INF2178 (Winter, 2024)
Michelle Lee (#1001448169)

INTRODUCTION

This report will explain the most significant findings for Assignment 1 for the INF2178H (2024) course. This assignment involves performing basic data analysis techniques as the following sections will outline and demonstrate.

Toronto City is notorious for its inability to address an ongoing homelessness and housing crisis that policy makers have yet to implement appropriate policies to curb the effects of. The dataset that this project uses - the Daily Shelter & Overnight Service Occupancy & Capacity (2021) - compiles information on Toronto's Shelter Support and Housing Administration division's Shelter Management Information System (SMIS) database and provides many variables that can be studied. The first section will provide a preliminary Exploratory data analysis (EDA) to study some variables of interest, before proposing some relevant research questions based on the EDA. Finally, the project will employ statistical methods like T-Tests to investigate the statistical significance between variables, before reporting key findings.

EXPLORATORY DATA ANALYSIS (EDA)

The following section will provide basic data visualizations and subsequent explanations to analyze what are the most significant preliminary findings from our dataset concerning measures such as central tendency and spread of the data points shown.

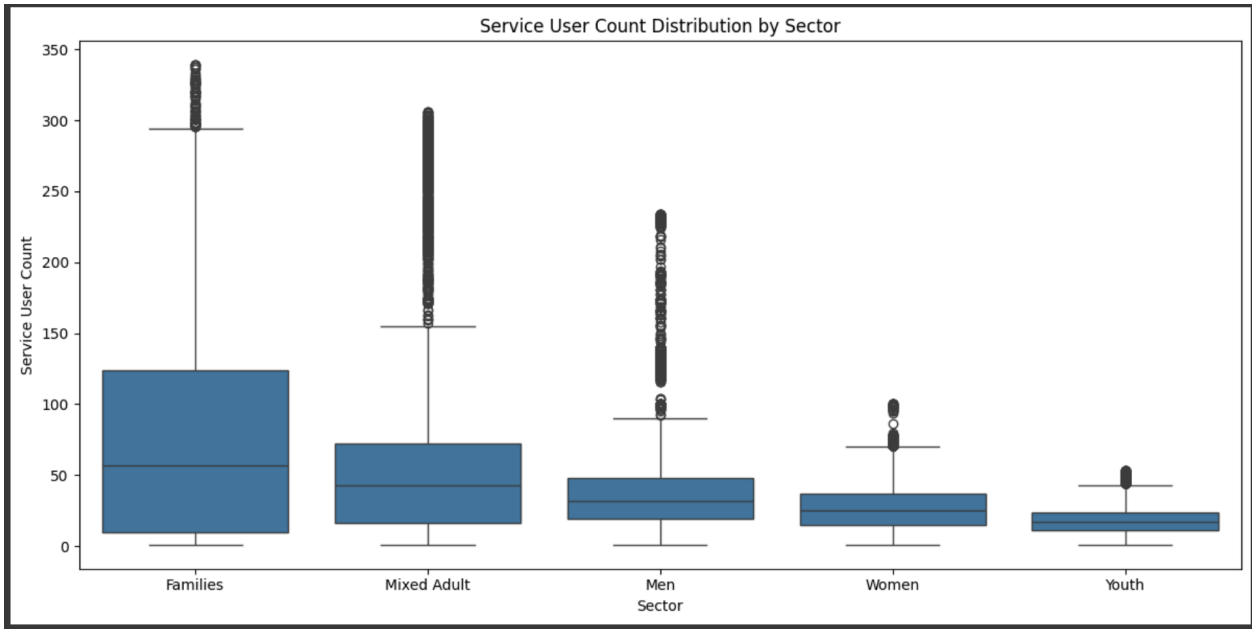

Figure 1: Service User Count Distributed by Sector

- Comment that the dark points are outliers; standardize the USER COUNT Numbers

Figure 1 shows a boxplot that displays the variables 'SERVICE_USER_COUNT' as distributed by 'SECTOR' - it shows the quantity of users using different shelter programs as distributed by sector type. As seen above, the sector type with the highest user count is families, followed by mix adults. The sector types with less user counts appear to be women and youth.

INF2178 (Winter, 2024)
Michelle Lee (#1001448169)

Figure 1's summary statistics are as follows:

Families: Median = 57, Mean ≈ 79.65
Men: Median = 32, Mean ≈ 39.87
Mixed Adult: Median = 43, Mean ≈ 62.11
Women: Median = 25, Mean ≈ 28.66
Youth: Median = 17, Mean ≈ 19.54

The Families sector shows a higher median and mean service user count compared to other sectors, meaning shelters targeting families tend to have a higher number of users overall. The significant difference between the median and mean suggests a right-skewed distribution, which means some family shelters have a significantly large number of users.

Meanwhile, Men and Mixed Adult sectors have relatively high medians and means for service user count. Women and Youth sectors, by comparison, have lower median and mean service user counts, which might mean that shelter programs servicing women and youth tend to serve fewer individuals on average. This might indicate a few possibilities: either there's less demand for or by these groups, or that there are less larger shelters servicing women and youth demographics at current.

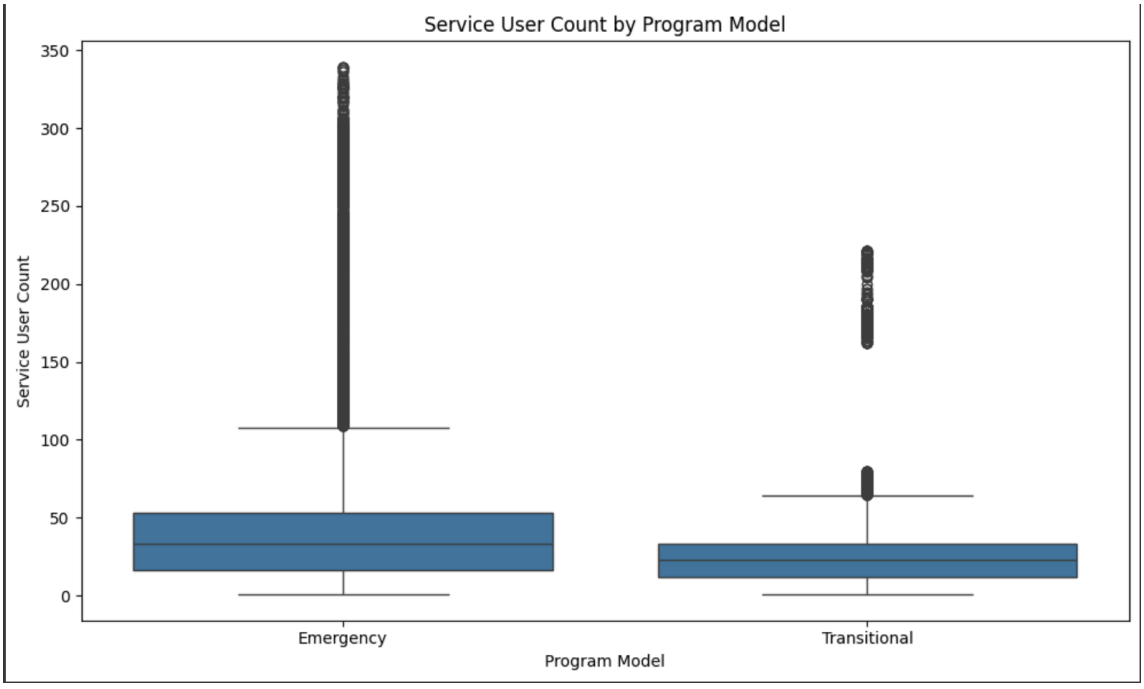

Figure 2: Service User Count by Program Model

Meanwhile, Figure 2 above shows a boxplot displaying 'SERVICE_USER_COUNT' by 'PROGRAM_MODEL'. As seen by the multiple black data points, there are outliers in both 'Emergency' and 'Transitional' program models, with 'Emergency' showing a higher range and potential outliers.

Figure 2's Summary Statistics are as follows:

Emergency Program Model: Median = 33, Mean ≈ 49.06
Transitional Program Model: Median = 23, Mean ≈ 30.99

The median service user count is higher for the Emergency Programs compared to Transitional programs (33 compared with 23 respectively), indicating that overall emergency programs may serve more users during a given period.

The InterQuartile Range (IQR) that measures spread is also wider for the 'Emergency' model, which suggests more variability in service user counts within this category, compared to transitional that has a narrower IQR conversely.

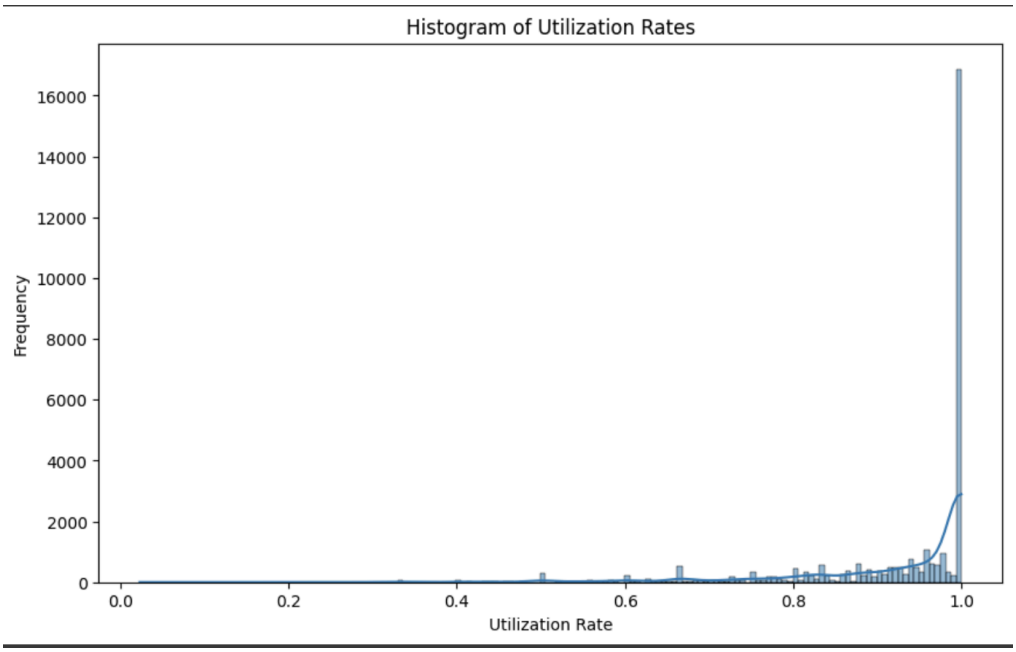

Figure 3: Histogram of Utilization Rates.

Figure 3 above shows a histogram - specifically, a frequency distribution that displays the utilization rate of shelters. The utilization rate is calculated by dividing all the data points for 'OCCUPIED_BEDS' by all the data points of 'CAPACITY_ACTUAL_BED', or:

$$df[UTILIZATION\ RATE] = \frac{df[OCCUPIED\ BEDS]}{df[CAPACITY\ ACTUAL\ BEDS]}$$

The histogram also shows an extreme rightward skew. As seen above, the rightward skew is most significant at a utilization of rate of close to 1.0. This means that, based on this dataset, a large proportion of shelters in Toronto are operating at close to 100% capacity. This is significant as it provides further evidence that supports existing reports and studies that highlight the multi-pronged issues concerning: 1) a decrease in the amount of shelters available, 2) the lack

of additional capacity in existing shelters to accommodate the rising population of homeless people in Toronto.

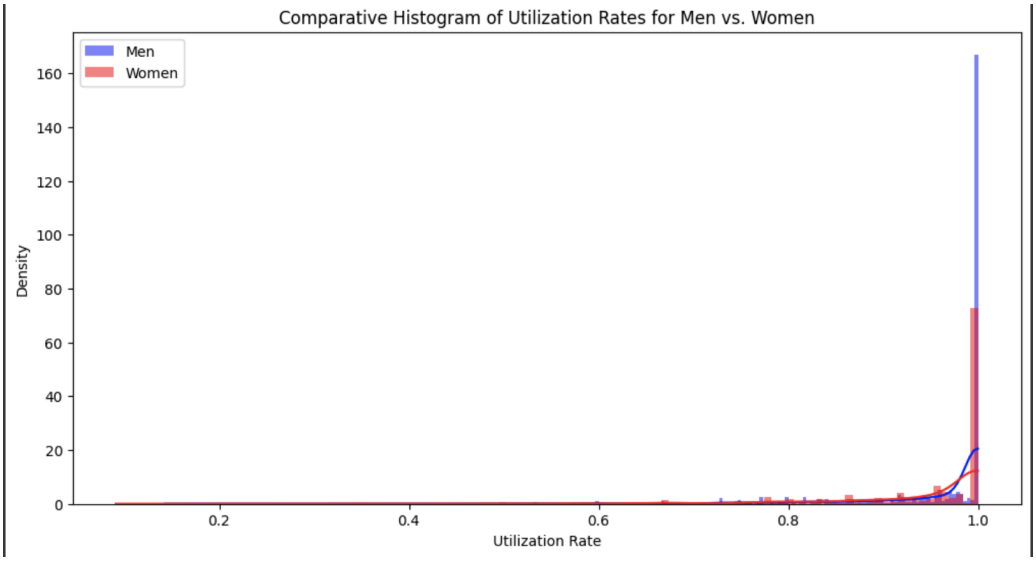

Figure 4: Comparative Histogram - Men vs Women Utilization rates

In addition, figure 4 splits utilization rates based on data available on gender. The summary statistics are as follows:

Men's Median and Mean: 1.0, 0.96
Women's Median and Mean: 1.0, 0.94

This means that both groups have a utilization rate close to 100% irrespective of gender, but the mean being lower than the median may indicate some underutilization in some programs.

RESEARCH QUESTIONS

Based on our EDA, as well as current studies and existing policy and research interests on Toronto's housing and homelessness problems, some more specific research questions that will be addressed in this study include:

1) Are there any statistically significant differences between Emergency and Transitional usage in Shelter Programs when using User Count as a dependent variable?
2) For program sectors, are there any statistical gender disparities between users in utilization rates, specifically between men and women?
3) Does the size of a shelter significantly affect its utilization rate?

The following sections will use T-Tests - a form of to investigate whether or not there are statistical relationships between the variables posed in our research questions.

DATA ANALYSIS

INF2178 (Winter, 2024)
Michelle Lee (#1001448169)

The following will apply T-Test techniques to investigate if there are any statistical significant relationships between variables as proposed in the previous research question portion. Specifically, I will be using *Welch's T-Test* to perform the subsequent tests. The rationale for this is because of the following assumptions:

i) **Sample Sizes**: the sample size of the groups sampled for each group is assumed to be different based on what datapoints we have from our dataset, making Welch's T-Test more suitable compared to others such as Student's T-Test.

ii) **Variation between samples**: we are also assuming that the variances are also different between samples, which is another key assumption of Welch's T-Test.

iii) **Independence**: We are assuming that each subject belongs to one group and do not overlap with one another.

T-Test #1: Emergency program usage vs Transitional usage - independent samples
This T-test was performed in order to investigate the first research question: to check if there are any statistically significant differences between Emergency and Transitional usage in Shelter Programs when using User Count as a dependent variable. The T-Test yielded the following results:

T-Statistic = 29.94

P-Value = 3.172×10^{-195}

Degrees of Freedom = 50940.0

From our results above, our test calculated a t-statistic of 29.9376 - which is a relatively high t-statistic value. This implies a statistically significant difference between the mean emergency programs users when compared with transitional programs users.

Meanwhile, the p-value from this test is 3.172×10^{-195} - which is an extremely small given normal alpha levels (0.05, 0.01). This means that the observed difference between each group's mean is not likely to have occurred by chance. This means that we can reject the null hypothesis and therefore conclude that there's a statistically significant difference in service user counts between the two types of programs.

A degrees of freedom value of 50940.0 is very high, indicating a large sample size which can provide more confidence in the statistical significance of the test result.

T-Test #2: Men vs Women in Utilization Rates, use T-test for independent samples
This T-Test was performed to help answer the second research question: are there any statistical gender disparities between users in utilization rates, specifically between men and women across program sectors? The T-Test yielded the following results:

T-Statistic = 12.92

INF2178 (Winter, 2024)
Michelle Lee (#1001448169)

P-Value = 6.491×10^{-38}
Degrees of Freedom = 11636.73

From the results above, our test outputted a t-statistic of 12.92. This is a relatively large t-statistic value, which indicates a significant difference between the means of Men and Women for utilization.

Meanwhile, the p-value for this test was 3.172×10^{-195} - which is relatively small and close to zero. This means that the chance of the observed difference occurring by chance is unlikely, and provides evidence for us to reject the null hypothesis and conclude that there is a statistical significance between the utilization rates between men and women sectors.

A degrees of freedom value of 11636.73 is very high, indicating, similar to the previous test, a large sample size which provides more confidence in the statistical significance of our results.

T-Test #3: Small vs Large shelters based on Utilization rate - independent samples test
This T-Test was performed to help answer the second research question: are there any statistical gender disparities between users in utilization rates, specifically between men and women across program sectors? The T-Test yielded the following results:

T-Statistic = -28.91
P-Value = 2.249×10^{-181}
Degrees of Freedom = 31254.83

From the results above, our test outputted a t-statistic of -28.91, which is a relatively significant value and indicates a statistically large difference between the two groups - small and large shelters. The negative sign suggests that the mean of smaller shelters is lower than the mean of the large shelters.

Meanwhile, the p-value for this test was 2.249×10^{-181} - which is relatively small and close to zero. This means that the chance of the observed difference occurring by chance is unlikely, and provides evidence for us to reject the null hypothesis and conclude that there is a statistical significance between small and large shelters.

A degrees of freedom value of 31254.83 is very high, indicating, similar to the previous tests, a large sample size which provides more confidence in the statistical significance of our results.

CONCLUSION:
In conclusion, from our data analysis above, the main trends seem to be that not only are Toronto's shelters operating at close to 100% capacity, but there are also disparities in Shelter utilization between emergency and transitional programs, and between different demographic groups (e.g., men vs. women). The size of shelters also affects their utilization rates, with larger shelters having different dynamics compared to smaller ones.