# Exploratory Project

Justin Vigil

2022-09-19

## Introduction

The depression data set explored depression among adults via a set of interviews. There are 294 observations with 37 variables.

I am going to be exploring the correlation between annual income and depression. The two variables that I will be looking at are INCOME and CASES.

**My research question is as follows**: *Is there a higher propensity of depression in people that make less than $30,000 dollars annually versus people that make more than $30,000 annually?*

## Univariate Exploration

The first thing that I need to do is clean my INCOME variable so that it is describes the observations as either > 30,000 or < 30,000.

```
depression$income_clean <- ifelse(depression$income < 30, "<30,000", ">30,000")
table(depression$income_clean, useNA = "always")
```

```
##
## <30,000 >30,000    <NA>
##     231      63       0
```

I need to refactor my cases variable so that it reads "Normal" and "Depressed", not 1 and 0.

```
table(depression$cases)
```

```
##
##   0   1
## 244  50
```

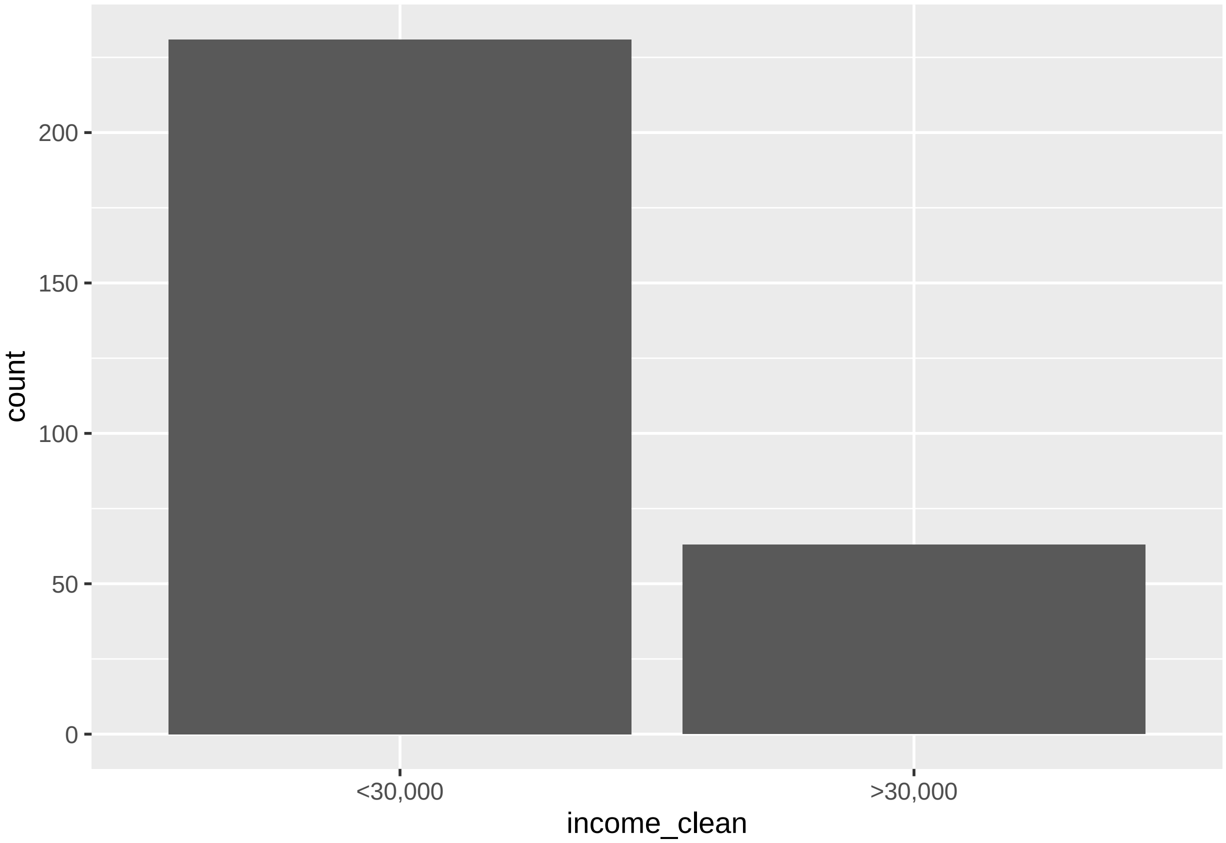
```
depression$cases_new <- factor(depression$cases, labels=c("Normal", "Depressed"))
table(depression$cases_new, depression$cases)
```

```
##
##               0   1
##   Normal    244   0
##   Depressed   0  50
```

Now that my income and the cases variables are cleaned, I am going to produce graphs of the income_clean and cases variables independently.

### Income

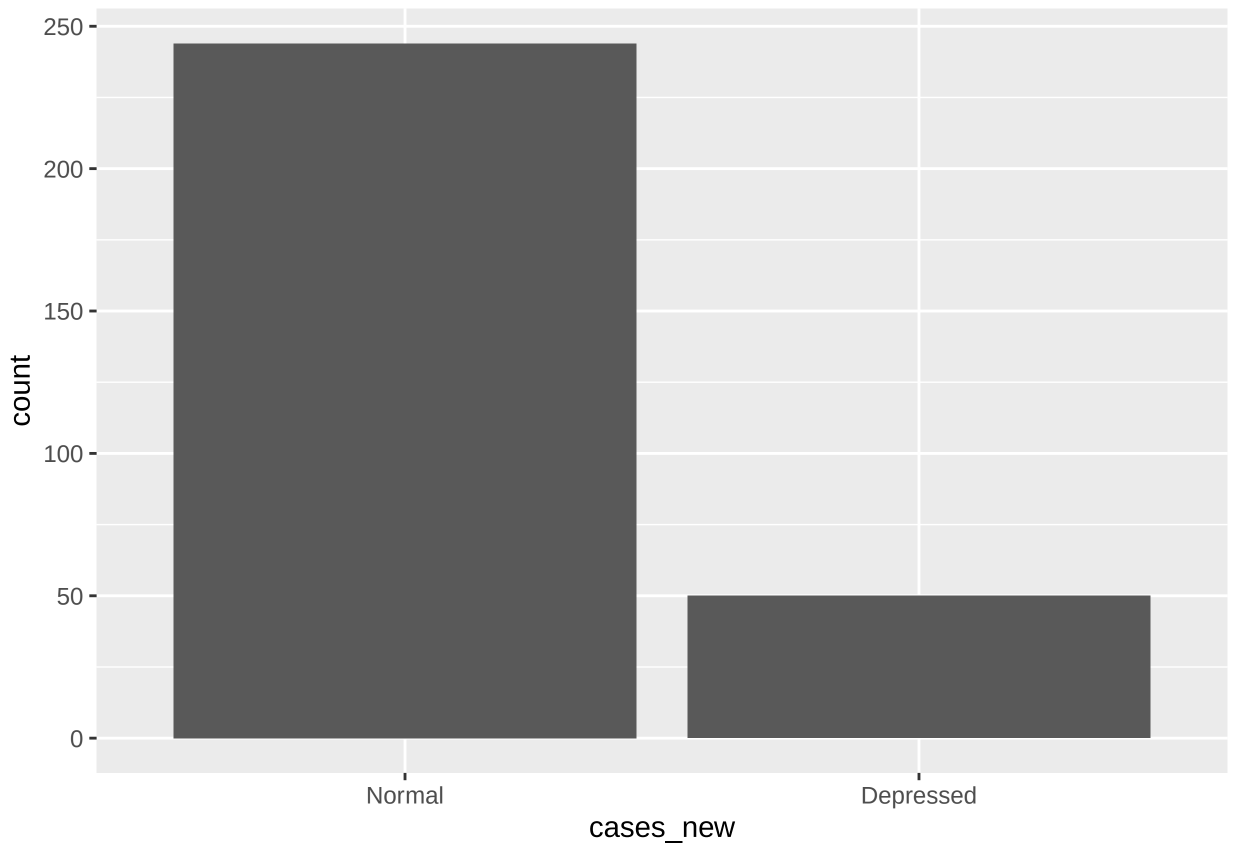
```
ggplot(depression, aes(x=income_clean)) + geom_bar()
```



This bar chart shows me that the majority of the observations have an income that is less than $30,000 annually.

### Depression(Cases variable)

```
ggplot(depression, aes(x=cases_new)) + geom_bar()
```



Again, the majority of the observations are considered "Normal" according to the survey.

## Bivariate Exploration

I want to look at both my variables in a table.

```
table(depression$cases_new, depression$income_clean)
```

```
##
##             <30,000 >30,000
##   Normal        184      60
##   Depressed      47       3
```

```
table(depression$cases_new, depression$income_clean) %>% prop.table()
```

```
##
##                <30,000    >30,000
##   Normal    0.62585034 0.20408163
##   Depressed 0.15986395 0.01020408
```

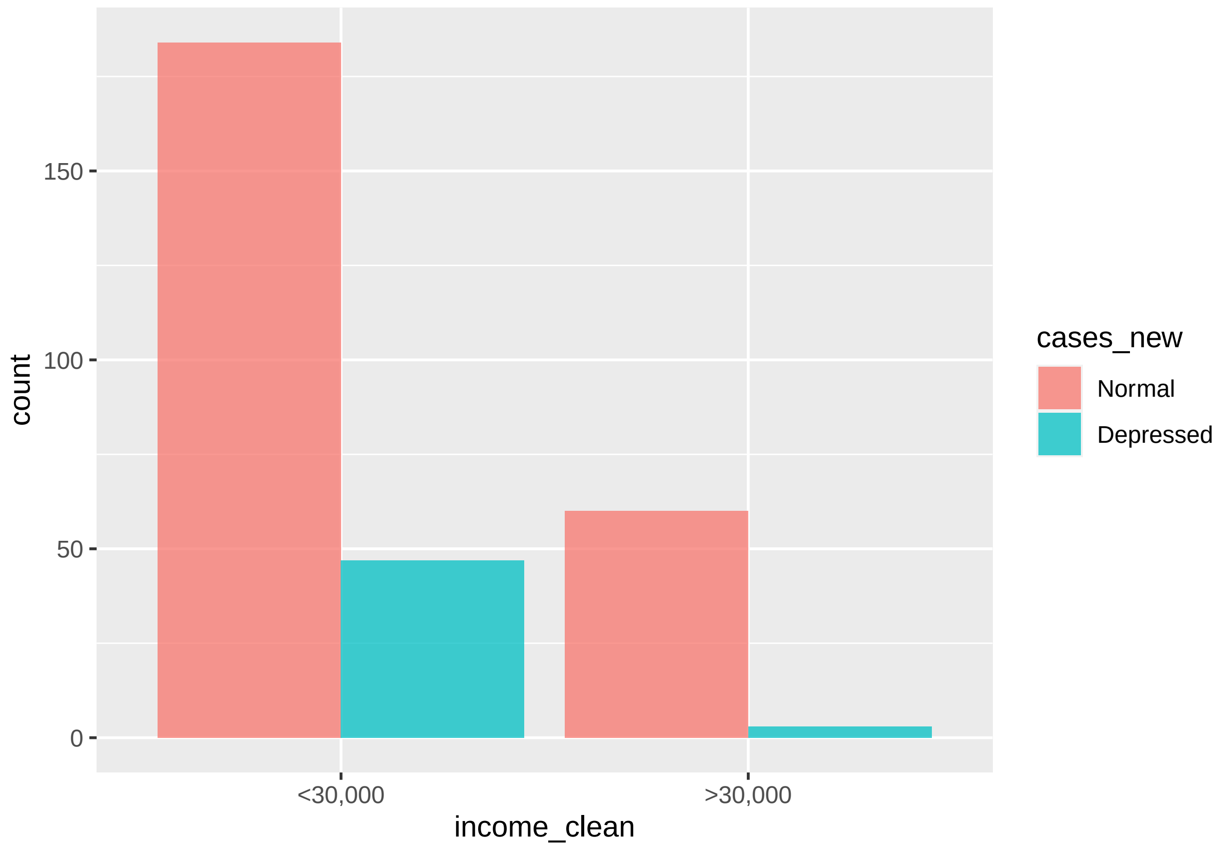
According to these statistics, about 63% of observations that earn less than $30,000 annually are considered "Normal". About 16% of observations that earn less than $30,000 annually are considered "Depressed". About 20% of observations that earn more than $30,000 annually are considered "Normal". About 1% of observations that earn more than $30,000 annually are considered "Depressed".

```
table(depression$cases_new, depression$income_clean) %>% prop.table(margin=2)
```

```
##
##                <30,000    >30,000
##   Normal    0.79653680 0.95238095
##   Depressed 0.20346320 0.04761905
```

According to this proportion table, about 20% of the people who earn less than $30,000 annually are depressed. For people who earn more than $30,000, the percentage of people who are depressed drops to about 5%.

```
ggplot(depression, aes(x=income_clean, fill=cases_new)) + geom_bar(alpha=0.75,  position="dodge")
```



This graphic further helps us see that there is a higher proportion of the people who earn less than $30,000 annually have depression.

## Conclusion

Since both my variables were categorical, I used bar charts for my univariate exploration. I also used a bar chart for my bivariate exploration, again because my variables were categorical. Through my bivariate exploration, I found that of people who earn less than $30,000 annually, there was a higher proportion of people with depression compared to the people who earn more than $30,000 annually. About 20% of the people who earn less than $30,000 annually have depression whereas only about 5% of the people who earn more than $30,000 annually have depression. This is something that I thought would be true, but after analyzing the data, it became clear that it was.