data analysis project

2024-02-29

JACK FERNEAU

Introduction: For my Data Analysis Project, I will be looking at the Depression Data. This data set comes from a first set of interviews in a prospective study of depression in the adult residents of Los Angeles County. It shows depression levels among 294 male and female and within this data set, there are 37 variables. For this data set, I am choosing the variables CESD, drink, and income. I would like to test to see who is more likely to get depression as well to see the correlation of different variables that are common in depressive people.

```
library(ggplot2)
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
```

```
depress <- read.delim("/Users/jferneau/Downloads/depressiondata.txt",header = TRUE, sep = "\t")
```

Univariate Exploration:

For my first variable, I wanted to look at the Income (in thousands of dollars per year) of those who are depressed, to see if there is a direct correlation between them.

```
table(depress$INCOME)
```

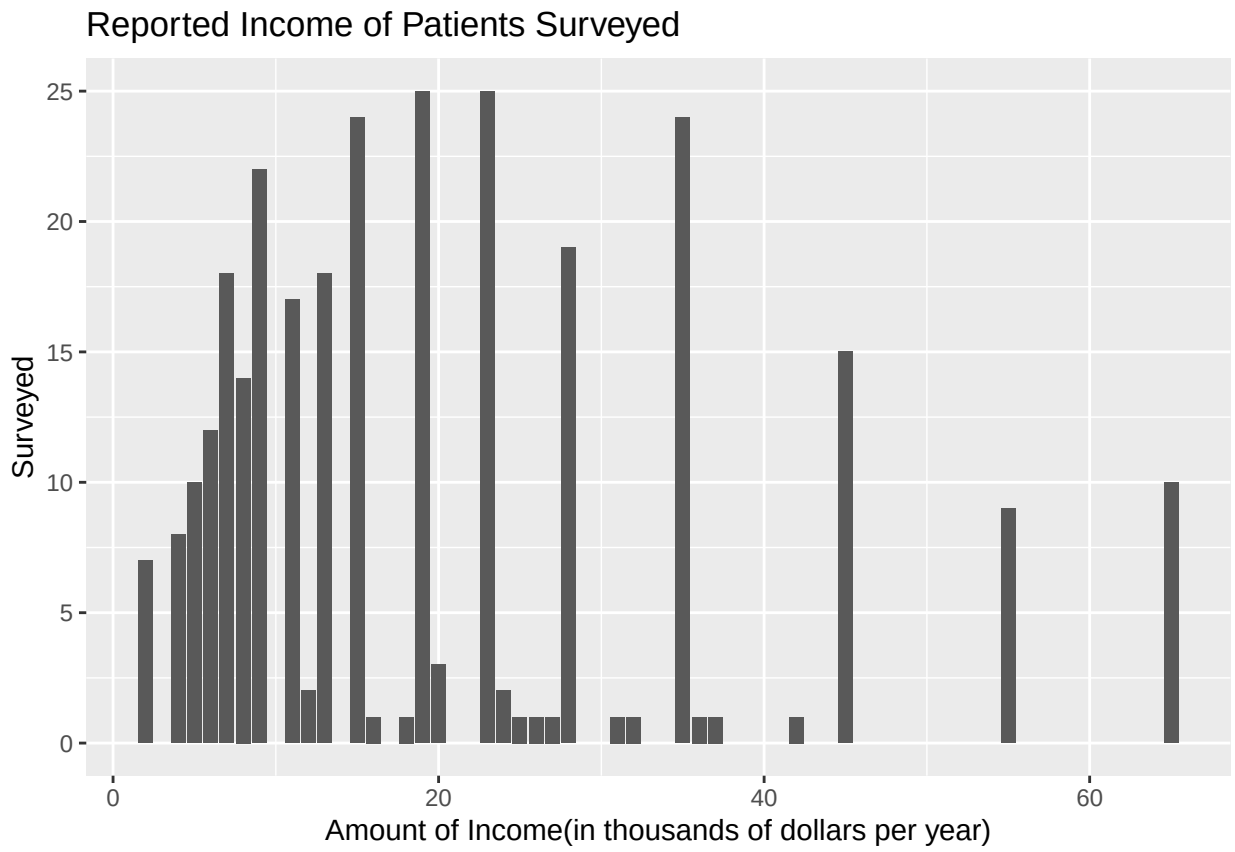
```
##
##  2  4  5  6  7  8  9 11 12 13 15 16 18 19 20 23 24 25 26 27 28 31 32 35 36 37
##  7  8 10 12 18 14 22 17  2 18 24  1  1 25  3 25  2  1  1  1 19  1  1 24  1  1
## 42 45 55 65
##  1 15  9 10
```

```
summary(depress$INCOME)
```

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

This illustrates how amount people make who took part of this study, with the average being 20,570/year and the minimum being 2,000/year while the maximum is 65,000/year.

```
ggplot(depress, aes(x=INCOME)) + geom_bar() + xlab("Amount of Income(in thousands of dollars per year)")
```



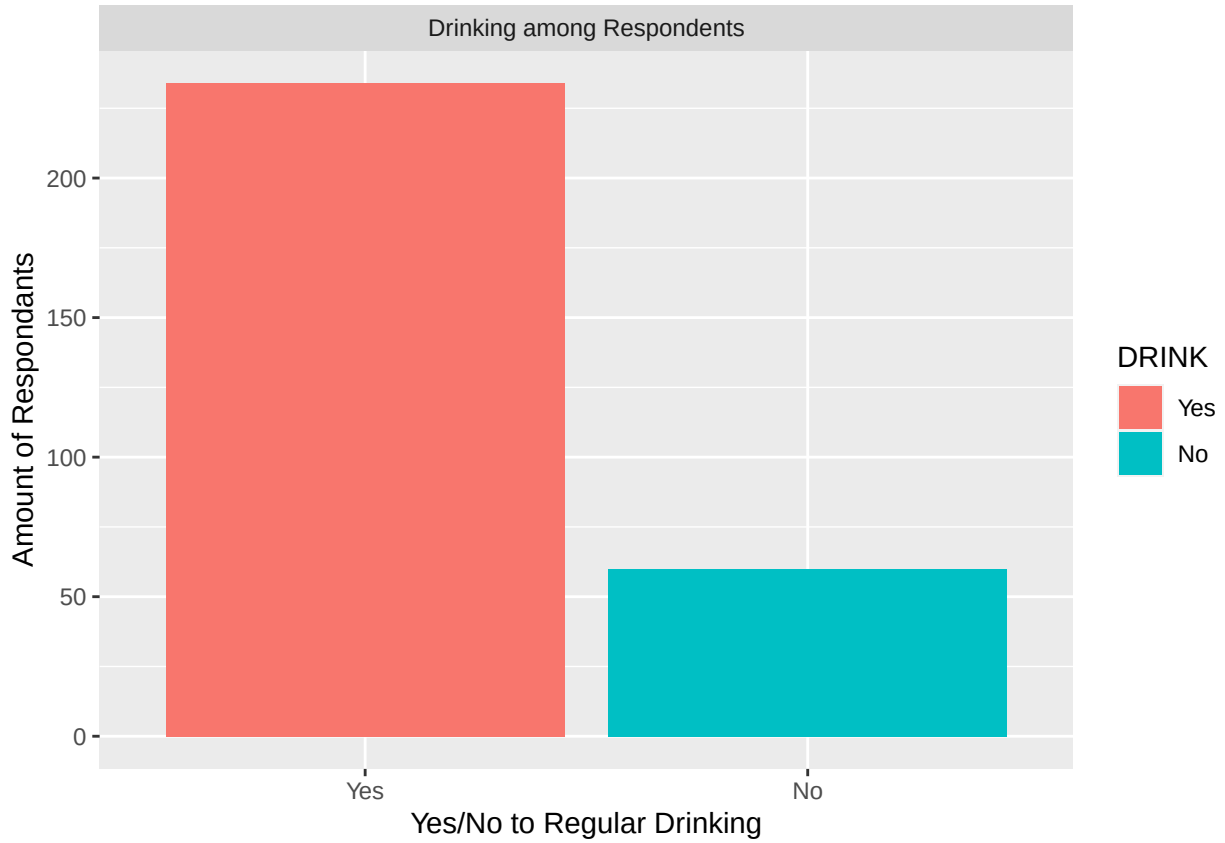
For my second variable, I wanted to look at whether or not drinking is a common variable among those suffering from depression

```
depress$DRINK <- factor(depress$DRINK, labels = c("Yes", "No"))
table(depress$DRINK)
```

```
##
## Yes  No
## 234  60
```

(when asked if the patients were regular drinkers)

```
ggplot(depress, aes(DRINK, fill = DRINK)) + geom_bar() + ylab("Amount of Respondants") + xlab("Yes/No t
```



From the results, it seems that a majority of the respondents are regular drinkers compared to those that do not drink regularly.

Now lets take a look at the reported depression levels of the respondents by investigating the CESD variable.

```
summary(depress$CESD)
```

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

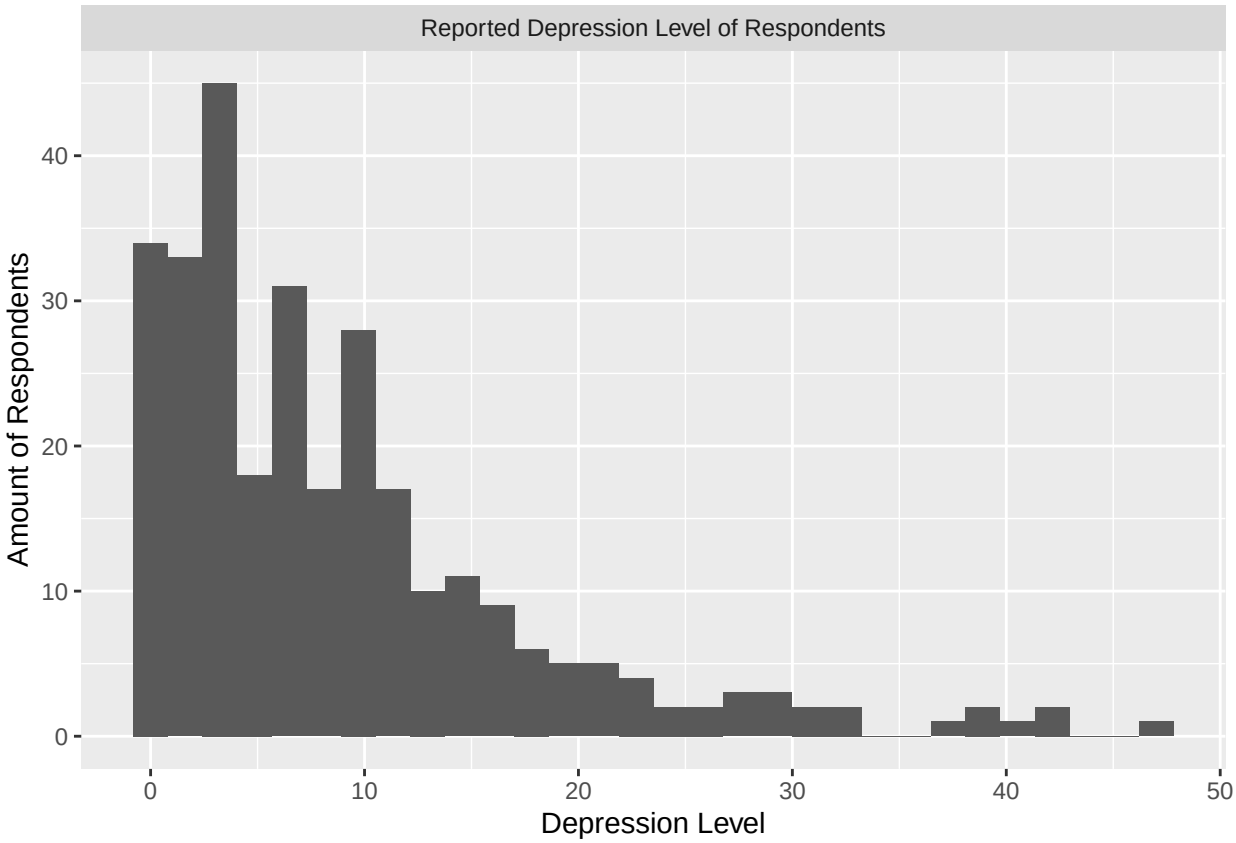
```
table(depress$CESD)
```

```
##
##  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 24 25 26
## 34 20 13 25 20 18 15 16 17 17 11  7 10 10  5  6  5  4  6  3  2  5  4  1  1  2
## 28 29 31 33 38 39 40 42 47
##  3  3  2  2  1  2  1  2  1
```

The lowest-level possible = 0 and the highest-level possible = 60

```
ggplot(depress, aes(x =CESD)) + geom_histogram() + ylab("Amount of Respondents") + xlab("Depression Leve
```

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

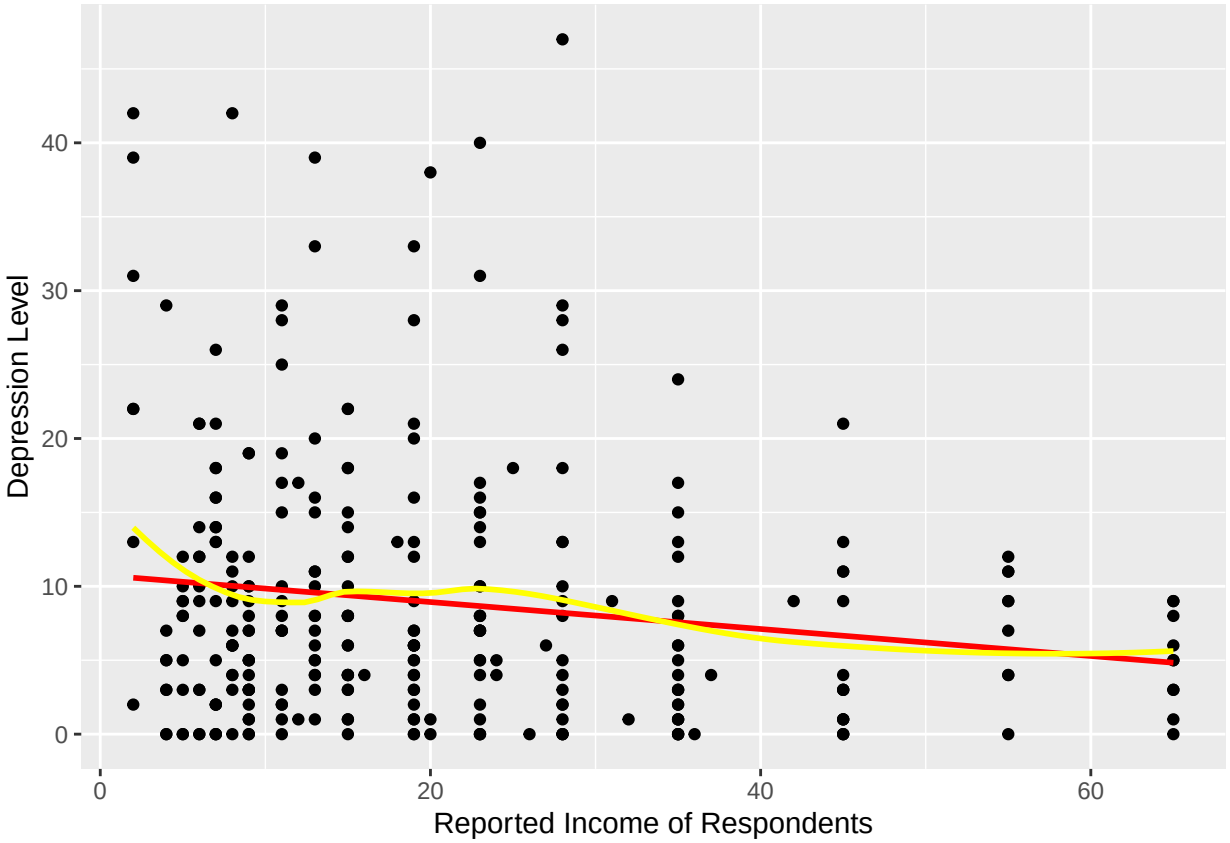


The first graph shows the reported income of respondents, the second graph illustrates whether or not respondents were regular drinkers, and the third graph is the frequency distribution of depression levels of respondents. The reported income and frequency of depression levels however, reported income had a smaller/less noticeable skew than depression levels. I will look further into the income variable in order to see if there is a significant relationship.

Bivariate Exploration:

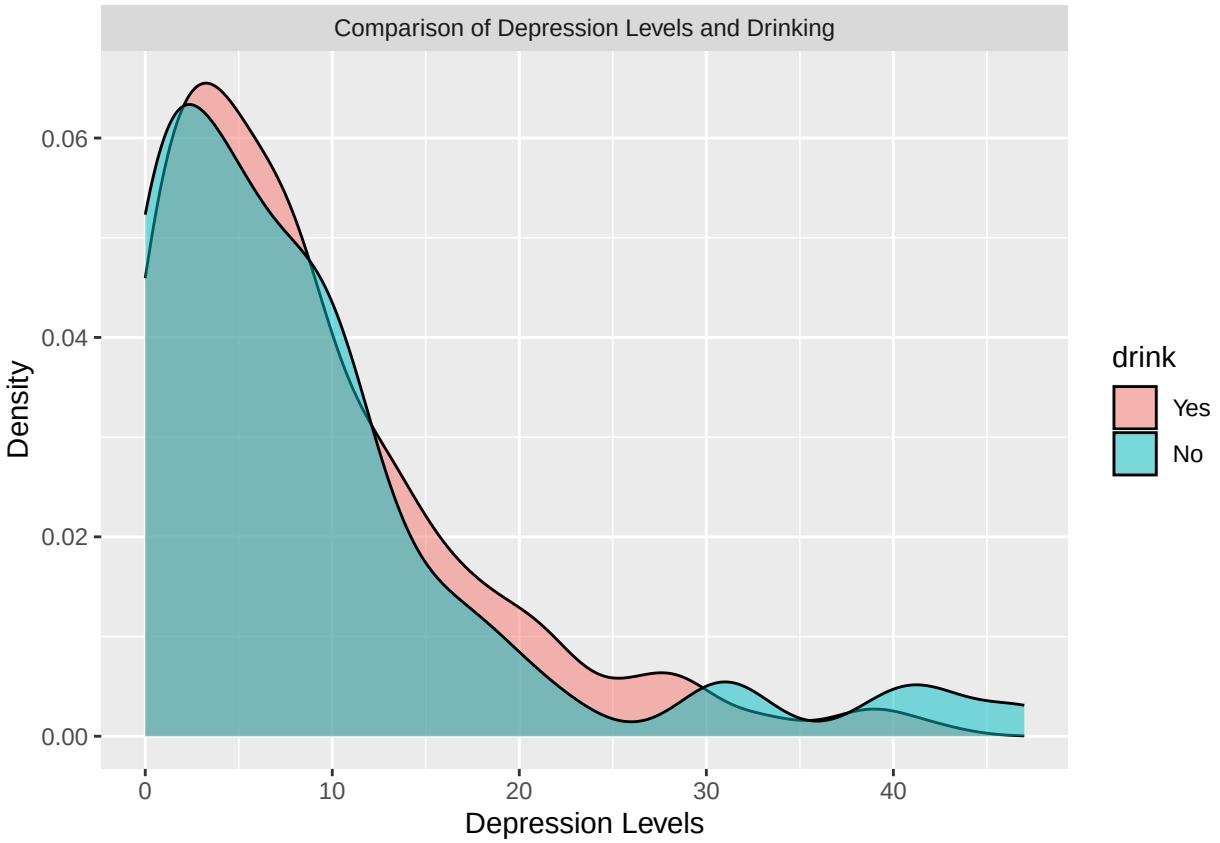
```
ggplot(depress, aes(x = INCOME, y = CESD)) + geom_point() + geom_smooth(se = FALSE, method = "lm", colo
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



This scatterplot demonstrates how there is not a significant relationship between reported income and depression levels in patients. Both lines are pretty much overlapping and with a very small negative relationship. Because there is no significant relationship, I will further investigate the drink variable as opposed to the income variable. The depression levels of respondents will continue to be studied.

```
ggplot(depress, aes(x = CESD, fill = DRINK)) + geom_density(alpha = 0.5) + scale_fill_discrete(name = "
```



This density graph shows a relationship between depression levels in those that said 'yes' to drinking regularly versus those that said 'no' to drinking regularly. In the graph, the 'yes' score is represented in red and 'no' scores are blue. The graph clearly shows that those who said yes to drinking regularly have higher depression scores than those that said no to drinking regularly. While the 'yes' category maybe only a little higher peak than the 'no' category, it still shows how respondents who drink regularly are more likely to have depression.

Conclusion: Initially, I thought depression levels would be interesting to compared to the variable of income. When I found out there wasn't really a correlation between the 2, I replaced the variable of income with drink. It was discovered from this data set that depression levels are actually different between those that said 'yes' to drinking regularly and those that said 'no' to drinking regularly, with the people who indulge in drinking overall are more depressed than those that do not.