Final Project: Exploratory Data Analysis of the Depression Data Set

Emiliana Lozano

2/25/2022

#Introduction

The Depression data set was gathered by interviewing adult residents in Los Angeles County as preliminary information for a future depression study. The data set is comprised of 294 observations and 37 variables. This exploratory analysis utilizes the "chronill" and "income" variables to determine possible correlation between presence of chronic illness in the past year and annual income.

```
depression<-read.table("/Users/emilianalozano/Documents/CSUC/CSUC_Spring_2022/MATH130/Data/depress_0812

head(depression)
```

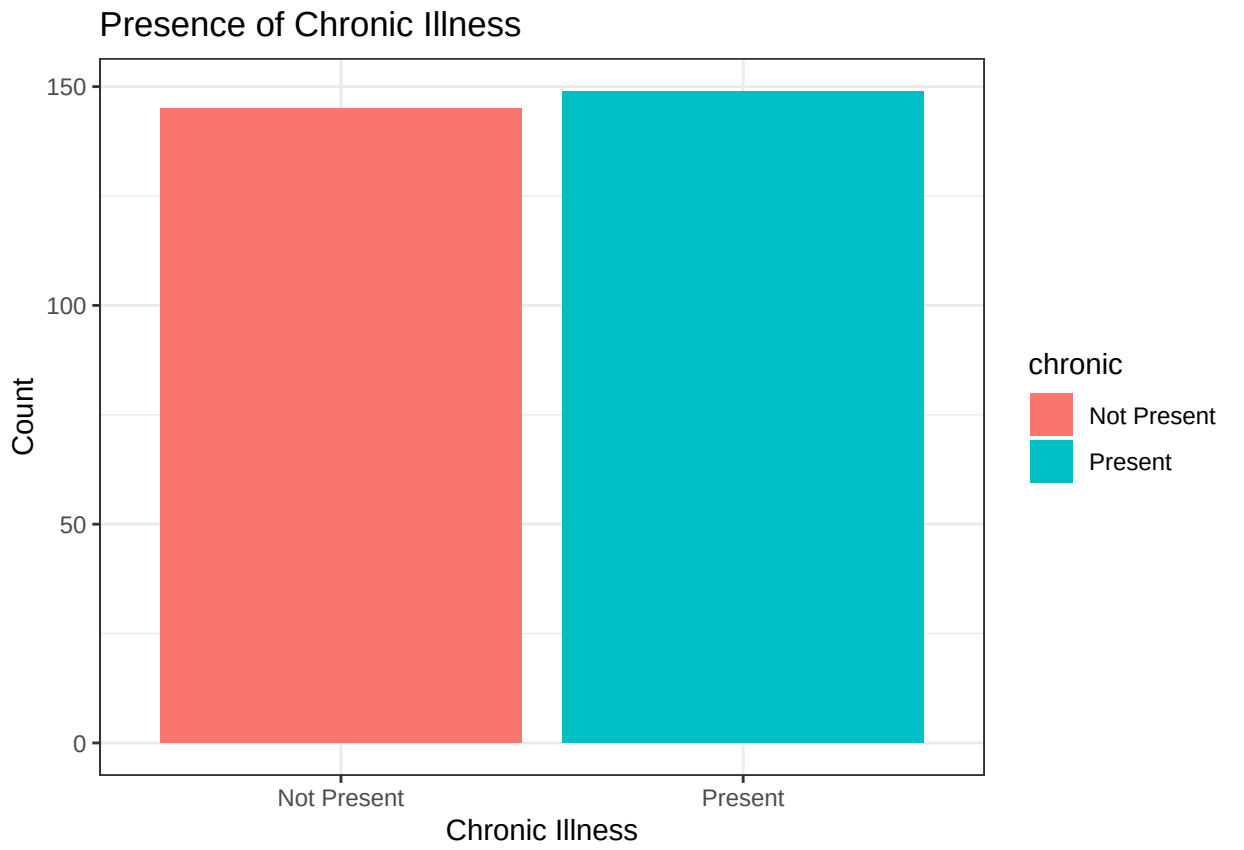
```
##   id sex age   marital       educat  employ income relig c1 c2 c3 c4 c5 c6 c7
## 1  1   1  68   Widowed      Some HS Retired      4     1  0  0  0  0  0  0  0
## 2  2   0  58  Divorced Some college      FT     15     1  0  0  1  0  0  0  0
## 3  3   1  45   Married      HS Grad      FT     28     1  0  0  0  0  1  0  0
## 4  4   1  50  Divorced      HS Grad   Unemp      9     1  0  0  0  0  1  1  0
## 5  5   1  33 Separated      HS Grad      FT     35     1  0  0  0  0  0  0  0
## 6  6   0  24   Married      HS Grad      FT     11     1  0  0  0  0  0  0  0
##   c8 c9 c10 c11 c12 c13 c14 c15 c16 c17 c18 c19 c20 cesd cases drink health
## 1  0  0   0   0   0   0   0   0   0   0   0   0   0    0     0     0      2
## 2  0  0   0   0   1   0   0   1   0   1   0   0   0    4     0     1      1
## 3  0  0   0   0   0   0   1   1   1   0   0   0   0    4     0     1      2
## 4  3  0   0   0   0   0   0   0   0   0   0   0   0    5     0     0      1
## 5  3  3   0   0   0   0   0   0   0   0   0   0   0    6     0     1      1
## 6  0  1   0   0   1   2   0   0   2   1   0   0   0    7     0     1      1
##   regdoc treat beddays acuteill chronill
## 1      1     1       0        0        1
## 2      1     1       0        0        1
## 3      1     1       0        0        0
## 4      1     0       0        0        1
## 5      1     1       1        1        0
## 6      1     1       0        1        1
```

#Univariate Exploration

```
chronic<-ifelse(depression$chronill ==1, "Present", "Not Present")
table(depression$chronill) # 0=FALSE/Not Present, 1=TRUE/Present
```

```
##
##   0   1
## 145 149
```

```
library(RColorBrewer)
ggplot(depression, aes(x=chronic, fill=chronic))+geom_bar()+theme_bw()+ggtitle("Presence of Chronic Illn
```

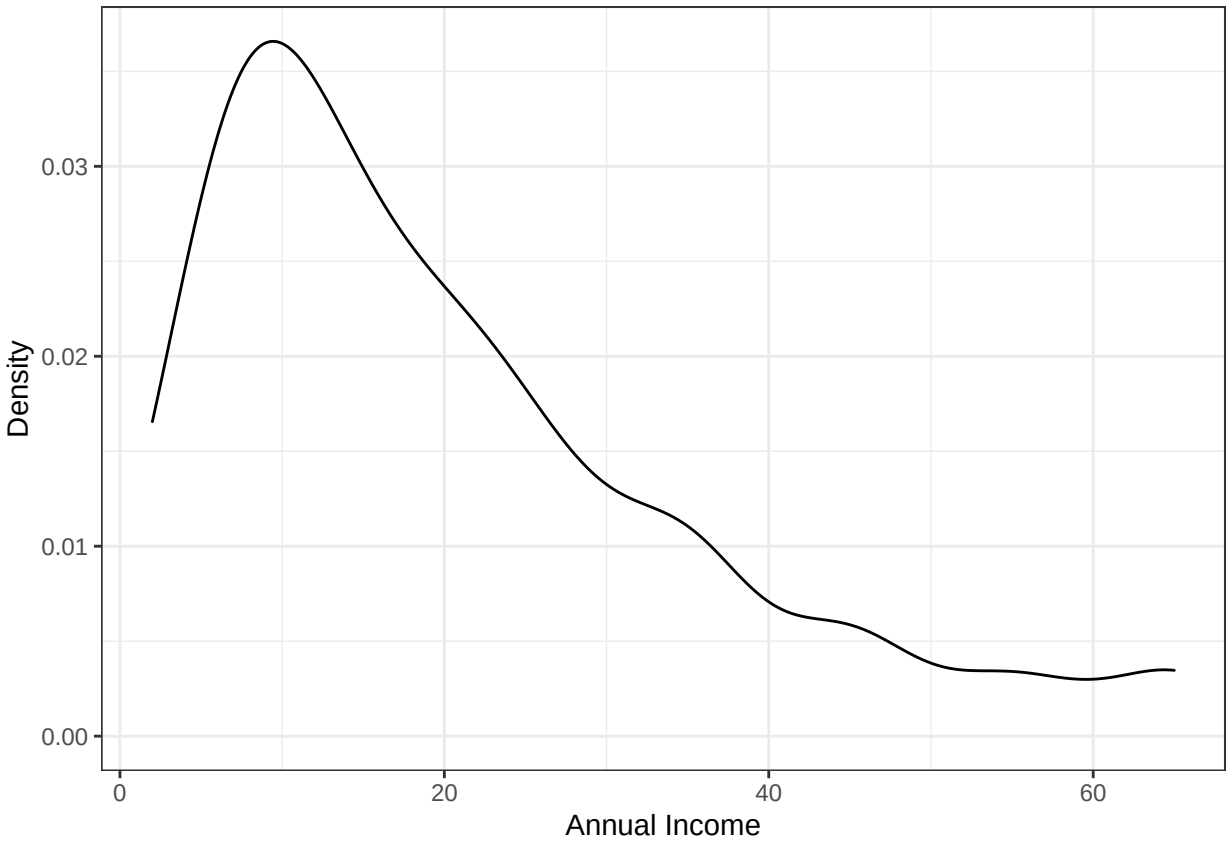


Of the participants interviewed for this data set, chronic illness within the past year was present in 149 individuals, and 145 individuals did not suffer from chronic illness.

```
summary(depression$income)
```

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

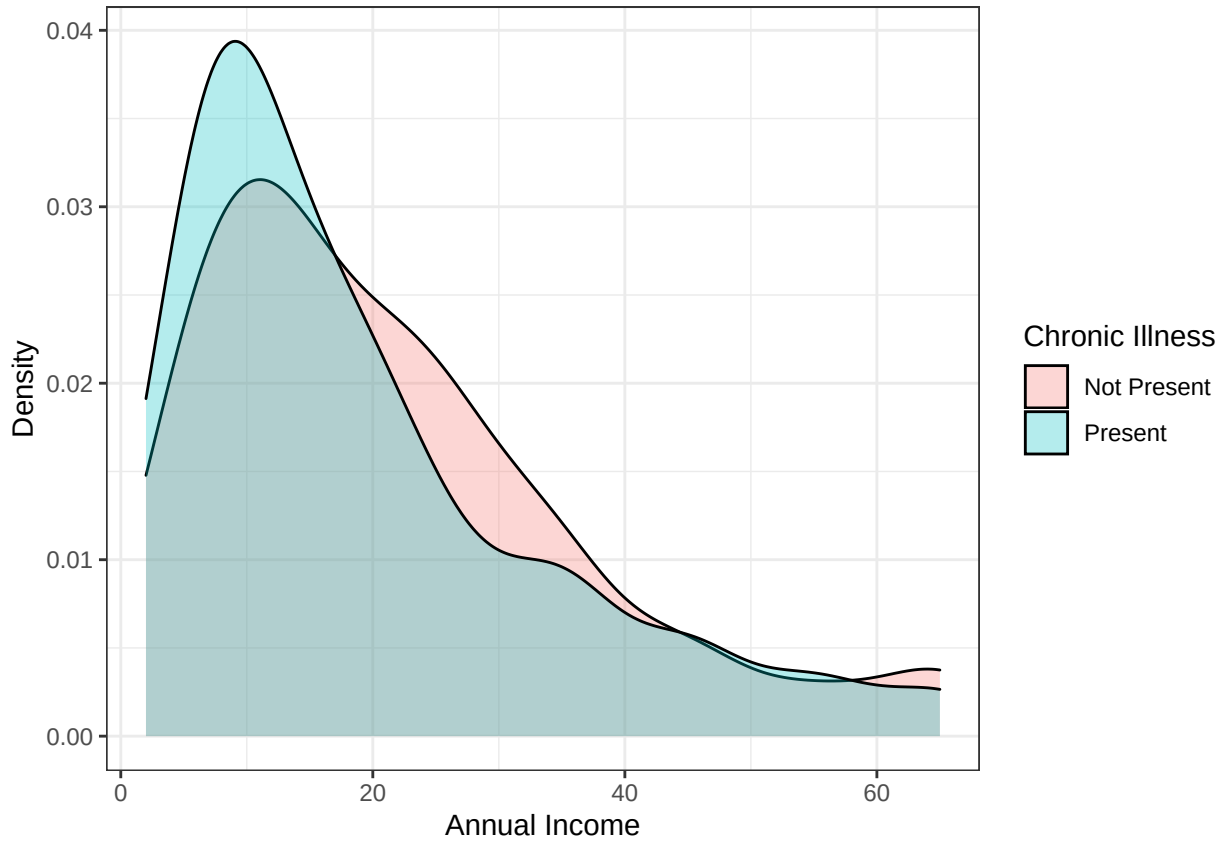
```
ggplot(depression, aes(x=income))+geom_density()+theme_bw()+xlab("Annual Income")+ylab("Density")
```



Of the participants interviewed for this data set, the lowest annual income was 2,000 dollars, and the highest was 65,000 dollars. The mean annual income was $20,570. The majority of the participants fell on the lower income end of the density plot.

#Bivariate Exploration

```
ggplot(depression, aes(x=income, fill=chronic))+geom_density(alpha=.3)+theme_bw()+xlab("Annual Income")+
```



This density plot compares the distribution of annual income to the presence of chronic illness within the past year. The peak presence of chronic illness was higher than the peak of no chronic illness presence.

#Conclusion

The peak presence of chronic illness occurred in lower income participants. The peak absence of chronic illness was also in lower income participants. This may be due to the average annual income of the entire sample size being on the low income end at $20,570 suggesting most of the participants were low income individuals. The peak presence was higher than the peak absence suggesting more individuals had chronic illness than not in the low income range. It may be possible that with a larger sample size with a wider range of income, a stronger correlation between annual income and chronic illness within the past year may be observed.