Apr 30th, 2018

airbnb

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Helping your family attending commencement to choose the best AirBNB near Columbia

Using up-to-date data on AirBNB listings located in the vicinity of the university, we want to draw insights allowing to pick the ideal location

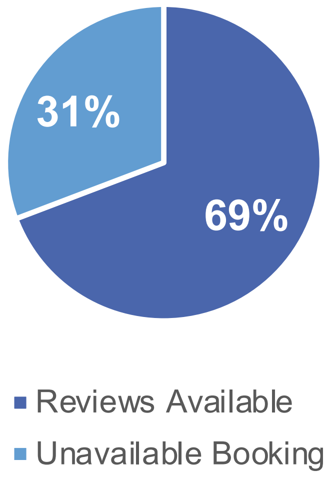
Columbia University
in the city of New York

- We started with more than 50K listings for NYC and decided to focus on 3 neighborhoods
- After that, we made an analysis and built our models with 5,296 datapoints corresponding to rooms available from **AirBNB data**, however, some of them are not longer available
- When we incorporated Ratings Analysis, we found data points for 3,666 of them and performed a new analysis of them
- Additionally, we performed sentiment analysis, price correlation, reviews correlation among other analysis
- Finally, we built an interactive dashboard to help with the decision

Manhattan Upper West Side

Morningside Heights

Harlem



Our first step was analyzing the data we gathered directly from AirBNB databases

The data we obtained, contained +50K datapoints for NYC with features including location, price and amount of reviews

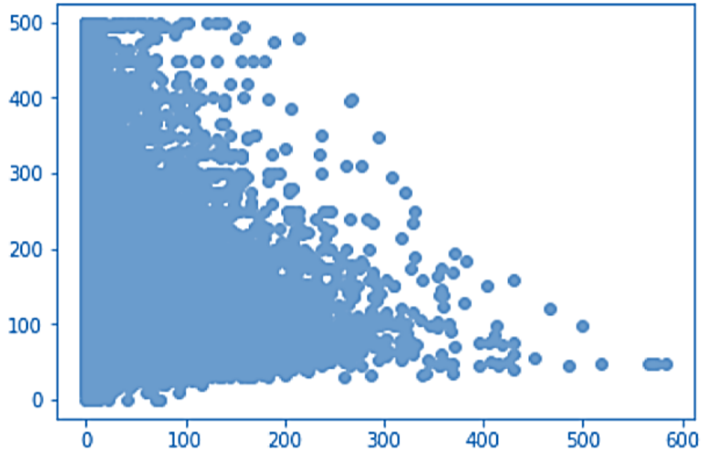
id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
2454	superCondo	2688	Ben	Manhattan	Midtown	40.75552	-73.9677	Entire home/apt	137	7	1	30-01-19	1	1	65
2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.9724	Private room	149	1	9	19-10-18	0.23	8	365
2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.9838	Entire home/apt	225	1	43	02-01-19	0.38	2	365
3330	++ Brooklyn Penthouse Guestroom ++	4177	Jbee	Brooklyn	Williamsburg	40.70856	-73.9424	Private room	70	5	39	07-12-18	0.35	3	290
3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.9419	Private room	150	3	0			1	365
3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.9598	Entire home/apt	89	1	231	17-01-19	4.37	1	162
5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.944	Entire home/apt	80	10	9	19-11-18	0.11	1	48
5099	Large Cozy 1 BR Apartment In Midtown East	7322	Chris	Manhattan	Murray Hill	40.74767	-73.975	Entire home/apt	185	2	70	02-12-18	0.59	1	0
5121	BlissArtsSpace!	7356	Garon	Brooklyn	Bedford-Stuyvesant	40.68688	-73.956	Private room	60	45	49	05-10-17	0.42	1	0
…															

As we can see, there are no features referring to customer experience (ratings or reviews), however, we aimed to find a model to predict price and estimate the best "value" option

We started by processing the data and looking for a correlation between the available features and the impact on the price

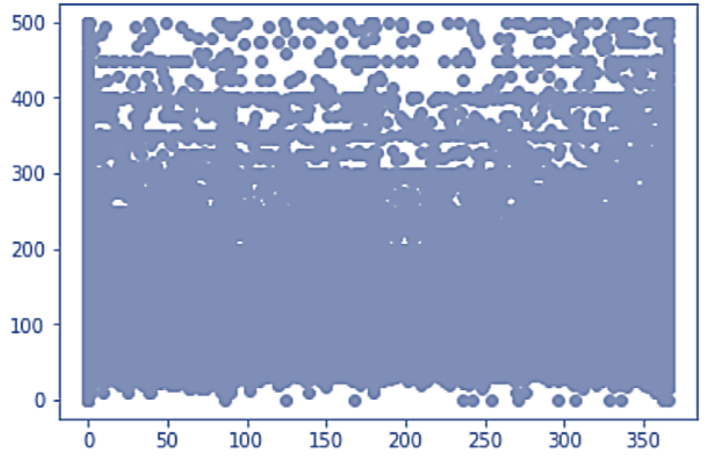
The three main features we looked at initially in order to try to understand the price were the reviews, the availability and the minimum nights required

# of Reviews vs Price



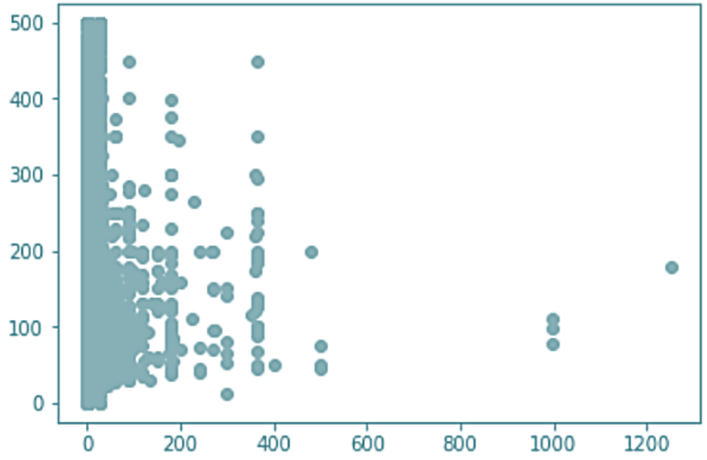
After this analysis, we can see that the number of reviews doesn't explain the price. We can drop this feature as it is not explanatory

Availability vs Price



We conclude that availability is not an explanatory feature for the price, it shows no identifiable pattern

Minimum Nights vs Price



We can see some slight relation in long stays, but for most listings which require few nights, it is not explanatory

After reviewing the original data, we found that potential relevant data was missing and we decided to gather it

We built a scraper for AirBNB and used it on the 5000+ data points corresponding to the 3 neighborhoods analyzed

Room	Stars	Accuracy	Communication	Cleanliness	Location	Check-In	Value
5203	5	10	10	10	10	10	10
5295	4.5	10	10	9	9	10	9
6021	4.5	9	10	10	9	10	9
9668	4.5	9	9	7	9	9	9
9704	5	10	10	9	10	10	10
…							

- We obtained the Rankings across 6 different categories as well as the overall Ranking (Stars)

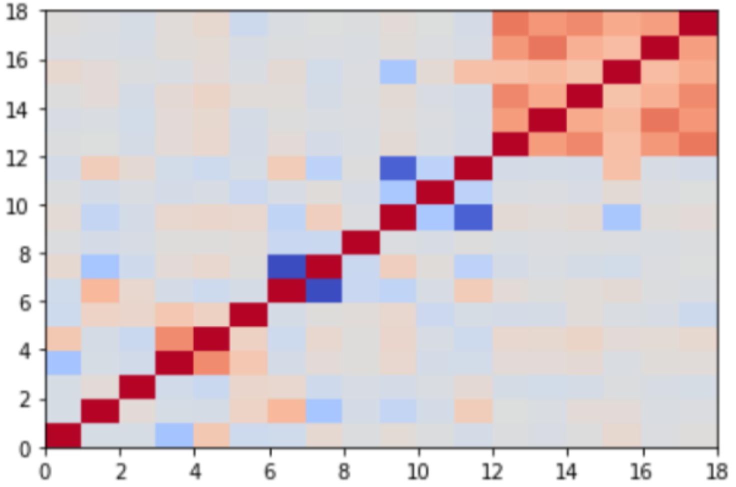
Room	Review 1	Review 2	Review 3	Review 4	Review 5	Review 6	Review 7	Description
5203	"fantastic. great stay and nice people "	"Great location, 5min walk to Columbia campus;...	"Great price for the room and location. My hos...	"Welcoming space, clean, great location, comfo...	"Hands down my favorite Airbnb listing where I...	"MaryEllen is a very nice, hospitable person. ...	"The whole family is amazing, including their ...	sectioned_description":{"access":"Guests will ...
5295	"Great option for a trip to NYC!!!!! Quick res...	"The host canceled this reservation 14 days be...	"Great place. Great location \nSuper Host. \n...	"This a perfect place! Perfect location (Uppe...	"Lena is a very thoughtful and attentive host....	"Great location."	"Lena was a terrific host, and her apartment i...	sectioned_description":{"access":null,"author_...
6021	"I would first like to start by saying that Cl...	"My stay was beyond pleasant! Claudio has a be...	"Claudio's place was great. Only 2 short block...	"This is one of the best Airbnb apartment I've...	"This is an amazing location - across the stre...	"The reservation was canceled 7 days before ar...	"Claudio's apartment was a great find! It is c...	sectioned_description":{"access":"Guests are w...
9668	"This is probably one of the best places to st...	"We met Donaldo and he was very sweet and prov...	"I didn't meet Trip, but his co-host managed e...	"Big applauds to a co-host Donald. He was very...	"pretty comfortable place and accessible to an...	"I'll begin by saying, "you get what you pay f...	"Trip was very flexible and accommodating!"	sectioned_description":{"access":"bedroom, bat...
...								

- We took the latest 7 reviews for each of the rooms as they are the most representative
- We also obtained the Description of the Listing
- With these features we were able to perform additional analysis

Model Performances

We implemented several models to regress the price of listings on collected features; we found out there was no real correlation between price and any of the explanatory variables

```
plt.pcolor(data_borough.corr(),cmap='coolwarm')
plt.show()
```



Heat Matrix of features correlation

Models : Linear Regression, KNN, Random Forest and Gradient Boosting Regression

Performances before tuning :

Training Set Mean Absolute Error: 59.9824

Test Set Mean Absolute Error: 54.0978

Performances after:

Training Set Mean Absolute Error: 28.2546

Test Set Mean Absolute Error: 31.7768

True test set	Predicted set
count 349.000000	count 349.000000
mean 115.647564	mean 114.561204
std 59.677574	std 40.456137
min 28.000000	min 39.824811
25% 70.000000	25% 78.232716
50% 99.000000	50% 108.601961
75% 150.000000	75% 151.403092
max 300.000000	max 194.025523

We performed Sentiment analysis in order to obtain an additional point of comparison for our friends and family

After performing sentiment analysis on the 3 neighborhoods we concluded that they have virtually the same perception

Morningside Heights

Positive	Negative	Neutral	Compound
26%	2%	72%	40%

Harlem

NEW YORK NY

Positive	Negative	Neutral	Compound
27%	2%	71%	41%

Manhattan Upper West Side

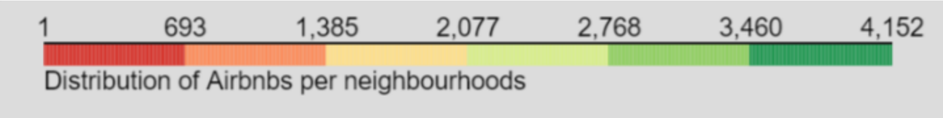
Positive	Negative	Neutral	Compound
27%	2%	71%	41%

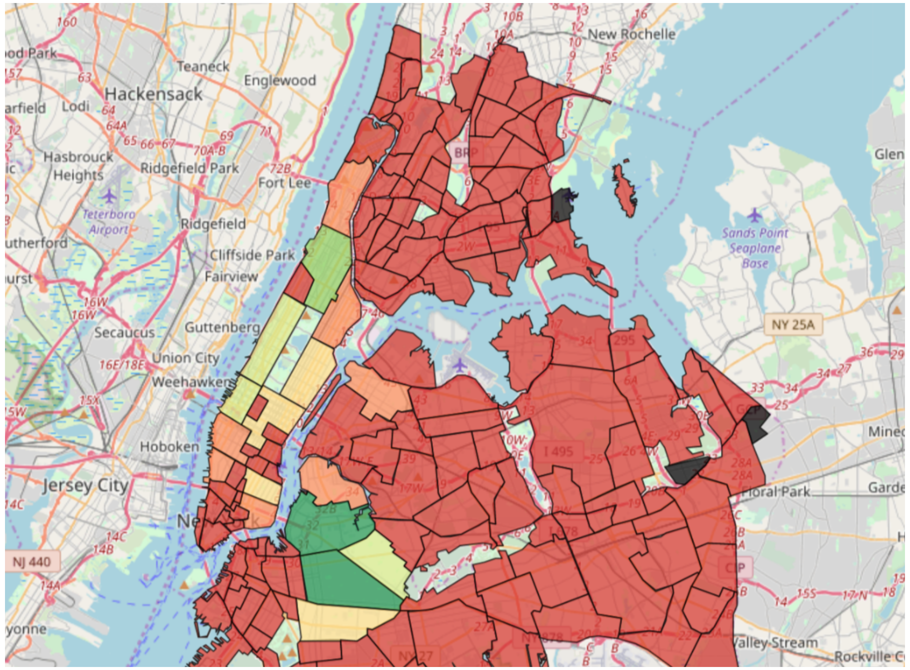
Given this similarity, the decision should depend on other factors such as room features, price, location, etc.

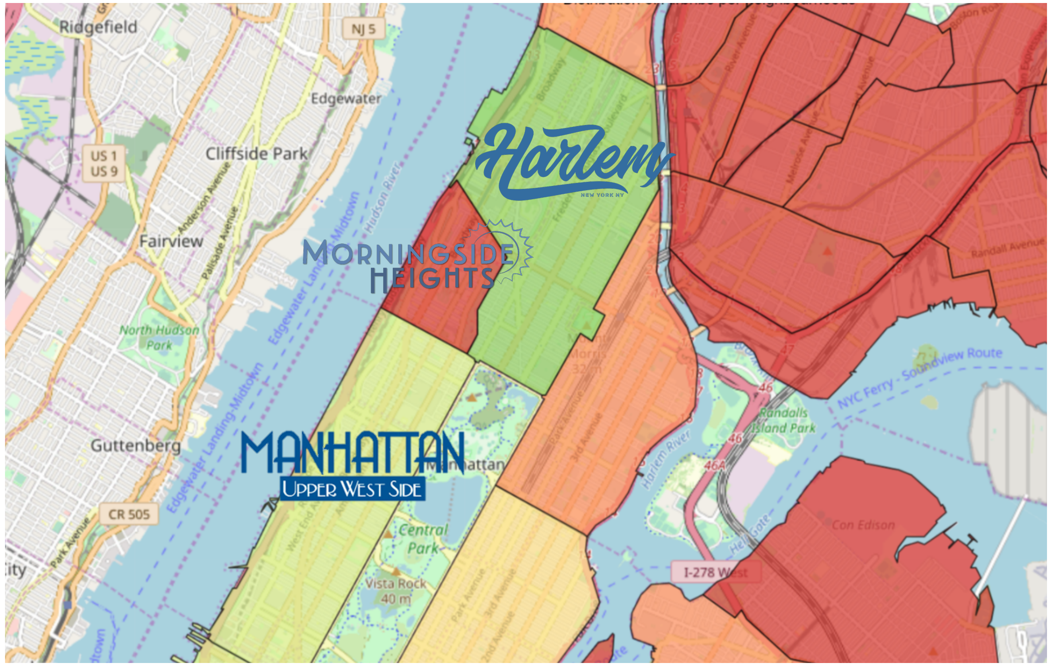
As a final step, we visualized additional information per neighborhood that could be useful in the choosing process

We plotted the number of listings per neighborhood to visualize where the concentration is highest and lowest, for example if you are looking for a less “touristy” area

Distribution of Listings





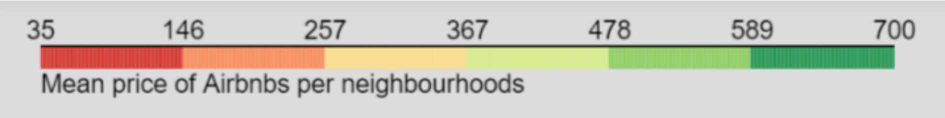


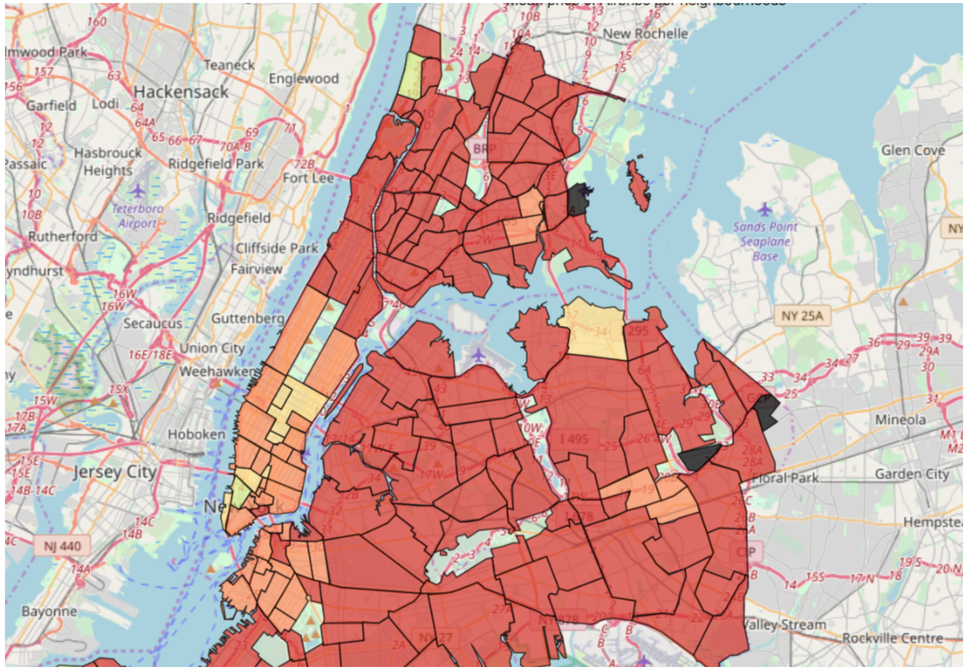
Harlem has the most listings available and Morningside shows a very limited amount of them, both due to its size and the “student” nature of the neighborhood

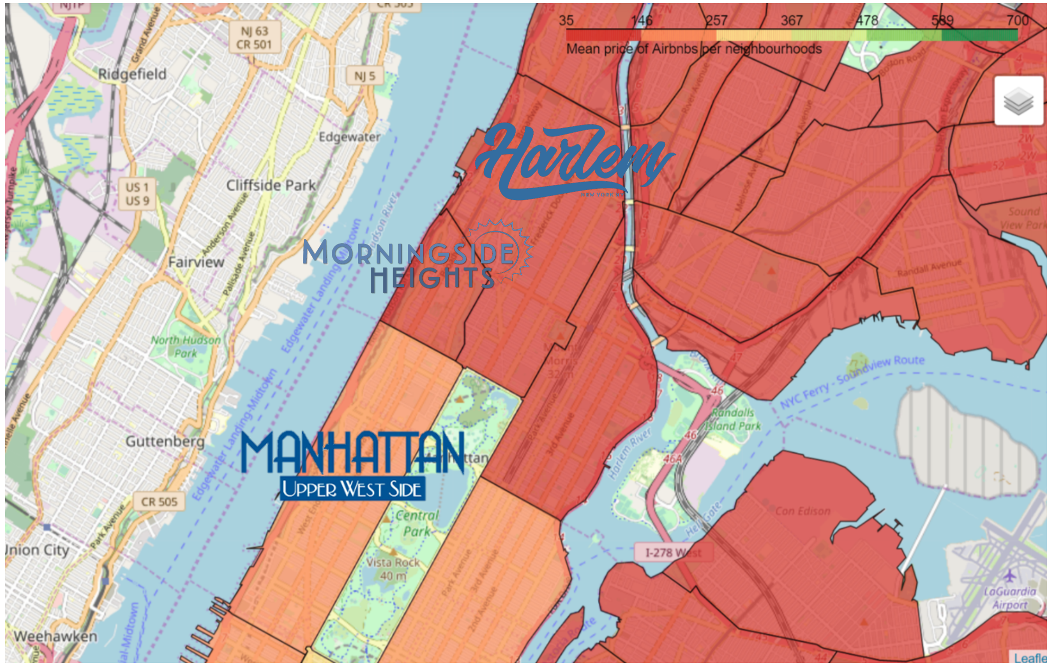
As a final step, we visualized additional information per neighborhood that could be useful in the choosing process

We plotted the average price of listings per neighborhood to visualize which tend to be more or less expensive, so that you can be within budget

Mean Price







We can observe than in average, UWS has slighter higher prices than the other 2 analyzed neighborhoods, this is expected as it is closer to midtown

In summary, we analysed and built several models to understand better the dynamics of AirBNB in NY

We focused on the 3 closest neighbourhoods to Columbia University to offer an option to friends and family attending commencement

The following are the main conclusions on the various analysis performed:

Reviews

- The amount of reviews and its consistency is one of the few available features that can help with a decision

Sentiment Analysis

- All 3 Target Neighborhoods share virtually the same perception based on the available reviews, therefore this is not a relevant decision maker in this particular situation

Price

- Price depends on the facilities and the services offered and not on reviews or location in this particular case

Neighbourhood

- Morningside Heights has the less amount of options, followed by UWS. Harlem has the most available listings

Manhattan is a very peculiar area in which **prices are not exactly correlated to ratings or any other feature we analyzed**

The High Demand experienced all-year round in almost every "central" neighborhood might be one of the main reasons for this

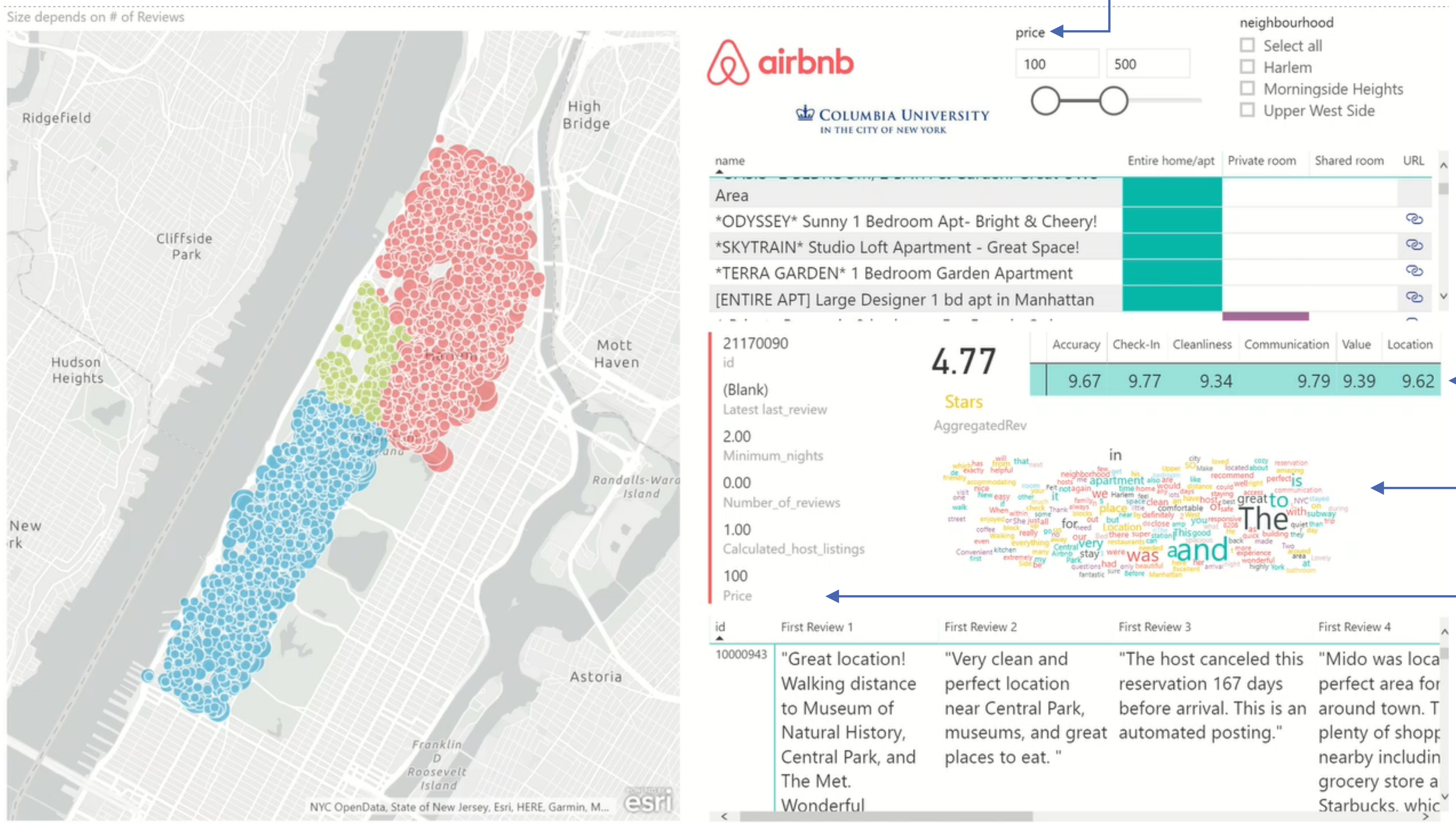
AirBNB appears to be a self-discriminating platform. We can conclude this from the listings available, the reviews are in general very good which might be an indicator of bad hosts and their listings dropping out of the platform

We decided to build an interactive Dashboard to help in the selection of the best experience

Power BI

As an alternative to our friends and family, we decided to implement an interactive tool to help make a decision

Due to the particular case of these neighborhoods and the complexity involved trying to predict the value we built a Dashboard to help ease the process of looking for the right option



Price Selection

Neighborhood Selection

Listing, Type of Room & URL

Rankings

Word Cloud Visual Aid

Listing Features

Reviews

***URL for Dashboard** (Map will not show in Web due to License Issues, but is available in original PowerBI file)*

https://app.powerbi.com/view?r=eyJrIjoiMzE0OTc4ZDEtZTE0ZC00NWU1LWIyYmYtMDk1NDAxNmQyN2QyIiwidCI6IjA2ZDdlNTFiLTU4M2MtNDA4ZC05ZDY3LTQyNzI3YzcwM2NlNCIsImMiOjR9

Thank You