Kevin Nitièma | 1002673298
INF2178H: Experimental Design for Data Science – Assignment 2

Exploring Child Care Access in Ontario

1.   Introduction

Access to affordable and high-quality child care is a pressing concern for families across Ontario due to limited availability and high fees. Despite efforts to improve access and reduce costs, 75% of families still find child care services unaffordable. In response, the provincial government pledged to create 100,000 new child care spaces by 2026 to ease the burden on families facing childcare expenses.

Central to Ontario's strategy are five key pillars: affordability, access, quality services, inclusion, and improved data reporting. This report analyses child care access in Ontario, specifically focusing on licensed centres in Toronto. Using the '*INF2178_A2_data.xlsx*' dataset from this GitHub repository, we aim to uncover trends and patterns to inform policy interventions for greater equity and accessibility in the child care system.

Our exploration will address two (2) fundamental research questions, serving as guiding principles in unravelling the troubling nature of Ontario's Child Care patterns. The research questions are as follows:

1.   **Research Question 1:** How do the quantities of available child care spaces differ across Ontario's Community Councils, and do these differences hold statistical significance?
2.   **Research Question 2:** Are there significant interactions between the types of Child Care clusters (such as infant, toddler, preschooler, etc.) and Ontario's Community Councils in determining the provision of child care services?"

Through a nuanced analysis of child care access, we aim to contribute meaningful insights that resonate with the overarching goal of ensuring every child in Ontario has access to safe, affordable, and high-quality early learning experiences.

2.   Data Cleaning and Data Wrangling

The raw dataset includes **1,063 entries (rows)** and **17 features (columns)**. Upon initial examination, minimal data cleaning was needed for our analysis. However, we identified discrepancies and introduced new features for future analysis. Below, we outline our observations and the added features:

A. Observations and Considerations:
1.   Since our analysis is mainly quantitative, we have reduced our working dataset to the following columns sourced from the raw dataset. Below we provided a short description of each column:
     - `ward:` City ward number.
     - `IGSPACE:` Child care spaces available for infants aged 0-18 months.
     - `TGSPACE:` Child care spaces available for toddlers aged 18-30 months.
     - `PGSPACE:` Child care spaces for preschoolers aged 30 months to grade 1.
     - `KGSPACE:` Child care spaces available for children in full-day kindergarten.
     - `SGSPACE:` Child care spaces available for children grade one and above.
     - `TOTSPACE:` Total number of child care spaces available across all age groups.

- ○ `subsidy:` Indicates whether the centre has a fee subsidy contract (Y/N).
- ○ `cwelcc_flag:` Indicates participation in CWELCC program (Y/N).

2. Only the the `BLDGNAME` columns seemed to have many missing values (NaN):
   - ○ `BLDGNAME` has **715 non-null**. We decided that `BLDGNAME` was insignificant to our data analysis, so no data cleaning was deemed necessary.

B. **Feature Engineering:**
   Here, we created **six (6) new features** to add to our dataset to aid in later analysis (we also later developed additional features which we address in our report). The features are as follows:

1. `IGSPACE_PF:` This feature represents the proportional frequency (percentage) of child care spaces available for infants in each child care centre. It indicates the relative space allocated to infants aged 0-18 months in a given centre. The subscript `"_PF"` stands for **proportional frequency**. We applied the same logic for the following features: `TGSPACE_PF, PGSPACE_PF, KGSPACE_PF, SGSPACE_PF`

   - ○ $IGSPACE\_PF\ (\%)\ =\ \dfrac{Number\ of\ child\ care\ spaces\ for\ infants}{Total\ number\ of\ child\ care\ spaces} \times 100$

2. `COMMUNITY_COUNCIL:` We categorised each entry by the corresponding Community Council ward. Our classification was informed by The City of Toronto, which served as our reference for ward clustering.
   - ○ `'ETOBICOKE' = [1, 2, 3, 5, 7]`
   - ○ `'NORTH_YORK' = [6, 8, 15, 16, 17, 18]`
   - ○ `'SCARBOROUGH' = [20, 21, 22, 23, 24, 25]`
   - ○ `'TORONTO_EAST_YORK' = [4, 9, 10, 11, 12, 13, 14, 19]`

3. Exploratory Data Analysis (EDA)

After adding additional features to our new working dataset, we proceeded with a comprehensive EDA to leverage insights that could potentially lead to **interesting research questions**. We started by describing our quantitative data (*Figure 1*) and plotting multiple figures including boxplots (*Figure 2*) , histograms (*refer to notebook*), and heatmap correlation matrices (*Figure 3),* to uncover patterns and trends across the dataset. Our detailed EDA is further documented in the accompanying code for reader reference.

	IGSPACE	TGSPACE	PGSPACE	KGSPACE	SGSPACE	TOTSPACE	IGSPACE_PF	TGSPACE_PF	PGSPACE_PF	KGSPACE_PF	SGSPACE_PF
mean	3.90	11.60	24.26	14.26	21.66	75.67	5.71	17.70	38.39	14.50	23.70
std	6.09	12.09	18.58	20.49	30.42	47.82	9.10	17.86	28.85	18.87	28.37
min	0.00	0.00	0.00	0.00	0.00	6.00	0.00	0.00	0.00	0.00	0.00
25%	0.00	0.00	16.00	0.00	0.00	43.00	0.00	0.00	15.69	0.00	0.00
50%	0.00	10.00	24.00	0.00	0.00	62.00	0.00	14.49	39.02	0.00	0.00
75%	10.00	15.00	32.00	26.00	30.00	97.00	11.43	30.81	56.14	30.23	46.58
max	30.00	90.00	144.00	130.00	285.00	402.00	50.00	100.00	100.00	100.00	100.00

Figure 1: Dataset Quantitative Data Statistics

Following our EDA, we found minimal correlation and strong skewness in most of our continuous variables. This prompted us to focus on analyzing the total spaces (`TOTSPACE`) in Child Care Centers across Ontario, particularly across different Community Councils: Etobicoke, North York, Scarborough, and Toronto & East York. However, we observed that the distribution of total space in Child Care Centers by Community Council was also largely right-skewed, as seen in *Figure 4.*
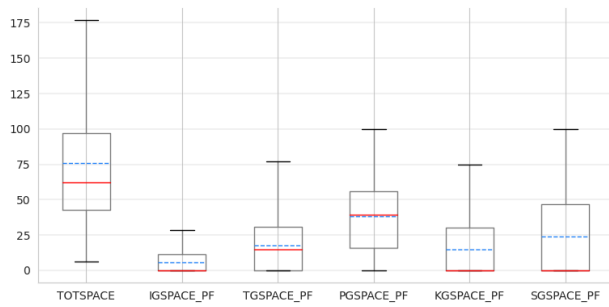
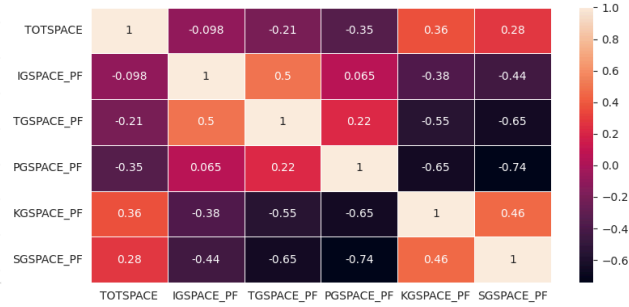Figure 2: Boxplot of Quantitative Data of interest   Figure 3: Correlation Matrix Heatmap
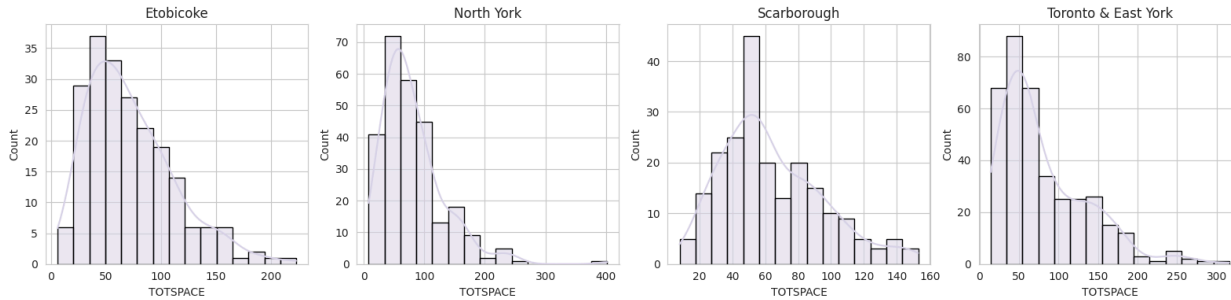

Figure 4: Distribution of Child Care Centre total spaces by Community Council

To address the issue of non-normality in our data, we applied a **square root transformation** (*Figure 5*) to the `TOTSPACE` variable. This transformation aims to balance the distribution by compressing larger values and expanding smaller ones. After transforming the values, we stored them in a new variable called `TOTSPACE_sqrt`. This adjustment aligns our data closer to the assumptions of statistical tests, hopefully enhancing the reliability of our analyses.

COMMUNITY_COUNCIL	TOTSPACE	TOTSPACE_sqrt
ETOBICOKE	164	12.81
TORONTO_EAST_YORK	62	7.87

Figure 5: Table with square root transformation (`TOTSPACE_sqrt`)

After further data wrangling, including outlier removal, we visualised our data with square root transformation, as depicted in *Figure 6*. This resulted in a slight improvement towards normality, although not achieving perfection. A visual comparison between the distributions in *Figure 5* and *Figure 6* shows promising progress. However, to ensure robustness, we will continue to assess normality through various methods throughout our study. Nonetheless, this sets us on a promising path for later analysis to formulate compelling research questions, as detailed later in this report.

Note: While the square root transformation applied to the `TOTSPACE` variable aims to address non-normality and align the data with the assumptions of statistical tests, we acknowledge that this transformation may influence the interpretation of results. Transforming the data alters the scale and distribution, which can impact the significance of findings. Therefore our interpretations will be made with caution by considering the implications of the transformation on the overall analysis.
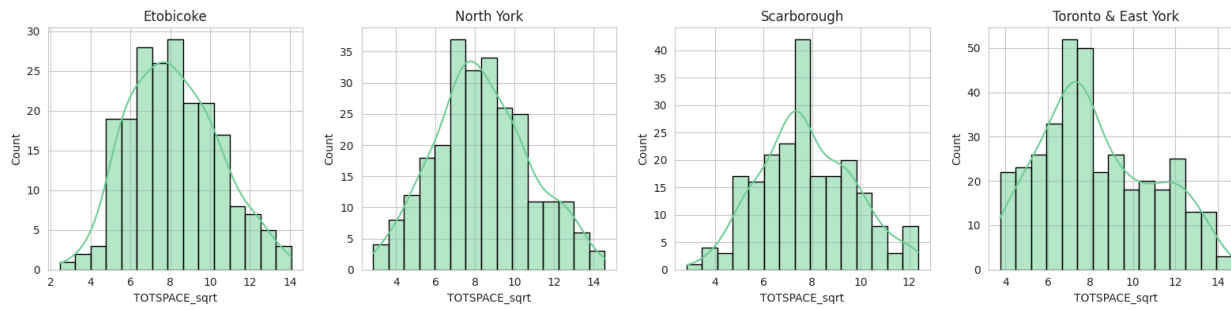
Figure 6: Distribution of Child Care Centre total spaces by Community Council (**after square root transformation**)

4.   Available Child Care space across Community Councils

Research Question #1: How do the quantities of available child care spaces differ across Ontario's Community Councils, and do these differences hold statistical significance?

To explore this question, we analysed the variation in the number of Child Care centres across Community Councils. In *Figure 7*, illustrating this disparity, we observe a higher count in the *Toronto & East York* community, surpassing its counterparts by **at least 100**. We speculate that this discrepancy might stem from the **denser population** in this area compared to others. *Scarborough* and *Etobicoke* exhibit relatively comparable numbers. Despite these observations, further analysis is required to understand how space availability differs across these regions. This exploration could provide valuable insights into where to prioritise the implementation of effective Child Care benefits, particularly in areas of higher demand.
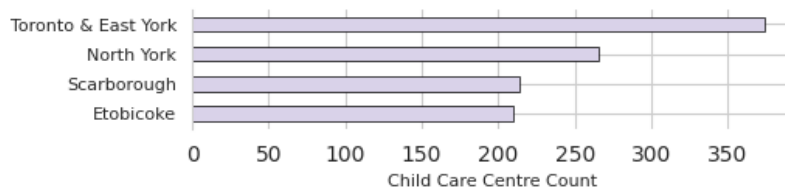

Figure 7: Count of Child Care Centres by Community Council

Next, to address the core of our research question, we conducted a **one-way ANOVA** statistical test to assess the significance of the differences in total space (`TOTSPACE_sqrt`) among different Community Councils. With an alpha level set at 0.05, and a recorded $p-value = 0.04$ (as shown in *Figure 8*), our findings indicate a statistically significant difference in Child Care space across the Community Councils.

	df	sum_sq	mean_sq	F	PR(>F)
COMMUNITY_COUNCIL	3.0	46.57	15.52	2.76	**0.0413**
Residual	1041.0	5863.27	5.63	NaN	NaN

Figure 8: One-Way ANOVA Results

Before proceeding with a post-hoc test to identify potential pairwise statistical differences between the Community Councils, we conducted a thorough review of our model assumptions and conditions, summarised as follows recognizing that our data satisfies the **assumption of independence**:
1. **Normal Distribution:** initially, we applied a square root transformation to our data, resulting in a somewhat normal distribution. Shapiro-Wilk tests confirmed the normal distribution of our dataset.
2. **Normal Distribution of Residual:** while our Q-Q plot (*Figure 9*) and our distribution of our residuals (*Figure 10)* show somewhat promising results considering the skewness of the original dataset, further

examination with Shapiro-Wilk tests on residuals indicated a significant departure from normality, **violating this assumption** (the p-value was far below our $\alpha = 0.05$ threshold).
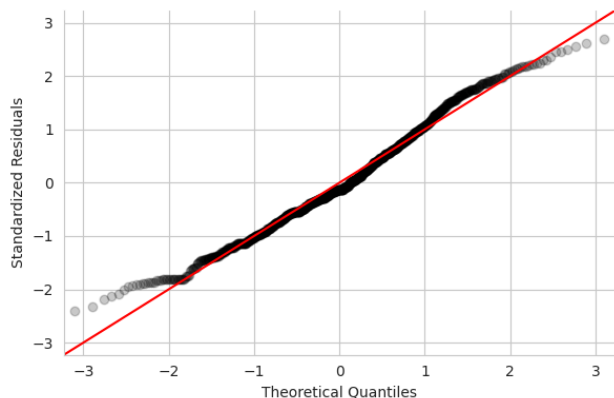

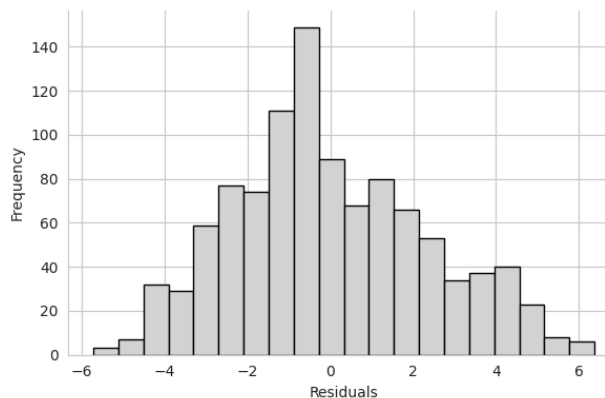Figure 9: Residuals Quantile-Quantile Plot

Figure 10: Residuals Histogram Plot

3. **Homogeneity of Variance:** We ran both Bartlett's test (assuming normality) and Levene's test (assuming non-normality) which both, again, yielded a p-value far below our $\alpha = 0.05$ threshold. **This finding violates the assumption of equal variance in our dataset.**

Review & Results:

Given these considerations, we chose to employ **Welch's one-way ANOVA**, a method robust to **violations of homogeneity of variances** and independent of the assumption of normality. Upon conducting our analysis, we obtained a $p - value = 0.013$ (with $\alpha = 0.05$), indicating statistical differences among Ontario's Community Councils regarding space availability in Child Care Centers.

Next, we proceeded to conduct a **Games-Howell test** as our post-hoc analysis as it's well-suited for situations **where these assumptions are not met**. Our analysis revealed significant differences between North York and Scarborough ($p - value = 0.02$) and between Scarborough and Toronto & East York ($p - value = 0.04$), indicating variations in Child Care space among Community Councils, with **Scarborough** as the region of interest.

A	B	mean(A)	mean(B)	diff	se	T	df	p-value	hedges
NORTH_YORK	SCARBOROUGH	8.36	7.78	0.58	0.20	2.91	469.20	**0.02**	0.26
SCARBOROUGH	TORONTO_EAST_YORK	7.78	8.29	-0.51	0.19	-2.66	553.17	**0.04**	-0.21

Figure 10: Residuals Histogram Plot

5. Child Care Centre Availability: Community Councils vs. Child Care Age Clusters

Research Question #2: Are there significant interactions between the types of Child Care clusters (such as infant, toddler, preschooler, etc.) and Community Councils in determining the provision of child care services?

In this study, we aimed to investigate the impact of **two categorical independent variables**, namely `COMMUNITY_COUNCIL` and `CHILD_CARE_CLUSTER` (i.e., infant, toddler, preschooler, kindergarteners, etc.), on the availability of space in child care facilities, represented by the **dependent variable** `AVAILABLE_SPACE_PF`.

Our approach was as follows:

1. **Square Root Transformation** was applied to `AVAILABLE_SPACE_PF` to get close to normal distribution.
2. **NaN values** in `AVAILABLE_SPACE_PF` were imputed using the mean value of the associated `CHILD_CARE_CLUSTER.`
3. **Assumptions** and **Conditions,** in this analysis, followed the same logic described in *Section 4*.
4. **Methodologies** involved two-way ANOVA and Tukey's tests to explore the effects of `COMMUNITY_COUNCIL` and `CHILD_CARE_CLUSTER` on `CHILD_CARE_CLUSTER`, including their interaction (*Fig. 11* and *12*).
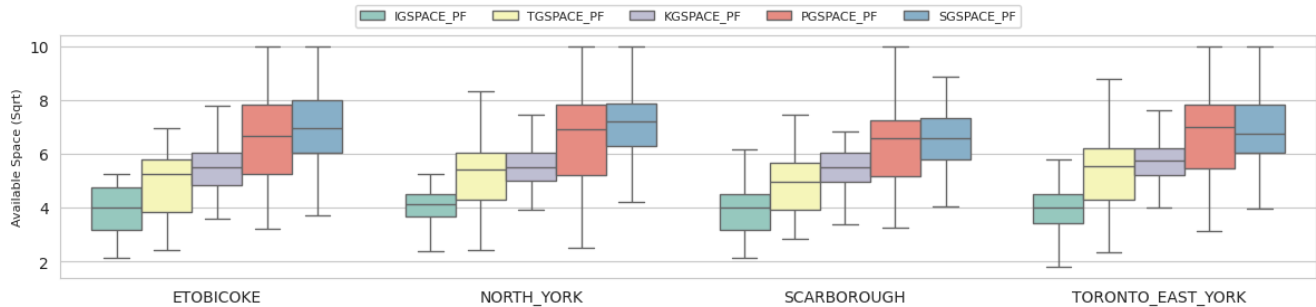

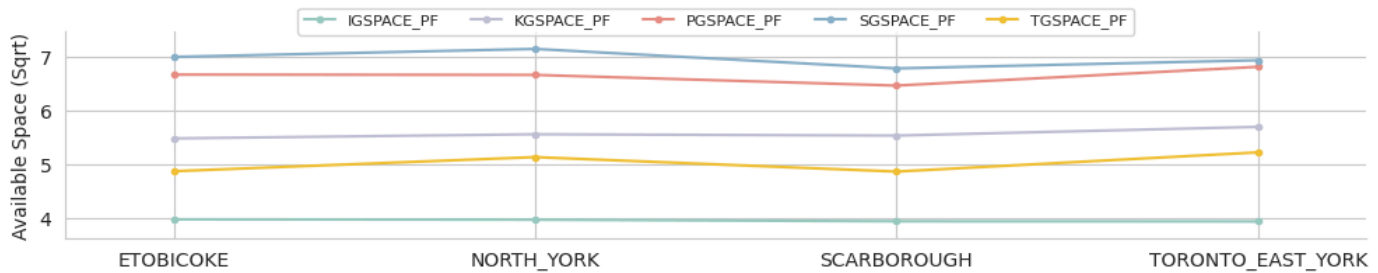Figure 11: Boxplot of Child Care Clusters spaces across Community Councils

Figure 12: Interaction Plot of Community Council and Child Care Cluster

Results:

The two-way ANOVA revealed **significant main effects** for both `COMMUNITY_COUNCIL` $(p-value = 0.04)$ and `CHILD_CARE_CLUSTER` $(F = 1625.89$ and $p-value$ close to 0) with $\alpha = 0.05$, indicating that there are differences in available space across different community councils and child care clusters. However, the non-significant interaction effect $(p-value = 0.34)$, suggests that space availability's relationship with community councils remains **consistent across child care clusters**.

Tukey's HSD test results revealed **significant differences** in `AVAILABLE_SPACE_PF` between Scarborough and Toronto & East York $(p-value = 0.003)$. Again, similar to our first study, here, **Scarborough resurfaces as a region of concern**. Our second Tukey's HSD test results unveiled significant disparities in `AVAILABLE_SPACE_PF` among all the child care clusters – i.e., `IGSPACE_PF, TGSPACE_PF, PGSPACE_PF, KGSPACE_PF, SGSPACE_PF` with $p-value = 0.001$ (with $\alpha = 0.05$) across all combinations pairwise comparisons.

6.   Conclusion

Our analysis pointed to **significant disparities in child care access** across Ontario, focusing particularly on **Scarborough** as a key area needing targeted interventions to address regional inequities. Our findings can encourage policymakers to pinpoint areas that require resource allocation to address disparities in child care. This study contributes to the ongoing dialogue on enhancing child care provision and ensuring every child in Ontario has access to **safe, affordable, and high-quality early learning experiences**. However, although robust to deviations, our model has **limitations** (square root interpretation). Additional data and further analysis could potentially yield better strategic decisions and recommendations to enhance child care access across Ontario.