INF 2178 Technical Assignment 2

Jinhang Luo (1005124777)

Faculty of Information
University of Toronto
INF2178: Experimental Design for Data Science
Prof.Shion Guha

March 9th, 2024

Introduction

1.1 Background

This dataset, sourced from Toronto-licensed child care centres, updated in February 2024, provides a comprehensive analysis of child care agencies across the city through the lens of licensed child care providers. The dataset is particularly noteworthy for its detailed breakdown of child care centres by operating auspice (i.e., non-profit, commercial or public) and its breakdown of child care centres capacity by age group (infants through elementary school-age children). These comprehensive data are invaluable to policymakers, educators, and practitioners in developing targeted strategies to strengthen childcare infrastructure and service, thereby closing gaps in accessibility and affordability and ultimately supporting the developmental needs of children and the economic needs of families in cities.

1.2 Dataset description

The dataset contains 1,063 records and can be found at the following link: https://open.toronto.ca/dataset/licensed-child-care-centres/ (Toronto Open Data, 2024). Numerous attributes are present in the dataset. However, our main focus is exploring the interaction between variables such as operating auspice, building type, and the total number of child care spaces.

1.3 Research question

In analyzing the dataset, we aimed to explore the structural and operational differences between these child care centres. In particular, to discover how different management models may affect the availability of childcare centres. Moreover, we seek to understand the impact of building type on these centres' capacity and service delivery. Consequently, two research questions have been formed.

> **1.** How does operating auspice (commercial, non-profit, or public) affect the total number of child care spaces?
> **2.** How do differences in building type (e.g., public elementary school, house) interact with operating auspices to affect the number of child care spaces (all age groups)?

Data Cleaning

We tidied up the dataset by removing unnecessary columns and renaming the remaining ones. There was no need to reshape the dataset, as it was already in an optimal format for our analyses.

Quantitative Analysis (One-way ANOVA)

We will use the one-way Anova Test to determine whether the operating auspice (commercial, non-profit, or public) would significantly influence the total number of child care spaces.

As seen in **Table 1**, the p-value is around 5.06e-10, and the F-statistic is about 21.84. Since the p-value is less than $\alpha = 0.05$, it can be shown that there is enough evidence to reject the

null hypothesis of Anova. It also follows that the operating auspice statistically affects the total number of child care spaces. On the other hand, since the F-statistic is a large number, which demonstrates that the differences between the means of the groups are equally statistically significant.

Table 1. The Oneway ANOVA Table (Total Child Care Spaces as Outcome Variable)

	Df	Sum sq	Mean sq	F value	Pr(>F)
Operating auspice	2	9.61e+04	48056.06	21.84	5.06e-10***
Residuals	1060	2.33e+06	2200.06	-	-

Source: Licensed Child Care Centres Dataset From *Toronto Open Data*.
*** p < 0.001; ** p < 0.01; * p < 0.05.

3.1 Assumption Checks

A. Normality of Residuals

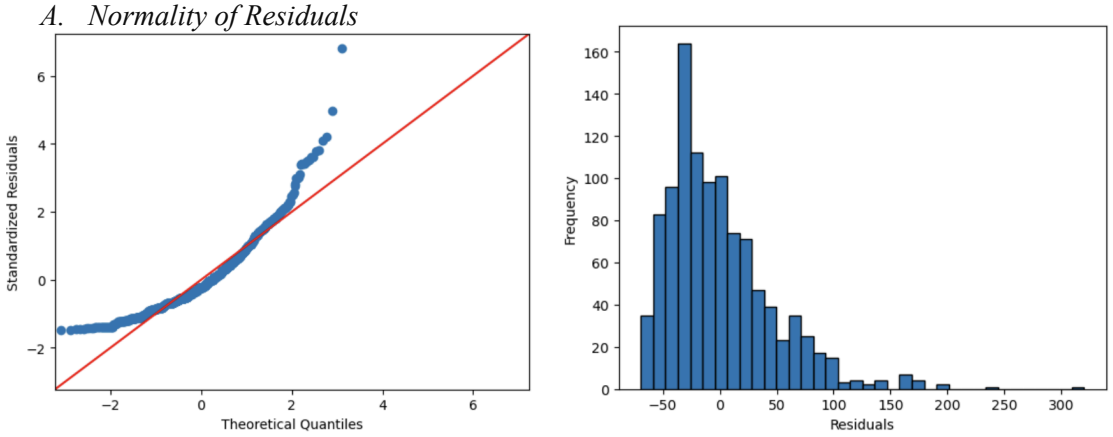

Figure 1: The Q-Q Plot and Histogram Plot from ANOVA Table

As shown in **Figure 1,** we used the Box-Cox transformation method since the residuals from the initial ANOVA model did not meet the normality assumption. This technique optimizes the data transformation to approximate a normal distribution, thereby improving the validity of subsequent ANOVA tests.

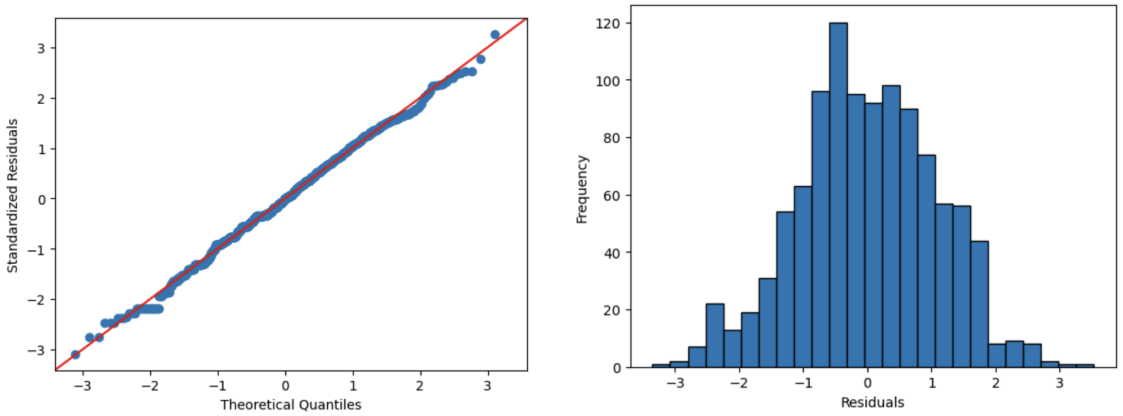

Figure 2: The Q-Q Plot and Histogram Plot from ANOVA Table (Transformed)

Table 2. The Shapiro-Wilk Test

	Test statistics (W)	P-Value
Value	0.997	0.044

Source: Licensed Child Care Centres Dataset From *Toronto Open Data*.

As shown in **Figure 2**, both diagnostic plots show that the assumption of normality of the residuals is reasonably well met. The Q-Q plot shows good agreement with the normal line, with essentially all points on the line. The shape of the histogram also supports the assumption that the residuals are normally distributed.

As shown in **Table 2,** the Shapiro-Wilk test has a test statistic (w) of 0.997 and a p-value of 0.044, which suggests that the residuals are very close to normality. However, the small significance suggests a slight deviation from perfect normality.

B. Homogeneity of Variances

As shown in **Table 3**, Bartlett's test statistic (T) for Bartlett's test is approximately 17.26 with a p-value of about 0.0002, which indicates a violation of the homogeneity assumption. However, considering the robustness of ANOVA to certain violations of its assumptions, we decided to continue with the analysis and will consider alternative approaches for specific explanations in the limitation section at the end.

Table 3. The Bartlett's Test

	Test statistics (T)	Degrees of freedom (Df)	P-Value
Value	17.26	2.00	0.0002

Source: Licensed Child Care Centres Dataset From *Toronto Open Data*.

3.2 Post-hoc tests (Tukey's HSD)

Tukey's HSD determines whether the differences between each group are statistically significant.

As shown in **Table 4**, with Total child care spaces as the dependent variable, there are statistically significant differences between Commercial and Non-Profit and between Non-Profit and Public. The notable exception is the comparison between Commercial and Public, where there is no statistically significant difference in the total number of child care spaces.

Table 4. The Tukey HSD Table (Total Child Care Spaces as dependent variable)

	Commercial : Non-Profit	Commercial : Public	Non-Profit : Public
Reject	True	False	True

Source: Licensed Child Care Centres Dataset From *Toronto Open Data*.

Quantitative Analysis (Two-way ANOVA)

To determine how the operating auspice (commercial, non-profit, or public) and type of building interactively influence the total number of child care spaces, we will use the two-way ANOVA Test. In other words, we will examine whether there is a significant interaction between these two factors that affect the total child care spaces.

As shown in **Table 5**, the p-value for the interaction between operating auspice and building type was 5.29e-08, indicating that the interaction had a statistically significant effect on the total number of child care spaces, with an F-statistic of approximately 2.95. The main effect of building type by itself was also significant, with a p-value of 4.28e-32 and an F-statistic as high as 33.9, suggesting that this factor had a significant effect on the total number of child care spaces. On the other hand, operating auspices produced a p-value of 1.00, indicating no significant effect when considered in isolation.

Table 5. The Two - way ANOVA Table (Total Child Care Spaces as Outcome Variable)

	Df	Sum sq	Mean sq	F value	Pr(>F)
Operating auspice	2	-1.61e-13	-8.06e-14	-8.17e-14	1.00e+00
Type of Building	29	9.70e+02	3.35e+01	3.39e+01	4.28e-32***
Operating auspice: Type of Building	58	1.69e+02	2.91e+00	2.95e+00	5.29e-08***
Residuals	1011	9.97e+02	9.86e-01	-	-

Source: Licensed Child Care Centres Dataset From *Toronto Open Data*.
*** p < 0.001; ** p < 0.01; * p < 0.05.

4.1 Assumption Checks

A. Normality of Residuals

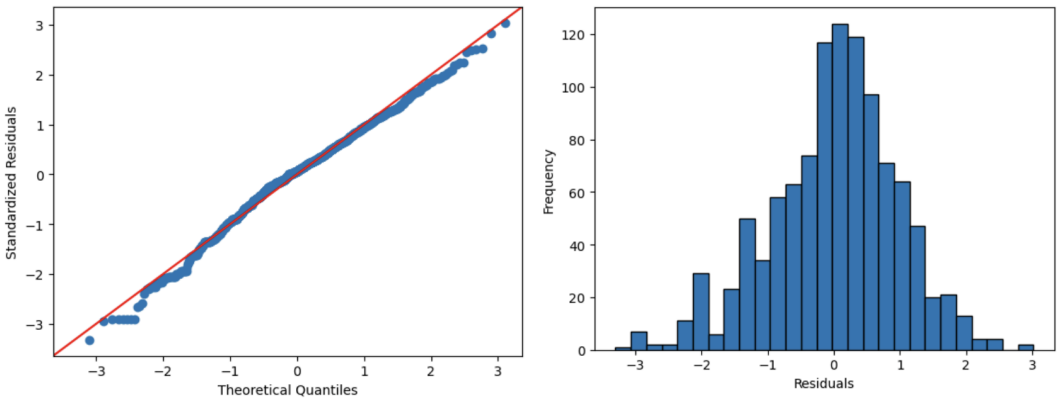

Figure 3: The Q-Q Plot and Histogram Plot from Two-way ANOVA Table (Transformed)

Table 6. The Shapiro-Wilk Test

	Test statistics (W)	P-Value
Value	0.99	3.05e-06

Source: Licensed Child Care Centres Dataset From *Toronto Open Data*.

As shown in **Figure 3**, both diagnostic plots show that the assumption of normality of the residuals is reasonably well met. The Q-Q plot shows good agreement with the normal line, with essentially all points on the line. The shape of the histogram also supports the assumption that the residuals are normally distributed.

As shown in **Table 6**, the Shapiro-Wilk test has a test statistic (w) of 0.99 and a p-value of 3.05e-06, which suggests that the residuals are normality.

B. *Homogeneity of Variances*

As shown in **Table 7**, Levene's test statistic (W) for Bartlett's test is approximately 1.89 with a p-value of about 0.0002, which indicates a violation of the homogeneity assumption. However, considering the robustness of ANOVA to certain violations of its assumptions, we decided to continue with the analysis and will consider alternative approaches for specific explanations in the limitation section at the end.

Table 7. The Levene's Test

	Test statistics (W)	Degrees of freedom (Df)	P-Value
Value	1.89	51	0.0002

Source: Licensed Child Care Centres Dataset From *Toronto Open Data*.

4.2 Post-hoc tests (Tukey's HSD)

Tukey's HSD determines whether the differences between each group are statistically significant.

As shown in **Table 8**, there is a significant difference in total child care spaces between building types. "True" indicates that there is a statistically significant difference in total child care space between the building types being compared, which suggests that the building type has a significant impact on the availability of child care centres." False," in other words, shows that there is no statistically significant difference between the building types, which suggests similarity in child care spaces between these types.

Table 8. The Tukey HSD Table (Total Child Care Spaces as dependent variable)

	Commercial Building	House	Other	Place of Worship	Public High School	Public Elementary School	Public Elementary (French)	Purpose Built
Catholic Elementary	True	True	True	True	True	True	True	False
Church	False	False	False	False	False	True	True	False
…	…	…	…	…	…	…	…	…
Public Elementary School	False	False	False	False	True	False	False	True

Source: Licensed Child Care Centres Dataset From *Toronto Open Data*.

4.3 Interaction Plot

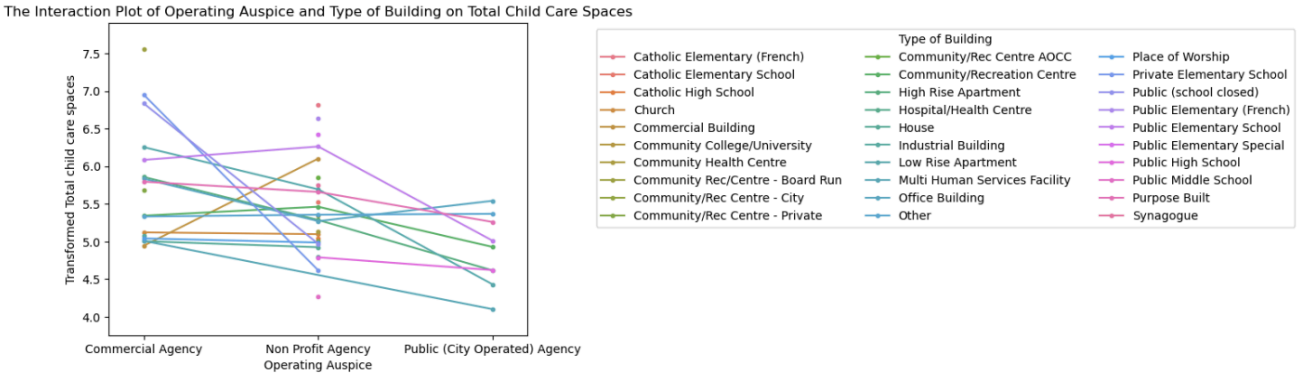

Figure 4: The Interaction Plot of Operating Auspice and Type of Building On Total Child Care Spaces

As shown in **Figure 4**, the interaction plot shows how building type and operating auspice combine to influence total child care space. Most building types are trending downward. For example, the total number of child care spaces in the public elementary school building type decreases significantly depending on the operating auspice from non-profit to public. This shows that public Agencies Are associated with fewer spaces in certain building types compared to non-profit Agencies. Moreover, most of the lines in the plot are intersecting, indicating that the relationship between operating auspices and total child care spaces is correlated and depends on the specific building type.

Limitation

1. The findings from this dataset may be limited to other cities or regions, given that they are derived from a specific dataset for the City of Toronto.

2. Bartlett's and Levene's tests indicate that the assumption of homogeneity of variance is violated. While ANOVA is robust to some deviations from this assumption, the extent of the deviation here may have influenced the investigation's results. Alternative methods, such as Welch's ANOVA or non-parametric tests, may be considered to address this issue in later stages.

Conclusion

In summary, our investigation demonstrates a complex dynamics relationship between operating auspices and building types for total child care spaces in Toronto. First, it is clear that non-profit agencies tend to provide more space than public agencies. Second, certain buildings (e.g., public elementary schools) show variability in offering child care space, with public institutions providing significantly less child care space compared to non-profit institutions. For policymakers and child care centres, these insights highlight the need for targeted strategies to address the unique resource and space challenges posed by different auspice types and buildings.

Appendix A

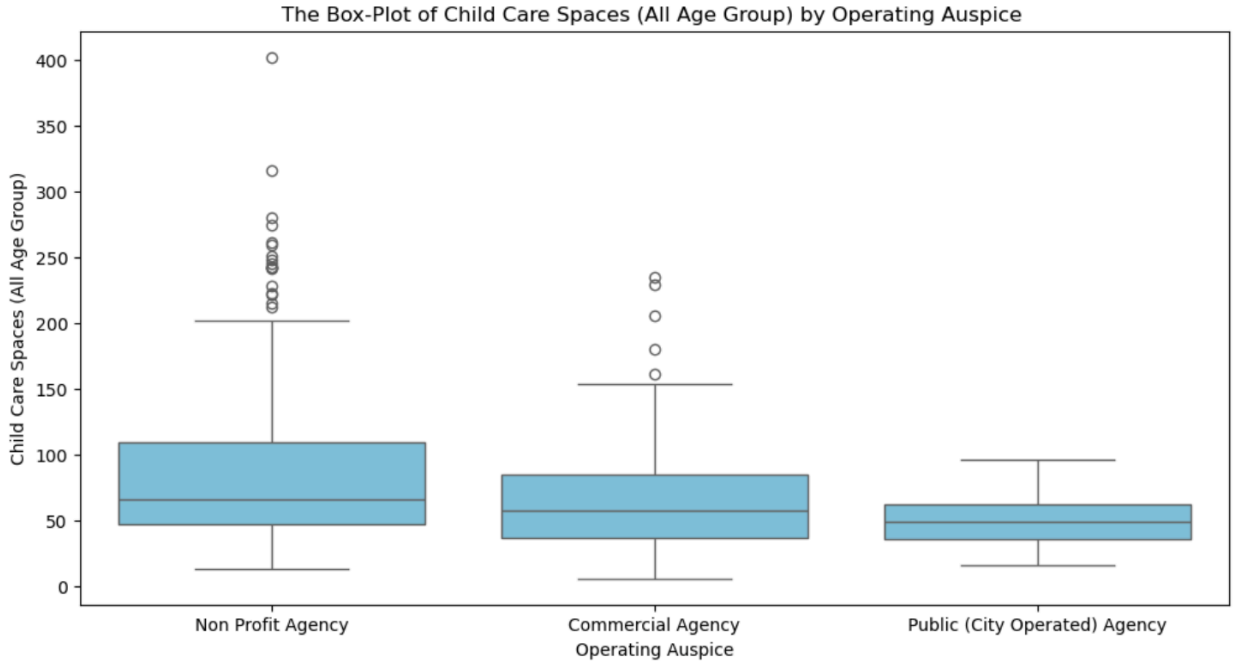

Figure A1. Boxplot of Total Child Care Spaces by Operating Auspice

The box plot illustrates the distribution of total child care center space under different operating supports (commercial, nonprofit, or public). As seen in the plot, the distribution is wider for nonprofit providers, whose values are generally higher than those of commercial and public providers, suggesting greater variability.

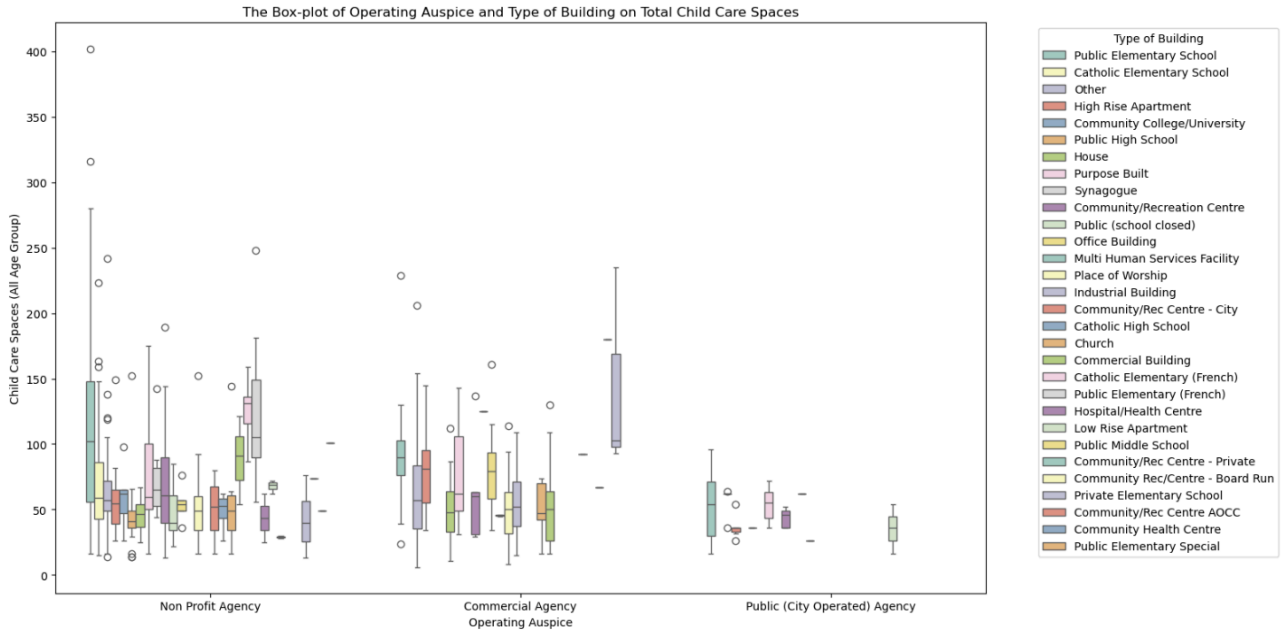

Figure A2. Boxplot of Total Child Care Spaces by Operating Auspice

This boxplot displays the interaction between operating auspice and building type on the total number of child care spaces. Each color represents a different type of building, and the boxes are grouped by operating auspice, providing a graphic comparison of the central tendency and variability within each group. It shows that the capacity of public elementary schools and commercial buildings varies widely, suggesting that there are differences in the capacity of child care centers across building types.

Reference List

Open data dataset. (2024). Retrieved from https://open.toronto.ca/dataset/licensed-child-care-centres/