UNIVERSITY OF TORONTO
FACULTY OF INFORMATION

INF 2178 - Experimental Design for Data Science

Technical Assignment 2

By

Charlie Zhang

Course Instructor: Shion Guha

Date: March 09, 2024

Exploring Toronto's Licensed Child Care Centres

1. Introduction

The exploration of Toronto's licensed child care centers through the dataset "INF2178_A2_data.xlsx" reveals critical insights into the city's child care facilities, pivotal for the well-being and development of its youngest residents. This analysis is particularly timely, as challenges such as availability, accessibility, and affordability of child care services become increasingly pressing against the backdrop of Toronto's evolving socio-economic landscape.

2. Data Cleaning and Data Wrangling

The dataset has 16 columns and 1063 rows, remarkably without any missing values. The columns present a mix of integers (int) and strings (str), indicating the presence of both continuous data and categorical variables. Here is an overview of each column based on the initial inspection and the provided data dictionary:

a. Observations and Considerations

1. **_id**: (int) A unique identifier for each row in the dataset.
2. **LOC_ID**: (int) The location identifier for each child care centre.
3. **LOC_NAME**: (str) The name of the child care centre.
4. **AUSPICE**: (str) The type of management running the centre, such as 'Non Profit Agency'.
5. **ADDRESS**: (str) The street address of the child care centre.
6. **PCODE**: (str) The postal code where the child care centre is located.
7. **ward**: (int) The ward number where the child care centre is situated.
8. **bldg_type**: (str) The type of building the child care centre is housed in.
9. **BLDGNAME**: (str) The name of the building, if applicable.
10. **IGSPACE**: (int) The number of infant spaces available at the centre.
11. **TGSPACE**: (int) The number of toddler spaces available at the centre.
12. **PGSPACE**: (int) The number of preschool spaces available at the centre.
13. **KGSPACE**: (int) The number of kindergarten spaces available at the centre.
14. **SGSPACE**: (int) The number of school-age spaces available at the centre.
15. **TOTSPACE**: (int) The total number of spaces available at the centre.
16. **subsidy**: (str) Indicates whether a subsidy is available at the centre ('Y' for yes or 'N' for no).
17. **cwelcc_flag**: (str) Indicates if the centre is part of the City of Toronto's Child Care Fee Subsidy Program ('Y' for yes or 'N' for no).

The dataset appears to be in good shape for our analysis with no missing values in the key columns of interest (AUSPICE, TOTSPACE, subsidy, and ward). The 'BLDGNAME' column has some missing values, but this won't affect our analysis.

The 'AUSPICE' column has three unique types: Non-Profit Agency, Commercial Agency, and Public (City Operated) Agency. For the 'subsidy' status, there are two unique values indicating whether subsidies are available ('Y' for yes, 'N' for no). The dataset covers child care centers across 25 different wards.

Next, I will proceed with the statistical analysis to address the research questions. I will start with Research Question 1, exploring the disparity in total child care center capacity based on the operating auspice (Non-Profit vs. Commercial vs. Public). I will use a one-way ANOVA for this purpose. Following that, I will tackle Research Question 2 by performing a two-way ANOVA to assess how subsidy availability and ward location interact to influence the total capacity.

b. **Feature engineering**
I created **totalCapacity** column to aid in my analysis. By summing up **TOTSPACE** for each unique **LOC_NAME** and assigning it back to the data frame, I calculated the total capacity available across all entries for each child care centre. It is crucial for understanding the overall capacity of each centre, especially if there are multiple entries for the same location.

3. Exploratory Data Analysis

I extensively analyzed the dataset and created bar graphs and boxplots to further explore the different features and columns to se how they varied across different levels. I started by summarizing quantitative data.

	LOC_ID	ward	IGSPACE	TGSPACE	PGSPACE	KGSPACE	SGSPACE	TOTSPACE
count	1063.0	1063.0	1063.0	1063.0	1063.0	1063.0	1063.0	1063.0
mean	8087.89	12.51	3.90	11.60	24.26	14.26	21.66	75.67
std	5151.25	7.03	6.09	12.09	18.58	20.49	30.42	47.82
min	1013.0	1.0	0.0	0.0	0.0	0.0	0.0	6.0
25%	1862.0	6.0	0.0	0.0	16.0	0.0	0.0	43.0
50%	8826.0	12.0	0.0	10.0	24.0	0.0	0.0	62.0
75%	13245.0	19.0	10.0	15.0	32.0	26.0	30.0	97.0
max	14504.0	25.0	30.0	90.0	144.0	130.0	285.0	402.0

Table 1 – summary of space types

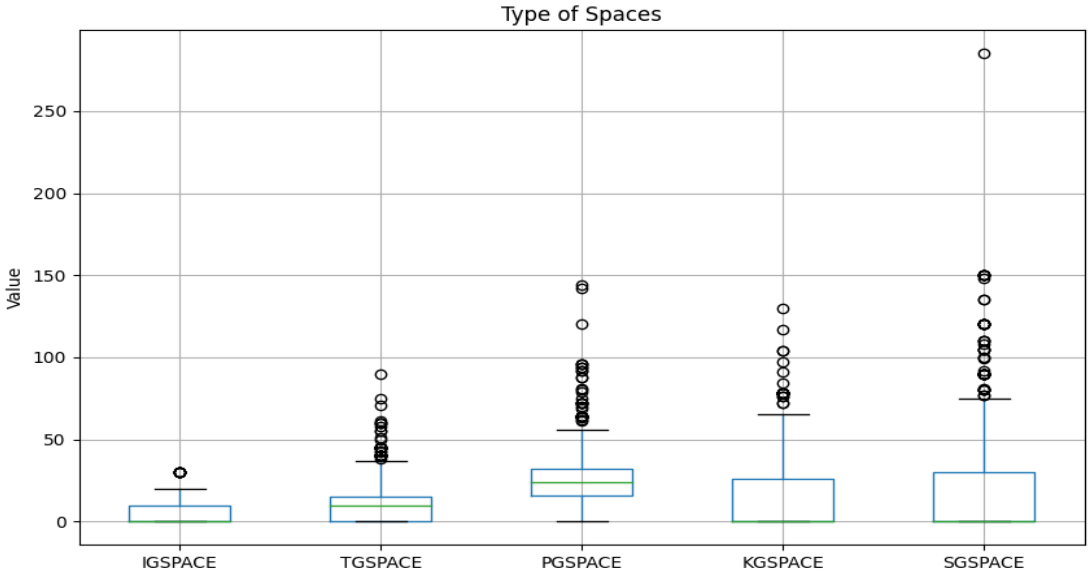

Figure 1 – boxplot for Type of Spaces

4. Total Capacity Across the different types of Auspices - One Way Anova

Research Question 1: Is there a disparity in total child care centre capacity based on the operating auspice?

	df	sum_sq	mean_sq	F	PR(>F)
AUSPICE	2.0	84461.991	42230.995	18.933	<0.001
Residual	1060.0	2364373.938	2230.541	NaN	NaN

Table 2 - Anova Table (One-Way Anova) of totalCapacity by Auspice

I conducted a one-way ANOVA test to analyze the total capacity across different levels of Auspice. The results indicate a significant difference in total capacity among the various auspice levels, with a p-value of less than 0.001.

Test statistic(w)	p-value
0.90279	<0.001

Table 3 - Shapiro Wilk Test Result

Parameter	Value
Test statistics (W)	15.79710
Degrees of freedom (Df)	2.00000
p-value	< 0.001

Table 4 - Levene's Test Result

After confirming the significance of the overall test, I proceeded to assess the assumptions of normality and homogeneity of variance. The Shapiro-Wilk test for normality and Levene's test for homogeneity of variance both resulted in p-values below 0.05, indicating deviations from these assumptions.

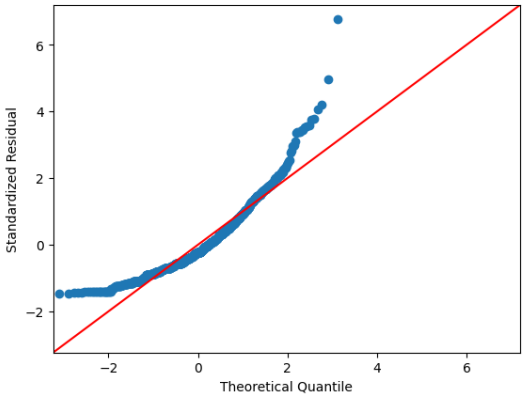

Figure 2 – boxplot for Type of Spaces

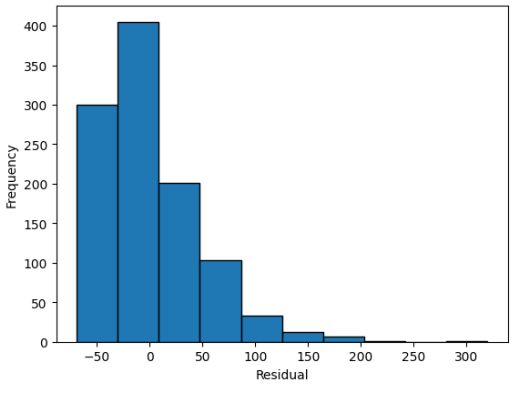

Figure 3 – QQ plot for normality check

Additionally, I conducted post-hoc tests to explore specific pairwise comparisons between groups. The results revealed significant differences in total capacity for all pairs.

group1	group2	meandiff	p-adj	lower	upper	reject
Commercial Agency	Non Profit Agency	15.3157	< 0.001	7.8488	22.7826	true
Commercial Agency	Public (City Operated) Agency	-19.0189	0.047	-37.8158	-0.2221	true
Non Profit Agency	Public (City Operated) Agency	-34.3346	< 0.001	-52.5698	-16.0994	true

Table 5 - Pairwise Comparison of Auspice (Post-hoc)

Based on these findings, I concluded that there is indeed a notable difference in total capacity across different types of **Auspice**. Notably, City Operated Auspice showed the lowest capacity among the categories. This suggests a need for resource reallocation, with a recommendation to direct more funding towards City Operated Child Centers to address the capacity gap and promote a more equitable distribution of childcare resources.

5. Differences in totalCapacity Across Wards and Subsidy Levels - Two-way ANOVA

Research Question: Does the total capacity of a child care center differ by ward and subsidy status?

In addressing this query, I utilized a two-way ANOVA framework, focusing on two primary independent variables: **Ward** and **Subsidy** status, against the dependent variable, **totalCapacity**. This analysis also probes the interaction effect between Ward and Subsidy statuses to understand their combined influence on child care center capacities.

For analytical clarity, I curated a dataframe, **ward_subsidy**, comprising these three critical columns. The Subsidy column was further refined to denote **Y: available** and **N: unavailable**, enhancing interpretability. Additionally, a separate dataframe, **ward_subsidy_2**, facilitated Exploratory Data Analysis (EDA), ensuring balanced treatment levels across all 25 wards, with each ward represented by child care centers both with and without subsidy.

Visualization Insights:

The boxplot underscores a discernible trend: centers with subsidies generally boast larger capacities than those sans subsidy, with notable variability across wards.

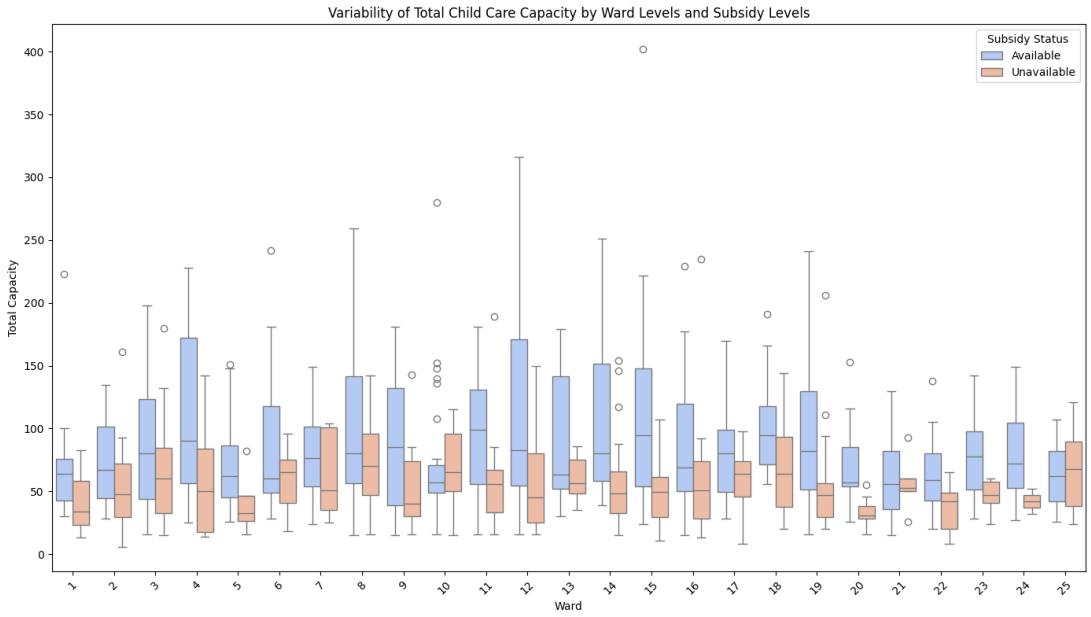

Figure 4 – Boxplot for Total Capacity by Ward and Subsidy Levels

Our interaction plot further delineates the nuanced differences in mean total capacities between subsidy levels across wards, highlighting exceptions such as Wards 10 and 21, where capacities align closely.

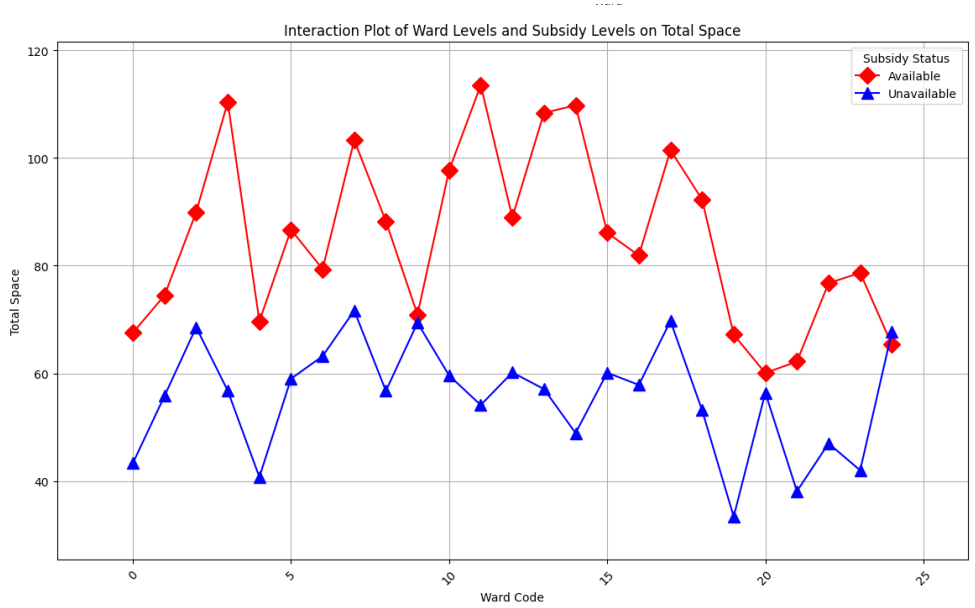

Figure 5 – Interaction Plot for Total Capacity by Ward and Subsidy Levels

Two-way ANOVA Findings:

index	sum_sq	df	F	PR(>F)	mean_sq
C(ward)	145685.34	24.0	2.98	< 0.001	6070.22
C(subsidy)	227852.67	1.0	111.72	< 0.001	227852.67
C(ward):C(subsidy)	55644.0	24.0	1.14	0.295	2318.5
Residual	2066082.64	1013.0	NaN	NaN	2039.57

Table 6 - Pairwise Comparison of Auspice (Post-hoc)

With a p-value less than 0.001, I conclude that total capacity significantly differs by ward. Similarly, a p-value less than 0.001 underlines significant differences in total capacity by subsidy status. The interaction's p-value exceeds 0.05, indicating insufficient evidence to claim that total capacity's variation by ward is dependent on subsidy availability.

Despite no significant ward-specific treatment levels, Tukey's HSD confirms subsidy status as a differentiator in total capacity. Furthermore, certain wards (e.g., 4, 14, and 15) revealed significant interactions between subsidy levels, suggesting localized disparities in capacity influenced by subsidy availability.

ANOVA Assumptions Testing:

Normality of Residuals: Both visual and statistical tests (e.g., QQ plots, histograms, Shapiro-Wilk test with $p < 0.001$) suggest deviations from normality, indicating the model's poor fit and skewed residuals.

Parameter	Value
Test Statistic	0.93512
p-value	<0.001

Table 7 - Shapiro Wilk Test Result

Parameter	Value
Test statistics (W)	3.024
Degrees of freedom (Df)	49
p value	< 0.001

Table 8 - Levene's Test Result

Equal Variance: Levene's Test, prompted by the residuals' non-normal distribution, yielded a p-value less than 0.001, leading us to reject the hypothesis of equal variance across treatments.

The two-way ANOVA analysis, coupled with assumption testing, suggests that while significant insights were gleaned regarding the impacts of ward and subsidy on child care center capacities, the foundational assumptions of ANOVA were not met. This discrepancy underscores the necessity for cautious interpretation of the results. Future analyses might consider data transformation or alternative statistical methods, like Friedman's Test, which do not presuppose equal variances, to validate these findings further.

6. Conclusion:

Toronto's licensed child care centers exhibit significant capacity disparities by auspice type, ward location, and subsidy availability. Non-Profit agencies generally boast higher capacities, while subsidy availability correlates with larger capacities. Targeted interventions are needed to address capacity gaps and ensure equitable access to childcare services. Policy efforts should prioritize funding for underserved areas and incentivize partnerships. Despite deviations from ANOVA assumptions, this analysis underscores the importance of data-driven decision-making to promote accessible and affordable childcare for all Toronto residents, enhancing early childhood development outcomes and family well-being.