INF2178 Assignment1 - An Exploration of Homelessness Trends in 2021
Mingrui Fu 1010506551

1. Introduction

In recent years, the issue of homelessness in Toronto has escalated, posing challenges to the city's shelter support system. Despite efforts to provide overnight and shelter services for the unhoused population, the demand often surpasses the available resources, leading to individuals being turned away due to insufficient shelter spaces. To gain insights into the shelter usage trends and understand the dynamics of the crisis, this exploratory data analysis focuses on a dataset titled "INF2178-A1-data.xlsx," which meticulously tracks daily occupancy and capacity of Toronto shelters throughout the year 2021. Through Python libraries, including pandas, numpy, seaborn, and matplotlib, the dataset is loaded and prepared for analysis.

1. Research Question and Method

The dataset offers a plethora of information, and certain columns stand out as potentially crucial for shedding light on the homeless situation. Key variables include CAPACITY_TYPE, PROGRAM_MODEL, SERVICE_USER_COUNT, CAPACITY_ACTUAL_BED, OCCUPIED_BEDS, CAPACITY_ACTUAL_ROOM, and OCCUPIED_ROOMS. With a particular emphasis on quantitative analysis through t-tests and exploratory data analysis using Python libraries like seaborn and matplotlib, this study aims to answer several research questions:

- How do categorical variables like CAPACITY_TYPE and PROGRAM_MODEL influence continuous variables such as SERVICE_USER_COUNT and OCCUPANCY_RATE?
- Can t-tests reveal significant differences in continuous variables based on shelter program characteristics?
- What insights can exploratory data analysis provide regarding shelter occupancy rates, and how can visualizations contribute to our understanding of the data?

The first step involves loading the data set and performing initial exploration to understand its structure, including data types, missing values, and basic statistics. Subsequently, relevant columns such as CAPACITY_TYPE, PROGRAM_MODEL, SERVICE_USER_COUNT, CAPACITY_ACTUAL_BED, OCCUPIED_BEDS, CAPACITY_ACTUAL_ROOM, and OCCUPIED_ROOMS are selected for in-depth analysis. Quantitative analysis using t-tests is employed to compare continuous variables based on categorical ones. For instance, t-tests can be conducted to explore differences in OCCUPANCY_RATE between different PROGRAM_MODEL categories or to investigate variations in OCCUPANCY_RATE based on CAPACITY_TYPE. Additionally, exploratory data analysis is performed using Python visualization libraries like seaborn and matplotlib. Visualizations, such as box plots, histograms, and count plots, are created to uncover patterns, trends, and potential outliers in the data. This narrative exploration aims to tell a preliminary story about shelter usage trends and highlight areas that may require further analysis to address specific research questions effectively. The details of how each step being operated is showing below:

Initial Exploration:
Understanding the dataset's structure is paramount. With 50,944 entries and 14 columns, the data covers various aspects of shelter occupancy. Initial exploration reveals information of the data set, including the size and type of each columns, summary statistics of each columns, including count, mean, standard deviation, min, max and quantiles, and also missing values, emphasizing the need for data cleanliness.

Selecting Relevant Columns and Exploring Data:
To streamline the focus, key columns such as 'CAPACITY_TYPE', 'PROGRAM_MODEL', and occupancy metrics are selected. This subset allows for a detailed exploration of service user counts, capacity, and occupancy across different program models and capacity types. Since both 'CAPACITY_TYPE’ and 'PROGRAM_MODEL' have more than one category, the first this is to group the data, then compute the statistics for the subgroups and draw the box plots for each of the service user counts, capacity, and occupancy. The occupancy rate is calculated by dividing occupied rooms or beds by actual capacity.

Perform T-Tests:
T-tests are employed to assess differences in service user counts and occupancy rates between different program models and capacity types. One of the conditions of using the T-test is samples are approximately normally distributed. The discrete variable does not satisfy this condition, so that is the reason why we need to create the continuous variable ‘OCCUPANCY_RATE’ to conduct T-tests. In this research, Welch’s T-test is used, since the sample sizes are unequal. Significant variations in service user counts between Emergency and Transitional program models are noted. Additionally, occupancy rates show distinctions between Bed Based and Room Based capacity types. The details of the result interpretation would be analyzed in the following Result section.

Visualizations:
Visualizations play a crucial role in storytelling. Boxplots and heatmaps are utilized to illustrate the distribution of service user counts, capacity, and occupancy across program models and capacity types. These visual aids enhance the understanding of shelter usage dynamics. Besides, the exploration data analysis is the initial step of the entire data analysis process, this visualization step is a good help for discovering more information other than the selected columns and providing foundation for following steps of the entire data analysis process.

2. Result

In this section, we analyze the result we found by using the methods mentioned in the above section. In the initial exploration step, there are totally 50944 entries and the missing values are a few, so the data set is very clean. However, we should notice that the missing values of ‘OCCUPIED_BEDS’ plus the missing values of ‘OCCUPIED_ROOMS’ are 50944, and the same things also happened to the missing values of ‘CAPACITY_ACTUAL_BED’ and ‘CAPACITY_ACTUAL_ROOM’, which means if the ‘CAPACITY_TYPE’ is bed based, then the data are recorded under the ‘CAPACITY_ACTUAL_BED’ and ‘OCCUPIED_BEDS’. if the ‘CAPACITY_TYPE’ is room based, then the data are recorded under the ‘CAPACITY_ACTUAL_ROOM’ and ‘OCCUPIED_ROOMS’. Additionally, from the summary statistics we could find that the mean of the service users of year 2021 is 45, it is less than the mean of actual room and bed capacities, which means usually these shelters provided enough spaces for the homeless people, but may be the opposite would happen on a specific day.

The box plots are used to examine the relationships between different program models/ capacity types (selected categorical variables) and service user/ room or bed capacity/ occupancy rate(selected discrete and continuous variables). Plots are shown below:

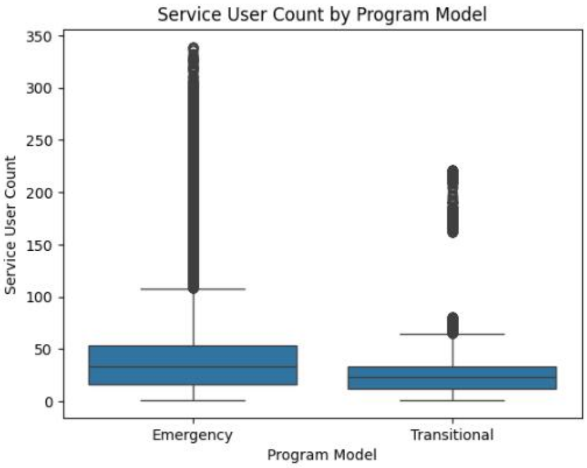

Fig 1: Service user count by model

For both two models, their centers are around 40, but emergency model has more service users around this center and it has more outliers than that of transitional model. Similar trend is also for the relationship between program model and ‘CAPACITY_ACTUAL_ROOM’,‘CAPACITY_ACTUAL_BED’,‘OCCUPIED_ROOMS’, ‘OCCUPIED_BEDS’. For capacity type, the center of service user is around 50 for the room based type and it is around 30 for the bed based type, more service users prefer room based type, plot is shown as below:

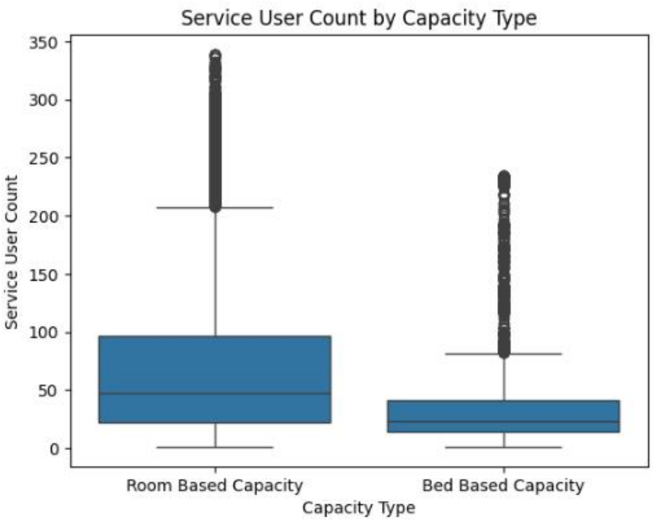

Fig 2: Service user count by capacity type

Similar trend is also for the relationship between capacity type and 'CAPACITY_ACTUAL_ROOM','CAPACITY_ACTUAL_BED','OCCUPIED_ROOMS', 'OCCUPIED_BEDS'. For occupancy rate, room based capacity and emergency model have higher rate, and the rates of bed based capacity and transitional model have more variance.

After the Welch's T-tests are conducted, in T-test of OCCUPANCY_RATE for different PROGRAM_MODEL categories, the p-value is close to 0, which means we should reject the null hypothesis, and there is a huge difference of occupancy rate between two models. In T-test of OCCUPANCY_RATE for different CAPACITY_TYPE categories, the p-value is also close to 0, which means we should reject the null hypothesis, and there is a huge difference of occupancy rate between two kind of capacities. These findings suggest potential factors influencing shelter usage trends.

To explore more information of the data set, some other kind of plots are shown below:

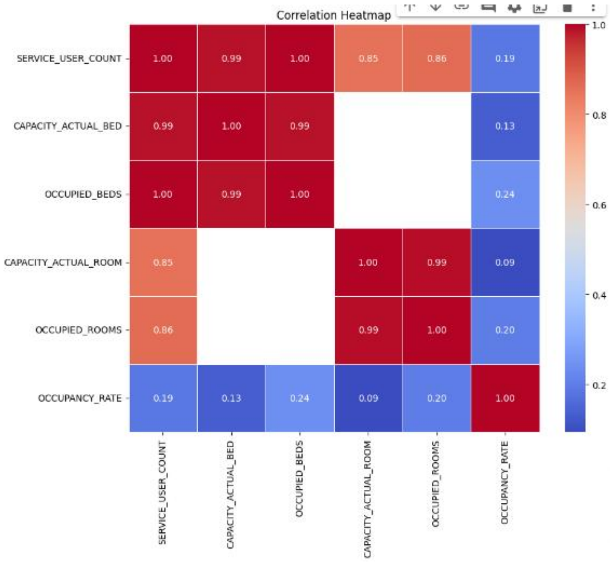

Fig 3: Correlation heatmap

We can find that those discrete variables in the data set affect each other, but occupancy rate has low correlation with those discrete variables. In the initial exploration step, we mentioned that usually these shelters provided enough spaces for the homeless people. In the histograms below, when count is lower than 50, the frequency of service user count is higher than the total of room and bed capacities, This conclusion is contrary to what was previously reached, and we need do more analysis.

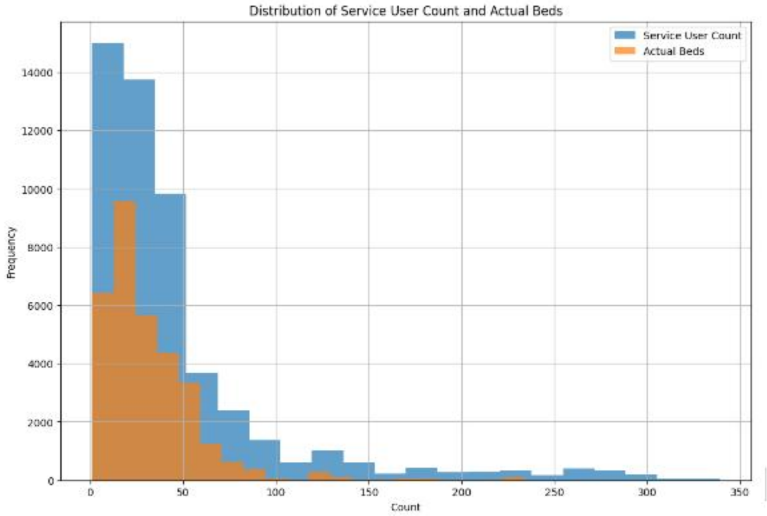

Figure 4: Distribution of Service User Count and Actual Beds

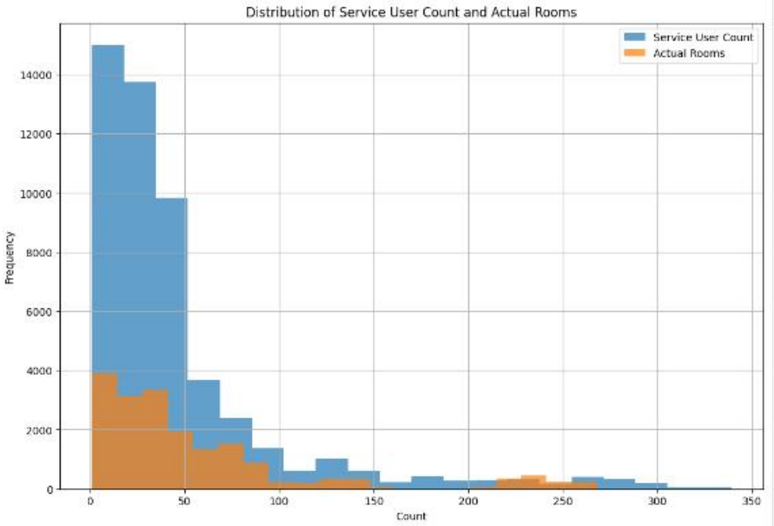

Figure 5: Distribution of Service User Count and Actual Rooms

3. Conclusion

In conclusion, this exploratory data analysis sheds light on the complex dynamics of shelter usage in Toronto throughout 2021. Leveraging the dataset "INF2178-A1-data.xlsx" and employing quantitative analysis through t-tests and exploratory data analysis using Python libraries, we have addressed key research questions regarding the influence of categorical variables like CAPACITY_TYPE and

PROGRAM_MODEL on continuous variables such as SERVICE_USER_COUNT and OCCUPANCY_RATE. The initial exploration emphasized the cleanliness of the dataset, but the presence of missing values and certain anomalies, like the inverse relationship between service user counts and room/bed capacities, calls for further scrutiny. The T-tests revealed significant variations in service user counts and occupancy rates based on program models and capacity types, suggesting potential factors influencing shelter usage trends.

Moving forward, several future steps could enhance our understanding. A temporal analysis exploring monthly or seasonal variations could unveil trends influenced by external factors. Incorporating geospatial data might reveal spatial patterns in shelter demand, while machine learning models could predict future occupancy and identify contributing factors. Qualitative analysis, involving interviews or surveys, would complement quantitative findings, providing a more comprehensive picture.