University of Toronto

INF2178

Experimental Design for Data Science

Technical Assignment I (EDA)

Instructor

Professor Shion Guha

Submitted by

Lam Hong Kevin Ching

1009243043

2024/02/04

Introduction

The dataset from the City of Toronto's Open Data portal represents a granular snapshot of the daily occupancy and capacity within the city's shelter system. It encompasses data on various organizations providing shelter services, detailed by program names and models such as emergency and transitional shelters. The data includes specific service areas like COVID-19 response, capturing occupancy rates, the number of service users, and the available capacity, whether room-based or bed-based. This information is essential for understanding the dynamics of shelter use and availability, reflecting the city's efforts to accommodate vulnerable populations and manage resources effectively. The data's regular updates ensure a current view of the shelter system, facilitating informed decision-making and policy development to support those in need of these critical services.

Research Objects

The primary objective of the research is to investigate the impact of factors such as SECTOR and CAPACITY_TYPE on the SERVICE_USER_COUNT within Toronto's shelter system, as detailed in the dataset from the City of Toronto's Open Data portal. The study seeks to discern patterns and correlations that indicate how different sectors (like families, mixed adults, men, youths, and women) and capacity types (room-based or bed-based) influence the number of service users. Furthermore, the research aims to explore the implications of PROGRAM_MODEL on SERVICE_USER_COUNT, aiming to uncover insights into how the emergency and transitional models impact on the number of service users.

Data Cleaning

The dataset read from the CSV file using the pandas package in Python comprises 50,944 rows, each corresponding to a data entry, and 14 variables that represent different attributes related to the shelter system's daily occupancy and capacity. These variables include date, organizational information, program details, sector categorization, service models, types of capacity, and actual usage statistics. This comprehensive dataset provides a foundation for an in-depth analysis of shelter services in Toronto.

When scanning the variables, the unique identifiers such as ORGANIZATION_NAME, PROGRAM_ID, and PROGRAM_NAME were deemed unhelpful for the purpose of discerning general patterns over the year. To streamline the analysis and focus on broader trends rather than specific events, the decision was made to exclude these columns from the dataset.

Upon refining the dataset and examining it for null values, it was observed that there were negligible missing values in columns such as PROGRAM_MODEL, OVERNIGHT_SERVICE_TYPE, and PROGRAM_AREA. However, substantial missing values were noted in the capacity-related columns. Specifically, there were 18,545 missing entries for both CAPACITY_ACTUAL_BED and OCCUPIED_BEDS, and 32,399 for CAPACITY_ACTUAL_ROOM and OCCUPIED_ROOMS.

Data Visualization

The project is primarily designed to investigate the variables affecting the number of service users in Toronto's shelter system. From the descriptive statistics (Table 1), the Families sector shows a higher average service user count (approximately 79.65) compared to the Individuals sector (which includes 'Men', 'Women', 'Mixed Adult', 'Youth') with an average of about 41.50. This suggests that family-oriented services tend to accommodate larger groups per service instance, potentially due to families being housed together. The standard deviation is also higher in the Families sector (approximately 76.73) than in the Individuals sector (around 47.98), indicating more variability in the number of service users among family services. The minimum count of service users is 1 for both sectors, but the maximum is significantly higher for the Families sector at 339, compared to 306 for the Individuals sector. This range suggests that the Families sector has instances of extremely high occupancy, which could be indicative of larger family units or perhaps a consolidation of services. The median (50th percentile) value for the Families sector is greater than that of the Individuals sector, indicating that the typical service instance for families involves more users.

	FAMILIES SECTOR	**INDIVIDUAL SECTOR**
MEAN	79.65	41.50
STD	76.73	47.99

MIN	1	1
25%	10	15
50%	57	27
75%	124	48
MAX	339	306

Table 1: Descriptive Statistics for Families and Individuals Sector

The histograms (Figure1) for the Families and Individuals sectors show the distribution of service user count. The Families histogram has a right-skewed distribution, indicating a concentration of lower service user counts with fewer instances of high user counts, consistent with the higher mean and maximum values in the descriptive statistics. The Individuals histogram shows a more pronounced peak at the lower end, with a rapid decline in frequency as the service user count increases, reflecting a lower average user count and less variability. Both distributions have long tails to the right, but the tail is more pronounced for the Families sector, corroborating the higher variability in family service usage.

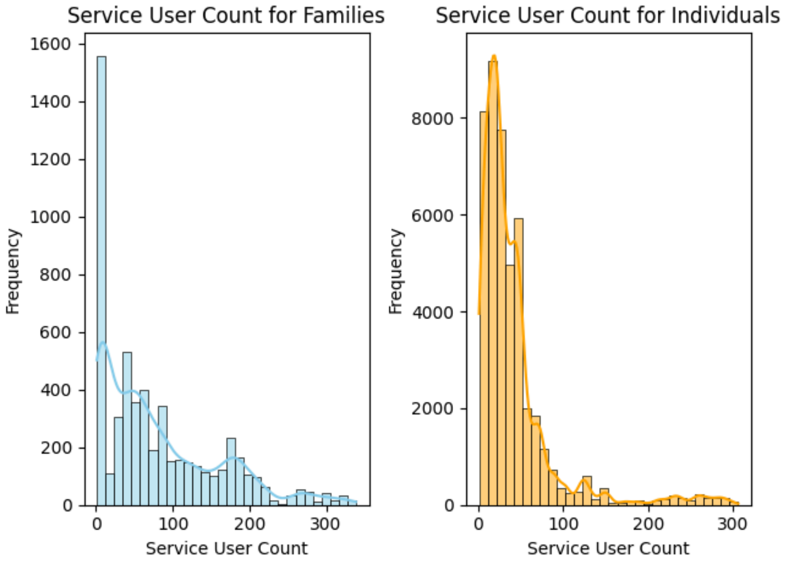

Figure 1: Service User Count for Families and Individuals.

Meanwhile the descriptive statistics for the Emergency and Transitional program models (Table 2) reveal distinct patterns in service utilization. Emergency shelters exhibit a higher average service user count, with a mean of approximately 49 individuals, compared to around 31 individuals for Transitional shelters. This difference signifies that Emergency services, typically designed to provide immediate, short-term relief, accommodate more individuals at any given time. The variation in user count is also more pronounced in Emergency shelters, as indicated by a higher standard deviation, suggesting that the number of individuals seeking emergency shelter can fluctuate greatly. The maximum service user count for both models indicates that there are instances when shelters operate at or near full capacity, but this is more common in the Emergency model. The median values corroborate that Emergency shelters generally serve more individuals than Transitional shelters, which aim to provide a more stable and longer-term solution to individuals in need.

In combination with the boxplot (Figure 2), Emergency models reveals a higher median and greater interquartile range, indicative of more users on average and greater variability in user numbers compared to Transitional models. The presence of numerous outliers in the Emergency model suggests instances of exceptionally high service user counts, aligning with the higher maximums observed in the descriptive statistics. The Transitional model shows a more compact distribution with fewer and lower outliers, reflecting a generally lower and more consistent service user count.

	EMERGENCY MODEL	TRANSITIONAL MODEL
MEAN	49.06	30.99
STD	55.91	36.43
MIN	1	1
25%	16	12
50%	33	23
75%	53	33
MAX	339	221

Table 1: Descriptive Statistics for Families and Individuals Sector

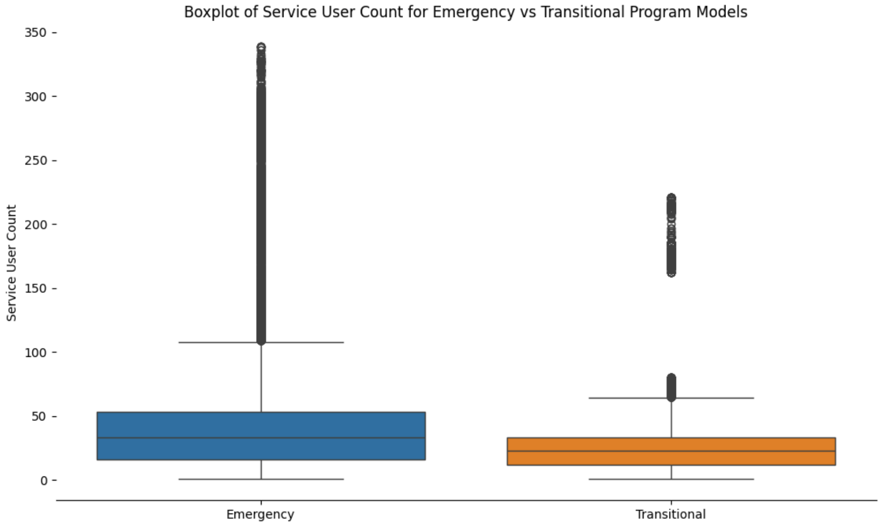

Figure 2: Boxplot of Service User Count for Emergency vs Transitional Program Models

T-tests

T-tests will statistically evaluate these hypotheses to determine if the differences in means are statistically significant, thereby influencing the service user count over the year. The hypothesis sets out to investigate the variance in service user counts between two distinct sectors: families and individuals.

> *H0: There is no significant difference in the mean service user count between family shelters and shelters for individuals.*
> *H1: There is a significant difference in the mean service user count between family shelters and shelters for individuals.*

According to the t-test results, with a statistic of approximately 36.49 and a p-value essentially at 0, strongly suggest rejecting the null hypothesis (H0). This implies there is a statistically significant difference in the mean service user count between family shelters and shelters for individuals. The large t-statistic indicates a considerable difference between the means of the two groups, and the extremely small p-value confirms that this difference is unlikely to have occurred by chance. The degree of freedom (df) for this test is around 6210, which is high and indicates a large sample size, lending credibility to the test's results. This analysis provides robust evidence that the demographic sector significantly affects service user counts in shelters.

To explore whether Emergency and Transitional program model would impact on service user count, the following hypothesis are set.

> *H0: There is no significant difference in the mean service user count between the emergency and transitional program models.*
> *H1: There is a significant difference in the mean service user count between the emergency and transitional program models.*

The t-test comparing Emergency and Transitional program models yields a t-statistic of approximately 38.85 and an extremely low p-value (essentially 0). This result leads to the rejection of the null hypothesis, indicating a statistically significant difference in the mean service user count between Emergency and Transitional programs. The high t-statistic suggests a substantial difference between the two program models' service user counts. With the degrees of freedom around 20758, the test has high reliability due to the large sample size. These results suggest that the type of program model—Emergency or Transitional—has a distinct impact on the number of service users in the shelters.

Conclusion

Ultimately, this report's analysis of the City of Toronto's shelter system provides crucial insights into the sector-specific and program model variances in service user counts. The findings from the t-tests reveal notable differences between the families and individuals sectors, and between emergency and transitional shelters. These variations highlight the diverse needs within the shelter-seeking population and underscore the importance of targeted service provision. The report not only contributes to a better understanding of the complexities within the shelter system but also aids in informing policy development and resource allocation. By acknowledging the distinct patterns of shelter use, stakeholders can better tailor their strategies to address the unique requirements of different groups effectively. In conclusion, this study serves as a valuable resource for ongoing efforts to optimize the shelter system in Toronto, ensuring that the needs of all individuals and families seeking shelter are met efficiently and compassionately.