# Airbnb Listing Analysis

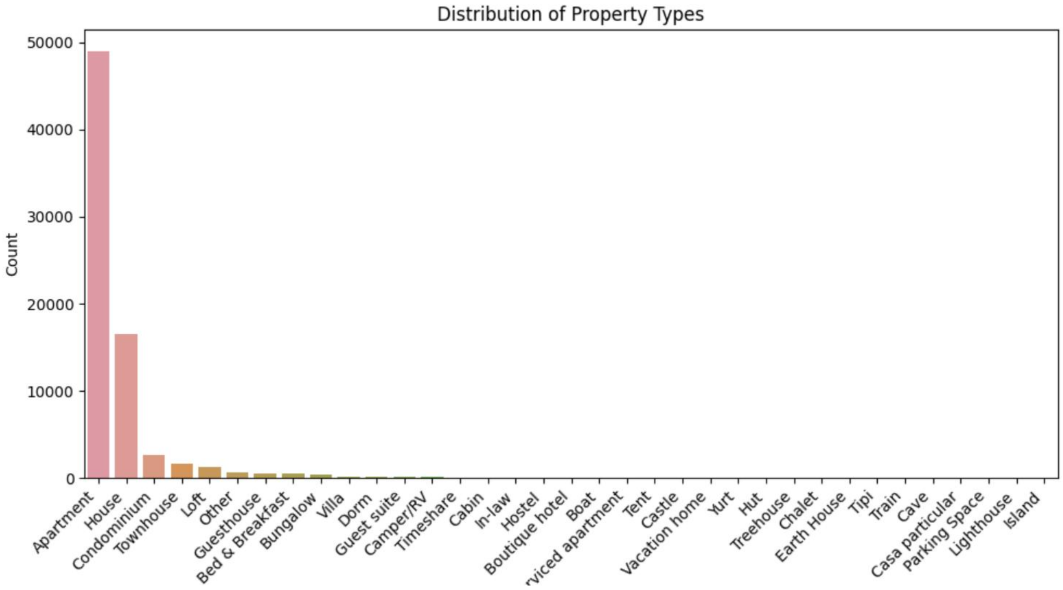
Explanation of Data and Data Collection Process

The AirBnB dataset contains 29 columns and 74111 rows of data. Each row of data consists of the specific features and details regarding each AirBnB listing. Here's a brief description of each feature:

- `id`: Unique identifier for each listing.
- `log_price`: The logarithm of the price of the listing.
- `property_type`: Type of the property (e.g., apartment, house, condominium).
- `room_type`: Type of the room (e.g., entire home/apt, private room, shared room).
- `amenities`: List of amenities available in the listing.
- `accommodates`: Number of people the listing can accommodate.
- `bathrooms`: Number of bathrooms in the listing.
- `bed_type`: Type of bed (e.g., real bed, sofa bed, pull-out sofa).
- `cancellation_policy`: Level of strictness of cancellation policy
- `cleaning_fee`: Whether a cleaning fee is charged for the listing.
- `city`: City where the listing is located.
- `description`: Description of the listing.
- `first_review`: Date of the first review for the listing.
- `host_has_profile_pic`: Indicates if the host has a profile picture.
- `host_identity_verified`: Indicates if the host's identity has been verified.
- `host_response_rate`: Response rate of the host to inquiries.
- `host_since`: Date the host joined Airbnb.
- `instant_bookable`: Indicates if instant booking is available for the listing.
- `last_review`: Date of the most recent review for the listing.
- `latitude`: Latitude coordinate of the listing's location.
- `longitude`: Longitude coordinate of the listing's location.
- `name`: Name or title of the listing.
- `neighbourhood`: Neighbourhood where the listing is situated.
- `number_of_reviews`: Total number of reviews for the listing.
- `review_scores_rating`: Rating score based on reviews.
- `thumbnail_url`: URL of a thumbnail image for the listing.
- `zipcode`: Zipcode of the listing's location.
- `bedrooms`: Number of bedrooms in the listing.
- `beds`: Number of beds in the listing.

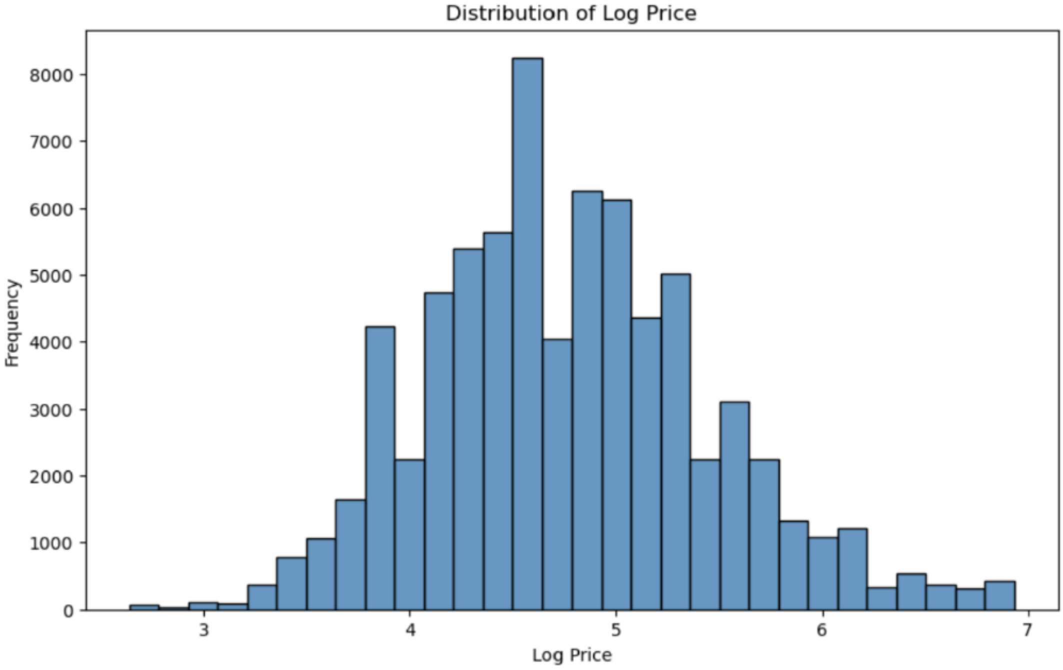
During our feature engineering process, we removed several features which are irrelevant to determining air bnb prices such as ‘id’, ‘zip code’, ‘name’ etc. As some of our models do not handle string values well, we used one-hot encoding to encode the categorical variables which had string values such as “bed _type” ， ‘city’ ， ‘room_type” . After cleaning and encoding the categorical variables, some features also had missing values which we imputed with the median value of the column.

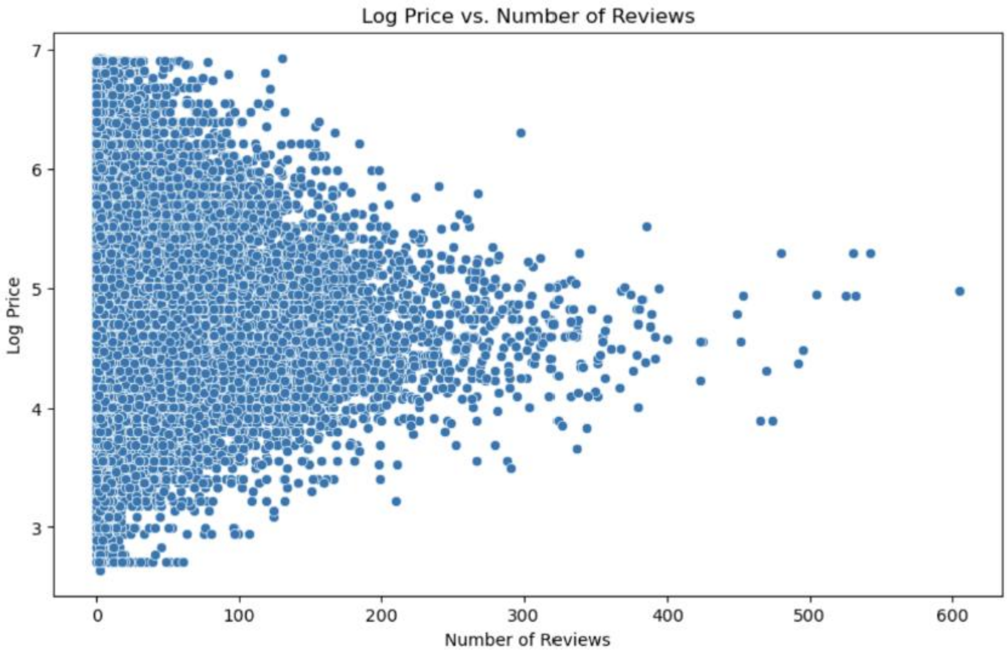
Exploratory Data Analysis



The count plot shows that Apartment is the most popular types of property, with around 49000 frequency. The variety of choice for property types is quite wide as there are Airbnb type such as Boat, Castle, Cave and Island. Although most of the listing are apartment or normal houses, there are still a variety of choice in the US even though the number is not that significant.

Understanding the distribution of property type can allow host and traveller to know about the demand and supply and thus giving them reference for how to set or choose appropriate price for a particular property type.

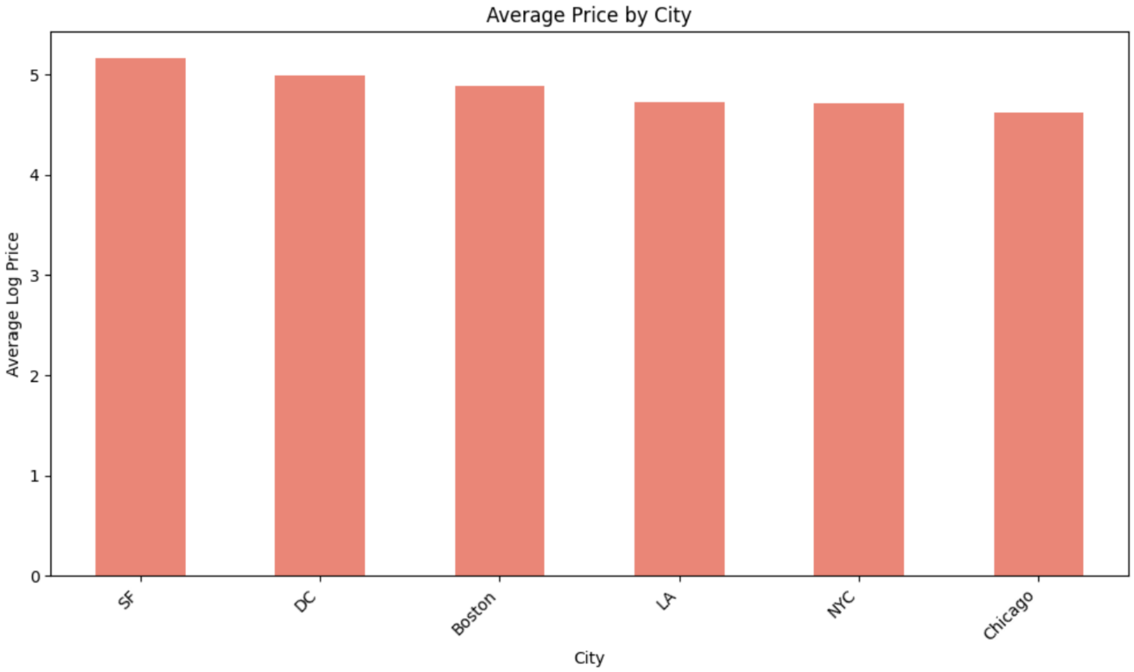




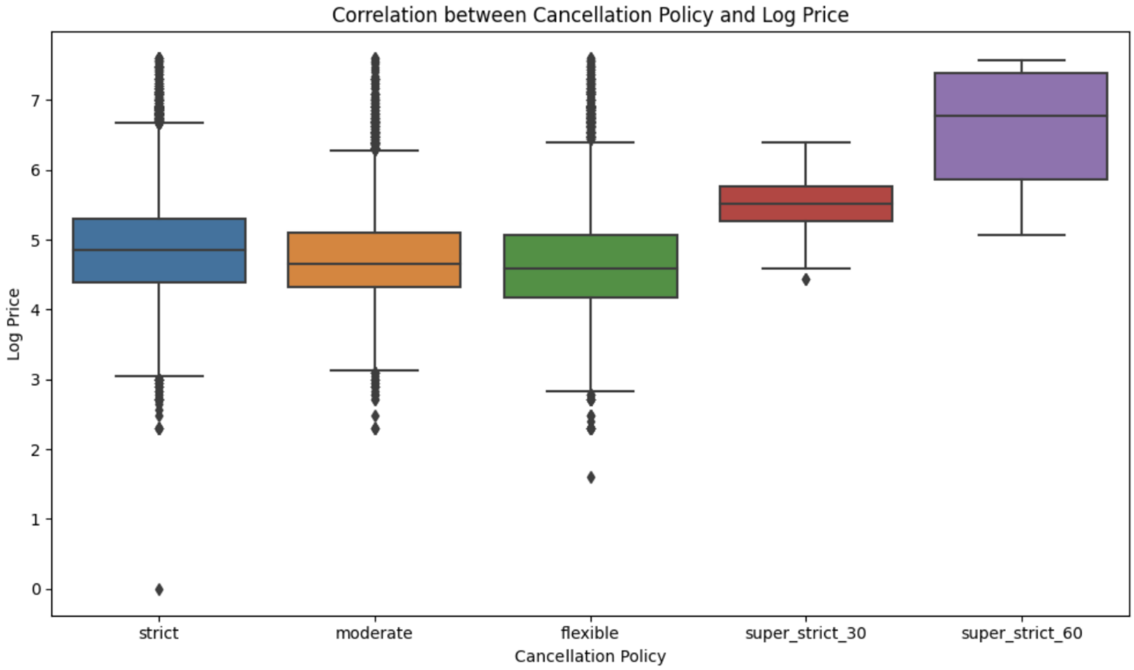
It is crucial for the hosts to understand the distribution of prices so that they can adjust their pricing strategies to increase competitiveness.

The above histogram reveals a peak frequency around a log price of 4.6, indicating a most common price range within the dataset.

This scatterplot shows that the highest number of reviews for a listing is 600, which the log-price lies on around 5. We can observe that the range of log price of around 4 to 5.5 tends to have some of the most reviewed Airbnb listings, which further proves that this range of log price is the most popular price range for both travellers and listeres.



Here is a bar chart for showing average price by city, we can discover that coastal area listing such as San Francisco have a higher average log price, which SF has a log price of over 5; while inland area listing like Chicago have a lower average log price, which the log price is around 4.7.



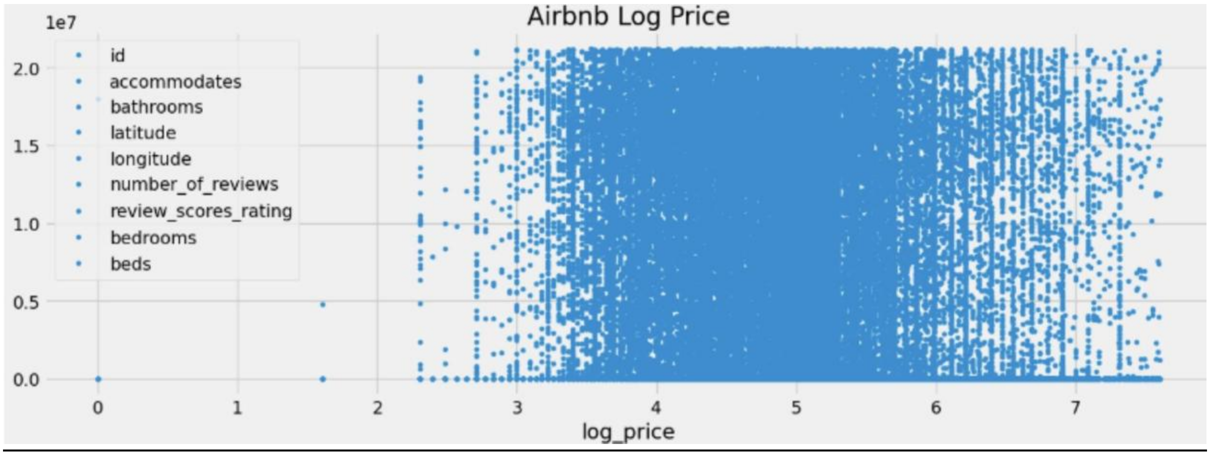
As for the cancellation policy boxplot, strict policy tends to lead to a higher log price as well. Although the log price range among the first three policy have only a slight difference, super_strict cancellation policy will lead to a very high log price. Like for the mean of super strict 60, the log price is already around 7. This may shows that an Airbnb with strict cancellation policy is more expensive since those Airbnb may provide a better living experience compared to the Airbnb with more flexible cancellation policy.

Therefore host and traveler can refer to these graph to check if their price setting or price of target listing is too high or too low. They can adjust and assess the price by criteria like city or cancellation policy

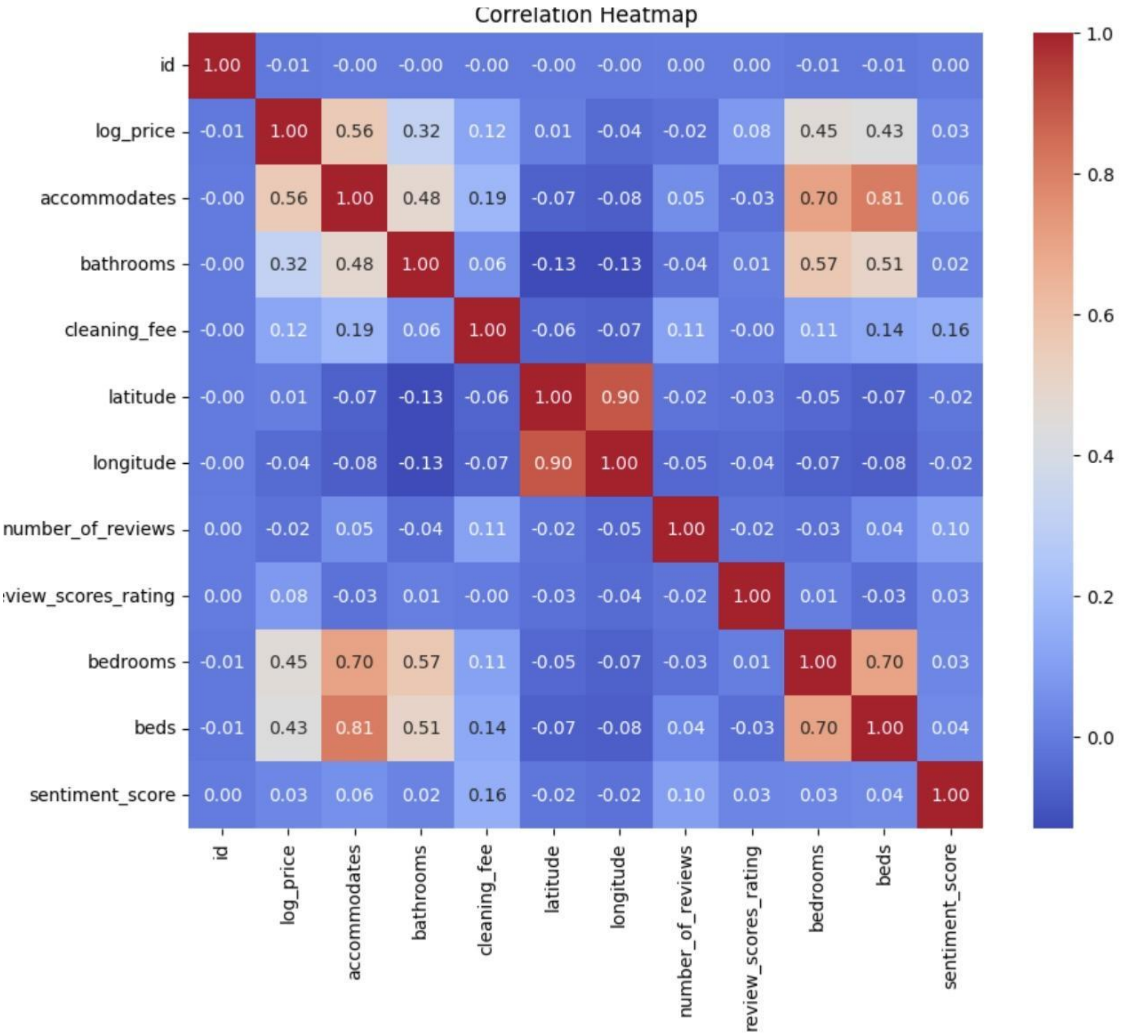
Analytical Procedure

- More than 80% of Airbnb listings have a review ratings higher than 91 which suggests that most listings have a positive sentiment
- Most listings are situated in NYC and LA which are densely populated and have many residential districts.

Result



There is the distribution of those variables. The majority of variables tend to fall within the moderate price range. It is more appropriate to consider our recommendations when the room price is within the moderate range.



There is the correlation analysis of those variables. Log price shows a high correlation with accommodates (0.56).Log price also displays moderate correlations with beds (0.43) and bedrooms (0.45)Furthermore, log price show slight correlations with rating (0.08) and sentiment score (0.03), implying that the se factors have a minor impact on pricing decisions.Conversely, log price shows the least correlation with longitudes (-0.04) and number of reviews (-0.02), indicating that these factors have minimal influence on pricing decisions.

To better price their Airbnb listings, hosts should focus on personal factors instead of environmental factors. Key factors to consider for pricing are accommodates, bedroom types, beds provided. For accommodates, larger accommodations may command higher prices. For bedroom types, different types of bedrooms for example, master bedroom, single bedroom can impact the pricing. Hosts should consider the availability of specific bedroom types. For beds provided, the number and types of beds available in the listing can influence the pricing. Hosts should also avoid unnecessary spending or cost-cutting on external factors that have minimal correlation with pricing decisions. Instead, focus on factors that directly affect the guests' experience and satisfaction. For example, hosts should upgrade their facilities instead of increase review numbers and boost ratings.

By considering these factors and guest needs, hosts can optimize their pricing strategies for maximum profitability while providing value to their guests.

Conclusion

- Linear Regression (Ridge Regularization)
    - Root Mean Squared Error (RMSE): 0.452
    - R-squared (R2): 0.576
    - Boston, NYC, DC had the most expensive airbnb listings on average.
    - LA and San Francisco had cheaper airbnb listings.
    - Eastern Cities have more expensive airbnb listings than western cities.

- XGBoost Model
    - Root Mean Squared Error (RMSE): 0.3778055901761709
    - R-squared (R2): 0.7036202949330803