Exploring Toronto Childcare Centres

Zhuoying Li

1004021202

Introduction

Research Questions

Question 1: Is there a significant difference between childcare spaces for different age groups when comparing various auspice (commercial, nonprofit, or public)?

Question 2: Is there a significant different between total available spaces for different operating auspice and the availability of a subsidy contract?

Data Cleaning

The dataset contains 1063 rows in total, with 17 columns. For our analysis, there are a few columns that are not needed, such as ID. Below is a brief description of all the columns we are interested in, and if they are renamed, the new column name.

- **AUSPICE** (renamed to **auspice**) – Operating auspice, such as commercial, nonprofit or public)
- **ward** – City ward number for the location.
- **IGSPACE** (renamed to **infant**) – Childcare space for 0-18 months.
- **TGSPACE** (renamed to **toddler**) – Childcare space for 18-30 months.
- **PGSPACE** (renamed to **preschool**) – Childcare space for 30 months to grade 1.
- **KGSPACE** (renamed to **kindergarten**) – Childcare space for full-day kindergarten.
- **SGSPACE** (renamed to **schoolage**) – Childcare space for children grade 1 and up.
- **TOTSPACE** (renamed to **total**) – Childcare space for all age group (sum of IGSPACE, TGSPACE, PGSPACE, KGSPACE, and SGSPACE).
- **subsidy** – Does the centre have a fee subsidy contract with the city?

ID column is kept as a unique primary key for each row, the rest columns are discarded. Following table shows the first 5 rows of cleaned data.

ID	AUSPICE	WARD	INFANT	TODDLER	PRESCHOOL	KINDERGARTEN	SCHOOLAGE	TOTAL	SUBSIDY
1	Non Profit Agency	3	0	20	32	52	60	164	Y
2	Non Profit Agency	8	0	0	12	26	45	83	Y
3	Non Profit Agency	25	0	10	16	26	50	102	Y
4	Non Profit Agency	10	10	15	40	0	0	65	Y
5	Non Profit Agency	20	0	10	16	0	0	26	Y

Figure 1: First 5 rows of cleaned data.

Exploratory Data Analysis (EDA)

First, let's do some exploratory data analysis by counting number of rows for various categorical variables.

For auspice, we got the following. There are very few public operated agencies, and quite a lot of non-profit agency in the dataset.

Non Profit Agency	703
Commercial Agency	321
Public (City Operated) Agency	39

For subsidy, which is a binary variable, we got the following. We can see most childcare centers have the subsidy agreement with the city.

Y	718
N	345

For ward, it is hard to look at a table because there are so many options, so a bar graph is drawn to better visualize count of rows by ward.

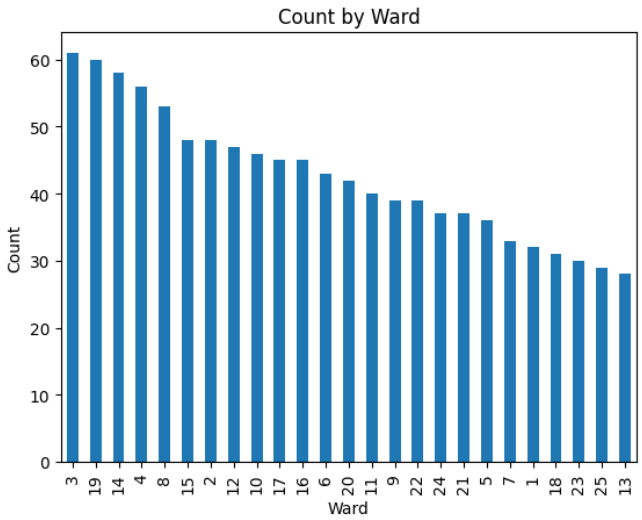

Figure 2: Count by Ward

Subsidy might have some effect on space available, so we draw a bar graph on both different age group capacity and the availability of a subsidy contract. From some preliminary analysis, seems like subsidy have significant positive impact on kindergarten and schoolage groups, while having negative impact on toddler and preschool age groups.

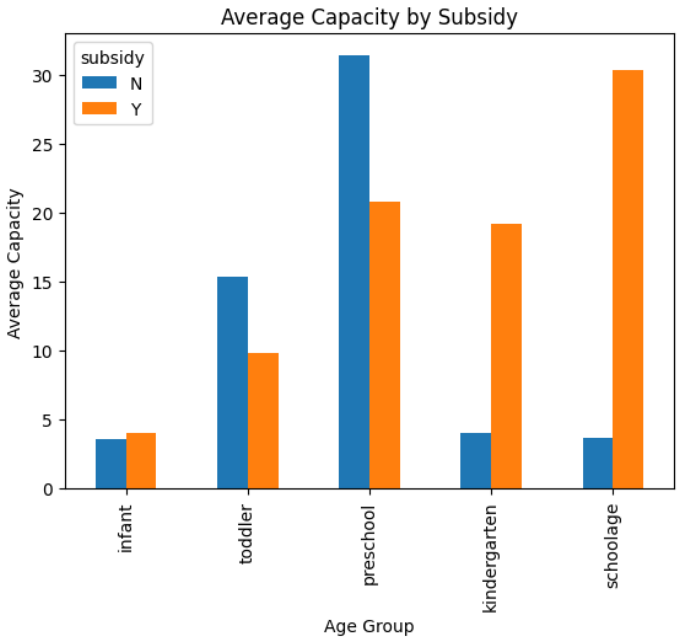

Figure 3: Average Capacity by Subsidy

Next, we draw a comparison bar graph on average total space available based on subsidy, grouped by auspice. We can see that on average, subsidy does have a positive impact on space available. One thing to note is that public agency always has subsidy.

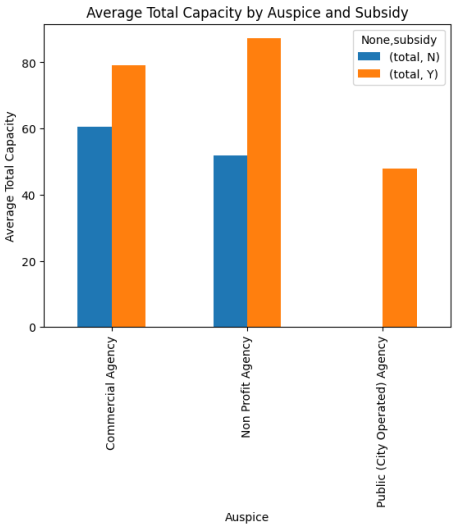

Figure 4: Average Total Capacity by Auspice and Subsidy

Finally, we draw a boxplot on the same group to see the mean and distribution, seems like almost none of the groups follow a normal distribution by looking at the box plot.

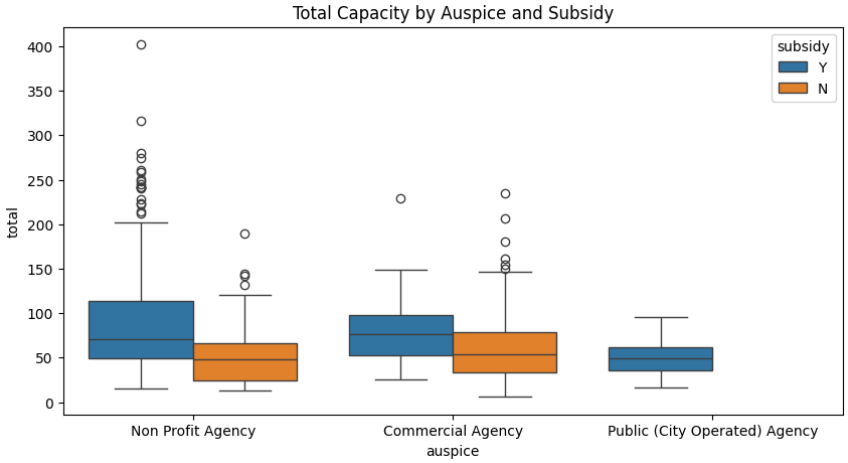

Figure 5: Total Capacity by Auspice and Subsidy

Space Available by Age Across Operating Auspices

In this section, we use one-way ANOVA to answer our first research question:

Question 1: Is there a significant difference between childcare spaces for different age groups when comparing various auspice (commercial, nonprofit, or public)?

H_0: There is no significant difference between mean of childcare spaces for different age groups, across different operating auspice (treatment).

H_1: At least one pair of means of the childcare spaces available is significantly different from each other.

A one-way ANOVA model is created using the formula “capacity ~ C(age_group)”, and produces the following ANOVA table.

	SUM_SQ	DF	F	PR(>F)
C(AGE_GROUP)	2.821233e+05	4.0	188.190768	4.517383e-151
RESIDUAL	1.990101e+06	5310.0	/	/

Figure 6: One-Way ANOVA Table

Because the P-value is much less than the significance level of 0.001, **we reject our null hypothesis**. This means there is sufficient evidence to suggest at least one of the mean childcare spaces is different from others across various operating auspices.

A TukeyHSD test is conducted to see our findings in detail, we can see in the following table, every P-value is < 0.05, therefore there is a significant variation across all pairs.

COMPARISON	STATISTIC	P-VALUE	LOWER CI	UPPER CI
0-1	-10.361	0.000 (< 0.05)	-12.653	-8.070
0-2	-20.362	0.000 (< 0.05)	-22.654	-18.071
0-3	-17.765	0.000 (< 0.05)	-20.056	-15.473
0-4	-7.704	0.000 (< 0.05)	-9.995	-5.412
1-0	10.361	0.000 (< 0.05)	8.070	12.653
1-2	-10.001	0.000 (< 0.05)	-12.292	-7.710
1-3	-7.404	0.000 (< 0.05)	-9.695	-5.112
1-4	2.658	0.014	0.366	4.949
2-0	20.362	0.000 (< 0.05)	18.071	22.654
2-1	10.001	0.000 (< 0.05)	7.710	12.292
2-3	2.597	0.017	0.306	4.889
2-4	12.659	0.000 (< 0.05)	10.367	14.950
3-0	17.765	0.000 (< 0.05)	15.473	20.056
3-1	7.404	0.000 (< 0.05)	5.112	9.695
3-2	-2.597	0.017	-4.889	-0.306
3-4	10.061	0.000 (< 0.05)	7.770	12.353
4-0	7.704	0.000 (< 0.05)	5.412	9.995
4-1	-2.658	0.014	-4.949	-0.366
4-2	-12.659	0.000 (< 0.05)	-14.950	-10.367
4-3	-10.061	0.000 (< 0.05)	-12.353	-7.770

Figure 7: Tukey HSD table

Next, we check the ANOVA assumptions. First is to check if residuals are normally distributed. This is checked by creating a QQ plot, shown below.

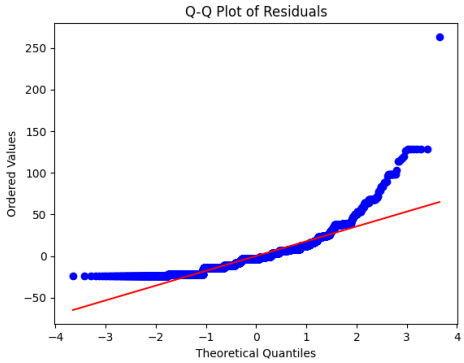

Figure 8: Residual QQ plot

From this plot, we can see the residual doesn't seem to be normally distributed as it shows a U pattern. We verify this finding by doing a Shapiro Wilk test, this shows a P-value of 2.428744671679071e-58, which is much less than 0.001, confirming that the model **is violating the assumption**.

Secondly, we check to see if variances are homogeneous, this is done using Levene's test because the data is not normal. This test yields a P-value of 5.564354307511864e-116, again much less than 0.001, therefore the model is **also violating our second assumption**.

Total Space Available Across Age Group and Availability of Subsidy Contract

In this section, we use two-way ANOVA to answer our second research question:

Question 2: Is there a significant different between total available spaces for different operating auspice and the availability of a subsidy contract?

- $H0_1$: There is no significant difference in the mean of total childcare spaces depending on its subsidy contract with the city.
- $H0_2$: There is no significant difference in the mean of total childcare spaces depending on its age group.
- $H0_3$: There are no interaction effects.

After running the two-way ANOVA using model formula "capacity ~ C(age_group) + C(subsidy) + C(age_group):C(subsidy)", we get the following ANOVA table.

	SUM_SQ	DF	F	PR(>F)
C(AGE_GROUP)	2.821233e+05	4.0	215.341584	1.758733e-171
C(SUBSIDY)	3.215308e+04	1.0	98.168364	6.073586e-23
C(AGE_GROUP):C(SUBSIDY)	2.204016e+05	4.0	168.230101	8.040792e-136

Figure 9: ANOVA Table

Given the values, we have the following conclusions for our hypothesis.

- **We reject $H0_1$** as 1.758733e-171 is significantly less than 0.001, which means there is evidence for a significant difference in the mean of total childcare spaces depending on its subsidy contract with the city.
- **We reject $H0_2$** as 6.073586e-23 is significantly less than 0.001, which means there is evidence for a significant difference in the mean of total childcare spaces depending on its age group.
- **We reject $H0_3$** as 8.040792e-136 is significantly less than 0.001, which means there is evidence for interaction effects.

Drawing the interaction plot confirms our finding, we can see the lines crossing, indicating interaction effect. We can also see significant differences between age groups.

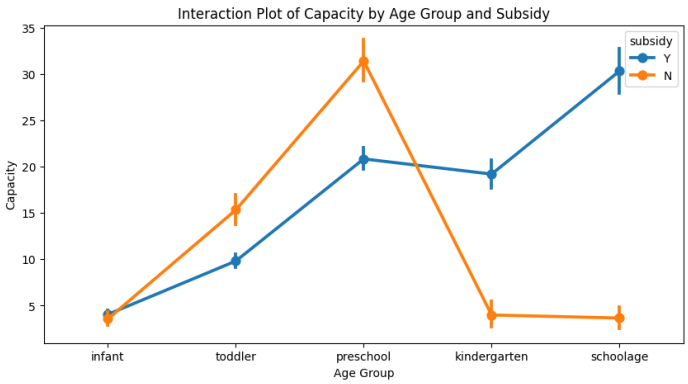

Figure 10: Interaction Plot

Running Tukey HSD test confirms our findings again, following are some sample rows from TukeyHSD test.

COMPARISON	STATISTIC	P-VALUE	LOWER CI	UPPER CI
0-1	0.486	1.000	-3.266	4.239
0-2	-5.755	0.000 (< 0.05)	-8.778	-2.732
0-3	-11.273	0.000 (< 0.05)	-15.026	-7.521
0-4	-16.777	0.000 (< 0.05)	-19.800	-13.754
0-5	-27.337	0.000 (< 0.05)	-31.089	-23.585
0-6	-10.001	0.000 (< 0.05)	-18.161	-12.115
0-7	-15.138	1.000	-3.686	3.818

Figure 11: Tukey HSD table (truncated)

Next, we check our assumptions for ANOVA, first we draw a QQ plot for the residuals, similar to one-way ANOVA, we see U shape in this graph as well.

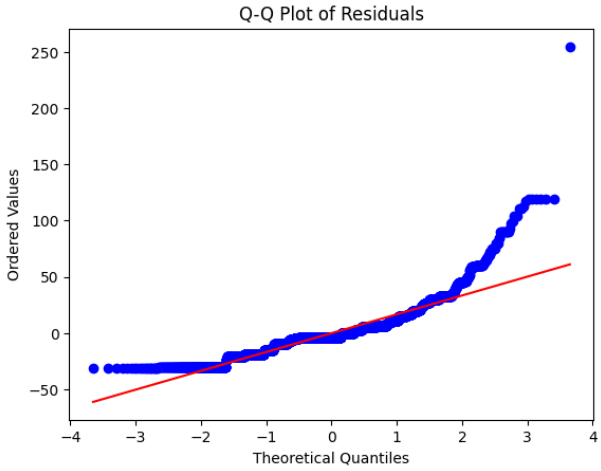

Figure 12: Residual QQ Plot

Doing a Shapiro-Wilk test shows a P-value of 1.4888348267066942e-56, indicating non-normal residuals. Thus, **violating our first assumption**.

Doing a Levene test shows a P-value of 1.02762669672092e-280, indicating non-homogeneous variance, **violating our second assumption.**

Conclusion & Discussion

From our exploration of the dataset, we conclude that it is likely the available childcare spaces are significantly different across various operating auspice. We also found that there is a significant interaction effect between total space available, available of subsidy contract, and age group, suggesting a subsidy contract may have positive influence on total childcare capacity.

However, we are unable to meet our assumptions for ANOVA for both tests, this means our result may not be valid, further analysis and potential transformation of data might be needed to obtain proper results.