Craig Bentzen

Airbnb Chicago Price Prediction

Problem Statement:

Getting started as a host on Airbnb is a series of daunting questions: What things are guests looking for in an Airbnb listing? How do I maximize the chance my guests will give me good reviews? How do I stand out from the other 100 or so Airbnb listings in the surrounding area? And the ultimate question - What should I charge per night for my listing? For the last question, Airbnb provides a tool that suggests a price per night. I want to examine the suggested price tool using data analysis and machine learning techniques on Airbnb listing data from Chicago.

After all the work and expense of setting up a listing, seeing what Airbnb recommended for my listing was fairly demoralizing and financially scary. Airbnb's price estimator tool recommended that I rent my listing for an average of $40 during the week and $60 on weekends. I couldn't figure out how Airbnb calculated these numbers so I immediately discounted them. Instead I looked at a local hotel's nightly price, discounted 10% and viola - nightly average price of $125. I didn't have immediate success but after one month I was making money and fully booked on weekends and partially booked during the week.

After six months of successful bookings and positive reviews, did the price estimator tool ever update its price? Not significantly. The highest suggested weekend price was $70. Still fairly far off from the $125 guests were willing to pay. I was always intrigued with why the price estimator tool priced my listing so poorly. To satiate my curiosity I want to gain insight into the following questions:

- Which features of an Airbnb listing are most impactful to price?
- Can I create a machine learning model to accurately predict the nightly rate hosts are listing their properties for?

Data Wrangling:

The source data for the Chicago Airbnb data set used in this analysis can be found on InsideAirbnb.com's website. Inside Airbnb provides monthly snapshots of the publically available Airbnb data for free use. Inside Airbnb segments the data by city for many of the worlds largest Airbnb Market's. I choose Chicago because I live in Chicago.

The source data is scrapped from Airbnb's website and certain fields are pre-aggregated for analysis. The raw data consisted of 8,519 records, one record per Chicago listing, with 72 columns or features.

The below list of cleansing and cleanup tasks were performed on the data:

- 1,396 records were dropped because the listing has yet to receive a review. A listing without reviews will be considered unused, fake, or the host is being paid off the platform. In any case, the price hasn't been confirmed by guest reviews, so we will remove these listings.
- 27 features contained null entries. Categorical/string fields were filled with 'none' and numerical fields were filled with '0'.
- Price column had to be converted to numeric data type and special characters were removed.
- 32 features or parameters were dropped for various reasons: measured the same thing, not helpful towards analysis, didn't have enough data points.
- Outliers were reviewed to determine if they are true outliers or bad data that should be removed.
- Low outliers on Price were removed because it doesn't make sense that any Airbnb listing should be 1 or 2 dollars a night.
- 4 word count features were created to summarize number of amenities, length of listing description, length of house rules, length of the about host section
- Number of reviews per month were rounded to the nearest whole number for easier segmentation
- During the data analysis phase, the neighborhood parameter was seen to be very influential to price (which seems logical). To explore the impact neighborhood had in combination with other features the following parameters were created: Average price by neighborhood, Average price by neighborhood and property type, Average price by neighborhood and number of accommodations, Average price by neighborhood, beds, and bathrooms, Average price by neighborhood and rounded reviews per month.

The final shape of the cleaned data set was 6,508 records and 42 parameters.

Exploratory Data Analysis:

As the target variable is price per night. The below figure displays the distribution of the nightly price. The distribution is highly skewed to the right and multi peaked. The peaks occur at somewhat regular intervals. For example: 200, 250, 300, 350 and 400 are all peaks. Hosts must round to these numbers at higher prices.

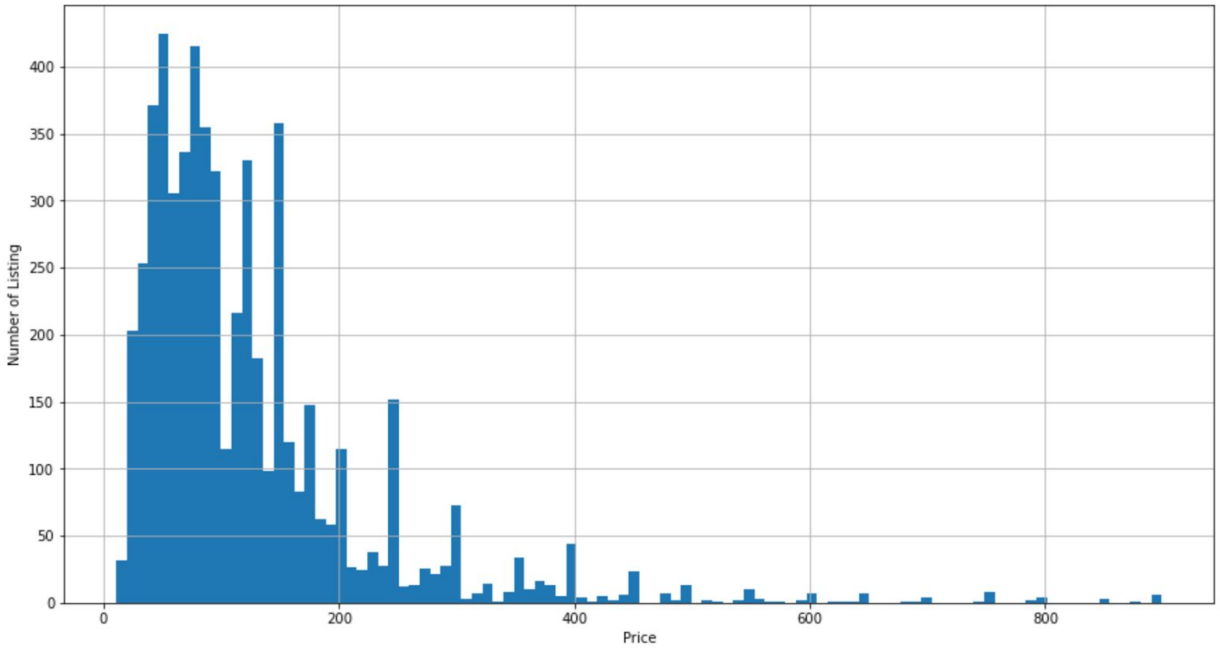


Figure XX: Histogram of Nightly Price

The first question I wanted to explore is how the neighborhood of an Airbnb listing affects price. The figure below displays the listing's neighborhood vs nightly price. The "Other" category are all neighborhoods that have less than 100 Airbnb listings.

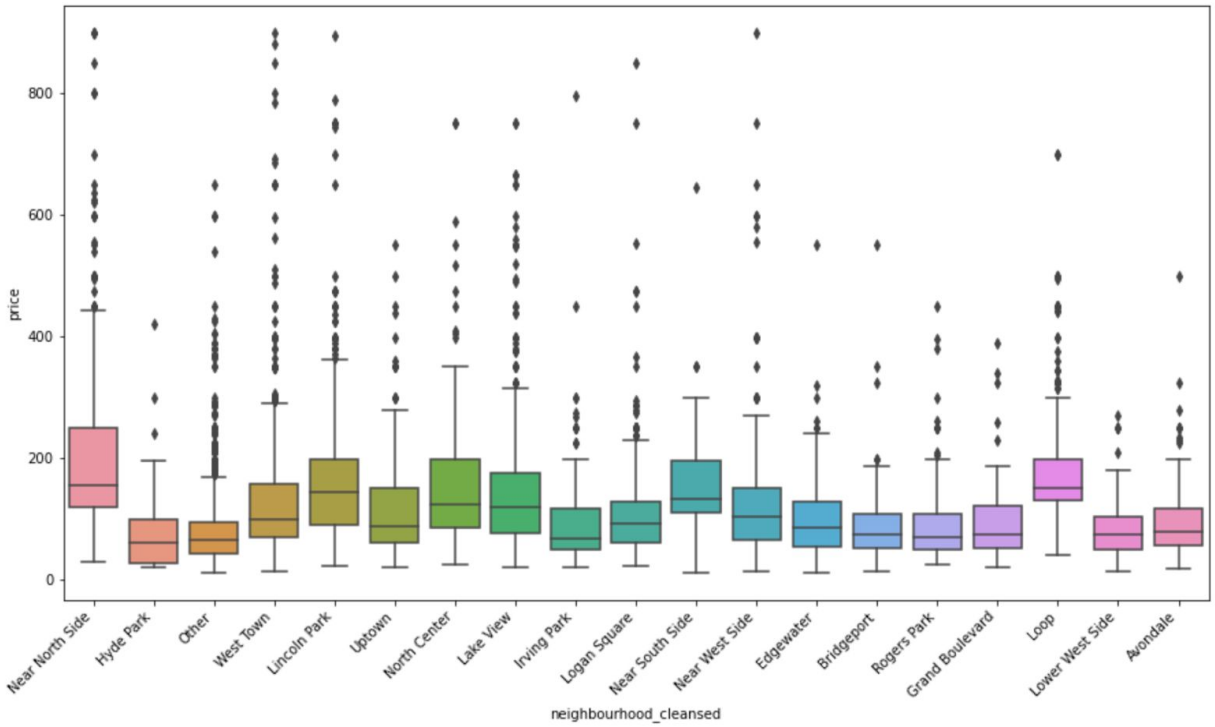


Figure 1: Box and whisker plot of Neighborhood vs. Nightly Price

There are many outliers in the above box and whisker plots. Is this bad data or the highly skewed nature of the distribution of price? In the next figure the price is segmented by the number of people the listing can accommodate.

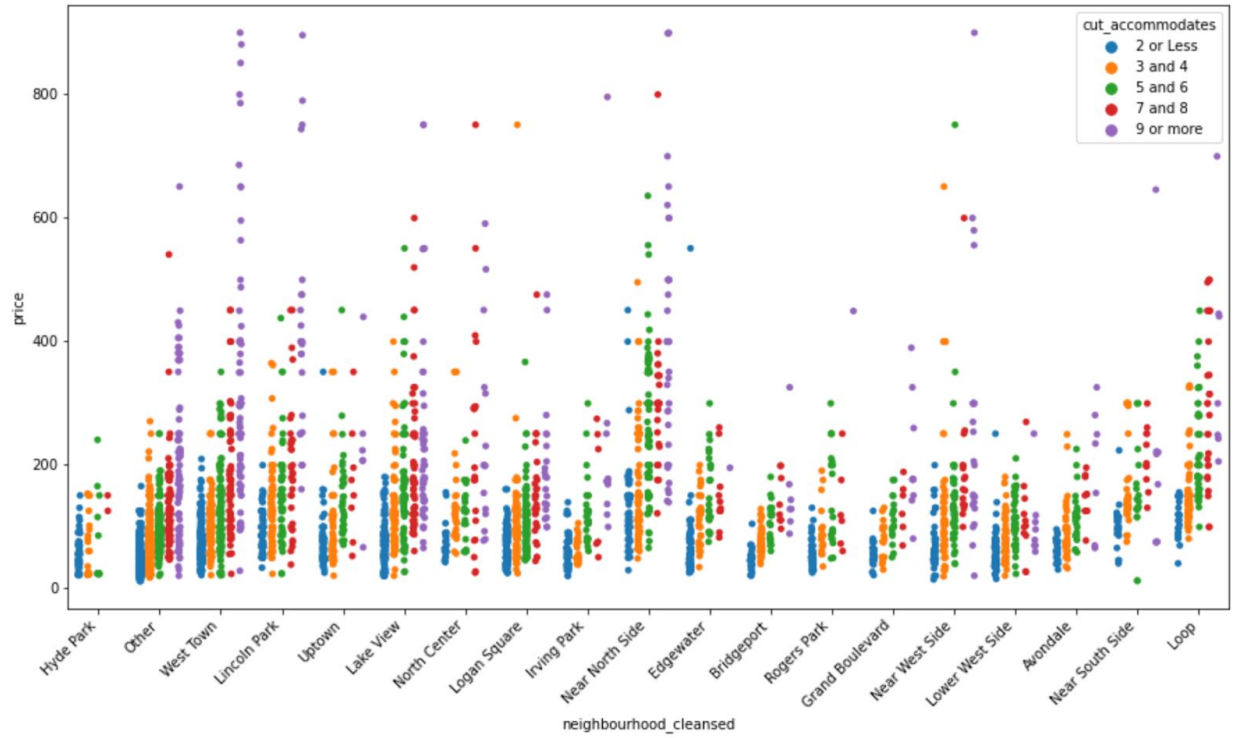


Figure XX: Scatter plot of Price vs Neighborhood and segmented by Accomodates

The vast majority of listings priced above $400 can accommodate 5 or more guests. For that reason we should leave the outliers in data because they are not bad data.

As a host, the takeaway is that listings with five or more guests have a significantly larger range of values.

The next figure segments price by the number of reviews per month.

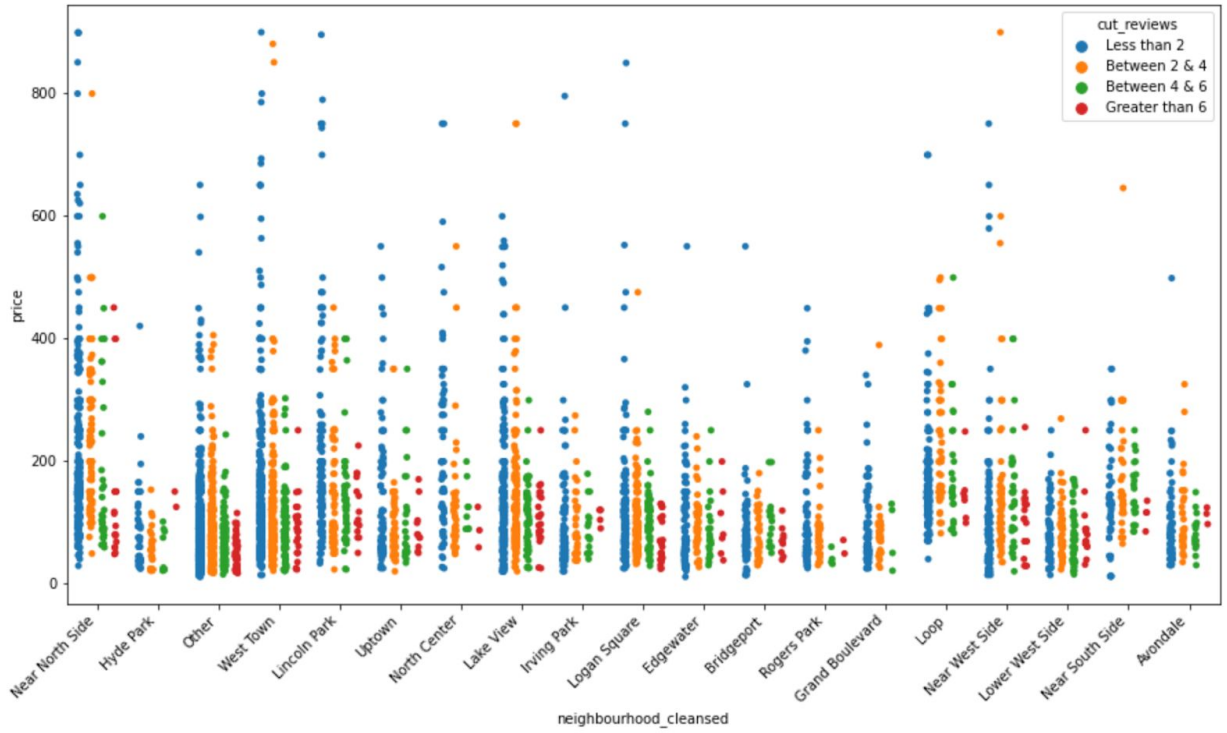


Figure XX: Scatter plot of Price vs Neighborhood segmented by Number of Reviews per Month

Almost all of the listings with six or more reviews a month have a price lower than $200 a night. Listings with a price of over $400 per night get rented at a decreased rate because the fast majority receive two reviews or less a month.

In the next figure I want to further examine the distribution of price in each neighborhood.

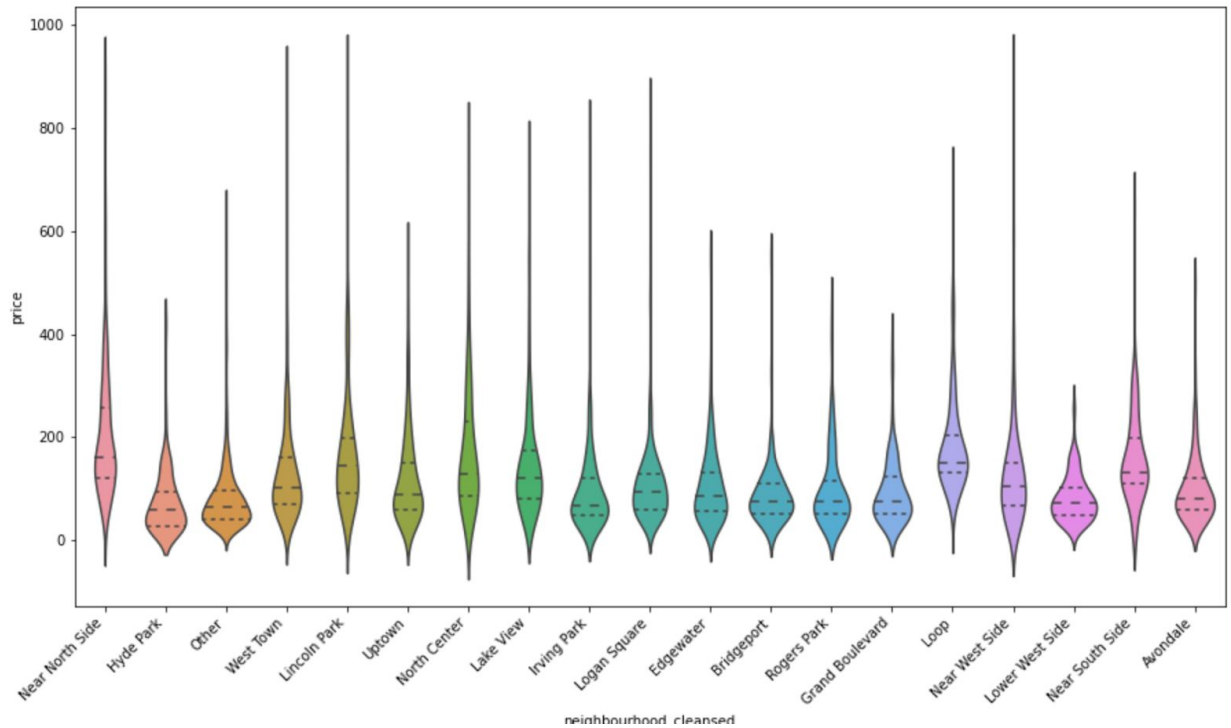


Figure XX: Violin Plot of Price vs Neighborhood

It is interesting to note in the violin plot that the most expensive neighborhoods to live in generally have the widest distributions. Near North Side, Lincoln Park, North Center, Lake View, and Near West Side are all expensive neighborhoods with a lot of tourism. As a host, the take away is that the price for these neighborhoods will have a large range of values that guests are willing to pay.

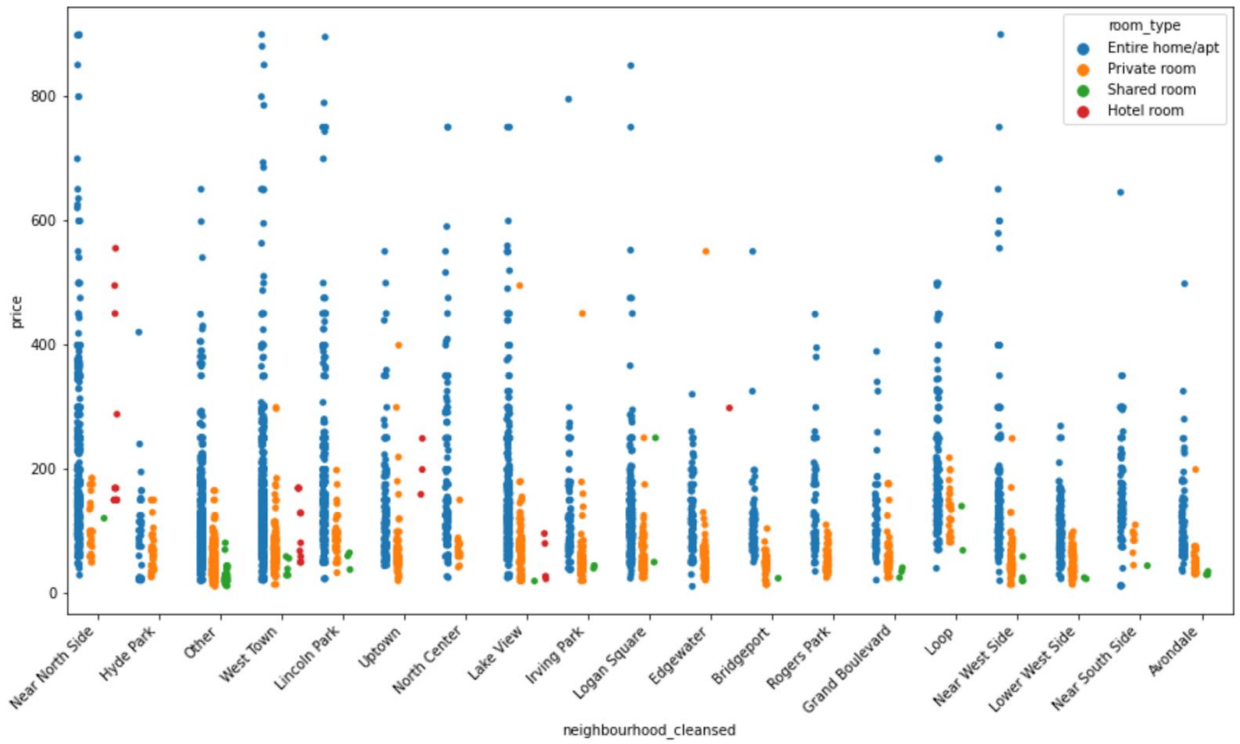


Figure XX: Scatter plot of Price vs Neighborhood segmented by Listing Type

The vast majority of listings on Airbnb are for Entire Homes and Apartments. Very few private rooms are rented for above $200.

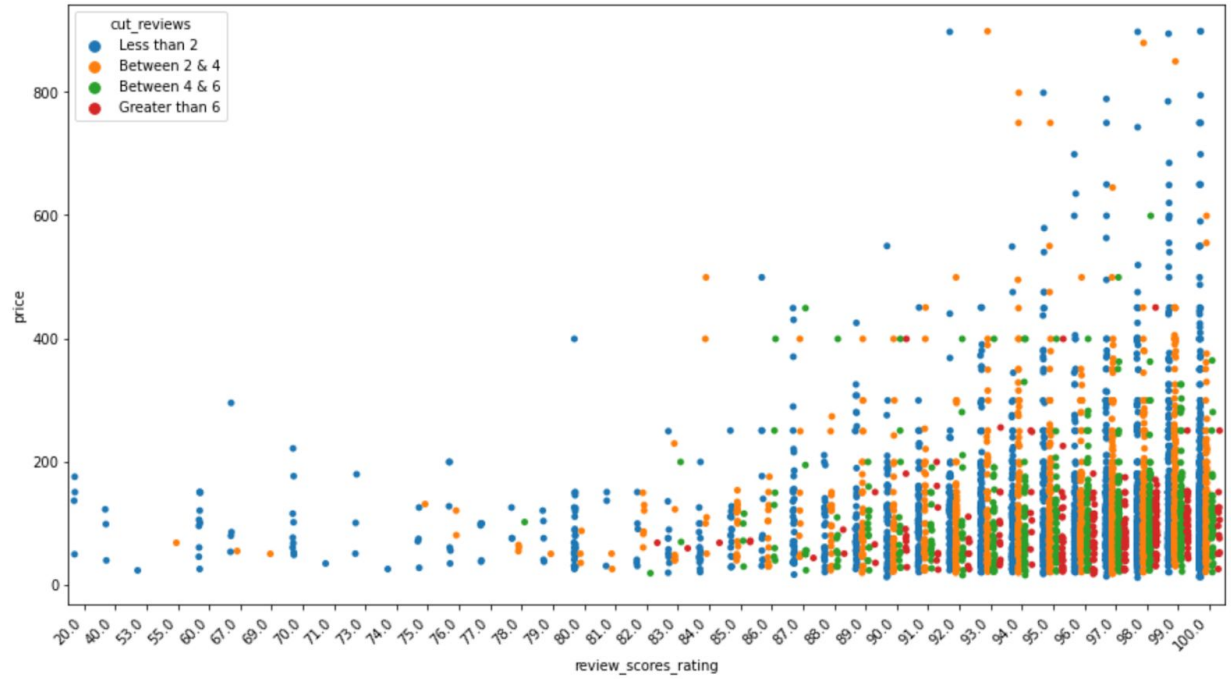


Figure XX: Price by Review Rating

The above figure shows the review rating and price segmented by the number of reviews. It is interesting to note that the majorting of listings with 6 or more ratings a month have a review rating of at least 90%. Very few listing that want to charge more than $400 a night have a review rating less than 90%.

Amenities are a free text field in the Airbnb platform and hosts and list as many amenities as they feel warranted. But do these amenities affect price at all. The below join plot does find a slight positive correlation to price.

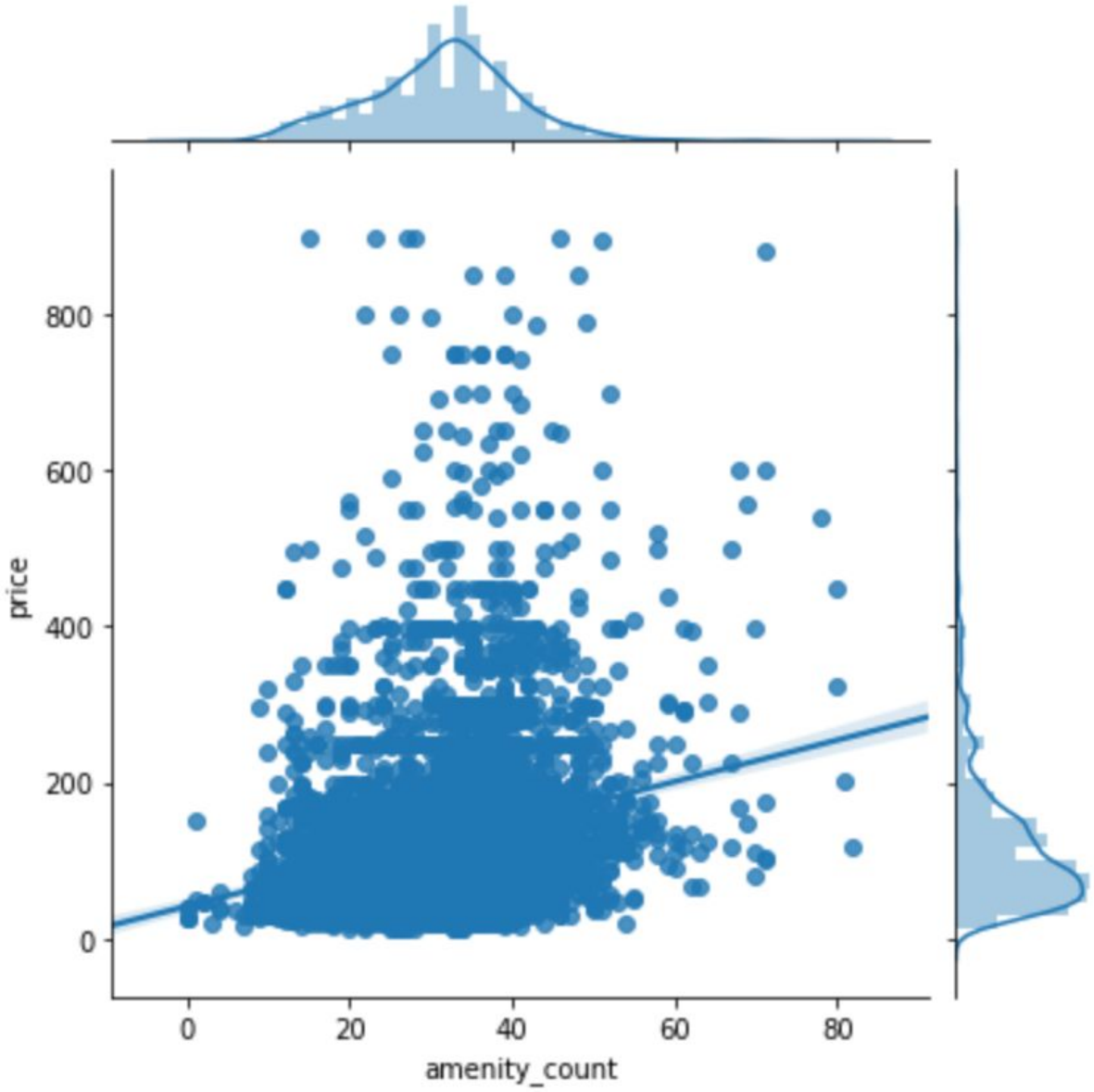


Figure XX: Joinplot of Amenities Count vs Price

Is it possible to glean any understanding form high performing Airbnb listings? I'm going to define a high performing Airbnb as any listing that has a greater than 98% rating and more than 3.2 ratings per month. The top 25% in each category.

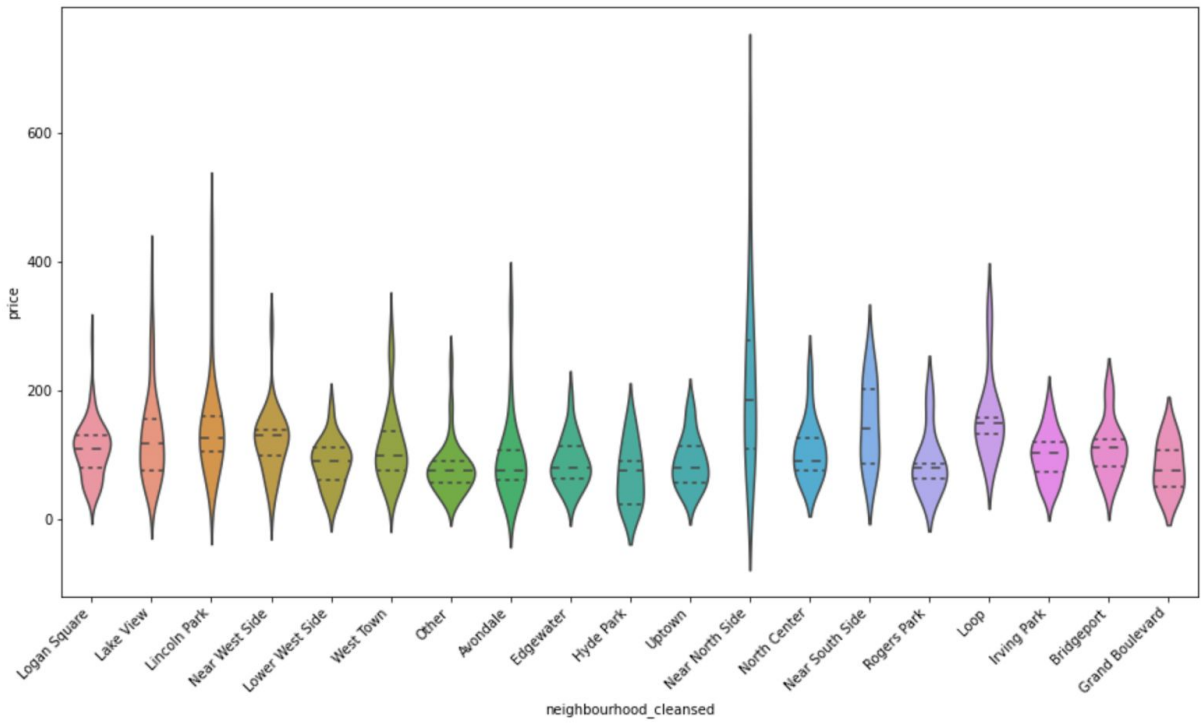


Figure XX: Violin Plot of Neighborhood vs Price

What is interesting is that there are fewer outliers and the distributions for the most part are flatter. This makes sense as high performing Airbnb's know their worth. Also the neighborhoods with the most tourists again have the flattest distributions and outliers.

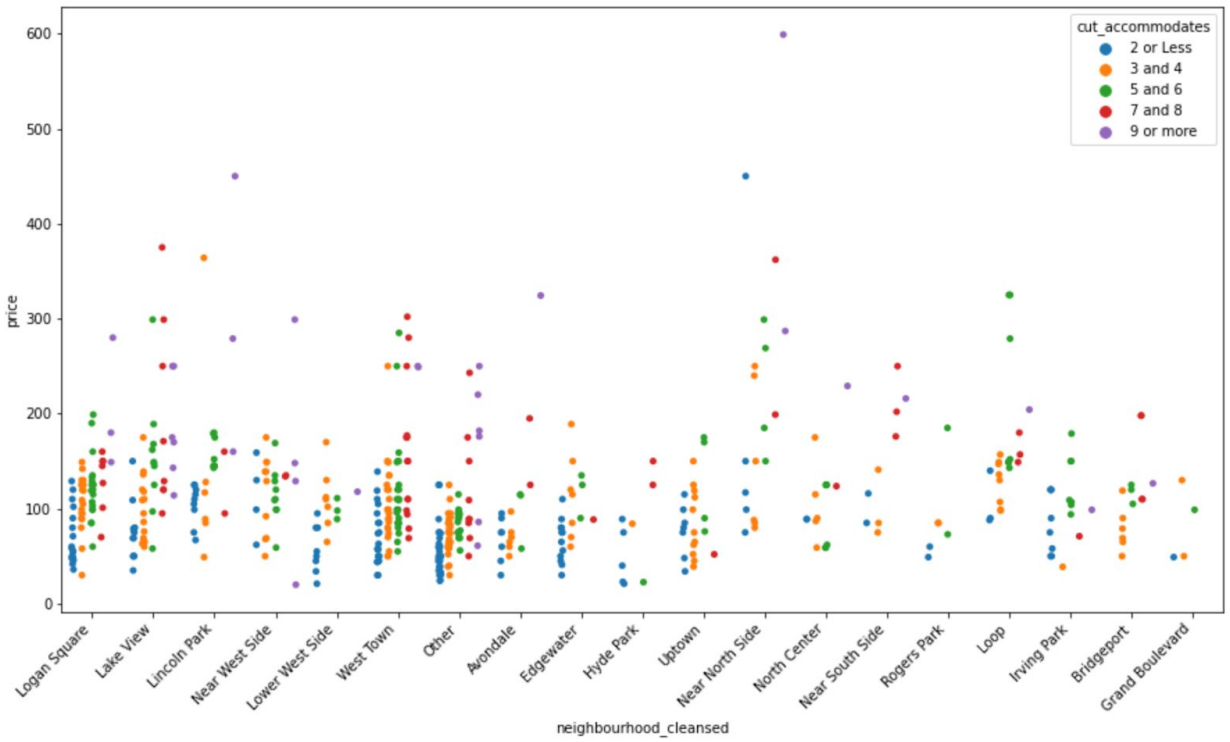


Figure XX: Price vs Neighborhood segmented by Accommodates of high performing Airbnbs

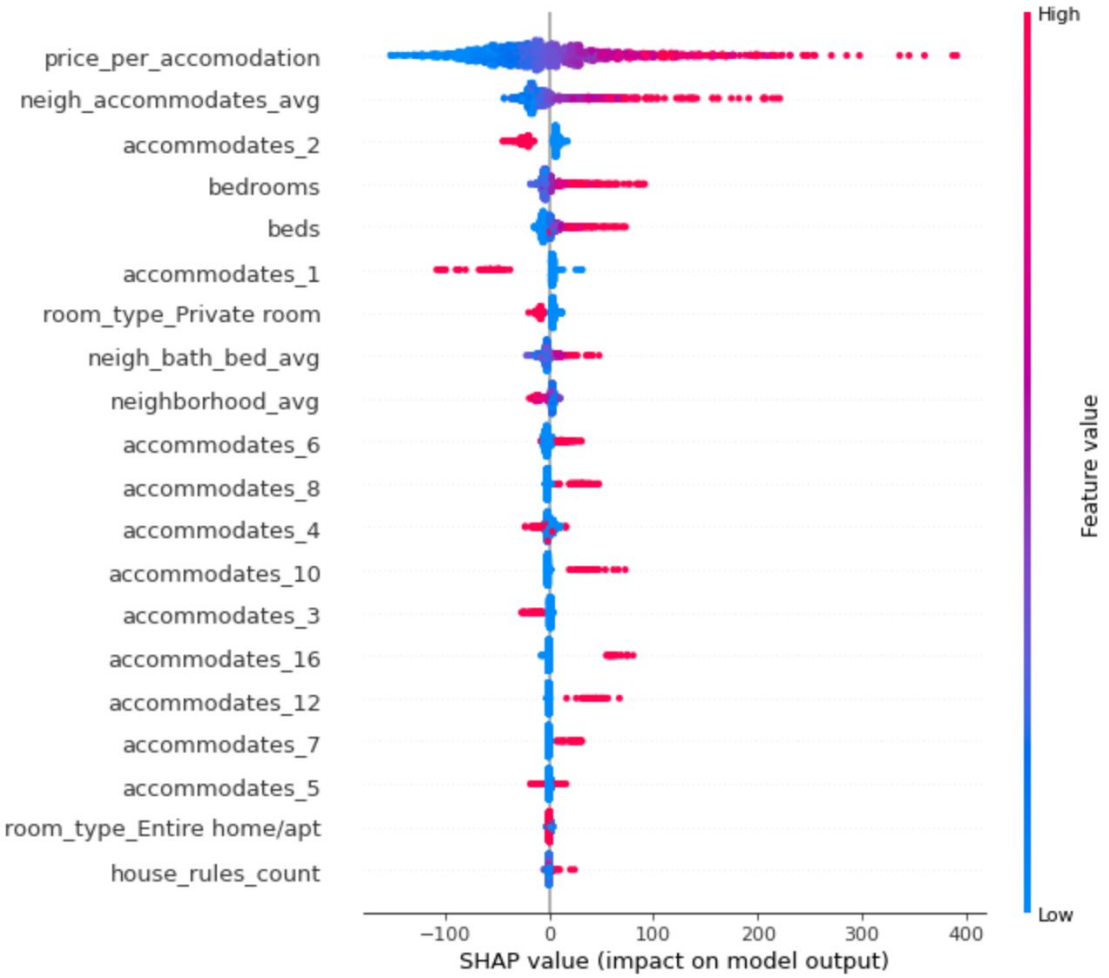
The above shows that Airbnb's that have more accommodations know to charge more on average.

Data Modeling:

The below table depicts various regression models and techniques. Using the below as a starting point I want to examine and tune some of the more effective models and find the most important features.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
0	CatBoost Regressor	6.1774	257.1911	15.6446	0.9760	0.0802	0.0473	4.6730
1	Light Gradient Boosting Machine	8.6614	434.9656	20.5553	0.9588	0.1035	0.0662	0.2723
2	Extra Trees Regressor	6.1278	511.5150	22.0489	0.9520	0.0969	0.0413	1.2469
3	Gradient Boosting Regressor	14.4668	611.6706	24.4608	0.9416	0.1680	0.1296	1.2304
4	Extreme Gradient Boosting	14.5458	620.7386	24.6437	0.9408	0.1704	0.1315	0.4080
5	Random Forest	9.4023	637.2742	24.6185	0.9404	0.1316	0.0687	2.0830
6	Decision Tree	11.5174	1088.9661	32.6613	0.8941	0.1745	0.0773	0.0678
7	Ridge Regression	23.7955	1715.9288	41.1775	0.8331	0.4090	0.2692	0.0156
8	Bayesian Ridge	23.7760	1716.1391	41.1827	0.8331	0.4050	0.2674	0.0326
9	Linear Regression	23.8735	1722.9069	41.2557	0.8324	0.4133	0.2726	0.0412
10	TheilSen Regressor	23.5460	1839.9391	42.6519	0.8243	0.3938	0.2561	5.7409
11	Orthogonal Matching Pursuit	25.2228	1860.5885	42.9165	0.8178	0.3975	0.2797	0.0151
12	Lasso Regression	24.7669	1914.3776	43.5291	0.8122	0.3648	0.2474	0.0312
13	Huber Regressor	26.5576	2196.2483	46.6308	0.7858	0.3282	0.2572	0.4688
14	Elastic Net	27.0952	2186.5079	46.5335	0.7847	0.3517	0.2627	0.0199
15	Passive Aggressive Regressor	33.1900	2863.5411	53.0105	0.7259	0.4531	0.3435	0.0419
16	AdaBoost Regressor	47.1985	3072.2750	55.3868	0.6969	0.5713	0.6964	0.8818
17	K Neighbors Regressor	34.5213	3384.0607	57.7661	0.6732	0.3913	0.3422	0.0674
18	Support Vector Machine	42.0357	5992.4331	76.4971	0.4422	0.4467	0.4017	1.8224
19	Lasso Least Angle Regression	60.8185	8182.8585	89.7399	0.2290	0.6650	0.7815	0.0145

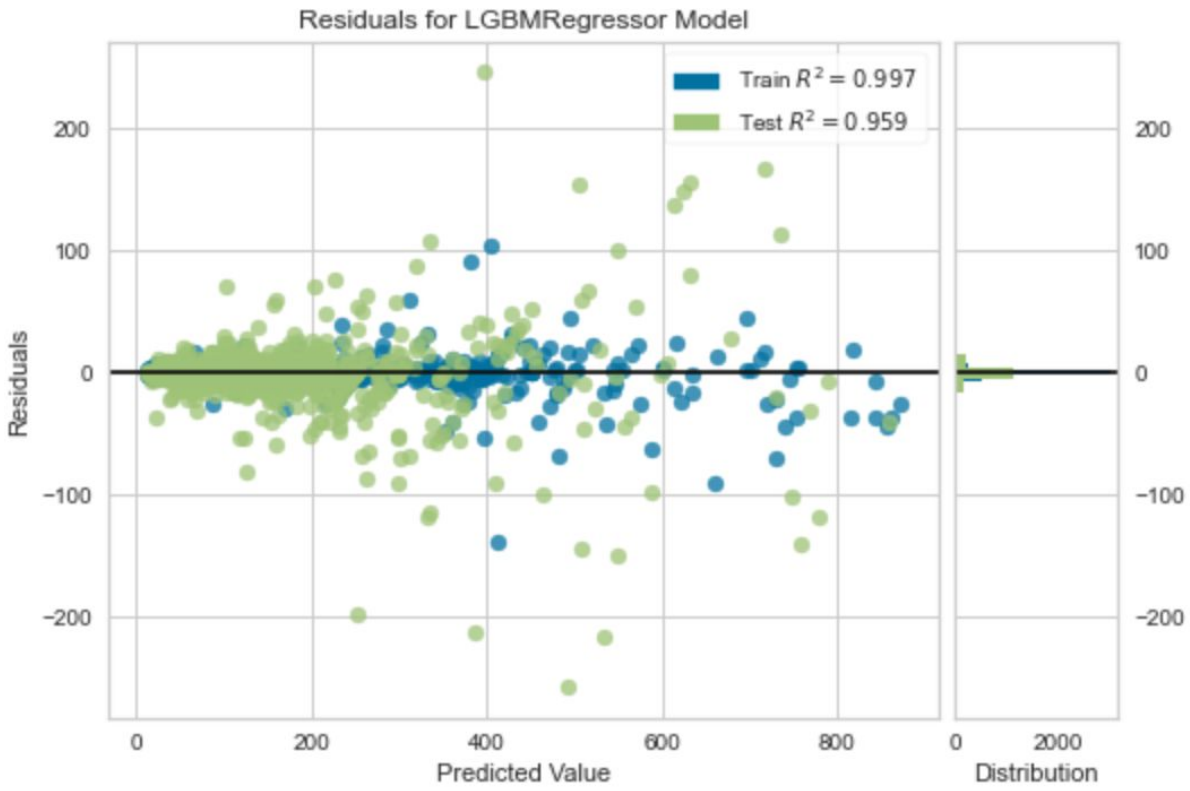
Initial impressions is that the data is modeled fairly well by a number of models. The Decision Tree based models are better at maximizing the R squared value and minimizing the Root Mean Squared Error.



The above figure displays the SHAP values for the CatBoost Regressor model. The SHAP values for the various features show which features continue the most to the model positively or negatively. To read a SHAP value chart. Values to the right of 0 on the X-axis contribute positively to the target variable and values to the left contribute negatively to the target. Values that are pink have a higher impact on the target variable than values that are blue. So as Price per Accommodation increases it positively affects price. Alternatively, when a listing only accommodates one guest, it negatively impacts price.

The means that the most important features that dictate price are the number of people that listing can accommodate.

These feature importance plots were similar across the top models.



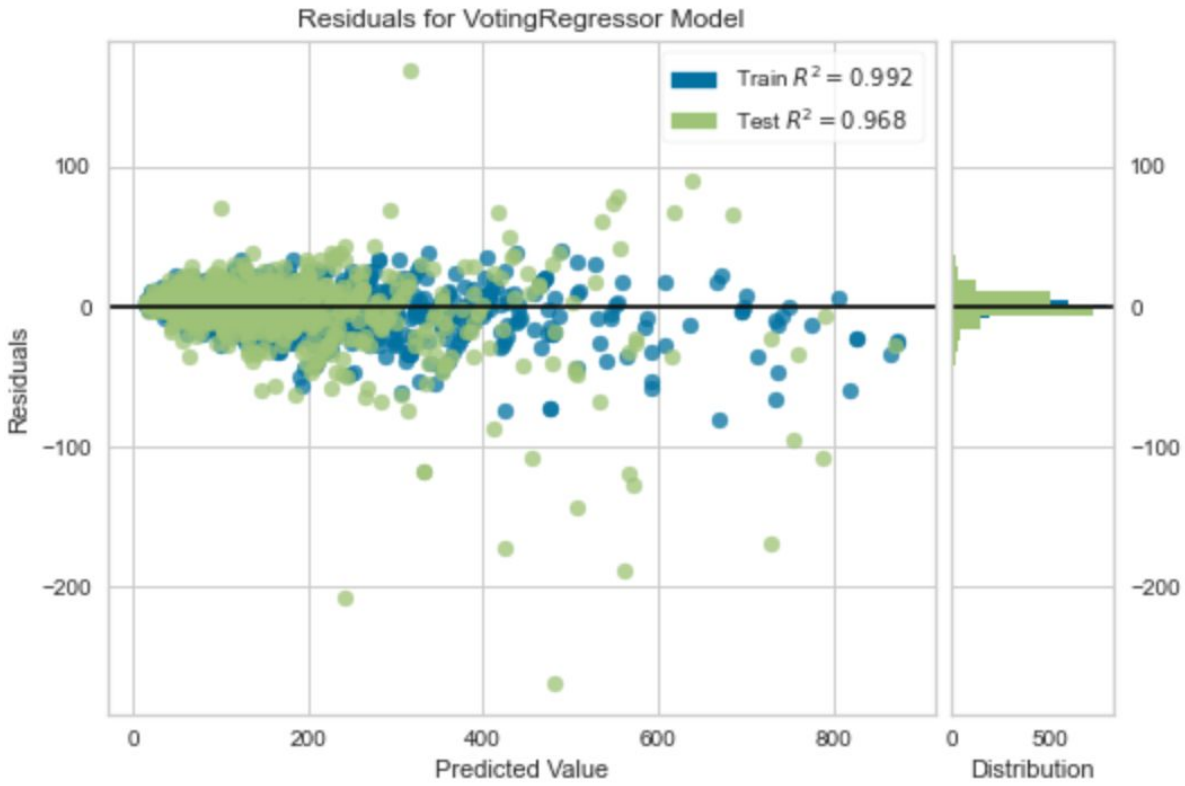
The above figure depicts the residual plot for the Light Gradient Boost Method Regressor Model. The residuals look fairly evenly distributed.

The features were normalized and the models were rerun. The following table is the output:

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
0	CatBoost Regressor	6.0711	257.2605	15.6163	0.9760	0.0783	0.0462	4.5625
1	Light Gradient Boosting Machine	8.7502	436.1221	20.6245	0.9587	0.1047	0.0670	0.2641
2	Extra Trees Regressor	6.3017	536.1047	22.5285	0.9501	0.0979	0.0420	1.3591
3	Gradient Boosting Regressor	14.4654	612.6271	24.4843	0.9415	0.1683	0.1298	1.2077
4	Extreme Gradient Boosting	14.5458	620.7386	24.6437	0.9408	0.1704	0.1315	0.4283
5	Random Forest	9.3873	638.6983	24.6391	0.9404	0.1313	0.0685	1.3923
6	Decision Tree	11.9502	1220.3340	34.5562	0.8816	0.1808	0.0809	0.0690
7	Ridge Regression	23.7920	1715.8695	41.1768	0.8331	0.4091	0.2690	0.0139
8	Bayesian Ridge	23.7749	1715.8887	41.1788	0.8331	0.4062	0.2677	0.0438
9	Huber Regressor	22.7970	1752.0048	41.6263	0.8309	0.3718	0.2356	0.5166
10	Orthogonal Matching Pursuit	25.2228	1860.5885	42.9165	0.8178	0.3975	0.2797	0.0136
11	Lasso Regression	24.5087	1919.5160	43.5779	0.8130	0.3609	0.2413	0.0217
12	Passive Aggressive Regressor	27.1944	2013.3742	44.7240	0.8042	0.4497	0.2920	0.0598
13	K Neighbors Regressor	29.9092	2571.0757	50.4922	0.7510	0.3374	0.2832	0.0862
14	Elastic Net	30.9808	2732.0976	51.8806	0.7407	0.3514	0.3149	0.0172
15	AdaBoost Regressor	48.5175	3249.5264	56.9515	0.6807	0.5799	0.7137	0.8562
16	Support Vector Machine	39.1599	6308.9208	78.3599	0.4157	0.4191	0.3476	1.7032
17	Lasso Least Angle Regression	60.8185	8182.8585	89.7399	0.2290	0.6650	0.7815	0.0133

As you can see the CatBoost Regressor is the best model with a Mean Absolute Error of 6.07, RMSE of 15.6 and R2 of .9760.

The below table is a residual plot of an ensemble model of the top five models of the previous table.


Residuals for VotingRegressor Model
Train $R^2 = 0.992$
Test $R^2 = 0.968$
Residuals
100
0
−100
−200
0
200
400
600
800
Predicted Value
0
500
Distribution

Conclusions:

With and R2 of .97 and Mean Absolute error of around $6.00 it is possible to build a fairly good model to predict Airbnb nightly price. The most important features relate to the number of accommodations and neighborhood. Creating features around the most important features allowed the model to become far more accurate.

Future Research:

I think it would be interesting to inspect the records that cause the most error. There were a lot of very low nightly prices that don't make a lot of sense to me, example a listing in a very nice neighborhood for $15 a night. This requires a lot of manual effort and review.

I also think gathering the photo data would be useful. Having just the count of photos and relative brightness might be useful enough. No one will rent an Airbnb with only two dark photos. Going down the rabbit hole you could use computer vision to confirm all the rooms, beds, and bathroom exist.

Doing some natural language processing on the title, description, about, and reviews would also be interesting. Do successful listing use similar phrases? Do some phrases contribute to price?