Problem #2: First is to explore the utilization of shelter resources based on two capacity types, using “Bed Based Capacity” as well as “Room Based Capacity”. To do this, I calculated occupancy rates for the dataset containing information about shelter programs, dividing the number of occupied beds by the actual bed capacity for “Bed Based capacity”. Similarly, diving the number of occupied rooms by the actual room capacity to find out “Room Based capacity”. Then to provide a single measure of occupancy rate by combining the above two capacity types, and check if the “Room occupancy rate” is not null and will take this value, otherwise fall back to “Bed occupancy rate”.

Later, to determine if there are statistically significant differences in occupancy rates across different categories defined by categorical variables such as capacity type, program model etc. in the shelter dataset. The process is to iterate over list of categorical variables (ex. Capacity type, program model, program area etc). And for each categorical variable, the goal is to identify the two most frequent categories. Then to separate the occupancy rates for each pair of categories. The last process is to perform t-test to see if statistically significant difference between each two sets of occupancy rates.

The code output is showing the name of the categorical variable being tested, the names of the top two categories, and the t-statistic and p-value for each comparison to determine statistical significance.

From output:
Capacity_Type: the t-statistic which comparing between Bed based capacity and Room based capacity is -4.4988, the negative t-statistic indicates that the mean occupancy rate for Bed based capacity is lower than that for “Room based capacity”.

The p-value (~0.000007) is significantly less than 0.05, means a statistically significant difference between the two capacity types, **so capacity type does have effect on the occupancy levels.**

Program_Model: is to compare the between the emergency and Transitional categories. T-statistic is very large (40.9811), this high positive value indicates the mean occupancy rate for “emergency” is significantly higher than “Transitional”. Since two have significant different occupancy rates, I can assume the **urgency of the program has influence** on how occupied they are. P-value which is 0, means a significant difference between those two groups.

Program_Area: comparing Base shelter and overnight Services System vs. cOVID-19 Response, this showed t-statistic -1.5867 & p-value is 0.113. the negative t-statistic shows the mean occupancy rate for “Base shelter overnight services system” might lower than “COVID-10 Response”. However, the p-value is larger than 0.05 indicates no significant difference in occupancy rates between the two areas. Therefore, **we cannot confidently tell** whether the occupancy rates between the two program areas are different or not.

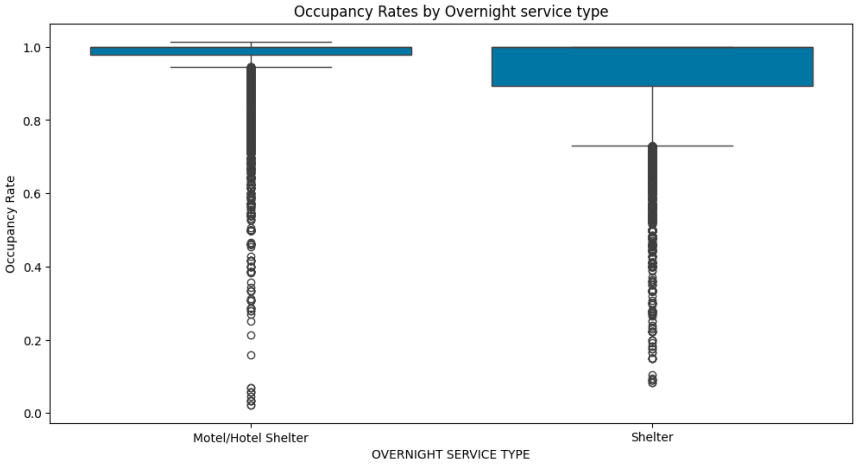
SECTOR: Mixed Adult vs. Men has t-statistic -34.12 and p-value nearly nil. It shows a statistically significant difference in occupancy rates between “mixed adult” and “Men”, with a lower

occupancy rate within “Mixed adult” s compared to “Men” sector. We can infer that the demographic factors does impact occupancy.

Overnight_Service_Type: Shelter Vs. Motel/Hotel Shelter has t-statistic -43.0280 and p-value of Nil. As p-value is less than 0.05, means a significant difference between the two categories, and the mean occupancy rate for Shelter is lower than the Model/Hotel Shelter. From this, we learned the **service environment have influence for shelter occupancy**.

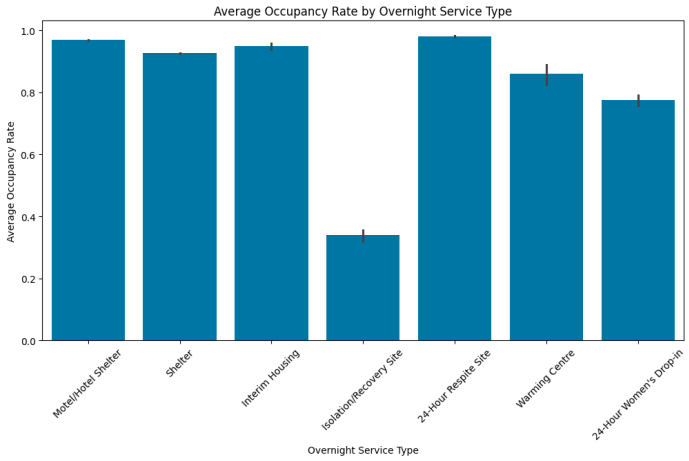
From above code output, we learned different type characteristics affect shelter occupancy, and there is significant difference in occupancy rates for most of the categorial variables (for output with p-value is less than 0.05 and have significant difference).

Problem #3: Then I would like to explore further with some visualization plots by comparing models on selected category.
1)Data analysis: the **Boxplot** for Overnight_Serivce_Type



It shows the occupancy rate for model/hotel shelter and Shelter. We can see the median occupancy rate (from the line inside the boxes) for Motel/Hotel Shelter is higher than Shelter, therefore Model/Hotel shelter has higher central tendency of occupancy rates. The IQR(interquartile range), the range between the first quartile (25% overall range, bottom of blue box) and the third quartile (75% range, top of the blue box), and Motel/Hotel shelter IQR is narrow than the shelter, as less variable changes, and most data are clustered to the range that close to rate 1.0, means there is less variability in occupancy rates in Motel/hotel shelter than Shelter. However, the Shelter has a wider IQR, it means data are having greater dispersion. And the outliers are more prevalent and spread out.

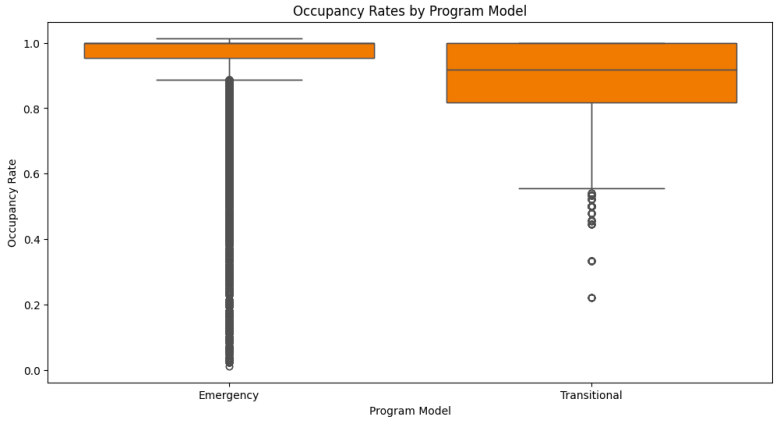
We can not observe standard deviation (sd) in a boxplot, but usually a wider IQR could be a sign of a higher sd. We also cannot tell the mean from boxplot, but usually close to the median in symmetric distribution. From boxplot, we can tell that Model/Hotel Shelters tend to have more consistent occupancy rates, while traditional Shelters have a wider range of occupancy rates and tends to have more outliners.



2)Bar chart

Per above chart showa the Motel/Hotel Shelter, Shelter, Interim Housing, and Warming Centre have relatively high average occupancy rates, with above 80% of occupancy rate. Isolation/Recovery Site and 24-Hour Women's Drop-in have relatively lower average occupancy rates, with Isolation/Recovery Site has the lowest occupancy rate.

3)Boxplot: Emergency vs Transitional

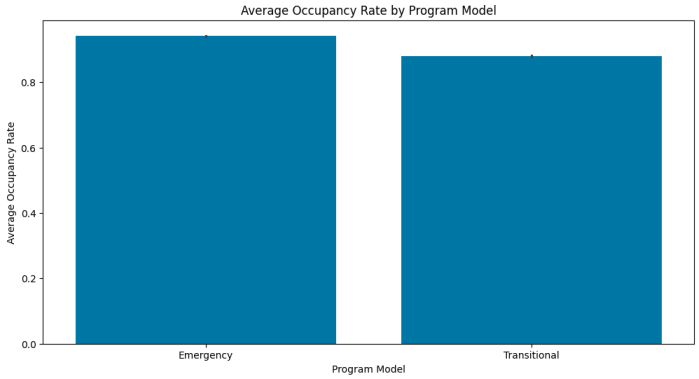


Median is the 50% line inside of box. For emergency, the median is close to the top of the box, which is above 0.9. For transitional, the median is slightly higher than 0.8 but should be lower than emergency. The emergency model has a smaller IQR, which is between Q1 and Q3 (25% to 75% range for orange box), as the box for emergency is very narrow, indicating low variability among the central 50% of occupancy rate. The data are clustered near the top of box for emergency model, and no outliners, it means all the data points fall within the range. However, the transitional model has several outliers on the lower end, therefore some occupancy rates are much lower than the typical rates for transitional group.

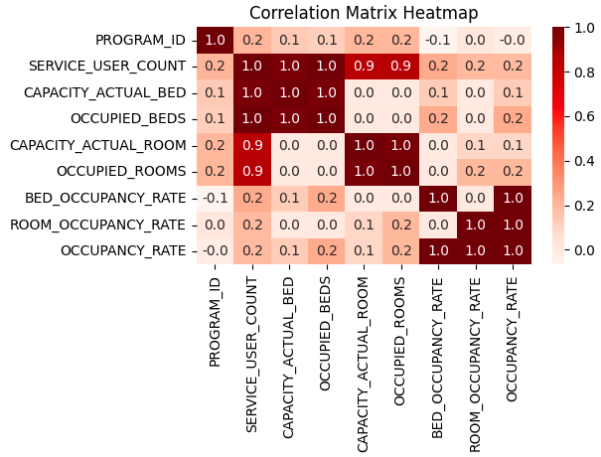
4) **Bar chart**: emergency vs transitional

The average (mean) occupancy rate are similar for emergency and transitional model, both above 0.8 and emergency has slightly higher occupancy rate. It also shows low variability within each group.

From above Boxplot and Bar chart, we can tell the while the mean rates are similar, the spread and variability of the rates between two model are different, traditional model having a wider spread and lower outlier. The emergency model has a more consistent occupancy rate which is near capacity, but transitional model has more variability and sometimes lower rates.



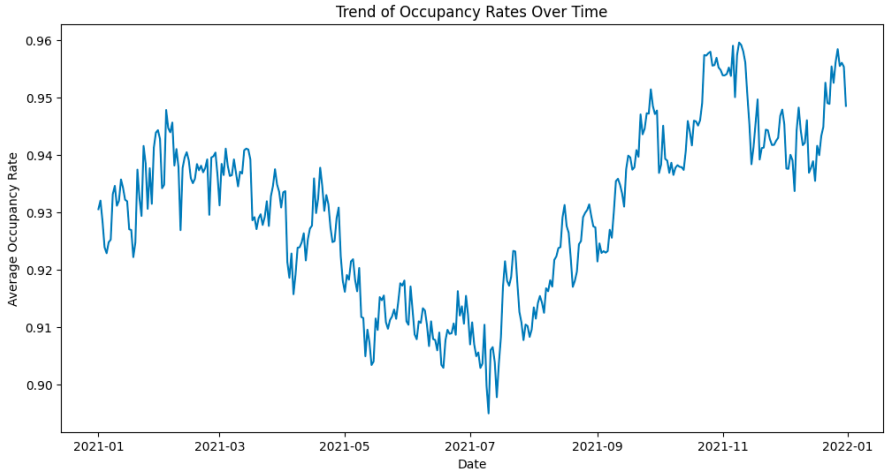
5)**heatmap correlations**



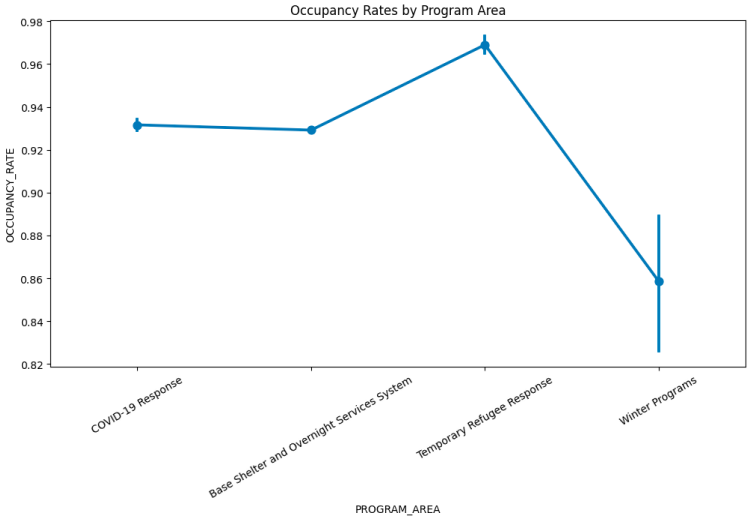
Per above 5) we can tell there are perfect correlation (100%) between Service_user_Count and both Capacity_actual_bed and Occupied_beds. This indicates that Service_user_count might direct reflect actual bed capacities or occpied beds. Also, Bed_occupancy_rate and Room_occupancy_rate have perfect correlation of 1 with occupancy_rate, this shows that occupancy rate might composite of two types of occupancy rates, and they are not independent metrics of each other. Further, there is weak correleation with Program_ID with all other metrics, which does make sense as Program_ID is likely a variable that serves as an identifier and does not have quantitive relations with all other feature variable. Lastly, there are no strong correlations indicating potential issues, as none of the feature variable shows a strong negative or correlations with others, so more data value might needed to explore its complex relationships, or there is potential issues in capacity management.

6)**Line graph** of avergae occupancy rates from Jan 21 to Feb 22
The occupancy rates shows increasing trend overall. So this is a growing demand for the sheleters and a reduction in the total number of available beds/rooms, leading to higher occupancy rates. Also, occupancy rates are lower during summer time (july) but are higher in

winter time, due to a inreased shelter demand. And an increase during end of the year indicates further weather changes have influence on the occupacy rate.



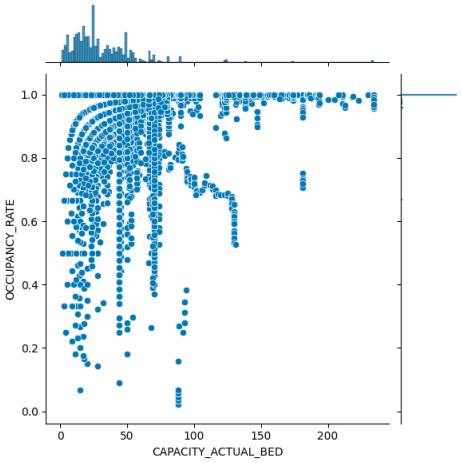
7) **line chart**



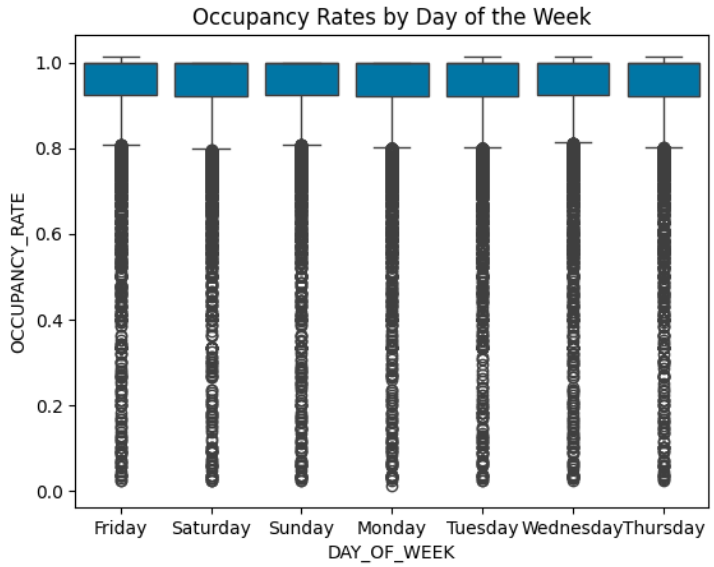
The occupancy rate are ranging from 0.82 to 0.98. The highest occupancy rate is temporary refugee response, at around 0.98, indicating high demand and capacity for those services. Covid 19 and base shelter and oversight service system have similar occupancy rate at nearly 0.92. shows hight utilization of the available capacity. However, winter program shows a decrease and this could due to reduced demand or increased capacity during winter (as discussed above in 6) line graph) holiday time as the rate tends to change according weather.

8) **scatter plot**

The actual bed capacity is ranging from 0 to 200, we can infer many facilities (especially those with smaller bed counts, which is between 0 and 100, are at near full occupancy in terms of capacity utlization. And larger facilities (100 to 200) has wider ranges of occupancy rate, therefore there is more challenging to maintain a high occupancy rate as size increase. Overall lower bed capacities have varying levels of operational efficiency and different level of demand.



9)**Boxplot**: days vs occupancy rate



The medians are very close to the top of box, therefore there is a high occupancy rates across all days, around over 80% of occupancy rate. The IQR (first quartile 25% to third quartile 75% percentile) are similar everyday, indicating consistent variablity in occupancy rates. As for the outliers, which is at the lower end and shows having unusual low occupancy rate, every day has outliers and this shows variability of each day. There is a high degree of consistency in median occupancy rate everyday, less significant deviation for particular day from the others. However having a significant variability in occupancy rates (per outliers at lower end), therefore we can tell that although high occupancy rate during the week, there could be certain days where occupancy is particularly low, which planner would need for better resouce allocation.