INF 2178 - Experimental Design For Data Science

Technical Assignment 1

By

Saad Umar

Course Instructor: Shion Guha

Date: Feb 04, 2024

Overview

The dataset is a compilation of records from various shelters across Toronto, and provides a rich source of information. It includes key details such as shelter types, capacities, occupancy rates, and demographic data. A preliminary look at the dataset reveals a wide range of shelters, catering to different sectors of the homeless population including men, women, youth, and families. The diversity of the shelters, ranging from emergency to transitional types, highlights the multidimensional approach taken to address the immediate and long-term needs of the homeless.

```
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   OCCUPANCY_DATE          50944 non-null  datetime64[ns]
 1   ORGANIZATION_NAME       50944 non-null  object
 2   PROGRAM_ID              50944 non-null  int64
 3   PROGRAM_NAME            50909 non-null  object
 4   SECTOR                  50944 non-null  object
 5   PROGRAM_MODEL           50942 non-null  object
 6   OVERNIGHT_SERVICE_TYPE  50942 non-null  object
 7   PROGRAM_AREA            50942 non-null  object
 8   SERVICE_USER_COUNT      50944 non-null  int64
 9   CAPACITY_TYPE           50944 non-null  object
 10  CAPACITY_ACTUAL_BED     32399 non-null  float64
 11  OCCUPIED_BEDS           32399 non-null  float64
 12  CAPACITY_ACTUAL_ROOM    18545 non-null  float64
 13  OCCUPIED_ROOMS          18545 non-null  float64
```

Fig 1

Process

The first step was to import the relevant libraries and the shelter dataset which was assigned to the *df* variable.

The next was the preprocessing of data. The main goal of data preprocessing is to remove any inconsistencies and improve the quality of the data. From figure 1, we can see that there are 50,944 records and most of the data is available. After finishing preprocessing, we moved towards the EDA and descriptive statistics, the purpose of this is to aid in describing and summarizing the data. This phase is crucial for understanding the dataset's characteristics. It typically includes summarizing statistics (like mean and median), examining distributions of variables, and identifying patterns or anomalies. Graphical representations such as histograms, box plots, and scatter plots were used to visualize these aspects. From our descriptive statistics, we got the following information, information from *table 1* later aided us with the t-tests to see if there is a significant difference in the occupancy rates between men and women, discussed later in this writeup.

	Room Occupancy rate for Men	Room Occupancy rate for Women
Min	0.82	0.38
Max	1.0	1.0
Mean	1.0	0.97
Median	1.0	1.0
IQR	0.0	0.01

Table 1

The histogram in figure 2 below shows that the service user count is right-skewed. The majority of the data points are concentrated on the left side of the histogram, indicating that most values are on the lower side. The central tendency of the distribution will ideally be represented by the median due to the influence of the right tail on the mean.
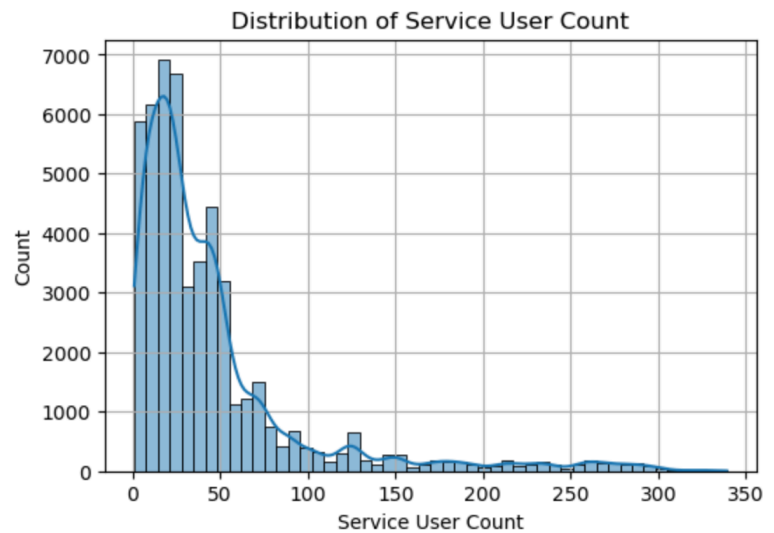
Figure 3 shows that the shelters being provided in Toronto are either in the capacity of rooms or beds. The most common capacity Type appears to be Bed Based Capacity dominating across the various sectors.
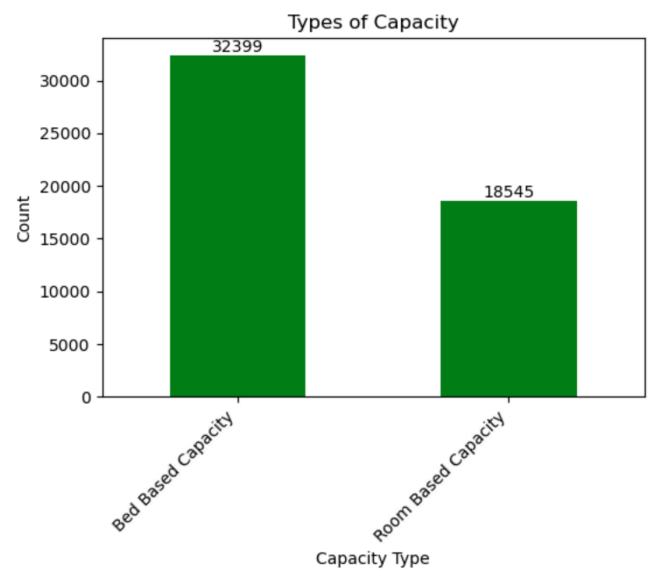

Fig 2

Figure 4 shows the trend of occupancy rates for bed based and room based shelters over the period of one year from January 2021 to January 2022. The trends show that there is comparatively a higher occupancy rate for room based shelters. The figure shows that there is an initial increase in the occupancy rates for both the capacity types, but we see that from April to June the occupancy rates fall a little and then rise. Since the occupational rates are fluctuating cyclically, it's not possible to exactly pinpoint the cause of it with our analysis. But during this time, the covid vaccines were being rolled out while the covid cases were reaching their second peak, hence the cyclic trend. However, this cannot be labeled as a confounding variable and cannot be linked to cause-and-effect. It might be one of the potential reasons contributing to this trend of shelter occupancy rates.

Moreover, also not to fall into any paradox, the figure 5 shows that the number of people using the shelter services have consistently been increasing from April to December.

Fig 3 above & Fig 4 below

2
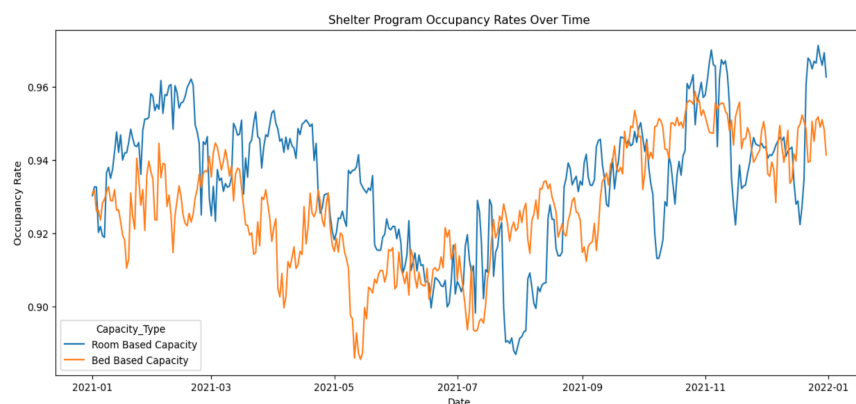
Insights from analysis

Figure 4 below displays the distribution across different sectors for two types of capacity: room-based and bed-based. On the x-axis, we have the two capacity types - room-based capacity and bed-based capacity. The y-axis represents the count of each capacity type being used. Each bar corresponds to a specific sector, including families, mixed adults, men, women, and youth.

Fig 5

The varying heights of the bars provide a visual comparison of counts across the different sectors for both room-based and bed-based capacities, offering insights into the distribution of services among diverse demographic groups.
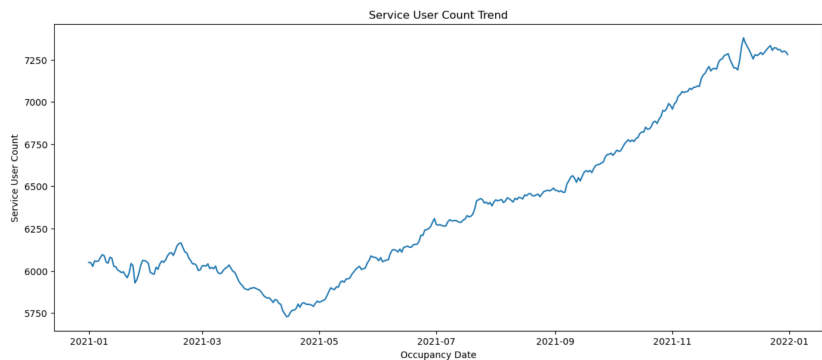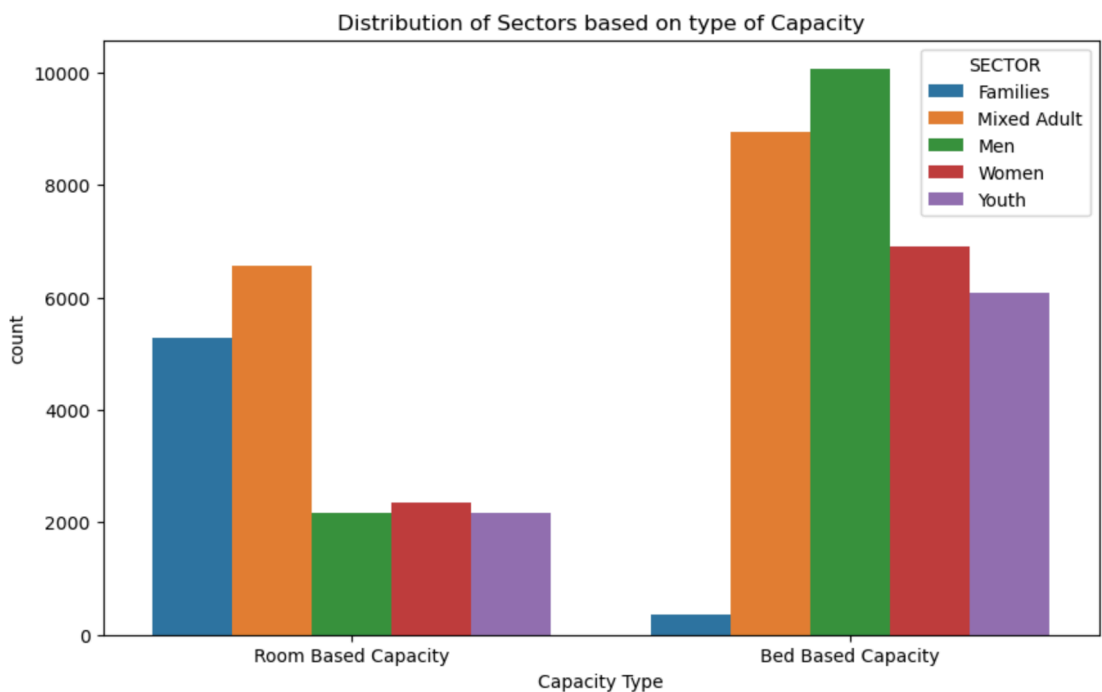

Fig 6

The gender-based analysis of the shelters presents another layer of complexity. It reveals significant differences in the number of bed based capacity for shelters between men and women. This could be indicative of broader societal issues such as employment challenges and social support structures. Women's bed based shelters, while fewer in number, highlight the need for targeted services, especially for those facing specific challenges like domestic violence.

	P- value	T-Statistic	Meaning
T-test 1: Compare type of capacity and its respective occupancy rate	6.8604e-06	-4.4988	H0 = There is no significant difference in the mean of occupancy rate for bed based and room based capacities H1 = There is a significant difference in the mean of bed based and room based capacities. Result: Since our alpha is (0.05), the p-value is less than our alpha so the test is significant and we can conclude there is significant difference in the mean between the two capacities.
T-Test 2: compare the Room Occupancy Rate between Men & Women	2.3324e-41	13.7292	H0 = There is no significant difference in the mean of room occupancy rate for men and women. H1 = There is a significant difference in the mean of room occupancy rate for men and women Result: Since our alpha is (0.05), the p-value is less than our alpha so the test is significant and we can conclude there is significant difference in the mean for room based capacity for men and women.
T-test 3: compare the Bed Occupancy Rate between Men & Women	6.4910e-38	12.9180	H0 = There is no significant difference in the mean of bed based occupancy rate between men and women. H1 = There is a significant difference in the mean of bed based occupancy rate for men and women. Result: Since our alpha is (0.05), the p-value is less than our alpha so the test is significant and we can conclude there is significant difference in the mean between men and women for bed based occupancy rate.

T-test 4: compare the service user count for the two different types of capacities	6.8605e-06	-4.4988	H0 = There is no significant difference in the mean of service user count for bed based and room based capacities
			H1 = There is a significant difference in the mean of service user count for bed based and room based capacities.
			Result: Since our alpha is (0.05), the p-value is less than our alpha so the test is significant and we can conclude there is significant difference in the mean between the two capacities.

Table 2

Delving into the impact of shelter types on service user counts, the study finds that room-based shelters tend to have more stable user counts compared to bed-based shelters. This suggests a preference among users for the privacy and stability offered by room-based accommodations. The analysis also touches upon how different shelter configurations impact the well-being of the homeless, pointing towards a need for varied resource allocation to cater effectively to the diverse needs of this population.
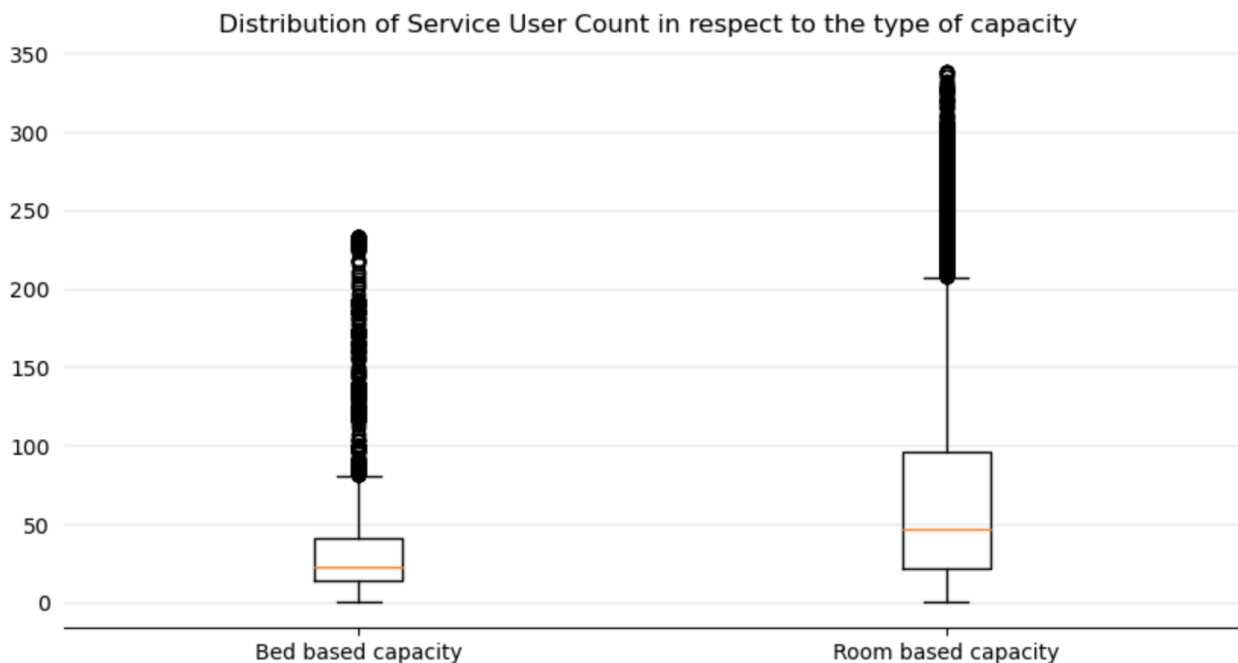

Fig 7

Possible Assumptions, limitations and further Insights
While performing the t-test, some assumptions were made: (a) The samples were independent of each other and there is no relationship between the observations of each group. (b) The data in each group is approximately normally distributed. (c) There are no significant outliers in the two groups. T-Test 2: which compares the Room Occupancy Rate between Men & Women shows that there is significant difference between the two. Meanwhile it shows the same for bed occupancy rate as well as the service user count for the different capacity types (t-test: 4). Several factors could contribute to these results, potentially: Men and women might have different preferences when it comes to shelter types or conditions. For example, women may prefer shelters with more privacy, safety measures, or specific amenities, leading to variations in occupancy rates. Shelter policies or regulations may vary based on gender. It's possible that certain shelters prioritize accommodating women differently than men, leading to disparities in occupancy rates. Differences in demographic characteristics, such as age, family status, or specific needs of men and women, could contribute to variations in their utilization of shelter services. Safety concerns, especially among women, may influence their decision to stay in shelters. Women might be more selective in choosing shelters based on safety features, impacting the occupancy rates. Availability and effectiveness of community outreach and support programs tailored for specific gender groups could affect the decision-making process of seeking shelter. Outreach efforts may be more effective for one gender, influencing their shelter usage. Cultural and social factors may play a role in shaping individuals' decisions regarding shelter use. Different cultural norms and expectations could contribute to variations in occupancy rates. Differences in the accessibility of shelter services for men and women could impact occupancy rates. For example, if there are more shelters for women in certain areas, it may influence their utilization.

Some limitations of using this dataset involved that the actual and occupied bed and room occupancies could not directly be used for t-test since they were discrete variables. Though it's theoretically possible to use a t-test with discrete variables, there are several limitations and assumptions associated with the t-test that may make it inappropriate and less reliable for discrete variables: e.g. t-test assumes that the data within each group follows a normal distribution. With continuous variables, the Central Limit Theorem often allows for the assumption of normality, even with non-normally distributed populations. However, this assumption is harder to justify for discrete variables.

Understanding the underlying reasons for the observed differences requires a comprehensive analysis of socio-economic, cultural, and policy-related factors. It is important to consider these potential explanations and, if possible, gather additional qualitative data or conduct further investigations to gain deeper insights into the specific dynamics contributing to the observed differences in Room and Bed Occupancy Rates between Men and Women in shelter programs.