Airbnb Recommendation System

Team 5 Henry Dimlow, Pandora Shou, Zayn Sui, Cindy Zhang

Our Team

Henry Dimlow

Pandora Shou

Zayn Sui

Cindy Zhang

Table of contents

01

Overview

Overview

Our purpose and the corresponding tasks to be done:

- Create the database on a more corporate-friendly level to boost its usability by company employees
  - What kinds of tables should be contain and what kinds of columns in each table?
  - What's the linkage between tables ?
- Make usage of the dataset to provide suggestions to 2 most important client group — customers and hosts
  - Recommendation system for customers
  - Decision making system for host
- Draw insights from database to learning the current airbnb market
  - characteristics of listings in different cities
  - host operation situation and profitability

02

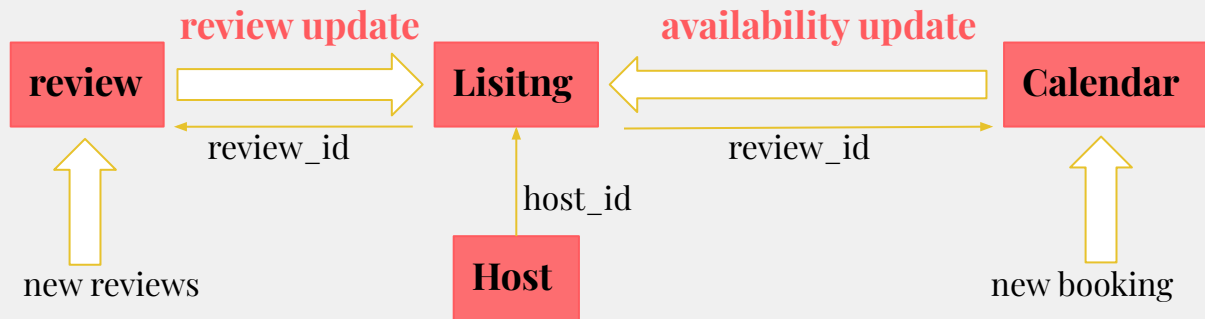# Data wrangling & Database Description

Database design & Data Wrangling

Create Tables

- Merge separate dataset for 8 cities
- Split host and listing from the original data
- Delete some text columns unhelpful
  (ie. listing descriptions)

Data Cleaning

- remove comma symbols in comments
- fill na value in rating-related columns with 0
- fill na value with 'unknown' in text columns
* We leave all the listing to keep dataset complete

review update   **availability update**

| review | → review_id | Lisitng | ← review_id | Calendar |

host_id ↑

new reviews

Host

new booking

- The Database have 4 tables: Listing, Host, Calendar, Review (2.49GB)
- Review and Calendar tables need to be updated frequently, which are also linked to some summaried columns (eg. availability in 30 days/ reviews in 30 days)

Data Source: Airbnb Inside — USA

Data storage and Processing

| approach	RDS	Hive
Dataset	Los Angeles(smaller)	California
Loading Path	local file – RDS	Local file – S3 – Hive
Processing	ER Model, Queries, Funtions	Queries,Link to Tableau
Queries	recommendation system (with sentiment analysis)	- recommendation system (without sentiment analysis) - Decision–making system

Database Description

Airbnb-California

598 neighbors **74609** listings **37975** hosts **170209** beds in total

Summary Statistics for values which is helpful to understand the overall market:

	min	Q1	median	Q3	max	stdv
Listings in a Neighborhood	1	13	41.5	143.5	2270	239.74
Hosts in a Neighborhood	1	18.82	3	12	771	52.96
Accommodates	0	2	3	6	16	2.85
Number of Beds	1	1	2	3	32	1.77
Price	0	$88	$150	$263	$999	$175.81
Price per Person	0	$33.75	$49.5	$71.4	$999	$41.45
Number of Reviews	0	1	9	46	1689	79.76
Rating	0	4.68	4.88	5.0	5.0	0.55

* We don't filter out the listing with 0 value which is unreasonable because this is a database for queries and those listing with error values are real. So in real business scenarios, those listings shouldn't be deleted but require adjustment.

03
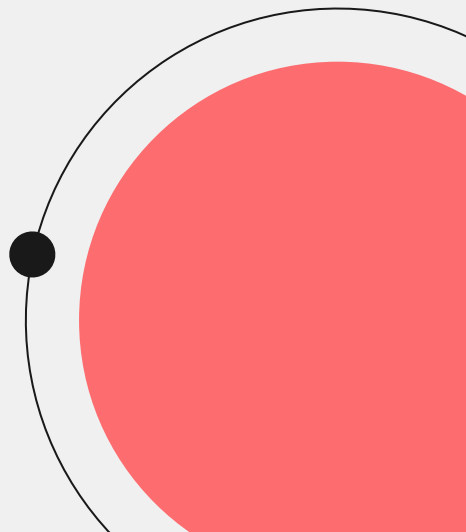
Problems Solved

Recommendation System for Customers

Customer Requirement:
City & Neighborhood
Price Range
Accommodates
Date Duration

Customer Priority:
Accuray
Communication
Cleanliness
Location

Sentiment Analysis for comments:
positive words
negative words

list of listing

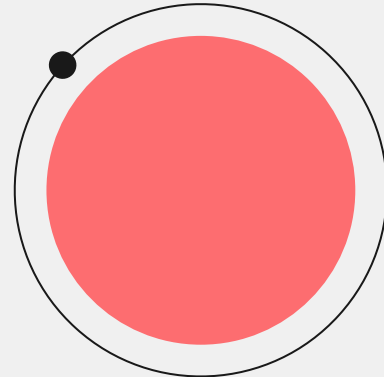**personalized rating(0-5)**

general rating(0-1)

*1

*5

order

Recommended Listings

Recommendation System for Customers

Customer Requirement:
Location: Hollywood
Accommodate: 2
Trip date: 2022-11-30 to 2022-12-03"
Price range: $80-220

Customer Priority:
Accuray
Cleanliness
Communication
Location

listing_id	overall_rating	
52763527	10	
50365526	10	
39301021	10	
48614491	10	
51752110	10	
33397503	10	
30680366	10	
25734648	9.943014569710801	
47287433	9.94	
27312821	9.935089302500469	
39732436	9.919	
53303493	9.916	
14520333	9.905000000000001	
54074518	9.9	
51691036	9.888	
52909541	9.888	
2144103	9.883278084714549	
49774867	9.867999999999999	
2326099	9.85776433195085	
5736708	9.85	
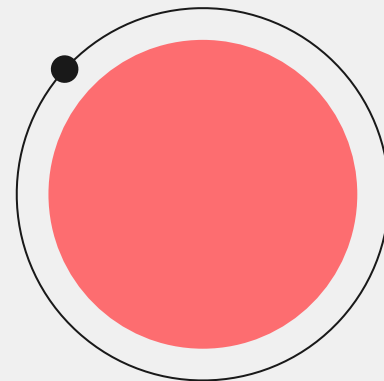
Decision-Making System for Hosts

Listing Conditions:
City & Neighborhood
Roomtype

Host Goal:
Price Grade(A/B/C/D/E)

Listing Conditions:
los angeles & Hollywood
Private room
Grade B

Results：
Estimated Occupany rate
Estimated Price
Estimated Revenue

Results：

Expected Occupancy Rate	69%
Expected Price	$87.93
Expected Revnue	$22408.94

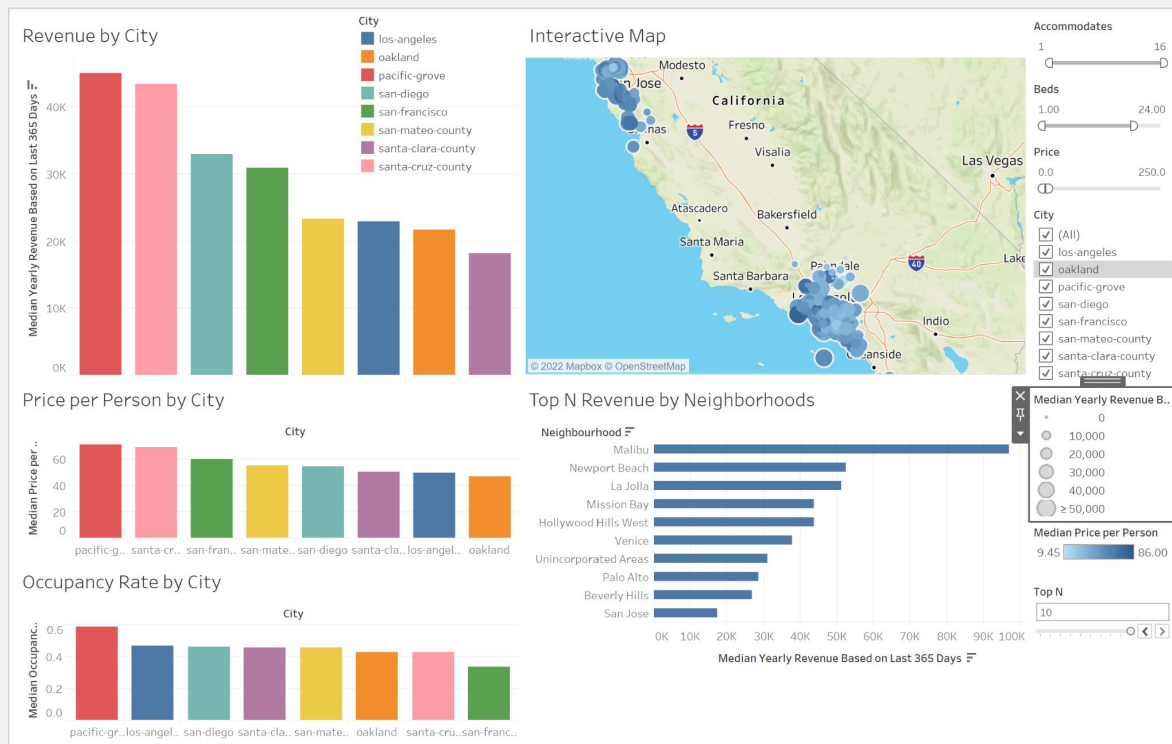*base on listings in dataset in the same area and same roomtype in next 12 months

04

Insights from Dataset

Insight from City Dashboard

- Price per Person by City

  - LA has the highest median price but is the second lowest median price per person

  - Pacific Grove has the highest price per person despite having the second lowest median price

- Occupancy Rate by City

  - Pacific-grove has by far the largest median occupancy rate

  - San Francisco has by far the lowest occupancy rate despite not standing out in terms of median price or Price per Person
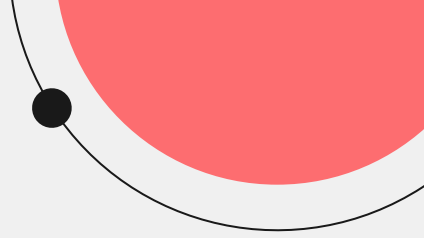
Host Insights Dashboard
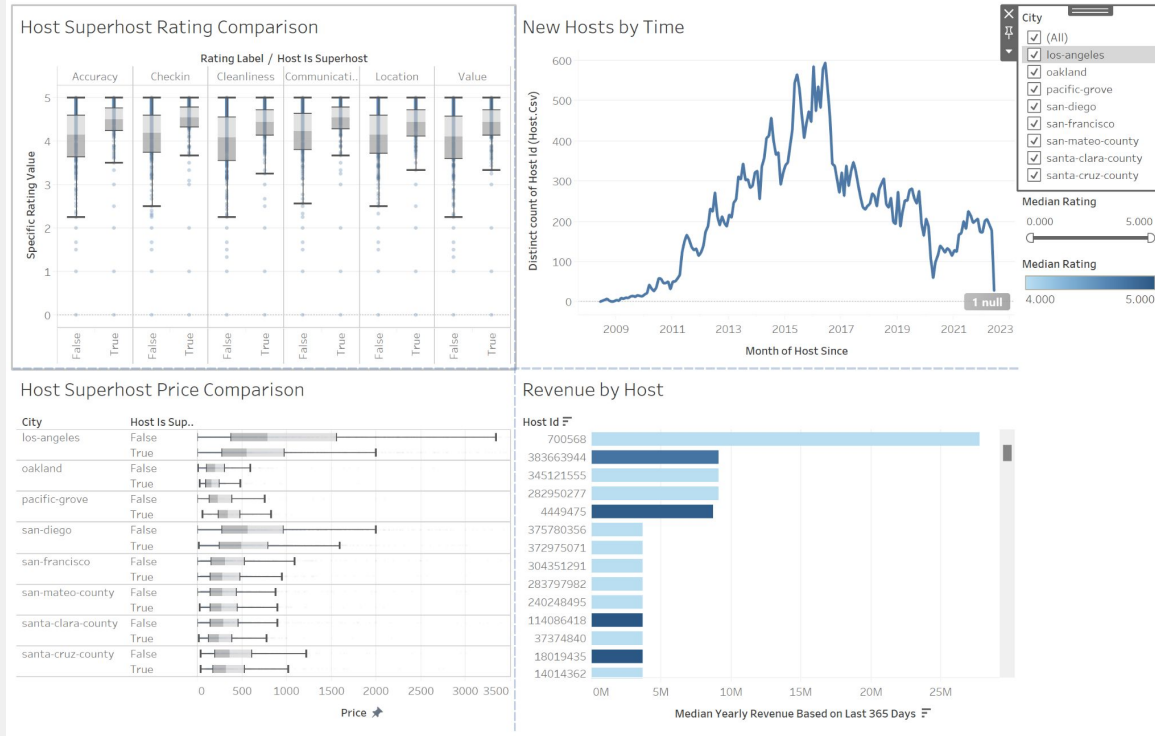
Insight from Host Dashboard

- New Hosts by Time

  - Steady increase in new hosts every year, large peaks of new hosts every summer

  - After 2016 new hosts peak and begin to decline

- Price Distribution by City and Superhost

  - LA has the largest range of IQR and the highest Median price. Oakland has the smallest range and lowest median

  - Pacific-Grove is the only city where superhosts' median price is higher than regular hosts

- Host Superhost Rating Comparison

  - Superhost's median rating is higher than non-superhosts and range is smaller

05

Application & Improvement

Application & Improvement

- Application in Airbnb
  - Embed Recommendation System in the **Airbnb application interface** to recommend listings
  - Embed Decision-making System in the **host interface** to help price their listing and analyze the potential values

- Improvement of models
  - **User priority** can be collected when they are new and used each time when they order a listing
  - More complicated analysis can be exhibited for host

- Generic Application
  - The logic behind our models can be generalized for many **platforms for both customers and merchants,** like lifestyle platforms and e-commerce websites

Thanks!

Do you have any questions?