HYPOTHESIS TESTING

AUSTRALIAN INSTITUTE OF HEALTH AND WELFARE

By Christian Orochi Themin

8 May 2020

Introduction

This report will compare the average length of stay between large hospitals and medium hospitals to determine whether there is any statistical difference.

The data used for investigation is collected from Australian Institute of Health and Welfare.

Problem Statement

Investigators want to know if large hospitals and medium hospitals had an effect in determining the patients' length of stay.

Using the methods of calculating the Descriptive Statistics, Visualisation using Histogram & Boxplot, we can compare between the two groups of hospitals.

Hypothesis Testing will be implemented for further testing and will discuss with Conclusion.

Data

The Average length of stay dataset can be downloaded from the website Australian Institute of Health and Welfare. The dataset contains over 30,000 reports from different hospitals in Australia and there are 13 variables in the dataset:

1. Reporting unit
2. Reporting unit type
3. State
4. Local Hospital Network
5. Peer group
6. Time period
7. Category
8. Total number of stays
9. Number of overnight stays
10. Percentage
11. Average length of stay (days)
12. Peer group average (days)
13. Total overnight patient bed days

Analysis

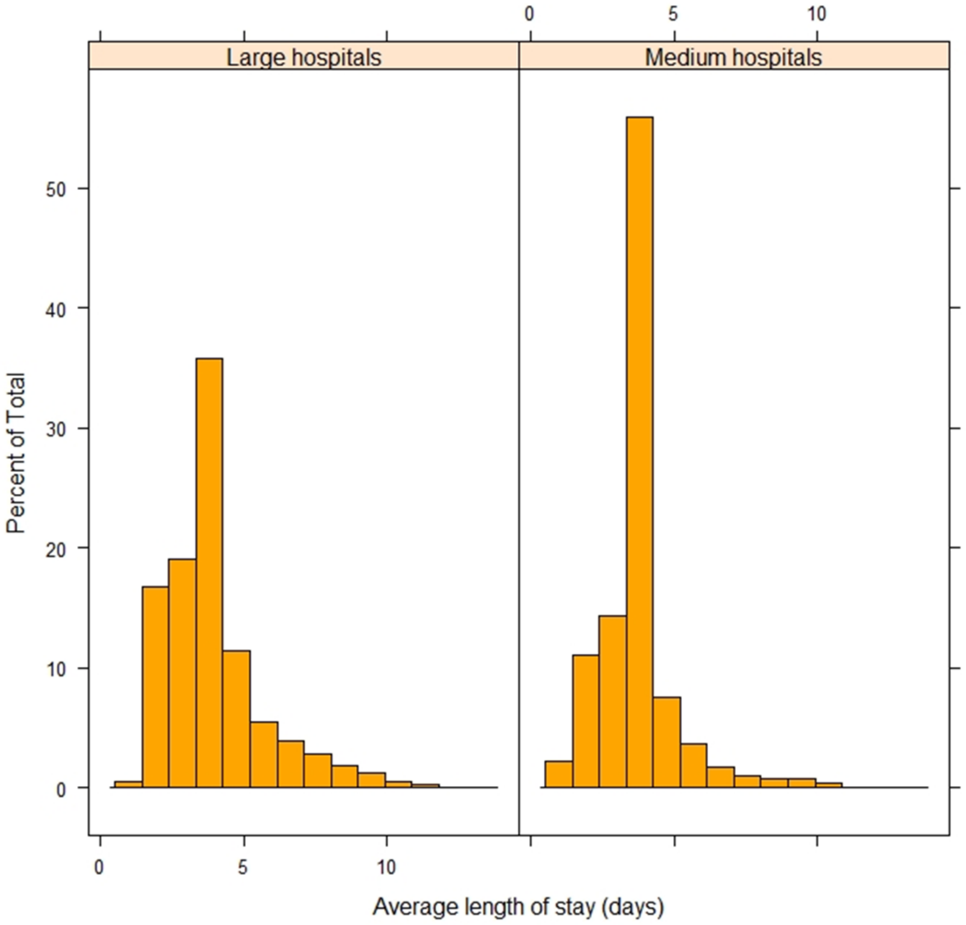
During the analysis process on the variable name Average length of stay (ALOS), there have been found around 19,770 missing values, it is more than 50% of the data.

In this report, it will be replaced with Median value of the variable which is 3.4.

The Descriptive Statistics are then summarised as below:

	Min	Q1	Median	Mean	Q3	Max	SD	n
Large hospitals	1.2	2.7	3.4	3.85	4.5	12.6	1.76	5692
Medium hospitals	1	3.1	3.4	3.57	3.6	13.2	1.4	3877

Analysis

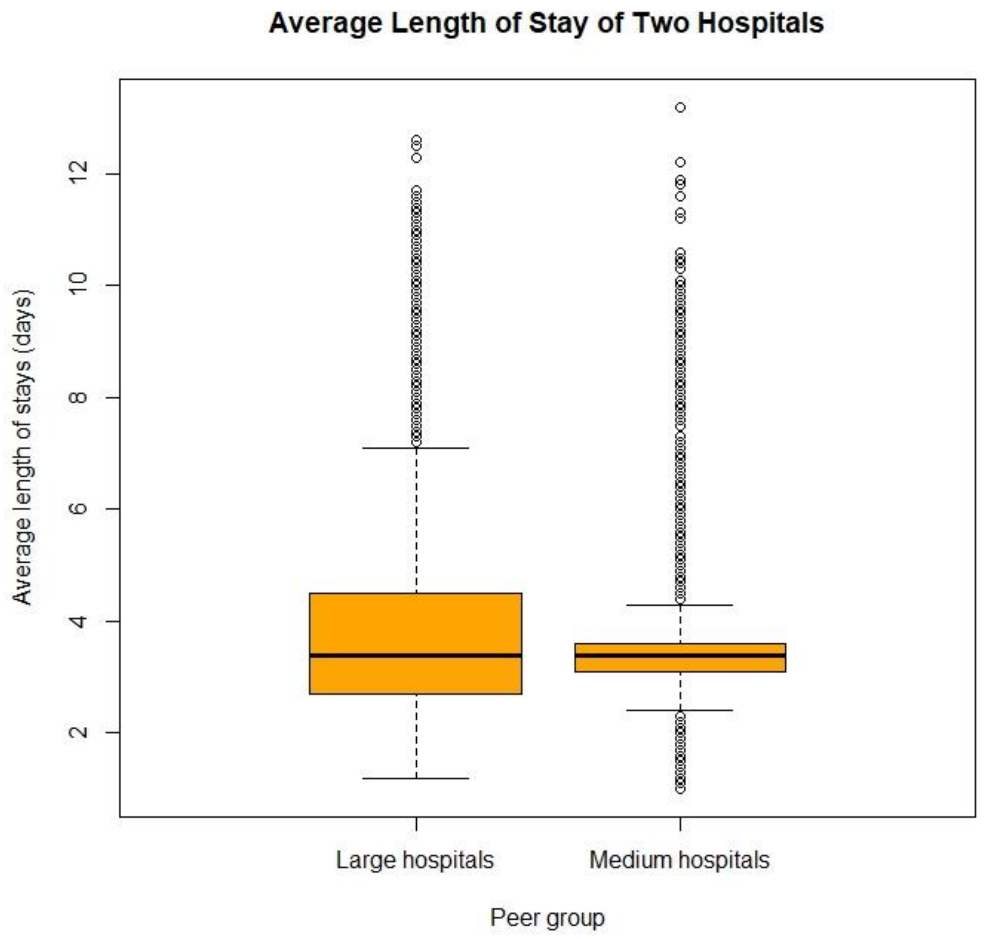


The histograms shows that Medium hospitals tend to have higher length of stays than the Large hospitals.

It is also right-skewed or unbalance distribution, possibly due to outliers.

We can identify this further using boxplot.

Analysis



The boxplot shows there are many outliers appears in both hospitals.

The outliers occurrence are likely due to the Missing value that was replaced with Median.

Analysis

The outliers produced poor measurements and it is unwanted for this report.

It is preferred to be removed by using the following method:

Large hospitals:

Lower outlier < Q1 – 1.5 * IQR = 2.7 – (1.5 * 1.8) = 0

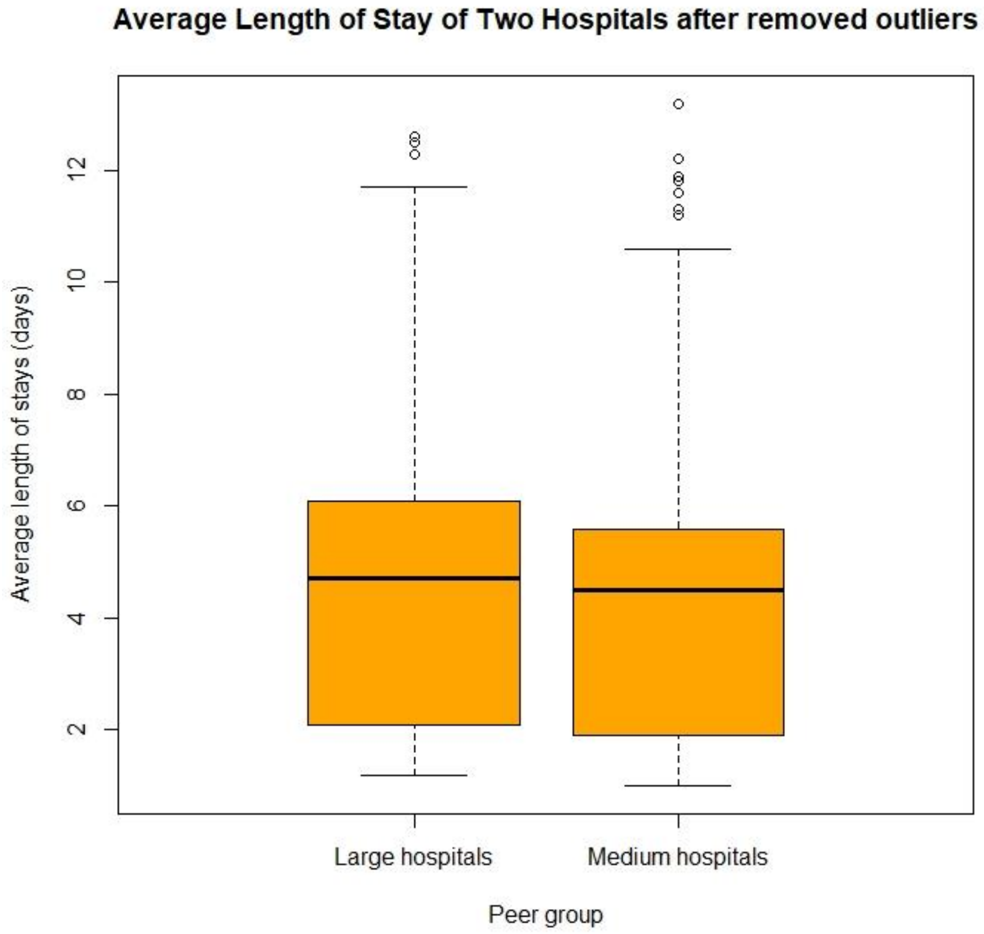
Upper outlier > Q3 + 1.5 *IQR = 4.5 + (1.5 *1.8) = 7.2

Medium hospitals:

Lower outlier < Q1 – 1.5 * IQR = 3.1 – (1.5 * 0.5) = 2.35

Upper outlier > Q3 + 1.5 *IQR = 3.6 + (1.5 *0.5) = 4.35

Analysis

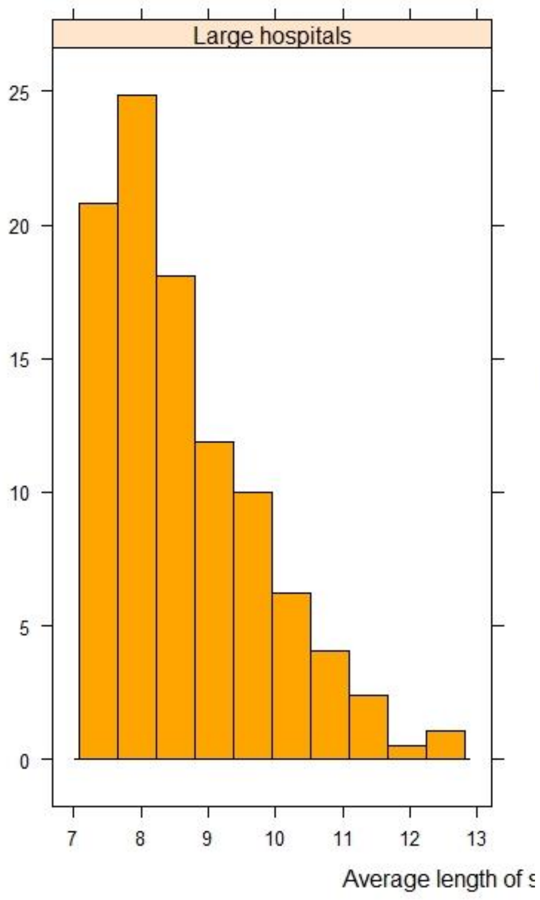


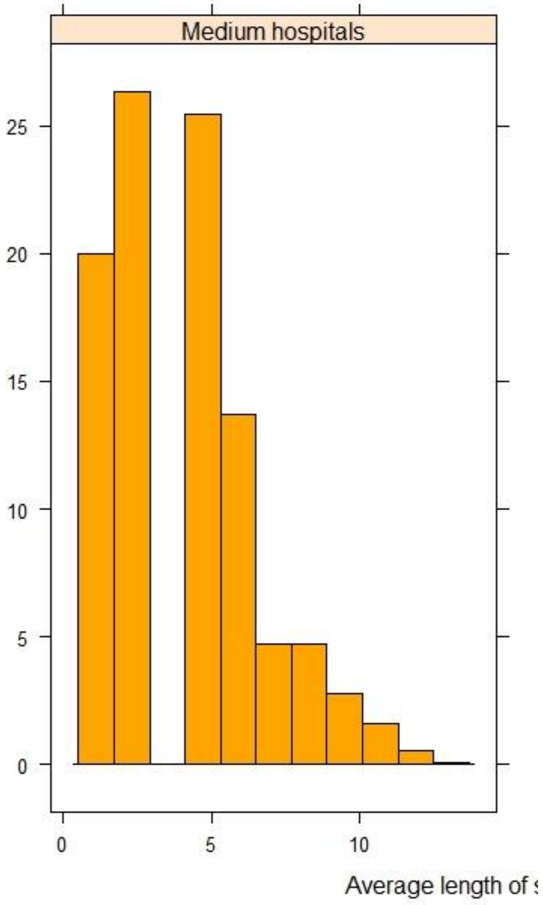
After removed some of the outliers, there are still a few outliers left.

This is due to a very large dataset with highly skewed distributions.

For this report, we will leave the result the way as it is.

Analysis





The histogram of removed outliers are still right-skewed distributed or positively distributed.

It is reasonable because in general, the length of hospital stay is better when its lesser, showing that the patient is recovered sooner.

Hypothesis Testing

It has been said that the Average length of stay is 3.4 days.

The investigator want to check with the assumption as below:

Ho : μ = 3.4

Ha : μ != 3.4

We will check for both large and medium hospitals.

Hypothesis Testing – Large Hospitals

$t = 19.507, p < 0.05$, 95% CI [3.81, 3.9]

A one-sample t-test was used to determine any significant different of the Average length of stay from the previous assumed mean of 3.4 days.

The sample's mean is resulted as 3.85 days with 95% CI.

The result is statistically significant higher than the assumption mean.

Hypothesis Testing – Medium Hospitals

$t = 7.67$, $p < 0.05$, 95% CI [3.53, 3.62]

A one-sample t-test was used to determine any significant different of the Average length of stay from the previous assumed mean of 3.4 days.

The sample's mean is resulted as 3.57 days with 95% CI.

The result is statistically significant higher than the assumption mean.

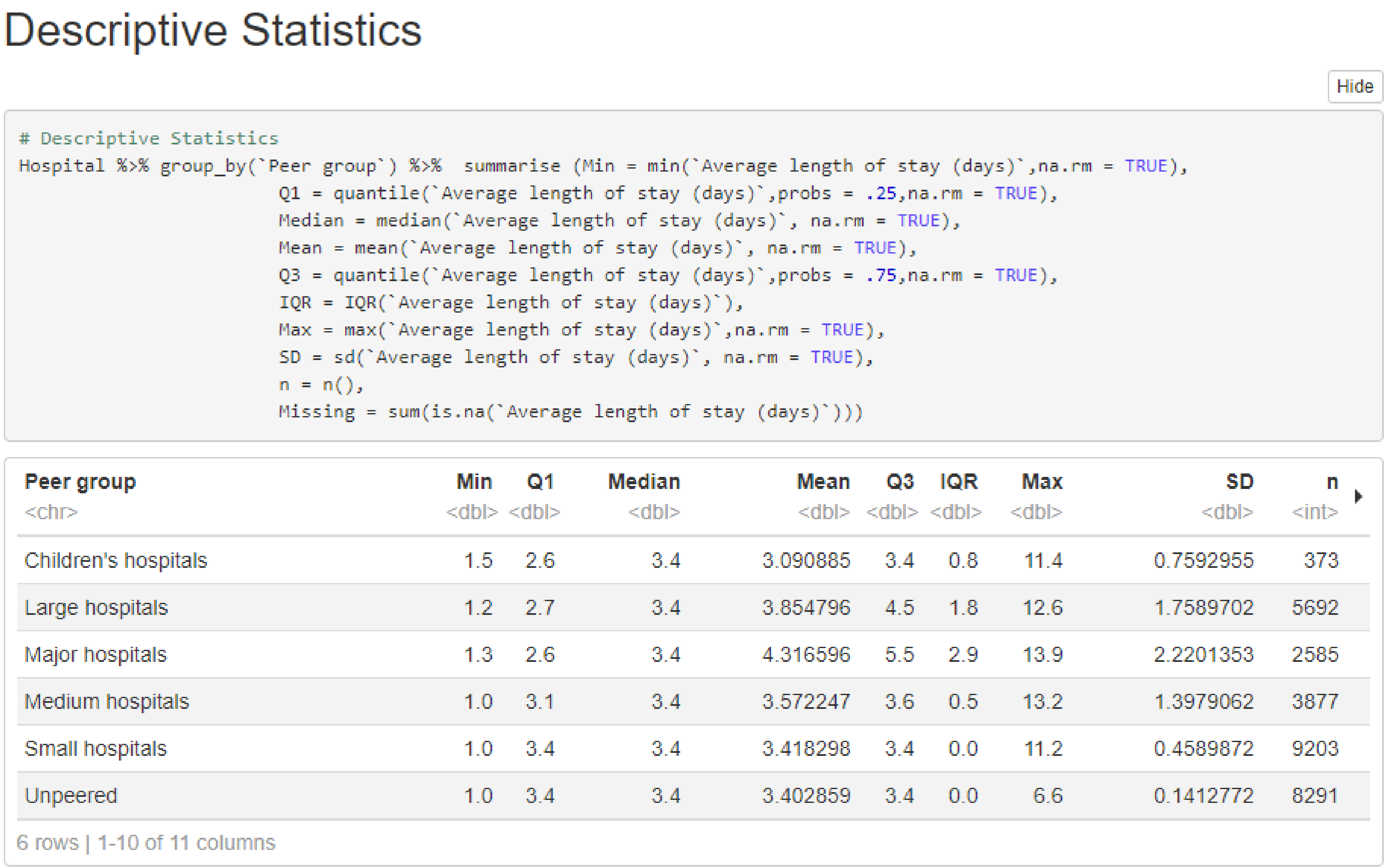
Conclusion

Several statistic tests and investigations have been conducted between the two group of hospitals. To conclude, there is a statistical significant difference in the average length of stay between the hospitals.

Despite the large dataset, there has been found more than 50% of missing values and produced many outliers when replacing with Median value.

Large hospitals have an average of at least 7 nights of stay and the longest stay is up to 13 days. While Medium hospitals starts from 0 night the least and the longest of 13 days but with a very low frequency compare to the Large hospitals.

Coding Implementation



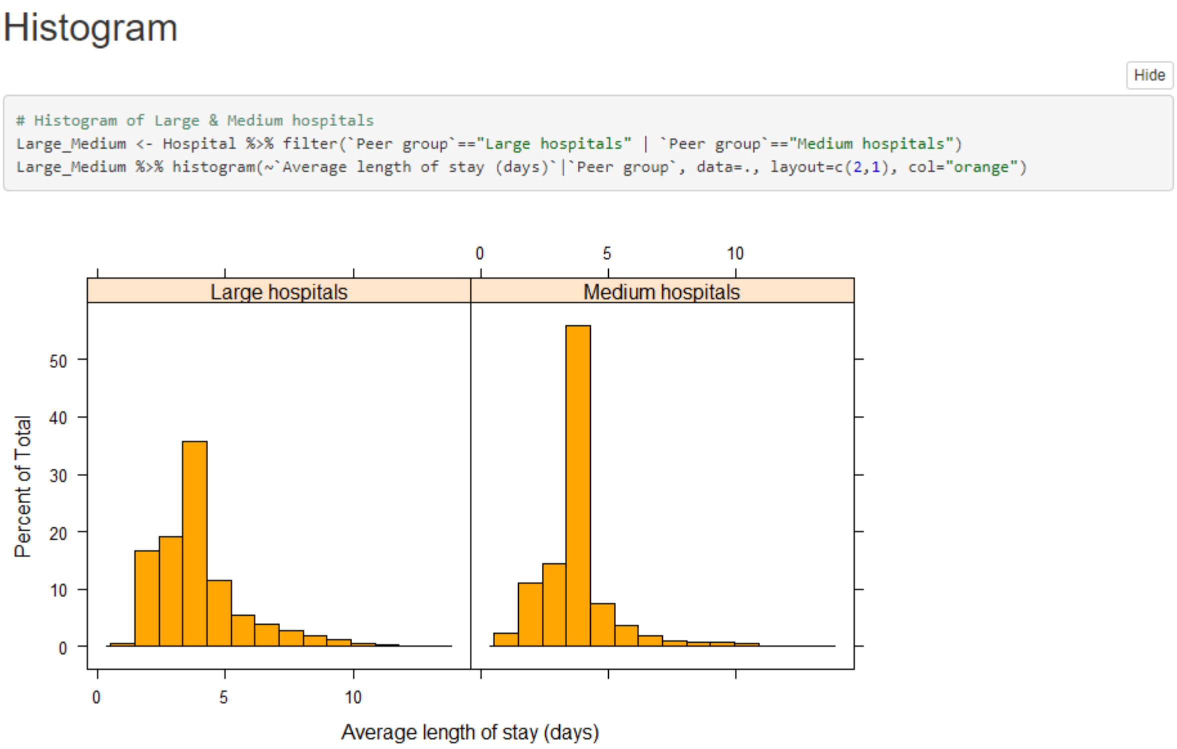
Descriptive Statistics

```
# Descriptive Statistics
Hospital %>% group_by(`Peer group`) %>%  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),
                        Mean = mean(`Average length of stay (days)`, na.rm = TRUE),
                        Q3 = quantile(`Average length of stay (days)`,probs = .75,na.rm = TRUE),
                        IQR = IQR(`Average length of stay (days)`),
                        Max = max(`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)`)))
```

Peer group	Min	Q1	Median	Mean	Q3	IQR	Max	SD	n
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
Children's hospitals	1.5	2.6	3.4	3.090885	3.4	0.8	11.4	0.7592955	373
Large hospitals	1.2	2.7	3.4	3.854796	4.5	1.8	12.6	1.7589702	5692
Major hospitals	1.3	2.6	3.4	4.316596	5.5	2.9	13.9	2.2201353	2585
Medium hospitals	1.0	3.1	3.4	3.572247	3.6	0.5	13.2	1.3979062	3877
Small hospitals	1.0	3.4	3.4	3.418298	3.4	0.0	11.2	0.4589872	9203
Unpeered	1.0	3.4	3.4	3.402859	3.4	0.0	6.6	0.1412772	8291

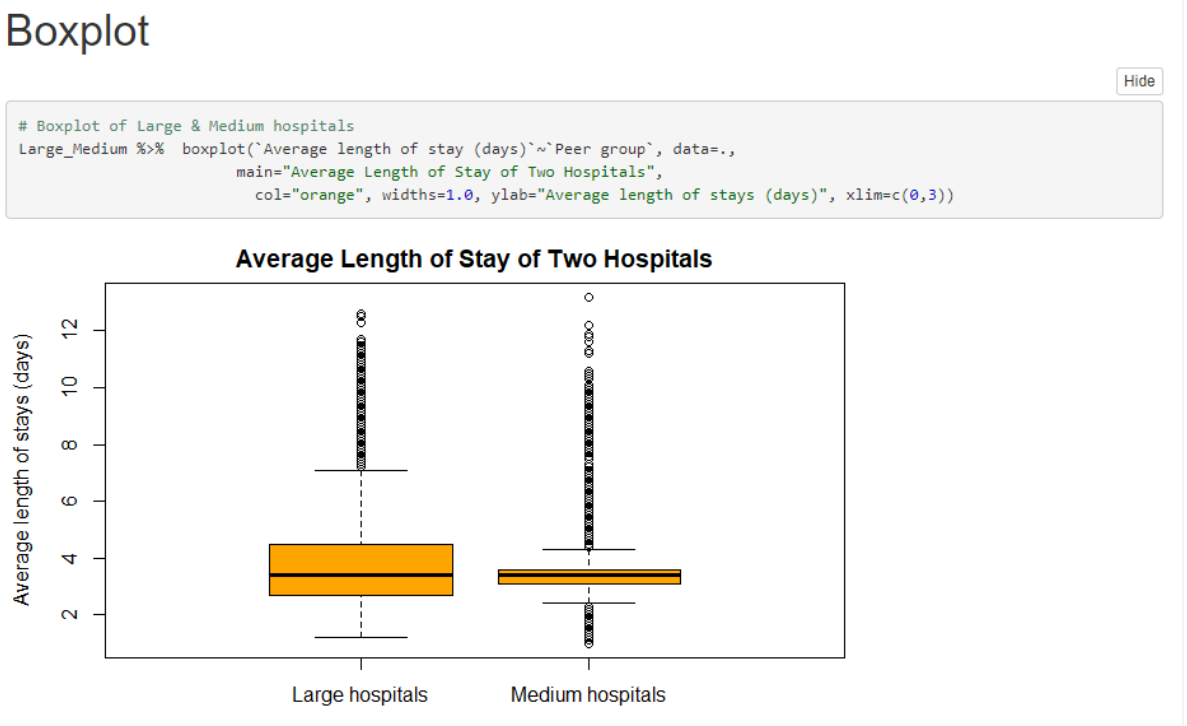
6 rows | 1-10 of 11 columns

Coding Implementation



Histogram

```
# Histogram of Large & Medium hospitals
Large_Medium <- Hospital %>% filter(`Peer group`=="Large hospitals" | `Peer group`=="Medium hospitals")
Large_Medium %>% histogram(~`Average length of stay (days)`|`Peer group`, data=., layout=c(2,1), col="orange")
```



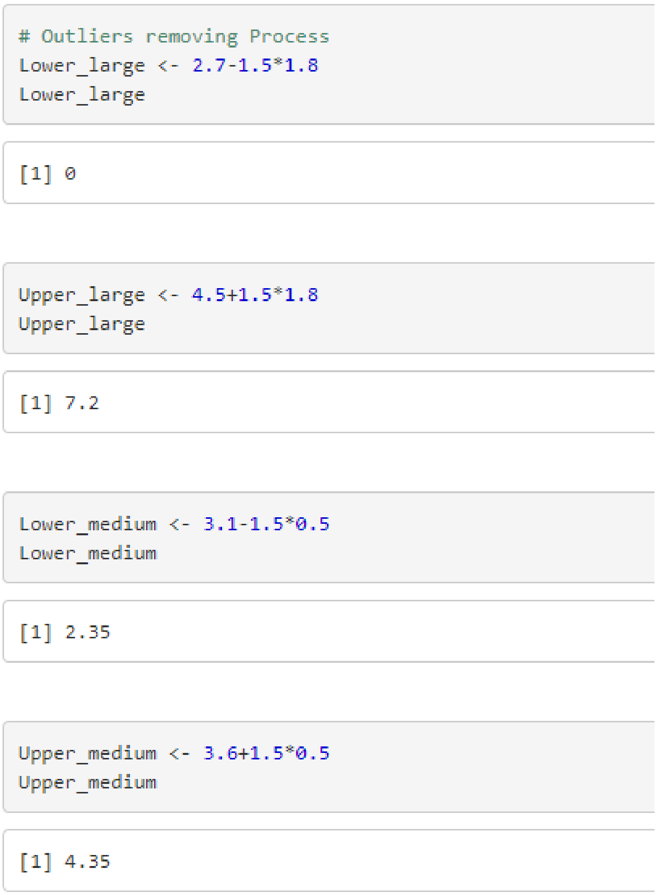
Boxplot

```
# Boxplot of Large & Medium hospitals
Large_Medium %>%  boxplot(`Average length of stay (days)`~`Peer group`, data=.,
                main="Average Length of Stay of Two Hospitals",
                  col="orange", widths=1.0, ylab="Average length of stays (days)", xlim=c(0,3))
```

Average Length of Stay of Two Hospitals
Average length of stays (days)
2
4
6
8
10
12
Large hospitals
Medium hospitals

Coding Implementation

Outliers Removing Process



```
# Outliers removing Process
Lower_large <- 2.7-1.5*1.8
Lower_large
```

```
[1] 0
```

```
Upper_large <- 4.5+1.5*1.8
Upper_large
```

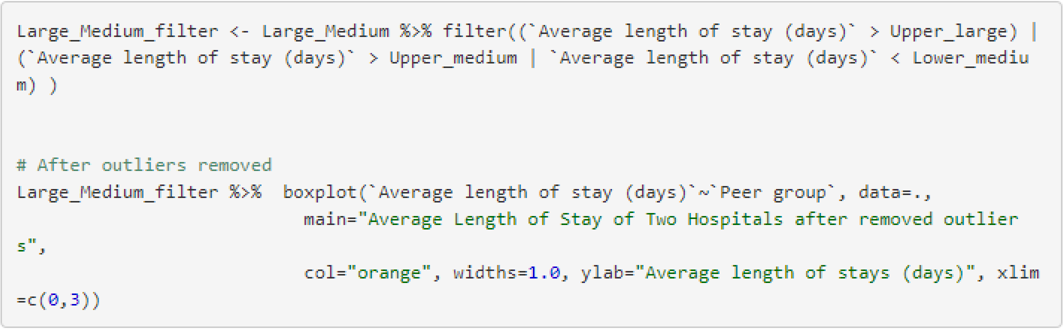
```
[1] 7.2
```

```
Lower_medium <- 3.1-1.5*0.5
Lower_medium
```

```
[1] 2.35
```

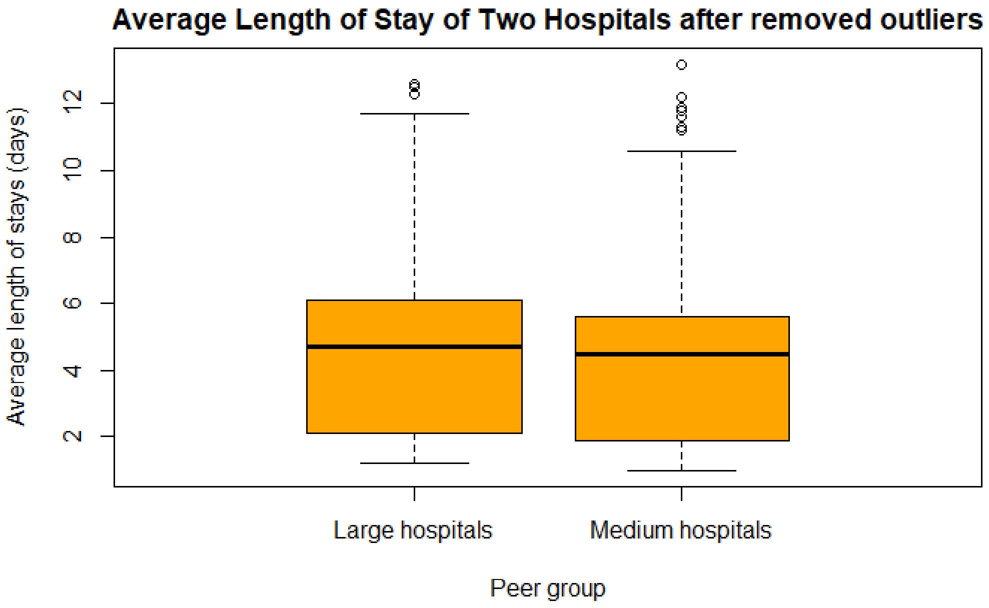
```
Upper_medium <- 3.6+1.5*0.5
Upper_medium
```

```
[1] 4.35
```

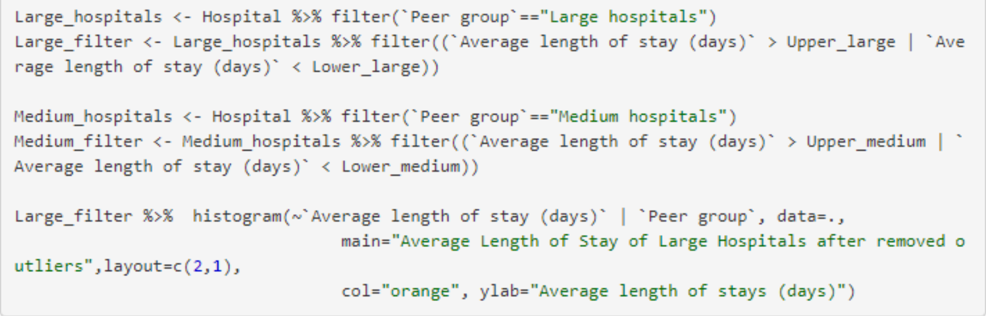


```
Large_Medium_filter <- Large_Medium %>% filter((`Average length of stay (days)` > Upper_large) |
(`Average length of stay (days)` > Upper_medium | `Average length of stay (days)` < Lower_mediu
m) )


# After outliers removed
Large_Medium_filter %>%  boxplot(`Average length of stay (days)`~`Peer group`, data=.,
                                  main="Average Length of Stay of Two Hospitals after removed outlier
s",
                                  col="orange", widths=1.0, ylab="Average length of stays (days)", xlim
=c(0,3))
```



Coding Implementation

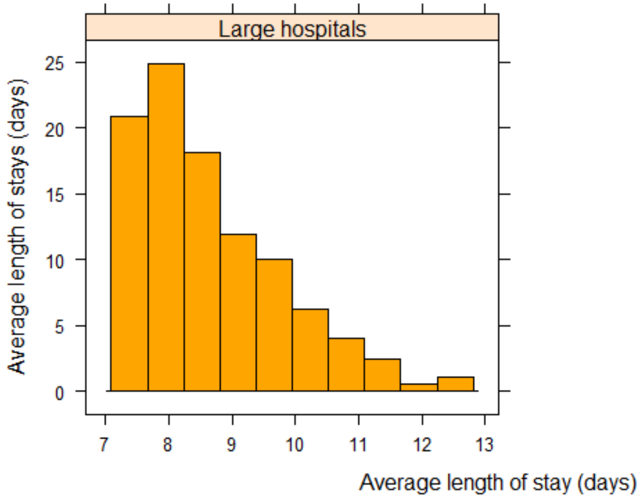


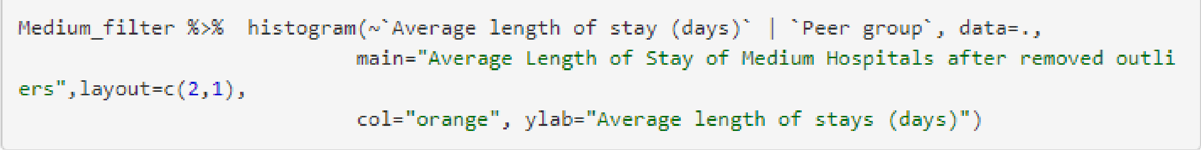
```
Large_hospitals <- Hospital %>% filter(`Peer group`=="Large hospitals")
Large_filter <- Large_hospitals %>% filter((`Average length of stay (days)` > Upper_large | `Ave
rage length of stay (days)` < Lower_large))

Medium_hospitals <- Hospital %>% filter(`Peer group`=="Medium hospitals")
Medium_filter <- Medium_hospitals %>% filter((`Average length of stay (days)` > Upper_medium | `
Average length of stay (days)` < Lower_medium))

Large_filter %>%  histogram(~`Average length of stay (days)` | `Peer group`, data=.,
                                 main="Average Length of Stay of Large Hospitals after removed o
utliers",layout=c(2,1),
                                 col="orange", ylab="Average length of stays (days)")
```

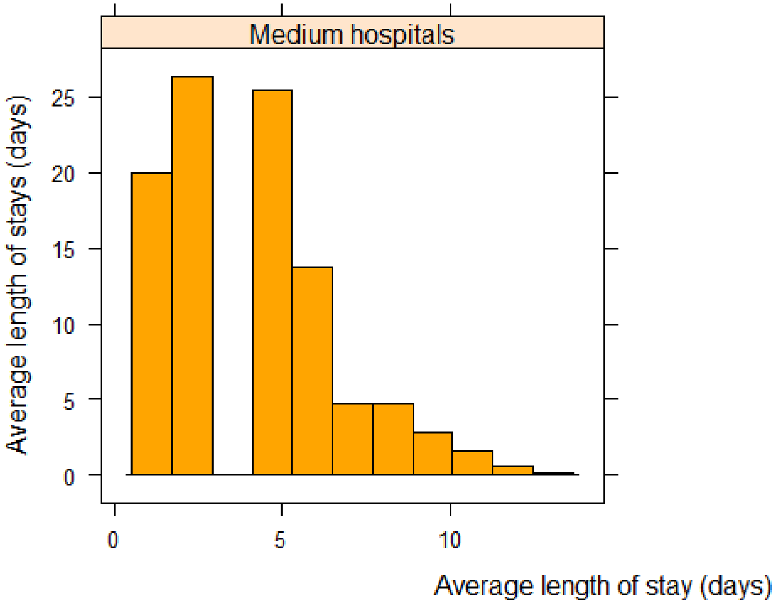
Average Length of Stay of Large Hospitals after removed outliers



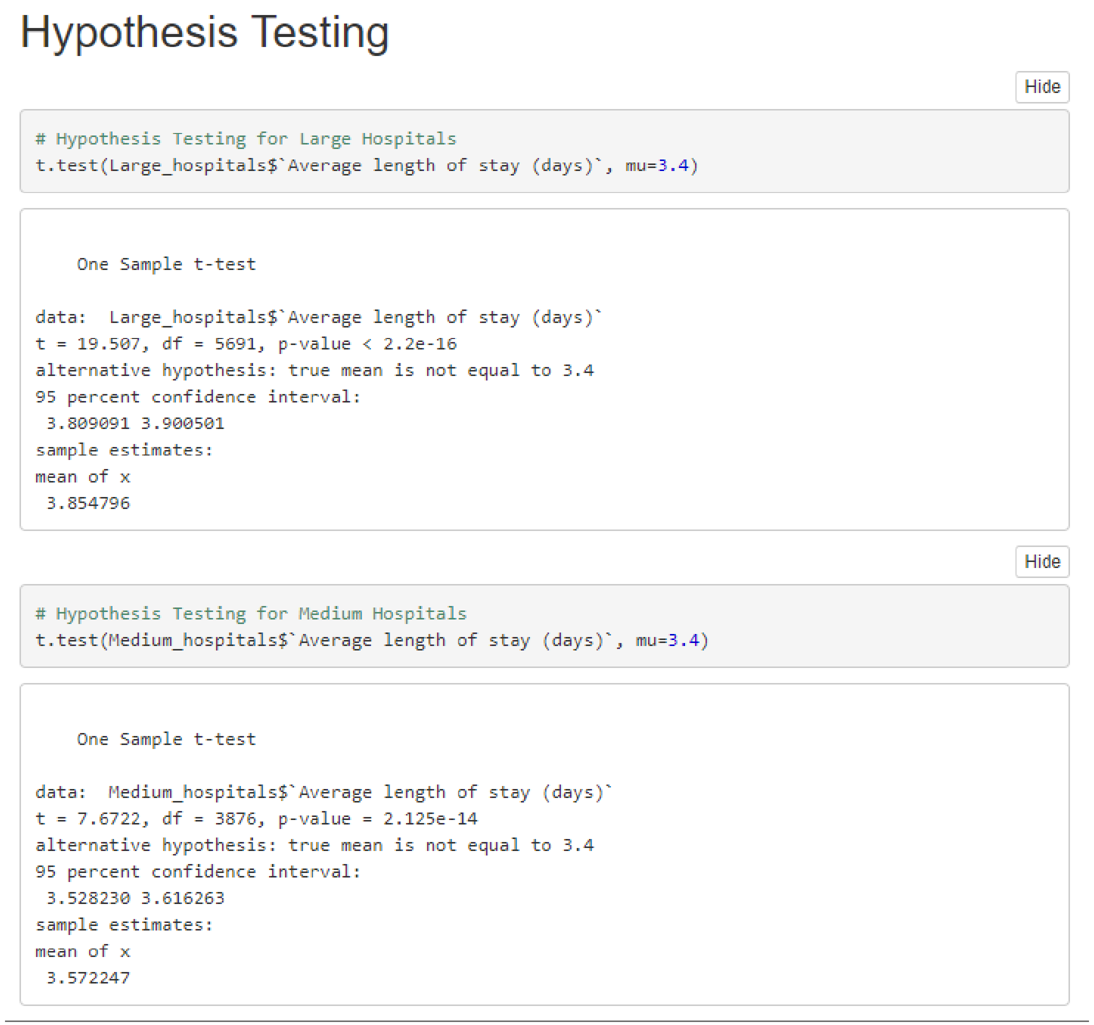


```
Medium_filter %>%  histogram(~`Average length of stay (days)` | `Peer group`, data=.,
                                 main="Average Length of Stay of Medium Hospitals after removed outli
ers",layout=c(2,1),
                                 col="orange", ylab="Average length of stays (days)")
```

Average Length of Stay of Medium Hospitals after removed outliers



Coding Implementation



Hypothesis Testing

```
# Hypothesis Testing for Large Hospitals
t.test(Large_hospitals$`Average length of stay (days)`, mu=3.4)
```

```

    One Sample t-test

data:  Large_hospitals$`Average length of stay (days)`
t = 19.507, df = 5691, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 3.4
95 percent confidence interval:
 3.809091 3.900501
sample estimates:
mean of x 
 3.854796 
```

```
# Hypothesis Testing for Medium Hospitals
t.test(Medium_hospitals$`Average length of stay (days)`, mu=3.4)
```

```

    One Sample t-test

data:  Medium_hospitals$`Average length of stay (days)`
t = 7.6722, df = 3876, p-value = 2.125e-14
alternative hypothesis: true mean is not equal to 3.4
95 percent confidence interval:
 3.528230 3.616263
sample estimates:
mean of x 
 3.572247 
```

REFERENCES

- RMIT University for Data Science Students

All works are done individually by Christian Themin

8 May 2020