#Student Names and ID: Jiachen Liu (1010182204)
#Instructor Name: Professor Shion Guha
#Course Code: INF2178
#Course Name: Experimental Design for Data Science
#Program: Master of Information
#Faculty: Faculty of Information
#School: University of Toronto

INF2178 Technical Assignment 1 Short Narrative

Data Preprocessing

This dataset is a large one with 14 columns and 50,944 rows, so it's important to preprocess the dataset suitable for late analysis. After the initial examination of the dataset, I found that it revealed missing values in columns such as "PROGRAM_NAME," "PROGRAM_MODEL," "OVERNIGHT_SERVICE_TYPE," and "PROGRAM_AREA."

These columns are all categorical variables and the missing values are not too many. To address these missing values, I performed a data cleaning step by using placeholder value "Unknown" to fill missing entries in all of the columns, which ensured that the dataset was more robust for subsequent analysis, allowing for a comprehensive exploration of shelter usage trends.

Exploratory Data Analysis (EDA)

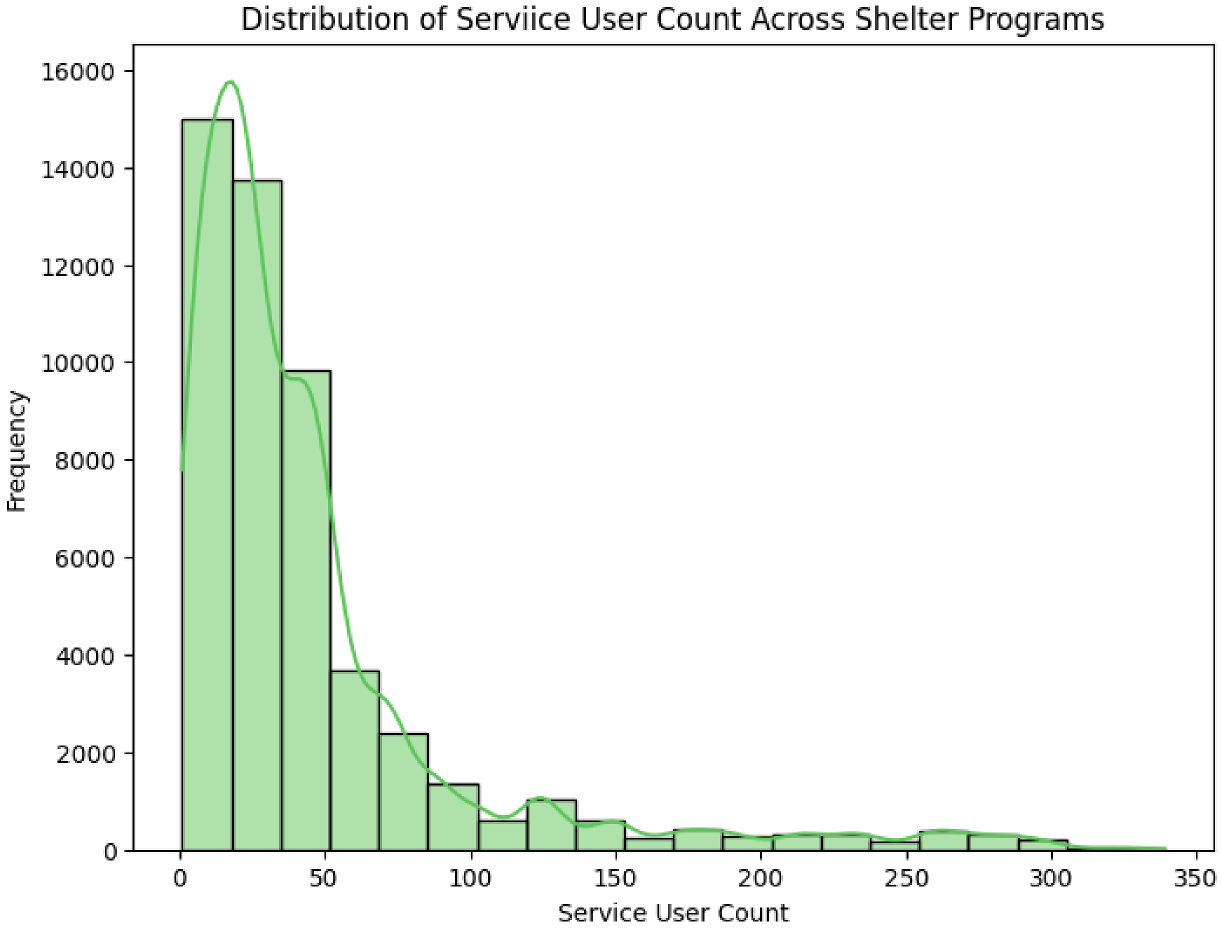
To further analyze the dataset, specifically the trends and distribution of certain columns, I developed 5 research questions and then conducted EDAs to answer them.

1.How are shelter programs distributed across different sectors?
2.What is the distribution of "SERVICE_USER_COUNT" across all shelter programs?
3.How does the overall occupancy of shelters change over the course of the year?
4.How do occupancy rates vary across different shelter sectors?
5.How do occupancy rates differ across different program areas?

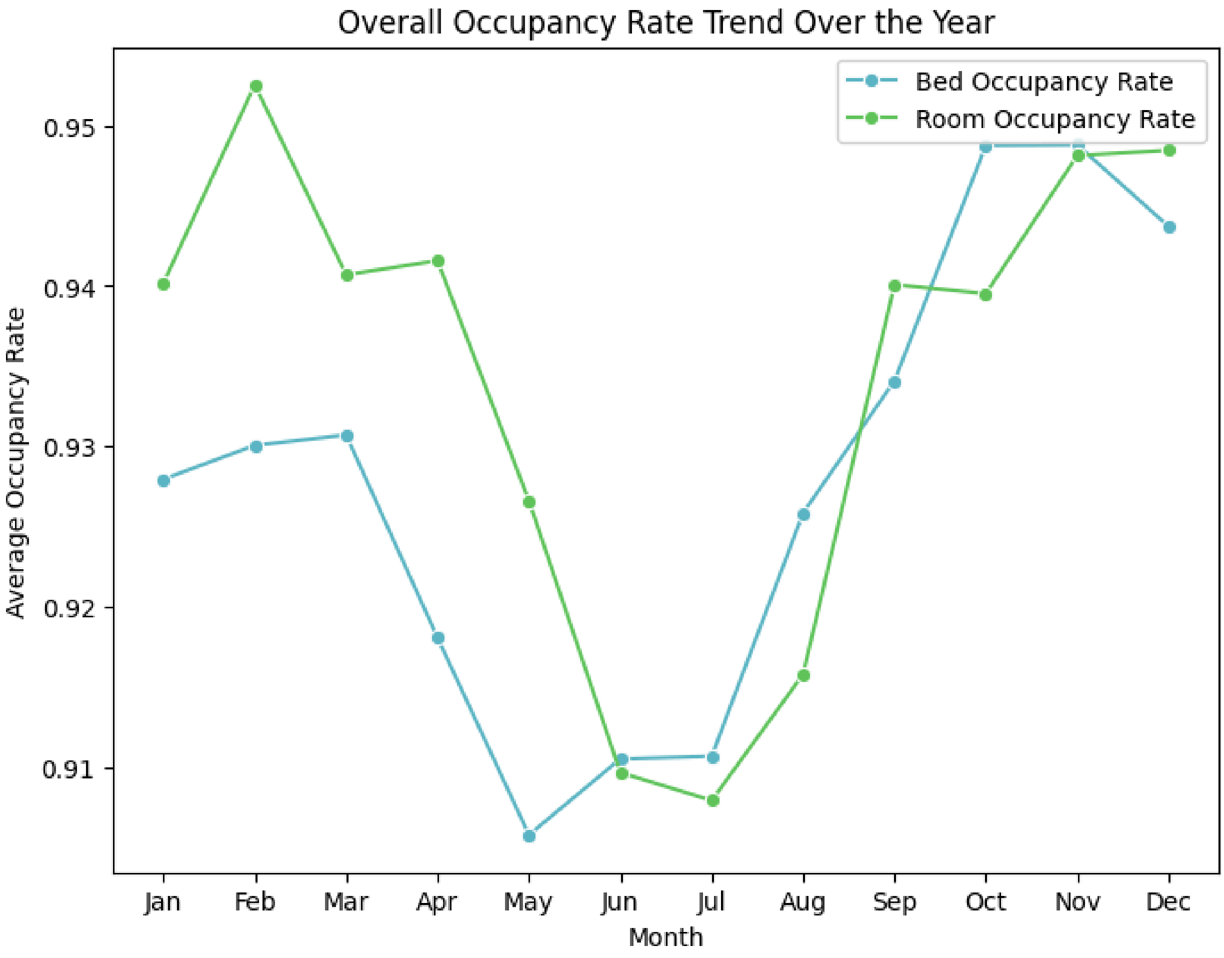
Since some of them require occupancy rates, which is not a continuous variable, I first calculated it before I did any analysis.

The first question focuses on shelter program distribution across sectors, and the bar chart with color-coded palettes highlighted the diversity in shelter programs catering to various demographics, including men, women, mixed adults, youth, and families. From the bar chart, I learned the basic demographics of shelter programs in terms of sectors. The "mixed adult" sector has the highest number of programs and the families sector the lowest. This indicates that there is a significant focus on providing shelter services to a diverse adult population, which could include both single adults and couples without children. Also, the numbers of the families sector while there are services available for families experiencing homelessness, the number of programs is less than those for other sectors. This could reflect a lower relative demand, or possibly indicate an area where additional resources might be needed. Also, this could possibly be the demographic makeup of the homeless population in Toronto.

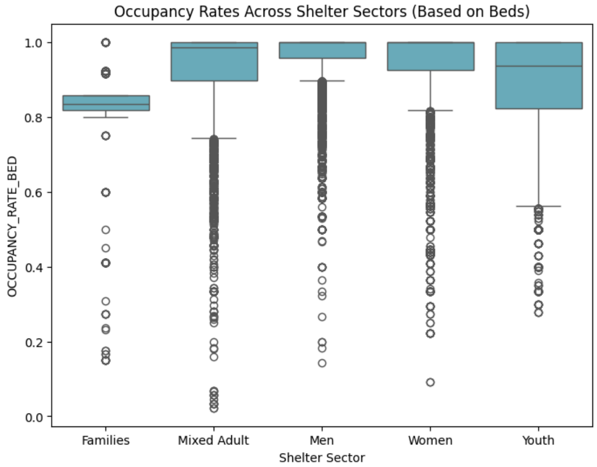
The second question delves into the distribution of shelter service user count. From the histogram we can see that the distribution of service user count is right-skewed, meaning that a large number of shelter programs have a relatively small number of users, while fewer programs have a large number of users. The large amount of presence of shelter programs with higher user counts could signal instances where facilities are at or near full capacity, which could correlate with the challenges mentioned in the assignment introduction, such as individuals being turned away due to lack of space. We can infer that while a substantial number of shelter programs operate with fewer users, there is a significant need for programs that can accommodate larger populations.

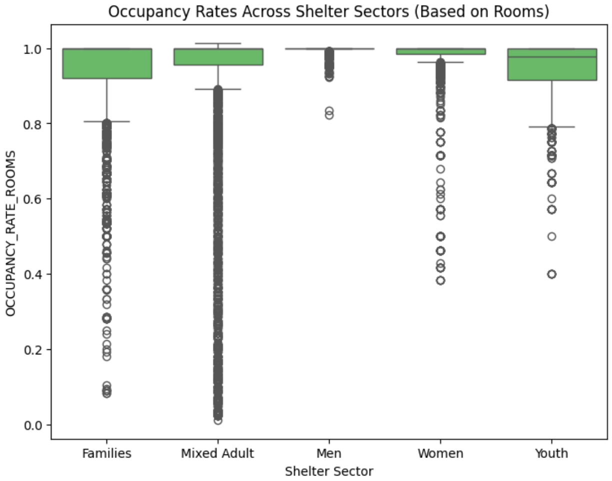


The third question was meant to explore the seasonal occupancy rate trend based on beds and rooms. There are clear seasonal patterns in shelter occupancy. The bed occupancy rate is generally lower than the room occupancy rate. Both bed and room occupancy rates appear to fluctuate throughout the year and experience a significant dip around the middle of the year, particularly in June and July. This suggestS that during warmer months, the demand for shelter decreases, possibly due to more people choosing to stay outdoors or the availability of temporary seasonal housing solutions. During the winter months, starting from September and peaking in December, there's a significant increase in occupancy rates towards the end of the year, which is a common trend as weather gets colder. Moreover, the rates are quite high throughout the year, hovering above 90%. This indicates a consistently high level of utilization of shelter services, suggesting that the shelter system is under constant demand and possibly close to its capacity limits, especially during the colder months.

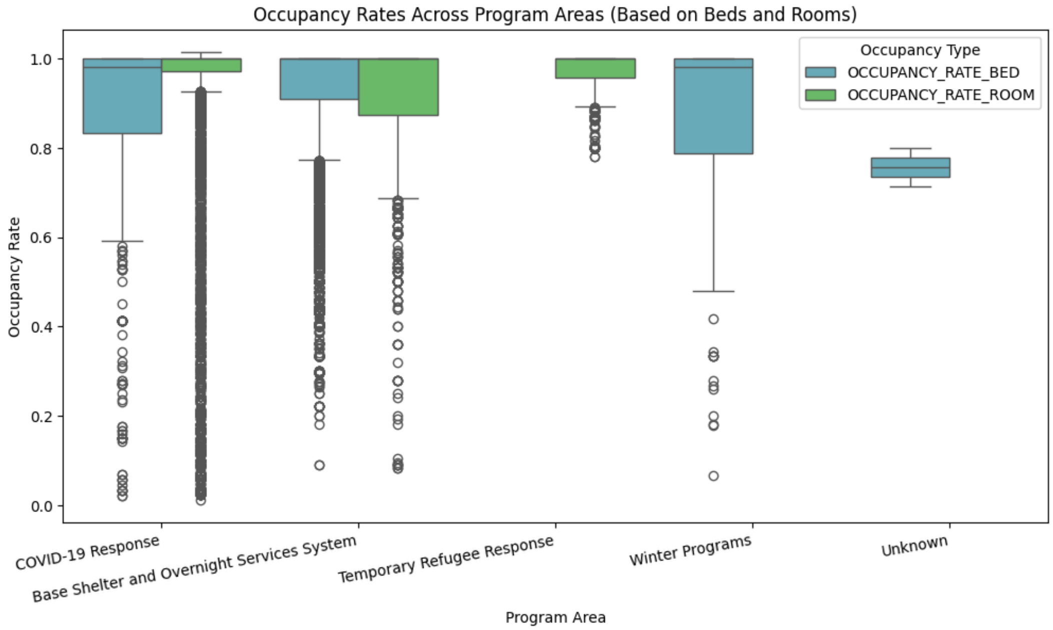


The fourth question focuses on occupancy rates across sectors. The consistent high occupancy rates in the "Youth" sector across both beds and rooms could signal a pressing need for more youth-targeted shelter services. The variability and outliers in the data suggest that while some programs might be at capacity or overutilized, others have potential to serve more users. From these two graphs, we learn that the shelter system in Toronto experiences different rates of occupancy in different sectors, and this varies further when comparing bed-based and room-based accommodations. Such insights can guide targeted interventions and policy decisions to address the specific needs of each sector, especially those with high demand and consistent occupancy (like the "youth" sector).





The fifth EDA task aims to see the occupancy rates across different program areas. I melted the data so I was able to create a side-by-side boxplot comparing bed and room occupancy rates across different program areas. There is a noticeable difference in occupancy rates across program areas, which reflects the diverse nature of the programs, the populations they serve, and their capacity to meet demands. In almost all program areas except for the COVID-19 response, the median room occupancy rate is higher than the bed occupancy rate. This could be due to a preference for room accommodations when available or reflect that family and group accommodations (which are typically room-based) are at a higher demand.



Quantitative Analysis Using T-Tests

To see whether occupancy rates based on both beds and rooms vary for the 2 types of program models(emergency and transitional), I conducted 2 t-tests comparing the mean bed occupancy rates/mean room occupancy rates between Emergency and Transitional shelter programs, respectively.The null hypothesis will be: there is no significant difference in bed/room occupancy rates between Emergency and Transitional programs.

For bed occupancy rates, the t-statistic of 2.828 suggests that there is a difference in the means of the two groups. However, since the P-value is greater than 0.05, we do not have enough evidence to reject the null hypothesis. This means that, based on the data and the t-test conducted, there is no significant difference in bed occupancy rates between Emergency and Transitional shelter programs.

For room occupancy rates, the t-statistic of 1.941 is lower than that for bed occupancy rates, which suggests a smaller difference between the group means for room occupancy rates. Similarly, because the P-value is greater than 0.05 significance level, we cannot reject the null hypothesis for room occupancy rates either. This indicates that the room occupancy rates for Emergency and Transitional programs are not significantly different from each other.

In conclusion, the t-tests results for both bed and room occupancy rates between Emergency and Transitional shelter programs show that the differences in the means of these groups are not statistically significant. This suggests that both Emergency and Transitional programs have similar occupancy rates in Toronto. However, the lack of statistical significance does not prove that the null hypothesis is true, it's just that we do not have sufficient evidence.

Conclusion

Based on the exploratory data analysis (EDA) and the t-test results conducted on the dataset tracking the daily occupancy and capacity of Toronto shelters for the year 2021, I learned several insights. The shelter system in Toronto shows a complex landscape with varying levels of demand and utilization across different sectors and program areas. The impact of external factors such as seasonality and the COVID-19 pandemic is evident in occupancy trends. The lack of significant differences in occupancy rates between Emergency and Transitional programs suggests that both program types are crucial components of the overall shelter strategy, serving the homeless population effectively without significant disparities in usage.

These findings underscore the necessity for a nuanced approach to shelter planning and resource allocation that considers the diverse needs of different groups in various situations. Continual monitoring and analysis of shelter data are essential for adapting to changing needs and ensuring that all individuals experiencing homelessness have access to shelter services.