Richa Rambhia
November 28, 2021

Module 4 R Practice
Report & Outputs

Introduction

Part 1:

The cats dataset in the package 'MASS' was used for the part 1 task of this assignment. The cats dataset has data attributes like group of people, body weight and height. This dataset gives us information about the group of people and their body weight and height on which the analysis can be done to check whether the body weight and height of the male group is higher or the female group. The group of people is divided into two categories in the dataset namely, the male group and the female group.

Part 2:

In the part 2 of the assignment two samples of data were given on which the analysis needs to be done. One of the two samples of data is the average sleeping quality scores in the week before the workshop and the other is the average sleeping quality scores in the week following the workshop. This dataset is to be analyzed to further know and evaluate if the meditation has an effect on the sleep quality or not. Here, the sleeping quality is ranked on the scale from 0-10 which means the higher the better.

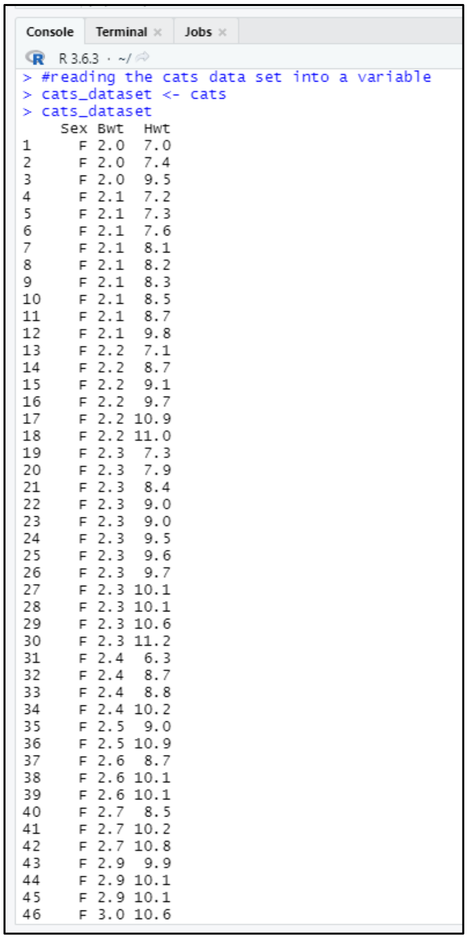
The analysis and hypothesis testing are to be done in both the parts to answer the questions for further analysis.

Part 1
Data Analysis

The package names 'MASS' was installed at first to access the cats dataset and the library for the same was imported. Once the libraries were imported, the cats dataset was read into a variable. To answer the questions about the dataset and analyze the dataset, it was first needed to describe the data and have an understanding about what the dataset exactly consists of after which one can start its analysis. Descriptive analysis is the basic step in the data analysis process where we describe the data such as displaying the column names, displaying the start and end records of the dataset, describing the dimensions of the dataset and also all the statistical value description such as min, max, mean, median, range and standard deviation. All of the statistical values mentioned can be performed using the function in r names summary. This function returns all the statistical values of the dataset. The output of the same is as shown below.

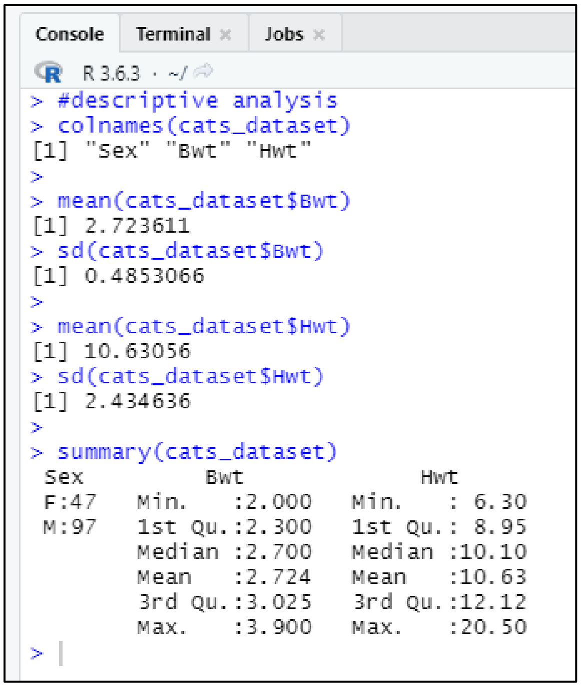
Output:

1. Reading the csv file



```
> #reading the cats data set into a variable
> cats_dataset <- cats
> cats_dataset
    Sex Bwt  Hwt
1     F 2.0  7.0
2     F 2.0  7.4
3     F 2.0  9.5
4     F 2.1  7.2
5     F 2.1  7.3
6     F 2.1  7.6
7     F 2.1  8.1
8     F 2.1  8.2
9     F 2.1  8.3
10    F 2.1  8.5
11    F 2.1  8.7
12    F 2.1  9.8
13    F 2.2  7.1
14    F 2.2  8.7
15    F 2.2  9.1
16    F 2.2  9.7
17    F 2.2 10.9
18    F 2.2 11.0
19    F 2.3  7.3
20    F 2.3  7.9
21    F 2.3  8.4
22    F 2.3  9.0
23    F 2.3  9.0
24    F 2.3  9.5
25    F 2.3  9.6
26    F 2.3  9.7
27    F 2.3 10.1
28    F 2.3 10.1
29    F 2.3 10.6
30    F 2.3 11.2
31    F 2.4  6.3
32    F 2.4  8.7
33    F 2.4  8.8
34    F 2.4 10.2
35    F 2.5  9.0
36    F 2.5 10.9
37    F 2.6  8.7
38    F 2.6 10.1
39    F 2.6 10.1
40    F 2.7  8.5
41    F 2.7 10.2
42    F 2.7 10.8
43    F 2.9  9.9
44    F 2.9 10.1
45    F 2.9 10.1
46    F 3.0 10.6
```

2. Descriptive Analysis

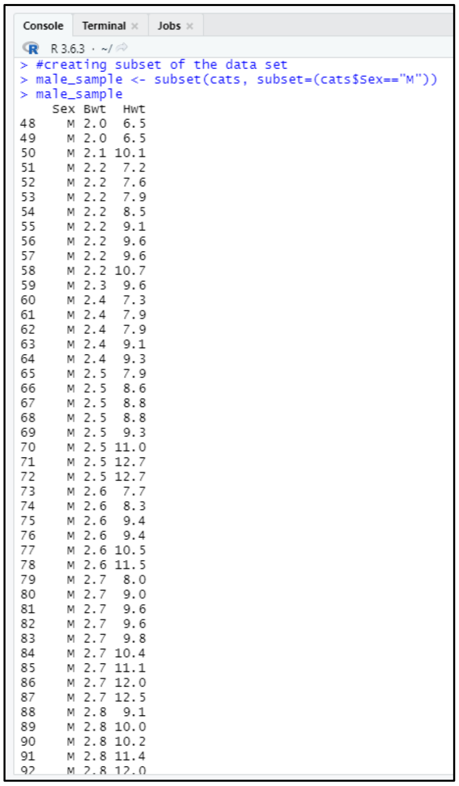


```
> #descriptive analysis
> colnames(cats_dataset)
[1] "Sex" "Bwt" "Hwt"
>
> mean(cats_dataset$Bwt)
[1] 2.723611
> sd(cats_dataset$Bwt)
[1] 0.4853066
>
> mean(cats_dataset$Hwt)
[1] 10.63056
> sd(cats_dataset$Hwt)
[1] 2.434636
>
> summary(cats_dataset)
 Sex         Bwt              Hwt
 F:47   Min.   :2.000   Min.   : 6.30
 M:97   1st Qu.:2.300   1st Qu.: 8.95
        Median :2.700   Median :10.10
        Mean   :2.724   Mean   :10.63
        3rd Qu.:3.025   3rd Qu.:12.12
        Max.   :3.900   Max.   :20.50
> |
```

In order to perform the analysis and hypothesis testing, it was important to create a subset of the dataset and create new variables. The group of people consisted of the male category and the female category which needed to be extracted so as the analysis on separate entities can be performed. Also, the body weight was the parameter on which the testing needs to be performed which was extracted as we needed to answer the question of whether the male and female cat samples have the same bodyweight.

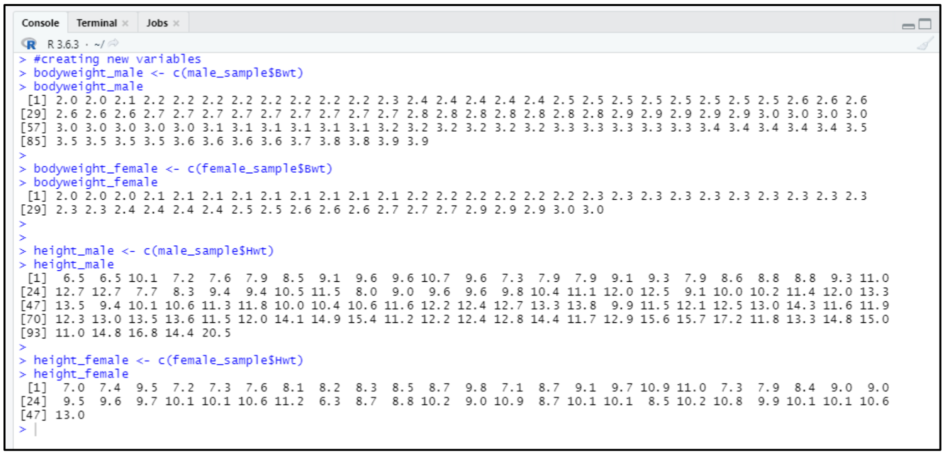
Output:

1. Creating subset of the dataset



```
> #creating subset of the data set
> male_sample <- subset(cats, subset=(cats$Sex=="M"))
> male_sample
    Sex Bwt  Hwt
48    M 2.0  6.5
49    M 2.0  6.5
50    M 2.1 10.1
51    M 2.2  7.2
52    M 2.2  7.6
53    M 2.2  7.9
54    M 2.2  8.5
55    M 2.2  9.1
56    M 2.2  9.6
57    M 2.2  9.6
58    M 2.2 10.7
59    M 2.3  9.6
60    M 2.4  7.3
61    M 2.4  7.9
62    M 2.4  7.9
63    M 2.4  9.1
64    M 2.4  9.3
65    M 2.5  7.9
66    M 2.5  8.6
67    M 2.5  8.8
68    M 2.5  8.8
69    M 2.5  9.3
70    M 2.5 11.0
71    M 2.5 12.7
72    M 2.5 12.7
73    M 2.6  7.7
74    M 2.6  8.3
75    M 2.6  9.4
76    M 2.6  9.4
77    M 2.6 10.5
78    M 2.6 11.5
79    M 2.7  8.0
80    M 2.7  9.0
81    M 2.7  9.6
82    M 2.7  9.6
83    M 2.7  9.8
84    M 2.7 10.4
85    M 2.7 11.1
86    M 2.7 12.0
87    M 2.7 12.5
88    M 2.8  9.1
89    M 2.8 10.0
90    M 2.8 10.2
91    M 2.8 11.4
92    M 2.8 12.0
```

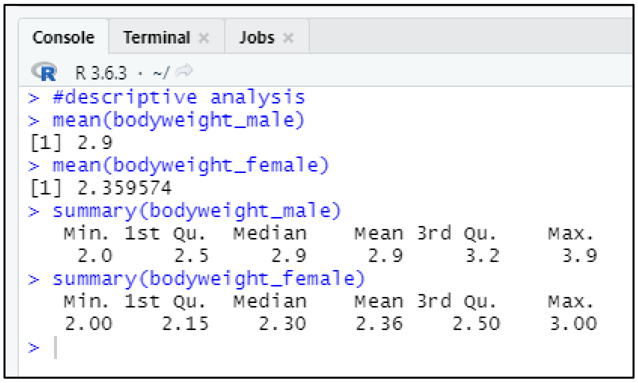
2. Creating new variables



```
> #creating new variables
> bodyweight_male <- c(male_sample$Bwt)
> bodyweight_male
 [1] 2.0 2.0 2.1 2.2 2.2 2.2 2.2 2.2 2.2 2.2 2.2 2.3 2.4 2.4 2.4 2.4 2.4 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.6 2.6 2.6
[29] 2.6 2.6 2.6 2.7 2.7 2.7 2.7 2.7 2.7 2.7 2.7 2.7 2.8 2.8 2.8 2.8 2.8 2.8 2.8 2.9 2.9 2.9 2.9 2.9 3.0 3.0 3.0 3.0
[57] 3.0 3.0 3.0 3.0 3.0 3.1 3.1 3.1 3.1 3.1 3.1 3.2 3.2 3.2 3.2 3.2 3.2 3.3 3.3 3.3 3.3 3.3 3.4 3.4 3.4 3.4 3.4 3.5
[85] 3.5 3.5 3.5 3.5 3.6 3.6 3.6 3.6 3.7 3.8 3.8 3.9 3.9
> 
> bodyweight_female <- c(female_sample$Bwt)
> bodyweight_female
 [1] 2.0 2.0 2.0 2.1 2.1 2.1 2.1 2.1 2.1 2.1 2.1 2.1 2.2 2.2 2.2 2.2 2.2 2.2 2.3 2.3 2.3 2.3 2.3 2.3 2.3 2.3 2.3 2.3
[29] 2.3 2.3 2.4 2.4 2.4 2.4 2.5 2.5 2.6 2.6 2.6 2.7 2.7 2.7 2.9 2.9 2.9 3.0 3.0
> 
> 
> height_male <- c(male_sample$Hwt)
> height_male
 [1]  6.5  6.5 10.1  7.2  7.6  7.9  8.5  9.1  9.6  9.6 10.7  9.6  7.3  7.9  7.9  9.1  9.3  7.9  8.6  8.8  8.8  9.3 11.0
[24] 12.7 12.7  7.7  8.3  9.4  9.4 10.5 11.5  8.0  9.0  9.6  9.6  9.8 10.4 11.1 12.0 12.5  9.1 10.0 10.2 11.4 12.0 13.3
[47] 13.5  9.4 10.1 10.6 11.3 11.8 10.0 10.4 10.6 11.6 12.2 12.4 12.7 13.3 13.8  9.9 11.5 12.1 12.5 13.0 14.3 11.6 11.9
[70] 12.3 13.0 13.5 13.6 11.5 12.0 14.1 14.9 15.4 11.2 12.2 12.4 12.8 14.4 11.7 12.9 15.6 15.7 17.2 11.8 13.3 14.8 15.0
[93] 11.0 14.8 16.8 14.4 20.5
> 
> height_female <- c(female_sample$Hwt)
> height_female
 [1]  7.0  7.4  9.5  7.2  7.3  7.6  8.1  8.2  8.3  8.5  8.7  9.8  7.1  8.7  9.1  9.7 10.9 11.0  7.3  7.9  8.4  9.0  9.0
[24]  9.5  9.6  9.7 10.1 10.1 10.6 11.2  6.3  8.7  8.8 10.2  9.0 10.9  8.7 10.1 10.1  8.5 10.2 10.8  9.9 10.1 10.1 10.6
[47] 13.0
> |
```

The descriptive analysis was also performed for the subset of the dataset and also on the new variables created.

Output:

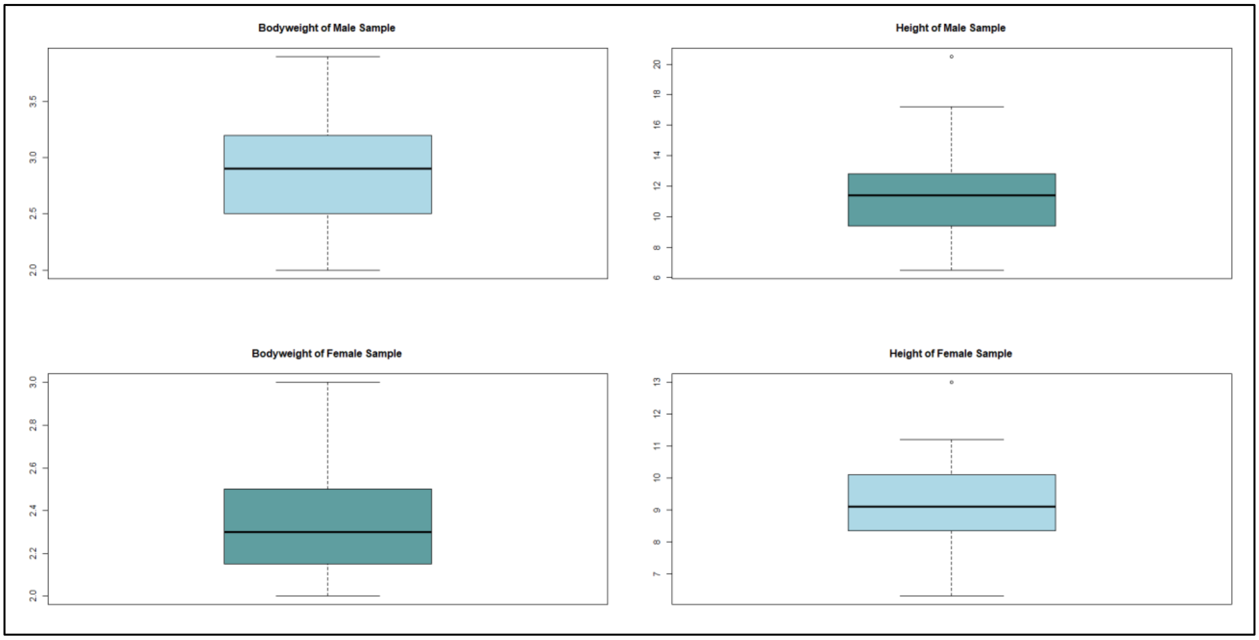


```
> #descriptive analysis
> mean(bodyweight_male)
[1] 2.9
> mean(bodyweight_female)
[1] 2.359574
> summary(bodyweight_male)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0     2.5     2.9     2.9     3.2     3.9
> summary(bodyweight_female)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00    2.15    2.30    2.36    2.50    3.00
> |
```

Data Visualization

Next step after the descriptive analysis is the data visualizations where the data understanding would be clearer. The visualizations created was on the body weight and height of the male and female cat sample of the dataset where boxplot was used to create the same. Boxplot helps in understanding the statistical values depicting groups of numerical data through their quartiles. It basically displays the dataset based on a five number summary i.e., the minimum, the maximum, the sample median and the first and third quartiles. Below is the boxplot shown for the two attributes of the male and female cat sample.

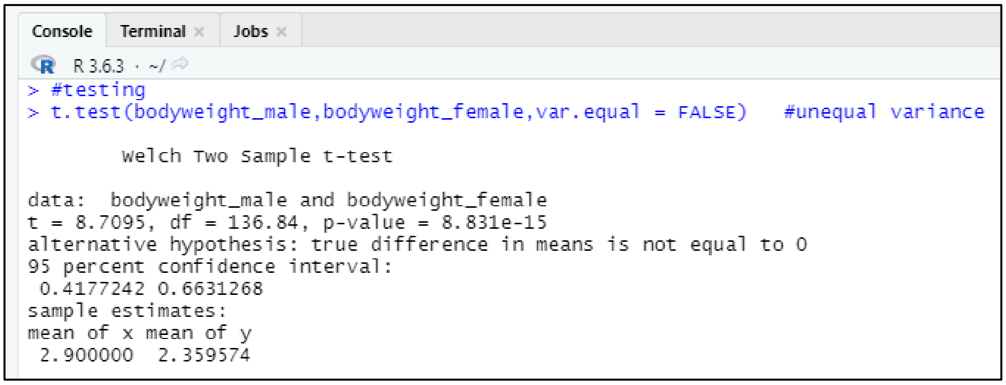
Output:



Hypothesis Testing

Hypothesis testing here would help in answering the questions termed and that would further help us in understanding the data and analyzing the dataset. T-test is used in the hypothesis testing which is a statistical test that compares the means of two samples. Here our question is to test whether the bodyweight of the male and female cat sample is same or not. This question could be answered by performing the hypothesis testing using the t-test. Since there are two attributes of the same group which are independent of each other, two-sample t-test is to be performed as two-sample t-test is performed when the two population means of the samples are independent of each other. The output shown below performs the two-sample t-test on the bodyweight of the male and female cat samples which answers the question.

Output:



```
> #testing
> t.test(bodyweight_male,bodyweight_female,var.equal = FALSE)   #unequal variance

        Welch Two Sample t-test

data:  bodyweight_male and bodyweight_female
t = 8.7095, df = 136.84, p-value = 8.831e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.4177242 0.6631268
sample estimates:
mean of x mean of y
 2.900000  2.359574
```

The output above describes the t-value, degree of freedoms, p-value and the mean of x and y. These values help us in reaching to a conclusion and in deciding whether to accept or reject the null hypothesis. The table below shows the test values computed.

t-test statistic	Degrees of freedom	p-value	95% confidence interval	Mean of x and y
8.7095	136.84	8.831e-15	[0.4177242 0.6631268]	mean of x: 2.900000 mean of y: 2.359574

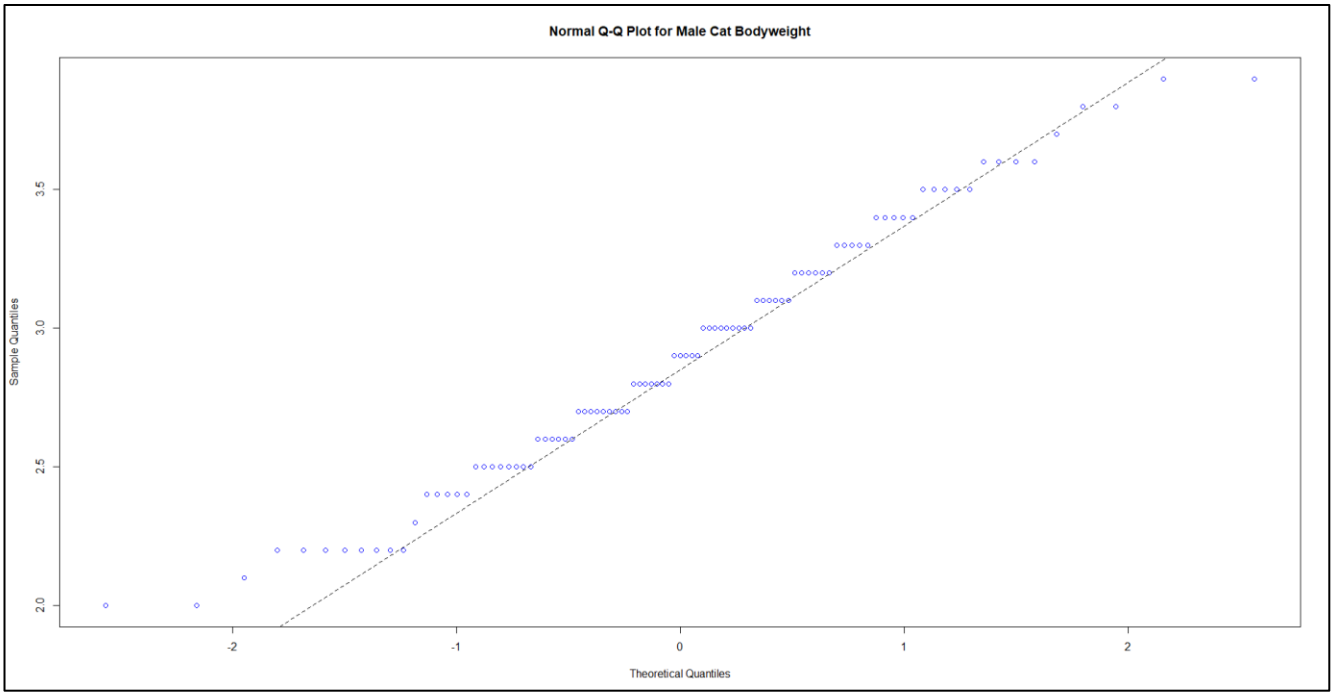
Now, since the p-value of the test is less than 0.05, we reject the null hypothesis. These results provide the means of both samples where there is no difference between the means and so it is seen that the bodyweight of sample A i.e., the male cat sample is different than the bodyweight of sample B i.e., the female cat sample. Therefore, the answer to our question is that male and female cat samples do not have the same bodyweight and the bodyweight of male cat sample is higher than the bodyweight of female cat sample with a p-value of 8.831e-15.

Data Visualizations for Testing:

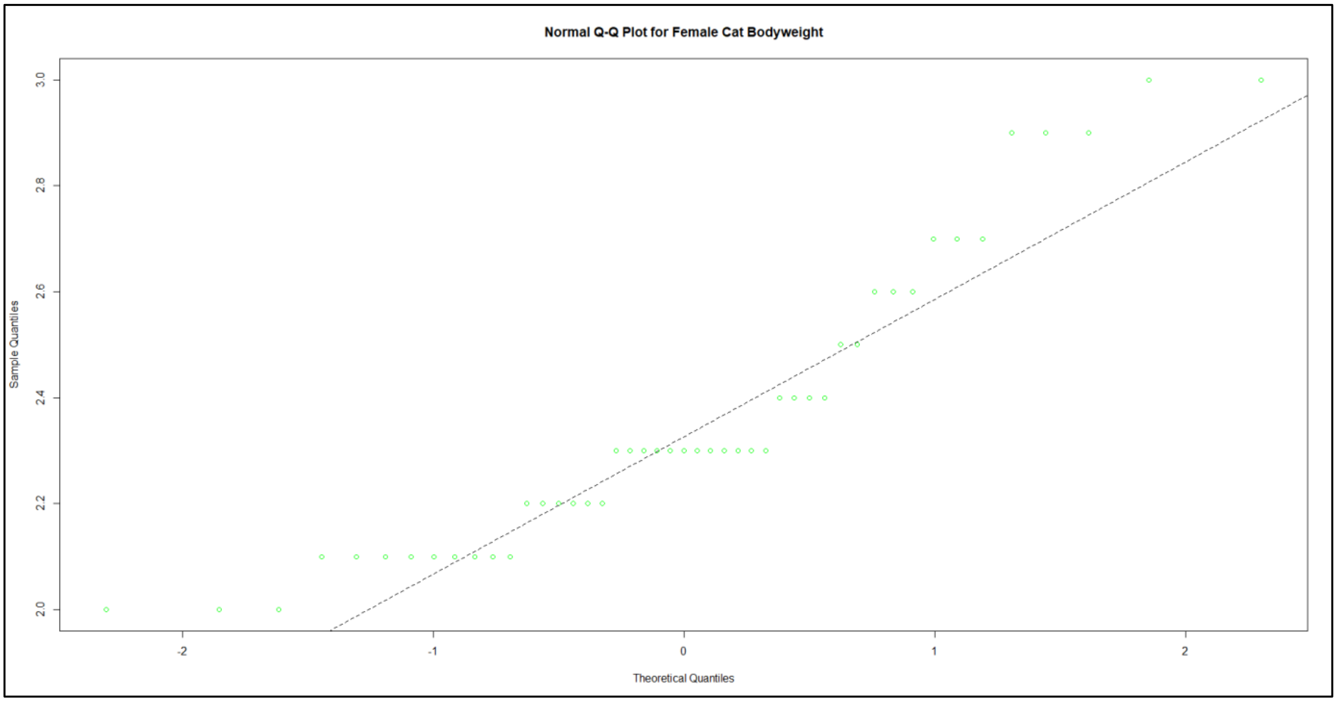
The normal q-q plot was created for visualizations of testing for the bodyweight of male cat sample and the bodyweight of female cat sample. Another visualization created was a boxplot which depicts the group of people along with their bodyweight value. This helps in understanding the statistical value of the bodyweight attribute for both the groups of people.

Output:

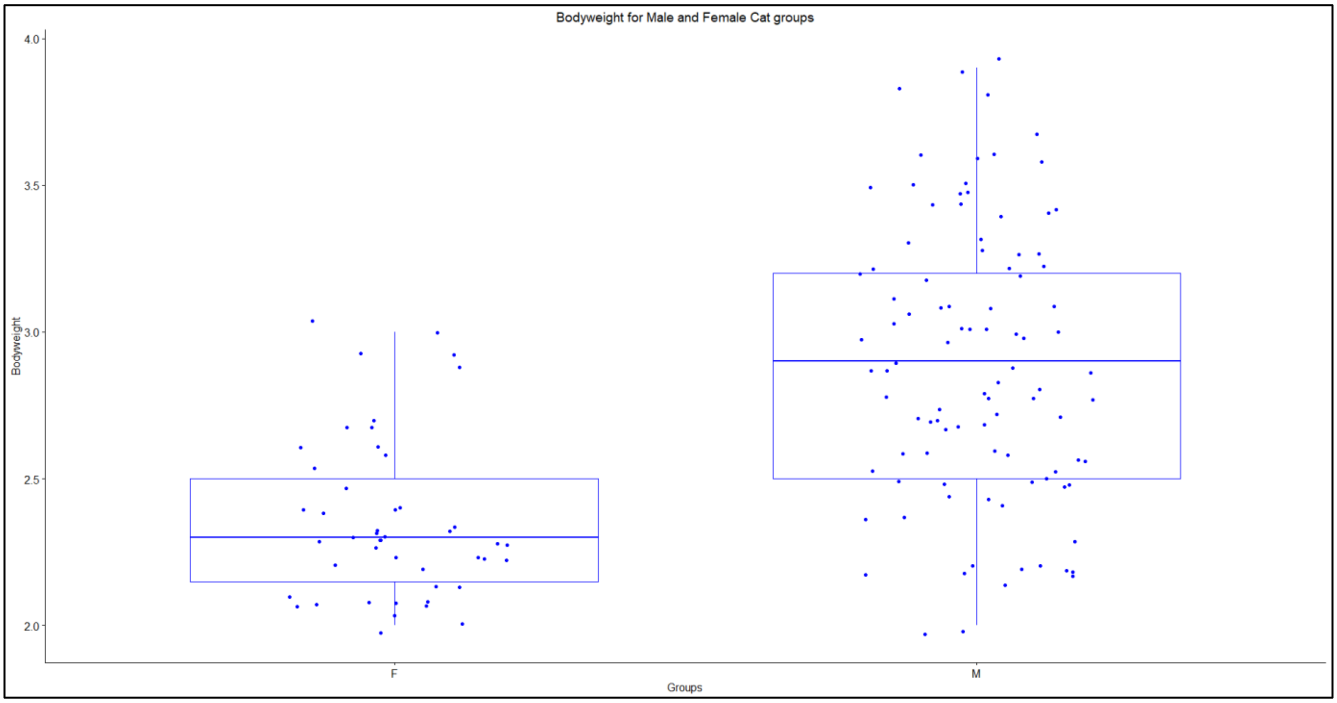
1. Graph 1: Normal Q-Q Plot
   a. Male Cat Bodyweight



b. Female Cat Bodyweight



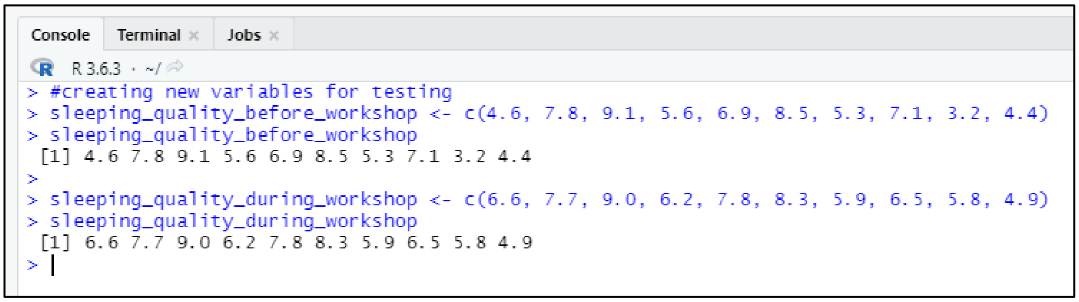
2. Graph 2: Boxplot



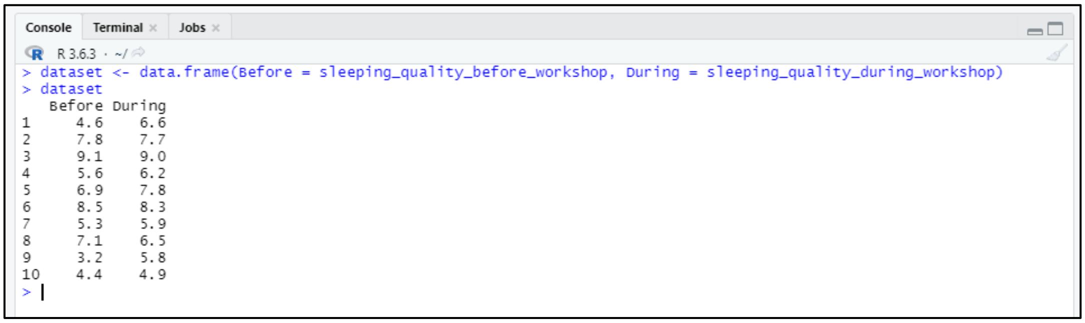
Part 2
Data Analysis

In the part 2 of the assignment, two samples of data were given which represents the sleeping quality before meditation and the sleeping quality during meditation. These samples of data were stored in a variable of the data frame. The two samples of data was then merged into one dataset with the help of the data.frame() function. The column names were named as ‘Before’ and ‘During’ which would help in further visualizations and analysis. The samples of data and the entire dataset created is as follows.

Output:



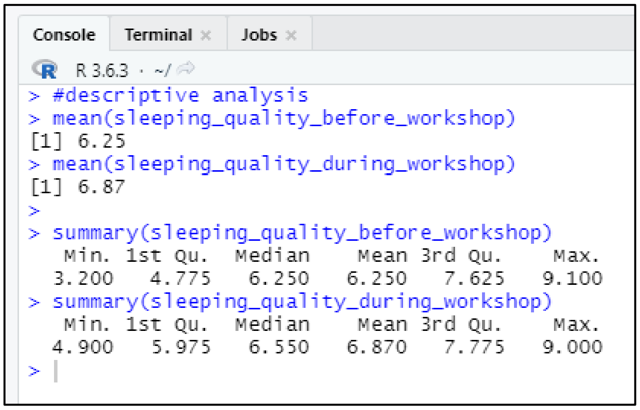
```
> #creating new variables for testing
> sleeping_quality_before_workshop <- c(4.6, 7.8, 9.1, 5.6, 6.9, 8.5, 5.3, 7.1, 3.2, 4.4)
> sleeping_quality_before_workshop
 [1] 4.6 7.8 9.1 5.6 6.9 8.5 5.3 7.1 3.2 4.4
>
> sleeping_quality_during_workshop <- c(6.6, 7.7, 9.0, 6.2, 7.8, 8.3, 5.9, 6.5, 5.8, 4.9)
> sleeping_quality_during_workshop
 [1] 6.6 7.7 9.0 6.2 7.8 8.3 5.9 6.5 5.8 4.9
> 
```



```
> dataset <- data.frame(Before = sleeping_quality_before_workshop, During = sleeping_quality_during_workshop)
> dataset
   Before During
1     4.6    6.6
2     7.8    7.7
3     9.1    9.0
4     5.6    6.2
5     6.9    7.8
6     8.5    8.3
7     5.3    5.9
8     7.1    6.5
9     3.2    5.8
10    4.4    4.9
> 
```

The descriptive analysis was performed for the respective attributes which were created above. Summary function helps in getting all the statistical values which returns the min, max, mean, median and range values. The output of the descriptive analysis is as shown below.

Output:



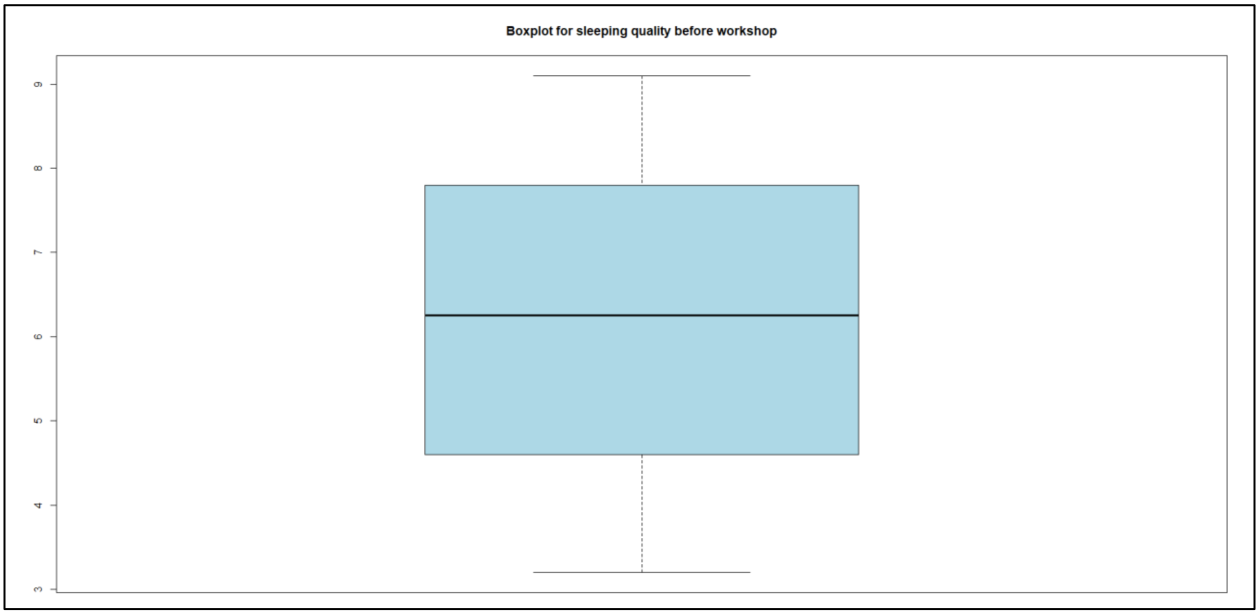
```
> #descriptive analysis
> mean(sleeping_quality_before_workshop)
[1] 6.25
> mean(sleeping_quality_during_workshop)
[1] 6.87
> 
> summary(sleeping_quality_before_workshop)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  3.200   4.775   6.250   6.250   7.625   9.100 
> summary(sleeping_quality_during_workshop)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  4.900   5.975   6.550   6.870   7.775   9.000 
> 
```

Data Visualization

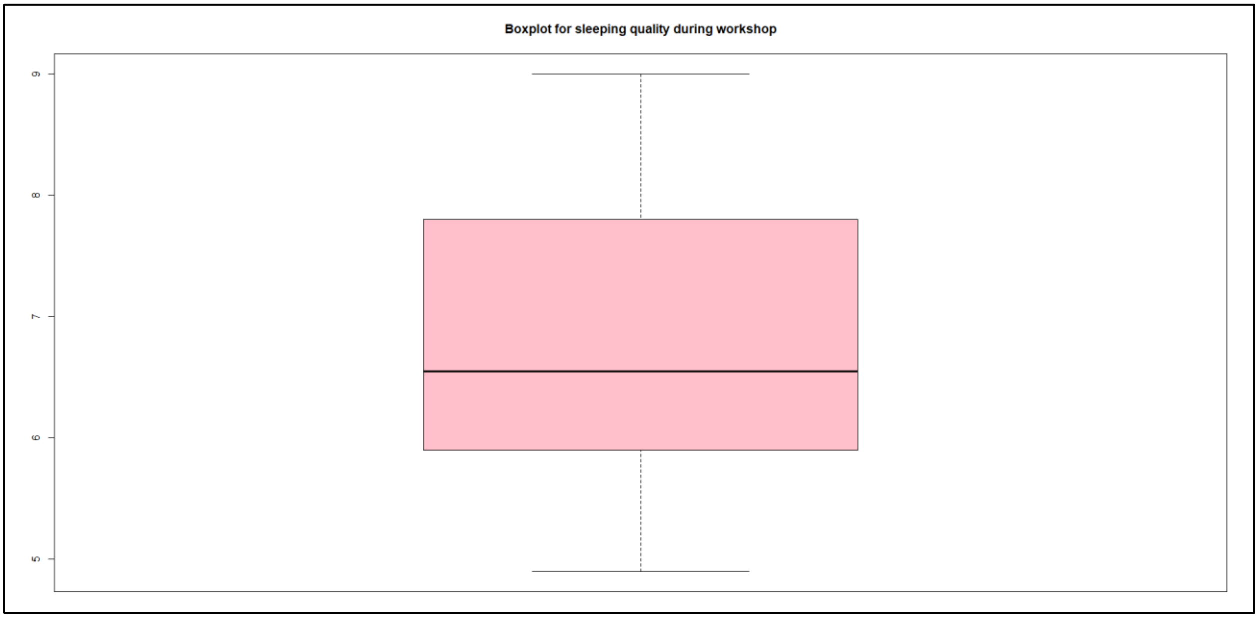
In this section, we plotted the data samples which were created using the boxplot and bar plot for the two samples of data i.e., sleeping quality before workshop and sleeping quality during workshop. The outputs of the same is as shown below.

Output:

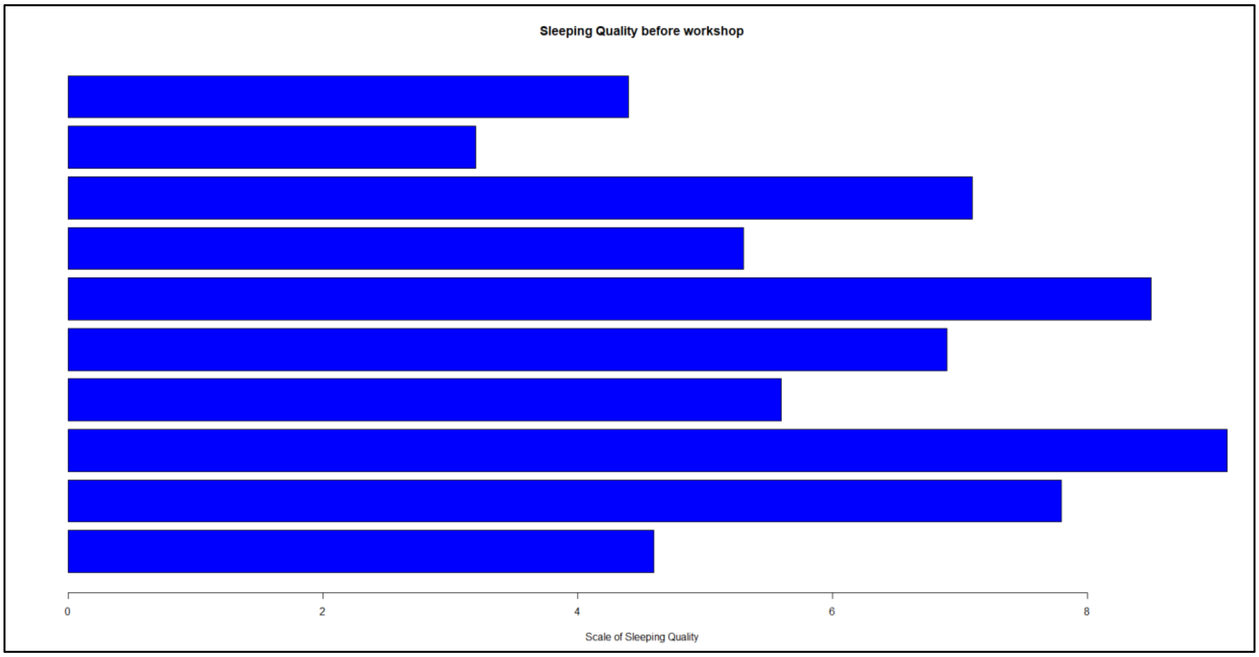
1. Boxplot for sleeping quality before workshop



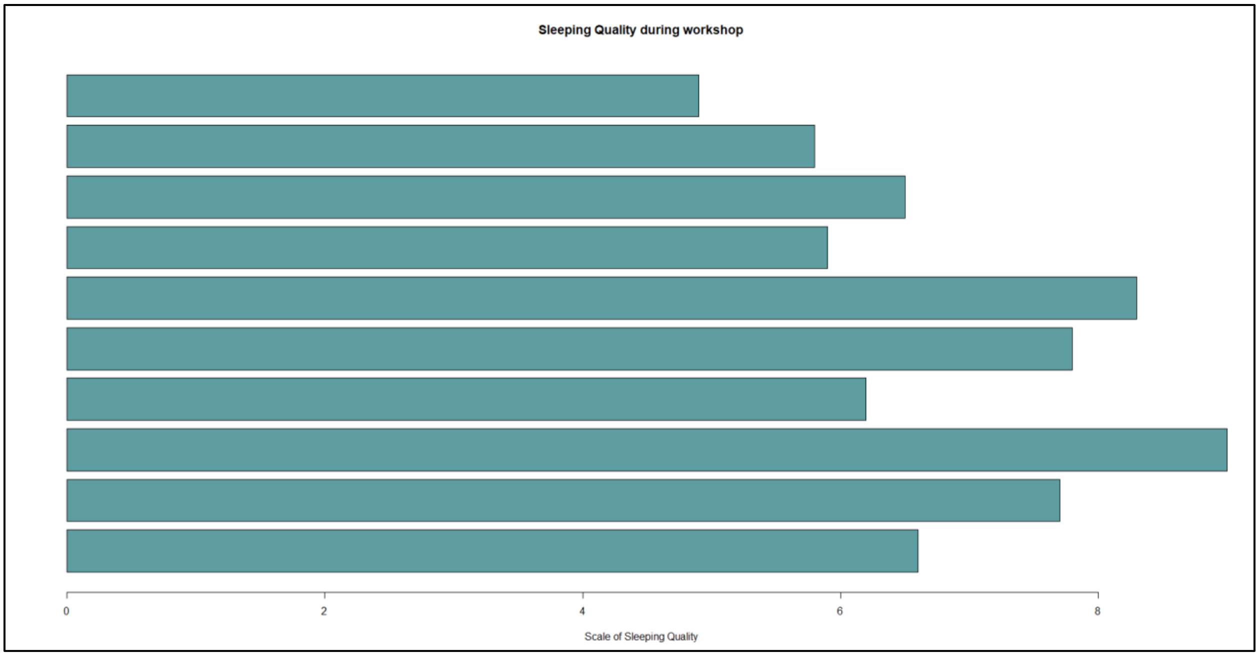
2. Boxplot for sleeping quality during workshop



3. Bar plot for sleeping quality before workshop



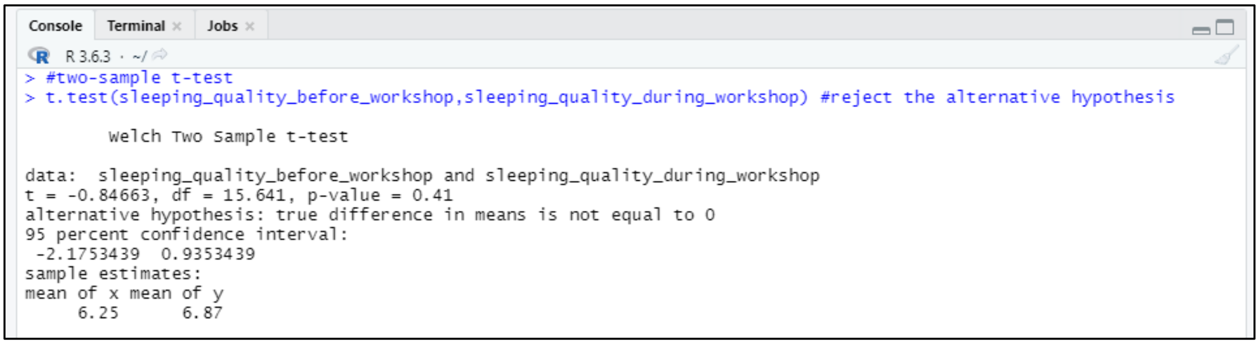
4. Bar plot for sleeping quality during workshop



Hypothesis Testing

For the part 2 of the assignment, a two-sample t-test was performed to know whether meditation has an effect on sleep quality or not. This analysis and testing was done using the t.test() in R and the output of which is as shown below.

Output:



```
> #two-sample t-test
> t.test(sleeping_quality_before_workshop,sleeping_quality_during_workshop) #reject the alternative hypothesis

        Welch Two Sample t-test

data:  sleeping_quality_before_workshop and sleeping_quality_during_workshop
t = -0.84663, df = 15.641, p-value = 0.41
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.1753439  0.9353439
sample estimates:
mean of x mean of y 
     6.25      6.87 
```

The output from the above snapshot gives us information about all the statistical values which would help in getting to a conclusion. Now, since the p-value of the test is greater than 0.05, we reject the null hypothesis and the answer to our question is that yes, it is true that meditation does improves the sleeping quality. The table below shows all the statistical values of the two-sample t-test.

t-test statistic	Degrees of freedom	p-value	95% confidence interval	Mean of x and y
-0.84663	15.641	0.41	[-2.1753439 0.9353439]	mean of x: 6.25 mean of y: 6.87

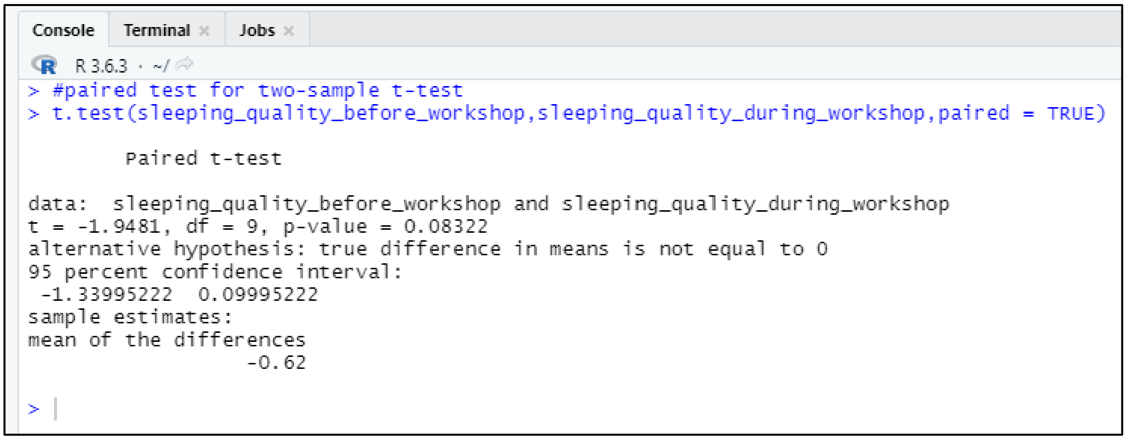
Now, since the p-value of the test is greater than 0.05, we fail to reject the null hypothesis that our data is normally distributed and so we can proceed with the paired t-test. Also, the level of significance was considered as 0.05 which is 95% confidence, but it was also observed that changing the level of significance from 0.05 to 0.1 does not change our conclusion.

Paired test for two-sample t-test:

Paired test is performed when we need to compare the means of two groups of observation where the observations are randomly assigned to each of the two groups. If the two samples are given, the observations of one sample can be paired with the observation of the other sample. Paired t-test can be used in making the observations on the same sample before and after an event.

Now, since, the two-sample t-test fails to reject the null hypothesis, paired t-test can be performed here. As, paired t-test can be performed on the same sample for an event before and after, the dataset we worked on also has the same events of dataset and so we can apply a paired t-test on the dataset. The output of the paired t-test for the dataset is as shown below.

Output:



```
> #paired test for two-sample t-test
> t.test(sleeping_quality_before_workshop,sleeping_quality_during_workshop,paired = TRUE)

        Paired t-test

data:  sleeping_quality_before_workshop and sleeping_quality_during_workshop
t = -1.9481, df = 9, p-value = 0.08322
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.33995222  0.09995222
sample estimates:
mean of the differences
                  -0.62

> |
```

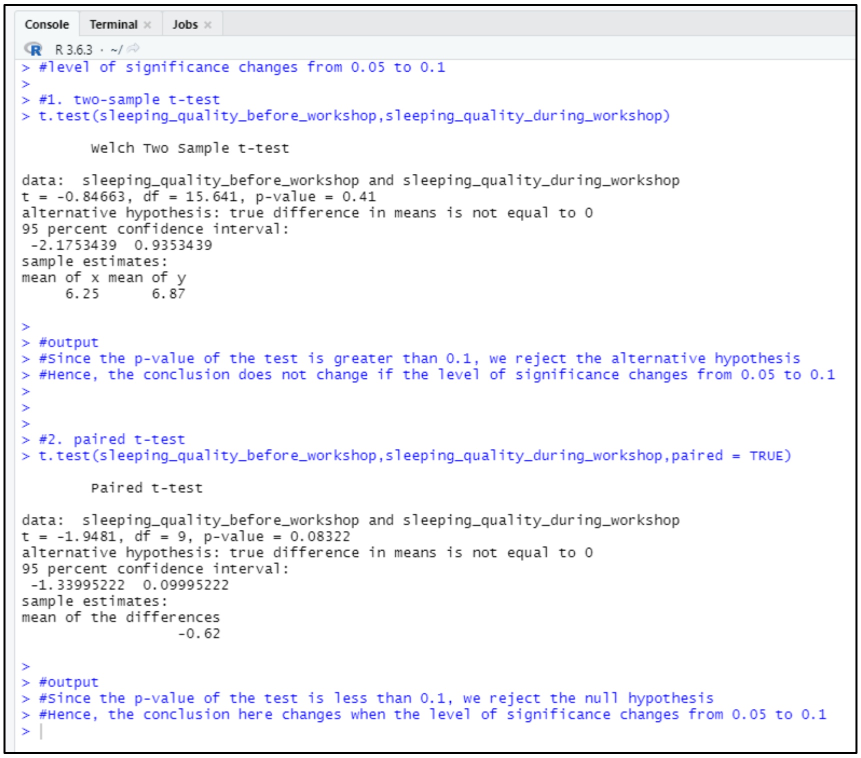
From the output it is observed that the p-value of the test is greater than 0.05, which means we reject the alternative hypothesis. The output also gives us the value of the mean of the differences which is -0.62. Therefore, we can conclude that the sleeping quality before meditation is different than that from the sleeping quality after meditation.

The answer to the question that whether this should be a paired test or not is that yes, it should be a paired test because if the two samples are given, then the observation of one sample can be paired with the observation of the other sample and this test can be used in making observations on the same sample before and after an event.

Level of Significance:

The level of significance is the measurement of statistical significance when the null hypothesis is implicit to be established or discarded. It determines the statistical significance of the result of the null hypothesis to be false. The level of significance known as alpha too was kept as 0.05 for the above tests but it was also observed that changing the value of alpha from 0.05 to 0.1 does not hamper our conclusion that we made earlier. This can be observed in the output below.

Output:

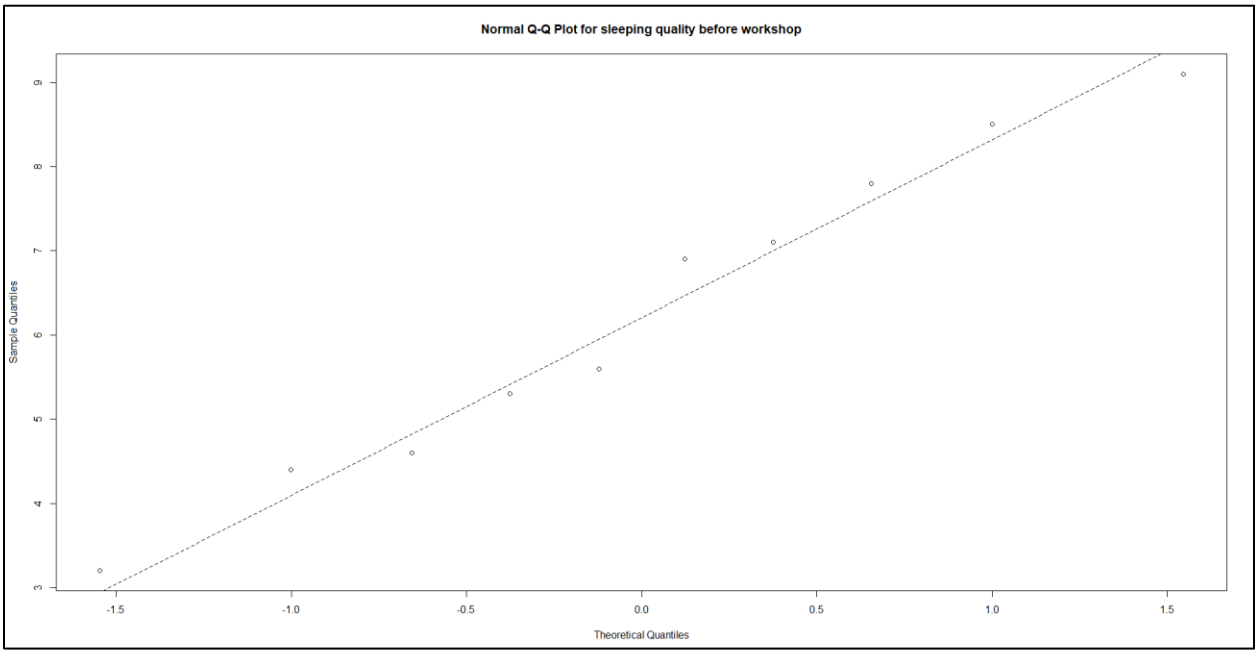


```
> #level of significance changes from 0.05 to 0.1
>
> #1. two-sample t-test
> t.test(sleeping_quality_before_workshop,sleeping_quality_during_workshop)

        Welch Two Sample t-test

data:  sleeping_quality_before_workshop and sleeping_quality_during_workshop
t = -0.84663, df = 15.641, p-value = 0.41
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.1753439  0.9353439
sample estimates:
mean of x mean of y
     6.25      6.87

>
> #output
> #Since the p-value of the test is greater than 0.1, we reject the alternative hypothesis
> #Hence, the conclusion does not change if the level of significance changes from 0.05 to 0.1
>
>
>
> #2. paired t-test
> t.test(sleeping_quality_before_workshop,sleeping_quality_during_workshop,paired = TRUE)

        Paired t-test

data:  sleeping_quality_before_workshop and sleeping_quality_during_workshop
t = -1.9481, df = 9, p-value = 0.08322
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.33995222  0.09995222
sample estimates:
mean of the differences
                  -0.62

>
> #output
> #Since the p-value of the test is less than 0.1, we reject the null hypothesis
> #Hence, the conclusion here changes when the level of significance changes from 0.05 to 0.1
> 
```

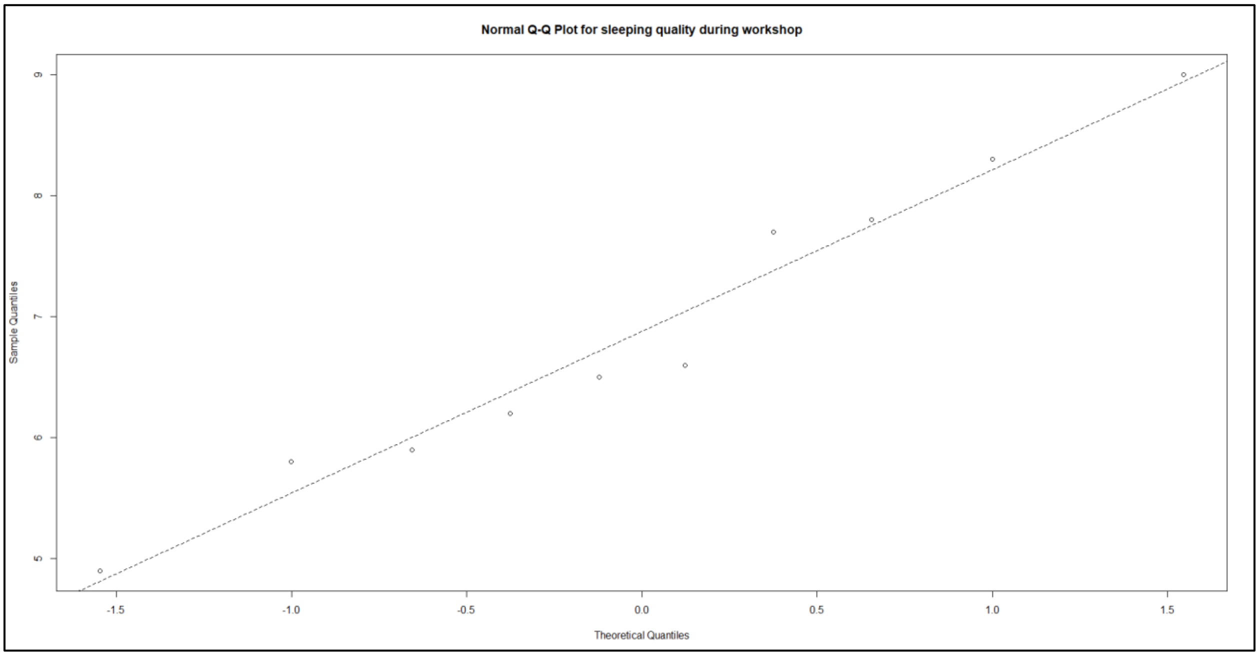
Therefore, the conclusion does not change if the level of significance changes from 0.05 to 0.1.

Data Visualizations for Testing:

1. Graph 1: Normal Q-Q plot for sleeping quality before workshop



2. Graph 2: Normal Q-Q plot for sleeping quality during workshop



Summary

The tasks in this assignment consisted of two parts where in the first part the aim was to find whether the bodyweight of the male cat sample and female cat sample is same or not and in the second part we had to test and find whether the sleeping quality before workshop is different or not than the sleeping quality during workshop. In the first part, we test with respect to the two-sample t-test with unequal variance and in the second part, we test with respect to the two-sample t-test and paired test to understand the before and after scenarios of the sleeping quality.

The analysis of the first part of the assignment speaks about whether the bodyweight of the male cat sample is different or same as compared to the bodyweight of the female cat sample. Here, from the dataset the samples were extracted as the testing and analysis needed to be done on two different groups of people for the same attributes. After the data samples were extracted, some visualizations were performed which gave us idea and understanding about what the data is like. Once the dataset was visualized, we began with the testing part where we performed two sample t-test and from the output it was observed and concluded that the bodyweight of the male cat sample is higher than the bodyweight of the female cat sample and hence both do not have the same bodyweight and we rejected the null hypothesis as the p-value of the test was greater than 0.05. Later, the data visualization for the testing was also performed where a normal q-q plot was plotted and then the bodyweight of the male and female group was plotted on the boxplot.

In the second part of the analysis, two samples of data were given which defines the sleeping quality before and during the meditation. Here, a data frame was created in which the data samples were stored on which the descriptive analysis and data visualizations were performed. For testing and analysis, two sample t-test was performed on the dataset which concluded that meditation does improves the sleeping quality since our alternative hypothesis was rejected as the p-value was greater than 0.05. Now, since in the two-sample t-test we fail to reject the null hypothesis, we can therefore proceed with the paired test. In the paired test, we test for the two events i.e., before and after the event. When performed the paired test, we observed that since the p-value of the test is greater than 0.05, we fail to reject the null hypothesis and thus we concluded that the sleeping quality is improved during the period of meditation.

The answer to the question whether it should be a paired test or not is that it should be a paired test because if two samples are given, then the observation of one sample can be paired with the observation of the other sample and this test can be used in making observations on the same sample before and after an event.