IEOR 142 Final Project Report

Nayan Chavan, Harry Li, Priya Kamdar, Nirmol Kaur, Ananya Raghavan

Motivation

Given the world we live in today, remote work, traveling, and finding unique stays has become the new norm. People turn to the comfort of others' homes and services like Airbnb as a means of exploring and enjoying stays in new cities. However, one of the main problems people face is not knowing the fair price for Airbnb rentals for different, unfamiliar neighborhoods. This often leads people to being overcharged and not being able to address this on their own.

Our project is based on helping people through the process of booking Airbnbs and helping them save money. Throughout the semester we focused on being able to predict the price of an Airbnb given particular features in different neighborhoods. Our goal was to help people figure out the predicted price for the Airbnb listing given the features of the surrounding area and how it compares to the actual price of the Airbnb rental unit they are looking at.

For this project, we chose to focus on predicting prices specifically in New York City. A study done by the Office of the New York State Comptroller showed that in 2019 New York City had 66.6 million visitors[1]. Now, with the world opening up again, this number is expected to increase. With so many unique neighborhoods and pockets within the city, we determined that it would be most interesting to get data on different factors including crime rates, median rent, and public housing to ultimately predict prices as accurately as possible. Furthermore, we expect that the research completed for New York City is not only relevant for this city, but can also be applied to other locations in the future.

Data

Data Collection

We used two different data sources to create our primary data frame that was used in modeling. The first dataset was obtained from **Inside Airbnb**[2]. This website obtains publicly available data from the Airbnb site. This data is from 2019 Airbnb listings in New York City. The data set includes features such as the listing's name, the host's name, the NYC borough, neighborhood, latitude and longitude, room type, and nightly prices. This data provides us with multiple features to effectively predict the price of a given Airbnb listing in NYC; if we increase the scope of our project, it could allow us to solve questions surrounding whether or not the listing is a value deal when compared to other listings with similar features. The second dataset we utilized is from the **NYU Furman Center[3].** This dataset includes profiles of each borough: Brooklyn, Bronx, Manhattan, Queens, and Staten Island. This contained many different indicators, and we selected five to focus on. These five include the following: Index of housing price appreciation, all property types, Serious crime rate (per 1,000 residents), Median rent, all (2020$), Housing Units, and Public housing (% of rental units).

---

[1] The Tourism Industry in New York City | Office of the New York State Comptroller
[2] Inside Airbnb. Adding data to the debate.
[3] New York Neighborhood Data Profiles – NYU Furman Center

Data Processing and Description

To clean and process the data, we first examined the datasets. Then, we renamed columns for readability, standardization, and subsequent data processing. We also found that many columns of the data were strings, so we converted all those columns to floats to access the quantitative data. After the data cleaning, we appended the indicators from the NYU Furman dataset profiles for each borough onto the first Airbnb dataset and named this 'listing_data'. This final dataset was used for our models. In short, this dataset was made up of publicly available Airbnb listing data for New York City and additional city data specific to each borough.

Analytics Models Explanation

We used five different methods to model the data and explain each method below. Additional information about our model can be found on our project GitHub[4].

OLS Model

After some additional data preprocessing, we split the data into a test and train set and then used the statsmodel library to run OLS. This gave us a very small R^2 value of 0.019, so we calculated the Variance Inflation Factor (VIF) to help tune the model. We found that one of the five independent variables (*index of housing price appreciation, all property types, serious crime rate (per 1,000 residents), median rent, all (2020$), housing units, and public housing (% of rental units*) were perfectly collinear with another which was resulting in 'inf' VIF scores. Therefore, we removed one variable at random ('*public_housing_percentage_of_rental_units*') and re-ran the OLS model. We saw a significant change in the VIF values after removing this variable and continued to remove variables with the highest VIF values until the remaining features all had VIF values under 5. This left us with 3 features: housing units, the serious crime rate per 1000 residents, and median rent. **Despite using VIF for feature selection, our OSR^2 value was only 0.0147 and our R^2 was .018.**

We added back additional variables (latitude, longitude, minimum nights, and availability) to the original list of 5 features to create a better model. These 4 variables were added after examining feature importance, which was calculated as a part of the Random Forest portion. We saw the R^2 value improve to 0.039 with the additional variables! After evaluating VIF scores, we ran a model with the following features: housing units, the serious crime rate per 1000 residents, median rent, latitude, longitude, minimum nights, and availability. This gave us an R^2 of .038 and OSR^2 of .034, which is a lot higher than before adding additional features. Finally, we used the get_dummies function to add in categorical geographic variables such as neighborhood and borough to tune the model. This final model gave us the **highest R^2 value of approximately 0.063 and OSR^2 value of .0633**. The low R^2 and OSR^2 values are most likely because OLS does not perform well with multicollinearity. Therefore, we may have had better performance with models that can better handle this such as Ridge Regression.

Logistic Regression Model

After setting up training data along with test data we were able to develop a logistic regression model. We focused on several features including housing units, crime rates, median rent, public housing percentage, and price appreciation index. We found that our model did not perform as

[4] github.com/nayanchavan/ieor142

accurately as we expected **giving us an accuracy of less than 1%**. The reason for this may be that logistic regression models learn linear decision surfaces that separate classes. It could be possible that our classes may not be linearly separable and therefore may result in lower accuracy. Moving forward, to improve our logistic regression model we could attempt changing and or adding features that may impact accuracy. In addition, we could focus on Hyper Parameter tuning and GridSearch along with carrying out feature scaling and normalization. For the sake of this project, we decided to turn to other tree-based models as they can learn rules from our data.

Random Forest Model

Before creating the Random Forest Model, we dropped some features that had little value to the result such as the hostname and host id. Additionally, we dropped some categorical variables that resulted in hundreds of features such as neighborhood, while still keeping relevant data as part of the dataset. Then we began by using the Random Forest Regressor without any cross-validation. **The OSR^2 value we obtained from this was 0.151**. Then, we dug deeper into this by looking at feature importance. This revealed that the longitude and latitude were the most important features, while the borough was the least important. To produce a more accurate regressor, we ran the Random Forest Regressor again with cross-validation (CV). We configured a 3-fold CV and found the best hyperparameters for the model. Using this method, we were able to **increase the OSR^2 value slightly to 0.164**. Finally, we created a **Random Forest Classifier and with an accuracy of approximately 6.08%**. In the future, we could be able to improve this model by increasing the number of trees used. Furthermore, the OSR^2 value and accuracy values may be low because we need to use more high-quality data and feature engineering in our original dataset.

DTC Model

The Decision Tree Classifier model uses the following features: index of housing price appreciation, all property types, serious crime rate (per 1,000 residents), median rent, all (2020$), housing units, and public housing (% of rental units). **This returns an accuracy of approximately 3.918%**. This model is often used for exploratory discovery. DTC models are considered greedy and deterministic in terms of the way that the data is handled which often results in overfitting of the data and therefore lower accuracy. Some ways that we could increase the accuracy of this model would be to add more data to the dataset and do more tuning to our feature selection. In addition, if we were to perform k-fold cross validation this would likely improve our accuracy as well.

Impact

The impact of our work with regards to the problem that we are trying to solve, of renters being uninformed of the fair price of an Airbnb in the area they are looking to rent, is high because our analysis can help renters choose where and what to rent in New York-based on a variety of important factors to them, such as crime rate, room type, and capacity based on the number of guests, housing type, and cost. The map in our GitHub notebook showing differences in Airbnb prices based on location is especially helpful for renters to identify New York boroughs that would be within their price range to rent, and subsequently, choose a location that fulfills other key criteria they prioritize. To expand the scope of our analysis, we can aggregate the data by housing type and guest capacity, then by cost, and display this in a map for short-term criteria.

This would allow people to look at the information that is more relevant to them rather than needing to still sort through Airbnb options for these two criteria after selecting based on price.

Findings

We discovered that Brooklyn and Manhattan are more expensive than Queens, the Bronx, and Staten Island. Furthermore, the five features we originally wanted to look at ['housing_units','index_of_all_housing_type_price_appreciation','serious_crime_rate_per_1000_residents','median_rent_all_2020','public_housing_percentage_of_rental_units'] actually did not have as strong of a correlation with Airbnb listing prices as we hypothesized. Our model accuracy improved by including geographical categorical variables such as borough and neighborhood.

We believe that this is likely because people looking at Airbnb listings are looking for short-term housing and probably will not evaluate all the aspects of a neighborhood as someone looking for long-term housing would do, such as crime rate, schooling zones, or the median rent, but rather by the closeby activities for people or amenities that they would want to take advantage of. Thus, the density map of prices showing the cost based on location will aid in the decision-making of which Airbnb to rent within a borough.

Future Scope

In this section, we will discuss ways in which we could have improved, refined, and expanded our research.

Improvements & Refinements

We found that it may have been more effective and impactful if we had re-run our models from the perspective of a potential Airbnb renter. For example, with one-hot-encoding on the OLS model, we found that some neighborhoods within boroughs had an extremely high, positive correlation with price. However, the borough itself had a relatively low correlation with price. Due to this, the average accuracy of how our model was performing on the entire dataset was skewed because of large outliers such as this. In hindsight, we could have chosen a specific listing with the features we chose to investigate, manually calculated the price our model equations produced, and then compared the calculated, predicted price to the actual listing price. Furthermore, upon finding that geographic variables improved our models significantly, we could have calculated the actual average price of Airbnb listings for each neighborhood or borough and then manually compared it to our predictive average prices. This may have given us a more realistic evaluation of the quality of the model and how it was proposed to be used in practice.

Expansion of Scope

If we were to expand the scope of our research, we believe that a lasso regression model would have been beneficial. A lasso regression model can reduce the number of predictors and features used[5] and is based on minimizing mean squared error[6]; therefore, we may have been able to determine a more accurate conclusion regarding the greatest influence on New York City Airbnb prices.

---

[5] Penalized Regression Essentials: Ridge, Lasso & Elastic Net - Articles

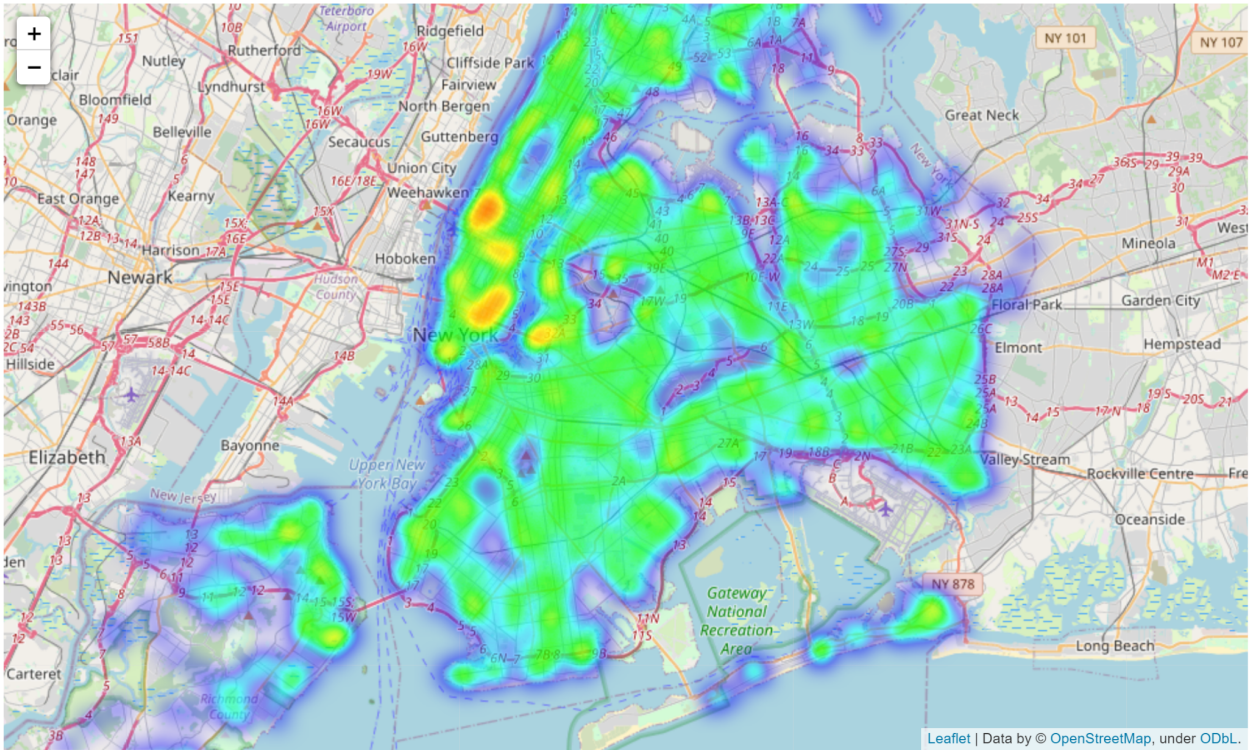
[6] Least Absolute Shrinkage and Selection Operator (LASSO) | Columbia Public Health

Appendix

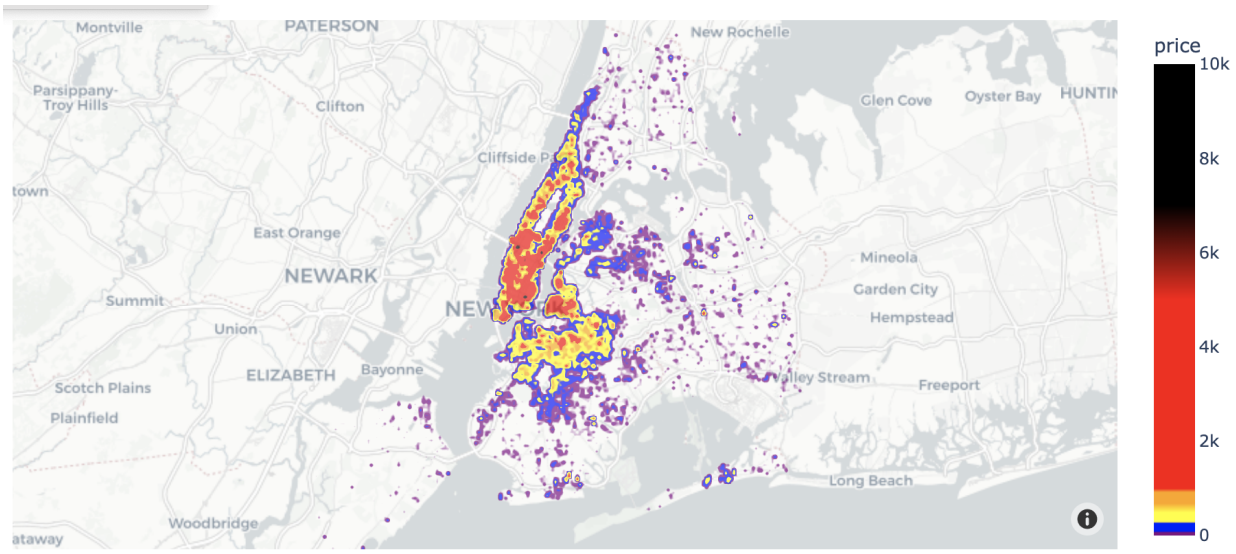
Please find visualizations and other findings below.

Appendix I

Below is a heat map of the density of locations of Airbnbs. Based on the map, there are more Airbnb locations in Manhattan than the other boroughs of New York we looked at.

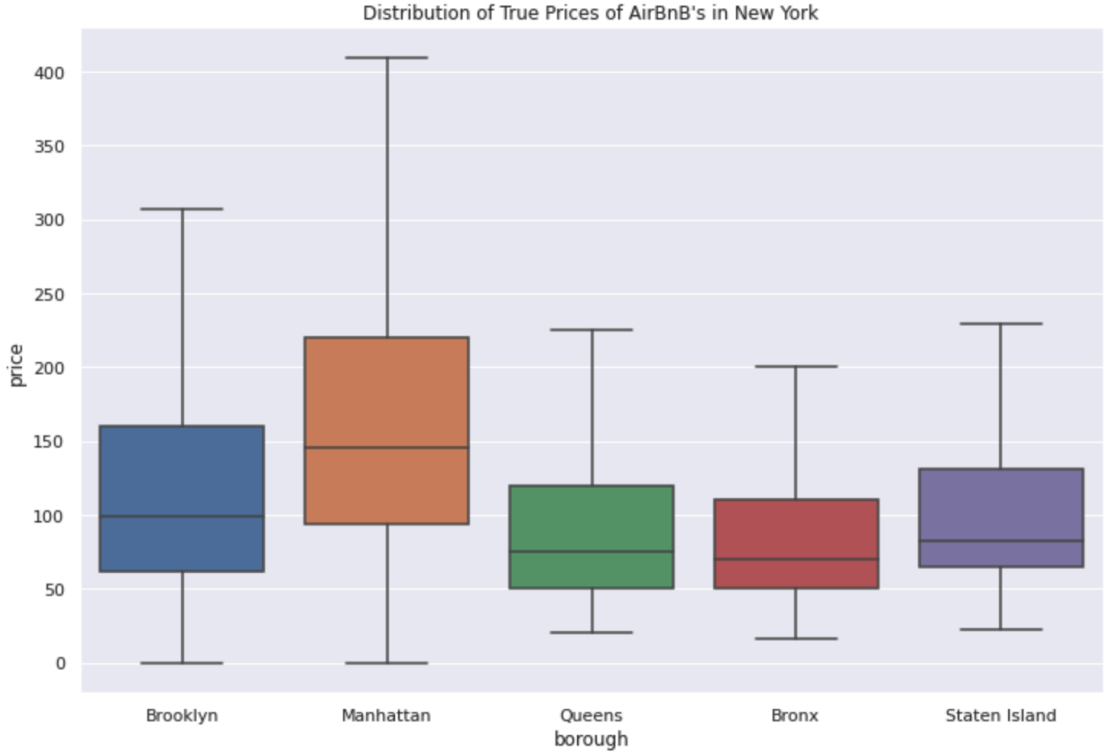


Below is a density map of price to show the average prices for a given area. When looking at this map, people can narrow down the places they are looking to rent based on whether or not the price is in their budget range.



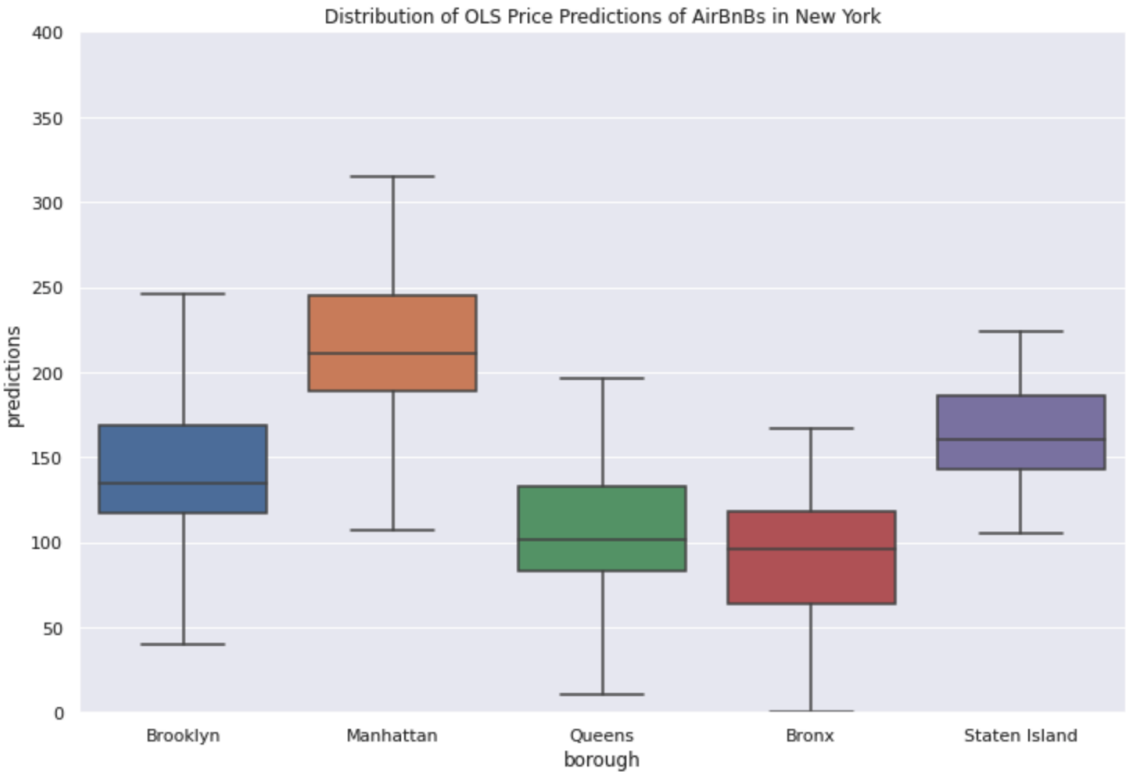
Appendix II

Here's a preliminary chart, displaying the distributions of *true* prices in each borough in New York. By *true,* we mean that this is the distribution of the y_test values.

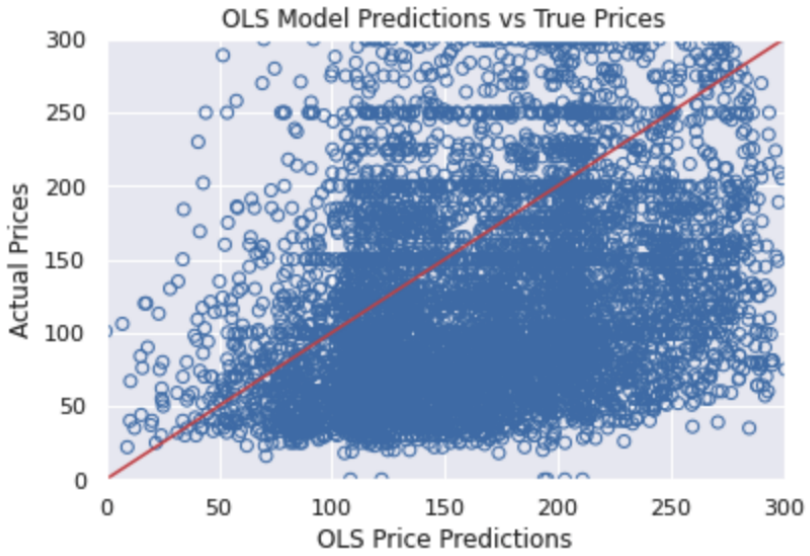


Appendix III

Here is a chart showing the distribution of price predictions of Airbnbs in New York, made by our OLS model.

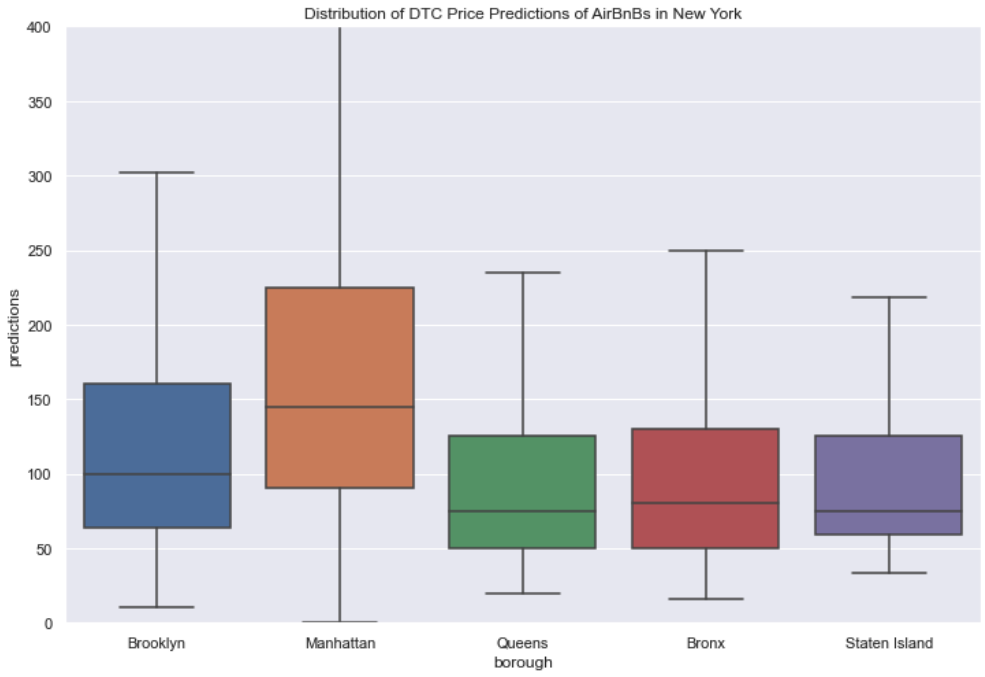


Here is a Scatter Plot displaying the OLS model's predictions, crossed with the actual prices.

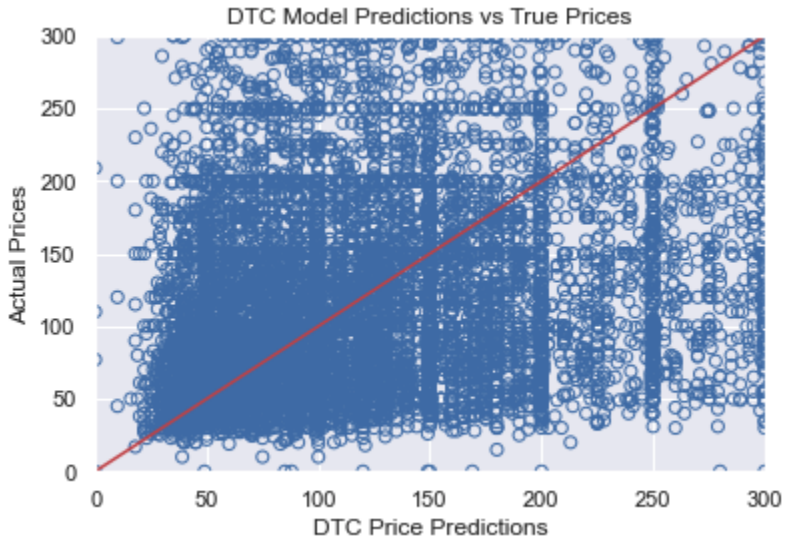


Appendix IV

Here is a chart showing the distribution of price predictions of Airbnbs in New York, made by our DTC model.

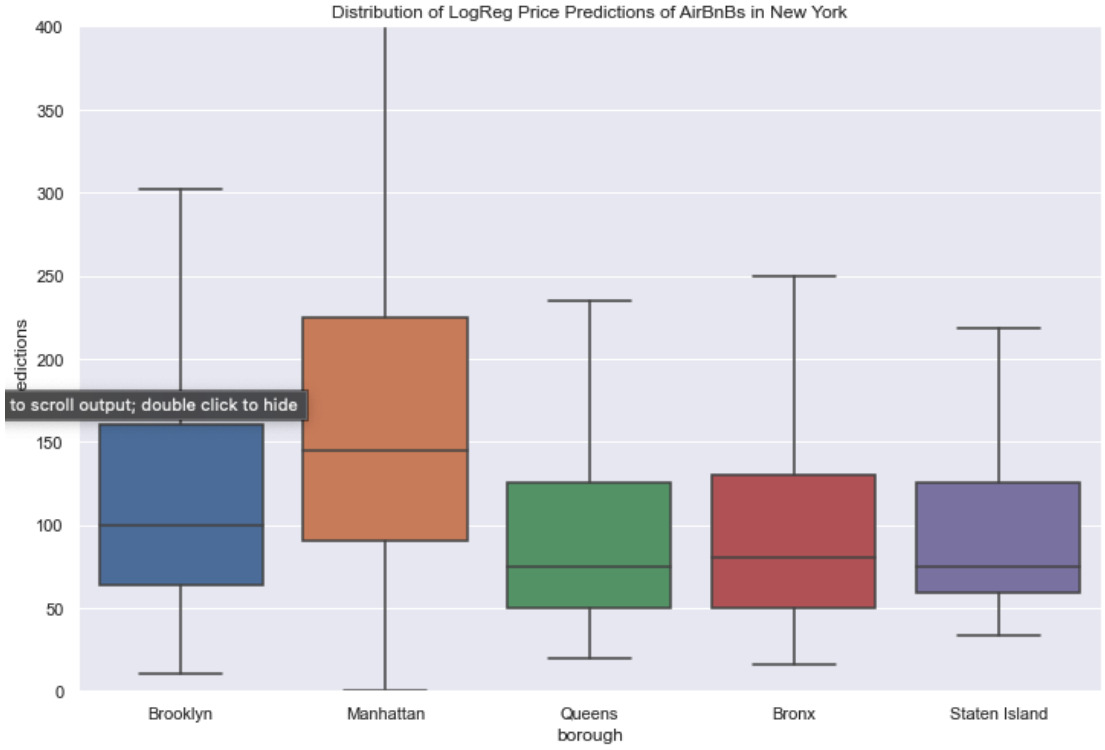


Here is a Scatter Plot displaying the DTC model's predictions, crossed with the actual prices.



Appendix V

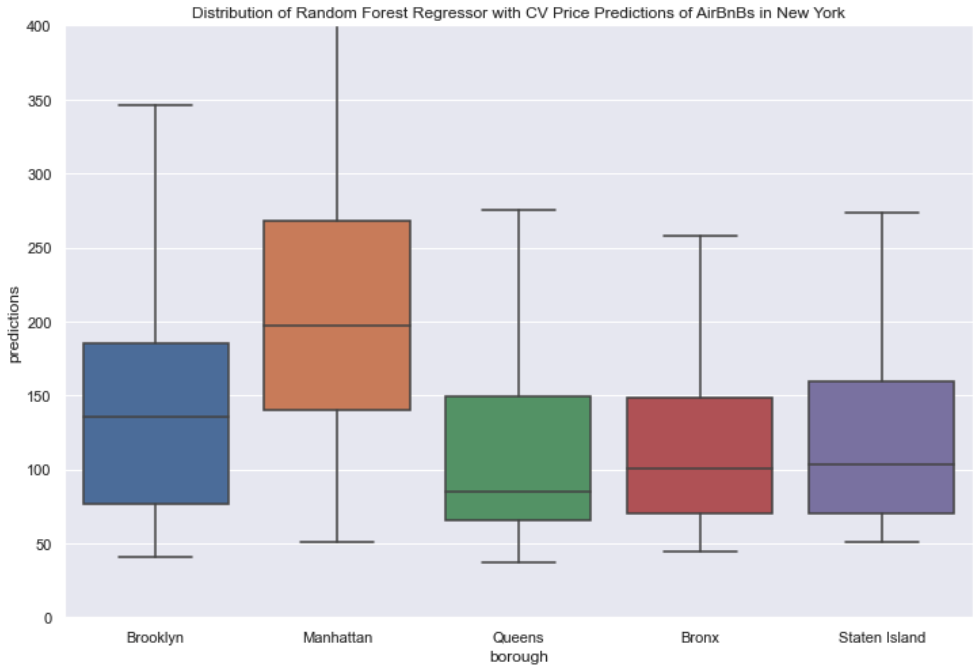
Here is a chart showing the distribution of price predictions of Airbnbs in New York, made by our LogReg model.



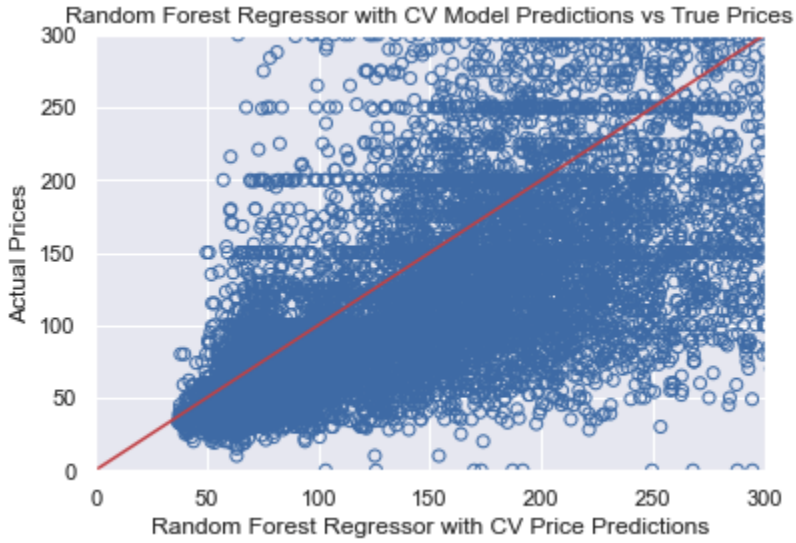
Here is a Scatter Plot displaying the LogReg model's predictions, crossed with the actual prices.

Appendix VI

Here is a chart showing the distribution of price predictions of Airbnbs in New York, made by our Random Forest Regressor with the CV model.

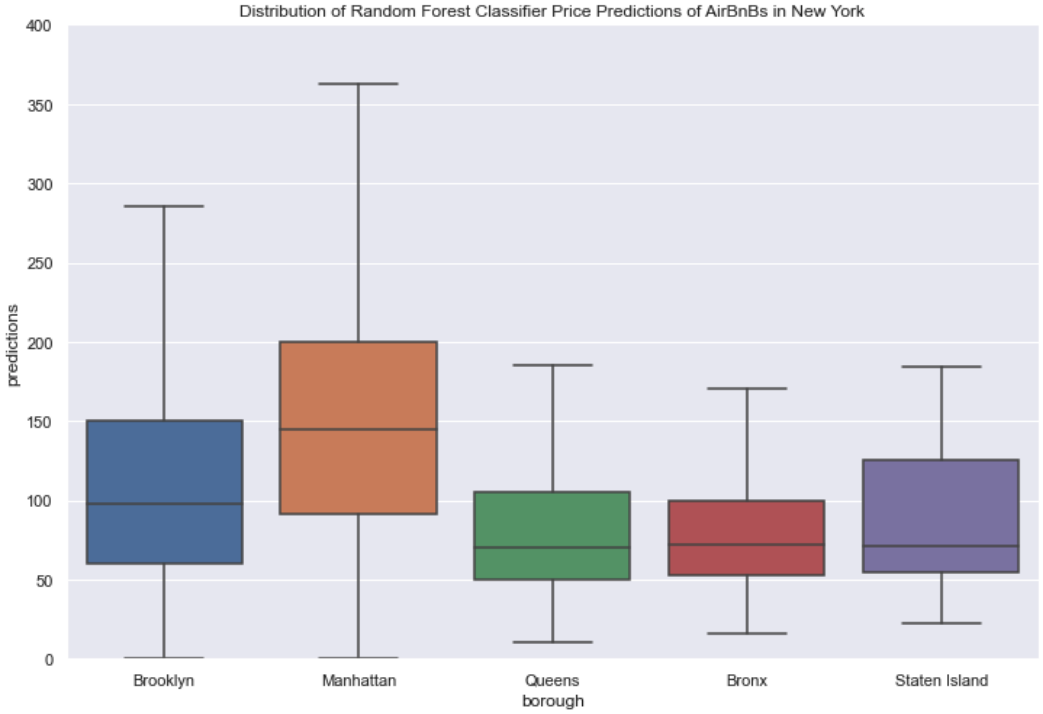


Here is a Scatter Plot displaying the Random Forest Regressor with the CV model's predictions, crossed with the actual prices.



Appendix VII

Here is a chart showing the distribution of price predictions of Airbnbs in New York, made by our Random Forest Classifier model.



Here is a Scatter Plot displaying the Random Forest Classifier model's predictions, crossed with the actual prices.