Introduction:
This report conducts an exploratory data analysis (EDA) on a dataset outlining Toronto's shelter usage patterns. The goal is to discover any statistical difference in the occupied rate, and contribute to enhancement of Toronto's shelter support system.

Data Selection
I will examine shelter usage trends by using a dataset titled INF2178_A1_data.xlsx. ('CAPACITY_TYPE', 'PROGRAM_MODEL', 'SERVICE_USER_COUNT', 'CAPACITY_ACTUAL_BED', 'OCCUPIED_BEDS', 'CAPACITY_ACTUAL_ROOM', 'OCCUPIED_ROOMS') will be the columns that I select to achieve my data analysis goals.

	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
...	...	...	...	...	...	...	...
939	Bed Based Capacity	Emergency	6	20.0	6.0	NaN	NaN
940	Bed Based Capacity	Emergency	23	23.0	23.0	NaN	NaN
941	Bed Based Capacity	Transitional	13	14.0	13.0	NaN	NaN
942	Bed Based Capacity	Emergency	10	10.0	10.0	NaN	NaN
943	Bed Based Capacity	Transitional	29	29.0	29.0	NaN	NaN

After loading the data, it shows there are some missing values, and I discovered the total amout of the missing value in each columns: In order to deal with these missing values, data cleaning is necessary.

```
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
```

Data Cleaning

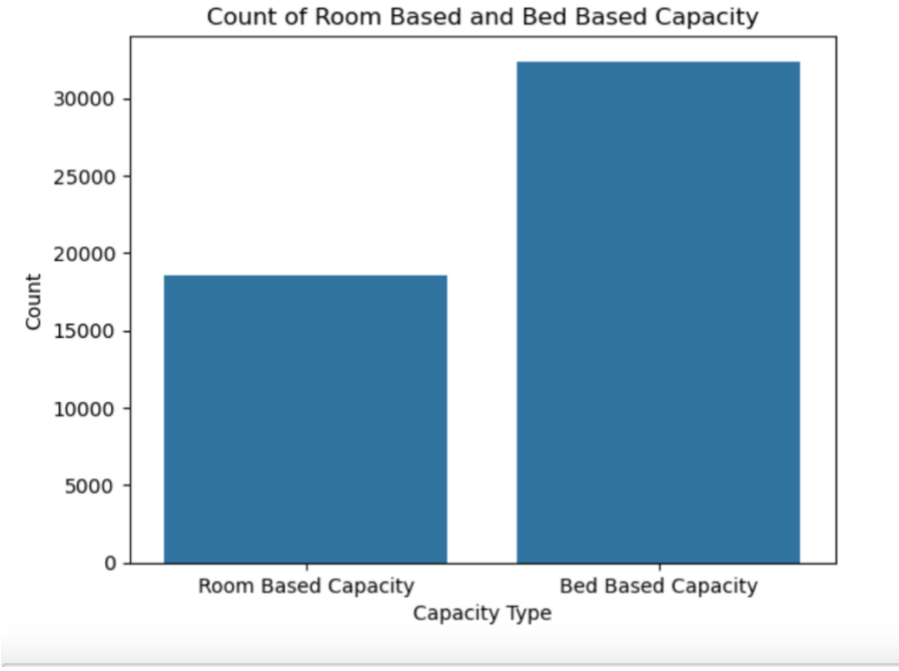
- I have set the capacity actual bed and occupied bed to 0 value for the room based capacity type, also the capacity actual room and occupied room to 0 value for bed based capacity type.
- Fills missing values in the 'PROGRAM_MODEL' column with the most frequently occurring value-model in that column.

	CAPACITY_TYPE	PROGRAM_MODEL	SERVICE_USER_COUNT	CAPACITY_ACTUAL_BED	OCCUPIED_BEDS	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS
0	Room Based Capacity	Emergency	74	0.0	0.0	29.0	26.0
1	Room Based Capacity	Emergency	3	0.0	0.0	3.0	3.0
2	Room Based Capacity	Emergency	24	0.0	0.0	28.0	23.0
3	Room Based Capacity	Emergency	25	0.0	0.0	17.0	17.0
4	Room Based Capacity	Emergency	13	0.0	0.0	14.0	13.0
...	...	...	...	...	...	...	...
939	Bed Based Capacity	Emergency	6	20.0	6.0	0.0	0.0
940	Bed Based Capacity	Emergency	23	23.0	23.0	0.0	0.0
941	Bed Based Capacity	Transitional	13	14.0	13.0	0.0	0.0
942	Bed Based Capacity	Emergency	10	10.0	10.0	0.0	0.0
943	Bed Based Capacity	Transitional	29	29.0	29.0	0.0	0.0

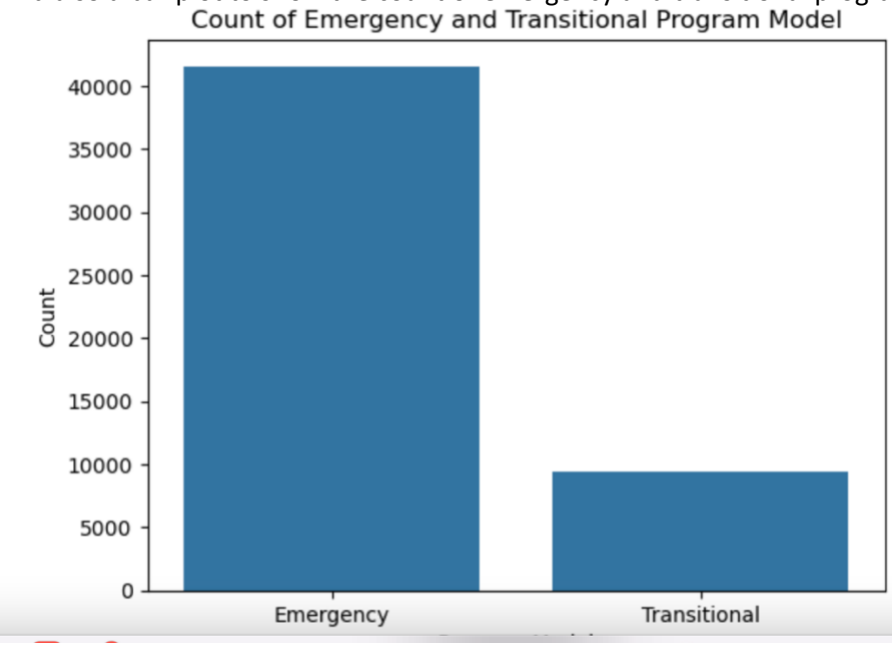
Now the data has been cleaned and It is the time to do the EDA.

EDA

First, I created a bar plot to show the count of room based capacity and bed based capacity type. Different capacity types show a significant difference, it leads my interest to discover the occupied rate for the capacity types.



And also a bar plot to show the count of emergency and transitional program model:



I created a column called OCCUPIED_BED_RATE to show the occupied rate for bed based capacity. And show the summary statistics :

CAPACITY_TYPE	PROGRAM_MODEL	SERVICE_USER_COUNT	CAPACITY_ACTUAL_BED	OCCUPIED_BEDS	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS	OCCUPIED_BED_RATE
Bed Based Capacity	Emergency	6	8.0	6.0	0.0	0.0	0.750000
Bed Based Capacity	Emergency	22	24.0	22.0	0.0	0.0	0.916667
Bed Based Capacity	Emergency	8	12.0	8.0	0.0	0.0	0.666667
Bed Based Capacity	Transitional	10	12.0	10.0	0.0	0.0	0.833333
Bed Based Capacity	Emergency	11	12.0	11.0	0.0	0.0	0.916667
...	...	...	...	...	...	...	...
Bed Based Capacity	Emergency	6	20.0	6.0	0.0	0.0	0.300000

```
bed based capacity 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
```

I also created a column called OCCUPIED_ROOM_RATE to show the occupied rate for room based capacity. And show the summary statistics:

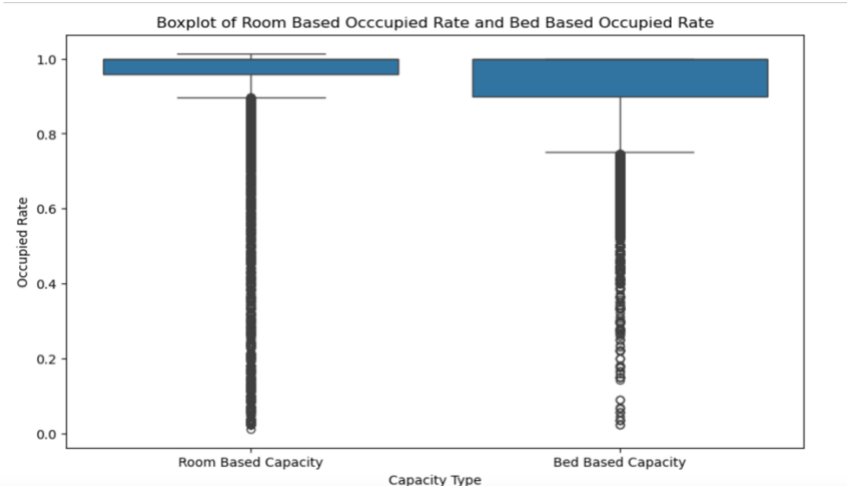
APACITY_TYPE	PROGRAM_MODEL	SERVICE_USER_COUNT	CAPACITY_ACTUAL_BED	OCCUPIED_BEDS	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS	OCCUPIED_ROOM_RATE
Room Based Capacity	Emergency	74	0.0	0.0	29.0	26.0	0.896552
Room Based Capacity	Emergency	3	0.0	0.0	3.0	3.0	1.000000
Room Based Capacity	Emergency	24	0.0	0.0	28.0	23.0	0.821429
Room Based Capacity	Emergency	25	0.0	0.0	17.0	17.0	1.000000
Room Based Capacity	Emergency	13	0.0	0.0	14.0	13.0	0.928571
...	...	...	...	...	...	...	...
Room Based Capacity	Emergency	128	0.0	0.0	128.0	128.0	1.000000

```
room based capacity 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
```

After showing occupied rate in room based and bed based capacity by separate columns. Now I created one column called OCCUPIED_RATE to show the occupied rate for bed based and room based capacity type.

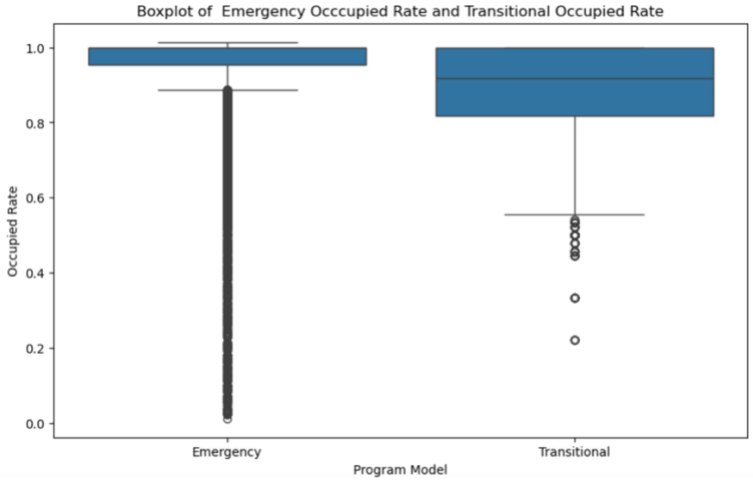
CAPACITY_TYPE	PROGRAM_MODEL	SERVICE_USER_COUNT	CAPACITY_ACTUAL_BED	OCCUPIED_BEDS	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS	OCCUPIED_RATE
Room Based Capacity	Emergency	74	0.0	0.0	29.0	26.0	0.896552
Room Based Capacity	Emergency	3	0.0	0.0	3.0	3.0	1.000000
Room Based Capacity	Emergency	24	0.0	0.0	28.0	23.0	0.821429
Room Based Capacity	Emergency	25	0.0	0.0	17.0	17.0	1.000000
Room Based Capacity	Emergency	13	0.0	0.0	14.0	13.0	0.928571
...	...	...	...	...	...	...	...
Bed Based Capacity	Emergency	6	20.0	6.0	0.0	0.0	0.300000
Bed Based Capacity	Emergency	23	23.0	23.0	0.0	0.0	1.000000

Now I am going to show boxplot of room based capacity occupied rate and bed based capacity occupied rate.



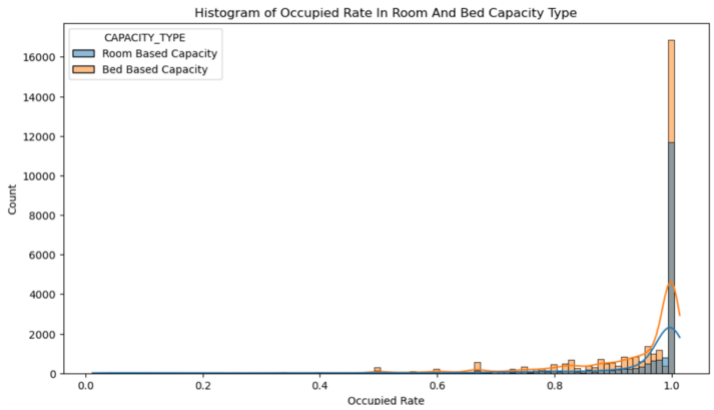
The boxplot illustrates room and bed occupancy. Rooms are displaying a narrow occupancy range and a high median. In contrast, beds have a broader range, occasionally being less utilized or showing variation in occupancy, as indicated by outliers.

Also the boxplot of occupied rate for different program model:



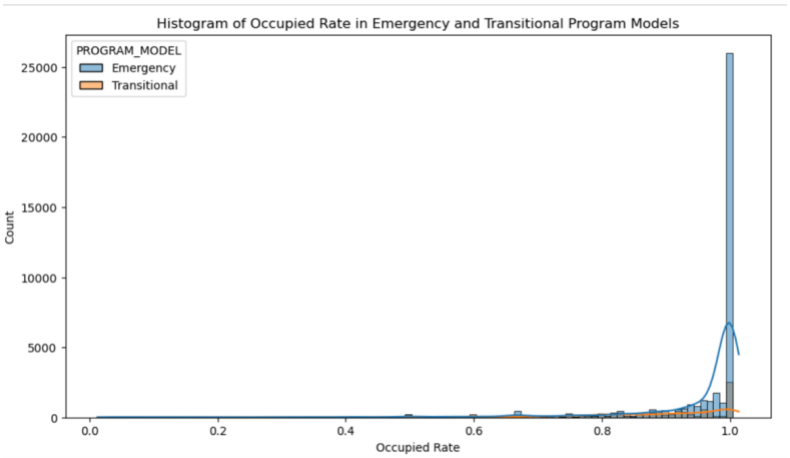
This boxplot shows the occupancy levels of emergency and transitional programs. Emergency programs consistently operate at full occupancy, indicating a continual high demand. In contrast, transitional programs exhibit a broader range of occupancy, with some below full occupancy.

The distribution of occupied rate in bed based and room based capacity:



Both histograms reveal high occupancy rates, particularly the room capacity histogram, which peaks at full occupancy, indicating frequent maximization of room resources. The bed capacity distribution displays a bit more variability in occupancy rates.

Also the distribution of occupied rate in emergency and transitional program model:



The graph reveals consistently high utilization of emergency and transitional occupied rates. Emergency model demonstrate a sharp peak at full occupancy, while transitional model display a high but less steep peak, suggesting diversity.

Now I come up the first interesting research question after the EDA process:

Research Question 1:

Is there a significant difference in occupied rate between room based capacity and bed based capacity in Toronto's shelter support program?

Null hypothesis : There is no significant difference in occupied rate between room based and bed based capacity.

Alternative hypothesis: There is a significant difference in occupied rate between room based and bed based capacity.

Result

I have performed two-sample t-test:

```
t-statistic = 4.854104599422829
p-value = 1.2128933183471424e-06
```

The t-statistic for the two-sample t-test is 4.8541, the positive t-statistics indicates two compared group has a significant difference, and the p value is really small, very close to zero and it's less than 0.05. There is strong evidence to reject the null hypothesis, as a result, there is a statistically significant difference in occupied rate between room based and bed based capacity.

I also performed the Welch's t-test:

```
t-statistic = 4.498751771925636
p-value = 6.860477551487939e-06
```

The t-statistic for the Welch's t-test is 4.4988, the positive sign of t-statistic supports that there is a significant difference in two compared groups, the p value is less than 0.05, it indicated that there is strong evidence to reject the null hypothesis, there is a statistically significant difference in occupied rate between room based and bed based capacity.

Research Question 2:
Is there a significant difference in occupied rate between emergency and transitional program model in Toronto's shelter support program?

Null hypothesis : There is no significant difference in occupied rate between emergency and transitional program model
Alternative hypothesis: There is a significant difference in occupied rate between emergency and transitional program model

Result
I have performed two-sample t-test:

```
t-statistic = 39.06876276218507
p-value = 0.0
```

The large positive value of t-statistics shows a significant difference between the two group that being compared. The very small value of p-value also indicated that the null hypothesis is rejected, therefore the occupied rate between emergency and transitional program model has the significant difference.

I also performed the Welch's t-test:

```
t-statistic = 40.97518639553636
p-value = 0.0
```

The t-statistic is 40.9752, the positive large value of t-statistic suggests a significant difference of occupied rate between two compared program model groups. Also the p-value indicated that the null hypothesis can be rejected, has strong evidence to support the alternative hypothesis.