Predicting rental prices and identifying ‘systemised’ listings using Airbnb data

Julia McAleenan

Part 1: Predicting Airbnb prices

Can we predict rental prices for Airbnb properties in London?
Which factors have the greatest influence on the price?

- Airbnb is an online marketplace which allows home-owners and renters to list their properties online, so that guests can pay to stay in them.
- Hosts are expected to set their own prices for their listings. Although Airbnb and other sites provide some general guidance, there are currently no free and accurate services which help hosts price their properties.
- Airbnb pricing is important to get right, particularly in big cities like London where there is a lot of competition. Understanding which are the main factors that influence the listing price could also help a host maximize their income.
- For guests, a tool which predicts typical rental prices could be used to identify good value properties.

The dataset

Inside Airbnb
Adding data to the debate

- From Inside Airbnb which regularly scrapes data from the major cities on Airbnb to facilitate public discussion on how Airbnb is really being used in these cities.
- Data scraped on 16th December 2020.
- Dataset contains approximately 77,000 listings for London including information on the property, the location, the host and ratings. In addition, over 1.1 million reviews associated with the listings.

Limitations

- The price is the price advertised by the host and not the price actually paid by a guest.
- Many of the features are dependent on the information entered by the host - not always consistent or correct.
- There is no clear way to identify ‘active’ listings.
- Data is from December 2020 - in the middle of the pandemic.

Data cleaning

Initial dataset ~ 77,000 → Active listings ~ 21,000 → Final dataset

Only consider active listings since inactive listings are likely to have unreliable data.

Assume that for a listing to be active:

- It must have been reviewed in the last 12 months
- It must have at least two nights booked in the next 90 days

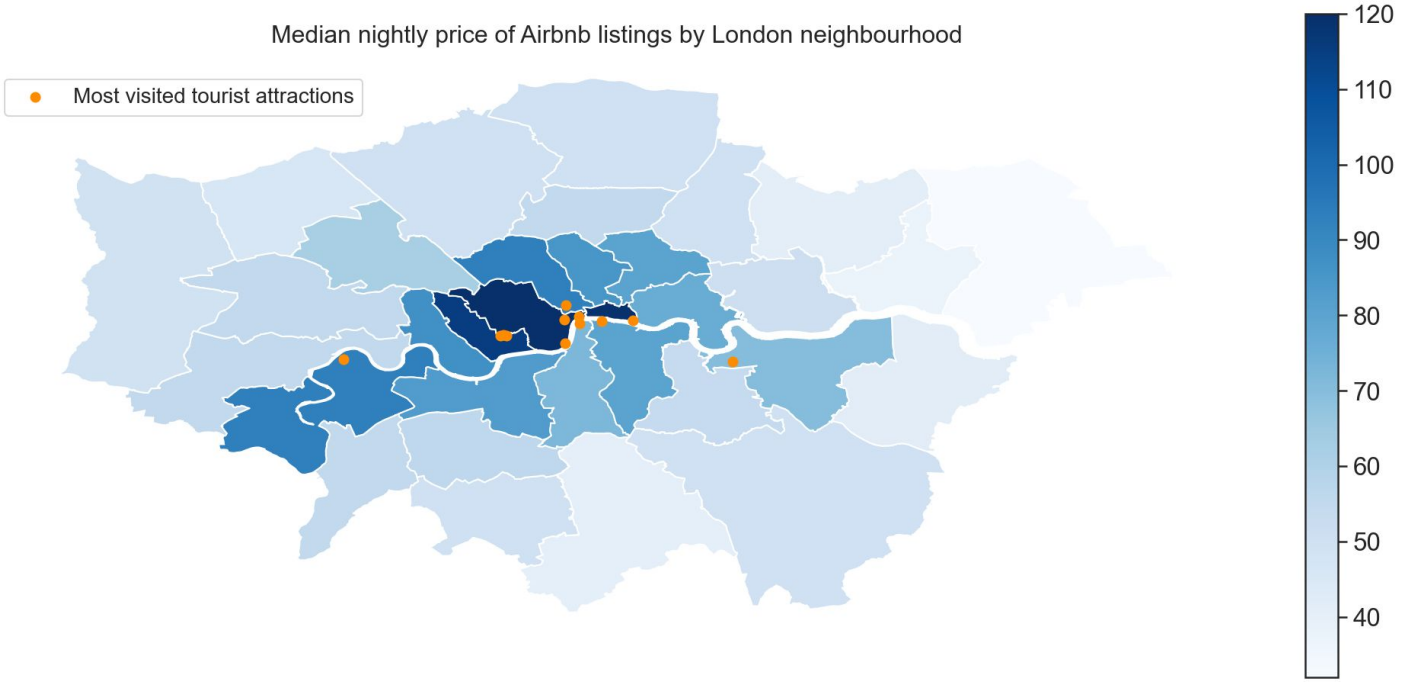
- Impute missing values
- Convert text information into numerical features for modelling (e.g. '2 private bathrooms' - create a number of bathrooms feature and a type of bathroom (private or shared) feature)
- Convert dates into numbers of days

33 features for modelling from original dataset, including:

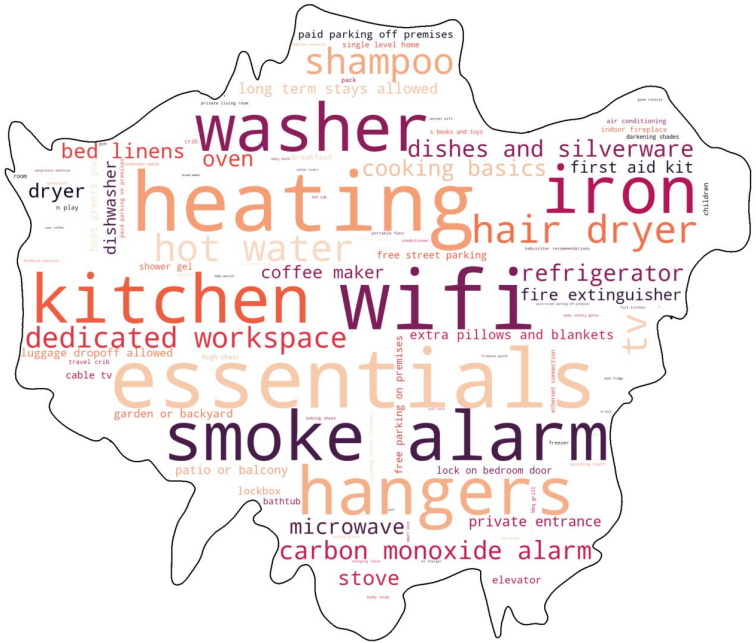
- # accommodates
- # bathrooms, # bedrooms, # beds
- property / room type
- neighbourhood
- review ratings, counts and how recent
- availability
- host number of listings, response time and rate

Feature engineering - distance to tourist attractions

The distances to each of the 12 most visited tourist attractions in London were calculated and added as new features.



NLP was used to engineer additional text features



Most common words / phrases in the list of amenities

Most common words / phrases in the listing name

Sentiment analysis using VADER

- There are ~1.1 million reviews associated with the listings.
- After dropping non-English reviews and automated postings, VADER was used to analyse sentiment of the remaining ~1 million reviews.
- VADER is a rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.

VADER examples:

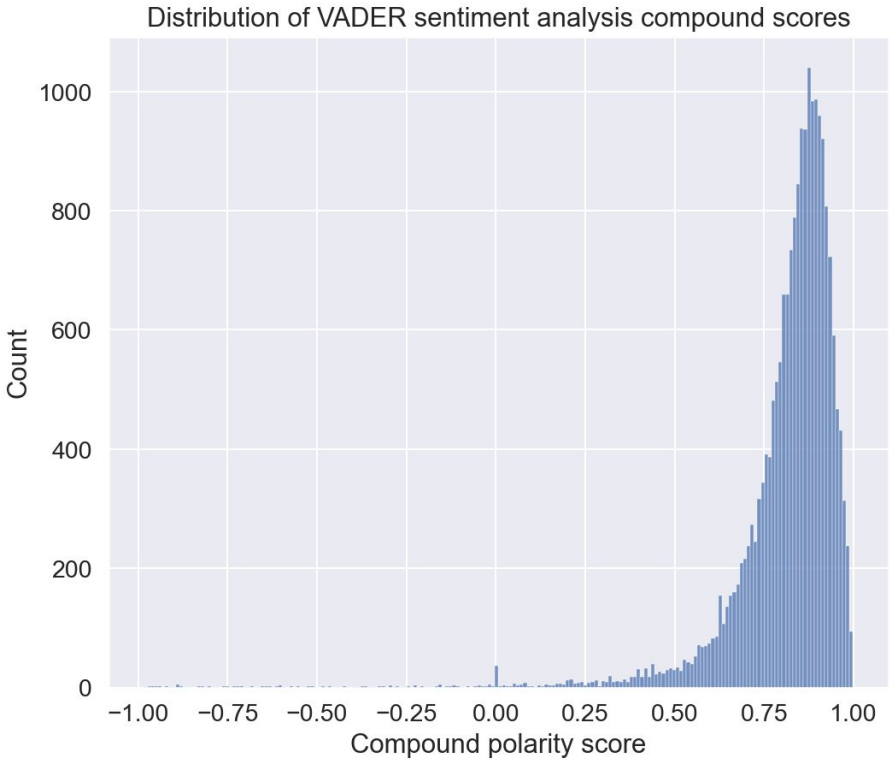
"We stayed with Adriano and Valerio for a week when first moving to London. The apartment is great and very clean compared to a lot of places we've seen in London. Situated very close to Brixton tube and good bus links to central London. Thanks guys"

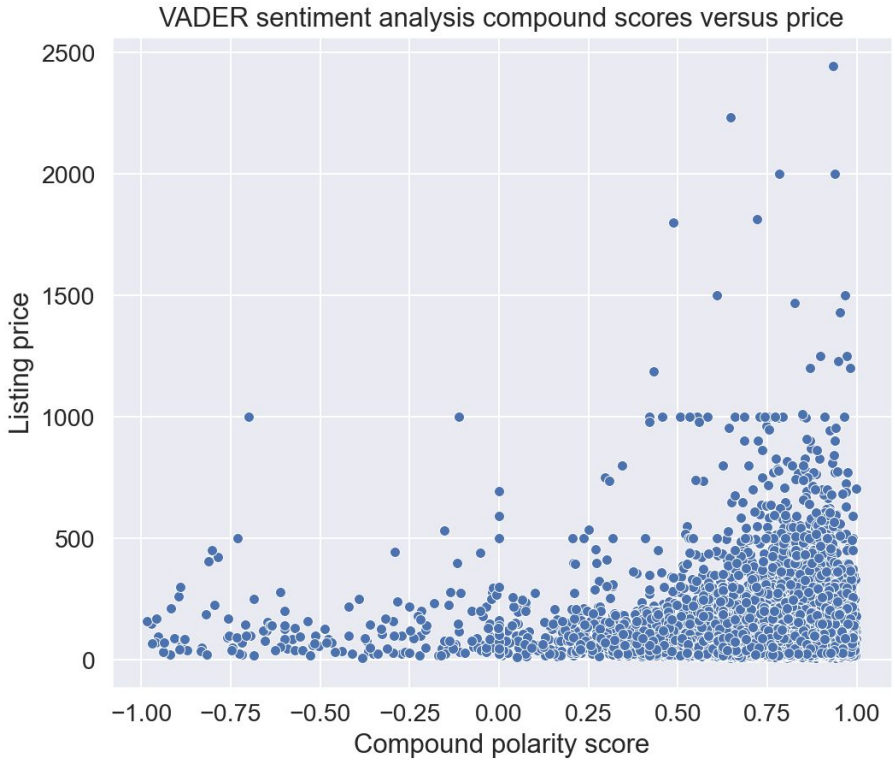
compound: 0.9214, neg: 0.0, neu: 0.752, pos: 0.248

"Overall just pretty gross place. The room itself was clean, but very small. The staircase was just filthy and smelled. The hall and bathroom smelled pretty bad, and so did the kitchen. There was no lock on the door which is a basic thing, also no WiFi whatsoever (no signal). Just overall not very pleasant."

compound: -0.8413, neg: 0.231, neu: 0.652, pos: 0.117

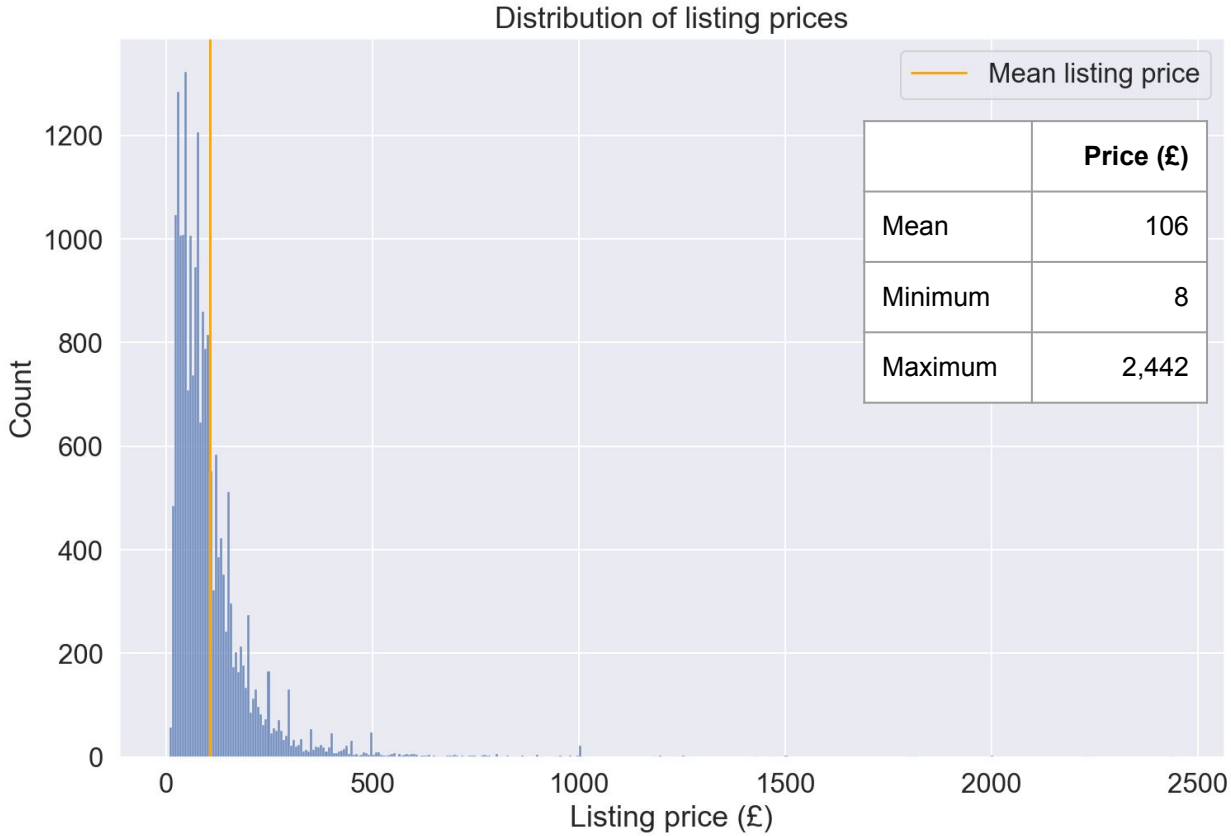
Sentiment in reviews tends to be very positive





Price distribution

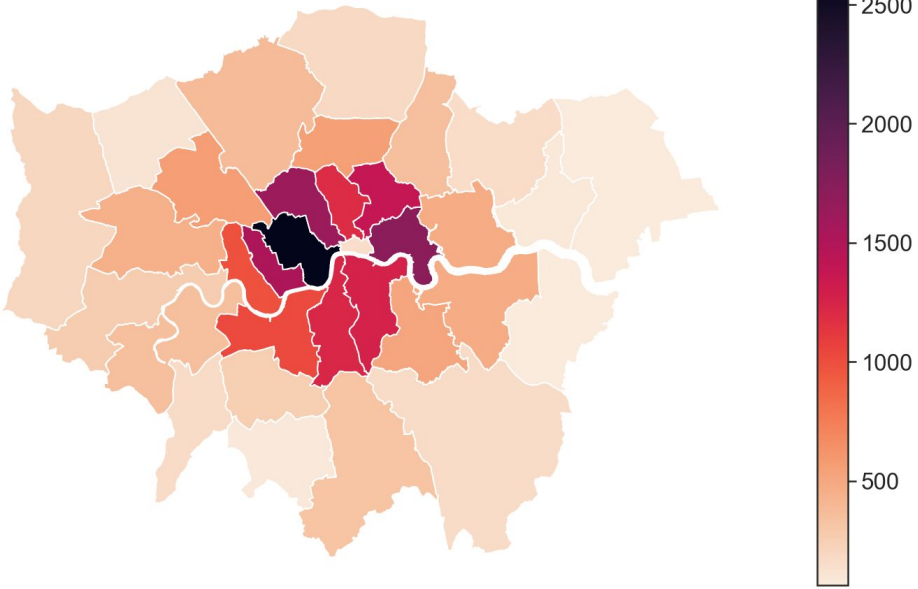
- The chart shows the distribution of daily rental prices in London.
- The distribution shows positive skew - it has a long right tail.
- For this reason, a power transformation to make the price more normal distribution-like will be considered for the modelling.



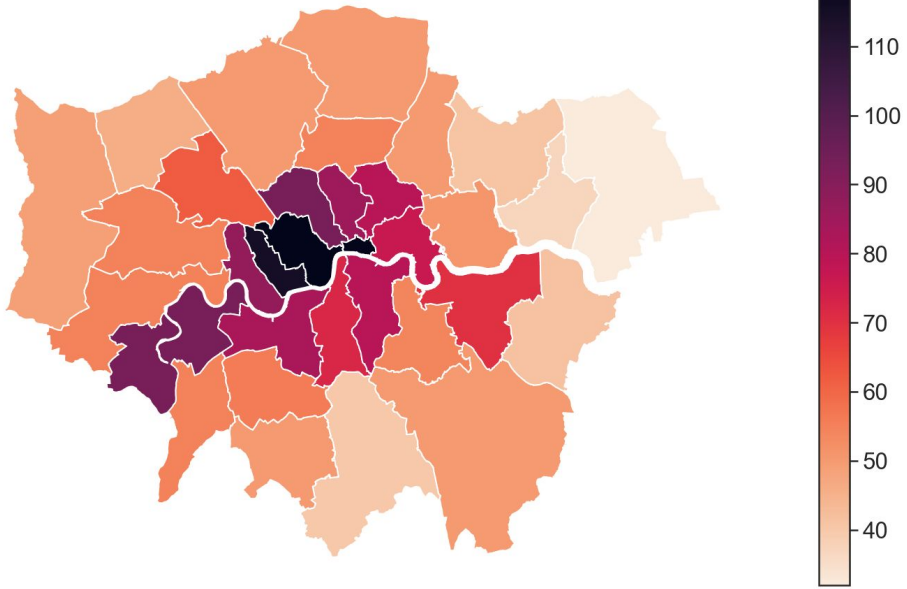
	Price (£)
Mean	106
Minimum	8
Maximum	2,442

The most expensive listings are in the centre

Number of Airbnb listings by London neighbourhood



Median nightly price of Airbnb listings by London neighbourhood



Room type

The categorical feature ‘room type’ should be a good predictor of price.

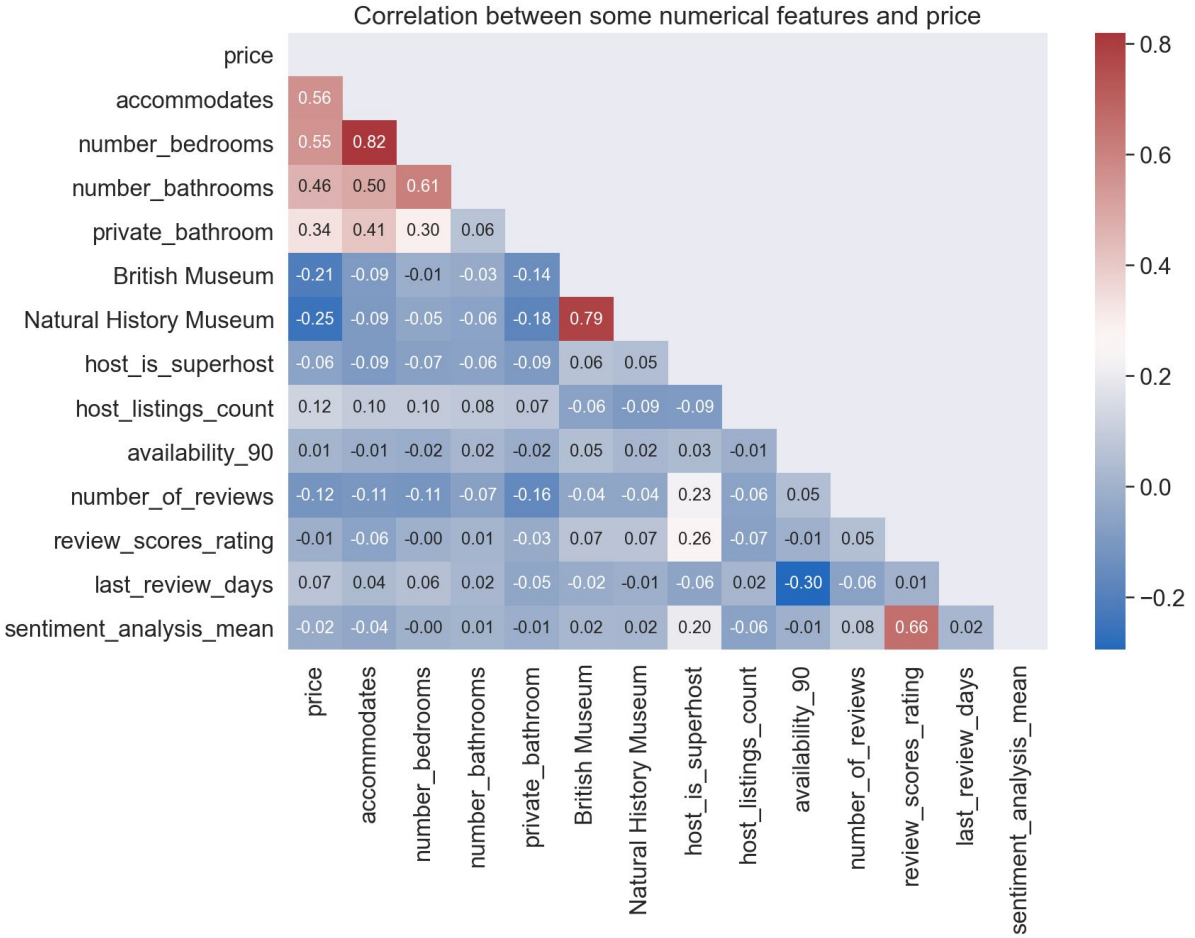
Further information on the room type is extracted from the ‘property type’ feature using NLP - this adds potentially important words such as:

- Hostel
- Guesthouse
- Loft
- Suite
- Cottage

Numerical features

The heatmap shows the correlation between a subset of the numerical variables and the price. Key points:

- The size of the property (number of bedrooms, number of bathrooms, how many people it accommodates) should be a reasonable predictor of price.
- Review ratings and sentiment have low correlation with the listing price.
- Location should play a role.



Modelling

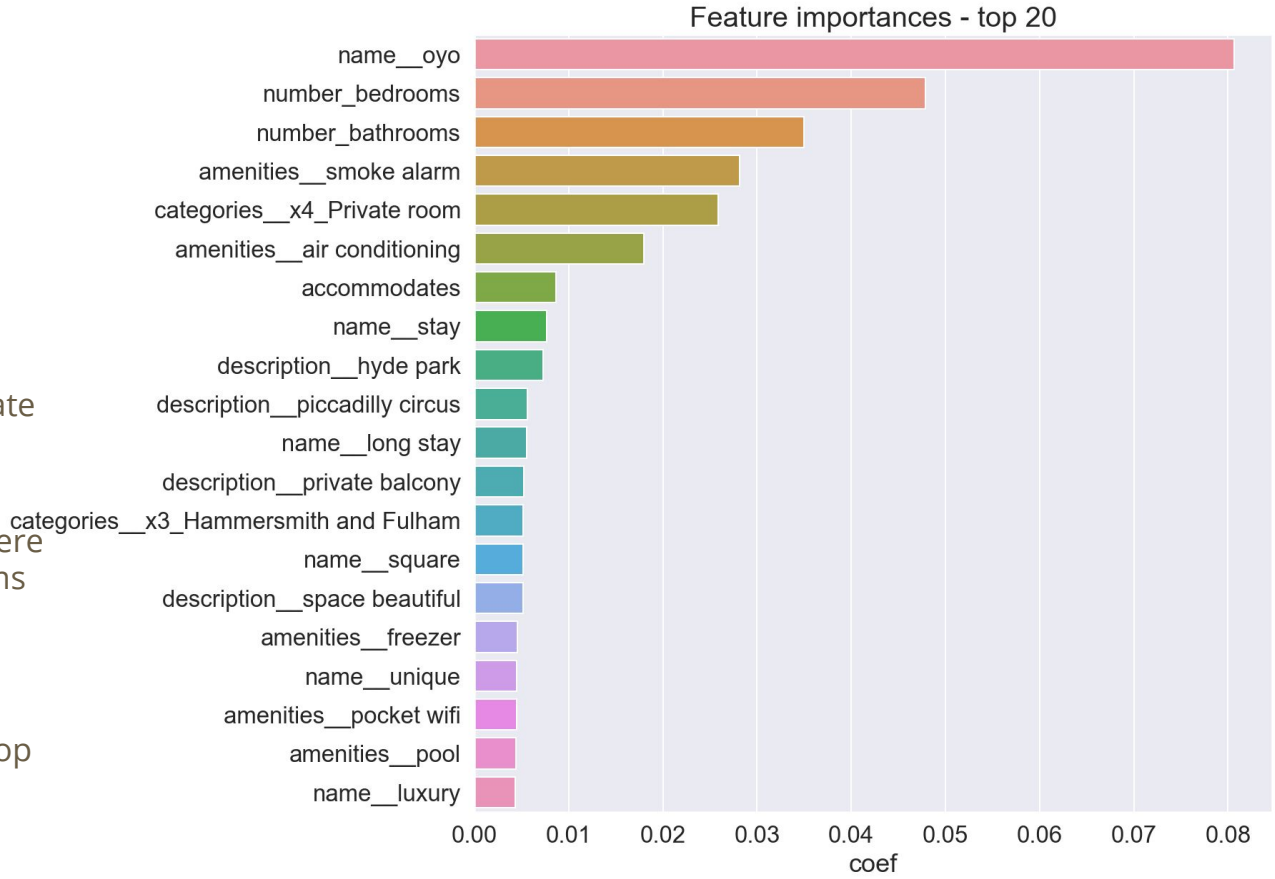
- Many different regression models were run and fine-tuned. The table shows a summary of the best performing models.
- XGBoost achieves the highest mean cross-validated R^2 score and the lowest mean absolute error.
- The neural network tried was a multilayer perceptron with 3 hidden layers. Further tuning could potentially improve this further.
- For comparison, the Random Forest model run without the additional engineered features gave an R^2 score ~0.07 lower.

Model	Mean CV score	MAE (on test set)
Ridge	0.556	37.8
Lasso (with power transform)	0.581	29.9
Random Forest	0.639	30.5
XGBoost	0.659	29.5
Gradient boosting	0.646	30.5
MLP neural network	0.637[1]	32.9

1. The score for the neural network is the R^2 score on the test set

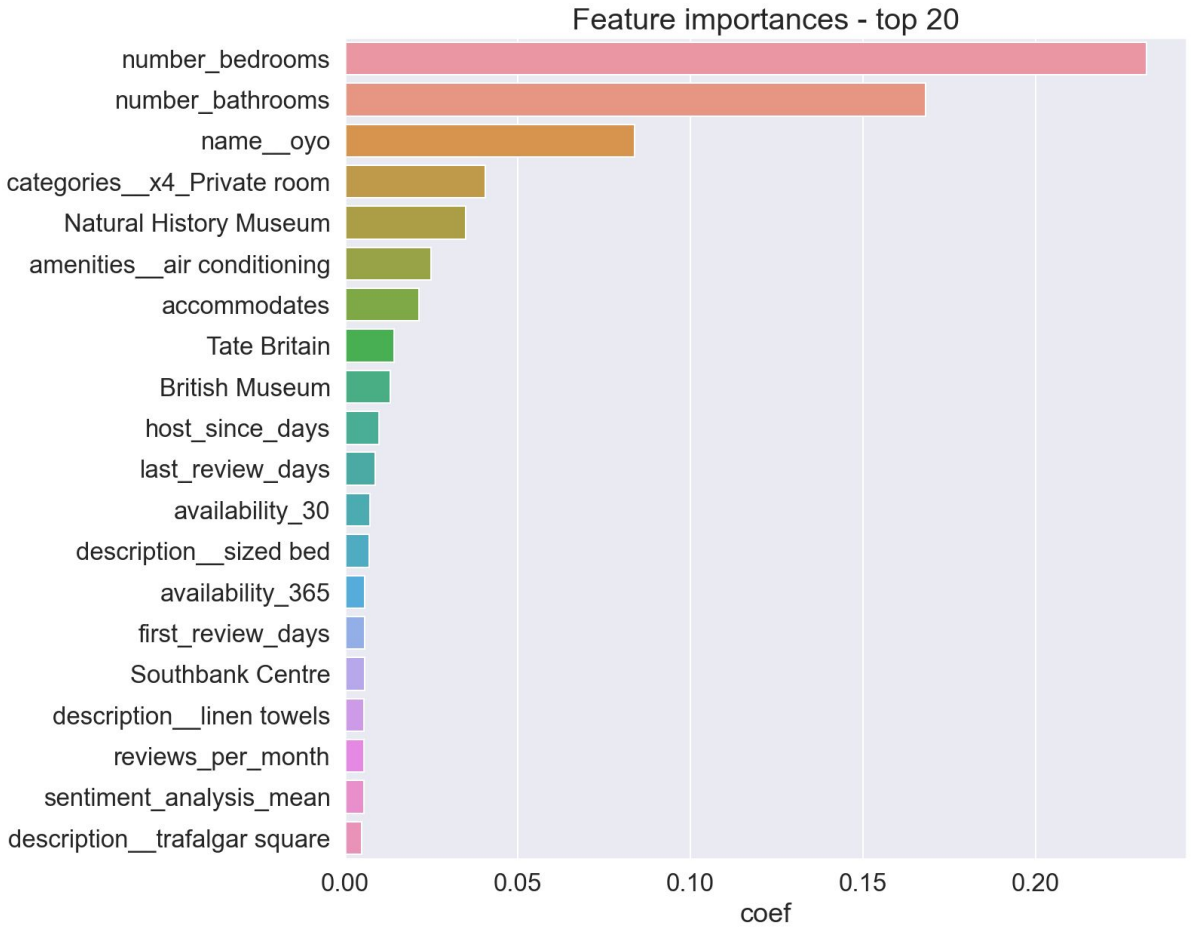
Feature importances - XGBoost model

- The size of the property is important in predicting the price (number bedrooms, number bathrooms, accommodates).
- Having a room type of 'Private room' is an important predictor.
- OYO is a chain of hotels. There are 62 listings for OYO rooms in the dataset with a mean price of £817.
- The features resulting from NLP make up many of the top feature importances.



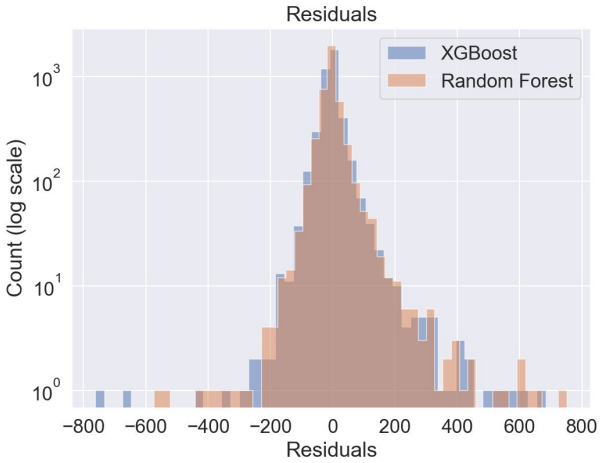
Feature importances - Random Forest model

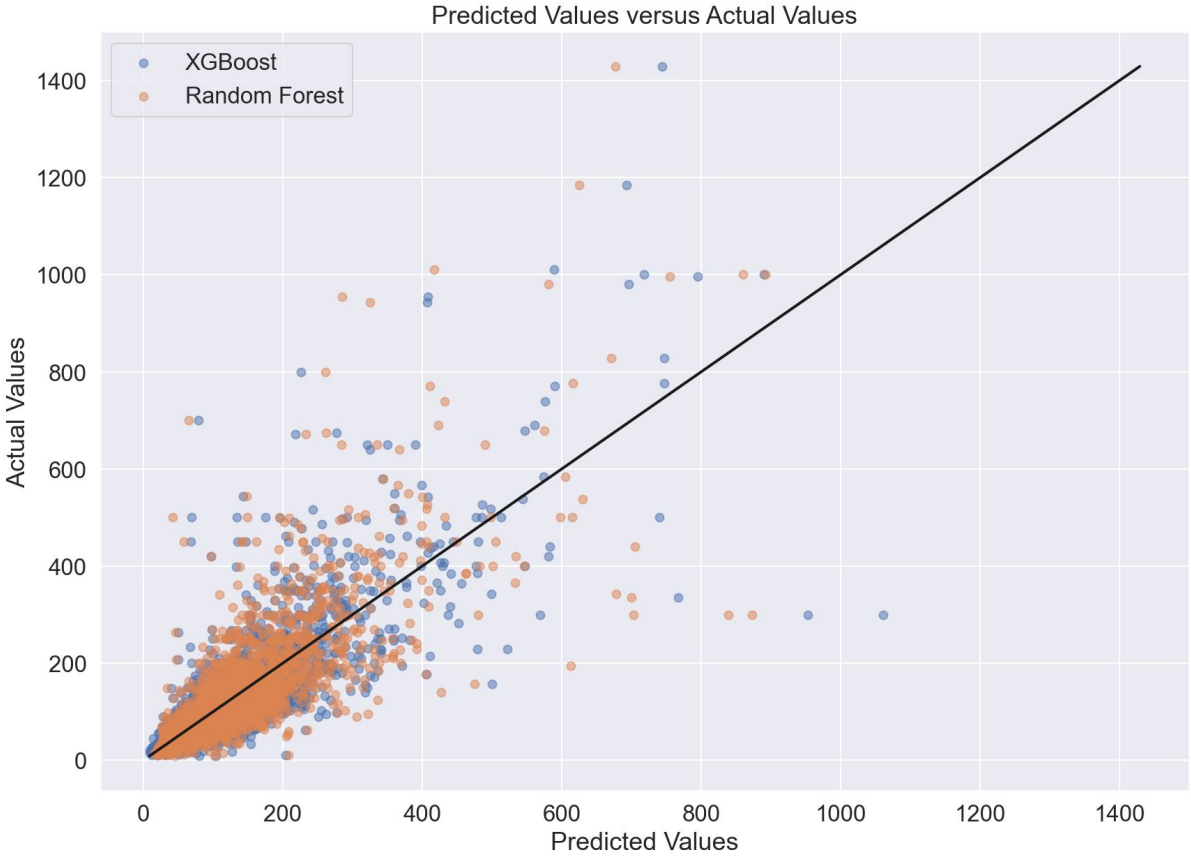
- Again the size of the property is important (number bedrooms, number bathrooms, accommodates).
- Having ‘OYO’ in the name is the third most important feature.
- Location features play a bigger role here (e.g. Natural History Museum).
- Other features such as sentiment analysis and availability are important.



Predicted values versus actual values

The chart shows predicted values versus actual values for both the XGBoost and Random Forest models. Both models struggle to predict the higher prices.





Conclusions

- The XGBoost model gave the highest R^2 score of 0.659 and mean absolute error of 29.5. The model struggles to predict the high prices where there is very little data.
- It would be interesting to run the analysis on potentially larger datasets at different points in time, for example December 2019, before the pandemic.
- Feature engineering - the extra features that I engineered added ~0.07 to the final R^2 score.
- Other features which could help to explain the variation in price:
    - Listing photos are potentially very important. Metrics on e.g. number of photos, quality of photos could help explain some of the variance.

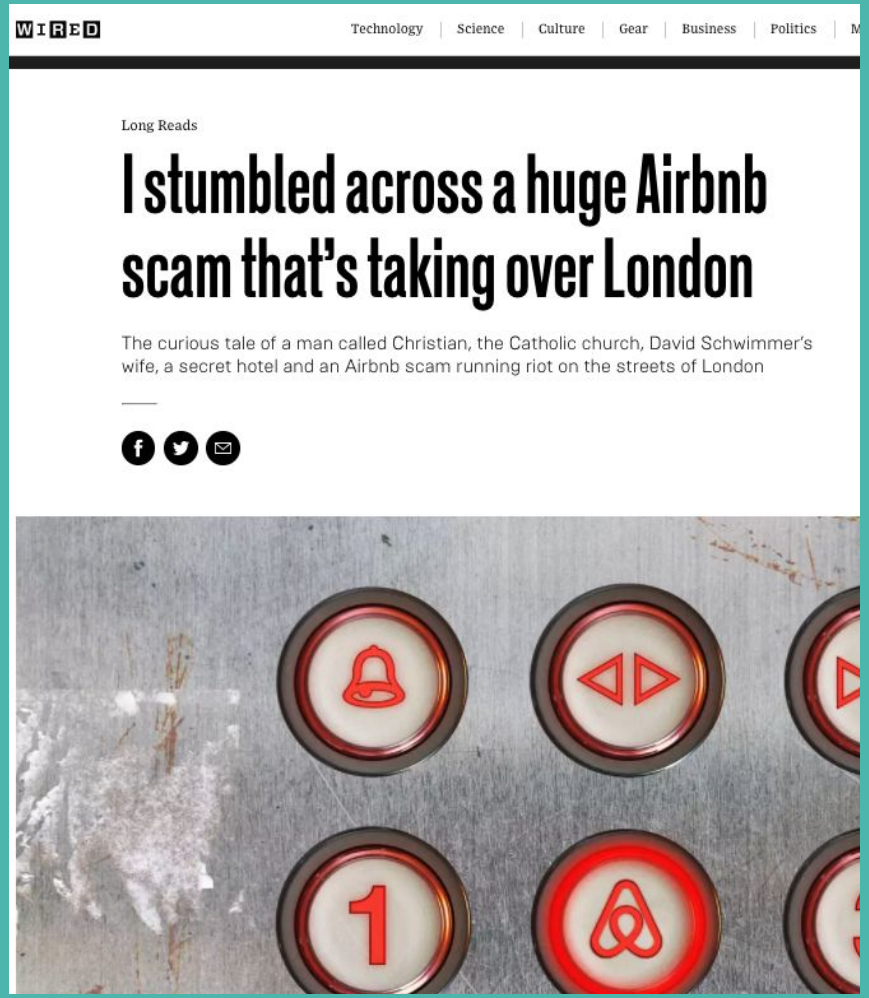
Part 2: Identifying ‘systemised’ listings

‘Systemised’ listings on Airbnb

This article was published in February 2020 and describes in great detail an Airbnb scam which the journalist uncovered in London.

The scam centres around a de facto hotel built for short-term rentals which breaches both Airbnb policies and local planning laws (London’s 90-day law).

There are duplicate listings for apartments, most listings use the same photos or mirror image photos and some of the listings don’t exist. Guests often complain of being put into a different apartment to the one they booked and sometimes apartments are double-booked. Many of the reviews and host profiles are fake or misleading.

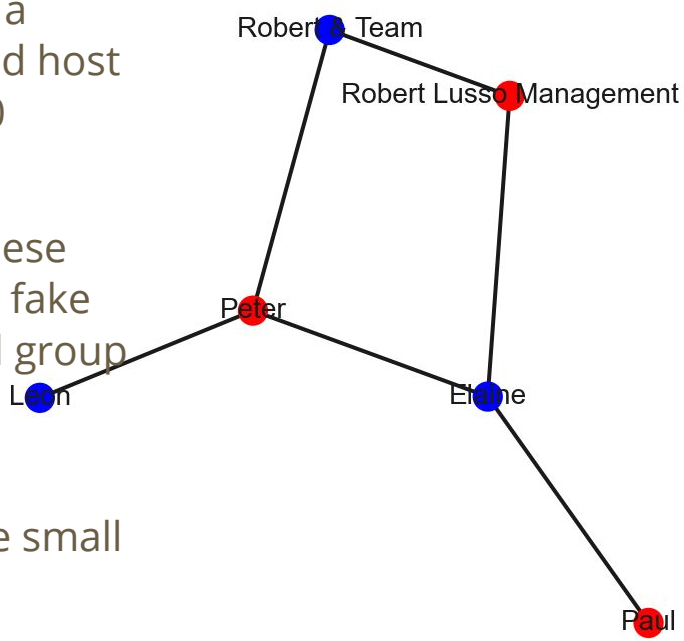


A network of connected hosts and reviewers

The article identifies a network of connected host accounts behind 200 ‘systemised’ listings.

Listings hosted by these accounts have many fake reviews from a small group of reviewers.

Can we use network analysis to find these small networks?



…. All these host profiles have a few things in common: they all use stock photography as their profile pictures, and they all use similar text in their bios. Before long, a network of connected host accounts emerges. As well as Robert & Team, Leon and Elaine & Team, there’s also Eveline, Natalia, Felly, Robert Lusso Management and Alex. Airbnb listings hosted by these accounts are littered with fake reviews. As well as Peter, Elaine and Alex Cosmin ….

…. Between them, they have received over 2,100 reviews on 200 listings, most of them in London….

…. All of these accounts are essentially one person, or at least one company. And yet they have all passed Airbnb’s account verification and safety processes ….

The data

Initial dataset ~ 1.7 million edges

Subset of edges ~ 10,000

Data taken from October 2019 (from the time of the research for the article) and October 2018 (to capture accounts which have been deleted).

Reviews data includes reviewer_id and listing_id. The listing_id can then be mapped to the host_id using listings data.

Problem: too many edges to be able to identify small networks easily.

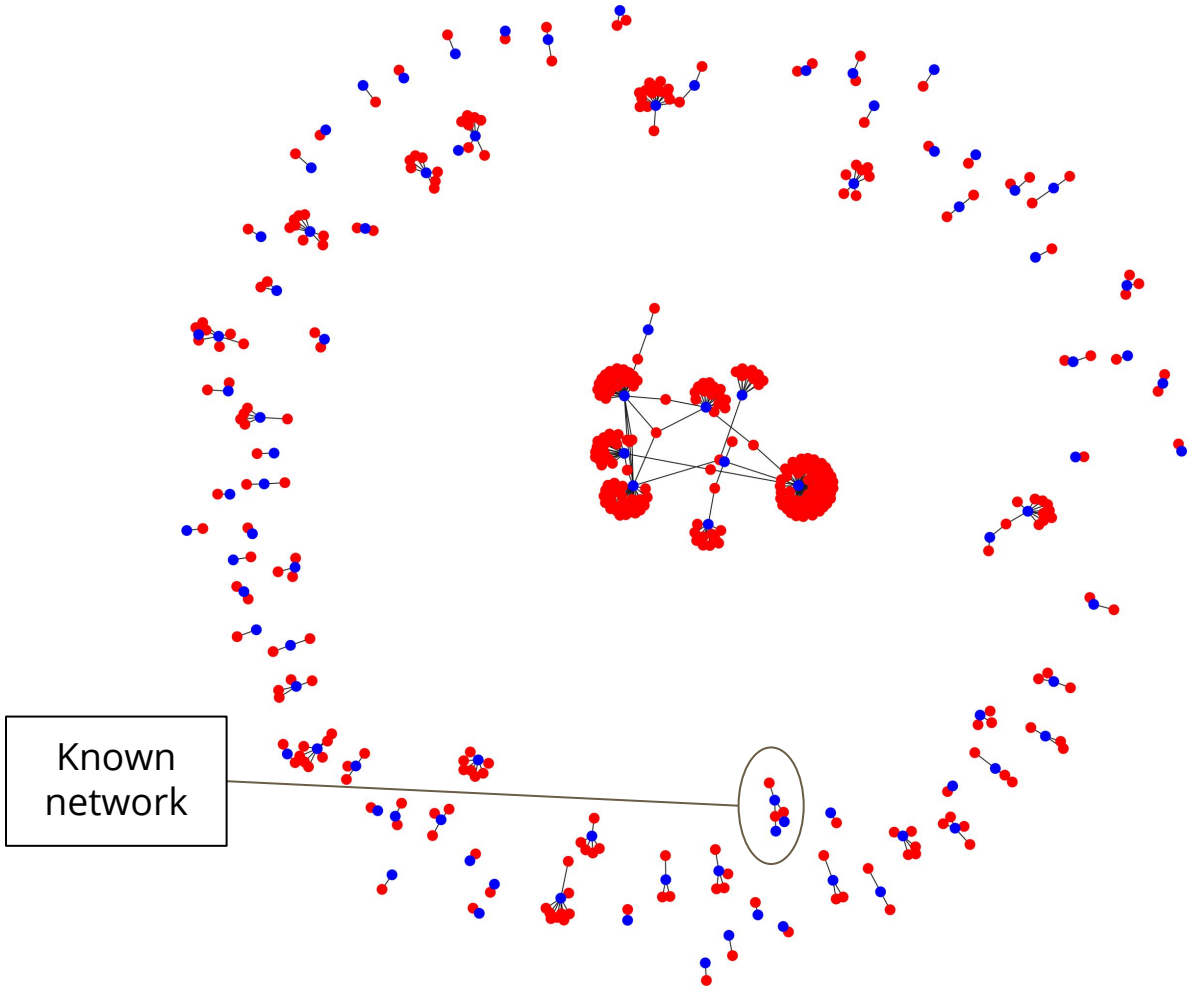
Subset on characteristics which are more likely to indicate fake reviews and 'systemised' listings:

- Only include hosts and reviewers where reviewers have left multiple reviews for a host (can be across different listings)
- Only include hosts with multiple listings
- Only include reviewers who have reviewed multiple hosts
- Exclude multiple reviews from a reviewer for a listing which occur on the same date

First test concept on a subset of 1,000 edges

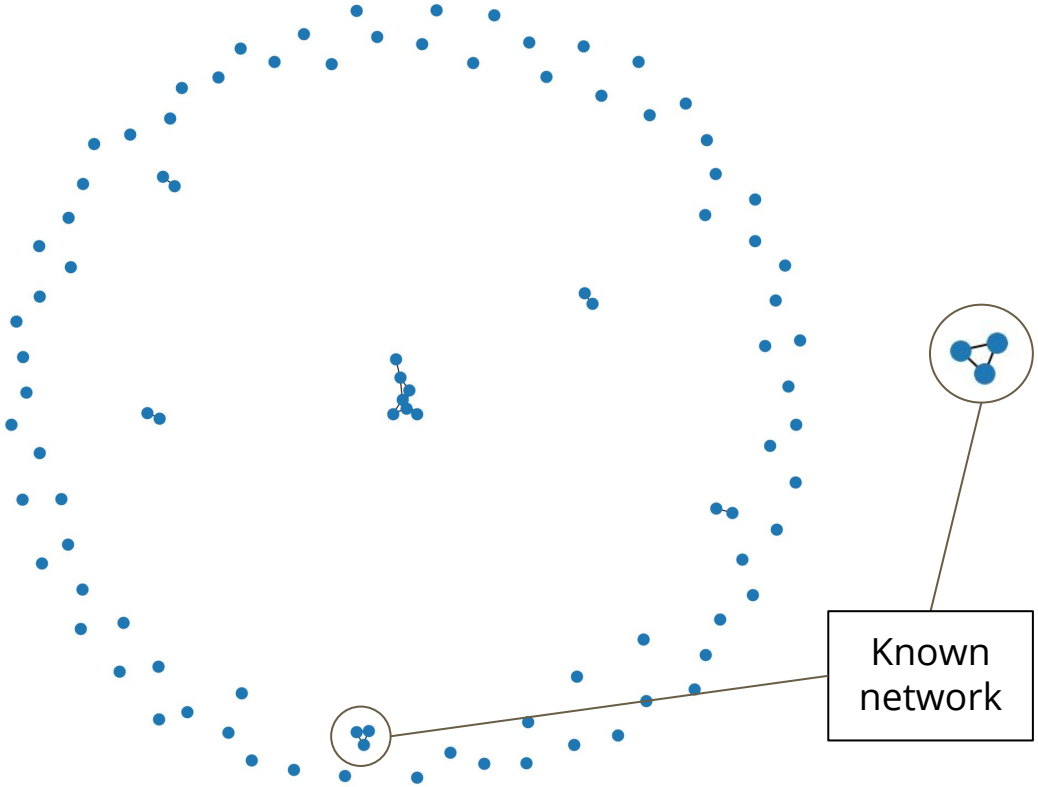
Step 1

- Bipartite network created as no direct links exist amongst hosts or amongst reviewers.
- **Red**: reviewers
- **Blue**: hosts



Step 2: Create projected graph on hosts

Step 3: Remove hosts with no links

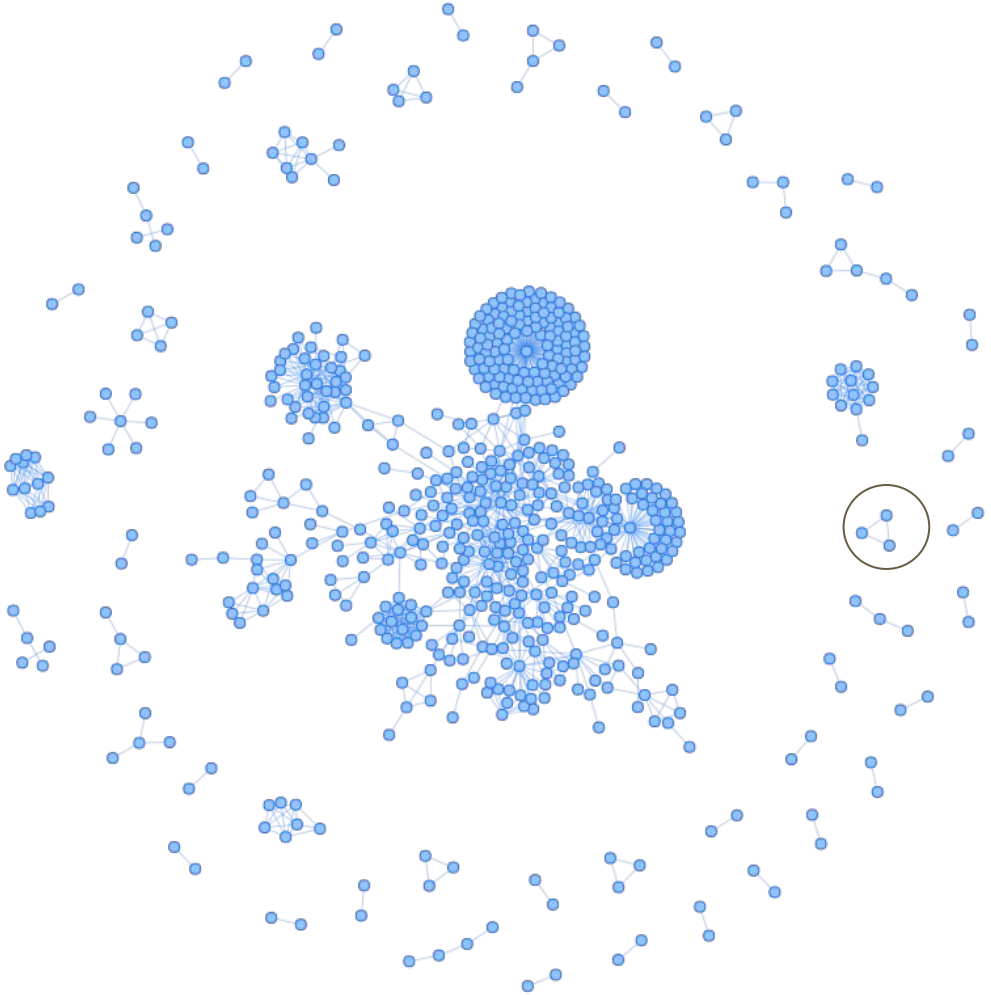


On full dataset (October 2019)

It is difficult to investigate the other networks as many accounts have been deleted.

Following a similar discovery of scam listings in the United States, Airbnb said in an email titled “In The Business of Trust” that it would review every single listing and host on its platform by December 2020.

Let’s look at the data.

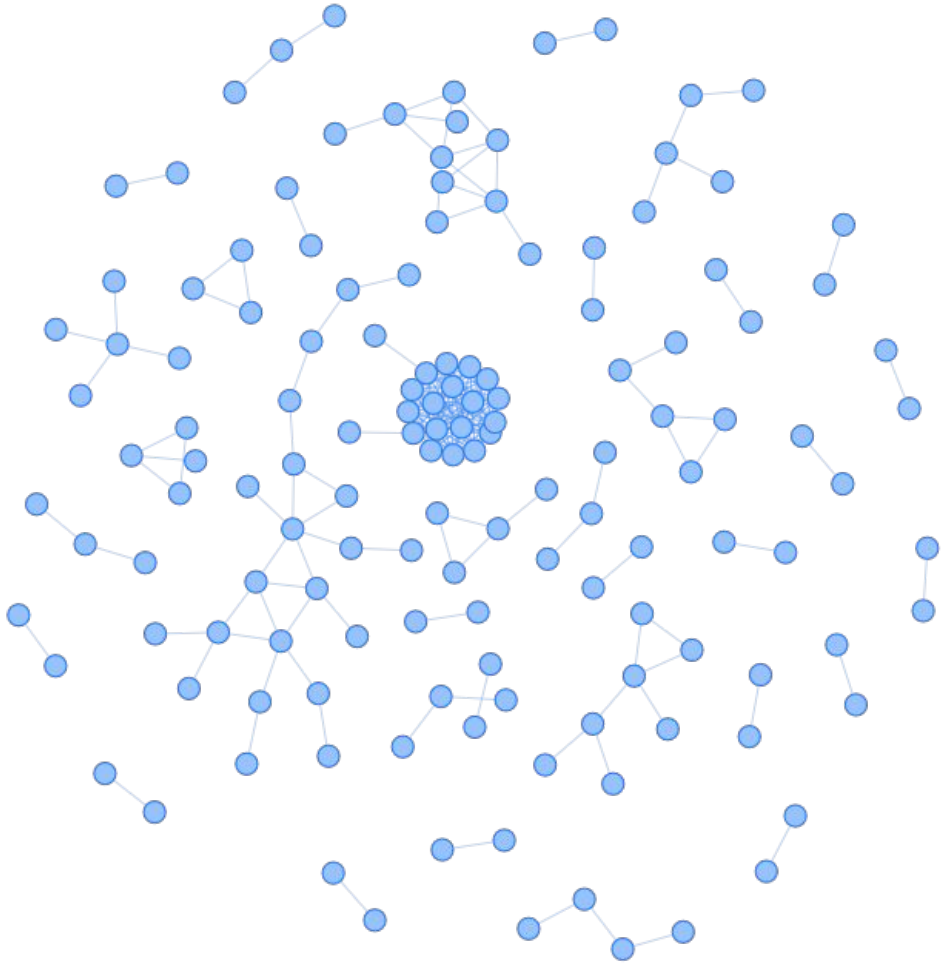


December 2020 data

Starting point: ~1.1 million edges

After subsetting on the characteristics which are more likely to indicate fake reviews and ‘systemised’ listings: ~3,600 edges

Let’s look at some of the components.

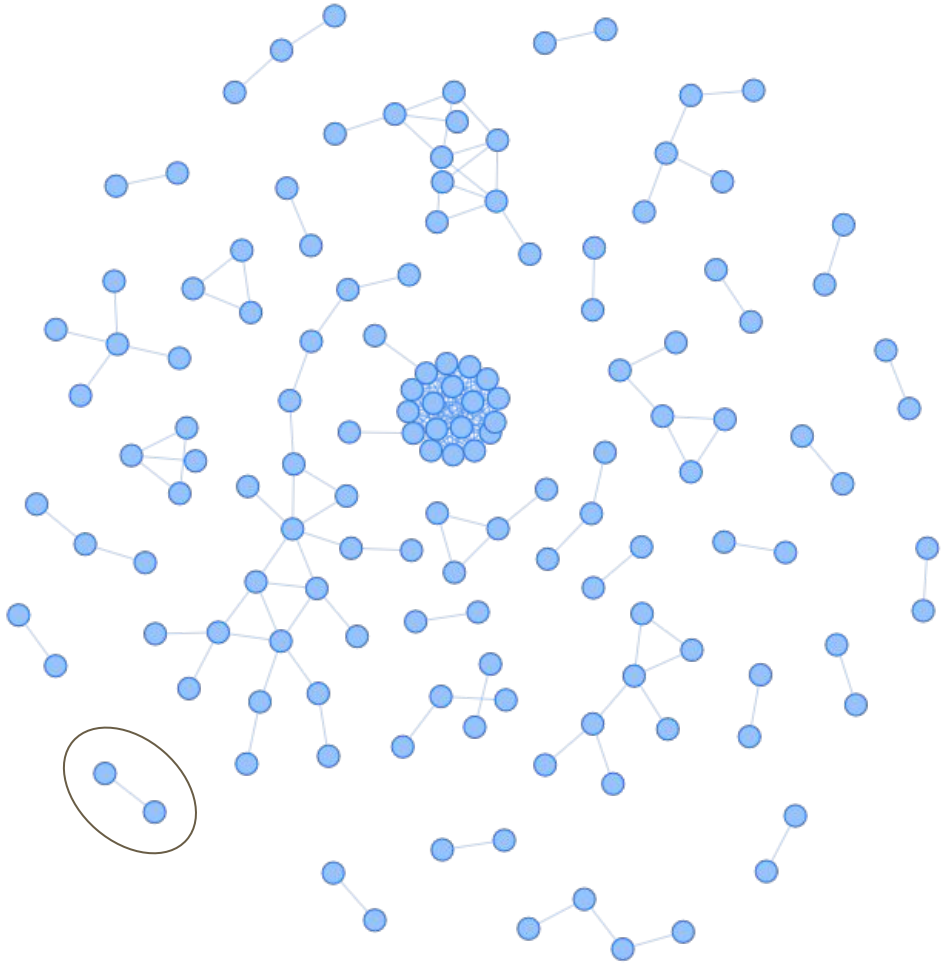


Simple example

These two host accounts have the same name.

The host profile photos are different but appear to be the same person.

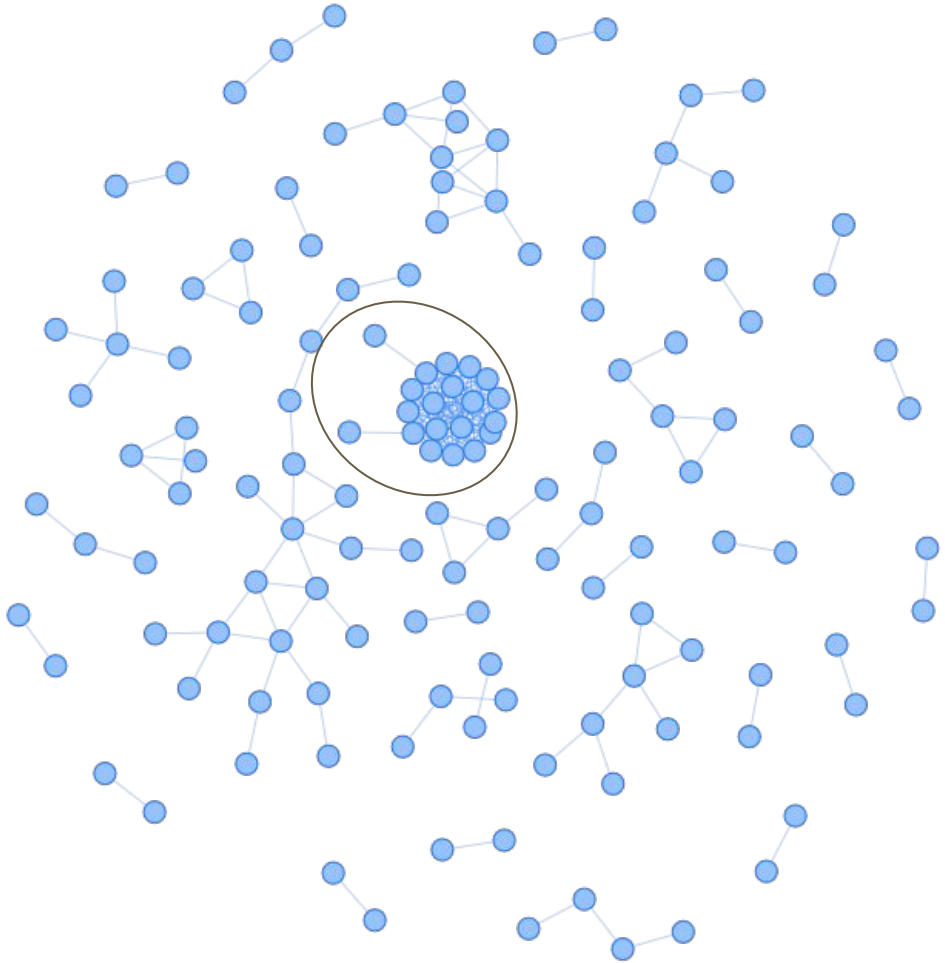
The two host accounts have listings which look like duplicates (very similar names, identical photos).



Complex example

Most of these host accounts have variations on the same name (e.g. X and Y, Y and X, X & Y, X/Y). Further investigation shows:

- Duplicate listings, some photos taken from different angles to make them look different
- Overall positive reviews but some very negative
- Multiple comments about not staying in the apartment that was booked
- Multiple reviews mention last minute cancellations due to leaks and quite a few complaints about poor customer service

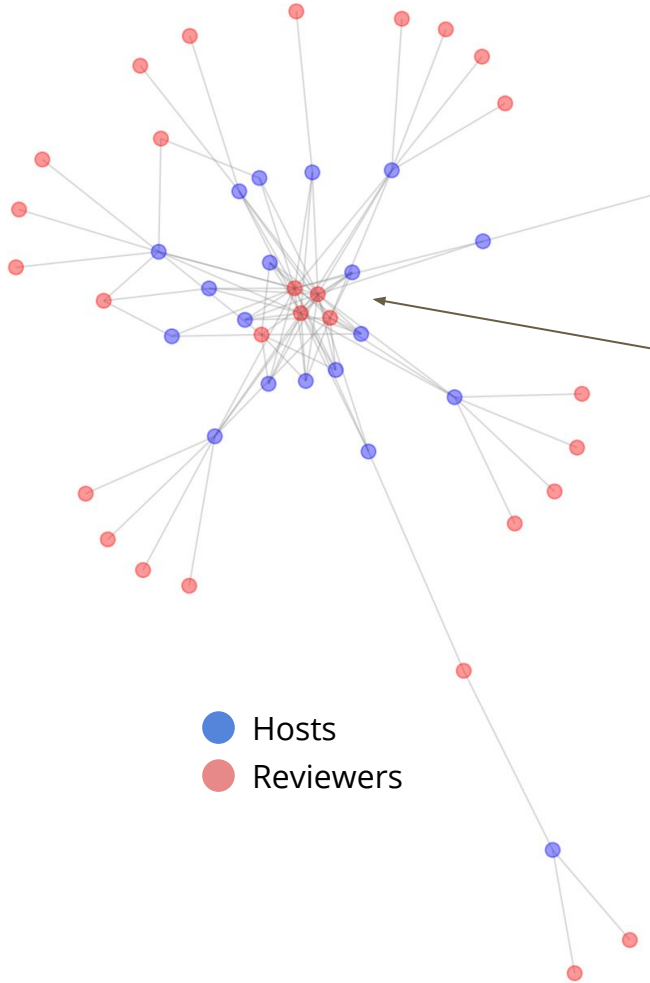


Complex example (cont'd)

The network graph shows this network with the reviewers who connect the hosts added back in.

There are 5 reviewers at the centre who have written many reviews for listings belonging to these hosts.

Accounts for Reviewer A and Reviewer B seem to have been deleted. The other 3 reviewers have many reviews from the 'connected' hosts with identical text.



Reviewer	# of reviews
Reviewer A	77
Reviewer B	73
Reviewer C	70
Reviewer D	43
Reviewer E	30

Questions