INF 2178 Assignment 1

By

Ziyu Zhang (1008838260)

Master of Information

Department of Information, University of Toronto

INF 2178H S Experimental Design for Data Science

Professor. Shion Guha

Introduction

In recent years, Toront has seen a marked increase in the population of homeless. Based on the dataset of shelter trends in Toronto, this article will examine the trend of different types of shelter in recent years. This dataset has covered a wide variety of data including different shelter programs, occupancy of shelter, total available shelters and different sectors. This dataset has categorized the shelter type into two categories. The first is room based capacity and another is bed based capacity, and these two categories are also the concentration of this essay.

Methods:

Before delving into the detailed data analytics, I have done some data preprocessing to make it easier to discover some potential trends. The first thing I did was to create a new column called occupancy rate, which is shown as the format of percentage.

ORGANIZATION_NAME	PROGRAM_ID	PROGRAM_NAME	SECTOR	PROGRAM_MODEL	OVERNIGHT_SERVICE_TYPE	PROGRAM_AREA	SERVICE_USER_COUNT	CAPACITY_TYPE	CAPACITY_ACTUAL_BED	OCCUPIED_BEDS	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS	OCCUPANCY_RATE
COSTI Immigrant Services	15371	COSTI North York West Hotel - Family Program	Families	Emergency	Motel/Hotel Shelter	COVID-19 Response	74	Room Based Capacity	NaN	NaN	29.0	26.0	89.655172
COSTI Immigrant Services	16211	COSTI North York West Hotel - Seniors Program	Mixed Adult	Emergency	Motel/Hotel Shelter	COVID-19 Response	3	Room Based Capacity	NaN	NaN	3.0	3.0	100.0
COSTI Immigrant Services	16192	COSTI North York West Hotel Program - Men	Men	Emergency	Motel/Hotel Shelter	COVID-19 Response	24	Room Based Capacity	NaN	NaN	28.0	23.0	82.142857
COSTI Immigrant Services	16191	COSTI North York West Hotel Program - Mixed Adult	Mixed Adult	Emergency	Motel/Hotel Shelter	COVID-19 Response	25	Room Based Capacity	NaN	NaN	17.0	17.0	100.0
COSTI Immigrant Services	16193	COSTI North York West Hotel Program - Women	Women	Emergency	Motel/Hotel Shelter	COVID-19 Response	13	Room Based Capacity	NaN	NaN	14.0	13.0	92.857143

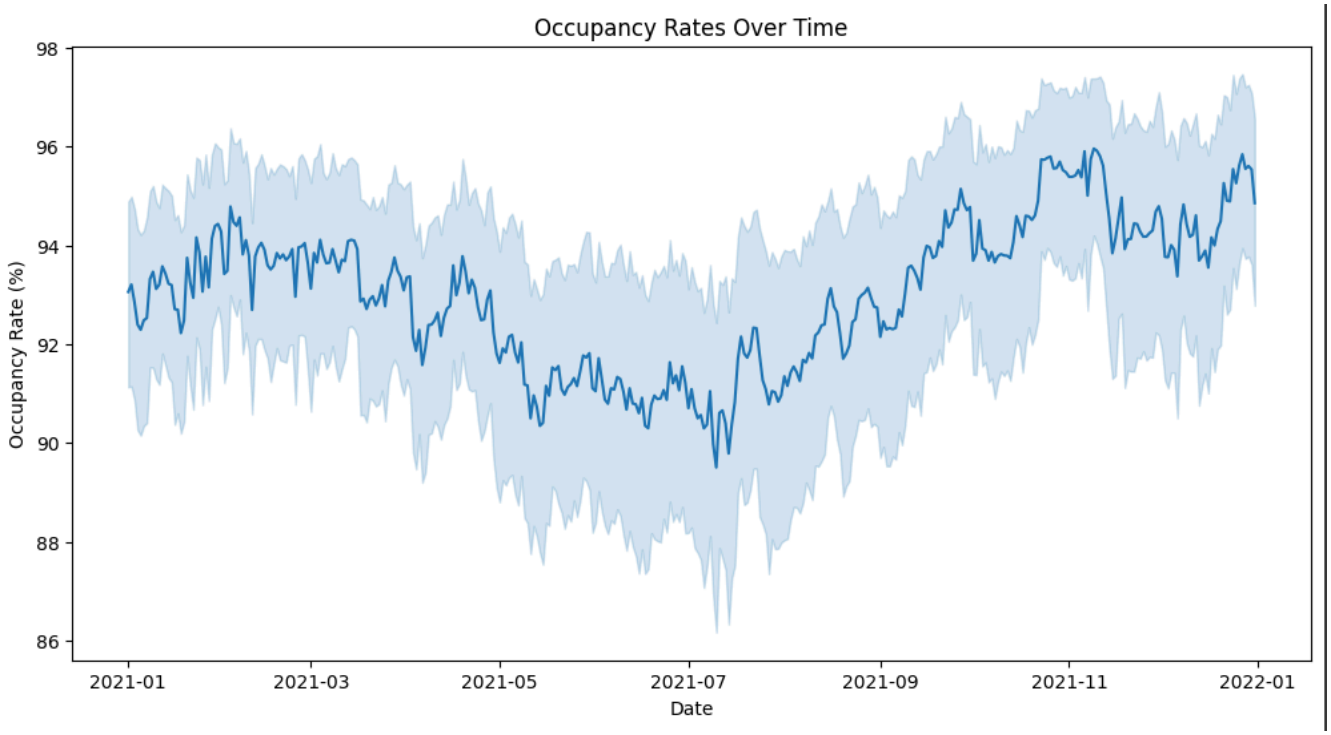
And then, I have calculated some summary statistics like minimum, maximum, mean, median and some percentiles.

```
Occupancy rate summary statistics


Min: 1.2
Mean: 93.01
Max: 101.41
25th percentile: 92.31
Median: 100.0
75th percentile: 100.0
Interquartile range (IQR): 7.69


Occupancy rate summary statistics
```

From the result, it can be easily seen that most shelters have the occupancy rate around 100 since the median is still 100 and mean is 93.01 while there are still some shelters that have extremely low occupancy rate like 1.2. However, the IRQ has indicated that there is not a big difference between 25th percentile and 75th percentile, which means those shelters with low occupancy rate may be outliers. After having the occupancy rate, I moved forward to the official EDA process and generated some useful graphs like box plots and line plots. Finally, I have conducted a two-sample t test, and Welch's t test to further prove the finding.

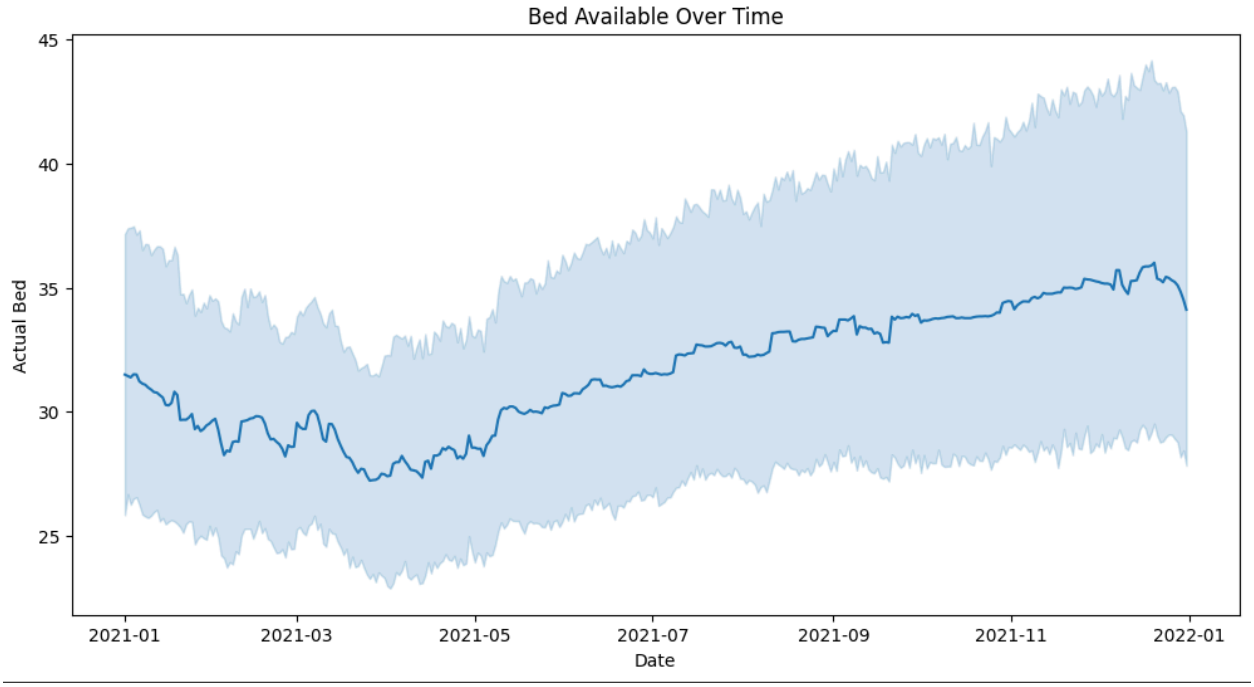
EDA

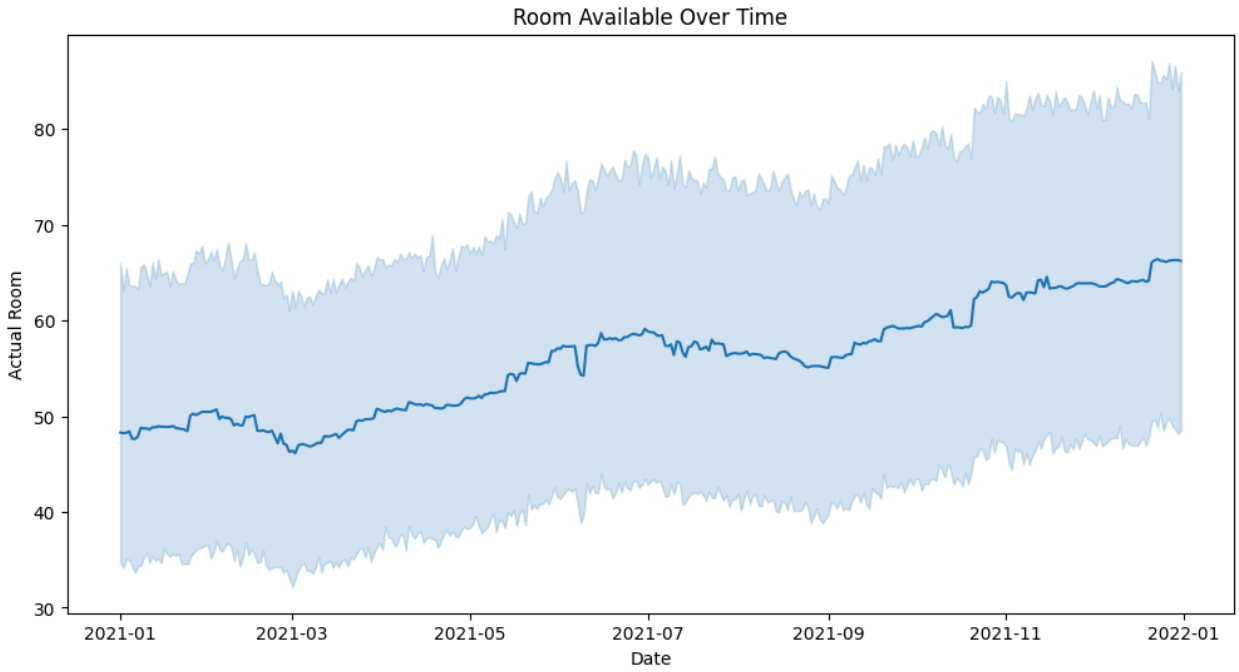
The first plot I conducted is a line plot, which is a time-series analysis to see the occupancy rate over time. For the purpose of readability, I have converted the datatype of occupancy date to datetime format and sorted it based on month from 2021-01 to 2022-01



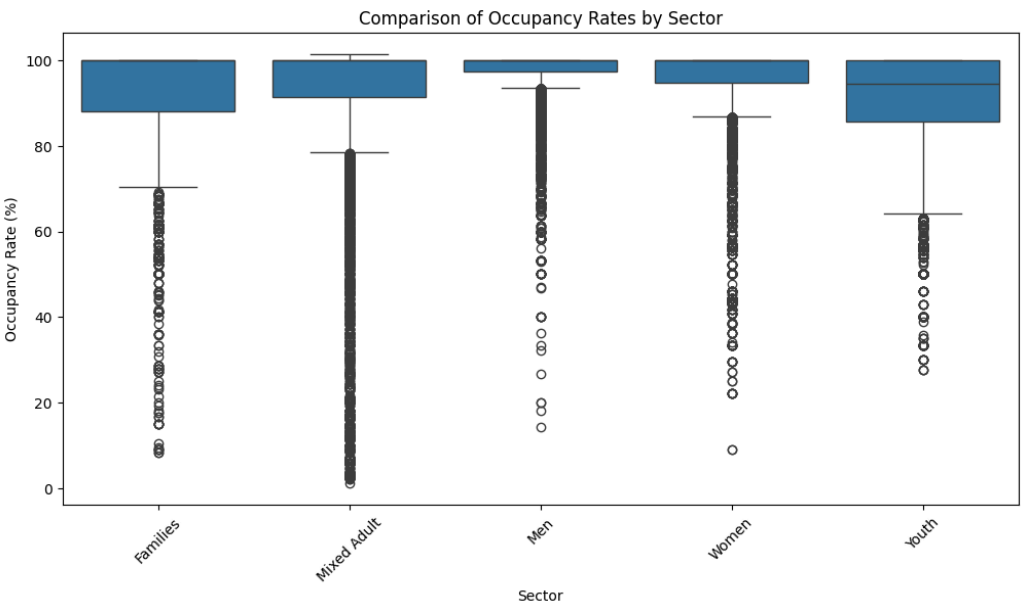
The x-axis of the graph represents the date while the y-axis of the graph displays the data of occupancy rate. From the graph, we can easily see that 2021-07 seems to be an important

focus point for the later research since there is a slightly decreasing trend before July 2021 while there is a significantly increasing trend after July 2021. Based on this finding, some hypotheses may be conducted like some important policy towards homeless people made in July 2021. The advantage of using line plots in a time-series analysis is that it can visualize the trend clearly and straightforward to use and understand. Moreover, line plots can help identify correlation and patterns for later use like making predictions. However, only visualizing the trend of occupancy rate may be easily biased since it can be affected by the overall capacity. If the overall capacity has decreased tremendously since January 2021, the overall trend of occupancy rate will still be increasing. Hence, I keep creating two line plots to show the overall bed and room capacity during the same time period since the shelter type can be either room based or bed based.





From these two graphs, we can find both room and bed availability are increasing over time while room availability shows a solid increasing trend, but the bed availability shows a slight decreasing trend around April 2021. Combining these three graphs, we can find both occupancy rate and shelter availability are increasing, which means the amount of homeless people keeps increasing since January 2021.



The fourth graph I have used is a boxplot. The advantage of boxplot is that it can display concise summary of data such as range and IQR and outlier detection. Moreover, we can easily compare distributions across different groups side by side by making multiple boxplots. In this plot, I have created the boxplot of occupancy rate for different sectors. We can observe that there are multiple outliers for each sector. However we can see that both the men and women sector will have a relatively high occupancy rate compared to families and youth. On the other hand, the mixed adult sector has displayed most outliers which means it has a wider distribution. Based on these findings, the shelter provider can make corresponding arrangements on shelter management for different sectors.

Two-sample t-test

The purpose of t-test is to find if there is a significant mean difference between transitional and emergency program models. Here I have used the Scipy package to complete two t-tests.

The first t-test I have conducted is an independent two-sample t-test. This test is used when comparing the means of two independent groups to see if there is a statistically significant difference between them. As a result, I got the T-statistics equal to -39.07 and a p-value of 0.

```
T statistics = -39.0749698065413
P value = 0.0
```

The T-statistic is a measure of difference between two groups in terms of the number of standard errors. Such a negative T-statistic indicated that the first group has a lower mean than the second group. The magnitude of the T-statistic is quite large, which suggests a substantial difference between the means of the two groups being compared. Meanwhile, a p-value is the probability of obtaining test results at least as extreme as the observed results, under the assumption that the null hypothesis (which is there is no difference between groups) is true. P-value shows exactly 0

here may be the result that statistical software output will round the result for p value less than 0.01 and 0.001, which means that the observed difference in means is highly statistically significant, and the likelihood that this difference occurred by chance is extremely small.

Given these results, I could be quite confident that there is a significant difference between the two groups. However, these results are contingent upon the quality of the data and the validity of the assumptions underlying the T-test (such as normality, independence, and equal variances). Hence, to further prove this conclusion, I kept doing the Welch's T-test by setting equal variances as false.

```
T statistics = -40.98111537219914
P value = 0.0
```

The similarity in results between the two tests suggests that the findings are robust, even when accounting for potential differences in variances between groups. This adds confidence to my previous conclusion.

Conclusion

This report includes the overall EDA process of investigating the trend of shelter for homeless people in Toronto, as well as two different t-tests to investigate the difference among shelter programs. As a result, there is a significant increasing trend in both homeless people and shelters in Toronto since January 2021, and shelters for men and women are easily occupied fully compared to families and youth. Last but not least, there is also a statistically significant difference in mean between different shelter programs.