CS274 ML Project 1 Part 1 Report

Group Members: Xinyue Liang, Peijun Qing

Kaggle Name: lxyskywalker, Kris55

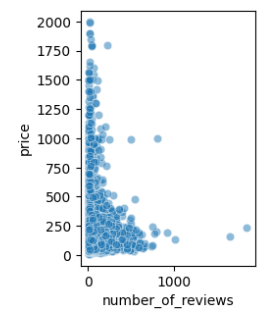
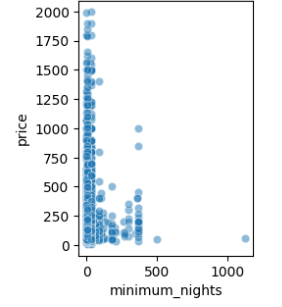
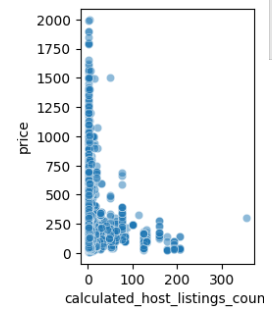
CS274 23Fall Section 2A

Professor: Sarah Masud Preum

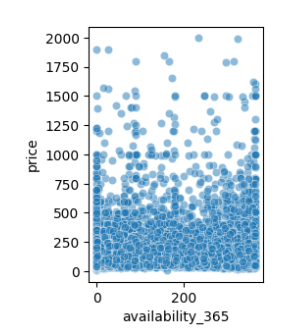
Word Count: 644

Part 1 Data:

The dataset we receive contains includes a training set and a testing set. Based on our observation, there are no missing datapoints in the given sets. Thus saved us the effort of filling out incomplete data (which is a situation that might be encountered in other projects.) The target variable is the price of the room. We presume that “neighbourhood group, location (the combination of latitude and longitude), room type, minimum\_nights, review information (except last review), host listings count and availability” could be useful features.

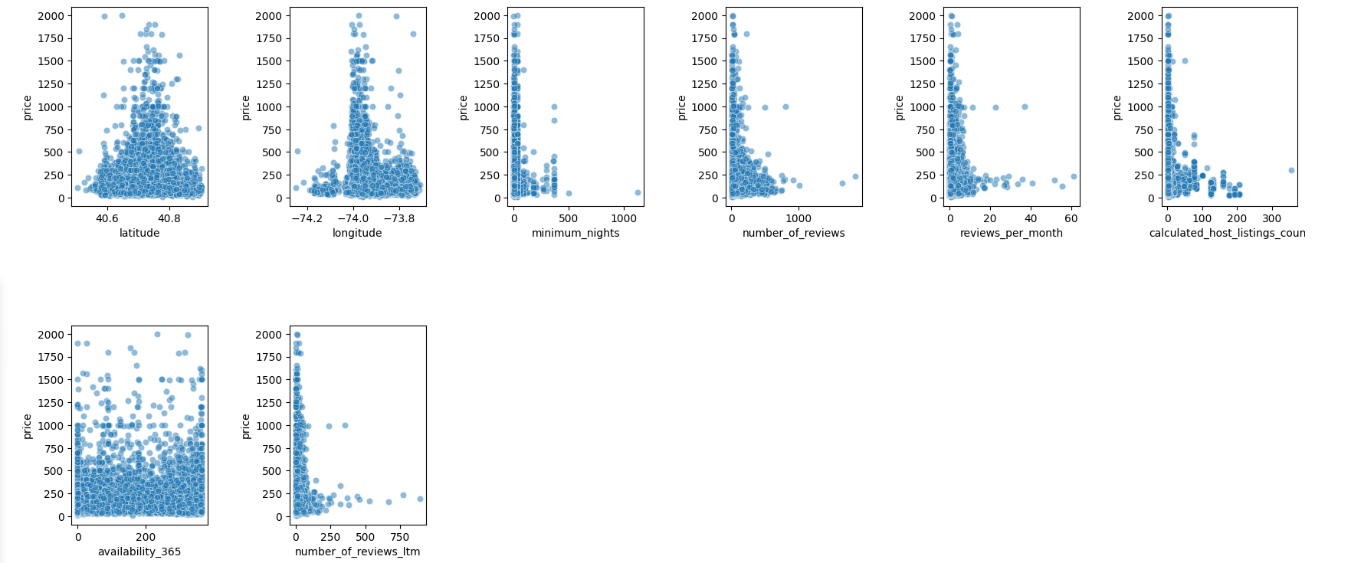


We visualized all the datapoints in relation to price. For instance, we can clearly see that the above 3 factors have a linear relationship with price. Calculated\_host\_listings\_count is a feature that indicates how many listings were under a host’s name. Intuitively we think that a host with multiple listings are more likely to have ways to cut cost and provide rooms with lower price. As for the minimum nights feature, which is a feature that indicates the shortest span customers need to rent the room. We suppose that the lower the minimum night requirement, the higher the cost because of the higher accessibility and frequency and cost of cleaning the room. And the number of reviews could be telling the number of people that rented the room.

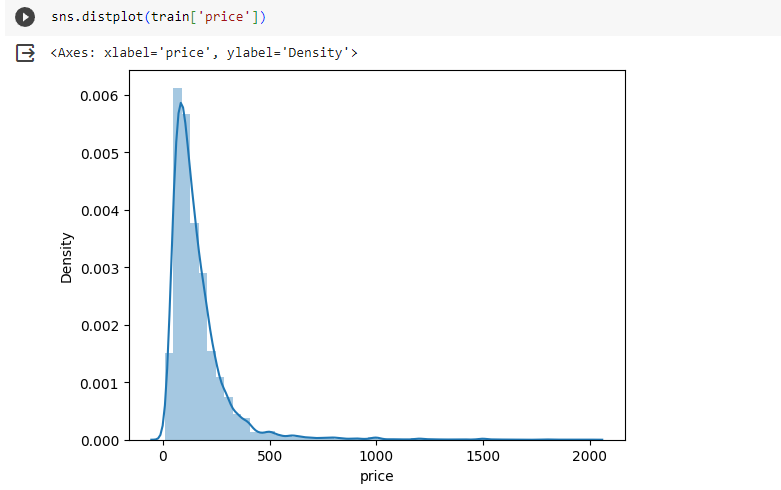


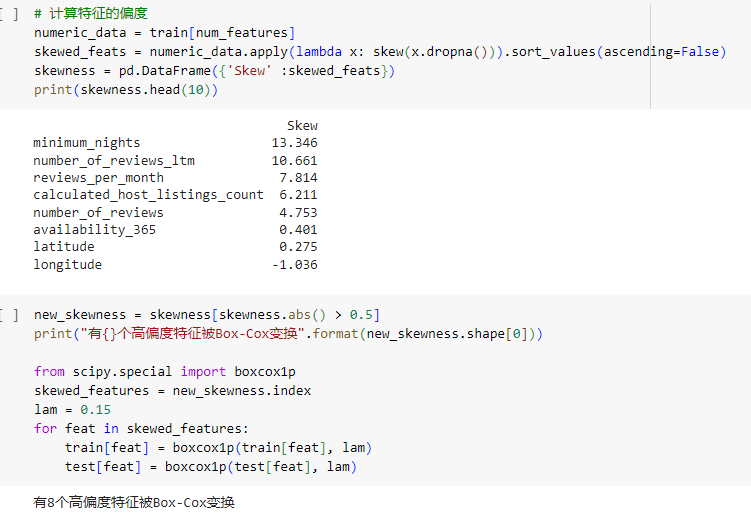
Similarly, the availability 365 also shows a linear relationship with price. It indicates the days of availability among a year. Add similar to the reasoning of minimum nights feature, higher the availability, higher the price.

Part 2 Methods:

After getting the basic idea of the datasets, we firstly dropped off the less irrelevant features: “'id','name','host\_name','neighbourhood','host\_id','license','last\_review'” Then we categorize all the features into numerical or string features. We ploted out all the numerical ones to further validate our hypothesis.

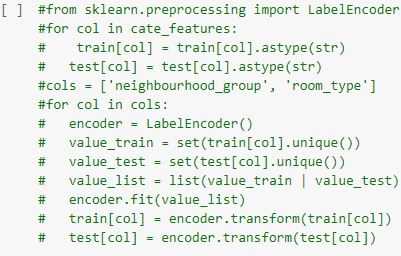
In the process we also have a few failed attempts on: removing the outliner datapoints (but the datapoints we removed were inter related), removing some of the number of comments features thinking that they might be duplicate with each other. However, all these attempts were negatively influencing the performance of our model, so we removed them.

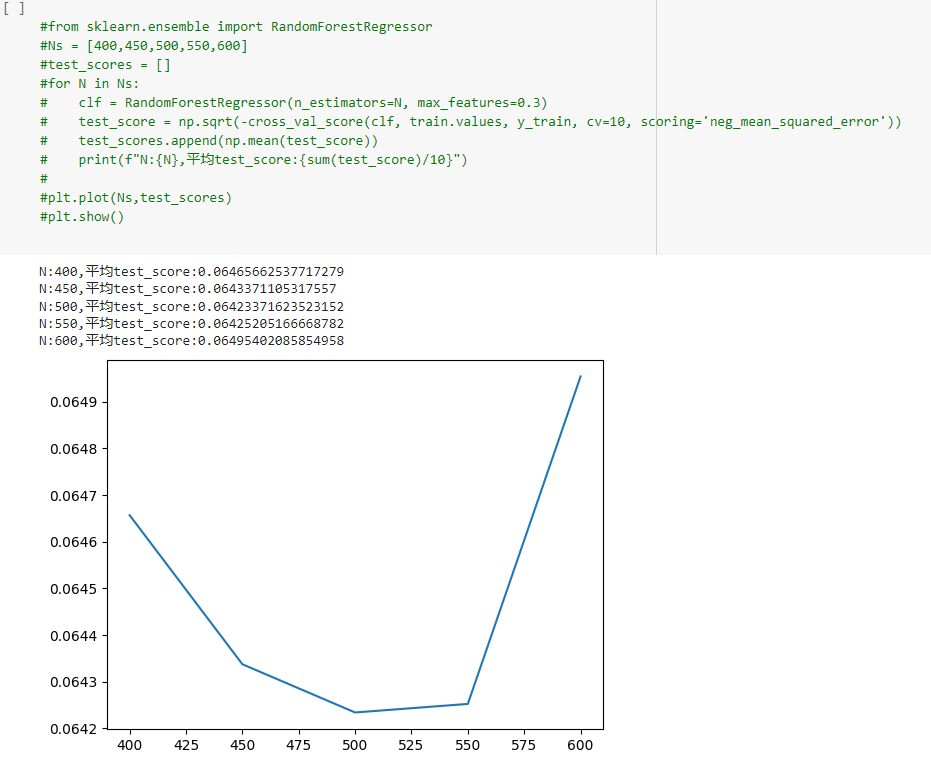
We also plotted out the graph of our target feature, price, to see the distribution of it.

The final step of handling the datasets was smoothen the skewness of the data. We first calculated the skewness then manage to achieve 0.15 as the following steps: 

The model we finally use is random forest. As the decision trees can only take numerical features, we converted the string ones using onehot encoding.

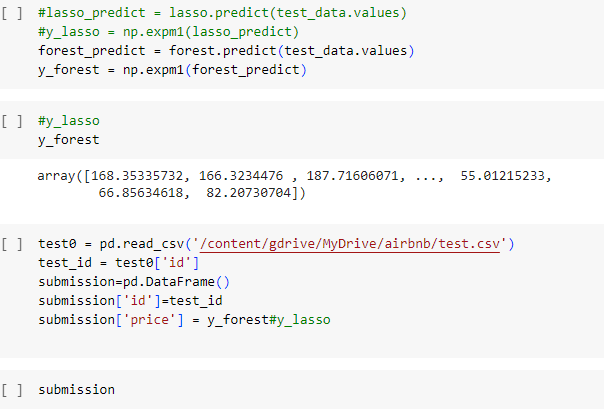


We also tried label encoder, but that method is slightly less effective

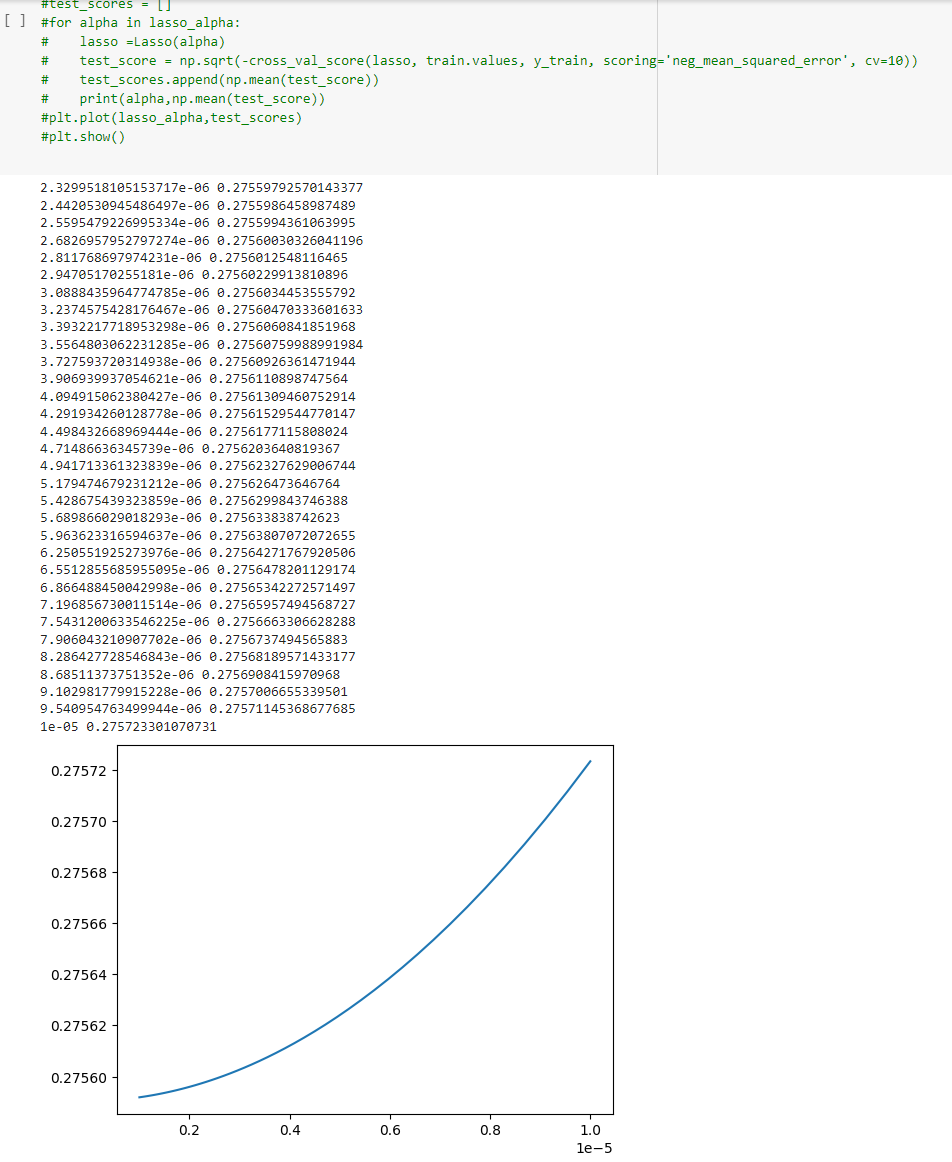
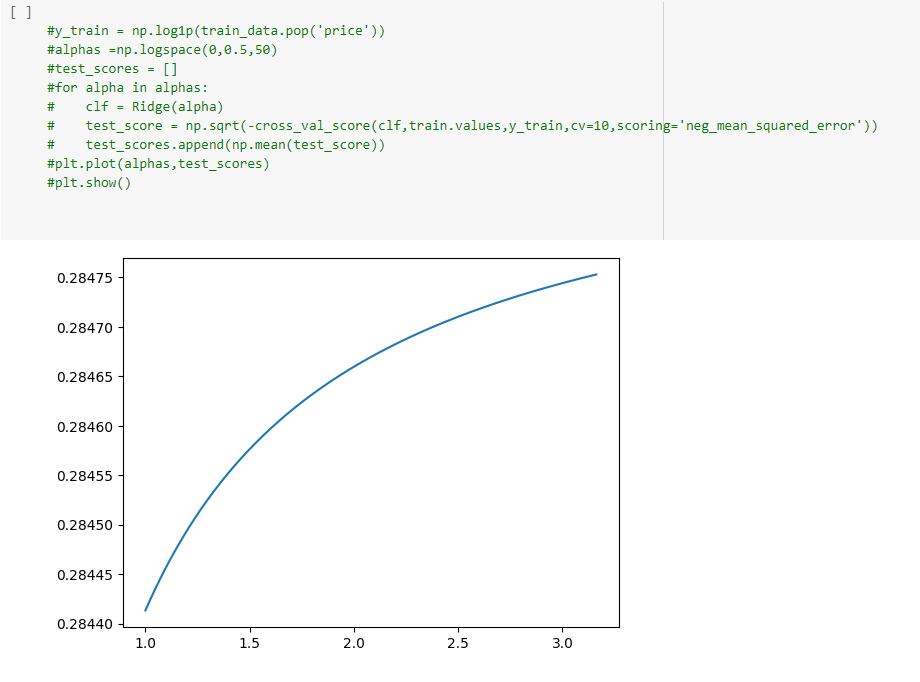
.

Then we implemented the random forest training with two hyper-parameters:, n\_estimators which is the number of trees in the forest, and it’s the primary hayer parameter we need to tune. The max\_features can be used to validate the best out comes of different tree numbers. Form the graph above we get the lowest cost with 500 trees.

Finally, we implemented the random forest with 500 trees.



One other models we tried but have worse performance are ridge and lasso, which are both transformations of standard linear regression. Here are the efforts we made to find the alpha of these two models:



Part 3 Results

The best Kaggle score we achieved using random forest is 110.98. The score we got by using lasso is 124.02 and we did not submit ridge regression result as we did not actually completed it.

As for the future works, we identified that there are several possible features that could be extracted from the name such as how many bedrooms, if the room is aimed for luxury/designer market and if the room is a apartment or house. We could construct many more features from these information. On the other hand, we might need to introduce things such as hot maps to deal with the longitude and latitude features.