New York City Airbnb Price Prediction & Analysis

Akshi Saxena *Rakshit Soni* *Rajat Manish Bhagat*

Summary

New York City has been a very popular tourist destination throughout the years. In 2023, the city predicted approximately 61 million tourists throughout the year. In the early years, these tourists used to live in hotels. As time progressed, the concept of Homestays became popular with people preferring the experience of living in their budget-friendly apartment with the comfort of a homely host rather than a room in the hotel. Today we know these as Airbnbs.

With New York being a popular tourist destination and one of the major cities in the world, staying here is expensive. Hence people visiting for a short period usually book Airbnb which provides more amenities than a regular hotel sometimes, ranging from automated check-in, furnished rooms, kitchens, private parking, and many more. Airbnbs come in many forms, such as Shared rooms to Houseboats!

We are trying to understand the characteristics of a listing and predict its price based on these factors. This will help hosts understand how to price their listings and which popular amenities can help improve the quality of their listings. On the flip side, it will also help users trying to book an Airbnb to understand if the property they have booked is appropriately priced or not.

The project uses data scraped from Airbnb which shows the listings in New York City for the year 2022 and its related data such as which amenities it provides, the price per night, reviews, ratings, property type, etc. The data is stored in a single file *listsings.csv* which has a little over 40,000 records.

Using the standard data science workflow, we first analyzed the data to identify the data types to see if we need to convert them into other types. Then we performed exploratory data analysis (EDA) to understand different trends and distribution of data and whether there were any missing data values and outliers that needed to be handled. We noticed that some transformations would have to be applied to the data to make it useful in a model. Based on this, we created multiple models to try and predict the price of Airbnb accommodations.

Data Source: http://insideairbnb.com/get-the-data/
Dataset:http://data.insideairbnb.com/united-states/ny/new-york-city/2023-03-06/data/listings.csv.gz

Repository:https://github.com/akshi-saxena/airbnb-price-prediction/blob/main/Improvement_Airbnb_Price_Prediction.ipynb

Methods

Method details are available separately.

Results

One of the first observations we made was that the majority of the Airbnbs listed were in the price range of $50-$300 per night. Very few exceeded this range with some exotic ones going for around $10,000 per night. (See Appendix item 11)

Through the correlation matrix (see Appendix item 5), we find the beds, the number of people the listing accommodates and the number of bathrooms have a positive correlation with the price and we also confirm these factors play a major role in estimating a price by looking at each model's importance.

Upon training 4 different types of models, we found the following features to be important.

1. Linear Regression
   We looked at the coefficients to determine which features were important. (see Appendix item 6). Locations like Staten Island, and Manhattan, and room types like shared/private seem to be the most important features.

2. Lasso Regression
   We looked at the absolute value of coefficients that were greater than zero to look at the important features. The important features are review scores and the number of beds in an AirBnB. (See Appendix item 8)

3. Decision Tree
   We looked at the impurity in nodes to determine the feature importances of the decision tree. The important features are the number of people a listing accommodates, host response rate, amenities like air conditioning, and the type of accommodation like a villa or hotel. (See Appendix item 9)

4. Random Forest Regressor
   The feature importances are calculated based on the decrease in the impurity of nodes that make up the random forest for a particular feature that is used in the split and we see

that amenities like hot water, parking, and the nicer areas in NYC are some important features. (See Appendix item 10)

The evaluation metrics (See Appendix item 12) show that the Random Forest Regressor performs the best and it may be best to assume there is some non-linear relationship in the data that is not being captured by the linear model. By looking at the feature importance of other models, we found some of the major factors that affect the price estimation of an Airbnb listing as mentioned above. Overall, the number of beds, the number of people the property can accommodate, the number of baths, and the location as well as the overall rating of the place are key factors in determining the price of the Airbnb listing.

Discussion

The project results can help anyone who is planning to book an Airbnb in New York City. Users can look for the amenities, number of beds and baths, security features, and the overall rating of the listing and determine if the pricing is fair or not. Hosts on the other hand can improve the price of their listing by providing more amenities as well as maintaining a good overall rating. The conclusions that we have provided can help people make better-informed decisions while booking an Airbnb.

There is a lot of scope for future improvements. The dataset that we worked with only contained details about listings in New York City at a point in time. Since the prices of these listings keep on changing based on the seasons (summer would be a better time to enjoy New York) or any event that is happening (concerts, sporting events, etc.), we can also include these changes and predict an optimal travel period also. We can also try using other models like KNN, XGBoost, etc. The created models can be applied to other cities to understand the pricing trend in different markets.

Statement of Contributions

Akshi Saxena

- Data Preprocessing
    - Dropped unnecessary features
    - Missing Value Imputation
- Exploratory data analysis
- Feature Selection and Modeling - Lasso Regression and Decision Trees
- Documentation

Rajat Bhagat

- Data Preprocessing

    - Text Features
    - Binary Features
- Exploratory data analysis
- Feature Selection and Modeling - Random Forest Regression
- Documentation

Rakshit Soni

- Data Preprocessing
    - Features that conveyed similar information
    - Categorical Features
- Exploratory data analysis
- Feature Selection and Modeling - Linear Regression
- Documentation

References

- https://towardsdatascience.com/how-to-tune-a-decision-tree-f03721801680
- https://machinelearningmastery.com/calculate-feature-importance-with-python/
- https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html
- https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/#:~:text=Lasso%20regression%20is%20a%20regularization,i.e.%20models%20with%20fewer%20parameters
- https://cs229.stanford.edu/proj2015/236_report.pdf

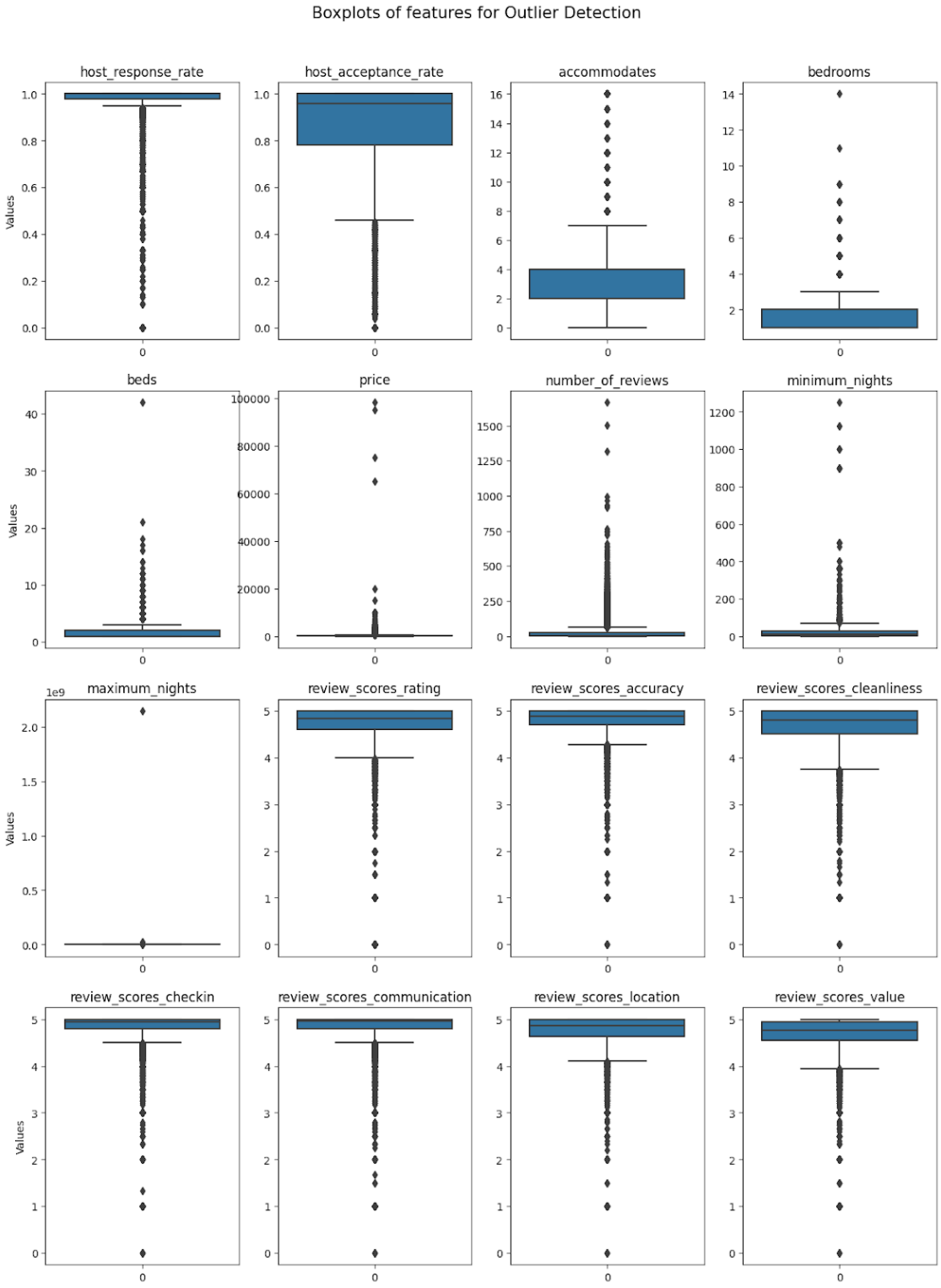
Appendix

1. Table showing the different values in *'property_type'* and to which category they were mapped.

Imputed Value	**Original Values**
Apartment	'Entire rental unit' 'Private room in rental unit' 'Entire serviced apartment' 'Shared room in rental unit' 'Room in aparthotel' 'Room in serviced apartment' 'Private room in serviced apartment' 'Entire home/apt' 'Shared room in serviced apartment'
House / Villa	'Private room in condo' 'Private room in loft' 'Entire loft' 'Private room in townhouse' 'Private room in home' 'Entire condo' 'Entire home' 'Entire townhouse' 'Entire guesthouse' 'Shared room in loft' 'Shared room in home' 'Entire place' 'Private room in guesthouse' 'Entire cottage' 'Tiny home' 'Entire bungalow' 'Shared room in condo' 'Shared room in townhouse' 'Private room in bungalow' 'Entire villa' 'Private room in villa' 'Private room in in-law' 'Shared room in guesthouse' 'Private room in tiny home' 'Private room in vacation home' 'Shared room in bungalow' 'Private room in earthen home' 'Private room in cottage' 'Entire vacation home'

	'Shared room in vacation home' 'Shared room'
Hotel	'Entire guest suite' 'Room in boutique hotel' 'Private room in bed and breakfast' 'Private room in guest suite' 'Private room' 'Room in hotel' 'Private room in hostel' 'Private room in resort' 'Shared room in guest suite' 'Entire bed and breakfast' 'Room in resort' 'Shared room in bed and breakfast' 'Private room in dorm' 'Room in bed and breakfast'
Other	'Private room in houseboat' 'Boat' 'Private room in religious building' 'Cave' 'Floor' 'Houseboat' 'Shared room in floor' 'Private room in floor' 'Private room in casa particular' 'Private room in tent' 'Private room in farm stay' 'Private room in barn' 'Lighthouse' 'Private room in train' 'Barn' 'Private room in lighthouse' 'Casa particular' 'Camper/RV' 'Private room in camper/rv' 'Private room in kezhan' 'Castle' 'Tent' 'Private room in minsu' 'Private room in tower' 'Shared room in casa particular' 'Shared room in shepherd's hut'

2. Boxplots for analysis



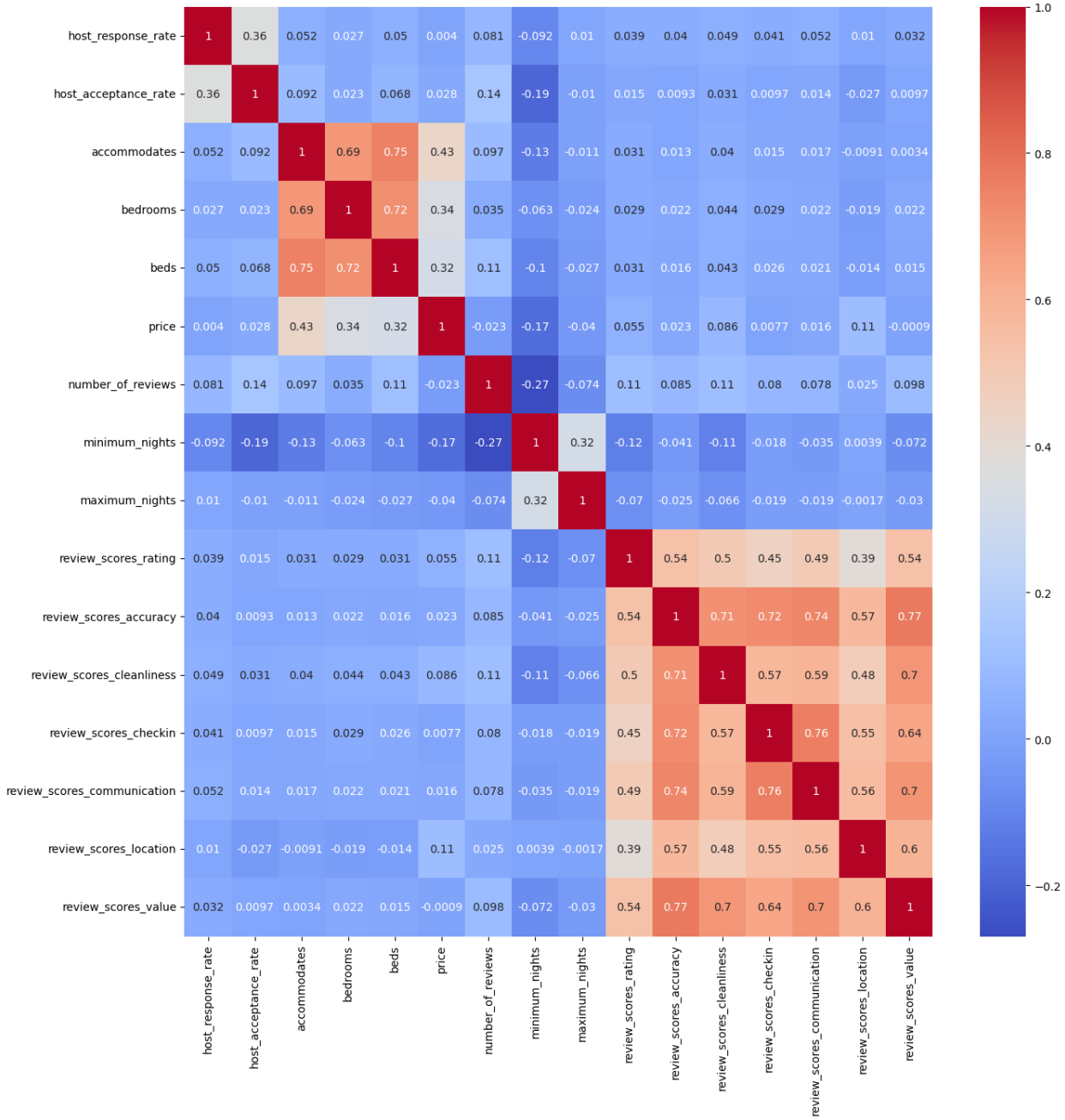
3. Word Cloud for Amenities

4. Columns Missing values that were Imputed and their methods

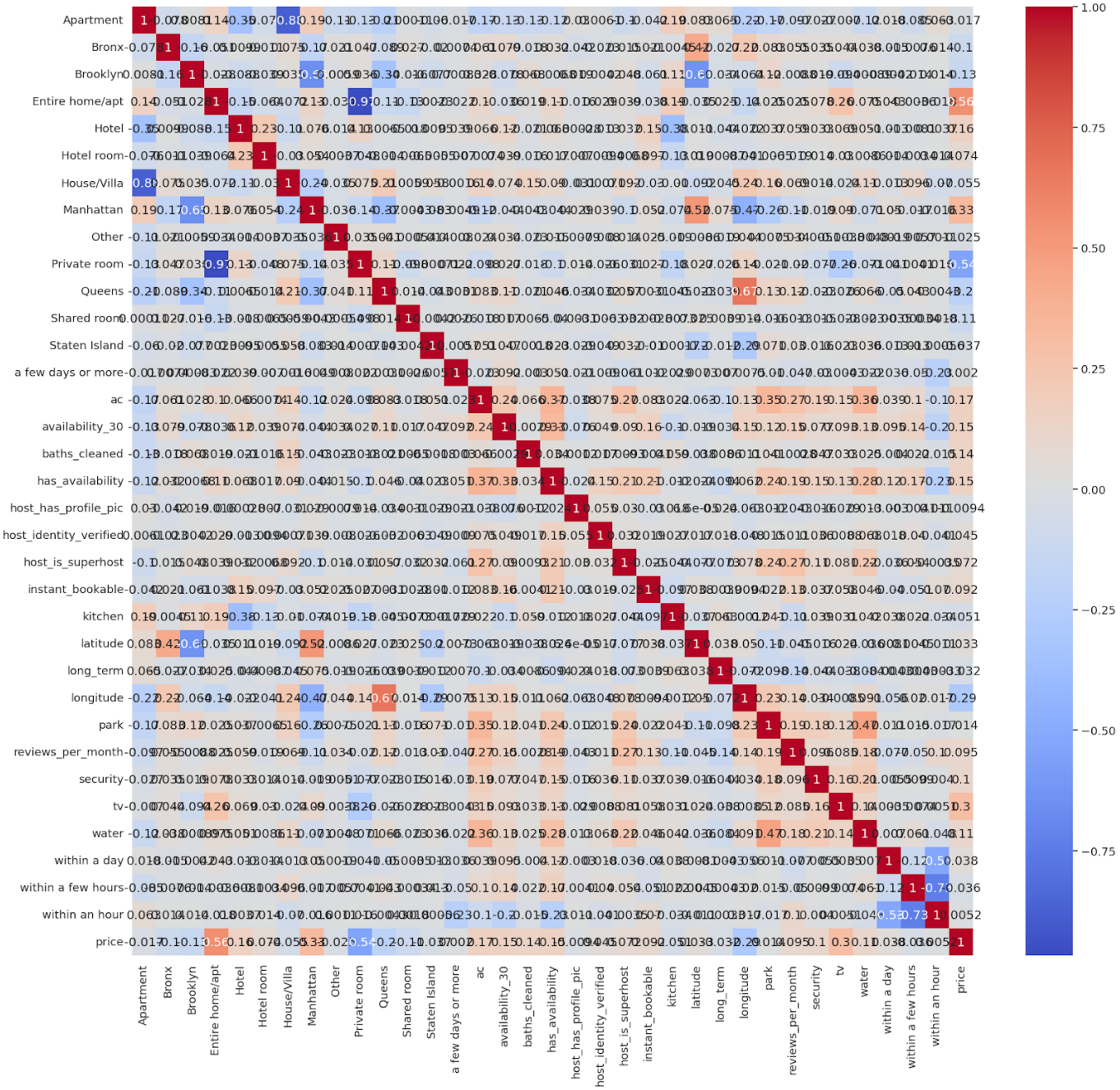
Feature	Missing Percentage	Imputation Method
host_response_time	32.85	One-Hot Encoding
host_response_rate	32.85	Mean value
host_acceptance_rate	29.4	Mean value
host_is_superhost	0.07	Most Frequent Value
host_identity_verified	0.01	Most Frequent Value
bathrooms_text	0.19	Mean value
bathrooms	100.0	Dropped
bedrooms	9.2	Most Frequent Value
beds	2.27	Most Frequent Value
calendar_updated	100.0	Dropped
license	100.0	Dropped
review_scores_rating	22.62	Mean value
review_scores_accuracy	23.69	Mean value
review_scores_cleanliness	23.67	Mean value
review_scores_checkin	23.7	Mean value

review_scores_communication	23.68	Mean value
review_scores_location	23.71	Mean value
review_scores_value	23.71	Mean value

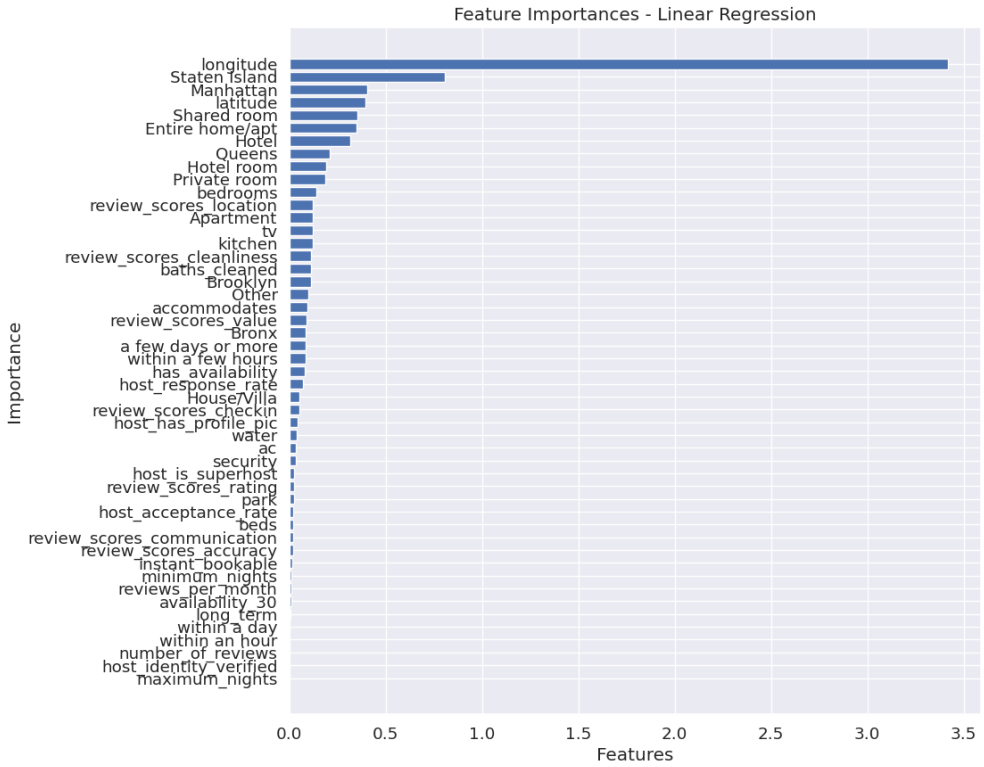
5. Numeric features correlation matrix



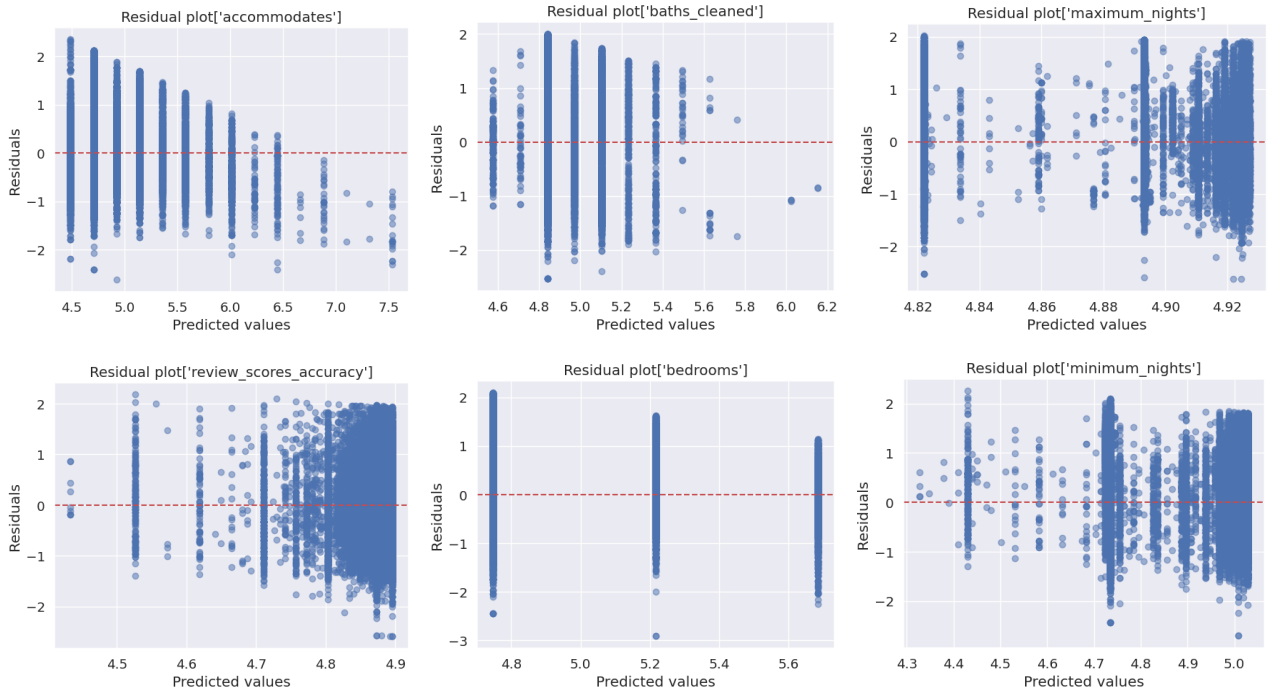
6. Categorical feature correlation matrix



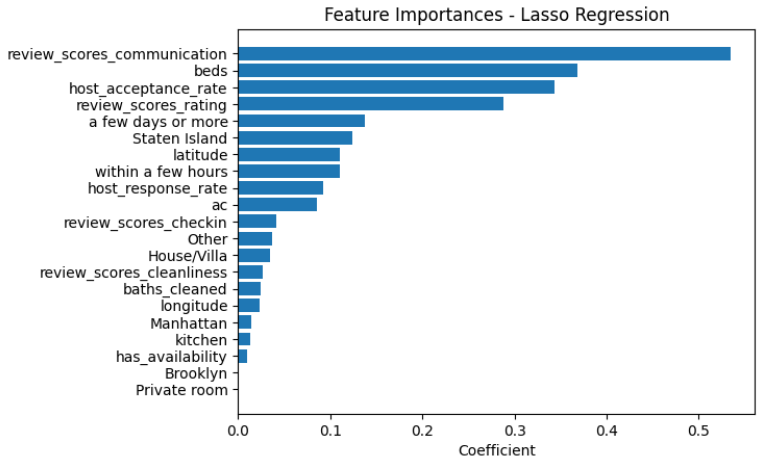
7. Feature Importance - Linear Regression



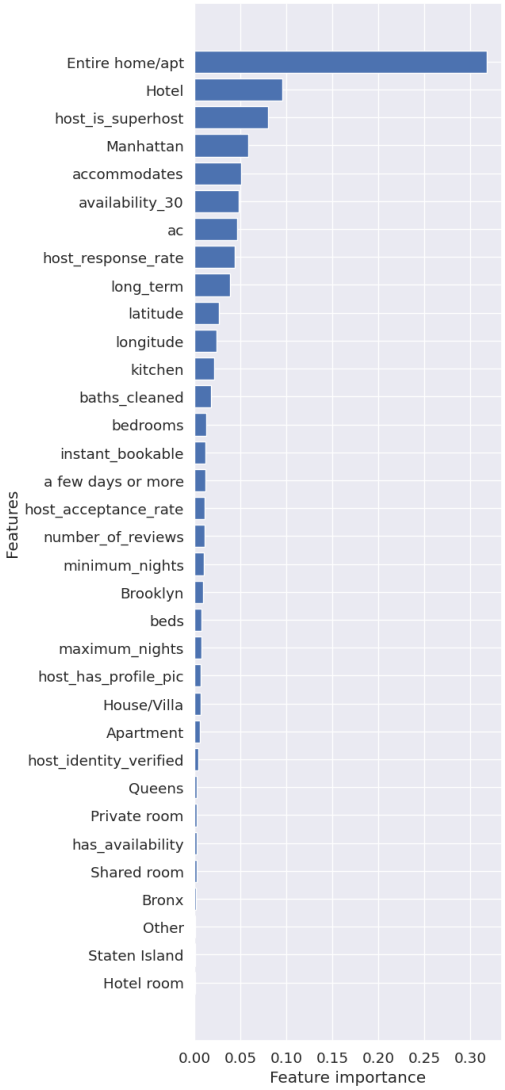
8. Residual Plots of some features



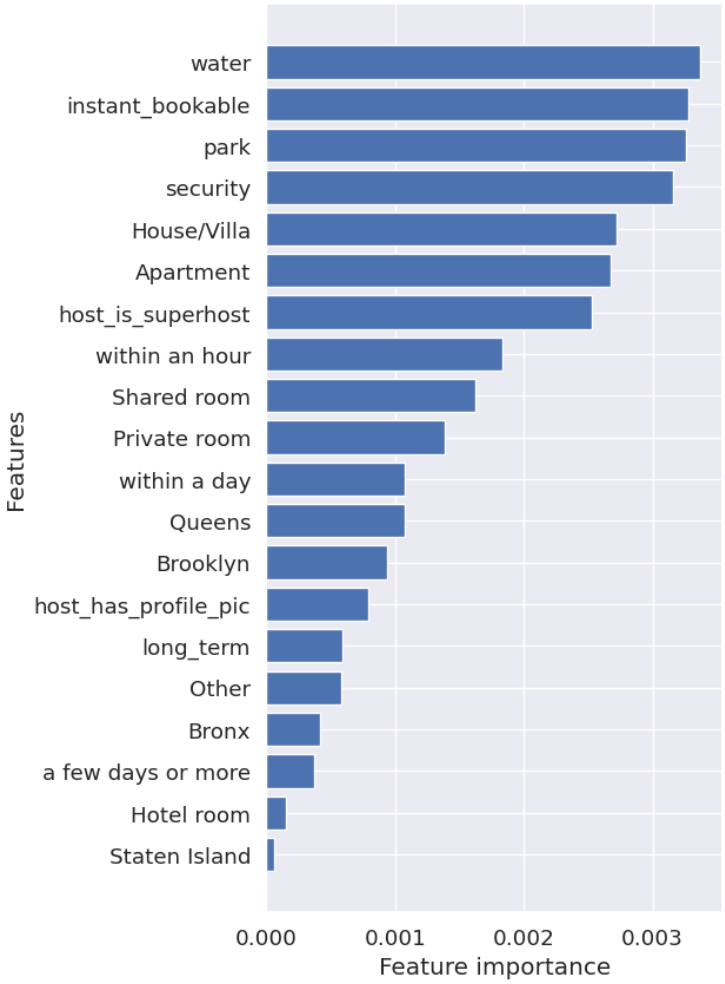
9. Feature Importance - Lasso Regression



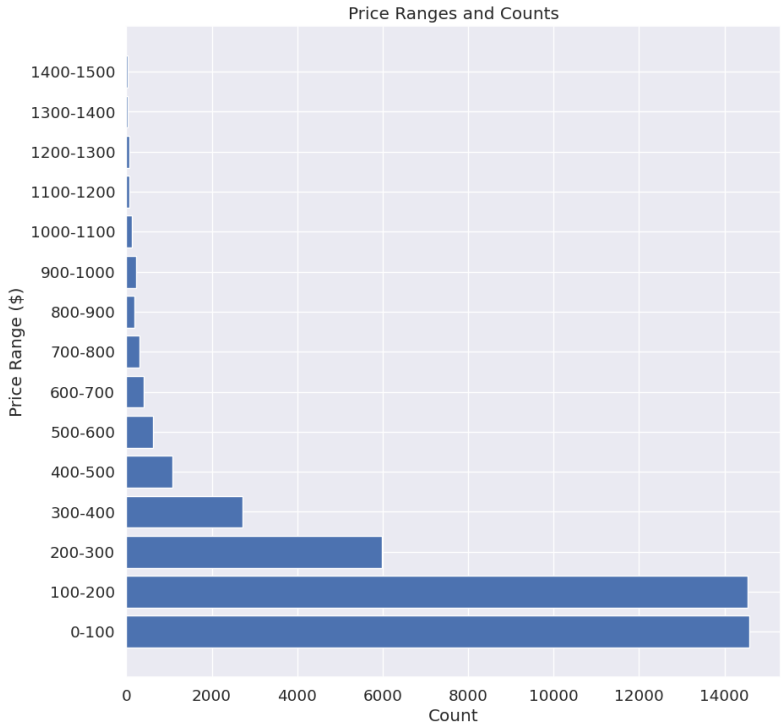
10. Feature Importance - Decision Tree



11. Feature Importance - Random Forest



12. Pricing Range distribution



13. Evaluation Results for all the models

Model	**All Features**		**5 Fold CV Mean Scores**	
	R-squared	**RMSE**	**R-squared**	**RMSE**
Linear Regression	0.629	0.427	0.629	0.426
Lasso Regression	0.629	0.427	0.629	0.427
Decision Tree Regression	0.631	0.426	0.634	0.424
Random Forest Regression	0.741	0.357	0.744	0.354