Toronto Licensed Child Care Analysis

INF2178: Experimental Design for Data Science

Instructor: Professor Shion Guha

Yuanyuan Pan 1003980150

Assignment Introduction

This assignment aims to conduct a statistical analysis on the licensed child care capacity and agencies by examining the data collected from Toronto licensed child care centers which contains relevant information for multiple age groups. Motivated by the current situation of licensed and unlicensed child care in Ontario which is high in fees and low in availability of spaces, the primary goal of this assignment is to investigate features that may influence the total child care spaces (TOTSPACE) provided by these child care centers. The study employs statistical techniques, specifically one-way ANOVA and two-way ANOVA, to gain insights into the impact of potential factors on the total child care spaces offered. This report will offer a comprehensive statistical analysis based on the trends of different plots and the resulting findings from two ANOVA tests.

The study addresses two fundamental research questions:

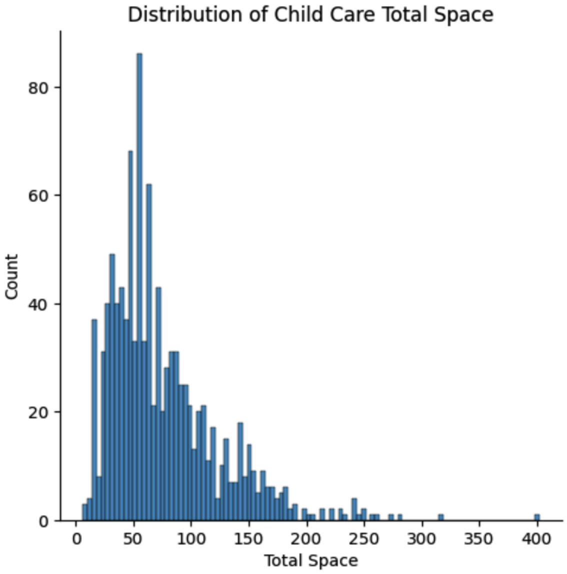
1. How does the operating auspice of child care centers influence the total spaces they offer?
2. How does the interaction between auspice and subsidy status influence the total child care spaces?

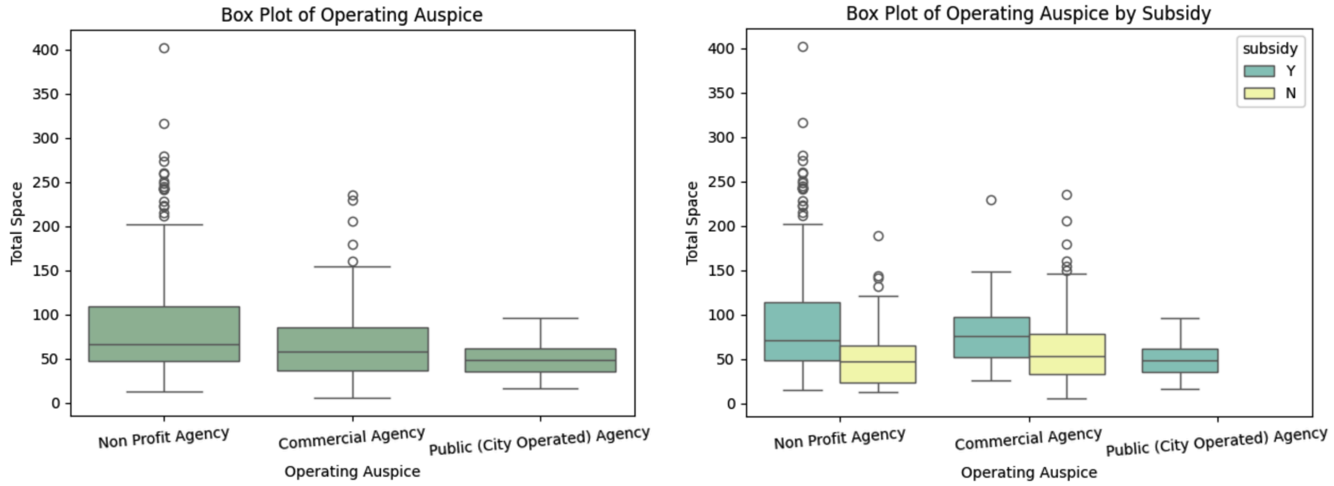
Data Pre-processing and Cleaning

In the data pre-processing step, the data named “INF2178_A2_data.xlsx” has been read into the dataframe. Only four variables have been kept for use in the later analysis, they are id, AUSPICE, TOTSPACE, and subsidy. ID indicates the id number of each row, AUSPICE represents the operating auspice, TOTSPACE represents the total number of child care spaces, and subsidy represents whether there is a subsidy for each agency.

Preliminary Analysis

The **histogram** on the right shows the distribution of child care total spaces for all age groups. By examining the above histogram, it is right skewed (positive skewness), which means that median could be more appropriate when measuring central tendency since mean could be influenced by outliers. This could also be seen as a unimodal since it has its highest peak at around 60. Other than that, potential outliers could also be detected.





The box plot on the left displays three different types of operating auspice. By examining the plot, it shows that non profit agency total space count is the highest, followed by commercial agency and public agency. Both non profit agencies and commercial agencies have an obvious trend of positive skewness (right skewed) while the public agencies tend to have a normal distribution on the total space count. Both non profit and commercial agencies have outliers.

The combined box plot on the right shows the distribution of total space count of different operating auspice by whether having subsidy. By examining the above plot, it shows that in general, child care spaces tend to have subsidies rather than not having it. For non profit agencies, agencies having subsidies are more in count than not having subsidies. Both having and not having subsidy agencies space count are positively skewed and having outliers. For commercial agencies, it seems that the space count for having and not having subsidies are similar. The trends for both of them are slightly positively skewed as well. For public agencies, it is very clear that all child care spaces under this agency have subsidies. It is also clear that the trend compared to other agencies is more normally distributed.

One-way ANOVA

According to what has been specified in research questions, this One-way ANOVA test focuses on examining the variation in total child care spaces across different auspices. This analysis aims to determine whether there are statistically significant differences in the mean total spaces among the auspices. In this ANOVA test, the independent variable is Operating Auspice (AUSPICE) which categorizes child care centers into three distinct types: Commercial, Non-Profit, and Public, and the dependent variable Total Child Care Spaces (TOTSPACE) which represents the child care spaces for all age groups. The hypothesis of the test specified as following:

Null Hypothesis (H0): There is no significant difference in mean total child care spaces among child care centers with different auspices (Commercial, Non-Profit, Public).

Alternative Hypothesis (H1): There is a significant difference in mean total child care spaces among child care centers with different auspices.

ANOVA Table:

	df	sum_sq	mean_sq	F	PR(>F)
C(AUSPICE)	2.0	9.611211e+04	48056.057145	21.843051	5.057716e-10
Residual	1060.0	2.332065e+06	2200.061571		

The above ANOVA table displays the degree of freedom, sum of squares and mean squares for both operating auspice and the residual. The sum of squares shows the sum of squared deviations of each group mean from the overall mean, where the mean squares reflects the variance between group means. The F-statistic is the ratio of variances and a higher value indicates more significant differences between group means. The **p-value** (PR(>F)) found in the above table is extremely low ($p < 0.001$) which suggests strong evidence against the null hypothesis. Therefore, it indicates that at least one pair of means among the operating auspice groups is significantly different.

Tukey's HSD Post Hoc Test

Since the conducted one-way ANOVA shows there is at least one pair of means that have significance difference among the operating auspice, the Tukey's HSD post hoc test is conducted to identify the specific pair(s) of groups that differ from each other in terms of mean total child care spaces.

group 1	group 2	Difference	p-value
Non Profit Agency	Commercial Agency	17.119417	0.001000
Non Profit Agency	Public (City Operated) Agency	34.334610	0.001000
Commercial Agency	Public (City Operated) Agency	17.215193	0.077966

The above test results of Tukey's HSD post hoc test pinpoint where the significant differences lie among groups. The difference represents the estimated difference in mean total child care spaces between two groups being compared. By examining the p-value of three pairs, it is obvious that there is a significant difference in mena total child care spaces between Non-Profit and Commercial Agencies, with Non-Profit Agency having a higher mean since the p-value is 0.001. Also, the mean total child care spaces for Non-Profit Agencies are significantly higher when compared to Public Agencies. For the pair Commercial Agency vs. Public Agency, although the mean difference is positive, the resulting p-value (0.078) shows that it is not statistically significant.

Checking Model diagnostics

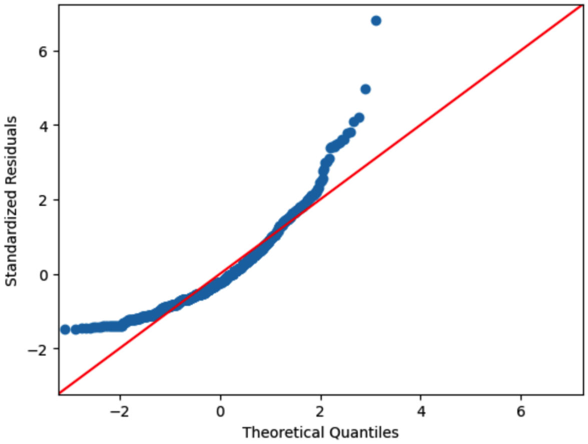
After conducting the ANOVA and Tukey's HSD Post Hoc test, assumptions need to be checked since it could help to ensure the validity and interpretability of the results.

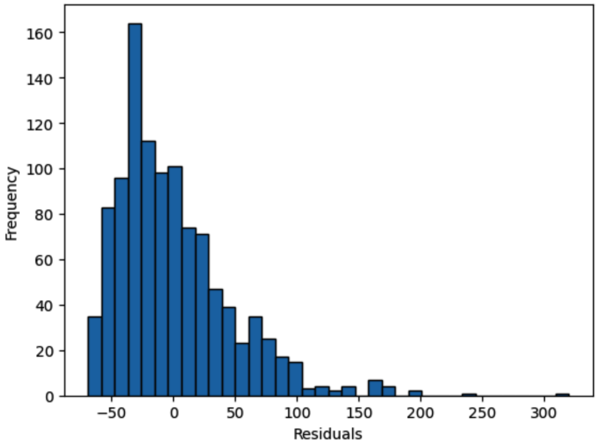
Quantile-Quantile (Q-Q) Plot:

The below Q-Q plot on the left provides a visual assessment of the normality assumption. By examining the plot, while the central region aligns with the theoretical quantiles, the curve on the two sides suggest that the residuals may not perfectly follow normal distributions. The deviation could be due to non-constant variance or outliers.

<u>Residual Distribution:</u>

The below histogram on the right shows the distribution of the residuals. The right-skewed histogram indicates that the residuals have a tendency to be larger on the positive side. The unimodal distribution suggests a central tendency, while the outliers on the right indicates potential instances where the model does not fully capture variability.





<u>Shapiro-Wilk Test:</u>

Moreover, the Shapiro-Wilk test has been conducted on the residuals, which will provide insights into how well the residuals conform to a normal distribution. The resulting **W = 0.902** while **p-value < 0.001**. The value of 0.902 is close to 1 which suggests that the residuals show some degree of adherence to a normal distribution. The p-value suggests strong evidence against the null hypothesis that the residuals are normally distributed.

<u>Levene's Test:</u>

In the previous section (preliminary analysis), the histogram shows that the sample is not normally distributed, hence a Levene's Test is conducted to assess the equality of variances across different groups.

Test statistics (W)	17.9271
p value	< 0.001

Above table shows the results of the test, since the p-value is less than the significance level, it indicates that there is strong evidence to reject the null hypothesis of equal variances. This violation of the assumption of homogeneity of variances may affect the reliability of the one-way ANOVA results.

Two-way ANOVA

The second research question aims to explore the interaction between operating auspice and subsidy status and its influence on the total child care spaces in Toronto. Two-way ANOVA is employed to investigate how these two independent categorical variables (operating auspice and subsidy) jointly impact the dependent variable, which is the total space of child care in Toronto. The hypothesis of this test are specified as following:

Null Hypotheses (H0):

There is no significant main effect of Auspice on the total childcare spaces.

There is no significant main effect of Subsidy Status on the total childcare spaces.

There is no significant interaction effect between Auspice and Subsidy Status on the total childcare spaces.

Alternative Hypotheses (Ha):

There is a significant main effect of Auspice on the total childcare spaces.

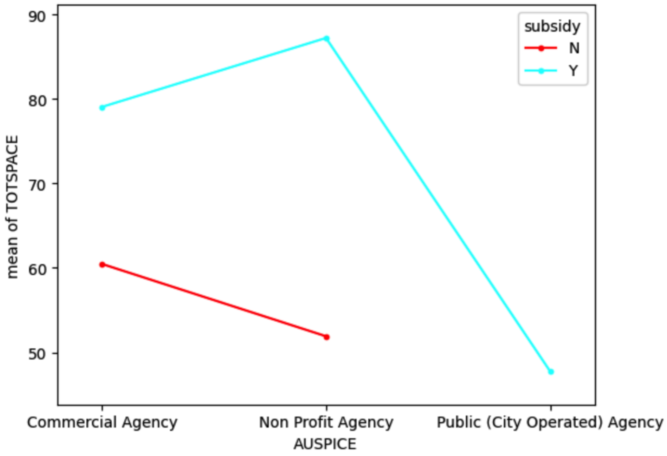
There is a significant main effect of Subsidy Status on the total childcare spaces.

There is a significant interaction effect between Auspice and Subsidy Status on the total childcare spaces.

ANOVA Table:

	df	sum_sq	mean_sq	F	PR(>F)
C(subsidy)	1.0	8.352744e+04	83527.441787	40.117876	3.529094e-10
C(AUSPICE)	2.0	8.568144e+03	4284.072217	2.057622	1.282684e-01
C(subsidy):C(AUSPICE)	2.0	5.603445e+04	28017.227037	13.456555	1.694282e-06
Residual	1058.0	2.202809e+06	2082.050461		

By examining the above ANOVA table, the p-value for both subsidy and the interaction between subsidy and auspice are less than 0.001, hence, the main effect of subsidy status is statistically significant which suggests that there is a significant difference in total childcare spaces based on whether a center has a subsidy contract or not, and The interaction effect between Subsidy Status and Auspice is also statistically significant which suggests that the combined influence of subsidy status and auspice on total childcare spaces is different from what would be expected if their effects were independent.



Interaction Plot:

The left interaction plot shows the effect between Auspice and Subsidy status on the response variable Total Child Care Spaces. The lines are not parallel suggests an interaction effect, meaning the effect of Auspice is dependent on whether there is a subsidy or not. For centers without a subsidy, the Total Spaces decrease from Commercial to Non-Profit. For centers with a subsidy, there is an increase from Commercial to Non-Profit and then a decrease from Non-Profit to Public.

Operating Auspice Post Hoc Test:

The test results for operating auspice shows that the p-value for both **Non Profit Agency vs. Commercial Agency (p=0.0043)** and **Non Profit Agency vs. Public Agency (p=0.0042)** are less than the significance level, indicating that there is significance in total child care spaces between these groups. On the other hand, **Commercial Agency vs. Public Agency (p=0.250)** is insignificant.

Subsidy Status Post Hoc Test:

The test results for subsidy status has a **p-value = 0.001**, which indicates a statistically significant difference between groups with and without subsidy.

Interaction (Auspice & Subsidy) Post Hoc Test:

The test results for the interaction effect of auspice and subsidy show only six pairs of significant results. By interpreting the p-values these six pairs of interactions all have a p-value less than 0.05, and they are shown as follow (Note: Y means with subsidy, N means without subsidy):
Non-Profit Agency (Y) vs. Non-Profit Agency (N): with a mean difference of 35.33.
Non-Profit Agency (Y) vs. Commercial Agency (N): with a mean difference of 26.76.
Non-Profit Agency (Y) vs. Public Agency (Y): with a mean difference of 39.46.
Non-Profit Agency (N) vs. Commercial Agency (Y): with a mean difference of 27.16.
Commercial Agency (Y) vs. Commercial Agency (N): with a mean difference of 18.60.
Commercial Agency (Y) vs. Public Agency (Y): with a mean difference of 31.29.
According to the above result, the non-profit agency and commercial agency with subsidy tend to have more effect when they have interactions with other groups.

Checking Model diagnostics

After conducting assumption checks, the results are very similar to the results of assumption checks in one-way ANOVA.

Conclusion

The result of one-way ANOVA on auspice and total spaces shows that there is a significant difference in the total child care spaces among different auspices, and Post hoc Tukey's HSD test revealed significant differences in total spaces lies in Non-Profit vs. Commercial agencies and Non-Profit vs. Public agencies. The result of two-way ANOVA shows that the interaction between auspice and subsidy status has a significant influence on the total child care spaces.The analysis indicates that both auspice and subsidy status play crucial roles in determining the total child care spaces. Non-Profit agencies tend to have significant differences compared to Commercial and Public agencies. Additionally, the interaction between auspice and subsidy further emphasizes the need for a nuanced approach when assessing child care spaces. This finding could be crucial for policymakers and organizations involved in child care when providing insights into potential features influencing total spaces and guiding decisions for resource support. However, since the ANOVA tests might have violated the normality and homogeneity of variance, the results from this report could lack of generalizability and reliability, further statistical method may be needed.