INF2178 Assignment 1 Narrative
Name: Samantha Chui
Student Number: 1002369126
Professor: Shion Guha

Data Pre-Processing
In order to prepare for analysis, I first previewed the dataset using shelter_df.head() and shelter_df.info().

I wanted to separate the column type based on whether it's a category-type column (i.e. it further describes the shelter) or if it's a data-type column (i.e. it contains the number of users). After changing column PROGRAM_ID to type str, only all data-type columns will be float or int whereas all other category-type columns are str or datetime.

I then wanted to count the number of unique values in the category-type columns. I created a function unique_categories that takes a dataset and returns the number of unique values for each non-float or non-int type column.

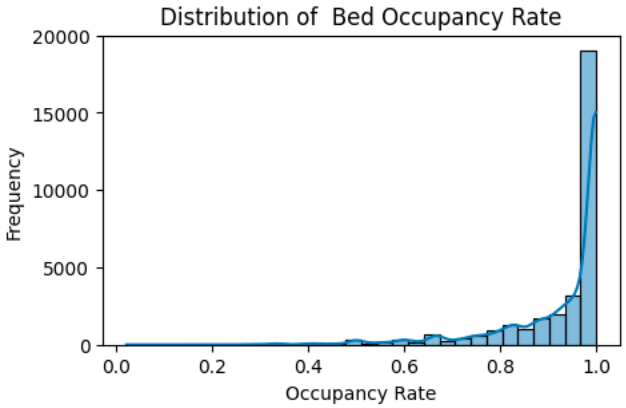
For the analysis, I wanted to look at:

- The difference in occupancy in bed vs. room type shelters
- The difference in number of users in the type of organization: City of Toronto vs. COSTI Immigrant Services

Bed vs. Room Occupancy
I created a filtered df bed_df that contains only data for bed type shelter. This df contains only the relevant columns: OCCUPANCY_DATE, OCCUPIED_BEDS, CAPACITY_ACTUAL_BED. I then created a new column OCCUPANCY_RATE which is the percent occupancy for each row of data.

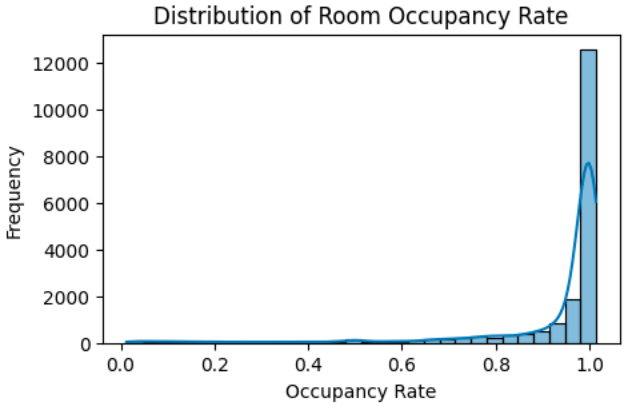
I wanted to further narrow it down by daily occupancy, however I need to first see how the data is distributed to determine how it should be aggregated. I created a function get_stats that takes a dataset and returns the mean, median, min and max value, and the IQR of the data.

Based on the summary statistics on OCCUPANCY_RATE, it showed that the majority of the shelter was near full capacity, although there were some days where the shelter was at less than 1% occupancy. This suggests that the distribution of the data is skewed - in this case, it may not be appropriate to use the mean to aggregate the daily occupancy rate. To explore this further, I plotted the data in a histogram to visualize the distribution.



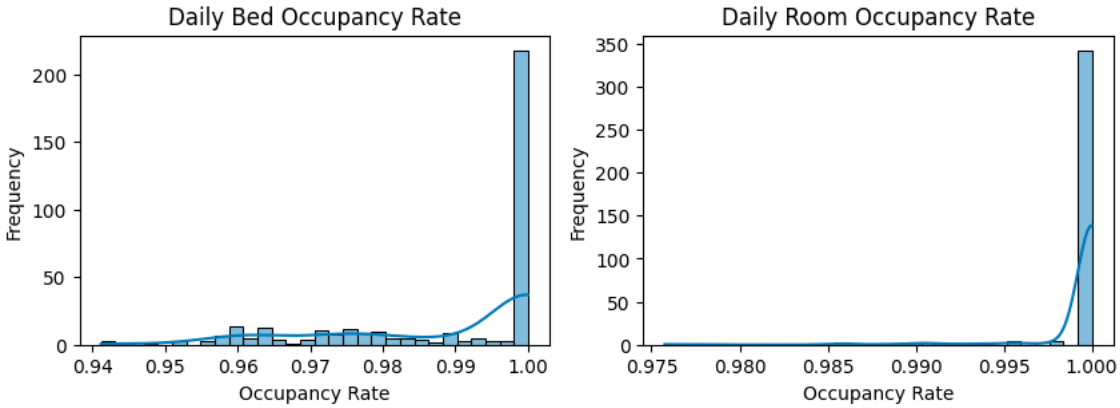
Indeed, it shows that the data does not follow a normal distribution - it is skewed to the right. The median is more appropriate because it is a robust measure of central tendency in instances where the data is skewed or there are outliers. The daily_beds is the aggregated daily median occupancy rate for bed based shelters.

I did the same for room based shelters. The summary statistics showed a similar trend where the majority of the shelters were near full capacity, although there were some instances where the shelter was at less than 1% occupancy. Similarly, the histogram showed that the data does not follow a normal distribution - it is skewed to the right.

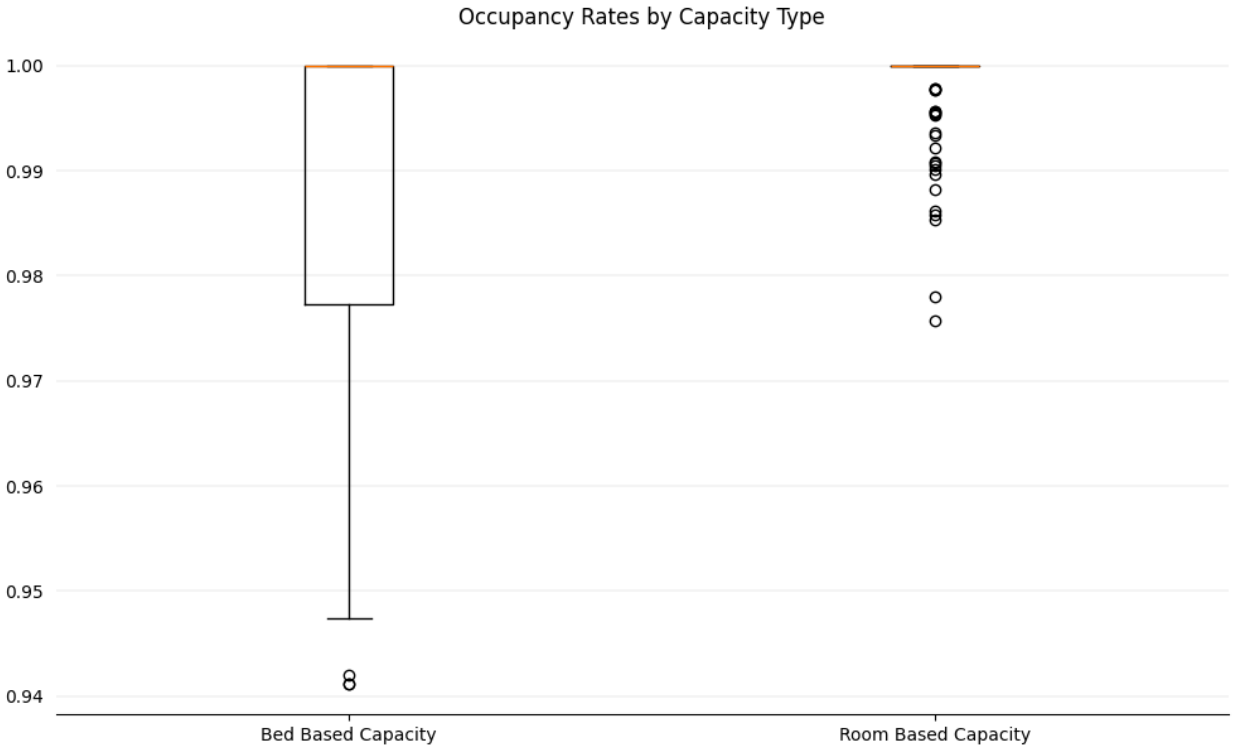


The df daily_beds contains the aggregated daily median occupancy rate for room based shelters.

I plotted the daily median occupancy rate as a histogram - this is similar to the plots above but with the data that will be used to conduct the actual statistical analysis:



This confirms that the data does not follow a normal distribution. I then plotted the data as a boxplot to visualize the variance of the data:



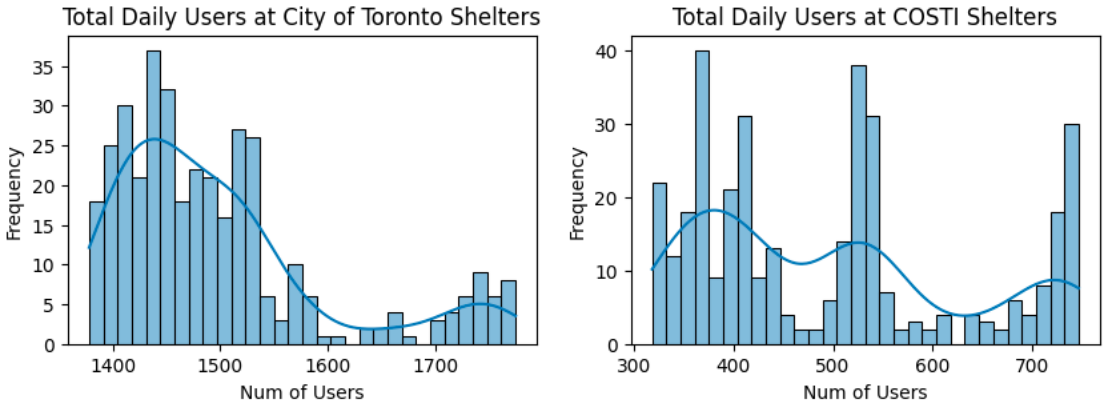
The boxplot shows that although both samples show a similar histogram shape, they do not share the same variance. In this case, a Welch's T-Test should be used. We can still assume that the samples follow a normal distribution because of Central Limit Theorem. The purpose of the t-test is to examine the mean differences between the two samples: bed vs room type shelters. The null hypothesis is that the two samples share the same mean, whereas the alternate hypothesis is that the two samples do not share the mean.

I created a function welchs_ttest which takes two samples and returns the t-statistic and p-value from the Welch's t-test. Based on the calculated p-value and the specified critical value (the default is a = 0.05), it also returns whether the null hypothesis should be accepted or rejected. In this case, since the p-value is less than 0.05, we reject the null hypothesis. This means that bed and room type shelters have different occupancy rates.

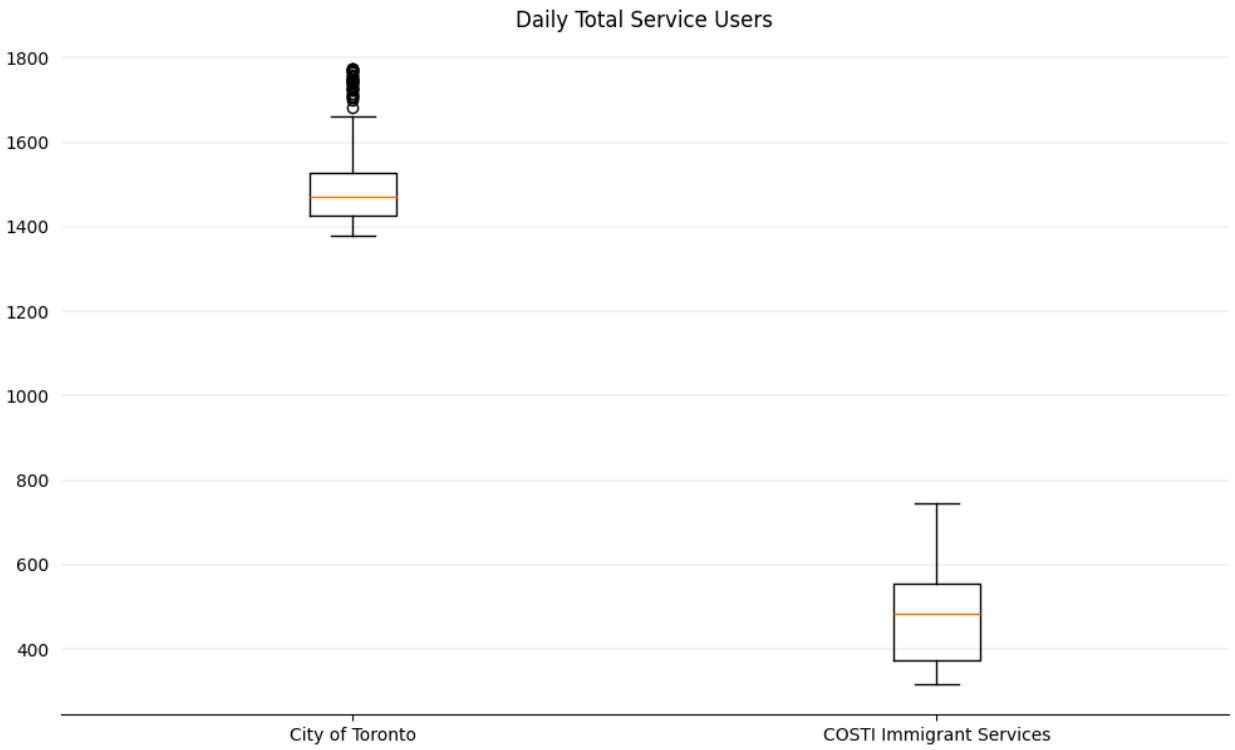
Service Users: City of Toronto vs. COSTI Immigrant Services

I created a filtered df toronto_df that contains only data from City of Toronto. This df contains the number of service users for each shelter affiliated with City of Toronto. I aggregated the data in a new df daily_toronto, which contains the daily total number of users accessing City of Toronto shelters. The summary statistics of this dataset was not very meaningful as, unlike the earlier analysis, the number of service users is a discrete variable.

I then created a similar filtered df costi_df, then aggregated the data as daily_costi, which contains the daily total number of users accessing COSTI shelters. Similarly, the summary statistics were generated but were not very meaningful. In order to better understand the dataset, I plotted a histogram:



The results showed that the two datasets are quite different: for City of Toronto shelters, there is a bimodal distribution whereas for COSTI shelters, there is a multimodal distribution. Next, I plotted the dataset as boxplots to see if the two samples would have equal variance:



The results showed that most of the City of Toronto shelters had a higher number of service users compared to COSTI shelters.

I used welchs_ttest again to examine the mean differences in service users between the two organizations. The null hypothesis is that City of Toronto and COSTI shelters share the same mean, whereas the alternate hypothesis is that the two organizations have different means. The results showed a p-value less than 0.05, suggesting that we should reject the null hypothesis.