Final

Bryanneh McKenzie

2025-06-28

Introduction: For my final project I chose the Depression data set from the first wave of prospective participants in a depression study. This study focused on the adult residents of Los Angeles County and had a total of 294 observation. My question today is how does educaiton level and gender influence depressive symptoms?

Prep:

```
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(forcats)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v ggplot2   3.5.2     v stringr   1.5.1
## v lubridate 1.9.4     v tibble    3.2.1
## v purrr     1.0.4     v tidyr     1.3.1
## v readr     2.1.5

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Univariate Exploration:

```
depress <- read.delim("/Users/bryannehs/Desktop/math130/data/Depress.txt")

depress <- depress %>%
  mutate( sex = factor(SEX, levels = 1:2, labels = c("Male", "Female")),
    education = factor(EDUCAT, levels = 1:7, labels = c("Less than HS", "Some HS", "HS Grad", "Some Col
    income_annual = INCOME * 1000)
```

Here I will see the cooralation between depression and education level.

```
table(depress$EDUCAT)
```

```
##
##   1   2   3   4   5   6   7
##   5  61 114  48  43  14   9
```

```
edu_tab <- table(
  factor(
    depress$EDUCAT,
    levels = 1:7,
    labels = c("Less than HS","Some HS","HS Grad","Some College","Bachelor's", "Master's","Doctorate")))
table(depress$education)
```

```
##
## Less than HS      Some HS      HS Grad Some College   Bachelor's     Master's
##            5           61          114           48           43           14
##    Doctorate
##            9
```

```
prop.table(edu_tab) * 100
```

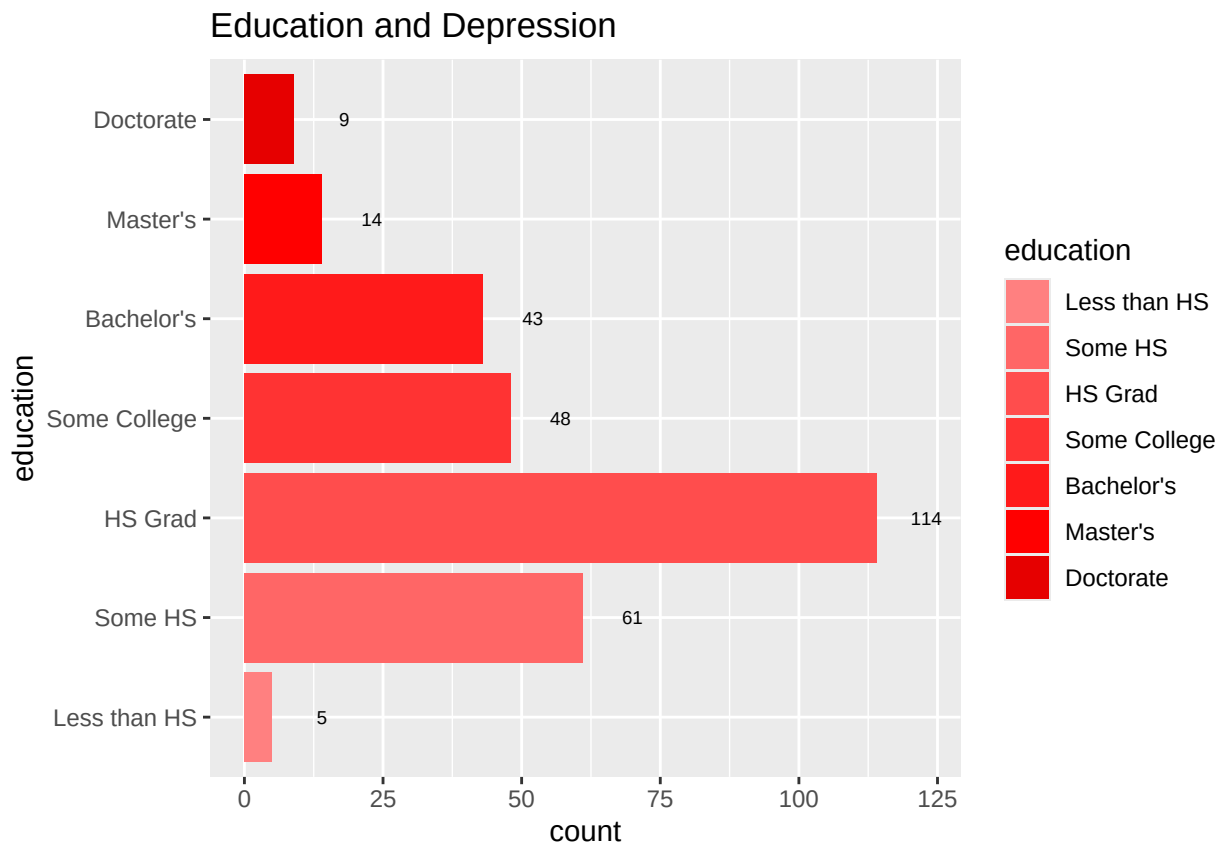
```
##
## Less than HS      Some HS      HS Grad Some College   Bachelor's     Master's
##     1.700680    20.748299    38.775510    16.326531    14.625850     4.761905
##    Doctorate
##     3.061224
```

```
summary(depress$SEX)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   1.622   2.000   2.000
```

```
library(ggplot2)
ggplot(depress, aes(x =education, fill = education)) +
  geom_bar() +
  scale_fill_manual(values=c("#ff8080","#ff6666","#ff4d4d" ,"#ff3333","#ff1a1a","#ff0000","#e60000")) +
  geom_bar(aes(y=..count..)) + ggtitle("Education and Depression") +
  geom_text(aes(y=..count.. +9, label=..count..), stat = 'count', size =2.5) +
  coord_flip()
```

```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Gender: Here I will test the coorelation between gender and depression.

```
sex_tab <- table(
  factor(
    depress$SEX,
    levels = 1:2,
    labels = c("Male", "Female")))
table(depress$SEX)
```

```
##
##   1   2
## 111 183
```

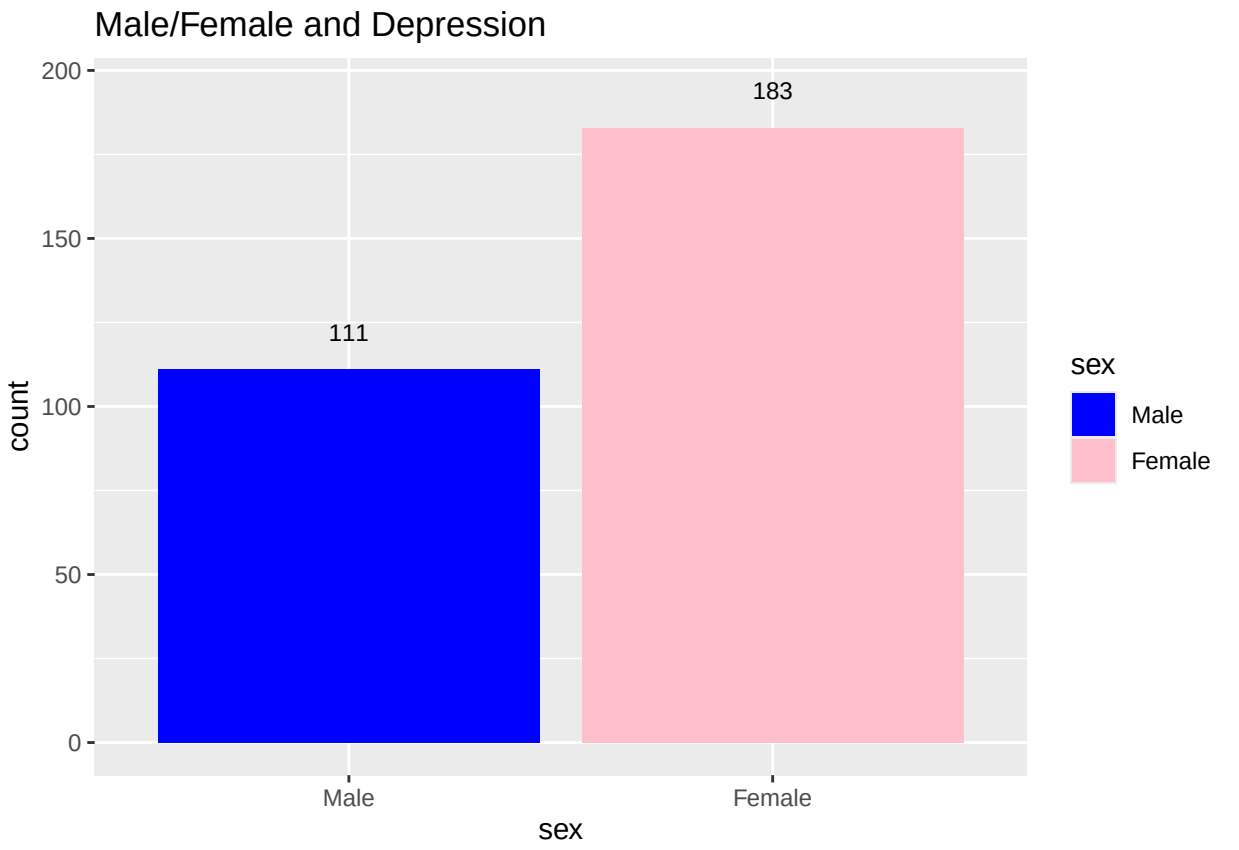
```
prop.table(sex_tab) * 100
```

```
##
##    Male  Female
## 37.7551 62.2449
```

```
summary(depress$SEX)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   1.622   2.000   2.000
```

```
depress$sex <- factor(depress$SEX,levels = c(1, 2), labels =c("Male", "Female"))

ggplot(depress, aes(x = sex,  fill = sex)) +
   geom_bar (aes(y=..count..)) +
  scale_fill_manual(values = c("blue", "pink")) +  ggtitle("Male/Female and Depression") +
  geom_text(aes(y=..count.. + 11, label=..count..),
            stat = 'count',
            size =3)
```



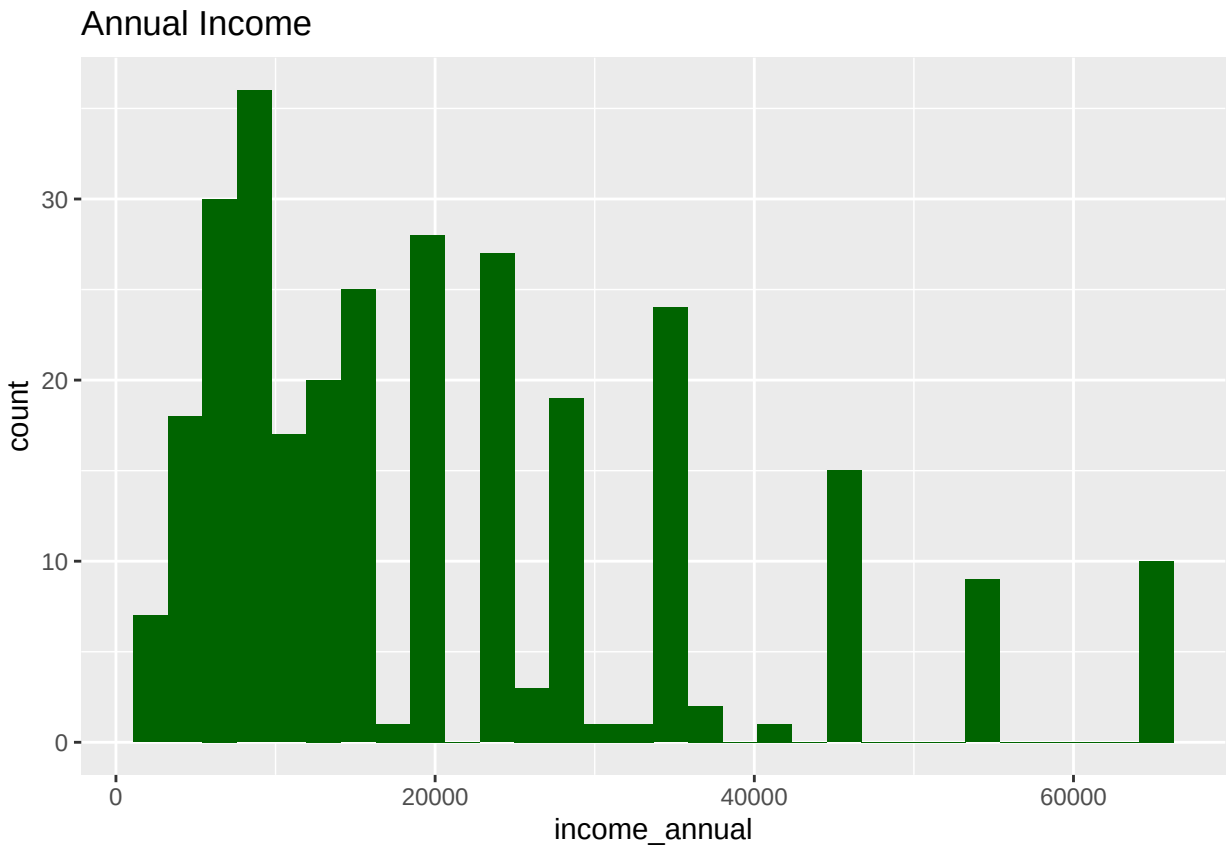
Income: Here I will test the coorelation between income and depression.

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

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

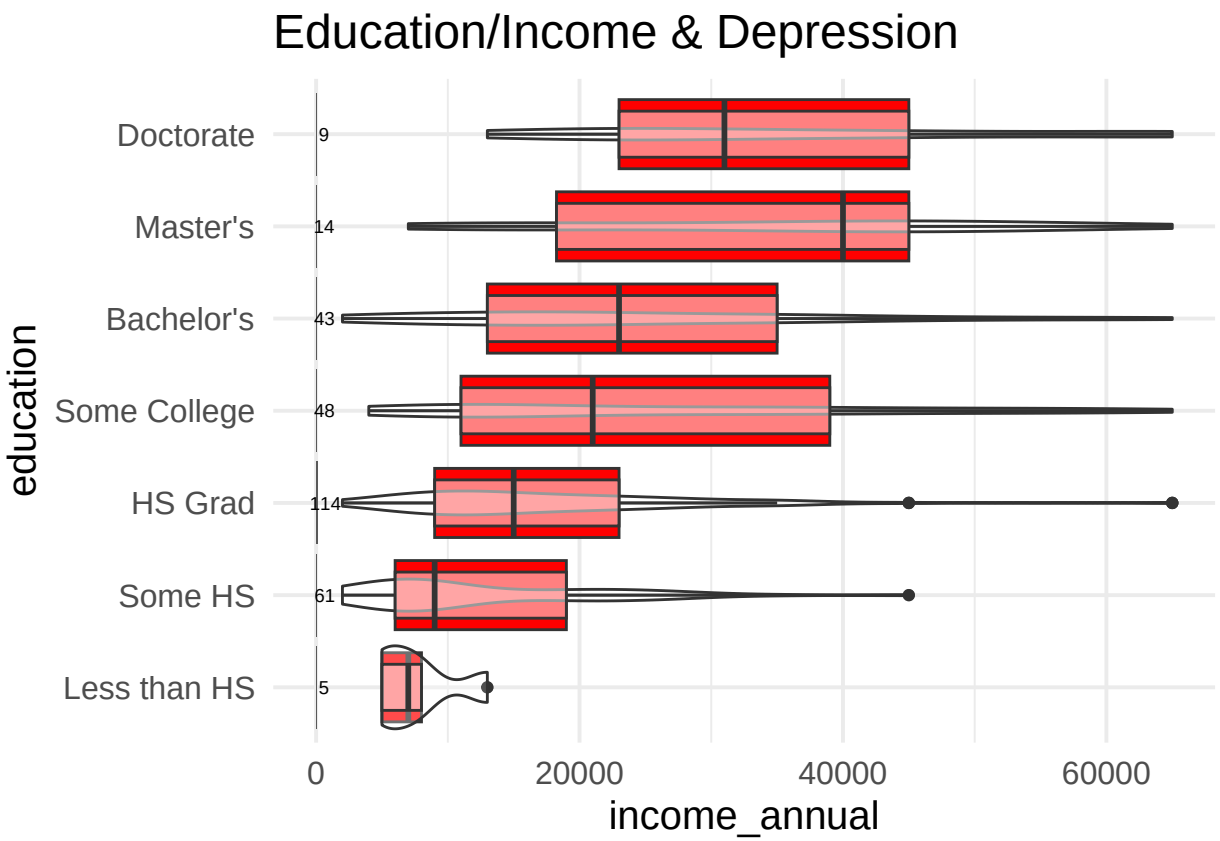
```
depress$income_annual <- depress$INCOME *1000

ggplot(depress, aes(x = income_annual)) + geom_histogram(fill="darkgreen") + ggtitle("Annual Income")
```

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



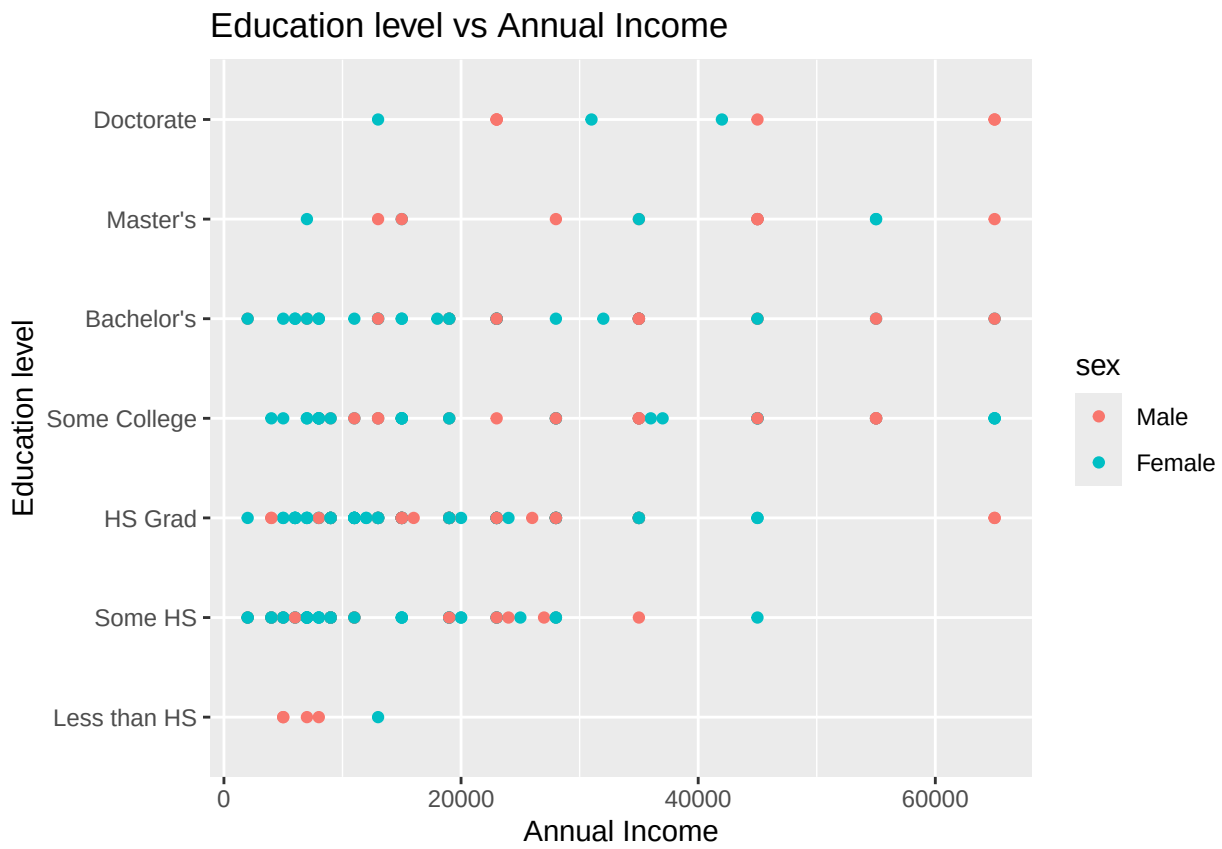
Now I am going to compare education and income levels

```
ggplot(depress, aes(x=education, y=income_annual))+
  geom_boxplot(fill= "red") +
geom_violin(alpha=.3) +geom_boxplot(alpha=.5, width=.5) +
geom_bar(aes(y=..count..)) + ggtitle("Education/Income & Depression") +
  geom_text(aes(y=..count.. +600, label=..count..), stat = 'count', size =2.5) +
  theme_minimal(base_size = 15) +
  coord_flip()
```



I like this but I want to do it again with a scatterplot and see the difference with gender

```
ggplot(depress,aes(x=education, y=income_annual, color=sex)) +
  geom_point() +
  ggtitle("Education level vs Annual Income") +
  xlab("Education level") +
  ylab("Annual Income") +
   coord_flip()
```



In conclusion from what I can observe without coming to any conclusions, there is a small a negative link between education and depression however, it appears similar to male and females. It appears that people with less than a high-school degree have the highest average scores. However when in regards to this same score, females on average have a higher average score while still following the same trend. Income also appears linked to depression scores.

‘‘‘

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.