Capstone Presentation

Edmond O’Driscoll

Hypothesis No.1

Q. Can I predict the likelihood an AirBnB listing will get a last-minute booking?

Or to put it another way ...

“Will you be a host or relaxing by the coast this Saturday?”

Where I got my AirBnB data

Inside Airbnb
Adding data to the debate

AIRDNA

airbnb

What data was available?

- Review Rating
- Amount of Reviews
- Location
- Amount of Images
- Host Listing Total,
- # Bedrooms and 38 more

A. Maybe … but not in time for this presentation

Current databases do not include enough daily and forecast observations, alongside simply not enough time to gather the information needed

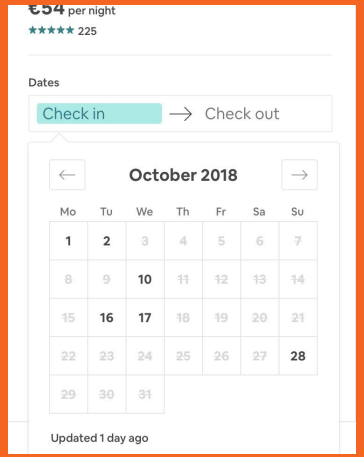
From failure … to insight!

The data activists who built Inside AirBnB have been using broad estimates to predict AirBnB occupancy levels.

They've called this the San Francisco Model.

Q. Why do they do this?

Problem: AirBnB does not share occupancy data.

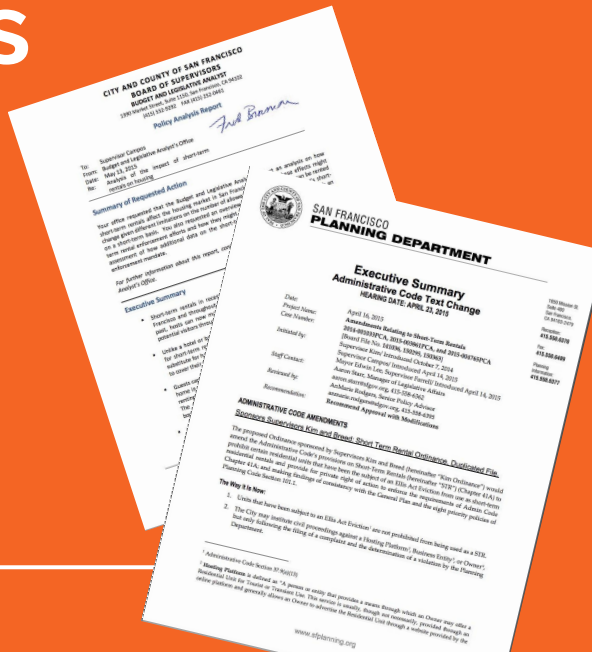


Calendar data is either:

Available
or
unavailable
(booked/not wanting to be booked)

How review-based predictions started with these two papers

- The Planning Department Investigation and;
- The Board of Supervisors Investigation



San Fran model fleshed out

- A review rate of 30.5%- 72% per stay
- Average length of stay 5.5 (in San Fran)
- Total overall occupancy up to 70%

Maybe delete

Inside AirBnB:
Murray Cox

Inside Airbnb's "San Francisco Model" uses a modified methodology as follows:

- A **Review Rate of 50%** is used to convert **reviews** to **estimated bookings.**
  - Alex Marqusee uses a review rate of 72%, however this is attributed to an unreliable source: Airbnb's CEO and co-founder Brian Chesky.
  - The Budget and Legislative Analyst's Office (page 49) also use a value 72% for their review rate, and in addition, introduce a higher impact model using a review rate of 30.5% - based on comparing public data of reviews to the The New York Attorney General's report on Airbnb released in October 2014.
  - Inside Airbnb analysis found that using a review rate 30.5% is more fact based, however probably not conservative enough, given that the Budget and Legislative Analyst's Office did not take into account missing reviews because of deleted listings. A review rate of 72% is unverifiable - therefore 50% was chosen as it sits almost exactly between 72% and 30.5%.

- An **average length of stay** is configured for each city, and this, multiplied by the **estimated bookings** for each listing over a period gives the **occupancy rate**
  - Where statements have been made about the average length of stay of Airbnb guests for a city, this was used.
  - For example, Airbnb reported 5.5 nights as the average length of stay for guests using Airbnb in San Francisco.
  - Where no public statements were made about average stays, a value of **3 nights per booking** was used.
  - If a listing has a **higher minimum nights** value than the average length of stay, the minimum nights value was used instead.

- The **occupancy rate** was **capped at 70%** - a relatively high, but reasonable number for a highly occupied "hotel".
  - This controls for situations where an Airbnb host might change their minimum nights during the high season, without the review data having a chance to catch up; or for a listing with a very high review rate.
  - It also ensures that the occupancy model remains conservative.
- **Number of nights** booked or available per year for the **high availability** and **frequently rented** metrics and filters were generally aligned with a city's short term rental laws designed to **protect residential housing**.

How does this break down for UK and Ireland ?

1 Review = 9.25 day Occupancy*

*Up to 70% occupancy per month

Hypothesis No.2

Q. Can I predict Occupancy rates better?

Find and subset the businesses...

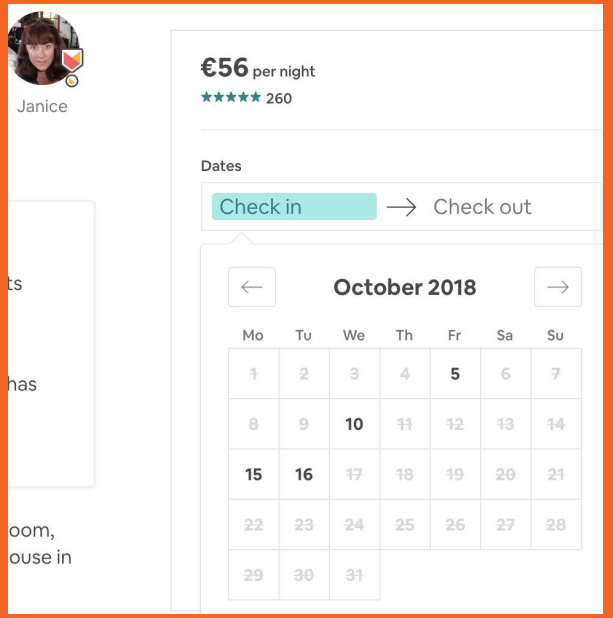
- Filter listings by those that are available to rent continuously for more than 365 days
- Filter the individual listings that the host has 20 or more listings overall
- Filter the listings by “Instant Book” … the secret sauce

Instant Book feature

- Customers can book instantly
- Host's calendar automatically updates
- Host hugely penalized if booking is not approved

Accurate Occupancy dates that say:

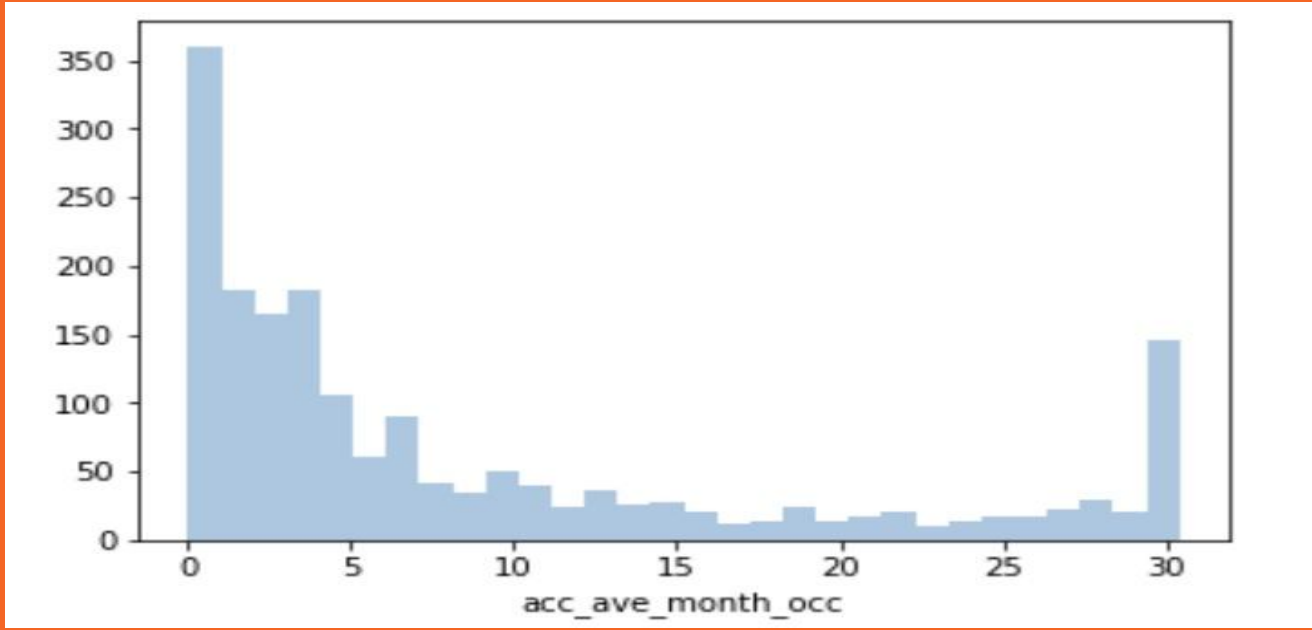
Available
or
Booked



Method of Analysis

- Obtained data ✓
- How & Why subset dataset ✓
- EDA
- Modelling
- Results

Distribution of Occupancy



Key Correlations to Occupancy

Modelling

- Linear Regression
- LassoCV
- RidgeCV

- Logistic Regression
- Random Forest

Modelling Results

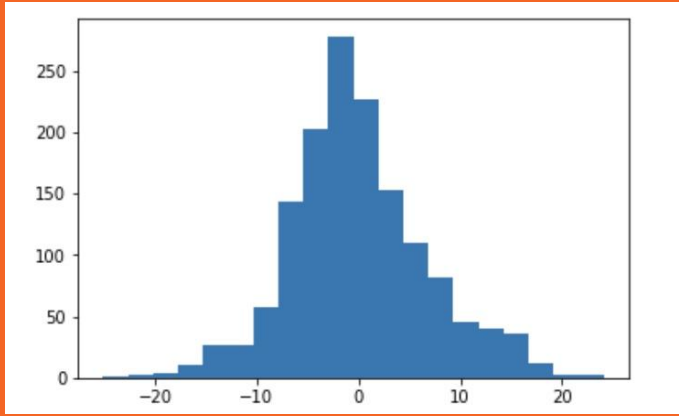
R2 score of Ed's model level:

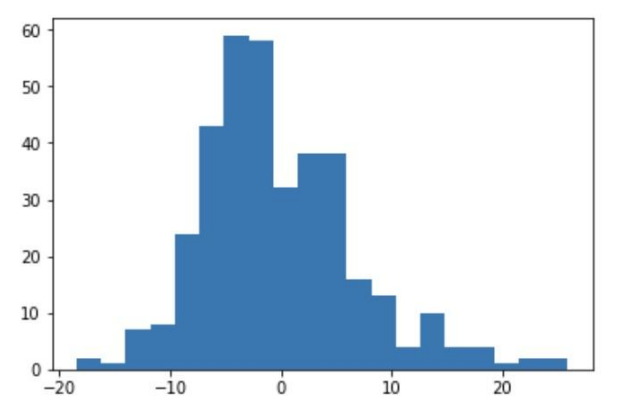
0.445

R2 score of Inside AirBnB's Occupancy level:

-1.274

Distribution of Prediction

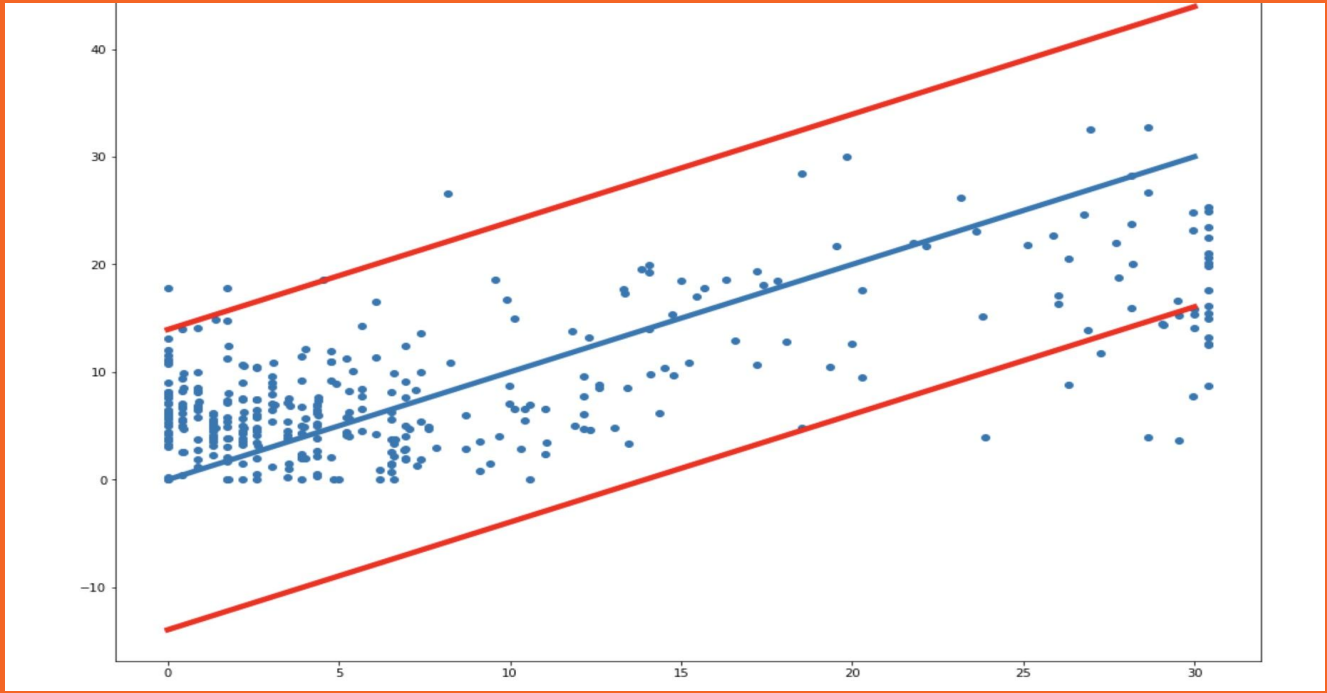




Train

Test

Predictions:



Mean Square Error

Inside AirBnB MSE: **200.9**

My MSE : **48.6**
(better by .75!)

Can be 95% confident that the true occupancy amount is between

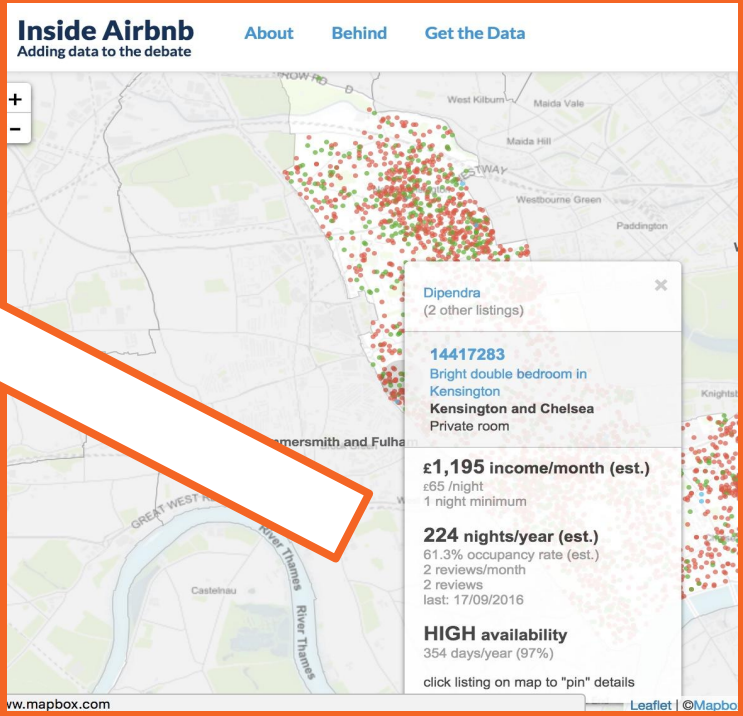
-6.97 & +6.97 of the prediction

(*As the model is built on professional AirBnB hosts who have fine-tuned their listings, these results are optimistic but since the same tools are available to all, they should be achievable for all)

*(** Predictions for the months of July and August)*

Ability to improve individual predictions

- 0 Days - 8.21 days/per month
- £0 - £533.65 income/per month
- Not £1,195 as Inside AirBnB predicts
- More realistic figure for hosts



In conclusion