Homework 5: Exploratory Data Analysis Project

Mikayla Perll

9/27/21

```
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(sjPlot)
```

```
## Registered S3 methods overwritten by 'parameters':
##   method                from
##   as.double.parameters_kurtosis    datawizard
##   as.double.parameters_skewness    datawizard
##   as.double.parameters_smoothness  datawizard
##   as.numeric.parameters_kurtosis   datawizard
##   as.numeric.parameters_skewness   datawizard
##   as.numeric.parameters_smoothness datawizard
##   print.parameters_distribution    datawizard
##   print.parameters_kurtosis        datawizard
##   print.parameters_skewness        datawizard
##   summary.parameters_kurtosis      datawizard
##   summary.parameters_skewness      datawizard
```

```
## Learn more about sjPlot with 'browseVignettes("sjPlot")'.
```

Introduction: Depression

For the exploratory data analysis project I will be analyzing the study of depression in adults living in Los Angeles County. The study looks at variables including sex, age, marital status, education, employment status and more. The questions individuals in the study were asked included general health questions and rating behaviors on a depression scale.

The variables I will be utilizing in this data set will be sex, employment status, cesd, and age.

1.Read in the 'Depression' data set.

```
depress <- read.delim("/Users/mikaylaperll/Desktop/math130/data/depress_081217.txt")
```

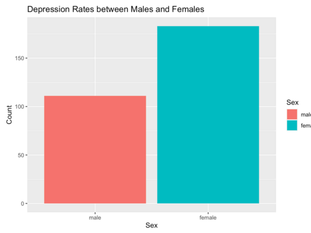
Univariate description of each variable being considered

Sex of Individuals

```
depress$sexrename <- factor(depress$sex, labels=c("male", "female"))
summary(depress$sexrename)
```

```
##   male female
##    111    183
```

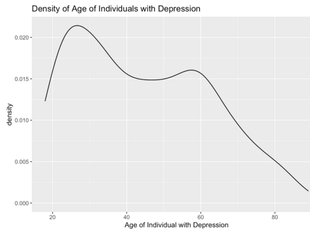
```
ggplot(depress, aes(x=sexrename, fill=sexrename)) + geom_bar() + xlab("Sex") + ylab("Count") + ggtitle("Depressio
n Rates between Males and Females") + scale_fill_discrete(name="Sex")
```



This bar graph shows the amount of males compared to females who are depressed. The data shows that females struggle far more with depression in the Los Angeles County. This study showed there are 183 females and 111 males who were depressed.

Age of Individuals

```
ggplot(depress, aes(x=age)) + geom_density() +xlab("Age of Individual with Depression") + ggtitle("Density of Age
of Individuals with Depression")
```



The density plot here shows the density by age of individuals with depression from this data set. You can see that the majority of younger individuals from 25-35 are depressed while the graph slopes down as age increases. The older individuals did not seem to be depressed.

Employment Status

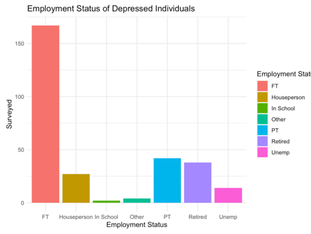
```
table(depress$employ)
```

```
##
##          FT Houseperson   In School       Other          PT     Retired
##         167          27           2           4          42          38
##       Unemp
##          14
```

```
summary(depress$employ)
```

```
##    Length     Class      Mode
##       294 character character
```

```
ggplot(depress, aes(x=employ, fill=employ)) + geom_bar() + xlab("Employment Status") + ylab("Surveyed") + ggtitle
("Employment Status of Depressed Individuals") +scale_fill_discrete(name="Employment Status") + theme_minimal()
```



The table above shows the employment status of all individuals in this study. Most of the participants are full time (167), very few are in school, unemployed, or other. The rest work part time, are retired, or are a house person.

CESD Depression Scores

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

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

This shows the summary statistics for the variable "cesd." The cesd variable is said to use a depression scale from 1-60 based on the questions given. We can see from the summary that the minimum score was 0 and the maximum was 47. The mean was calculated at 8.884.

Bivariate Analysis

Sex and Cesd Scores

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

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

```
summary(depress$sexrename)
```

```
##   male female
##    111    183
```

```
ggplot(depress, aes(x=cesd, fill=sexrename)) + geom_density() + scale_fill_discrete(name="Sex") + xlab("Cesd Scor
e") + ylab("Density") + ggtitle("Relationship between Sex and Cesd scores") + theme_light()
```

According to the graph, there is no significant difference between sex and the cesd scored test. We can see that the males and female density plots overlap, but we cannot tell if the females or males had higher or lower scores on the test.

Age and Cesd Scores

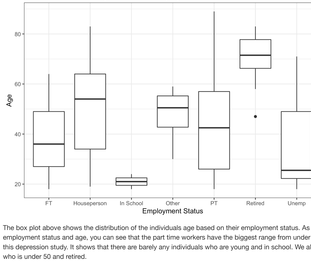
```
summary(depress$age)
```

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

```
summary(depress$employ)
```

```
##    Length     Class      Mode
##       294 character character
```

```
ggplot(depress, aes(y=age, x=employ)) + geom_boxplot() + theme_bw() + xlab("Employment Status") + ylab("Age") + g
gtitle("The Distrubution of Age Based on Employment Status")
```



The box plot above shows the distribution of the individuals age based on their employment status. As there really is no correlation between employment status and age, you can see that the part time workers have the biggest range from under 20 to almost 90 years old participants in this depression study. It shows that there are barely any individuals who are young and in school. We also see that there is an outlier of someone who is under 50 and retired.