Exploration of the Center childcare capacity rate in Toronto

1. Introduction

Toronto has a shortage of licensed childcare services for children- the following study will tackle the problem of finding affordable and quality childcare and collects information on the operation and capacity of these centres for multiple age groups. The dataset used in the study is called "INF2178_A2_data.xlsx". It contains several attributes of Childcare centres such as 'Operating auspice', 'Childcare spaces for infants 0-18 months', 'Childcare spaces for toddlers 18-30 months', 'Childcare spaces for all age groups' and some others related variables.

This report will offer a comprehensive data analysis on childcare center capacity in Toronto. One-way-anova and two-way anova will be the main statistical method that to be used to discover the impact of potential variables on the childcare capacity rate.

This analysis will focus on investigating the childcare space capacity rate for toddlers 18-30 months, to precisely discover the impact of different attributes on this capacity rate.

Our exploration will address two fundamental research questions:

Research question: 1. Is there a significant difference in childcare capacity rate for toddlers 18-30(TGSPACE_RATE) under different types of operations auspice. (AUSPICE)?

2. Does the different types of operation auspice(AUSPICE) and whether Centre has a fee subsidy contract-Yes/No(subsidy) cause a interaction impact and a significant difference on childcare capacity rate for toddlers 18-30 months(TGSPACE_RATE)?

2. Data Cleaning and Data Wrangling

The raw dataset has a total of 17 columns with 1063 rows, we have reduced our working data to 4 columns.

AUSPICE- Operating auspice (Commercial, Non Profit or Public)
TGSPACE- Child care spaces for toddlers 18-30 months
TOTSPACE- Child care spaces for all age groups
subsidy - Centre has a fee subsidy contract (Yes/No)

After we checked, there is no missing value in the dataset.

Feature Engineering:

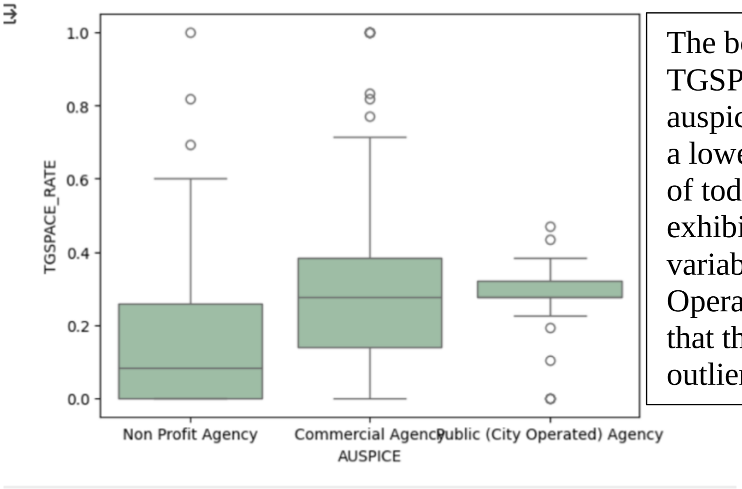
Adding a new columns named: TGSPACE_RATE- Calculate the Childcare spaces for toddlers 18-30 months capacity rate within total space in all age groups(TOTSPACE)

TGSPACE_RATE = TGSPACE / TOTSPACE

Now we have totally five columns and 1063 rows in the dataset.

3. One-way-anova data analysis

Now perform the one-way-anova test to discover the first research question.



The box plot provides a visual comparison of TGSPACE_RATE among different Operating auspice types (AUSPICE). Non profit agency shows a lowest median, indicate that a lower capacity rate of toddlers 18-30 months. Commercial Agency exhibit a higher median and a wide range, shows the variability of toddler capacity rate. Public(City Operated) Agency has a narrow range which implies that the lower variation of the rate. There are some outliers between all of the three types of auspice.

We will now check anova table statistics:

	df	Sum_sq	Mean_sq	F	PR(>F)
C(AUSPICE)	2.0	4.358882	2.179441	78.268	1.977e-32
Residual	1060.0	29.516651	0.027846	NaN	NaN

High F-statistics(78.268) shows that there is a significant difference of capacity rate of toddlers 18-30 months in non profit, commercial and public(city operated) agency, This statistical evidence also reinforced by the p-value which is far below 0.05.

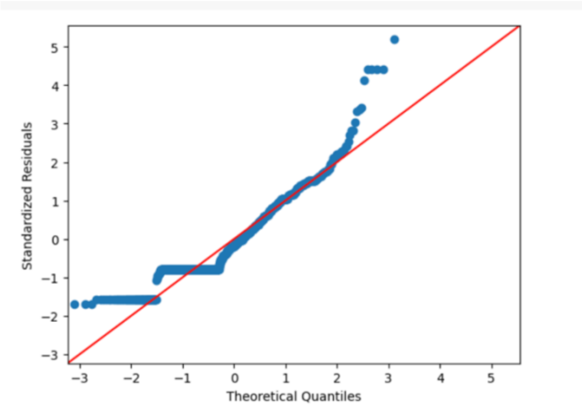
After, post hoc test using Tukey's HSD

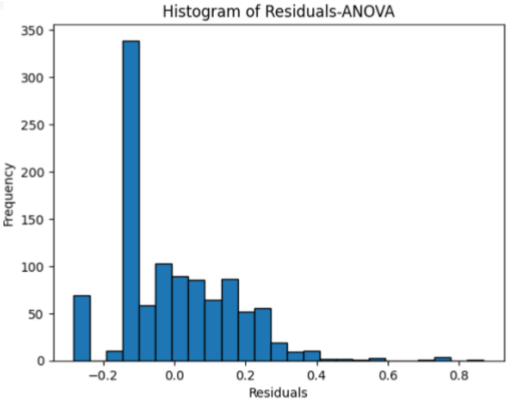
	Group1	Group2	Diff	Lower	Upper	q-value	p-value
0	Non Profit Agency	Commercial Agency	0.133137	0.106754	0.159520	16.750	0.001
1	Non Profit Agency	Public (City Operated) Agency	0.151498	0.087067	0.215929	7.805	0.001
2	Commercial Agency	Public (City Operated) Agency	0.018361	-0.048054	0.084778	0.918	0.773

The Tukey's HSD post-hoc test results show significant differences in capacity rate of toddlers 18-30 months among the Operating auspice (Commercial, Non Profit or Public) groups. Non-Profit Agencies and Commercial Agencies have a small mean difference of 0.133137, with p-value-0.001, suggesting a meaningful discrepancy in toddler capacity rate. For Non-Profit vs. Public (City Operated) shows a mean difference-0.151498, supported with a significant p-value of 0.001, indicating a significant difference in capacity rate. However, commercial and Public agencies does not have strong evidence to show a significant difference in toddlers capacity rate, reflected by a p-value of 0.772.

We need to check the assumptions for anova:

Normality assumption.





The Q-Q plot displays the standardized residuals vs the theoretical quantiles to visualize the normality, as the point deviate clearly in the tails, this indicates the potential violations of normality, also with the presence of outliers.

The histogram of the distribution of anova residuals is also created, which clearly shows a skewness-skewed to the right, it shows a deviation of normal distribution.

Shapiro Wilk test for normality

From this test for the normality-Shapiro wilk test, the w statistics is approximately 0.921, and the p value is smaller than 0.001, this result shows significant evidence to reject the normality.

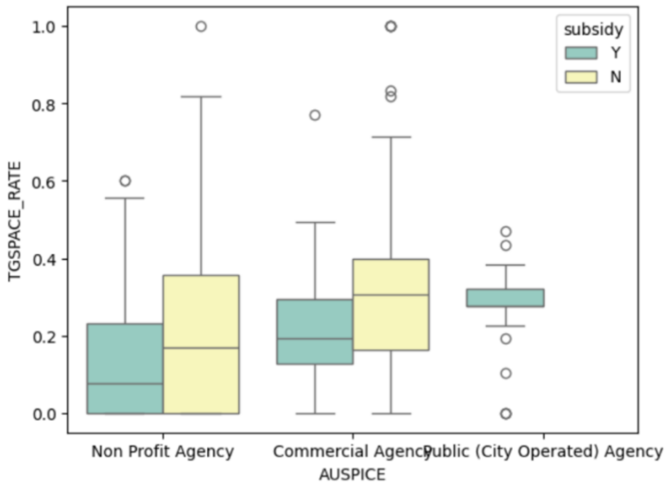
Equal variance assumption (Homogeneity of variance)

The Levene's test is applied when the sample is not normally distributed, for the result, p-value is smaller than 0.05, this implies the significant discrepancy about variance in this data group. This suggested that the assumption of equal variance is violated.

From the above result, the assumptions are not being met, the result of one-way-anova might not be perfectly reliable.

4. Two-way-anova data analysis

Using two-way-anova test for second research question.



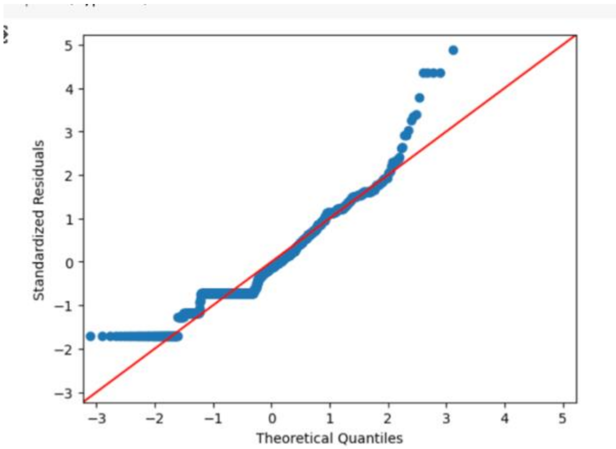
The boxplot displays the distribution of the toddlers 18-30 months capacity rate (TGSPACE_RATE) across different Operating auspice types (AUSPICE), with an additional variable indicating whether the childcare center has a fee subsidy contract (subsidy). Non-Subsidized agency usually has a higher median and variation of the capacity rate. Commercial Agency shows a more noticeable difference of capacity rate between subsidized and non-subsidized childcare centers, with supported by non-subsidized has higher median, also by the wider discrepancy between those two medians. Public(City Operated) agency demonstrate a smallest capacity rate difference between non-subsidized and subsidized groups.

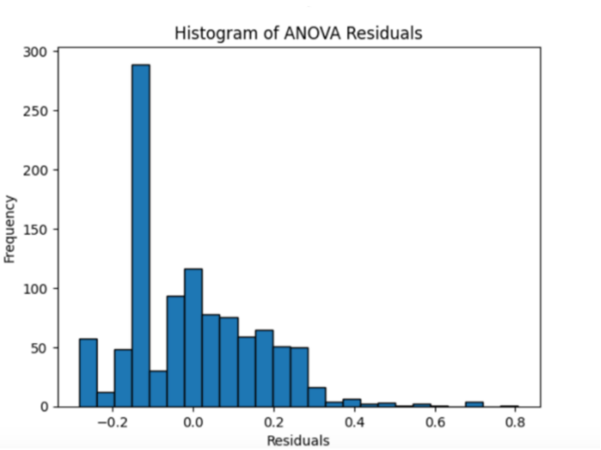
Two-way-anova test statistics results:

index	df	sum_sq	mean_sq	F	PR(>F)
C(AUSPICE)	2.0	1.80576235279101	0.902881176395505	33.214	1.023e-14
C(subsidy)	1.0	0.6058499739648906	0.6058499739648906	22.287	2.663e-06
C(AUSPICE):C(subsidy)	2.0	0.16243506108011316	0.08121753054005658	2.988	0.051
Residual	1058.0	28.760089459189878	0.027183449394319357	NaN	NaN

The two-way anova result present the analysis of the effect of Operating auspice types (AUSPICE), whether center has a fee subsidy contract (subsidy), and their interaction effect on the toddlers 18-30 months capacity rate (TGSPACE_RATE). For auspice types, the F-statistic is 33.214, and the p-value is extremely low, smaller than 0.01, suggesting a highly significant effect of the type of childcare agency on the capacity rate, the subsidy status also has a significant effect, with an F-statistic of 22.29 and a p-value below 0.001. However, the interaction effect between these two variables has an F-statistic of 2.988 and a p-value of approximately 0.05, this suggests that the combined effect of operation auspice type and subsidy status on toddler capacity rates is existed but weaker compared to their individual effects.

Now we check the assumption of two-way-anova:
Normality assumption.
Equal variance assumption (Homogeneity of variance)





From the QQ-plot, as the point are deviating, this shows the violation of normality assumption, also the p value from Shapiro Wilk test is smaller than 0.001, this result shows significant evidence to reject the normality, this also supported by the histogram-right skewed.

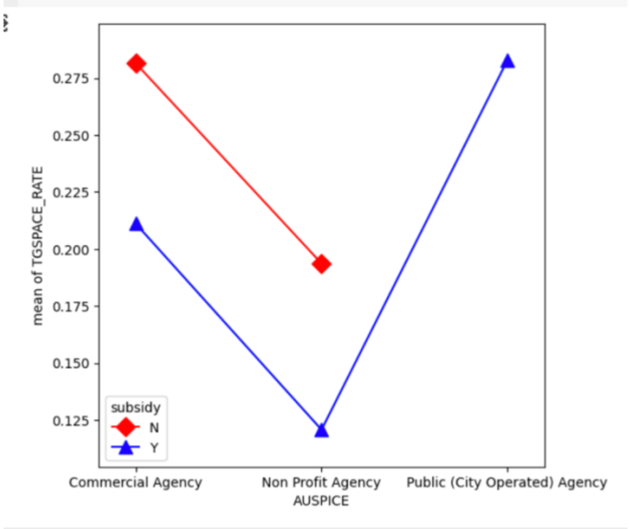
After we use the Levene's test, it has the result of test-statistics 14.23 and p value is really small(<0.001), indicate that the assumption of equal variance is violated in this anova model.

Post-hoc test results:

index	group1	group2	Diff	Lower	Upper	q-value	p-value
0	Non Profit Agency,Y	Non Profit Agency,N	0.07295415520233192	0.02254770547645476	0.12336060492820908	5.843	0.001
1	Non Profit Agency,Y	Commercial Agency,Y	0.09044648920131326	0.03379698424431655	0.14709599415830998	6.446	0.001
2	Non Profit Agency,Y	Commercial Agency,N	0.1608233967525002	0.1250404073544212	0.1966063861505792	18.146	0.001
3	Non Profit Agency,Y	Public (City Operated) Agency,Y	0.16208349154028076	0.08430379995583946	0.23986318312472205	8.414	0.001
4	Non Profit Agency,Y	Public (City Operated) Agency,N	0.0	-Infinity	Infinity	0.0	0.9

From this test result, The toddlers 18-30 months capacity rate of Non profit agency with subsidized has significant difference than Non profit agency without subsidized, subsidized commercial agency, non-subsidized commercial agency, subsidized public agency. With a mean difference of 0.073, 0.09, 0.161, 0.162, also supported by p values which are all smaller than 0.05(0.001). However, for non profit subsidized agency vs non-subsidized public agency, we do not have enough evidence to show that there is significant difference in capacity rate, also shown by p value = 0.9.

At the end, we look the interaction plot:



This interaction plot show the relationship between Operating auspice types (AUSPICE), Subsidy status(subsidy), and their interaction effect on the rate of toddlers capacity. The plot shows that non profit agency with subsidy has lower rate of toddler capacity than without subsidy. Commercial Agencies exhibit a lower rate with subsidy compared to without. There is no non-subsidy showing in public agency, might because of it belongs to the government. Overall, the capacity rate of commercial agency is higher than non profit agency generally.

5. Conclusion

For the first question, we are using the one-way anova, we have concluded that the toddlers capacity rate with non profit agency vs commercial agency and non profit agency vs public agency indicated a significant difference, but for commercial agency vs public agency, we do not have enough evidence to show the significant capacity rate difference between.

For the second research question, two-way-anova results indicated that both types of operation auspice and subsidy status cause a significant effect on the toddlers capacity rate. The interaction effect of these two variables also suggest a potential impact for the capacity rate.

Pay attention to the assumption for one-way and two-way anova are not perfectly met, the normality assumption and equal variance assumption are violated. Which implies that the result might not be reliable and 100 percent accurate, need to be paid extra attention.