Executive Summary: Predicting Airbnb Prices in Vancouver

1. Project Problem Statement:

The underlying question of this project is to predict the prices of Airbnb listings in Vancouver accurately. By applying data science techniques, the goal is to understand the factors influencing rental prices and develop a model that adds business and societal value. This project aims to benefit both hosts and travelers in the competitive short-term rental market by providing insights into optimal pricing strategies and informed decision-making.

2. Background on the Subject Matter Area:

With the rise of the sharing economy, Airbnb has become a popular choice for travelers seeking affordable and unique accommodations, and for hosts looking to monetize their properties. Previous attempts to address this problem have been made, but advancements in data science and machine learning allow for more accurate predictions and deeper insights into pricing trends. In addition, those attempts were in different cities.

3. Details on Dataset:

The dataset used for this project was sourced from "http://insideairbnb.com/vancouver/," providing information on various attributes of Airbnb listings in Vancouver. The dataset includes features like property type, location, availability, review scores, amenities, and prices. Insight Airbnb collects this data to facilitate public discussion. It consists of 5975 rows and 75 columns. However, it is essential to note that there is no historical data available, and the dataset only includes information on listings present on the Airbnb website in the last year.

4. Summary of Cleaning and Preprocessing:

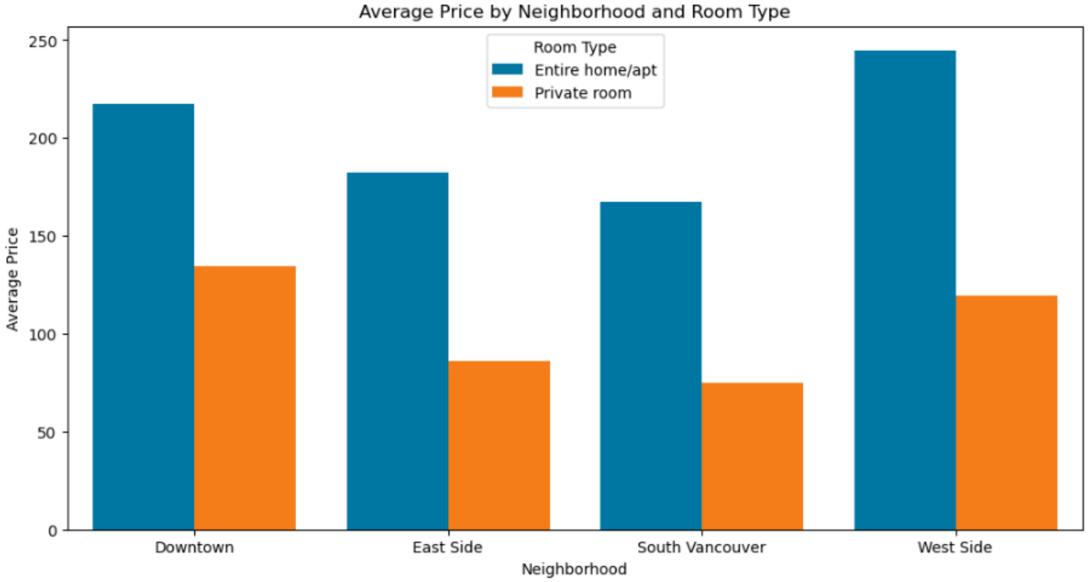
Data cleaning and preprocessing were crucial to ensure dataset quality and optimize model performance. Columns with a large number of missing values or unclear metadata were dropped to avoid complexity. The remaining columns were transformed, when necessary, into numeric format using one-hot encoding or order labeling.

To deal with missing value, first, we did feature engineering. For example, the column ‘bedrooms’ was populated from the free text ‘description’ by using regex. Secondly, we did some research. Finally, in the absence of another solution, we applied a mean.

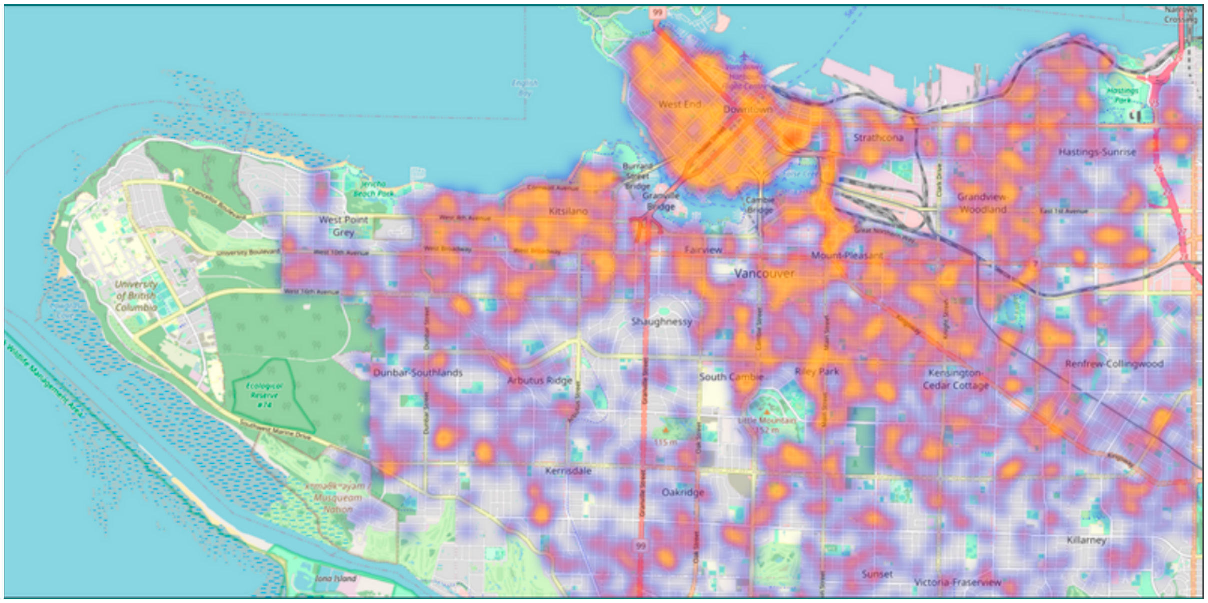
To handle outliers, the 3-sigma rule was applied, reducing their influence on model training. Logarithmic transformations were used to improve data distribution and reduce skewness in the price distribution.

To address multicollinearity, a threshold of -0.7 or 0.7 was used to determine high correlation between independent variables. In cases of high correlation, the correlation with the dependent variable (log price) was compared for each pair of correlated variables. The variable with the lowest correlation to log price was removed from the dataset. However, despite this step, further tests revealed that multicollinearity was still present in the remaining variables.

Exploratory data analysis (EDA) uncovered insights on Airbnb prices in Vancouver, influenced by factors like property location and number of rooms. Downtown had the highest number of listings and prices for private bedrooms, while the West Side had the most expensive entire homes on average.



We also discovered that a majority of the Airbnb is along the seawall from Downtown to Kitsilano, which is a tourist hotspot.



Given the presence of multicollinearity and limited linear relationships between the independent and dependent variables, we anticipate that traditional linear regression may not yield satisfactory performance in our prediction task. Therefore we used alternative modeling techniques to ensure accurate predictions in our regression task.

5. Insights, Modeling, and Results:

The XGBoost model emerged as the best performer, achieving the lowest Mean Squared Error (MSE) and the highest R-squared value of 64%. Hyperparameter tuning had a noticeable impact on some models, with the decision tree model showing significant improvement, while the random forest model already performed well with default parameters. In addition, other modeling techniques were also explored, including K-Nearest Neighbors (KNN), Linear Regression, and Ridge Regression.KNN.

6. Findings and Conclusions:

The results of this project align with our initial goals, as the XGBoost model demonstrated excellent predictive performance. However, predicting prices for two distinct types of properties (entire home/apt and private room) posed challenges, reflected in the R-squared value not being higher. Despite this, the insights gained from the feature importance shed light on key factors affecting prices, such as the number of bathrooms, accommodation capacity, and availability. Each additional bathroom in an Airbnb listing is associated with an approximate 8.73% increase in the predicted price. The number of guests an Airbnb listing can accommodate plays a crucial role in determining the predicted price. For each extra guest the listing can accommodate, there is an estimated 7.37% increase in the predicted price.. The availability of the listing for the next 30 days also impacts the predicted price. Increasing the availability by one unit leads to an estimated 1.29% increase in the predicted price.

The presence of some amenities has also notable effects on the predicted price. Having a dishwasher in the Airbnb listing is linked to an approximate 2.45% increase in the predicted price. Providing shampoo as an amenity in the Airbnb listing leads to an estimated 1.49% increase in the predicted price. The presence of a refrigerator in the Airbnb listing is associated with an approximate 1.10% increase in the predicted price. A dedicated workspace in the Airbnb listing leads to an estimated 0.91% increase in the predicted price, assuming all other factors remain constant.
All approximate increases are assuming all other factors remain constant.

Understanding how features influence the predicted price can be valuable in making informed decisions about pricing and booking an Airbnb listing.

7. Practical Applications and Future Directions:

The practical value of this project is evident in its ability to assist both hosts and travelers in the Vancouver Airbnb market. Hosts can optimize pricing strategies based on key features, and travelers can make more informed decisions when booking accommodations. The project's results can also be leveraged by stakeholders in the tourism industry to understand rental trends and make data-driven decisions.

Future steps may involve expanding the dataset to include historical data, refining the model to handle different types of properties more effectively, and incorporating neighborhood-specific trends for more accurate predictions. Additionally, exploring advanced modeling techniques and incorporating external factors like local events and seasonal trends could further enhance predictive performance.

In conclusion, this project highlights the power of data science in understanding Airbnb pricing patterns and offers valuable insights for both hosts and travelers. By leveraging advanced modeling techniques, this project contributes to a more transparent and efficient short-term rental market in Vancouver, benefiting the entire Airbnb community.