Quantitative Investigation on the car parking situation at Chadstone shopping Centre

Yuan Wei

University of the Tasimania
ywei10@utas.edu.com

Abstract. The purpose of this report is to investigate the car parking situation at Chadsone shopping centre using quantitative tools and techniques. The t-test, ANOVA test, normality test is performed regarding differences between school holiday and non-school holiday; and the differences between four different terminal carparks.

Keywords: ANOVA test, t-test, normality test.

1 Introduction

The purpose of this report is to investigate the car parking situation at Chadsone shopping centre using quantitative tools and techniques. The dataset includes a whole-year-record of the number of cars entering each of four different carparks every day, there are Northeast, southeast, south and southwest carparks respectively. And two questions are expected to be analyzed based on the dataset, one is "Are their differences between car park usage during school holidays and non-school holidays?" and the other one is "Are there differences in the number of cars using different terminal carparks". A range of the statistical analysis techniques is applied to the existing dataset to examine the two questions. The main tool used for analyzing the data is Python and Excel, data preparation has conducted in the excel while python is used to perform the test and generate the results. In the next section, description and illustration of the selected data set is discussed with appropriate charts. And in the main body of the report, the analytics results are discussed regarding holiday vs non-holiday questions and different terminal carparks questions.

2 Description and illustration of the selected dataset

The whole dataset has 5 columns which are Date, NE, SE, S, SW respectively. It collected the number of cars entering these four car parks with respect to the date, from 29/01/2018 to 28/01/2019. Refer to the table one, it can be discovered that the mean of the NE, SE, S, SW are different. The count of the four-data set is 365. And the range of the

Table 1. Table captions should be placed above the tables.

Indicator	**NE**	**SE**	**S**	**SW**
Mean	5969.01	2066.21	1343.75	6123.89
Range	8317	2849	1470	7742
Minimum	3482	1143	865	3835
Maximum	11799	3992	2335	11577
Sum	2178689	754168	490467	2235220
Count	365	365	365	365

The box plot and error bar plot is displayed below that show the median, interquartile range, maximum, minimum, and outliers that should be excluded.

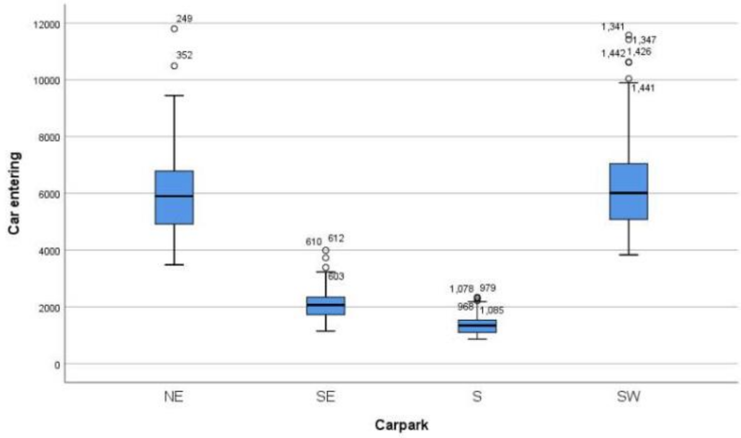

Fig.1. Box-plot of carparks

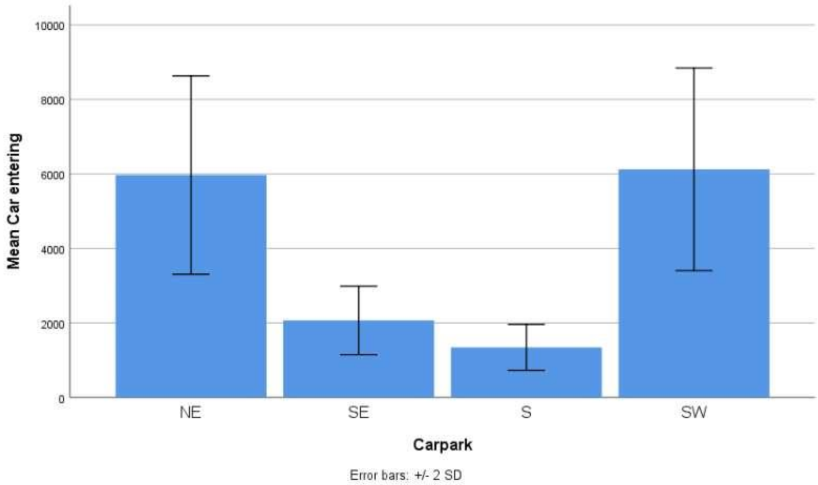

Fig.2. Error bar chart

3 Are there differences between car park usage during the school holidays and non-school holidays?

Data selection

According to the question, it focuses on the car entering the carparks at school holiday and non-school holiday. The School terms have been specified in the question where Term 1 from 29/01/18 to 29/03/18, Term 2 from16/04/18 to 29/06/18, Term 3 from16/07/18 to 21/09/18, and Term 4: from 08/10/18 to 21/12/18. Thus, we need to select the holiday and non-holiday intervals from the dataset and since it is only requiring estimate the difference between holidays and non-holidays, not studying the individual car parks difference. combining the holiday of four terms to form holiday interval; combining the school days of four terms to form non-holiday intervals, and also combining the count of the number of cars entering the four carparks, NE, SE, SW, and S together. Although the two-sample size is different, the non-holiday sample size is bigger than the non-holiday sample as shown in the table below:

Table 2. Selected school holiday and selected Non-school holiday data

Selected School Holiday data		
Holiday Interval	**Date**	**Sum of NE, SE, S, SW**
T1holiday	30/03/18	20266
	…	…
	15/04/18	17113
T2 holiday	30/06/18	18502
	…	…
	15/07/18	19528
T3 holiday	22/09/18	19791
	…	…
	07/10/18	22347
T4 holiday	22/12/18	20404
	…	…
	28/01/19	16319
Selected Non-School Holiday data		
Non-holiday interval	**Date**	**Sum of NE, SE, S, SW**
T1 Non-holiday	29/01/18	14782
	…	…
	29/03/18	13416
T2 Non-holiday	16/04/18	16011
	…	…
	29/06/18	15314
T3 Non-holiday	16/07/18	14570
	…	…
	21/09/18	16198
T4 Non-holiday	08/10/18	15878
	…	…
	21/12/18	14085

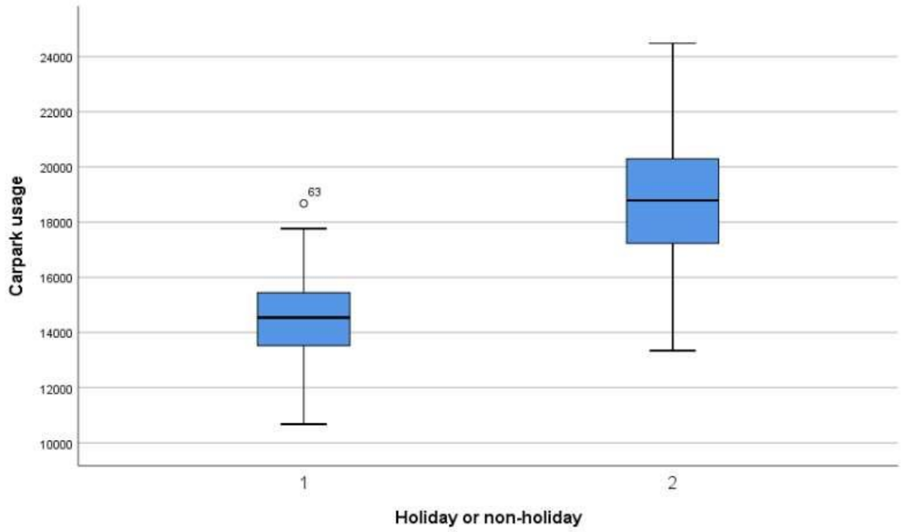

Fig.3. Box plot for holiday and non-holiday

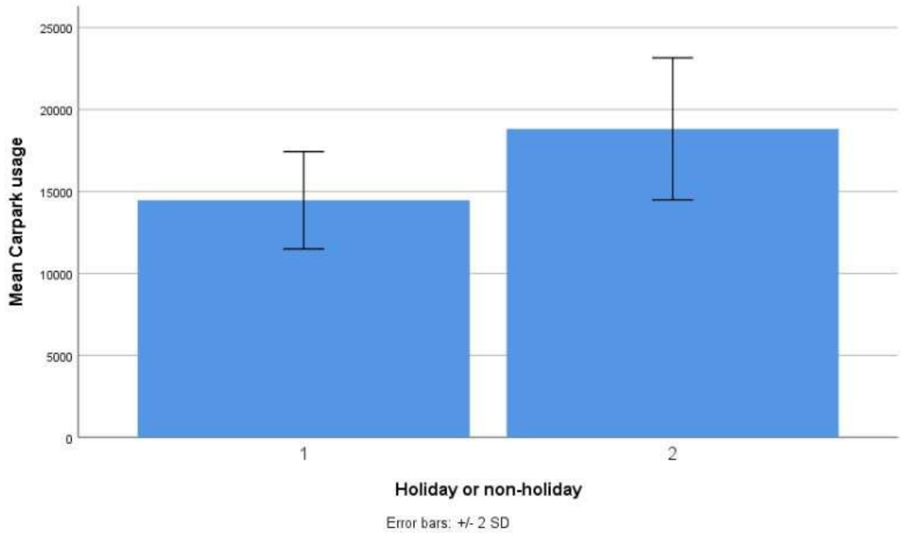

Fig.4. Error bar chart

As required, the t-test is performed to estimate the selected data. Since two independent samples were taken from two different populations: the cars entering carparks in holidays and non-holidays. Thus, a two-sample T-test is appropriate for this scenario, and also, we would like to know the difference between the mean, estimate the difference between the means of these populations. The null hypothesis would be that there's no difference, the mean difference between holidays and non-holidays would be equal to zero:

$$H_0: \text{Mean}_{\text{Holiday}} - \text{Mean}_{\text{Non-Holiday}} = 0$$

And the alternative hypothesis would be that there's the difference, the mean difference between holidays and non-holidays would not be equal to zero:

$$H_a: \text{Mean}_{\text{Holiday}} - \text{Mean}_{\text{Non-Holiday}} \neq 0$$

Results and discussion

Table 3. Holiday vs non-Holiday t-test

Holiday vs Non-holiday
Two Sample t-test
H_0: Mean $_{\text{holiday}}$ – Mean $_{\text{non-holiday}}$ = 0 H_a: Mean $_{\text{holiday}}$ – Mean $_{\text{non-holiday}}$ $\neq$ 0
t = -21.18, p-value < 0.01
alternative hypothesis: true difference in mean $\neq$ 0

As shown in table 3, If that p-value is below the significance level, then this was a unlikely scenario, reject the null hypothesis, which would suggest the alternative. But if your p-value is greater than your significance level, then you would fail to reject your null hypothesis, and so you would not have sufficient evidence to conclude the alternative [1]. As assumed that working with a significance level of 0.05, refer to the table, the p-value is less than 0.01, which is less than our alpha value, 0.05. Therefore, the null hypothesis is rejected and accept the alternative hypothesis. And to answer the question, there is sufficient evidence to conclude that there is a difference between differences between car park usage during school holidays and non-school holidays.

The discussion regarding whether the t-test is appropriate for this data

The normality test is performed to determine whether the data is normally distributed:

- Null hypothesis (H0): Sample follows a normal distribution
- Alternative hypothesis (Ha): Sample doesn't follow a normal distribution

Table 4. Holiday vs non-Holiday normality-test

Sample	w	pvalue
Holiday	1.00	0.99
Non-Holiday	1.00	0.63

According to the result of the normality test of the holiday vs non-holiday dataset, we can draw the following conclusion. For the holiday, p-value is 0.99, which is greater than the alpha value, 0.05. Thus, the null hypothesis is accepted and the alternative hypothesis is rejected.

A holiday sample is normally distributed. While the non-holiday has p-value of 0.63 which is greater than 0.05, which means the null hypothesis is accepted and the alternative hypothesis is rejected. Non-Holiday sample is also normally distributed. Therefore, the t-test is appropriate for this question, two individual sample collected from a separate population. Thus, used two-sample t-test. And t-test is a parametric test that assumes the distribution of the data is normal which matches the normality test results where the two data are normally distributed

4 Are there differences in the number of cars using different terminal carparks?

Data selection

As the question asks about the difference of terminal carparks, the whole dataset is used in this case. The different terminals include the NE, SE, SW, S, and the count of the number of the cars entering these four carparks respect to the date from 29/01/2018 to 28/01/2018.

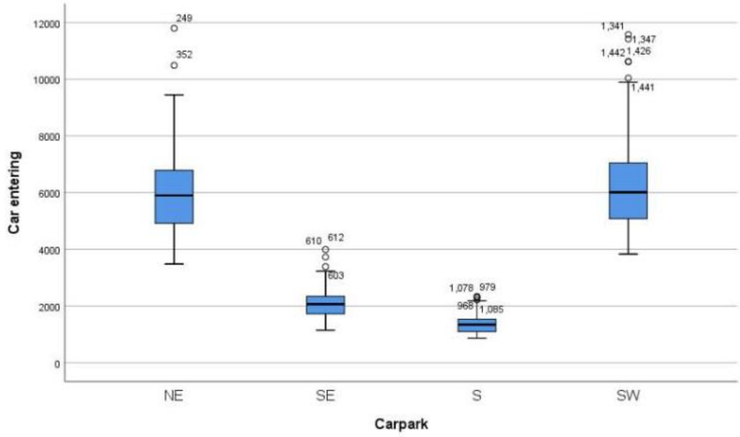

Fig.5. Box-plot of carparks

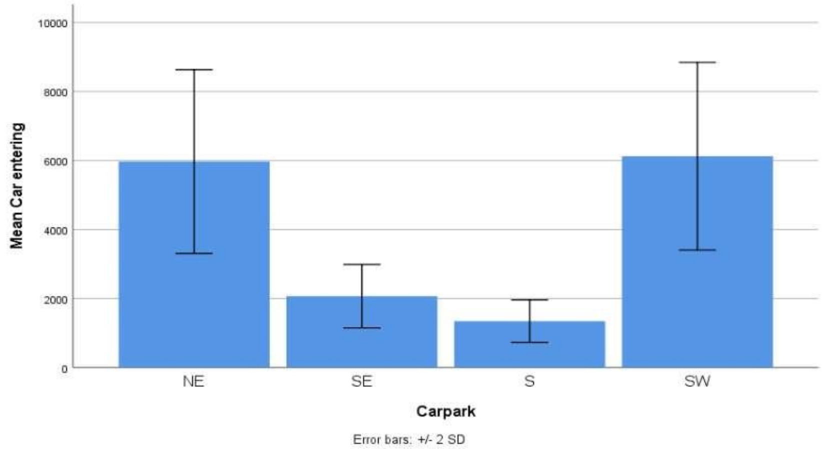

Fig.6. Error bar chart

Results and discussion

As required, ANOVA-test is performed to estimate the difference between the four-terminal carparks. The null hypothesis would be that there's no difference between NE, SE, SW, S, and the alternative hypothesis would be that there's a difference between NE, SE, SW, S:

Table 5. NE, SE, S, SWANOVA-test

Sample	**ANOVA test**	
	t	**pvalue**
NE, SE, S, SW	2367.46	0

According to the results shown above, the p value is equal to 0.00, is less than the alpha 0.05. Therefore, the null hypothesis is rejected and accept the alternative hypothesis. And to answer the question, there is a significant difference between the average usage of these four carparks. However, when the pvalue equal to 0, it might show that the result might be faulty.

The t-test is conducted to between each pair of the carparks to determine whether these carparks have any difference, thus there are 6 tests between these carparks displayed respectively below. And to generalize the hypothesis based on the previous presented ANOVA test, the null hypothesis would be that there's no difference between any two of the carparks, and the alternative hypothesis would be that there's the difference between any two of the carparks:

- Null hypothesis (H0): Sample follows normal distribution
- Alternative hypothesis (Ha): Sample doesn't follow normal distribution

Table 6. NE, SE, S, SW T-test

Sample	T-test	
	t	**pvalue**
NE, SE	52.92	<0.01
NE, S	64.33	<0.01
NE, SW	-1.55	0.12
SE, S	24.96	<0.01
SE, SW	-53.99	<0.01
S, SW	-65.48	<0.01

According to the results shown above, the p value for test set (NE, SE), (NE, S), (SE, S), (SE, SW), (S, SW) are all less than 0.01 which is less than the alpha 0.05, means that the null hypothesis is rejected and accept the alternative hypothesis. There is a significant difference between average usage of (NE, SE), (NE, S), (SE, S), (SE, SW), (S, SW). However, the p value of NE and SW carpark is 0.12 which is greater than the 0.05. Thus, the null hypothesis fails to reject which means NE and SW are no difference. However, there would be Type I error appears if running the t-test over the same data over multiple times, the error is increased by 5% inclemently which resulting in unacceptable errors [2].

Generally, the normality test is performing before the t-test or ANOVA test to determine whether the data is normal or not which determine the test is appropriate or not. Thus, a normality test has performed to the NE, SW, S SE data. The hypothesis would be:

- Null hypothesis (H0): Sample follows normal distribution
- Alternative hypothesis (Ha): Sample doesn't follow normal distribution

Table 7. NE, SE, S, SW Normality-test

Sample	Normality Test	
	w	**pvalue**
NE	0.97	<0.01
SE	0.97	<0.01
S	0.96	<0.01
SW	0.96	<0.01

As shown in the table, it is clear that all the samples have a pvalue that is less than 0.01 which is less than the 0.05 and means the null hypothesis is rejected. Therefore, the data of NE, SE, S, SW are not normal.
Therefore, the kruskal-wallis test is used to replace the ANOVA test because it is suitable for not normally distributed data. And the null hypothesis would be that there's no difference between NE, SE, SW, S, and the alternative hypothesis would be that there's the difference between NE, SE, SW, S:

Table 8. NE, SE, S, SW Kruskal-Wallis test

Sample	Kruskal-Wallis test	
	h	**pvalue**
NE, SE, S, SW	1186.54	0.01

Pvalue is 0.01 which is less than 0.05, the null hypothesis rejected which means these four car parks are different.

And the u-test is performed to replace the t-test since u-test is a non-parametric test which suitable for the data is not normal:

Table 9. NE, SE, S, SW U-test

Sample	U-test	
	u	**pvalue**
NE, SE	20	<0.01
NE, S	0	<0.01
NE, SW	62121	0.06
SE, S	12029.5	<0.01
SE, SW	6	<0.01
S, SW	0	<0.01

According to the results shown above, the p value for test set (NE, SE), (NE, S), (SE, S), (SE, SW), (S, SW) are all less than 0.01 which is less than the alpha 0.05, means that the null hypothesis is rejected and accept the alternative hypothesis. There is a significant difference between average usage of (NE, SE), (NE, S), (SE, S), (SE, SW), (S, SW). However, the p value of NE and SW carpark is 0.06 which is greater than the 0.05. Thus, the null hypothesis fails to reject which means there is a statistically significant difference between NE and SW.

References

1. Dallal, Gerard E. The Little Handbook of Statistical Practice. (2012).
2. Betz, M.A., Gabriel, K.R., Type I Errors and Analysis of Simple Effects, Journal of Educational Statistics. (2018)
3. Montgomery. The analysis of variance; Practical interpretation of results; Comparing means with a control (2001)