Math130 EDA Project

2-24-22

Introduction

For my Exploratory Data Analysis Project, I will be analyzing data from the depression data set and specifically looking at the variables "age", "chronill", and "cesd". I chose these variables because I was interested to see if aging and chronic illnesses has any affect on depression levels in people. The variable "cesd" measures depression levels from 0 to 60 in each observation. The variable "chronill" determines whether the observations have a chronic illness (yes or no). While "age" is the other variable being analysed.

```
depress <- read.table("/Users/allisoncarmichael/Downloads/math130/data/depress_081217.txt", header=TRUE
str(depress)
```

```
## '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 ...
```

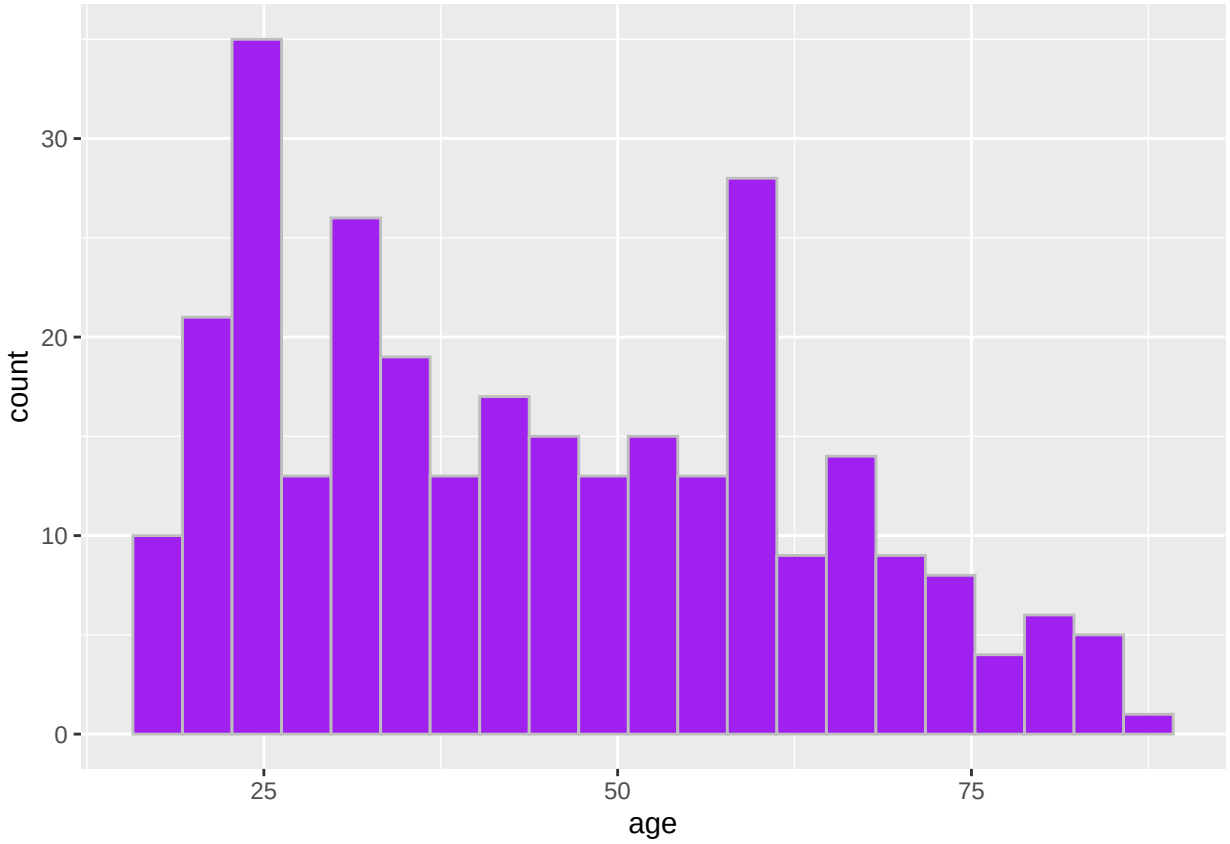
Univariate Exploration

AGE:

```
library(ggplot2)
```

```
## Warning in register(): Can't find generic `scale_type` in package ggplot2 to
## register S3 method.
```

```
ggplot(depress) +
geom_histogram(aes(x=age),
binwidth = 3.5, fill = "purple" , color = "grey")
```




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

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

This summary of the "age" variable shows us the youngest age out of the 294 observations recorded was 18 while the oldest age recorded was 89. The average age across observations was roughly 44.41 years old.

```
var(depress$age)
```

```
## [1] 327.0832
```

The variance of the "age" variable is 31.

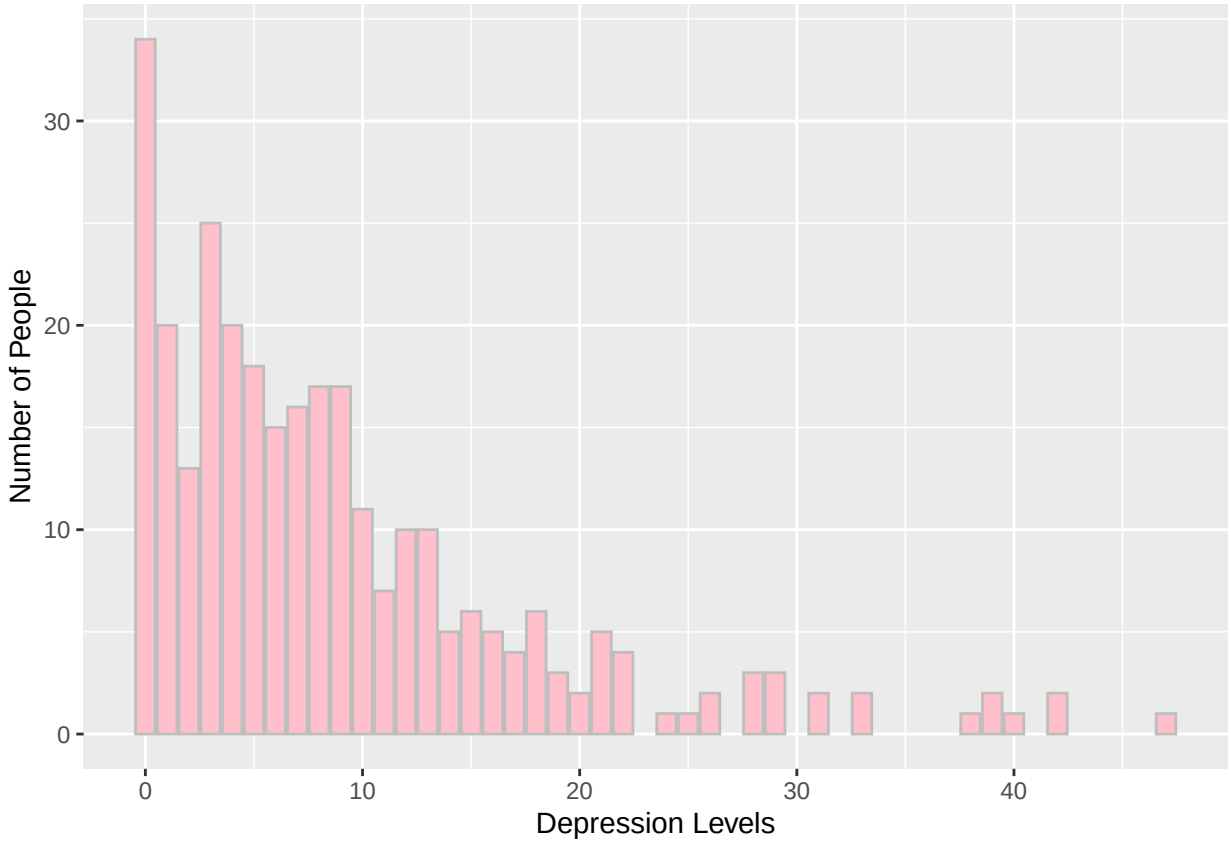
```
sd(depress$age)
```

```
## [1] 18.08544
```

Standard deviation for the "age" variable is 18.08

CESD: Level of Depression

```
ggplot(depress, aes(x=cesd)) + geom_bar(fill='pink', color='grey')+ylab("Number of People")+xlab("Depres
```



The bar graph shows that out of the 294 observations in that data set, less people had high cesd levels. The majority of the people had cesd scores of 10 or lower, according to left skew pattern of the bar graph.

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

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

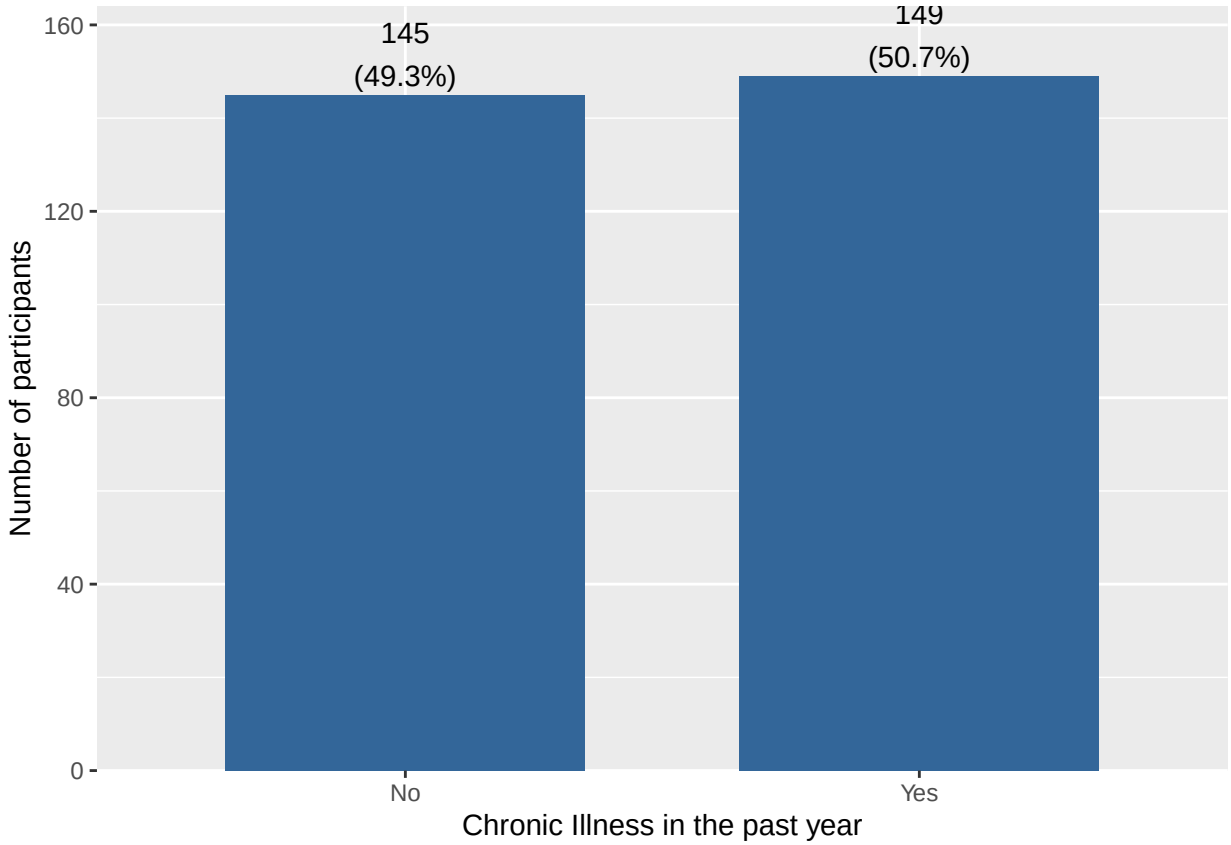
The summary of the "cesd" variable shows us that the average depression level across the data set is 8.88. The lowest depression levels were 0 while the highest levels were 47.

CHRONILL:

```
depress$chronill <- factor(depress$chronill, labels=c("No","Yes"))
table(depress$chronill)
```

```
##
##  No Yes
## 145 149
```

```
library(sjPlot)
plot_frq(depress$chronill)+xlab("Chronic Illness in the past year")+ylab("Number of participants")
```

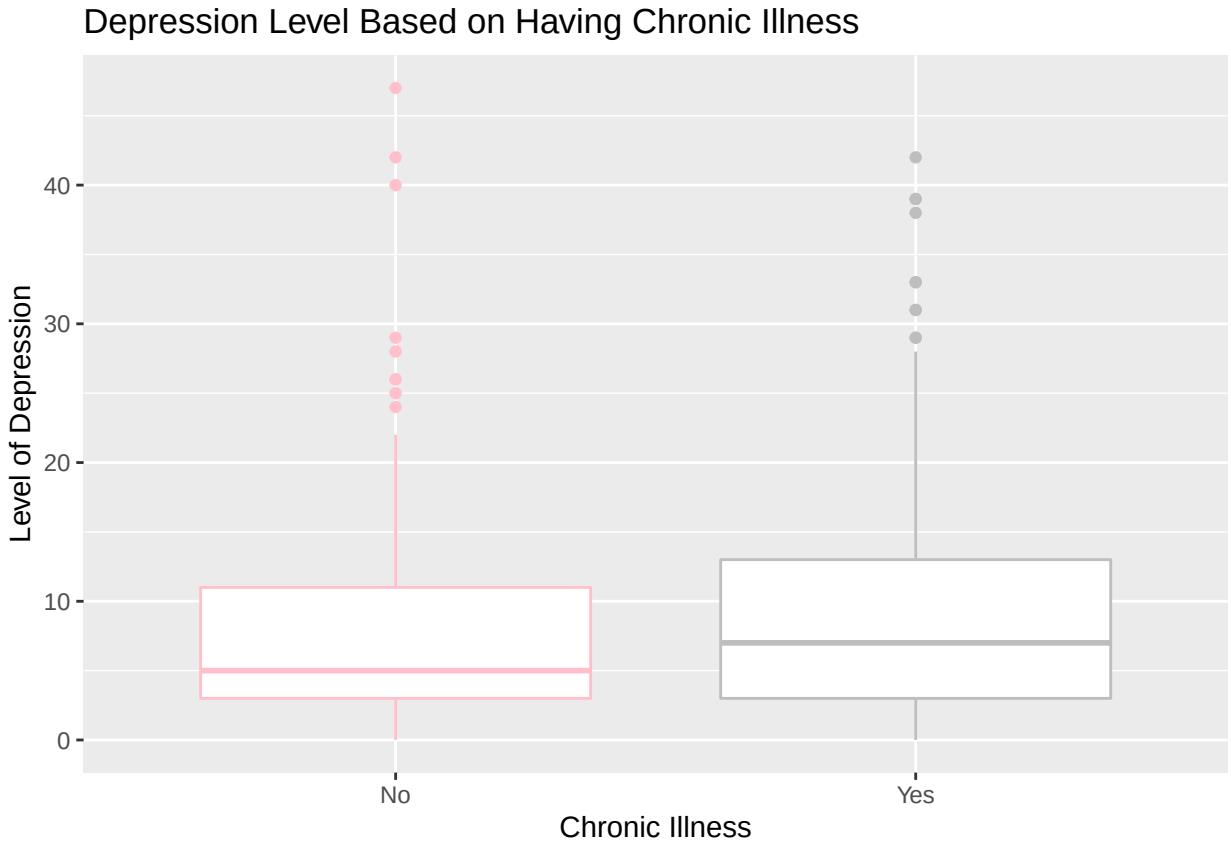


The table and the bar chart clearly show that 145 observations said "no" while 50.7% of the observations (149) said "yes" to having a chronic illness.

Bivariate Exploration

CESD vs. CHRONILL

```
ggplot(depress, aes(x=chronill, y=cesd, col=chronill)) + geom_boxplot()+
xlab("Chronic Illness")+ ylab("Level of Depression")+
ggtitle("Depression Level Based on Having Chronic Illness")+
scale_color_manual(values = c("pink", "grey"), guide = "none")
```



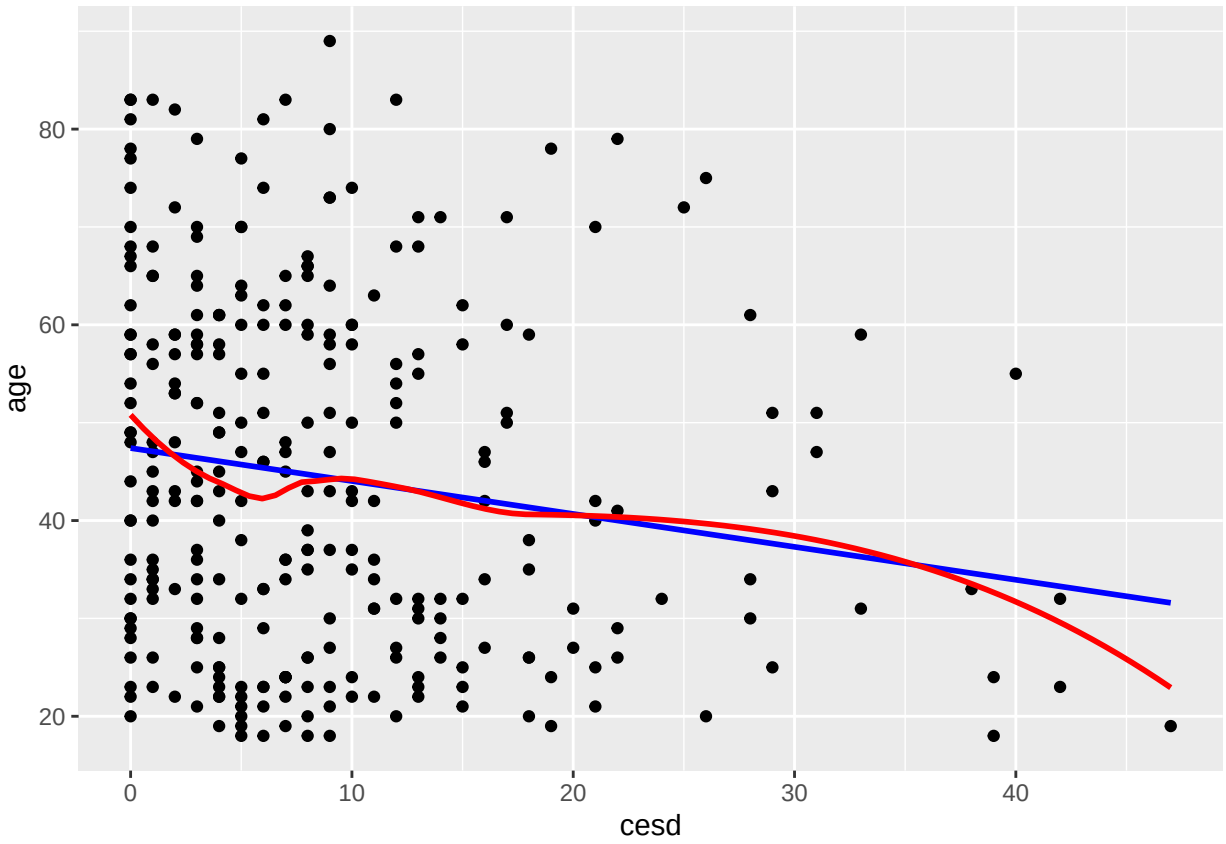
This box plot represents the level of depression each observation has according to whether or not they have a chronic illness. We can see that highest level of depression in observations who have a chronic illness is roughly 13 while the lowest is around 8. People who don't have a chronic illness have a high of about 11 and a low of about 9.

CESD vs. AGE

```
ggplot(depress, aes(x=cesd, y=age)) + geom_point() +
  geom_smooth(se=FALSE, method="lm", color="blue") +
  geom_smooth(se=FALSE, color="red")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



According to the scatter plot, the common cesd level across all ages was about 19 and below. We can see it is very common for a person of any age to have a cesd level of about 3-5.

Conclusion

I found that people with chronic illnesses have a higher chance of having depression and the older you get, the lower your cesd levels will be. You can see these findings in my comparison graphs.