Survery on Mental Health Across Programs of Study in University

Deanna King, Jim Moroney, Pooja Patel, Christopher Wilhite

2022-11-17

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, no such values were found.

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   :20.53
##                     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
##
##
##
##
```

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, 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 attacks?` becomes `Panic`
- `Did you seek any specialist for a treatment?` becomes `Treatment`

```
#code for that
```

Table Manipulation

```
#code for that
```

We also saw it necessary to Jim - add attribute for "STEM? (y/n)" to the dataset. this can be done entirely manually, just use your best judgement on which one is stem.

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. The mean is approximately (add mean!).

```
#boxplot(studentData_df$age) #jim - i can't figure this out haha i am so sorry
```

CGPA Distribution

The following histogram shows the frequency of each CGPA interval for participants. Most participants have a CGPA within the range of (add mode!).

```
#hist(studentData_df$cgpa) #can't do this until data has been wrangled
```

Something with the STEM major vs mental health (depression and anxiety and panic)

Jim- i can't do this until data has been wrangled. i think one of those matrix things could work for this one

Something to do with gpa vs depression

Jim - same thing as above, maybe a normal plot?

Something to do with year in school vs panic attacks

Jim - same thing again, i definitely think a normal plot for this one

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