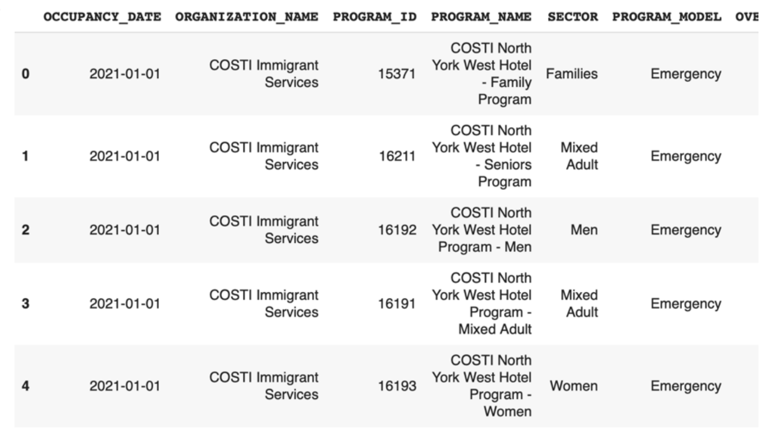
Technical Assignment 1:
Examining Toronto's Shelter Usage Trends

Devin W. de Silva
Faculty of Information
School of Graduate Studies, University of Toronto
INF2178: Experimental Design for Data Science
Professor Shion Guha
February 4, 2024

Over the past years, I've read quite a bit about Toronto's growing homelessness issues and the overflowing of shelters, but by exploring it through the data analysis below, I realized that there's a lot of subtleties relating to shelter use in Toronto.

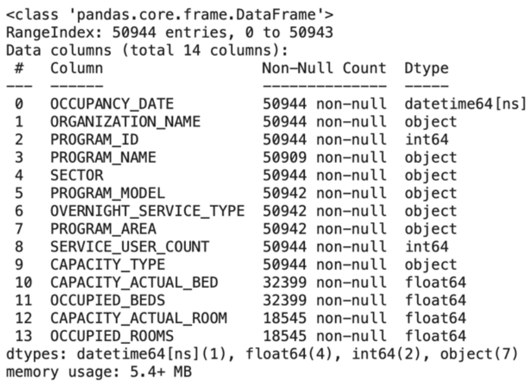
1. Loading and Cleaning the Dataset:

I begin by perusing the Excel document itself and loading the data to the Colab file. Before any data exploration could happen, I need to clean the dataset, to remove the many NaN values, and keep only the information relevant to my analysis. Mainly, I merged the four columns "CAPACITY_ACTUAL_BEDS", "OCCUPIED_BEDS", "CAPACITY_ACTUAL_ROOMS", "OCCUPIED_ROOMS" to the two simple columns "Capacity" and "Occupied". This removes much NaN values and we could still identify and sort all the data by "CAPACITY_TYPE" into the two categories of "room" and "bed". I then calculated the occupancy rate by dividing "Occupied" by "Capacity".

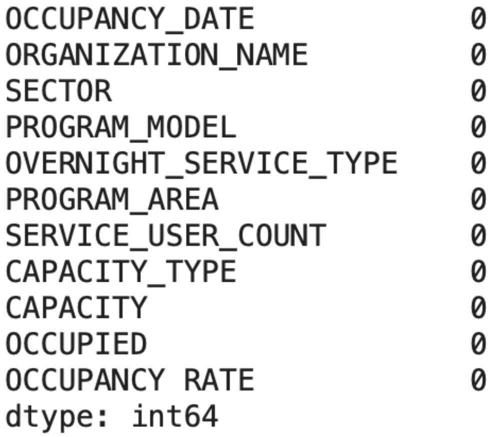


	OCCUPANCY_DATE	ORGANIZATION_NAME	PROGRAM_ID	PROGRAM_NAME	SECTOR	PROGRAM_MODEL	OVE
0	2021-01-01	COSTI Immigrant Services	15371	COSTI North York West Hotel - Family Program	Families	Emergency	
1	2021-01-01	COSTI Immigrant Services	16211	COSTI North York West Hotel - Seniors Program	Mixed Adult	Emergency	
2	2021-01-01	COSTI Immigrant Services	16192	COSTI North York West Hotel Program - Men	Men	Emergency	
3	2021-01-01	COSTI Immigrant Services	16191	COSTI North York West Hotel Program - Mixed Adult	Mixed Adult	Emergency	
4	2021-01-01	COSTI Immigrant Services	16193	COSTI North York West Hotel Program - Women	Women	Emergency	

Identifying column data types to clean and dataset with NaN values removed:

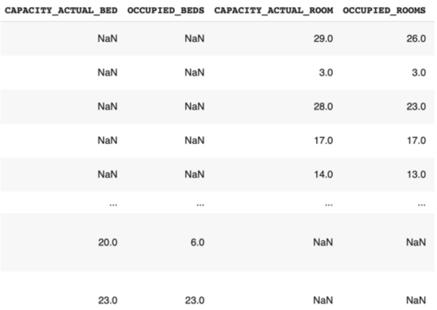


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50944 entries, 0 to 50943
Data columns (total 14 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   OCCUPANCY_DATE         50944 non-null  datetime64[ns]
 1   ORGANIZATION_NAME      50944 non-null  object
 2   PROGRAM_ID             50944 non-null  int64
 3   PROGRAM_NAME           50909 non-null  object
 4   SECTOR                 50944 non-null  object
 5   PROGRAM_MODEL          50942 non-null  object
 6   OVERNIGHT_SERVICE_TYPE  50942 non-null  object
 7   PROGRAM_AREA           50942 non-null  object
 8   SERVICE_USER_COUNT     50944 non-null  int64
 9   CAPACITY_TYPE          50944 non-null  object
 10  CAPACITY_ACTUAL_BED    32399 non-null  float64
 11  OCCUPIED_BEDS          32399 non-null  float64
 12  CAPACITY_ACTUAL_ROOM   18545 non-null  float64
 13  OCCUPIED_ROOMS         18545 non-null  float64
dtypes: datetime64[ns](1), float64(4), int64(2), object(7)
memory usage: 5.4+ MB
```



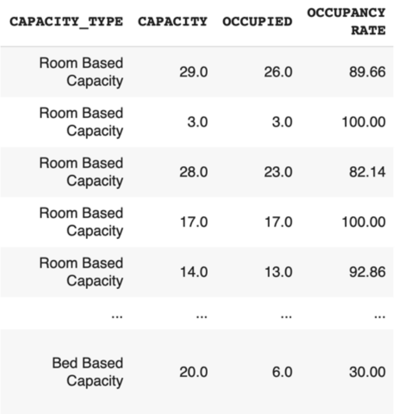
```
OCCUPANCY_DATE            0
ORGANIZATION_NAME         0
SECTOR                    0
PROGRAM_MODEL             0
OVERNIGHT_SERVICE_TYPE    0
PROGRAM_AREA              0
SERVICE_USER_COUNT        0
CAPACITY_TYPE             0
CAPACITY                  0
OCCUPIED                  0
OCCUPANCY RATE            0
dtype: int64
```

Capacity and Occupied before cleaning:



CAPACITY_ACTUAL_BED	OCCUPIED_BEDS	CAPACITY_ACTUAL_ROOM	OCCUPIED_ROOMS
NaN	NaN	29.0	26.0
NaN	NaN	3.0	3.0
NaN	NaN	28.0	23.0
NaN	NaN	17.0	17.0
NaN	NaN	14.0	13.0
...	...	...	...
20.0	6.0	NaN	NaN
23.0	23.0	NaN	NaN

Capacity and Occupied after cleaning:

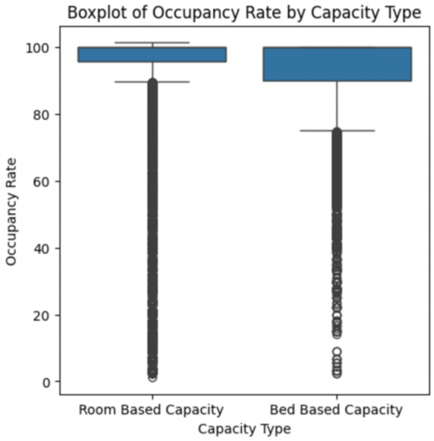


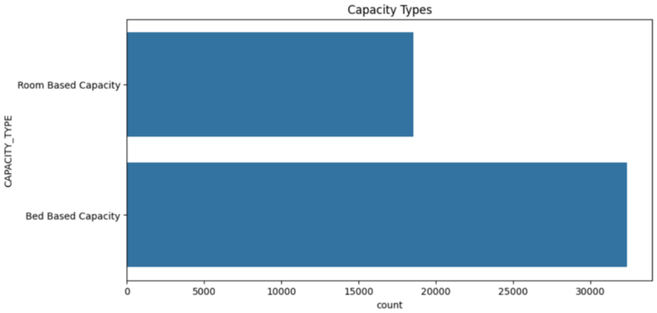
CAPACITY_TYPE	CAPACITY	OCCUPIED	OCCUPANCY RATE
Room Based Capacity	29.0	26.0	89.66
Room Based Capacity	3.0	3.0	100.00
Room Based Capacity	28.0	23.0	82.14
Room Based Capacity	17.0	17.0	100.00
Room Based Capacity	14.0	13.0	92.86
...	...	...	...
Bed Based Capacity	20.0	6.0	30.00

The first main chunk of data exploration I completed has to do with identifying the differences between the room and bed-based capacities. I started by performing a t-test to identify whether there is a significant difference between the occupancy rates of the two capacities. The results are T = 4.845526591877832 with p =1.2664700757084775e-06

The results reject the null hypothesis that there is no significant difference, given the large T and the very small p value. There is an evident difference in occupancy rates.

My next step, therefore, is to illustrate this difference with figures. The boxplot below shows that room-based capacity is generally indeed higher and less variable than those for bed-based capacity, even though both have plenty of outliers. The other chart shows the absolute numbers of the two, with there being a larger absolute amount of bed based capacity.

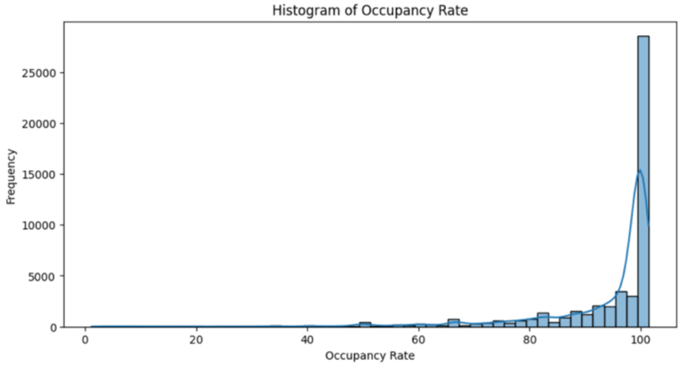




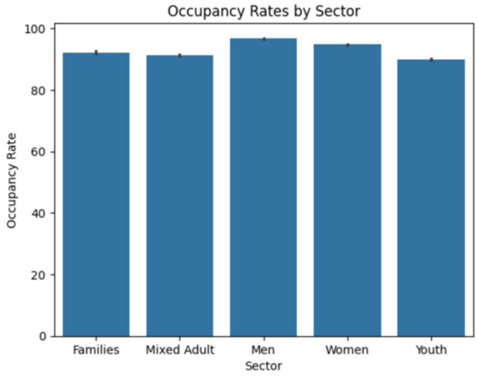
I proceeded to learn more about occupancy rates of the shelters over the course of 2021, beginning by collecting some summary statistics:

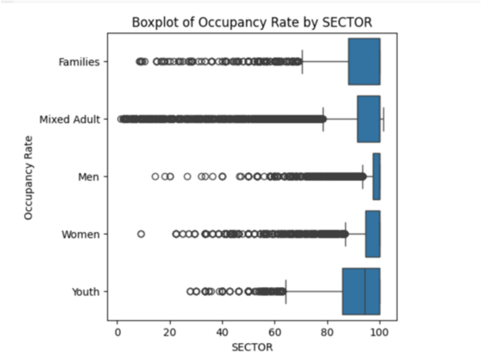
Mean: 93.01
Median: 100.0
Min Value: 1.2
Max Value: 101.41
25th Quartile: 92.31
75th Quartile: 100.0

With the overwhelming majority of the data falling well beyond the 90th percentile, this shows a very challenging situation at the shelters, as they shelters are consistently almost fully occupied. There are, however, a few extremely low outliers dragging the mean below the median (which may be a better measure of central tendency here). I created the histogram below to illustrate this point more visually.

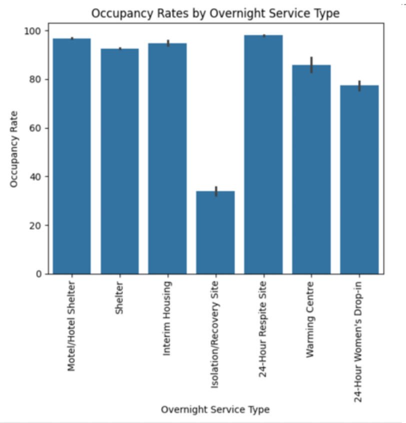


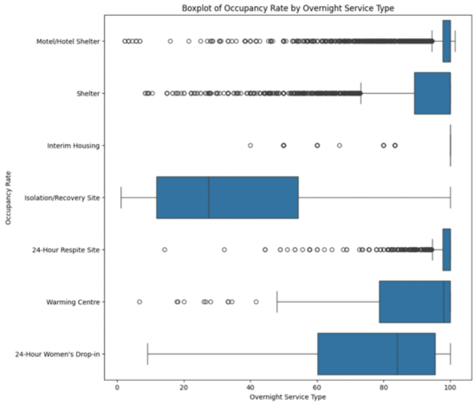
Now, I try to scrutinize hidden subtleties in the occupancy rates in more detail, beginning by looking dividing it up by sectors. The rates shown on the bar chart seem fairly consistent visually, and the boxplot shows us a bit more information about how the spread (e.g. Occupancy for men seem to be usually nearing full, despite many outliers, while the youth occupancy rates are more variable)



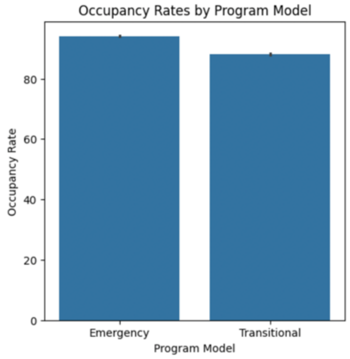


In the same way as above, I created plots for overnight service type too. Here, we can see that most service types still have an occupancy rate over 80% over the year, with one evident exception, that of isolation/recovery site, which only holds around 35 % capacity. The boxplot for this site also seem much less skewed than the others.





I proceeded to learn more about the occupancy rates by program model, by plotting it and doing a t-test too. The t-statistic is 39.075125805845744 while the p value is 0, showing a significant difference in occupancy rates between emergency and transitional models.

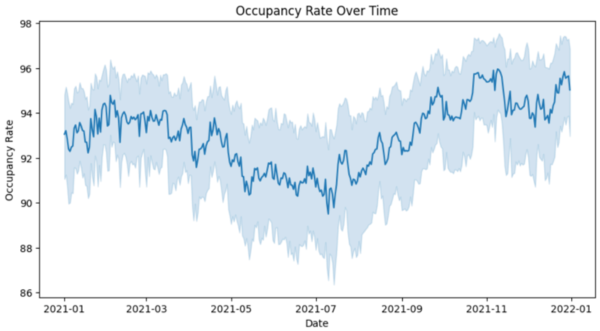


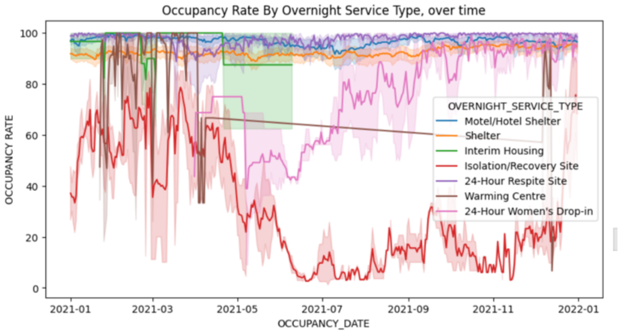
Next, I furthered my exploration of the above data by adding the dimension of time into account. This first chart below shows a fluctuating pattern over the year, dipping down during the summer months, but clearly rising over the autumn and fall to a higher level than when the year first started.

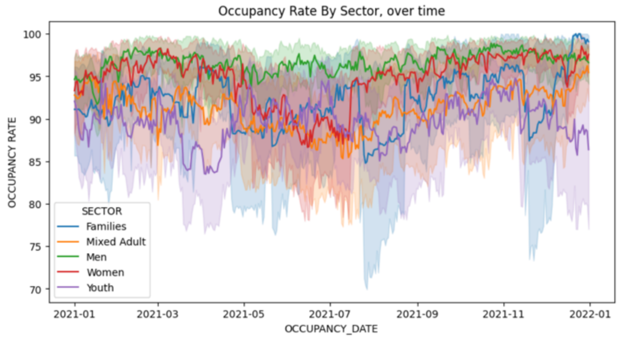
When overnight service type is taken into account, we see that most types are almost full occupied throughout the year, the only two with evident seasonal changes are 24-Hour women drop in, which massively increased since about June, while the Isolation and Recovery Site reduced markedly since about the same time, but rebounded quickly nearing the end of the year – perhaps corresponding to the COVID situation in Toronto at the time.

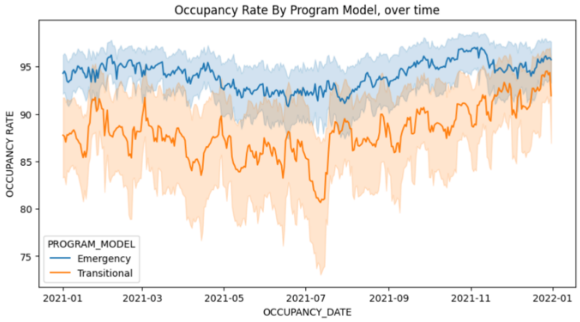
The different sectors also show some seasonal fluctuations and illustrates that men and women's occupancy rates are generally higher than the other sectors over the year. I conducted a t-test here too between men and women occupancy rates, with t=15.257267854146386 and p value=2.76838362506134e-52, meaning the difference in men's and women's shelter occupancy rates themselves are also statistically significant.

The occupancy rate by program model confirms the t test statistic I conducted earlier, showing two very distinct lines between emergency and transitional.


Occupancy Rate Over Time


Occupancy Rate By Overnight Service Type, over time


Occupancy Rate By Sector, over time

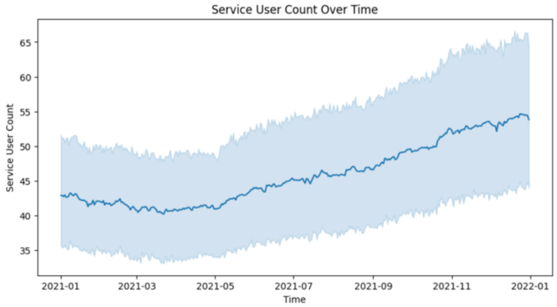

Occupancy Rate By Program Model, over time

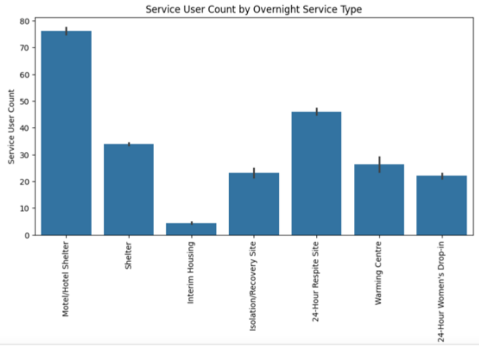
The last section of data exploration I included in this project has to do with service user count averaged per program. Again, I collected the summary statistics to begin with. It is clear that the data is skewed with a mean higher than the median.

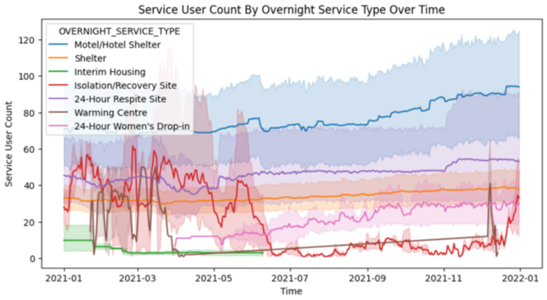
Mean: 45.73, Median: 28.0, Min Value: 1, Max Value: 339, 25th Quartile: 15.0, 75th Quartile: 51.0.

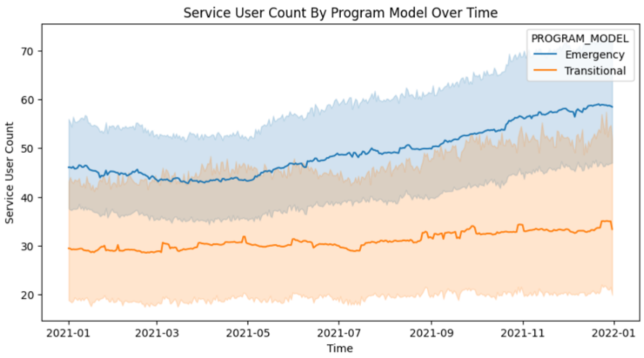
The first graph shows the alarming gradual increase in service user count over time.
If we filter these by overnight service type, we find that many are housed in hotels and motels, while relatively few are housed in interim housing. Plotting this over time shows that almost all types of service increased in usage over the year, with some fluctuating more than others, such as the isolation/recovery site, which reduced to nearly zero during the summer and fall, before rebouncing quickly in the winter.

The same increasing trend is shown in both program models too, and the average absolute amounts of emergency programs have always been markedly higher than the transitional ones. To confirm this, I conducted a t-test, with t=29.937570467283667 and p-value of 3.1720139638162956e-195, showing that there is indeed a significant difference in count of the the two program models, rejecting the null hypothesis.









All these analyses proves one key point – most homeless shelters in this city are already or nearing the brink of being overwhelmed, despite occasional seasonal changes. It is imperative to devote more resources to resolve this worsening situation and assist those most in need.