Analysis of Daycare Space Availability by Auspice and Subsidy Status

Introduction

This study investigates the utilization of different types of daycare auspices and amount of daycare spots available. This research focuses on the number of daycare spaces (**TOTSPACE**) by various daycare auspices. Particularly, we are seeking evidence that auspice typology (**Non-Profit Agency**, **Commercial Agency**, and **Public (City Operated) Agency**) and subsidy status (subsidized **Y** or not subsidized **N**) bring the condition of total spaces in daycare centers to a significant change.

Through a detailed analysis of the data available in the 'INF2178_A2_data.xlsx' we expect large fluctuations in daycare spaces across different auspice types and subsidy statuses and we plan to present findings in a manner that makes them widely understandable.

Our investigation revolves around pivotal questions:

1. **How do the total daycare spaces may differ among various auspice types?**
2. **How does the intersection of auspice type and subsidy status affect the availability of daycare spaces?**

With this investigation as the vehicle, we hope to achieve the goal of enlightening daycare space structures and data-driven recommendations for equal distribution of these spaces would be among the outputs.

Methodology

There were **17 columns** of attributes including **1064 records** of base information of both subsidized and non-subsidized base facilities managed by three auspices. In the **TOTSPACE** variable, the number of total spaces available in any facility noted will be displayed. I performed ANOVA to assess the average **AUSPICE** effect and **subsidy** levels.

To guarantee the rightfulness of the ANOVA assumptions, I used the Shapiro-Wilk test to check the normality of residuals and Levene's test for the equality of variances. Outliers were determined by applying the IQR method for a more accurate normality test, which helped determine whether outliers were influencing the normal distribution.

Results

Before examining the first research question, I first present an overview of the general distribution of spaces within the dataset. The **TOTSPACE** variable, representing the total number of daycare spaces available, shows the following:

Name	TOTSPACE
Count	1063
Mean	75.674506
Std	47.816518
Min	6
25%	43
50%	62
75%	97
Max	402

Assumptions:

Normal Distribution

- A non-normal distribution of the residuals was confirmed through Shapiro-Wilk's test. (w = 0.901775598526001, p< less than 1.4964898448030214e-25).

Homogeneity of Variances

- Levene's test for homogeneity of variances across the **AUSPICE** groups yielded a result (W = 9.1994, p = 0.0001) that was significant, thus, indicating heterogeneity of variances, and violation of ANOVA assumptions.

Although the Shapiro-Wilk test indicated a significant deviation from normality, and Levene's test revealed heterogeneity of variances across the AUSPICE groups, we proceeded with the ANOVA analysis. This decision was based on the consideration of the large sample size, which can mitigate the impact of these violations on the ANOVA's robustness.

Residual Analysis

One-Way ANOVA

- A one-way ANOVA was utilized in testing for the effects of **AUSPICE** on **TOTSPACE**. Results showed us a result with statistical significance (F(2, 1060) = 21.843, p < 5.057716e-10), so we had to reject the null hypothesis that the AUSPICE has no impact on the **TOTSPACE**.

One-Way ANOVA Results:	sum_sq	df	F	PR(>F)
C(AUSPICE)	9.611211e+04	2.0	21.843051	5.057716e-10
Residual	2.332065e+06	1060.0	NaN	NaN

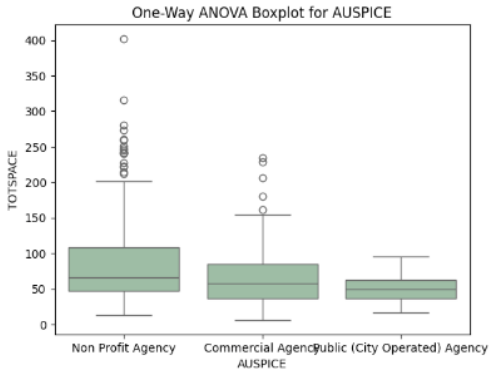

Figure 1: One-way ANOVA Boxplot for AUSPICE

- Q-Q plot and histogram were used to check the normal distribution of the standardized residues of the one-way ANOVA. The Q-Q plot did not follow the 45-degree reference line by far, thus the histogram showed right skewness which is a sign of the distribution that did not obey normality.

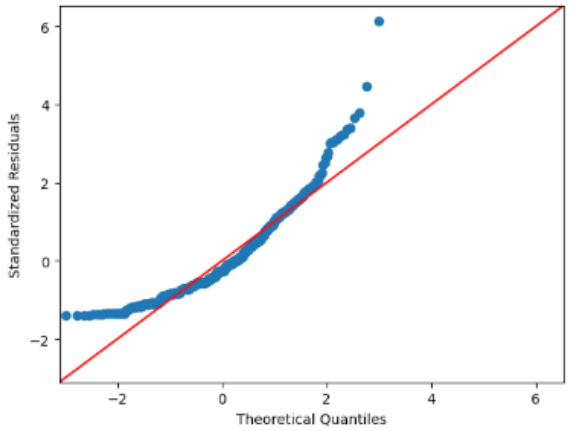

Figure 2: One-way ANOVA Q-Q plot

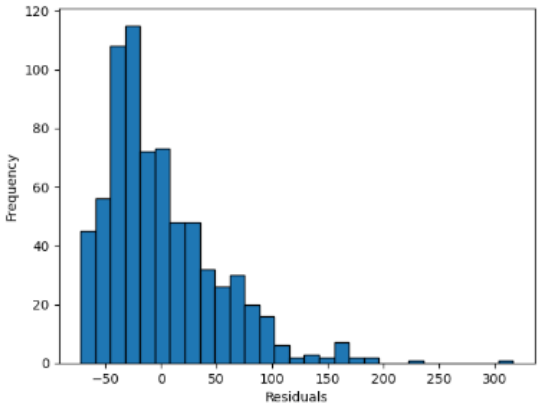

Figure 3: One-way ANOVA Histogram plot

- In Tukey's HSD post hoc test, these agencies had a significant difference in the **TOTSPACE** category ($p < 0.00001$) in the **Non-Profit Agency** group, compared to both **Commercial Agency** ($p = 0.006071$) and **Public (City Operated) Agency** ($p = 0.005901$) groups. No significant difference was found regarding comparisons of **Commercial Agency** and **Public (City Operated) Agency** ($p = 0.272554$).

Tukey's HSD Summary:

	group1	group2	Diff	Lower
0	Non Profit Agency	Commercial Agency	16.806538	3.993722
1	Non Profit Agency	Public (City Operated) Agency	36.177966	8.673910
2	Commercial Agency	Public (City Operated) Agency	19.371429	-10.141900

	Upper	q-value	p-value
0	29.619353	4.356853	0.006071
1	63.682022	4.369046	0.005901
2	48.884757	2.180132	0.272554

Outlier Removal and Normality Reassessment

- Upon the completion of a preliminary analysis, outliers were detected and deleted from the dataset. This approach is intended to lower the bias from the extreme values that might cause misleading results and not meet the ANOVA assumptions.
- Even the removal of outliers would not correspond to the Shapiro-Wilk test normality test for the residuals that establish a deviation from normality ($w = 0.9491212368011475$, $p = 2.4891868476754512e\text{-}18$). This p-value is still significantly low and hence after removing the outliers the data set did not conform to a normal distribution. Thus, outliers do not influence the normal distribution in this case.

  Shapiro-Wilk test Results

  - The Shapiro-Wilk test statistic: 0.9491212368011475
  - P-value: 2.4891868476754512e-18

Two-Way ANOVA

- Subsequently, a two-way ANOVA was carried out which revealed a significant main effect for **AUSPICE** and **subsidy**, and a significant interaction effect between these two variables ($p < 0.001$ and $p < 0.001$ respectively,) thus, **AUSPICE** impact on **TOTSPACE** may differ depending on the subsidy status.

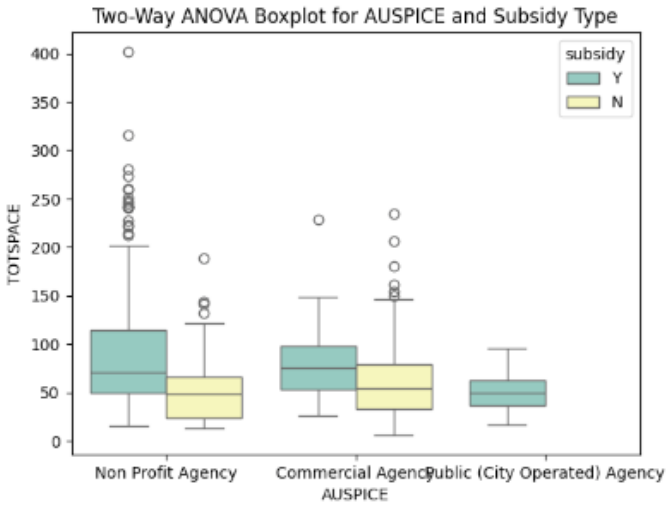

Figure 5: Two-way ANOVA Boxplot for AUSPICE and Subsidy Type

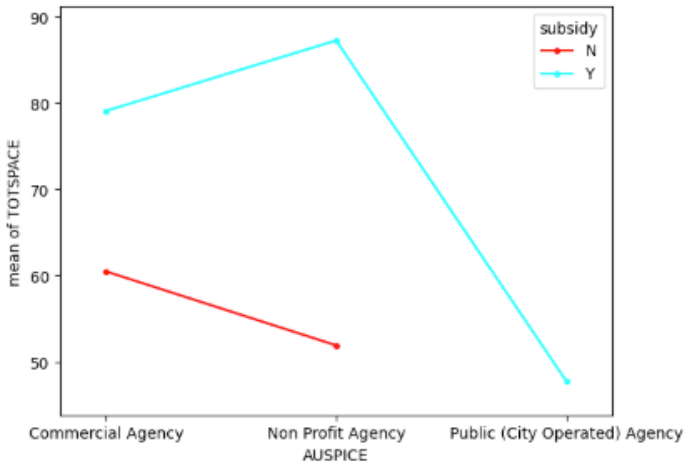

Figure 6: Two-way ANOVA Interaction plot for AUSPICE and Subsidy Type

Discussion

This finding showed a variation of daycare spaces for the different auspices and the difference between subsidized and non-subsidized facilities that had statistical significance. Among others, **Non-Profit Agency** facilities gave the impression of diverse in

terms of space compared to other facilities possibly indicating operational or funding discrepancies.

The result of the analysis was a statistically significant variation in the number of daycare places according to their auspice type and the existence of subsidizing or not. Indeed, **Non-Profit Agency** premises largely differed from the others in the aspect of spaciousness. Even without the removal of strong deviations, normality was still reviewed, thus making it clear that other statistical techniques for example, data transformation or non-parametric tests should be sought. These discoveries, therefore, underline the general characteristics helping to model daycare space allocation and further stress the use of proper statistical methods, which take into consideration the fact about data distribution by nature.

The significant p-values in Shapiro-Wilk and Levene's tests do not validate the normality and equal variances assumptions underlying ANOVA, thus, indicating the need for data reassessing or opting for different methods, like the data transformation or non-parametric tests.

These findings indeed emphasize the intricate interplay among the authority type, funding source, and the establishment of daycare places. Future research may well investigate the causes of this difference and it is recommended to use statistical approaches with the invariance-heterogeneity variation feature that can handle the data that is non-Gaussian.