Student ID: 1001420223
Student Name: Aaron Chen
Instructor: Shion Guha
Course: INF2178 Experimental Design for Data Science
Assignment: Technical Assignment 2

Exploring Licensed Child Care Centres Data in Toronto

1. Introduction

Accessing affordable and accessible child care is often challenging for families in Ontario due to high fees and limited availability. To address this issue, the provincial government pledged to add 100,000 new child care spaces within a decade starting from 2016. The dataset under examination comprises licensed child care centers in Toronto, which were updated in February 2024.

This report entails a comprehensive data analysis of Toronto's licensed child care centers with the objective of under covering whether there is difference between spacing options with different child care centers. We will delve into the dataset named 'INF2178_A2_data.xlsx'.

Research Questions:

1. **Research Question 1**: How does center space differ between Operating auspice, is there a significant difference?
2. **Research Question 2**: How does center space differ between Operating auspice and subsidy, is there a significant difference?

2. Data Cleaning and Data Wrangling

The dataset comprises 17 columns and 1063 rows. Initial inspection revealed null data only in 'BLDGNAME' column, which is non-essential for our analysis. Thus, no further cleaning was needed

Interested Columns:
AUSPICE: Operating auspice (Commercial, Non Profit or Public)
TOTSPACE: Child care spaces for all age groups
subsidy: Centre has a fee subsidy contract (Yes/No)

3. Exploratory Data Analysis (EDA)

Categorical Variables:

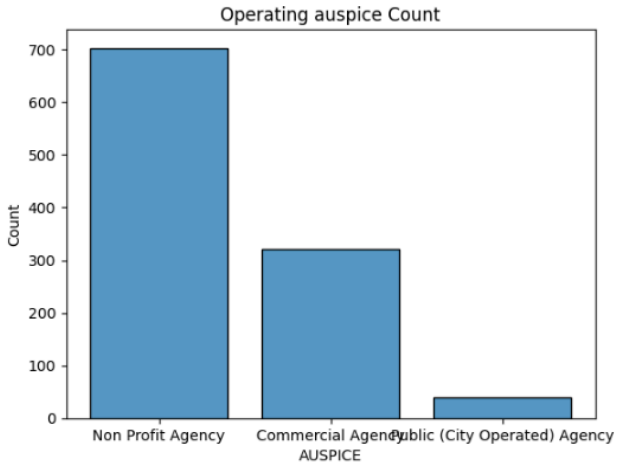
```
Operating auspice
['Non Profit Agency' 'Commercial Agency' 'Public (City Operated) Agency']
city ward number
[ 3  8 25 10 20 24  6 19  4  1 14  5  7 17  2 15  9 18 11 12 21 23 22 13
 16]
If centre has a fee subsidy contract
['Y' 'N']
```

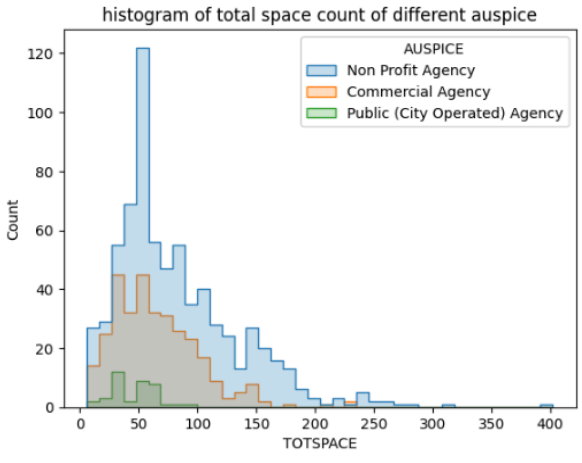
We see that there are 3 types of operating auspice (Non Profit Agency, Commercial Agency, Public (City Operated) Agency) and a condition on whether or not a center has a fee subsidy. There is also 25 ward number representing the city block the center is located in.

Continuous Variables:

	IGSPACE	TGSPACE	PGSPACE	KGSPACE	SGSPACE	TOTSPACE
mean	3.89	11.60	24.26	14.26	21.66	75.67
std	6.09	12.09	18.58	20.49	30.42	47.82
min	0	0	0	0	0	6
25%	0	0	16	0	0	43
50%	0	10	24	0	0	62
75%	10	15	32	26	30	97
max	30	90	144	130	285	402

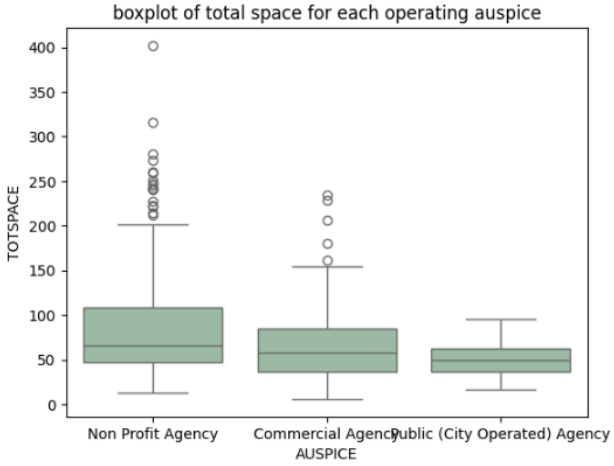
The TOTSPACE have similar mean and median but from the above table we can note a few points. The 25% percentile of SGSPACE, KGSPACE, TGSPACE, and IGSPACE are all 0 meaning that at least 25% of the centers don't provide spaces for children out of the 30 months up to grade one range (preschoolers).

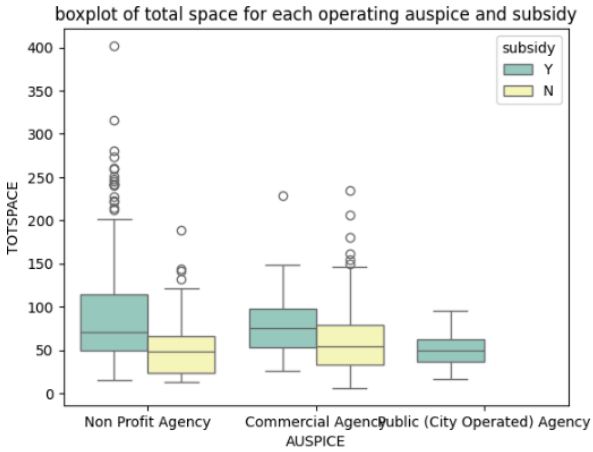




Looking at the above graph we can see a distribution of data we have between types of AUSPICE, we can see that most of our data consist of Non Profit Agency.

From the histogram we can see that all three types seem to be having the same variance, with Public (City Operated) Agency having too little data compared to the other two.





Looking at the Box plot we see that the median of the three groups are pretty much the same, suggesting they have the same central tendency. We also can see with more data, the variance of Non Profit Agency type is greater than the other two. One thing to note for the Non Profit Agency type is that the distribution looks skewed, and it might be mostly due to outliers. The spread of the three groups looks roughly the same, suggesting it passes the Homogeneity of Variance assumption.

Looking at the boxplot for when we consider subsidy, we see that for a public agency there is no center without subsidy. Overall the distribution looks normal even with the subsidy split, and also satisfying the assumption of anova.

4. ANOVA

An ANOVA was conducted to examine the differences in total space capacity among different licensed child care centers categorized by AUSPICE. The results of the ANOVA test are presented in the table below:

	df	sum_sq	mean_sq	F	PR(>F)
C(AUSPICE)	2.0	96112.11	48056.06	21.84	5.06e-10
Residual	1060	2332065.26	2200.06	NaN	NaN

The ANOVA results indicate a significant difference in space capacity between AUSPICE categories ($p < 0.001$), suggesting rejection of the null hypothesis.

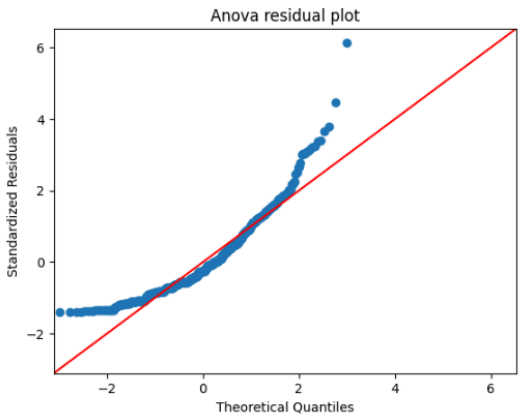
To further explore the significant difference identified in the ANOVA, Tukey's Honestly Significant Difference (HSD) test was performed. The results of the pairwise comparisons are summarized below: NPA(Non Profit Agency), CA(Commercial Agency), PA(Public Agency)

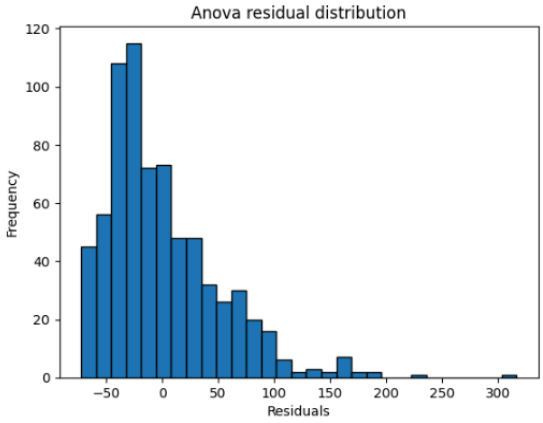
	group1	group2	Diff	Lower	Upper	q-value	p-value
0	NPA	CA	16.81	3.99	29.62	4.36	0.0061
1	NPA	PA	36.18	8.67	63.68	4.37	0.0059
2	CA	PA	19.37	-10.14	48.88	2.18	0.2726

The results reveal a significant difference in mean space capacity between NPA/CA and NPA/PA, as indicated by p-values less than 0.05. This might be caused by a significant difference in NPA compared to the other two.

Assumption Checks:

Residual Normality:





Shapiro Wilk test result

Statistic	P Value
0.9018	1.49e-25

Examination of the residual plot suggested not normally distributed. Subsequently, the Shapiro Wilk test was conducted, revealing a significant departure from normality ($p < 0.01$)

Homogeneity of Variances:

Since the sample is not normally distributed, we will use the Levene's test to assess homogeneity of variances. The result is below

	Parameter	Value
0	Test Statistics(W)	9.19
1	Degrees of freedom	2.0

2	P value	0.0001

The test yielded a p-value less than 0.05, indicating violation of the assumption of homogeneity of variances. Given the violation of assumptions, caution is warranted in interpreting the ANOVA results.

We will tackle our second research problem using two-way ANOVA.

	df	sum_sq	mean_sq	F	PR(>F)
C(AUSPICE)	2.0	8567.99	4283.99	2.05	0.13
C(subsidy)	1.0	83527.44	83527.44	40.12	3.53e-10
C(AUSPICE):C(subsidy)	2.0	56034.45	28017.23	13.46	1.69e-06
Residual	1060	2202809.39	2082.05	NaN	NaN

The ANOVA results indicate a significant difference in space capacity between AUSPICE and subsidy categories ($p < 0.001$), suggesting rejection of the null hypothesis.

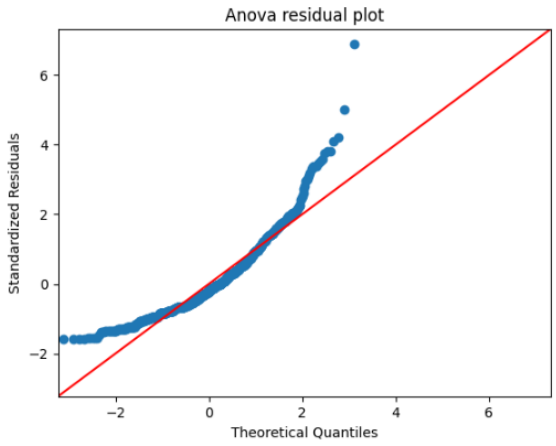
Below is the post hoc test of the two way ANOVA model, N/Y indicating No subsidy and Yes subsidy. (since the table is too long I will just paste the screenshot)

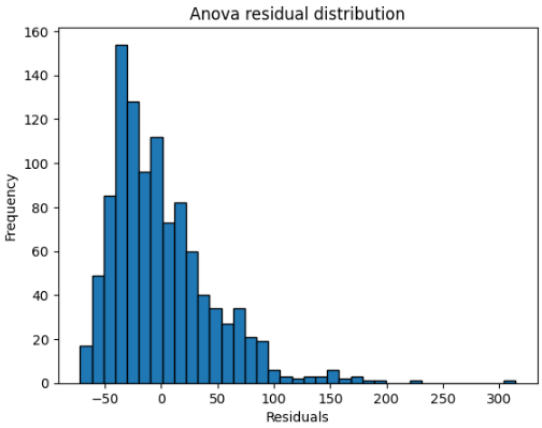
	group1	group2	Diff	Lower	Upper	q-value	p-value
0	(Non Profit Agency, Y)	(Non Profit Agency, N)	44.99	25.58	64.39	9.37	0.00
1	(Non Profit Agency, Y)	(Commercial Agency, Y)	0.15	-28.47	28.76	0.02	0.90
2	(Non Profit Agency, Y)	(Commercial Agency, N)	28.37	11.19	45.55	6.67	0.00
3	(Non Profit Agency, Y)	(Public (City Operated) Agency, Y)	40.75	8.31	73.20	5.08	0.00
4	(Non Profit Agency, Y)	(Public (City Operated) Agency, N)	0.00	-inf	inf	0.00	0.90
5	(Non Profit Agency, N)	(Commercial Agency, Y)	44.84	11.40	78.28	5.42	0.00
6	(Non Profit Agency, N)	(Commercial Agency, N)	16.62	-7.78	41.01	2.75	0.38
7	(Non Profit Agency, N)	(Public (City Operated) Agency, Y)	4.23	-32.54	41.01	0.47	0.90
8	(Non Profit Agency, N)	(Public (City Operated) Agency, N)	0.00	-inf	inf	0.00	0.90
9	(Commercial Agency, Y)	(Commercial Agency, N)	28.23	-3.98	60.43	3.54	0.12
10	(Commercial Agency, Y)	(Public (City Operated) Agency, Y)	40.61	-1.76	82.97	3.87	0.07
11	(Commercial Agency, Y)	(Public (City Operated) Agency, N)	0.00	-inf	inf	0.00	0.90
12	(Commercial Agency, N)	(Public (City Operated) Agency, Y)	12.38	-23.27	48.04	1.40	0.90
13	(Commercial Agency, N)	(Public (City Operated) Agency, N)	0.00	-inf	inf	0.00	0.90
14	(Public (City Operated) Agency, Y)	(Public (City Operated) Agency, N)	0.00	-inf	inf	0.00	0.90

The results reveal a significant difference in index 0,2,3,5 (NPA, Y vs NPA, N)(NPA, Y vs CA, N)(NPA, Y vs PA, Y) (NPA, N, CA, Y). Combining this with the result we got from the one-way anova, we can say NPA is definitely causing the mean diff problem.

Assumption Checks:

Residual Normality:





Statistic	P Value
0.9018	1.53e-25

$p<0.05$, reject null hypothesis thus residual not normal

Homogeneity of Variances:

	Parameter	Value
0	Test Statistics(W)	9.2
1	Degrees of freedom	2.0
2	P value	0.0001

$P<0.05$, reject null hypothesis thus homogeneity of Variances violated.

Like one-way anova, two-way anova yield almost the same result, the assumption check also didn't pass so we need to take caution on interpreting the result.

5. Conclusion

In this report, we analyzed Toronto's licensed child care centers dataset to explore spacing differences across operating auspices and subsidy statutes. We addressed two research questions:

1. Operating Auspices: A one-way ANOVA revealed significant differences in space capacity among operating auspices ($p<0.001$), particularly between NPA and others.
2. Operating Auspices and Subsidies: A two-way ANOVA showed significant differences in space capacity based on both operating auspices and subsidy statues($p<0.001$)

However, violated assumptions regarding normality and homogeneity of variances urge caution in interpreting the results for both ANOVA tests.