Student Name: Nianchuer Liu
Student Number: 1010332454
Email:nianchuer.liu@mail.utoronto.ca

Experimental Design for Data Science -A1

1. Examine the dataset

From the preview of the data, we can identify the following columns:

- `OCCUPANCY_DATE`: The date of the record.
- `ORGANIZATION_NAME`: The name of the organization providing the service.
- `PROGRAM_ID`: A unique identifier for the program.
- `PROGRAM_NAME`: The name of the program.
- `SECTOR`: The sector the program serves (e.g., Families, Mixed Adult, Men, Women).
- `PROGRAM_MODEL`: The model of the program (e.g., Emergency).
- `OVERNIGHT_SERVICE_TYPE`: The type of overnight service provided (e.g., Motel/Hotel Shelter).
- `PROGRAM_AREA`: The area of the program, for instance, related to COVID-19 response.
- `SERVICE_USER_COUNT`: The count of service users.
- `CAPACITY_TYPE`: The type of capacity (e.g., Room Based Capacity).
- `CAPACITY_ACTUAL_BED`: The actual bed capacity.
- `OCCUPIED_BEDS`: The number of occupied beds.
- `CAPACITY_ACTUAL_ROOM`: The actual room capacity.
- `OCCUPIED_ROOMS`: The number of occupied rooms.

For the assignment, the columns of interest are `CAPACITY_TYPE`, `PROGRAM_MODEL`, `SERVICE_USER_COUNT`, `CAPACITY_ACTUAL_BED`, `OCCUPIED_BEDS`, `CAPACITY_ACTUAL_ROOM`, and `OCCUPIED_ROOMS`.

Since we've been asked to conduct t-tests, we'll focus on continuous variables that we can compare across different categories. From the provided columns, `SERVICE_USER_COUNT`, `CAPACITY_ACTUAL_BED`, `OCCUPIED_BEDS`, `CAPACITY_ACTUAL_ROOM`, and `OCCUPIED_ROOMS` are continuous, while `CAPACITY_TYPE` and `PROGRAM_MODEL` are categorical. We can conduct t-tests to compare, for example, the means of `OCCUPIED_ROOMS` between different `CAPACITY_TYPE` groups or between different `PROGRAM_MODEL` groups.

2. Compute the shelter program occupancy rates

The first step will be to compute the shelter program occupancy rates, which can be defined as the number of occupied beds or rooms divided by the actual capacity.

	BED_OCCUPANCY_RATE	ROOM_OCCUPANCY_RATE
count	32399.000000	18545.000000
mean	0.927885	0.934087
std	0.122562	0.163241
min	0.022727	0.012048
25%	0.900000	0.958333
50%	1.000000	1.000000
75%	1.000000	1.000000
max	1.000000	1.014085

The occupancy rates for beds and rooms have been computed and added to the dataset. Here are some statistics about these new continuous variables:

- **OCCUPANCY_RATE_BEDS**: This rate ranges from about 2.27% to 100%, with an average of approximately 92.79% and a standard deviation of 12.26%.
- **OCCUPANCY_RATE_ROOMS**: This rate ranges from about 1.20% to 101.41%, with an average of approximately 93.41% and a standard deviation of 16.32%.

3. T-test

To perform t-tests on the provided dataset, my approach involves several key steps:

1. **Creating Continuous Variables**: Since t-tests compare means of continuous variables across different groups, I compute new continuous variables if necessary. In this context, I calculate the occupancy rates (both for beds and rooms) as they offer a meaningful continuous measure for comparison.
2. **Selecting Categorical Variables for Comparison**: I choose categorical variables that might show different behavior in terms of occupancy rates. These could include `CAPACITY_TYPE`, `PROGRAM_MODEL`, or any other categorical variable of interest.
3. **Performing T-Tests**: I conduct t-tests to compare the mean occupancy rates between different groups defined by the selected categorical variables. The t-test helps in understanding whether the differences in means are statistically significant.

Here are the results of the t-tests along with their corresponding null and alternative hypotheses:

1. **T-Test for `OVERNIGHT_SERVICE_TYPE` with `BED_OCCUPANCY_RATE`**:
    - **Null Hypothesis**: There is no significant difference in bed occupancy rates between different `OVERNIGHT_SERVICE_TYPE`s.
    - **Alternative Hypothesis**: There is a significant difference in bed occupancy rates between different `OVERNIGHT_SERVICE_TYPE`s.
    - **T-Statistic**: -29.9999
    - **P-Value**: Approximately : Approximately 1.00e-194
    - Given the extremely low p-value, we reject the null hypothesis. This suggests that there is a statistically significant difference in bed occupancy rates between different

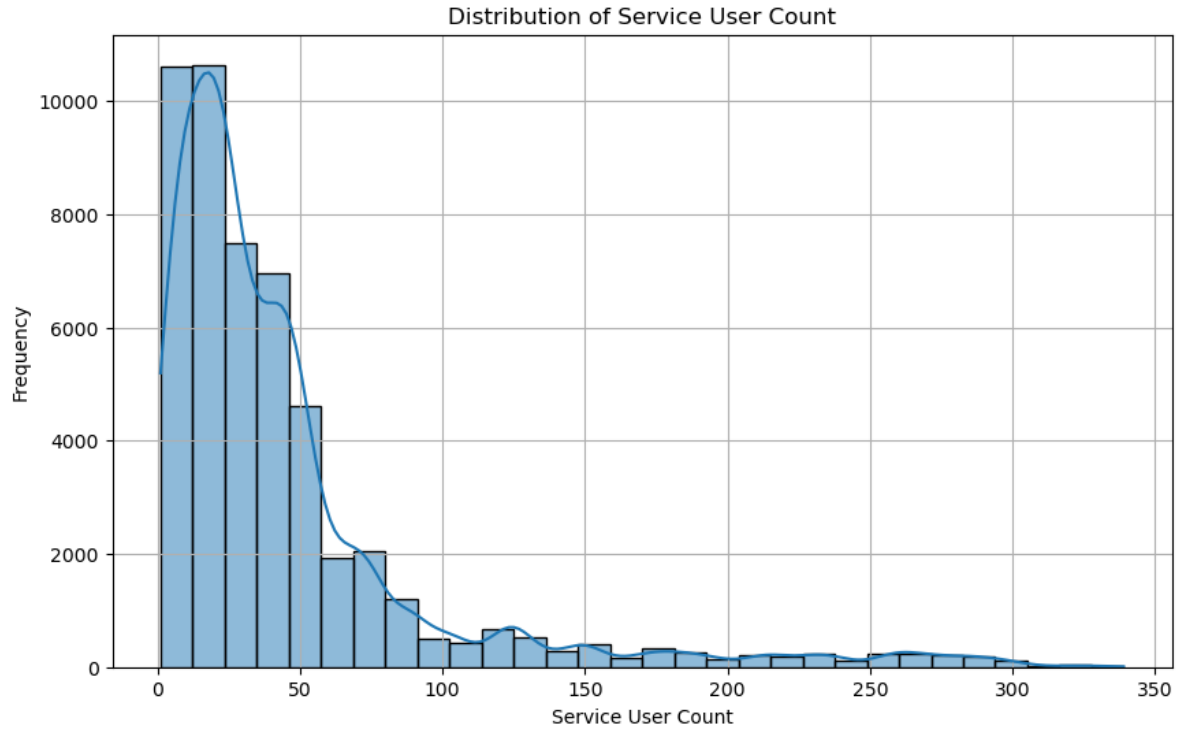
`OVERNIGHT_SERVICE_TYPE`s.

2. **T-Test for `SECTOR` with `ROOM_OCCUPANCY_RATE`**:
    - **Null Hypothesis**: There is no significant difference in room occupancy rates between different `SECTOR`s.
    - **Alternative Hypothesis**: There is a significant difference in room occupancy rates between different `SECTOR`s.
    - **T-Statistic**: 8.6608
    - **P-Value**: Approximately 5.29e-18
    - This result also leads to the rejection of the null hypothesis, indicating a significant difference in the means.
3. **T-Test for `PROGRAM_MODEL` with `BED_OCCUPANCY_RATE`**:
    - **Null Hypothesis**: There is no significant difference in bed occupancy rates between different `PROGRAM_MODEL`s.
    - **Alternative Hypothesis**: There is a significant difference in bed occupancy rates between different `PROGRAM_MODEL`s.
    - **T-Statistic**: 38.7807
    - **P-Value**: Approximately 1.26e-321
    - The extremely low p-value suggests a statistically significant difference, thus leading to the rejection of the null hypothesis.

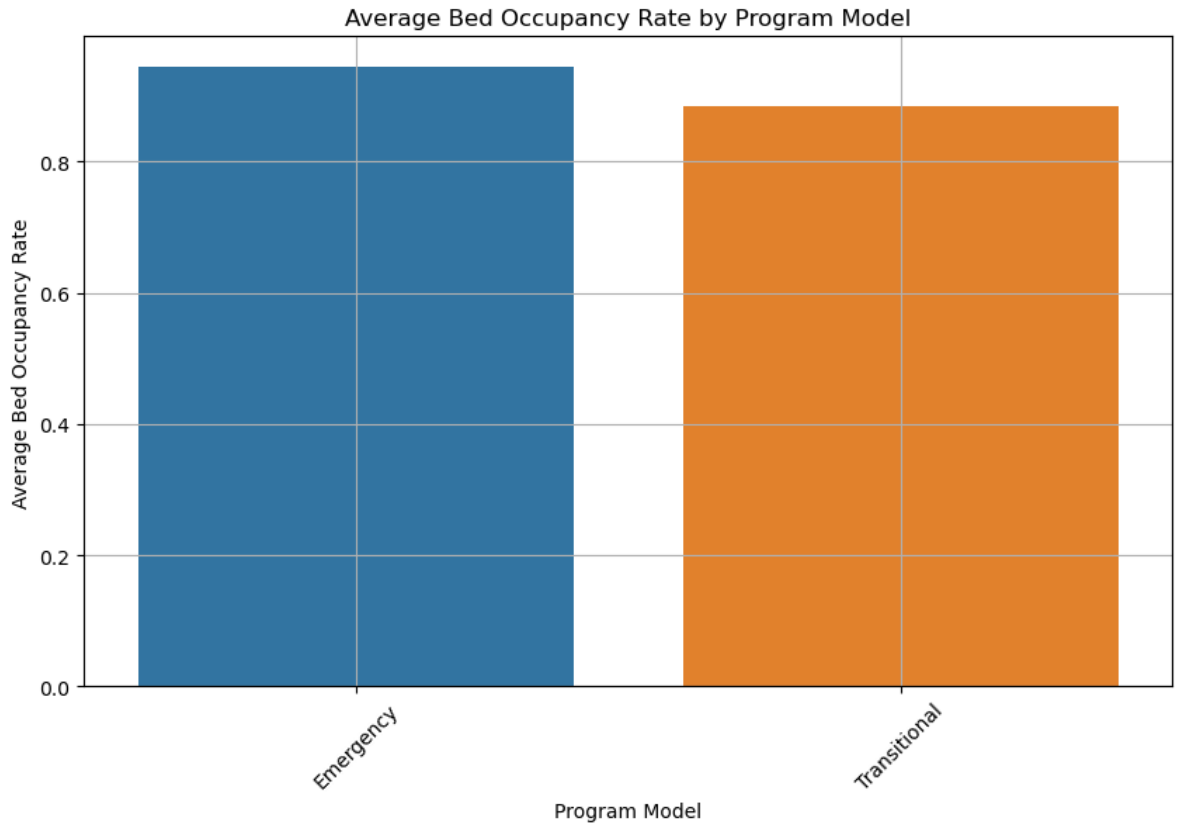
4. EDA

The exploratory data analysis (EDA) of the shelter data reveals several interesting points that could lead to further investigation:

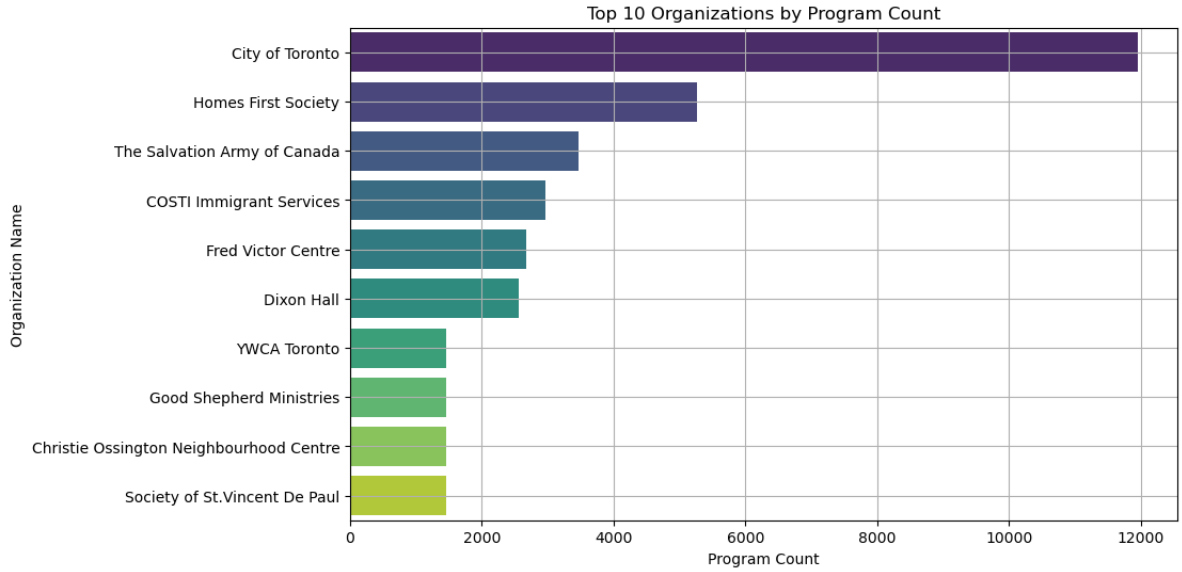
1. **Service User Count**: The histogram of `SERVICE_USER_COUNT` indicates a right-skewed distribution. Most programs have a relatively low number of users, with few programs accommodating a very high number of users. This could suggest that a small number of programs are shouldering a much larger burden in terms of user numbers, which might affect the quality of service or resource allocation.



1. **Bed Occupancy Rate by Program Model**: The bar chart displaying the average bed occupancy rate by `PROGRAM_MODEL` shows that 'Emergency' programs have a higher occupancy rate than 'Transitional' ones. This could indicate a higher demand or a more efficient usage of available resources in emergency shelters.



1. **Programs per Organization**: The bar chart for the number of programs per organization shows that the `City of Toronto` has the highest number of programs, followed by `Homes First Society` and `The Salvation Army of Canada`. This concentration of programs in a few organizations might point to the centralization of services and the potential for economies of scale, but it may also raise questions about diversity of service provision and risk concentration.



Summary Statistics:

- The `PROGRAM_ID` suggests that the data might cover a range of programs within a specific identification range.
- The `SERVICE_USER_COUNT` has a wide range, with a mean significantly higher than the median, again indicating a skewed distribution.
- For `CAPACITY_ACTUAL_BED` and `OCCUPIED_BEDS`, the means are quite close, suggesting high occupancy rates.
- `CAPACITY_ACTUAL_ROOM` and `OCCUPIED_ROOMS` also show high utilization.

Further Analysis:

- **Temporal Trends**: Analysis of `OCCUPANCY_DATE` could reveal seasonal patterns or trends over time in program usage and occupancy rates.
- **Resource Allocation**: A deeper look into the data could assess whether resources like beds and rooms are being allocated efficiently across different programs and organizations.
- **Program Effectiveness**: Examining outcomes for service users, such as duration of stay or follow-up status after leaving the shelter, could provide insight into the effectiveness of different program models.
- **Demographic Factors**: Incorporating demographic data of service users could help in understanding who is being served by these programs and whether some groups are over- or under-represented.