Suha Lee
1005155626
04-02-2024

INF2178 Technical Assignment 1 – Narrative

The shelter usage trends dataset examines active shelter services in Toronto and their usage in the year 2021. Featuring a wide range of shelters covering various service types and targeting different demographics, the dataset provides insight into the needs of the homeless population in Toronto. As general motivation to direct my research questions, I am interested in discerning patterns that could suggest trends in demand for different services based on personal and environmental factors. By identifying areas of increased demand, we can begin considering prescriptive measures to address the needs of the growing homeless population as well as predicting and accounting for additional strain on the system before it occurs.

We begin our analysis by preprocessing and cleaning the dataset. Since we are interested in overall service use across the observed facilities instead of comparing individual programs, we drop columns concerning program identification such as ORGANIZATION_NAME. For comparison across programs, we are more concerned with occupancy rates as opposed to total service user counts. Since we do not have an understanding of the scope of the coverage of the dataset, it is difficult to extrapolate meaning by strictly comparing SERVICE_USER_COUNT. External factors such as location and funding can both greatly impact these values, not to mention unintentional bias that could have been introduced in the reporting and collection of the dataset. Instead, we introduce our own OCCUPANCY_RATE variable by dividing the reported occupancy values by the corresponding capacity values. This provides a metric that we can use to compare usage rates across the observations in the dataset. We note that there are instances where observations are missing relevant values or display values outside of the valid range. Since these observations make up an insignificant portion of the total dataset, we elect to adjust or drop the observations as appropriate without greatly impacting the data.

In our exploratory data analysis (EDA), we begin by producing summary statistics for each of the numerical variables present in the dataset, as shown in Table 1 below.

	SERVICE_USER_COUNT	CAPACITY_ACTUAL_BED	OCCUPIED_BEDS	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS	OCCUPANCY_RATE
count	50942.000000	32397.000000	32397.000000	18545.000000	18545.000000	50942.000000
mean	45.728515	31.628145	29.781400	55.549259	52.798382	0.930148
std	53.326660	27.128189	26.379825	59.448805	58.792876	0.138786
min	1.000000	1.000000	1.000000	1.000000	1.000000	0.012048
25%	15.000000	15.000000	14.000000	19.000000	16.000000	0.923077
50%	28.000000	25.000000	23.000000	35.000000	34.000000	1.000000
75%	51.000000	43.000000	41.000000	68.000000	66.000000	1.000000
max	339.000000	234.000000	234.000000	268.000000	268.000000	1.000000

Table 1: Summary Statistics for Numerical Variables

A key point of interest is that while all other metrics tend to deviate greatly, the overall OCCUPANCY_RATE is heavily skewed towards full capacity. This indicates that of the services

observed, regardless of the supply the demand will rise to meet it. Using histograms (Figure 2) we can observe the different distributions.

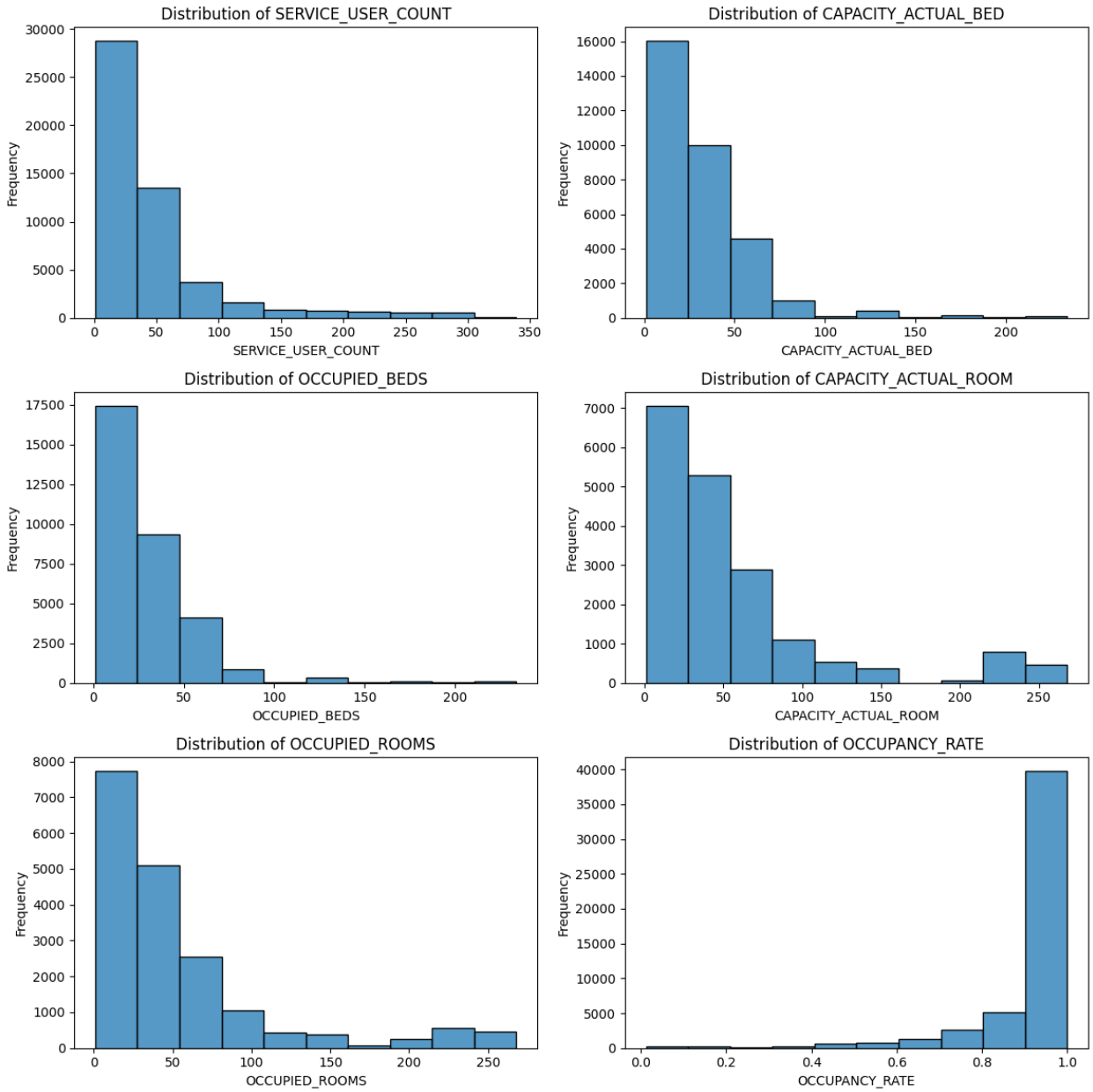

Figure 2: Distributions of the Numerical Variables

Figure 2 demonstrates the tendency for shelters to skew towards smaller sizes in terms of capacity. Regardless of the capacity, the occupancy rates tend towards 100%, as seen by the severely left skewed OCCUPANCY_RATE distribution and the occupancy histograms that closely mirror their corresponding capacity histograms. We can attempt to visualize this growth in the homeless population by plotting a trend of the total SERVICE_USER_COUNT over the course of the year.

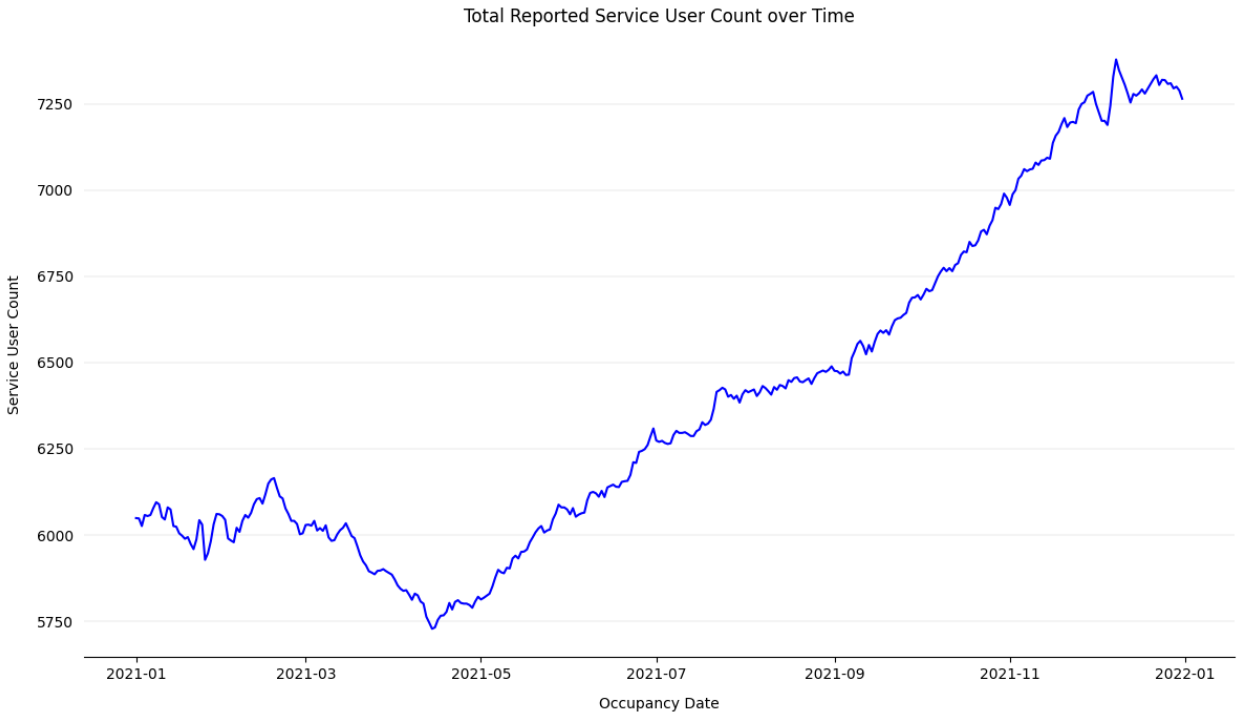

Figure 3: Total SERVICE_USER_COUNT across 2021

Aside from a slight dip around April 2021, we see significant, steady growth over time. Due to the limitations of the dataset, we are unable to discern the exact factors influencing this. To further pursue this line of inquiry we may consider additional sources examining legal, economical, and health related events that coincide with this period.

Returning our focus to comparisons of OCCUPANCY_RATE across different factors, we visualize differences in distributions by producing boxplots of OCCUPANCY_RATE divided by different classes of the categorical variables. Namely, we are interested in examining differences in SECTOR, PROGRAM_MODEL, OVERNIGHT_SERVICE_TYPE, PROGRAM_AREA, and CAPACITY_TYPE. SECTOR refers to the demographic (age and sex) of the population the shelter provides services for. PROGRAM_MODEL, OVERNIGHT_SERVICE_TYPE, and PROGRAM_AREA are all classifications of the different needs and circumstances the service in question fulfills. Finally, CAPACITY_TYPE indicates whether the shelter operates on a per-bed or per-room basis.

Looking at the boxplots by SECTOR (Figure 4) we observe clear differences in the deviations for OCCUPANCY_RATE. While all services skew towards full capacity regardless of demographic, some services such as those that target men do so more severely. This may reflect a greater demand for shelter by individuals that identify as such, implying that either these groups are at greater risk or that they seek shelter services as a solution at a greater rate. Similarly, we observe visual differences in the boxplots for CAPACITY_TYPE (Figure 5). Room-based capacity services are skewed more heavily towards full capacity than those that operate under bed-based capacity. This may indicate a preference for service users to seek out shelters that afford them additional privacy, suggesting a greater demand for such services.

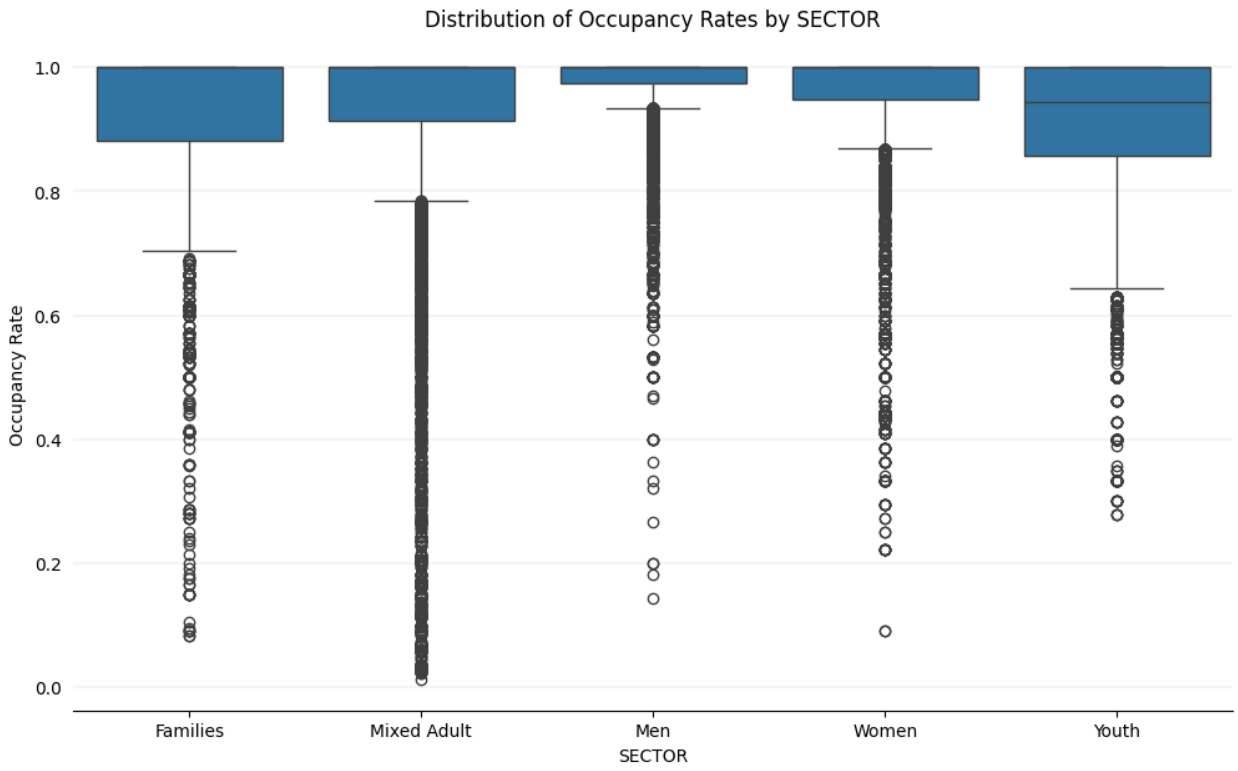

Figure 4: Boxplots of OCCUPANCY_RATE by SECTOR

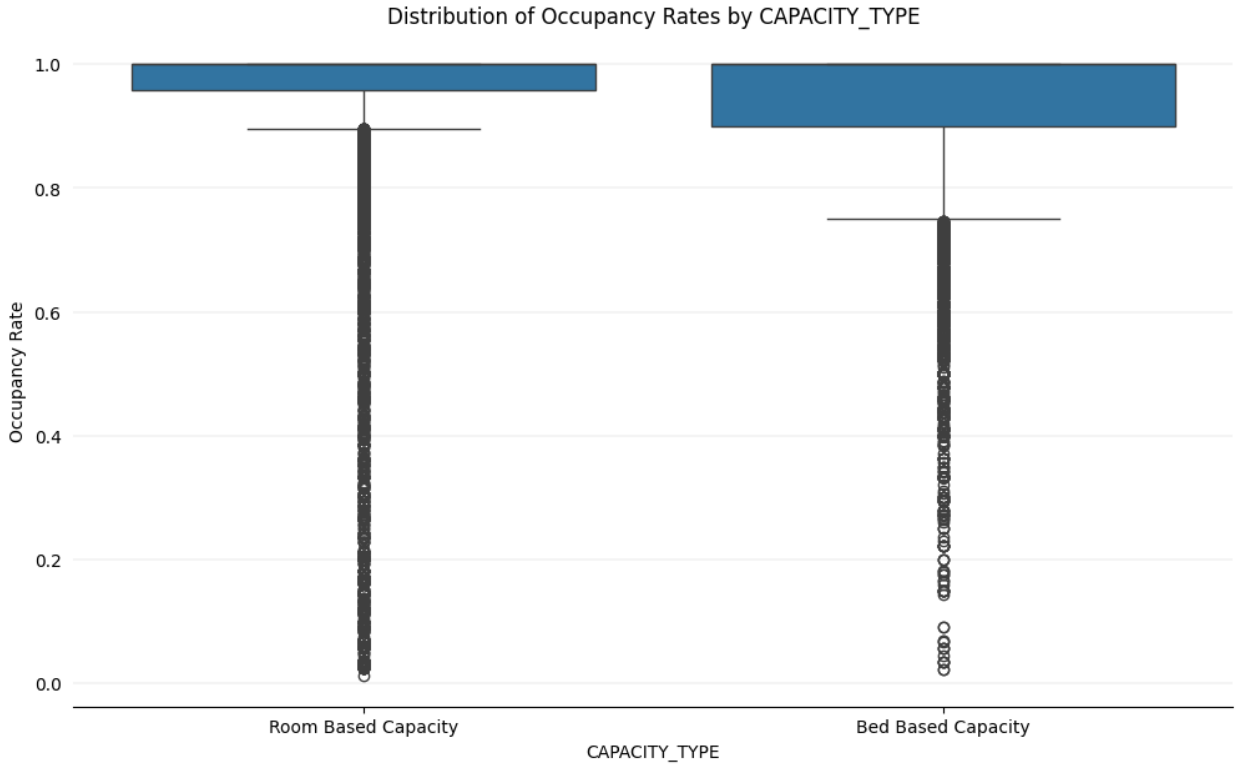

Figure 5: Boxplots of OCCUPANCY_RATE by CAPACITY_TYPE

We can also consider environmental factors such as weather conditions by plotting these differences across time, with month serving as a proxy for general climate.

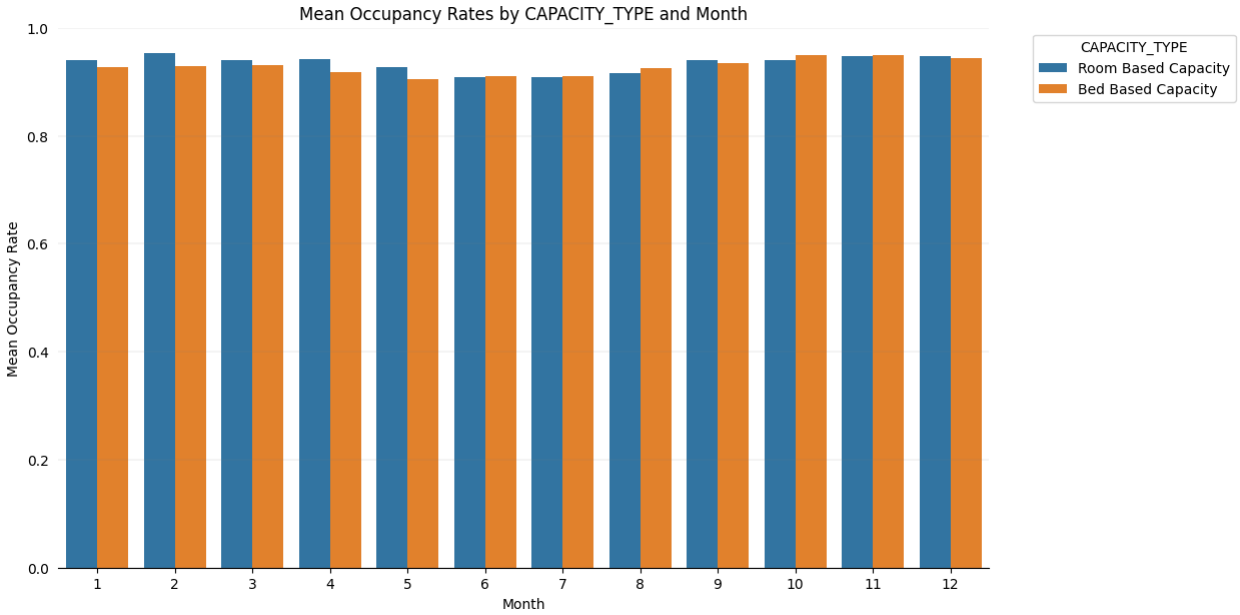

Figure 6: Mean OCCUPANCY_RATE by CAPACITY_TYPE over Time

We can further quantify this preference by calculating mean OCCUPANCY_RATE for each capacity type for different SECTOR classes (Figure 7).

CAPACITY_TYPE	Bed Based Capacity	Room Based Capacity
SECTOR		
Families	0.788377	0.930657
Men	0.960008	0.996345
Mixed Adult	0.920785	0.899819
Women	0.939212	0.969882
Youth	0.880683	0.945445

Figure 7: Pivot Table of mean OCCUPANCY_RATE by SECTOR and CAPACITY_TYPE

The pivot table above displays some level of differences in the occupancies of services based on capacity type by different demographics. Notably, families show a great preference for room-based capacity services, which makes intuitive sense.

In order to gain a more confident understanding of the differences in service demand, we can conduct T-Tests to gauge the observed differences in mean OCCUPANCY_RATE across the different classes. Before conducting a T-Test we need to check for independence, normality, and equal variances. Given the large sample size of observations across different programs and shelter, we can generally assume that one day's observations are independent of another. As observed in the boxplots, occupancy rates are generally heavily skewed towards 1 with a wide range of outliers falling below it. The distributions are not normal. Furthermore, we are unable to state that there are equal variances in occupancy rates across the different groups we intend to observe. Given that we fail normality and equal variances, we should consider using a Welch's T-Test. Welch's T-Test does not assume equal variances, and it is more robust to distributions differing from a

normal one. In addition to the large sample size, trimming the observations to the IQR could help account for the heavily skewed distributions.

```
Classes with statistically significant differences between capacity types:

Families (P-Value: 1.071241219604091e-200)
Mixed Adult (P-Value: 9.06778874233819e-91)
Men (P-Value: 4.309098186784349e-46)
Women (P-Value: 7.259738868334228e-55)
Youth (P-Value: 2.1885743598728367e-76)
Emergency (P-Value: 3.4076023397627547e-47)
Transitional (P-Value: 1.6852906550908115e-103)
Motel/Hotel Shelter (P-Value: 2.3062042116848434e-55)
Shelter (P-Value: 2.4468835591529988e-09)
COVID-19 Response (P-Value: 7.675627869347647e-28)
```

Figure 8: Groups Displaying Statistically Significant Differences in Mean OCCUPANCY_RATE by CAPACITY_TYPE

Based on the results shown in Figure 8, we once again see that age/sex demographics tend to play heavily into a user's choice of service. Furthermore, we see key program types such as emergency and COVID-19 Response displaying heavy preference. In the case of the latter, we can assume additional factors such as health concerns may play a considerable role.

Overall, the findings of the analysis suggest that there is merit in pursuing further research towards identifying preferences for service users. As people in need, they face a level of hardship and undue stress that could result in further harm. Identifying comfort zones for these individuals could help promote the use of such services to those that truly need them, as well as helping identify areas for improved focus and efficiency for planners and regulators.

In terms of limitations the dataset is quite limited in terms of the scope of information provided. As discussed above, SERVICE_USER_COUNT is a difficult metric to judge independently due to the lack of information on other factors that could influence such numbers – factors that are not directly related to the service users themselves. Additionally, as the dataset only covers a single year, it is impractical to attempt to extrapolate results as we may need to consider how factors specific to that time could have influenced the data. For further steps, I would suggest expanding the dataset in terms of time and number of services observed. In terms of my specific research motivations, I would pursue additional experiments concerning the service users themselves, such as conducting interviews to gauge satisfaction rates and areas of need.