
Airbnb Price in New York City
Kin Sun

Introduction

- I want build a model to predict a property's per night price on Airbnb.
- Prospective clients are any individuals, property management companies, existing Airbnb hosts, new investors—anyone who just wants to start getting involved with Airbnb.

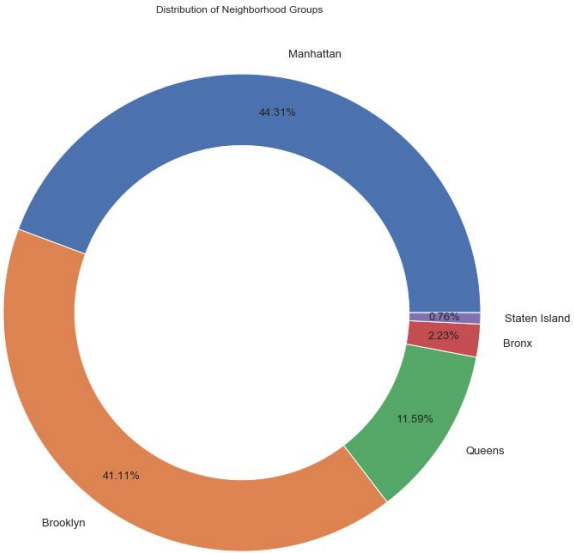
Data Wrangling

- Data source:
  https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data
- Dropped features,
- Investigated Nulls and filled Nulls with 0

Data Description (Final)

Feature	Description
neighbourhood	area
room_type	listing space type
price	price in dollars, per night
minimum_nights	amount of nights minimum
number_of_reviews	number of reviews
reviews_per_month	number of reviews per month
calculated_host_listings_count	amount of listing per host
availability_365	number of days when listing is available for booking

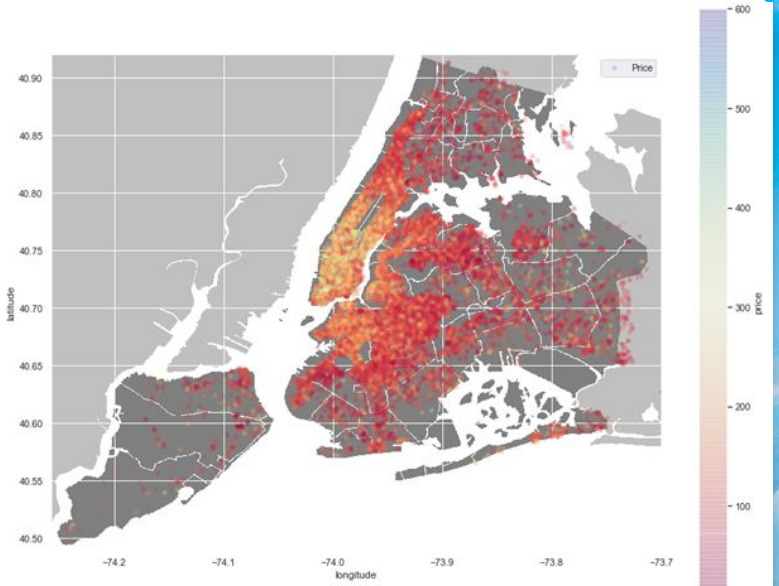
Data Description (Final)



This is the distribution of data in each New York City Borough

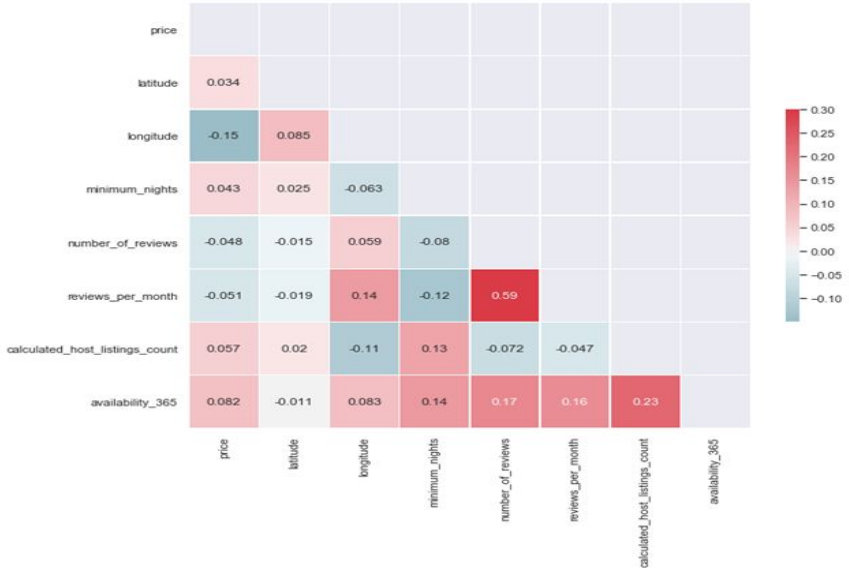
Exploratory Data Analysis

The heat map shows that Downtown Manhattan is significantly more expensive than any other areas



Exploratory Data Analysis (Cont..)

We see that the correlations are not strong between any of our variables, with exception of “reviews_per_month” and “number_of_reviews,” but it is still not very strong



Modeling

- Used pd.get_dummies on categorical variables
- Split dataset to 75% training and 25% testing
- Models used:
  - Linear Regression
  - Ridge and Lasso Regression (Explored)
  - Random Forest Regression

Modeling (Cont..)

- GridSearchCV for hyperparameter tuning
- Model evaluation:
  - Root Mean Square Error (RMSE)
  - R^2
  - Mean Absolute Percentage Error (MAPE)

Linear Regression Coefficient

- **Goal:** Lower RMSE and MAPE, higher R^2
- Linear Regression Model has almost identical error, fit and accuracy
- Based on our criteria, Random Forest is the best model

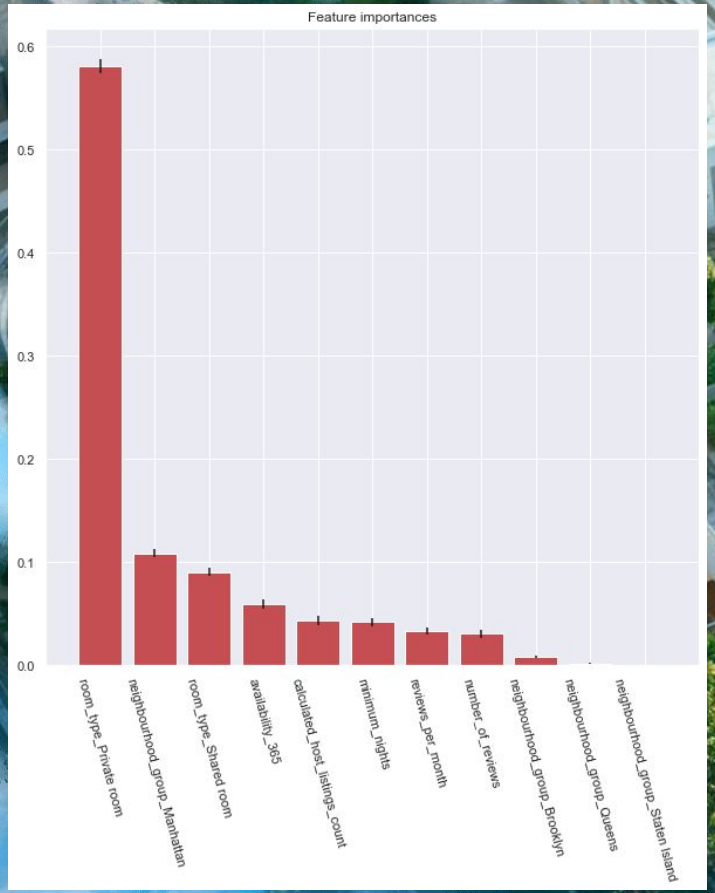
	RMSE	R2	MAPE
LR_Test	53.490615	0.451247	37.961281
Ridge_Test	53.490610	0.451247	37.964141
Lasso_Test	53.489367	0.451273	37.967859
RF_Test	51.004630	0.501069	35.520892

Linear Regression Coefficients

	Features	Coefficients	F-statistics	P-values
0	minimum_nights	-0.242368	32.901	0.0
1	number_of_reviews	-0.047594	37.329	0.0
2	reviews_per_month	-0.825077	93.175	0.0
3	calculated_host_listings_count	0.127663	1036.314	0.0
4	availability_365	0.055885	133.450	0.0
5	neighbourhood_group_Brooklyn	22.851811	1158.846	0.0
6	neighbourhood_group_Manhattan	54.261456	4140.439	0.0
7	neighbourhood_group_Queens	11.299690	1000.244	0.0
8	neighbourhood_group_Staten Island	-1.378366	63.120	0.0
9	room_type_Private room	-80.827102	16462.094	0.0
10	room_type_Shared room	-104.342369	655.038	0.0

- Positive coefficients will have a positive impact on price
- Conversely, negative coefficients will have negative impact on price
- We see that private room, shared room and Manhattan have high impacts

Random Forest Feature Importance



- "room_type_Private room" has the highest importance in our Random Forest model
- Note that unlike Linear Regression coefficients, the feature importances do not tell us whether a feature has a positive or a negative impact

Recommendations

- **Location**: Manhattan tend of have higher listing price, while Bronx and Staten Island are the lowest of the five boroughs
- **Room type:** Price of an entire house/apartment is higher than private room, which is also higher than shared room
- **Worst case scenario**: On the right are the 5th and 95th Percentile of residuals for the models. Worst case scenario in prediction for Random Forest model will either overshoot by $71.26 or undershoot By $102.67. Same logic follows for the Linear Regression model.

Random Forest
0.05 -71.255439
0.95 102.666099
Name: price, dtype: float64

Linear Regression:
0.05 -72.439109
0.95 109.387873
Name: price, dtype: float64

Conclusions and future work

Overall, the Linear Regression Model and Random Forest Model have no overfitting problem, but the R-Squared, RMSE, and MAPE aren't stellar. There are other models or implementations I could try:

- Split the data by boroughs and run models for each of them
- Since we know the coordinate of the listing, we can determine the distance of the closest:
  - Downtown
  - Park
  - Beach
  - Shopping center
  - Crime rate associated with the coordinates' zip code