THE UNIVERSITY OF SYDNEY

QBUS6810 Statistical Learning and Data Mining

(2019S2)

Group Project

Due date:

Friday 8 Nov 2019

Group Members:

480458801

490158104

490202429

490230916

490258138

Contents

1. Introduction

The report aims to establish a predictive model of the nightly rent price for hosts, property managers, and real estate investors based on data collected from Airbnb. A more accurate model can provide better guidance to the host during pricing and revenue forecasting. The features are mainly captured by Exploratory Data Analysis (EDA). OLS regression, Lasso regression, Ridge regression, Elastic Net, and Light GBM were all tried, but the discussions of Ridge regression and Elastic Net are in more detail in this report. The final model uses ridge regression because RMSE and R^2 perform better. The report also includes findings and explanations and attempts to find insights that can help the host to make a decision.

2. Problem Background

In the past decade, a multitude of online platforms of tourism has become popular with the rise of the sharing economy (Cheng 2016). They provide a marketplace that bridges the gap between the host and guest with the support of networking technology, rather than owning real estate. With this platform, the host can use their homes more efficiently and receive revenue from the lodging, while the guest could have more opportunities when looking for short-term or tourism-related rentals. Airbnb, founded in 2008, is one of the online marketplaces that arranges accommodation mainly for homestays and travel experiences. As of October 2019, 2 million people book their accommodation with Airbnb every day, which is a huge success. Airbnb shows 'the global community of hospitality' which changed the way travel accommodation (Roelofsen & Minca, 2018).

Since Airbnb is a service platform, hosts decide their daily prices at their discretion, and without financial background or lack of awareness of market trends, it is difficult to price. To help solve this problem, Airbnb attempts to reduce the error in daily price forecasts by analyzing market trends and industry experience (AirDNA, 2019). Reasonable pricing can generate good revenue for the host, provide a comfortable experience for guests, and bring higher traffic to the Airbnb platform. Under this background, the report tries to establish models to predict the price and find helpful insights for host.

3. Data Description

3.1 Describe the data

The data set is the real data scraped from Airbnb, which includes 83 variables in both train and test data set. The sample size is 9838 in the train set while 22957 in the test set. The 'price', priced in the Australian dollar per night, is the response variable that needs to predict. In the data sets, there are both unstructured data, such as 'summary', 'space' and 'description', and structure data, including the types of numerical and category. The number of variables for each type of data is listed in Table 1. From the description of data, the variable mainly related to hosting, location, room, reviews, and others. The details of each variable are attached to

Appendix 1.

Types	Amount
Numerical	39
Text	21
Categorical	13
Boolean	7
Date	3

Table.1: Summary of data types

3.2 Data Processing

3.2.1 Missing values

The data set misses a significant amount of values. In this processing, some variables have been removed, but other variables have been retained because they still have actual information. The variable, 'host_acceptance_rate' is dropped as it misses all the values. Besides, some repetitive and useless variables with some missing values are also removed. For example, the variables like 'background_offered' become worthless since all values are the same. Some variables are similar meanings with others, for example, 'street' and 'smart_location', and some variables like 'host_id' do not affect the prediction of the price.

Other missing values for these variables are populated in different ways. First, the missing value of numerical variables' square_feet', 'cleaning_fee_perc', and 'reviews_per_month' are filled with mean. Second, the missing value of variables related to the number of rooms, including 'beds', 'bedrooms', and 'bathrooms', are replaced by the median. Third, the missing values of variables related to review scores, like 'review_scores_rating', is replaced by zero as there are no scores. Finally, for text, fill the missing value of 'cancellation_policy' with 'flexible' as there no additional requirement and fill the missing value of 'host_neighbourhood' and 'city' by the information provided in 'host_location' and 'street'.

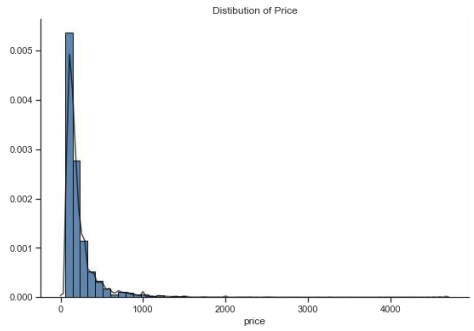
All the missing values are filled after this processing, and the same processing is done in the test data set. The new train set and test set have 58 variables and the same sample size as before.

4. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a method of capturing the main features of the data set after data processing.

4.1 Distribution Plot of price

The distribution of price (the response) is right-biased distribution, according to Figure 1. The use of log transformation reduces the skewness and improves the interpretation of the pattern. From Figure 2, the distribution of price after the log transformation is closer to a normal distribution.



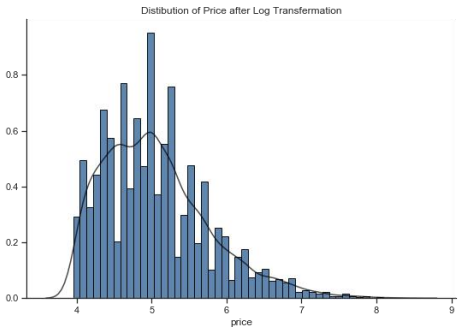
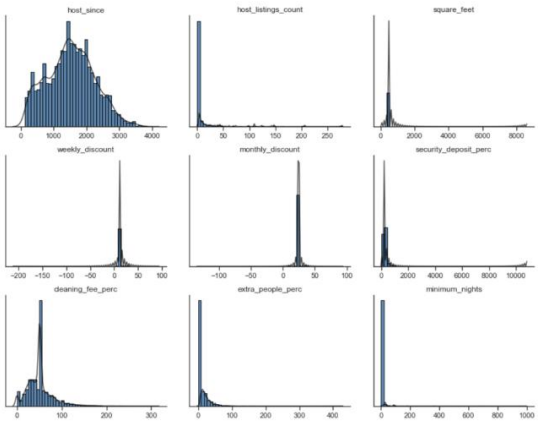


Figure 1: Distribution of Price

Figure 2: Distribution of Price after log transformation

4.2 Distribution Plot

Using the 'plot_dists' function in 'Statlearning' to draw the distribution of continuous variables and the distribution of discrete variables shown in the following Figure 3 and 4. Here only shows the first 9 pictures, the whole result shown in Appendix 2.1 and 2.2. All features shown asymmetrical state, including the most of features contain considerable degree of right skew and several features contain degree of left skew.



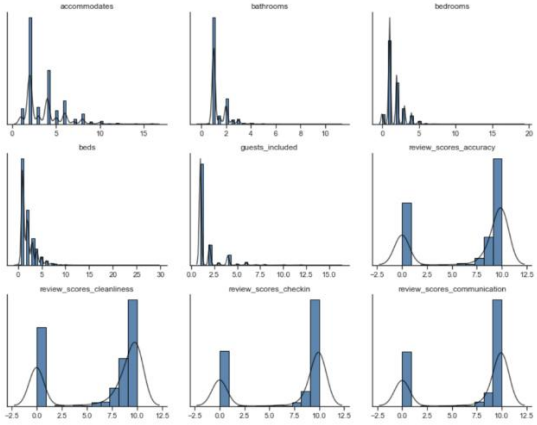


Figure 3: Distribution of continuous variables

Figure 4: Distribution of discrete variables

4.3 Regression Plot

The following Figure 5 and 6 (the whole results see in appendix 2.3 and 2.4) describe the linear regression relationship between continuous variables and discrete variables and price by using the 'plot_regression' function in 'Statlearning'. The performance of regression for discrete variables slightly better than continuous variables. However, the performance of individual variable is not satisfied enough.

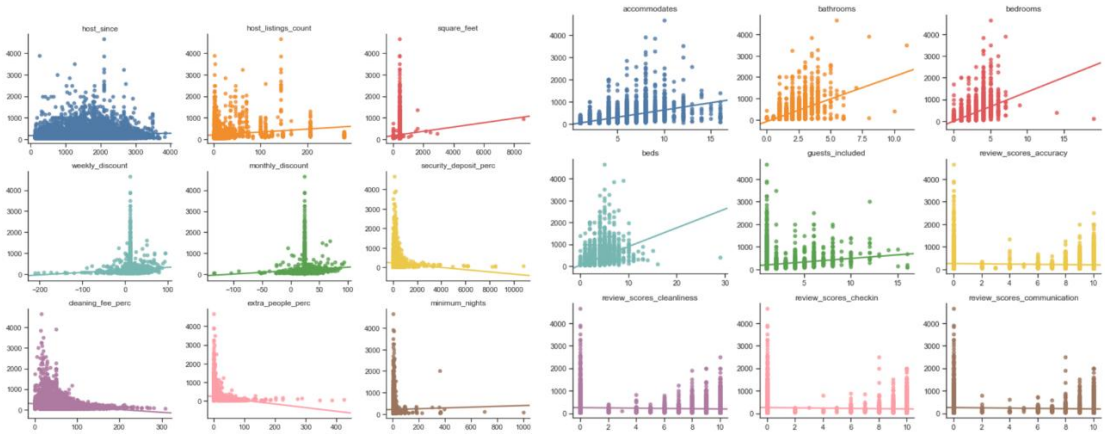


Figure5: Regression plot for continuous variables *Figure6:* Regression plot for discrete variables

4.4 Box Plot

The boxplot (Figure 7) presents the relationship between the dummy variables and the independent variables. It shows that there are some outliers in these data.

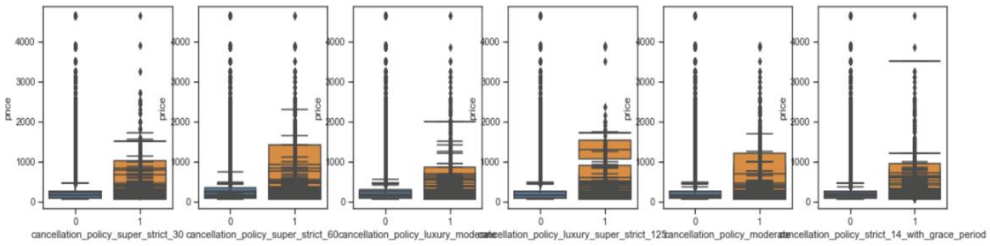


Figure 7: the box plot

4.5 Correlation

Using ‘correlation’ function in ‘Panda’ to calculate the correlation between each variables and ‘price’. The partly result is below:

```
room_type_Private room                          -0.31071
property_type_Apartment                         -0.24430
cleaning_fee_perc                               -0.16614
extra_people_perc                               -0.14228
calculated_host_listings_count_private_rooms    -0.13298
                                                    ...
beds                                             0.53169
accommodates                                     0.57891
bathrooms                                        0.58251
bedrooms                                         0.59917
price                                            1.00000
```

Figure 8: Correlation

Then, in terms of the lager number of dependent variables, select the correlation of variables in training dataset and testing dataset more than 0.1 and less than -0.1, and using ‘heatmap’ function in ‘Seaborn’ to draw Figure 9, which shows the more accurate correlation between left every two variables. The rule is: the whiter the color, the stronger the correlation, or the darker the color purple, the weaker the correlation. For instance, there is a strong correlation between ‘calculated_host_listings_count_ entire_homes’ and ‘calculated_host_listings_count’, which reached the highest 1.0.

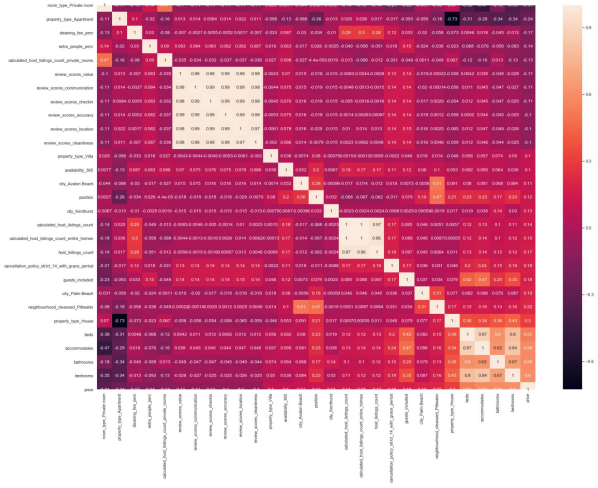


Figure 9: Heatmap

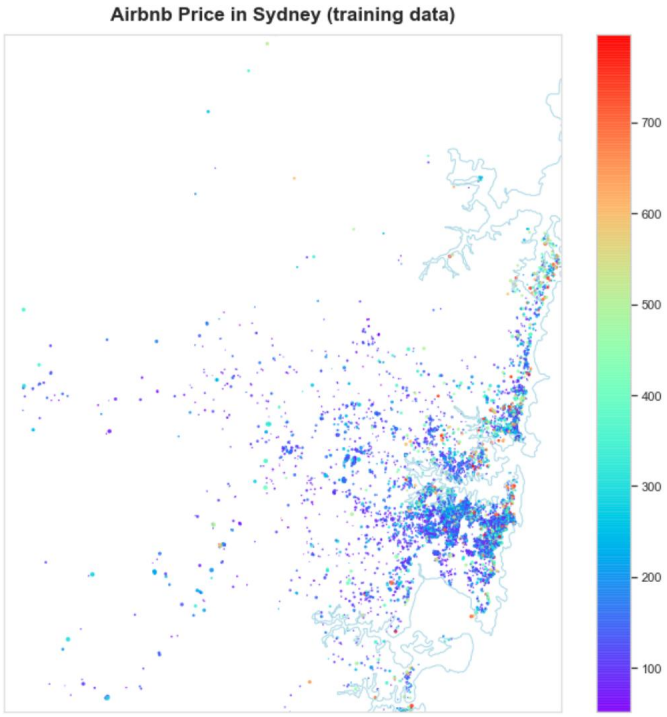


Figure 10 Map of Sydney

4.6 Map

In order to visually analysis the distribution of Airbnb in Sydney, firstly, using the ‘describe’ function in ‘Pandas’ to select the four extreme values of latitude and longitude variables. Secondly, using the ‘map’ function in ‘mpl toolkits.basemap’ to draw a map of the distribution of Airbnb in Sydney shown in Figure 10, which is based on ‘Data Analytics Using Open-Source Tools’ by Jeffrey Strickland. The map shows that the redder the color, the price is higher; the more concentration the point, the numbers of housing is lager; the bigger the point, the accommodates are bigger. For example, the numbers, price and accommodates of the coastal areas are higher than the inland areas. Thus, the latitude and longitude variables may be useful in our forecast model.

4.7 Text processing using WordClouds

A word cloud is a tool-weighted list for visualizing text data, and the frequency is the most basic type used to mine text data (Jin, 2017). In these two images, the font size indicates the number of words that appear in the 'summary, and 'summary' and 'amenities'. The font becomes more significant as the frequency increases.

Figure 11: WordClouds of ‘Summary *Figure 12:* WordClouds of ‘Summary’ and ‘amenities’

With WordClouds, 'apartment' and 'bedroom' most often appear in the 'summary', followed by words' cafe', 'home', and 'new'. These high-frequency words are mainly the descriptions of the house, such as gardens, balconies, balconies, and lofts, as well as their old and new, size, or privacy. The characteristics of the home are always the most concerned, and high-frequency words are also related to neighborhood environments such as cafes and beaches. In WordClouds of 'summary' and 'amenities', 'TV', 'Wifi', 'Kitchen', and 'Washer' are high-frequency words. This result is in line with Airbnb's 'the global community of hospitality'. In addition to facilities such as TV and Wifi, home-like living makes short-term or travel-related rentals more focused on kitchen and washing machine functions.

Therefore, the host is most concerned about the characteristics of the house in the 'summary', and the neighborhood environment is often included. Also, features such as the kitchen and washing machine will be the amenities that lodging usually provides, compared to hotels.

5. Feature Engineering

In terms of some models cannot handle too many numerical variables or categorical variables, both them will be not conducive to analysis, so feature engineering needs to process.

5.1 Dummy variables processing

A dummy variable is a dichotomous quantitative variable created to represent categorical data (Hardy 1993). Moreover, one dummy variable is removed for each category variable in the model to avoid multicollinearity problems. For example, there are seven categories under 'cancellation_policy', but only six dummy variables are reserved. Also, due to the inconsistency of the number of categories in the training set variable and the test set variable, the number of variables in two data sets is different after creating dummy variables. Here, delete the variables that are not the same to keep the consistency.

5.2 Other Data processing

Date-related data, true or false data, and longitude and latitude data are also processed before analysis to ensure the accuracy predictions.

The date related data like 'first_view' are converted into date format and replaced by the gap of days between the date and 01/11/2019 (the set end date). The number of gap days is easier to handle in further analysis. In true or false data, 't' and 'f' in variables such as 'host_is_superhost' and 'require_guest_phone_verification', are replaced with '1' and '0', respectively. Longitude and latitude are replaced by the geodesic distance from the position of the sample to the mean longitude-latitude position of samples. This change expands the difference between the samples.

5.3 Handling Text

In this part, because text-based data need further processed, or will impact on the accuracy of the later model. Express different features by using different feature engineering strategies to contribute to choose significant variables. In addition, text type data and categorical data will be transformed to numeric data in order to fit the later model. Based on this, we use the two methods to process the different text variables.

To ‘amenities’, we use the ‘CountVectorizer’ function in ‘sklearn.feature_extraction.text’ to select special characters, such as ‘{ }’, ‘( )’ in this column. Then, replace these special characters by blank space, which can make this column more cleaner and became a plain text-based data.

As for ‘summary’, we use the frequency-inverse document frequency (TF-IDF) method to deal with the text-based data. TF-IDF usually present the importance and the frequency of a particular word in the document. In generally, the higher frequency a word, which means it is more important, but some words like ‘a’, ‘the’, ‘they’ cannot be regarded as essential, as they do not cover real meaning. Thus, TF-IDF can help to remove the frequency but meaningless words and is used to increase the weight of more useful words for the later model.

Finally, the contain of text in the training dataset and the testing dataset may not be the exact same, so there may be extracted the different types of word. Therefore, we delete the data that they are in the training dataset but not in the testing dataset, and also do the same processing data in the testing dataset.

After the series of data pre-processing, exploratory data analysis and handling text processing, we use ‘describe’ and ‘boxplot’ function in ‘matplotlib.pylot’, which can more precise select the outliers in the training dataset and delete them.

5.4 Log Transformation

Firstly, we attempt to switch the latitude and the longitude variables into a new correct ‘position’ variable in the map which may directly impact on the prices, by calculating the distance between the longitude and latitude and the average point, and the distance between all of these points and a particular point is obtained, in which the performance of this variable is much better than the latitude and the longitude.

Secondly, in order to make the training data exhibit the normally distribution, we choose to transfer the variables that the skewness more than 0 into log variables, because Box strictly requires the data to be positive, or box-cox is not data frame after processing. Since log transformation only works on right skewness, we first remove the data that the skewness less than 0. Therefore, estimating the skewness of the variables by using ‘skew’ function from ‘scipy.stats’. Then, using log transformation to process the variables that the skewness more than 0. After log transformation, the skewness of all variables less than 0. It will be more useful for later modeling.

5.5 Scaling

We can process the scaling in the dataset, and some features are different from varying sizes, because the data of different features themselves may differ greatly, the scaling process can make the values more uniform to convenient build the later model. If we process scaling first than log transformation, the values of variables will become bigger and the scaling process is meaningless. Thus, we use the ‘MinMaxScaler’ and ‘StandardScaler’ function in

'sklearning.preprocessing' to deal with the training dataset. In order to make a comparison with the later model, we tried these two possibilities, MinMax processing and Standardisation, and then observe which one can make the performance of model will be better.

6. Methodology

Overall, there are five models applied and two models will be focused to discuss in the report, which are Ridge Regression and Elastic Net.

6.1 Ordinary Least-Squares Regression

Ordinary least-squares (OLS) regression is one of the most adopted statistical techniques in econometrics. It is widely used to find appropriate parameters of a linear regression by minimizing the sum of the squared errors (the difference between observed values and predicted values) (Hutcheson, 1999). In the case of a model with p explanatory variables, the model is defined as

$$Y = \beta_0 + \sum_{j=1..p} \beta_j X_j + \varepsilon$$

The coefficient β which fits the model best is defined as

$$\widehat{\beta}_{ols} = argmin \sum_{j=1}^{p} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2$$

6.1.1 OLS Result Analysis

For the given dataset, it is obvious that the dependent variable price is closely related to the other features. Thus, OLS regression is chosen to be first regression model. The OLS regression is simply performed by adopting linear regression library. The obtained RMSE output is 0.3329, which is relatively acceptable. However, the tested RMSE output derived from cross validation shows a negative value of -3.3001. This strongly implies that an existing issue of overfitting. This a common issue while performing multiple linear regression because the model tends to be overfitted when more features are added to the model. This leads to a bad performance of using test dataset to make predictions. Therefore, OLS model is not selected as final model.

Model	RMSE	R squared	MAE	Test RMSE
OLS Regression	0.3329	0.7467	0.2524	3.3001
Lasso Regression	0.3467	0.7252	0.2642	0.3892
Ridge Regression	0.3414	0.7336	0.2599	0.3889
Elastic Net	0.3470	0.7247	0.2643	0.3883
LightGBM	0.2299	0.8791	0.1711	0.5455

Table 2: Results of Models

6.1.2 OLS Model Limitations

One of the essential assumptions of applying OLS is linearity of independent variables. In the real-world situation, nevertheless, it is much more complicated and very unlikely to have simple linear relationship. This means even with an infinite number of training points, the model will often fail to do a good jot to make a forecast. In addition, the variable selection for OLS model is time consuming when the number of independent variables is large. So it is likely unavoidable to cause overfitting or underfitting problem (Backward, 2009).

6.2 Lasso Regression

Lasso Regression (Least Absolute Shrinkage and Selection Operator) obtains a more refined model by constructing a penalty function, so that it compresses some coefficients and sets some coefficients to zero. Therefore, the advantage of subset shrinkage is preserved, which is a method of processing biased estimates with complex collinear data. Lasso Regression is a little different from Ridge regression in that it uses absolute values instead of squares (Yang & Wen, 2018).

6.2.1 Lasso Result Analysis

The RMSE and R squared result from lasso regression is 0.3467 and 0.7252 respectively. The model performance is lower than ridge regression. As lasso and ridge regression operates similarly but with different penalty term. Based on RMSE result, ridge regression is more desirable.

6.3 Ridge Regression

Ridge regression is an advanced technique for dealing with multi-collinearity issue existing in multiple independent variable regression model. Multi-collinearity is the existence of near-linear relationships among the independent variables. The occurrence of multi-collinearity results in large variance in least square estimations, which indicates a deviation from true values. Hence, in order to control the variance, it is effective to impose a constraint on regression coefficients as not to get un boundedly large (Ehsanes, Golam, & Mohammad, 2019). The loss function of ridge regression model is defined as

$$J_w = \min \{||Xw - y||^2 + \alpha||w||^2\}$$

The constant alpha represents a coefficient of penalty term. The influence of penalty term becomes more significant as the increase of alpha. Oppositely, the regression becomes linear square regress as alpha approaching zero.

6.3.1 Ridge Result Analysis

In prior to build ridge regression model, it is critical to decide the value of alpha. From Figure 13, it can be observed that the coefficients of independent variable start to shrink while increasing ridge constant which is defined as alpha. More importantly, it is clear that the coefficients are stabilized when ridge constant reaches approximately 0.42. The value of ridge constant cannot be too large, otherwise, the penalty term dominates the output and the coefficient will finally approach to zero (Ehsanes, Golam, & Mohammad, 2019). Thus, ridge constant selection is the key point of accurately training the ridge regression model. This is achieved in programming by creating a list to store all possible attempts with a logged scale as shown in Figure 14.

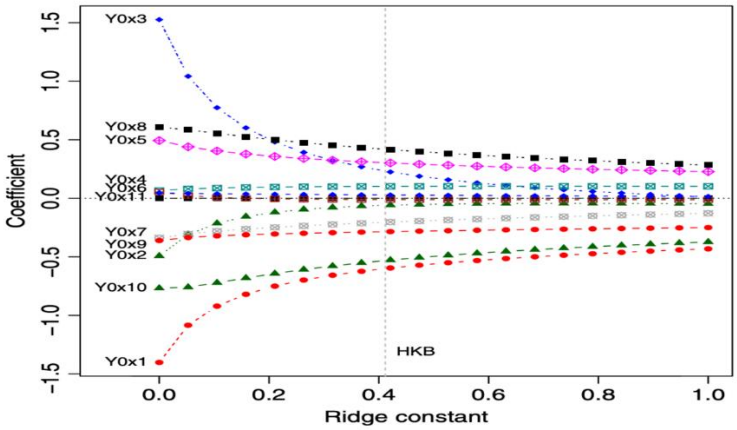


Figure 13: Ridge constant vs Variable coefficient

```
alphas = list(np.logspace(-15, 15, 151, base=2))
ridge = RidgeCV(alphas=alphas, cv=5)
ridge.fit(X_train, y_train)
```

Figure 14: Setting Information 1

According to Table 2, the obtained RMSE and cross-validated RMSE is 0.3414 and 0.3889 respectively. The tested error demonstrates a slightly better performance than lasso regression. This is also proved by a greater R squared value of 0.7336. The main purpose of this project is to help hosts to better understand the pricing and make price prediction for the owners and investors. So the accuracy of price forecast is extremely important, which is determined by correctly trained regression model with proper variable coefficients. The implementation of ridge regression effectively solved the issues of OLS regression like overfitting and collinearity. Also, by contrast with lasso regression, a better performance of ridge regression is verified. Thus, it is selected as one of the final models.

6.3.2 Ridge Model Limitations

Ridge regression trades variance for bias, meaning that the outputs from ridge regression is

not unbiased. This would arise a question that how much bias is acceptable in order to decrease variance. The optimal solution for this question is using cross-validated sum of squared residual to determine the appropriate ridge constant. Normally, the obtained R squared value of ridge regression is slightly lower than R squared value of OLS regression (Ehsanes, Golam, & Mohammad, 2019).

6.4 Elastic Net

The emergence of elastic net is firstly due to the critiques on lasso regression, whose variable selection can be too dependent on data. Elastic networks as a compromise between ridge and lasso regression. The goal of minimizing the loss function is defined as

$$L_{enet}(\widehat{\beta}) = \frac{\sum_{i-1}^{n}(y_i - x_i\widehat{\beta})^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\right)\sum_{j=1}^{m}\widehat{\beta}^2 + \alpha\sum_{j=1}^{m}|\widehat{\beta}_{j|}$$

where α (alpha) is the mixing parameter between ridge and lasso

Theoretically, elastic net regression has significantly positive effect on dealing with severe multi-collinearity. The performance of elastic net regression is very close to lasso regression when alpha equal to one. To the contrary, the performance is close to ridge regression when alpha equal to zero (Alhamzawi & Mohammad, 2017.).

6.4.1 Elastic Net Result Analysis

The L1ratio (alpha) is set to 11 attempts in a range of 0 to 1 as shown in Figure 15. The cross-validation size is set to 5, which is exact same when doing ridge regression. The obtained optimal alpha is 0.8. The RMSE and cross-validated RMSE is 0.3707 and 0.3883 respectively. The result is very close to the result from ridge regression. The lower cross-validated RMSE indicates a less overfitted situation. The RMSE and R squared result reflects a more biased prediction. The obtained result from elastic net is the best result compared with other four regression model. It demonstrates the significance of overcoming overfitting and collinearity issue. From the perspective of real-world situation, a more unbiased forecast would be more useful for advising to the host about the pricing.

```
enet = ElasticNetCV(l1_ratio=[0.01,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,0.99], cv=5)
enet.fit(X_train, y_train)
```

Figure 15: Setting Information 2

6.4.2 Elastic Net Model Limitations

One major limitation of elastic net regression is the computational cost. It is required to cross-validate the relative weight of L1 vs L2 penalty, which greatly increases computational cost.

6.5 Light GBM

Light GBM is a gradient boosting framework using tree based leaning algorithm. Figure 16 shows the main difference between light GBM and the other algorithm. Light GBM chooses the leaf with maximum delta loss to grow. In this way, it reduces more loss when growing same leaf. The main characteristics of light GBM is high speed and ability of handling large size data with lower memory to run (Ke, Meng, Finley, Wang, & Chen, 2017).

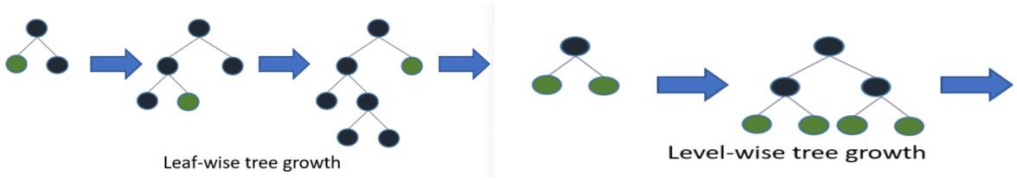


Figure 16: Tress growth

6.5.1 LightBGM Result Analysis

The obtained RMSE and R squared is 0.2299 and 0.87919 respectively. The R squared value is the highest among the five selected models. But a cross-validated RMSE indicates a potential overfitting issue.

7. Predictive Model Validation

General performance of all the 5 models are presented in the following table. According to the result for the test set on Kaggle, the optimal model should be the Elastic Net model, which has a l1 regularization ratio equals to 0.8. In general, RMSE of test sets of Lasso model, Ridge model and Elastic net model are pretty close. But test sets' RMSE of the Light Gradient Boosting model (LightGBM) and OLS Regression model are really high, which indicates these two models are less accurate than the other three models.

However, based on root-mean-square error (RMSE) of training sets, the best model should be the Light Gradient Boosting model (LightGBM). It has great performance on the training data but a poor explanation for the test figures. This problem might be caused by the overfit problem, which indicates that this model has a low variance but a high bias (Table 2).

8. Data Mining and Conclusion

8.1 Insights

Eight insights are listed based on the interpretation of Elastic Net model.

1. The higher the "weekly_discount" is, the higher the "price".

2. With increase of the number of “accommodates”, “bathrooms” and “bedrooms”, the “price” will grow as well.

3. “position” has a negative relationship with “price”. “price” will drop when the distances between central location and other accommodations increase.

4. If “property_type” of accommodations is boat, “price” will increase.

5. If “city” of accommodations is Riverview, “price” will increase.

6. When “calculated_host_listings_count” and “cleaning_fee_perc” increases, “price” will decrease.

7. In text fields “summary”, when it contains words like “view”, “beach”, “living”, “pool”, “sydney”, “new”, “price” tends to increase.

8. Words in column “summary” such as “we”, “room”, “quiet”, “bus”, “place” will bring down “price”.

8.2 Detailed explanations

Figure 17 from python illustrates the ranking about the absolute value for the coefficients of features from Elastic Net model. As can be seen in the output the most significant feature of this model is “weekly_discount” which has a coefficient equal to 0.90349. Another 4 features which are most closely related to the model are “accommodates”, “bathrooms”, “bedrooms” and “position”.

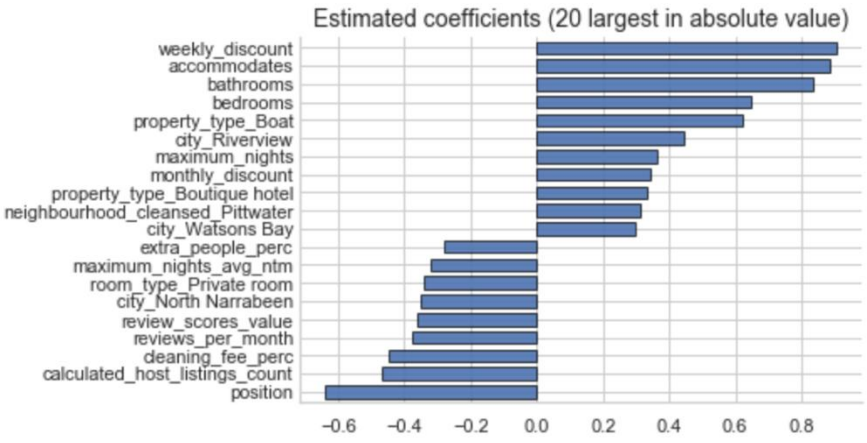


Figure 17: Estimated Coefficients

Since the dependent variable of this model is the log transformation of “Price”, the higher the coefficient, the higher the price. Among the top 5 features which have the highest absolute values of coefficients, position is a little different from the others. Since “weekly_discount”, “accommodates”, “bathrooms” and “bedrooms” have positive estimated coefficients, the

higher these features are, the higher the price is. Referring to “position”, it has the lowest coefficients among all the features, which shows that “price” will decrease with the increase of “position”. As “position” is a feature combining “longitude” and “latitude”. It measures the distance between the point which has an average longitude and latitude and other points. As a result, points which are far away from the average point will lead to a lower price. Besides, when “property_type” is boat, “price” will increase in a large degree as well. Another two importance features which will bring down “price” are “calculated_host_listings_count” and “cleaning_fee_perc”.

In addition, *Table 3* shows the ranking of words in text field features which come from column “summary”. From this table, it is obvious that word “views” has the most significant influence on “price”, following by “we”, “room”, “quiet”. However, the effects of word “we”, “room”, “quiet”, “bus”, “place” are negative while other words shown in the table have positive effects on “price”.

Ranking	**Word in "summary"**	**estimated coefficients**
1	views	0.226036
2	we	-0.122768
3	room	-0.115218
4	quiet	-0.113814
5	beach	0.111132
6	living	0.097159
7	bus	-0.090442
8	pool	0.087865
9	sydney	0.085906
10	place	-0.081962
11	new	0.079865

Table 3: Estimated Coefficients

Reference List

AirDNA, 2019. *AirDNA | Short-Term Rental Data & Analytics | Airbnb & Vrbo*. [online] Available at: https://www.airdna.co/ [Accessed 7 Nov. 2019].

Alhamzawi, R., & Mohammad, H. M. (2017.). The Bayesian elastic net regression. *Communications in Statistics - Simulation and Computation.*

Backward, C. (2009). Ordinary Least Squares Linear Regression: Flaws, Problems and Pitfalls.

Cheng, M. (2016). Sharing economy: A review and agenda for future research. *International Journal of Hospitality Management*, 57, pp.60-70.

Ehsanes, S. A., Golam, K., & Mohammad, A. (2019). *Theory of ridge regression estimators with applications*. Hoboken, NJ: Wiley Blackwell.

Hardy, Melissa A. (1993) Regression with dummy variables. Newbury Park, [Calif.] ;: SAGE.

Hutcheson, G. (1999). *Ordinary Least-Squares Regression*. SAGE Publications, Ltd.

Jin, Y. (2017). Development of Word Cloud Generator Software Based on Python. *Procedia Engineering*, Volume 174, 2017, pp.Pages 788-792.

Ke, G., Meng, Q., Finley, T., Wang, T., & Chen, W. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *31st Conference on Neural Information Processing Systems*. Long Beach.

Roelofsen, M. and Minca, C, 2018. The Superhost. Biopolitics, home and community in the Airbnb dream-world of global hospitality. *Geoforum*, 91, pp.170-181.

Yang, X., & Wen, W. (2018). Ridge and Lasso Regression Models for Cross-Version Defect Prediction. *IEEE Transactions on Reliability*, 67(3), 885-896.

Appendix

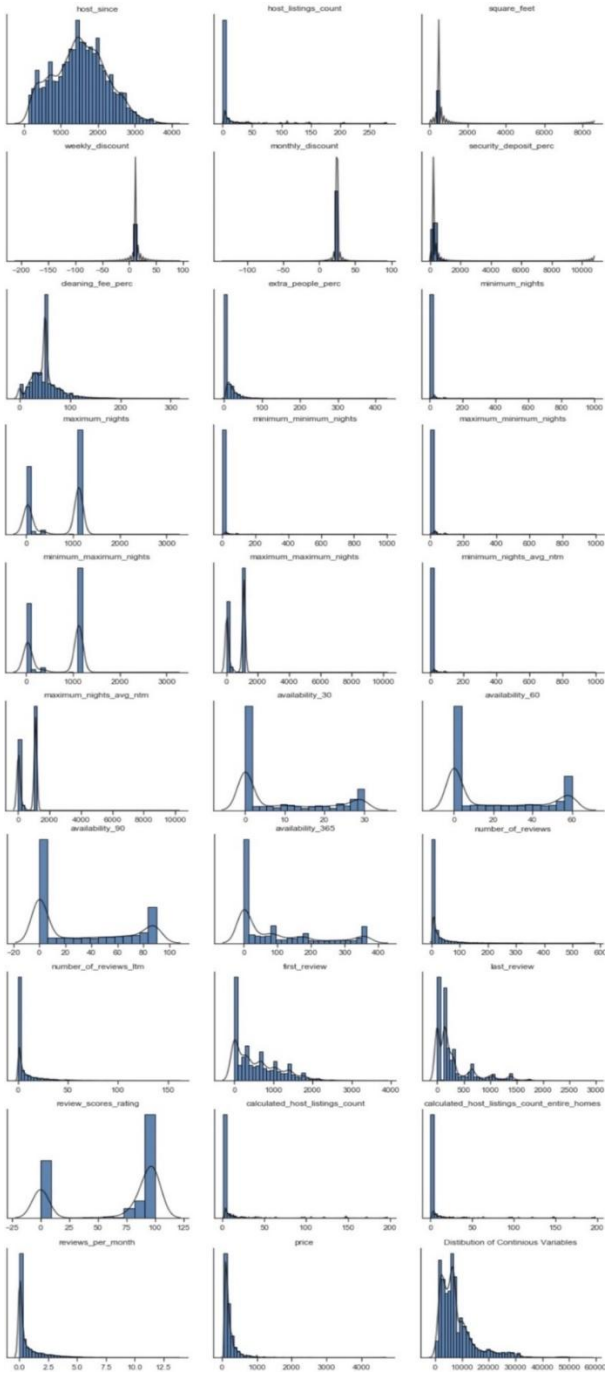
Appendix 1

Index	Predictors	Types	The number of missing value
0	name	Text	0
1	summary	Text	0
2	space	Text	2654
3	description	Text	0
4	experiences_offered	Text	0
5	neighborhood_overview	Text	3438
6	notes	Text	5538
7	transit	Text	3369
8	access	Text	3892
9	interaction	Text	3878
10	house_rules	Text	3987
11	host_id	Numerical	0
12	host_since	Date	0
13	host_location	Text	8
14	host_about	Text	4363
15	host_response_time	Categorical	4566
16	host_response_rate	Numerical	4566
17	host_acceptance_rate	Numerical	9838
18	host_is_superhost	Boolean	0
19	host_neighbourhood	Text	2915
20	host_listings_count	Numerical	0
21	host_total_listings_count	Numerical	0
22	host_verifications	Text	0
23	host_identity_verified	Boolean	0
24	street	Text	0
25	neighbourhood	Text	1151
26	neighbourhood_cleansed	Text	0
27	city	Text	6
28	zipcode	Numerical	31
29	smart_location	Text	0
30	latitude	Numerical	0
31	longitude	Numerical	0
32	is_location_exact	Categorical	0
33	property_type	Categorical	0
34	room_type	Categorical	0
35	accommodates	Numerical	0
36	bathrooms	Numerical	5
37	bedrooms	Numerical	3

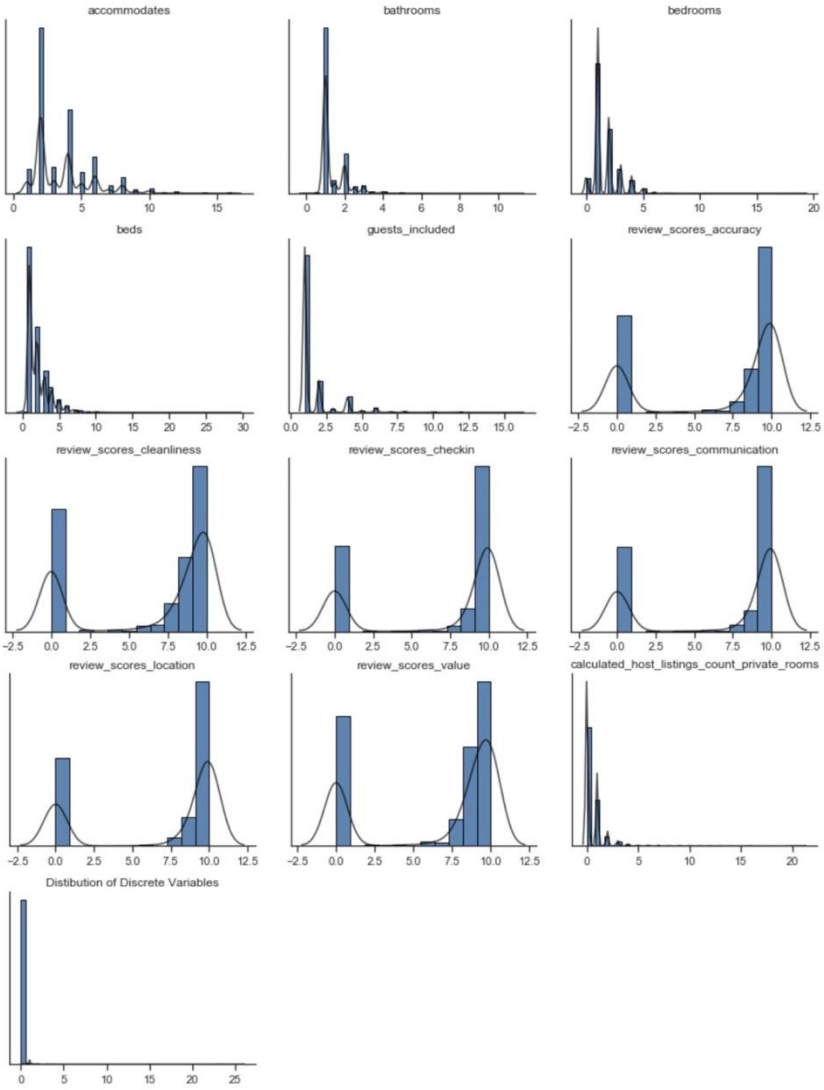
38	beds	Numerical	14
39	bed_type	Categorical	0
40	amenities	Text	0
41	square_feet	Numerical	9779
42	weekly_discount	Numerical	8940
43	monthly_discount	Numerical	9299
44	security_deposit_perc	Numerical	3303
45	cleaning_fee_perc	Numerical	2445
46	guests_included	Numerical	0
47	extra_people_perc	Numerical	0
48	minimum_nights	Numerical	0
49	maximum_nights	Numerical	0
50	minimum_minimum_nights	Numerical	0
51	maximum_minimum_nights	Numerical	0
52	minimum_maximum_nights	Numerical	0
53	maximum_maximum_nights	Numerical	0
54	minimum_nights_avg_ntm	Numerical	0
55	maximum_nights_avg_ntm	Numerical	0
56	availability_30	Numerical	0
57	availability_60	Numerical	0
58	availability_90	Numerical	0
59	availability_365	Numerical	0
60	number_of_reviews	Numerical	0
61	number_of_reviews_ltm	Numerical	0
62	first_review	Date	2698
63	last_review	Date	2698
64	review_scores_rating	Categorical	2950
65	review_scores_accuracy	Categorical	2957
66	review_scores_cleanliness	Categorical	2954
67	review_scores_checkin	Categorical	2959
68	review_scores_communication	Categorical	2956
69	review_scores_location	Categorical	2960
70	review_scores_value	Categorical	2962
71	requires_license	Boolean	0
72	instant_bookable	Boolean	0
73	is_business_travel_ready	Boolean	0
74	cancellation_policy	Categorical	1
75	require_guest_profile_picture	Boolean	0
76	require_guest_phone_verification	Boolean	0
77	calculated_host_listings_count	Numerical	0
78	calculated_host_listings_count_entire_homes	Numerical	0
79	calculated_host_listings_count_private_rooms	Numerical	0

80	calculated_host_listings_count_shared_rooms	Numerical	0
81	reviews_per_month	Numerical	2698
82	price	Numerical	0

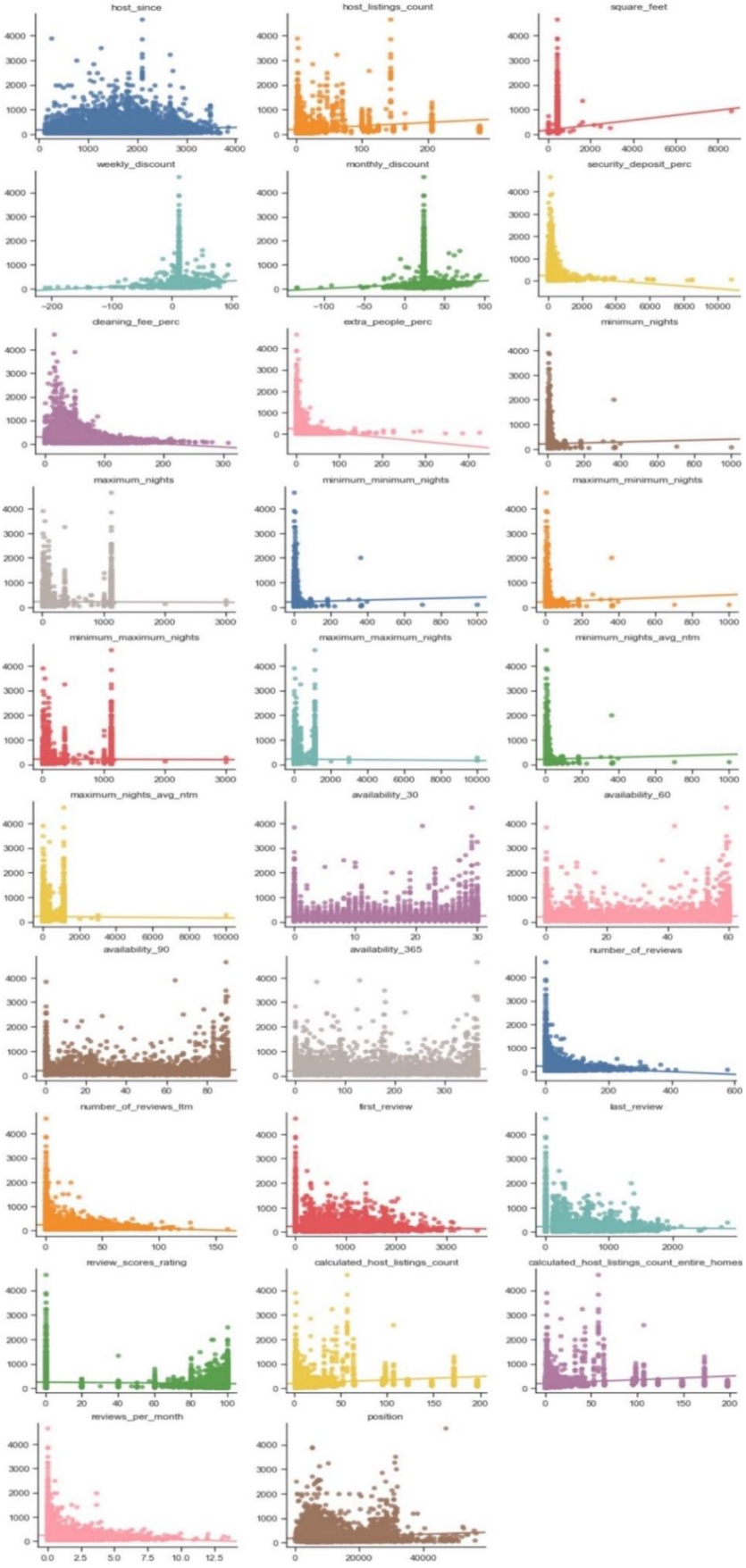
Appendix 2.1



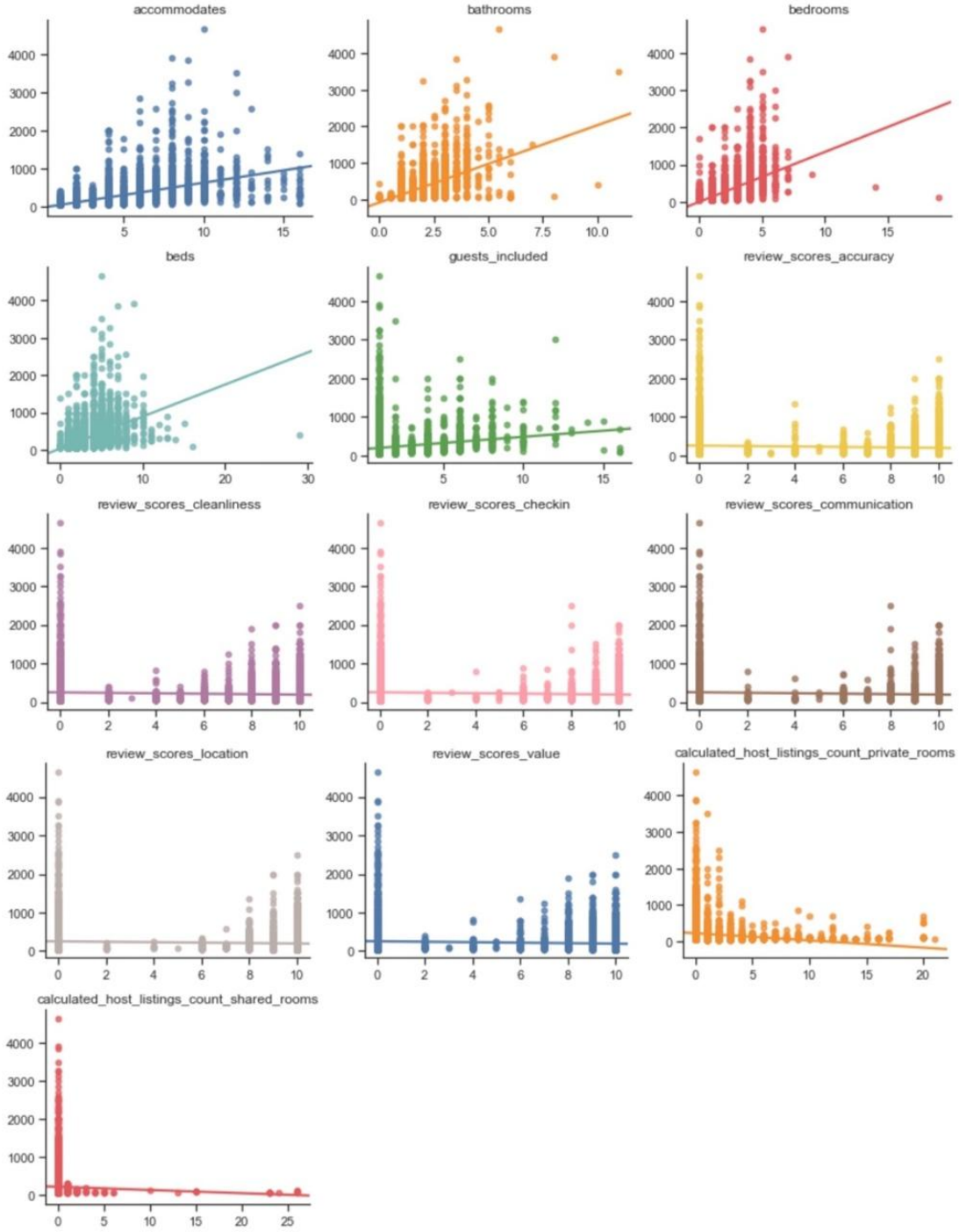
Appendix 2.2


accommodates
bathrooms
bedrooms
beds
guests_included
review_scores_accuracy
review_scores_cleanliness
review_scores_checkin
review_scores_communication
review_scores_location
review_scores_value
calculated_host_listings_count_private_rooms
Distibution of Discrete Variables

Appendix 2.3


host_since
host_listings_count
square_feet
weekly_discount
monthly_discount
security_deposit_perc
cleaning_fee_perc
extra_people_perc
minimum_nights
maximum_nights
minimum_minimum_nights
maximum_minimum_nights
minimum_maximum_nights
maximum_maximum_nights
minimum_nights_avg_ntm
maximum_nights_avg_ntm
availability_30
availability_60
availability_90
availability_365
number_of_reviews
number_of_reviews_ltm
first_review
last_review
review_scores_rating
calculated_host_listings_count
calculated_host_listings_count_entire_homes
reviews_per_month
position

Appendix 2.4


accommodates
bathrooms
bedrooms
beds
guests_included
review_scores_accuracy
review_scores_cleanliness
review_scores_checkin
review_scores_communication
review_scores_location
review_scores_value
calculated_host_listings_count_private_rooms
calculated_host_listings_count_shared_rooms