# JSC370 Final Report

Aditya Khan

2024-04-30

## 1. Introduction

### 1.1. Background and Research Objective

When trying to analyse how pricing works in the rental market, the immediate factors behind listing price that immediately come to mind are related to characteristics of the property itself. The intuition would be that other factors like the characteristics of the neighbourhood the listing is in probably would not be as important. The reason being that consumers are less likely to have a strong understanding of characteristics of neighbourhoods if they are travelling. Hence in theory, renters are not required to adjust their price (to a great degree) for the characteristics of the neighbourhood the property is in. This would be in stark contrast to the housing market.

Our objective is to test this theory: to what extent does the characteristics of the neighbourhood actually impact the price of a rental listing? We choose to restrict our setting to AirBnBs and specifically those in Toronto. AirBnBs are chosen because the data is easily obtainable online. Toronto is chosen as the city, due to the fact that it has a good mix of both affluent neighbourhoods (e.g. Trinity Bellwoods), and less affluent ones.

We choose the neighbourhood characteristics we study according to the following heuristic: what neighbourhood characteristics could potentially impact prices? First, we consider the safety of the neighbourhood. Secondly we consider the median household income of the neighbourhood. Thirdly, we consider the proximity of the place to downtown. The last is important to consider, because proximity to downtown is an indicator for things like quick access to amenities (downtown has a high concentration of shops) and access to luxuries.

With that said, there is clear confounders that we need to adjust for: namely the property type. Different property types would generally fetch different prices.

In light of the above discussion, our research question is hence: to what extent is there an association between price of an AirBnB listing in Toronto, and 1) the safety of the neighbourhood the listing is in, 2) the median household income of the neighbourhood, and 3) proximity to downtown? Does this association (if it exists) differ by property type?

As we look the data further, our exploration may inform the addition of further covariates to any model we fit, to answer our research question above.

### 1.2. Data Description

The primary dataset that we use, is AirBnB data from Toronto that was scraped on February 14, 2024 by InsideAirBnB. This data is cross-sectional. Each row in this dataset is a listing, and each column is some feature about the listing. There are 20630 observations in this dataset, and 75 features. Notably, all variables that we listed in our research question are present in this dataset except for two:

- the amount of crime in a neighbourhood, and
- the median household income of the neighbourhood.

To find the amount of crime in each neighbourhood, we use data the "Major Crime Indicators (MCI) Open Data" from the Toronto Police Service. Each observation is a major crime (all since 2014). There are 372899 observations. Each column is some feature about the crime. There are 31 columns. Notably, the columns include which neighbourhood the crime was committed in. We choose to focus on MCIs (rather than petty crimes), because major crimes are the most likely to affect public perception of the safety of a neighbourhood, and hence, prices of a listing.

The median household income, is obtained from City of Toronto's Open Data. This data contains summary statistics for each neighbourhood in the city, based on the 2021 Canadian census. Each row is a relevant summary statistics (of which there are 2603) and each column is a neighbourhood (of which there are 158).

# 2. Methods

## 2.1. Wrangling Each Dataset Separately

All of the datasets were downloaded from the websites mentioned in the references sections. In the references section, [1] is the website to download the AirBnB data. Navigate to the section titled Toronto, and download "listings.csv.gz", under February 14, 2024. [3] is the website to download the MCI Open Data. [4] is the website to download the Neighbourhood Profile data. For the latter two, there is a download button on the website. The former two datasets are in CSV format. The final dataset mentioned, from [4], is in XLSX format.

Altogether, we have three datasets that need to be merged. But before we do that, we need to fix underlying issues that are intrinsic to the dataset. After that is done, we fix issues that can be identified only after merging, and compute any remaining key variables that we need. As a technical note, all wrangling and analysis done below is with the `data.table` library in `R` to maximise speed.

### 2.1.1. Wrangling the Crime Data

We first wrangled the crime data. This data contained a number of issues. In terms of missing data, the data description for this dataset mentions that geographical missing data is specified as "NSA" for neighbourhood data if either the police division for the observations is also "NSA" or the X, Y coordinates are 0 or outside of Toronto. Taking these conditions into account, we replace occurrences of "NSA" and unrealistic longitude, latitude, X, and Y data (those that are clearly outside Toronto) with NA.

Another notable issue, was that the occurrence dates for crimes in this dataset (given by "OCC_DATE") were not of type datetime (but rather strings). Furthermore, these dates were incorrect, since for all crimes, the time occurred is listed as occurring at 5 am. Hence, we recalculate the date correctly as a datetime variable.

Finally, we recall that the point of this dataset was to get an understanding of how safe a neighborhood is. Duly, we create a new variable "crime_count" which is defined as the amount of crimes that were committed per day in a given neighbourhood, within a year of February 14, 2024; the day the AirBnB listings are from. The reason why we only interested in recent criminality, is that it would provide a more accurate understanding of how safe the neighbourhood is at the point of time the listing was given.

### 2.1.2. Wrangling the Income Data

Next we wrangle the neighbourhood profile data from Open Data from City of Toronto. In this data, we are only interested in obtaining the median household income for each neighbourhood, so we only clean this

particular subset of the data. There was no missing data, and all of the data appeared to be realistic. The first issue though, was that the data had neighbourhoods as the columns in the dataset, and not under just one column. Since this would be a problem when we merge with the other datasets on a *single* neighbourhood column, we transposed the data such that all of the neighbourhoods would be under one column, and the corresponding income value would be in a second column. The second issue was that the median household incomes were strings, and not numerics. Hence, we converted them to the correct type.

#### 2.1.3. Wrangling the AirBnB Data

Finally, we wrangled the AirBnB data. In the AirBnB data, a number of observations under specific columns were empty or put as missing under the string "N/A", not actually formally given the NA variable. Hence, we made that change. Another issue, was that prices were of type character with a dollar sign pre-pended to the string. We hence fixed that problem by removing the dollar sign, and converting the price to a numeric. Finally, a number of numerical variables like the minimum nights a renter had to stay, were represented by strings, rather than integers. So we converged those to integers.

In terms of unrealistic data, one accommodation set its rent price to just $1. This observation also required the renter to say over 1000 nights. This is rather unrealistic, so we removed this observation.

Finally, we consider two further changes we need to make to the variables to ensure easy interpretability and agreement with the other datasets.

Firstly, recall that we consider property type as a variable we need to adjust for. The problem with the property type as it stands, is that it is a categorical variable with 58 levels. This is bad for interpretability. Hence, we combine property types under 13 few broad categories.These categories were chosen by analysing the prefixes of each category (e.g. 13 of the original categories had "entire" followed by some house type like villa, so these are classified under "entire property").

The second consideration was the neighbourhoods that were listed in the observations for this data. Toronto underwent a change in the scheme it used for defining neighbourhoods in 2021, going from 140 to 158 neighbourhoods. To align with the current neighbourhood scheme, we choose to do our analysis using the 158 neighbourhoods; duly the other datasets follow that scheme. However, a few observations in the dataset still used the 140 scheme. Due to the difficulty of mapping exactly which neighbourhood those observations would lie in under the current scheme, we chose to discard those observations.

### 2.2. Merging and Finalising the Data

#### 2.2.1. Computing Some Remaining Variables

After we wrangled and cleaned the data separately, we inner joined all of the data on their neighbourhood columns. After merging, there was one final variable that remained to be computed. We had to compute the proximity to downtown for each property - this is a key predictor in our research question. To compute the distance, we first note that the latitude and longitude for downtown Toronto is 43.6515 and -79.3835 respectively (see [2]). So to compute how many kilometres away a given property is, we compute the Haversine distance between the property's coordinates, and downtown's.

#### 2.2.2. Finalising Data Cleaning and Wrangling

The first thing we note in the final cleaning step, is that the range for price is 13-999, whereas it is 57200-222000 for median household income. Since the median household incomes are a couple of order of magnitudes higher, we choose to scale down the median household incomes by dividing by 100, for the purpose of interpretability; this is since we eventually intend to fit a regression model with price as the response.

Next, we note that the only levels in the categorical variable that represents property type that have more than 21 observations are: "Shared room", "Private room", and "Entire property". Since all of the rest of the levels have miniscule amounts of data, we remove those observations, to minimise extreme class imbalance.

At the end of the cleaning process, we end up with eleven columns, and 8982 observations. We specify the variables we will be using in the rest of the analysis and what they represent below.

| Variable | Information | |
|---|---|---|
| **Variable** | **Meaning** | **Role** |
| **nbd** | Neighborhood | Categorical |
| **nbd_desc** | Neighbourhood Desc. Given by Host | Text Data |
| **lat** | Latitude | Numerical Predictor |
| **lon** | Longitude | Numerical Predictor |
| **property_type** | Property Type | Categorical Confounder |
| **price** | Price of Property | Numerical Response |
| **nbd.crime_count** | Crimes in Neighborhood (/day) in Past Year | Numerical Predictor |
| **distance_from_downtown** | Distance from Downtown (km) | Numerical Predictor |
| **nbd.med_hh_scaled** | Scaled Med. Household Income in Neighborhood (/$100) | Numerical Response |

**Table 1: Variable Descriptions**

## 2.3. Methods for Data Exploration of Variables

Now we explain how we initially analysed the variables that we used our exploration.

Our analysis will start by analysing summaries about each of the numerical and categorical variables we are interested in. Once we have that introductory understanding of the properties of the variables, we begin to explore relationships between the variables.

The first thing we do, is some text analysis. In our data, we have access to descriptions that the host of the property wrote, to describe it: `nbd.desc` in our nomenclature. So to get an introductory understanding of whether neighbourhoods impact price, by seeing whether the top tokens in the descriptions differ between different listing price levels.

That only provides us with a broad idea of whether we can expect any stark differences between price levels, at least from the perspective of the hosts. So we turn our attention to analysing how the distribution of price changes (or doesn't change) with the predictors. Namely, we plot a stacked histogram of price, with respect to the property type. This gives us an indication of how much of an effect our confounder has.

We then plot boxplots of the price distribution, over different levels of median household income and crime counts (which are two of our predictors of interest) to see if it seems to have any notable impact on price. To solidify our understanding of whether it does, we follow that visualisation up, with a scatterplot involving crime counts, income, and price. Our final visualisation is a map of each neighbourhood in Toronto, coloured by the price. By marking downtown on the map, we get a rough idea of whether proximity to downtown actually matters.

After getting a grasp of the relationships of interest from the visualisations, we formally test the relationships by building a number of models, both for the purpose of examining the relationship price has with the predictors in our research question, and for the the purpose of prediction. The former directly answers our research question. The latter indirectly answers our research question (if a model is predictive, we can expect some broad association between the covariates and the response).

We build a number of different models to answer our question

- A mean model with price as the response and neighbourhood as the "treatment". The rationale for building this separately is mentioned in section 3.
- A linear regression model to directly answer our research question.
- A linear mixed model to account for the neighbourhood level covariates (e.g. neighbourhood household income) in our data.
- A regression tree to predict price.
- Bagging Random Forest, Boosting models to predict price.

Throughout the model building process, we highlight important insights we obtain about the relationship between the covariates and the response. At the end, we also choose the best model, primarily on the basis of comparing RMSE.

# 3. Preliminary Results

## 3.1. Summary Statistics

| | price | distance_from_downtown | nbd.crime_count | nbd.med_hh_scaled |
|---|---|---|---|---|
| **# of Values** | 8982.00 | 8982.00 | 8982.00 | 8982.00 |
| **# Null** | 0.00 | 0.00 | 0.00 | 0.00 |
| **# Missing** | 0.00 | 0.00 | 0.00 | 0.00 |
| **Min** | 13.00 | 0.41 | 0.15 | 572.00 |
| **Max** | 999.00 | 27.63 | 3.44 | 2220.00 |
| **Range** | 986.00 | 27.22 | 3.29 | 1648.00 |
| **Sum** | 1228045.00 | 70779.16 | 8151.45 | 7861873.00 |
| **Median** | 100.00 | 6.10 | 0.64 | 850.00 |
| **Mean** | 136.72 | 7.88 | 0.91 | 875.29 |
| **SE of Mean** | 1.23 | 0.06 | 0.01 | 2.01 |
| **95% CI of Mean** | 2.41 | 0.12 | 0.01 | 3.94 |
| **Variance** | 13543.97 | 33.70 | 0.37 | 36343.89 |
| **SD** | 116.38 | 5.81 | 0.61 | 190.64 |
| **Coef. of Variation** | 0.85 | 0.74 | 0.67 | 0.22 |

**Table 2: Summary Statistics For Numerical Variables**

In table 2, we see that none of the variables have any missing values. For all of the variables, the mean is larger than the median. As well, the mean for all variables appears to be closer to the min value, than the max. Both of these facts indicate that all of the distributions are to some degree, right-skewed. In fact, we can separately compute the number of outliers for the relevant variables as follows:

- Price has 628 outliers
- Crimes per day in neighbourhood in past year has 272 outliers
- Scaled neighbourhood household income has 454 outliers
- Distance from downtown has 198 outliers

The reason why we did not remove outliers, is that the outliers make sense for all of the variables. Most properties that are listed are likely normal houses or apartments, so the right skew occurs due to the occasional lister who puts up a property for rent that is very upscale. A similar rationale explains why household income is right skewed. As for distance to downtown, the right skew also makes sense. One would expect that most properties in Toronto proper, would be centred around Downtown, so for most places, the

distance would be small, and hence, a right skew. For crimes per day, most places would be relatively safe, except for a handful of neighbourhoods.

Now we look at the one categorical variable represented in our research question.

| | Type | Count | Proportion |
|---|---|---|---|
| 5 | Entire property | 5101 | 0.5679136 |
| 6 | Private room | 3783 | 0.4211757 |
| 7 | Shared room | 98 | 0.0109107 |

**Table 3: Summary Statistics For Property Type**

We see that about 56% of the data properties allow for the rent of the entire property, 42% allow for a private room, and just 1% allow for a shared room. This makes sense, since most consumers would prefer living with some degree of privacy. So, listings that allow for that are likely to be more common, just to align with demand.

## 3.2. Visualisations

Now we do visualisations. We start by analysing the neighbourhood descriptions that hosts write on the listing page, detailing qualities of the neighbourhood. Particularly, we want to see if the words they use differ by price level. Here, we defining a price to be "low" (if between 0-quartile 1), "medium" (if between quartile 1-3), or "high" (above quartile 3).


**Figure 1: Word Clouds of Neighbourhood Descriptions, by Price Level**

We see in Figure 1 that the words used are almost the same across all three price levels. Namely, there is an emphasis on walkability, restaurants and shops, and being close to parks. On one hand, this may just be telling us that hosts like to emphasise those facts irrespective of price levels since they know they will appeal to the consumer.

On the other hand, it is notable that all of the things that are emphasised in the word clouds are things that characterise downtown: namely quick access to amenities and easy walkability. If we interpret these words as what hosts think will appeal to a customer, clearly that means the customer demands these things. And if they're found in downtown, then we would expect demand for AirBnBs near downtown to be higher - and hence, raise the cost of the listing. So this can suggest in a weak sense that proximity to downtown may have some influence on price (in fact, this is verified explicitly later in the map visualisations on the website).

One thing that this text analysis does show, is that specific amenities could potentially be predictive of

price. Namely, restaurants, shops and parks. These feature prominently in the wordclouds. Hence, we create binary variables `near.rest`, `near.shop`, `near.park` for those amenities respectively, evaluating to 1 if the term shows up in the neighbourhood description and 0 otherwise.

Beyond what we did above, we can also do a more granular analysis, by additionally seeing if the word clouds change, as we consider only listings of a certain property type. In our case, subsetting our data only for listings that yield the "entire property" or a "private room" yield nearly identical wordclouds to that in figure 1.


**Figure 2: Word Clouds of Neighbourhood Descriptions for Shared Room Listings, by Price Level**

The wordcloud for shared rooms listings differs quite a bit from figure 1 though. This can be for two reasons. The first reason could be that there are not many observations for shared room listings, so perhaps the

wordclouds are not representative. Alternatively, the wordclouds to capture some difference in neighbourhood description, compared to the other property types. In that case, it justifies our inclusion of property type as a confounder in predicting price.

In a similar way, we can also subset for listings in neighbourhoods of a certain crime or income level, to check if the neighbourhood descriptions differ. To do this, we define crime levels and income levels in a similar way to how we defined it for categorical price levels.

However after doing that, we find that there is no substantial difference in wordclouds across the different income and crime levels. This tells us that, at least from the perspective of the listers, the crime or income level does not really impact the neighbourhood description across price levels.

The next visualisation we observe is the stacked histogram of price, by property type. The idea is that we want to verify our previous findings: that is, if price changes by property.

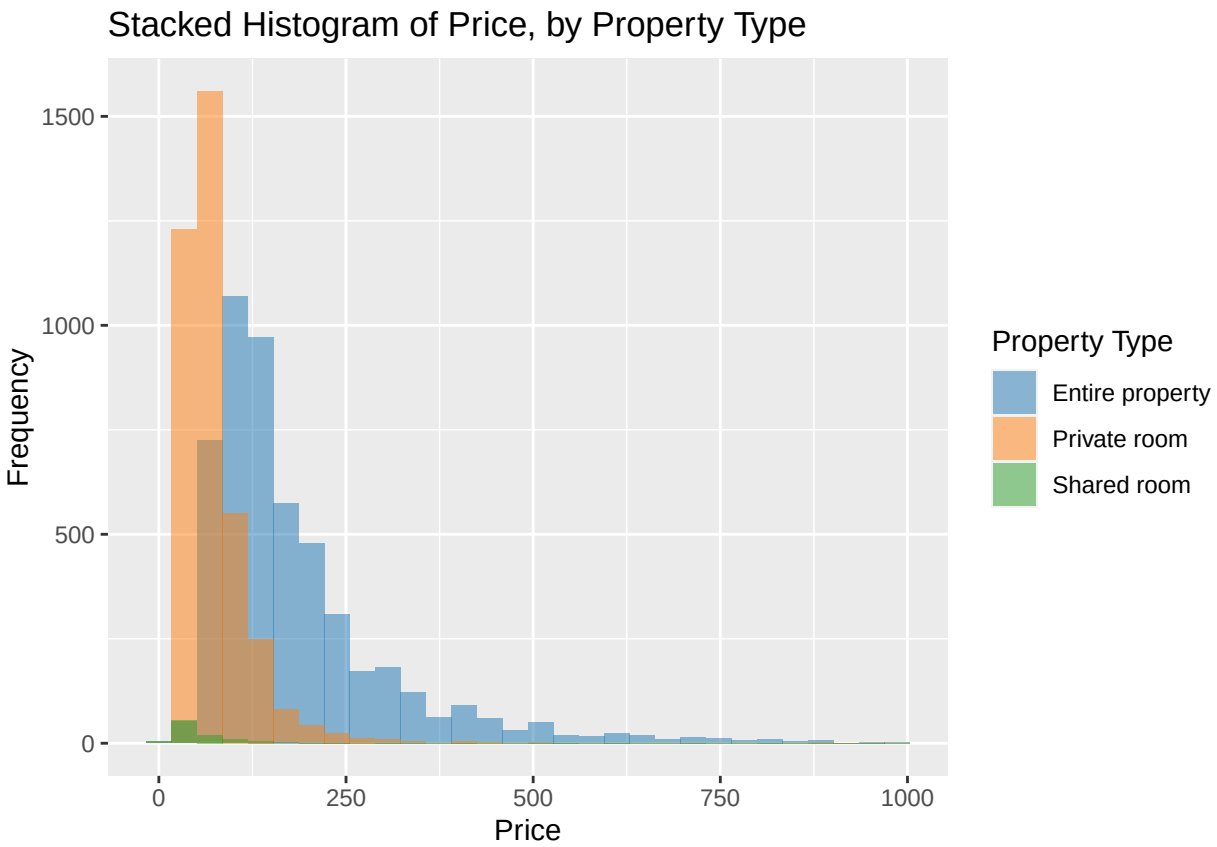


**Figure 3: Stacked Histogram of Price, by Property Type**

As we can see in the above distribution, it does to some degree. The distribution for a private room's price is must narrower and clustered around lower prices, in contrast to entire property. This makes sense, since renting an entire property would make sense to cost a lot more. Most importantly, it justifies our choice to adjust for property type, because as we can see, it does have an impact on our desired response.

Next we verify whether the price distribution changes according to different levels of scaled median household income, and crime. This can help us inform some of the lack of differences we saw in the wordclouds, when subsetting for different income levels, and crime levels.

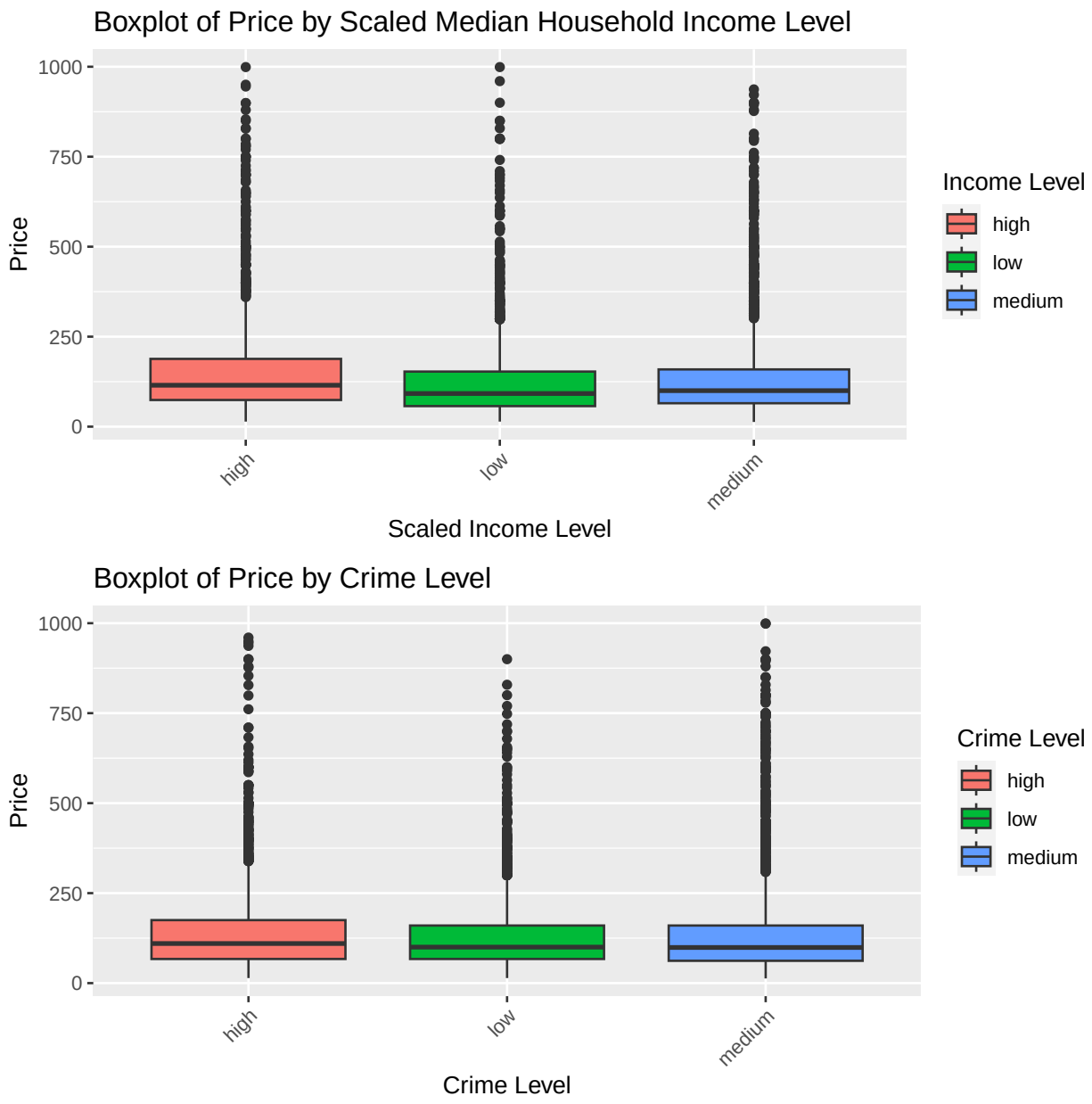


**Figure 4: Boxplot of Price, by Scaled Median Household Income Level and Crime Level**

We note here that the levels for scaled median household income and crime are defined in exactly the same way as they were for price level.

Analysing the boxplots, there appear to be a substantial amount of outliers for price. However, the actual distribution seems not to really differ between either crime level, or scaled household income. What this suggests, is that even if there is a significant relationship between these variables and price, the actual rate at which price changes with respect to these variables probably will not be that much.

To verify that claim, we now plot a scatterplot with price as the response.

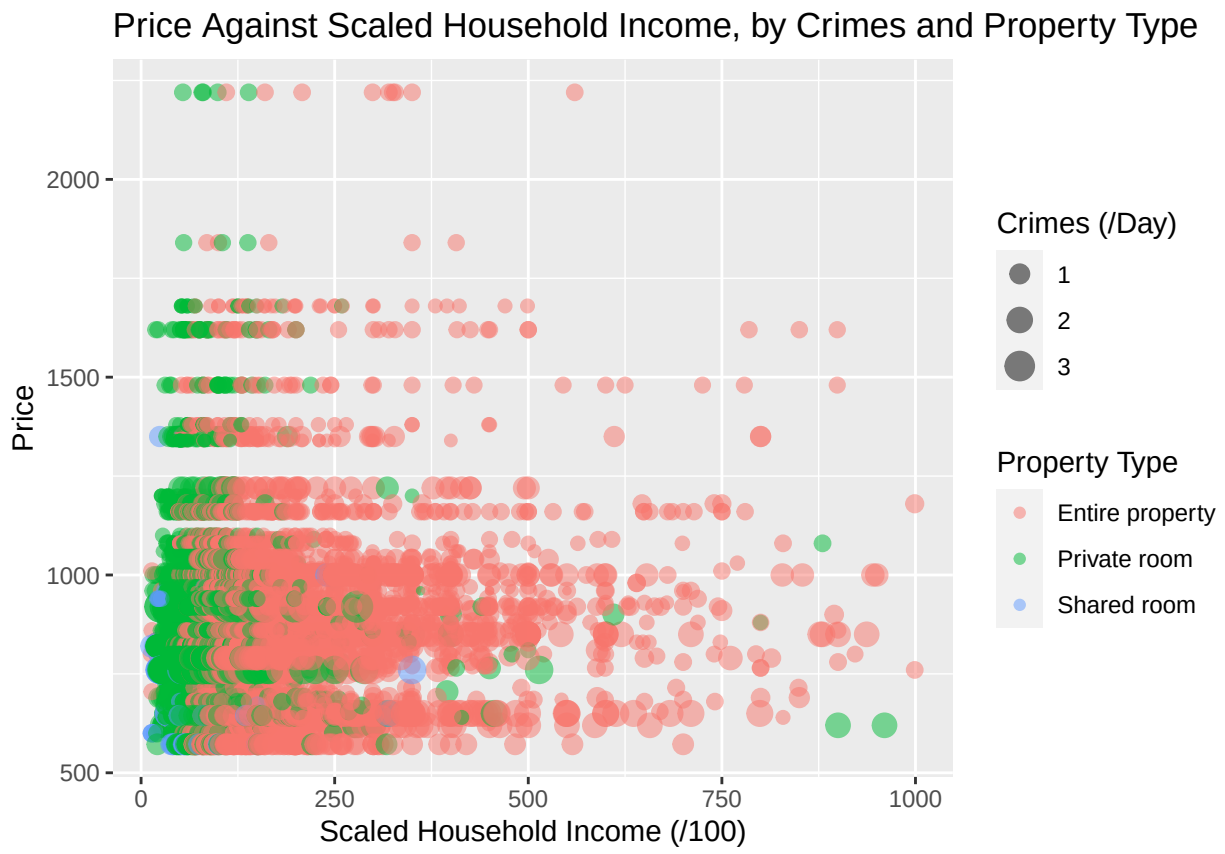


**Figure 5: Price Against Scaled Household Income, by Crimes and Property Type**

Here, we plot price against scaled household income. The size of the dots correspond to increasing crimes levels (that is, crimes committed per day). The colour corresponds to the property type.

We see that there is at best a weak positive linear association between price and scaled household income. Of course, if there was going to be any association at all, then it makes sense that it would be positive, since if the neighbourhood is affluent, the property is likely to be upscale, and hence cost more. But the fact that it seems that it is not too strong, agrees with our explanation in the introduction of this report.

In terms of crimes per day, we the majority of the large circles (which mean more crime per day) appear to be clustered around the bottom of the plot - i.e. where price is low. This would give some credence to the claim that more crime leads to lower prices - which is not unreasonable to claim since one would have likely have more crime in less affluent areas. With that said, this does not provide any conclusive evidence towards a relationship, particularly since the small circles (which represent less crimes) appear to be evenly spread across the chart.

Finally, it is worth noting that the majority of the private room listings appear to be in places where household income is lower. On the other hand, those renting out the entire property appear to be evenly spread. There seems to be minimal difference with respect to price though.

One variable that we have not yet explored yet is the distance to downtown. We can analyse this alongside the neighbourhood level variables of income level and crime, in a 3D scatterplot, coloured by price. This visualisation is found on the website associated with this project. On this visualisation, we see that in general, the majority of the more expensive listings (those of a lighter colour) are found relatively close to downtown, and in neighbourhoods with a slightly lower crime count. Interestingly, the median household income of the neighbourhood does not appear to have any large impact on price, as the proportion of relatively expensive listings (those of a lighter colour) appears to only increase by a little bit, across levels of income. This in particular, agrees with what we found in figure 4 and 5.

Our final visualisation is a map where each marker is on a neighbourhood, and written on the marker is the mean price for an AirBnB listing in that neighbourhood. The darker the colour, the higher the price. Hovering over each marker will yield the neighbourhood name. This visualisation is again found on the website.

We can see that the less costly places to get an AirBnB are on the outskirts of city. And it does appear to be true in general, that as you get closer to downtown (which is marked red on the map), the markers get darker - and hence of higher price. So it does seem to suggest that proximity to downtown does matter towards prices.

We verify all of these assertions now, by modelling.

## 3.3. Model Fitting

Throughout the fitting process, for any hypothesis test, we choose to test at a standard significance level of $\alpha = 0.05$. For any model that is not the mean model, we also use it as a predictive model. So, we fit those models on a training set, and evaluate on a test set. The test set consists of 70% of the dataset.

### 3.1.1 Mean Model

The first thing we do, is fit a simple mean model. Here, all we check is the hypothesis $H_0$ : all neighbourhood price means are the same, against $H_1$ : there is at least one neighbourhood that differs in mean.

Our rationale for doing this is twofold. The first is that it gives us a general understanding of whether neighbourhood does have an impact on price. This can be particularly enlightening, when we do further post-hoc analysis on the ANOVA results. The second reason this is useful, is that, as explained in the next subsection, we cannot actually include neighbourhood as a predictor in the linear regression model. So, this simple mean model allows us to still fit a model with neighbourhood as a predictor.

Upon fitting the model, we find that the p-value is rounded to zero. So at least, we can conclude that there is a signficant difference between at least one neighbourhood's mean AirBnB's price, against the others. Since some neighbourhoods do have different prices than others, it is reasonable to conclude that neighbourhood does matter to some degree.

Using these results, we proceed to do some post-hoc analysis, using a Tukey HSD test. This is a test that allows us to assess the significance of pairwise comparisons of mean neighbourhood prices. Namely, $H_0 : \mu_i = \mu_j$ and $H_0 : \mu_i = \mu_j$, for neighbourhood $i \neq j$.

Upon doing the Tukey test, we find that only around 6% of the pairwise comparisons are actually significant. Furthermore, among signficant comparisons, the absolute difference in mean price is on average, 98.61. On the other hand, among insignficant comparisons, the absolute difference in mean price is on average, 35.82.

This tells a couple of things. Firstly, most neighbourhoods are actually not that different from each other, when it comes to price. However, for the neighbourhoods that are significantly in price the difference in mean neighbourhood AirBnB price on average, appears to be quite a lot.

One final piece of information we can garner from the post-hoc analysis, is a rough measure of how different a given neighbourhood is, compared to any other neighbourhood in Toronto. We can do this, by counting the number of significant comparisons a neighbourhood had in the Tukey Test.

We display the results of this computation as a map, posted on the website. We find that the neighbourhood "most different" from other neighbourhoods is Forest Hill South. It had 53 signficant comparisons in the Tukey Test. An interesting observation is that the majority of the "most different" neighbourhoods are close to downtown. To there appears to be some impact of proximity to downtown, on the price of a listing, at least on a neighbourhood level.

### 3.3.2. Linear Regression Model

We build the linear regression model in the following way. Firstly, we fit a full model with every predictor mentioned in table 1, including the three new binary variables we defined when analysing the wordclouds. At this point, we address any issues with the fit, and refit the model. Then, we use stepwise variable selection (with AIC as the criterion) to select variables. Finally, we do some quick comparisons of the full model with the reduced model, to choose a finalised linear regression model.

Now we proceed with building the model. After fitting the initial model, we find that the design matrix was singular. Some further experimentation showed that the reason, was the inclusion of neighbourhood as a predictor. The reason this could have caused issues is either because 1) some neighbourhoods had very small number of listings, or 2) amongst neighbourhoods, the difference in price was in general quite small. This could cause instability in the numerical estimates for the coefficients.

Either way, this meant that we had to remove neighbourhood as a predictor. Upon refitting the full model without neighbourhood, we find that the fit no longer had any singularities. Furthermore, the variance inflation factor of each predictor was close to 1, so multicollinearity was minimal.

Then we proceeded with stepwise selection (using AIC as the criterion). The selected variables can be found in table 5. After fitting the model, we compared the reduced model with the full model. The comparison results are below.

| Model | AIC | BIC | Adjusted R-Squared | RMSE on Test Data |
|---|---|---|---|---|
| Full | 76496.90 | 76577.85 | 0.216 | 94.90209 |
| Reduced | 76492.16 | 76552.87 | 0.216 | 94.87833 |

**Table 4: Linear Regression Model Comparison Results**

In the above table, we see that across all metrics, the reduced model is at least as good as the full model. Namely, the AIC, BIC, and RMSE are all lower for the reduced model. It also matches the full model on adjusted $R^2$. Therefore, the reduced linear regression model is chosen. The model summary is below.

| | Coefficient | Std. Error | P-value |
|---|---|---|---|
| Intercept | 8018.3723463 | 2162.2217734 | 0.0002104 |
| Latitude | -134.2013351 | 29.1221668 | 0.0000041 |
| Longitude | 25.3390129 | 17.0796879 | 0.1379717 |
| Property Type: Private Room | -102.5670897 | 2.8594730 | 0.0000000 |
| Property Type: Shared Room | -109.2911255 | 13.5118714 | 0.0000000 |
| Crimes Per Day in Past Year | 5.0751628 | 2.3277686 | 0.0292745 |
| Distance From Downtown | -0.9757519 | 0.2449887 | 0.0000689 |
| Median Household Income Scaled | 0.0495176 | 0.0074168 | 0.0000000 |

**Table 5: Linear Regression Results**

We see that the p-value for all of the coefficients significant, except for longitude. So it is in fact the case that these predictors have a significant association with price. Looking at the specific values now, we see that in the presence of all other predictors, for a one unit increase in the distance from downtown, we expect that the price goes down by \$0.98. This agrees with the results we saw from figure 5, and with the rationale we gave for why this can make sense.

We also see that in the presence of all the other predictors, for a \$100 increase in medium household income, the price of the listing hoes up by on average, \$0.05. The weak positive trend agrees with what we concluded under figure 4, and the rationale given there for why this can make sense.

Finally of note, is that in the presence of the other predictors, for one more crime per day in the past year in that neighbourhood, the price of a listing in the neighbourhood increases by $5.08 on average. This is very counterintuitive and almost contradictory to what we expected from figure 4, and from intuition (in section one). The only way that this can make sense, is that if we consider downtown to be more crime prone than the rest of the city. Then since proximity of downtown increases price, so would crime.

### 3.3.3. Linear Mixed Model

Next we fit a linear mixed model. The reason why we do this, is that is a natural grouping structure in our data: neighbourhoods. Furthermore, in our data, we have both individual (i.e. AirBnB listing) level predictors, as well as group level covariates (e.g. neighbourhood crime level). So it is a natural choice to fit a model that accounts for this. Namely, we account for this grouping structure by considering all of the covariates in the full linear regression model, with an additional correlated random slope for the neighbourhood crime level and neighbourhood income level covariates, with respect to the neighbourhood.

| | Coefficient | Std. Error | DF | t-value | P-value |
|---|---|---|---|---|---|
| Intercept | 6633.2621599 | 3023.0838600 | 6163 | 2.1942038 | 0.0282580 |
| Latitude | -115.6085007 | 41.6137014 | 6163 | -2.7781355 | 0.0054837 |
| Longitude | 18.0596991 | 22.7446698 | 6163 | 0.7940190 | 0.4272150 |
| Property Type: Private Room | -103.2126527 | 2.9167141 | 6163 | -35.3866191 | 0.0000000 |
| Property Type: Shared Room | -108.5336366 | 13.5617237 | 6163 | -8.0029382 | 0.0000000 |
| Crimes /Day in Past Year | 3.2821630 | 3.6765545 | 113 | 0.8927280 | 0.3739001 |
| Distance From Downtown | -0.8376924 | 0.3359832 | 6163 | -2.4932567 | 0.0126838 |
| Med. Household Income Scaled | 0.0446758 | 0.0115721 | 113 | 3.8606367 | 0.0001888 |
| Restaurant in Text Description | -2.1996709 | 3.8130260 | 6163 | -0.5768833 | 0.5640394 |
| Shop In Text Description | 1.1847149 | 3.8713469 | 6163 | 0.3060214 | 0.7595987 |
| Park in Text Description | -1.3799599 | 3.5357750 | 6163 | -0.3902850 | 0.6963393 |

**Table 6: Linear Mixed Model Results**

The estimated coefficients are all relatively similar to those that were obtained from fitting the linear regression. The difference is that the estimates are uniformly somewhat closer to 0, than those of the linear regression. However, the standard errors of significant estimates are slightly higher. Furthermore, neighbourhood crime is no longer significant. So at least when it comes to examining the relationship between the predictors and the response, this model may not be as good.

### 3.3.3. Machine Learning Models

Now we fit three different machine learning models. First, we fit a basic decision tree, using all of the predictors. We fit this using parameters of minsplit = 10 (i.e. minimum observataions for a node to be split), minbucket = 3 (i.e. minimum observations in a leaf), and do 10 cross validations. In particular, we choose a complexity parameter based off of what minimises xerror. The optimal complexity parameter we obtained was 0.0002.

The second model that we fit was a bagging model. The third model that we fit was a random forest model. And the fourth and final model fitted, was a gradient boosting model. We note that for the gradient boosting model, we chose to use 1000 trees and 5-fold cross validation.

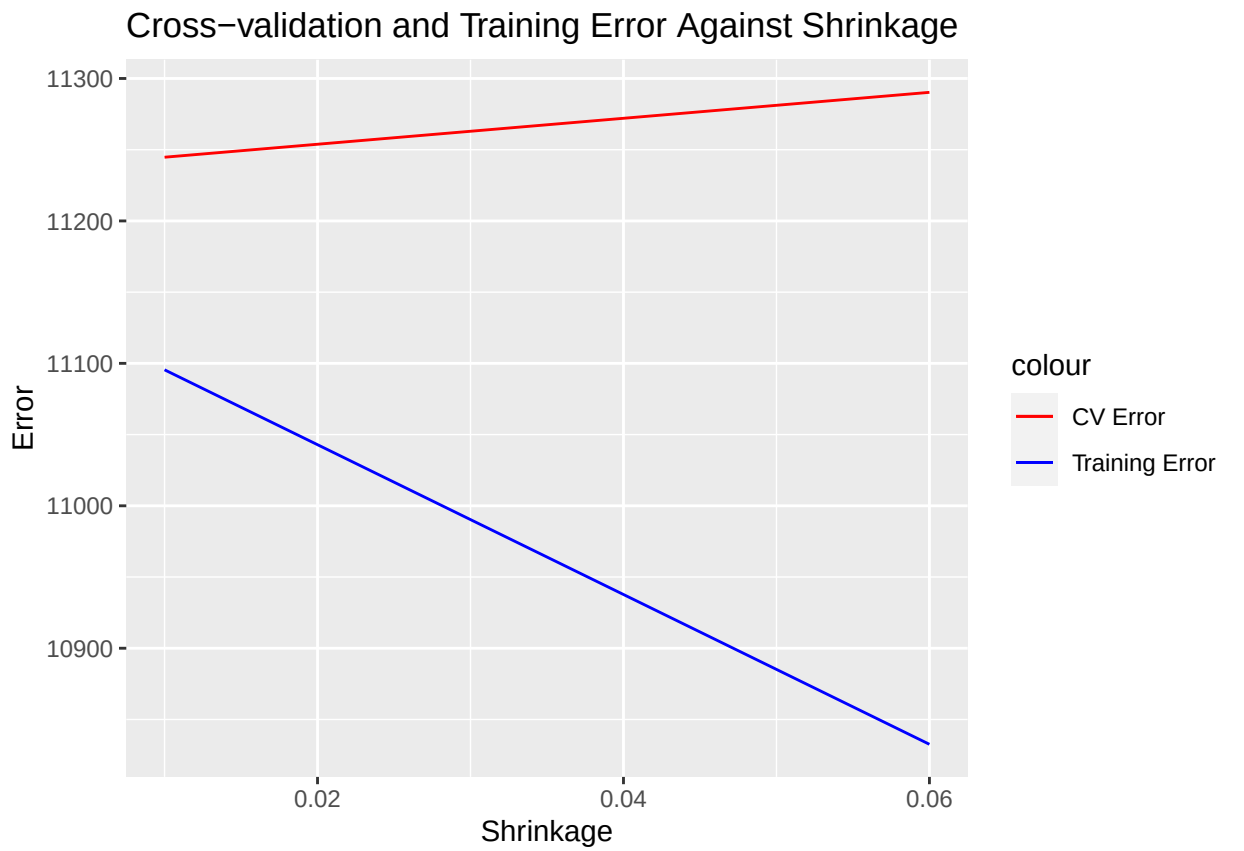


**Figure 6: Shrinkage Against CV and Train Error**

We chose the shrinkage parameter, by plotting the CV and train error for values of the parameter between 0.01 and 0.5. We found what while the training error continues to decrease as the parameter increases, the CV error slowly increases. To try and minimise both, we chose a shrinkage parameter in the middle of the range we were looking at: 0.25.

After fitting the bagging, random forst, and boosting models, we plotted their variable importance plots.

```
grid.arrange(boost_imp, rf_imp, bag_imp, nrow = 1)
```

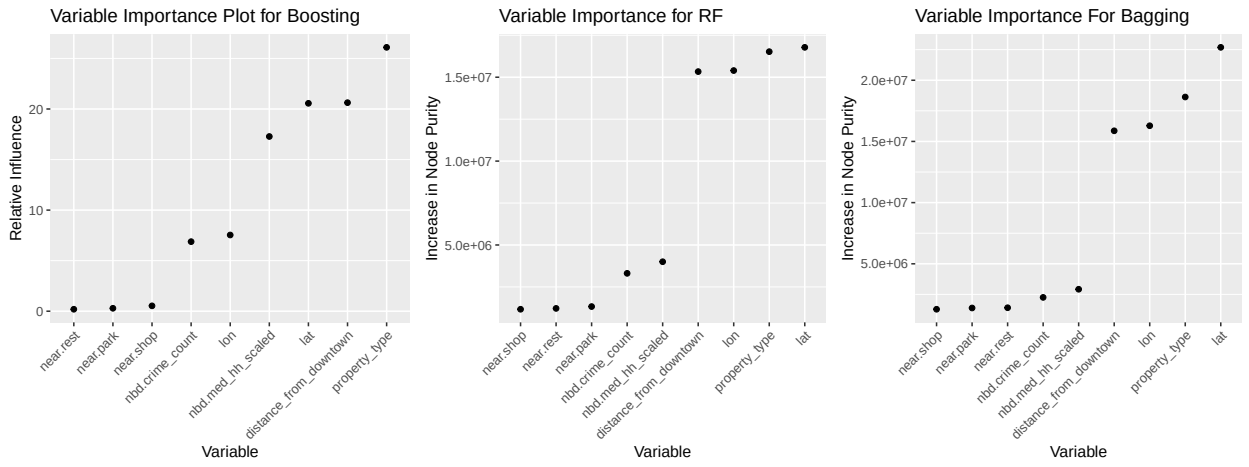


**Figure 7: Variable Importance Plots for Bagging, Random Forest, and Boosting**

The variable importance plots for the three different models are almost identical. The fact that they agree, gives us confidence that this is a good estimation of how important they actually are to the price of an AirBnB listing. Based off of the variable importance plot, we see that the four most important predictors are latitude, longitude, property type and the distance from downtown. The general idea we get from this, is hence that the location and property type of the listing are what impacts prices the most. In comparison, other neighbourhood characteristics like neighbourhood crime or income do not really appear to impact the price. This in general, appears to agree with a lot of the exploratory visualisations that we saw, where proximity to downtown generally appeared to be particularly important (e.g. the leaflet map displaying prices on the website). It also agrees with what we saw in figure 4, where the distribution of price did not really appear to change much with income and crime levels.

#### 3.3.4. Comparison of All Predictive Models

We now make a comparison of the predictive performance of all of the models that we fit (except the full linear regression one, since the reduced one was better). Particularly, we test them on their RMSEs on the test data.

| Model | RMSE on Test Data |
|---|---|
| Linear Regression | 94.87833 |
| Linear Mixed Model | 94.72725 |
| Classification Tree | 172.29376 |
| Bagging | 99.08127 |
| Random Forest | 97.77682 |
| Gradient Boosting | 94.68930 |

**Table 7: Comparison of RMSEs of Fitted Models**

Comparing RMSEs, we find that all of the models have rather similar errors, except for the classification tree, which performs almost two times worse than the others. In terms of the model that minimises the RMSE, gradient boosting performs the best. So at least on the basis of predictive power, that model would be preferred. In general though, all of the RMSEs but that of the classification tree are good, in the sense that they are both lower than the mean and median price values (see table 2). So they are relatively reasonable.

In terms of which model provides the most insight into the relationship between the response and the predictors, it would have to be the the gradient boosting model (in fact, bagging and the random forest model do equally well in this regard). The reason that the linear regression model is not as great, is that although it gives very straightforward interpretations of the effects of each of the predictors, some of the coefficient results are counterintuitive - such as the fact that more neighbourhood crime appears to correlate with higher prices. This problem extends to the linear mixed model, which has the additional issue of having higher standard errors on each of the coefficient estimates, compared to that of the linear regression ones. In comparison, the variable importance plot of the gradient boosting model is not only intuitive in that it makes sense, it also agrees strongly with our exploratory visualisations. So since it is likely to be the most accurate.

## 4. Conclusion and Limitations

The conclusion that we arrive at, is that although neighbourhoods do have a significant impact on the price (from the ANOVA model), it is questionable how much of an impact that income and crime levels of a neighbourhood have on price. This is not only seen in the visualisations, where any association appears to be weak, but also in the variable importance plots, where these covariates appear to have a weak impact on predicting price. On the other hand, proximity to downtown does matter. So we can say that the

closer a neighbourhood is to downtown, the higher we would expect the prices to be. This is corroborated by the leaflet map displaying the results of the post-hoc Tukey test, showing that the "most different" neighbourhoods to others in price, happened to be closer to downtown in general.

There are some limitations to our work though. For instance, we were only able to consider a limited subset of neighbourhood characteristics that could potentially have an impact on the price of a listing. For instance, one could think that walkability of a neighbourhood might be important to price - this was a word that showed up a lot in the neighbourhood description text wordclouds we constructed. We unfortunately did not have the data for this.

There are further limitations to our analysis. For instance, the median household income comes from the 2021 census (which is the most recent time we can find the relevant data), and yet the rest of our data is from 2023-24. This is an unavoidable issue with the scope of our project.

# 5. References


1. Get the Data. Inside Airbnb. (n.d.). http://insideairbnb.com/get-the-data/
2. Latitude.to. (n.d.). Latitude and longitude of Downtown Toronto. Latitude.to, maps, geolocated articles, latitude longitude coordinate conversion. https://latitude.to/articles-by-country/ca/canada/7404/downtown-toronto
3. Major crime indicators open data. Toronto Police Service Public Safety Data Portal. (n.d.). https://data.torontopolice.on.ca/datasets/TorontoPS::major-crime-indicators-open-data/about
4. Open data dataset. City of Toronto Open Data Portal. (n.d.). https://open.toronto.ca/dataset/neighbourhood-profiles/


Note: For posterity, much of the Final Report was adapted from my Midterm Report (as was mentioned that we could, from the pdf).