Final Project

Samantha Romero_sromero

9/25/2020

Introduction

For this project the data was acquired from 'https://norcalbiostat.netlify.app/teaching/data/#depression'. This was a potential study on depression in adult residents of Los Angeles County with 294 observations. Within the depression data set cesd, sex, and health will be explored. CESD denotes depression levels from a range of 0 (lowest level) to 60 (highest level). Sex is the gender of the participants and health is categorized as "Excellent", "Good", "Fair", or "Poor". I will be exploring the relationships between level of depression and health status, as well as, level of depression and gender.

```
library(sjPlot)
library(ggplot2)
depress <- read.table(
  "C:/Users/Sam/Documents/Fall2020/math130/Final_Project/depress_081217.txt",
  header=TRUE, sep="\t")
head(depress)
```

```
##   id sex age   marital       educat  employ income relig c1 c2 c3 c4 c5 c6 c7
## 1  1   1  68   Widowed      Some HS Retired      4     1  0  0  0  0  0  0  0
## 2  2   0  58  Divorced Some college      FT     15     1  0  0  1  0  0  0  0
## 3  3   1  45   Married      HS Grad      FT     28     1  0  0  0  0  1  0  0
## 4  4   1  50  Divorced      HS Grad   Unemp      9     1  0  0  0  0  1  1  0
## 5  5   1  33 Separated      HS Grad      FT     35     1  0  0  0  0  0  0  0
## 6  6   0  24   Married      HS Grad      FT     11     1  0  0  0  0  0  0  0
##   c8 c9 c10 c11 c12 c13 c14 c15 c16 c17 c18 c19 c20 cesd cases drink health
## 1  0  0   0   0   0   0   0   0   0   0   0   0   0    0     0     0      2
## 2  0  0   0   0   1   0   0   1   0   1   0   0   0    4     0     1      1
## 3  0  0   0   0   0   0   1   1   1   0   0   0   0    4     0     1      2
## 4  3  0   0   0   0   0   0   0   0   0   0   0   0    5     0     0      1
## 5  3  3   0   0   0   0   0   0   0   0   0   0   0    6     0     1      1
## 6  0  1   0   0   1   2   0   0   2   1   0   0   0    7     0     1      1
##   regdoc treat beddays acuteill chronill
## 1      1     1       0        0        1
## 2      1     1       0        0        1
## 3      1     1       0        0        0
## 4      1     0       0        0        1
## 5      1     1       1        1        0
## 6      1     1       0        1        1
```

Univariant

Variable: Sex

Relabeled the sex variable from 0 to male and 1 to female.

```
depress$sex_fac <- factor(depress$sex, labels= c("Male", "Female"))
table(depress$sex, depress$sex_fac, useNA="always")
```

```
##
##        Male Female <NA>
##   0     111      0    0
##   1       0    183    0
##   <NA>    0      0    0
```

```
plot_frq(depress$sex_fac)+ylab("Number of People")+ xlab("Gender")+
  ggtitle("Gender Distribution")+ scale_fill_manual(values=c("green", "blue"))
```

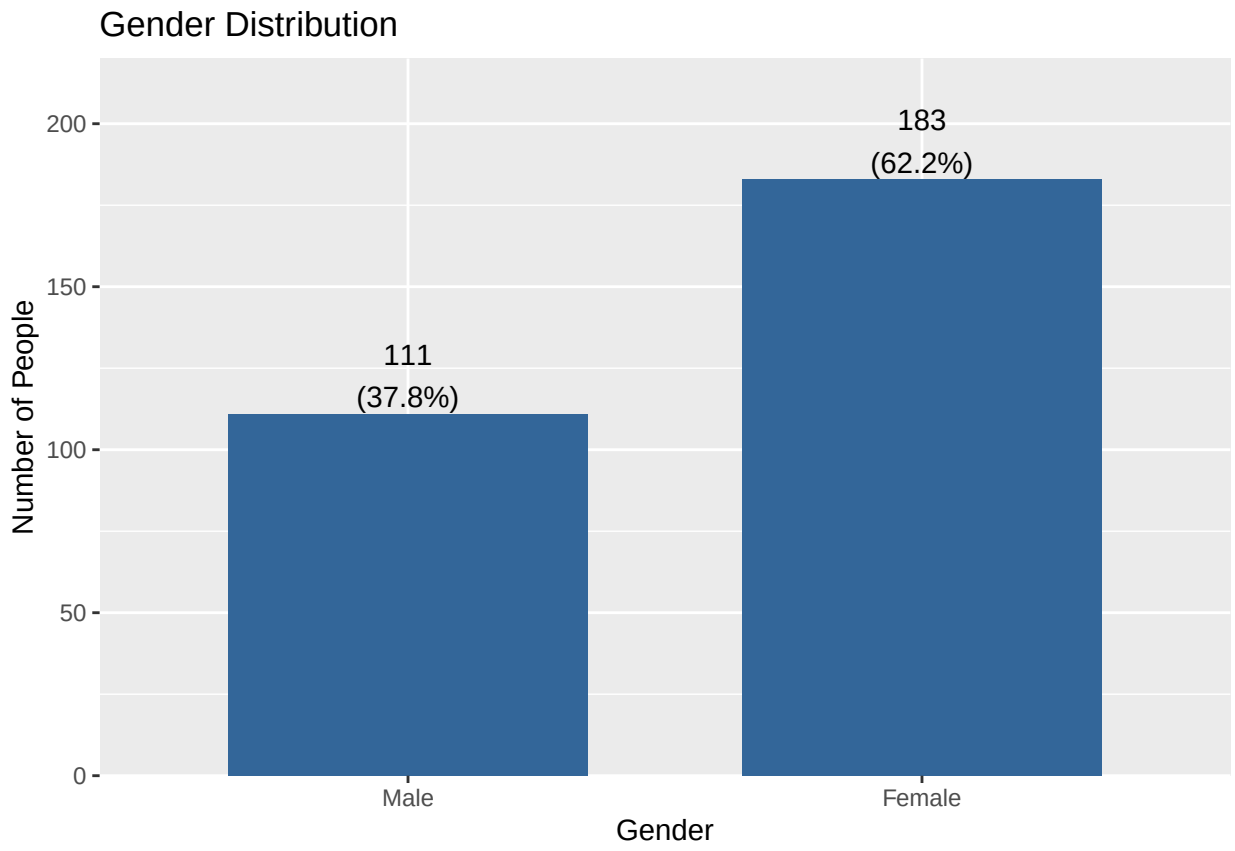

Figure 1. The chart shows that of the 294 observations 111 are males which accounts for 37.8% of the sample size. While females are 183 of the total number of observations, they accounts for 62.2% of the sample size.

Variable: cesd

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

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

```
sd(depress$cesd)
```

```
## [1] 8.823655
```

```
ggplot(depress, aes(x=cesd, fill=cesd))+geom_histogram(color="red")+
  ylab("Number of People")+ xlab("Depression Level")+
  ggtitle("Depression Level Distribution")+ theme_dark()
```

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

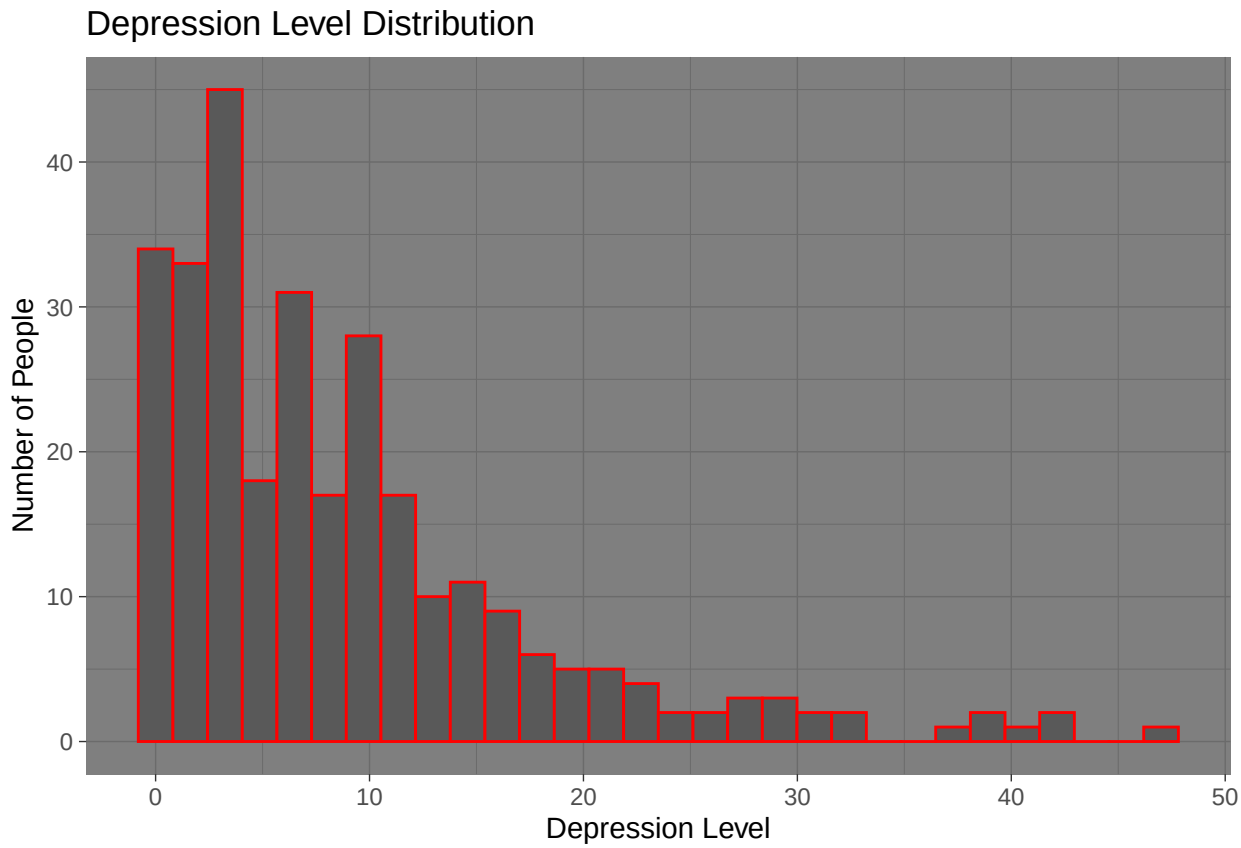

Figure 2. The graph is skewed to the right,which indicates outliers in the higher depression levels. The greatest level of depression in the data set is 47 and there are only a few of them. The average depression level is 8.88 with a standard deviation of 8.82.

Variable: health

```
table(depress$health)
```

```
##
##   1   2   3   4
## 130 115  35  14
```

```
depress$health_fac <- factor(depress$health, labels=c(
  "Excellent", "Good", "Fair", "Poor"))
table(depress$health_fac)
```

```
##
## Excellent      Good      Fair      Poor
##       130       115        35        14
```

```
library(RColorBrewer)
ggplot(depress, aes(x=health_fac, fill= health_fac))+geom_bar()+
  scale_fill_brewer(palette="Set3", guide=FALSE)+ xlab("Health Status")+
  ylab("Number of People")+ggtitle("State of Health")+
  geom_text(aes(label=..count..), stat='count', size= 5)
```

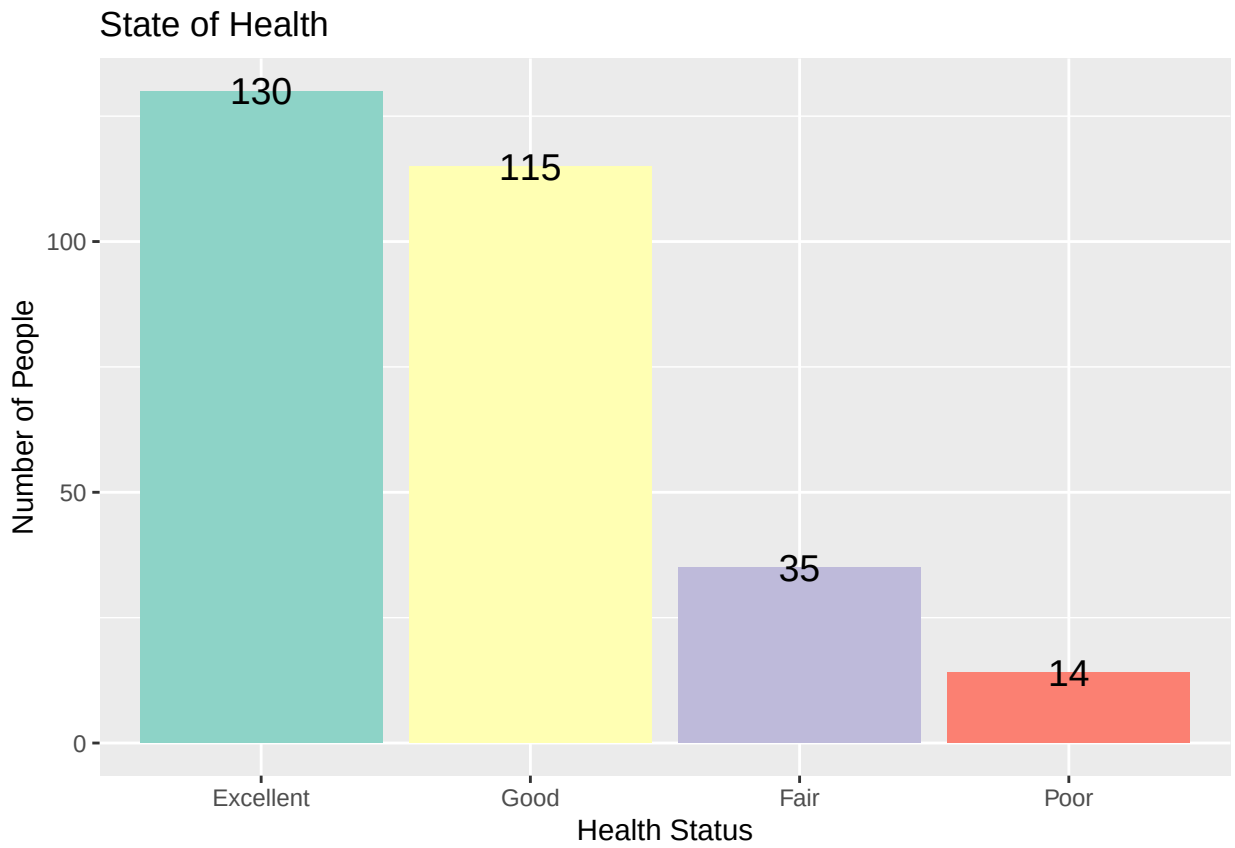

Figure 3. 130 people are in "Excellent" health, 115 people are in "Good" health, 35 people are in "Fair" health and 14 people are in "Poor" health.

Bivariate

Sex vs.CESD

```
ggplot(depress, aes(x=sex_fac, y=cesd, col=sex_fac)) + geom_boxplot()+
  xlab("Gender")+ ylab("Level of Depression")+
  ggtitle("Depression Level Based on Gender")+
  scale_color_manual(values = c("cyan", "magenta"), guide= FALSE)
```

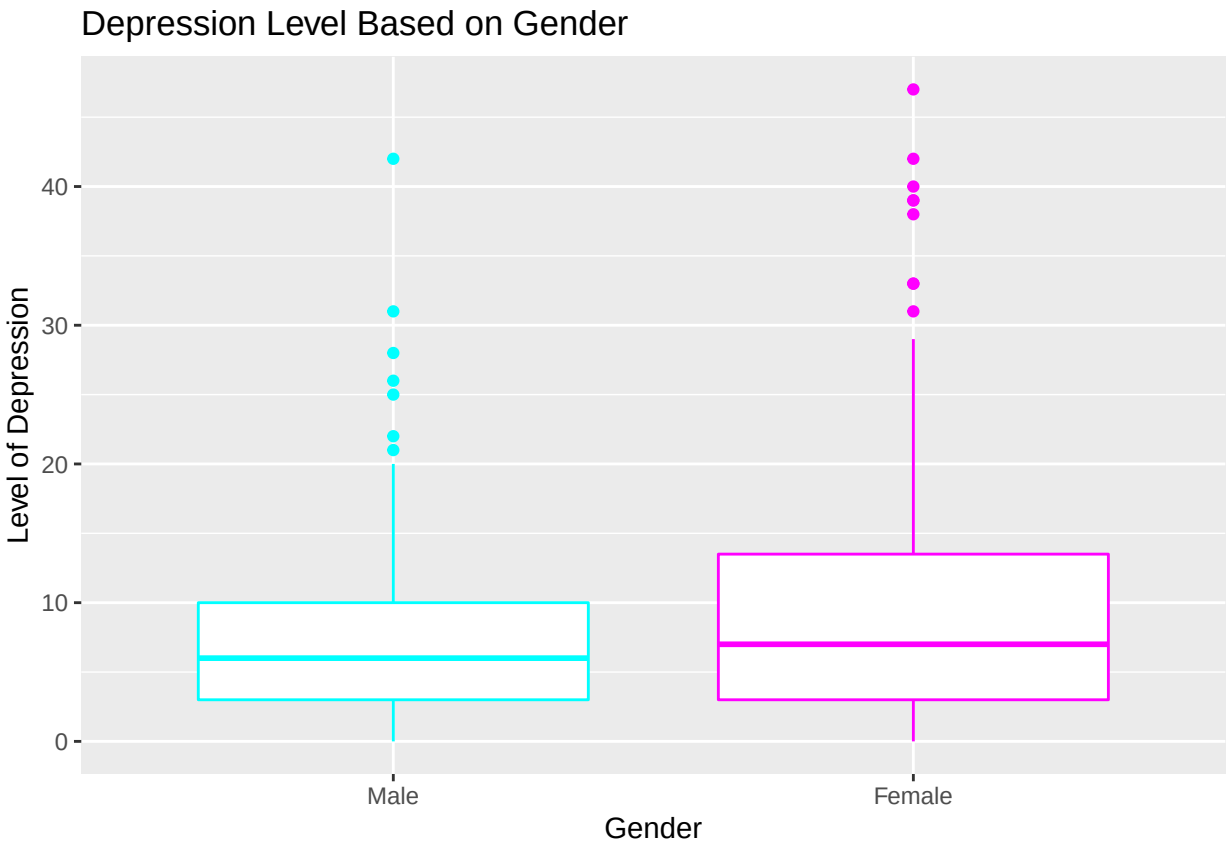

Figure 4. For the male variable the bulk of them have a depression level below 10 and above approximately 4. For the female variable the bulk is just below 15 and above approximately 4. Both have a few possible outliers of depression levels above 20.

Health vs. CESD

```
ggplot(depress, aes(y=cesd,x=health_fac, fill=health_fac))+ geom_boxplot() +
  scale_fill_brewer(palette="Set3", guide=FALSE)+
  xlab("Health Status")+ ylab("Depression Level")
```

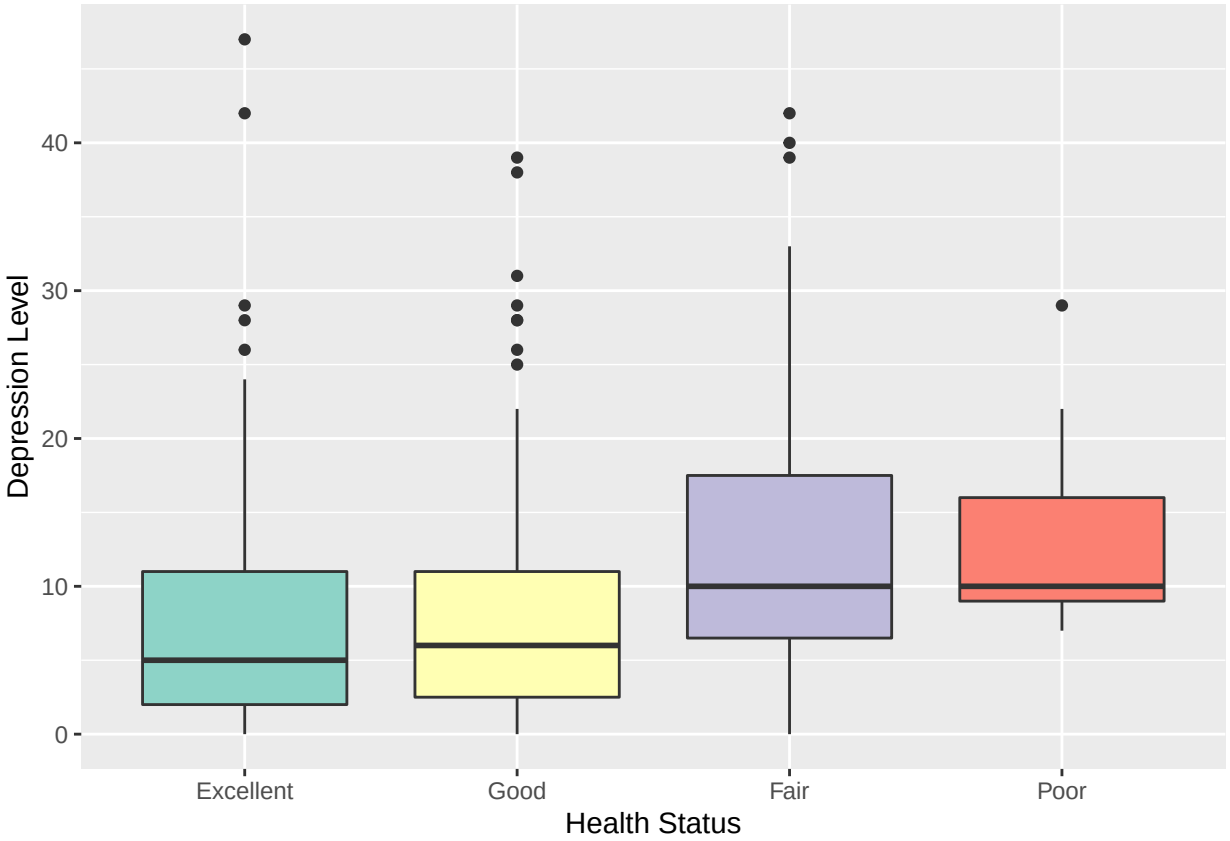

Figure 5. The "Excellent" status bulked at approximately between 2.5 and 10.25 for depression levels. The "Good" health status is bulked approximately between 3 and 10 for depressions. The "Fair" health status has the largest bulk and starts higher than "Excellent" and "Good" with it's bulk between 5 and 17 of the depression levels. The "Poor" health status's bulk is higher than "Excellent" and "Good" as well but has a smaller bulk than the rest. The bulk is between 9 and 16 for depression level.