Candice Huo-Understanding Childcare Centers in Toronto

1. Project overview

In Toronto, it is difficult to find childcare in Toronto. This study aims to use data analysis to gain an in-depth understanding of the trends, characteristics, and key factors of Toronto's Childcare Centers, and to provide strong support for relevant policies and social intervention.

This study uses a childcare data set from the City of Toronto, which includes multiple key variables, such as Childcare spaces for different age groups, information of the childcare center etc.

2. Data cleaning, preprocessing

To ensure the quality and availability of data, we do data cleaning and preprocessing including the following two parts:

- Delete unnecessary columns such as '_id', 'LOC_ID', 'LOC_NAME', 'ADDRESS', 'PCODE', 'BLDGNAME'.
- Since there are no missing values or outliers, we will use the original dataset.

3. One-Way ANOVA and Two-Way ANOVA/Research Question1

Research Question 1: What is the relationship between Childcare spaces for all age groups and other factors such as AUSPICE, cwelcc_flag and subsidy. For this research question, we perform One-Way ANOVA and Two-Way ANOVA to explore the data.

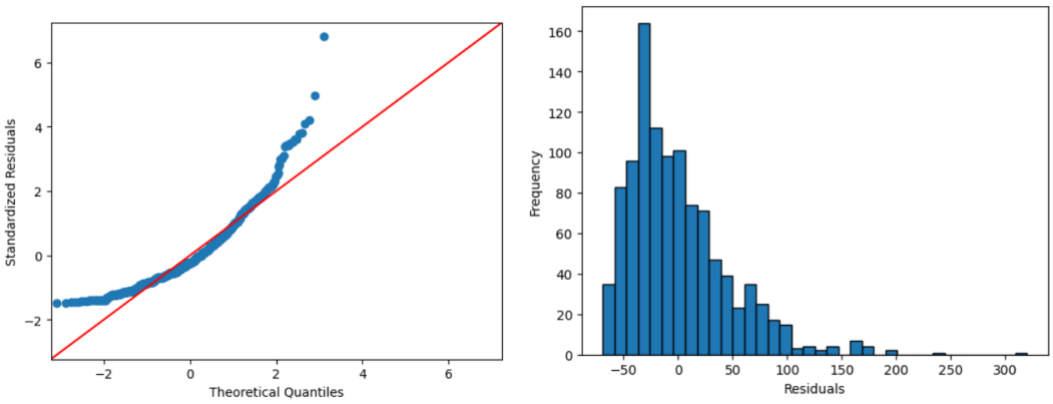
- One-Way ANOVA

The one-way ANOVA is a parametric test used to test for a statistically significant difference of an outcome between 3 or more groups. From the table, we can know that 'AUSPICE' has three different values, so in this part, we calculate One-Way ANOVA on TOTSPACE based on AUSPICE'. The result is shown below.

	F-statistic	p-value
One-way ANOVA results	21.84	5.06e-10

The F-statistic is the ratio of the variation between sample means to the variation within samples. The p-value is the probability of obtaining an F-statistic as extreme or more extreme than the observed one, assuming the null hypothesis is true.

In the second part, we examine the residuals of the model by plotting the Quantile-Quantile plot and histogram, and we can see that the residuals of the model meet the assumptions of the linear regression model,



In the third part, we get Shapiro-Wilk test result and Levene's test result

Shapiro-Wilk test	Test statistic (W): 0.902	p-value: 1.496e-25
Levene's test	Test statistic (W): 17.9271	Degrees of freedom: 2.0000

We can see from the table above that the Shapiro-Wilk test suggests that the residuals do not follow a normal distribution and Levene's test indicates that the variances of the residuals are not equal across different levels of the factor.

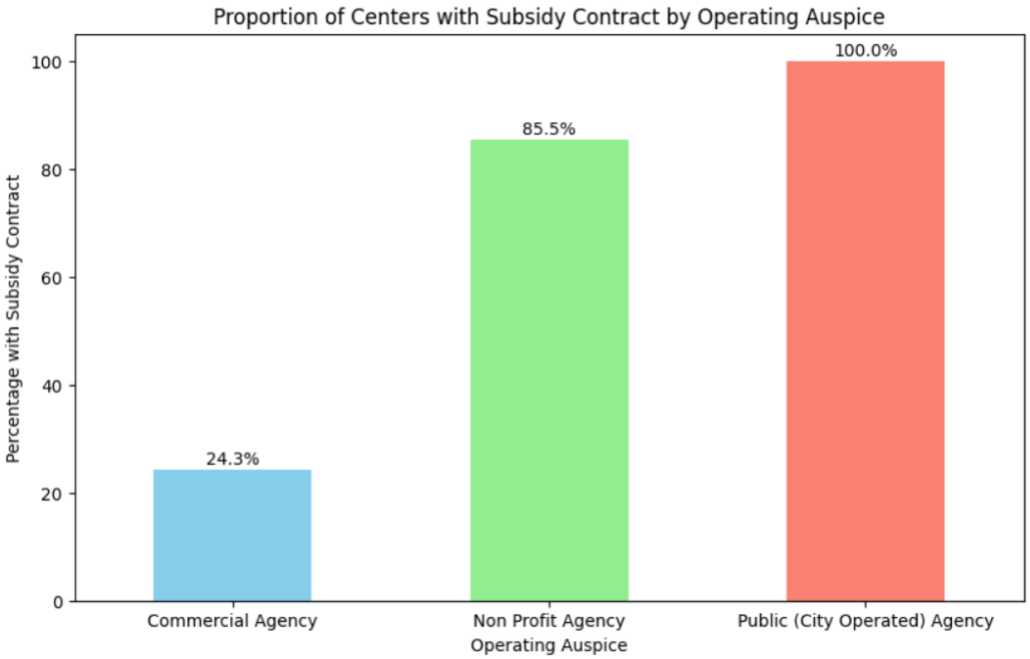
Conclusion: the F-statistic is 21.84, which means that the variation between sample means is much larger than the variation within samples. The p-value is 5.06e-10, which means that the probability of getting such a large F-statistic by chance is very low, so we can reject the null hypothesis and conclude that there is a statistically significant difference between these group means.

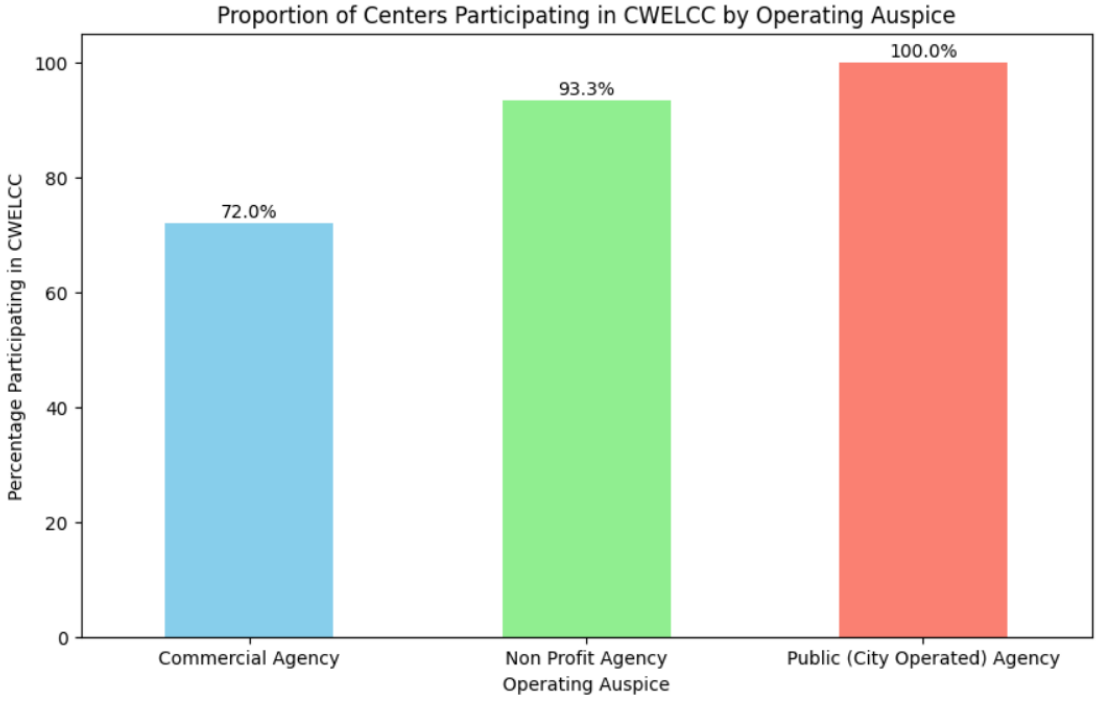
- Two-Way ANOVA

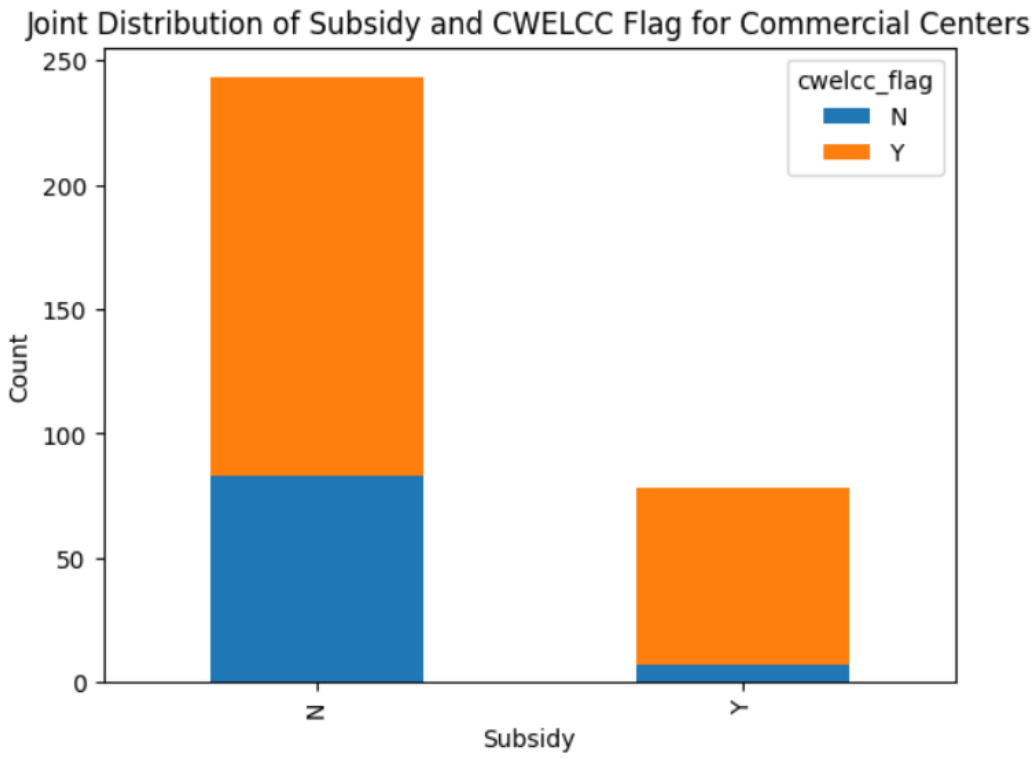
A two-way ANOVA is used to determine whether there is a statistically significant difference between the means of three or more independent groups that have been split on two factors. In this experiment, we explore the ANOVA results under variable cwelcc_flag, subsidy, and the interaction effect between cwelcc_flag and subsidy.

	df	sum_sq	mean_sq	F	PR(>F)
C(cwelcc_flag)	1.0	6.72e+03	6723.43	3.18	7.49e-02
C(subsidy)	1.0	9.82e+04	98161.81	46.38	1.63e-11
C(cwelcc_flag):C(subsidy)	1.0	1.91e+04	19108.68	9.03	2.72e-03
Residual	1058.0	2.24e+06	2116.69		

Besides calculating One-Way ANOVA and Two-Way ANOVA**,** we also visualize the proportion of Centers with Subsidy Contract and proportion of Centers Participating in CWELCC to further explore the data. For commercial agency, we visualize the joint distribution of Subsidy and CWELCC Flag.





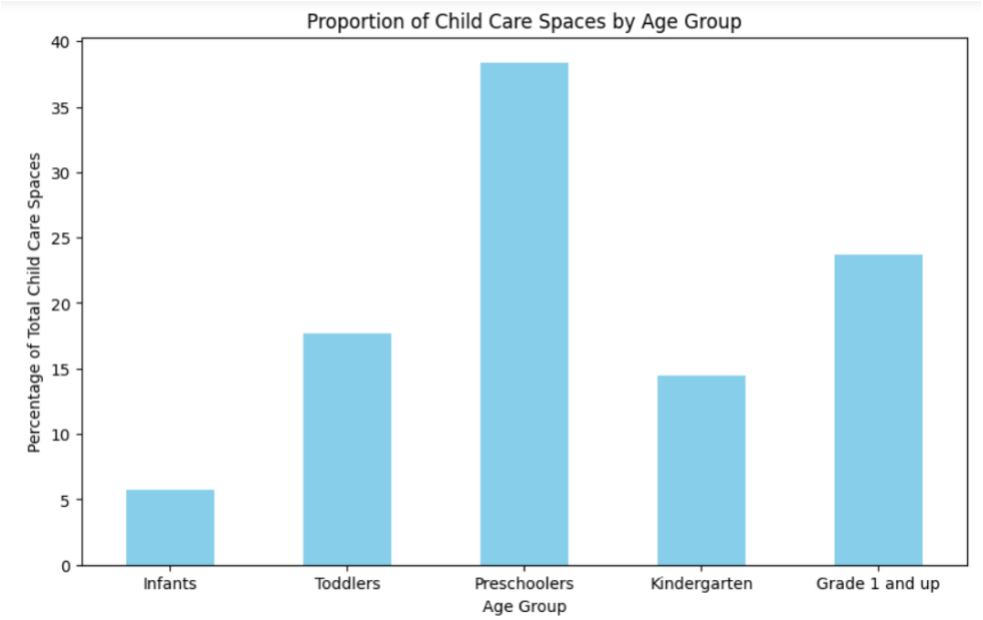


Conclusion: Based on the provided ANOVA results and plot, we can see that subsidy has a significant effect on the data, while the individual effect of cwelcc_flag on the data is not significant enough. This is mainly because most centers have already participated in CWELCC and only 25% of commercial agencies have a fee subsidy contract. However, when considering the interaction between cwelcc_flag and subsidy, this also significantly affects the data since there is an imbalance in the joint distribution of subsidy and cwelcc_flag values for commercial childcare centers.

4. Exploratory data analysis/Research question 2

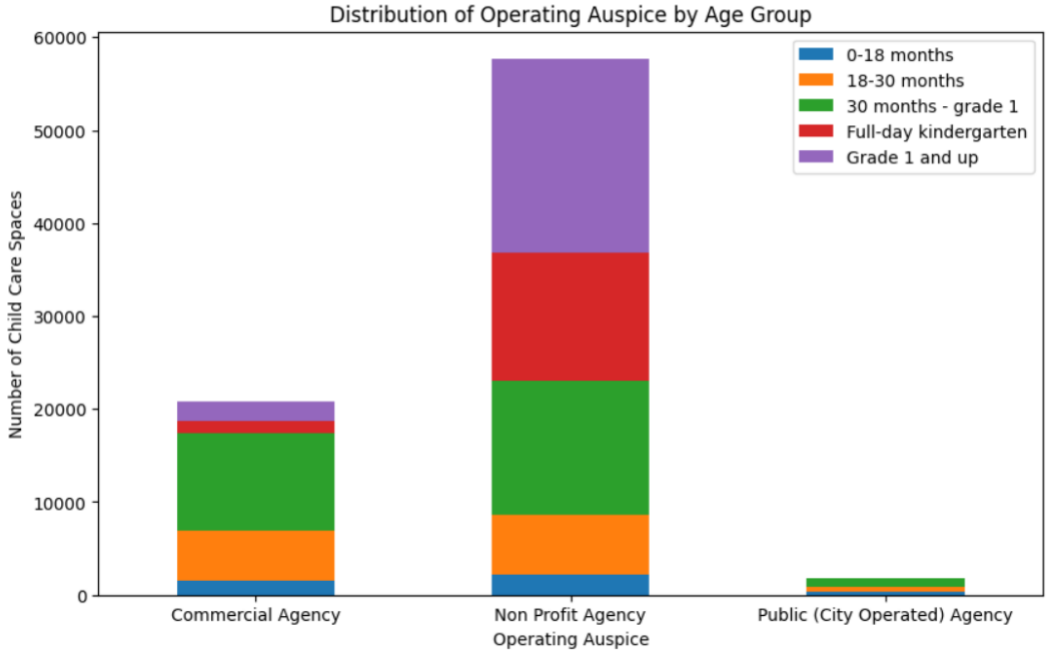
Research Question 2: How childcare center provides childcare spaces for different age groups. To explore this question, we visualize some key factors of the data.

1) First, we explored the proportion of different age group childcare spaces.



In this plot, we can see that childcare spaces in the centers are primarily focused on preschool-aged children which has nearly 40%, with a relatively balanced distribution across other age groups except infants.

2) Distribution of Operating Auspice by Age Group:



In summary, most childcare centers operate on a non-profit basis, with a smaller proportion operating for profit and public entities being relatively rare. The

distribution of childcare spaces between profit and non-profit centers varies across age groups, with non-profit organizations being more prevalent child in Full-day kindergarten and Grade 1 and up. Public centers have a limited presence and only provide space for babies.

Conclusion: Based on the above analysis, the reasons and factors influencing the selection of different childcare centers by various age groups may vary:

1. Commercial organizations prioritize offering childcare services for younger children, this is may because older children, particularly school-age children, may require additional resources and facilities.
2. Non-profit organizations may have a mission or philosophy that serves a wide range of age groups, including those with specific developmental needs or those traditionally underserved by childcare services. As a result, they may offer a more diverse range of programs catering to different age groups.
3. Not too many spaces are provided for infants, this is mainly because infants need higher dependency and need for individualized care.
4. Non-profit organizations are more inclined to accept government funding and support, and participate in government subsidy programs, such as the CWELCC program, compared to for-profit entities, which participate to a lesser extent.

5. Further analysis

Currently we are exploring One-Way ANOVA and Two-Way ANOVA for variable TOTSPACE based on different conditions. We also explore the reason why different age groups choose different kinds of childcare centers by visualization some result from the dataset. In future analysis, we can explore **Geospatial Analysis**: Conducting geospatial analysis to identify spatial patterns and disparities in the distribution of childcare centers across different City wards. This could help identify underserved areas and inform targeted interventions to improve access to childcare services.