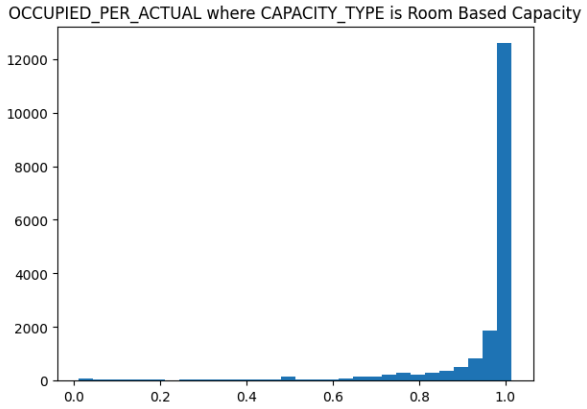
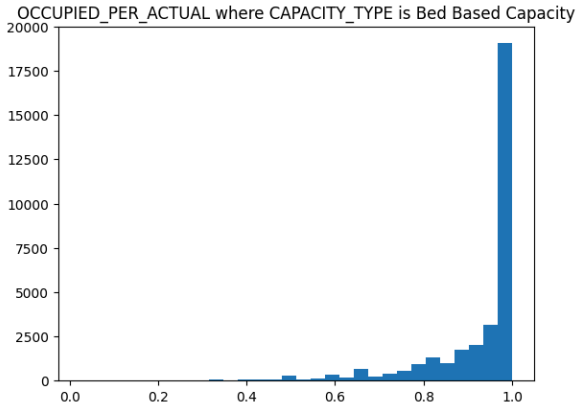
INF2178
Chi-shiun Yang
1009916897
Jan 29, 2024
Assignment 1

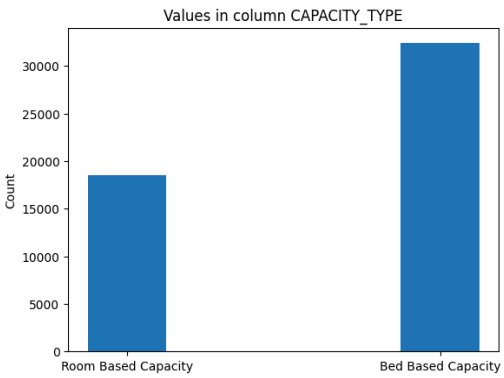
After loading the data, I realized that some column are not very useful for analysis, so only the columns PROGRAM_MODEL, SERVICE_USER_COUNT, CAPACITY_TYPE, CAPACITY_ACTUAL_BED, OCCUPIED_BEDS, CAPACITY_ACTUAL_ROOM, and OCCUPIED_ROOMS were used in the assignment.

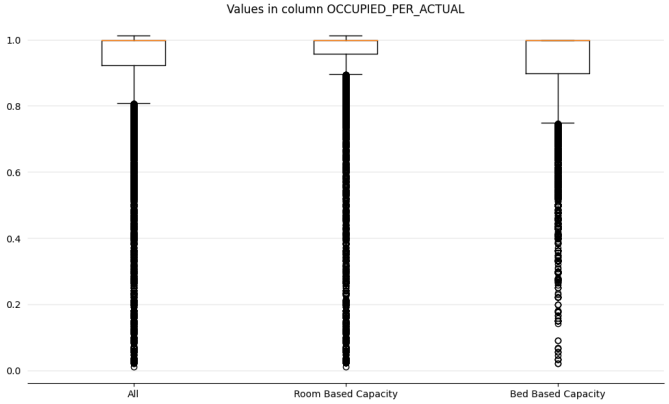
Before the EDAs and t-tests, I created two new rows for the data frame so that less calculation will be done while analyzing. First, the OCCUPIED_PER_ACTUAL column. It used the value in the occupied column divided by the value of the capacity column. And the USER_PER_OCCUPIED column used the service user column divided by the occupied column. In this column, we can see that the number of users per occupied bed is 1, which is predictable and reasonable.

The first part of the analysis is the value of OCCUPIED_PER_ACTUAL based on different CAPACITY_TYPE. There are 2 capacity types, room based and bed based. Here are their histogram of two capacity types, bar plot for their amount, and the box plots along with the population.









From the bar plot, we can see that most shelters are bed based. And from the histograms, we can see that both types of shelters are skewed to the left and most shelters have high capacity. For the box plot, we would like to analyze it with the data summary shown below.

	Population	Room Based Capacity	Bed Based Capacity
Min	0.01	0.01	0.02
Q1 (25th percentile)	0.92	0.96	0.90
Median	1.00	1.00	1.00
Q3 (75th percentile)	1.00	1.00	1.00
Max	1.01	1.01	1.00
IQR (Interquartile range)	0.08	0.04	0.10
Mean	0.93	0.93	0.93
Standard deviation	0.14	0.16	0.12

With this information, we can see that the median is 1, which means more than 50% of the shelters are almost full, and from the IQR, we can see that the bed based capacity area is much higher than the room based ones, but the standard deviation shows the opposite results. This means that as a whole, the room based capacity is more wide spread, and the central portion of the bed based capacity is more wide spread. Also, from the box plot, we see that there are many outliers in the datasets, so maybe the shelter needs to be examined about why they have much lower occupation. Now, to see whether the capacity type makes the mean of the value different, we would like to apply one sample t-test for each type, and then apply Welch's 2 sample t-test to compare the mean between them. The results are as follows:

```
T-test with population mean (Room Based Capacity) on OCCUPIED_PER_ACTUAL
t-statistic = 3.41
p-value = 0.0
At a 95.0% confidence level, we reject the null hypothesis.
```

```
T-test with population mean (Bed Based Capacity) on OCCUPIED_PER_ACTUAL
t-statistic = -3.11
p-value = 0.0
At a 95.0% confidence level, we reject the null hypothesis.
```

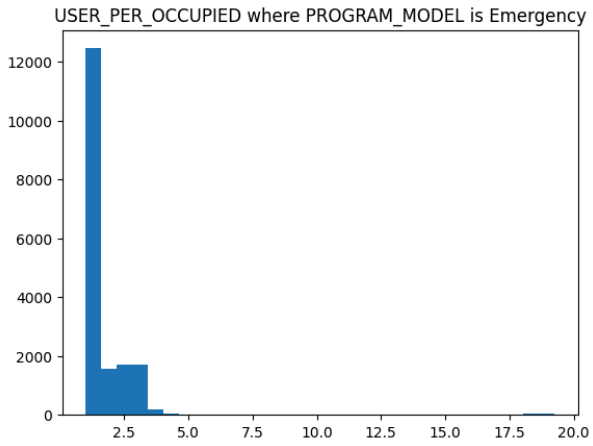
```
Two sample t-test on two categories
t-statistic = 4.5
p-value = 0.0
At a 95% confidence level, we reject the null hypothesis.
```

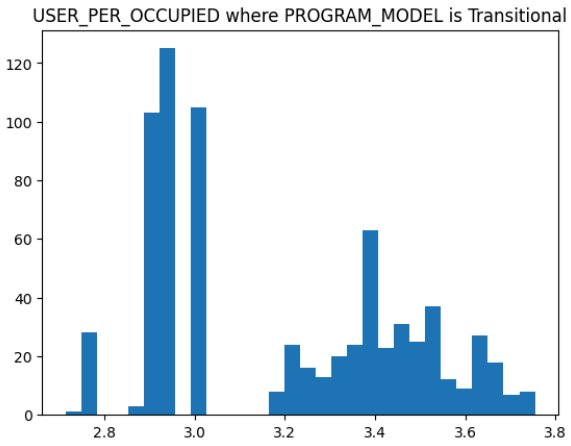
From the very low p-values, we can see that the mean are not equal for both capacity types. By comparing the magnitude of the t-statistics values, we can see that the bed based mean is slightly closer to the population mean than room based, and the bed based mean is lower than population mean whereas room based mean is higher than it.

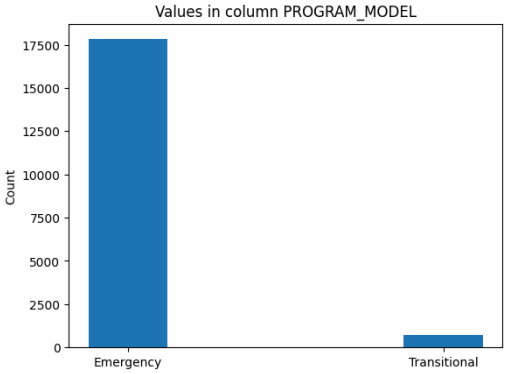
For the next part, we would like to compare the user per occupied room between different types of program models of the room based capacity ones. Below is the table of the amount of shelters in each capacity type and program model. We can see that most shelters are emergency models, and the proportion of emergency models in room based capacity is much higher than that of bed based capacity.

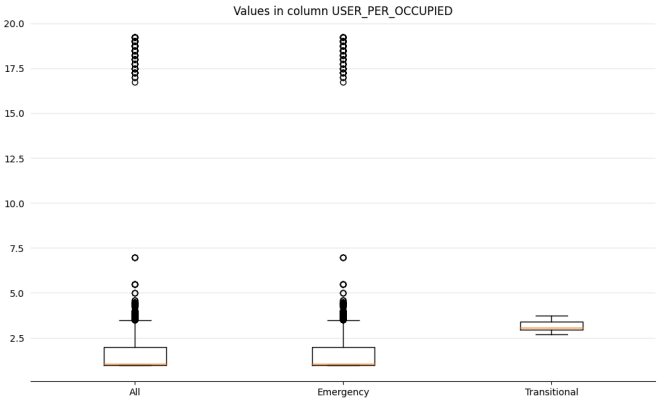
CAPACITY_TYPE	Bed Based Capacity	Room Based Capacity
PROGRAM_MODEL		
Emergency	23726	17815
Transitional	8671	730

And then we used the same method with the previous section. From the plots and the table below, we can see that the transitional models have a distribution that is more normal, and it does not have any outliers. Furthermore, the transitional models have the highest minimum, Q1, median, Q3, maximum, and mean. This means that the amount of users per room is much higher than the emergency ones and the low standard deviation and IQR shows us that there is only a small difference between the users for each occupancy.









	Population	Emergency	Transitional
Min	1.00	1.00	2.71
Q1 (25th percentile)	1.00	1.00	2.95
Median	1.07	1.05	3.08
Q3 (75th percentile)	2.00	2.00	3.42
Max	19.25	19.25	3.75
IQR (Interquartile range)	1.00	1.00	0.47
Mean	1.70	1.64	3.19
Standard deviation	1.65	1.65	0.28

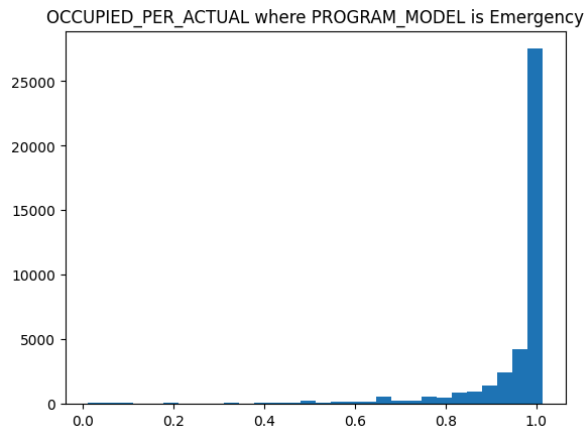
Although we can already see that the mean is different in the summary table above, we still applied t-test to them. The results are as expected, due to the large number of emergency models, the magnitude of the t-statistics is lower than that of the transitional model. However, the mean is still different. In this analysis, I think the reason why the emergency models have many outliers might be that there are many factors that might cause people to need a shelter such as temperature, rain, or other non regular events.

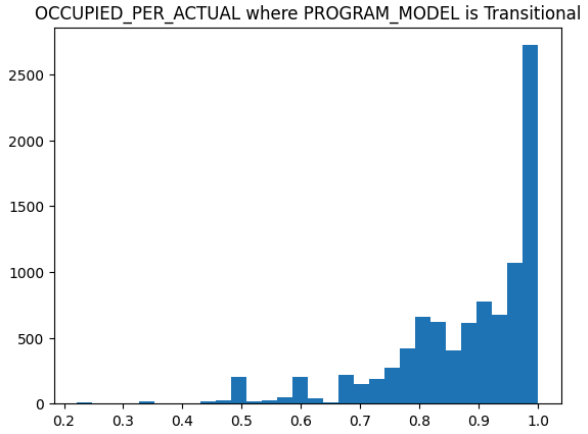
```
T-test with population mean (Emergency) on USER_PER_OCCUPIED
t-statistic = -4.79
p-value = 0.0
At a 95.0% confidence level, we reject the null hypothesis.
```

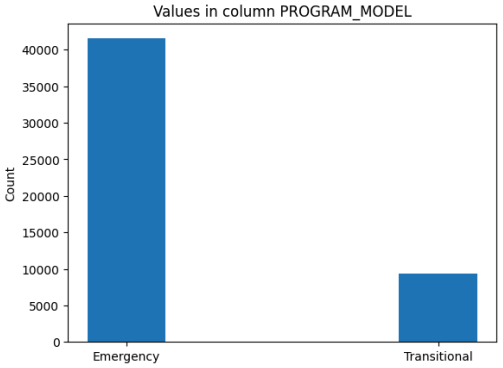
```
T-test with population mean (Transitional) on USER_PER_OCCUPIED
t-statistic = 146.07
p-value = 0.0
At a 95.0% confidence level, we reject the null hypothesis.
```

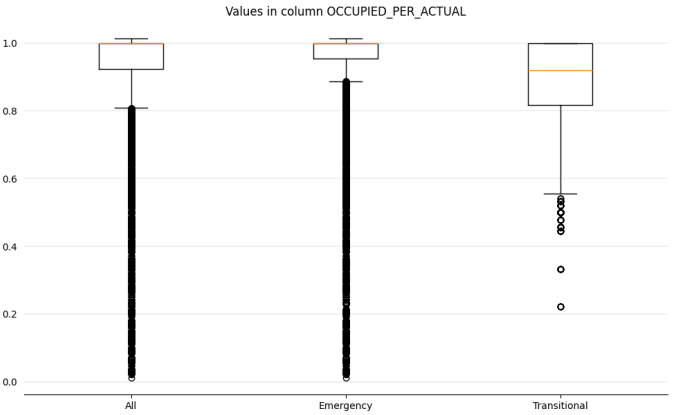
```
Two sample t-test on two categories
t-statistic = -96.64
p-value = 0.0
At a 95% confidence level, we reject the null hypothesis.
```

Now, similar to the first analysis, we would like to see if the mean of occupancy rate differs between different program models. There are also many outliers in the emergency models, and a much larger number of emergency models cause the population to be distributed more alike to it. A difference is that the mean of occupancy in the emergency model is much higher than that of the transitional model. From the t-test, we can also see that both models' mean is different from the population, and that between models are also different.









	Population	Emergency	Transitional
Min	0.01	0.01	0.22
Q1 (25th percentile)	0.92	0.95	0.82
Median	1.00	1.00	0.92
Q3 (75th percentile)	1.00	1.00	1.00
Max	1.01	1.01	1.00
IQR (Interquartile range)	0.08	0.05	0.18
Mean	0.93	0.94	0.88
Standard deviation	0.14	0.14	0.13

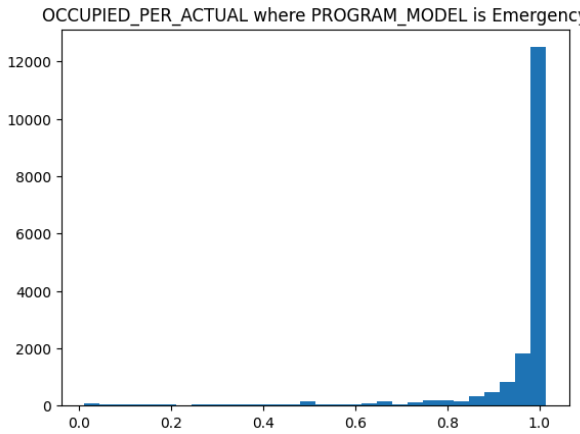
```
T-test with population mean (Emergency) on OCCUPIED_PER_ACTUAL
t-statistic = 16.79
p-value = 0.0
At a 95% confidence level, we reject the null hypothesis.
```

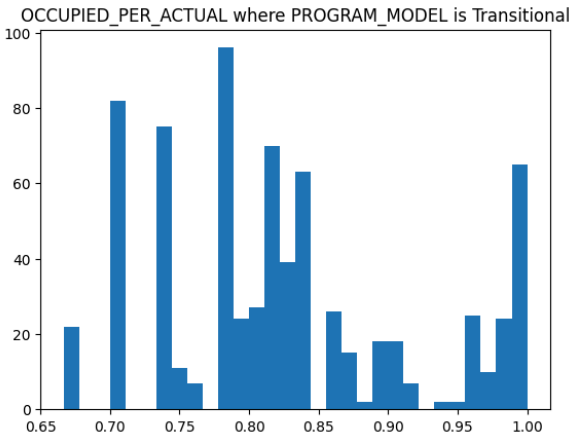
```
T-test with population mean (Transitional) on OCCUPIED_PER_ACTUAL
t-statistic = -37.45
p-value = 0.0
At a 95% confidence level, we reject the null hypothesis.
```

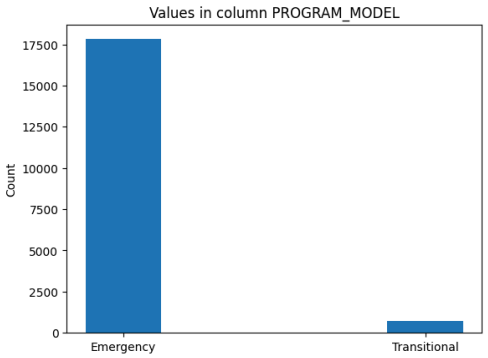
```
Two sample t-test on two categories
t-statistic = 40.98
p-value = 0.0
At a 95% confidence level, we reject the null hypothesis.
```

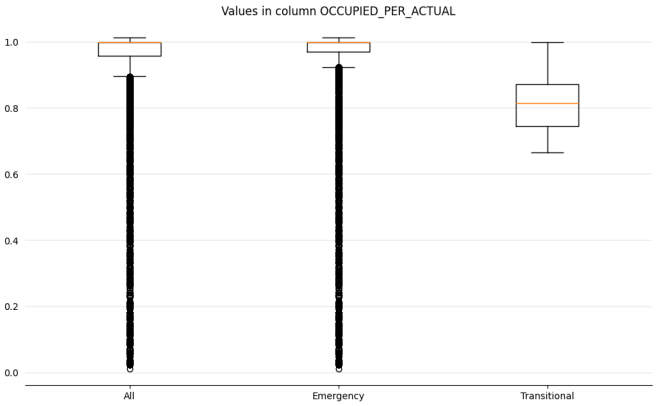
Lastly, we compute if the mean of occupancy rate differs between different program models for room and bed based capacity types. From the results of room based capacity types, we can see the same results as above, larger amount , higher mean, and wider spread of

emergency models, better distribution in transitional models, and the means are different in all three datasets.









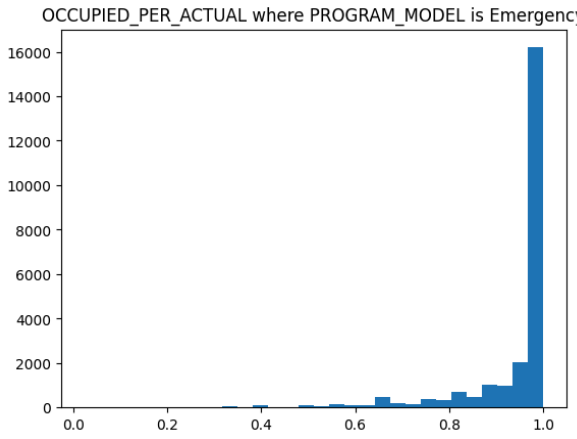
	Population	Emergency	Transitional
Min	0.01	0.01	0.67
Q1 (25th percentile)	0.96	0.97	0.75
Median	1.00	1.00	0.81
Q3 (75th percentile)	1.00	1.00	0.87
Max	1.01	1.01	1.00
IQR (Interquartile range)	0.04	0.03	0.12
Mean	0.93	0.94	0.82
Standard deviation	0.16	0.16	0.09

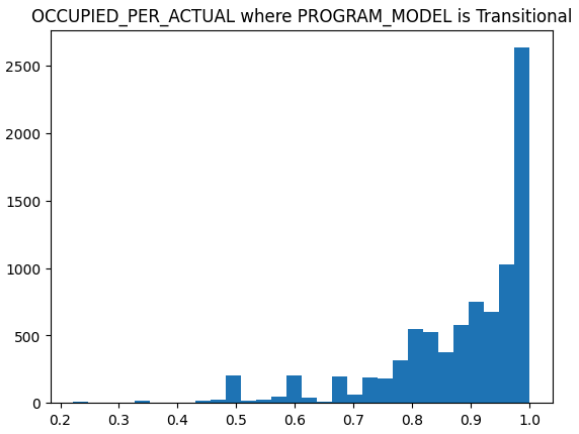
```
T-test with population mean (Emergency) on OCCUPIED_PER_ACTUAL
t-statistic = 7.03
p-value = 0.0
At a 95% confidence level, we reject the null hypothesis.
```

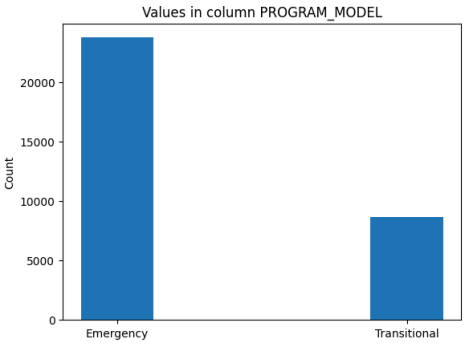
```
T-test with population mean (Transitional) on OCCUPIED_PER_ACTUAL
t-statistic = -31.17
p-value = 0.0
At a 95% confidence level, we reject the null hypothesis.
```

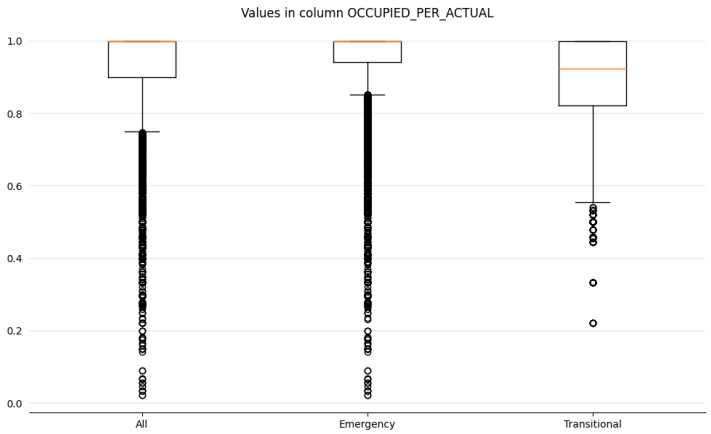
```
Two sample t-test on two categories
t-statistic = 31.71
p-value = 0.0
At a 95% confidence level, we reject the null hypothesis.
```

Below is the same EDAs and t-test of the bed based capacity. We can see the similar results as above, such as the wider spread as well as the mean of occupancy of emergency models and the means are different in all three datasets. However, the distribution transitional models are also skewed to the left, and due to the larger proportion of transitional models in bed based capacity, the t-statistics value for 1 sample t-tests are more similar compared to that of room based capacity.









	Population	Emergency	Transitional
Min	0.02	0.02	0.22
Q1 (25th percentile)	0.90	0.94	0.82
Median	1.00	1.00	0.92
Q3 (75th percentile)	1.00	1.00	1.00
Max	1.00	1.00	1.00
IQR (Interquartile range)	0.10	0.06	0.18
Mean	0.93	0.94	0.89
Standard deviation	0.12	0.12	0.13

```
T-test with population mean (Emergency) on OCCUPIED_PER_ACTUAL
t-statistic = 17.94
p-value = 0.0
At a 95% confidence level, we reject the null hypothesis.
```

```
T-test with population mean (Transitional) on OCCUPIED_PER_ACTUAL
t-statistic = -32.12
p-value = 0.0
At a 95% confidence level, we reject the null hypothesis.
```

```
Two sample t-test on two categories
t-statistic = 36.78
p-value = 0.0
At a 95% confidence level, we reject the null hypothesis.
```

In summary, from the capacity type, we can see that the room based ones have slightly higher occupancy rate, and there are a large number of outliers might be caused by weather or other emergency situations. This can also be seen while we are comparing program models. Most of the shelters are emergency models and the mean occupancy rate is much higher than transitional models. Furthermore, more research can be done for the shelters that have lower occupancy. After finding the reason, the problem of not having enough space can be improved.