2178 Assignment 1
Yiwen Mei
1010159766

Background setting

In the wake of a chilly winter in 2023, the City of Toronto found itself grappling with questions about the efficiency of its shelter program. With the thawing snow revealing not just the onset of spring but also the need for policy enhancements, city officials embarked on a data-driven journey to shed light on two critical aspects of the shelter system:

- the occupancy rates across different types of shelter capacities
- the demographic distribution of service users by gender.

This narrative unfolds the analytical process that was undertaken to glean insights from the shelter data, insights that would be instrumental in guiding future decisions aimed at optimizing the shelter program for the diverse needs of Toronto's residents.

Research Questions

The exploration was rooted in two central research questions.

1. Did room-based and bed-based capacities within shelters experience similar occupancy rates?
2. Was there a difference in the service user count between the male and female shelter sectors?

The answers to these questions promised to inform a broader understanding of resource utilization and demographic service patterns within the shelter system, paving the way for more informed, equitable, and efficient resource allocation.

Data Preparation and Cleaning

The analytical process began with a scrupulous phase of data preparation and cleaning. The dataset, rich with various categorical variables, needed to be sifted through to ensure accuracy and relevance. Variables such as organization names, sectors, program models, types of services, program areas, and capacity types were each examined for unique categories, with any anomalies or null values addressed promptly. This phase set the stage for a clean, reliable dataset that would underpin the integrity of the subsequent analysis.

Occupancy Rate: Room vs Bed

With a pristine dataset at hand, the analysis proceeded to focus on occupancy rates. A subset of data encompassing capacity types, sectors, program models, and the count of actual versus occupied beds and rooms was isolated for detailed scrutiny. Occupancy rates were meticulously calculated for both room and bed capacities, with the ultimate aim of merging these into a singular, comprehensive metric. This consolidation of occupancy rates was a pivotal step, allowing for a direct comparison between the two types of shelter capacities. To get some full insight of the data, let’s take a look at the statistics of the two types of shelters’ occupancy rate:

```
OCCUPIED_BEDS_RATE summary statistics:
Min: 0.02
Mean: 0.93
Max: 1.0
25th percentile: 0.9
Median: 1.0
75th percentile: 1.0
Interquartile range (IQR): 0.1
```

```
OCCUPIED_ROOM_RATE summary statistics:
Min: 0.01
Mean: 0.93
Max: 1.01
25th percentile: 0.96
Median: 1.0
75th percentile: 1.0
Interquartile range (IQR): 0.04
```

T-test:

A statistical comparison was conducted using the two-sample t-test to understand if the occupancy rates for room-based and bed-based capacities significantly differed. Since the data are fairly large, it is reasonable to assume that data are normally distributed because of the central-limit-theorem. We will start by doing a two-sample t-test to get a general sense first, and then we check if the variance of two groups is equal. If not, we need to do a Welch T-test instead.

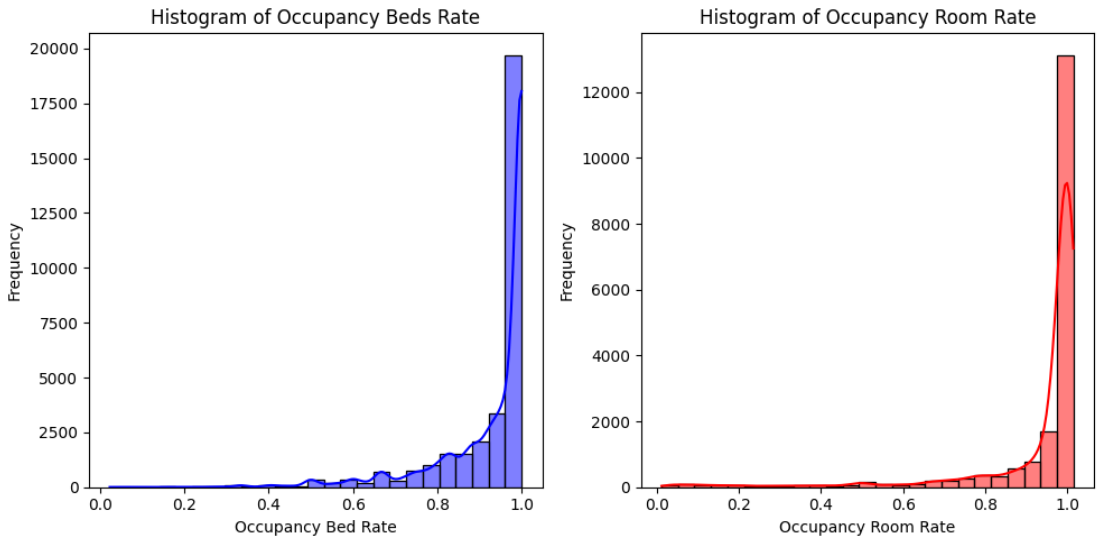
H_0: The occupancy of bed-based shelter and room-based shelter are the same.
H_1: There is a statistically significant difference between the occupancy of bed-based shelter and room-based shelter.

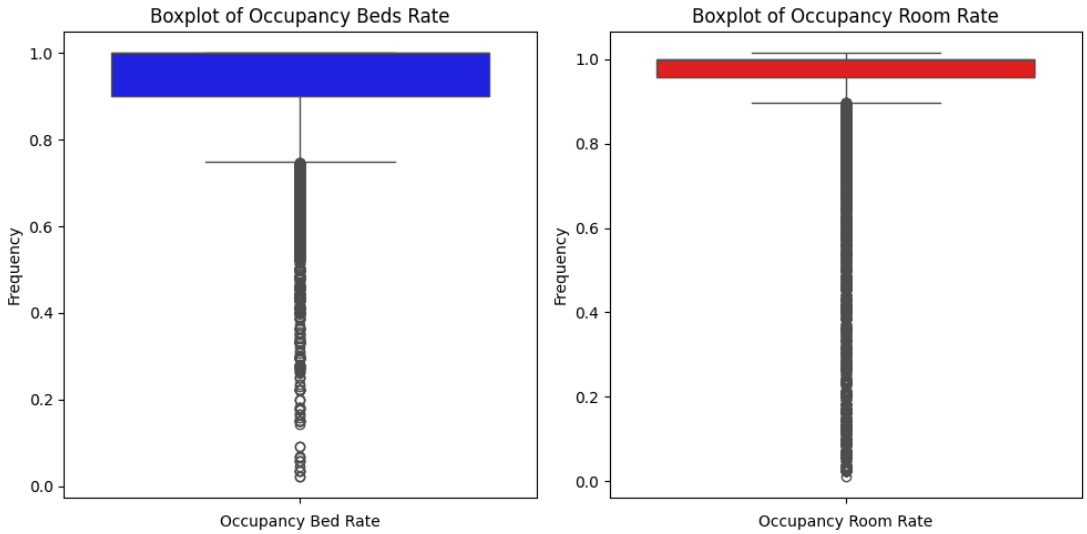
```
2-sample t-test result:
t-statistic: 4.854104599422829
p-value: 1.2128933183471424e-06
```

The t-test number suggests that we have found strong evidence to reject our null hypothesis. The p-value looks promising, but we assumed the variance of the two groups are equal. However, are they really equal?

Visualization:

Both histograms and boxplots are excellent at comparing medians, distributions, range, etc. between two groups of interest.





After carefully examining the above plots, we have the following conclusions:

1. Both the beds and rooms have high occupancy rates, with medians close to full capacity. This indicates efficient utilization of both types of shelter capacities.
2. The variance in the occupancy rate for beds seems to be higher than that for rooms because the spread of outliers and the width of the whiskers in the boxplot is greater for beds. This is also visible in the histograms where the beds rate has a more pronounced tail.
3. Both groups exhibit a left skewness in their distributions, more so in the case of beds, which has a heavier tail of low occupancy rates.

Unfortunately, the two-sample t-test equal variance assumption is not met, and we need to do Welch T-test instead.

Welch T-test:

```
Welch's t-test result:
t-statistic: 4.498751771925636
p-value: 6.860477551487939e-06
```

The extremely small p-value suggest that the observed difference in the means of these two groups is significant and not likely due to random chance. We should still reject our null hypothesis. This also implies that the type of capacity (room or bed) has a statistically significant impact on the occupancy rate, and therefore, the city of Toronto might consider this finding in their strategy to improve the shelter program's efficiency.

Gender User Count Analysis

The narrative then turned its lens towards the gender distribution of shelter service users. Although the specifics of this analysis were not extracted, one could anticipate the approach. The analysis would involve aggregating the service user counts by gender and applying appropriate statistical measures to identify any notable differences. This comparison would be critical in highlighting any gender-based disparities in service utilization, offering a clear-eyed view of how the shelter's resources were being accessed by different demographics.

The gender user count analysis was not merely a statistical exercise; it was an inquiry into the fabric of the shelter program, seeking to ensure that services were equitably utilized by all genders. Should the analysis uncover a significant discrepancy in service user counts between genders, it would raise important questions about the accessibility and inclusivity of the shelter system, prompting discussions on how best to address these disparities. To get some full insight of the data, let's take a look at the statistics of the user count for different gender:

```
Service User Count for Men Sector
Min: 1
Mean: 39.87
Max: 234
25th percentile: 19.0
Median: 32.0
75th percentile: 48.0
Interquartile range (IQR): 29.0
```

```
Service User Count for Woen Sector
Min: 1
Mean: 28.66
Max: 100
25th percentile: 15.0
Median: 25.0
75th percentile: 37.0
Interquartile range (IQR): 22.0
```

T-test:

We want to test if the number of men and women using the shelter services is the same or if there's a noticeable difference. Since the data is fairly large, we can assume that it would follow a normal distribution due to the central limit theorem. We will start by doing a two-sample t-test to get a general sense first, and then we check if the variance of two groups user counts are equal. If not, we need to do a Welch T-test instead.

H_0: The service user count for men and women in the shelters is the same.
H_1: There is a significant difference between the number of men and women using the

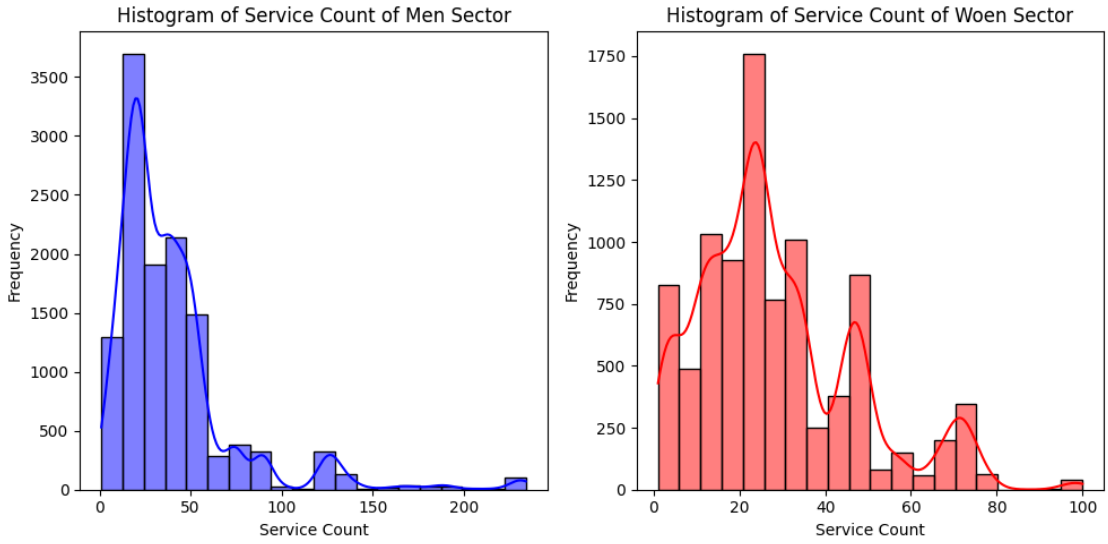
shelter services.

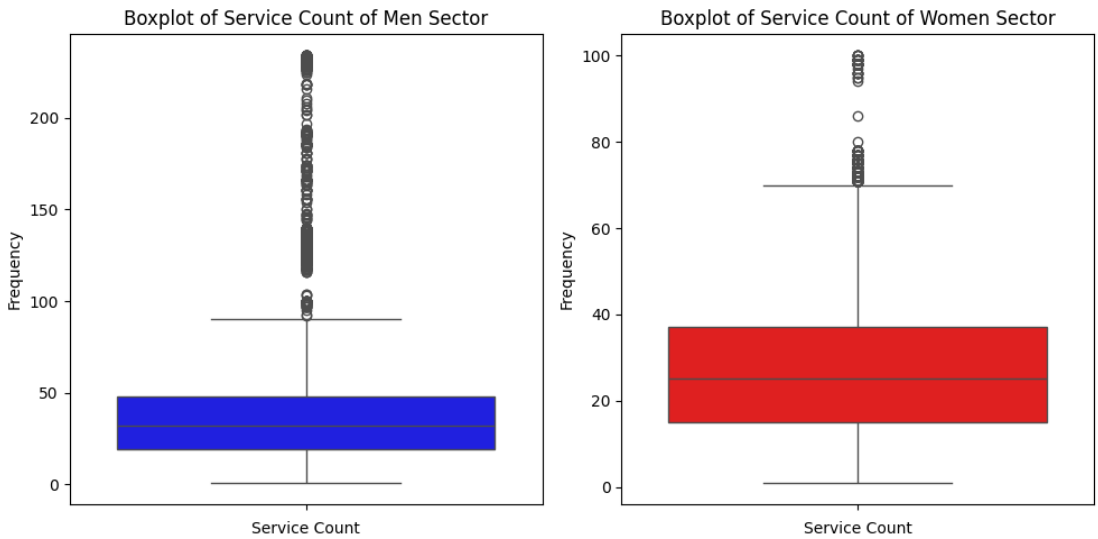
```
2-sample t-test result:
t-statistic: 28.24955042305982
p-value: 2.022346382679794e-172
```

The t-test number suggests that we have found strong evidence to reject our null hypothesis. The p-value looks promising, but we assumed the variance of the two groups are equal. However, are they really equal?

Visualization:

Both histograms and boxplots are excellent at comparing medians, distributions, range, etc. between two groups of interest.





After carefully examining the above plots, we have the following conclusions:

1. The variance (spread of the data) in the men's sector appears to be larger than that of the women's sector, which is evident from both the boxplot and the histogram.

2. The median service count for women is higher than for men, indicating that on average, women are using more services.
3. The men's sector has a greater number of outliers, which could indicate that there are specific instances where men use a lot more services than usual.

In summary, the service usage patterns differ between the men's and women's sectors, with men showing a wider range and more variability in service counts, and women showing higher median service usage and less variability. These differences could be important for the city of Toronto when considering how to allocate resources and optimize services for different genders within the shelter system. On the other hand, the two-sample t-test equal variance assumption is not met, and we need to do Welch T-test instead.

Welch T-test:

```
Welch's t-test result:
t-statistic: 30.50421164151915
p-value: 1.0532742925183108e-199
```

The extremely small p-value mean that the average service user count is significantly different between the male and female sectors. We should still reject our null hypothesis. This finding could prompt further investigation into why this difference exists and whether any policy changes or resource allocations are needed to address it.

Insights and Conclusion

The statistical analysis of shelter occupancy rate and gender-based shelter usage count revealed significant differences. Welch's t-test, accounting for variance inequality, confirmed substantial and consistent disparities in occupancy rates between room-based and bed-based shelters. On the other hand, the analysis of gender-based usage patterns showed that men had greater variability and outliers in service counts, while women exhibited higher median usage with less variability.

These statistical findings have real-world implications for Toronto's shelter system. Significant gender-based differences in service usage call for targeted strategies to ensure equitable access and utilization of resources. These insights can inform policy and operational refinements, enhancing the effectiveness and equity of the shelter program to meet the needs of all service users and foster a supportive environment.