# Introduction

Toronto is the largest city in Canada with apartment rental rates among the highest in the country. The rental market can be competitive and fast moving, requiring rental decisions to be made quickly. There may not always be time for prospective renters to explore areas in which they are considering renting. Furthermore, it may be difficult to evaluate value of a rental given the size and diversity of the market.

This project aims to develop a tool to estimate rental cost based on information about the apartment itself as well as the amenities in the surrounding area. Such a tool could be used by prospective renters to compare the expected price to the asking price and aid in making rental decisions.

# Data

The rental data used in this project come from the data set [here](https://www.kaggle.com/rajacsp/toronto-apartment-price). It is a 2018 data set that includes:

* number of bedrooms,
* number of bathrooms,
* existence of a den,
* address,
* latitude,
* longitude, and
* price

of 1124 rental apartments. In addition to the data about the apartments themselves, Foursquare location data is used to evaluate the desirability (and therefore rental price) of an area based on the prevalence of restaurants in that area. For each area, the top 100 venues along with the venue category and coordinates were pulled from Foursquare. This data set included 194 unique venue categories (e.g. park, sandwich place, tech startup, etc.). After review of the categories, restaurants were considered to be any venue category containing the keywords:

* Restaurant
* Joint (e.g. ‘BBQ Joint’)
* Place (e.g. ‘Taco Place’)
* Coffee Shop
* Spot (e.g. ‘Breakfast Spot’)

Restaurants were used as a proxy for desirability assuming that areas with a larger number of restaurants in their top 100 venues would be more desirable.

# Methodology

## Data Exploration and Cleaning

As a first step, the rental data was cleaned up by removing duplicates and outliers. The data was then mapped to show the geographic spread, as shown in Figure 1 below.

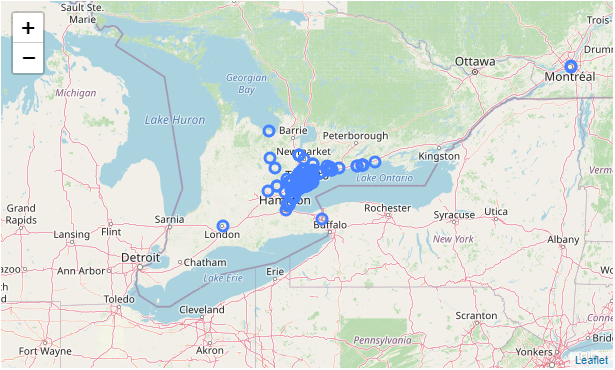


Figure 1 – Geographic locations of raw data set

It can be seen in Figure 1, that some of the entries in the data set were not in Toronto. Instead, some were as far away as Montreal. The data were further filtered to include only listings where the postal code was in a borough including the word ‘Toronto’ based on this [Wikipedia page](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M).

The resulting data set included 559 entries within the city of Toronto shown in Figure 2.

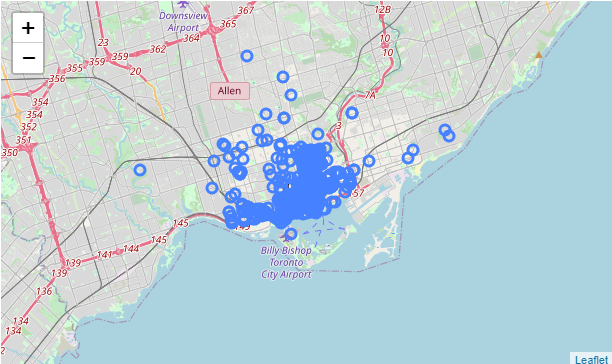


Figure 2 – Geographic locations of data set in Toronto only

## Neighborhood Clustering

It would be possible to collect Foursquare location data for each individual entry. However, with 559 entries, it is more efficient to group the listings based on location and retrieve location data for the group. Using only the Wikipedia data cited above, it would be possible to cluster entries based on postal codes, boroughs, or neighborhoods. However, there are only 4 boroughs and there are 39 postal codes (even more neighborhoods). It would be preferable to have a number of groups between 4 and 39. Therefore, kmeans was used to create 10 distinct clusters based on location. Figure 3 shows the resulting clusters.

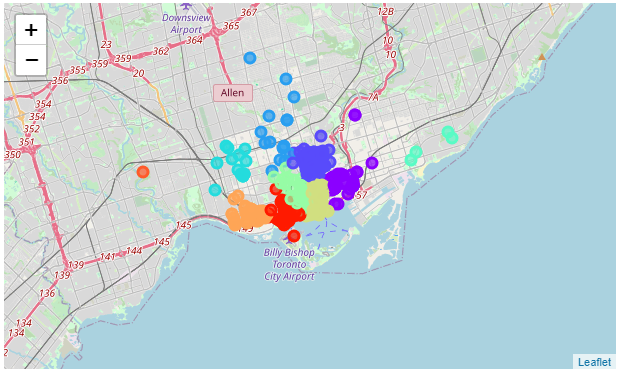


Figure 3 – 10 clusters of Toronto rental data

## Foursquare Location Data

The centroids of the 10 clusters were used to retrieve the top 100 venues within a 2 km radius of the centroid. This returned 1000 venues of 194 unique categories. Using the keywords outlined in the Data section, the number of restaurants in the top 100 venues was counted for each cluster as shown in Table 1.

Table 1 – Number of restaurants among top 100 venues per cluster

|  |  |
| --- | --- |
| **Cluster** | **Number of Restaurants (in top 100 venues)** |
| 0 | 30 |
| 1 | 39 |
| 2 | 31 |
| 3 | 35 |
| 4 | 43 |
| 5 | 36 |
| 6 | 25 |
| 7 | 28 |
| 8 | 42 |
| 9 | 40 |
| Total | 349 |

Table 1 also shows that restaurants account for 349 of the top 1000 venues.

## Distance to Centre

It was hypothesized that, in addition to the features of the rental apartment, and the neighborhood, the price may also depend on the location relative to the centre of Toronto. Therefore, the distance between each rental and the centre was calculated using geopy geodesic distance.

## Model Development

Based on the aforementioned data collection and analysis, the following model inputs were available for 559 unique entries:

* Number of bedrooms
* Existence of a den
* Number of bathrooms
* Number of restaurants within to 100 nearby venues
* Distance to city centre

and the target variable was price.

The data were split into a training set and a testing set and were fit using a linear regression model, with increasing orders of polynomials to determine the best fit.

# Results

Figure 4 below shows the R2 values for the models of increasing order.

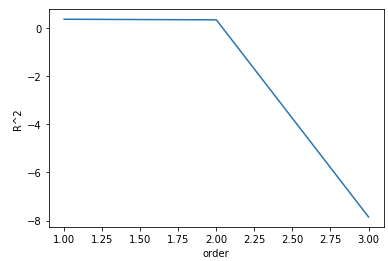


Figure 4 – R2 values for testing data with increasing polynomial order

It can be seen that the best fit was with a first order polynomial with an R2 value of 0.364. Therefore, the resulting model is:

Using the above model, the distribution of the predicted values in the test set can be compared to the actual values as shown in Figure 5.

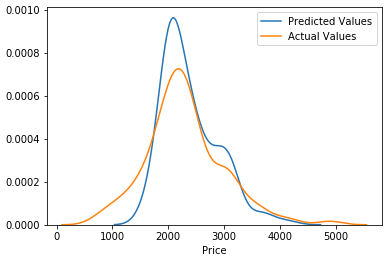


Figure 5 – Predicted versus actual testing data distribution for linear model

# Discussion

An R2 value of 0.364 indicates that the model does not account for a significant portion of the variability in rental prices. Therefore, the predictions of rental price based on the model inputs alone will not be very accurate.

The distribution shown in Figure 5 shows that the model predicts a smaller range of rental prices than is seen in the dataset. The model fit may be improved if additional rental data were available.

The original hypothesis that an area with more restaurants in the top 100 venues would be a more desirable place to live (and therefore more expensive) was not supported by the data. The negative coefficient associated with number of restaurants indicates that the price goes down as the number of restaurants goes up. The hypothesis that the further from the centre, the lower the price was supported as can be seen by the negative coefficient.

# Conclusion

The rental price of an apartment in Toronto was not well predicted by:

* number of bedrooms
* existence of a den
* number of bathrooms
* number of restaurants within to 100 nearby venues, and
* distance to city centre

based on the available data. The expectation that more restaurants in the area would increase the rental price was not found to be true. Therefore, it is expected that there may be better ways to leverage the Foursquare data to evaluate neighborhoods, their amenities and their impact on price.