Name: Berke Derin Berktay

Exploring The Capacities of Torontonian Child Care Facilities

1. Introduction

Due to various reasons including but not limited to high costs and low availability of spaces, throughout the recent years the increase of difficulty for many Torontonian families to find government funded child care facilities for their children unfortunately became a trend and it continues to be a growing issue amongst the communities in Toronto. This study aims to find out the main reasons and factors that contribute to this trend by observing certain relationships between different traits of the child care centers located in Toronto. The dataset that this study uses contains several information related to each center. Figure 1 below contains all of the variables that the dataset contains and their corresponding meanings.

Column	Description
_id	Unique row identifier for Open Data database
LOC_ID	Unique identifier
LOC_NAME	Name of the child care centre
AUSPICE	Operating auspice (Commercial, Non Profit or Public)
ADDRESS	Address street number, street name, street type, street direction, and unit
PCODE	Address postal code
ward	City ward number
bldg_type	Type of building
BLDGNAME	Name of the building the child care centre is located in
IGSPACE	Child care spaces for infants 0-18 months
TGSPACE	Child care spaces for toddlers 18-30 months
PGSPACE	Child care spaces for preschoolers 30 months up until they enter grade one
KGSPACE	Child care spaces for children in full-day kindergarten
SGSPACE	Child care spaces for children grade one and up
TOTSPACE	Child care spaces for all age groups
subsidy	Centre has a fee subsidy contract (Yes/No)
cwelcc_flag	'Y' indicates space participates in CWELCC, blank indicates it does not'

Figure 1: All of the Columns and Their Descriptions

This study aims to answer the following four research questions using the dataset:
Research Question 1: Is there a significant interaction between the type of auspice and the availability of subsidy on child care spaces?
Research Question 2: How does the auspice of a child care center and participation in the CWELCC together influence the total number of child care spaces?

Research Question 3: Does the presence of a subsidy and participation in the CWELCC program interact to affect the capacity of child care centers in Toronto?

Research Question 4: Does the presence of a subsidy and participation in the CWELCC program interact to affect the capacity of primary school age children care centers in Toronto?

The EDA's and variable counts were considered when determining research questions.

2. Data Cleaning and Data Wrangling

The first thing that we do is to check the null and zero values to see if there are any entries that do not make sense. After checking, we see that the only column that has zero or empty entries is the building name column, which is a useless variable in our case, so we move on. The next step is to perform feature engineering to get rid of useless columns and/or create new ones. We see that the only new column that we might want to create is a total space column that sums all of the row's spaces for each category, but it is already provided to us. Considering our research questions, and the fact that only the metrics that are categorical can be used to observe their relation with the spaces, other than the space metrics, we should only keep the categorical variables that can relate to the research questions that we come up with both our intuition and the exploratory data analysis results. Though we also keep the building type since it might be used later on depending on the amount of categories that it contains. Therefore we delete the many columns to reduce complexity.

3. Exploratory Data Analysis

The first thing that we do is to observe the amount of unique variables that our categorical remaining variables have. The results are seen in the below figure:

```
['Non Profit Agency' 'Commercial Agency' 'Public (City Operated) Agency']
['Public Elementary School' 'Catholic Elementary School' 'Other'
 'High Rise Apartment' 'Community College/University' 'Public High School'
 'House' 'Purpose Built' 'Synagogue' 'Community/Recreation Centre'
 'Public (school closed)' 'Office Building'
 'Multi Human Services Facility' 'Place of Worship' 'Industrial Building'
 'Community/Rec Centre – City' 'Catholic High School' 'Church'
 'Commercial Building' 'Catholic Elementary (French)'
 'Public Elementary (French)' 'Hospital/Health Centre'
 'Low Rise Apartment' 'Public Middle School'
 'Community/Rec Centre – Private' 'Community Rec/Centre – Board Run'
 'Private Elementary School' 'Community/Rec Centre AOCC'
 'Community Health Centre' 'Public Elementary Special']
```

Figure 2: Unique columns of the remaining categorical variables

Name: Berke Derin Berktay

Due to the fact that the substantial amount of unique variables of the building type might make the study very complex, a more detailed research is required, and therefore for this specific research, we exclude it from our analysis. Let us see the graphs of the comparisons of the remaining three categorical variables, which are the cwelcc type, subsidy type, and the auspice type, and their relation with the total space availability and see if there are differences in the amount of availability between certain groups. We aldo craft the main statistical features of the subgroups of the categorical variables. The resulting figures are below:

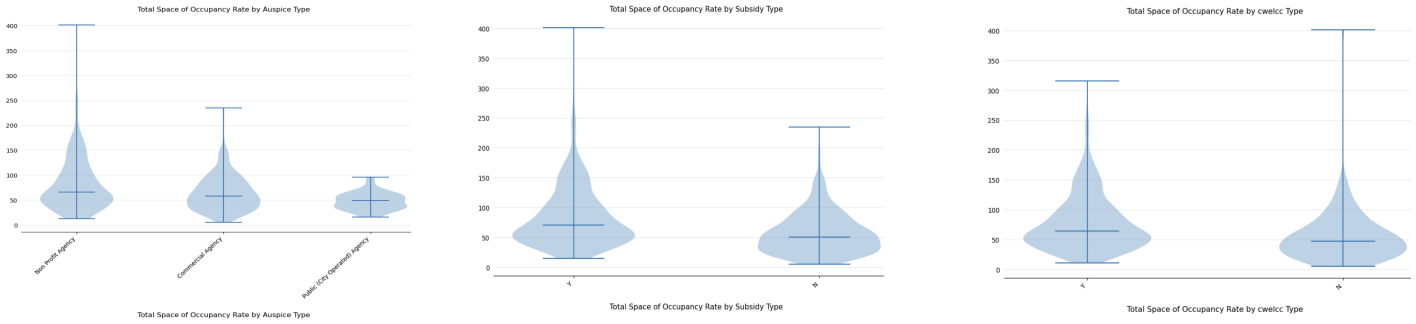


Figure 3: Exploratory Graphs for occupancy vs auspice (left), vs subsidy (left), and vs cwelcc (right) type

```
 Non profit agency summary statistics
Min: 13
Max: 402
Mean: 82.1
25th percentile: 47.0
Median: 66.0
75th percentile: 109.0
IQR : 62.0

Commercial agency summary statistics
Min: 6
Max: 235
Mean: 64.98
25th percentile: 37.0
Median: 58.0
75th percentile: 85.0
IQR : 48.0

Public agency summary statistics
Min: 16
Max: 96
Mean: 47.77
25th percentile: 36.0
Median: 49.0
75th percentile: 62.0
IQR : 26.0
```

```
Subsidized Group summary statistics
Min: 11
Max: 316
Mean: 78.78
25th percentile: 46.0
Median: 65.0
75th percentile: 101.0
IQR : 55.0

Not Subsidized Group summary statistics
Min: 6
Max: 402
Mean: 54.68
25th percentile: 28.0
Median: 48.0
75th percentile: 65.0
IQR : 37.0
```

```
cwelcc group summary statistics
Min: 11
Max: 316
Mean: 78.78
25th percentile: 46.0
Median: 65.0
75th percentile: 101.0
IQR : 55.0

Not cwelcc group summary statistics
Min: 6
Max: 402
Mean: 54.68
25th percentile: 28.0
Median: 48.0
75th percentile: 65.0
IQR : 37.0
```

Figure 4: Summary Statistics for each type of auspice(left), subsidy(middle), and cwelcc (right) category

Both the graphs and the results of the statistical data suggest that for all three variables, there is a significant difference between their inner groupings. More precisely, for the subsidy groupings: The non profit agency centers overall have higher total space than the commercial agency centers and the commercial agency centers have a higher overall total space than the public agencies. The subsidized group overall has a higher total space than the non subsidized group, and the group with cwelcc overall has a higher total space than the group without cwelcc. These facts suggest that a one-way ANOVA test is needed for each variable to see if there is significant difference between the groupings in terms of the total space means. On top of this, it is also a possibility that these variables also effect each other in terms of the outcome of total space for the centers, and therefore further two-way ANOVA tests are needed to see such effects. Hence, we came up with the first three research questions that we would like to answer. On top of this, regarding question 1, understanding if subsidies impact capacity differently across various types of operating auspices can inform where financial support may be most effective. This question probes whether the type of administration (non-profit, commercial, public) combined with financial aid can significantly boost the number of child care spots, guiding resource allocation and support measures. Regarding question 2, this question seeks to determine the combined impact of operational type and a specific policy program (CWELCC) on child care availability. By analyzing these

two factors together, we can assess whether policy programs are equally effective across different operational models or if they benefit one type more than others. Regarding question 3, exploring the interaction between subsidies and CWELCC participation aims to understand if these initiatives have a synergistic effect on increasing capacity. This analysis could highlight the success or limitations of stacking policy interventions and whether they cumulatively lead to a significant increase in available child care spaces. Down the line, we also want to look at question 4, which only considers the children that are grade one and up since the government can consider allocating less space to them down the line since the infants should be the priority.

4. One-Way ANOVA Results

The following standard procedure was used to perform one-way anova:

Step 1: An Ordinary Least Squares regression model is fitted using the variable in question as the predictor for TOTSPACE. An ANOVA table is generated from the model to test whether the mean TOTSPACE is the same across different categories. This test determines if there are statistically significant differences in the total spaces offered by centers with different operating auspices.

Step 2: After finding significant differences with ANOVA, a post-hoc Tukey HSD test is conducted to determine which pairs of the categories are significantly different. It helps to identify which specific groups' means (in terms of TOTSPACE) differ from each other when you have more than two groups.

Step 3:A Q-Q plot and histogram are used to check the normality of residuals. These diagnostics are crucial to validate the assumptions of ANOVA. The Q-Q plot should show points following the line if residuals are normally distributed, and the histogram should look roughly bell-shaped.

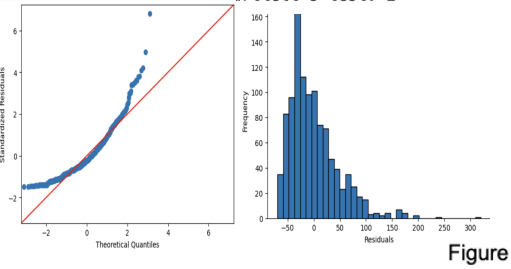
Step 4: We conduct Shapiro-Wilk tests if residuals from the OLS model are normally distributed and Bartlett's test to check if variances are equal across groups when data is normally distributed. We use Levene's test as an alternative for checking homogeneity of variances that is less sensitive to departures from normality. These tests verify the assumptions underlying ANOVA. ANOVA assumes that the residuals are normally distributed and that variances are homogeneous across groups. Violations of assumptions make the ANOVA less credible. The following two figures are the results for each relation:

	df	sum_sq	mean_sq	F	PR(>F)
C(AUSPICE)	2.0	9.611211e+04	48056.057145	21.843051	5.057716e-1
Residual	1060.0	2.332065e+06	2200.061571	NaN	NaN

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	Non Profit Agency	Commercial Agency	17.119417	9.703599	24.535235	7.662434	0.001000
1	Non Profit Agency	Public (City Operated) Agency	34.334610	16.224077	52.445142	6.292710	0.001000
2	Commercial Agency	Public (City Operated) Agency	17.215193	-1.453146	35.883531	3.060857	0.077966

```
#assumption 1: residuals are normally distributed: Shapiro Wilk test
w, pvalue = stats.shapiro(model.resid)
print(w,pvalue)
```

```
0.901775598526001 1.4964898448030214e-25
Bartlett's test statistic: 89.58603867335393, p-value: 3.520779884632816e-20
Degrees of freedom for Bartlett's test: 2
```



	Parameter	Value
0	Test statistics (W)	17.9271
1	Degrees of freedom (Df)	2.0000
2	p value	0.0000

	df	sum_sq	mean_sq	F	PR(>F)
(subsidy)	1.0	1.607654e+05	160765.392950	75.227653	1.550892e-17
Residual	1061.0	2.267412e+06	2137.051825	NaN	NaN

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	Y	N	26.265831	20.323646	32.208016	12.266022	0.001

```
Shapiro-Wilk test statistic: 0.8925793766975403, p-value: 1.4240313682847578e-26
Bartlett's test statistic: 49.081859352317615, p-value: 2.45500187845506e-12
Degrees of freedom for Bartlett's test with subsidy: 1
```

	Parameter	Value
0	Test statistics (W)	22.9889
1	Degrees of freedom (Df)	1.0000
2	p value	0.0000

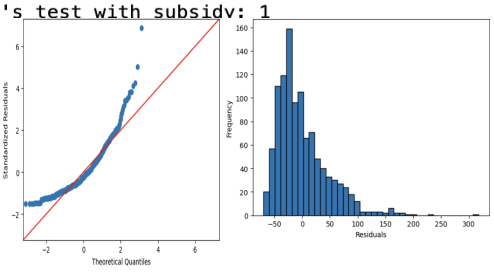


Figure 5: Results for the Space vs Auspice (right) and Space vs Subsidy Type

Name: Berke Derin Berktay

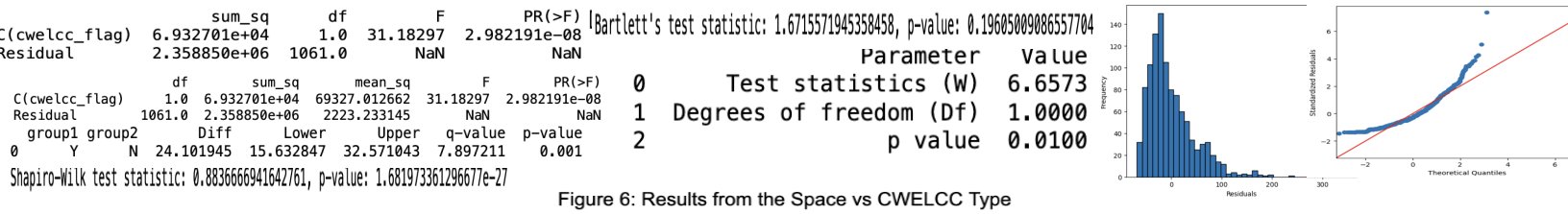


Figure 6: Results from the Space vs CWELCC Type

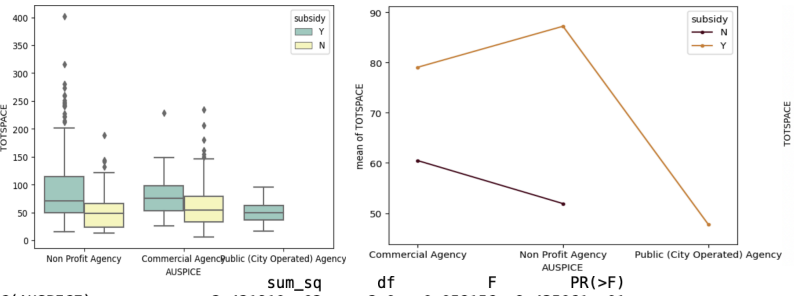
The results for the Space vs
Auspice relationship suggest that the subsidy groups significantly affect TOTSPACE, as indicated by ANOVA with a p-value far below the 0.05 threshold. Tukey's HSD confirms this further, pinpointing the difference between specific groups. However, the residuals from the model do not follow a normal distribution (graph 1 not following the line and graph 2 is skewed) (Shapiro-Wilk test p-value: ~0), and the variances across groups are not homogeneous (Bartlett's and Levene's test p-values: ~0), challenging some of the ANOVA assumptions. Similarly,the results for the Space vs Subsidy relationship suggest that the type of subsidy significantly impacts TOTSPACE, as shown by the ANOVA, which reports a p-value well below the standard threshold of 0.05. This finding is further corroborated by Tukey's HSD, which identifies a significant mean difference between the subsidy groups Y and N. The diagnostics for the model, including the Q-Q plot and histogram, visual tools for assessing the distribution of residuals, suggest deviation from normality, a condition supported by the Shapiro-Wilk test (p-value: ~0) and the same graph observation. Additionally, tests for homogeneity of variances, specifically Bartlett's and Levene's, indicate significant differences between the variances of subsidy groups (p-values: ~0). For the third relation, the results for the Space vs CWELCC Flag relationship suggest that the cwelcc_flag groups have a significant impact on TOTSPACE, as the ANOVA results indicate a p-value significantly less than 0.05. Tukey's HSD shows a marked mean difference between the groups, which substantiates the ANOVA findings. The model diagnostics imply non-normal distribution of residuals, with the Shapiro-Wilk test yielding a p-value effectively at zero. However, Bartlett's test shows a p-value of 0.196, indicating that variances across the cwelcc_flag groups could be considered homogeneous, unlike the Levene's test, which suggests otherwise with a p-value of 0.01. Therefore, while the impact of cwelcc_flag on TOTSPACE is statistically significant, the assumption of equal variances is ambiguous. Despite the significant findings for all the three cases, the model's underlying assumptions are violated, implying a need for cautious interpretation of the results or the consideration of alternative statistical methods

5. Two-Way ANOVA Results

Here is the procedure that we took to conduct the two-way anova:
Step 1: Generate boxplots for space across different categories of categorical variable 1, further split by the categorical variable 2. This helps visualize the distribution, including central tendency and outliers, within and across groups and gives us an intuitive answer regarding the significance and whether or not the effect is the same as both categorical variables change.
Step 2: An OLS regression model is fitted with space as the dependent variable and categorical variable 1, categorical variabl 2, and their interaction as independent variables. This step prepares for ANOVA analysis by modeling the relationships.
Step 3: Using anova_lm, ANOVA is performed on the fitted model to test the significance of categorical variable 1, categorical variable 2, and their interaction on space. The output includes sum of squares, degrees of freedom, F-statistic, and P-values.
Step 4: An interaction plot is created to visualize how the relationship between categorical variable 1 and space varies with categorical variable 2. This highlights potential interaction effects between variables.

Name: Berke Derin Berktay

Step 5:Multiple pairwise comparisons are performed using Tukey's HSD to pinpoint significant differences between group means for categorical variable 1,categorical variable 2, and their combinations. This helps in understanding specific group differences after a significant ANOVA.
Here are the results for all 4 two-way ANOVAs that each correspond to their research questions:



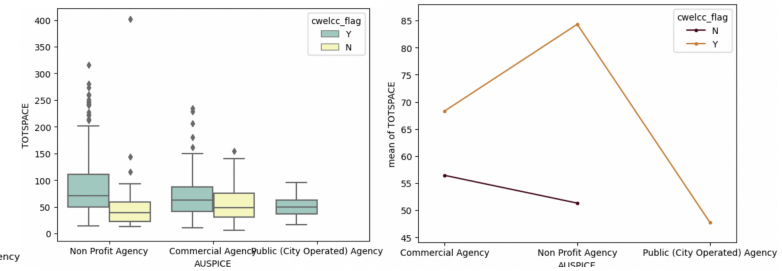
	sum_sq	df	F	PR(>F)
C(AUSPICE)	2.421910e+02	2.0	0.058156	9.435061e-01
C(subsidy)	1.192246e+05	1.0	57.257129	8.291626e-14
C(AUSPICE):C(subsidy)	3.183621e+02	2.0	0.076446	7.822266e-01
Residual	2.203038e+06	1058.0	NaN	NaN

	df	sum_sq	mean_sq	F	PR(>F)
C(AUSPICE)	2.0	2.421910e+02	121.095486	0.058156	9.435061e-01
C(subsidy)	1.0	1.192246e+05	119224.582872	57.257129	8.291626e-14
C(AUSPICE):C(subsidy)	2.0	3.183621e+02	159.181060	0.076446	7.822266e-01
Residual	1058.0	2.203038e+06	2082.266185	NaN	NaN

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	Non Profit Agency	Commercial Agency	17.119417	9.703599	24.535235	7.662434	0.001000
1	Non Profit Agency	Public (City Operated) Agency	34.334610	16.224077	52.445142	6.292710	0.001000
2	Commercial Agency	Public (City Operated) Agency	17.215193	-1.453146	35.883531	3.060857	0.077966

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	Y	N	26.265831	20.323646	32.208016	12.266022	0.001

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	(Non Profit Agency, Y)	(Non Profit Agency, N)	35.327657	21.376765	49.278548	10.224012	0.001000
1	(Non Profit Agency, Y)	(Commercial Agency, Y)	8.165515	-7.513254	23.844284	2.102713	0.650319
2	(Non Profit Agency, Y)	(Commercial Agency, N)	26.764597	16.861011	36.668182	10.911306	0.001000
3	(Non Profit Agency, Y)	(Public (City Operated) Agency, Y)	39.460387	17.933459	60.987315	7.400955	0.001000
4	(Non Profit Agency, Y)	(Public (City Operated) Agency, N)	0.000000	-inf	inf	0.000000	0.900000



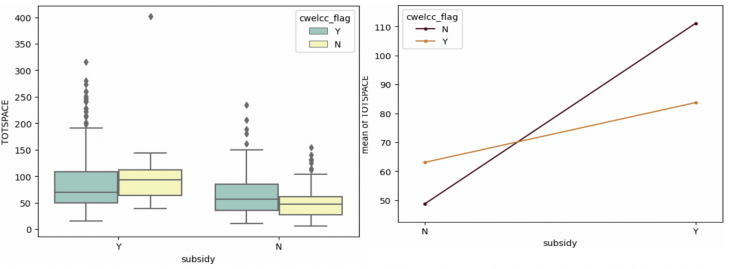
	sum_sq	df	F	PR(>F)
C(AUSPICE)	7.201242e+04	2.0	16.673931	7.417898e-08
C(cwelcc_flag)	4.522732e+04	1.0	20.944088	5.288385e-06
Residual	2.286838e+06	1059.0	NaN	NaN

	df	sum_sq	mean_sq	F	PR(>F)
C(AUSPICE)	2.0	7.201242e+04	36006.212191	16.673931	7.417898e-08
C(cwelcc_flag)	1.0	4.522732e+04	45227.322754	20.944088	5.288385e-06
Residual	1059.0	2.286838e+06	2159.431484	NaN	NaN

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	Non Profit Agency	Commercial Agency	17.119417	9.703599	24.535235	7.662434	0.001000
1	Non Profit Agency	Public (City Operated) Agency	34.334610	16.224077	52.445142	6.292710	0.001000
2	Commercial Agency	Public (City Operated) Agency	17.215193	-1.453146	35.883531	3.060857	0.077966

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	Y	N	24.101945	15.632847	32.571043	7.897211	0.001

Figure 7: Two- Way Anova results for the space in terms of agency and subsidy types (left) and space in terms of agency and cwelcc type (right)



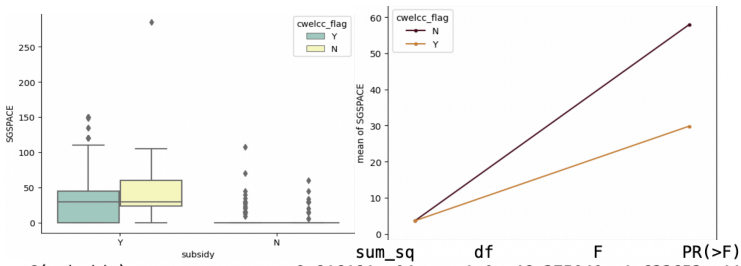
	sum_sq	df	F	PR(>F)
C(subsidy)	9.816181e+04	1.0	46.375040	1.633653e-11
C(cwelcc_flag)	6.723429e+03	1.0	3.176381	7.499648e-02
C(subsidy):C(cwelcc_flag)	1.910868e+04	1.0	9.027604	2.721895e-03
Residual	2.241580e+06	1059.0	NaN	NaN

	df	sum_sq	mean_sq	F
C(subsidy)	1.0	9.816181e+04	98161.809243	46.375040
C(cwelcc_flag)	1.0	6.723429e+03	6723.428955	3.176381
C(subsidy):C(cwelcc_flag)	1.0	1.910868e+04	19108.682949	9.027604
Residual	1059.0	2.241580e+06	2116.694877	NaN

	PR(>F)
C(subsidy)	1.633653e-11
C(cwelcc_flag)	7.499648e-02
C(subsidy):C(cwelcc_flag)	2.721895e-03
Residual	NaN

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	Y	N	26.265831	20.323646	32.208016	12.266022	0.001

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	Y	N	24.101945	15.632847	32.571043	7.897211	0.001



	sum_sq	df	F	PR(>F)
C(subsidy)	9.816181e+04	1.0	46.375040	1.633653e-11
C(cwelcc_flag)	6.723429e+03	1.0	3.176381	7.499648e-02
C(subsidy):C(cwelcc_flag)	1.910868e+04	1.0	9.027604	2.721895e-03
Residual	2.241580e+06	1059.0	NaN	NaN

	df	sum_sq	mean_sq	F
C(subsidy)	1.0	141017.145514	141017.145514	184.976568
C(cwelcc_flag)	1.0	1408.473768	1408.473768	1.847539
C(subsidy):C(cwelcc_flag)	1.0	8685.150906	8685.150906	11.392582
Residual	1059.0	807330.132897	762.351400	NaN

	PR(>F)
C(subsidy)	6.029422e-39
C(cwelcc_flag)	1.743593e-01
C(subsidy):C(cwelcc_flag)	7.641314e-04
Residual	NaN

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	Y	N	26.654007	23.086173	30.22184	20.730831	0.001

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	Y	N	14.744037	9.349548	20.138526	7.584464	0.001

Figure 8: Two- Way Anova results for the space in terms of cwelcc and subsidy types (left) and grade one or above space in terms of cwelcc and subsidy type (right)

Name: Berke Derin Berktay

The boxplot visualizes total space (TOTSPACE) variations across agency types (AUSPICE) and subsidy statuses, the interclass differences can already be seen. ANOVA results highlight subsidy status as a significant factor affecting total space ($p < 0.05$), unlike agency type or their interaction, which show no significant effect. Tukey's HSD tests further dissect these relationships, revealing significant differences in total space between Non-Profit and both Commercial and Public Agencies. However, no significant difference is observed between Commercial and Public Agencies. Subsidy status, when analyzed separately, shows a clear impact on total space, with significant differences between subsidized and non-subsidized agencies, as confirmed by Tukey's HSD test. Lastly, a complex Tukey's HSD analysis combining agency type and subsidy status identifies significant differences in certain group pairs, but also uncovers computational issues in others, indicating potential data or assumption-related problems. On the second case the ANOVA analysis shows that both agency type (AUSPICE) and the CWELCC flag (cwelcc_flag) significantly affect total space, with p-values indicating strong statistical significance for both factors. Tukey's HSD tests for agency types demonstrate significant differences in total space between Non-Profit Agencies and both Commercial and Public Agencies. The difference between Commercial and Public Agencies is not significant, as the p-value is above the 0.05 threshold. When analyzing the CWELCC flag with Tukey's HSD, there is a significant difference in total space between facilities with the CWELCC flag and those without, reinforcing the ANOVA findings. The interaction plot shows these relationships, showing how total space varies by agency type within each CWELCC category. Overall, both the type of agency and the CWELCC flag significantly impact total space, with clear differences identified between specific groups in the Tukey's tests. In the third case, ANOVA shows that subsidy status has a significant impact on total space ($p < 0.05$), while the CWELCC flag's effect is marginal ($p = 0.075$). The interaction between subsidy status and the CWELCC also significantly affects total space, indicating that the impact of subsidy status might vary depending on the CWELCC flag. Tukey's HSD tests for subsidy status confirm a significant difference in total space between subsidized and non-subsidized centres, highlighting the substantial impact of subsidies on space. Similarly, Tukey's HSD analysis for the CWELCC flag shows a significant difference in total space between facilities with and without the CWELCC, despite the marginal p-value in the ANOVA, suggesting that the CWELCC flag still influences total space. The interaction plot visualizes how total space varies with subsidy status within each category of the CWELCC flag and backs up our claims. Lastly, in the fourth case, focusing on school-age space (SGSPACE), the ANOVA analysis reveals that subsidy status has a highly significant impact ($p < 0.00001$), while the effect of the CWELCC flag is not significant ($p > 0.05$). However, the interaction between subsidy status and the CWELCC flag is significant, indicating that the influence of subsidy on SGSPACE varies depending on the CWELCC flag status. Tukey's HSD tests for subsidy status show a significant difference in SGSPACE between subsidized and non-subsidized facilities, emphasizing the strong effect of subsidies. The Tukey's HSD test for the CWELCC flag, despite the ANOVA's non-significant result for the CWELCC flag alone, indicates a significant difference in SGSPACE between facilities with and without the CWELCC flag, suggesting a practical impact that wasn't highlighted by the ANOVA. The interaction plot demonstrates how SGSPACE varies with subsidy status across CWELCC flag categories.

6. Conclusion

The study on Torontonian child care facilities using two-way ANOVAs revealed that subsidies significantly increase child care spaces, particularly in non-profit centers, and their impact is further amplified when combined with CWELCC program participation. While the type of operating auspice and CWELCC participation independently contribute to capacity expansion, their interaction shows no significant variance in effectiveness across different auspices. Notably, for primary school-aged care, the synergy between subsidies and CWELCC participation markedly enhances space availability. These findings underscore the critical role of subsidies and highlight the potential of targeted policy interventions, such as the CWELCC program, to effectively increase child care capacities in Toronto.