A Data-Driven Analysis of Vancouver's Rental and Airbnb Properties with BC Assessment, Vancouver Cultural Spaces, and Translink Datasets

1 Motivation and Background

The sharing economy has experienced significant growth in recent years and Airbnb has become increasingly popular for its flexibility and convenience. Over time, it has revolutionized the way people have been traveling and seeking accommodations. Meanwhile, Airbnb has become an alternative way for property owners to earn income in addition to traditional rentals. To better help property owners, guests, as well as policymakers, and researchers to understand the factors that drive pricing and review ratings in the Airbnb market, and help compare its profitability with traditional rentals, we conducted this data-driven analysis using the Vancouver area as an example. This research will also benefit researchers that are interested in the impact of short-term rentals on housing markets, local economies, and cultural landscapes.

We aimed to identify the key determinants that influence the pricing of Airbnb listings in Vancouver and how they affect review ratings. As a popular travel destination with a diverse cultural scene, Vancouver presents an interesting case study to explore the interplay between various factors. In order to obtain a better understanding, we used data from multiple sources, including BC Assessment, Translink data, and Vancouver Cultural Spaces data to analyze interdependencies between different factors, such as property characteristics, location, proximity to cultural attractions & public transportation, and host profile. We have also integrated rental price data crawled from Craigslist, which allows us to compare revenue between Airbnb and traditional rentals, and such comparisons will serve as a guide for homeowners in their investment decisions.

Our findings can serve as valuable guidance for property owners to make profitability comparisons, optimize pricing strategies, improve listing attributes, and enhance their guest experiences; and also for guests seeking affordable accommodations and high-quality experiences; and policymakers responsible for overseeing short-term rentals to ensure a balanced and sustainable housing market.

We have developed a platform based on our research. This tool will help property owners to determine whether a property is more profitable as an Airbnb or as a traditional rental. It will also provide pricing forecasts and review score forecasts for potential new Airbnb hosts to help them improve their guest experiences.

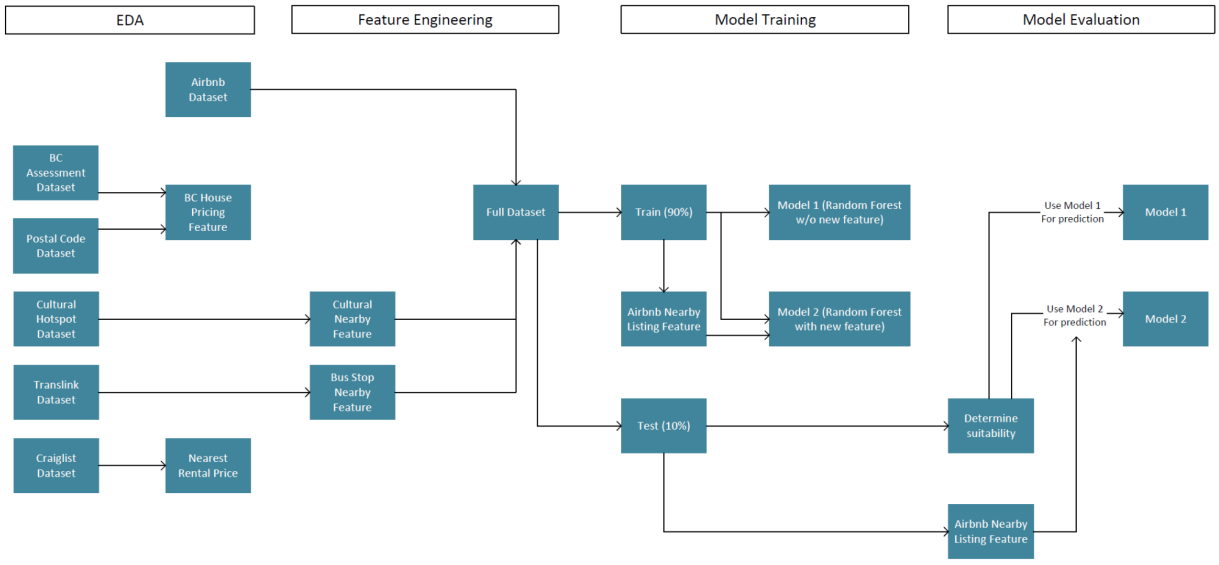
In this report, we present our methodology, data sources, and analytical techniques, followed by a detailed discussion of the results and their implications for the Airbnb market in Vancouver. Our investigation offers a deeper understanding of the factors that shape the pricing dynamics and review ratings of Airbnb listings, as well as the profitability comparison between two accommodation rental approaches, contributing to the ongoing discourse on the sharing economy and its impact on urban communities.

2 Problem Statement

In this project, we aim to address the following research questions. Based on the location and attributes of a particular property, we will try to answer whether it's more profitable as a traditional home rental or as an Airbnb. We will also investigate what are the key factors that influence the pricing and review ratings of Airbnb listings?

By addressing these two research questions, we hope to provide useful guidance for hosts to make more informed decisions. Meanwhile, by providing more transparency in the Airbnb marketplace, this research will benefit hosts, guests, and researchers, and thereby contribute to a more sustainable and equitable marketplace for short-term rentals.

3 Data Science Pipeline



The Airbnb price forecast and review forecast is generated using machine learning models, while the rent forecast is generated directly using the Craigslist dataset by calculating the average rent near the geographic coordinates of the property (see section 3.7). Detailed implementations are provided below sections.

3.1 Data Sources

We used datasets from Airbnb, Translink, City of Vancouver, Craigslist and Statistics Canada. Most of the datasets we used in this project were available directly from their providers. The rental price we used was crawled from Craigslist, we crawled 7500 rental listings in Vancouver from their website that consist of geographic coordinates, rental price and other metadata. We also conducted a small-scale data crawling from Wikipedia in the EDA phase, but the associated analysis did not yield significant results and is therefore not mentioned in this report. A description of the datasets used is given below.

3.1.1 Vancouver Airbnb Dataset

The Vancouver Airbnb Dataset provides detailed information about Airbnb listings in Vancouver, including property types, locations, prices, and host information.

3.1.2 BC Assessment

The BC Assessment dataset contains information about property tax rates and assessments in British Columbia, which enabled us to analyze the distribution of property taxes in Vancouver.

3.1.3 Postal Code Conversion File

The Postal Code Conversion File contains geographic information of postal codes, together with the BC Assessment data, we managed to match postal codes with longitude and latitudes.

3.1.4 Vancouver Cultural Spaces

The Cultural Spaces dataset contains information about cultural spaces in Vancouver, including their type, primary use, ownership, and geographic coordinates.

3.1.5 GTFS Static Data (Translink)

The GTFS Static Dataset from Translink provides geographical information about trips and stops for public transit in the Vancouver region.

3.1.6 Craigslist Rental Data

The Craigslist Rental Data was crawled from Craigslist and provides 7500 rental information in Vancouver, including geographic location, rental price, and other metadata.

3.1.7 Data Source Links

Dataset	Link
Vancouver Airbnb Dataset	http://insideairbnb.com/get-the-data/
BC Assessment	https://opendata.vancouver.ca/explore/dataset/property-tax-report/information/
Postal Code Conversion File	https://abacus.library.ubc.ca/dataset.xhtml?persistentId=hdl:11272.1/AB2/KBP0AM

Vancouver Cultural Spaces	https://opendata.vancouver.ca/explore/dataset/cultural-spaces/export/?disjunctive.type&disjunctive.primary_use&disjunctive.ownership
GTFS Static Data (Translink)	https://www.translink.ca/about-us/doing-business-with-translink/app-developer-resources/gtfs/gtfs-data
Craigslist Rental Data	Crawled

3.2 Data Preprocessing

In this section, we cleaned and integrated data from various sources, including Airbnb listings, BC Assessment, Translink Data, Craigslist Rental Data, and Vancouver Cultural Spaces. This helped us transform raw data into a format that is appropriate for analysis.

3.2.1 Data Cleaning

In our data cleaning process, we performed two main procedures: addressing missing values and outliers. While cleaning the Airbnb dataset, we found a number of missing values in their neighborhood information. To rectify this, we leveraged external crawled data from Wikipedia to match longitude and latitude coordinates and successfully imputed the missing data in the neighborhood column. Additionally, we took steps to remove outliers from all data sources to ensure the reliability and robustness of our analysis.

3.2.2 Data Integration

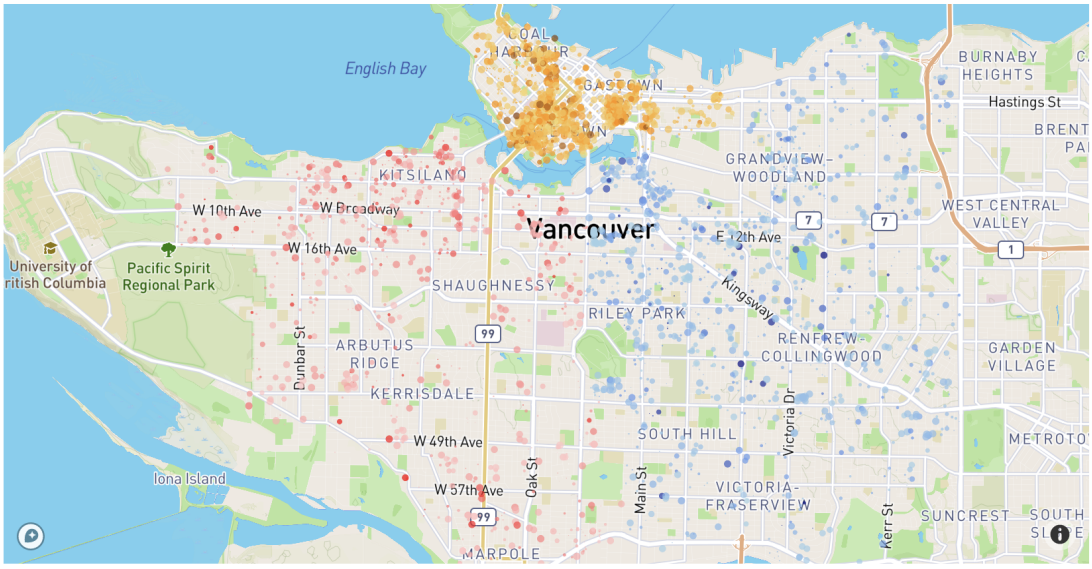
We combined five distinct datasets to provide a comprehensive analysis of the factors influencing Airbnb pricing. In order to integrate data from sources, we first attempt to identify common attributes among the datasets, yet there are no explicit common keys for merging among most of the datasets, hence we used geographic coordinates as a basis to establish connections between datasets. After identifying these connections, we integrated the datasets to create a reliable and meaningful unified dataset. This enabled us to effectively examine the various factors that contribute to the pricing dynamics and review ratings of Airbnb properties in Vancouver effectively.

3.3 Exploratory Data Analysis

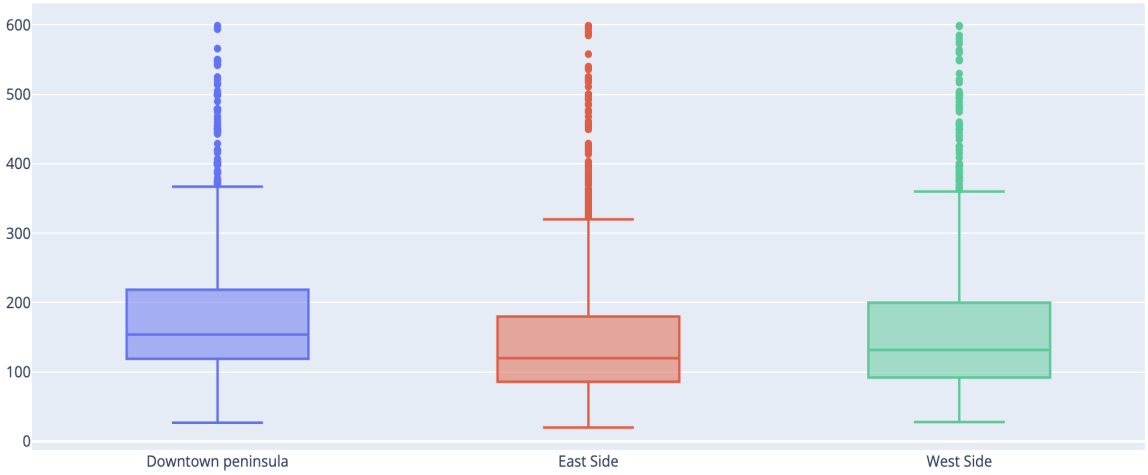
3.3.1 Density and distribution of prices

We analyzed regional pricing differences in the Vancouver Airbnb market. Thereby providing valuable insights for hosts to determine competitive pricing strategies and for guests seeking accommodations that match their budget preferences.

In our analysis, we sought to provide a price density overview at the city level by categorizing the neighborhoods into three distinct regions based on their geospatial information. This clustering approach allowed us to effectively identify and compare each region's pricing trends.



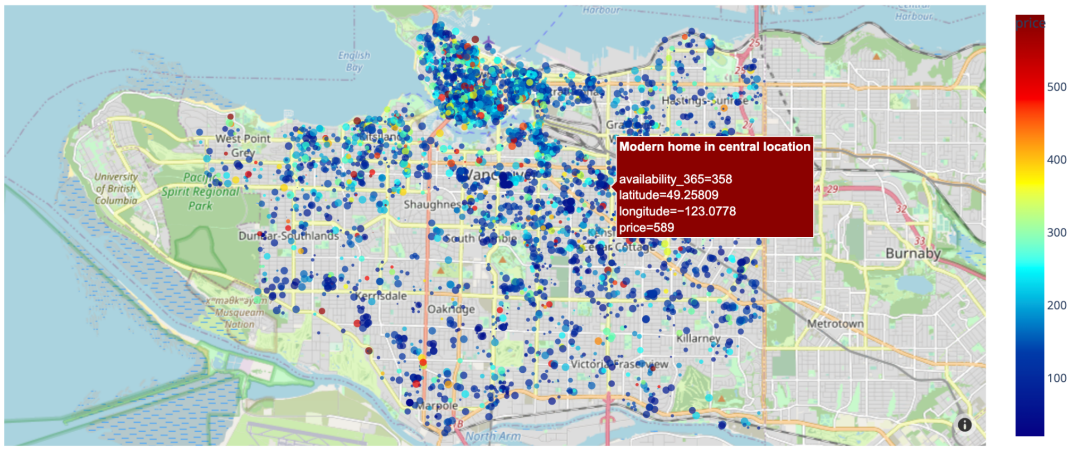
We plotted the price density distribution based on our previously clustered areas as below.



We found that the Downtown peninsula region exhibited relatively high Airbnb prices, positioning it as a more premium accommodation option for tourists and visitors. On the other hand, the East Side region demonstrated relatively lower prices, making it a more affordable option for budget-conscious travelers.

3.3.2 Geographic location and distribution of prices

In addition to our regional analysis, we generated a heat map of Airbnb listings using their latitude and longitude to visualize the spatial distribution of accommodations and pricing trends.



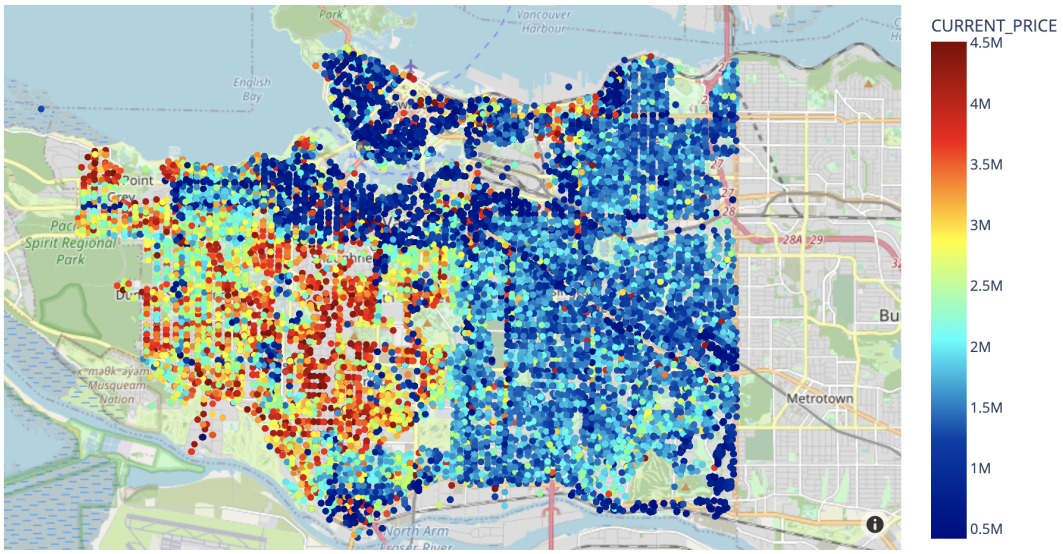
Our heat map highlighted a pattern where high-priced listings were predominantly situated near the waterfront, offering picturesque lake views and easy access to popular recreational areas.

This finding emphasizes the appeal of waterfront housing for tourists and visitors, which often commands premium pricing due to the desirable location and amenities. Hosts with properties in these areas should use their unique selling points to attract potential guests, while travelers should consider the trade-offs between price and location when choosing accommodations in Vancouver.

3.3.3 Housing Prices in Vancouver

1. Price Distribution

We generated a heat map for Vancouver housing property price using data from BC Assessment to visualize the price distribution of properties in different regions. Our heatmap highlighted a pattern where high-priced properties mainly seat in the west part of the lower mainland.



After matching with the neighborhood information, we found these neighborhoods usually have a higher property price: West Point Grey, Dunbar-Southlands, Arbutus-Ridge, Shaughnessy, South Cambie, Kerrisdale, and Oakridge.

It's noticeable that the price distribution of housing properties in Vancouver is not similar to the price distribution of Airbnb listings that we have seen before.

One possible reason is that the Airbnb market operates under different dynamics and influences than the real estate market. Airbnb listings are short-term rentals, the hosts may adjust their prices according to the demand and competition in their specific area, which may not necessarily correlate with the long-term real estate value of the property.

Another reason could be the difference in the type and quality of properties listed on Airbnb compared to the overall real estate market. Some properties with higher property values may not be listed on Airbnb, while some properties with lower values but with desirable amenities may command higher Airbnb prices.

2. Average Nearest Property Price & Airbnb Prices

To further explore the relationship between the property price and Airbnb listing price, we calculated the price of the 5 nearest properties within a 150m radius of each Airbnb listing.

The haversine formula was employed to calculate the distance between locations, and the R-tree data structure was used to reduce computational complexity.

We then performed correlation analysis on different price groups to reveal the relationship between property price and Airbnb listing price in each property price range.

We can see that although the correlation values are small, there is still a pattern with lower price range properties, the Airbnb listing price tends to have a positive correlation with the property price, and as the property enters higher and higher price ranges, the Airbnb listing price will

have an increasingly negative correlation with the property price. Furthermore, properties valued over 1 million dollars may not be economically viable for short-term rentals due to the high investment costs involved.

This observed pattern may be due to several factors:

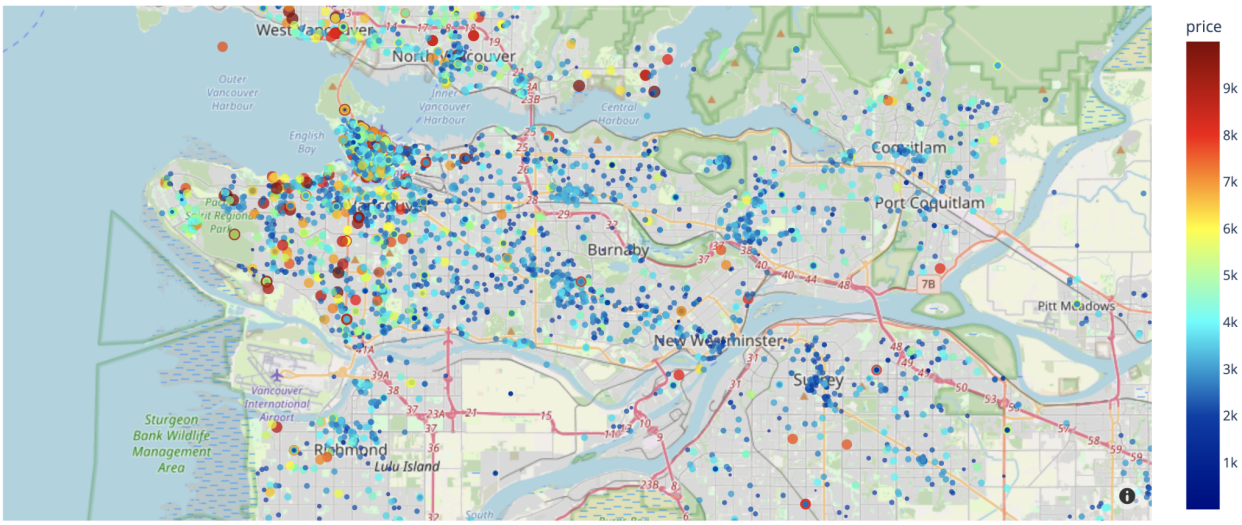
In lower price ranges, the property and Airbnb listing prices may be influenced by similar factors such as location, amenities, and demand. Therefore, as the property price increases, the Airbnb listing price may also increase to reflect the perceived value of the property.

However, at higher price ranges, properties may be purchased for investment purposes or as a luxury item, which may not necessarily translate into higher Airbnb rental rates. Additionally, in high-end markets, luxury hotels may be in a position to compete with Airbnb rentals, making high-end properties more difficult to command high rents on the Airbnb market.

This finding may also indicate that the higher the value of the property the less suitable it may be for Airbnb.

3.3.4 Rental Prices in Vancouver

We analyzed the rental prices in the Vancouver rental market using our crawled Craigslist rental data to provide insights on the profitability of traditional home rental. We have generated a heat map below for Vancouver house rental prices to visualize the price distribution of properties in different regions.



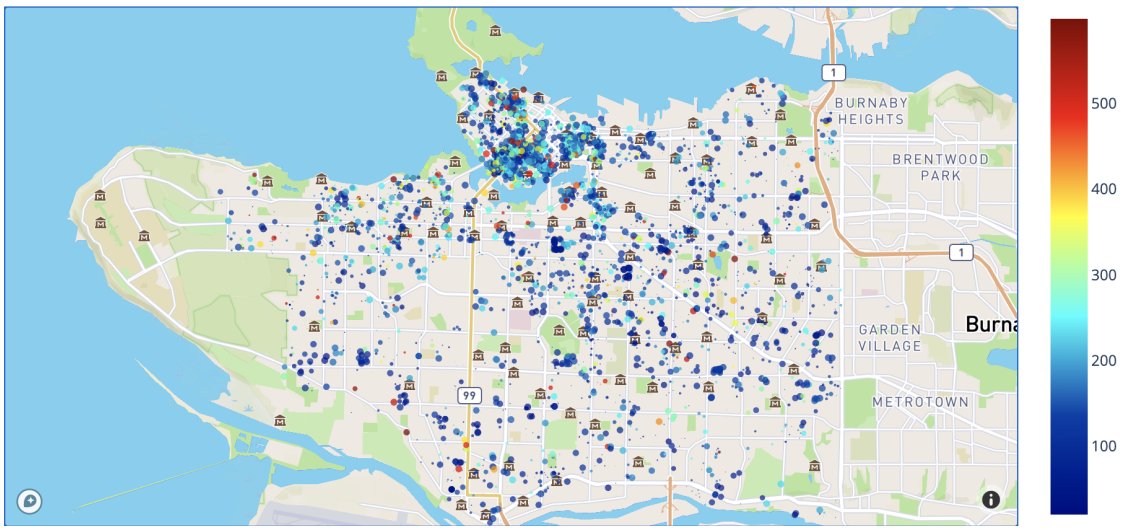
It is noticeable that the distribution of rental prices across areas is similar to the pattern of housing prices we have seen in the previous section. It appears that the neighborhood that has a higher housing price tends to have a higher rental price.

This similarity may be due to the fact that the factors that can affect both rents and home prices include common aspects like location, quality of the neighborhood, proximity to schools and public transportation, local amenities, and the demand for housing in the area. Therefore we can see that having higher home prices will also make rents more expensive.

3.3.5 Airbnb Listings and Cultural Spaces

The Vancouver Cultural Spaces dataset provides the location and other attributes of 1,942 cultural spaces in the city of Vancouver. We integrated the geographic information of the cultural spaces in Vancouver into the Airbnb dataset so that we could provide valuable insights into the relationship between cultural spaces and short-term rental prices. These geographic coordinates were later also used for feature engineering, which will be discussed in the next section.

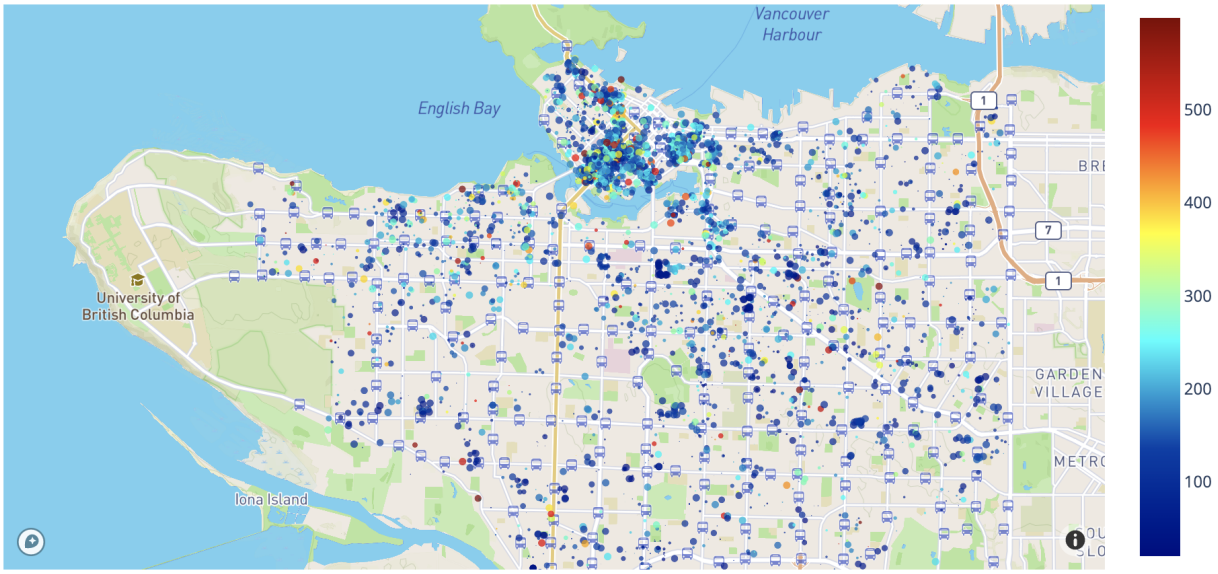
Cultural spaces, such as museums, art galleries, theaters, and historical landmarks, can attract tourists and visitors to an area, which may increase demand for short-term rentals in the neighborhood, leading to higher prices. We plotted some of the cultural spaces on the map as well as the Airbnb listings in the map below.



3.3.6 Airbnb Listings and Translink Bus Stops

The Translink GTFS Static dataset provides the location and other attributes of more than 8,000 bus stops in the city of Vancouver. We integrated the geographic information of the bus stops into the Airbnb dataset so that we could provide valuable insights into how the proximity to public transportation affected Airbnb listing prices. These geographic coordinates were later also used for feature engineering, which will be discussed in the next section.

Access to transportation could be a desirable feature that hosts can leverage to increase their prices. We plotted some of the bus stops on the map as well as the Airbnb listings in the map below.



3.4 Feature Engineering

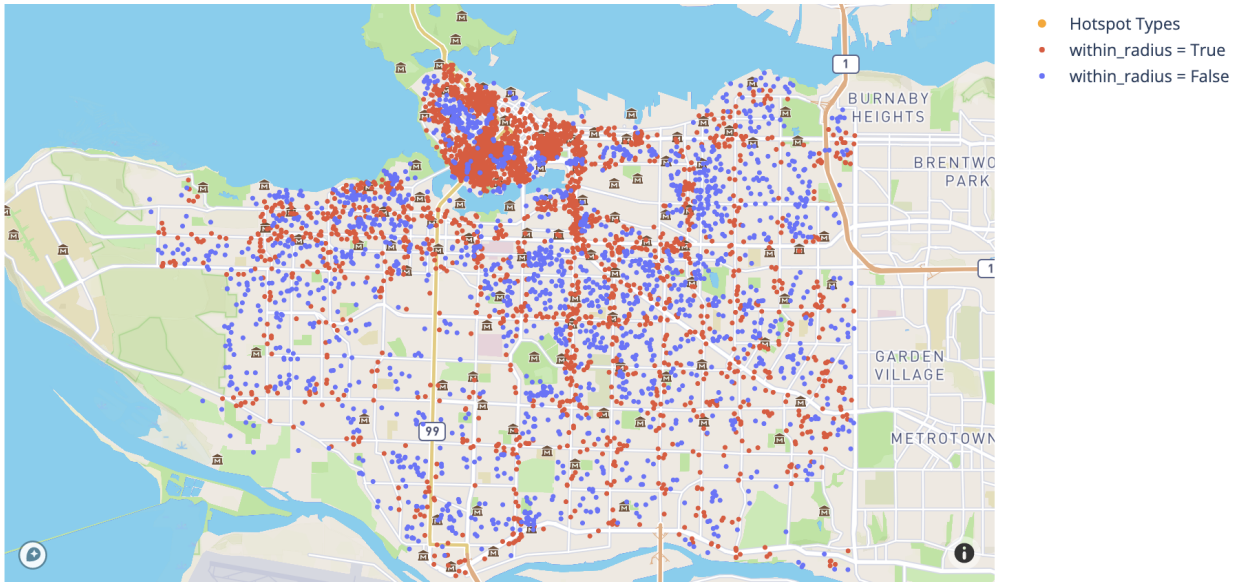
Based on our EDA results, we performed feature engineering based on the Vancouver Cultural Space dataset and Translink's GTFS Static dataset.

Due to the lack of explicit common keys among datasets, we used geographic coordinates as a basis to establish connections between datasets.

Through these feature engineering steps, we enriched our dataset with new features, allowing for more in-depth analysis and potentially improving the performance of our subsequent predictive models.

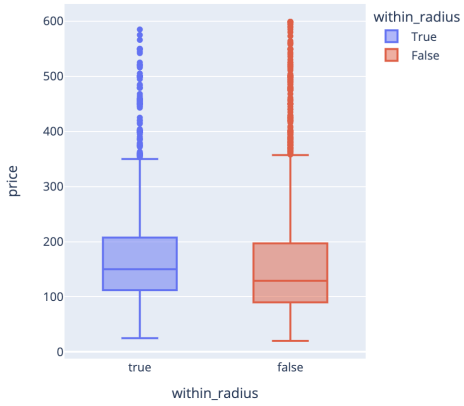
3.4.1 Proximity to Cultural Spaces

We created a new feature to assess whether a cultural space is located within a 200-meter radius of an Airbnb listing, by calculating the distance between each listing and its nearest cultural site. This approach allows us to determine the potential influence of proximity to cultural attractions on the pricing and desirability of Airbnb listings in a systematic manner.



T-test

After creating the feature, we conducted a T-test to determine the statistical significance of the feature. The purpose of this test was to assess whether the difference in the price of listings for two groups was large enough. By doing so, we aimed to validate the usefulness of the engineered feature in our predictive model.



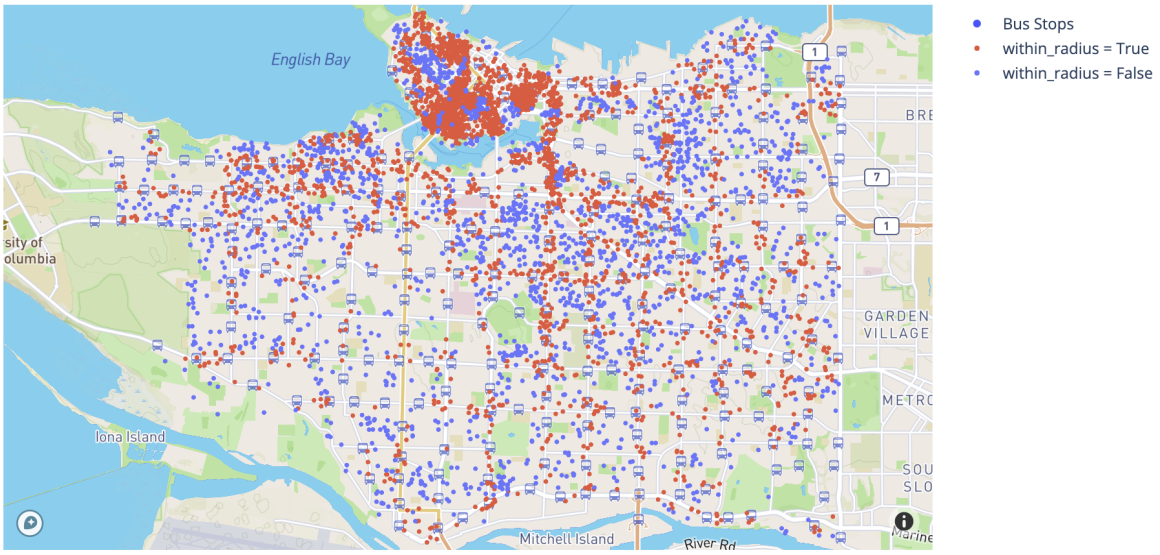

```
t-statistics: 6.057268498533037
P-value: 1.4759769407912069e-09
Reject the null hypothesis. There is a significant difference.
```

The t-statistic of 6.0573 and The p-value of 1.4759e-09 indicate that the prices of Airbnb listings differ significantly depending on the availability of cultural space within its 200-meter radius.

3.4.2 Proximity to Translink Bus Stops

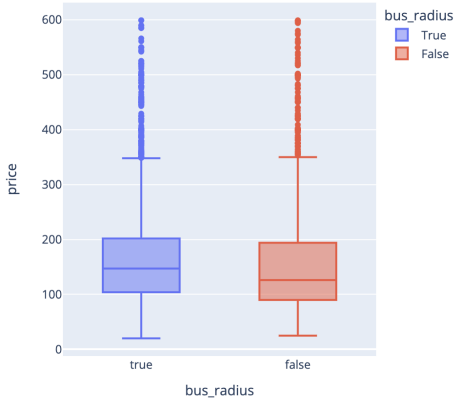
We created a new feature to assess whether a bus stop is located within a 150-meter radius of an Airbnb listing, by calculating the distance between each listing and its nearest bus stop. This

approach allows us to determine the potential influence of access to public transportation on the pricing and desirability of Airbnb listings in a systematic manner.



T-test

After creating the feature, we conducted a T-test to determine the statistical significance of the feature. The purpose of this test was to assess whether the difference in price of listings for two groups was large enough. By doing so, we aimed to validate the usefulness of the engineered feature in our predictive model.




```
t-statistics: 3.6757578650293503
P-value: 0.00023939528388292777
Reject the null hypothesis. There is a significant difference.
```

The t-statistic of 3.675 and The p-value of 0.000239 indicate that the prices of Airbnb listings differ significantly depending on the availability of bus stops within its 150-meter radius.

3.4.3 Average Price of Nearest Listings

We calculated the average price of the 8 closest Airbnb listings within a 200-meter radius of each Airbnb listing used. This feature allows us to use the prices of existing listings to guide the pricing of their neighboring listings, while the average price of eight nearby listings ensures that this feature can be extended to data that the model has not seen before, without easily overfitting during the training process.

It is worth noting that not all Airbnb listings have at least 8 nearby listings within a 200-meter radius. (i.e. this feature **cannot be guaranteed to exist** for every listing in) Therefore, we have trained **one model** that utilizes this feature for listings that have this feature. At the same time, we have also trained **another model** that does not use this feature for listings that don't have.

3.4.4 Average Review Score of Nearest Listings

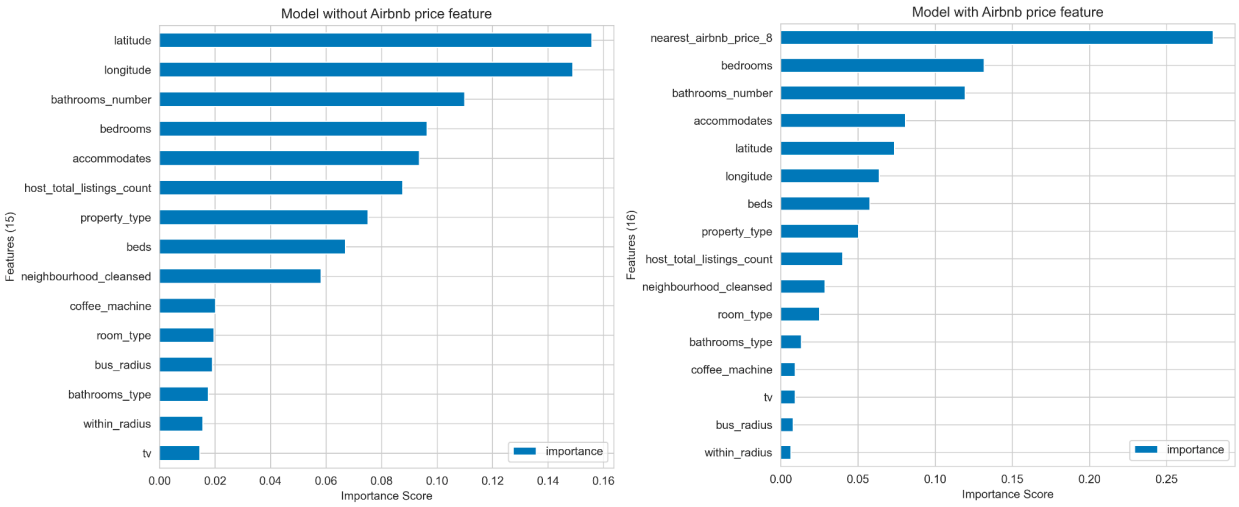
We also calculated the average review score of the 8 closest Airbnb listings within a 200-meter radius of each Airbnb listing used. This feature can reflect a general visitor's impression of the neighborhood. This allowed us to use the nearby review scores to forecast new listing review scores. In the following modeling phase, we trained models using both the dataset with and without the inclusion of this new feature.

Similar to the previous feature, this feature is **not guaranteed** to be present for every listing. Therefore, we also trained two models, **one using** this feature and **one not**.

3.5 Price Prediction Model

3.5.1 Feature Selection

We used mutual information (MI) to select the top 15 features that had the highest MI value. We plotted the feature importance for both models below to see the relative contribution of each feature toward the model's predictions. These feature importances gave us valuable insights into the significance of each feature when predicting Airbnb listing prices. We were able to create a model that only uses 15 features for the model trained without considering average nearest listings price, and 16 features for the model trained with average nearest listings prices. Both models perform better than the full model with 42 features.



3.5.2 Model Training and Selection

We trained and compared five models, namely KNN, SVR, Random Forest, Gradient Boosting Tree, and Lasso Regression using the training data with a 20 fold cross-validation, based on their Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). And we choose the Random Forest and Gradient Boosting Tree model based on their performance.

Price Models	**RMSE**	**MAE**
Radial Support Vector Machine	61.28	45.08
Random Forest	**55.17**	**42.36**
Gradient Boosting Tree	**52.91**	**39.61**
Lasso Regression	58.14	45.40
K-Nearest Neighbors Regression	56.78	42.77

3.5.3 Model Tuning

We performed model tuning through grid search over a range of hyperparameter values. We tuned the two hyperparameters of both models with all combinations of values for a total of 72 experiments. Then, we selected the best-performing parameters for both models and moved them to the evaluation stage.

3.5.4 Model Evaluation

1. Model trained without the *average nearest listings price* feature

The Random Forest model, which was trained without the nearest listings price feature, achieved a Root Mean Squared Error (RMSE) of 53.11, outperforming the gradient boosting model by a slight margin.

Additionally, the Mean Absolute Error (MAE) of the random forest model was approximately 40.16, which is also slightly lower than the corresponding error metric for the gradient-boosting tree model.

Price Models	**RMSE**	**MAE**
Random Forest	53.11	40.16
Gradient Boosting Tree	54.06	40.78

2. Model trained with the *nearest listings price* feature

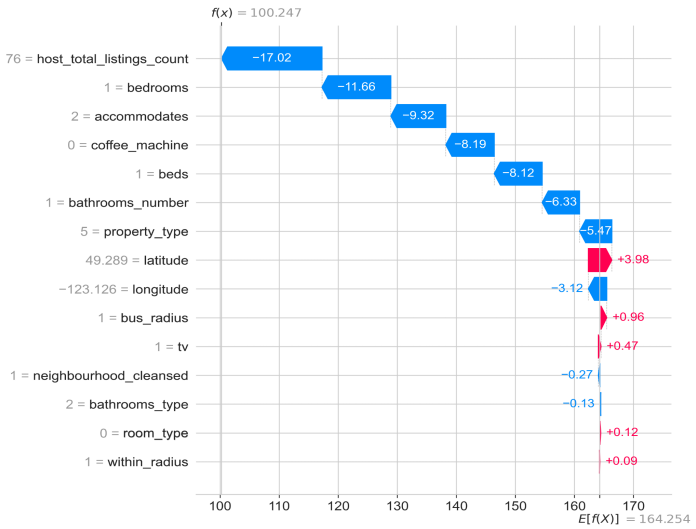
When the Random Forest model was trained with the nearest listings price feature, its Root Mean Squared Error (RMSE) was approximately 49.20, which is lower than that of gradient boosting tree, with a value of 50.55. Additionally, the Mean Absolute Error (MAE) of the random forest model was approximately 31.99, which is lower than the corresponding metric for gradient boosting tree.

Price Models	**RMSE**	**MAE**
Random Forest	49.20	31.99
Gradient Boosting Tree	50.55	33.15

3.5.5 Model Interpretability

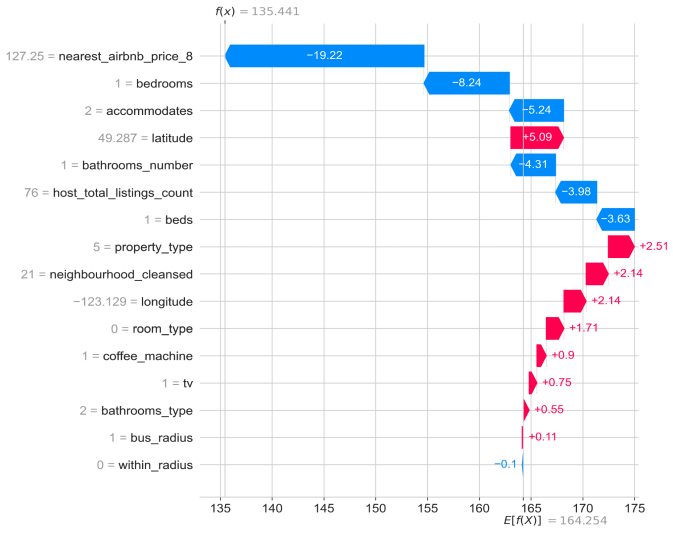
As the random forest regression model demonstrated better performance than the gradient boosting tree model, we utilized the SHAP library to interpret the model output by displaying the contributions of each individual feature to the predicted value.

Model trained without nearest listings price



For this specific instance, the predicted value is 100.247, which is substantially lower than the base value of 164.254. Host_total_listings_count is the most significant negative contributor, which could suggest that the more listings a host manages, the lower the quality of the accommodations and, therefore, the lower the prices. On the other hand, the feature latitude is the most substantial positive contributor, with a shapely value of 3.98. However, due to the combined negative contributions outweighing the positive contributions, the final predicted value is lower than the average.

Model trained with nearest listings price



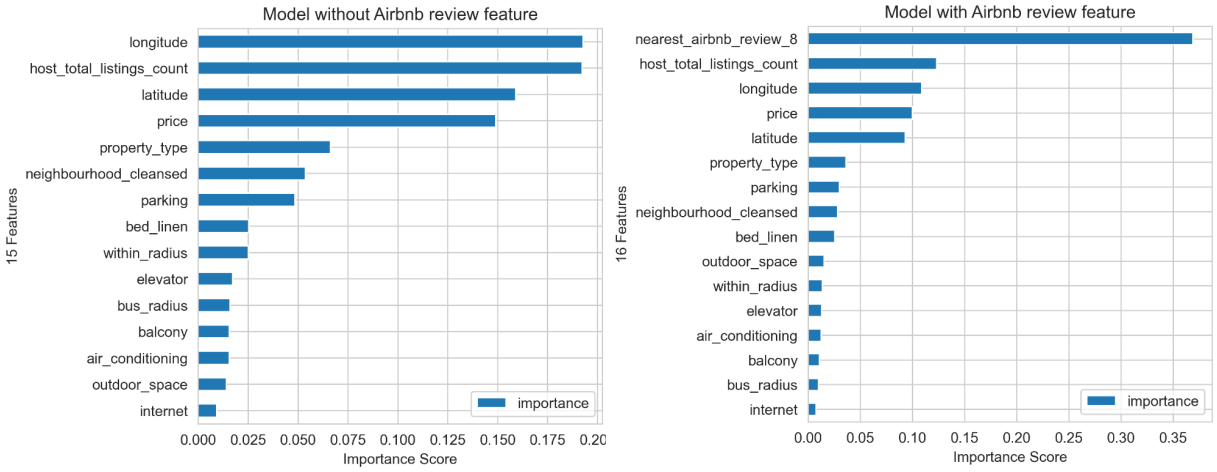
The predicted value for this instance is lower than the base value, indicating significant negative contributions. The largest negative contributor is nearest_airbnb_price_8 with a Shapley value of -19.22, and smaller bedroom number also has a negative impact. Overall, negative contributions outweigh the positive, leading to a lower predicted value. This suggests

nearest_airbnb_price_8 captures an underlying factor, as Airbnb prices are often related to nearby listings.

3.6 Review Prediction Model

3.6.1 Feature Selection

We employed mutual information (MI) to select the top 15 features with the highest MI value for both models. Below are the feature importance plots for the models, which show the relative contribution of each feature to the model's predictions. These feature importances provided useful insights into the significance of each feature when predicting Airbnb listing reviews. The final models used only 15 or 16 features, depending on whether we considered the nearest listings review score, but still outperformed the full model that used 43 features.



3.6.2 Model Training and Selection

We trained and compared the same five models: KNN, SVR, Random Forest, Gradient Boosting Tree, and Lasso Regression. We assessed their performance based on the RMSE and MAE metrics. And we choose the Random Forest and Gradient Boosting Tree model based on their performance.

Review Models	RMSE	MAE
Radial Support Vector Machine	0.35	0.22
Random Forest	**0.33**	**0.20**
Gradient Boosting Tree	**0.34**	**0.21**
Lasso Regression	0.36	0.22
K-Nearest Neighbors Regression	0.36	0.23

3.6.3 Model Tuning

To optimize our models, we conducted a grid search on both models, varying their hyperparameters over a range of values. We constructed a grid of four hyperparameters, and trained and evaluated both models with every combination of these hyperparameters, leading to a total of 72 experiments. We chose the optimal hyperparameters for each model based on their performance on cross-validation, with a focus on achieving the lowest average RMSE score and MAE. We then moved both models with their tuned hyperparameters to the evaluation phase.

3.6.4 Model Evaluation

1. Model trained without *nearest listings review* feature

The Root Mean Squared Error (RMSE) for both models trained without considering the nearest listings review score is around 0.53, which measures the average difference between the predicted and actual Airbnb listing reviews.

Similarly, the Mean Absolute Error (MAE) for both models is approximately 0.28, which is another evaluation metric that calculates the average absolute difference between the predicted and actual Airbnb listing reviews

Review Models	**RMSE**	**MAE**
Random Forest	0.53	0.28
Gradient Boosting Tree	0.52	0.26

2. Model trained with *nearest listings review* feature

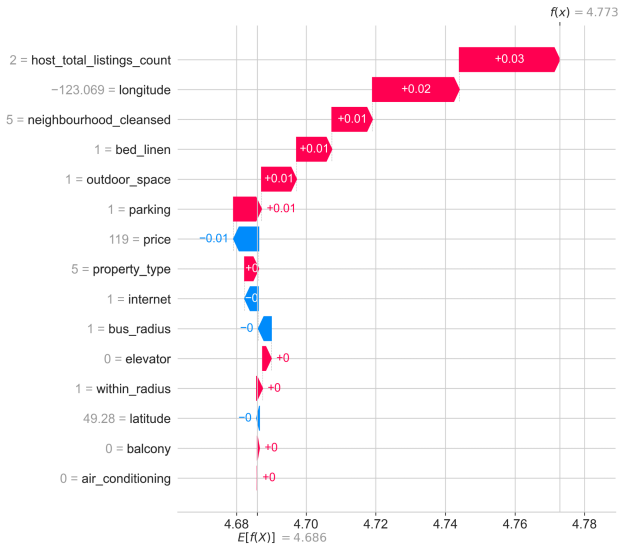
The Gradient Boosting model trained with nearest_aribnb_review_8 feature slightly outperforms the Random Forest model in terms of RMSE, with the Root Mean Squared Error (RMSE) of approximately 0.26, and the Mean Absolute Error (MAE) of approximately 0.19.

Review Models	**RMSE**	**MAE**
Random Forest	0.27	0.19
Gradient Boosting Tree	0.26	0.19

3.6.5 Model Interpretability

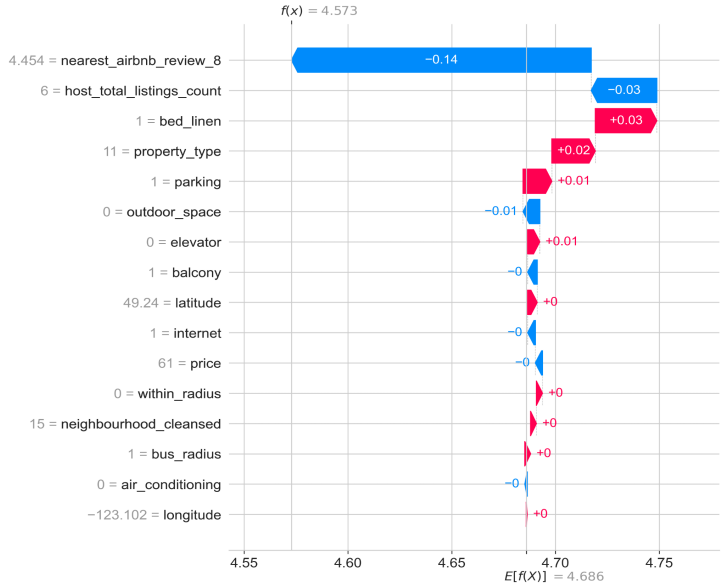
Since the gradient boosting regression model slightly outperformed the random forest model, we used the SHAP library to interpret the output of the model by quantifying feature importance with Shapley values.

1. Model trained without *nearest listings review* feature



For this instance, our model predicted a review score of 4.773, which is higher than the mean base value of 4.686. The feature with the largest positive contribution to the prediction is host_total_listings_count with a value of 2, indicating that managing two listings at the same time tend not to lead bad reviews. On the other hand, the feature with the largest negative contribution is price, with a value of 119, implying that a high price of $119 per night could result in a negative review. The positive contributions outweigh the negative contributions, resulting in the final predicted value being higher than the mean.

2. Model trained with *nearest listings review* feature



The shap.force_plot shows that the predicted value for this instance is slightly below the mean base value, indicating overall negative contributions to the prediction. The largest negative contributor is nearest_airbnb_review_8, while host_total_listings_count also has a negative impact. Bed_linen has a positive contribution, indicating its importance to the accommodation experience. Overall, the negative contributions outweigh the positive contribution, suggesting that nearest_airbnb_review_8 helps capture underlying patterns. This makes sense as reviews of an Airbnb are typically related to nearby Airbnb reviews.

3.7 Rental Prediction

The rent forecast is generated directly using the Craigslist dataset by calculating the average rent of the eight nearest rental listings within 200 meters of the geographic coordinates of the property.

Rental prediction relies on the Craigslist dataset, however, the description of each rental information on Craigslist varies greatly and it is actually difficult to extract valuable property attribute information. In fact, we did not find a rental data source that provides a large amount and complete attribute information (e.g., Zillow has relatively complete attribute information, but the amount of property data is small and difficult to crawl, while Craigslist has sufficient amount of property data, but the property attribute information is difficult to extract). We crawled almost all listings on Vancouver Craigslist to create a rental dataset as large as possible to make up for the lack of other housing properties.

We also discussed other limitations regarding the profitability comparison (between Airbnb and traditional) in Section 7 - Limitations.

4 Methodology

In our methodology, we utilized a variety of tools and analysis methods to address the various challenges presented by our project.

4.1 Data Preprocessing

We used Python's Pandas library for handling and cleaning the raw datasets. This included dealing with missing values, imputing data, and removing outliers. We chose Pandas because of its powerful data manipulation capabilities and its wide adoption in the data science community.

4.2 Data Integration

We used geographic coordinates to merge the datasets due to the lack of explicit common keys among most of the datasets. The proximity analysis based on geospatial data used the Haversine formula to calculate distances using longitude and latitude, which helped measure the proximity to transit stations or cultural spaces and also helped to locate the nearby listings.

The R-tree data structure was also largely employed to reduce the complexity of proximity computation. R-tree is a spatial data structure that is widely used to process geospatial data, it can organize spatial data in a hierarchical structure that can efficiently search for data within a specific spatial range.

4.3 Exploratory Data Analysis

We employed various visualization tools, such as Matplotlib, Plotly and OpenStreetMap, to explore the data and identify trends, patterns, and relationships between variables. This step facilitated a deeper understanding of the data and informed our feature selection process for building machine learning models.

4.4 Feature Selection

We used Scikit-learn for feature selection and all the rest of the machine learning related tasks. Mutual information (MI) was used to select the top features that had the highest MI value. MI calculates variable dependency based on entropy from k-nearest neighbor distances and does not assume linearity between variables (unlike some other methods such as F-values). Mutual information scores indicate a greater impact on the model's performance, it also provided insights into the most influential factors that affect Airbnb listing prices. By performing feature selection, we managed to half the number of features while remaining the performance, resulting in well-performing models while reducing their complexity.

4.5 Model Training and Selection

To assess the effectiveness of different machine learning models, we trained five different models using Scikit-learn with 20 fold cross-validation to obtain dependable performance

estimates. RMSE and MAE were used as performance metrics. We eventually selected a random forest and a gradient boosting tree based on their performance.

4.6 Model Tuning

We used Scikit-learn to perform grid searches to tune our models. A total of 144 (72 for price prediction models, 72 for review prediction models) experiments have been conducted. Throughout the tuning process, we have been using CPUs, and in the future, we will use GPUs for tuning to enable parallel computing.

4.6 Model Evaluation and Validation

We utilized Scikit-learn for implementing these machine learning models, and we evaluated their performance using standard metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

RMSE measures the average difference between predicted and actual Airbnb listing prices, with a lower value indicating a better model fit and a higher value indicating a larger discrepancy between the predicted and actual values.
MAE is another evaluation metric that measures the average absolute difference between the predicted and actual Airbnb listing prices. Similarly to RMSE, a lower MAE value indicates a better model fit.

5 Evaluation

The research project provides valuable insights for various stakeholders, such as property owners, guests, policymakers, and researchers, by identifying key determinants influencing Airbnb listing prices and review ratings in the Vancouver area.

The project's structure allows for a comprehensive analysis of the factors driving pricing and review ratings in the Airbnb market. Data preprocessing ensures clean and integrated data for analysis, while exploratory data analysis offers an in-depth analysis of the distribution of Airbnb listing prices, housing prices, and rental prices, and also looked into how cultural space and public transportation affected the Airbnb listing prices. This helps property owners understand regional pricing differences and make informed decisions.

Feature engineering enhances the dataset with new features, such as proximity to transit stations and cultural spaces, which can be crucial factors for guests and property owners. Machine learning models, like decision trees and random forests, predict Airbnb listing prices and rank influential factors, providing insights into the market's driving forces.

The evaluation and validation of the models using standard metrics like RMSE and MAE ensure the accuracy and reliability of the project's findings. These insights can be beneficial for

researchers studying the impact of short-term rentals on housing markets, local economies, and cultural landscapes.

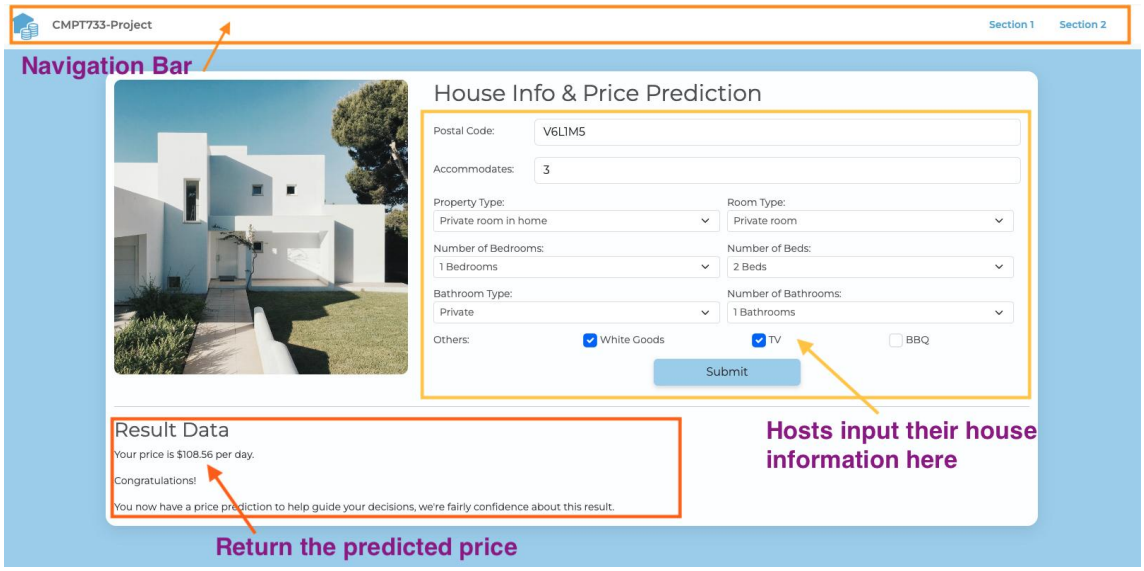
Overall, the project effectively explores the interplay between various factors that affect pricing and review ratings in the Airbnb market, making it a useful reference for property owners, guests, and policymakers. Additionally, the comparison between Airbnb and traditional rentals can guide homeowners in their investment decisions.

6 Data Product

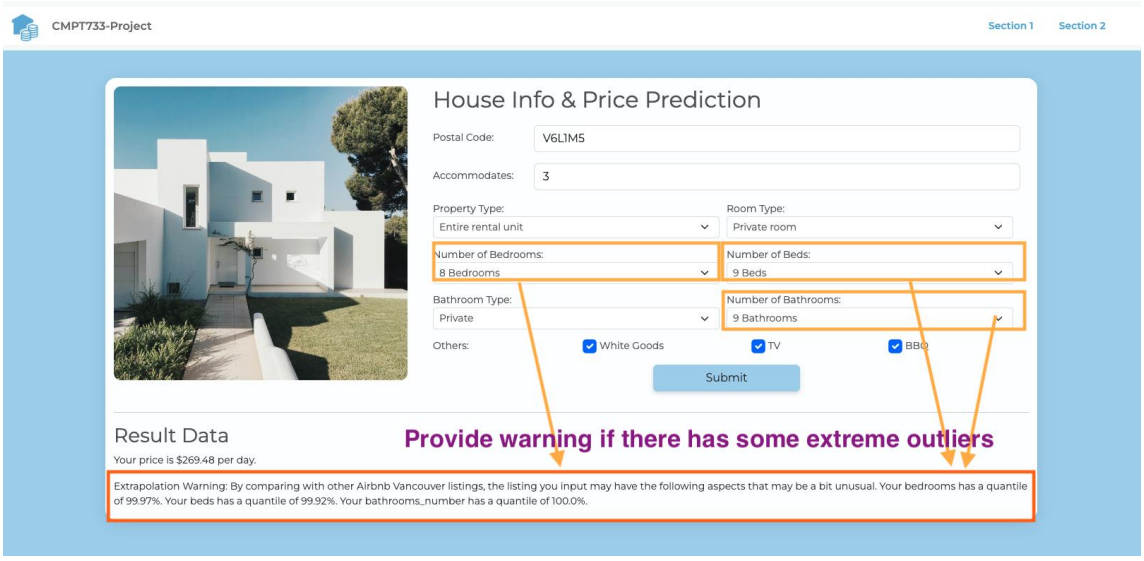
Our data product is designed to help Airbnb hosts price their listings more accurately and predict reviews based on a range of inputs. Using machine learning algorithms and data from Airbnb and other Vancouver-related datasets, our approach analyzes a wide range of factors, including location, amenities, and so on. And our aim is to provide an efficient and easy-to-use product for Airbnb hosts, as well as property managers who handle multiple listings.

To demonstrate our data product, here is a simple walkthrough, showing how a host inputs their data into the system and receives personalized pricing and review predictions.

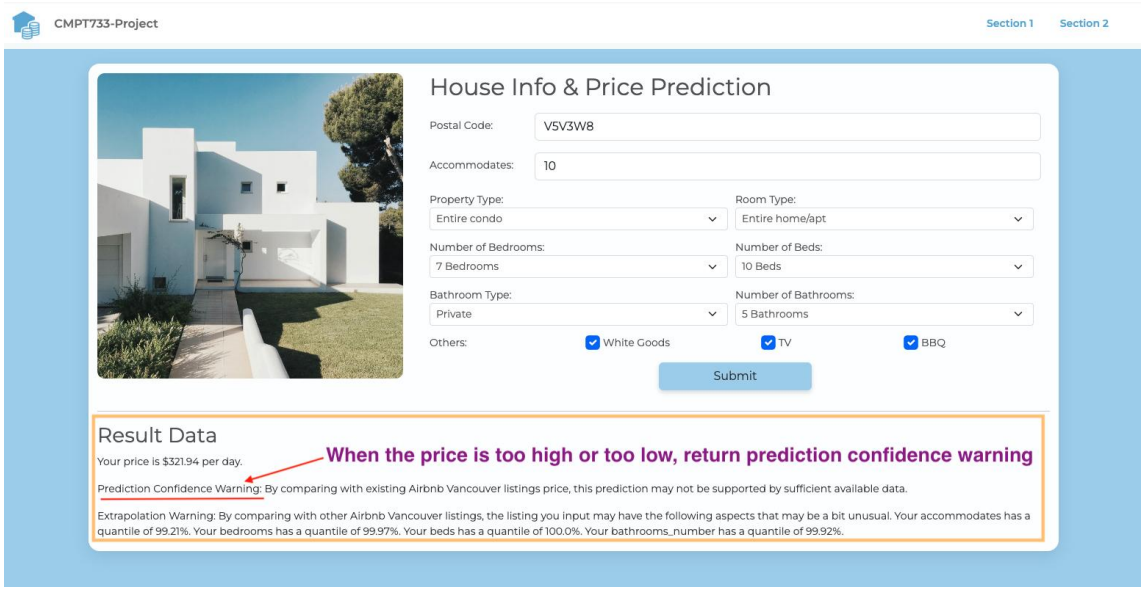
Case 1 - Price Prediction



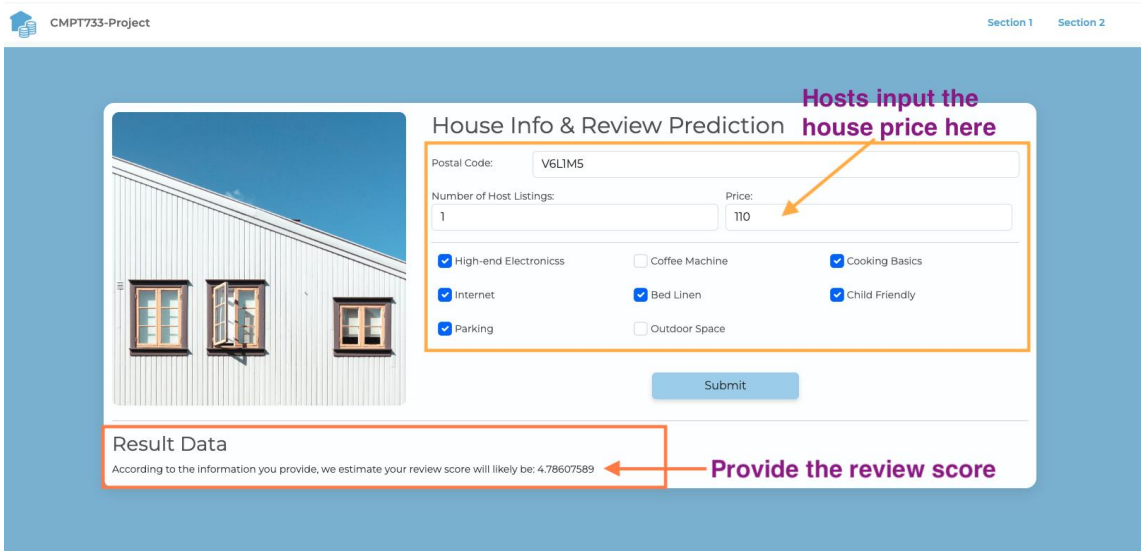
Case 2 - Price Prediction With Warning (When receiving radical input)



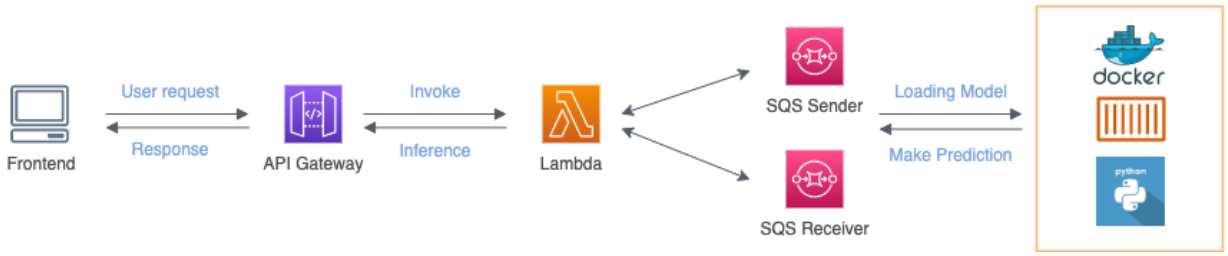
Case 3 - Price Prediction With Warning (When made aggressive prediction)



Case 4 - Review Score Prediction



Platform architecture:



7 Limitations

Although we attempted to compare the profitability of using the property for rental and for Airbnb listing, the real cost of rental properties and Airbnb listings may remain unclear. The income comparison between the two options may become unreliable due to many hidden unforeseeable factors. For example, running an Airbnb listing may involve variable operational costs like cleaning and maintenance fees, while traditional rentals usually don't involve such fees during the lease period. Another factor is the occupancy rate, the occupancy of Airbnb may fluctuate due to seasonality which may consequently affect the hosts' income, whereas traditional rentals typically guarantee a consistent income through 100% occupancy over the lease period.

8 Lessons Learnt

One of the things we learned is that when integrating data from different sources, we don't necessarily have to find a common key and merge on a one-to-one correspondence. We can

integrate data within a much coarser granularity, which will give us the opportunity to work with more data sets.

In this project, we also learned that we can analyze Airbnb listings not only from its own datasets, but also from multi-angel aspects, such as bus stops, cultural spaces, house prices, and rental rates. This approach enhances our data exploration and analysis capabilities, providing a more comprehensive understanding of the factors affecting Airbnb listings.

Another thing we have learned is the importance of continuous improvement and a growth mindset. Throughout the project, we have been updating our teaching team on our progress and have received a lot of inspiration from our professors, and we have been using this feedback constantly to improve our project to further make our project more meaningful and expand our scope.

9 Summary

Our findings emphasized the impact of location, amenities, and host profiles on Airbnb pricing, offering guidance for property owners to optimize their pricing strategies and improve guest experiences. We also highlight the distinct dynamics in the Airbnb market compared to traditional housing and rental markets, with waterfront properties and cultural spaces attracting premium pricing. Furthermore, we discovered that higher-value properties may not be suitable for Airbnb due to economic viability.

Using our research, we developed a platform to help property owners determine the profitability of traditional rental and Airbnb and also provide pricing forecast review scores for potential hosts to enhance their guest experiences. This project contributed to the ongoing discourse on the sharing economy and its impact on urban communities, offering valuable insights for various property owners.