EDA4_JAM

Jordan McKenzie

2025-06-27

```
depress <- read.delim("/Users/jordenzie/Desktop/math130/data/Depress.txt")

str(depress)
```

```
## 'data.frame':    294 obs. of  37 variables:
##  $ ID      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ SEX     : int  2 1 2 2 2 1 2 1 2 1 ...
##  $ AGE     : int  68 58 45 50 33 24 58 22 47 30 ...
##  $ MARITAL : int  5 3 2 3 4 2 2 1 2 2 ...
##  $ EDUCAT  : int  2 4 3 3 3 3 2 3 3 2 ...
##  $ EMPLOY  : int  4 1 1 3 1 1 5 1 4 1 ...
##  $ INCOME  : int  4 15 28 9 35 11 11 9 23 35 ...
##  $ RELIG   : int  1 1 1 1 1 1 1 1 2 4 ...
##  $ C1      : int  0 0 0 0 0 0 2 0 0 0 ...
##  $ C2      : int  0 0 0 0 0 0 1 1 1 0 ...
##  $ C3      : int  0 1 0 0 0 0 1 2 1 0 ...
##  $ C4      : int  0 0 0 0 0 0 2 0 0 0 ...
##  $ C5      : int  0 0 1 1 0 0 1 2 0 0 ...
##  $ C6      : int  0 0 0 1 0 0 0 1 3 0 ...
##  $ C7      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ C8      : int  0 0 0 3 3 0 2 0 0 0 ...
##  $ C9      : int  0 0 0 0 3 1 2 0 0 0 ...
##  $ C10     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ C11     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ C12     : int  0 1 0 0 0 1 0 0 3 0 ...
##  $ C13     : int  0 0 0 0 0 2 0 0 0 0 ...
##  $ C14     : int  0 0 1 0 0 0 0 0 3 0 ...
##  $ C15     : int  0 1 1 0 0 0 3 0 2 0 ...
##  $ C16     : int  0 0 1 0 0 2 0 1 3 0 ...
##  $ C17     : int  0 1 0 0 0 1 0 1 0 0 ...
##  $ C18     : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ C19     : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ C20     : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ CESD    : int  0 4 4 5 6 7 15 10 16 0 ...
##  $ CASES   : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ DRINK   : int  2 1 1 2 1 1 2 2 1 1 ...
##  $ HEALTH  : int  2 1 2 1 1 1 3 1 4 1 ...
##  $ REGDOC  : int  1 1 1 1 1 1 1 2 1 1 ...
##  $ TREAT   : int  1 1 1 2 1 1 1 2 1 2 ...
##  $ BEDDAYS : int  0 0 0 0 1 0 0 0 1 0 ...
##  $ ACUTEILL: int  0 0 0 0 1 1 1 1 0 0 ...
##  $ CHRONILL: int  1 1 0 1 0 1 1 0 1 0 ...
```

Introduction

Hi! For my project I will be utilizing the 'Depression' dataset, which includes information on people's depression scores (CESD), their age, how they rated their own health, their level of education, etc. I wanted to see if there's any connection between these things — like does feeling healthy or having more education mean people feel less depressed? Let's explore.

I will mainly be focusing on these three variables...

CESD: This is the depression score. A higher score means more symptoms of depression. AGE: How old the person is. EDUCAT: What education experience this person has.

Research Question: Does age and education level relate to overall depression scores (CESD)?

Let's get this party started (and try not to get too depressed)!

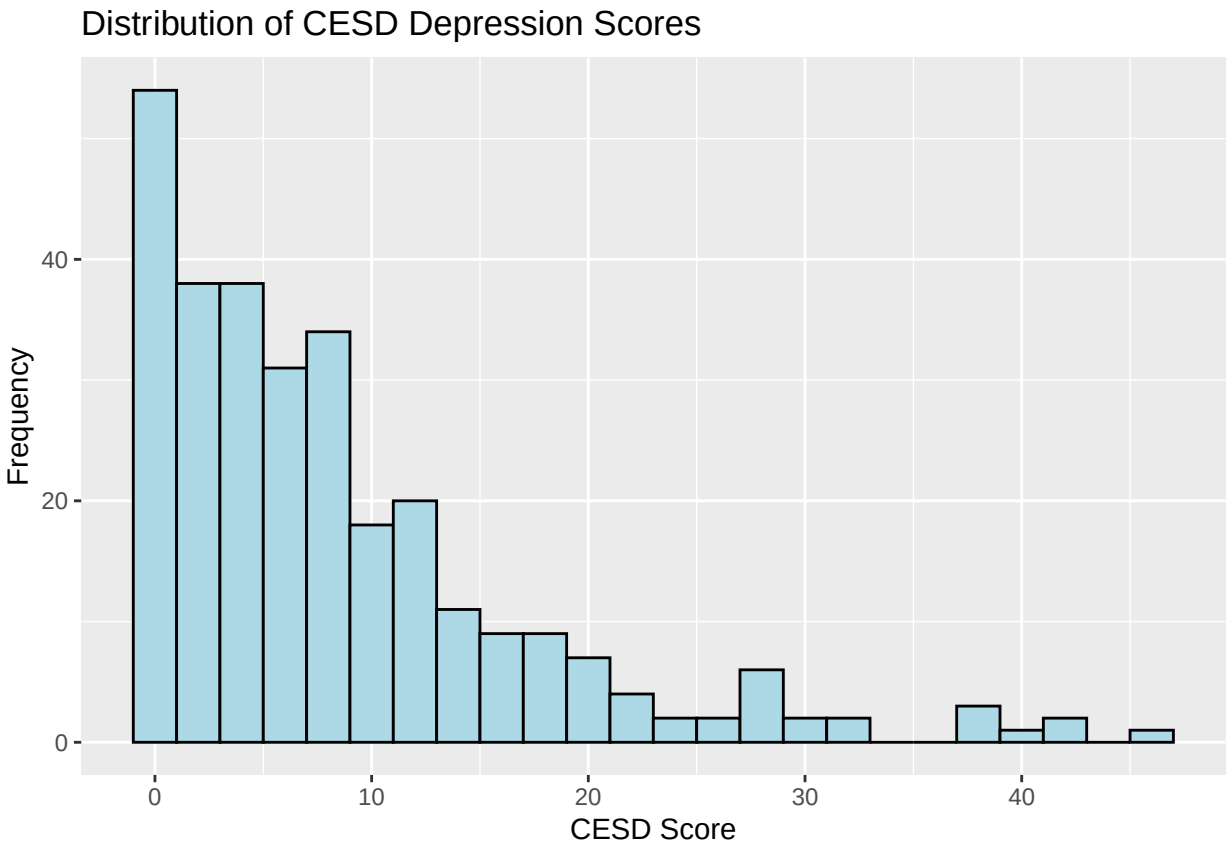
Univariate Exploration

```
summary(depress$CESD)
```

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

```
library(ggplot2)

ggplot(depress, aes(x = CESD)) +
  geom_histogram(binwidth = 2, color = "black", fill = "lightblue") +
  ggtitle("Distribution of CESD Depression Scores") +
  xlab("CESD Score") +
  ylab("Frequency")
```



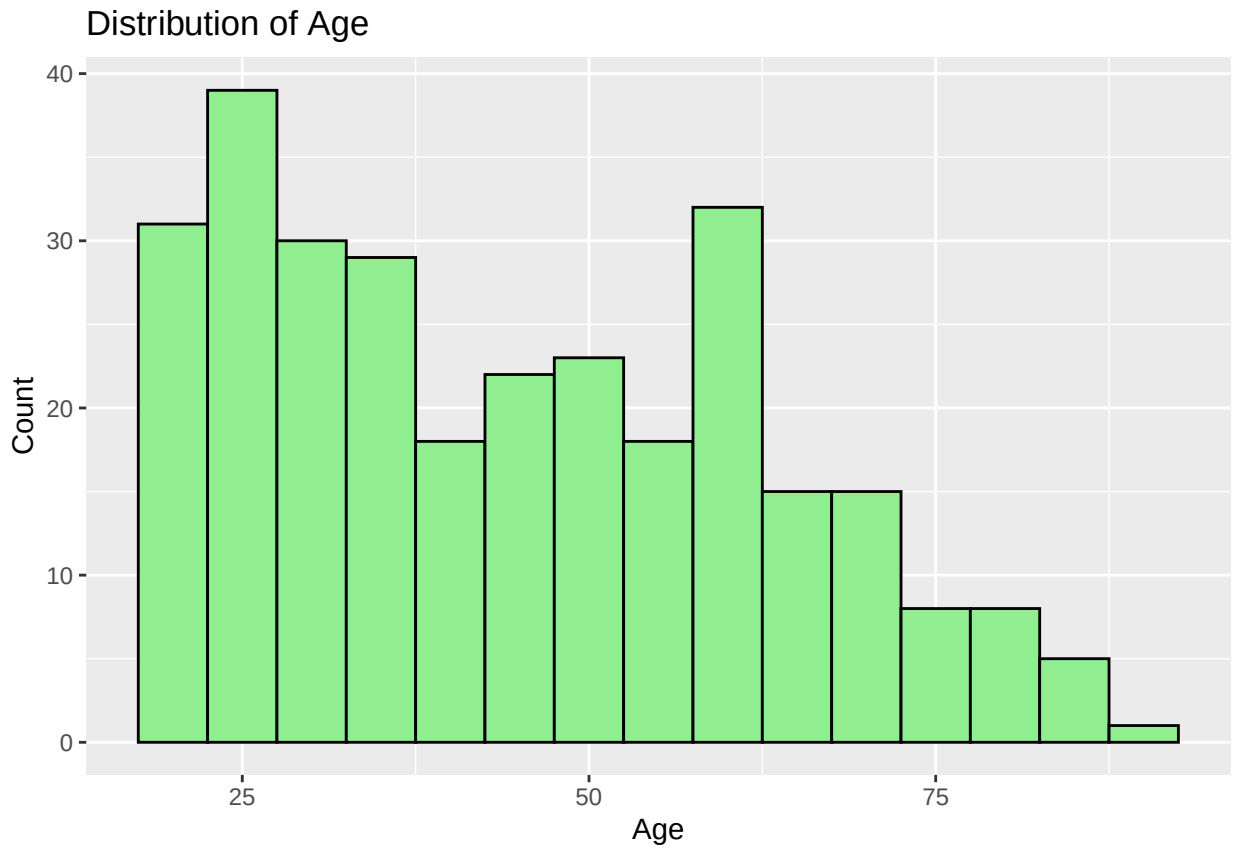
It looks like only a few people scored high and many folks scored in the lower 1-15 range.

Now let's take look at age:

```
summary(depress$AGE)
```

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

```
ggplot(depress, aes(x = AGE)) +
  geom_histogram(binwidth = 5, fill = "lightgreen", color = "black") +
  ggtitle("Distribution of Age") +
  xlab("Age") +
  ylab("Count")
```

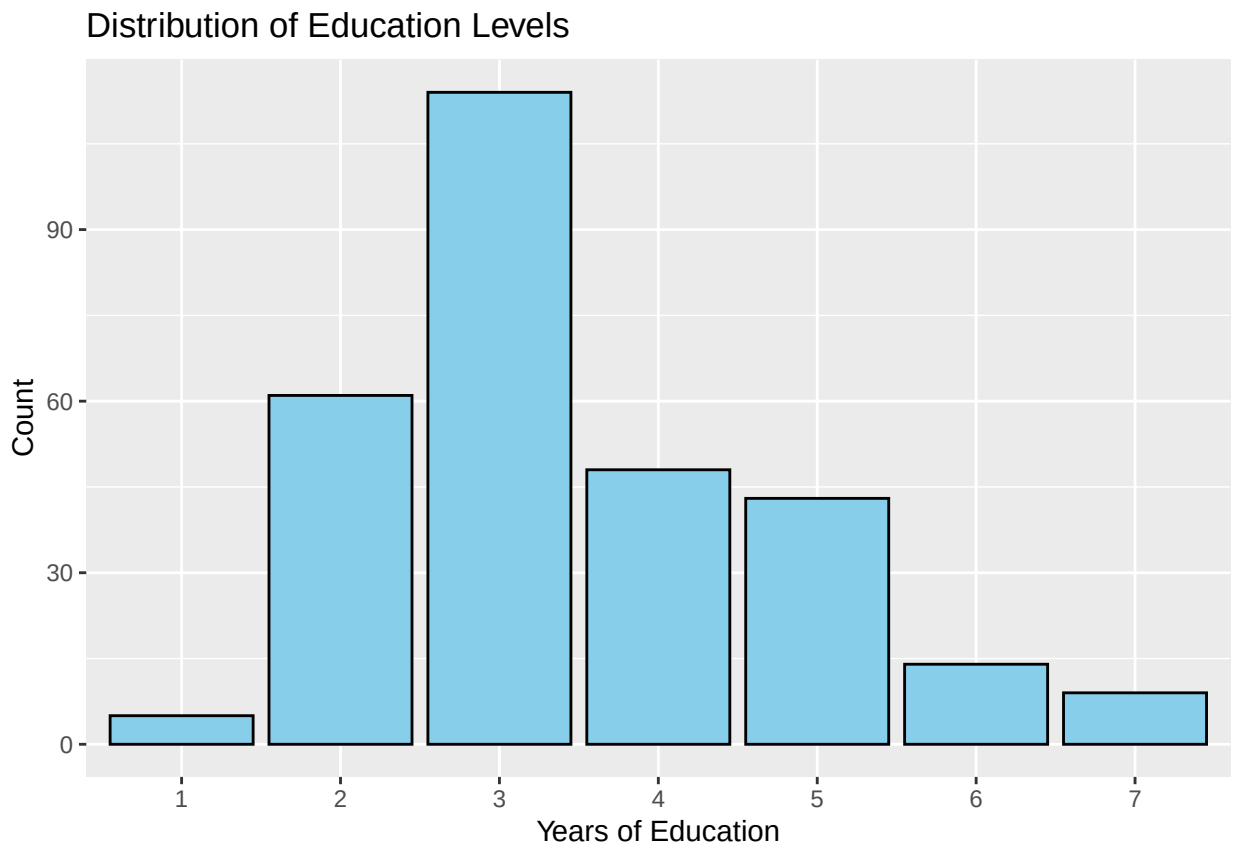


Here's the education spread:

```
table(depress$EDUCAT)
```

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

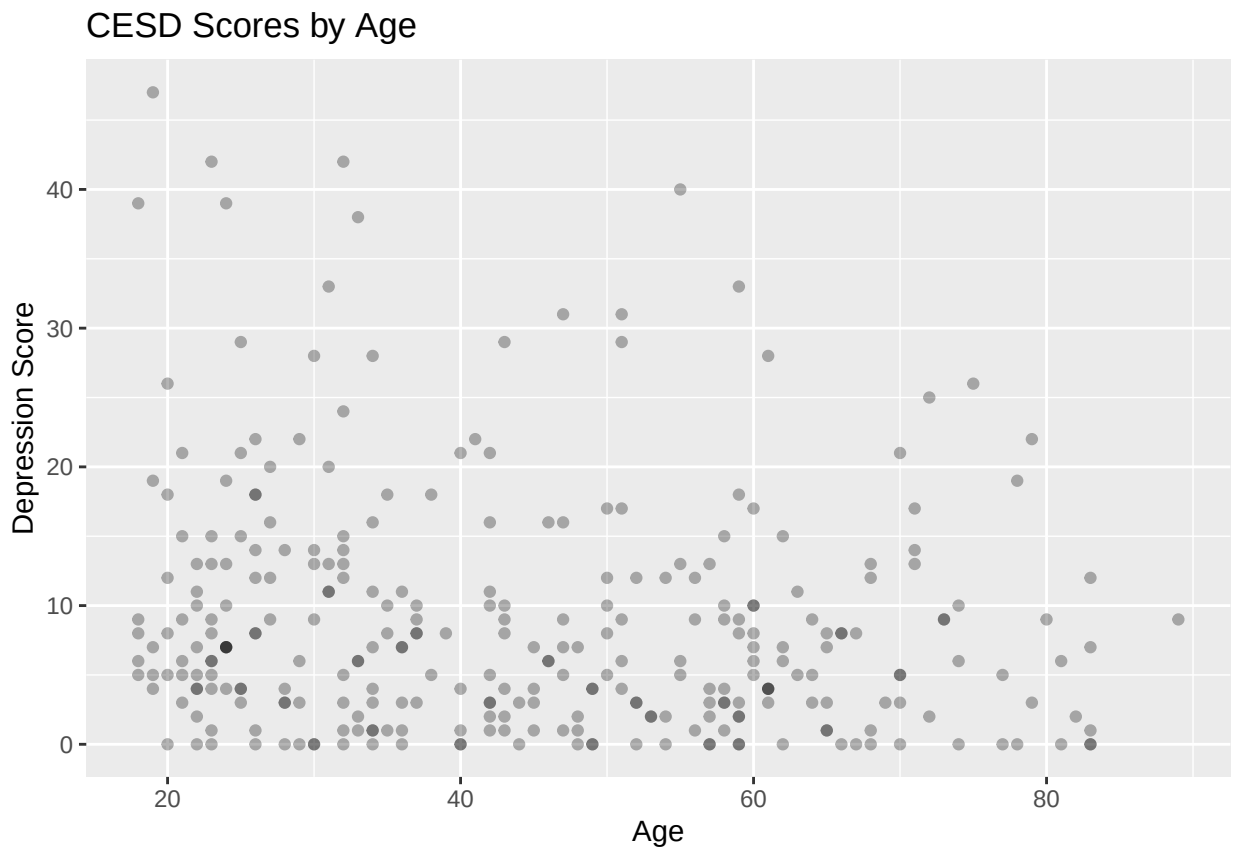
```
ggplot(depress, aes(x = factor(EDUCAT))) +
  geom_bar(fill = "skyblue", color = "black") +
  ggtitle("Distribution of Education Levels") +
  xlab("Years of Education") +
  ylab("Count")
```



Bivariate Exploration

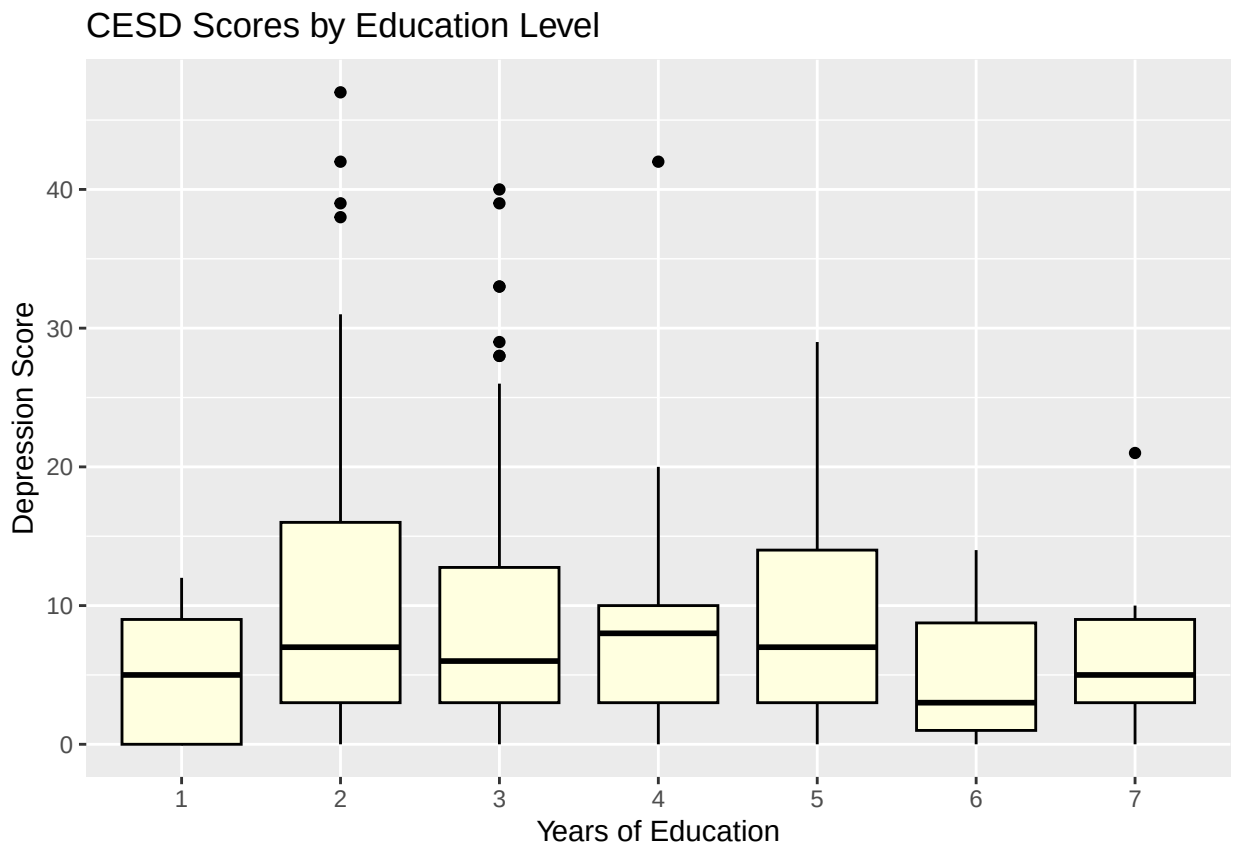
Let's check out CESD by AGE using a scatterplot:

```
ggplot(depress, aes(x = AGE, y = CESD)) +
  geom_point(alpha = 0.3) +
  ggtitle("CESD Scores by Age") +
  xlab("Age") +
  ylab("Depression Score")
```



And CESD scores by EDUCATION level:

```
ggplot(depress, aes(x = factor(EDUCAT), y = CESD)) +
  geom_boxplot(fill = "lightyellow", color = "black") +
  ggtitle("CESD Scores by Education Level") +
  xlab("Years of Education") +
  ylab("Depression Score")
```



It appears higher depressive scores are corresponding with lower education numbers.

Let's check out some summary stats for CESD by EDUCATION level:

```
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
```

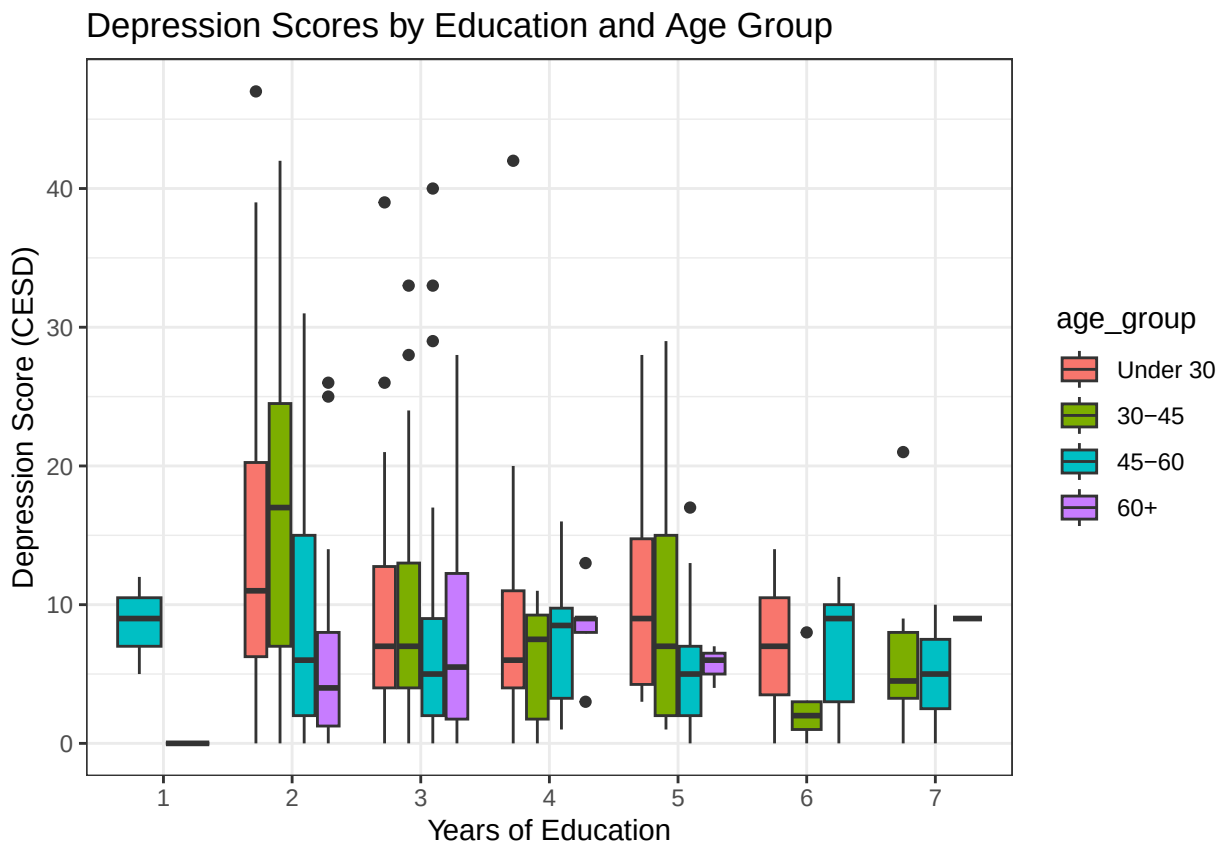
```
depress %>%
  group_by(EDUCAT) %>%
  summarise(
    mean_CESD = mean(CESD, na.rm = TRUE),
    sd_CESD = sd(CESD, na.rm = TRUE),
    count = n()
  )
```

```
## # A tibble: 7 x 4
##   EDUCAT mean_CESD sd_CESD count
##    <int>     <dbl>   <dbl> <int>
## 1      1      5.2     5.36     5
## 2      2     10.8    11.7     61
## 3      3      8.97    8.64   114
## 4      4      7.83    6.97    48
## 5      5      9.33    7.75    43
## 6      6      4.71    4.86    14
## 7      7      6.78    6.51     9
```

Here's one last plot to wrap it all together. This graph shows CESD depression scores by both age group and education level.

```
depress <- depress %>%
  mutate(age_group = cut(AGE, breaks = c(0, 30, 45, 60, 100),
                         labels = c("Under 30", "30-45", "45-60", "60+")))

ggplot(depress, aes(x = factor(EDUCAT), y = CESD, fill = age_group)) +
  geom_boxplot() +
  ggtitle("Depression Scores by Education and Age Group") +
  xlab("Years of Education") +
  ylab("Depression Score (CESD)") +
  theme_bw()
```



Conclusion

In this project, I looked at whether age and education might relate to how depressed someone feels (based on CESD scores).

For education, there wasn't a very clear trend — having more years of school didn't always mean lower depression. But people with lower education had more ups and downs in their scores.

With age, the scores were kind of all over, though younger people sometimes had slightly higher CESD scores.

The last plot with both age group and education showed that younger people with less education tended to have higher and more spread out depression scores. Older adults were a bit more steady, especially those with more education.

This was just a first look, but it gave me a better idea of what groups might be feeling more down — and it's something that could be worth exploring further.

I hope you had as much fun as I did!