Understanding Data - Assignment 1

Title

Predicting Airbnb Prices: Understanding What Drives Short-Term Rental Costs

Research Question

Which listing features most influence Airbnb prices, and how well can we predict these prices using regression models?

Explanation and Relevance

Airbnb prices can vary significantly, even in the same city or neighborhood. A one-bedroom apartment in one area might cost twice as much as a similar one nearby. These price differences typically result from factors such as location, room type, guest capacity, and the number of reviews or ratings.

The goal of this project is to identify the most significant factors and create a model that can predict listing prices based on their details. By examining thousands of Airbnb listings, we can see how certain features, such as proximity to popular areas, more space, or higher reviews, affect the final price.

This kind of analysis can benefit both hosts and guests. Hosts can use the findings to set fair and competitive prices, while guests can understand why one listing costs more than another. For anyone interested in how data can clarify everyday situations, this project demonstrates how we can use available information to make better choices.

Theory and Background

Predicting prices is a common issue in data science, often addressed with regression analysis. Regression helps us see how various factors, or features, relate to the value we want to predict, which in this case is the price of an Airbnb listing. By examining patterns in past data, we can estimate how much each feature affects the overall price.

In this project, we use two types of regression models: Linear Regression and Random Forest Regressor. Linear Regression assumes a straightforward relationship between features and price. If one feature increases, the price adjusts predictably. This model is simple, easy to understand, and serves as a good starting point for most prediction tasks.

However, real-world data, like Airbnb listings, tend to be more complex. Price does not always rise or fall in a straight line with features like bedrooms or ratings. That's where Random Forest comes in. This model combines many smaller decision trees to capture patterns that are non-linear or show interactions between features. It can make more accurate predictions even when the relationships between inputs and price are complicated.

Both models operate under supervised learning, where we give the model examples of listings (the input) and their actual prices (the output). The model learns from this information and then predicts prices for new listings it hasn't encountered before.

Understanding Data - Assignment 1

By comparing these two methods, we can understand the limitations of simple models and the benefits of more flexible ones. This comparison shows how data-driven approaches can clarify and predict Airbnb pricing in a clear and reliable manner.

Problem Statement

The main goal of this project is to predict the nightly price of an Airbnb listing based on the information in its public data. Airbnb listings include various details, such as the type of property, its location, the number of guests it can accommodate, and how well it's rated by past visitors. By examining these details, we aim to identify which features significantly affect the price and how accurately a model can predict it.

- **Inputs (Features):**
  - ➔ **neighbourhood_cleansed:** The area or neighborhood where the listing is located.
  - ➔ **room_type:** The kind of space being offered (Entire home/apartment, Private room, etc.).
  - ➔ **accommodates:** The number of guests the listing can host.
  - ➔ **bedrooms, bathrooms_text:** Indicators of the size and comfort level of the listing.
  - ➔ **Number_of_reviews, review_scores_rating:** Measures of popularity and guest satisfaction.
  - ➔ **availability_365:** How many days the listing is available for booking during the year.

- **Output (Target):**
  - ➔ **price:** A continuous numeric value that represents the predicted nightly price of the listing in USD.
- **Example:**

  For a listing with the following details:

  - ➔ **Location:** Downtown Boston
  - ➔ **Room Type:** Entire home/ apartment
  - ➔ **Accommodates:** 3 guests
  - ➔ **Bedrooms:** 1
  - ➔ **Bathrooms:** 1
  - ➔ **Number of Reviews:** 45
  - ➔ **Review Score Rating:** 4.85
  - ➔ **Availability:** 120 Days

  The model might predict a price of $210 per night.

  This setup allows the model to learn from hundreds or thousands of listings, identifying pricing patterns, and then use that knowledge to predict prices for new or unseen listings.

Understanding Data - Assignment 1

Problem Analysis

The Airbnb dataset includes numerical, categorical, and text-based information, which creates several challenges during preparation for modeling. Some listings are missing details, like the number of reviews or bathrooms. Others have outliers, or extremely high or low prices, that can skew results. We need to handle these issues carefully to ensure the model learns the right patterns instead of being affected by incomplete or unusual data.

To tackle these challenges, we cleaned and transformed the data before modeling. We filled in missing numeric values using the median and replaced missing categorical values with the most common category. We cleaned prices by removing symbols like "$" and "," and clipped extreme values between the 5th and 95th percentiles to lessen the influence of luxury or unusually cheap listings. We converted categorical variables, such as room type and neighborhood, into numerical form using One-Hot Encoding. This change helps machine learning models process them effectively.

The approach to solving the problem followed a structured data science workflow:

- **Data Understanding and Cleaning** - Inspecting data quality, handling missing values, and formatting fields.
- **Feature Engineering** - Selecting relevant columns and converting categorical data into usable numeric forms.
- **Modeling** - Applying regression algorithms (Linear Regression and Random Forest Regressor) to learn the relationship between features and price.
- **Evaluation** - Measuring performance using Root Mean Squared Error (RMSE) and R-squared to compare accuracy between models.

The main data science principles used here are supervised learning, feature transformation, and model evaluation. Supervised learning allows the model to learn from past examples, such as listings with known prices. Feature transformation makes sure the data is in the correct format for machine learning algorithms. Model evaluation metrics offer a way to measure performance objectively and confirm that the predictions are sensible in real-world terms.

Solution Explanation

This project uses a straightforward, repeatable process to turn raw Airbnb listings into trustworthy price predictions. Each step has a specific goal and builds on the one before it, making the final results easy to trust and explain.

Step-by-step solution (Low-Level Explanation)

- **Load the data**
  - ➔ Read the Airbnb listings file and keep only the columns needed for prediction (price, location, room type, capacity, reviews, availability).
- **Clean and Standardize**
  - ➔ Convert the price from a string (e.g., "$129") to a number.
  - ➔ Parse bathrooms_text into a numeric bathrooms value.
  - ➔ Handle missing values (median for numbers, most frequent for categories).
  - ➔ Clip extreme prices to reduce the influence of outliers.

Understanding Data - Assignment 1

- **Transform Features**
  - ➔ One-hot encode categorical features (room_type, neighbourhood) so models can use them.
  - ➔ Keep numeric features as numbers and ensure consistent dtypes.
- **Split the data**
  - ➔ Split into training (to learn) and test (to evaluate) sets so we can judge how well the model works on unseen listings.
- **Train two Models**
  - ➔ Linear Regression as the interpretable baseline.
  - ➔ Random Forest Regressor to capture non-linear patterns and interactions.
- **Evaluate and Compare**
  - ➔ Compute RMSE (average dollar error) and R^2 (variance explained) on the test set.
  - ➔ Pick the model with the lower RMSE and higher R^2 as the stronger performer.
- **Interpret Results**
  - ➔ Review the Random Forest feature importances to see which factors matter most (location, room type, capacity).
  - ➔ Check that the findings match common sense and the patterns seen in the EDA.

Pseudocode (High-Level Explanation)

- load listings.csv
- select [price, neighbourhood, room_type, accommodates, bedrooms, bathrooms_text, number_of_reviews, review_scores_rating, availability_365]
- price ← clean_currency (price)
- bathrooms ← parse_number(bathrooms_text)
- impute_missing(numeric=median, categorical=most_frequent)
- price ← clip(price, p5, p95)
- X, y ← build_features_and_target(encode_one_hot=[neighbourhood, room_type])
- X_train, X_test, y_train, y_test ← train_test_split(X, y, test_size=0.2, seed=42)
- models ← {LinearRegression(), RandomForestRegressor()}
- for m in models:
  - ➔ fit m on (X_train, y_train)
  - ➔ y_pred ← m.predict(X_test)
  - ➔ rmse, r2 ← metrics(y_test, y_pred)
- Pick the model with the best rmse (lowest) and r2 (highest)
- Inspect feature_importances (if tree model) and summarize drivers of price

Why this approach is sound (Logical Reasoning)

Data leakage is avoided. All preprocessing, such as imputation and encoding, is conducted in a model pipeline. The model learns only from the training data and then applies this learning to the test data.

- **Fair evaluation:** A held-out test set estimates how the model will perform on new listings. This prevents overly optimistic results.

- **Suitable metrics:** RMSE shows the average dollar error, which is easy to understand. R^2 indicates how much of the price variation is explained.

- **Baseline vs. advanced model:** Starting with Linear Regression provides a clear benchmark. Random Forest demonstrates the benefit of managing non-linear relationships.

- **Consistency checks:** The most important features from the model—neighbourhood, room_type, accommodates, and bedrooms—align with the trends observed during exploratory data analysis. This boosts confidence in the results.

- **Robustness steps:** Clipping outliers and using sensible imputations reduce the model's sensitivity to unusual entries and missing data.

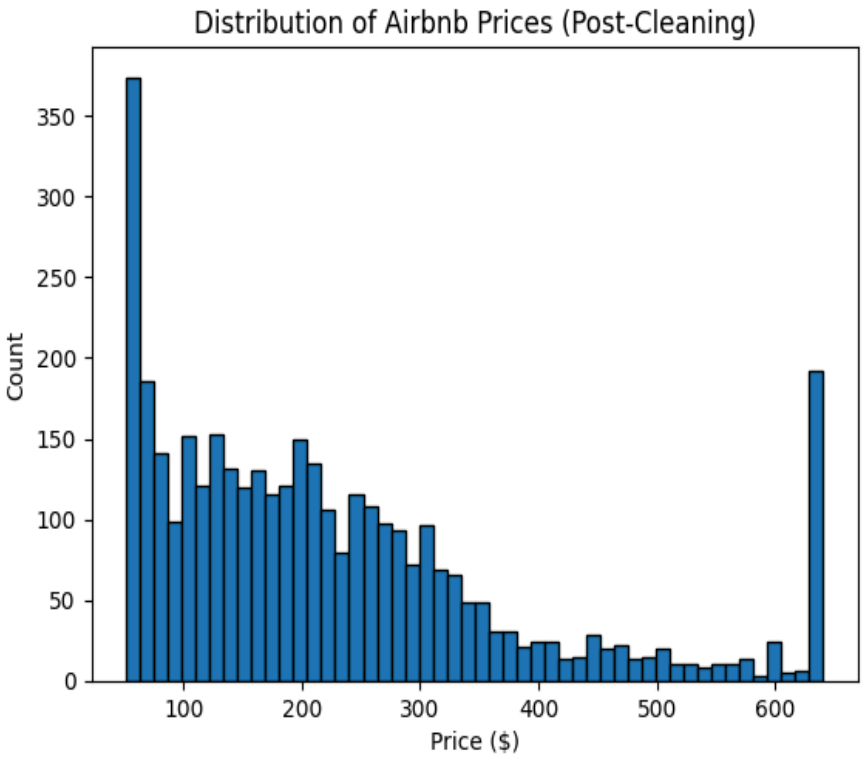
Together, these choices make the solution easy to follow, repeat, and defend. The pipeline is clean, the evaluation is fair, and the insights align with how Airbnb pricing works in practice.

Results and Data Analysis

After cleaning and preparing the Airbnb dataset, the final version had 3,695 valid listings. These listings were used to examine pricing patterns and train regression models. The analysis focused on understanding how different features, such as room type, location, and number of guests, influence the nightly price.

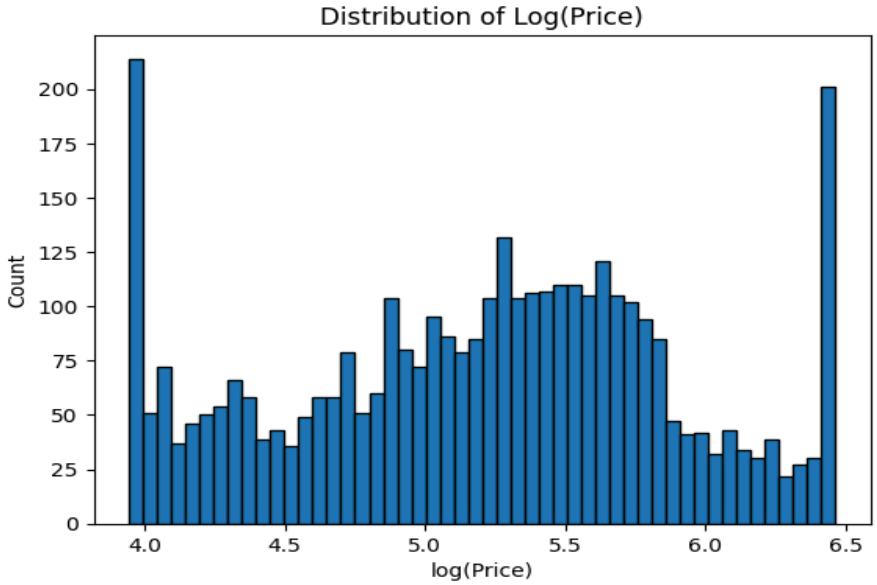
1) **Data Summary and Preparation Results**
   Before modeling, the dataset was reduced to the most relevant columns and cleaned for consistency.
   - **Original records:** 4,560
   - **After cleaning and outlier removal:** 3,695
   - **Columns used:** price, neighbourhood_cleansed, room_type, accommodates, bedrooms, bathrooms, number_of_reviews, review_scores_rating, availability_365

Understanding Data - Assignment 1



2) Price Distribution and Skew

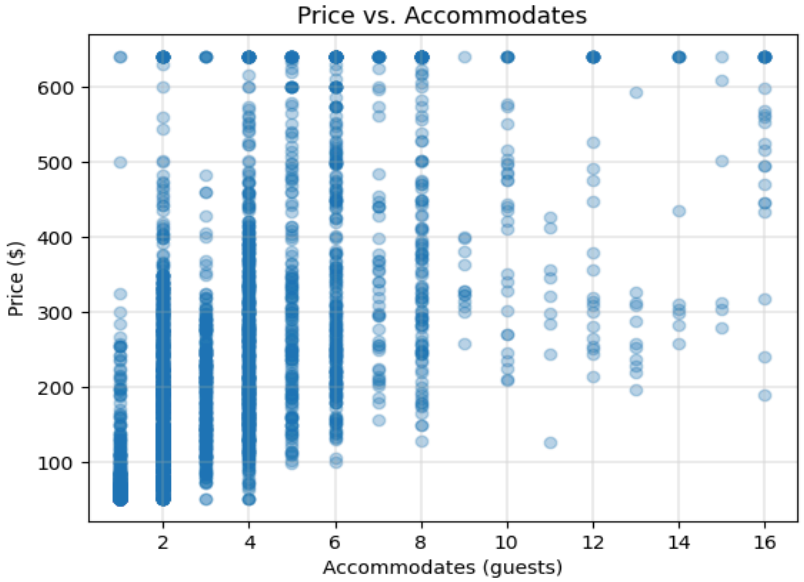
A histogram of the price column showed that Airbnb prices were right-skewed. This means most listings were moderately priced, while a few luxury listings cost much more. Using a log transformation helped to visualize prices more evenly and revealed the central pricing range across listings.

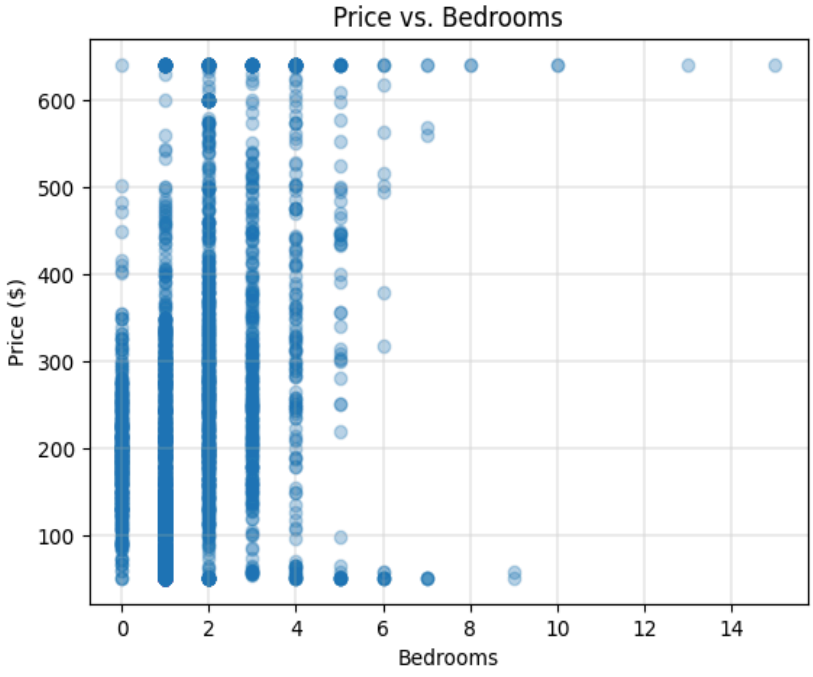


Understanding Data - Assignment 1

3) Relationships Between Price and Capacity Features

Scatter plots comparing price with features like accommodations and bedrooms showed that larger listings usually cost more. However, the increase is not consistent. Prices rise sharply up to a certain point and then level off. This confirms that while capacity matters, it's not the only factor that affects price.



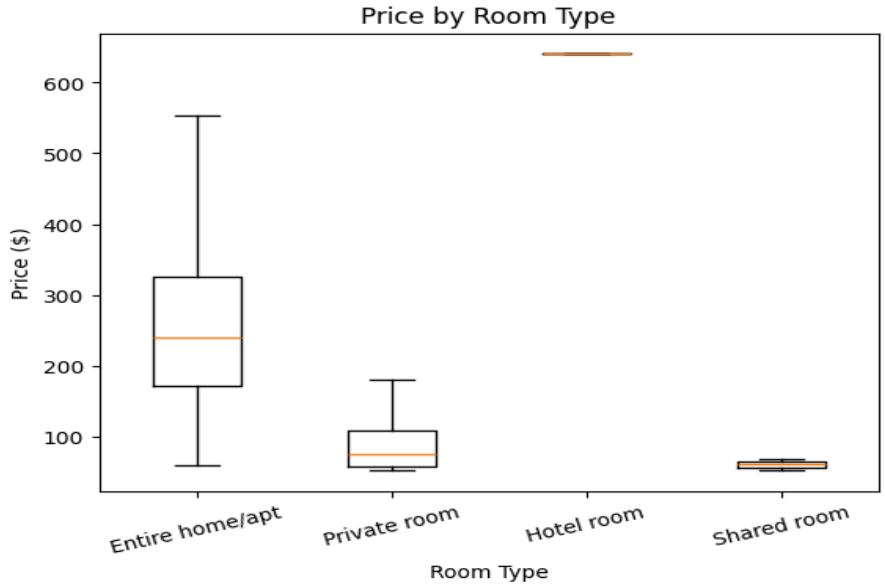


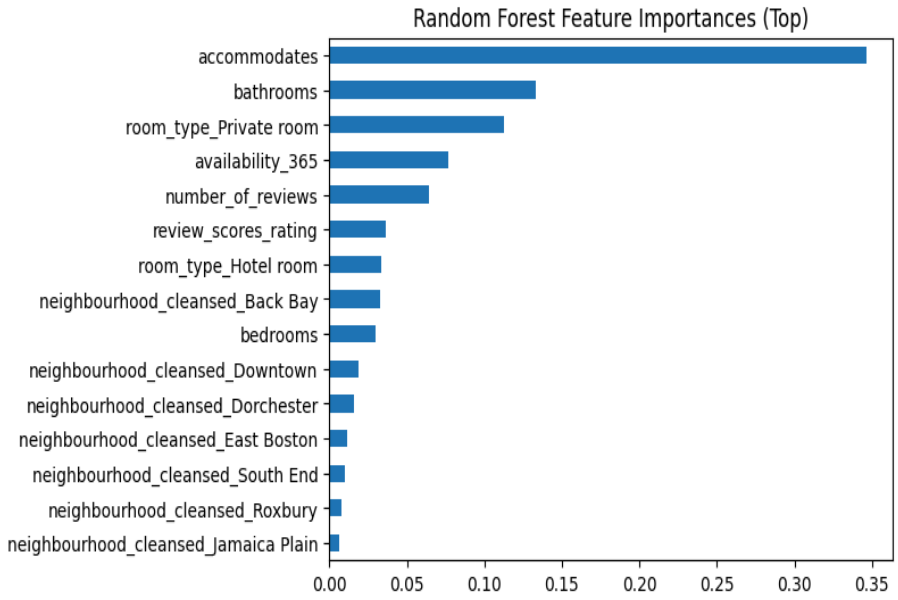
Understanding Data - Assignment 1

4) Categorical and Location-Based Differences

Boxplots and bar charts showed clear differences among property types and neighborhoods.

- Entire homes and apartments consistently had the highest prices.
- Private rooms and shared rooms were much cheaper.
- Some neighborhoods had higher average prices because they were close to city centers or tourist attractions.



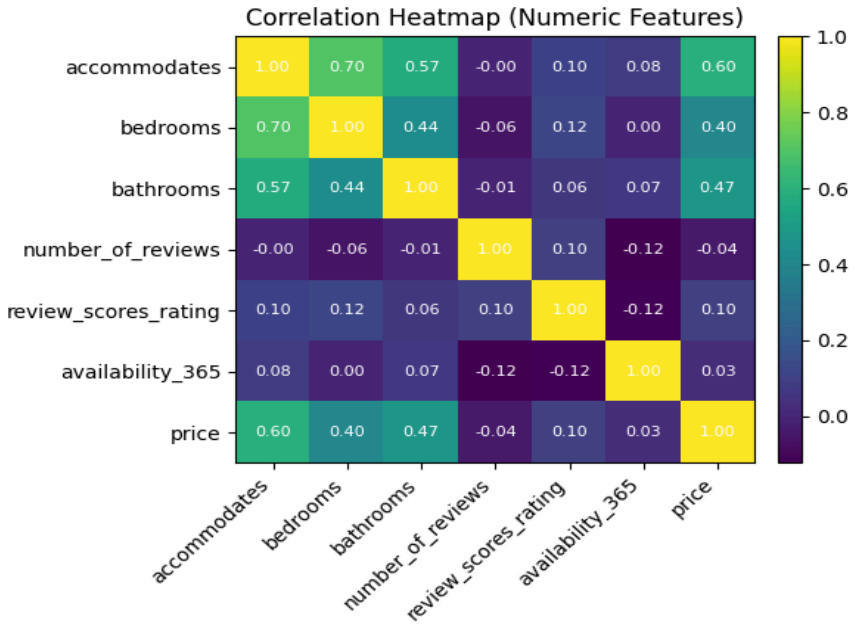


Understanding Data - Assignment 1

5) Correlation Heatmap and Feature Relationships

The correlation matrix highlighted that capacity-related features such as `accommodates`, `bedrooms`, and `bathrooms` were positively correlated with price.

Features like `number_of_reviews` and `availability_365` had weaker relationships, showing that guest volume alone doesn't directly determine price.



6) Model Performance

Two Regression models were trained and evaluated:

Model	RMSE ↓	R-Squared ↑	Interpretation
Linear Regression	98.04	0.600	Captures 60% of price variation; interpretable baseline.
Random Forest Regressor	87.08	0.685	Captures 68.5% of price variation; handles non-linear patterns

The Random Forest Regressor achieved a lower RMSE, meaning it has a smaller average error, and a higher R^2, which shows a better fit. This indicates that it understands complex relationships in the data better. These results match the theory. Random Forest models perform better than simple linear models when the relationships between features are non-linear or involve interactions.

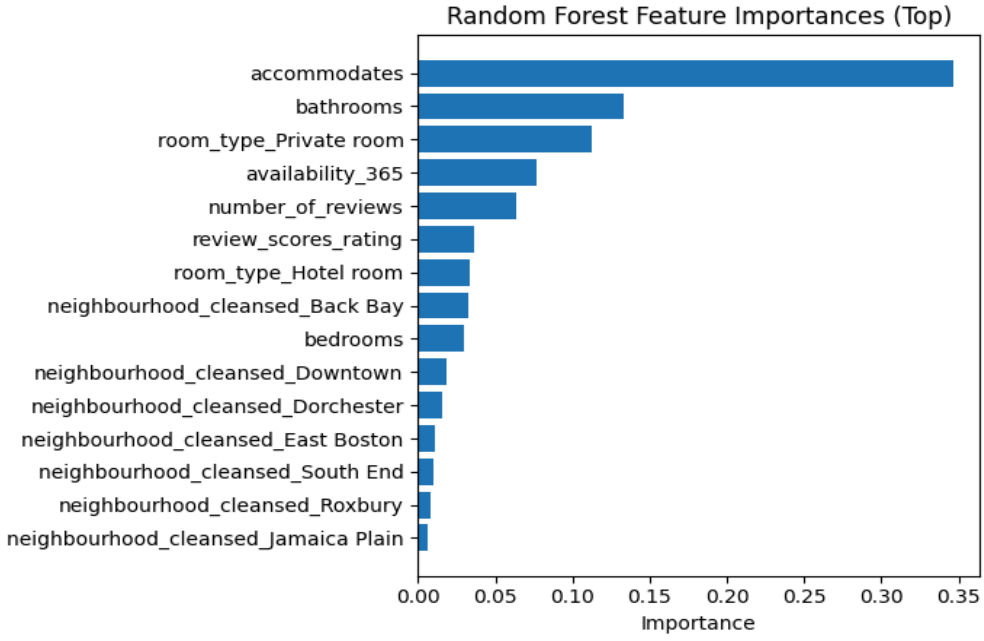
	feature	importance
0	accommodates	0.346290
1	bathrooms	0.132933
2	room_type_Private room	0.112513
3	availability_365	0.076477
4	number_of_reviews	0.063943
5	review_scores_rating	0.036728
6	room_type_Hotel room	0.033488
7	neighbourhood_cleansed_Back Bay	0.032458
8	bedrooms	0.030066
9	neighbourhood_cleansed_Downtown	0.018849
10	neighbourhood_cleansed_Dorchester	0.015539
11	neighbourhood_cleansed_East Boston	0.011397
12	neighbourhood_cleansed_South End	0.009925
13	neighbourhood_cleansed_Roxbury	0.007672
14	neighbourhood_cleansed_Jamaica Plain	0.006045

7) Feature Importance and Interpretation

The feature importance chart from the Random Forest model showed which factors affect Airbnb prices the most. The top predictors included:

- **neighbourhood_cleansed:** Location is the strongest factor in determining price.
- **room_type:** Entire homes/apartments cost the most, while shared rooms cost the least.
- **accommodates, bedrooms, and bathrooms:** Higher capacity generally increases price.

These findings support what was seen during the exploratory data analysis. The main drivers of price relate to location, room type, and property size.



Understanding Data - Assignment 1

9) Discussion and Insights

The results show a clear connection between the theoretical understanding of regression and the observed data.

Linear Regression captures broad, predictable trends but has difficulty with irregularities.

Random Forest manages complex, non-linear pricing behavior more effectively, which improves prediction accuracy.

The models show that pricing is influenced by a mix of location, property type, and capacity. Reviews and availability have smaller, yet still noticeable effects.

These findings offer a practical and understandable view of how data-driven models can help explain and predict Airbnb prices.

10) References & License

1) Inside Airbnb. (2025). *Get the Data.* Retrieved from https://insideairbnb.com/get-the-data.html - Primary data source containing public Airbnb listing information such as price, location, room type, and host details.

2) Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., et al. (2011). *Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research*, 12, 2825–2830 - Library used for data preprocessing, regression modeling, and evaluation.

3) Hunter, J. D. (2007). *Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering*, 9(3), 90–95 - Used for creating plots and visualizations to support exploratory analysis and presentation of results.

4) Pandas Development Team. (2024). *pandas: Powerful Python Data Analysis Toolkit.* Retrieved from https://pandas.pydata.org - Used for data cleaning, feature selection, and transformation throughout the analysis.

5) Breiman, L. (2001). *Random Forests. Machine Learning*, 45(1), 5–32 - Theoretical foundation for ensemble-based regression modeling and feature importance analysis.

6) NumPy Developers. (2024). *NumPy: Fundamental Package for Scientific Computing in Python.* Retrieved from https://numpy.org – Used for numerical operations and array-based data handling during preprocessing and evaluation.