Regression Project:
AirBnB Pricing Analytics

QBUS3820 | Group 13

490465932
500591890
490151943

27th May 2022

Pages: 20

Table of Contents

Table of Contents 2

Introduction 3

Problem Formulation & Objectives 4

Exploratory Data Analysis 4

Univariate Analysis 4

Bivariate 9

Feature Engineering 14

Methodology 14

Results 18

Data Mining: What are the best hosts doing? 20

References 24

Appendix A: Variables 26

Appendix B: Univariate Analysis for Continuous Variables 28

Appendix C: Univariate Analysis of Text Variables 30

Appendix D: Bivariate Analysis of Continuous and Categorical Variables 33

Appendix E: Bivariate Analysis of Text Variables 36

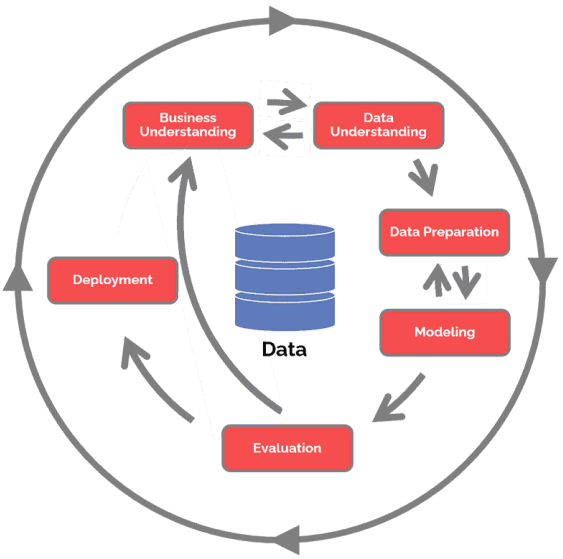
Appendix F: Training and Test Results 38

Introduction

Project

Airbnb is an online platform that connects hosts of rental properties and lodgings with holiday seekers. The current business model involves Airbnb osts deciding on their own pricing strategies, with 3% of the rental price charged as a fee from the company. Lodging offerings on the Airbnb platform vary widely in their characteristics, with some accommodation in a single studio, while others have over ten rooms. As such, the price of rentals also differs significantly across these categories. The core of this project is identifying the factors which correlate with high accommodation fees.

Methodology



The methodology used for this project is the CRISP-DM process through which the team iterates through six key steps. Business understanding is gained by examining the project specifications as well as performing research on Airbnb and short-term rental pricing trends. Similarly, data understanding occurs through a close investigation of the data through exploratory data analysis. Data preparation occurs through the extrapolation of missing data, checking of data entries and removal of unnecessary variables. In the next stage, linear, tree-based and stacking modelling techniques are used to predict Airbnb pricing. These models are then evaluated using cross-validation based on three carefully chosen error metrics. The model which was evaluated to return the most favourable results, was chosen to be the benchmark model. Implementation of the benchmark model will help hosts on pricing as well as owners and investors to predict potential revenue.

Results

Through feature analysis, the key predictors identified were bedrooms, accommodates, beds, latitude, longitude, acceptance_percent, review_rating, host_has_profile_pic, host_is_verified, room_type, has_availability and instant_bookable. These predictors were found to be correlated with price and thus used as the features in the modelling process.

In the modelling process, six types of models ordered based on predictive performance included: Model stacking, optimised forest, random forest, decision tree, ridge regression, and multiple linear regression. Model stacking, which involved the stacking of the optimised forest model with the ridge regression model, performed the best across all three metrics of RMSE, MSE and Adj R^2. Conversely, as the simplest model, multiple linear regression was set as the benchmark model and performed the most poorly.

Problem Formulation & Objectives

This project focuses the analysis of Airbnb accommodation data to explore factors that influence and contribute to accommodation prices on the Airbnb platform. The audience of this report includes a variety of stakeholders including hosts, property investors and Airbnb management. As such, the insights and considerations from this report will be focused on the perspective of increasing revenue rather than from a customer's perspective.

In this regression project, the two objectives are:

1. Modelling the price of accomodation using the provided features.
2. Providing meaningful insights as to the characteristics and behaviours of the best hosts.

To achieve the first objective, six models will be built and evaluated. The models are multiple linear regression (MLR), ridge regression, decision tree, random forests, optimised forest and model stacking. These models will be evaluated on RMSE, MAE, and Adjusted R^2 to determine the best model.

Through data mining, this report will outline key characteristics of the best Airbnb hosts, which is defined as having the highest accommodation prices. These insights can be used to inform the decisions of existing and future Airbnb property investors.

Key areas for consideration include:

- Dealing with redundant variables and missing values.
- Determining the correlation between price and other predictors.
- Exploring which variables have the greatest effect on Price.
- Choosing the appropriate features for modelling.
- Examining the differences between linear regression models and non-linear models
- Choosing the most appropriate model.
- Generating implementable insights for Airbnb hosts.

Exploratory Data Analysis

Univariate Analysis

Continuous and Categorical Analysis

For the univariate variables, we try to perform an exploratory analysis for the data of Airbnb rentals in Sydney. From figure 1, it shows the distribution status of the host rentals in NSW map. It can be noticed that a large amount of rentals are concentrated on the coastline, which is distributed along the coastline. This map reflects a radial distribution from the sea to the city.

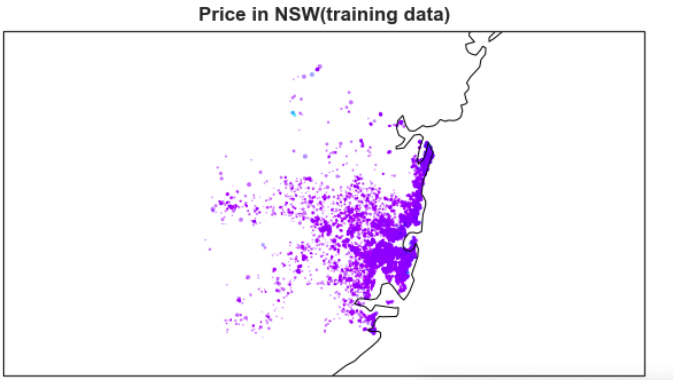


Figure 1: The density of location in NSW map

Our team examines the distribution of the response variable -- daily price in local currency. As can be seen from the histogram (Figure 2), the distribution of the price is right-skewed. So, we treat it into the log transformation, which can make the distribution of the 'price' be close to the normal distribution. After the log transformation, the points approximately follow a relatively normal distribution without any notable outliers.

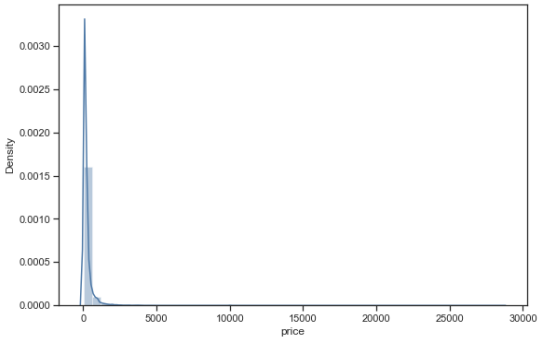


Figure 2: Histogram of the price

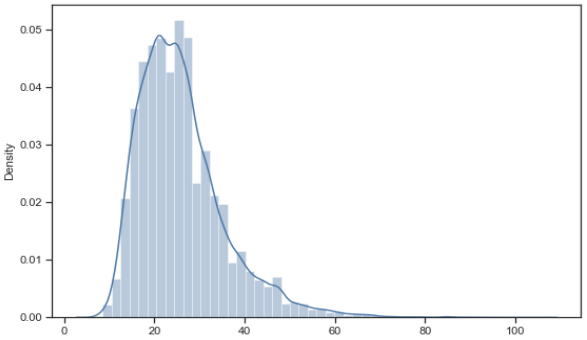
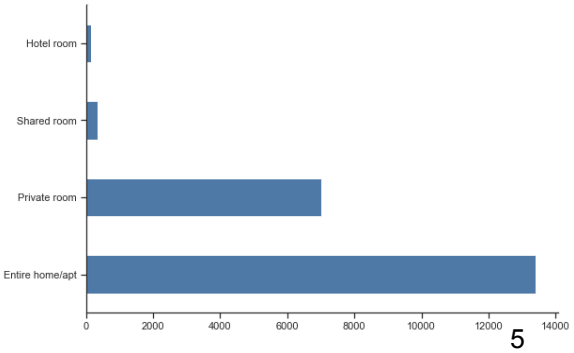


Figure 3: Histogram of the log_price

For the continuous variables, we also examine the plot of different continuous variables; such as Review Ratings and Acceptance Percent. From the plot and dataset insights, we could know that most of them are left-skewed, and at the same time, some of them have a number of missing values. It implies that we may deal with missing values for filling their mean though some missing values are important.

In the categorical variables, we use a predictor to show its distribution – room type. It could be noticed that most of the room type is the entire home/apt. And the lowest one is the room. So, this if it is an entire home/apt, the more customers would choose it, and therefore, the price of it may become higher and competitive.

Moreover, we also need to take the discrete variable into account. We examine the distribution of the 'number of reviews', which can be extremely right-skewed. When wen trying to do the transformation (for log transformation and Yeo-Johnson transformation). The distribution of the Number of Reviews seems to not change into the normal distribution, which has lower bias and variance.



Additionally, we could also find that from the crosstab plots that most of discrete variable has a specific value which is extremely higher than other values. It may indicate that the price probably not change with the discrete variables. Figure 4: Plot of room type

From the Airbnb dataset, we could know that there are four binary variables: host_is_superhost, host_has_profile_pic, has_availability and instant_bookable. Three of these binary variables from the crosstab plots have a large amount of “True” or “False”. We can say that most of host is not super host. Most of hosts have profile picture. A large amount of hosts’ home has availability for customers. And most of guests can automatically book the listing without the host requiring to accept their booking request. It is an indicator of a commercial listing.

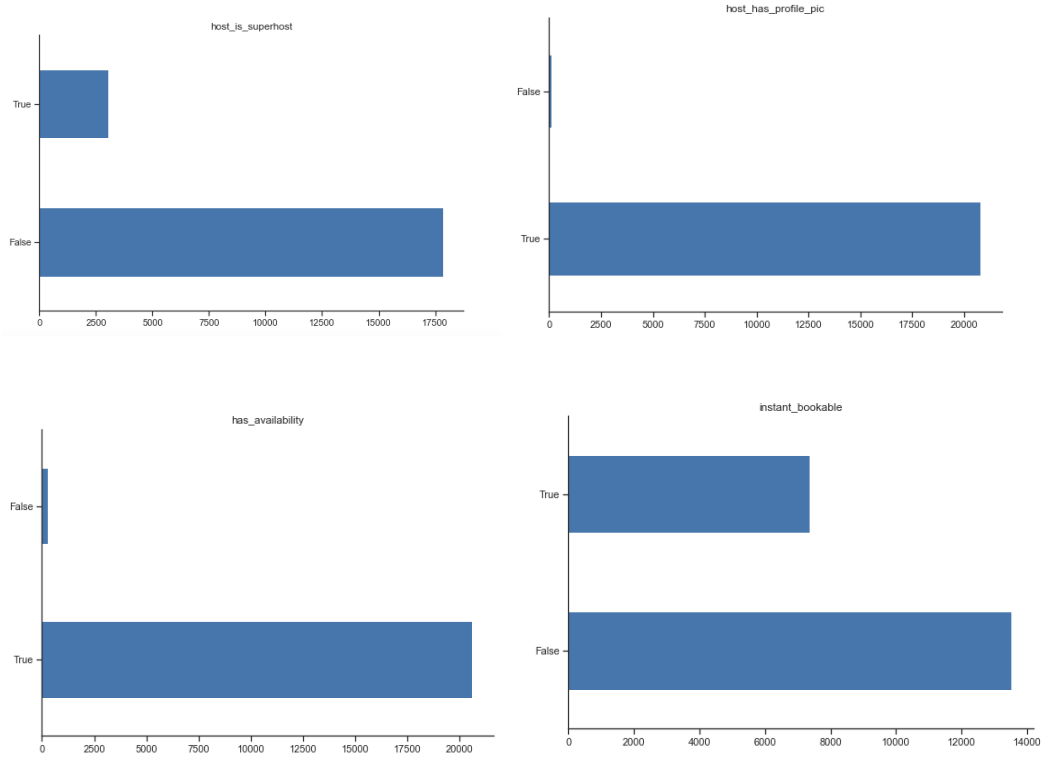


Figure 5: Plots for binary predictors

Through the exploratory analysis for univariate variables, we generally have the knowledge of each variable. Univariate analysis explores each variable in a data set separately. It looks at the range of values, as well as the dominant tendency of the values. This can give better understanding when doing the EDA of bivariate variables in machine learning.

Text

Word cloud

Based on the word cloud in Figure 6 and Figure 7, for variables ‘name’ and ‘description’, ‘apartment’ seems to be a popular choice of word which is followed by ‘beach’ and ‘bedroom’. Since these two variables have similar frequent vocabularies, it may imply multicollinearity. That said, it may be noted the hosts favour numerous words that relate to the beach which may imply that a large proportion believes that it attracts customers.

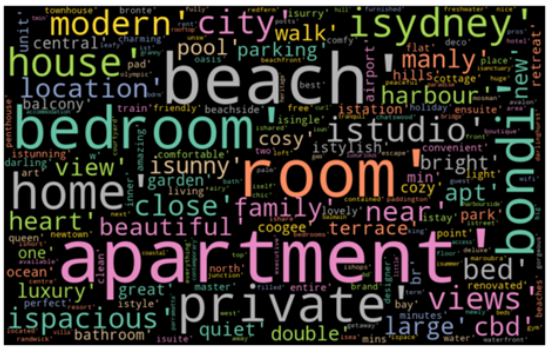


Figure 6: Word cloud based on listing ‘name’

Figure 7: Word cloud on listing ‘description’

Interestingly, the majority of hosts tend to leave their ‘neighbourhood overview’ and ‘host about’ section empty which can be seen in Figure 8 and Figure 9. As for the neighbourhood overviews that were filled out, listings are often advertised by noting down possible leisure activities. For example, eating at ‘restaurants’ or ‘cafes’, otherwise going to the ‘beach’ or on a ‘walk ‘ tend to be popular vocabularies.

Figure 8: Word cloud based on neighbourhood overview

Figure 9: Word cloud based on host about

As expected there are a few common words used for listings when describing ‘amenities’, found in Figure 10. Most interestingly, ‘allowed’ is a very common word used by the hosts although it is not a physical amenity that may be provided by them. Possible connotations for ‘allowed’ are ‘pets allowed’ or ‘smoking allowed’. This implies that the hosts believe that words that implies autonomy have previously attracted or would attract customers. Other common words include convenience and leisure such as ‘washer’, ‘tv’ or ‘wifi’, similar to the previous word clouds.

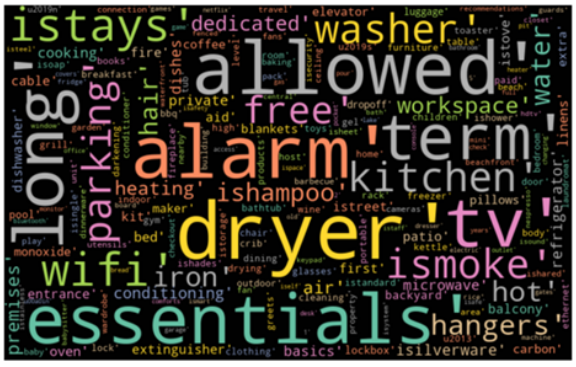


Figure 10: Word cloud based on 'amenities'

Sentence length

Analysis of sentence length for the text variables only found insightful results for three variables: name, description, and amenities. The sentence length of the listing name exhibits a normal distribution that has been positively skewed. Interestingly, sentence length of the listing name is relatively higher than expected with an average of approximately 50 words. This may be attributed to the fact that hosts use the names of a listing to advertise its opportunities within its names Figure 11.

Similarly, in Figure 12, it can be seen that hosts were using amenities to describe all the opportunities their Airbnb provided in comparison to other listings. Based on this inference, it may also be inferred that there would be a positive relationship between amenities length and pricing, as the amenities length represents benefits which would likely lead to an increased price.

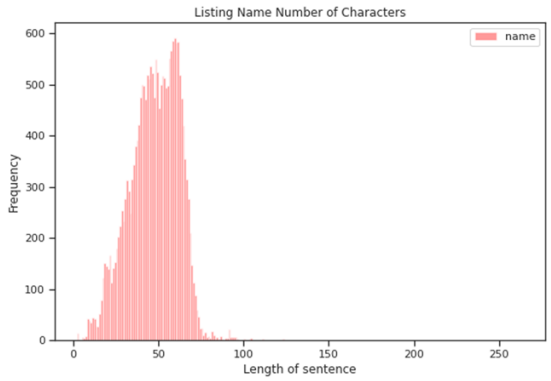


Figure 11: Sentence Length for Listing Name

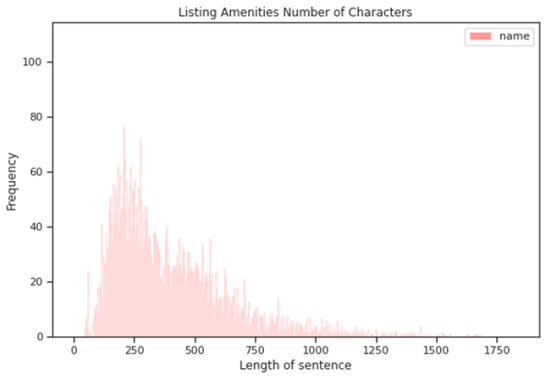


Figure 12: Sentence Length for Amenities

As for description found in Figure 13, there was a uniformed distribution of sentence length, with a large proportion of missing. This abundance lack in description may be attributed to the use of different languages, such as Korean or Chinese. Whereas the uniformed distribution implies that the hosts lack knowledge of the optimal description length. This is likely because while a short description requires less work to read, it has less freedom to advertise the listing.

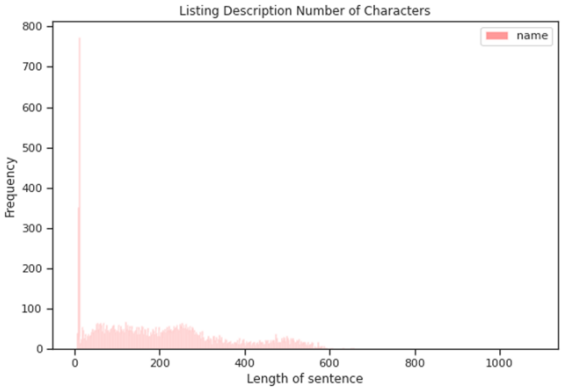


Figure 13: Sentence length for listing description

Sentiment analysis

The only notable sentimental analysis between text variables, that is the majority of hosts writes down a positive description of their listing, seen in Figure 14. Whereas for their own description and neighbourhood overview, the sentiment tends to be neutral. For host description, this may be attributed to the lack of text or effort in leaving a positive impression. Meanwhile, neighbourhood overview is attributed to describing the surrounding activities or opportunities which tend to exhibit neutral to slightly positive sentiment.

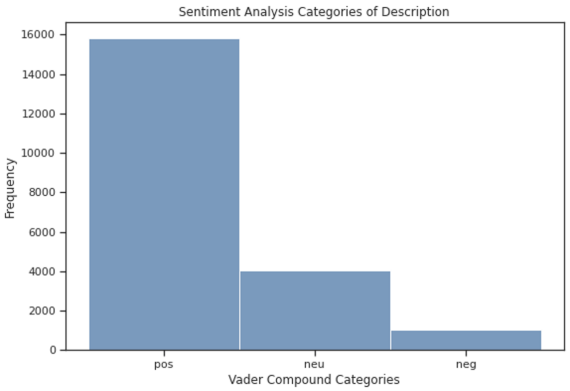


Figure 14: Sentiment Analysis for Description

Bivariate

Numerical

Beds, Bedrooms and Accommodates

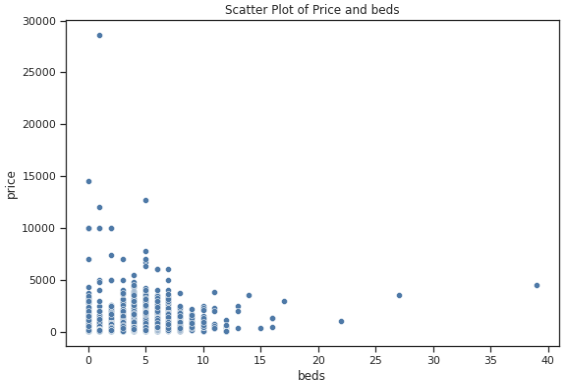


Figure 13: Scatterplot of price and beds

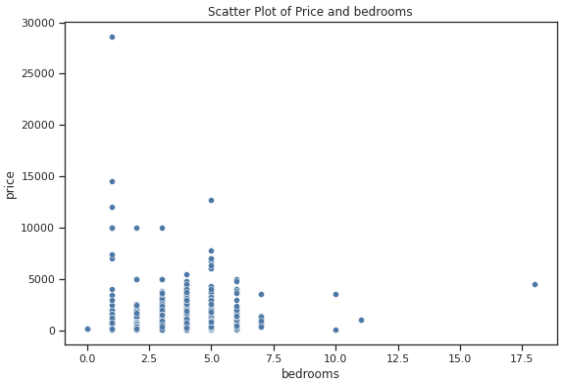


Figure 14: Scatterplot of price and bedrooms

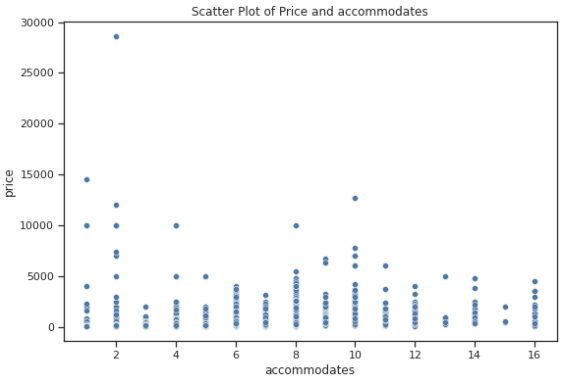


Figure 15: Scatterplot of price and accommodates

The above plots show a weak correlation between these features and price, suggesting a potential non-linear relationship. Interestingly in the accommodates plot, it appears that the price spikes for even numbered accommodates and dips for odd numbered accommodates. There also appears to be some multicollinearity between these three features, with general increases in price in the four to six range for beds and bedrooms, and assuming double beds, an increase in the eight to ten range for accommodates.

Latitude and Longitude

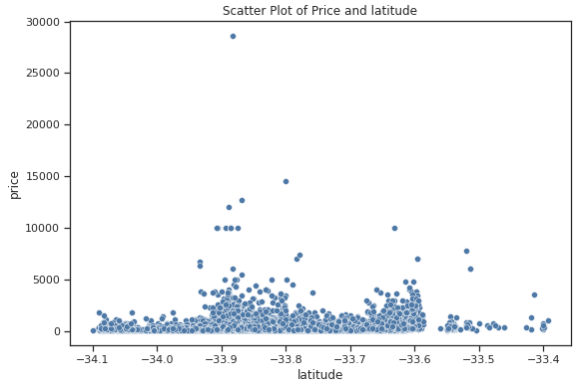


Figure 16: Scatterplot of price and latitude

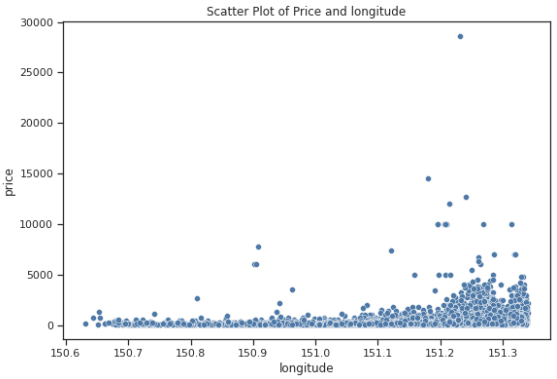


Figure 17: Scatterplot of price and longitude

As seen in the scatterplots, there appears to be a weak relationship between price and latitude and longitude. In the latitude plot, there are two increases in price at around -33.9 and -33.6, which correspond with particular waterfront suburbs such as Bondi. Similarly, in the longitude plot, an increase in longitude is correlated with an increase in price, suggesting accommodation close to the coast has higher prices.

Acceptance percent

Figure 18: Scatterplot of price and acceptance percent

The scatterplot suggests that as acceptance percent increases, the price of the accommodation will also increase, thus there may be a linear relationship. However there are clear spikes in price at acceptance rates of 30% and 50% which weakens the above relationship.

Review rating

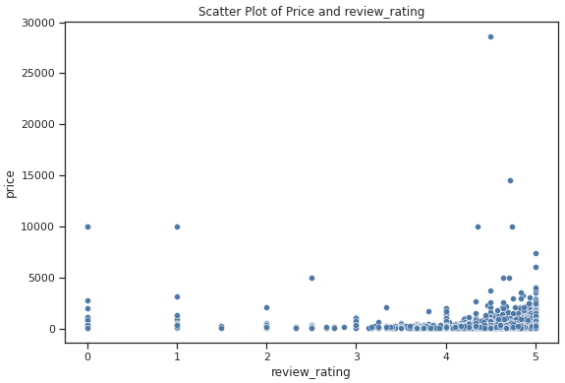


Figure 19: Scatterplot of price and review rating

As seen in the scatterplot, a higher review rating is generally correlated with a higher accommodation price, thus indicating potential exponential relationship between review rating and price. A bivariate analysis of other review features and price, as seen in appendix D, suggests that the relationship between reviews and price are quite similar.

Categorical

Response time

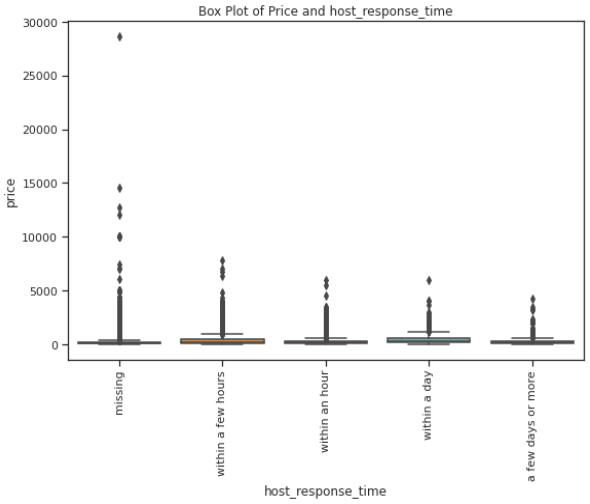


Figure 20: Boxplot of host response time

As seen in the above boxplot, there are a significant number of missing data points, suggesting that this metric is not recorded for all accommodation. Though the distribution of price varies across the different response time categories, there does not appear to be a clear correlation between shorter response times and higher prices.

Host profile picture and verification

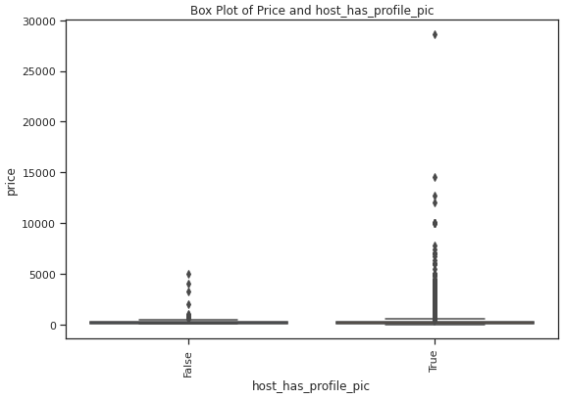


Figure 21: Boxplot of host profile picture

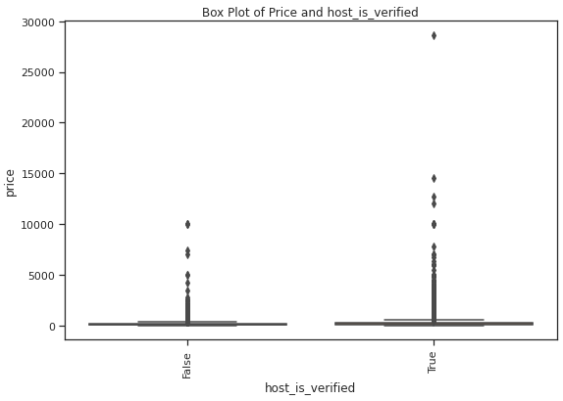


Figure 22: Boxplot of host verification

The profile picture boxplot shows a significantly larger distribution of price for hosts with profile pictures. Similarly, the boxplot for verification shows a larger distribution of prices for verified hosts compared to non-verified hosts. This suggests profile picture and verification are correlated with higher accommodation prices.

Room type

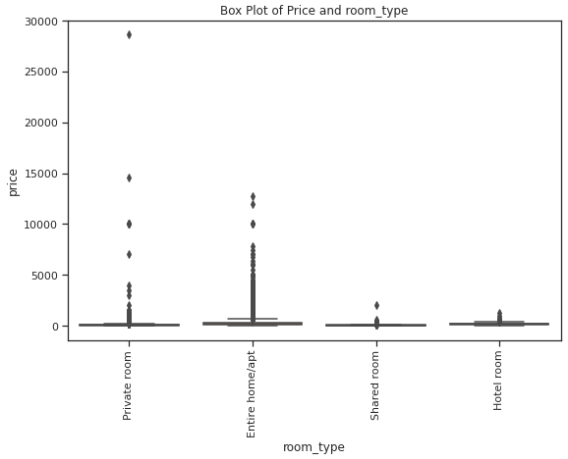


Figure 23: Boxplot of room type

As seen in the above boxplot, the distribution of price for each room type differs significantly across categories. Entire home appears to have the highest average prices, with shared room and hotel room having the lowest average prices. This suggests that room type is correlated with the accommodation price.

Has availability and instant bookable

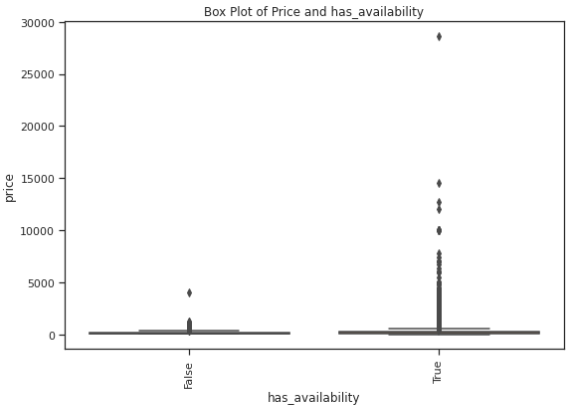


Figure 24: Boxplot of has availability

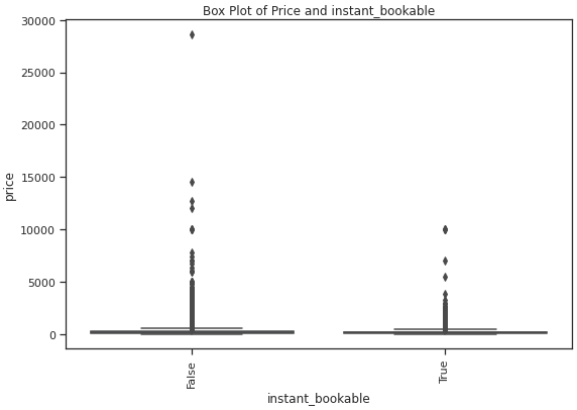


Figure 25: Boxplot of instant bookable

The availability boxplot shows that accommodation with availability has a larger distribution of prices and a higher average price. Conversely, the instant bookable boxplot suggests that non-instant bookable accommodation is correlated with higher accommodation prices.

Text

Length and Sentiment Analysis against Price

Person's correlation coefficient (r) and Spearman's rank-order correlation (rho) was calculated to quantify the linear and non-linear relationship between the text variables (hotjar, n.d.; Swapnilbobe, 2021). Pearson's provides information on the strength and direction of the linear relationship between two numerical variables (SOS, n.d.) Similarly, Spearman's rho measures the strength and direction between variables, but with monotonic relationship – that is an increase in one variable when the other increases but in a non-linear manner (Swapnilbobe, 2021). Accordingly, both metrics are able to determine if two numerical variables are significantly related, thus its predictive value to one another (SOS, n.d.).

Firstly, the relationship between price and the sentiment analysis of the text variables were evaluated; however, both returned weak correlations (Appendix E). Similarly, the relationship between the text length of the text variables and price were evaluated but returned no notable results. This implies that the text variables may not have predictive value towards the price of each listing, regardless of the sentiment it expresses nor the length it exhibits.

Feature Engineering

Various columns were dropped for differing reasons (Appendix A). Firstly, the variables that consisted of URLs, because there are other variables which indicate whether its content exists, meanwhile no other information may be extracted from them – such as 'picture_url'. Secondly, the index variables such as 'host_id', were removed as they do not provide any significant insight. For the same reason, information regarding scraping was removed as well as it does not provide information about the AirBnB, such as 'last_scraped. Then variables which exhibit similar information, that were in similar metrics were removed, such as 'host_total_listing_count' in comparison to 'host_listing_count', or 'availability_60' in comparison to 'availability_30'. Finally, empty variables were dropped because they do not consist of any data despite their label, such as 'bathrooms' and 'calendar_update'. In effect, out of the 73 results, only 54 remained.

Some variables were renamed to avoid confusion, such as making the spelling of 'neighbourhood' to be consistent. A type inference was made and the data type was corrected accordingly - 28 were numerical, 13 were categorical with 5 being binary, 1 was a time variable, 5 was text variables, and 2 were coordinates.. Some variables are corrected or converted into Boolean variables, such as 'licence'. For numerical variables, symbols were removed then converted into floats. Text variables were cleaned and converted into two numerical variables.

In total there were 20880 observations, however, there was a large number of missing values. The proposed modelling techniques are unable to account for missing values and will omit the observation if missing values, thus resulting in a loss of information and a degraded predictive performance (Kumar, 2020). For the text data, a 'missing' string imputed as missingness was data itself, such as for the text variables. Missing numerical values were imputed with either the median or the mean of the remaining values in the column. This method is a lack of consideration of potential covariance between features.

Methodology

Based on the nature of the data, there were six models that were chosen. This included a Decision Tree, a Random Forest, an Optimised Random Forest, a Multiple Linear Regression, a Ridge Regression and a Model Stack. These models were trained using cross validation, and also evaluated with it on RMSE, MASE, and Adj R2.

Cross validation:

Cross-validation (CV) is an evaluation technique used to help train and test a model performance through splitting the data set allowing multiple iterations of evaluations as demonstrated in Figure 26 (Lyashenko & Jha, 2022). The stratified k-Fold was the chosen variation of the standard k-fold CV because it is effective in cases of target imbalance. The technique splits the dataset on 'k' folds, which was chosen to be 10, so that each contains the same percentage of samples for each

target class as the complete set (Lyashenko & Jha, 2022). This technique was used for all the models when fitting and evaluating the models for the training data.

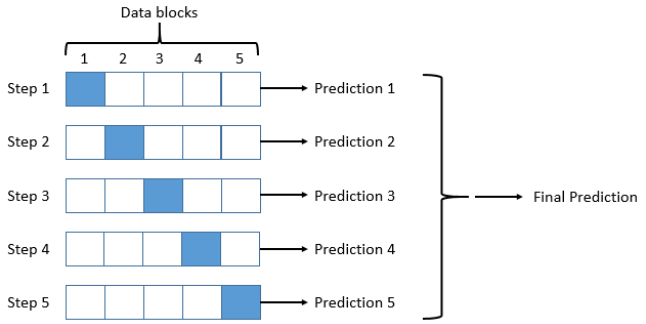


Figure 26: Stratified k-Fold CV

Model Evaluation Metrics

The function of the evaluation metrics intends to work on providing constructive feedback on the predictive value of the models. After we have built our regression models, we can receive the supported feedback from the evaluation metrics (Srivastava, 2019). This provides indication on whether changes to our model results in improvement, allowing the determination of the best predictive model. Besides RMSE evaluation metrics, we still have confusion matrix, log-loss and AUC-ROC matrix to evaluate the regression models. However, we decide to choose RMSE and others popular matrix to evaluate, which is more visualised and comprehensible to the hosts.

The Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) and Adjusted R squared (Adj. R2) were used to compare the models. Root Mean Squared Error measures the standard deviation of residuals. The Mean Absolute Error represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset (Ch

ugh,2020). The Adjusted R-Squared is an improved version of the R-Squared, which penalises addition of new predictors to the model to accommodate for complexity and overfitting (Srivastava, 2019). They were calculated as follows:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \qquad Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1} \qquad \mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n} |y_j - \hat{y}_j|$$

Figure 27: Formula for evaluation metric

Consideration of different evaluation metrics is required to evaluate the models. The choice of metric entirely depends on the type of model and the implementation plan of the regression or classification model.

Linear model:

Multiple Linear Regression

Multiple Linear Regression is a basic modelling method that estimates the relationship between two or more features and the dependent variable. MLRs are highly interpretable with the coefficient of each feature describing the impact of the individual feature on the response variable, keeping all other variables constant (Fabien, 2019). However, MLR relies on the assumptions of homoscedasticity, independence of observations, normality and linearity, which may not always be satisfied, thus undermining the validity of the results (Bevans, 2022).

The MLR model takes on the below formula:

$$price = -70951.211 + 135.640 * bedrooms + 53.395 * accommodates - 20.625 * beds + 542.819 * latitude + 590.996 * longitude - 1.169 * acceptance\ percent + 0.560 * review\ rating - 106.675 * host\ has\ profile\ pic + 18.592 * host\ is\ verified + 30.827 * has\ availability - 21.310 * instant\ bookable + 114.504 * room\ type_{hotel\ room} + 36.772 * room\ type_{private\ room} + 46.365 * room\ type_{shared\ room}$$

Ridge Regression

Ridge regression is a model tuning tuning method that performs L2 regularisation to analyse data with high multicollinearity. By introducing a penalty term to the cost function, ridge regression shrinks the parameters and reduces model complexity by shrinking the coefficients (Jain, 2017). The hyperparameter, λ, controls the impact of the penalty term in the model. When λ = 0, the model produced is similar to the MLR model as no penalty is applied, however as λ → ∞, the penalty increases and the coefficients approach zero. As a result, the model trades lower variance for increased bias (Engati, n.d.).

The Ridge Regression model takes on the equation below:

$$price = -70528.757 + 135.692 * bedrooms + 53.338 * accommodates - 20.600 * beds + 537.855 * latitude + 587.086 * longitude - 1.171 * acceptance\ percent + 0.578 * review\ rating - 105.420 * host\ has\ profile\ pic + 18.619 * host\ is\ verified + 30.712 * has\ availability - 21.338 * instant\ bookable + 113.411 * room\ type_{hotel\ room} + 36.530 * room\ type_{private\ room} + 45.984 * room\ type_{shared\ room}$$

Tree-based model:

Decision tree

The decision tree algorithm is a tree-based model with a set of binary rules used to predict the response variable. At each node a split in the data occurs based on the chosen feature for that node and this process repeats until predictions are reached in the root node (Drakos, 2019). In comparison to other models, the decision tree is beneficial as it is easily interpretable and can capture non-linear relationships. However, decision trees are prone to overfitting, with small changes in data having the potential to significantly alter the tree structure and subsequent predictions (Gurucharan, 2020).

The depth of the decision tree was tuned by iterating through tree depths between 0 and 20. Based on the metric of adjusted R^2, the best hyperparameter for tree depth is 4 as seen in the figure below.

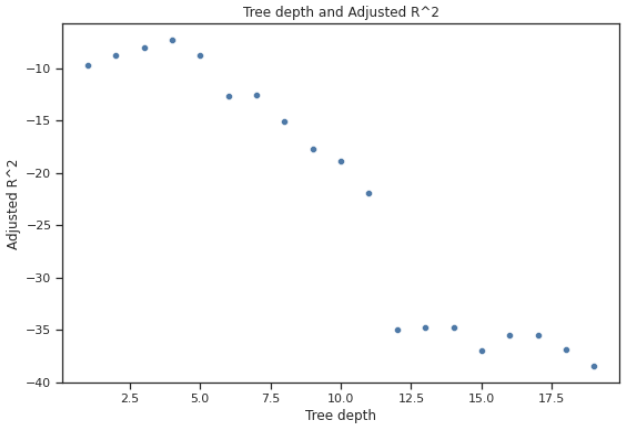


Figure 28: Plot of Decision Tree Depth and Adj R^2

Random Forest

Random forest is an ensemble learning method that combines several decision trees to output an averaged prediction. This mitigates the overfitting nature of individual decision trees and

can improve predictive accuracy (Kho, 2018). Furthermore, random forest models are able to better handle missing values compared to decision trees and are less impacted by small changes in the dataset. However, due to the ensemble learning process, random forests often have low interpretability and thus may not be suitable for detecting key business insights (Mbaabu, 2020). The benchmark random forest model for this project was built using a maximum depth of 10.

Optimised Forest

The random forest model above was tuned to build the optimal forest model with the hyperparameters chosen using a grid search cross validation. The model was tuned on half of the training dataset due to computational complexity and found the best parameters to be number of estimators = 91 and maximum depth = 8.

Ensemble model:

Model stacking

Model stacking is an ensemble machine learning algorithm that combines models using an aggregated method, which can be seen in Figure 29. This is done by utilising a meta-model to combine the predictions of other machine learning algorithms which are called based-models. With ensemble models, although they are able to improve the overall performance, there are a few things to consider. Firstly, model ensembles have very low interpretability due to their complexity. Secondly, ensemble models must be fitted carefully, otherwise they are prone to overfitting. Finally, model training and inference are more time and computationally consuming which can ultimately impact its practical use (Alhamid, 2021). Therefore, although the ensemble model has very good predictive value, it may be difficult to implement due to its computational and time cost.

Figure 29: Model Stacking Architecture

The stacking model constructed utilised linear regression as its meta-model as the target variable is numerical (Brownlee, 2020). The base-models consist of one tree-based model and one regression model which performed the best, to incorporate the unique insight the differing model offers. Accordingly, the current stack model used a regression meta-model which incorporated the optimised random forest model and the ridge regression model as its based-model.

As previously mentioned, based on hyperparameter tuning the optimised random forest was given the max depth of 8, max features of , and 91 number of estimators. Whereas, the ridge regression used the learning rate = 0.99 based on hyperparameter tuning. After being fitted by the training set, the prediction of the optimised random forest was multiplied by the coefficient of 1.191, meanwhile the prediction of the ridge regression was multiplied by -0.102.

Results

Evaluation Scores and Model Comparison

Through cross-validation, a distribution of each evaluation metrics was collected for each model for both the train and test data (Appendix F). Based on the collected error values, a box plot was made for each metric as shown in Figure 30..

Average evaluation score	**RMSE**	**MAE**	**Adj. R2**
Decision Tree	322.658	122.566	-7.271
Random Forest (RT)	306.314	110.996	-6.454
Optimised RT	295.393	110.904	-5.932
Multiple Linear Regression (MLR)	343.917	143.824	-8.397
Ridge Regression	343.902	143.726	-8.396
Model Stacking	**293.766**	**110.163**	**-5.856**

Table 30: Average evaluation scores for test data for each model

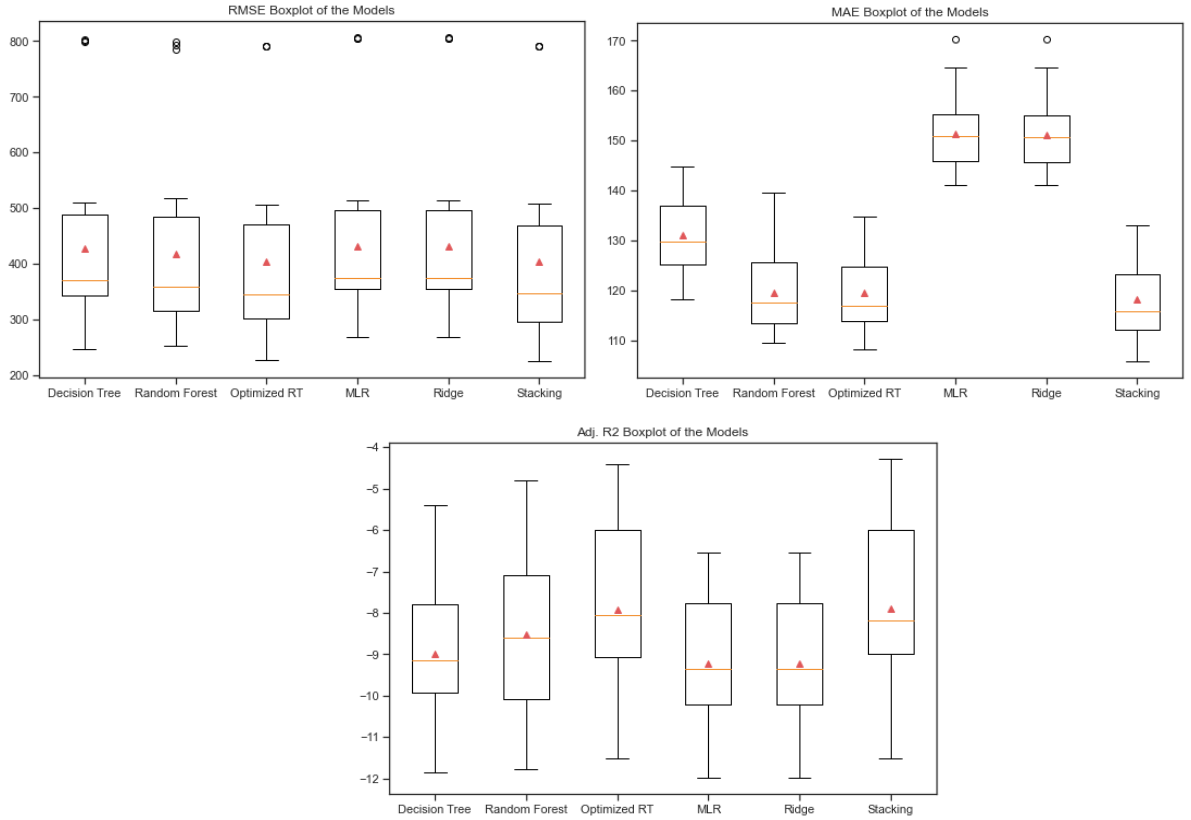


Figure 31: Model metrics boxplots

Based on the MAE box plots seen in Figure 31, it is visible that the tree-based model and the model stack performed significantly better than the regression model. This indicates that the tree based models are more accurate in predicting values than the regression models. That said, when comparing the MAE to the RMSE in Figure 31, the models seem to generally produce the same error.

However, the difference between these two metrics is that RMSE is able to penalise large errors more appropriately. This implies that the regression models may have produced large errors.

The Adj. R2 returns an interesting result, as it takes into account model complexity. Based on the Adj.R2 box plot in Figure 31, it may be seen that the magnitude of the regression models' Adj. R2 are in general greater than the other models, except the Decision Tree. This is interesting as the Random Forest and Optimised RT had both their hyperparameters tuned which likely prevented them from being penalised by addition of unnecessary features. That said, it was surprising to find that the Ridge Regression returned evaluations nearly identical to the MLR. Perhaps its penalty term was not optimised properly to prevent the addition of unnecessary variables.

Ultimately, comparison of model evaluation found that the model stack returned the least errors out of all metrics, which can be seen in Table #. It has incorporated the Optimised RT, the best tree-based model, as well as the ridge regression, a linear regression able to moderate the prediction, which allowed it to exhibit the best predictability. That said, the reason that model stack was chosen over the optimised RT was due to its ability to return a preferable Adj. R2 value, indicating that the incorporation of the ridge regression allowed it to better prevent overfitting, which tree-based models tend to be susceptible towards, as well as preventing it from overestimating the predictions.

It may also be valuable to mention that between the evaluation scores based on the training and testing dataset were relatively similar. Infact, there was a slight improvement in predictability for the test scores (Appendix F). This is a good indication that the models constructed have good generalizability and did not overfit the training model.

Hence, the best model based on predictive power was the ensemble model. Between the tree-based models and the linear models, the tree-based models were able to perform better. This was attributed to a lack of optimization as well as feature engineering and selection. However, stacking of both types of models returned the best results as it further moderates for overfitting.

Benchmark and Best Model

In this project, the multiple linear regression model was established as the benchmark, with the model stacking evaluated to be the best model. When compared against the MLR model, the evaluation metrics for model stacking were significantly improved.

Benchmark model: Multiple Linear Regression

The benchmark model was determined to be the MLR due to its simplicity. Though the decision tree was also considered for the benchmark model, it was determined that the use of hyperparameter tuning to determine the tree-depth added another layer of complexity and thus MLR remained as the ideal benchmark model. Furthermore, the MLR model can be easily interpreted for business insights, with the coefficient of each feature representing the change in price as a result of a unit change in that feature, ceteris paribus.

The equation for the benchmark model:

$$price = -70951.211 + 135.640 * bedrooms + 53.395 * accommodates - 20.625 * beds + 542.819 * latitude + 590.996 * longitude - 1.169 * acceptance\,percent + 0.560 * review\,rating - 106.675 * host\,has\,profile\,pic + 18.592 * host\,is\,verified + 30.827 * has\,availability - 21.310 * instant\,bookable + 114.504 * room\,type_{hotel\,room} + 36.772 * room\,type_{private\ room} + 46.365 * room\,type_{shared\,room}$$

The intercept of the model is -70951.211, with longitude having the large coefficient of 590.996, closely followed by the coefficient of latitude at 542.819. This suggests that the coordinates

of the accommodation are strongly correlated with higher prices. On the other hand, acceptance percent had the lowest coefficient of -1.169, indicating a weak correlation between price and acceptance percent.

Best model: Model Stack

Based on the nature of model stacking and its ability to incorporate unique aspects of other models, as well as return the most favourable evaluation scores, it was determined to be the best model. When comparing the model stacking against the benchmark MLR model, RMSE improved by 50.151, MAE improved by 33.661 and adjusted R^2 improved by 2.541. This suggests that the modelling methods and model selection used to build the model stack are effective in improving predictive performance.

The model stack has two main estimators, the optimised random forest model and ridge regression model. Based on the coefficients, the optimised forest has a greater influence on the predicted value, while the ridge regression moderates the predicted value.

Model	Coefficient	Hyperparameters
Optimised Random Forest	1.191	Depth = 8, features = 4, estimators = 91
Ridge Regression	-0.102	Alpha = 0.09

Data Mining: What are the best hosts doing?

To examine, ‘what are the best hosts doing?’ Both price and occupancy rate have been investigated in conjunction with other variables to extract insight. Accordingly, the ‘best hosts’ were characterised by high prices, assuming that the properties are priced based on their value.

Beds, Bedrooms, Accommodates & Room Type

Through data mining, it is proposed that the best hosts leverage the unique characteristics of Airbnb by offering a large rental area and the opportunity to immerse in the community.

As seen in the bivariate analysis of bedrooms and price, accommodation having around five bedrooms is correlated with higher prices. Furthermore, in the Ridge Regression model, the bedroom feature has a coefficient of 135.692, which indicates it is a large contributor to increased prices. This is likely due to the lack of competitors in this area of the short-term rental market as hotels generally have smaller suites with one or two bedrooms. A similar trend can be seen for the number of beds offered, with around five beds correlated with higher prices. Interestingly, the price spikes for properties that are able to accommodate an even number of guests and dips for properties with an odd number of guests.

When exploring room type, entire homes and private rooms were generally correlated with higher prices in comparison to shared rooms and hotels. This is likely due to the perception that alternative accommodation for shared rooms and hotel room types, such as hostels and non-Airbnb listed hotels, are more economical. As noted by Airbnb (2021), the differentiating characteristic of the platform’s accommodation is enabling guests to “experience a deeper connection to the community … and the people”, a unique opportunity more suited to entire homes and private rooms.

The above data mining suggests that the best hosts leverage two distinctive selling points of the Airbnb accommodation, namely being able to accommodate a larger number of guests than traditional hotel suites and enabling guests to experience the local community at an economical price.

Instant bookable

The bivariate analysis in this project suggests that hosts who use the instant bookable feature in Airbnb may be associated with lower accommodation prices. The instant book feature allows guests to book a property without requiring the host to manually approve the booking, thus increasing the efficiency of the reservation process (iGMS, 2020). However, the use of instant booking may give guests a negative impression of the accommodation, implying that there is a lower effort put in by the host for the pre-accommodation experience. Furthermore, instant booking may suggest that the host is willing to accept any guest to stay, thus reducing the sense of exclusivity associated with the property and resulting in a decreased prices (Nodifi, 2021).

Host has profile picture and host is verified

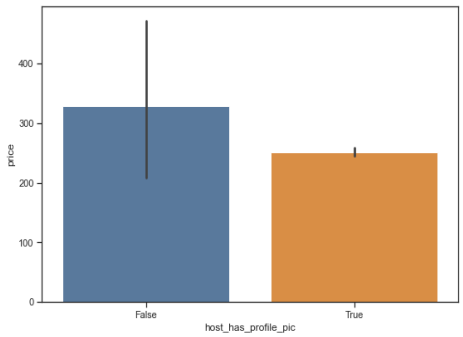


Figure 32: Boxplot of price and profile picture

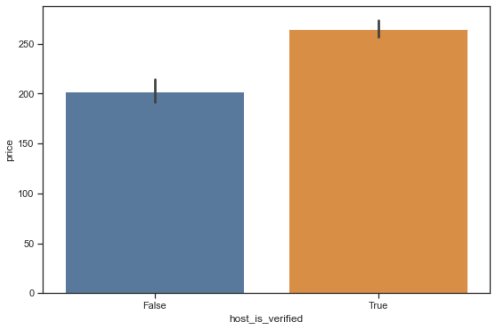


Figure 33: Boxplot of price and verification

From the bar charts in Figure 32 and 33, the 'True' value is in the large amount of these two features.

host_has_profile_pic

If the host has the profile picture online, the price is distributed from 200 dollars to 500 dollars approximately. However, if the host doesn't have a profile picture online, the price of the rentals only stays between 240 dollars and 250 dollars approximately. This price will not increase to a higher price (such as to 500 dollars). This is because the profile picture could represent whether this rental has qualification to operate. It may have more legal power to attract and convince customers to rent his rentals.

host_is_verified

If the host is verified online, the price is distributed above 250 dollars approximately. However, if the host is not verified, the price of the rentals may only change between 170 dollars and 220 dollars approximately. The lowest price of the host is verified is much higher than the highest price of the host is not verified. This price distribution shows that if the host is verified, they can set the price at a higher price and customers will also trust them.

In summary, it is suggested that the hosts need to try to apply the verification first, and then upload the profile picture online. These actions may cause higher profit. It proves your rentals are legal and reasonable.

Host location:

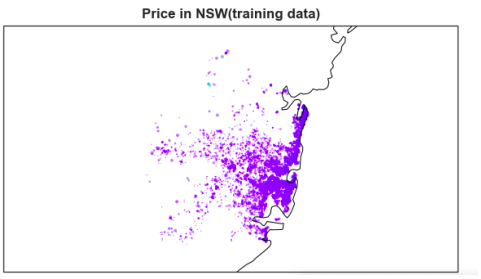


Figure 34: Location of AirBnB, colored by price

From Figure 34, it may be inferred that a large amount of rentals are within the city and distributed along the coastline. This may be because along the coastline, the house has better views of the beach as well as more leisure activities and better restaurants. With the effect of tourism in the seaside, the prices near the sea are higher as well. This is further supported by the text analysis that found a lot of AirBnB hosts choose to advertise the leisure activity surrounding their property, especially characteristics relating to the coastline such as the beach or walking.

On the other hand, properties within the city centre are also found to be extremely popular. This may be attributed to the night life or the quality food that may be found within these areas. This may also be attributed to the facilities such as transportation, or entertainment such as shopping centres, that are more readily available within these areas.

The heavy demand for rental properties coupled with low interest rates reflects that this is an important opportunity for hosts to meet the continuous demand for coastal properties as the population increases (Kollosche M, 2022). As more hosts and investors look to purchase or develop on the coast in Airbnb, the price levels will most likely increase gradually in the coming year. Therefore, it is integral for hosts and investors to snatch up properties within the city or along the coastal area.

Limitation

A limitation to consider for the suggested recommendation is that price is not a holistic measurement of ‘best host’. Other criterias such as occupancy rate or reviews should be considered because price may not necessarily be accurate. It is assumed that the price of the AirBnB is reflective of the listing value, thus reflects the effort of the host which equates to being the Best Host’. However, it is possible that hosts inflate the cost of their AirBnB even though the quality of their listing is not great. Moreover, the listing itself is not the only component that makes a good host. That is, reviews should be taken into consideration as response rate or acceptance rate of listing request is just as important as the quality of the listing itself. This is because lack of timeliness may cause hosts to lose customers or leave negative reviews. That said, there are some fixed factors which hosts cannot change, such as the environment. Naturally, listings at favourable locations such as in the city or the coastline, would have greater traffic. This means that regardless of the effort of the host to improve the listing, these places will naturally garner more attention. Therefore, future data mining should take into consideration more variables, such as location and reviews, when evaluating what the ‘best hosts’ are currently doing.

Conclusion

In conclusion, in an attempt to gain insight regarding what the ‘best hosts’ are doing, the impact of various predictor variables on ‘price’ was evaluated. Firstly, allowing a large number of guests and immersion into local culture. Secondly, hosts who utilised the instant bookable feature. Thirdly, hosts that have a profile picture and those that are verified. Finally, the hosts that obtained and rented out AirBnB within the city or near coastal areas as they provide leisure activity. That said, it must be considered that the metric used to determine the best host is price which is not holistic. Nonetheless, based on the AirBnB price, these are the features which best hosts exhibit and it is recommended that other hosts attempt to follow suit to improve their performance.

References

Airbnb. (2021). What Makes Airbnb, Airbnb. Retrieved from Airbnb: https://news.airbnb.com/what-makes-airbnb-airbnb/

Alhamid, M. (2021). Ensemble Models. Retrieved from Towards Data Science: https://towardsdatascience.com/ensemble-models-5a62d4f4cb0c

Bevans, R. (2022). Multiple Linear Regression | A Quick and Simple Guide. Retrieved from Scribbr: https://www.scribbr.com/statistics/multiple-linear-regression/

Brownlee, J. (2020). Stacking Ensemble Machine Learning With Python. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-pyn/

Drakos, G. (2019). Decision Tree Regressor explained in depth. Retrieved from GDCoder: https://gdcoder.com/decision-tree-regressor-explained-in-depth/

Engati. (n.d.). What is Ridge Regression?. Retrieved from: https://www.engati.com/glossary/ridge-regression

Fabien, M. (2019). Interpretability and explainability (Part 1). Retrieved from Explorium: https://www.explorium.ai/blog/interpretability-and-explainability-part-1/

Gurucharan, M. (2020). Machine Learning Basics: Decision Tree Regression. Retrieved from Towards Data Science: https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda

hotjar. (n.d.). What is a heatmap? Retrieved from hotjar: https://www.hotjar.com/heatmaps/

iGMS. (2020). What Is Airbnb Instant Book and How to Use It Effectively. Retrieved from: https://www.igms.com/what-is-instant-book-on-airbnb/

Jain, S. (2017). A comprehensive beginners guide for Linear, Ridge and Lasso Regression in Python and R. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/

Kollosche Marketing. (2022). Kollosche team. Retrieved from https://www.kollosche.com.au/media-centre/gold-coast-rental-demand-on-the-rise/

Kumar, S. (2020). 7 Ways to Handle Missing Values in Machine Learning. Retrieved from Towards Data Science: https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e

Lyashenko, V., & Jha, A. (2022). Cross-Validation in Machine Learning: How to Do It Right. Retrieved from Neptune Blog: https://neptune.ai/blog/cross-validation-in-machine-learning-how-to-do-it-right

Mbaabu, O. (2020). Introduction to Random Forest in Machine Learning. Retrieved from Section: https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning

Nodifi. (2021). Members Only: How Exclusivity in Marketing Creates Hype. Retrieved from: https://nodifi.com.au/resources/members-only-how-exclusivity-in-marketing-creates-hype/

SOS. (n.d.). Pearson Correlation and Linear Regression. Retrieved from Statistics Online Support: http://sites.utexas.edu/sos/guided/inferential/numeric/bivariate/cor/

Swapnilbobe. (2021). Spearman's Correlation. Retrieved from Analytics Vidhya: https://medium.com/analytics-vidhya/spearmans-correlation-f34c094d99d8

Wikipedia. (2022). Spearman's rank correlation coefficient. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

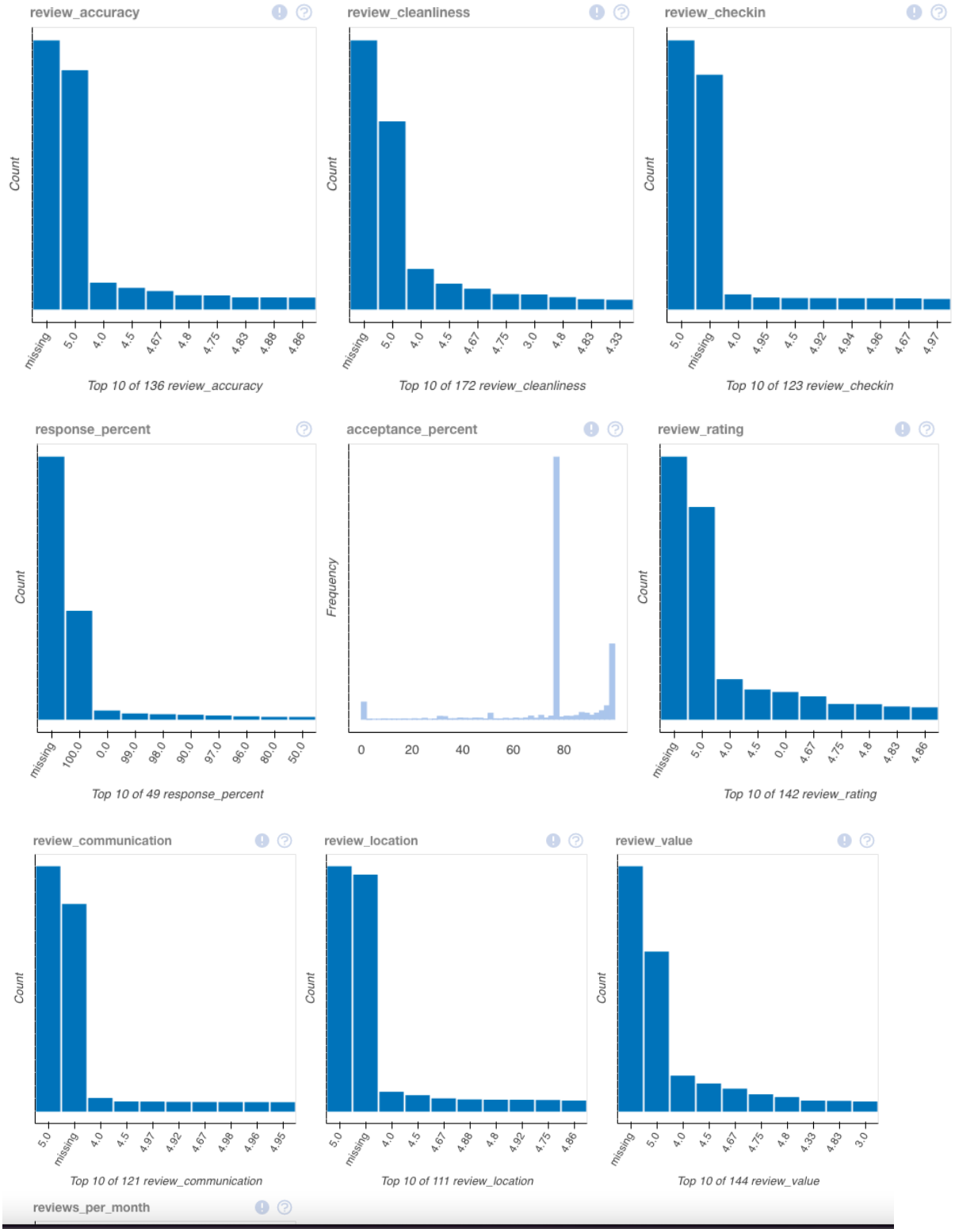
Appendix A: Variables

Variables	Dropped	Rename
id	Yes	-
listing_url	Yes	-
scrape_id	Yes	-
last_scraped	Yes	-
name	No	-
description	No	-
neighborhood_overview	No	Neighbourhood overview
picture_url	Yes	-
host_id	Yes	-
host_url	Yes	-
host_name	No	-
host_since	No	-
host_location	No	-
host_about	No	-
host_response_time	No	-
host_response_rate	No	-
host_acceptance_rate	No	-
host_is_superhost	No	-
host_thumbnail_url	Yes	-
host_picture_url	Yes	-
host_neighbourhood	No	-
host_listings_count	Yes	-
host_total_listings_count	No	-
host_verifications	No	-
host_has_profile_pic	No	-
host_identity_verified	No	-
neighbourhood	Yes	-
neighbourhood_cleansed	No	neighbourhood
neighbourhood_group_cleansed	Yes	-
latitude	No	-
longitude	No	-
property_type	No	-
room_type	No	-
accommodates	No	-
bathrooms	Yes	-

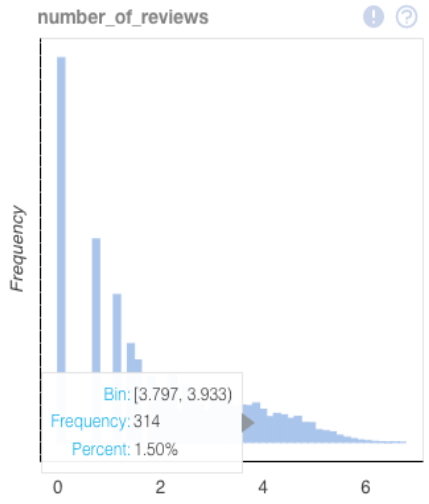
bathrooms_text	No	-
bedrooms	No	-
beds	No	-
amenities	No	-
price	No	-
minimum_nights	No	-
maximum_nights	No	-
minimum_minimum_nights	No	-
maximum_minimum_nights	No	-
minimum_maximum_nights	No	-
maximum_maximum_nights	No	-
minimum_nights_avg_ntm	No	-
maximum_nights_avg_ntm	No	-
calendar_updated	Yes	-
has_availability	No	-
availability_30	No	-
availability_60	Yes	-
availability_90	Yes	-
availability_365	No	-
calendar_last_scraped	Yes	-
number_of_reviews	No	-
number_of_reviews_ltm	Yes	-
number_of_reviews_l30d	No	-
first_review	No	-
last_review	No	-
review_scores_rating	No	-
review_scores_accuracy	No	-
review_scores_cleanliness	No	-
review_scores_checkin	No	-
review_scores_communication	No	-
review_scores_location	No	-
review_scores_value	No	-
license	No	-
instant_bookable	No	-
calculated_host_listings_count	No	-
calculated_host_listings_count_entire_homes	No	-
calculated_host_listings_count_private_rooms	No	-
calculated_host_listings_count_shared_rooms	No	-
reviews_per_month	No	-

Appendix B: Univariate Analysis for Continuous Variables

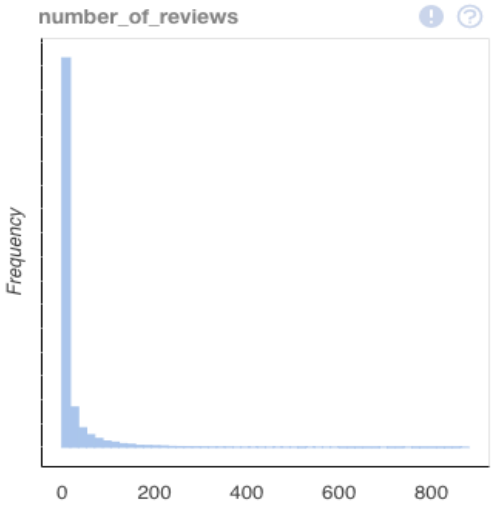
Plot of continuous variables



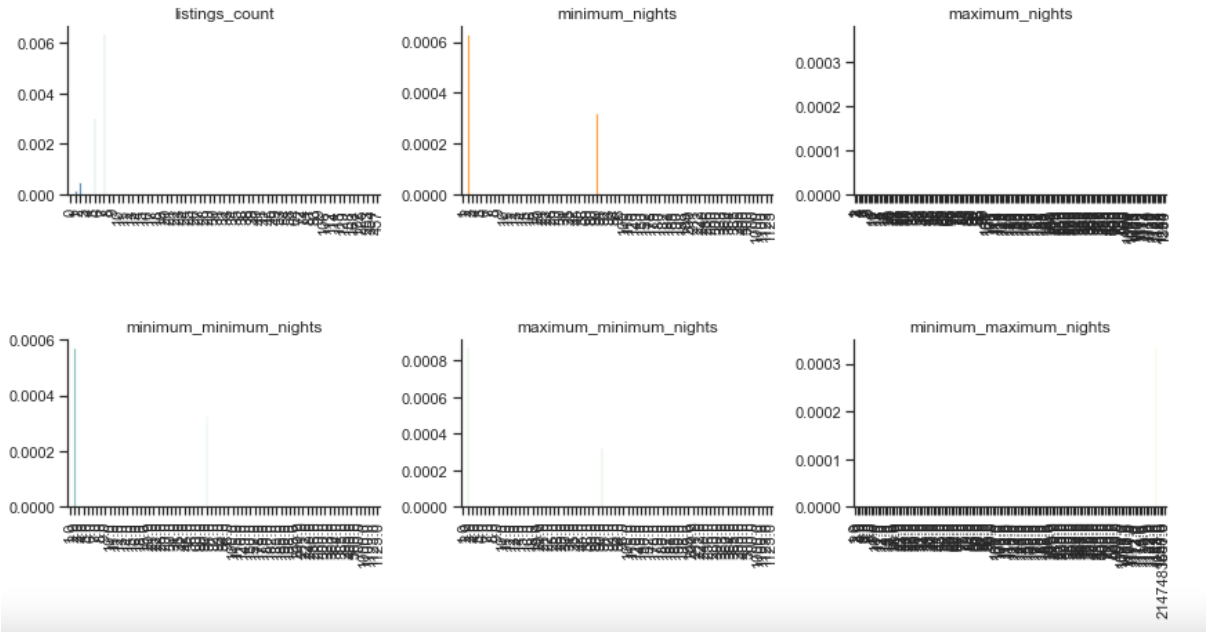
The plot of number_of_reviewers



the log transformation for number_of_reviewers



Plots for discrete variables

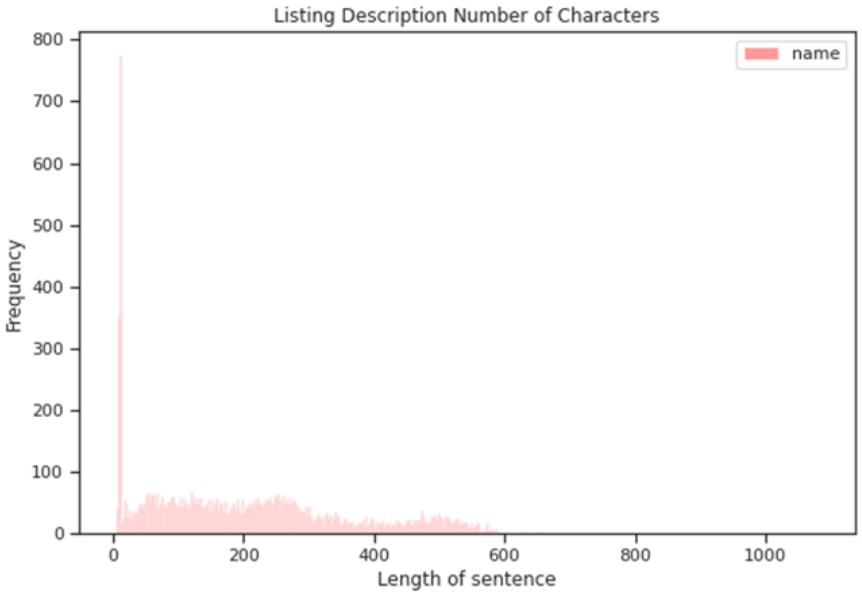


Appendix C: Univariate Analysis of Text Variables

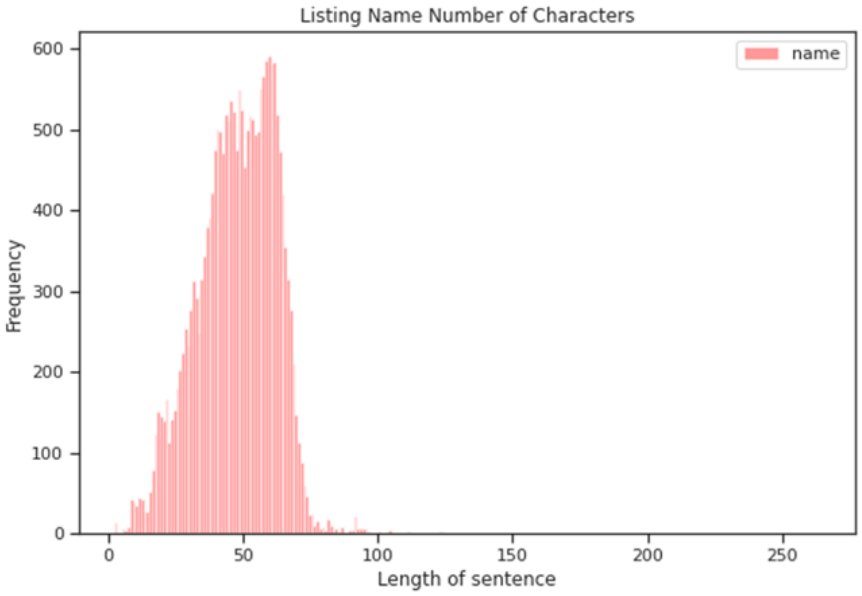
WORD LENGTH

	name	description	neighbourhood_overview	host_about	amenities
nobs	20880	20880	20880	20880	20880
minmax	(2, 264)	(2, 1083)	(2, 1128)	(2, 4070)	(2, 1833)
mean	47.044157	221.256466	135.636638	183.270738	405.54478
variance	214.71523	23731.96454	27394.72897	90798.155875	66585.567802
skewness	0.204324	0.567335	1.74765	3.078497	1.391397
kurtosis	5.482322	-0.541551	3.535091	15.234867	2.238502

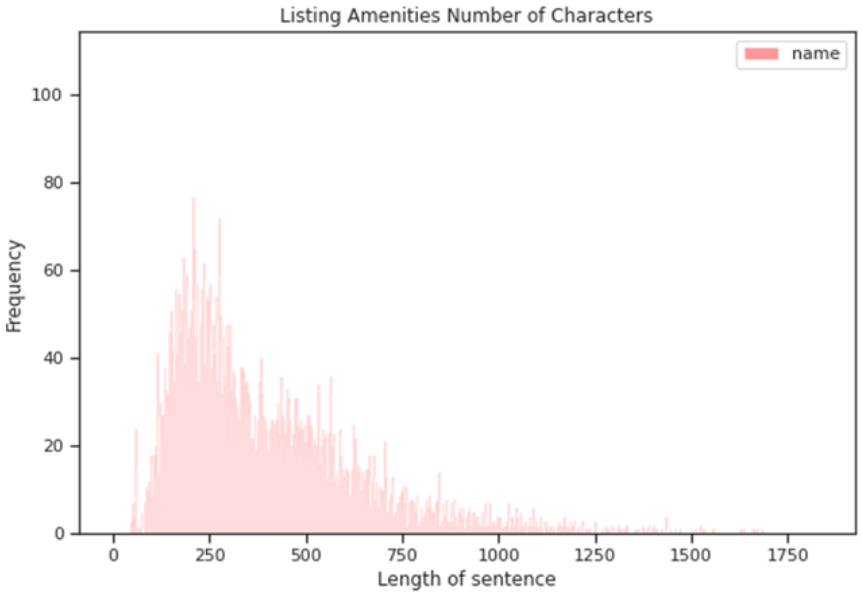
Description:



Name



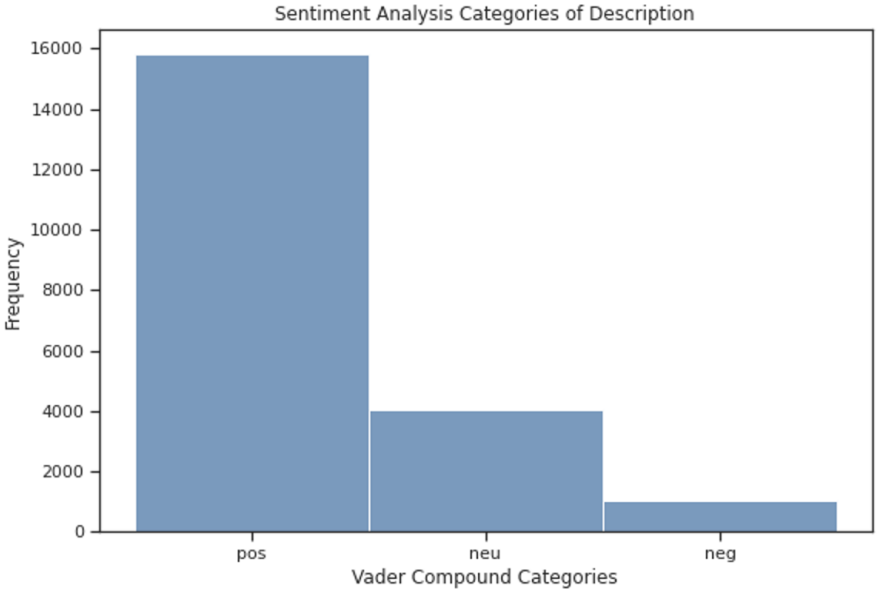
Amenities



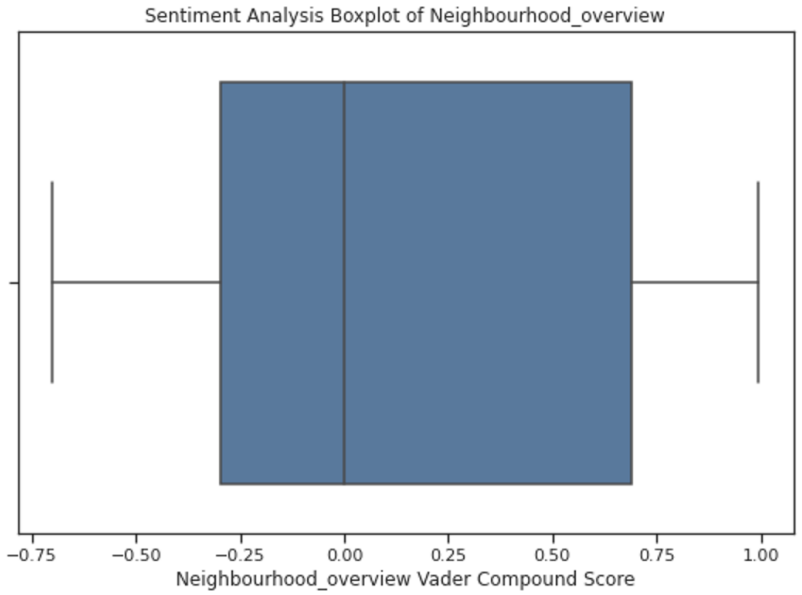
SENTIMENT ANALYSIS

	nobs	minmax	mean	variance	skewness	kurtosis
neighbourhood_overview	20880	(-0.7003, 0.9935)	0.18421	0.238532	0.353603	-1.542209
host_about	20880	(-0.8316, 0.9992)	0.261759	0.323385	0.180849	-1.827304
descrip	20880	(-0.7176, 0.9948)	0.541832	0.151903	-0.700831	-0.866962

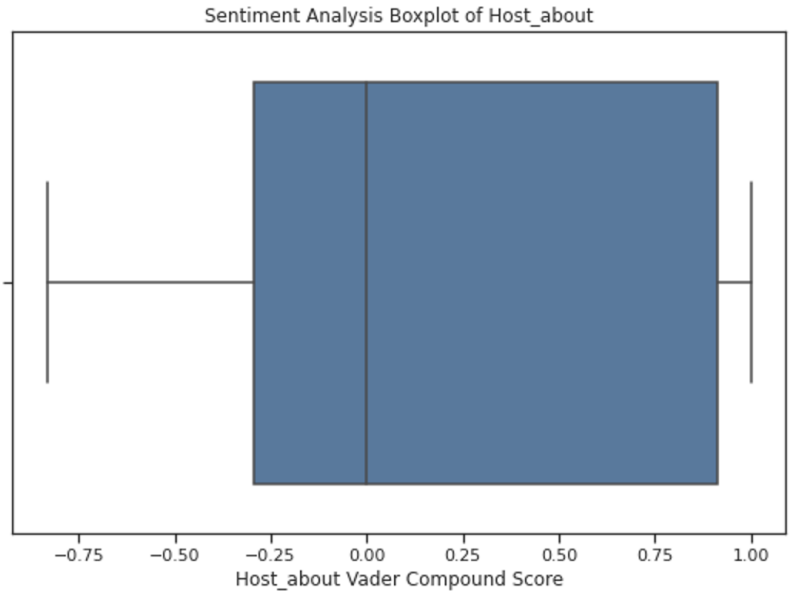
Description:



Neighbourhood overview:

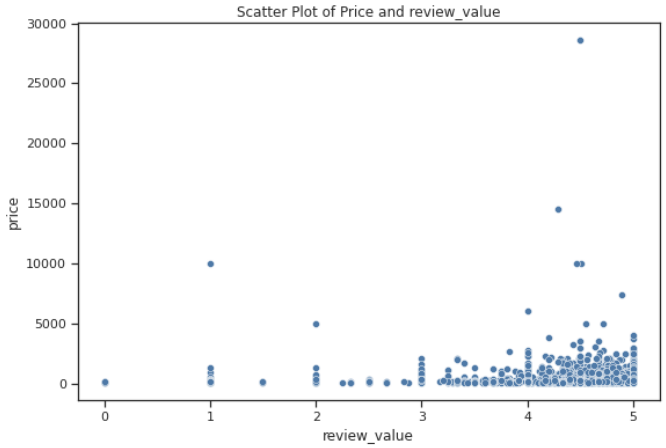


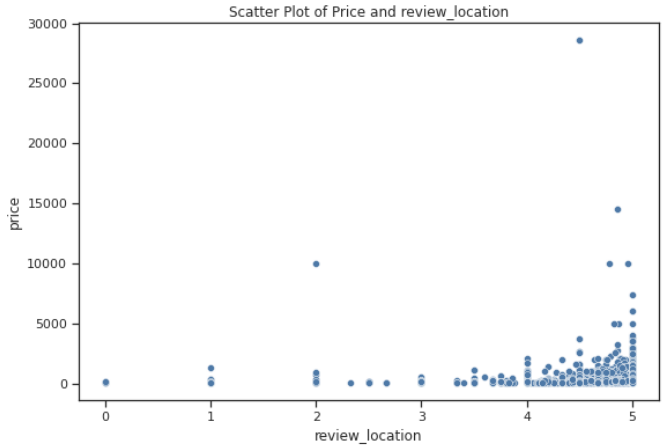
Host about:

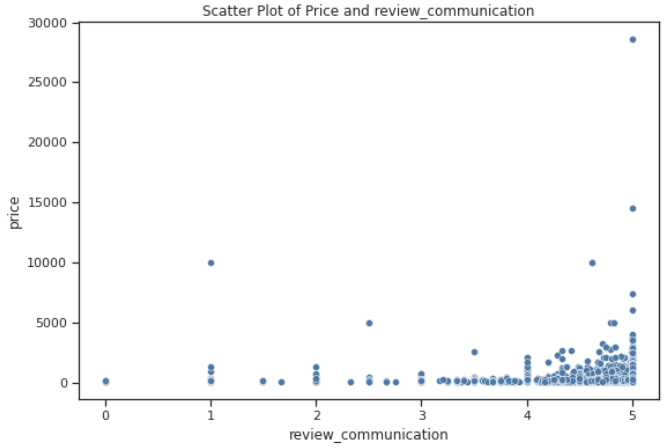


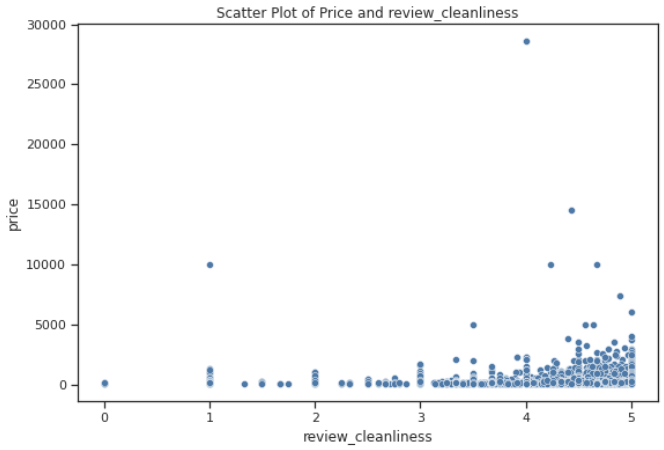
Appendix D: Bivariate Analysis of Continuous and Categorical Variables

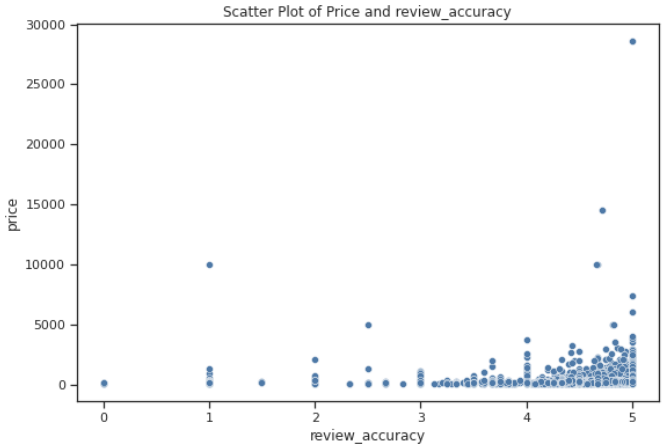
Scatterplots of review features and price



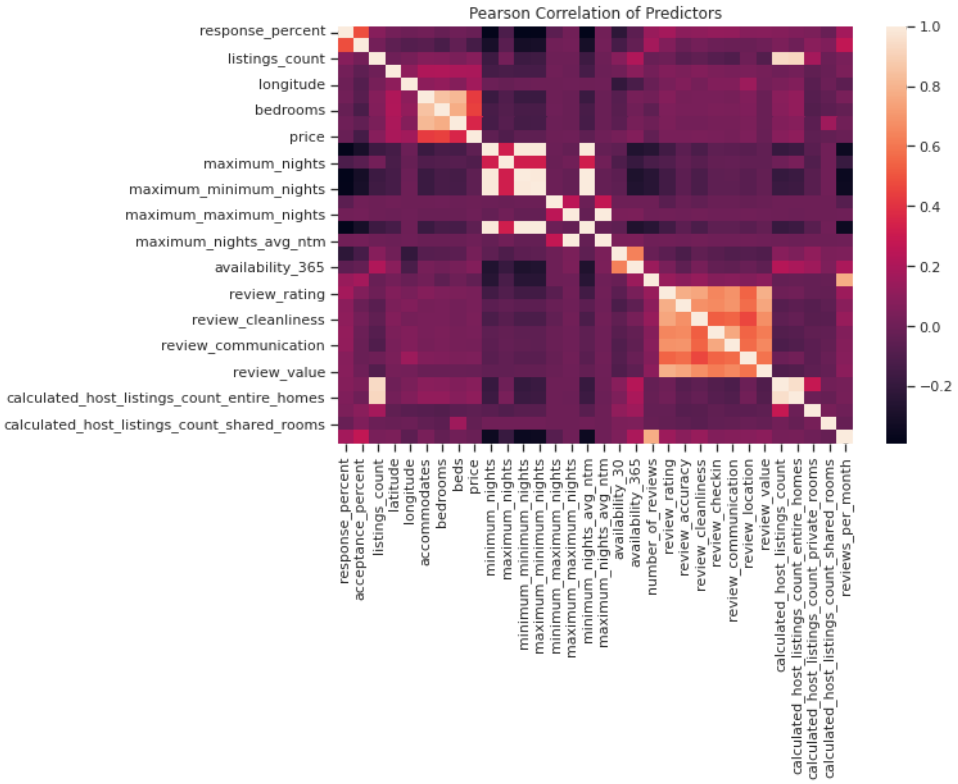




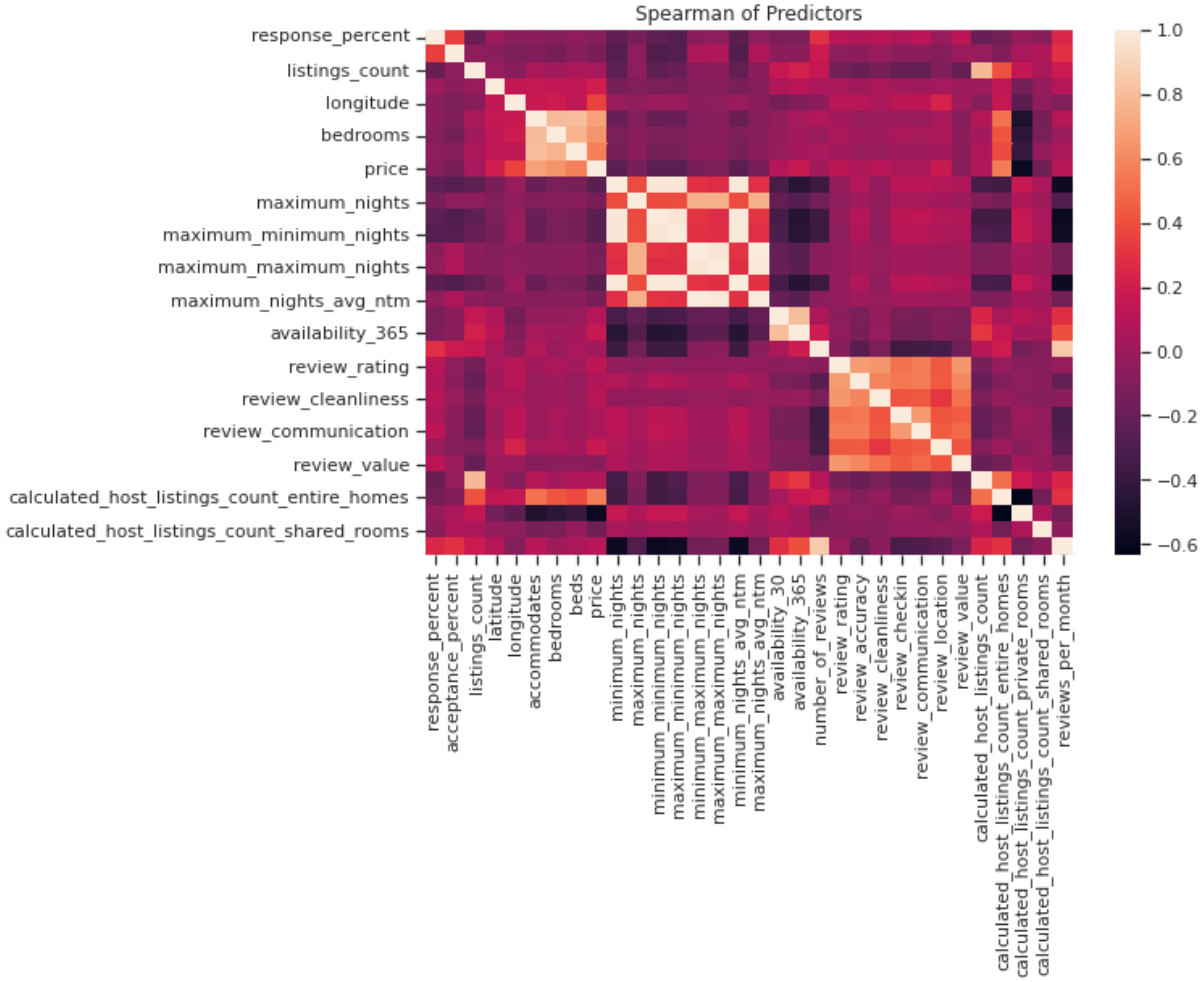




Pearson Correlation of Predictors

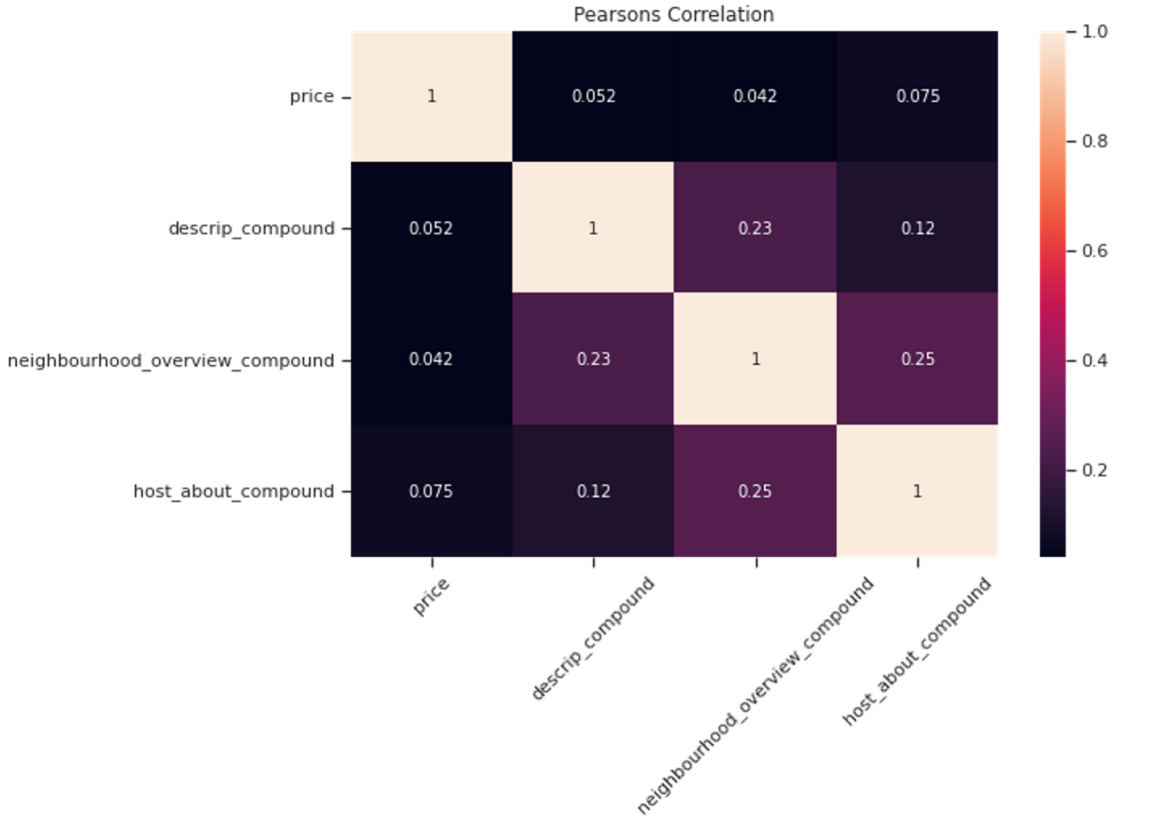


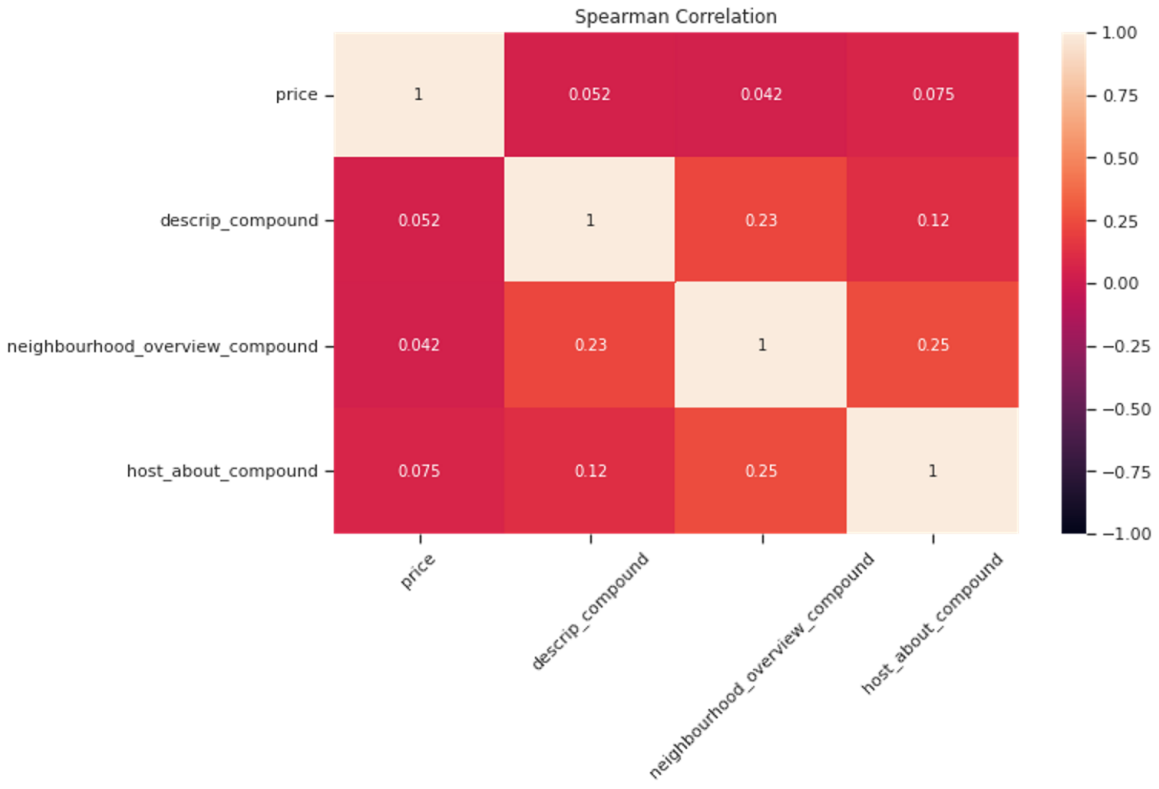
Spearmans Rank of Predictors



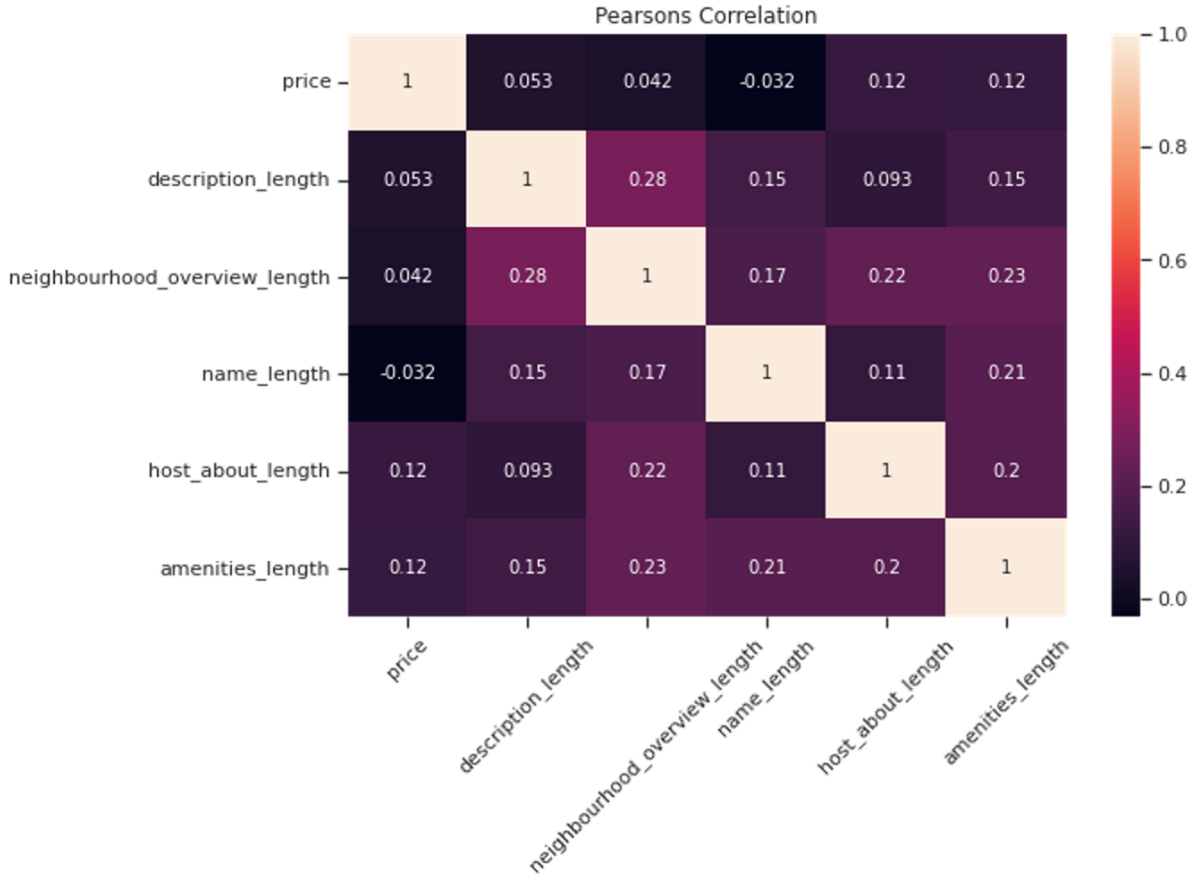
Appendix E: Bivariate Analysis of Text Variables

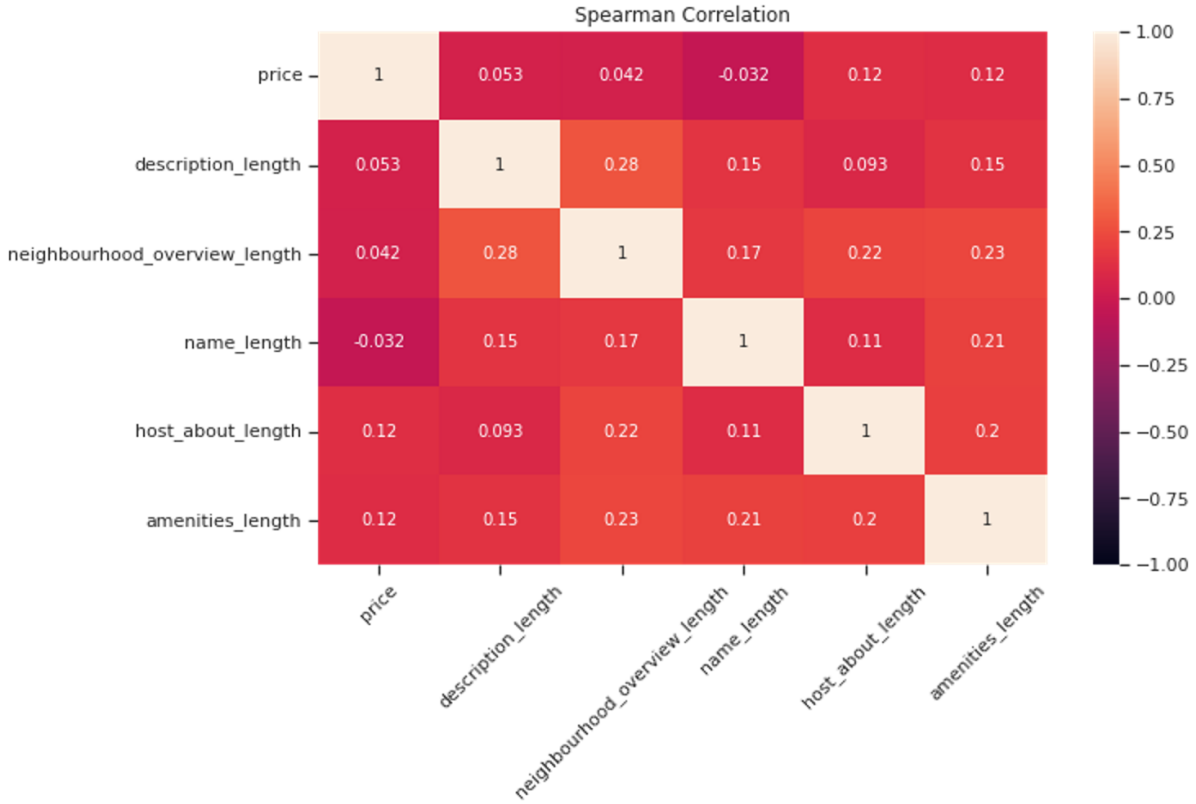
Sentiment Analysis and Price:





Text Length and Price:





Appendix F: Training and Test Results

Train:

Average evaluation score	**RMSE**	**MAE**	**Adj. R2**
Decision Tree	427.403	131.168	-8.984
Random Forest (RT)	417.232	119.509	-8.517
Optimised RT	405.001	119.592	-7.910
Multiple Linear Regression (MLR)	432.443	151.005	-9.223
Ridge Regression	432.442	151.188	-9.223
Regression Stacking	404.262	118.272	-7.886

Test:

Average evaluation score	**RMSE**	**MAE**	**Adj. R2**
Decision Tree	322.658	122.566	-7.271
Random Forest (RT)	306.314	110.996	-6.454
Optimised RT	295.393	110.904	-5.932
Multiple Linear Regression (MLR)	343.917	143.824	-8.397
Ridge Regression	343.902	143.726	-8.396
Regression Stacking	**293.766**	**110.163**	**-5.856**