Team Project – Final Report

Maximizing *Airbnb* Revenue Through AI Derived Listing Price Estimates

Exec Summary & Company Background

The objective of this project was to identify an automated method to best estimate a suggested booking price for hosts that will maximize their earnings, and by extension maximize *Airbnb* revenue.

Airbnb is a community-based online platform for listing and renting local homes, with bookings made through its website and on mobile devices through its app. The company connects hosts and guests, facilitating the process of renting without owning any rooms itself. With over 150 million users and 7 million listings in 34 thousand cities as of 2020, the company is the largest of its kind.

In finding accommodation, guests can search for rooms using filters such as lodging type, dates, location, and price, and can even search for specific types of homes, such as bed and breakfasts or vacation homes. Hosts provide prices and other details for their rental or event listings, such as the allowed number of guests, home type, rules, and amenities. Following a stay, guests have the option to leave both reviews and comments about the accommodation, visible for prospective guests to read.

We compared the seasonality trends from 2020 to 2021. The business insights from seasonality for 2021 show that prices for Toronto continue to rise after the summer whereas prices for Vancouver drop after the summer. We recommend considering lowering prices in Toronto if demand is not significant in the Fall or offering discounts to increase booking. We recommend increasing marketing and partnering with tourist spots to attract customers in the Fall in Vancouver.

Problem Identification & AI Business Case

Airbnb's revenue model is commission based, with the company receiving a 6-12% fee from the guest for booking an accommodation, while additionally charging the host a 3% fee for each successful transaction.

As *Airbnb's* revenue is directly tied to the quantity and price point of successful accommodation bookings, it is in the company's financial interest to facilitate the most bookings at the highest price point possible. A key factor in maximizing revenue is in striking a balance between rental prices hosts would like to charge with a price guests would be willing to pay. Should prices be set too high, guests will be less likely to purchase accommodations, leading to fewer transactions and therefore lower revenue. Likewise, should hosts set their prices too low, the company will miss out on potential revenue that would have otherwise been earned from a higher transaction price.

The business problem addressed in this project tackled this price-setting dilemma by identifying an automated method to best estimate a suggested booking price for hosts that will maximize their earnings, and by extension maximize *Airbnb* revenue. To add some Canadian relevancy to the project, data from the most recent listings in Toronto and Vancouver will be used, taking into consideration factors such as location, amenities, date, reviews, and any more relevant variables. Pertinent questions to be answered will include understanding the effects of seasonality on price, as well as determining which listing factors have the greatest impact on price. Ultimately, we endeavoured to determine whether prices can be predicted based on the information acquired in the datasets.

Description of Data Used

The datasets were acquired from *Inside Airbnb*, an online tool that collects publicly available data from *Airbnb* listings, grouped based on city. The two datasets to be gathered for each city are titled: *Calendar* data and *Listings* data, with a common dependent variable of listing price.

> *Calendar* data is organized by listing identification number and compares listing price to the time of the year in which the room was listed. This data will be used to determine the effects of seasonality on listing price. The dataset is complete, with no missing or null values.
>
> The *Listings* data contains the bulk of the information and variables, from host information like average response time and length of experience as a host, to neighbourhood information, property type, accommodation size and capacity, the maximum and minimum lengths of a stay, and review data like numerical review scores and quantity of reviews per month. While many null values exist in this dataset, the bulk of those fall under variables less relevant to the objective of listing price prediction.

In Airbnb terminology, *listings prices* are the prices set by hosts for potential bookings at their accommodations. This information is public, and differs from *transaction prices*, the prices of accommodations actually purchased by guests. Transaction data is private and confidentiality information, and is available internally to Airbnb.

For the purpose of this project, it is important to understand that listing dates fall in the future, as only future dates can be booked by guests and not dates that have already occurred in the past. (Consider how if attempting to book a summer cottage, only prices for June 2021 will be displayed rather than any prices for June 2020). These prices are not predictive, as they are simply the prices that hosts have already decided to set for prospective guests in the upcoming season. Whether guests decide to purchase at those prices is a different matter.

Unlike listings data, transaction data is only available retrospectively, that is after a transaction has already occurred. To conclude, while transaction data for future dates can only even be predictive, listings data is not.

For comparative purposes, we also gathered archived listings data scraped one year ago. These listing dates were set by hosts in the first two months of 2020, reflecting a hospitality industry operating in a pre-COVID world, and cover a date range of February 2020 to January 2021 inclusive.

Analysis of Methods Applied & Discussion of Results

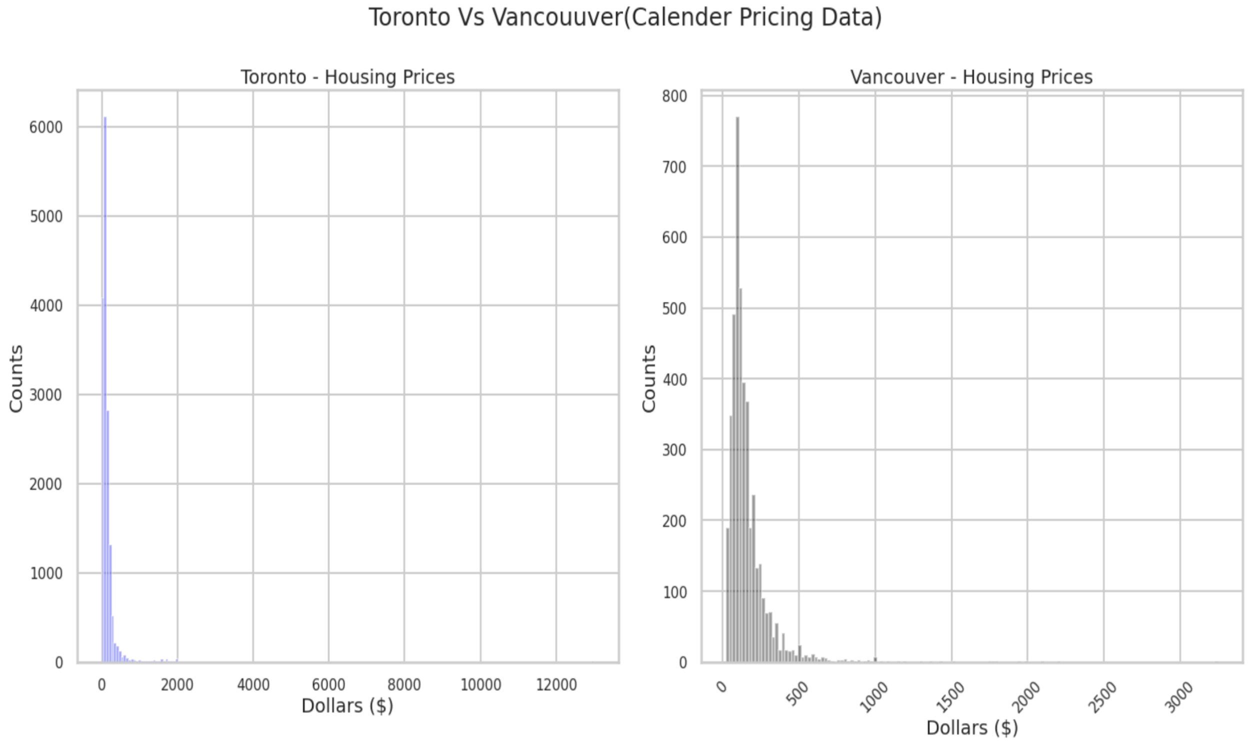
Methods:

- Histogram in different scales for both cities
- Kernel density estimation
- Data Visualization in trajectory plot

After initial cleaning the data, we checked for patterns in seasonality for either market by plotting a histogram with listing prices. We checked the prices less than and greater than $1000.

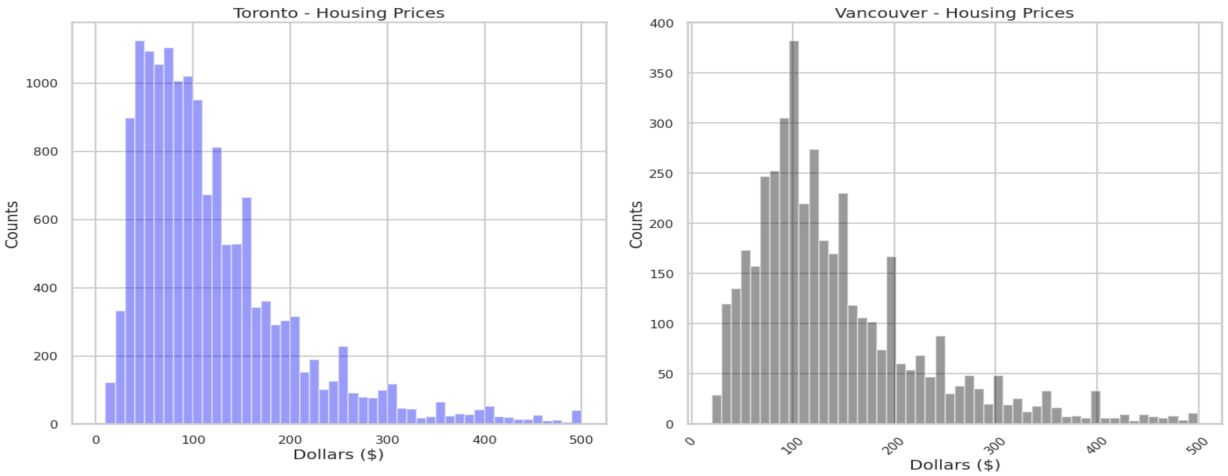
The following section describes the pricing distribution of the listings and the seasonality trends for both cities.

Overall listing price distributions for Toronto and Vancouver are as follows:

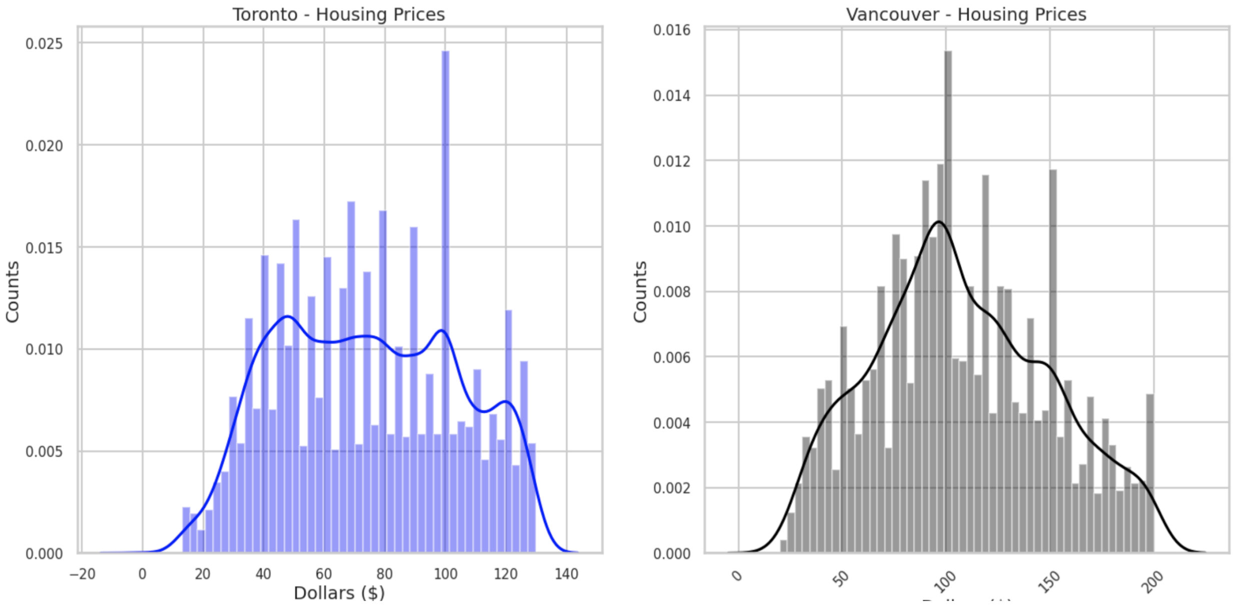


As a first step, we plotted the histogram to see the distribution of listing prices in both cities. The blue color is Toronto, the black is Vancouver. As shown in the graph, overall, Toronto has more house listings than Vancouver's. The y-axis of Toronto is significantly higher than Vancouver's. Both of the distributions are right skewed. Most prices fall under $1000 in Toronto most under $500 in Vancouver, there are long tails for both graphs. Toronto's data is more concentrated under $200 while Vancouver's is spread out under $500.

In Toronto, the highest price reaches around $12,000 and the highest price for Vancouver is over $3,000. Since the tail is very long, to dig into the data distribution details more clearly, we narrow down the range of the price under $500. This step can give us more insights on the main distribution of the price in both cities.

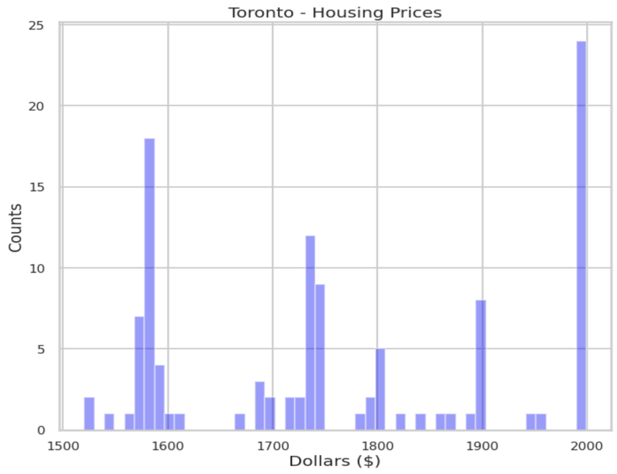


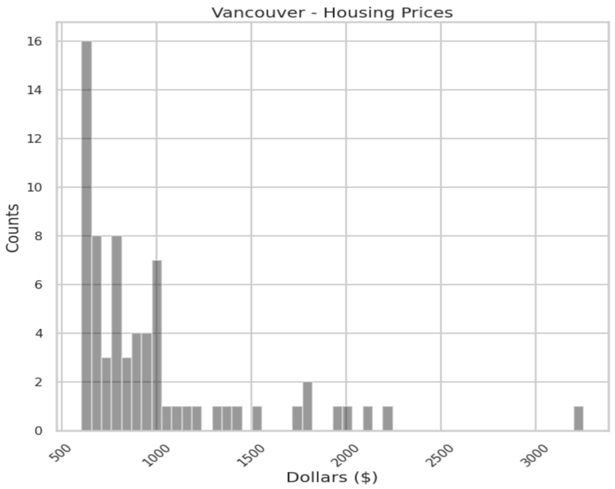
The price under $500 for both cities are still rightly skewed in the visualization. We narrowed down the range to $200 for Vancouver and $140 for Toronto and implemented the kernel density plot to normalize and smooth the graphs. The distribution for price under the threshold for Toronto and Vancouver is almost normally distributed.

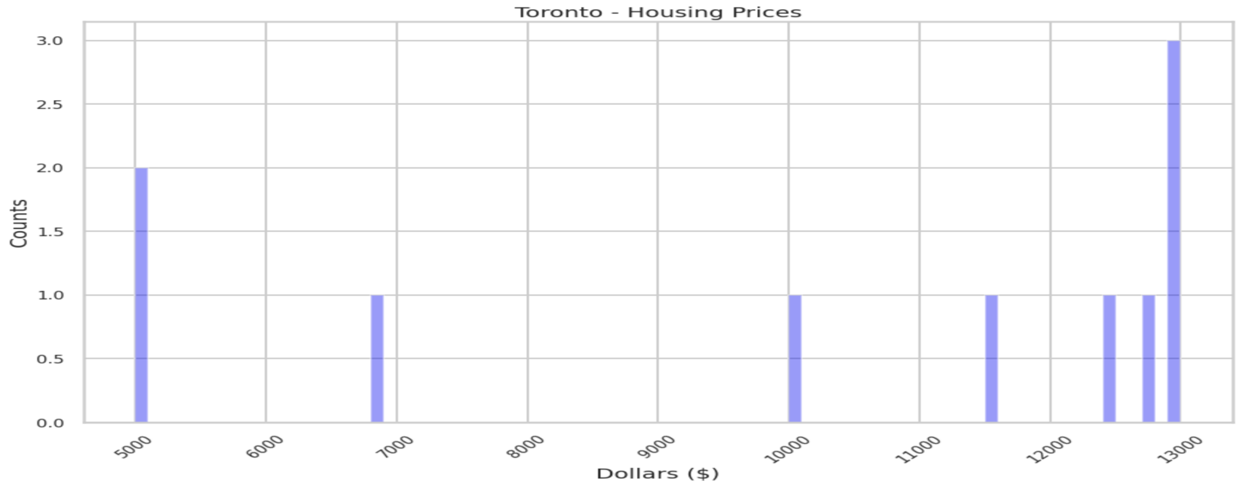


To better understand the distribution, we also generated an outlier distribution in histogram.

The outlier price limit we set for Toronto is $1000 to $2000, with another threshold above $4000. For Vancouver, the price limit sets greater than $600.





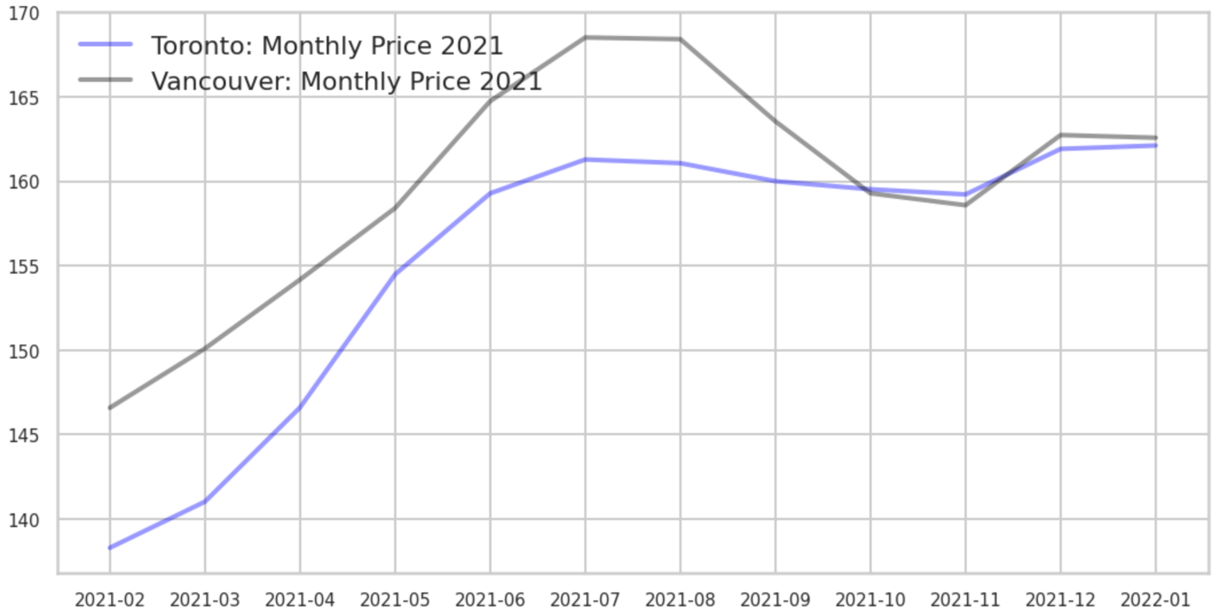


As shown in the visualizations, in Toronto most outliers fall into the range from $1,500 to $2,000, a few more fall in the range above $5,000. In Vancouver, most outliers are concentrated in the range between $500 to $1,000, the rest are above the $1,000 range and each have a count equal or lower than 2.

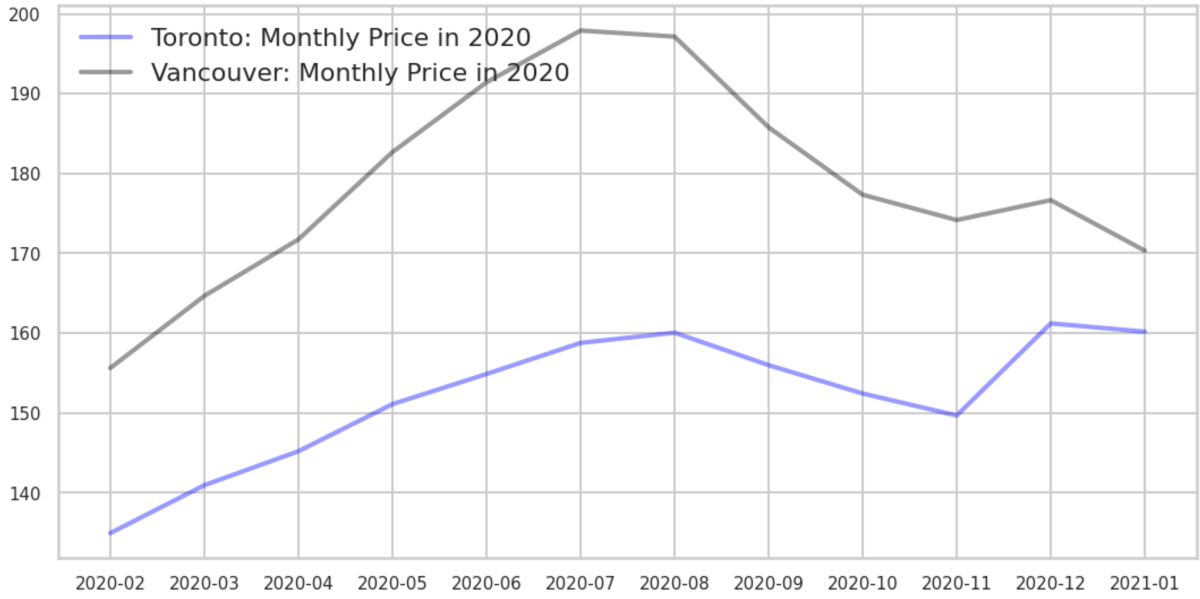
Such a distribution graph would help us understand the pricing data better and later to reduce the outliers, do the modeling, and the outlier analysis.

Existing Listing Price with Seasonality for 2021 and Comparison with 2020:

As a goal, we summarized the updated datasets that start from February 2021 to January 2022 for both Toronto and Vancouver. Since the datasets have included the listing price posted by the host throughout the 2021, it aims to provide insights to those new entry hosts and let them better understand the price trend with seasonality. We also include the datasets for 2020 to be compared with our 2021 trend.



When examining results, the seasonal trend for Toronto and Vancouver are very similar which makes sense as they are both popular Canadian cities and travel destinations. Throughout the year, the prices for Vancouver are much higher than Toronto in 2020 and 2021. There is an increasing trend from February to July for both Toronto and Vancouver. The highest price for Vancouver is during the summer, the same as Toronto. After July, the Toronto trend line is stable, while Vancouver has slightly decreasing trend in autumn. The average price for both cities ranges from $140 to $168.



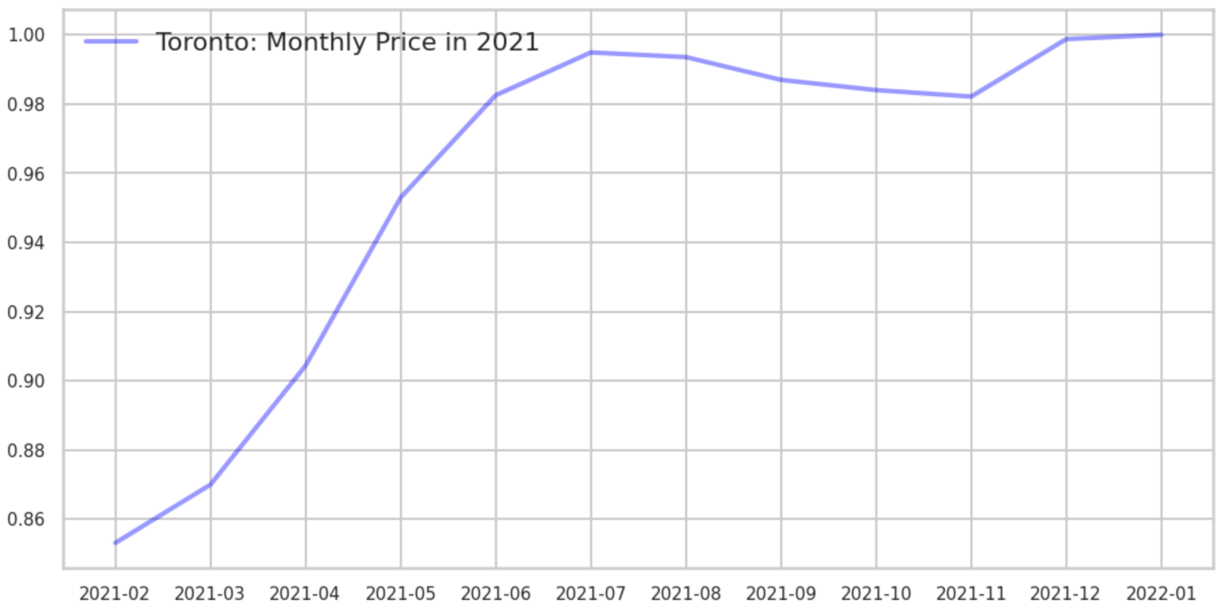
Compared with 2020, the overall trends for both Toronto and Vancouver are similar. It is interesting to find that Toronto has the highest price in December in both years, which might be related to high customer demand near Christmas and New Year's Eve.

Then we examined each city separately. Firstly, we adjusted the scale to range from 0 to 1 using the following formula:

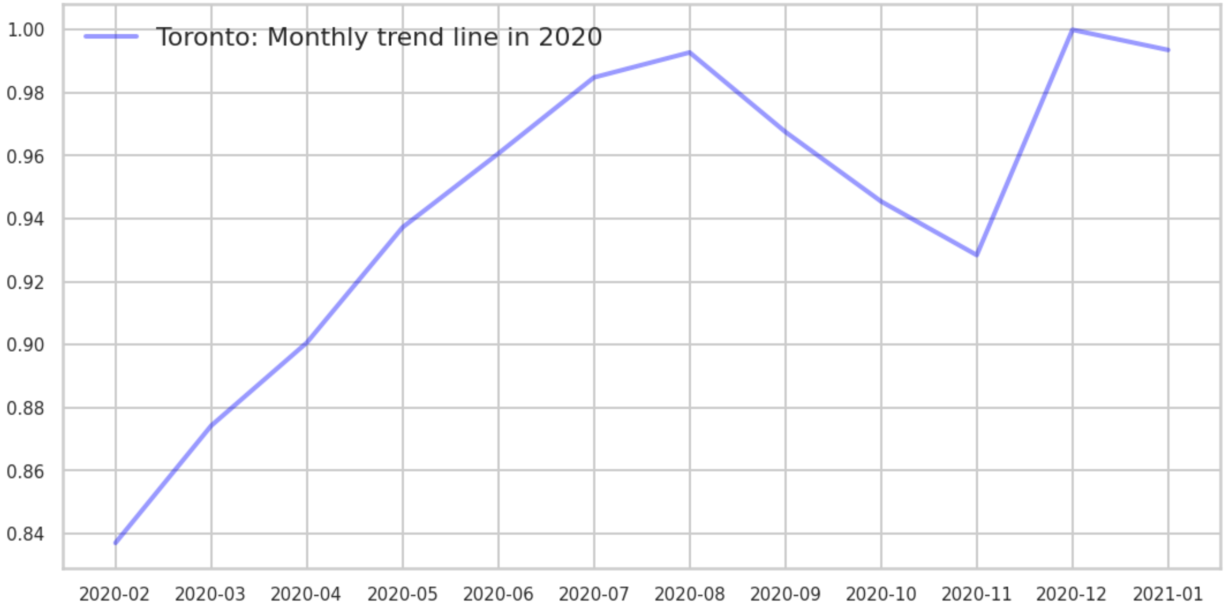
$$\textit{Index} = \frac{\textit{Price in each Month}}{\textit{Max Price in the Year}}$$

We then plotted two graphs as seen below, one for each city.

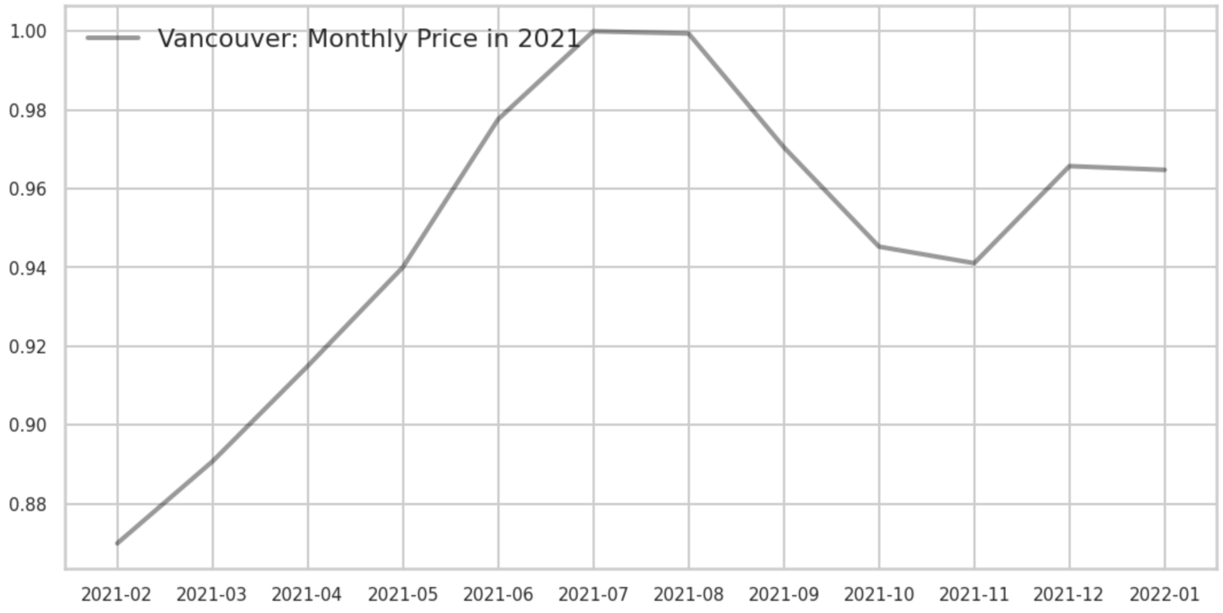
For Toronto, the trend is increasing at the beginning of February and reaches 0.994 in July. After July, the trend is stable until the end of year. The highest price is in January, fitting with findings mentioned above.



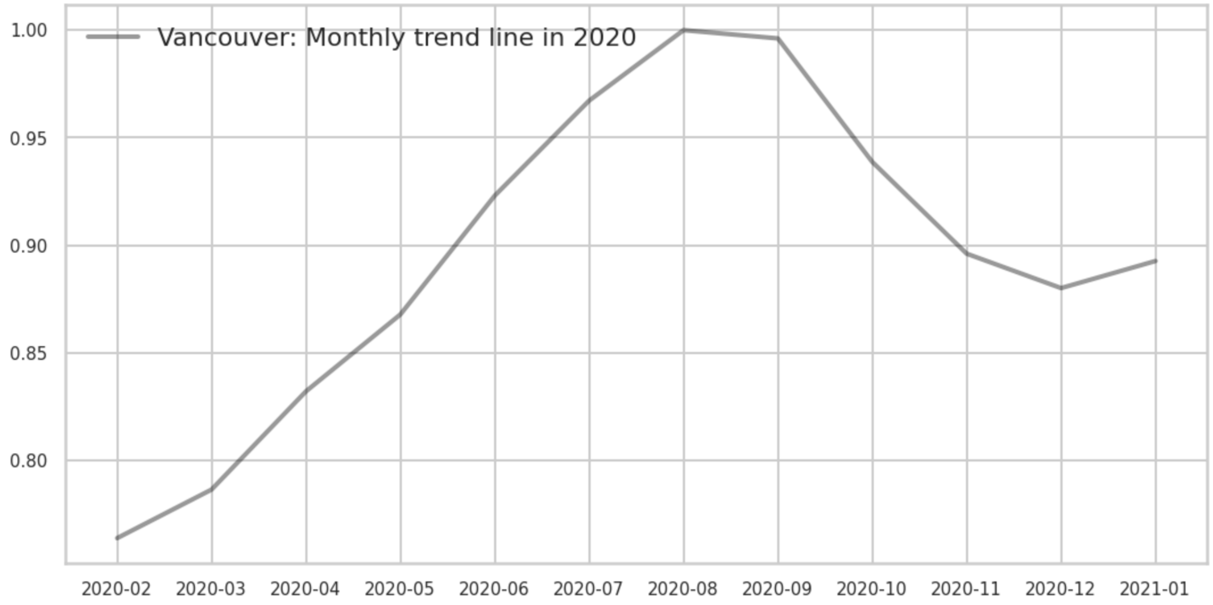
Compared with 2020, Toronto listing price drops significantly in November and reaches to the highest in December. However, for both years, December and January have the highest average listing price in Toronto across all the listings.



For the 2021 listing price in Vancouver, it starts to increase since February similar to Toronto listing price trend. During the summer season, the prices are the highest throughout the entire year. At the beginning of autumn, the price drops until November and reaches back in December.



In 2020, it is very similar to the trend as we observed in 2021. The average listing price starts to increase in February and reaches the top during the summer. Once the summer is near to the end, the price is going to drop. However, in 2020, the average price in December is lower than November and the graph goes to a local minimum.



Overall, the current existing listed price shows the trend of listing price from the existing hosts in 2021. Compared with the year 2020, we did not find any major differences. The city of Toronto and the city of Vancouver are likely to have the highest price in July and December individually, due to the peak season for holidays and vacations.

Discussion of How Results Address the Problem and any Business Insights Gained

Prices are overall predictable due to seasonality. Overall, the prices listed by the host for 2021 are similar for both Toronto and Vancouver. There are more listings already posted by the host in Toronto than Vancouver. As we can see from the plots, the listing price is lowest in February since the cold weather for both Toronto and Vancouver. The price is increasing in spring and summer. Since the weather is becoming warmer, more visitors plan their vacation. Therefore, the price and bookings will be competitive during the peak season for both cities.

Prices continue to rise in Toronto after summer in 2021. From a business perspective, this can lead to potentially fewer bookings in the Fall because of higher prices in the off-peak tourist season. We recommend considering lowering prices if demand is not significant in the Fall or offering discounts to increase booking. For the summer of 2021, accommodations in Vancouver are more expensive than Toronto and prices drop in Vancouver after the Summer. Airbnb should increase marketing and partner with tourist spots to attract customers in the Fall since Vancouver is warmer than Toronto in the Fall and this can be an effective marketing technique to attract tourists from Toronto and other Canadian cities. Then, after increasing demand, they should increase prices later on to avoid losing potential revenue.

In addition, during the pandemic, more companies/schools are allowing their employees/students to work/study from home and continue to do so meaning less accommodations will likely be used for co-ops and student housing in 2021. Therefore, there may be less demand in the near future for accommodations for non vacation purposes so the prices in the Fall may decline further and so going forward, Airbnb prices may be even lower throughout the seasons due to this potential change in future demand because of the pandemic. The uncertainty of the effects of the pandemic could make consumer demand fluctuate and therefore, these changes need to be monitored by Airbnb in the future model.

Data Modeling and Evaluation

In the data modeling and evaluation part, a machine learning model is created to predict the price that Airbnb should charge when a property is added, and feature importance will also be evaluated in this step. First, we do the data cleaning in both calendar and listing dataset. We only remain the number in the price column and convert the column time's data type into datetime64. When cleaning the listing data, we keep the selected column and drop columns with NA beyond a certain threshold. Then we drop rows that have NA values with a threshold=1 and fill the remaining NA using 0 values. Next, we join the calendar and listing datasets together and drop the price and neighborhood variables.

In the machine learning model selection, we decide to test the random forests algorithm because it can be well applied to create linear or non-linear regression models with many features. Meanwhile, this model can be trained and get results in a relatively fast speed and evaluate feature importance. However, some disadvantages like occasional overfitting and outputs of feature importance may be skewed towards the categorical variables that are also worth being noticed.

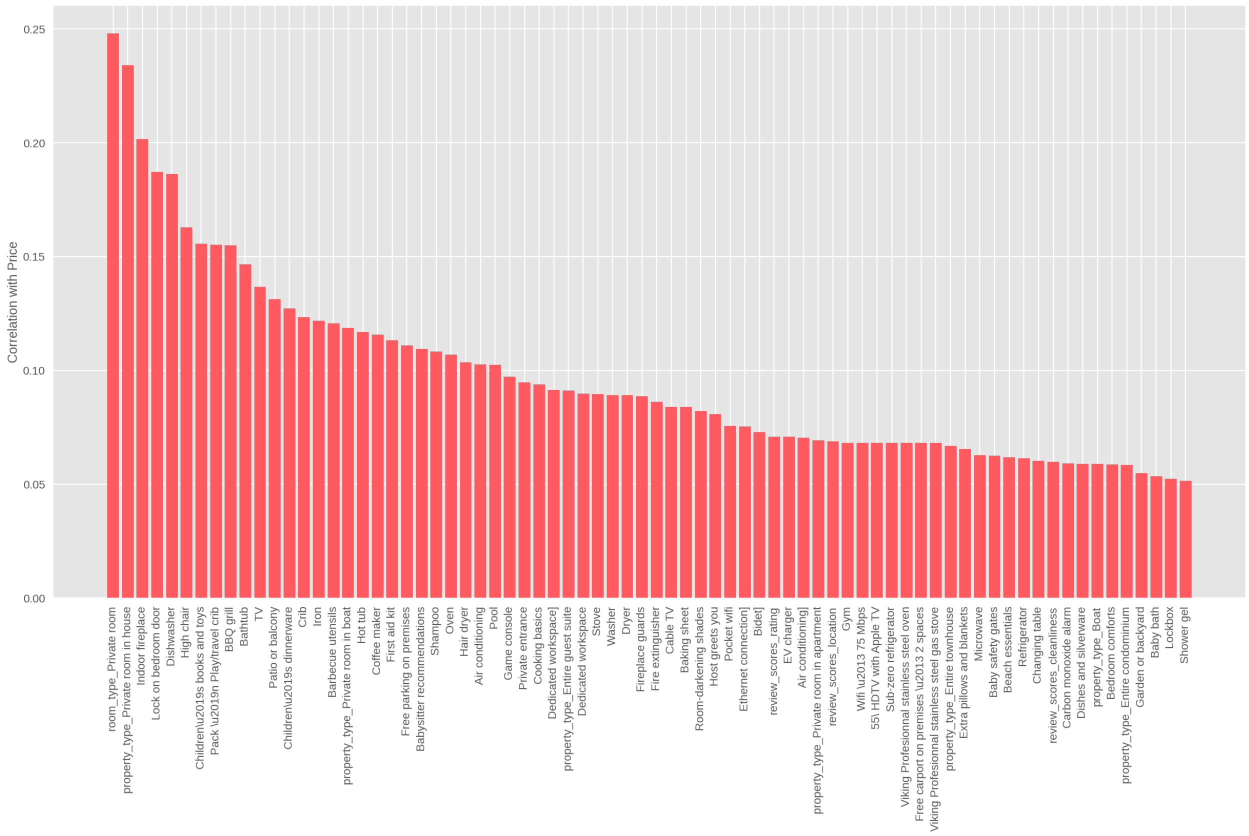
In the model testing part, we first used Gridsearch for the parameter setting. The RMSE of the training set is around 72% but the RMSE of the testing set is just around 26%. The result shows that the model is overfitting. We think it is because we have too many independent variables all trying to influence the dependent variable.

It's also worth pointing out that calculating the RMSE has the benefit of penalizing large errors but in this case it's possible that using MAE is a better function as it might be fine that some of the data points (the outliers) are way off from our prediction. Given this it might be worth removing these and re-testing. Then we can see from calculating the MAE the absolute error is around 58.72% with a percentage error of 34.85% and the Accuracy is around 65.15%which again isn't perfect but isn't as skewed by outliers as by the RMSE.

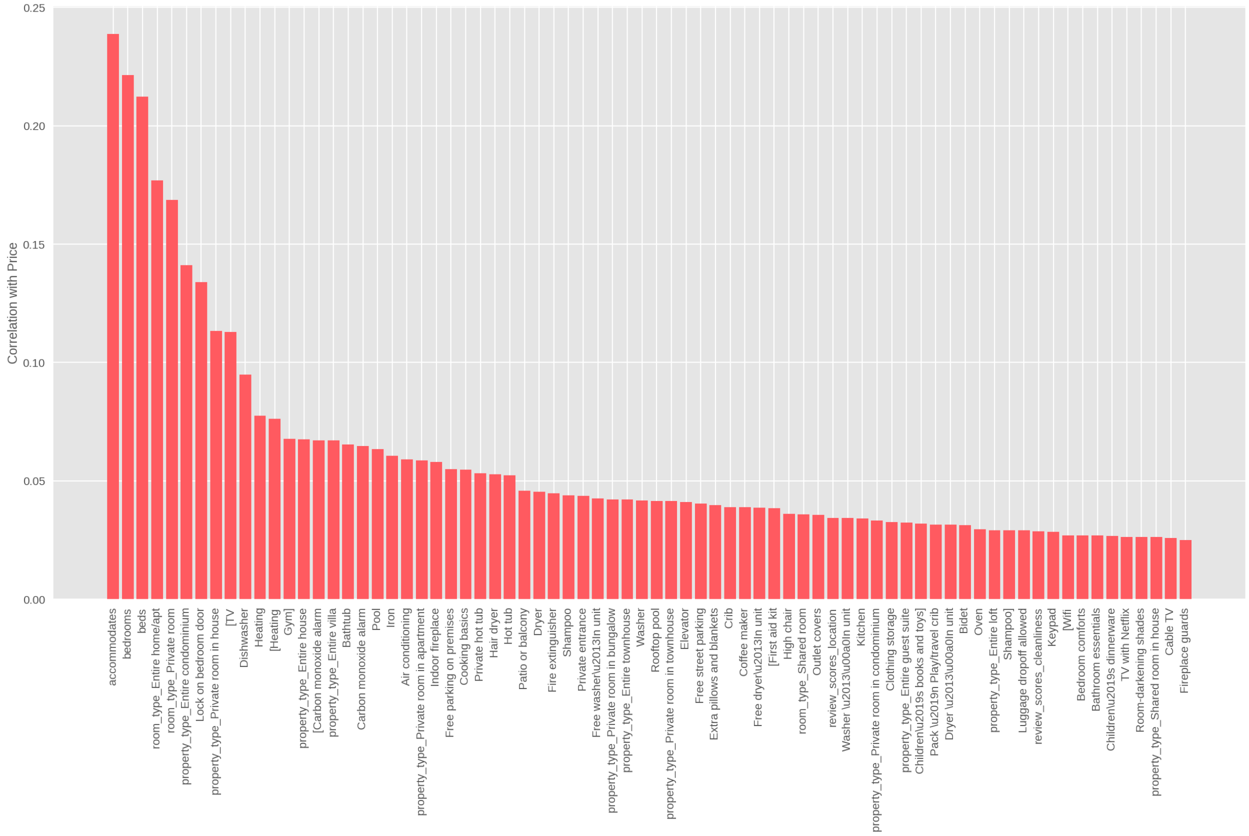
Removing the outliers increased the accuracy and the model is now probably being as skewed by the outlier data points. The next step will be to perform feature scaling, this will ensure that variables with a bigger variance don't have a bigger influence on the data. Based on the result, the model is improving.

Reducing Number of Features

For the next step, we need to reduce the number of features to avoid overfitting and increase the accuracy. Also, we need to find out which features are mostly related to the price. The first method to remove unnecessary columns was to remove uncorrelated columns to the price column. We calculated the correlation between each feature and price and eliminated elements with less than 0.05 correlation. It reduced the number of columns from 527 to 80 in the Vancouver data set and from 702 to 74 in Toronto. For finding the threshold, we plotted features and their correlation. To find the threshold, the correlation of each column and price were plotted, then we found a point that the correlation drops drastically from that point.



Toronto dataset - Correlation for each features against price

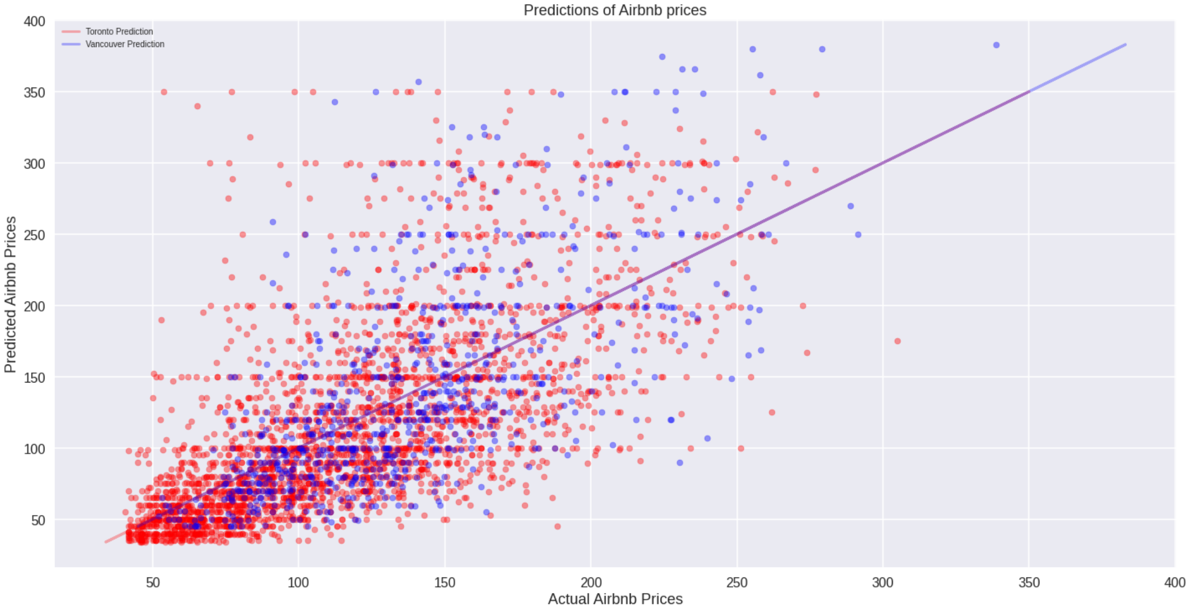


Vancouver dataset - Correlation for each features against price

We utilized RFE (Recursive feature elimination) and the Forward Selection method to filter more columns in the next step. We narrowed down the number of features to 50 for Toronto and 48 features for Vancouver using this method. With the help of this method, we could enhance the accuracy by around 1%. As a result, we reached 73.13% and. 72.18% accuracy for Toronto and Vancouver, respectively.

Due to the large size of data, this process took several hours to train. To make the training process faster, we utilized the PCA (Principal Component Analysis) method to compress the features more. The principal component analysis is a technique for reducing the dimensionality of such datasets, increasing interpretability and minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance.

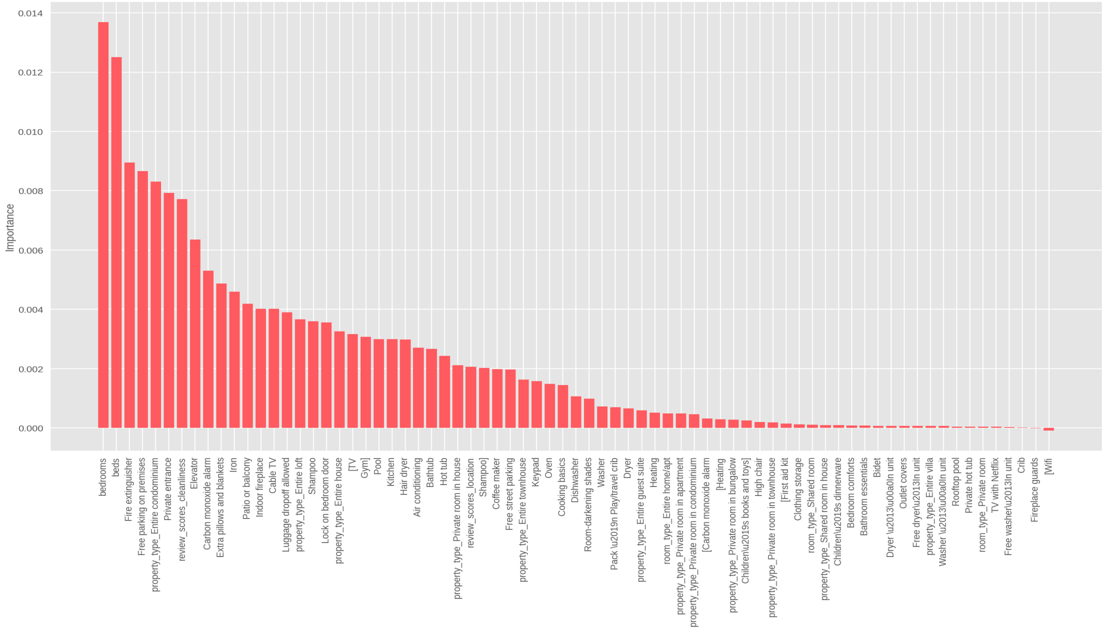
In the below chart, the true prices and the predicted prices have been plotted. The line on 45 degrees shows the perfect model to compare it with our model.



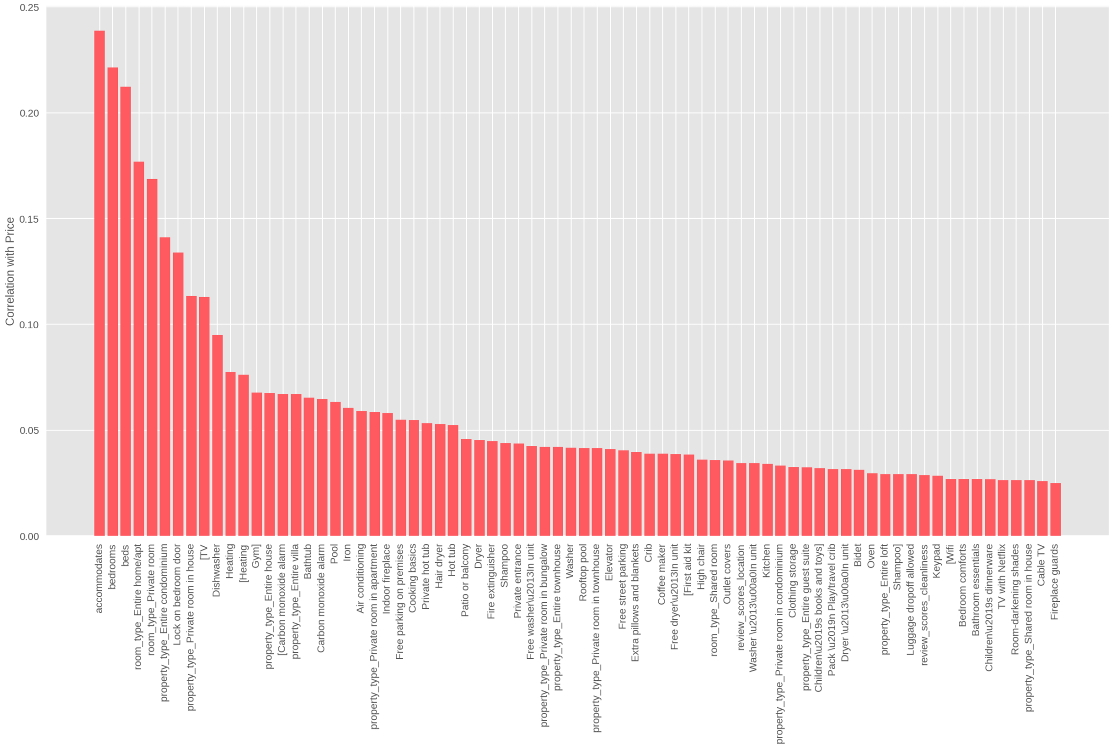
Model accuracy plotted

Feature Importance

In order to obtain the importance of each variable in determining the price, the difference in error was measured once with that variable and once without considering that variable in the training process. In this case, the most important feature is the one that reduces the error of the model the most. The importance of each feature was plotted.



Toronto dataset - Importance for each features against price



Vancouver dataset - Importance for each features against price

As we can see in Toronto's chart, the number of bedrooms, number of beds, Free Parking, Private Entrance, Cleanliness Review Score, Elevator, and so on are variables that affect the price the most. The interesting point is that almost all the most important Vancouver features are practically identical to Toronto's important features, except some minor differences like the balcony and patio that count as an important factor in Vancouver, but they are not so crucial in Toronto.

Discussion of AI Implementation in Airbnb

In determining how best to apply the model, we turned to the five major steps of implementation in the Intelligent Systems Implementation Process.

Having developed a business case and needs assessment, we discover that the model would best be applied as a feature to host accounts, guiding them to a preferred listing price that would maximize revenue for themselves and therefore the company.

With respect to preparation and acquisition, the data needed to build the model is readily available for Airbnb, and can likely be created in-house. In improving the model and further refining the prediction accuracy, relevant internal data, not made available in the datasets we used, could be incorporated. For example, while the data we used was *listings* data, in other words the prices hosts have set for their accommodations, Airbnb can take advantage of *transaction* data, internal data of which accommodation listings resulted in a purchase. While we have assumed that most hosts would set their prices at such a level as to maximize their revenue by facilitating the greatest amount of purchases for their rooms, it is possible that not all listings were purchased for every day that they were made available.

An important factor which highly influences the price is location. Due to the computational limit, we did not include this as a factor in the model. However the location also highly affects the price especially if the location is near a tourist attraction or in downtown. We recommend Airbnb to take location into account in their dataset for future analysis.

Finally, the model needs to be tested against its purpose – does it actually help maximize revenue? As discussed, while we've trained it on real listings price data, we have assumed that as an average on the whole, hosts have historically chosen optimal prices for their listings. To test, a control would be needed followed by a test of the model, a test with suggested prices *higher* than the model, and finally a test with suggested prices *lower* than the model. This would indicate if our model is under-suggesting, over-suggesting, or correctly suggesting listings prices that maximize revenue. These tests could be conducted in a variety of sample cities to see if similar conclusions are reached across geographies.

References:

Nath, Trevir. "How Airbnb Makes Money." *Investopedia*, Investopedia, 11 Dec. 2020, www.investopedia.com/articles/investing/112414/how-airbnb-makes-money.asp.

"Get the Data." *InsideAirbnb*, 2 Mar. 2021, insideairbnb.com/get-the-data.html.

AirbnbEng. "How We Deliver Insights to Hosts." *Medium*, Airbnb Engineering & Data Science, 13 Dec. 2016, medium.com/airbnb-engineering/how-we-deliver-insights-to-hosts-7d836520a38.

Wergieluk, Julian. "Histograms vs. KDEs Explained." *Medium*, Towards Data Science, 30 Apr. 2020, towardsdatascience.com/histograms-vs-kdes-explained-ed62e7753f12.

Brems, Matt. "A One-Stop Shop for Principal Component Analysis." *Medium*, Towards Data Science, 10 June 2019, towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c.

Brownlee, Jason. "How to Calculate Feature Importance With Python." *Machine Learning Mastery*, 20 Aug. 2020, machinelearningmastery.com/calculate-feature-importance-with-python/.