Predicting the Nightly Rate to List an Airbnb in New York

Background, Motivation, & Business Question

Our team is looking to create a model that helps decide how much a person should list their New York rental on AirBnb for. With over 36,230 airbnb listings in New York as of 2021 (Peck, 2022), overpricing a rental can lead to lost business. Underpricing, on the other hand, can lead to lost profits. Pricing the listing competitively is key for success in the short-term and holiday rental market.

The dataset used in this report is a cross-sectional sample of New York Airbnb rentals listed in 2019. The data is sourced from Kaggle in the form of a csv file (Dgomonov, 2019). It contains 16 features of interest, including our y (dependent) variable, price.

Based on the dataset we can use machine learning to predict the price at which a new lister should offer their property for rental on AirBnb in New York. We compared multiple regression models to arrive at the one which explains the listed prices most effectively.

Our statistical question is to predict the price of a rental for a given location, listing features, neighborhood, type of listing space and number of reviews. It addresses the business question as it will help a new lister to determine the price that they should list their Airbnb rental for since the predicted price will be based on similar listings that are currently on Airbnb.

The statistical question falls short of the business question because it does not consider what the optimal listing price is for a unit, the statistical question only aims to predict what the listing price would be based on current listings. This fails to take into account different pricing strategies, such as preferred vacancy rate or demand for a particular unit.

We recognize our dataset's shortcomings to be the following:

1. Physical properties of rental are not included: The dataset does not contain the size (square footage) or quality of construction of rental for observations.
2. Popularity of rentals not included effectively: The dataset does not account for the popularity of AirBnb "superhosts", popular rentals, average reviewer ratings/sentiment and recency of reviews.
3. Amenities included in the rental are not considered: Amenities such as cleaning services, concierge services, or physical amenities such as access to pools, hot tubs etc. are not available in the dataset. Flexibility and Payment Schedule not mentioned:

1.

AirBnb hosts have varying cancellation and payment schedules. Cancellation policies vary from flexible to extremely strict. (https://www.igms.com/airbnb-cancellation-policy/). Payment schedules too vary from immediate collection at booking to collection the day of check in. These features are not available in the dataset.

4. No. of Beds not available: Our dataset does not mention the total number of beds and/or the maximum number of guests allowed per listing.
5. The dataset does not contain how often a listing sits vacant or the percentage of time that a unit is vacant. There is likely some correlation between the number of reviews and how often a listing is rented, however we do not know how long each listing has been on Airbnb. The model could be skewed by listings that are priced too high and are never rented or by units priced too low that are always booked.

Exploratory Data Analysis

Prior to building a model to predict price, we sought out to explore the training data to better understand its features and observations. Please note that the training/test split was done prior to this step. The key findings from our exploration are listed below.

1. There a relatively small number of highly priced observations that skew the data

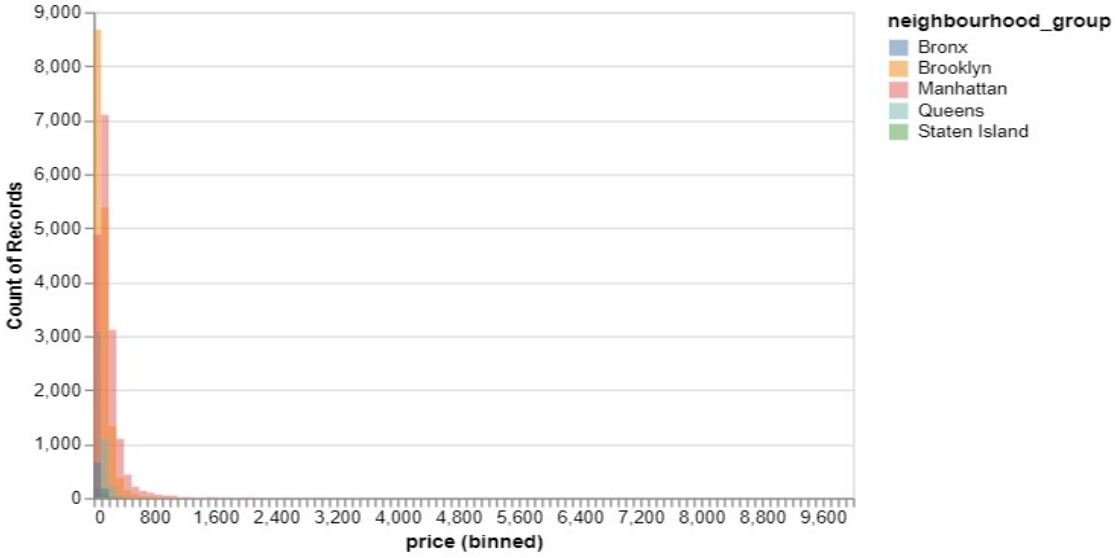

Fig. 1

Figure 1 above shows a distribution of the data by neighbourhood group. There is evident skew in the data, i.e. a few outliers with extremely high nightly prices which affect the mean price of the dataset. Hence, we believe that median is a better measure of central tendency. The plot of data distribution for price <= $1000/night is given below in figure 2 for further exploration of prices by neighbourhood group. From the image below we can see that different

neighbourhood groups have tendencies to be priced differently suggesting that this could be a good variable to use in our model.

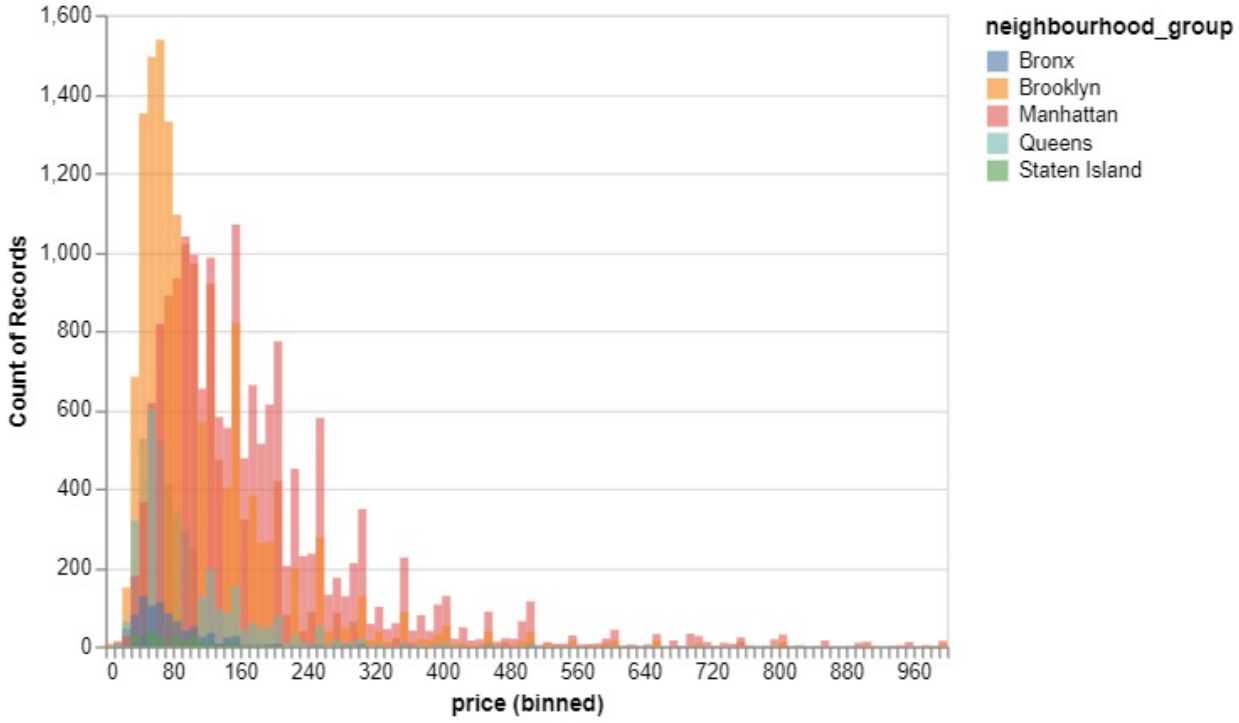

Fig. 2

The price difference by neighbourhood can be visualized on a map of NYC. Figure 3 below shows the listings plotted coloured by price quartile, with a clear distinction that Lower Manhattan has higher prices relative to the rest of the areas around it. One of the key features that is distinguishable from this plot is that while there are general trends in price by latitude and longitude, listings of a similar price are somewhat grouped together, suggesting that a nearest neighbours algorithm could be good at determining the price of a unit.

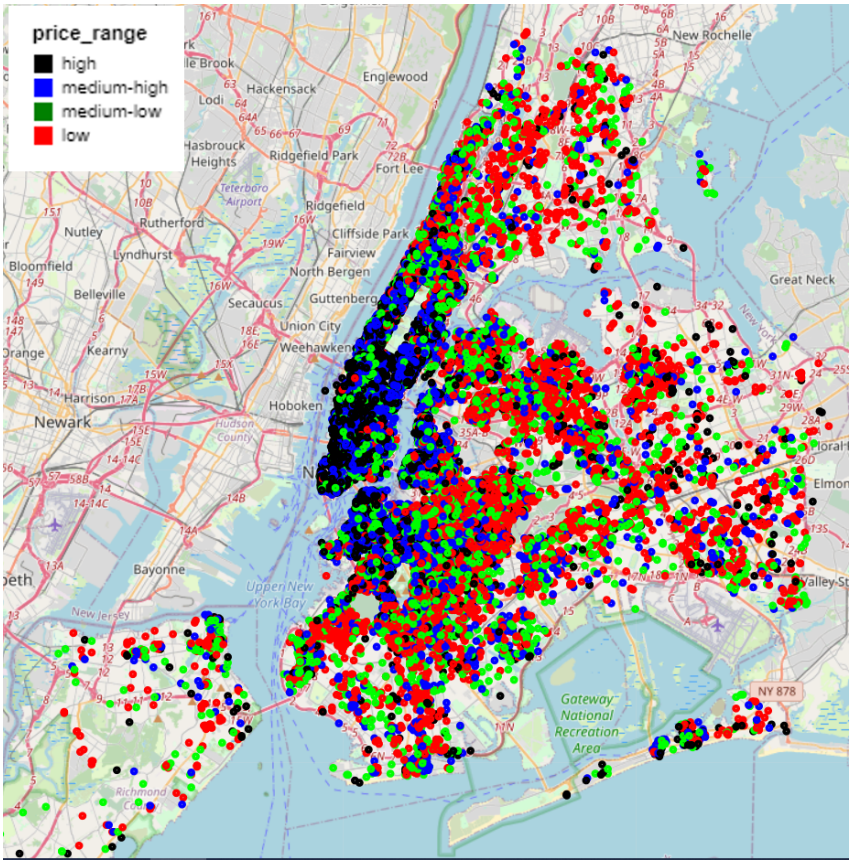

Fig.3

2. Median price and minimum stay per neighbourhood group

neighbourhood_group	price	minimum_nights
Bronx	65.0	2.0
Brooklyn	92.0	3.0
Manhattan	150.0	3.0
Queens	75.0	2.0
Staten Island	75.0	2.0

Fig. 4

The table in Figure 4 illustrates the difference in median prices for each neighbourhood group. Manhattan seems to be the most expensive, followed by Brooklyn, with the Bronx at the lowest nightly price point.

3. Distribution of pricing by type of listing

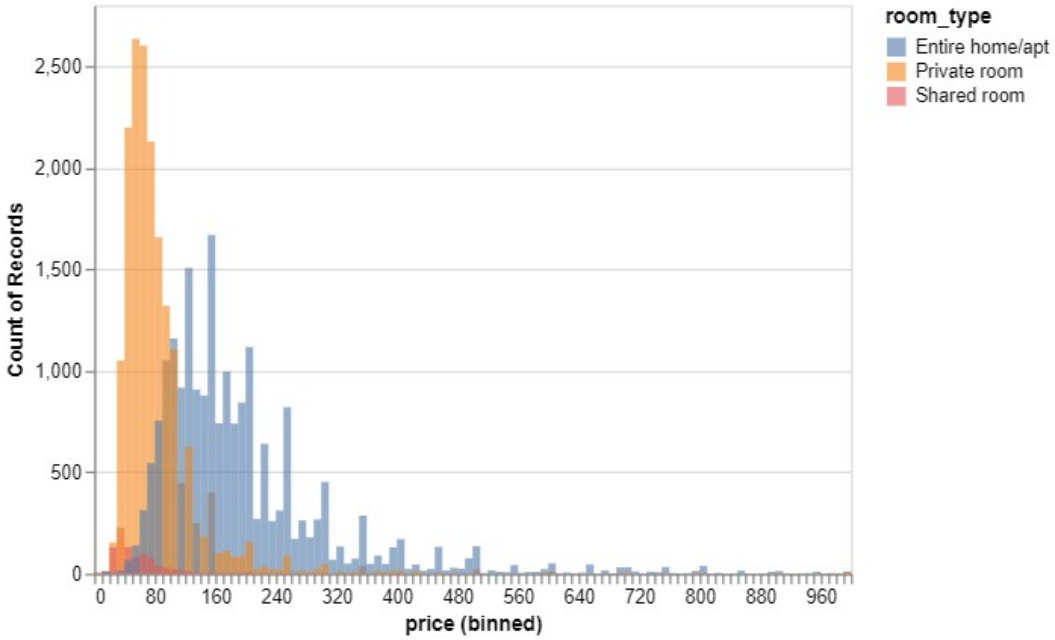

Fig. 5

Across all locations, shared rooms are priced the lowest, followed by private rooms. Entire homes/apartments are, as expected, the highest price. This histogram plot also shows that there is a clear boundary in price between units that are the entire home vs private room, meaning that this will be an important feature in predicting the price of a listing.

4. Correlation of numerical observations

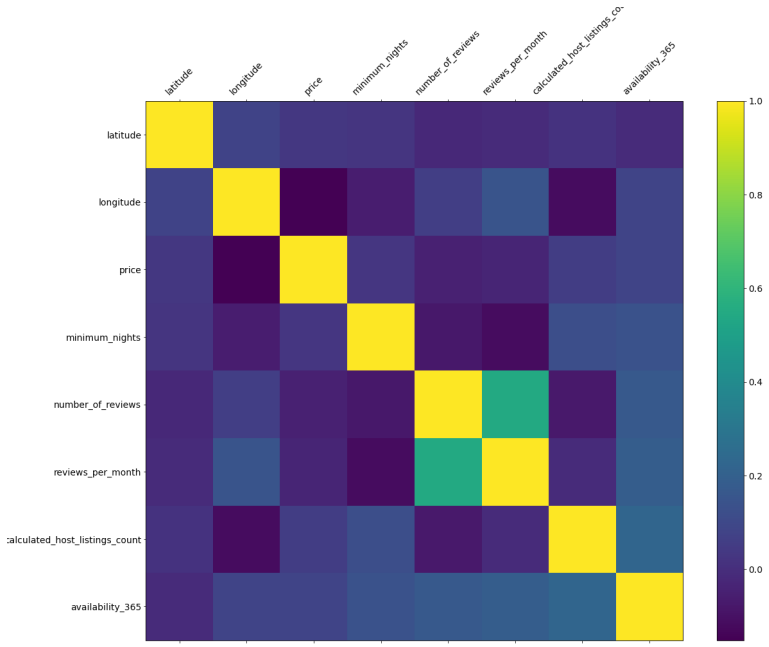

Fig. 6

None of the numerical features seem to have a high correlation with the price, which signifies that a simple OLS regression will likely not perform well on our dataset.

Method

Data Preprocessing

In order to effectively use our dataset without excluding too much information from what is available, we took the following steps:

1. Dropped the columns: 'id', 'host_id', 'host_name', and 'last_review'. 'id', 'host_id', 'host_name' were dropped as they were simply identifiers of existing data which did not add to the descriptive value of the model. 'Last_review' was dropped as it would not be available to a new unit being priced.
2. Split into training and test: We split the dataset into training and test datasets (80-20 split respectively).
3. Explored the 'training' data: Next step we took was to explore the training dataset, key findings from which are listed in the Data Exploration section above.
4. Imputed features: The 'reviews_per_month' column had 10,052 null values. However, valuable information could be gathered from the remaining observations. Hence, we imputed data for said column with constant strategy, filling in missing values as 0.
5. Scaled features: Numeric features ('latitude', 'longitude' , 'minimum_nights' , 'number_of_reviews', 'reviews_per_month', 'calculated_host_listings_count', 'availability_365') were scaled for the purposes of using the dataset in a kNN regression. If left unscaled, the distances in kNN calculations would be disproportionately influenced by a feature with a higher range than the others.
6. One-hot encoding carried out: Categorical features ('neighbourhood_group', 'neighbourhood', 'room_type') were transformed. One-hot encoding was used to create a new binary column for each category in each of the mentioned categorical columns.

A pipeline was used to carry out transformations on both training and test datasets without breaking the golden rule (i.e. to not let test data seep into training data).

Feature Selection:

Upon model testing we found that having neighbourhood_group, neighbourhood , latitude, and longitude was redundant. The variables all measure different granularities of location and hence we chose to drop neighbourhood and neighbourhood_group since these create bins out of the location. The decision to keep latitude and longitude was made as we believe within neighbourhoods, the proximity to landmarks, public transit, tourist destinations etc. might play a role in pricing, which would be better captured by a more granular measure (i.e. geographical coordinates).

Dummy models:

A dummy model was created to extract a baseline for our final model. A dummy regression model was created with strategy mean (i.e., it predicts all prices as the mean price for the dataset). This model resulted in an accuracy close to zero on the training dataset and shows similar values on the test dataset- that is to say it was not capable of predicting the price accurately.

Running the models:

kNN regression, Decision tree regression and Ridge regression models were run on the training dataset with all observations. KNN was selected as a model since it works well with non-linear features in a dataset and could with groupings of observations. The KNN model works to compare observations close to an observation of interest in order to generate a regression line for the dataset.

The Decision Tree regression model was also chosen to deal with the non-linearity of the data, with the idea that splitting on key features could generate an accurate model. In running the decision tree model, it was found that a higher number of splits quickly resulted in overfitting on the training data.

The Ridge regression model was chosen to compare as a baseline for the other two models since it is easy to understand, and both the KNN and Decision Tree models should be able to be tuned for better performance compared to the Ridge model since the dataset is non-linear.

Improving the models:

Upon running our models on the complete dataset, which we found to be skewed during exploration, we found that our model was unable to predict the price accurately.The best performing model using the Randomized GridSearch with cross validation for hyperparameter tuning was the KNN regression model using 100 neighbours with 0.120 and a test score of 0.100 using KNN's R-squared metric. Since R2 is a measure of the variance explained in the model, our model explained 12% of the variance in price of an Airbnb listing, suggesting that there are more factors that affect the price of a listing in our dataset. Given that our dataset did not have certain features listed which may contribute to pricing decisions for the outliers (such as luxury amenities, views etc.), we calculated the interquartile range and removed observations which were outliers for price. Refitting our model on the reduced data set greatly improved performance of the model. Intuitively this result makes sense. At a lower price range people will value an airbnb listing based on different features than for higher priced listings. For example location and room type are clearly important factors in the price of an Airbnb listing, but to differentiate between the wide range of listings containing the same approximate location and room type you need other factors such as the amenities (kitchen, pool, hot tub, etc) which are not in the dataset.

The final model is better suited to predict prices for the wider range of listings with average features.

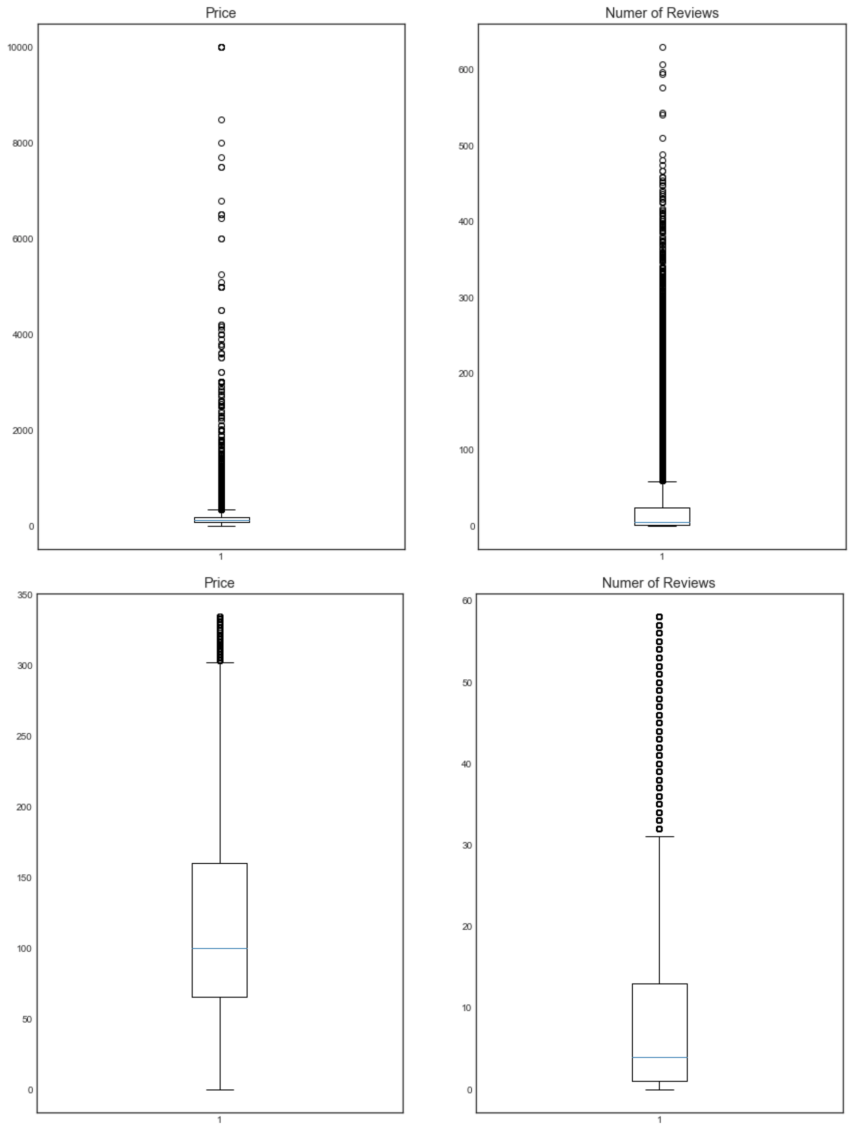

Fig. 8 Distribution of data points before & after removing the outliers

Model Performance

Of the current features, only a subset (type of listing, location) seems to have a significant effect on the price.

Hyperparameter tuning

To tune the hyperparameters in the models, RandomGridSearch CV was used in order to decrease the computational complexity of hyperparameter testing. In all cases the models were run with an initial set of hyperparameters and then based on the average train and test scores for the models the set of hyperparameters was adjusted. In the case of a high train score and a low test score the set of hyperparameter values tested was adjusted to make the model more general. Ultimately, for all models the best average test score after cross validation determined the best hyperparameters.

In the case of the KNN model, adding more neighbours makes the model more specific to the training data while decreasing the number of neighbours makes the model more general. The final value of n selected in the KNN model was 100 which is moderately specific. For the decision tree model, limiting the max_features reduces the complexity and therefore makes the model more general. In order to make the decision tree model generalize well the max_features in the decision tree model were limited to 4 splits. For the ridge regression model, a higher level of alpha made the model more general, since it decreases the coefficients in the linear model.

For our Random Forest regressor model, we investigated the impact of different values for the hyperparameters on that model to achieve the best results. According to our randomized cross validation results the optimum maximum depth and number of estimators were calculated to be 80 and 200 respectively.

Model selection

The best prediction score belongs to the “Random Forest” model with 56% accuracy on the test set. The best accuracy for the other models was less than 55%. Therefore, we chose the random forest regressor model as our final prediction model.

Results

Our final selected model accounts for 56% percent of the pricing decision. While it is a good starting point for pricing decisions, the model is not comprehensive enough to base the pricing decision on in entirety. Adding data such as average customer ratings, popularity (Superhost/Hot Property), types of amenities, number of beds, etc. is needed to improve the model and improve the accuracy of the prediction. Since the features in our model only explain part of the factors that go into the price of an Airbnb unit, adding in additional features that people consider when pricing their listing would improve the accuracy of the model.

References

1. Peck, B. (2022, February 8). *Hotels vs. airbnb in NYC: What's the difference?* Investopedia. Retrieved February 13, 2022, from https://www.investopedia.com/articles/personal-finance/010215/hotels-vs-airbnb-new-york-city-visitors.asp

2. Dgomonov. (2019, August 12). *New York City airbnb open data*. Kaggle. Retrieved February 13, 2022, from https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data