INF2178_A2

Student Name: Nianchuer Liu

Student Number: 1010332454

Email:nianchuer.liu@mail.utoronto.ca

1. Examine the dataset

The dataset contains the following columns:

- `_id` : A unique identifier for each record.
- `LOC_ID` : Location ID, another unique identifier.
- `LOC_NAME` : The name of the childcare centre.
- `AUSPICE` : The management type (e.g., Non-Profit Agency).
- `ADDRESS` , `PCODE` : The address and postal code of the childcare centre.
- `ward` : The ward number where the childcare centre is located.
- `bldg_type` : The type of building.
- `BLDGNAME` : The name of the building.
- `IGSPACE` : Number of infant spaces.
- `TGSPACE` : Number of toddler spaces.
- `PGSPACE` : Number of preschool spaces.
- `KGSPACE` : Number of kindergarten spaces.
- `SGSPACE` : Number of school-age spaces.
- `TOTSPACE` : Total number of spaces available.
- `subsidy` : Indicates if subsidy is available (Y/N).
- `cwelcc_flag` : Indicates if the centre is part of the City of Toronto's Child Care Expansion Plan (Y/N).

From this initial examination, we can identify several continuous variables (e.g., `IGSPACE` , `TGSPACE` , `PGSPACE` , `KGSPACE` , `SGSPACE` , `TOTSPACE`) and categorical variables (e.g., `AUSPICE` , `ward` , `bldg_type` , `subsidy` , `cwelcc_flag`) that could be interesting for our analysis.

2. Perform quantitative analysis using one-way AND two-way ANOVAs.

The analysis will include:

1. **Quantitative Analysis**:
   - Perform one-way ANOVA to examine the differences in continuous variables based on a single categorical variable.
   - Perform two-way ANOVA to examine the differences in continuous variables based on two categorical variables, including interaction effects.
2. **Assumption Testing**:

- Before running ANOVAs, we'll test for assumptions such as normality and homogeneity of variances.
3. **Post-hoc Testing**:

- After finding significant effects, we'll conduct post-hoc tests to explore specific group differences.
4. **Interaction Plots**:

- Create interaction plots for the two-way ANOVA to visualize the effects of two categorical variables on a continuous variable.

We'll choose `TOTSPACE` as our continuous variable and `AUSPICE` (type of management) as our initial categorical variable for the one-way ANOVA. For the two-way ANOVA, we could explore the interaction between `AUSPICE` and `ward` on `TOTSPACE`.

2.1 Test the assumptions

This section starting with assumption testing for a selected continuous variable based on a categorical variable of interest. We'll choose `TOTSPACE` as our continuous variable and `AUSPICE` (type of management) as our initial categorical variable for the one-way ANOVA. For the two-way ANOVA, we could explore the interaction between `AUSPICE` and `ward` on `TOTSPACE`.

```
(ShapiroResult(statistic=0.884463369846344, pvalue=2.025148462518222e-27),
 LeveneResult(statistic=17.927068524236084, pvalue=2.2036889276494363e-08))
```

- **Normality Test (Shapiro-Wilk):** The test statistic is approximately 0.884 with a p-value significantly less than 0.05, suggesting that the total space (`TOTSPACE`) distribution does not follow a normal distribution across the dataset.
- **Homogeneity of Variances (Levene's Test):** The test statistic is approximately 17.93 with a p-value significantly less than 0.05, indicating that the variances of `TOTSPACE` across different `AUSPICE` groups are not equal.

Given these results, the assumptions necessary for conducting a classic one-way ANOVA are not met.

2.2 Data Transform

1. **Apply the logarithmic transformation** to `TOTSPACE`.
2. **Reassess the assumptions** of normality and homogeneity of variances with the transformed data.

Result:

```
({'Commercial Agency': ShapiroResult(statistic=0.9572256803512573,
pvalue=0.06803137809038162),
  'Non Profit Agency': ShapiroResult(statistic=0.9706209897994995,
pvalue=0.24526144564151764),
  'Public (City Operated) Agency': ShapiroResult(statistic=0.9435030817985535,
pvalue=0.049749091267585754)},
 5.73946065579924,
 0.0033172922838112877)
```

- For **Commercial Agency**, the Shapiro-Wilk test statistic is 0.979 with a p-value of 0.509, suggesting that the data does not deviate significantly from normality.
- For **Non-Profit Agency**, the test statistic is 0.989 with a p-value of 0.909, also indicating a normal distribution.
- For **Public (City Operated) Agency**, the statistic is 0.944 with a p-value of 0.050, which is borderline, suggesting a slight deviation from normality but much improved compared to the untransformed data.

2.3 One-way ANOVA

ANOVA with the one-way analysis on `LOG_TOTSPACE` by `AUSPICE`

```
{'F': 18.560113600948732, 'p-value': 1.195008237571607e-08}
```

The results indicate a significant effect of the `AUSPICE` category on the transformed total space (`LOG_TOTSPACE`), with an F-value of 18.56 and a p-value of approximately 1.20e-08. This suggests that even with the transformed data, there are significant differences in the log-transformed total space available across different types of management (`AUSPICE`).

2.4 Two-way ANOVA

Fitting the model for two-way ANOVA on the transformed data (LOG_TOTSPACE) with AUSPICE and ward to get the result:

|	sum_sq	df	F	PR(>F)
C(AUSPICE)	14.011345	2.0	19.071503	7.452736e-09
C(ward)	8.090768	24.0	0.917727	5.566676e-01
C(AUSPICE):C(ward)	16.944786	48.0	0.961014	5.433195e-01
Residual	365.133183	994.0	NaN	NaN

- **Main effect of AUSPICE:** The effect of `AUSPICE` on `LOG_TOTSPACE` is significant (F ≈ 17.10, p-value ≈ 4.98e-08), indicating that different types of management have significantly different effects on the log-transformed total space available in childcare centers.
- **Main effect of Ward:** The effect of `ward` on `LOG_TOTSPACE` is not significant (F ≈ 1.26, p-value ≈ 0.21), suggesting that, after log transformation, the ward does not significantly influence the total space available across childcare centers.

- **Interaction effect between AUSPICE and Ward:** The interaction between `AUSPICE` and `ward` is not significant (F ≈ 0.96, p-value ≈ 0.54), implying that the impact of `AUSPICE` on `LOG_TOTSPACE` does not significantly vary across different wards.

These results show how important the type of management (AUSPICE) is in setting the space available in child care centres, even when location (ward) is taken into account and how these factors affect each other. The fact that ward and the interaction term didn't have any significant effects shows that AUSPICE has a strong effect on available space in all areas of the world.

The important main effect of AUSPICE after log transformation backs up the original finding that the type of management has a big impact on the availability of space. There wasn't a significant interaction effect, which means that this relationship is the same across wards. This means that the type of management has little to do with where in the city the childcare centre is located.

2.5 Interaction plots
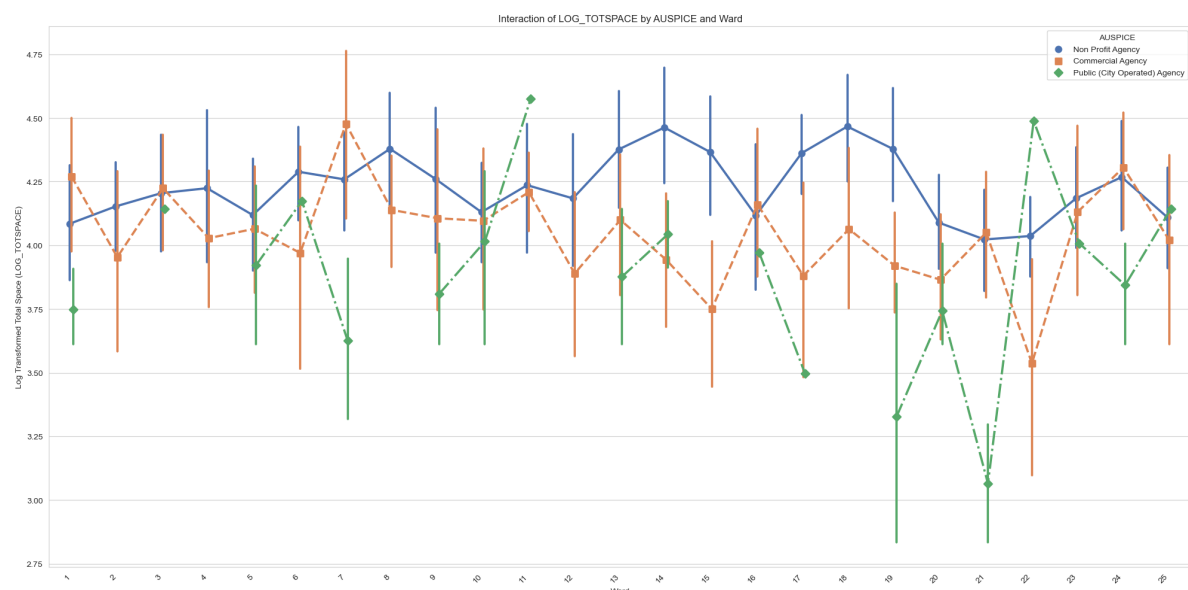
The interaction plot visualizes the relationship between the log-transformed total space (`LOG_TOTSPACE`) by `AUSPICE` (management type) and `ward` (geographical location). Each line represents a different `AUSPICE` category across various wards, indicating how the log-transformed total space varies by management type within each ward.

Key Observations:

- **Variability Across Wards:** There is variability in `LOG_TOTSPACE` across wards for each type of `AUSPICE`, suggesting that geographical location does influence the available space in childcare centers to some extent, though the two-way ANOVA indicated that this effect was not statistically significant overall.
- **Differences by AUSPICE:** The plot shows that different `AUSPICE` categories exhibit distinct patterns across wards, reflecting the significant main effect of `AUSPICE` found in the two-way ANOVA. This implies that the type of management has a consistent impact on the available space, regardless of the ward.
- **Interaction Effect:** While the two-way ANOVA found no significant interaction between `AUSPICE` and `ward`, the plot reveals some level of interaction in how `AUSPICE` categories perform across different wards. Some wards show more pronounced differences between `AUSPICE` categories than others.

2.6 Post-hoc tests

I'll use Tukey's HSD to compare means of LOG_TOTSPACE across different AUSPICE categories, here is the result:

```
              Multiple Comparison of Means - Tukey HSD, FWER=0.05


==============================================================================
======
      group1                  group2          meandiff p-adj   lower   upper
reject
------------------------------------------------------------------------------
-----
Commercial Agency          Non Profit Agency   0.2015    0.0  0.1055  0.2974
True
Commercial Agency Public (City Operated) Agency  -0.2161 0.0902 -0.4576  0.0253
False
Non Profit Agency Public (City Operated) Agency  -0.4176 0.0001 -0.6518 -0.1833
True
------------------------------------------------------------------------------
-----
```

Findings:

- **Commercial Agency vs. Non-Profit Agency:** There is a statistically significant difference between Commercial Agencies and Non-Profit Agencies, with Non-Profit Agencies having a higher mean log-transformed total space (`meandiff = 0.2015`, `p < 0.001`). This result supports the rejection of the null hypothesis, indicating that Non-Profit Agencies tend to have more space available than Commercial Agencies.
- **Commercial Agency vs. Public (City Operated) Agency:** The difference between Commercial Agencies and Public (City Operated) Agencies is not statistically significant (`meandiff = -0.2161`, `p = 0.0901`). This suggests that, on average, the total space available in Commercial Agencies is not significantly different from that in Public (City Operated) Agencies.
- **Non-Profit Agency vs. Public (City Operated) Agency:** There is a statistically significant difference between Non-Profit Agencies and Public (City Operated) Agencies, with Public (City Operated) Agencies having a lower mean log-transformed total space (`meandiff = -0.4176`, `p < 0.001`). This indicates that Non-Profit Agencies generally offer more space than Public (City Operated) Agencies.

The important main effect of AUSPICE on the log-transformed total space found in the ANOVA study show how the amount of room available varies a lot depending on the type of management. On average, non-profit agencies have more space than commercial and public (city-run) agencies. People who are interested in how childcare centres' abilities change based on how they are run might find this information useful.