Survery on Mental Health Across Programs of Study in University

Mental Health Maniacs - Deanna King, Jim Moroney, Pooja Patel, Christopher Wilhite

2022-11-22

Packages

In order to better interpret the dataset, we utilize the `pander` package for table creation and manipulation. Likewise, we use `insert other packages` for `insert reasoning`.

```
library(pander)
```

Student Mental Health Data

All data required for this interpretation was obtained from the International Islamic University in Malaysia. This dataset is publicly available via Kaggle, and contains the following features:

- `Timestamp` - time at which the survey was completed
- `Choose your gender` - gender (male or female) of the participant
- `Age` - age of the participant at the time of survey completion
- `What is your course?` - program in which the participant is majoring
- `Your current year of Study` - how many years the participant has attended university
- `What is your CGPA?` - current grade point average (or the ratio of grade points earned to grade points attempted), calculated on a 0.0-4.0 scale
- `Marital Status` - describes whether or not the participant is married
- `Do you have Depression?` - states whether or not the participant has depression
- `Do you have Anxiety?` - states whether or not the participant has anxiety
- `Do you have Panic attacks?` - states whether or not the participant experiences panic attacks
- `Did you seek any specialist for a treatment?` - states whether or not the participant sought professional treatment for any mental health concerns

```
studentData_df <- read.csv(file="./StudentMentalHealth.csv")
#str(studentData_df)
summary(studentData_df)
```

```
##   Timestamp         Choose.your.gender      Age        What.is.your.course.
##  Length:101         Length:101         Min.   :18.00   Length:101
##  Class :character   Class :character   1st Qu.:18.00   Class :character
##  Mode  :character   Mode  :character   Median :19.00   Mode  :character
##                                        Mean   :20.53
##                                        3rd Qu.:23.00
##                                        Max.   :24.00
##                                        NA's   :1
##  Your.current.year.of.Study What.is.your.CGPA. Marital.status
```

```
##  Length:101                 Length:101         Length:101
##  Class :character           Class :character   Class :character
##  Mode  :character           Mode  :character   Mode  :character
##
##
##
##
##  Do.you.have.Depression. Do.you.have.Anxiety. Do.you.have.Panic.attack.
##  Length:101              Length:101           Length:101
##  Class :character        Class :character     Class :character
##  Mode  :character        Mode  :character     Mode  :character
##
##
##
##
##  Did.you.seek.any.specialist.for.a.treatment.
##  Length:101
##  Class :character
##  Mode  :character
##
##
##
##
```

```
#pander(studentData_df)
```

Data Cleaning

We thoroughly examined the data to ensure that no noisy or missing data values were present. More specifically, we ensured that no negative values existed in our numerical attributes (age, year of study, CGPA), and that no missing values were present in any tuple. Upon examination, only one column was found to have any missing data: `Age`. Though it is not particularly robust, we opted to fill in this missing data with a simple mean of the column.

```
studentData_df$Age[is.na(studentData_df$Age)] <- floor(mean(studentData_df$Age, na.rm=TRUE))
```

In order to further facilitate our analysis of this data, we deemed it appropriate to drop the `Timestamp` column, as it provided no relevant information to what we were looking for and seemed to be more of a vanity metric for the circumstances in which the data was originally acquired.

```
studentData_df = subset(studentData_df, select = -c(1))
summary(studentData_df)
```

```
##  Choose.your.gender      Age        What.is.your.course.
##  Length:101         Min.   :18.00   Length:101
##  Class :character   1st Qu.:18.00   Class :character
##  Mode  :character   Median :19.00   Mode  :character
##                     Mean   :20.52
##                     3rd Qu.:23.00
##                     Max.   :24.00
##  Your.current.year.of.Study What.is.your.CGPA. Marital.status
##  Length:101                 Length:101         Length:101
```

```
##  Class :character           Class :character   Class :character
##  Mode  :character           Mode  :character   Mode  :character
##
##
##
##  Do.you.have.Depression. Do.you.have.Anxiety. Do.you.have.Panic.attack.
##  Length:101              Length:101           Length:101
##  Class :character        Class :character     Class :character
##  Mode  :character        Mode  :character     Mode  :character
##
##
##
##  Did.you.seek.any.specialist.for.a.treatment.
##  Length:101
##  Class :character
##  Mode  :character
##
##
##
```

Due to the method in which this survey was conducted, some features of the data were able to be entered in an non-deterministic manner. The column `Your.current.year.of.Study` suffers from this the most, as random capitalization in the responses creates several different "bins" of responses for data that is otherwise meant to be the same. To fix this, we elected to simple cast all characters in this column to an uppercase state to remove any ambiguity.

```
studentData_df$Your.current.year.of.Study <- toupper(studentData_df$Your.current.year.of.Study)
```

Data Wrangling

Renaming Columns

Wrangling for the most part consisted of making the data look more presentable and easier to parse for our exploratory analysis and display purposes. Through some minor idiosyncrasies of the method through which this data was obtained, lengthy and oddly formatted names currently index most of our columns; for these reasons we gave the each column a less verbose name that still unambiguously indicated what data said column held.

- `Choose.your.gender` becomes simply `Gender`
- `Age` - age of the participant at the time of survey completion
- `What.is.your.course.` - is simplified into `Major`
- `Your.current.year.of.Study` - has been summarily shortened to `Year`
- `What.is.your.CGPA.` - similarly shortened to just `GPA`

The following attributes have been shortened to just their respective affects. It is assumed that the names are preceeded by, "is," or, "has," before each condition (i.e. "has Anxiety).

- `Marital.status` becomes `Married`
- `Do.you.have.Depression.` becomes `Depressed`
- `Do.you.have.Anxiety.` becomes `Anxiety`
- `Do.you.have.Panic.attack.` becomes `Panic`
- `Did.you.seek.any.specialist.for.a.treatment.` becomes `Treatment`

With all of these, we were seeking the simplicity of single word names.

```
colnames(studentData_df)[colnames(studentData_df) == 'Choose.your.gender'] <- 'Gender'
colnames(studentData_df)[colnames(studentData_df) == 'What.is.your.course.'] <- 'Major'
colnames(studentData_df)[colnames(studentData_df) == 'Your.current.year.of.Study'] <- 'Year'
colnames(studentData_df)[colnames(studentData_df) == 'What.is.your.CGPA.'] <- 'GPA'
colnames(studentData_df)[colnames(studentData_df) == 'Marital.status'] <- 'Married'
colnames(studentData_df)[colnames(studentData_df) == 'Do.you.have.Depression.'] <- 'Depressed'
colnames(studentData_df)[colnames(studentData_df) == 'Do.you.have.Anxiety.'] <- 'Anxiety'
colnames(studentData_df)[colnames(studentData_df) == 'Do.you.have.Panic.attack.'] <- 'Panic'
colnames(studentData_df)[colnames(studentData_df) == 'Did.you.seek.any.specialist.for.a.treatment.'] <-
```

Categorizing ambiguous data

Given that we are looking for various correlations between Science, Technology, Engineering, and Mathematics (STEM) majors and the various mental health issues that might affect them, the detailed knowledge of what major a student is enrolled in doesn't interest us as data - we only care whether or not it's considered STEM. Unfortunately there is no easy algorithmic way to do this, we had considered using various "sounds-like" libraries and methods, but decided for a dataset this small that it'd be best to just manually build a new column by hand. This included looking for what the responses from the `Major` column in our dataset correlated to and simply populating a new column with "Yes" or "No" before adding it to our dataset. We added this column as `STEM`. We have included the first 5 rows as a sample of the data set, full data set is listed in Appendix A at the end of this document.

```
studentData_df['STEM'] <- c('Yes','No','Yes','No','Yes','Yes','No','Yes','No','No',
                            'Yes','Yes','Yes','Yes','No','Yes','No','No','Yes','Yes',
                            'Yes','Yes','Yes','Yes','Yes','No','Yes','Yes','Yes','Yes',
                            'No','Yes','Yes','Yes','Yes','Yes','No','Yes','No','Yes',
                            'Yes','Yes','No','Yes','No','Yes','Yes','Yes','Yes','Yes',
                            'No','Yes','Yes','Yes','Yes','Yes','No','No','Yes','No',
                            'Yes','Yes','Yes','Yes','Yes','No','Yes','No','No','Yes',
                            'Yes','No','Yes','Yes','Yes','Yes','Yes','No','Yes','Yes',
                            'No','No','Yes','No','Yes','Yes','No','Yes','Yes','Yes',
                            'No','Yes','Yes','No','No','No','Yes','Yes','No','No',
                            'Yes')
pander(studentData_df[1:5,])
```

Table 1: Table continues below

Gender	Age	Major	Year	GPA	Married	Depressed
Female	18	Engineering	YEAR 1	3.00 - 3.49	No	Yes
Male	21	Islamic education	YEAR 2	3.00 - 3.49	No	No
Male	19	BIT	YEAR 1	3.00 - 3.49	No	Yes
Female	22	Laws	YEAR 3	3.00 - 3.49	Yes	Yes
Male	23	Mathemathics	YEAR 4	3.00 - 3.49	No	No

Anxiety	Panic	Treatment	STEM
No	Yes	No	Yes
Yes	No	No	No
Yes	Yes	No	Yes

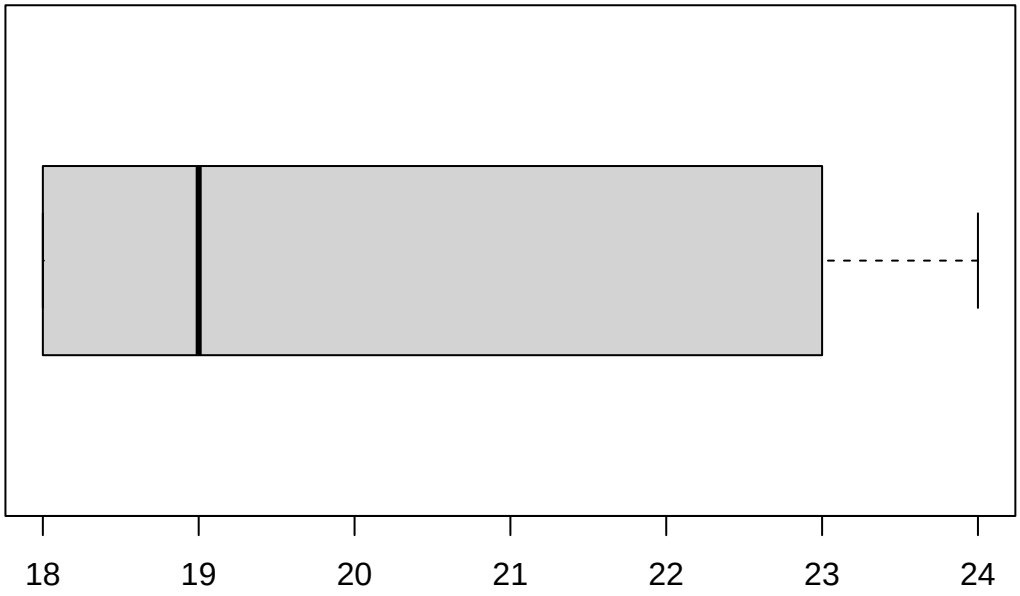
Anxiety	Panic	Treatment	STEM
No	No	No	No
No	No	No	Yes

Exploratory Data Analysis

Age Distribution

The following boxplot shows the distribution of each participant's age. The youngest participants are 18 years old, and the oldest are 24 years old. Given that our mean is 20.5247525, we can be assured that our data is fairly representative of the average college student.

```
boxplot(studentData_df$Age, horizontal = TRUE)
```



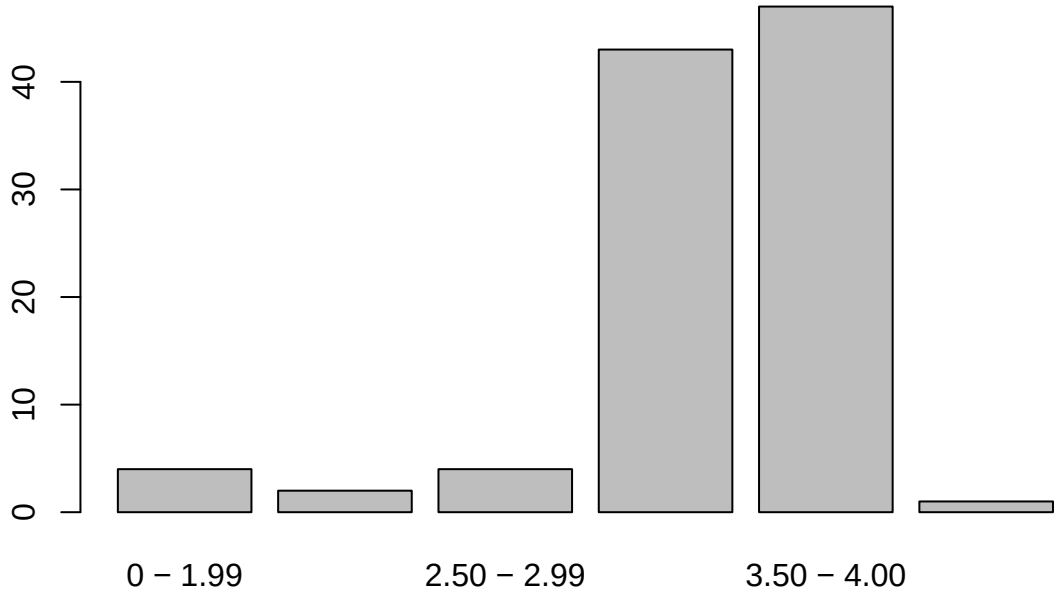

```
summary(studentData_df$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   18.00   19.00   20.52   23.00   24.00
```

CGPA Distribution

The following histogram shows the frequency of each CGPA interval for participants. Most participants have a CGPA within the range of 3.00 - 3.49.

```r
gpaTable <- table(studentData_df$GPA)
barplot(gpaTable)
```



A look at key data characteristics

Distribution of Gender and Major

Below we can see displayed in a contingency table two key aspects of our data set. Represented in the rows is the distribution of male and female students, with the columns showing how many of each are majoring in a STEM field or not. With the "Sum" features of this table we can see that nearly 75 percent of our student sample is female, with just over two thirds of all students majoring in some STEM field. This could indicate that we are not accurately representing both males and females with this data, this will be kept in mind moving forward.

```r
genderSTEMtable <- addmargins(with(studentData_df, table(studentData_df$Gender, studentData_df$STEM), k
pandoc.table(genderSTEMtable, style = "grid", caption = "STEM Majors by Gender")
```

```
##
##
## +------------+----+-----+-----+
## |       | No | Yes | Sum |
## +============+====+=====+=====+
## | **Female** | 28 | 47  | 75  |
## +------------+----+-----+-----+
```

```
## |  **Male**  | 5  | 21  | 26  |
## +------------+----+-----+-----+
## |  **Sum**   | 33 | 68  | 101 |
## +------------+----+-----+-----+
##
## Table: STEM Majors by Gender
```

The data we're working with indicates that 62.6666667% of all female students are in STEM, while 80.7692308% of all male students represented are majoring in some STEM field.

Depression by Gender and Major

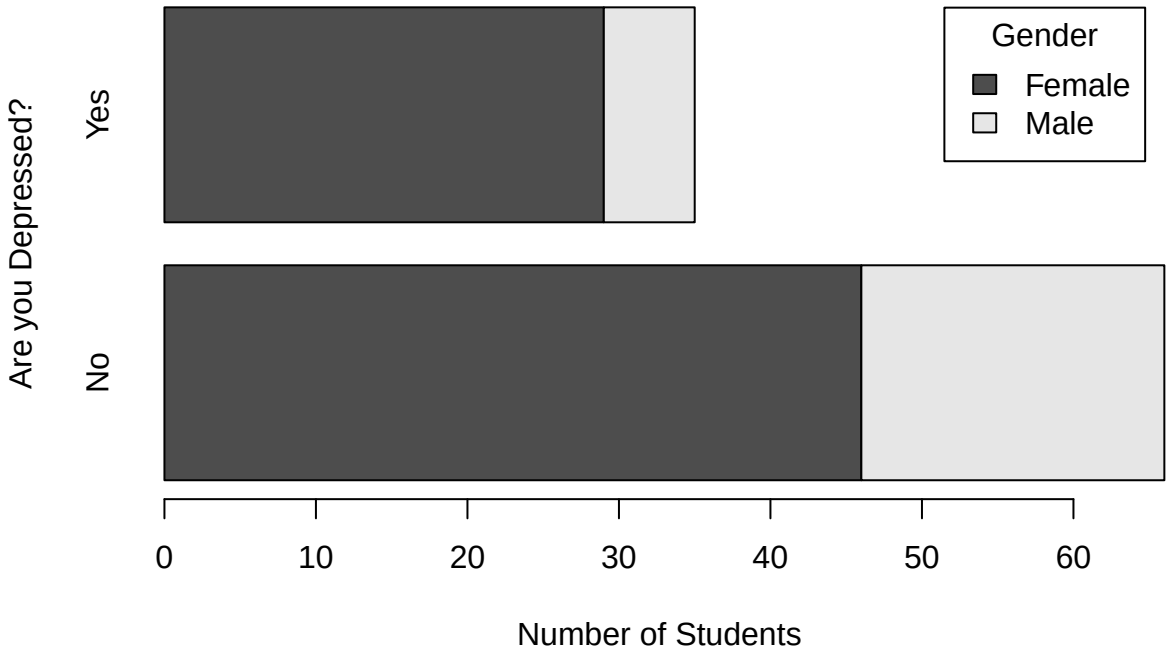
Given that our data is so heavily skewed in samples toward females, it would be interesting to take a look at how self reported depression stacks up by both gender and major. Below is once again a contingency table that represents gender by row and whether or not the student reported as depressed by column.

```
genderDepressedtable <- addmargins(with(studentData_df, table(studentData_df$Gender, studentData_df$Depr
pandoc.table(genderDepressedtable, style = "grid", caption = "Depressed Students by Gender")
```

```
##
##
## +------------+----+-----+-----+
## |       | No | Yes | Sum |
## +============+====+=====+=====+
## | **Female** | 46 | 29  | 75  |
## +------------+----+-----+-----+
## |  **Male**  | 20 |  6  | 26  |
## +------------+----+-----+-----+
## |  **Sum**   | 66 | 35  | 101 |
## +------------+----+-----+-----+
##
## Table: Depressed Students by Gender
```

The data we're working with indicates that 38.6666667% of all female students reported as depressed, while 23.0769231% of all male students indicated in the survey that they are depressed.

```
barplot (with(studentData_df, table(studentData_df$Gender, studentData_df$Depressed)), horiz=TRUE, xlab
```

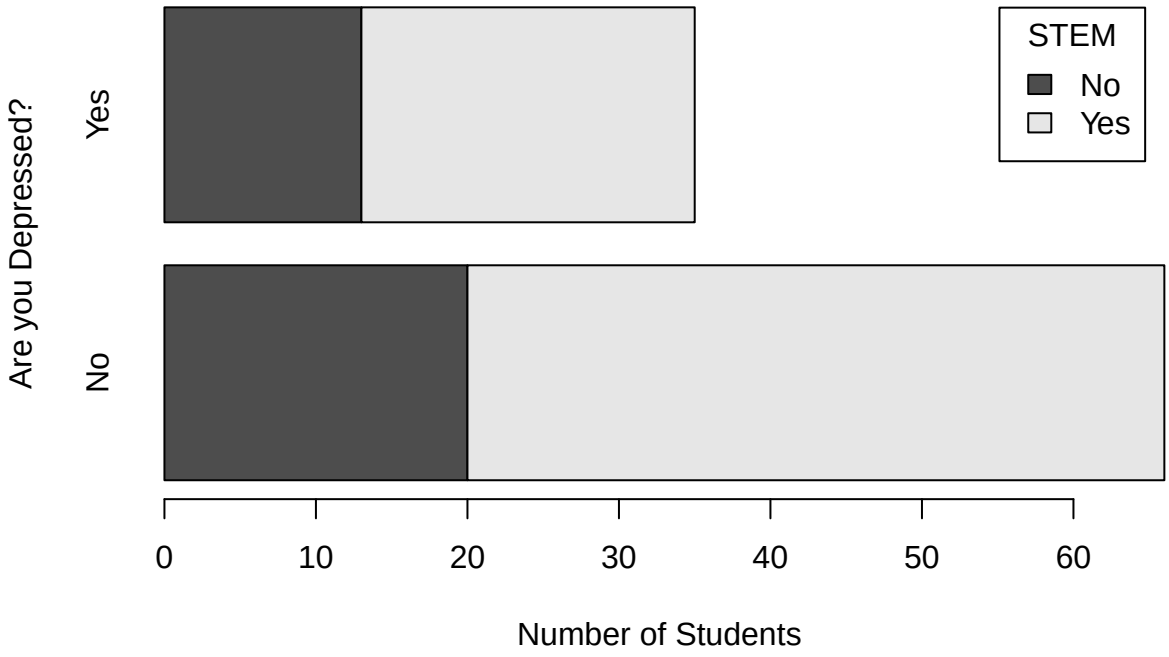


Perhaps a more relevant exploration of this data, especially for the purposes of this survey, would be the relationship between being a STEM major and reporting as Depressed. Our table below once again represents whether a student is depressed or not by column, but this time the rows give us the quality of being in a STEM related major.

```
stemDepressedtable <- addmargins(with(studentData_df, table(studentData_df$STEM, studentData_df$Depresse
pandoc.table(stemDepressedtable, style = "grid", caption = "Depressed Students by Major")
```

```
##
##
## +---------+----+-----+-----+
## |    | No | Yes | Sum |
## +=========+====+=====+=====+
## | **No**  | 20 | 13  | 33  |
## +---------+----+-----+-----+
## | **Yes** | 46 | 22  | 68  |
## +---------+----+-----+-----+
## | **Sum** | 66 | 35  | 101 |
## +---------+----+-----+-----+
##
## Table: Depressed Students by Major
```

```
barplot (with(studentData_df, table(studentData_df$STEM, studentData_df$Depressed)), horiz=TRUE, xlab =
```



Something to do with gpa vs depression

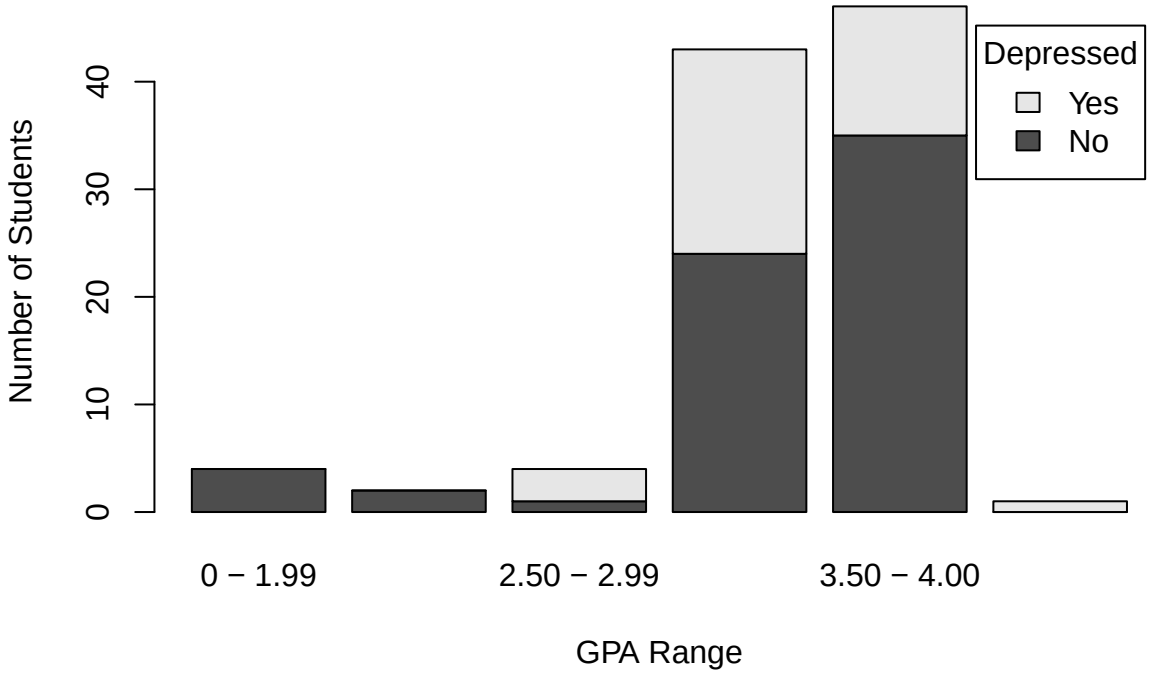
Jim - same thing as above, maybe a normal plot? stacked box plot, y depressed x gpa range

```
gpaDepressiontable <- with(studentData_df, table(studentData_df$GPA, studentData_df$Depressed), keepAtt
pandoc.table(gpaDepressiontable, style = "grid", caption = "Depressed Students by GPA")
```

```
##
##
## +-----------------+----+-----+
## |            | No | Yes |
## +=================+====+=====+
## |  **0 - 1.99**   | 4  |  0  |
## +-----------------+----+-----+
## | **2.00 - 2.49** | 2  |  0  |
## +-----------------+----+-----+
## | **2.50 - 2.99** | 1  |  3  |
## +-----------------+----+-----+
## | **3.00 - 3.49** | 24 | 19  |
## +-----------------+----+-----+
## | **3.50 - 4.00** | 35 | 12  |
## +-----------------+----+-----+
## | **3.50 - 4.00** | 0  |  1  |
## +-----------------+----+-----+
##
```

```
## Table: Depressed Students by GPA
```

```
barplot (t(gpaDepressiontable), horiz=FALSE, ylab = "Number of Students", xlab = "GPA Range", legend.te
```

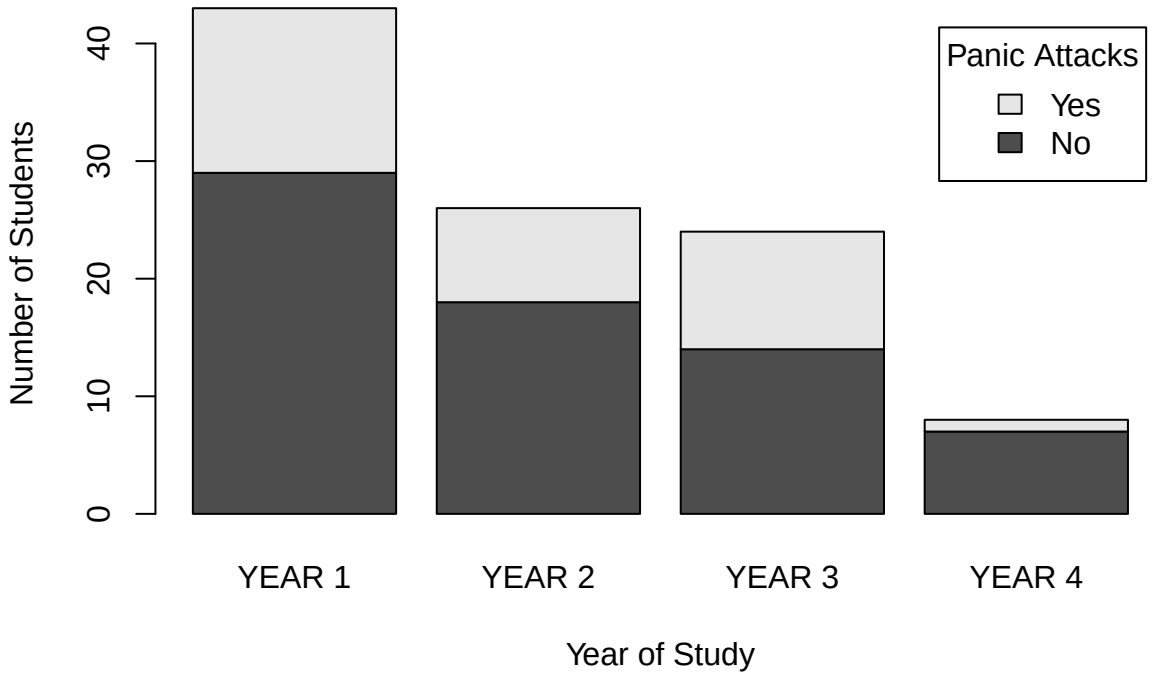


Something to do with year in school vs panic attacks

```
yearPanictable <- with(studentData_df, table(studentData_df$Year, studentData_df$Panic), keepAttrs = TR
pandoc.table(yearPanictable, style = "grid", caption = "Panic in Students by Year")
```

```
##
##
## +------------+----+-----+
## |       | No | Yes |
## +============+====+=====+
## | **YEAR 1** | 29 | 14  |
## +------------+----+-----+
## | **YEAR 2** | 18 |  8  |
## +------------+----+-----+
## | **YEAR 3** | 14 | 10  |
## +------------+----+-----+
## | **YEAR 4** | 7  |  1  |
## +------------+----+-----+
##
## Table: Panic in Students by Year
```

```
barplot (t(yearPanictable), horiz=FALSE, ylab = "Number of Students", xlab = "Year of Study", legend.te
```



Conclusions

Should answer the following questions:

Are there any unexpected patterns or relationships in your data? Does there appear to be any cause/effect phenomena? Can you suggest hypotheses for these relationships? Which variables are important? Does the data contain any anomalies or outliers? What assumptions are you making about the data, and can you verify these speculations?

Jim - i can write this after all of the data has been wrangled and charts have been made :) thank you!! <3