Experimental Design for Data Science A1 INF 2178 Xiaoxin Zhou 1003549742

Introduction and research questions:

Before reading the dataset, I understood that it is a record of shelter services that support the city's homeless by providing a place to stay overnight. However, due to not having enough shelter spaces to accommodate many homeless people, they are often being turned away. **The research questions** can be:

Determine which program experiences the highest homeless visitation rates and assess the potential costs if demand exceeds supply.

Explore whether arrangements based on rooms or beds influence the preferences of the homeless in choosing overnight accommodation.

Can we identify the reasons for supply not meeting demand through a T-test?

	OCCUPANCY_DATE	ORGANIZATION_NAME	PROGRAM_ID	PROGRAM_NAME	SECTOR	PROGRAM_MODEL	OVERNIGHT_SERVICE_TYPE	F	PROGRAM_AREA	SERVICE_USER_COUNT	CAPACITY_TYPE	CAPACITY_ACTUAL_BED	OCCUPIED_BEDS	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS
0	2021-01-01	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
1	2021-01-01	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
2	2021-01-01	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
3	2021-01-01	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
4	2021-01-01	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
...	...	...	...	...	...	...	...		...	...	...	...	...	...	...
50939	2021-12-31	YWCA Toronto	14671	YWCA Davenport - Youth	Youth	Emergency	Shelter	(	Base Shelter and Overnight Services System	6	Bed Based Capacity	20.0	6.0	NaN	NaN
50940	2021-12-31	Youth Without Shelter	12292	Youth without Shelter Emergency Shelter Program	Youth	Emergency	Shelter	(	Base Shelter and Overnight Services System	23	Bed Based Capacity	23.0	23.0	NaN	NaN
50941	2021-12-31	Youth Without Shelter	12291	Youth without Shelter Stay In School Program	Youth	Transitional	Shelter	(	Base Shelter and Overnight Services System	13	Bed Based Capacity	14.0	13.0	NaN	NaN
50942	2021-12-31	YouthLink	14891	YouthLink Emergency Program	Mixed Adult	Emergency	Shelter	(	Base Shelter and Overnight Services System	10	Bed Based Capacity	10.0	10.0	NaN	NaN
50943	2021-12-31	YouthLink	14911	YouthLink Transitional Program	Mixed Adult	Transitional	Shelter	(	Base Shelter and Overnight Services System	29	Bed Based Capacity	29.0	29.0	NaN	NaN

50944 rows × 14 columns

Fig 1 After loading the dataset, it is evident that it contains 50,944 rows and 14 columns. These columns are as follows:

OCCUPANCY_DATE: 365 days date

ORGANIZATION_NAME: 35 different organizations

PROGRAM_ID: numbers

PROGRAM_NAME: 167 different programs

SECTOR: ['Families', 'Mixed Adult', 'Men', 'Women', 'Youth'] different sector of homeless

PROGRAM_MODEL: ['Emergency', 'Transitional', nan] different types of program model

OVERNIGHT_SERVICE_TYPE: 8 overnight service type includes NAN

PROGRAM_AREA: 5 programs area includes NAN

SERVICE_USER_COUNT: number of user (homeless)

CAPACITY_TYPE: ['Room Based Capacity', 'Bed Based Capacity']

CAPACITY_ACTUAL_BED: number of beds have

OCCUPIED_BEDS: number of beds in use

CAPACITY_ACTUAL_ROOM: number of rooms have

OCCUPIED_ROOMS: number of rooms in use

Each column represents a specific attribute of the data, offering a comprehensive view of shelter usage trends in Toronto. This dataset provides a rich source of information for analysis, encompassing various aspects such as occupancy dates, organizational details, program specifics, sector classifications, and capacity metrics.

Pre-Process/Clean Data:

- Remove meaningless columns as:
    - RGANIZATION_NAME, PROGRAM_ID, PROGRAM_NAME, OVERNIGHT_SERVICE_TYPE, SECTOR, OCCUPANCY_DATE, PROGRAM_AREA
- CAPACITY_TYPE split the rooms and beds.
- PROGRAM_MODEL split the Emergency and Transitional model.
- Drop NAN data if the dataset have empty

Data Visualization:

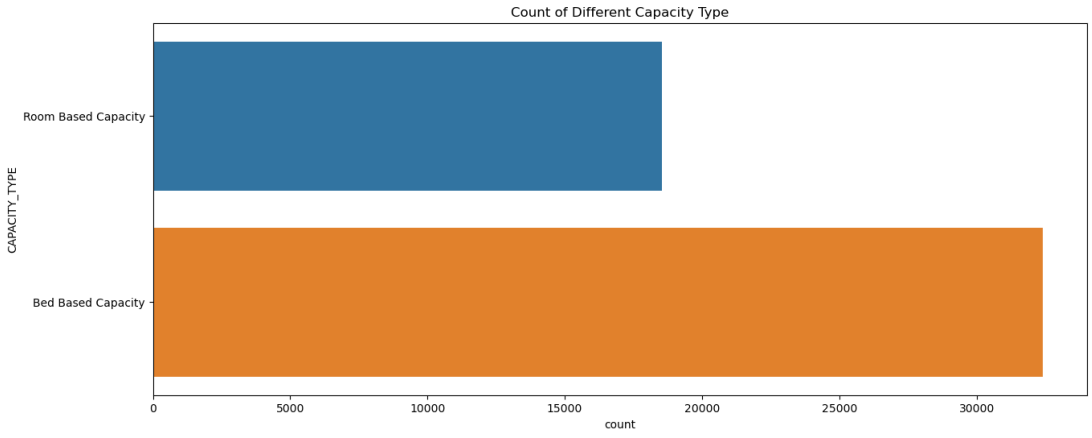

Fig 2 To visualize the number of rooms and beds, we divided the data into different subsets based on capacity type. The next process will divide the data into different subsets based on the program model.

From Figures 3 and 4, we can clearly observe that the bed and room capacities for 'Emergency' and 'Transitional' program types are significantly different.

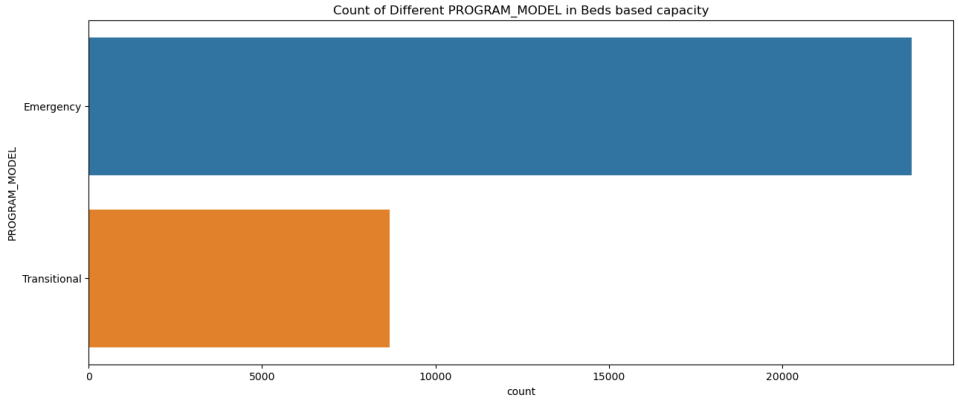

Fig 3 visualizes the varying room capacities across different program models.

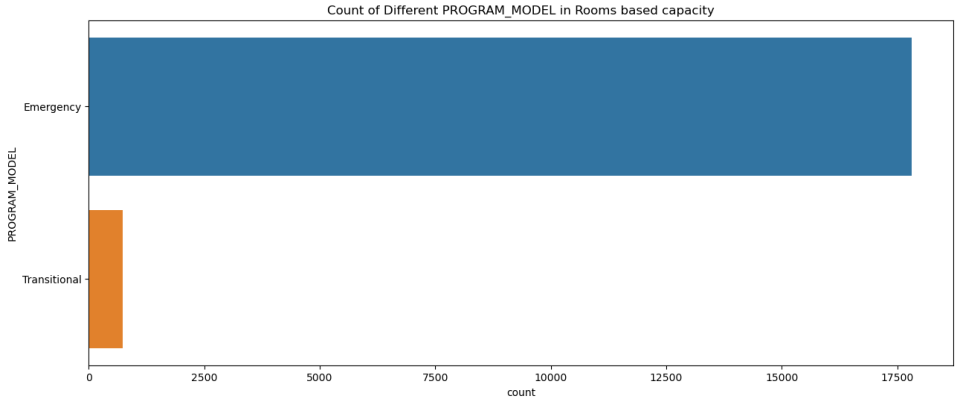

Fig 4 visualizes the varying beds capacities across different program models.

Quantitative Result:

In this section, I will carry out outlier detection and remove any outliers before conducting further analysis. Initially, we will review a box plot in Figure 5 to visualize the data prior to the detection and removal of outliers. Subsequently, in the following section, we will remove the outliers and apply the box plot once again to assess the changes in the data distribution in Figure 6.

In Figure 7, I will present correlation heatmaps for all four datasets: dataset_bed_Emergency, dataset_bed_Transitional, dataset_room_Emergency, and dataset_room_Transitional. Based on the results, it becomes evident that, across all subsets, the actual number of service users, actual beds, and occupied beds are strongly correlated.

In the final stage of our quantitative analysis, I will conduct t-tests on the subsets. The goal is to analyze the differences between bed and room capacities and the variations between different program models. To answer the research question. To compare the visitation rate, we can perform t-tests between the different program models for both room and bed capacity.

Hypothesis for visitation rates same capacity with different program

H0 (null hypothesis): There is no significant difference in the mean visitation rate or preference between the two groups being compared (Emergency vs Transitional in bed-based capacity, Emergency vs Transitional in room-based capacity).

H1 (alternative hypothesis): There is a significant difference.

	T-statistic	P-value
Bed-based Capacity (Emergency VS Transitional)	27.414	1.36×10 ^−163
Room-based Capacity (Emergency VS Transitional)	-18.898	6.50×10 ^−79

The positive t-statistic indicates that the mean visitation rate for the Emergency program model in bed-based capacity is significantly higher than for the Transitional model. The extremely small p-value, which is practically zero, suggests that this difference is highly statistically significant. This result allows us to confidently reject the null hypothesis (H0) and accept the alternative hypothesis (H1), indicating a significant difference in visitation rates between the Emergency and Transitional programs in the context of bed-based capacity. The negative t-statistic here suggests that the mean visitation rate for the Emergency program model in room-based capacity is significantly lower than for the Transitional model. Again, the extremely small p-value reinforces the statistical significance of this finding. This result leads us to reject the null hypothesis in favor of the alternative hypothesis, indicating a significant difference in visitation rates, but in the opposite direction compared to the bed-based capacity. Owing to the inherently transient nature of emergency accommodation in Bed-based Capacity settings, there is a markedly higher turnover rate compared to Room-based Capacity arrangements. This elevated fluidity is a direct consequence of the short-term lodging typically offered in emergency scenarios. As a result, bed-based facilities frequently experience a higher rate of vacancy. This pattern of occupancy suggests that while these beds are crucial for providing immediate shelter in urgent situations, their usage fluctuates significantly, often leading to periods where many beds remain unoccupied. Understanding this dynamic is essential for optimizing the allocation and management of these critical resources, ensuring that they are available for those in need while minimizing inefficiencies in their utilization.

<u>Hypothesis for visitation rates same program with different capacity</u>

H0 (null hypothesis): There is no significant difference in the mean visitation rate between the Emergency program in bed-based capacity and the Emergency program in room-based capacity. (bed-based capacity vs room-based capacity in Emergency, bed-based capacity vs room-based capacity in Transitional).

H1 (alternative hypothesis): There is a significant difference.

	T-statistic	P-value
Emergency Program (Beds VS Rooms)	-75.758	0
Transitional Program (Beds VS Rooms)	-105.348	0

The results of the t-test corroborate our previous speculation that room-based capacity is more popular than bed-based capacity. The fact that all the t-statistics are negative, and the p-values are essentially zero, suggests that these differences are highly statistically significant. This allows us to

confidently reject the null hypothesis (H0) and accept the alternative hypothesis (H1), indicating the significant difference in visitation rates between room-based and bed-based capacities in both Emergency and Transitional program contexts. These results highlight a clear preference or need for room-based accommodation over bed-based ones in both Emergency and Transitional programs.

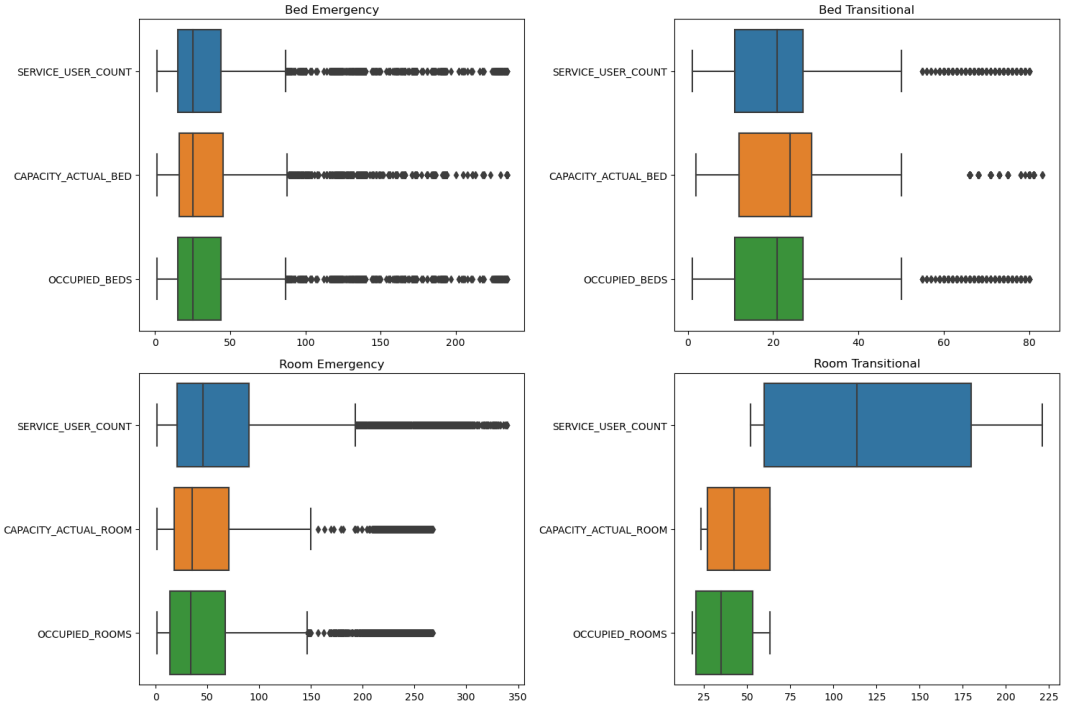

Fig 5 Box plot(before) for bed Emergency, bed Transitional, room Emergency, room Transitional.

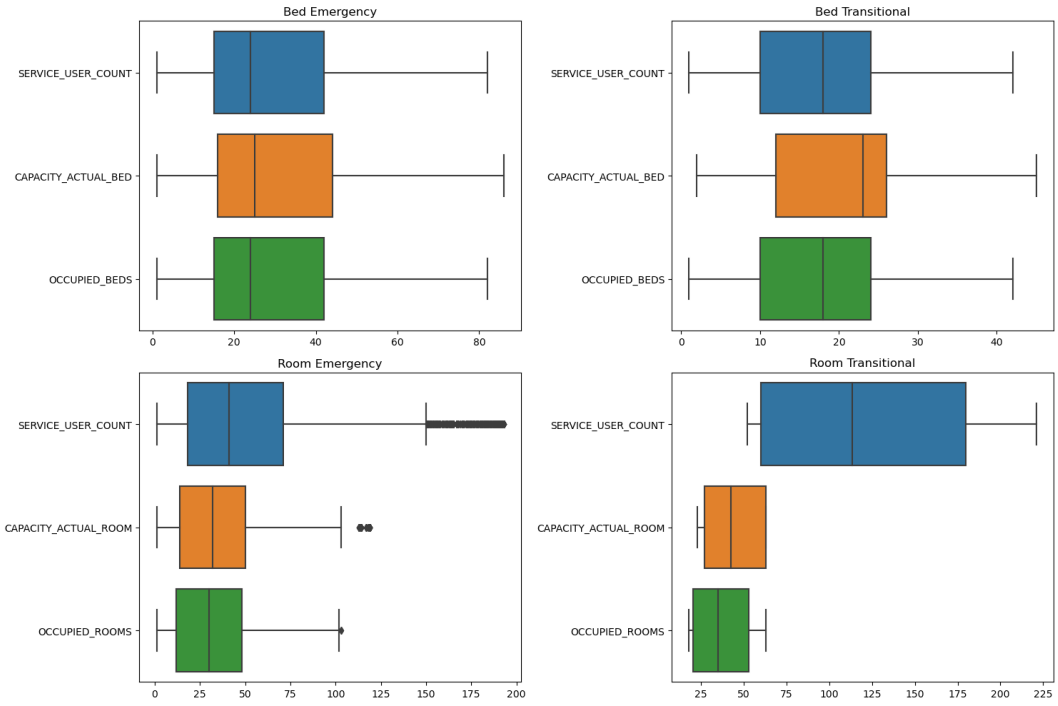

Fig 6 Box plot(after) for bed Emergency, bed Transitional, room Emergency, room Transitional.

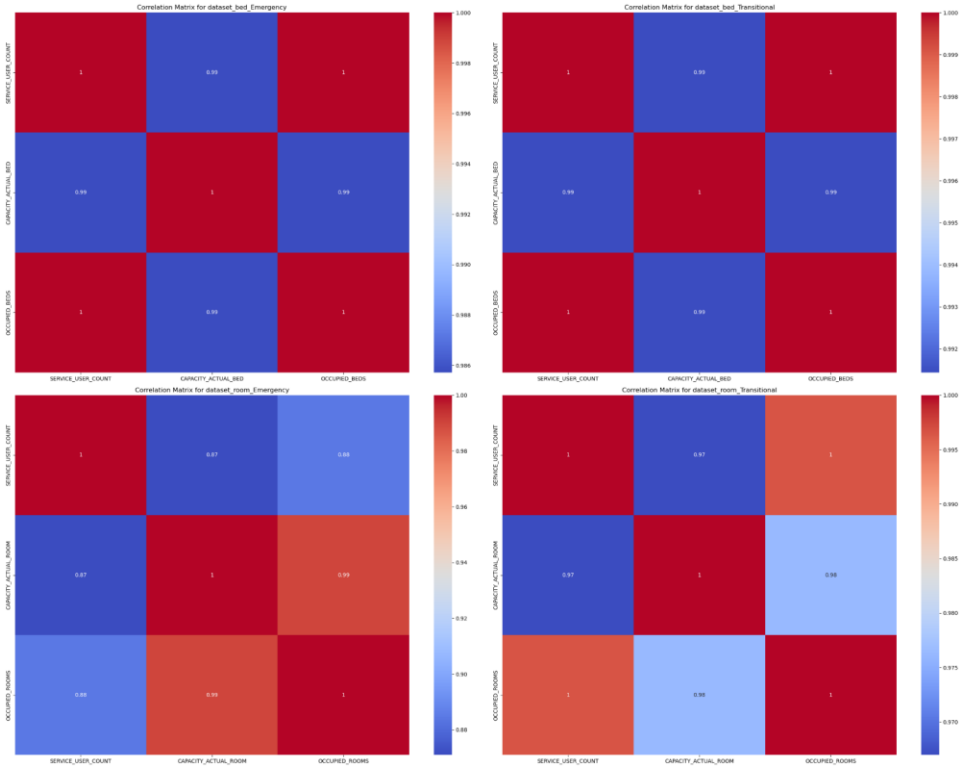

Fig 7 Correlation heatmaps for all four datasets: dataset_bed_Emergency, dataset_bed_Transitional, dataset_room_Emergency, and dataset_room_Transitional.