INF 2178H: Experimental Design for Data Science
Technical Assignment 1
Student name: Ruochen Zhao
Student ID: 1009882478
Submission Date: 2024/02/04

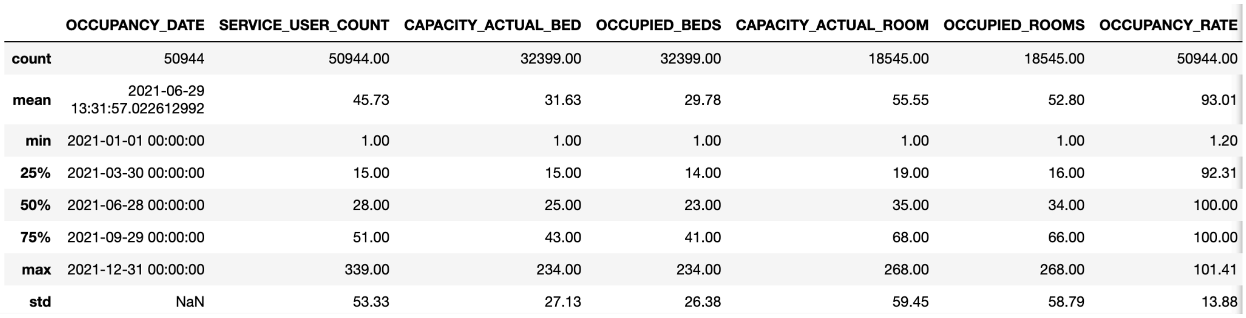
1. Introduction

Toronto's struggle with a substantial population lacking stable housing has become a more severe social problem, a trend highlighted by a 2021 report from the City of Toronto. To dive deeper into this issue, a thorough analysis was performed to examine the shelter usage trends, which records the daily shelter utilization and capacity throughout the year. This investigation aims to construct a detailed account of shelter occupancy over the day of 2021 and scrutinize the daily average occupancy. It is hypothesized that different capacity types will yield different usage of the facility, providing critical insights into how the construction for shelters vary with different. With the aid of Python's data analysis libraries, this data has been analyzed efficiently for a deeper exploration.

2. Data Preparation and Exploratory Data Analysis

2.1 Exploring relationship between daily occupancy rate for bed-based capacity and room-based capacity

Our dataset comprised daily records spanning the entire year of 2021, with a comprehensive lists of different shelter organizations and characteristics of these programs. Since we are only interested in data related to the shelter usage information, we will focus on studying columns including OCCUPANCY_DATE, PROGRAM_MODEL, SERVICE_USER_COUNT, CAPACITY_TYPE, CAPACITY_ACTUAL_BED, OCCUPIED_BEDS, CAPACITY_ACTUAL_ROOM and OCCUPIED_ROOMS. At the end of the dataset, 'OCCUPANCY_RATE' is created to calculate the daily occupancy rate for each type by dividing each row's data for 'OCCUPIED ROOMS' with 'CAPACITY_ACTUAL_ROOM' and each row's data for 'OCCUPIED BEDS' with 'CAPACITY_ACTUAL_BED'. Below is the summary statistic table for these columns.



	OCCUPANCY_DATE	SERVICE_USER_COUNT	CAPACITY_ACTUAL_BED	OCCUPIED_BEDS	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS	OCCUPANCY_RATE
count	50944	50944.00	32399.00	32399.00	18545.00	18545.00	50944.00
mean	2021-06-29 13:31:57.022612992	45.73	31.63	29.78	55.55	52.80	93.01
min	2021-01-01 00:00:00	1.00	1.00	1.00	1.00	1.00	1.20
25%	2021-03-30 00:00:00	15.00	15.00	14.00	19.00	16.00	92.31
50%	2021-06-28 00:00:00	28.00	25.00	23.00	35.00	34.00	100.00
75%	2021-09-29 00:00:00	51.00	43.00	41.00	68.00	66.00	100.00
max	2021-12-31 00:00:00	339.00	234.00	234.00	268.00	268.00	101.41
std	NaN	53.33	27.13	26.38	59.45	58.79	13.88

Table 1: Summary statistic

Since this is time-series data, we need to check the stationarity of it first, and I choose to draw the trend of daily mean occupancy rate for each capacity type. From the line chart below, it can be seen that there is a fluctuation in occupancy rates for both types of capacity throughout the year. However, the daily change is not so fluctuational according to the summary table above, which shows that the more than 50% of the occupancy rate is 100% or slightly above 100%, and most data fluctuates by no more than 5%. With this information, we can use a t-test in the next section to see if there is a difference between the daily occupancy rate for these two different capacity types.

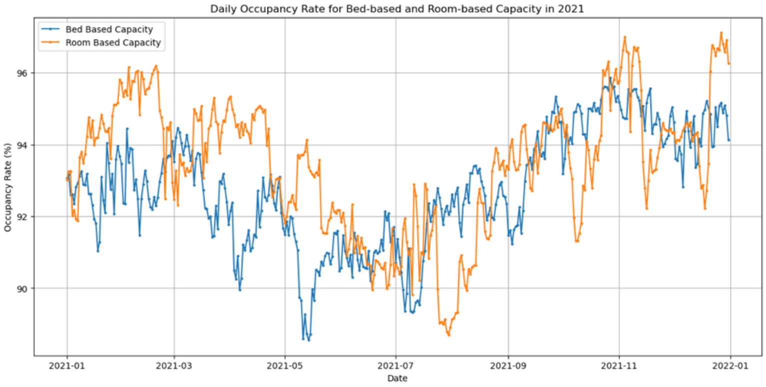


Figure 1: Daily Occupancy Rate for Bed-based and Room-based Capacity in 2021

Another interesting feature I noticed from Figure 1 is the overlapping parts in July, it seems that in July the two capacity types have a similar occupancy rate each day. In order to preliminarily show this information, figure 2 below shows a box plot to compare the occupancy rate within the month July for both occupancy types. Both types of capacities have median occupancy rates that are relatively high. However, the range of occupancy rates seems to be wider for the bed-based capacity compared to the room-based capacity, suggesting more variability in the occupancy rates for bed-based services within the month. This creates some uncertainty for the observation that there seems no difference in mean between the two capacity types. Therefore, a t-test is needed in the later section to test this hypothesis.

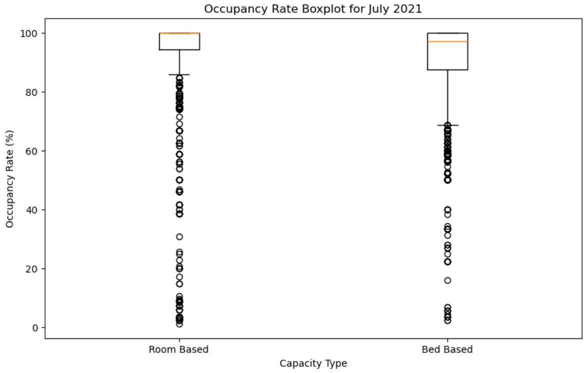


Figure 2: Occupancy Rate Boxplot for July 2021

2.2 Exploring the service user count and occupancy rate for two different program models (emergency and transitional)

Our next step of analysis is to see the user number for emergency and transitional programs, where emergency means it can be accessed by any individual or family experiencing homelessness with or without a referral. One the other hand, a transitional shelter program means that it can only be accessed by referral only. From the histogram we generated, it can be seen that the service user count for both emergency and transitional program models appears to be fairly consistent month over month. In addition, the emergency program model seems to have a higher user count compared to the transitional program model each month, which could indicate a greater need or preference for immediate, short-term shelter solutions within the community.

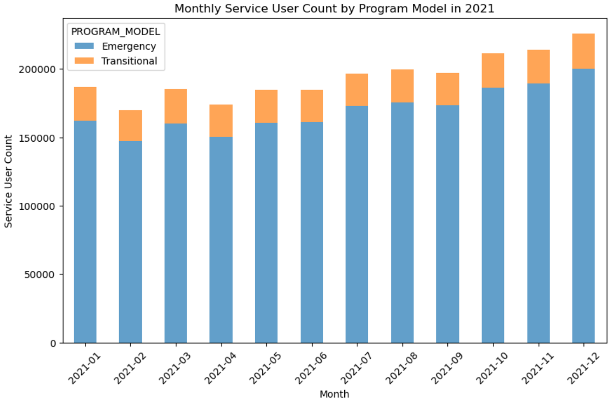


Figure 3: Monthly Service User Count by Program Model in 2021

Another interesting relationship could be explored between the occupancy rate and program model. Similarly, we could ask if there is a difference between the occupancy rate for different program model. From Figure 4 below we can see that the emergency program model appears to have a higher median occupancy rate compared to the transitional program model, where the median is indicated by the line within the box. In addition, the emergency program model shows less variability in occupancy rates, as the box is shorter, which indicates that the values are more clustered around the median. On the other hand, the transitional program model has a longer box, suggesting a wider range of occupancy rates. In the next section, a t-test will be conducted as well to prove our observation from the box plot.

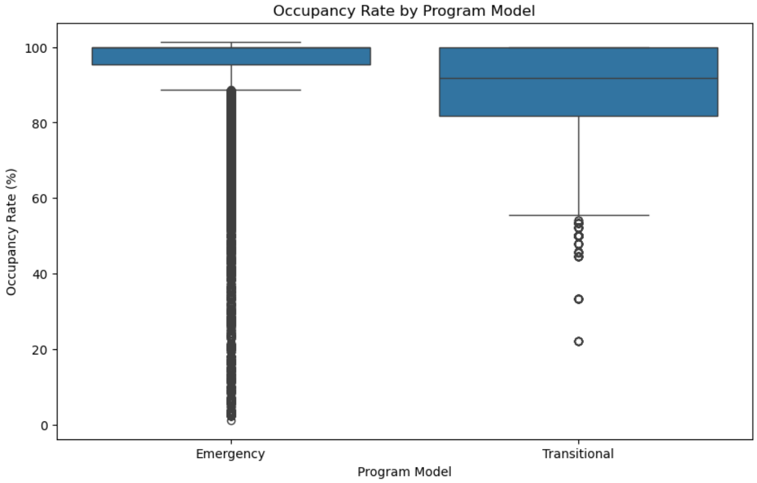


Figure 4: Occupancy Rate by Program Model

When comparing the daily occupancy rate for two program models in a line chart, the figure below shows a similar stationary trend as in Figure 1. However, there is a clear separation for occupancy rate for the two program models, where emergency model shows a higher occupancy rate throughout the year. This might indicate a higher or more consistent demand for emergency shelters. In addition, the emergency program appears to have less variability in its occupancy rate compared to the transitional program, as indicated by the smoother line. This shows that usage for these two models may be influenced by different factors, which could be a further research question.

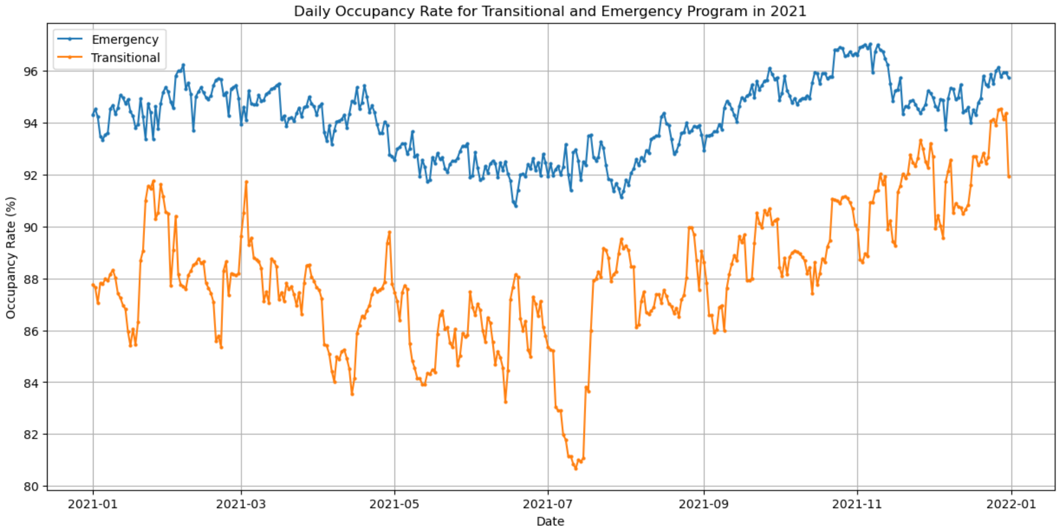
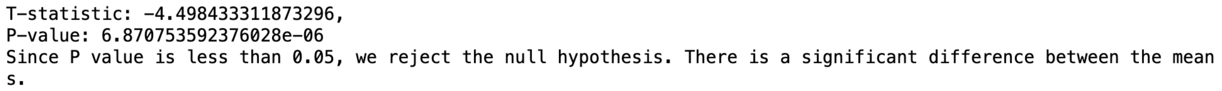


Figure 5: Daily Occupancy Rate for Transitional and Emergency Capacity in 2021

3. Quantitative Analysis using t-tests

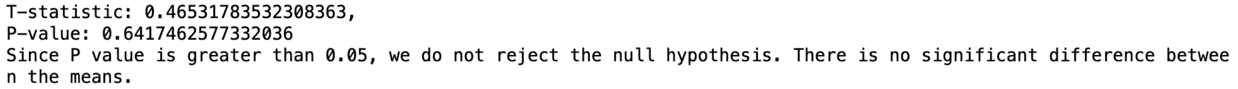
3.1 Statistical Significance of Occupancy Rates Across Shelter Capacities

The next step is to conduct a more formal statistical analysis, and a Welch two sample t-test (given the possible unequal sample size) was employed to see if the differences observed between bed and room occupancy rates were statistically significant. Here the Null Hypothesis (H0) is that there is no significant difference between the daily mean occupancy rate for bed-based capacity and room-based capacity, and the Alternative Hypothesis (H1) is that there is a significant difference between the means. From the t-test result below, it can be seen that the P-value is far less than 0.05, which means we reject the null and state that there is a difference between these means.



```
T-statistic: -4.498433311873296,
P-value: 6.870753592376028e-06
Since P value is less than 0.05, we reject the null hypothesis. There is a significant difference between the mean
s.
```

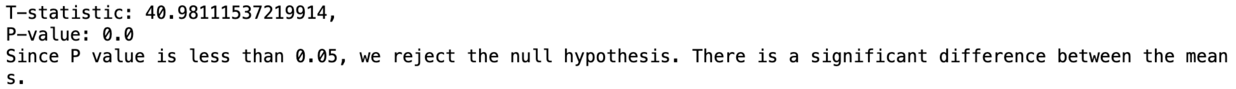
Another feature is the overlapping parts in July in terms of daily occupancy rate for bed-based and room-based capacities, and our next research question could be: will there be no difference between the means of daily usage rate for the two different capacities? Our Null Hypothesis (H0) is that there is no significant difference between the mean daily occupancy rate for bed-based capacity and room-based capacity, while the Alternative Hypothesis (H1) states that there is a difference between the means. The result indicates that since the P-value is greater than 0.05, there is no difference between the mean of the usage rate for bed-based and room-based capacity types. This means during the summer time, both room-based and bed-based capacities show a similar daily usage.



```
T-statistic: 0.46531783532308363,
P-value: 0.6417462577332036
Since P value is greater than 0.05, we do not reject the null hypothesis. There is no significant difference betwee
n the means.
```

3.2 Statistical Significance of Occupancy Rates for Program Models

In this section, we will conduct a t-test to answer the question: whether there is a difference in daily occupancy rate between different program models (emergency and transitional). Here our Null Hypothesis is that there is no difference between the means of daily usage (occupancy rate) for different program models (emergency and transitional), and the Alternative Hypothesis is that there is a significant difference between the means. By conducting the two-sample t-test we know that the P value is 0.0, which is less than 0.05. Therefore, we reject the null hypothesis and state that there is evidence supporting the alternative hypothesis.



```
T-statistic: 40.98111537219914,
P-value: 0.0
Since P value is less than 0.05, we reject the null hypothesis. There is a significant difference between the mean
s.
```

4. Conclusion

The above data analysis conducted on Toronto's shelter usage for the year 2021 has yielded several significant insights into the occupancy patterns. To begin with, the comparison between bed-based and room-based types indicated a discernible variance in occupancy rates. Both categories exhibited a seasonal trend, with a dip in demand during the summer months contrasted by a heightened necessity for shelter services in the winter. This seasonal fluctuation highlights the influence of external factors on shelter utilization and underscores the importance of tailoring shelter capacity planning to these cyclical demand changes.

Furthermore, the statistical evaluation, underpinned by Welch's two-sample t-test, established a clear differentiation in occupancy rates between emergency and transitional programs, with the former sexhibiting consistently higher rates. This distinction underscores the greater demand for emergency shelters within the city's shelter system, possibly reflecting the immediate nature of the housing crises that residents face.

The statistical findings also pointed to the absence of a significant difference in the mean daily usage rates between bed-based and room-based capacities during the month of July, a period where the usage patterns appeared to converge. This observation prompts a deeper examination of seasonal factors or operational dynamics that may influence shelter demands in different ways at various times of the year.

In light of these findings, policymakers and service providers are encouraged to consider the differing demands for shelter types when planning resource allocation and program development.

The evidence supports the need for a robust and responsive shelter system that can adapt to the fluctuating demands for emergency and transitional housing, as well as bed-based and room-based capacity.

Lastly, while this analysis has shed light on key aspects of shelter usage, it also opens avenues for further research. Future studies could explore the underlying causes of the observed occupancy trends, the impact of policy changes on shelter demands, and the role of other social services in shaping these patterns. As the city continues to address its housing challenges, data-driven insights such as these will be invaluable in guiding efforts to provide safe, accessible, and adequate shelter for all residents in need.