INF2178- Assignment 1

Student Name: Tony Yang
Student Number: (1003289229)

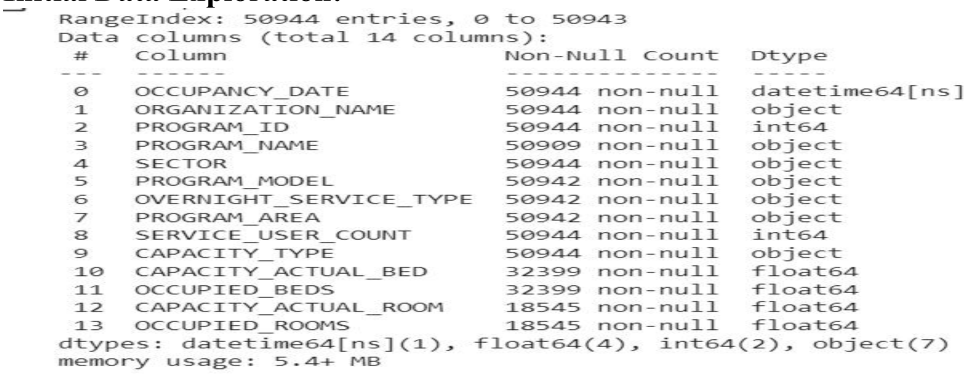
Introduction:

Over the past few years, Toronto has experienced a significant rise in its homeless population. Toronto city has tried their best in shelter support for homeless people,but there are often not enough shelter spaces still. Our goal is to analyze the shelter dataset and investigate the trend in shelter usage. Upon initial examination of the dataset, the following research questions have been formulated.

Research Questions:

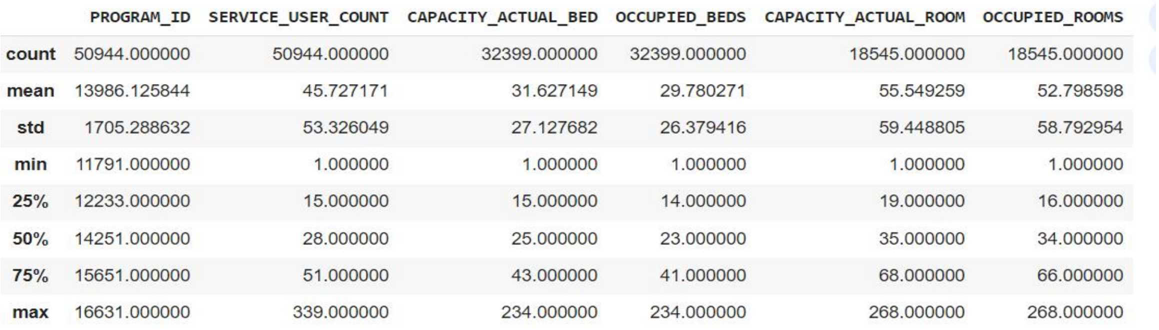
1.How does shelter occupancy and capacity differ throughout the year?
2. Are there differences in occupancy rates and capacity utilization between different capacity type.
3. Is there a statistically significant difference in the capacity, occupancy ,service user count and occupancy rates between bed-based and room-based shelters in Toronto?

Initial Data Exploration:



```
RangeIndex: 50944 entries, 0 to 50943
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
```

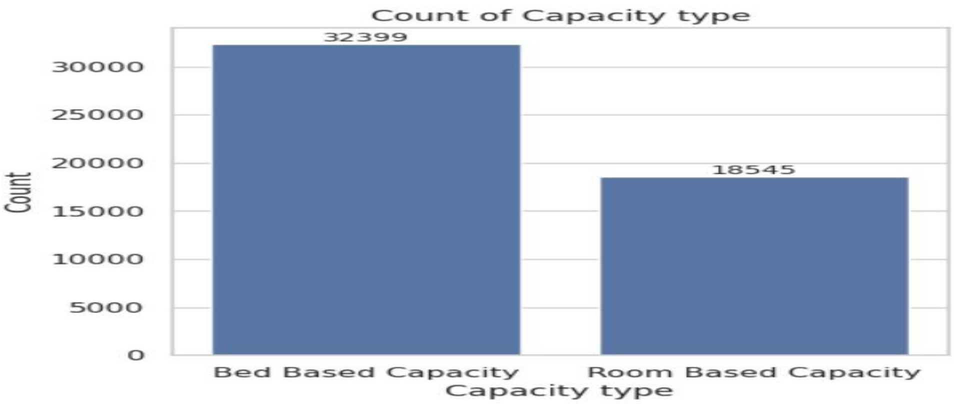
At the first step, I started examining the dataset. The dataset contains 50944 data points with 14 columns as the figure show above. I also computed statistics for all numerical columns include central tendency, dispersion, and shape as figure show below. I will focus on examination of these numerical columns and the CAPACITY_TYPE.



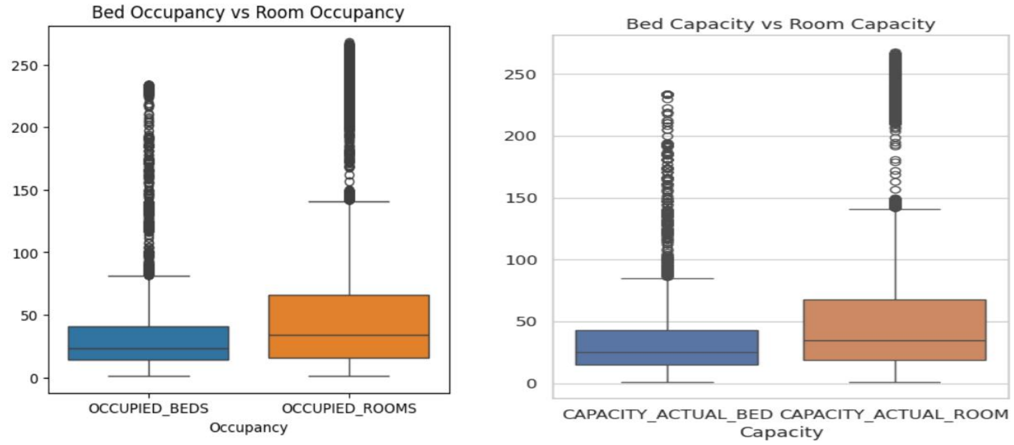
	PROGRAM_ID	SERVICE_USER_COUNT	CAPACITY_ACTUAL_BED	OCCUPIED_BEDS	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS
count	50944.000000	50944.000000	32399.000000	32399.000000	18545.000000	18545.000000
mean	13986.125844	45.727171	31.627149	29.780271	55.549259	52.798598
std	1705.288632	53.326049	27.127682	26.379416	59.448805	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	16631.000000	339.000000	234.000000	234.000000	268.000000	268.000000

Data Virtualization:

I start visualizing the capacity_type column by counting the occurrences of each type, it helps us understand the distribution of shelter types. The Bed Based type has count of 32399 whereas Room based type has count of 18545.

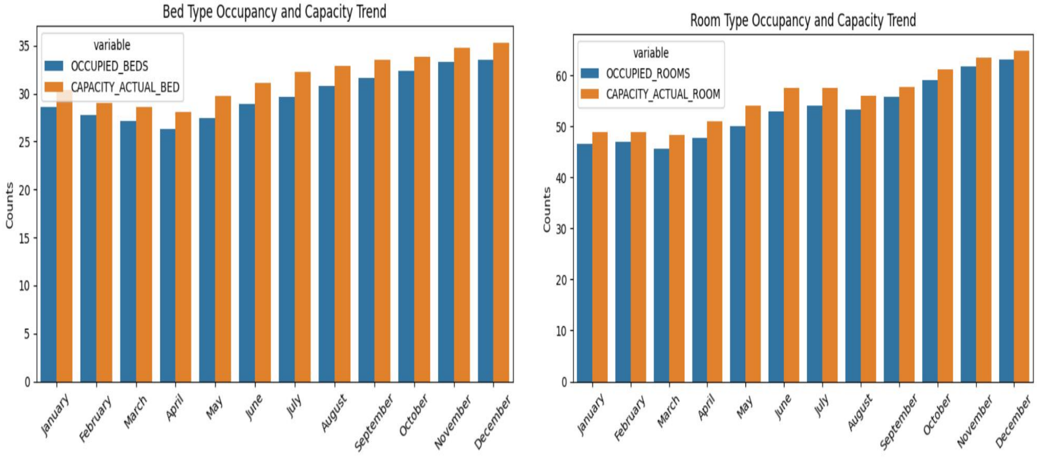


Following that observation, I see that there are two distinct types of shelters: Bed-based and Room-based. To help in future analysis, I created independent DataFrames corresponding to these two shelter types. I examined the occupancy and capacity for both types through boxplots as the figures show below.



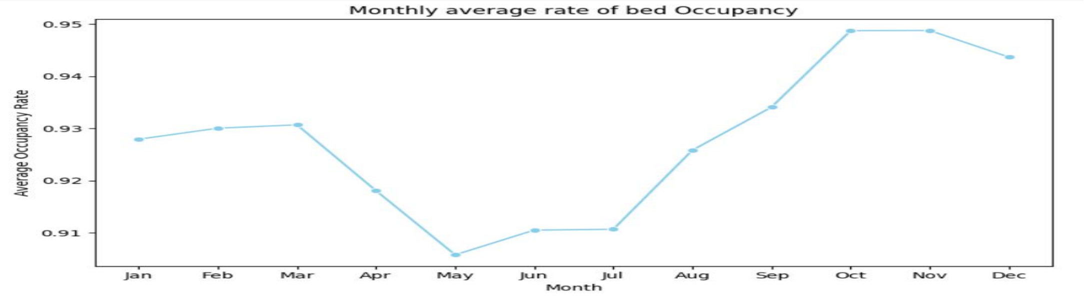
From the plot, I have obtained that for bed-based shelters, the mean occupancy is approximately 29.78, with a standard deviation of 26.38 whereas the actual capacity of approximately 31.63, with a standard deviation of 27.13. For room-based shelters, it has a higher mean occupancy of about 52.80, with a larger standard deviation of 58.79. And room-based shelters has a higher average actual capacity of about 55.55, with a larger standard deviation of 59.45.

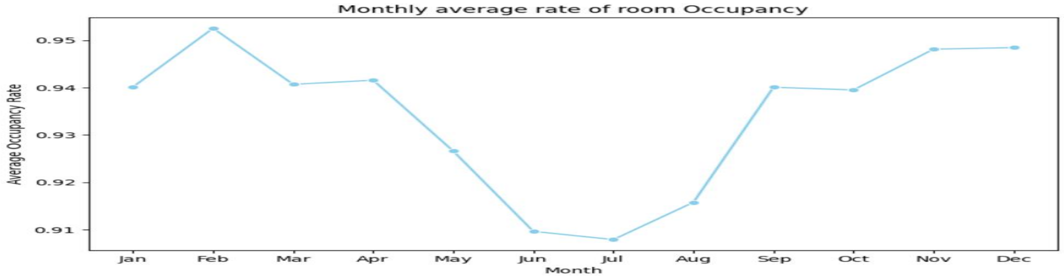
Following this, I fetch out and created the month column from the date column and proceed to generate trends for both Bed-based and Room-based occupancy and capacity group by month.



From the plots, we can see that capacity and occupancy throughout the year shows a continued and significant trend of growth. The increases in occupancy and capacity highlight the continued need for shelter support within the community.

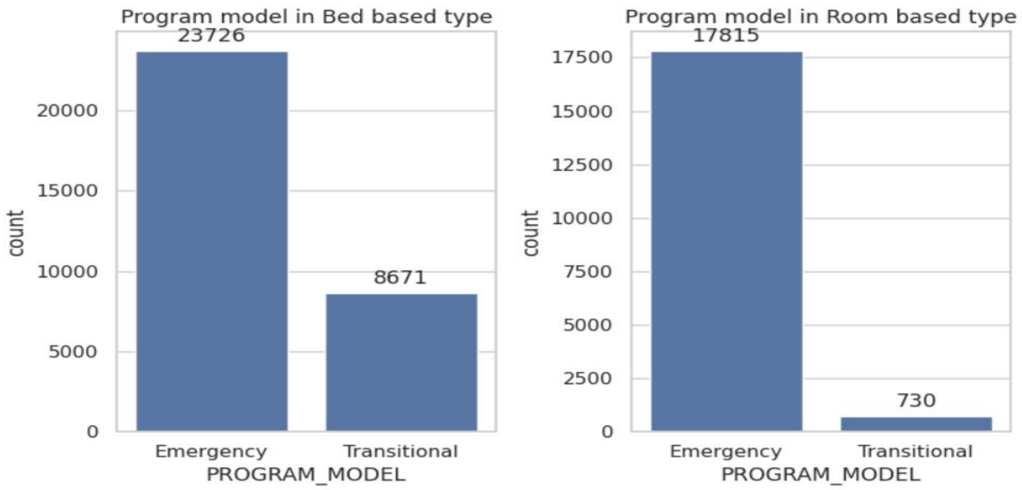
Followed by the trend, I also calculated and plot the occupancy and capacity rate which indicates the proportion of available shelter space that is being utilized as the figures show below:



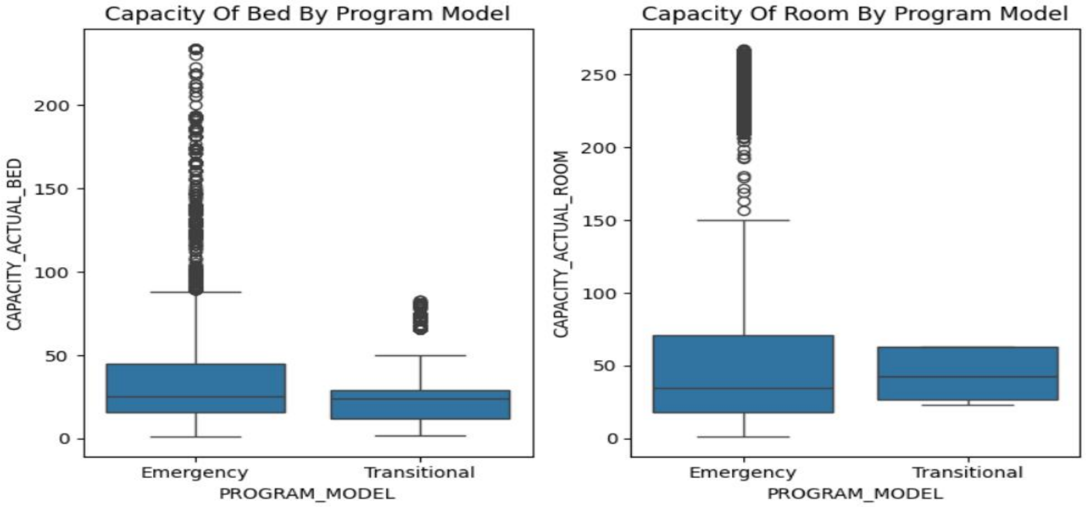


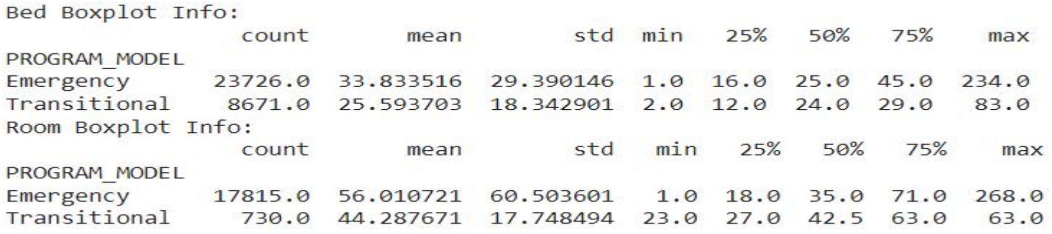
From the visualizations, it shows that both bed-based and room-based shelters have similar occupancy rates consistently exceeding 90%. However, there is difference in the month of the lowest occupancy rates: for bed-based shelters, it occurs in May, whereas for room-based shelters, it is in July. These trends reflect a high level of utilization across the period, suggesting that shelters are operating at or near full capacity with limited room for additional occupants.

Move on, I shift my focus to examining the program model for both bed and room types. I proceed by generating a histogram to visually represent the counts for both emergency and transitional program models as the plot show below.

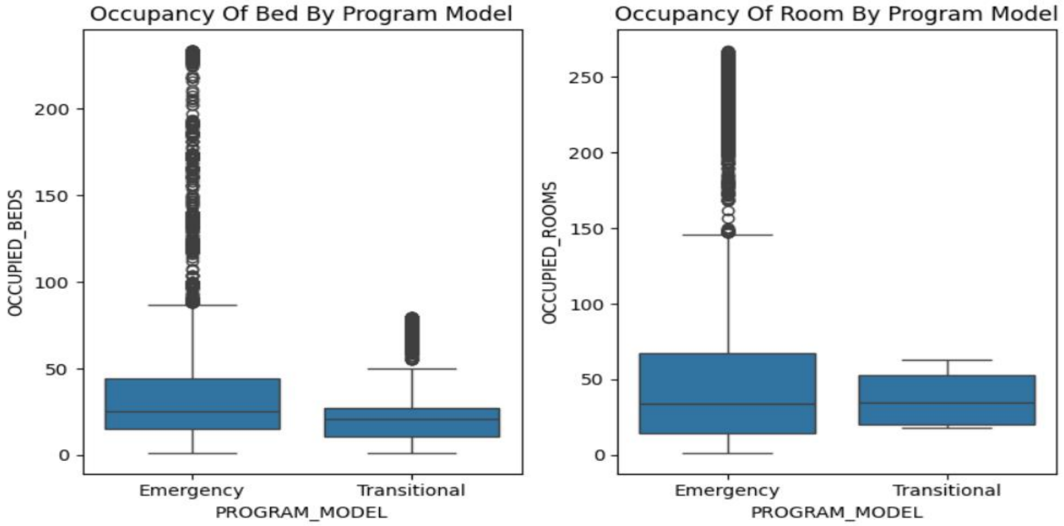


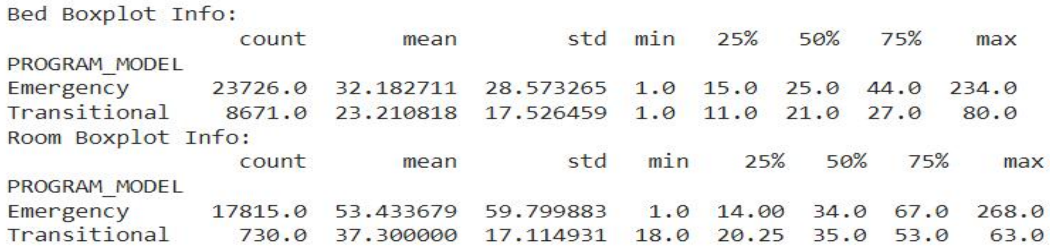
From the plots, it shows that both bed-based and room-based shelters have higher counts for the Emergency model compared to the Transitional model. Afterward, I have created a boxplot to illustrate the differences between these two program models concerning bed-based and room-based capacity and occupancy with statistical information. This analysis aims to provide insights and understanding the shape of the distribution of program model as the figure show below:





```
Bed Boxplot Info:
                   count       mean        std  min   25%   50%   75%    max
PROGRAM_MODEL
Emergency        23726.0  33.833516  29.390146  1.0  16.0  25.0  45.0  234.0
Transitional      8671.0  25.593703  18.342901  2.0  12.0  24.0  29.0   83.0
Room Boxplot Info:
                   count       mean        std   min   25%   50%   75%    max
PROGRAM_MODEL
Emergency        17815.0  56.010721  60.503601   1.0  18.0  35.0  71.0  268.0
Transitional       730.0  44.287671  17.748494  23.0  27.0  42.5  63.0   63.0
```

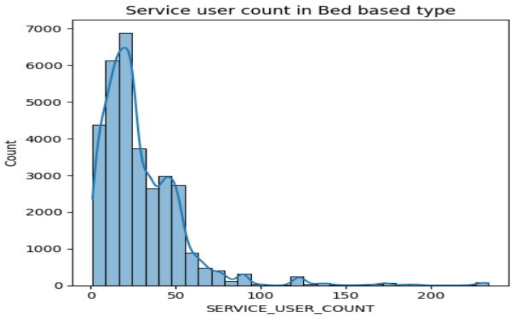


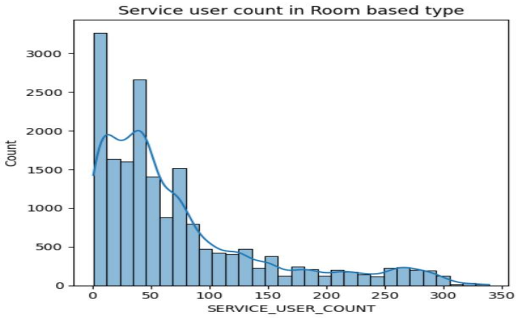


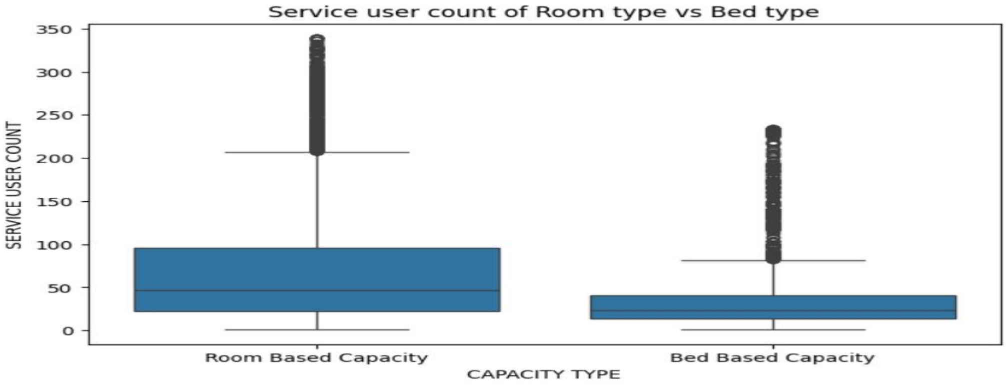
```
Bed Boxplot Info:
                   count       mean        std  min    25%   50%   75%    max
PROGRAM_MODEL
Emergency        23726.0  32.182711  28.573265  1.0   15.0  25.0  44.0  234.0
Transitional      8671.0  23.210818  17.526459  1.0   11.0  21.0  27.0   80.0
Room Boxplot Info:
                   count       mean        std   min    25%   50%   75%    max
PROGRAM_MODEL
Emergency        17815.0  53.433679  59.799883   1.0  14.00  34.0  67.0  268.0
Transitional       730.0  37.300000  17.114931  18.0  20.25  35.0  53.0   63.0
```

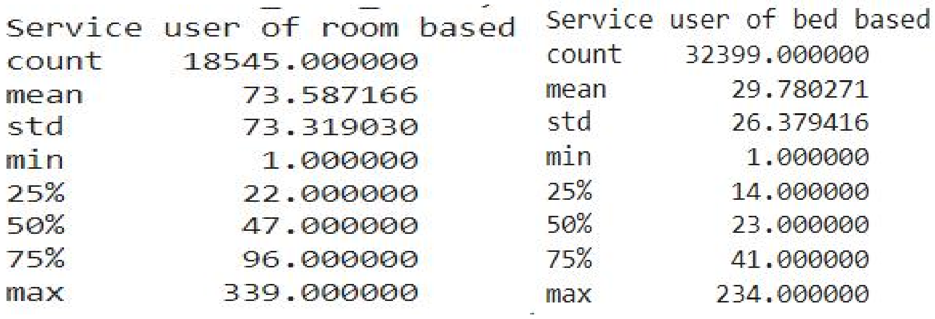
From the statistic, we can see that in both bed-based and room-based shelters, the Emergency program model demonstrates higher average occupancy and capacity compared to the Transitional model. These statistics underscore a consistent trend of higher occupancy and capacity in Emergency programs across both shelter types.

Now, I move over to examine the distribution of service user count across different capacity types by plotting both histograms and boxplots as the figures show below:









```
Service user of room based      Service user of bed based
count     18545.000000          count    32399.000000
mean         73.587166          mean        29.780271
std          73.319030          std         26.379416
min           1.000000          min          1.000000
25%          22.000000          25%         14.000000
50%          47.000000          50%         23.000000
75%          96.000000          75%         41.000000
max         339.000000          max        234.000000
```

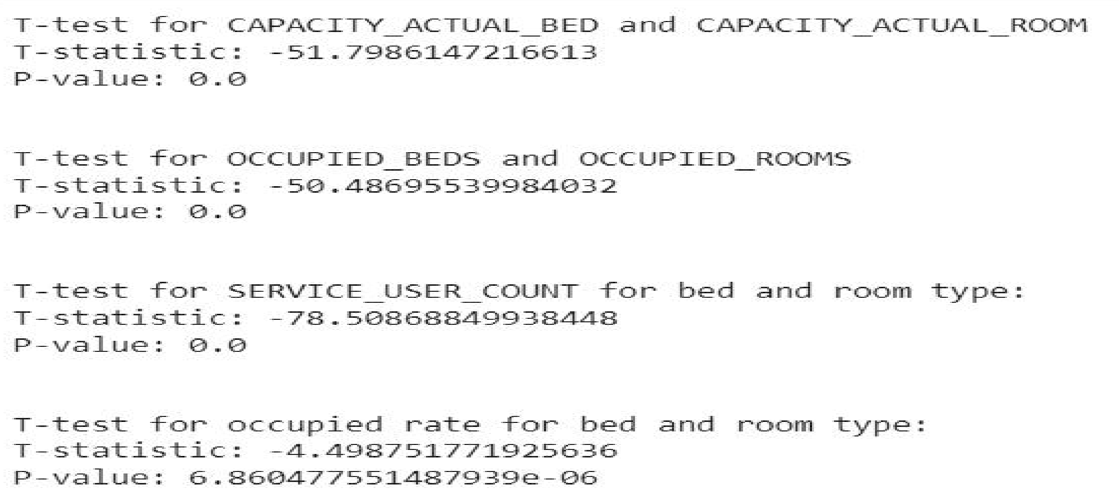
From the histogram, it shows that the service user count is right-skewed for both types indicating there is a lot outliers. From boxplot information, bed-based shelters displaying higher service user counts compared to room-based shelters. Interestingly, the room-based type has a higher mean service user count of 73, whereas the bed-based type has a lower mean of 29.

T-Test

Exploratory data analysis has been completed, and now we will conduct t-tests to determine if there is a significant difference between Bed-based and Room-based types. We will perform t-tests on occupancy, capacity, service user count, and occupancy rate.

Hypothesis:
H0: There is no significant difference in (Occupancy, Capacity, Service User Count, Occupancy Rate.) means between Bed-based and Room-based types.
H1: There is a significant difference in (Occupancy, Capacity, Service User Count, Occupancy Rate.)means between Bed-based and Room-based types.

T test Result:



```
T-test for CAPACITY_ACTUAL_BED and CAPACITY_ACTUAL_ROOM
T-statistic: -51.7986147216613
P-value: 0.0


T-test for OCCUPIED_BEDS and OCCUPIED_ROOMS
T-statistic: -50.48695539984032
P-value: 0.0


T-test for SERVICE_USER_COUNT for bed and room type:
T-statistic: -78.50868849938448
P-value: 0.0


T-test for occupied rate for bed and room type:
T-statistic: -4.498751771925636
P-value: 6.860477551487939e-06
```

For CAPACITY_ACTUAL_BED and CAPACITY_ACTUAL_ROOM, the t-statistic is -51.8 and the p-value is $0 < 0.05$ indicates that we reject the null hypothesis and conclude that there is a significant difference in capacity means between Bed-based and Room-based types.

For OCCUPIED_BEDS and OCCUPIED_ROOMS, the t-statistic is -50.5 and the p-value is $0 < 0.05$ indicates that we reject the null hypothesis and conclude that there is a significant difference in occupancy means between Bed-based and Room-based types.

For SERVICE_USER_COUNT, the t-statistic is -78.5 and the p-value is $0 < 0.05$ indicates that we reject the null hypothesis and conclude that there is a significant difference in service user count means between Bed-based and Room-based types.

For occupancy rate, the t-statistic is -4.5 and the p-value is 6.86e-06 < 0.05 indicates that we reject the null hypothesis and conclude that there is a significant difference in occupancy rate means between Bed-based and Room-based types.

Conclusion:
The results highlights the urgent need to develop strategies in managing bed-based and room-based shelters to adequately address the diverse needs of homeless people. The significantly higher capacity and occupancy rates observed in room-based shelters suggest that certain populations may prefer such facilities or be better suited for longer-term stays. Understanding these differences can perform better resource allocation to optimize shelter services and supports for people experiencing homelessness.