Jingxian Chen
1005354572

INF2178 Assignment1: Analyzing Shelter Usage Trends in Toronto: A 2021 Perspective

Background and Introduction

In the face of Toronto's growing homelessness, the city's shelters are vital yet strained. This study has analyzed how different shelter programs are used and how fully they are occupied. The findings reveal distinct usage patterns by concentrating on the correlation between service user numbers and program models, and the relationship of occupancy rates to the type of occupancy—specifically bed versus room utilization.

Data Preparation and Methodology

- In the analysis, I focused on variables critical to understanding shelter capacity and utilization, such as CAPACITY_TYPE, PROGRAM_MODEL, SERVICE_USER_COUNT, CAPACITY_ACTUAL_BED, OCCUPIED_BEDS, CAPACITY_ACTUAL_ROOM, and OCCUPIED_ROOMS.
- The dataset underwent a rigorous cleaning process to ensure accuracy in these variables. The analytical approach included constructing correlation heatmaps to identify relationships, performing Welch's and two-sample t-tests to detect differences in service user numbers across shelter models, and generating box plots for a visual distribution comparison. These methods together aimed to provide a comprehensive overview of shelter usage and capacity dynamics in Toronto.

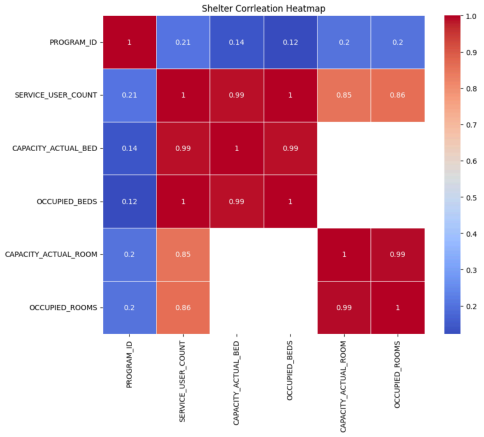
Research Questions

1. Do Emergency and Transitional shelter programs in Toronto demonstrate a statistically significant difference in their average daily service user counts, and if so, what does this suggest about the utilization patterns of these two types of shelter services?
2. How do the occupancy rates and numbers differ between bed-based and room-based shelter capacities, and what are the implications of these differences for managing the efficiency of shelter utilization in Toronto?

Key Observations from the Heatmap:

1. SERVICE_USER_COUNT, CAPACITY_ACTUAL_BED, and OCCUPIED_BEDS have very high positive correlations with each other (close to 0.99), indicating that as the service user count increases, the actual bed capacity and the number of occupied beds tend to increase proportionally.

2. CAPACITY_ACTUAL_ROOM and OCCUPIED_ROOMS also display a very high positive correlation (close to 0.99), suggesting a similar relationship between room capacity and room occupancy.
3. OCCUPANCY_RATE_BED and OCCUPANCY_RATE_ROOM do not seem to have strong correlations with other variables, suggesting that the rate of occupancy may be influenced by factors not represented in this correlation matrix.
4. PROGRAM_ID has very low correlations with all other variables, indicating that the program identifier does not have a linear relationship with counts or capacities.



Service Users by Program Model

Exploratory data analysis

1. **Statistical Summary for Emergency and Transitional Programs:**

- **Emergency**: The service user count has a wide range with a minimum of 1 and a maximum of 339, a median of 33, and a mean of 49.06, suggesting a right-skewed distribution given the mean is higher than the median. The standard deviation is 55.92, indicating a high level of variability in service user count.
- **Transitional**: Compared to emergency programs, transitional programs have a lower range of service user count (min 1 to max 221), a median of 23, and a lower mean of 30.99, which also suggests a right-skewed distribution. The standard deviation is 36.43, which is smaller than that of emergency programs, indicating less variability.

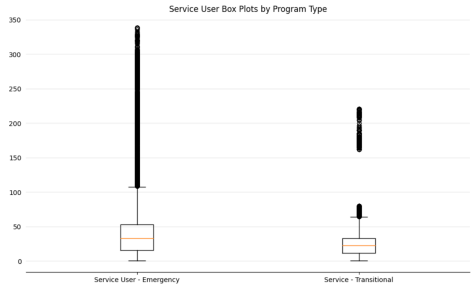
```
Emergency
Min: 1
Max: 339
Median: 33.0
Mean: 49.06
Standard Deviation: 55.92


Transitional
Min: 1
Max: 221
Median: 23.0
Mean: 30.99
Standard Deviation: 36.43
```

2. **Box Plots Observations:**

- The box plot for the Emergency program type shows a larger interquartile range (IQR), with more outliers extending above the upper whisker. This is

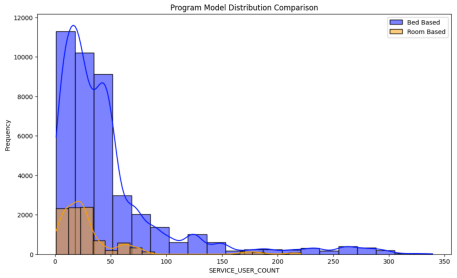
consistent with the higher variability and wider range observed in the statistical summary.

- The Transitional program type's box plot has a narrower IQR and fewer outliers, which correlates with its smaller standard deviation and indicates that the service user numbers are more concentrated around the median.



3. **Histogram Comparison:**

- The histogram comparison between Bed Based and Room Based services shows a higher frequency of lower service user counts for Bed Based services, which rapidly decreases as the service user count increases. This suggests that most of the bed-based services have a lower number of users, with fewer instances of very high usage.
- The Room Based services also show a decrease in frequency as service user count increases, but without the sharp decline seen in Bed Based services, indicating a more uniform distribution across different service user counts.



T-test Analysis

1. **Rationale for Using Two-Sample t-Test:** Despite the data not adhering to a normal distribution, the Two-Sample t-Test is applied under the justification provided by the Central Limit Theorem. This theorem states that as sample sizes increase, the distribution of sample means will approximate a normal distribution, rendering the Two-Sample t-Test appropriate for large sample sizes as its validity is less sensitive to the shape of the distribution.
2. **Differentiating Welch's t-Test from Two-Sample t-Test:** Welch's t-Test is utilized when the assumption of equal variances between the two groups is untenable. In contrast, the Two-Sample t-Test presumes homogeneity of variances across groups. When the equality of variances between groups is uncertain, Welch's t-Test serves as a more conservative approach.
3. In this instance, Welch's t-Test yields a t-statistic of 38.85 with a p-value near zero, strongly indicating a significant difference in service user numbers between Emergency and Transitional programs.

The Two-Sample t-Test corroborates this finding with a t-statistic of 29.94 and an exceedingly small p-value (3.17e-195). These significant statistical outcomes suggest that the difference in service user counts between Emergency and Transitional housing programs is not due to random chance, and it may have substantial implications for the allocation and management of housing resources.

```
Welch's t-statistic: 38.85174699254652
Welch's p-value: 0.0
Two-sample t-statistic: 29.937570467283667
Two-sample p-value: 3.1720139638162956e-195
```

Shelter Occupancy Number by Capacity Type Analysis

Based on the analysis of the shelter occupancy rate discussed above, I would like to go further to determine the trend of the shelter occupancy number to gain a more comprehensive analysis.

Exploratory data analysis

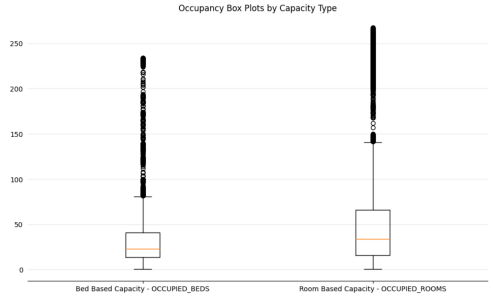
1. **Statistical Summary for Occupied Beds and Rooms Number:**

- **Occupied Beds**: The data for bed-based capacity (OCCUPIED_BEDS) shows a minimum occupancy of 1 and a maximum of 234, with a median occupancy of 23. The mean occupancy stands at 29.78, slightly above the median, indicating a right-skewed distribution. The standard deviation is 26.38, suggesting a moderate spread around the mean. This variability points to fluctuating occupancy levels, which could reflect diverse individual needs or inconsistent admission rates.
- **Occupied Rooms**: For room-based capacity (OCCUPIED_ROOMS), the occupancy figures span from a minimum of 1 to a maximum of 268. The median of 34 is lower than the mean of 52.8, again suggesting a right-skewed distribution but with a larger spread, as evidenced by a standard deviation of 58.79. The wider dispersion suggests that room occupancy levels are less predictable and could be influenced by factors such as family size or the specific requirements of room-based services.

```
OCCUPIED_BEDS
Min: 1.0
Max: 234.0
Median: 23.0
Mean: 29.78
Standard Deviation: 26.38


OCCUPIED_ROOMS
Min: 1.0
Max: 268.0
Median: 34.0
Mean: 52.8
Standard Deviation: 58.79
```

2. **Box Plot Observations:**

- The box plots for each capacity type visually underscore the data's skewness. For bed-based capacity, the interquartile range is compact, with a significant number of outliers indicating days of high occupancy. Conversely, the room-based capacity shows a broader interquartile range, implying a more diverse range of occupancy levels, and a cluster of extreme values reflecting occasional high occupancy days.



Shelter Occupancy Number by Capacity Type Analysis

Exploratory data analysis

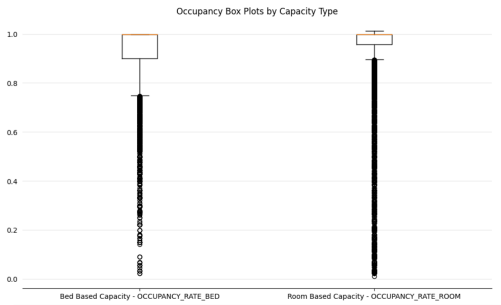
1. **Statistical Summary for Occupied Beds and Rooms Rate:**

- **Occupancy Rate for Beds**: The occupancy rate for beds (OCCUPANCY_RATE_BED) presents a minimum rate of 0.2 and a maximum of 1.0, indicating that on some days, at least 20% of bed capacity was occupied, with full occupancy on other days. The median is at full capacity (1.0), and the mean is close to it (0.93), suggesting that bed capacities are typically near or at full occupancy. A standard deviation of 0.12 points to a relatively tight concentration of rates around the mean.
- **Occupancy Rate for Rooms**: Room-based occupancy (OCCUPANCY_RATE_ROOM) shares a similar profile, with nearly full occupancy on average (mean of 0.93) and a median of 1.0. However, a slightly wider standard deviation of 0.16, compared to bed-based occupancy, indicates a slightly greater variability in room occupancy rates.

```
OCCUPANCY_RATE_BED
Min: 0.02
Max: 1.0
Median: 1.0
Mean: 0.93
Standard Deviation: 0.12


OCCUPANCY_RATE_ROOM
Min: 0.01
Max: 1.01
Median: 1.0
Mean: 0.93
Standard Deviation: 0.16
```

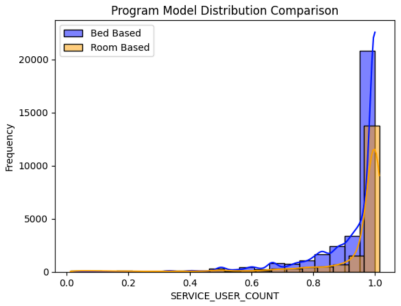
2. **Box Plot Analysis:**

- The box plots for both bed and room occupancy rates predominantly cluster at the higher end of the scale, with the majority of data points (as indicated by the compact boxes) showing high occupancy rates. There are outliers present at the lower end for both types of capacities, but they are more prominent in room-based capacities, suggesting occasional underutilization.



3. **Program Model Distribution Comparison:**

- The histogram illustrates that both bed-based and room-based services typically operate at high service user counts, with a sharp increase in frequency as occupancy rates approach 1.0. This indicates that both types of shelter capacities are frequently operating at or near full capacity.



T-test Analysis

1. The Welch's t-test, which does not assume equal variances, shows a significant difference with a t-statistic of -4.4987 and a p-value of approximately 6.86e-06. The Two-Sample t-test, assuming equal variances, confirms this difference with a t-statistic of -4.8541 and a p-value of approximately 1.21e-06.
2. Both tests indicate a statistically significant difference between bed-based and room-based occupancy rates, suggesting they are not the same and this difference is not due to random chance. The negative t-statistics suggest that bed-based capacities have a lower mean occupancy rate compared to room-based capacities. This information could be vital for resource distribution and policy planning in shelter management.

```
Welch's t-statistic: -4.498751771925636
Welch's p-value: 6.860477551487939e-06
Two-sample t-statistic: -4.854104599422829
Two-sample p-value: 1.2128933183471424e-06
```

Insight and Conclusion

The analysis confirms significant differences in daily user counts between Emergency and Transitional shelters, with Emergency usage being highly variable and Transitional more stable. Occupancy for both bed and room capacities often reaches full, with room-based occasionally overextended, suggesting the need for flexible management. These findings should guide strategic planning in Toronto's shelter system to address the distinct needs of Emergency and Transitional services.