INF2178 A2 - Technical Assignment
Anshuta Kulkarni

Introduction

The dataset used in this assignment contains information on Toronto-licensed child care centres, including centre details and capacity based on child age groups. The aim of the assignment is to examine the data to study for differences in child care centre capacity across various factors to address low space availability and help families in Toronto access affordable child care.

Data Pre-Processing

To access the variables more efficiently, I renamed the age-based capacity continuous variables to their relevant age groups, such as 'Infants_Cap' (i.e. Infant child care centre capacity), 'Toddlers_Cap' and so on, including the overall capacity as 'Total_Cap'. I also renamed the building type categorical variable to 'Building_Type' for readability. Following this, I converted a few numerical columns to string type to maintain their information, without it affecting the dependent continuous variables.

Exploratory Data Analysis

In order to study the capacity of centers relative to various age groups, I constructed a summary statistics table, replacing '0' values with NaN to avoid skewing descriptive statistics. The results showed centres for Preschoolers, Kindergarteners, and Grade One and above having significant capacity. The distribution of total capacity across centers showed a right skew, indicating that most centers in the data could accommodate between 30 to 50 children. Moreover, the analysis also highlighted that the largest center in the entire dataset in terms of total capacity, considering operational auspice and building type, was a Public Elementary School. Additionally, to study which ward had the most centres and biggest capacity - the data showed ward number 3 having the most number of centres but interestingly, ward number 14 having the highest capacity among all the other wards. Lastly, it was observed that a majority of the centers operated under subsidized fee contracts and were participants in CWELLC, underlining an existing trend towards support programs and financial assistance in the childcare sector in Toronto.

Hypotheses of Statistical Tests used in the assignment -

1. **Tukey-HSD Test (For pairwise comparisons)**
    - **Null Hypothesis -** The means of the tested groups are equal
    - **Alternative Hypothesis -** The means of the tested groups are not equal
2. **Shapiro Wilk Test (For testing normality)**
    - **Null Hypothesis -** Data is drawn from normal distribution
    - **Alternative Hypothesis -** Data is not drawn from normal distribution
3. **Levene's Test (For testing homogeneity of variances)**
    - **Null Hypothesis -** Samples from populations have equal variances
    - **Alternative Hypothesis -** Samples from populations do not have equal variances

One Way ANOVA - I

RQ : Does the age of the child have an effect on the capacity of the centres?

Null Hypothesis (H0) : The centre capacity means are the same across all age groups
Alternate Hypothesis (HA) : At least one centre capacity mean of an age group is different from those of other age groups

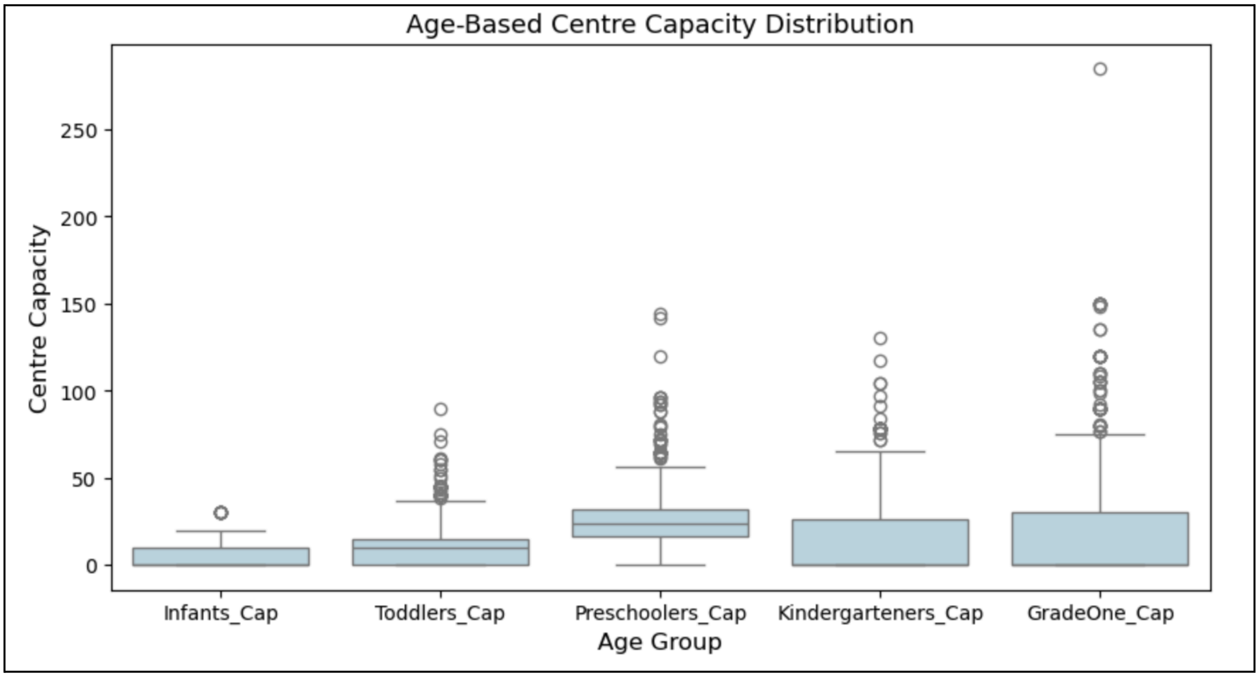


Figure 1. Child care centre capacity distribution based on age group

Results from ANOVA : p-value < 0.001, F-statistic : 188.191
Given the p-value < 0.05, the conclusion is that centre capacity has significant differences across the various age groups, similar to observations from Figure 1.

Post-hoc Test
To assess differences in age-based center capacities, I conducted post-hoc tests utilizing the Tukey-HSD Test. The findings indicated that all comparisons yielded a p-value < 0.05, suggesting significant differences in mean capacity across all pairwise comparisons of age-based centers.

Testing Assumptions

- Assumption 1: Residuals are normally distributed
  - The plots and Shapiro-Wilk Test conclude that the residuals are not normally distributed since the p-value < 0.001, and therefore, we reject the null hypothesis.
- Assumption 2: Homogeneity of Variances
  - Given the data did not pass the normality assumption check, I used Levene's Test, which resulted in a p-value < 0.001, and since p-value < 0.05, we reject the null

hypothesis and conclude that the age-based centres' capacities have unequal variances.

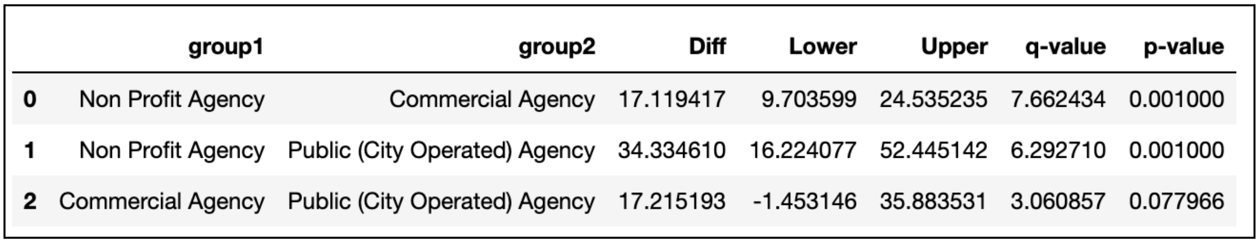
One Way ANOVA - II

RQ : Does the operating auspice have an effect on the total capacity of the centres?

Null Hypothesis (H0) : The centre total capacity means are the same across all auspice groups
Alternate Hypothesis (HA) : At least one centre's total capacity mean of an age group is different from those of other auspice groups

Results from ANOVA : p-value < 0.001, F-statistic : 21.843
With a p-value < 0.05, it can be concluded that there are significant differences in the means of center total capacities across the various operating auspice groups.

Post-hoc Test



	group1	group2	Diff	Lower	Upper	q-value	p-value
0	Non Profit Agency	Commercial Agency	17.119417	9.703599	24.535235	7.662434	0.001000
1	Non Profit Agency	Public (City Operated) Agency	34.334610	16.224077	52.445142	6.292710	0.001000
2	Commercial Agency	Public (City Operated) Agency	17.215193	-1.453146	35.883531	3.060857	0.077966

Table 1. Tukey-HSD Test Results for significant differences in capacity means of auspice groups

Utsing the Tukey-HSD Test as the post-hoc analysis to examine differences in capacity means across auspice groups, results indicated significant differences in total capacity means for only two comparisons: Non Profit Agency - Commercial Agency, and Non Profit Agency - Public (City Operated) Agency. This observation was also evident in the plotted distribution of capacity for each auspice group, where the Non Profit Agency exhibited the highest mean total capacity.

Testing Assumptions

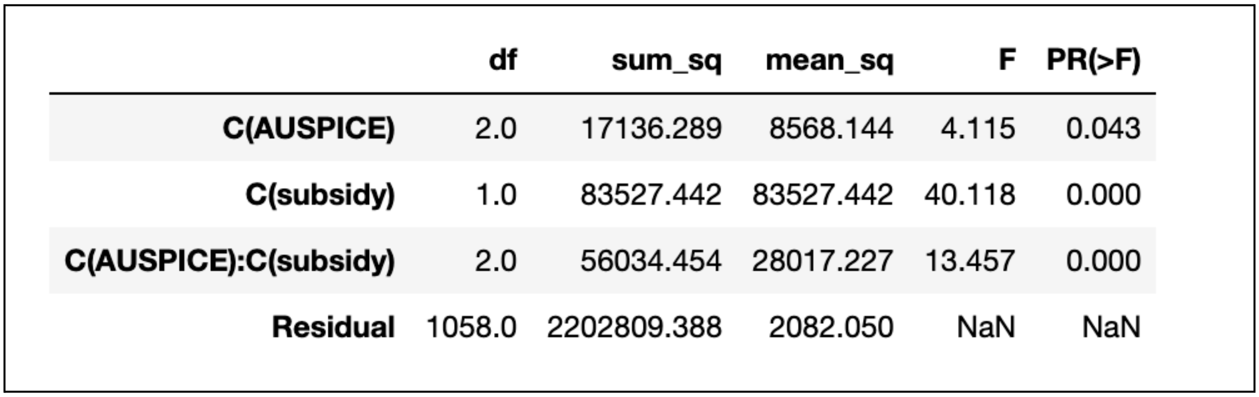
- Assumption 1: Residuals are normally distributed
    - The plots and Shapiro-Wilk Test conclude that the residuals are not normally distributed since the p-value < 0.001, and therefore, we reject the null hypothesis as the p-value < 0.05.
- Assumption 2: Homogeneity of Variances
    - Given the data did not pass the normality assumption check, I used Levene’s Test, which resulted in a p-value < 0.001, and since p-value < 0.05, we reject the null hypothesis and conclude that the centres' capacities of different operating auspice groups have unequal variances.

Two-Way ANOVA - I

RQ : Does the operating auspice and subsidy status of a fee contract have an effect on the total capacity of the centres?

1. **Null Hypothesis (H0) : There is no effect of operating auspice group on the centre total capacity**
   Alternate Hypothesis (HA) : There is a significant effect of operating auspice group on the centre total capacity
2. **Null Hypothesis (H0) : There is no effect of subsidy status on the centre total capacity**
   Alternate Hypothesis (HA) : There is a significant effect of subsidy status on the centre total capacity
3. **Null Hypothesis (H0) : There is no effect of the interaction of operating auspice group and subsidy status on the centre total capacity**
   Alternate Hypothesis (HA) : There is a significant effect of the interaction of operating auspice group and subsidy status on the centre total capacity

Results from ANOVA :



	df	sum_sq	mean_sq	F	PR(>F)
C(AUSPICE)	2.0	17136.289	8568.144	4.115	0.043
C(subsidy)	1.0	83527.442	83527.442	40.118	0.000
C(AUSPICE):C(subsidy)	2.0	56034.454	28017.227	13.457	0.000
Residual	1058.0	2202809.388	2082.050	NaN	NaN

Table 2. Two-Way ANOVA Results for effect on corresponding factors on total capacity

Main Effect Auspice - p-value = 0.043, F-statistic = 4.115. Since the p-value < 0.05, we reject the null hypothesis (1 - H0), and conclude that the operating auspice group does have a significant effect on the centre's total capacity.
Main Effect Subsidy - p-value < 0.001, F-statistic = 40.118. Since the p-value < 0.05, we reject the null hypothesis (2 - H0), and conclude that the subsidy status group does have a significant effect on the centre's total capacity.
Interaction Effect Auspice & Subsidy - p-value < 0.001, F-statistic = 13.457. Since the p-value < 0.05, we reject the null hypothesis (3 - H0), and conclude that the interaction of operating auspice and subsidy status group does have a significant effect on the centre's total capacity.

Interaction Plot

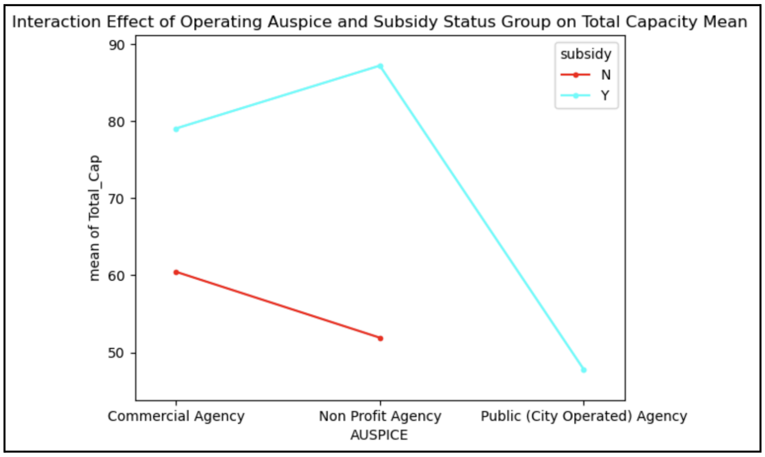


Figure 2. Child care centre capacity distribution based on age group

As seen in Figure 2, the interaction plot shows a significant interaction effect between auspice and subsidy groups, confirming the two-way ANOVA results.

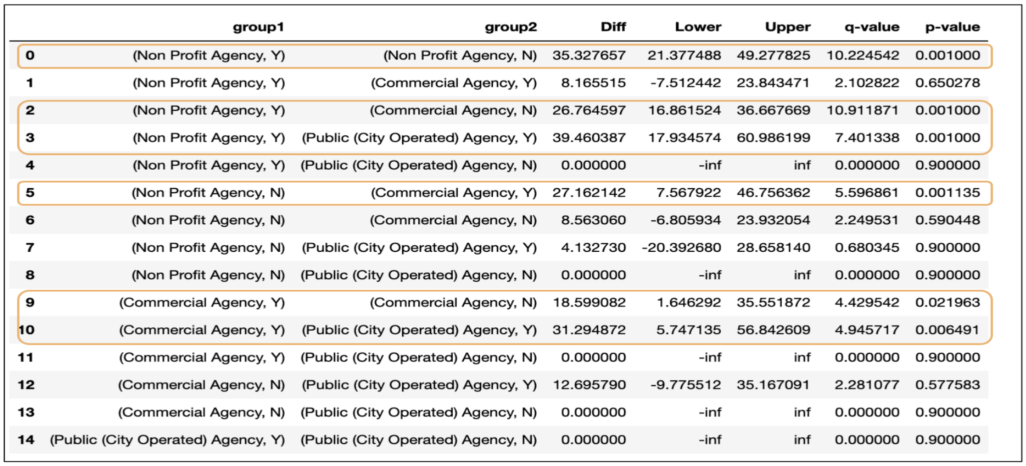
Post-hoc Test

I conducted three sets of post-hoc tests using the Tukey-HSD test to analyze the main effects and interaction effect on the centre's total capacity.

For the main effect of operating auspice, consistent with Table 1 findings, significant differences in capacity means were observed for only two comparisons: Non Profit Agency - Commercial Agency, and Non Profit Agency - Public (City Operated) Agency.

For the main effect of subsidy status, with a p-value = 0.001, we conclude that there are significant differences in the centre’s total capacity means between the two groups.

For the interaction effect of operating auspice and subsidy status group, the results were the following -



	group1	group2	Diff	Lower	Upper	q-value	p-value
0	(Non Profit Agency, Y)	(Non Profit Agency, N)	35.327657	21.377488	49.277825	10.224542	0.001000
1	(Non Profit Agency, Y)	(Commercial Agency, Y)	8.165515	-7.512442	23.843471	2.102822	0.650278
2	(Non Profit Agency, Y)	(Commercial Agency, N)	26.764597	16.861524	36.667669	10.911871	0.001000
3	(Non Profit Agency, Y)	(Public (City Operated) Agency, Y)	39.460387	17.934574	60.986199	7.401338	0.001000
4	(Non Profit Agency, Y)	(Public (City Operated) Agency, N)	0.000000	-inf	inf	0.000000	0.900000
5	(Non Profit Agency, N)	(Commercial Agency, Y)	27.162142	7.567922	46.756362	5.596861	0.001135
6	(Non Profit Agency, N)	(Commercial Agency, N)	8.563060	-6.805934	23.932054	2.249531	0.590448
7	(Non Profit Agency, N)	(Public (City Operated) Agency, Y)	4.132730	-20.392680	28.658140	0.680345	0.900000
8	(Non Profit Agency, N)	(Public (City Operated) Agency, N)	0.000000	-inf	inf	0.000000	0.900000
9	(Commercial Agency, Y)	(Commercial Agency, N)	18.599082	1.646292	35.551872	4.429542	0.021963
10	(Commercial Agency, Y)	(Public (City Operated) Agency, Y)	31.294872	5.747135	56.842609	4.945717	0.006491
11	(Commercial Agency, Y)	(Public (City Operated) Agency, N)	0.000000	-inf	inf	0.000000	0.900000
12	(Commercial Agency, N)	(Public (City Operated) Agency, Y)	12.695790	-9.775512	35.167091	2.281077	0.577583
13	(Commercial Agency, N)	(Public (City Operated) Agency, N)	0.000000	-inf	inf	0.000000	0.900000
14	(Public (City Operated) Agency, Y)	(Public (City Operated) Agency, N)	0.000000	-inf	inf	0.000000	0.900000

Table 3. Tukey-HSD Test Results for significant differences in capacity means of different pairwise comparisons of auspice and subsidy status groups

As shown in Table 3, the pairwise comparisons showing significant differences in total capacity means are highlighted in orange. With p-values < 0.05, we reject the null hypothesis, concluding significant differences in total capacity means between these pairwise groups. The findings imply that child care centers primarily operated by Non Profit Agencies or Commercial Agencies, particularly those with predominantly subsidized fee contracts, significantly impact total capacity.

Testing Assumptions

- Assumption 1: Residuals are normally distributed
  - The plots and Shapiro-Wilk Test conclude that the residuals are not normally distributed since the p-value < 0.001, and therefore, we reject the null hypothesis.
- Assumption 2: Homogeneity of Variances
  - Given the data did not pass the normality assumption check, I used Levene's Test, which resulted in a p-value of NaN, rendering the test inconclusive.

Conclusion

The results suggest that a majority of the capacity centres having higher capacity are those operated by Non-Profit agencies having subsidized fee contracts, with varying capacity means across age groups. Given these findings, it would help to further study this issue by tracking daily occupancy in these centres, to compare the values across time to help predict the upcoming usage and effectively invest the funds provided by the provincial government.

References

- Python Graph Gallery - https://python-graph-gallery.com
- Stack Overflow - https://stackoverflow.com
- Seaborn Documentation - https://seaborn.pydata.org/archive/0.11/tutorial/categorical.html
- Pandas DF Documentation - https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html
- How to Perform ANOVA in Python, Renesh Bendre, https://www.reneshbedre.com/blog/anova.html
- Displaying Multiple DataFrames Side By Side in Jupyter Lab/Notebook, Liu Zuo Lin, https://python.plainenglish.io/displaying-multiple-dataframes-side-by-side-in-jupyter-lab-notebook-9a4649a4940
- Introduction to ANOVA for Statistics and Data Science (with COVID-19 Case Study using Python), https://www.analyticsvidhya.com/blog/2020/06/introduction-anova-statistics-data-science-covid-python/