INF2178 Technical Assignment 2

Name: KA YUEN LEE

Student Number: 1010073974

Professor: Shion Guha

1. Introduction

In recent years, Toronto has experienced a substantial evolution in the dynamics of childcare services, the increasing costs and fluctuating availability of childcare options have become critical factors affecting families across the city. Recognizing the need for a granular analysis of this important service, this report embarks on an in-depth examination of Toronto's licensed child care centers. Leveraging the comprehensive dataset 'INF2178_A2_data.xlsx', this report aims to illuminate the operational patterns and capacity metrics of these centers, providing insights into the availability of childcare spaces for different age groups and the distribution of these services across the city's diverse neighborhoods.

There are 2 research questions that will be discussed in the following analysis:

> Research Question 1: Considering the auspice under which child care centers operate (Commercial, Non-Profit, Public), what is the average capacity and capacity per center within these auspice categories, and are there significant differences in capacities?
>
> Research Question 2: Is there an interaction effect between the auspice under which child care centers operate (Commercial, Non-Profit, Public) and their participation status in the CWELCC system on the centers' capacity (TOTSPACE)?

By addressing these questions, we can have a report that not only charts the current landscape of childcare services in Toronto but also serves as a foundational analysis for policy recommendations and future planning.

2. Data cleaning

The raw dataset has a total of 17 columns with 1063 entities (rows). Our dataset comprises numerous variables, encompassing unique identifiers, operational details, and capacity metrics for childcare centers in Toronto. With a focus on the questions at hand, we focus on the following columns that were pertinent to our analysis:

A. Observation and Considerations:

a). After initial review, we observed that the dataset is good, with a wealth of information across several dimensions. For this study, we concentrated on the following columns with short description :

- ✧ TOTSPACE: childcare spaces for all age groups.
- ✧ AUSPICE: Operating auspice (Commercial, Non-Profit, or Public)
- ✧ Cwelcc_flag: Space participates in CWELCC or not (Yes/No)

b). The column 'BLDGNAME' has 348 missing values. However, these missing data do not significantly impede our capacity to conduct a thorough analysis as our research questions do not directly rely on this column.

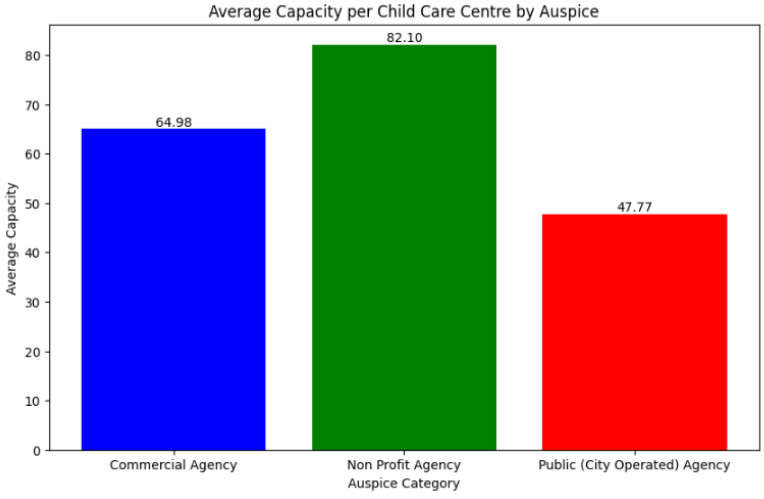
B. Feature Engineering:

In this part, we created one feature into our dataset that reflects the average capacity per center within the three auspice categories: Commercial, Non-Profit, and Public. This new feature will enable us to quantitatively compare the capacity across these categories:

$$Average\ Capacity\ per\ Auspice = \frac{\sum(Total\ Spaces\ in\ Each\ Centre\ within\ Auspice)}{Count\ of\ Centres\ within\ Auspice}$$

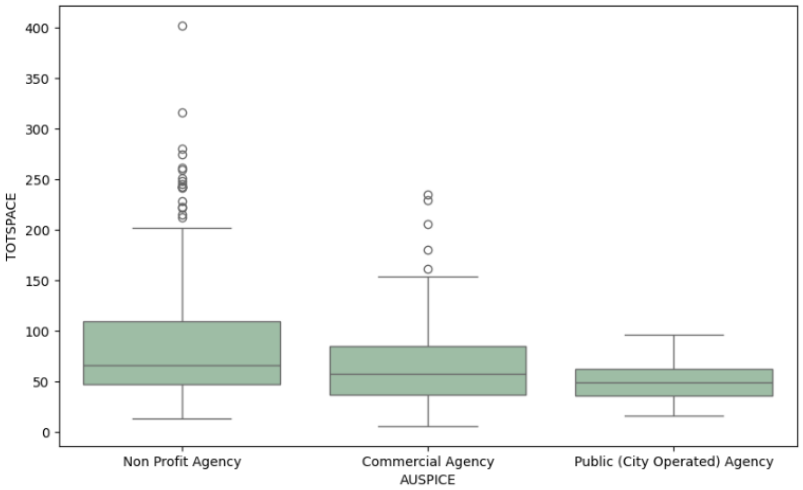
3. Exploratory Data Analysis

Research Question 1: Considering the auspice under which child care centres operate (Commercial, Non-Profit, Public), what is the average capacity and capacity per centre within these auspice categories, and are there significant differences in capacities?



The bar chart represents the average capacity of childcare centers categorized by their operating

auspice. These differences in average capacities suggest that there is variability in the size of childcare centers based on their auspice type, and the Non-Profit Agencies have the highest capacity.



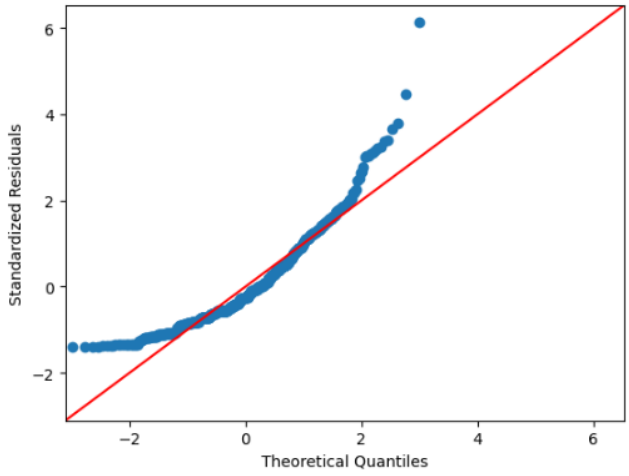
The box plot shows the total space distribution across the three categories of childcare center auspices. Overall, this plot could suggest that Non-Profit Agencies tend to have a higher capacity and greater variability in the size of their centers compared to Commercial and Public agencies.

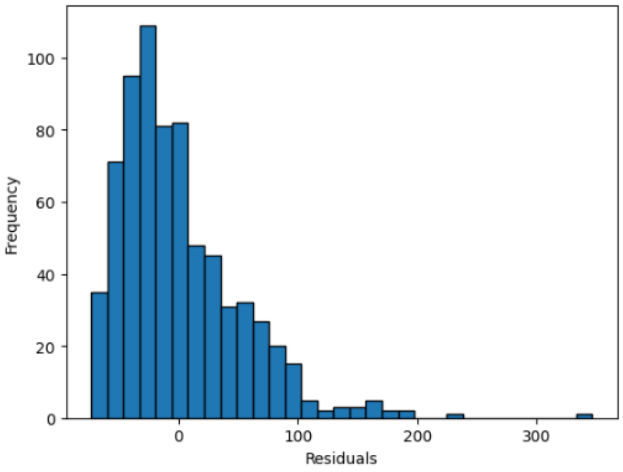
	df	sum_sq	mean_sq	F	PR(>F)
C(AUSPICE)	2	9.61E+04	48056.06	21.843051	5.06E-10
Residual	1060	2.33E+06	2200.062	NaN	NaN

The results from the one-way ANOVA underscore the visual observations from the box plot. With an F-statistic value of approximately 21.84 and a highly significant p-value of approximately 5.057716e-10, we reject the null hypothesis, which posited that the mean total space is the same for all auspice types. Instead, we conclude that the total space provided by childcare centers is indeed affected by the type of auspice under which they operate.

Group1	Group2	Diff	Lower	Upper	q-value	p-value
Non-Profit Agency	Commercial Agency	16.807	3.994	29.619	4.357	0.006
Non-Profit Agency	Public (City Operated) Agency	36.178	8.674	63.682	4.369	0.006
Commercial Agency	Public (City Operated) Agency	19.371	-10.142	48.885	2.18	0.273

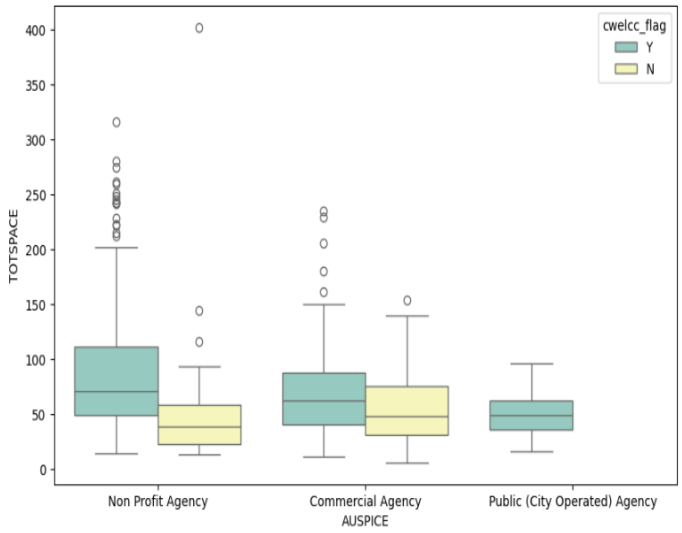
Due to the means are different, Tukey's HSD test was conducted for post-hoc comparison. The Tukey HSD test shows that Non-Profit Agencies have a significantly higher average capacity compared to both Commercial Agencies and Public Agencies, while Commercial Agencies and Public Agencies do not have a significant difference in their average capacities.

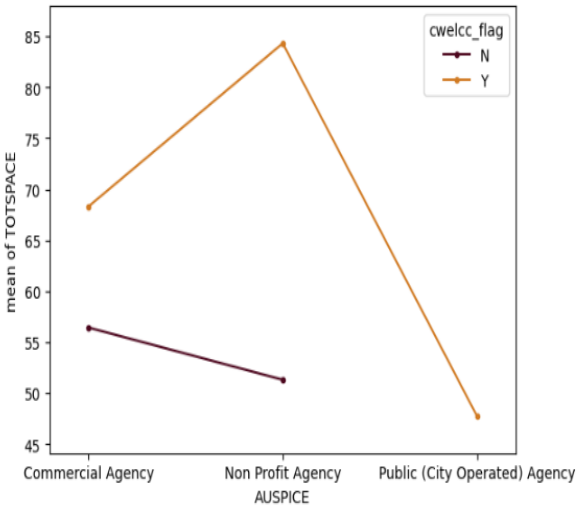




These results show that the assumption of normality is violated for the data, as both the QQ-plot and the Shapiro-Wilk test (with a w value of 0.902 and a p-value of 1.50e-25) demonstrate significant deviation from a normal distribution. This non-normality is also visually supported by the skewness observed in the QQ-plots and histograms. Moreover, the p-value is very low (0.0001) from Levene's test, so the assumption is crucial for the validity of ANOVA, is not met. Based on the above results, two assumptions are not satisfied with our data. Despite these violations of ANOVA's underlying assumptions, the descriptive statistics may still suggest that non-profit agencies have a higher average total space compared to other types of agencies.

Research Question2： Is there an interaction effect between the auspice under which child care centers operate (Commercial, Non-Profit, Public) and their participation status in the CWELCC system on the centers' capacity (TOTSPACE)?





The box plot suggests that childcare centers participating in the CWELCC system may have greater

variability in their capacity sizes compared to those not participating. To validate and assess the significance of this observation, a two-way ANOVA will be conducted. This analysis will help determine if the differences observed are statistically significant, considering the interaction between CWELCC participation and the type of auspice under which the centers operate.

From the Interaction Plot, the considerable difference in slopes between the Y and N lines, especially in Non-Profit and Public Agencies, indicates there may be an interaction effect, where the influence of CWELCC participation on total space capacity differs depending on the type of auspice. Also, it shows that participating in CWELCC has a higher mean.

	df	sum_sq	mean_sq	F	PR(>F)
C(AUSPICE)	2	1.08E+05	54167.406	25.189	6.10E-07
C(cwelcc_flag)	1	3.77E+04	37688.323	17.526	3.07E-05
C(AUSPICE):C(cwelcc_flag)	2	29495.61	14747.806	6.8580	0.0010984
Residual	1058	2.28E+06	2150.461	NaN	NaN

The two-way ANOVA results indicate a significant interaction between the type of auspice and participation in the CWELCC system, as demonstrated by the interaction term (C(AUSPICE):C(cwelcc_flag)). The significant p-value associated with the interaction term, which is below the alpha level of 0.05, suggests that the total space capacity of child care centers is likely influenced by whether they participate in the CWELCC program.

Also, Tukey's HSD test was conducted for post-hoc comparison. The Tukey HSD test shows significant differences in the mean total space (TOTSPACE) among child care centers categorized by their auspice type and participation in the CWELCC system. Particularly emphasizing that Non-Profit Agencies participating in CWELCC report higher capacities. Moreover, the Shapiro-Wilk test (with a w value of 0.897 and a p-value of 4.58e-26) demonstrates a significant deviation from a normal distribution. And Levene's test yields a p-value < 0.0001, so the assumption is crucial for the validity of ANOVA, is not met. In the result, two assumptions are not satisfied with our data.

Conclusion

In conclusion, this comprehensive analysis has explained the dynamic factors affecting the capacity of Toronto's licensed child care centers. The statistical evidence from ANOVA and Tukey's HSD post hoc tests confirms that both the type of auspice and CWELCC participation significantly influence the centers' capacity. Notably, Non-Profit Agencies participating in CWELCC tend to offer greater space. These insights are crucial for policymakers, suggesting the need for some strategies such as encouraging more organization participation in government initiatives like CWELCC.