# AIRBNB PRICE PREDICTI0N

# Table of Contents


**Introduction** **2**

**Business Implications** **2**

**Initial Data Cleaning** **3**

**Exploratory Data Analysis** **6**

*Amenities* *6*

*Correlations* *8*

**Prediction – Random Forest Method** **8**

**Prediction – Linear Regression** **10**

**Classification – Logistic Regression** **11**

**Conclusions and Key Takeaways** **12**

*Model Recommendation* *12*

**Challenges and Drawbacks** **13**

**Appendix A - Correlation between numeric variables** **14**

**Appendix B - Correlation between categorical variables** **15**

**Appendix C - Correlation between all variables** **16**

**Appendix D - Order of Impurity in Random Forest Model** **16**

**Appendix E - Pairwise plots of final variables** **17**

**Appendix F - Correlations between final variables** **18**

**Appendix G – Correlation coefficients** **18**

# Introduction

Airbnb is an online service that allows people to offer rooms for booking and also allows consumers to book these rooms and stay for a temporary period. The hosts have the freedom to set the price as they wish, however, these prices need to be selected based on what they have to offer the consumers. The goal of this analysis is to determine the factors that are considered when setting a price for an Airbnb, and to find out whether a house or room falls in the upper price range or lower price range, depending on the features offered by the host.

# Business Implications

Hosts who list their properties on Airbnb have the flexibility to set the market price (cost per night) for their property; However, their lack of information and understanding of the pricing dynamics of this market have limited their ability to maximize their Return on Investment (ROI). It is difficult to determine the ideal price, because if the price is too high, no one will book and if the price is too low, the host will lose money.

On average, people in the US spend around $130 on hotels per night. On analysis of the Airbnb data, it can be seen that just under half the listings in the Los Angeles area are above this price. Our aim is to determine exactly what features are prompting people to increase the price beyond what people are willing to pay on average. With this information, people who would like to list their properties will be able to determine the most optimal price on their property in order to attract more people as well as generate a profit.

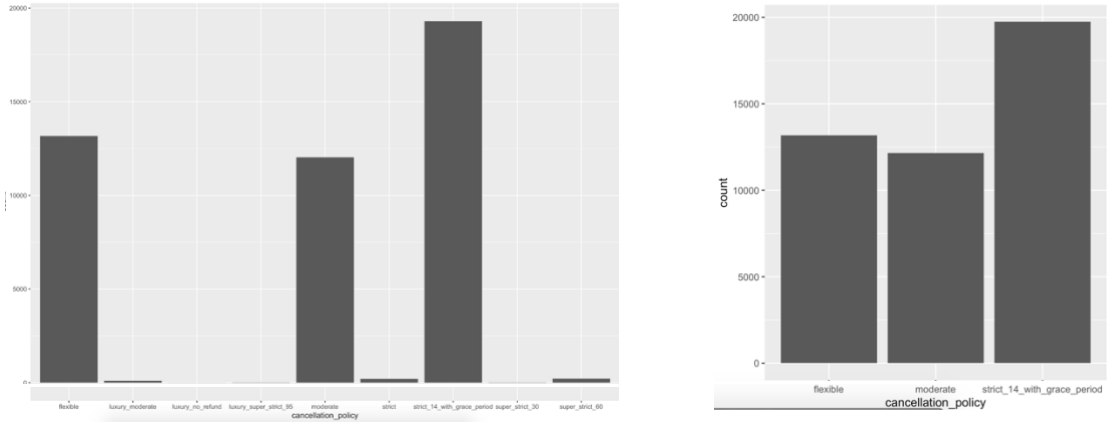
The dataset contains data from the Airbnb website for different listings for the Los Angeles Area. It has a combination of numeric and categorical data, and the dependent variable for this analysis will be “Price” which is a numeric variable.

We would like to achieve the following:

1. Run correlations on the different features of an Airbnb listing and determine how they affect the price.
2. Conduct Exploratory Analysis to determine the most popular features in an Airbnb listing.
3. Run Logistic Regression, Linear Regression and Random Forest predictive models to predict the most optimal base price brackets for and Airbnb host.

# Initial Data Cleaning

For our analysis, the Airbnb data from Los Angeles was used. The property_type variable contains too many different options, and most of them have low frequency occurrence as shown in Graph 1. “Apartment” and “House” are the two dominating classes for the feature, hence all the other classes were classified into these two, and an additional class “Other” was created for those that belonged in neither category (shown in the right graph 2). The number of cancellation policies were also contained within 3 categories as the other categories had a very low frequency of entries.

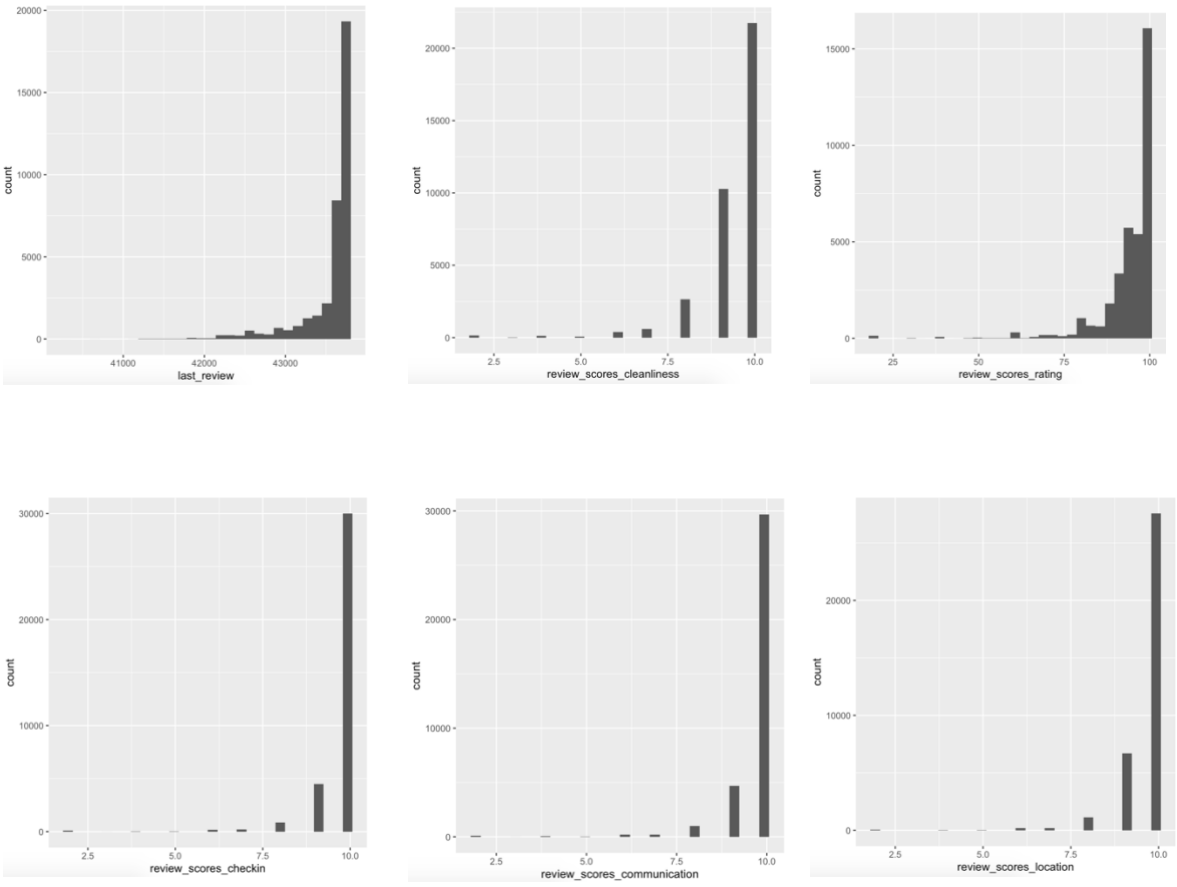



Numeric null values were either deleted, or populated with the average value of the respective columns. Rows that contained categorical Null values were removed from the dataset before analysis.

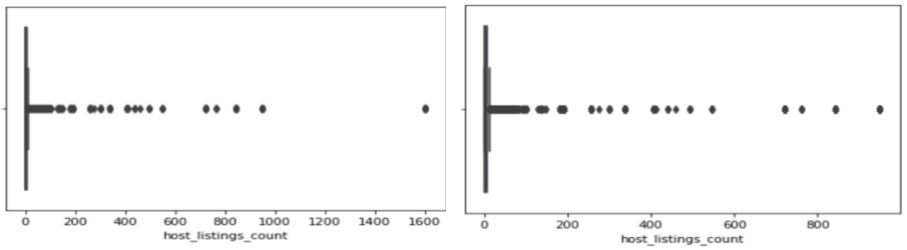
Correlation between the different variables in the dataset is determined for the following reasons:

i. To determine the correlation between the independent variables, and the dependent variable (Price).
ii. To determine any multi-collinearity that exists between the variables, and make note of it during the analysis.
iii. To remove any columns that would not add value to the analysis or predictions.

The results of the correlations between the different variables can be found in Appendix A, B and C. Based on the analysis of the results, the columns maximum_nights, minimum_nights, host_listings_count were all dropped, as they were the least correlated to the price
From the data, the reviews were all left-skewed as shown in Graph 5. It is likely that those that gave a good review in one category ended up giving similar reviews in all categories. This also meant there was a high amount of multi-collinearity between all these variables. The columns reviews_per_month, review_scores_value, review_scores_communication, review_scores_checkin, review_scores_rating, review_scores_cleanliness, review_scores_location were all dropped for the final analysis.




The next step was to determine any major outliers in the data. The variable host_listings_count had an outlying value of 1603. All these rows were removed from the final dataset.




The amenities were also converted to dummy variables and each amenity added as an individual column.

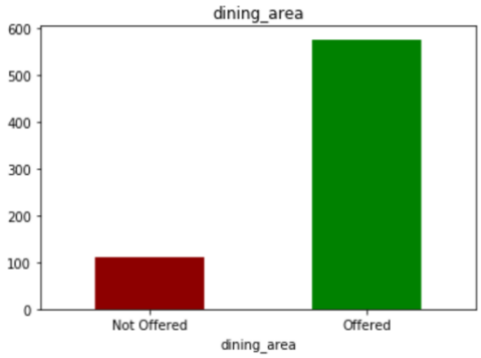
The features used for the final analysis are:

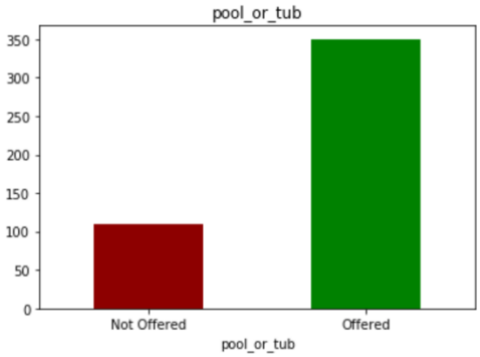
| Variable | Description |
|---|---|
| Id | Listing ID |
| property_type | The type of the property listed – House, Apartment or Other. |
| room_type | Room type being listed – whether it was the entire property, a private room or a shared room. |
| accommodates | Number of people the room or property being listed accommodates |
| bathrooms | The number of bathrooms inside room or property. |
| bedrooms | The number of bedrooms inside room or property. |
| beds | The number of beds inside room or property. |
| amenities | A set containing the different amenities offered. |
| price | Price of the property for one night (dependent variable) |
| security_deposit | Security deposit amount in USD. |
| cleaning_fee | Cleaning Fee in USD |
| guests_included | Number guests included in the booking fee. |
| extra_people | Price per additional guest after above number is satisfied. |
| availability_365 | The number of days the listing is available for booking in the next 365 days. |
| number_of_reviews | The number of reviews for the property. |

# Exploratory Data Analysis

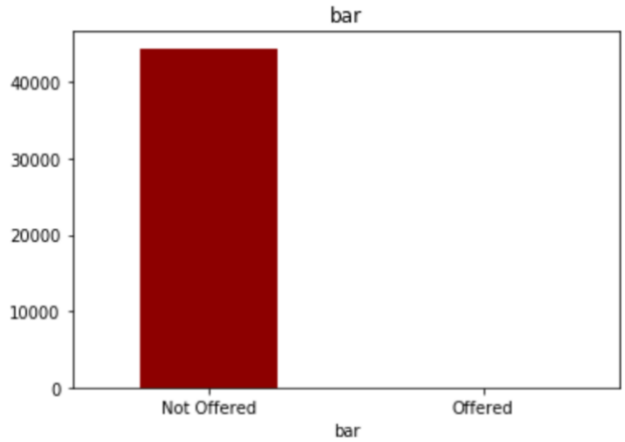
## Amenities

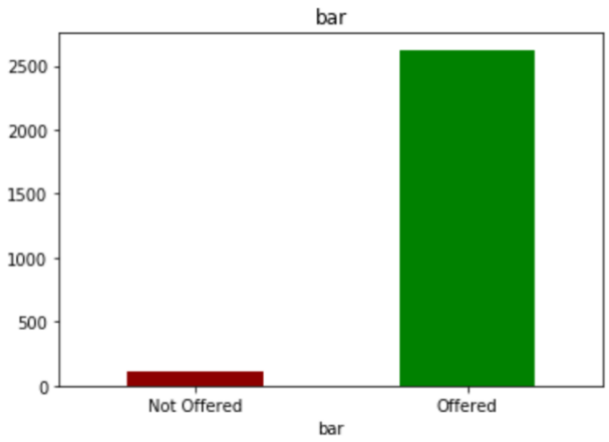
Hosts often provide extra amenities that they believe guests would appreciate. The goal of the exploratory analysis is to determine how the various amenities have an affect on the cost.
**Highly Correlated Amenities:** Based correlation analysis between the different amenities and price, we find that the amenities “dining area”, “pool or tub”, and “bar” have highest correlation with price. The analysis shows that the median price of listings that have a regular or formal dining area is higher by a significant amount of around $500.
Places that offer Swimming pools, spas or saunas also have a higher median price than those that don’t.




Those that offer bars as amenities also have a significantly higher price. The graph on the left shows the number of people offering a bar as an amenity, and the number of people who are not. Despite most people not offering a bar, the median price of places that do offer is very high. The reason for its affect on the price could be because those that offer bars are offering upper-scale accommodation with other factors that would drive up the price.

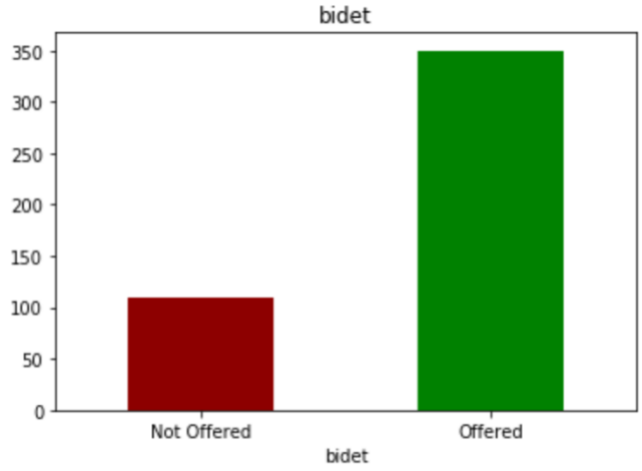



**Low correlated Amenities** There are also amenities, like the bidet, which seem to drive up the price of houses, however, the correlation analysis shows insignificant correlation between this amenity, and the price. The reason for this difference in price could again be due how upscale the rest of the accommodation is.

```
In [46]: correlations_amenities = amenities_data.corr()
         print(amenities_data[amenities_data.columns[1:]].corr()['price'][:].sort_values())
```

```
bidet                             0.036291
```




**Services:** There are services that many hosts offer, which surprisingly seem to have no effect on the median price. Gestures such as 24-hour check-in, self check-in, and the host greeting you on arrival do not seem to affect the price of the Airbnb and should not be reasons to increase the price. The graphs below show that there is not a significant difference between the prices of accommodations whether these services are offered or not, and there are other more important factors to consider if hosts would like to increase the prices of their accommodation.



## Correlations

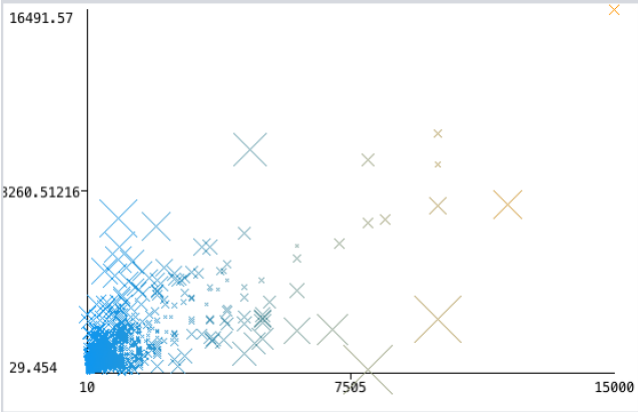
A pairplot is a concise way to check the relationship between variables. In our case above, we can see within a 12x12 pairplot that there is a relationship between the dependent variable

price and the independent variables such as accommodates, bathrooms, bedrooms, extra_people, last_review and number_of_reviews (see Appendix E).

Correlations between the variables were also generated to ensure they were significant. (see Appendix F)

# Prediction – Random Forest Method

The categorical variables used for this predictions were property_type and room_type. The model was generated using 80% of the data for training and the remaining as test data. The number of iterations is set to 100. The results are as shown below.

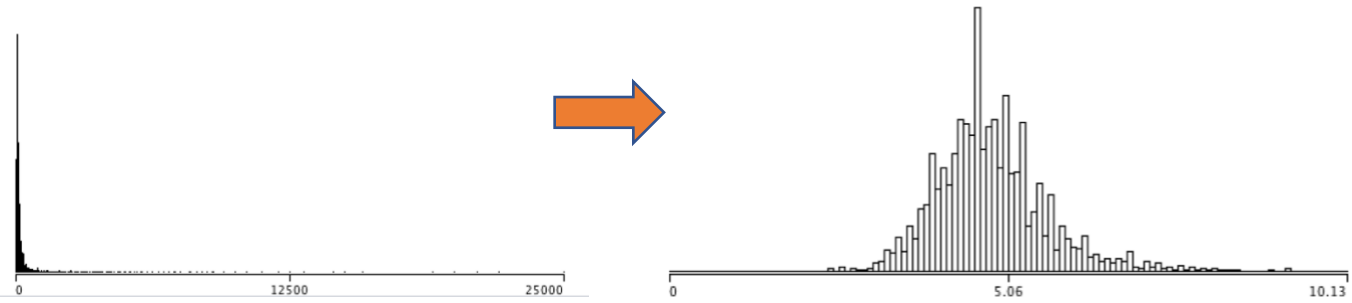



```
=== Summary ===

Correlation coefficient                  0.7977
Mean absolute error                    138.2411
Root mean squared error                351.0769
Relative absolute error                 73.3825 %
Root relative squared error             66.5719 %
Total Number of Instances             8893
```

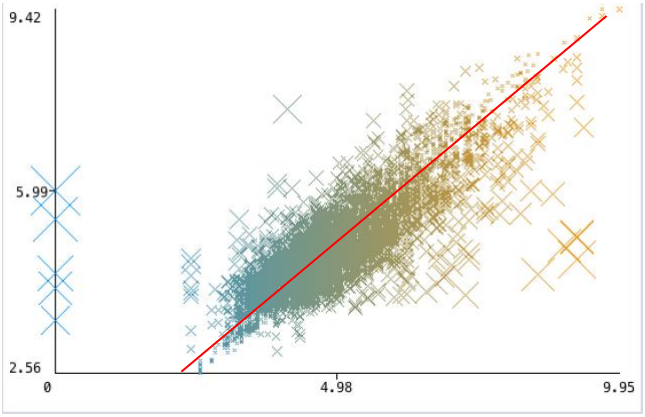
The graph was difficult to interpret due to the price variable being extremely right skewed. In order to improve readability, the price was transformed using log-transform so that it was not right-skewed. The change in distribution is as shown below.

```
clean_data_transformed = pd.DataFrame(clean_data)
clean_data_transformed['price_transformed'] = np.log(clean_data.price + 1)
clean_data_transformed.describe()
```




We did further prep-processing; for the column "room_type" there were 3 possible values: “Entire home/apt”, “Private room”, and “Shared room”. We created dummy variables for this column. The "property_type" column had categories “Apartment”, “House”, and “Other”. We

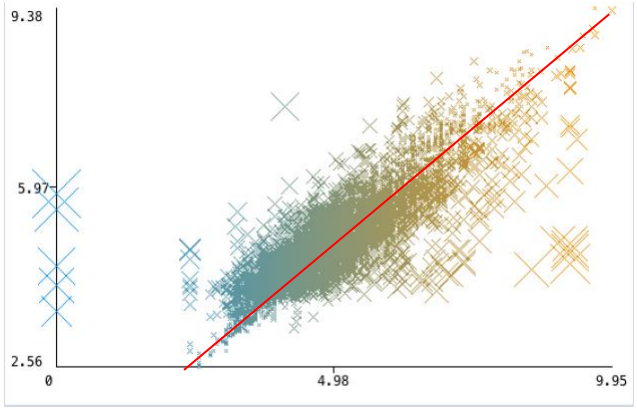
also created dummy variables for those values. Keeping all other parameters the same, we ran the model again, and the Relative absolute error decreased by 27.58%.




```
=== Summary ===

Correlation coefficient                  0.871
Mean absolute error                      0.2922
Root mean squared error                  0.4188
Relative absolute error                 46.2092 %
Root relative squared error             49.1362 %
Total Number of Instances            13339
```

After the above model was generated, we ran it again, this time with the amenities columns. This showed a slight improvement in the model, with a small decrease in our error rates.

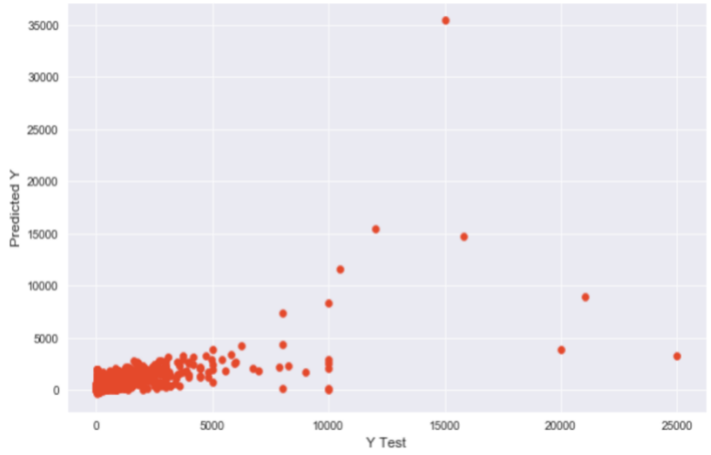



```
=== Summary ===

Correlation coefficient                  0.8788
Mean absolute error                      0.2819
Root mean squared error                  0.4091
Relative absolute error                 44.5668 %
Root relative squared error             47.896  %
Total Number of Instances            15118
```

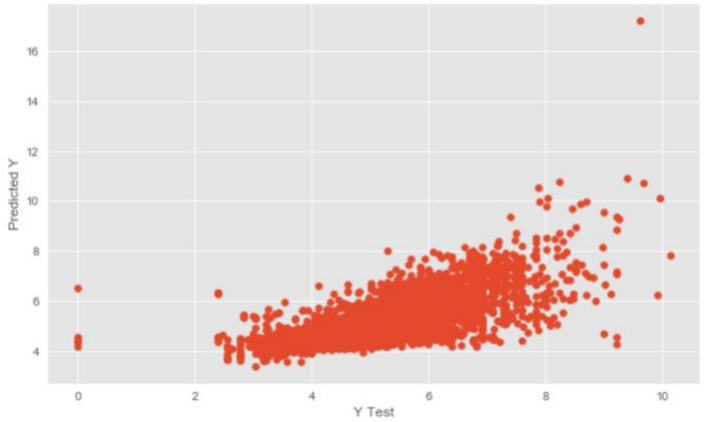
# Prediction – Linear Regression

Linear regression was run with Price as the target variable, and the other variables as predictors. The following results were generated (before transformation of price variable).




| | | | |
|---|---|---|---|
| **Dep. Variable:** | price | **R-squared:** | 0.464 |
| **Model:** | OLS | **Adj. R-squared:** | 0.464 |
| **Method:** | Least Squares | **F-statistic:** | 3496. |

After transformation:




| | | | |
|---|---|---|---|
| **Dep. Variable:** | price_transformed | **R-squared:** | 0.567 |
| **Model:** | OLS | **Adj. R-squared:** | 0.567 |
| **Method:** | Least Squares | **F-statistic:** | 5302. |

```
MAE: 0.4052
MSE: 0.3175
RMSE: 0.563
```

We were able to significantly improve our linear model through feature engineering, just as we did for the Random Forest Model, by creating dummy variables for the categorical attributes, and by using the amenities column in the Linear Regression Analysis. The following results were generated, using a 66% split of the data:

```
MAE: 0.3183
MSE: 0.1943
RMSE: 0.4408
```

# Classification – Logistic Regression

We conducted Logistic Regression in order to help hosts decide whether their house needs to be priced in the higher range or lower range. We chose a value of $135, which is the average price of Hotels in the US, for splitting the price into two halves. For any prices equal to or larger than $135, we ranked them as “High”, the others ranked as “Low”. The bins were chosen to be 10 with equal frequency, with cross-validation folds = 10. The accuracy achieved was 80.14%.

```
=== Confusion Matrix ===

     a     b   <-- classified as
 17971  4506 |     a = High
  4325 17663 |     b = Low
```

```
=== Summary ===

Correctly Classified Instances       35634               80.1394 %
Incorrectly Classified Instances      8831               19.8606 %
Kappa statistic                          0.6028
Mean absolute error                      0.2704
Root mean squared error                  0.368
Relative absolute error                 54.0839 %
Root relative squared error             73.6045 %
Total Number of Instances            44465
```

We changed bins from 10 to 15 and tried both equal frequency and unequal frequency and found out that they have no influence on the accuracy. The percentage split was set as 70% and the accuracy increased to 82% after training the data.

```
=== Summary ===

Correctly Classified Instances       10927               81.9177 %
Incorrectly Classified Instances      2412               18.0823 %
Kappa statistic                          0.6129
Mean absolute error                      0.2539
Root mean squared error                  0.354
Relative absolute error                 53.0637 %
Root relative squared error             72.3031 %
Total Number of Instances            13339

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.900    0.302    0.818      0.900    0.857      0.618    0.896     0.922     Low
                 0.698    0.100    0.821      0.698    0.754      0.618    0.896     0.866     High
Weighted Avg.    0.819    0.222    0.819      0.819    0.816      0.618    0.896     0.900

=== Confusion Matrix ===

    a    b   <-- classified as
 7221  806 |    a = Low
 1606 3706 |    b = High
```

Although our prediction accuracy was high, by splitting our data this way, there is a lot of information loss. It may be more beneficial to have more categories for splitting our data.

# Conclusions and Key Takeaways

The coefficients of the regression equation generated (see Appendix G), and the exploratory analysis show the following:

1. For every extra person a host can accommodate, the average price of the property is expected to increase by around $23, keeping all other variables constant.
2. For every increase in the number of beds in a room, the average price of the property goes down by aorund $33, keeping all other variables constant.
3. The p-value for the extra_charge variable is quite high, showing that the extra charge that is added for every additional guest added to the booking does not affect the base price of the property. Therefore, hosts that are flexible with adding more guests should carefully consider the charge per guest, instead of raising the base price of the airbnb.
4. Adding extra services like greeting your guests on arrival or 24-hour checkin should not increase the value of the property significantly.
5. Places with a larger number of bathrooms have a much higher price on average compared to the places with a large number of bedrooms.
6. The type of property, whether its an apartment, house or any other type of property does not have a significant influence on the price, as the p-value is quite large for these variables.

## Model Recommendation

| | **Random Forest** | Linear Regression |
|---|---|---|
| Mean Absolute Error | **0.28** | 0.31 |
| Root Mean Squared Error | **0.41** | 0.44 |
| Mean Squared Error | **0.16** | 0.19 |
| | | |

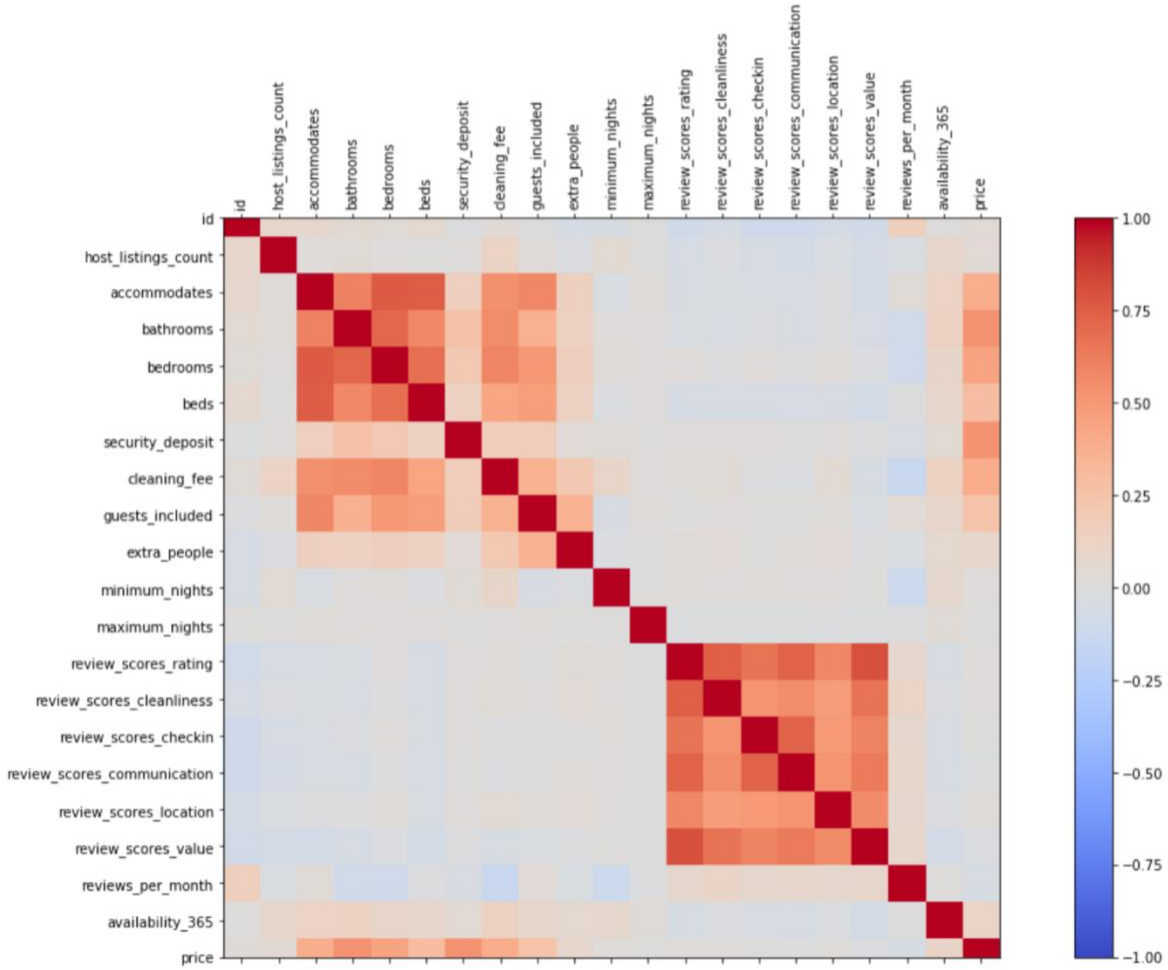
The **Random Forest** predictions have lesser error rates compared to the Linear Regression Models, with the same dataset, pre-processing steps taken before running each model, and split of data between training and test (66%). Therefore, we recommend the Random Forest Model for predicting the Airbnb House Prices with the features we have used.

Our analyses were done using a combination of Weka and Python. Weka was particularly useful for the Random Forest implementation as we could focus on tweaking different parameters to achieve higher accuracy. Linear regression was done using Python. The coefficients generated using Weka and Python were similar.

# Challenges and Drawbacks

1. A challenge we faced was interpreting the results after normalising the price variable, as we they no longer represented the original values.
2. We were missing some important data in our dataset, such as seasons – as prices tend to vary depending on the time of the year, and occasions. Having this information is essential to achieving a higher accuracy on our price predictions.
3. We removed location from our final dataset as there were over 400 categories for this. Finding some way to include location in our future models would allow us to predict the prices more accurately.
4. Our model was better at predicting the lower values, but the accuracy of predictions for the higher-priced houses was low as seen from the Actual vs Predicted graphs. This high price for the property is most likely due to its location, which we are not analysing in our model.

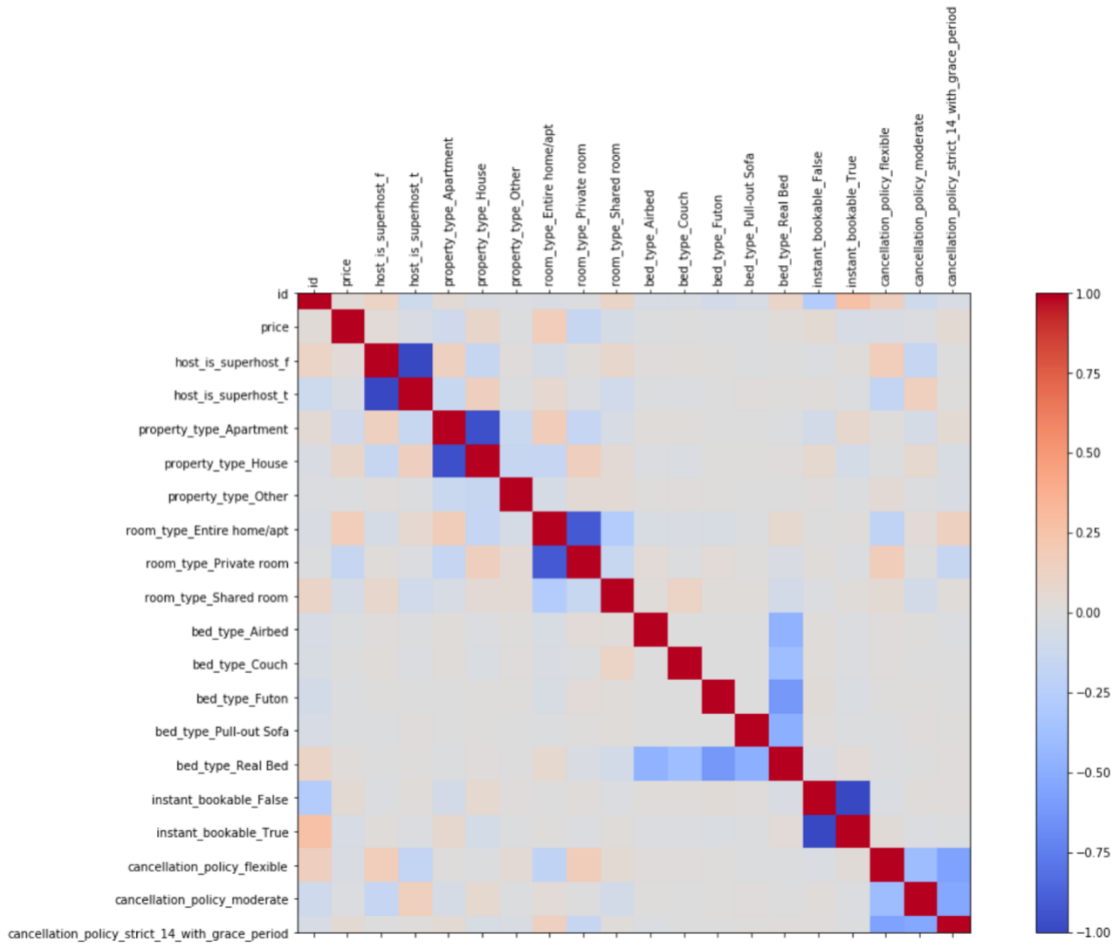
# Appendix A - Correlation between numeric variables




```
print(numeric_data[numeric_data.columns[1:]].corr()['price'][:].sort_values())
```

```
reviews_per_month            -0.040387
review_scores_value          -0.016975
review_scores_communication  -0.008702
review_scores_checkin         0.000514
maximum_nights                0.007057
review_scores_rating          0.007932
minimum_nights                0.009897
review_scores_cleanliness     0.010610
review_scores_location        0.019488
host_listings_count           0.042013
extra_people                  0.082354
availability_365              0.102457
guests_included               0.243109
beds                          0.292671
accommodates                  0.387066
cleaning_fee                  0.394526
bedrooms                      0.443785
security_deposit              0.526929
bathrooms                     0.528871
price                         1.000000
Name: price, dtype: float64
```

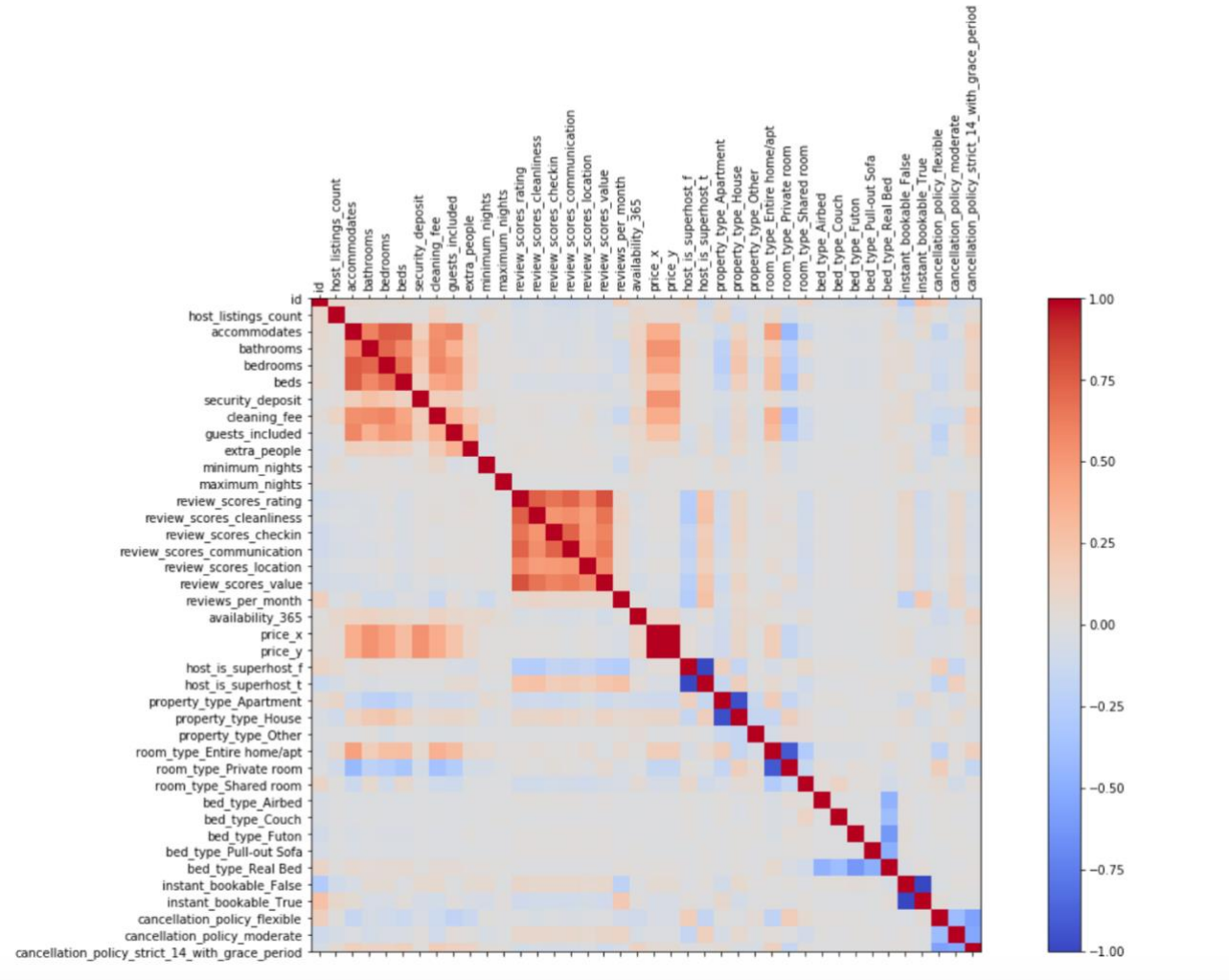
# Appendix B - Correlation between categorical variables




```
print(cat_data_dummies[cat_data_dummies.columns[1:]].corr()['price'][:].sort_values())
```

```
room_type_Private room                            -0.150893
property_type_Apartment                           -0.096241
room_type_Shared room                             -0.054893
instant_bookable_True                             -0.051524
cancellation_policy_flexible                      -0.033489
host_is_superhost_t                               -0.031779
cancellation_policy_moderate                      -0.018193
property_type_Other                               -0.014532
bed_type_Futon                                    -0.011641
bed_type_Airbed                                   -0.010057
bed_type_Pull-out Sofa                            -0.009981
bed_type_Couch                                    -0.007440
bed_type_Real Bed                                  0.019830
host_is_superhost_f                                0.031709
cancellation_policy_strict_14_with_grace_period    0.047028
instant_bookable_False                             0.051524
property_type_House                                0.100040
room_type_Entire home/apt                          0.169090
price                                              1.000000
Name: price, dtype: float64
```

## Appendix C - Correlation between all variables

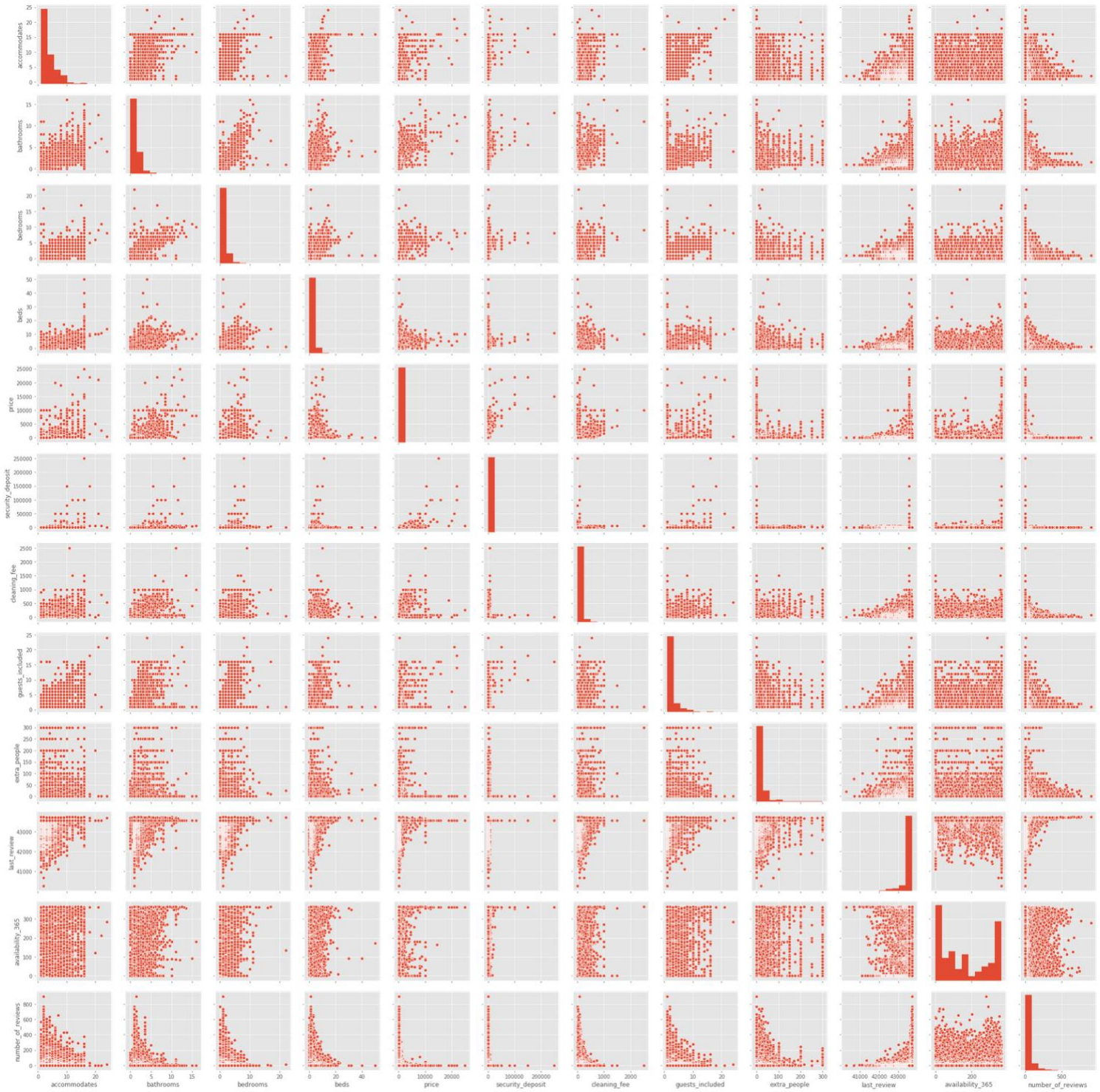



## Appendix D - Order of Impurity in Random Forest Model

```
Attribute importance based on average impurity decrease (and number of nodes using that attribute)

    205.79 (  1829)  Entire home/apt
     38.08 (  2735)  Private room
      8.9  ( 59721)  bedrooms
      3.83 ( 94572)  bathrooms
      3.26 (137781)  accommodates
      1.53 (290106)  cleaning_fee
      1.02 (139026)  beds
      0.83 ( 34033)  House
      0.73 ( 34809)  Apartment
      0.69 (263612)  security_deposit
      0.53 (181959)  extra_people
      0.52 (100423)  guests_included
      0.47 (372694)  availability_365
      0.43 (297298)  number_of_reviews
      0.37 (361649)  last_review
```

# Appendix E - Pairwise plots of final variables



## Appendix F - Correlations between final variables


## Appendix G – Correlation coefficients

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | 143.5256 | 242.982 | 0.591 | 0.555 | -332.723 | 619.774 |
| **accommodates** | 22.9371 | 1.535 | 14.947 | 0.000 | 19.929 | 25.945 |
| **bathrooms** | 194.7892 | 3.268 | 59.601 | 0.000 | 188.383 | 201.195 |
| **bedrooms** | 39.1173 | 3.461 | 11.303 | 0.000 | 32.334 | 45.900 |
| **beds** | -33.2851 | 2.007 | -16.581 | 0.000 | -37.220 | -29.351 |
| **security_deposit** | 0.1120 | 0.001 | 113.828 | 0.000 | 0.110 | 0.114 |
| **cleaning_fee** | 0.6178 | 0.032 | 19.385 | 0.000 | 0.555 | 0.680 |
| **guests_included** | -15.1588 | 1.505 | -10.073 | 0.000 | -18.109 | -12.209 |
| **extra_people** | 0.0709 | 0.083 | 0.858 | 0.391 | -0.091 | 0.233 |
| **last_review** | -0.0125 | 0.007 | -1.681 | 0.093 | -0.027 | 0.002 |
| **availability_365** | 0.0958 | 0.015 | 6.249 | 0.000 | 0.066 | 0.126 |
| **number_of_reviews** | -0.0681 | 0.036 | -1.884 | 0.060 | -0.139 | 0.003 |
| **Apartment** | 59.7187 | 81.069 | 0.737 | 0.461 | -99.178 | 218.615 |
| **House** | 57.8881 | 81.302 | 0.712 | 0.476 | -101.465 | 217.241 |
| **Other** | 25.9188 | 81.571 | 0.318 | 0.751 | -133.962 | 185.800 |
| **Entire_home_apt** | 123.2389 | 11.624 | 10.602 | 0.000 | 100.455 | 146.023 |
| **Private_room** | 106.1641 | 11.419 | 9.297 | 0.000 | 83.782 | 128.546 |