

Regression Analysis of Airbnb Ratings

Darrick Suen, Harrison Kim

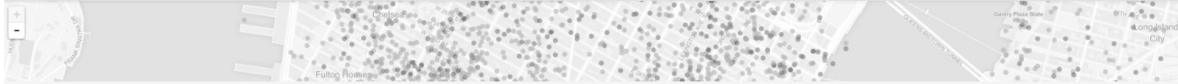
Problem

Goal: To be able to provide insight/advice on how to improve listings while minimizing investment

Analysis Method: Regression

Inside Airbnb
Adding data to the debate

About Behind Get the Data



Get the Data

The data behind the Inside Airbnb site is sourced from publicly available information from the Airbnb site. The data has been analyzed, cleansed and aggregated where appropriate to facilitate public discussion. See more [disclaimers here](#), and a data dictionary [here](#).



If you would like to do further analysis or produce alternate visualisations of the data, it is available under a [Creative Commons CC0 1.0 Universal \(CC0 1.0\)](#) "Public Domain Dedication" license.

If you have any questions, or would like to request data you don't see here, please contact data@insideairbnb.com. Please let us know who you are, your interest in the data and Airbnb. We prioritise requests based on their alignment with the project's mission - to provide free data that quantifies the impact of short-term rentals on housing and residential communities; and also provides a platform to support advocacy for appropriate and effective policies to protect our cities from the impacts of short-term rentals.

Only the last 12 months worth of data for each location is hosted here, to control data download expenses (in response to the abuse of researchers downloading large datasets, often-times for purposes totally removed from the mission of the project).

If your purpose of use for the data aligns with the mission of the project and you require access to archived data, please ask. Archived data will always be available to housing and community activists.

Some data requests require work, and if your request does not align with the mission of the project, you may be asked to [donate](#) to the project to contribute to the project's sustainability.



Datasets

Website: <http://insideairbnb.com/get-the-data.html>

- Will use all listing from the 28 available cities in United States
- ~200k Samples, 74 Columns
- 509MB Total

Feature Examples:

- Flexible: Amenities, Price, Host Response, Min/Max Nights, Description
- Fixed: # Bed/Bath rooms, Location, Superhost, Host Since

Targets: Review Rate, Review Scores

	Files	File Size (MB)
0	listings_Ashville.csv	6.709891
1	listings_Austin.csv	25.272316
2	listings_Boston.csv	9.063423
3	listings_Broward County.csv	29.132234
4	listings_Cambridge.csv	2.240993
5	listings_Chicago.csv	18.179963
6	listings_Clark County.csv	24.450829
7	listings_Columbus.csv	3.586148
8	listings_Dallas.csv	10.462730
9	listings_Denver.csv	9.762857
10	listings_Hawaii.csv	68.024691
11	listings_Jersey City.csv	3.265589
12	listings_Los Angeles.csv	79.754568
13	listings_New Orleans.csv	17.421437
14	listings_New York City.csv	86.223243
15	listings_Oakland.csv	5.828382
16	listings_Pacific Grove.csv	0.551775
17	listings_Portland.csv	10.880926
18	listings_Rhode Island.csv	8.818789
19	listings_Salem.csv	0.455231
20	listings_San Diego.csv	26.274057
21	listings_San Francisco.csv	18.288557
22	listings_San Mateo.csv	7.369669
23	listings_Santa Clara.csv	14.114541
24	listings_Santa Cruz.csv	3.619012
25	listings_Seattle.csv	12.508978
26	listings_Twin Cities.csv	10.284313
27	listings_Washington.csv	21.438617

Potential Challenges

1. Discover most relevant features
2. Addressing categorical features
3. Text Analysis for description feature
4. Multiple datasets being combined - Ensure train and test data come from all datasets
5. Columns containing lists
 - Example: Amenities: ["Hot water", "Luggage dropoff allowed"....."Keypad", "Smoke alarm"]
6. Ambiguous Columns