University of Toronto

INF2178

Experimental Design for Data Science

Technical Assignment 2

Instructor

Professor Shion Guha

Submitted by

Lam Hong Kevin Ching

1009243043

2024/03/09

1. Introduction

In the bustling city of Toronto, childcare services are a pivotal component of the urban infrastructure, supporting families and contributing to the economic vitality by enabling parents to participate in the workforce. As the demand for accessible and quality childcare continues to grow, understanding the landscape of these services, including the capacity and factors influencing availability, becomes crucial for stakeholders at all levels. This report utilizes a comprehensive dataset provided by the City of Toronto, which offers an in-depth look at licensed childcare centers across the city. The dataset includes information on each center's location ID, location name, address, operating auspice, each type of room space, the availability of subsidies, and the implementation of the Canada-Wide Early Learning and Childcare (CWELCC) agreement flag—a marker of centers participating in a national effort to reduce childcare costs for families, etc.

The research seeks to address two critical questions within this context:

1. **Research Question 1:** Does the operating auspice of childcare centers significantly affect the total number of childcare spaces available for all age groups?
2. **Research Question 2:** How do subsidy availability and the CWELCC flag impact the total space available in childcare centers, and is there an interaction effect between these two variables on space availability?

By investigating these questions, this report aims to uncover patterns and potential disparities in childcare space availability, offering insights that could inform policy decisions, guide parents in making informed choices about childcare, and assist providers in identifying areas for expansion or improvement.

2. Data Cleaning and Data Wrangling

The raw dataset comprises 17 columns and 1063 rows. Upon the initial examination, I determined that extensive data cleaning was not required for the analysis. In the process of preparing the dataset for analysis, several columns were dropped to focus the study on variables directly relevant to the research objectives. This decision was based on the rationale that these variables were not essential for examining the core relationships of interest.

A. Observations and Considerations:

The following are short descriptions of each column:

- **AUSPICE**: The management type of the childcare center (e.g., Non-Profit Agency).
- **bldg_type**: The type of building the childcare center is located in.
- **IGSPACE, TGSPACE, PGSPACE, KGSPACE, SGSPACE**: The number of spaces for infant, toddler, preschool, kindergarten, and school-age groups, respectively.
- **TOTSPACE**: Total number of spaces available at the childcare center.
- **subsidy**: Indicates if the childcare center is subsidy eligible (Y/N).
- **cwelcc_flag**: Indicates if the childcare center is part of the Canada-wide Early Learning and Childcare system (Y/N).

B. Feature Engineering:

The only new feature added to the dataset is indicated as follow:

- **interaction**: This column represents the interaction term between the subsidy status (indicating whether a childcare center is subsidy eligible) and the cwelcc_flag (indicating whether a childcare center is part of the Canada-wide Early Learning and Childcare system).

3. Exploratory Data Analysis (EDA)

The descriptive statistics (Table1) and box plot (Figure 1) visualization collectively reveal distinct patterns in the availability of childcare spaces across different age groups. Infant spaces are notably scarce, with an average of fewer than 4 per center, and a maximum offering of 30, highlighting a significant disparity in provision across centers. In contrast, toddler spaces are more abundant, with an average of 11.6 and some centers providing as many as 90 spaces. Preschool spaces are even more prevalent, with an average of 24.3 spaces and a peak availability of 144, indicating that centers are more likely to cater to this age group. Kindergarten spaces, with an average of 14.3 and a maximum of 130, show moderate availability, while school-age spaces present a wider variability in distribution, with an average of 21.6 and some centers offering up to 285 spaces.

	IGSPACE	TGSPACE	PGSPACE	KGSPACE	SGSPACE	TOTSPACE
MEAN	3.89	11.60	24.25	14.25	21.26	75.67
STD	6.09	12.08	18.57	20.49	30.42	47.81
MIN	0	0	0	0	0	6
25%	0	0	16	0	0	43
50%	0	10	24	0	0	62
75%	10	15	32	26	30	97
MAX	30	90	144	130	285	402

Table 1: Dataset Quantitative Data Statistics

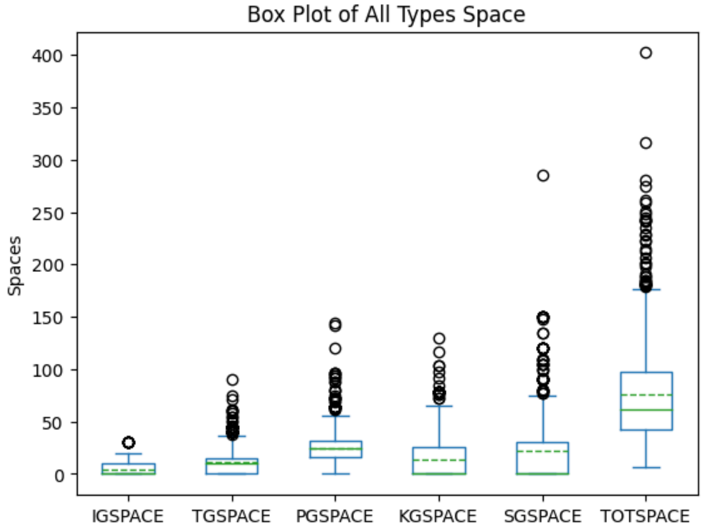

Figure 1: Box Plot of All Types Space

The box plot underscores these findings, showcasing a range of distribution skews and outliers, particularly within the toddler, preschool, and school-age groups, where numerous centers exceed the norm significantly, indicative of a right-skewed distribution. This skewness is also reflected in the position of the medians, which are notably lower within the box, especially for infant and toddler spaces, suggesting that the majority of childcare centers offer fewer spaces for these age groups. The total spaces box plot reveals a broad spectrum of center capacities, with a median that, while lower relative to the age-specific categories, indicates a general trend toward centers with lower to medium total capacities, despite the presence of some centers with significantly higher totals. This overall variability, from small to large centers, points to a diverse childcare landscape.

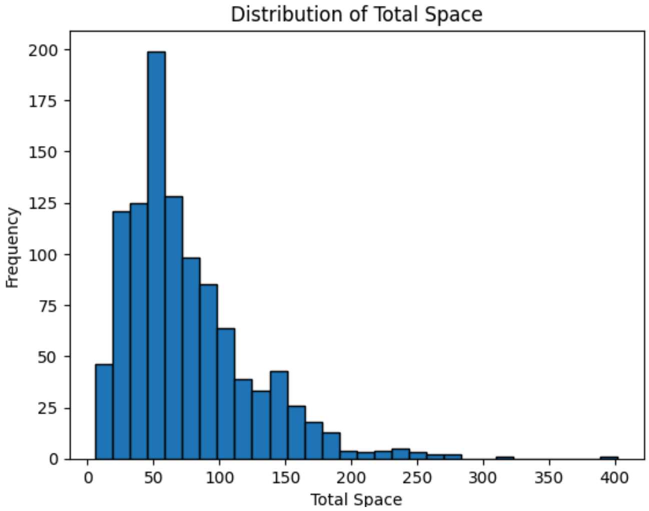

Figure 2: Distribution of Total Space

Meanwhile, the histogram (Figure 2) displays the distribution of total spaces across childcare centers. The distribution is right-skewed, as indicated by the longer tail extending towards the higher number of spaces. The majority of childcare centers have a lower number of total spaces, with the frequency peaking sharply for centers with fewer than 50 spaces and then gradually declining as the number of spaces increases. Very few centers have more than 250 spaces, suggesting that such high capacity is uncommon. The histogram reveals that while there is a range of capacities, there is a clear concentration of centers with fewer spaces. This pattern suggests that smaller centers constitute a significant portion of the sample, with larger centers being relatively rare. The presence of outliers on the higher end could indicate that a small number of centers cater to a much larger population, or possibly that they have more resources or infrastructure to offer a greater number of spaces.

4. Operating Auspice and Total Space

Research Question 1: Does the operating auspice of childcare centers significantly affect the total number of childcare spaces available for all age groups?

The first research question probes the influence of the operating auspice on the availability of childcare spaces across all age groups in childcare centers. The operating auspice refers to the administrative body managing the center. The underlying hypothesis is that the management type might correlate with the total number of spaces offered, potentially due to differences in funding, resources, or operational objectives. To investigate this, a one-way ANOVA test is employed. In this context, it serves to compare the mean number of total spaces available in childcare centers across different categories of auspices.

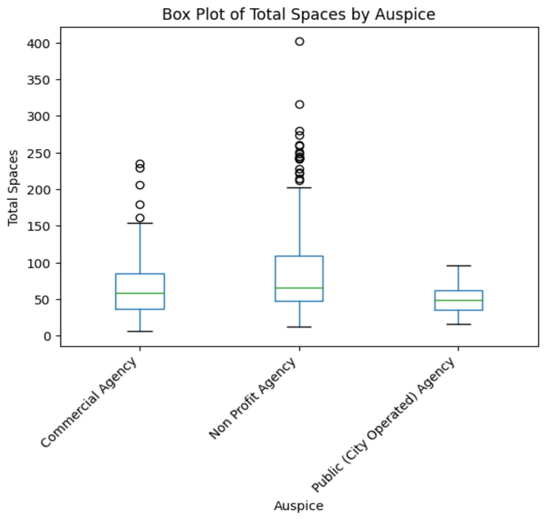

Figure 3: Box Plot of Total Space by Auspice

Before conducting One-way ANOVA, I drew a box plot (Figure 3), which illustrates the distribution of total spaces available in childcare centers grouped by operating auspice. The median total spaces for Commercial and Non-Profit agencies are similar, with Non-Profit Agencies displaying a slightly higher median, suggesting that, on average, they may offer more spaces than Commercial Agencies. However, the Non-Profit category also exhibits greater variability in the number of spaces, as indicated by the IQR, and has more outliers, showing that some Non-Profit Agencies offer a much larger number of spaces compared to their peers. In stark contrast, Public Agencies show a much lower median and a tighter IQR, which indicates less variability and generally fewer spaces offered. The absence of outliers for Public Agencies suggests a more consistent service provision across this category.

	F	**PR(>F)**
C(AUSPICE)	21.843051	5.057716e-10

Table 2: One-way ANOVA Test Results

Table 3: Tukey’s HSD Test Results

Group 1	**Group 2**	**P-value**
Non-Profit	Commercial	0.001000
Non-Profit	Public (City Operated)	0.001000
Commercial	Public (City Operated)	0.077966

The results from the one-way ANOVA test (Table 2) reveal a statistically significant difference in the number of total childcare spaces available based on the operating auspice of the childcare centers. The F-statistic value of approximately 21.84, coupled with a p-value significantly less than 0.001, strongly indicates that the average total spaces differ among commercial, non-profit, and public (city-operated) agencies. To further dissect these differences, a Tukey’s HSD test was conducted. The Tukey HSD results (Table 3) elucidate that non-profit agencies significantly differ from commercial agencies in their mean total spaces, with a p-value of 0.001. Similarly, non-profit agencies also differ significantly from public (city-operated) agencies, evidenced by an identical p-value of 0.001. However, when comparing commercial agencies with public (city-operated) agencies, the p-value of 0.077 suggests a marginal difference that does not reach conventional levels of statistical significance, implying that the mean total spaces provided by these two types of auspices may not be significantly different. This nuanced analysis underscores the influence of operating auspice on childcare space availability, with non-profits standing out distinctly from the other types, while commercial and public agencies exhibit similar capacity profiles.

5. Subsidy and CWELCC across Total Space

Research Question 2: How do subsidy availability and the CWELCC flag impact the total space available in childcare centers, and is there an interaction effect between these two variables on space availability?

The second research question seeks to understand the impact of subsidy availability and the CWELCC program on the total space available in childcare centers, as well as to explore the potential interaction effect between these two factors. Crucially, an interaction effect would suggest that the influence of one factor depends on the level of the other, indicating a more complex relationship between these variables and space availability. To address this multifaceted question, a two-way ANOVA test is utilized. This statistical method extends the one-way ANOVA to examine not only the individual effects of two independent variables on a dependent variable but also how these variables might interact with each other.

The results (Table 4) from the two-way ANOVA indicate significant effects for both the availability of subsidy and the presence of the CWELCC flag on the total space in childcare centers, as well as a significant interaction between these two factors. The F-statistic for the subsidy variable is quite large (approximately 46.38), with a p-value smaller than 0.05, indicating a strong effect of subsidy availability on total space. The CWELCC flag also has a significant effect, albeit weaker, with a p-value just below the 0.05 threshold. Most intriguing is the interaction term's F-statistic and associated p-value (approximately 9.03 and 0.0027 respectively), which suggest that the impact of one factor depends on the level of the other.

The Tukey's HSD test (Table 5) for interaction provides a detailed look at pairwise comparisons between different combinations of subsidy availability and CWELCC status. The results show significant differences in total space availability between centers with neither subsidy nor CWELCC and those with either or both (p-values all below 0.05), except for the comparison between centers with only a subsidy and those with both a subsidy and CWELCC, where the p-value is above 0.05, indicating no significant difference in total space availability between these two groups.

	F	PR(>F)
C(SUBSIDY)	46.375040	1.633653e-11
C(CWELCC_FLAG)	3.176381	7.499648e-02
C(SUBSIDY): C(CWELCC_FLAG)	9.027604	2.721895e-03

Table 4: Two-way ANOVA Test Results

Group1	Group2	p-adj	reject
N-N	N-Y	0.0289	True
N-N	Y-N	0.0000	True
N-N	Y-Y	0.0000	True
N-Y	Y-N	0.0015	True
N-Y	Y-Y	0.0000	True
Y-N	Y-Y	0.1456	False

Table 5: Tukey's HSD test for the Interaction

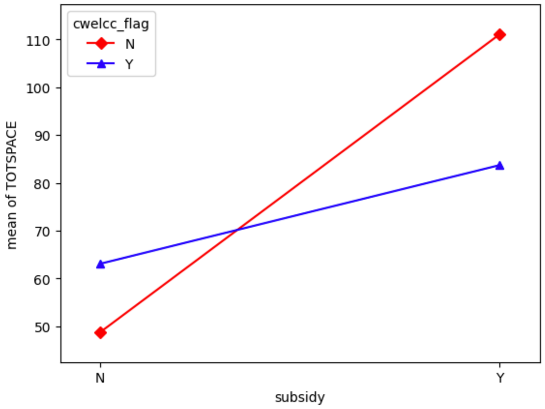

Figure 4: Interaction Plot

The interaction plot illustrates the relationship between subsidy availability, participation in the CWELCC program, and the mean total number of spaces available in childcare centers. The upward trend in both lines indicates that childcare centers with subsidy availability generally offer more spaces than those without, irrespective of their CWELCC status. Notably, the slope of the line for centers with CWELCC status is steeper, which suggests that the increase in mean total spaces associated with subsidy availability is more pronounced for centers participating in the CWELCC program. This reflects a potential interaction effect, where the impact of subsidy on space availability is enhanced by participation in the CWELCC program. Furthermore, the crossing of the lines indicates that for centers without subsidy, CWELCC participation is associated with a lower mean number of spaces, while for those with subsidy, CWELCC participation is associated with a higher mean number of spaces. This crossover interaction suggests that the effect of one factor on total spaces is not consistent across levels of the other factor, affirming the presence of an interaction effect as previously indicated by the two-way ANOVA results.

6. Limitation and Conclusion

The analysis conducted through One-way and Two-way ANOVA provided insightful findings regarding the factors affecting childcare space availability in Toronto. However, a limitation in this study stems from the fact that both ANOVA tests did not fully meet the assumption checks. While this situation is commonly encountered in real-world data and analyses, it underscores the inherent limitations of applying ANOVA tests. Such deviations from assumptions may influence the interpretation of the results, necessitating a cautious approach when generalizing findings. This limitation reflects a broader challenge in statistical analyses, where ideal conditions are seldom met in practical scenarios, thus highlighting the importance of considering alternative methods or supplementary analyses to validate findings.

This report has explored significant aspects of childcare space availability in Toronto, uncovering the influence of operating auspice, subsidy availability, and CWELCC participation on the distribution of childcare spaces. The findings from both One-way and Two-way ANOVA tests reveal meaningful patterns and disparities that can inform stakeholders, including policy decisions, parental choices, and provider strategies. Despite the limitations related to the assumptions of ANOVA, the study contributes valuable insights into the childcare landscape. It suggests areas for policy intervention and highlights the critical role of subsidies and CWELCC in enhancing childcare space availability. Future research should address the identified limitations and explore more comprehensive models to further understand the dynamics of childcare provision.