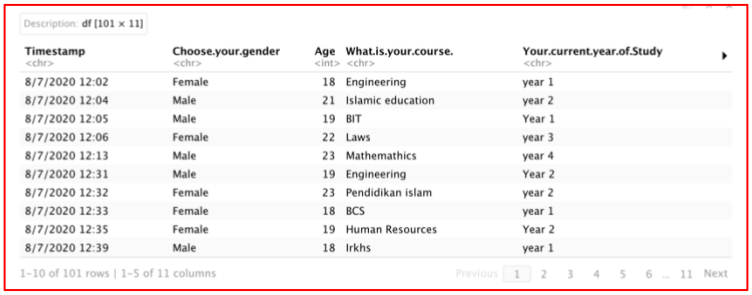
Shamim Sherafati
R Practice – ALY 6010

Module 2
Date: 2022/11/13

Data Cleaning

First, I imported libraries and then After importing the dataset “Student Survey.csv”, I started to clean the data for further analysis.



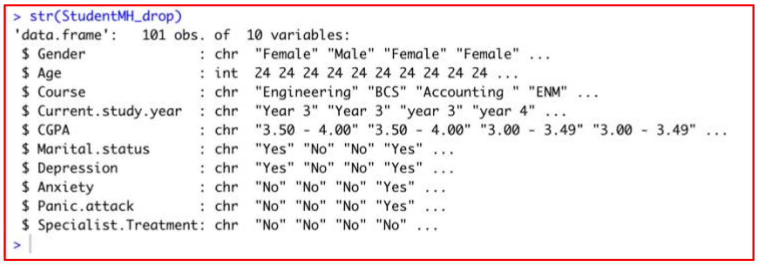
Description: df [101 × 11]

Timestamp <chr>	Choose.your.gender <chr>	Age <int>	What.is.your.course. <chr>	Your.current.year.of.Study <chr>
8/7/2020 12:02	Female	18	Engineering	year 1
8/7/2020 12:04	Male	21	Islamic education	year 2
8/7/2020 12:05	Male	19	BIT	Year 1
8/7/2020 12:06	Female	22	Laws	year 3
8/7/2020 12:13	Male	23	Mathemathics	year 4
8/7/2020 12:31	Male	19	Engineering	Year 2
8/7/2020 12:32	Female	23	Pendidikan islam	year 2
8/7/2020 12:33	Female	18	BCS	year 1
8/7/2020 12:35	Female	19	Human Resources	Year 2
8/7/2020 12:39	Male	18	Irkhs	year 1

1-10 of 101 rows | 1-5 of 11 columns

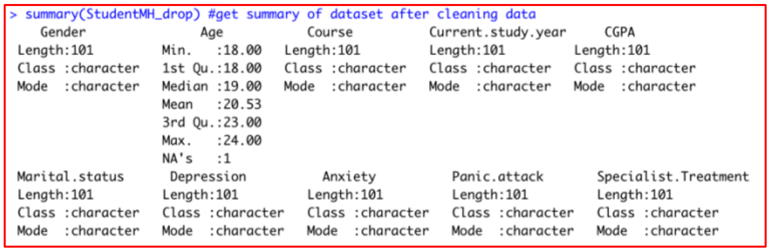
Firstly, I sorted my data with descending age column and then the column ‘Timestamp is dropped as it has no impact towards the comparison of data. Later, columns: ‘Choose.your.gender ‘, ‘What.is.your.course. ‘, ‘Your.current.year.of.Study ‘, ‘What.is.your.CGPA. ‘, ‘Do.you.have.Depression. ‘, ‘Do.you.have.Anxiety. ‘, ‘Do.you.have.Panic.attack. ‘, ‘Did.you.seek.any.specialist.for.a.treatment. ‘were renamed to ‘Gender’, ‘Course’, ‘Current.study.year’, ‘CGPA’, ‘Depression’, ‘Anxiety’, ‘Panic.attack’ and ‘Specialist.Treatment’ respectively to be more specific.

So, After cleaning data, it becomes 101 rows with 10 variables.



```
> str(StudentMH_drop)
'data.frame':	101 obs. of  10 variables:
 $ Gender              : chr  "Female" "Male" "Female" "Female" ...
 $ Age                 : int  24 24 24 24 24 24 24 24 24 24 ...
 $ Course              : chr  "Engineering" "BCS" "Accounting " "ENM" ...
 $ Current.study.year  : chr  "Year 3" "Year 3" "year 3" "year 4" ...
 $ CGPA                : chr  "3.50 - 4.00" "3.50 - 4.00" "3.00 - 3.49" "3.00 - 3.49" ...
 $ Marital.status      : chr  "Yes" "No" "No" "Yes" ...
 $ Depression          : chr  "Yes" "No" "No" "Yes" ...
 $ Anxiety             : chr  "No" "No" "No" "Yes" ...
 $ Panic.attack        : chr  "No" "No" "No" "Yes" ...
 $ Specialist.Treatment: chr  "No" "No" "No" "No" ...
> 
```

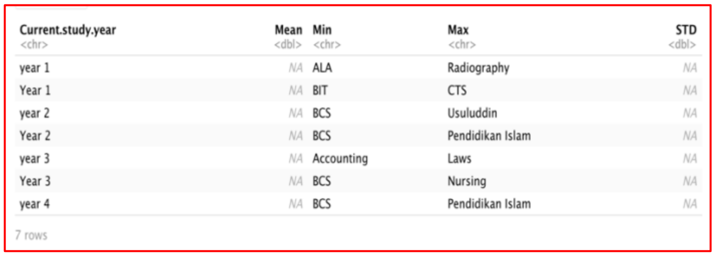
By getting summary of whole data, I gained information about each columns to start analysing them.



```
> summary(StudentMH_drop) #get summary of dataset after cleaning data
    Gender               Age            Course          Current.study.year     CGPA          
 Length:101         Min.   :18.00   Length:101         Length:101         Length:101        
 Class :character   1st Qu.:18.00   Class :character   Class :character   Class :character  
 Mode  :character   Median :19.00   Mode  :character   Mode  :character   Mode  :character  
                    Mean   :20.53                                                           
                    3rd Qu.:23.00                                                           
                    Max.   :24.00                                                           
                    NA's   :1                                                               
 Marital.status      Depression          Anxiety          Panic.attack       Specialist.Treatment
 Length:101         Length:101         Length:101         Length:101         Length:101          
 Class :character   Class :character   Class :character   Class :character   Class :character    
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character    
```

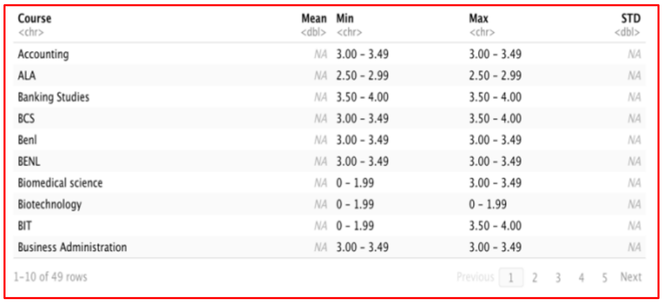
Mean, Min, Max, STD

First, I choose Courses based on different year of study to compare them; BCS is one of the course which can get both in first and 4th year of study but it's the minimum of it which means the least proportion will get this course, while Radiography and CTS for first year and Pendidikan islam makes the most proportion of courses in these two years of study.



Current.study.year	Mean	Min	Max	STD
<chr>	<dbl>	<chr>	<chr>	<dbl>
year 1	NA	ALA	Radiography	NA
Year 1	NA	BIT	CTS	NA
year 2	NA	BCS	Usuluddin	NA
Year 2	NA	BCS	Pendidikan Islam	NA
year 3	NA	Accounting	Laws	NA
Year 3	NA	BCS	Nursing	NA
year 4	NA	BCS	Pendidikan Islam	NA

7 rows



Course	Mean	Min	Max	STD
<chr>	<dbl>	<chr>	<chr>	<dbl>
Accounting	NA	3.00 - 3.49	3.00 - 3.49	NA
ALA	NA	2.50 - 2.99	2.50 - 2.99	NA
Banking Studies	NA	3.50 - 4.00	3.50 - 4.00	NA
BCS	NA	3.00 - 3.49	3.50 - 4.00	NA
Benl	NA	3.00 - 3.49	3.00 - 3.49	NA
BENL	NA	3.00 - 3.49	3.00 - 3.49	NA
Biomedical science	NA	0 - 1.99	3.00 - 3.49	NA
Biotechnology	NA	0 - 1.99	0 - 1.99	NA
BIT	NA	0 - 1.99	3.50 - 4.00	NA
Business Administration	NA	3.00 - 3.49	3.00 - 3.49	NA

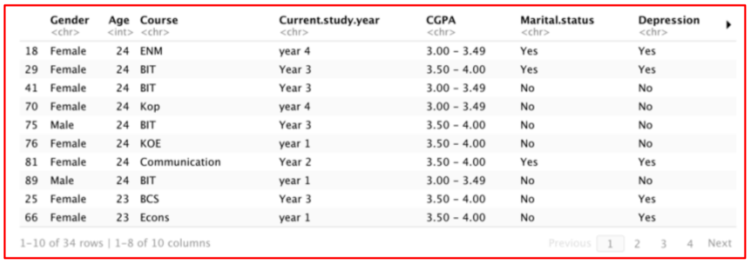
1-10 of 49 rows

Previous
1
2
3
4
5
Next

Here, I compare different courses based on the GPA score which students got. Most of the got high score in Banking studies and BCS which their maximum grade is 3.50-4.00 and their minimum grade is in the range of 3.00-3.49 which means students perform well in these two courses.
However, Biotechnology is one the course which seems most of the students has problem with it as its GPS score range is between 0-1.99.

Subsets of data

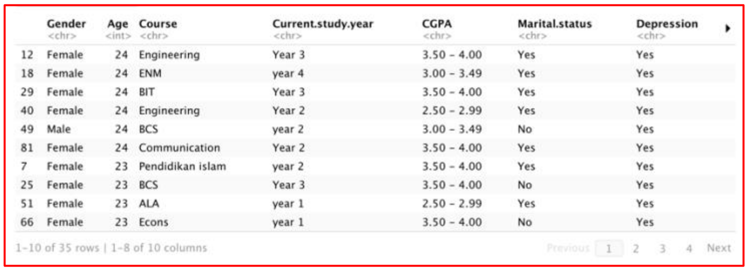
Here I created 4 subset of different variables to compare mental health in students;



	Gender	Age	Course	Current.study.year	CGPA	Marital.status	Depression
	<chr>	<int>	<chr>	<chr>	<chr>	<chr>	<chr>
18	Female	24	ENM	year 4	3.00 - 3.49	Yes	Yes
29	Female	24	BIT	Year 3	3.50 - 4.00	Yes	Yes
41	Female	24	BIT	Year 3	3.00 - 3.49	No	No
70	Female	24	Kop	year 4	3.00 - 3.49	No	No
75	Male	24	BIT	Year 3	3.50 - 4.00	No	No
76	Female	24	KOE	year 1	3.50 - 4.00	No	No
81	Female	24	Communication	Year 2	3.50 - 4.00	Yes	Yes
89	Male	24	BIT	year 1	3.00 - 3.49	No	No
25	Female	23	BCS	Year 3	3.50 - 4.00	No	Yes
66	Female	23	Econs	year 1	3.50 - 4.00	No	Yes

1-10 of 34 rows | 1-8 of 10 columns

In first chart, I filtered the survey of anxiety to 'yes' then I noticed that 34 of students has anxiety problem and the range of students with anxiety is between 18-24 years old with different variables which seems anxiety is not specify to one specific group.

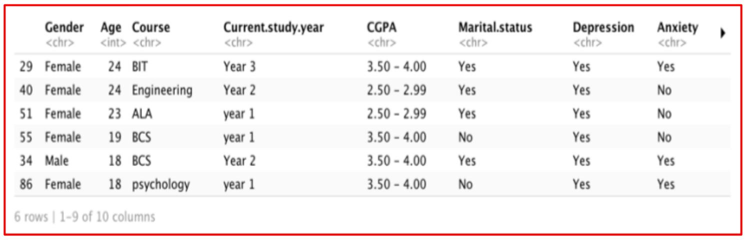


	Gender	Age	Course	Current.study.year	CGPA	Marital.status	Depression
	<chr>	<int>	<chr>	<chr>	<chr>	<chr>	<chr>
12	Female	24	Engineering	Year 3	3.50 - 4.00	Yes	Yes
18	Female	24	ENM	year 4	3.00 - 3.49	Yes	Yes
29	Female	24	BIT	Year 3	3.50 - 4.00	Yes	Yes
40	Female	24	Engineering	Year 2	2.50 - 2.99	Yes	Yes
49	Male	24	BCS	year 2	3.00 - 3.49	No	Yes
81	Female	24	Communication	Year 2	3.50 - 4.00	Yes	Yes
7	Female	23	Pendidikan islam	year 2	3.50 - 4.00	Yes	Yes
25	Female	23	BCS	Year 3	3.50 - 4.00	No	Yes
51	Female	23	ALA	year 1	2.50 - 2.99	Yes	Yes
66	Female	23	Econs	year 1	3.50 - 4.00	No	Yes

1-10 of 35 rows | 1-8 of 10 columns

This trend is also the same in the survey of depression when I filter it to 'yes' and they number is 35 from 101 students.

However, based on these two surveys, I filter Specialist.Treatment to 'yes' to see how many of them with anxiety and depression problem attend to see a doctor for treatment and I noticed that only 6 students from 35 students attend for treatment and within these 6 students, 5 of them are female. Also , most of these 6 students are depressed and got panic attack.

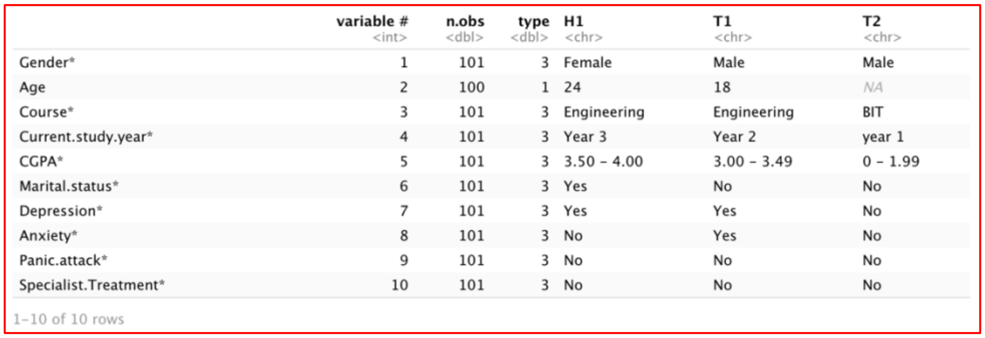


	Gender	Age	Course	Current.study.year	CGPA	Marital.status	Depression	Anxiety
	<chr>	<int>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>
29	Female	24	BIT	Year 3	3.50 - 4.00	Yes	Yes	Yes
40	Female	24	Engineering	Year 2	2.50 - 2.99	Yes	Yes	No
51	Female	23	ALA	year 1	2.50 - 2.99	Yes	Yes	No
55	Female	19	BCS	year 1	3.50 - 4.00	No	Yes	No
34	Male	18	BCS	Year 2	3.50 - 4.00	Yes	Yes	Yes
86	Female	18	psychology	year 1	3.50 - 4.00	No	Yes	Yes

6 rows | 1-9 of 10 columns

Describe Data

I filter my dataset here to describe my data from the first and two last rows to gain information about all my variables that I see the variable in different columns are actually the same dispreading.

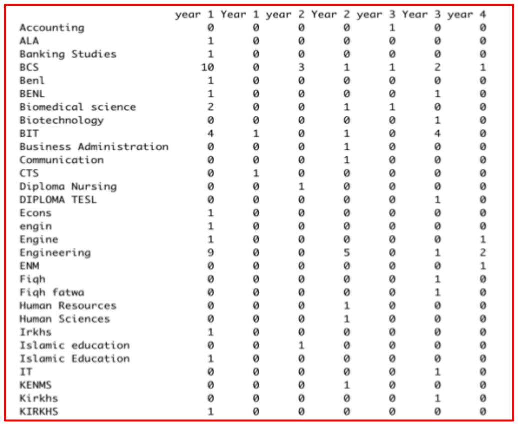


	variable #	n.obs	type	H1	T1	T2
	<int>	<dbl>	<dbl>	<chr>	<chr>	<chr>
Gender*	1	101	3	Female	Male	Male
Age	2	100	1	24	18	NA
Course*	3	101	3	Engineering	Engineering	BIT
Current.study.year*	4	101	3	Year 3	Year 2	year 1
CGPA*	5	101	3	3.50 - 4.00	3.00 - 3.49	0 - 1.99
Marital.status*	6	101	3	Yes	No	No
Depression*	7	101	3	Yes	Yes	No
Anxiety*	8	101	3	No	Yes	No
Panic.attack*	9	101	3	No	No	No
Specialist.Treatment*	10	101	3	No	No	No

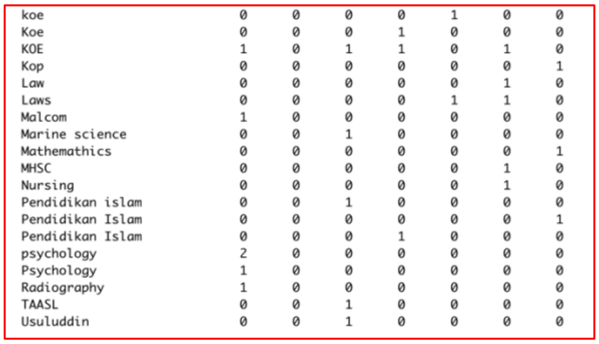
1-10 of 10 rows

Table For Course and Current.study.year

Now I created a table to explain different courses which must taken each years;
As we can see, students must take 41 courses for their first year and 26 courses for their second year and this trend decreased in other years to 24 and 8 for 3rd and 4th year respectively. So, it means that they most difficulties are in their first year of study and for BCS and Engineering which these courses have taken more than others.



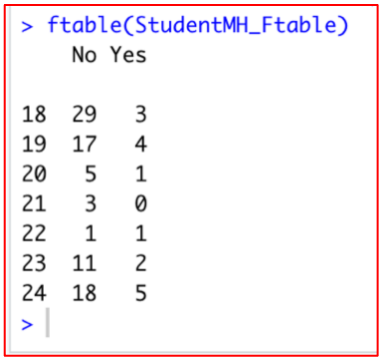
```
                          year 1 Year 1 year 2 Year 2 year 3 Year 3 year 4
Accounting                     0      0      0      0      1      0      0
ALA                            1      0      0      0      0      0      0
Banking Studies                1      0      0      0      0      0      0
BCS                           10      0      3      1      1      2      1
Benl                           1      0      0      0      0      0      0
BENL                           1      0      0      0      0      1      0
Biomedical science             2      0      0      1      1      0      0
Biotechnology                  0      0      0      0      0      1      0
BIT                            4      1      0      1      0      4      0
Business Administration        0      0      0      1      0      0      0
Communication                  0      0      0      1      0      0      0
CTS                            0      1      0      0      0      0      0
Diploma Nursing                0      0      1      0      0      0      0
DIPLOMA TESL                   0      0      0      0      0      1      0
Econs                          1      0      0      0      0      0      0
engin                          1      0      0      0      0      0      0
Engine                         1      0      0      0      0      0      1
Engineering                    9      0      0      5      0      1      2
ENM                            0      0      0      0      0      0      1
Fiqh                           0      0      0      0      0      1      0
Fiqh fatwa                     0      0      0      0      0      1      0
Human Resources                0      0      0      1      0      0      0
Human Sciences                 0      0      0      1      0      0      0
Irkhs                          1      0      0      0      0      0      0
Islamic education              0      0      1      0      0      0      0
Islamic Education              1      0      0      0      0      0      0
IT                             0      0      0      0      0      1      0
KENMS                          0      0      0      1      0      0      0
Kirkhs                         0      0      0      0      0      1      0
KIRKHS                         1      0      0      0      0      0      0
```



```
koe                            0      0      0      0      1      0      0
Koe                            0      0      0      1      0      0      0
KOE                            1      0      1      1      0      1      0
Kop                            0      0      0      0      0      0      1
Law                            0      0      0      0      0      1      0
Laws                           0      0      0      0      1      1      0
Malcom                         1      0      0      0      0      0      0
Marine science                 0      0      1      0      0      0      0
Mathemathics                   0      0      0      0      0      0      1
MHSC                           0      0      0      0      0      1      0
Nursing                        0      0      0      0      0      1      0
Pendidikan islam               0      0      1      0      0      0      0
Pendidikan Islam               0      0      0      0      0      0      1
Pendidikan Islam               0      0      0      1      0      0      0
psychology                     2      0      0      0      0      0      0
Psychology                     1      0      0      0      0      0      0
Radiography                    1      0      0      0      0      0      0
TAASL                          0      0      1      0      0      0      0
Usuluddin                      0      0      1      0      0      0      0
```

Frequency Tables For age and Marital status

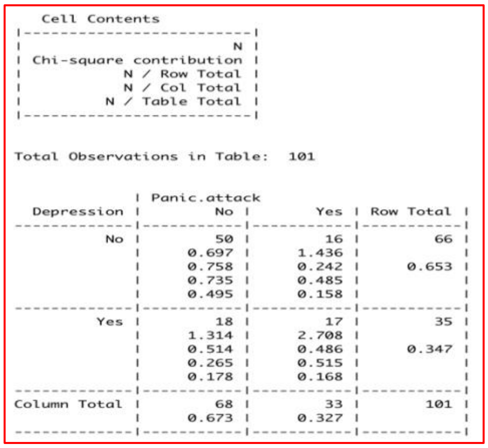
The table below shows that 9 of 101 married students are in the age range of 19 and 24 years old, and none of the 21 years old students are married



```
> ftable(StudentMH_Ftable)
    No Yes

18  29   3
19  17   4
20   5   1
21   3   0
22   1   1
23  11   2
24  18   5
>
```

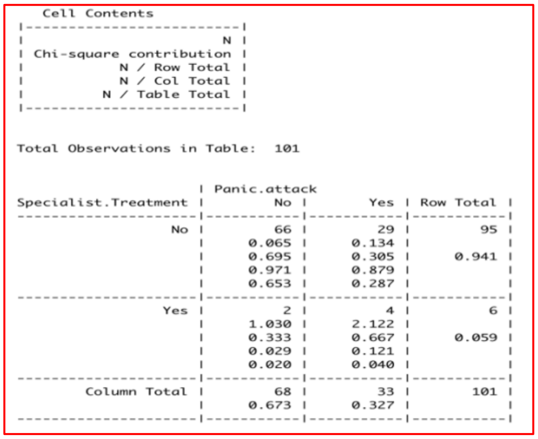
Cross Table for Specialist.Treatment, Depression, Panic.attack

The cross table below depicts two mental health problems and proportion of attending for treatment for them. As we can see, half of the students who got panic attack, are depressed too. Which this means getting panic attack has direct connection with depression.

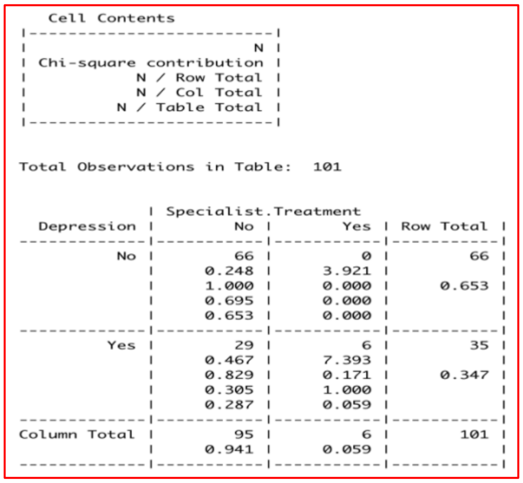


```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  101


             | Panic.attack
  Depression |        No |       Yes | Row Total |
-------------|-----------|-----------|-----------|
          No |        50 |        16 |        66 |
             |     0.697 |     1.436 |           |
             |     0.758 |     0.242 |     0.653 |
             |     0.735 |     0.485 |           |
             |     0.495 |     0.158 |           |
-------------|-----------|-----------|-----------|
         Yes |        18 |        17 |        35 |
             |     1.314 |     2.708 |           |
             |     0.514 |     0.486 |     0.347 |
             |     0.265 |     0.515 |           |
             |     0.178 |     0.168 |           |
-------------|-----------|-----------|-----------|
Column Total |        68 |        33 |       101 |
             |     0.673 |     0.327 |           |
-------------|-----------|-----------|-----------|
```

However, its different for students who are depressed. As the survey shows, from 29 of students who are depressed, only 6 of them goes for treatment which this proportion is not really good and should be considered.



```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  101


                     | Panic.attack
Specialist.Treatment |        No |       Yes | Row Total |
---------------------|-----------|-----------|-----------|
                  No |        66 |        29 |        95 |
                     |     0.065 |     0.134 |           |
                     |     0.695 |     0.305 |     0.941 |
                     |     0.971 |     0.879 |           |
                     |     0.653 |     0.287 |           |
---------------------|-----------|-----------|-----------|
                 Yes |         2 |         4 |         6 |
                     |     1.030 |     2.122 |           |
                     |     0.333 |     0.667 |     0.059 |
                     |     0.029 |     0.121 |           |
                     |     0.020 |     0.040 |           |
---------------------|-----------|-----------|-----------|
        Column Total |        68 |        33 |       101 |
                     |     0.673 |     0.327 |           |
---------------------|-----------|-----------|-----------|
```

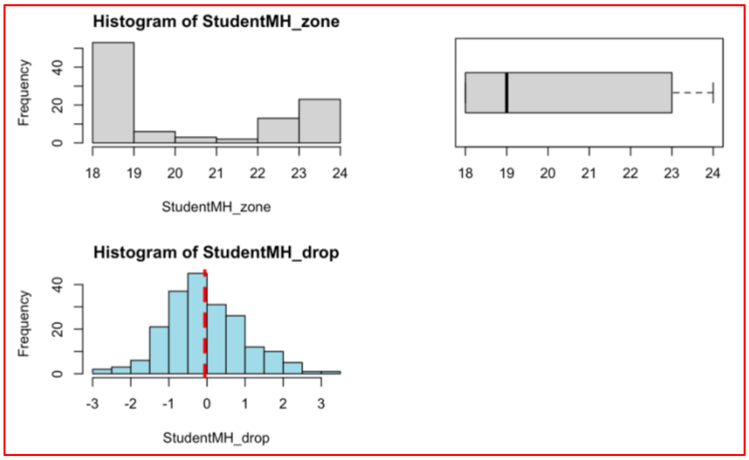


```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  101


             | Specialist.Treatment
  Depression |        No |       Yes | Row Total |
-------------|-----------|-----------|-----------|
          No |        66 |         0 |        66 |
             |     0.248 |     3.921 |           |
             |     1.000 |     0.000 |     0.653 |
             |     0.695 |     0.000 |           |
             |     0.653 |     0.000 |           |
-------------|-----------|-----------|-----------|
         Yes |        29 |         6 |        35 |
             |     0.467 |     7.393 |           |
             |     0.829 |     0.171 |     0.347 |
             |     0.305 |     1.000 |           |
             |     0.287 |     0.059 |           |
-------------|-----------|-----------|-----------|
Column Total |        95 |         6 |       101 |
             |     0.941 |     0.059 |           |
-------------|-----------|-----------|-----------|
```

As we can see from the third table, only 4 of 29 students who got panic attack attend for a specialist for treatment.

So, to conclude, we can realize that as this two problems have connection with others, students must consider to attend for treatment.

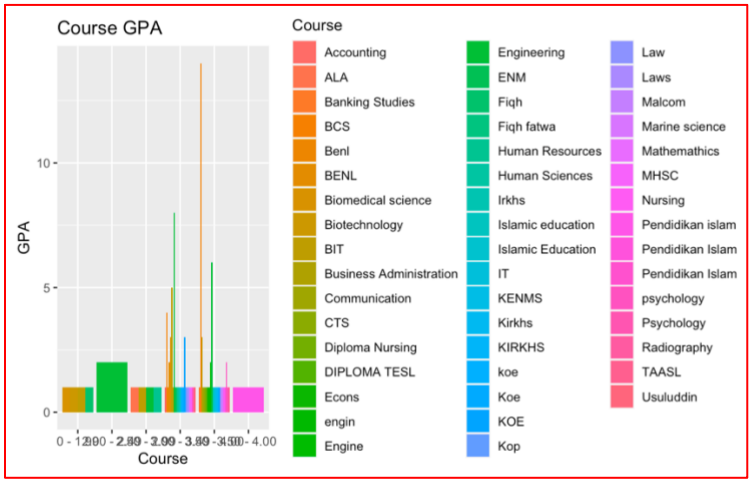
Diagram :

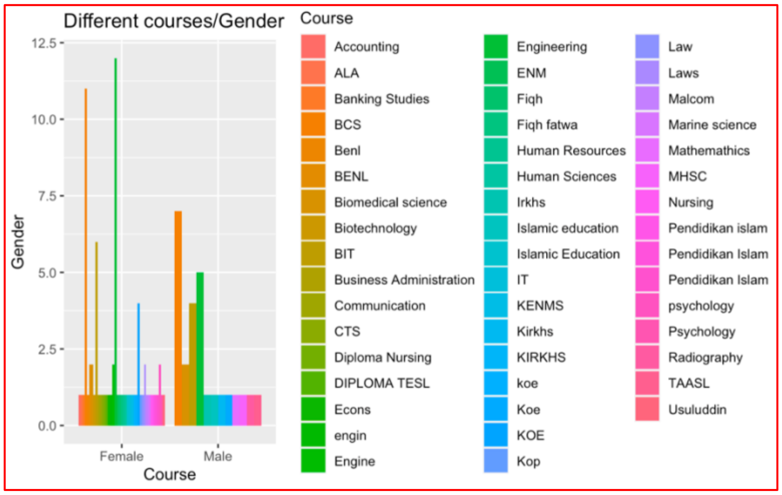


In this three table which I used 'par' and 'abline' function to create them, I showed 'Age' diagram proportion in three different way. As we can see, the age of students from the survey which we had, shows that its in the range of 18 to 24 and as the diagrams below depict, most of them are in the age of 18-19 years old.

Different courses GPA Score

The diagram below depicts different course GPA score; As we can see, TAASL, Radiography, Psychology, Nursing and MHSC are the courses which the survey shows students get higher grade on them. While Biotechnology, BIT and Business administration seems to be hard for students as they got 0-1.90 and this trend follows by Engineering, ENM, Engine and Econs with 1.90-2.00 that scored. Other courses seems to be moderate for students.

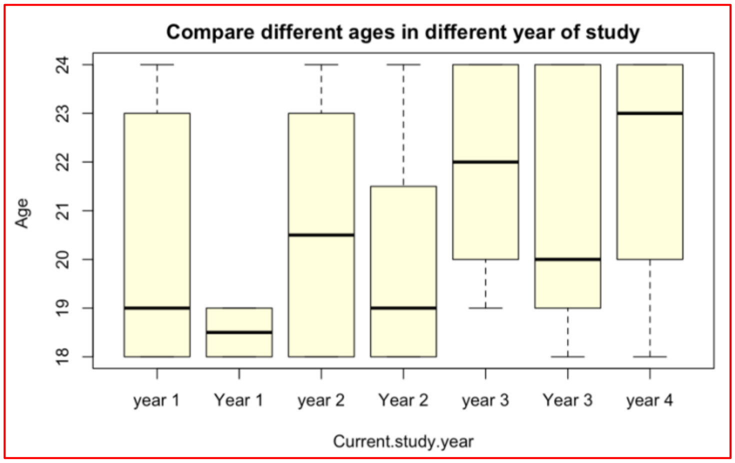




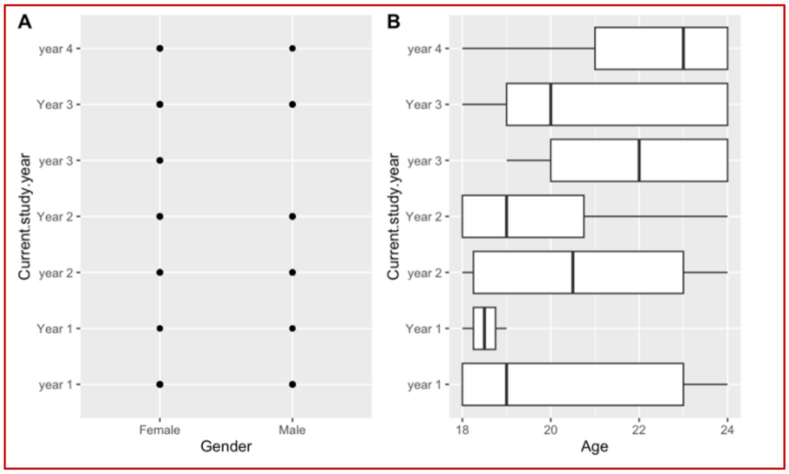
The bar plot below shows the students gender in different course which they got; Almost the same course they took but Banking Studies and Engineering is the most taken course and in female students these two courses are more interested and have higher counted in compare with male students.

Compare different ages in different year of study

One of the question which asked from students was their current year of study. So, I created a boxplot for this and found that as it can be expected, most of the first year students are 18-19 years old and this trend increase by increasing their year of study but there is some exception that in 3rd year of study, both 22 and 20 years old students study. Also, it can be seen that in their last year of study, most of them are 24th.



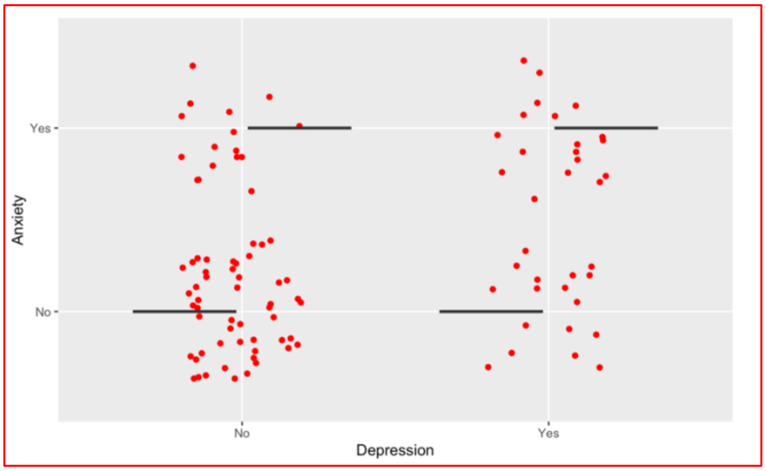
Now, I combine both age and gender with current year of study in one diagram as we can see below.



Jitter chart

The main reason to use a jitter plot rather than a strip plot is when you have too many marks overlapping and either you want to be able to select any individual mark (which is difficult or even impossible if the marks overlap entirely).

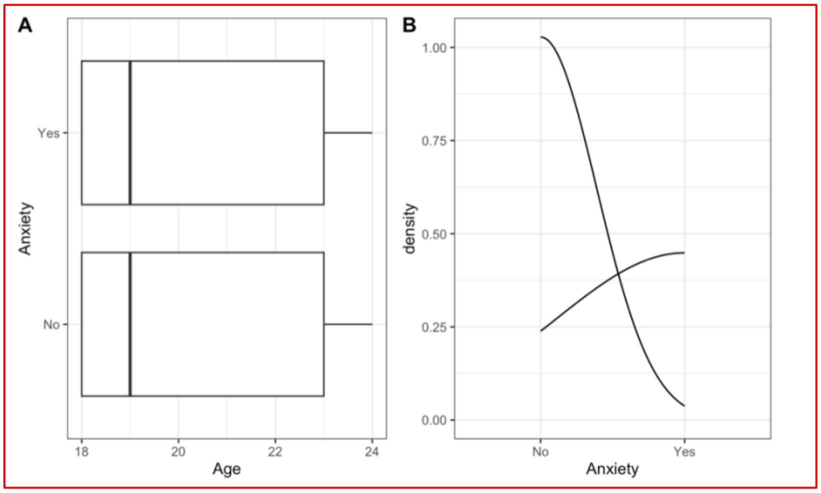
For example, I created a jitter plot and will explain its detail.



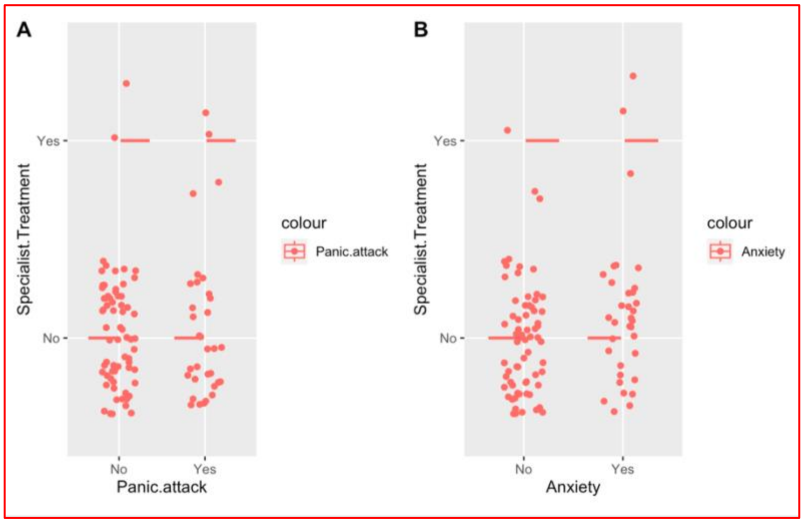
This jitter chart represent Depression and Anxiety; I use boxplots to detect outliers as well. Here we can seen that those students who are depressed, can also have anxiety as well and this trend is also the same for those who have anxiety. So, it can be conclude that, the relation between anxiety and depression is direct.

Impact of anxiety on different ages

Here in these charts, I depict the impact of anxiety in different ages; Most of the students who have anxiety are in the age range of 18-20 years old and from the density chart we can see that the level of anxiety is increased which is something that may base of the challenges that student may be faced to.



Now, in the jitter chart below that depicts the relation between anxiety and panic attack with treatment;



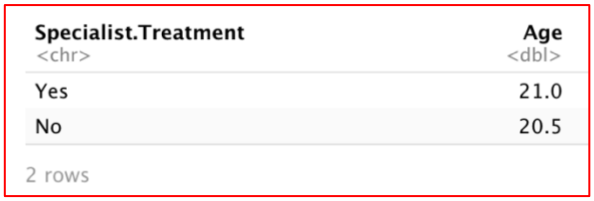
As we can seen from the chart, unfortunately both anxiety students and those with panic attack didn't attend to a specialist for treatment which is something serious that must be consider.

Average of ages who go for Specialist.Treatment

Here I create a table to choose the average of ages who go for Specialist.Treatment based on the survey.

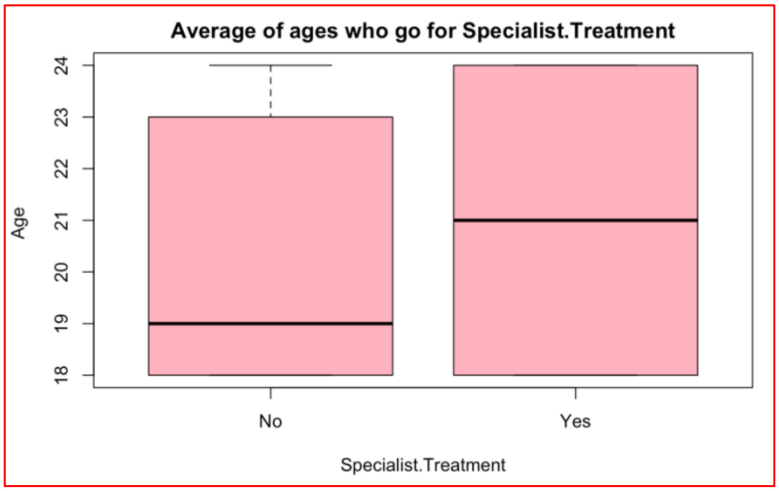
The average age of students who attend to Specialist Treatment is 21 years old and this average for those who don't go for treatment is about 20 years old.

Now, based on this information I created a boxplot to depict it more accurately and as we can see from the chart, it seems as their age increased they became more willing to go for treatment.

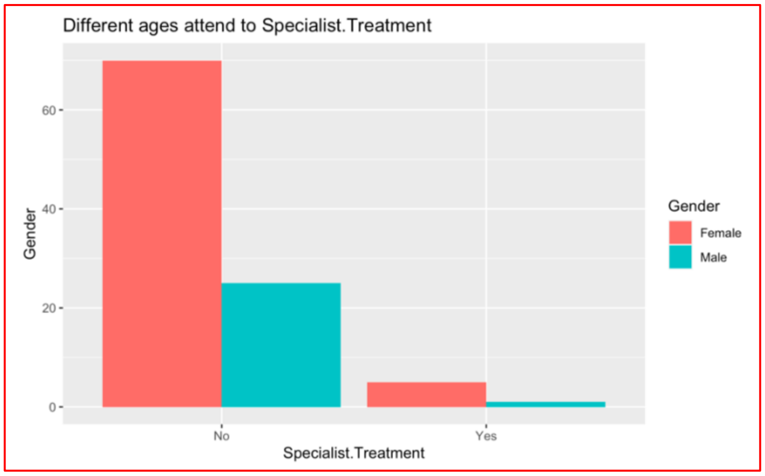


Specialist.Treatment <chr>	Age <dbl>
Yes	21.0
No	20.5

2 rows



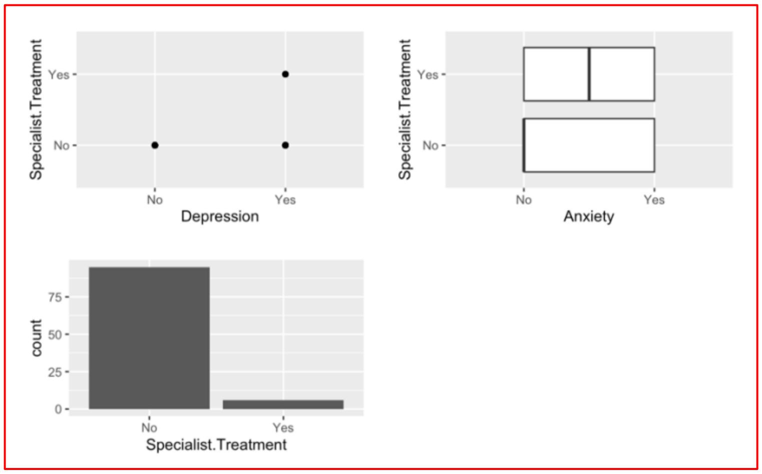
Different ages gender attend to Specialist.Treatment



In this stage I compare different gender who attend for treatment; As we can see the proportion of those students who attend for treatment are more within female students rather than the male students.

check if those with depression, anxiety attend to Specialist.Treatment or not

Finally, I created diagrams to check if those students with depression, anxiety attend to Specialist Treatment or not;
Those students who have depression are more likely to attend for treatment in compare with those with anxiety that its moderate between them to go or not. However in total, most of the students based on survey are not willing to go for treatment for their mental health problem.



References:

1. Aphalo, Pedro J. 2017. Ggpmisc: Miscellaneous Extensions to ’Ggplot2’. https://CRAN.R-project.org/package=ggpmisc.
2. Attali, Dean. 2017. GgExtra: Add Marginal Histograms to ’Ggplot2’, and More ’Ggplot2’ Enhancements. https://github.com/daattali/ggExtra.
3. Wilke, Claus O. 2017. Ggridges: Ridgeline Plots in ’Ggplot2’. https://CRAN.R-project.org/package=ggridges.
4. Wickham, Hadley, and Winston Chang. 2017. Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics.
5. Schloerke, Barret, Jason Crowley, Di Cook, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Joseph Larmarange. 2016. GGally: Extension to ’Ggplot2’. https://CRAN.R-project.org/package=GGally.