Final Project

Nicole Paulson

February 19, 2019

```
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)
```

```
depress <- read.delim("/Users/nicolepaulson/Documents/MATH130/depress_081217.txt", header=TRUE,sep="\t"
```

1. Introduction of the Data

Depresstion Data

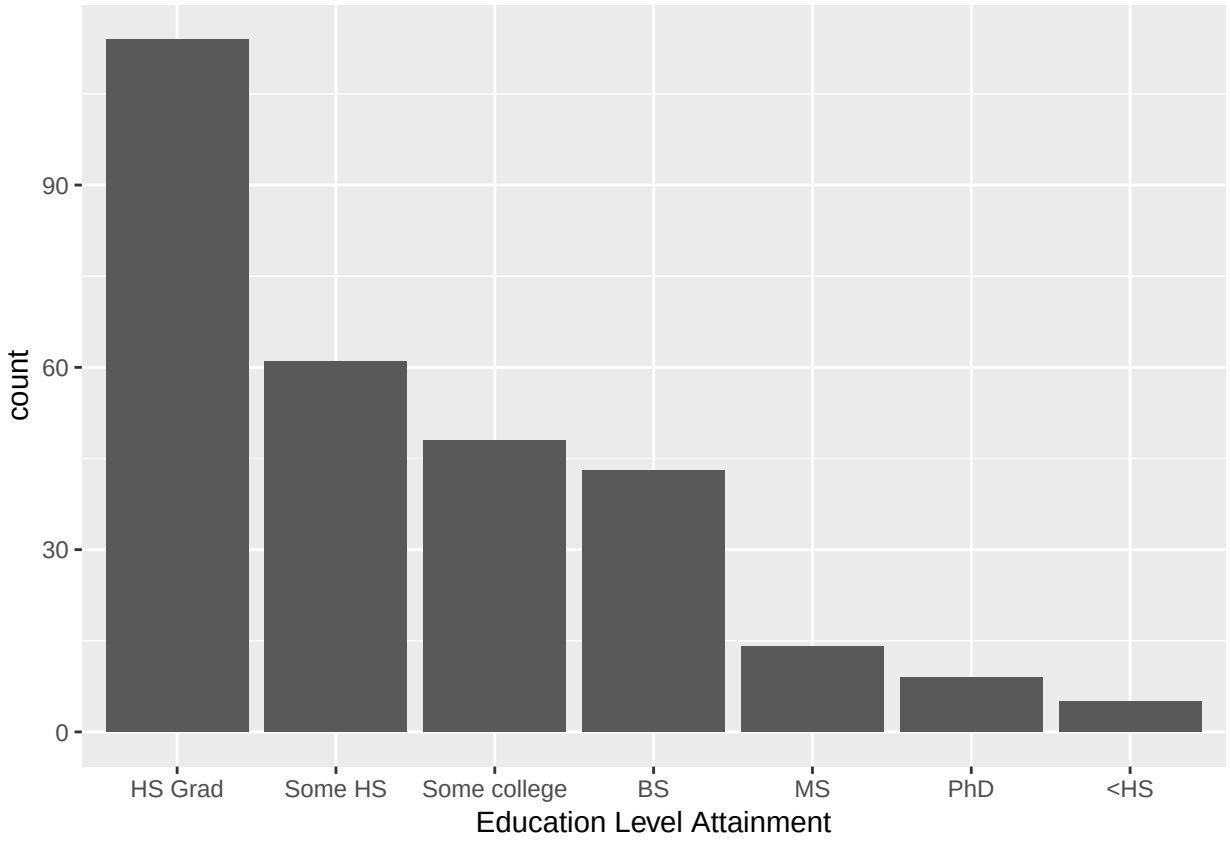
This dataset is a set of interviews of a prospective study about depression among adults in Los Angeles. There are 294 observations and 37 variables. Some of the variables I will be looking at will be depression level (cesd), education (educat), and income.

2. Univariate Descriptions

```
summary(depress$educat)
```

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

```
ggplot(depress, aes(x = forcats::fct_infreq(depress$educat))) + geom_bar() + xlab("Education Level Atta
```



The educat variable, which is categorical, is a measure of education level attainment among those interviewed for the dataset. The most represented group in this dataset is high school graduates (114 people), and the lowest represented group is less than high school (5 people).

```
summary(depress$income)
```

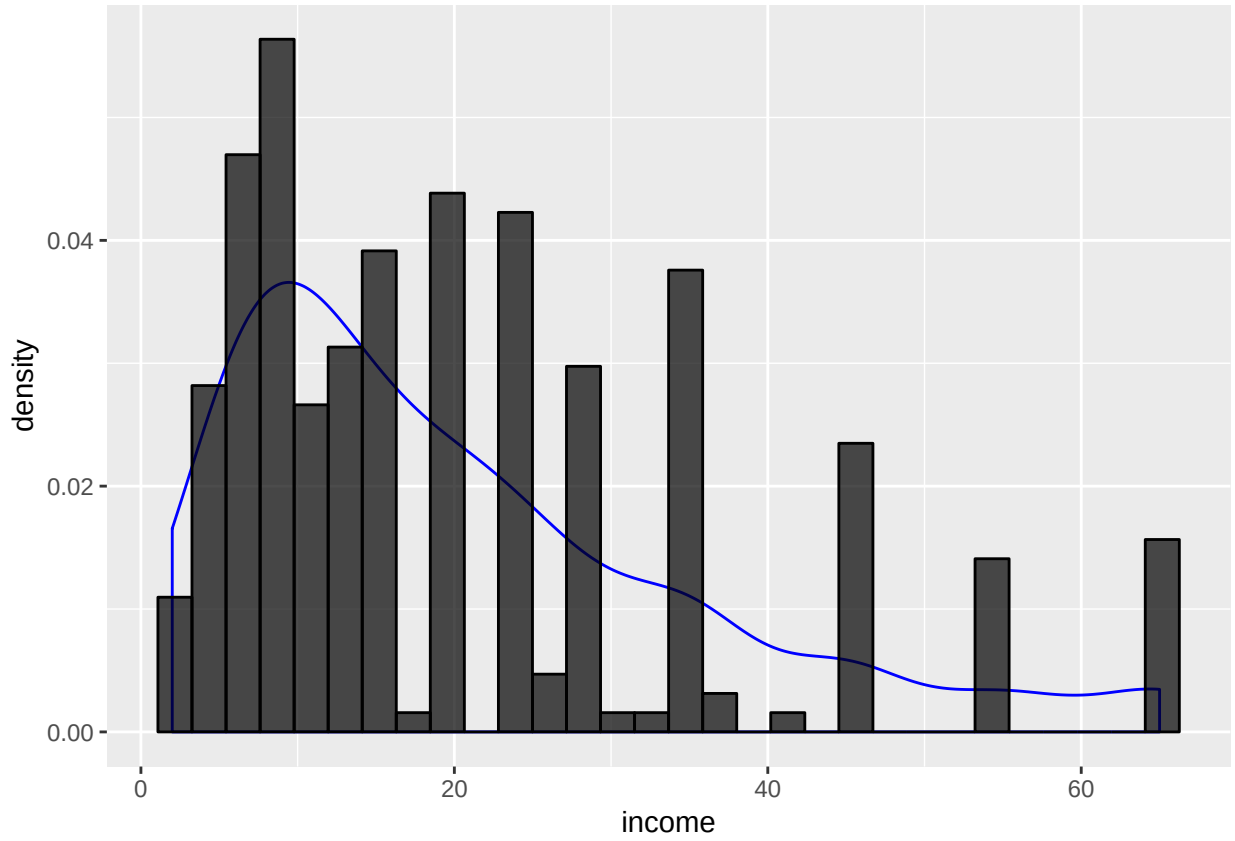
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    9.00   15.00   20.57   28.00   65.00
```

```
mean(depress$income)
```

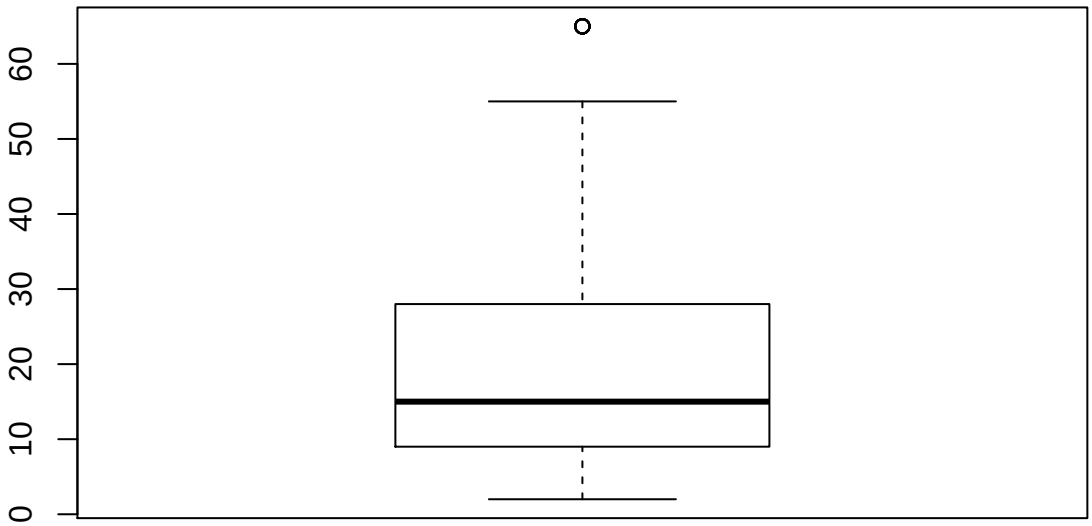
```
## [1] 20.57483
```

```
ggplot(depress, aes(x=income)) + geom_density(col="blue") +
  geom_histogram(aes(y=..density..), colour="black", fill="black", alpha = 0.7)
```

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




```
boxplot(depress$income)
```



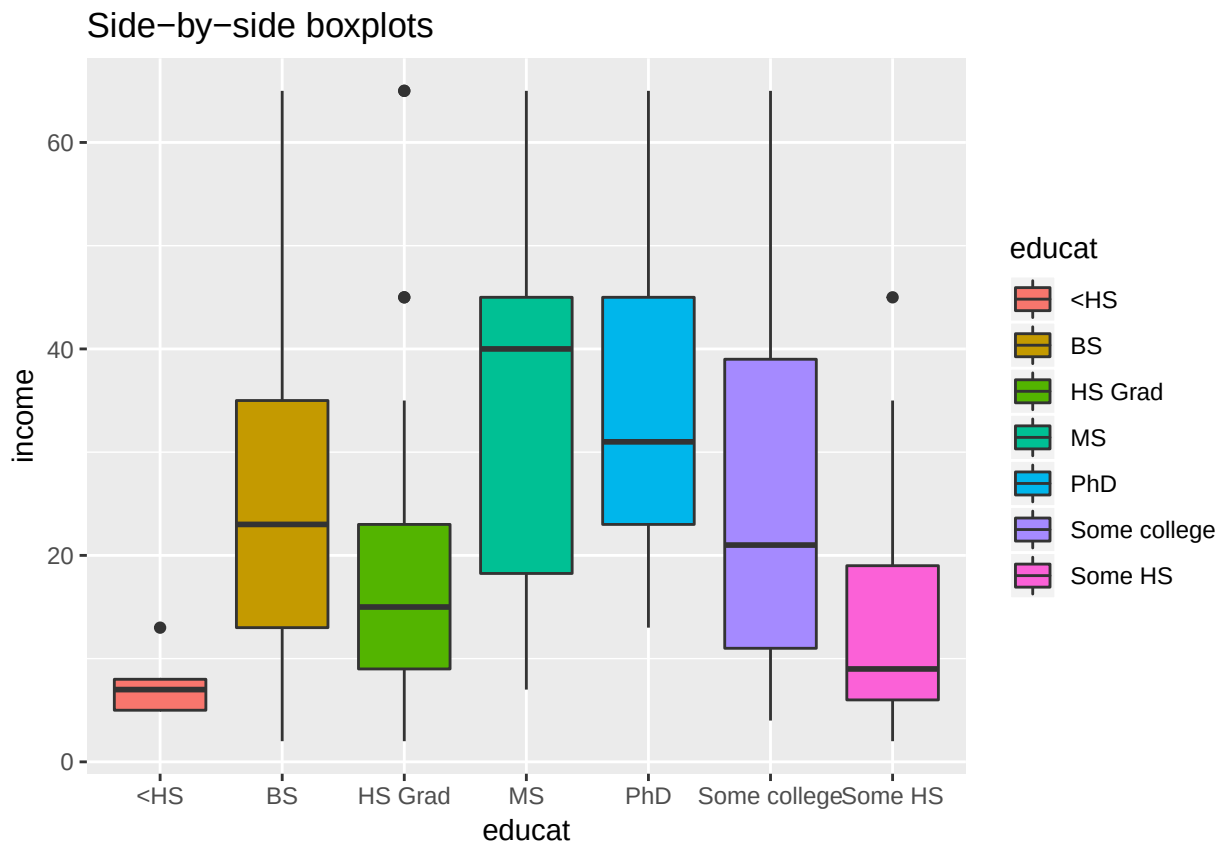
The income variable, which is quantitative, has a range from 2 to 65. Meaning the lowest income is $2,000 and the highest income is $65,000 among the people in this interview sample. The mean, or average income is $20,574, and the histogram is approximately skewed right. It is unimodal, with most of the income between 2,000 to 20,000. The box plot shows that the median is 15 (or $15,000), and there is an outlier at 65 (or $65,000).

3. Bivarariate Descriptions

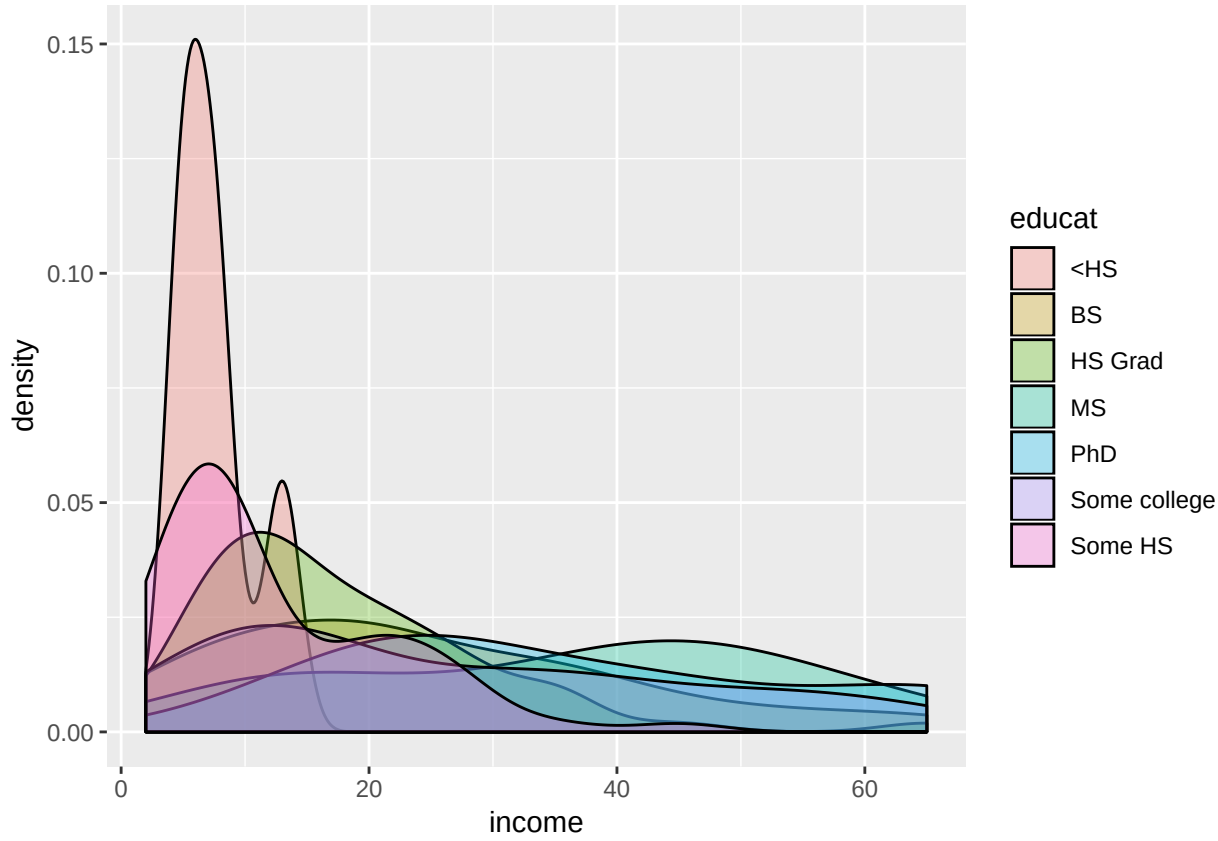
```
depress %>% group_by(educat) %>% summarise(min=min(income),
                                           max=max(income))
```

```
## # A tibble: 7 x 3
##   educat         min   max
##   <fct>        <dbl> <dbl>
## 1 <HS              5    13
## 2 BS               2    65
## 3 HS Grad          2    65
## 4 MS               7    65
## 5 PhD             13    65
## 6 Some college     4    65
## 7 Some HS          2    45
```

```
ggplot(depress,
       aes(y=income,
           x=educat,
           fill=educat)) +
  geom_boxplot() +
  ggtitle("Side-by-side boxplots")
```




```
ggplot(depress, aes(x=income, fill=educat)) + geom_density(alpha=.3)
```



The density plot shows that those with less than a high school education have the lowest income. It is unimodal and skewed right. The mean, or average income, is approximately $20,574. The side-by-side boxplot shows that those with PhDs have the highest range of income, from $13,000 to $65,000. The only groups that make the highest amount of income are those with some high college to those with PhDs, however there are outliers among high school graduates who also make $65,000.

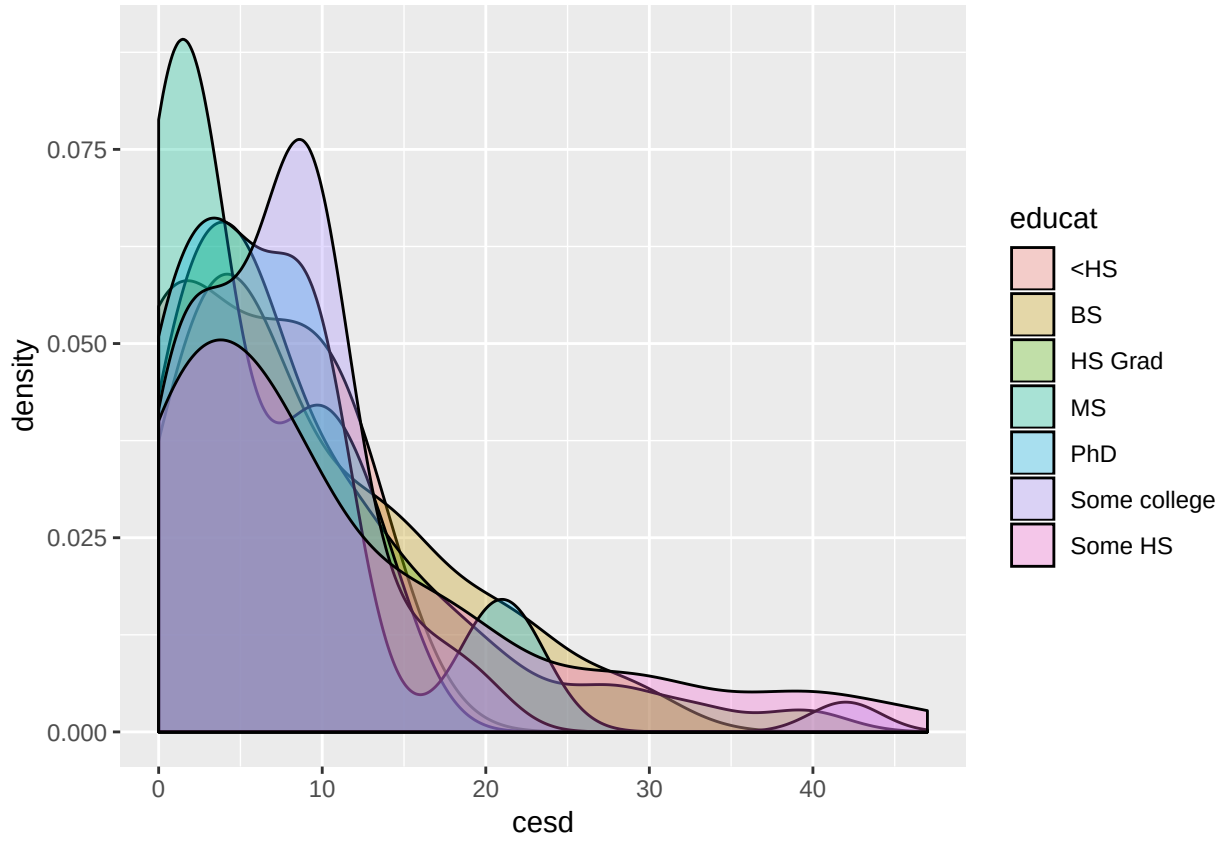
```
summary(depress$cesd)
```

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

```
mean(depress$cesd)
```

```
## [1] 8.884354
```

```
ggplot(depress, aes(x=cesd, fill=educat)) + geom_density(alpha=.3)
```



The mean of cesd, which is a continuous variable that measures how people have felt or behaved in the past week, is 8.88. The density plot is unimodal and skewed right. Cesd ranges from 0 (lowest amount of depression) to 60 (highest amount of depression). In this particular dataset, the cesd ranges from 0-47. I tested educational attainment level against cesd to see the relationship. Most educational attainment levels fall around the mean of the 8.8 cesd, however, those with masters have the lowest amount of depression because of the peak around 20-25.