How to Succeed as an Airbnb Host
——Evidence from Machine Learning and Text Analysis Approaches

Yutian Lai
Master of Computational Social Science Program, University of Chicago

E-mail: ylai@uchicago.edu
Date: June 3rd, 2020

Research Question

Q1: What factors are the driving forces behind Airbnb review ratings?

Q2: Are there any geospatial differences?

- Review scores are metric for the satisfaction-level of the guests.
- Hosts can condition their property/self attributes
- Airbnb can incentivize their hosts to change

Data Collection and Pre-processing

http://insideairbnb.com/get-the-data.html (April 2020)

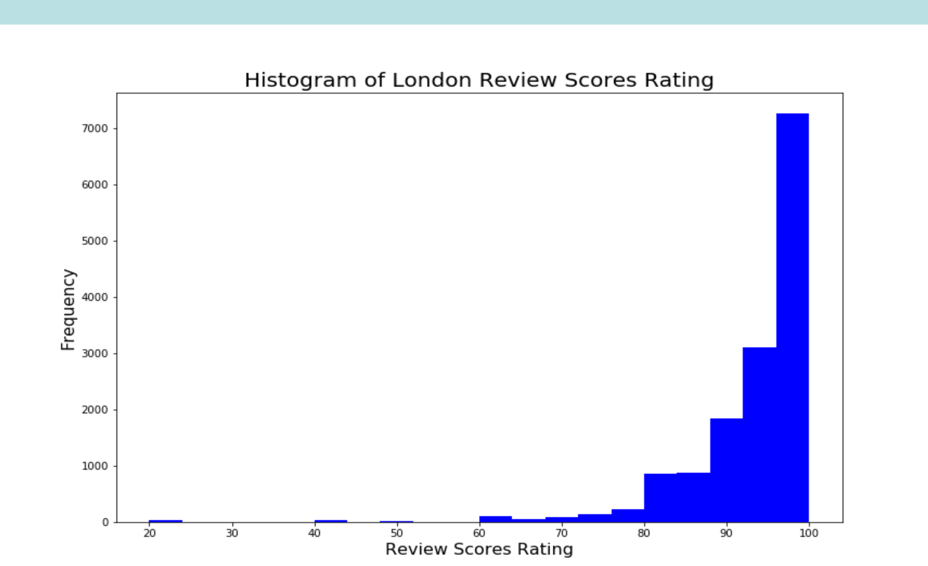
1. Drop/Create/Modify attributes

Numerical: 'beds', 'bedrooms', 'guests_included', 'bathrooms', 'review_scores_communication', etc.

Categorical: 'host_has_profile_pic', 'TV' 'host_identity_verified', 'property_type', etc.

Text: 'house_description', 'self_about'

Response Variable: 'review_scores_rating'



2. Heatmaps---No feature removed

3. Data normalization + Text: uni-/bi-gram

City	Sample Size	Feature Number
London	14676	10051
New York	8427	10052
Tokyo	5288	10049

4. Pipeline Construction

- Pipeline1: (numerical, categorical, text(bag of words))
- Pipeline2: (numerical, categorical, text(TF-IDF))

Machine Learning Approach and the Results

Models:

- Ridge regression (both pipelines)
- Lasso regression (both pipelines)
- Random forest (the first pipeline)
- Gradient boosting(the first pipeline)

Hyperparameter tuning:

- randomized search method

Optimization:

- 5-fold cross-validation with 10 iterations on the training set

Performance metric:

- r2 (Main metric)
- Mean squared error (MSE)
- Mean absolute error (MAE)

London

Model Name	Training R2	Test R2	MSE	MAE
Ridge1	0.84	0.61	29.1	3.66
Ridge2	0.81	0.66	25.55	3.3
Lasso1	0.68	0.66	25.33	3.2
Lasso2	0.67	0.65	25.86	3.24
Random Forest	0.9	0.68	24.02	2.94
Gradient Boosting	0.84	0.69	23.1	2.93

Top five features

Review scores for cleanliness(+)
Review scores for communication(+)
Review scores for location(+)
Number of reviews(+)
Verification as super host(+)

London-specific features

Property/neighborhood attributes

bathrooms, beds, bedrooms, guest capacity, beautiful kitchen

New-York

Model Name	Training R2	Test R2	MSE	MAE
Ridge1	0.74	0.68	12.56	2.39
Ridge2	0.69	0.67	12.73	2.37
Lasso1	0.66	0.65	13.4	2.43
Lasso2	0.65	0.64	14.02	2.47
Random Forest	0.87	0.62	14.78	2.38
Gradient Boosting	0.86	0.65	13.49	2.31

Top five features

Review scores for cleanliness(+)
Review scores for communication(+)
Review scores for location(+)
Verification as super host(+)
Flat(unigram)(-)

New York-specific features

Host attributes

Time as host, host acceptance rate (almost in every model)

Tokyo

Model Name	Training R2	Test R2	MSE	MAE
Ridge1	0.74	0.71	14.05	2.63
Ridge2	0.7	0.71	14.24	2.68
Lasso1	0.67	0.7	14.72	2.72
Lasso2	0.67	0.7	14.85	2.73
Random Forest	0.87	0.68	15.52	2.54
Gradient Boosting	0.86	0.7	14.51	2.59

Top five features

Review scores for cleanliness(+)
Review scores for communication(+)
Review scores for location(+)
Verification as super host(+)
Meter(unigram)(-)

Tokyo-specific features

Location

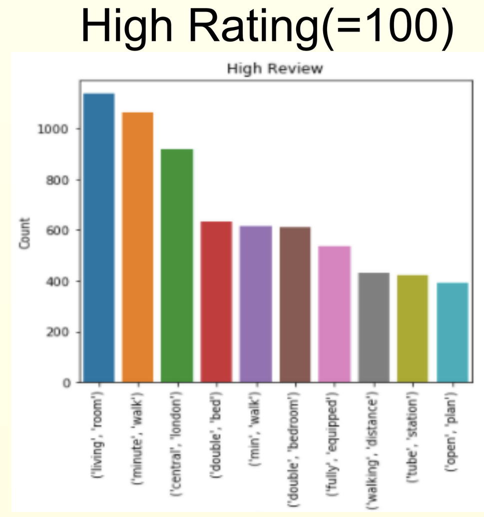
Asakusa/Ueno(-), Narita(-), Shibuya (+),Ginza(+)

Text Analysis(Bigram analysis) Approach and the Results

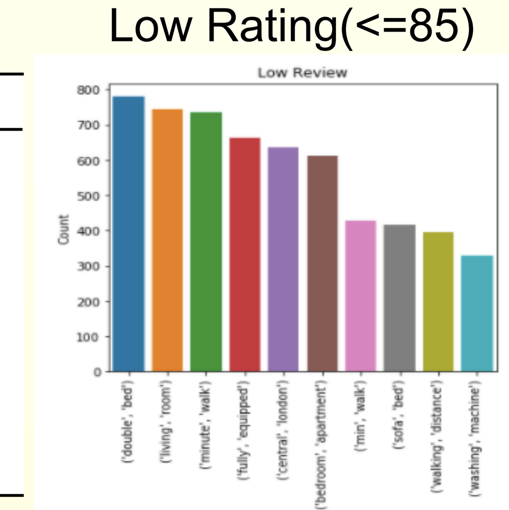
House Description

High Rating(=100) | Low Rating(<=85)

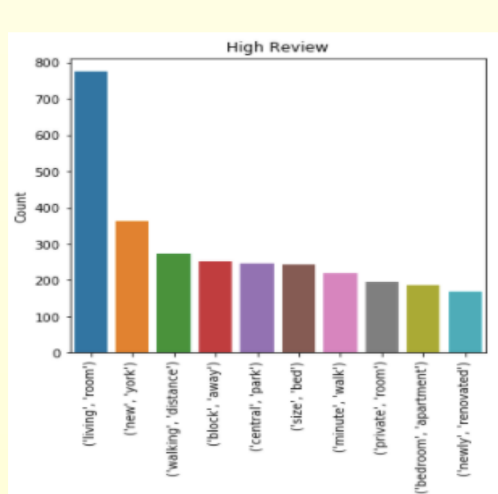
London



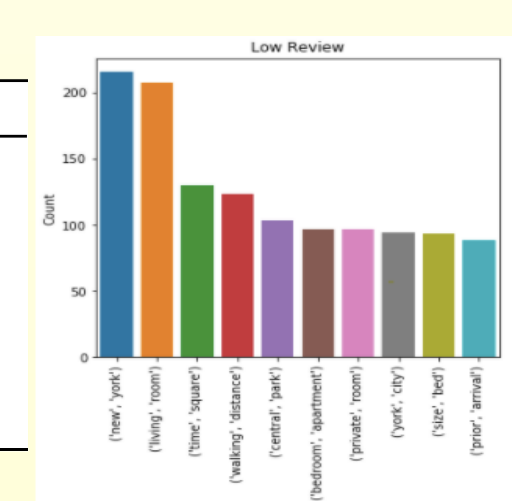
- **Amenities**: double bed
  Location: central London
- **Neighborhood**: min walk
- **Transportation**: tube station



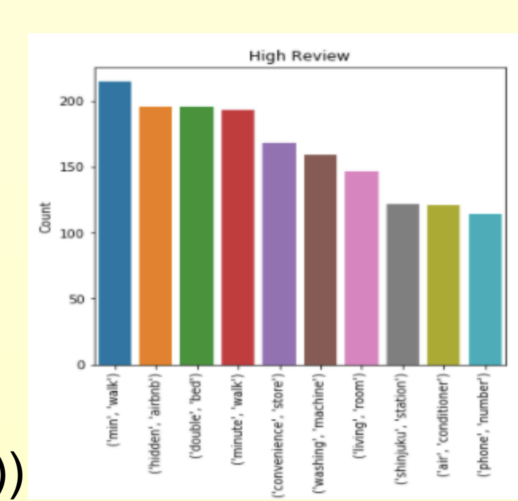
New York



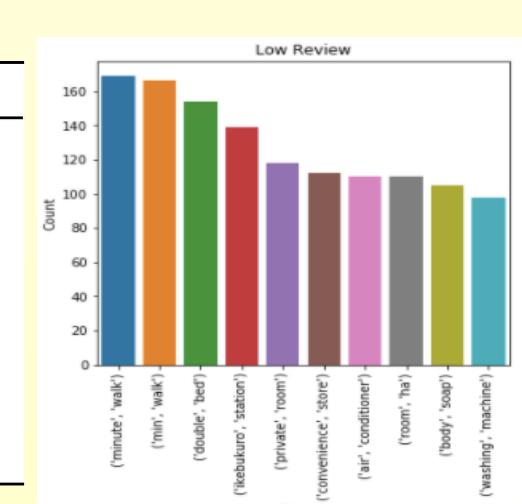
- **Amenities**: living room, newly renovated (high)
- **Location**: New York
- **Neighborhood**: min walk



Tokyo



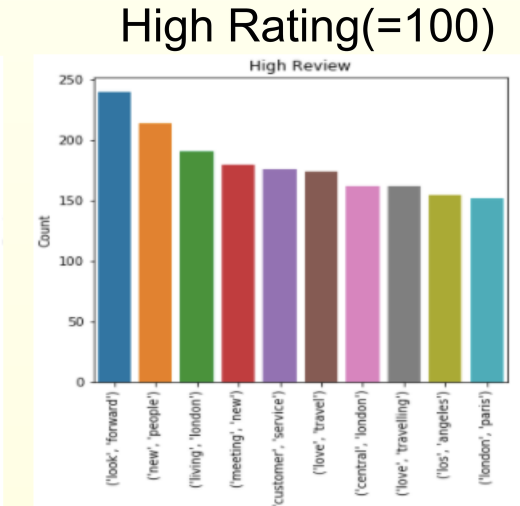
- **Amenities**: double bed
  Location: Ikebukuro station (low), Shinjuku station (high)
- **Neighborhood**: min walk



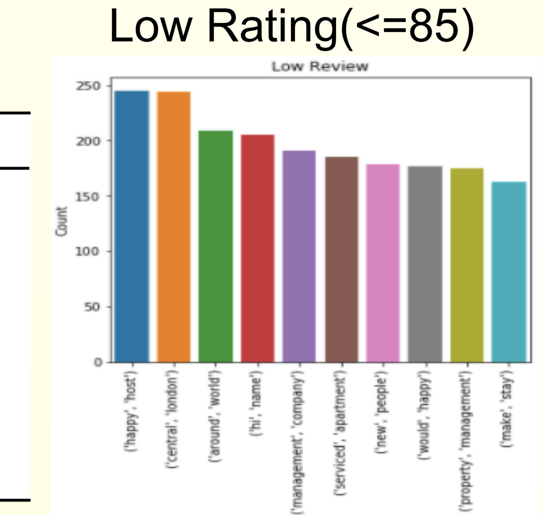
Host Description

High Rating(=100) | Low Rating(<=85)

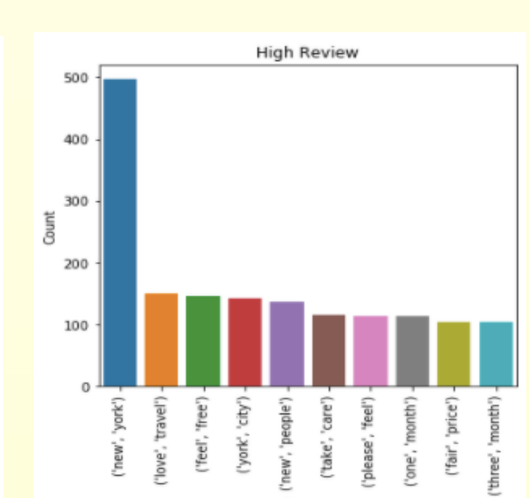
London



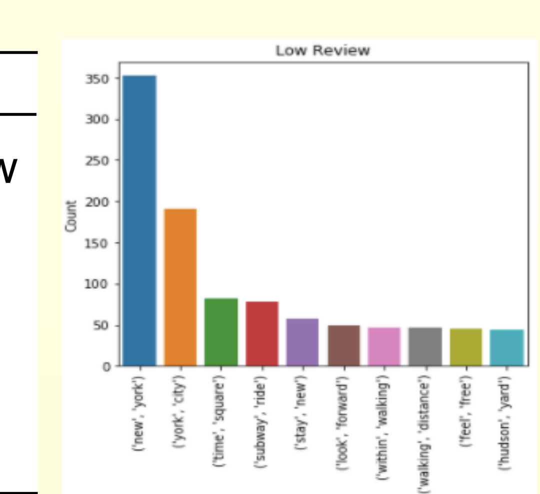
- **Self-background**: management company
- **Welcome message:** would happy
- **Resonance (high)**:love travel



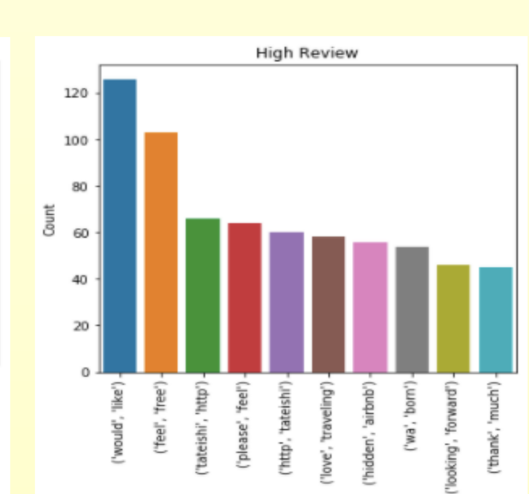
New York



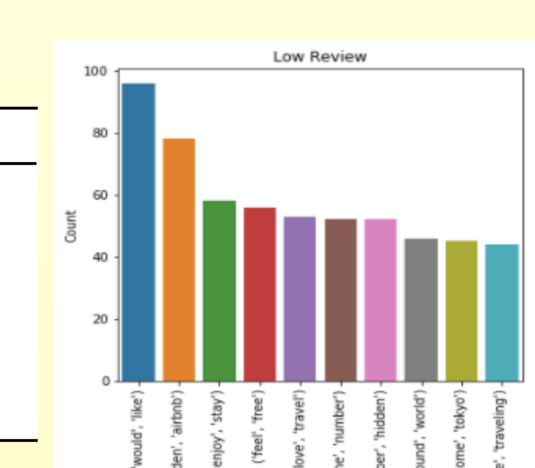
- **Surroundings (low)**:New York, time square
- **Welcome message (high)**: feel free
- **Resonance (high)**:love travel



Tokyo



- **Welcome message (more frequent than the other two cities):** welcome Tokyo
- **Resonance**: love travel



Conclusion

- **Cleanliness, communication, location and verification of super host** are the main influencing factors for the rating score
- Three localities display significant differences in terms of factors framing the results("amenities, host and location" debate (Cheng & Jin, 2019))
  London: property
  New York: host
  Tokyo: location
- Successful hosts should strategically arouse resonance in guests

Limitations

The need to analyze more different, or even cross-cultural regions, adding to the literature regarding the convergence/divergence of Airbnb guests' traveling experiences (Brochado et al., 2017).

References

- Cheng, M., & Jin, X. (2019). What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management*, *76*, 58-70.
- Brochado, A., Troilo, M., & Aditya, S. (2017). Airbnb customer experience: evidence of convergence across three countries. *Annals of Tourism Research*, *63*, 210-212.

Acknowledgements

I sincerely thank Dr. Evans and all classmates in MACS 30250-Perspectives on Computational Research in Economics for your helpful comments and suggestions.