Student Mental Health Data Visualization

Melvis Onoriode

2022-05-19

```
health <- read.csv("Student Mental health.csv")
```

Data Cleaning

```
names(health) <- c('Timestamp', 'Gender', 'Age', 'Course', 'Year', 'CGPA', 'Married', 'D
epression', 'Anxiety', 'Panic_Attack', 'Treatment')
indx <- apply(health, 2, function(x) any(is.na(x)))
indx
```

```
##    Timestamp       Gender          Age       Course         Year         CGPA
##        FALSE        FALSE         TRUE        FALSE        FALSE        FALSE
##      Married   Depression      Anxiety Panic_Attack    Treatment
##        FALSE        FALSE        FALSE        FALSE        FALSE
```

```
# There is missing data in our AGE column
which(is.na(health$Age))
```

```
## [1] 44
```

```
# observation 44 -> since it is only one individual with missing data, we'd look to repl
ace the missing with the median age of this group to keep the numbers clean rather than
 using a decimal as a mean.
health[44,'Age'] <- median(health$Age, na.rm =T)
```

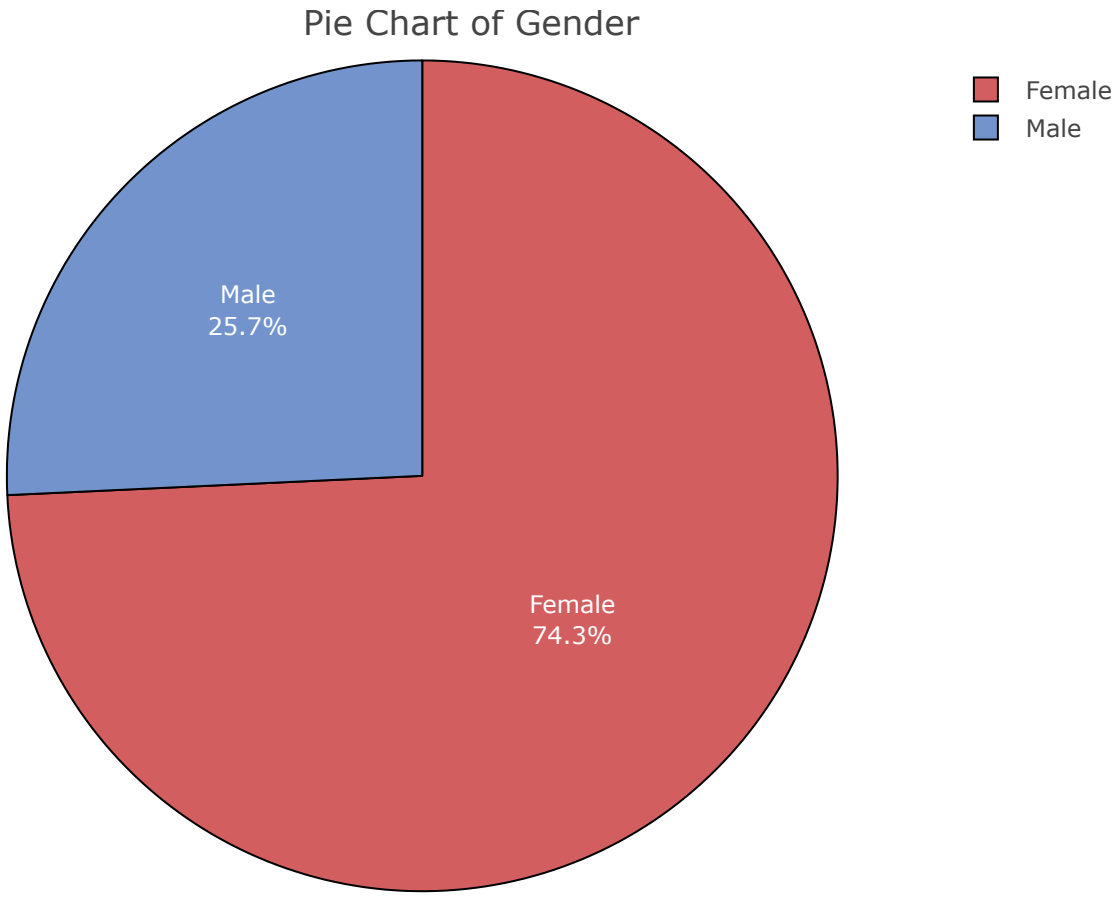
Data Visualization (Overview)

Gender Distribution

```
Health_SummaryStat <- health %>%
  group_by(Gender) %>%
  summarise(count = n(),
            percentage = round((n()/ nrow(health)), digits = 4))
Health_SummaryStat
```

```
## # A tibble: 2 × 3
##   Gender count percentage
##   <chr>  <int>      <dbl>
## 1 Female    75      0.743
## 2 Male      26      0.257
```

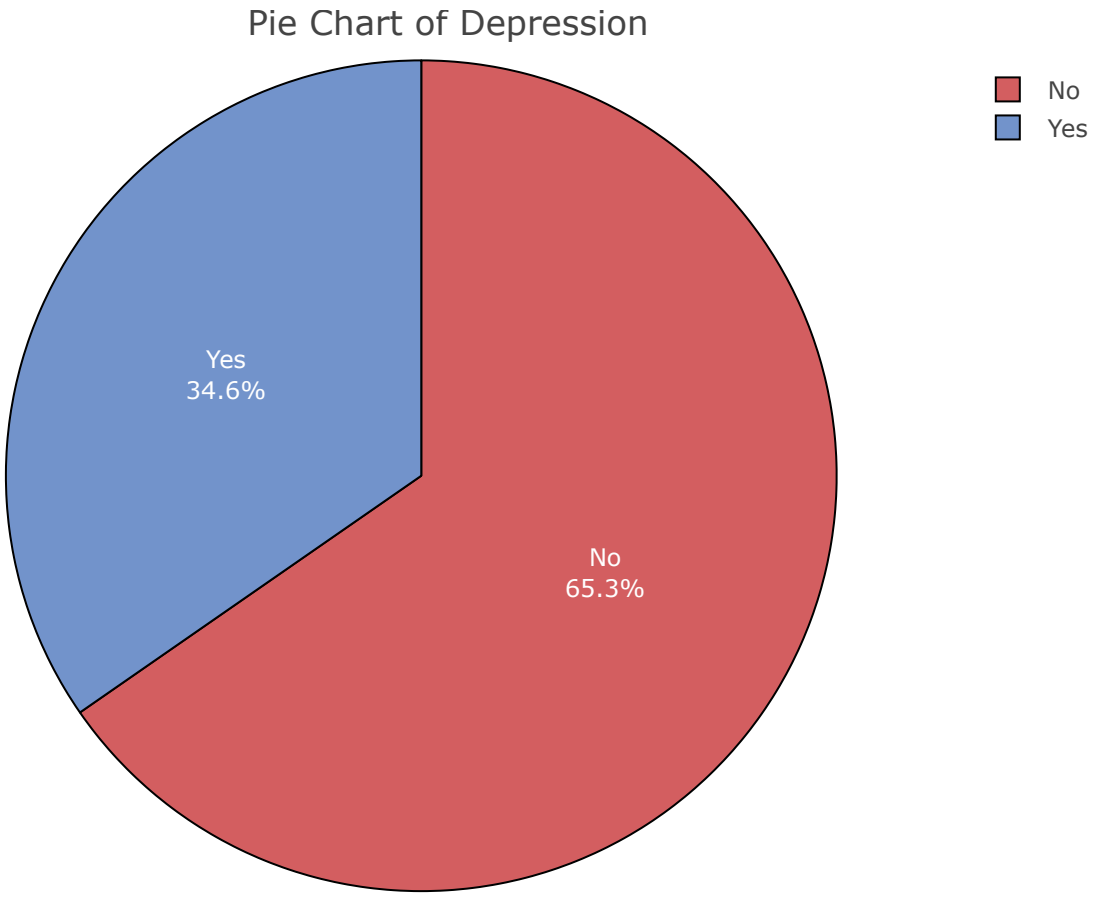
```
colors <- c('rgb(211,94,96)','rgb(114,147,203)')
Gender_PieChart <- plot_ly(data = Health_SummaryStat, labels = ~Gender, values = ~percen
tage,
                type = 'pie', sort = F,
                textposition = 'inside',
                textinfo = 'label+percent',
                insidetextfont = list(color = 'White'),
                hoverinfo = 'text',
                text = ~count,
                marker = list(colors = colors,
                line = list(color = 'Black', width = 1)),
                showlegend = TRUE)
Gender_PieChart <- Gender_PieChart %>% layout(title = 'Pie Chart of Gender')
Gender_PieChart
```



74.3% of observations were female compared to 25.7%

Depression

```
Health_SummaryStat2 <- health %>%
  group_by(Depression) %>%
  summarise(count = n(),
            percentage = round((n()/ nrow(health)), digits = 4))
Depression_PieChart <- plot_ly(data = Health_SummaryStat2, labels = ~Depression, values
 = ~percentage,
                type = 'pie', sort = F,
                textposition = 'inside',
                textinfo = 'label+percent',
                insidetextfont = list(color = 'White'),
                hoverinfo = 'text',
                text = ~count,
                marker = list(colors = colors,
                line = list(color = 'Black', width = 1)),
                showlegend = TRUE)
Depression_PieChart %>% layout(title = 'Pie Chart of Depression')
```



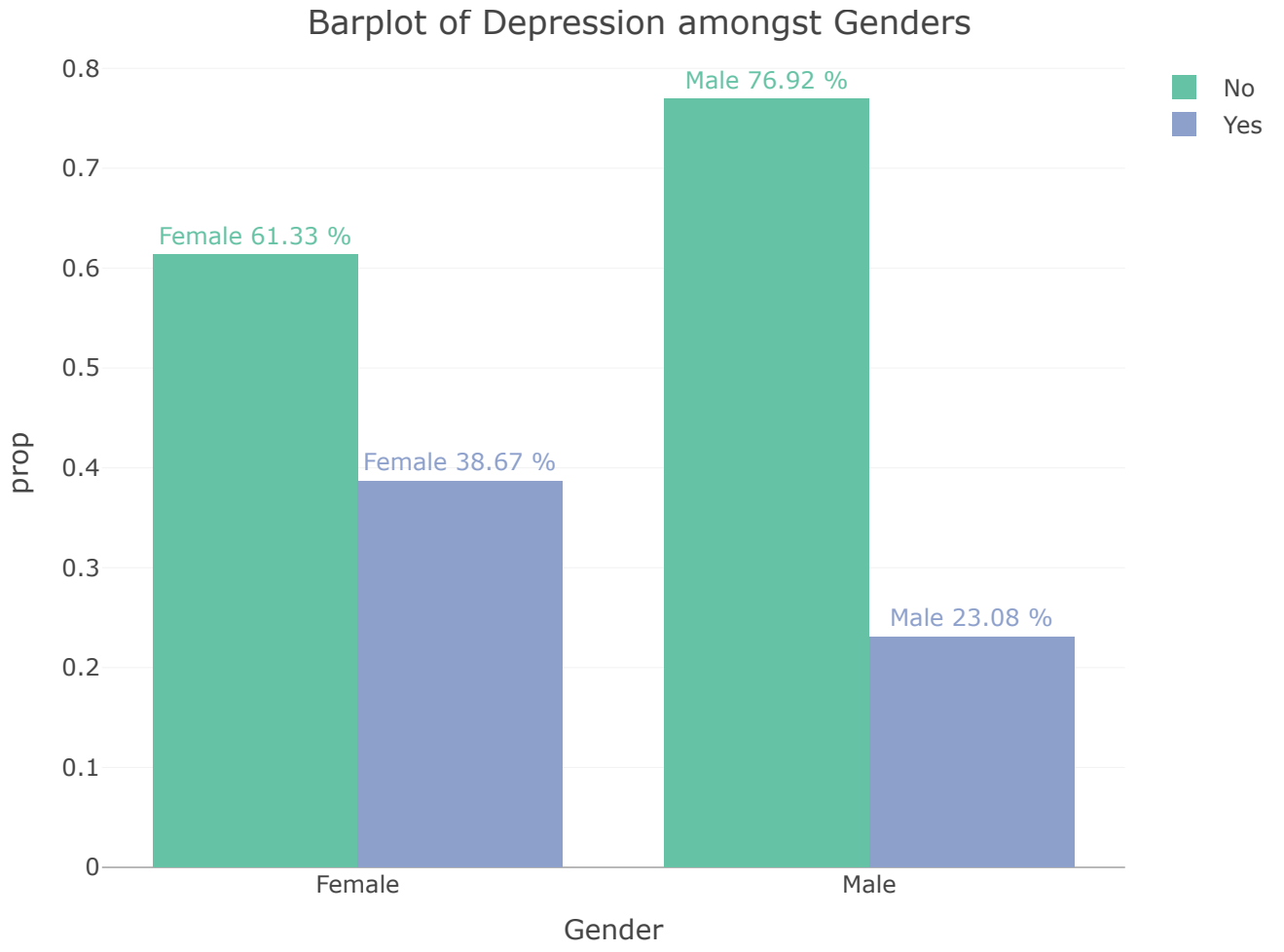
34.6 % of sample had depression compared to an average 5.0% rate amongst adults in the population.

Depression vs Gender

```
health %>%
  count(Gender, Depression, sort = F) %>%
  group_by(Gender) %>%
  mutate(prop = round((n / sum(n)),digits = 4)) %>%
  plot_ly(x = ~Gender, y=~prop, color = ~Depression, type = "bar",
          text = ~paste(Gender, prop*100 ,'%'),
          textposition = 'outside') %>%
    layout(barmode = 'Stacked',
           title = 'Barplot of Depression amongst Genders')
```

```
## Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning r
equested palette with 3 different levels

## Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning r
equested palette with 3 different levels
```



Females tended to on average have more cases of depression compared to male. (38.67% vs 23.08%)

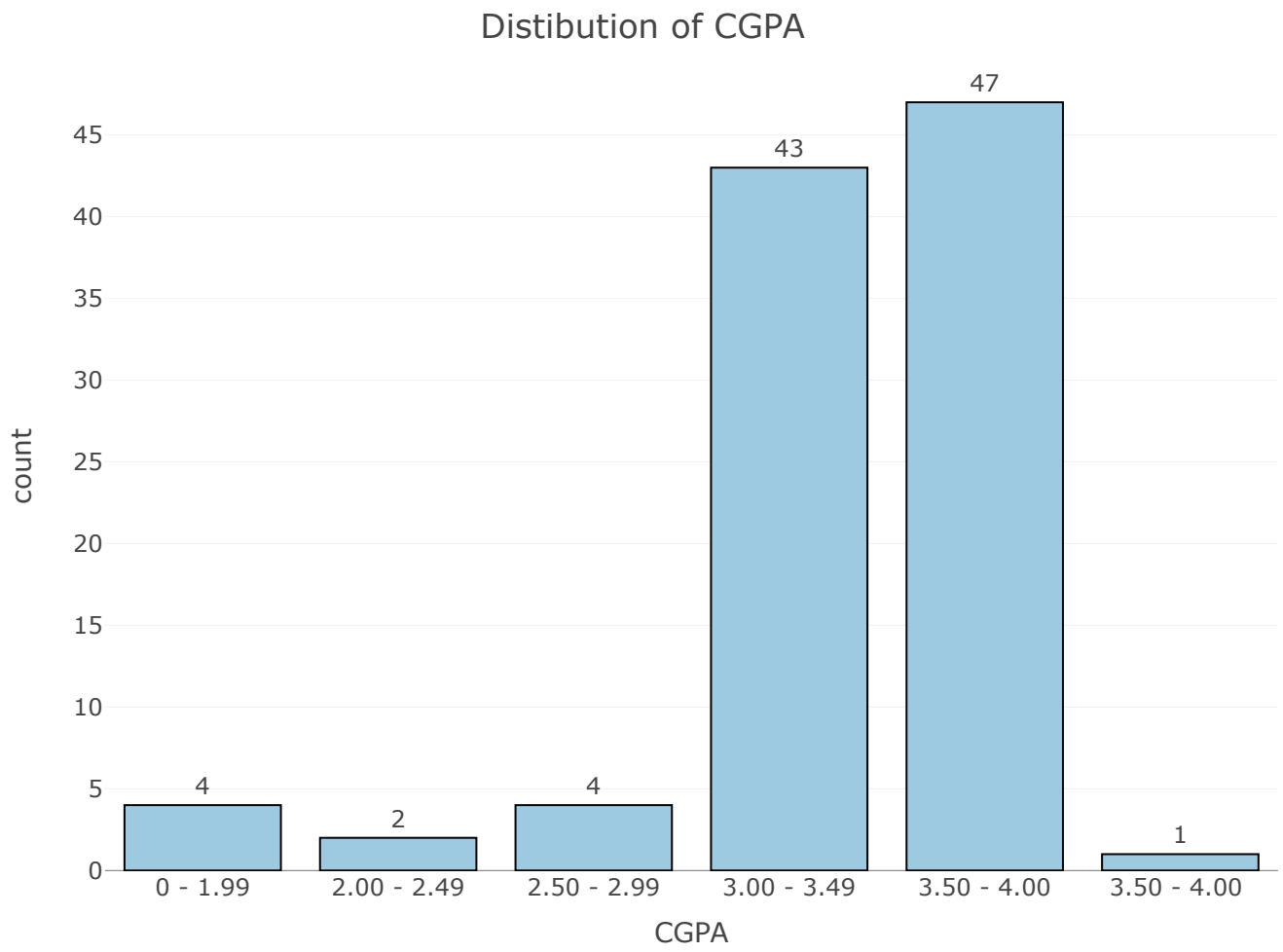
GPA Distribution

Looking to see if there is a positive association between GPA and Depression (i.e higher CGPA -> see higher proportion depressed? Just initial thoughts)

```
health$CGPA <- as.factor(health$CGPA)
levels(health$CGPA)
```

```
## [1] "0 - 1.99"     "2.00 - 2.49"  "2.50 - 2.99"  "3.00 - 3.49"  "3.50 - 4.00"
## [6] "3.50 - 4.00 "
```

```
# Levels are in order ~
health %>%
  group_by(CGPA)%>%
  summarize(count = n()) %>%
  plot_ly(x =~CGPA, y=~count, type = 'bar',
        text = ~count,
        textposition = 'outside',
        marker = list(color = 'rgb(158,202,225)',
          line = list(color = 'black',
                 width = 1.0))) %>%
  layout(title = 'Distibution of CGPA')
```

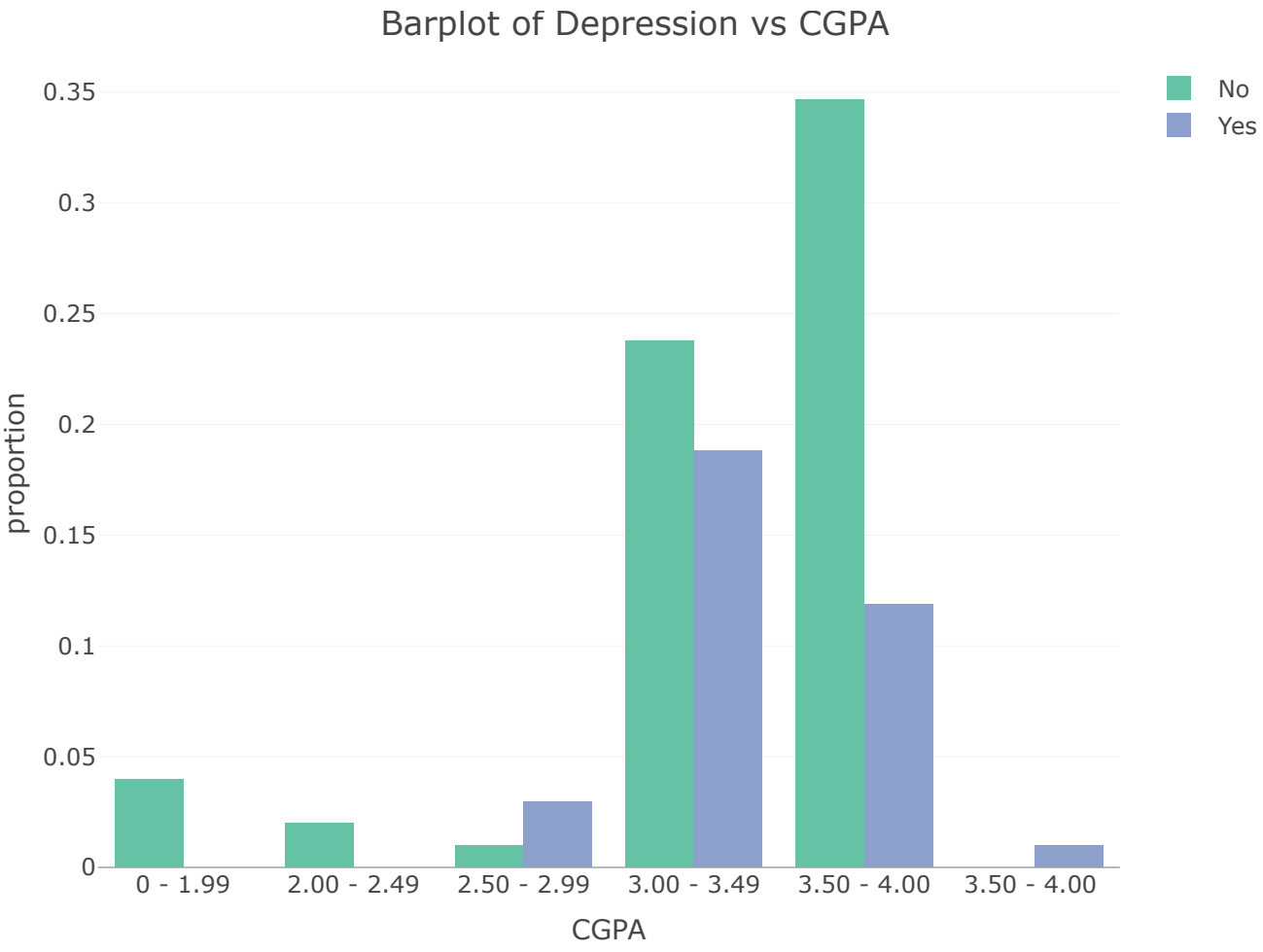


CGPA vs Depression

```
health %>%
  count(CGPA, Depression, sort = F) %>%
  mutate(proportion = round((n/sum(n)),digits=4)) %>%
  plot_ly(x =~CGPA, y=~proportion, color = ~Depression, type = 'bar') %>%
  layout(barmode = 'Group',
         title = 'Barplot of Depression vs CGPA')
```

```
## Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning r
equested palette with 3 different levels

## Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning r
equested palette with 3 different levels
```



We aren't able to see any clear patterns with CGPA and Depression, but we do notice a non-monotonic increase in depression rate as CGPA increases. This might be due to inadequate group sizes, and if given a more percise CGPA value, we would of been able to get a clearer picture. But with this data set we do notice that CGPA has a slight positive associate with depression rate.

Major Courses

Since there are many different coures we'd look at the top ones and see if the depression rates vary amongst them.

```
health %>%
  group_by(Course) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  filter(count >2)
```

```
## # A tibble: 5 × 2
##   Course             count
##   <chr>              <int>
## 1 BCS                   18
## 2 Engineering           17
## 3 BIT                   10
## 4 Biomedical science     4
## 5 KOE                    4
```

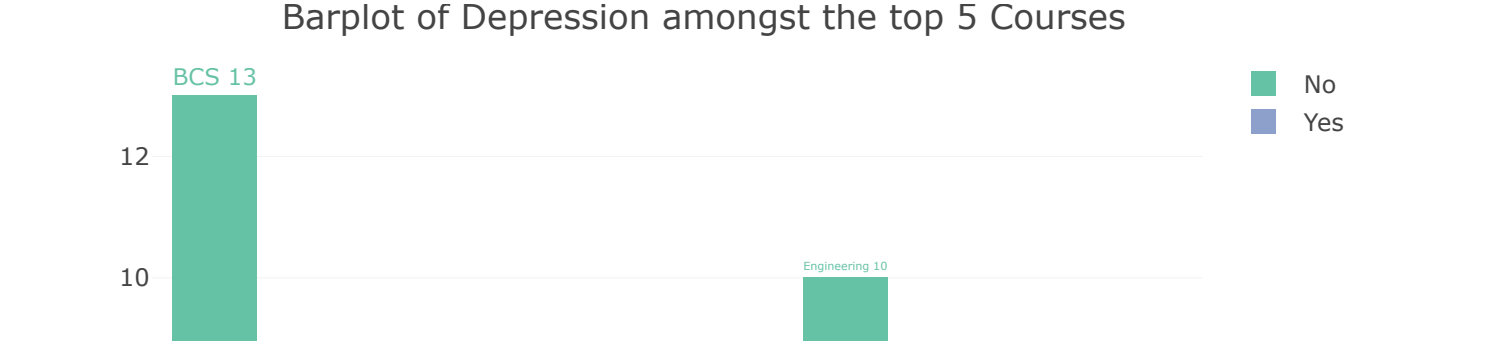
```
## Lets look at coures BCS, Engineering, BIT, Biomedical science , and KOE
```

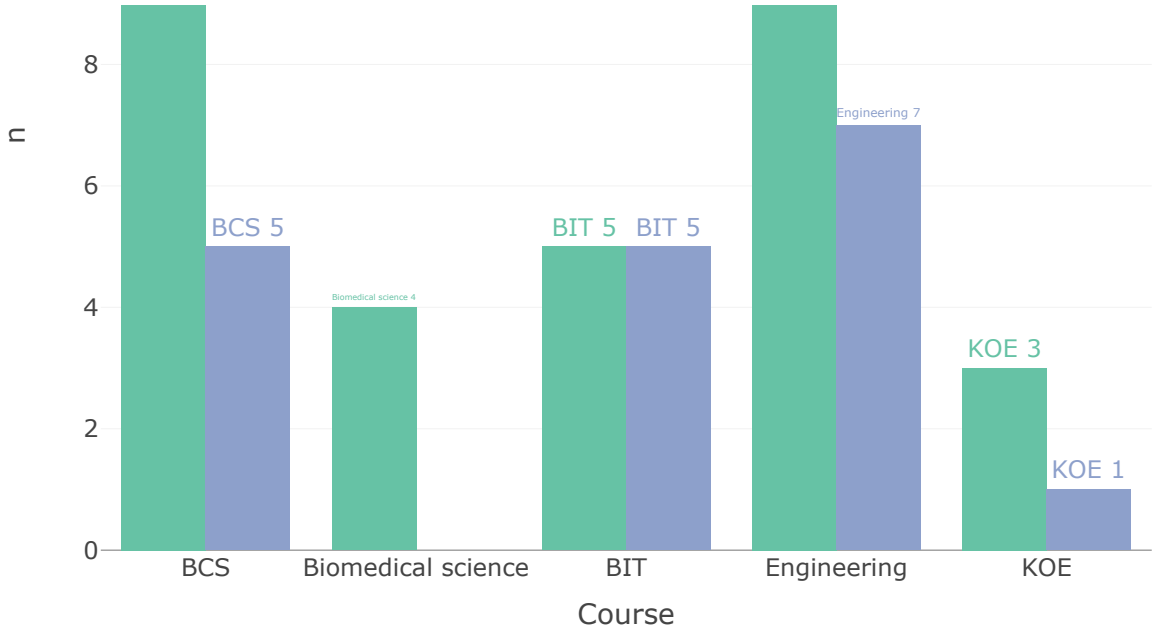
Top Major Courses vs Depression

```
health %>%
  filter(grepl('BIT|KOE|BCS|Engineering|Biomedical science', Course)) %>%
  count(Course, Depression, sort = T) %>%
  group_by(Course) %>%
  mutate(prop = round((n / sum(n)),digits = 4)) %>%
  plot_ly(x = ~Course, y=~n, color = ~Depression, type = "bar",
          text = ~paste(Course, n),
          textposition = 'outside') %>%
  layout(barmode = 'Stacked',
         title = 'Barplot of Depression amongst the top 5 Courses')
```

```
## Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning r
equested palette with 3 different levels

## Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning r
equested palette with 3 different levels
```





Engineering students seem to have the most cases of depression, and surprisingly 0 came out of the Bio medical group. Group sizes too small to make any conclusive associations.