Technical Assignment 2:
Exploring Childcare Availability in Toronto

Devin W. de Silva
Faculty of Information
School of Graduate Studies, University of Toronto
INF2178: Experimental Design for Data Science
Professor Shion Guha
March 9, 2024

Childcare spaces available in Toronto has been severely underfunded and undersupplied. The dataset we are working with lists the childcare centres across the city, providing details such as their auspice, location, ward, type of building they're housed in, and the number of spaces available for different age groups (including infant, toddler, preschool, kindergarten, school-age), as well as whether they are subsidized or not.

By analyzing the spatial distribution of available spaces, we can better understand and quantify the gap between supply and demand for affordable childcare and the areas we need to act to bridge this gap.

Given the sheer scale of this dataset, it is difficult to provide a comprehensive analysis in this timeframe. For this analysis, we are specifically looking at the infant spaces specifically, as well as their auspice type and distribution. We are trying to answer three interrelated questions:

a) **Is there a significant difference in the number infant spaces available between different wards?**
- One-way ANOVA with ward as the independent variable and IGSPACE as the dependent variable.

b) **Does the type of auspice (Non-Profit vs. Commercial vs. City Operated) have a significant effect on the infant space?**
- One-way ANOVA with AUSPICE as the independent variable and IGSPACE as the dependent variable.

c) **Is there an interaction effect between the auspice type and the ward on the number of infant space available?**
- Two-way ANOVA with AUSPICE and ward as the independent variables and IGSPACE as the dependent variable.

TECHNICAL ASSIGNMENT 2

The raw data contains 11 columns with a total of 1063 columns. For the purpose of this analysis, not a lot of data cleaning would be necessary. I did, nonetheless, drop several columns that seem relatively less relevant to the analysis, namely '_id', 'LOC_ID', 'ADDRESS', 'PCODE', because they are unique identifiers for records or locations are not of much analytical value for our statistical tests that aim to understand trends. I also noticed that null values only exist in the BLDGNAME column, and this may disrupt the calculations, and since the building name itself is not exactly relevant to the analysis either, I proceeded to drop that column as well. I then double checked to confirm that there are no more null values in the dataset. I used the describe() method to gain some preliminary statistical information about the dataset, as shown below.

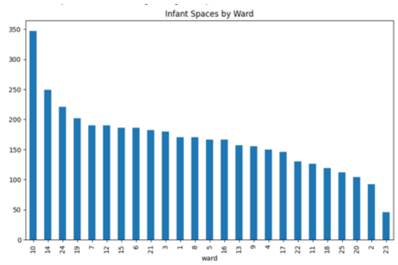
Index	IGSPACE	TGSPACE	PGSPACE	KGSPACE	SGSPACE	TOTSPACE	Total
mean	12.51	3.90	11.60	24.26	14.26	21.66	75.67
std	7.03	6.09	12.09	18.58	20.49	30.42	47.82
min	1.00	0.00	0.00	0.00	0.00	0.00	6.00
25%	6.00	0.00	0.00	16.00	0.00	0.00	43.00
50%	12.00	0.00	10.00	24.00	0.00	0.00	62.00
75%	19.00	10.00	15.00	32.00	26.00	30.00	97.00
max	25.00	30.00	90.00	144.00	130.00	285.00	402.00

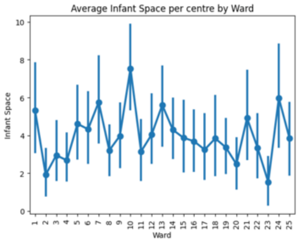
While the assignment notes that we could calculate overall centre capacity numbers, as far as I can tell, that number is already given in TOTSPACE, so I instead opted for calculating the proportion of each group space in relation to the total spaces and created new columns. For example, I tried to calculate the proportion of infant spaces as a percentage of total spaces in every centre: *df['p_IG'] = df['IGSPACE'] / df['TOTSPACE'].round(2)*. I wanted to know this because I was initially thinking about also analyzing whether the building type has related to the proportion of infant spaces available, but given the scope of the project I preferred to keep them more relevant to the more pertinent questions listed above. For this analysis, we focus mainly on the following columns:

IGSPACE	**Childcare spaces for infants 0-18 months**
ward	**City ward number**
AUSPICE	**Operating auspice (Commercial, Non-Profit or Public)**

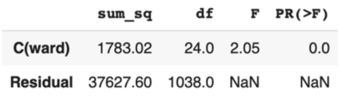
Research Question 1: Is there a significant difference in the number infant spaces available between different wards? This is a one-way ANOVA with ward as the independent variable and IGSPACE as the dependent variable.

- Null Hypothesis (H0): There is no significant difference in the mean number of infant spaces (IGSPACE) across different wards.
- Alternative Hypothesis (H1): There is a significant difference in the mean number of infant spaces (IGSPACE) across different wards.



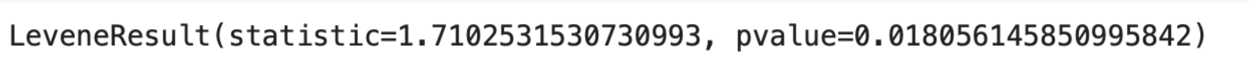


I first created a bar chart showing the number of infant spaces across different wards and a line chart showing the average infant space per centre by ward. As one can tell, ward 10 has by far the most infant spaces at nearly 350 while ward 25 has the least at less than 50.[1] The one way ANOVA test, unsurprisingly, showed that there are indeed significant differences in infant space availability across different wards (F(24, 1038) = 2.05, p = 0.002).



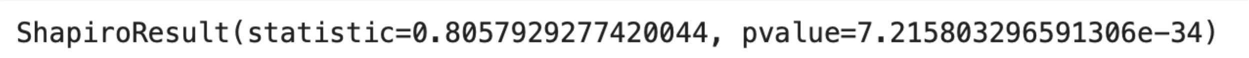
	sum_sq	df	F	PR(>F)
C(ward)	1783.02	24.0	2.05	0.0
Residual	37627.60	1038.0	NaN	NaN

Since this is a one-way ANOVA test, there is no interaction effect, so I have not included any interaction plots. I proceeded to check whether or not the assumptions for the one way ANOVA tests, by first testing the criteria of homogeneity of variances using Levene's test:



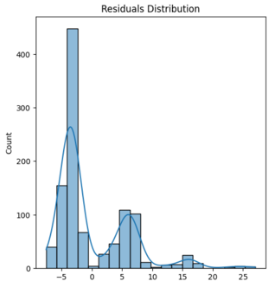
```
LeveneResult(statistic=1.7102531530730993, pvalue=0.018056145850995842)
```

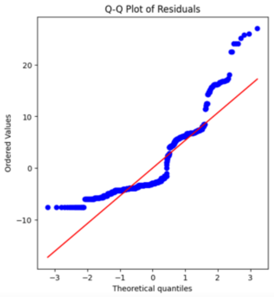
As one can tell, the p value for the result is around 0.02 which is smaller than the alpha value of 0.05, which means that the homogeneity assumption is not met. I proceeded to test the normal distribution of residuals using the Shapiro-Wilk test, and again, this second assumption is not met either.



```
ShapiroResult(statistic=0.8057929277420044, pvalue=7.215803296591306e-34)
```

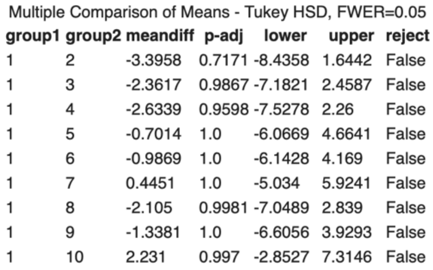
To confirm the test visually, I created a normal distribution and a Q-Q plot showing how skewed the residuals are, and how the residuals really fit the linear regression line.





I also ran a post-hoc Tukey HSD test, and it shows that the null hypothesis actually stands for the overwhelming majority of cases, the statistically difference in means is only evident in a number of comparisons, such as those between ward 10 and ward 23.

[1] I looked up the location of those wards and it comes as no surprise that ward 10 is the affluent Spadina-Fort York and and ward 25 is the relatively more deprived Scarborough North. Of course, we need to know per capita figures to before making conclusive suggestions about the relation between wealth and infant space availability.



Multiple Comparison of Means - Tukey HSD, FWER=0.05

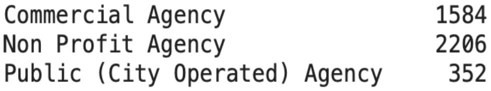
group1	group2	meandiff	p-adj	lower	upper	reject
1	2	-3.3958	0.7171	-8.4358	1.6442	False
1	3	-2.3617	0.9867	-7.1821	2.4587	False
1	4	-2.6339	0.9598	-7.5278	2.26	False
1	5	-0.7014	1.0	-6.0669	4.6641	False
1	6	-0.9869	1.0	-6.1428	4.169	False
1	7	0.4451	1.0	-5.034	5.9241	False
1	8	-2.105	0.9981	-7.0489	2.839	False
1	9	-1.3381	1.0	-6.6056	3.9293	False
1	10	2.231	0.997	-2.8527	7.3146	False

Research Question 2: Does the type of auspice (Non-Profit vs. Commercial vs. City Operated) have a significant effect on the infant space?

Again, this would be a one-way ANOVA, with AUSPICE as the independent variable and IGSPACE as the dependent variable.

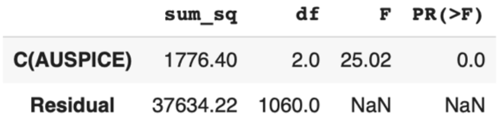
- Null Hypothesis (H0): There is no significant effect of the type of auspice (Non-Profit, Commercial, City Operated) on the number of infant spaces.
- Alternative Hypothesis (H1): There is a significant effect of the type of auspice on the number of infant spaces.

I first identified the total number of auspice infant spaces by category. A majority them are run by non-profit agencies.



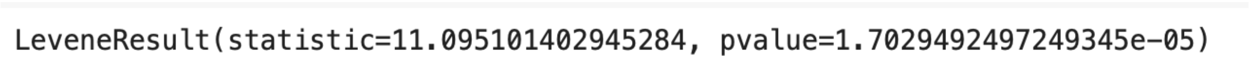
```
Commercial Agency               1584
Non Profit Agency               2206
Public (City Operated) Agency    352
```

The one-way ANOVA for auspice type on infant space also showed a significant effect (F(2, 1060) = 25.02, p < 0.001). This tells us that the management type of the childcare centre does influence infant space availability.



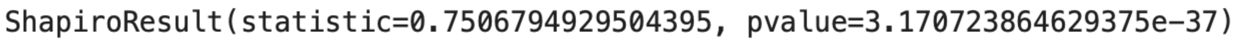
	sum_sq	df	F	PR(>F)
C(AUSPICE)	1776.40	2.0	25.02	0.0
Residual	37634.22	1060.0	NaN	NaN

Since this is a one-way ANOVA test, there is no interaction effect, so I have not included any interaction plots in this test either. I proceeded to check whether or not the assumptions for the one-way ANOVA tests, by testing the criteria of homogeneity of variances using Levene's test:

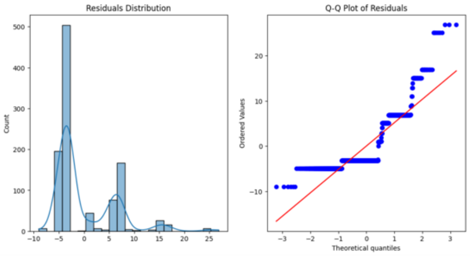


```
LeveneResult(statistic=11.095101402945284, pvalue=1.7029492497249345e-05)
```

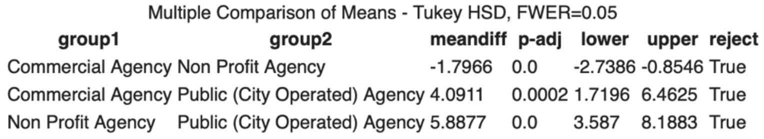
As one can tell, the p value for the result is less than 0.001 which is much smaller than the alpha value of 0.05, which means that the homogeneity assumption is not met. I proceeded to test the normal distribution of residuals using the Shapiro-Wilk test, and as with the first ANOVA, this second assumption is not met either. This is clearly visible visually in the distribution and residual Q-Q plots below.



```
ShapiroResult(statistic=0.7506794929504395, pvalue=3.170723864629375e-37)
```



Finally, I confirmed the significant effect the type of agency has on the number of infant spaces. The "reject" column is "True" for all comparisons means that the null hypothesis of equal means can be rejected in each case.



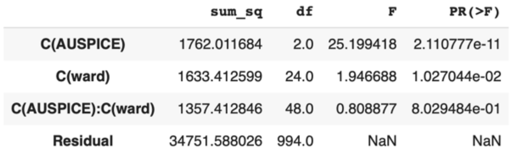
```
                 Multiple Comparison of Means - Tukey HSD, FWER=0.05
       group1                  group2             meandiff  p-adj   lower   upper  reject
Commercial Agency Non Profit Agency                -1.7966   0.0    -2.7386 -0.8546 True
Commercial Agency Public (City Operated) Agency    4.0911    0.0002 1.7196  6.4625  True
Non Profit Agency    Public (City Operated) Agency 5.8877    0.0    3.587   8.1883  True
```

Research Question 3: Is there an interaction effect between the auspice type and the ward on the number of infant space available?

This is a two-way ANOVA with AUSPICE and ward as two independent variables and IGSPACE as the dependent variable.

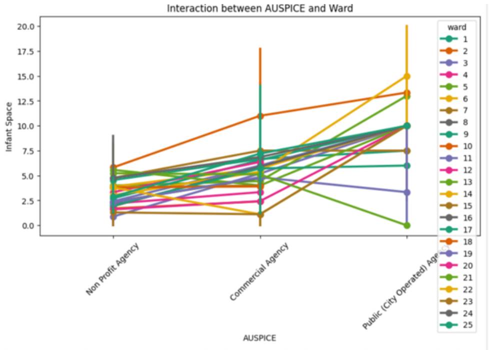
- Null Hypothesis (H0): There is no interaction effect between auspice type and ward on the number of infant spaces.
- Alternative Hypothesis (H1): There is an interaction effect between auspice type and ward on the number of infant spaces.

Unlike the two one-way ANOVAs, the two-way ANOVA test found that there is no significant interaction effect between ward and auspice type (p = 0.8029), so null hypothesis is accepted. It means that the two factors independently affect infant space availability and doesn't influence each other's effects.



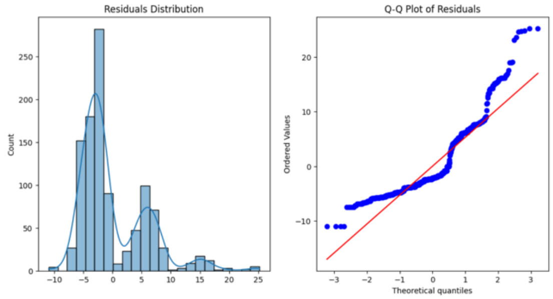
	sum_sq	df	F	PR(>F)
C(AUSPICE)	1762.011684	2.0	25.199418	2.110777e-11
C(ward)	1633.412599	24.0	1.946688	1.027044e-02
C(AUSPICE):C(ward)	1357.412846	48.0	0.808877	8.029484e-01
Residual	34751.588026	994.0	NaN	NaN

Here is the interaction plot between auspice and ward with an interaction effect added by ward:

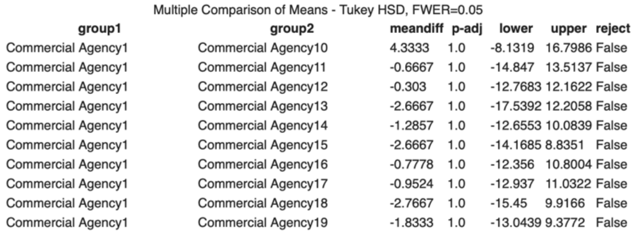


As with the Shapiro-Wilk test conducted before, the normal distribution of residuals assumption in this test is rejected, as the p value is extremely small at p > 0.001. The lack of a normal distribution graph below confirms the results.

```
Text(0.5, 1.0, 'Q-Q Plot of Residuals')
```



Lastly, I included a post-hoc Tukey test to confirm that there is genuinely no interaction effect between auspice and ward on the number of infant space available, as every single grouping is rejected. Given the sheer number of wards, the Tukey is very extensive, so I am only including the first few lines:



Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Commercial Agency1	Commercial Agency10	4.3333	1.0	-8.1319	16.7986	False
Commercial Agency1	Commercial Agency11	-0.6667	1.0	-14.847	13.5137	False
Commercial Agency1	Commercial Agency12	-0.303	1.0	-12.7683	12.1622	False
Commercial Agency1	Commercial Agency13	-2.6667	1.0	-17.5392	12.2058	False
Commercial Agency1	Commercial Agency14	-1.2857	1.0	-12.6553	10.0839	False
Commercial Agency1	Commercial Agency15	-2.6667	1.0	-14.1685	8.8351	False
Commercial Agency1	Commercial Agency16	-0.7778	1.0	-12.356	10.8004	False
Commercial Agency1	Commercial Agency17	-0.9524	1.0	-12.937	11.0322	False
Commercial Agency1	Commercial Agency18	-2.7667	1.0	-15.45	9.9166	False
Commercial Agency1	Commercial Agency19	-1.8333	1.0	-13.0439	9.3772	False

Conclusion

The first ANOVA showed us how there is significant differences in infant space availability across wards and makes it clear that there is significant geographical difference in access to infant care. The second ANOVA shows us that auspice type has a significant effect on infant spaces, and this suggests that the funding models of childcare centres do indeed make a difference to availability of spaces for infants.

The last two-way ANOVA, however, shows that there is no significant interaction effect between ward and auspice type, and tells us that we can see the two factors as being independent of one another.

It is worthwhile to note that, for the ANOVAs we conducted, the assumptions of homogeneity of variances and normal distribution of residuals generally not met, which means that the results may not be entirely reliable, despite the relatively large sample sizes.

With that said, childcare providers could consider these factors we analysed today as a starting point to identify where we could start in improving access to infant care services and research this dataset further to clarify any reliability concerns, and in turn, help in the development of targeted strategies to address gaps in childcare availability across agencies and geography.