INF2178 Experimental Design For Data Science

Technical Assignment #2

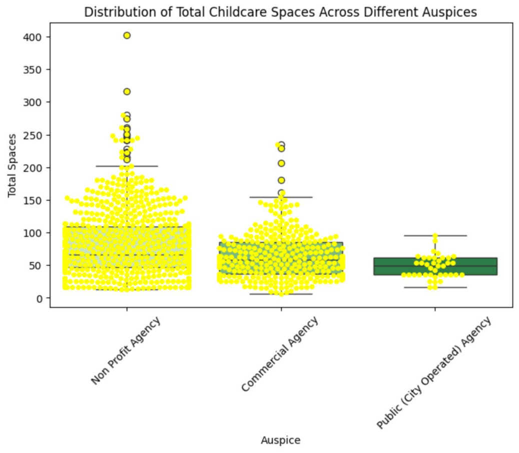
Student Name: Liguangxuan He | Student ID: 1006141809

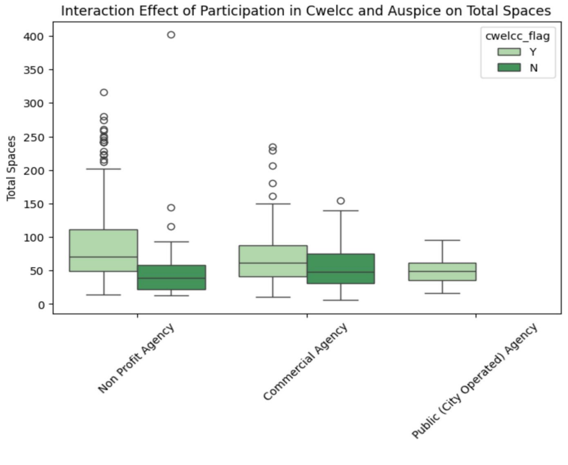
Data Cleaning & Data Wrangling

The raw dataset under consideration comprises 17 columns and 1063 entries. The initial exploration revealed the necessity of minor data cleaning to ensure seamless usability in future analyses. Handling missing values in ‘BLDGNAME’ column were replaced with ‘Not Specified’ to maintain data integrity. Converting ‘PCODE’ column to string type for consistency and ease of use. Renaming column ‘_id’ to ‘ID’ for improved clarity. Removing duplicate rows to avoid potential errors due to non-unique indexes. Filtering missing values in columns of interest such as ‘AUSPICE’, ‘cwelcc_flag’, and ‘TOTSPACE’ to ensure removal of any missing values to prepare for subsequent analysis.

- **Observations of columns in interest:**
  - 'AUSPICE' describes the management type of child care centers, with entries showing 321 commercial, 703 non-profit, and 39 public agencies, revealing their operational and financial framework.
  - 'cwelcc_flag' denotes enrollment in the CWELCC program, with 'Y' for 926 participating centers and 'N' for 137 non-participating, important for evaluating the program's effects.
  - 'TOTSPACE' counts the total available child care spots, a key metric to analyze child care capacity variations by agency type and CWELCC involvement.
- **Research Questions:**
  - One-way ANOVA: Is there a significant difference in total childcare spaces among different types of managing agencies?
  - Two-way ANOVA: Does the interaction between managing agency type and participation in the CWELCC program have significant effect on total childcare spaces, beyond the main effects of each factor individually?

Prior Visualization


Distribution of Total Childcare Spaces Across Different Auspices


Interaction Effect of Participation in Cwelcc and Auspice on Total Spaces

Two boxplots were generated to gather insights for ANOVA analysis. The first one displays total childcare spaces across different auspices, aiding in assessing variance and central tendency under various management types. This aligns with one-way ANOVA. The second boxplot depicts total childcare spaces across combinations of auspices and CWELCC participation, relevant to two-way ANOVA. These visuals offer clarity on the interaction effect between CWELCC participation and auspices on childcare spaces, aiding in statistical analysis and interpretation.

The first graph illustrates the total childcare spaces categorized by managing agency type. Non-Profit Agencies exhibit a wide range and higher median of total spaces compared to Commercial and Public Agencies, indicating greater variability in capacity, especially with significant outliers suggesting some Non-Profit Agencies offer many more spaces. Commercial Agencies have a narrower spread and fewer outliers, generally offering fewer spaces. Public Agencies show the most consistent distribution with fewer and lower total spaces.
The second graph explores how CWELCC program participation affects space availability across agency types. Non-Profit Agencies participating in CWELCC tend to have higher median space numbers compared to non-participating ones, suggesting the program's impact on space availability. Similarly, for Commercial Agencies, CWELCC participation appears to be linked to increased space availability, although less pronounced than in Non-Profit agencies. Public Agencies show minimal variation in space numbers based on CWELCC participation, indicating consistent effects across these agencies. These visual insights lay the groundwork for a two-way ANOVA analysis to statistically assess the main effects and interaction effect of agency type and CWELCC participation on space availability.

1.1 One-way ANOVA Analysis

For the one-way ANOVA test, I will compare the total number of spaces across the different auspices. This analysis help us understand if the type of agency managing the childcare centre has a significant effect on the availability of childcare spaces.

	df	sum_sq	mean_sq	F	PR(>F)
C(AUSPICE)	2.0	96112.11	48056.06	21.84	0.0
Residual	1060.0	2332065.26	2200.06	NaN	NaN

Table 1. One-way ANOVA table

The one-way ANOVA analysis compares the total number of childcare spaces across different auspices to determine if the type of agency managing the childcare center significantly affects space availability. The null hypothesis (H0) posits that there is no significant difference in mean total childcare spaces among the auspices, while the alternative hypothesis (H1) suggests that there is a significant difference. The obtained F-statistic of 21.84 with a corresponding p-value less than 0.05 (significance level < 0.05) rejects the null hypothesis, indicating that at least one auspice significantly differs from the others in terms of mean total childcare spaces. Additionally, the mean square values provide insight into the variance within groups (mean_sq = 48056.06) and between groups (mean_sq = 2200.06), indicating that there is substantially more variability in total childcare spaces between auspices than within them. This suggests that the type of agency managing the childcare center indeed has a significant effect on childcare space availability, as evidenced by the ANOVA results.

1.2 One-way ANOVA: post hoc test using Tukey's HSD

group 1	group 2	Diff	Lower	Upper	q-value	p-value
A	B	17.12	9.70	24.54	7.66	0.00100
A	C	34.33	16.22	52.45	6.29	0.00100
B	C	17.21	-1.45	35.88	3.06	0.07797

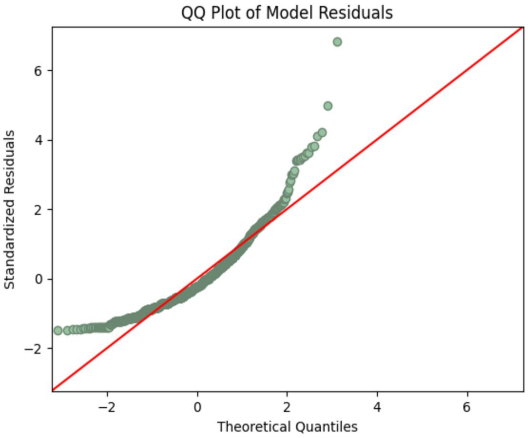
Table 2. Post Hoc Test for One-way ANOVA analysis, A-Non-Profit Agency, B-Commercial Agency, and C-Public (City Operated) Agency

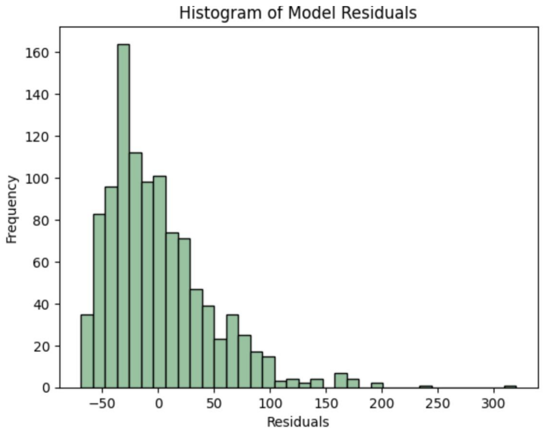
The post hoc test results from Table 2, conducted using Tukey's HSD (Honestly Significant Difference) method, provides insights into specific pairwise comparisons between different auspices following the one-way ANOVA analysis. The test results indicate significant differences in mean total childcare spaces between Non-Profit Agencies (A) and both Commercial Agencies (B) (difference = 17.12, $p < 0.001$) and Public Agencies (C) (difference = 34.33, $p < 0.001$). Additionally, there is a significant difference in mean total childcare spaces between Commercial Agencies and Public Agencies (difference = 17.21, $p = 0.078$). However, while the difference between Commercial and Public Agencies is notable, it does not reach statistical significance at the conventional threshold of $p < 0.05$. These findings highlight specific pairwise differences in childcare space availability among different types of managing agencies, further supporting the results of the one-way ANOVA analysis.

1.3 One-way ANOVA Analysis Assumption Check

Assumption 1: Normality of residual

The quantile-quantile (qq) plot and histogram for the one-way ANOVA residuals indicate non-normality, the qq-plot reveals a noticeable deviation from the expected line at both end edges (above the red line), with a pronounced skew in the upper tail, while the histogram shows a right-skewed distribution, lacking symmetry around the central peak. These suggest the presence of positive skewness and outliers, potentially compromising the ANOVA's assumptions.





Assumption 2: Homogenity of variances (Levene's test)

From table 1.3.2 (below), The levene's test results, with a test statistic W of 17.93 and p-value of almost 0, suggest that the assumption of homogenity of variances has been violated for the one-way ANOVA test. The small p-value which is less than 0.05 indicates that there are statistically significant differences in the variances across the groups being compared. The finding means that the group variances are not equal and, therefore, the basic assumption required for conducting a valid ANOVA is not met. States there's need to consider alternative statistical methods that do not assume equal variances.

Parameter	Value
Test Statistic (W)	17.9271
Degrees of freedom (Df)	2.0000
p value	0.0000

Table 1.3.2 Levene's test result

2.1 Two-way ANOVA Analysis

For the two-way ANOVA test, I will examine the interaction effect between 'cwelcc_flag' and 'AUSPICE' on the total spaces. This analysis will explore whether the combination of participating in cwelcc and managing agency type affects the number of available childcare spaces.

	df	sum_sq	mean_sq	F	PR(>F)
C(AUSPICE)	2.0	108334.81	54167.41	25.19	0.0
C(cwelcc_flag)	1.0	37688.32	37688.32	17.53	0.0
C(AUSPICE):C(cwelcc_flag)	2.0	29496.61	14747.81	6.86	0.0
Residual	1058.0	2275187.35	2150.46	NaN	NaN

Table 3. Two-way ANOVA table

In examining the influence of managing agency type, CWELCC participation, and their interaction on childcare space availability, we confront specific hypotheses. The null hypotheses assert that managing agency type and CWELCC participation neither individually nor interactively influence the total childcare spaces, suggesting no significant differences or effects. Conversely, the alternative hypotheses propose that managing agency type significantly impacts childcare spaces, that CWELCC participation has a significant effect, and crucially, that the impact of CWELCC participation notably varies with the managing agency type. The analysis yields F-statistics (25.19 for agency type, 17.53 for CWELCC participation, and 6.86 for their interaction) and p-values almost zero across the board, compellingly suggesting that not only do managing agency type and CWELCC participation independently affect childcare space availability but also their combined interaction significantly influences space provision. This robust statistical evidence leads to the rejection of all null hypothesis in favor of the alternative hypotheses, affirming significant variations in childcare spaces based on agency type, CWELCC participation, and the specific interaction between these factors.

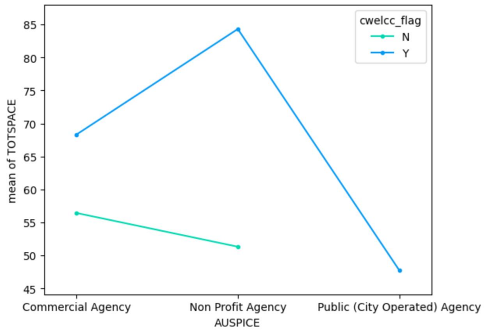


Figure. Interaction plot

The interaction plot shows the average total childcare spaces for different types of managing agencies, highlighting how CWELCC participation affects space availability. Non-Profit Agencies in CWELCC report higher averages, while Public Agencies show the opposite, especially those in CWELCC, suggesting that CWELCC's benefits on space availability aren't consistent across all agency types. This variation illustrates that the impact of CWELCC is moderated by the agency's managing structure.

2.2 Two-way ANOVA: post hoc test using Tukey's HSD

group 1	group 2	Diff	Lower	Upper	q-value	p-value
A	B	17.12	9.79	24.45	7.75	0.00
A	C	34.33	16.43	52.24	6.36	0.00
B	C	17.22	-1.24	35.67	3.10	0.07

Table 4. Post Hoc Test #1 for two-way ANOVA analysis, A-Non-Profit Agency, B-Commercial Agency, and C-Public (City Operated) Agency

The Post Hoc Test #1 linked to the two-way ANOVA suggests Non-Profit Agencies have significantly more childcare spaces compared to Commercial and Public Agencies, with mean differences of 17.12 and 34.33, respectively, both with strong statistical significance. In contrast, Commercial and Public Agencies do not differ significantly in the number of spaces offered, as indicated by a p-value above the 0.05 threshold.

group 1	group 2	Diff	Lower	Upper	q-value	p-value
Y	N	24.1	15.77	32.43	8.03	0.0

Table 5. Post Hoc Test # 2 for two-way ANOVA analysis

The results from Post Hoc Test #2, linked to the two-way ANOVA analysis, reveal that centers participating in CWELCC ('Y') have on average 24.1 more childcare spaces than those not participating ('N'). The difference is statistically significant, with a p-value close to 0, indicating an extremely low probability that such a difference in means could occur by chance. This underlines the effectiveness of CWELCC participation in increasing available childcare spaces.

group 1	group 2	Diff	Lower	Upper	q-value	p-value
(A, Y)	(A, N)	32.99	13.00	52.98	6.66	0.00
(A, Y)	(B, Y)	15.99	5.86	26.12	6.38	0.00
(A, Y)	(B, N)	27.88	12.99	42.76	7.56	0.00
(A, Y)	(C, Y)	36.54	14.72	58.36	6.76	0.00
(A, Y)	(C, N)	0.00	-inf	inf	0.00	0.90
(A, N)	(B, Y)	17.00	-4.19	38.18	3.24	0.20
(A, N)	(B, N)	5.11	-18.71	28.94	0.87	0.90
(A, N)	(C, Y)	3.55	-25.13	32.23	0.50	0.90

(A, N)	(C, N)	0.00	-inf	inf	0.00	0.90
(B, Y)	(B, N)	11.88	-4.57	28.33	2.92	0.31
(B, Y)	(C, Y)	20.55	-2.37	43.47	3.62	0.11
(B, Y)	(C, N)	0.00	-inf	inf	0.00	0.90
(B, N)	(C, Y)	8.66	-16.72	34.04	1.38	0.90
(B, N)	(C, N)	0.00	-inf	inf	0.00	0.90
(C, Y)	(C, N)	0.00	-inf	inf	0.00	0.90

Table 6. Post Hoc Test #3 for two-way ANOVA analysis, A-Non-Profit Agency, B-Commercial Agency, and C-Public (City Operated) Agency

Post Hoc Test #3, accompanying the two-way ANOVA, focuses on the interaction between agency type and CWELCC participation ('Y' for Yes, 'N' for No). The results demonstrate a significant difference between Non-Profit Agencies participating in CWELCC (A, Y) when compared to all other groups, except for Public Agencies not participating in CWELCC (C, N), which shows no difference (Diff = 0.00, p-value = 0.90). This significant difference is highlighted by the high q-values and a p-value of practically zero, emphasizing the positive impact of CWELCC participation for Non-Profit Agencies on the total number of childcare spaces.

Furthermore, the results do not show a significant difference between Non-Profit Agencies not participating in CWELCC (A, N) and other groups, indicating that without CWELCC participation, the agency type does not make a significant difference in childcare space availability. There are no significant differences among Commercial Agencies with or without CWELCC participation (B, Y vs. B, N) and between them and Public Agencies (C, Y and C, N). This suggests that the CWELCC program's benefits might be more pronounced within Non-Profit Agencies regarding the provision of childcare spaces.

Conclusion

The analysis conclusively illustrates that both the type of managing agency and participation in the CWELCC program significantly influence the availability of childcare spaces, with Non-Profit Agencies particularly benefiting from CWELCC participation. The one-way ANOVA highlighted a distinct variance in space availability across different agency types, confirmed by post hoc tests that detailed significant differences between Non-Profit and other agency types. The two-way ANOVA further emphasized the complexity of these relationships, revealing a pronounced interaction effect where CWELCC's impact varied by agency type. Particularly, Non-Profit Agencies participating in CWELCC were shown to offer significantly more childcare spaces, showcasing the program's potential in enhancing childcare space availability within specific organizational contexts.