Samantha Chui (1002369126) INF2178 Winter 2024

A2: Exploring Licensed Child Care Centres in Toronto

1. Introduction

This report offers an analysis of all licensed child care centres in Toronto using a dataset from the Toronto Open Data Catalogue titled '*INF2178_A2_data.xlsx*', dated February 2024. The goal of the analysis is to identify certain trends in how a type/category of centre operates. This information can be used to inform parents which child care centre they may wish to enroll their child in based on their particular needs.

2. Data Pre-Processing

The dataset consists of 16 columns and 1063 rows of data, with no null values. The data is either type int or str (object). Based on the data dictionary provided, we renamed columns as necessary into more meaningful phrases. I decided to distinguish categorical variables from data variables by the column type, where only columns containing data will be type int. This means that columns ID and Ward were transformed into type str. This is based on personal preference to easily group identify all categorical columns during analysis.

In order to identify potential variables for analysis, I wanted to first identify how many unique values are in each category. I created a function unique_categories which takes a dataframe, extracts only columns of type str (as a way to distinguish a category), and calculate the number of unique values of that column. The function returns a table of each column that was identified as a category and the respective number of unique values.

3. One-Way ANOVA

Research Question: Does child care capacity differ by age group?

In a one-way ANOVA, multiple groups are compared to one another to see whether they are significantly different. I created a dataframe age_group which contains the 5 age group columns: Infants, Toddlers, Preschool, Kindergarten, and Grade_School. Total_Spaces is excluded since it is the aggregated total of these age groups, and I would like to focus on capacity of each service/age group offered by child care centres in Toronto. The dataframe needs to be further cleaned as the 0 values left as is will skew the distribution of the data. Instead, it should be converted to NaN. The function replace_nan takes a dataframe and replaces each cell value of 0 with NaN, and returns the modified dataframe. This was applied to age_group.

As part of exploratory data analysis (EDA), each age group should be assessed for normal distribution and equal variances before computing a one-way ANOVA. The function summary_table was created to take a dataframe and compute the mean, median, min and max value, and IQR of each column, while ignoring NA values. These statistics are returned as a table. Applying this function to age_group gives the following output:

Statistic	**Infants**	**Toddlers**	**Preschool**	**Kindergarten**	**Grade_School**
Count	359.000	670.000	854.000	461.000	505.000
Mean	11.538	18.404	30.196	32.876	45.596
Median	10.000	15.000	24.000	26.000	30.000
Min_Value	4.000	2.000	1.000	5.000	6.000

Max_Value	30.000	90.000	144.000	130.000	285.000
Quartile_1	10.000	10.000	20.000	24.000	30.000
Quartile_3	10.000	20.000	38.000	48.000	60.000

Table 1: Summary statistics of child care capacity by age group.

From the summary statistics, we can see that capacity generally increases as we move to older age groups. For example, the median capacity of Grade_School is triple the value of that for Infants. This suggests that the age groups may not have equal variance. A histogram of each age group can provide further insight into the distributions.

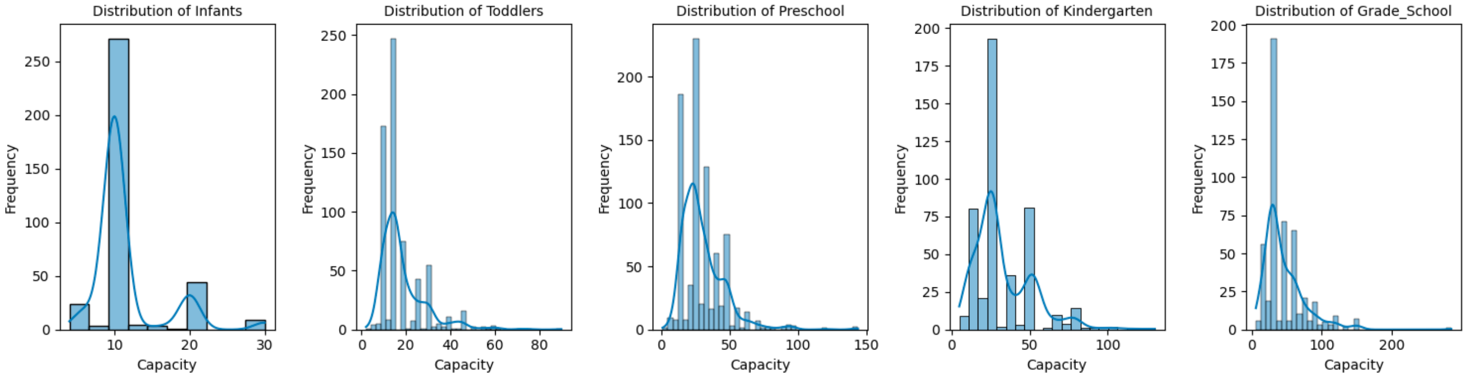

Figure 1: Distribution of child care capacity by age group.

The histograms show that each age group is right-skewed, although the intensity of and number of peaks differs between each group. In summary, the visualization suggests that all age groups do not follow a normal distribution. However, we can apply Central Limit Theorem (CLT) and still assume a normal distribution due to a sufficiently large sample size.

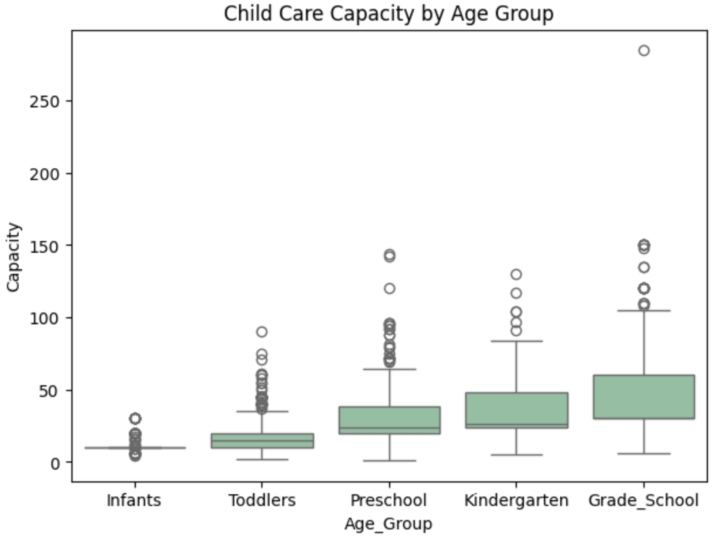

Figure 2: Variability of child care capacity by age group.

The boxplot is consistent with what we saw from the summary table: capacity increases as we move to older age groups. This suggests that the number of spaces offered varies much greater for older age groups. It also shows that there are a lot of outliers in each age group.

	df	sum_sq	mean_sq	F	PR(>F)
C(Age_Group)	4.0	330261.920	82565.480	264.388	< 0.001
Residual	2844.0	888150.516	312.289	NaN	NaN

Table 2: One-way ANOVA results for child care capacity across age group levels.

A one-way ANOVA was computed using age group as a treatment level. Based on a significance level of α = 0.05, the results indicated a p-value of less than 0.001. This suggests that there is sufficient evidence to reject the null hypothesis that child care capacity does not differ by age group.

group1	group2	Diff	Lower	Upper	q-value	p-value
Infants	Toddlers	6.867	3.712	10.022	8.402	0.001
Infants	Preschool	18.658	15.624	21.692	23.738	0.001
Infants	Kindergarten	21.339	17.943	24.734	24.260	0.001
Infants	Grade_School	34.058	30.729	37.388	39.482	0.001
Toddlers	Preschool	11.791	9.302	14.280	18.284	0.001
Toddlers	Kindergarten	14.472	11.553	17.391	19.139	0.001
Toddlers	Grade_School	27.192	24.349	30.034	36.926	0.001
Preschool	Kindergarten	2.681	< 0.001	5.469	3.712	0.066
Preschool	Grade_School	15.400	12.693	18.108	21.955	0.001
Kindergarten	Grade_School	12.720	9.613	15.827	15.802	0.001

Table 3: Post-hoc results using Tukey's HSD to test effect of age group levels on child care capacity.

Post-hoc testing was done to compare each treatment level. The results of Tukey's HSD reveals that the majority of age groups differ from each other at a statistically significant level (where p < 0.05). The only exception is preschool and kindergarten, where the p-value was greater than 0.05.

Testing ANOVA Assumptions

Assumption 1: The residuals follow a normal distribution

This assumption will be tested using both visualization methods and statistical tests.

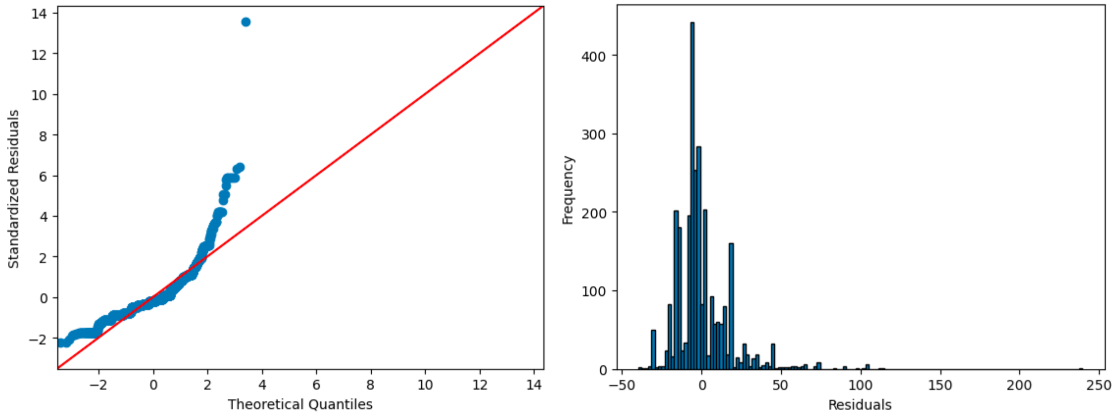

Figure 3: Distribution of residuals as a QQ plot (left) and histogram (right) for one-way ANOVA.

The QQ plot shows a curve upwards rather than a diagonal line. There are many points both above and below the reference line. This suggests that the model is not well-fitted. The

histogram also shows a right skew. Overall, the plots of the residuals suggest that it does not follow a normal distribution.
The Shapiro-Wilk test is used as a test for normality, where the null hypothesis is that the data follows a normal distribution. Our test yielded a p-value of less than 0.001, meaning we reject the null hypothesis: the residuals do not follow a normal distribution.

Assumption 2: The levels have an equal variance
Levene's Test is an appropriate statistical test for equal variance when the sample does not follow a normal distribution. The null hypothesis is that the treatment levels have equal variance. The test yielded a p-value of less than 0.001, meaning we reject the null hypothesis: the treatment levels do not have equal variance.

In conclusion, although our results of one-way ANOVA suggest that the capacity of child care centres differs according to the age group, these results must be interpreted with caution as the assumptions for testing have not been met. It is highly suggested to use another statistical test where equal variance is not required, such as Welch's ANOVA.

4. Two-Way ANOVA

Research Question: Does total capacity of a child care centre differ by ward and subsidy?
In a two-way ANOVA, there are two treatments/independent variables; in this context, we have chosen to examine Ward and Subsidy against Total_Space. We are also testing for an interaction effect between the two treatment levels. I created a dataframe ward_subsidy with these 3 columns, then replaced the values in column Subsidy to "Available" or "Unavailable" so that it is more meaningful. I created a separate dataframe to aggregate the data by counts, sorted by ward in ascending order. The dataframe ward_subsidy_2 is for EDA only, as it allows me to check that there are an equal number of treatment levels for each variable: for each of the 25 wards, there are centres with or without subsidy.

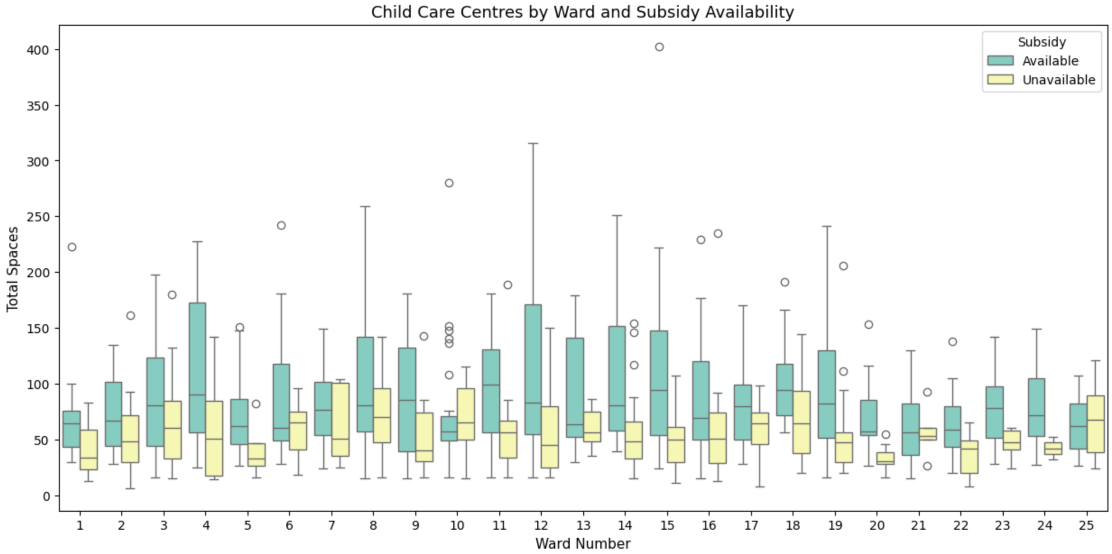

Figure 4: Variability of total child care capacity by ward levels and subsidy levels.

The boxplot of the data shows that centres with subsidy tend to have larger capacity compared to centres without subsidy. It also shows that the range varies as we move across wards.

	df	**sum_sq**	**mean_sq**	**F**	**PR(>F)**
C(Ward)	24.0	145685.344	6070.223	2.976	< 0.001
C(Subsidy)	1.0	227852.673	227852.673	111.716	< 0.001
C(Ward):C(Subsidy)	24.0	55644.003	2318.500	1.137	0.295
Residual	1013.0	2066082.638	2039.568	NaN	NaN

Figure 5: Two-way ANOVA results for total child care capacity across ward levels and subsidy levels.

A two-way ANOVA was computed using a significance level of α = 0.05. For total space by ward, the p-value of less than 0.001 suggests that we reject the null hypothesis: total capacity does differ by ward. For total space by subsidy, the p-value of less than 0.001 suggests that we reject the null hypothesis: total capacity does differ by subsidy. Lastly, for the interaction between ward and subsidy, the p-value greater than 0.05 suggests that we do not have sufficient evidence to reject the null hypothesis: total capacity by ward does not dependent on total capacity by availability of subsidy. To further study the effect of the treatments, 3 rounds of post-hoc testing were done using Tukey's HSD based on the 3 null hypotheses. Since there are 25 wards, the resulting dataframe of the pairwise comparisons is very large and thus was filtered to show only those where a significant effect was found (based on a p-value less than 0.05).

For total space by ward, there were no treatment levels with a p-value of statistical significance. For total space by subsidy, the p-value of Tukey's HSD is less than 0.05. The results of post-hoc testing suggest that we reject the null hypothesis: total capacity does differ by subsidy. Lastly, in post-hoc testing for the interaction effect there were multiple treatment levels where the p-value was less than 0.05, suggesting statistical significance. The majority of the interactions involve different wards and/or subsidy type; only wards 4, 14, and 15 show a significant interaction within the same ward where the subsidy level differed. Interestingly, the geographical location of each ward pair where the p-value was less than 0.05 were not close to each other. The interaction of the two variables can also be visualized using an interaction plot.

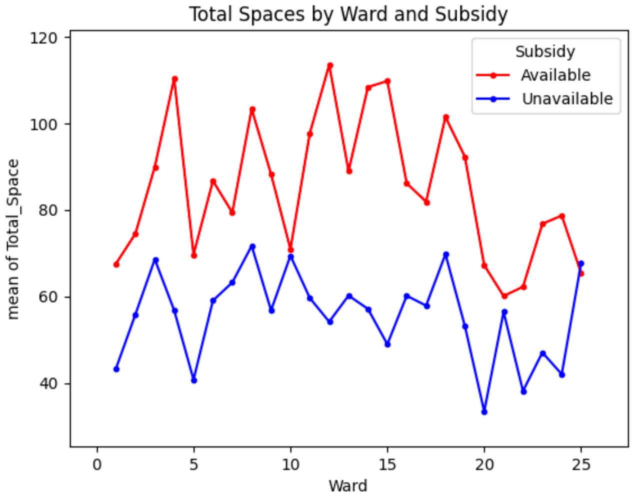

Figure 6: Interaction plot of ward levels and subsidy levels on total space.

The interaction plot shows a significant difference between the means of each Subsidy level across most wards, with the exception of Ward 10 and 21; for these wards, the mean total space is almost equal. As we move from Wards 1 to 24, centres with subsidy tend to have higher capacity on average compared to centres without subsidy. In Ward 25, however, we see the opposite effect: the mean total space is marginally higher for centres without subsidy than those with subsidy. There doesn't seem to be a distinguishable pattern regarding the increase/decrease of total space as we move across treatment level Ward. We can see that as we move across the wards, some increase in both subsidy available and subsidy unavailable centres, while for others, centres with subsidy increase while centres without subsidy decrease or vice versa.

Testing ANOVA Assumptions

Assumption 1: The residuals follow a normal distribution

This assumption will be tested using both visualization methods and statistical tests.

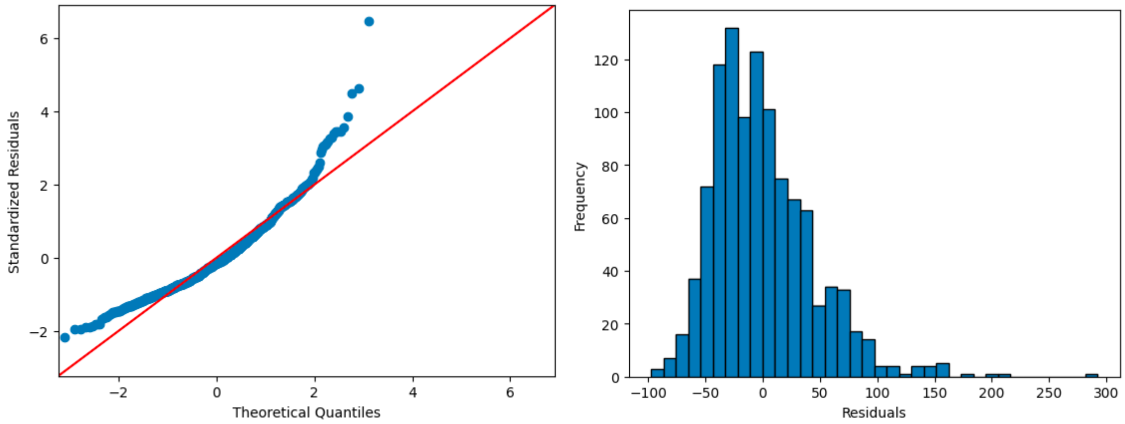

Figure 7: Distribution of residuals as a QQ plot (left) and histogram (right) for two-way ANOVA.

The QQ plot shows a curve upwards rather than a diagonal line along the reference line. This suggests that the model is not well-fitted. Furthermore, the histogram has a right-skewed. This suggests that the residuals do not follow a normal distribution.

The Shapiro-Wilk test yielded a p-value of less than 0.001, meaning we reject the null hypothesis: the residuals do not follow a normal distribution.

Assumption 2: The ward levels and subsidy levels have equal variance

Using Levene's Test again, due to non-normal distribution of the residuals, we found a p-value of less than 0.001, meaning we reject the null hypothesis: the treatment levels do not have equal variance.

In conclusion, our testing for ANOVA assumptions confirm that our dataset is not suitable for analysis using two-way ANOVA. It is possible to transform the data and re-do the tests to see if the assumptions are now met. Another option is to use another statistical test where equal variance is not required, such as Friedman's Test. Overall, the results of our two-way ANOVA testing is not reliable and should be interpreted with caution.