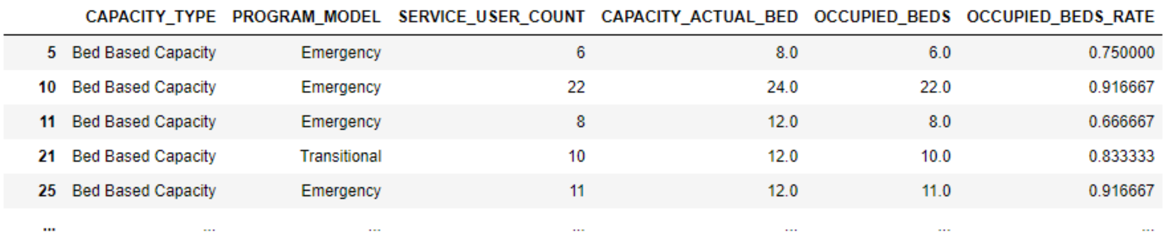
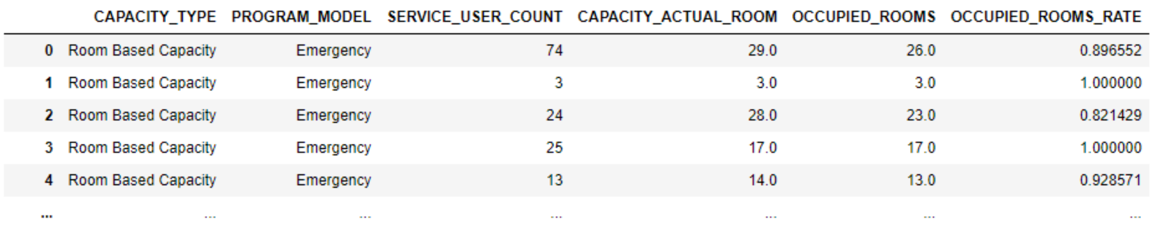
INF2178 Assignment 1 – Yuyang Liu

For the purpose of this project, we will be examining the dataset called “INF2178_A1_data”. This project is intended to conduct statistical t-test on various variables to see the difference and perform exploratory data analysis (EDA) on the variables by building some data visualizations.

To begin with, I grouped the original dataset by the variable “CAPACITY_TYPE” into two separate sub-datasets, calculate relevant occupancy rate and add it as a new column.



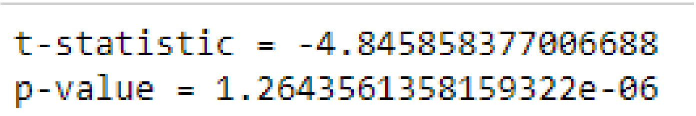
	CAPACITY_TYPE	PROGRAM_MODEL	SERVICE_USER_COUNT	CAPACITY_ACTUAL_BED	OCCUPIED_BEDS	OCCUPIED_BEDS_RATE
5	Bed Based Capacity	Emergency	6	8.0	6.0	0.750000
10	Bed Based Capacity	Emergency	22	24.0	22.0	0.916667
11	Bed Based Capacity	Emergency	8	12.0	8.0	0.666667
21	Bed Based Capacity	Transitional	10	12.0	10.0	0.833333
25	Bed Based Capacity	Emergency	11	12.0	11.0	0.916667
...	...	...	...	...	...	...



	CAPACITY_TYPE	PROGRAM_MODEL	SERVICE_USER_COUNT	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS	OCCUPIED_ROOMS_RATE
0	Room Based Capacity	Emergency	74	29.0	26.0	0.896552
1	Room Based Capacity	Emergency	3	3.0	3.0	1.000000
2	Room Based Capacity	Emergency	24	28.0	23.0	0.821429
3	Room Based Capacity	Emergency	25	17.0	17.0	1.000000
4	Room Based Capacity	Emergency	13	14.0	13.0	0.928571
...	...	...	...	...	...	...

Then I removed the null values remaining in the two sub-datasets to finish off the data cleaning process.

To analyze the dataset statistically, first of all, I was interested in finding if there is a real difference between the mean occupancy rates of the bed-based capacity and room-based capacity, so I conducted a two sample t-test and obtained the following results.

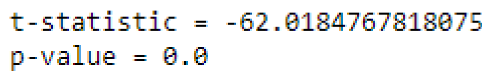


```
t-statistic = -4.845858377006688
p-value = 1.2643561358159322e-06
```

The negative t-statistic indicates that the bed occupancy rate is lower than the room occupancy rate, and an extremely small p-value tells that the result is statistically significant under any alpha level. We have strong evidence to reject the null hypothesis by claiming that there is probably a true difference between the two rates.

I also conducted two-sample t-test on several other interesting topics, including:

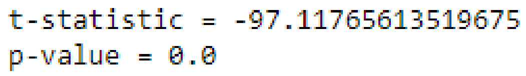
A two-sample t-test on 'CAPACITY_ACTUAL_BED' and 'CAPACITY_ACTUAL_ROOM':



```
t-statistic = -62.0184767818075
p-value = 0.0
```

From the t-test result (Both t-statistic and p-value), we know that there is probably a real difference between the mean actual capacity of the bed type shelter and the mean room type shelter, and the room type has a lot more capacity than the bed type shelter. This makes sense because usually a room can fit more people than a bed does.

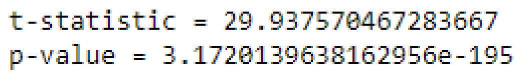
A two-sample t-test on 'SERVICE_USER_COUNT' between different 'CAPACITY_TYPE':



```
t-statistic = -97.11765613519675
p-value = 0.0
```

From the result, we see that the room type shelter can accommodate and serve more people than the bed type shelter on average and the result is statistically significant, this is reasonable, and its interpretation is just like above, the room type shelter is probably larger and it can fit more homeless people.

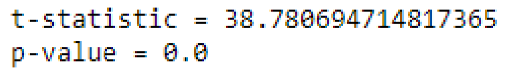
A two-sample t-test on 'SERVICE_USER_COUNT' between different 'PROGRAM_MODEL':



```
t-statistic = 29.937570467283667
p-value = 3.1720139638162956e-195
```

I was also interested in finding out if the number of served homeless is different between emergency and transitional shelter programs. The result has shown that there is indeed a mean difference, the emergency program can serve more people than the transitional program. This aligns with the program definition, as transitional programs require referral, this set threshold to limit the number of users.

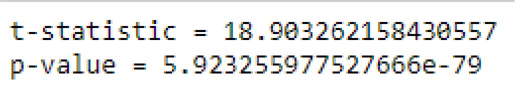
A two-sample t-test on 'OCCUPIED_BEDS_RATE' between different 'PROGRAM_MODEL':



```
t-statistic = 38.780694714817365
p-value = 0.0
```

By focusing on the bed occupancy rate specifically, we can tell there is a true mean difference between the emergency and the transitional programs. This is due to the referral threshold of the transitional programs, this blocks off and limits the number of homeless people that can use the shelter service.

A two-sample t-test on 'OCCUPIED_ROOMS_RATE' between different 'PROGRAM_MODEL':

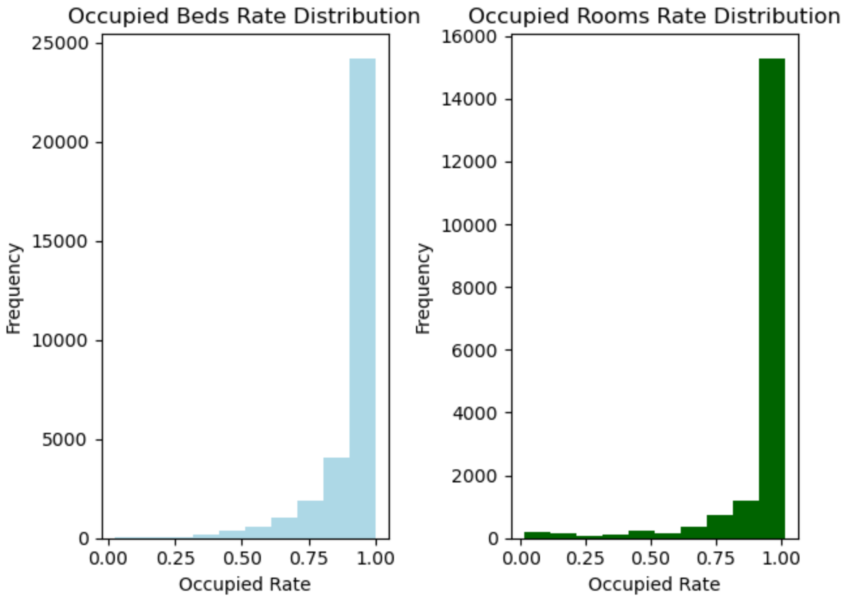


```
t-statistic = 18.903262158430557
p-value = 5.923255977527666e-79
```

By focusing on the room occupancy rate specifically, we can also tell there is a true mean difference between the emergency and the transitional programs. This aligns with the previous result and the interpretation is the same.

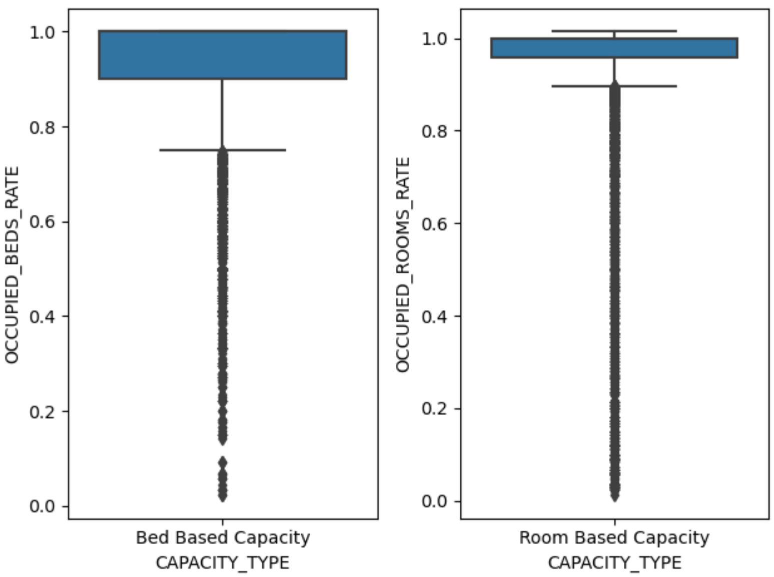
After analyzing the t-tests, I was interested in building some visualizations in an attempt to get to learn more about the dataset. The constructed graphs are shown below:

Histogram for 'OCCUPIED_BEDS_RATE' and 'OCCUPIED_ROOMS_RATE' side by side:



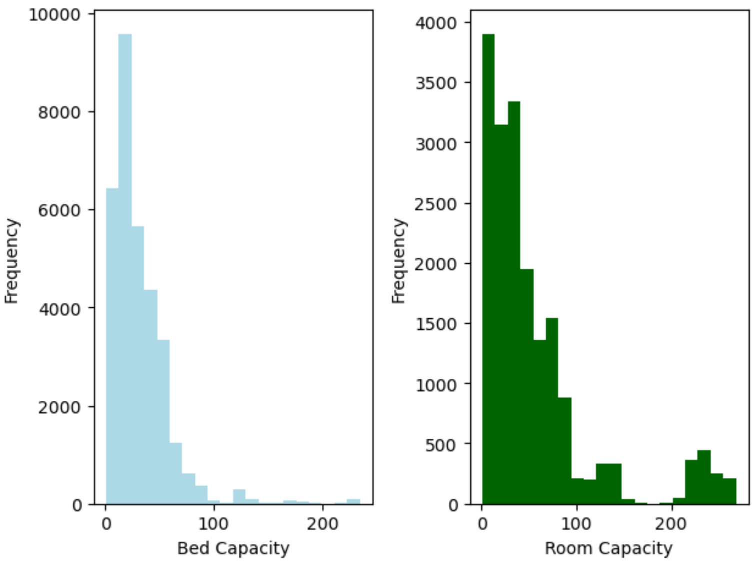
We can observe a fairly similar distribution of the two occupancy rates. Both the bed-mode and the room-mode have a very high utilization rate. However, there are a few exceptions for the room-mode, this may be attributed to that the room-based shelters are for family, whereas the majority of homeless are not in the unit of family, therefore there are some room-based shelters have low occupancy rate. Both histograms skewed to the left, it indicates that majority of the time the shelters are being utilized at high or even full capacity.

Boxplot for 'OCCUPIED_BEDS_RATE' and 'OCCUPIED_ROOMS_RATE':



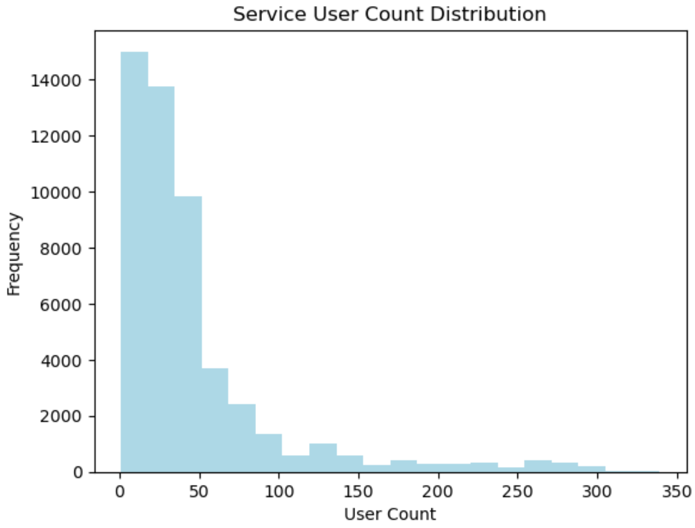
By looking at the boxplot of the two rates based on capacity type, we see that room-based shelters have smaller variance than the bed-based shelters, but it has more outliers that experience lower occupancy rate as well.

Histogram for 'CAPACITY_ACTUAL_BED' and 'CAPACITY_ACTUAL_ROOM':



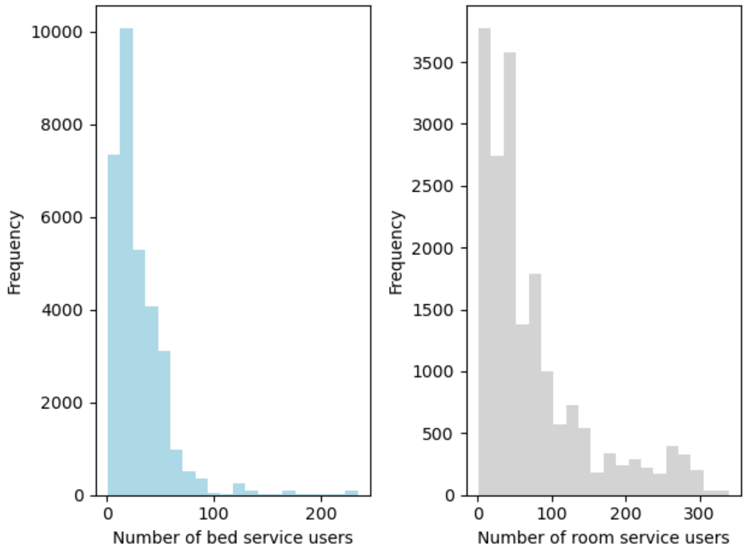
Majority of the shelters have less than 100 people capacity, this applies to both the bed and the room-based shelters. But the room-based shelters have some exceptions that can accommodate more than 200 homeless people as we can see a clear small peak on the right hand side of the room-based capacity histogram.

Histogram for 'SERVICE_USER_COUNT':



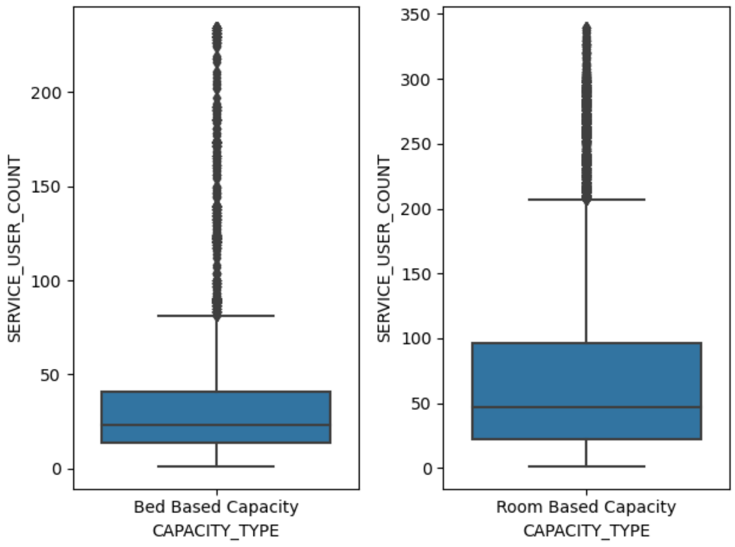
We can see a similar trend as the previous graph. Most of the shelters have served less than 100 people overnight due to capacity limitations. Some exceptions have accommodated more than 200 homeless, and majority of these shelters are room-based.

Histogram for 'SERVICE_USER_COUNT' between different 'CAPACITY_TYPE':



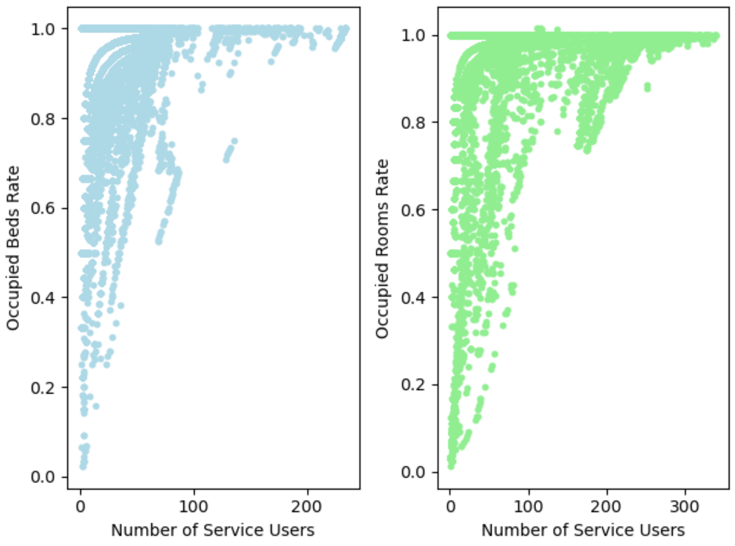
Similar implications as above graphs, some room-based shelters are able to accommodate more than 200 people, whereas the bed-based shelters have a clear cluster less than 100 people. Both histograms are right skewed because most shelters are small in size and have limited capacity.

Boxplot for 'SERVICE_USER_COUNT' between different 'CAPACITY_TYPE':



Boxplots have shown us a clearer picture of the capacity for two different shelter modes. We can see a larger variance for room-based shelters which extends to have higher capacity up to almost 350 people. Whereas more than three fourths of the bed-based shelters can only serve less than 50 people by looking at the blue box on the graph.

Scatterplots for 'SERVICE_USER_COUNT' and the occupancy rates:



The relationship between the number of people served and the occupancy rate is positive for both scatterplots. This makes intuitive sense as more people being served overnight, the occupied rate is higher in the shelters.