AirBnB Data Analysis

By: Matthew Edelmann

About; What is AirBnB

AirBnB is a company founded in 2008. It allows home owners/renters to temporarily use their places as a hotel. The user has to put their own value per night for the guest to pay.

Problem Statement

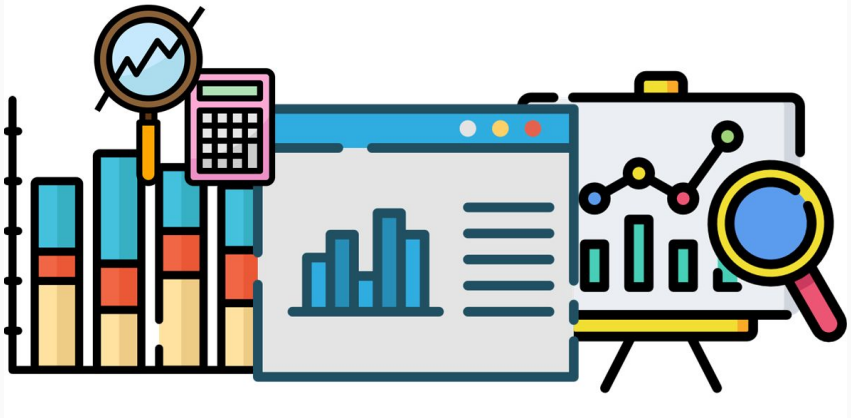
I wanted to see if I could predict the price of an AirBnB in the New York City area. I was aiming at someone who has a space in the New York City area, and may want to use it for an AirBnB. I will try to give them an estimated price based on the borough they live in and other factors. I will also give them the errors with the model, as there are many factors I may not have such as their rent price or living expenses.

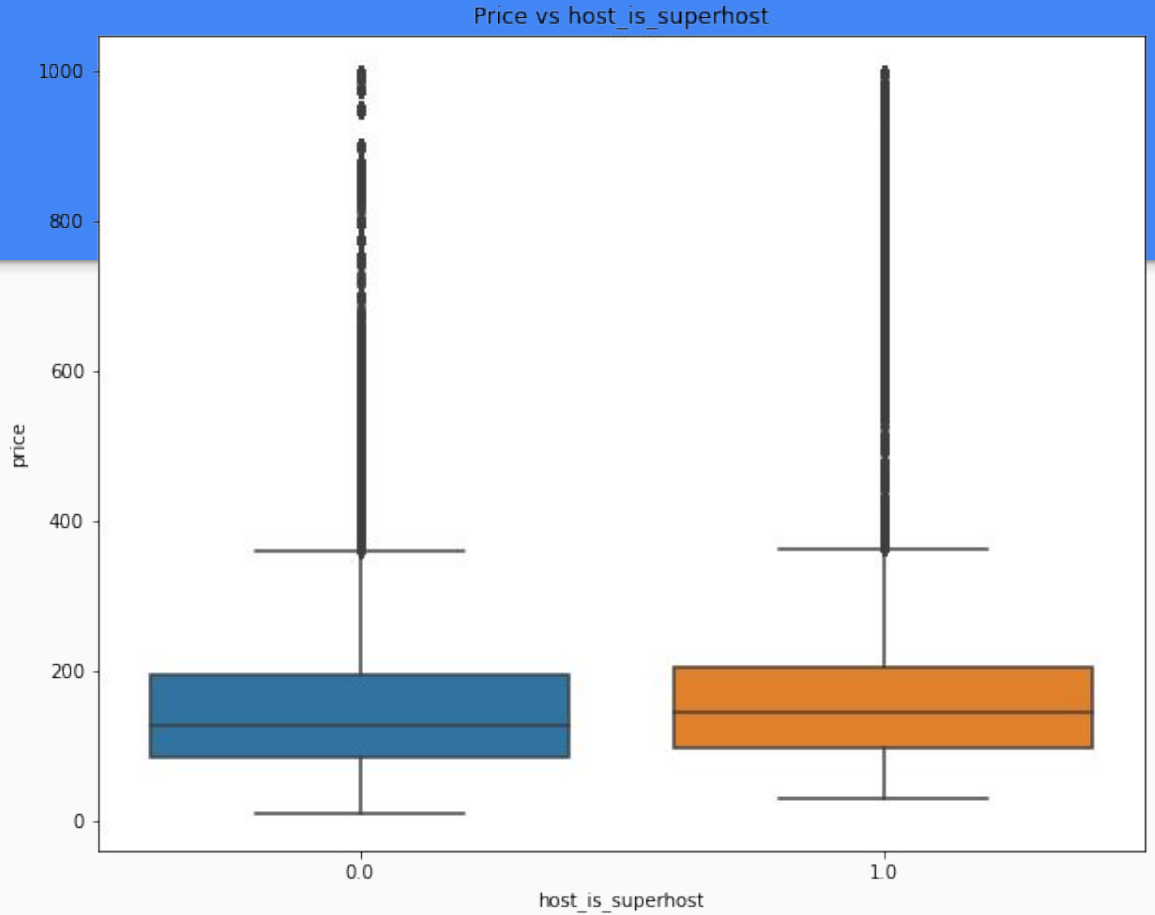
The Data

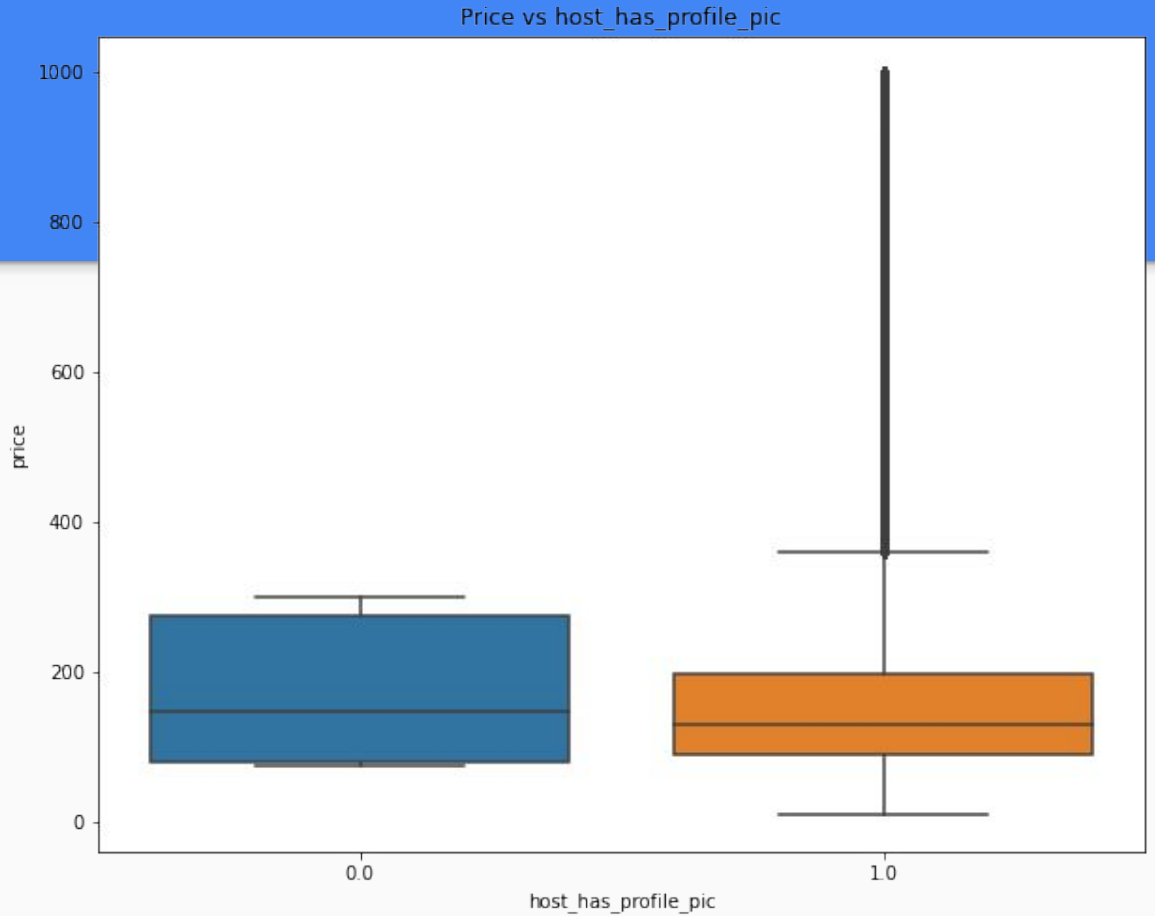
airbnb

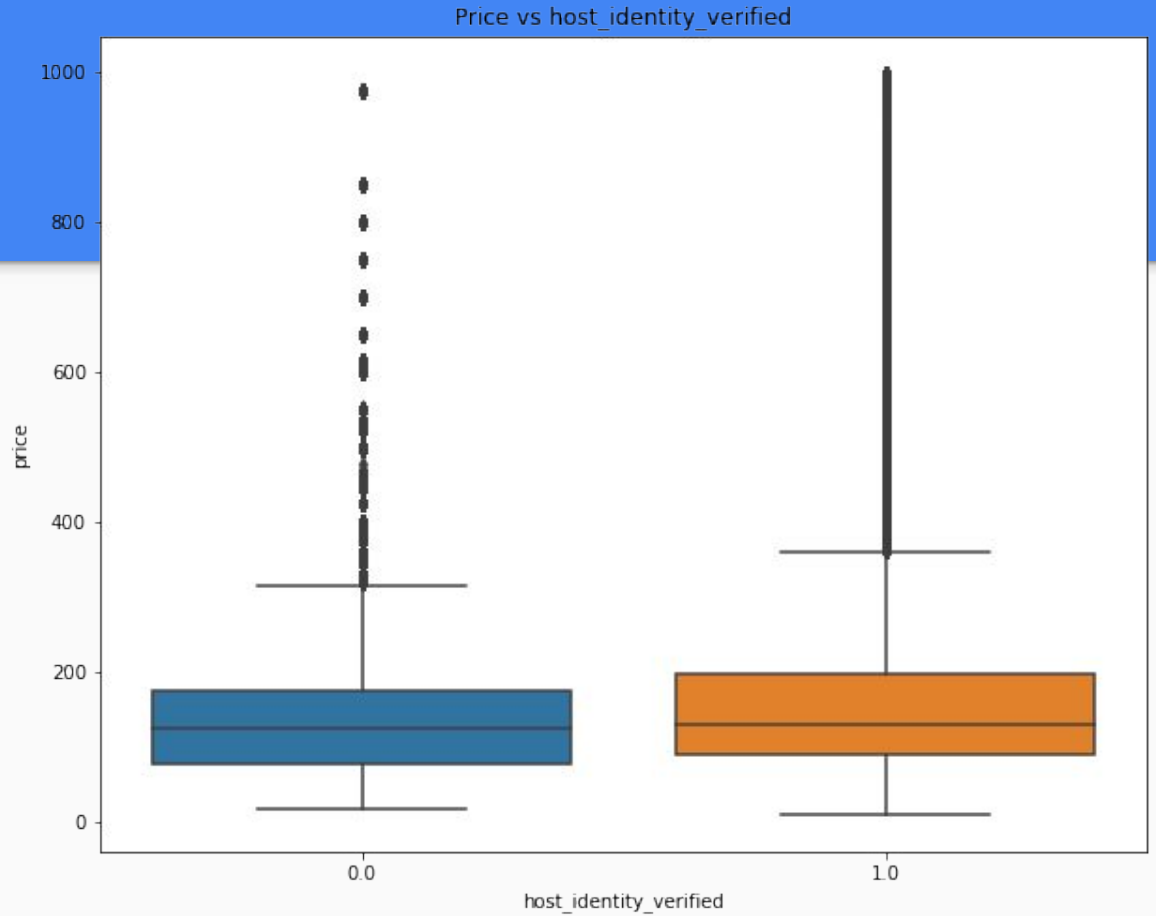
EDA

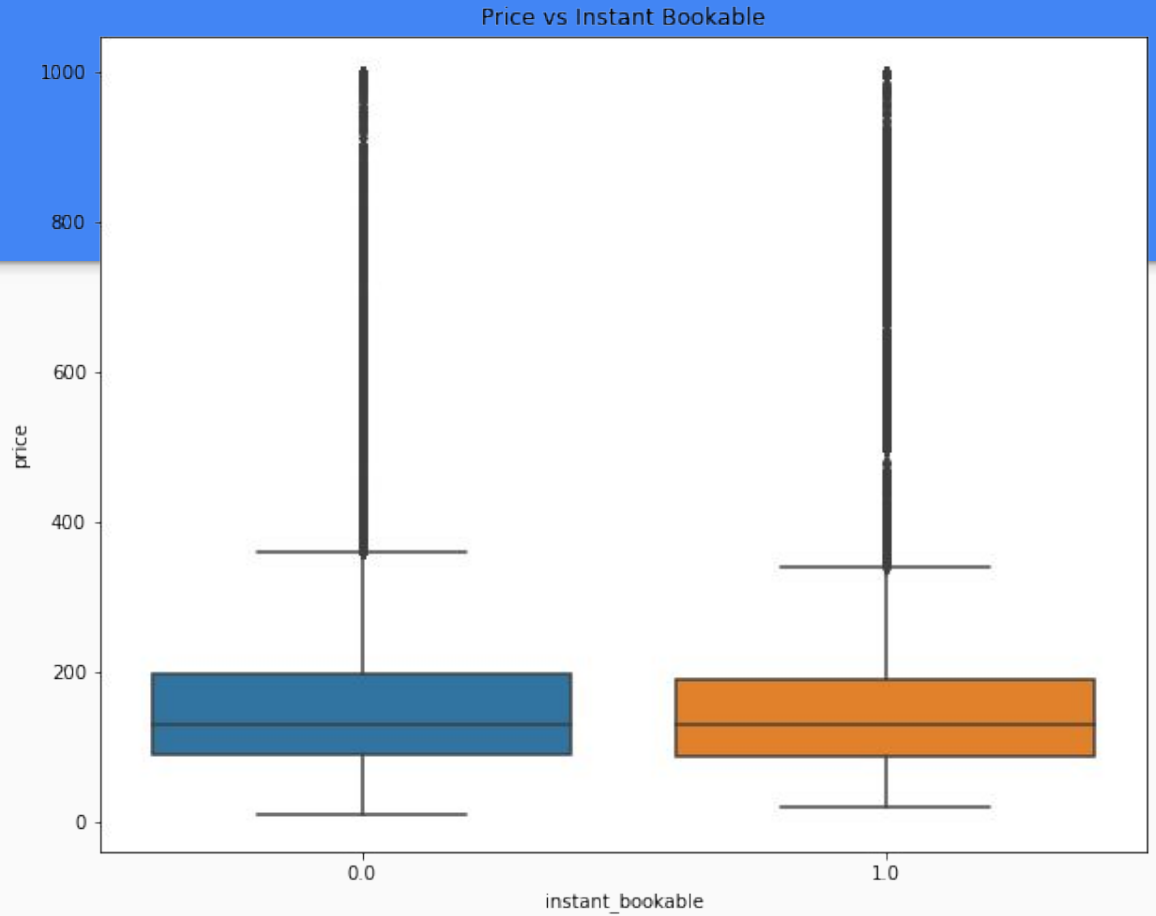
I will be showing some boxplots and barcharts with some interesting finding.

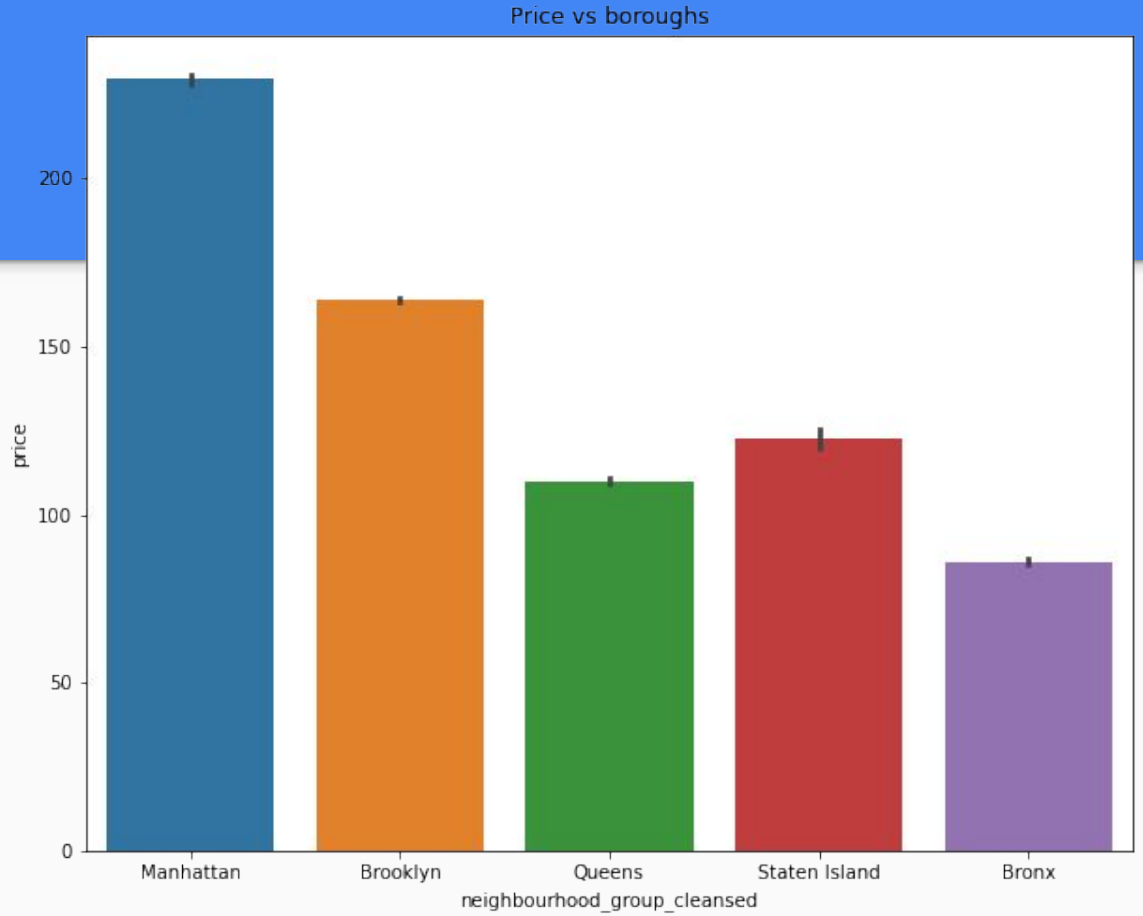


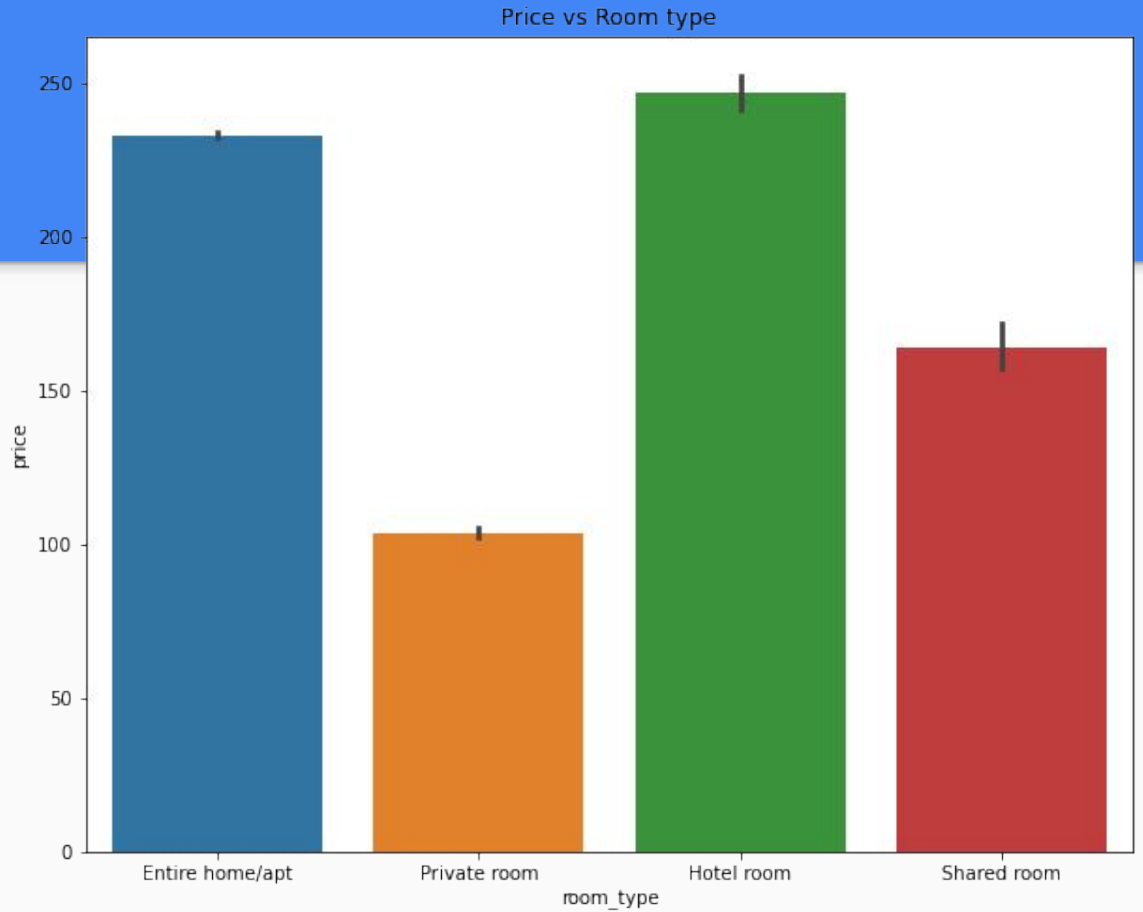

Price vs host_is_superhost
1000
800
600
400
200
0
price
0.0
1.0
host_is_superhost


Price vs host_has_profile_pic
1000
800
600
400
200
0
price
0.0
1.0
host_has_profile_pic


Price vs host_identity_verified
1000
800
600
400
200
0
price
0.0
1.0
host_identity_verified


Price vs Instant Bookable
1000
800
600
400
200
0
price
0.0
1.0
instant_bookable


Price vs boroughs
200
150
100
50
0
price
Manhattan
Brooklyn
Queens
Staten Island
Bronx
neighbourhood_group_cleansed


Price vs Room type
250
200
150
100
50
0
price
Entire home/apt
Private room
Hotel room
Shared room
room_type

Modelling

- I created a different model for each Borough
- I made 5 types of model for each: linear regression, decision tree, bagging regressor, random forest, and extra trees. I'll be showing the best model for each Borough
- I will be judging the models on R^2, Mean squared error, Root mean squared error, and the mean of the residuals

Manhattan

Baseline:

MSE	RMSE	Residuals
293408.19	541.67	126.11

Bagging Regressor:

R^2:

Score on training set: 0.47280

Score on testing set: 0.59523

MSE	RMSE	Residuals
88183.01	296.96	13.9

Staten Island

Baseline:

MSE	RMSE	Residuals
14305.23	119.6	55.72

Decision Tree:

R^2:

Score on training set: 0.97333

Score on testing set: 0.97445

MSE	RMSE	Residuals
337.14	18.36	7.56

Bronx

Baseline:

MSE	RMSE	Residuals
1481.31	38.49	29.7

Decision Tree:

R^2:

Score on training set: 0.98882

Score on testing set: 0.98901

MSE	RMSE	Residuals
16.11	4.01	2.13

Brooklyn

Baseline:

MSE	RMSE	Residuals
51664.94	227.3	79.99

Random Forest:

R^2:

Score on training set: 0.98578

Score on testing set: 0.98443

MSE	RMSE	Residuals
705.02	26.55	7.7

Queens

Baseline:

MSE	RMSE	Residuals
12372.58	111.23	47.75

Decision Tree:

R^2:

Score on training set: 0.92299

Score on testing set: 0.92385

MSE	RMSE	Residuals
913.73	30.23	18.66


Streamlit

Conclusion

In conclusion, I was able to make a model that did as I wanted it to. My goal was to estimate the price of an AirBnB in the New York area, and have a potential AirBnB'r use it to determine the estimated price of there place. I was able to do this in all 5 boroughs, with all except Manhattan having an R^2 score of over 92%. Overall, I would say the project was a success.

Going forward

If given more time I would try to use GridSearch models to get even more accurate models. I would also try to make the Streamlit app even cleaner looking and a bit more user friendly. If possible, I might try to look for more data to include things like average income of the neighborhood or square feet of the AirBnB. I think that could be valuable information in creating a good model.

Questions?