EDA_AyeshaAhmad

Ayesha Ahmad

2024-09-28

```
library(ggplot2)
```

Introduction

In this Explanatory Variable Analysis Project, I analyzed the Depression Data Set. The data set is from the first set of interviews in the propseptive study of depression in adults of Los Angeles. There was 294 observations and ten variables in this data set all together. I will be focusing on the variables, Age and Education. My reserach question is whether or not higher education amongst different ages of people affects depression rates.

```
depression <- read.delim("/Users/ayeshaahmad/Desktop/Depress.txt")
```

```
head(depression)
```

```
##   ID SEX AGE MARITAL EDUCAT EMPLOY INCOME RELIG C1 C2 C3 C4 C5 C6 C7 C8 C9 C10
## 1  1   2  68       5      2      4      4     1  0  0  0  0  0  0  0  0  0   0
## 2  2   1  58       3      4      1     15     1  0  0  1  0  0  0  0  0  0   0
## 3  3   2  45       2      3      1     28     1  0  0  0  0  1  0  0  0  0   0
## 4  4   2  50       3      3      3      9     1  0  0  0  0  1  1  0  3  0   0
## 5  5   2  33       4      3      1     35     1  0  0  0  0  0  0  0  3  3   0
## 6  6   1  24       2      3      1     11     1  0  0  0  0  0  0  0  0  1   0
##   C11 C12 C13 C14 C15 C16 C17 C18 C19 C20 CESD CASES DRINK HEALTH REGDOC TREAT
## 1   0   0   0   0   0   0   0   0   0   0    0     0     2      2      1     1
## 2   0   1   0   0   1   0   1   0   0   0    4     0     1      1      1     1
## 3   0   0   0   1   1   1   0   0   0   0    4     0     1      2      1     1
## 4   0   0   0   0   0   0   0   0   0   0    5     0     2      1      1     2
## 5   0   0   0   0   0   0   0   0   0   0    6     0     1      1      1     1
## 6   0   1   2   0   0   2   1   0   0   0    7     0     1      1      1     1
##   BEDDAYS ACUTEILL CHRONILL
## 1       0        0        1
## 2       0        0        1
## 3       0        0        0
## 4       0        0        1
## 5       1        1        0
## 6       0        1        1
```

The variable I focused on in this data set was age and education.

AGE Continuous

EDUCAT 1 = Less than high school
2 = Some high school

3 = Finished high school
4 = Some college
5 = Finished bachelor's
degree
6 = Finished master's
degree
7 = Finished doctorate

Univariate Variable

The first variable that I chose to analyze is Age. Age is continous in this case and is shown as the last birthday they celebrated.

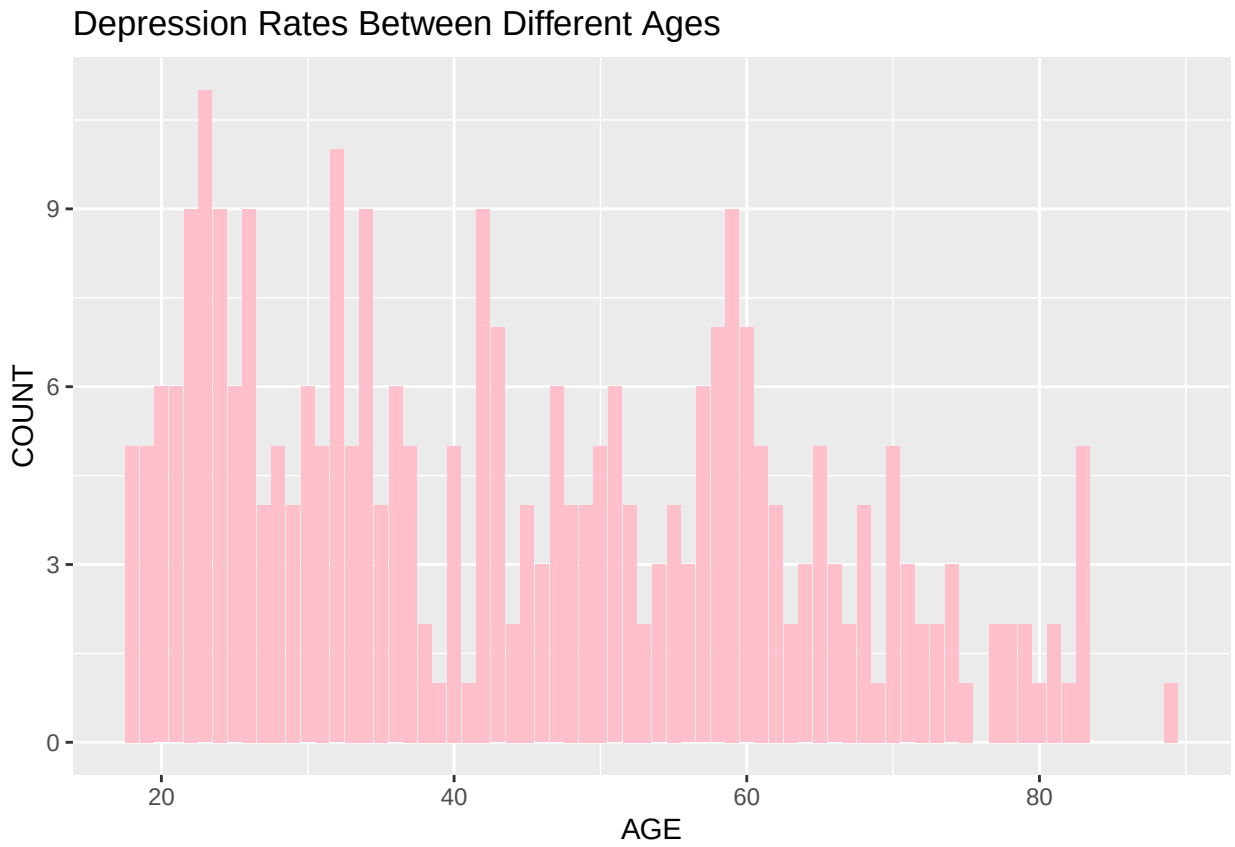
```
table(depression$AGE)
```

```
##
## 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43
##  5  5  6  6  9 11  9  6  9  4  5  4  6  5 10  5  9  4  6  5  2  1  5  1  9  7
## 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69
##  2  4  3  6  4  4  5  6  4  2  3  4  3  6  7  9  7  5  4  2  3  5  3  2  4  1
## 70 71 72 73 74 75 77 78 79 80 81 82 83 89
##  5  3  2  2  3  1  2  2  2  1  2  1  5  1
```

```
summary(depression$AGE)
```

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

```
ggplot(depression, aes(x=AGE)) + geom_bar(fill="pink") + xlab("AGE") + ylab("COUNT") + ggtitle("Depressi
```



The bar graph shows the diversity between the depression rates between the different ages of people who were part of this study. From this bar graph, it can be predicted that the highest depression rates comes from those in their early twenties, as well as early thrities. It can also be seen that the least depression rates are from those who are in their late thirties and early forties.

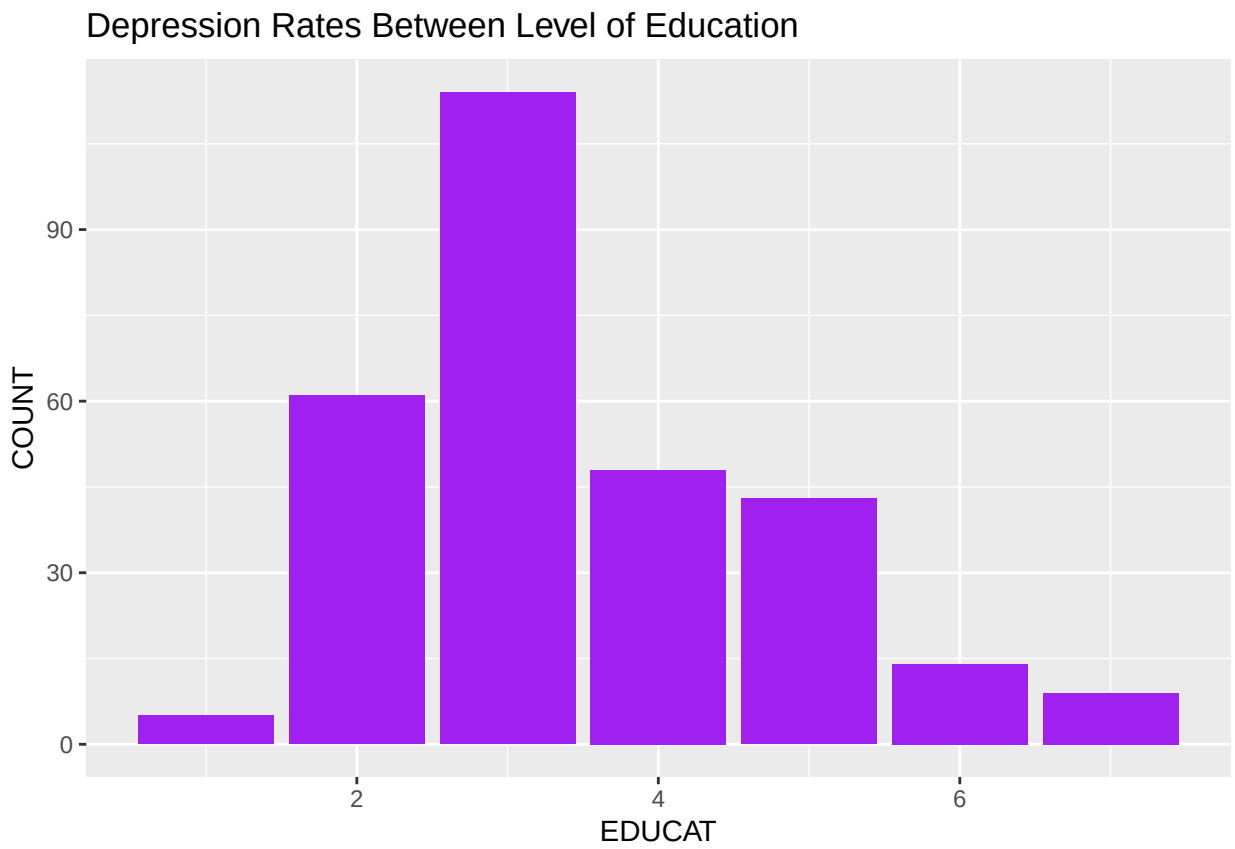
```
table(depression$EDUCAT)
```

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

```
summary(depression$EDUCAT)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    3.00    3.00    3.48    4.00    7.00
```

```
ggplot(depression, aes(x=EDUCAT)) + geom_bar(fill="purple")+  xlab("EDUCAT") + ylab("COUNT") + ggtitle(
```



The bar graph represnts the depression rates betwwen the different levels of education comapleted. The values are rpesented here, 1 = Less than high school, 2 = Some high school, 3 = Finished high school, 4 = Some college, 5 = Finished bachelor's degree, 6 = Finished master's degree, 7 = Finished doctorate. We are able to conclude that just from this variable, the most depression rates come from those who have only finished high school. The least rates of depression are from those who have less than a high school eduction or from those who have finshed a doctorate's degree.

Bivariate Exploration

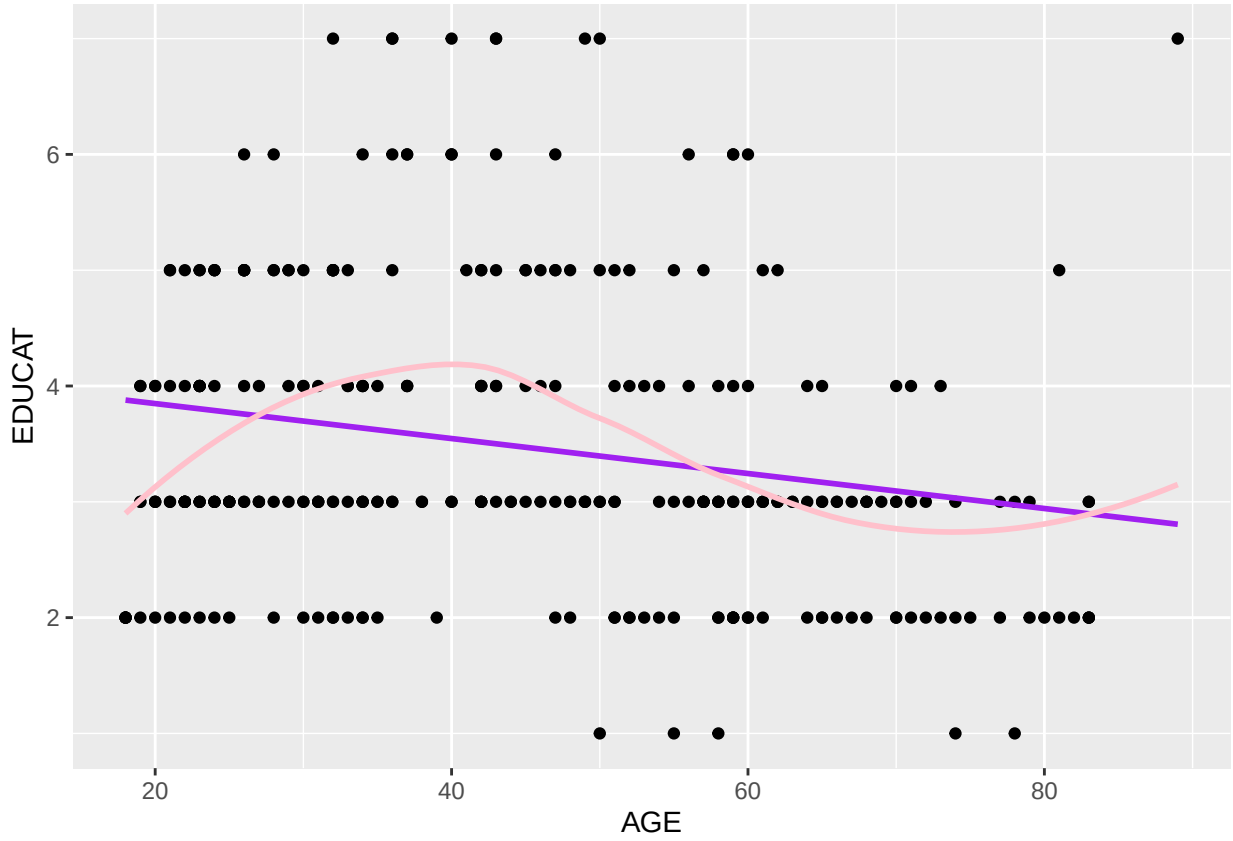
```
table(depression$AGE, depression$EDUCAT)
```

```
##
##      1 2 3 4 5 6 7
##   18 0 5 0 0 0 0 0
##   19 0 1 2 2 0 0 0
##   20 0 1 3 2 0 0 0
##   21 0 1 2 1 2 0 0
##   22 0 1 6 1 1 0 0
##   23 0 1 3 5 2 0 0
##   24 0 1 4 1 3 0 0
##   25 0 1 5 0 0 0 0
##   26 0 0 2 1 5 1 0
##   27 0 0 3 1 0 0 0
##   28 0 1 1 0 2 1 0
##   29 0 0 1 1 2 0 0
##   30 0 1 2 2 1 0 0
```

```
##   31 0 1 3 1 0 0 0
##   32 0 2 2 0 5 0 1
##   33 0 1 1 2 1 0 0
##   34 0 2 3 3 0 1 0
##   35 0 1 2 1 0 0 0
##   36 0 0 2 0 1 1 2
##   37 0 0 0 3 0 2 0
##   38 0 0 2 0 0 0 0
##   39 0 1 0 0 0 0 0
##   40 0 0 2 0 0 2 1
##   41 0 0 0 0 1 0 0
##   42 0 0 4 3 2 0 0
##   43 0 0 1 2 1 1 2
##   44 0 0 2 0 0 0 0
##   45 0 0 1 1 2 0 0
##   46 0 0 1 1 1 0 0
##   47 0 1 1 1 2 1 0
##   48 0 1 2 0 1 0 0
##   49 0 0 3 0 0 0 1
##   50 1 0 2 0 1 0 1
##   51 0 2 2 1 1 0 0
##   52 0 2 0 1 1 0 0
##   53 0 1 0 1 0 0 0
##   54 0 1 1 1 0 0 0
##   55 1 1 1 0 1 0 0
##   56 0 0 1 1 0 1 0
##   57 0 0 5 0 1 0 0
##   58 1 2 3 1 0 0 0
##   59 0 4 2 1 0 2 0
##   60 0 2 3 1 0 1 0
##   61 0 1 3 0 1 0 0
##   62 0 0 3 0 1 0 0
##   63 0 0 2 0 0 0 0
##   64 0 1 1 1 0 0 0
##   65 0 2 2 1 0 0 0
##   66 0 1 2 0 0 0 0
##   67 0 1 1 0 0 0 0
##   68 0 1 3 0 0 0 0
##   69 0 0 1 0 0 0 0
##   70 0 2 2 1 0 0 0
##   71 0 1 1 1 0 0 0
##   72 0 1 1 0 0 0 0
##   73 0 1 0 1 0 0 0
##   74 1 1 1 0 0 0 0
##   75 0 1 0 0 0 0 0
##   77 0 1 1 0 0 0 0
##   78 1 0 1 0 0 0 0
##   79 0 1 1 0 0 0 0
##   80 0 1 0 0 0 0 0
##   81 0 1 0 0 1 0 0
##   82 0 1 0 0 0 0 0
##   83 0 3 2 0 0 0 0
##   89 0 0 0 0 0 0 1
```

```
ggplot(depression, aes(x=AGE, y=EDUCAT)) + geom_point() + geom_smooth(se=FALSE, method="lm", color="purp
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



The scatterplot brings both variables, education, and age, into consideration. It can be seen that the level of education that has the most depression was 3, which represents "finished high school" while the ages that contributes most to the depression rates is the early tweinties to the early thirties. The relationshop between the lowest rates can also be seen through this scatterplot. The education levels of "finshed doctorate" and "some high school" were the least depressed, especially those in their late thirties and early forties.

Conclusion

In conclusion, it can be stated that education levels has a correlation to age when it comes to depression rates. In this study, the most depression rates were seen amongst those who had finished high school and were in their early twenties. Through the diffferent grpahs and displays that we looked at, we were also able to conclude that the least depression rates came from those in their late thirties and early forties and have finshed some high school or finshed a doctorate degree. The younger ages had more depression rates than those who were older in age and had completed more education. The research question of whether or not a higher education leads to higher depression rates can be answered after looking at the analysis. A higher education amongst different ages does not lead to higher depression rates, but rather a lower education has led to higher depression rates.