Code

AKSHAT VIJAYVARGIA/S3826627

REQUIRED PACKAGES

Hide

```
library(knitr)
library(ggplot2)
library(readr)
library(tidyr)
library(deductive)
library(validate)
library(Hmisc)
library(stringr)
library(lubridate)
library(outliers)
library(MVN)
library(MASS)
library(caret)
library(dplyr)
```

DATASET 1- Average Length Of Stay

This data is provided and collected by AIHW(Australian Institute of Health and Welfare), which was founded on 5 June'1987 and is a national agency, which collects and provides facts, information and statistics for Australia's health wealth and welfare.

Hide

```
library(readxl)
ALS <- read_excel("C:/Users/61422/OneDrive/Desktop/AA/average-length-of-stay-multil
evel-data.xlsx",
    skip = 12)
```

```
Expecting logical in O30023 / R30023C15: got '‡'Expecting logical in O30024 / R3002
4C15: got '‡'Expecting logical in O30025 / R30025C15: got '‡'Expecting logical in O
30026 / R30026C15: got '‡'Expecting logical in O30027 / R30027C15: got '‡'Expecting
logical in O30028 / R30028C15: got '‡'New names:
* `` -> ...9
* `` -> ...11
* `` -> ...13
* `` -> ...15
* `` -> ...17
* ...
```

After importing, we will preprocess the data.While preprocessing the data, I noticed that a variable was not in right datatype, so with the help of as.numeric() function I changed it's type. And just to be sure, I used str() function to check the datatype.

```
NAs introduced by coercion
```

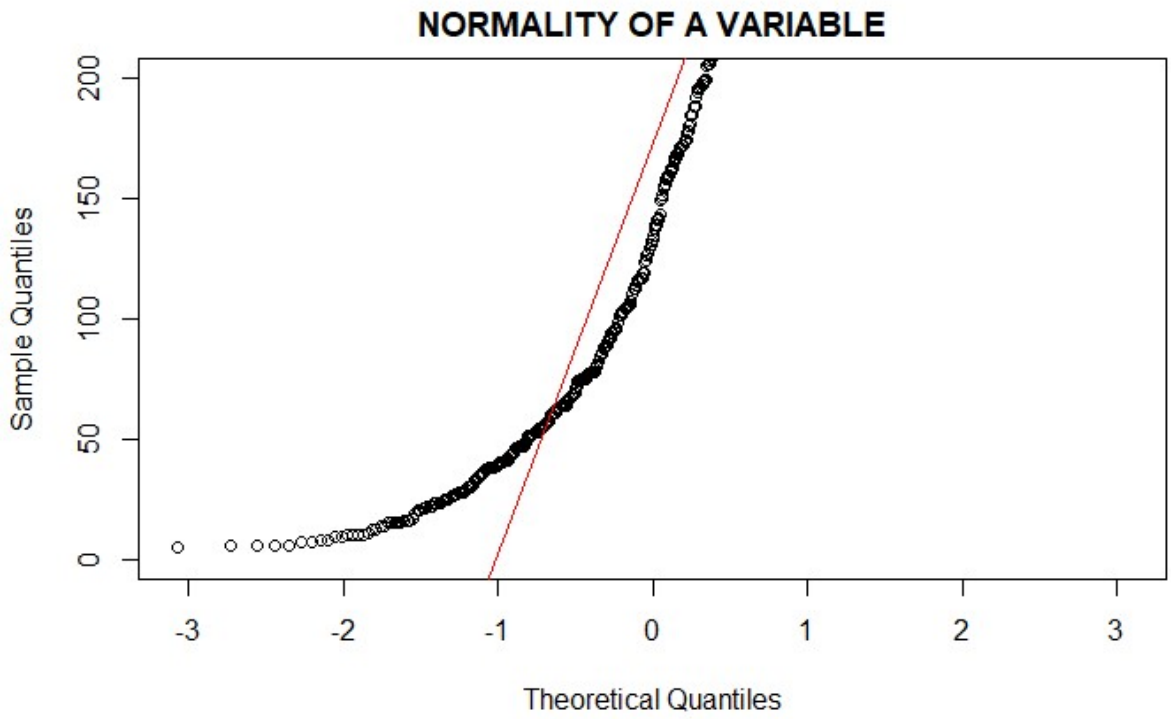
```
tibble [499 x 3] (S3: tbl_df/tbl/data.frame)
 $ State                      : chr [1:499] "NSW" "NSW" "NSW" "NSW" ...
 $ Local Hospital Network (LHN): chr [1:499] "South Western Sydney" "South Western
Sydney" "South Western Sydney" "South Western Sydney" ...
 $ Total number of stays      : num [1:499] 1105 1095 1090 1108 1179 ...
```

DATASET 1 - ANALYSIS AND VISUALIZATION

Lets visualize the Average Length of Stay by using qqnorm(). We are using qqnorm() because it will not only make it easier for us understand but also will tell us about the normal distribution of the data.

Hide

```
qqnorm(DF$`Total number of stays`, main="NORMALITY OF A VARIABLE",ylim = c(0,200))
qqline(DF$`Total number of stays`, col="red")
```



According to the graph, our data is not normally distributed, but it is a large data with more than 50 observations(around 500 observations) and ttest is valid for large samples from non-normal distributions. Now lets jump into our one sample ttest. To have the clear picture.

For database 1, we will use One sample Ttest. H0-> Mean length of stay is not different from

4.5days HA-> Mean length of stay is different from 4.5days.

Hide

```
t.test(DF$`Total number of stays`,mu=4.5)
```

```
    One Sample t-test

data:  DF$`Total number of stays`
t = 16.648, df = 473, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 4.5
95 percent confidence interval:
 204.0744 257.4910
sample estimates:
mean of x
 230.7827
```

The result of the One sample Ttest is t is 16.648, degree of freedom is 473 and p value is <2.2e-16 or very less than our significant value 0.05.According to the result, we can reject the null hypothesis(H0) and with 95% confidence we can conclude that the mean length of stay is different from 4.5days.

DATASET 1- DISCUSSION

In this investigation, we found that the average length of stay of south western Sydney that we were all claiming to be 4.5 days, is not true. As per the results of the hypothesis testing the mean length of stay in south western Sydney is different from 4.5 days. The strength of my investigation is that I picked the data from AIHW(Australian Institute of Health and Welfare), which is a Government statutory agency and jurisdiction of Commonwealth of Australia, so my dataset is pure and authentic. The limitations of the investigation is that I am restricted to only one factor i.e Average length of stay(days), but to evaluate the number of patients and their stay includes price difference, quality of service, behavior of doctors etc.Therefore my investigation may not fully say that this 4.5 days mean length of stay is right.

DATASET 2- Tutorials

This data is provided by RMIT University for our better understanding to the real world tasks. The data has some variable factors, which can affect the variable Student performance with different degrees.

Hide

```
library(readr)
TUT <- read_csv("C:/Users/61422/OneDrive/Desktop/Assignment 4b-1.csv")
```

```
Parsed with column specification:
cols(
  Gender = [32mcol_double()[39m,
  IQ = [32mcol_double()[39m,
  Profession = [32mcol_double()[39m,
  Advice = [32mcol_double()[39m,
  School = [32mcol_double()[39m,
  `Score without after tutorial` = [32mcol_double()[39m,
  `Score without before tutorial` = [32mcol_double()[39m
)
```

Before going further, lets check the datatypes.

Hide

```
str(TUT)
```

```
tibble [1,290 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Gender                      : num [1:1290] 1 2 1 1 2 1 2 1 1 1 ...
 $ IQ                          : num [1:1290] 91 117 105 110 101 93 111 74 96 111
...
 $ Profession                  : num [1:1290] 3 3 3 3 3 3 5 2 3 5 ...
 $ Advice                      : num [1:1290] 4 2 1 4 2 6 6 2 2 5 ...
 $ School                      : num [1:1290] 1 1 1 1 1 1 1 1 1 1 ...
 $ Score without after tutorial : num [1:1290] 42 38 43 37 35 41 45 50 44 44 ...
 $ Score without before tutorial: num [1:1290] 50 13 27 44 35 55 53 19 35 27 ...
 - attr(*, "spec")=
  .. cols(
  ..   Gender = [32mcol_double()[39m,
  ..   IQ = [32mcol_double()[39m,
  ..   Profession = [32mcol_double()[39m,
  ..   Advice = [32mcol_double()[39m,
  ..   School = [32mcol_double()[39m,
  ..   `Score without after tutorial` = [32mcol_double()[39m,
  ..   `Score without before tutorial` = [32mcol_double()[39m
  .. )
```

Gender is factor yet in numerical, we will use as.factor() to convert it.

Hide

```
TUT$Gender<-as.factor(TUT$Gender)

# CHECKING STR() AGAIN
str(TUT)
```

```
tibble [1,290 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Gender                       : Factor w/ 2 levels "1","2": 1 2 1 1 2 1 2 1 1 1
...
 $ IQ                           : num [1:1290] 91 117 105 110 101 93 111 74 96 111
...
 $ Profession                   : num [1:1290] 3 3 3 3 3 3 5 2 3 5 ...
 $ Advice                       : num [1:1290] 4 2 1 4 2 6 6 2 2 5 ...
 $ School                       : num [1:1290] 1 1 1 1 1 1 1 1 1 1 ...
 $ Score without after tutorial : num [1:1290] 42 38 43 37 35 41 45 50 44 44 ...
 $ Score without before tutorial: num [1:1290] 50 13 27 44 35 55 53 19 35 27 ...
 - attr(*, "spec")=
  .. cols(
  ..   Gender = [32mcol_double()[39m,
  ..   IQ = [32mcol_double()[39m,
  ..   Profession = [32mcol_double()[39m,
  ..   Advice = [32mcol_double()[39m,
  ..   School = [32mcol_double()[39m,
  ..   `Score without after tutorial` = [32mcol_double()[39m,
  ..   `Score without before tutorial` = [32mcol_double()[39m
  .. )
```

Before jumping directly to conclusion, lets first see how are summary statistics of both the scores before tutorial and after tutorial are.

Hide

```
# SUMMARY STATISTICS
summary(TUT$`Score without after tutorial`,na.rm=TRUE)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  33.00   37.00   41.00   41.17   44.00   55.00 
```

Hide

```
summary(TUT$`Score without before tutorial`,na.rm=TRUE)
```

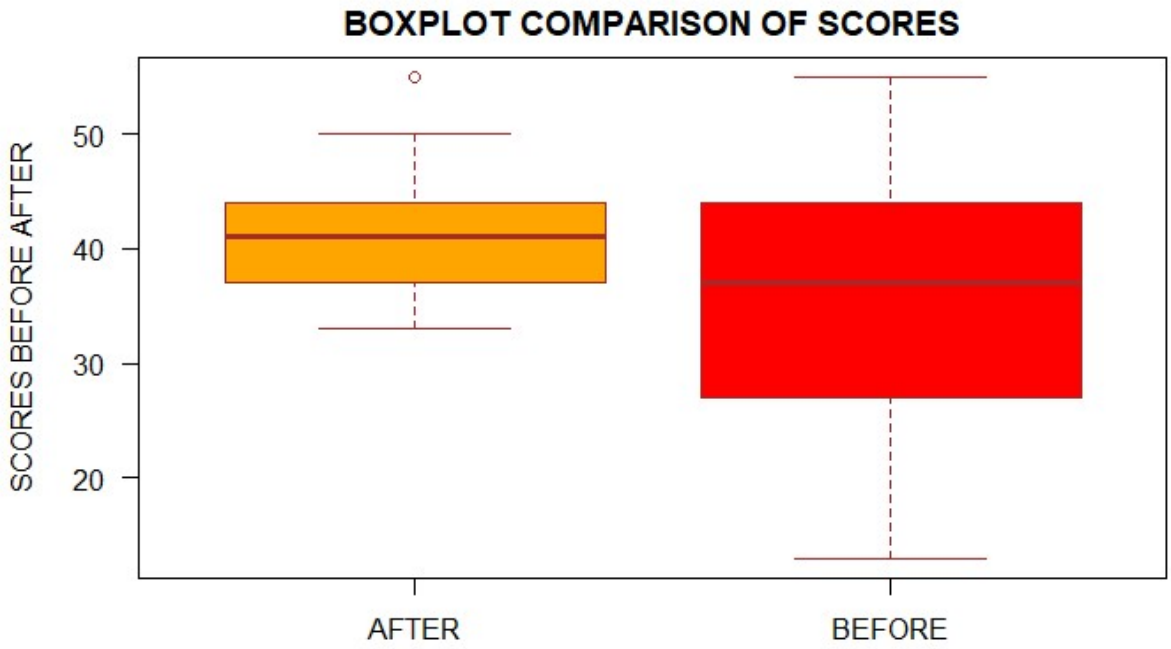
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  13.00   27.00   37.00   35.82   44.00   55.00 
```

DATASET 2 - ANALYSIS AND VISUALIZATION

After calculating Summary() and SD() of both the scores before and scores after. Now, our data is set for visualization. We will use box plot to compare the scores.

Hide

```
boxplot(TUT$`Score without after tutorial`,TUT$`Score without before tutorial`, mai
n = "BOXPLOT COMPARISON OF SCORES", ylab="SCORES BEFORE AFTER", at = c(1,2), names
= c("AFTER","BEFORE"), las = 1, by=1, col = c("orange","red"), border = "brown", ho
rizontal = FALSE, notch = FALSE )
```



According to the box plot and summary statistics, the mean of Scores after tutorial is slightly greater than the mean Scores before tutorial.As of now, with the box-plot visualization, we can see that the scores are good after tutorial than before tutorial, we can vaguely say that, but to be precise We will use hypothesis testing. I am using Paired T.test for both the variables.

DATASET 2 - HYPOTHESIS TESTING

H0= There is no difference in the scores. HA= There is difference in the scores.

Hide

```
t.test(TUT$`Score without after tutorial`,TUT$`Score without before tutorial`,paire
d = TRUE, alternative = "two.sided")
```

```
    Paired t-test

data:  TUT$`Score without after tutorial` and TUT$`Score without before tutorial`
t = 19.144, df = 1289, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.802104 5.898671
sample estimates:
mean of the differences
               5.350388
```

The result of the test, shows our p-value is less than 2.2e-16 and 95% confidence interval is from 4.802104 to 5.898671. According to the result, we know that p-value is less than our significance level which is 0.05alpha and 95% confidence interval misses our null hypothesis(H0). Therefore, we can reject null hypothesis(H0). According to the test result above, we can say that there is difference in the scores before and after tutorials. We can conclude that students, who attended the lectures score more than students, who did not attend the lectures.

DATASET 2 - DISCUSSION

In this investigation, we found that the scores of students, who attended the lectures are more than the scores of students, who did not attend the lectures. The strength of my investigation is that, I got data from RMIT University, which one of the prestigious university in Australia. So, I can say that the dataset is pure and authentic. The limitations of the investigation is that I am restricted to only two factors i.e Scores before and scores after tutorials, but to evaluate we should have considered the student’s IQ, course difficulty etc..Therefore my investigation may not fully say who scores more.