**AirBnB Pricing Analysis**

**IDS 561: Big Data**

**Spring 2023**

**Final Project Report**

Group Number 13

Ming Lin (677365261)

Jinrong Qiu (673250167))

Yixiao Pan (672848449)

Shyam Patel (678481784)

**Problem Setting**

Airbnb is an online platform that allows people to rent out their homes, apartments, or other types of accommodation to travelers. The company was founded in 2008 and has since become one of the world's largest hospitality companies, with over 4 million hosts in more than 220 countries. What sets Airbnb apart from traditional hotel accommodations is the personalized experience that it offers travelers. Instead of staying in a standard hotel room, Airbnb allows travelers to rent out a wide range of unique and authentic accommodations, from cozy apartments in urban areas to luxurious villas in secluded areas.

Airbnb has also disrupted the traditional hotel industry by offering more competitive prices for travelers. With a wide range of accommodation options available on the platform, travelers can find lodging that fits their budget and preferences. Moreover, Airbnb's review system and verified profiles help build trust between hosts and travelers, creating a more welcoming and safe environment for all parties involved. Another key feature of Airbnb is its commitment to sustainability and responsible travel. The company has introduced various initiatives to reduce its carbon footprint and promote sustainable tourism, such as encouraging hosts to use eco-friendly cleaning products and offering experiences that highlight local culture and traditions.

One of the key factors that determine the price of an Airbnb listing is the features of the property. Features such as the number of bedrooms, location, amenities, and overall quality can significantly influence the price of a listing. However, determining which features have the most significant impact on the price can be a complex task. This is where machine learning techniques come into play. Machine learning is a subset of artificial intelligence that enables computers to learn from data without being explicitly programmed. It involves building models that can make predictions or decisions based on patterns in the data. In the context of Airbnb, machine learning can be used to analyze data about various features of listings and identify which ones have the most significant impact on the price.

The problem setting for this task is to build a machine learning model that can predict the log price of an Airbnb listing based on its features. The input data for this model would be a dataset of Airbnb listings and their associated features, such as the number of bedrooms, location, amenities, and overall quality. The output would be a predicted log price for each listing based on its features.

**Data Description**

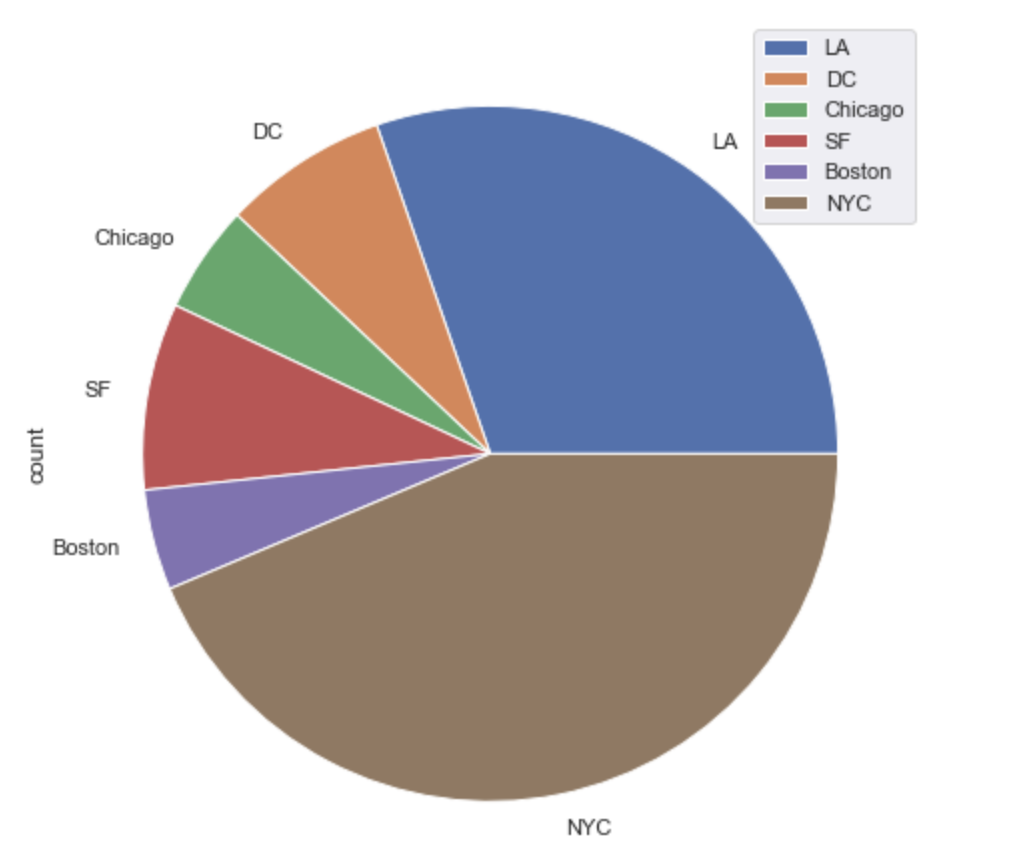
The Airbnb dataset available on Kaggle is a comprehensive collection of information on over 74,000 Airbnb listings in major cities around the world. The dataset is rich in detail and includes information about the properties, hosts, neighborhoods, and various other attributes. The dataset provides a wealth of information that can be used to analyze and explore the world of Airbnb rentals.

The dataset includes 16 columns and 74,104 observations. The columns in the dataset are:

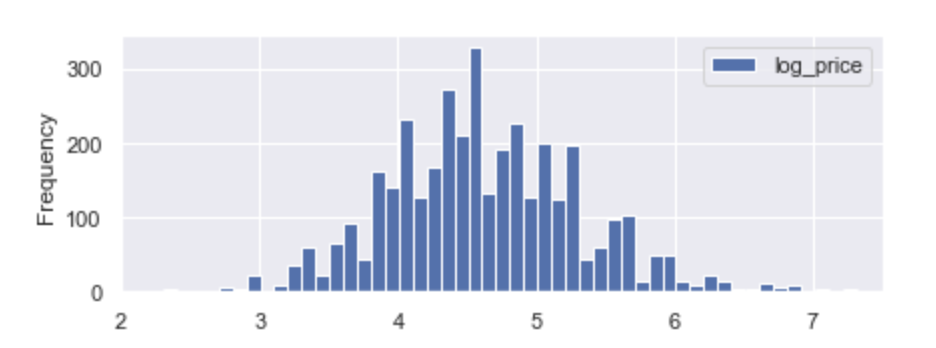
1. ID: This column contains a unique identifier for each listing. It can be used to identify each listing individually.
2. Name: This column contains the name of the property. It can be used to identify the property and differentiate it from other listings.
3. Host ID: This column contains a unique identifier for each host. It can be used to identify the host of each listing.
4. Host Name: This column contains the name of the host. It can be used to identify the host and differentiate them from other hosts.
5. Neighbourhood Group: This column contains the name of the neighborhood group (if applicable). It can be used to group listings by neighborhood.
6. Neighbourhood: This column contains the name of the neighborhood. It can be used to identify the location of each listing.
7. Latitude: This column contains the latitude coordinates of the property. It can be used to plot each listing on a map.
8. Longitude: This column contains the longitude coordinates of the property. It can be used to plot each listing on a map.
9. Room Type: This column contains the type of room (private room, shared room, or entire home/apt). It can be used to group listings by type of room.
10. Price: This column contains the nightly price for the listing. It can be used to analyze the pricing of listings.
11. Minimum Nights: This column contains the minimum number of nights required to stay. It can be used to analyze the length of stay for listings.
12. Number of Reviews: This column contains the number of reviews for the listing. It can be used to analyze the popularity of each listing.
13. Last Review: This column contains the date of the last review. It can be used to analyze the recency of reviews.
14. Reviews Per Month: This column contains the average number of reviews per month. It can be used to analyze the frequency of reviews.
15. Calculated Host Listings Count: This column contains the number of listings for each host. It can be used to analyze the activity of hosts on the platform.
16. Availability 365: This column contains the number of days the property is available for rent in a year. It can be used to analyze the availability of listings.

This dataset is an excellent resource for those interested in analyzing the Airbnb market. The dataset can be used to identify popular neighborhoods, determine pricing strategies, and explore relationships between host characteristics and the success of their listings. Moreover, the dataset can be used to develop machine learning models that can predict the price of a listing based on its features, identify the features that influence the popularity of a listing, and even make recommendations to hosts on how to improve their listings. Overall, the Airbnb dataset is a valuable resource for anyone interested in the sharing economy and the world of vacation rentals.

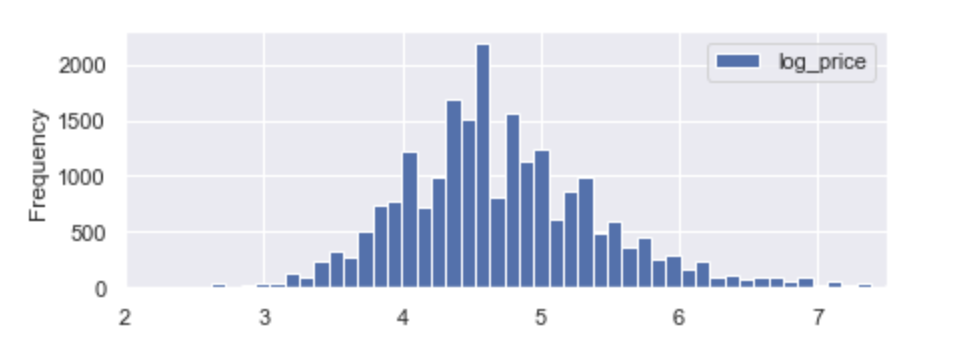
The samples we used in this dataset were from Los Angeles, Washington DC, Chicago, San Francisco, Boston, New York City, most of the samples were from Los Angeles, and New York City.



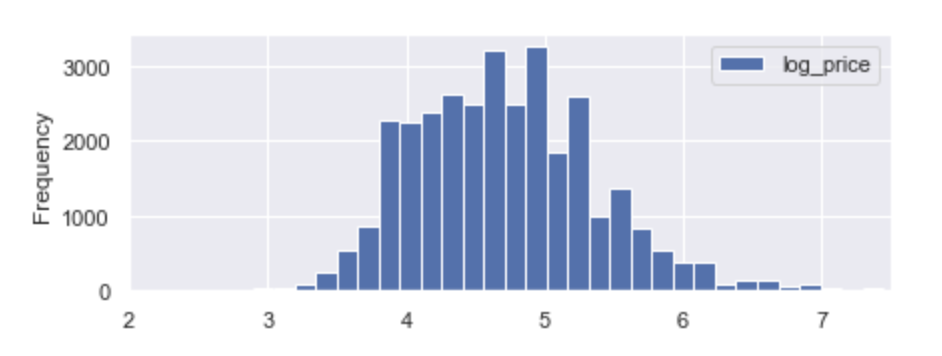
The distribution of prices in each city are shown below. Boston has the average highest price in this dataset. The price distribution of Chicago, Washington DC, and Las Angeles have basically the same price distribution.



Chicago

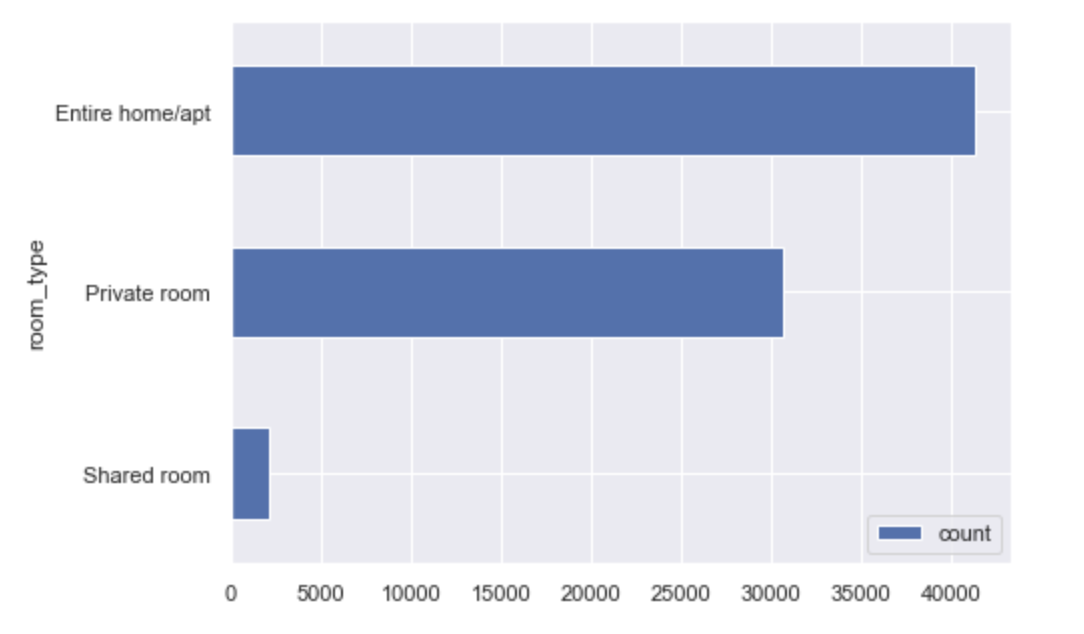
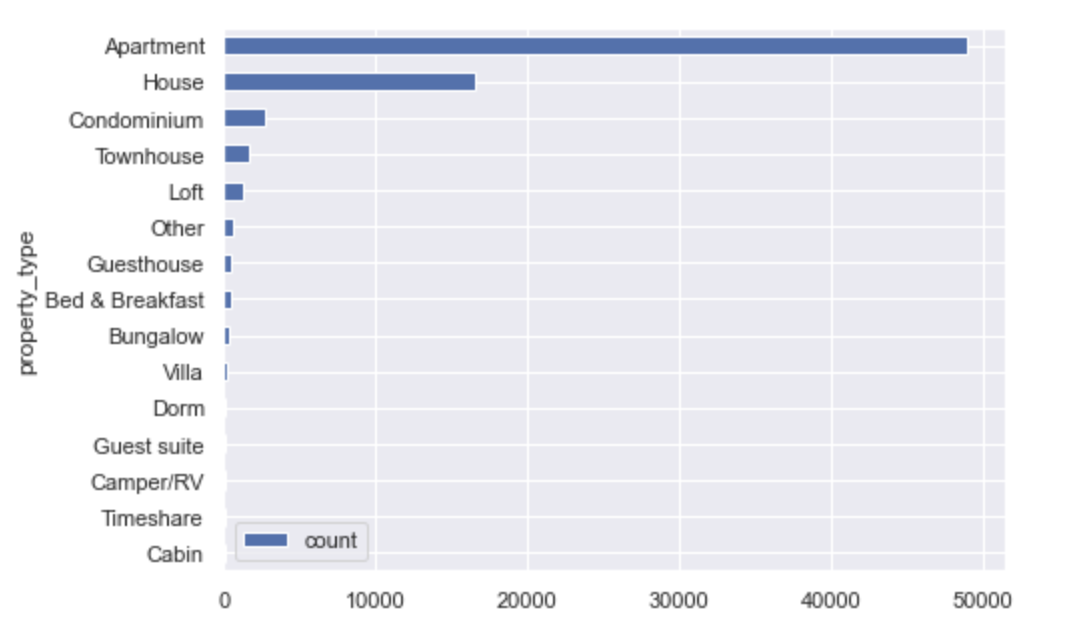


Las Angeles



New York City

As for the property type, most of the samples ordered apartment and house, and for the room type, we could see most of the customers prefer the entire home or apartment and private room.



**Techniques**

Data preparation is a crucial step in any data analysis or machine learning project. In the case of the Airbnb dataset, several steps can be taken to clean and preprocess the data before using it for analysis or modeling.

One important step is to identify and handle missing values in the dataset. Missing values can cause errors in analyses and modeling and can lead to inaccurate results. To handle missing values, we can first identify the columns with missing values and decide on an appropriate strategy to handle them. For example, we can drop rows with missing values in specific columns like 'bedrooms' since it is a crucial feature for rental properties. Next, we can remove unnecessary columns from the dataset that are not needed for analysis or modeling, such as 'id', 'host\_identity\_verified', 'name', 'description', and 'thumbnail\_url'. Removing these columns can help reduce the dimensionality of the dataset and improve the performance of machine learning models. To further clean the dataset, we can remove duplicate rows, which can occur due to multiple entries of the same property or host. Removing duplicates can improve the accuracy of the analysis and modeling.

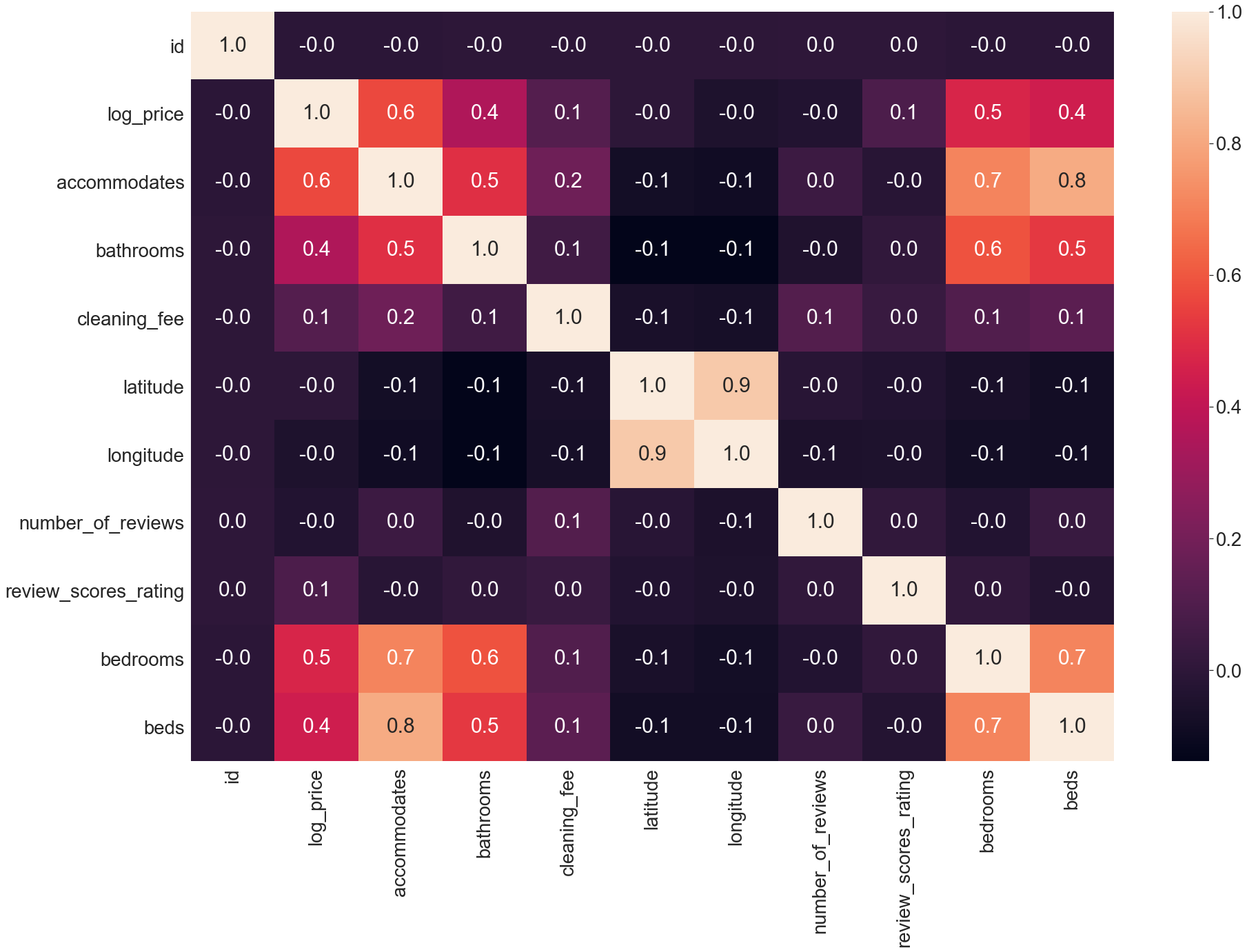
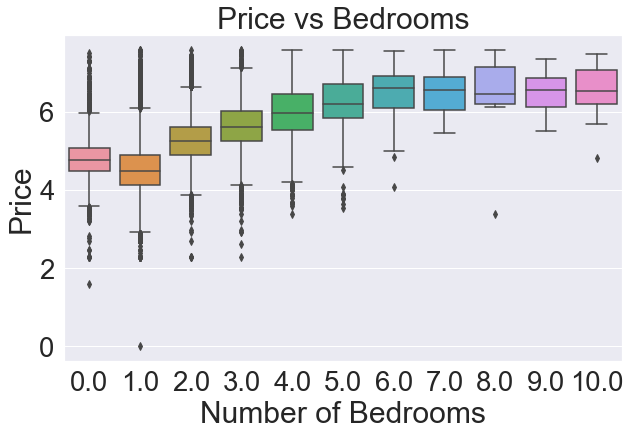
We can also convert the data type of certain columns to datetime format, such as 'last\_review', 'first\_review', and 'host\_since', to make it easier to analyze and manipulate these features.

Additionally, we can drop features with too many unique values, as they may not provide enough information to improve the accuracy of the models. For example, the feature 'neighbourhood' may have too many unique values to be a useful feature. To convert the response rate feature to a number, we can remove the percent sign and convert it to a float data type. We can also replace the missing value for 'host\_has\_profile\_pic' with 'f' to ensure consistency in the dataset. By performing these data preparation steps, we can obtain a clean and well-structured dataset that is suitable for analysis and modeling.

We also used the following models in our analysis:

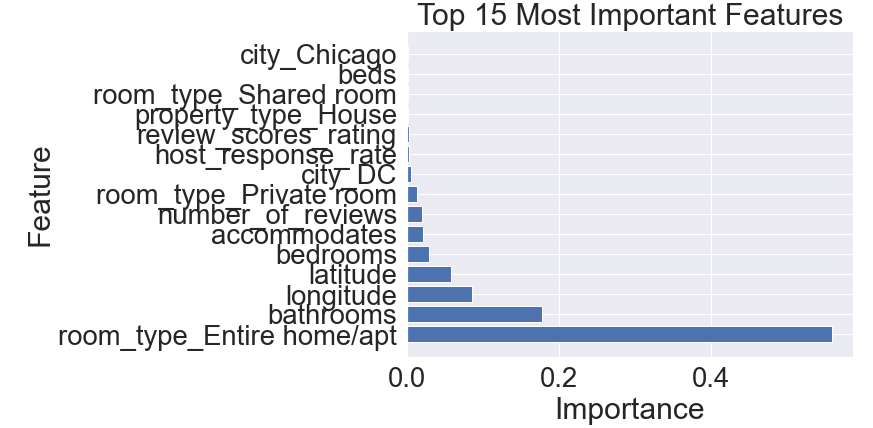
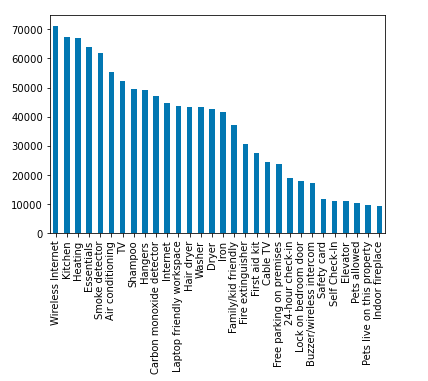
1. **Linear Regression:** Linear regression is a simple yet powerful algorithm used for predicting a continuous numerical value. It works by fitting a straight line through a set of data points in a way that minimizes the sum of the squared differences between the predicted values and the actual values. Linear regression assumes that there is a linear relationship between the input variables and the output variable.
2. **Lasso Regression:** Lasso regression is a type of linear regression that adds a penalty term to the objective function, which helps to reduce the complexity of the model and avoid overfitting. It does this by shrinking the coefficients of some of the input variables towards zero, effectively selecting only the most important features for prediction. Lasso regression is especially useful when dealing with datasets that have a large number of input variables.
3. **Decision Tree:** A decision tree is a tree-shaped model used for classification and regression analysis. It works by recursively splitting the data into subsets based on the most significant attribute or feature. The goal is to create a tree that predicts the target variable by making decisions based on the input variables. Decision trees are simple and easy to understand, making them a popular choice for data analysis and machine learning.
4. **Random Forest:** A random forest is an ensemble learning method that uses multiple decision trees to improve the accuracy of predictions. It works by creating a set of decision trees using a subset of the input variables and data points, and then averaging the results of the individual trees to make a final prediction. Random forests are highly accurate and can handle large datasets with many input variables.
5. **K-Nearest Neighbors (KNN):** KNN is a simple, non-parametric algorithm used for classification and regression analysis. It works by finding the K nearest data points to a new input and predicting the output variable based on the most common or average output variable value of its K neighbors. KNN is computationally efficient and can handle non-linear data, making it a popular choice for data analysis.
6. **Support Vector Machine (SVM):** SVM is a powerful algorithm used for classification and regression analysis. It works by finding the hyperplane that best separates the input data into different classes or groups. SVM can handle non-linear data by using kernel functions to map the data into a higher-dimensional space, making it easier to separate the data. SVM is known for its ability to handle complex datasets and high-dimensional data

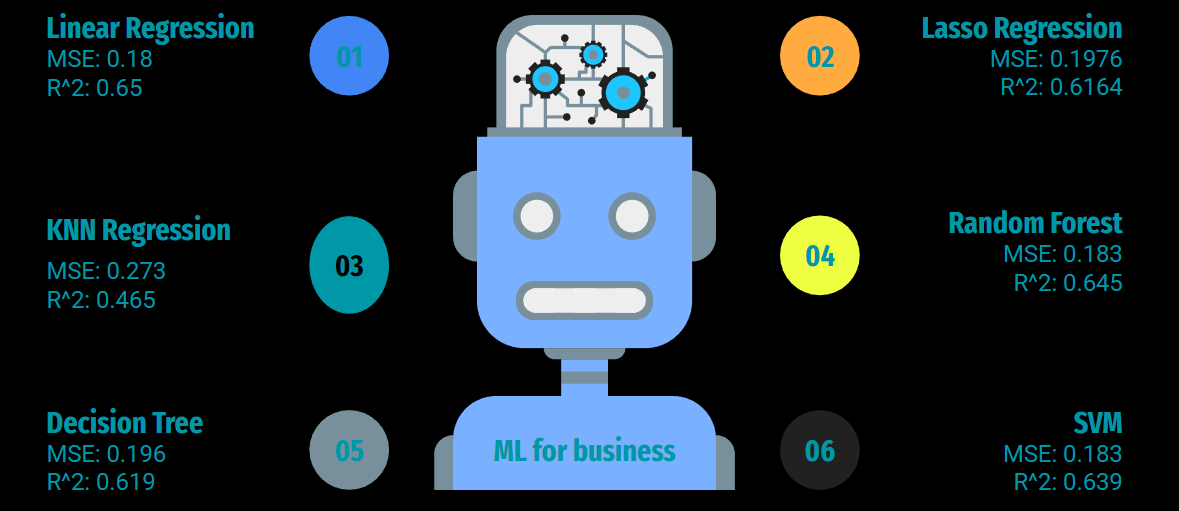
**Results (Still need to describe the visualizations and why they matter)**

****

The Amenities in column means what will the place offer, and it is shown as a list. We break the list and count all the features, trying to find what features that the place in the dataset would offer the most. On the one hand, we want to figure out what features play an important role in a customer's mind, on the other hand, we also want to find out what features that the place will offer most. As the hist graph shown below, we found that wireless internet got the first place, kitchen and heating got the second and third highest in the dataset. We believe wireless internet and kitchen are more important for the group of customers, as for the heating, we think that might because half our samples in the dataset are from Chicago, NYC, and Boston, the reason for why heating got the high rank is because those city have a long period in winter, and those cities will be cold.

Through the feature selection we found out the top 15 most important features based on average count. Wireless Internet has become the top most feature requirement while getting an apartment. Whereas, indoor fireplace became the least preferred. This shows the change in needs of people over time.

****

****

As we can see most models perform generally pretty well. Some notable results include Linear Regression, Random Forest, and SVM having a higher combined MSE and R^2 compared to the other models. Between these 3 models we ended up with choosing Random Forest as our final model as it has the highest combination of MSE and R^2.

**Role of Team Members**

**Ming** - Team Organization, Model Building and Analysis, Report Consolidation

**Jin** - Data cleaning, Data Visualizations, Model Buildings and Analysis

**Shyam** - Slide allocation, and goal setting.

**Peter** - Data Exploration, Model Building and analysis