Predicting Airbnb Prices in New York

Federico Anzil - February 2020

Introduction - Business Problem

Airbnb allows hosts to rent their properties to Airbnb users. Host set their own prices per night. Price per night is the most important variable for defining the hosts' revenue. While **Airbnb hosts** can experiment with different prices to find the 'best price' for their properties, new hosts can have a hard time in finding their optimal price, resulting in loses because of days with too high prices and low occupation, or too low prices and a hidden opportunity cost.

The objective of this project is to create a predictive model of Airbnb prices to solve this problem. Also, model outcome will sever as a basis to better understand pricing determinants or be used for finding low price opportunities by **Airbnb guests**.

Using this models, we will be able to predict the price of a property taking into account features like room types, number of bathrooms, number of bedrooms, how many people is allowed, and so on.

Another objective is to find how do near venues influence pricing. For instance, how will be the price influence is the property is near a coffee shop, a museum or a transport hub? By incorporating these features, our model will be useful for predicting how will the price change after the introduction or closure of a nearby venue.

The Data

While Airbnb doesn't release public data, the private webpage named Inside Airbnb releases data obtained from scraping the official Airbnb website. Inside Airbnb

webpage contains several datasets for large cities of the world, including New York City. Latest datapoint at the moment of writing is from 04 December, 2019.

For each city, we find the following datasets:

- listings: detailed listing data
- calendar: calendar data
- reviews: review data

For our model, we will be using only the listing data. This data contains 106 columns. Some of the columns I will be focusing are:

- Price
- Latitude
- Longitude
- Accommodates
- Review Scores
- Room Type
- Amenities
- Extra People
- Bathrooms
- Bedrooms

	id	listing_url	scrape_id	last_scraped	name	summary	space	description	experiences_offered	neighborhood_overvie
0	2595	https://www.airbnb.com/rooms/2595	20191204162729	2019-12-07	Skylit Midtown Castle	Beautiful, spacious skylit studio in the heart...	- Spacious (500+ft²), immaculate and nicely fu...	Beautiful, spacious skylit studio in the heart...	none	Centrally located in th heart of Manhattan ju
1	3831	https://www.airbnb.com/rooms/3831	20191204162729	2019-12-07	Cozy Entire Floor of Brownstone	Urban retreat: enjoy 500 s.f. floor in 1899 br...	Greetings! We own a double-duplex brownst...	Urban retreat: enjoy 500 s.f. floor in 1899 br...	none	Just the right mix of urba center and local n
2	5099	https://www.airbnb.com/rooms/5099	20191204162729	2019-12-06	Large Cozy 1 BR Apartment In Midtown East	My large 1 bedroom apartment has a true New Yo...	I have a large 1 bedroom apartment centrally l...	My large 1 bedroom apartment has a true New Yo...	none	My neighborhood Midtown East is calle Murr
3	5121	https://www.airbnb.com/rooms/5121	20191204162729	2019-12-06	BlissArtsSpace!	NaN	HELLO EVERYONE AND THANKS FOR VISITING BLISS A...	HELLO EVERYONE AND THANKS FOR VISITING BLISS A...	none	Na
4	5178	https://www.airbnb.com/rooms/5178	20191204162729	2019-12-05	Large Furnished Room Near B'way	Please don't expect the luxury here just a bas...	You will use one large, furnished, private roo...	Please don't expect the luxury here just a bas...	none	Theater district, mar restaurants around her

5 rows × 106 columns

For data related to near venues, I will use the Foursquare Developer API, that allows us to get data like category and rating from near venues, passing a latitude and longitude.

You can see a sample of the data that can be extracted using the Foursquare API in the following image:

```
nearby_venues.head(10)
```

	name	categories	lat	lng
0	East River Esplanade	Pedestrian Plaza	40.704847	-74.004593
1	Crown Shy	Restaurant	40.706187	-74.007490
2	Westville Wall Street	American Restaurant	40.704760	-74.006732
3	McNally Jackson	Bookstore	40.706502	-74.003513
4	South Street Seaport	Harbor / Marina	40.706896	-74.003671
5	Adel's best #1 Halal Food Cart	Falafel Restaurant	40.705609	-74.005599
6	Imagination Playground at Burling Slip	Playground	40.706233	-74.004215
7	Pier 15	Pier	40.704742	-74.003407
8	Trading Post	American Restaurant	40.705940	-74.003983
9	Dig Inn	American Restaurant	40.706106	-74.007290

As we can see, we can get the venue location and category.

Methodology

The Airbnb dataset contained 106 columns but for this analysis only the ones that I considered more important were kept. Some variables, are not related with this analysis, like the URL or the id. For other variables, like the listing summary, or the description of the space, could be included using other methods out of the scope of this report.

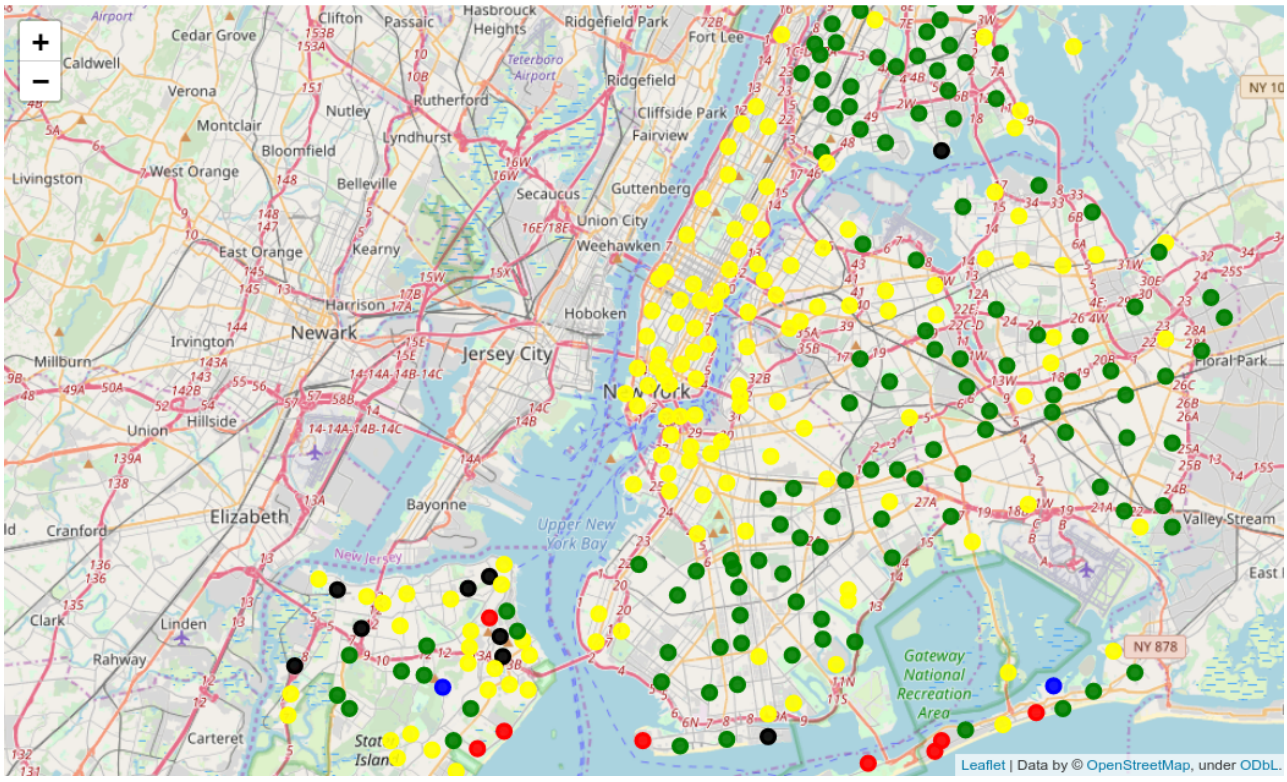
After this initial cleaning, the dataset consisted on 22 features, including latitude and longitude. The location information is important, because we will use this information to merge this dataset with data from Foursquare, to take into account the venues near the listing. I expected that the most common kind of venues where a listing is located could play a role in the price.

Venues Clustering

Using the Fourquare API and a , I created 5 clusters to differentiate neighborhoods according to the most common venues.

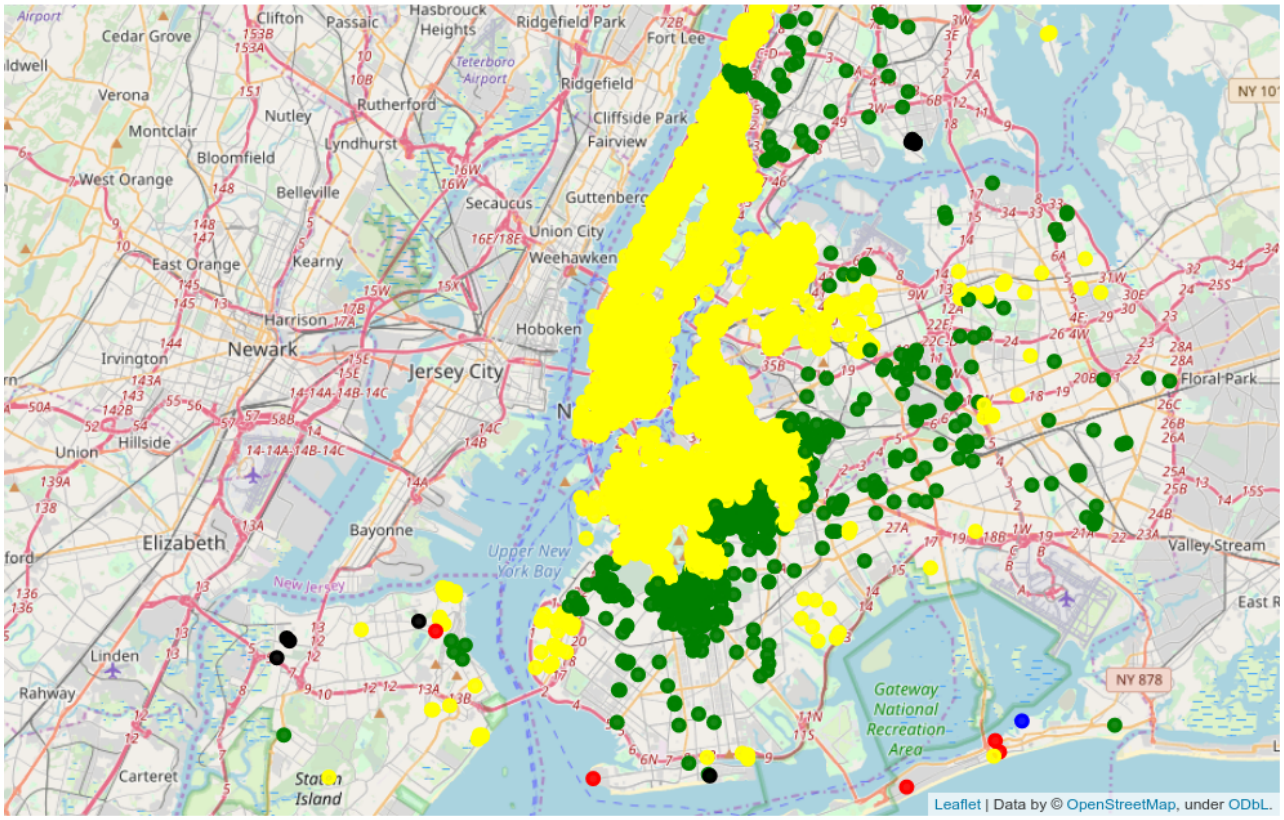
For example, in cluster 1 we have relatively more pizza places, banks and pharmacies. Cluster 2 has tons on Italian restaurants, bars, bodegas and cafes. Cluster 3 has more parks and farmer markets; and so on.

Using the Folium library, I created a map to visualize clusters:



As we can see, there are two big clusters and 3 smaller ones. Further analysis on the topic could focus only on certain areas where Airbnb is more popular (like Manhattan) and take into account venues that could be more important to tourists, like transport hubs, museums and parks. Another approach could take into account venues density or distance to important venues instead.

After running our model, I appended the cluster label to the generated venues dataset, that contains information of venues with latitude, longitude and cluster. This allowed me to merge this data with the Airbnb dataset, and add cluster data to each Airbnb listing.



Most of Airbnb properties are located on clusters 1 and 2:

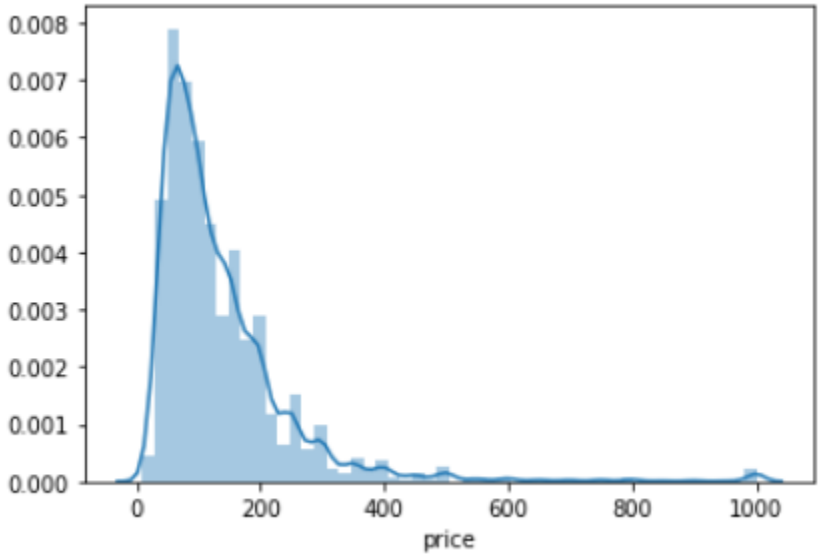
```
cluster_2    36641
cluster_1     7139
cluster_4       77
cluster_0       63
```

```
cluster_3       21
```

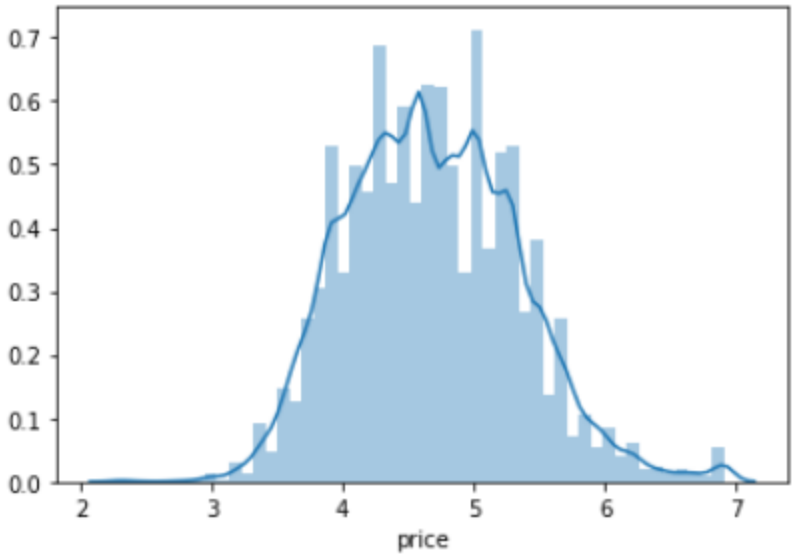
Data Preparation

Logaritmic Transformation

Looking at the distribution of the price (our target variable) we see that a logarithmic transformation could be better to achieve a normal distribution.



So, I applied a logarithmic transformation to the price all the non-dummy numerical features. Afterwards, this is the distribution of the price:



Data Cleaning and Prepossessing

Some of the steps of the data cleaning involved:

- Removing listing with price = 0
- Replacing outliers with price too low (<$10) or too high(>$1000)
- Filtering amenities: I took into account only a subset of amenities, that include security deposit, if extra people is allowed, cleaning fee, air conditioning, parking on premises, and so on. As we will see below, cleaning fee is an important determinant of the price.

For prepossessing, clusters and amenities were converted to dummy variables.

You can find all the cleaning and prepossessing steps on the Jupyter Notebook.

Modelling

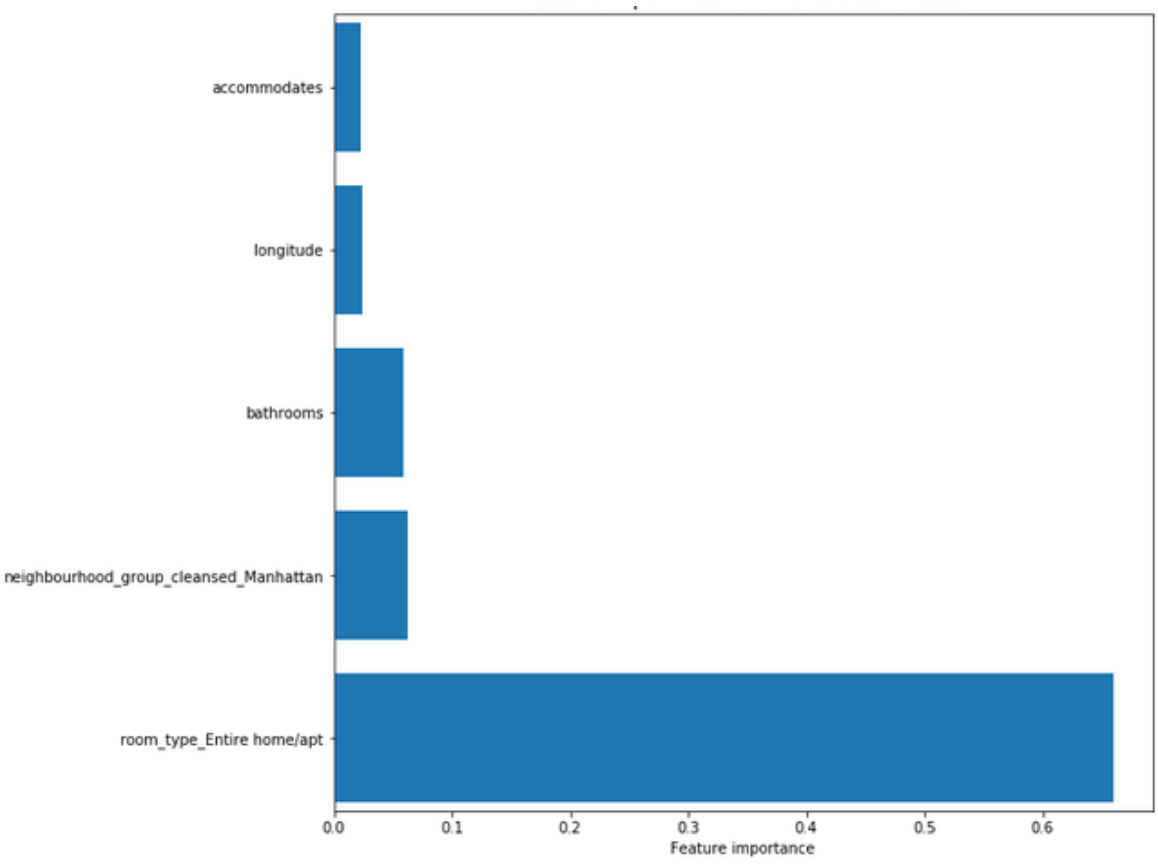
I used a Gradient Boosting method from the XGBoost library. This algorithm gained popularity because of in the data science world because of it's ability to come up with very accurate models on a wide array of business situations. Feel free to find more about XGBoost and how to use it at https://www.datacamp.com/community/tutorials/xgboost-in-python

Results

The model result was able to achieve an r**2 of 0.70

Here is what I found:

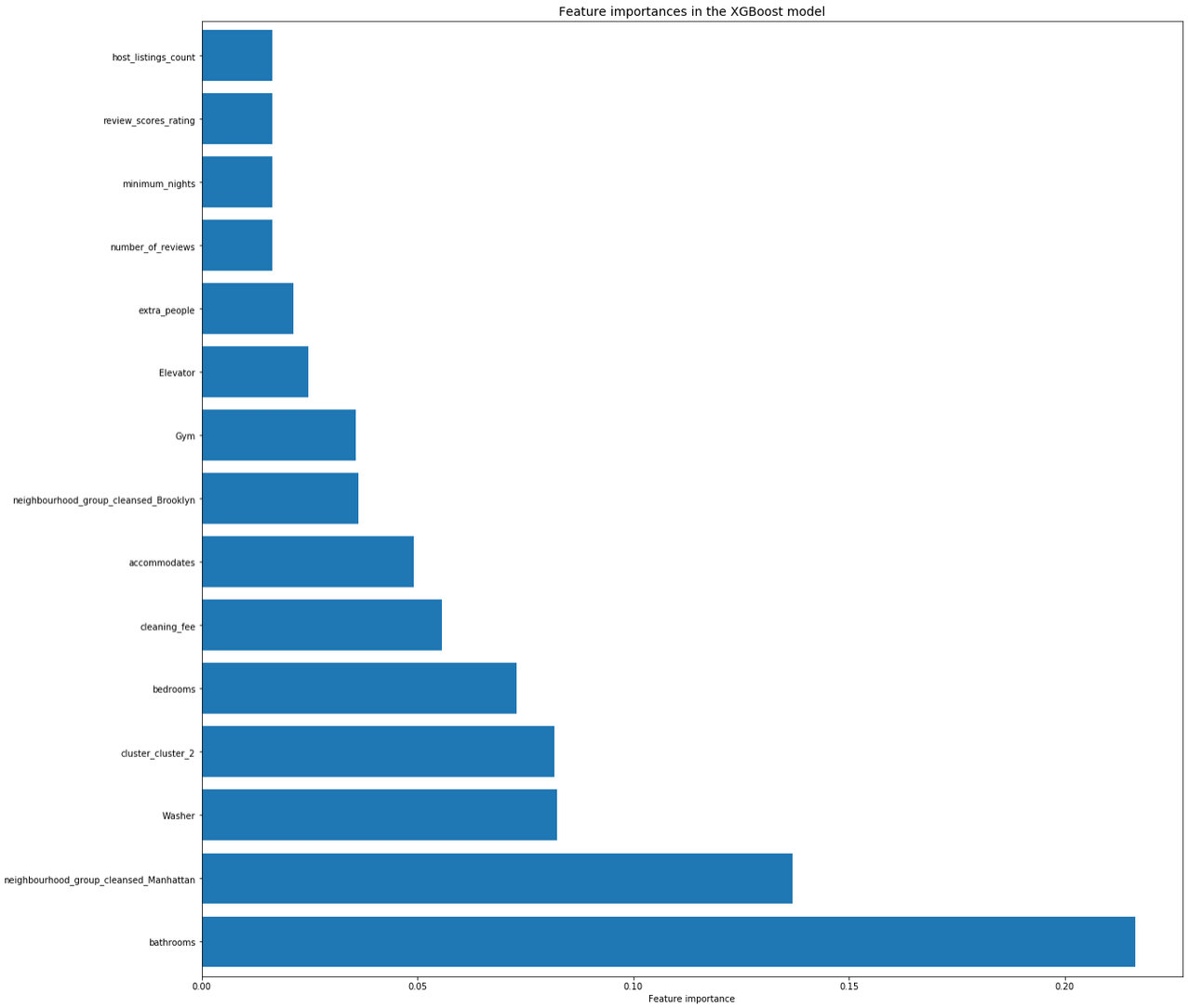
As expected, the most important determinant of the price was the feature ‘room_type_Entire home/apt’ , followed by the neighborhood and the number of bathrooms.



Of course, comparing different properties according if it is the entire home/apt vs shared properties is like comparing oranges and apples, so afterwards I focused only on

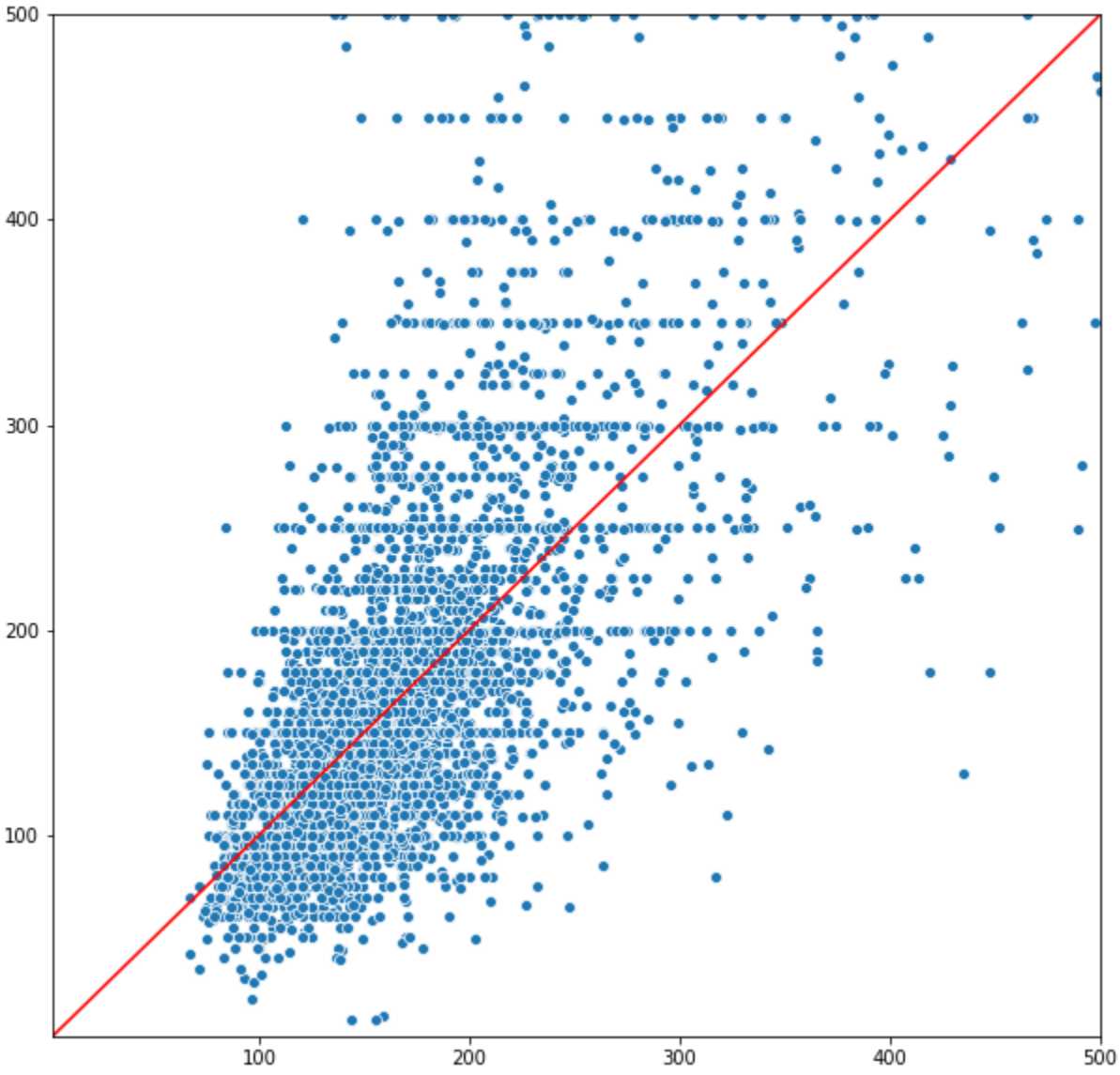
This time, the R**2 lowered to 0.49 but we have more meaningful information from the model.

Taking into account only entire home/apartments, the most important features are number of bathrooms, neighborhood (location), washer machine and cluster(location).



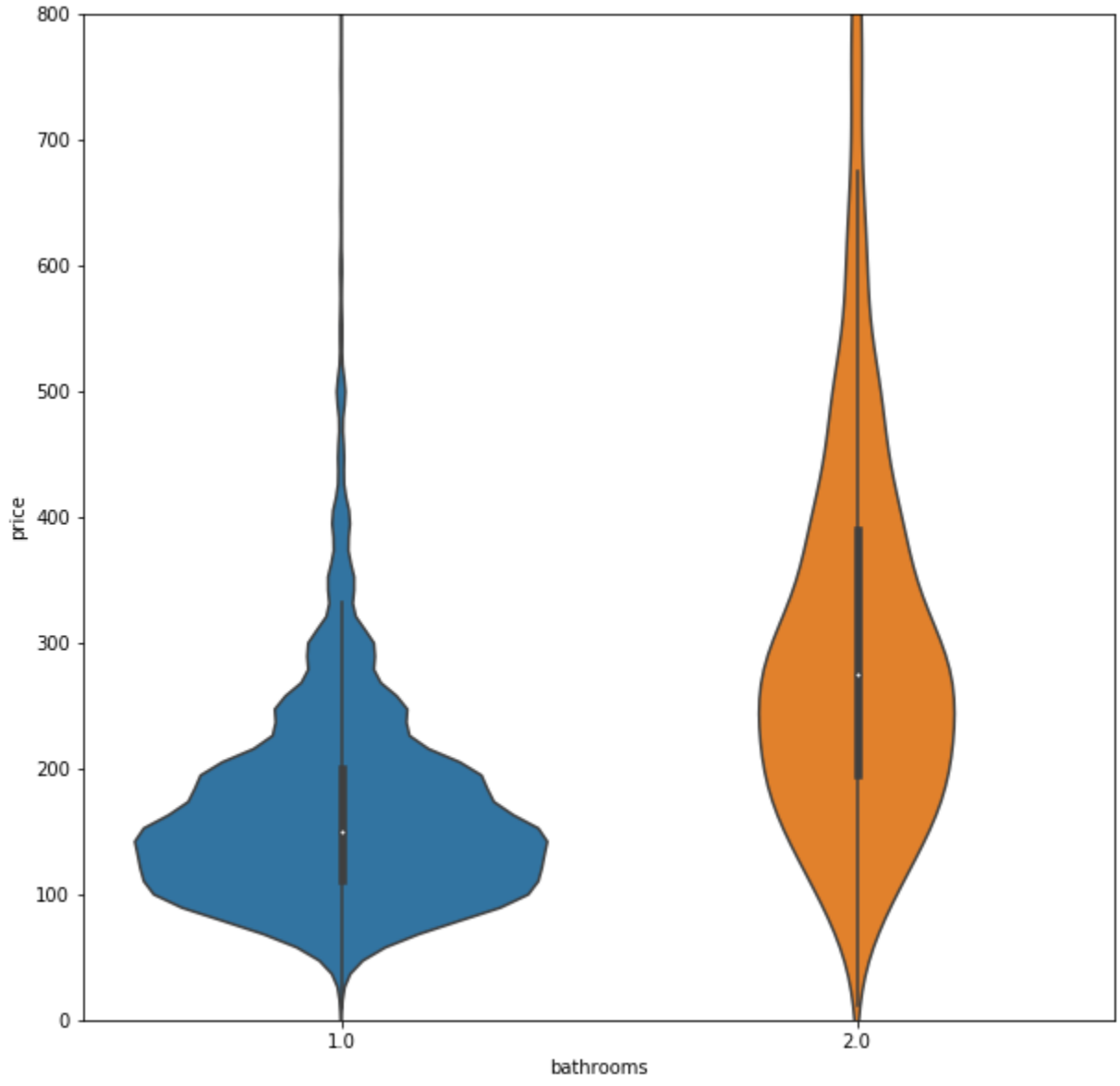
And here is the decision tree:

In the following plot we can see a visualization of the accuracy of the model. In this plot, logarithmic scales were converted back to the original scale:

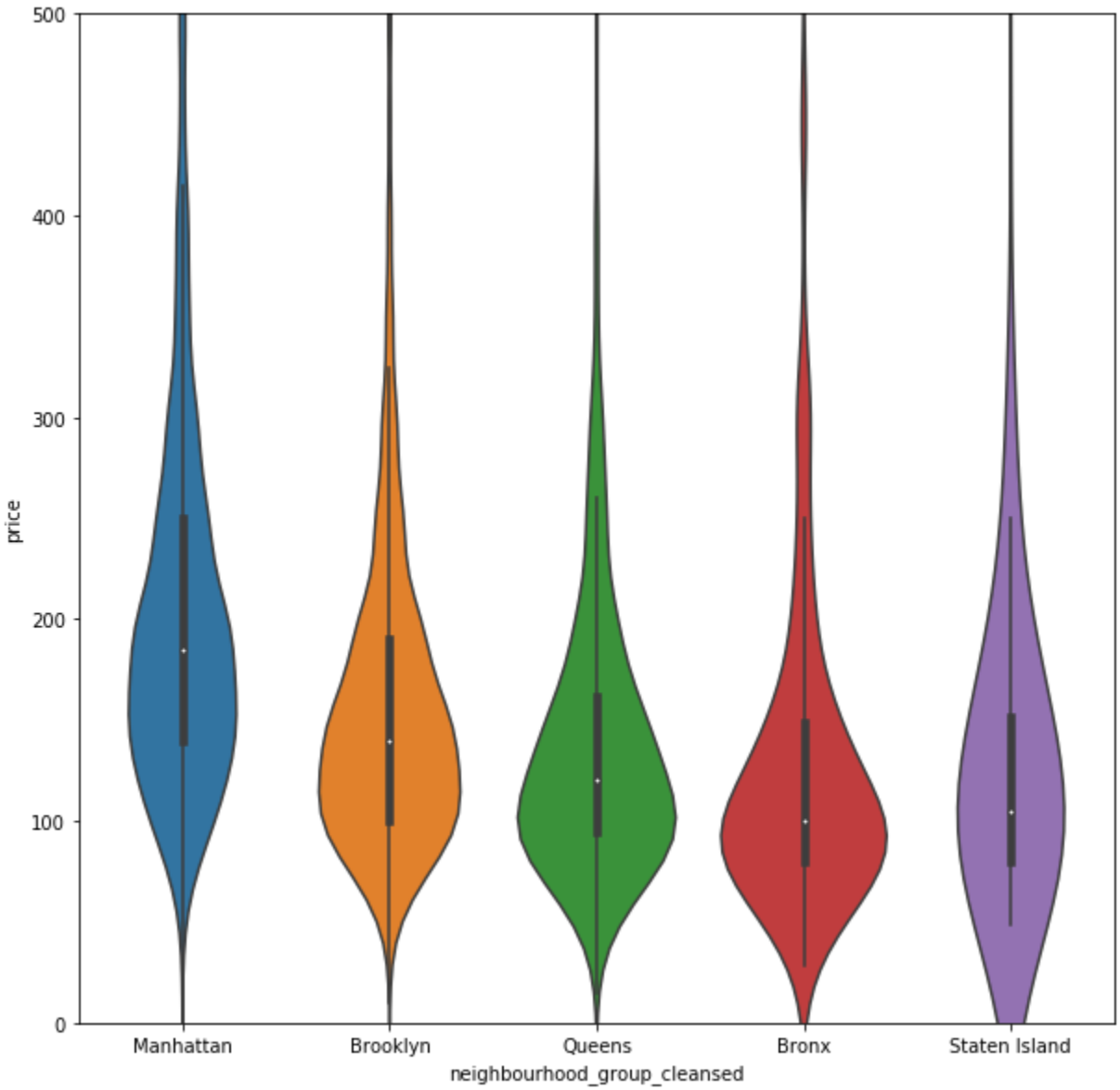


Discussion

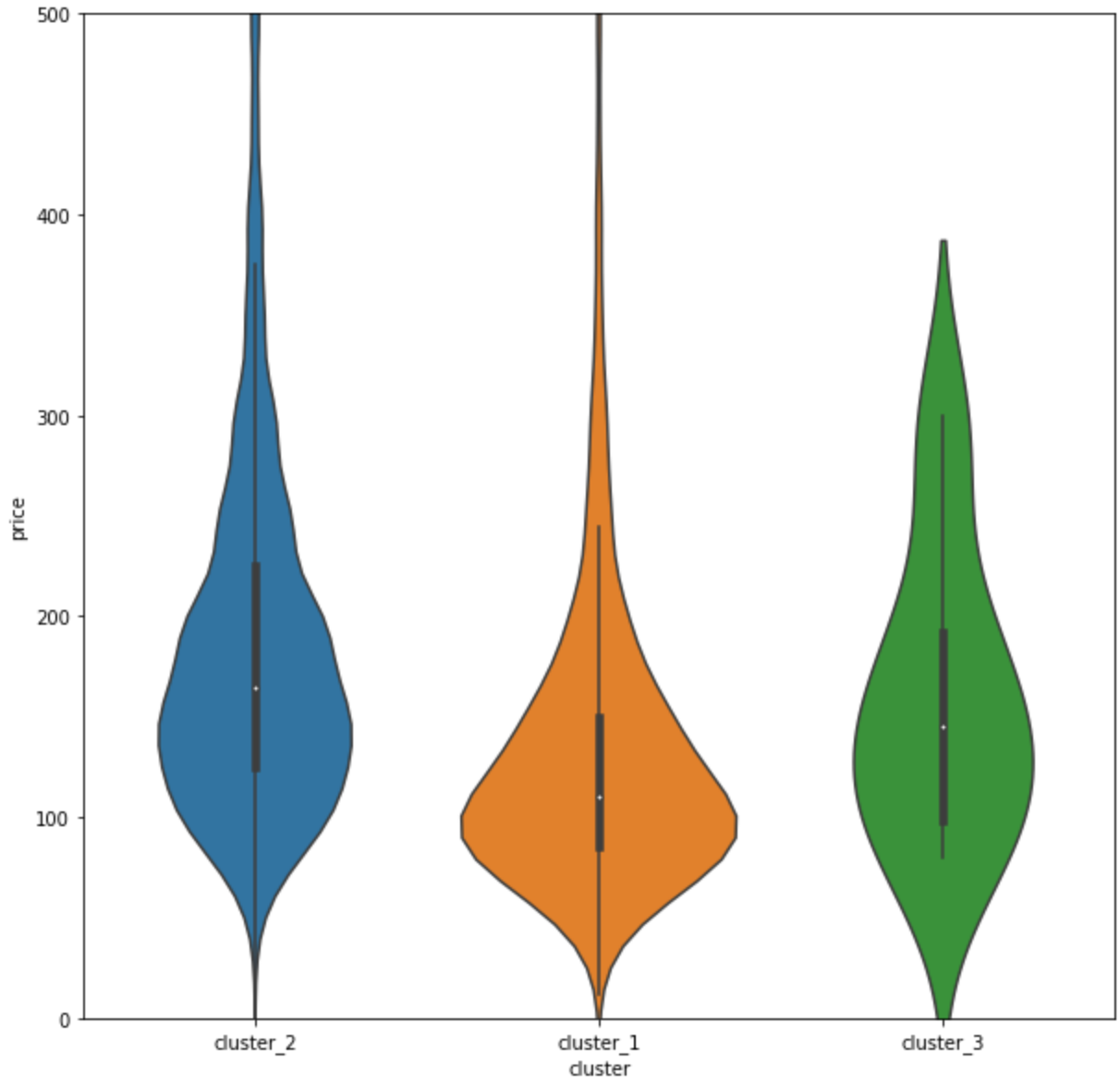
Here we can see the relation between bathrooms and price (only for 1 or 2 bathrooms, and whole apartment or house):



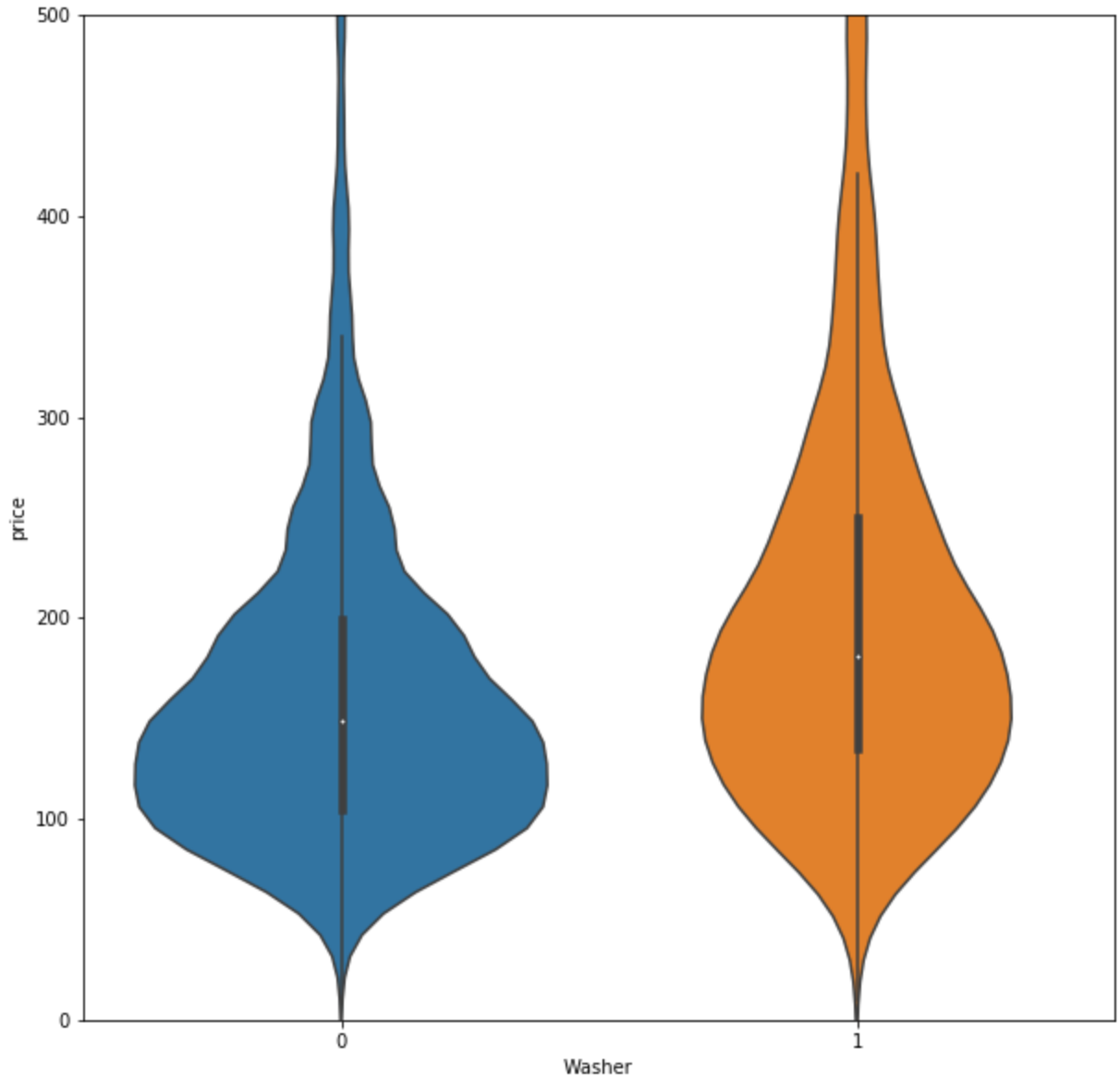
Relation between neighborhood and price:



And the cluster:


500
400
300
200
100
0
price
cluster_2
cluster_1
cluster_3
cluster

And washer:



As we can see, the model allows us to find some important features to predict the price, besides if the property is shared or not.

Interestingly, and unexpectedly for me, reviews scores didn't play such a big role. Neither did the number of people that can be accommodated in a property.

Most important features were number of bathrooms, bedrooms and washer amenity.

Location did play a big role. Prices in Manhattan and in cluster 2 tend to be higher.

Conclusions and Recommendations

The XGBoost model was able to explain half of the price variation given the features. Some important features were location related, like neighborhood and other were amenities like washer machine and cleaning fee.

Some variables not included in this study could add more explanation power to the model.

Future similar modelling should take seasonality into account.

Also, as we could see, location and near venues seem to play a role. It would be interesting to include related variables like distance to important features like museums, airports and so on. And also include other location variables like crime rate per neighborhood.