Experimental Design for Data Science
INF 2178 Technical Assignment
Nargiz Guliyeva

The increasing homelessness crisis in Toronto emphasizes the pressing need for effective interventions. This analysis aims to explore the dataset to define the research questions, find insights, and determine the significant differences between models and capacities of homeless shelters. Considering that the dataset has been significantly skewed, a non-parametric test has been employed to compare service users by program model and capacity type. Both tests suggested the result that reject null hypothesis.

INTRODUCTION

The rise of the homeless population highlights the urgent need for effective interventions. According to City of Toronto, more than 8,000 people experiencing homelessness in Toronto (City of Toronto, (2023)

As the demand for shelter services increases, the support systems face challenges in accommodating the needs of families and individuals. The Shelter, Support & Housing Administration division of the City of Toronto delivers housing and homelessness services in partnership with community agencies, which aim to prevent homelessness by helping people to access emergency shelters and to find and keep housing (Jadidzadeh, and Kneebone, 2018).

The presented dataset contains information required to examine organizations, their programs, and features, and additionally, it also provides information on the occupancy, the capacity and other characteristics.

METHODOLOGY

The Exploratory Data Analysis (EDA) and t-test have been employed on the dataset. Preceding EDA, data preprocessing steps have been executed. Three additional variables have been derived from the existing ones: 'month,' indicating the monthly data; 'combined capacity' – merged bed and room capacity; and 'combined occupied' - merged bed and room occupancy.

To address missing variables, a few datapoints under categorical variables, have been dropped. Numerical variables (CAPACITY_ACTUAL_BED, OCCUPIED_BEDS, CAPACITY_ACTUAL_ROOM, OCCUPIED_ROOMS) complement each other, so the missing values under these four variables have been filled with zero. Additionally, combined capacity and occupancy columns have been created to consolidate the values for these variables.

Graphical and non-graphical analysis were conducted to examine both categorical and numerical variables. The insights from those analyses have been summarized under the "finding" sections in the Python notebook. For graphical univariate and multivariate analysis, bar charts, heatmaps, boxplots, and histograms have been used.

The distribution for each of the numerical variables has been plotted. The histograms for the float variables have been highly skewed, and none of them followed a Gaussian distribution. The attempt has been made to normalize data using Quantile Transformation, which is advantageous in handling skewed data; however, the data has been highly distorted and the new distributions do not resemble a normal distribution.

In this regard, the dataset does not meet all the assumptions for the independent t-test, i.e., Independent and Identically distributed, Gaussian distributed, equal variance, continuous variable, and increase the possibility of Type I error (Herzog et al., 2019). Given the violation

of the assumptions and the likelihood of Type I error, the Mann-Whitney, a non-parametric test, has been used to compare program models and capacity types based on the number of service users (West, 2021; Herzog et al., 2019).

RESULTS

The analysis extensively explored the dataset, and it has been reflected in the Python notebook. Only key results have been included in this paper. The dataset recorded data between January 1, 2021, and December 31, 2021. The dataset has 13 columns and 50,944 rows. Originally, the dataset had one time-related variable, 7 object variables, 2 integers, and 4 float variables. As discussed above, three additional variables have been derived.

The emergency programs substantially prevail over transitional programs, and the number of programs that offer bed-based capacity is almost twice as high as room-based capacity (Figure 1, Figure 2). Program model and capacity type have been further explored by adding numerical variables on user count, occupancy, and capacity.

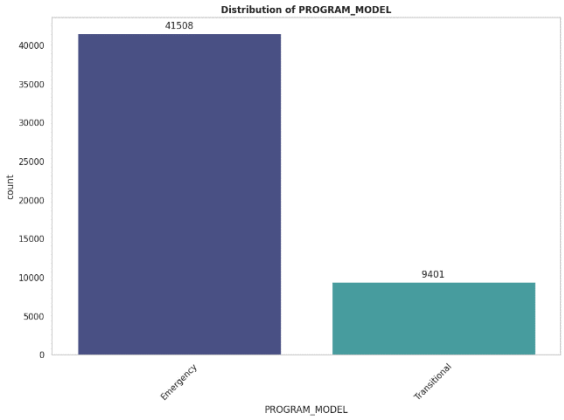


Figure 1.

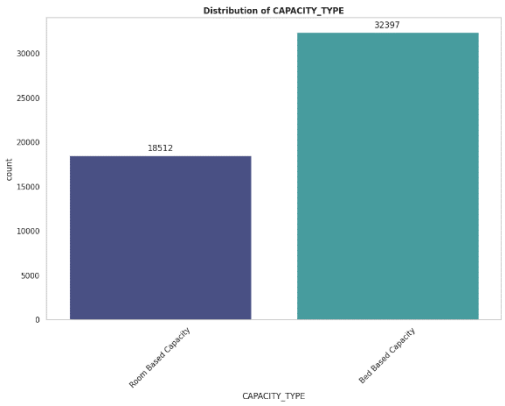


Figure 2

The analysis shows that the use of the room-based capacity emergency capacity is much higher than any other categories (Figure 3). Based on the findings key research questions have been formulated: (1) Is there a statistically significant difference in the service users between emergency and transitional program models in the homeless shelters? (2) Is there a significant difference in the service user counts between shelter programs with Bed-Based Capacity and those with Room-Based Capacity?

Considering that the data is skewed, the Mann-Whitney U statistic non-parametric test has been used instead of t-test. The results of it have been presented in the next section of the paper.

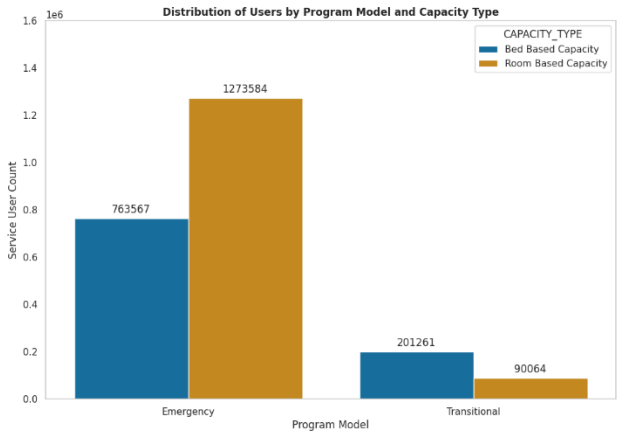


Figure 3

According to the analysis, the top 10 organizations provide support to 85% of the clients, with the top 3 organizations being the City of Toronto, Homes First Society, and the Salvation Army of Canada (Figure 4) Upon identifying the City of Toronto as the leader in providing services, further analysis has been conducted to identify successful programs in the City of Toronto. Based on the boxplot, Birkdale Residence Bedded Program, Family Residence - Main Program, and SSHA Downtown Women Program are leading in the number of individuals served (Figure5).

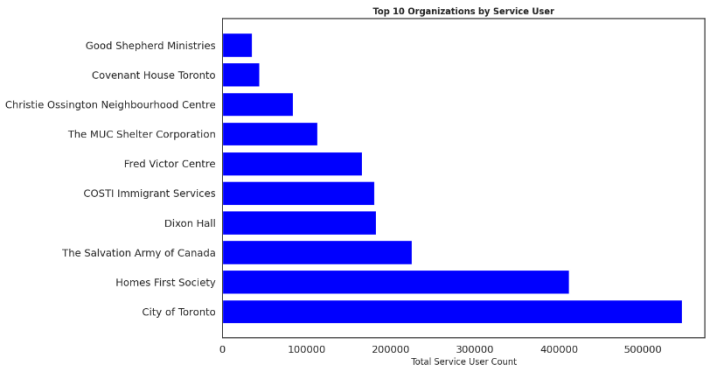


Figure 4

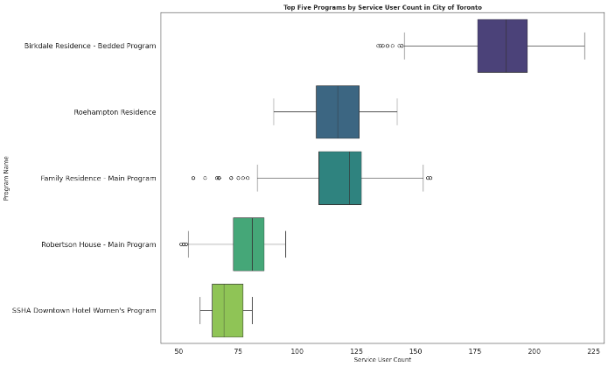


Figure 5

The majority of the organizations provide emergency services, except for MUC Shelter Corporation, where the transitional program prevails, and Good Shepherd Ministries, which has an equal capacity for emergency and transitional programs (Figure 6). Another finding is that most services are provided to mixed adults, while a few organizations specialize in serving specific groups. For example, Covenant House Toronto provides services to youth, and Good Shepherd Ministries provides services to men (Figure 7).

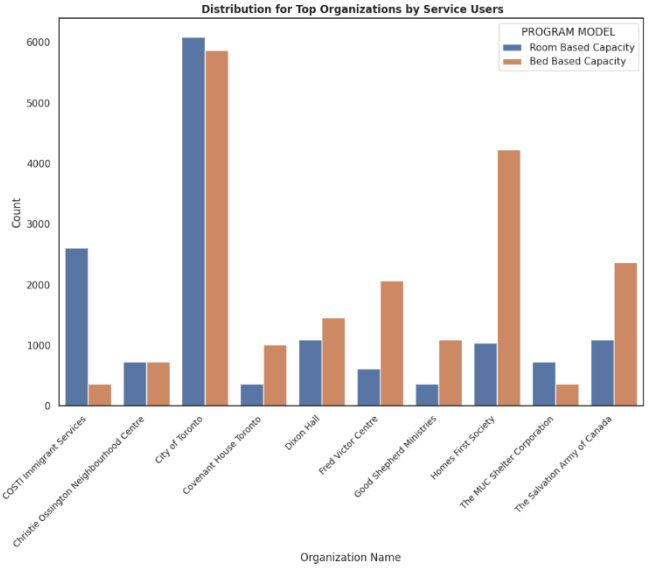


Figure 6

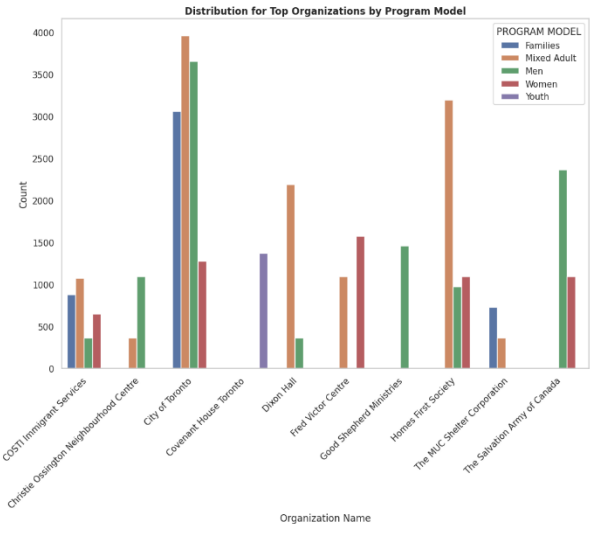


Figure 7

The organizations and their programs can be further reviewed if matched with the efficiency data or qualitative data is collected to complement these findings. The potential research area can be identifying features of successful interventions and scaling them further.

Most organizations primarily provide services for mixed adults, specifically co-ed or all-gender facilities, followed by services for men, women, youth, and families. The table and bar plot illustrates that the most utilized service category, comprising 28%, is mixed adult room capacity. This is followed by family's room-based capacity at 19% and men bed-based capacity at 15%. (Figure 8, Table 1).

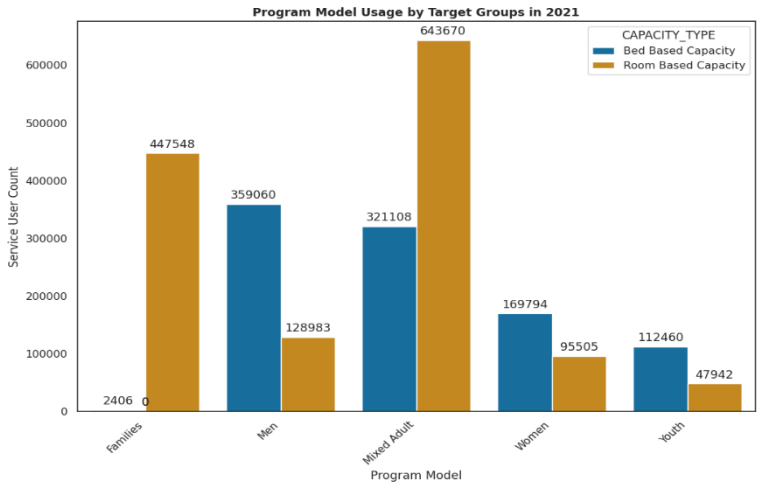


Figure 8

	SECTOR	CAPACITY_TYPE	SERVICE_USER_COUNT	Percentage_Total	Percentage_INGroup
0	Families	Bed Based Capacity	2406	0.0	1.0
1	Families	Room Based Capacity	447548	19.0	99.0
2	Men	Bed Based Capacity	359060	15.0	74.0
3	Men	Room Based Capacity	128983	6.0	26.0
4	Mixed Adult	Bed Based Capacity	321108	14.0	33.0
5	Mixed Adult	Room Based Capacity	643670	28.0	67.0
6	Women	Bed Based Capacity	169794	7.0	64.0
7	Women	Room Based Capacity	95505	4.0	36.0
8	Youth	Bed Based Capacity	112460	5.0	70.0
9	Youth	Room Based Capacity	47942	2.0	30.0

Table 1

The emergency model for the mixed adult has the highest occupancy followed by men and women. The statistical significance of this difference can further explored with ANOVA and other statistical tests.

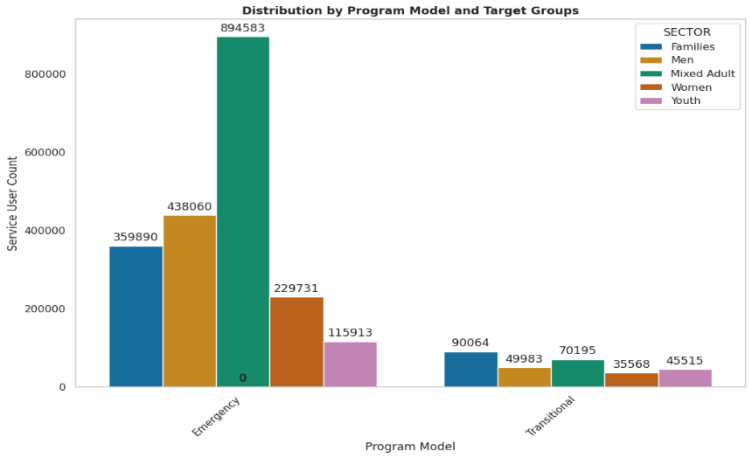
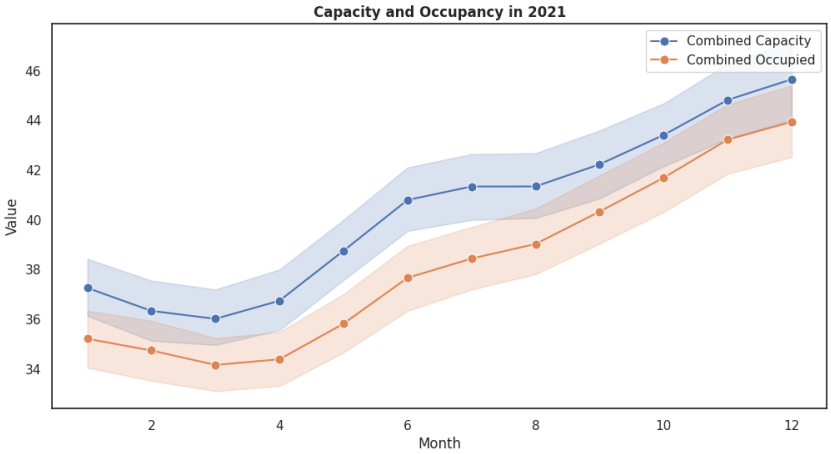


Figure 9

The capacity of the organizations is not completely utilized, as on average, 5% of the spaces (beds or rooms) are not filled. The highest unutilized capacity is observed in June, with an unutilized capacity of 7.7%, and the lowest is in November, with 3.6% unused capacity (Figure 9, Table 2). The further analysis can help to identify the programs with the unutilized capacity and optimize the number of the intakes in the shelter.



MONTH	
1	5.5
2	4.4
3	5.1
4	6.4
5	7.6
6	7.7
7	7.0
8	5.6
9	4.5
10	4.0
11	3.6
12	3.7

Figure 10 **Table 2**

The median and IQR data for bed-room capacity-occupancy and user count are almost equal while the total of the user counts are substantially higher (Figure 10). Additionally, service users numbers are much more skewed. It is important to highlight the difference between combined occupancy and service user counts: combined occupancy reflects the occupancy number of beds and rooms, while service user counts represent the occupancy number of individuals. Considering that some rooms can accommodate more than 1 individual, the number of users is higher.

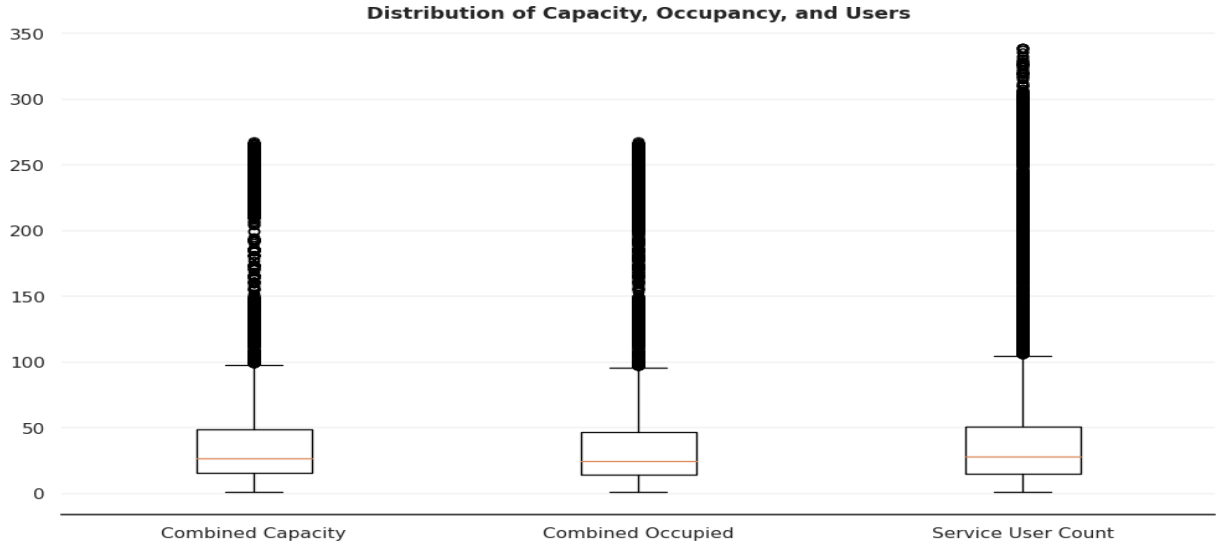


Figure 11

Statistical test

Based on the plotted histograms, none of the numerical variables are equally distributed. As discussed in the 'Method' section this violates the assumption for the t-test and increases the likelihood of Type I error (Figure 11). Considering this, Mann-Whitney U statistic has been applied instead of t-test.

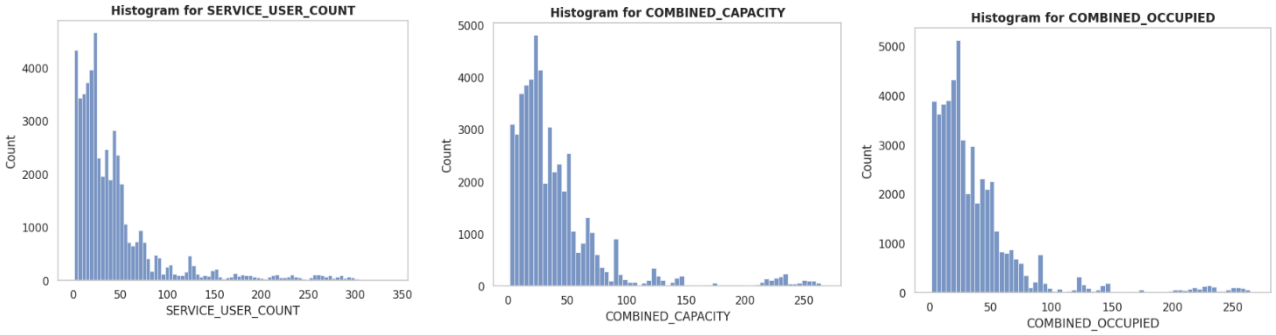


Figure 12

The first question is whether there is statistically significant difference in the service users between emergency and transitional program models in the homeless shelters. The calculated Mann-Whitney U statistic is 241,789,588.5, and the associated p-value is approximately 2.78 * 10^ (− 288). The p value is extremely low; therefore, the null hypothesis, i.e. there is no significant difference in the occupancy rate between Emergency and Transitional program models, has been rejected.

The second question is whether there is a significant difference in the service user counts between shelter programs with Bed-Based Capacity and those with Room-Based Capacity. The Mann-Whitney U statistic is 181,722,970, and the p-value associated with this test is reported as 0.0. Given that the p-value is less than the significance level (p < alpha), the null hypothesis (There is no significant difference in the service user counts between shelter programs with Bed-Based Capacity and those with Room-Based Capacity) is rejected.

DISCUSSION AND LIMITATIONS

Several limitations should be considered in interpreting the findings of this analysis and further expanding the research. Moreover, the non-normal distribution and high skewness of numerical values challenge the assumptions of traditional parametric tests, highlighting the need for careful interpretation. The dataset's inherent biases and limitations in establishing causal relationships emphasize the complexity and overall challenge in drawing definitive conclusions. Despite statistical significance being achieved in hypothesis tests, the practical significance requires careful consideration and further analysis. Lastly, the impact of missing data and the choice of imputation methods on the analysis were not extensively explored, potentially introducing biases.

Further analysis could be significantly enhanced with additional data on the daily number of new admissions. The computation with the current variables can potentially lead to an overcount in the 'service_user_count' variable. This, in turn, affects the accuracy of occupancy measures. Furthermore, a more comprehensive understanding of shelter dynamics could be achieved by incorporating data on the average length of stay per program. This will support in identifying how quickly are individuals discharged from the programs and could as serve additional in defining the success of programs. Finally, other statistical tests, e.g. ANOVA, can be used to identify differences within target groups.

REFERENCE

City of Toronto (2023) Housing & Homelessness Research & Reports, City of Toronto. Available at: https://www.toronto.ca/city-government/data-research-maps/research-reports/housing-and-homelessness-research-and-reports/ (Accessed: 03 February 2024).

Fay, M. P., & Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. Statistics surveys, 4, 1.

Herzog, M. H., Francis, G., & Clarke, A. (2019). Understanding statistics and experimental design: how to not lie with statistics (p. 142). Springer Nature.

Jadidzadeh, A., & Kneebone, R. (2018). Patterns and intensity of use of homeless shelters in Toronto. Canadian Public Policy, 44(4), 342-355.

Mukhiya, S. K., & Ahmed, U. (2020). Hands-On Exploratory Data Analysis with Python: Perform EDA techniques to understand, summarize, and investigate your data. Packt Publishing Ltd.

West, R. M. (2021). Best practice in statistics: Use the Welch t-test when testing the difference between two groups. Annals of clinical biochemistry, 58(4), 267-269.