ALY6010 MODULE 3

WEEK 3

Name : shreyansh Bhalodiya
NUID : 002664707

Dataset Selected:

```
head(df)
```

A data.frame: 6

	Timestamp	Choose.your.gender	Age	What.is.your.course.	Your.current.year.of.Study	What.is.your.CGPA.	Marital.status	D
	<chr>	<chr>	<int>	<chr>	<chr>	<chr>	<chr>	
1	8/7/2020 12:02	Female	18	Engineering	year 1	3.00 - 3.49	No	
	8/7/2020							

Dimensions of Dataframe : `101 * 11`

Renaming data frame columns to short names.

```
colnames(df)[colnames(df) == "Choose.your.gender"] <- "gender"
colnames(df)[colnames(df) == "What.is.your.course."] <- "course"
colnames(df)[colnames(df) == "Your.current.year.of.Study"] <- "year_of_study"
colnames(df)[colnames(df) == "What.is.your.CGPA."] <- "CGPA"
colnames(df)[colnames(df) == "Marital.status"] <- "Marital_status"
colnames(df)[colnames(df) == "Do.you.have.Depression."] <- "Depression"
colnames(df)[colnames(df) == "Do.you.have.Anxiety."] <- "Anxiety"
colnames(df)[colnames(df) == "Do.you.have.Panic.attack."] <- "attack"
colnames(df)[colnames(df) == "Did.you.seek.any.specialist.for.a.treatment."] <- "specialist_treatment"
```

Part 1
One Sample T test

We have a student age column for all 4 years of students. Ideally their age will be in range 18-24 let's assume their mean is 23 and do one sample test.

Null Hypothesis : mean is equal to 23
Alternative Hypothesis: Mean is not equal to 23

```
## One test sample test

# One-sample t-test
res <- t.test(df$Age, mu = 23)
# Printing the results
res
```

```
        One Sample t-test

data:  df$Age
t = -9.8947, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 23
95 percent confidence interval:
 20.03468 21.02532
sample estimates:
mean of x
    20.53
```

Results

We can clearly see that the mean is not equal to 23 hence our alternative hypothesis is true following is our interpretations for other factors used in hypothesis testing.

```
# mean is equal to 23
# t is the t-test statistic value (t = -57.966),
# df is the degrees of freedom (df= 99),
# p-value is the significance level of the t-test (p-value =  2.2e-16).
# conf.int is the confidence interval of the mean at 95% (conf.int = [20.03468,
21.02532]);
# sample estimate is the mean value of the sample (mean = 20.53).
```

Part 2

Hypothesis testing for p-value using two sample test

we can see that our values are not aligned in year of study columns. First, let's give a common name for each unique year.

```
df["year_of_study"][df["year_of_study"] == "year 1"] <- "Year 1"
df["year_of_study"][df["year_of_study"] == "year 2"] <- "Year 2"
df["year_of_study"][df["year_of_study"] == "year 3"] <- "Year 3"


table(df$year_of_study) ### rechecking again
```

```
Year 1 Year 2 Year 3 year 4
    43     26     24      8
```

Problem statement

Let's assume that the age of males and females in the same year of study is the same. We will conduct the T-test for year 1 and year 2 assuming that mean age for male and female students is the same for each year.

Year 1

```
df_1 <- df[which(df$year_of_study=='Year 1'),]
df_1_m <- df_1[which(df_1$gender=='Male'),]
df_1_f <- df_1[which(df_1$gender=='Female'),]

t.test(df_1_m$Age, df_1_f$Age)
```

```
        Welch Two Sample t-test

data:  df_1_m$Age and df_1_f$Age
t = -0.23321, df = 13.617, p-value = 0.8191
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.064885  1.660845
sample estimates:
mean of x mean of y
 19.88889  20.09091
```

Year 2

```
df_2 <- df[which(df$year_of_study=='Year 2'),]
df_2_m <- df_1[which(df_2$gender=='Male'),]
df_2_f <- df_1[which(df_2$gender=='Female'),]

t.test(df_2_m$Age, df_2_f$Age)
```

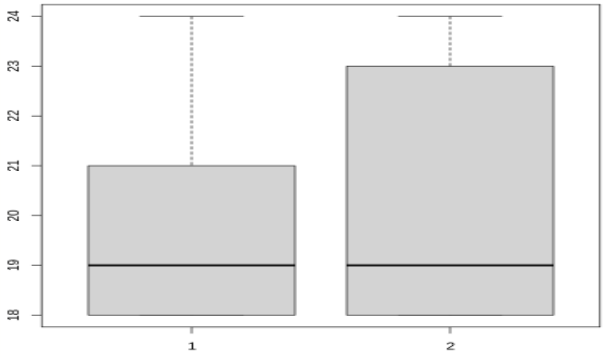
```
        Welch Two Sample t-test

data:  df_2_m$Age and df_2_f$Age
t = 0.55155, df = 20.743, p-value = 0.5872
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.224595  2.107712
sample estimates:
mean of x mean of y
 19.72727  19.28571
```

For both year students we are assuming that the mean of male and female is the same.
But the same is not true as in year 1 we can see that the mean of male and mean of female is 19.88 and 20.09 which is definitely closer but not equal. We are getting a p-value as 0.8191 which is greater than 0.05.

On the similar lines mean for year 2 male and female students is not the same it's more closer as compared to year 1. For year 2 we got a p-value as 0.5872 which is more than 0.05 but less than 0.8191 the main reason is the mean age for year 2 students is more close as compared to year 1 students.

Overall we must be wondering that means are almost small but still we are high P-value. The main reason is that the overall age of almost all the students is the same hence we have very small width for assumptions. Hence even a smaller difference in mean ages is giving high value in P-value. We can see the same iin boxplots below.

YEAR 1



YEAR 2