1. Introduction

In the province of Ontario, finding qualified or unqualified childcare services can be challenging due to high fees and limited childcare spaces. According to the Toronto Children's Services department, 75% of families cannot afford childcare, prompting the provincial government to pledge the addition of 100,000 new childcare spaces from 2016 to 2026. This report aims to delve into the situation of licensed childcare centers in Toronto. Utilizing a dataset named "INF2178 A2 data.xls," which was updated in February 2024, this analysis seeks to provide insights into the operation and capacity of these centers across various age groups.

Our exploration will address two basic research questions to gain insight into the situation of qualified childcare centers in Toronto:

1. research question 1：To explore whether the effect of different child care age stages on space size significant
2. research question 2: Main Effect Hypothesis 1: Whether there is a significant difference in space sizes across different AUSPICE categories
3. research question 3: Main Effect Hypothesis 2: Whether there is a significant difference in space sizes across different age stages
4. research question 4: Interaction Effect Hypothesis: Whether there is a significant interaction effect between AUSPICE and age_stage_treatments

2. Data Cleaning and Data Wrangling

This original dataset has 17 columns and 1,063 entities (rows). To better analyze the research, we have adjusted the original dataset:

2.1 Since our analysis is quantitative in nature, I have considered deleting unimportant columns and keeping the following to make the data clearer and easier to understand: “AUSPICE”, “IGSPACE”,“TGSPACE”, “PGSPACE”, “KGSPACE”,“SGSPACE”, “TOTSPACE”, “subsidy”, “cwelcc_flag”

2.2 I have conducted a preliminary exploration of the reduced data. based statistical summary:

- The dataset has a total of 1063 entries with 10 features.
- Seven features are recognized as integer types (int64) and four features are recognized as object types.
- There are no significant missing values as each feature has 1063 non-null values.
-

3. Data Engineer

In order to better explore whether the effect of different age stages on the size of the space is significant, we created new continuous variables using all age stages, there are a total of five age stages: 'IGSPACE_proportion', 'TGSPACE_proportion', 'PGSPACE_proportion', 'KGSPACE_proportion', and 'SGSPACE_proportion' as our categorical variables for one-way ANOVA (treatments), and their corresponding values representing the size of the space are our continuous variables.

4. ANOVA

4.1 EDA

The boxplot reflects (Figure 1) varying degrees of variability in the distribution of space sizes across different age stages. Additionally, there are multiple outliers in each age stage, suggesting that some childcare centers have unusually high space sizes for these age groups. This could be attributed to special service requirements, funding sources, or operational models, representing a societal response; therefore, we consider retaining these outliers .

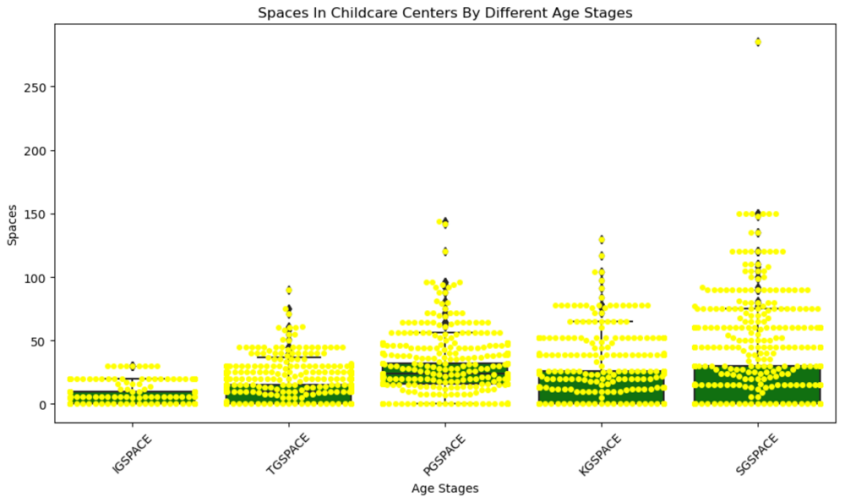

Figure 1

4.2 One-way

- H0: There is no significant difference in space size among different childcare age stages.
- H1: At least one age stage exhibits statistically significant differences in space size.

	Df	Sum_sq	Mean_sq	F	PR(>F)
C(age_stage_treatments)	4.0	2.821233e+05	70530.816839	188.190768	4.517383e-151
Residual	5310.0	1.990101e+06	374.783617	NaN	NaN

Table 1:One ANOVA

The ANOVA results(Table1) show 4 degrees of freedom as 5 categories in age_stages with an F-statistic of 188.190768 and a corresponding p-value: 4.517383e-151<0.05, which suggests that we have sufficient evidence to reject the null hypothesis that there is no significant difference in space size among different child care age stages.

4.2.1 Post Hoc Test

Next, we got a better understanding of the relationship between these categories through the post hoc test:

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	IGSPACE	TGSPACE	7.703669	5.412308	9.995029	12.974001	0.001000
1	IGSPACE	PGSPACE	20.362183	18.070822	22.653543	34.292619	0.001000
2	IGSPACE	KGSPACE	10.361242	8.069881	12.652602	17.449707	0.001000
3	IGSPACE	SGSPACE	17.764817	15.473456	20.056177	29.918310	0.001000
4	TGSPACE	PGSPACE	12.658514	10.367153	14.949874	21.318618	0.001000
5	TGSPACE	KGSPACE	2.657573	0.366213	4.948933	4.475706	0.013527
6	TGSPACE	SGSPACE	10.061148	7.769787	12.352508	16.944309	0.001000
7	PGSPACE	KGSPACE	10.000941	7.709580	12.292301	16.842912	0.001000
8	PGSPACE	SGSPACE	2.597366	0.306006	4.888726	4.374309	0.017028
9	KGSPACE	SGSPACE	7.403575	5.112214	9.694935	12.468603	0.001000

Table 2: Post Hoc Test

Post Hoc Test (Table 2) shows that differences in the space size among all age stages were significant (p-value < 0.05) and were weaker for TGSPACE and KGSPACE; PGSPACE and SGSPACE compared to the other stages. The result implies that almost every age stage is significantly different from at least one other age stage in terms of the number of childcare spaces at different ages.

4.2.2 Residual Analysis

I will perform a residual analysis by QQ plot (Figure2), Histogram (Figure3) and Shapiro-Wilk test to check that the assumptions of the ANOVA model are met.

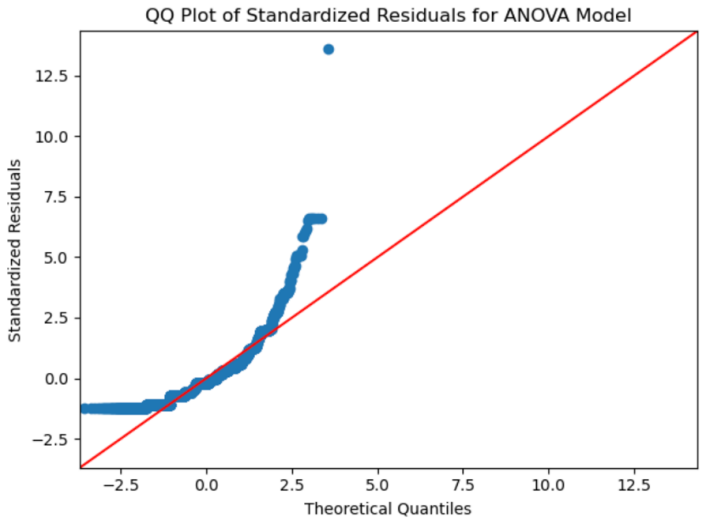

Figure 2

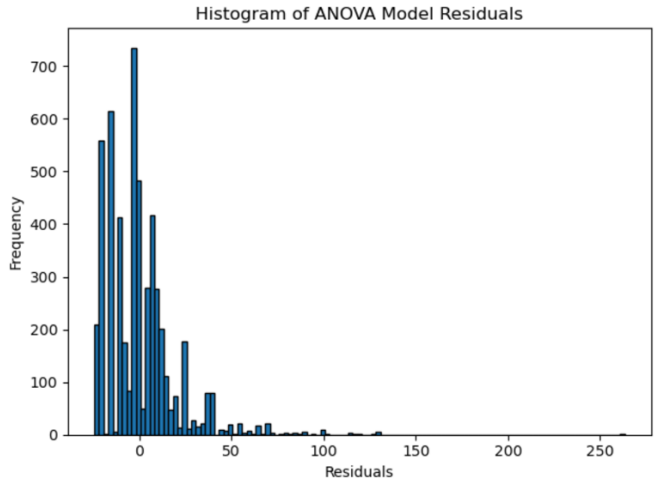

Figure 3

The Quantile-Quantile Plot(Figure2) shows that most of the data points follow this line in the center portion, but there are deviations at the ends, especially in the upper right corner of the graph. This indicates that there are some extreme values of the data away from the center that are much larger than would be expected from a normal distribution.

The histogram (Figure3) shows that the distribution of the residuals of the data is skewed, which also suggests that the assumption of normality of the data may have been violated.

Assumption1: residuals are normally distributed: Shapiro Wilk test:

H0:residuals are normally distributed; H1:residuals are not normally distributed

Shapiro-Wilk test statistic	p-value
0.8427016139030457	<0.001

Table3: Shapiro Wilk test

Shapiro Wilk test (Table3) shows that p-value<0.001, implying that we have enough evidence to reject the null hypothesis that the residuals are normally distributed.

Since our data do not currently conform to a normal distribution, we performed homogeneity test of variance using the Levene test, which is less dependent on the normal distribution of the data.

4.2.3 Variances are Homogenous

Assumption2: variances are homogenous: Levene's test when the sample is not normally distributed:

H0:the variances are consistent across groups;H1:the variances are inconsistent across groups.

	Parameter	Value
0	Test statistics (W)	142.6228
1	Degrees of freedom (Df)	4.0
2	p value	<0.001

Table 4: Variances are Homogenous

Variances are Homogenous (Table 4) shows that the p-value is less than 0.05, implying that we have enough evidence to reject the null hypothesis that the variances are consistent across groups.

4.3 Two-way ANOVA

We add another new categorical variables for the Two-way ANOVA component: AUSPICE (Commercial, Non Profit, Public)

4.3.1 EDA

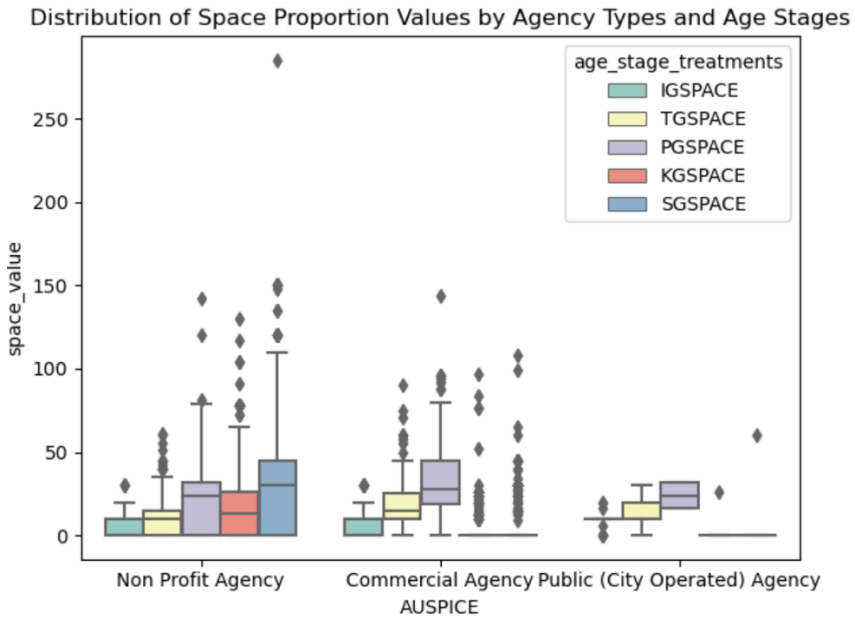

Figure 4

The boxplot (Figure 4) shows that the differences in the distribution of space size values among different age stage treatment groups within various AUSPICE categories may indicate

the presence of potential interaction effects. For example, the differences between age stages were greater within nonprofit organizations.

4.3.2 Two-Way ANOVA

- Main Effect Hypothesis 1:H0: There is no significant difference in space sizes between different AUSPICE categories; H1: There is a significant difference in space sizes for at least one of AUSPICE categories.
- Main Effect Hypothesis 2: H0: There is no significant difference in space sizes between different age stages; H1: There is a significant difference in space sizes for at least one of the ages.
- Interaction Effect Hypothesis: H0: There is no significant interaction effect on space sizes between AUSPICE and age_stage_treatments; H1: There is a significant interaction effect on space sizes between AUSPICE and age_stage_treatments.

	sum_sq	df	F	PR(>F
C(AUSPICE)	1.922242e+04	2.0	29.104369	2.685589e-13
C(age_stage_treatments)	2.821233e+05	4.0	213.579208	3.726614e-170
C(AUSPICE):C(age_stage_treatments)	2.206458e+05	8.0	83.519093	9.972931e-131
Residual	1.750233e+06	5300.0	NaN	NaN

Table 5: Two-way ANOVA

- C(AUSPICE): p-value<0.05, indicating that we have sufficient evidence to reject the null hypothesis that there is no significant difference in space sizes between different AUSPICE categories..
- C(age_stage_treatments): p-value<0.05, indicating that we have sufficient evidence to reject the null hypothesis that there is no significant difference in space sizes between different age stages.
- C(AUSPICE):C(age_stage_treatments) indicates the interaction effect between AUSPICE and age stages. The p-value <0.05 suggests that there is an interaction between the two categorical variables, meaning that the effect of space size in different age stages within various AUSPICE categories is interdependent.

4.3.3 Interaction plot

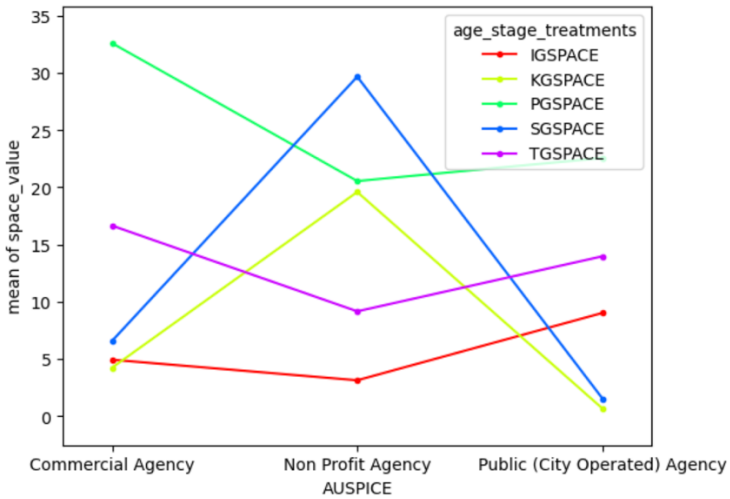

Figure 5

This interaction plot (Figure 5) shows that different AUSPICE categories behave significantly differently in space across age stages. For example, for TGSPACE, nonprofit organizations have a smaller average space size than commercial organizations, while public organizations show a relatively larger value.

4.3.4 Tukey HSD

group1	group2	Diff	Lower	Upper	q-value	p-value
Non Profit Agency	Commercial Agency	3.423883	2.140456	4.707311	8.844824	0.001000
Non Profit Agency	Public (City Operated) Agency	6.866922	3.732600	10.001244	7.263738	0.001000
Commercial Agency	Public (City Operated) Agency	3.443039	0.212180	6.673898	3.533178	0.033472

Table 6: Tukey HSD- AUSPICE categories

Tukey HSD- AUSPICE categories (Table 6) shows that the results of the post-hoc test showed all p-values < 0.05 for comparisons between different AUSPICE categories, suggesting that AUSPICE type is an important factor influencing the size of the space, further supporting that we have sufficient evidence to reject the null hypothesis.

group1	group2	Diff	Lower	Upper	q-value	p-value
IGSPACE	TGSPACE	7.703669	5.552803	9.854535	13.821471	0.001000
IGSPACE	PGSPACE	20.362183	18.211316	22.513049	36.532634	0.001000
IGSPACE	KGSPACE	10.361242	8.210376	12.512108	18.589533	0.001000
IGSPACE	SGSPACE	17.764817	15.613951	19.915683	31.872592	0.001000
TGSPACE	PGSPACE	12.658514	10.507648	14.809380	22.711163	0.001000
TGSPACE	KGSPACE	2.657573	0.506707	4.808439	4.768062	0.006750
TGSPACE	SGSPACE	10.061148	7.910282	12.212014	18.051121	0.001000
PGSPACE	KGSPACE	10.000941	7.850075	12.151807	17.943102	0.001000
PGSPACE	SGSPACE	2.597366	0.446500	4.748232	4.660042	0.008773

Table 7: Tukey HSD- age stages

Tukey HSD- age stages (Table 7) shows that The results of the post hoc test showed all p-values < 0.05 for comparisons between different age stages, suggesting that age stage is an important factor influencing the size of the space, further supporting that we have sufficient evidence to reject the null hypothesis that there is no significant difference in space sizes between different age stages

group1	group2	Diff	Lower	Upper	q-value	p-value
(Non Profit Agency, IGSPACE)	(Non Profit Agency, TGSPACE)	6.024182	2.735427	9.312937	8.789533	0.001
(Non Profit Agency, IGSPACE)	(Non Profit Agency, PGSPACE)	17.408250	14.119495	20.697006	25.399362	0.001
(Non Profit Agency, IGSPACE)	(Non Profit Agency, KGSPACE)	16.452347	13.163592	19.741102	24.004660	0.001
(Non Profit Agency, IGSPACE)	(Non Profit Agency, SGSPACE)	26.529161	23.240405	29.817916	38.707151	0.001
(Non Profit Agency, IGSPACE)	(Commercial Agency, IGSPACE)	1.796599	-2.356899	5.950098	2.075566	0.900

Table 8: Tukey HSD

Tukey HSD (Table 8) shows tahtThe results show significant differences in space dimensions at different ages in different types of organizations, but there are still cases where the results of the effect of the type of AUSPICE may be inconsistent at different ages, such as the space of IGSPACE in nonprofit agency with the space of IGSPACE in commercial agency-value= 0.900 > 0.05, which means that we have evidence in support of the null hypothesis.