Shelter Usage Data Analysis

INF2178

Experimental Design for Data Science

Instructor: Professor Shion Guha

Assignment Introduction

This assignment aims to conduct an exploratory data analysis and quantitative analysis of the shelter usage trends by examining the daily occupancy and capacity of Toronto shelters for the year 2021. Through investigating the occupancy rates and serving user counts for different types of capacity and program modal, this analysis tries to identify potential challenges, and highlight areas for improvement in the existing shelter organizations, so that future established shelters could accept more homeless when they need overnight and shelter services.

Data Pre-processing and Cleaning

The dataset that was used in this assignment is a shelter usage data containing 13 features, including occupancy date, organization name, program ID, program name, sector, program model, overnight service type, program area, capacity type, service user count, capacity actual bed, occupied beds, capacity actual room, and occupied rooms. However, to achieve the goal of this assignment, only 8 features will be chosen from the original dataset for later analysis, and they are program model, capacity type, service user count, capacity actual bed, occupied beds, capacity actual room, and occupied rooms.

First, these 8 features were extracted from the original dataset and they were assigned to a new dataframe. Since this assignment is interested in investigating the occupancy rate of the shelters, a new variable has been created, that is, occupancy rate. The occupancy rate variable was created according to the capacity type; if the type is room based capacity, the occupancy rate was calculated through occupied rooms divided by capacity actual room; if the type is bed based capacity, the occupancy rate was calculated through occupied beds divided by capacity actual bed. To double make sure occupancy rate could be used for later calculation, it was then changed to the type float.

Moreover, for later conveniences when conducting exploratory data analysis (EDA) and t-tests, two separate sub-dataframe were created, they are room based dataframe, which only containing data of room based capacity type, and bed based dataframe, which only containing data of bed based capacity type. Now all the pre-processing and cleaning have been done, ready for later EDA and t-tests.

Exploratory Data Analysis (EDA)

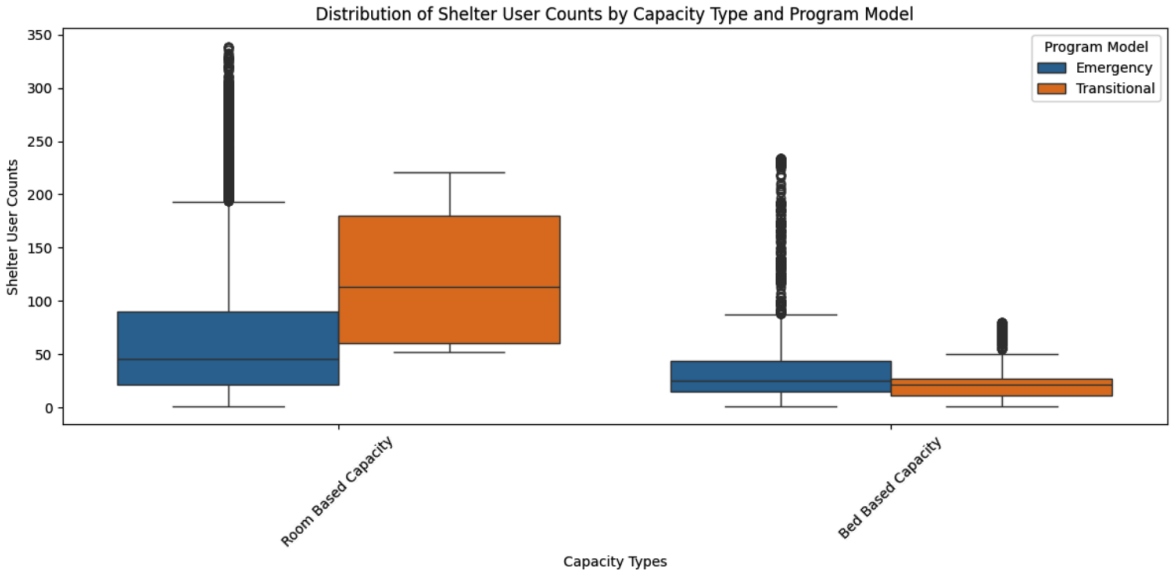
- **Summary statistics**

	Room Based: Service User Count	Bed Based: Service User Count		Room Based: Occupancy Rate	Bed Based: Occupancy Rate
Min	**1**	**1**		**0.01**	**0.02**
Mean	**73.59**	**29.78**		**0.93**	**0.93**
Max	**339**	**234**		**1.0**	**1.0**
25th Percentile	**22.0**	**14.0**		**0.96**	**0.9**
Median	**47.0**	**23.0**		**1.0**	**1.0**
75th percentile	**96.0**	**41.0**		**1.0**	**1.0**
IQR	**74.0**	**27.0**		**0.04**	**0.1**
Standard deviation	**73.32**	**26.38**		**0.16**	**0.12**

By examining above summary statistics of **service user counts** for both room based and bed based capacity, room based capacity is much higher for its mean, max, and median compared to bed based capacity. This shows that room based capacity tends to accept more homeless than bed based capacity in each service. Comparing two standard deviations, room based has more variability. For both capacity types, the mean is bigger than the median, which means that both data are positively skewed (right skewness). This means that mean is more influenced by extreme values, and hence median is more suitable to be the measure of central tendency since it is more robust.

By examining above summary statistics of **occupancy rate** for both room based and bed based capacity, both capacity is very identical for its mean, max, and median. This suggests that both capacities have similar occupancy rates. Comparing two standard deviations, room based has slightly more variability. Again, for both capacity types, the mean is slightly bigger than the median, which means that both data are slightly positively skewed (right skewness), and median is more suitable to be the measure of central tendency.

- **Box plots**



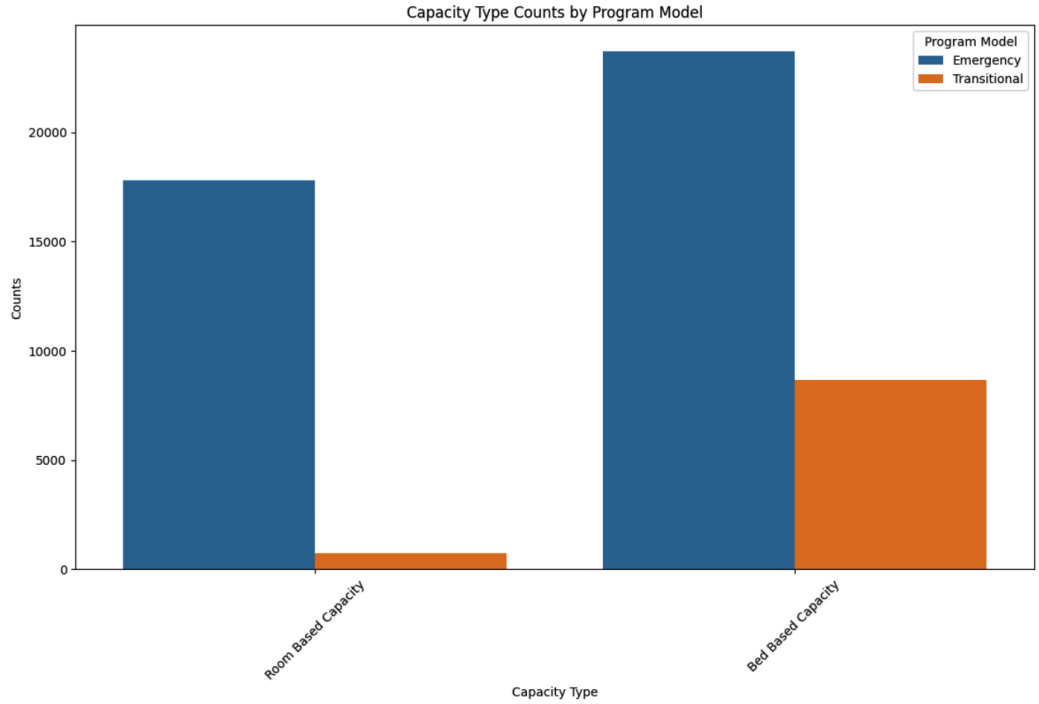
The above combined boxplot shows the distribution of shelter user counts by different capacity type and program model. By examining the resulting boxplots as a whole, it is obvious that the shelter user count of room based capacity is higher than bed based capacity in general.

For room based capacity, the spread of the middle 50% of the data is about between 25 and 90 for the emergency program and is about between 60 and 190 for the transitional program. The latter has a bigger range and higher shelter user counts compared to prior. However, the prior has much more outliers. The median for transitional programs is about 120 and is about 40 for emergency programs, which is also a big difference.

For bed based capacity, the spread of the middle 50% of the data is about between 2 and 50 for the emergency program and is about between 15 and 25 for the transitional program. The prior has a bigger range. The median for transitional programs is about 20 and is about 25 for emergency programs, which is almost the same.

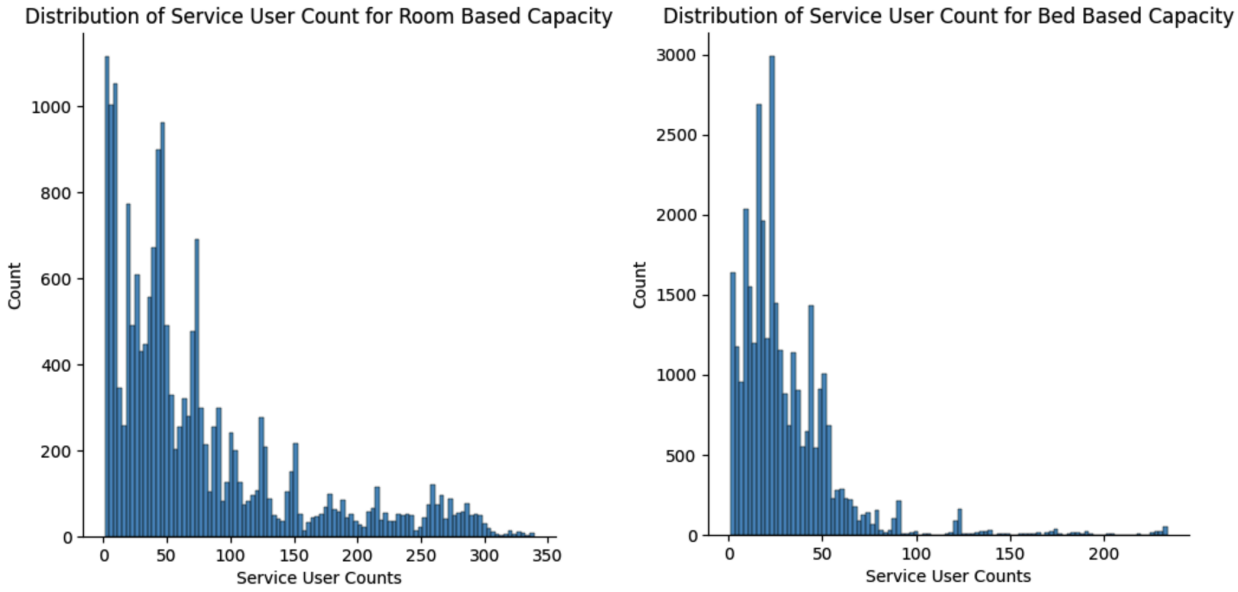
All the boxplot has a right skewness(positively skewed). For both the room based and bed based capacity, large numbers of outliers were found for both capacities, especially for emergency models. The spread of the emergency program on shelter user counts is much wider regardless of the capacity type, which indicates a higher variability.

- **Bar plots**



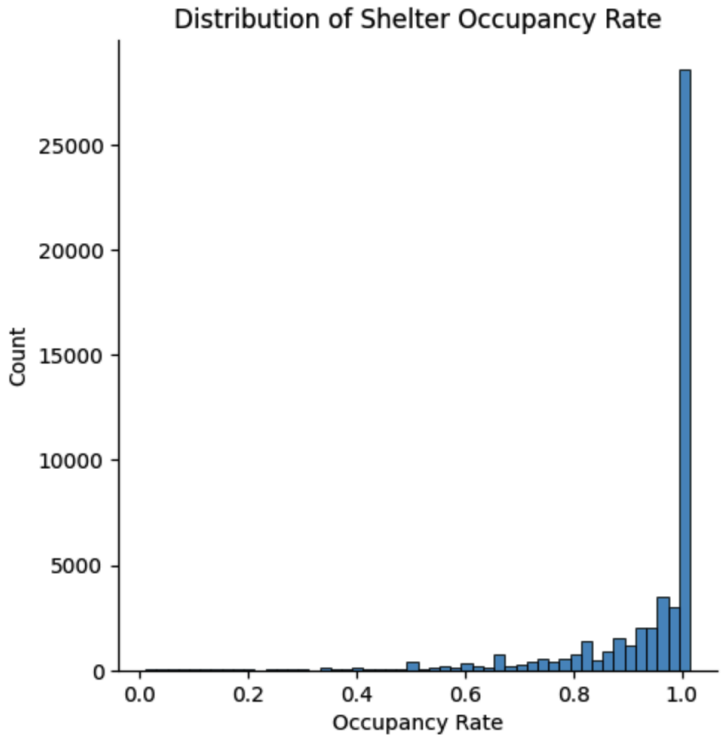
The above combined bar plots show the counts of capacity, categorized by its type and the program model. As a whole, it is obvious that the bed based capacity type is much more than room based capacity type in count. Moreover, both capacities tend to construct emergency programs rather than transitional ones since emergency program counts are much higher than transitional program counts regardless of the capacities.

- **Histograms**



The above histogram **on the left** shows the distribution of service user count for **room based capacity**. As shown above, it is positively skewed (right skewness), meaning the mean is greater than the median and the median could be more appropriate in terms of measuring central tendency since the mean could be influenced by outliers. Also, the above histogram could be seen as a unimodal which has the highest peak above 1000 count at service user count around 5. Nevertheless, there is another second high peak at closing to 1000 count at service user count around 48. The histogram has multiple small peaks after the first main one, hence it may also be considered as a multimodal histogram. Most of the data distributed before 100 service user count, which suggests that most of the room based capacity has its occupancy less than 100 count.

The above histogram **on the right** shows the distribution of service user count for **bed based capacity.** As shown above, it is positively skewed (right skewness) as well, meaning the mean is greater than the median and the median could be more appropriate in terms of measuring central tendency. Also, the above histogram could be seen as a unimodal which has the highest peak almost reaching 3000 count at service user count around 25. Most of the data distributed before 50 service user count, which suggests that most of the bed based capacity has its occupancy less than 50 count. Compared to the previous histogram, bed based capacity is much bigger in its counts at its highest peak than room based.



The Left histogram shows the distribution of shelter occupancy rate as a whole. It is very clear that the distribution is extremely left skewed (left skewness), showing most of the data points are distributed after 0.8. This suggests that the mean is less than the median, and it is more reasonable to measure the central tendency through the median since it is more robust. It is a unimodal histogram. This suggests that regarding the capacity type, most of the shelters are fully occupied.

Quantitative Analysis: T-Tests

In this analysis, two t-tests have been conducted, both are one-tailed t-test:

1. Comparing **the mean of occupancy rate** between room based and bed based capacity
   Null hypothesis: The mean of x (room based occupancy rate) is less than or equal to the mean of y (bed based occupancy rate).
   Alternative hypothesis: There is a significant difference and the mean of x (room based occupancy rate) is greater than the mean of y (bed based occupancy rate).
2. Comparing t**he mean of service user count** between room based and bed based capacity
   Null hypothesis: The mean of x (room based service user count) is less than or equal to the mean of y (bed based service user count).
   Alternative hypothesis: there is a significant difference and the mean of x (room based service user count) is greater than the mean of y (bed based service user count).

	Occupancy Rate T-Test		**Service User Count T-Test**
t-statistic	4.854		97.123
p-value	6.064e-07		0.0

The above table shows the test results for both t-tests. Both are independent two sample t-tests, and they are one-tailed since this study is only interested in investigating whether room based capacity has greater mean on each feature than the bed based capacity. For both t-tests, p-value is much less than 0.05 (assumed significance level), hence we have strong evidence to reject both of the null hypotheses. This suggests that the mean of room based occupancy rate is significantly higher than bed based, and the mean of room based service user count is significantly higher than bed based. The t-statistics suggests the standard errors under the assumption that the null hypothesis is true.

Conclusion

In conclusion, by EDA, room based capacity tends to accept much more homeless when compared to bed based capacity. Regarding the capacity types, there are much more emergency programs than transitional programs, however, room based capacity has a bigger interquartile range and value in transitional programs than in emergency programs. The occupancy rate for shelter usage is extremely high; most of the shelters are fully occupied, which means increased homeless could face a problem of not having a place for overnight stay. By t-tests, room based capacity has significantly higher mean of occupancy rate and service user count. Hence, as a whole, shelter organizations may need to consider changing their strategies in the future, building more room based capacities, and increasing the amount of transitional programs. In future study, handling outliers may need to be considered in the data cleaning section.