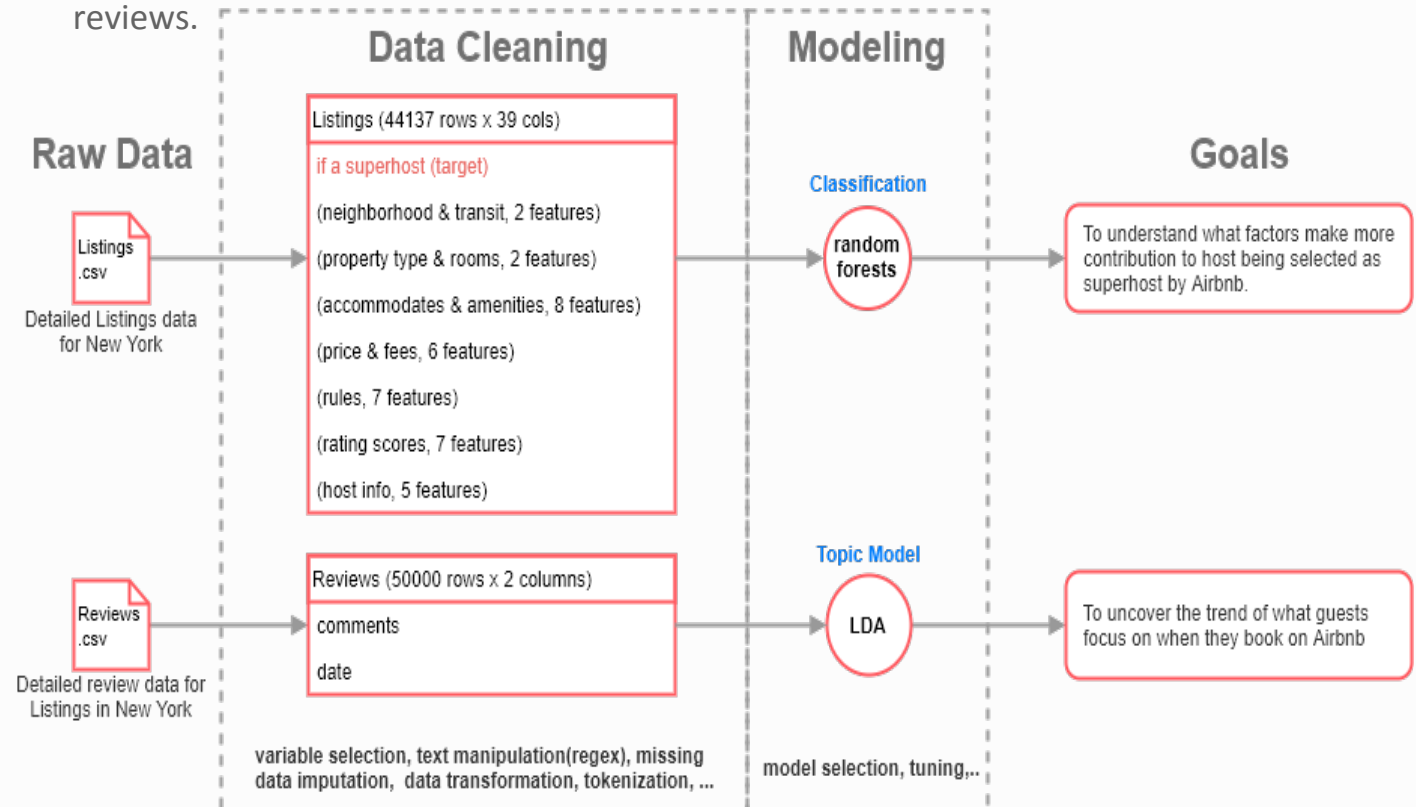# INTRODUCTION

If you ever scrolled through Airbnb you have surely seen some renters as 'Superhosts'. However, no exact rules telling how to get the badge. The project aims to understand what does it mean to be a 'Superhost' and What characteristics do the existing 'Superhosts' have. Therefore, the renters on Airbnb can know what they exactly need in order to become 'Super'.

- A random forests model was applied to listings data to classify if a host being selected as 'Superhost' by Airbnb;
- The second goal was achieved by fitting topic model to extract topical trend with guest reviews.

Data Cleaning

Modeling

Raw Data

Listings .csv — Detailed Listings data for New York

Listings (44137 rows x 39 cols)
if a superhost (target)
(neighborhood & transit, 2 features)
(property type & rooms, 2 features)
(accommodates & amenities, 8 features)
(price & fees, 6 features)
(rules, 7 features)
(rating scores, 7 features)
(host info, 5 features)

Classification
random forests

Goals
To understand what factors make more contribution to host being selected as superhost by Airbnb.

Reviews .csv — Detailed review data for Listings in New York

Reviews (50000 rows x 2 columns)
comments
date

Topic Model
LDA

To uncover the trend of what guests focus on when they book on Airbnb

variable selection, text manipulation(regex), missing data imputation, data transformation, tokenization, ...

model selection, tuning,...

Yiqiao Chen

DATA MINING ON LISTINGS

- Applied Random Forests (500 trees, 'gini') to listings data and calculated the importance for each feature;
- Selected the top 15 important and plotted the bar chart;
- According to the plot we can see can image a 'Superhost' to be like:

1) **reviews**
   - Having enough reviews and a good overall rating;
   - Sharing accurate information;
   - Providing a valuable booking and well located place.
   - Being helpful and responsive;
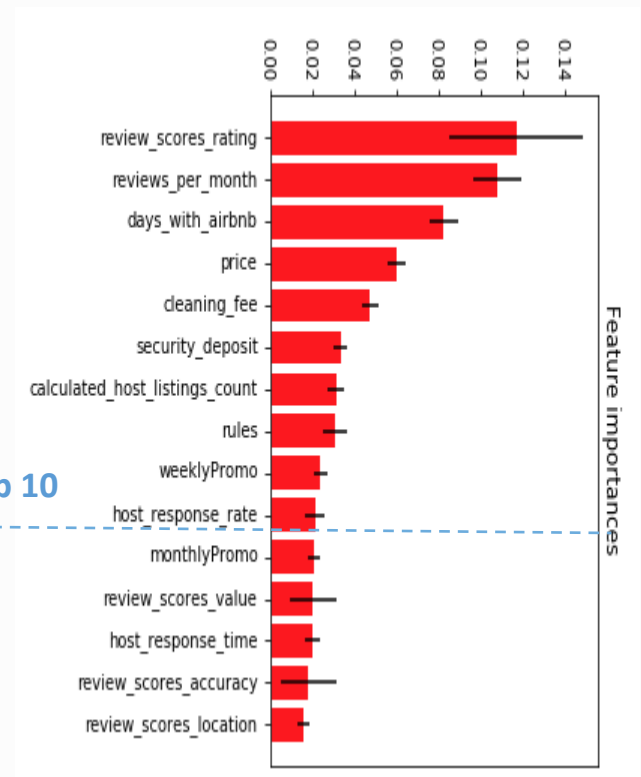   - Specifying rules clearly but not excessively;

2) **price and promotions:**
   - Offering attractive daily price options;
   - Charging reasonable cleaning fee and deposit;
   - Giving discounts for staying longer (week/month).

3) **activities:**
   - Having a long time experience hosting on Airbnb;
   - Avoiding too many listings under his/her name;

Top 10

Yiqiao Chen

TEXT MINING ON REVIEWS

- 50k comments were randomly sampled from 840k (too big), 10k for each year (2013 – 2017, 5 years);
- Documents (comments) were tokenized, cleaned and converted to bag-of-words;
- LDA was applied to get topics (topics = 10, iterations = 400, passes = 20, chunk size = 2000);
- Most meaningful topics (5 of out 10) were selected and their trend were plotted;
- As we can see, guests care more about:
    1) Convenient external environment;
    2) Comfortable interior design;
    3) Efficient host service.

**Trends of Topics 2013-2017 in New York
(% of mentioning in sampled reviews)**

Accommodates

Host service

Location & transit

Topic0	Topic1	Topic2	Topic3	Topic4
"good"	"easy"	"great"	"host_canceled"	"question"
"room"	"restaurant"	"home"	"day"	"responsive"
"bathroom"	"new_york"	"nice"	"arrival"	"helpful"
"night"	"great"	"location"	"automated"	"easy"
"kitchen"	"city"	"stay"	"posting"	"host"
"floor"	"subway"	"comfortable"	"check"	"time"
"bedroom"	"brooklyn"	"host"	"reservation"	"quick"
"apartmen"	"park"	"apartment"	"host"	"said"
"space"	"location"	"recommend"	"arrival_automated"	"needed"
"bed"	"day"	"clean"	"canceled"	"check"