Exploratory Data Analysis and Statistical Insights on Shelter Program Occupancy

.ipynb file - https://colab.research.google.com/drive/1SvhfebCjqxz9A_E4LeaVKlwGxHhXyy42?usp=sharing

INF2178 - Technical Assignment 1 Narrative Report

Fatima Ashfaq

1010784732

February 4, 2024

1. Introduction

The exploration and analysis of the City of Toronto's Shelter Program dataset aims to uncover meaningful insights into the occupancy trends, program models, and capacity types. This narrative provides a detailed walkthrough of the entire process including initial processing, non-graphical and graphical EDA, and statistical experiments. Since Google Colab was utilised to perform these analyses, the dataset was uploaded to Google Drive to derive actionable insights for stakeholders and policymakers.

2. Initial Processing

The "OCCUPANCY_DATE" column underwent date conversion to enable temporal analysis. Unrequired columns such as "PROGRAM_ID" and "SECTOR" were removed to provide clearer focus on the required columns for the statistical analysis. Figure 1 displays the first few data entries of the loaded and processed dataframe.

	OCCUPANCY_DATE	ORGANIZATION_NAME	PROGRAM_NAME	PROGRAM_MODEL	OVERNIGHT_SERVICE_TYPE	PROGRAM_AREA	SERVICE_USER_COUNT	CAPACITY_TYPE	CAPACITY_ACTUAL_BED	OCCUPIED_BEDS	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS
0	2021-01-01	COSTI Immigrant Services	COSTI North York West Hotel - Family Program	Emergency	Motel/Hotel Shelter	COVID-19 Response	74	Room Based Capacity	NaN	NaN	29.0	26.0
1	2021-01-01	COSTI Immigrant Services	COSTI North York West Hotel - Seniors Program	Emergency	Motel/Hotel Shelter	COVID-19 Response	3	Room Based Capacity	NaN	NaN	3.0	3.0
2	2021-01-01	COSTI Immigrant Services	COSTI North York West Hotel Program - Men	Emergency	Motel/Hotel Shelter	COVID-19 Response	24	Room Based Capacity	NaN	NaN	28.0	23.0
3	2021-01-01	COSTI Immigrant Services	COSTI North York West Hotel Program - Mixed Adult	Emergency	Motel/Hotel Shelter	COVID-19 Response	25	Room Based Capacity	NaN	NaN	17.0	17.0
4	2021-01-01	COSTI Immigrant Services	COSTI North York West Hotel Program - Women	Emergency	Motel/Hotel Shelter	COVID-19 Response	13	Room Based Capacity	NaN	NaN	14.0	13.0

Figure 1 - df.head() after initial processing

3. Non-Graphical Exploratory Data Analysis

In this portion of the analysis, descriptive statistics was utilised. The first method df.describe() is used to compute the statistics for all quantitative variables in the dataset. The second method, which makes used of the defined function 'get_summary_statistics' computes the mean, median, minimum value, and maximum value for the variable of interest. In preparation and insight generation for this analysis,

this latter method was used to examine which variable, “SERVICE_USER_COUNT” or “OCCUPIED_BEDS”, should be used for the graphical EDA and for answering the research questions. Due to the presence of several null values in the “OCCUPIED_BEDS” column, the “SERVICE_USER_COUNT” column was selected.

4. Graphical Exploratory Data Analysis

This section of the analysis made use of visualisations to unravel insights, patterns, and relationships within the data.

The first visualistion, figure 2, made use of a boxplot to identify any outliers, if present, and determine if different capacity types, room-based capacity and bed-based capacity, exhibited similar variance patterns. This boxplot illustrates the variances across different capacity types, providing useful insights into the distribution of occupancy rates. As seen in the visualisation below, the boxplot displays a significant number of outliers - hence the variations between and within the two capacity types may be different.

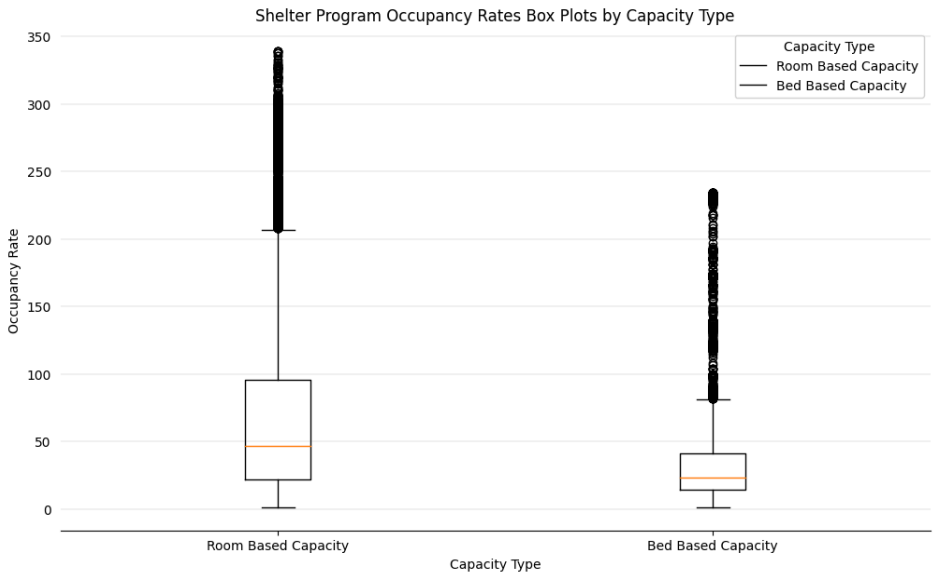

Figure 2 - Shelter Program Occupancy Rates by Capacity Type

The second boxplot, a winsorized boxplot, figure 3, was generated to address the outliers detected in the initial boxplot. This boxplot was generated after winsorizing the “SERVICE_USER_COUNT” variable. Inspiration and tutorial to address the outliers was gained from the SciPy v1.12.0 manual accessed from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mstats.winsorize.html. This technique of using the ‘mstats.winsorize’ function helped mitigate the effect of the extreme values on the analysis. However, since there are still several outliers present in the bed-based capacity type, this research will make use of Levene’s test for equal variances to determine the appropriate statistical test to gain answers to the formulated research questions.

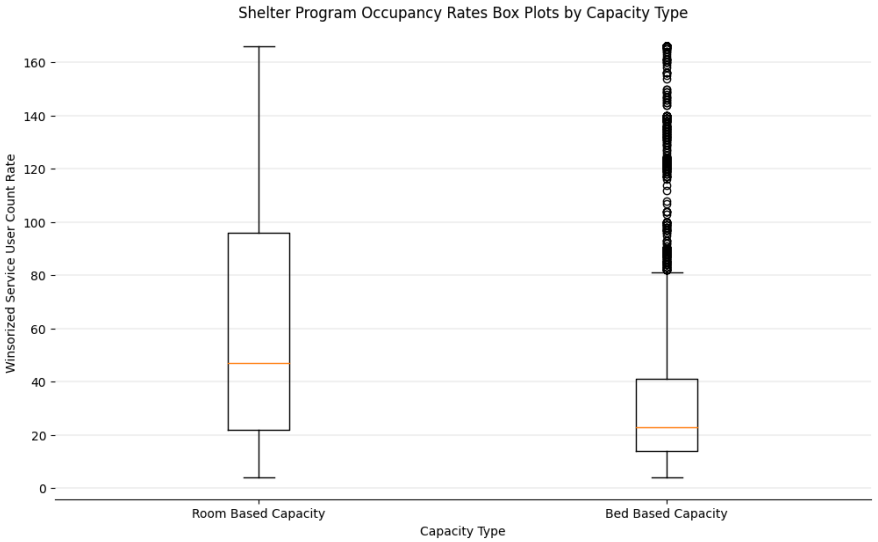

Figure 3 - Winsorized Shelter Program Occupancy Rates by Capacity Types

As seen in figure 4, a time series analysis was also performed to explore temporal patterns. The line graph showcases the trends in the daily admissions, providing insights into the fluctuations of the occupied beds over time and uncover the patterns in the daily intake of individuals within the shelter program. Upon the first glimpse of the graph, it is evident that the intake of individual only increased over time. The lowest daily intake throughout all the organisations was recorded between the months April and May of 2021 and the highest intake was recorded in December 2021. This could be explained with the impact of COVID-19 on the incomes and employment status of individuals.

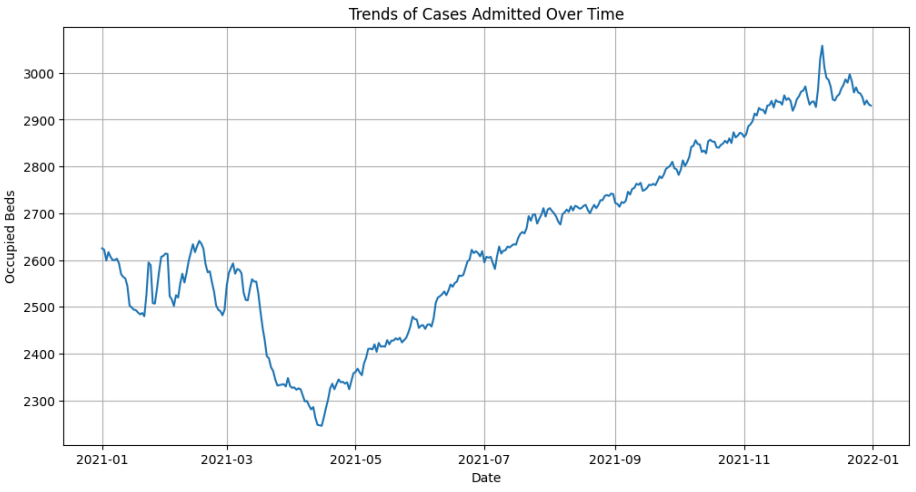

Figure 4 - Trend of Cases Admitted Over Time

Furthermore, a barplot was also constructed to visualise the intake trends in the two program models; Emergency and Transitional, as shown in figure 5. The emergency model bar is significantly larger than than the bar for transitional model's intake. This surge in the emergency model's intake could be, once again, due to the impact of COVID-19 as many families and individuals lost their income

streams and hence fell victim to homelessness. The intake in emergency model is accessible to everyone, with or without a referral whereas the transitional model only accepts intakes that have a referral.

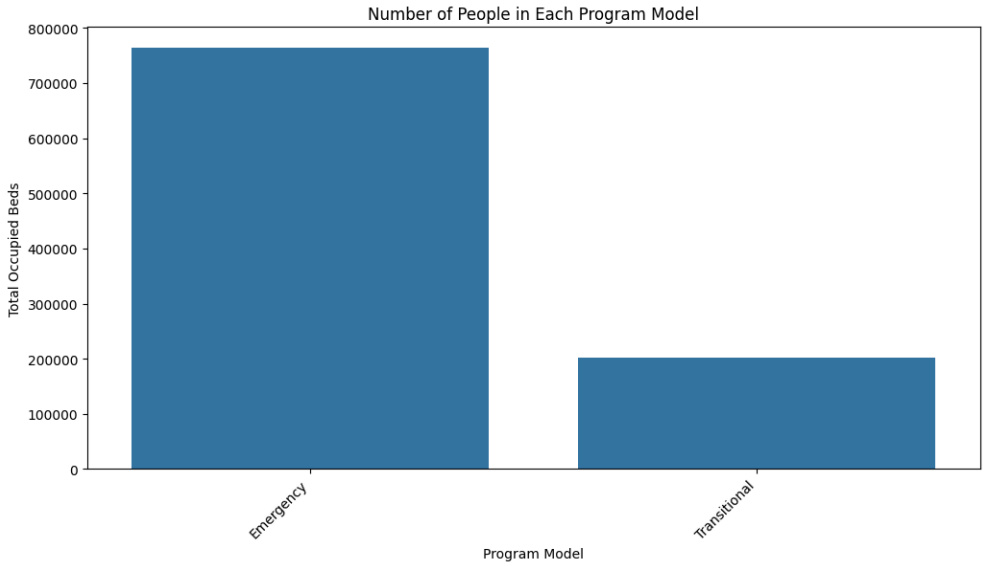

Figure 5 - Number of People in Each Program Model

Another barplot was created to examine the number of admissions across the different shelter organisations. As seen in figure 6, City of Toronto has the highest admissions of individuals and Women's Hostels Inc. has the least.

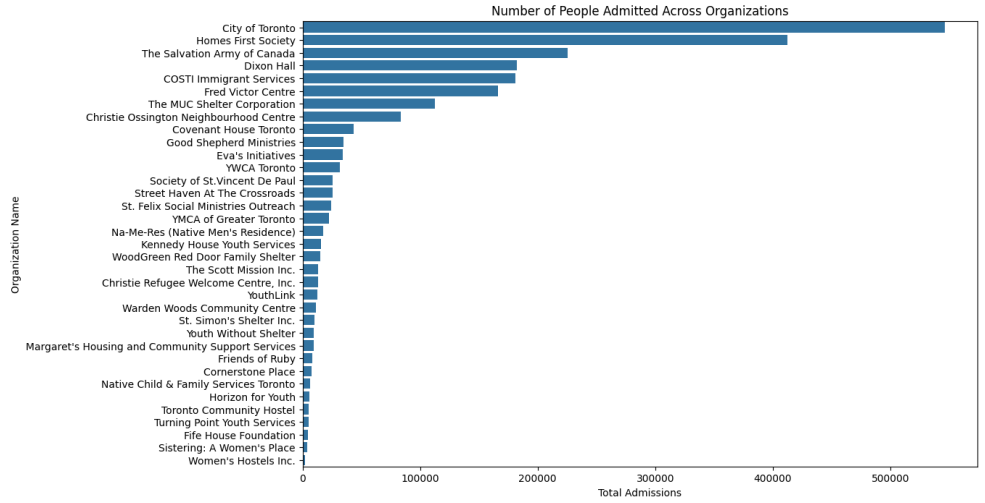

Figure 6 - Number of People Admitted Across Organisations

5. Statistical Experiments

Three research questions were formulated, and Welch's t-tests were conducted to answer them. These tests provided statistical evidence of differences in occupancy rates for specific categories.

5.1. Research Question 1 - Occupancy Trends by Program Model

This first research question sought to determine if different program models exhibited significant differences in occupancy rates. Since the p-value obtained for Levene's test for equality of variances - insights of this statistical test was gained from https://www.statology.org/levenes-test-python/ - was extremely small, close to zero (**2.7468122789149392e-111**), this indicated that the variances of the two groups/samples are significantly different. Hence, a Welch's t-test was applied as it is more robust to the deviations between the groups/samples. The extremely low p-value of **4.7779490253035953e-247** from Welch's t-test indicates a significant difference in the occupancy rates between the two models (emergency & transitional), hence providing insights to achieve better program planning and resource allocations between the two. This leads to the rejection of the null hypothesis as there is a significant difference in the occupancy rates across program models.

```
Levene's test for equality of variances:
T-statistics=  506.4029672064086
P-value = 2.7468122789149392e-111

Welch's T-test results for Program Models:
T-statistic: 33.950217918446846
P-value: 4.7779490253035953e-247
```

Figure 7 - Statistics Computed For Research Question 1: Occupancy Trends by Program Model

5.2. Research Question 2 - Occupancy Rates in COVID-19 Response Programs

This second research question investigated whether the occupancy rates differed significantly between programs operating as COVID-19 responses and other programs. Similar to the first research question, the p-value obtained for the Levene's test for equality of variances was extremely small (**1.79153e-319**) hence leaning towards the application of Welch's t-test due to its robust nature. The t-test p-value of **4.167030326879367e-100** indicates a substantial difference in occupancy rates, signifying the unique challenges and dynamics associated with COVID-19 response programs.

```
Levene's test for equality of variances:
T-statistics=  1493.57191966262
P-value = 1.79153e-319

Welch's T-test results for COVID-19 Response vs. Other Programs:
T-statistic: 23.10668268905625
P-value: 4.167030326879367e-100
```

Figure 8 - Statistics Computed For Research Question 2: Occupancy Rates in COVID-19 Response Programs

5.3. Research Question 3 - Shelter Capacity vs. Occupancy

This research question explores the significance of differences in occupancy rates between programs with room-based and bed-based capacity. Once again, the Levene's test p-value of **0.0** indicates

unequal variances, and the p-value obtained from Welch's t-test (**0.0**) confirmed a significant difference between these two capacity types.

```
Levene's test for equality of variances:
T-statistics=  8706.219887257426
P-value = 0.0

Welch's T-test results for Room-based vs. Bed-based Programs:
T-statistic: 78.50868849938448
P-value: 0.0
```

Figure 9 - Statistics Computed For Research Question 3: Shelter Capacity vs. Occupancy

6. Conclusion

This comprehensive analysis of shelter program data unveiled intricate patterns, trends, and statistical differences across the different aspects. The combination of non-graphical and graphical EDA, along with Welch's T-tests provided a holistic understanding of occupancy dynamics. These findings hold immense value for stakeholders, assisting them in making informed decisions, optimising resource allocation, and addressing unique challenges posed by different program models and capacities. The narrative serves as a roadmap for future analyses and interventions in the context of shelter programs, emphasizing the importance of data-driven decision-making.

7. References

Herzog, M. H., Francis, G., & Clarke, A. (2019, August 13). *Understanding Statistics and Experimental Design*. Springer. http://books.google.ie/books?id=4yCpDwAAQBAJ&pg=PR4&dq=https://link.springer.com/book/10.1007/978-3-030-03499-3&hl=&cd=1&source=gbs_api

Z. (2020, July 10). *How to Perform Levene's Test in Python*. Statology. https://www.statology.org/levenes-test-python/

scipy.stats.mstats.winsorize — SciPy v1.12.0 Manual. (n.d.). https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mstats.winsorize.html

How to t-test by group in a pandas dataframe? (n.d.). Stack Overflow. https://stackoverflow.com/questions/45015038/how-to-t-test-by-group-in-a-pandas-dataframe