Technical Assignment #1

Jiayin Huo

University of Toronto

INF 2178

Professor Shion Guha

2024/2/1

Understanding Homelessness Trends in Toronto

1. Project overview

In Toronto, the homeless population is increasing year by year, which is causing social concern. This study aims to use data analysis to gain an in-depth understanding of the trends, characteristics and key factors of Toronto's homelessness problem, and to provide strong support for relevant policies and social intervention.

This study uses a homeless shelter data set from the City of Toronto, which includes multiple key variables, such as the number of service users, bed availability, shelter type, etc. The data cleaning process includes handling missing values, outliers.

2. Data cleaning, preprocessing

To ensure the quality and availability of data, we do data cleaning and preprocessing including the following two parts:

- Delete unnecessary columns such as 'PROGRAM_ID', 'PROGRAM_NAME', 'ORGANIZATION_NAME', 'OCCUPANCY_DATE'.
- Fill missing values for column 'PROGRAM_MODEL', 'OVERNIGHT_SERVICE_TYPE', 'PROGRAM_AREA' to retain useful information.

3. T-test

The t-test (t-test) is a statistical method used to compare whether there is a significant difference in the means of two samples. It is widely used in research and experiments, especially in small sample situations. The basic idea of t-test is to compare the means of two samples to determine whether they come from the same population. Specific analysis methods include:

- One-sample t-test: Used to test whether the mean of a specific variable is equal to a given value.
- Independent samples t-test: Used to compare whether there are significant differences in means between different groups.

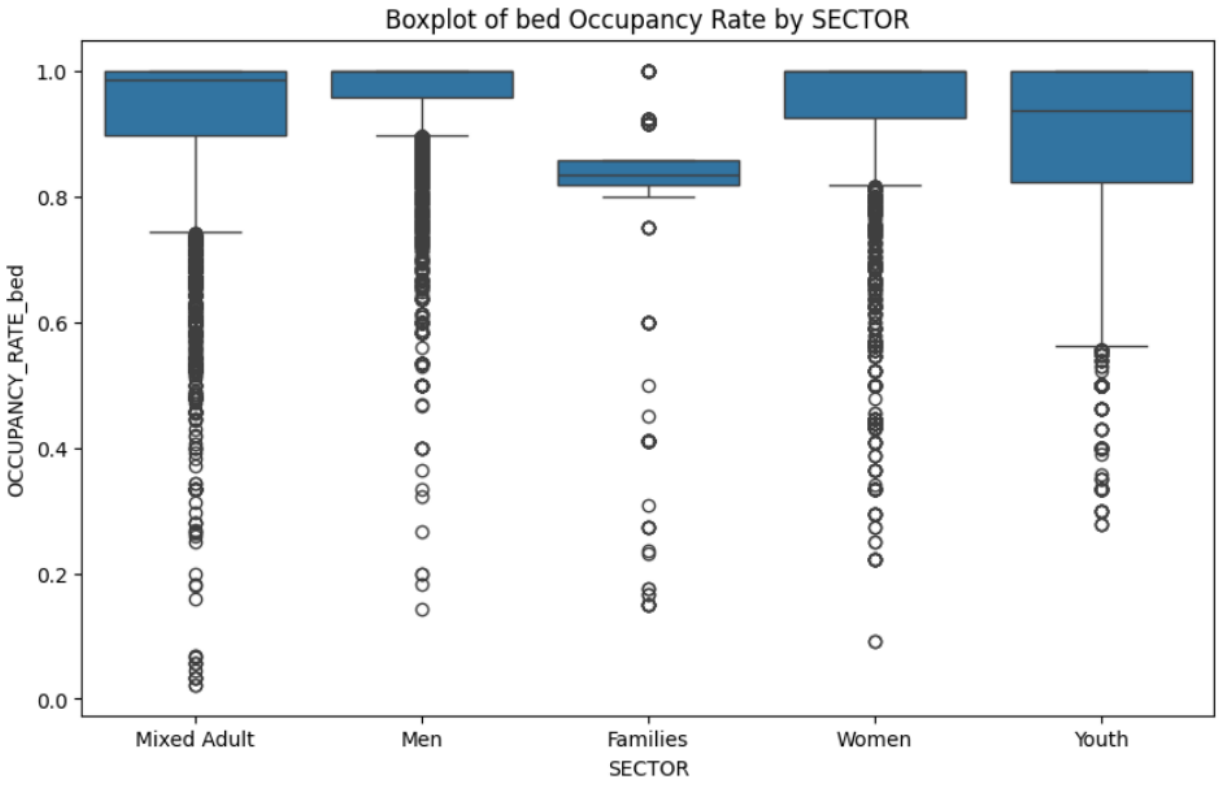
To gain a deeper understanding of homelessness, we use independent samples t-test methods and boxplots on shelter program occupancy rates.

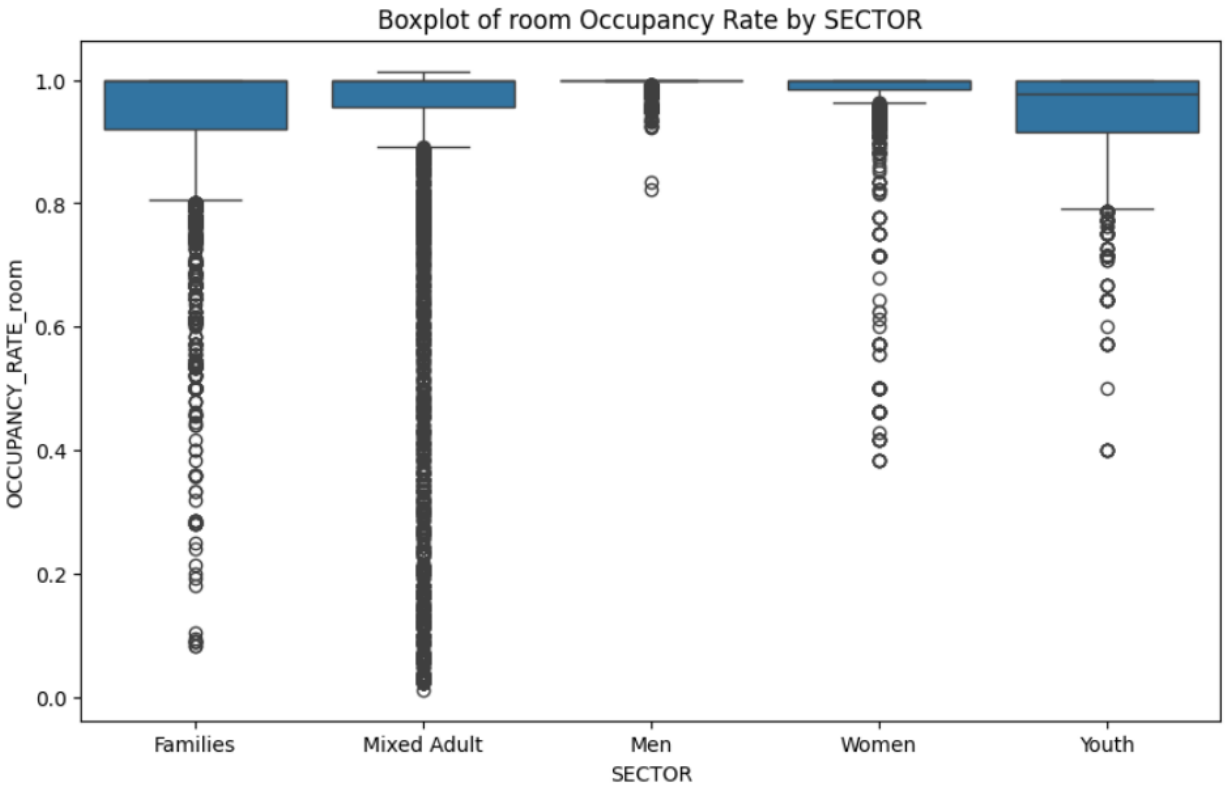
From the table, we can know that ‘CAPACITY_TYPE’ has two different values, so in this part, we calculate t-test methods and boxplots on room and bed shelter program occupancy separately.

A. In first attempt, we calculate the t-test value based on variable Sector 'Mixed Adult' and 'Men'.

	T-statistic	P-value
Bed OCCUPANCY_RATE	24.765	1.00e-132
Room OCCUPANCY_RATE	9.175	5.33e-20

We also try the boxplot for both shelter program occupancy on Sector



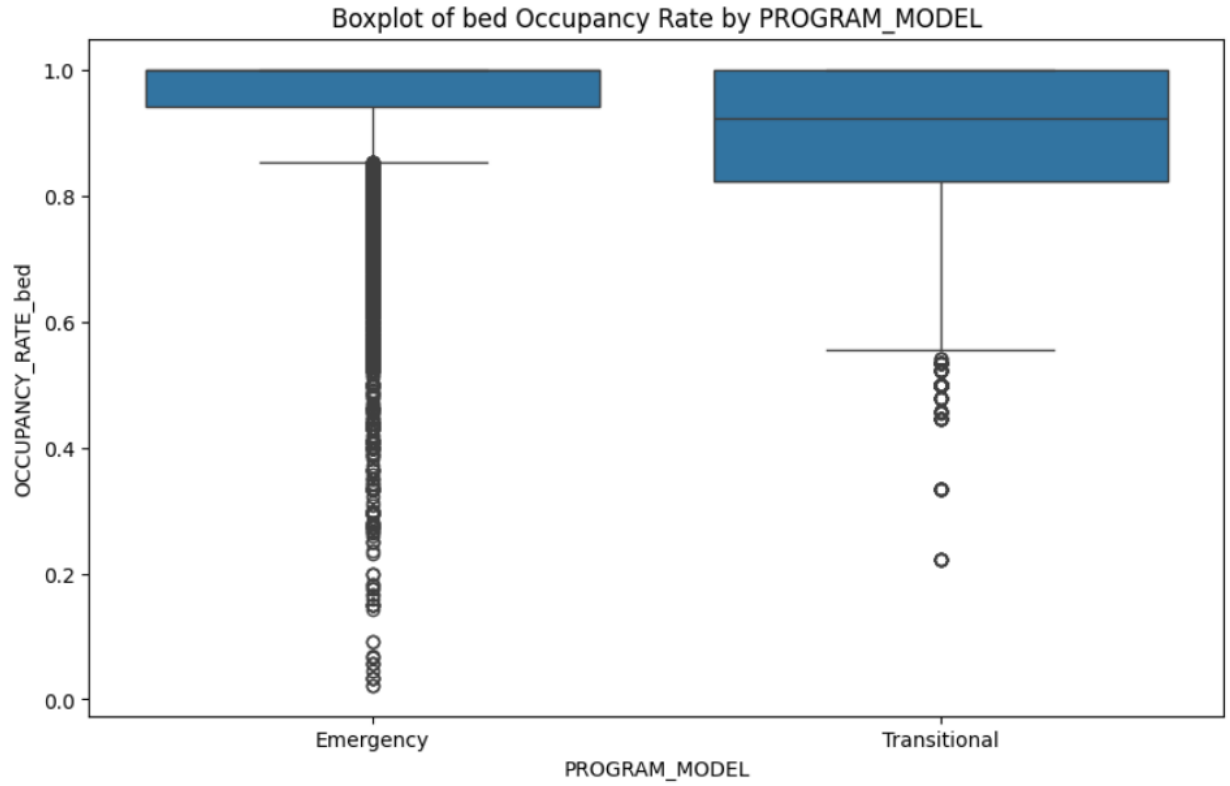


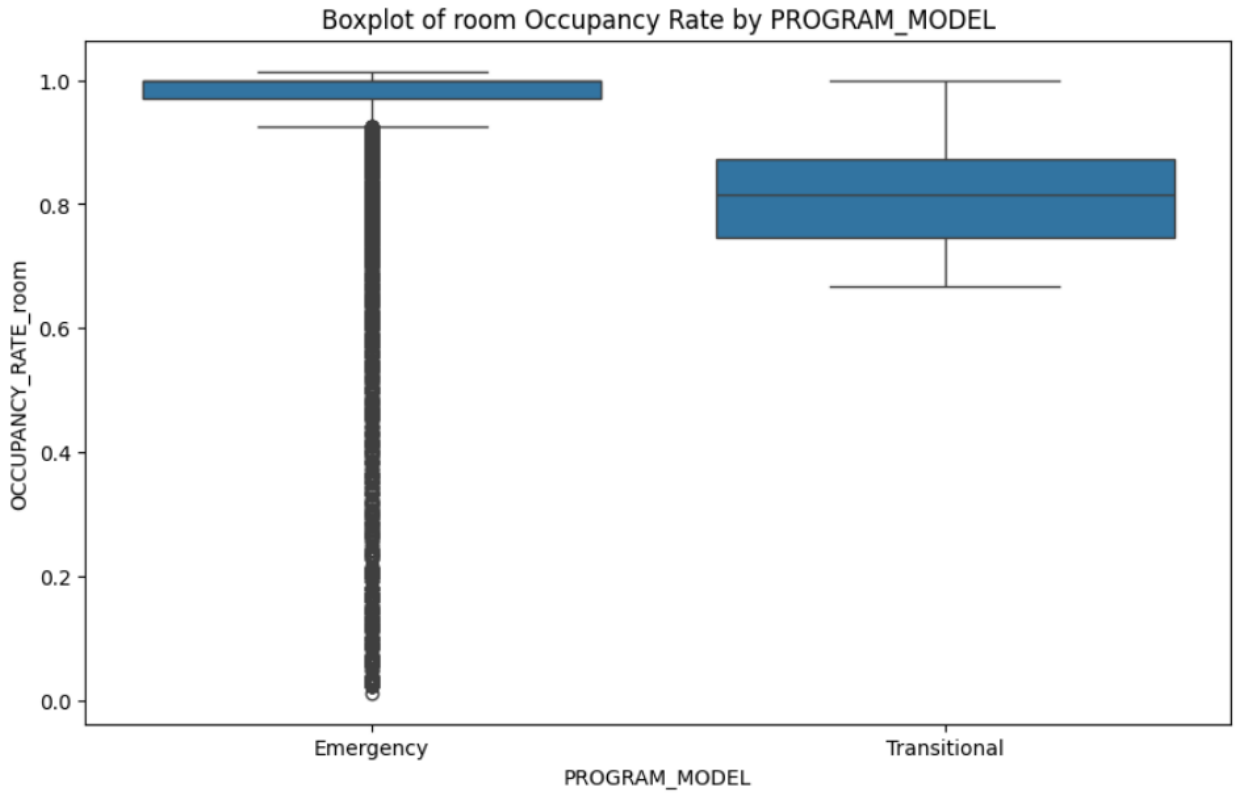
According to the definition of t-test, we can get a conclusion that the difference on both shelter program occupancy is big on Sector 'Mixed Adult' and 'Men'.

B. In the second attempt, we calculate the t-test value based on variable PROGRAM_MODEL ' Emergency' and ‘Transitional’, the result is shown below.

	T-statistic	P-value
Bed OCCUPANCY_RATE	36.774	1.0223e-282
Room OCCUPANCY_RATE	31.711	4.4252e-150

The boxplot is shown below.


Boxplot of bed Occupancy Rate by PROGRAM_MODEL
OCCUPANCY_RATE_bed
1.0
0.8
0.6
0.4
0.2
0.0
Emergency
Transitional
PROGRAM_MODEL

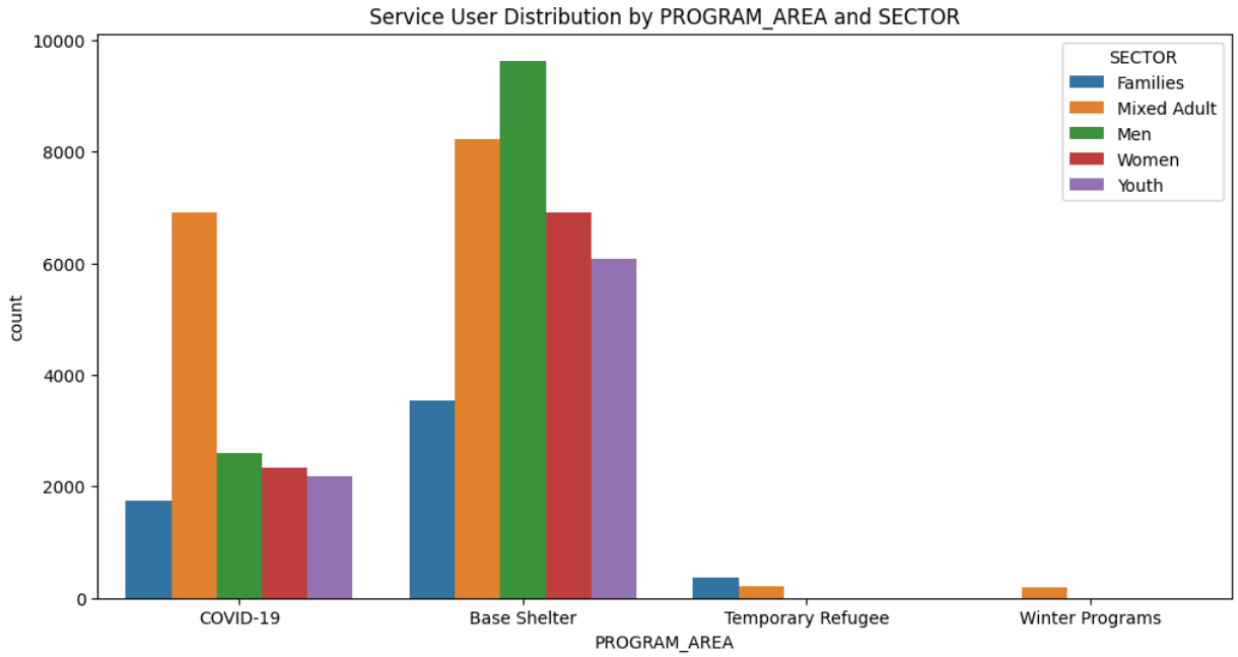

Boxplot of room Occupancy Rate by PROGRAM_MODEL
OCCUPANCY_RATE_room
1.0
0.8
0.6
0.4
0.2
0.0
Emergency
Transitional
PROGRAM_MODEL

In conclusion, the small p-values strongly suggest that there is a statistically significant difference between the groups being compared. The large t-statistics indicate that the observed differences are substantial relative to the variability within each group. So, there is a meaningful distinction between the compared groups on Occupancy Rate.

4. Exploratory data analysis

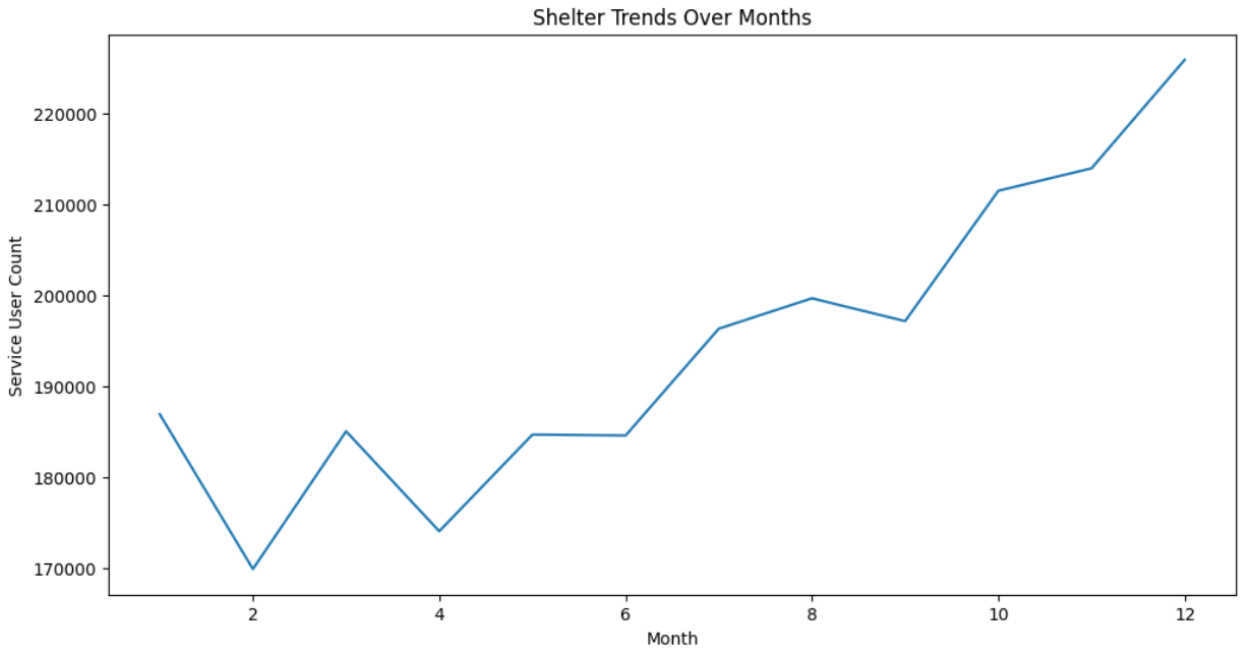
To further explore the dataset, we make the following attempts:

1) First, we explored the Service User Distribution and Program Area Relationship, we use countplot to visualize the distribution of service user groups across different program areas.



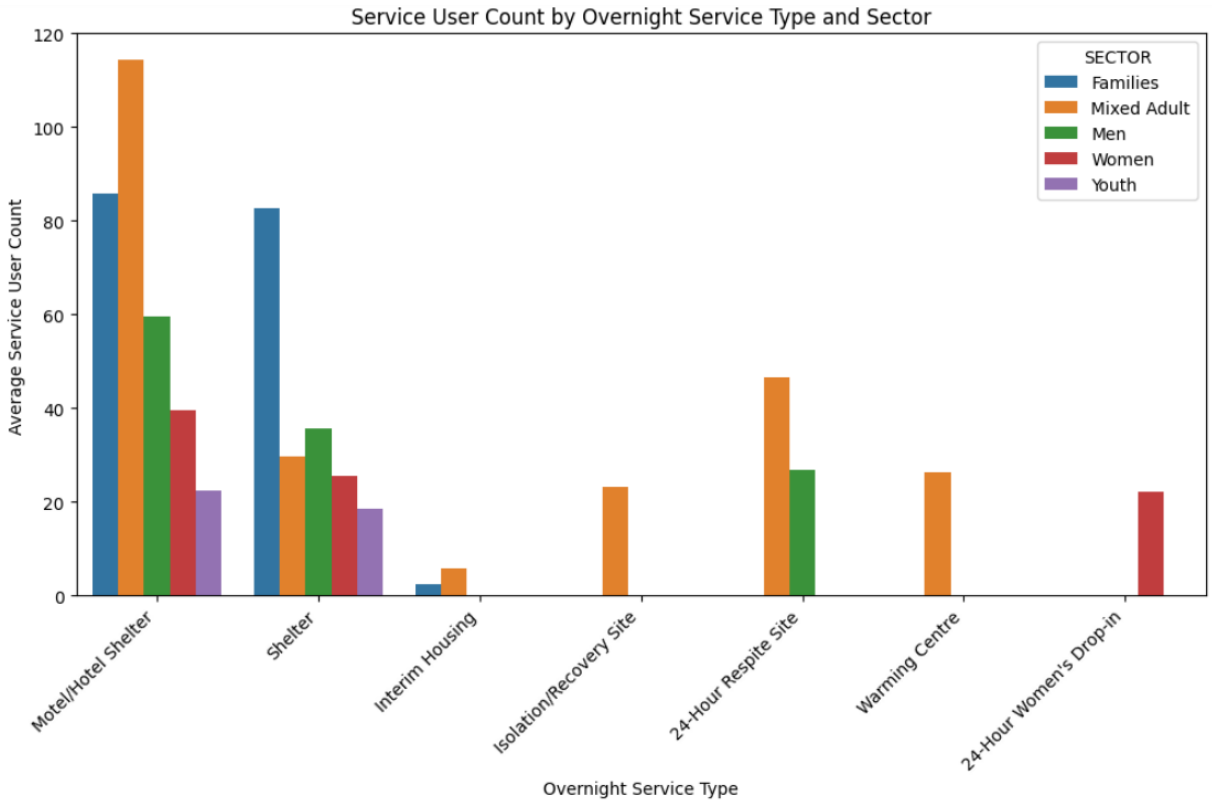
This visualization provides insights into the quantity distribution of various service user groups in different program areas, we can see that Winter Programs and Temporary Refugee has the lowest count of people.

2) Shelter Trends Over Time:



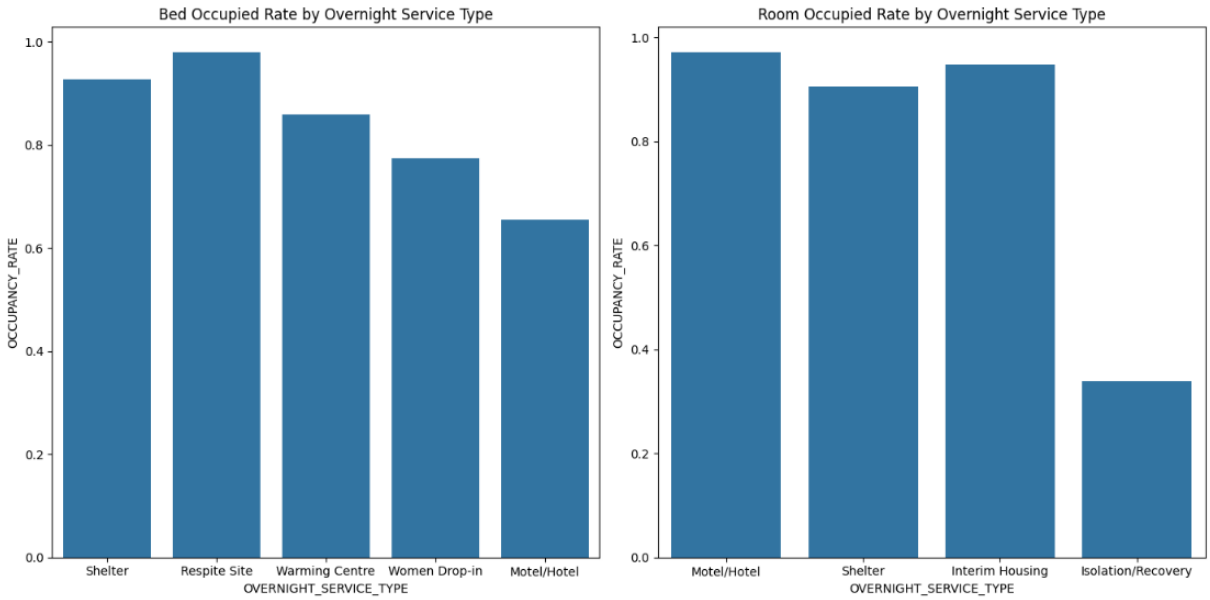
From the chart, we can see that there is a continuous growth in service user count over the months, suggesting an increasing demand for shelter services.

3) Service user counts across different overnight service types and sectors



This bar plot illustrates the distribution of average service user counts across different overnight service types and sectors. From this plot, we can see that most people will choose Motel/Hotel Shelter and Normal Shelter.

4) Bed/Room Occupied Rate by Overnight Service Type



From the data, we know we have 7 types of OVERNIGHT_SERVICE_TYPE, 'Motel/Hotel Shelter', 'Shelter', 'Interim Housing', 'Isolation/Recovery Site', '24-Hour Respite Site', 'Warming Centre', "24-Hour Women's Drop-in".

From the plot, we can find that 'Motel/Hotel Shelter', 'Shelter' provide both beds and rooms. 'Interim Housing', 'Isolation/Recovery Site' only provides rooms, '24-Hour Respite Site', 'Warming Centre', "24-Hour Women's Drop-in" only provide beds. 'Isolation/Recovery Site' has the lowest Occupied Rate. For bed occupied rate, 'Motel/Hotel Shelter' has the lowest, which means people do not want to choose motel or hotel as a shelter.

5. Further analysis

a. Examine the relationships between numerical variables and identify strong correlations, which may suggest variables that are highly interrelated. Visualize the correlation matrix using a heatmap.
b. Explore time-series patterns and trends in shelter occupancy rates.

c. Assess the importance of different features in predicting occupancy rates.