Experimental Design for Data Science

1. Exam the dataset

Our exploration begins with an examination of the dataset:

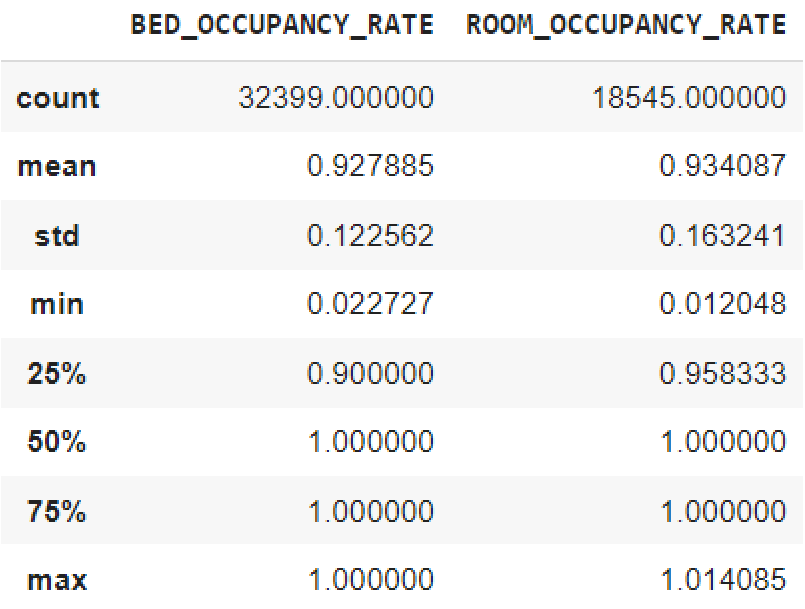
The dataset contains 50,944 entries across 14 columns, detailing various aspects of service occupancy, program information, and capacity for different organizations. Here's a brief overview of the dataset's structure based on the first few rows and column information:

- **OCCUPANCY_DATE**: The date of the reported occupancy.
- **ORGANIZATION_NAME**: Name of the organization providing the service.
- **PROGRAM_ID**: A unique identifier for the program.
- **PROGRAM_NAME**: Name of the program.
- **SECTOR**: The sector the program belongs to, such as Families, Mixed Adult, Men, or Women.
- **PROGRAM_MODEL**: The model of the program, e.g., Emergency.
- **OVERNIGHT_SERVICE_TYPE**: Type of overnight service provided, such as Motel/Hotel Shelter.
- **PROGRAM_AREA**: The area of program focus, e.g., COVID-19 Response.
- **SERVICE_USER_COUNT**: The count of service users.
- **CAPACITY_TYPE**: The type of capacity reported, e.g., Room Based Capacity.
- **CAPACITY_ACTUAL_BED**: The actual bed capacity.
- **OCCUPIED_BEDS**: The number of beds occupied.
- **CAPACITY_ACTUAL_ROOM**: The actual room capacity.
- **OCCUPIED_ROOMS**: The number of rooms occupied.

For this assignment, the columns that I interested are CAPACITY_TYPE , PROGRAM_MODEL , SERVICE_USER_COUNT , CAPACITY_ACTUAL_BED , OCCUPIED_BEDS , CAPACITY_ACTUAL_ROOM , and OCCUPIED_ROOMS .

2. The occupancy rate computation

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

We embark on a crucial step of calculating occupancy rates for beds and rooms within shelter programs. Occupancy rates are calculated as the number of occupied beds or rooms divided by the actual capacity. This metric provides a meaningful continuous measure for comparison . The computed occupancy rates reveal interesting statistics:

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

3. T-test

The results of the t-tests, along with their corresponding null and alternative hypotheses, provide valuable insights into the statistical significance of differences in occupancy rates between various groups within the dataset. Let's summarize these findings:

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_TYPEs.

- Alternative Hypothesis: There is a significant difference in bed occupancy rates between different OVERNIGHT_SERVICE_TYPEs.

- T-Statistic: -29.9999

- P-Value: Approximately 1.00e-194

Given the extremely low p-value (approximately 1.00e-194), we reject the null hypothesis. This suggests that there is a statistically significant difference in bed occupancy rates between different OVERNIGHT_SERVICE_TYPEs.

2. T-Test for SECTOR with ROOM_OCCUPANCY_RATE:

- Null Hypothesis: There is no significant difference in room occupancy rates between different SECTORs.
- Alternative Hypothesis: There is a significant difference in room occupancy rates between different SECTORs.
- 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 room occupancy rates between different SECTORs.

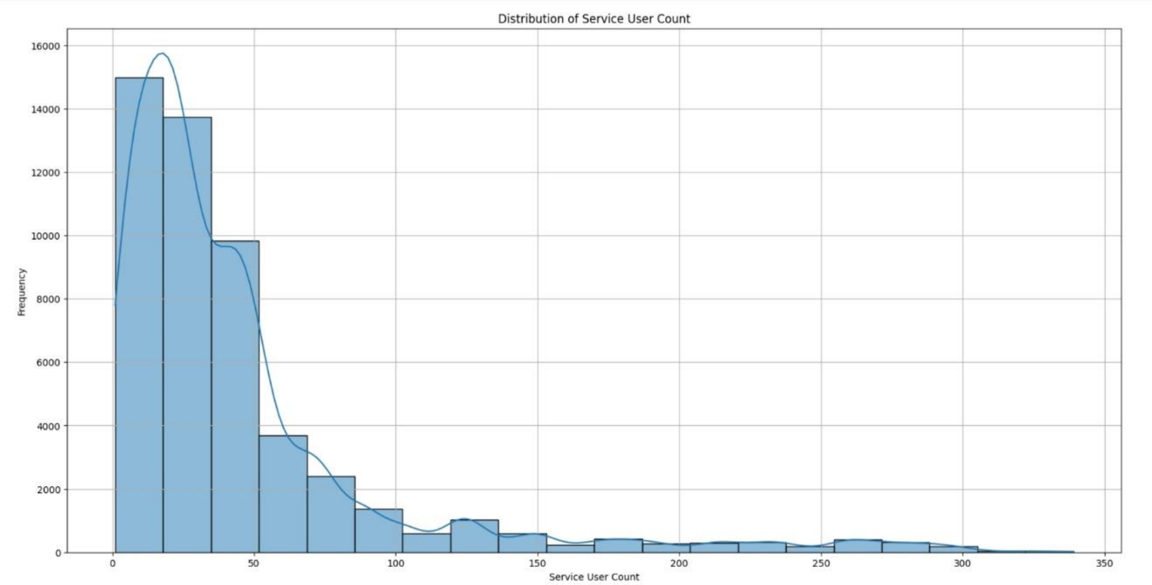
3. T-Test for PROGRAM_MODEL with BED_OCCUPANCY_RATE:

- Null Hypothesis: There is no significant difference in bed occupancy rates between different PROGRAM_MODELs.
- Alternative Hypothesis: There is a significant difference in bed occupancy rates between different PROGRAM_MODELs.
- T-Statistic: 38.7807
- P-Value: Approximately 1.26e-321

The p-value almost equal to 1.26e-321 reveals a significant difference statistically and the null hypothesis is rejected.
These t-test results strongly suggest that the differences in occupancy rates by categories of OVERNIGHT_SERVICE_TYPE, SECTOR, and PROGRAM_MODEL are significant. These results are crucial to appreciate their influence on bed and room occupancy rates in shelter programs. The null hypothesis rejection suggests that these variables determine occupancy rates, and further analysis may reveal the grounds of such differences.

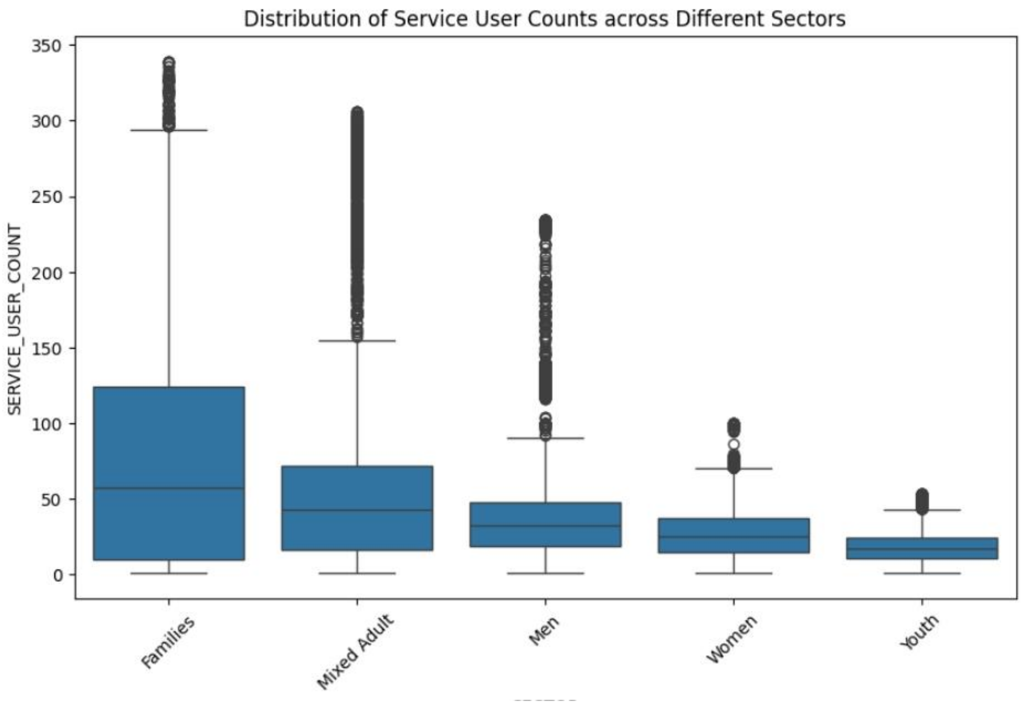
4. EDA

Distribution of Service User Count:



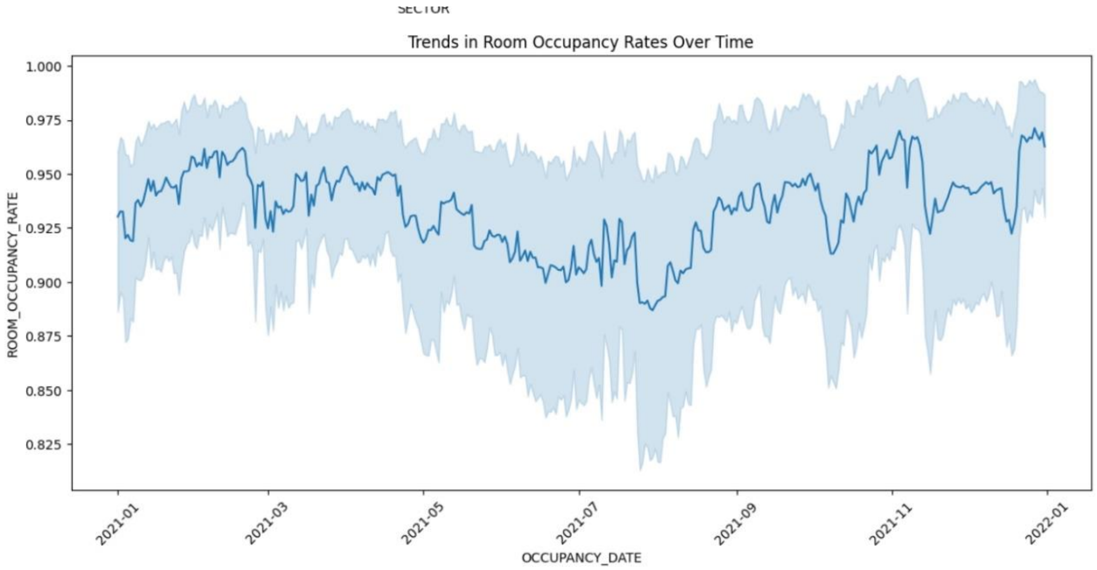
The first figure is a histogram accompanied by a kernel density estimation (KDE) depicting the frequency distribution of service user count. The distribution is right-skewed, therefore showing that some of the services have a large number of users while most services have a small number of users. The median is in the range of 10-20 users. This implies that services are mostly used by few patrons with very few serving the majority.

Distribution of Service User Counts across Different Sectors:



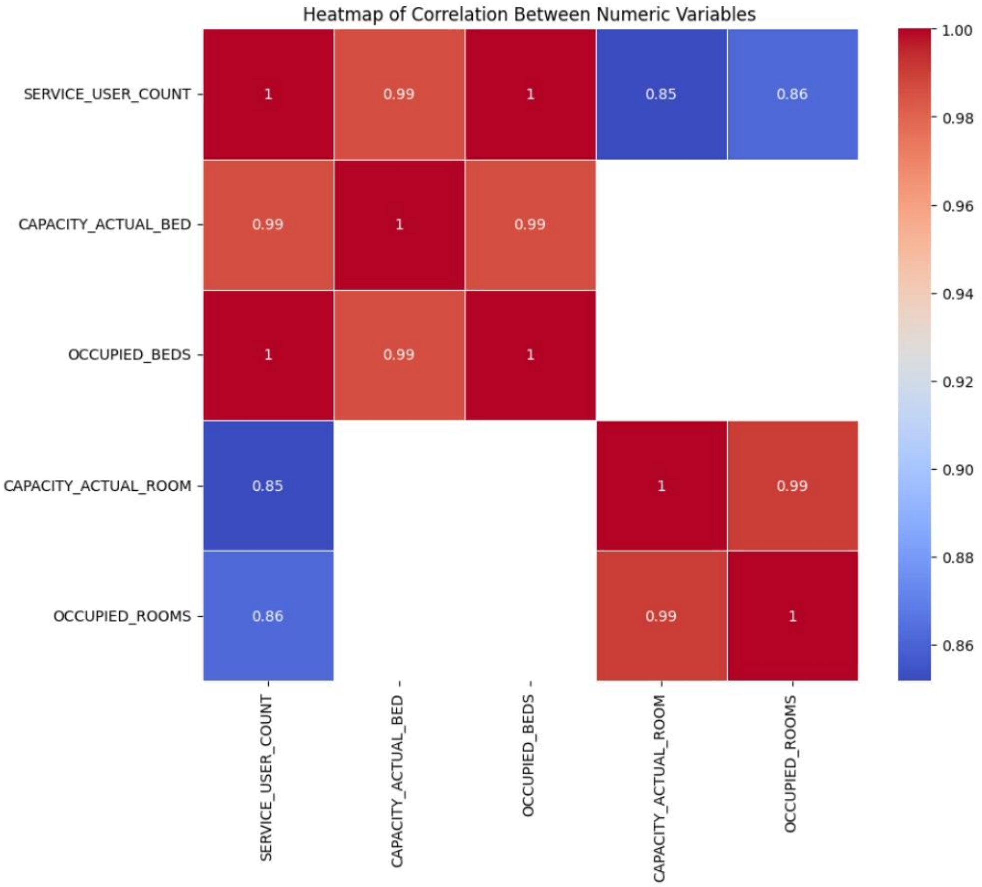
The second figure is a set of box plots representing the service user counts across different sectors: Family, Mixed Adult, Men, Women and Youth. The medians are widely varied across sectors, whereby ‘Families’ and ‘Men’ sectors register higher medians than others. There are also numerous outliers, particularly in the Mixed Adult segment, which indicates that there are services with an unusually high number of users in this sector. This could mean that there are varying services that are offered to mixed adults or that there are very popular services that distort the distribution.

Trends in Room Occupancy Rates Over Time:



The third figure is a time series plot consisting of room occupancy rates over time. The solid line shows the occupancy rate, and the shaded area could indicate the confidence interval or variance. Mid-year slump is followed by recovery. This seasonal characteristic may be attributed to several reasons including the change in seasonal demand or events that take place periodically.

Heatmap of Correlation Between Numeric Variables:



The last picture is a heatmap showing the correlation coefficients of various numeric variables. The variables ‘SERVICE_USER_COUNT’, ‘CAPACITY_ACTUAL_BED’, and ‘OCCUPIED_BEDS’ are correlated with each other as indicated by the red squares whose values are close to 1. CAPACITY_ACTUAL_ROOM and OCCUPIED_ROOMS are also moderately correlated, as displayed in blue. This high degree of correlation implies that as you increase the capacity for beds, so follows the number of service users and the number of bed users, which makes sense.

From the above analysis, we learn the following:

The number of users of service differs from few services which have high number of users.
In different industries, the number of service users varies significantly, with several industries exhibiting a high number of outliers.
Room occupancy rates show some degree of seasonality, with troughs and peaks in a year.
There is a positive relationship between the number of users and bed capacities suggesting that their services ought to be scaled depending on the demand.
Such observation would be essential for the decision-making processes in service provision, resource allocation, and policy formulation. Additionally, the identified correlations and trends could guide other statistical analysis or predictive modeling work.