Student: Yixin Chang (1005991651)
INF2178 - Assignment 2

Analyzing the Proportion of Child Care Spaces for Preschoolers in Toronto

1. Introduction

In recent years, the availability of childcare spaces has become a critical concern for parents, policymakers, and researchers. The dataset "INF2178_A2_data.xlsx" collects information of all Licensed Child Care Centers in the City of Toronto. This report focuses specifically on childcare spaces for preschoolers aged 30 months up until they enter grade one. It's an important period of growth and discovery, where every day brings new learnings, friendships, and adventures.
Our exploration will address two research questions:

1. **Research Question 1:** How does the proportion of childcare spaces for preschoolers (aged 30 months up until they enter grade one) among the total childcare spaces in childcare centers vary by different operation auspice? Are there any significant differences among commercial agency, nonprofit agency or public agency?
2. **Research Question 2:** How does the proportion of childcare spaces for preschoolers (aged 30 months up until they enter grade one) among the total childcare spaces in childcare centers vary by subsidy (whether the center has a fee subsidy contract) and CWELCC flag (whether the space participates in CWELCC)?

2. Data Cleaning and Data Wrangling

We create and add a new feature to our dataset for further analysis:
PG_RATE: this feature refers to the proportion of childcare spaces for preschoolers (aged 30 months up until they enter grade one) among the total childcare spaces in childcare centers.

$$PG\_RATE = \frac{PGSPACE}{TOTSPACE}$$

- PGSPACE: Childcare spaces for preschoolers 30 months up until they enter grade one
- TOTSPACE: Childcare spaces for all age groups

We also reduce our dataset to the following columns:

- AUSPICE: operating auspice (Commercial, Non Profit or Public)
- subsidy: whether the center has a fee subsidy contract (Yes/No: Y/N)
- cwelcc_flag: whether the space participates in CWELCC (Yes/No: Y/N)
- PG_RATE: the result of PGSPACE/TOTSPACE, indicating the proportion of childcare spaces for preschoolers (aged 30 months up until they enter grade one) among the total childcare spaces in childcare centers

The current dataset has 1063 rows and 4 columns. After critically reviewing the data, we found that there are no missing values (NaN) in the current dataset. Then, we can work on our current dataset for further analysis.

3. One-way ANOVA

Research Question 1: How does the proportion of childcare spaces for preschoolers (aged 30 months up until they enter grade one) among the total childcare spaces in childcare centers

(PG_RATE) vary by different operation auspice? Are there any significant differences among commercial agency, nonprofit agency or public agency?

To explore this research question, we use boxplots and one-way ANOVA. Firstly, the following boxplot (Figure 1) visualizes the distribution of the PG_RATE variable, which represents the ratio of PGSPACE over TOTSPACE, for different categories of operating auspices: Non-Profit Agency, Commercial Agency, and Public (City) Operated Agency.

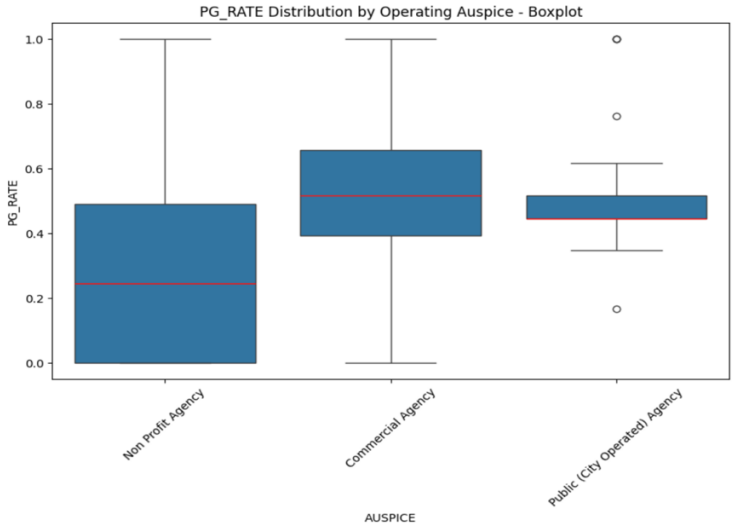

Figure 1: Boxplot of PG_RATE by AUSPICE

Based on Figure 1, where the red line denotes the median, Commercial Agency with the highest median of PG_RATE allocates a higher proportion of their total space to PGSPACE. Furthermore, Public (City operated) Agency shows a more consistent PG_RATE across their data points, as reflected by the narrowest IQR. The data points of Non-Profit Agency are mostly clustered around the median, which indicates a relative consistency in the proportion of PGSPACE within this category. These observations of the PG_RATE distribution show a clear distinction in how PGSPAC' is assigned across different agency types.

To further analyze our Research Question 1, we use the one-way ANOVA. Figure 2 is the ANOVA table that shows the results. The null hypothesis (H_0) is that there are no differences in the means of PG_RATE among different AUSPICE categories (Non-Profit, Commercial, Public). The alternative hypothesis (H_A) is that at least one AUSPICE category has a different mean of PG_RATE compared to the others.

	DF	Sum of Squares	Mean Square	F-ratio	P-value
C(AUSPICE)	2.0	12.759	6.379	89.386	< 0.001
Residual	1060.0	75.651	0.071		

Figure 2: One-way ANOVA table that analyzes the effect of AUSPICE on PG_RATE

Since the p-value < 0.001 < 0.05 (the significant level $\alpha = 0.05$), we reject the null hypothesis. This result suggests that there is a statistically significant difference in the means of PG_RATE among different AUSPICE categories. However, it's important to note that while ANOVA indicates a significant difference, it does specify where that difference lies. This is where post-hoc testing can help identify the specific group differences. Therefore, we use post-hoc tests, such as Tukey's HSD, to further explore which specific AUSPICE categories differ from each other, and the test results are shown in Figure 3.

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	Non Profit Agency	Commercial Agency	0.234	0.192	0.277	18.427	< 0.001

1	Non Profit Agency	Public (City Operated) Agency	0.204	0.101	0.307	6.572	< 0.001
2	Commercial Agency	Public (City Operated) Agency	0.030	-0.076	0.137	0.944	0.762

Figure 3: Post-hoc test results using Tukey's HSD

In these pairwise comparisons, Non Profit Agency has a significantly lower mean PG_RATE than both Commercial Agency and Public Agency (p-value < 0.001 < 0.05). Nevertheless, the difference between the mean PG_RATE of Commercial Agency and Public Agency is not statistically significant (p-value = 0.762 > 0.05). This suggests that both Commercial and Public Agencies allocate a larger portion of their space to preschool-aged children than Non-Profit Agencies.

Checking assumptions is an important step when performing one-way ANOVA. Firstly, QQ-plot from standardized residuals is used in Figure 4 for normality assumption.

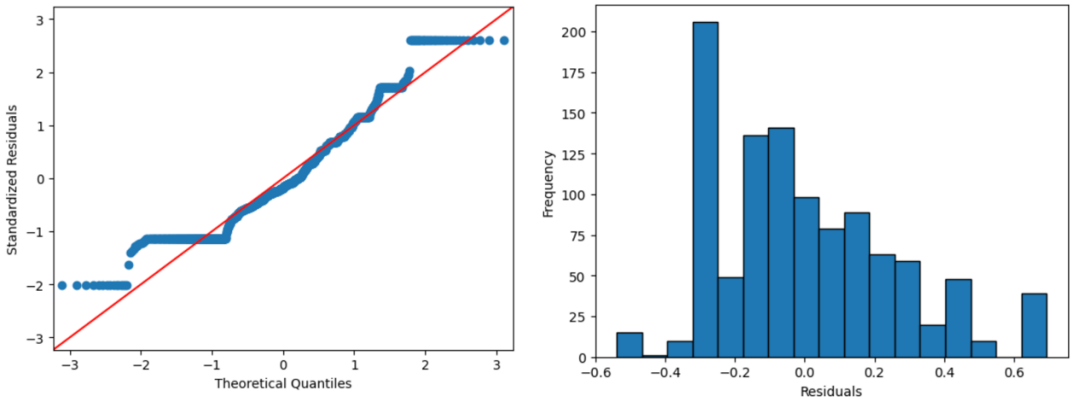

Figure 4: QQ-plot and Histogram from standardized residuals

In the QQ-plot, if the residuals are normally distributed, they should roughly follow the red reference line. Based on the above figure, we can find that the residuals may not be perfectly normally distributed. Furthermore, the histogram is right skewed. This again indicates a potential violation of the normality assumption of the residuals.
To further check whether residuals are normally distributed, we conduct Shapiro-Wilk test, and the results are displayed in Figure 5.

Test Statistics: W	**p-value**
0.947	< 0.001

Figure 5: Results of Shapiro-Wilk test

Since the p-value < 0.001 < 0.05, we reject the null hypothesis and conclude that the residuals are not normally distributed.
Because the data is not drawn from normal distribution, Levene's test is used to check the second assumption, the Homogeneity of variances. Figure 6 shows the results of this test.

	Parameter	**Value**
0	Test statistics (W)	14.986
1	Degree of freedom (Df)	2
2	p value	< 0.001

Figure 6: Results of Levene's test

As the p value (< 0.05) is significant, we reject null hypothesis and conclude that the variances in PG_RATE are not equal across the different AUSPICE groups. The assumption of homogeneity of variances has been violated.

We find that both assumptions – normality of data and homogeneity of variances are violated. This may affect the validity of the ANOVA results and lead to incorrect results. Another point is that one-way ANOVA may not be the most appropriate test, so alternative statistical methods can be considered. However, our sample size is 1063, which is large. The results of ANOVA might still provide some useful information.

4. Two-way ANOVA

Research Question 2: How does the proportion of childcare spaces for preschoolers (aged 30 months up until they enter grade one) among the total childcare spaces in childcare centers (PG_RATE) vary by subsidy (whether the center has a fee subsidy contract) and CWELCC flag (whether the space participates in CWELCC)?

To explore this research question, we use boxplots and two-way ANOVA. Firstly, the following boxplot (Figure 7) visualizes the distribution of PG_RATE based on two factors: the presence of a fee subsidy (subsidy) and participation in the CWELCC program (cwelcc_flag). In this boxplot, 'Y' indicates 'Yes' (the presence of a subsidy or CWELCC participation), while 'N' indicates 'No'.

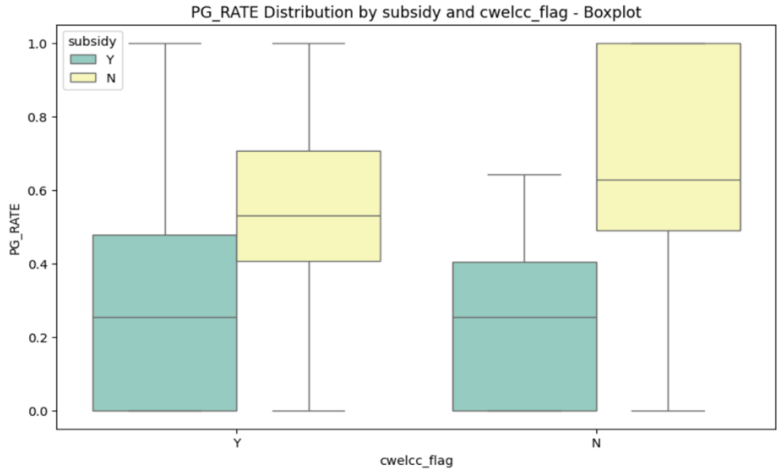

Figure 7: Boxplot of PG_RATE by subsidy and cwelcc_flag

Based on the above graph, the median of PG_RATE appears similar regardless of CWELCC participation for centers with a subsidy. However, for centers without a subsidy, the median of PG_RATE seems to be higher for those not participating in CWELCC. In both types of center participation in CWELCC, the median of PG_RATE for centers without fee subsidy contracts is higher than those of centers with fee subsidy contracts. All types spread widely, and PG_RATE for centers without subsidy contracts and participating in CWELCC has the narrowest IQR.

To further analyze the Research Question 2, we use the two-way ANOVA. From two-way ANOVA, we can test three hypotheses: The effect of 1) subsidy status, 2) CWELCC participation, and 3) the interaction between subsidy status and CWELCC participation on the proportion of childcare spaces for preschoolers (PG_RATE). The two-way ANOVA table is displayed in Figure 8:

	DF	Sum of Squares	Mean Square	F-ratio	P-value
C(cwelcc_flag)	1.0	0.294	0.294	4.727	0.030
C(subsidy)	1.0	15.056	15.056	242.443	< 0.001
C(cwelcc_flag):C(subsidy)	1.0	0.180	0.180	2.906	0.089
Residual	1059.0	65.765	0.062		

Figure 8: Two-way ANOVA table examining the impact of two independent variables - subsidy status (subsidy) and CWELCC participation (cwelcc_flag) - on the dependent variable PG_RATE.

The p value obtained for subsidy and cwelcc_flag is statistically significant with $p < 0.05$. We conclude that whether the center has a fee subsidy contract significantly affects PG_RATE, and whether centers participate in CWELCC significantly affects PG_RATE. However, interaction of both subsidy status and CWELCC participation does not significantly affect PG_RATE ($p = 0.089 > 0.05$). We also use the interactive plot to visualize the interaction effects in Figure 9.

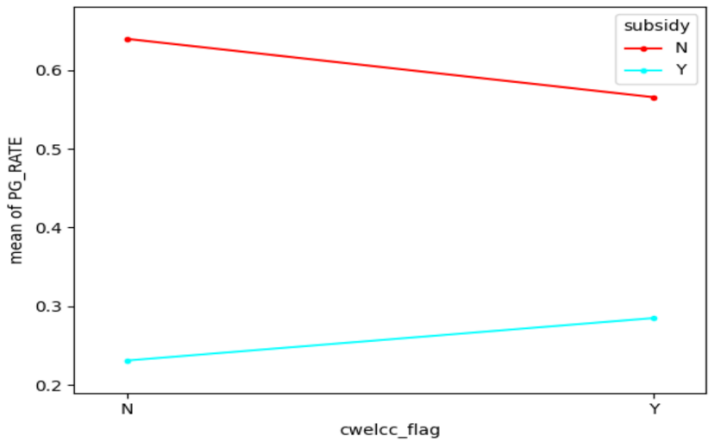

Figure 9: The interactive plot shows the relationship between CWELCC participation (cwelcc_flag), subsidy status (subsidy), and the mean of PG_RATE.

From the above graph, the interaction effect seems to be significant because two lines are not parallel and do not cross each other. This is in contrast of the results obtained in two-way ANOVA table. Therefore, we will do post-hoc tests like Tukey's HSD to see details, and check assumptions of two-way ANOVA to validate the reliability of its results. The following Figures 10, 11, 12 show the results of Tukey's HSD tests.

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	Y	N	0.249	0.204	0.294	15.447	< 0.001

Figure 10: Tukey's HSD for CWELCC Flag 'cwelcc_flag'

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	Y	N	0.308	0.276	0.340	26.722	< 0.001

Figure 11: Tukey's HSD for Subsidy status 'subsidy'

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	(Y, Y)	(Y, N)	0.281	0.231	0.330	20.672	< 0.001
1	(Y, Y)	(N, Y)	0.054	-0.126	0.233	1.094	0.851
2	(Y, Y)	(N, N)	0.355	0.293	0.417	20.687	< 0.001
3	(Y, N)	(N, Y)	0.335	1.152	0.518	6.657	< 0.001
4	(Y, N)	(N, N)	0.074	0.002	0.146	3.751	0.040

Figure 12: Tukey's HSD for different combinations (cwelcc_flag, subsidy)

These results suggest that both subsidy status and CWELCC participation independently have significant effects on PG_RATE. For the interaction part, the lack of a significant difference between groups (Y, Y) and (N, Y) indicates that the presence of CWELCC does not significantly change PG_RATE for centers that have a fee subsidy contract; however, other groups show significant impacts on PG_RATE.

To check assumptions for two-way ANOVA, we firstly use QQ-plot to identify normality. The following Figure 13 is for the residual plot and histogram:

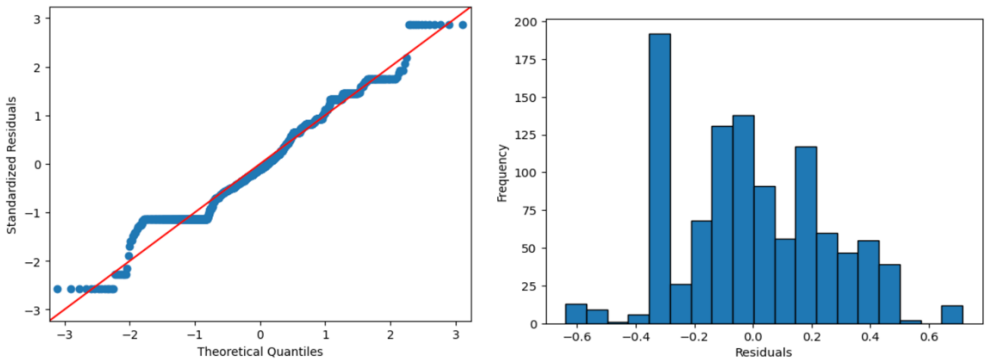

Figure 13: QQ-plot and Histogram from standardized residuals

Based on the above figure, we can find that some points do not follow the red line, and the histogram is slightly right skewed. This indicates a potential violation of the normality assumption of the residuals.
Additionally, we perform Shapiro-Wilk test, and the results are displayed in Figure 14.

Test Statistics: W	p-value
0.972	< 0.001

Figure 14: Results of Shapiro-Wilk test

Since the p-value < 0.001 < 0.05, we reject the null hypothesis and conclude that the residuals are not normally distributed.
Because the data is not drawn from normal distribution, Levene's test is used to check the second assumption, the Homogeneity of variances. Figure 15 shows the results of this test.

	Parameter	Value
0	Test statistics (W)	5.544
1	Degree of freedom (Df)	3
2	p value	0.001

Figure 15: Results of Levene's test

The p-value = 0.001 < 0.05, so we reject null hypothesis and conclude that the assumption of homogeneity of variances has been violated.
Both assumptions of two-way ANOVA are violated, so the results might be not reliable and other methods should be considered for further analysis.

5. Conclusion

For the first research question, we discover that the distribution of preschool-age (30 months to grade one) childcare spaces differs significantly across various types of operating auspices, with public and commercial agencies offering a higher percentage of spaces for this age group than their non-profit agencies. For the second research question, having a fee subsidy contract and participating the CWELCC program both independently contribute to variations in the proportion of childcare spaces for preschoolers (30 months to grade one) among total space. Additionally, the interaction of these elements might have a slightly significant influence. However, both research questions faced limitations as the data did not meet the normality and homogeneity of variances assumptions required for one-way and two-way ANOVA. The results might be not accurate, and there is a need to consider different statistical methods for a more thorough analysis.