Exploratory Data Analysis Project

Justin Ricketts

March 7, 2022

Introduction

```
DEPRESS<-read.table("/Users/justinricketts/Documents/College Work/MATH 130/Data/depress_081217.txt",
                    header=TRUE,sep="\t")
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 ...
```

```
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)
```

The data set that I chose to analyze is the Depression, which is from interviews of adults in Los Angeles County that have depression. There were 294 observations and 37 variables, but I will only be selecting 2. The two that I selected are Income and Level of Depression. Income describes the income in thousands of dollars per year for the adult. Level of Depression describes to what level the adult is depressed, with 0 being the lowest level and 60 being the highest level.

Univariate Analysis of Variables

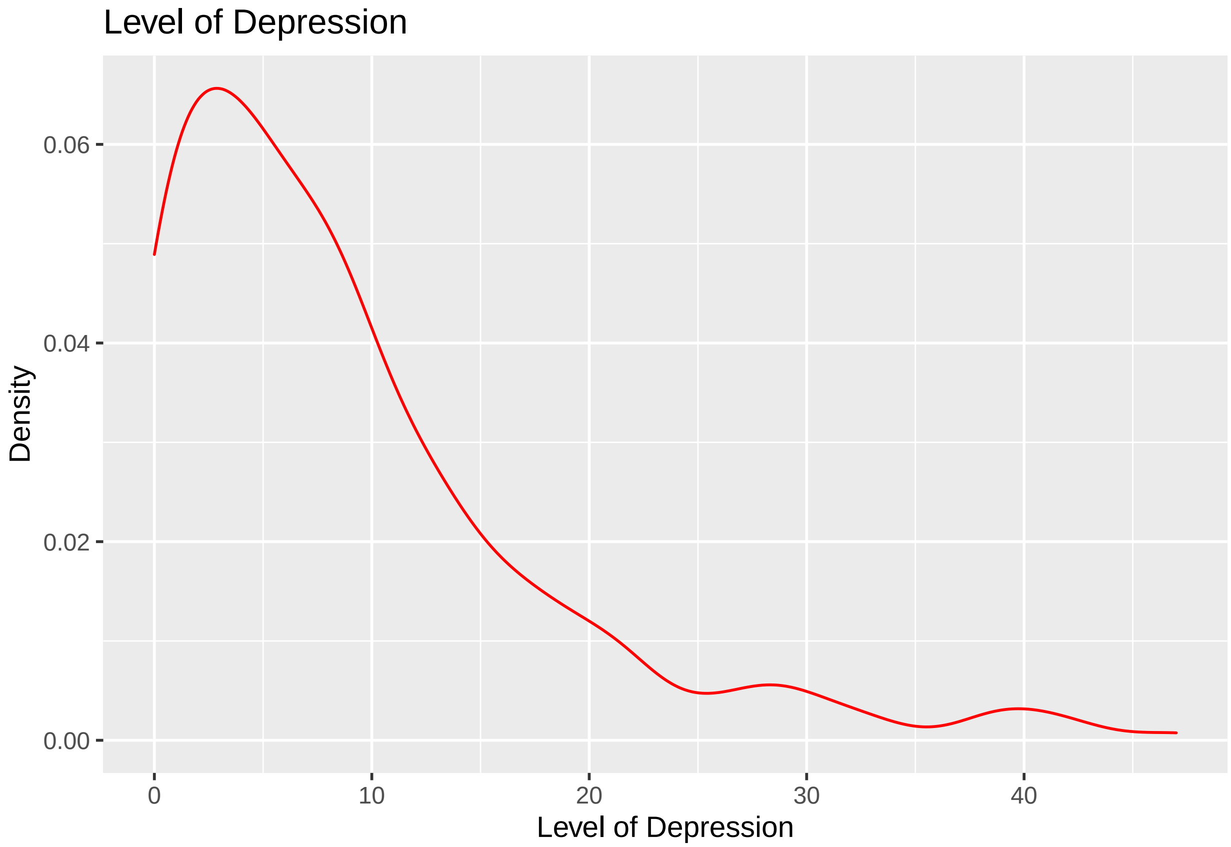
1. Level of Depression

```
summary(DEPRESS$cesd)
```

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

These summary statistics explain that the median Level of Depression among the observations was 7.000, while the highest level of depression was 47.000 and the lowest level was 0.000.

```
ggplot(DEPRESS, aes(x=cesd))+geom_density(col="red") +
  xlab("Level of Depression")+ylab("Density") +
  ggtitle("Level of Depression")
```



The graph shows that most of the participants had a low or no level of depression. We see the density decrease as the depression level increased.

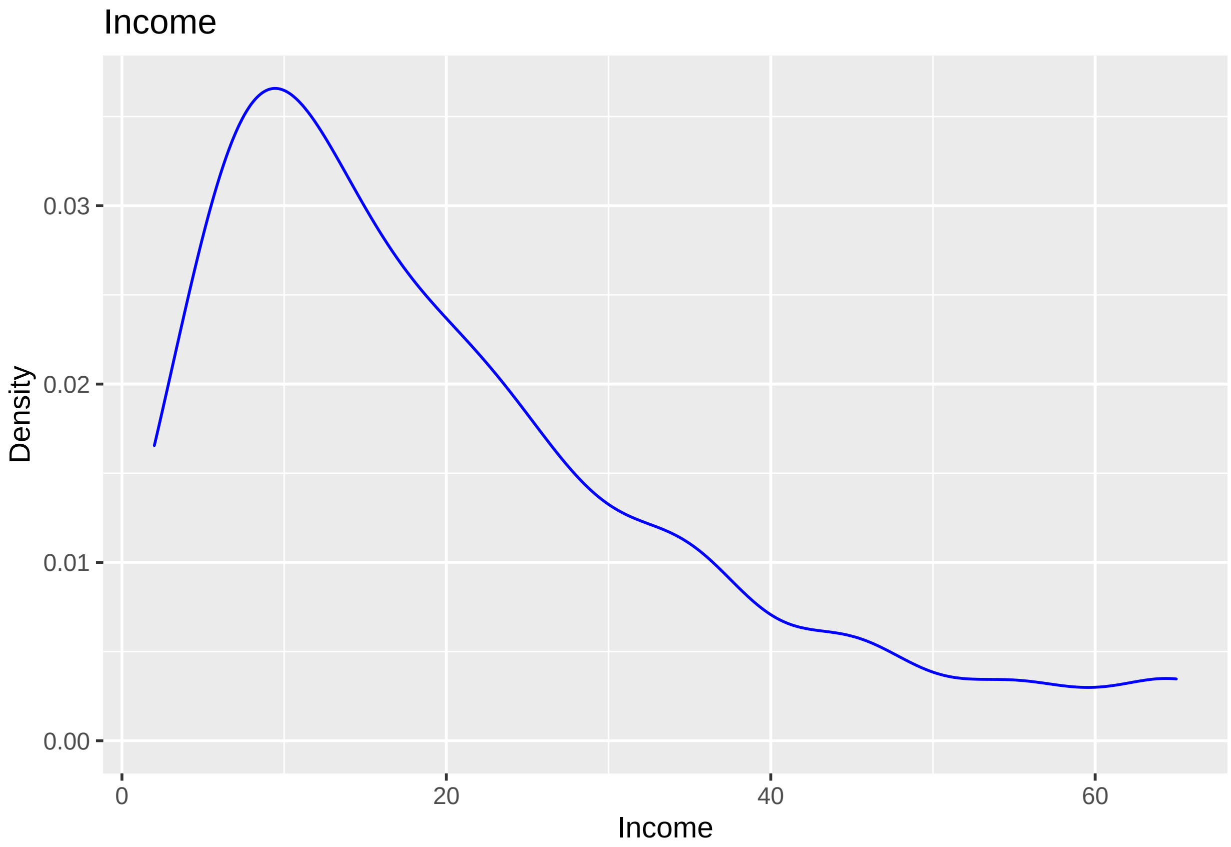
2. Income

```
summary(DEPRESS$income)
```

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

These statistics show that the median level is 15.00, and the highest level was 65.00 and the lowest level was 2.00.

```
ggplot(DEPRESS, aes(x=income))+geom_density(col="blue") +
  xlab("Income")+ylab("Density") +
  ggtitle("Income")
```



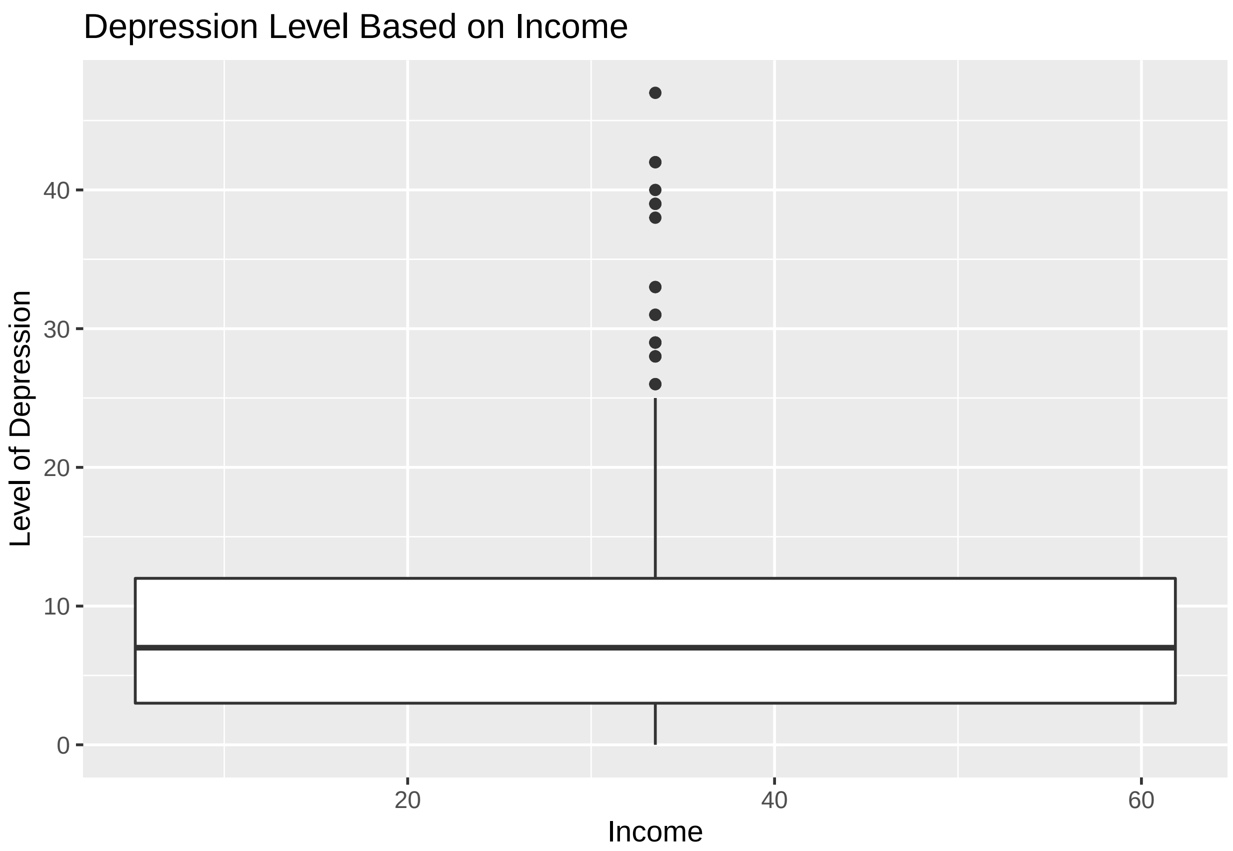
This graph shows the income that the adults made per year in thousands of dollars.

Bivariate Analysis

Level of Depression & Income

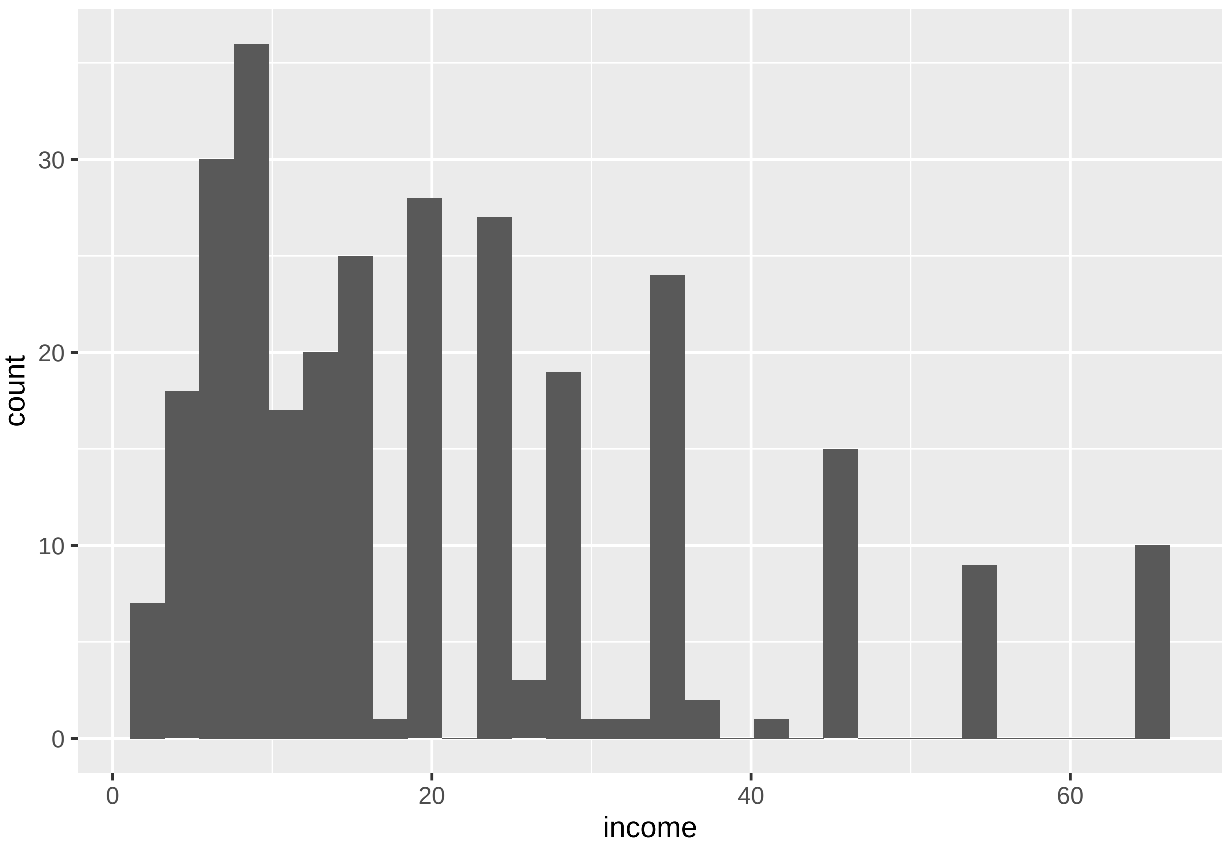
```
ggplot(DEPRESS, aes(x=income, y=cesd)) + geom_boxplot()+ xlab("Income")+ylab("Level of Depression")+
  ggtitle("Depression Level Based on Income")+scale_color_manual(values = c("red", "blue"),
                                                                  guide= "none")
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



```
ggplot(DEPRESS,aes(x=income))+geom_histogram()
```

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



The graph shows that the income was the highest between 0 and 20. This histogram is consistent with the results that were discovered in my boxplot.

Conclusion

I discovered that income did not necessarily affect the level of depression that the adults experienced. The boxplot represents the depression level based off the income and the histogram shows the number of adults and their annual income in thousands of dollars. There is no sufficient evidence to justify that income and level of depression are correlated.