**Ensemble Project: Airbnb Price Prediction**

—------------------------

Haiwei FU, Mengyu LIANG, Nhat Mai NGUYEN, Jinji SHEN and Vanshika SHARMA

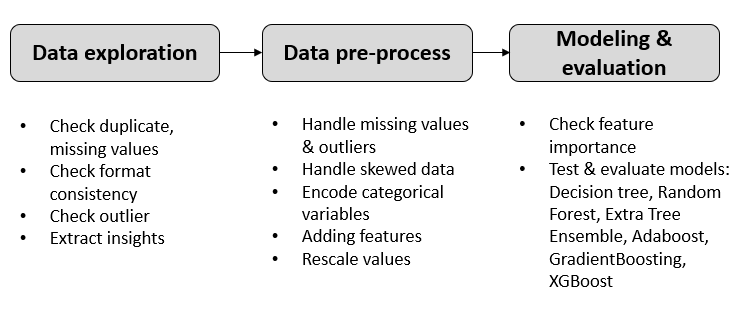
**1. INTRODUCTION**

Predicting Airbnb prices in New York City can be a fascinating project because it offers an excellent opportunity to study and understand the factors that impact Airbnb prices in one of the most vibrant and diverse cities in the world. The project is relevant to real-world applications, and the insights obtained can have a practical impact in the hospitality industry. For example, the hosts and guests can make more informed decisions, leading to better experiences and more effective use of resources.

In the context of this project, we can explore the relationship between different features such as location, amenities, and host characteristics, and how they affect the price of a listing. Additionally, we can investigate how different ensemble methods like Random Forest, Gradient Boosting, etc. can improve the accuracy and robustness of the predictions.

**2. METHODOLOGY**

Our approach to the problem is illustrated as followed:

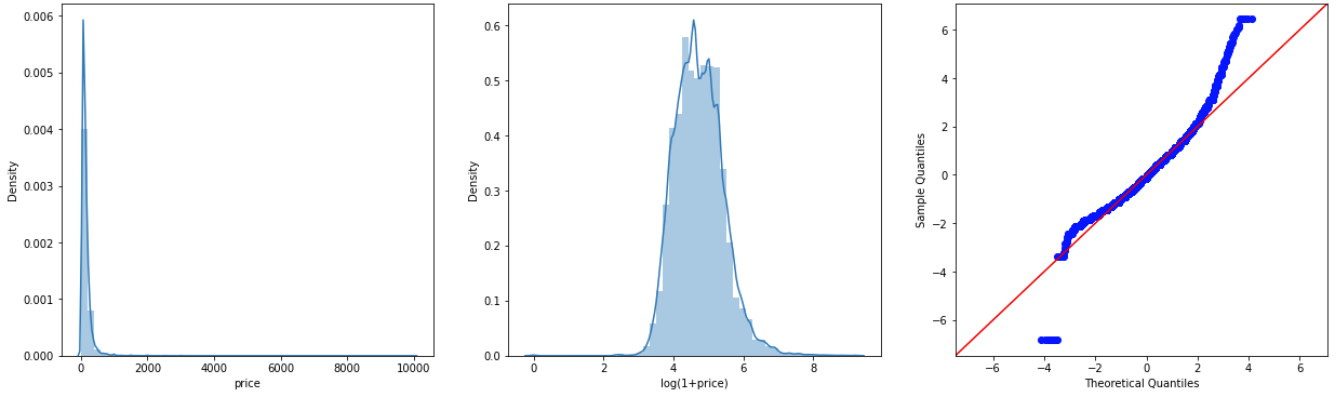


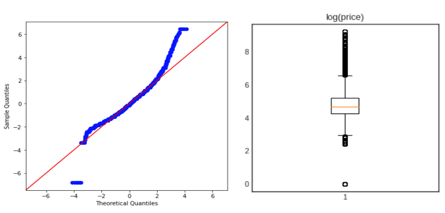
**2.1 Data exploration & pre-process**

From our check, there are no duplicates or format inconsistency but there are about 10k null data in last\_review and reviews\_per\_month. Since these imply the airbnb does not have review, we impute these missing values by 0.

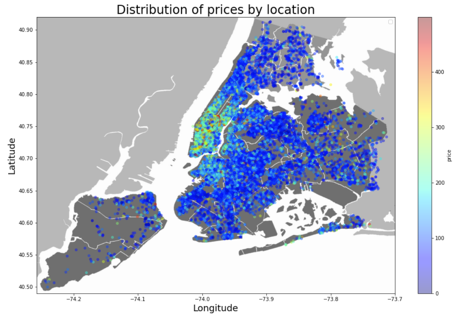
The distribution of price is highly skewed. By taking logarithm, the distribution is closer to normal distribution. Given that many statistical models and machine learning algorithms assume that the target variable follows a normal distribution, we use log(price) as target to ensure the performance of models. Normally, any observations that are more than 1.5 IQR below Q1 or more than 1.5 IQR above Q3 are considered outliers. In this case, we can see that log(price) has the huge number of outliers (22% belongs to "outlier" category). However, such huge number of outliers sometimes can be useful info that reflect the underlying distribution. So, we decide to create 2 scenarios to test out:

* + Consider such 22% data as outliers => remove them
  + Consider such 22% data as useful info for prediction => keep them





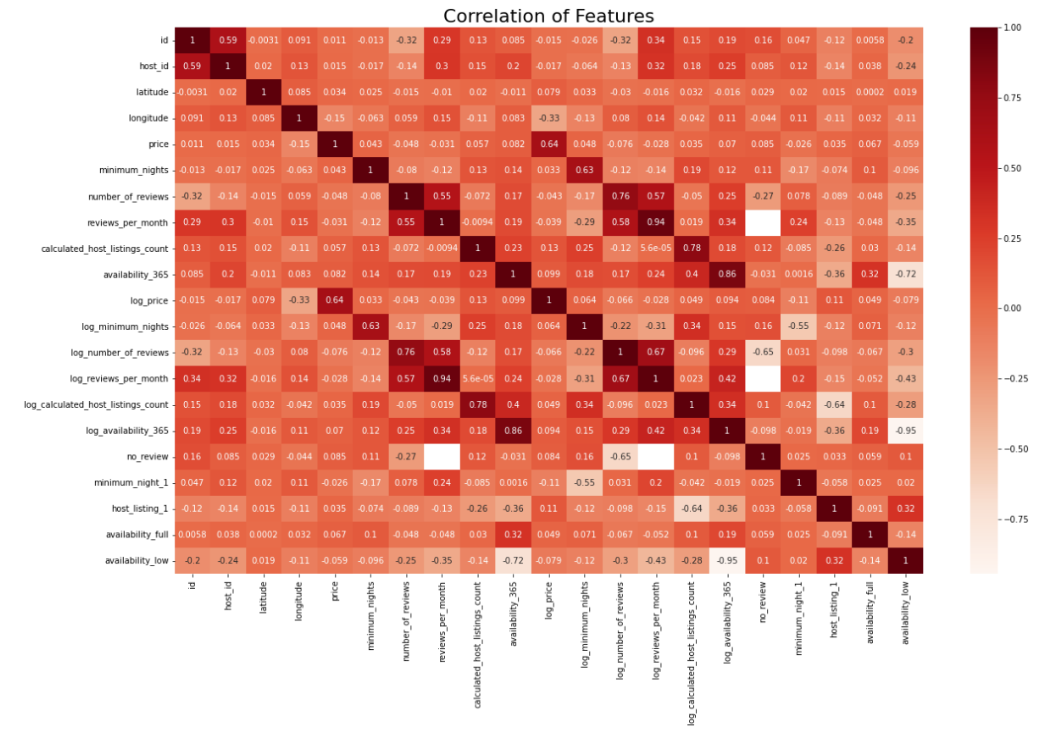
The price highly depends on the location of the apartment. Here, the prices of airbnbs in Manhattan region are higher than other regions





Except for location features, distributions of other numerical variables are high skewed. Though taking logarithm does not help normalize these variables, we keep them as potential features. Besides, since the variables are highly concentrated in certain values, we think other yes/no features such as no\_review, fully\_available, low\_available, etc. can provide some some information for predicting price, we added them to the feature pool.

Most of the variables are uncorrelated (except for newly added features and original features that these new features are extracted from)



After encoding categorical data and rescaling values to avoid the dominance of large-value features, we measure the feature importance by impurity reduction via default random forest model. Location info and room type info seem to have high predicting power. Using such insight, we proceed with the modeling phase.

**2.2 Modelling & evaluation**

Decision tree: a supervised learning algorithm used in machine learning for classification and regression analysis. It creates a tree-like structure where each internal node represents a feature or attribute, each branch represents a decision rule, and each leaf node represents a class label or a numerical value. Decision trees work by recursively splitting the data into smaller subsets based on the most informative features until a stopping criterion is met, such as reaching a maximum depth or minimum number of samples per leaf. They are popular for their interpretability and ability to handle categorical and numerical data, but they can be prone to overfitting and instability.

Random forest

…

Evaluation metrics

**3. CONCLUSION**

**REFERENCES**

[1] <https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data>