

# EDA:

Below, an examination of some categorical columns of interest is presented. It's notable that two rows contain NaN values in the Program Models column, which were subsequently removed. Consequently, there are two capacity types, two program models, and four program areas.

```
Capacity Types:
['Room Based Capacity' 'Bed Based Capacity']
Program Models:
['Emergency' 'Transitional' nan]
Program Models cleared null
['Emergency' 'Transitional']
Program Areas
['COVID-19 Response' 'Base Shelter and Overnight Services System'
 'Temporary Refugee Response' 'Winter Programs']
```

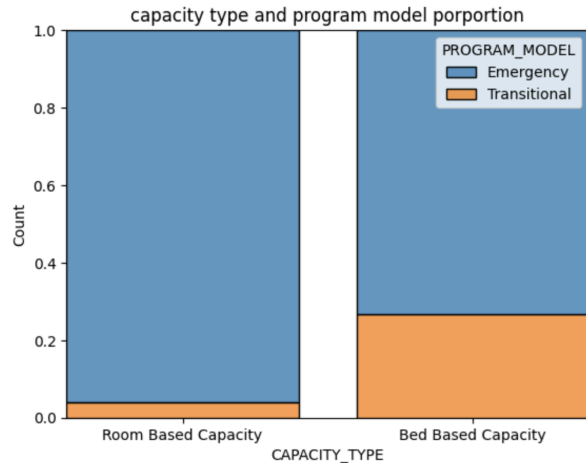
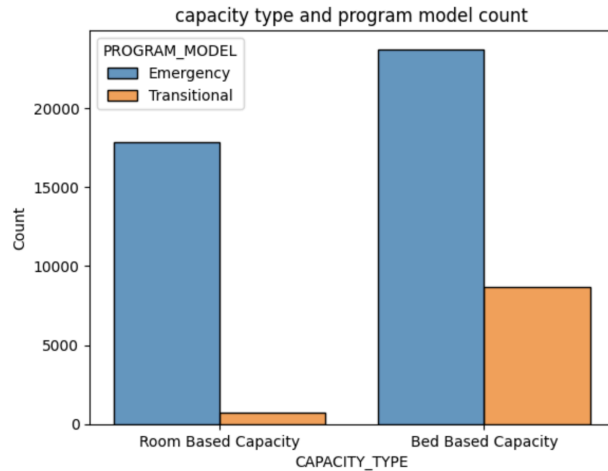
Below is a summary of the data: a total of 50942 rows. From the service user count, one can observe that the mean is significantly distant from the median, hinting that the distribution of service user count is heavily skewed.

Upon examining the bed and room columns, it is evident that there are 32392 rows of bed capacity type and 18545 rows of room capacity type. The two capacity types are represented by their corresponding columns (i.e., capacity\_actual\_bed and occupied bed for bed capacity type).

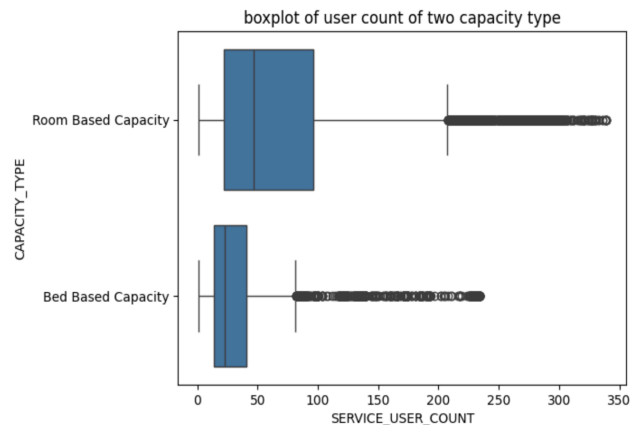
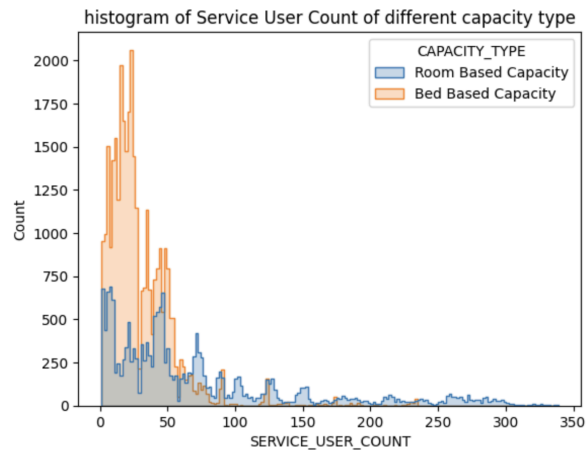
Analyzing the means of the occupied and actual columns, it can be observed that most shelters, on average, are near full capacity.

	PROGRAM_ID	SERVICE_USER_COUNT	CAPACITY_ACTUAL_BED	OCCUPIED_BEDS	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS
count	50942.000000	50942.000000	32397.000000	32397.000000	18545.000000	18545.000000
mean	13986.022005	45.728515	31.628145	29.781400	55.549259	52.798598
std	1705.241577	53.326660	27.128189	26.379825	59.448805	58.792954
min	11791.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	12233.000000	15.000000	15.000000	14.000000	19.000000	16.000000
50%	14251.000000	28.000000	25.000000	23.000000	35.000000	34.000000
75%	15651.000000	51.000000	43.000000	41.000000	68.000000	66.000000
max	16611.000000	339.000000	234.000000	234.000000	268.000000	268.000000

Below, a bar plot displays the distribution of the program model within each capacity type. It is evident that most room-based capacities are used for emergency models.

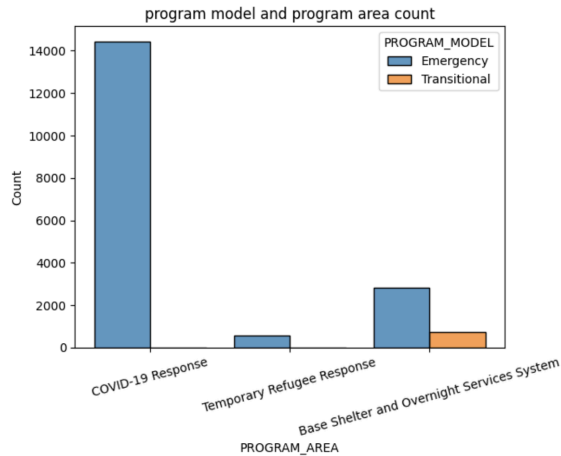


Before delving into each capacity type individually, let's examine another comparison between the two types. The histogram below illustrates that bed-type shelters tend to have a greater number of shelters with a small user count, whereas most shelters with high user counts belong to the room capacity type. (intuitively correct)



Next, let's delve into the exploration of room-based capacity. It's notable that all transitional models within the room type belong to the Base Shelter and Overnight Services System program area.

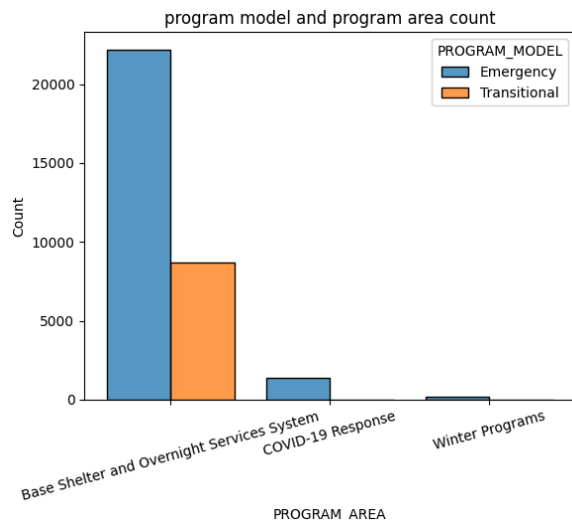
Upon calculating the occupancy rate, it's evident that the average occupancy rate is near 1. From the plot and count below, we observe that approximately one-third of the rooms are not fully occupied, causing the box plot to extend fully to the right extreme.



	SERVICE_USER_COUNT	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS	OCCUPANY_RATES
count	18545.000000	18545.000000	18545.000000	18545.000000
mean	73.587166	55.549259	52.798598	0.934087
std	73.319030	59.448805	58.792954	0.163241
min	1.000000	1.000000	1.000000	0.012048
25%	22.000000	19.000000	16.000000	0.958333
50%	47.000000	35.000000	34.000000	1.000000
75%	96.000000	68.000000	66.000000	1.000000
max	339.000000	268.000000	268.000000	1.014085

room count: 18545  
not full room count: 7024

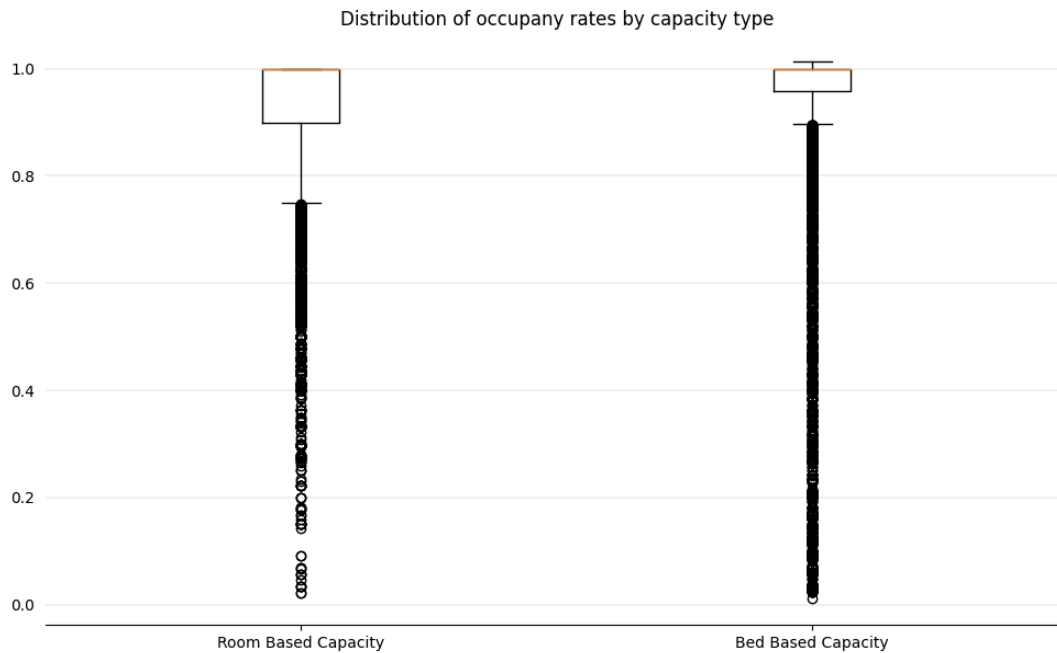
Next, let's examine the bed capacity type. The data is filtered out, and occupancy is calculated. A similar trend is observed in the occupancy rate being near zero. However, the bed capacity type has significantly fewer service user counts compared to the room type, which is intuitively correct.



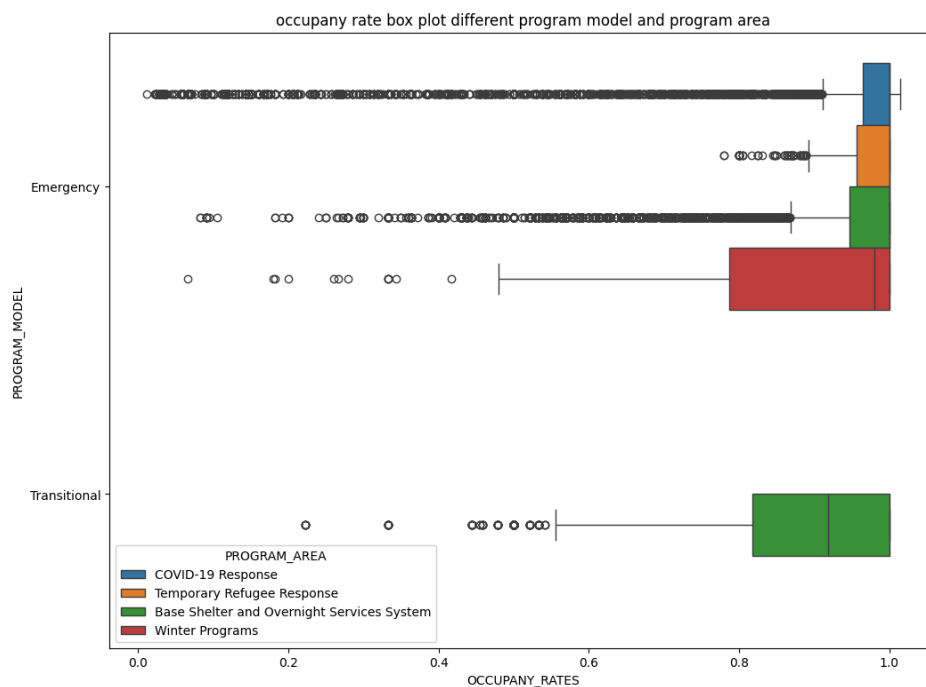
	SERVICE_USER_COUNT	CAPACITY_ACTUAL_BED	OCCUPIED_BEDS	OCCUPANY_RATES
count	32397.000000	32397.000000	32397.000000	32397.000000
mean	29.781400	31.628145	29.781400	0.927895
std	26.379825	27.128189	26.379825	0.122558
min	1.000000	1.000000	1.000000	0.022727
25%	14.000000	15.000000	14.000000	0.900000
50%	23.000000	25.000000	23.000000	1.000000
75%	41.000000	43.000000	41.000000	1.000000
max	234.000000	234.000000	234.000000	1.000000

bed count: 32397  
not full bed count: 15582

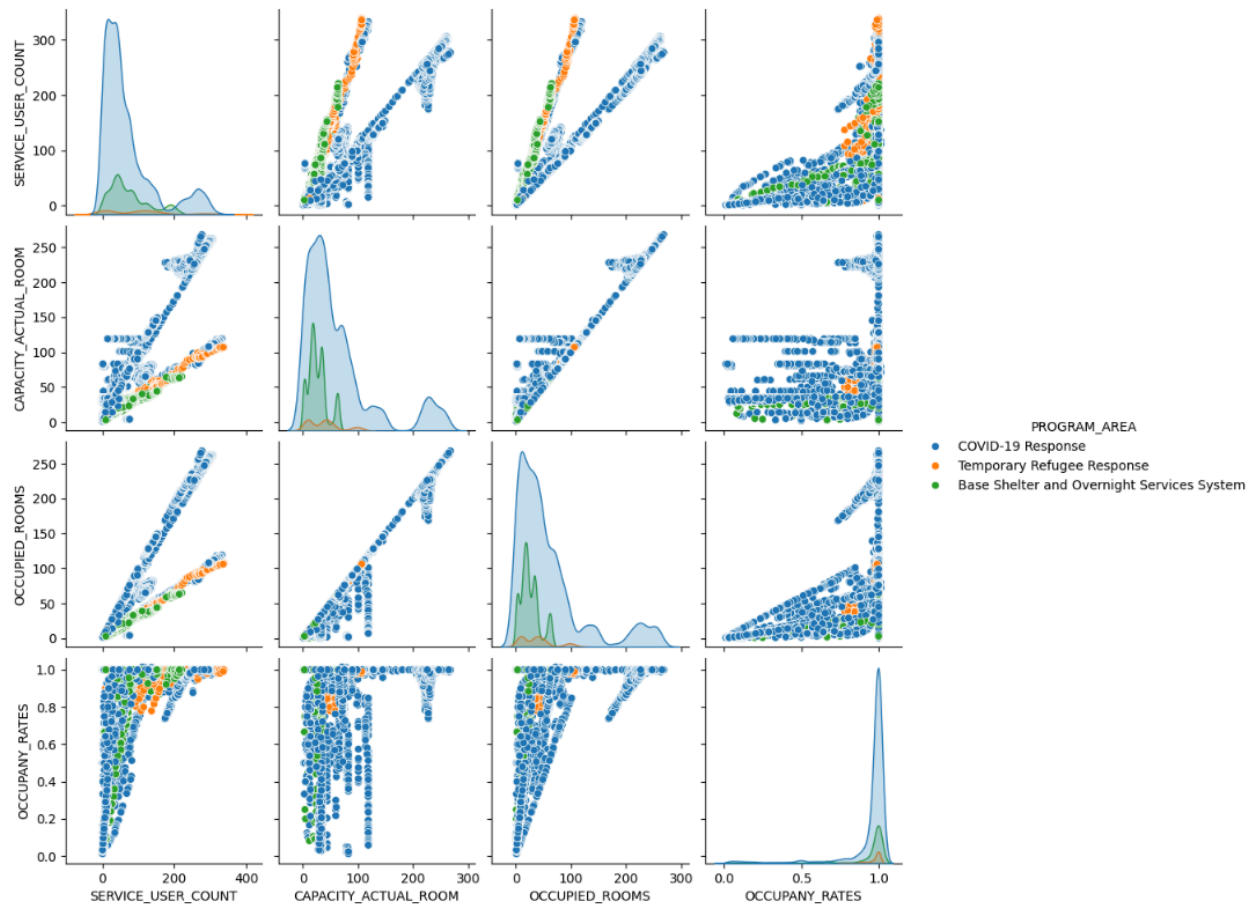
After calculating the occupancy rate, the box plot of the two occupancy rates is drawn. As expected, the box is near 1.0 with numerous outliers below (heavily skewed left). This implies that most of the shelters are either full or nearly full this year.



Below, a comparison of occupancy rates between different program models and program areas is presented. The Covid-19 response has the smallest box towards the right extreme. This indicates that Covid-19 response shelters are the fullest this year compared to other shelters. Also note that there is only a base shelter and overnight services system for transitional program model.



A paired scatter plot of each numerical variable is plotted for each capacity type. No interesting patterns emerge in the bed type data, but some intriguing patterns are observed when examining room data. It appears that there is a different relationship between Covid-19 response shelters and the other two shelter types when comparing room and user count. In row 1 column 3, it's evident that other shelters have a larger slope compared to Covid-19 response in this relationship. Thus, we can infer that Other Shelter tends to accommodate more users compared to Covid-19 shelters.



## T-test:

After examining the data, several research questions come to mind that could be followed up with a t-test. Let's go through them one by one.

Is the occupancy rate the same for bed capacity type and room capacity type? Is the occupancy rate 1 (all full)?

```
ttest on two capacity type on the overall dataset occupancy rates
t-statistic = 4.845858377006688
p-value = 1.2643561358159322e-06
```

```
sample t-test check if occupancy rates population mean is 1
t-statistic = -113.59634972790509
p-value = 0.0
```

Null hypothesis is rejected in both cases since  $p\text{-value} < 0.05$ .

Is the user count the same for bed capacity type and room capacity type?

```
ttest on two capacity type on the overall dataset user count
t-statistic = 97.11765613519675
p-value = 0.0
```

Null hypothesis is rejected  $p\text{-value} < 0.05$

Same question but for a different program model instead of different capacity type.

```
ttest on two program model on the overall dataset user count
t-statistic = 29.937570467283667
p-value = 3.1720139638162956e-195
```

```
ttest on two program model on the overall dataset occupancy rates
t-statistic = 39.07496980654121
p-value = 0.0
```

Both rejected  $p\text{-value} < 0.05$

Are all covid-19 response shelters at full capacity?

```
sample t-test check if occupancy rates population mean is 1
Covid-19 model full capacity?
t-statistic = -49.340636491758765
p-value = 0.0
```

Rejected  $p\text{-value} < 0.05$

To conclude, the dataset comprises two capacity types, two program models, and four program areas, encompassing 50,942 rows. Room-type shelters exhibit significantly higher user counts compared to bed-type shelters. Notably, 95% of room-type shelters are designated for emergency use. The Base Shelter and Overnight Services System program area exclusively applies to Transitional models. Almost all shelters operate at full capacity, with Covid-19 response shelters averaging closer to 100% occupancy compared to others. Scatter plots illustrate that Covid-19 response shelters accommodate fewer people per room than other shelters.

T-tests reveal significant differences in occupancy rates between the two shelter types, despite both being near 100% capacity. Additional tests indicate that Covid-19 response shelters are not operating at full capacity, even in the most severe year.