Name: Qi Li
Student ID: 1005299302
Course: INF2178-Experimental Design for Data Science
Instructor: Shion Guha
Content: Technical Assignment I

Research Questions

1. Evaluate the adequacy of current shelter establishments distinguished by capacity types by comparing the number of room-base and bed-based units with the total number of users occupied for each type to assess the alignment between shelter capacity and the actual demand.
2. Evaluate the adequacy of current shelter establishments distinguished by program models by comparing the number of emergency and transitional units with the total number of users occupied for each model to assess the alignment between shelter capacity and the actual demand.
3. What practical scale of building shelters can be determined by evaluating the distribution of the service_user_count variable, considering differences in capacity types, program models, or both?
4. What is the temporal pattern of occupancy rate in 2021 across both capacity types?
5. Are there significant differences in the occupancy rates among shelters with distinct capacity types or program models?

By answering these questions, I aim to provide insightful applications in resource allocation in Toronto and potentially answer the question of why homeless people in Toronto are often turned away despite available shelter spaces.

EDA

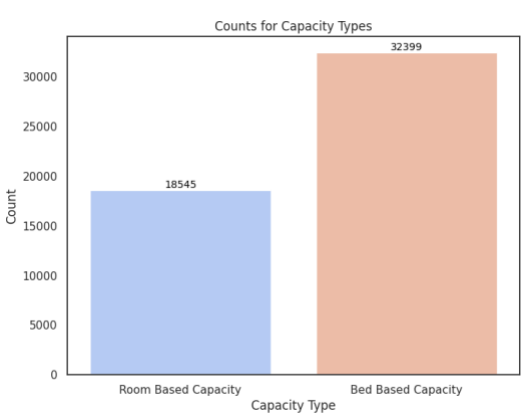

Fig 1. Count plot by two capacity types

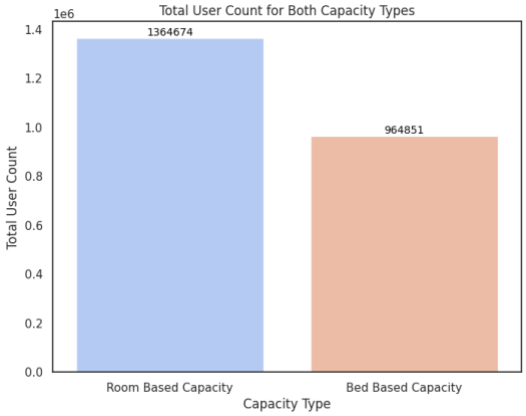

Fig 2. Bar plot of the total users for two capacity types

There is a notable imbalanced distribution of homeless shelter capacity types, with 32,399 units of bed-based and 18,545 units of room-based. To assess the adequacy of this establishing plan made by the government of Toronto, it's important to compare it with the actual utilization measured by the total service-users counts for each type of capacity provided in Fig 1 and Fig 2. In 2021, the total number of room-based capacity users was significantly higher at 1,364,674 individuals than the total users of bed-based capacity at 964,851 individuals. The disproportionate relationship of total users and total capacities distinguished by capacity type suggests an inefficient allocation that potentially explains the problem of the homeless turning away despite available shelter spaces. A more responsive strategy should emphasize constructing more room-based capacity due to its high demand.

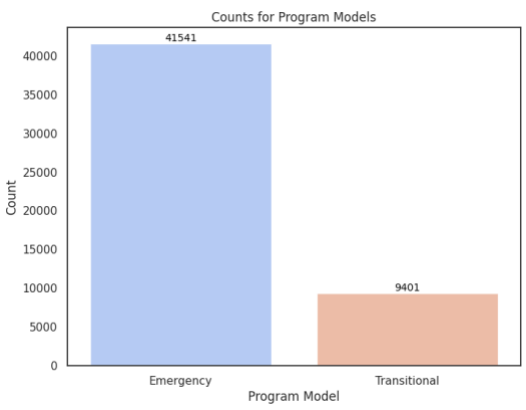

Fig 3. Count plot by two program models

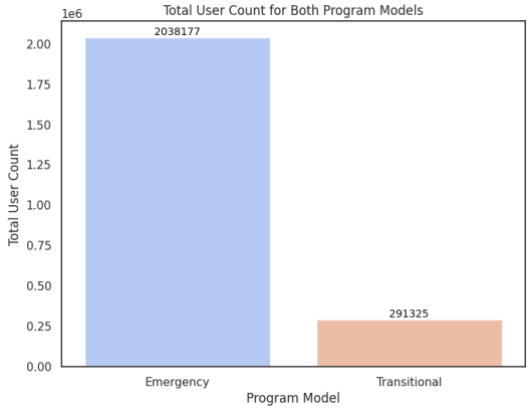

Fig 4. Bar plot of total users for two program models

A similar concern isn't evident when comparing the relationship between total users and program models. As depicted in Fig 3 and Fig 4, there are more programs designed to be emergency (41,541) compared to transitional programs (9,401). This correctly reflects the excessive usage of emergency programs with 2,038,177 individuals over the transitional programs with 291,325 individuals in 2021.

Through an examination of the distribution of the service_user_count variable, I aim to determine the practical scale of building shelters, considering distinctions in capacity type, program models or a combination of both.

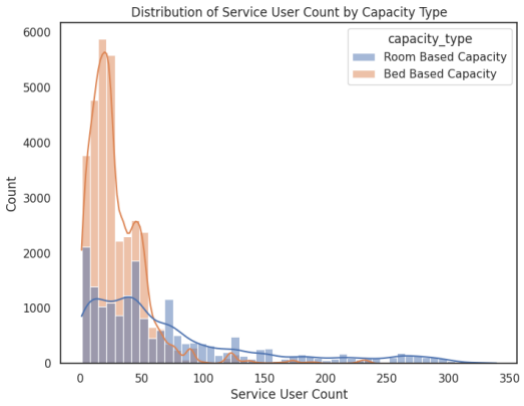

Fig 5. Histogram of service_user_count by capacity types

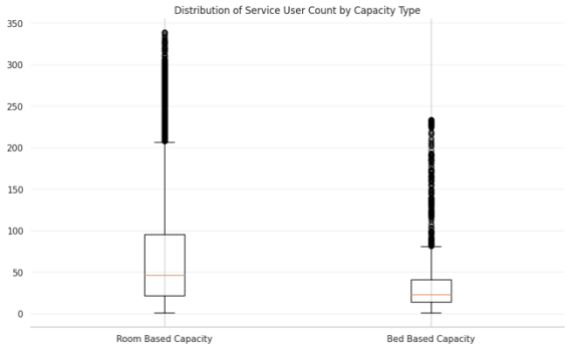

Fig 6. Boxplot of service_user_count by capacity types

As illustrated in Fig 5, the distributions of service_user_count are right-skewed regardless of the capacity type, with values predominantly concentrated between 0 and 100 on the x-axis. Yet the distribution for bed-based capacities appears notably flatter. The box plot presented in Fig 6 reveals that the more extreme skewness of bed-based capacity is not solely attributable to its higher unit count. Considering the information given in the summary statistics, the service_user_count for bed-based capacity possesses a lower IQR and median (Note: the IQR and median are more appropriate measurements for the central tendency and spread here due to the high volume of outliers).

Based on the distribution pattern, a more effective construction plan should address downsizing bed-based capacity with target accommodation under 100 individuals per facility while it's best to build room-based capacity with various sizes and the distribution depends on actual needs among different areas of Toronto.

Such distinctions in terms of IQR and median are not significant when contrasting two program models. Hence, the height difference in the histogram may be due to the difference in the unit counts. One would conclude that capacity type is the more important factor in deciding the size of shelters.

Then I calculate the occupancy rate for both capacity types and investigate how the rate fluctuated in 2021.

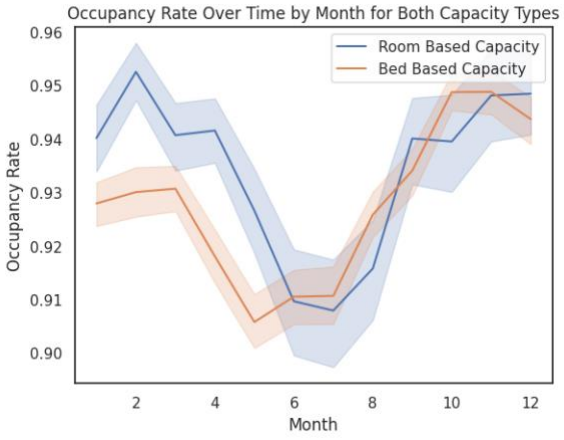

Fig 7. Line chart of occupancy rate for both capacity types in 2021

From Fig 7, there is a significant variation throughout the year 2021. Irrespective of the capacity type, the occupancy rate is relatively low from April to September and higher in the remaining months on average. The pattern aligned with temperature changes in Toronto. During the warmer period, fewer homeless people seek refuge as the temperature allows them to stay on the streets. In contrast, warm shelter places are in high demand in the winter. From a city planning standpoint, the government can shut down facilities having continuously low occupancy rates

between April and September to cut costs and reopen them to provide more shelter spaces for homeless people when the weather is severe. This dynamic strategy coincides with the seasonal variations in demand can help to improve the effectiveness and cost-efficiency of the shelter management.

T-tests

One Sample T-test

First, I want to know if the true population mean of the daily service user count is equal to 45. This only involves one continuous variable so the One Sample t-test would be suitable for this. Assumptions (Herzog et al, 2019) checks:

1. The data is independent and identically distributed.
2. By the Central Limit Theorem, population distribution is Gaussians
3. The dependent variable is ratio scaled.
4. The sample size is fixed.

The hypothesis test would be:

$$H0: \mu = \ 45$$

$$H1: \mu \neq 45$$

From my code, the calculated t-statistic = 3.0778231986607936 and the p-value = 0.002086292851627579 is lower than the conventional significance level of 0.05. Therefore, the null hypothesis is rejected and concludes that we have significant evidence to say the population mean of the service user count is not equal to 45.

Welch's T-test

Now I want to study if there is a significant difference in the average occupancy rate between different program models or different capacity types. The four assumptions are all met the same

as for the One Sample t-test. Since it's a hypothesis test comparing two samples, I checked the variances of both capacity types/program models are not equal. Hence, the Student's T-test is omitted and I chose Welch's T-test.

Null Hypothesis: There is no significant difference in the mean occupancy rate between the two capacity types/program models.

Alternative Hypothesis: There is a significant difference in the mean occupancy rate between the two capacity types/program models.

The obtained t-statistic = 4.498751771925636 and p-value = 6.860477551487939e-06 are for the test between bed-based occupancy rate and room-based occupancy rate. The P-value is less than the significant level of 0.05 so the null hypothesis is rejected and conclude that we have enough evidence to say the means occupancy rate of bed-based and room-based capacities are not equal. Additionally, the t-statistic = 40.981115372199206 and p-value = 0.0 which is less than the significant level of 0.05 for the comparison between the mean occupancy rate of emergency and transitional program models. So, the null hypothesis is rejected and conclude that we have enough evidence to say the means occupancy rate between emergency and transitional models are not equal.

Reference

Herzog, M. H., Francis, G., & Clarke, A. (2019) *Understanding Statistics and Experimental Design : How to Not Lie with Statistics*. Springer International Publishing.