2021 Toronto Shelter Service Data Analysis

INF2178 Assignment 1 write up
Chenyang Pan
1005131554

Nowadays, although many shelter services are provided by different programs and institutions for homeless people in the city of Toronto, there are still unhoused and homeless people who do not have a chance to have an overnight shelter in the city due to the increased homeless people in Toronto. The goal of this study is to perform an in depth exploratory data analysis toward the data, INF2178_A1_data.xlsx, which contains the information of capacity and occupancy of overnight shelter programs of 2021 in Toronto, and provide suggestion for future overnight shelter institutions accordingly in order to let more unhoused and homeless people have opportunity to have overnight shelter.

In the data preprocessing section, 7 features of our interest are selected and abstract from 14 features. Out of these 7 features, “PROGRAM_MODEL”, the type of shelter program, and “CAPACITY_TYPE”, the type of shelter capacity, are both binary categorical variables, where the programs are in two types, Emergency and Transitional, and can also be divided into bed based capacity shelter and room based capacity shelter. “SERVICE_USER_COUNT”, “CAPACITY_ACTUAL_ROOM”, “OCCUPIED_ROOMS”, “CAPACITY_ACTUAL_BED”, “OCCUPIED_BEDS”, are numerical variables where “SERVICE_USER_COUNT” is the number of user in the shelter program, “CAPACITY_ACTUAL_ROOM” and “OCCUPIED_ROOMS” are the room capacity and number of room occupied for room based capacity shelters, “CAPACITY_ACTUAL_BED” and “OCCUPIED_BEDS” are the bed capacity and number of beds occupied for bed based capacity shelters. I first had an overview of the entire dataset, and decided to divide the data set into two separate data sets by the type of capacity they are, room or bed. Then I calculated and added the occupancy rate, calculated as occupied room or bed divided by capacity of room or bed, for both separated room data and separated bed data. Doing this, I can look at each capacity type of shelter individually and make comparison easier.

In the first step of Exploratory data analysis I made a table using the aggregated data to see the number of user service count in program mode type emergency and transitional separately(Shown in Table 1), and calculated the proportion of each type model and found 81.55% are Emergency and 18.45% is transitional. So we know the majority shelter program in the city is Emergency.

Table1

PROGRAM_MODEL	Count
Emergency	41541
Transitional	9401

Then, another similar table can be made based on capacity type to observe data distribution shown in Table 2, and also calculated proportion of each type; 63.6% for bed based and 36.4% for room based. This tells us that the number of bed based capacity programs is more than room based.

Table1

CAPACITY_TYPE	Count
Bed based capacity	32399
Room based capacity	18545

Next, I made the summary statistics including, mean, median, min, max, first and third quartile, IQR, spread and standard deviation for each numerical feature in both room based capacity data and bed based capacity data and which can be summarized in Table 3 and table 4 shown below. What can be found from the following summary is that, the mean and median of service users in room shelters is much larger than the number of service users in bed shelters, which implies that room shelters may have larger space to accept more people. Also the IQR, standard deviation of actual capacity room and occupied room in room shelter are all greater than actual capacity bed and occupied beds in bed shelters. Surprisingly, the IQR of occupancy rate (bed) is larger than occupancy rate room, but its spread is smaller than room. This may imply the occupancy rate (room) has some outliers. Also, one last things is that even though the median for both occupancy rate are 1 but occupancy rate (room) has slight greater mean than occupancy bed which may imply people may prefer room based shelter

Table 3 (room based capacity)

	Service user count (room)	Capacity room	Occupied room	Occupancy rate (room)	Service user count (room)	Capacity bed	Occupied beds	Occupancy rate (bed)
min	1	1	1	0.012	1	1	1	0.023
mean	73.587	55.549	52.799	0.934	29.781	31.628	29.781	0.928
max	339	268	268	1.014	234	234	234	1
spread	338	267	267	1.002	233	233	233	0.977
25th percentile	22.0	19	16	0.958	14	15	14	0.9
median	47	35	34	1	23	25	23	1
Standard deviation	73.319	59.449	58.793	0.163	26.38	27.128	26.38	0.123
75th	96	68	66	1.0	41	43	41	1.0

percentile								
IQR	74	49	50	0.042	27	28	27	0.1

Then, I made two bar plots to make a more apparent visual comparison of different program models and capacity type. Shown as Figure 1 and 2 below. The barplot of the program model and the barplot of the capacity type once again shows the huge visual difference of number between each type of program model.

Figure 1

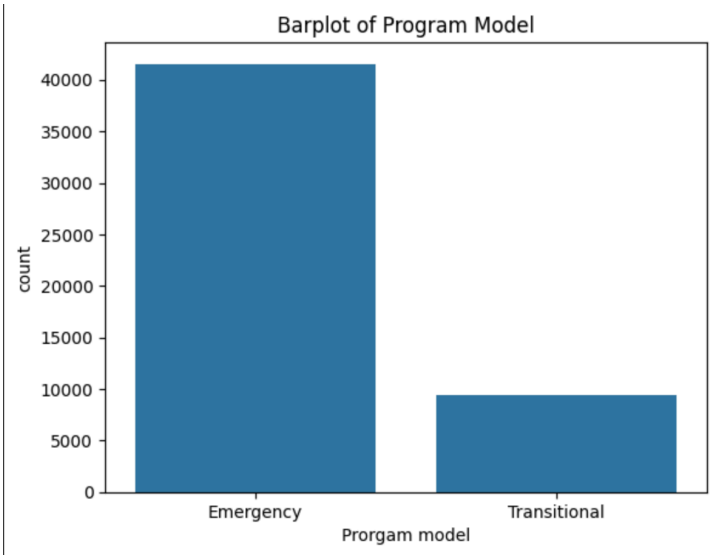
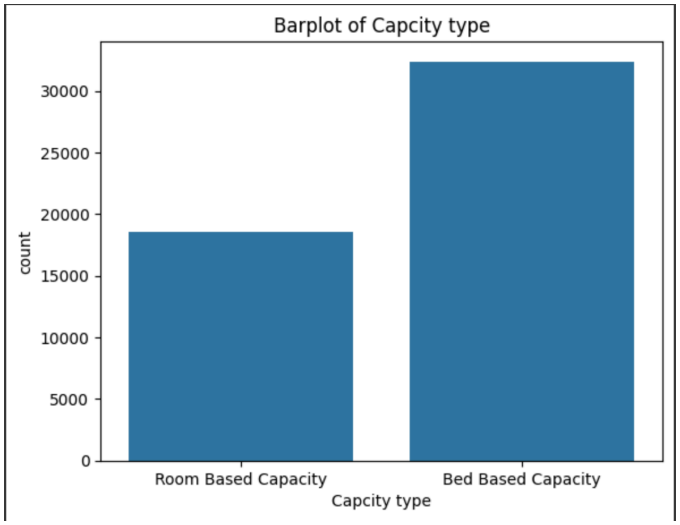


Figure 2



Next, I made two histograms, Figure 3 and 4, where Figure 3 shows the distribution of occupancy rate (room) and Figure 4 shows distribution of occupancy rate (bed). Both histograms are unimodal and left skewed. The two histograms give us some insight that the occupancy rate for both room and bed based capacity shelters are mostly 1 or close to 1.

Figure 3

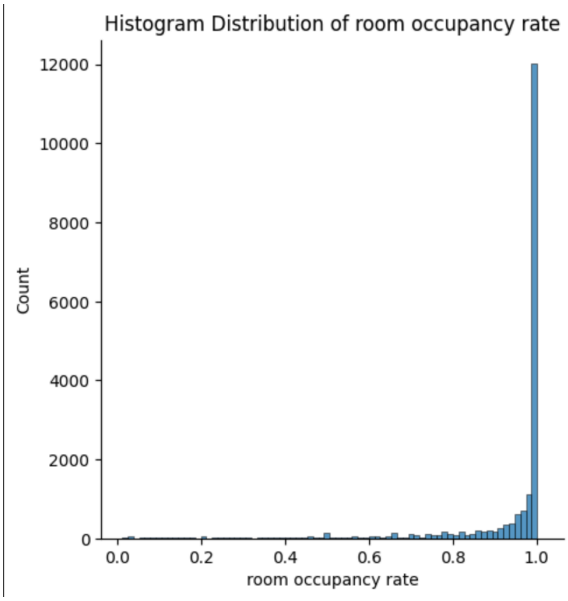
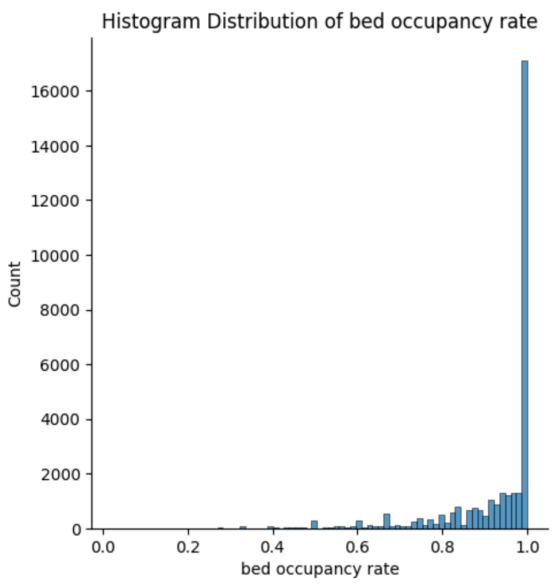


Figure 4



The next two histograms show the distribution of service user count for room based capacity and bed based capacity. These graphs are made for comparison. As shown in Figure 5 and 6. Both histograms are right skewed and the histogram of service users count (room) is more spread out. These two histograms clearly show the room based shelter accepts more people.

Figure 5

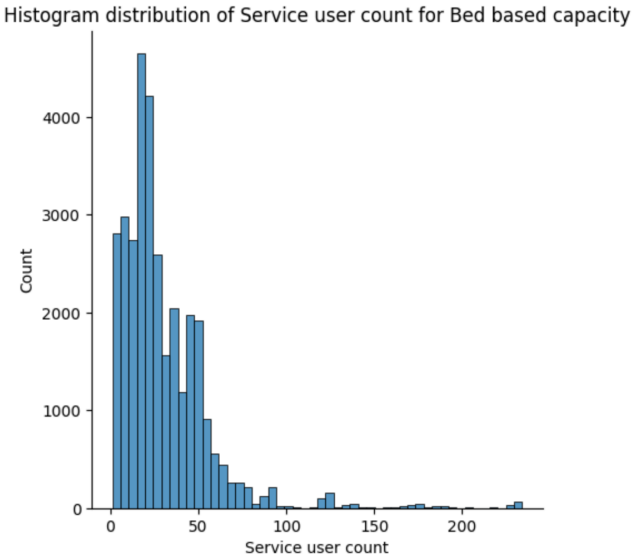
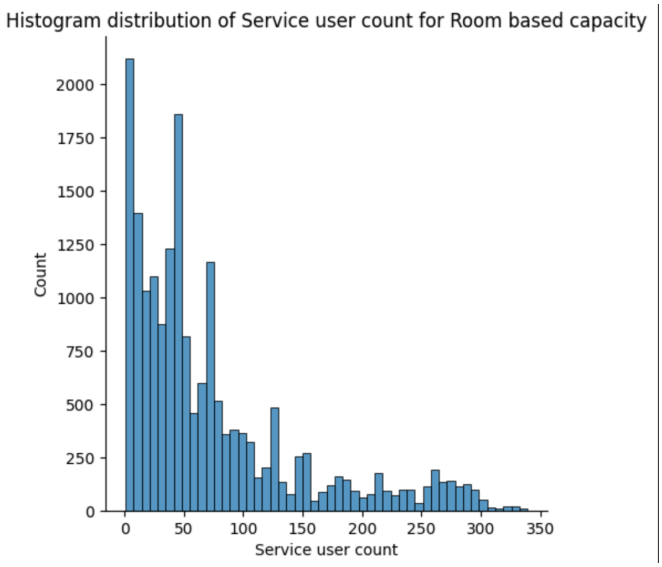
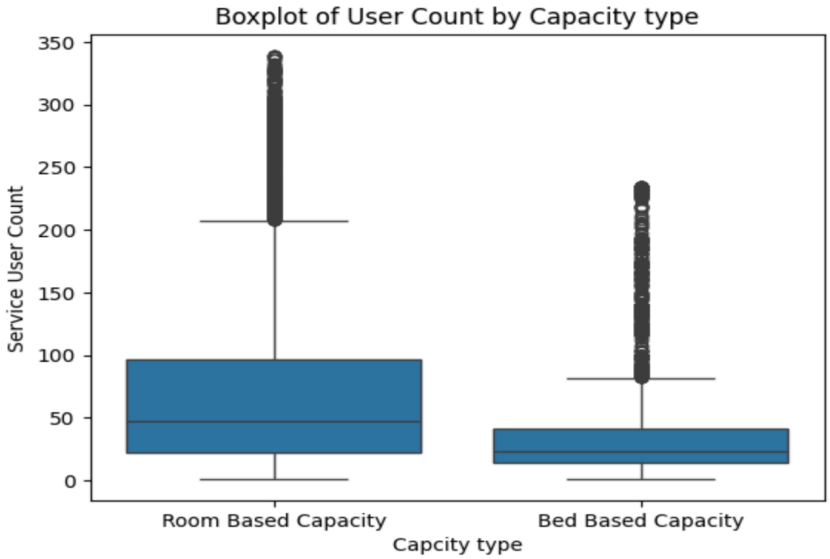


Figure 6



I also made a side by side boxplot of service user count by capacity type, as shown below in Figure 7, we can detect many outliers for user count of both capacity types. The boxplots show visually the IQR and median user of the room based capacity is larger than the bed based capacity. One thing to notice is that , since there are too many outliers, we may consider not removing these outliers since removing a large amount of data may affect the quality of the overall data.

Figure 7



Finally, I also made two scatterplots, shown below in Figure 8 and Figure 9. The first scatterplot is service user count of room based shelter vs the actual capacity of room, and the second plot is service user count of bed based shelter vs the actual capacity of bed. Two strong positive correlations can be detected from two scatterplot, which can tell us that the homeless are willing to stay in the city shelters when there is enough space.

Figure 8

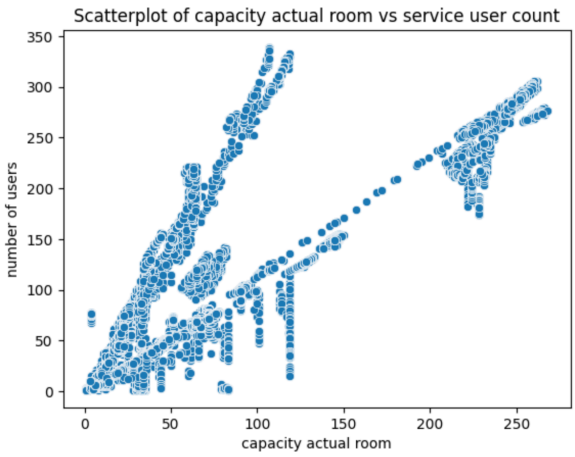
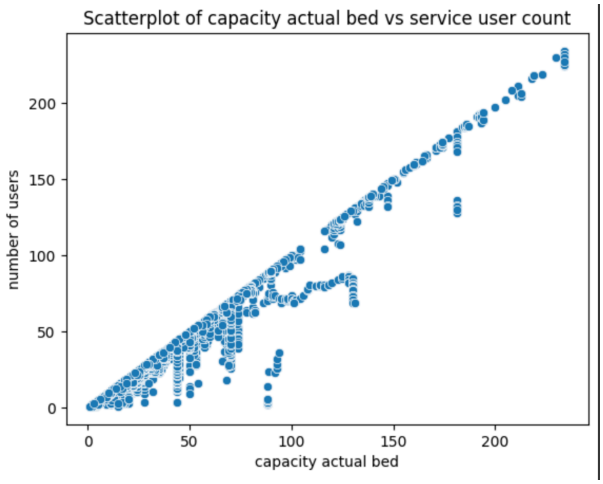


Figure 9



After the EDA, we have found some important implications from the data, which is that, it seems room based capacity shelters can accept more homeless people and people seem to prefer it over bed based shelter (based on comparing occupancy rate). However, there are more bed based capacity shelters in the city than room based capacity shelters.

I have also conducted two t-tests. The first t-test is a one sided Welch's T-test. THe null hypothesis is that the average service user of room type capacity and bed type capacity are the same and the alternative hypothesis is that the average users of room type capacity is greater than bed type capacity users. I chose Welch's t-test since the dataset does not have equal variance. The t statistic is 78.5 and the p value equals to 0. Since the p value is much smaller to the significance level of 0.05, we reject the null hypothesis.

The second t-test is a two sample one sided t-test. The null hypothesis is that the occupancy rate of two types of capacities are the same and the alternative hypothesis is that the occupancy rate of room based capacity type is greater than bed type. The calculated t statistic is 4.85 and the p value is 6.32e-7 which is again much more smaller than the significant level of 0.05, so that we reject the null hypothesis

The results of the two t-tests are expected since they match what we have observed from EDA. Thus, we can draw a conclusion that, in order to let more homeless people be accepted in the overnight shelters, I suggested, more room based capacity shelters should be built since the room based capacity shelters can accept more people and people prefer it over bed based capacity shelters.