STAT 450 Project: Predicting Condo Prices

David Yin, 13922159

Summary

The client is a real estate agent in Metro Vancouver who actively buys and sells mainly condo homes for her clients. She is interested in knowing the potential values of condo homes in three neighborhoods: Collingwood, Metrotown, and Whalley. Statistical methods were performed on the data provided to determine trends in condo price and estimate future condo price per square foot. Exploratory analysis has shown that the BC Investment Immigration Program preceeded a large boost in condo price. A forecast using Seasonal ARIMA has been fairly accurate for Collingwood and Whalley but not so much for Metrotown. Random forest, which predicts price based on multiple characteristics associated with a condo, has given us a prediction error of 6.7% for Collingwood, 7.8% for Metrotown, and 7.7% for Whalley.

Introduction

Our aim is to predict the monthly resale condo price per square foot in 2018, for each of the three sub-regions in the Greater Vancouver area. They are Collingwood, Metrotown and Whalley. We also seek to identify which region have a high growth potential. Additionally, we hope to analyze the potential sales growth of condos of different sizes in 2018.

Once we have accomplished these goals, we plan on comparing our predictions for December 2017 and January 2018 with market data to see how well our models work. Moreover, we would like to see if our predictions are consistent with those done by Statistics Canada or other institutes.

Data Description

To address these problems, we will use sales records on MLSLink (https://idp.gvfv.clareitysecurity.net). We intend on obtaining information on condos in the three aforementioned regions from 01/01/2006 to 12/31/2017. The variables we intend on using, from this dataset, are:

- Price
- Days on Market
- Total Bedroom
- Total Floor Area
- Age
- Year Built
- Address
- Bylaw Restrictions (such as pet allowance).

We hope to also collect external data on variables whose characteristics we believe have an association with condo prices. They include but are not limited to:

- Population (census data) - available for years 2006, 2011 and 2016

- Foreign exchange rate (forex.com)
- Government policies, such as the 15% foreign buyer tax implemented in 2016
- Interest rate (bank of canada)
- School ranking in the regions (various sites)

We did some data cleaning to make sure that outliers were excluded from our model. For example, abnormally old condos were excluded, as well as those without a price shown.

Methods

Exploratory Analysis

Before predicting the future of Greater Vancouver's real estate market, it is better for us to understand what has happened in the past. Hence, we used exploratory data analysis (EDA) to visualize the main characteristics of our data. EDA mainly focuses on checking assumptions, handling missing values and making necessary transformations of variables, which make it a good approach to explore the data and see what it suggests before model fitting or hypothesis testing. For example, by plotting the price of condos with respect to their age, we found that the condos with age of 999 were extremely old and would affect the accuracy of our model fitting, so we removed the data of such condos as outliers.

The visualization package we used in R is called "ggplot", which allows us to construct the initial plot object and add other components to the plot. We included external factors such as foreign currency exchange rates and government policies in the plots. Again, it helped us visualize their influences on the condo market.

ARIMA

To analyze the time series data, we decided to use AutoRegressive Integrated MovingAverage (ARIMA) model with an assumption that the price per square foot depends on both previous prices and white noise. We mainly used auto.arima function, which is part of R's forecast package, to make predictions.

auto.arima searches through multiple combinations of ARIMA parameters and selects the set that optimizes the model fit criteria. In addition, auto.arima is flexible enough to handle seasonal effect and trend in the time series data.

For each district, the time series data from January 2006 to September 2017 was used to fit the ARIMA model. Then, we used the fitted model to predict the price per square foot from October to December in 2017.

The blue line in the following three plots represents the prediction generated by the fitted ARIMA models. The black line represents the actual observed values. The dark grey and light grey shaded regions represent 95% and 80% confidence interval, respectively.

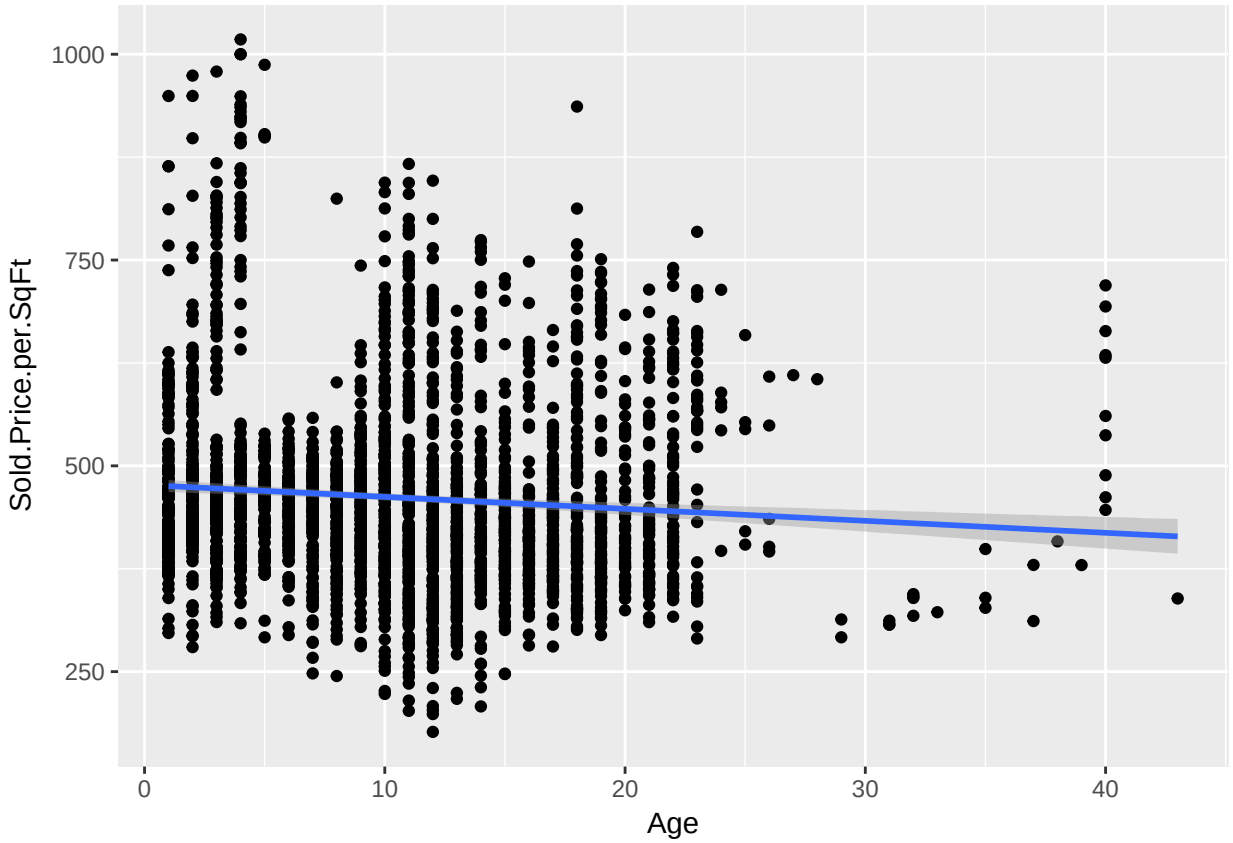
Random Forest

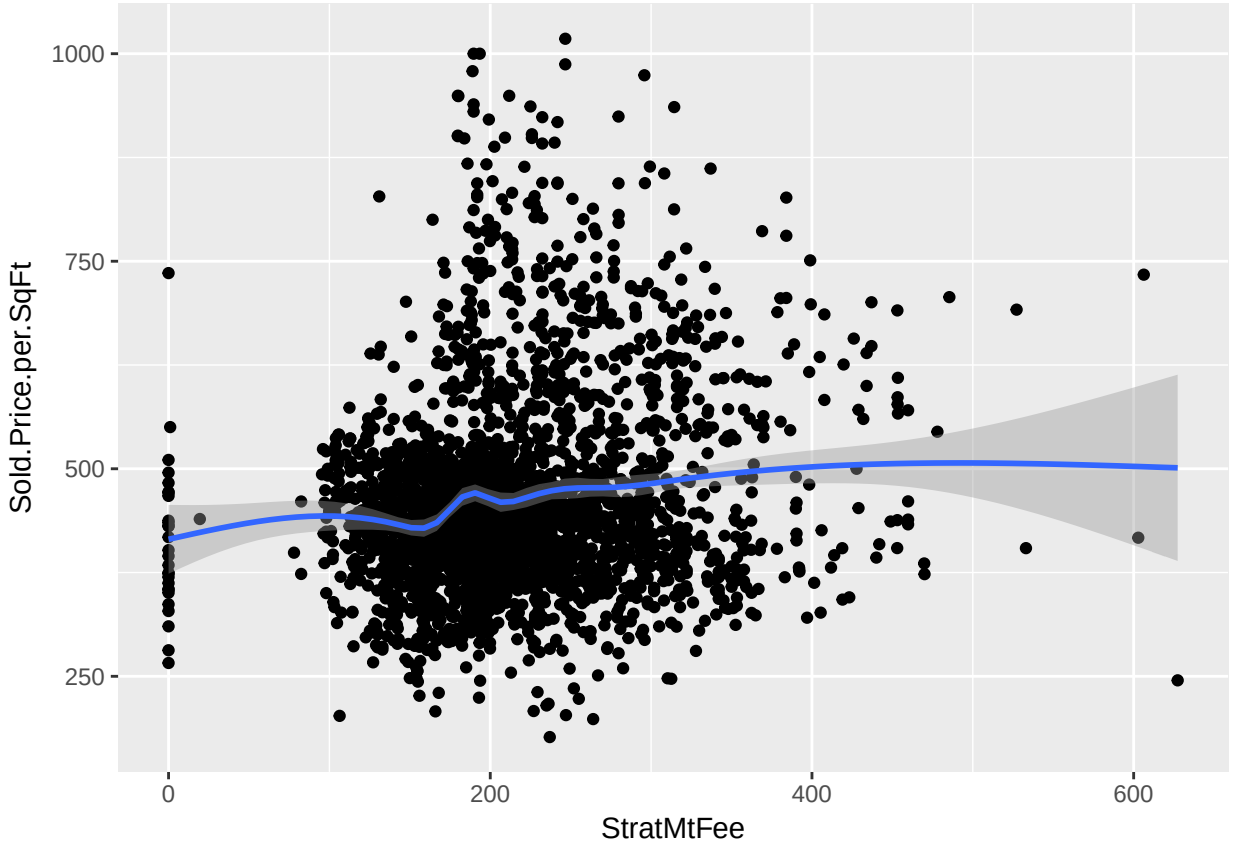
We would also like to look at how different variables can help us predict condo prices. Unlike ARIMA which is fitted to time series data, a random forest allows us to analyze our data from a different perspective. Random forest is a statistics method used to build predictive models for classification or regression models. It involves the use of ensembling the predictions of multiple decision trees. We will use it to predict monthly average price per square foot based on the relevant characteristics identified previously in our *Data Description* section.

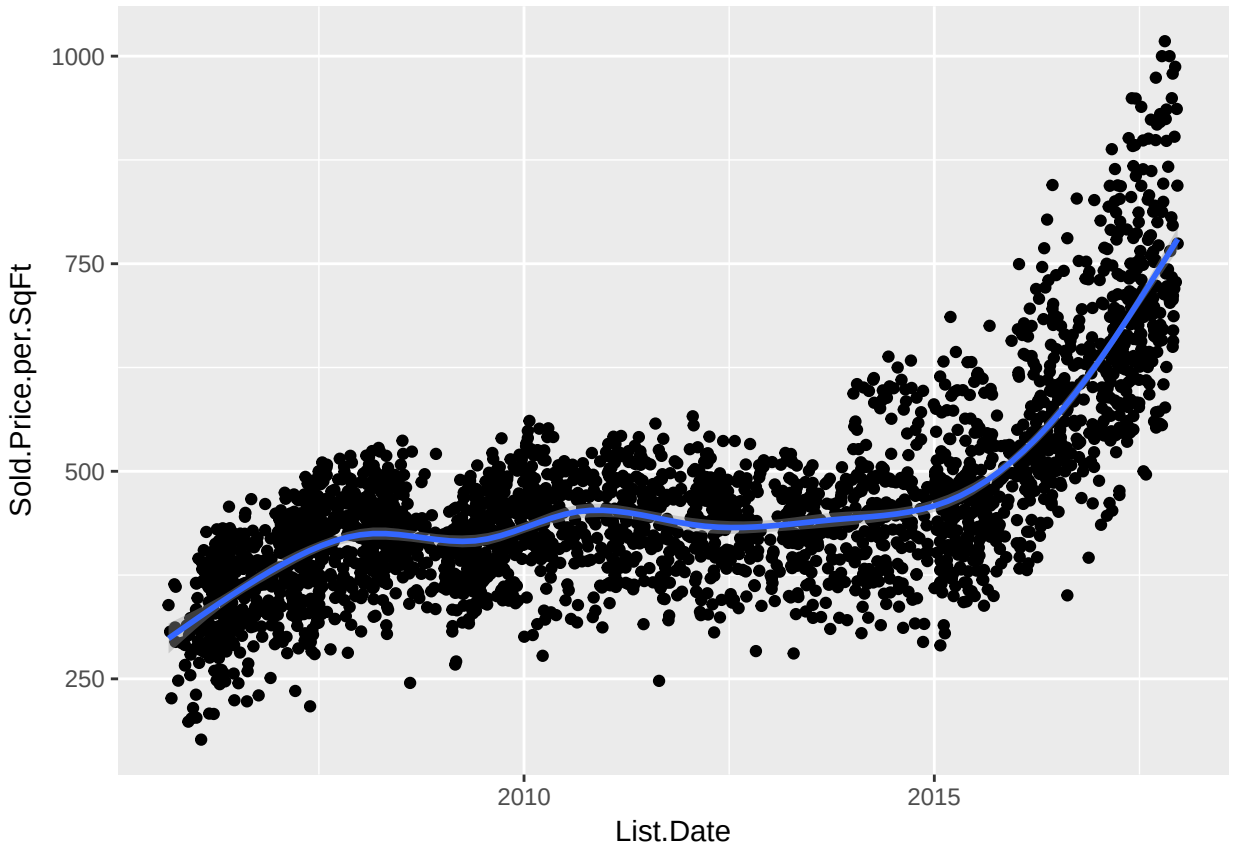
To assess the performance of our model, we will use a technique called cross-validation. Cross-validation allows us to break our data into random training and testing sets, using the training set to predict on the testing set. This will be done several times, with random training and testing sets each time, so we can get a measure of the prediction error for our random forest model.

Again unlike our ARIMA prediction, instead of predicting an average market price for condos, our random forest model requires a specific condo with all of the relevant charateristics listed in our variable list (see results section below) in order to predict the price of that specific condo.

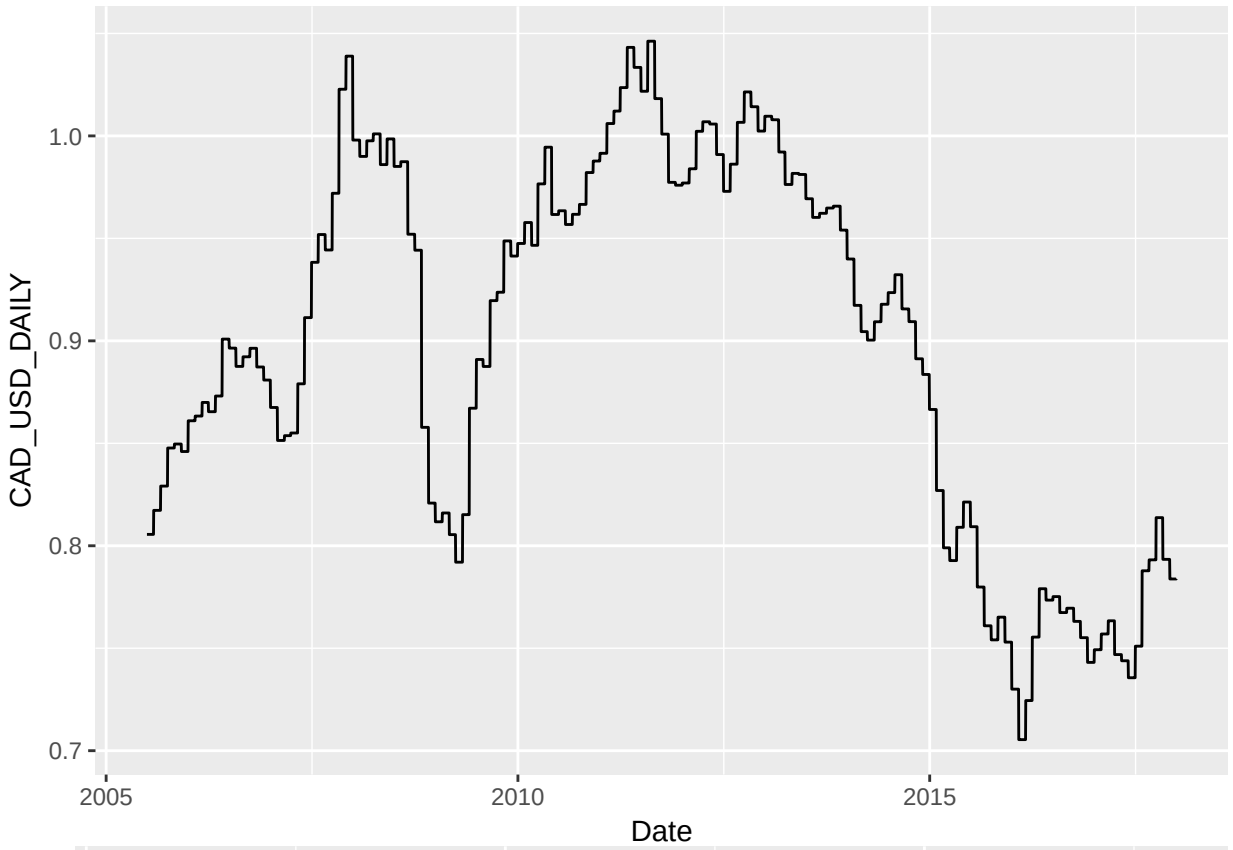
Results: Exploratory Analysis

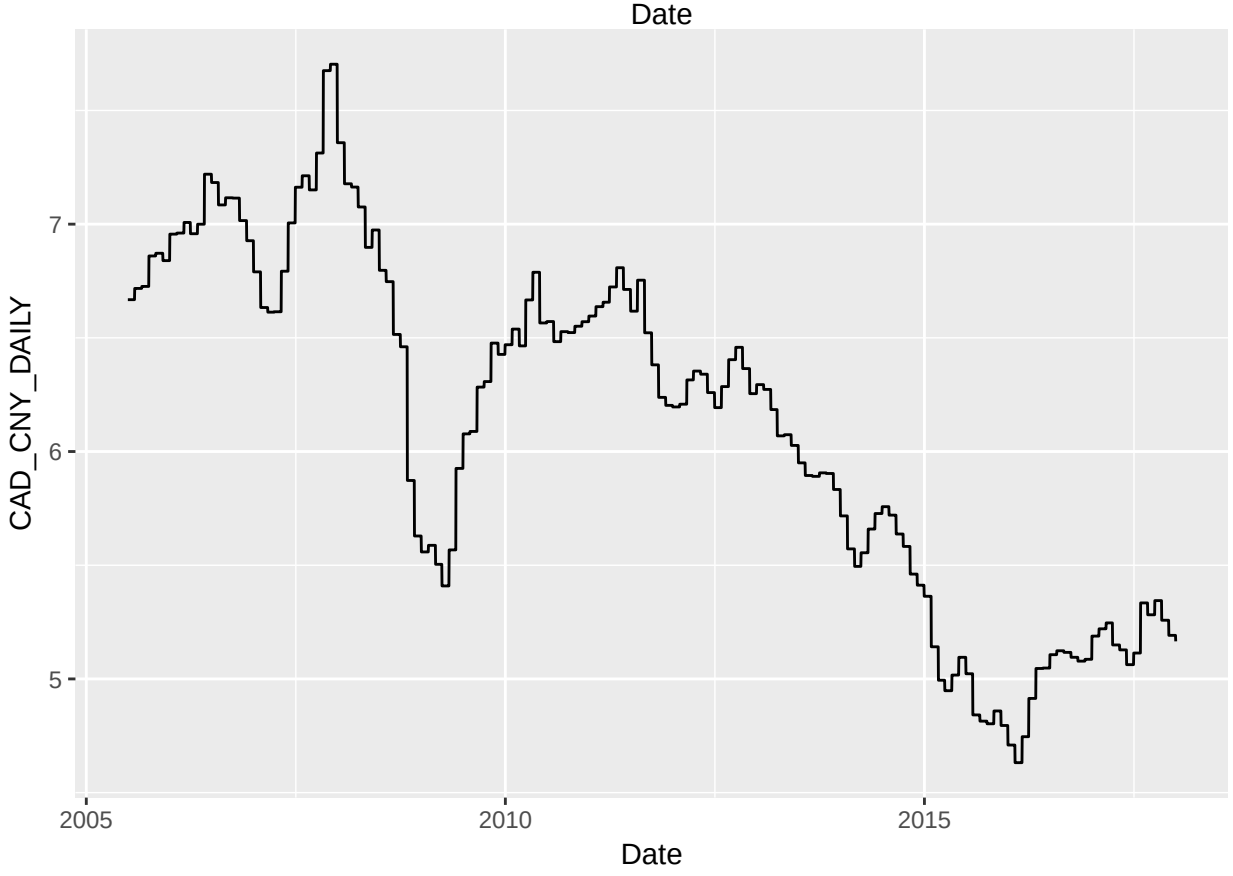


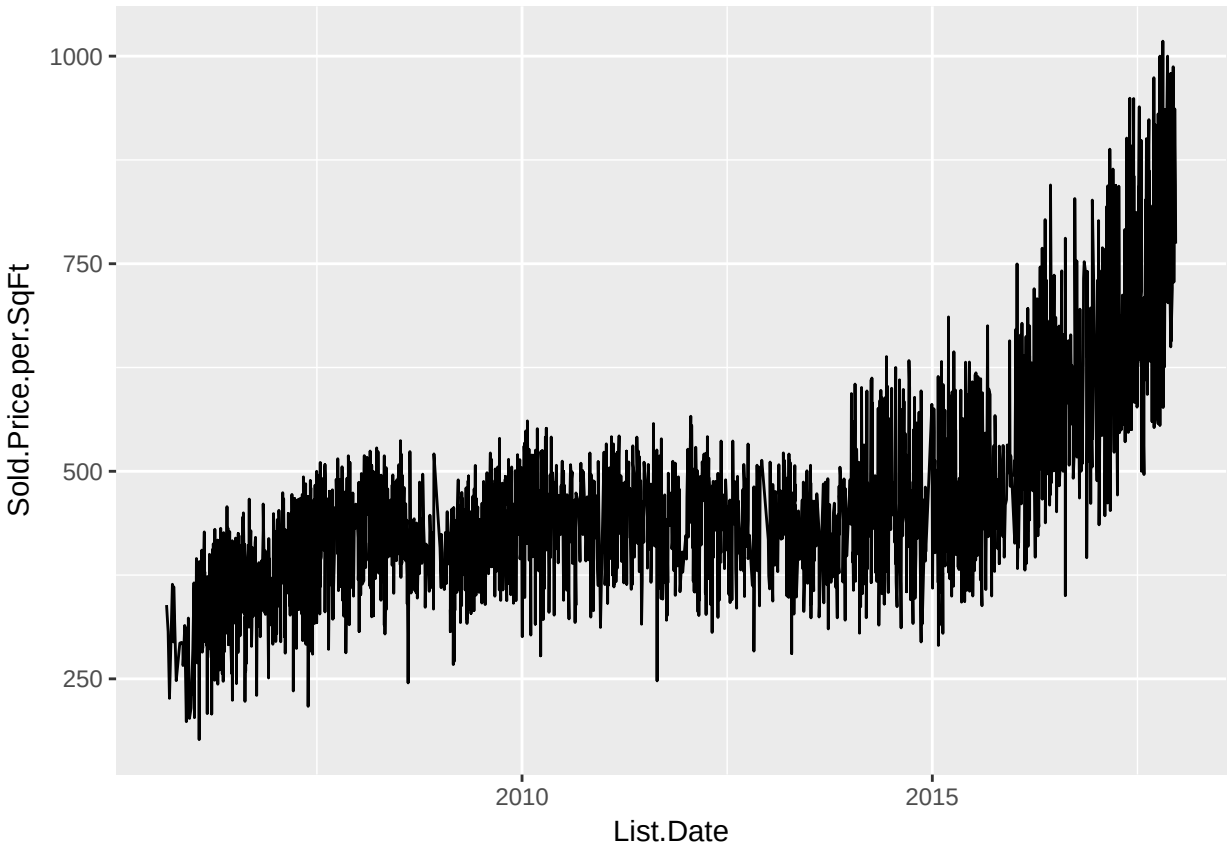




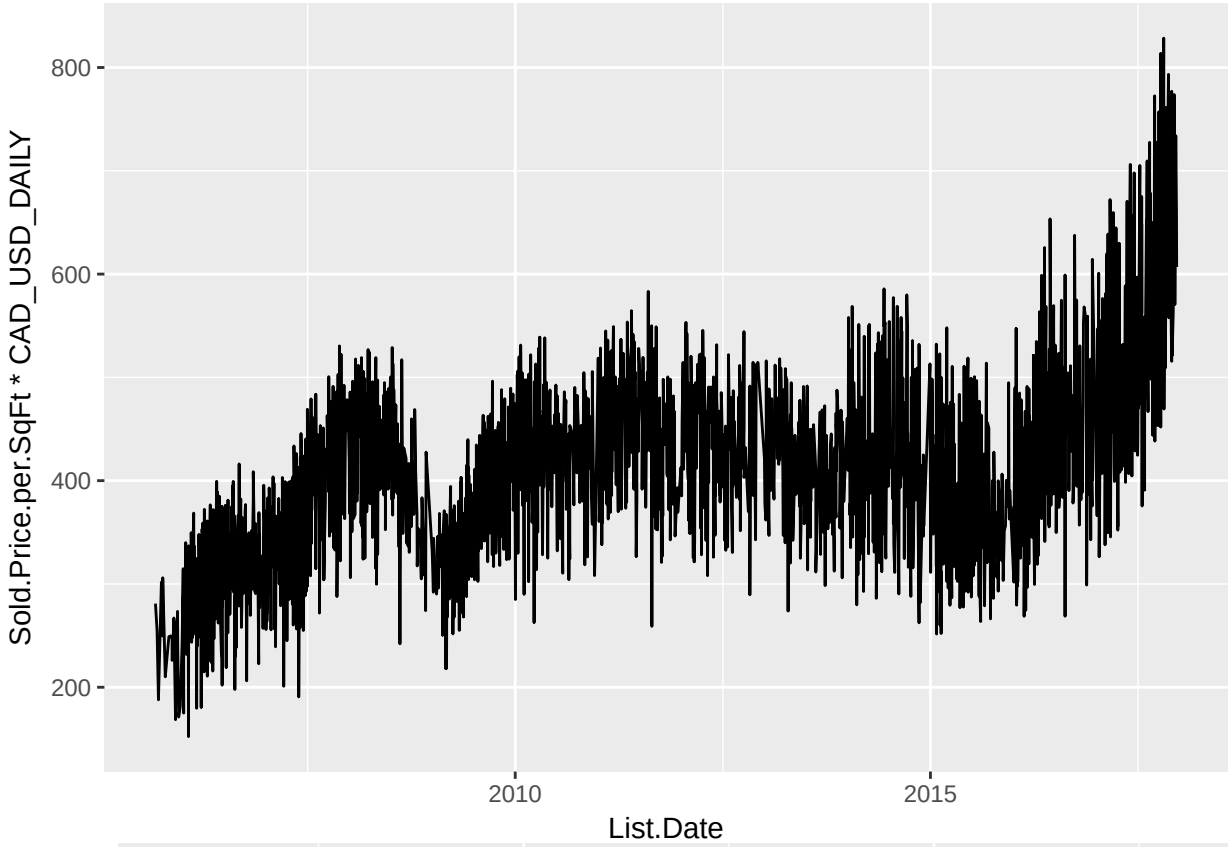
We loaded two history records of currency exchange rate, Canadian Dollar (CAD) versus US Dollar (USD), and CAD versus Chinese Yuan (CNY). We chose Chinese Yuan because Chinese citizens made up the largest buyer group in Vancouver's property market. Both of USD and CNY became stronger against CAD after 2014, due to Canadian economic recession. Meanwhile, we noticed an opposite trend in Vancouver's housing market. The housing price was stable before 2014, since then, it went up rapidly.

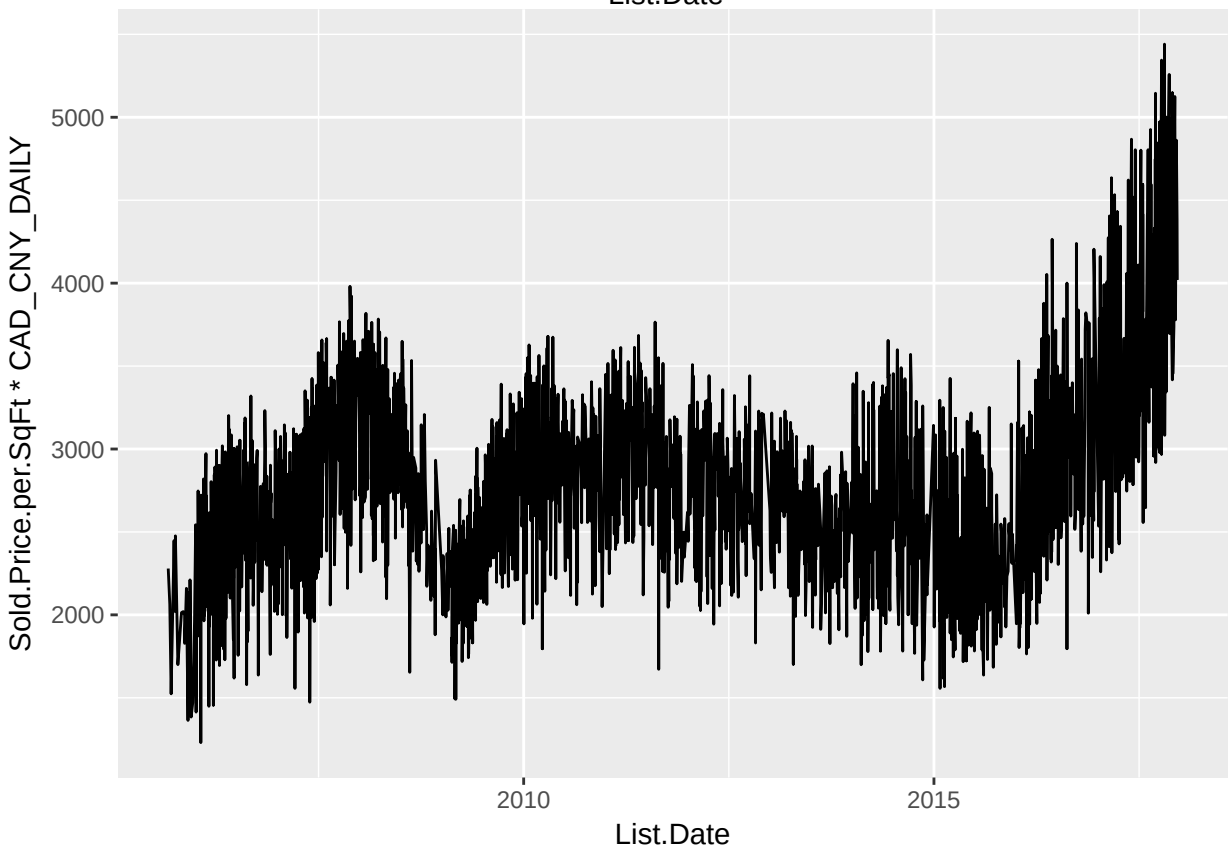

CAD_USD_DAILY
1.0
0.9
0.8
0.7
2005
2010
2015
Date


CAD_CNY_DAILY
7
6
5
2005
2010
2015
Date



We then plotted the product of condo price and exchange rates, to see if the change in Foreign Exchange Market was the factor that caused the growth of Vancouver's real estate market. A fairly straight line (note: add a geom_smooth line later) could be expected if the assumption was true. However, the upward trend was still obvious for both plots. Clearly, there were other factors contributing to the rapid increase of housing price.





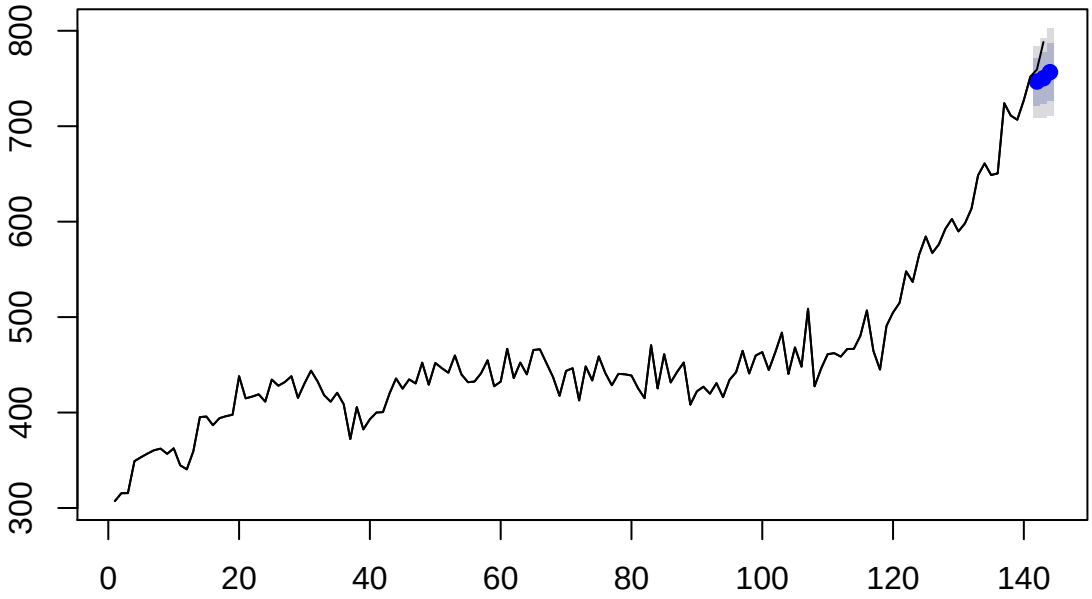
Due to the poor housing affordability, the City of Vancouver have come up with some policies trying to

address this issue. During the past two years, they have implemented "First Time Home Buyers' Program", "Empty Homes Tax", "15% Foreign Buyer Tax". The client was interested to see how these policies affected the condo market and if the foreign buyer tax would continue affecting the market as the tax rate would increase five more percent this year (note: to-do prediction). We also included "BC Investment Immigration Program" and the 2008 Financing Crisis since it was believed that they had influences on Vancouver's real estate market.

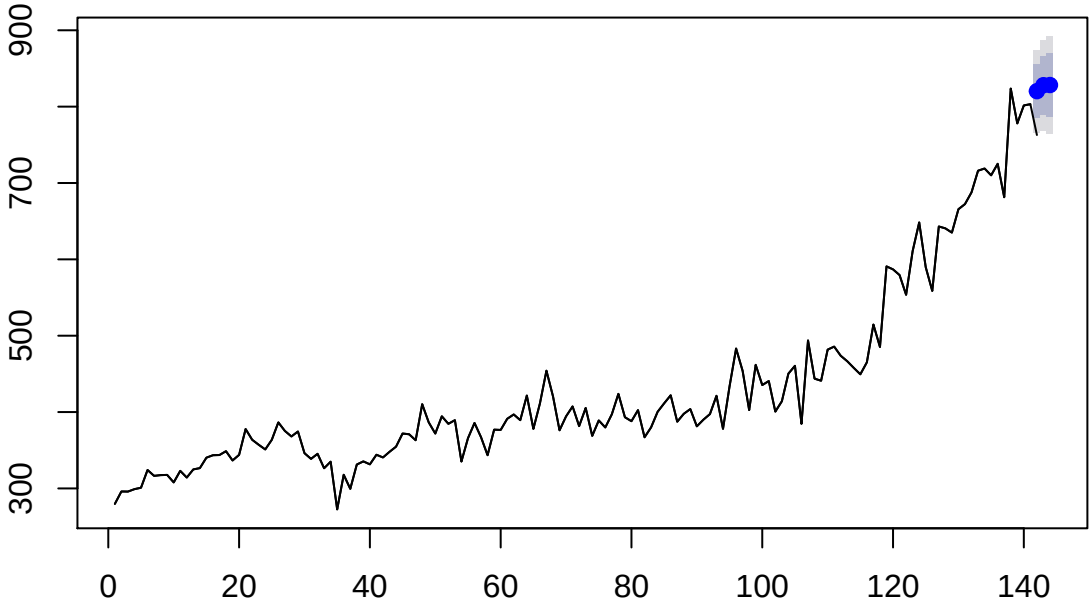
Further analysis needed here...

Results: ARIMA

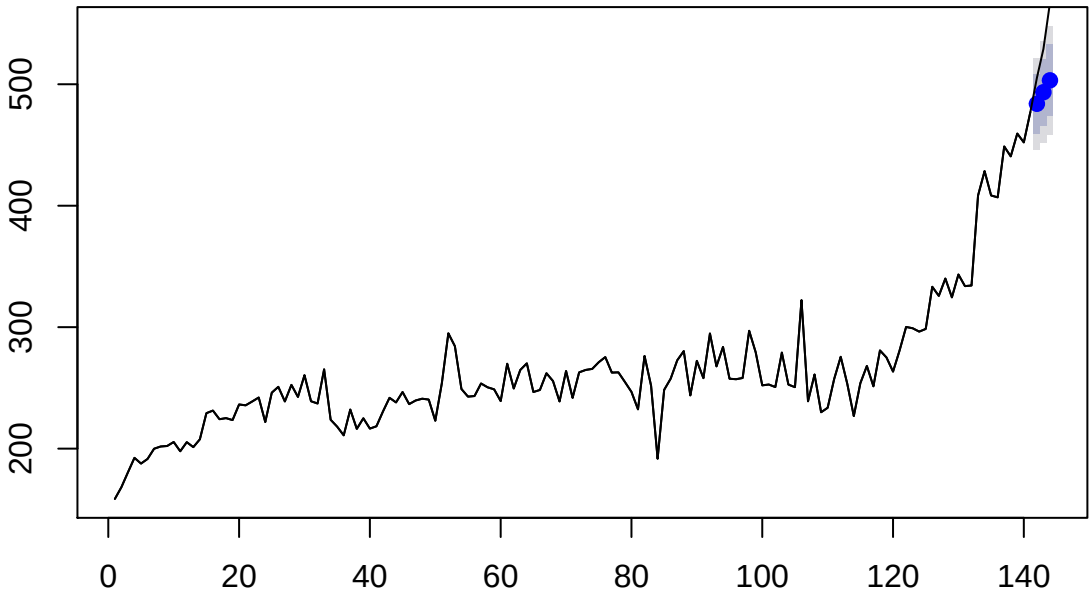
Collingwood



Metrotown



Whalley



For Collingwood and Whalley, our fitted ARIMA models were able to predict prices within the 95% and 80% confidence bounds. For Metrotown, however, the fitted ARIMA model was not able to make accurate predictions. We believe that inaccuracies in these ARIMA predictions can be due to new events that cannot be captured in our model, such as a new government policy or a change in the economy.

Results: Random Forest

We supplied variables that could be related to condo prices in order to build a random forest model. We excluded variables that we think are irrelevant for prediction, such as agent name and MLS number. The remaining variables that were considered in the model are:

- Days on Market
- Previous Status
- Strata Maintenance Fee
- Number of Bedrooms
- Number of Bathrooms
- Age
- Locker
- Number of Parking Spaces
- Previous Sold Price / Status
- Sold Year
- Sold Month
- Total Floor Area
- Bylaw Restrictions (such as pet allowance)

```
## [1] "RMSE for Collingwood: "
```

```
## [1] 0.06679834
```

```
## [1] "RMSE for Metrotown: "
```

```
## [1] 0.07812502
```

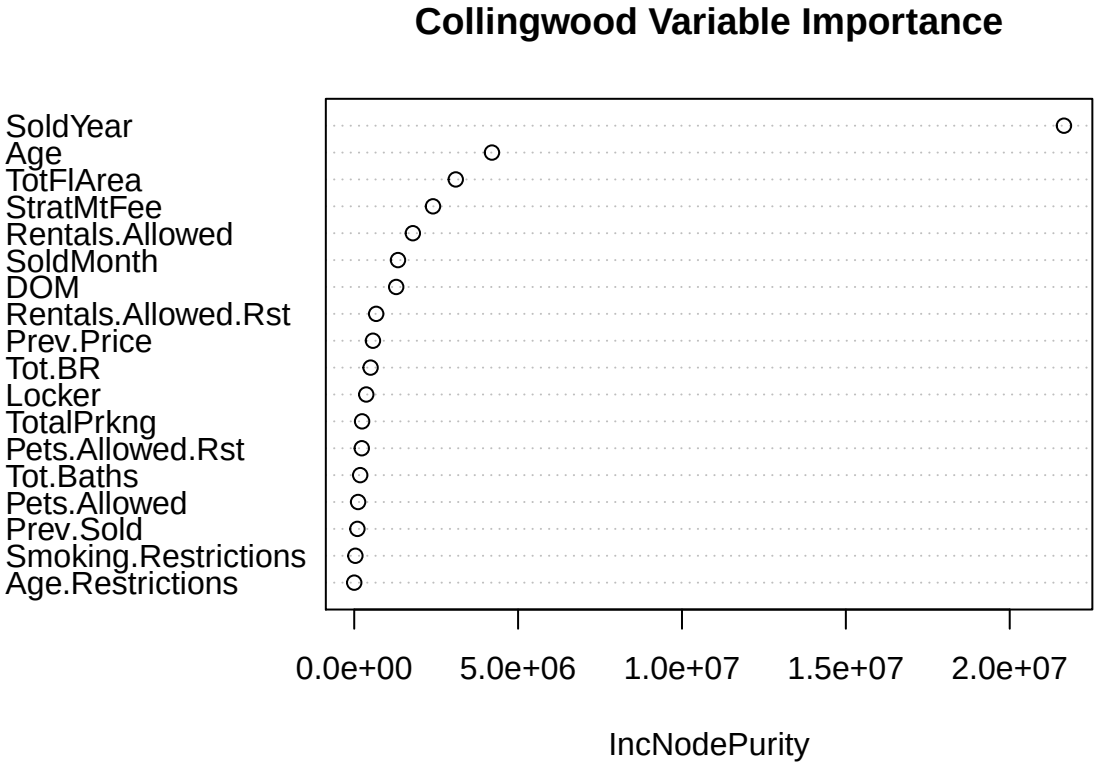
```
## [1] "RMSE for Whalley: "
```

```
## [1] 0.07740411
```

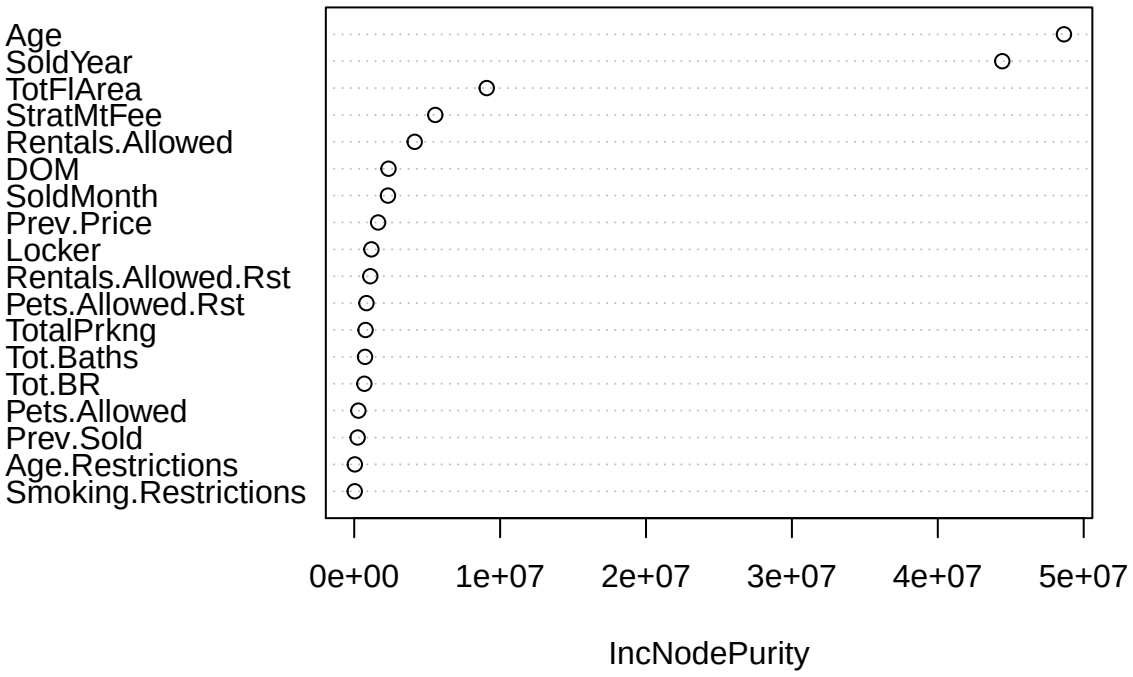
The root mean squared error is a good indicator of the relative error of our prediction. Using random forest with 100 trees, along with cross validation, we have found that our root-mean-squared prediction error is **6.7% for Collingwood, 7.8% for Metrotown, and 7.7% for Whalley**.

To put that into perspective, if we consider a standard 2 bed 2 bath new condo with various characteristics in Collingwood, our prediction might give a price of $750 per sqft, but that is subject to an error of 0.067*750 ~= $50. So the predicted price is anywhere between $700 and $800 per sqft. In our next report, we will attempt to minimize this prediction error by including external variables and reviewing the data processing.

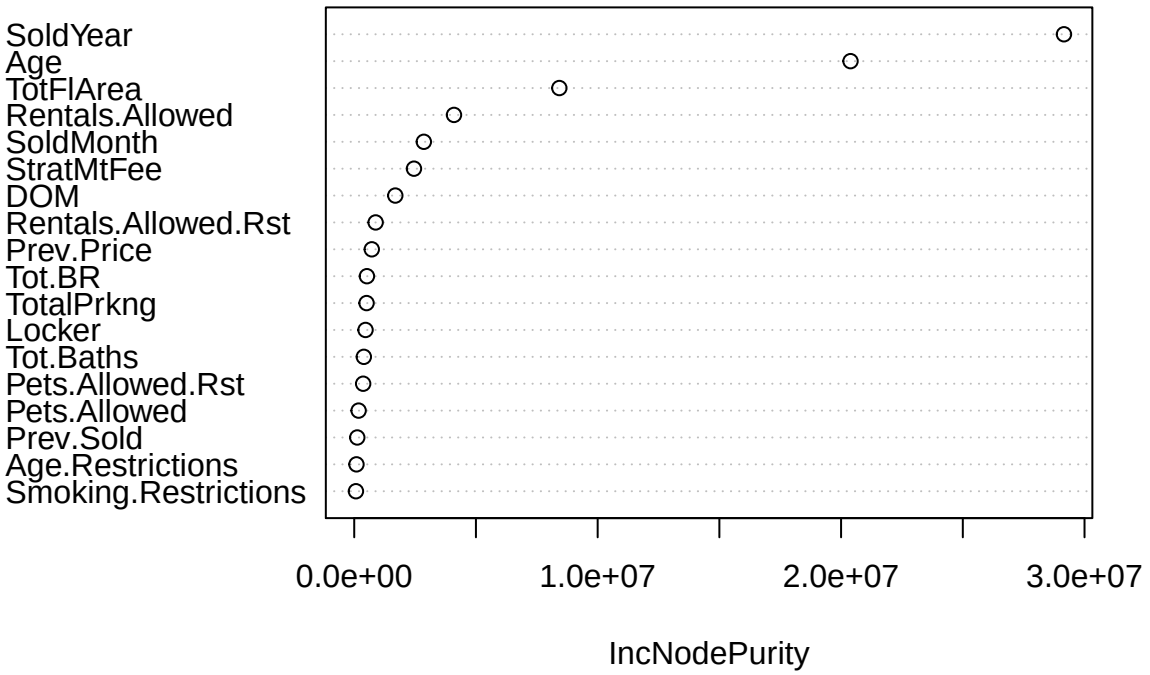
Let us take a look at the variable importance of these random forest models by district:




```
## [1] "INSERT INDENT"
```




```
## [1] "INSERT INDENT"
```



We can see that the sold year comes out as one of the most impactful variable in the random forest model for all districts. This is expected, since as we know, all condo prices have increased rapidly over the past few years. However, notice that some of the other variables that stand out are age, total floor area, strata fee, and in particular, whether rentals are allowed. We expected some of the top variables to show up are

floor area, age, but we did not expect the restriction on rental to be near the top of the list of top variables. This can certainly be a variable to consider when predicting the price of a potential condo. In many of these cases, total bedroom and bathrooms did not appear near the top of the list. We believe this is due to their overlapping similarities with total floor area, because condos with greater floor area tend to have more bedrooms and bathrooms.

Conclusions

Among all the external factors we analyzed, we have seen that the currency exchange rate, BC Investment Immigration Program and the 15% Foreign Buyer Tax had the most powerful impact on the real estate market. Surprisingly, the Foreign Buyer Tax boosted the condo market for all three sub-regions, which was the opposite of what it supposed to do. On the other hand, "First Time Home Buyers' Program" and "Empty Homes Tax" did not seem to improve the housing affordability. We would like to perform further tests to see it their influences were statistically significant.

The fitted ARIMA model was able to make relatively accurate predicitions for Collingwood and Whalley, but it did not perform well for Metrotown. One of the limitations of ARIMA is that if a new event that has not been observed in the past occurs, ARIMA can fail to capture it and make a poor prediction. We believe that this was the case for predicting the price for Metrotown. We will perform further analysis to check this. In addition, we intend to include additional external variables, such as government policies, when fitting the ARIMA model in order to visualize their effects on the trend.

Random forest have given us a preliminary performance with a root-mean-squared prediction error of 6.7% for Collingwood, 7.8% for Metrotown, and 7.7% for Whalley. To improve the model, we will include external variables such as interest rate and school rating. For our next report, we are also looking to see if the distance to skytrain has a relationship with the price.