Course: INF2178 Assignment1

Student Name: Jianheng Chen

Student number: 1005680746

In this assignment, we are studying the dataset of daily occupancies and capacities of Toronto shelters in 2021. My choice of variables are capital types, program model, service user count, capacity actual bed, occupied beds, capacity actual room, occupied rooms, which are the same as suggested in the instruction (the other variables are more like indexes, or a bunch of names that could be a bit irrelevant to the questions I would like to study).

After importing the dataset, I first try to see the first 10 rows and last 10 rows and attributes of the variables in the dataset.

```
Data columns (total 14 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   OCCUPANCY_DATE          50944 non-null  datetime64[ns]
 1   ORGANIZATION_NAME       50944 non-null  object
 2   PROGRAM_ID              50944 non-null  int64
 3   PROGRAM_NAME            50909 non-null  object
 4   SECTOR                  50944 non-null  object
 5   PROGRAM_MODEL           50942 non-null  object
 6   OVERNIGHT_SERVICE_TYPE  50942 non-null  object
 7   PROGRAM_AREA            50942 non-null  object
 8   SERVICE_USER_COUNT      50944 non-null  int64
 9   CAPACITY_TYPE           50944 non-null  object
 10  CAPACITY_ACTUAL_BED     32399 non-null  float64
 11  OCCUPIED_BEDS           32399 non-null  float64
 12  CAPACITY_ACTUAL_ROOM    18545 non-null  float64
 13  OCCUPIED_ROOMS          18545 non-null  float64
dtypes: datetime64[ns](1), float64(4), int64(2), object(7)
memory usage: 5.4+ MB
None
OCCUPANCY_DATE            365
ORGANIZATION_NAME          35
PROGRAM_ID                169
...
OCCUPIED_BEDS             191
CAPACITY_ACTUAL_ROOM      221
OCCUPIED_ROOMS            248
dtype: int64
```

From this I know most of the numerical data will be duplicated as the unique numbers are very little compared to the whole observations. The attributes of the variables of interest are good to use and no transformation is required. Then I can conduct selection for those variables of my interest.

After this, I start my process for data cleaning. I check the NAs in the dataset, which prompt me the result of:

```
CAPACITY_TYPE                0
PROGRAM_MODEL                2
SERVICE_USER_COUNT           0
CAPACITY_ACTUAL_BED      18545
OCCUPIED_BEDS            18545
CAPACITY_ACTUAL_ROOM     32399
OCCUPIED_ROOMS           32399
dtype: int64
```

From this I realize that can observe that 'Room based capacity' has no Bed information, 'Bed based capacity' has no Room information, which is reasonable. In addition, there are 2 rows of information missing in the program model. I choose to remove the two-missing row in program model, and this will not affect my observations for the dataset. For the other missing values, I replace them with 0, which will better describe the actual status of Room and Bed availabilities.

Following this, I divide occupied bed by capacity actual bed to derive the occupancy rate of bed. With the same process I get the occupancy rate of room. Monitoring the first 10 rows:

BED_OCCUPANCY	ROOM_OCCUPANCY	OCCUPANCY_RATE
NaN	0.896552	0.896552
NaN	1.000000	1.000000
NaN	0.821429	0.821429
NaN	1.000000	1.000000
NaN	0.928571	0.928571
0.75	NaN	0.750000
NaN	0.956522	0.956522
NaN	0.956522	0.956522
NaN	1.000000	1.000000
NaN	0.975610	0.975610

The data looks reasonable (0 divide by 0 will prompt NAs). And I combine these two columns to a new column named occupancy rate, which will remove this NAs problem.

I then move to t-test. By there, the t-test of choice will be two sample t-test. The assumption of normality (large enough to fit CLT), independent (proven above), equal variance, quantitative are

fitted and I want to see the mean difference in different groups. For unequal sample size and variance, I will use Welch's t-test in support. The first group of data I want to study is the mean difference between occupancy types. As a result,

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

```
t-statistic = -4.491108081297825
p-value = 7.1112242415027885e-06
```

In both t-tests the p-value are greater than 0.05, the alpha level of choice, from which I will conclude that there is a significant difference between the bed and room.

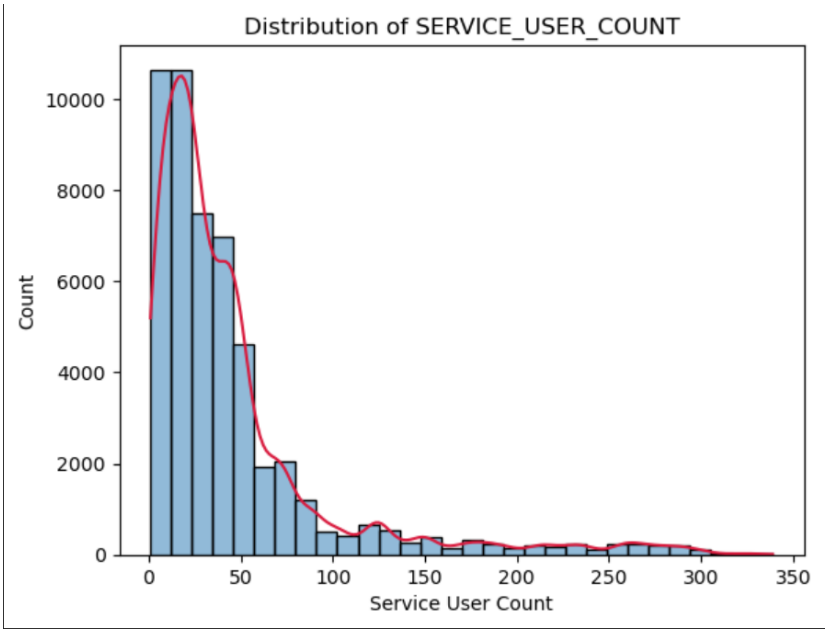
In addition, I am also interested in the mean difference between different program models.

```
t-statistic = 39.07496980654136
p-value = 0.0
```

```
t-statistic = 40.981115372199206
p-value = 0.0
```

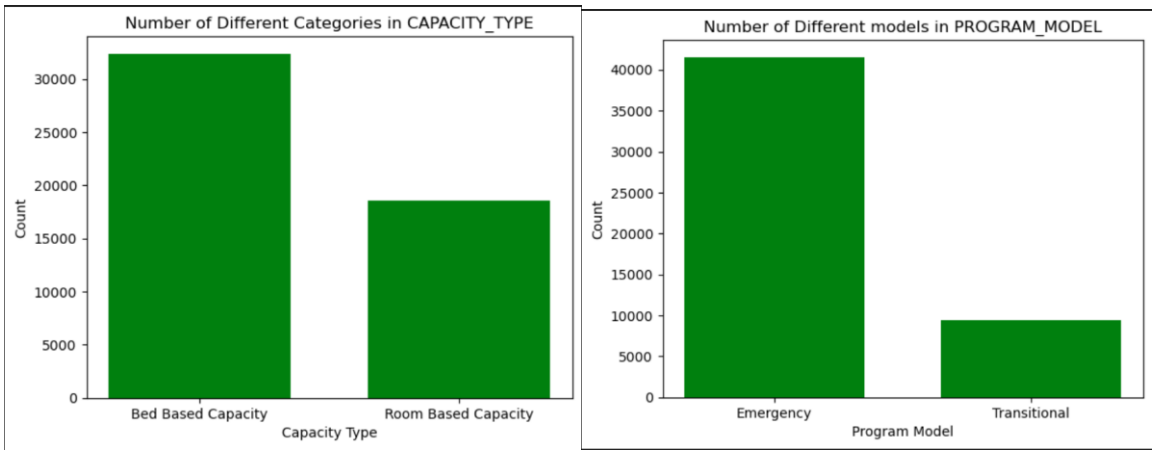
To my surprise, in both t-tests the p-value are very closed to 0, which is smaller than the significance level. It means that there is no significant difference between the emergency and transitional.

Then we will move to the EDA section, I first check the distribution of the service count user.



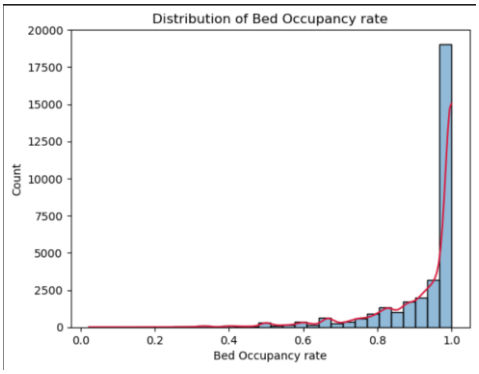
It shows a very severe trend of right skewed, most of the service users are very limited. 10 has the highest frequency, which means that few users are staying in an overnight program as of the occupancy time and date. In addition, there could be some extreme level exist as there is a very long tail.

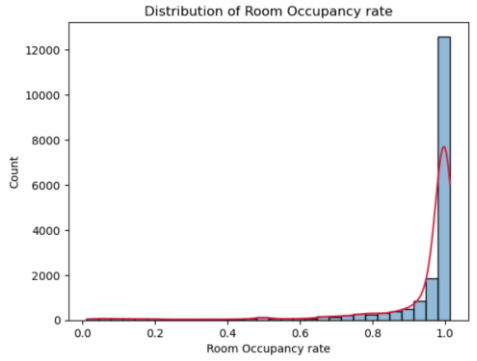
Following this, I want to see the difference in numbers between different categorical variables.



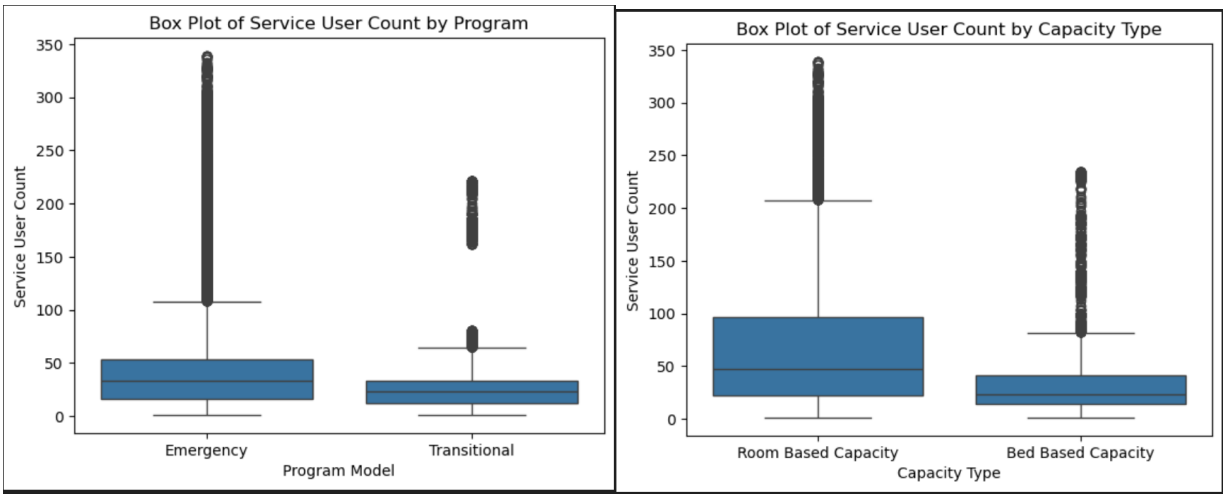
These give a direct visualization for the size of Bed based capacity is greater than Room based capacity (almost 1.8 times), and Emergency model is almost 4times the size of Transitional model. Followingly, the data solely based on the number of beds/rooms is ineffective in understanding the occupancy situations. So, I will get summary statistics for only occupancy columns.

```
Bed Occupancy 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
Setosa summary statistics


Setosa summary statistics


Room Occupancy 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
Setosa summary statistics
...
Setosa summary statistics
```

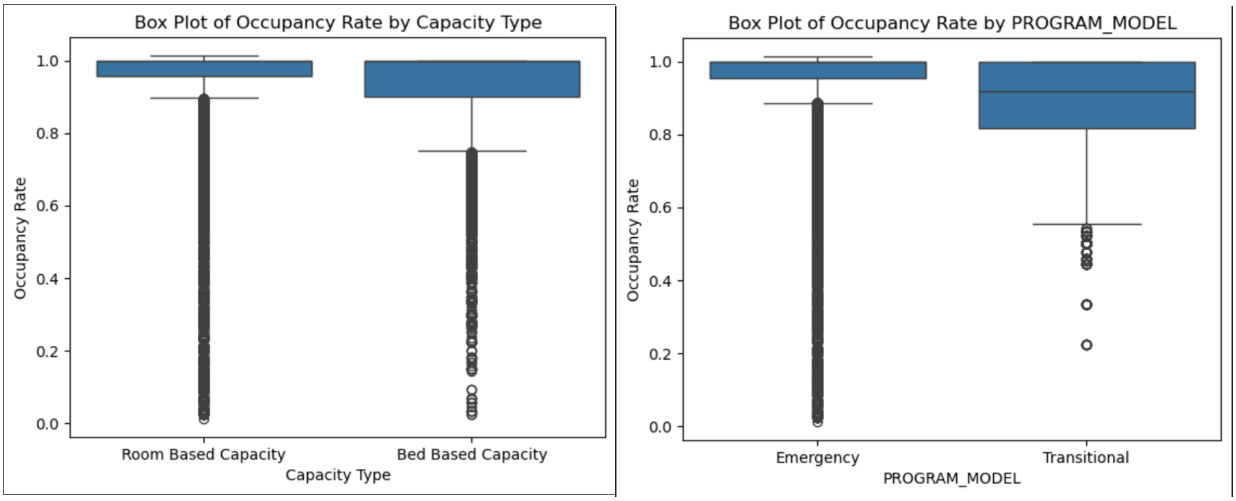




As can be seen, both Bed and Room Occupancy has mean of 0.93 and the 25th percentiles are around 0.9, which means the data are also skewed. To be noticed, the maximum in Room occupancy is 1.01 which is greater than 1, which means that there are some outliers existing.



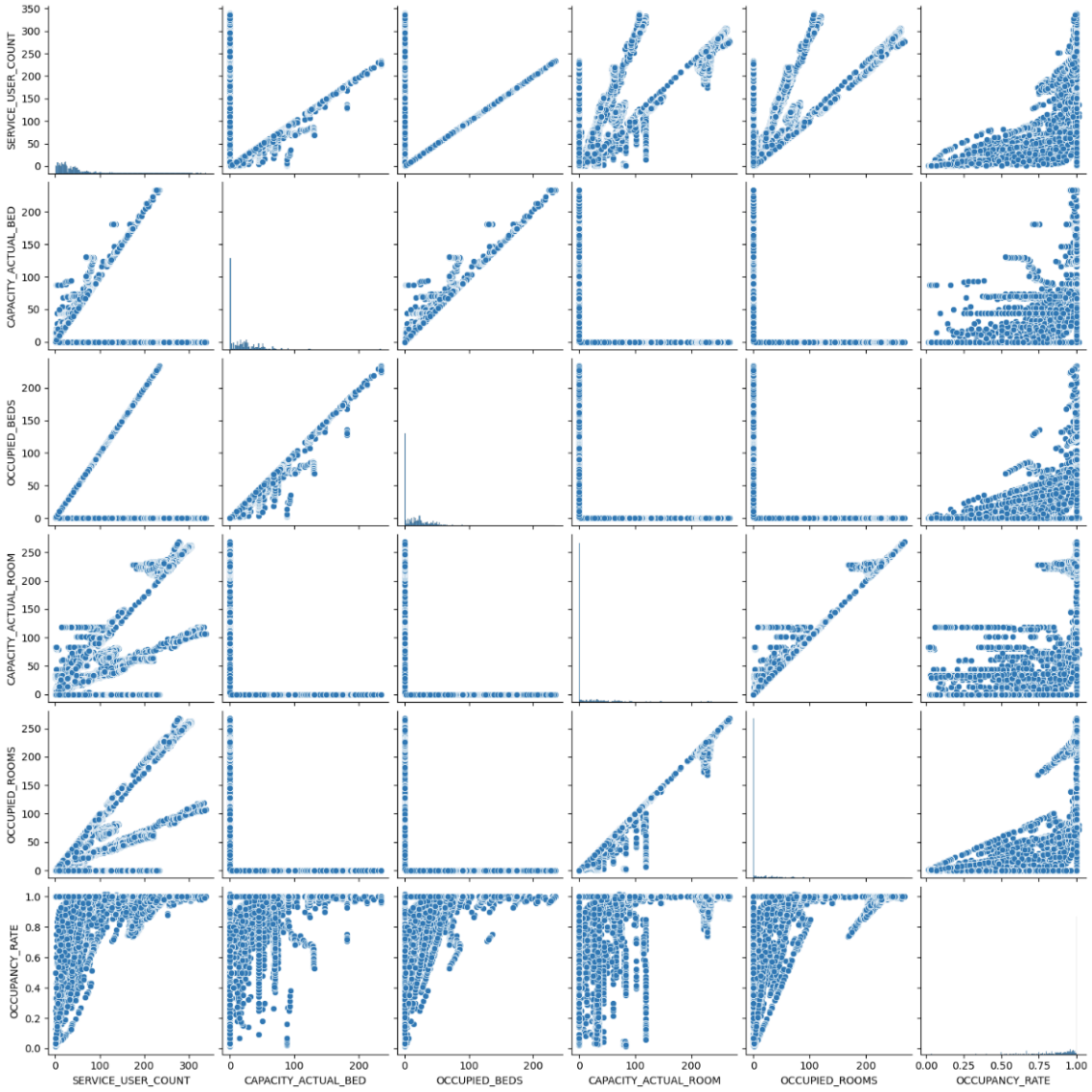
Moving to the box plot of the service user count. We can see all the outliers are appeared above the upper whiskers, in both program model and capacity type cases. Transitional data has apparently less outliers compared with the emergency data. Digging into this, emergency category has a wider range and higher median than the transitional program, suggesting that the central tendency of user counts is higher compared to the transitional program. Room based capacity shows a same status compared to Bed based capacity service that service user count is higher in rooms than in beds.



Moreover, when I study the box plot of the occupancy rates, all outliers are living below the lower whiskers in both cases. It is noticeable that Bed based capacity also shows fewer extreme data compared to Room based capacity, while transitional program has much fewer outliers compared to emergency program. Combing with the EDA we have above, we know that both Room and Bed based capacity has median near to 1. Comparing the wider interquartile range, we

can see more variability exist in the Bed based occupancy rates. For another thing, the transitional program has a noticeably lower median occupancy rate, suggesting lower average occupancy than the emergency program. Additionally, the interquartile range for transitional program is wider, indicating greater variability in occupancy rates.

Lastly, we will move on to the bivariate pair plot.



Most of the graphs show positive correlations with each other as observable in upward trends. It is not surprising that there is no correlation between room and bed data as they are independent. As occupancy rate are derived from bed and room data, the data is clustered near the value of 1.