EDA_mdalberti

Melissa D'Alberti

2022-09-25

Introduction

This exploratory data analysis project will focus on the depression data set. The depression data set is from the first set of interviews of a prospective study of depression in the adult residents of Los Angeles County and it includes 294 observations. The code book indicates that those that were interviewed were asked a variety of lifestyle questions about gender, income, religion and much more. For this project I will analyze the income variable and the age variable which are both numerical variables.I'm interested in how income and age relate to a patient being depressed or not, is there any correlation?

```
library(openintro)
```

```
## Loading required package: airports

## Loading required package: cherryblossom

## Loading required package: usdata
```

```
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)
depress <-read.delim("/Users/rosadalberti/Desktop/Math130/data/depress_081217.txt", header=TRUE,sep="\t
dim(depress)
```

```
## [1] 294  37
```

Univariate Exploration

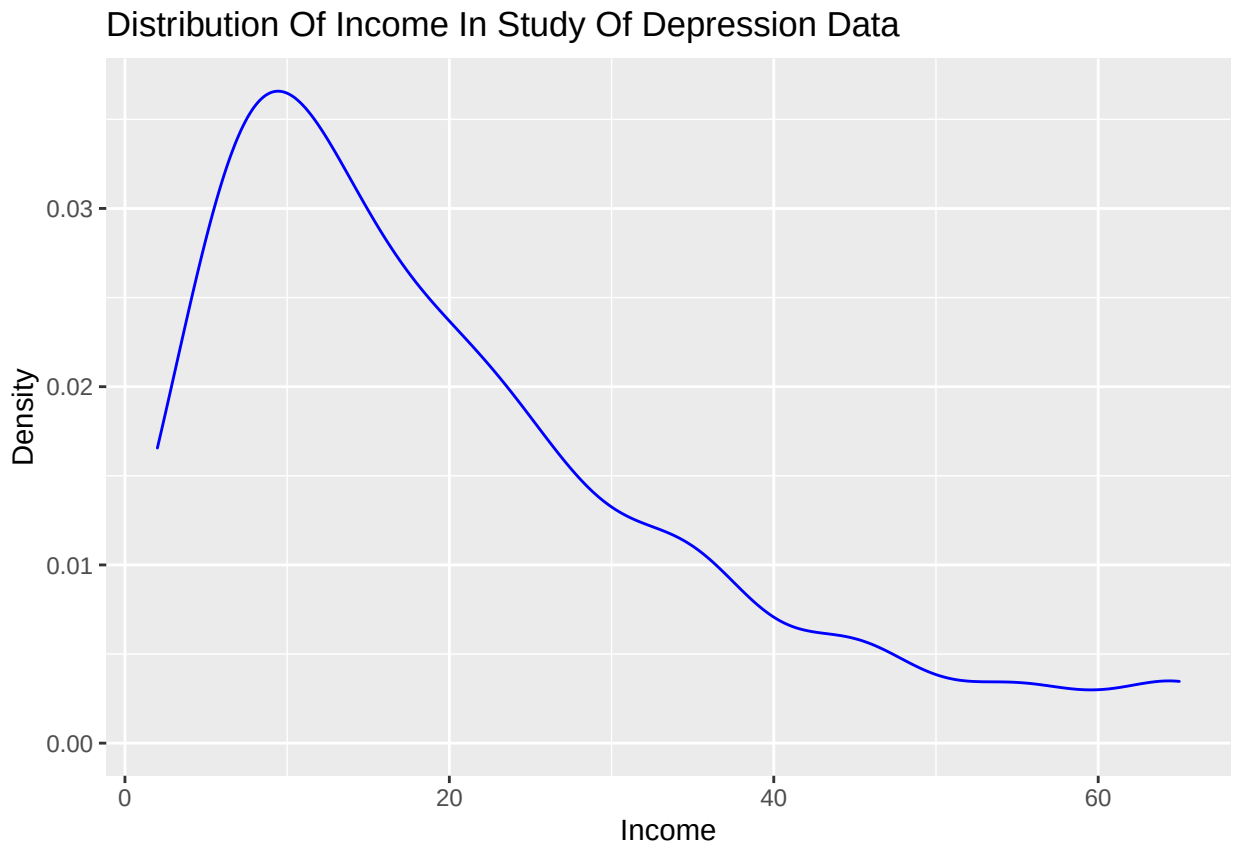
Income

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

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

This variable is both numerical and continuous. We now know based off of this table that the minimum income in the thousands is 2.00 with a maximum income amount of 65.00. The mean income of the data set depress is 20.57. Importantly, these values represent a number in the thousands.

```
ggplot(depress, aes(x=income))+geom_density(col="blue") +
xlab("Income")+ylab("Density") +
ggtitle("Distribution Of Income In Study Of Depression Data")
```



This distribution shows that there could be a correlation between lower income and a higher depression density with a relatively steep decline as income increases indicating that when income grows then depression decreases.

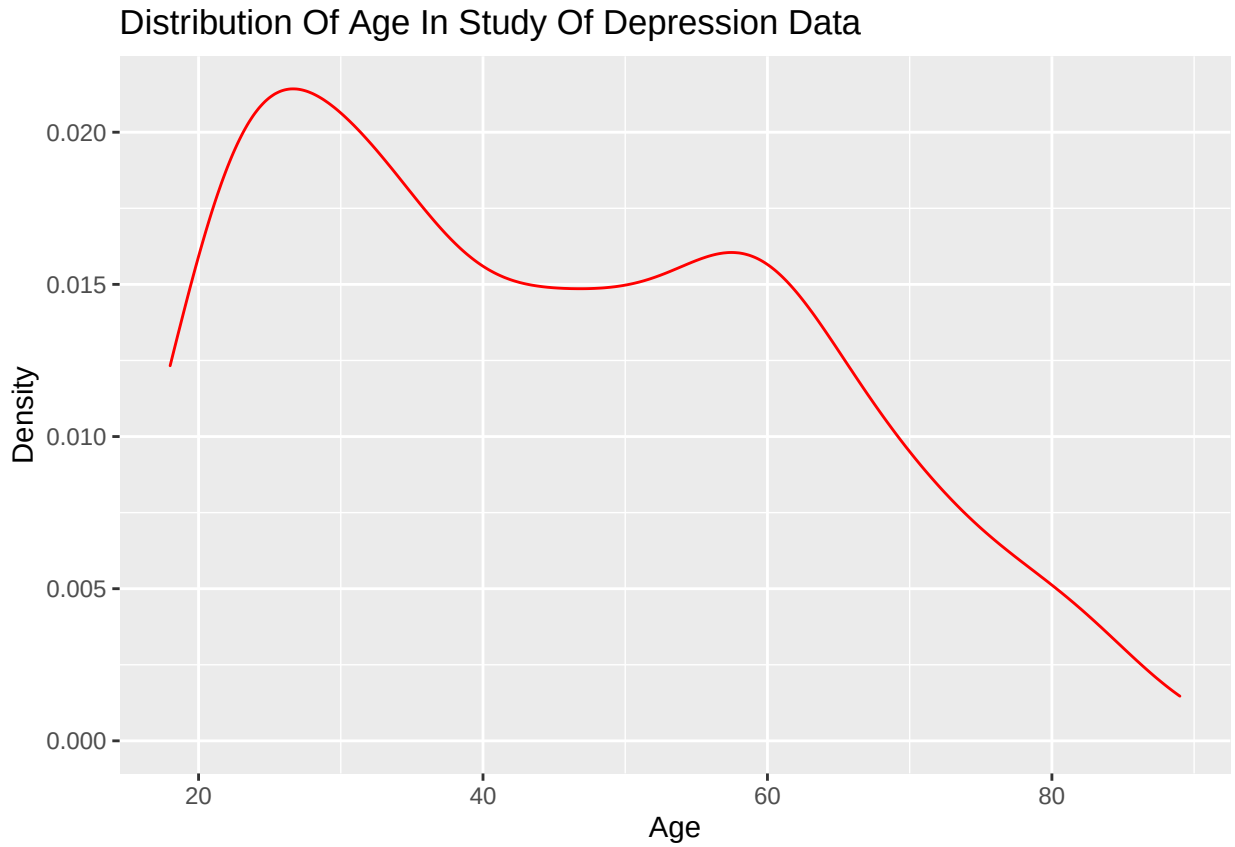
Age

```
summary(depress$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   28.00   42.50   44.41   59.00   89.00
```

This variable that we are examining is also numerical and continuous. The average of the age variable is 44.41. This also has a large range of numbers with a minimum age of 18.00 and maximum age of 89.00.

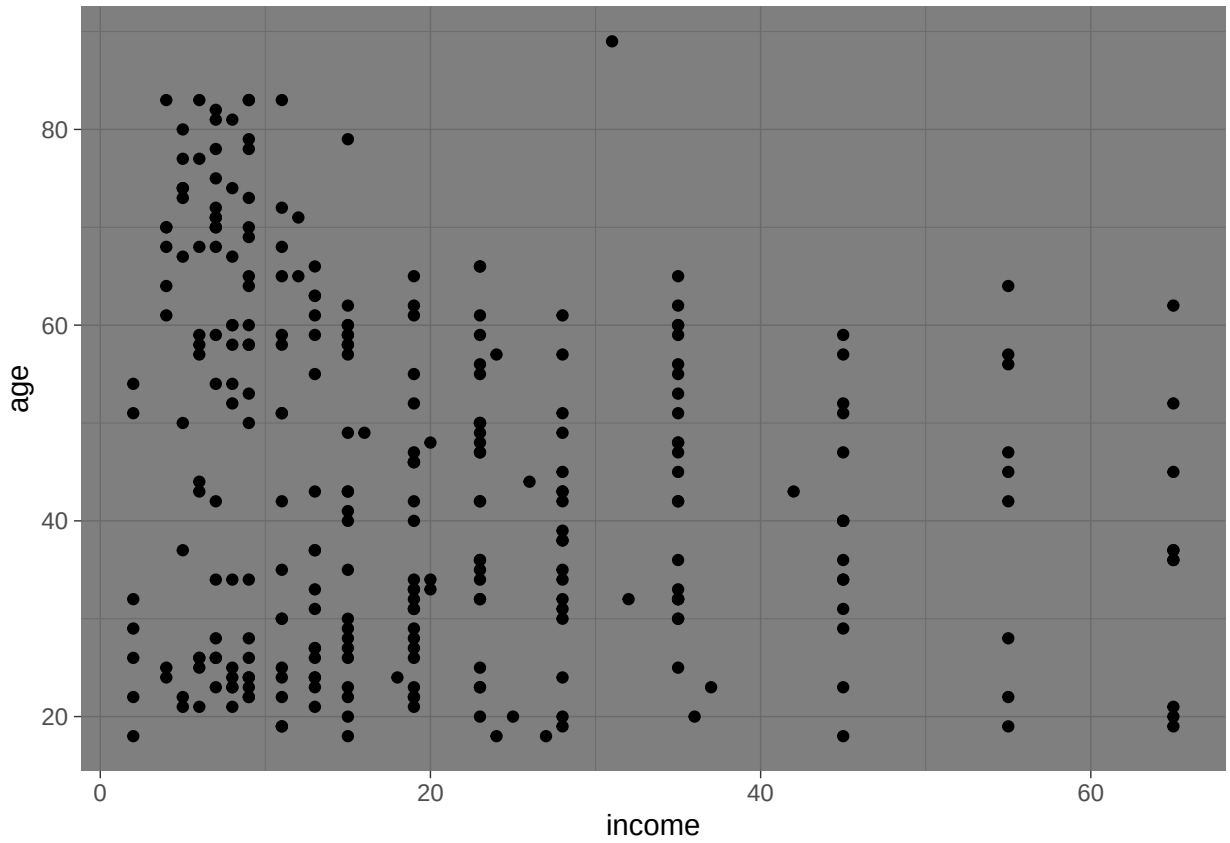
```
ggplot(depress, aes(x=age))+geom_density(col="red") +
xlab("Age")+ylab("Density") +
ggtitle("Distribution Of Age In Study Of Depression Data")
```



Similarly to the income variable, this plot shows a correlation between younger age and higher depression.In comparison, a steep drop on the distribution shows older subjects of the survey with a much lower density of depression. One interesting thing to note is the dip in the middle of ages 40 to 50 and once again an increase before the steep drop.

Bivariate Exploration

```
ggplot(depress, aes(x=income, y=age)) + geom_point()+theme_dark()
```



This summary of the variables age and income shows that most of the data points lie within the lowest income bracket of 0 to 20,000 despite the age values. This could signify a correlation between lower incomes and depression among all ages in the data set. Points on the plot become less dense as income grows.

Conclusion

After exploring variables by using plots and summary statistics, I was able to see strong correlations between low income and high depression. Surprisingly, there was a large density in depression among a large range of ages. Those that had the lowest incomes despite ages seemed to have the most depression density. Ultimately, the data did appear to support my predictions.