Airbnb Reviews Topic Modeling

Latent Dirichlet Allocation & KMeans

Daiki Minaki

Objective

Understand Airbnb review patterns to see what people remember about their stay and how it affected their ratings.

Problem

1. Airbnb has millions of reviews, but there is currently no way to sort through the topics of reviews without reading them individually.

2. Some listings have hundreds of reviews. Categorizing reviews would allow users to look closely into the areas they want to know about.

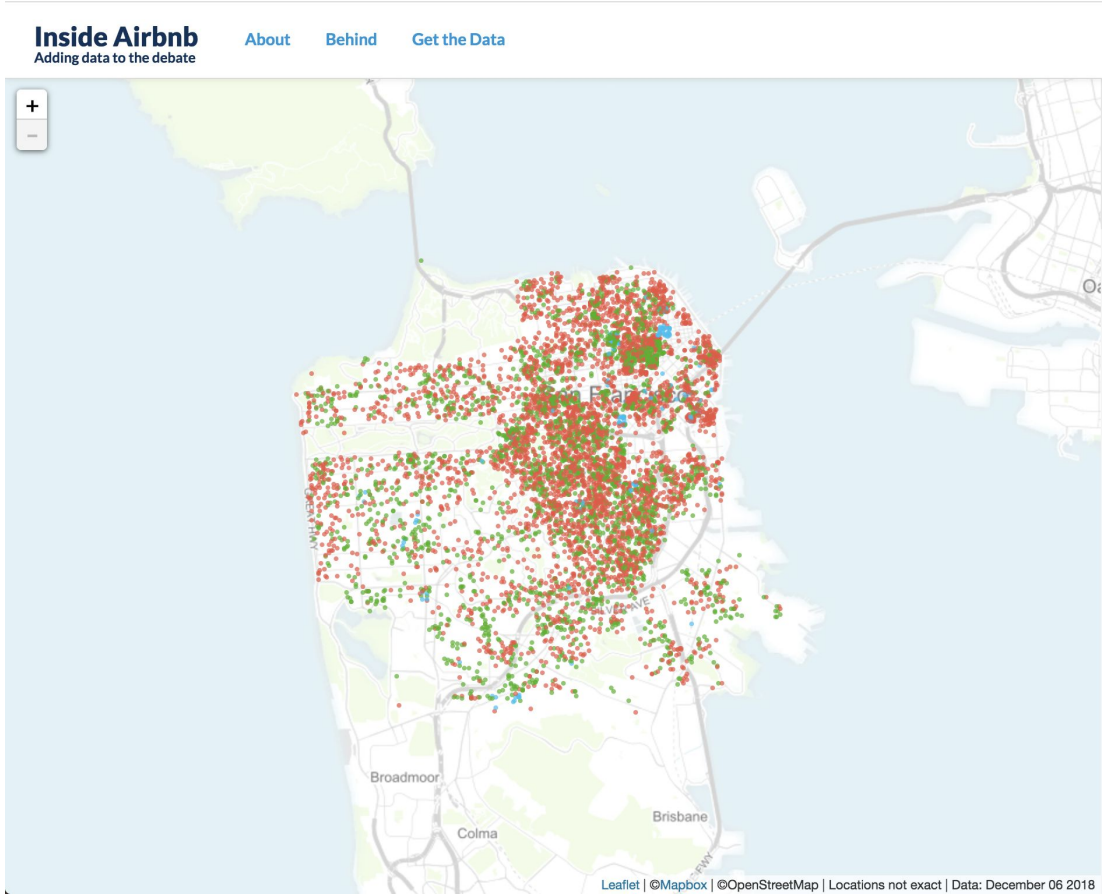
Approach

Extract Topics from Airbnb reviews using Latent Dirichlet Allocation and find patterns in review topics and listings.

About the Data

Source: InsideAirbnb.com/

- Listings.csv
- Reviews.csv
- Calendar.csv
- Neighborhood.geojson



Outline

1. Data Collection
2. Data Cleaning & Wrangling
3. Exploratory Data Analysis
4. Topic Modeling
5. Tag Review Topics
6. Analysis Using Review Content

Data Collection

Scraping the ‘Get the Data’ Page of Inside Airbnb

1. Collect Download Links for All Cities

2. Use Download Links to **Download** and **File** Accordingly (collection.py)

- **Download** CSV File Directly as CSVs
- **Download** Compressed Files, **Unzip**, and **Save** as CSV
- **Download** Geo Files as geojson

*** For This Project I will be focusing on San Francisco Listings.**

Data Cleaning & Wrangling

Cleaning Review Data

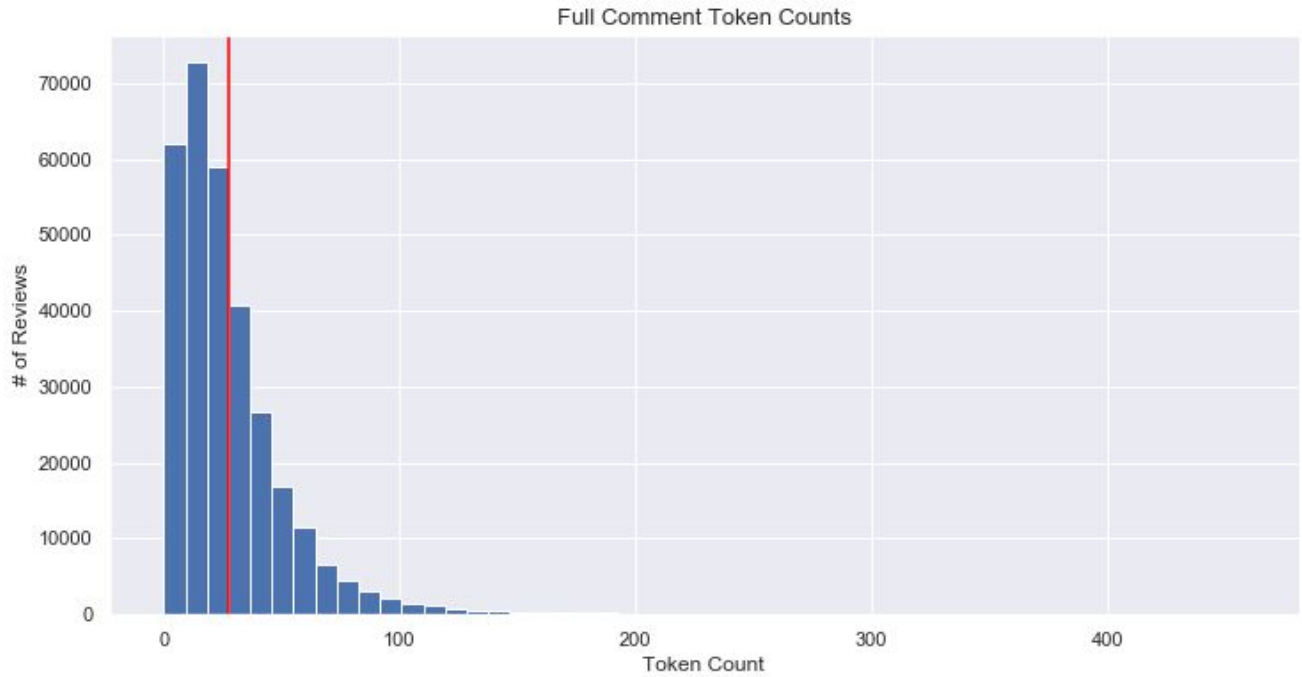
- Lemmatize
- Remove Stopwords
- Remove non-English Reviews
- Tokenization & POS Extraction

Wrangling Data

- Full Review Tokens
- Noun, Verb, Adjective Tokens
- Name Entity (NER) Tokens

Exploratory Data Analysis (EDA)

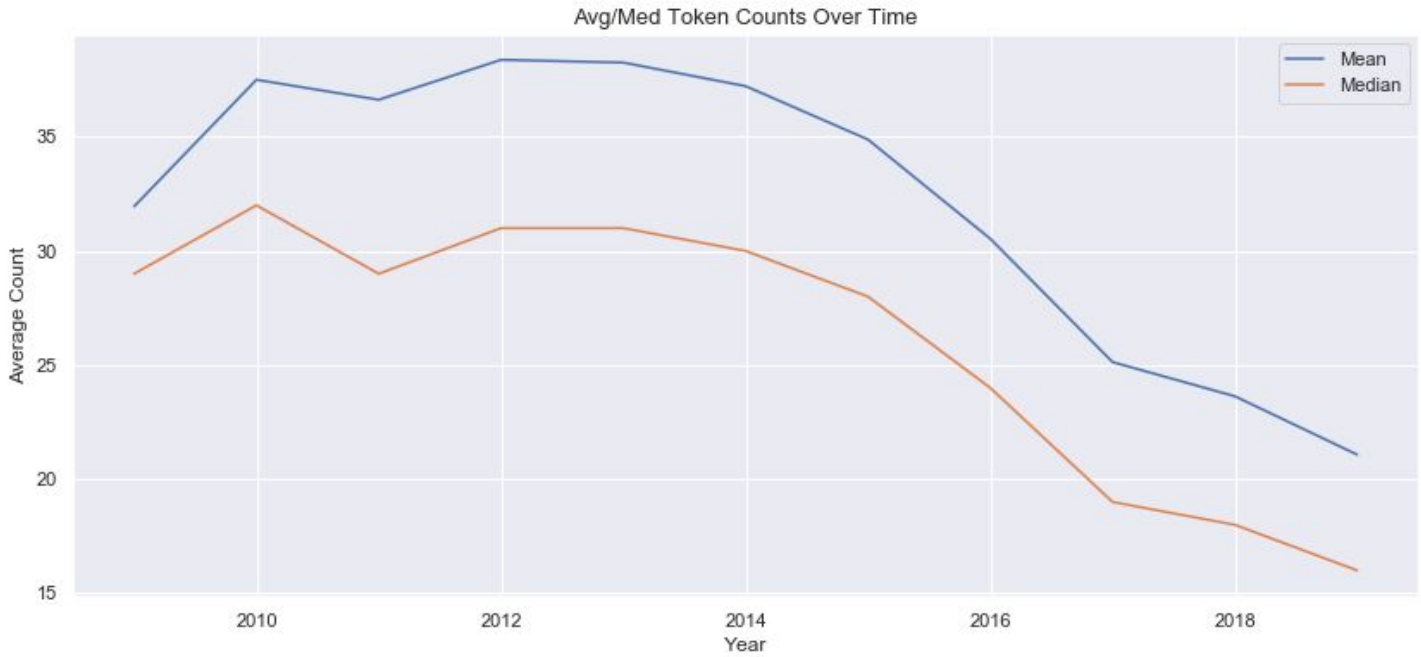
Typically how long are reviews?



Average review length (without stopwords) is around 30 words.

EDA Continue...

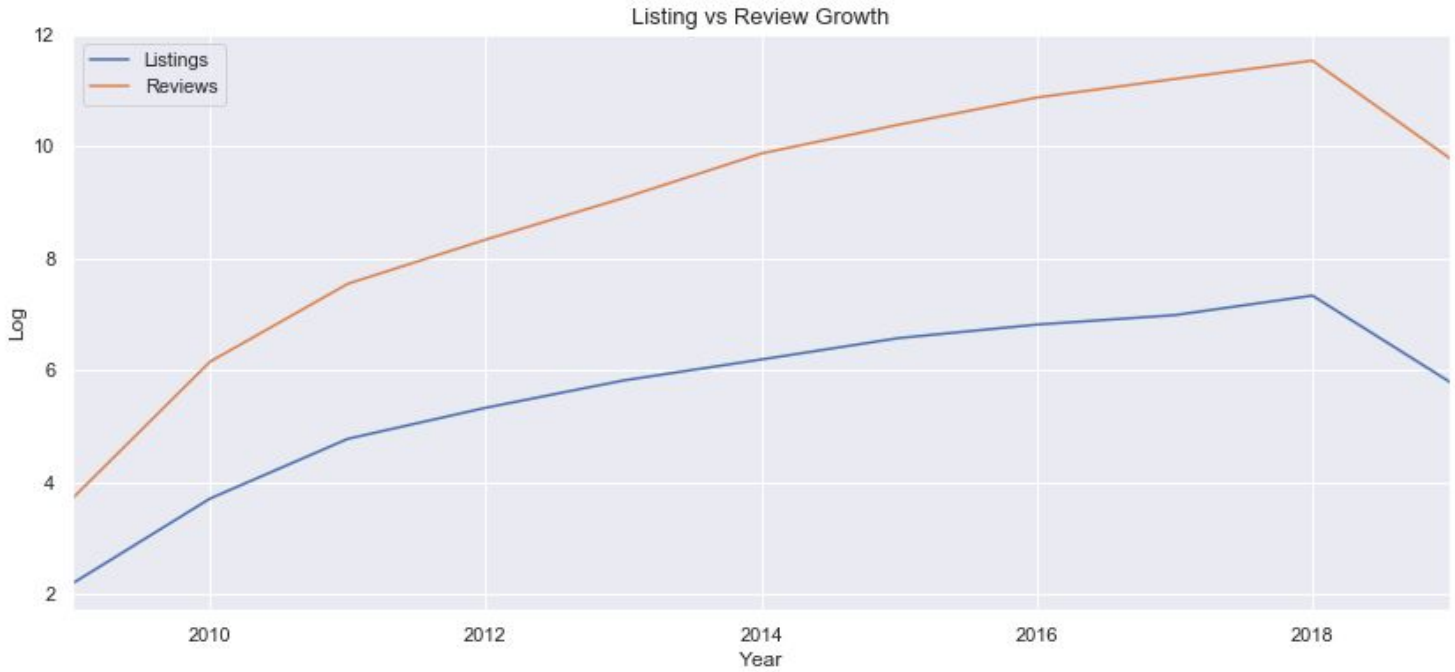
Trends in Review Length



Average review length has fallen year after year.

EDA Continue..

Listing vs Review Growth



Listing and Review growth is roughly parallel.

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation is a generative statistical model that finds **topics in text data** by looking at **recurring patterns across documents** and weighing them based on how distinct each pattern is.

Why LDA?

- With **over 300,000 reviews** we need an algorithm that will look at overall trends in text and summarize and tag individual documents.
- LDA is good at generalizing topics with some cost in terms of computing time.

Topic Selection

Parameters: 50 Topics, 10 Words, 50 Passes

Data: Review Tokens from over 300,000 Reviews

Topics:

- 10 Review Topics Chosen
- Top Value for each Topic Selected

```
Checkin & Communication
17: check, late, even, early, arrived, time, accommodating, let, flight, last
23: easy, check, communication, access, check-in, clean, super, communicate, made, instruction
27: quick, question, always, quickly, respond, response, available, responded, message, john

Host
7: great, gave, local, tip, recommendation, host, city, area, helpful, provided
8: really, enjoyed, stay, thank, much, hospitality, cat, appreciated, thanks, staying
38: wonderful, host, stay, comfortable, clean, beautiful, home, location, lovely, would

Noise
2: n't, place, would, room, night, bit, stay, noise, one, nice

Location
4: city, place, quiet, great, perfect, neighborhood, space, spot, studio, stay
22: great, location, place, stay, host, clean, would, comfortable, super, excellent
34: downtown, walk, minute, bus, uber, easy, get, away, bart, city
37: walking, distance, within, restaurant, square, location, union, many, wharf, fisherman
44: restaurant, great, shop, bar, close, coffee, walk, store, nearby, location

Come Again
41: stay, back, place, definitely, would, time, come, next, trip, visit
49: recommend, would, place, highly, stay, definitely, great, clean, staying, anyone

Parking
35: parking, car, street, family, find, easy, house, space, park, spot

Cleanliness
47: nice, place, clean, really, super, room, stay, bed, house, comfortable

House
10: room, bathroom, private, bedroom, clean, kitchen, living, space, bed, shared
14: kitchen, bed, comfortable, well, towel, shower, bathroom, space, everything, need
39: home, feel, felt, like, made, welcome, make, beautiful, comfortable, safe

Value
40: good, location, value, price, communication, facility, place, service, reasonable, elizabeth

Accuracy
33: everything, needed, exactly, need, described, apartment, picture, stay, perfect, sure
```

Sample Review Topics

Review ID: 6660
Review:
Returning to San Francisco is a rejuvenating thrill but this time it was enhanced by our stay at Holly and David's be autifully renovated and perfectly located apartment. You do not need a car to enjoy the City as everything is within walking distance - great restaurants, bars and local stores. With such amenable hosts and a place to stay that enhanc es one's holiday, we will be returning again and again.

Topics

Host: 8.0	Noise: 4.0
Come Again: 6.0	House: 3.0
Accuracy: 6.0	Parking: 1.0
Location: 6.0	Value: 1.0
Cleanliness: 5.0	Checkin & Communication: 1.0

Review ID: 11519
Review:
We were very pleased with the accommodations and the friendly neighborhood. Being able to make a second bed out of th e futon couch was particularly helpful. Having a full kitchen, a lovely walkout garden, and TV + DVD + FM stereo were added bonuses. Holly and David were most gracious and met our every request. Being within walking distance of both th e Haight Street and Castro Street scenes was great. The only negative for us was the difficulty in finding on-street parking, due to both the density of the neighborhood and the construction work going on. A few evenings we had to par k about 4 blocks away. For people who need to use private vehicles, this problem does restrict their plans somewhat. For people who can use public transportation and/or walk, this is no problem.

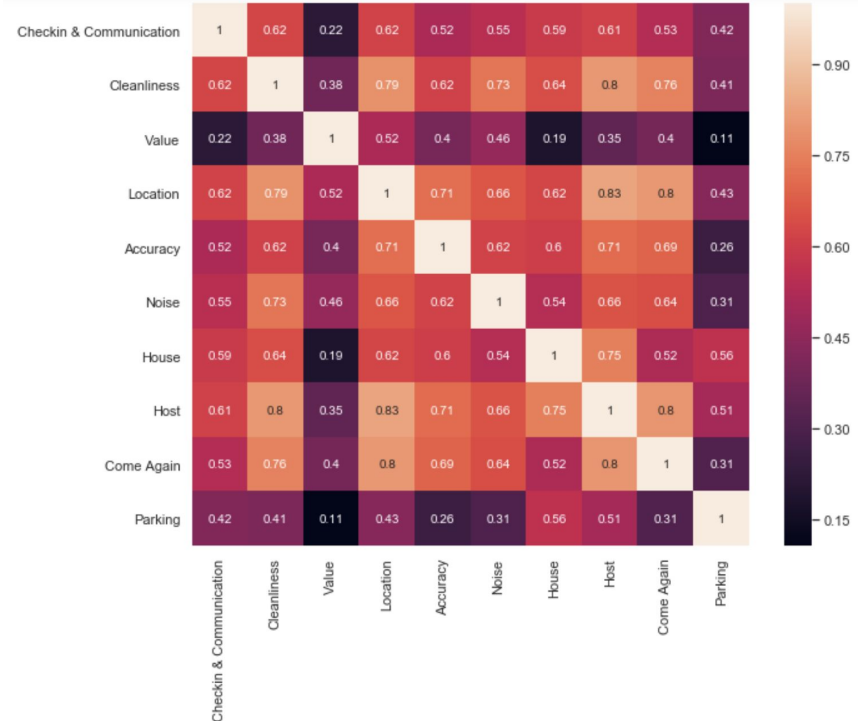
Topics

Location: 9.0	Come Again: 3.0
Cleanliness: 6.0	Accuracy: 2.0
Host: 5.0	Checkin & Communication: 2.0
House: 5.0	
Parking: 4.0	

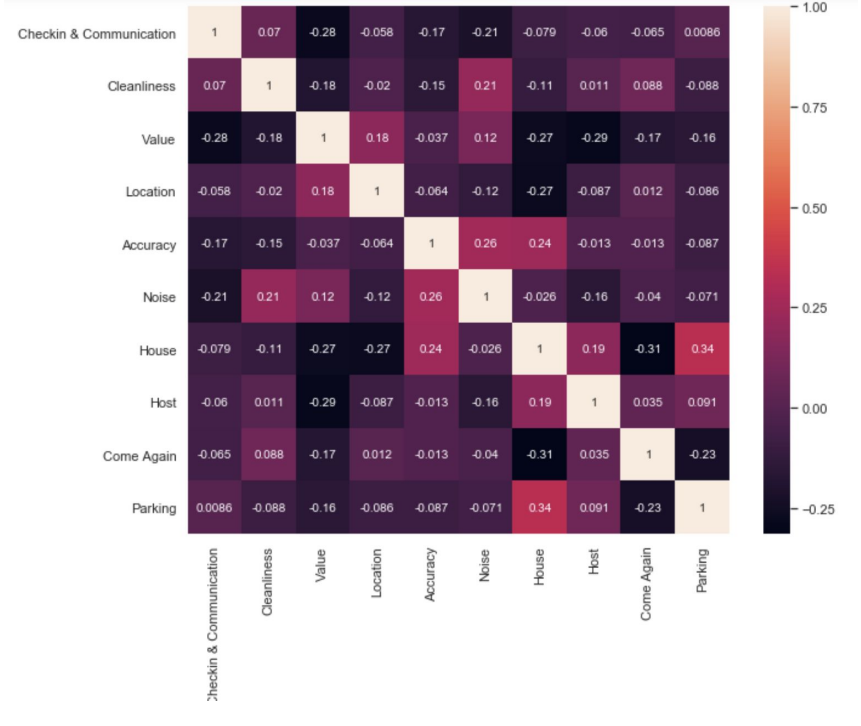
Normalizing Review Topic Scores

*Normalize Topic Scores To Get Similar Ranges

Before Normalization



After Normalization



Topic Analysis & Clustering

Are there any patterns in the kind of reviews that listings receive?

K Means Clustering

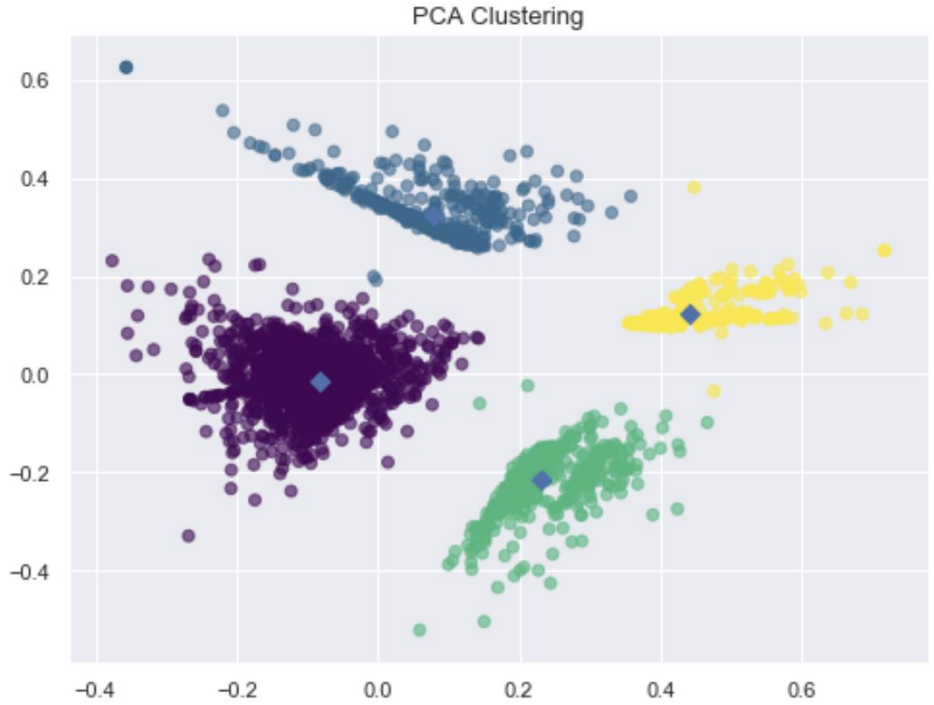
Distinct Clusters of Listings

Data: Review Topics by Listing

PCA: Top 3 Features

Clusters: Knee/Elbow Method

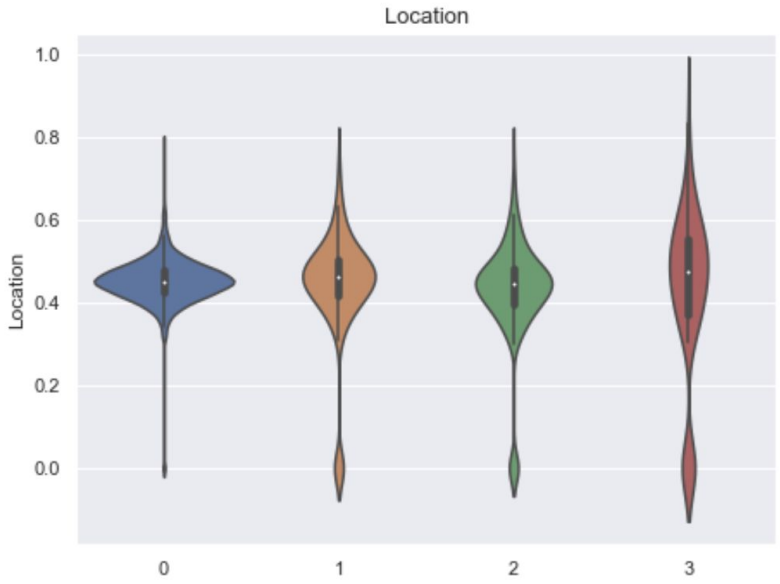
N_clusters: 4 Clusters

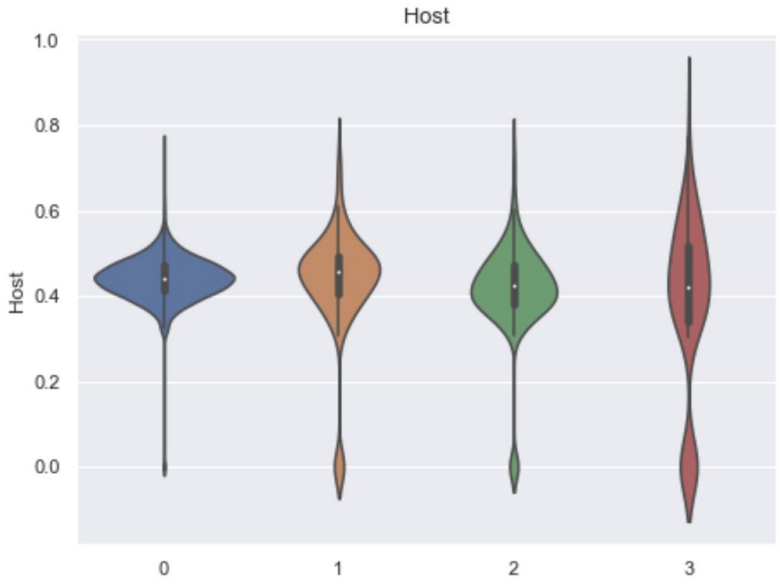


Most Review Topics Were Similar

Most Review Topics were Similar Across Clusters.

Ex. Location, Host, Value, Accuracy, House, etc





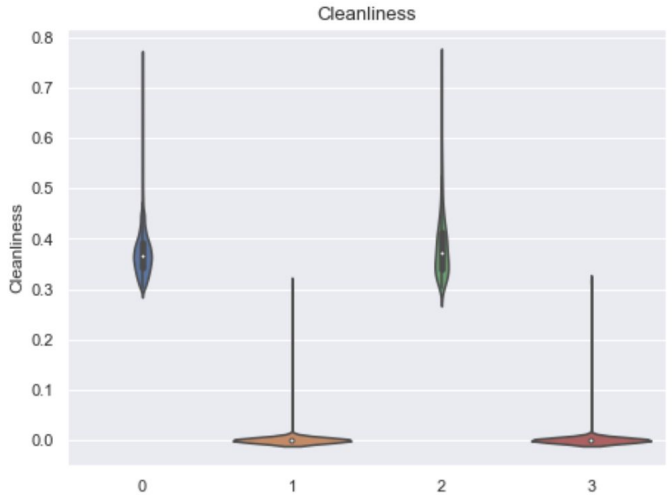
Key Topic Differences

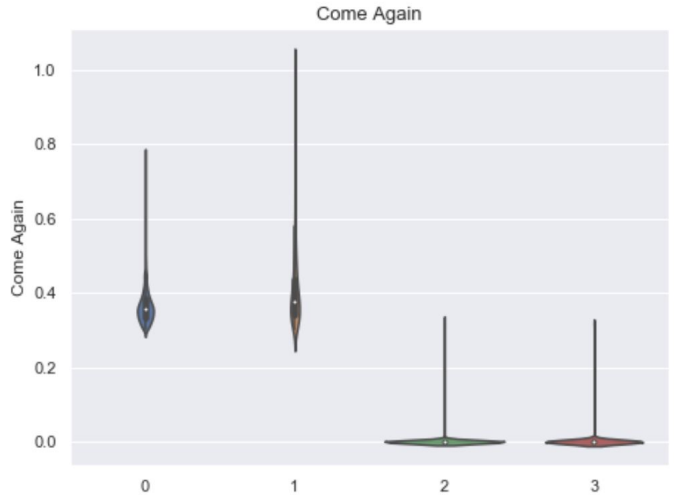
Cleanliness

- Clusters 0 and 2 are High
- Clusters 1 and 3 are Low

Come Again

- Clusters 0 and 1 are High
- Clusters 2 and 3 are Low





Cluster Features

Cluster 1:

- **Low** Cleanliness
- **High** Come Again

Cluster 0:

- **High** Cleanliness
- **High** Come Again

Cluster 3

- **Low** Cleanliness
- **Low** Come Again

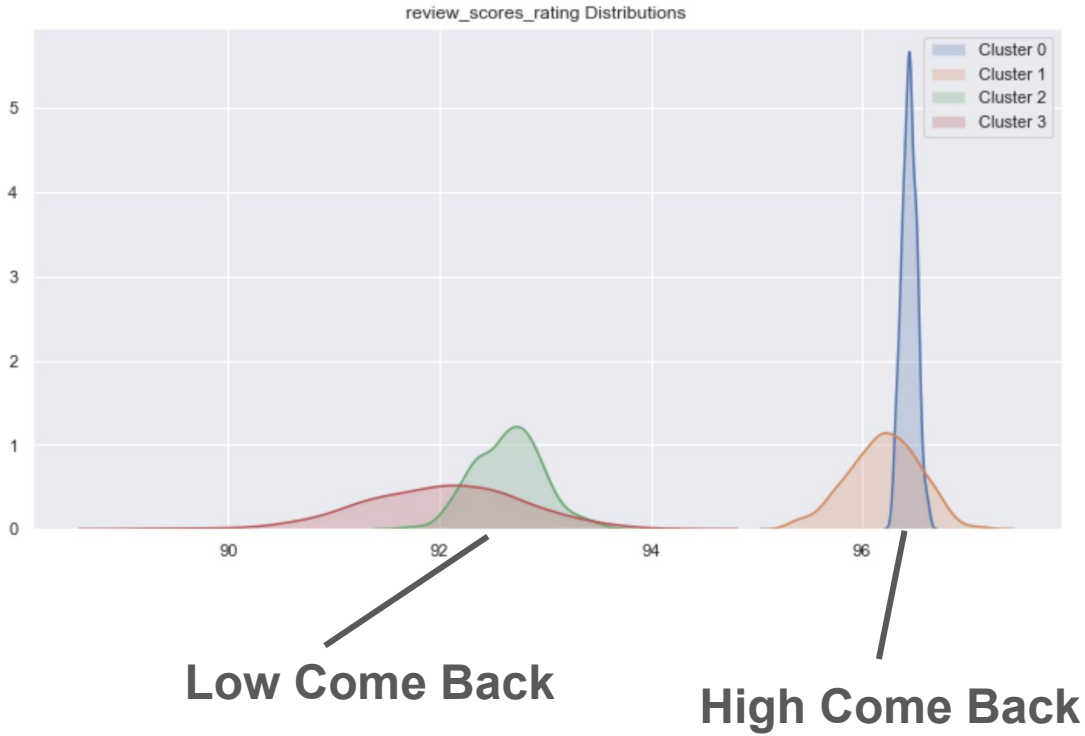
Cluster 2:

- **High** Cleanliness
- **Low** Come Again

Ratings by Clusters

Cluster 0 and 1 have ***higher ratings*** than **Cluster 2 and 3**.

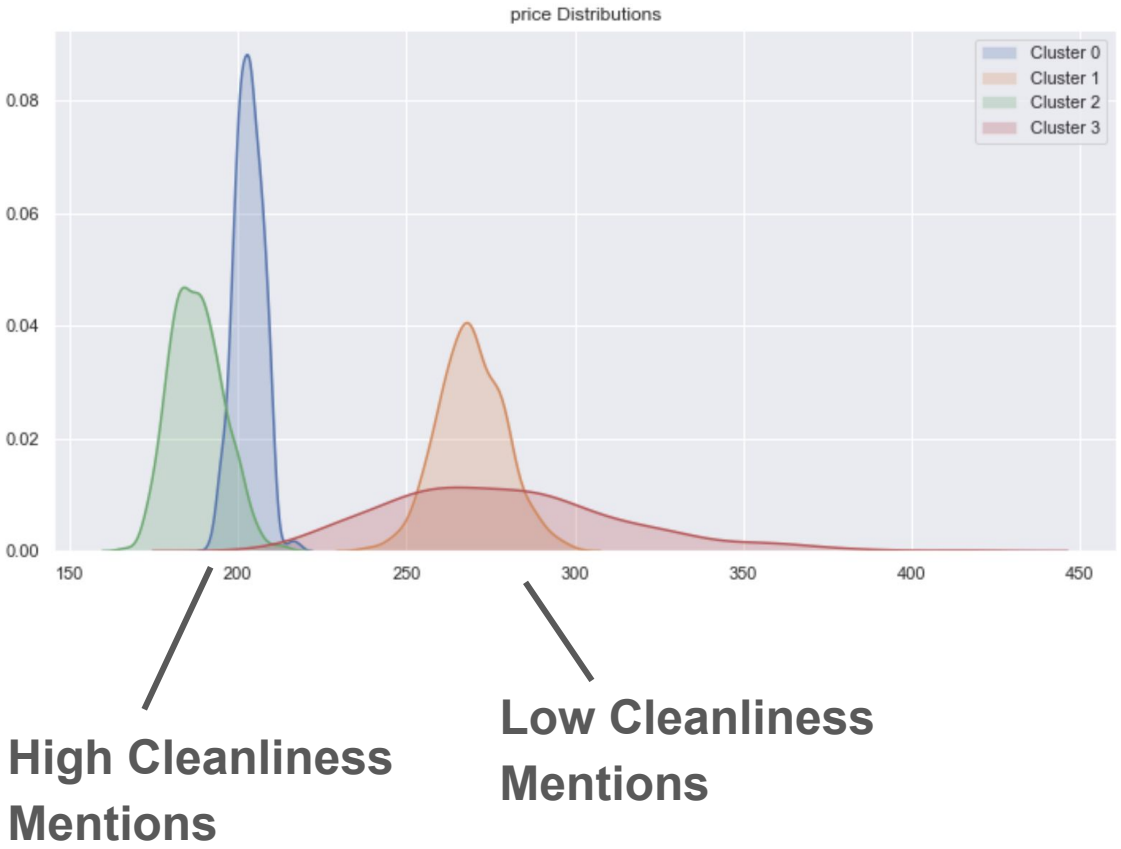
More "Come Back" Reviews, Higher Ratings.



Price by Clusters

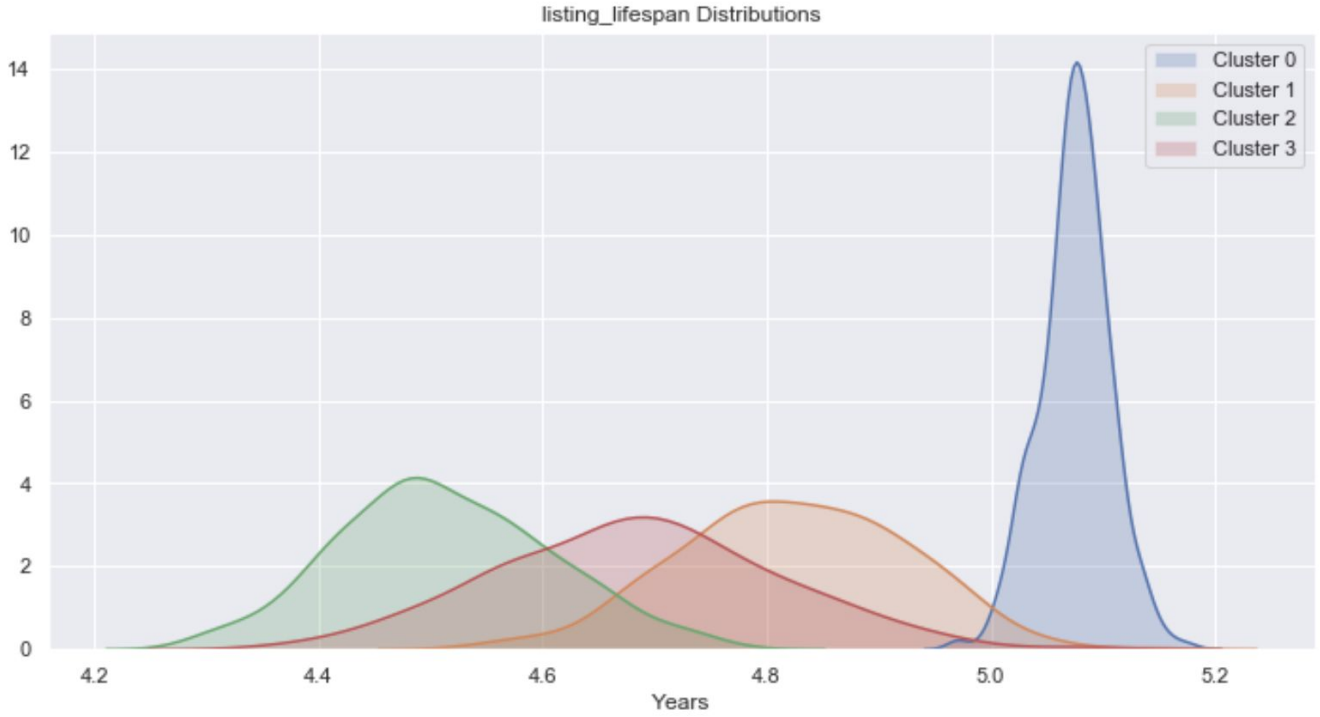
Reviewers tend to remember **Cleanliness More** for **Lower Price Listings**

***Do Higher Prices expect High Cleanliness (Less Mentions)?**



Listing Lifespan by Clusters

Lifespan of Listings are Fairly Similar



Reviews by Clusters

Number of reviews varies with no relation to listing lifespan.

Even with lower reviews, **Cluster 1 (Orange)** was able to offer an experience where people wanted to **Come Back** and give a **high Rating.**

Outcomes & Tips

- There are **Four Main Clusters** of listing reviews.
- **Tip For Guest:** Listings with “**Come Back**” reviews are likely **Higher Rating**.
- **Tip For Hosts:** Guests notice “**Cleanliness**” More for **Lower Priced** Listings.

How to Improve the Project

- **More Cities** and **Reviews**. One of the limiting factors for this was computing power and time as even with only 300,000 reviews LDA took around 9 hours.
- **Test More Variations of Number of Topics, Passes, Words,** and **Tokens.**
- **More Time and Strategy** in **Topics Selection.**
- **More Specific Case Studies**: Review Topics of **“Bad”** Listing.
- **Deeper** investigation on **Listing Descriptions and Information.**

References and Resources

GitHub Repository

https://github.com/daikiminaki/Capstone_2_Airbnb_Review_Topic_Modeling

Detailed Report

Email

dminaki95@gmail.com