A Closer Look at Property Prices in Toronto

Author: Barzin Doroodgar
Date: June 7, 2020

Table of Contents

Introduction

Toronto has one of the most expensive real estate markets in North America [1]. Regardless of the high prices, it remains as one of the most desired cities to live in. In this report, we will take a closer look at the property prices in Toronto and try to understand what is it about a given neighbourhood that impacts the property prices. More specifically, the problem we will be addressing here is the relationship between the venues in each neighbourhood and how they might affect the property prices across Toronto.

The audience of this report is anyone who might be interested in investing in the Toronto real estate market, or any professional who would like advice their clients on where a new investment might make sense.

Data

The data we will be examining comes from the House Sales in Ontario from kaggle [2]: https://www.kaggle.com/mnabaee/ontarioproperties/data. This dataset lists the property prices in Ontario along with address, area name, and latitude and longitude data on each property. Furthermore, we will be using the the Foursquare API to get some useful location data about each property. The combination of these two dataset will allow us to gain some useful insight about the relationship between property prices and some of their attributes. Another price of data that was used as a feature was the total count of unique venues in each neighbourhood. This can be achieved by grouping the data location by neighbourhood and getting a total count.

The images below show examples of the property prices dataset from kaggle and the location data from Foursquare:

	Unnamed: 0	Address	AreaName	Price ($)	lat	lng
0	0	86 Waterford Dr Toronto, ON	Richview	999888	43.679882	-79.544266
1	1	#80 - 100 BEDDOE DR Hamilton, ON	Chedoke Park B	399900	43.250000	-79.904396
2	2	213 Bowman Street Hamilton, ON	Ainslie Wood East	479000	43.251690	-79.919357
3	3	102 NEIL Avenue Hamilton, ON	Greenford	285900	43.227161	-79.767403
4	6	#1409 - 230 King St Toronto, ON	Downtown	362000	43.651478	-79.368118

Figure 1: Raw property price data from kaggle

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.654260	-79.360636	Morning Glory Cafe	43.653947	-79.361149	Breakfast Spot
1	Regent Park, Harbourfront	43.654260	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa
2	Garden District, Ryerson	43.657162	-79.378937	Ryerson Image Centre	43.657523	-79.379460	Art Gallery
3	Garden District, Ryerson	43.657162	-79.378937	Balzac's Coffee	43.657854	-79.379200	Coffee Shop
4	St. James Town	43.651494	-79.375418	Gyu-Kaku Japanese BBQ	43.651422	-79.375047	Japanese Restaurant

Figure 2: Sample location data from Foursquare

The data that we need from the property price dataset are the AreaName, Price, and latitude and longitude columns. Feeding the latitude and longitude values into Foursquare API, we can obtain data on the venues in each of the property's surrounding neighborhood. Finally combining the dataset together can start analyzing the data.

First the property price dataset was cleaned. The dataset contains property prices for the entire province of Ontario. Therefore, the first step was the use the address column to filter the dataset with address belonging to the city of Toronto. Then the dataset was checked for any NaN values (which there was none). On further analysis of the dataset, it was found that some property values were very low (one property was as low as $25). This is either an error in the dataset, or we may have item that don't really qualify as a true property (houses, commerical spaces, etc.) Thus, the dataset was filtered on price to only hold properties with a price of over $50,000. The location data from Foursquare did not need much cleaning.

Methodology

First, some elementary data analysis was performed on the data to just explore the dataset and make sure everything made sense. Our property price dataset contained a total of 4906 properties (after the mentioned filters were applied). These properties were in 212 different neighbourhoods. As a first step, the property prices were averaged over each neighbourhood (AreaName). A histogram of the mean property prices was plotted to see the mean price distribution.

We can see from the histogram that most of the prices are under $600,000. The majority of the prices are under $200,000. This makes sense if we think about the number of small apartment and condo units that exist in Toronto, compared to large detached houses that can be priced for over $500,000. We also have a few expensive properties around $1.4 million dollars. These can be luxury houses in the Midtown Toronto area.

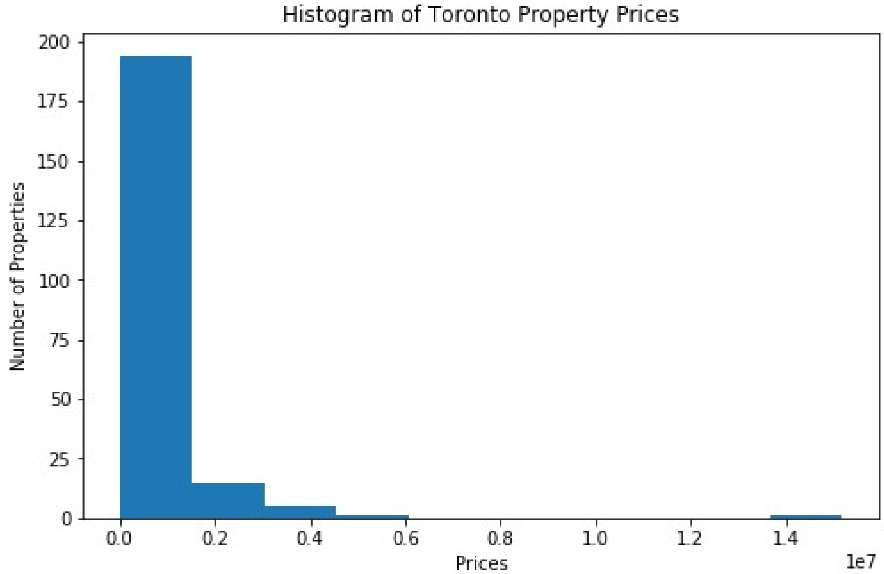

Figure 3: Histogram of property prices in Toronto

To dig deeper into the average property prices in each neighbourhood, we used the Folium library to generate a map of Toronto. Blue circled were added to the map to indicate the neighbourhoods in Toronto and a label was added to indicated the neighbourhood name and average price for that neighbourhood.

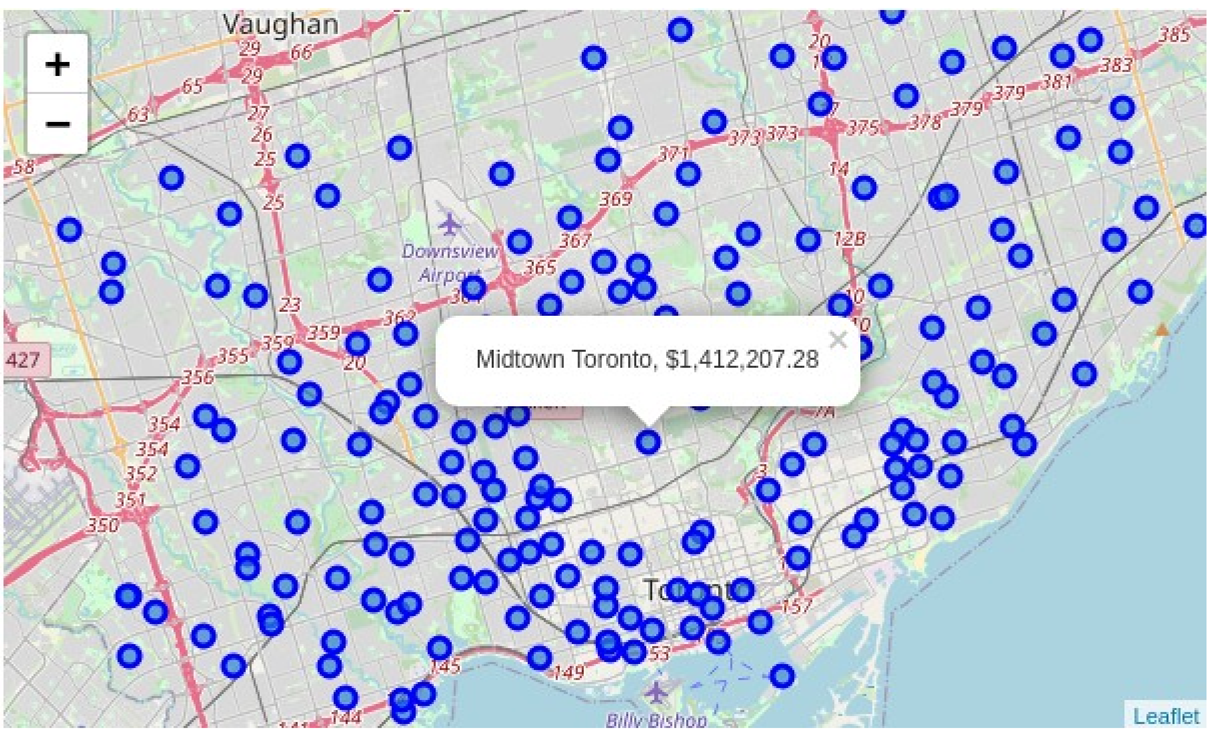

Figure 4: Map of Toronto with average prices in each neighbourhood

As expected, the Midtown Toronto area was the outlier in the dataset with an average price of $1.4 million.

Next, the location data was obtained for each neighbourhood and examined. First thing that was noticed was that Foursquare API only returned 84 rows of data, and not the 212 rows for each neighbourhood. This may be due to the close proximity of some of the neighbourhoods. That is if the latitude and longitude of the neighbourhoods are very close, Foursquare may not have unique location data for those neighbourhoods.

The nest step was to get a total count of venues for each neighbourhood. This was done by simply grouping the venues by their neighbourhoods, and getting the total count for each.

The column that will be used in data analysis is the last column, which tells us how many unique categories of venues exists in each neighbourhood.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Alderwood	1	1	1	1	1	1
1	Amesbury	2	2	2	2	2	2
2	Bay Cloverhill	15	15	15	15	15	15
3	Bayview Woods - Steeles	1	1	1	1	1	1
4	Belgravia	2	2	2	2	2	2

Figure 5: Venue category count for each neighbourhood

Next, the venue names were one-hot-encoded so that we can easily used them in data analysis later on.

	Neighborhood	ATM	Accessories Store	Adult Boutique	American Restaurant	Arts & Crafts Store	Asian Restaurant	BBQ Joint	Bagel Shop	Bak
0	Richview	0	0	0	0	0	0	0	0	
1	Downtown	0	0	0	0	0	0	0	0	
2	Old East York	0	0	0	0	0	1	0	0	
3	Dorset Park	0	0	0	0	0	0	0	0	
4	Morningside	0	0	0	0	0	0	0	0	

Figure 6: One hot encoded venues for each neighbourhood

Finally, all the datasets were merged to include the neighbourhood, average price, venue count, and all one hot encoded venue features in a single dataframe. The final dataframe was created as below:

	Neighborhood	Price	Latitude	Longitude	Venue Count	ATM	Accessories Store	Adult Boutique	Ame Resta
0	Alderwood	9.931799e+05	43.603214	-79.545025	1	0	0	0	
1	Amesbury	7.945000e+04	43.704548	-79.482700	2	0	0	0	
2	Bay Cloverhill	2.490000e+05	43.665531	-79.385147	15	0	0	0	
3	Bayview Woods - Steeles	1.290222e+06	43.792517	-79.390080	1	0	0	0	
4	Belgravia	1.365250e+06	43.697310	-79.445359	2	0	0	0	

Figure 7: Final merged dataframe.

Data Analysis

The purpose of our data analysis is to find relationships between a property's price and the different features. These features as stated earlier include the different kinds of venues that may exists around the property as well as the total count of venues in the property's vicinity.

The data was analyzed in two steps. The first step was to get a correlation matrix on the data. This is a quick and easy way to find which features of the dataset are most correlated to the priced. The second step was to used a lasso logistic regression to find which features have to most impact on the price of the properties.

Pearson Correlation

The Pearson correlation can be used to find the correlation of each feature with any other feature in the dataset [3]. The thing we're more interested in, however, is the correlation of the price with the other features. Correlation can be negative or positive and has a value between -1 and 1. A positive correlation means that the value of the price increasing feature (or existence of the feature in case of one-hot-encoded features). A negative correlation, on the other hand, means that the price decreases with the increase or existence of that feature. And a correlation near zero means that there is little to no correlation.

Logistic Regression with Lasso Regularization

Logistic regression is a machine learning algorithm that has embedded feature selection properties. That is the algorithm naturally maps the target value with a weighted combination of features [4]. The higher the weight of the feature, the more impact it has on the target value. Lasso Regression, is a regularization method that forces a lot of weights to be zero. Therefore, it naturally eliminates the weights that have insignificant contributions to predicting the the target value. To use the logistic regression method to select the features that have the most impact, we first binned the property prices to 5 categories: “very low”, “low”, “medium”, “high” and “very high”. This allows to simplify the target value. That is, instead of looking at a continuous value, the algorithm looks at price ranges.

Results and Discussion

Pearson Correlation

The image below shows the top positively corrected features to the property price. We see some nice venues like ice cream shop, playground, cafe, boutique, restaurants, nail salon, yoga studio, and cheese shop. These venues are all strong indicators of wealthy neighbourhoods. A property surrounded by these hip venues is usually one that is situated in a nice neighbourhood.

Ice Cream Shop	0.555208
Playground	0.325820
Gas Station	0.304605
Bubble Tea Shop	0.164910
Café	0.161757
Boutique	0.156044
Speakeasy	0.120337
Men's Store	0.060565
Restaurant	0.054502
Nail Salon	0.045300
Bar	0.040667
Sandwich Place	0.032995
Bank	0.028493
Yoga Studio	0.028243
Cheese Shop	0.027636
Latitude	0.022561
Gastropub	0.021897
Music Venue	0.019851
Venue Count	0.010476

Figure 8: Positively correlated features to price

One interesting result here is the Latitude of the property which indicates the north-south position of the property. This also makes a lot of sense because usually as we move from the north of Toronto (e.g. Richmond Hill) to downtown, the prices increase. This is due to the reduced amount of real estate in the downtown core as well as other factors, like proximity to the subway line and financial district. Also, note that the venue count is a top contributor to the increase in price. This also is not surprising, since the more types of venues there is a around a property, the more accessible it is. That is a property which only has two types of venues around it should be less valuable than one with many different venues (restaurants, shops, move theatre, bars, etc.)

The features with a negative correlation to the price are shown. Notice that some of the positively features such as cafe and restaurants, have been replaced by pizza place, fast food restaurant and wings joints. Also, note that there are a lot of different cuisines from different cultures, such as Vietnamese, Korean, Middle Eastern, and Caribbean restaurants. This tells us about the property prices in various areas of Toronto that is home to a lot of the immigrants in the city.

Figure 9: Negatively correlated features to price

Pizza Place	-0.161243
Fast Food Restaurant	-0.131237
Arts & Crafts Store	-0.127560
Moving Target	-0.102029
Intersection	-0.096330
Bakery	-0.095666
Convenience Store	-0.090548
Wings Joint	-0.088101
Korean Restaurant	-0.077799
Shoe Store	-0.075897
Historic Site	-0.069627
Vietnamese Restaurant	-0.068871
Caribbean Restaurant	-0.061951
Middle Eastern Restaurant	-0.059073
Sushi Restaurant	-0.057590
Farmers Market	-0.049677
Longitude	-0.046943
Lounge	-0.046001
Shopping Mall	-0.044079

Another interesting find here is the Longitude that has been identified as a negatively correlated feature. That is, we move farther east or west in the city, the property prices reduce. This may be due to the transit proximity of the properties. As you may know the main subway line in Toronto runs south to north. There is a line that runs from west to east but it is mainly located in the downtown area. That is any property north of downtown does not have direct access to the subway line. It can also be due to other factors, such as poorer and less developed neighbourhoods, in the fast west and east corners of Toronto.

Logistic Regression with Lasso Regularization

Lastly, a logistic regression classifiers was run to predict the price categories (very low", "low", "medium", "high" and "very high"). In addition to the logistic regression model, a "SelectFromModel" object from Scikit-learn's feature selection library was used to extract the features with the highest weights from the logistic regression model. Logistic regression was used with lasso regularization to further emphasize important features and demote less important weights. By running the logistic regression model on our binned price values, three features were selected to have the most impact. These features were "Venue Count", "Cafe" and "Pizza Place". Looking at our highest positively and negatively correlated features from the previous section, we can see that the logistic regression model with lasso regularization has identified, "Venue Count" and "Cafe" as the best positively correlated feature, and "Pizza Place" as the top negatively correlated feature, confirming our results from the Pearson correlation analysis.

Future Recommendations

Here we have done a simple analysis on the property prices in the city of Toronto. There are a lot of other techniques that can be performed to further analyze the data and potentially find other useful insights on the factors affecting Toronto property prices. The first improvement that can be done, is to get more recent data. The property price data was from 2016, and had some property values in the dataset that needs further investigation. Also, we have used the location data from a 2018 version of Foursquare. Although two years does not change a lot of the landscape in Toronto, in terms of the venues that may exists in a neighbourhood, it would be better to have two fully synchronized datasets for prices, and location data working together. Finally, we have only explored two methods for analyzing the impact of features on property prices. There are many other algorithms such as Chi-Squared, Recursive Feature Elimination and Tree methods for feature selection.

Conclusion

We have explored the property prices in the city of Toronto. Property prices were obtained from an open source dataset from kaggle and combined with location data for the property coordinates from Foursquare. Pearson correlation as well as logistic regression with lasso regularization was used to find the features that most positively and negatively impact property prices in the city of Toronto. It was found the high end shops such as cafes and boutiques as well as playgrounds have a positively correlation with the property prices. It was also found that the number of different venues around a property as well as it's latitude (north-south position) have a positive effect on the property prices. It was also found that lower end pizza and fast food places were negatively correlated with the property prices. Other negatively correlated factors were identified as the longitude (east-west position) of property within the city.

References

[1] Zolo, June 2020, accessed June 7, 2020, <https://www.zolo.ca/toronto-real-estate/trends>

[2] Kaggle, 2016, accessed June 7, 2020, <https://www.kaggle.com/mnabaee/ontarioproperties>

[3] Pandas, 2014, accessed June 7, 2020, <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html>

[4] scikit-learn, 2019, accessed June 7, 2020, <https://scikit-learn.org/stable/modules/linear_model.html>