Exploratory Data Analysis

Alisha Fang

2/21/2022

Intro

The data set I will be analzying is a study of depression in adult residents who are living in the Los Angeles County. I will be observing the variables education, income in thousands of dollars per year, and cases of depression. I am interested in observing the relationship of one's education with their income and how it all relates towards an individual having depression.

```
depression <- read.table("/Users/alisha/Desktop/math 130/data/depress.txt", header=TRUE,
sep="\t")
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)
library(forcats)
library(RColorBrewer)
```

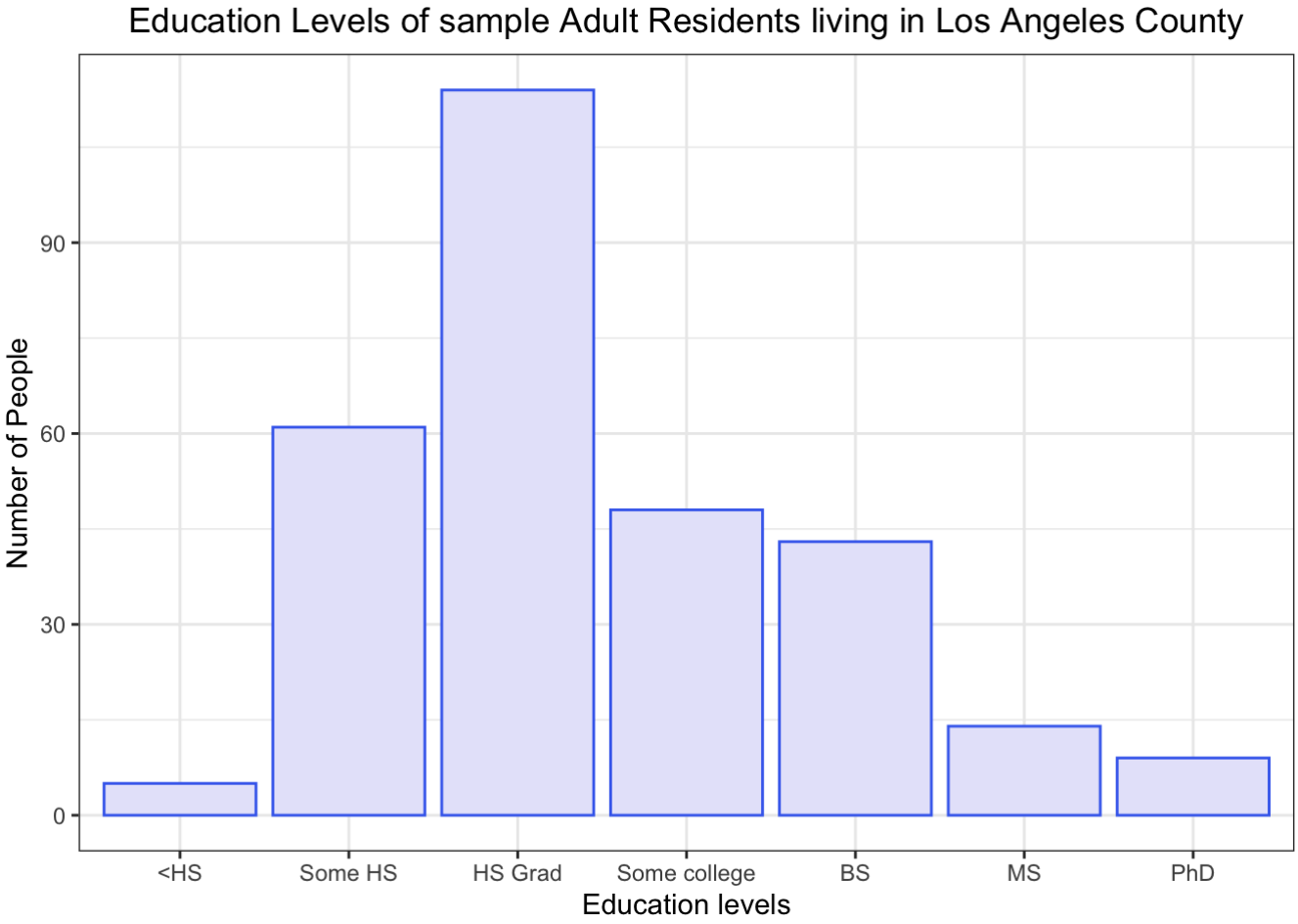
Univariate Analysis of Variables

Levels of Education

```
Depression<- depression
Depression$education<- depression$educat %>% fct_relevel("<HS","Some HS", "HS Grad", "So
me college", "BS", "MS", "PhD")
table(Depression$education)
```

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

```
ggplot(Depression, aes(x=education))+
  geom_bar(color="royalblue2", fill= "lavender")+
  xlab("Education levels")+
  ylab("Number of People")+
  ggtitle("Education Levels of sample Adult Residents living in Los Angeles County")+
  theme_bw()+
  theme(plot.title = element_text(hjust=0.5))
```



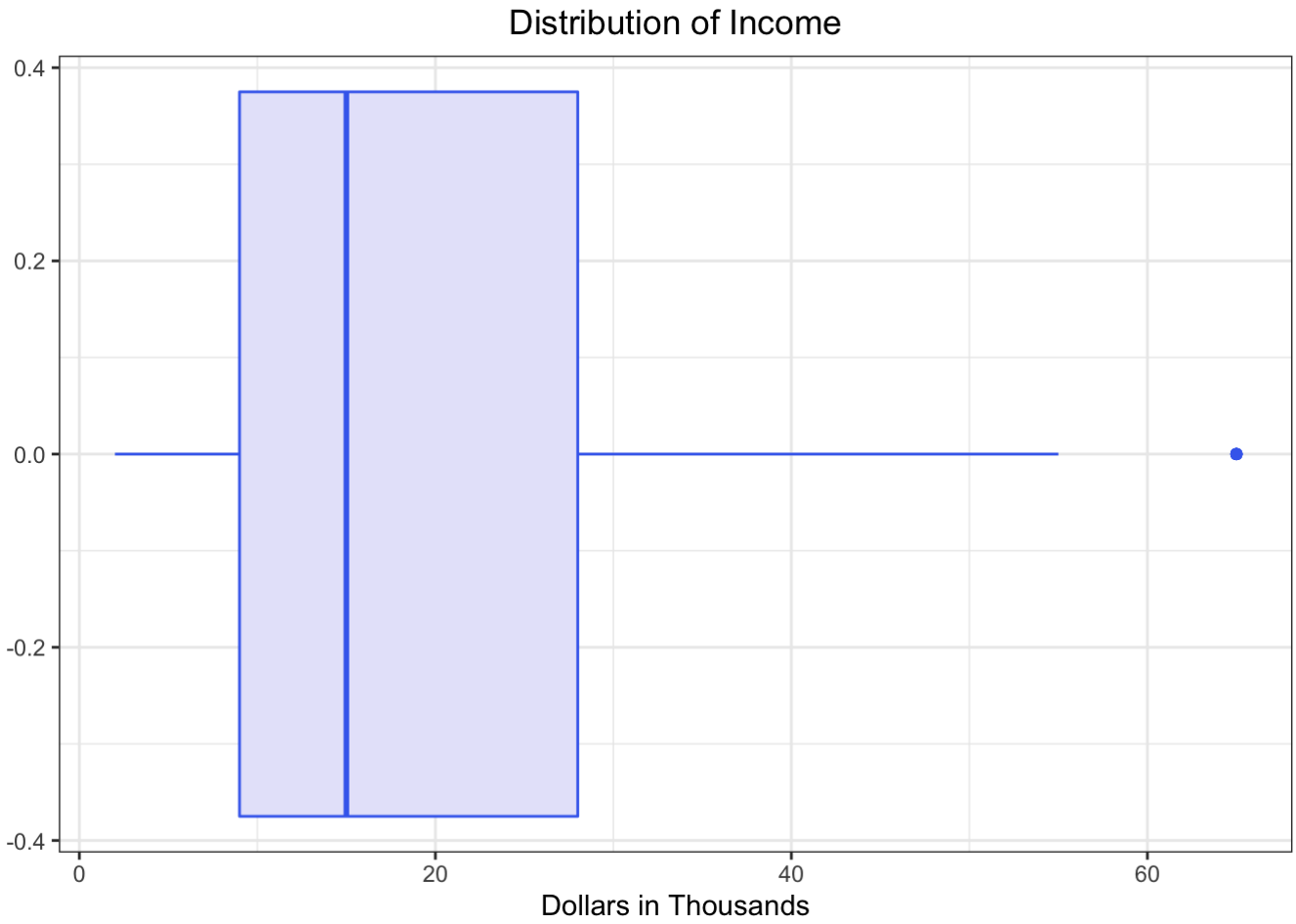
This is a barchart and table that represents the education levels of our sample. Most of our sample population contains individuals who will be from those who have some HS experience and HS Grad.

Income Level

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

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

```
ggplot(Depression, aes(x=income))+
  geom_boxplot( color= "royalblue2", fill="lavender")+
  ggtitle("Distribution of Income")+
  xlab("Dollars in Thousands")+
  theme_bw()+
  theme(plot.title = element_text(hjust=0.5))
```



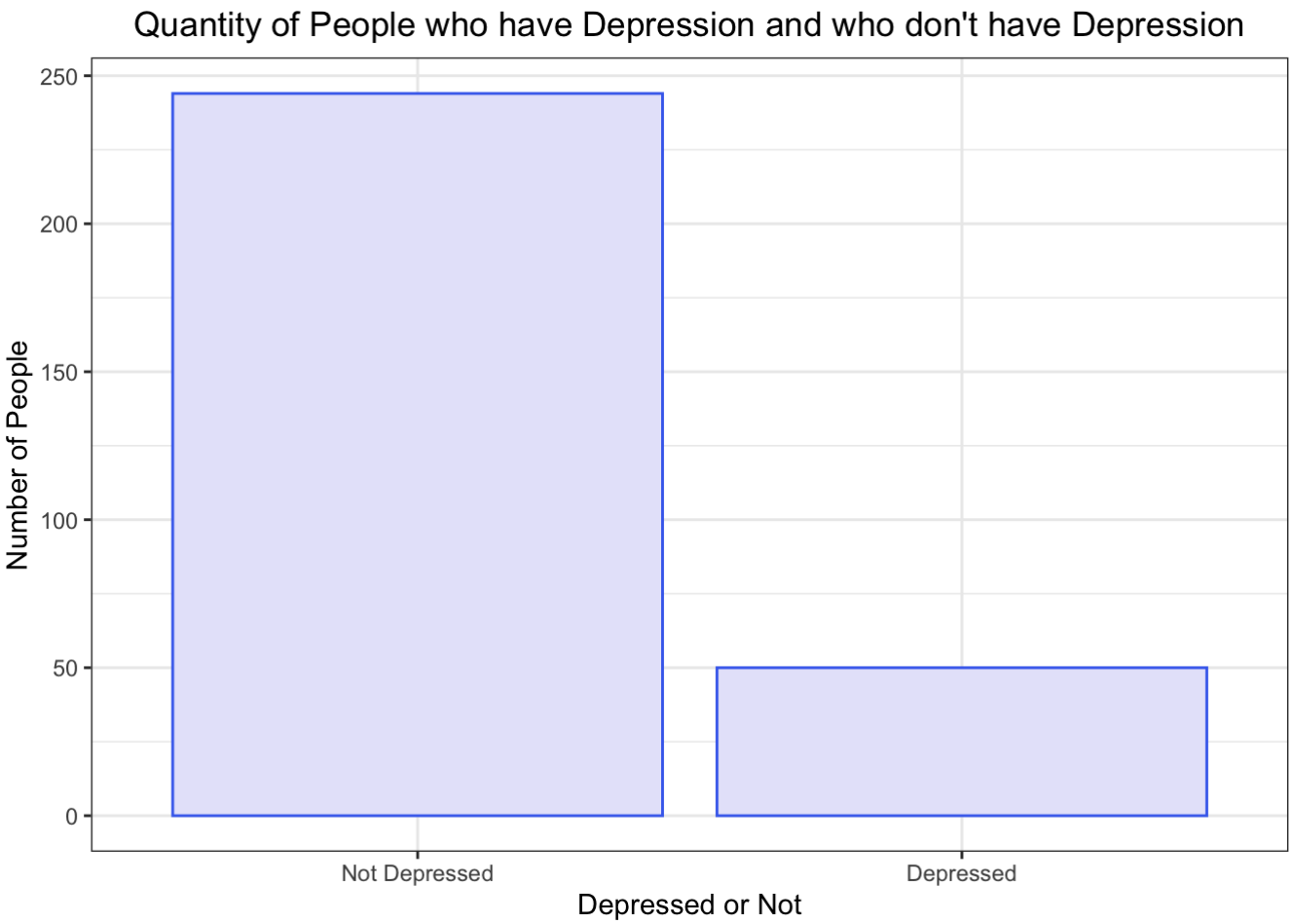
Here is a boxplot based on a table of our income of our sample. We can see that the amount made by 50% (Q1-Q3)of our sample lies between $9,000 and $28,000.

To be depressed or not to be depressed?

```
Depression$depress <- factor(depression$cases, labels=c("Not Depressed", "Depressed"))
table(Depression$depress)
```

```
##
## Not Depressed     Depressed
##           244            50
```

```
ggplot(Depression, aes(x=depress))+
  geom_bar(color= "royalblue2", fill= "lavender")+
  ggtitle("Quantity of People who have Depression and who don't have Depression")+
  xlab("Depressed or Not")+
  ylab("Number of People")+
  theme_bw()+
  theme(plot.title = element_text(hjust=0.5))
```

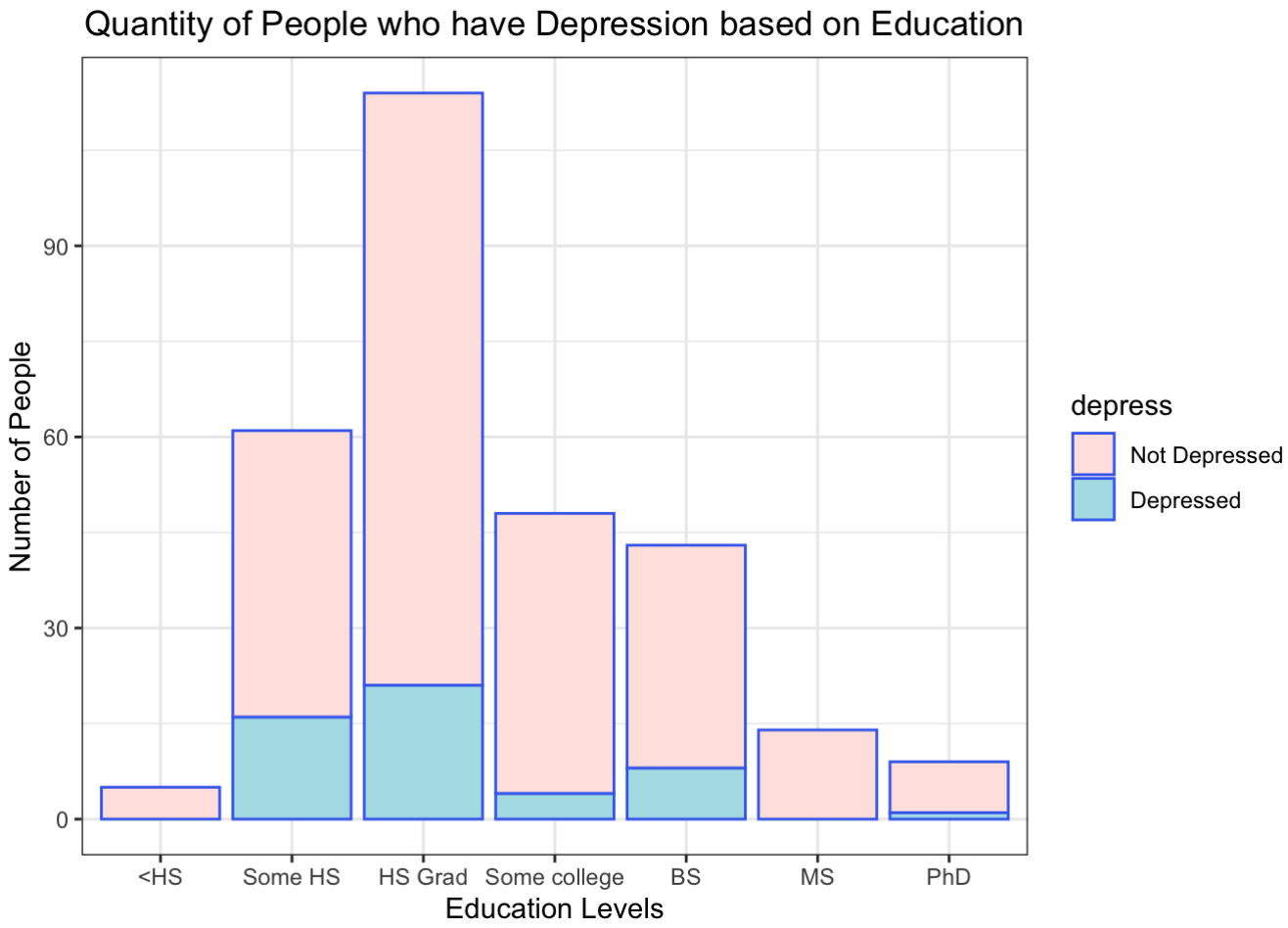


Here is a barchart to help us visualize the difference between how many people are depressed from our sample and who are not. From the data, we can see that a greater portion of our sample are individuals who are not depressed.

Bivariate Exploration

Comparing Education and Depression

```
ggplot(Depression, aes(x=education, fill=depress)) +
  geom_bar(color="royalblue2")+
  scale_fill_manual(values=c("mistyrose", "powderblue"))+
  ggtitle("Quantity of People who have Depression based on Education")+
  xlab("Education Levels")+
  ylab("Number of People")+
  theme_bw()+
  theme(plot.title = element_text(hjust=0.5))
```



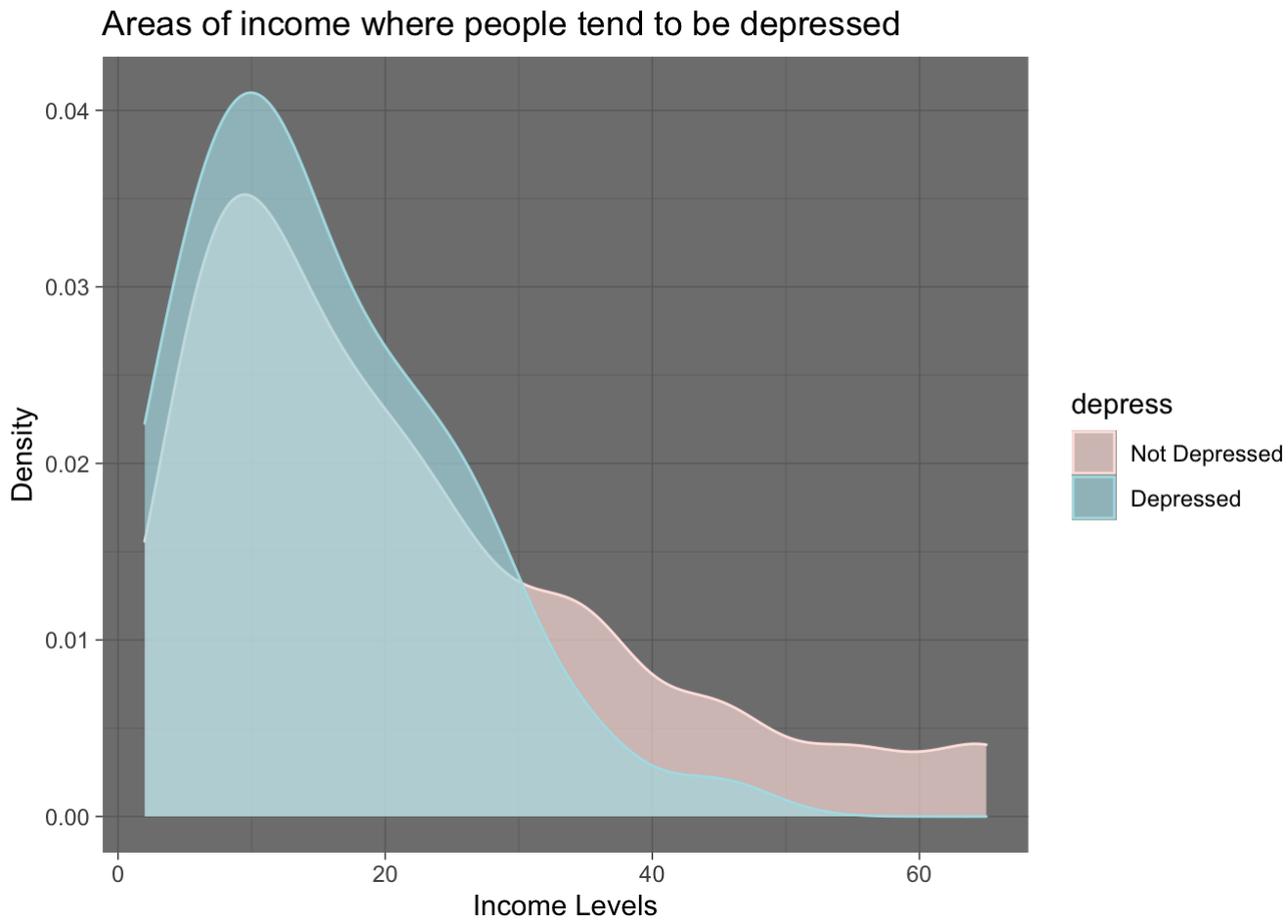

```
table(Depression$education, Depression$depress)
```

```
##
##                 Not Depressed Depressed
##   <HS                       5         0
##   Some HS                  45        16
##   HS Grad                  93        21
##   Some college             44         4
##   BS                       35         8
##   MS                       14         0
##   PhD                       8         1
```

We can observe from this data that a great portion of those who are depressed are those who have some HS and HS Graduate experience. There is an outlier where an individual with a PhD has depression.

Comparing Income levels with Depression

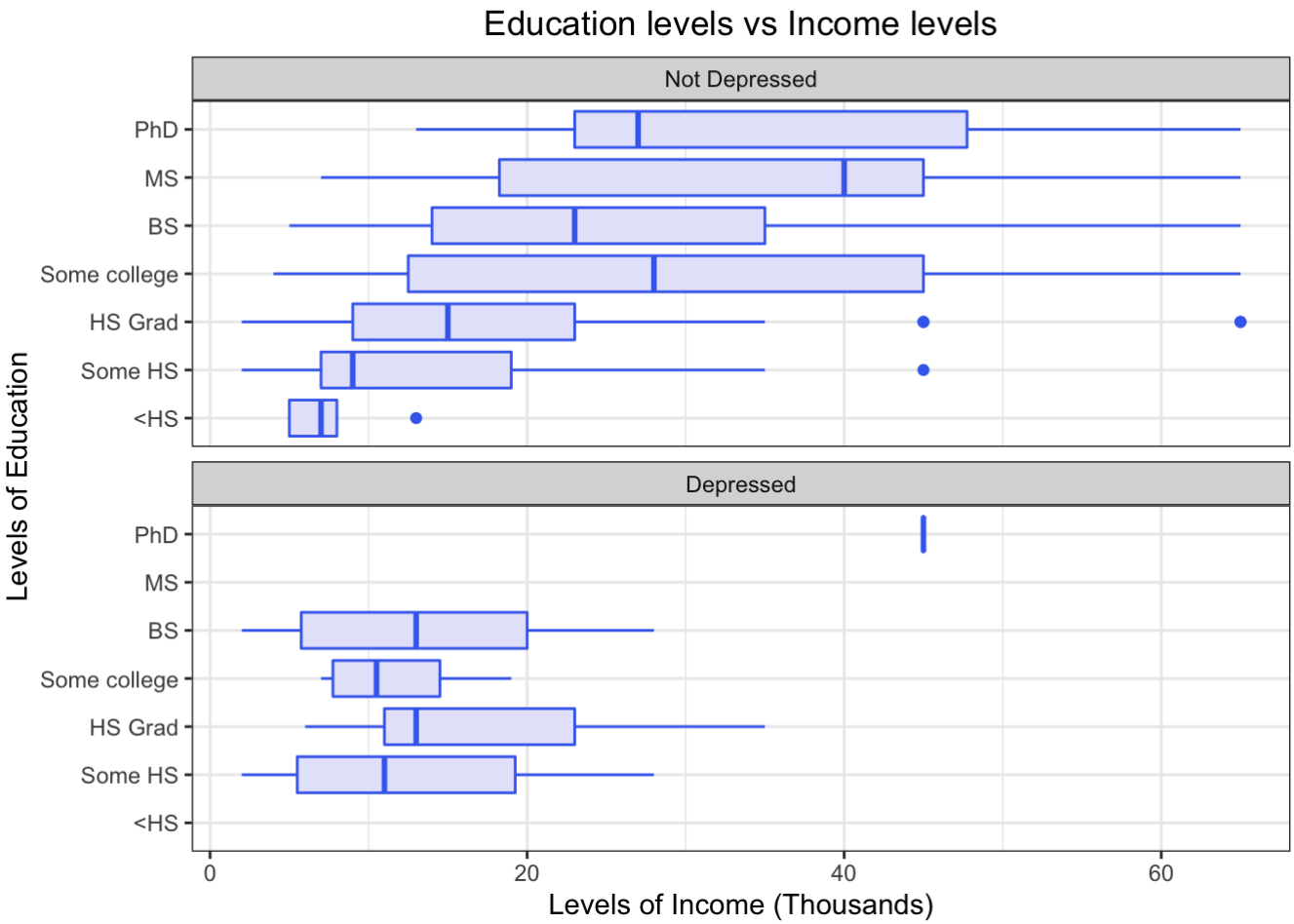
```
ggplot(Depression, aes(x=income, color=depress, fill= depress)) +
  geom_density(alpha=0.65)+
  scale_color_manual(values=c("mistyrose", "powderblue"))+
  scale_fill_manual(values=c("mistyrose", "powderblue"))+
  ggtitle("Areas of income where people tend to be depressed")+
  xlab("Income Levels")+
  ylab("Density")+
  theme(plot.title = element_text(hjust=0.5))+
  theme_dark()
```



We can also see that of those with depression also tend to have lower incomes. At the same time, those who aren't depressed also tend to have lower incomes as well.

Comparing Income and Education levels in those who are and aren't Depressed

```
ggplot(Depression, aes(x=education, y= income))+
  geom_boxplot(color="royalblue2", fill= "lavender")+
  coord_flip()+
  facet_wrap(~depress, ncol=1)+
  scale_color_discrete(guide="none")+
  ggtitle("Education levels vs Income levels")+
  ylab("Levels of Income (Thousands)")+
  xlab("Levels of Education")+
  theme_bw()+
  theme(plot.title = element_text(hjust=0.5))
```



From this dataset we can see that the distribution of income for those based on their level of education and whether or not they are depressed. Most of those who are depressed do have a lower income and a lower education level.

Conclusion

While it was harder to see in the density plot the relationship for income and depression. While we cannot say that having a lower income and lower education will result in depression, those who are depressed will typically meet those qualities. There are still plenty of individuals who meet those qualities but aren't depressed. Therefore, levels of education and levels of income may not be the best way to identify if an indivdual is depressed or not. However, they can still contribute to that possible. Though from the diagram, we can also surprisingly see that levels of education do not necessarily dictate what an individual's income may be.