Exploratory Data Analysis

James Lawrence

03/08/2022

```
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(RColorBrewer)
library(knitr)
library(ggplot2)
```

Data Set

(Description from Dr. Donatello's Teaching Website, linked in 'References')

The depression data set is from the first set of interviews of a prospective study of depression in the adult residents of Los Angeles County and includes 294 observations. More details on the origin and study design can be found in Practical Multivariate Analysis, 5th edition by Afifi, May and Clark.

```
Depression<-read.table("/Users/jameslawrence/Documents/math130/data/depress_081217.txt",
dim(Depression)
```

```
## [1] 294  37
```

Introduction

This data presents a unique opportunity to explore a chronic disease's relationship with a number of variables, including income level. I anticipate seeing a negative correlation between income level and depression, and a similar relationship as a function of income.

UNIVARIATE DESCRIPTION:

INCOME:

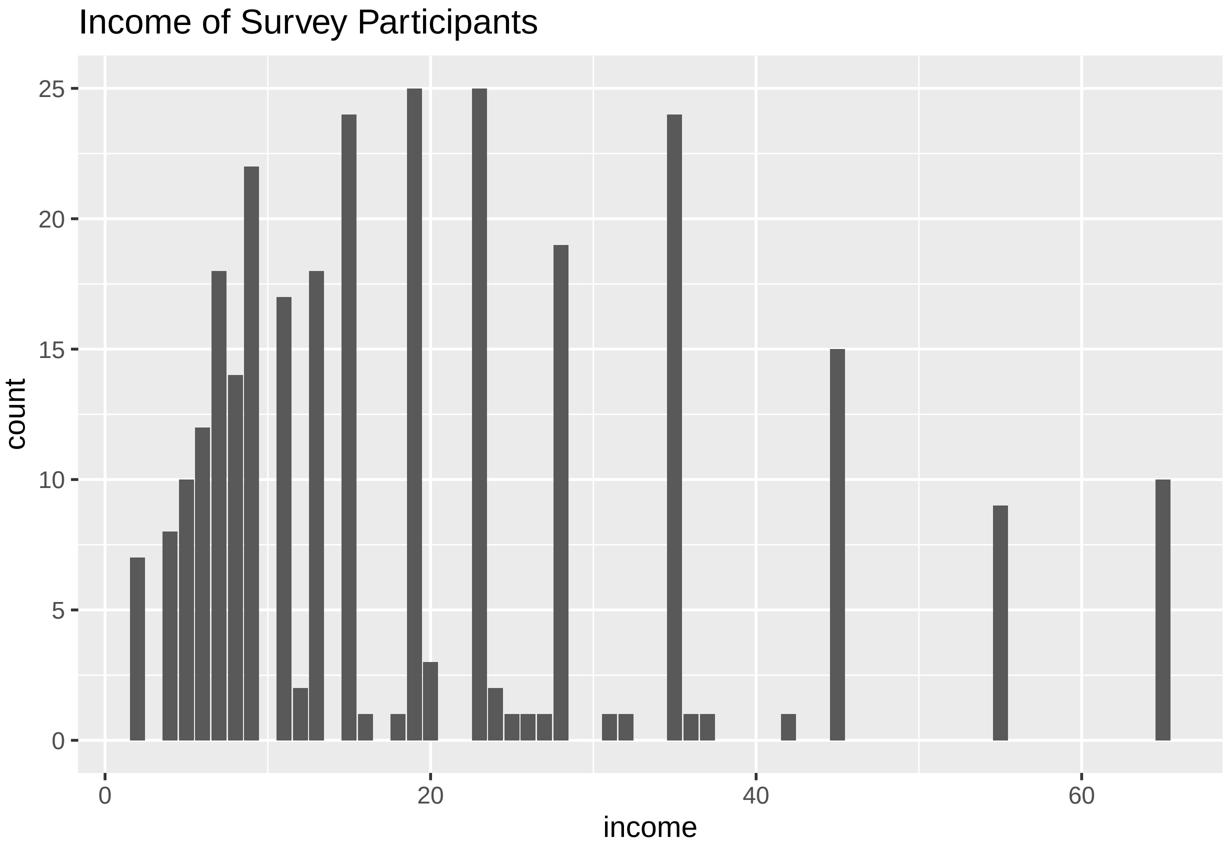
The data below shows the the distribution of average income per year in thousands of dollars across the data set, establishing a maximum income of $65,000 and a mean of approximately $20,000 annually for the participants from Los Angeles County.

```
summary(Depression$income)
```

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

The plot below shows the general distribution of income among survey participants.

```
ggplot(Depression, aes(x=income)) +
  geom_bar() +
  ggtitle("Income of Survey Participants")
```



Income here is shown to be skewed leftward, and thus towards a much lower amount by count of survey participants, with only 18 individuals making more than $50,000 per year. This is reflected in both the summary table and the graph above.

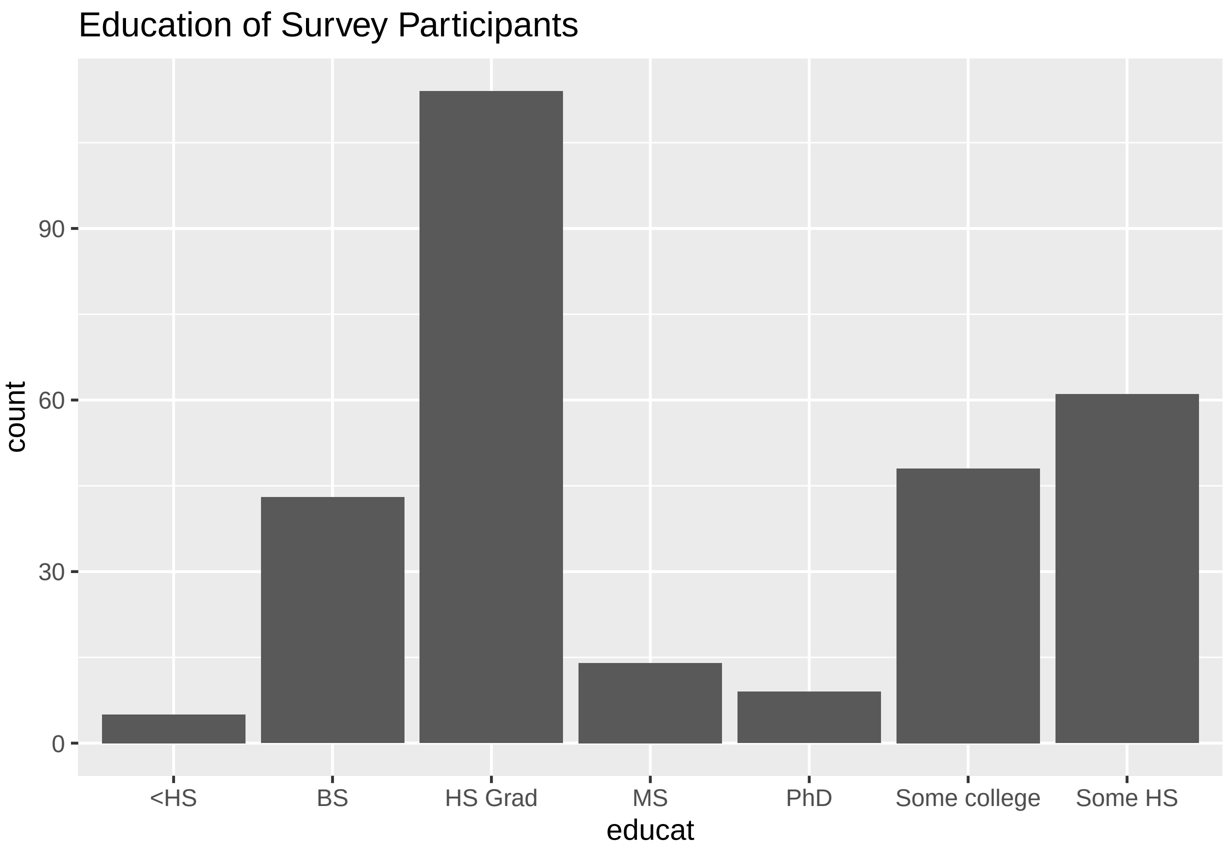
EDUCATION:

The plot and table below details the education level by number of participants across the individuals surveyed.

```
table(Depression$educat)
```

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

```
ggplot(Depression, aes(x=educat)) +
  geom_bar() +
  ggtitle("Education of Survey Participants")
```



The number of PhD and Master's Degree holders within this data set is also relatively low, showing a total of 23 individuals holding either degree out of a total of 294 individuals surveyed, or about 8% of our sample.

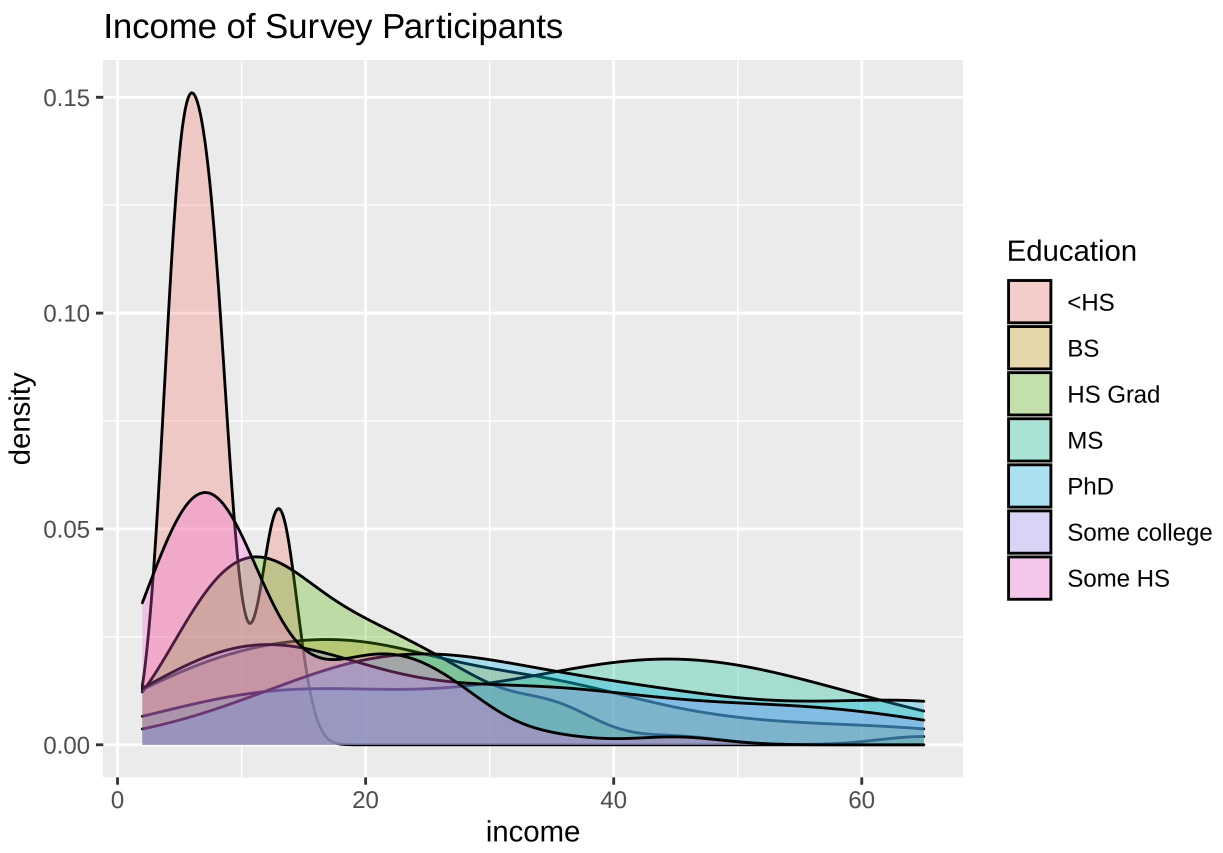
##Bivariate Exploration:

Now that we've got an idea of the demographics that were examined in this study, we can explore the relationship between these variables and the extent of depression.

Income, Education, and Depression

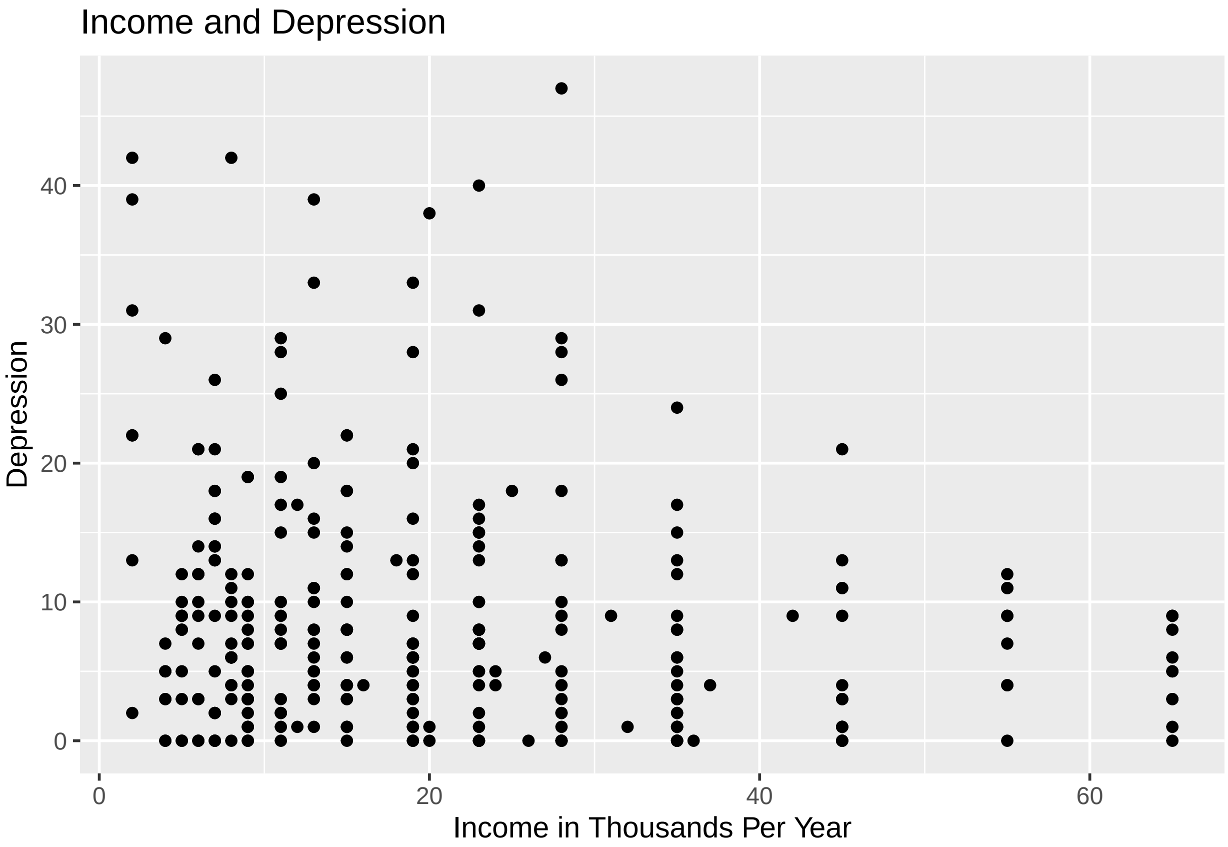
The plot below describes the income level of survey participants as a function of education level, showing a high amount of individuals with an education below high school making less than $20,000 annually.

```
ggplot(Depression, aes(x=income, fill=educat )) + geom_density(alpha=.3) + scale_fill_discrete("Educati
  ggtitle("Income of Survey Participants")
```



An examination of depression and income is shown below:

```
ggplot(Depression, aes(x=income, y=cesd)) + geom_point() +
  ggtitle("Income and Depression") +
  xlab("Income in Thousands Per Year") +
  ylab("Depression")
```



It's apparent that the majority of depression as shown by CESD score is relegated to lower income.

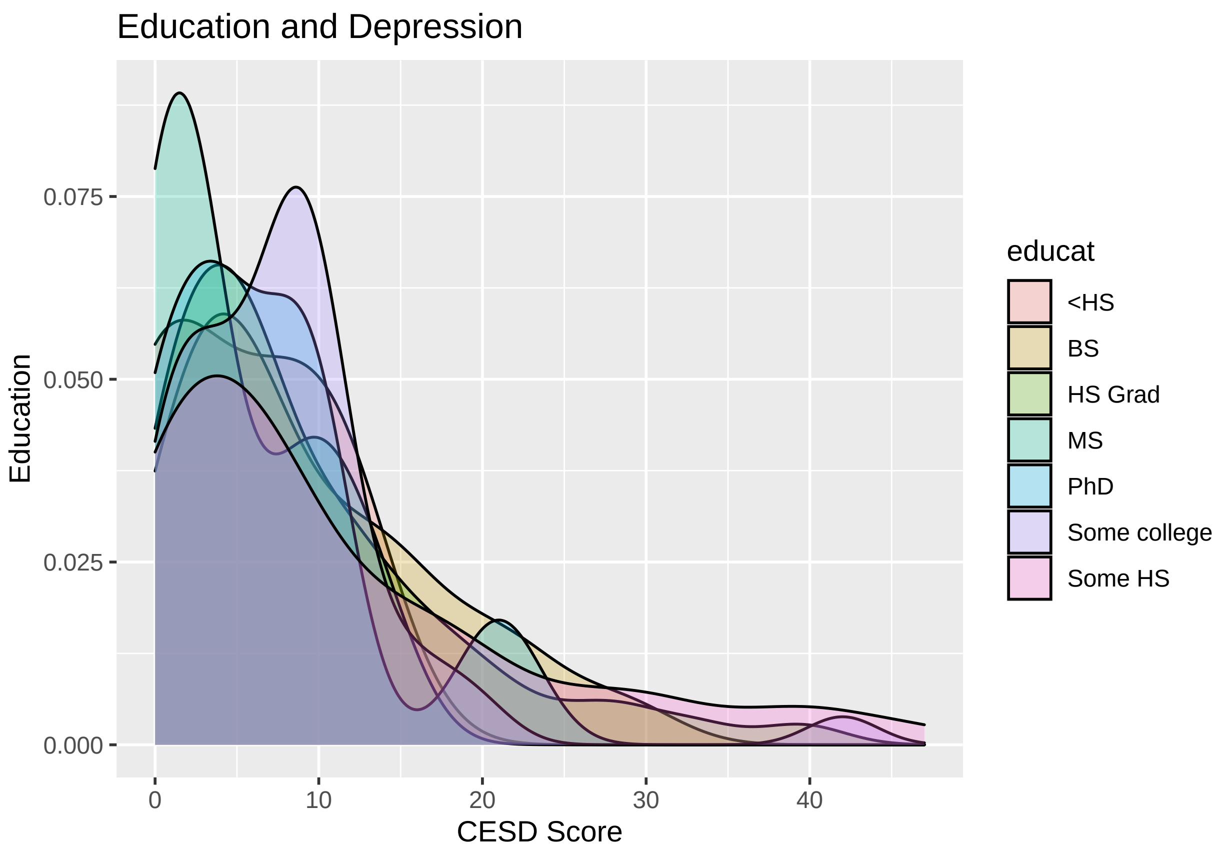
```
mean(Depression$income)
```

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

But, as shown by the mean above, this may also be skewed as factor of average income.

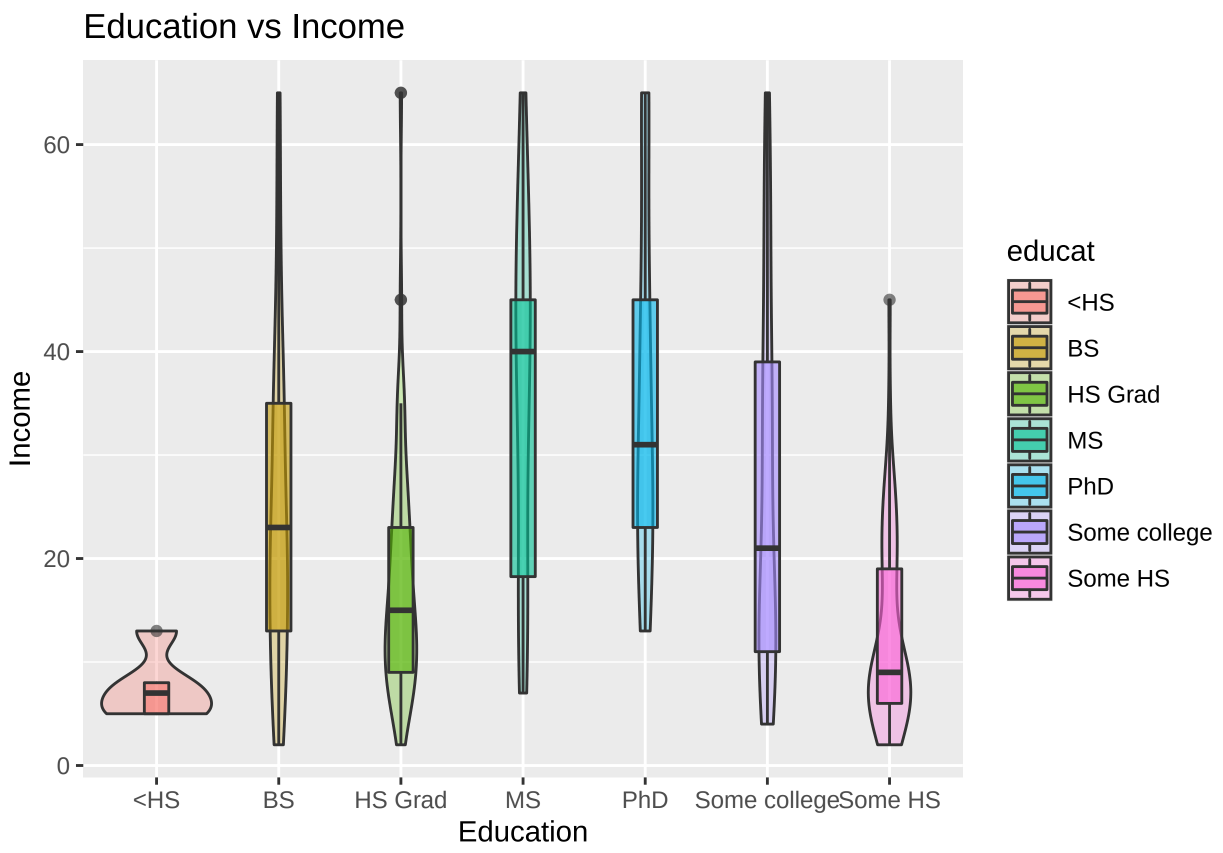
Education level and depression are examined in a similar manner, showing increased CESD with decreased income:

```
ggplot(Depression, aes(x=cesd, fill=educat)) + geom_density(alpha = .25) +
  ggtitle("Education and Depression") +
  xlab("CESD Score") +
  ylab("Education")
```



When taken in comparison, and factoring in the notion that lower income may be correlated to lower levels of education, there may be an association between the two variables. Further analysis is conducted below between the two variables:

```
ggplot(Depression, aes(x=educat, y=income, fill=educat)) +
        geom_violin(alpha=.3) +
        geom_boxplot(alpha=.6, width=.2) +
          xlab("Education") +
            ylab("Income") +
              ggtitle("Education vs Income")
```



Which shows that the MS and PhD routes tend to pay off.

In short, seeing the relationship between income and education shows the observer of the data the possible relationship between education, income, and depression that might arise. With regards to the hypothesized connections stated earlier, there seems to be support for negative correlations between both depression and income, as well as between depression and education level.

References

```r
links<-c("https://www.norcalbiostat.com/data/")

links
```

```
## [1] "https://www.norcalbiostat.com/data/"
```