To: Professor Qi Wu

From: Shuyi Yu, Tongrui Yu, Hongyang Zhou

Date: Apr 22th, 2022

Re: Revenue optimization for Airbnb listings

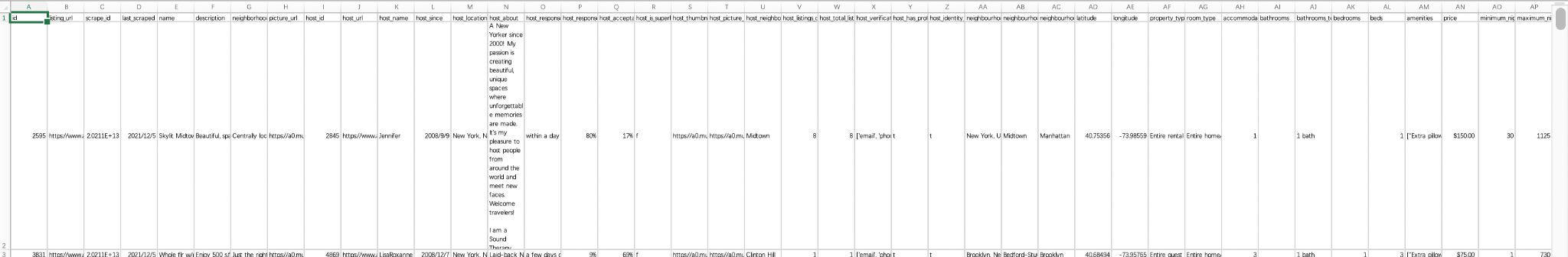
**Introduction**

Airbnb is an O2O service platform that connects travelers with homeowners who have rooms available for rent, providing users with a wide range of accommodation information. Our goal is to help Airbnb find the optimal listing price and get the maximum revenue. Based on what we learned in class, the focus would be on the price and demand. Tried to figure out how the factors impact the demand and price and build the model for them. We were going to prepare the listing data, and demand data and take descriptive statistics analysis at first, followed by classification using a clustering model. Then determined the prediction model and estimated demand function. The next step is to optimize the price to maximize the revenue. Finally, the sample from the validation dataset would be tested by the previous model to obtain the optimal price.

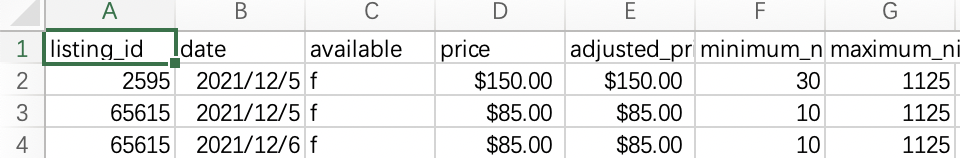
**Data Description**

The data comes from the Inside Airbnb website (<http://insideairbnb.com/get-the-data/>). We searched the data of New York City and downloaded it. The dataset has two major spreadsheet listings and calendar. Listing data includes some important host information, such as host id, location, room information, review contents, etc. The calendar dataset involves the availability of the listing, price, and nights.

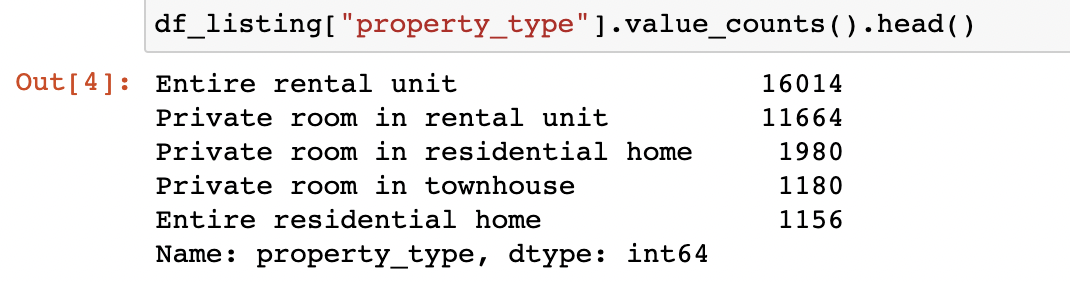
Listing table:



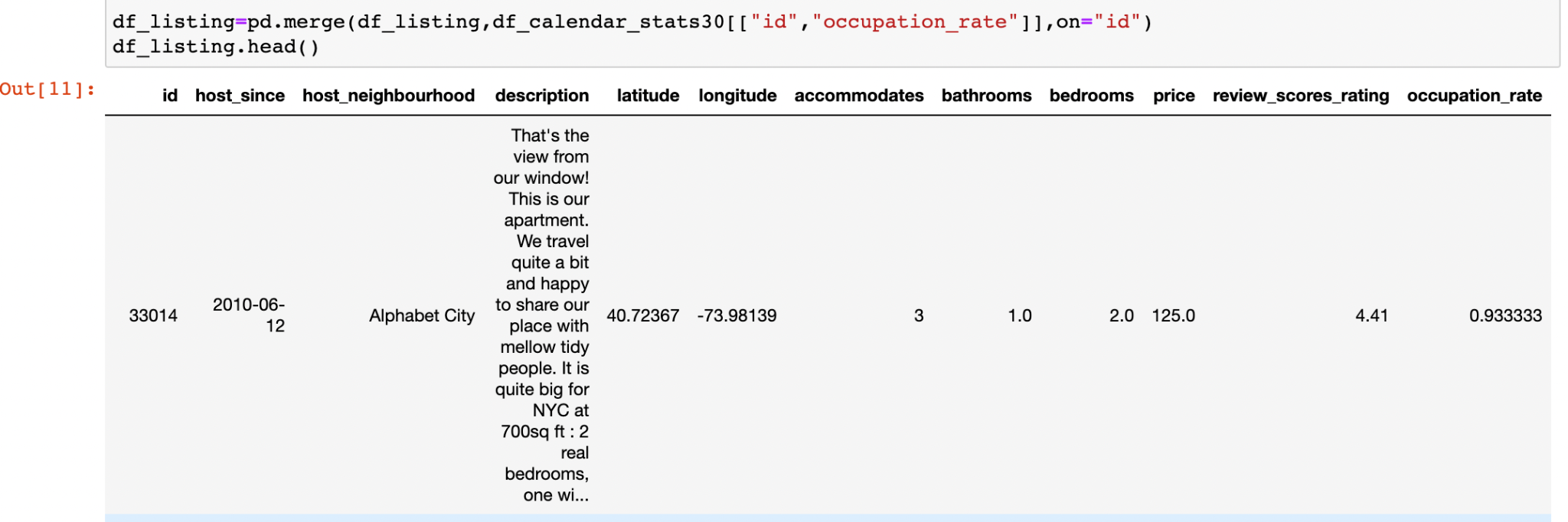
Calendar table:



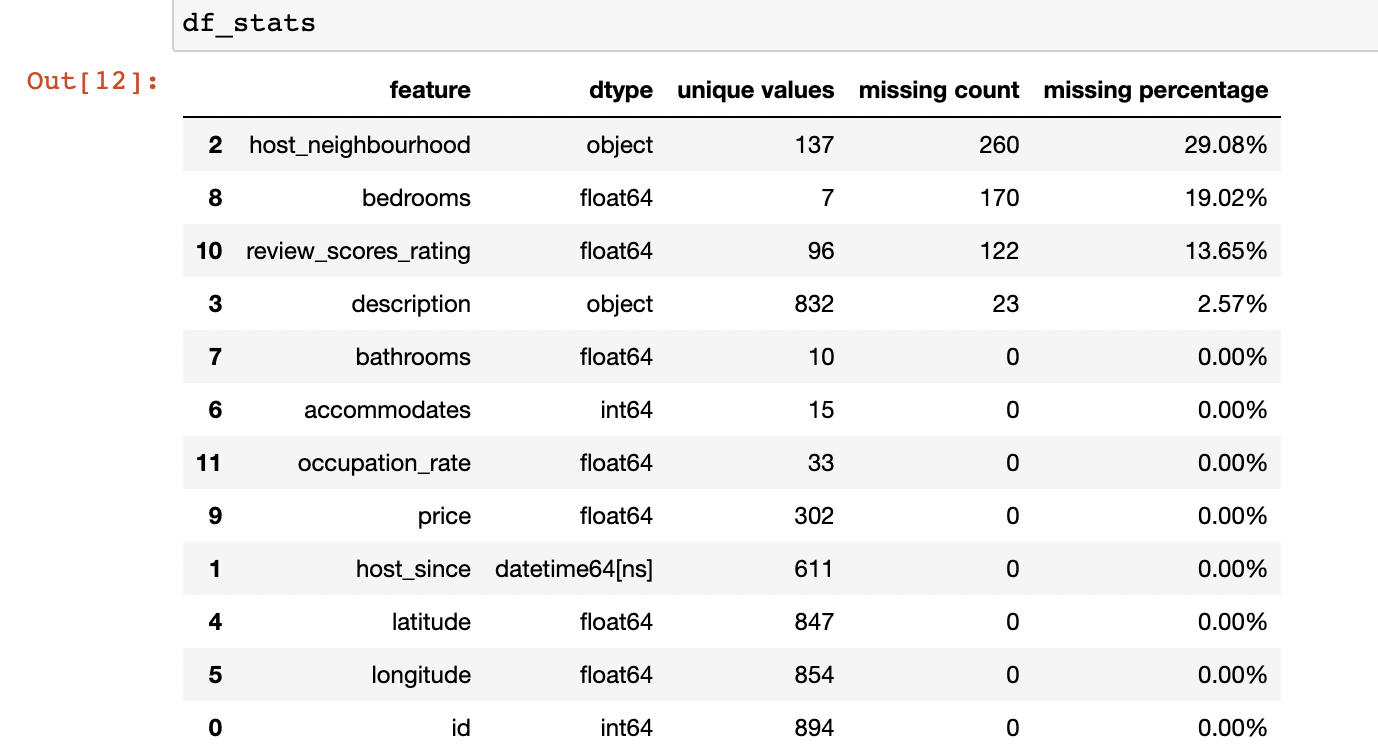
Data preparation is tough work, especially when the data comes from the real business. To begin with, we chose the entire rental unit property as the target. Not only because the number of such properties is the largest 16014, but also because the relationship between price and demand is more obvious if limited to only one type of property. We selected the data with host\_since before 2021-01-01, since the listing date is beginning from 2021-12-04. If we choose the host\_since before 2021-01-01, it represents that the duration of listing exceeds 1 year and has the representativeness. We dropped so many useless variables and kept some basic variables, such as host id, host\_since, neighborhood, description, latitude, longitude, accommodates, bathrooms, price, and rating scores.



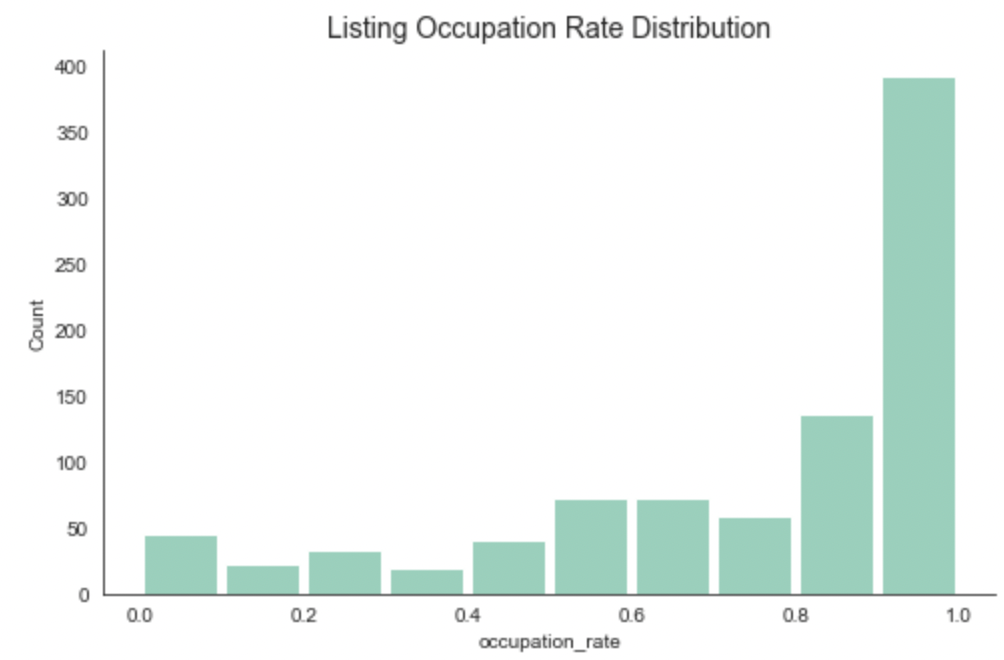
Prepare the demand data from the calendar table. The value f in the available columns represents not available for renting and value t stands for the listing is available for rent. We computed the demand, and occupation rate, in a month after 2021-12-04. Occupation rate = not available days / (not available days + available days). Then merge the demand data and previous listing data. At last, we checked the missing values, unique values count, and data type.



Missing Values



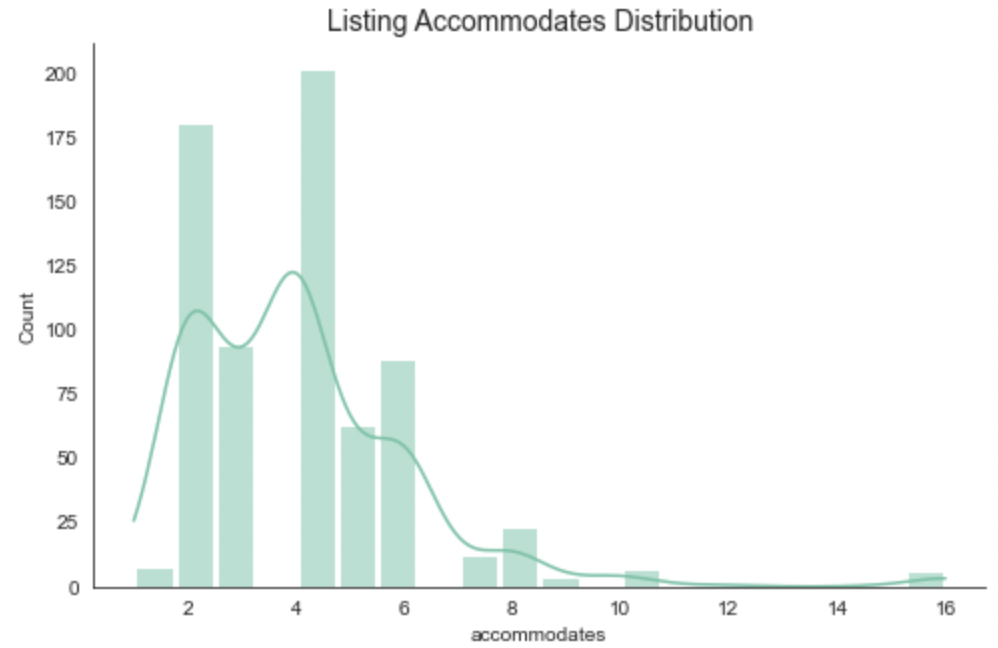
Subsequently conducted the descriptive analysis of the key variables. The distribution of occupation rates shows that most occupation rates exceed 600, accounting for nearly 67%. For the listing price distribution, most prices accumulate from $0 to $500. The mode of price is around $180. But there are a few extremely high prices that can be identified as the outlier. Utilized IQR standard principles to find the lower and upper bound, then deleted all the outliers. The range of review\_score\_rating is from 4.0 to 5.0. The whole distribution is skewed. The accommodates follows an approximately bimodal distribution. Most listings can accommodate 2 and 4 customers. The vast majority of bedrooms are under 3. One bedroom up 61.33%. Two bedrooms occupy 31.83%.



图表, 直方图

描述已自动生成

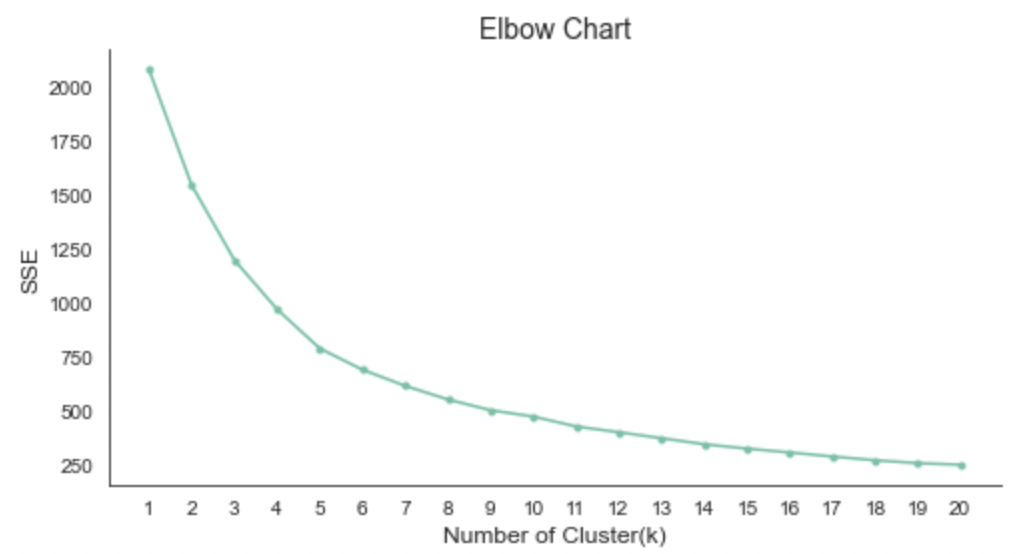
图表, 旭日形

描述已自动生成

**Cluster Analysis**

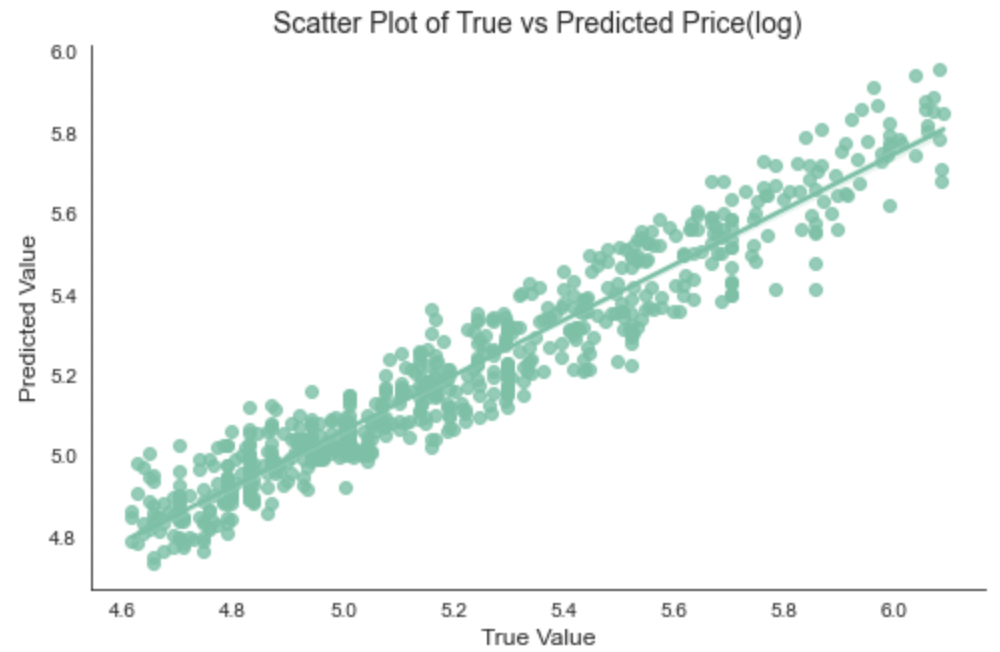
Clustering is the task to divide the data into several groups based on the data’s attributes. The classification method we utilized is K-Means. The principle of K-means clustering is to randomly select K objects, and the instance is more likely to fall into the same category as the point close to it. The center point of each class is taken as a new object and iterated continuously until the classification is completed. We made the assumption that similar listings are applied to the same demand function. Three features latitude, longitude, and accommodates were used in the cluster analysis. Latitude and longitude stand for the geological location. Accommodate is used to evaluate the size of the house. Generally speaking, the larger size house, the more travelers can be accommodated.

Before the cluster analysis, standardization of all columns is the first step. The reason why we take the standardization before is that the difference in measurement units will produce very different clustering results. Search the optimal number of clusters by elbow chart. When the number of clustering reaches 5 and the value of K increases, SSE decreases and becomes flat. We determined that the number of clusters equals 5. The numbers of clusters1 to 5 are 207, 263, 126, 58, and 38.



**Price Prediction model**

We chose the latitude, longitude, accommodates, bathroom, bedrooms, review score, host neighborhood, and description these features to predict the price via the training random forest model. To begin with, we needed to fill in the missing values and take the log transformation of the price because the original price follows the skewed distribution. For the bedrooms variable, the missing values are replaced with the mode. The mean of review scores took place of the missing values. Based on the features we selected, used on-hot encoding to code the characteristic of categories. With regard to description columns, tfidf was used to process the text contents. After setting up the model, we used GridSearchCV, a built-in class of sklearn, to adjust the hyperparameters to find the optimal values. The r squared of this price prediction model is 0.8620. In order to test the accuracy of the prediction roughly, he predicted and true values are basically linear, indicating that the model prediction is not bad.

.

**Demand Estimation**

Before building the model, plot the scatter plot of price and occupation rate, review score rating and occupation rate in each cluster. Only if there is a negative correlation, the maximum value can be found optimally. Every cluster has its own demand function through OLS estimation.

**D=**

**i=the number of clusters**

| Cluster | Intercept | (ln(price)) | (review rating) | (bathrooms) | (bedrooms) |
| --- | --- | --- | --- | --- | --- |
| 1 | 1.569 | -0.184 | 0.36 | -0.016 | -0.007 |
| 2 | 1.697 | -0.251 | 0.074 | -0.012 | 0.014 |
| 3 | 1.293 | -0.201 | 0.093 | -0.058 | 0.09 |
| 4 | 2.556 | -0.474 | 0.07 | -0.044 | 0.149 |
| Å5 | 1.173 | -0.452 | 0.31 | 0.143 | 0.037 |

**Revenue Optimization**

The optimization was solved by the scipy.optimize package, designed for minimizing the objective function. But for our question, we need to solve for maximizing the profit, so we need to multiply the original objective function by -1. In this case, the minimum value solved by scipy.optimize implies the maximum revenue of the previous objective function.

The adjusted objective function = -1\* (

)\*log(price)

Before processing optimization, we also defined some constraints and limit the range of variables’ values. Constraint1: The demand (occupation rate) should be larger or equal to 0. Constraint 2: The demand (occupation rate) should be smaller or equal to 1. Constraint 3: price \* demand should be larger or equal to 0. The values range for review\_score\_rating, bathrooms, and bedrooms should be within (1, 5), (0, 5), and (0, 10). After the optimization is terminated successfully, return the occupation rate and optimized price (after transformation). Saved the results in the dataframe for later use conveniently.

It is weird that cluster 3, 4, and 5 have the same optimized price and revenue. We came up with several reasons why this is the case maybe that the demand function is not accurate enough. There are many factors that affect demand, and our model does not include all significant factors, such as amenities near the listing, accessibility, and other factors that are not easily quantifiable. Before building the model, we plotted the scatterplot between DV and IDV, the linear relationship is not very significant and obvious. It is also possible that the model is not the best model and should perhaps consider polynomials, cross terms, etc. From our perspective, it is probably not accurate enough to only depend on the cluster model to find the optimal price. If it is possible, we can dive deeper into the machine learning in researching the effects of different variables.

The another reason is that the amount of demand data is not large enough. Maybe we can compute the demand in the next 3 months, but it is more likely to add the seasonal factors.

Optimization algorithm: we only chose scipy.optimize.minimize as our solving tool, if we want to improve the accuracy of the optimization result, we will try different algorithm in optimization, such as SGD, etc, to improve the accuracy of optimization results. These are the improvements we can tried to make in the future.

Consequently, the optimal price can be determined by optimal price of the corresponding cluster and prediction of the price. Each way has the same weight in calculating the final optimal price.

| Cluster | Occupation ratio | Optimized price | Revenue |
| --- | --- | --- | --- |
| 1 | 87.45% | $116 | $3,043.26 |
| 2 | 100.00% | $123 | $3,690.00 |
| 3 | 100.00% | $403 | $12,090.00 |
| 4 | 100.00% | $403 | $12,090.00 |
| 5 | 100.00% | $403 | $12,090.00 |

**Conclusion of Findings**

When we want to help the new listing to set the price, we should identify the cluster which the listing belongs to. Then find the optimal price of the corresponding cluster. And next, use the price prediction model to forecast the price. Combine the above two methods to give the final price. Assume that each method has a weight of 50%. We choose a sample listing from the dataset to identify the cluster by K-Means. After the process, this sample belongs to cluster 4. Looking up the optimal price and demand of cluster 4 that equal 403 and 100%. Then use the features of this listing in the price prediction model and get the ln(price) = 5.097. The price is exp(5.097) = $163.53. The final price is 214 by combining the two ways and the estimated maximum revenue for next month is $6420. We suggest the listing consider the price we computed already and keep the high occupation rate in the next month and get the estimated maximum revenue.

**Appendix**

Operation instructions (how to execute the program):

Our program only has two large tables and one Jupyter notebook file. If you don’t have Jupyter notebook, you can download the anaconda and launch the Jupyter notebook inside Anaconda. Then click the airbnb listing price optimize with ipynb extension on the main page. Before running the code, please change the path of the file when you need to read the data. After settling down, click the run button, Jupyter will automatically help you run the code in the cell. Jupyter Notebook is an easy tool for anyone.