Classifying Airbnb Reviews With Natural Language Processing

Brenner Heintz

Data

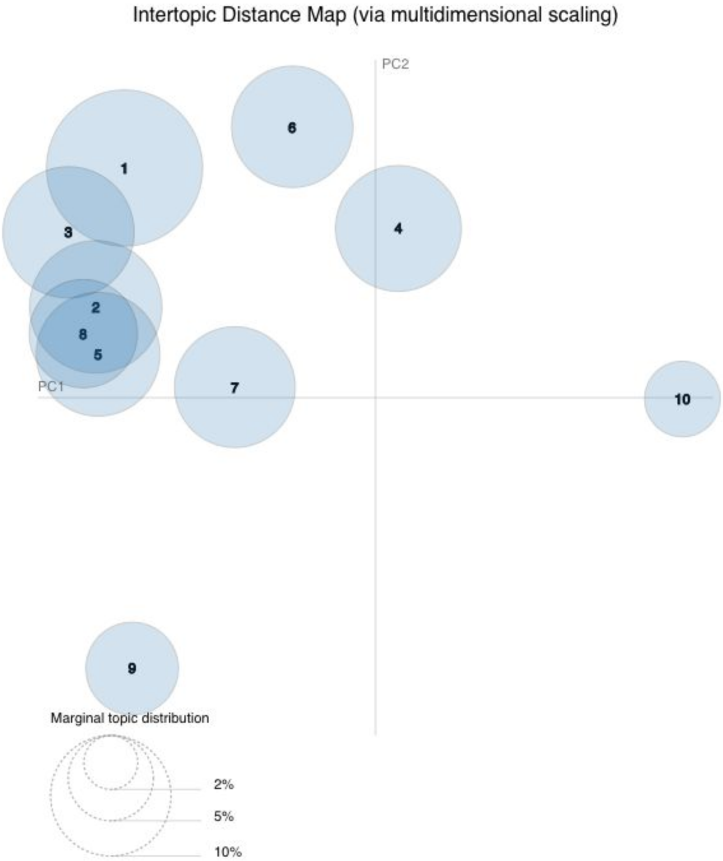
- InsideAirbnb.com
- Portland, Oregon
- 4,911 listed rental properties
- 318,262 reviews

Goals

- **Classify customers into groups using NLP** to analyze their reviews.
- Use those groups (along with listing data) to **predict the price of Airbnb rentals.**

Methodology

- Data munging & preprocessing
  - Punctuation, etc.
  - Stop words
- Topic modeling: Latent Dirichlet Allocation (10 topics)
  - Stemming
  - Bi-grams



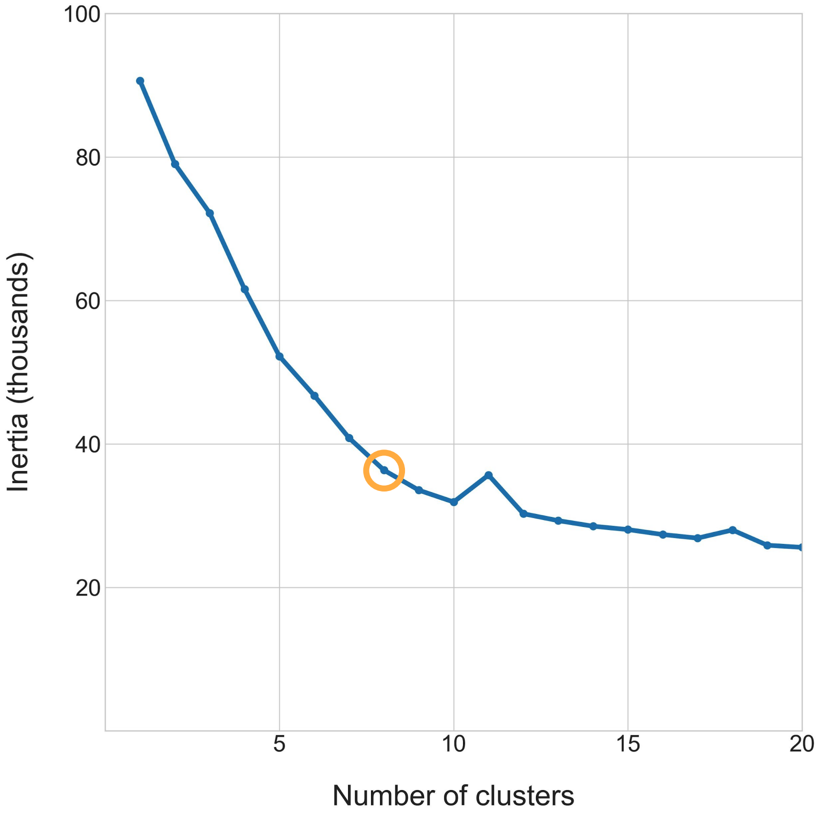
Topic Modeling

airbnb

Topic	**“Close to Everything”**	**“Feels Cozy”**	**“Host with the Most”**	**“Holistic Experiences”**
Words	**great location** **downtown** **walking distance** **neighborhood** **restaurants** **food**	**feels (like) home** **perfect** **cottage** **space** **comfortable**	**host** **gracious** **wonderful host** **easy (to) communicate** **loved**	**great experience** **really nice** **exactly (as) described** **best** **good** **time**
Example Quote	“...the apartment is really close to bus lines...we took public transportation to go everywhere…” “It's a great location, close to restaurants and bars.” “Peninsula Park, New Seasons Market and access to I-5 are also really close to the apartment”	“Very comfortable and felt like home.' “...this was a magical, cozy space that made us dream of creating a similar converted-attic space for our kids some day…” “In fact, it didn't feel like a basement at all!...cohesive design sense that is modern yet cozy”	“Thank you for hosting and I appreciate the good communication…” “Leslie was super accommodating...They were super attentive and responded really quickly....	“Excellent first experience with Airbnb, exactly as described...“ “Overall, we had a great experience and couldn’t be happier with Sara’s condo.”

K-Means Clustering

- Used K-Means to create 8 clusters based upon LDA probabilities
- Fed into supervised learning algorithm to predict rental prices

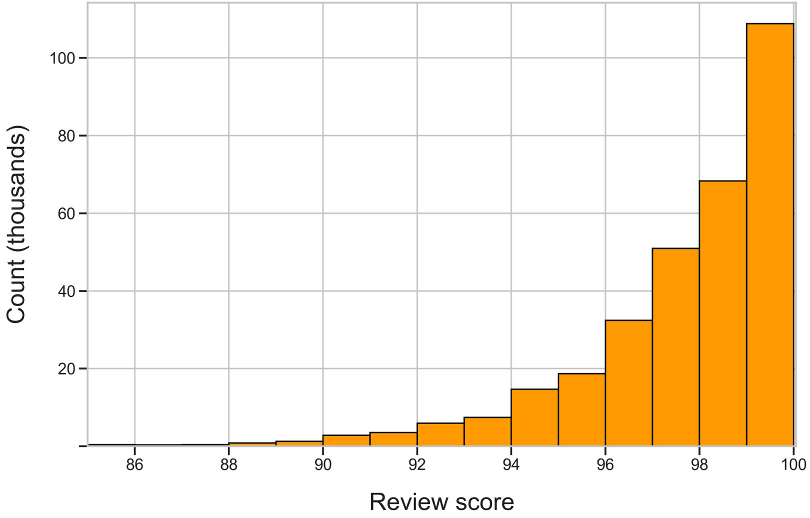


Regression Presented Challenges

- Best R^2 score = Random Forests (0.14)
- Failure to reject null hypothesis → cannot accurately predict rental prices based upon current features
- Still, topic modeling may provide some insight

Lessons Learned

- Reviews on Airbnb carry very little information, and are *extremely* skewed
  - May be biased by “quid pro quo”
  - Company may ferret out bad reviews or intervene with unhappy customers
- Topic modeling is difficult for reviews, since they are inherently limited in terms of topic scope



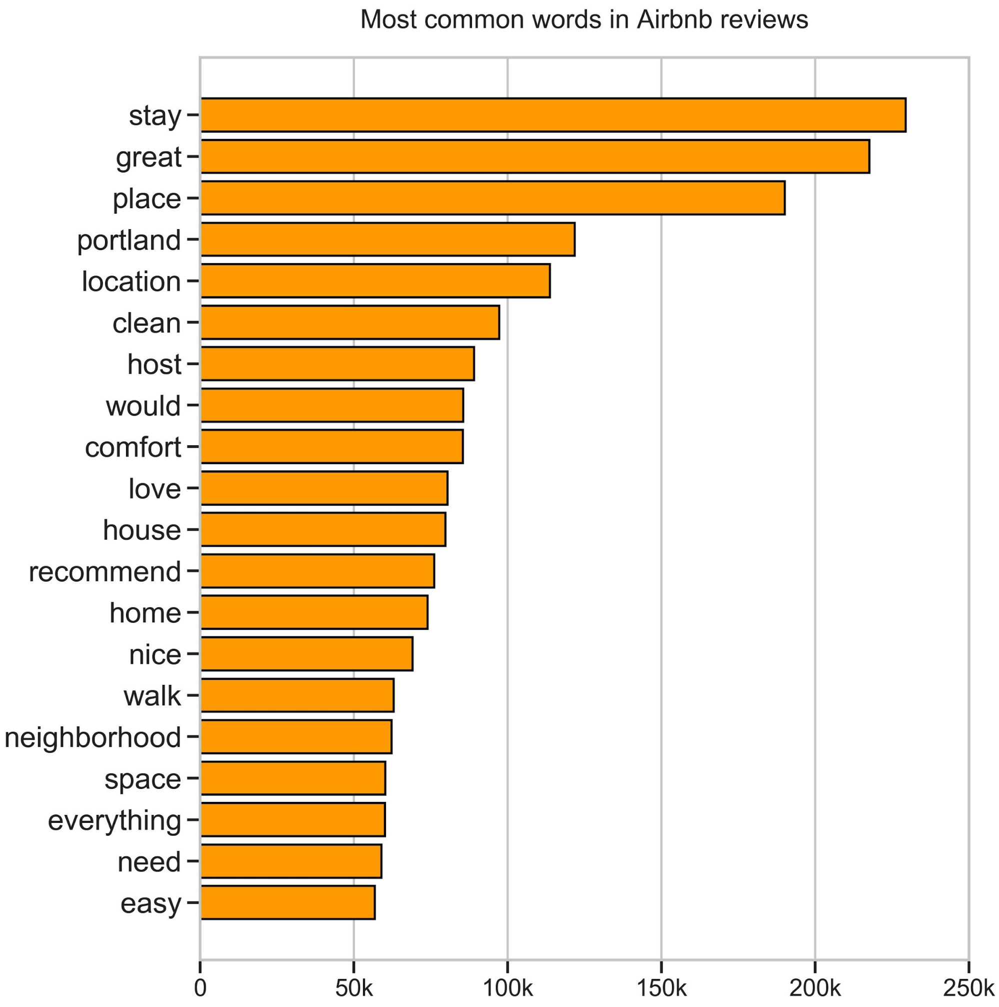
Future Work

- Look at markets outside of Portland with wider price ranges
- Bring in features from property descriptions
- Focus analysis on bad reviews, and the factors that are most likely to cause them
    - Likely to produce more robust topic modeling results

airbnb

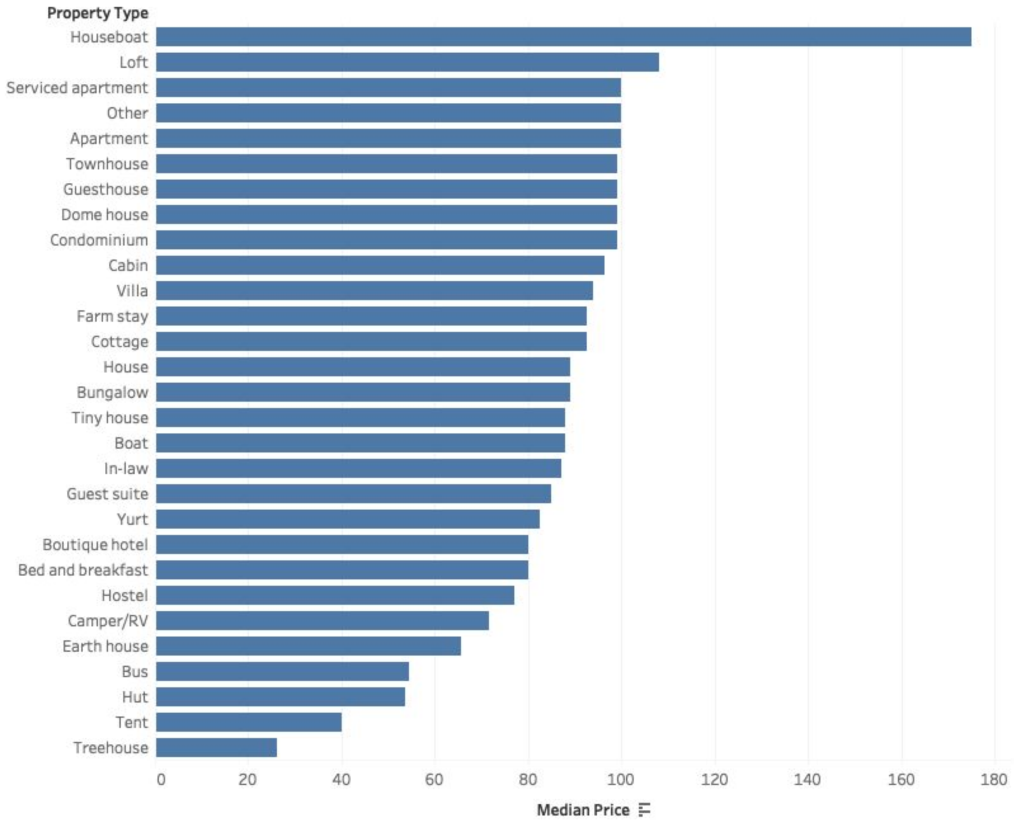
Thank You

Brenner Heintz, Metis ‘18
brenner.heintz@gmail.com
github.com/athena15

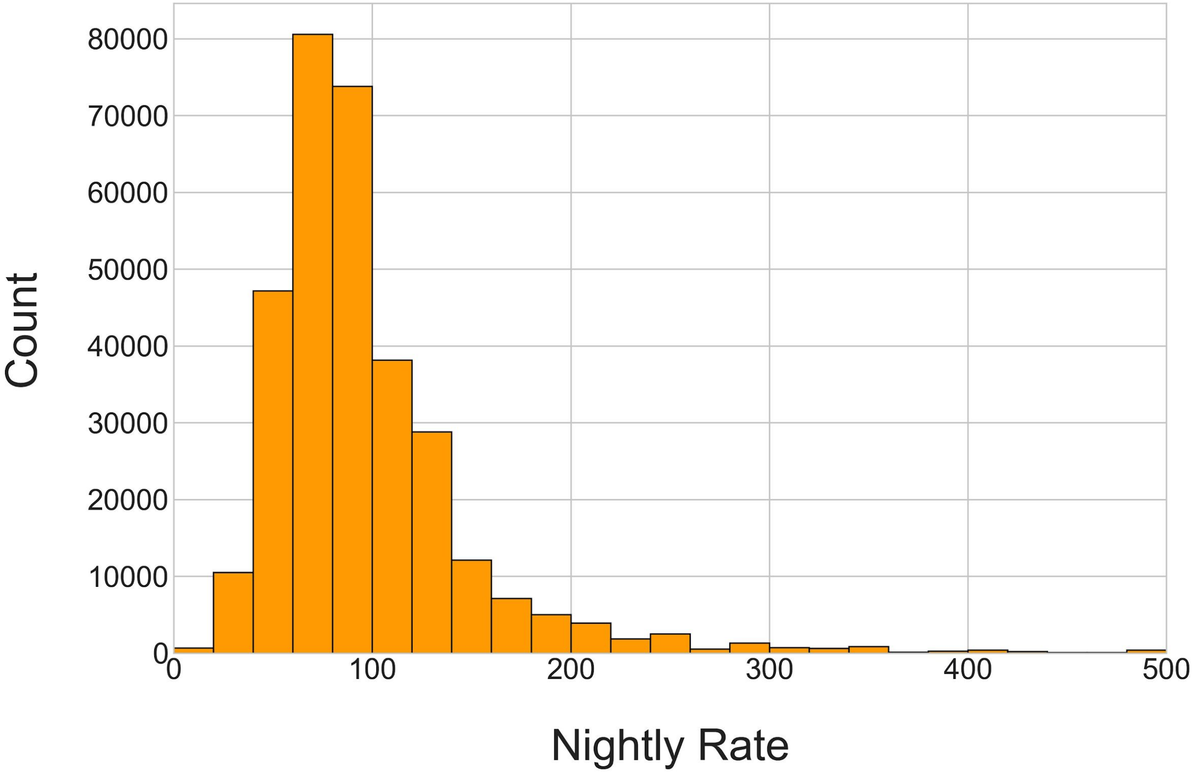

Most common words in Airbnb reviews
stay
great
place
portland
location
clean
host
would
comfort
love
house
recommend
home
nice
walk
neighborhood
space
everything
need
easy
0
50k
100k
150k
200k
250k

airbnb

Average Cost by Property Type

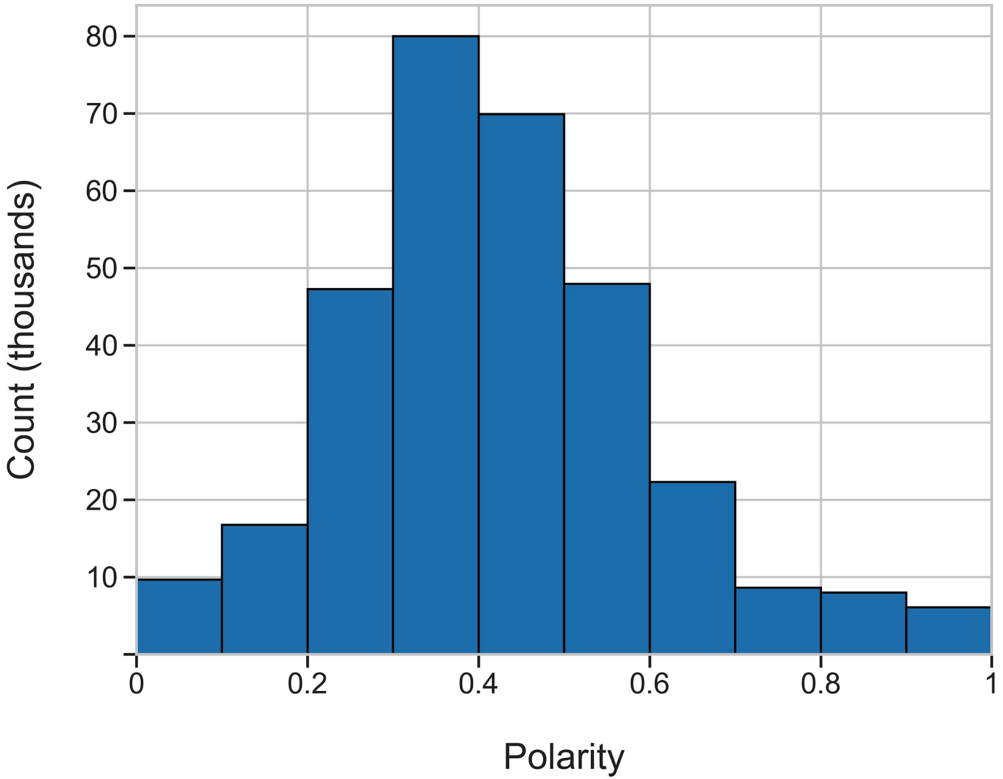


Average Portland Airbnb Nightly Rate

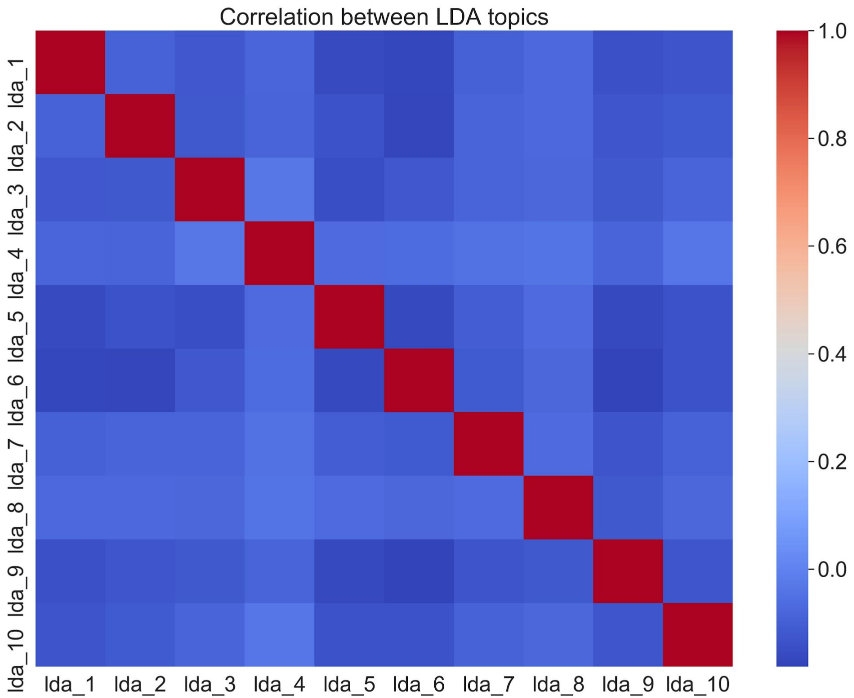


Review Sentiment (Polarity)

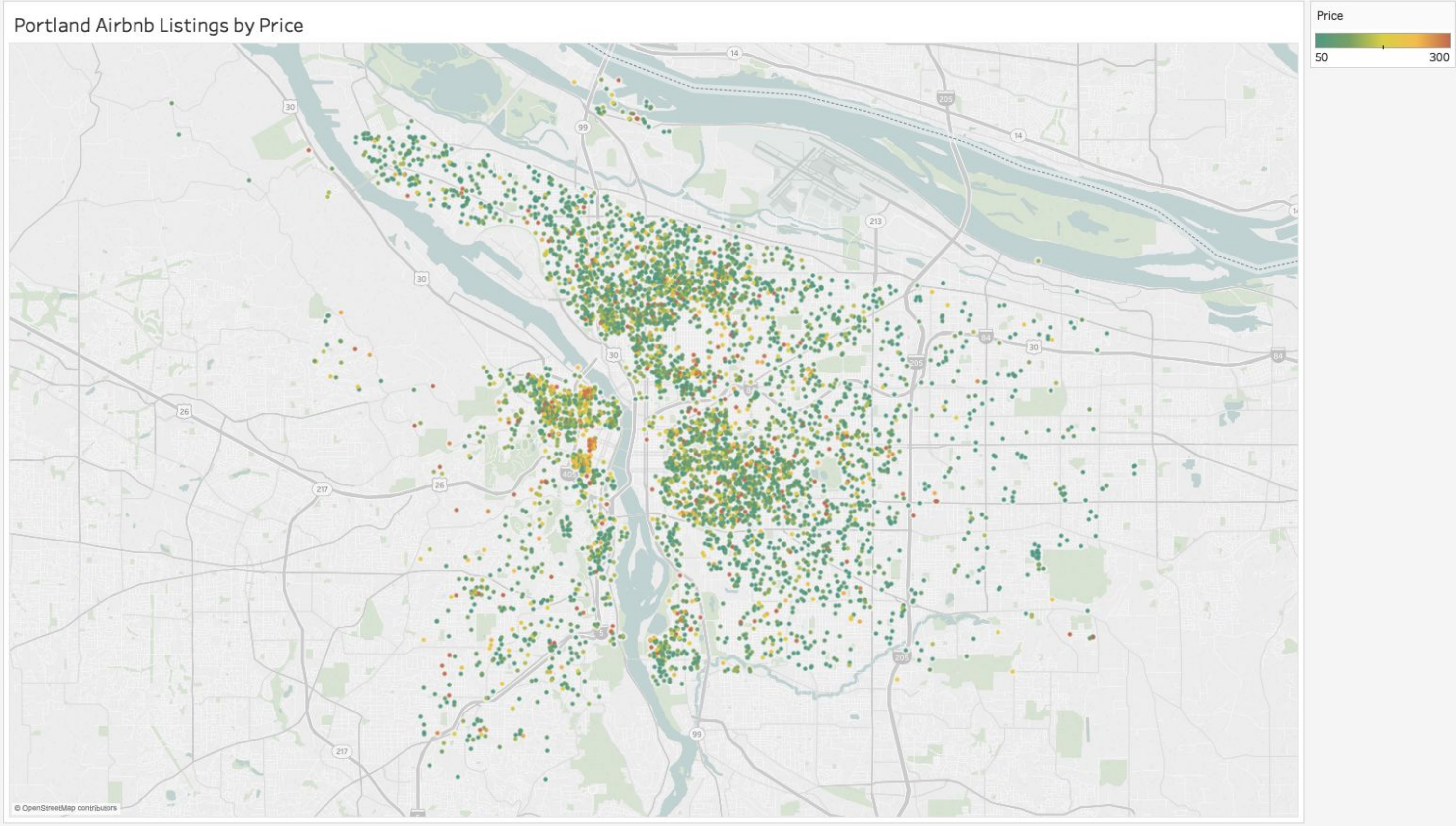
airbnb



Correlation between LDA topics

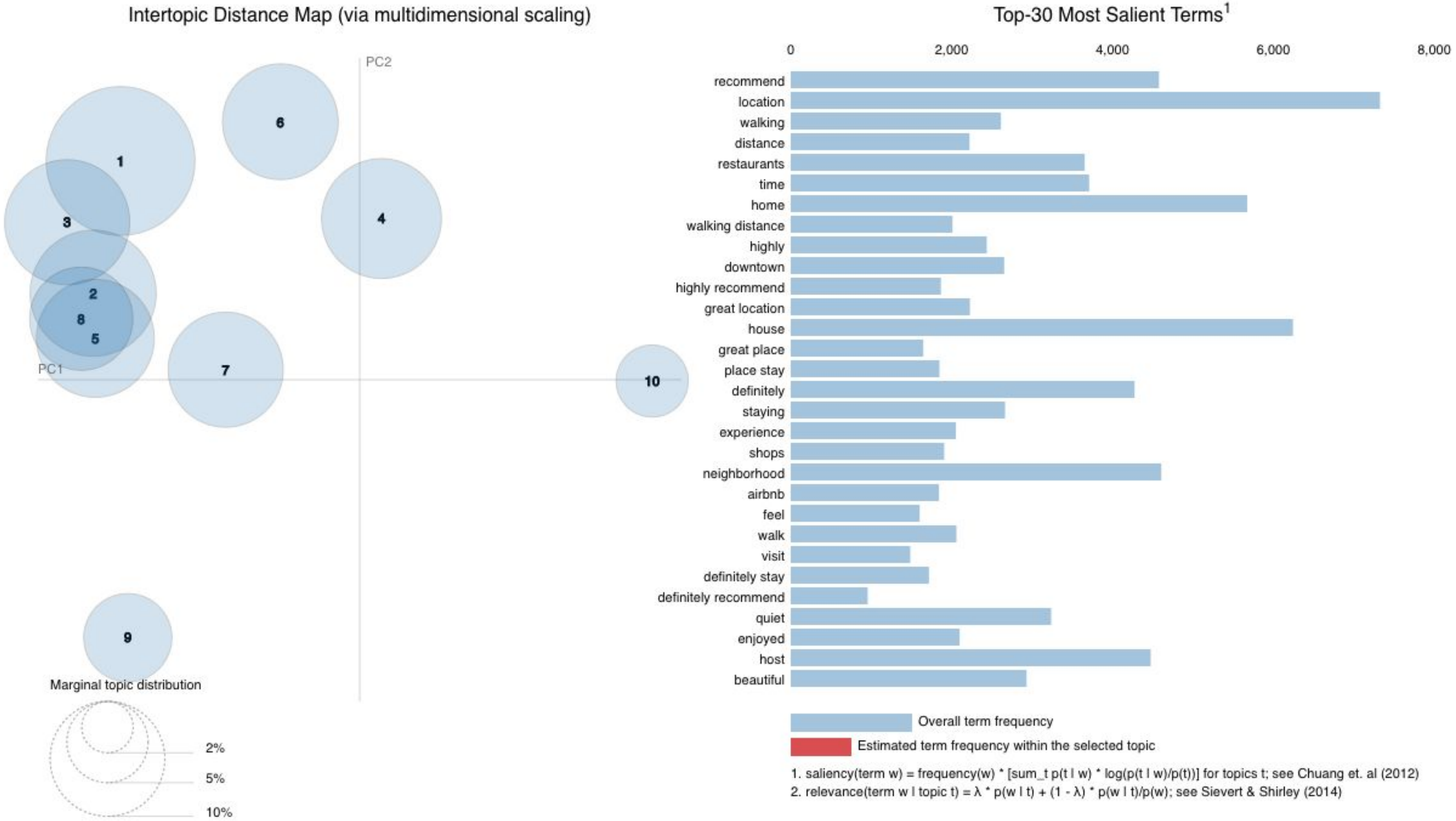


Airbnb Listings Mapped by Price



Portland Airbnb Listings by Price

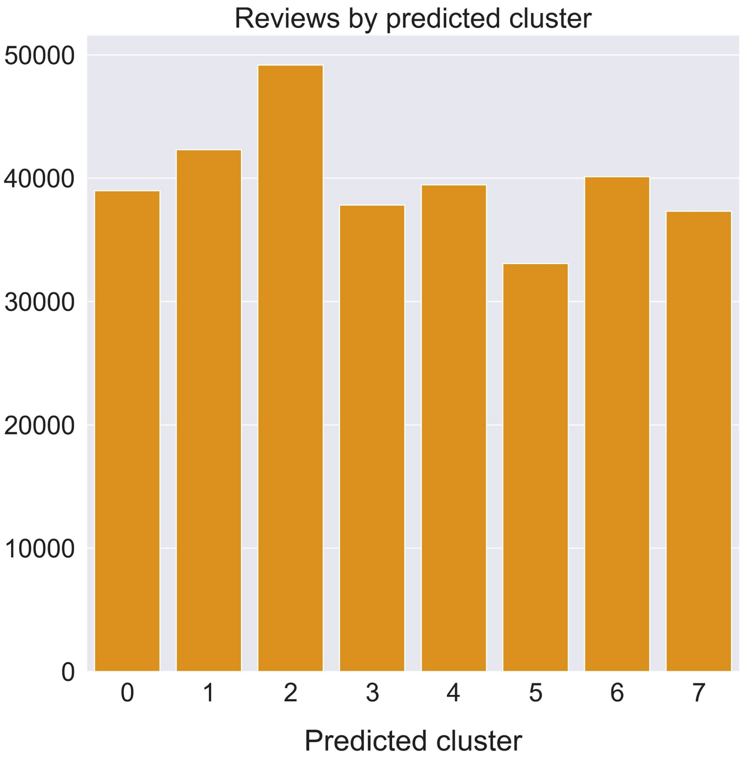
Latent Dirichlet Allocation

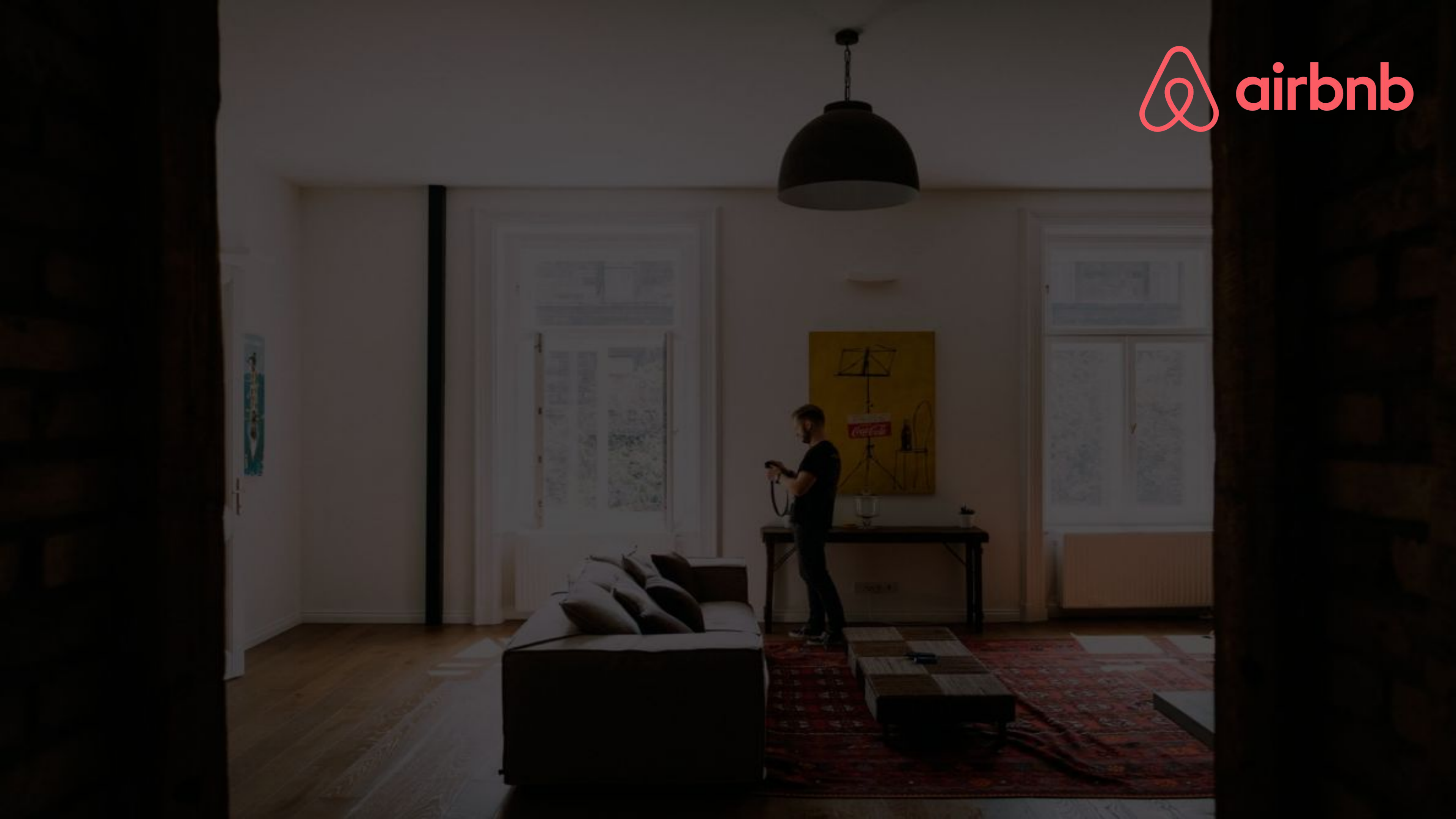

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Salient Terms[1]

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Reviews Per K-Means Cluster




airbnb