Unraveling the Enigma of Child Care Spaces

INF2178 - Technical Assignment 2
Fatima Ashfaq - 1010784732
Professor Shion Guha
March 9, 2023

Link to ipynb code file:
https://colab.research.google.com/drive/1fI7ytS1VHCY1O1SNOvpiCdLLE4vwJvUI?usp=sharing

1. Introduction

The investigation into the dynamics of child care spaces was initiated with a clear purpose: to elucidate the nuanced patterns underlying child care provision across varying auspices. This endeavor was driven by a series of meticulously crafted research questions aimed at discerning the impact of CWELCC (Canada-Wide Early Learning and Child Care) participation, subsidy status of centers, and their potential interaction on the distribution of child care spaces.

2. Research Questions

The research questions examined and tested in the statistical analyses include

1. Auspice variation: How do child vare spaces differ among different auspices?
2. CWELCC and Subsidy Influence - Is there a discernible difference in child care spaces concerning CWELCC participation and subsidy status?
3. Interaction Effects: Do interaction effects manifest between CWELCC participation and subsidy status regarding child care spaces?

3. Exploratory Data Analysis (EDA)

In any statistical analysis, a crucial first step is to thoroughly explore the dataset to gain insights into its structure, distribution, and potential patterns. In this analysis, both non-graphical and graphical exploratory analysis was conducted to understand the dataset's characteristics fully.

3.1 Non-Graphical EDA

Descriptive statistics play a fundamental role in unravelling the essence of a dataset, offering insights into its central tendencies, variabilities, and ranging. The comprehensive set of statistics computed for each numerical variable in the dataset as shown in *table 3.1* provide a robust understanding of the distribution of child care spaces across the different age groups and aid in the identification of potential outliers.

Upon conducting the non-graphical EDA, profound insights emerged regarding the distribution of child care spaces across various age groups. Analysis of the descriptive statistics reveals significant variations in child care space availability, underscoring the importance of understanding these nuances. For instance, the mean and median values for each age group offer valuable insight into the typical capacity offered, while the range between the minimum and maximum values shed light on the extent of variability within each category.

	IGSPACE	TGSPACE	PGSPACE	KGSPACE	SGSPACE	TOTSPACE
Count	1063	1063	1063	1063	1063	1063
Mean	3.90	11.60	24.26	14.26	21.66	75.67
Std	6.09	12.09	18.58	20.49	30.42	47.82
Min	0.00	0.00	0.00	0.00	0.00	6.00
25%	0.00	0.00	16.00	0.00	0.00	43.00
50%	0.00	10.00	24.00	0.00	0.00	62.00
75%	10.00	15.00	32.00	26.00	30.00	97.00
Max	30.00	90.00	144.00	130.00	285.00	402.00

Table 3.1 - Descriptive Statistics Computed For Each Numerical Variable

By examining these statistics, it becomes evident that certain age groups exhibit wider spreads in child care space availability compared to others, highlighting potential areas for further investigation and analysis.

3.2 Graphical EDA

Visualisations play a crucial role in uncovering patterns, trends, and relationships within the data. In this analysis, various graphical techniques are employed to provide a comprehensive exploration of the dataset.

Histograms are utilised to visualise the distribution of child care spaces for each age group. Figure 3.2 represents the total distribution of all the child care spaces for all age groups, which like all other histograms in the code file displays a unimodal distribution skewed to the right.

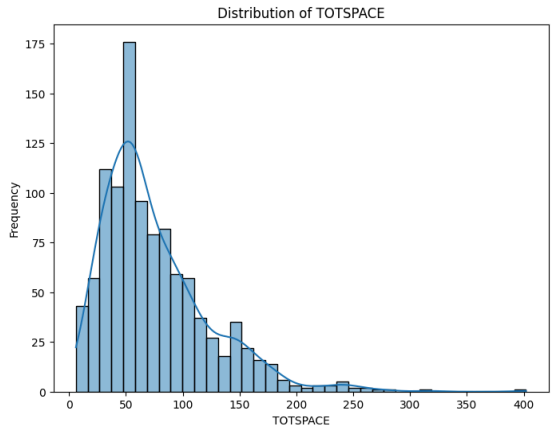

Figure 3.2 - Histogram for TOTSPACE (Child care spaces for all age groups)

Furthermore, bar charts are employed to visualise the distribution of the dataset's categorical variables such as auspice, subsidy status, and CWELCC participation. These visualisations offer a quick overview of the frequency of each category within the variables, enabling the identification of imbalances or dominant categories within the dataset. The bar chart for the auspice, figure 3.3, illustrates that non-profit agencies have the highest child care spaces with approximately 700 instances in the dataset, and public city operated agencies have the lowest with only 100 instances (approximately) in the dataset. In addition, figure 3.4, displays the bar chart for the frequency of subsidy, and the centers that are subsidized demonstrate a higher care space than those that do not receive subsidy. Figure 3.5 summarises that CWELCC participating centers have a higer child care capacity than those that do not. Finally, figure 3.6 highlights that preschoolers (PGSPACE) age group has the highest number of mean child care spaces among all other groups.

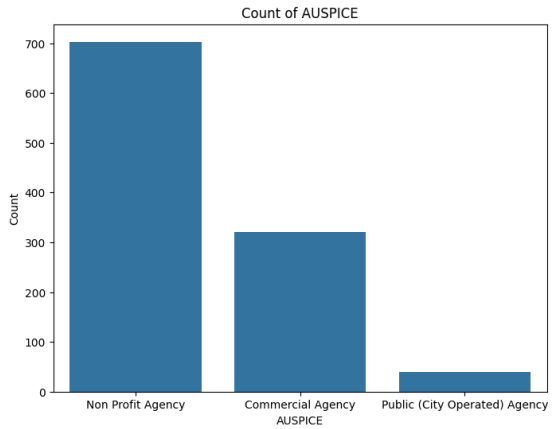

Figure 3.3 - Bar Chart of Auspice Variable

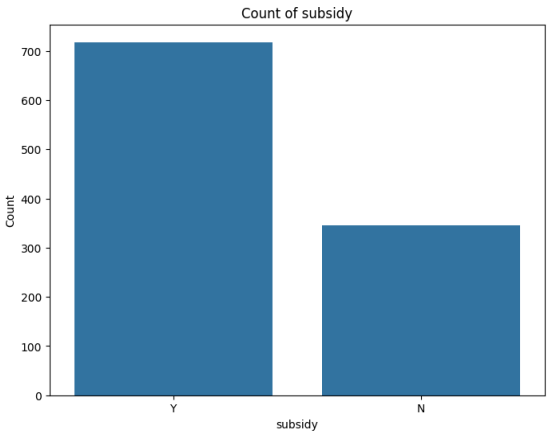

Figure 3.4 - Bar Chart of Subsidy Variable

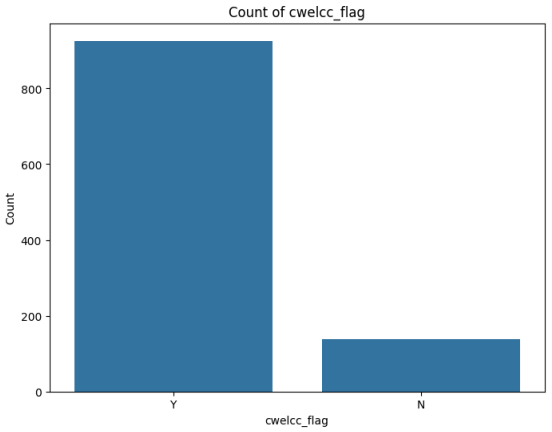

Figure 3.5 - Bar Chart of CWELCC Variable

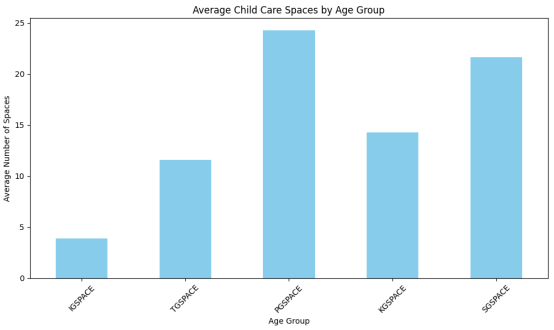

Figure 3.6 - Average Child Care Spaces by Age Group

Boxplots serve as effective tools for visualising the distribution of child care capacity across different asupsices and subsidy statutes as seen in figure 3.6. Non-profit agencies, while dominating the child-care capacity is also displaying the most outliers in both, subsidised and not subsidised conditions. The public city operated agencies does not present any outliers exceeding 1.5 times the upper and lower quartile unlike the other auspice categories. Furthermore, figure 3.8 demonstrates the distribution of child care spaces among the various age groups, which as previously discussed, highlights the significantly high capacity in the preschooler age group (PGSPACE) as seen in the close and tight spread of the strip plot points on the boxplot.

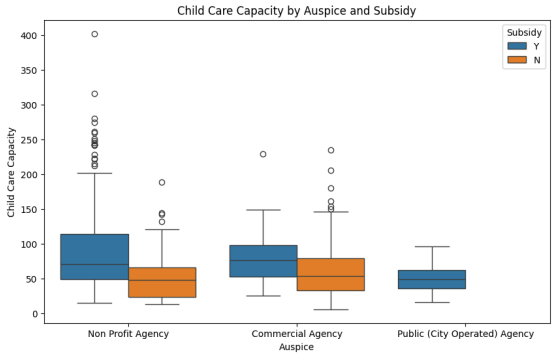

Figure 3.7 - Box Plot of Child Care Capacity by Auspice and Subsidy

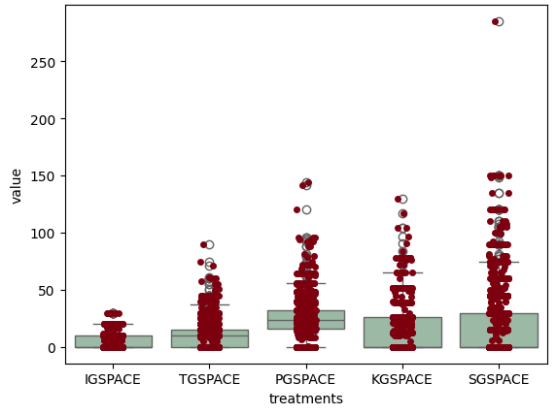

Figure 3.8 - Box Plot of Distribution of Child Care Spaces Across Different Age Groups

4. One-Way ANOVA

A One-Way Analysis Of Variance (ANOVA) was conducted to elucidate the impact of auspice on child care spaces. The hypotheses of the test are as follows;

1. Null Hypothesis (H0): The mean child care spaces are equal across all auspices.
2. Alternative Hypothesis (H1): At least one of the mean child care spaces significantly differs from the others across auspices.

Based on the one-way ANOVA results (table 4.1) with a p-value of 4.517383e-151, which is significantly lower than the predetermined significance level of 0.001, the null hypothesis is rejected. Therefore, it can be concluded that there is sufficient evidence to suggest that at least one oof the mean child care spaces significantly differs from the others across auspices.

Source	Degrees of Freedom	Sum of Squares	Mean Square	F-Value	P-Value
C(treatments)	4.0	2.821233e+05	70530.816839	188.190768	4.517383e-151
Residual	5310	1.990101e+06	374.783617	NaN	NaN

Table 4.1 - One-Way ANOVA Results Table

A TukeyHSD post hoc test was conducted to examine the significance of differences between groups in child care space availability. The results, detailed in Table 4.2, reveal p-values below 0.05 for all group comparisons, indicating significant differences in child care space availability between groups, with notable variations across comparisons ($p < 0.05$).

Group 1	Group 2	Mean Difference	95% CI (Lower)	95% CI (Upper)	Q-value	P-value
IGSPACE	TGSPACE	7.70	5.41	9.99	12.97	0.001
IGSPACE	PGSPACE	20.36	18.07	22.65	34.29	0.001
IGSPACE	KGSPACE	10.36	8.07	12.65	17.45	0.001
IGSPACE	SGSPACE	17.76	15.47	20.06	29.92	0.001
TGSPACE	PGSPACE	12.66	10.37	14.95	21.32	0.001
TGSPACE	KGSPACE	2.66	0.37	4.95	4.48	0.014
TGSPACE	SGSPACE	10.06	7.77	12.35	16.94	0.001
PGSPACE	KGSPACE	10.00	7.71	12.29	16.84	0.001
PGSPACE	SGSPACE	2.60	0.31	4.89	4.37	0.017
KGSPACE	SGSPACE	7.40	5.11	9.69	12.47	0.001

Table 4.2 - TukeyHSD Post Hoc Test For One-Way ANOVA

Furthermore, the model diagnostics were checked to see if the ANOVA assumptions were satisfied. The first assumption - residuals are normally distributed - was checked through the creation of a QQ plot (figure 4.3), histogram (figure 4.4) and Shapiro Wilk test (table 4.5). The QQ plot clearly shows a non-normal distribution of the residuals as it displays a ‘U” shaped pattern. The histogram of the residuals further solidifies this result as it is highly skewed to the right. Finally, the Shapiro Wilk test produces a p-value of less than 0.001 ($p < 0.001$) which once again provides evidence into the violation of the first assumption.

The second assumption - variances are homogeneous - was checked using a Levene’s test instead of Bartlett’s test as the data is not normally distributed. This assumption was also found to be violated by the model as the test produced a p-value of less than 0.001 ($p < 0.001$).

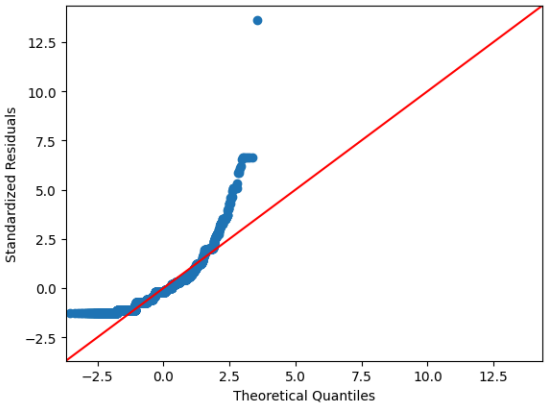

Figure 4.3 - One-Way ANOVA Residual QQ Plot

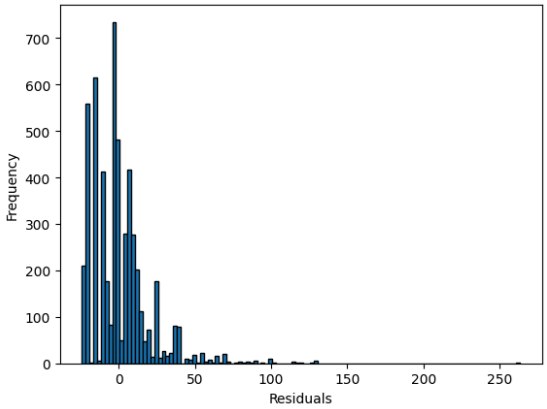

Figure 4.4 - One-Way ANOVA Residual Histogram

Parameter	Value
Test statistics (W)	142.6228
Degrees of freedom	4.0000
p-value	0.0000

Table 4.5 - Shapiro Wilk Test Results

5. Two-Way ANOVA

A two-way ANOVA was conducted on the data to explore the effects of CWELCC participation, subsidy status, and their interaction on the total child care spaces available. The statistical model that was employed considers both main effects and the interaction effect between CWELCC participation and subsidy status. The hypotheses being tested are as follows:

1. ***Null Hypothesis 1 (H0 1)***: There is no significant difference in the child care spaces according to the space's participation in CWELCC.
2. ***Null Hypothesis 2 (H0 2)***: There is no significant difference in the child care spaces according to the subsidy fee contract status of the center.
3. ***Null Hypothesis 3 (H0 3)***: There are no interaction effects.

To determine whether to reject or fail to reject the null hypotheses based on the ANOVA table (table 5.1), the p-values associated with each main factor and interaction term are examined as shown below:

1. The first null hypothesis revolving around CWELCC participation is not rejected as the p-value of the for the CWELCC factor is approximately 0.075 which is greater than the predetermined significant level ($p > 0.05$). There is no significant difference in child care spaces based on CWELCC participation.
2. The second null hypothesis on the subsidy status and its effect on child care spaces is rejected as the p-value for the subsidy factor is less than 0.001 ($p < 0.001$) indicating sufficient evidence against H02 as there is a significant difference in child care spaces based on subsidy fee contract status.
3. The third null hypothesis - no interaction effects - is rejected as the p-value for the interaction term is approximately 0.00272 which is less than the 0.05 significance level ($p < 0.05$). There is a significant interaction effect between CWELCC participation and subsidy status on total child care spaces.

Factor	Degrees of Freedom	Sum of Squares	F-Statistic	P-Value
CWELCC Participation	1	6723.429	3.176	7.499648e-02
Subsidy Status	1	98161.81	46.375	1.633653e-11
CWELCC*Subsidy Interaction	1	19108.68	9.028	2.721895e-03
Residual	1059	2241580	NaN	NaN

Table 5.1 - Two-Way ANOVA Results Table

Tukey's HSD pairwise post hoc test (table 5.2) was conducted to further explore the significant interaction effect. It was found that centers with both CWELCC participation (Y) and a subsidy fee contracts (Y) have a significantly higher mean difference in child care spaces compared to those with CWELCC participation but without subsidy contracts (N) with a p-value of less than 0.001 ($p < 0.001$). Similarly, centers with both CWELCC participation (Y) and no subsidy contracts (N) exhibit a significantly higher mean difference in child care spaces compared to those without CWELCC participation and

without subsidy contracts (N,N) with a p-value < 0.001. Furthermore, there is also a significant mean difference in child care spaces between centers with CWELCC participation and those without, regardless of subsidy contract status. However, there are no significant differences in child care spaces between centers with subsidy contracts (Y) and those without (N), irrespective of CWELCC participation status. Overall, these results suggest that the combined influence of CWELCC participation and subsidy contract status significantly effects the total child care spaces available in centers.

Group 1	Group 2	Difference	Lower Bound	Upper Bound	Q-Value	p-Value
(Y, Y)	(Y, N)	29.13	14.09	44.17	7.05	0.001
(Y, Y)	(N, Y)	30.32	-9.03	69.67	2.81	0.195
(Y, Y)	(N, N)	40.49	22.08	58.90	8.01	0.001
(Y, N)	(N, Y)	59.45	18.03	100.87	5.23	0.001
(Y, N)	(N, N)	11.36	-11.13	33.85	1.84	0.555
(N, Y)	(N, N)	70.81	28.05	113.57	6.03	0.001

Table 5.2 - Tukey HSD Post Hoc Test Results for Two-Way ANOVA

The interaction plots visually explore the interplay between CWELCC participation, subsidy status, and their combined effect on the total child care spaces available. In Figure 5.3, the interaction plot of auspice and subsidy reveals that public city-operated agency centers exhibit the lowest care capacity, with or without subsidy, with no child care spaces available in public centers without subsidy. Non-profit agencies, particularly when receiving subsidy, demonstrate the highest capacity among the three types, with approximately 90 spaces, significantly lower (approximately 50 spaces) when not subsidized. Commercial agencies has the highest average capacity among centers without subsidy, approximately 60 spaces, but the second lowest when subsidized. Similarly, Figure 5.4, illustrating the interaction plot of auspice and CWELCC participation, exhibits analogous patterns to the interaction plot of auspice and subsidy on total child care capacity. However, Figure 5.5, depicting the interaction plot of CWELCC participation and subsidy status, shows two lines mirroring each other. Centers with CWELCC participation and subsidy display lower spaces than those with subsidy but no CWELCC participation. In contrast, centers without subsidy participating in CWELCC exhibit a higher space capacity, albeit lower than the other line, compared to centers without CWELCC participation or subsidy.

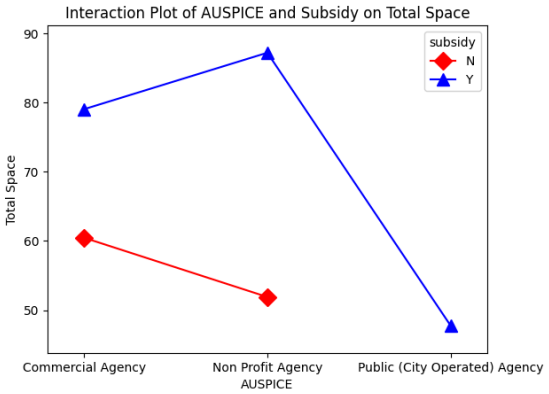

Figure 5.3 - Interaction Plot of Auspice and CWELCC on Total Child Care Space

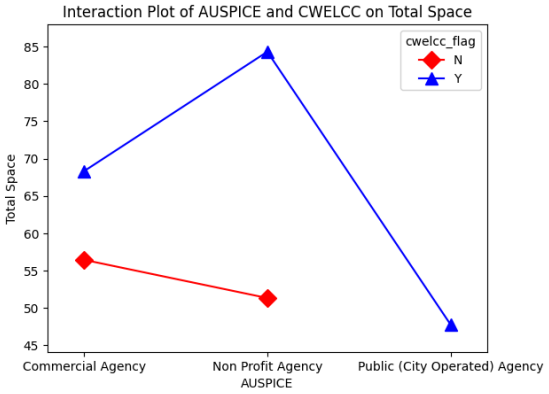

Figure 5.4 - Interaction Plot of Auspice and CWELCC on Total Child Care Spaces

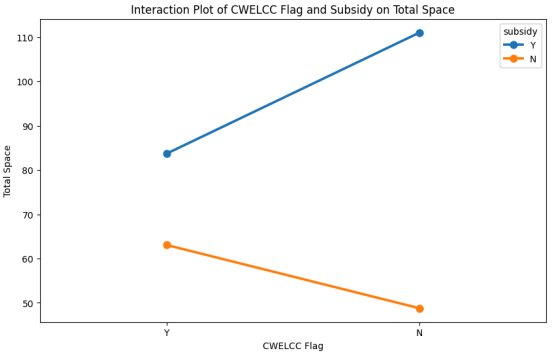

Figure 5.5 - Interaction Plot of CWELCC and Subsidy on Total Child Care Spaces

Similar to the one-way ANOVA, assumption checks were performed to ensure the validity and reliability of the two-way ANOVA result and help maintain the integrity of the overall study. The assumptions remain the same as those in the one-way ANOVA, and the same methods were used to check them. Similar to the prior check in the one-way anova, the QQ plot (figure 5.6) and histogram (figure 5.7) of the residuals indicate deviation from normality. This violation of assumption of normality is further solidified by the Shapiro-Wilk test which produced a statistically significant result with a p-value less than 0.001 hence warranting a cautious interpretation of the ANOVA outcomes. Levene's test

was used to check whether the assumption of homogeneity of variances was satisfied. The test produced a statistically significant evidence of unequal variances with a p-value of 0.002 ($p < 0.05$), reinforcing the need for careful consideration when interpreting the ANOVA results.

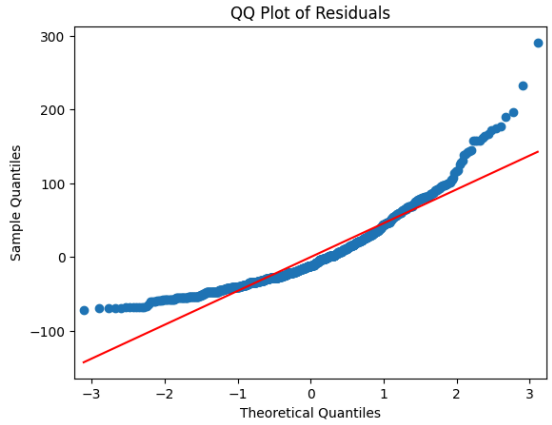

Figure 5.6 - QQ Plot of Two-Way ANOVA Residuals

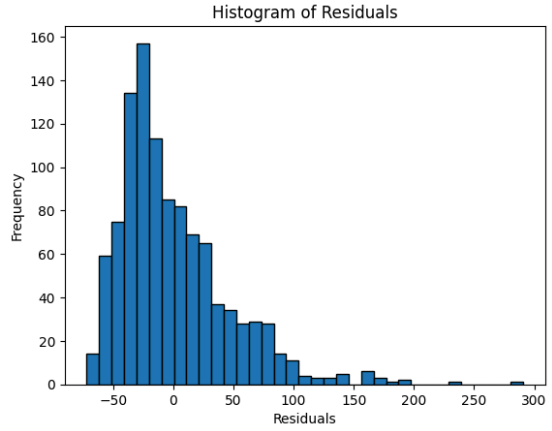

Figure 5.7 - Histogram of Two-Way ANOVA Residuals

6. Limitations

The inability to meet the assumptions of the ANOVA model poses a notable limitation to this analysis. While ANOVA is robust to moderate violations of assumptions, severe departures from normality and homogeneity of variances can lead to inflated type 1 error rates and compromised statistical validity. As such, caution must be exercised in the interpretation of the ANOVA results, and alternative analytical approaches such as non-parametric tests or transformations of the data, may be warranted to mitigate the impact of these violations on the conclusions drawn from this study.

7. References

10.2.1 - ANOVA Assumptions | STAT 500. (n.d.). PennState: Statistics Online Courses. https://online.stat.psu.edu/stat500/lesson/10/10.2/10.2.1#:~:text=Assumptions%20for%20One%2DWay%20ANOVA,The%20data%20are%20independent.

Frost, J. (2023, October 26). *Using Post Hoc Tests with ANOVA*. Statistics by Jim. https://statisticsbyjim.com/anova/post-hoc-tests-anova/

Bedre, R. (2023, November 12). *How to Perform ANOVA in Python*. RS Blog. https://www.reneshbedre.com/blog/anova.html

https://scikit-posthocs.readthedocs.io/_/downloads/en/stable/pdf/

Two-Way ANOVA in python, Tukey's HSD test fo…

30. Two Way ANOVA in Python || Dr. Dhaval Ma…