Assignment 1 Narrative

Zhuoying Li
1004021202

When the data was first presented, I attempted to find some basic information about the dataset, and gather some initial descriptive statistics.
Following is a list of variables

- OCCUPANCY_DATE - date string
- ORGANIZATION_NAME - categorical variable
- PROGRAM_ID - categorical variable
- PROGRAM_NAME - categorical variable
- SECTOR - categorical variable
- PROGRAM_MODEL - categorical variable
- OVERNIGHT_SERVICE_TYPE - categorical variable
- PROGRAM_AREA - categorical variable
- SERVICE_USER_COUNT - numerical variable
- CAPACITY_TYPE - categorical variable
- CAPACITY_ACTUAL_BED - numerical variable
- OCCUPIED_BEDS - numerical variable
- CAPACITY_ACTUAL_ROOM - numerical variable
- OCCUPIED_ROOMS - numerical variable

Then I tried to find the mean and median for all numerical variables
Mean

- SERVICE_USER_COUNT - 45.727171
- CAPACITY_ACTUAL_BED - 31.627149
- OCCUPIED_BEDS - 29.780271
- CAPACITY_ACTUAL_ROOM - 55.549259
- OCCUPIED_ROOMS - 52.798598

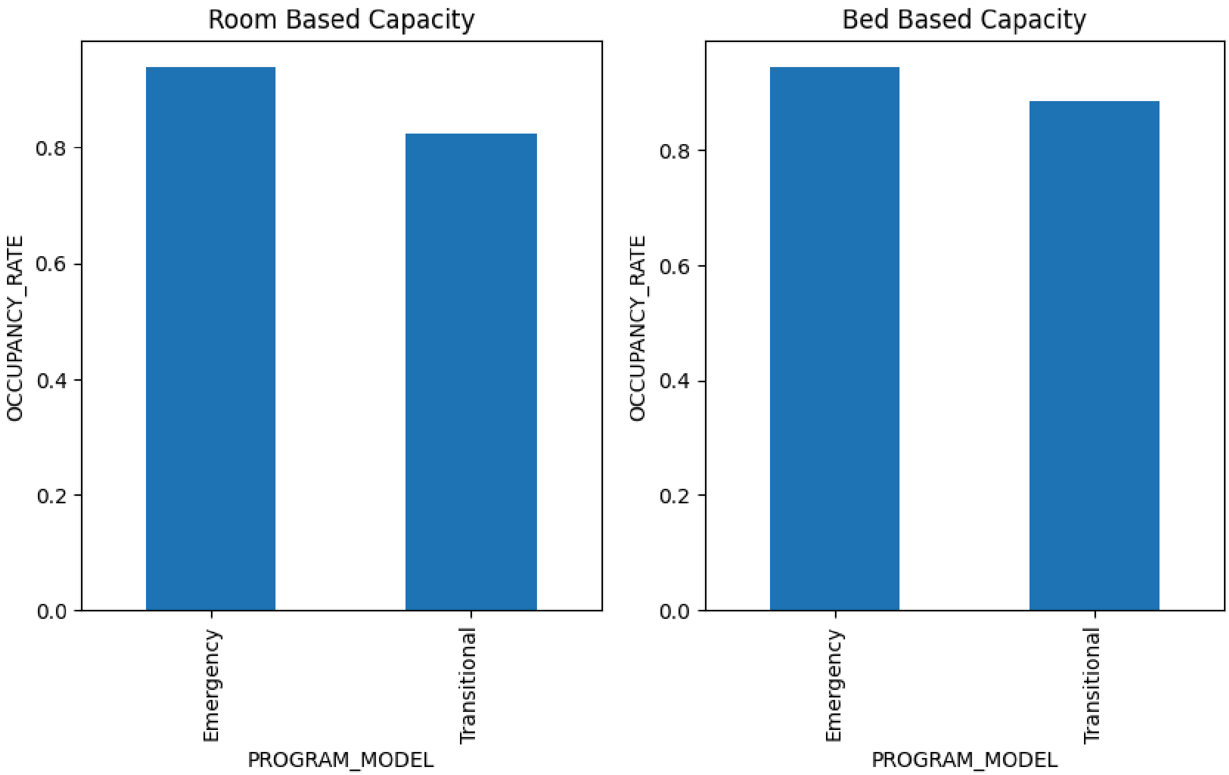
Median

- SERVICE_USER_COUNT - 28.0
- CAPACITY_ACTUAL_BED - 25.0
- OCCUPIED_BEDS - 23.0
- CAPACITY_ACTUAL_ROOM - 35.0
- OCCUPIED_ROOMS - 34.0

We can see there are quite a bit of difference between the mean and median, suggesting the data might be heavily skewed and have a lot of outliers.

Then, I attempted to find out the unique possible values for all categorical variables, so when we proceed with further exploration, we are able to refer back to these values.

Once this has been obtained, I computed a new numerical variable called OCCUPANCY_RATE, this number is calculated using room capacity for room based capacity type rows, and bed capacity for bed based capacity type rows. With this new variable, here is a bar plot based on PROGRAM_MODEL



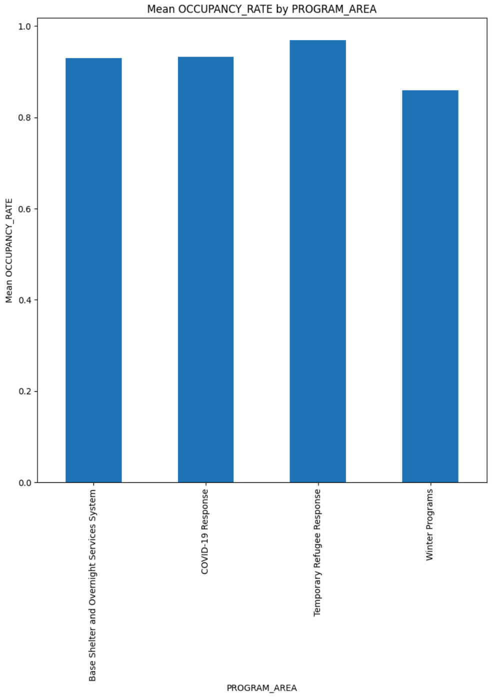
Transitional programs seems to have less occupancy rate compared to emergency, regardless of the capacity type, so, I did a t-test for both capacity types to see if there is really a significant difference.

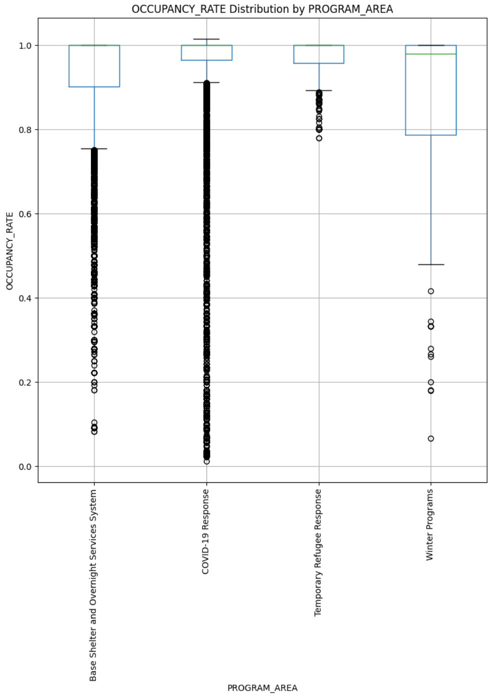
For both, the t-test returned < 0.05 p-value
Room based
T-stat: 31.71080126309493
P-value: 4.4252019739840735e-150
Bed based
T-stat: 36.78483679745313
P-value: 7.273950955976339e-283

Thus we are confident to reject the null hypothesis, meaning there are significant differences between mean occupancy rate of Emergency and Transitional programs.

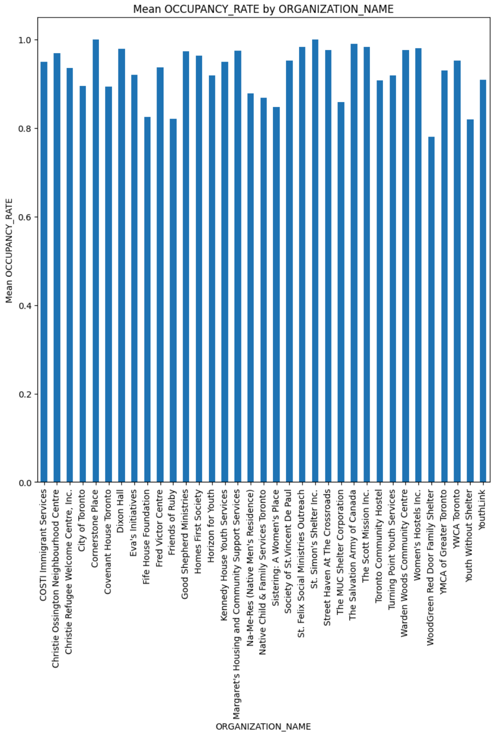
At this point, I want to explore the relationship between more categorical variables and the occupancy rate, so I started to draw both the bar graph and box graph for each categorical variable.

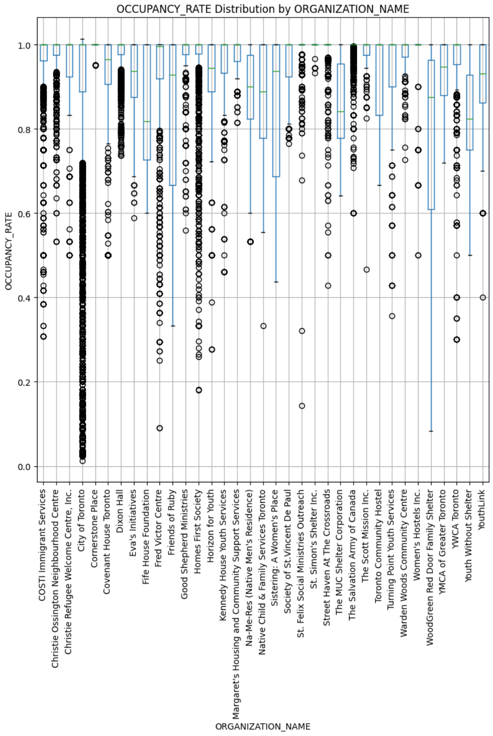
Program Area



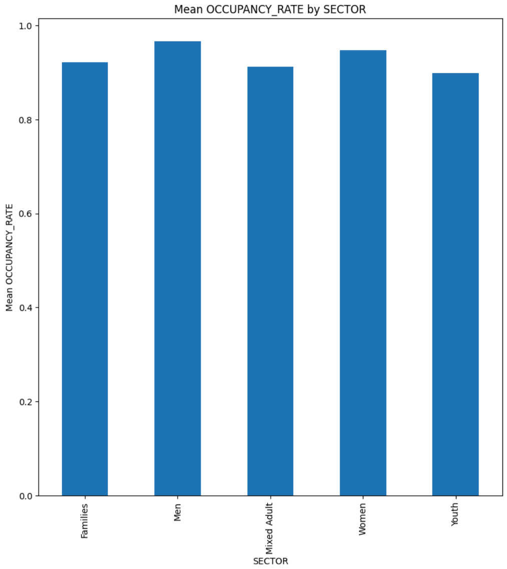


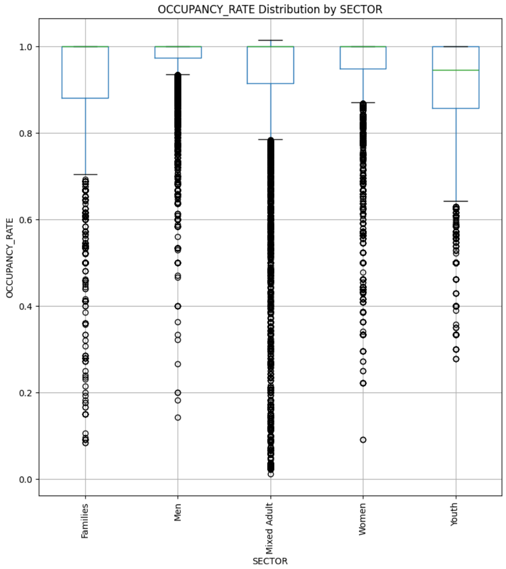
Organization Name



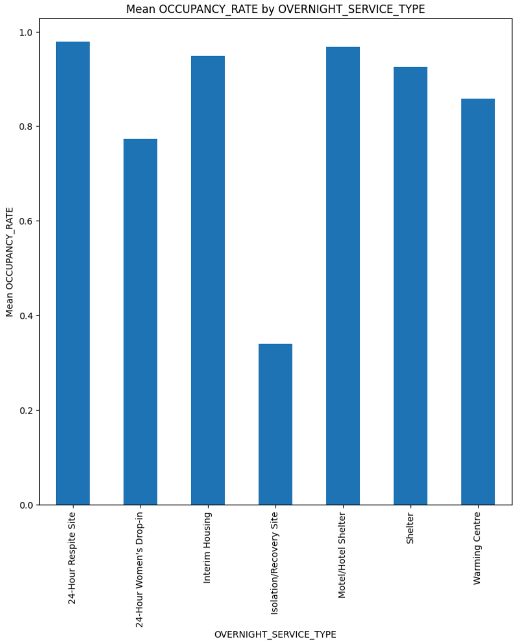


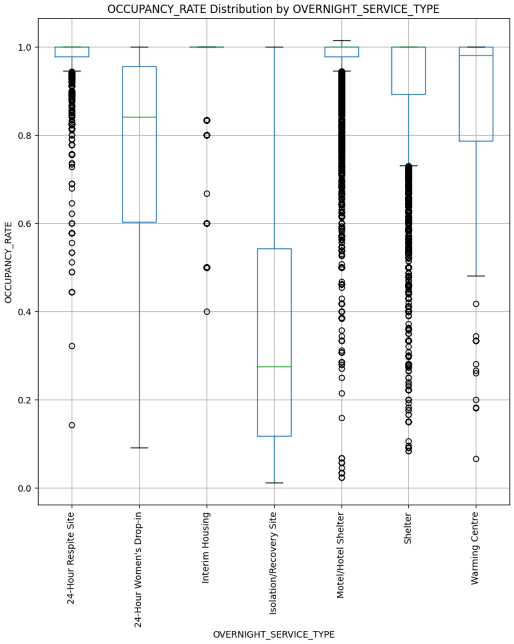
Sector



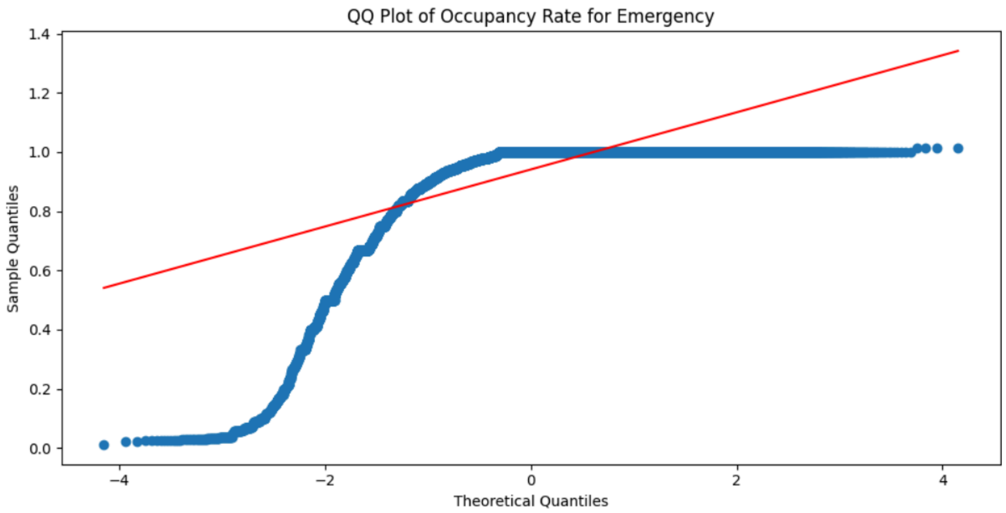


Overnight Service Type



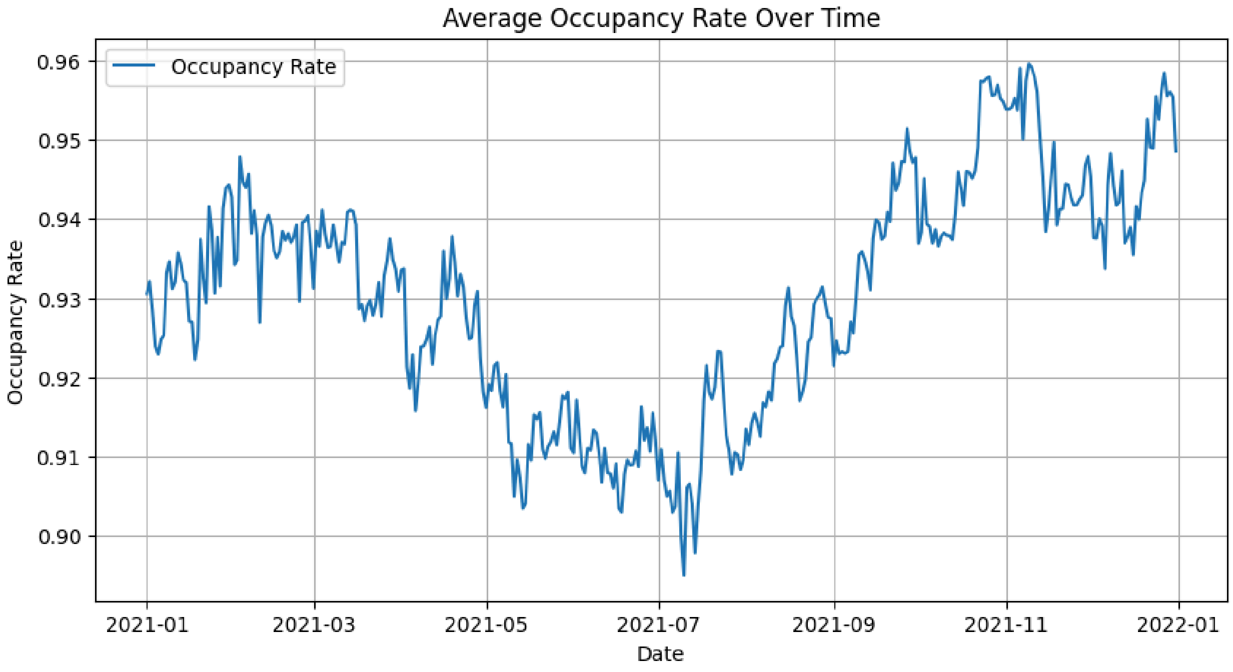


At this point, we can see almost all categories are heavily skewed, almost none of the categories follow a normal distribution, this is again confirmed by drawing some QQ plots. Following is an example QQ plot for emergency programs, we can see it definitely is not a normal distribution.



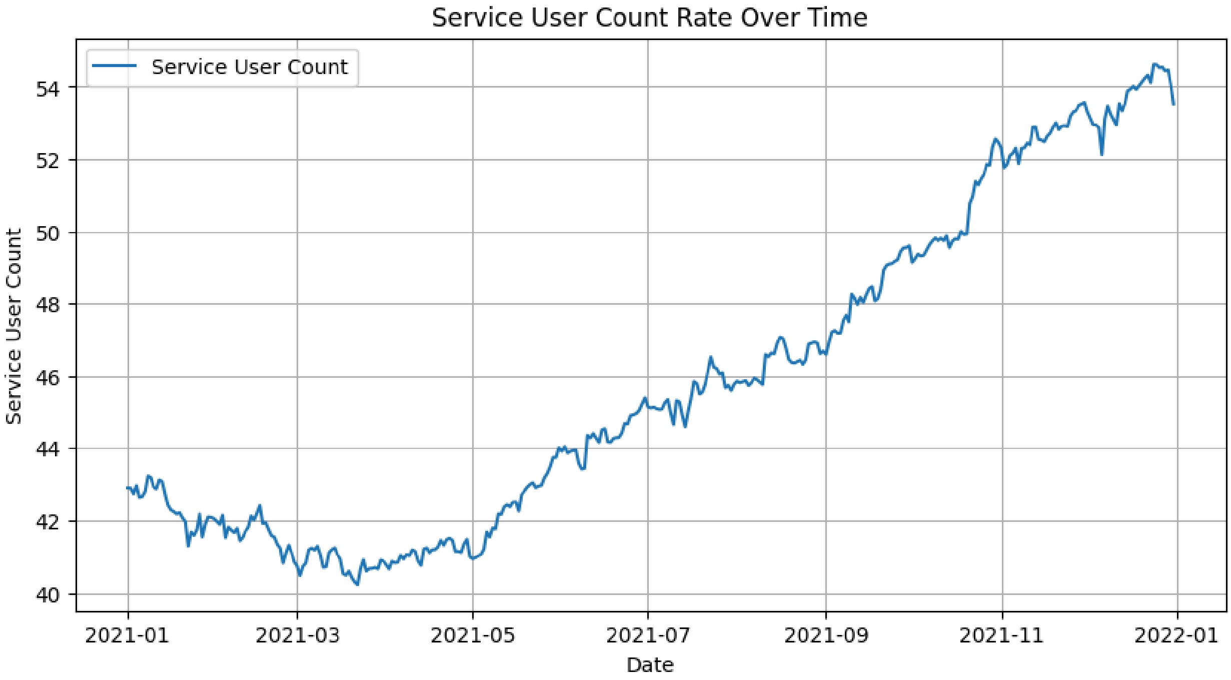
Because of this, it made little sense to continue doing t-tests, as we don't have the proper distribution, the result will be unreliable. I still did a couple in the code, but we won't describe them here.

The final question I want to explore is if there is a relationship between occupancy rate and the time, since the first column of the data is date. Thus I drew a line graph, grouping each point based on date.



This graph shows that there seems to be higher occupancy rate during winter-times, however it is important to note the range, as the average occupancy rate is only changing from around 0.90 to 0.96, so not a significant change.

I also drew the same graph for service users count.



This showed consistent growth between January 2021 and January 2022. Since the occupancy rate didn’t change much during thai time period, it indicates some growth in actual capacity throughout the year.