Airbnb Price Prediction

Toh Jun Meng (U2322727H)
Low Kan Yui (U2322359G)
Natanael Tan Tiong Oon (U2322658G)

Introduction

What is Airbnb?

- A platform connecting hosts renting out homes to potential guests as an alternative to hotels.

Wide Variety of Pricing Factors:

- Listings vary in type, size, amenities, location, and more, influencing prices.

Problem Statement:

- **Challenge**: Help prospective Airbnb hosts set competitive prices for their listings.
- **Goal**: Use data to identify key factors influencing prices and suggest a pricing strategy.

Dataset

Source:

- Airbnb Prices Dataset from Kaggle.

Coverage:

- **57128 Data Points**
- Listings from **6 major US cities**:
  - **New York City**
  - **Los Angeles**
  - **San Francisco**
  - **Washington DC**
  - **Boston**
  - **Chicago**

Target Variable:

- **log_price**: The logarithmic transformation of listing prices.

Features:

- Categorical: property_type, room_type, bed_type, etc.
- Numerical: bathrooms, bedrooms, number_of_reviews, etc.
- Geospatial: latitude, longitude.
- Date Time: first_review, last_review.

Data Cleaning

Dropped Features:

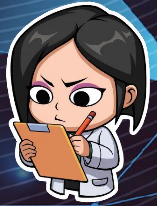
- **Host Features**:
  - host_has_profile_pic, host_identity_verified, host_response_rate, host_since
  - **Reason**: Minimal impact on pricing based on exploratory analysis.
- **NLP Features**:
  - thumbnail_url, amenities, description, name, neighbourhood
  - **Reason**: Encoding results in multiple features, leading to the **curse of dimensionality**.

Handling Categorical Data:

- **Room Type**: Condensed into Private & Shared
- Grouped less frequent types into **"Other"** category based on frequency thresholds.

Handling Time Data:

- Calculated **review_duration** (days between the first and last review).
- **Dropped** first_review and last_review due to negligible correlation with price.



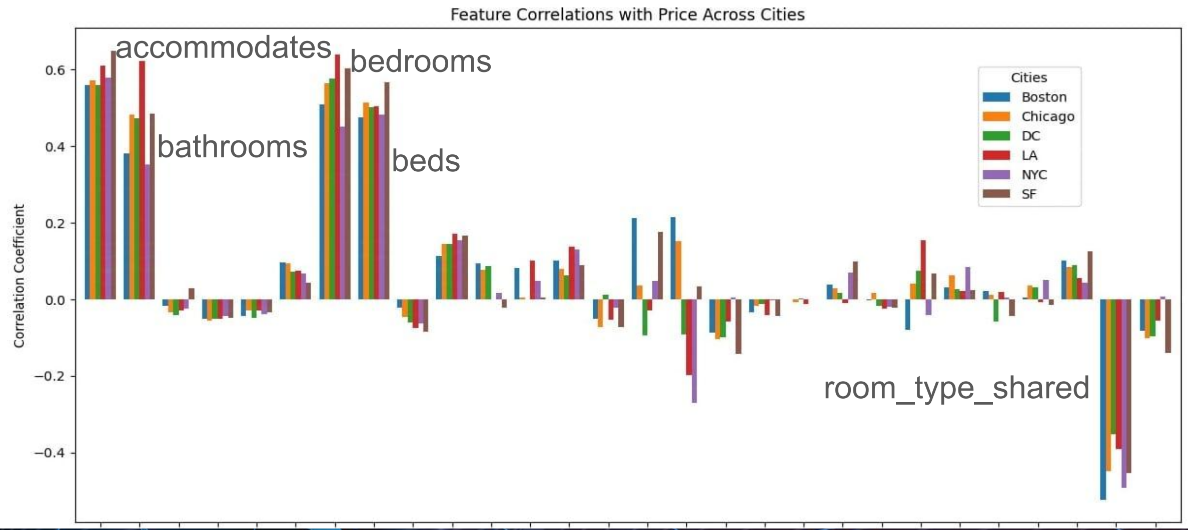
Exploratory Data Analysis Stage 1: Correlation

Occupancy and Room Features:

- Features related to **occupant capacity** (e.g., number of bedrooms, beds, and bathrooms) show a **positive correlation with price** across multiple cities.

Variability by City:

- Correlation strength varies by city, indicating that location-specific factors play a role in pricing.



Exploratory Data Analysis

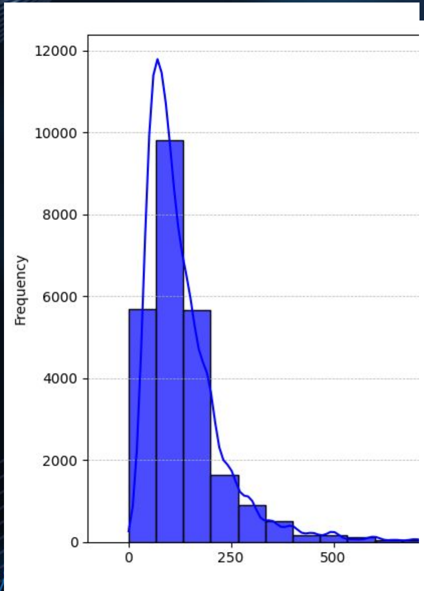
Stage 2: Investigating Price Patterns with Dimensionality Reduction

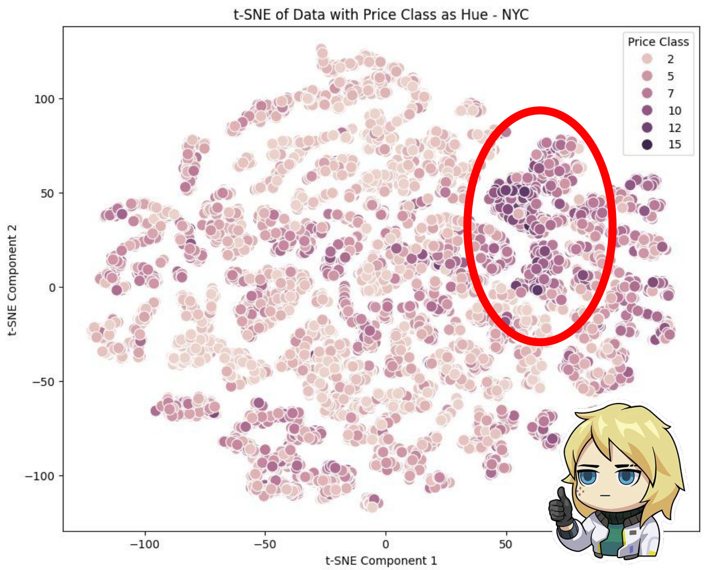
- Initial histogram of price provided basic distribution insights but was limited in revealing relationships with other features.
- To explore deeper patterns in high-dimensional data, we applied **PCA** (Principal Component Analysis) and **t-SNE** (t-distributed Stochastic Neighbor Embedding).

PCA and t-SNE for Visualization:

- **PCA**: Reduces features to principal components that capture the most variance in data.
- **t-SNE**: A non-linear technique focused on preserving local structure, allowing us to see clustering patterns.

In order to visualise patterns we use **K-Means clustering on price** to obtain **Price Classes** as a **Hue**.



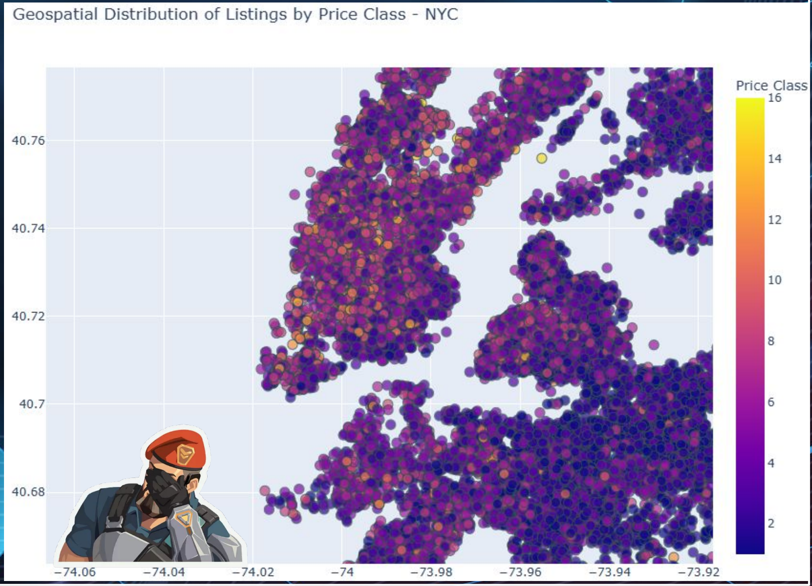


Exploratory Data Analysis

Stage 3: Analyzing Geospatial Patterns in Pricing

Map Visualization: The map (as shown) displays Airbnb listings in **New York City**, color-coded by `price_class`.

- Listings are color-coded from purple (lower price classes) to yellow (higher price classes), making it easy to spot higher-priced areas.



Insights from NYC Map

- Higher price_class listings are concentrated in certain areas, such as central Manhattan.
- This pattern suggests that **location within the city** plays a significant role in pricing.

Further Analysis

- We should analyze proximity to transport hubs, landmarks, and attractions to quantify these geospatial influences on pricing.

Feature Engineering

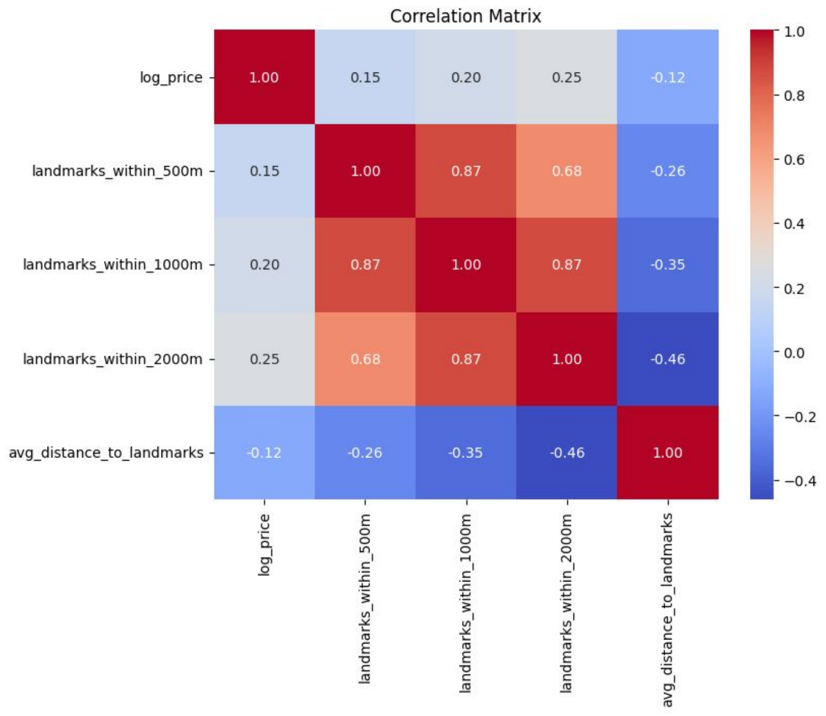
- Hypothesis: nearness to famous landmarks → higher value for tourism
  - Use web scraping to find the top landmarks for all 6 cities from wikipedia
  - Used regex to extract coordinate data
  - For each airbnb, calculate the average haversine distance to landmarks in the city
- Hypothesis: nearness to metro stations → higher property value
  - Manually gathered coordinates of metro stations for all 6 cities
  - For each airbnb, calculate the nearest metro station for its city

Landmarks

General Insights:

- **Average**: Weak negative correlation with price
- **Distance-Specific**: Weak positive correlation with price

This suggest that dense clusters of landmarks or their proximity could reflect the centrality and desirability of a location which might influence price.

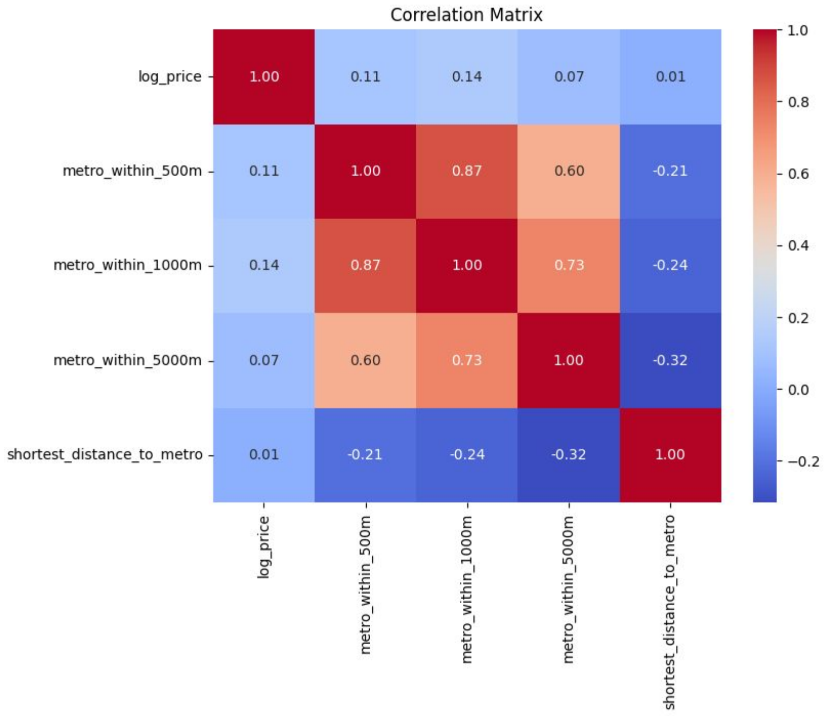


Metro Station

General Insights:

- **Shortest distance**: Very weak positive correlation with price
- **Distance-Specific**: Weak positive correlation with price

In general, the location of metro stations may play a role in Airbnb prices, but may only be more useful when paired with other data



Machine Learning Solution

Data **without** Feature Engineering

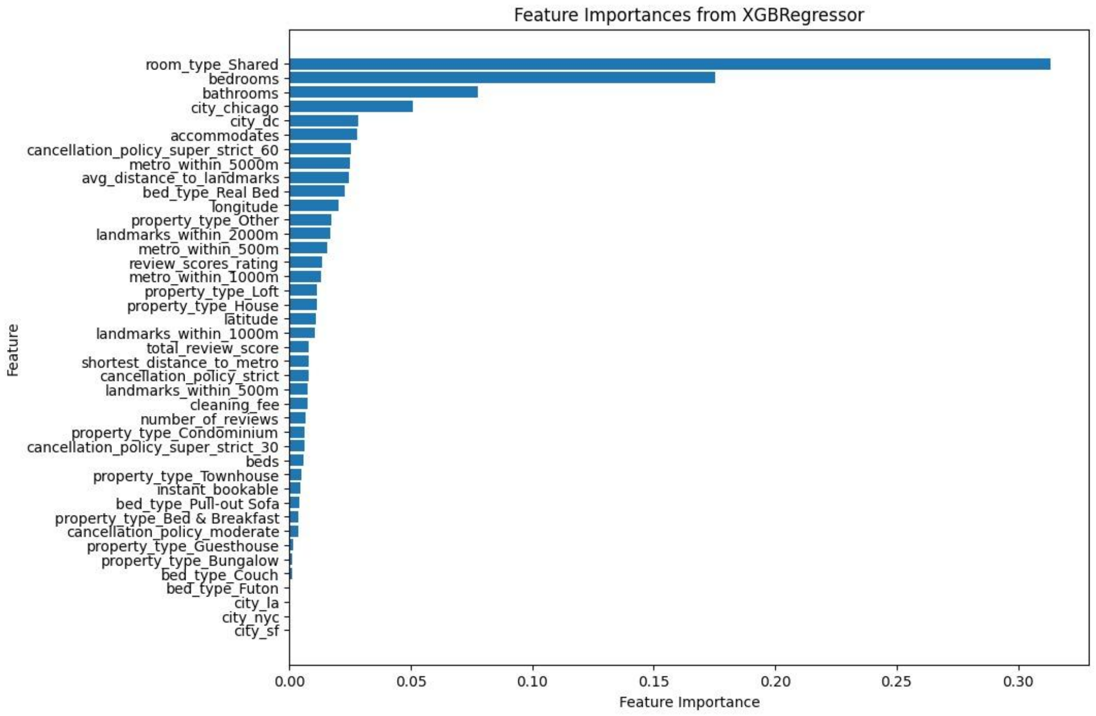
- Linear Regression
  - R^2: 0.520
  - RMSE: 93.06
- Decision Tree Regressor
  - R^2: 0.213
  - RMSE: 119.12
- Random Forest Regressor
  - R^2: 0.643
  - RMSE: 80.26
- XGBoost
  - R^2: 0.652
  - RMSE: 79.16

Data **with** Feature Engineering

- Linear Regression
  - R^2: 0.555
  - RMSE: 89.59
- Decision Tree Regressor
  - R^2: 0.286
  - RMSE: 113.72
- Random Forest Regressor
  - R^2: 0.654
  - RMSE: 78.97
- XGBoost
  - R^2: 0.666
  - RMSE: 77.57

Insights

- **Importance of city:** high importance of Chicago and DC
- **Importance of engineered features:** relatively high importance of metro within 5000m and average distance to landmarks



Conclusion

Outcome

- **Solving the problem**: Moderately high R^2 score of 0.67
- **Other insights**: Our hypotheses on landmarks and metro stations were useful

Learning Points

- **Exploratory data analysis:** Dimensionality reduction techniques such as PCA and t-SNE to cluster data points
- **Feature engineering**: Scraping internet data to augment our dataset
- **Model optimization**: Advanced model XGBoost that uses gradient boosting to make highly accurate predictions