Data Analysis Project

Eric Ruiz-Cardenas

2/26/2021

1. Introduction

The data that I will be analyzing today is the depression data set which contains 294 observations and 37 variables. This data is derived from a prospective study that collected information from its first set of interviewed adult residents from Los Angeles County regarding depression. More information about this study can be found in Practical Multivariate Analysis, 5th edition by Afifi, May and Clark. I will explore the variables of income status, employment status, and depression level (CESD).

```
depress <- read.delim("C:/Users/ericc/OneDrive/Desktop/math130/data/depress_081217.txt", header=TRUE,se

library(ggplot2)
library(knitr)
library(forcats)
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
```

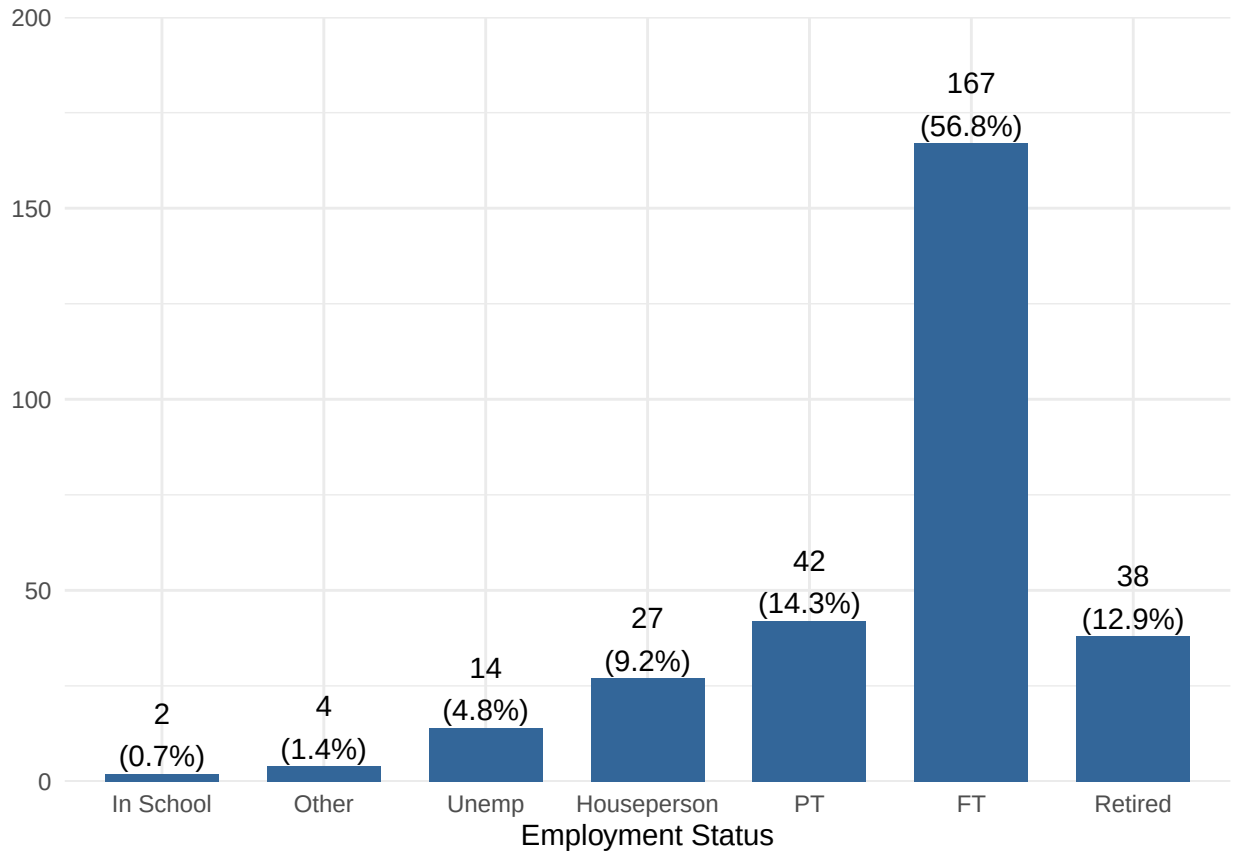
```
library(sjPlot)
```

2. Unvivariate Descriptions

```
depress$employ %>% fct_relevel("In School", "Other","Unemp", "Houseperson",
                               "PT", "FT", "Retired") %>% table()
```

```
## .
##   In School       Other       Unemp Houseperson          PT          FT
##           2           4          14          27          42         167
##     Retired
##          38
```

```
depress$employ <- factor(depress$employ, levels=c("In School", "Other", "Unemp",
        "Houseperson", "PT", "FT", "Retired"))
plot_frq(depress$employ) + xlab("Employment Status") + theme_minimal()
```



The bar graph shows the employment status representation among all 294 observations studied. The employment status ranged all the way from 2 responses of being still "In School" and 167 responses of being employed full time "FT". 56.8% of the adults studied replied as being employed as full time.

```
summary(depress$income)
```

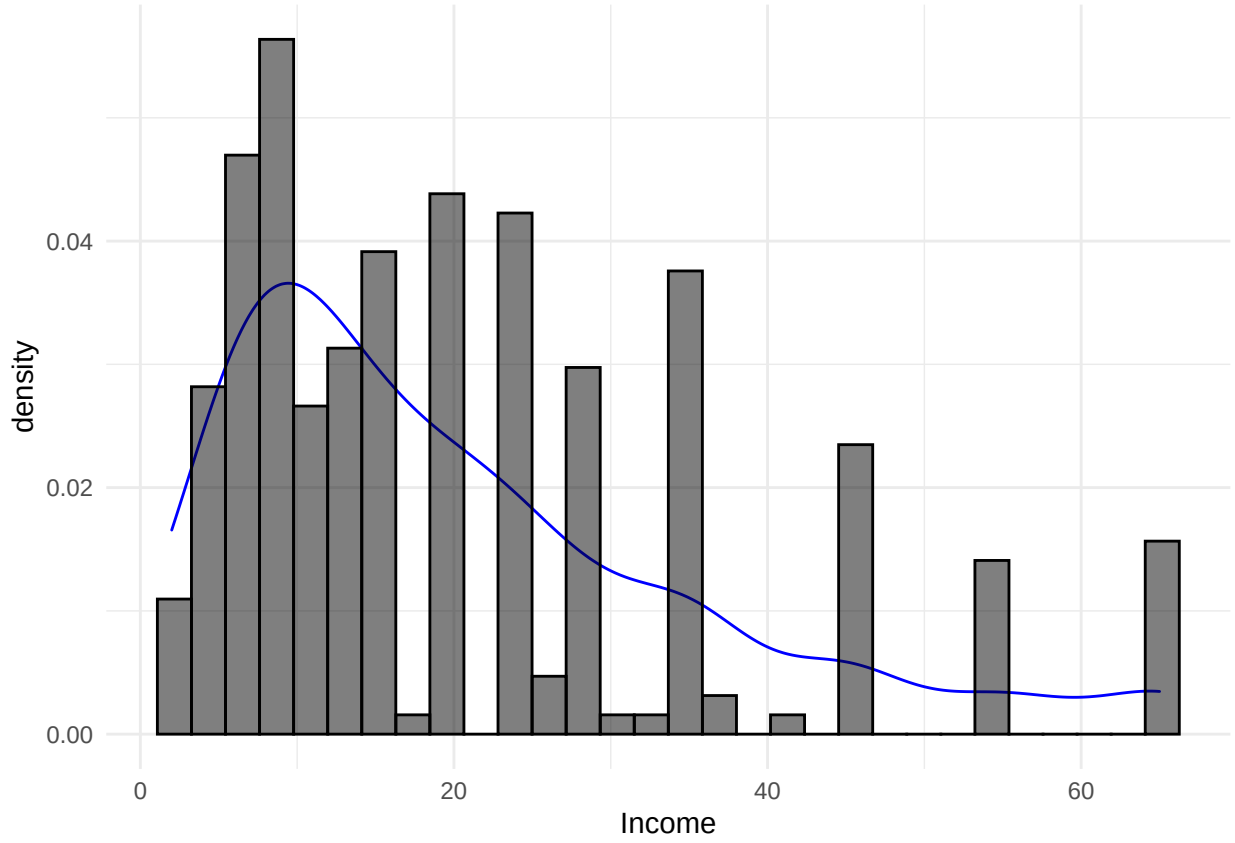
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    9.00   15.00   20.57   28.00   65.00
```

```
mean(depress$income)
```

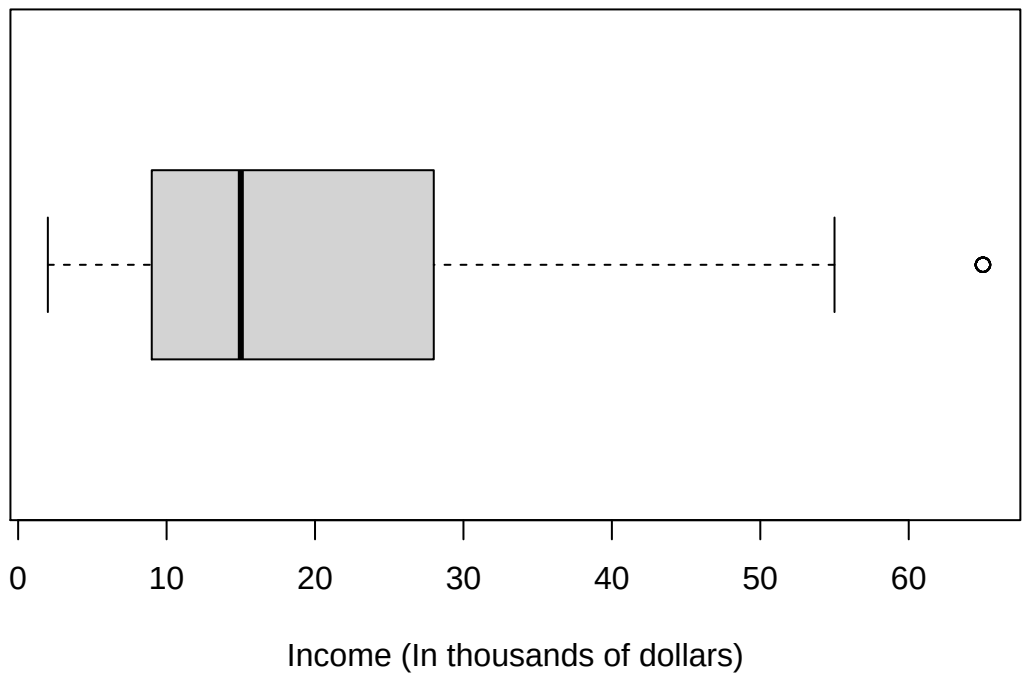
```
## [1] 20.57483
```

```
ggplot(depress, aes(x=income)) + geom_density(col="blue")+ geom_histogram(aes(y=..density..),
        colour="black", fill="black", alpha=0.5) + xlab("Income") + theme_minimal()
```

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




```
boxplot(depress$income, horizontal= TRUE, main= "Distribution of Income Status",
        xlab="Income (In thousands of dollars)")
```

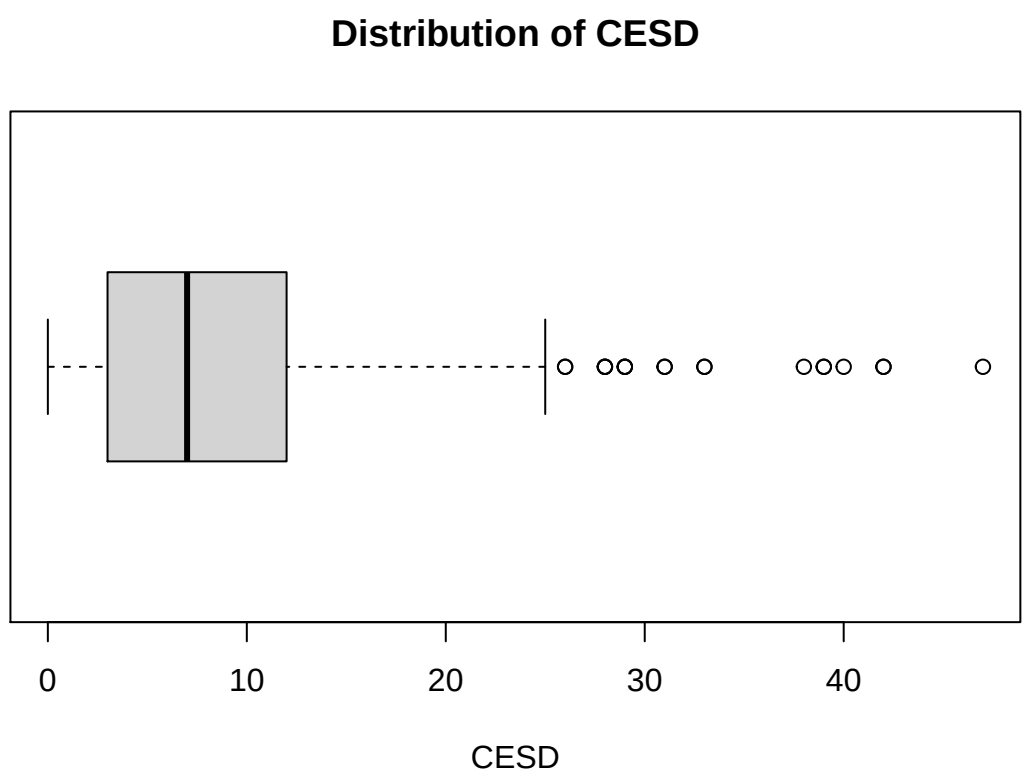


From the density plot and box plot, we can see the data for the income variable is unimodal and skewed right. The range stretched from $2,000 to $65,000.The mean average income was 20.57 ($20,570) and the median lied shortly below that at 15 ($15,000). The boxplot also shows an outlier at 65 ($65,000). The income of the majority of the adults interviewed lied heavily below $20,000 thus explaining why the graph is skewed right.

```
summary(depress$cesd)
```

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

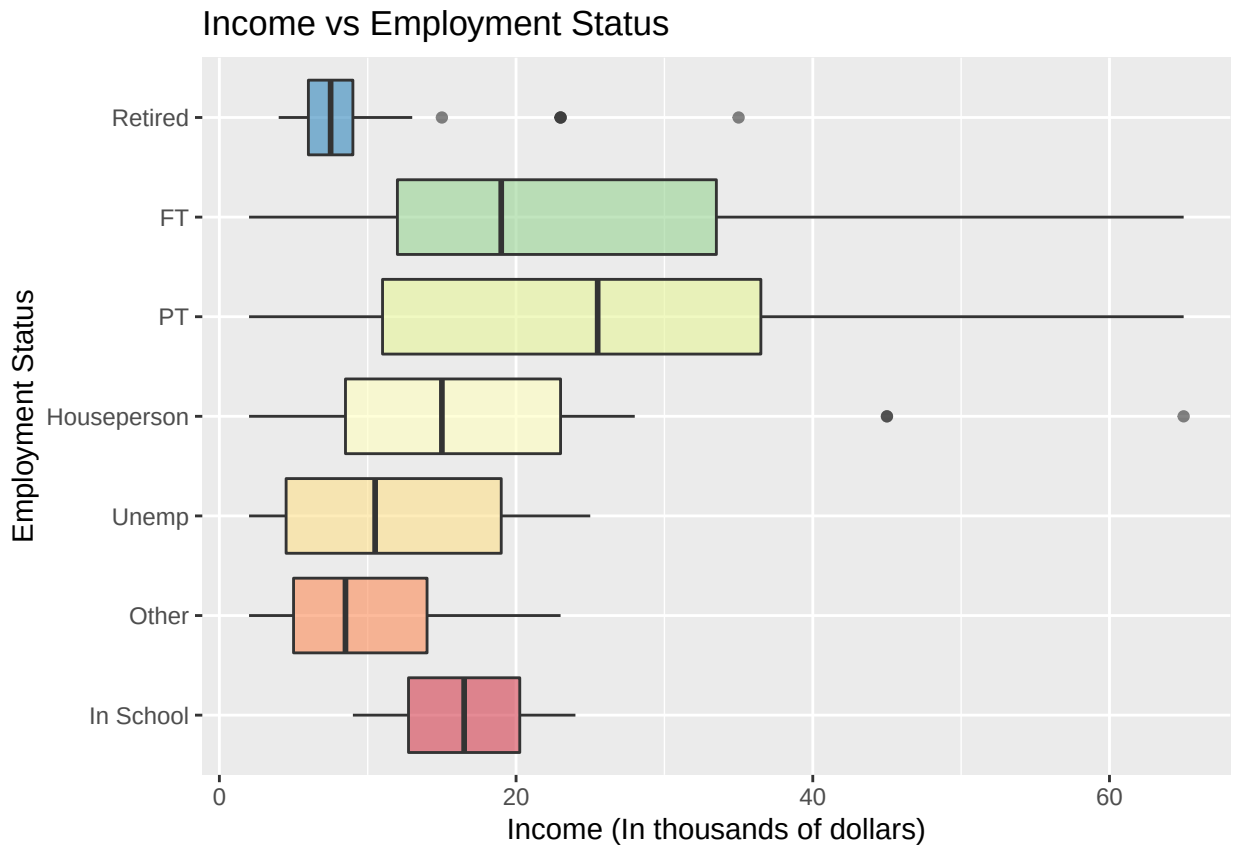
```
boxplot(depress$cesd, horizontal= TRUE, main= "Distribution of CESD", xlab="CESD")
```



CESD is defined as the sum of the C1-C20 (20 total cards) that depicted a level of 1-3 determining how the individual felt or behaved a certain number of times a week. The mean CESD or depression score that the individuals felt rested at about 8.88. The majority of CESD scores seem to fall from 0-10. The overall range of all individuals interviewed extended from a score of 0 to a max of 47. After around 25 we see an extensive amount of outliers which also contributes to the skewed right boxplot. (From the codebook for depression, depressed is where CESD >=16 and Normal<16)

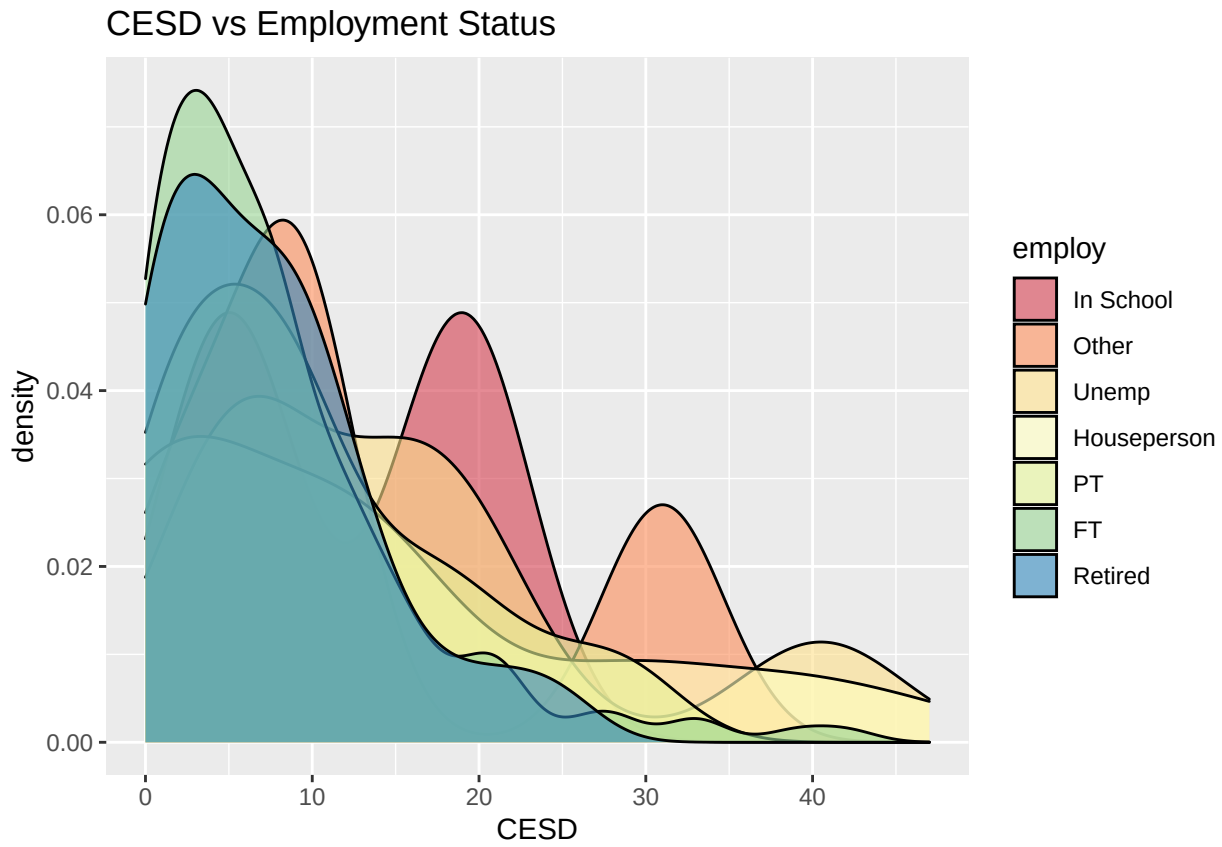
Bivariate Comparison

```
ggplot(depress, aes(x=employ, y=income, fill=employ,)) + geom_boxplot(alpha=0.6) +
        theme(legend.position="none") +
scale_fill_brewer(palette="Spectral") +ylab ("Income (In thousands of dollars)") +
        xlab ("Employment Status") +ggtitle("Income vs Employment Status") +coord_flip()
```



The grouped box plot shows the comparison between employment status and income status. Those that had an employment status of full-time or part-time, had the largest range in income status for their category. Comparing the full time and part however we can see that the median lies higher among part-time workers compared to full-time workers. Also, the boxplots are less skewed right for the part-time workers compared to the full-time workers. The lowest income concentration can be seen by those whose employment status is considered retired.

```
ggplot(depress, aes(x=cesd, fill=employ)) +
         geom_density(alpha=0.6) +
         scale_fill_brewer(palette="Spectral") +
         xlab ("CESD") +
         ggtitle("CESD vs Employment Status")
```



The overlaid density plot shows that employment status has an influence on the CESD depression scores. All employment statuses with the exception of this being "In school", and "Other", have a large concentration of a CESD score sitting at or below 10 which aligns with the mean cesd score of 8.88 that we had calculated earlier. The density plot depicts a bimodal distribution of income status among those who identified as "Other" the case could be due to the result of a low representation of those who identified as "Other" in this study.

Summary

In regards to the relationship between the income status and the employment status, we saw that all employment status average income lay heavily between $2,000 to $20,000. When we analyzed the variables of employment status against Cesd depression scores we saw that there was a small correlation. If we were to rate the employment status as a scale of working or not we can extract that those who identified as having a job or retired were more likely to score lower on the cesd score compared to those who worked minimally ("In School", "Other", "Unemp") or not at all. Due to confounding variables and the small sample size to fully represent a population in some variables, we can not statistically draw any conclusions about the variables we explored today.