Kevin Nitièma | 1002673298
INF2178H: Experimental Design for Data Science – Assignment 1

Exploring Shelter Usage Trends in Toronto

1. Introduction

Over the past few years, Toronto has faced a notable rise in its homeless population. While the City of Toronto's shelter support system strives to offer overnight and shelter services to those without homes, a pressing challenge persists – a frequent inability to accommodate all individuals seeking shelter due to unavailable spaces. This pressing issue led us to understand and address the patterns within Toronto's shelter system.

This report offers a comprehensive exploratory data analysis of Toronto's shelter usage throughout 2021 with a goal to uncover the underlying trends surrounding homelessness in the city. To unravel these patterns, we delve into the dataset named '*INF2178_A1_data.xlsx*' (accessible in this GitHub repository), tracking the daily occupancy and capacity of Toronto shelters for the specified year.

Our exploration will address three fundamental research questions, serving as guiding principles in unraveling the troubling nature Toronto's homelessness patterns:

1. **Research Question 1:** How does shelter occupancy vary over the months of the year, and are there any significant differences between cold and warm seasons?
2. **Research Question 2:** Are there significant differences in the occupancy rates of emergency and transitional shelter programs, and if so, in which direction?
3. **Research Question 3:** Do programs with room-based capacity exhibit different occupancy characteristics compared to programs with bed-based capacity?

By addressing these questions, we aim to contribute insights into the dynamics of homelessness in Toronto and provide a deeper understanding that can inform more effective interventions.

2. Data Cleaning and Data Wrangling

The raw dataset has a total of **14 columns** with **50,944** entities (**rows**). After initial review of the dataset, we were confident that not much data cleaning was deemed necessary for the scope of our analysis. However, we noted any observed discrepancies and defined new features necessary for future analysis. Below we outlined our observations and all new features added to our dataset:

A. Observations and Considerations:

1. Since our analysis is quantitative, we've reduced our working dataset to the following columns source from the raw dataset. Below we provided a short description of each column:
    - `OCCUPANCY_DATE:` date of the collected record collected;
    - `PROGRAM_MODEL:` classified as either Emergency or Transitional;

- `SERVICE_USER_COUNT:` daily count of the number of service users;
- `CAPACITY_TYPE:` classified as bed-based or room-based capacity;
- `CAPACITY_ACTUAL_BED:` number of beds offered by the shelter;
- `OCCUPIED_BEDS:` number beds showing as occupied by a shelter;
- `CAPACITY_ACTUAL_ROOM:` number of rooms offered by the shelter;
- `OCCUPIED_ROOMS:` number rooms showing as occupied by a shelter.

2. The following columns seemed to have many missing values (`NaN`):
   - `CAPACITY_ACTUAL_BED` and `OCCUPIED_BEDS` each have **32399 non-null**
   - `CAPACITY_ACTUAL_ROOM` and `OCCUPIED_ROOM` each have **18545 non-null**
     However, after critically reviewing the data, we noted that these pairs of columns are complimentary. This means that a missing pair of data is accounted for by the other pair.

B. Feature Engineering:

Here, we created three (3) new features to add to our dataset to aid in later analysis. The features are as follows:

1. `OCCUPANCY_RATE:` this feature refers to the percentage of occupied beds or rooms in a shelter; it measures how fully utilized the available capacity is at a given time.
   - $Occupancy\_Rate\ (\%) = \frac{Number\ of\ Occupied\ Beds\ or\ Rooms}{Total\ Capacity\ (Beds\ or\ Rooms)} \times 100$
2. `OCCUPANCY_MONTH:` this feature refers to the month in which the data was collected. We sourced this from the `OCCUPANCY_DATE` column. Each month corresponds to a number.
   - E.g.: January -> 1, February -> 2 … December -> 12
3. `OCCUPANCY_SEASON:` this feature refers to the season in which the data was collected. We sourced also this from the `OCCUPANCY_DATE` column
   - SPRING: March 20 - June 20
   - SUMMER: June 21 - September 21
   - FALL: September 22 - December 20
   - WINTER: December 21 - March 19

3. Exploratory Data Analysis (EDA)

After adding our new features to our new working dataset, we proceeded with a comprehensive EDA to leverage insight that could potentially lead to **interesting research questions**. We started by describing our quantitative data as shown in *Figure 1* below. Additionally, we employed boxplots seen in *Figure 2* to visually represent the distribution of these features, after removing outliers. This descriptive analysis offered a **clearer understanding of the general trends within each feature**.

	SERVICE_USER_COUNT	CAPACITY_ACTUAL_BED	OCCUPIED_BEDS	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS	OCCUPANCY_RATE
count	50944	32399	32399	18545	18545	50944
mean	45.72717101	31.62714899	29.780271	55.54925856	52.798598	93.01423308
std	53.32604926	27.12768152	26.37941613	59.44880523	58.7929541	13.87878024

min	1	1	1	1	1	1.2
25%	15	15	14	19	16	92.31
50%	28	25	23	35	34	100
75%	51	43	41	68	66	100
max	339	234	234	268	268	101.41

Figure 1: Dataset Quantitative Data Statistics

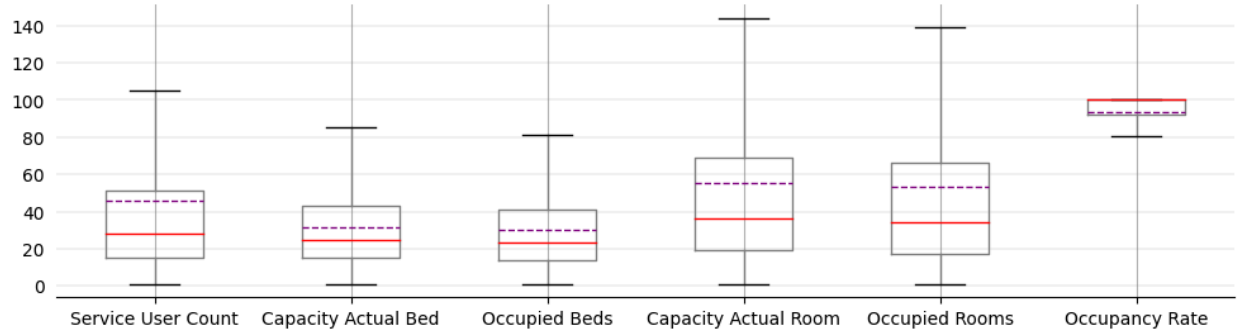


Figure 2: Boxplot of Quantitative Data Without Outliers

Based on the boxplot above, where the median is denoted in red and the mean in a dotted purple line, we observed that the distribution patterns of `CAPACITY_ACTUAL_BED` align remarkably with `OCCUPIED_BEDS`, and similarly, `CAPACITY_ACTUAL_ROOM` aligns with `OCCUPIED_ROOM`. This apparent similarity suggests a robust relationship between these paired variables. The observation may hint at the possibility that **shelters consistently operate at, or near, their full capacity**.

To validate our hypothesis and explore potential relationships in other features, we generated a correlation matrix (*Figure 3*) to provide a holistic overview of the relationships between variables. As seen on the correlation matrix, there exists a very **strong positive relation** (r = 0.99) between the paired features. This figure proved instrumental in validating our initial assumptions regarding the overwhelming nature of shelters.

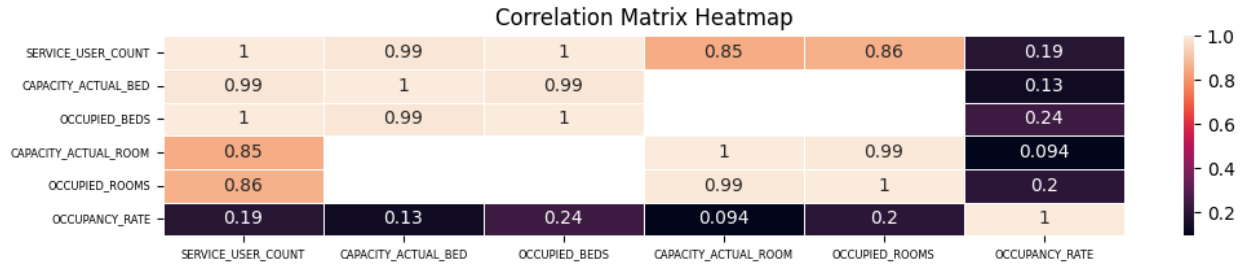


Figure 3: Correlation Matrix of Quantitative Features

Next we looked more closely at the `OCCUPANCY_RATE` feature since it provided an accurate representation of a shelter's activity, regardless of its capacity type (i.e., bed or room). The `OCCUPANCY_RATE` distribution below (*Figure 4*) is highly left-skewed, indicating a prevalence of higher frequencies gravitating towards full capacity. This skewness in the histogram serves as a visual confirmation of the **dominant pattern of shelters operating at their maximum occupancy**.

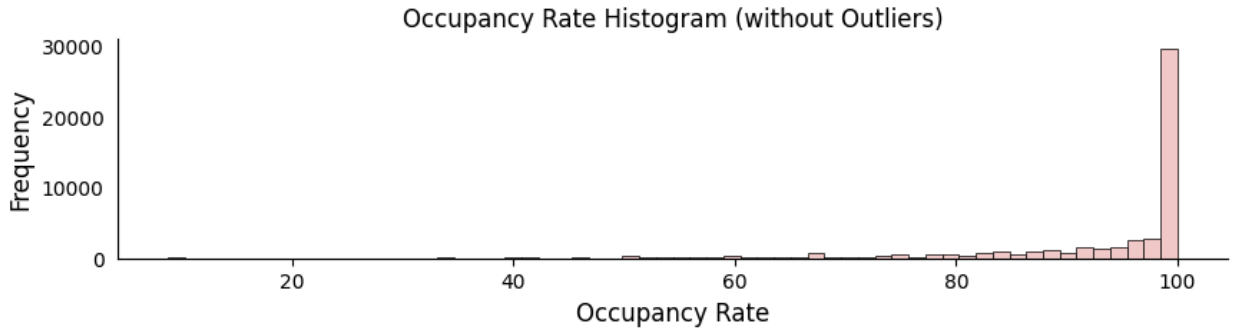


Figure 4: Correlation Matrix of Quantitative Features

Additionally, our analysis revealed a troubling trend in 2021: **four (4) shelters exceeded their capacity limits**. Again, this highlights the significant challenges shelters face due to overwhelming demand surpassing available resources. (*Figure 5*)

OCCUPANCY_DATE	PROGRAM_MODEL	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS	OCCUPANCY_RATE (%)
2021-06-11 0:00:00	Emergency	79	80	101.27
2021-06-22 0:00:00	Emergency	71	72	101.41
2021-06-23 0:00:00	Emergency	73	74	101.37
2021-06-24 0:00:00	Emergency	71	72	101.41

Figure 5: Shelters Exceeding Capacity in 2021

Having performed a thorough examination of our dataset, our discoveries led to three key research questions. These questions aim to deepen our understanding of (1) the variation in occupancy rates across different seasons of the year, (2) the comparative analysis between two program models — Emergency and Transitional, and (3) the distinctions between bed- versus room-based programs.

4. Occupancy Rate across Seasons

Research Question #1: How does shelter occupancy vary over the months of the year, and are there any significant differences between cold and warm seasons?

For this analysis, we aimed to first investigate how shelter occupancy varies over the months of the year. From *Figure 6*, which illustrates the month occupancy rate, we observe a noticeable dip in the warmer months of the year (depicted in yellow).

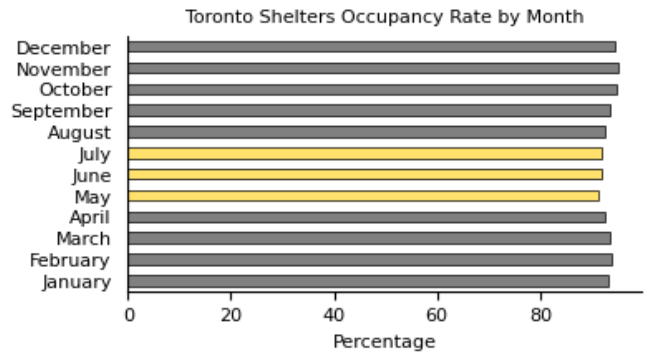
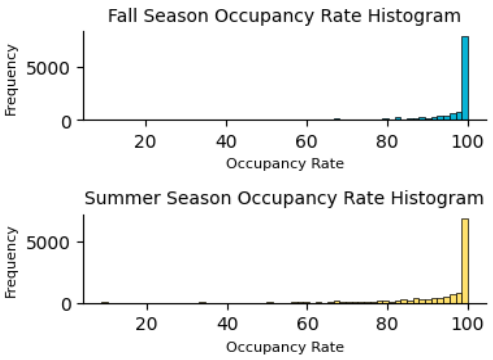


Figure 6: Monthly Toronto Shelters Occupancy Rate



` Figure 7: Fall and Summer Occupancy Rate Histograms

OCCUPANCY_SEASON	OCCUPANCY_RATE (%)
FALL	94.786313
SPRING	92.11799
SUMMER	92.524339
WINTER	93.760738

To expand our understanding of this variation, we calculated the mean `OCCUPANCY_RATE` for each season of the year (*Figure 8*). Based on our results, it seems that **colder months** (Fall and Winter) have a **higher occupancy rate than warmer months**.

Figure 8: Table of Average (mean)Occupancy Rate by Season

To investigate this hypothesis, we compared the `OCCUPANCY_RATE` between **FALL** and **SUMMER** in 2021 using an **independent two-sample t-test.**.

Note: While all conditions and assumptions necessary for conducting a t-test are met, it's crucial to acknowledge that our data does **not adhere to a normal distribution**, as indicated in *Figure 7*. However, since our sample size is very large, the t-test can be somewhat robust to deviations from normality due to the central limit theorem.

T-test Results

- **T-statistic:** 13.709974341349328 | **P-value:** 1.2648784270620004e-42

Our conducted t-tests revealed a significant difference (with $\alpha = 0.05$), pointing towards colder seasons, specifically Fall, consistently manifesting higher shelter occupancy rates. This observation confirms the heightened demand for shelter during colder months and sheds light on how seasonal changes can influence the utilization of shelters by people without housing.

5. Program Model and Occupancy Rates

Research Question #2: Are there significant differences in the occupancy rates of emergency and transitional shelter programs, and if so, in which direction?

To explore this question, we first visualized the mean occupancy rates of transitional programs versus emergency programs (*Figure 9*), suggesting a potential distinction between the two models.

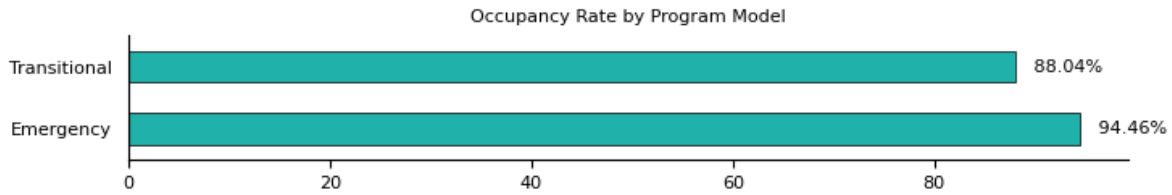


Figure 9: Bar chart comparison of average occupancy rate by program model

We then compared the occupancy rates between emergency and transitional shelter programs using another **independent two-sample t-test**. **Note** that the **assumptions and conditions** for this test align with those of the first research question.

T-test Results

- **T-statistic:** 43.79810975299453 | **P-value:** 0.0

The outcomes of our t-tests presented strong evidence of a significant difference (with $\alpha = 0.05$), with emergency programs consistently boasting higher occupancy rates compared to their transitional counterparts. This statistical insight validates and quantifies the observed trend by emphasizing the role that emergency shelters play in promptly addressing the needs of individuals experiencing homelessness in Toronto.

6. Capacity Type Comparison

Research Question #3: Do programs with room-based capacity exhibit different occupancy characteristics compared to programs with bed-based capacity?

Similar to the methods used in our two previous research questions, when comparing programs with bed-based versus room-based capacity, our **t-tests** suggested a statistically significant difference (with $\alpha = 0.05$) between bed-based and room-based capacity types. Recognizing that our data is left skewed (*Figure 10*), our **assumptions and conditions** for this test align with those of the previous research questions – i.e. the large sample size contributes to the robustness of the t-test, making it less sensitive to deviations from normality.

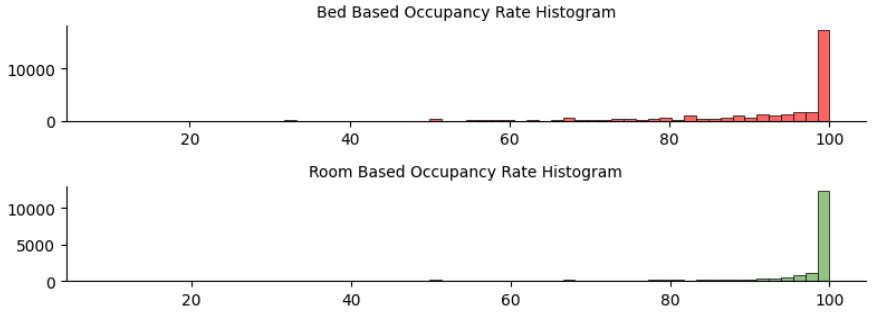


Figure 10: Histograms of occupancy rate for bed-based (top) and room-based (bottom) capacity type

T-test Results

- **T-statistic:** -9.802846881903566 | **P-value:** 1.175399592234721e-22

The negative t-statistic of approx. -9.80 suggests that the choice between bed-based and room-based capacity significantly influences the occupancy rates, with room-based capacity programs exhibiting higher rates compared to their bed-based counterparts.

7. Conclusion

Through quantitative analysis and visual exploration, we gained valuable insights about shelter usage trends in Toronto. Our findings showed a substantial impact of **weather conditions**, **program models**, and **accommodation types** on shelter occupancy rates. Perhaps, this knowledge can guide policymakers in bettering the effectiveness of shelter services such that they can align with the needs of Toronto's unhoused population. Although this analysis provides key insights into shelter occupancy, a potential avenue for future exploration involves re-examining the skewed distributions (as discussed above) using a **log-based model** to potentially derive even more valuable insights from the data.