CX1015 - Introduction to Data Science and Artificial Intelligence
Mini Project

AirBnB
open data from Seattle

a presentation by Lim Jia En, Mohit Prashant and Suhana Gupta

Presentation Outline

The Seattle AirBnB Dataset

Calendar

A list of dates on which particular AirBnBs (listing id) were booked and the prices they were booked at.

Listing

A list of listing ids and their corresponding information (price, host, location, description, superhost status, rating, amenities).

Reviews

A list of reviews and their corresponding information (listing id, date, reviewer).

Objective

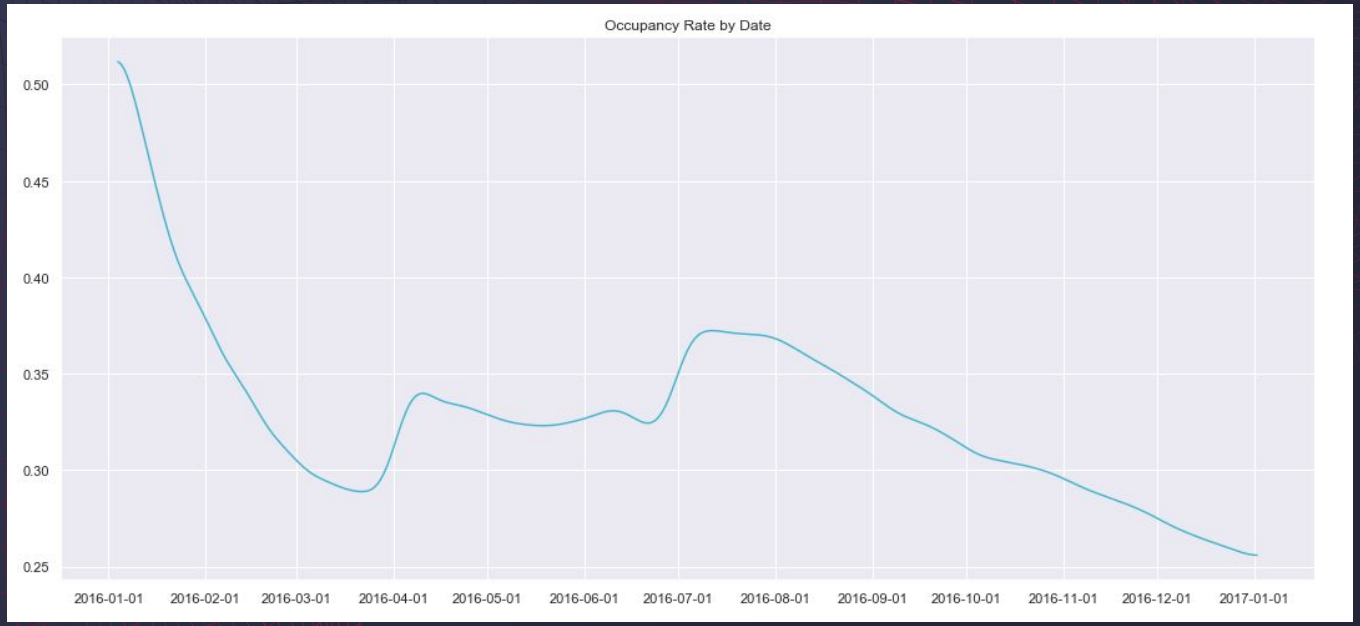
Find out what makes an AirBnB listing popular from the data presented to us.

Preliminary Exploration

01

SEASONAL TREND IN HOUSING AND OCCUPANCY

timeplot and explanation



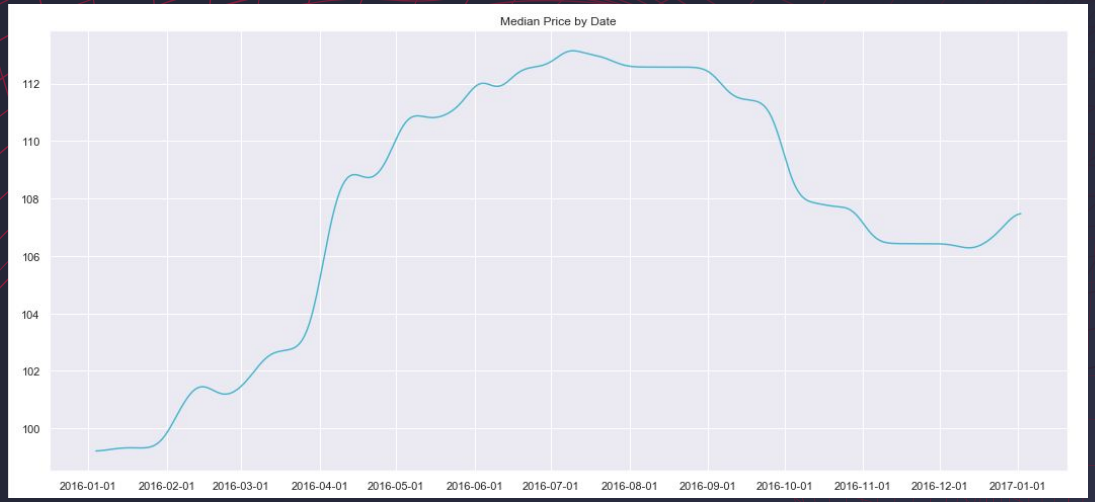
- highest occupancy: Jan, 2016
- lowest occupancy: Jan, 2017
- sporting events such as the NFL have a great impact on occupancy rates
- spring and summer are the most popular seasons to rent an AirBnB in Seattle
- winter is the least popular season as many don't wish to go on holidays

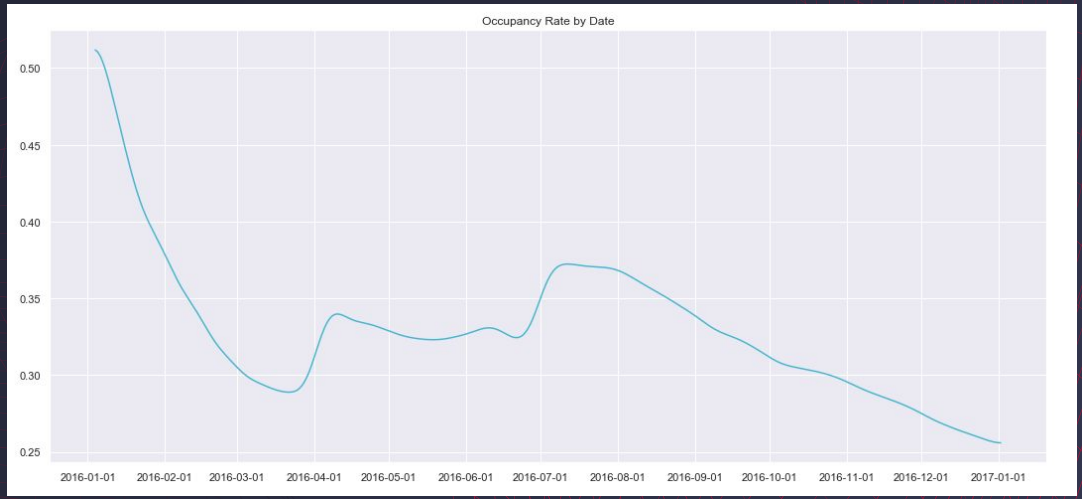
1.1

TREND BETWEEN THE OCCUPANCY RATE & PRICES OF AIRBNB

timeplot and explanation

- there is a correlation between price and occupancy rate
- the median price of rental increases when the occupancy rate is peaking
- during popular vacation seasons, such as spring and summer, AirBnB owners will increase the prices of AirBnBs
- during unpopular seasons such as winters, AirBnB owners will reduce their prices to attract customers





02

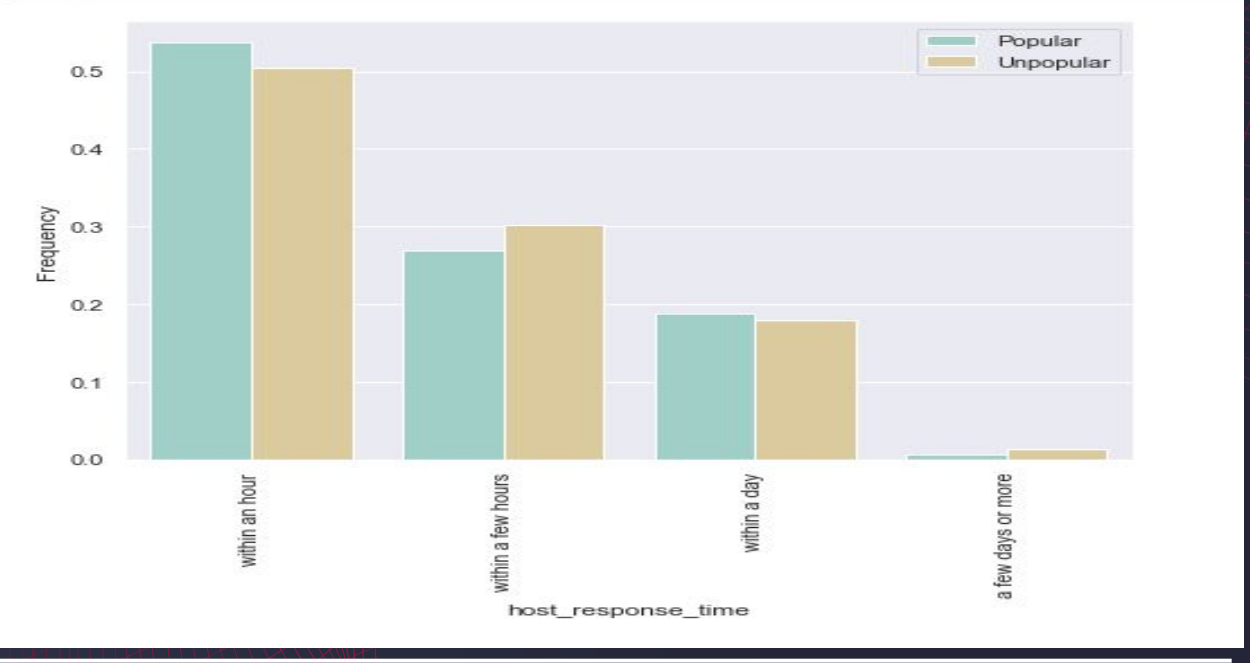
POPULAR VS UNPOPULAR AIRBNB

What are the features/traits that
makes a good and bad Airbnb?

2.1

HOST RESPONSE RATE

How fast the host replied to interested applicants?



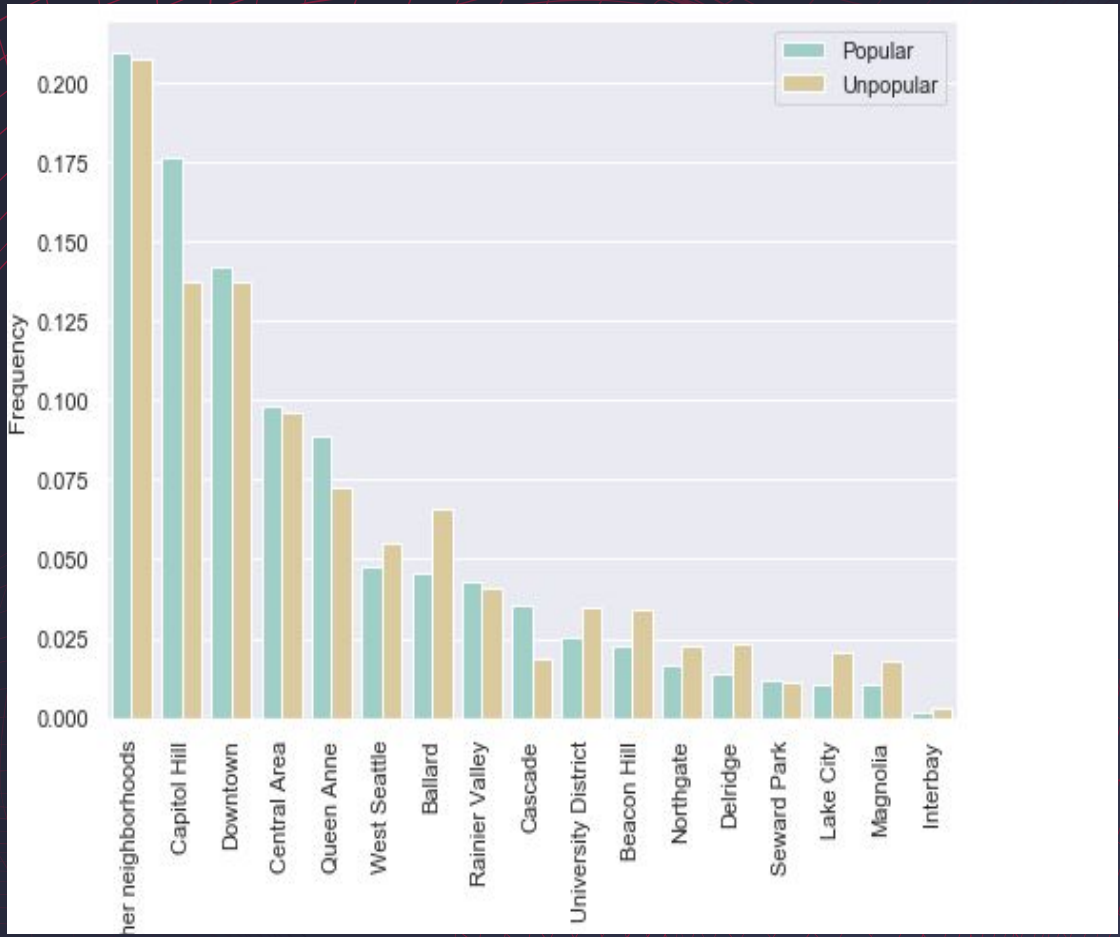
- Fast replies
- Provides clarity for interested people who are branching into the Airbnb Industry

2.2

NEIGHBOURHOODS

A District/Community within a town or a city

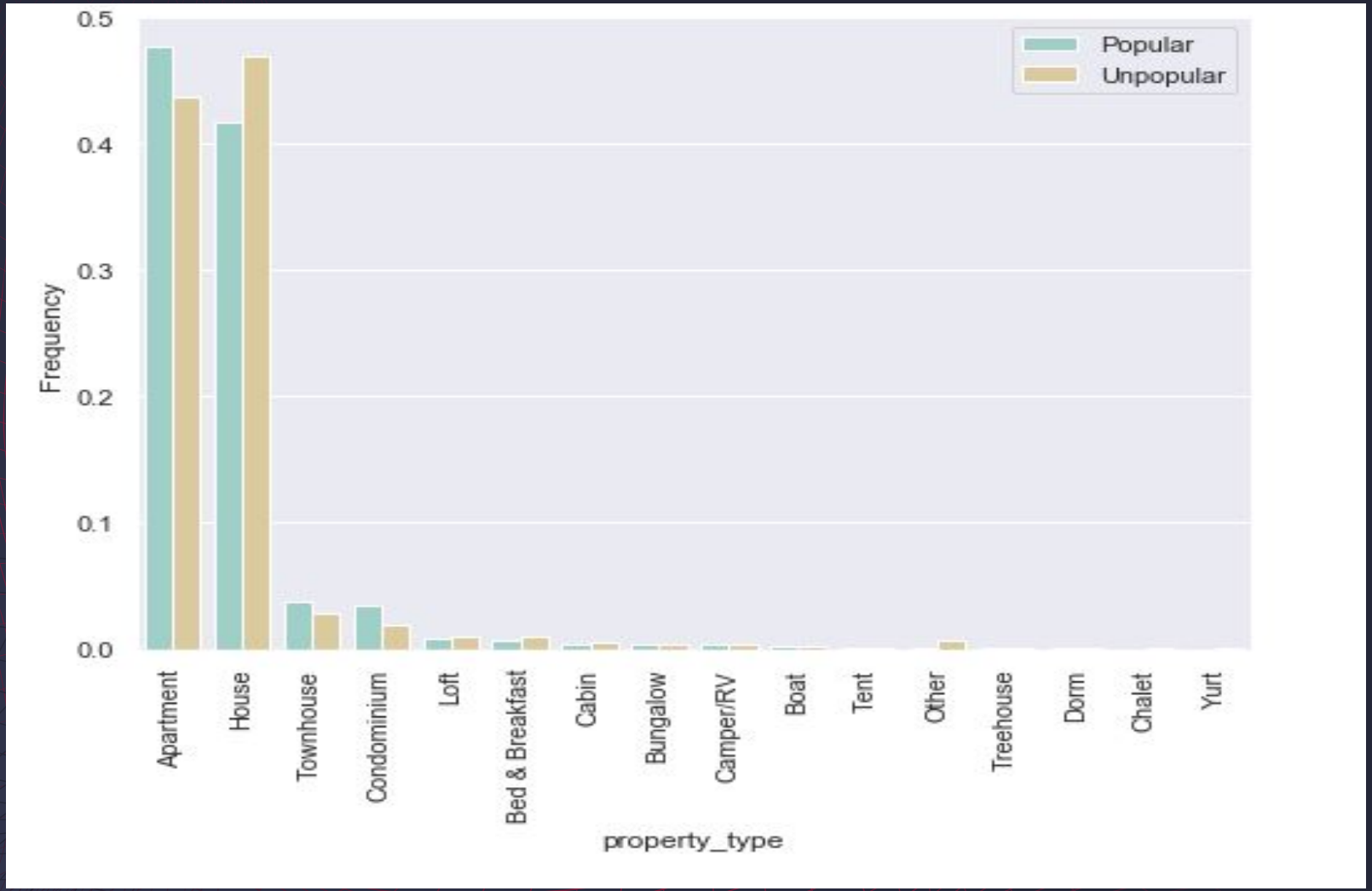
- Capitol Hill neighbourhood is the most popular
- Research also shows that there are many entertainment outlets
- For eg. Lake City is the least popular as it is not a neighbourhood that caters to tourist



2.3

PROPERTY TYPE

Different type of properties/
setting of Airbnb houses.

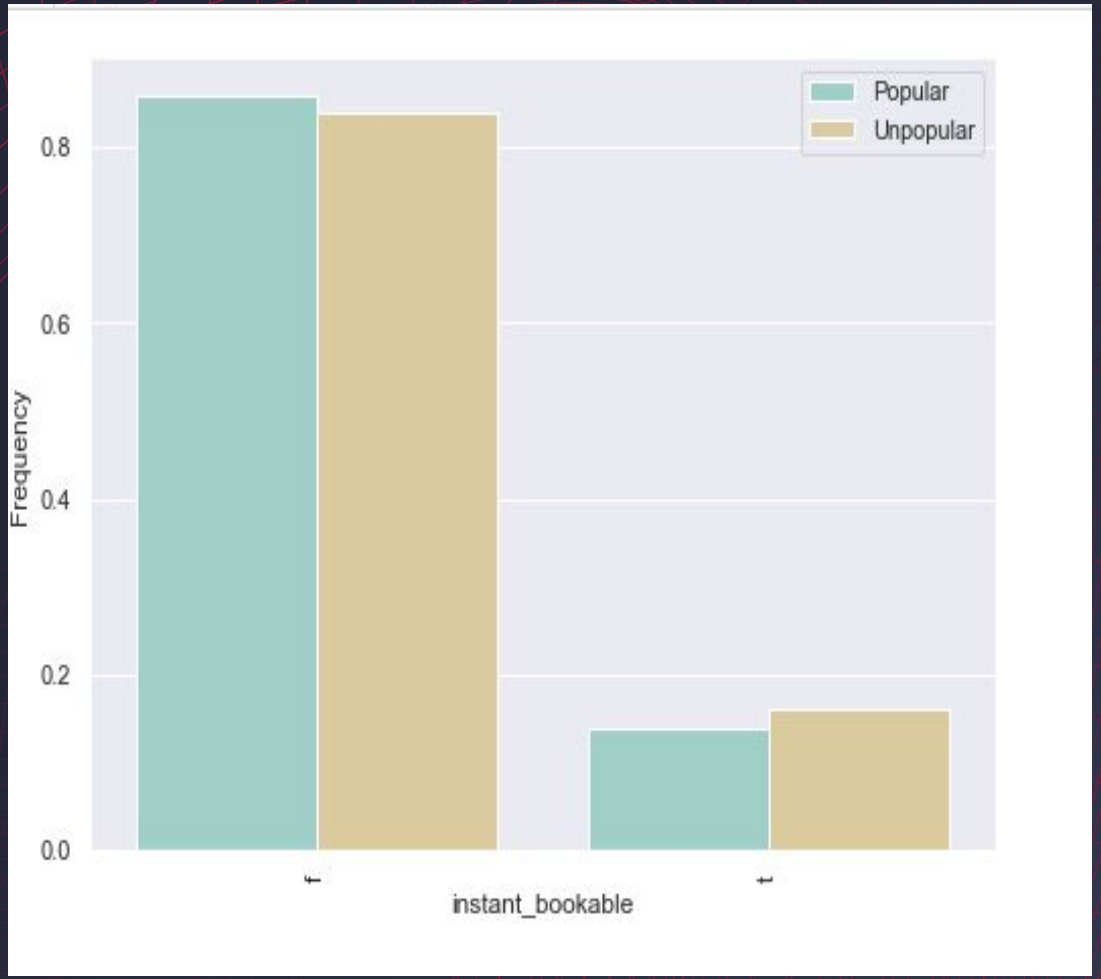

Popular
Unpopular
Frequency
0.5
0.4
0.3
0.2
0.1
0.0
Apartment
House
Townhouse
Condominium
Loft
Bed & Breakfast
Cabin
Bungalow
Camper/RV
Boat
Tent
Other
Treehouse
Dorm
Chalet
Yurt
property_type

2.4

INSTANT BOOKABLE

Booking that does not require the approval of the owners

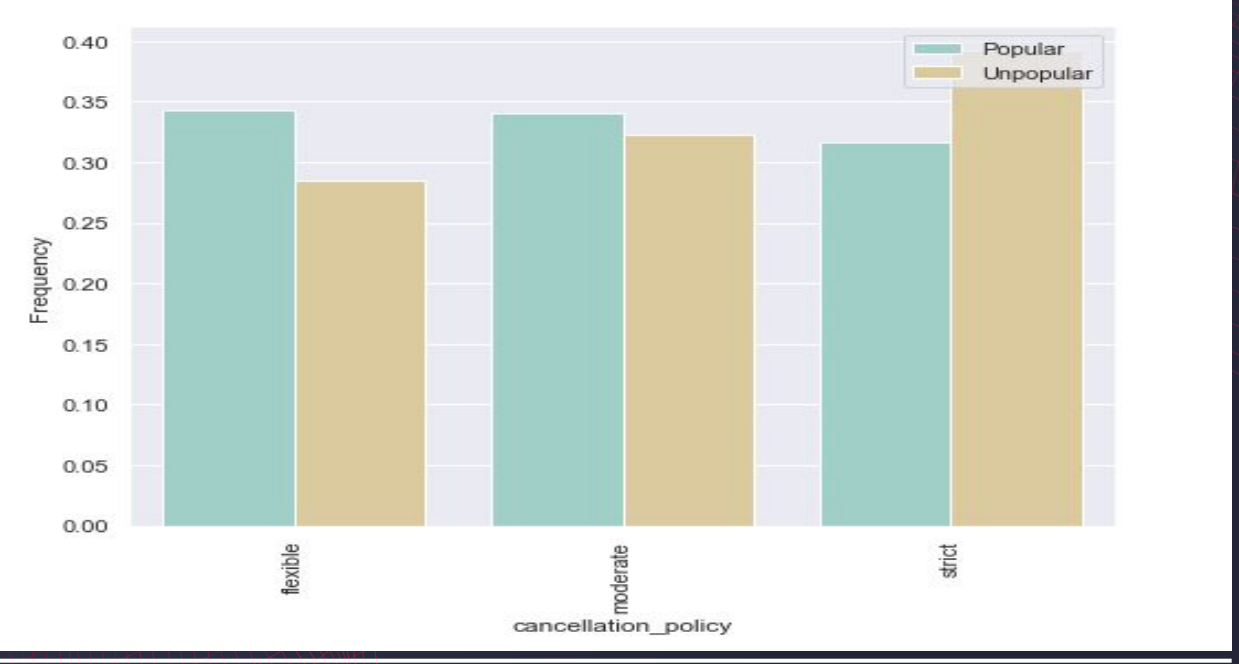
- Results of Instant Bookable in popular and unpopular Airbnb is about the same
- Conclusion, Instant Bookable does not impact heavily on applicants wanting to book an Airbnb



2.5

CANCELLATION POLICY

Terms and Condition for
cancellation of bookings

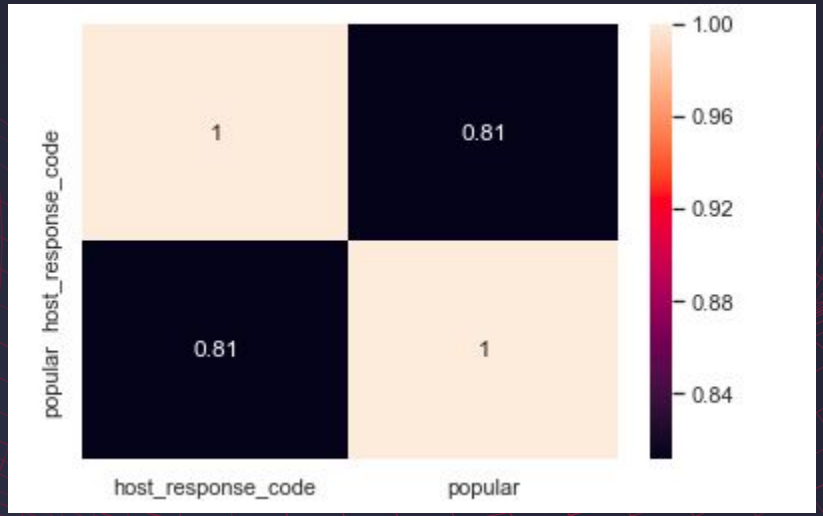


- More flexible cancellation entices applicants to book their stay

03 CORRELATION COEFFICIENTS OF FEATURES

- able to do that for host_response_time, instant_bookability, and cancellation_policy as those could be mapped from categorical to quantitative data in sorted order
- for neighborhood and property_type data, the same calculation was not possible as there's no real way to numerically sort these types as they are not ordinal categorical variables
- instead for these two data types, we found the top 5 neighborhoods and top 5 property types that had the highest number of popular airbnbs per our calculations

3.1
HOST RESPONSE RATE



We see fairly good correlation between higher response codes (meaning faster responses) to popularity rates ~ 0.81

3.2

NEIGHBOURHOODS

```
Broadway
Belltown
Minor
Wallingford
Fremont
```

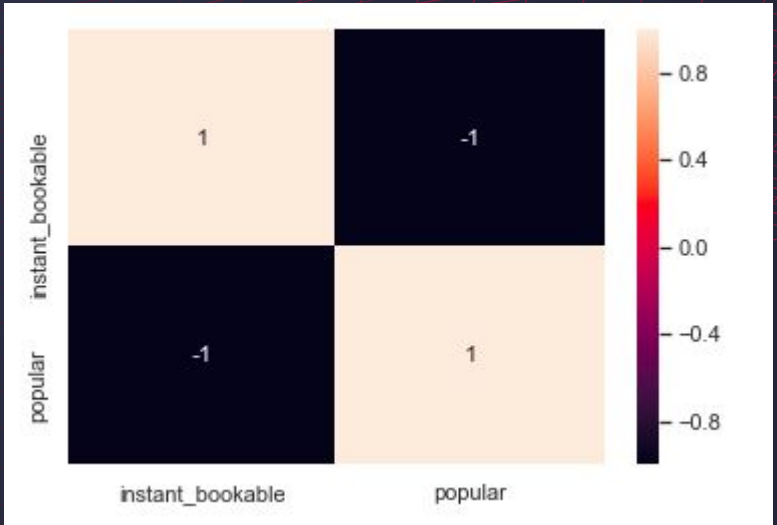
We can now see the top 5 neighborhoods had the highest number of popular airbnbs

3.3
PROPERTY TYPE

```
Apartment
House
Condominium
Townhouse
Bed & Breakfast
```

We can now see the top 5 property types that had the highest # of popular airbnbs

3.4
INSTANT BOOKABLE



We see a direct (-1) negative correlation between instant bookability and popularity. This directly makes sense because
those that are instantly bookable, likely have little to no demand, meaning high availability and thus low popularity
by our calculations.

3.5

CANCELLATION POLICY

We see surprisingly good correlation between higher cancellation codes (meaning moderate or strict policies) to popularity
~ 0.74

ML Problem

How do the reviews a superhost receive vary from a regular host and can we predict if a host is a superhost based solely on reviews?

What type of ML problem is it?

Classification of Superhost vs. Not-Superhost

using textual data (listing reviews).

	listing_id	id	date	reviewer_id	reviewer_name	comments
0	7202016	38917982	2015-07-19	28943674	Bianca	Cute and cozy place. Perfect location to every...
1	7202016	39087409	2015-07-20	32440555	Frank	Kelly has a great room in a very central locat...
2	7202016	39820030	2015-07-26	37722850	Ian	Very spacious apartment, and in a great neighb...
3	7202016	40813543	2015-08-02	33671805	George	Close to Seattle Center and all it has to offe...
4	7202016	41986501	2015-08-10	34959538	Ming	Kelly was a great host and very accommodating ...

Data Cleaning

Textual data is not in a state where it can be directly worked with and several steps of cleaning and preparation are required.

Tokenization

“Perfect location to do this and that” → [‘Perfect’, ‘location’, ‘to’, ‘do’, ‘this’, ‘and’, ‘that’]

This is necessary to uniformize the data and make the handling of data easier. Furthermore, the separation of lexical pieces is required to conduct analysis on sentences as a whole..

Lemmatization

original_word	lemmatized_word
trouble	trouble
troubling	trouble
troubled	trouble
troubles	trouble

This is necessary to uniformize the data by changing the tense and inflection such that adjacent concepts are made comparable by the parser.

Removing Stop Words

comments
Cute and cozy place. Perfect location to every...
Kelly has a great room in a very central locat...
Very spacious apartment, and in a great neighb...
Close to Seattle Center and all it has to offe...
Kelly was a great host and very accommodating ...

cleancomments
[cute, cozy, place, perfect, location, everyth...
[kelly, great, room, central, location, beauti...
[spacious, apartment, great, neighborhood, kin...
[close, seattle, center, offer, ballet, theate...
[kelly, great, host, accommodating, great, nei...

This is necessary to remove unwanted information/information with limited value/information that adds noise to the data.

Stop word examples: to, and, has, a, in, all, it, he, she...

Model Used:
Random Forest with Adapted TF-IDF

TF-IDF

Term Frequency X Inverse Document Frequency

A word is more important if it is more frequent in a particular class of documents than others.

Adapting TF-IDF

A vector was created with all words and all instances of the word were counted.

word	0	1
great	52401	32780
stay	42993	30159
place	43290	26659
seattle	32492	23563
's	32614	21157
clean	25166	15774

A score was assigned to each word based on the count using this formula:

word	0	1	total	score
nappers	60	0	60	-0.273479
merchant	79	1	80	-0.273202
bunkhouse	96	0	96	-0.304602
adin	94	0	94	-0.303206
posting	1164	101	1265	-0.300781
temple	112	2	114	-0.287117

Score= $-(log(0.5 + a/(a+b)))^3 * log(0.5 + |b-a|)$

Words implying superhost were positively scored, others were negatively scored. The strength of the implication was dependent on the magnitude of the score

Scoring Reviews

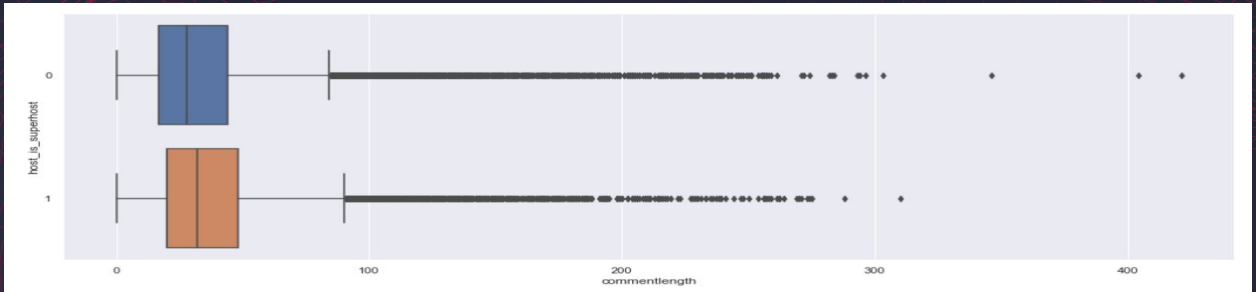
The score of each review is the sum of the scores of the words comprising it.

cleancomments
[cute, cozy, place, perfect, location, everyth...
[kelly, great, room, central, location, beauti...
[spacious, apartment, great, neighborhood, kin...
[close, seattle, center, offer, ballet, theate...
[kelly, great, host, accommodating, great, nei...

host_is_superhost	score	predict
0	-0.050922	0
0	-0.607278	0
0	-0.440571	0
0	-0.285162	0
0	-0.273824	0

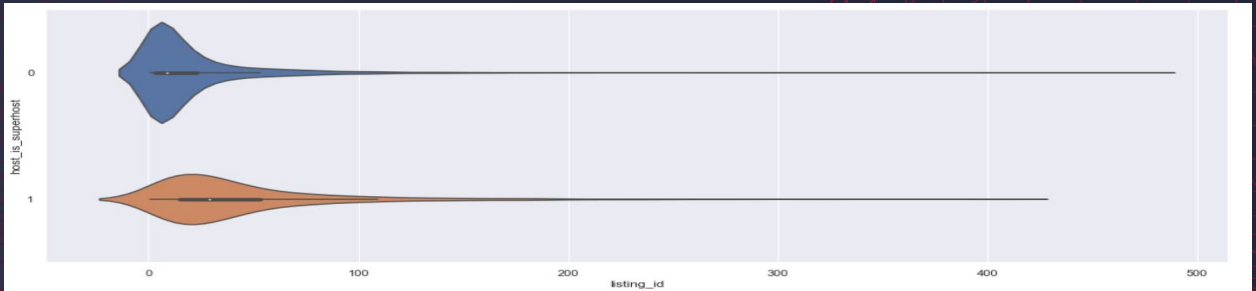
```
Accuracy :  0.6869069090308967
False Positives :  0.0032888920323938184
False Negatives :  0.30980419893670946
True Positives :  0.07311006589572208
True Negatives :  0.6137968431351747
```

Observations from Further Exploration



Comment Length vs. Superhost Status

Review Frequency per Listing vs. Superhost Status



Random Forest Classification

An ensemble classifier that instantiates multiple tree models and analyses them to find the optimal splits and weightages.

The Model Abstraction

[Data]

↓

Split by TF-IDF Scores

↓

Split by Comment Length

↓

Split by Review Frequency

↓

[Classify]

```
Accuracy :  0.7759545449187207
False Positives :  0.03789888130518325
False Negatives :  0.186146573776096
True Positives :  0.19676769105633554
True Negatives :  0.5791868538623852
```

CONCLUSION

- the main objective of our project wa to better understand the AirBnB Seattle data set.
- conducted initial exploratory analysis
- aim: differentiate between a super host and a normal host (analysed factors such as reviews left by guests, listing frequency
- also tested the popularity vs unpopularity of various listings (compared frequency of occupancy against various factors such as property type and neighborhood)
- beyond the scope of our course: we used NLTK to clean our dataset to better understand our machine learning problem.

THANK YOU!

any questions?