Name: Yun Zhou (1009709442)
Assignment 2: INF2178 Technical Assignment 2

Explore Toronto licensed childcare centres

1. Introduction

The dataset focused on licensed childcare centers in Toronto, specifically analyzing operational aspects and capacity for various age groups. This dataset catalogs detailed information about child care centers, including their demographic profiles, capacities, and locations.

The exploration will address the following research questions, aims to understand the impact of different variables on childcare center capacities, utilizing statistical methods and visualizations to extract insights from the data.

Data preparation: First create a new Data Frame, then converting “AUSPCIE” & ‘subsidy’ to categorical variables. 'AUSPICE' contains values that categorize childcare centers by their managing authority (e.g., Non-Profit, For-Profit). 'TOTSPACE' was already a numeric type.

2. Data story

The raw dataset has a total of 17 columns with 1063 rows, presenting information on childcare centers, encapsulating various characteristics such as the type of auspice (e.g., Non-Profit Agency), related to location and facility type.

From initial observations, the dataset reveals a range of childcare providers, from non-profit agencies to other types of management. This diversity suggests a multifaceted approach to childcare, each with its own set of benefits, challenges, and operational models. Also, the inclusion of a subsidy indicator (yes/no) in the dataset hints at the financial accessibility of these centers to the broader population. Furthermore, there is capacity variations, and the dataset allows for analysis of how childcare availability differs by region, auspice type, and subsidy status, potentially indicating disparities in access or demand.

To delve deeper, below are the research questions that this dataset can help answer:

Research Question 1: Whether there's a significant difference in the number of total spaces ('TOTSPACE') across different categories of childcare centers ('AUSPICE') and/or subsidy status ('subsidy').
Research Question 2: Does data meet assumptions required for conducting ANOVA?
Research Question 3: Is there an impact of two categorical independent variables, 'AUSPICE' and 'subsidy', and their interaction on the dependent variable 'TOTSPACE’?
Research Question 4: Which specific pairs of childcare agency types and subsidy statuses differ significantly in the average number of spaces provided?
Research Question 5: Do data meet assumptions required for conducting two-way ANOVA?

3. Data analysis

Research Question #1: Do different types of childcare agencies ('AUSPICE') or subsidy status vary in the total number of spaces ('TOTSPACE') they offer?

To do this, we examine the 'AUSPICE' variable to compare the mean 'TOTSPACE' across different categories Commercial Agency, Non-Profit Agency, and Public Agency. Then performs **one way ANOVA** test on "TOTSPACE" based on categorical variable 'AUSPICE' and "subsidy', to determine if these differences are statistically significant.

ANOVA test on 'AUSPICE': The total variation attributed to the difference between auspice types is approximately 96,112. The residual variation is about 2,332,065. There are 2 degrees of freedom for the auspice types (since there are three categories - 1) and 1060 degrees of freedom for the residuals. The F-statistic is 21.843, this measure of how much the group means vary relative to the variation within the groups. The p-value is extremely small (5.06e-10), which indicates that there is a statistically significant difference between the mean 'TOTSPACE' of at least two of the auspice categories.
Further is to perform Tukey's HSD **post-hoc tests** to identify above significant differences, by comparing every pair of groups within the 'AUSPICE' variable. The summary table is the result from Tukey's HSD test, it shows a significant mean difference of 17.119 between Non-Profit Agency and Commercial Agency in the number of 'TOTSPACE', with a very low p-value (0.001), indicating this is likely a true difference and not due to random chance. The mean difference between Non Profit Agency and Public (City Operated) Agency is 34.33, which is also significant (p-value 0.001), suggesting a substantial difference in 'TOTSPACE' between these two types of agencies. The comparison between Commercial Agency and Public (City Operated) Agency shows a mean difference of 17.21, but the p-value is 0.08, which is above the 0.05 threshold for significance. This means that while there is a difference, we cannot be as confident that it isn't due to chance. Overall, there significant differences in the total number of spaces ('TOTSPACE') between Non Profit Agencies and both Commercial Agencies and Public (City Operated) Agencies, with Non Profit Agencies having more spaces on average. The difference between Commercial and Public Agencies is not statistically significant at the 0.05 level, therefore no substantial difference in the total spaces offered between these two types of agencies.

ANOVA test on 'subsidy': The variation is attributed to the difference between subsidized and not subsidized is approximately 160,765, with the residual variation being around 2,267,412. There is 1 degree of freedom for the subsidy status (since there are two categories - 1) and 1061 degrees of freedom for the residuals. The F-statistic is 75.23, indicating a strong relationship between subsidy status and 'TOTSPACE'. The p-value is even smaller than for 'AUSPICE' (1.550892e-17), strongly suggesting that there is a statistically significant difference between the mean 'TOTSPACE' for subsidized versus non-subsidized centers.
Following the same **post-hoc test** on the **'TOTSPACE'**, result shows a statistically significant mean difference of 26.265831 in 'TOTSPACE' between subsidized (Y) and not subsidized (N) groups. The 95% confidence interval for this difference is between 20.32 and 32.208, since does not include 0 therefore the difference is significant. The q-value is quite large at 12.266, the p-value is very small (0.001), (p-value<0.05), the difference is statistically significant.

Conclusion: Overall significant differences in the average number of total spaces available at childcare centers based on their auspice type and whether they are subsidized.

ANVOA boxplot:
Further, to **visualize data distribution of 'AUSPICE'**, I generate a box plot of the 'TOTSPACE' values grouped by the different types of 'AUSPICE'. The x parameter is the categorical variable and the y parameter sets the numerical variable ('TOTSPACE'). From boxplot **result,** we know that Non-Profit Agencies tend to have a wider range of 'TOTSPACE' compared to Commercial and Public agencies, with a higher median and more outliers. This suggests that Non-Profit Agencies vary more in size and have several centers with a large number of total spaces. Commercial Agencies have fewer outliers and a lower median, suggesting more uniformity and generally fewer spaces available. Public (City Operated) Agencies have the smallest range and median, indicating the most consistency in the number of spaces offered, which tend to be fewer than the other types of agencies.

To **visualize distribution of 'subsidy'**, same logic as above**.** This box plot has two categories: 'N' (not subsidized) and 'Y' (subsidized). The **boxplot result** shows that ('Y') generally have a higher median number of spaces available than ('N'), with a greater range and more outliers. This indicate subsidized centers are larger on average, with some having a very high capacity, whereas non-subsidized centers are smaller on average with less variability in size.

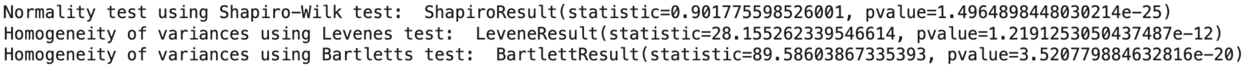
Research Question 2: Does data meet assumptions required for conducting ANOVA? Use two **plots**, Shapiro-Wilk, Levene's, and Bartlett's tests to validate AVOVA result.

1. **Q-Q Plot**: the plot result shows a curve, indicating that the residuals have a distribution that deviates from normality, especially at the lower and higher ends (evident from the points deviating from the red line).
2.**Histogram**: shows a right-skewed distribution of residuals, as the longer tail on the right side of the histogram. This means larger positive residuals than expected in a normal distribution. So the assumption of normally distributed residuals might be violated. This can impact the validity of the ANOVA results since ANOVA assumes that the residuals are normally distributed.

Check assumptions for 'subsidy':
3.**Shapiro-Wilk Test**: to check whether the residuals from the ANOVA model are normally distributed, W (test statistic) is 0.893, p-value is very small (1.424e-26), lower than alpha 0.05. Therefore, the residuals do not follow a normal distribution.
4.**Levene's Test**: W (test statistic) is 22.989, p-value is very small (1.862e-06), indicating that the variances of 'TOTSPACE' are not equal between the subsidized and non-subsidized groups.
5.Bartlett's Test: W (test statistic) is 49.082, p-value is very small (2.455e-12), also suggesting that the variances are not equal.
Overall, the above results indicate the assumptions of normality of residuals and homogeneity of variances are violated for the variable 'subsidy' of its effect on 'TOTSPACE'. This means the data may not meet the necessary conditions for conducting a traditional ANOVA. Given these violations, it is recommended to apply a different statistical method that doesn't assume

normality and equal variances (like a non-parametric test), or use a more robust form of ANOVA that can handle these issues for further analysis.

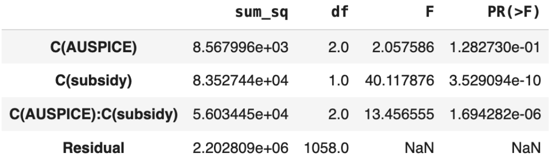
Check the assumptions for 'AUSPICE' variable's effect on 'TOTSPACE':



```
Normality test using Shapiro-Wilk test:  ShapiroResult(statistic=0.901775598526001, pvalue=1.4964898448030214e-25)
Homogeneity of variances using Levenes test:  LeveneResult(statistic=28.155262339546614, pvalue=1.2191253050437487e-12)
Homogeneity of variances using Bartletts test:  BartlettResult(statistic=89.58603867335393, pvalue=3.520779884632816e-20)
```

The test statistic from Shapiro-Wilk Test is 0.9018. The p-value is 1.496e-25 (much less than alpha 0.05), thefore the residuals do not follow a normal distribution. The test statistic is 28.155 in Levene's Test and p-value of 1.219e-12 (< 0.05), the assumption of homogeneity of variances is not met. Bartlett's Test shows test statistic value is 89.586 and p-value is 3.521e-20, again indicating that the variances are not equal across the groups. The results of above three tests all indicate significant p-values, meaning that the normality of residuals and homogeneity of variances assumptions are violated for the 'AUSPICE' variable in the context of its effect on 'TOTSPACE'.

Research Question 3: Is there an impact of two categorical independent variables, 'AUSPICE' and 'subsidy', and their interaction on the dependent variable 'TOTSPACE’?



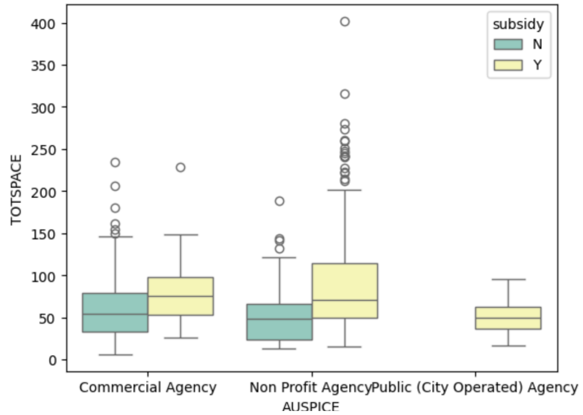
	sum_sq	df	F	PR(>F)
C(AUSPICE)	8.567996e+03	2.0	2.057586	1.282730e-01
C(subsidy)	8.352744e+04	1.0	40.117876	3.529094e-10
C(AUSPICE):C(subsidy)	5.603445e+04	2.0	13.456555	1.694282e-06
Residual	2.202809e+06	1058.0	NaN	NaN

First, per **two-way ANOVA table** output, p-value for the main effect of 'AUSPICE' is 0.128 (p value>0.05), suggesting that differences in 'TOTSPACE' across different 'AUSPICE' categories are not statistically significant when controlling for 'subsidy' and the interaction effect. The p-value for 'subsidy' is extremely small (3.53e-10), indicating a significant difference in 'TOTSPACE' between subsidized and non-subsidized groups. The interaction between 'AUSPICE' and 'subsidy' has a p-value of 1.69e-06, therefore the effect of 'subsidy' on 'TOTSPACE' may differ depending on the 'AUSPICE' category.
Overall, the two-way ANOVA results suggest that 'subsidy' status and the interaction between 'subsidy' status and 'AUSPICE' category have significant effects on 'TOTSPACE'. However, the main effect of 'AUSPICE' category alone does not significantly affect 'TOTSPACE'.

Second, the **interaction plot for 'AUSPICE' and 'subsidy'** shows an upward trend from 'N' to 'Y', indicating that, on average, subsidized centers ('Y') have more spaces than non-subsidized centers ('N'), regardless of the 'AUSPICE' category. The lines are not parallel therefore there is an interaction effect. Also, for Commercial and Non-Profit Agencies -the subsidy has a comparable effect on 'TOTSPACE', The slope for Public (City Operated) Agency is noticeably different, implying the subsidy status does not substantially change the mean 'TOTSPACE' for this category. Therefore, the interaction effect is consistent with two-way ANOVA results, and indicated a significant interaction between 'AUSPICE' and 'subsidy'.

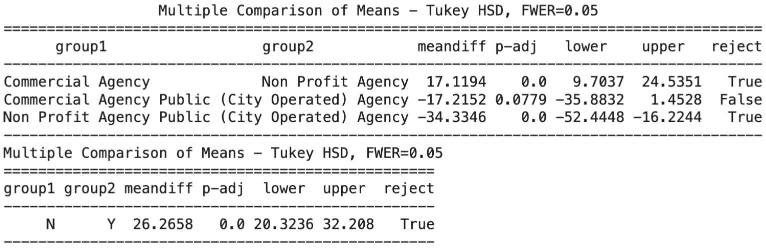
Further, to provide insights into potential interaction effect between 'AUSPICE' and 'subsidy', we create box plot to visualize whether certain types of agencies have more spaces available than others and whether subsidy status is associated with difference in number of spaces.

```
<Axes: xlabel='AUSPICE', ylabel='TOTSPACE'>
```



The two-way grouped box plot as above shows median 'TOTSPACE' for Non Profit Agencies is higher than for Commercial Agencies and Public (City Operated) Agencies. The variability in 'TOTSPACE' (as indicated by the IQR and whiskers) is also greatest among Non Profit Agencies. Subsidized centers (indicated by the lighter color) generally seem to have more 'TOTSPACE' across all types of 'AUSPICE', with Non Profit Agencies showing the most pronounced difference between subsidized and non-subsidized centers. There are many outliers in the Non Profit Agency category, indicating some Non Profit Agencies have a significantly higher number of 'TOTSPACE' than the typical centers. Overall, the 'AUSPICE' type and 'subsidy' status are both factors that affect the total number of spaces available at childcare centers, with subsidies playing a potentially substantial role within each 'AUSPICE' category.

Research Question 4: Which specific pairs of childcare agency types and subsidy statuses differ significantly in the average number of spaces provided? In Two-way ANOVA



```
                 Multiple Comparison of Means - Tukey HSD, FWER=0.05                    
========================================================================================
     group1                   group2             meandiff p-adj   lower    upper  reject
----------------------------------------------------------------------------------------
Commercial Agency             Non Profit Agency  17.1194    0.0   9.7037  24.5351   True
Commercial Agency Public (City Operated) Agency -17.2152 0.0779 -35.8832   1.4528  False
Non Profit Agency Public (City Operated) Agency -34.3346    0.0 -52.4448 -16.2244   True
----------------------------------------------------------------------------------------
Multiple Comparison of Means - Tukey HSD, FWER=0.05
==================================================
group1 group2 meandiff p-adj  lower  upper  reject
--------------------------------------------------
     N      Y  26.2658   0.0 20.3236 32.208   True
--------------------------------------------------
```

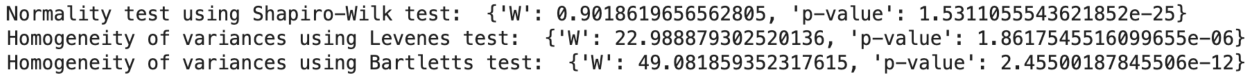
Post-hoc tests for 'AUSPICE' by comparing mean ‘TOTSPACE’ between each pair of ‘AUSPICE’ groups. Result shows a statistically significant mean difference of 17.1194 spaces between Commercial Agencies and Non-Profit Agencies, with the non-profit agencies having more spaces on average. The comparison between Commercial Agencies and Public (City Operated) Agencies is not statistically significant (p = 0.0779), indicating that any observed difference in mean 'TOTSPACE' might be due to random chance. Non-Profit Agencies have significantly fewer spaces than Public (City Operated) Agencies, with a mean difference of -34.3346.

Post-hoc Test for 'subsidy': compares the mean 'TOTSPACE' between subsidized (Y) and non-subsidized (N) groups and a statistically significant mean difference of 26.2658 spaces, with subsidized centers having more spaces on average.

Overall, the post-hoc analysis indicates the type of agency ('AUSPICE') has a significant effect on the mean number of total spaces, with non-profit agencies having more spaces than commercial ones and public agencies having fewer spaces than non-profits.

<u>Tukey HSD Post-hoc Test</u> Result: There is a significant difference in 'TOTSPACE' between 'Non Profit Agency, Y' and 'Non Profit Agency, N', and between 'Non Profit Agency, Y' and 'Commercial Agency, N', as well as with 'Public (City Operated) Agency, Y'. This suggests that for Non-Profit Agencies, subsidy status makes a significant difference in 'TOTSPACE'. There are no significant differences in 'TOTSPACE' for the 'Public (City Operated) Agency' when comparing subsidized to non-subsidized, as indicated by the difference of 0 and large p-values. Several comparisons have a mean difference of 0.000 and an infinite confidence interval, indicating that was no variation between these groups.
Overall, the subsidy status has a significant effect on the total spaces available in childcare centers and that this effect interacts with the type of agency operating the center.

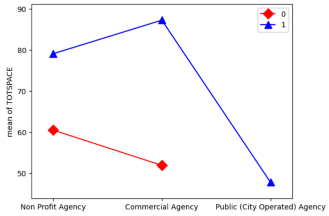
Research Question 5: Do data meet assumptions required for conducting two-way ANOVA?



```
Normality test using Shapiro-Wilk test:  {'W': 0.9018619656562805, 'p-value': 1.5311055543621852e-25}
Homogeneity of variances using Levenes test:  {'W': 22.988879302520136, 'p-value': 1.8617545516099655e-06}
Homogeneity of variances using Bartletts test:  {'W': 49.081859352317615, 'p-value': 2.45500187845506e-12}
```

<u>Check assumption for 'subsidy'</u>: Shapiro-Wilk Test has test statistic W near 0.902. p-value 1.53e-25, the residuals of the model do not follow a normal distribution, thus violating the normality assumption. Levene's Test result: W is approximately 22.989, p-value 1.86e-06, the group variances are not equal across the 'subsidy' groups. Bartlett's Test has test statistic near 49.082 and p-value 2.455e-12, suggesting variances are not equal. The results indiciate significant violations of the normality of residuals and the homogeneity of variances for the 'subsidy' variable. Two-way ANOVA assumptions are not met.
<u>Check assumption for 'AUSPICE'</u>: ShapiroResult has t-statistic=0.902 &pvalue=1.531e-25. The very small p-value indicates the residuals from the two-way ANOVA model do not follow a normal distribution, violating key assumption of ANOVA. Levene's test result has t-satistic=28.155, & pvalue=1.219e-12, suggesting the variances of 'TOTSPACE' are not equal across the different 'AUSPICE' categories. This is another violation of ANOVA assumptions as ANOVA assumes equal variances (homoscedasticity). Therefore, the assumptions necessary for conducting a traditional two-way ANOVA are not met.

Further to visualize how the mean of dependent variable 'TOTALSPACE' changes across categorical variable 'AUSPICE' and 'subsidy' by conducting interaction plot: For Non-Profit Agencies, there is a marked increase in mean 'TOTSPACE' from non-subsidized to subsidized. For Commercial Agencies, the mean 'TOTSPACE' is higher for non-subsidized than subsidized, which is a unique trend compared to the other categories. The effect of 'subsidy' is not consistent across 'AUSPICE' categories as from non-parallel lines. Also, the impact of subsidy on total available spaces is significantly different across the types of childcare agencies.


mean of TOTSPACE
Non Profit Agency
Commercial Agency
Public (City Operated) Agency