Introduction

The dataset used in this assignment contains daily occupancy information along with capacity of beds/rooms at various Toronto shelters in 2021, with key variables such as occupancy date, organization name, program model, etc. The aim of the assignment is to study shelter trends and make inferences from the same through Exploratory Data Analysis (EDA) and quantitative analysis.

Data Pre-Processing

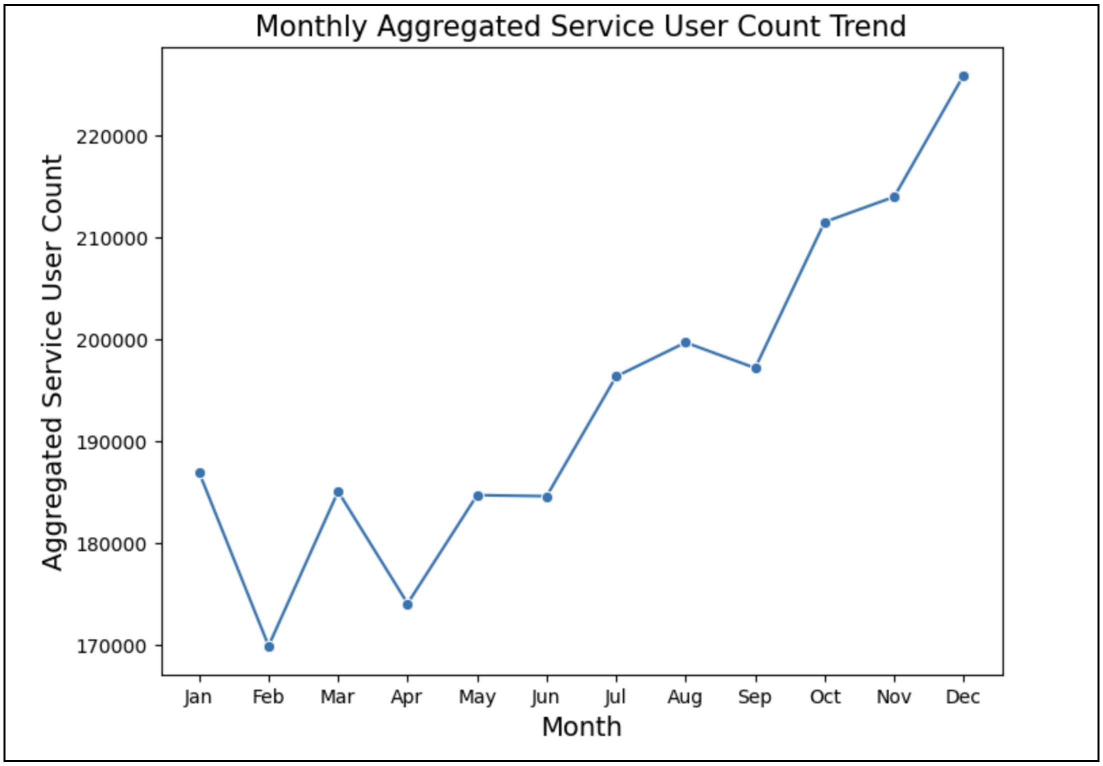
In order to accurately track shelter occupancy rates across both capacity types, bed and room based capacity, I created a new variable that calculated the ratio using the number of occupied beds/rooms divided by the actual capacity of the beds/rooms, and was then used in the analysis to track occupancy trends across various categorical variables and quantitative analyses. To further study these trends across time, I extracted the month from the occupancy date column and converted it into the full name of the month, such as January, February, etc.

EDA

To get an overview of the distribution of service users across all shelters within the dataset, I constructed a histogram plot based on the service user count. The graph showed peaks concentrated around 0-100 service user counts, indicating a higher count of service users falling within this range. This potentially highlights the shelters' capabilities of serving smaller to medium-sized populations of service users in the dataset.

Next, in order to track the most widely used organizations providing shelter in the data set, I tried to identify any trends in the service users based on the organization name. As there was no further data on if the service users were new or existing for the organization, with a high likelihood of overlap between the two, simply aggregating the user counts seemed to be a biased method of ranking the organizations. Therefore, for a clearer view of how the service users were distributed across various organizations over time, I calculated the median of the service user counts based on the organization names, which would give comparable central values. The graph confirmed these as the top three shelter organizations - WoodGreen Red Door Family Shelter, The MUC Shelter Corporation and The Salvation Army of Canada.

Further, to understand the trend of service users with time, I plotted a line graph to visualize the overall distribution of aggregated service users counts over each month through 2021.

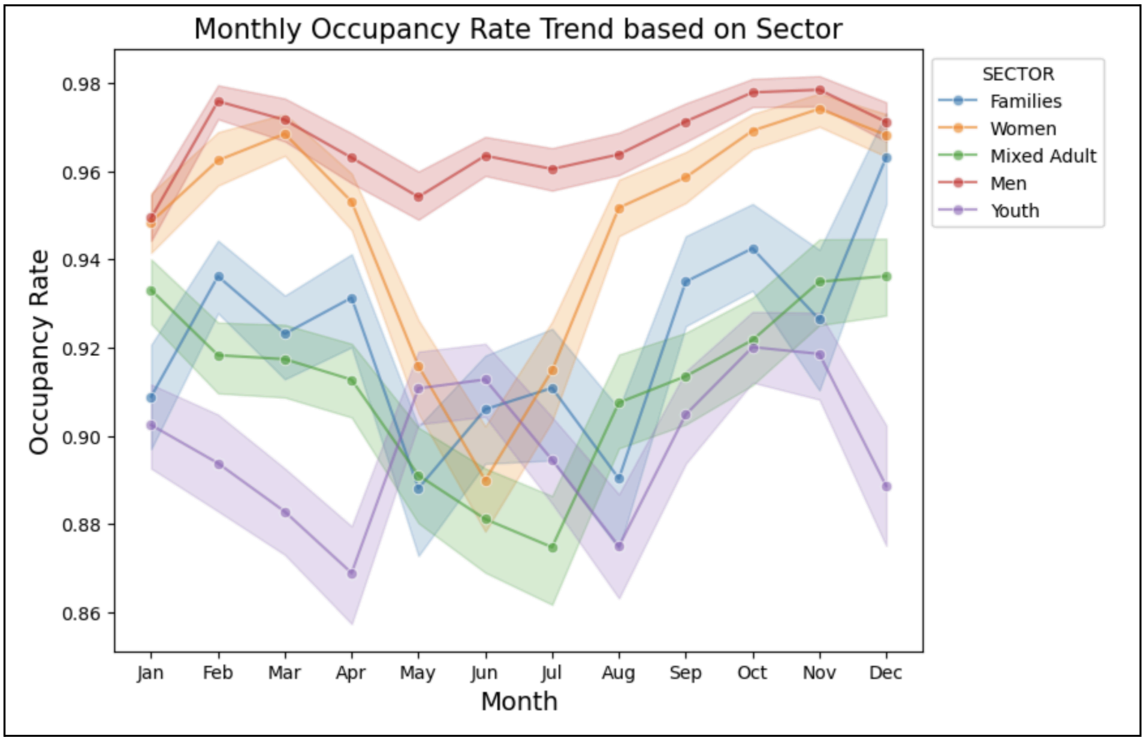


The graph shows a dip in the months of February and April, followed by a steady increase of service users towards the end of the year with the highest count in December.

Following this, to understand the overall distribution of service users based on the kind of overnight shelter accommodation and the sector categorization, I created a grouped barchart which showed a higher majority of the service users using Motel/Hotel Shelters and Shelters compared to the other overnight accommodation categories. The graph also showed that within Motel/Hotel Shelters, the highest count of service users belonged to the Mixed Adult sector, while the highest count among Shelter service users belonged to those in the Men sector. Additionally, the graph also depicts that although Mixed Adult service users are notably the highest count overall, especially in Hotel/Motel Shelters, the data shows a higher count of all other sector categories in Shelter, implying higher usage in this type of accommodation for non-Mixed Adult service users.

Further, it was observed that there was a higher presence of service users in room-based accommodations than bed-based accommodations based on capacity type. This suggests that the shelter system and organizations likely provide more facilities with room-based capacities to possibly accommodate for more service users.

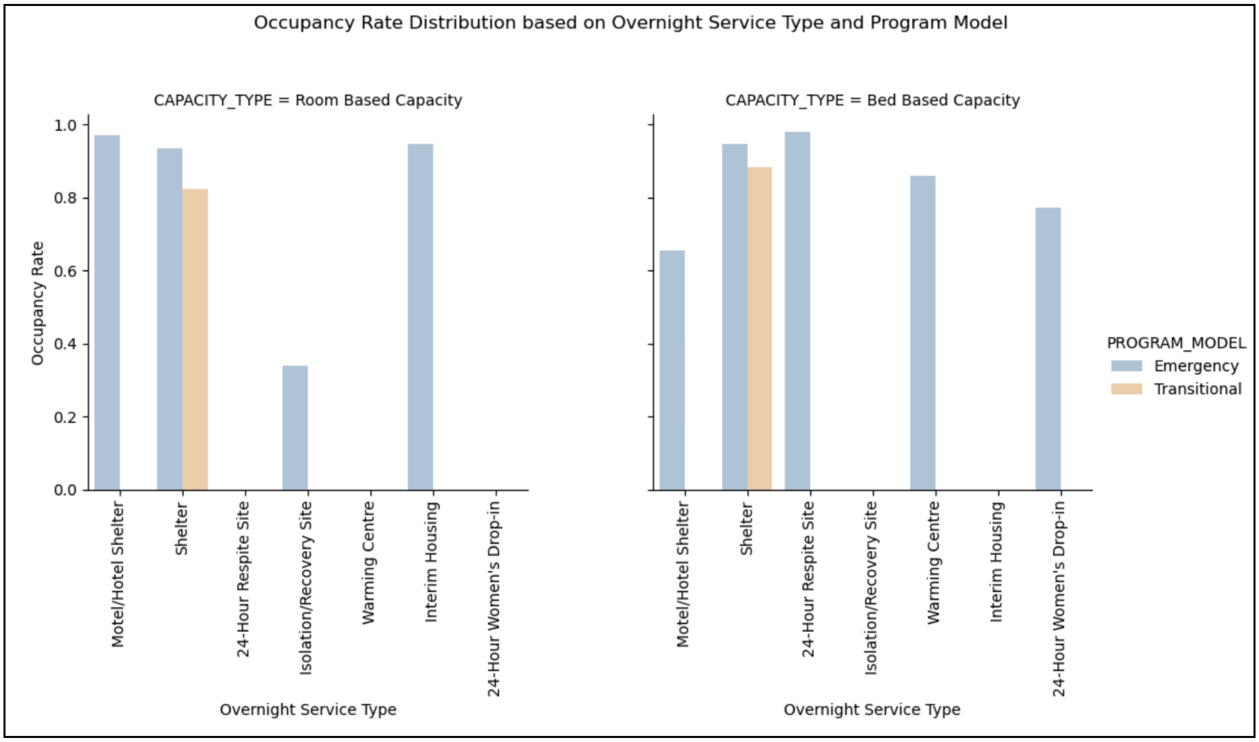
Next, to analyze the occupancy rate trends across the months of 2021 for each of the six sector categorizations, I created a grouped line plot to easily compare the trend across months among the sectors.



The graph shows a steady high occupancy rate for Men service user groups across all months, but a significant drop in the middle of the year for Women, Mixed Adult and Families service user groups, followed by a steady increase in these same user groups towards the end of the year. For the users in the Youth sector, the data showed a drop in the month of April, followed by a stable increase towards the middle of the year, followed by another drop during August and then December - This sector category has the lowest occupancy rate compared to the other categories, and seems to follow a different trend, especially during the middle and the end of the year. This highlights the diversity in occupancy trends across various sector groups, requiring different levels of support and resources across the year.

As each program in the data set was further categorized into one of four program areas, I used a box plot to understand the service user distribution in each program area. The graph shows the smallest distribution for Winter Programs, which is likely due to its low availability for all year round services. The graph also shows a low variability for Base Shelter and Overnight Services System, with a high count of outliers. Additionally, the graph shows a higher variation in the distribution of the remaining two program areas, with Temporary Refugee Response having the highest spread, followed by COVID-19 Response shelters having a significant amount of outliers. Given this dataset of year 2021 and the rampant spread of COVID-19 a couple of years prior to this, these outliers in the two program areas may be further studied.

Following this, to understand the occupancy trends based on type of capacity at the accommodation and program model - I created comparable bar charts to study occupancy rate for any trends among the combinations of these categories.



As expected, the graph shows a clear distinction between the room-based and bed-based type of capacities. The data shows that all service users in the 'Transitional' program model use Shelters, while 'Emergency' program model service users vary across other service type categories based on type of capacity. Among room-based capacity, Model/Hotel Shelter and Interim Housing have the highest occupancy rates, while Isolation/Recovery Site has the lowest occupancy rate. For bed-based capacity, 24-Hour Respite Site and Shelter have the highest occupancy rates, with Motel/Hotel Shelter having the lowest occupancy rate. Given these findings, it would make sense for the Isolation/Recovery Site having the overall lowest occupancy rate in 2021, as this service type was primarily aimed for service users suffering from COVID-19.

Quantitative Analysis using t-tests

For the following analyses, I used Welch's t-test for its robust nature handling unequal sample sizes and unequal variances.

1. **Based on Program Model**

   The results from the t-test concluded that there are significant differences between mean occupancy rates of the two types of program models - Transitional or Emergency.

2. **Based on Program Frequency**
   For this analysis, I created a dummy variable which categorized the program area as 'Regular' (Base Shelter and Overnight Services System) and 'Temporary' (COVID-19 Response, Temporary Refugee Response, Winter Programs) based on their description in the dataset file. The results from the t-test concluded that there was not enough evidence to confirm that there are significant differences in the mean occupancy rates between regular and temporary programs.

3. **Based on Type of Capacity**
   The results from the t-test concluded that there are significant differences between mean occupancy rates of the two types of capacity - Bed-based or Room-based capacity.

Future Scope

In the overall monthly distribution of service users, the data seems to show its peak in December, followed by a low count in January. It would be interesting to see if this is a trend seen across the previous years, and further, study the impact of any resources that could be of use to service users during this time. Additionally, the data shows differing trends in occupancy rates across sectors, especially Youth service users - it would help to further track these trends to provide insight into efficient resource allocation during these peak times of the year.

Conclusion

To summarize, this assignment has provided valuable insight into understanding the importance of EDA and quantitative analysis. Discovering trends through EDA gave a deeper understanding of various aspects related to shelter occupancy rates, service user distribution, diverse sector groups and trends over time. Further, testing for significant differences among groups also provides key information to help make informed decisions to help reduce homelessness and improve support to this population. This assignment serves as an example on the application of data science in a social science context, giving potential to help improve the world through analytics.

References

- Python Graph Gallery - https://python-graph-gallery.com
- Stack Overflow - https://stackoverflow.com
- Seaborn Documentation - https://seaborn.pydata.org/archive/0.11/tutorial/categorical.html
- Pandas DF Documentation - https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html