Study Report Toronto Shelter Using Trend

Chenwei Zhu 1006003619

In recent years, recessionary economic conditions, and significant increases in the cost of rent have led to a significant rise in the number of homeless people in Toronto. (Fred Victor, 2024) Although the City of Toronto's shelter support system provides overnight and shelter services to homeless people, those seeking shelter are often turned away due to a lack of shelter space. This study will use data analysis to examine trends in shelter utilization to provide a better understanding of shelter use for the purpose of assisting the homeless.

The current study was divided into two parts, data preparation summary and experimental data analysis. In the data preparation section, the data set was initially observed and cleaned to identify variables that might be useful. The data set consists of 50944 data items, 14 variables, seven of which are categorical variables, six are numerical variables, and one is a date time value.

```
Data columns (total 14 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   OCCUPANCY_DATE          50944 non-null  datetime64[ns]
 1   ORGANIZATION_NAME       50944 non-null  object
 2   PROGRAM_ID              50944 non-null  int64
 3   PROGRAM_NAME            50909 non-null  object
 4   SECTOR                  50944 non-null  object
 5   PROGRAM_MODEL           50942 non-null  object
 6   OVERNIGHT_SERVICE_TYPE  50942 non-null  object
 7   PROGRAM_AREA            50942 non-null  object
 8   SERVICE_USER_COUNT      50944 non-null  int64
 9   CAPACITY_TYPE           50944 non-null  object
 10  CAPACITY_ACTUAL_BED     32399 non-null  float64
 11  OCCUPIED_BEDS           32399 non-null  float64
 12  CAPACITY_ACTUAL_ROOM    18545 non-null  float64
 13  OCCUPIED_ROOMS          18545 non-null  float64
```

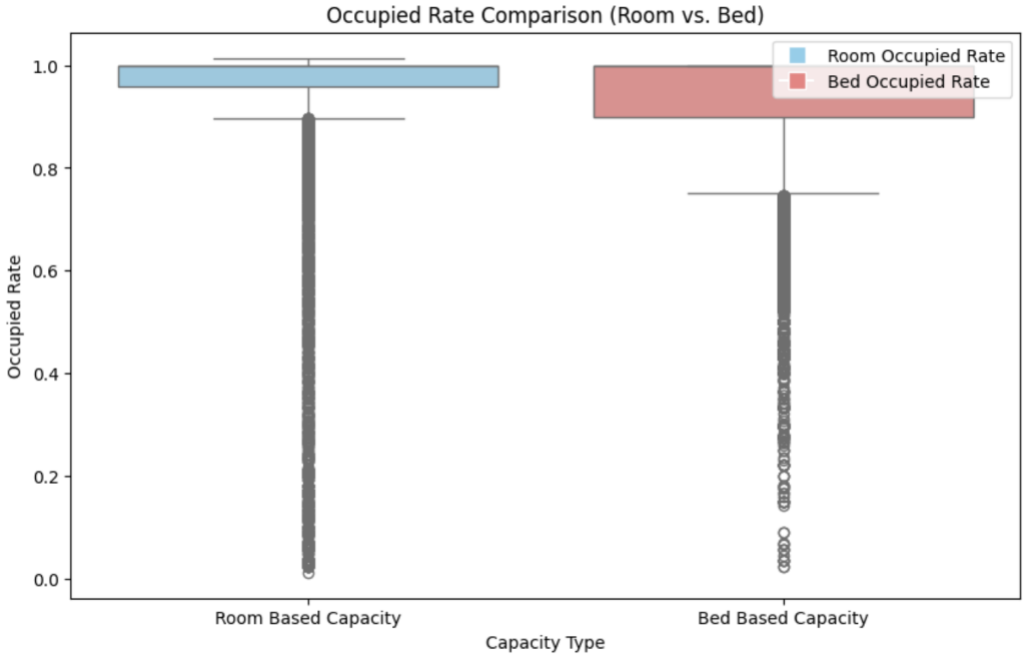
The first step in the data cleaning process was to look at the null values (n/a). The dataset categorized the shelters into two categories according to the "CAPACITY TYPE", "ROOM BASED CAPACITY " and "BED BASED CAPACITY ". Therefore, when checking the null value (n/a), it is also necessary to check the two " CAPACITY TYPE " separately. In the case of " ROOM BASED CAPACITY " shelters, the "CAPACITY_ACTUAL_BED" and "OCCUPIED_BEDS" are correctly shown as empty, and vice versa for "BED BASED CAPACITY". As a result of checking, there are two data items in the data about "BED BASED CAPACITY ". where the “PROGRAM MODEL” is empty. Since the amount of data is large and removing these two data will not affect the analysis result, these two data are removed from the data set directly.

CAPACITY_TYPE	PROGRAM_MODEL	SERVICE_USER_COUNT	CAPACITY_ACTUAL_BED	OCCUPIED_BEDS	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS
Bed Based Capacity	2	0	0	0	32399	32399
Room Based Capacity	0	0	18545	18545	0	0

To discover trends in shelter usage, shelter utilization rates need to be calculated for comparison. The occupied rate of each shelter is obtained by dividing the number of occupied shelters by the actual capacity of each shelter. For a shelter of type "ROOM BASED CAPACITY", divide "OCCUPIED_ROOMS" by "CAPACITY_ACTUAL_COMPACT_ROOM" to get the occupied

rate ("ROOM_OCCUPIED_RATE") of that shelter. For a shelter of type "BED BASED CAPACITY", divide "OCCUPIED_BEDS" by "CAPACITY_ACTUAL_BED" to get the occupied rate of the shelter ("BED_OCCUPIED_RATE "). This is the first part of this research. Moving on to the second part of this study, the experimental data analysis. In this step the data will be visualized and exploratory analysis will be performed to explore correlations, relationships, and trends between the data to help determine the direction of further analysis.

Known from the dataset that shelters are categorized into 'Bed Based Capacity' and 'Room Based' according to their 'capacity type'. Capacity' and 'Room Based'. Is there any difference in occupancy rates between the two different types of shelters. To visualize the occupancy rate of the two types of shelters, boxplot is used to visualize the data. [Figure 1]



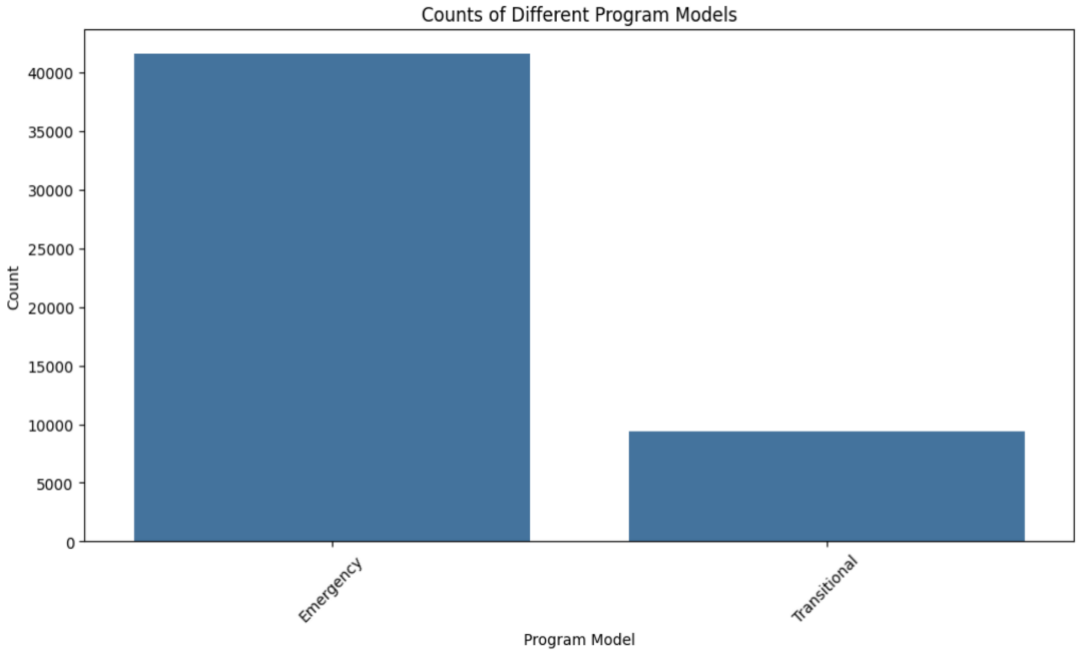
[Figure 1]

The boxplot shows the type of ‘Room Based Capacity’ shelter has high median, close to the maximum, a generally high occupancy rate. the type of ‘Bed Based Capacity’ shelter has lower median compared to the Room Occupied Rate, indicating a lower overall occupancy rate. The graph shows that there is a difference in occupancy between the two types of shelters, but whether this difference is statistically significant [i.e., whether it is statistically significantly different] requires further testing.

To verify whether there was a statistically significant difference in occupancy between the two different types of shelters, a one-sided t-test was chosen to be used. Come up with the Null Hypothesis: “There is no significant difference between the occupied rate of Room Based Capacity shelter and the occupied rate than Bed Based Capacity shelter.” After running an one sided t-test(α=0.05), the p-value = 6.321780679079661e-07 (p-value < 0.05), t-statistic = 4.845858377006688. There is enough evidence to reject the null hypothesis. Since this t-test is a one-tailed test, as indicated by the halving of the p-value. The direction of the test is determined

by the sign of the t-statistic: The t-statistic is positive, therefore the Room Based Capacity shelter has the higher occupied rate than Bed Based Capacity shelter.
The results of this analysis show that 'Room Base Capacity' shelters have higher occupancy rates than 'Bed Base Capacity' shelters. This means that homeless people preferred 'Room Base Capacity' shelters to 'Bed Base Capacity' shelters.

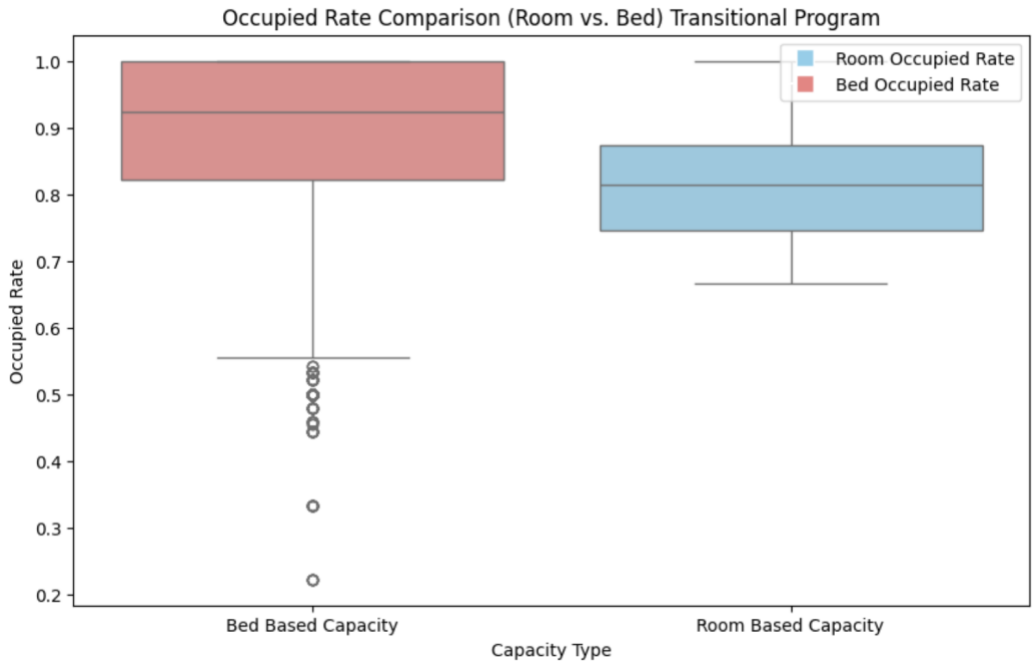
The dataset also shows that the shelter's program model is categorized as 'Emergency' and 'Transitional'. The number of `emergency' shelters is much larger than the number of `transitional' shelters. [Figure 2]



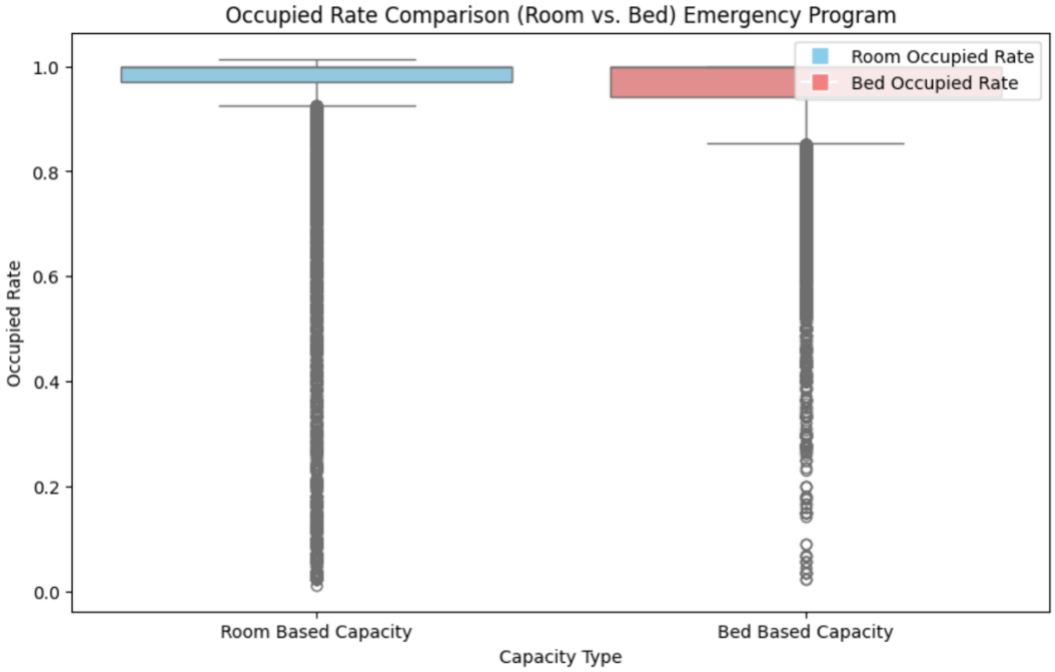
[Figure 2]
So it makes me wonder, as we now know that the occupancy rate of "Room Base Capacity" shelters is significantly higher than that of "Bed Base Capacity" types of shelters, whether this trend persists in a subset of different program models, and whether Simpson's Paradox exists.

Therefore, an in-depth subset was chosen for the study. The dataset was first categorized into "Emergency" and "Transitional" based on the project model. Then its 'Bed Occupied Rate' and 'Room Occupied Rate' were visualized and plotted on Boxplot respectively. [Figure 3] [Figure 4]

Both boxplots show the type of ‘Bed Based Capacity’ shelter has high median, close to the maximum, a generally high occupancy rate. the type of ‘Room Based Capacity’ shelter has lower median compared to the Room Occupied Rate, indicating a lower overall occupancy rate. This shows that 'Bed Base Capacity' shelters have a higher occupancy rate than 'Room Base Capacity' shelters, both in the 'Emergency' and 'Transitional' shelter subsets. 'Bed Base Capacity' shelters have a higher occupancy rate than 'Room Base Capacity' shelters in both the 'Emergency' and 'Transitional' shelter subsets. This is contrary to the conclusions of previous studies at the aggregate level.



[Figure 3]



[Figure 4]

For this I chose to perform a One-sided t test on each of the two subsets. For the subset with 'Transitional' model, Null Hypothesis: The mean ROOM_OCCUPIED_RATE is equal to the mean BED_OCCUPIED_RATE in the Transitional Shelter. The test result shows p-value =1.512326360327876e-36 (p-value < 0.05). T-statistic= -12.624991306569669. There is enough evidence to reject the null hypothesis. So, there is a statistically significant difference, with the

BED_OCCUPIED_RATE being higher than the ROOM_OCCUPIED_RATE in Transitional Shelter.

For the subset with ‘Emergency’ model, Null Hypothesis: The mean ROOM_OCCUPIED_RATE is equal to the mean BED_OCCUPIED_RATE in the Emergency Shelter. The p-value is 0.0001955539665925258 3(p-value <0.05), T-statistic: -3. 546302026901181.Therefore, there is a statistically significant difference, with the BED_OCCUPIED_RATE being higher than the ROOM_OCCUPIED_RATE in Emergency Shelter.

In summary, it was found in this study that at the aggregate level, 'Room Base Capacity' shelters have higher occupancy rates than 'Bed Base Capacity' shelters. However, when broken down into the two subsets of the Program Model, all the subsets show that the occupancy rate of 'Bed Base Capacity' shelters is higher than that of 'Room Base Capacity' shelters. Simpson's Paradox emerges, so the previously out trend for the overall trend (homeless people preferring Room Base Capacity) is no longer reliable, and it is possible that confounding variables are causing interference. Since the trend was the same for all subsets, I believe that the trend in the subset in this study is more reliable, i.e., shelters that prefer Bed Base Capacity.

Since Bed Base Capacity shelters are highly utilized and can usually accommodate more people. Therefore, it is recommended that the government increase the number of Bed Base Capacity shelters to accommodate more homeless people with limited resources to help them survive the winter.

Reference

Homelessness in Toronto - facts and statistics. Fred Victor. (2024, January 25). https://www.fredvictor.org/facts-about-homelessness-in-toronto/

Sprenger, J., & Weinberger, N. (2021, March 24). *Simpson's paradox*. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/paradox-simpson/