Airbnb Reviews Topic Modeling and Analysis

Latent Dirichlet Allocation & KMeans

by

Daiki Minaki

Summary

Airbnb listings have 6 types of star ratings (Accuracy, Cleanliness, Checkin, Communication, Location, Value). In this project I will use Latent Dirichlet Allocation to categorize reviews so users can find reviews specific to the rating they want more information about.

Why it is worth solving.

In San Francisco alone there are over 300,000 reviews on Airbnb. Reviews are typically about areas of the stay that were most memorable for guests (good or bad) and provide a lot of insight on what a listing actually has to offer. By understanding what guests are saying about different listings and finding trends in the way people review we can better understand the relationship between guest and host.

The goal is to use Latent Dirichlet Allocation and K-Means Clustering to first find out what reviews are talking about, group them and then understand the differences in the way that reviewers review listings.

Approach

1. Data Cleaning/Wrangling (Tokenization, POS, NER)
2. Exploratory Data Analysis
3. Topic Modeling
4. Topic Analysis
5. Clustering by Topics
6. Outcome & Conclusion

The Data

Source: InsideAirbnb.com
The data consists of reviews, calendar, and listing data from different cities around the world. For the sake of this project I will only be focusing on AirBnB listings and reviews in San Francisco but have set up the project so it can be reproduced using any city that someone may choose to analyze and categorize.

Data Wrangling

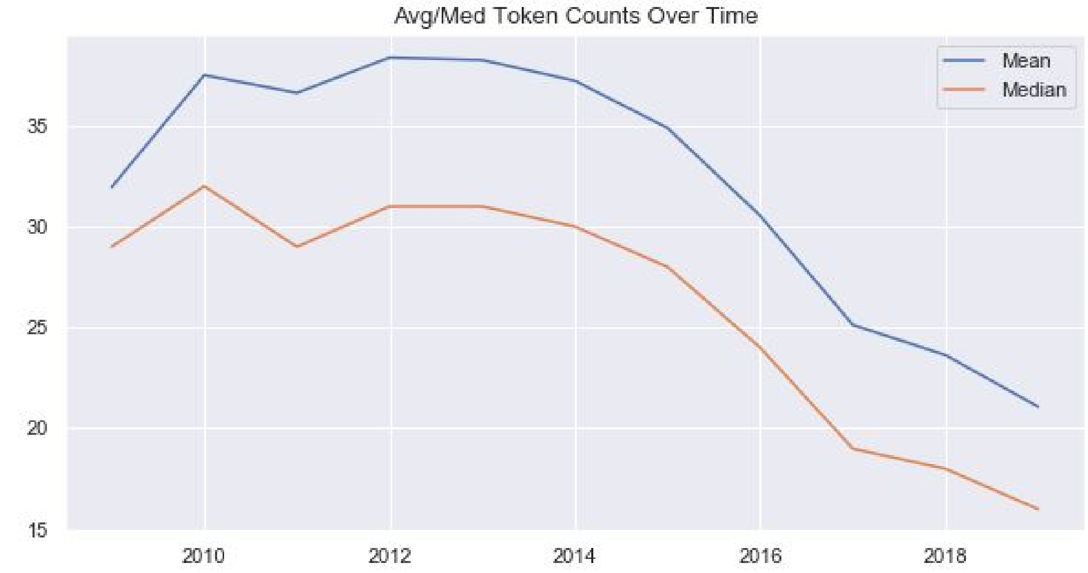
The data wrangling portion focuses on the reviews.csv as we are first going to focus on topic modeling and categorizing reviews. After each review has been categorized and labeled it may then be required to do some more wrangling with listing or calendar data to prepare for more in depth application of the results of the topic modeling.

To prepare the data for topic modeling I first cleaned, lemmatized and tokenized the raw review data. To test different parts of speech and variations I also filtered out name entities to create a column with non-name entity tokens as well as POS extraction to get the verbs, nouns, and adjectives for each review respectively. I then also did a count on each token set to get an idea of how long each review was.

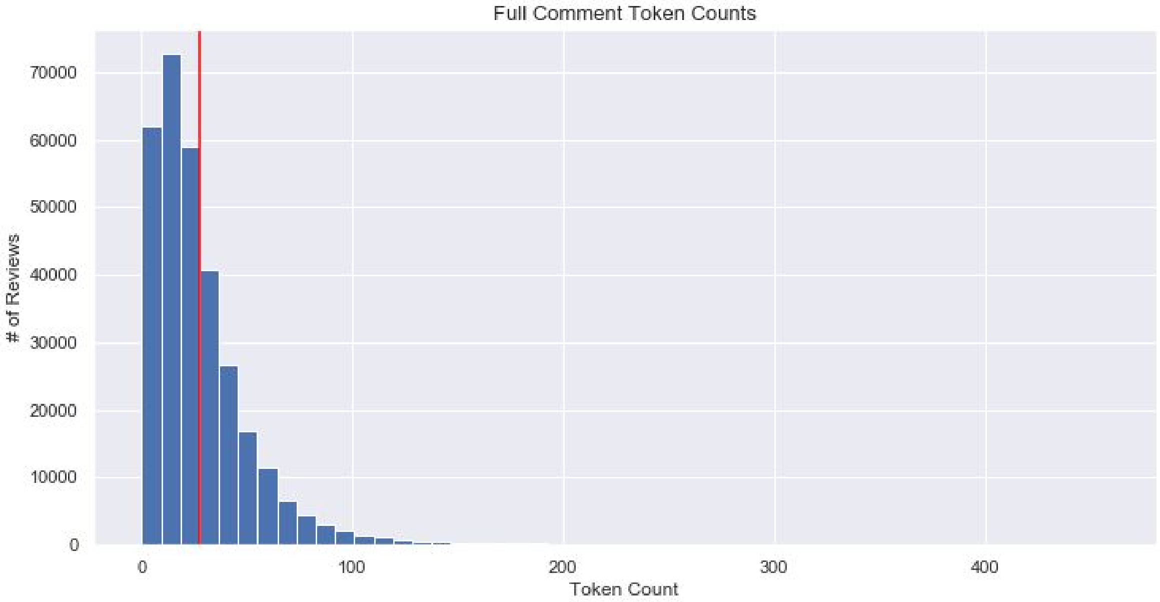
Exploratory Data Analysis (EDA)

Initial EDA for this project is very minimal mainly consisting of token count distributions. There will be more analysis on the data after categorizing the review data. The EDA mainly consisted of trying to understand how long reviews were and if reviews had in some way changed over time.

Since 2012, the number of reviews per year is declining.



Most reviews are very short with an average of about 30 words per review overall.



Topic Modelling

For Topic Modeling I used Latent Dirichlet Allocation to group topics found in the review data. I tried to choose topics based on the six topics that AirBnB has chosen to rate each listing: Accuracy, Cleanliness, Communication, Checkin, Location, and Value but I also added Transport as this seemed to be common topics found in many of the models that I tested.

I tested the different inputs for the model to find the best categorizations for the model using tokens as follows: Full Review, Non-Name Entities, Adjectives, Verbs, Nouns, Non-Name Entities Plus Adjectives. For the most part I used 50 Topics and 10 Words with 50 passes with the exception of adjectives which I used 5 words as the counts overall were very low.

In the end I chose the **Full Review Tokens input** model as it had the best topic breakdowns overall, however, some models did seem to perform better for specific topics.

Final Topics:

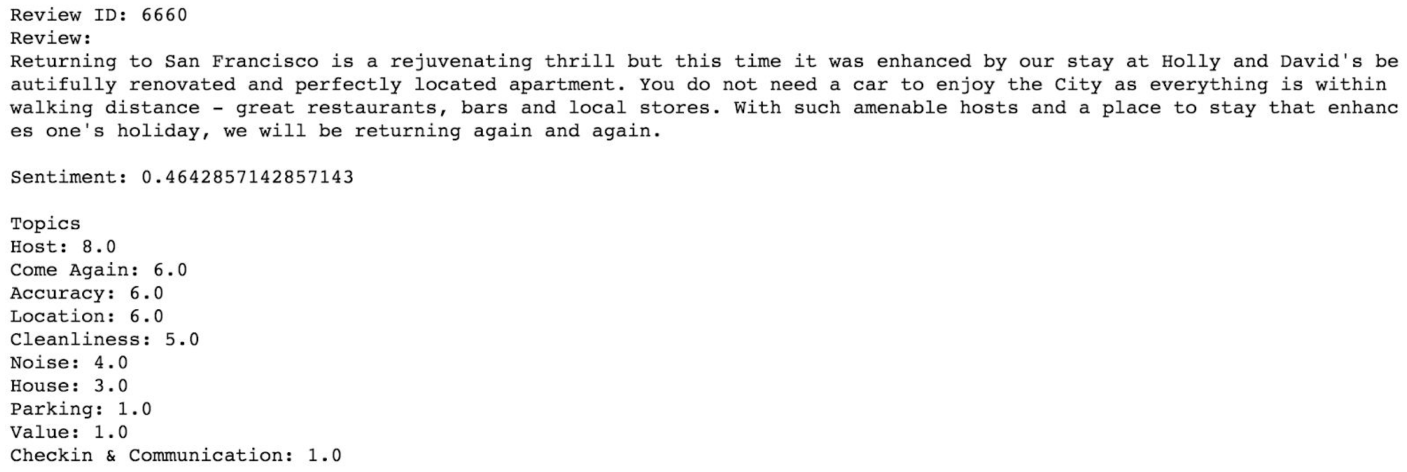
```
Checkin & Communication
17: check, late, even, early, arrived, time, accommodating, let, flight, last
23: easy, check, communication, access, check-in, clean, super, communicate, made, instruction
27: quick, question, always, quickly, respond, response, available, responded, message, john

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
```

Tagging Reviews with Selected Topics

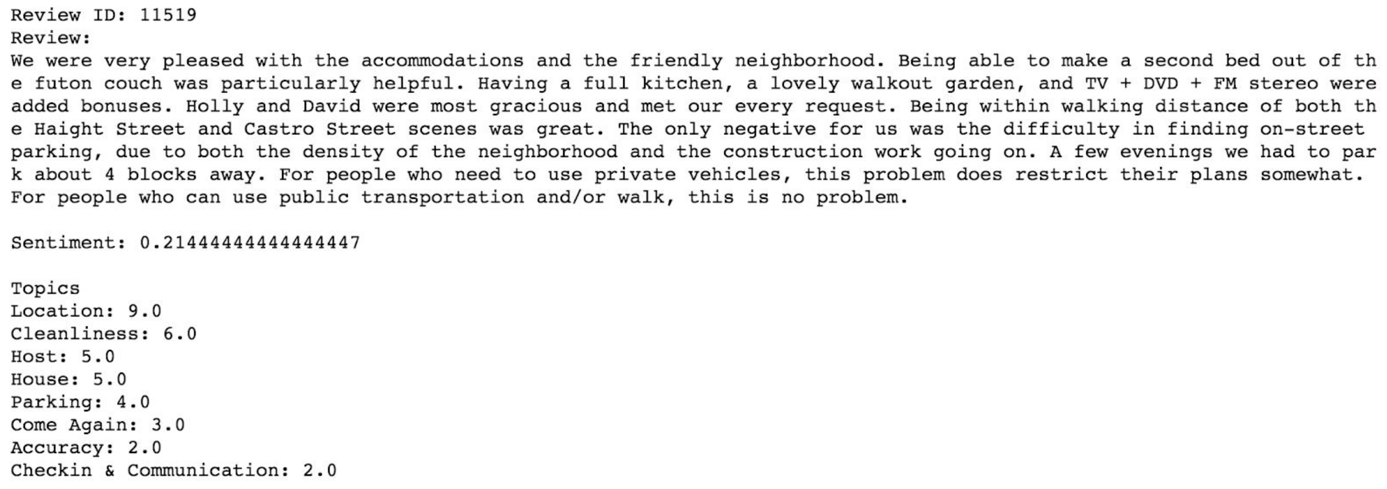
The next step in the process is to actually tag each review by the topics that have been selected. It may be the case that I will have to tune some of the topics based on how they actually tag the reviews.

The two reviews below show some examples of reviews as tags given to each. Location, Host, Cleanliness, and Come Again were the most talked about topics overall.

Tagged Reviews:



```
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.

Sentiment: 0.4642857142857143

Topics
Host: 8.0
Come Again: 6.0
Accuracy: 6.0
Location: 6.0
Cleanliness: 5.0
Noise: 4.0
House: 3.0
Parking: 1.0
Value: 1.0
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.

Sentiment: 0.21444444444444447

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

Topic Analysis and Clustering

“Are there any trends in the way listings are reviewed?”

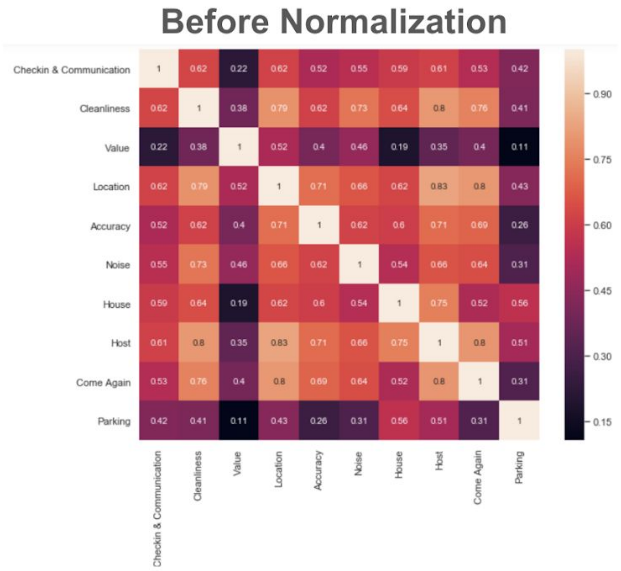
Clustering Listings by Review Content

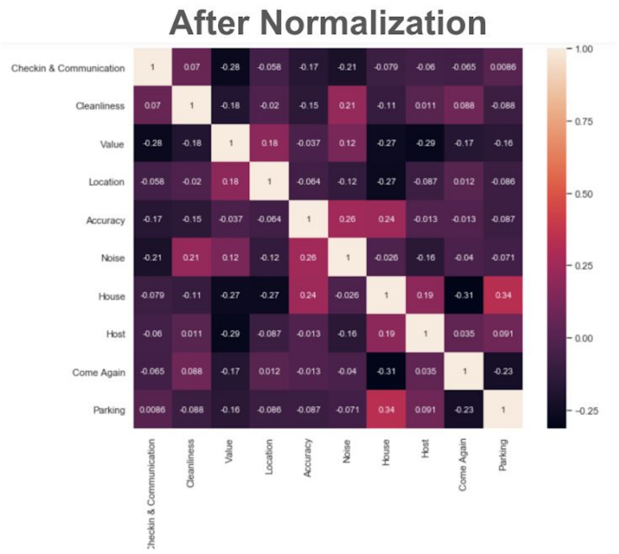
The topics assigned during the LDA is only really as good as the information and new insights we are able to get out of the availability of the new data. For this project I was interested in understanding the possible trends and patterns in how people are reviewing listings. To do this I decided to see if there were any groupings of review topics by listing types.

Normalizing Review Topic Weights

To give each topic the same range of values I decided to normalize the weights for each topic which giving each topic a value between 0 and 1. This helps us compare between topics more effectively and allows us to understand if a specific review is more important or if the topic is just talked about more often in general.

Normalization seems to have an adverse effect on correlation between topics as shown below with some level of correlation between topics like Host and Come Again in the heatmap before normalization and much less correlation with normalization.

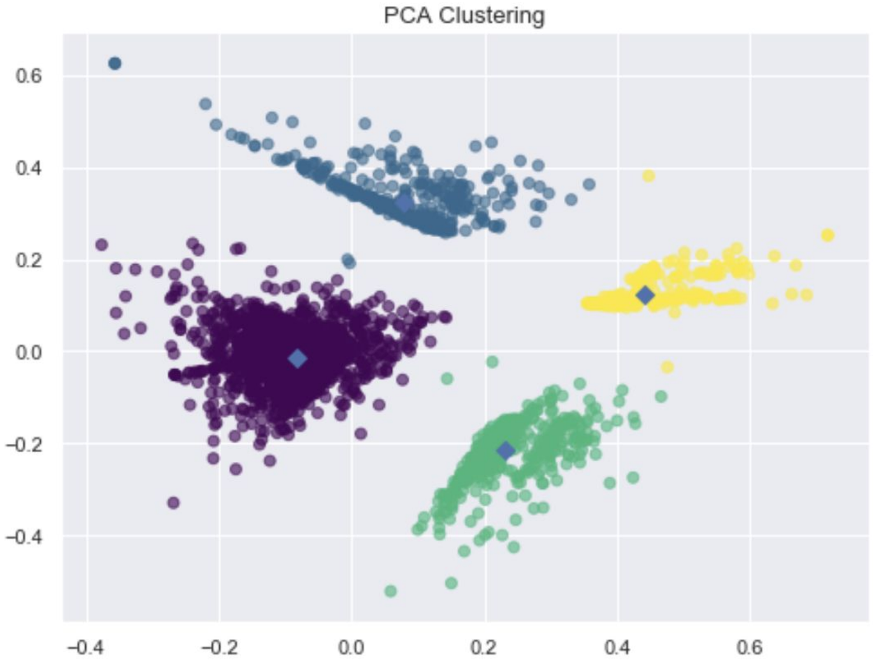




PCA & K-Means

With the normalized review topics I was interested in finding clusters in the topic behavior of different listings. Using K-Means I was able to end up with four distinct clusterings of listings. I decided to do PCA to both reduce dimensionality and to make it possible to visualize the clusterings as shown below.

To find the clusters I first did PCA on the topic data and used the top 3 features in a K-Means model to cluster the data. Choosing the number of clusters was fairly simple as the PCA visualization showed 4 clear clusterings but I also decided to use Knee/Elbow Method to find the best n_clusters.

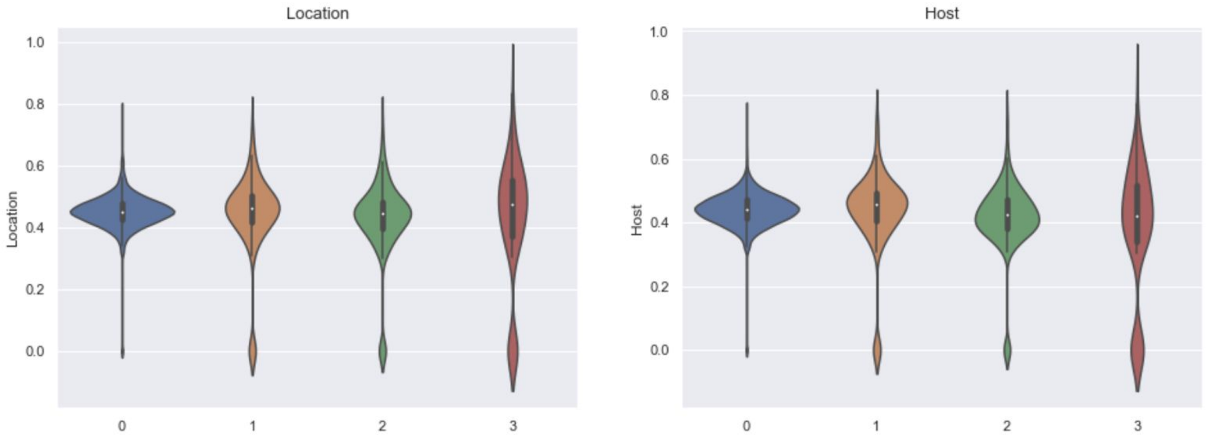


Cluster Analysis

With four clear clusterings it was now time to understand what these clusters were made of and what the key differences were in the four groups of listings.

Most Review Topics were Similar Across Clusters.

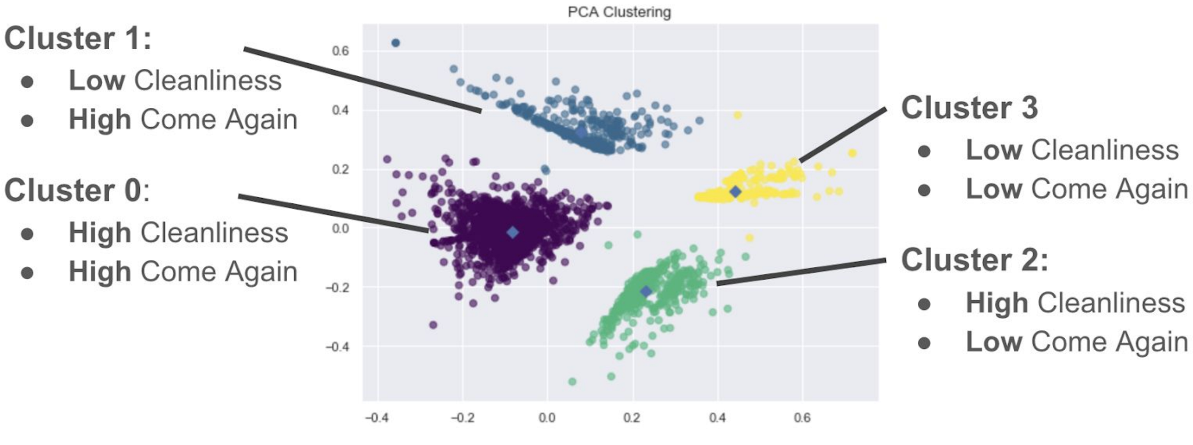
Out of the 10 topics that were assigned during the LDA, 8 of them had similar distributions and ranges. This is not ideal but it does allow us to focus on the two topics which we now know are responsible for splitting the data into four clusters.



Some of the similar topics include: Ex. Location, Host, Value, Accuracy, House, etc

Key Differences between Cluster

The two topics that impacted the clusterings were “Cleanliness” and “Come Again”.



Cluster 1:

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

Cluster 0:

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

Cluster 3

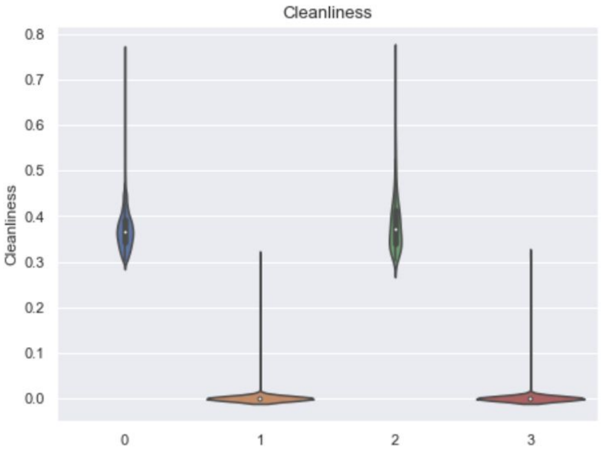
- **Low** Cleanliness
- **Low** Come Again

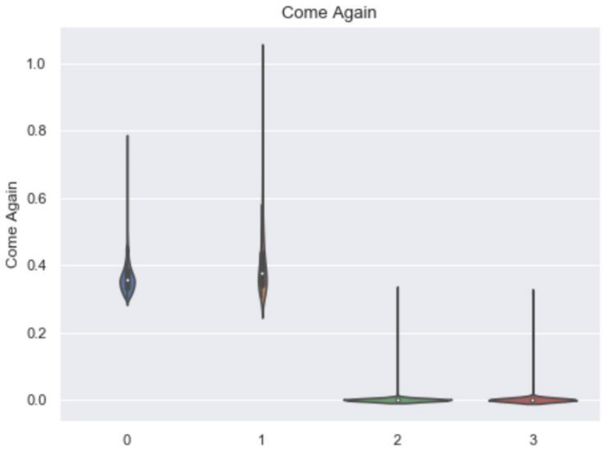
Cluster 2:

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

“Cleanliness” and “Come Again”

Visualized by topics weights by cluster it is clear that “Cleanliness” is talked about at a high rate in cluster 0 and 2 but is not talked about as much in clusters 1 and 3. “Come Again” on the other hand is high for clusters 0 and 1 but low for 2 and 3. The different levels of “Cleanliness” and “Come Again” end up being the difference in the four K-Means Clusters.





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

Analyzing Clusters by Listings

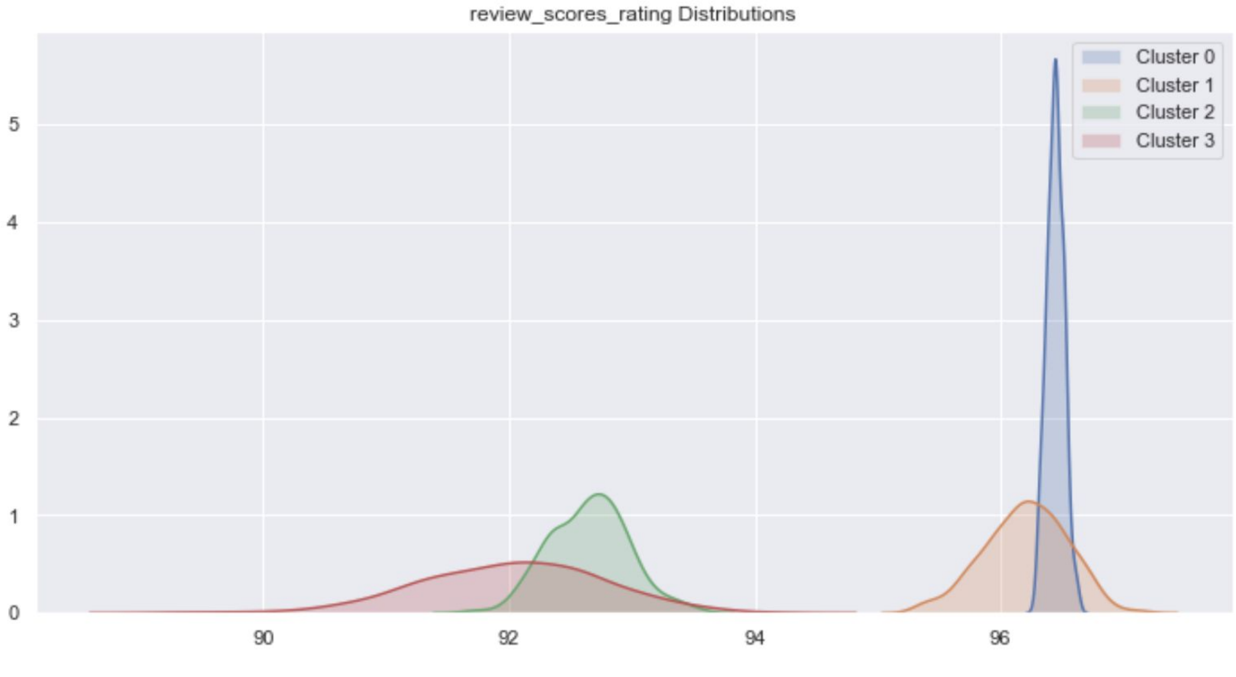
Now comes the question why the clusters are the way they are. What makes people talk more about “Come Again” when going to listings in clusters 0 and 1? What makes people talk about “Cleanliness” more when reviewing listings in clusters 0 and 2?

By understanding the behavior of reviews for certain listings we can likely come away with insights that will allow us to better understand what expectations different listings might have on guests and what reviewers end up remembering about listings.

Listing Rating

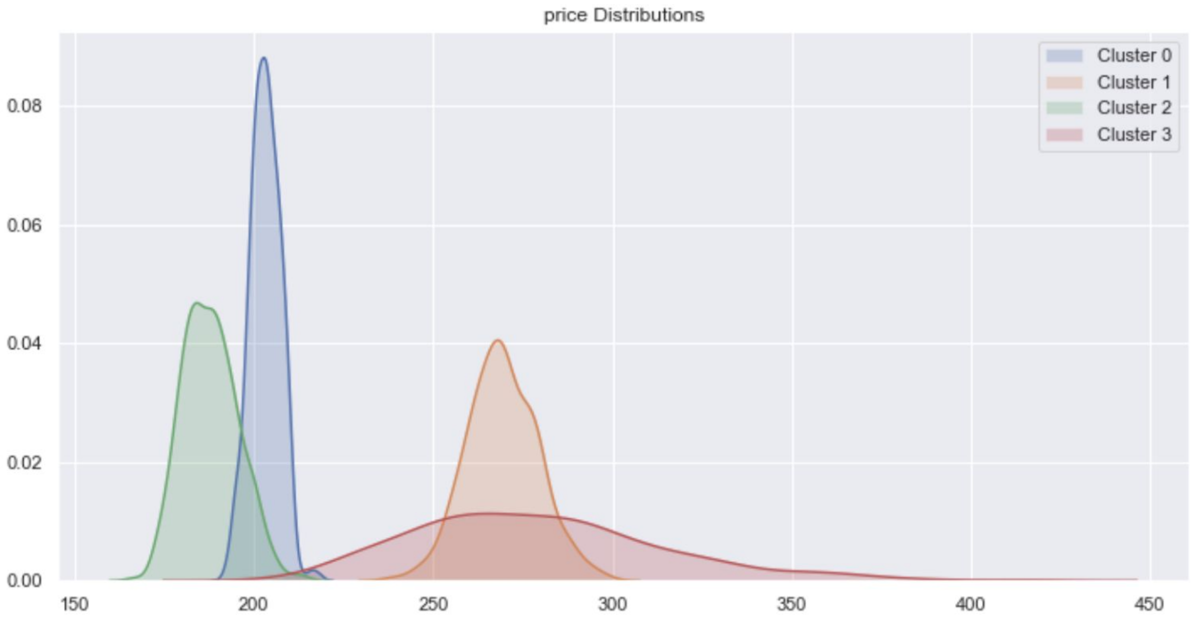
Looking at the distribution of listing ratings of the four clusters we see that Clusters 0 and 1 have a statistically significantly higher Review Rating Score than Clusters 2 and 3. Clusters 0 and 1 have different “Cleanliness” topic weights but both have high “Come Back” Reviews.

Cluster 0 and 1 have ***higher ratings*** than **Cluster 2 and 3**.
More “Come Back” Reviews, **Higher Ratings**.



Price Distribution

When looking at price we see that Clusters 1 and 3 tend to be more expensive and Clusters 0 and 2 less so. Reviewers tend to remember **Cleanliness More** for **Lower Price Listings.**



Conclusion

Using Latent Dirichlet Allocation (LDA) I was able to successfully pull different topics from review data and use these topics to gain insight on how guests review listings on Airbnb. To get a sense of some overarching trends in the type of reviews that were left for different listings I performed K-Means clustering which clustered the listings into four main groups.

The key differences in the review behavior for the four clusters mainly dealt with whether "Cleanliness" and "Come Again" were mentioned in the reviews of the listings or not. Using this I also did some analysis to see what the reason could be for the difference and whether there were any interesting relationships between listing features and the mention of the different topics. In the end it seems like there is a clear relationship between reviewers leaving "Come Back" reviews and a listing with high ratings and listings with lower prices with more mentions of cleanliness.

Reflection and Improvements

If there was more time and if I was to do this project over I would focus on a few main parts:

1. More data. The complete Airbnb dataset consists of cities from around the world with even more reviews, listings, and guests. For the sake of this project I decided to stick to San Francisco as there were already over 300,000 reviews. Even with only 300,000 reviews the LDA was time consuming (about 9 hours) when done on a laptop making it difficult to iterate on LDA models.
2. Better preprocessing. One important lesson I learned while using LDA was how much the input can change the topics that are identified in the output. Given another opportunity to do an LDA project I would likely spend more time preprocessing the data possibly using more cities with some sort of filtering to lessen the actual number of reviews.
3. Depth. Given more time it would also be possible to look more deeply into actual description and listing details when looking at listings and their corresponding reviews. Depth would also mean looking at more specific cases such as negative reviews and seeing what topics dominate in specific reviews and listings. This also relates to the first point of needing to use more data and more time.

References and Resources

GitHub Repository
https://github.com/daikiminaki/Capstone_2_Airbnb_Review_Topic_Modeling

GitHub Presentation
https://github.com/daikiminaki/Capstone_2_Airbnb_Review_Topic_Modeling/blob/master/Capstone_2_Airbnb_Review_Topic_Modeling_Presentation.pdf

Email Address
dminaki95@gmail.com