

# **Pricing Model for Airbnb Properties**

Alvin Du, Gao Jie, Scott Yim, Alex Barry

## **Business Understanding / Case / Problem Definition**

As a new Airbnb host in New York City, there are many factors influencing how I would decide to price my property. From the attributes of the physical property itself like location and size, amenities like air conditioning and wifi, and whether I allow guests or add a cleaning fee, there are many considerations to take into account, especially if I want to maximize value while still remaining competitive amongst neighboring Airbnb properties.

The business problem we are addressing is helping Airbnb hosts understand the numerical value of their apartments (target), to directly inform how to strategically price their rental properties (action) based on the type of rental and included features / amenities. To do this, our team has tested multiple models that use historical Airbnb data to identify what features of a property affect price. Ultimately, the main question we are trying to answer is as follows: Is there a price that our data model can recommend to maximize value creation and profit?

While New York City is a good pilot location given the number of visitors and tourists who travel to the city annually, we see this as a universal and scalable problem across the world. Airbnb is a global company headquartered in San Francisco, California with 150 million users in over 65,000 cities and 2 million booking transactions per day<sup>1</sup>. With an average per night price of \$80, Airbnb uses a revenue share model where they take 3% in a service fee of the booking transaction amount<sup>2</sup>. If we can optimize prices for both Airbnb and the Airbnb host, there could be tremendous upside potential for everyone given the volume of transactions that occur on a daily basis.

---

<sup>1</sup> <https://muchneeded.com/airbnb-statistics/>

<sup>2</sup> <https://ipropertymanagement.com/research/airbnb-statistics>

Assuming that the predicted rental price is market equilibrium price, the hosts would maximize their profit under a recommendation we could provide to them, so the overall business transaction value would increase. This would add business value by educating hosts on the value of their property, and also allow Airbnb to collect a larger service fee amount from the percentage they collect from each transaction / stay. When a host enters the features of his/her house, our data model would output a suggested price, which would enable them to price strategically relative to the market's historical values based on what their property has to offer.

The data mining problem is to analyze historical Airbnb data across many features and identify patterns related to price, and to predict rental price for a property. By performing this supervised modeling analysis using real and recent Airbnb datasets, we see this data mining problem helping inform Airbnb and its hosts in two major ways. First, for those who own their properties and can actually make alterations, this information may help them in deciding to add amenities or renovate their apartments to drive the price up (e.g. adding air conditioning or washer / dryer, more beds, allowing pets, including a parking pass, etc.). Second, for those who may not own their properties, it is still helpful for them to know the value of their apartment in the market, and there may be adjustments they can make even as renters (e.g. adding a window air conditioning unit, including or removing the cleaning fee, allowing extra guests, etc.). Overall, if pricing is competitive and optimized, Airbnb can collect the highest percentage of each booking transaction. As a result, we believe this analysis and model will help drive revenue for Airbnb and enable New York City Airbnb hosts to set their properties at the most competitive prices.

## Data Understanding

We were able to construct a final dataset comprising two different datasets from Kaggle.com and InsideAirbnb.com. Because we had many features we wanted to examine with respect to New York City apartments, there wasn't a single dataset that encompassed all of the information we needed to perform a robust analysis, but luckily were able to find two fairly large datasets with a common unique identifier that we could merge on.

Our first dataset is from Kaggle.com and includes 2019 NYC Airbnb data ([link](#)). The dataset has 48,900 rows and 16 total columns of features, from which the informative attributes included location, price, property type, review, availability. The types of features and dimensionality of the possibly informative attributes (about 10 of the 16) were as follows:

1. Neighborhood (categorical, 5 options) = dimensionality of 4
2. Neighborhood (categorical, 221 options) = 220
3. Latitude (numeric) = 1
4. Longitude (numeric) = 1
5. Room type (categorical, 3 options) = 2
6. Price (numeric) = 1
7. Minimum nights (numeric) = 1
8. Number of reviews (numeric) = 1
9. Reviews per month (numeric) = 1
10. Availability 365 (numeric) = 1

This was an ideal dataset from the perspective of number of instances and baseline features, but our goal for the project was to also analyze *features and amenities of the apartment itself*, which

this first dataset didn't include. As a result, we did more research to find another supplementary dataset that we could combine with the dataset from Kaggle.com.

The second dataset that we identified is from InsideAirbnb.com ([link](#)) and includes 2019 NYC Airbnb features. This dataset has 48,300 rows and 106 columns of features, including number of beds, bathrooms, bedrooms, bed\_type, amenities, guests\_included, cleaning\_fee, extra\_people, and review\_scores\_rating, amongst others. There was some overlap between this dataset and the one from Kaggle.com, but some of the types of features and dimensionality of the possibly informative and new attributes of this dataset were as follows:

1. Number of beds (numeric) = 1
2. Number of bathrooms (numeric) = 1
3. Number of bedrooms (numeric) = 1
4. Bed type (categorical, 5 options) = 4
5. Guests included (numeric) = 1
6. Cleaning fee (numeric) = 1
7. Extra people (numeric) = 1
8. Amenities (converted to binary) = 1 each
  - a. Air conditioning
  - b. Wifi
  - c. Cable TV
  - d. Heating
  - e. Washer
  - f. Dryer

- g. Elevator
- h. Parking
- i. Pets allowed
- j. Bathtub

## **Data Preparation**

The scraped data from Airbnb in both datasets was not formatted in a way that was ready for data mining. In preparing our data, we did all of our preprocessing in Excel and Python, which were both powerful tools in helping us produce the format required for data mining in WEKA and preparing the necessary arff file.

Our first step with the two datasets was merging on the unique 'id' identifier to get them into one file to identify where there were matches across the two datasets. We did this using a vlookup and then removed the instances that didn't yield a match across the two datasets. Once we had a single fileset, we used Python to filter the variables we thought might be interesting and put some extra effort to convert text variables which were amenities and host\_verification. On the one hand, we created binary variables for each feature so that we were able to estimate the individual factor effect. For example, if the housing was equipped with a hot tub, then the binary variable "hot tub" is one, otherwise is 0. On the other hand, we also calculated the number of verifications and amenities for each listing. For example, if the host had three types of verification on his listing, such as email address, Facebook page, and phone number, then the total\_verification was assigned 3. Although the second approach assumed that different types of

amenities and different types of verification are substitutes, the total numbers for amenities and for verification still are fair proxies in our case.

After the Python work, we went back to Excel and the csv file to add filters to the column headers to identify potential issues like stray characters or anomalies in the data. Through this, we discovered the description titles of the properties were not the cleanest, including many instances showing characters like: ~, :::, \*, etc. Additionally, there were random and stray problems like inadvertent zeroes, negative values, etc. Since WEKA would not be able to process these characters, we went through and first performed a “Find and Replace” to adjust some. However, due to the size of the file itself, “Find and Replace” would not always execute on the entire dataset, and we had to do much of this processing manually on 5,000 rows at a time, after which we’d sort and delete the erroneous rows. We decided to remove most of the N/A, zero, and negative values entirely because of the sheer size of instances we had to work with, but likely would have kept them by simply replacing with blanks if we had fewer data available. See screenshots below for snapshots into our data preparation process.

counter	id	name	longitude	latitude	room_type	price	minimum
16	27	~*Spacious	40.81305	-73.95466	Private room	52	2
34	137	~*Private B	40.79295	-73.93997	Private room	69	2
63	169	~*Large Ro	40.67868	-73.97307	Private room	120	3
70	176	51485 Lower East S	40.72319	-73.99201	Private room	83	1
72	1330	652515 COLUMBUS	40.76758	-73.98722	Entire home	85	30
73	1331	652648 GRAMERCY	40.74189	-73.97833	Entire home	87	30
74	1332	652691 COLUMBUS	40.76934	-73.98464	Entire home	95	30
105	1363	671633 REDUCED PR	40.77382	-73.95088	Entire home	141	2
91	2583	1631845 ~*1 bedroo	40.71915	-73.99309	Entire home	150	3
15	3221	2137111 Upper West	40.78689	-73.96985	Entire home	225	2
30	3236	2148025 Bed & Bagel	40.69346	-73.90765	Entire home	225	4
79	3286	2193902 East Village	40.72422	-73.98211	Entire home	225	2
402	3613	2496301 Comfy, Room	40.69587	-73.9096	Entire home	108	3
114	4034	3026848 NYC Finest li	40.74318	-73.97193	Entire home	150	30
30	4050	3046941 Awesome 2	40.74462	-73.97488	Entire home	100	30
165	4085	3091202 Great 1bdM	40.74175	-73.97741	Entire home	84	30
127	4147	3171234 ~ PRIME loca	40.71765	-73.9624	Entire home	425	5
403	4634	3645284 Reno 2BR~St	40.76043	-73.95992	Entire home	160	30
406	4637	3646551 Gramercy~R	40.73568	-73.98062	Entire home	150	30
466	4701	3703633 Reno 2BR~pi	40.73268	-73.98689	Entire home	195	30
83	4825	3817666 ColumbusCr	40.7652	-73.98612	Entire home	150	30
385	4827	3818279 Reno1BR~Ri	40.74445	-73.97181	Entire home	200	30
195	5038	3984168 Union Squar	40.73256	-73.98558	Entire home	170	30
195	5241	4172947 Your Own 2 i	40.6942	-73.95269	Entire home	95	7
79	5428	4338177 Sleeps 41 Pri	40.73947	-74.00114	Entire home	160	30
196	5648	4509500 Prime Chelse	40.73984	-74.0006	Entire home	150	30
192	5649	4509560 Reno Studio	40.73505	-73.98212	Entire home	140	30
353	5805	4621217 Gramercy~N	40.73346	-73.98216	Entire home	140	30
54	5806	4621713 Prime Union	40.7339	-73.98548	Entire home	185	30
123	6587	5231817 UES LRG 1Bf	40.76201	-73.96084	Entire home	150	30
43	6607	5259407 2BR Prime V	40.73226	-74.00644	Entire home	250	30
127	6692	5358354 New BLDG~T	40.76685	-73.98644	Entire home	170	30

Figure 1. Data Cleaning in Excel

airbnb[['host_verifications','amenities']]		
	host_verifications	amenities
0	['email', 'phone', 'google', 'reviews', 'jumio...	{'Cable TV', 'Internet, Wifi', 'Air conditioning', K...
1	['email', 'phone', 'reviews', 'kba']	{TV, 'Cable TV', 'Internet, Wifi', 'Air conditioning...
2	['email', 'phone', 'facebook', 'reviews', 'kba']	{Internet, Wifi, 'Air conditioning', Kitchen, Elev...
3	['email', 'phone', 'reviews', 'jumio', 'govern...	{TV, 'Cable TV', 'Internet, Wifi, Kitchen, 'Buzzer/w...
4	['email', 'phone', 'facebook', 'reviews', 'off...	{Wifi, 'Air conditioning', Kitchen, 'Pets live on...
5	['email', 'phone', 'facebook', 'reviews']	{TV, Wifi, 'Air conditioning', 'Paid parking off ...
6	['email', 'phone', 'facebook', 'google', 'revi...	{Internet, Wifi, 'Air conditioning', 'Paid parkin...
7	['email', 'phone', 'manual_online', 'reviews', '...	{TV, 'Cable TV', Wifi, 'Air conditioning', Kitchen...
8	['email', 'phone', 'reviews', 'jumio', 'govern...	{Internet, Wifi, 'Air conditioning', Kitchen, Door...
9	['email', 'phone', 'reviews', 'kba']	{Internet, Wifi, 'Air conditioning', Breakfast, 'B...
10	['email', 'phone', 'facebook', 'reviews', 'jum...	{Internet, Wifi, 'Air conditioning', 'Free street...

```

a1=[ "Wifi", "Airconditioning", "Kitchen", "Heating", "Bathtub", "Washer", "Dryer",
      "Elevator", "Parking", "Smokedetector", "Petsallowed", "TV" ]

for a in a1:
    airbnb[a]=airbnb2[ 'amenities' ].str.contains(a,na=False)
    airbnb.loc[airbnb[a],a]=1
    airbnb.loc[airbnb[a]==False,a]=0

v1=[ 'email', 'phone', 'facebook', 'reviews', 'jumio',
      'offline_government_id', 'kba', 'selfie', 'government_id', 'identity_manual' ]
for v in v1:
    airbnb[v]=airbnb[ 'host_verifications' ].str.contains(v,na=False)
    airbnb.loc[airbnb[v],v]=1
    airbnb.loc[airbnb[v]==False,v]=0

```

Figure 2. Data Cleaning in Python

## Modeling + Evaluation

While we were able to upload our preprocessed input into WEKA and create the arff file, executing models on the dataset was a challenge. We encountered several attempts where WEKA would throw an error while it was running a model and per the professor's recommendations, we used "MergeInfrequentNominalValues" to reduce dimensionality and removed features that were likely going to cause noise without being informative at all (e.g. property description). Following these steps enabled us to eliminate the noise in our dataset and run the models successfully.

Our target was price, which is a numeric feature type, is shown below in the two images. The left image shows the skewed attributes of pricing and was our first attempt at plotting the



results. To avoid distorting our analysis, we decided to use the natural log of the pricing as our target variable. This distribution is shown on the right.

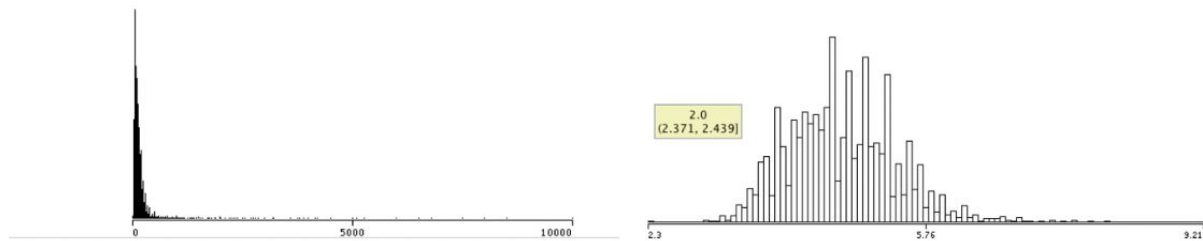


Figure 3. Distribution of Price and Log(price)

In modeling our data, we used three different types of models: Linear regression (LR), REPTree, and M5P -- all with 10-fold cross-validation. By using cross validation we are addressing the potential problem of overfitting. With more generalized evaluation and testing, the models can better handle new data and give more accurate results. Models too specific to our dataset will cause overfitting, which can produce misleading R-squared values, regression coefficients, and p-values.<sup>3</sup> “A regression model that becomes tailor-made to fit the random quirks of one sample is unlikely to fit the random quirks of another sample.”<sup>4</sup>

In defining the three models we used, linear regression is the most classic and common predictive model for a numeric target variable.<sup>5</sup> REPTree is a decision tree that outputs a numeric value and is used as a regression, while the M5P classifier is a decision tree in which each tree node has a linear regression<sup>6</sup>. From our research and understanding, we can think of M5P as a combination of a decision tree and linear regression model.

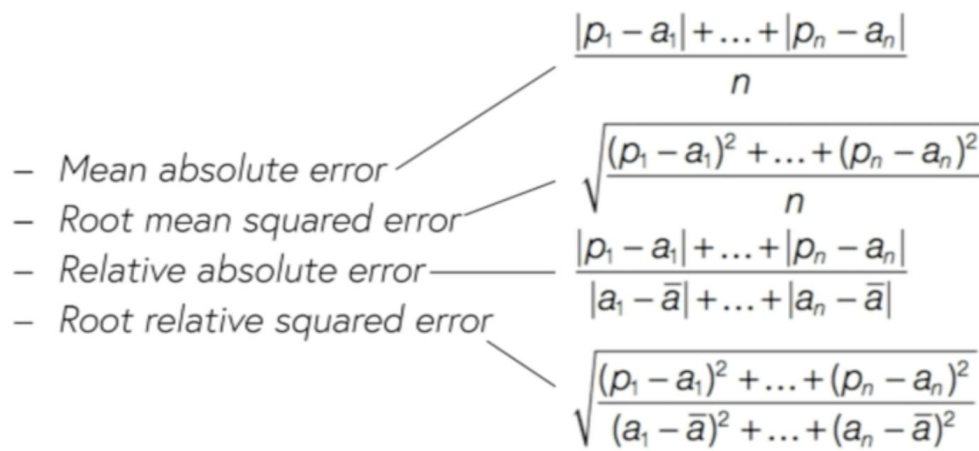
<sup>3</sup> <https://statisticsbyjim.com/regression/overfitting-regression-models/>

<sup>4</sup> <https://statisticsbyjim.com/regression/overfitting-regression-models/>

<sup>5</sup> <https://www.statisticssolutions.com/what-is-linear-regression/>

<sup>6</sup> [https://en.wikipedia.org/wiki/Weka\\_\(machine\\_learning\)](https://en.wikipedia.org/wiki/Weka_(machine_learning))

This excludes classification models as choices for evaluation and therefore disqualifies analysis using the lift curve and AUC. Fortunately, after we performed a regression in WEKA, there were four errors measures we could use for evaluation that WEKA output: “Mean absolute error,” “Root mean squared error,” “Relative absolute error,” and “Root relative squared error.” Below are the definitions for these four types of errors<sup>7</sup>:



- Mean absolute error  $\frac{|p_1 - a_1| + \dots + |p_n - a_n|}{n}$
- Root mean squared error  $\sqrt{\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{n}}$
- Relative absolute error  $\frac{|p_1 - a_1| + \dots + |p_n - a_n|}{|a_1 - \bar{a}| + \dots + |a_n - \bar{a}|}$
- Root relative squared error  $\sqrt{\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \dots + (a_n - \bar{a})^2}}$

Figure 4. Definitions for Error Measures

Similar to accuracy for a classification model, these four error evaluations show prediction performance of our models (REPTree, LR, and M5P).

In addition to error evaluations, R-Squared is also another evaluation method for a regression. Generally, Adjusted R-squared is a better indicator than R-Squared. This is because Adjusted R-Squared calibrates for the number of coefficients relative to the sample size in order to correct it for bias.<sup>8</sup> While an R-Squared evaluation is benefited by adding more and more explanatory variables, the Adjusted R-Squared penalizes this method of adding variables. In the

<sup>7</sup> <https://www.futurelearn.com/courses/data-mining-with-weka/0/steps/25396>

<sup>8</sup> <https://people.duke.edu/~rnau/rsquared.htm>

equation below,  $k$  is the number of independent variables in the equation.

**Adjusted  $R^2$  as a tool for model assessment**

The adjusted  $R^2$  is computed as

$$R_{adj}^2 = 1 - \frac{Var(e_i)/(n - k - 1)}{Var(y_i)/(n - 1)} = 1 - \frac{Var(e_i)}{Var(y_i)} \times \frac{n - 1}{n - k - 1}$$

where  $n$  is the number of cases used to fit the model and  $k$  is the number of predictor variables in the model.

Figure 5. Definition for Adjusted R-Squared

The evaluation of the models we tested are listed in the table below. As shown in the data, the relatively higher correlation coefficient, lower error term, and higher adjusted R-Squared translates to being a better performing model. In this sense, regardless of which criteria we use to compare the different models, M5P emerges as the best model among these three. However, all three models perform well in prediction. As an aside, it is important to note that WEKA only reports 5% significant factors in linear regression and also automatically solves any multicollinearity problems with the explanatory variables. WEKA automatically performs feature selection to only select relevant attributes.

Table 1. Evaluation Results for Models

Measure	REPTree	Linear Regression	M5P
<b>Correlation coefficient</b>	0.7745	0.7976	0.813
<b>Mean absolute error</b>	0.3048	0.2935	0.2784
<b>Root mean squared error</b>	0.4343	0.413	0.3991

<b>Relative absolute error</b>	56.3156 %	53.8541%	51.2568%
<b>Root relative squared error</b>	63.4333 %	60.3128%	58.2828%
<b>Adjusted R-Squared</b>	59.32%	63.21%	65.07%

## Analysis + Insights

Although M5P is a better model for predicting, REPTree and Linear Regression are able to tell us what factors are the best predictors of pricing. By that, linear regression and REPTree can give us estimates for the parameters that we can understand and analyze. As for the M5P, it is much harder to understand and to interpret the results. Having interpretable results is crucial to helping hosts make strategic pricing decisions based on amenities, room type, and location. In the REPTree model, the nodes at the top of the decision tree have the most explanatory power. In this visualization, we found that room type, neighbourhood, and review scores rating were at the top.

```

REPTree
=====
room_type = Private room
|
| neighbourhood_merged_infrequent_values = Highbridge or Clason Point or Eastchester or Woodlawn or Un
|
| review_scores_rating < 5.5 : 4.43 (86/0.4) [30/0.57]
|
| review_scores_rating >= 5.5
|
| neighbourhood_group = Manhattan : 4.43 (9/0.31) [4/0.14]
|
| neighbourhood_group = Brooklyn
|
| extra_people < 27.5 : 4.07 (15/0.21) [2/0.08]
|
| extra_people >= 27.5 : 4.76 (4/0.03) [2/0.05]
|
| neighbourhood_group = Queens
|
| cleaning_fee < 95
|
| Heating < 0.5 : 4.86 (3/0.15) [1/0.3]
|
| Heating >= 0.5
|
| availability_365 < 315
|
| cleaning_fee < 55
|
| calculated_host_listings_count < 4.5
|
| total_host_verifications < 5.5
|
| | minimum_nights < 3.5 : 4.12 (24/0.02) [12/0.25]
|
| | minimum_nights >= 3.5 : 3.83 (3/0) [2/0.02]
|
| total_host_verifications >= 5.5 : 3.87 (17/0.07) [8/0.06]
|
| calculated_host_listings_count >= 4.5
|
| minimum_nights < 1.5 : 3.97 (3/0.01) [3/0.09]
|
| minimum_nights >= 1.5 : 3.71 (6/0.01) [2/0.09]

```

Figure 6. Results for REPTree Model

Moreover, the linear regression model validates these findings with room type having the greatest percent effect on pricing. With shared room as the dummy variable (benchmark), a private room can be priced 38% higher and an entire home apartment can be priced 92% higher than a listing that is a shared room with similar amenities. It is fair to say that the most important indicator of Airbnb pricing is room type.

Table 2. Parameters for Room\_type

Room_type	Parameter
Private Room	0.3804
Entire Home Apartment	0.9201

Below are the parameters for different boroughs in the linear regression model, where the benchmark is the Bronx. For example, housing in Queens is priced 13% higher than in the Bronx on average. On the other hand, Staten Island prices for listings are 1.8% lower than in the Bronx. These threshold effects on location can also be captured in bed types. In the table below on bed types, futons were chosen as the benchmark feature. So listings that included couches benefitted with pricing 14% higher than similar listings with futons.

Table 3. Parameters for Borough

Borough	Parameter
Queens	0.1305
Staten Island	-0.0181

Brooklyn	0.2153
Manhattan	0.4468

Table 4. Parameters for Bed\_type

Bed_type	Parameter
Couch	0.1424
Airbed	0.0533
Pull-out sofa	0.0854
Real-bed	0.0391

#### Other Interesting Findings

From our linear regression model, we found estimates that matched our intuition on pricing. For example, a listing on Airbnb that includes an elevator can be priced 11% higher than a similar listing that includes the same amenities but does not include an elevator. An interesting finding is that a kitchen is negatively correlated with price. That is, if a kitchen is included as part of the amenities, the price decreases by 9%. This can be explained as a unique effect of tourists and travellers. Tourists and guests might not think a kitchen is useful if they're typically planning on eating out at the many restaurants in New York City while they're in town. As a standard for alternative comparison, this makes sense since hotels do not carry personal kitchens and instead often have restaurants for their guests to dine at. As before with the boroughs, a

listing in Manhattan can be priced 45% higher than a similar listing in the Bronx. These findings make sense and match our understanding of pricing here in New York City.

Table 5. Parameters for Amenities

Amenities	Parameter	Amenities	Parameter
Air-conditioning	0.0595	Washer	0.0478
Kitchen	(-0.0924)	Dryer	0.0216
Heating	(-0.0131)	Elevator	0.1132
Bathtub	0.0467	Parking	(-0.0071)
CableTV	0.0676		

## Deployment /Applications

In considering deployment of our model, we feel we could plug into Airbnb's robust data science team which, at the time of [this video](#), has a department size of more than 150 data scientists<sup>9</sup>. With access to the original (not scraped) data, we think Airbnb's data science team could improve our model exponentially and scale it out to other markets around the world. From a business perspective, making pricing recommendations would not only drive more revenue, but help create a better experience for the host by taking the guessing out of their pricing decision. With access to such powerful and influential tools comes responsibility: we feel that there are ethical implications for Airbnb to ensure their pricing is objective, factual, and transparent.

---

<sup>9</sup> "Scaling Knowledge: A Look at Airbnb Data Science Team" <https://www.youtube.com/watch?v=6QVXPnrSbLU>

Because Airbnb takes a percentage of the booking transaction amount from hosts, the company would be incentivized to recommend higher and higher prices since that would increase their revenues. However, it is important to take this into account so the model is not deployed with this intention nor used to take advantage of higher pricing since this will affect the end user—many of whom choose Airbnb for the economic value as a more affordable alternative to hotels. To mitigate this, we could set controls and thresholds to maintain the pricing recommendation outputs within certain minimum and maximum ranges based on metrics like mean or median of similar properties. Additionally, another plausible deployment method would be for a third-party or standalone company or group to maintain the technology, as opposed to having Airbnb's team managing it directly. Of course, this would be a more difficult sell since Airbnb would likely want the control of upkeeping the model internally, but still something to consider.

### **Reflection + Learnings**

Our team learned a lot during the course of this project. Beyond the fact that if we want to be Airbnb hosts, we should rent out properties in Manhattan with air conditioning in elevator buildings, we worked well as a team. We found immense value in having diverse and balanced skillsets between Scott's Excel work and strong business background, as well as Gao's Python training. We combined all of our skills in order to preprocess our complicated and large datasets, from merging the two datasets to deleting erroneous characters to converting to a format that WEKA could accept as an input. We performed our analysis both electronically and in person through group meetings, and were very communicative throughout, which helped make the



project move along smoothly. One of the key components of our project that we were lucky to have was time—we started weeks in advance and needed all of it since there were many hiccups with our dataset that we could not have anticipated. Ultimately, we were able to create what we believe is a valuable and relatable final report, since most people have used Airbnb in the past, and are proud of the work we have done and the final presentation we gave in class. It was a great capstone to end the course!