Sheng Zhang
eily.zhang@mail.utoronto.ca

An Investigation of the Child Care Service in Ontario

1. Introduction

Childcare accessibility remains a prominent concern in Ontario due to high costs and limited availability. This report investigates the factors impacting child care services in the region, focusing on locational and categorical components of child care centres. Based on the investigation of the dataset "INF2178_A2_data.xlsx," the report employs exploratory data analysis (EDA) to address the following research questions:

1. Does the auspice type of the child care centre (non profit agency, commercial agency, public agency) have a significant effect on the total capacity number of the child care centre?
2. Does the ward code of the child care centre (1, 2, …, 25) have a significant effect on the total capacity number of the child care centre?
3. How does both the auspice type and the ward code of the child care centre affect the total capacity number of the child care centre?

With these research questions in mind, we can start our data analysis.

2. Data Cleaning and Data Wrangling

The initial dataset has 17 columns and 1063 rows. By checking the data and the data dictionary, it becomes evident that certain columns lack relevance for our data analysis objectives. Thus, we will focus on the following key columns to initiate our analysis:

- AUSPICE: operating auspice (Commercial, Non Profit, or Public)
- ward: city ward number (From 1 to 25)
- IGSPACE: child care spaces for infants 0-18 months
- TGSPACE: child care spaces for toddlers 18-30 months
- PGSPACE: child care spaces for preschoolers from 30 months to grade one
- KGSPACE: child care spaces for children in full-day kindergarten
- SGSPACE: child care spaces for children grade one and up
- TOTSPACE: child care spaces for all age groups

To streamline our analysis and focus on the total capacity number, we'll introduce a new variable:

- Total_Capacity_Number: the total child care space (defined by the sum of IGSPACE, TGSPACE, PGSPACE, KGSPACE, SGSPACE, TOTSPACE)

Additionally, for clarity and consistency, we'll rename the 'AUSPICE' column to 'Auspice_Type' and the 'ward' column to 'Ward_Code'. Consequently, the columns for our exploratory data analysis in the subsequent stage will be: Auspice_Type, Ward_Code, and Total_Capacity_Number.

3. Exploratory Data Analysis(EDA)

First we construct a summary table to check the statistics of the dataset [Figure 1].

Summary Statistics for Total Capacity Number	count	mean	std	min	25%	50%	75%	max	Total_Capacity_Number
Total_Capacity_Number	1063	151.3490122	95.63303567	12	86	124	194	804	9145.677511

Figure 1: Summary statistics for the variable total capacity number in the dataset

Also, because we care about two variables that might influence the total capacity number, auspice type and ward code, we plot the bar graphs for the total capacity number across auspice types (for one-way ANOVA) [Figure 2] and ward codes (for one-way ANOVA) [Figure 3].

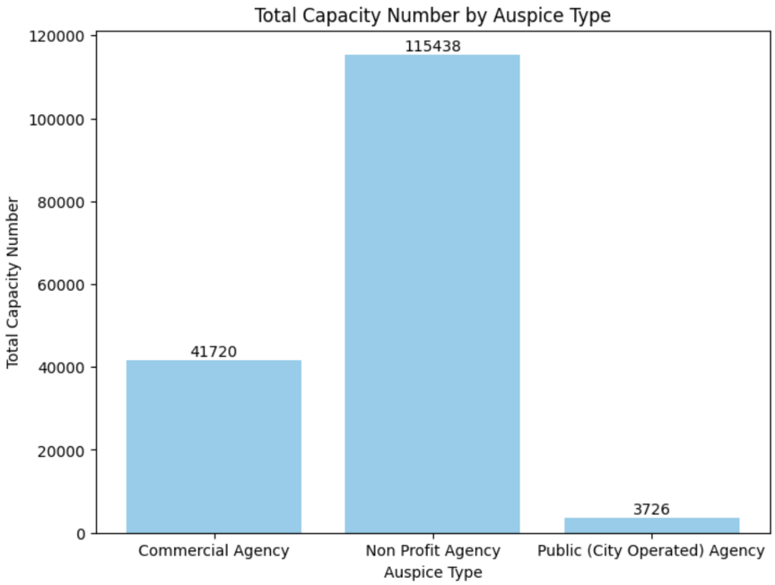
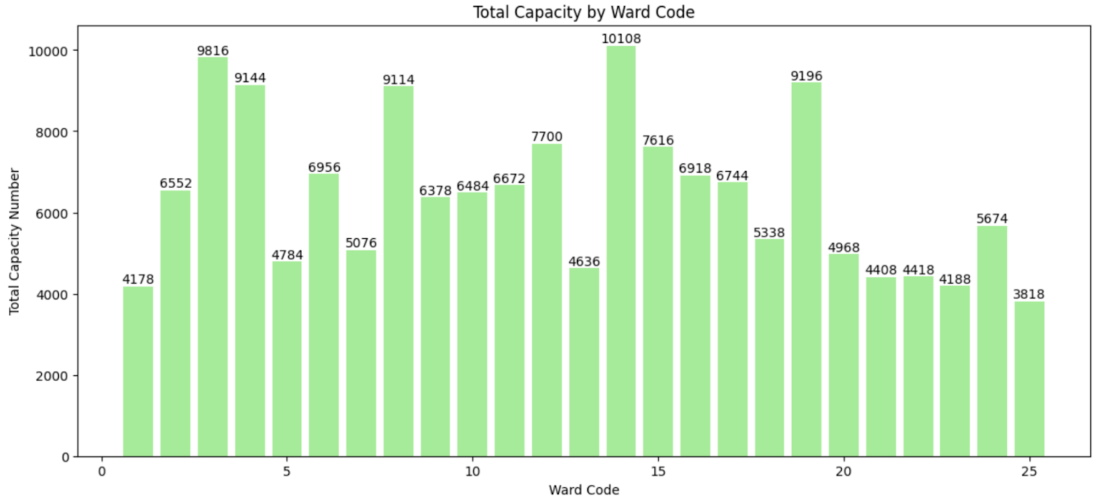


Figure 2(left):
Bar graph for the total capacity number across auspice types.

Figure 3(down):
Bar graph for the total capacity number across ward codes



The bar graphs show that non-profit agencies typically offer more capacity spaces than the other two types. Certain ward codes, like 3 and 14, also tend to have more capacity spaces compared to others. To further analyse, we'll create box plots to examine the distribution of total capacity numbers across different variables.

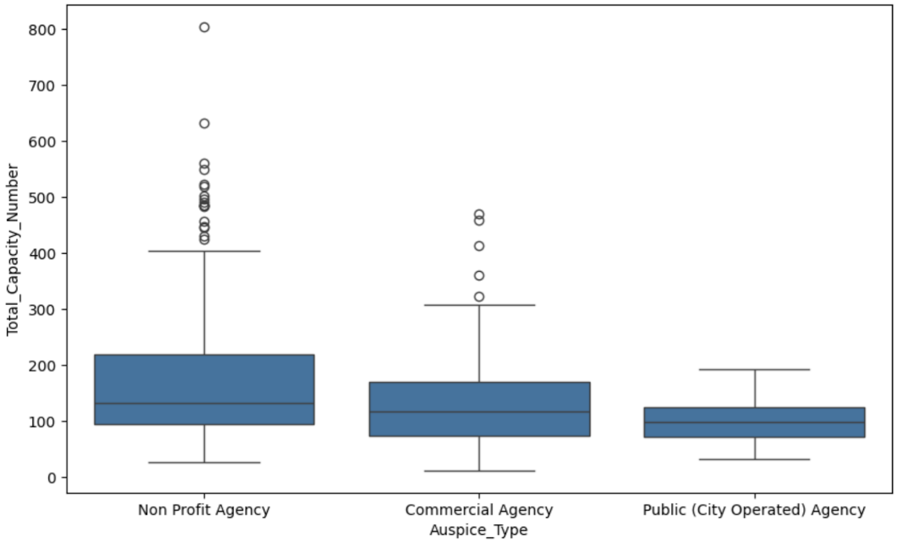


Figure 4:
Box plot for total capacity number between each auspice type

We'll include box plots for total capacity numbers by auspice type (one-way ANOVA) [Figure 4], by ward code (one-way ANOVA) [Figure 5], and by auspice type and ward code (two-way ANOVA) [Figure 6]. Note that outliers are represented by dots and the median is indicated by the solid line within the boxes.

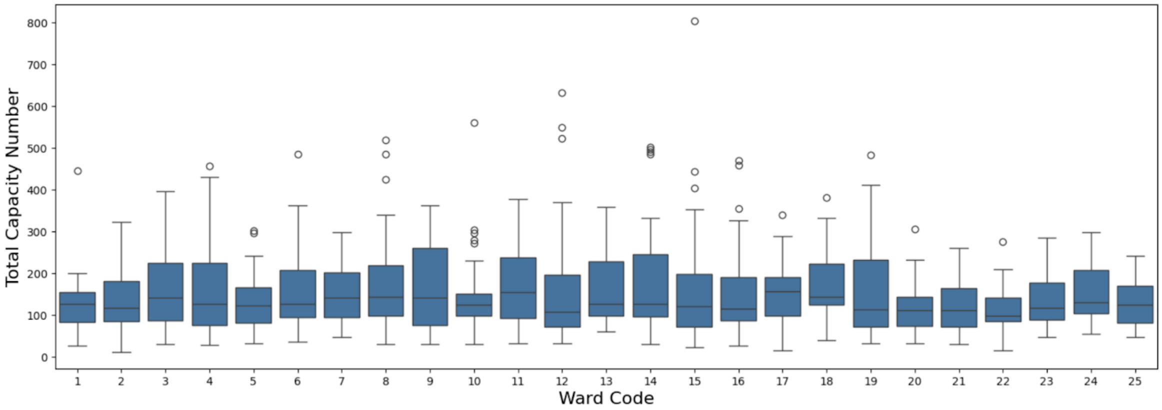


Figure 5:
Box plot for total capacity number between each ward code

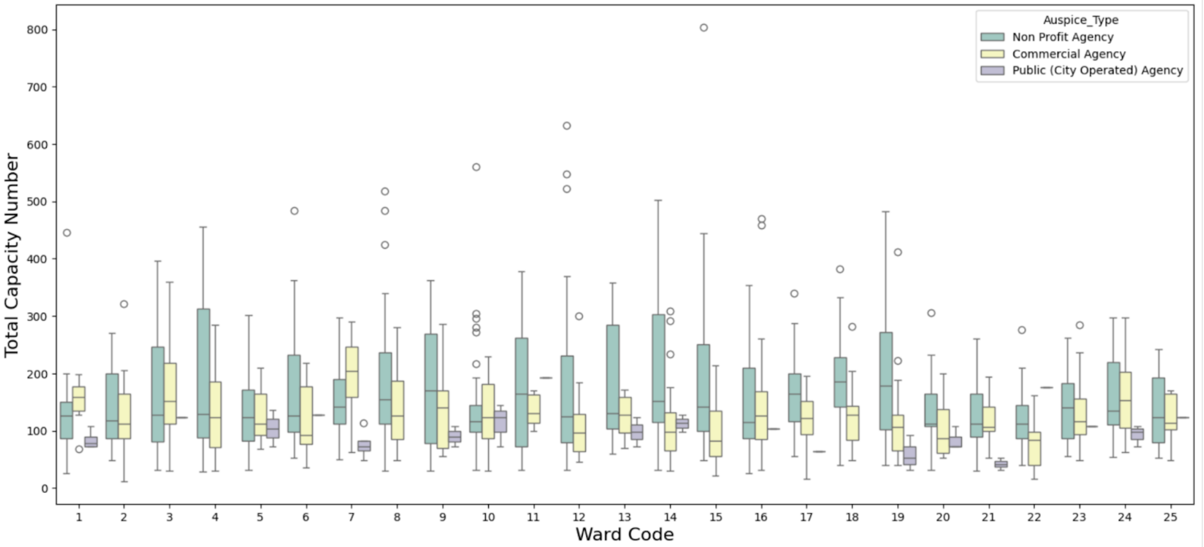


Figure 6:
Box plot for total capacity number between each ward code AND each auspice type

In Figure 4, the boxes representing auspice types show similar shapes, with the non-profit agency's box being slightly different. This suggests there may be something of interest, but we need to implement ANOVA to confirm. In Figures 5 and 6, while some boxes resemble each other, most display significant differences. This indicates that certain groups may not contribute meaningfully or could strongly influence the results. Further analysis is needed to draw conclusions.

Now, we're prepared for deeper analysis. With each variable having at least three groups, we'll employ one-way and two-way ANOVA to investigate how different variables influence the continuous variable of total capacity number. All data are randomly drawn and independent, meeting the assumption of independence. Our significance level, or alpha-level, for all analyses will be set at 0.001. Detailed ANOVA analyses for each research question and their assumptions will be covered in subsequent sections.

4. Total Capacity Number across Auspice Type

Research Question 1: Does the auspice type of the child care centre have a significant effect on the total capacity number of the child care centre?

Since we have only one independent variable (the auspice type) with at least three groups, we'll utilise one-way ANOVA for this research question. The null hypothesis posits that the auspice type of the child care centre has no impact on the total capacity number. As previously mentioned, the chosen alpha-level is 0.001. Thus, we'll present the table of the one-way ANOVA in Figure 7.

	df	sum_sq	mean_sq	F	PR(>F)
C(Auspice_Type)	2	384448.457	192224.229	21.843	<0.001
Residual	1060	9328261.059	8800.246	NaN	nan

Figure 7: One-way ANOVA table for auspice type

The ANOVA p-value is below our chosen significance level of 0.001, leading us to reject the null hypothesis. This means that the auspice type significantly impacts total capacity. To pinpoint which groups differ, we use Tukey's HSD post-hoc test. See Figure 8 for the results.

group1	group2	Diff	Lower	Upper	q-value	p-value
Non Profit Agency	Commercial Agency	34.239	19.407	49.07	7.662	0.001
Non Profit Agency	Public Agency	68.669	32.448	104.89	6.293	0.001
Commercial Agency	Public Agency	34.43	-2.906	71.767	3.061	0.078

Figure 8: Post-hoc table for auspice type

From the table above, we conclude that the total capacity number for non-profit agencies significantly differs from both commercial and public agencies, while the total capacity number for commercial agencies does not differ from that of public agencies.

Next, we'll examine the assumptions for one-way ANOVA. The independence assumption was met during data collection. However, we need to check the assumptions for normality and homogeneity of variance:

1. Assumption 1: residuals are normally distributed
   We perform the Shapiro-Wilk test and obtain a p-value of <0.001, indicating deviation from normality. Thus, the assumption of normality is not met.
2. Assumption 2: homogeneity of variance
   As the normality assumption is not met, we use Levene's test and obtain a p-value of <0.001, suggesting unequal variance. Therefore, the assumption of homogeneity of variance is also not met.

As all assumptions of one-way ANOVA, except for independence, are not met, the previous ANOVA results might be inaccurate or misleading. To improve accuracy, we could consider alternative tests like Welch or Forsythe tests, or use another alpha-level. Additionally, comparing other factors may help avoid violations of assumptions in future analyses.

5. Total Capacity Number across Ward Code

Research Question 2: Does the ward code of the child care centre (1, 2, …, 25) have a significant effect on the total capacity number of the child care centre?

Similar to the previous analysis, we'll employ one-way ANOVA since there's only one independent variable (ward code) with at least three groups. The null hypothesis posits that the ward code of the child care centre has no impact on the total capacity number, with the alpha-level set at 0.001. The ANOVA table is presented in Figure 9 below.

	df	sum_sq	mean_sq	F	PR(>F)
C(Ward_Code)	24	314392.256	13099.677	1.447	0.0759
Residual	1038	9398317.26	9054.256	NaN	nan

Figure 9: One-way ANOVA table for ward code

The ANOVA p-value exceeds our chosen significance level of 0.001, leading us to fail to reject the null hypothesis. This suggests that the auspice type doesn't significantly affect total capacity number. While a post-hoc test isn't necessary, we'll conduct it as a precaution. We'll use Tukey's HSD test, and the results are presented in Figure 10 below. Due to the large number of comparisons for a variable with 25 groups, the table only displays results where the p-value for the post-hoc test is less than 0.001. If any two groups have a p-value below 0.001, indicating a significant difference, we wouldn't expect them to appear here since we failed to reject the null hypothesis for ANOVA.

group1	group2	Diff	Lower	Upper	q-value	p-value

Figure 10: Post-hoc table for ward code

The table remains empty when applying the condition for a p-value less than 0.001, indicating no significant differences between any two groups. This aligns with our conclusion from the ANOVA table.

Moving forward, let's assess the assumptions for one-way ANOVA. The independence assumption holds from the data collection phase. However, we need to check for normality and homogeneity of variance:

1. Assumption 1: residuals are normally distributed
   The Shapiro-Wilk test yields a p-value of <0.001, indicating deviation from normality. Therefore, the assumption of normality is not met.
2. Assumption 2: homogeneity of variance
   As normality is not met, we use Levene's test and obtain a p-value of <0.001, suggesting unequal variance. Thus, the assumption of homogeneity of variance is also not met.

As all assumptions of one-way ANOVA, except for independence, are not met, the previous ANOVA results might be inaccurate or misleading. To improve accuracy, we could consider alternative tests like Welch or Forsythe tests, or use another alpha-level. Additionally, comparing other factors may help avoid violations of assumptions in future analyses.

6. Total Capacity Number across Auspice Type and Ward Code

Research Question 3: How does both the auspice type and the ward code of the child care centre affect the total capacity number of the child care centre?

Although we will use two-way ANOVA here as there are two independent variables (auspice type and ward code) and at least three groups, the general strategy is also similar to the one-way ANOVA in the previous sections. However, in two-way ANOVA, we have three null hypotheses simultaneously (with the alpha-level 0.001):

1. There is no significant difference in the total capacity number of child care centres across different auspice types.
2. There is no significant difference in the total capacity number of child care centres across different ward codes.
3. There is no interaction effect between auspice type and ward code on the total capacity number of child care centres.

We've encountered a familiar situation. Based on our past one-way ANOVA analyses, it seems likely we'll reject the first null hypothesis and not reject the second one. But to be sure, we need to conduct a two-way ANOVA, which is presented in Figure 11.

index	df	sum_sq	mean_sq	F	PR(>F)
C(Auspice_Type)	2	396596.959	198298.479	22.624	<0.001
C(Ward_Code)	24	363329.411	15138.725	1.727	0.0299
C(Auspice_Type):C(Ward_Code)	48	331555.485	6907.406	0.788	0.832
Residual	994	8712393.625	8764.984	NaN	nan

Figure 11: Two-way ANOVA table for auspice type and ward code

We find that the auspice type has a p-value less than 0.001, leading us to reject the first null hypothesis, indicating a significant difference in total capacity numbers. However, both ward code and the combination of auspice type and ward code show p-values greater than the alpha-level 0.001. Therefore, we fail to reject both the second and third null hypotheses. This means there's no significant difference in total capacity numbers across different ward codes, nor is there an interaction effect between auspice type and ward code on total capacity numbers in child care centres.

To determine which specific groups differ, we apply Tukey's HSD post-hoc test, the results of which can be seen in Figure 12 below. Due to the large number of comparisons for a variable with 25 groups and a variable with 3 groups, we've constrained the table to show only combinations that reject the null hypothesis in the post-hoc test.

group1	group2	meandiff	p-adj	lower	upper	reject

Figure 12: Post-hoc table for auspice type and ward code

We've encountered an interesting finding. Although we've established a significant difference in total capacity numbers, there's nothing remarkable when we look closer. This implies that while there's evidence indicating that not all group means are equal, the discrepancies between specific groups aren't significant enough to be detected by Tukey's HSD test. This might be because of smaller sample sizes or weaker effects between specific group pairs, which may not reach significance after adjusting for multiple comparisons.

Next we will check the interaction by making an interaction plot as shown in Figure 13.

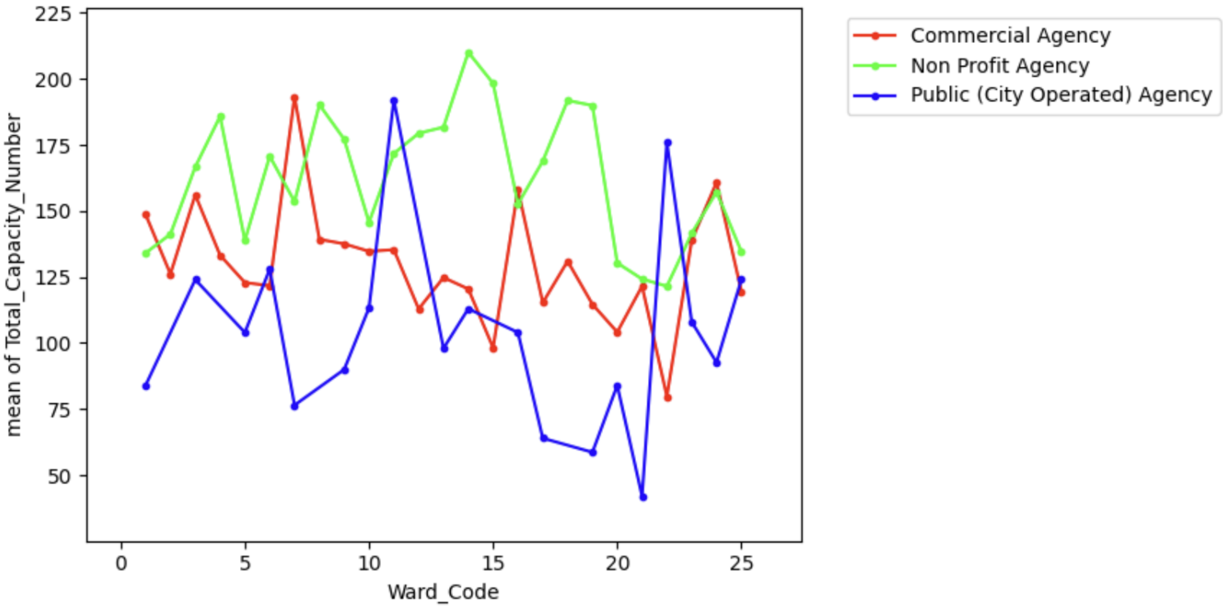


Figure 13: Interaction plot for auspice type and ward code

There is another strange observation here, the plot shows that there is interaction effect between any two groups but this contradicts with our third null hypothesis that we fail to reject. So we will have to check the assumptions. We have:

1. Assumption 1: residuals are normally distributed
   We perform the Shapiro-Wilk test and find a p-value of <0.001, indicating a deviation from normal distribution.
2. Assumption 2: homogeneity of variance
   Using Levene's test, we find p-values of <0.001 for most group combinations, suggesting unequal variances.

As all assumptions of one-way ANOVA, except for independence, are not met, the previous ANOVA results might be inaccurate or misleading. To improve accuracy, we could consider alternative tests like Welch or Forsythe tests, or use another alpha-level. Additionally, comparing other factors may help avoid violations of assumptions in future analyses.

7. Conclusion

After the two one-way ANOVA and one two-way ANOVA analyses, even though the assumptions for the ANOVA are not met, we could still make rough predictions based on what we have, that auspice type is highly possible to be a factor that influences the total capacity type. So in the future, the government might consider this as a component in the construction of the child care centres and parents can also choose a child care centre based on this factor.