0   0   0   0   0   0   0   0   0    6     0     1      1      1     1
## 6   0   1   2   0   0   2   1   0   0   0    7     0     1      1      1     1
##   BEDDAYS ACUTEILL CHRONILL
## 1       0        0        1
## 2       0        0        1
## 3       0        0        0
## 4       0        0        1
## 5       1        1        0
## 6       0        1        1
```

```
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)
```

## Introduction

I will be completing my exploratory data analysis on the Depression data set. This data set is a prospective study on depression completed in Los Angeles County. It includes 294 observations and 37 variables. The variables I will be considering during this analysis are CESD, ACUTEILL, and CHRONILL. I am interested in seeing if there was any correlation between CESD and the presence of an acute illness in the past two months or chronic illness in the past year.

## Univariate Exploration

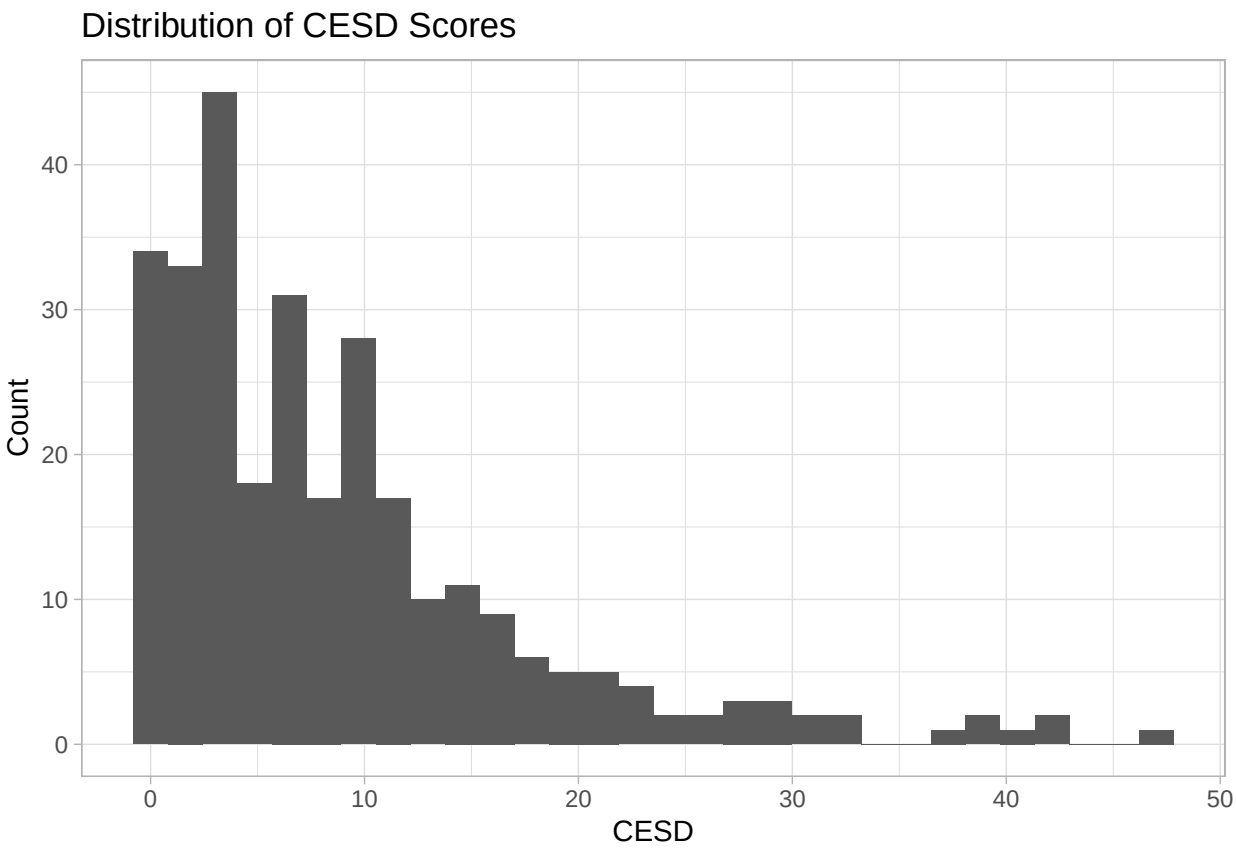
### CESD

```
summary(depress$CESD)
```

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

```
ggplot(depress, aes(x=CESD)) + geom_histogram() + theme_light() + ggtitle("Distribution of CESD Scores")
```

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



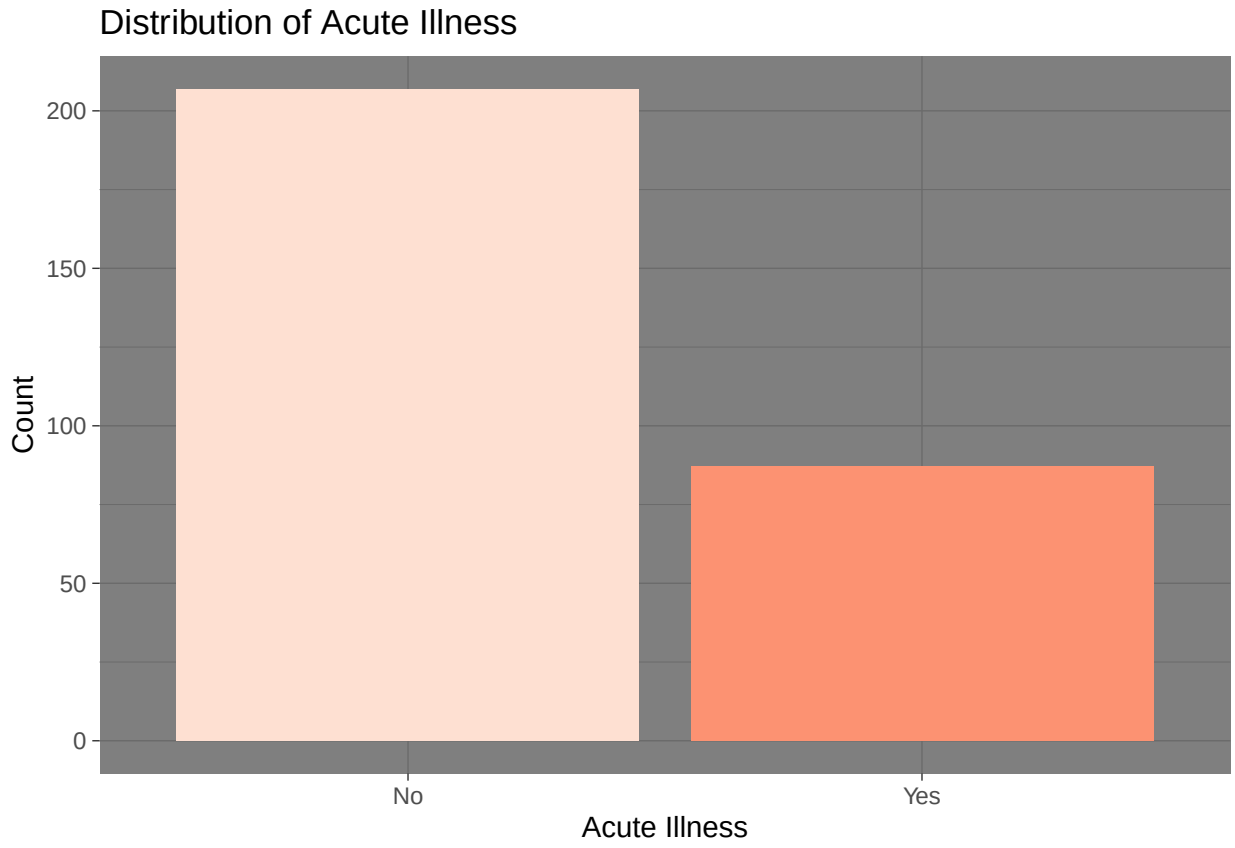
The CESD variable represents the total score out of 60 on a depression questionnaire. There is 20 questions on the questionnaire and each question has an individual score range of 0-3, resulting in the maximum score of 60. A CESD of >16 indicates depression. By looking at my graph and summary statistics I can see that the maximum score only reached 47 and the mean score was only 8.884.

**Acute Illness**

```
depress$Acute<- ifelse(depress$ACUTEILL==1, "Yes", "No")
table(depress$Acute) %>% prop.table()*100
```

```
##
##       No      Yes
## 70.40816 29.59184
```

```
ggplot(depress, aes(x=Acute, fill=Acute)) + geom_bar() + theme_dark() + ylab("Count") + xlab("Acute Illn
```



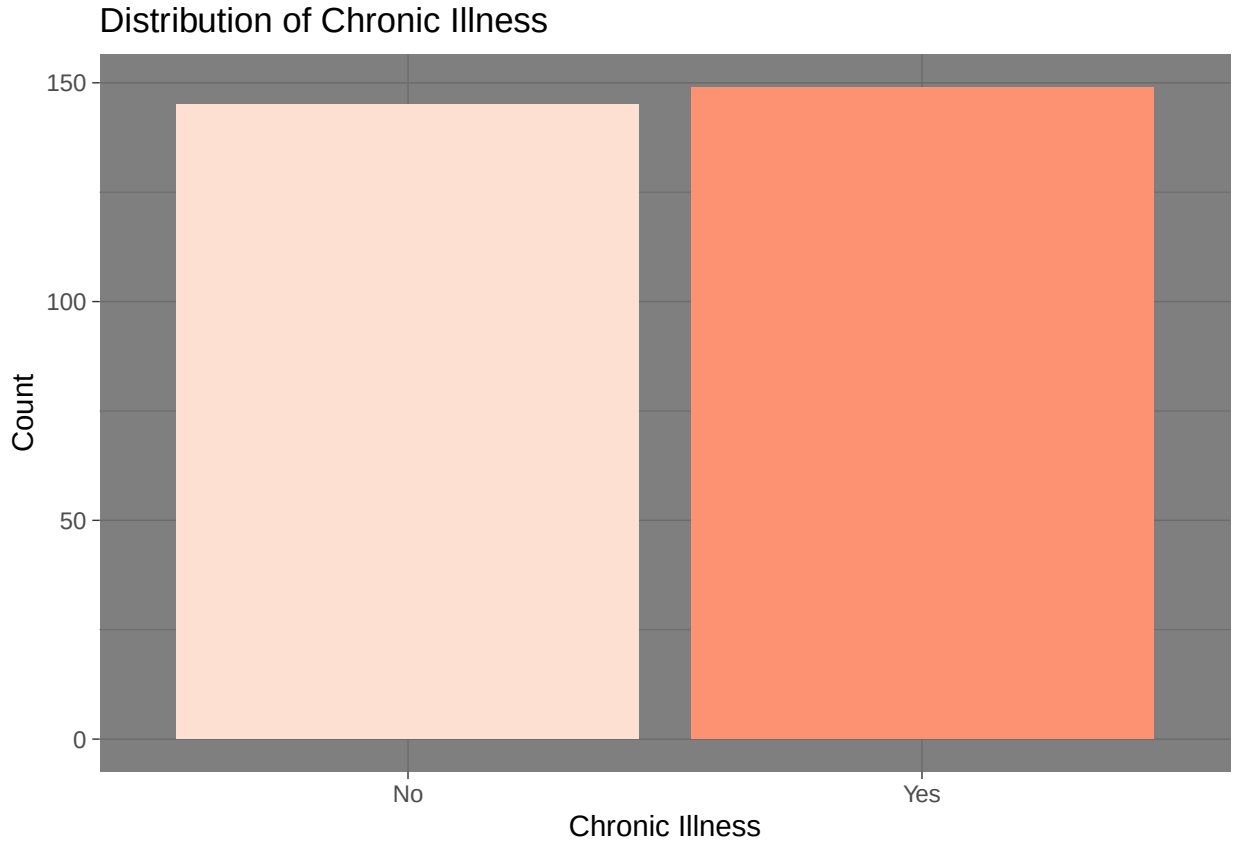

The acute illness variable shows whether or not a person has experienced any acute illness within two months of when the survey was administered. Approximately 29.6% of survey respondents said they had experienced an acute illness within two months and 70.4% of respondents said they have not experienced an acute illness within the past two months.

### Chronic Illness

```
depress$Chronic <- ifelse(depress$CHRONILL==1,"Yes","No")
table(depress$Chronic) %>% prop.table()*100
```

```
##
##       No      Yes
## 49.31973 50.68027
```

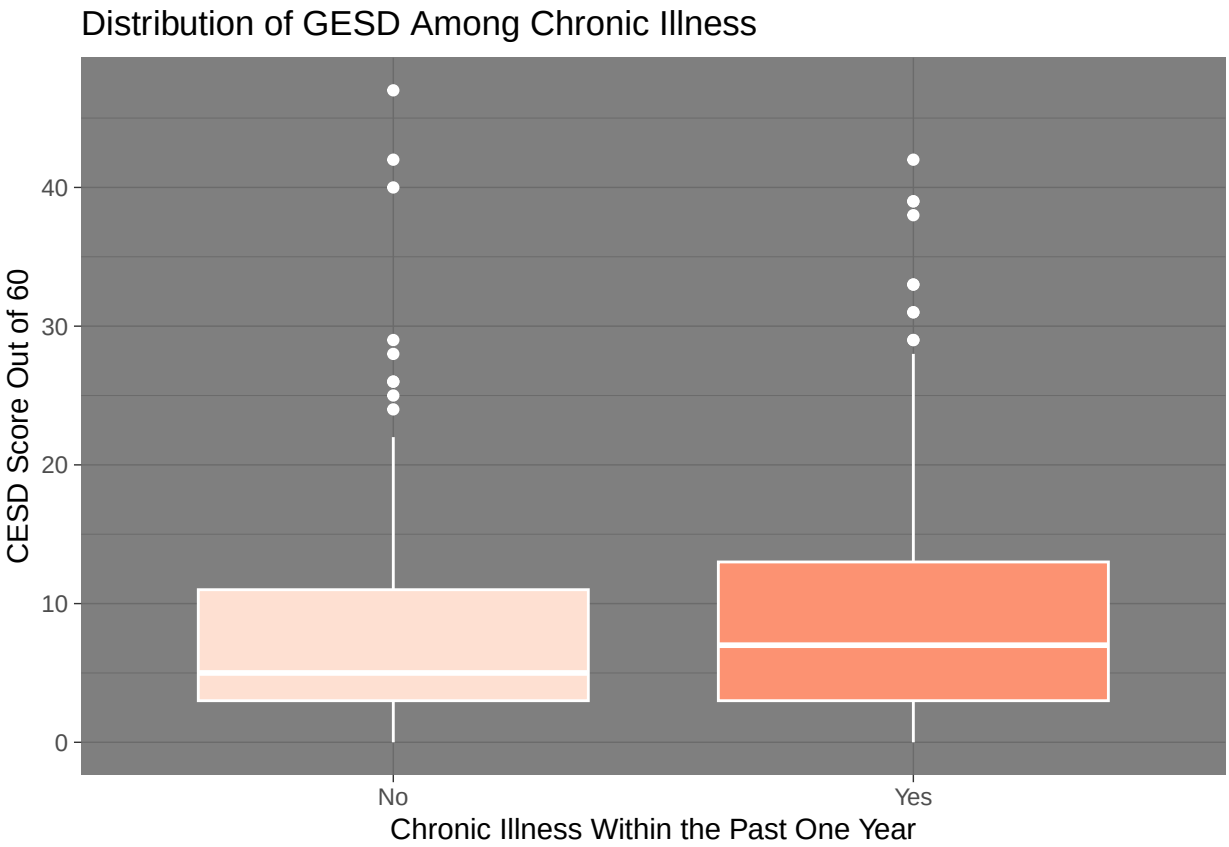
```
ggplot(depress,aes(x=Chronic, fill=Chronic)) + geom_bar() + theme_dark() + ylab("Count") + xlab("Chroni
```




The chronic illness variable shows whether or not a person has experienced any chronic illness within a year of when the survey was administered. Approximately 50.7% of survey respondents said they had experienced a chronic illness within the year and 49.3% of respondents said they have not experienced a chronic illness within the year.

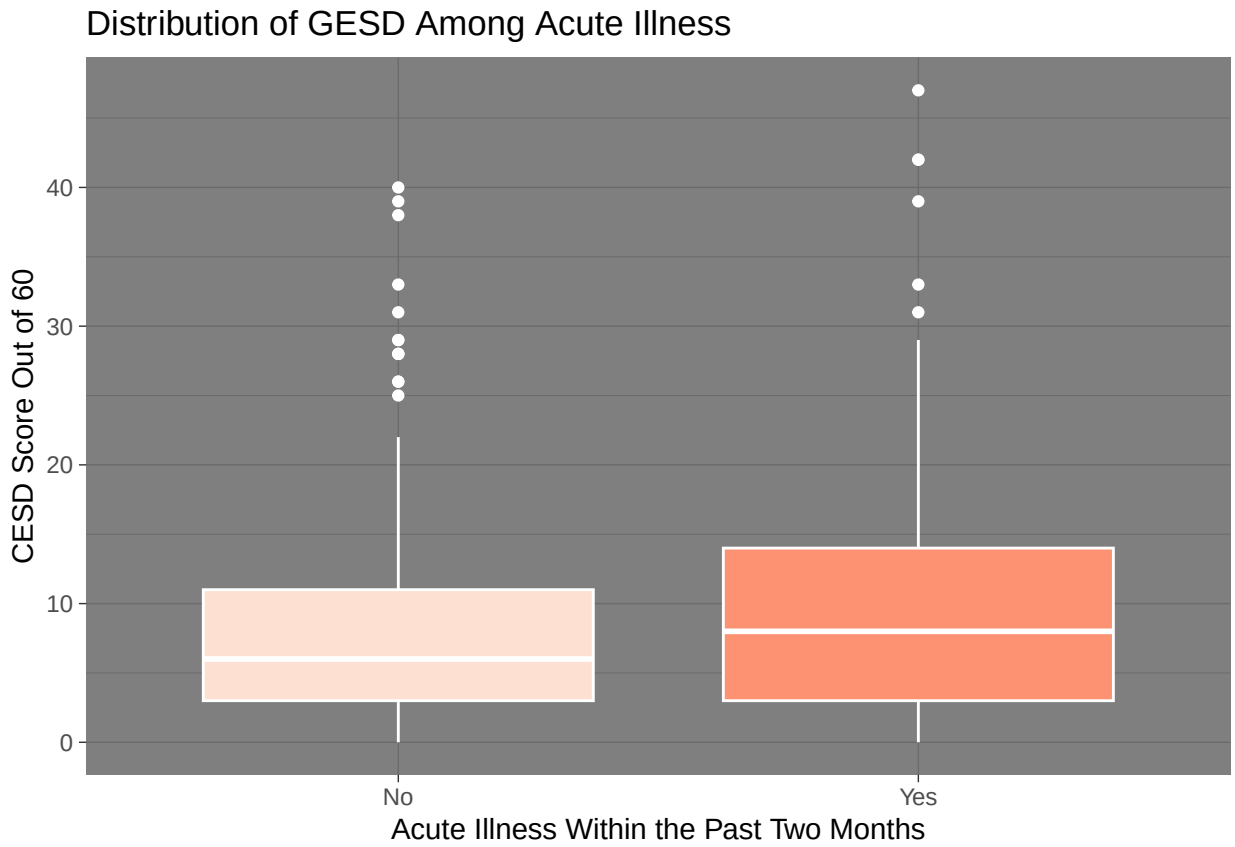
## Bivariate Exploration

```
ggplot(depress, aes(y=CESD, x=Chronic, col=Chronic,fill=Chronic)) + geom_boxplot() + ggtitle("Distributi
```



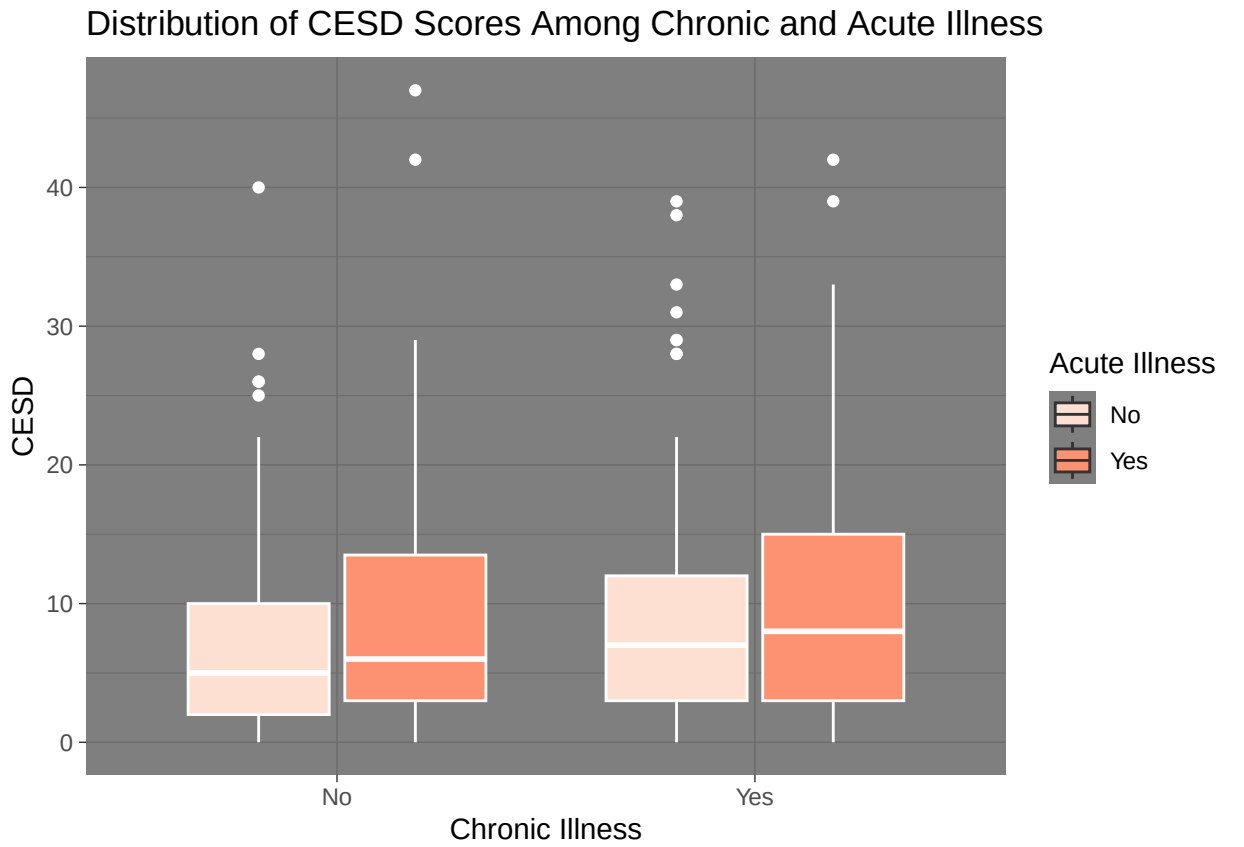

In the above boxplot we can see that the group with a chronic illness has a slightly higher mean CESD score than those without a chronic illness. Interestingly the group without a chronic illness has the outlier with the highest CESD score.

```
ggplot(depress, aes(y=CESD, x=Acute, col=Acute, fill=Acute)) + geom_boxplot() + ggtitle("Distribution o
```




In the above boxplot we can see that those who experienced an acute illness had a higher mean CESD score compared to those who did not experience an acute illness.

```
ggplot(depress, aes(y=CESD, x=Chronic, col=Chronic, fill=Acute))  + geom_boxplot() + xlab("Chronic Illn
```




The above boxplot we can see CESD scores compared with both chronic and acute illnesses. We can see here that the group that experienced both a chronic illness in the past year and an acute illness in the past two months had the highest mean CESD score. We can also see that all three groups that experienced either only chronic illness, only acute illness, and both chronic and acute illness had a higher average CESD score than the group who experienced no illnesses.

```
tapply(depress$CESD, depress$Chronic, mean) %>% round(2)
```

```
##   No  Yes
## 8.06 9.68
```

The table above shows me that the average CESD score for those who experienced a chronic illness in the past year was 9.68, while the average CESD score for those without a chronic illness was 8.06.

```
tapply(depress$CESD, depress$Acute, mean) %>% round(2)
```

```
##    No   Yes
##  8.28 10.32
```

The table above shows me that the average CESD score for those who experienced an acute illness in the past two months was 10.32, while the average CESD score for those without an acute illness was 8.28.

## Conclusion

In conclusion, the groups who experienced chronic and/or acute illness had slightly higher average CESD scores than those who did not experience any illnesses, which aligned with my prior prediction.