Analysis on Admitted patients

Shonil Dabreo

Analysis on Admitted patients

- Introduction
- Problem Statement
- Data
- Descriptive Statistics and Visualization
- Hypothesis Testing
- Discussion
- References

Introduction

- Admitted patients are patients who undergo a public or private hospital’s formal admission process to receive treatment and/or care. The details of the admitted patients are recorded which include their No of overnight stays, Type of the hospital, Name of the hospital, etc.
- The problem is to understand the statistical difference in the average length of stay (days) between large and medium hospitals.
- Statistics is undertaken to compare the difference in the average length of stay (days) between both hospitals. The results of the statistics are interpreted by which patients might choose one hospital over the other.

Problem Statement

- The goal is to understand if there is any statistical significant difference in the average length of stay (days) between large and medium hospitals which might make patients choose one over the other.
- The problem is established by assuming a null hypothesis that the mean average length of stay (days) for large hospitals is equal to the mean average length of stay (days) for medium hospitals.
- T-tests are statistical hypothesis tests which is conducted to test if the hypothesis is plausible.

Data

Data preprocessing is extremely important because it allows improving the quality of the raw experimental data. The primary aim of preprocessing is to eliminate those small data contributions associated with the experimental error.

Data Cleaning is one of the Data preprocessing step for collecting and preparing the data.

Steps of Data Cleaning:-

- Eliminating records or attributes with missing values
- Identify and remove outliers
- Feature Engineering

Data

Eliminating records or attributes with missing values

- The data is noisy which contains the textual information about the data source. This textual information is removed (i.e. skipped) while retrieving the data.
- There are certain attributes which contains missing values in all the rows. These attributes are ignored by selecting attributes which includes some values in it.

Data

Identify and remove outliers

The average length of stay (days) attribute includes NP values (i.e. Reported values which didn’t meet the criteria).

The NP values include criteria where patients reported deaths or transfers to another hospital. Therefore, Replacing NP values with any constant value would be illogical. The NP values could be assumed as outliers.

Data

Identify and remove outliers

- The total count of NP values is calculated by filtering the NP values from average length of stay (days) attribute.
- The average length of stay (days) attribute is converted to numeric. This changes all the NP values to NA (i.e. Not Applicable) values.
- These NA values are then ignored by filtering average length of stay (ALOS) attribute values which are other than NA values.

Data

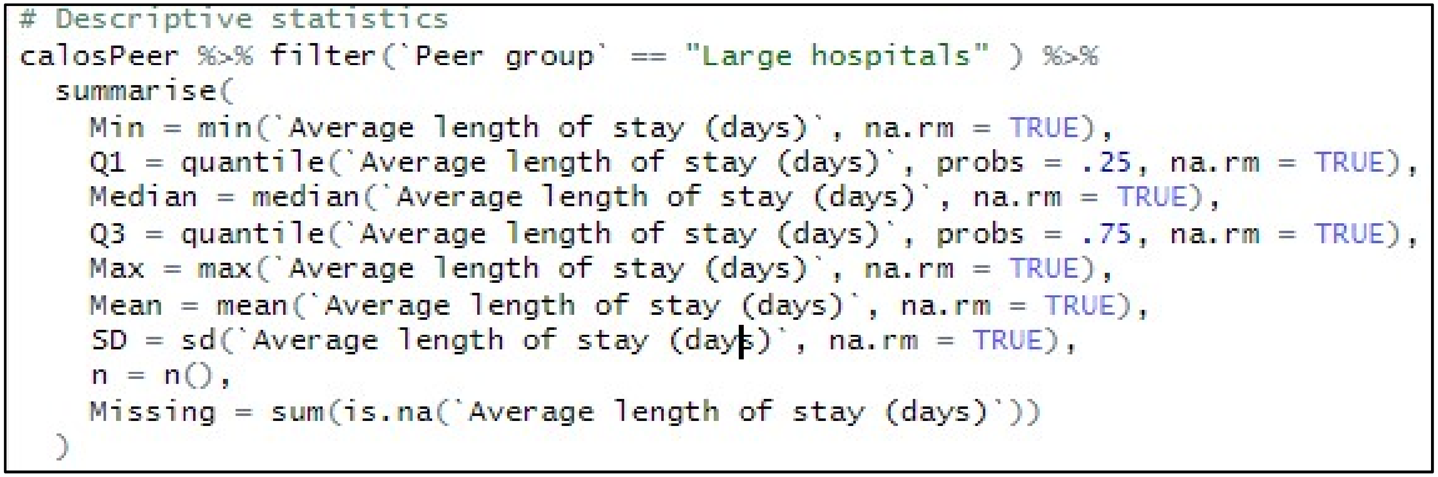
Feature Engineering

- The attributes of interest are selected to calculate the statistical information.
- The ‘average length of stay (days)’ and ‘Peer group’ are the attributes which are selected to perform the statistics.
- ‘average length of stay (days)’ is calculated as the number of bed days for overnight stays divided by the number of overnight stays.
- ‘Peer group’ is the type of hospital where patients were admitted such as Major hospital, Large hospital, Medium hospital, Children’s hospital, Small hospital and Unpeered.

Descriptive Statistics and Visualization

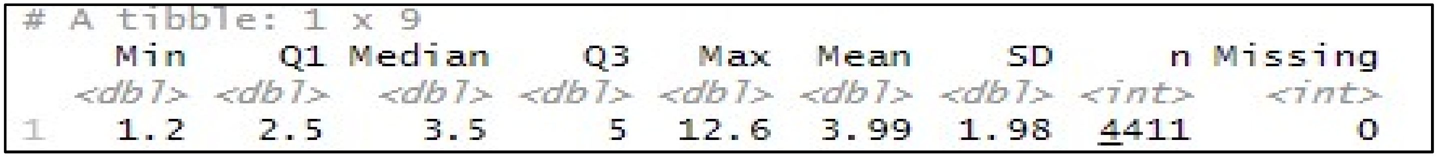
Descriptive Statistics

- Descriptive statistics of average length of stay (days) in Large hospitals is calculated by filtering the average length of stay (days) based on Peer group ‘Large hospital’.



```
# Descriptive statistics
calosPeer %>% filter(`Peer group` == "Large hospitals" ) %>%
  summarise(
    Min = min(`Average length of stay (days)`, na.rm = TRUE),
    Q1 = quantile(`Average length of stay (days)`, probs = .25, na.rm = TRUE),
    Median = median(`Average length of stay (days)`, na.rm = TRUE),
    Q3 = quantile(`Average length of stay (days)`, probs = .75, na.rm = TRUE),
    Max = max(`Average length of stay (days)`, na.rm = TRUE),
    Mean = mean(`Average length of stay (days)`, na.rm = TRUE),
    SD = sd(`Average length of stay (days)`, na.rm = TRUE),
    n = n(),
    Missing = sum(is.na(`Average length of stay (days)`))
  )
```

Code



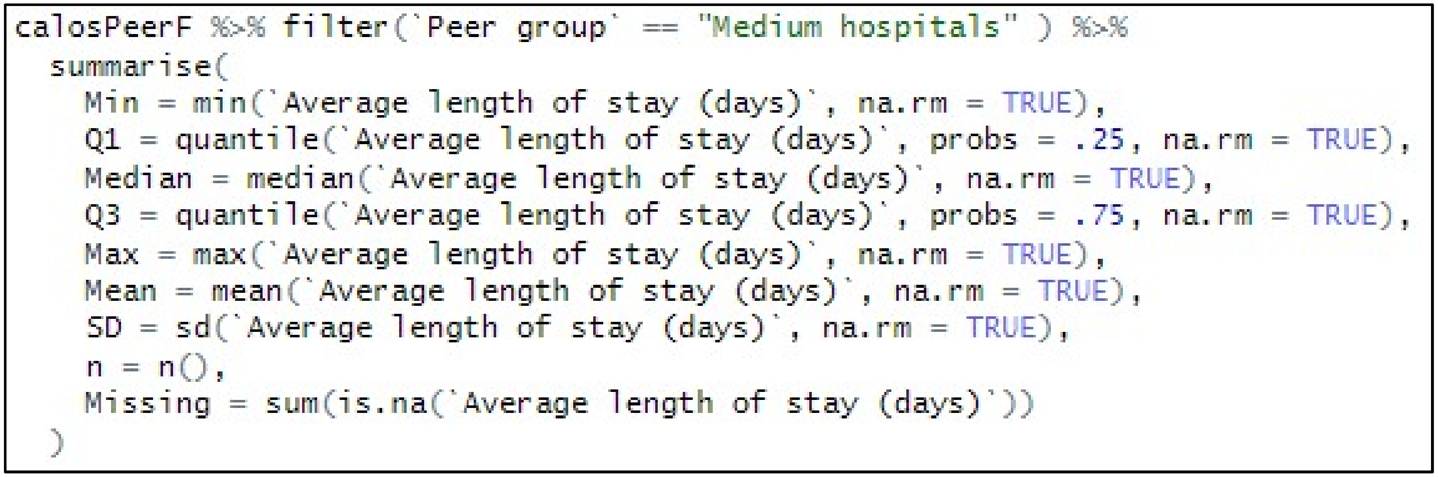
```
# A tibble: 1 x 9
    Min    Q1 Median    Q3   Max  Mean    SD     n Missing
  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <int>
1   1.2   2.5    3.5     5  12.6  3.99  1.98  4411       0
```

Output

Descriptive Statistics and Visualization

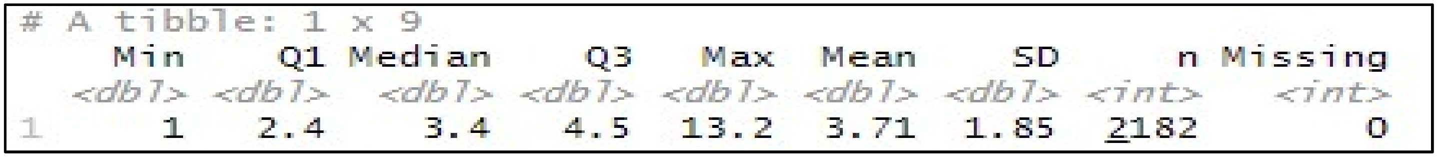
Descriptive Statistics

- Descriptive statistics of average length of stay (days) in Medium hospitals is calculated by filtering the average length of stay (days) based on Peer group ‘Medium hospital’.



```
calosPeerF %>% filter(`Peer group` == "Medium hospitals" ) %>%
  summarise(
    Min = min(`Average length of stay (days)`, na.rm = TRUE),
    Q1 = quantile(`Average length of stay (days)`, probs = .25, na.rm = TRUE),
    Median = median(`Average length of stay (days)`, na.rm = TRUE),
    Q3 = quantile(`Average length of stay (days)`, probs = .75, na.rm = TRUE),
    Max = max(`Average length of stay (days)`, na.rm = TRUE),
    Mean = mean(`Average length of stay (days)`, na.rm = TRUE),
    SD = sd(`Average length of stay (days)`, na.rm = TRUE),
    n = n(),
    Missing = sum(is.na(`Average length of stay (days)`))
  )
```

Code

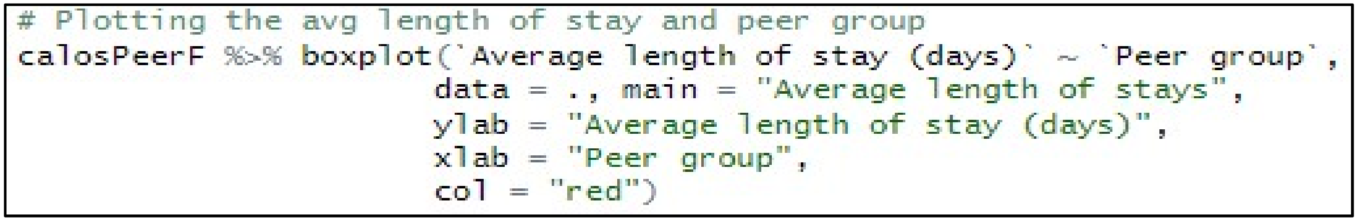


```
# A tibble: 1 x 9
    Min    Q1 Median    Q3   Max  Mean    SD     n Missing
  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <int>
1     1   2.4    3.4   4.5  13.2  3.71  1.85  2182       0
```

Output

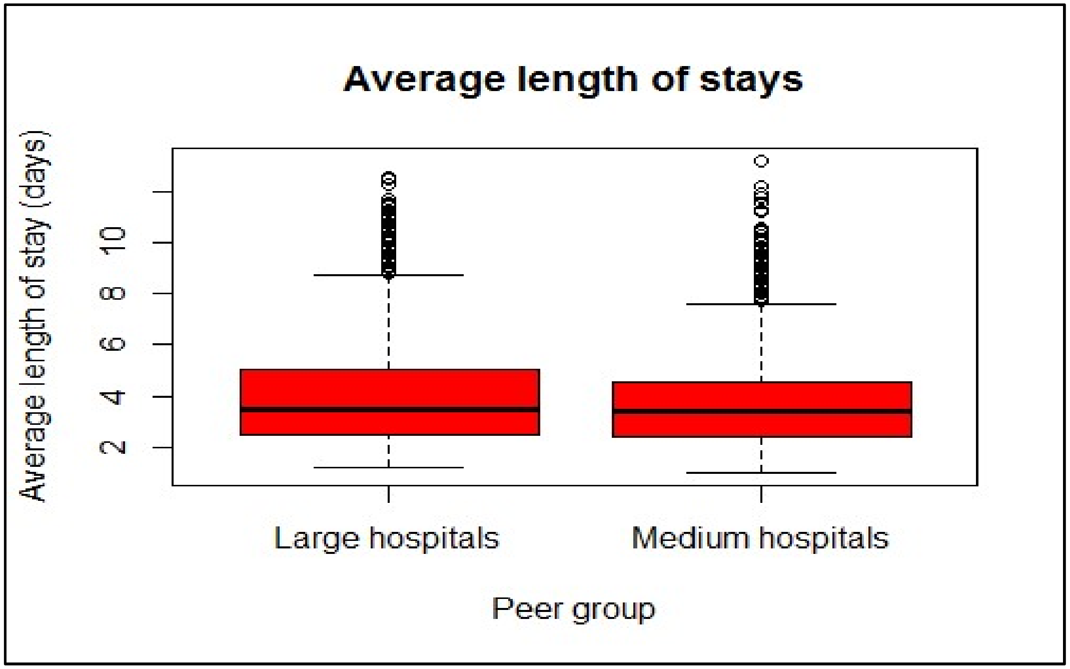
Descriptive Statistics and Visualization

Visualization



```
# Plotting the avg length of stay and peer group
calosPeerF %>% boxplot(`Average length of stay (days)` ~ `Peer group`,
                       data = ., main = "Average length of stays",
                       ylab = "Average length of stay (days)",
                       xlab = "Peer group",
                       col = "red")
```

Code



Output

Descriptive Statistics and Visualization

Visualization

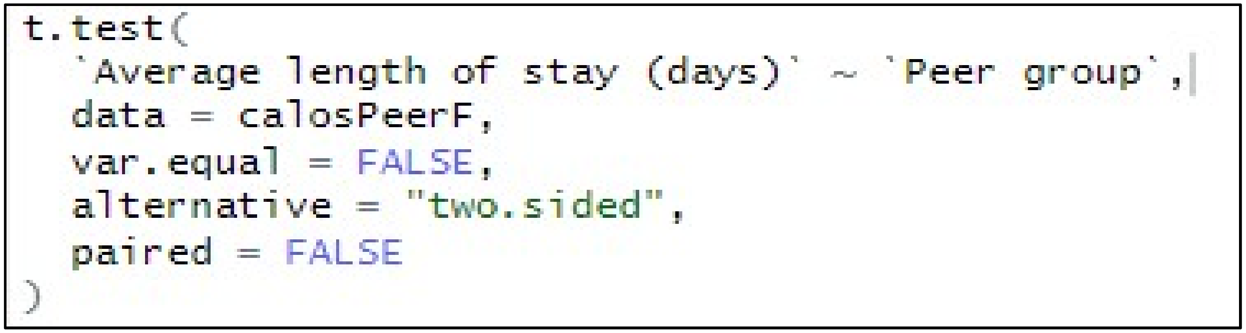
- Box plot of the average length of stay (days) in Large hospitals and Medium hospitals is plotted for visual analysis.
- Looking at the output, we can say that the mean of average length of stay (days) in Large hospitals is greater than the Medium hospitals.
- Many patients have similar average length of stay (days) in Large hospitals in certain parts of the scale which is Quartile group 3. But in other parts of the scale the average length of stay (days) of patients are variable.
- The outliers are visible from the upper whisker which are separately shown as plotted points.

Hypothesis Testing

Null Hypothesis is an assumption on which statistical T-test is computed to find out if the plausible hypothesis is true or not.

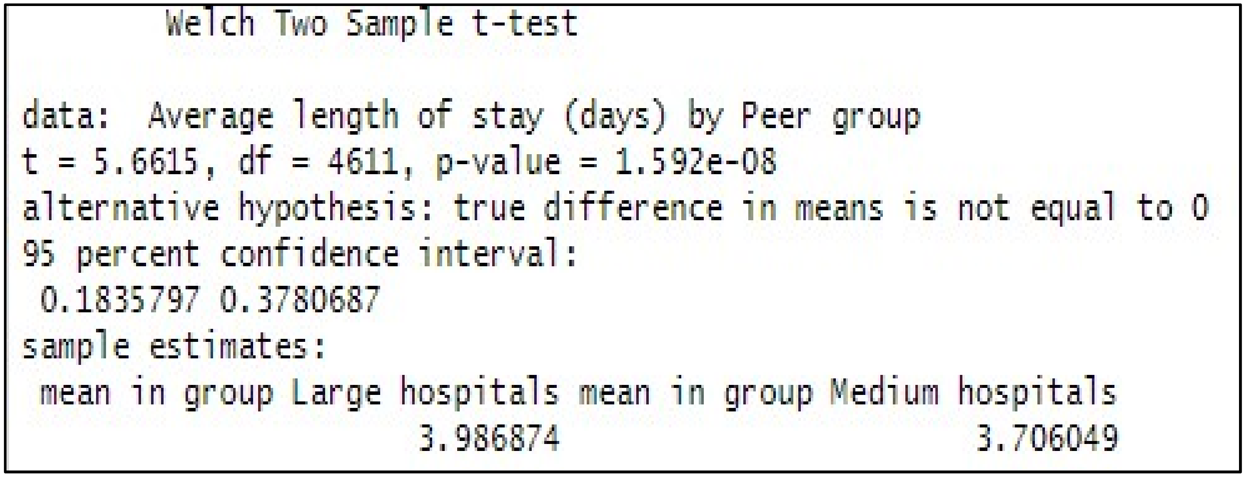
Null Hypothesis (H0 = 0):
Mean of average length of stay (days) in large hospitals is equal to the mean of medium hospitals.

Hypothesis Testing



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

Code



```
	Welch Two Sample t-test

data:  Average length of stay (days) by Peer group
t = 5.6615, df = 4611, p-value = 1.592e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1835797 0.3780687
sample estimates:
 mean in group Large hospitals mean in group Medium hospitals
                      3.986874                       3.706049
```

Output

Hypothesis Testing

T-test is conducted and result of the t-test for hypothesis is displayed.

- **t** is the t-test statistic value which is 5.6615
- **df** is the degrees of freedom which is 4611
- **p-value** is the significance level of the **t-test** (p-value = 0.0001592).
- **conf.int** is the **confidence interval** of the mean at 95% ([0.1835797, 0.3780687]);
- **sample estimates** is the mean value of two independent samples (mean = 3.986874, 3.706049).

Interpretation:

The p-value of the test is 0. 0001, which is less than the significance level alpha = 0.05. Therefore, Null hypothesis is rejected.

Thus, it can be concluded that mean of average length of stay (days) in large hospitals is significantly different than the mean of medium hospitals.

Discussion

- The mean of average length of stay (days) in large hospitals is greater than the mean of average length of stay (days) medium hospitals.
- The degree of freedom is higher which means more power to reject a false null hypothesis and find a significant result
- The limitation is that the outliers were removed for t-test hypothesis testing.
- As a result, we can say that patients might choose Large hospital over the Medium hospital.

References

- Science direct. Data Pre-processing [online]. Available at <https://www.sciencedirect.com/topics/engineering/data-preprocessing / > [Accessed 10 May 2020]

Thank You