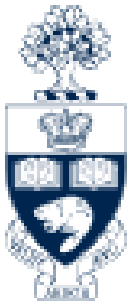

UNIVERSITY OF TORONTO
FACULTY OF INFORMATION

INF 2178 - Experimental Design For Data Science

Technical Assignment 2

By

Saad Umar

Course Instructor: Shion Guha

Date: March 09, 2024

Exploring Toronto's Licensed Child Care Centres

1. Introduction

The provision of child care centers play a pivotal role in the city of Toronto to shape the well-being and development of its youngest residents. A comprehensive dataset on licensed child care centers offers a profound insight into the operational capacities of these facilities catering to diverse age groups. However, amidst the diversity of these facilities, challenges persist. The availability, accessibility, and affordability of quality child care services remain areas of concern for families navigating the city's dynamic socio-economic landscape.
This report offers a comprehensive data exploratory analysis of Toronto's Licensed child care centers with a goal to uncover the underlying trends surrounding these centers in the city of Toronto. We delve into the dataset, named: '*INF2178_A2_data.xlsx*' to unravel the nuanced dimensions of Toronto's child care system, shedding light to address three fundamental research questions, serving as guiding principles in discovering the patterns of Toronto's child care centers:

1. **Research Question 1:** Does the total capacity of the shelters vary depending on the type of auspice, is there a significant difference in the total capacity of the different types of auspice?
2. **Research Question 2:** Do the different types of Auspices and Wards 3,8, and 25 have a significant difference on the total capacity of childcare centers
3. **Research Question 3:** Do the different types of Auspices and being a subsidized center have a significant difference on the total capacity of childcare centers.

By addressing these questions, we aim to contribute insights into dynamics of child care centers in Toronto and provide a deeper understanding that can inform more effective interventions.

2. Data Cleaning and Data Wrangling

The raw dataset has a total of 17 columns with 1063 entities (rows). After initial review of the dataset, I was confident that not much data cleaning was necessary for the scope of my analysis. However, I defined one new feature which might be necessary for future analysis. Below I have outlined my observations and the new feature which I added to the dataset:

a. Observations & Considerations

1. My analysis is quantitative and I've only used specific columns from the dataset to perform the analysis. However, I did not drop the other columns from the dataset as I feel those columns are needed for a more detailed analysis, which I plan to perform in the future. Below I have provided a short description of each columns that I have used in my analysis so far:
    - LOC_NAME: Name of the child care center
    - AUSPICE: Operating auspice (Commercial, Non Profit, or Public)
    - Ward: City Ward number
    - TOTSPACE: Child care spaces for all age groups
    - Subsidy: If the center has a fee subsidy contract (Yes/No)

2. The following column seemed to have many missing values:
    - BLDGNAME: has 715 non-null
      However, I did not perform data cleaning on this column since it was not involved in my analysis. On the contrary, I will be required to clean this column if I perform further analysis using this dataset.

b. Feature Engineering:

I created (1) new feature to add to the dataset to aid in my analysis. The feature is as follows:

1. **totalCapacity**: This column adds the values in the **TOTSPACE** columns given that the **LOC_NAME** column has the same name. There were 5 such instances where the LOC_NAME was not unique. Specifically, the **_id** of these rows were: [953, 1006], [860, 1026], [889, 998], [699, 1031], and [683, 1007].
   Thus, for these rows, where the LOC_NAME was the same, I added up the values in the TOTSPACE column and displayed it in the totalCapacity column. All other rows had the same value in this new column as the value in their respective TOTSPACE columns.

3. Exploratory Data Analysis

I performed an extensive EDA to leverage insight that could potentially help me derive insightful research questions. I started by summarizing quantitative data and then using a bunch of bar plots and boxplots to see how different features varied and how the distributions differed across different levels.

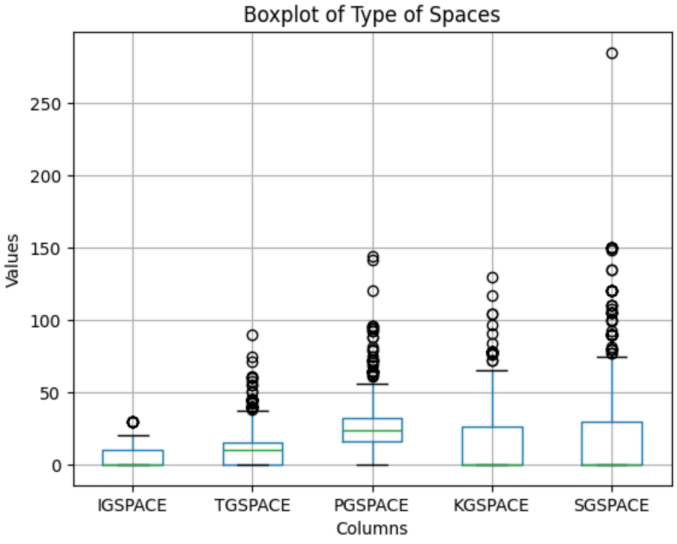

Fig 1 - Distribution of the Different type of Spaces

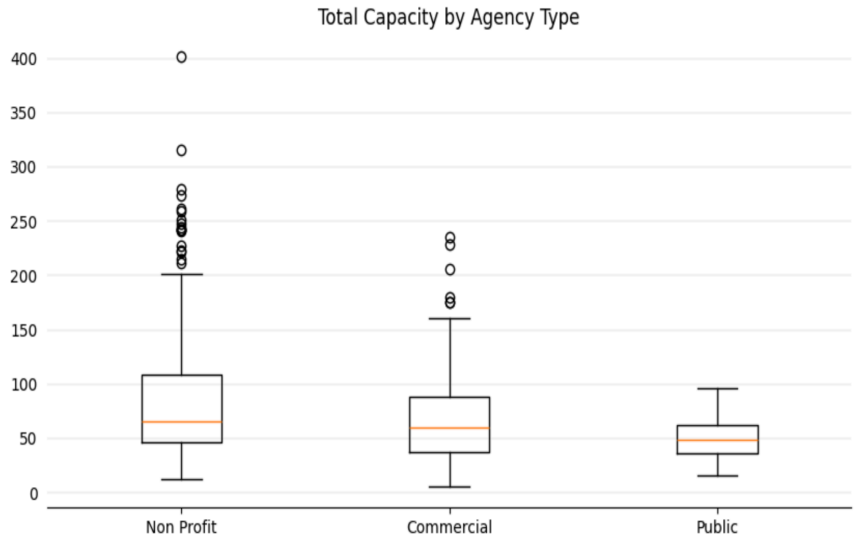

Fig 2 - Distribution of different types of Auspices by their total capacity

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

Research Question 1: Does the total capacity of the shelters vary depending on the type of auspice, is there a significant difference in the total capacity of the different types of auspice?

	df	sum_sq	mean-sq	F	PR(>F)
Auspice	2.0	84462	42231	18.93	<0.001
Residual	1060.0	2364374	2231	NaN	NaN

Table 1 - Anova Table (One-Way Anova) of totalCapacity by Auspice

I used the **one-way anova** test to determine if there is a significant statistical difference between the three different levels of Auspice. Using table 1, where I have a p-value of less than 0.001, I concluded that my null hypothesis, which is, “there is no significant difference in the total capacity of the three different auspice levels”, is rejected and there is significant difference. The **second step** was to check for the assumptions: I used the **Shapiro Wilk** test to check for the normality. In table 2 below, we can see that the p-value is less than the significance level of 0.05, hence we can reject the null hypothesis of the data being normal. These results were also backed by the ‘w test statistic’ and the QQ plot in Figure 3, as we can see that the data does not seem to follow a normal distribution. The other assumption test I used was the **Levene’s test** to check for variance homogeneity. From table 3 below, we can see that the p-value is less than the significance level of 0.05, hence we can reject the null hypothesis that, “the variance across the three groups are equal” and thus, we can conclude that the variances across the different levels of auspice are not equal. The **third step** was to perform the post-hoc tests. These tests are used to explore specific pairwise comparisons between groups when the overall omnibus test indicates a significant difference. From Table 4 below, we can observe that p-value is less than 0.05 (significance level) for all the different pairs, hence we reject the null hypothesis that there is no significant difference in the total capacity of the different types of auspice. This is true for all the pairs, as we can see in the table below.

Test statistic(w)	p-value
0.903	< 0.001

Table 2 - Shapiro Wilk Test Result

Parameter	Value
Test Statistic (W)	15.797
Degrees of Freedom	2.00
P Value	< 0.001

Table 3 - Levene’s Test Result

Conclusion: There is a significant difference in the total Capacity of the different types of Auspice. The analysis of total capacity across various auspice types reveals a notable disparity. Specifically, City Operated Auspice exhibits the lowest capacity among the different categories. This implies a distinct need for resource reallocation, suggesting that the government should consider directing more funding towards City Operated Child Centers. By addressing this capacity gap, the government can potentially enhance the effectiveness and reach of childcare

services, promoting a more balanced and equitable distribution of resources within the childcare system.

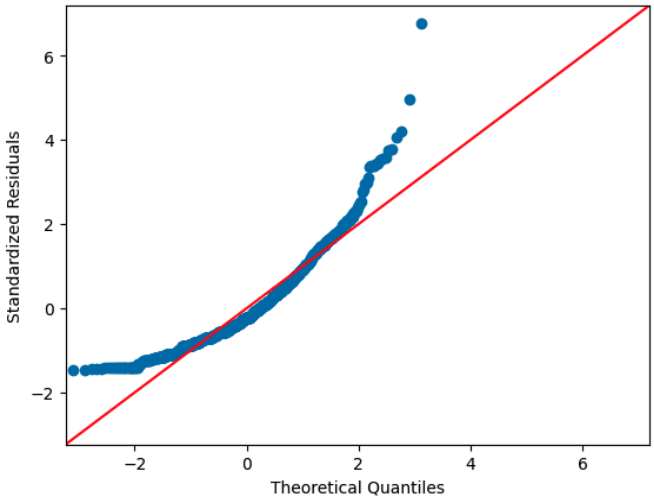

Fig 3 - QQ plot to check Normality

Group1	Group 2	meandiff	p-adj
Commercial	Non Profit	15.32	<0.001
Commercial	Public	-19.02	0.047
Non-Profit	Public	-34.33	0.001

Table 4 - Pairwise Comparison of Auspice (Post-hoc)

5. Two Way Anova of Total Capacity with Auspice & Wards # 3,8, 25

Research Question 2: Do the different types of Auspices and Wards # 3,8, and 25 have a significant difference on the total capacity of childcare centers, and if there is any interaction.

	df	p-value
Auspice	2.0	0.105
Ward	2.0	0.176
Auspice:Ward	4.0	0.709
Residual	135.0	NaN

Table 5 - Anova Table

Table 5 above, shows that the p-values of auspice and wards (3,8, 25) are greater than 0.05 (our significance levels), hence, we fail to reject the null hypothesis, that there is no significant difference of total capacity between the different auspice types located in wards 3, 8, and 25. Additionally, the interaction term (Auspice:Ward) is also greater than 0.05, thus there is not sufficient evidence to claim that there is a statistically significant interaction between auspice and wards 3, 8, 25 (this also follows the theory that if any one of the features is not significant, then the interaction is also not significant).
Figure 4 shows that there is one interaction between wards 3 and 8, however this interaction is not statistically significant, as the values from Table 5 suggest. Furthermore, Shapiro-Wilk and Levene's tests were performed to check for the normality and variance homogeneity assumptions. The Shapiro-Wilk test indicated that normal distribution is not followed as the p-value was less than 0.001 and this was also backed by the QQ plot. Finally, I performed Tukey's post-hoc test. The results showed that none of the pairwise comparisons were statistically significant and hence we failed to reject the null hypothesis that there is no significant difference in the total capacity for the specific wards and auspices. In table 6 below, I have displayed few of the results of the Tukey's test due to space constraints.

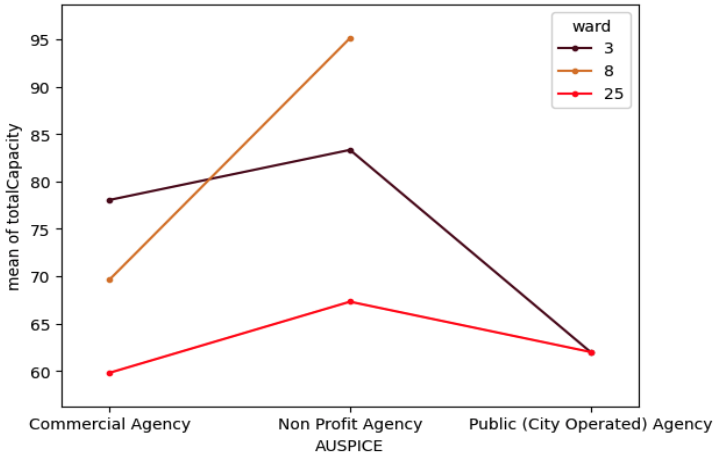

Fig 4 - Interaction Plot of Auspice & Ward

Group1	Group 2	mean diff	p-adj
(Commercial,25)	(Commercial,3)	18.24	0.993
(Commercial,25)	(Commercial,8)	9.83	1.000
(Commercial, 25)	(Non-Profit,8)	35.32	0.7711

Table 6 - Pairwise Comparison of Auspice & Ward 3,8,25

Conclusion: In light of the non-significant differences in total capacity among various auspice types in specific wards, the government might explore fostering collaboration and resource-sharing initiatives among these childcare facilities. This approach could lead to a more streamlined and efficient utilization of resources, potentially improving overall childcare services without favoring specific auspice categories. Additionally, considering a comprehensive assessment of childcare needs and demand across different wards may guide future resource allocation strategies for a more inclusive and responsive childcare system.

6. Two Way Anova of Total Capacity with Auspice & Subsidy

Research Question 3: Do the different types of Auspices and being a subsidized center have a significant difference on the total capacity of childcare centers, and if there is any interaction.

	df	p-value
Auspice	2.0	0.253
Subsidy	1.0	<0.01
Auspice:Subsidy	2.0	<0.01
Residual	1058.0	NaN

Table 7 - Anova Table of Auspice & Subsidy

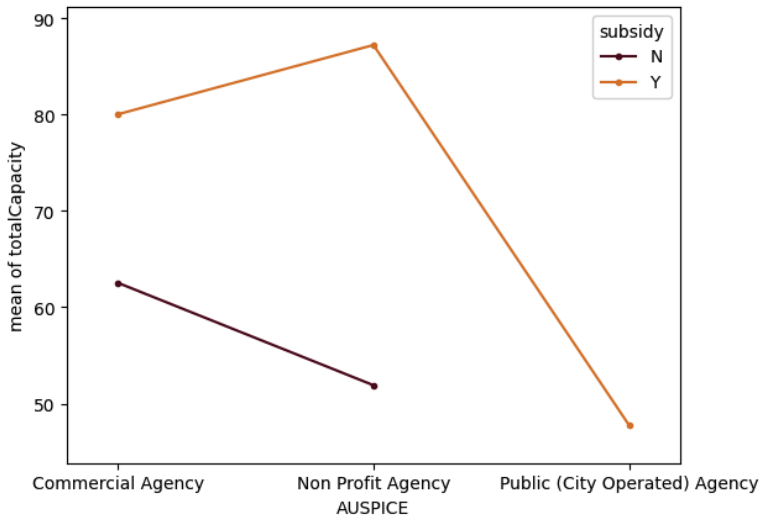

Fig 5 - Interaction Plot of Auspice & Subsidy

Table 7 above, shows that the p-values of auspice is greater than 0.05 (our significance levels) and of subsidy is less than 0.05. Additionally, the interaction term (Auspice:Subsidy) is less than 0.05, however, since one of our feature already has a p-value greater than 0.05, we can claim that the interaction term is not statistically significant, rendering the p-value of the interaction term in this case as meaningless, which is backed by the interaction plot in figure 5. Furthermore, Shapiro-Wilk and Levene's tests were performed to check for the normality and variance homogeneity assumptions. The Shapiro-Wilk test (table 9) indicated that normal distribution is not followed as the p-value was less than 0.001 and this was also backed by the QQ plot. Finally, I performed Tukey's post-hoc test. The results showed that some of the pairwise comparisons were statistically significant and some were not. In table 8 below, I have displayed few of the results of the Tukey's test due to space constraints.

Group1	Group 2	mean diff	p-adj	Reject
(Commercial,N)	(Commercial,Y)	17.50	0.029	True
(Commercial,N)	(Non Profit,N)	-10.63	0.287	False
(Commercial, N)	(Non-Profit,Y)	24.69	<0.001	True

Table 8 - Pairwise Comparison of Auspice & Subsidy (Post hoc)

Test statistic(w)	p-value
0.903	< 0.001

Table 9 - Shapiro Wilk Test Result

Conclusion: Subsidized Commercial Agencies exhibit a significantly higher total capacity compared to their non-subsidized counterparts, suggesting a positive influence of subsidies in this auspice category. However, non-subsidized Commercial Agencies demonstrate a notably lower capacity compared to Non-Profit and Public (City Operated) Agencies. The presence of subsidies significantly increases the capacity of Non-Profit and Public (City Operated) Agencies, emphasizing the effectiveness of subsidies in fostering higher capacities in these auspice categories. In crafting childcare policies, the government could consider targeted subsidy allocations to maximize the overall capacity and efficiency of childcare centers.

7. Conclusion

In conclusion, this comprehensive analysis of Toronto's licensed child care centers provides valuable insights into the city's childcare landscape. Challenges related to the accessibility, affordability, and quality of services persist, emphasizing the need for strategic interventions. The exploration of three key research questions highlighted significant variations in total capacity based on auspice types, suggesting a potential reallocation of resources to address disparities, particularly in City Operated Auspice. Collaborative resource-sharing initiatives among childcare facilities were recommended to optimize efficiency. The examination of subsidies revealed their substantial impact on total capacity, with subsidized Commercial Agencies exhibiting higher capacity. Tailoring subsidy allocations to Non-Profit and Public (City Operated) Agencies could enhance overall childcare services. These findings offer a foundation for evidence-based policymaking and further research to foster a more equitable and effective childcare system in Toronto.

Note: Figures are small throughout the report because of space constraints, please refer to the .ipynb file for a clearer version of the figures.