



Uncovering Main Topics in AIRBNB Reviews

LDA and LSA Methods

Group 1

Michael Arons

Malgorzata Klesna-Blat

Jiesen Ou

Wendy Zhai

AGENDA



Introduction and EDA



Topic Modeling with SASEM and Python



Delivering Insights



Limitations/Future Use

Introduction and Exploratory Data Analysis



What Is the Plan?



Corpus of Documents

1

AirBNB Review
Richard was a fantastic host! He made sure that we had everything we needed and even sent recommendations on exploring Nashville. The studio itself was clean, modern, and comfortable. LOVED IT! Perfect for a weekend trip and no more than a 10 minute
Steve was very helpful as a host, he was on-site to assist with showing the unit. He also provided Netflix credentials for my viewing pleasure. Very clean apartment with extremely modern amenities.
Great location, cute place. We had an awesome stay. Thank you!
My stay here was great. Very convenient to the 16th and Mission BART station and just a one-minute walk to a good bar/restaurant (The Monk's Kettle). I generally felt safe here and it was super quiet at night. Would definitely stay here again!
Steve was a great host— prompt communication and very accommodating when we needed a place to leave our bags for a few hours after checkout.
The apartment was nearby great bars and restaurants while being calm and quiet at night. Highly recommended.
My parents were the actual guests during this stay. From their perspective, the place was clean and located walking distance from BART which offers a lot of convenience. It was comfortable despite a basement feel with low ceilings in doorways. My father
Great host and great space. Clean, new, and quiet. Great neighborhood too. Will definitely seek it again when returning to sf.
小屋很温馨,地理位置特别棒!房东主人非常和善。推荐这间小屋!



Create Topics from Reviews Using LSA and LDA

2

S Topic id	S Term
topic_1	stay, host, recommend, clean, nice, definitely...
topic_2	walk, location, close, apartment, restaurant, ...
topic_3	host, stay, house, welcome, time, arrival, ho...
topic_4	stay, clean, comfortable, apartment, location,...
topic_5	apartment, clean, night, stay, little, nice, park...



S Topic id	S Term	S Topic
topic_1	stay, host, recommend, clean, nice, def...	Service
topic_2	walk, location, close, apartment, restau...	Location
topic_3	host, stay, house, welcome, time, arriv...	Meeting the host
topic_4	stay, clean, comfortable, apartment, lo...	Amenities



Insights, Answering key Questions

3

After figuring what the topics are based on the terms we can answer questions such as what do reviewers care about? If you are someone with a listing what should you put in your description to attract customers? And other questions. We can also compare our two different outputs from LSA and LDA



Latent Dirichlet Allocation (LDA)

Unsupervised learning technique

Input: Corpus of documents transformed into **Bag of Words** in a matrix format

Topics are created by generative probabilistic model

Output: A group of topics with keywords and weights assigned to each document and how aligned the document is to each topic

Latent Semantic Analysis (LSA)

Unsupervised learning technique

Input: Corpus of documents transformed into **Bag of Words** in a matrix format

Topics are created by dimension reduction using SVD

Output: A group of topics with keywords and weights assigned to each document and how aligned the document is to each topic



airbnb


AirBNB Review Data

Number of Reviews 

1.5M

Date Range 

Jan 2016 - June 2017

Country 

US

Key Locations 

NYC 20 % Data

LA 14% Data

Reviews Data

listing_id	id	date	reviewer_id	reviewer_name	comments
488835	104522929	2016-09-27	66003975	Carole	First time using Airbnb and couldn't be...
13546118	112010466	2016-11-03	27252960	Alexis	L'appartement est trÃ's sympa.
13546118	137928884	2017-03-18	111268477	Donovan	Place is pretty nice and Sergio was a bi...
549036	78711836	2016-06-08	25834	Armen	We enjoyed our trip to Barcelona and p...
2643611	84818807	2016-07-09	72772795	Rebecca	Patrica was the perfect host. She meet...
549036	82818700	2016-06-29	3838994	Jouni	Very comfortable room in a clean apart...
2643611	95293023	2016-08-19	83384637	Bryce	Such a lovely apartment and wonderful...
2643611	111071972	2016-10-30	26017119	Sourma	Great flat in a lovely location.

Listing Data

ID	Listing Url	City	State	Zipcode	Beds	Price
11124183	https://www.airbnb.com/rooms/111241...	Austin	TX	78746	5	[Null]
15359208	https://www.airbnb.com/rooms/153592...	West Lake Hills	TX	78746	4	[Null]
15450282	https://www.airbnb.com/rooms/154502...	Austin	TX	78723	1	63
15967775	https://www.airbnb.com/rooms/159677...	Austin	TX	78746	7	[Null]
17479642	https://www.airbnb.com/rooms/174796...	Austin	TX	78723	1	350
17513218	https://www.airbnb.com/rooms/175132...	Austin	TX	78759	1	45

Join by Specific Fields

Left	Right
1 Reviews_listing_id	Listing_ID
*	



Reviews_comments	Listing_City	Listing_State	Listing_Zipcode	Listing_Bedrooms	Listing_Price
This is a nice house located perfectly in...	San Diego	Ca	92037	3	395
The apartment is very nice! I reached it...	New York	Ny	10033	1	49
Gregorio is a gracious host who welco...	New York	Ny	10033	1	49
The room is so nice and clean, the coz...	New York	Ny	10033	1	49
It's a great room.	Brooklyn	Ny	11221	1	60
I had a great stay at Daniel, all what yo...	Brooklyn	Ny	11221	1	60



Where Are AIRBNB Reviews Coming From?

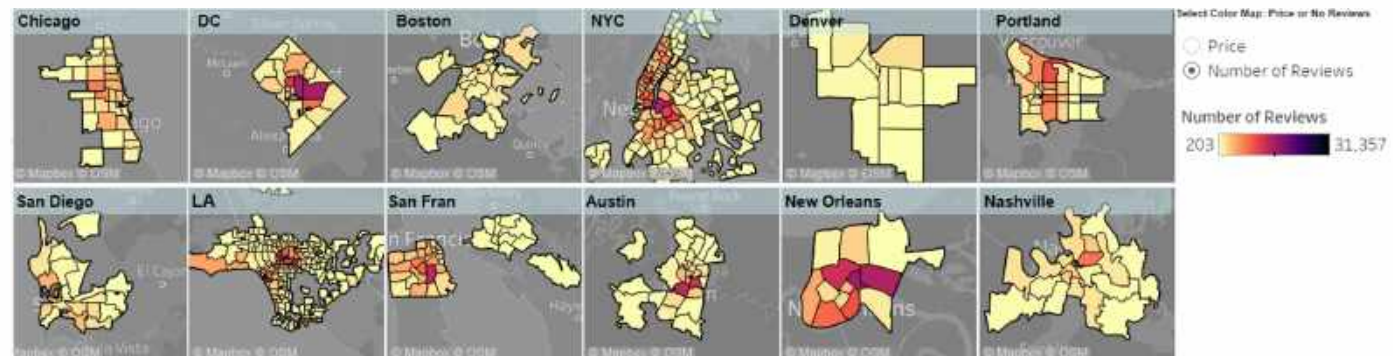
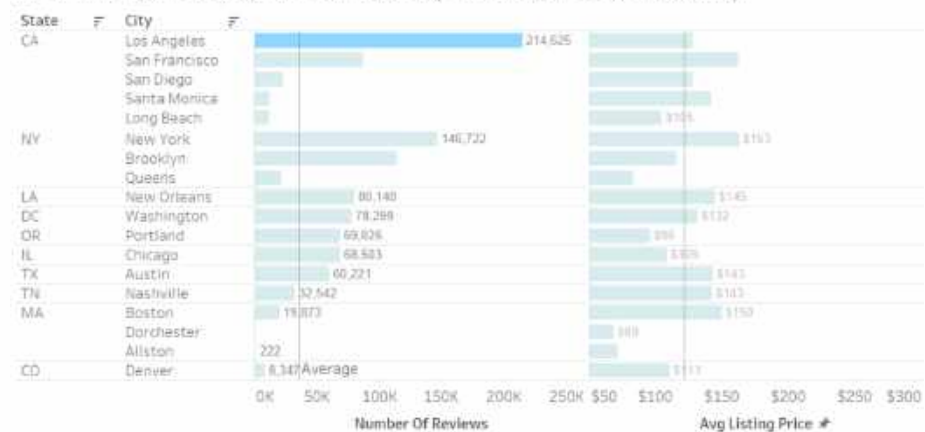
NYC accounts for over **20%** of total reviews in our dataset.

Los Angeles and New York City account for over **34%** of total reviews in our dataset

We notice that for each of the main areas as we get closer to the city center the number of reviews increase greatly.

Also, we notice that the number of reviews is **very concentrated in a few key zip codes** (especially LA and NYC)

Price and Reviews, By State and City (Filtered on > 200 Reviews)

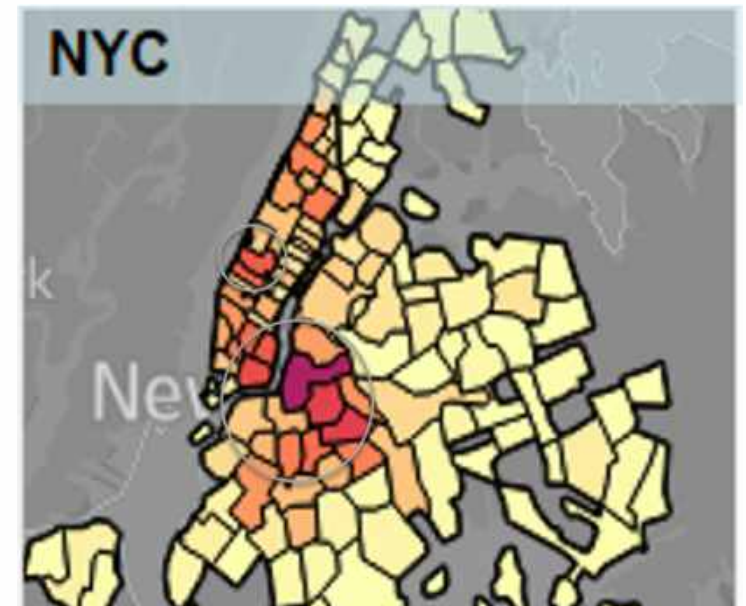
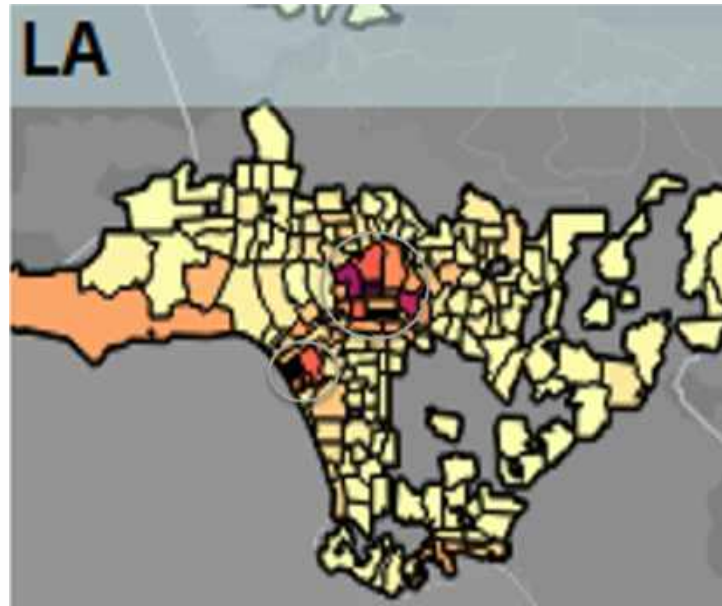


Number of Reviews zoomed in for NYC and LA provides a closer look at reviews in the top two cities in the dataset.

The darker the color of the area, the greater the number of reviews for that zip code in the dataset.

LA has 2 very dark spots indicating a very high number of reviews in these neighborhoods
NYC has several neighborhoods with a high number of reviews

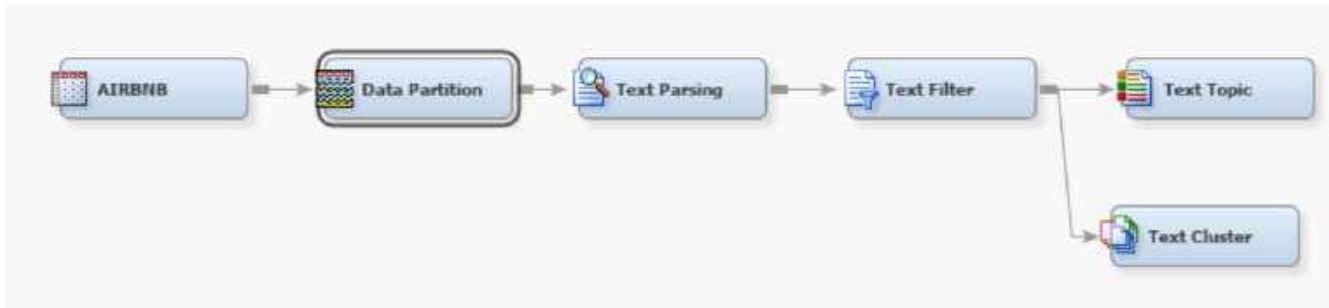
Number of Reviews
203 31,357



Topic Modeling with SASEM and Python



Topic Modeling with LSA



Data: review_comments;
127,158 rows

The challenge with SVD:
hard to determine the
optimal number of
dimensions.

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
User	9	0.069	0.600+apartment	Apartment	461	27983
User	11	0.080	0.500+great location,+location	Location	223	20680
User	1	0.066	0.300+great host,+host	Host	217	20525
User	3	0.082	0.300+automate,+cancel,+post,+reserva...	Service	11	2511
User	4	0.072	0.300+home,+house	Home/house	554	28484
User	7	0.064	0.030+room,+nice,+bed,+bathroom,+ho...	Room	693	26448
User	6	0.066	0.020+distance,walking,+walking distan...	Distance	372	15064
Multiple	5	0.060	0.007 f,â de, tr, est	Foreign language1	420	6228
Multiple	8	0.057	0.007+experience,first,airbnb,+good,+time	Airbnb experience	363	19414
Multiple	10	0.062	0.007+place,+stay,+great place,+definite...	Place/ stay	380	26780
Multiple	12	0.078	0.007+nice,+recommend,highly,+clean...	Nice/recommend	335	30789
Multiple	2	0.111	0.006 â, â, * âââ, âœâ	Foreign language 2	27	6755



Topic Modeling with LDA

Data Pre-Processing

- Tokenization
- Words that have fewer than 3 characters are removed
- All stopwords are removed
- Words are lemmatized and stemmed

Loading **gensim** and **nltk** libraries

```
stemmer = PorterStemmer()
def lemmatize_stemming(text):
    return stemmer.stem(WordNetLemmatizer().lemmatize(text, pos='v'))
def preprocess(text):
    result = []
    for token in gensim.utils.simple_preprocess(text):
        if token not in gensim.parsing.preprocessing.STOPWORDS and len(token) > 3:
            result.append(lemmatize_stemming(token))
    return result
```

Topic Modeling with LDA

Bag of Words on the Data set

- Pre-process the comment text, saving the results as 'processed_docs'
- Create a dictionary from 'processed_docs' containing normalized words and their integer ids

```
processed_docs = text.map(preprocess)
processed_docs[:10]
```

```
0 [locat, state, near, vanderbilt, walk, distanc...
1 [gregorio, place, great, drive, ohio, fantast,...
2 [gregorio, nice, friendli, hospit, nice, talk,...
3 [greg, welcom, kind, quiet, stay, felt, cozi, ...
4 [love]
5 [absolut, recommend, daniel, home, impecc, cle...
6 [daniel, great, host, welcom, sociabl, apart, ...
7 [great, valu, kind, host, clean]
8 [love, bedroom, condo, need, famili, kid, rang...
9 [great, stay, apart, perfect, children, day, a...
```

```
dictionary = gensim.corpora.Dictionary(processed_docs)
count = 0
for k, v in dictionary.iteritems():
    print(k, v)
    count += 1
    if count > 10:
        break
```

```
0 abl
1 away
2 beer
3 breakfast
4 cold
5 comfort
6 distanc
7 euro
8 excel
9 final
10 foot
```

Topic Modeling with LDA

Gensim filter_extremes and doc2bow

- Filter out tokens that appear in:
 - less than 15 documents or
 - more than 30% documents
 - keep only the first 100,000 most frequent tokens
- For each document, create a dictionary reporting how many words and how many times those words appear.
- Then check the document selected earlier.

```
dictionary.filter_extremes(no_below=15, no_above=0.3, keep_n=100000)
```

```
bow_corpus = [dictionary.doc2bow(doc) for doc in processed_docs]
bow_doc_4310 = bow_corpus[4310]
for i in range(len(bow_doc_4310)):
    print("Word {} (\"{}\") appears {} time.".format(bow_doc_4310[i][0],
                                                    dictionary[bow_doc_4310[i][0]],
                                                    bow_doc_4310[i][1]))
```

```
Word 88 ("love") appears 1 time.
Word 98 ("home") appears 1 time.
Word 102 ("provid") appears 1 time.
Word 117 ("need") appears 1 time.
Word 229 ("thing") appears 1 time.
Word 240 ("amaz") appears 2 time.
Word 295 ("coffe") appears 1 time.
Word 322 ("trip") appears 1 time.
Word 631 ("austin") appears 1 time.
```

Topic Modeling with LDA

Running LDA using Bag of Words

Train the LDA model using `gensim.models.LdaMulticore` and save it to 'lda_model'.
Let the number of topics equal 12.

```
lda_model = gensim.models.LdaMulticore(bow_corpus, num_topics=12, id2word=dictionary, passes=2, workers=2)
```

TopicID	Terms	Topic Label
0	0.022*sehr + 0.020*para + 0.019*casa + 0.015*todo + 0.011*bien + 0.010*metro + 0.009*como + 0.009*excelent + 0.009*all + 0.009*apartamento	Foreign language
1	0.053*restaur + 0.053*walk + 0.029*shop + 0.026*distanc + 0.024*close + 0.019*bar + 0.018*street + 0.017*easi + 0.017*neighborhood + 0.016*park	Food and shopping
2	0.019*room + 0.016*night + 0.015*arriv + 0.014*check + 0.012*time + 0.011*clean + 0.011*apart + 0.009*late + 0.009*need + 0.009*good	Check-in
3	0.033*park + 0.025*beach + 0.022*nice + 0.021*quiet + 0.020*clean + 0.019*comfort + 0.015*room + 0.015*neighborhood + 0.015*privat + 0.014*easi	Park and beach
4	0.093*apart + 0.026*subway + 0.020*clean + 0.018*walk + 0.015*close + 0.015*nice + 0.015*station + 0.013*easi + 0.013*need + 0.013*help	Location
5	0.022*home + 0.018*love + 0.018*perfect + 0.016*beauti + 0.014*walk + 0.013*time + 0.012*amaz + 0.010*better + 0.010*look + 0.009*citi	Great stay 1
6	0.019*kitchen + 0.016*room + 0.014*home + 0.013*comfort + 0.013*coffe + 0.012*like + 0.011*hous + 0.011*love + 0.010*bathroom + 0.010*need	Kitchen and room
7	0.035*question + 0.028*clean + 0.026*respond + 0.024*easi + 0.023*commun + 0.021*check + 0.019*quick + 0.019*help + 0.017*exactli + 0.017*describ	Communication
8	0.042*appart + 0.039*reserv + 0.035*nou + 0.031*post + 0.030*cancel + 0.028*arriv + 0.025*autom + 0.025*pour + 0.023*day + 0.021*bien	Reservation
9	0.072*hous + 0.022*love + 0.020*austin + 0.018*perfect + 0.017*need + 0.017*definit + 0.016*home + 0.015*time + 0.014*visit + 0.014*clean	Great stay 2
10	0.050*recommend + 0.026*highli + 0.024*experi + 0.023*home + 0.023*clean + 0.023*definit + 0.019*help + 0.019*welcom + 0.019*comfort + 0.016*accommod	Recommend
11	0.089*nice + 0.042*clean + 0.040*good + 0.035*room + 0.029*time + 0.028*thank + 0.019*help + 0.019*perfect + 0.016*comfort + 0.014*condo	Nice and clean



Topic Modeling with LDA

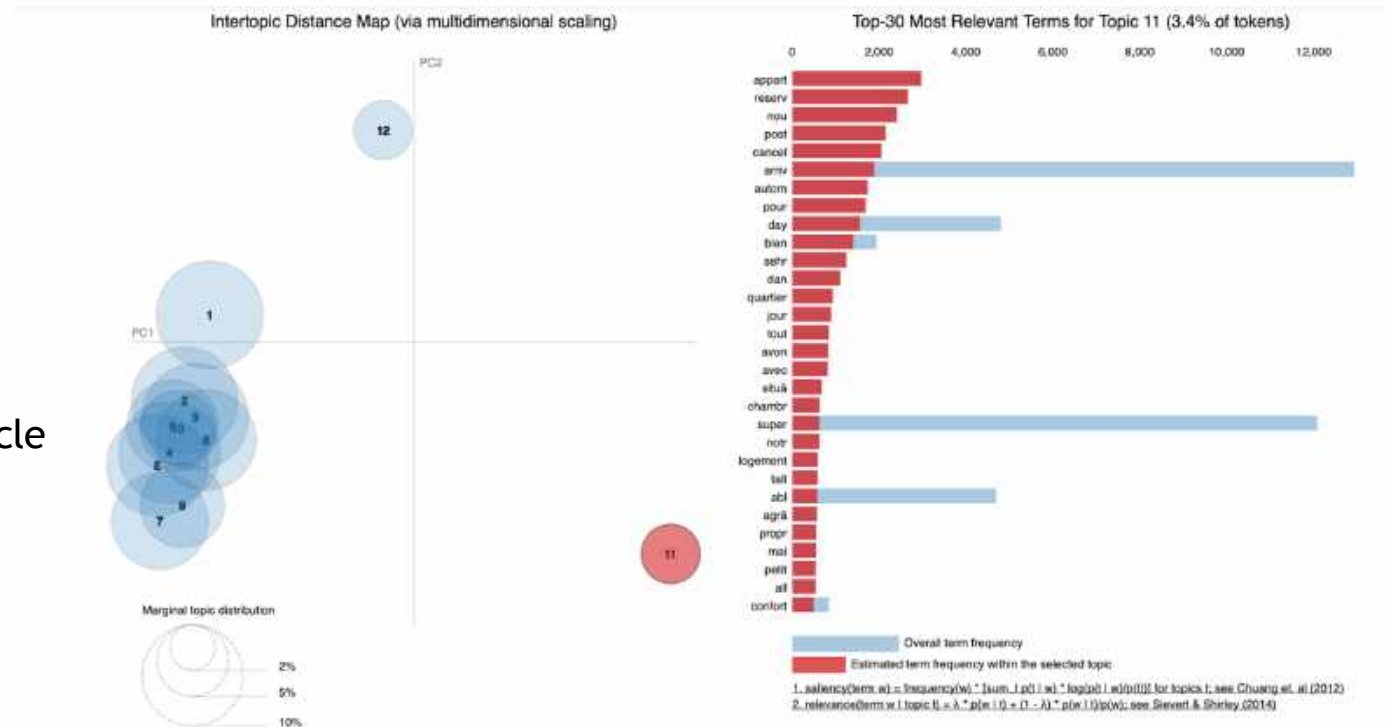
Visualization

pyLDavis allows for data visualization in an interactive format.

Left: a plot of the "distance" between all of the topics.

The relative size of a topic's circle corresponds to the relative frequency of the topic in the corpus.

Right: a bar chart showing top terms.



Takeaways and Insights

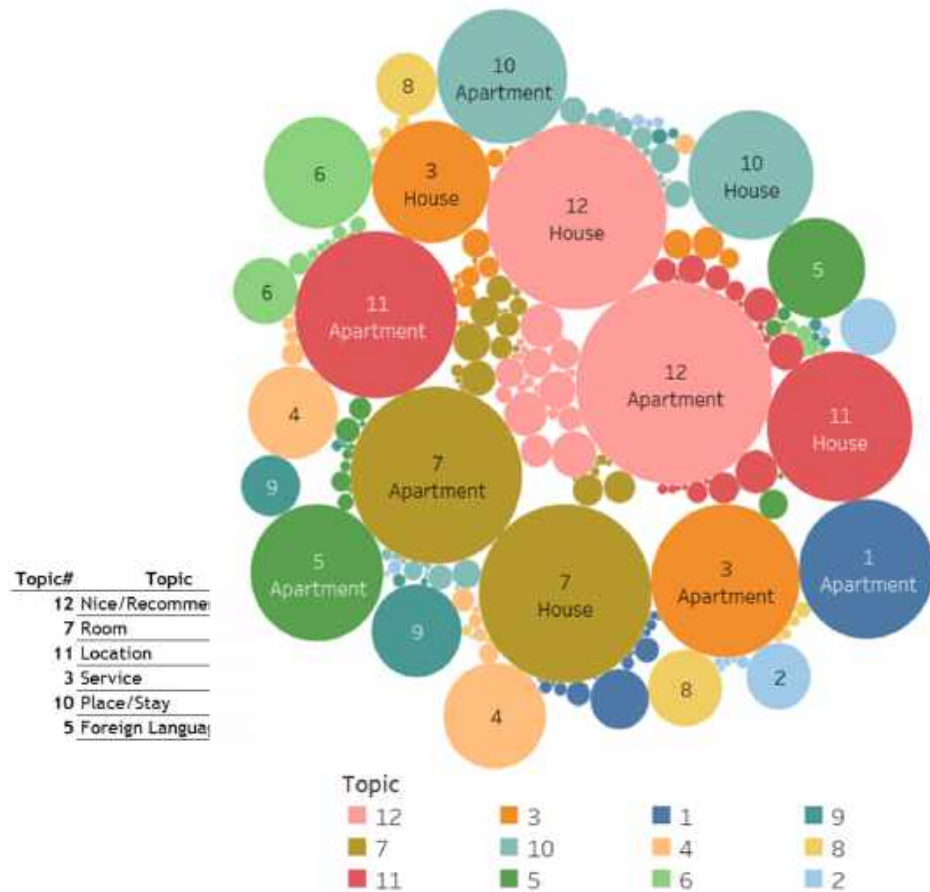




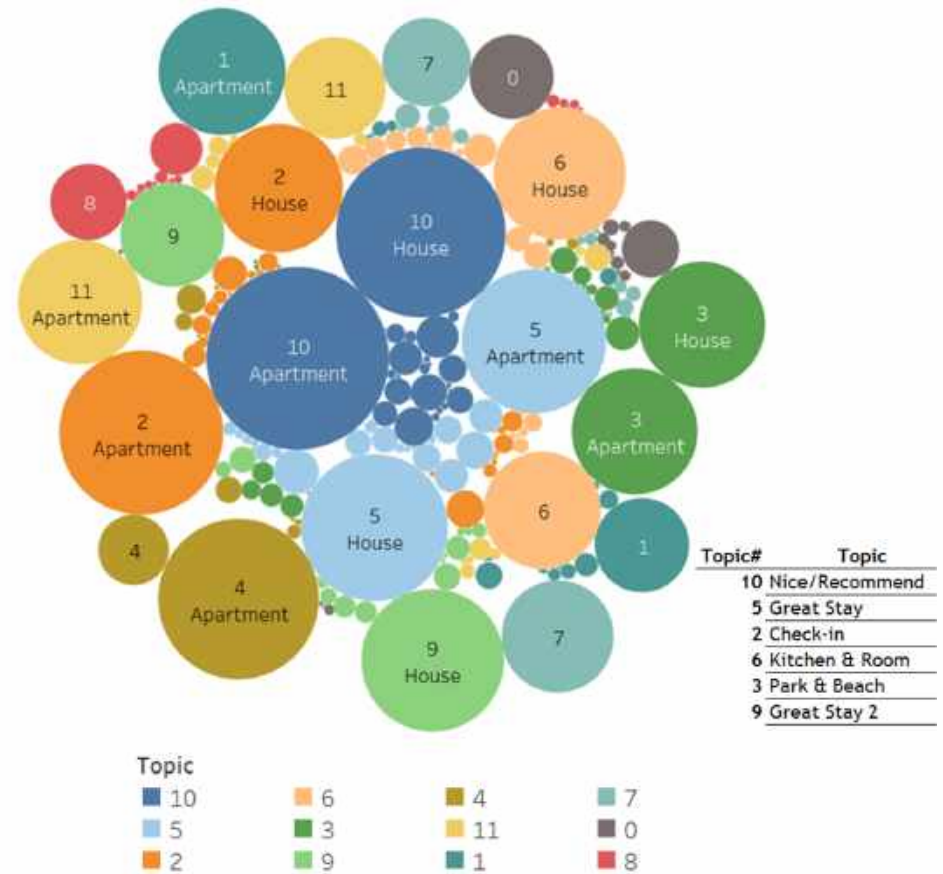
Topic#	Topic	SASEM Topic Terms
1	Host	+great host,+host
2	Foreign Language 2	â,ă,°,âșă,âceă
3	Service	+automate,+cancel,+post,+reservation,+arrival,+day
4	Home/House	+home,+house
5	Foreign Language 1	f,â,de,tră,est
6	Distance/Neighborhood	+distance,walking,+walking distance,+restaurant,+neighborhood
7	Room	+room,+nice,+bed,+bathroom,+house
8	Airbnb Experience	+experience,first,airbnb,+good,+time
9	Apartment	+apartment
10	Place/Stay	+place,+stay,+great place,+definitely,austin
11	Location	+great location,+location
12	Nice/Recommend	+nice,+recommend,highly,+clean,+apartment

Topic#	Topic	Python Topic Terms
0	Foreign Language	sehr, para, casa, todo, bien, metro, como, excelent, all, apartamento
1	Food and Shopping	restaur, walk, shop, distanc, close, bar, street, easi, neighborhood, park
2	Check-in	room, night, arriv, check, time, clean, apart, late, need, good
3	Park and Beach	park, beach, nice, quiet, clean, comfort, room, neighborhood, privat, easi
4	Location	apart, subway, clean, walk, close, nice, station, easi, need, help
5	Great Stay	home, love, perfect, beauti, walk, time, amaz, better, look, citi
6	Kitchen and Room	kitchen, room, home, comfort, coffe, like, hous, love, bathroom, need
7	Communication	question, clean, respond, easi, commun, check, quick, help, exactly, describ
8	Reservations and Cancellations	appart, reserv, nou, post, cancel, arriv, autom, pour, day, bien
9	Great Stay 2	hous, love, austin, perfect, need, definit, home, time, visit, clean
10	Recommend	recommend, highli, experi, home, clean, definit, help, welcom, comfort, accommod
11	Nice and Clean	nice, clean, good, room, time, thank, help, perfect, comfort, condo

SASEM Reviews by Topic & Property Type



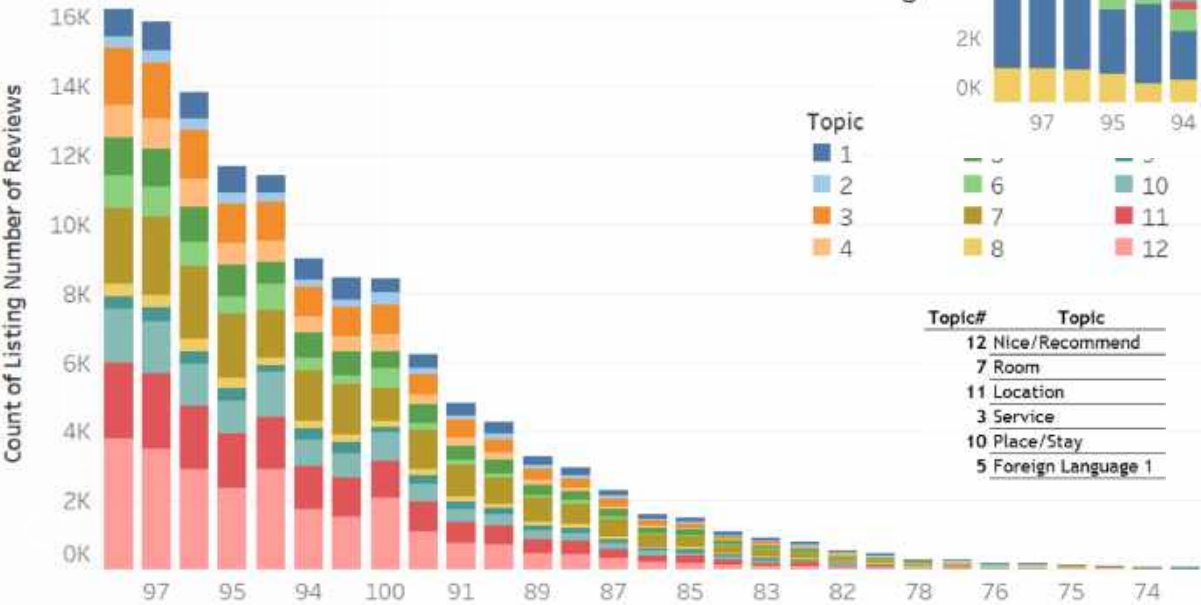
Python Reviews by Topic & Property Type





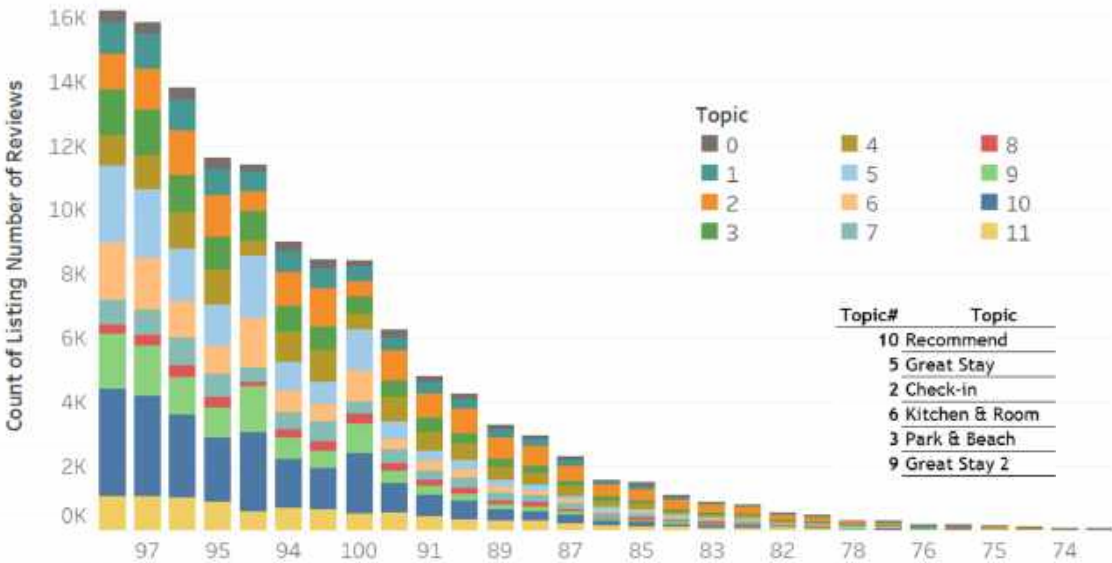
SASEM Reviews by Review Score & Topic

Listing Review Scores Rating

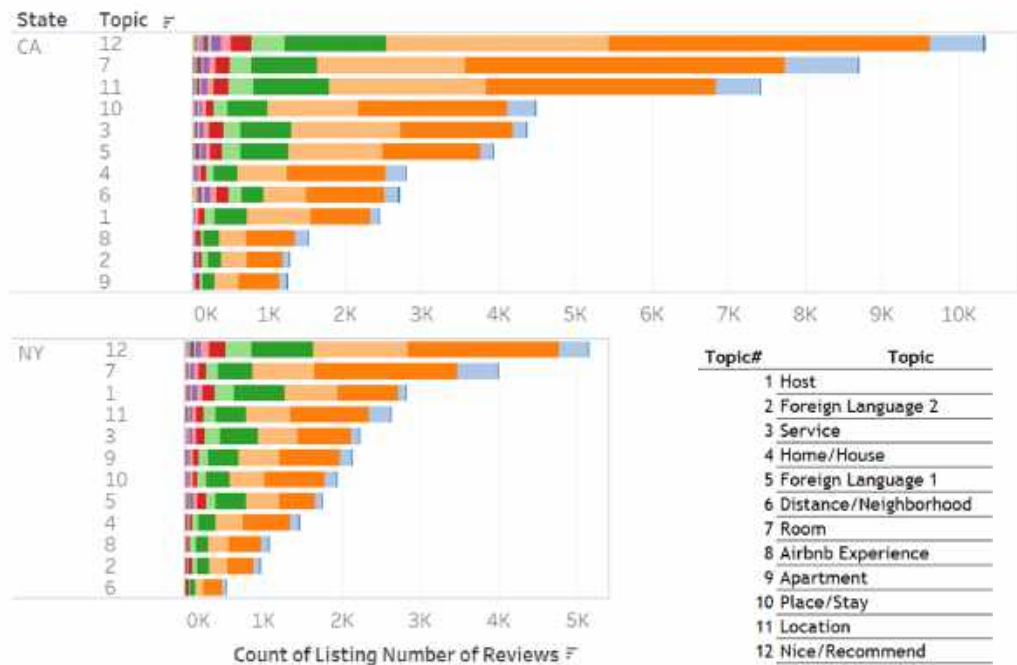


Python Reviews by Review Score & Topic

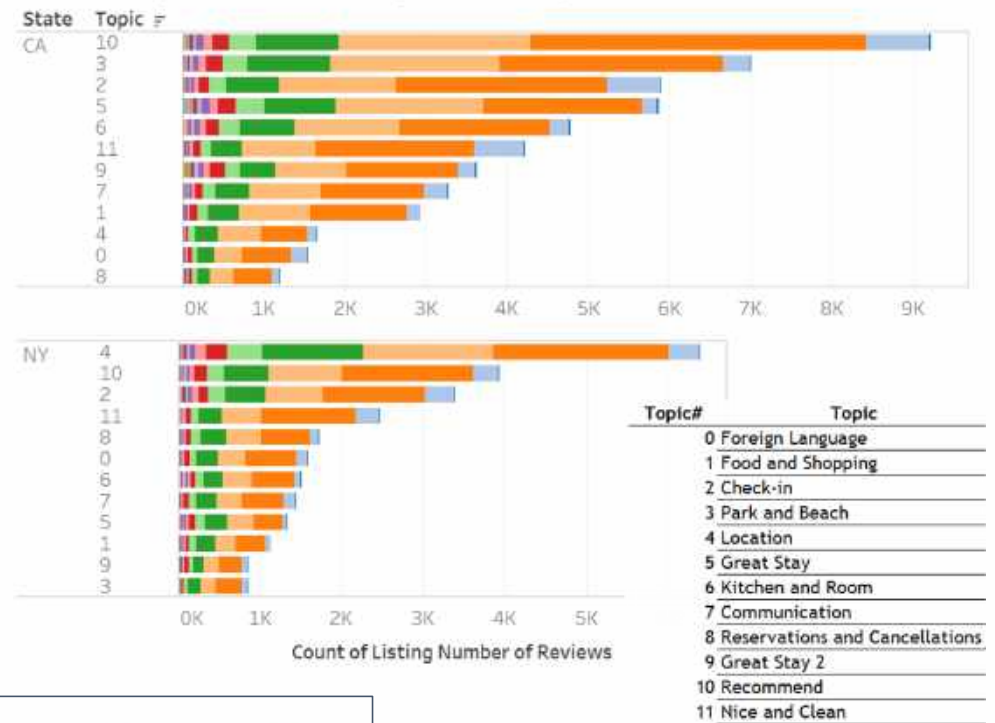
Listing Review Scores Rating



SASEM Reviews by State, Topic & Price - CA & NY



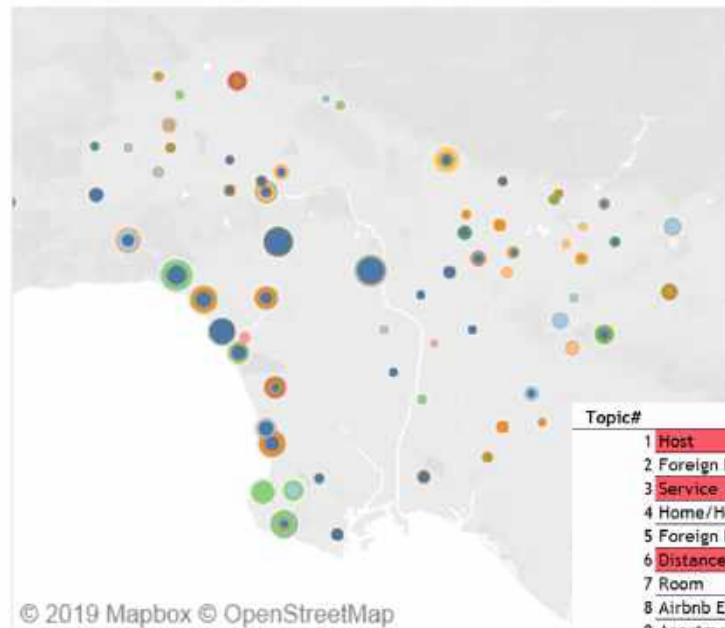
Python Reviews by State, Topic & Price - CA & NY



Listing Price BINS



SASEM Topic by Price - CA



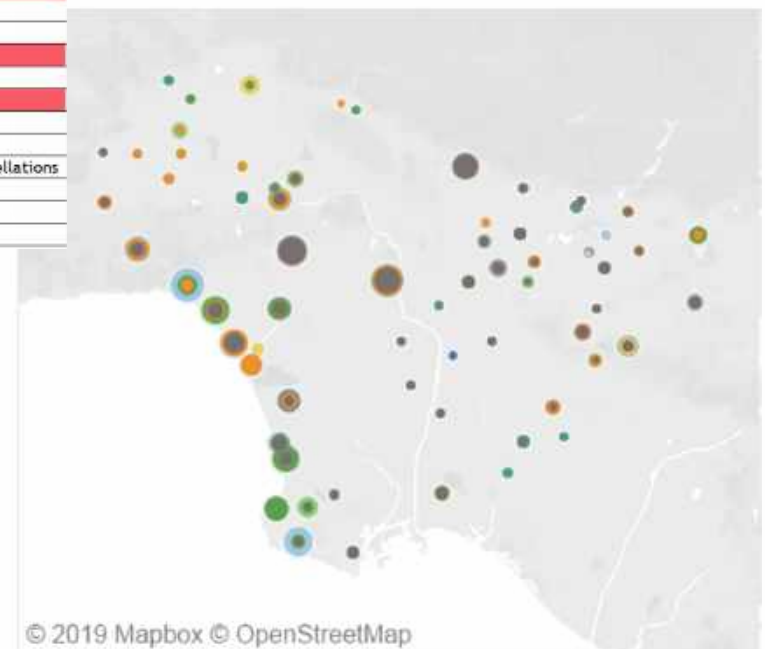
Topic



Topic#	Topic
0	Foreign Language
1	Food and Shopping
2	Check-in
3	Park and Beach
4	Location
5	Great Stay
6	Kitchen and Room
7	Communication
8	Reservations and Cancellations
9	Great Stay 2
10	Recommend
11	Nice and Clean

Topic#	Topic
1	Host
2	Foreign Language 2
3	Service
4	Home/House
5	Foreign Language 1
6	Distance/Neighborhood
7	Room
8	Airbnb Experience
9	Apartment
10	Place/Stay
11	Location
12	Nice/Recommend

Python Topic by Price - CA



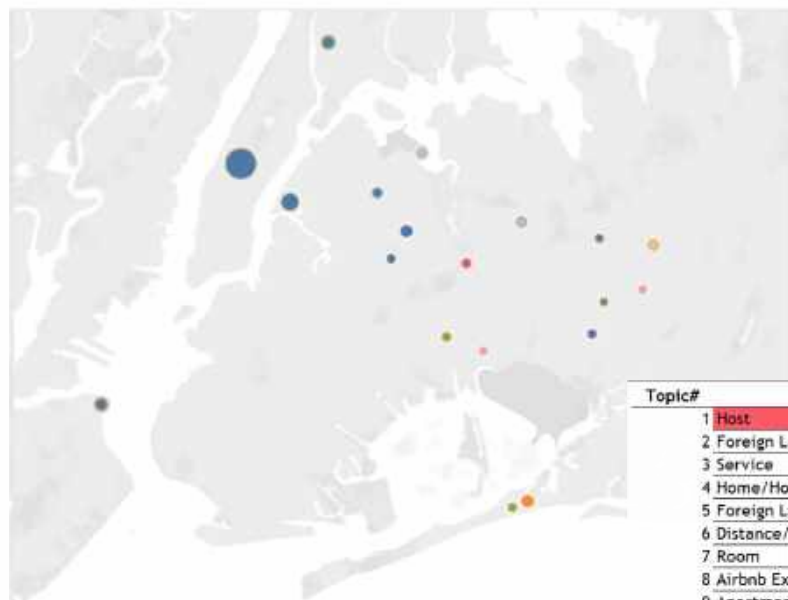
Topic



Listing Price BINS

• Null	• 100	• 250	• 400	• 550	• 700	• 850
• 0	• 150	• 300	• 450	• 600	• 750	• 900
• 50	• 200	• 350	• 500	• 650	• 800	• 950

SASEM Topic by Price - NY



© 2019 Mapbox © OpenStreetMap

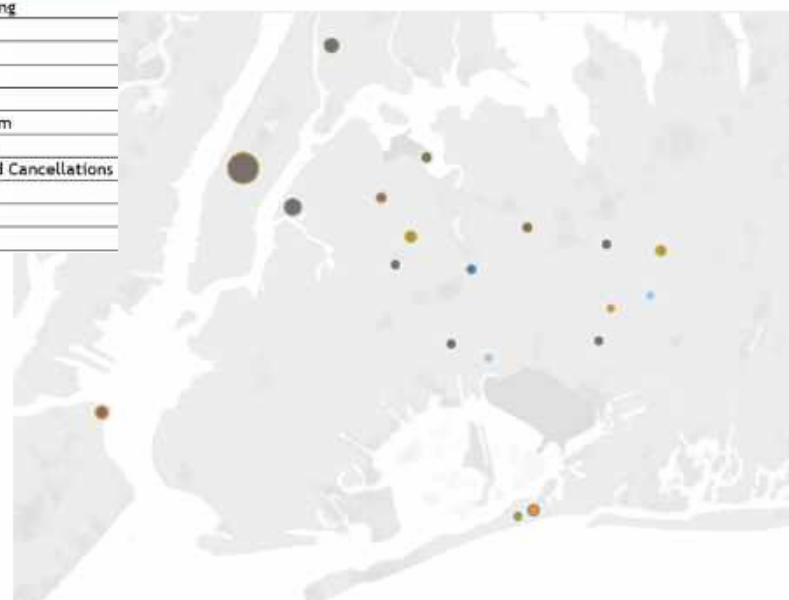
Topic

- | | | |
|---|---|----|
| 1 | 5 | 9 |
| 2 | 6 | 10 |
| 3 | 7 | 11 |
| 4 | 8 | 12 |

Topic#	Topic
0	Foreign Language
1	Food and Shopping
2	Check-in
3	Park and Beach
4	Location
5	Great Stay
6	Kitchen and Room
7	Communication
8	Reservations and Cancellations
9	Great Stay 2
10	Recommend
11	Nice and Clean

Topic#	Topic
1	Host
2	Foreign Language 2
3	Service
4	Home/House
5	Foreign Language 1
6	Distance/Neighborhood
7	Room
8	Airbnb Experience
9	Apartment
10	Place/Stay
11	Location
12	Nice/Recommend

Python Topic by Price - NY



© 2019 Mapbox © OpenStreetMap

Topic

- | | | |
|---|---|----|
| 0 | 4 | 8 |
| 1 | 5 | 9 |
| 2 | 6 | 10 |
| 3 | 7 | 11 |

Listing Price BINS

- | | | | | | | |
|--------|-------|-------|-------|-------|-------|-------|
| • Null | • 100 | • 250 | • 400 | • 550 | • 700 | • 850 |
| • 0 | • 150 | • 300 | • 450 | • 600 | • 750 | • 900 |
| • 50 | • 200 | • 350 | • 500 | • 650 | • 800 | • 950 |



Limitations and Future Use



Limitations

Software: SASEM can only take a proportion of data

Under-reporting: Survival bias for listings

Dictionary: Words alone do not provide adequate information

Meaning: 'Apple' is different from 'Apple'

Context: 'Recommend' is one of common words used in NEGATIVE reviews

More than words: Emoji — :-) :> XD (° ▽ °)



For the Future

Customers & Owners

Streamline searching process

Sentiment analysis to uncover true customers' demand

ANOVA and hypothesis test to find connections among topics

Prediction for price based on topics as predictors

US

Data mining is a subject of art & science

Human analysis still necessary

Reviews

★ 4.93 201 reviews

Hello, MA710 X Q

Cleanliness	5.0	Communication	5.0
Check-in	4.9	Value	4.9
Accuracy	4.9	Location	4.8

Topic#	Topic
12	Nice/Recommend
7	Room
11	Location
3	Service
10	Place/Stay
5	Foreign Language 1



THANK YOU!

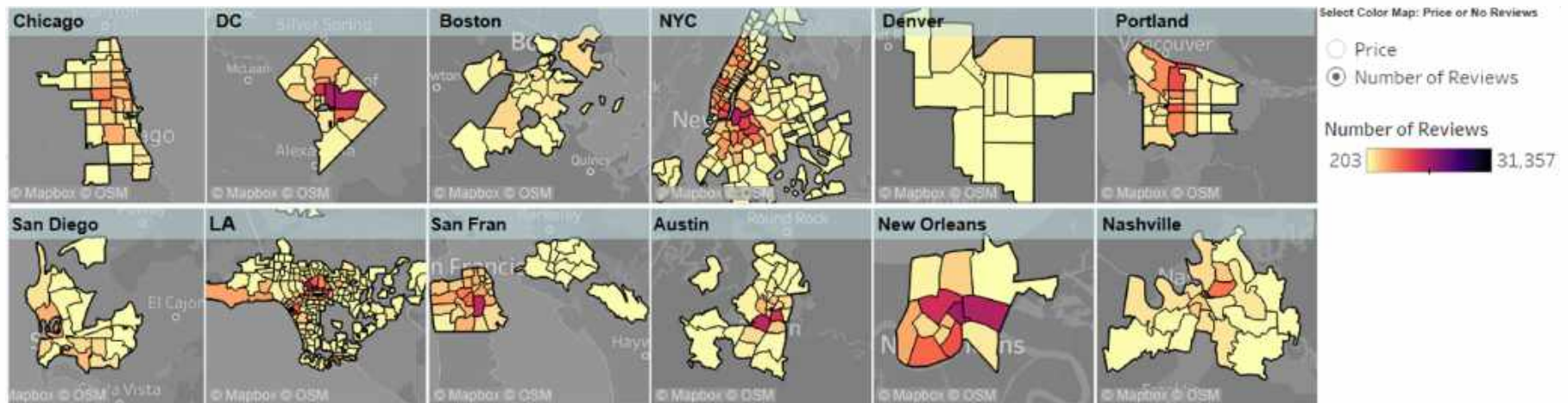


APPENDIX

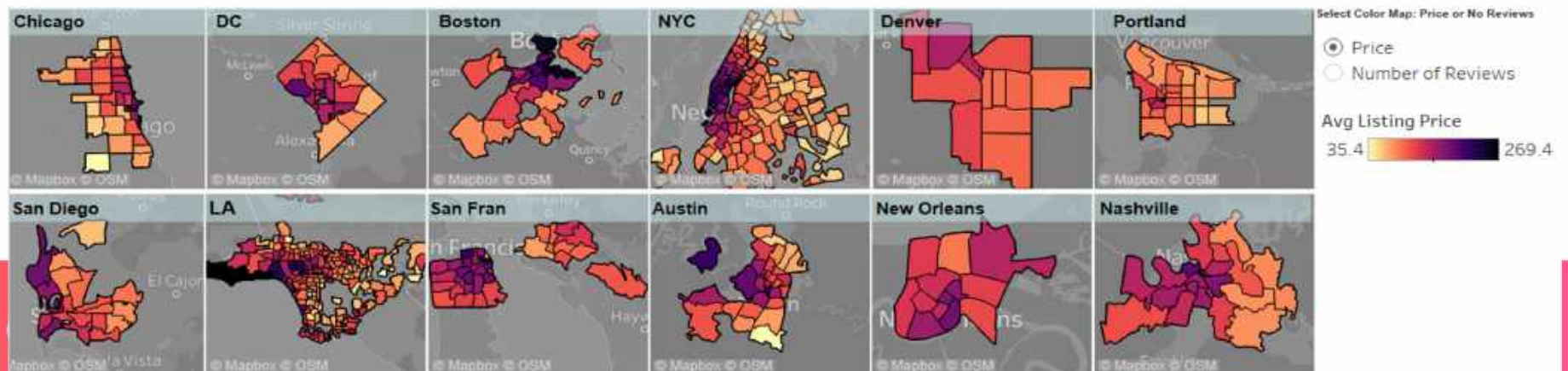


Reviews and Average Price by Zip Code

We notice that for each of the main areas as we get closer to the city center number of reviews increases greatly. Also, we notice that the number of reviews is **very concentrated** in a few key zip codes (especially in LA and NYC).



We notice some very dark spots in LA in Malibu, Boston (seaport areas) and some areas in NYC indicating very expensive listings. We also notice again that moving away from the center of the city usually results in lower costs, especially in NYC and Chicago.

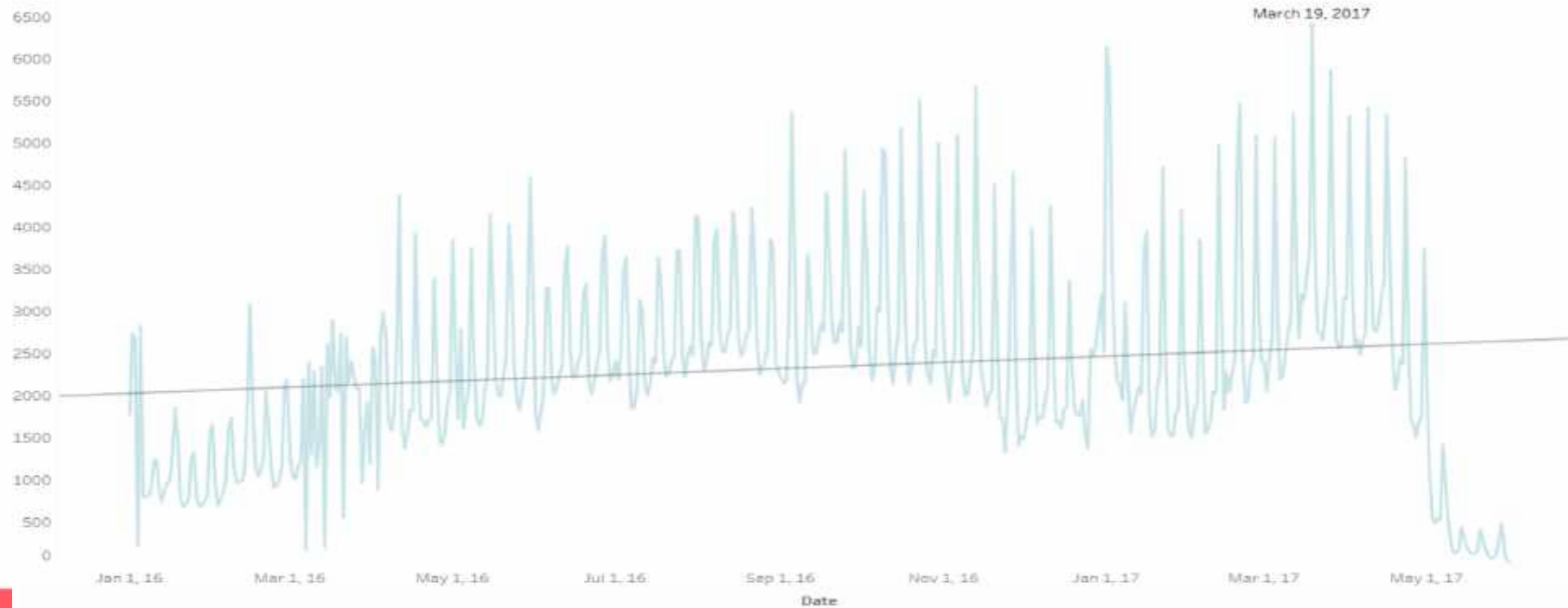


Reviews by Day

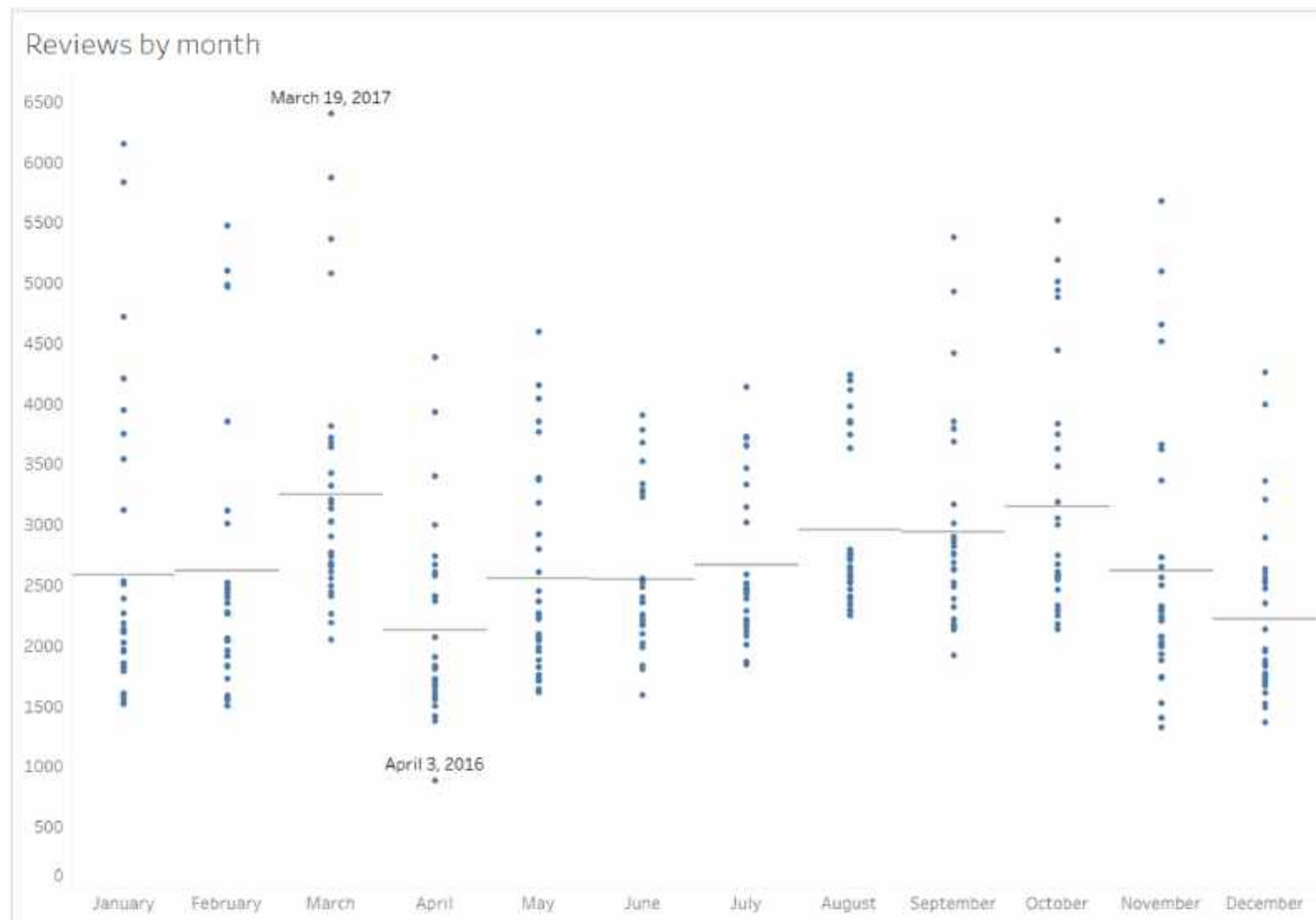
Reviews by day indicate a cyclical pattern where the first 2 months and the last month have a huge drop off.

We examine reviews by day and month to further investigate the cyclical nature of the number of reviews.

Airbnb Reviews by Day



Reviews by Day - cont.



Reviews by Day - cont.

The cyclical nature seems to be largely contributed to the days of the week people submit reviews.

It makes sense that people stay at an AirBNB on the weekend, and will submit a review on the Sunday or Monday after finishing the trip.

Here, we filter on the date range Apr 1, 2016 - Mar 31, 2017 because of the drop in number of reviews in the first couple months and last month.

