Student Name: Sheng Zhang
Student Number: 1006842543
Email: eily.zhang@mail.utoronto.ca

Assignment 1: Narrative

The columns that I am interested in analysing are CAPACITY_TYPE(the type of capacity of the shelter), PROGRAM_MODEL(the program model of the shelter), SERVICE_USER_COUNT(the number of accepted homeless people of the day), CAPACITY_ACTUAL_BED(the number of available beds), OCCUPIED_BEDS(the number of occupied beds), CAPACITY_ACTUAL_ROOM(the number of available rooms), and OCCUPIED_ROOMS(the number of occupied rooms) where the groups are PROGEAM_MODEL and all other columns are the attributes for these two groups. The adjusted data frame is:

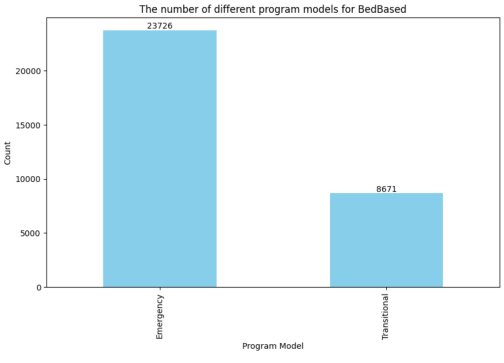
	CAPACITY_TYPE	PROGRAM_MODEL	SERVICE_USER_COUNT	CAPACITY_ACTUAL_BED	OCCUPIED_BEDS	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS
0	Room Based Capacity	Emergency	74	NaN	NaN	29.0	26.0
1	Room Based Capacity	Emergency	3	NaN	NaN	3.0	3.0
2	Room Based Capacity	Emergency	24	NaN	NaN	28.0	23.0
3	Room Based Capacity	Emergency	25	NaN	NaN	17.0	17.0
4	Room Based Capacity	Emergency	13	NaN	NaN	14.0	13.0
...	...	...	...	...	...	...	...
50939	Bed Based Capacity	Emergency	6	20.0	6.0	NaN	NaN
50940	Bed Based Capacity	Emergency	23	23.0	23.0	NaN	NaN
50941	Bed Based Capacity	Transitional	13	14.0	13.0	NaN	NaN
50942	Bed Based Capacity	Emergency	10	10.0	10.0	NaN	NaN
50943	Bed Based Capacity	Transitional	29	29.0	29.0	NaN	NaN

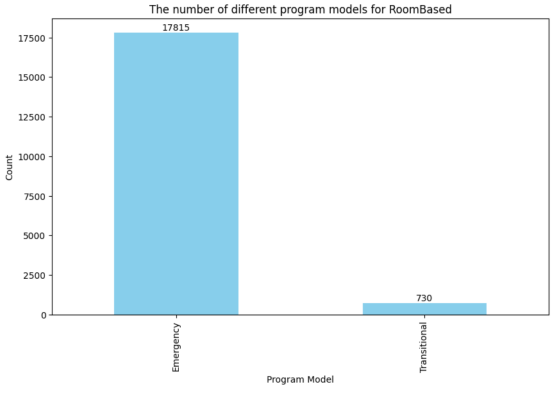
Because there are two types of capacity, so I create two separate data frames for each capacity type:

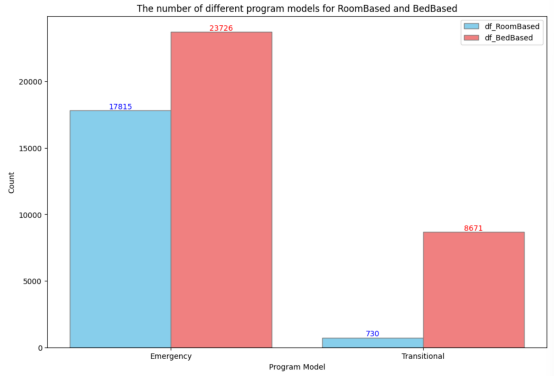
	CAPACITY_TYPE	PROGRAM_MODEL	SERVICE_USER_COUNT	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS
0	Room Based Capacity	Emergency	74	29.0	26.0
1	Room Based Capacity	Emergency	3	3.0	3.0
2	Room Based Capacity	Emergency	24	28.0	23.0
3	Room Based Capacity	Emergency	25	17.0	17.0
4	Room Based Capacity	Emergency	13	14.0	13.0
...	...	...	...	...	...
50920	Room Based Capacity	Emergency	128	128.0	128.0
50923	Room Based Capacity	Emergency	76	76.0	76.0
50927	Room Based Capacity	Emergency	10	3.0	3.0
50932	Room Based Capacity	Emergency	74	23.0	22.0
50934	Room Based Capacity	Emergency	27	28.0	27.0

	CAPACITY_TYPE	PROGRAM_MODEL	SERVICE_USER_COUNT	CAPACITY_ACTUAL_BED	OCCUPIED_BEDS
5	Bed Based Capacity	Emergency	6	8.0	6.0
10	Bed Based Capacity	Emergency	22	24.0	22.0
11	Bed Based Capacity	Emergency	8	12.0	8.0
21	Bed Based Capacity	Transitional	10	12.0	10.0
25	Bed Based Capacity	Emergency	11	12.0	11.0
...	...	...	...	...	...
50939	Bed Based Capacity	Emergency	6	20.0	6.0
50940	Bed Based Capacity	Emergency	23	23.0	23.0
50941	Bed Based Capacity	Transitional	13	14.0	13.0
50942	Bed Based Capacity	Emergency	10	10.0	10.0
50943	Bed Based Capacity	Transitional	29	29.0	29.0

Then I want to investigate which program model is more prevalent as a starter, so I have the three bar graphs, one for RoomBased, one for BedBased, and one more cross-comparison of the two capacity type:







So far, I see that both BedBased and RoomBased shelters have more emergency models compared to transitional models, and the BedBased capacity type has a higher number in both models compared to the RoomBased capacity type. As such, the only sign from the bar graph above is that the emergency model is a more prevalent and preferred type of shelter type for both the program runner and the homeless. However, I still need to dig into the other numerical variables other than this categorical variable.

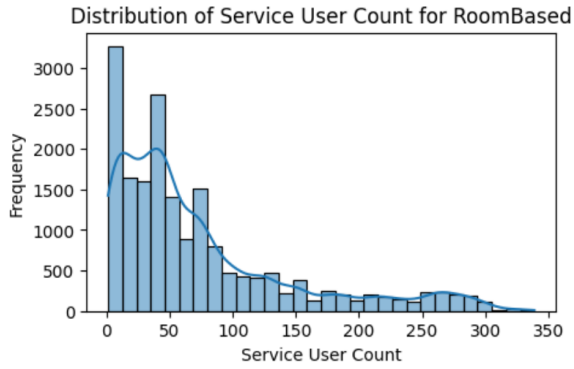
Now, the research question in my mind is whether the homeless tend to choose the shelters by capacity type, that is, whether capacity type is a factor that influences the homeless' choice of staying.

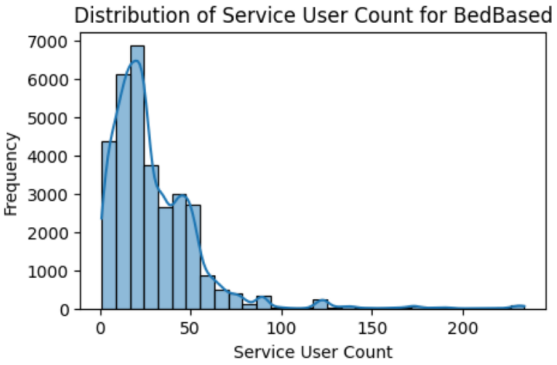
First, I have the summary statistics tables for RoomBased (Figure on the left), and BedBased(Figure on the right) to have a brief understanding about the numeric data. From the tables we can see that for all three variables SERVICE_USER_COUNT, CAPACITY_ACTUAL, and OCCUPIED, their variances are not equal when we compare the same variable between RoomBased and BedBased, which implies that only Welch's t-test can be applied here if I want to proceed such statistical tests.

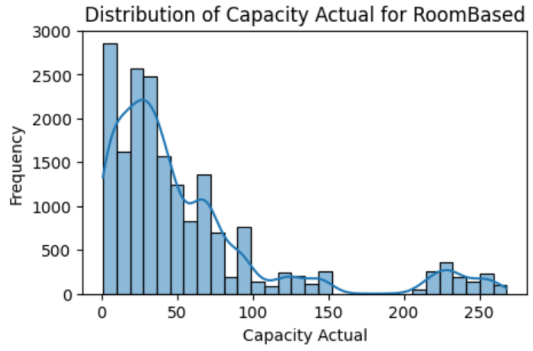
Summary Statistics for RoomBased	SERVICE_USER_COUNT	CAPACITY_ACTUAL	OCCUPIED
count	18545.000000	18545.000000	18545.000000
mean	73.587166	55.549259	52.798598
std	73.319030	59.448805	58.792954
min	1.000000	1.000000	1.000000
25%	22.000000	19.000000	16.000000
50%	47.000000	35.000000	34.000000
75%	96.000000	68.000000	66.000000
max	339.000000	268.000000	268.000000
variance	5375.680185	3534.160443	3456.611452

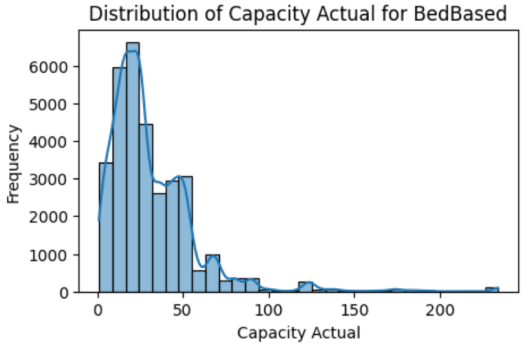
Summary Statistics for BedBased	SERVICE_USER_COUNT	CAPACITY_ACTUAL	OCCUPIED
count	32399.000000	32399.000000	32399.000000
mean	29.780271	31.627149	29.780271
std	26.379416	27.127682	26.379416
min	1.000000	1.000000	1.000000
25%	14.000000	15.000000	14.000000
50%	23.000000	25.000000	23.000000
75%	41.000000	43.000000	41.000000
max	234.000000	234.000000	234.000000
variance	695.873596	735.911104	695.873596

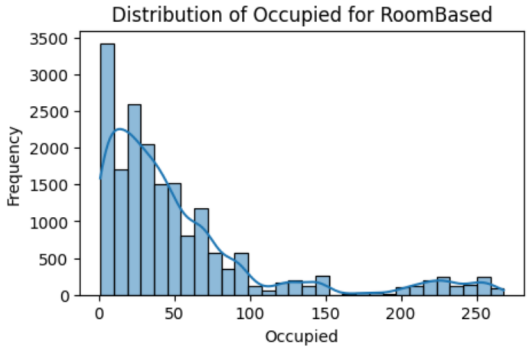
Then for both RoomBased and BedBased, I want to plot the distribution graph (histogram) for all three variables SERVICE_USER_COUNT, CAPACITY_ACTUAL, and OCCUPIED to check the curves that describe the distribution of their data. Also, I want to plot the multi-variable box plots that cross-compare BedBased and RoomBased for the three variables.

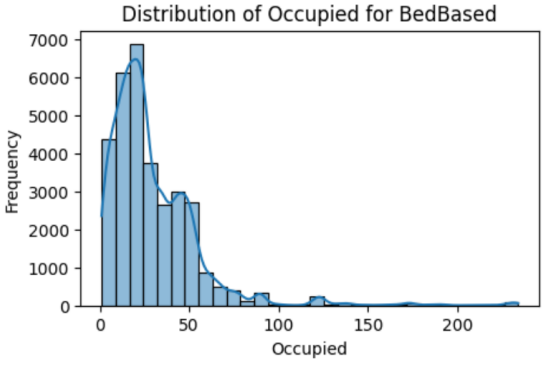

Distribution of Service User Count for RoomBased
Frequency
Service User Count

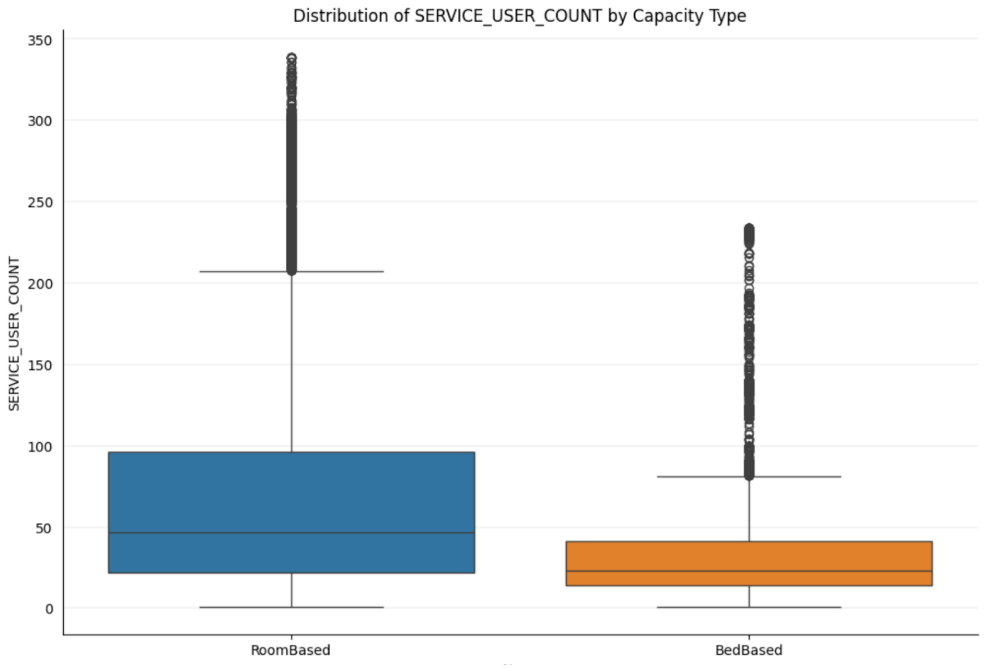

Distribution of Service User Count for BedBased
Frequency
Service User Count

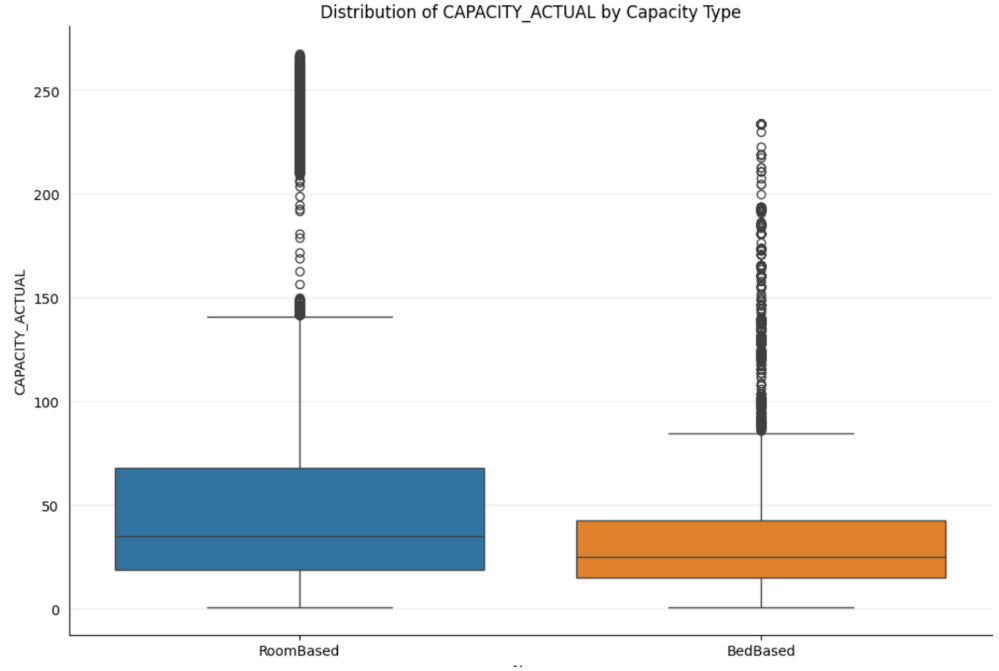

Distribution of Capacity Actual for RoomBased
Frequency
Capacity Actual

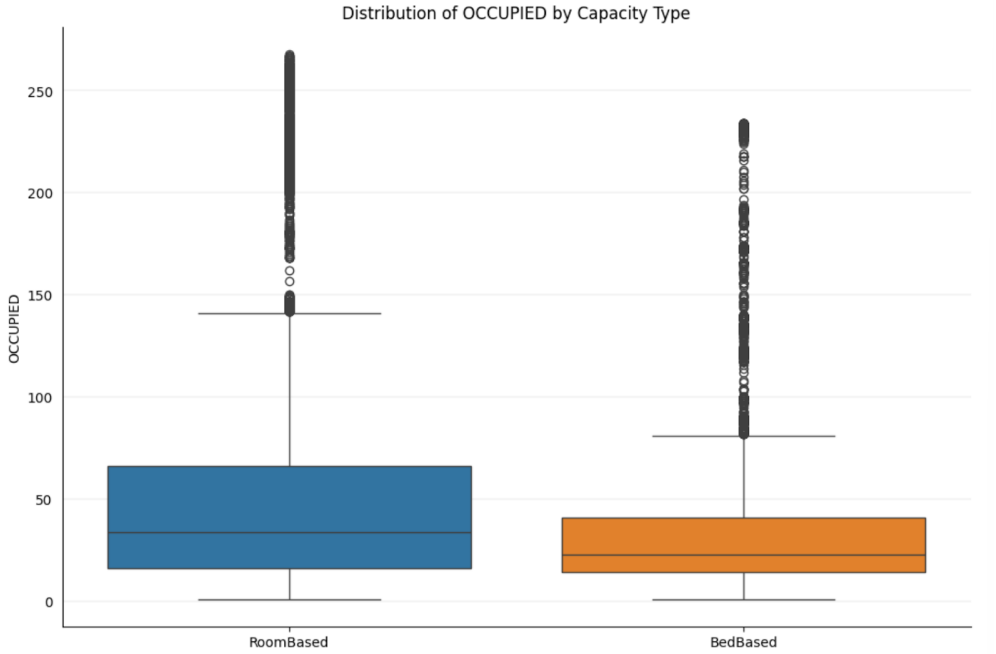

Distribution of Capacity Actual for BedBased
Frequency
Capacity Actual


Distribution of Occupied for RoomBased
Frequency
Occupied


Distribution of Occupied for BedBased
Frequency
Occupied


Distribution of SERVICE_USER_COUNT by Capacity Type
SERVICE_USER_COUNT
RoomBased
BedBased





From the histograms I can see that all of the six histograms are skewed and from the box plots we can also see that there are many data points that greater than the 75% quartile, so there are many data points with high values causing the skewed distribution and making them not so close to normal distributions.

So the next step is to perform t-test, since the variances for all three variables are not equal in BedBased and RoomBased as shown in the summary statistics tables previously. Thus I can only perform Welch's t-test here as it is designed for performing the unequal variable between-subject independent t-test.

After performing the Welch's t-test, I have the following results:

```
Welch's t-statistic for SERVICE_USER_COUNT = 78.50868849938448
p-value for SERVICE_USER_COUNT = 0.0


Welch's t-statistic for CAPACITY_ACTUAL = 51.7986147216613
p-value for CAPACITY_ACTUAL = 0.0


Welch's t-statistic for OCCUPIED = 50.48695539984032
p-value for OCCUPIED = 0.0
```

The high t-statistics and the low p-values suggest that there is a significant difference in the variables (SERVICE_USER_COUNT, CAPACITY_ACTUAL, and OCCUPIED) between these two capacity types. Specifically, SERVICE_USER_COUNT that shows the number of accepted homeless people of that day, CAPACITY_ACTUAL that shows the available beds/rooms in the shelters, and OCCUPIED that shows the number of occupied beds/rooms are all related to the capacity types. In other words, the capacity type is a factor that influences the homeless' choice of staying in shelters.

As a discussion, the further shelter programs can pay attention to the specific capacity type before they get started so that the government or the charity can make sure that the shelters are not “under-occupied” or “over-occupied” and to maximise the utilisation rate of the shelter programs to provide the homeless people with a warm and comfortable place to stay.