Toronto Rental Analysis

Analyzing Rental Listings from Craigslist
Author: Michael Parravani

Overview

Craigslist rental data was extracted from March 19 to April 19th, and analyzed using various methods in an attempt to attract insights. The main findings were that most listings within the city are 1 or 2 bedroom units, 1 bedroom units have the highest cost/bedroom, and the cost of unit is correlated to the population density of the area. From a geographic analysis of the data, most of the listings were found to be in the downtown core, with the highest price per bedroom generally in that area as well.

Limitations of Analysis

The data was extracted over a 1 month period of time, and from only one source – so it is a quite limited sample. Any findings from this analysis are stated with the understanding that the scope of the data is limited. Also, the pricing used is the listed asking price, not the actual agreed to price. These can differ widely depending on the area, time of year, etc. Lastly, there were many listings that had multiple units highlighted in each post (for ex. multiple listings for 1 building made in the same post). This muddies the analysis as it is difficult to find and extract these listings.

Methodology

To allow for brevity of this final report, please see the readme file for an overview of methodology.

Results

The below are a selection of the results of the analysis. For full results, please see the Craigslist Rental EDA.ipynb and Craigslist Detailed Analysis.ipynb notebooks.

Listing Analysis

The availability of rental units were highest among 1 and 2 bedroom apartments, with 3 bedrooms and studio (shown as 0 below) following by a substantial margin.

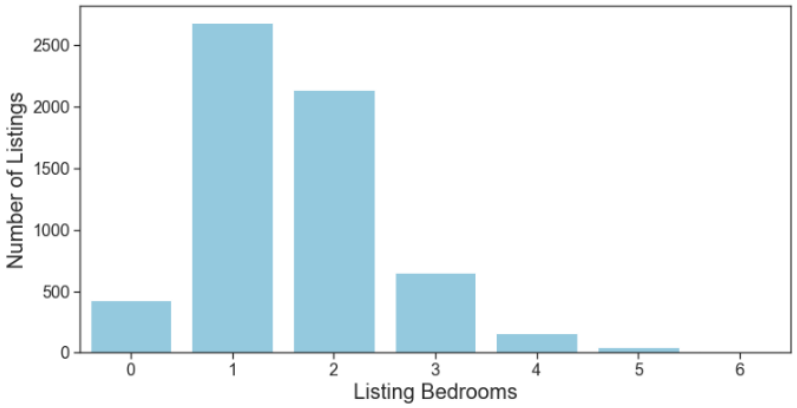
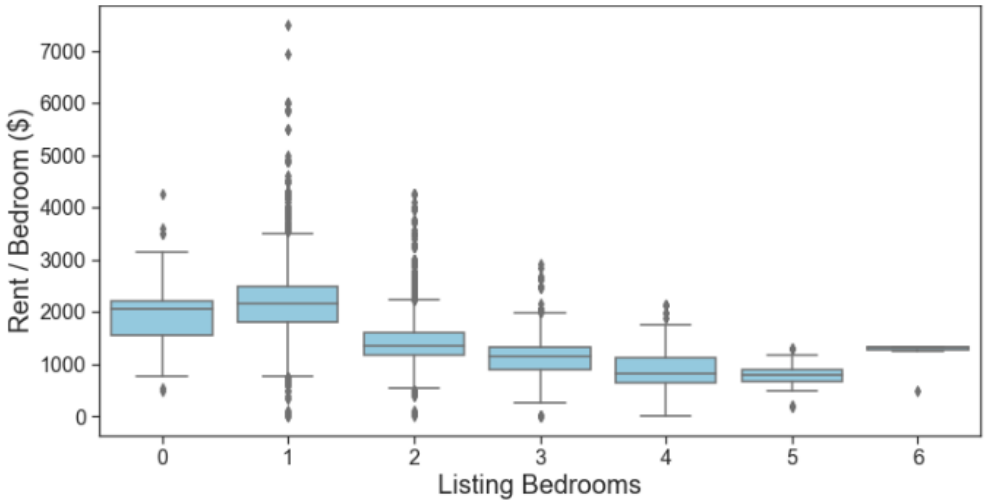


Figure 1: Count of listings by Bedroom Type (0 represents studio apartment)

Looking at rental costs per bedroom, the 1 and studio apartments were a substantial premium to 2+ bedroom units. This is the premium people pay to "live alone". What's interesting as well, is that the median price for studio and 1 bedroom apartments are quire similar, though the distribution of 1 bedroom units sits slightly higher.



The median size for all listed units (that noted an area) was 750sqft. Given that a high proportion of the units were 1 or 2 bedroom, this result is expected. It can be seen in the below distribution that though 750sqft is the median, there are a substantial number of units ~600sqft as well. This peak range of the two bins likely represents both 1 and 2 bedroom listings.

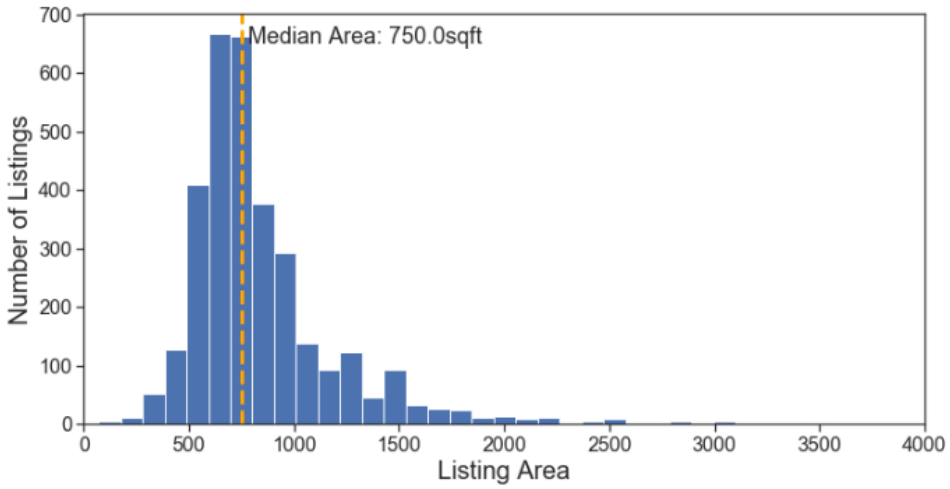


Figure 2: Count of units by listing area (sqft)

Census data was imported from SimplyAnalytics, and a population density was calculated per census tract area (density being average # of people per household). There was a noticeable negative correlation between the price per bedroom and population density of a particular area. This indicates that people prefer to live in areas that have a more sparse population (less people per housing unit).

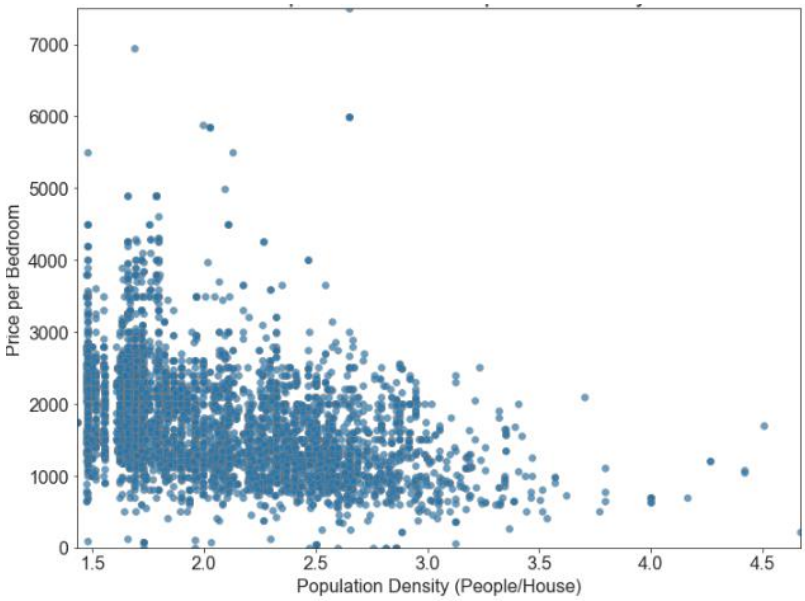


Figure 3: Rental costs vs Population Density, based on census tract area

Geographic Analysis

Most listings were in the downtown core, as well as Yonge/Eglington area. This is shown in the two plots below (a scatter plot and hexbin plot of listings by location).

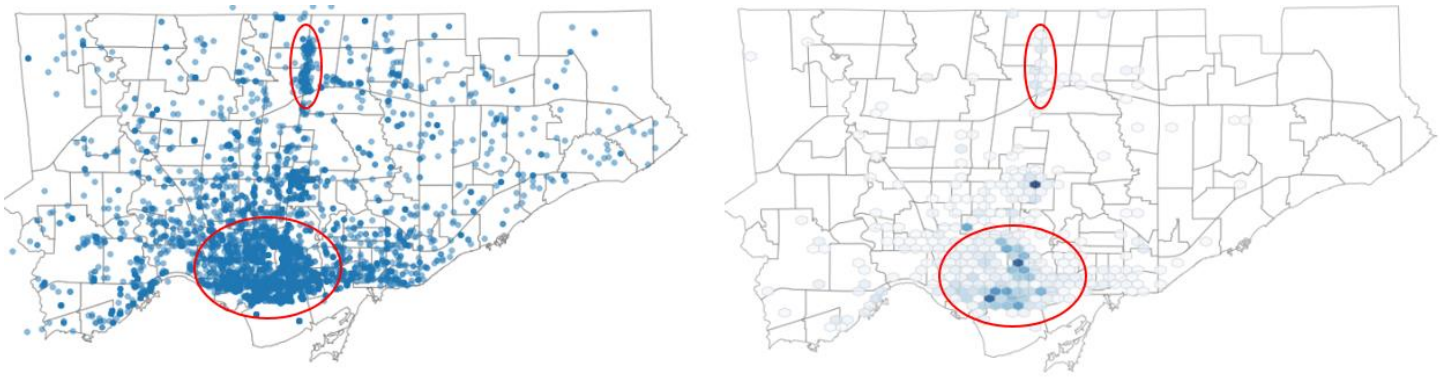


Figure 4: Scatter and Hexbin plots of listings by geographic location

It can be seen from the below plot in the areas of high listing density (notably downtown proper), that the listings are generally lower bedroom units. Note, the below chart includes studio apartments as 1 bedroom units.

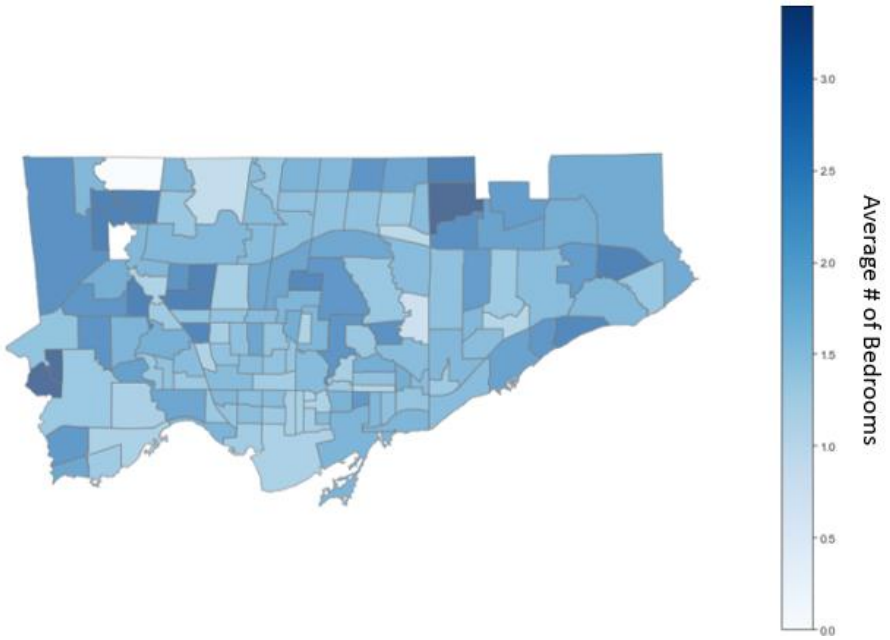


Figure 5: Average number of bedrooms by neighborhood

Natural Language Processing

Natural Language Processing of the body and title of each post produced a resultant tokenized listing and title. These were analyzed for their sentiment to determine if there was an impact on listing cost ($/bedroom). As seen below, there is little to no correlation. Notably a few average priced listings that were described with an extremely high sentiment while the highest cost units had an average sentiment score.

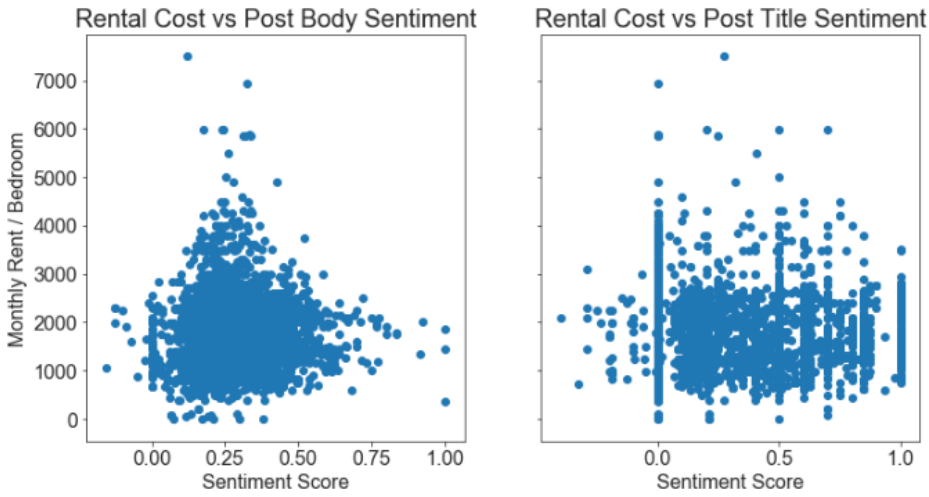


Figure 6: Sentiment score vs rent / bedroom for post body and title

The listings were broken into 4 quartiles based on their cost per bedroom. The text from the body of the posts were then analyzed to see if there were any differences in the items mentioned for higher cost posts. From the below, the higher cost listings (quarter 2+) have words like *view, ceiling and open* – likely describing the spacious feel of the units. The 3-4th quartiles have a higher emphasis on 'appliance', which is likely to emphasize quality appliances in the higher cost units. Lastly, quartiles 2-4 have prominence in washer and dryer,

while the 1st quartile has high prominence to 'laundry' - which could be ‘in building laundry' vs 'in suite washer/drier'.

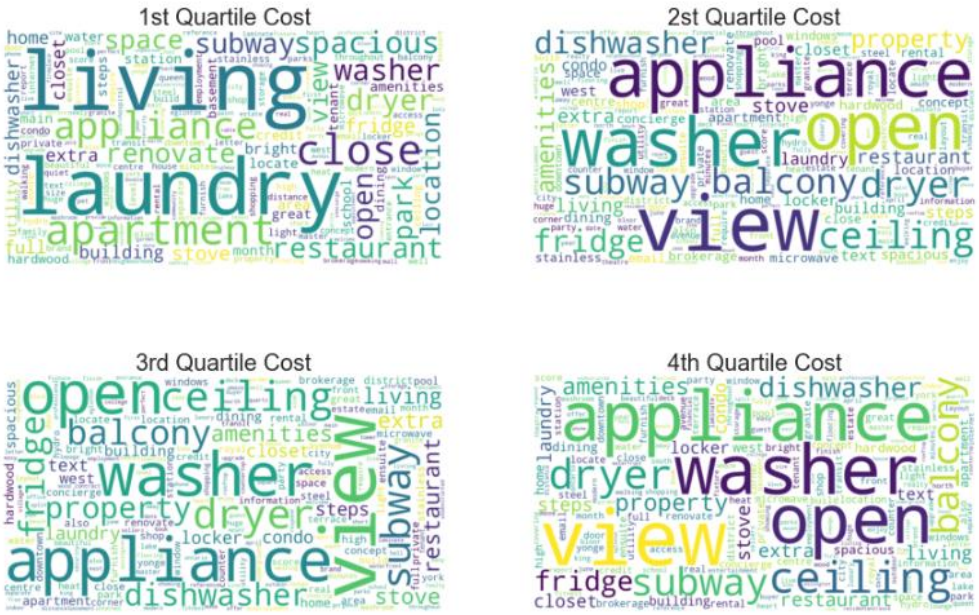


Figure 7: Word clouds for listing body, separated by cost/bedroom quartile

Model Results

Three models were built to predict price; linear regression, decision tree regression and random forest regression. The random forest regression had the highest r^2 with a value of 0.68. The input parameters were: area, population density, area median income, neighborhood ID, bedroom count, basement and studio apartment indicators (Boolean). The accuracy of the model can be improved with tuning parameters, and potentially trying further models, however the feature importance notably had the listing area (sqft) and area population density with the highest impact on resulting rental unit cost, while basement and studio features seemed to have little impact. This indicates that basement and studio rentals are priced the same as above ground and 1 bedrooms if their square footage and locations are comparable.

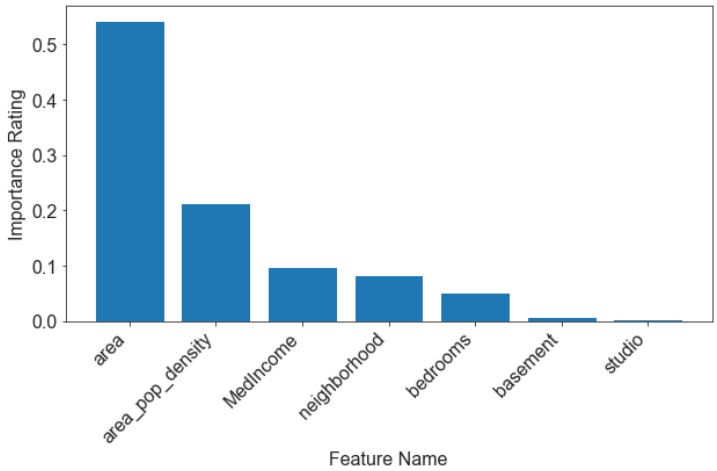


Figure 8: Random Forest Model - feature importance