Technical Assignment 1

Ying Du (Amelia)

Faculty of Information, University of Toronto

INF2178 LEC0101

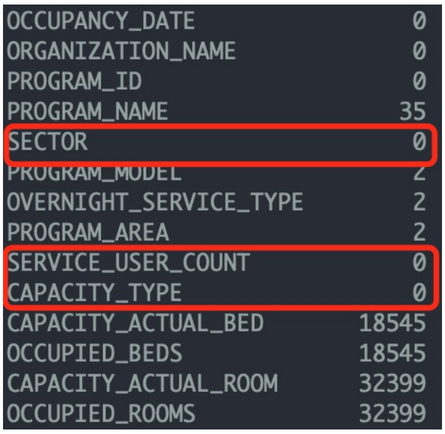
Shion Guha

January 24, 2024

In this data analysis, my primary focus will be on examining key indicators of shelter usage. The columns I will concentrate on include: 'SECTOR', 'PROGRAM_MODEL', 'SERVICE_USER_COUNT', 'CAPACITY_TYPE', 'CAPACITY_ACTUAL_BED', 'OCCUPIED_BEDS', 'CAPACITY_ACTUAL_ROOM', and 'OCCUPIED_ROOMS'. These variables are essential in understanding the dynamics of shelter occupancy and utilization.

Data Examination

My initial step in this analytical process will be to meticulously filter and check any missing data from all categories, especially my focused key indicators.



```
OCCUPANCY_DATE              0
ORGANIZATION_NAME           0
PROGRAM_ID                  0
PROGRAM_NAME               35
SECTOR                      0
PROGRAM_MODEL               2
OVERNIGHT_SERVICE_TYPE      2
PROGRAM_AREA                2
SERVICE_USER_COUNT          0
CAPACITY_TYPE               0
CAPACITY_ACTUAL_BED     18545
OCCUPIED_BEDS           18545
CAPACITY_ACTUAL_ROOM    32399
OCCUPIED_ROOMS          32399
```

Figure1. Missing Data Summary

Figure 1 shows that the non-missing and meaningful categories include ‘SECTOR’, ‘SERVICE_USER_COUNT’, and ‘CAPACITY_TYPE’.

I calculated the occupancy rates of different capacity types. I observed that the average occupancy rates are remarkably high. For bed-based capacity, the average occupancy rate is around 92.8%, and for room-based capacity, it is slightly higher at around 93.4%.

T-Test Analysis

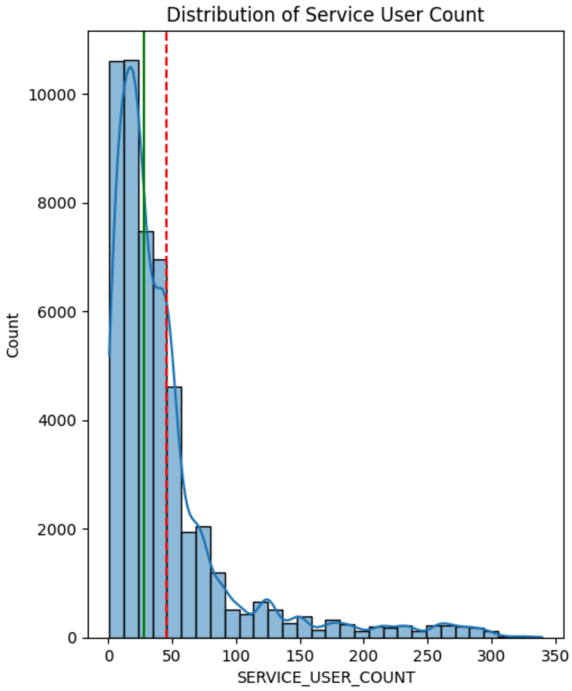
In this T-Test analysis, I initially focused on understanding the impact of different capacity types on service user counts. Utilizing the filtered categories, the analysis revealed a p-value of 0. This result indicates a statistically significant difference in the ‘SERVICE_USER_COUNT’ between the various ‘CAPACITY_TYPE’ types. A notable positive t-statistic was observed, suggesting that the mean value for Room-Based Capacity is higher than that for Bed-Based Capacity. The significant t-statistic value of 78.51 implies that this difference is not only statistically significant but also potentially of practical importance.

Comparison	T-Statistic	P-Value
Families vs Mixed Adult	15.12	4.69e-51
Families vs Men	37.26	2.14e-276
Families vs Women	49.08	0.00
Families vs Youth	58.35	0.00
Mixed Adult vs Men	35.15	5.15e-264
Mixed Adult vs Women	57.39	0.00
Mixed Adult vs Youth	75.08	0.00
Men vs Women	30.50	1.05e-199
Men vs Youth	59.48	0.00
Women vs Youth	38.43	2.33e-309

Figure2. T-test Result for Sector

Subsequently, my analysis shifted towards examining how various ‘SECTOR’ types influence ‘SERVICE_USER_COUNT’. The results in Figure 2 showed that all pairwise comparisons have extremely small p-values, confirming statistically significant differences in mean values between the groups. Notably, the sectors labelled 'Families' and 'Mixed Adult' consistently exhibited higher mean values compared to other sectors in all comparisons. The magnitude of these differences varied, with some of the largest differences observed in comparisons involving the 'Youth' sector. This suggests significantly lower mean values for the 'Youth' sector compared to others. The considerable magnitudes of the t-statistics across these comparisons reinforce the conclusion that these differences are not only statistically significant but also practically significant, indicating meaningful disparities in service user counts across these sector groups.

Exploratory Data Analysis (EDA)



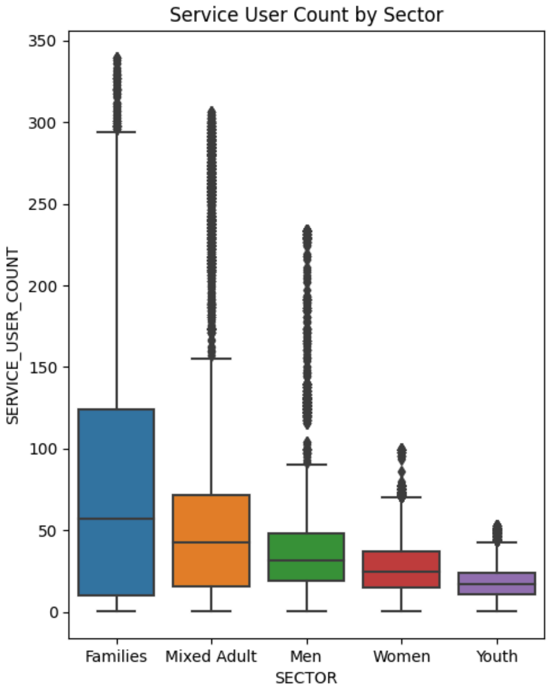


Figure3. Distribution of Service User Count

Figure4. Service User Count by Sector

From the histogram of Service User Counts in Figure 3, we can see a central tendency with a mean(in red) of approximately 45.73 and a median(in green) of 28. This disparity between the mean and median points to a right-skewed distribution, suggests that while the majority of shelters have lower user counts, there are a notable few with significantly higher counts, thereby elevating the average. The distribution's pronounced right skewness is further evidenced by a long tail extending towards higher service user counts, indicating the presence of outliers - shelters with exceptionally high user numbers. In conclusion, we can say that most shelters have low to moderate service user counts, with a smaller proportion experiencing very high counts.

Figure 4 shows the boxplot by sector, and we can see there are distinct patterns that emerged across different shelter sectors. Compared to other sectors, 'Families' and 'Mixed Adult' show broader interquartile ranges (IQR) in service user counts, suggesting that the middle 50% of the service user counts are spread out over a wide range of values. The spread suggests a diverse set of families and mixed adult shelters in terms of service user count. The century tendency lines of ‘Families’, ‘Mixed Adult’, ‘Men’, and ‘Women’ toward the lower quartile of the boxes indicate that the majority of shelters in those sectors cater to fewer users than the overall median. The ‘Youth’ ’sector is marked by the tightest spread and fewest outliers, indicating a uniformity in user counts among youth shelters. The median is centrally located within the IQR, pointing to a more symmetrical distribution within this sector. The narrow spread in the youth sector could reflect either a standardized service provision across these shelters or a lower demand, aligning with the high average occupancy rates calculated above.

I also noticed that there are numerous outliers exceeding the upper whisker for all the sectors, which also aligns with the conclusion from the histogram that smaller proportions are experiencing very high counts.

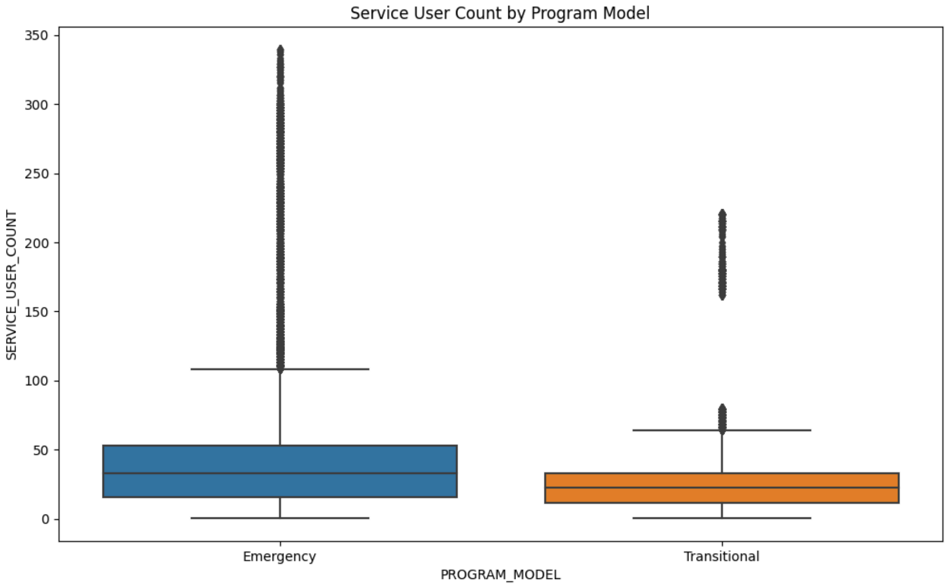


Figure5. Service User Count by Program Model

My intention was to employ a t-test to investigate whether different program models influence the service user counts. Nonetheless, the sample sizes for the two groups, 'Emergency' with 41,541 and 'Transitional' with 9,401, are notably imbalanced, which could potentially compromise the reliability of the test results due to unequal variance.

In Figure 5, the 'Emergency' shelters' service user counts exhibit a broader spread, signifying a greater variation in the number of service users across these shelters. Conversely, the 'Transitional' shelters demonstrate a central tendency that is skewed slightly towards the upper quartile. This indicates that a larger proportion of 'Transitional' shelters serve a higher number of users relative to the median of the overall data. Additionally, the 'Transitional' group has fewer outliers and a narrower IQR, suggesting that these shelters operate with a more consistent number of service users. This could imply that 'Transitional' shelters have a more uniform level of operation, possibly due to standardized processes.

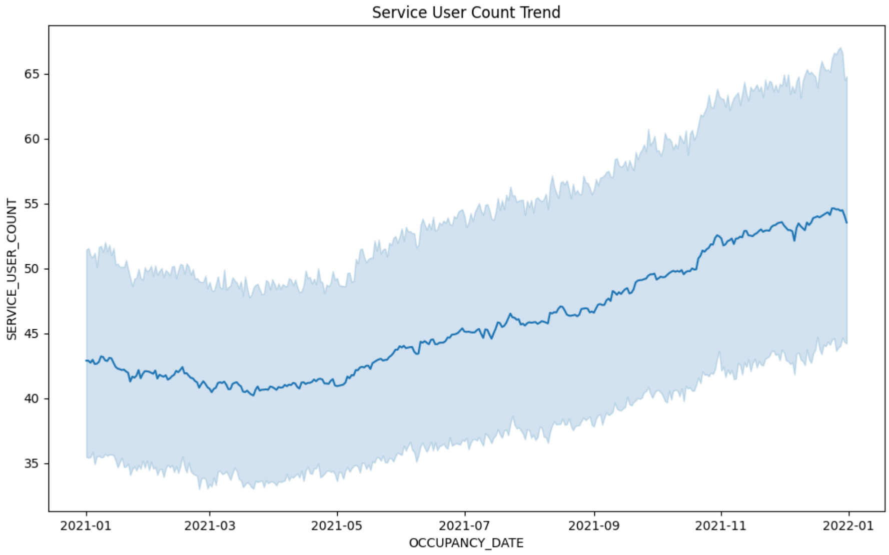


Figure6. Service User Count Trend

In conclusion, my analysis also focuses on the trends in service user counts over time. Figure 6 presents a line plot that tracks the changes in 'SERVICE_USER_COUNT'. The visual representation in this diagram clearly indicates an upward trend starting from May 2021. This observation suggests that there has been a consistent increase in the number of users at shelters since that time.

Data Reference

Python tutorials: Learn, build, & practice python programming. DataCamp. (n.d.). https://www.datacamp.com/tutorial/category/python