INF2178-A1

Yixin Chang
(1005991651)

Research Questions:

1. Which program model, Emergency or Transitional, is more popular, specifically in terms of the number of service users and the occupancy rate?
2. Which capacity type, Bed Based Capacity or Room Based Capacity, is more popular, specifically in terms of the number of service users and the occupancy rate?

Reason why I choose these questions: Because of a significant increase in the city's homeless population and lack of shelter spaces, the City of Toronto's shelter support system may consider providing more shelter spaces. These questions can help the system to decide which type of shelters should be increased based on the demand.

Dataset:

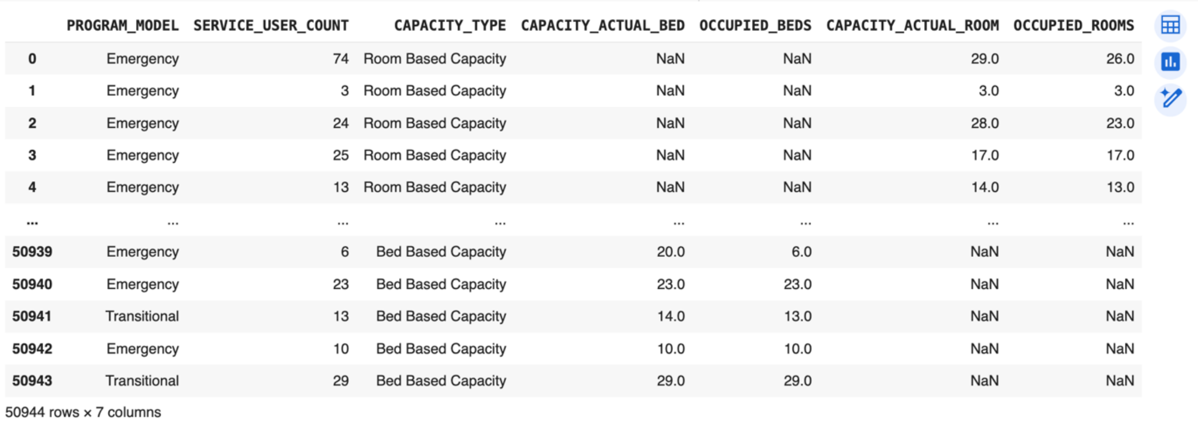
INF2178_A1_data.xlsx

Process, results and analysis:

Data Cleaning:

After loading the dataset in the python, I perform data cleaning to create a data frame for easy analysis. Therefore, I drop the unnecessary columns and check if there are some missing values that I need to deal with.

Table 1

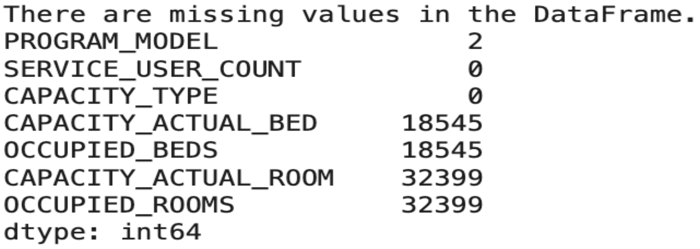


	PROGRAM_MODEL	SERVICE_USER_COUNT	CAPACITY_TYPE	CAPACITY_ACTUAL_BED	OCCUPIED_BEDS	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS
0	Emergency	74	Room Based Capacity	NaN	NaN	29.0	26.0
1	Emergency	3	Room Based Capacity	NaN	NaN	3.0	3.0
2	Emergency	24	Room Based Capacity	NaN	NaN	28.0	23.0
3	Emergency	25	Room Based Capacity	NaN	NaN	17.0	17.0
4	Emergency	13	Room Based Capacity	NaN	NaN	14.0	13.0
...	...	...	...	...	...	...	...
50939	Emergency	6	Bed Based Capacity	20.0	6.0	NaN	NaN
50940	Emergency	23	Bed Based Capacity	23.0	23.0	NaN	NaN
50941	Transitional	13	Bed Based Capacity	14.0	13.0	NaN	NaN
50942	Emergency	10	Bed Based Capacity	10.0	10.0	NaN	NaN
50943	Transitional	29	Bed Based Capacity	29.0	29.0	NaN	NaN

50944 rows × 7 columns

Table 1 shows the data frame that I will use for the analysis. After I remove the columns that I will not use, 50944 rows and 7 columns remain.

Results for checking missing values:



```
There are missing values in the DataFrame.
PROGRAM_MODEL               2
SERVICE_USER_COUNT          0
CAPACITY_TYPE               0
CAPACITY_ACTUAL_BED     18545
OCCUPIED_BEDS           18545
CAPACITY_ACTUAL_ROOM    32399
OCCUPIED_ROOMS          32399
dtype: int64
```

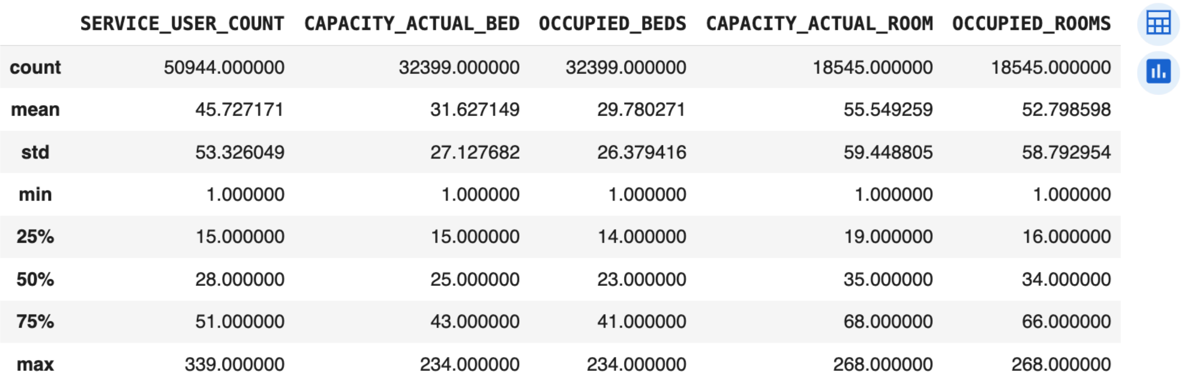
From the above results, we can find missing values in some specific columns. Last four columns have missing values because of different capacity types. If the capacity type is Bed

Based Capacity, CAPACITY_ACTUAL_BED and OCCUPIED_BEDS have values but CAPACITY_ACTUAL_ROOM and OCCUPIED_ROOMS are “NaN”, and vice versa. This format makes sense and has no effect on computations. When conducting analysis, I will also consider the CAPACITY_TYPE column. Although there are 2 missing values in the PROGRAM_MODEL column, SERVICE_USER_COUNT and CAPACITY_TYPE do not have missing value and contain values that I need. Therefore, in this part, I keep Table 1 unchanged and use it for further steps.

EDA:

1. Overall descriptive statistics:

Firstly, I generate descriptive statistics for numerical columns in the current data frame.

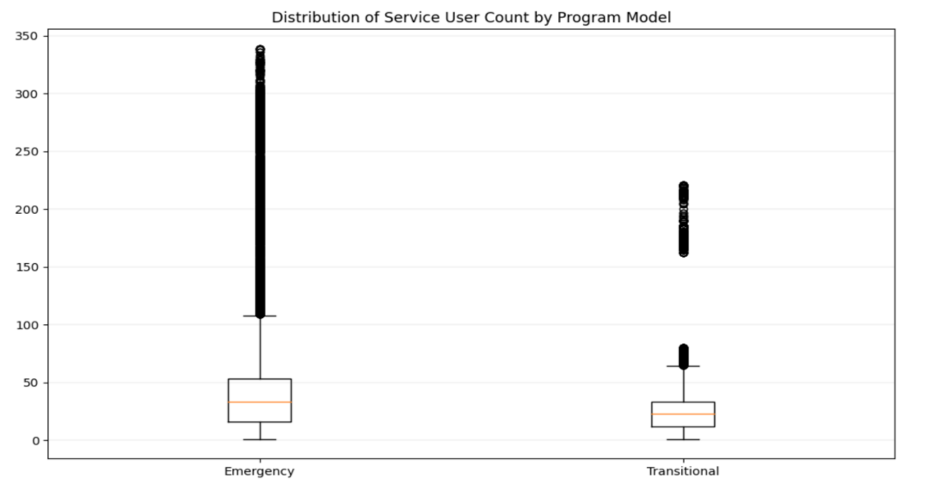


	SERVICE_USER_COUNT	CAPACITY_ACTUAL_BED	OCCUPIED_BEDS	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS
count	50944.000000	32399.000000	32399.000000	18545.000000	18545.000000
mean	45.727171	31.627149	29.780271	55.549259	52.798598
std	53.326049	27.127682	26.379416	59.448805	58.792954
min	1.000000	1.000000	1.000000	1.000000	1.000000
25%	15.000000	15.000000	14.000000	19.000000	16.000000
50%	28.000000	25.000000	23.000000	35.000000	34.000000
75%	51.000000	43.000000	41.000000	68.000000	66.000000
max	339.000000	234.000000	234.000000	268.000000	268.000000

The standard deviation is large in each column, which shows significant variability in the data. We can see that Bed Based Capacity is provided more times than Room Based Capacity, whereas the means of the number of beds provided or occupied are lower than those of rooms. In this case, I tentatively guess that there is a preference or greater need for Room Based Capacity than Bed Based Capacity. However, we cannot explore the distribution of service user count based on different program models and different capacity types in the above table. To make more analysis and comparisons, I use boxplots.

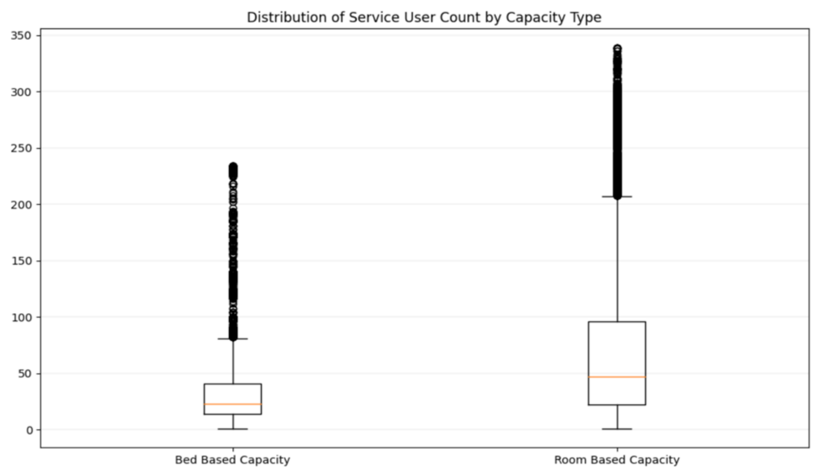
2. Service User Count:

Program Model: Emergency vs Transitional



The boxplot compares the distribution of service user count by program model, Emergency and Transitional. Transitional model has a more compact range of service user counts, indicating less variability compared to Emergency model. Both models have many outliers, but Emergency model has more. Furthermore, Emergency model's median is a little higher than Transitional model's, and its IQR is wider. This means the data of Emergency model is more dispersed. Hence, Emergency model is more flexible and popular for homeless population with different conditions, and it serves more users.

Capacity Type: Bed Based Capacity vs Room Based Capacity

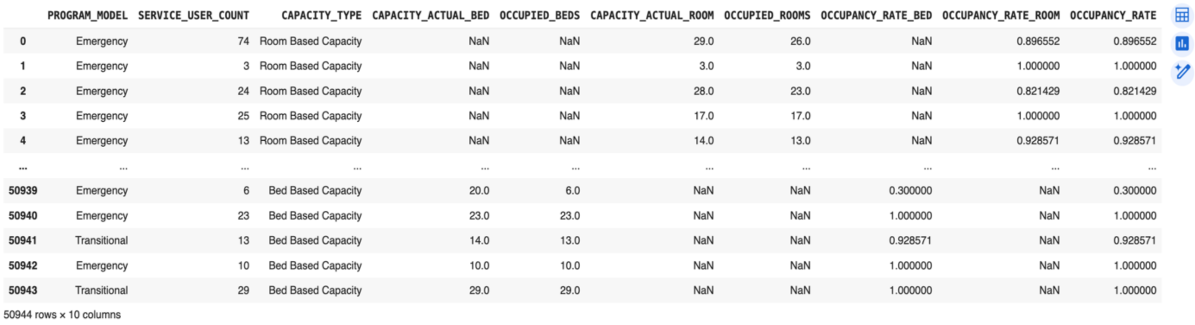


The boxplot comparing the distribution of service user count by capacity type, Bed Based Capacity and Room Based Capacity. Room Based Capacity has a higher median and a larger IQR, which means it can service more individuals. Both types have many outliers. It is clear that Room Based Capacity has a much larger range and higher amount of service user count than Bed Based Capacity's. This fact makes sense because rooms can accommodate many people at a time. Therefore, Room Based Capacity are more in demand.

3. Occupancy Rate:

Since there is no occupancy rate in the original table (Table 1), I calculate and add three occupancy rate columns for further analysis. Moreover, the occupancy rate is a continuous variable that I can use for t-tests. Here is a screenshot of the current data frame:

Table 2



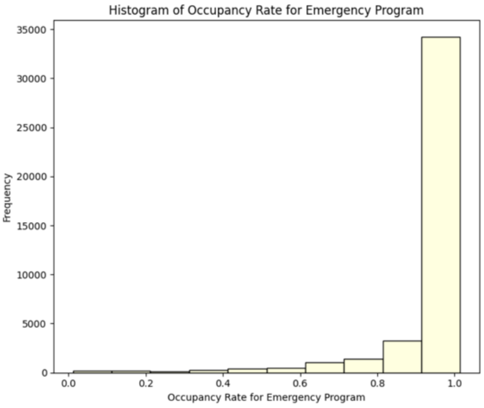
	PROGRAM_MODEL	SERVICE_USER_COUNT	CAPACITY_TYPE	CAPACITY_ACTUAL_BED	OCCUPIED_BEDS	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS	OCCUPANCY_RATE_BED	OCCUPANCY_RATE_ROOM	OCCUPANCY_RATE
0	Emergency	74	Room Based Capacity	NaN	NaN	29.0	26.0	NaN	0.896552	0.896552
1	Emergency	3	Room Based Capacity	NaN	NaN	3.0	3.0	NaN	1.000000	1.000000
2	Emergency	24	Room Based Capacity	NaN	NaN	28.0	23.0	NaN	0.821429	0.821429
3	Emergency	25	Room Based Capacity	NaN	NaN	17.0	17.0	NaN	1.000000	1.000000
4	Emergency	13	Room Based Capacity	NaN	NaN	14.0	13.0	NaN	0.928571	0.928571
...	...	...	...	...	...	...	...	...	...	...
50939	Emergency	6	Bed Based Capacity	20.0	6.0	NaN	NaN	0.300000	NaN	0.300000
50940	Emergency	23	Bed Based Capacity	23.0	23.0	NaN	NaN	1.000000	NaN	1.000000
50941	Transitional	13	Bed Based Capacity	14.0	13.0	NaN	NaN	0.928571	NaN	0.928571
50942	Emergency	10	Bed Based Capacity	10.0	10.0	NaN	NaN	1.000000	NaN	1.000000
50943	Transitional	29	Bed Based Capacity	29.0	29.0	NaN	NaN	1.000000	NaN	1.000000

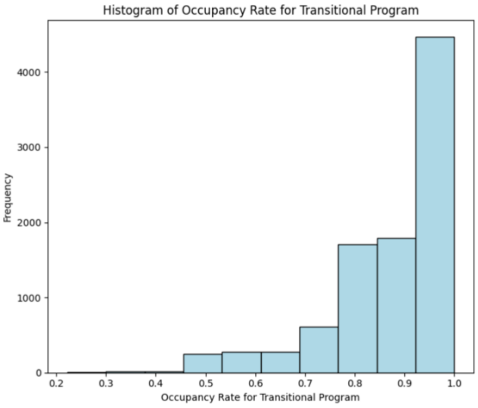
50944 rows × 10 columns

OCCUPANCY_RATE_BED contains occupancy rates for Bed Based Capacity, and OCCUPANCY_RATE_ROOM contains occupancy rates for Room Based Capacity. OCCUPANCY_RATE contains occupancy rates for both types, and this column is helpful when I compare the distribution of occupancy rate between different program models. I use Table 2 for further steps.

Program Model: Emergency vs Transitional

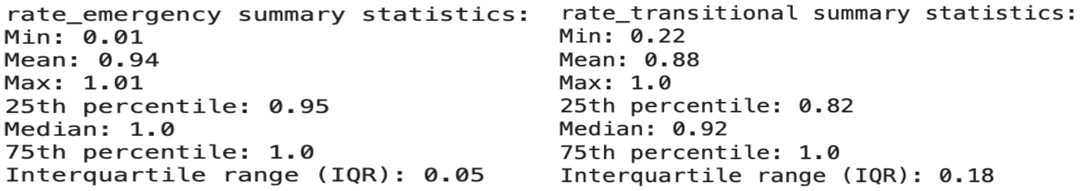
Histograms





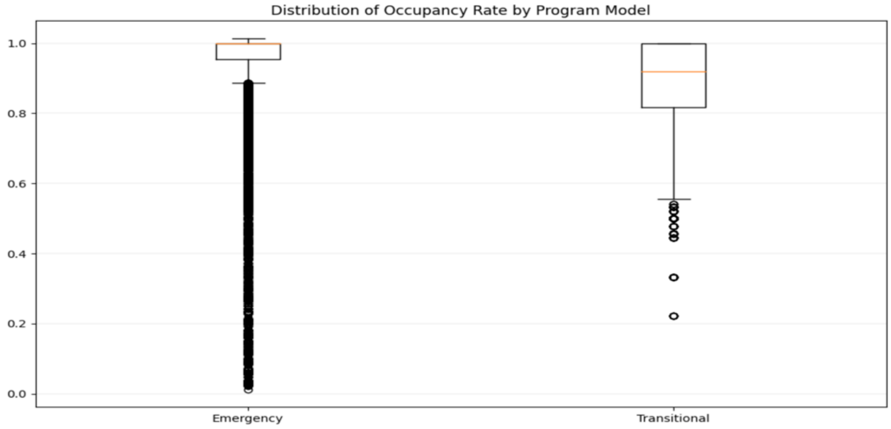
The occupancy rate distributions for Emergency and Transitional programs are shown in above two histograms. For Emergency Program, the occupancy rate is mostly clustered at 1, with approximately 34000 rates falling between 0.9 and 1. The distribution for Transitional Program is more spread out, with the highest frequency at 1 and roughly 4500 rates falling between 0.92 and 1. These indicate that Emergency Program are typically fully occupied and more popular. More Emergency shelters are needed.

Summary Statistics:



```
rate_emergency summary statistics:   rate_transitional summary statistics:
Min: 0.01                            Min: 0.22
Mean: 0.94                           Mean: 0.88
Max: 1.01                            Max: 1.0
25th percentile: 0.95                25th percentile: 0.82
Median: 1.0                          Median: 0.92
75th percentile: 1.0                 75th percentile: 1.0
Interquartile range (IQR): 0.05      Interquartile range (IQR): 0.18
```

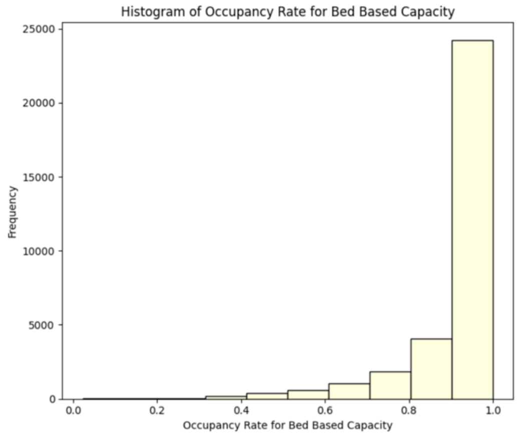
Boxplots:

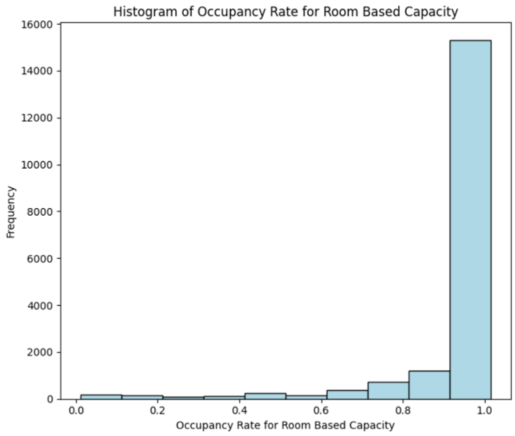


Both summary statistics and boxplots show that the mean and the median of the occupancy rate for Emergency Program are higher than those of Transitional Program. Transitional Program has a much wider IQR. From the boxplots, Emergency Program has much more outliers than Transitional Program’s. The box of Emergency Program is close to 1, which represents full occupancy. We can find the maximum occupancy rate for Emergency Program is 1.01. This is strange because rate is usually not greater than 1. It could be caused by a lack of capacity or incorrect data collecting. Overall, it indicates the need for additional emergency shelters to meet the demand.

Capacity Type: Bed Based Capacity vs Room Based Capacity

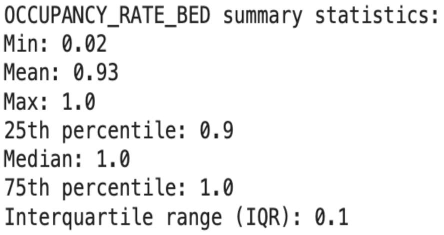
Histograms



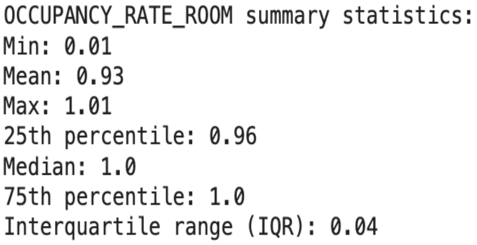


Above two histograms display the occupancy rate distributions for Bed Based Capacity and Room Based Capacity. Both histograms show the highest frequency at 1, but Room Based Capacity has a wider range of lower occupancy rates than Bed Based Capacity. Overall, both capacity types are highly utilized, it is hard to determine which one is more popular and needs further analysis.

Summary Statistics:

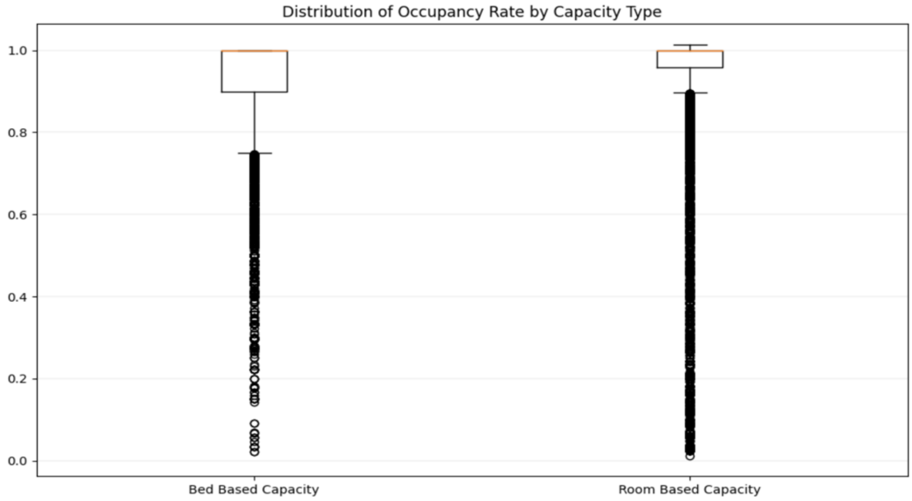


```
OCCUPANCY_RATE_BED summary statistics:
Min: 0.02
Mean: 0.93
Max: 1.0
25th percentile: 0.9
Median: 1.0
75th percentile: 1.0
Interquartile range (IQR): 0.1
```



```
OCCUPANCY_RATE_ROOM summary statistics:
Min: 0.01
Mean: 0.93
Max: 1.01
25th percentile: 0.96
Median: 1.0
75th percentile: 1.0
Interquartile range (IQR): 0.04
```

Boxplots:

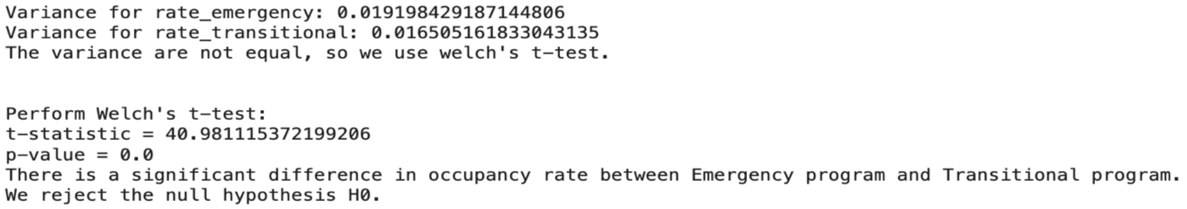


From both summary statistics and boxplots, we find means and medians of the occupancy rate for both capacity types are the same and close to 1. The IQR of Room Based Capacity is very narrow, implying that room occupancy rates are nearly always at 100%. Both types have many outliers, but Room Based Capacity seems to have more. We may conclude that both capacity types are in high demand, particularly Room Based Capacity, which has a narrow IQR.

t-tests:

To further analyze and examine the existence of the difference in occupancy rate distribution depending on different program models and capacity types, I use two t-tests. The occupancy rate is continuous, and each rate is independent. Assumptions are satisfied in both tests.

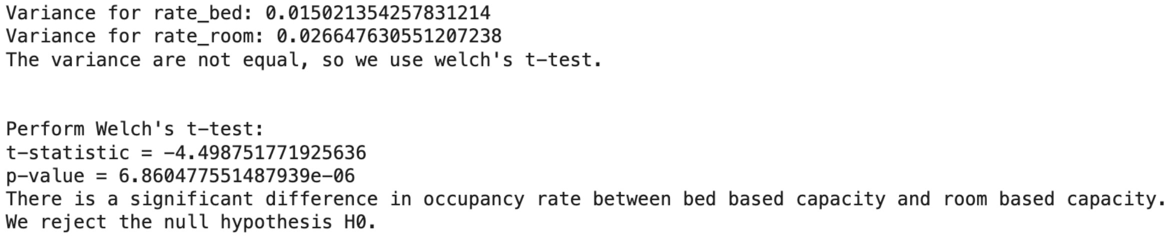
Program Model: Emergency vs Transitional
H0: There is no significant difference in occupancy rate between Emergency program and Transitional program.
Results:



```
Variance for rate_emergency: 0.019198429187144806
Variance for rate_transitional: 0.016505161833043135
The variance are not equal, so we use welch's t-test.


Perform Welch's t-test:
t-statistic = 40.981115372199206
p-value = 0.0
There is a significant difference in occupancy rate between Emergency program and Transitional program.
We reject the null hypothesis H0.
```

Capacity Type: Bed Based Capacity vs Room Based Capacity
H0: There is no significant difference in occupancy rate between Bed Based Capacity and Room Based Capacity.
Results:



```
Variance for rate_bed: 0.015021354257831214
Variance for rate_room: 0.026647630551207238
The variance are not equal, so we use welch's t-test.


Perform Welch's t-test:
t-statistic = -4.498751771925636
p-value = 6.860477551487939e-06
There is a significant difference in occupancy rate between bed based capacity and room based capacity.
We reject the null hypothesis H0.
```

Conclusion

For program model, Emergency Program is more popular due to its large service user counts and high occupancy rate compared to Transitional Program. Therefore, we should consider increasing emergency shelters as it has more demands.
For capacity type, both Bed Based Capacity and Room Based Capacity are popular. However, shelter system should still increase Room Based Capacity as it accommodates more people.