MCH project

2023-09-15

##Introduction: The dataset I will be using is the depression set. Will be exploring variables sex, religion, chronic illness. My research question is there a correlation between age and education for those who suffer with depression.

```
depress<- read.table( "C:/Users/mchoe/OneDrive/Documents/Chico/Fall 2023/Math 130/Data/depress_081217.t
library(ggplot2)
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(forcats)
```

##Univariate Exploration

```
head(depress)
```

```
##   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
```

###Variable 1 Age

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

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

```
mean(depress$age)
```

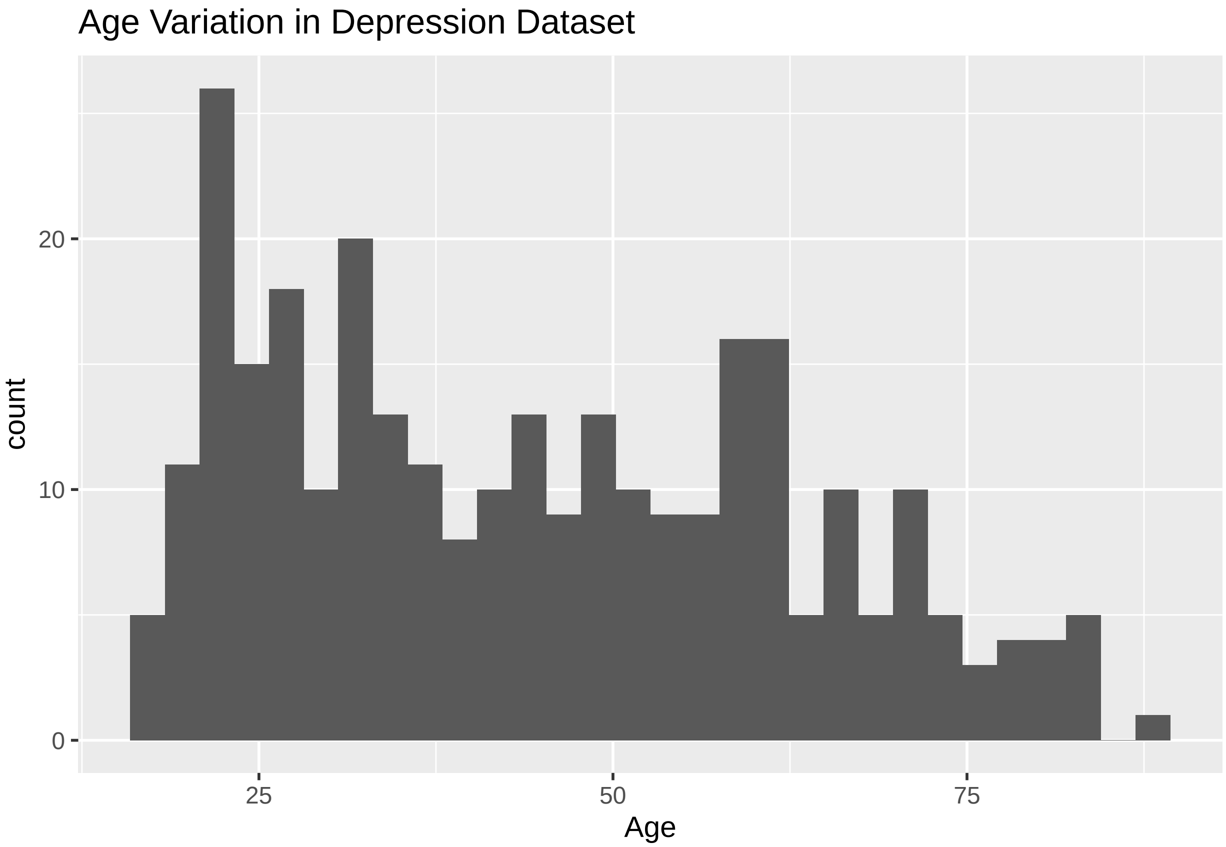
```
## [1] 44.41497
```

```
sd(depress$age)
```

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

```
ggplot(depress, aes(x=age, fill=age)) +  geom_histogram(bins=30)+ xlab("Age")+ggtitle("Age Variation in
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
```



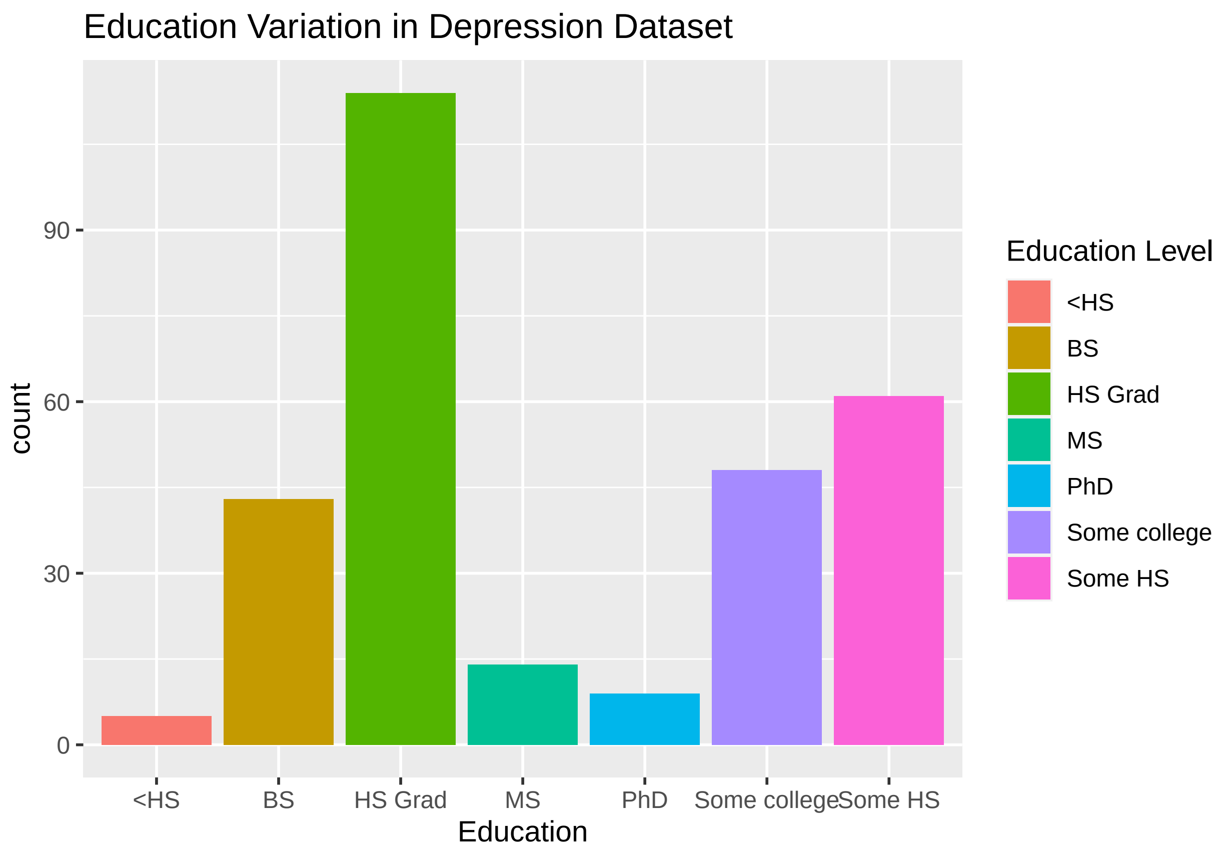
The youngest age in the dataset is 18 years old and the oldest is 89 years old. The mean age is 44 with a standard deviation of 18. The standard deviation is large because there is a wide range of age for people who experience depression in the dataset.

###Variable 2, Education

```
depress$Education_Level <-depress$educat
table(depress$educat)
```

```
##
##          <HS            BS       HS Grad            MS           PhD Some college
##            5            43           114            14             9           48
##      Some HS
##           61
```

```
ggplot(depress, aes(x=educat, fill=educat)) +  geom_bar()+xlab("Education")+ggtitle("Education Variatio
```



Looking at a table of the education variable we can see that the majority of participants are high school graduates with 114 out of 294. Less than 5 participants had less than high school education. There may be a gap in the data here as those with less than high school education probably suffer with depression but they may not be available for the survey's data collection. The second biggest group of education was those who received some high school education, with 61 out of 294 participants. This means they at least completed some high school. The third biggest group of education was those who attended some college, 48 out of 294 participants did. Slightly below this at 43 out of 294 is those who completed college education for a bachelors of science. 14 participants received a masters in science and 9 participants received a PhD.

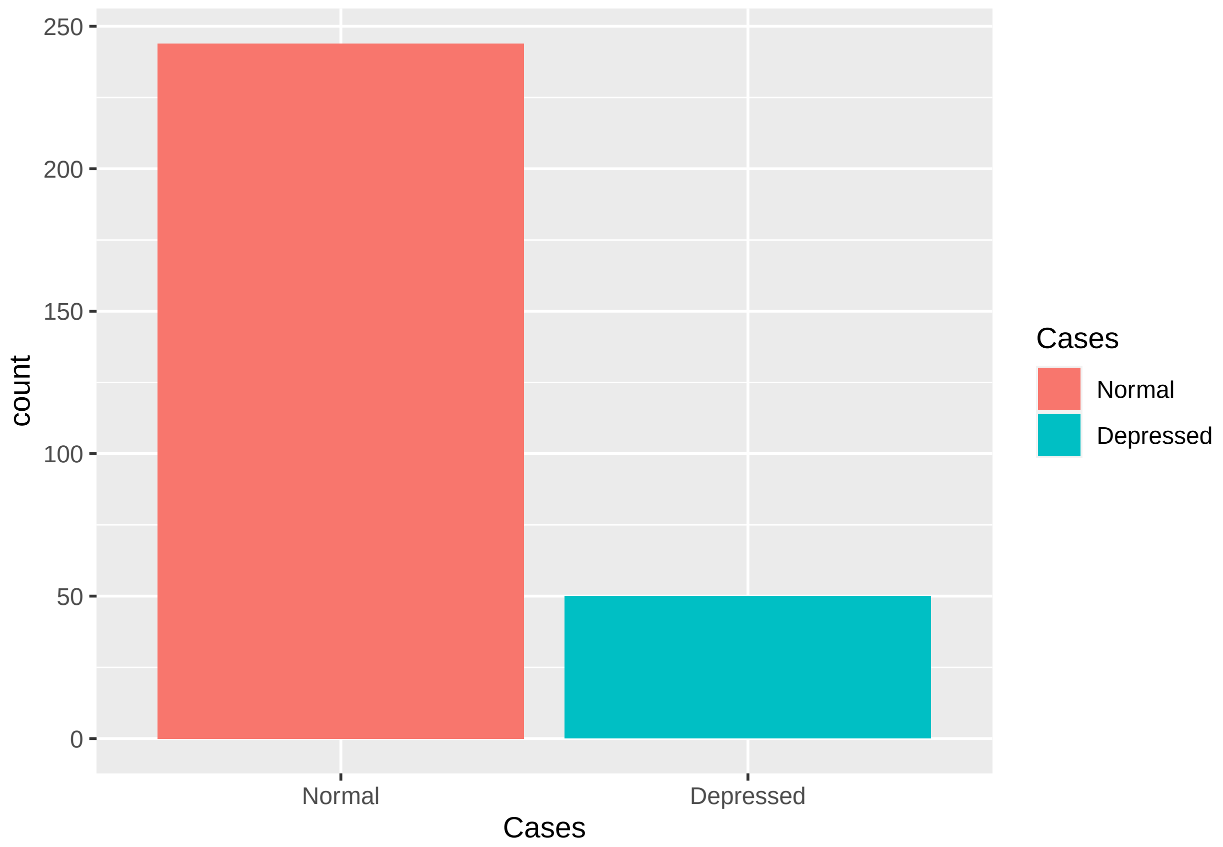
###Variable 3, Cases; Normal or Depressed

```
depress$cases<- as.character(depress$cases)

depress$casesforcats <- fct_recode(depress$cases, "Normal" = "0", "Depressed" = "1")
depress$Cases<-depress$casesforcats
table(depress$casesforcats)
```

```
##
##    Normal Depressed
##       244        50
```

```
ggplot(depress, aes(x=casesforcats, fill=casesforcats)) +  geom_bar()+xlab("Cases")+scale_fill_discrete
```



Out of 294 participants, 244 were normal (not experiencing depression), and 50 were depressed. This means the data will be skewed towards a correlation between age and education for those normal. The cases variable will be our baseline to compare the correlation between age and education for those who suffer with depression.

##Bivariate Exploration

###Age vs Education by Cases

```
by_educat<-group_by(depress, Cases,Education_Level)
summarise(by_educat,avg_age=round(mean(age, na.rm=TRUE), digits=2) )
```

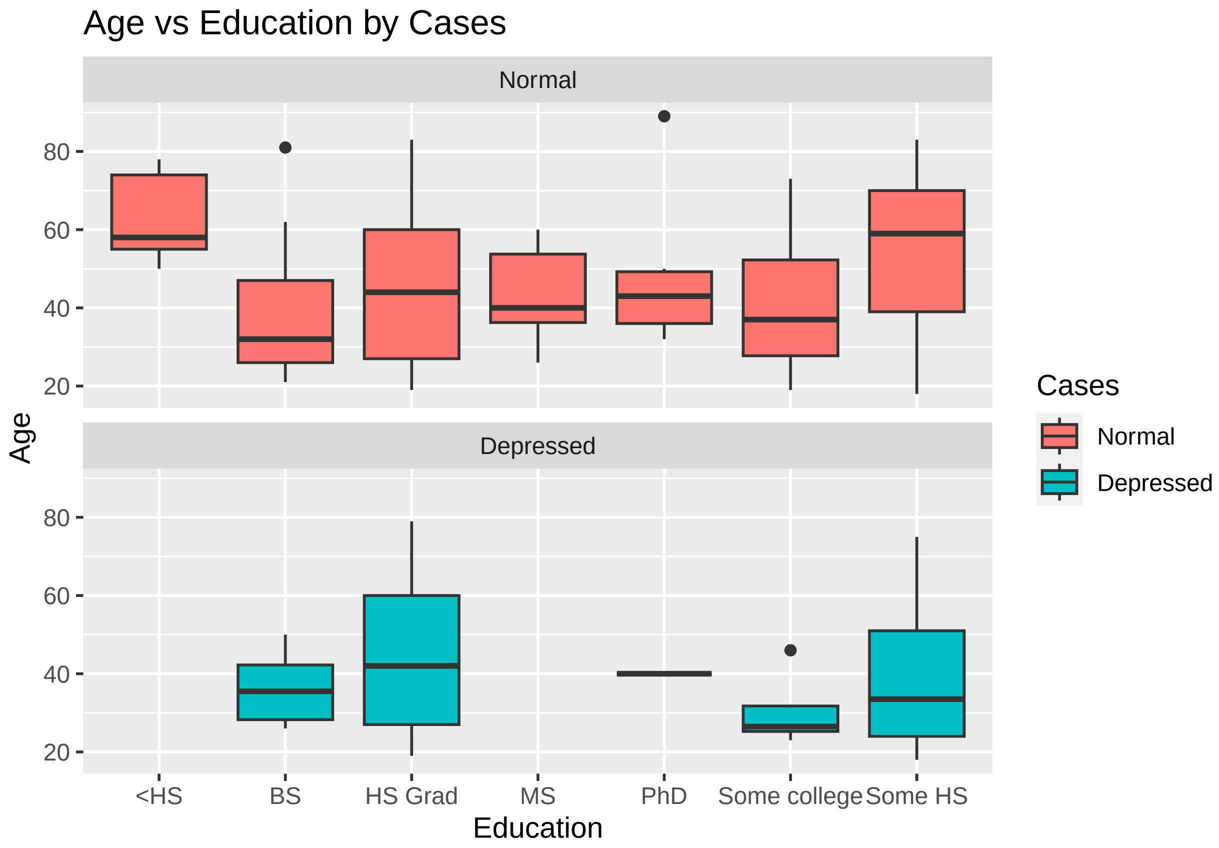
```
## ‘summarise()‘ has grouped output by ’Cases’. You can override using the
## ‘.groups‘ argument.
```

```
## # A tibble: 12 x 3
## # Groups:   Cases [2]
##    Cases     Education_Level avg_age
##    <fct>     <chr>             <dbl>
##  1 Normal    <HS                63
##  2 Normal    BS                 37.5
##  3 Normal    HS Grad            44.9
##  4 Normal    MS                 43
##  5 Normal    PhD                47.2
##  6 Normal    Some HS            55.4
##  7 Normal    Some college       40.0
##  8 Depressed BS                 35.9
##  9 Depressed HS Grad            45.1
## 10 Depressed PhD                40
## 11 Depressed Some HS            38.9
## 12 Depressed Some college       30.5
```

```
table(depress$casesforcats, depress$educat)
```

```
##            
##              <HS BS HS Grad MS PhD Some college Some HS
##    Normal      5 35      93 14   8           44      45
##    Depressed   0  8      21  0   1            4      16
```

```
ggplot(depress, aes(x=educat, y=age, fill=casesforcats)) + geom_boxplot()+xlab("Education")+ylab("Age")
```



```
depress$Cases<-depress$casesforcats
```

####Grouped Summary Statistics Analysis There are seven total education levels, there are "Normal" participants in each level. The average age's range from 40 to 63 years old. The "Depressed" participants are only in five out of 7 levels, they are not in the levels as follows; less than high school, masters degree in science. The average age's range from 30 to 45 years old. We can see that the age range for depressed participants is younger with some overlap at maximum age for depressed with the minimum age for normal.

####Bivariate Histogram Analysis Those who graduated form high school have around equal amounts of normal and depressed participants. For both the region from lower to upper quartile ranges with those from age ~25 to ~60 years old.Those who had less than high school education were all normal, keeping in mind there were only 5 out of 294 participants in this category, this is not statistically significant. Similar occurrences are shown with the categories who achieved PhD and masters of science, which were respectively, 9 and 14. Out of those who received a bachelors in science, there is a greater percentage of those normal over those depressed. For those depressed the region from lower to upper quartile ranges from ~25 to ~40 years old. Compared to those normal the region from lower to upper quartile ranges from ~25 to ~50 years of age. For those with some high school education there were more younger people depressed versus more older people who were normal(not depressed). The depressed ranged from ~25 to ~50 years old versus the normal ranging from ~40 to ~70 years old. A similar trend is observed for those who have some college education, there are more younger people depressed than older. Those who have some college education but are depressed the region from lower to upper quartile ranges with those from ~25 to ~30 years old. While in the same category, those who are normal(not depressed) the region from lower to upper quartile range from ~30 to ~50 years old.

##Conclusion Those who graduated from high school have an equal chance of being depressed or normal(not depressed). The education categories of some high school education, PhD and masters of science had small populations to survey. This means these categories cannot be analyzed with any statistical significance, as there is a gap in the data. It is impossible for little to none participants who receive some high school education, a PhD or a masters of science, to experience little to none depression. We can say the age range for the level of bachelors of science education, the normal category is older but not younger than the depressed category. We can say that for those who received some high school education, people who are younger experience more depression than people who are older. Some reasoning behind this can be younger people are more impressionable to society ideals, in which one is supposed to get and education and without it you are "failing at life". When one grows older they have pushed past these ideals and have already developed a living, and possibly a family, leading to confidence in ones ability and being normal(not depressed). We can apply the same concept to with some college education, for the reasoning of why younger participants experience more depression than older participants. When one has set out to accomplish something and fail to finish that goal that can put a lot of pressure on one's mental health. We can say that the average age's of depressed participants is lower than the average age of normal participants. With some overlap between the maximum average age for depressed and the minimum average age for normal. Based on my research question the age correlation with education for those depressed is that younger participants experience more depression than older participants.