Big Data Major Research Paper

By
Rachael Joan Dias
0651897

Addressed to,

Professor Dr. Zeinab Noorian
Trent University

AMOD-5610H-A-F01-2019GF-PTBO

Analysis of Zoocasa Housing Dataset and Development of a Rent Appraisal Model

1. Abstract

A common problem that landlords face is accurately estimating the rent that they should charge. Generally, the price of a property is determined by a property appraiser. Unfortunately, these appraisers may be biased because of personal gains from the buyer or seller. Therefore, an automatic prediction system is helpful in estimating the price as a third party and which may be less biased this is the primary motivation for this project. In addition to this there are many factors such as location and nearby amenities provided by the location that are important in estimating the price.

2. Literature Review

Though the price of a real estate property is linked with the economy it is still quite hard to accurately determine the price based on the vast data that is available [1]. From previous research on the Boston Housing Dataset we know that the price of a house is dependent on its size and geographical location [3], [4]. In paper [2] around 25,000 listing of houses were collected from websites such as Centris.ca and duProprio.com these listings had around 100 features/attributes which were supplemented with data from Montreal Open Data Portal and Statistics Canada, in addition to this the distance from police station and fire station was added. The asking price of the house was predicted with an error of 0.0985 using an ensemble of kNN and Random Forest Methods. Principal Component Analysis (PCA) was used to reduce the dimensionality though it did not improve the performance of the kNN algorithm [2]. The data used in the paper had very few missing values, therefore they were able to achieve a very high accuracy.

Paper [5] uses feature selection and extraction methods with Support Vector Regression both methods give the same R-square score. Again, this paper asserts the fact that the living area and location are crucial in determining the price of a property. However, this paper did not use p-value to test the significance of the model. My study aims to collect data of houses that are available for rent in Toronto and GTA apply statistical techniques to find a correlation between the price, size and location. Finally, use these attributes to build a prediction model.

3. Introduction

The project that I had proposed to undertake as part of this course was the analysis of real estate listings and the development of a machine learning model that will estimate the price of a rental in a neighborhood. The initial phase of the project dealt with scraping housing data from online rental portals such as Realtor.ca and Zoocasa using Python's BeautifulSoup and Selenium libraries. My primary motivation for this project is to address a common problem faced by house owners, that is deciding the rent they need to charge their tenants. Since the prices estimated by property appraisers may be biased because of their own vested interested from buyers or sellers. An automated prediction model can serve as a third-party recommendation system which is less biased. There are many housing datasets available on Kaggle, but I was interested in collecting data for the city of Toronto and GTA.

In addition, to the listings acquired geographical information such as latitude and longitude was integrated using the Google Maps API in Python. In order to enrich the dataset walkscore was also collected for every address listed using the walkscore API in Python. Walkscore is a number which gives the walkability score of an address, it basically tells us how easily errands can be accomplished on foot from an address.

90-100: Walker's Paradise (Daily errands do not require a car)
70-89: Very Walkable (Most errands can be accomplished on foot)
50-69: Somewhat Walkable (Some errands can be accomplished on foot)
25-49: Car-Dependent (Most errands require a car)
0-24: Car-Dependent (Almost all errands require a car) [6]

A house that is situated in a rich neighborhood and that is close to nearby amenities (grocery store, bus terminal etc.) is more likely to cost more than houses that are situated in a poor neighborhood and away from the amenities. Thus, location and nearby amenities greatly influence the price of a rental.

From the information collected the exploratory plots show the variation of price based on location, variation of walkscore based on price, comparison between the leased and sold price, price based on the type of house. I had also proposed that I would integrating the housing dataset with sociodemographic data from Statistic Canada and the Major Crime Indicator dataset. However, I was unable to integrate sociodemographic information because I did not get access to the licensed Postal Code Conversion Files of Statistics Canada. Also, the Major Crimes Indicator dataset consisted of data for the last 10 years whereas the data that I collected was only 1 month of data.

Statistical analysis was performed on the dataset using R to determine the correlation coefficient R between the continuous variables like bedrooms, bathrooms, garage spaces etc. and the price. Finally, a basic multiple regression model was built in Python using the Scikit-learn package.

4. Data Description

The dataset consists of 17318 listings in total which were scraped from Zoocasa over the month of October 2019, it has 34 attributes of houses which are currently up for rent in Toronto and GTA. The MLS number is a unique code that is provided by Toronto's Real Estate Board to identify a property listing. Some of the other fields which were present for the listing give basic information about the house such as the address, sold price, asking price, no. of bedrooms, bathrooms, parking spots, garage spaces, age, type of house, etc. The lot size indicates the total land, it includes living area inside a house, the backyard and frontyard. On the other hand, size only indicates the living area inside the house. DateAvailable tells us when the property will be available for the tenants to move in and DaysActive is the number of the listing has been up on the site. The listings also had additional features such the amenities provided fireplace, central vacuum, basement, driveway, heating, AC and fuel. To enrich the dataset latitude, longitude and walkscore are also added. Fields such as GEO_NAME and Zipcode are derived from the address.

```
MLS              object
Address          object
Price           float64
ListPrice       float64
Beds              int64
Baths             int64
ParkingSpots      int64
Description      object
Type             object
Levels           object
Size             object
DateAvailable    object
DaysActive      float64
LaundryLevel     object
CentralVac       object
Fireplace        object
Acreage          object
LotSize          object
Garage           object
Exterior         object
Age              object
Basement         object
Driveway         object
GarageSpaces    float64
Heating          object
AC               object
Fuel             object
lat_lng          object
lat             float64
lng             float64
walkscore         int64
GEO_NAME         object
ZIP_CODE         object
Size1           float64
dtype: object
```

Figure 1: List of all attributes of the Zoocasa dataset

5. Methodology

Active and leased listing were scraped over a 1-month period from online rental portal Zoocasa with the help of Python's BeautifulSoup and Selenium libraries. Below are screenshots of the data.

MLS	Address	Price	ListPrice	Beds	Baths	ParkingSpots	Description	Type	Levels	...	Heating	AC	Fuel
E4579636	1168 Craven Rd Toronto, Ontario M4J0A2 Canada	3400.0	3550.0	3	3	1	Beautiful 3 Bedroom, 3 Bath Townhome, Steps To...	Townhouse	3-Storey	...	Forced Air	Central Air	Gas
W4592514	75 Bloomsbury Ave Brampton, Ontario L6P1S6 Canada	2800.0	2800.0	5	4	2	Location! Location! Location! Ravine Lot: Imma...	Detached	2-Storey	...	Forced Air	Central Air	Gas
W4599249	4132 Lastrada Hts Mississauga, Ontario L5C3W3 ...	2750.0	2700.0	3	3	2	Charming 3 Bedrm Detached Home In Sought After...	Detached	2-Storey	...	Forced Air	Central Air	Gas

Figure 2: Sample Data

Extra geographic information Latitude and Longitude were collected using the GoogleMaps API and this data was also used to collect the walkscore of every address with the Walkscore API in Python.

Description	Type	Levels	...	Heating	AC	Fuel	lat_lng	lat	lng	walkscore	GEO_NAME	ZIP_CODE	Size1
3 Bedroom, 3 Bath nhome, Steps To...	Townhouse	3-Storey	...	Forced Air	Central Air	Gas	[43.6824564, -79.3254843]	43.682456	-79.325484	94	Toronto	M4J0A2	1300.0
Location! Location! Ravine Lot: Imma...	Detached	2-Storey	...	Forced Air	Central Air	Gas	[43.7811467, -79.7240157]	43.781147	-79.724016	5	Brampton	L6P1S6	NaN
3 Bedrm Detached e In Sought After...	Detached	2-Storey	...	Forced Air	Central Air	Gas	[43.57436790000001, -79.66361239999999]	43.574368	-79.663612	73	Mississauga	L5C3W3	1750.0

Figure 3: Sample Data with geographic columns

The Zoocasa housing dataset has a normal distribution which is right skewed. Outlier were detected and removed for the analysis as they would have skewed the results. Raw HTML files generally contain errors, these errors may have been present in the original records or may be due to data entry errors. Unfortunately, many of the attributes had missing values.

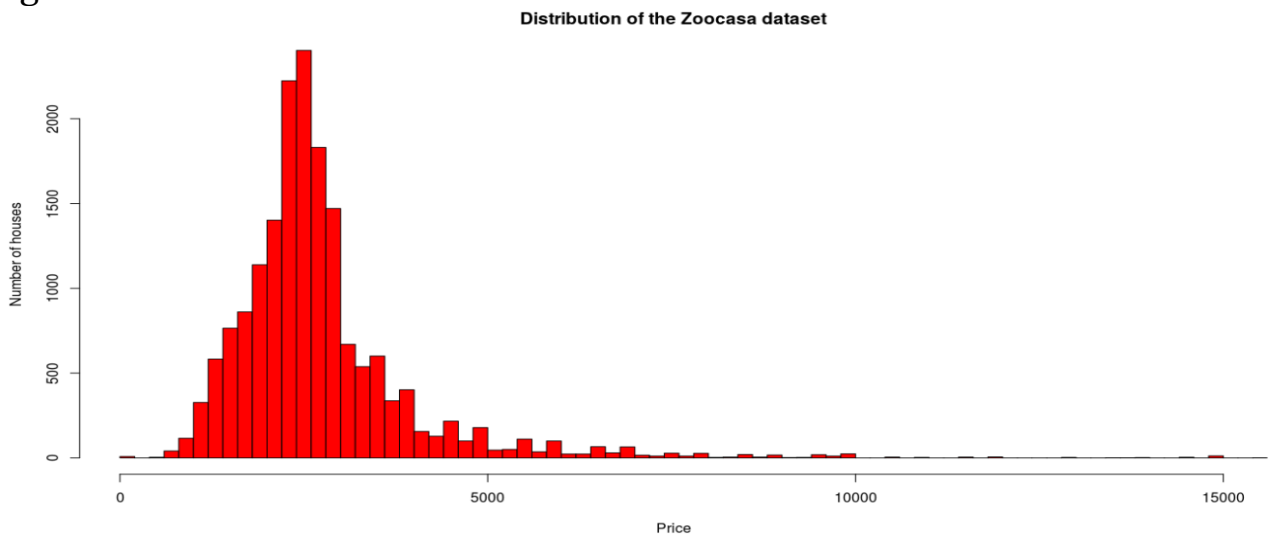

Figure 4: Price Distribution of the Zoocasa dataset

From the listing collected we can see that most of the listings belong to Toronto followed by Markham, Oakville and Brampton. The figure below shows the distribution of listings.

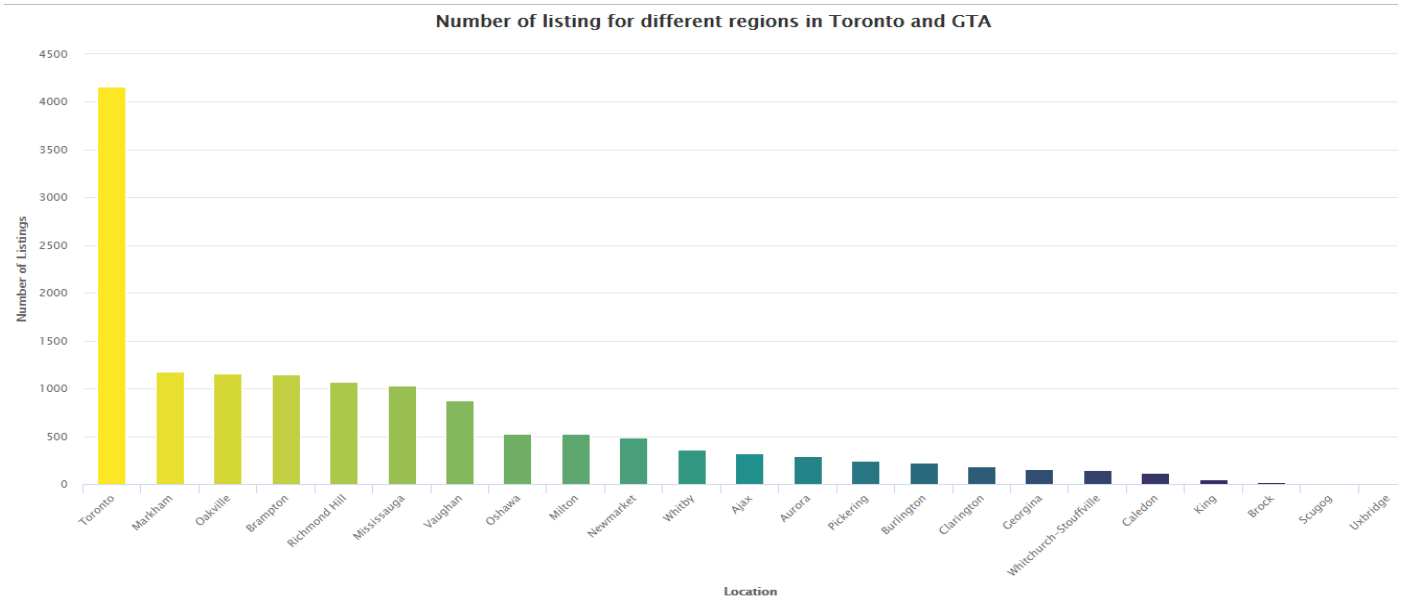

Figure 5: Number of listings vs location

In terms of price Oakville has the most expensive rentals as compared to any of the other locations. Vaughan, Mississauga and Burlington are 2nd, 3rd and 4th in terms of price. Toronto holds the 5th place, unfortunately the listings did not indicate which neighborhood the houses belong to.

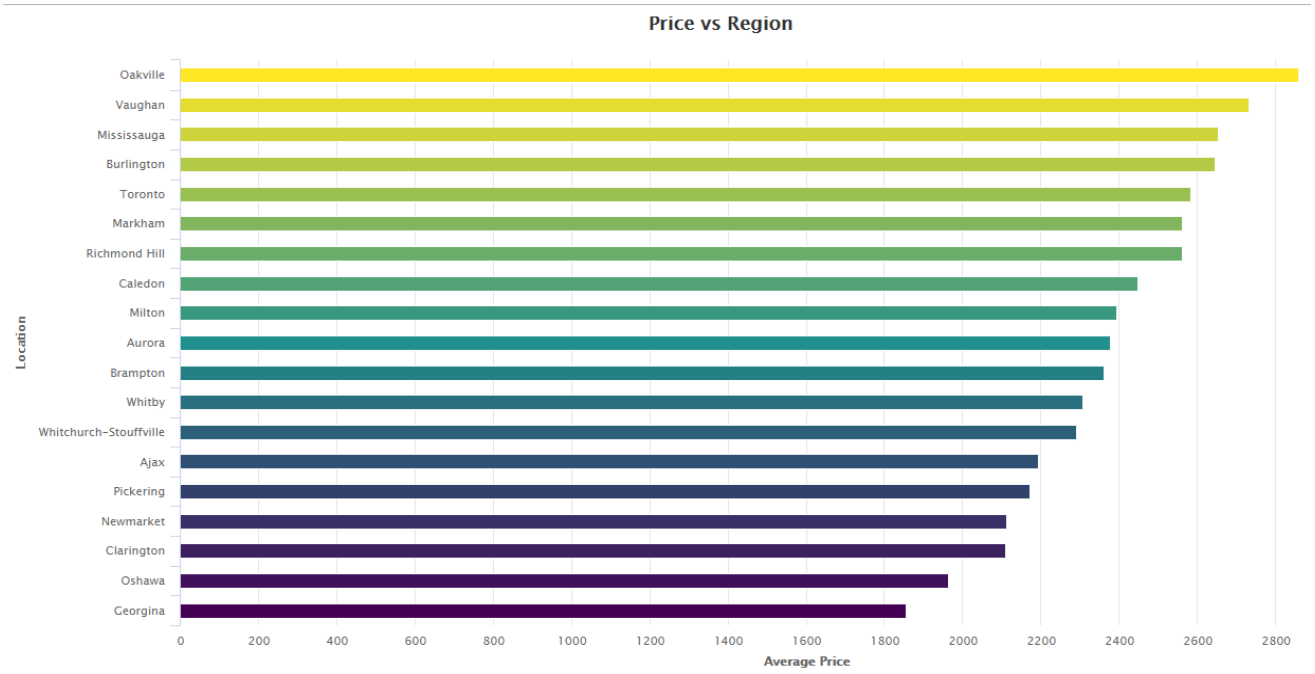

Figure 6: Price vs. Location

As expected the walkscore of Toronto is significantly higher when compared to the GTA.

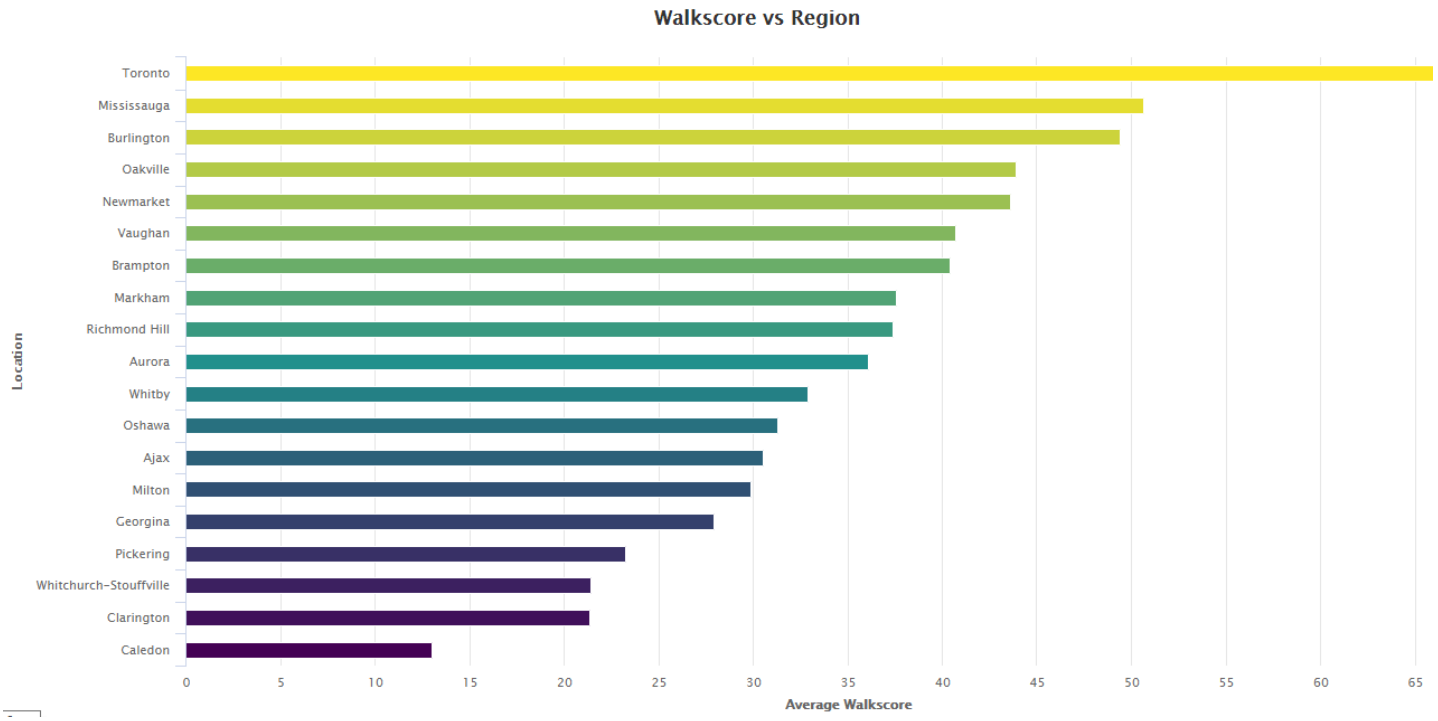

Figure 7: Walkscore vs. Location

Most of the houses that were listed are detached houses followed by Townhouses and Semi-Detached houses.

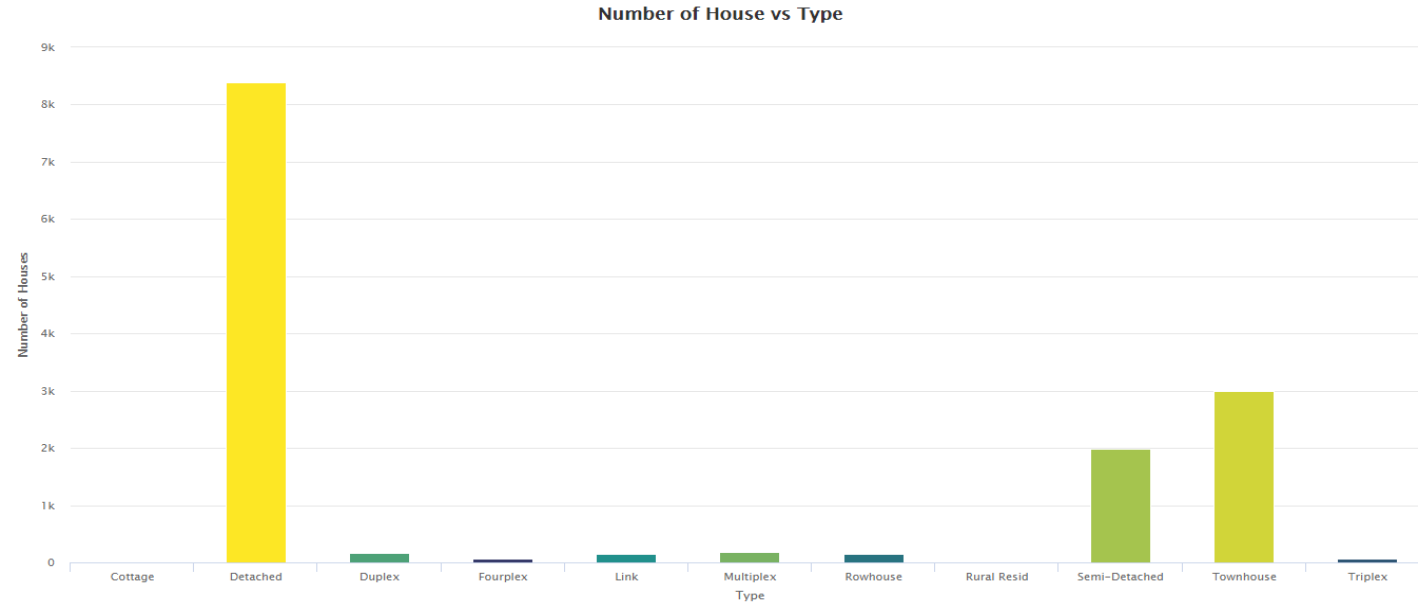

Figure 8: Number of Houses based on type

In the figure below we can see the average price of houses depending on the type. Many of the detached and semi-detached houses had only basements that were rented this explain the low average price. The problem with the Type attribute is that it did not accurately indicated the Type of the house some Rowhouses were listed as Townhouses and vice versa. Text processing would have helped to accurately determined the type of the house. Also, many of the detached and semi-detached houses rented only their basements but there was no indication of this in the Type column. Therefore, more pre-processing is required to accurately determined the house type from the dataset.

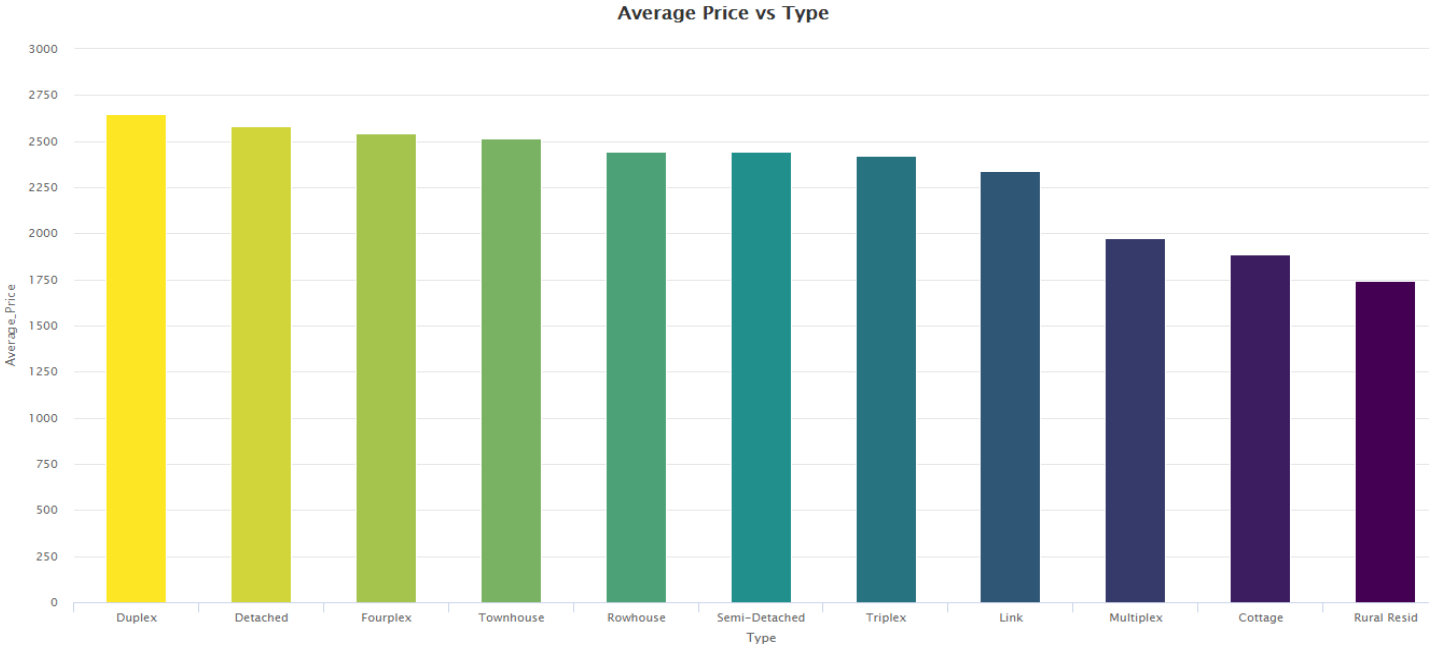

Figure 9: Average price of houses based on type

The figure below shows the comparison between the Sold Price and Asking Price. All the points that fall on red line are listings that have the same asking price and sold price, while the points below the red line are overpriced properties and those above the red line are underpriced properties.

Figure 10: Comparison between sold price and asking price

The figure below visualizes the correlation between average price and size. We can see that there is a moderate correlation between the price and size.

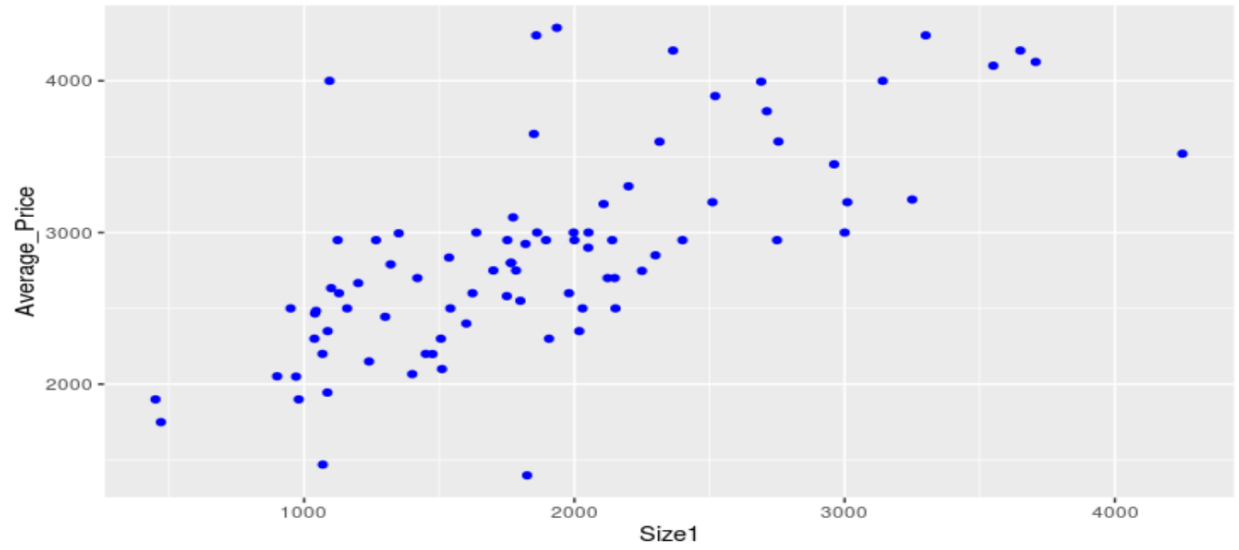

Figure 11: Scatter plot of average price vs. size

There is a downhill trend and a negative correlation coefficient of -0.2686063. This indicates that there is a tradeoff between size and walkscore. Staying close to the downtown area means that there will be an increase in the walkscore and decrease in the size of the house.

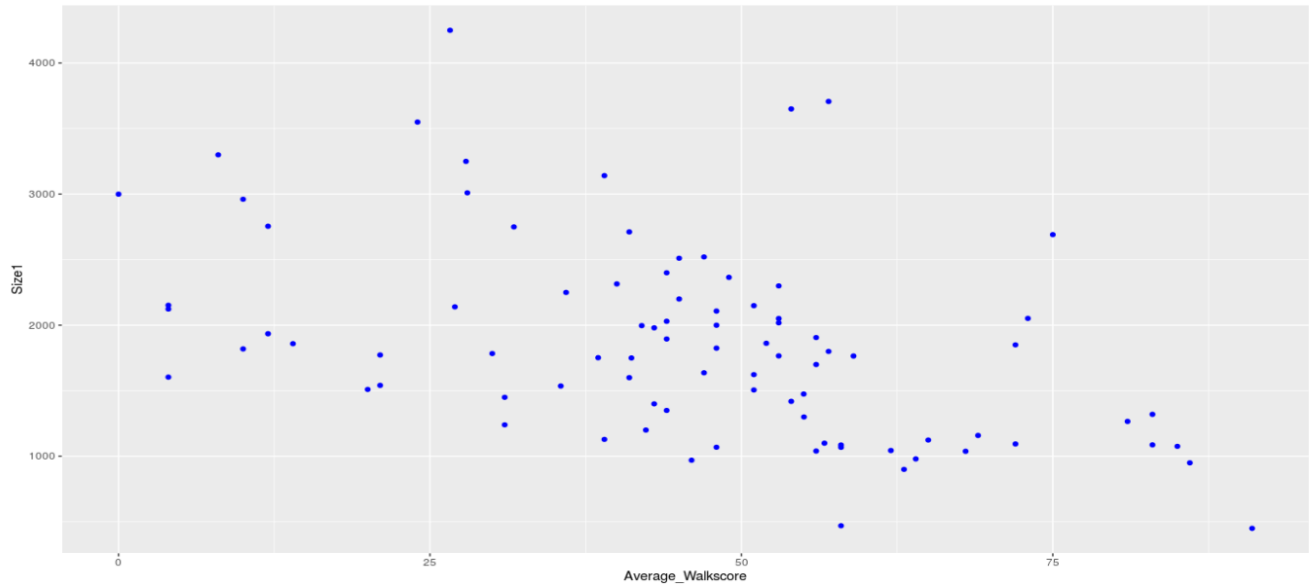

Figure 12: Scatter plot of size vs average walkscore

Variable	Correlation Coefficient
Beds	0.4208781
Baths	0.424842
Parking Spots	0.3141041
Garage Spaces	0.3451582
Size	0.5207605
Walkscore	0.1383759

Table 1: Correlation coefficient between all the continuous variables and price

The attributes beds, baths, parking spots, garage spaces, size, walkscore, central vacuum and fireplace have a p-value which is less than a significance level of alpha=0.5, so they are statistically significant and can be used to build a linear regression model.

```
Call:
lm(formula = b$Price ~ b$Beds + b$Baths + b$ParkingSpots + b$GarageSpaces +
    b$Size1 + b$walkscore + b$CentralVac + b$Fireplace + b$GEO_NAME +
    b$Heating + b$Type)

Residuals:
     Min       1Q   Median       3Q      Max
-2860.48  -178.84   -14.68   152.25  1511.99

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             270.18506  334.70963   0.807 0.419581
b$Beds                  124.57613    9.47066  13.154  < 2e-16 ***
b$Baths                 126.32292    7.52601  16.785  < 2e-16 ***
b$ParkingSpots           43.91540    4.78592   9.176  < 2e-16 ***
b$GarageSpaces           52.35234    9.98314   5.244 1.64e-07 ***
b$Size1                   0.22640    0.01149  19.697  < 2e-16 ***
b$walkscore               1.73198    0.24804   6.983 3.32e-12 ***
b$CentralVacYes          52.10288   14.07145   3.703 0.000216 ***
b$FireplaceYes           99.13836   11.88821   8.339  < 2e-16 ***
```

Figure 13: output of model summary in R, p-values for all continuous variables and categorical variables central vacuum and fireplace

Similarly, the p-value for different locations are statistically significant.

```
b$GEO_NAMEAurora               -10.82235   42.90342  -0.252 0.800860
b$GEO_NAMEBrampton             189.08033   34.34010   5.506 3.87e-08 ***
b$GEO_NAMEBrock               -714.36663  101.81356  -7.016 2.62e-12 ***
b$GEO_NAMEBurlington           468.18847   44.13592  10.608  < 2e-16 ***
b$GEO_NAMECaledon               54.34574   56.08385   0.969 0.332593
b$GEO_NAMEClarington          -270.68608   48.07856  -5.630 1.91e-08 ***
b$GEO_NAMEGeorgina            -395.33433   61.84118  -6.393 1.80e-10 ***
b$GEO_NAMEKing                 332.48841   86.29243   3.853 0.000118 ***
b$GEO_NAMEMarkham              124.26350   33.77515   3.679 0.000237 ***
b$GEO_NAMEMilton               191.16571   36.07412   5.299 1.22e-07 ***
b$GEO_NAMEMississauga          443.47948   35.67130  12.432  < 2e-16 ***
b$GEO_NAMENewmarket           -195.55498   39.72098  -4.923 8.82e-07 ***
b$GEO_NAMEOakville             571.25106   34.42771  16.593  < 2e-16 ***
b$GEO_NAMEOshawa              -261.65538   39.70533  -6.590 4.91e-11 ***
b$GEO_NAMEPickering            107.84573   47.61427   2.265 0.023561 *
b$GEO_NAMERichmond Hill        179.33492   34.44636   5.206 2.01e-07 ***
b$GEO_NAMEScugog               815.60970  230.65696   3.536 0.000410 ***
b$GEO_NAMEToronto              807.35293   35.38378  22.817  < 2e-16 ***
b$GEO_NAMEUxbridge             252.92651  322.76990   0.784 0.433309
b$GEO_NAMEVaughan              358.86692   34.85811  10.295  < 2e-16 ***
b$GEO_NAMEWhitby               -69.09375   39.88109  -1.732 0.083255 .
b$GEO_NAMEWhitchurch-Stouffville -115.71530   54.41281  -2.127 0.033506 *
```

Figure 14: output of model summary in R, p-values all the location variables

Type of heating is statistically significant, unfortunately the Type variable doesn't show a statistical significance. Here the Adjusted R-squared tells us that roughly 65% of the prices can be accurately predicted by our regression model.

```
b$HeatingFan Coil              164.10927  107.99567   1.520 0.128684
b$HeatingForced Air            132.83860   84.34492   1.575 0.115340
b$HeatingHeat Pump             337.49121  110.07526   3.066 0.002182 **
b$HeatingOther                 414.66022  156.69834   2.646 0.008168 **
b$HeatingRadiant               442.01432  101.32305   4.362 1.32e-05 ***
b$HeatingWater                 372.15414  103.10425   3.609 0.000310 ***
b$TypeDetached                 437.18589  321.71040   1.359 0.174233
b$TypeDuplex                   757.21100  330.63762   2.290 0.022059 *
b$TypeFourplex                 769.55667  344.30441   2.235 0.025460 *
b$TypeLink                     418.52896  324.77550   1.289 0.197579
b$TypeMultiplex                642.65718  335.03386   1.918 0.055151 .
b$TypeRowhouse                 437.46229  324.10336   1.350 0.177161
b$TypeSemi-Detached            412.00856  321.95408   1.280 0.200713
b$TypeTownhouse                440.81968  321.72303   1.370 0.170698
b$TypeTriplex                  365.88508  344.28496   1.063 0.287958
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 320.8 on 4475 degrees of freedom
  (9729 observations deleted due to missingness)
Multiple R-squared:  0.6588,    Adjusted R-squared:  0.6554
F-statistic:   192 on 45 and 4475 DF,  p-value: < 2.2e-16
```

Figure 15: output of model summary in R, p-values for categorical variables heating and type

We can see that there is a moderate correlation between all the continuous variables and the price. I chose the all the continuous variables and few of the categorical variables to build a multiple regression model.

Continuous variables:

- Bedrooms
- Bathrooms
- Parking Spots
- Garage Spaces
- Size
- Walkscore

Categorical variables:

- Location
- ^Fireplace
- Heating
- Type of House

General form of multiple regression equation,

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

Initial data consisted of 17318 records. After removing outliers and missing values only 5671 records were left for training the model. All records without Size were dropped and values for GarageSpaces were imputed using the most common value. Dummy variable coding or one hot encoding was used for the GEO_NAME column.

This figure shows the correlation after discarding outliers, missing values and imputing values that are missing.

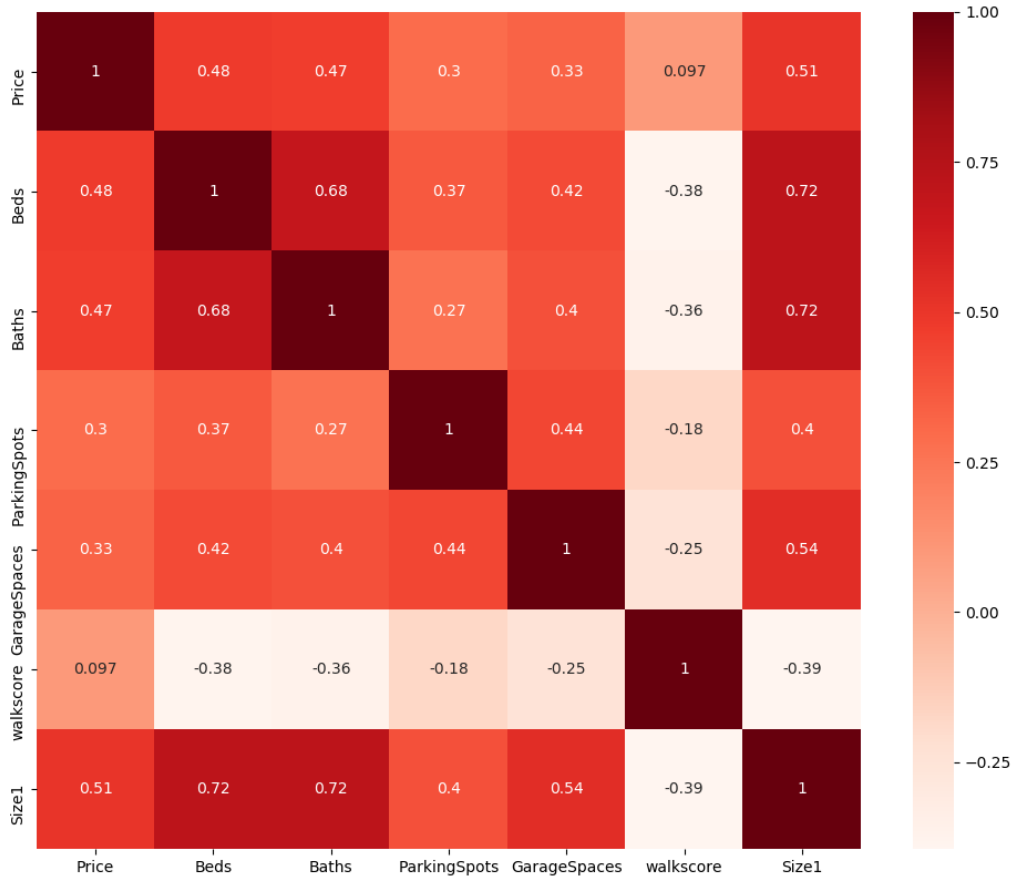

Figure 16: Correlation Matrix

6. Results

R-squared is the coefficient of multiple determination for multiple regression. The R-squared indicates that roughly 68% of the houses listed can be accurately predicted by the model, higher the R-squared better the model fits the data. The mean absolute error measures the magnitude of errors in a set of predictions, it is the absolute difference between actual value and the predicted value, this multiple regression model has a mean absolute error of 244.35. On the other hand, it has a root mean squared error of 334.88 which is the square root of the squared difference between the actual rent and predicted rent. Similarly, a mean squared error of 112146 is the average squared difference between the predicted rent and actual rent.

Parameter	Score
R-squared	68.33712509378525
Mean Absolute Error	244.35071105085967
Mean Squared Error	112146.36572225354
Root Mean Squared Error	334.8826148402654

Table 2: R-squared, mean absolute error, mean squared error and root mean squared error

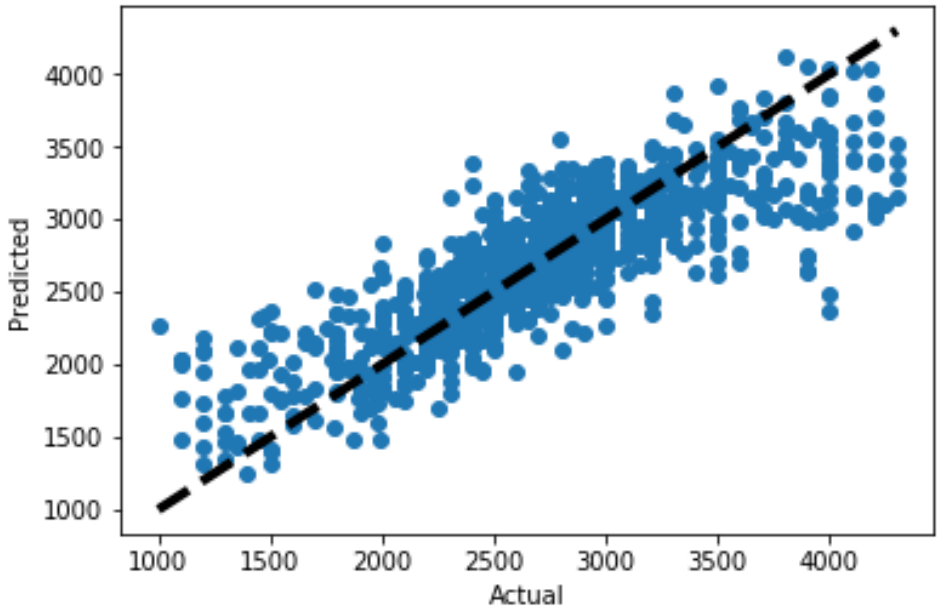

Figure 17: Actual vs Predicted values

7. Conclusion

From the data collected we can conclude that the most expensive regions are Oakville, Vaughan and Mississauga. While Clarington, Oshawa and Georgina are much lower in price in comparison to other regions in the GTA. The walkscore for Toronto is much higher in comparison to the GTA. Mississauga, Burlington and Oakville have higher walkscores when compared to the regions in GTA. As the walkscore increase there is a decrease in the size of the house, walkscore and size are negatively correlated. The negative correlation between size and walkscore clearly indicates that there is a trade-off between size and walkscore, when renting a house, a compromise must be made between the size and the price of the house. There is a very weak correlation with the walkscore and price, this may also be because some of the houses didn't have the walk score computed. The multiple linear regression model trained on the data collected was able to predict the rent with roughly 68% accuracy.

8. Future Research

Since the price of property is highly correlated with the economy, enriching the dataset by incorporating Statistic Canada's demographic data and the Major Crimes Indicator dataset would have really helped build a better model. There are many demographic features such as the age, average income, employment, unemployment rate, crime rates etc. of different location which would help make a more accurate prediction. I also believe that performing text analysis on the description of listings would help to derive more insight. I used the correlation coefficients 'R' to determine important continuous feature, however I randomly selected the categorical feature since there were too many of them and they required additional preprocessing. Thus, feature importance or selection techniques and feature engineering techniques will also help build a better predictive model. Finally, experimenting with other models such as Random Forrest Regression, Ensemble, kNN etc will help identify the right model for prediction.

9. References

[1] R. J. Shiller, “Understanding recent trends in house prices and home ownership,” National Bureau of Economic Research, Working Paper 13553, Oct. 2007. DOI: 10.3386/w13553. [Online]. Available: http://www.nber.org/papers/w13553.

[2] Nissan Pow, Emil Janulewicz and Liu (Dave) Liu, Applied Machine Learning Project 4 Prediction of real estate property prices in Montr´eal

[3] D. Belsley, E. Kuh, and R. Welsch, Regression Diagnostics: Identifying Influential Data and Source of Collinearity. New York: John Wiley, 1980.

[4] J. R. Quinlan, “Combining instance-based and modelbased learning,” Morgan Kaufmann, 1993, pp. 236–243.

[5] J. Y. Wu, “Housing Price prediction Using Support Vector Regression.”

[6] https://www.walkscore.com/methodology.shtml