Rental prices of apartments in Toronto

OBJECTIVE

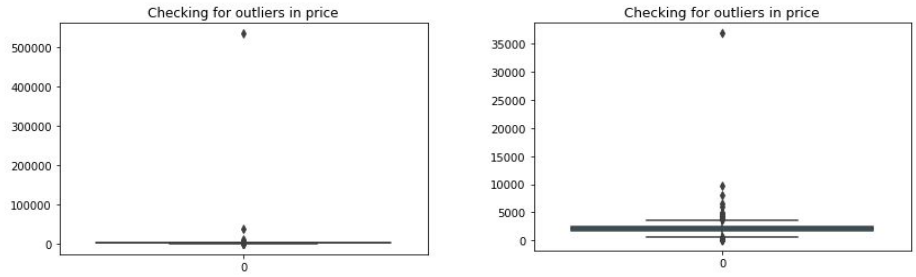
This report is the analysis of the cost of rental apartments within Toronto. The analysis is divided into two different segments. One of the sections will demonstrate hypothesis testing, while the other section is to build a predicting model for the pricing of Toronto rental apartments.

- Problem Statement: Someone searching for a one bedroom apartment to rent in downtown Toronto or nearby places.
    - Null Hypothesis: The cost of rent of apartments located in Downtown Toronto, are in the same range as the areas outside downtown ($\mu = \mu 0$). The outside places may include North York, Etobicoke, Mississauga etc.
    - Alternative Hypothesis: The cost of rent of apartments located in Downtown Toronto is higher than the apartments located outside downtown. ($\mu > \mu 0$).

- Predictive model: To build a predictive model that would take into consideration all the independent variables that are found within our data set, to establish or determine the cost of rent for an apartment.

DATA PREPARATION

The data set we used for this analysis was readily available on Kaggle. The data quality procured was very clean, without any null values. We had to remove a few duplicate records and a few records that were outside of the GTA, as this analysis focuses within the GTA. This data also consisted of some outliers, that gave of rental prices of $36,900 and $535,000, and so it was decided to remove these outliers as we did not consider them to be realistic.



There were a few challenges within the data set. For example, the Address attribute, the information was not always complete. To solve this issue, we added a new attribute named ZIP which extracts the first three characters of the postal code found in the Address attribute. Another attribute we created was Borough, which is based on classifying the record into Downtown Toronto, North York, Central Toronto, etc.

	Bedroom	Bathroom	Den	Address	Lat	Long	Price	ZIP	Borough
3	1	1.0	0	89 Chestnut St, Toronto, ON M5G 1R1, Canada	43.654155	-79.385211	550	M5G	Downtown Toronto
4	1	1.0	0	, Toronto m5s1x6 ON, Canada	43.665956	-79.404799	650	M5S	Downtown Toronto
5	1	1.0	0	, toronto m4b2z5 ON, Canada	43.705190	-79.323847	700	M4B	East York
6	1	1.0	0	, M5A 2V3, Toronto, ON	43.654228	-79.367015	700	M5A	Downtown Toronto
7	1	1.0	0	, Toronto M5G 1B1 ON	43.654723	-79.381400	700	M5G	Downtown Toronto

We used python in order to prepare the data, perform analysis and modelling. In python, we used libraries such pandas, matplotlib and seaborn.

ANALYSIS

Hypothesis Testing:

H0: The cost of rent of a one bedroom apartment located in Downtown Toronto is in the same cost range as the area outside of Toronto downtown.

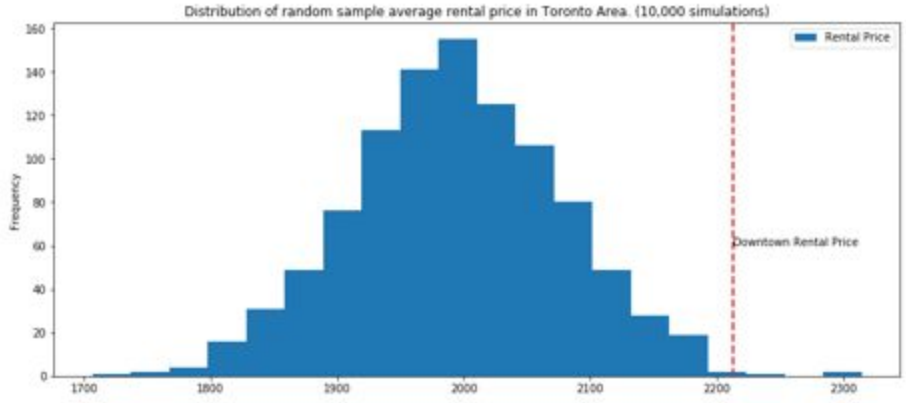
H1: The cost of rent of a one bedroom apartment located in Downtown Toronto is higher than the apartments located outside of the downtown region. There is a significant difference between the prices between them.

For this analysis, we prepared a sample of 40 apartments. The mean price of the sample was $2213.

In the following equation, Z is the Z-score of the sample mean, $\bar{x}$ is the sample mean, μ is the overall mean, Sx is the total standard deviation, n is the number of the samples taken for the hypothesis and CI is the Confidence Interval.

$$Z = (\bar{x} - \mu)/ (Sx/\sqrt{n}) = (2213 - 1993)/(542/\sqrt{40}) = 2.567 \quad (CI = .95)$$

If α <=.025, there is a significant difference between the prices of Downtown Toronto and nearby areas. Using the normal probability table, the value of p is found to be 0.995. The value of alpha is less than 0.025.



MODELLING

The model we built in order to predict the target variable (i.e. price of the rental apartment) is a multi linear regression. There are five independent variables that are taken into consideration: number of bedrooms, bathrooms, dens, and latitude and longitude of the location. We would model this data set using two different methods we learned in this course.

APPROACH 1: Multi Linear Regressions using OLS results.

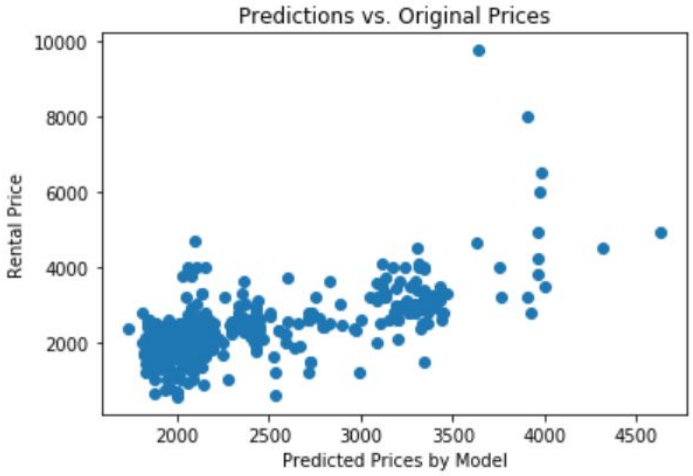
Firstly, in order to standardize the scale (eliminate the multicollinearity), the latitude and longitude values of all rental units were subtracted from the Toronto City Hall (Long = -79.3832, Lat =

43.6532) which is located in Downtown and also considered as our 'reference point'. On running the OLS, it was observed that the 'p-value' for longitude was calculated to be 0.820. As the value was greater than 0.05, it is statistically acceptable to remove this variable from the model. And as a result the new model generated using OLS was:

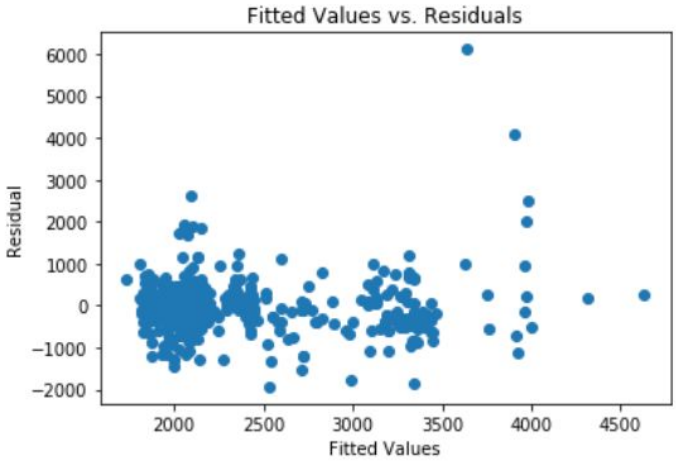
(I) *Price = 760.02 + 576.61 x Bedrooms + 675.45 x Bathroom + 279.13 x Den - 1.085e+04 x (Lat - 43.6532)*

Evaluating the model:

1. The rental prices and predicted prices on a scatterplot ideally should contain all the data on the 45 degree line (x=y). But for this model the data falls below the 45 degree line. It can be concluded that the model seems to overpredict the price.

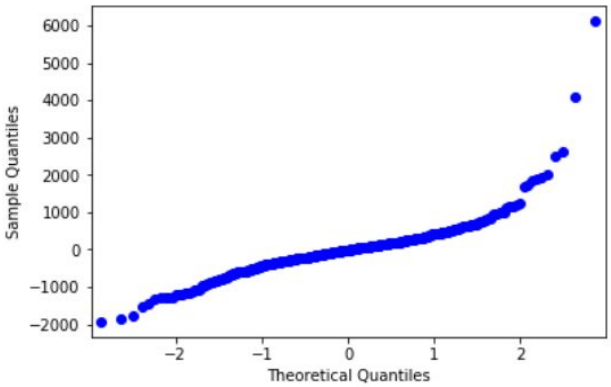


2. The scatter plot between residual and fitted values is used to test that the errors in the model are normally distributed. There is some little to no correlation in the model, but no definite pattern is observed. Hence, it is safe to say that the errors are distributed normally.

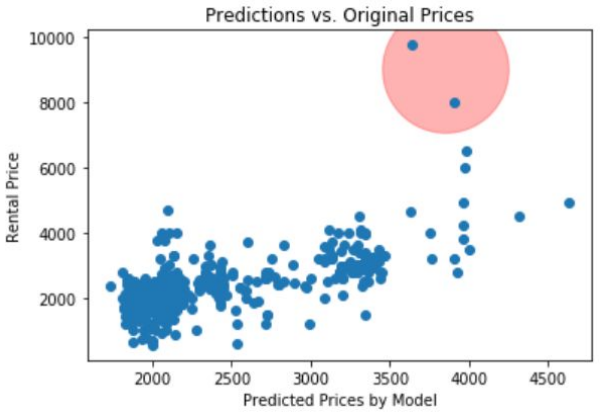


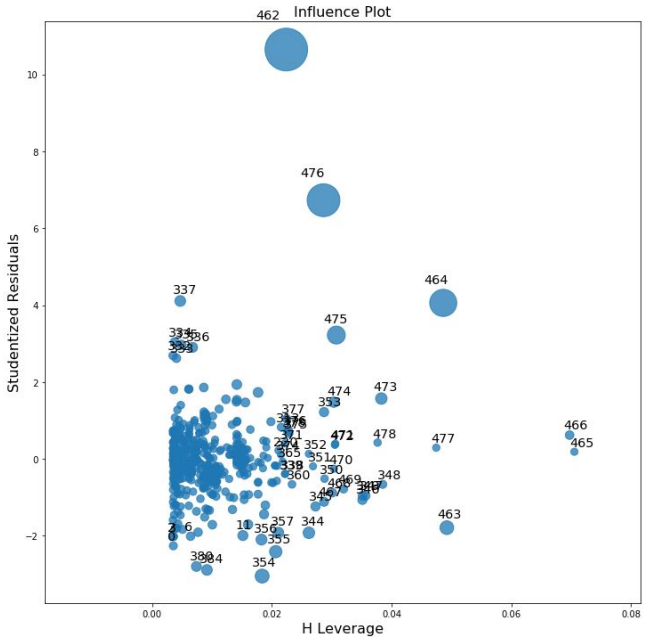
3. Q-Q plots take the sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution. The straighter the line, the more normal the

distribution. The line is reasonably straight but curves up and then down which suggests that the data is wide and flat in its distribution.



Outliers & Leverage Points:



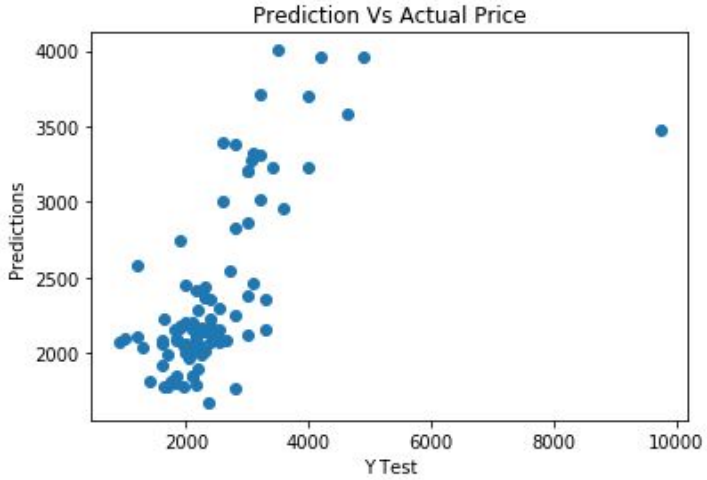


On visualizing the graph, it is evident that the points indicating the predicted price to be more than $7800, and in our case this would be considered an 'outlier'. Furthermore, the high leverage points also did not fit well with the model either. Both the outlier and leverage points (7) were removed in order to acquire a better fit for the model. The new model generated using OLS became:

(II) *Price = 946.8469 + 568.41 x Bedrooms + 507.01 x Bathroom + 294.14 x Den - 8983.18x (Lat - 43.6532)*

APPROACH 2: Testing/Training the Model.

In the second technique to model the data, training was performed on 80% of the data available. Based on the training, the testing was done on the remaining 20%. The data set is randomly distributed into the two segments, and standardized using *StandardScaler*. Next, GridSearch provided the best parameters to train the data ('copy_X': True, 'fit_intercept': True, 'normalize': True). Based on the algorithm (*LinearRegression*) and the parameters, the predictions are made for the test data set. The accuracy of the predicted price vs the actual price values was calculated to be 49.88%. Moreover, the values for Mean Square Error was 260692.35224264543 and Root Mean Square Error was 510.5804072255862.



CONCLUSIONS

I. Hypothesis Testing: The null hypothesis states that the prices for one bedroom apartments are in the same range for Downtown Toronto and nearby locations. But on performing hypothesis testing, it is safe to reject it. The 'p-value' calculated was found to be less than α. Hence we accept the alternative hypothesis which claims the price for the rental apartments in Downtown Toronto are more than locations further from our reference point, Toronto City Hall.

II. Modelling using OLS Results: After removing the outliers and influential points, the R^2 value increased from 0.431 to 0.469. The increase significantly made the model a much better fit for the data.

III. Testing/Training Model using Linear Regression as a Classifier: The result using the best parameters gave an accuracy score of 49.88%.

IV. We were unable to build a successful model using the above methods. There were some missing variables in the data set which could have helped in order to predict the target variable in a more effective and efficient way.

A. The age of the rental apartments has a major impact on the price. Units can be recently built, 10 years old or 25 years old. Newly built apartments comes with many amenities.
B. Access to transit and nearby plaza (grocery/convenience store) can also play a major role in the prediction of the price.
C. The size of the apartment was also not taken into consideration. Some apartments tend to have larger bedrooms or living rooms.
D. An addition of a balcony can play an important role in estimating the target variable.
E. Lastly, the neighborhood can be more considered more ‘unsafe’, based on criminal activities in the vicinity.