Price Prediction & Demand Forecasting for AirBNB Listings

Group Name: G

Group Members:

First name	Last Name	Student number
Dev	Makwana	C0885064
Joel	Crasto	C0883863
Krina	Patel	C0886861
Mahaveersinh	Chauhan	C0884854
Trushna	Patel	C0886910

Submission date: 14/08/2023

Table of Contents

Abstract

The world has undergone a transformative shift in the world of travel and accommodation in recent years and Airbnb has been a prominent player in this online marketplace for renting out homes/villas/studios. Millions of travelers have connected through this medium with unique lodging experiences in over 220 countries and regions. Airbnb has revolutionized the way people plan and experience their trips with a diverse range of listings and prices. In this project, we explore the world of Airbnb data analysis with the goal of helping people make wise decisions when selecting their ideal holiday rental. The goal of this project is to give people a better economic experience when they're making travel and holiday plans. To help visitors choose the best lodging that matches their interests and requirements, we aim to give a thorough analysis of the Airbnb listings, considering elements like amenities, reviews, and locations. To achieve this goal, we have cleaned and processed the Airbnb data, ensuring that the insights derived from our analysis are accurate and reliable. We recognize the significance of data integrity, as it forms the foundation upon which our recommendations and suggestions rest. To accomplish this, the data cleaned and processed the Airbnb data, ensuring the reliability and accuracy of the conclusions drawn from our analysis. We are aware of the importance of data integrity because it serves as the core basis for our suggestions and recommendations. We painstakingly cleaned and preprocessed the dataset, handling missing values, deleting duplicates, and standardizing the data formats to ensure consistency and relevancy. With useful insights into Airbnb listings, our research aims to enhance traveler's experiences by providing a predicted price based on certain predictors such as room type, accommodation, bedrooms, beds, review scores, minimum nights of stay, and so on. We want to improve the overall travel experience and ensure that every visitor gets an approximation of the spending based on their taste in their ideal location.

Introduction

In the era of technological advancements, data-driven insights have become one of the most important tools for making decisions across various industries. The realm of travel is one of the industries, wherein platforms like Airbnb revolutionize how people explore and experience travel destinations worldwide.

Our project is a comprehensive journey through the entire data analysis pipeline, aiming to provide users with well-curated and reliable insights into Airbnb listings. We prepared the dataset by dropping irrelevant columns, handling null values, and identifying and removing outliers. This process ensures that our analysis is based on a clean and robust foundation, enabling us to derive meaningful conclusions and recommendations for the users. Feature engineering is one of the most important aspects of our project, as we manipulate and transform raw data into informative representations with accuracy. Through techniques like the label and one-hot encoding, we converted categorical variables into numerical formats, making it easier to use these variables in predictive models. This transformation enhances the overall performance and accuracy of our analysis. Textual data is significant in Airbnb reviews and descriptions, as it contains valuable information about user experiences. To extract meaningful insights from textual data, we implemented Principal Component Analysis (PCA) for textual columns. By reducing the dimensionality of the text data, we can capture the most relevant information while reducing noise, making it suitable for analysis and visualization.

In our project, we utilize various visualizations, such as heatmaps and bar graphs, to present key patterns and relationships within the data as they help us identify correlations between variables allowing easy comparison and understanding of categorical distributions. Feature scaling is another essential step in our analysis, where we standardize numerical features to ensure fair comparisons and contributions from different attributes leading to more accurate predictions. To enhance the predictive capability of our analysis, we employ an Extra Tree Classifier enabling us to identify the most important features influencing Airbnb recommendations and provides valuable insights into the key drivers behind user preferences. In conclusion, our project embarks on a comprehensive exploration of Airbnb data by extracting and interpreting meaningful patterns from the data, to offer users tailored recommendations for their ideal Airbnb accommodation.

Methodology

The goal of an Airbnb dataset for price prediction and demand forecasting is to develop accurate models that can predict the prices of Airbnb listings and forecast the demand for those listings over a specific time period. This type of analysis is valuable for both hosts and guests, as it can provide insights into optimal pricing strategies and help travelers make informed decisions about when and where to book accommodations. The basic steps implemented for creating the project is outlined below:

1. Dataset Collection and Details

In the pursuit of unraveling the dynamics of the Canadian hospitality landscape within the context of Airbnb listings, a comprehensive dataset was meticulously curated from the esteemed Inside Airbnb platform. This repository, renowned for its robust and granular data offerings, served as the primary source for extracting insights that are central to this study.

The dataset encapsulates a rich needlepoint of information spanning seven prominent Canadian cities, each contributing a distinct thread to the larger narrative. The cities under scrutiny are Toronto, Winnipeg, New Brunswick, Vancouver, Victoria, Quebec, and Montreal – representing a diverse cross-section of geographic, cultural, and economic facets within the nation.

COLUMN	**DESCRIPTION**
id	Airbnb's unique identifier for the listing
name	Name of the listing
host_response_time	The average time taken by the host to respond to inquiries from potential guests
host_response_rate	The percentage of inquiries to which the host responds
host_acceptance_rate	The percentage of booking requests accepted by the host
latitude	The latitude coordinate of the property's location
longitude	The longitude coordinate of the property's location
property_type	The type or category of the property (e.g., apartment, house, etc.)
room_type	The type of room offered (e.g., entire home, private room, shared room)

COLUMN	DESCRIPTION
accommodates	The maximum number of guests the property can accommodate.
bathrooms	The number of bathrooms available in the property
bedrooms	The number of bedrooms available in the property
beds	The number of beds available in the property
amenities	The list of additional amenities provided with the property
price	The cost of renting the property
maximum_nights	The maximum number of nights allowed for booking
availability_30	The number of days the property is available for booking in the next 30 days
availability_60	The number of days the property is available for booking in the next 60 days
number_of_reviews	The total number of reviews for the property
review_scores_rating	The overall rating score given by guests in reviews
instant_bookable	Indicates whether the property can be instantly booked without host approval

Table 1 Dataset columns and description

2. Analysis Process

I. Data Cleaning

The dataset from Airbnb contains many null values that must be treated to get better results. Per the discussions and mutual decision, we removed them to avoid any incorrect predictions. Following are the steps that we followed as a part of Data cleaning:

- **Drop Columns:** To streamline the dataset and focus only on relevant information, we dropped columns that provided little or no valuable insights. Columns with a high amount of missing data or which were irrelevant to our analysis being done, such as picture_url, neighbourhood_group_cleansed, calendar_updated, and license were removed. Additionally, to avoid noise in the data, we dropped columns such as property_type so as to improve the prediction ability of the model.

- **Missing Values:** Handling missing values is crucial for ensuring the dataset's integrity and accuracy. During our data preprocessing, we encountered data (columns) that were irrelevant to the analysis being conducted. We identified columns with missing data, including listing_url, source, scrape_id, last_scraped host_thumbnail_url, and more.

- **Imputing Missing values from Name of Listings:** To enrich and clean the dataset further, we modified information from the names of the listings to impute missing values in related columns. We extracted relevant details from the listing names to impute the property_type column. This approach proved useful to fill in missing values and enhance the dataset's completeness and maintain the integrity of the data.

- **Converting String Values to Numeric:** To enable numerical analysis and modeling, we transformed categorical columns, such as host_response_rate, host_acceptance_rate, price into numeric representations for easy understanding of the data. We employed label encoding and one-hot encoding to implement it effectively on the data. For ordinal categorical variables, host_response_time, room_type, has_availability, instant_bookable we mapped them to numerical values using label encoding.

- **Detect and Remove Outliers:** Detecting and handling outliers is essential for maintaining the accuracy, integrity of the data and helps in better performance of the model. We identified outliers in numerical columns, such as price, using statistical techniques like interquartile range (IQR). Through visualizations like boxplots, we gained insights into data distribution and identified potential outliers in column instant_bookable which affected the price of the listing. We then removed outliers to avoid skewing the analysis and ensure robust model predictions to better predict the model.

II. Exploratory Data Analysis

a. **Correlation Matrix of our Data:** The correlation coefficients between several variables are shown in a table called a correlation matrix. Its value typically ranges from -1 to 1.

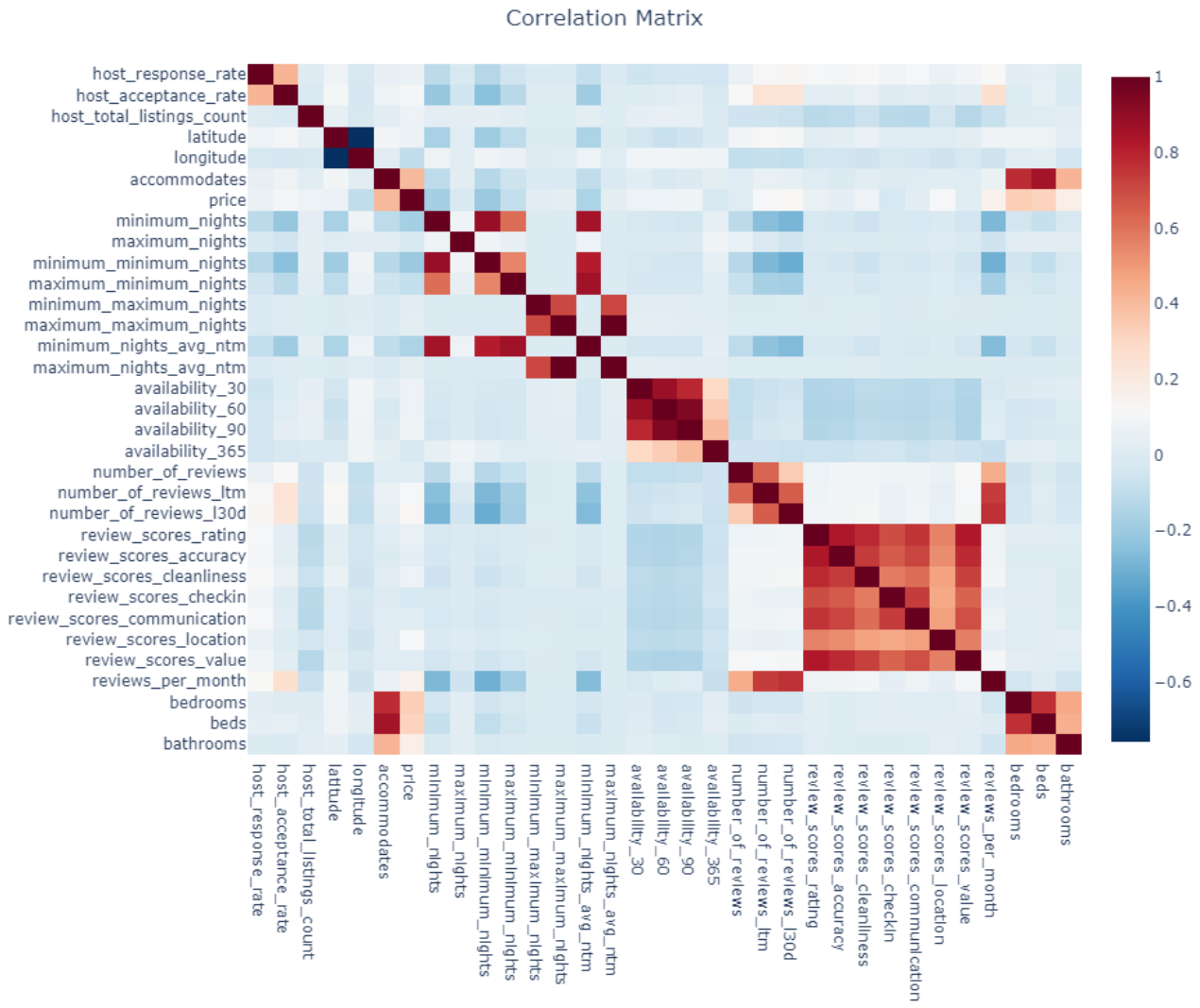

Figure 1 Correlation Matrix of the Dataset

Based on the correlation matrix of our dataset, following are the correlated features:

Correlated Features
Host_response_time - Host_response_rate - Host_acceptance_rate
Price - Accomodates - Room_Type
Bathroom - Bedroom - Beds
Miinimum Nights - Maximum Nights - Minimum Minimum Nights - Minimum Maximum Nights - Maximum Minimum Nights - Maximum Maximum Nights
Availability_30 - Availability_60 - Availability_90 - Availability_365
Number_of_reviews - Number_of_reviews_l30d - Number_of_reviews_ltm
Review_score_rating - Review_score_accuracy - Review_score_loneliness - Review_score_checkin - Review_score_communication - Review_score_value - Review_score_loaction

Table 2 Correlated Features from the dataset

b. Impact of Room Type on Price: There are more listings for entire home and private rooms as compared to hotels and shared rooms. However, there are only two box components for hotel as room type, one has the highest price and the

other ranges between around 100 to 300 dollars. Whilst there exists many box data points for shared room but they mostly lie between the price of 30 to 130 dollars, which makes it a little lower than private room prices being obvious as these rooms are shared**.** Mostly, the listings corresponding to entire home/apt and private rooms there are many outliers which we have removed later on.

Figure 2 Impact of Room Type on Price

c. **Price affected by Host Response Time:** It is observed that the higher listings have quick responses. Hoevere, there are many outliers that are removed later on.

Figure 3 Price affected by Host Response Time

d. **Price affected by Instant Booking:** The graph states that the higher prices make higher chances of instant booking as these might be the favorable places and listings of most of the customers.

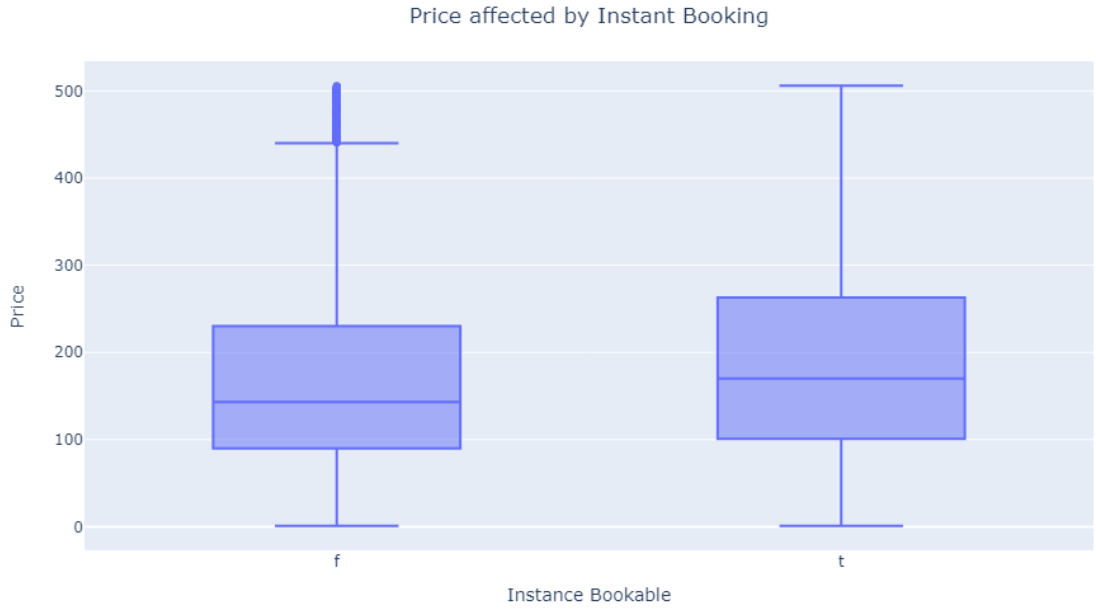

Figure 4 Price affected by Instant Booking

e. **Trend in Price based on Availability for 30 Days:** It is noted that as the availability ranges from 1 to 20 days, the price is observed to be increased with minor fluctuations

while the availability range from 20 to 30 days noted significant decrease in prices for most of the cities except for Winnipeg and New Brunswick.

Figure 5 Trend in Price based on Availability for 30 Days

f. Total listings corresponding to different cities: The following pie chart describes the distribution of listings corresponding to different cities. Toronto had the highest number of listings , followed by Montreal and Vancouver whilst Quebec and Winnipeg has the least counts.

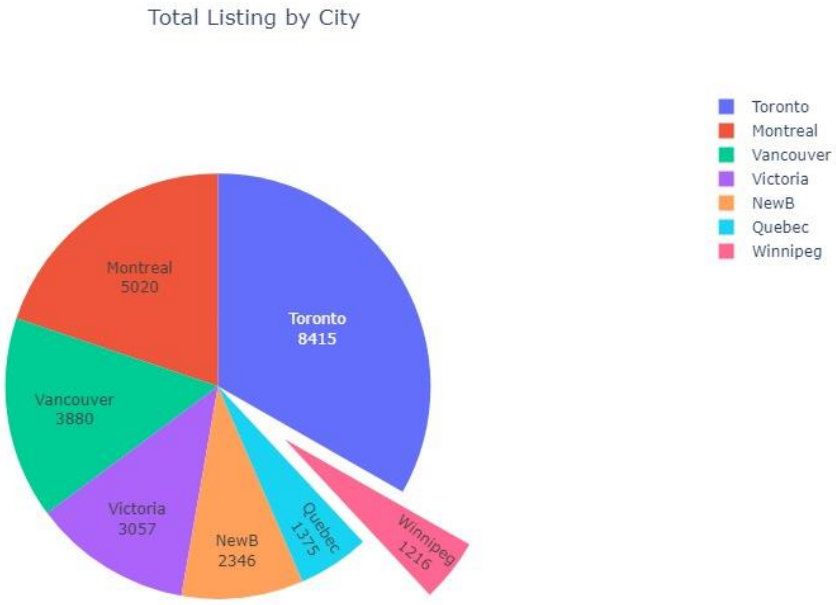

Figure 6 Total listings corresponding to different cities

g. Average of all Ratings by City: The average rating for listings in different cities was calculated and is around 4.6 for all cities.

Figure 7 Average of all Ratings by City

h. Monthly Average Pricing in cities: The following stacked bar graph plots the prices for all months of the year for all cities. The prices are at their peak in the months of summer (i.e., May, June, July, August) while its comparatively lesser in other months.

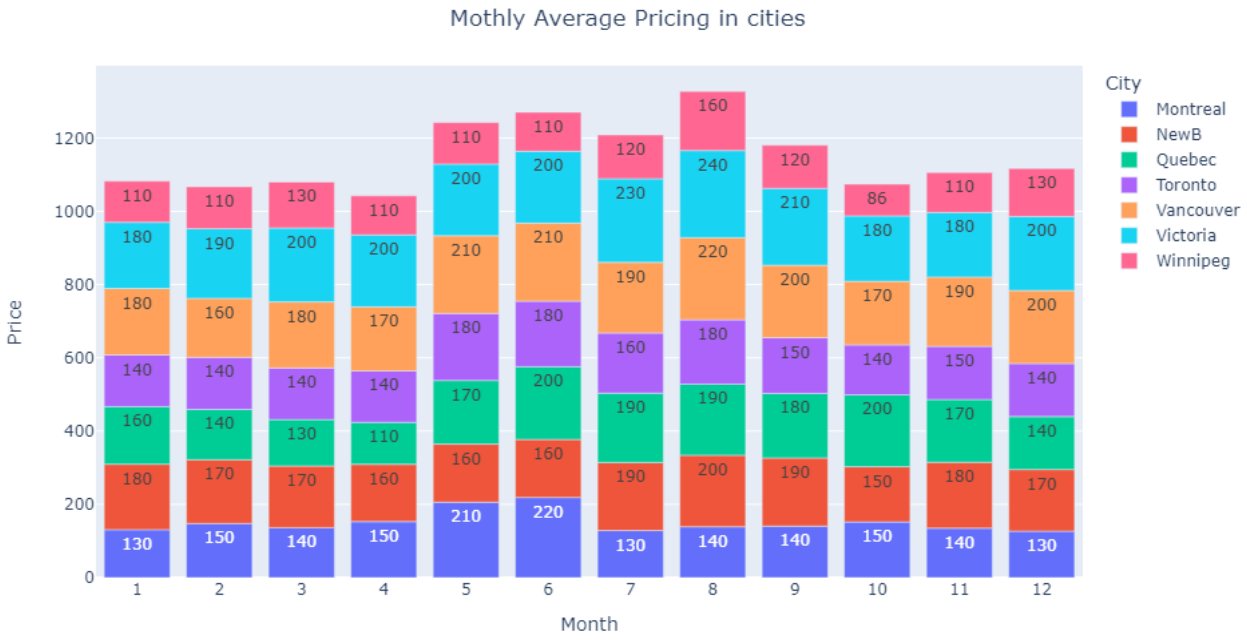

Figure 8 Monthly Average Pricing in cities

i. **Price Distribution over the Seven Cities:** Below graph plots the count of specific prices for different cities. Most prices are in the range of 0 to 200 with 100 dollars being the most listings for all cities.

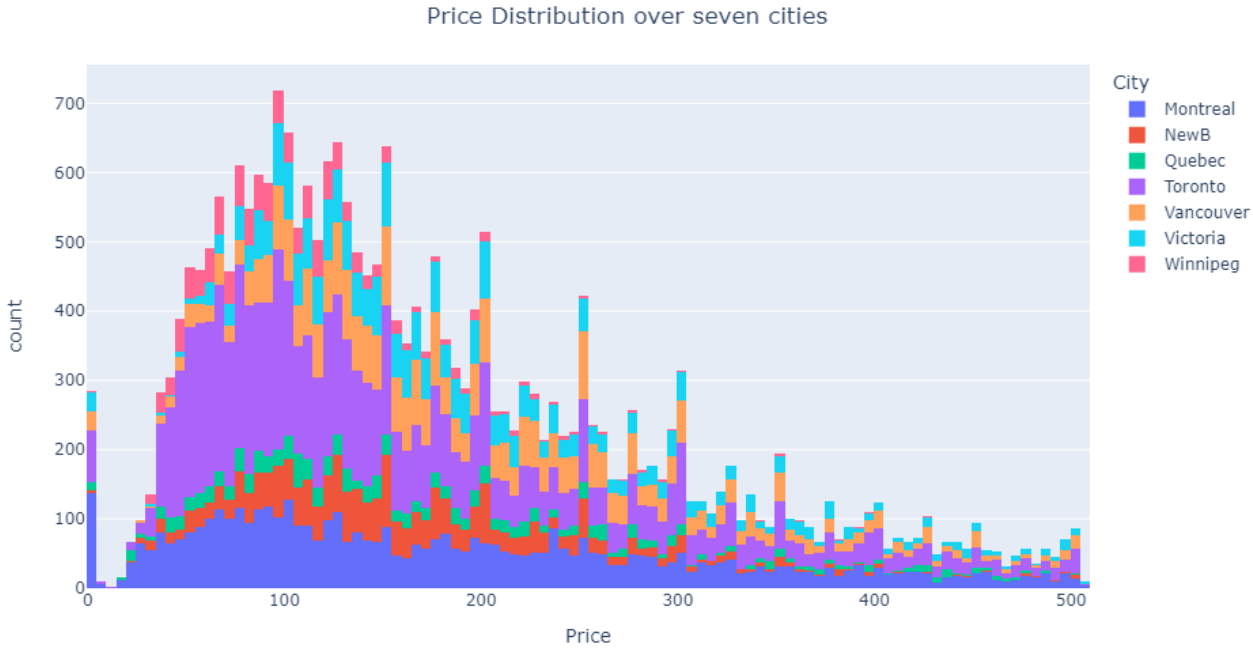

Figure 9 Price Distribution over the Seven Cities

j. **Based on Latitude and Longitude, distribution of all Listings on World Map:** Following map shows the distribution of listings with its prices at different places

based on the longitudes and latitudes on world map. The listings that are facing the sea or beach are more costlier than those inside the city.

Figure 10 Based on Latitude and Longitude, distribution of all Listings on World Map

k. **Average of all Price Variations according to the Type of Property:** Average price for the Private rooms is very less as compared to those where the entire property and hotels are listed.

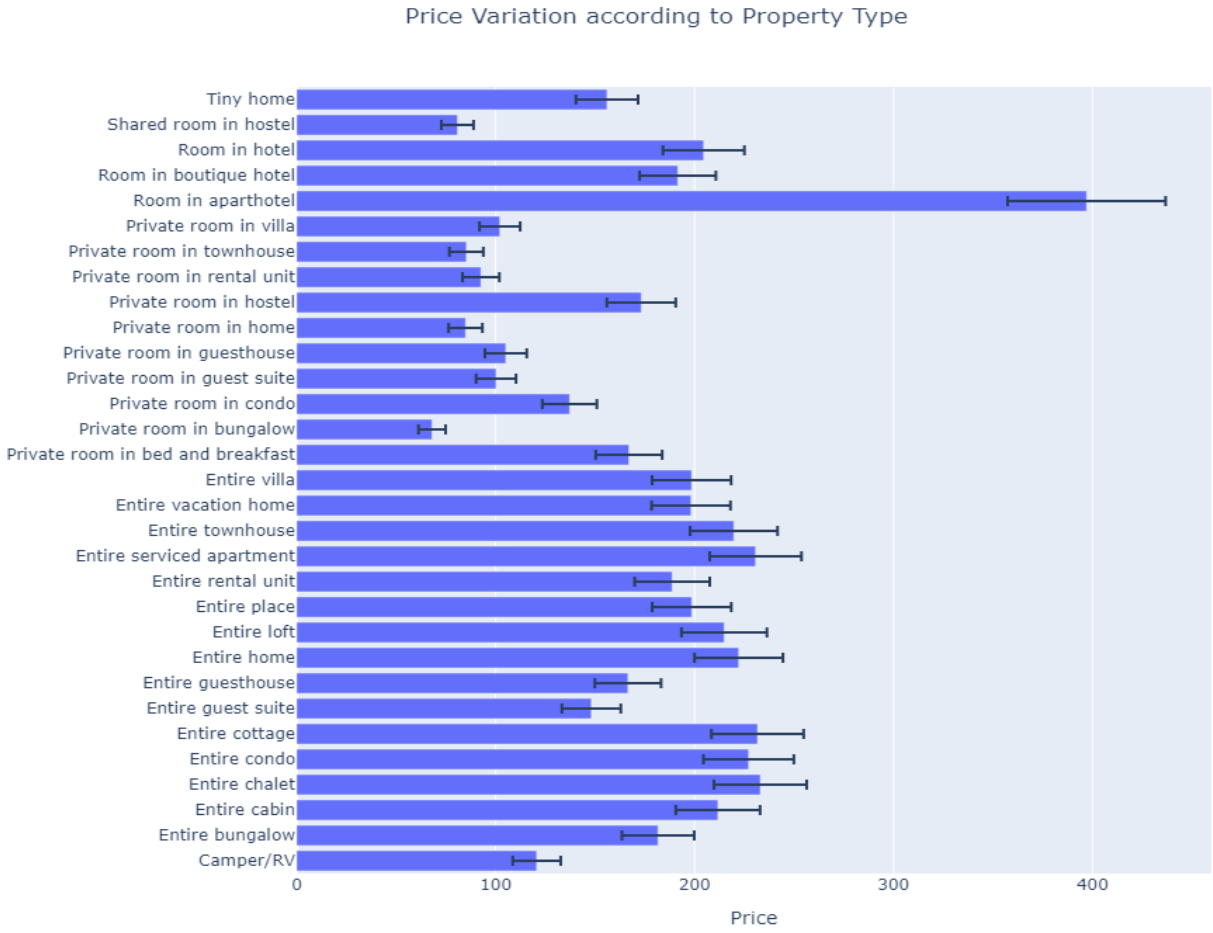

Figure 11 Average of all Price Variations according to the Type of Property

l. **Count of Common Amenities available in most of the Listings:** The amenities such as smoke alarm, kitchen, dishes and essentials are the most common amenities available in most of the listings. While dedicated workspace, self-check-in and heating are comparatively less common.

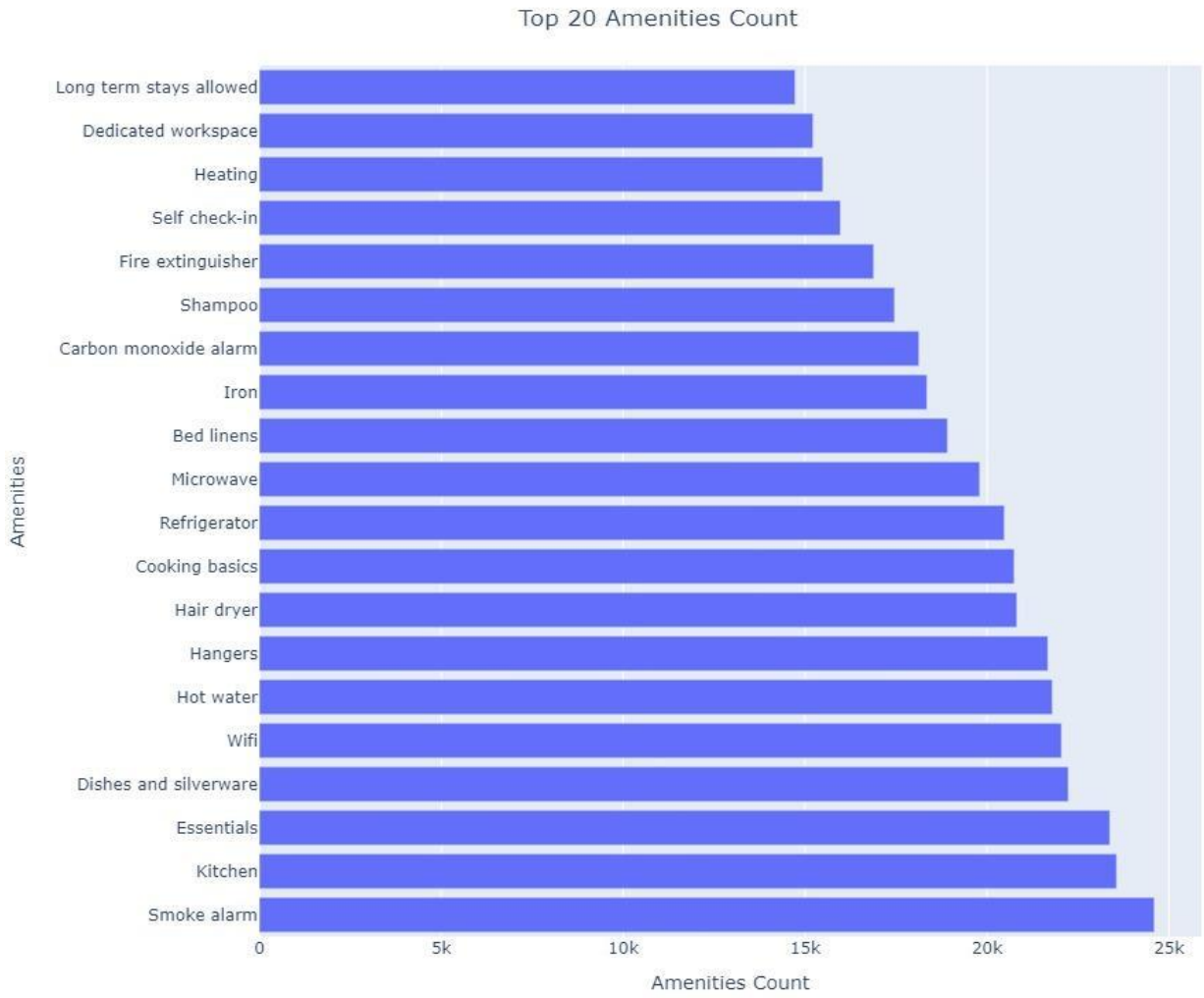

Figure 12 Count of Common Amenities available in most of the Listings

III. Feature Engineering

a. **Label Encoding:** Label encoding is used to convert categorical variables into numerical format, where each category is assigned a unique integer label. In our dataset, we applied label encoding to various columns such as "host_response_time", "room_type", "has_availability", and "instant_bookable", which contained categories like "within an hour," "Entire home/apt", "30", "F/T", etc. By mapping these categories to numeric values (0,

1, 2, respectively), we enabled the incorporation of this feature into predictive models which resulted in better analysis.

b. **One-Hot Encoding:** One-hot encoding is a technique used to create binary columns for each category within a categorical variable. For columns with multiple categories, we create new binary columns, where a value of 1 indicates the presence of that category, and 0 indicates its absence. We employed one-hot encoding on the "property_type" column, which had various property types like "Apartment," "House," "Condo," etc. We were effectively able to capture the categorical data converted into our dataset which helped us ensure that our predictive models are able to accurately intake the unique attributes of each property type and hence providing us with a better model to predict.

c. **Principal Component Analysis:** The execution of PCA initially resulted in MemoryError. Thus, it was required to find an alternative to reduce the dimensions of the TF-IDF result. The covariance matrix is computed in traditional PCA, which can be computationally and memory-intensive for big datasets. On the other hand, incremental PCA processes the data in smaller batches or chunks, making it possible to use it effectively on huge datasets.

IV. Prediction Model

a. **CatBoost:** CatBoost algorithm is based on principles of gradient boosting and decision trees. It considers the weak models and combines them using a greedy approach to build a strong prediction model. The trees produced by CatBoost act as regularization to avoid overfitting and provide more accurate predictions.

b. **XGBoost:** XGBoost is an ensemble supervised learning algorithm with regularization term and loss function. The objective of this algorithm is to find the difference between actual and predicted values. It is widely used due to its high performance and efficiency. It is based on gradient boosting where it combines several weak learners (decision trees) to create a powerful regression model.

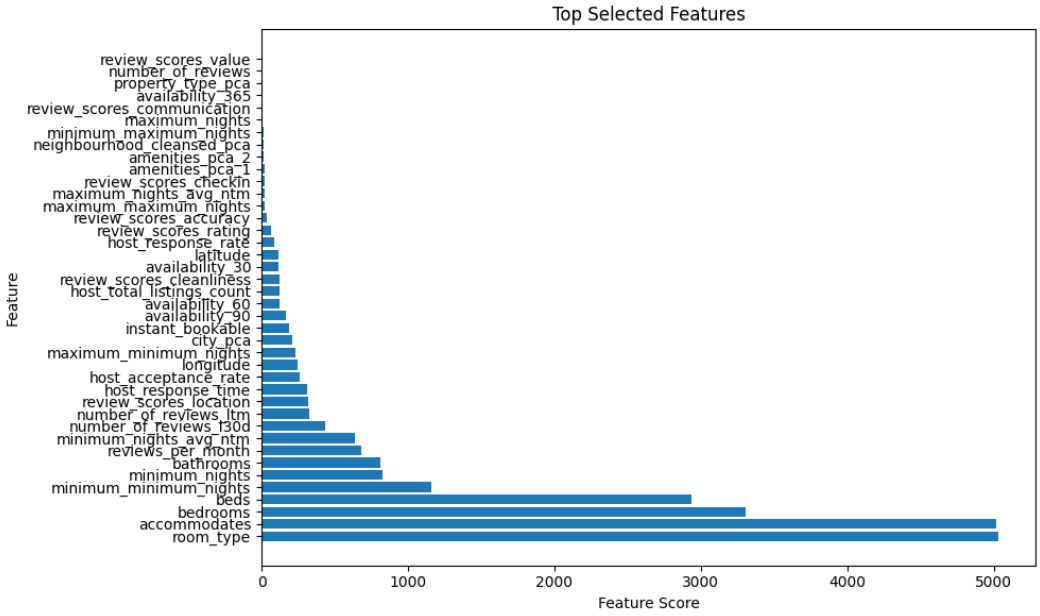

Figure 13: Select K-best features for Price Prediction

V. Demand Forecasting

Forecasting demand (Vareskic, 2021) involves using a mix of carefully gathered information from the Inside Airbnb dataset. This information includes historical booking patterns, trends that change over time, seasonal shifts, and local events. All these details come together to tell a story about how the demand for accommodations changes. We use advanced computer techniques, like studying how things change over time and using smart programs, to turn this dataset into models that can predict what travelers will do in the future. These models help us see patterns in traveler behavior that we might not notice otherwise. Demand forecasting is like a guiding light. It helps hosts, travelers, investors, and local authorities make better choices in a busy and always-changing market. At its heart, it's about using data to make educated guesses about how many people will want to stay in different cities in Canada – places like Toronto, Winnipeg, New Brunswick, Vancouver, Victoria, Quebec, and Montreal. Below a screenshot of the graph depicts the k-best features for demand forecasting.

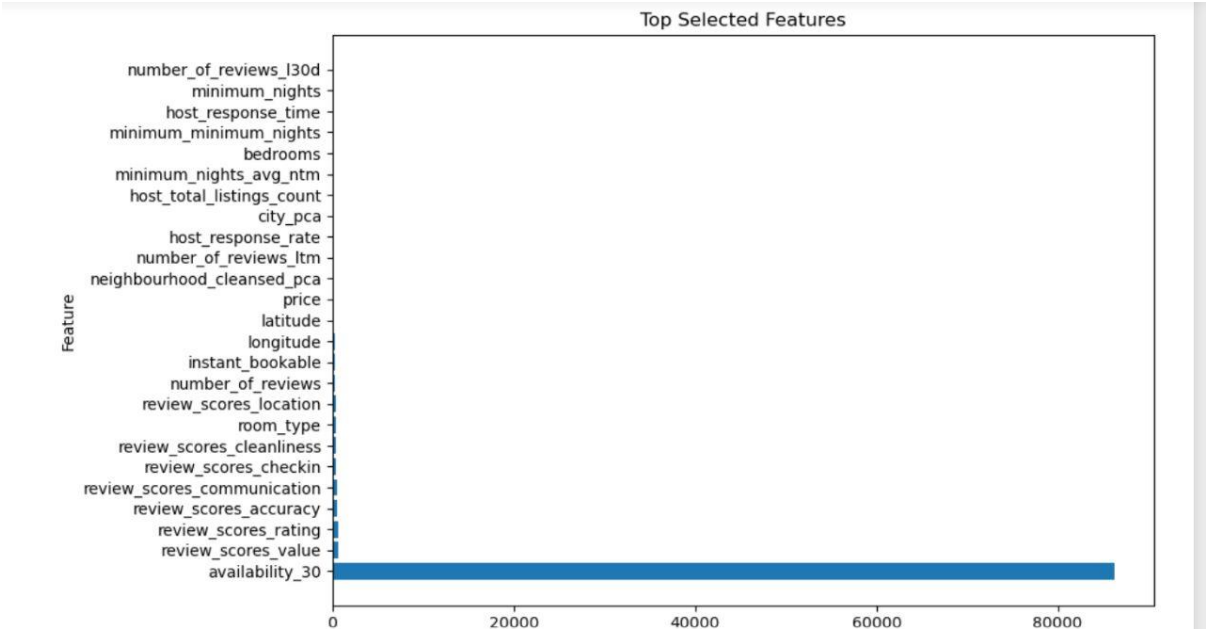

Figure 14 Select K-best features for Demand Forecasting

Results of Demand Forecasting

- The CatBoost model performs well in demand forecasting with a low Mean Squared Error (MSE) of 65.46 and a Mean Absolute Error (MAE) of 6.23. Its R2 test score of 0.79 indicates that around 79% of the variability in the demand can be explained by the model. The R2 train score of 0.85 highlights a balanced fit, demonstrating that the model effectively captures the underlying patterns in the training data.

- The XGBoost model also shows promising results with an MSE of 68.45 and an MAE of 6.33. Its R2 test score of 0.78 implies that approximately 78% of the demand variability is accounted for. The model's R2 train score of 0.88 indicates successful generalization and a robust grasp of demand dynamics during training.

- The Random Forest model delivers competitive performance with an MSE of 68.58 and an MAE of 6.43. Its R2 test score of 0.78 suggests that around 78% of the demand variability is captured by the model. The impressive R2 train score of 0.97 indicates that the model effectively adapts to the training data, although overfitting is visible.

Results of Price Prediction Model

The XGBoost model demonstrates strong performance with a low Mean Squared Error (MSE) of 3681.74 and a Mean Absolute Error (MAE) of 41.70. It showcases excellent predictive capabilities as evident from its high R2 test score of 0.70, indicating that around 70% of the

variance in the target variable is explained by the model. However, its very high R2 train score of 0.99 suggests a possibility of overfitting which is less generalized on new data.

Algorithms	MSE	MAE	R2 Score
CatBoost Train	1254.99	25.27	0.90
CatBoost Test	3772.46	42.50	0.69
XGBoost Train	28.24	4.01	0.99
XGBoost Test	3681.74	41.70	0.70

Table 2 Result matrix of Price Prediction Model

The CatBoost model achieves competitive results with an MSE of 3772.46 and an MAE of 42.50. Its R2 test score of 0.69 reflects its ability to capture patterns in the data, accounting for about 69% of the variance. The model's R2 train score of 0.90 highlights a balanced fit that avoids extreme overfitting, making it a comparatively reliable predictor for unseen data.

Conclusion and Future Work

In this price prediction and demand forecasting project, we set out to develop accurate models that could effectively predict prices and forecast demand for a given product or service. Throughout the project, we explored various data sources, preprocessing techniques, and modeling approaches to achieve our objectives. Our findings and results provide valuable insights into the feasibility and challenges of price prediction and demand forecasting in a dynamic market environment. successfully gathered and cleaned a comprehensive dataset, which included relevant features such as historical prices, economic indicators, seasonality factors, and promotional events. While our project achieved promising results, there are several avenues for future exploration and enhancement. Continuously refining and fine-tuning our models can lead to even better predictive accuracy. Experimenting with different hyperparameters, architecture designs, and ensemble methods may yield incremental improvements. Additionally, exploring additional data sources or deriving new features could enhance the model’s predictive power. Incorporating external data like social media trends, competitor pricing, and consumer sentiment could capture more nuances in the market. By continuing to refine our models, explore new data sources, and enhance interpretability, we can contribute to more informed decision-making and improved operational efficiency in the dynamic landscape of market dynamics and consumer behavior.

References

Vareskic, V. (2021, February 1). Demand Forecast using Machine Learning with Python. Medium. https://medium.com/vm-programming/demand-forecast-using-machine-learning-with-python-e8a4dca5aa0a

N. (2022, September 12). Demand Forecasting: Everything You Need to Know. Oracle NetSuite . https://www.netsuite.com/portal/resource/articles/inventory-management/demand-forecasting.shtml

Usage examples. (n.d.). Usage Examples - Python Package | CatBoost. https:///en/docs/concepts/python-usages-examples

Catboost with Python: A Simple Tutorial | AnalyseUp.com. (n.d.). Catboost With Python: A Simple Tutorial | AnalyseUp.com. https://www.analyseup.com/python-machine-learning/catboost-python-tutorial.html

A Step-by-Step Explanation of Principal Component Analysis (PCA). (2023, March 29). Built In. https://builtin.com/data-science/step-step-explanation-principal-component-analysis

sklearn.decomposition.PCA. (n.d.). Scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.decomposition.PCA.html

S, P. (2022, July 6). *An Introductory Note on Principal Component Analysis*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/07/principal-component-analysis-beginner-friendly/

Bar. (n.d.). Bar Charts in Python. https://plotly.com/python/bar-charts/