INF2178 A2 - Exploring Licensed Childcare Centers in Toronto

Yuyang Liu, 1005965617

Introduction & Research Question

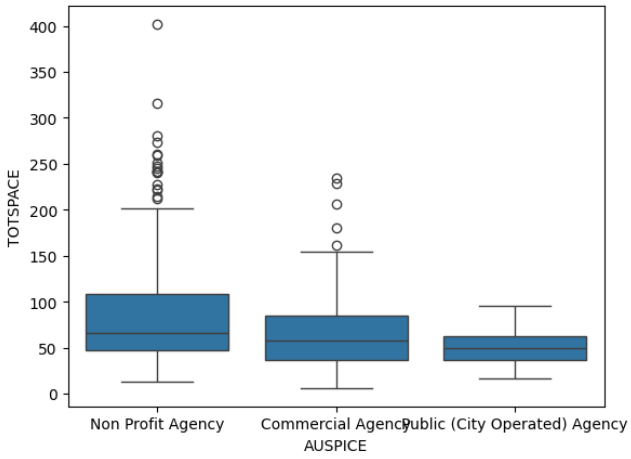
For the purpose of this project, we will be examining the dataset called "INF2178_A2_data". The dataset collects information on the operation and capacity of these centers for multiple age groups (updated February 2024). It has been difficult to find proper child care for children in Ontario due to high fees and low availability of spaces. This project is intended to explore some trends and conduct analyses in Toronto licensed child care centers.

The research questions we will be interested in include the followings:

1. *How does operating auspice affect the space availability of child care?*
2. *What is the impact of auspice and subsidy on child care spaces? Is there a significant interaction effect between auspice and subsidy on space availability?*
3. *What is the impact of auspice and CWELCC participation on child care spaces? Is there a significant interaction effect between the two variables on space availability?*

One-Way ANOVA Test – Auspice on Child Care Space

First, before conducting an ANOVA test, let's build a boxplot to view the data distribution of total child care space based on different categories of auspice.



From the boxplot we can see that total space for non-profit agency is greater than the total space for commercial and public agencies. Non-profit agency has some outliers that have very large capacity.

Then, we conduct a one-way ANOVA test of operating auspice on child care space availability. The result is shown in the below table:

	sum_sq	df	F	PR(>F)
C(AUSPICE)	9.61e+04	2	21.84	5.06e-10
Residual	2.33e+06	1060	NaN	NaN

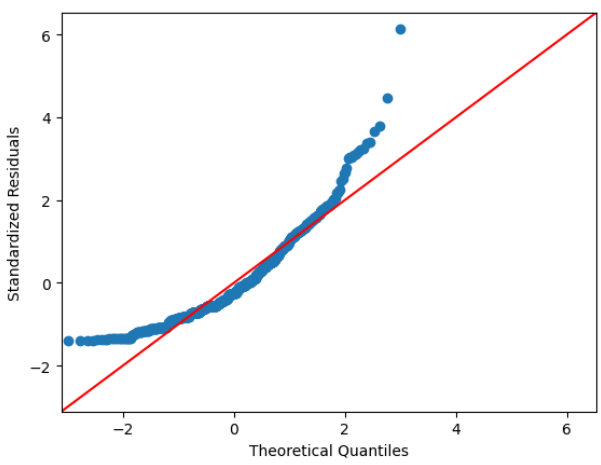
The result shows a p-value smaller than any significant level, it indicates that child care spaces do have significant difference across auspice categories.

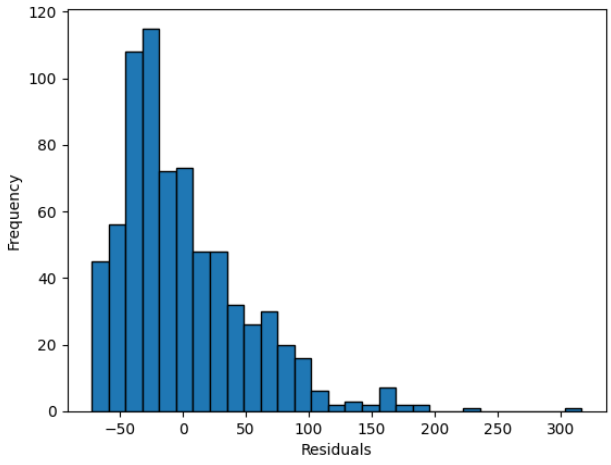
To further explore the trend, we will also conduct a post hoc test. The result is shown below:

Group 1	Group 2	Diff	Lower	Upper	q-value	p-value
NPA	CA	16.81	3.99	29.62	4.36	0.006
NPA	PA	36.18	8.67	63.68	4.37	0.006
CA	PA	19.37	-10.14	48.88	2.18	0.273

The first two rows shows a significant p-value, indicating that non-profit agency to have a significant difference with the other two types of agencies in terms of child care capacity.

We need to test the assumptions to see the fit of the ANOVA test, we first plot the dot plot and histogram of residuals to observe the trend. We can clearly see that residuals are not normally distributed, the first assumption is violated.





We then conduct the Shapiro Wilk test and obtain a statistic of 0.902 and a p-value of 1.496e-25. This implies that the residuals are not normally distributed, aligning with the plots we observe from above.

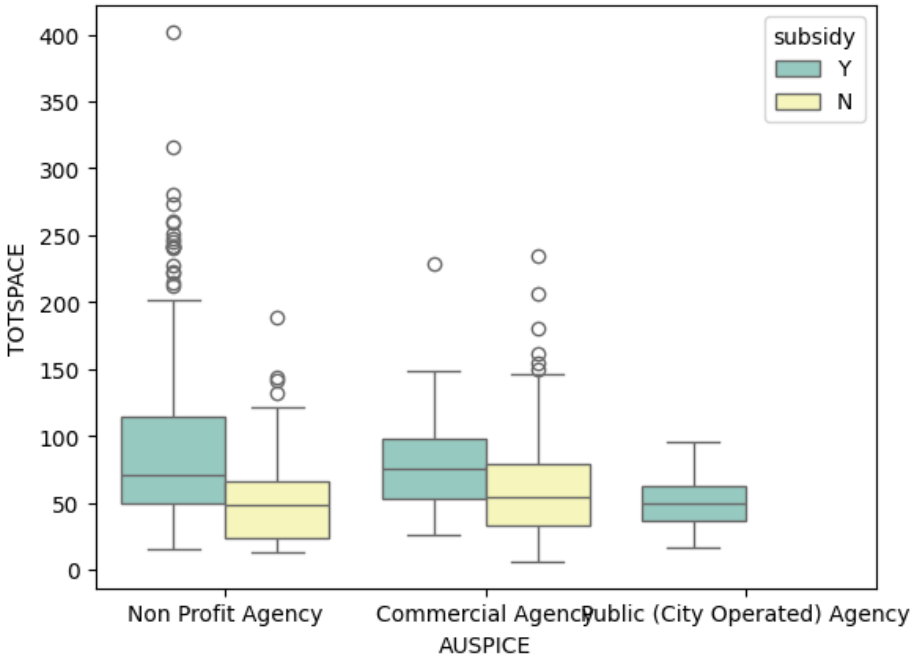
A Levene's test is conducted to test assumption 2 for data that is not normally distributed.

Parameter	Value
Test statistics (W)	9.199
Degrees of freedom (Df)	2.000
p-value	0.0001

The p-value is smaller than 0.001, the second assumption of homogeneity of variances is also violated. Two assumptions are violated, therefore the test results should be interpreted with caution.

Two-Way ANOVA Test – Auspice & Subsidy on Child Care Space

In addition of the one-way ANOVA, we are also interested in a two-way ANOVA of the impact of auspice category and subsidy on child care capacity. Similar to the one-way ANOVA, we first construct a boxplot to view the distribution.



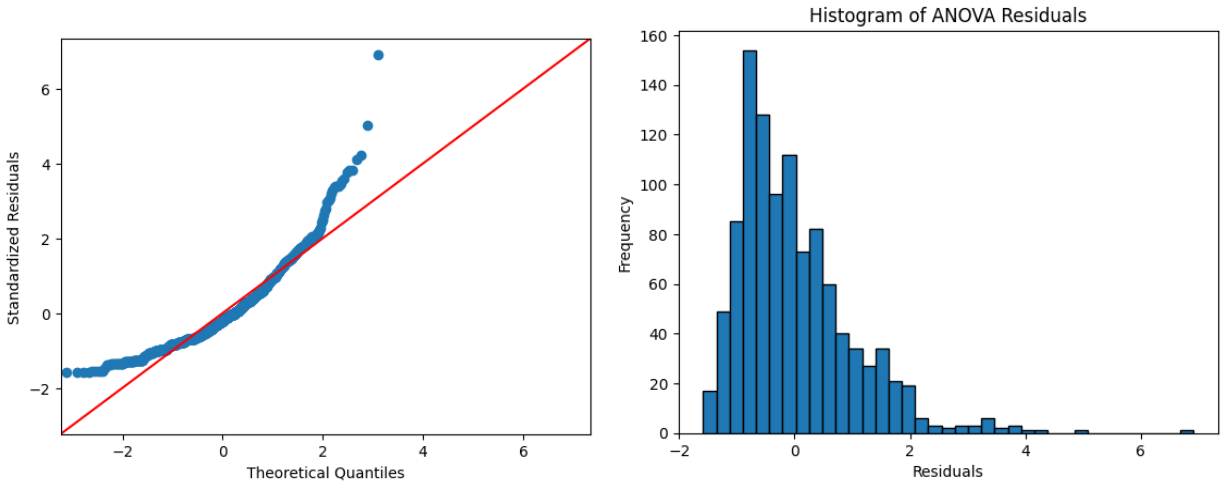
At a first glance, we can see that public agency does not have non-subsidy category, which makes sense as they are operated by the city, which automatically receive subsidy from the government. The other two categories of auspice do not exhibit a distinct difference, but the non-profit agency have some outliers on the top.

The two-way ANOVA test results are shown below:

	Sum_sq	df	F	PR(>F)
C(AUSPICE)	8.57e+03	2	2.06	1.28e-01
C(subsidy)	8.35e+04	1	40.12	3.53e-10
C(AUSPICE):C(subsidy)	5.60e+04	2	13.46	1.69e-06
Residual	2.20e+06	1058	NaN	NaN

The two-way ANOVA test incorporates both auspice categories and subsidy variables, and show that the results are significant as p-value is smaller than 0.001. It means that even when we include the subsidy variable, we can still conclude that different auspice types have distinct capacity availability.

As usual, we construct a dot plot, a histogram, and a Shapiro Wilk test to view the trend of residuals and to test assumption 1. Looking at the plots and p-value of the Shapiro Wilk test, we can see assumption 1 is violated. The residuals are not normally distributed.



Shapiro Wilk test provides a statistic of 0.902 and a p-value of 1.53e-25.

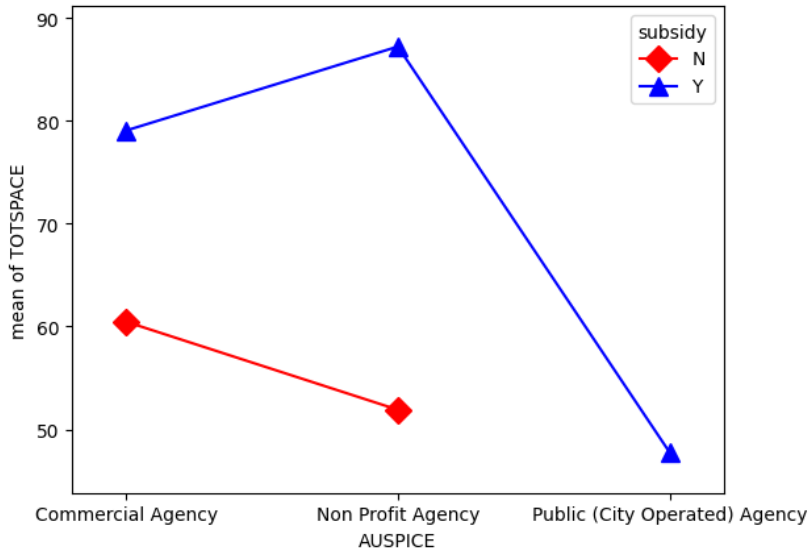
Then we perform the Levene's test for homogeneity of variances. We get a test statistic of 12.98 and a p-value of 2.515e-10. This means the second assumption is also violated.

We then conduct a post hoc test using Tukey's HSD, obtaining the following results:

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	(Non Profit Agency, Y)	(Non Profit Agency, N)	44.986164	25.584897	64.387430	9.370204	0.001000
1	(Non Profit Agency, Y)	(Commercial Agency, Y)	0.145138	-28.465689	28.755965	0.020500	0.900000
2	(Non Profit Agency, Y)	(Commercial Agency, N)	28.370552	11.192465	45.548638	6.674107	0.001000
3	(Non Profit Agency, Y)	(Public (City Operated) Agency, Y)	40.752830	8.307960	73.197700	5.075887	0.004757
4	(Non Profit Agency, Y)	(Public (City Operated) Agency, N)	0.000000	-inf	inf	0.000000	0.900000

The first, third, and forth row have p-value less than 0.01. Looking at the group comparison, we see that not only non-profit agency differs from the other two agencies, but also within the non-profit agency, with or without subsidy affects the child care space ablativity as well.

We are also interested in the interaction plot. We construct the plot and observe that subsidy type does not meet each other. Indicating a difference between the categories.



Two-Way ANOVA Test – Auspice & CWELCC participation on Child Care Space

Moreover, we would like to explore another two-way ANOVA of the impact of auspice category and CWELCC participation on child care capacity. Going through all the step from the previous two-way ANOVA test, we obtain a very similar results (Due to space limit, please refer detailed graphs and statistics to the code).

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	(Non Profit Agency, Y)	(Non Profit Agency, N)	31.483813	4.510872	58.456754	4.716927	0.011515
1	(Non Profit Agency, Y)	(Commercial Agency, Y)	15.792764	-2.624464	34.209991	3.465246	0.140465
2	(Non Profit Agency, Y)	(Commercial Agency, N)	24.032200	-1.788679	49.853080	3.761168	0.085078
3	(Non Profit Agency, Y)	(Public (City Operated) Agency, Y)	37.832200	4.565676	71.098725	4.595728	0.015278
4	(Non Profit Agency, Y)	(Public (City Operated) Agency, N)	0.000000	-inf	inf	0.000000	0.900000

The results also show evidence that non-profit agency has different space availability than the other two types of operating auspice. However, the p-values for this two-way ANOVA test and post hoc test show that participation of CWELCC may or may not have significant different child care spaces, depending on the significant level chosen. All p-values are greater than the 0.01 significant level, but some of them are less than 0.05.

When testing the assumptions of normality of residuals and homogeneity of variances. Both assumptions are unmet. This imposes a question regarding the fitness of the model.

Conclusion

In conclusion, the results from the tests we conducted in this project have shown that non-profit agency seems to have significant difference in child care space availability comparing to public agency and commercial agency. The two-way ANOVA test of auspice & subsidy on total space shows significant difference in both auspice and subsidy types. But the two-way ANOVA test of auspice & participation of CWELCC on total space may or may not exhibit significant difference depending on the significance level chosen.

However, all three tests conducted in this experiment do not meet the assumptions of residuals normality and variances homogeneity. As a result, all the outcomes should be interpreted with caution and we have to rethink and further explore the fitness of the model we utilized.