Data Science 2: Statistics for Data Science

Term Project - Group Two Report
Members: Walid Al Gherwi, Bryan Mallinson, Steven McAvoy, Vernon Naidoo, Mahshad Najafi Ragheb, Mohit Ramnani
Date: August 10 2020

Homeless Shelter Occupancy Analysis

Main Source of Data: City of Toronto Open Data Portal

https://open.toronto.ca/dataset/daily-shelter-occupancy/

Introduction

The ability to accurately anticipate the demand for shelter spaces can literally be a matter of life and death for society's most vulnerable people. This ability is required in order to have excess capacity in the short term for sudden spikes in demand (due to Weather) or to build (or remove) capacity in the long term due to economic factors like unemployment or inflation or the rising cost of housing.

Other factors like crime can also be correlated with homeless shelter usage - do factors like increases in domestic violence impact demand, and if so for all shelter types of specific types like family or women's shelters?

This paper will provide statistical analyses and commentary on these topics with collected insights and conclusions at the end.

Table of contents

References

Ali Jadidzadeh & Ron Kneebone, *Shelter from the Storm: Weather-induced patterns in the use of emergency shelters*, February 2015, The School of Public Policy Research Papers Volume 8, Issue 6 University of Calgary.

Daniel Wong, *Toronto And Vancouver See Cost Of Living Rise Over 20% Faster Than The Rest Of Canada,* Better Dwelling, March 26, 2018, https://betterdwelling.com/toronto-and-vancouver-see-cost-of-living-rise-over-20-faster-than-the-rest-of-canada/#

Unit 1: Objectives of this study

The Objective of this study is to examine the factors potentially increasing the demand for shelter beds as indicated by the daily occupancy. While we cannot prove direct causation, we can look to demonstrate strong correlation,

Below are the factors we chose to examine, with the sub-objectives for each factor:

Correlation of Available and Occupied Beds

Null Hypothesis: There is no statistically significant correlation between capacity and occupancy.

- If true, these two measures should have some degree of independence.
- If the current supply is adequate for the demand, increases in capacity would leave beds vacant.

Alternative Hypothesis: A statistically significant correlation exists between capacity and occupancy.

- If true, since capacity is controlled by shelter administration, further study may be warranted into quantifying the underlying need of homeless individuals as well as the number who are denied service each night.

CPI (Consumer Price Index), Rental Costs, Population and Shelter Occupancy

Examine the relationships between the cost of living in terms of CPI and occupancy. CPI represents the cost of a representative sample of goods and services, including the cost of rent. However, the cost of rent is directly tied to housing and its availability and was studied specifically.

Null Hypotheses: Neither CPI or rental costs are correlated with increased shelter occupancy.

Hypothesis 1: CPI is correlated with increased shelter occupancy

Hypothesis 2: Homeless shelter occupancy in the Toronto region is correlated between average cost of accommodation and population of Ontario.

Crime & shelter occupancy

The objective of analyzing Toronto policing data and comparing the main shelter dataset to:

1. Explore any correlations with either causing homelessness or criminal incidents resulting from increased homelessness;
2. Develop a predictive model using Bayesian inference to examine the likelihood of criminal activity based on proximity to shelters. i.e. Test if it is more likely for criminal activity to occur near a shelter.

Weather & shelter occupancy

A reasonable hypothesis is that inclement weather would make sleeping outdoors less appealing and persuade more people to make use of homeless shelters. We start with a null hypothesis and two hypotheses to test:

Null Hypothesis: Weather does not significantly impact demand for homeless shelters

Hypothesis 1: Negative temperatures correlate with higher demand for homeless shelters

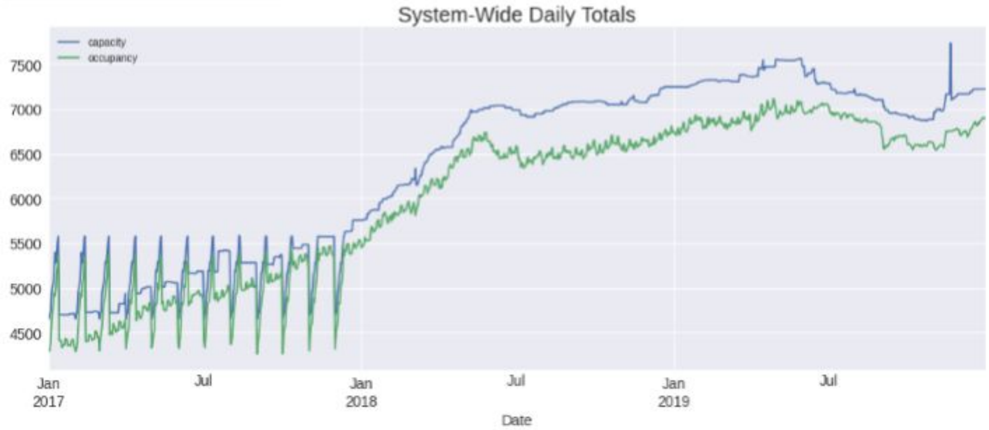
Hypothesis 2: Days with precipitation correlate with higher demand for homeless shelters

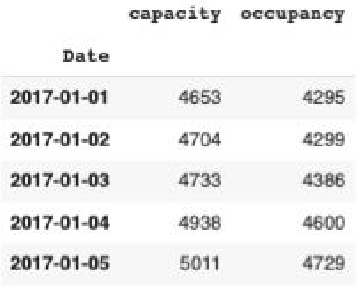
Unit 2: Dataset Preparation

Shelter Data Cleaning

Staff in (“homeless”) shelters in the City of Toronto, record both the number of beds available and the number occupied on a daily basis. Although the daily numbers in the datasets examined ranged from approximately 4500 to 7500, the data is reported at a very granular level, in some cases recording individual beds. Details provided include each facility’s name, location and managing organization as well as the sector being served (men, women, youth, etc.). The annual datasets, published via the Toronto Open Data Portal provided just under 40,000 detailed records per year.

This study of 36 months of data, captured between Jan, 1st, 2017 and Dec 31st, 2019, analyzed unusual trends within data subsets and examined possible correlations with external factors related to the environment, the economy and indicators of human behaviour, such as crime reports and population trends. The direction of the investigation was guided by papers written by experts in the field and by interviews with a co-chair of several shelters who reviewed our results and provided guidance, informed by years of experience working in this field.

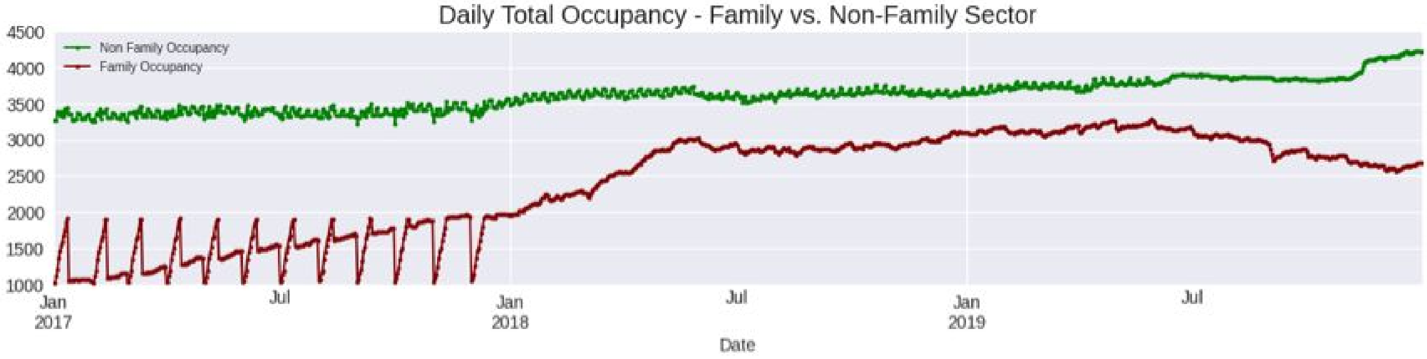




	capacity	occupancy
Date		
2017-01-01	4653	4295
2017-01-02	4704	4299
2017-01-03	4733	4386
2017-01-04	4938	4600
2017-01-05	5011	4729

Exclusion of Questionable Data: Family Sector

During Exploratory Data Analysis of the individual subsets within the prevailing time series trends, a specific data subset was identified with a predominant atypical pattern that significantly impacted (skewed) the overall data trends. Discussions with a subject matter expert revealed that the accuracy of family sector data has been in question for some time, due to a less stable delivery model and unsupervised data collection. Many of the family spaces are provided in motel and hotel rooms without sufficient oversight by shelter staff. The group decided to focus the study on the men, women, youth and co-ed sectors, which make up 3000 to 4000 beds per night.



Within the target sectors – men, women, youth, co-ed and family – an oscillating pattern was visible in 2017 but not seen in the other years or sectors. Overall, the non-family sectors were consistent with each other and had more steady daily trends across the three years reviewed.

There are missing values in the shelter postal code. In some cases, the shelter postal code already existed in the dataset in other rows. We replaced the missing values with the existing values. If the postal code was not available in the dataset, it was found by searching the shelter's address in a Google search.

CPI Data

The CPI data used in this analysis is downloaded from the Bank of Canada (CPI, 2000 to Present, Bank of Canada). CPI monthly data from the Bank of Canada is available as a CSV file and did not require further cleaning. The CPI data is aligned with the monthly averages of shelter data using pandas join method.

Rental Data

The data for average rent in Toronto was extracted from the Rental Market Report Archive of Toronto Regional Real Estate Board (TRREB). The archive consists of quarterly reports from every year. The numbers from these quarterly reports were picked and recorded in a csv file for use in the analysis.

The shelter occupancy data was resampled from daily to quarterly to match the date intervals of TRREB Rental Market data.

The data from TRREB Report Archive contained a different time interval from that of the shelter data. A conversion had to be made manually to change the format in rental data from Year-Quarter (Eg. 2019-Q3 for third quarter of 2019) to the last day of the quarter in yyyy-mm-dd like in shelter occupancy data in order for the merge function to work.

Ontario Population Data

The Ontario population data was extracted from Statistics Canada open data site (https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000901) . The dataset contains quarterly population data from the national census. The data was extracted to a csv file for use in the analysis.

Crime Incident Data:

In addition to the main shelter data set, the following data source was used, Toronto Police Service: Public Safety Data Portal, MCI 2014 to 2019 (https://data.torontopolice.on.ca/datasets/mci-2014-to-2019)

The data quality was excellent as it was provided through the Toronto Police Service and the nature of policing work also necessitates accurate documentation. However, it is noted that Toronto Police Service offsets the locations to the nearest intersection, for protection of privacy purposes, so the proximity model may be impacted as a result.

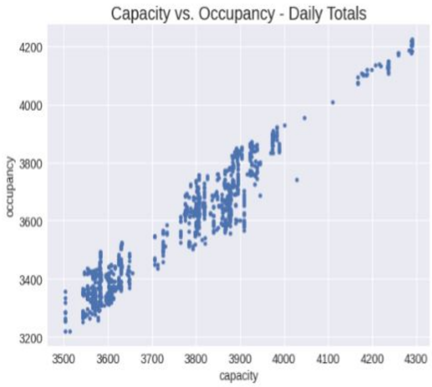
Data preparation required converting occurrence dates from strings to date-times, and removal of unnecessary columns. Columns kept were: Major Crime Indicators (MCI), offense, premise type, unique event ID, occurrence date, latitude and longitude. Note that there were no nulls in the original dataset.

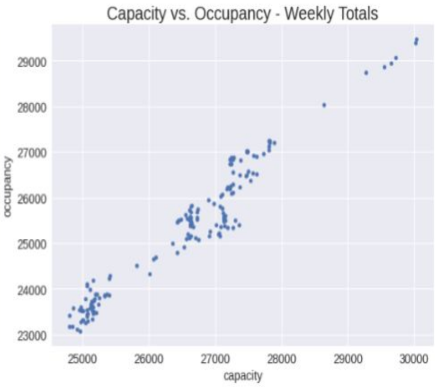
The data was prepared for the proximity Bayesian inference model by using the library Geopy. With Geopy, both the addresses of each shelter and the latitude and longitudes for each criminal occurrence could be converted to geocoded objects, in which the distance between each could then be obtained. A simple user defined function was written to append a column to the crime occurrence dataset for both: distance the crime occurred from the nearest shelter and that nearest shelter's address.

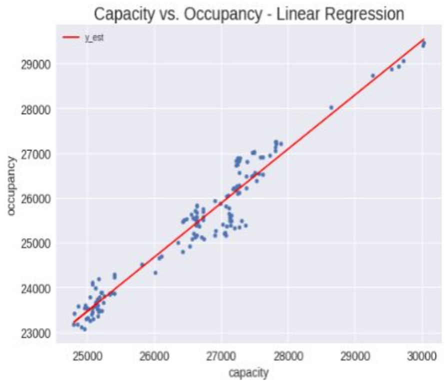
Weather
Data detailing daily temperature minimums and total precipitation from Jan 1, 2017 to Dec 31, 2019 was obtained from the Government of Canada: (https://climate.weather.gc.ca/historical_data/search_historic_data_e.html). The three CSV files (one per calendar year) were concatenated, and the columns that did not contain any data were removed as were a small number of days (rows) where minimum temperature or precipitation data was not collected. The weather data was indexed and merged with the shelter data by date.

Unit 3: Analysis: Correlation of Available and Occupied Beds

There are the two observations published by The City of Toronto for shelters, the number of shelter beds available ("capacity") and the number occupied ("occupancy"). Unfortunately, no rigorous measure is available for the number of people in need of shelters, that is, the number who could accept a bed if available. The various factors that may influence shelter use was of interest to this study group, as well as the question of whether the shelter system was adequately addressing the need for housing. In the plot of capacity vs. occupancy, seasonality is clearly visible within each week. Data was resampled to a weekly frequency to address this issue before regression. Partial weeks at the extreme ends of the range were dropped.

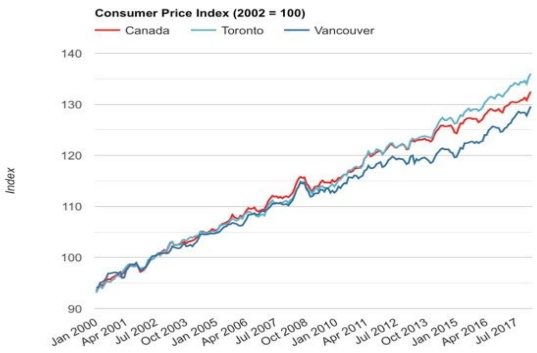






Ordinary Least Squares Regression (OLS) showed a strong linear relationship between capacity and occupancy with no homoscedasticity (nearly normal residuals) and no relationship between fitted or predictor values and residuals. Larger residuals showed lower leverage.

Unit 4: Analysis: Economic factors and homeless shelter occupancy



In this section, two analyses related to potential economic factors impact on shelter occupancy are considered: cost of living given in terms of the Consumer Price Index (CPI) and housing.

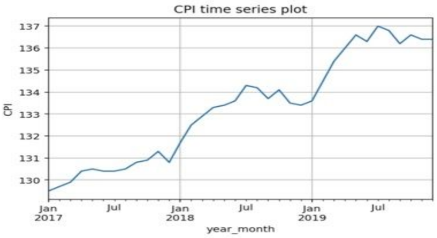
4.1 Consumer Price Index (CPI) analysis

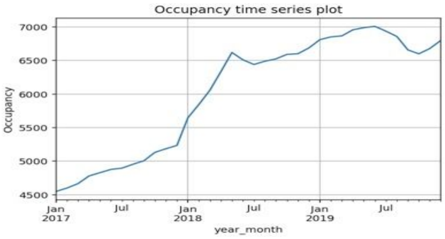
CPI is used to measure cost of living in Canada which tracks the change in costs of a basket of goods/services people use/consume, such as gasoline, food, vehicles, mortgage interest costs, etc. The index is proportional to the cost of living. When cost of living goes up, the index goes up. When cost of living goes down, the index goes down. (Ref. Article from Better Dwelling)

The CPI data used in this analysis is downloaded from the Bank of Canada (CPI, 2000 to Present, Bank of Canada). No complete CPI data for Toronto was found, however, as can be seen in the trends below Toronto CPI trends with Canada CPI and therefore without loss of generality we can use Canada's CPI in our analysis.

4.1.1 Exploratory Analysis of CPI and shelter occupancy data

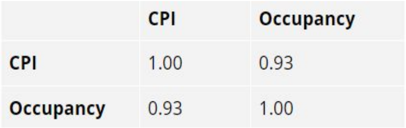
To explore the data, time series plots, scatter plots, and correlation analysis is performed. The time series plots for CPI and occupancy data are given on the left.



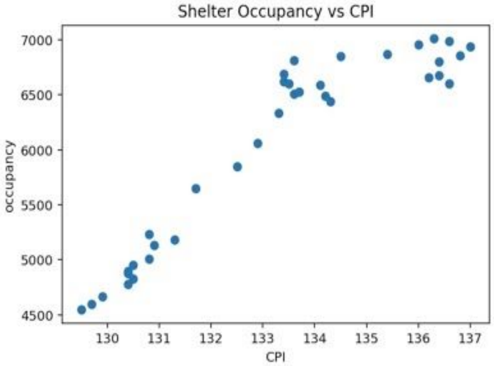


From the time series plots, both occupancy and CPI data trends together from Jan 2017 to Jan 2019 after which the occupancy data slope changed and the trend slowed down as CPI continued to trend up. This is due to the behavior observed in family shelters as explained in the shelter data preparation.

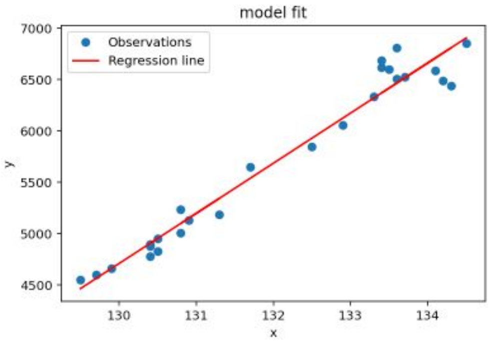
Next the correlation analysis for shelter occupancy versus CPI is provided.

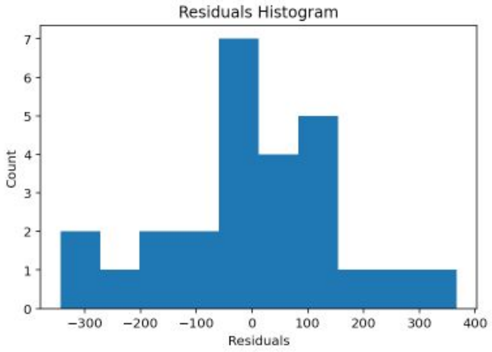


	CPI	Occupancy
CPI	1.00	0.93
Occupancy	0.93	1.00



From the correlation analysis we can see there is a strong positive correlation between CPI and occupancy. Similar to what we observed previously in the time series plot, the slope of linear relationship changes suddenly at values of CPI > 135 which is mainly related to the influence of family shelters on the data. If we try to model occupancy using CPI as a predictor and since both are time series data, we need to exclude sudden changes and build a model to predict the gradual change. The linear regression to model occupancy with CPI is considered next. The ordinary least squares (OLS) method from statsmodels is used on the data after removing the outlier portion and the model fit and residuals are shown below.

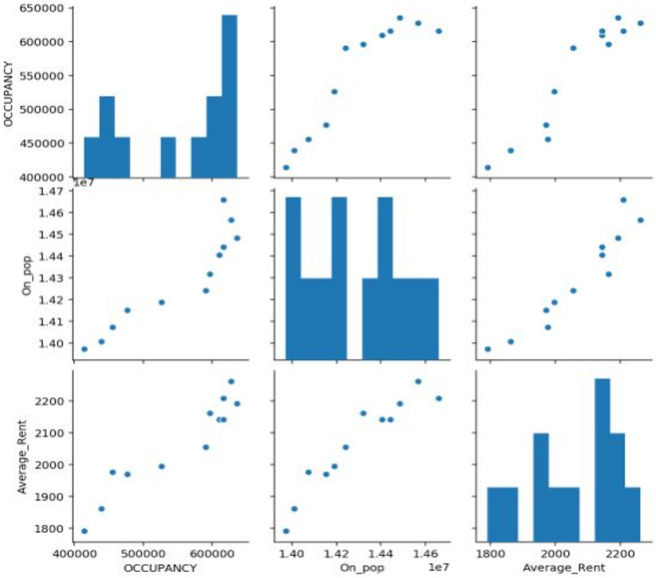


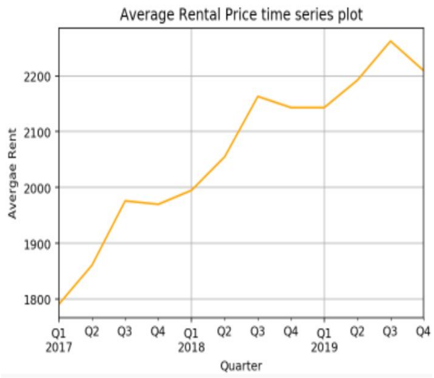


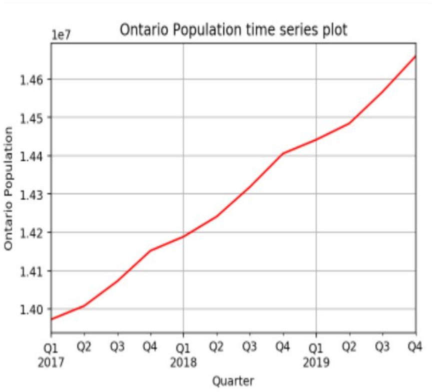
From the OLS regression results we can see the linear model fits the data well supported by both R-squared and Adj. R-squared of 0.962 and 0.961; respectively. The overall probability (F-stat) is very small and the model coefficients p-values are statistically significant ($P < 0.05$).

4.2 Rental Market and Population Growth Analysis

The data was first analyzed by developing several time series plots for average rental prices, the Ontario population, and shelter occupancy. The time series plots can be found below:







To help determine if there may be a linear correlation between each of the data sets, a scatter matrix plot for average rent, Ontario population and shelter occupancy was created to check correlations between the variables visually.

Next, an ordinary least squares (OLS) Regression Summary gave insights into the correlations between the shelter occupancy, and average rent and population growth. Although the value of R^2 was significant at 0.906, the p-value of Ontario population was very high (0.550), making it a statistically insignificant variable for our analysis as this was much greater than the threshold of 0.05.

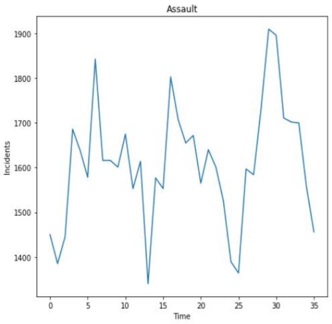
In order to improve the model, the population parameter was removed from the analysis and another OLS Regression Analysis was performed between just Average Rent and Shelter occupancy. This gave some promising results with a significant **R^2 value at 0.903** and very low p-values (<0.05).

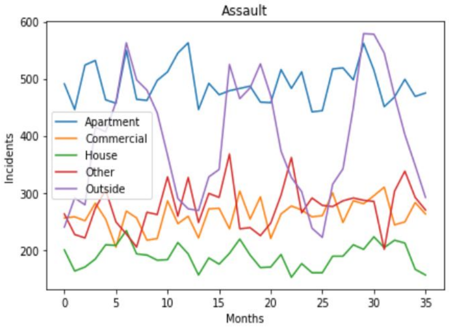
From the analysis we can reject the null hypothesis and say that Ontario population is not correlated with shelter occupancy. There is a strong correlation between average quarterly rent and shelter occupancy in Toronto with both showing an upward trend in the past 3 years.

Unit 5: Crime and homeless shelter occupancy

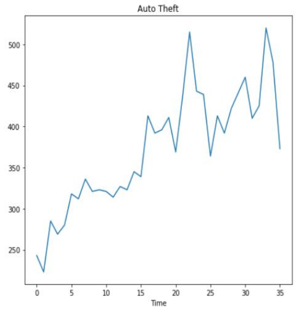
Analysis and model:

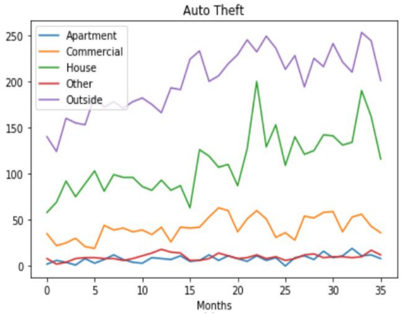
<u>Exploration of the Policing Data:</u>





After preprocessing the crime incident data, each major crime indicator (Assaults, Auto thefts, Break and Entering, Robbery and Theft over) was plotted as function of time and further categorized into each premise type, as





seen to the left for assaults and auto thefts (see Jupyter notebook for all MCI).

Total assault incidents seem to display seasonal variation i.e. higher occurrence in summer vs in winter. However, seasonality only seems to pertain to those assaults occurring outside. Total auto thefts seem to have a general growth over time for both outdoor and those occurring in a house.

While auto thefts seem to increase over time, in a similar manner to shelter use, it is challenging to infer if there is direct causation or a separate underlying cause for both trends.

Markov Chain Monte Carlo Model for Bayesian Inference on Likelihood of Crimes near Shelters

For the purposes of developing a proximity-based model the crime incident data was limited to occurrences between 2017 to 2019 and only the outdoor crimes (with the exception of breaking and entering).

To test the potential impact shelter locations may have on the surrounding environment, the distance from the location a crime occurred to the nearest shelter was examined to help evaluate whether crimes are more or less likely to occur near shelters, with a possible inference that those individuals in the shelter may be committing the crimes.

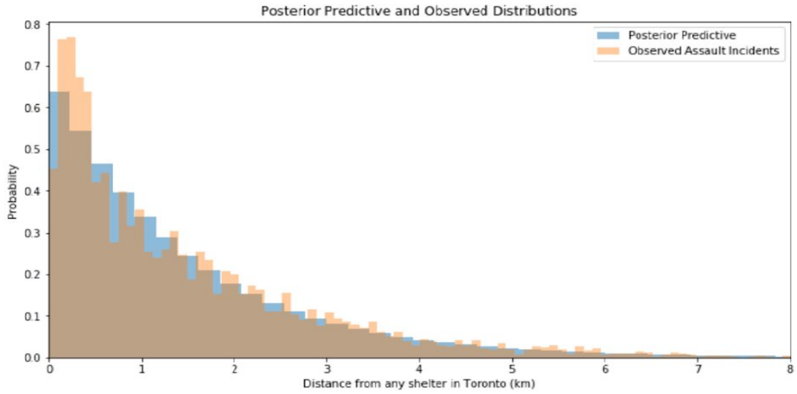
An analogous example to explain why an exponential distribution might be used to infer that a shelter location is a source of potential criminal activity is how heat dissipates from a flame as exponentially decaying as the distance from a shelter increases from the flame[1]. Likewise, if the likelihood of criminal incidents decays exponentially as the distance from a shelter increases it could be inferred that the shelter is a source of the potential criminal incidents.

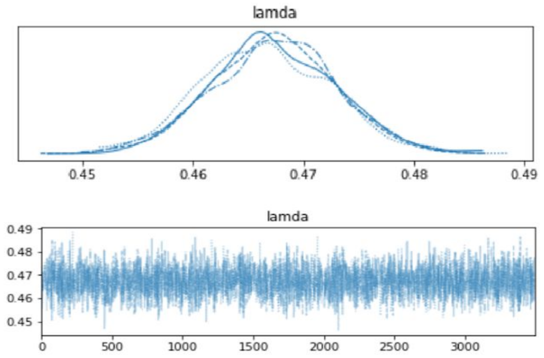
The distances of each criminal incident under each major crime indicator (MCI) category were plotted and used to fit an exponential distribution to the data using the Markov Chain Monte Carlo method. This method is used to estimate the value for lambda (λ) in the below exponential equation:

$$f(x \mid \lambda)=\lambda \exp\{-\lambda x\}$$

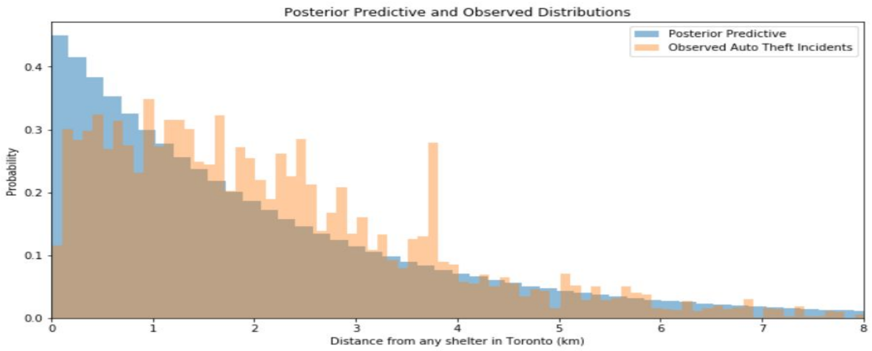
If the posterior exponential distribution fits the data then the highest density interval would be relatively small, the model would converge onto a constant value for lambda, and the distribution would visually match the data. The following posterior distribution, and Markov chain traceplots were obtained for assaults (please refer to Jupyter notebook for all other MCIs).

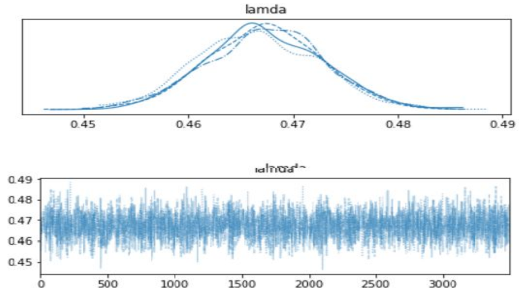
[1] Labovská, Zuzana & Labovský, Juraj. (2016). Estimation of thermal effects on receptor from pool fires. Acta Chimica Slovaca. 9. 10.1515/acs-2016-0029.





The above assault data traceplots also indicate that the model has converged as the bottom left graph seems to be varying randomly around a central constant. Note that the convergence could have been improved if more samples were used, however the limitations of the computer used meant that fewer samples were taken.





The above auto theft data traceplots also seem to indicate that the model has converged as the bottom left graph seems to be varying randomly around a central constant. However, the fit does not appear good as the assault distribution, as the predicted distribution seems to overpredict the observed data within 1km of the shelter. This 1km distance is the most important part of the model, for the purposes of drawing our conclusions on the connection between shelter occupants and criminal incidents, as distances beyond this 1km radius make the inference that the criminal incident was perpetrated by a shelter resident less likely.

Bayesian Inference Model Lambda Calculation Summary Table:

MCI	**Premise Type**	**Mean**	**Standard Deviation**	**HDI_3%**	**HDI_97%**	**r_hat**
Assault	Outdoor	0.691	0.006	0.681	0.702	1.0
Auto Theft	Outdoor	0.467	0.006	0.457	0.478	1.01
Robbery	Outdoor	0.566	0.007	0.553	0.581	1.0
Break and Enter	All[2]	0.545	0.004	0.538	0.552	1.0
Theft Over[3]	Outdoor	0.511	0.016	0.482	0.543	1.0

[2] All premises were included for breaking and entering as most events are logged based on the premise that was broken into - not outdoors.
[3] Theft over is defined as a theft of over $5,000 (excluding auto theft).

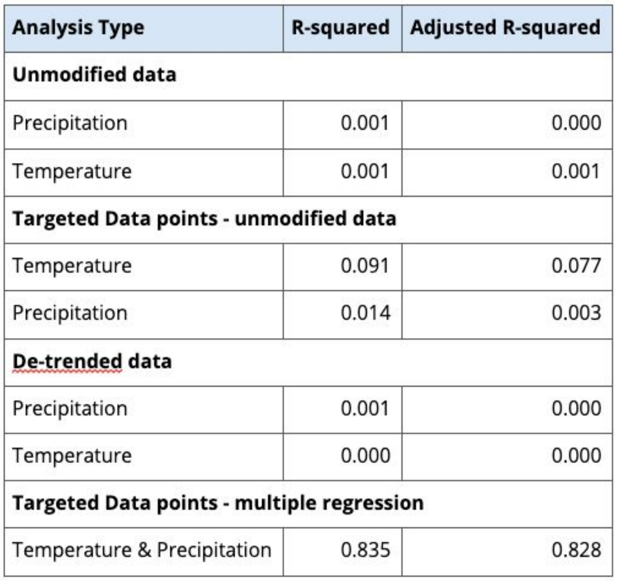
As seen in the above summary table, the exponential distribution seems to fit the data as the r_hat convergence diagnostics are all 1, with the exception of auto thefts. The Highest Density Interval (HDI) all seems fairly small for each row, indicating high confidence that the exponential distribution fits the data.

Given how well the exponential distribution fits the data, it is possible to infer that the likelihood of a criminal incident occurring decreases exponentially as the distance from the nearest shelter increases. However, without individual identification connecting shelter occupants directly with crime incidents, the conclusion may only be coincidental with other factors and it may not be shelter occupants committing these crimes.

Unit 6: Weather-induced patterns in the use homeless shelters

Analysis & model

OLS Regression Analysis was used to measure test hypotheses 1 & 2. Initial analysis with the unmodified data yielded R^2 & Adjusted R^2 values that were close to 0 (refer to table below).



Analysis Type	R-squared	Adjusted R-squared
Unmodified data		
Precipitation	0.001	0.000
Temperature	0.001	0.001
Targeted Data points - unmodified data		
Temperature	0.091	0.077
Precipitation	0.014	0.003
De-trended data		
Precipitation	0.001	0.000
Temperature	0.000	0.000
Targeted Data points - multiple regression		
Temperature & Precipitation	0.835	0.828

Next OLS Analysis was performed using targeted data points; days where the temperature was below -10 degrees Celsius and days where precipitation was greater than 10 mm (individually using OLS). This time, the R^2 & Adjusted R^2 values were slightly better, but not significant. The null hypothesis was not disproved.

The primary trend in the data is that occupancy is growing; the secondary trend is that there is seasonality in the data. Both are long term trends, and not helpful to analyzing daily spikes in data.

The occupancy data was de-trended using multiplicative and additive methods. There was no discernable difference in the residuals or autocorrelation. The data from the multiplicative de-trending was selected.

Using de-trended data, the OLS analysis was run again on temperature and precipitation individually. Again, the R^2 & Adjusted R^2 values were insignificant.

In an article about the influence of weather on shelter occupancy in Calgary, Alberta it was not extreme cold or precipitation alone that caused spikes in occupancy, but rather temperatures close to 0 or just below and precipitation - sleet or freezing rain - drives spikes in demand[4]. Applying this same approach and using Multiple Regression on Temperature and Precipitation the R^2 & Adjusted R^2 values were significant at 0.837 & 0.830 respectively. This suggests that the combination of temperatures near 0 degrees Celsius with significant precipitation (> 2mm) correlate to increased homeless shelter occupancy.

This correlation could be used by shelter system planners to prepare days ahead for potential spikes in demand using weather forecasts.

A future study could delve deeper into the nature of the correlation and looking at the impact of the variables using either backward elimination or forward selection.

[4] Jadidzadeh, Ali & Kneebone, Ron, "SHELTER FROM THE STORM: WEATHER-INDUCED PATTERNS IN THE USE OF EMERGENCY SHELTERS", The School of Public Policy Research Papers Volume 8, Issue 6 University of Calgary, February 2015.

Unit 7: Conclusions

We return to the objectives we started with and provide the results found:

Correlation of Available and Occupied Beds

Null Hypothesis: There is no statistically significant correlation between capacity and occupancy.

Alternative Hypothesis: A statistically significant correlation exists between capacity and occupancy.

The Null Hypothesis is rejected, as a statistically significant correlation was found between capacity and occupancy. Since capacity is controlled by shelter administration, further study is warranted into quantifying the underlying need of homeless individuals as well as the number who are denied service each night.

CPI (Consumer Price Index), Rental Costs, Population and Shelter Occupancy

Null Hypothesis: Neither CPI nor rental costs are correlated with increased shelter occupancy.

Hypothesis 1: CPI is correlated with increased shelter occupancy

Hypothesis 2: Rental costs are correlated with increased shelter occupancy

The Null Hypothesis is rejected as the linear regression analysis revealed statistically significant positive correlation between CPI & rental costs on one hand and shelter occupancy on the other. There is a strong correlation between average quarterly rent and shelter occupancy in Toronto with both showing an upward trend in the past 3 years.

Crime & shelter occupancy

Based on a series of Bayesian inference models, it appears that for all outdoor crimes, with the possible exception of auto thefts, the likelihood of a crime's occurrence increases exponentially with the decreasing distance to the nearest shelter. This model outcome may play a useful role in city planning decisions in establishing new shelter locations and in policing patrol route decision making.

Weather & shelter occupancy

Null Hypothesis: Weather does not significantly impact demand for homeless shelters

Hypothesis 1: Negative temperatures correlate with higher demand for homeless shelters

Hypothesis 2: Days with precipitation correlate with higher demand for homeless shelters

The Null Hypothesis is rejected, however not by Hypotheses 1 or 2 but by a third hypothesis which was established during the analysis. That hypothesis, that a specific combination of temperature and precipitation is correlated with increased shelter occupancy was proven using multiple regression.

Shelter occupancy is influenced by long-term factors such as the cost of living (CPI), the cost of rental units and the presence of crime in the community near the shelter. These factors have ramifications for city government and non-profit organizations who must make decisions on policy and funding based on information such as this. Shelter occupancy is also influenced on a short-term basis by the weather and the people who manage the day-to-day operations of homeless shelters must use weather forecasts to ensure homeless shelters can accommodate sudden spikes in demand because of specific weather conditions.