# Exploratory Data Analysis Project

Jessica Rualo

2025-06-28

## Introduction

The data set I'm using from the ones provided is the Depression set. My research question is, how are age and general health status related to depression among these patients? I'm interested in finding out to see if those variables are related to depression. For example, if their health is poorer than others, does that affect their chance of having depression or depression symptoms. The first variable is CESD which is the continuous measure of depressive symptoms, ranging from 0 to 60. Having a higher score indicates more symptoms. The second variable is age. The third variable is health. It ranges from 1 being excellent and 4 being poor. Out of all the variables, I chose these because I think that there could be a good chance that they correlate with each other.

```
depression <- read.csv("C:/math130/data/Depress.csv", header=TRUE)
```

## Univariate Exploration

I'm going to calculate summary statistics for each of my variables. After, I will use histograms for my numeric variables and then a bar chart for my categorical variable.

```
summary(depression$CESD)
```

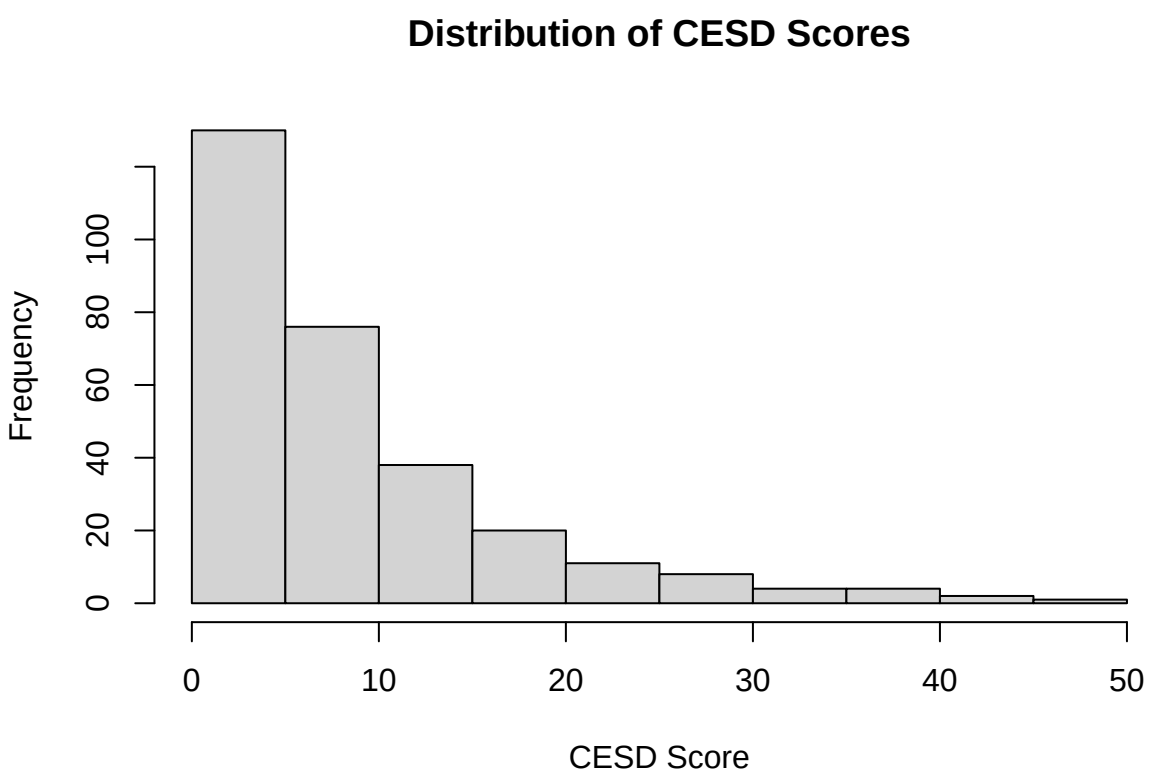
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.000   7.000   8.884  12.000  47.000
```

```
sd(depression$CESD, na.rm = TRUE)
```

```
## [1] 8.823655
```

This shows the average score is about 9 which is good since those with the most depression symptoms is 60.

```
hist(depression$CESD, main= "Distribution of CESD Scores", xlab = "CESD Score")
```




This graph shows that depressive symptoms are skewed right. There are less people with more depression symptoms.

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

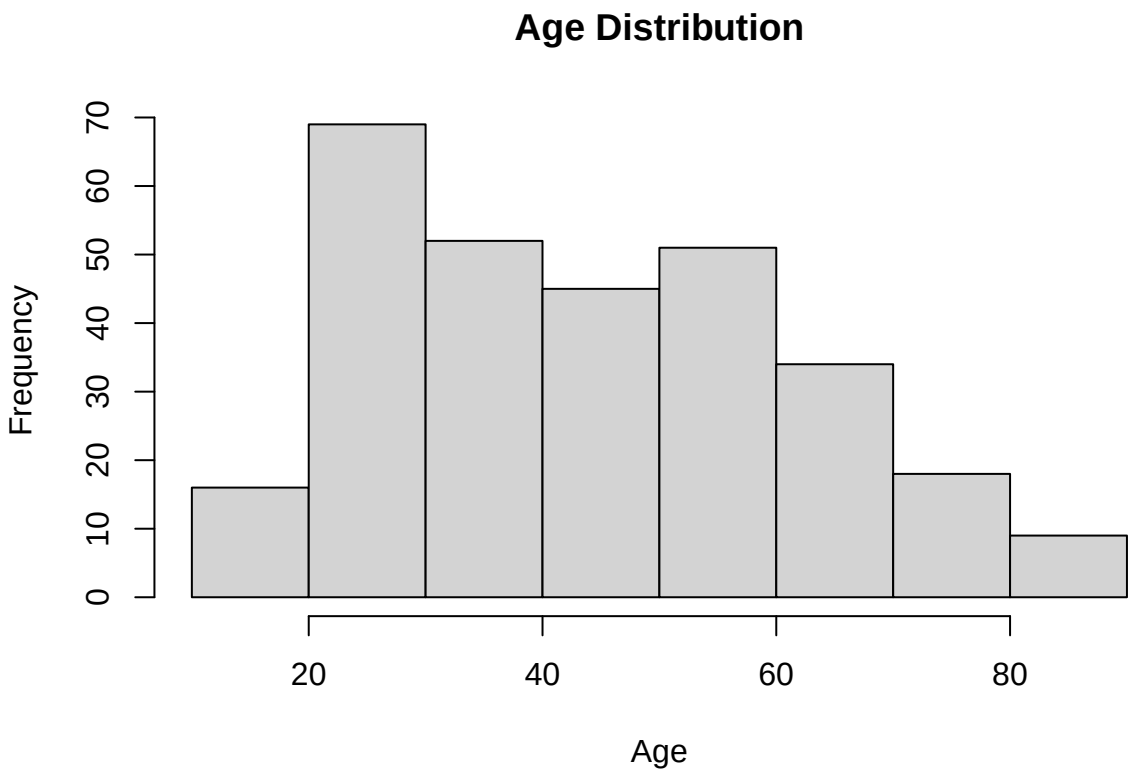
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   28.00   42.50   44.41   59.00   89.00
```

```
sd(depression$AGE, na.rm = TRUE)
```

```
## [1] 18.08544
```

The average age is about 44 years old; the ages can ran from 18 to 89 years old.

```
hist(depression$AGE, main = "Age Distribution", xlab = "Age")
```




There are more people that are around 20-30 years old in this data set.

For my single categorical variable, I'm looking at the

```
table(depression$HEALTH)
```

```
##
##   1   2   3   4
## 130 115  35  14
```

```
prop.table(table(depression$HEALTH))*100
```

```
##
##         1         2         3         4
## 44.217687 39.115646 11.904762  4.761905
```

This shows how many people reprted in each category, the top row being the actual number of participants and the second row showing the proportions in percentages. Poor health is rare with almost 5% and most people have excellent health.

```
barplot(table(depression$HEALTH), names.arg = c("Excellent", "Good", "Fair", "Poor"), main = "Self-Rate
```


This shows how many people reported each health status.

## Bivariate Exploration

1. CESD vs. Health I'm going to compare CESD which is my categorical variable and Health which is my numeric variable.

```
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
```

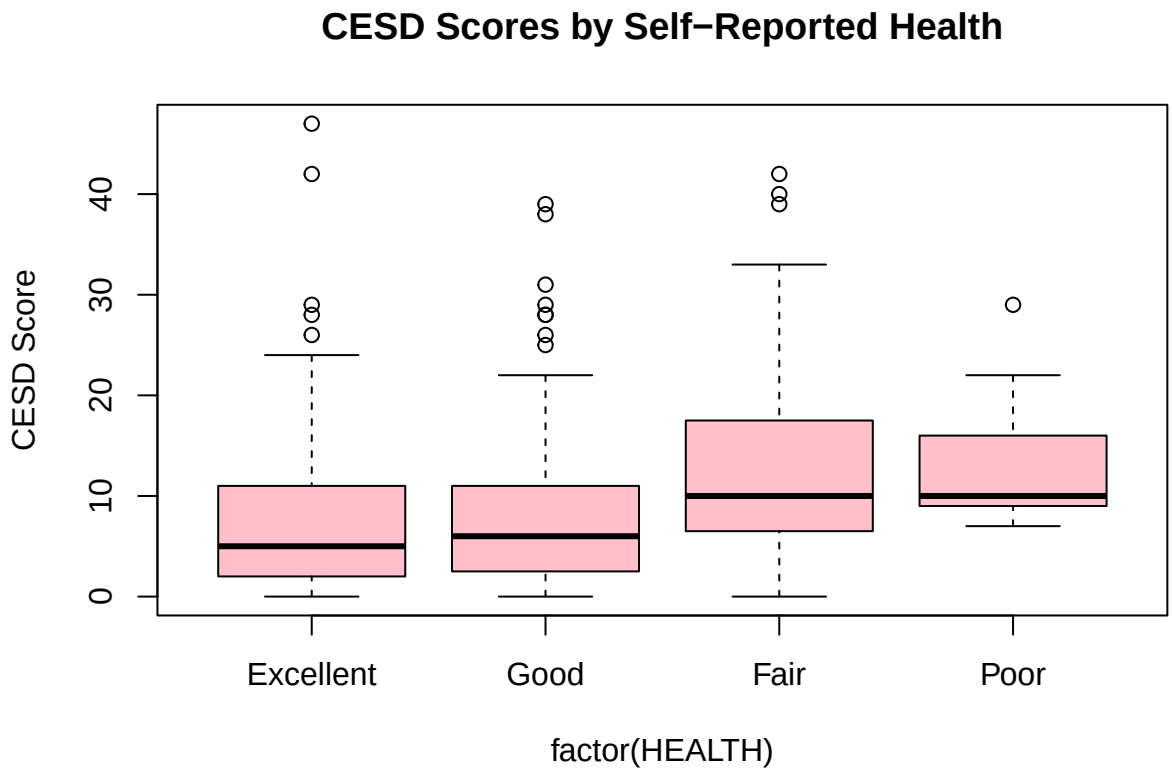
```
depression %>% group_by(HEALTH) %>%summarise(N = n(),Mean_CESD = mean(CESD, na.rm = TRUE), SD_CESD = sd
```

```
## # A tibble: 4 x 4
```

```
##   HEALTH     N Mean_CESD SD_CESD
##    <int> <int>     <dbl>   <dbl>
## 1      1   130      7.64    8.16
## 2      2   115      8.23    8.23
## 3      3    35     13.9    11.6
## 4      4    14     13.2     6.34
```

These results tell us that those with who rated their health as Fair also had the highest CESD scores while those who rated their health as Excellent had the lowest CESD scores. While the those who rated their health as Poor averaged a 13.2, it's still on the higher end. This can suggest a negative relationship between the participants' overall health and mental health.

```
boxplot(CESD ~ factor(HEALTH), data = depression, names = c("Excellent", "Good", "Fair", "Poor"),main =
```




I used a box plot to see the distribution of the data between CESD score and Health. There are a lot of outliers with higher scores in the Excellent category; however, it doesn't affect the overall result. The minimum score for the Poor category is the highest between the four categoies.

2. CESD vs. Age The second set I'm comparing is between both of my numeric variables which are CESD and age,

```
cor(depression$AGE, depression$CESD, use = "complete.obs")
```

```
## [1] -0.1641447
```

The correlation between age and CESD score was -0.16 which shows a weak negative relationship.

```
library(ggplot2)

ggplot(data = depression, mapping = aes(x = depression$AGE, y = depression$CESD)) + geom_point(color =
```

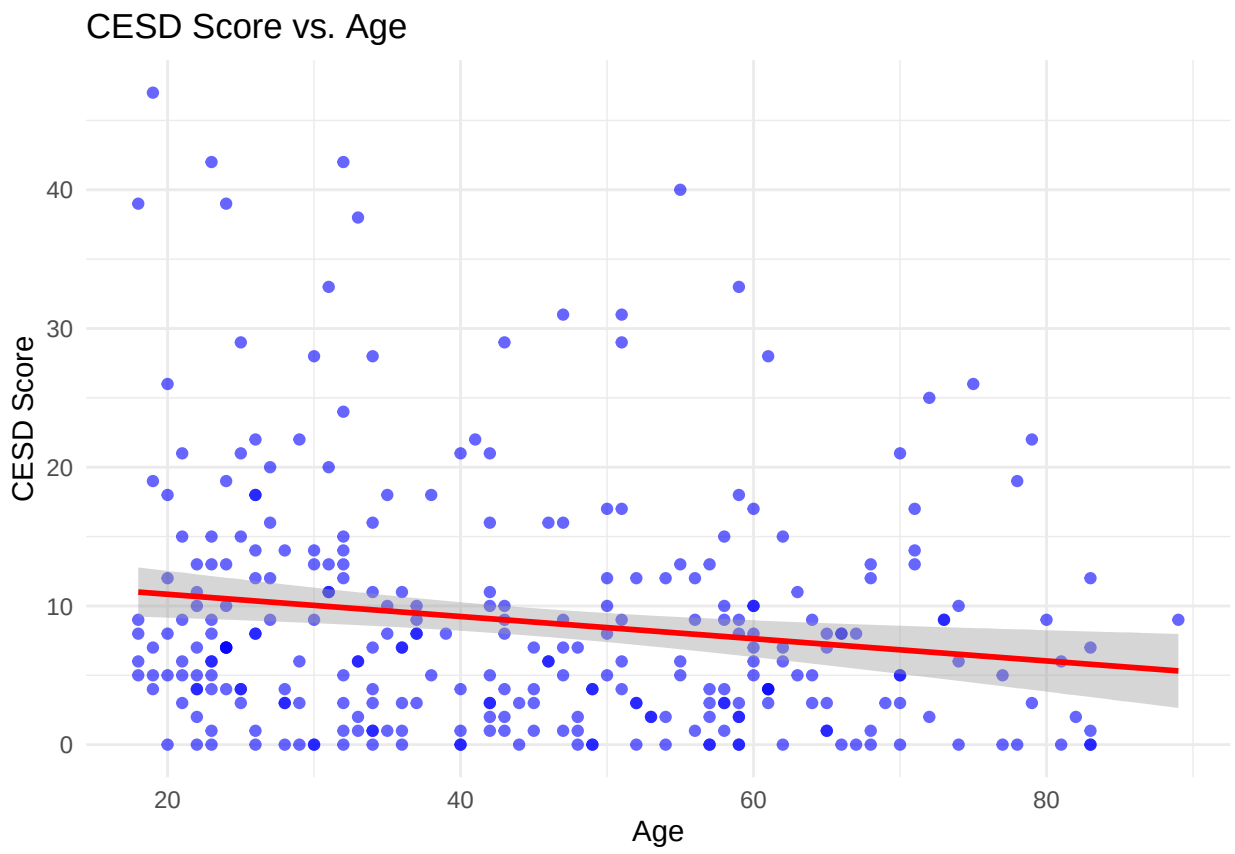
```
## Warning: Use of ‘depression$AGE‘ is discouraged.
## i Use ‘AGE‘ instead.
```

```
## Warning: Use of ‘depression$CESD‘ is discouraged.
## i Use ‘CESD‘ instead.
```

```
## Warning: Use of ‘depression$AGE‘ is discouraged.
## i Use ‘AGE‘ instead.
```

```
## Warning: Use of ‘depression$CESD‘ is discouraged.
## i Use ‘CESD‘ instead.
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```




This scatter plot shows the relationship between age and CESD score. There is a negative relationship because of the downward slope. This comparison also shows us a weak negative relationship between age and depressive symptoms.

## Conclusion

In this study, I looked at how depression scores relate to age and self-reported health. Most people said their health was good or excellent. People who reported fair to poorer health had higher depression scores, which means worse health is linked to more depressive symptoms. Age showed a small negative relationship with depression scores, meaning older people tended to have slightly fewer symptoms, but this was a weak connection. Overall, self-rated health seems more important for depression than age in this group. These results help us understand that how people feel about their health relates to their mental well-being.