AIRBNB PRICE PREDICTION

Rachana Thota (G01237600), Srujan Reddy Tekula (G01240653)

December 9, 2020

Abstract

Airbnb has become increasingly popular among travellers for accommodation across the world. Accordingly, there are large datasets being collected from the Airbnb listings with rich features. In this project, we aim to predict Airbnb listing price in major cities – New York City (NYC), San Francisco, Washington DC, Los Angeles, Chicago and Boston with various machine learning approaches. In addition to it, we aim to apply these ML approaches for these six cities individually and compare if they are consistent with the combined set. With one of our best approach – Regressor models, we have achieved RMSE score values of 120.38538 with XG boost and R^2 score of 0.597 with ridge regression method on the train data set and RMSE score of 20.96259 in comparison of NYC test dataset and combined test data set. We find that regressor models that is trained on the combined dataset of all six cities, rather than the individual datasets, is more generalizable to predict the price in a different city. With better price suggestion estimates, Airbnb home providers can reach an equilibrium price that optimizes profit and affordability. The objective of this project is to build a model that predicts the optimal price of a property taking into account different listing features. The end goal is that users can know what features of an Airbnb listing are most important as well as how prices should be fluctuating based on amenities, property type etc.

1 Introduction

Unlike hotels, which have their own pricing system, Airbnb prices are usually determined by the hosts empirically. It poses challenges for the new hosts, as well as for existing hosts with new listings, to determine the prices reasonably high yet without losing popularity. On the consumers' side, though they can compare the price across other similar listings, it is still valuable for them to know whether the current price is worthy and if it is a good time to book the rooms. This project is used to determine and predict the price of Airbnb listings in major US cities like Boston, Chicago, Washington DC, Los Angeles, New York, San Francisco.

The price for Airbnb renting depends on multiple factors, and we divide the input type into 3 categories, including continuous, categorical, set (amenities) features. We have extracted more than 60 features from the dataset. Here we only list a few of them that are both representative and important for the task, such as room size {accommodates, bathrooms, bedrooms, beds, ...}, location {neighbourhood, latitude, longitude, ...}, facilities {amenities, property type, ...}, and booking related {availability, cancellation policy, host response rate, ...}. The ground-truth label is the actual base price, and we use a variety of regression approaches including linear regression, ridge regression, lasso regression, as well as XG boost, to predict the value.

2 Problem Statement

Airbnb is a home-sharing platform that allows homeowners and renters ('hosts') to put their properties ('listings') online, so that guests can pay to stay in them. Hosts are expected to set their own prices for their listings. Although Airbnb and other sites provide some general guidance, there are currently no free and accurate services which help hosts price their properties using a wide range of data points. Airbnb pricing is important to get right particularly in big cities like New York where there is lots of competition and even small differences in prices can make a big difference. It is also difficult thing to do correctly – price too high and no one will book. Price too low and you will be missing out on a lot of potential income.

This project aims to solve this problem, by using regressor models to predict the price for properties across major cities in the United States. By examining a large data set of past home rentals and finding patterns and statistical relationships between house's characteristics and its price including patterns that might not be evident to a human who's looking at the data. We have explored the preparation and cleaning of Airbnb data and conducted some exploratory data analysis. With this the users can know what features of an Airbnb listing are most important as well as how prices are fluctuating based on amenities, location etc.

2.1 Notations

RMSE : Root Mean Square Error
R^2 : R-Squared score
LA : Los Angeles
DC : Washington DC
SF : San Francisco
NYC : New York City

3 Literature Review

As far as we are aware from our literature search, there are no published studies that apply data mining techniques to the similar data on these six major cities in US, combinedly and separately and compare those two. This projects data has largely been used for visualizations and analysis of listing types in each city.

Airbnb price prediction becomes popular due to the availability of large datasets in many cities. In 2015, Li et al [1] used Multi-Scale Affinity Propagation for price recommendation and show that it largely improves the precision of the reasonable price prediction. In 2017, Wang et al [2] worked-on Airbnb datasets from 33 cities and identified the 25 price determinants from a sample of 180,533 accommodation rental offers using ordinary least squares and quantile regression analysis. A similar work by Teubner et al [3] extracts reputation-related features and investigate its effect on pricing with linear regression. In 2019, Kalehbasti et al [4] used multiple machine learning approaches and sentiment analysis on predicting Airbnb price in NYC dataset, and they achieved 0.6901 R^2 value on the test dataset.

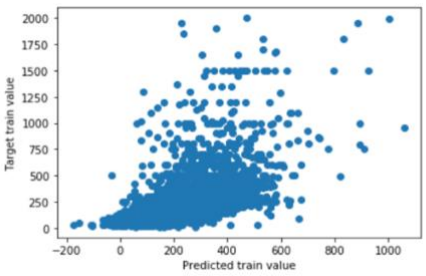
Apart from the published studies, there were few competitions held on Kaggle for a similar data set, among which most of them have used the data to visualize and analyse the data based on different features available. Few have developed models to predict the price on all the major cities in US combinedly. Also, among them, most of the models are developed in R language.

As part of our project, we are applying various regression models to predict the price for properties in major US cities combinedly and separately and compare the results. We developed our project using Python language.

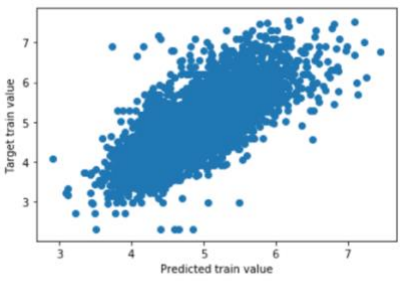
4 Methods and Techniques

As the target variable to be predicted is the price, which is a continuous value, we can apply regression models like linear regression, ridge regression etc., but unable to apply classification models like k-nearest neighbors, svc etc.,

For the results in 5.3 section, we used the default settings for all machine learning approaches imported from the sklearn package. Note that we have experimented with different parameters, like learning rate, n-estimators in XG Boost., and the output metrics stay largely unchanged.



Scatter plot of actual vs predicted value as integer



Scatter plot of actual vs predicted value as log

We have applied these machine learning approaches with different label transformations, with the target variable as an integer and as logarithmic value and we found that logarithmic transformation largely improves the model performance

Building a Machine Learning Model

We experimented with machine learning models for price prediction. As this is the regression task, the evaluation metric chosen was mean squared error (MSE). For accuracy, we have calculated r squared value for each model produced.

Linear regressor: We have used this model to define the relation between features and target variable through an equation that tries to represent the relationship between one dependent and multiple independent variables.

Ridge regressor: As per the evaluation metrics, there is a slight improvement in this model because the value of the R-squared has been increased. This reduced the model complexity by coefficient shrinkage because it uses L2 regularization technique.

Lasso regressor: This model is quite similar to ridge, but after evaluation metrics we have observed that both the RMSE and R-square for this model has increased. Therefore, lasso model is predicting better than both linear and ridge. This model is generally used when there are more number of features as it automatically does feature selection.

XG Boost: XG Boost is a powerful approach for building supervised regression models. It contains loss function and a regularization term. It tells about the difference between actual values and predicted values, i.e., how far the model results are from the real values. Compared to other regressor model, this gave best metrics.

5 Discussion and Results

5.1 Datasets

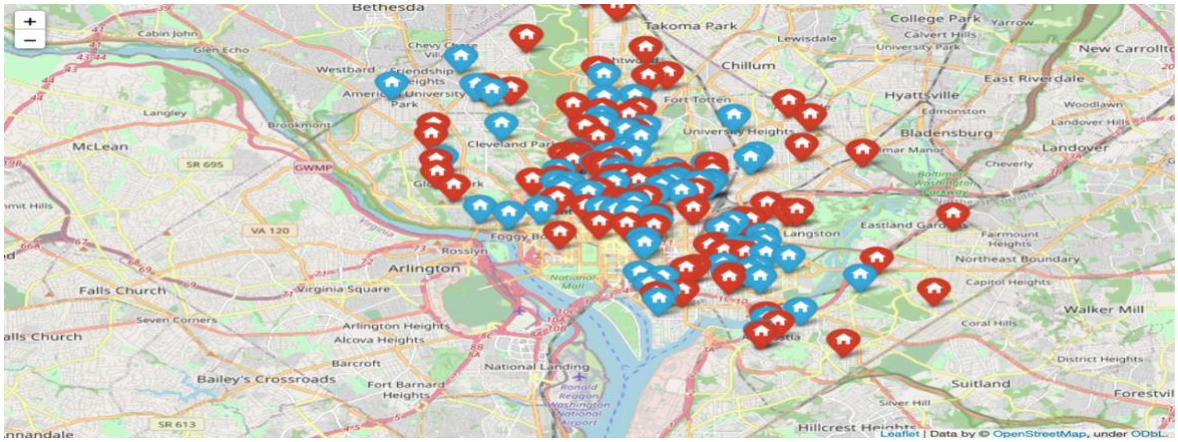
- The data for this project is taken from Kaggle and is also available in http://insideairbnb.com.
- Each row in a data set is a listing available for rental in Airbnb's site for the specific city.
- The columns describe different characteristics of each features. Some of the important features are: amenities, location, city, property type, number of bedrooms and bathrooms.
- A sample data is as follows:

id	property_typ	room_type	amenities	accommodat	bathrooms	bed_type	cancellation_	cleaning_fee	city	description	first_review	host_has_pr	host_identity
3895911	Apartment	Private room	{TV,"Cable T	2	1	Real Bed	flexible	TRUE	LA	Close to SM	23/10/16	t	f
9710289	Apartment	Entire home/	{TV,"Cable T	3	1	Real Bed	moderate	TRUE	NYC	This apartme	12/09/16	t	t
9051635	Apartment	Private room	{"Wireless Ir	1	1	Real Bed	moderate	TRUE	SF	Spacious 1 b	13/11/16	t	t
708374	Apartment	Entire home/	{TV,"Cable T	1	1	Real Bed	strict	TRUE	LA	Very clean 1	01/11/15	t	t
626296	Apartment	Entire home/	{TV,Internet,	2	1	Real Bed	flexible	TRUE	NYC	My apartment is airy & lig		t	t
3309829	Townhouse	Private room	{TV,Internet,	3	2	Real Bed	moderate	TRUE	LA	One private r	14/02/15	t	t

host_respons	host_since	instant_book	last_review	latitude	longitude	name	neighbourho	number_of_r	review_score	thumbnail_u	zipcode	bedrooms	beds
100%	13/08/16	f	26/02/17	34.0283724	-118.49445	Santa Monic	Santa Monic	6	97	https://a0.m	90403	1	1
100%	04/12/13	f	16/10/16	40.7203801	-73.942329	Bright, charn	Williamsburg	2	80	https://a0.m	11222	1	1
100%	02/08/11	f	17/11/16	37.785434	-122.47028	Private room	Richmond Di	2	100	https://a0.m	94118	1	1
100%	27/06/12	f	01/03/17	33.9760264	-118.46347	Marina del R	Marina Del R	7	94	https://a0.m	90292	0	1
	12/01/16	f		40.7355734	-74.005996	Bright Studic	West Village	0		https://a0.m	10014	1	1
100%	14/07/12	f	06/11/16	34.0848346	-118.09719	Private Room +Queen size		8	98	https://a0.m	91776	1	1

- The number of records in training data set is 74112 and test data set is 25459.
- The number of columns in training data set is 29 and test data set is 28 (excluding target variable price)
- The number of listings per individual cities in train and test data is as follows,

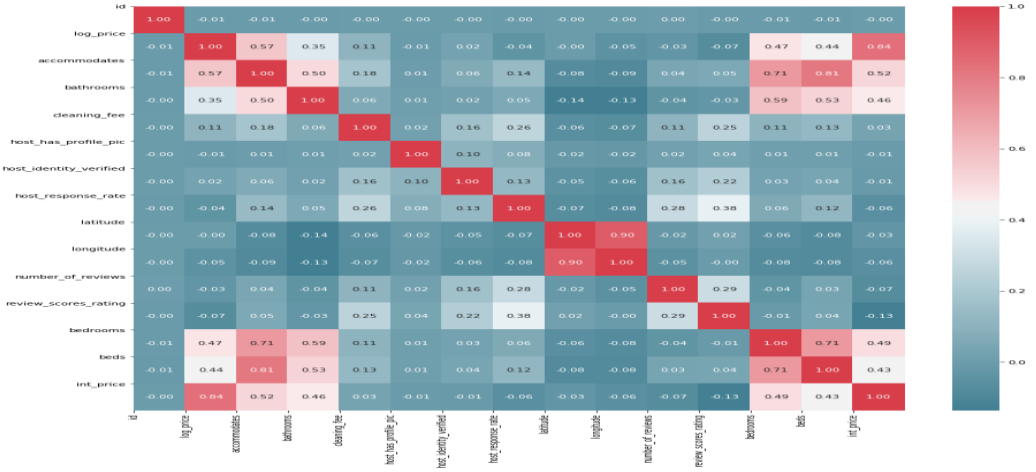
- The distribution of train and test data in certain area in the city of DC as an example is as follows, with train data points in red and test data points in blue colour.



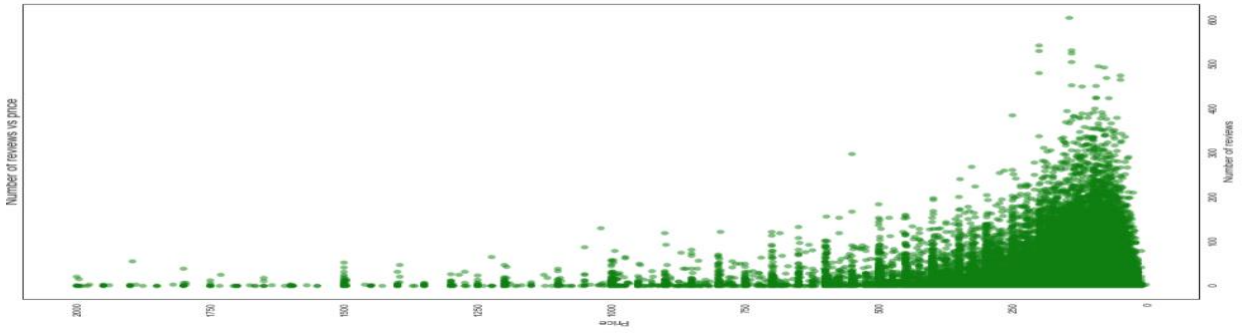
Data Cleaning: Missing values in both train and test data are found and filled them with appropriate values.

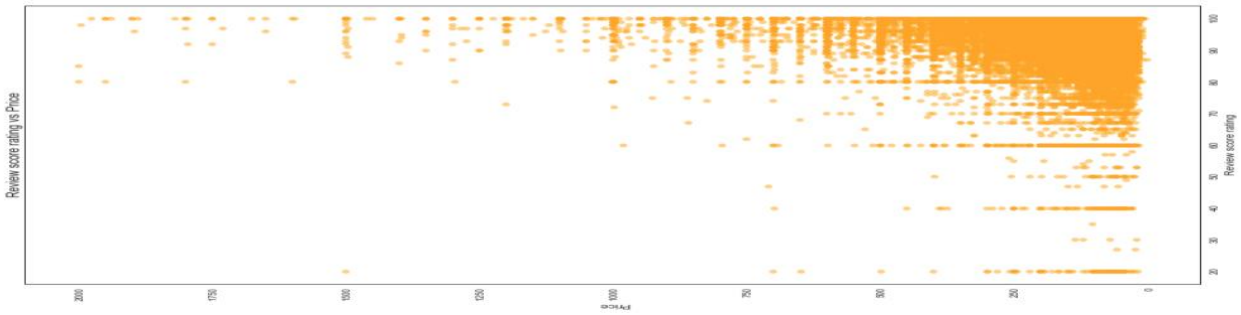
Feature Selection:

- Some of the features, as below, which doesn't show any significant differences in the price prediction are eliminated.
  - name
  - thumbnail_url
  - description
  - host_since
  - first_review
  - last_review
  - instant_bookable
- Based on the low correlation values between the numerical features and the log price (target variable), below features are eliminated/dropped.
  - id
  - host_has_profile_pic
  - host_identity_verified
  - host_response_rate
  - latitude
  - longitude



- Though number_of__reviews and review_score_rating have low correlation values, they are not dropped as they have significant impact in price as shown in below plot.





Feature Engineering

The dataset is classified into three sets of data: categorical, continuous and a set of categorical values.

Categorical Data: Features like, property_type, room_type, bed_type, city, cancellation_policy are converted into numeric format by one-hot encoding using pd.get_dummies().

Set of Categorical Data: Only one feature, amenities, comes under this category where a bunch of categorical values are arranged as a set, which are split into a numpy array and concatenated to the data set in binary format.

Continuous Data: As the continuous data is in already in numeric format, no changes are applied on this data. The features like, bathrooms, accommodates etc., comes under continuous data type.

- This above data is converted into vectorized form using PCA – principal component analysis.

5.2 Evaluation Metrics

We have used RMSE and R^2 metrics in this project to evaluate the training set of combined train data and to compare the difference in the predicted price value of the test data between the combined test data set and individual cities, used RMSE.

5.3 Experimental Results

We have experimented our data with different combinations of features and labels as follows,

1. *Feature: Continuous, Label: Int Price and Log Price* – Applied different models by using only the continuous features of the data set to predict int price and log price. Below is the table with RMSE and R^2 metric values on train data set.

Model	Feature	Label	RMSE	R^2
Linear Regression	C	Int price	132.84386	0.37916
Ridge Regression	C	Int price	132.84382	0.37916
Lasso Regression	C	Int price	132.84386	0.37916
XG Boost	C	Int price	122.32714	0.47380
Linear Regression	C	Log price	0.55951	0.39153
Ridge Regression	C	Log price	0.55951	0.39153
Lasso Regression	C	Log price	0.55951	0.39152
XG Boost	C	Log price	0.52189	0.47066

C – Continuous

We find that there isn't much difference in prediction between linear, ridge and lasso when only continuous features are used and a significant difference in XG boost model compared with other three.

Below table shows the RMSE scores of individual cities b/w the predicted price of individual test data and combined test data.

Model	Feature	Label	City	RMSE
Lasso Regression	C	Int price	NYC	17.58
			DC	79.01
			SF	39.07
			LA	20.64

			Boston	33.80
			Chicago	27.51
Lasso Regression	C	Log price	NYC	0.07675
			DC	0.22390
			SF	0.06753
			LA	0.09171
			Boston	0.09482
			Chicago	0.08989
XG Boost	C	Int price	NYC	22.98217
			DC	68.18182
			SF	44.42859
			LA	33.24297
			Boston	45.07194
			Chicago	38.25363
XG Boost	C	Log price	NYC	0.08573
			DC	0.16223
			SF	0.14307
			LA	0.10035
			Boston	0.15437
			Chicago	0.14181

C - Continuous

Though the RMSE score of DC is relatively less in XG boost model, it is higher for other cities compared to Lasso regression.

2. ***Feature: Continuous and Categorical, Label: Int Price and Log Price*** **–** Applied different models by using only the continuous and categorical features of the data set to predict int price and log price.

Below is the table with RMSE and R^2 metric values on train data set.

Model	Feature	Label	RMSE	R^2
Linear Regression	C + O	Int price	127.20432	0.429983
Ridge Regression	C + O	Int price	127.28782	0.43012
Lasso Regression	C + O	Int price	127.29901	0.43002
XG Boost	C + O	Int price	**120.38538**	0.49015
Linear Regression	C + O	Log price	0.47348	0.56572
Ridge Regression	C + O	Log price	0.47339	0.56444
Lasso Regression	C + O	Log price	0.47352	0.56420
XG Boost	C + O	Log price	0.47548	0.56064

C – Continuous
O – Categorical with One-hot encoding

We find that the RMSE score of XG Boost on continuous and categorical features to predict int price is the least score in all of our experimental results.

Below table shows the RMSE scores of individual cities b/w the predicted price of individual test data and combined test data.

Model	Feature	Label	City	RMSE
Lasso Regression	C + O	Int price	NYC	20.96259
			DC	83.82553
			SF	42.42641
			LA	21.46555
			Boston	39.60631
			Chicago	35.22584
Lasso Regression	C + O	Log price	NYC	0.08871
			DC	0.21349
			SF	0.09121
			LA	0.07887
			Boston	0.09980
			Chicago	0.12598

XG Boost	C + O	Int price	NYC	31.42483
			DC	78.30696
			SF	65.79058
			LA	45.75817
			Boston	48.77458
			Chicago	46.51064
XG Boost	C + O	Log price	NYC	0.11753
			DC	0.23444
			SF	0.16775
			LA	0.13820
			Boston	0.22293
			Chicago	0.19129

C – Continuous
O – Categorical with One-hot encoding

3. ***Feature: Continuous, Categorical and with Amenities, Label*****:** ***Int Price and Log Price*** **–** Applied different models on using the continuous, categorical and amenities features of the data set to predict int price and log price.

Below is the table with RMSE and R^2 metric values on train data set.

Model	Feature	Label	RMSE	R^2
Linear Regression	C + O + A	Int price	124.35031	0.45023
Ridge Regression	C + O + A	Int price	125.22260	0. 0.4485
Lasso Regression	C + O + A	Int price	125.23594	0.44843
XG Boost	C + O + A	Int price	121.16856	0.48376
Linear Regression	C + O + A	Log price	0.45640	0.59667
Ridge Regression	C + O + A	Log price	0.45532	**0.59707**
Lasso Regression	C + O + A	Log price	0.45552	0.59664
XG Boost	C + O + A	Log price	0.47860	0.55482

C – Continuous
O – Categorical with One-hot encoding
A – Amenities in binary format

We find that the R^2 score of Ridge regression on continuous, categorical and amenities features to predict int price is the highest score in all of our experimental results.

This table shows the RMSE scores of individual cities b/w the predicted price of individual test data and combined test data.

Model	Feature	Label	City	RMSE
Lasso Regression	C + O + A	Int price	NYC	22.00909
			DC	94.11743
			SF	49.71378
			LA	24.65948
			Boston	46.94188
			Chicago	47.81433
Lasso Regression	C + O + A	Log price	NYC	0.08960
			DC	0.23541
			SF	0.10593
			LA	0.09691
			Boston	0.12727
			Chicago	0.15368
XG Boost	C + O + A	Int price	NYC	38.05930
			DC	97.41006
			SF	71.97715
			LA	54.39210
			Boston	51.72756
			Chicago	59.61996
XG Boost	C + O + A	Log price	NYC	0.15310
			DC	0.28381

			SF	0.24399
			LA	0.17877
			Boston	0.27207
			Chicago	0.23277

C – Continuous
O – Categorical with One-hot encoding
A – Amenities in binary format

6 Conclusion

Overall, we have performed extensive feature selection and engineering, and experimented with various machine learning approaches in predicting Airbnb listing price. We showed that XG boost and lasso regression out-perform other approaches, and achieve r-squared value around 0.6. We also showed that by training the datasets from different cities individually.

- Linear regression worked correctly when only continuous values are included, with categorical values are included and K-fold is applied, few splits are seen with very higher values of RMSE and R^2 in the range of e+17.
- When tried to implement the regression models, the values were not consistent and are largely differed from each run.
- Unable to apply k-nearest or svc models on this data set as the target variable is a continuous value and can be applied on regression models. Though random forest model is working of this data set, it is taking a lot of time in order to compute the results and the results were comparatively similar to XG boost.

6.1 Directions for Future Work

- In future work, we would like to improve the performance with extra feature extraction and hyper-parameter tuning.
- Would like to work on the implementation of regression models for such a huge data set with such different data types
- Augment the model with natural language processing (NLP) of listing descriptions and/or reviews, e.g., for sentiment analysis or looking for keywords.
- In addition to predicting base prices, a sequence model could be created to calculate daily rates using data on seasonality and occupancy, which would allow the creation of actual pricing software.

7 References

[1] Yang Li, Quan Pan, Tao Yang, and Lantian Guo. Reasonable price recommendation on airbnb using multi-scale clustering. In 2016 35th Chinese Control Conference (CCC), pages 7038–7041. IEEE, 2016.

[2] Dan Wang and Juan L Nicolau. Price determinants of sharing economy-based accommodation rental: A study of listings from 33 cities on airbnb. com. International Journal of Hospitality Management, 62:120–131, 2017.

[3] Timm Teubner, Florian Hawlitschek, and David Dann. Price determinants on airbnb: How reputation pays off in the sharing economy. Journal of Self-Governance & Management Economics, 5(4), 2017.

[4] Pouya Rezazadeh Kalehbasti, Liubov Nikolenko, and Hoormazd Rezaei. Airbnb price prediction using machine learning and sentiment analysis. arXiv preprint arXiv:1907.12665, 2019.

[5] Kaggle. Airbnb price prediction, 2018.

[6] Price Prediction in the Sharing Economy: A Case Study with Airbnb data

[7] Emily Tang, Kunal Sangani Neighborhood and Price Prediction for San Francisco Airbnb Listings

[8] https://github.com/amayumradia/AirBnB-pricing-prediction/blob/master/Airbnb%20Data%20Exploration.ipynb

[9] https://github.com/Dima806/Airbnb_project/blob/master/airbnb_final_analysis_v3.ipynb

[10] https://airbnb-pricing-prediction.herokuapp.com/