Yunfan Liu
#1010006459
INF2178 Assignment1

The dataset provides a daily list of active overnight shelters and related services in the Shelter Management Information System (SMIS) database of the Toronto 2021 Shelter Support and Housing Administration. The data contains 50,944 rows and 14 variables: OCCUPANCY_DATE, ORGANIZATION_NAME, PROGRAM_ID, PROGRAM_NAME, SECTOR, PROGRAM_MODEL, OVERNIGHT_SERVICE_TYPE, ROGRAM_AREA, SERVICE_USER_COUNT, CAPACITY_ACTUAL_BED, OCCUPANCY_BEDS, CAPACITY_ACTUAL_ROOM, and OCCUPANCY_ROOMS. In this study, I will focus on the following variables: OCCUPANCY_DATE, PROGRAM_MODEL, SERVICE_USER_COUNT, CAPACITY_ACTUAL_BED, OCCUPANCY_BEDS, CAPACITY_ACTUAL_ROOM, and OCCUPANCY_ROOMS, to study the trend of shelter usage.

Data Preprocessing

First, I got ideas of the distribution of the data from the data description (Figure 1). The data shows that the sheltered housing service fluctuates significantly in terms of the number of service users, beds, and room capacity. The average number of users served is 45.72, but the standard deviation is 53.33, which means that sometimes only a very small number of users are served to provide, while sometimes many users may be served. For beds, while the average number of actual beds was 31.63 and the average number of occupied beds was 29.78, suggesting that most beds are usually occupied, there was still a high degree of variability in the number of beds (with a standard deviation of 27.13 beds). The room data showed a similar pattern, with a mean actual room capacity of 55.55 and a mean number of occupied rooms of 52.79, accompanied by great volatility (standard deviation of nearly 59 rooms). Additionally, there is a significant difference between the maximum values of the number of beds and rooms and the 75th percentile, highlighting high occupancy rates at certain extremes. These extreme values may point to peaks in service utilization caused by specific events or special circumstances.

	SERVICE_USER_COUNT	CAPACITY_ACTUAL_BED	OCCUPIED_BEDS	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS
count	50944.000000	32399.000000	32399.000000	18545.000000	18545.000000
mean	45.727171	31.627149	29.780271	55.549259	52.798598
std	53.326049	27.127682	26.379416	59.448805	58.792954
min	1.000000	1.000000	1.000000	1.000000	1.000000
25%	15.000000	15.000000	14.000000	19.000000	16.000000
50%	28.000000	25.000000	23.000000	35.000000	34.000000
75%	51.000000	43.000000	41.000000	68.000000	66.000000
max	339.000000	234.000000	234.000000	268.000000	268.000000

Figure 1

Then, I performed some initial preprocessing on the dataset. According to the statistical summary of the data (Figure2), it contains two objects, four floats, one integer, and one datetime. I first set the OCCUPANCY_DATE as the index to facilitate subsequent exploration of the time series data provided by the shelter. Regarding the missing value, the variable PROGRAM_MODEL contains two missing elements, which is a small number relative to the size of the entire dataset. Therefore, I considered deleting the rows that contain these missing values. Second, the variables

'CAPACITY_ACTUAL_BED', 'OCCUPIED_BEDS', 'CAPACITY_ACTUAL_ROOM', and 'OCCUPIED_ROOMS' also have missing values. However, there is a complementary relationship between BED and ROOM; that is, the shelter either provides only beds, or only rooms. In other words, these missing values do not represent data errors. Consequently, I'm considering using 0 to indicate that either a bed or room was not provided. (Figure 2)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50944 entries, 0 to 50943
Data columns (total 8 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   OCCUPANCY_DATE        50944 non-null  datetime64[ns]
 1   CAPACITY_TYPE         50944 non-null  object
 2   PROGRAM_MODEL         50942 non-null  object
 3   SERVICE_USER_COUNT    50944 non-null  int64
 4   CAPACITY_ACTUAL_BED   32399 non-null  float64
 5   OCCUPIED_BEDS         32399 non-null  float64
 6   CAPACITY_ACTUAL_ROOM  18545 non-null  float64
 7   OCCUPIED_ROOMS        18545 non-null  float64
dtypes: datetime64[ns](1), float64(4), int64(1), object(2)
memory usage: 3.1+ MB
```

Figure2

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 50942 entries, 2021-01-01 to 2021-12-31
Data columns (total 7 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   CAPACITY_TYPE         50942 non-null  object
 1   PROGRAM_MODEL         50942 non-null  object
 2   SERVICE_USER_COUNT    50942 non-null  int64
 3   CAPACITY_ACTUAL_BED   50942 non-null  float64
 4   OCCUPIED_BEDS         50942 non-null  float64
 5   CAPACITY_ACTUAL_ROOM  50942 non-null  float64
 6   OCCUPIED_ROOMS        50942 non-null  float64
dtypes: float64(4), int64(1), object(2)
memory usage: 3.1+ MB
```

Figure 2

EDA

First, I initiated my exploratory analysis. Initially, I examined the trends of service user monthly by summarizing the number of service user count each month. The 'Monthly Service User Trends' graph (Figure 3) illustrates an overall gradual upward trend in the demand for shelter services. From the start of the year to mid-year, experienced several fluctuations, with notably low counts in February, after which the counts gradually increased. There was a marked increase in demand during the second half of the year, peaking in November and December. This pattern may reflect the impact of certain seasonal factors or an increase in demand due to specific events or changes in policy.

Yunfan Liu
#1010006459
INF2178 Assignment1

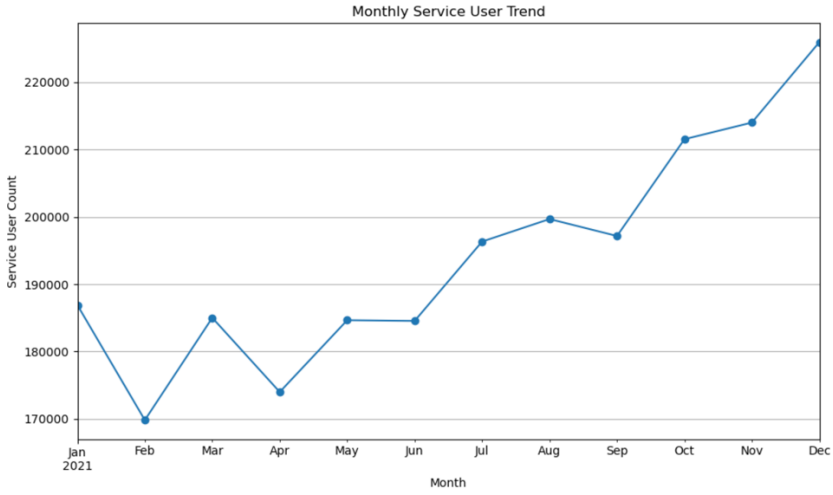

Figure3

Next, I analyzed the monthly trends in utilization based on different capacity types. Before conducting the trend analysis, I examined the distributions of room capacity, room occupancy, bed capacity, and bed occupancy separately using box plots (Figure 4). The plots reveal that the distributions of these four variables are right-skewed and contain lots of outliers. Given the social nature of these data, such outliers may reflect the real-world diversity; in other words, these outliers are likely real and valid observations rather than being measurement errors or input mistakes. Secondly, retaining outliers at the EDA stage provides a comprehensive understanding of the data's characteristics, leading me to decide to keep them.

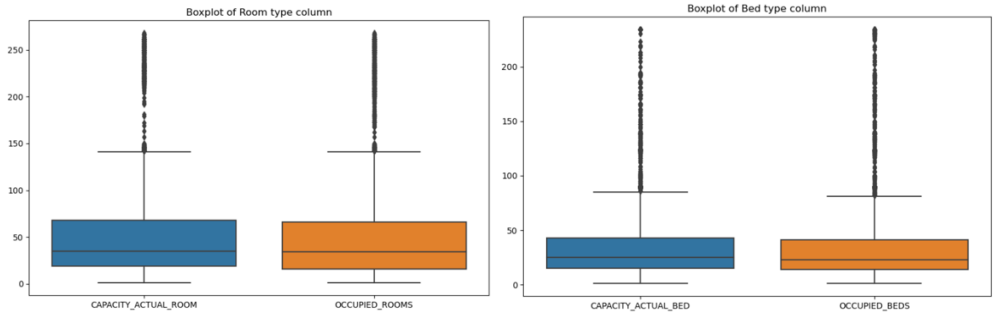

Figure 4

Subsequently, I aggregated the monthly room and bed utilization data and compared them using a line graph (Figure 5) that illustrates the different trends in utilization for the two capacity types (room-based and bed-based) over the course of a year. Overall, both room-based and bed-based utilization rates exhibit a brief upward trend at the beginning of the year. After February, both types of shelter capacity show a downward trend, maybe attributable to the summer season, which brings warmer temperatures and reduces reliance on shelter. However, from June until the end of the year, both room-based and bed-based utilization rates continue to rise, peaking especially at year-end. This may be due to increased demand for shelter as the winter weather gets colder. It worth to say, room-based utilization is generally lower than bed-based utilization before June but exceeds bed-based utilization after June. This difference could be attributed to seasonal effects, market prices, or changes in holiday patterns.

Yunfan Liu
#1010006459
INF2178 Assignment1

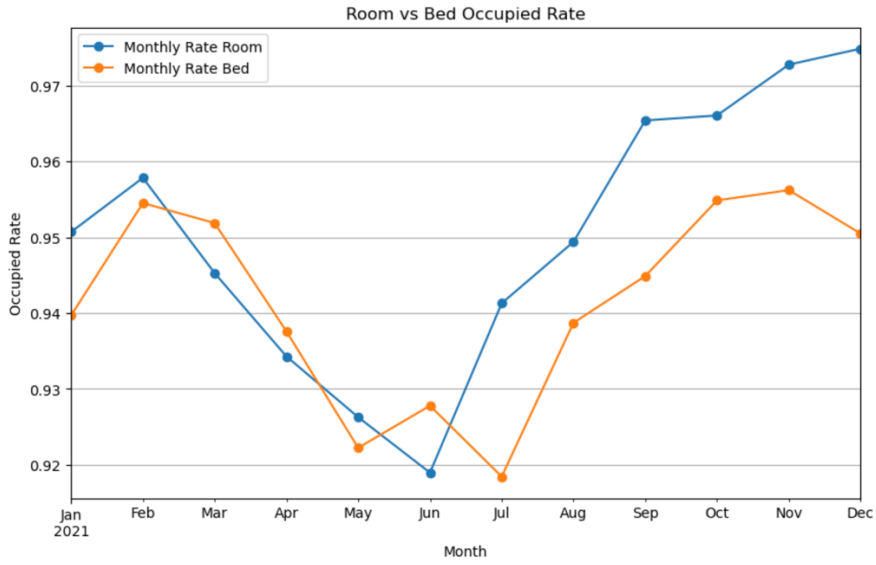

Figure 5

I also analyzed the average occupancy rates based on different program models. The grouped bar charts (Figure 6) indicate that the emergency model exhibits extremely high bed and room occupancy rates, reflecting its crucial role in addressing emergency needs. The transitional model also demonstrates high bed occupancy, but relatively low room occupancy, suggesting that this type of shelter facility relies more on beds than rooms to provide services. These findings imply that both models experience high occupancy rates for accommodation resources, which are nearly at full capacity. This situation may necessitate service providers considering the expansion of capacity or the improvement in the allocation of accommodation resources to meet the ongoing high demand.

	Program Model	Average Bed Occupancy Rate	Average Room Occupancy Rate
0	Emergency	0.9435	0.9386
1	Transitional	0.8852	0.8232

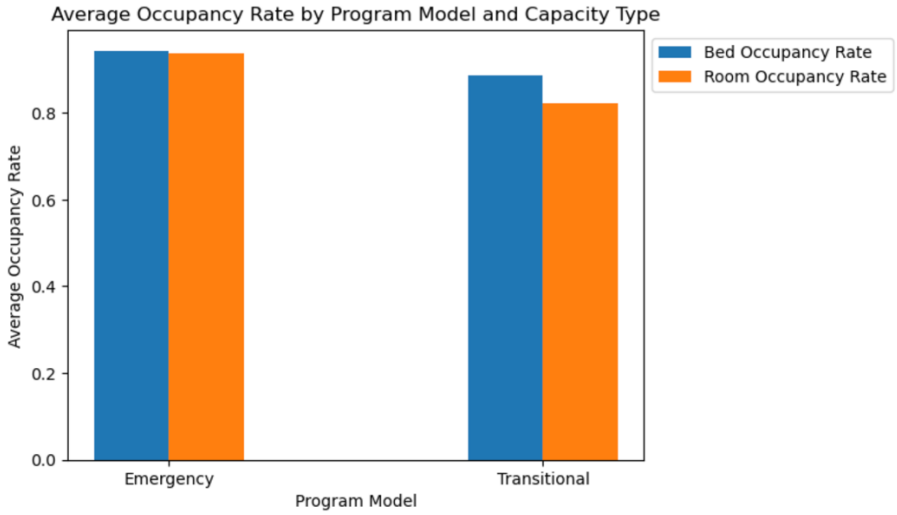

Figure 6

T-test

Finally, I compared the bed and room occupancy rates for the emergency and transitional models to determine if the differences between the two were significant. Before this, I conducted a normality test and a variance chi-square test. Given that the

Yunfan Liu
#1010006459
INF2178 Assignment1

sample size is large, according to the central limit theorem, the distribution of the sample mean tends to be normally distributed, even if the raw data itself is not. Additionally, the variance chi-square test shows a p-value ≤ 0.05, indicating sufficient evidence of unequal variances. At the Welch's t-test stage, the results reveal that there is a significant difference ($p < 0.05$) in bed occupancy between the Emergency Program model and the Transitional Program model, suggesting that these two models have differing effects on meeting bed demand. Similarly, for room occupancy, a significant difference ($p < 0.05$) is observed between the Emergency Program model and the Transitional Program model, indicating that these two models exhibit different effects in satisfying room demand. However, due to the large sample size in this case, even a small actual difference may still result in statistical significance. Therefore, further analysis of the effect size is necessary to assess the true significance of this difference, to provide more accurate planning for shelter services and resource allocation.