Data Science 2: Statistics for Data Science

Group 2 Members: Walid Al Gherwi, Bryan Mallinson, Steven McAvoy, Vernon Naidoo, Mahshad Najafi Ragheb, Mohit Ramnani

Date: August 10 2020

Homeless Shelter Occupancy Analysis - Discussion Summary

Data Sources

The primary data source is the City of Toronto's data on homeless shelter occupancy (https://open.toronto.ca/dataset/daily-shelter-occupancy/). The data spans the years 2017-2019. The data frequency was daily, broken out by shelter location and classified by type (men's, women's, family etc.). 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 the cost of living or the rising cost of housing. The relationship of crime to the location of homeless shelters is another area that is explored.

To facilitate our analysis, we researched and located the following additional data sets:

- Consumer Price Index (CPI) data downloaded from the Bank of Canada (CPI, 2000 to Present, Bank of Canada)
- Data for average rent in Toronto from the Rental Market Report Archive
- Ontario population from Government of Canada https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000901
- Toronto Police Service: Major Crime Indicators (MCI) https://data.torontopolice.on.ca/datasets/mci-2014-to-2019
- Daily weather data for the city of Toronto from Government of Canada data portal: (https://climate.weather.gc.ca/historical_data/search_historic_data_e.html)

Analysis

Ordinary Least Squares (OLS): This approach was used for CPI, rental and Weather analysis. We examined the requirements for applying the model then assessed the validity by examining the regression results like R^2, Adjusted R^2, the confidence interval, Kurtosis value, model coefficient p-value, and visual checks such as residual plots.

Bayesian inference model (Markov Chain Monte Carlo): Markov Chain Monte Carlo method was used to develop a bayesian inference model for the likelihood of a criminal incident occurring as a function of the distance from a shelter. The chosen distribution to model the likelihood was an exponential distribution and the fit of the posterior distribution was assessed by examining the rhat convergence diagnostic, size of highest density interval, and visual inspection of the model as compared observed data.

Multiple Regression: This approach was used for the weather analysis given that the variables used (temperature and precipitation) are not highly correlated and that we had a suitable sample size. We evaluated the R^2 and Adjusted R^2, and the Durbin-Watson test for homoscedasticity.

Challenges

The team faced a number of challenges, particularly in the beginning with the data.
Our main data set on shelter occupancy is high quality with few missing values, but there were trends in the data that interfered with the analysis. For example, the data related to family shelters was erratic and noisy, especially in the first year. To overcome this, we interviewed someone who works in the shelter system who explained how family shelter capacity and occupancy are calculated, leading us to drop the family classification from our data set. For the weather analysis the data was further detrended (using the multiplicative method) to remove long term growth and seasonal trends.
CPI data was not publicly available for Toronto (but summary trend data was), so the team compared the national trend with the Toronto trend and found that the national data was a reliable proxy.
The city of Toronto only publishes population data annually, so the team used Ontario's quarterly population data as a proxy.

Average rent data was not readily available on toronto public data sites, however, we did manage to find some quarterly reports on Toronto Regional Real Estate Board's report archive. These reports were in a pdf

format with one pdf file per quarter. In order to use the data in the analysis, the numbers for average rent for each quarter were picked and typed into a csv file manually.

The crime occurrence data available was extensive and had to be mapped to the closest shelter. The data was prepared using the Geopy library . With Geopy, both the addresses of each shelter; and the latitude and longitudes for each criminal occurrence could be converted to geocoded objects and a user defined function was used to get the distance between each of these objects. The function was then made to determine the minimum distance from every criminal occurrence to the nearest shelter.

Challenges in Analysis revolved around both finding the right testing method and the right criteria to test. For the weather analysis, we started out testing individual variables like temperature and precipitation, and not seeing any correlation. It was only when reading a study about homelessness in Calgary that we got the idea to test the combination of temperature and weather under specific conditions using Multiple Regression, which did find a correlation.

Conclusions

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 & 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.

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.

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.