Analysis on San Francisco Airbnb Open Data

Problem

Thanks to Airbnb, people now have more choices of accommodation when they travel to another city or a foreign country. Airbnb provides travellers with the opportunity to stay in a unique home and to deeply experience the local life. This analysis focuses on one of the most important aspects of an Airbnb listing - the price. I'm going to build a model that can predict the price of a listing in San Francisco based on its location, amenities, reviews and any other information that may apply.

Travellers who are planning a trip to San Francisco and Airbnb hosts in San Francisco will be interested in this analysis. Possible application of this analysis can be: finding the best price of a listing in a neighborhood and setting price based on the details of the listing.

Data

The data I used is from insideairbnb.com. The dataset 'listings' contains the detailed information of 7151 listings in San Francisco area on Airbnb, including location, host info, amenities, property type etc. (7151 records with 106 columns in total).

Data Cleaning Process

Listings table has 106 columns and almost half of them have missing values. I took the following steps to deal with the missing values:

A. Values in some columns such as **thumbnail_url**, **medium_url** and **x1_picture_url** are all missing, these columns are dropped from listings

B. **square_feet** column contains only 126 records which is less than 2% of total, dropped from listings

C. Some columns are not related to predicting the listing price, such as **scrape_id**, **last_scraped**, **picture_url** etc., are dropped from listings

D. Some columns contain duplicate information with other columns, such as id and **listing_url**, **host_total_listing_count** and **host_listing_count** etc., one of each pair is dropped from listings

E. Some columns contains descriptive scripts written by hosts which can be reflected by other numeric columns, such as **name**, **summary**, **description** etc. are dropped from listings

F. The rest columns with missing values

a. **house_rules**: missing value can be a sign of no additional rule, so I created a column named 'additional_house_rule' with value True in rows where house_rules is not missing and False in rows where house_rules is missing value. Then I dropped house_rules column
b. **host_location, jurisdiction_names**: these columns seem to be missing at random, rows with missing values are dropped
c. **zipcode**: used longitude and latitude to get the corresponding zipcode
d. **bathrooms, bedrooms, beds**: filled missing value with median
e. **weekly_price, monthly_price**: multiplied price by 7 to get weekly_price, and multiplied price by 30 to get monthly_price
f. **security_deposit, cleaning_fee**: filled missing value with 0
g. **host_response_time, host_response_rate, first_review, last_review, review_scores_rating, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value, reviews_per_month**: the missing value of these columns may flag a new host or a new listing, I filled missing values in review score columns with median, in reviews_per_month with 0, in first_review and last_review with date of today
h. **license**: missing value shows that the property doesn't have license. I created a column named 'have_license' with boolean values based on license column and dropped license column from listings

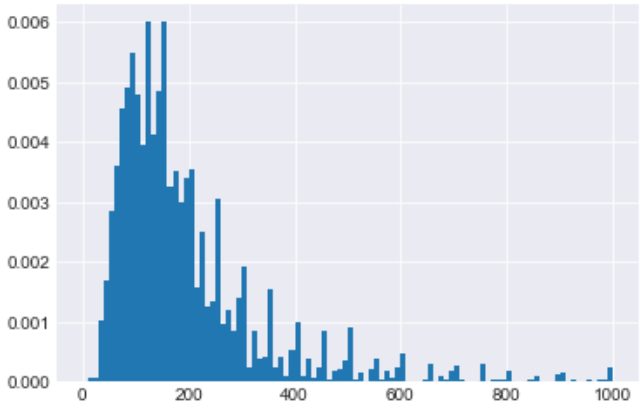
Exploratory Data Analysis

The dataset 'listings' consists of price, host features, property features and reviews of listings. In this section, I'm going to explore each part of the data and see if there is any trend, pattern or relationship.

Price

First let's check out the distribution of listing prices.

According to the histogram below, the distribution of price is right skewed with most listings priced under 200 dollars.



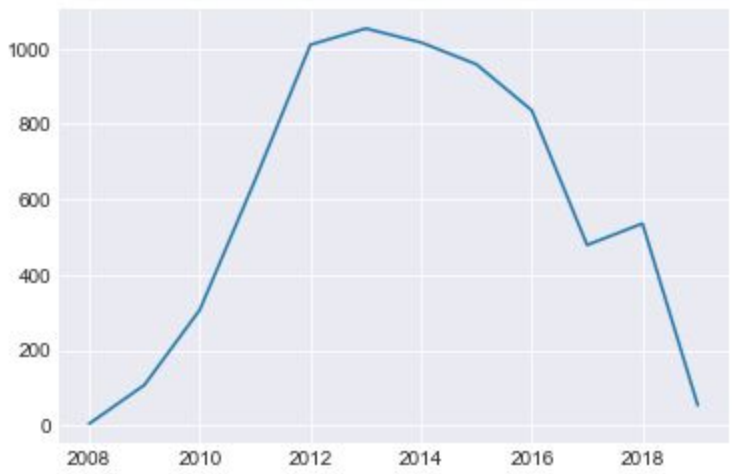
Host

- Host start year

  The line plot below shows the number of hosts joined per year. We can see that in 2008-2012 there was a fast increase in new hosts, and 2013 had the most newly-joined hosts. After 2013, number of new hosts dropped until 2018.

  The trend seems to match the home price trend, in 2008-2013 there was a continuous decline in home price after the market crash in 2008, and starting from 2013, the home market in San Francisco recovered with price increasing rapidly. This may indicate that the number of new hosts in Airbnb market is negatively related to the home price index, when home price is increasing, owners may tend to sell their extra properties instead of renting them out.

  Another possible explanation is that, after Airbnb was founded in 2008, the company experienced an increase in new hosts for 5 years till 2013, and the market calmed down and number of new hosts dropped after 2013.



- Superhost

  Superhosts are hosts who meet certain criteria such as 4.8+ overall rating, 10+ stays during the past year etc.. Airbnb states that superhosts earn up to 22% more than other hosts.

  The boxplots below show that the distribution of price is quite similar regardless of superhost or not. However superhosts have a higher median number of reviews (which is related to the number of bookings) than non-superhosts. Therefore superhost may not be related to higher price or lower price, but may be related to more reviews (more bookings).

To better understand the relationship between superhost and price, I did a t-test to compare the mean of listing price across superhost and non-superhost group:

H_0: $\mu_{superhost} = \mu_{normal}$
H_A: $\mu_{superhost} > \mu_{normal}$

t-statistic = 1.12, p-value = 0.13 > 0.05, fail to reject the null hypothesis.

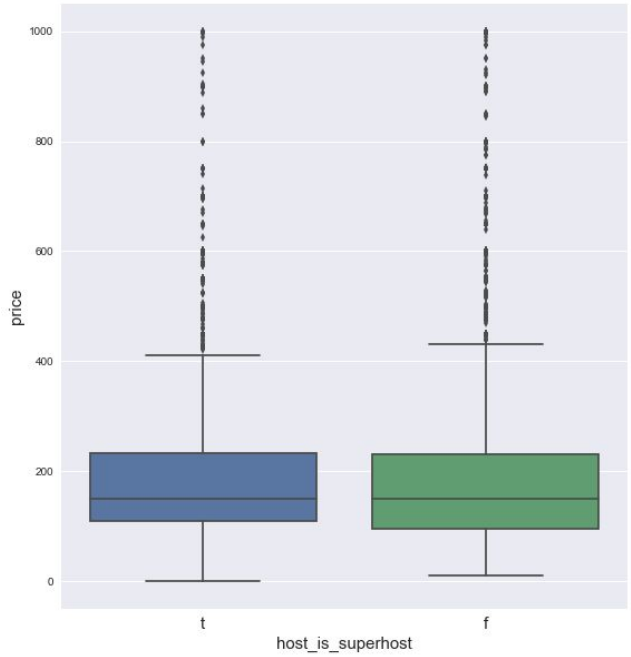
The mean listing price are the same for superhost and non-superhost.

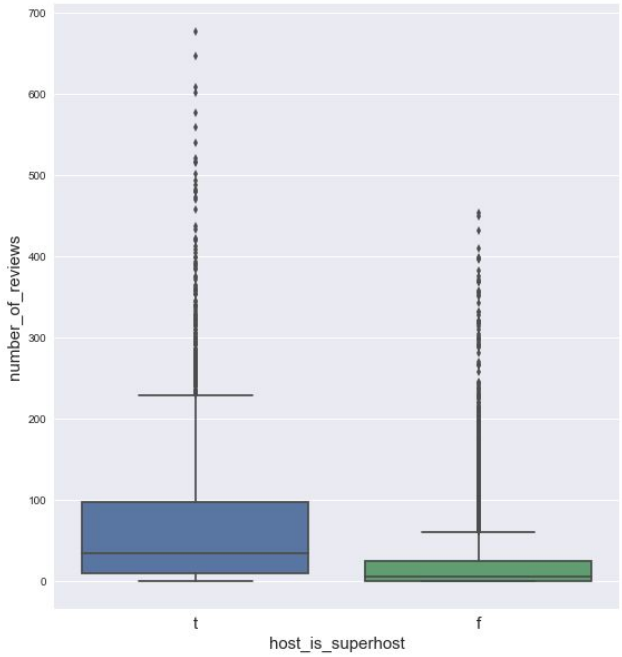
I also did a t-test to compare the mean number of reviews of two groups:

H_0: $\mu_{superhost} = \mu_{normal}$
H_A: $\mu_{superhost} > \mu_{normal}$

t-statistic = 25.63, p-value = 7.64e-139 << 0.05, reject the null hypothesis.

We have enough evidence to reject the null hypothesis, that is to say, the mean number of reviews(which can be seen as an indicator of number of booking) of superhosts is higher than that of normal hosts.

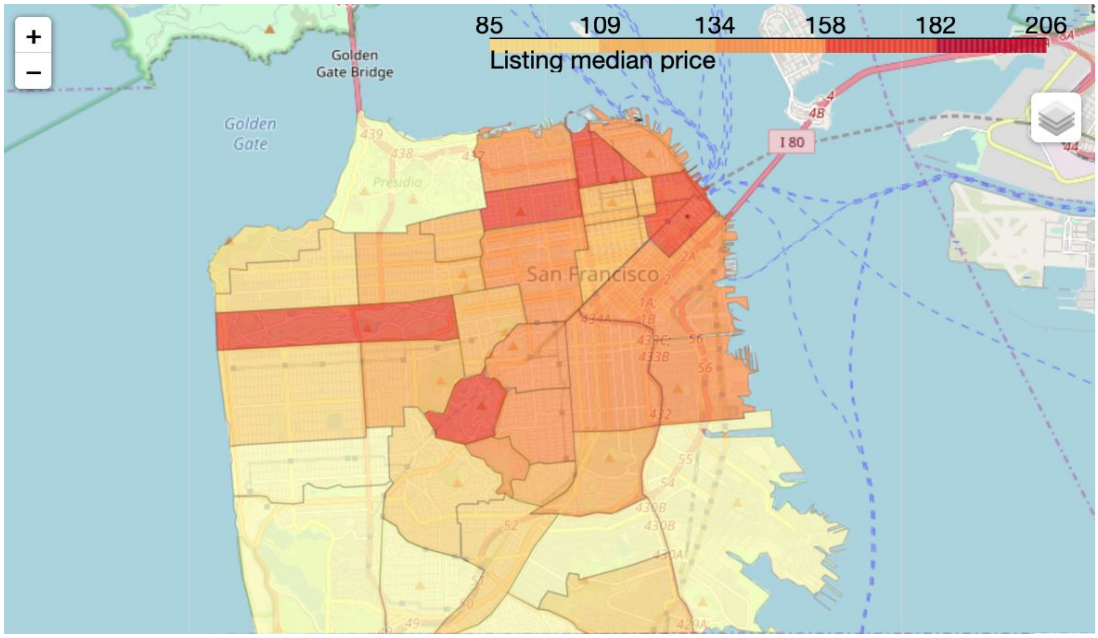




Property

- Neighborhood

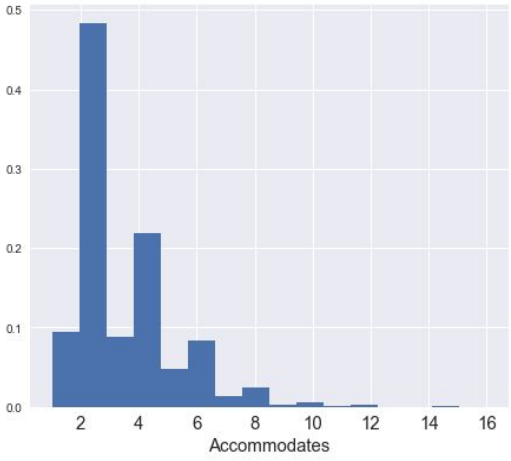
The map below shows the median listing price by neighborhoods in San Francisco. According to the map, listings in the northeastern part of SF have a relatively higher median price, and southern part has the lower median listing price.

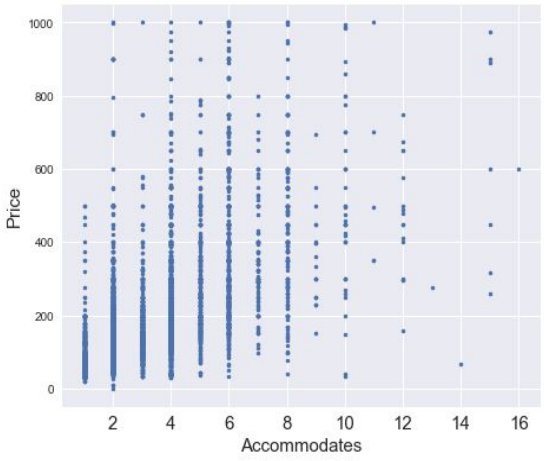


- Number of people the listing can accommodate

The number of people a listing can accommodate reflects the size of the listing. The histogram below shows the distribution of 'accommodates' column. Nearly half of the listings can accommodate only 2 people, and around 80% of the listings can accommodate 4 or less people.

According to the scatter plot on the right, there is a positive relationship between price and accommodates. The relationship between price and accommodates is intuitive: the more people can a listing accommodate, the higher the price tends to be.

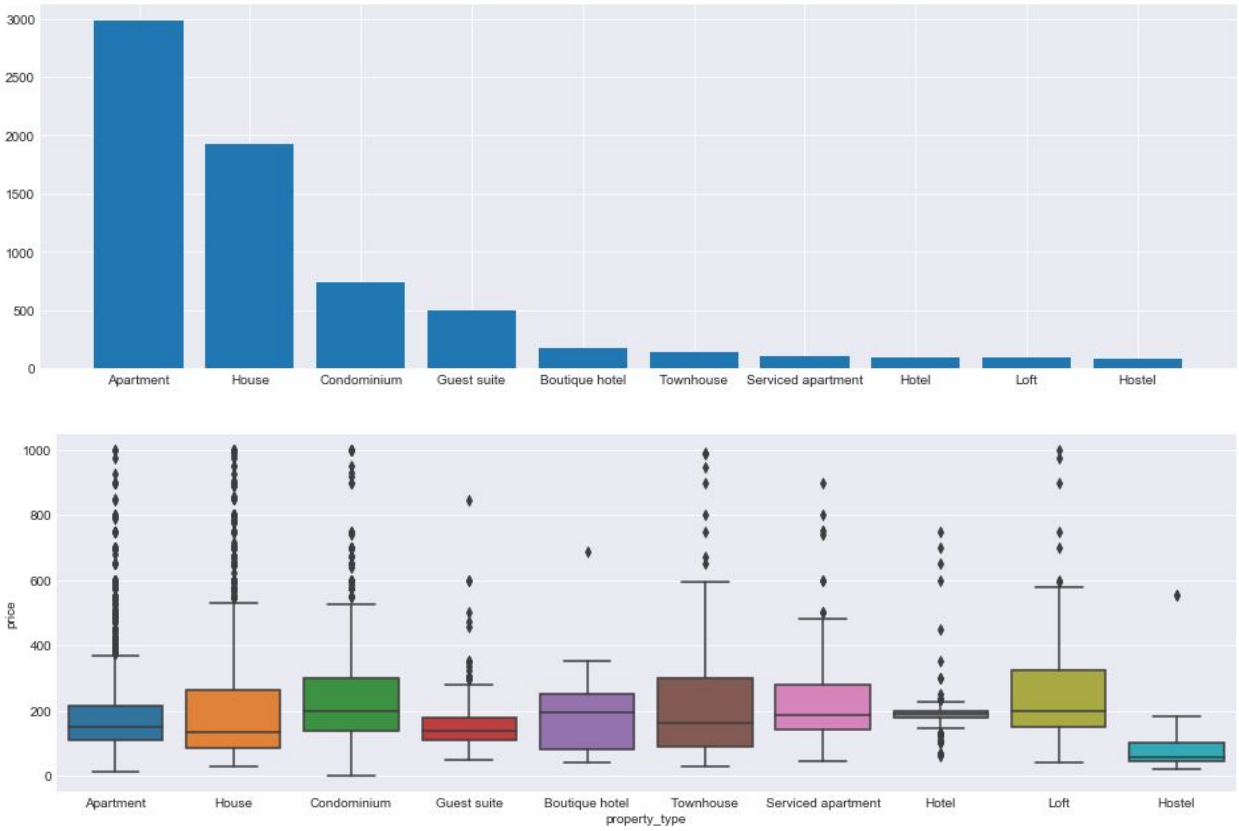




- Property type

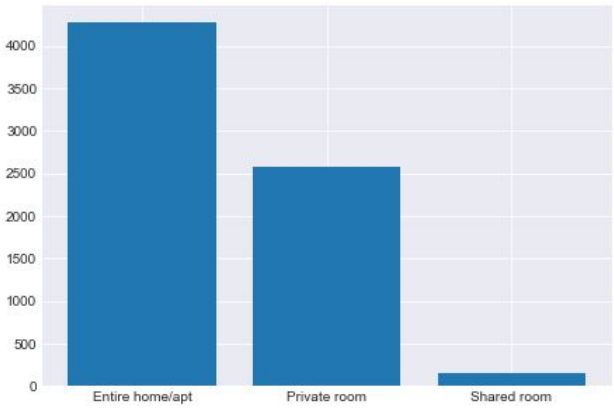
  The bar plot below lists the top 10 property types across the city. Apartment and house are the 2 most common options.

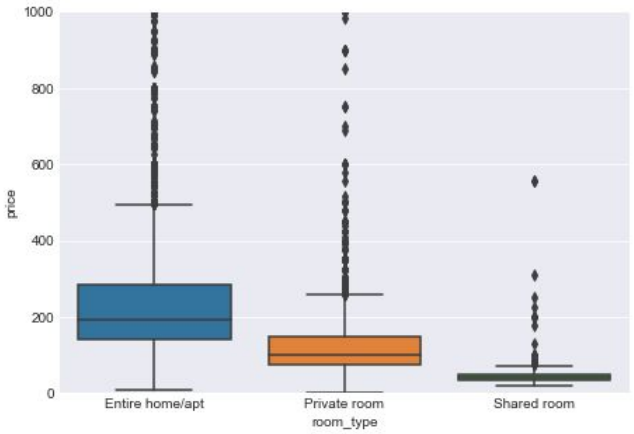
  The boxplot shows the distribution of listing prices by property type. We can see that condominiums, boutique hotels, and loft have relatively higher median price, and hostels are usually lower priced.



- Room type

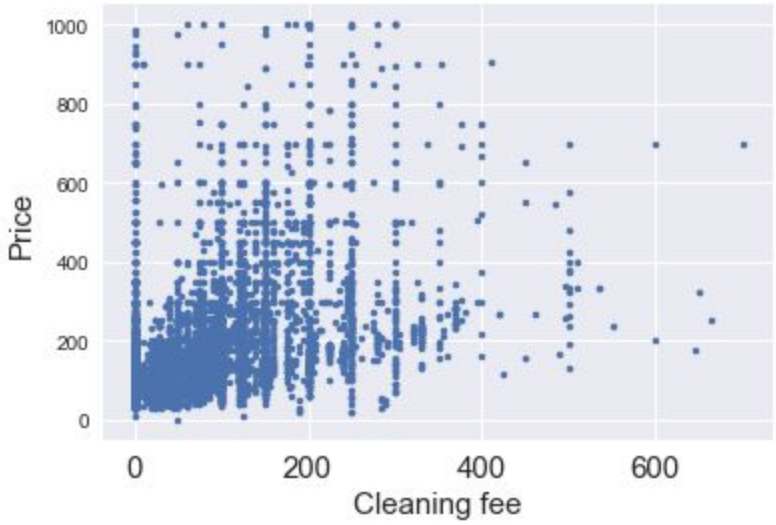
  The bar plot and boxplot show that most of the listings are entire home/apt with a highest median price among three room types. Shared rooms are not common comparing to entire home/apt and private room, but the price seems to be really nice - the median price of shared rooms is about 1/4 of that of entire home/apt!





- Cleaning fee

  Cleaning fee is an extra cost for Airbnb customers. Usually they need to pay a one-time cleaning fee when they make reservations, no matter how many nights they will stay. The following scatterplot shows that there is a positive correlation between cleaning fee and price. It is intuitive as the more cleaning fee it charges, the more labor it requires to clean the listing, and usually it indicates that the listing is either large or complicated.



Reviews

Reviews measure how much customers like the listing. I'm going to check:

1. review score, the rating of the listing
2. number of reviews, an alternative of number of bookings

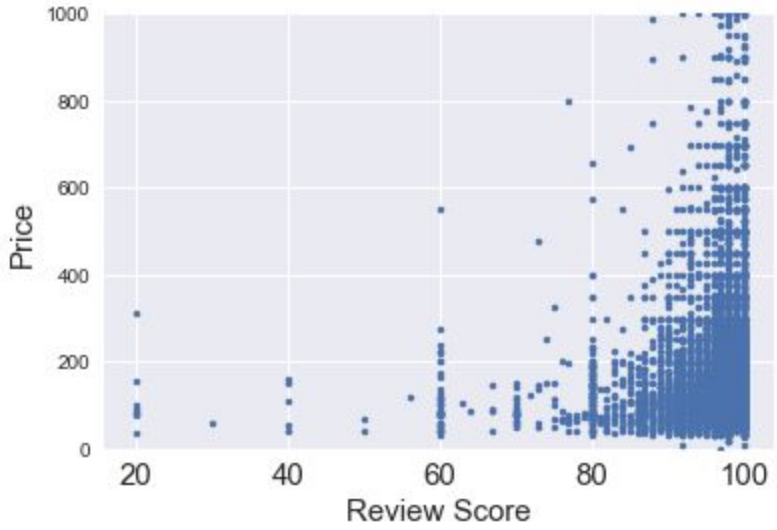
- Review score

  Review score indicates how customers like the listing. The scatter plot below shows a positive correlation between them, that is, a higher rated listing is likely to have a higher price.

  Due to the large sample size, it's hard to identify the relationship between reviews scores and listing price from the scatter plots. So I did a hypothesis testing on the significance of correlation.

  > *H0: price and review scores are independent*
  > *HA: there is a dependency between the samples*
  >
  > *Spearman's rank correlation = 0.22, p-value = 5.58e-78 << 0.05*
  > *Kendall's rank correlation = 0.16, p-value = 1.97e-78 << 0.05*

  We have enough evidence to reject the null hypothesis, that's to say, price is correlated with review scores.



- Number of reviews

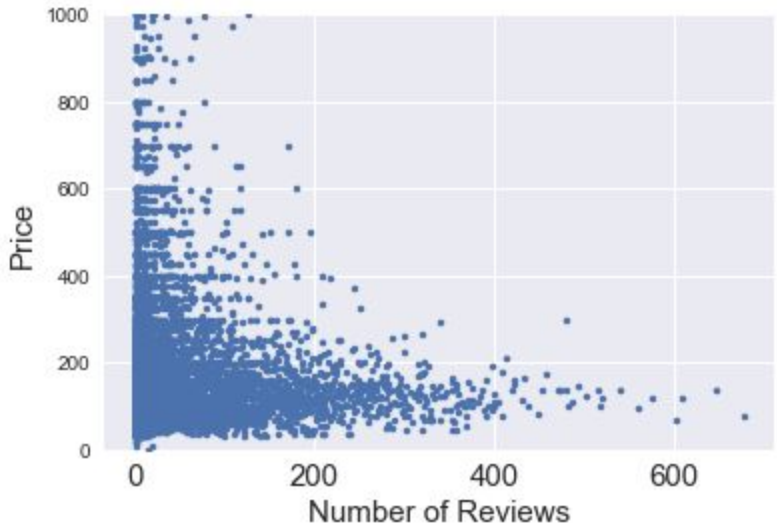
  Number of reviews can somewhat represent the popularity of the listing. The scatterplot below shows that price is negatively correlated with number of reviews, which indicates that people prefer booking listings with lower price.

  Hypothesis testing on correlation between price and number of reviews:

  > *H_0: price and number of reviews are independent*
  > *H_A: there is a dependency between the samples*
  >
  > *Spearman's rank correlation = -0.11, p-value = 9.22e-20 << 0.05*

Kendall's rank correlation = -0.07, p-value = 9.99e-19 << 0.05

We have enough evidence to reject the null hypothesis, that's to say, price is correlated with number of reviews.



Conclusion

The number of new Airbnb listings in the SF market might depends on the home price in that area. A booming market of Airbnb may associated with a recession in home market.

Superhosts earn more than regular hosts because they have more bookings instead of higher price!

Listing features play an important role in deciding the price. The location/neighborhood, size, property type, and room type are all factors that have effects. Cleaning fee is excluded from nightly price, but is positively correlated with list price, so be prepared to pay high cleaning fee for an expensive stay!

Review score is positively correlated with price and number of reviews is negatively correlated with price. That means, people are more likely to book low price listings but they may have more fun with a higher priced listing.

Model Selection

Predicting listing price based on listing features is a supervised learning problem, regression models are used in this project. The following steps are followed to model the data:

- **Data processing:** the preprocessed dataset is split into a training set (70%) and a testing set (30%), and K-Fold Cross Validation is applied on training set

    - The dataset consists of numeric features and categorical features
    - The following features are excluded from numeric features:
        - Index, id, host_id: indices and ids of listings and hosts
        - Zipcode, latitude, longitude: provide information of the listing locations which is similar to what 'neighbourhood' feature does
        - Weekly_price, monthly_price: calculated based on nightly price which is the dependent variable in the model
        - Security_deposit, cleaning_fee: usually set by host together with nightly price, should not be included in the model as predictive features
    - Categorical features are included based on the statistical analysis result above
- **Feature selection/reduction:** backward elimination and Principal Component Analysis (PCA) are used
- **Modeling:** Linear Regression models and Random Forest Regression models are used to fit the data
- **Evaluation:** R-squared of regression models and Mean-square-error (MSE) are used as metrics to compare the performance of different models

Linear Regression:

- **Model 1**: Linear regression on numeric features

  Model 1 is trained using numeric features in training set. Backward elimination based on the p-value of coefficient of each feature in the regression model is applied in order to select features.

- **Model 2, 3, 4, 5**: Linear regression on scaled/normalized numeric features

  Backward elimination is applied on each model to select the significant features.
    - Model 2 uses Standard scaled data
    - Model 3 uses MinMax scaled data
    - Model 4 uses Robust scaled data
    - Model 5 uses normalized data

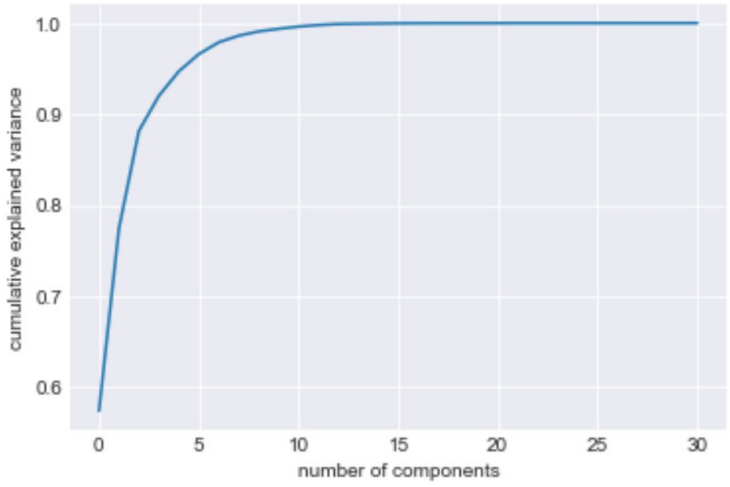
  Measurements show that the Robust scaler works better than other scaler/normalizer but no better than the original data, therefore, the original data will be used instead.

- **Model 6**: Linear regression on numeric data after removing highly correlated features

  Model 6 is trained on features selected from Model 1 without features which have high correlation (Pearson's correlation coefficient >= 0.8) with other features and lower correlation with price:
    - Beds: highly correlated with 'accommodates' (Pearson's correlation coefficient = 0.8237), keep 'accommodates' in the model
    - Calculated_host_listings_count_entire_homes: highly correlated with 'calculated_host_listings_count' (Pearson's correlation coefficient = 0.9846), keep 'calculated_host_listings_count' in the model

- **Model 7**: Linear regression on numeric data after PCA

  The figure below shows that 10 may be a good number of components in this case. Model 7 is trained on the 10 components transformed by PCA.



- **Model 8**: Linear regression on features from model 6 and categorical features

  The data used in model 8 includes the following categorical features:
  - Neighbourhood
  - Property_type
  - Room_type

  All categorical features are processed with One Hot Encoding.

Random Forest Regression:

- **Model 9**: Random Forest Regression on numeric features and selected categorical features

  The data used in model 9 includes all numeric features before backward elimination and the following categorical features:
  - Neighbourhood
  - Property_type
  - Room_type
  - Host_response_time
  - Host_is_superhost
  - Instant_bookable
  - Is_business_travel_ready
  - Cancellation_policy

GridSearchCV is applied to tune hyperparameters such as number of estimators, maximum features at every split, maximum number of levels in tree, and minimum number of samples required at each leaf node.

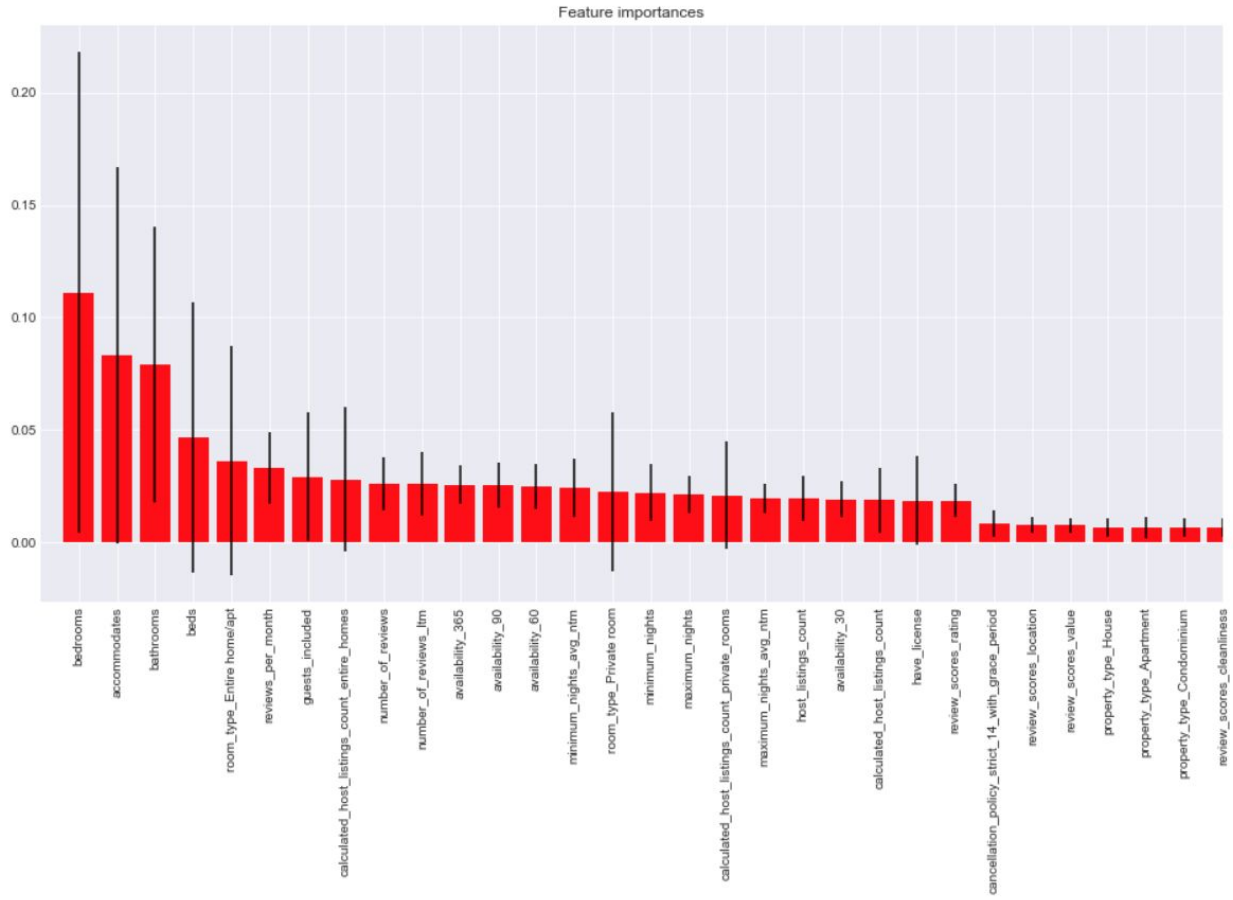
Model Evaluation and Selection

The table below shows the R-squared and MSE of each model on both training set and testing set. According to the metrics, Model 9 has the best performance among all of the models.

Model	Details	R^2 training	R^2 testing	MSE training	MSE testing
1	Linear Regression (LR) on numeric features	0.4791	0.4981	11120.12	10540.95
2	LR on Standard scaled numeric data	0.4791	0.4981	11119.31	10542.13
3	LR on MinMax scaled numeric data	0.4791	0.4981	11119.31	10542.13
4	LR on Robust scaled numeric data	0.4791	0.4981	11120.12	10540.95
5	LR on normalized numeric data	0.3089	0.3392	14699.55	10629.94
6	LR on non-collinearity numeric data	0.4751	0.4939	11200.50	10629.94
7	LR on numeric data after PCA	0.0914	0.0986	19259.13	18931.74
8	LR on data from model 6 plus 2 categorical features	0.5715	0.5940	9553.42	8526.51
9	Random Forest Regression on numeric and categorical features	0.9472	0.6315	1111.65	7739.23

Conclusion

The Random Forest Regressor does the best work on this data, the following graph shows the feature importance and the corresponding standard deviation of the Random Forest model:



Feature ranking:

1. feature bedrooms (0.111054)

2. feature accommodates (0.083135)

3. feature bathrooms (0.079152)

4. feature beds (0.046686)

5. feature room_type_Entire home/apt (0.036040)

6. feature reviews_per_month (0.032954)

7. feature guests_included (0.029107)

8. feature calculated_host_listings_count_entire_homes (0.027959)

9. feature number_of_reviews (0.025851)

10. feature number_of_reviews_ltm (0.025715)

The top 5 features are number of bedrooms, number of people the listing can accommodate, number of bathrooms, number of beds, and type of the room. So we can conclude that the size and capacity of a listing play the most important role in pricing, other features such as number of reviews, number of listings a host has (may be an indicator of if the host is experienced) and number of guests included also contribute to the price of a listing.

Limitations and Next Steps

There is an overfitting problem in Model 9, as the R^2 on training set is 0.95 vs. that on testing set is 0.63, and the MSE of prediction on testing set is much greater than that on training set. This is mainly because the minimum number of samples required at each leaf node is lowered to 1. Increasing the minimum_samples_leaf can reduce the overfitting problem, however, the prediction R^2 drops and MSE increases at the same time.

Following are some more approaches to try in the future:

- Based on the feature importance graph, we can create a model with top features so that overfitting problem may be reduced
- Some categorical features (e.g. ‘neighbourhood’) have many levels but only a few of them are significant, so grouping levels based on some specific rules may help us find the underlying relationship
- Categorizing price into several levels and turn the problem into a classification problem