Assignment 2 : Applied Analytics

Code

Yogesh Haresh Bojja (s3789918)

INTRODUCTION :

We have been provided with the data from Australian institute of health and welfare which calculates Average length of stay(ALOS). This data includes different 'Reporting unit' and its corresponding 'Reporting unit type' located across different states in Australia. Each reporting unit is categorised into a different 'Peer group'. This data shows the 'Number of overnight stays' and 'Total overnight patients bed' in the 'Time period' of 1 year in each 'Category' of the reporting unit. Average length of stay(ALOS) is calculated as the 'Total overnight patients bed' divided by the 'Number of overnight stays'.

PROBLEM STATEMENT :

This investigation will seek to understand if there is any statistical significant difference in the average length of stay (ALOS) between large and medium hospitals which might make patients choose one over the other. We are going to investigate this with the help of the following procedure. First we are going to check whether two samples are normally distributed or not by analyzing Q-Q plot, based on the analysis of the Q-Q plot we will take into consideration whether 'Central Limit theorem' proves that the distribution is normal or not considering the sample sizes. Once the data is proved to be normally distributed we can apply the 'Levene's test' to compare the variances of the samples and based on that we decide which t-test is to be performed for testing hypothesis. We are going to test the hypothesis for 95% confidence hence our area of significance(α) is 0.05.
Let group 1 be for Large Hospital and group 2 be for Medium Hospital.Thus the hypothesis can be written as :
$H_0 \rightarrow \mu_1 - \mu_2 = 0$
$H_A \rightarrow \mu_1 - \mu_2 \neq 0$
Two tailed testing will be applicable here.

DATA :

1) Loading data and dividing the data :
Data is taken from Australian institute of health and welfare. Data is loaded with read_excel() function from readxl package.

Hide

```
# Loading the data
library(readxl)
dataset <- read_excel("average-length-of-stay-multilevel-data.xlsx", skip = 12)
```

Columns I,K,M,O,Q,S are empty columns in the excel sheet. Hence we get warnings after executing read_excel(). To eliminate those we are negating indexes of those columns from our dataset. In this way the unwanted columns are removed. We have created 2 datasets one contains all the observations of Large Hospitals while the other dataset contains all the observations of the Medium Hospitals with the help of subset function. Subset function selects the observations in the dataset based on the condition specified.

Hide

```
dataset <- dataset[-c(9,11,13,15,17,19)]
#divide the data
large_hosp <- subset(dataset, dataset$`Peer group`=='Large hospitals')
medium_hosp <- subset(dataset, dataset$`Peer group`=='Medium hospitals')
head(large_hosp, 2)
```

Reporting unit <chr>	**Reporting unit type** <chr>	**State** <chr>	**Local Hospi** <chr>
Albury Wodonga Health [Albury Campus]	Hospital	NSW	Albury Wodo
Albury Wodonga Health [Albury Campus]	Hospital	NSW	Albury Wodo

2 rows | 1-4 of 13 columns

Hide

```
head(medium_hosp, 2)
```

Reporting unit <chr>	**Reporting unit type** <chr>	**State** <chr>	**Local Hospital Network (LHN)** <chr>	**F** <c
Armidale Hospital	Hospital	NSW	Hunter New England	M
Armidale Hospital	Hospital	NSW	Hunter New England	M

2 rows | 1-5 of 13 columns

2) Handling missing values :

Both the datasets contain NP and '-' values. We have removed the NP values of both the dataset by specifying the condition. Even after removing NP's we still have '-' values remaining in the dataset. As we require our 'Average length of stay(days)' to be numeric for the further analysis we cast it to numeric with as.numeric(). While casting '-' gets coerced to NA hence we remove them from the dataset by specifying the condition is.na(). is.na() identifies whether the value is NA or not and returns true if it's NA else returns false.

Hide

```
# removing NA
large_hosp <- large_hosp[large_hosp$`Average length of stay (days)`!='NP',]
large_hosp$`Average length of stay (days)` <- as.numeric(large_hosp$`Average length of
stay (days)`)
```

```
NAs introduced by coercion
```

Hide

```
large_hosp <- large_hosp[!is.na(large_hosp$`Average length of stay (days)`),]
medium_hosp <- medium_hosp[medium_hosp$`Average length of stay (days)`!='NP',]
medium_hosp$`Average length of stay (days)` <- as.numeric(medium_hosp$`Average length
of stay (days)`)
```

```
NAs introduced by coercion
```

Hide

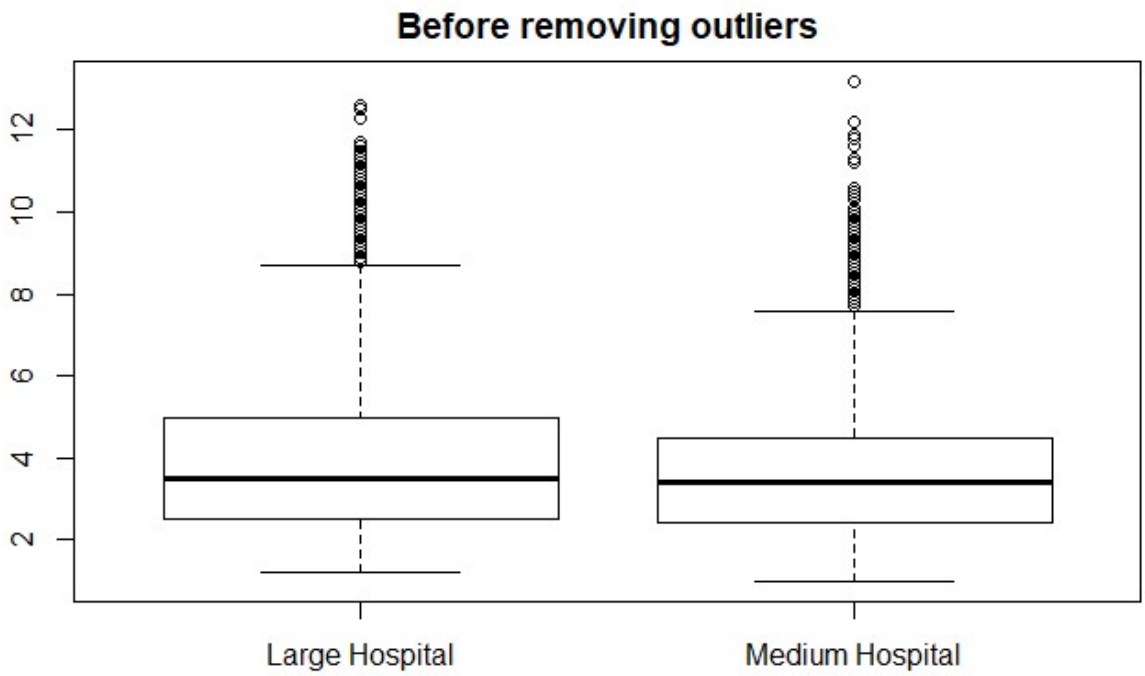
```
medium_hosp <- medium_hosp[!is.na(medium_hosp$`Average length of stay (days)`),]
```

3) Removing outliers:

Boxplot gives us the summary of the data visually. Hence by plotting a boxplot of 2 datasets we can see that there are many outliers. We need to remove them. We know that the efficient way to remove outliers is by eliminating all the values which falls above (Q3 + IQR * 1.5) and which falls below (Q1 - IQR * 1.5). As (Q1 - IQR * 1.5) is negative and there are no negative values in the 'Average length of stay(days)' we will only remove the values which exceed (Q3 + IQR * 1.5). We calculate Q3 with the help of quantile() function setting probs = 0.75. IQR is calculated with the function IQR(). Once we do this process and display a new boxplot of 'Average length of stay(days)' we can see that many outliers have been eliminated.

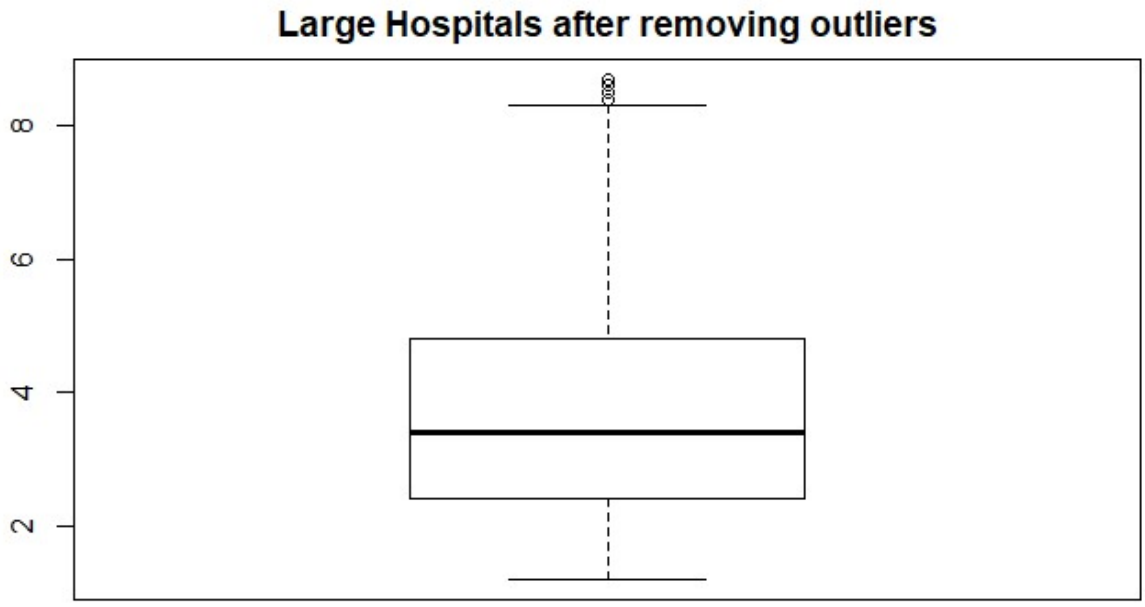
Hide

```
boxplot(large_hosp$`Average length of stay (days)`, medium_hosp$`Average length of sta
y (days)`, main = "Before removing outliers", names = c("Large Hospital", "Medium Hosp
ital"))
```



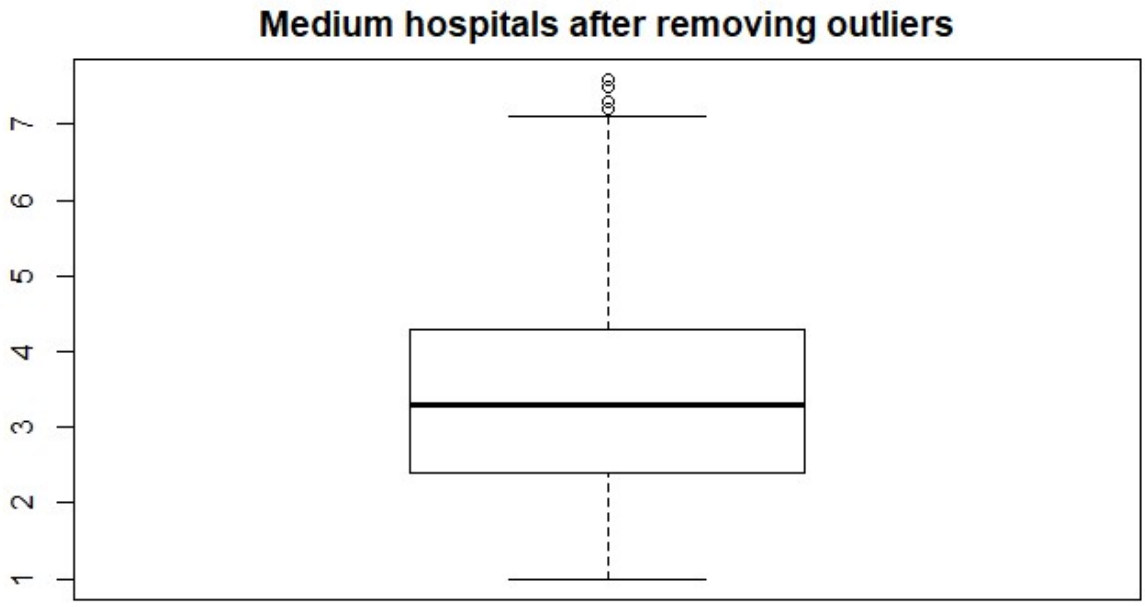
Hide

```
q3 <- quantile(large_hosp$`Average length of stay (days)`,probs = .75,na.rm = TRUE)
q <- IQR(large_hosp$`Average length of stay (days)`)
limit_max <- q3+1.5*q
large_hosp <- large_hosp[large_hosp$`Average length of stay (days)`<limit_max,  ]
boxplot(large_hosp$`Average length of stay (days)`, main = "Large Hospitals after remo
ving outliers")
```



Hide

```
q3 <- quantile(medium_hosp$`Average length of stay (days)`,probs = .75,na.rm = TRUE)
q <- IQR(medium_hosp$`Average length of stay (days)`)
limit_max <- q3+1.5*q
medium_hosp <- medium_hosp[medium_hosp$`Average length of stay (days)`<limit_max,  ]
boxplot(medium_hosp$`Average length of stay (days)`, main = "Medium hospitals after re
moving outliers")
```



DESCRIPTIVE STATISTICS AND VISUALIZATION :

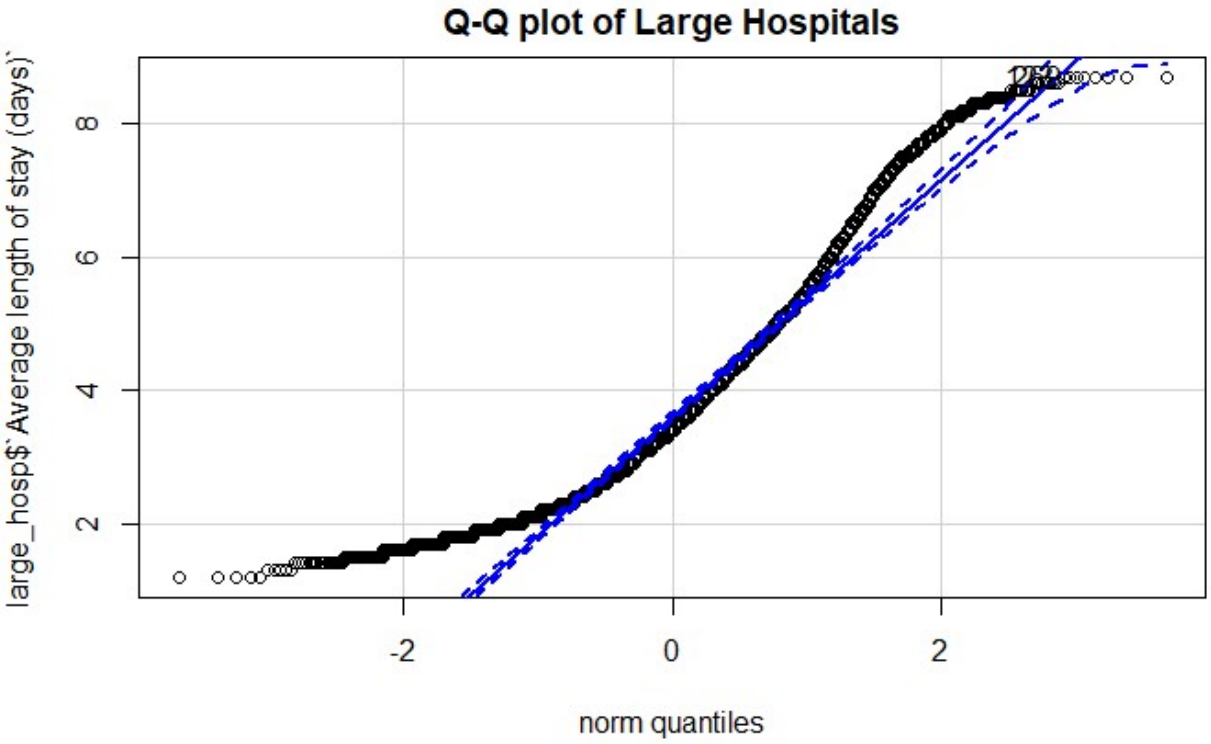
1) Analyzing the normality by Q-Q plot :

When the data is less than 30, two sample t-test and one sample t-test assumes that the sample is drawn from the normally distributed population. This assumption may be erroneous hence the best approach is to see data visually. Once the data is seen normal then we can proceed with t-test. For the perfectly normally distributed data the data points should be scattered on the diagonal, the shape deteriorates as the distribution gets away from the normal distribution. In such cases the shape may resemble ‘S’ or any other non-linear shape. From the figures below even though both the curves are slightly ‘S’ shaped we can see that the shape of the Q-Q plot for Large Hospitals is more ‘S’ shaped than that of the Medium Hospitals. This means that distribution for both the types of hospitals is not following the normal distribution.

Hide

```
qqPlot(large_hosp$`Average length of stay (days)`, dist="norm", main = "Q-Q plot of La
rge Hospitals")
```

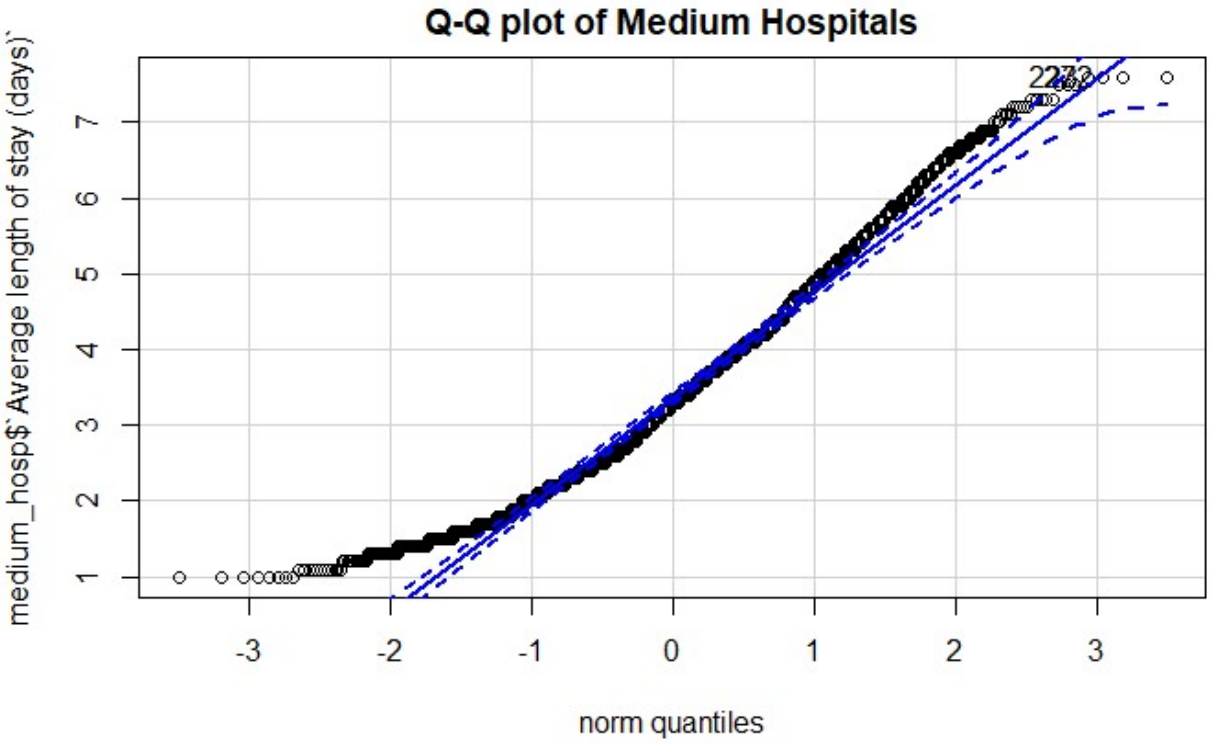
```
[1] 123 269
```



Hide

```
qqPlot(medium_hosp$`Average length of stay (days)`, dist="norm", main = "Q-Q plot of M
edium Hospitals")
```

```
[1] 223 272
```



HYPOTHESIS TESTING :

1) Central Limit Theorem :

Central Limit Theorem states that the sampling distribution of means is always distributed normally regardless of the underlying population when the sample size of the population > 30. In our case the sample size of Large Hospitals is 4265 and the sample size of Medium Hospitals is 2073 which are more than 30. Hence we can conclude that due to the large sample sizes according to the Central Limit theorem the distribution tends to be normally distributed and we can apply t-test for the hypothesis. Even though the Q-Q plot shows us that the data may not be normally distributed still we can perform t-test considering the Central Limit Theorem.

2) Homogeneity of variance :

$\mathbf{H_0 \rightarrow \sigma_1^2 = \sigma_2^2}$

$\mathbf{H_A \rightarrow \sigma_1^2 \neq \sigma_2^2}$

Assumption of equal variance between two samples is tested by Levene's test. Null hypothesis assumes the variances of both the groups are the same while we are rejecting it by proving the variances are not the same. By the help of Levene's test we find p-value, if the p-value is less than the area of significance then we need to reject the null hypothesis. If p-value is more than area of significance then Levene's test is not proved statistically significant. P-value of levene's test is 2.2e-16 which is less than the area of significance(0.05) hence its statistically significant. According to Levene's test we can say that variances of both the groups are different.

As Levenes test requires the columns to be numeric we have changed Large Hospitals to 1 and Medium Hospitals to 2 from the 'Peer group' column by replace() function. With as.numeric() the column gets casted to numeric datatype. factor() categorizes the data into two groups. LeveneTest() function is used to perform Levene's test. This function is present in car package.

Hide

```
df$`Peer group` <- replace(df$`Peer group`, df$`Peer group`=='Medium hospitals', 2)
df$`Peer group` <- replace(df$`Peer group`, df$`Peer group`=='Large hospitals', 1)
df$`Peer group` <- as.numeric(df$`Peer group`)
df$`Peer group` <- factor(df$`Peer group`)
leveneTest(df$`Average length of stay (days)` ~ df$`Peer group`, data=df, center = mea
n)
```

```
Levene's Test for Homogeneity of Variance (center = mean)
        Df F value    Pr(>F)
group    1  90.677 < 2.2e-16 ***
      6336
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3) Two sample t-test : Based on the Levene’s test we choose two sample t-test of unequal variance for hypothesis. With the help of this test we are trying to reject the null hypothesis. A Sample’s standard deviation is known, by the help of standard deviation and sample size we can calculate the degree of freedom. As we are looking here for areas of significance equal to 0.05 we find a 95% confidence interval. With the help of sample means, sample sizes and standard deviations we find the value of t. If t falls outside the 95% CI range then we have strong evidence to reject the null hypothesis else we fail to reject it. We get the value of t = 9.2315, degree of freedom = 4897 and 95% CI in range [0.289, 0.445]. As we are dealing with unequal variances var.equal = False. As the test is two tailed ‘alternative’ parameter is set as ‘two.sided’.

Hide

```
t.test(
  df$`Average length of stay (days)` ~ df$`Peer group`,
  data = df,
  var.equal = FALSE,
  alternative = "two.sided"
)
```

```

    Welch Two Sample t-test

data:  df$`Average length of stay (days)` by df$`Peer group`
t = 9.2315, df = 4897.1, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2891542 0.4450801
sample estimates:
mean in group 1 mean in group 2
       3.787667        3.420550
```

DISCUSSION :

A two-sample t-test was used to test for a significant difference between the mean Average length of stay(days) of Large hospitals and Medium hospitals. While the Average length of stay(days) for both the hospitals exhibited evidence of non-normality upon inspection of the normal Q-Q plot, the central limit theorem ensured that the t-test could be applied due to the large sample size in each group. The Levene's test of homogeneity of variance indicated that equal variance can not be assumed. The results of the two-sample t-test assuming unequal variance found a statistically significant difference between the mean Average length of stay(days) of Large hospitals and Medium hospitals, t(df=4897)=9.2315, p=2.2e-16, 95% CI for the difference in means [0.289 0.445].

Strength : Due to the large sample size our hypothesis could be tested properly as Central Limit Theorem overturned Q-Q plots verdict.

Limitation : t-test compares mean of only two samples.

Future scope : We have tested the hypothesis for ALOS between Large and Medium hospitals, in future we can test the hypothesis between different peer groups. We have omitted the NP values in the above testing but in future we can replace them with some other techniques like mean, median, or computed values from machine learning algorithms.

Conclusion : As 't' falls in the rejection region we can say that there is statistical significance to reject the null hypothesis. As the mean of Large hospitals(3.78) is greater than Medium Hospitals(3.42) we can conclude that stay at Large hospitals is more than that of the medium hospitals.

So the one take home message is that Large Hospitals admit the patients for a longer period of time than the medium hospitals.

REFERENCES :

- Applied Analytics Module 7 notes
- (R Documentation and manuals | R Documentation, 2020)
- (t.test function | R Documentation, 2020)
- (Admitted patients - Australian Institute of Health and Welfare, 2020)