# Exploratory Data Analysis on Shelter's Capacity

## Introduction:

This report presents a comprehensive exploratory data analysis (EDA) of a dataset detailing Toronto's shelter usage trends. The dataset includes various metrics such as occupancy date, organization name, service user count, available beds and rooms, etc. Our analysis aims to uncover insights into the capacity utilization of the shelters and to statistically assess any differences in capacity rates. So that we can help develop Toronto's shelter support system more efficiently.

## Data preparation:

Since we only focus on the 'CAPACITY_TYPE', 'PROGRAM_MODEL', 'SERVICE_USER_COUNT', 'CAPACITY_ACTUAL_BED', 'OCCUPIED_BEDS', 'CAPACITY_ACTUAL_ROOM' and 'OCCUPIED_ROOMS', other columns are deleted from the dataset. After loading the columns that we need, there are some missing values, so we can calculate the total amount of missing values in each column to clean the data.

```
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
```

Then, we look at the basic statistics for the data frame to get a sense of how to clean the data.

| | SERVICE_USER_COUNT | CAPACITY_ACTUAL_BED | OCCUPIED_BEDS | CAPACITY_ACTUAL_ROOM | OCCUPIED_ROOMS |
|---|---|---|---|---|---|
| **count** | 50944.000000 | 32399.000000 | 32399.000000 | 18545.000000 | 18545.000000 |
| **mean** | 45.727171 | 31.627149 | 29.780271 | 55.549259 | 52.798598 |
| **std** | 53.326049 | 27.127682 | 26.379416 | 59.448805 | 58.792954 |
| **min** | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| **25%** | 15.000000 | 15.000000 | 14.000000 | 19.000000 | 16.000000 |
| **50%** | 28.000000 | 25.000000 | 23.000000 | 35.000000 | 34.000000 |
| **75%** | 51.000000 | 43.000000 | 41.000000 | 68.000000 | 66.000000 |
| **max** | 339.000000 | 234.000000 | 234.000000 | 268.000000 | 268.000000 |

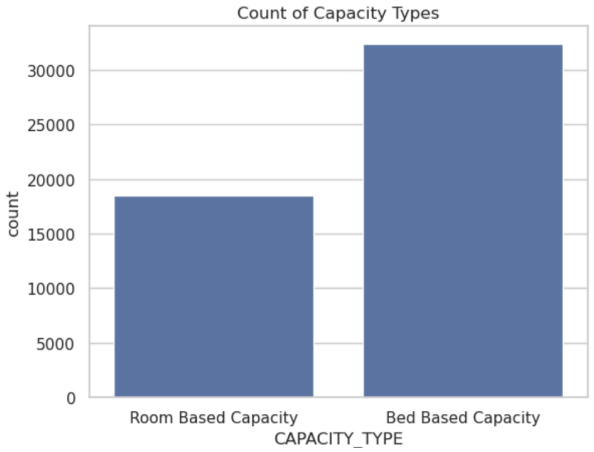
- For PROGRAM_MODEL, we fill the missing values by the most frequently occurring value which is its mode.
- For rows where CAPACITY_TYPE is 'Room Based Capacity', fill the columns about the bed with 0, and fill the columns about the room correspondingly.

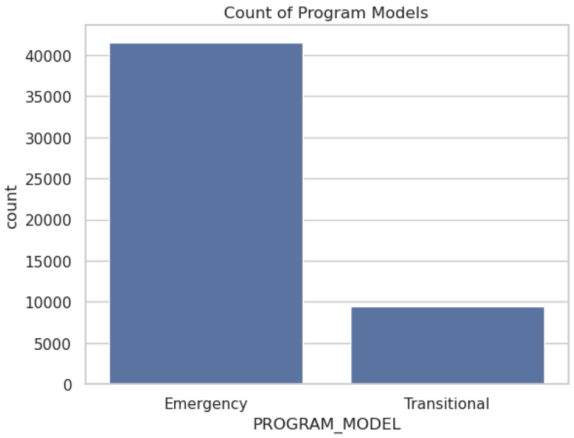
After that, we can check the number of missing values in each column was 0, so we can start doing the exploratory data analysis.

## Exploratory Data Analysis (EDA):

At the beginning of our data exploration, we look at two main things in our dataset: the types of capacity (like room or bed) and the types of programs (like emergency or transitional) we

have. We use simple bar charts to see how many of each type there are. This helps us understand what our data is mostly about and points us to what we should look at next.




- There are more bed-based types than room-based and the program model mainly belongs to emergency.

Next, we calculate and analyze how full the facilities are by looking at the occupancy of beds and rooms. We created a new column called ‘CAPACITY_RATE’ to show the actual occupancy based on the two different capacity types.

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

After looking at the overall capacity rate, we want to calculate the summary statistics to get a further understanding of the capacity rate based on the two different capacity types and two different program models.

```
Room 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
```

```
Bed 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
```

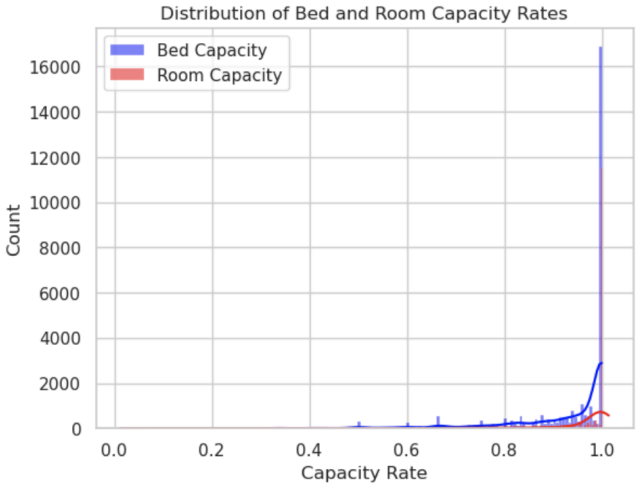
- For room occupancy rates, we find they range from a minimum of 1% to a maximum of 101%, with an average of 93%. Most rates cluster around 96% to 100%, indicating high and consistent usage of room capacity across facilities.
- Bed occupancy rates vary from 2% to 100%, averaging at 93%. The rates mainly fall between 90% and 100%, showing a widespread high utilization of bed capacity.

```
 Emergency Capacity summary statistics
Min: 0.01
Mean: 0.94
Max: 1.01
25th percentile: 0.95
Median: 1.0
75th percentile: 1.0
Interquartile range (IQR): 0.05
```

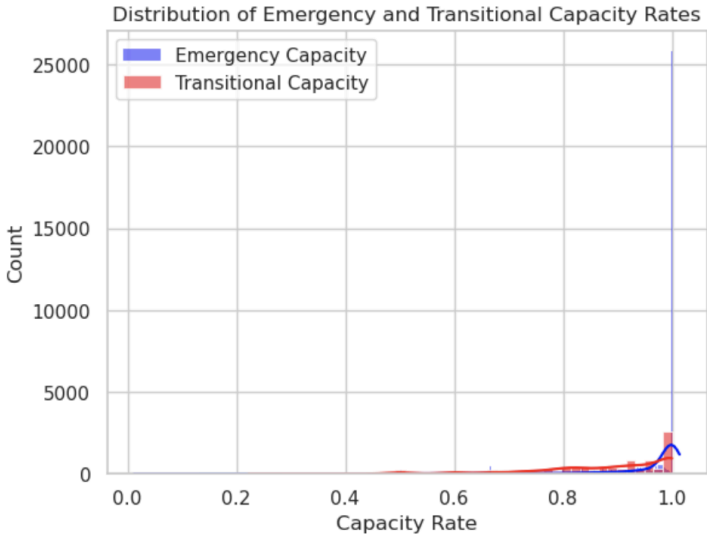
```
 Transitional Capacity summary statistics
Min: 0.22
Mean: 0.88
Max: 1.0
25th percentile: 0.82
Median: 0.92
75th percentile: 1.0
Interquartile range (IQR): 0.18
```

- Emergency capacity rates span from 1% to 101%, with an average of 94%. The majority of these rates are tightly grouped from 95% to 100%, reflecting a consistently high demand for emergency capacity.
- Transitional capacity rates range from 22% to 100%, averaging at 88%. These rates generally vary more widely, between 82% and 100%, indicating a diverse use of transitional capacity across different settings.

Then, we will create histograms to compare the distribution of bed and room capacity rates, and the distribution of emergency and transitional capacity rates.

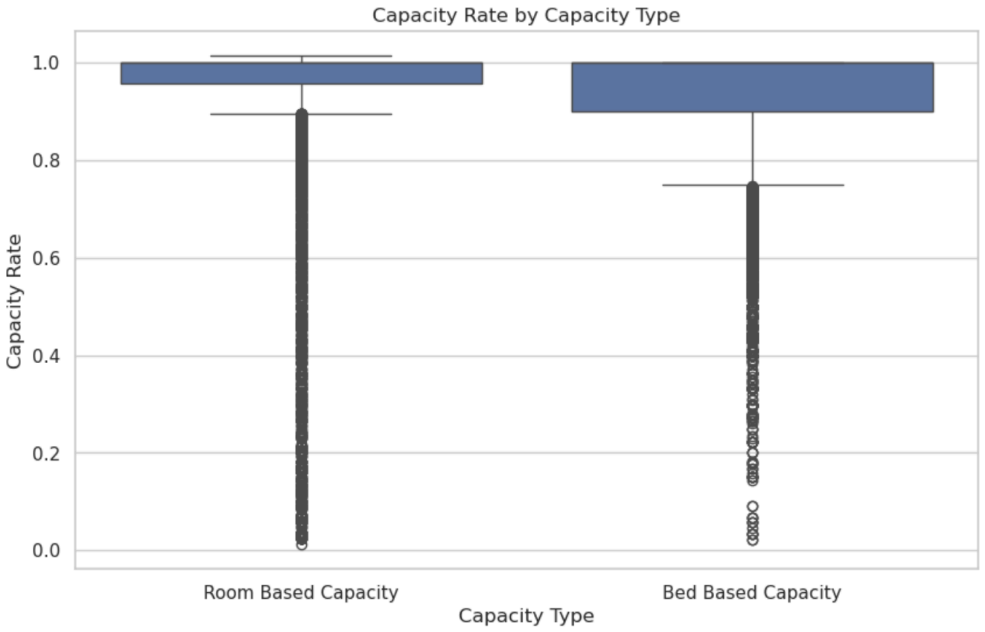



- The graph shows a high frequency of rates close to 1.0, indicating a tendency towards high occupancy. The room capacity, in particular, has a notable peak at full occupancy, which suggests that room resources are often maximized. The bed capacity distribution is slightly more spread out, indicating a bit more variability in bed occupancy rates.

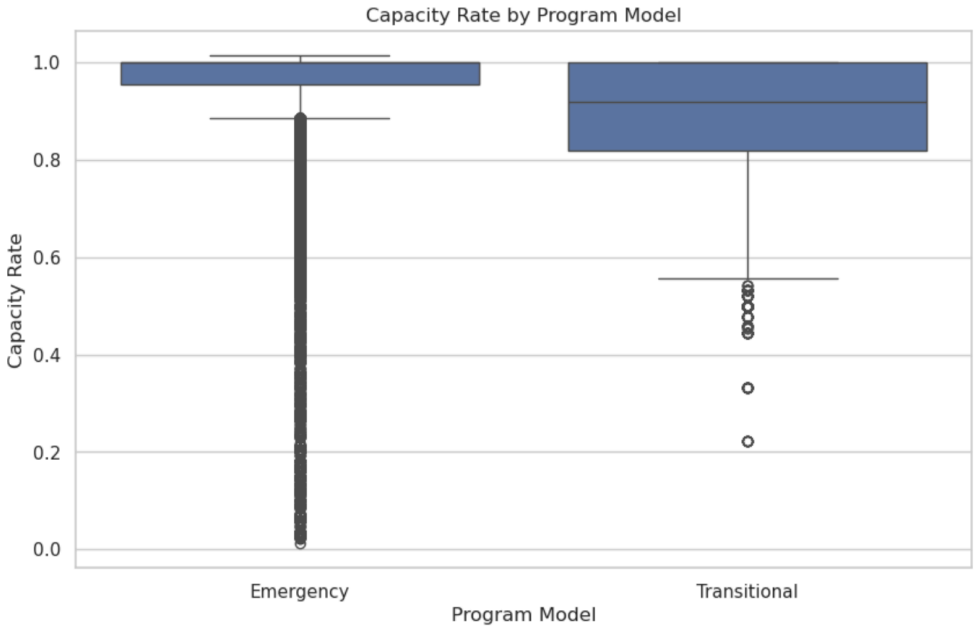


- The chart shows emergency and transitional capacity rates are mostly high, with a sharp peak at full occupancy for emergency services, indicating strong demand, while transitional services show a high but less sharp peak, suggesting varied but substantial use.

We can also look at the boxplots and similar patterns can be discovered in the histogram and summary statistics.




- The boxplot shows a generally high occupancy level for both, with room-based capacity often reaching full occupancy, while bed-based capacity demonstrates a wider range of occupancy rates.




- Emergency capacity rates are high, mostly at full capacity, whereas transitional rates have a broader spread, indicating more variation in their occupancy levels.

Based on the exploratory data analysis of capacity rates in different program models and types, we can formulate two research questions:

1. Is there a significant difference in capacity rate between the two capacity types?
2. Is there a significant difference in capacity rate between the two program models?

## T-test:

To determine the significant difference in capacity rate between room-based and bed-based types, we need to formulate the hypothesis.

- Null hypothesis: There is no significant difference in capacity rate between room-based and bed-based capacity types.
- Alternative hypothesis: There is a significant difference in capacity rate between room-based and bed-based capacity types.

By performing the two-sample t-test, we get the following result:

```
Two-sample t-test results:
t-statistic = 4.854104599422829
p-value = 1.2128933183471424e-06
```

The two-sample t-test result, with a t-statistic of 4.854 and a very small p-value (significantly less than 0.05), indicates a statistically significant difference in capacity rates between the room-based and bed-based capacity types. This suggests that the observed difference in capacity rates is unlikely to be due to random chance.

By performing the Welch's t-test, we get the following result:

```
Welch's t-test results:
t-statistic = 4.498751771925636
p-value = 6.860477551487939e-06
```

The Welch's t-test result, with a t-statistic of approximately 4.499 and a p-value around 6.86e-06, confirms the statistically significant difference in capacity rates between the two types. We can reject the null hypothesis that there is no difference between the capacity rates of the two capacity types, implying that the difference in capacity rates is indeed significant.

To determine the significant difference in capacity rate between the emergency and transitional program models, we need to formulate the hypothesis.

- Null hypothesis: There is no significant difference in capacity rate between the emergency and transitional program models.
- Alternative hypothesis: There is a significant difference in capacity rate between the emergency and transitional program models.

By performing the two-sample t-test, we get the following result:

```
Two-sample t-test results:
t-statistic = 39.06876276218507
p-value = 0.0
```

The two-sample t-test results, with a t-statistic of 39.069 and a p-value of 0.0, indicate a highly significant difference in capacity rates between the two program models. The extremely low p-value, effectively zero, strongly suggests that the observed differences in

capacity rates are not due to random chance but are statistically significant. This means there's a clear and significant difference in how the two program models utilize capacity.

By performing the Welch's t-test, we get the following result:

```
Welch's t-test results:
t-statistic = 40.97518639553636
p-value = 0.0
```

The Welch's t-test results, showing a t-statistic of 40.975 and a p-value of 0.0, further confirms a statistically significant difference in capacity rates between the two program models. We can reject the null hypothesis that there is no difference between the capacity rates of the two program models, implying that the difference in capacity rates is indeed significant.