Final Summary Report

Introduction and Objective

It is quite interesting to use analytic models to price the property. In this project. We focus on the renting properties in GTA. The dataset is scraped from trebhome.com using BeautifulSoup and Requests with Python. The objective of this project is to find which properties are the best for investments. The result of this study can give some information to the property investors to price the renting properties, answering questions like

1. What is the best rental price for the property? If too low, no profit gain; if too high, the property may not be rented.
2. What is the best property that I need to buy for investment based on the predicted rental price?

Method/ Data

The dataset is scraped from trebhome.com posted on March 20 to March 24, 2019 using Beautifulsoup and Requests with Python. The filters used are area" Toronto" , min_price"$300" , and max_price "$50,000" (Excluded locker and parking spot lease priced below $300) .The dataset contains 3164 records and 6 variables with duplicates. The variable names are ID, Address, Bedrooms, Bathrooms, Type, Price. The final dataset is written as rental1.csv file.

The Dataset is cleaned and analysed with R. The statistics, frequency, summary table, decision tree, Random Forest, and Regression models were generated in this project.

Interpretation of Results

Statistics

Statistics about Price, Bedrooms , Bathrooms .

Price:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
650	2100	2475	3022	3200	22500

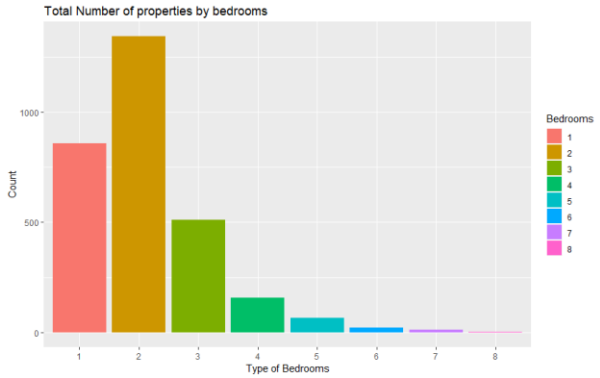
Bedrooms:

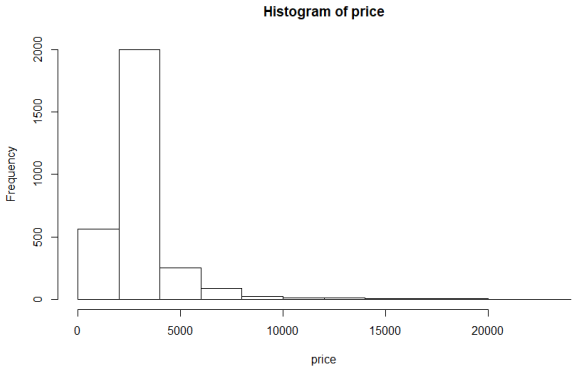
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	1.000	2.000	2.104	3.000	8.000

Bathrooms:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	1.000	1.000	1.634	2.000	8.000

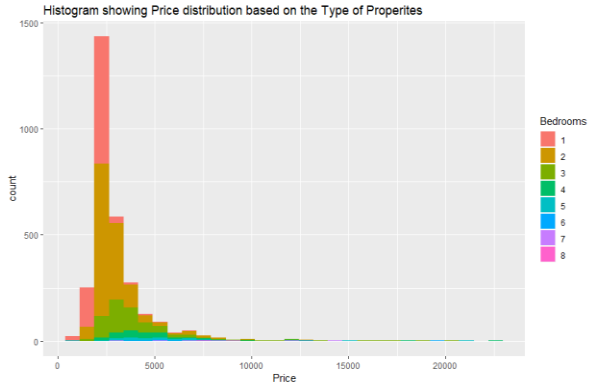
Frequency and Price Distribution

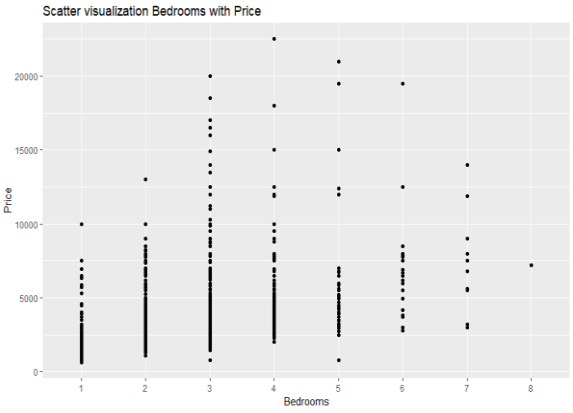


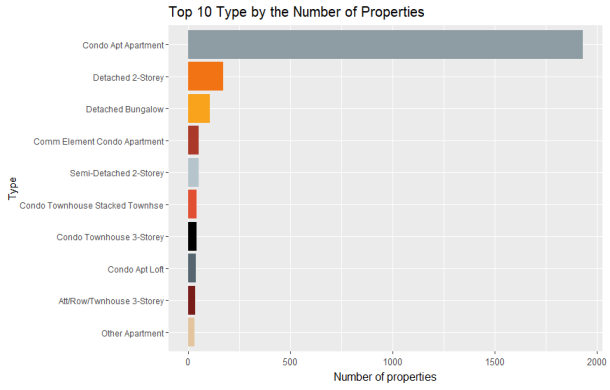


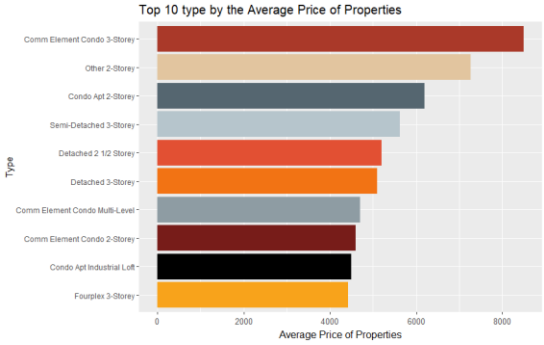
The highest frequency is 2-bedroom type.

The Price distribution is right skewed.









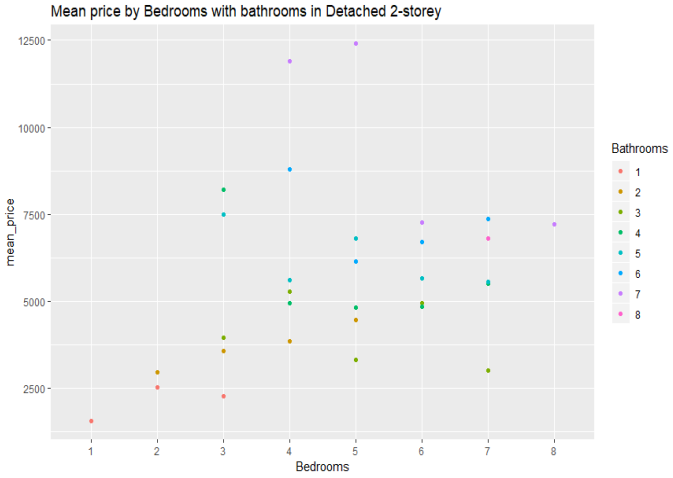
Summary Table

Average Price by Bedrooms (High Level)

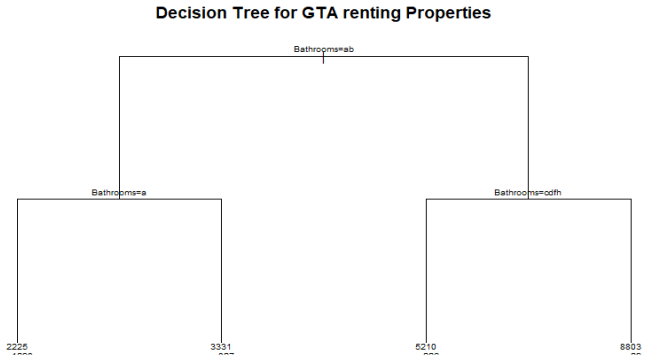
Bedrooms	mean_price2
1	2080.24
2	2816.92
3	4209.52
4	5707.04
5	6445.36

6	7533.64
7	7732.81
8	7200.00

Visualize Avg Price by Bedrooms and bathrooms in Detached 2-storey



Modelling (Decision Tree vs. Random Forest with ntree=500)



It can be concluded random forest(ntree=500) performs better than a single decision tree because both the mae or rmse show less in random forest model while compared to a single decision tree.

Regression Model

Model 1: Look at normal Price with Bedrooms and Bathrooms

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	830.75	64.97	12.787	<2e-16	***
Bedrooms	119.35	40.31	2.961	0.0031	**
Bathrooms	1186.27	47.39	25.033	<2e-16	***

Residual standard error: **1355** on 2369 degrees of freedom
Multiple R-squared: 0.4118, **Adjusted R-squared: 0.4113**
F-statistic: 829.4 on 2 and 2369 DF, p-value: < 2.2e-16

Model 2: Look at log(Price) with Bedrooms and Bathrooms

To overcome heteroskedasiticity with building log(Price).

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.34605	0.01389	528.679	<2e-16	***
Bedrooms	0.07407	0.00862	8.593	<2e-16	***
Bathrooms	0.25249	0.01013	24.912	<2e-16	***

Residual standard error: **0.2897** on 2369 degrees of freedom

Multiple R-squared: 0.4874, **Adjusted R-squared: 0.487**

F-statistic: 1126 on 2 and 2369 DF, p-value: < 2.2e-16

It shows linear_model2 is better than linear_model1. Linear_model 2 has Adjusted R-squared: 0.487, p-value: < 2.2e-16. (but still not good enough, further study needed)

The relationship shows price with bedrooms and bathrooms is
log(Price)=0.07407*number of Bedrooms +0.25247* number of Bathrooms +7.34605.
This shows that the number of Bedrooms has stronger positive relationships with the renting Price than the number of bathrooms.

This model shows rmse=2312.589 (actual=regression_test$Price,predicted=test_predictions).

Discussion and Challenges

The aim of this analysis is to answer a question of "Which ones are the best for investments?" Finding "BEST" is hard and it is a subjective matter.

Therefore, rather than concluding which ones are the best for investments, it is much wiser to perform further research about the Toronto area. Since in this analysis we are missing some potentially important variables related to properties. It is possible that the properties have higher rental prices because of better location, low crime rate, convenient transportation, and higher standard interior decorations of property etc.

Also, another challenge is that the most important variable to determine the renting price is Bedrooms in Decision Tree model; However, the most important variable to determine the renting price is Bedrooms. This project did not figure out why this discrepancy happens. It requires further studies.