INF2178 Assignment 1

Zonglin Li
1004910117

Research Question: The shelter occupancy rate in the Toronto area is affected by the period

Dataset Introduction

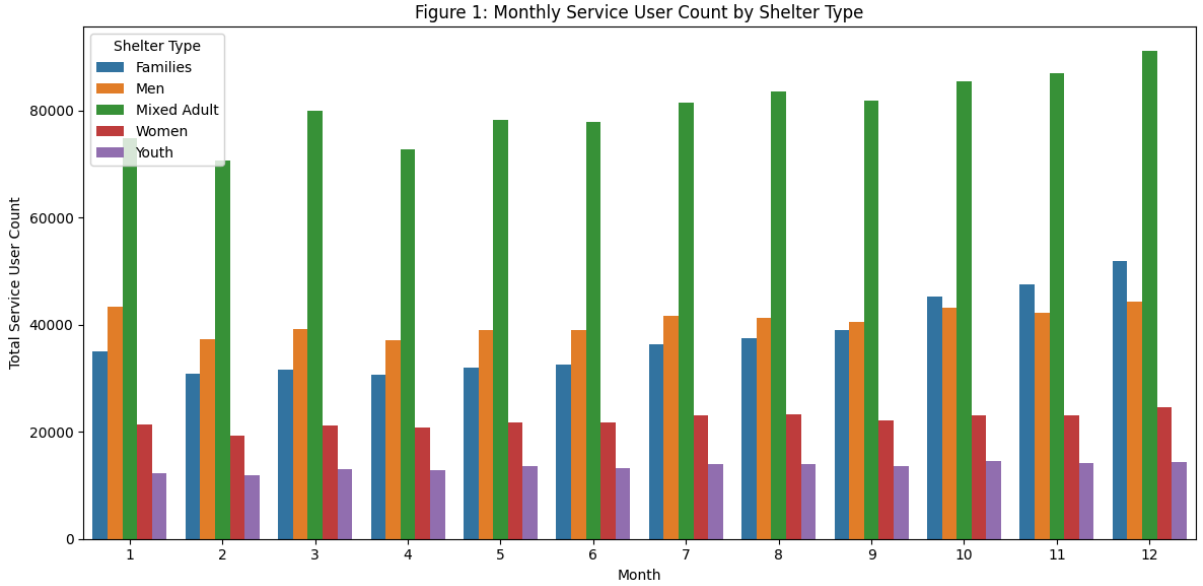
In this assignment, I will statistically analyze a dataset on Daily Shelter & Overnight Service Occupancy & Capacity collected by the City of Toronto to examine whether or not daily shelter occupancy is affected by seasonal variations. This dataset from the Toronto government's official website collects daily updates on shelter and overnight service programs administered by Shelter Support and Housing Management for 2021, including information on the program's operator, location, classification, occupancy, and capacity. The 2021 Shelter Daily Update Summary dataset has a total of 14 different fields, each representing data on the daily operations of the shelter, such as bed occupancy, geographic location, shelter type, and other data types, and the dataset contains a total of 50,944 records, with each record in the dataset representing the operations of a specific shelter on a specific date in 2021. status of a specific shelter on one particular date in 2021.

Modifications to dataset

Since my research topic compares and statistically analyzes quantitative data, some of the categorical variables in the dataset will not appear in my research. I chose to modify the original dataset by using the SELECT COLUMN FUNCTION to keep only 'CAPACITY_TYPE,' 'PROGRAM_MODEL,' 'SERVICE_USER_COUNT,' 'CAPACITY_ACTUAL_BED,' 'OCCUPIED_BEDS,' 'CAPACITY_ACTUAL_BED,' 'OCCUPIED_ROOMS,' 'SECTOR,' 'OCCUPANCY_DATE'. I found that there are two types of CAPACITY_TYPE in the dataset, one for Bed Based Capacity and one for Room Based Capacity, and to make it easier to standardize the calculations, I decided to add a new variable called ‘occupancy rate’ to the original dataset. The occupancy rate is obtained by dividing the number of beds/rooms occupied by the shelter on that day by the number of beds/rooms available and then multiplying by 100 so that it will be more convenient to use the occupancy rate for t-tests and other statistical methods in future calculations. Also, to better visualize the data, I created a new column called 'month' by extracting the month information from OCCUPANCY_DATE. This way, the data is categorized by month to visualize the data better.

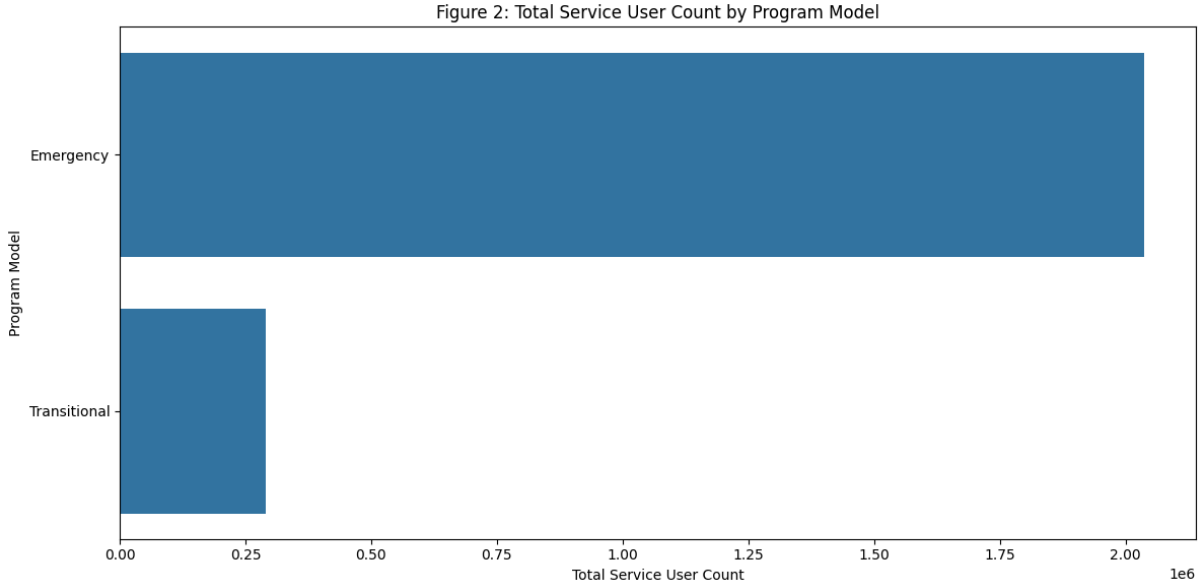
A big picture of shelter usage

Figure 1: Monthly Service User Count by Shelter Type



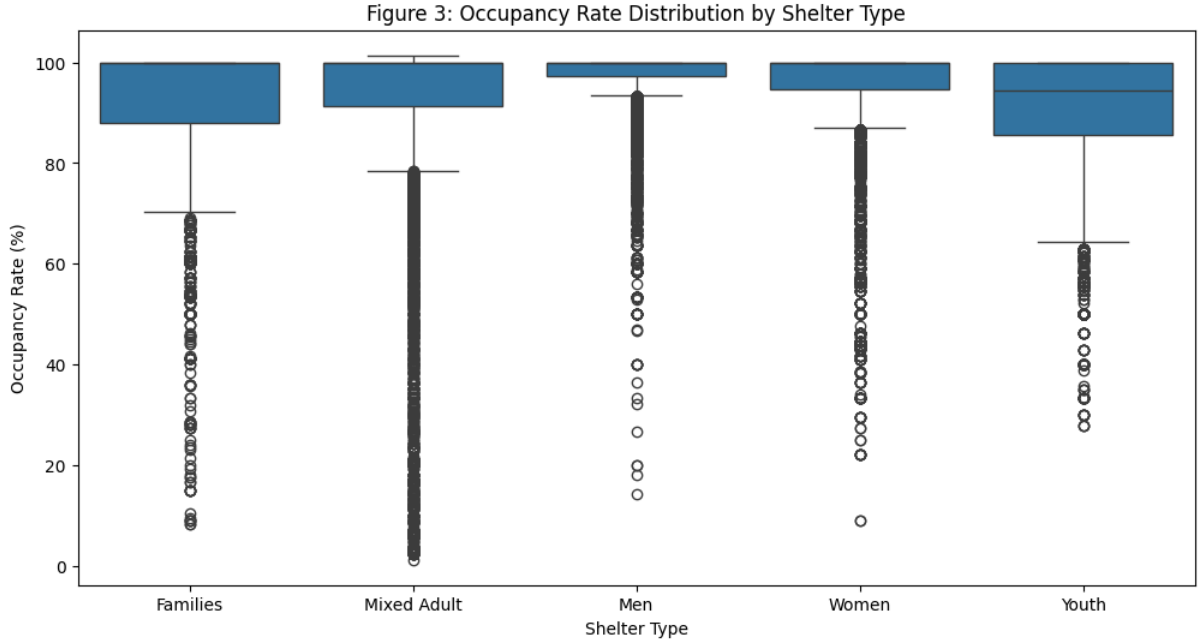
This bar chart shows the total number of users served per month in 2021 for different shelter types, distinguished by different colors. The Y-axis shows the month, and the X-axis shows the number of users served. It is easy to observe that the total number of users of sheltered housing services, specifically for mixed adults, is the highest every month. In December, the number of service users in all types of sheltered housing was the highest for 2021.

Figure 2: Total Service User Count by Program Model



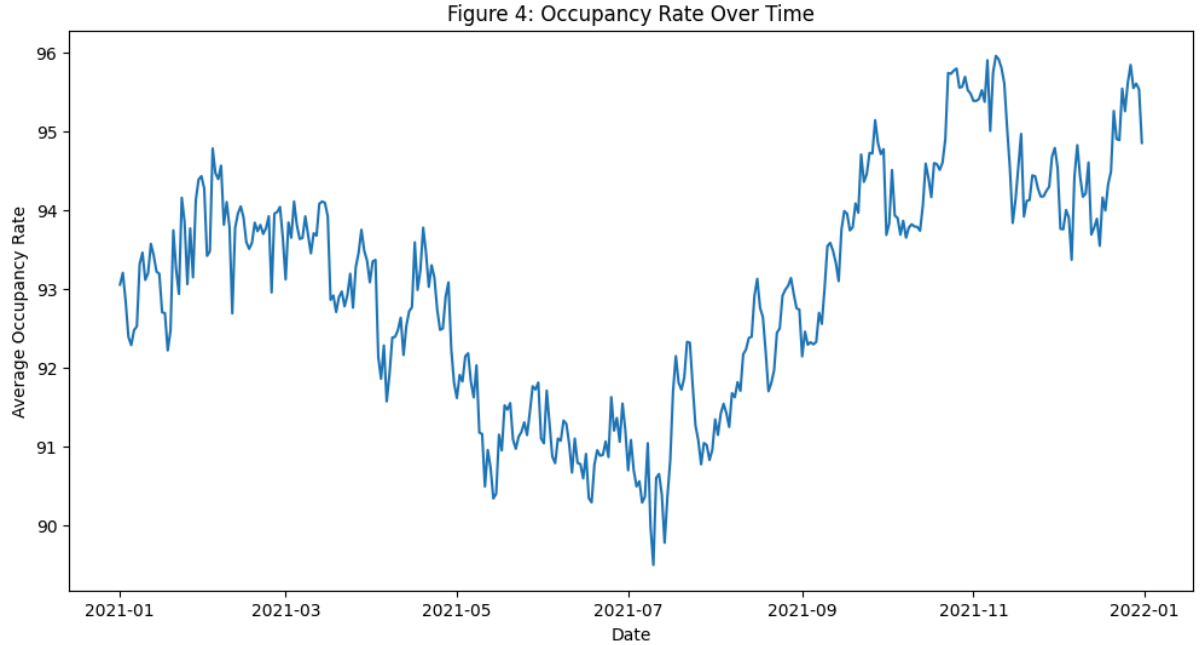
We then counted the total number of emergency shelter and transitional shelter service users and turned the data into a bar graph. We found that shelters that do not require referrals, like any homeless person or family, receive about 90% of the users. In comparison, those requiring referrals and needing to be screened for user criteria make up only 10%, suggesting that even fewer homeless people are seeking help through transitional shelters. Combined with the results of the second bar graph, we can see that the low number of users of shelters that serve specific genders and families may be because most of the homeless are screened out. After all, they cannot qualify for shelter programs. In contrast, the high number of users of shelters that provide shelter services like mixed-gender may be mostly emergency-level shelters.

Figure 3: Occupancy Rate Distribution by Shelter Type



I made boxplots for each shelter type; the occupancy rate is on the Y axis. The graph found that the median daily occupancy rate for almost all types of shelters was 100%, while the mean was also 90%. The following mixed adult type of shelter showed a larger interquartile span, indicating more significant variability in occupancy. Since we have nearly 100% median across all kinds of shelters, the demand for shelters will be high in 2021.

Figure 4: Occupancy Rate Over Time



In Figure 4, I created a line graph to reflect changes in occupancy rate over time; the Y axis represents the average occupancy rate, and the X axis is date by month. Figure 4 shows that from February through July 2021, shelter occupancy gradually declined to a yearly low of about 90 below shelter occupancy. We find shelter occupancy declines month by month throughout the summer of 2021 and by a more significant margin than in the other seasons. From the end of July through January 2022, shelter occupancy rises rapidly. It reached a high of 96% in December 2021, with the most significant increase in shelter occupancy

from September 2021 through December 2021, the most significant increase in shelter occupancy for the year. Looking at the line graph of shelter occupancy over time, I found that the summer of 2021 and the winner of 2021 were the seasons of decreasing and increasing shelter occupancy, respectively. Such a finding interested me in examining whether there is a statistical difference between the average occupancy of the same shelters in the summer and the winter months. Therefore, I will conduct a paired samples t-test in the upcoming study.

Paired t-test analysis

The reason I chose the paired t-test is that we are comparing the differences in test results for the same set of samples under different conditions, and according to my research question, I am comparing whether there is a difference between the mean values of the same shelters in the summer and winter of 2021. First, I define summer as June to August and winter as December to February. First, I filter the occupancy of these seasons and remove the nulls. I use my own 'month' and 'occupancy' columns in this step. Column. After calculating the T-statistic is -13.732150632203595, the P-value is 9.256519912353863e-43, we get a small negative t-value, which means the average occupancy in summer is lower than the average occupancy in winter. The P-value is much less than the 95% level of significance. This represents a statistically significant difference between the mean occupancy of the same shelters in summer and winter, which warrants further research.

```
Families - Summer vs. Winter: T-statistic=-6.301147864895831, P-value=3.4240591253713255e-10
Mixed Adult - Summer vs. Winter: T-statistic=-9.99188699937482, P-value=2.2893269759118674e-23
Men - Summer vs. Winter: T-statistic=-1.1295823645188705, P-value=0.2586962292557958
Women - Summer vs. Winter: T-statistic=-11.135510141256827, P-value=1.9108280299836974e-28
Youth - Summer vs. Winter: T-statistic=-0.24141315095392604, P-value=0.8092470288426543
```

I then used the same t-test methodology to test the sectors of the different hospitality shelters. I found that the p-values of the shelters specializing in males and the shelters serving adolescents were above the 0.05 level of significance, which suggests that the mean values of these two categories of shelters are not affected by the season.

Result

Through quantitative analysis and data visualization of the dataset on the daily situation of shelters, we obtained some instead striking findings, in which the number of users of mixed-gender shelters is much larger than that of other types of shelters in all 12 months of 2021, while only about 10% of the homeless can move into shelters that require referrals, and based on the occupancy rate over time A line graph of occupancy rates over time also observes a pattern in occupancy rates over time. A paired t-test with a p-value shows that the overall shelter's average occupancy rate varies by season, and further research is needed to determine why this is the case.

Discussion

We found significant differences in the average occupancy of shelters across seasons through T-testing, but is this just due to the period as a factor? With more in-depth investigation, we need to consider more relevant factors, such as the most crucial difference between seasons, which is the temperature. We commonly accept that Toronto has more extreme cold weather in winter; for the homeless community, living outdoors during the low temperature is very likely to lead to illness or even death, so it may lead to the homeless being more willing to go to shelters to seek help, and in the summer time because of the temperature outdoor are not as

dangerous to spend the night outside because of the warmer temperatures. Hence, occupancy rates tend to decrease during the summer months.