INF2178
Chi-shiun Yang
1009916897
Mar 03, 2024
Assignment 2

1. Introduction

Due to the high fee and low availability of spaces for children, it is difficult for families in Ontario to find child care for their kids. To find solutions to this difficulty, we would like to explore the data collected from the licensed child care in Toronto and see if there are any ways to improve the problem.

There are three main research questions in our exploration:

1. Does operating auspices of the children care center cause differences in total space available?
2. Does operating auspices of the children care center cause differences in the percentage of space available in different age groups?
3. Do the whether having a fee subsidy contract and/or participating in CWELCC have interaction between them while exploring the total space available in the children care center?

2. Data Cleaning and Wrangling

The raw data contains 17 columns and 1063 rows. After reviewing the dataset, some columns are not used in this analysis while some columns are created for easier calculation.

The new created columns IG_rate, TG_rate, PG_rate, KG_rate, and SG_rate are the percentage of child care space in different age group from columns calculated by the space for each group, such as the column IGSPACE, divided by the total number of availability, the column TOTSPACE.

Beside the new create column and the TOTSPACE for the total number of space, the column AUSPICE (the operating auspice), subsidy (whether the center has a fee subsidy contract), and cwelcc_flag (whether the center participates in CWELCC) are used.

3. Explordinary Data Analysis (EDA)

For each research question, we would draw the box-plot for the column to visualize our dataset. To make sure running one-way and two-way ANOVA (analysis of

variance) in this case is reliable, we have to draw some plots to make sure that the dataset satisfied the assumptions. The assumptions of ANOVAs are that each factor level has a normal population distribution (we use q-q plots, histograms, and Shapiro Wilk test to check residuals are normally distributed), the distributions have the same variance (we are using the Bartlett's test or the Levene's test), and the data are independent (this is true for this dataset). On the other hand, if the p value is less than our significance level (0.05 in this analysis), we reject the null hypothesis that the mean between groups are the same, and we would then use post-hoc tests to see where the difference is.

Research question 1

To discover whether operating auspices of the children care center cause differences in total space available, we would like to use one-way ANOVA with categories from the column AUSPICE to the value in column TOTSPACE.

From the box plot below, we can see that the distribution is slightly different between each group of operating auspices. The p-value for this one-way ANOVA test is 5.057716e-10, which is much less than 0.05. Therefore, we reject the null hypothesis and say that the mean of total space is different when grouped by the operating auspices of the children care center.

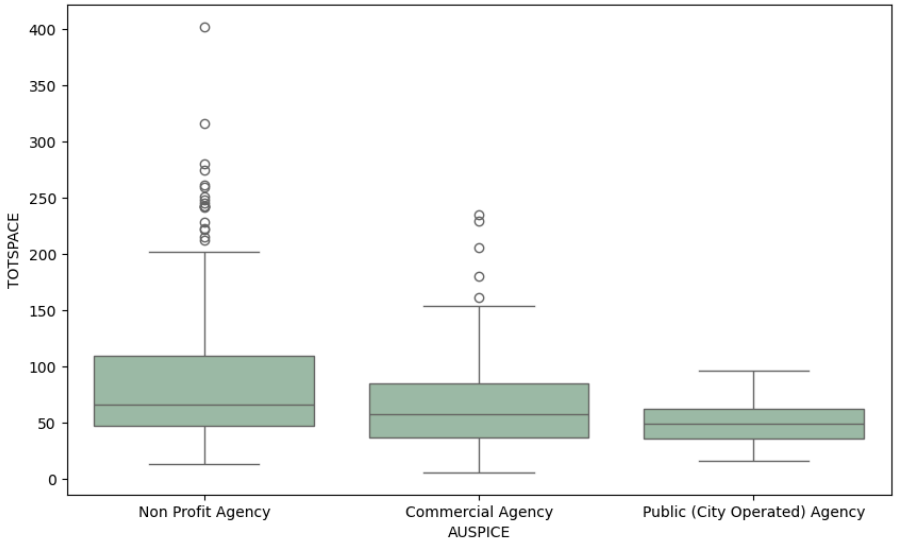

Figure 1: box-plot of total space grouped by auspices.

Since the means are different, we applied a post hoc test using Tukey's HSD. From the summary table, we can see that the p-value is higher than 0.05 between commercial agency and public agency. This tells us that the main difference occurs for the non profit agency group.

group 1	group 2	Diff	Lower	Upper	q-value	p-value
Non Profit Agency	Commercial Agency	16.81	3.99	29.62	4.36	0.01
Non Profit Agency	Public (City Operated) Agency	36.18	8.67	63.68	4.37	0.01

Commercial Agency	Public (City Operated) Agency	19.37	-10.14	48.88	2.18	0.27

Table 1: Tukey summary of total space between different auspice groups.

From the q-q plots and the histogram, we can see that the samples are skewed to the right. And since the Shapiro Wilk test has a p-value 1.50e-25, we use the Levene's test to see if the variances are homogeneous. The p-value is 0.0001 so from these results, the assumptions are not met so the reliability of the test is not very high but we can still somewhat say that the non profit agencies have higher total space.

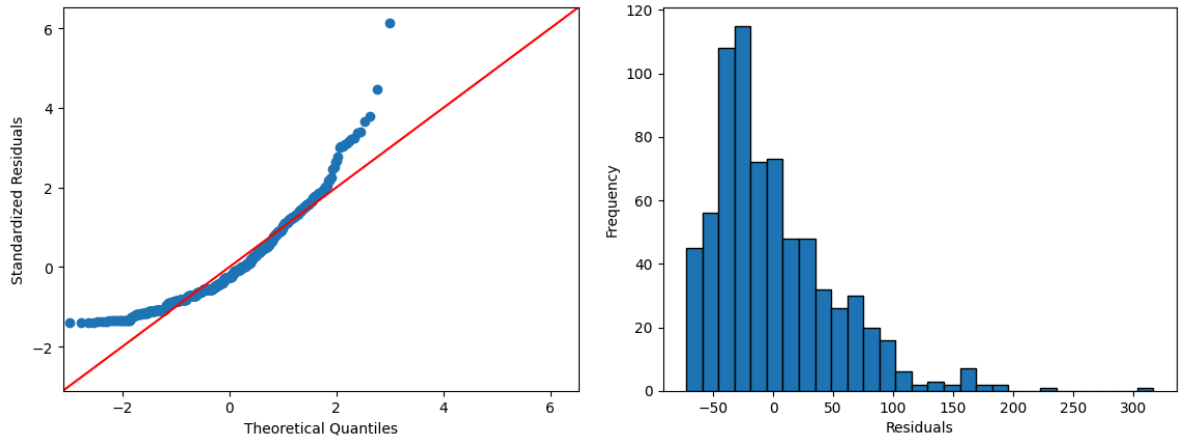

Figure 2: q-q plot and histogram of residuals.

Research question 2

To discover whether operating auspices of the children care center cause differences in the percentage of space available in different age groups, we use the same method as research question 1 but instead of using the value in column TOTSPACE, we use that of IG_rate, TG_rate, PG_rate, KG_rate, and SG_rate. The age of each column are:

- IG: infants 0-18 months
- TG: toddlers 18-30 months
- PG: preschoolers 30 months up until they enter grade one
- KG: children in full-day kindergarten
- SG: children grade one and up

We draw the same plots and run the same tests as we did in research question 1 just with different columns of the value we want to explore. From the ANOVA summaries, all p-values are less than 0.05, which tells us that there are differences between the mean of different operating auspices. Here are the p-values of the Tukey summary for each column. We can see that except for the group of infants 0-18 months, the commercial agency and public agency groups are more similar, which have the same result as in research question 1. The reason why the group non profit agency and commercial agency of infants 0-18 months are similar is that they have very few spaces so that their mean are both close to 0.

group 1	group 2	IG	TG	PG	KG	SG
Non Profit Agency	Commercial Agency	0.900	0.001	0.001	0.001	0.001
Non Profit Agency	Public (City Operated) Agency	0.001	0.001	0.001	0.001	0.001
Commercial Agency	Public (City Operated) Agency	0.001	0.241	0.529	0.582	0.398

Table 2: Tukey summary of space in different ages of children between different auspice groups.

When testing the assumptions of the ANOVAs, the p-value of the Shapiro Wilk test are all less than 0.05. Along with the q-q plot and histogram of residuals, we can say that the samples are not normally distributed. But for the p-value of Levene's test gathered in the table below, the variance of infants 0-18 months, toddlers 18-30 months, and preschoolers 30 months up until they enter grade one are the same.

IG	TG	PG	KG	SG
0.1196	0.0113	0.0739	<0.0001	<0.0001

Table 3: P-value of Levene's test of space between different auspice groups for all ages of children.

Research question 3

To discover whether having a fee subsidy contract and/or participating in CWELCC have interaction between them while exploring the total space available in the children care center, we would like to apply two-way ANOVAs by categories in the columns subsidy and cwelcc_flag to the value in column TOTSPACE.

We also use bar plots and the interaction plot to visualize the total space grouped by columns subsidy and cwelcc_flag. From this, it seems that participating in CWELCC has a higher mean.

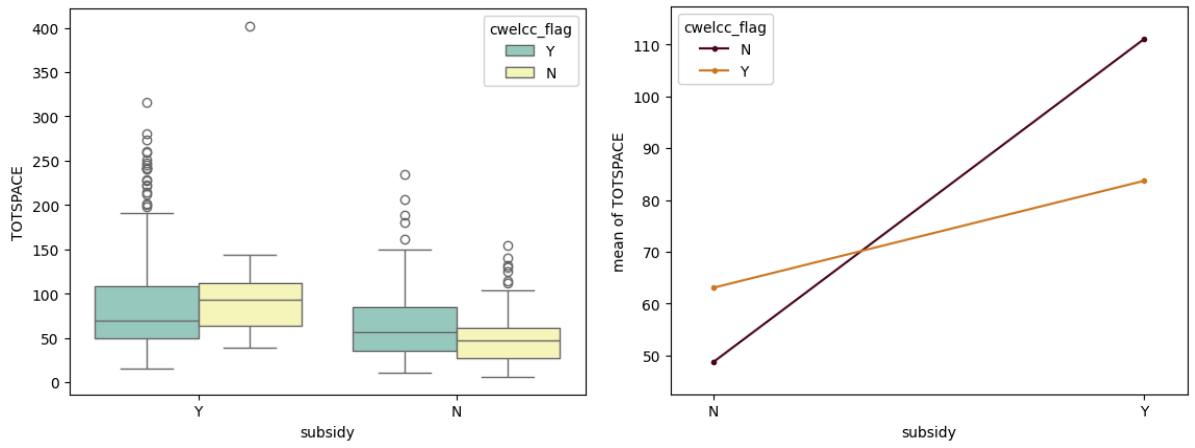

Figure 3: box-plots and interaction plot of total space grouped by subsidy and cwelcc_flag.

From the anova summary table, we can see that all p-values are less than 0.05, which rejects the null hypothesis and says that whether having a fee subsidy

contract and whether participating in CWELCC both have an effect on total space. Furthermore, there is an interaction effect between these two groups on total space.

	df	sum_sq	mean_sq	F	PR(>F)
C(subsidy)	1.0	9.81e+04	98161.80	46.37	1.63e-11
C(cwelcc_flag)	1.0	6.72e+03	6723.42	3.17	7.49e-02
C(subsidy):C(cwelcc_flag)	1.0	1.91e+04	19108.68	9.02	2.72e-03
Residual	1059.0	2.24e+06	2116.69	NaN	NaN

Table 4: Two-way ANOVA summary table of subsidy and cwelcc_flag on TOTSPACE.

We also checked the assumptions in this analysis. From the q-q plots and the histogram below, we can see that they are skewed to the right. And since the Shapiro Wilk test has a p-value 3.34e-26, we use the Levene's test to see if the variances are homogeneous. The p-value is 0.0001 so from these results, the assumptions are not met so the reliability of the test is not very high.

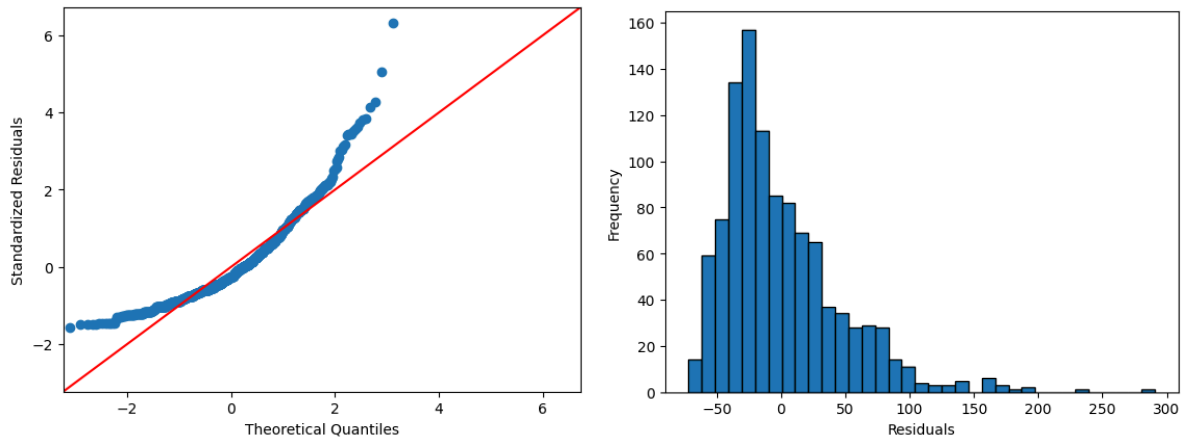

Figure 4: q-q plot and histogram of residuals.

4. Conclusion

From the results of the three research questions we have answered, the first conclusion is that the mean of total space is different when grouped by the operating auspices of the children care center, and the non profit agency having the highest total space are more different from other two groups. We got the same results in the second research question when separating the rate of availability by the children's age. Last but not least, from the third research question, we have that whether having a fee subsidy contract and participating in CWELCC both have an effect on total space and there is an interaction effect between them. A limitation of this analysis is that the distributions are not normal in all cases, and the variance are not homogeneous in most cases, which lowers the reliability of the ANOVAs' results. Combining other methods to solve this problem will give us a more accurate analysis.