★★★★★

Airbnb Analytics

Team 6: Zayn Sui, Sylvie Zhou, Joy Zhu, Mild Trakarnsakdikul

Business Understanding

Airbnb is a community-based, two-sided online marketplace that facilitates and connects **people who want to rent out their home** with **people who are looking for lodging** in a specific location around the world.

Their revenue is from:

- Commission host
  - Everytime guest makes payment, Airbnb takes 10% of the payment amount as commission.
- Guest Transaction fee
  - When travelers make payments for stays, they are charged a 3% fee for the transaction.

Business Question

Superhosts are hosts that provide outstanding hospitality, which means being highly-rated, experienced, reliable, and responsive..

The definition of a superhosts is vague without any qualitative values to base on.

- Explore what features make a superhost, and if there are any features that are weighted more than others.
- Allow us to use the findings to help the host better align themselves to become superhost on Airbnb.

What criteria can be used to more effectively identify and qualify Airbnb hosts as "Superhosts"?

Data Understanding

We are exploring Airbnb's United States data. The dataset is separated by the state and city the listing is located in.

For feasibility we decided to focus on the West Coast data:

- 131,388 observation
  - 45,595 superhost (35%)
  - 85,793 normal host (65%)
- 58 features
  - Property and host features
- 6 States
  - California, Colorado, Hawaii, Nevada, Oregon, Washington
- 14 Cities

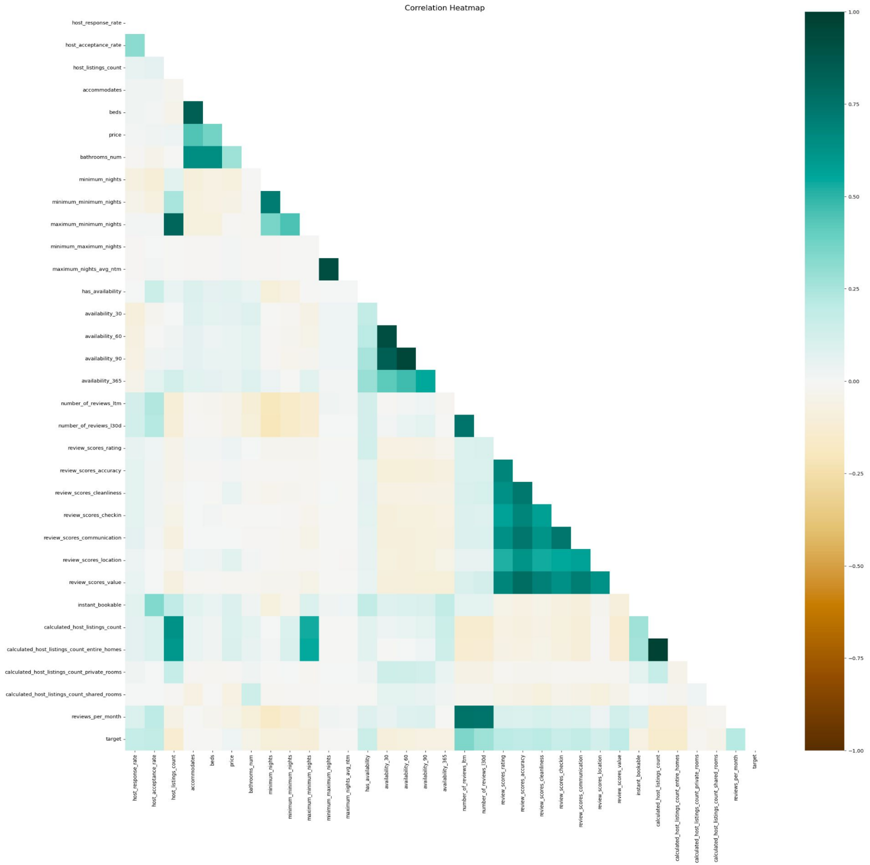
Source: http://insideairbnb.com/get-the-data/

★★★★★

Data Preparation

Numerical & Categorical Features Preprocessing

- Drop multi correlated variables:
  - maximum_nights_avg_ntm, minimum_maximum_nights, availability_60, calculated_host_listings_count, etc.
- Drop 848 rows with null value in target variable
- Replace nulls with mean for numerical variables



Text Preprocessing

- Combining all text data
  - Property Name, Description, Neighborhood Overview and Host About
- Cleaning the text data
  - Remove all non letters
  - Convert everything to lowercase
  - Remove stopwords and repeated words (br, airbnb, etc.)
  - Convert the words to their stem (paradise -> paradis)

Original Text	Clean Text
Cottage by the Redwoods This is a very private cute cozy small bohemian style retreat next to a creek under the redwood and oak trees. Across the street is a 40 acre park for hiking biking walking frisbee etc. including a dog park<br /><br /><b>The space</b><br />Lovely private setting by the creek, across the street from 40 acre park with trails and playing fields, beach easy walk or bike ride away, shops, restaurants and theater nearby. no traffic very peaceful.<br /><br /><b>Guest access</b><br />There is ample parking in front of the orchard. Bus stop to Santa Cruz or Capitola etc. is a five minute walk down the street.<br /><br /><b> Other things to note</b><br />Recycling and composting as well as minimal use of plastics Are greatly appreciated!	cottag redwood privat cute cozi small bohemian style retreat next creek redwood oak tree across street acr park hike bike walk frisbe etc includ dog park space love privat set creek across street acr park trail play field beach easi walk bike ride away shop restaur theater nearbi traffic peac guest access ampl park front orchard bu stop santa cruz capitola etc five minut walk street thing note recycl compost well minim use plastic greatli appreci quiet neighborhood street light sidewalk near end road littl traffic yet minut freeway health food store restaur minut away easygo environmentalist musician educ progress activ play music garden teach work homestead

★★★★★

Modeling & Evaluation

Precision Evaluation

Predicting a host as a Superhost when they are not (false positives) can lead to disappointment and dissatisfied guest.

- False positives could be costly for Airbnb
  - Guest may not return or pivot to hotels
- Precision priorities the proportion of correctly predicted superhosts among all the positive predictions
  - A high precision score indicates the model making fewer false positives predictions

NLP Modeling

Convert all the text data into numerical vector.

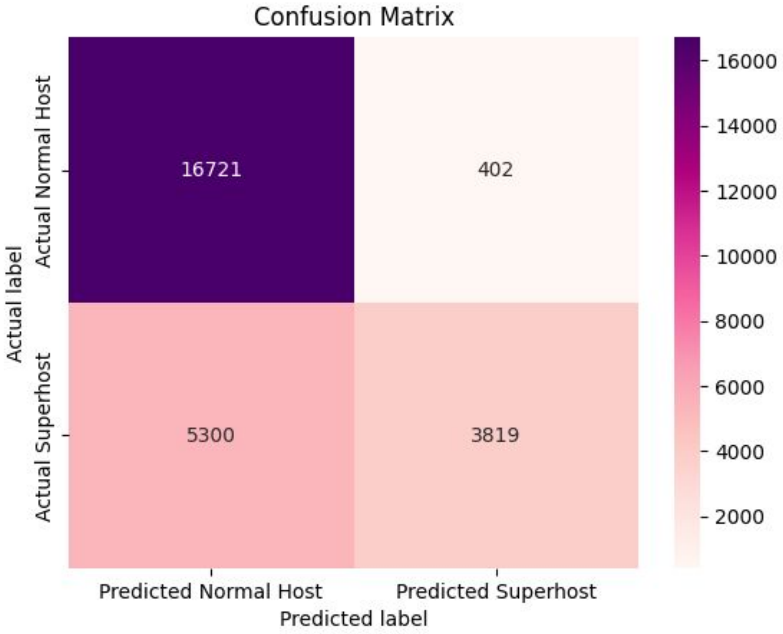
- CountVectorizer
  - Counts the number of times each word appears
- TfidfVectorizer
  - Assigns weights to each words based on its frequency

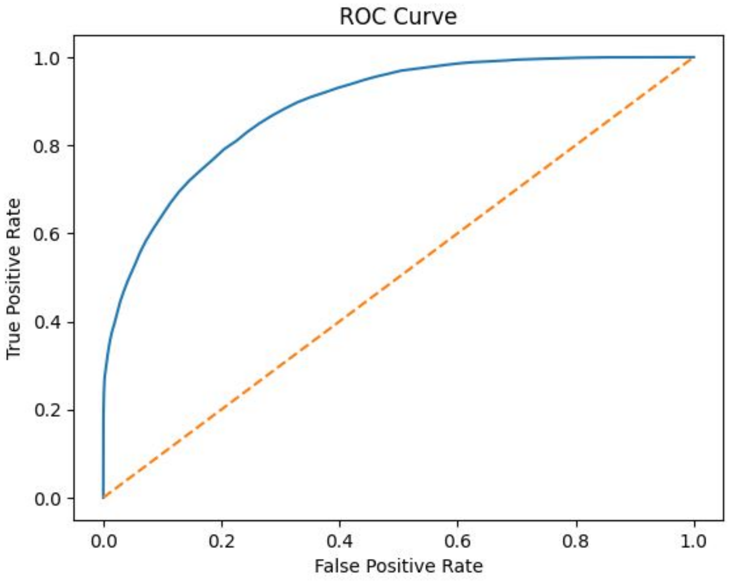
Each Model was train with the CountVectorizer and TfidVectorizer data

- Classification Modeling
  - Logistic regression, Decision Tree Classifier, KNN and Naive Bayes
- Ensemble Methods
  - CatBoost and Random Forest Classifier

Best NLP Model

Random Forest Classifier gave us the best **precision of 0.9** and accuracy of 0.78

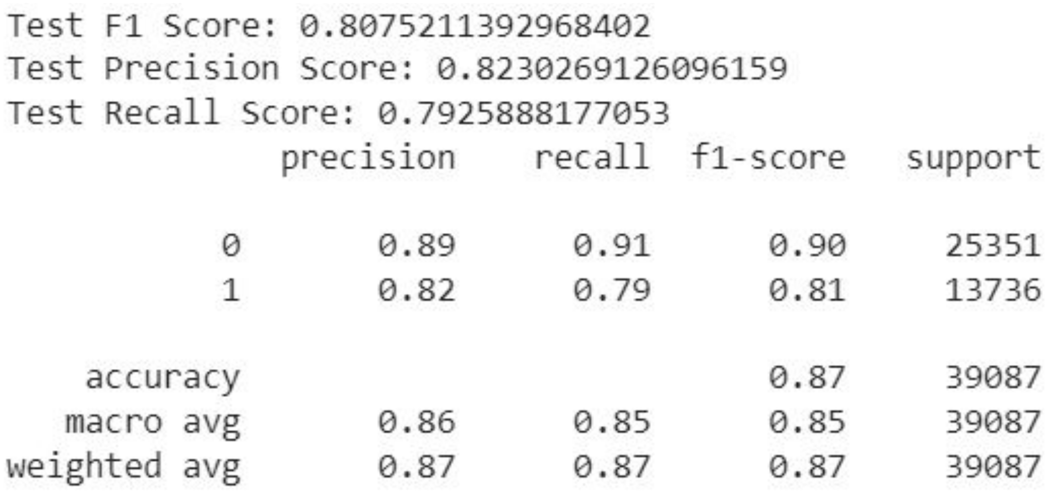




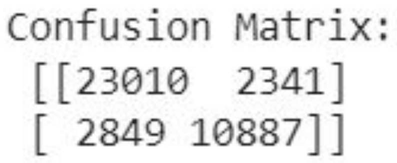
AUC: 0.70

Classification Modeling

- Modeling: Decision Tree, Logistic Regression, Naïve Bayes, **CatBoost**
- CatBoost precision score: 0.82

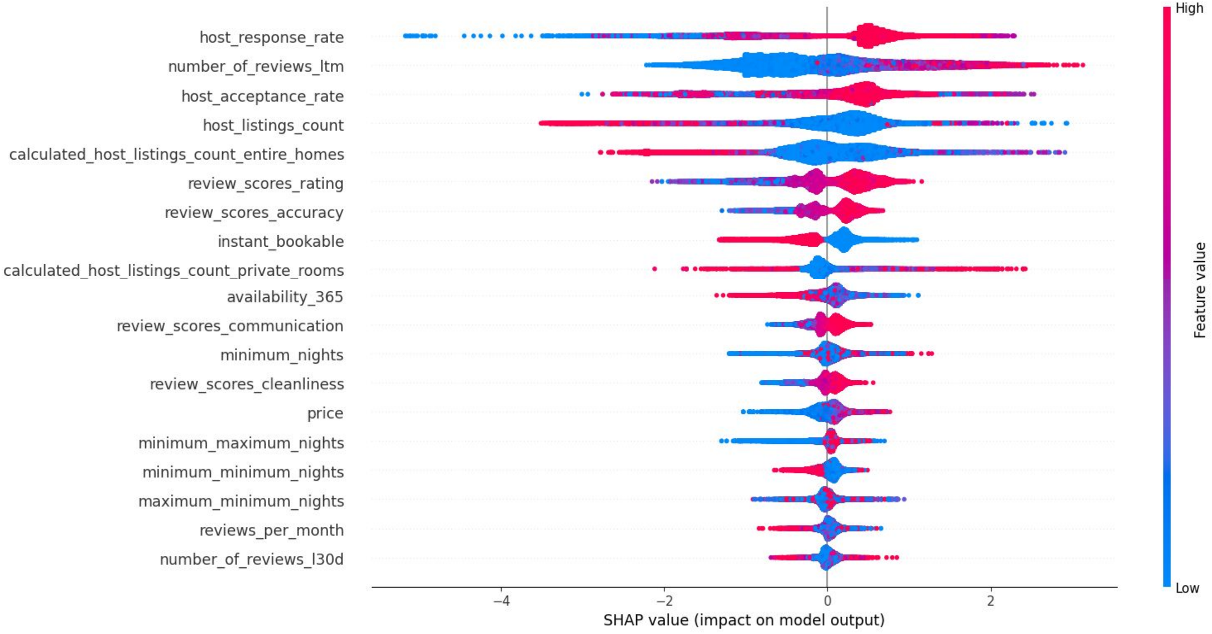


```
Test F1 Score: 0.8075211392968402
Test Precision Score: 0.8230269126096159
Test Recall Score: 0.7925888177053
              precision    recall  f1-score   support

           0       0.89      0.91      0.90     25351
           1       0.82      0.79      0.81     13736

    accuracy                           0.87     39087
   macro avg       0.86      0.85      0.85     39087
weighted avg       0.87      0.87      0.87     39087
```

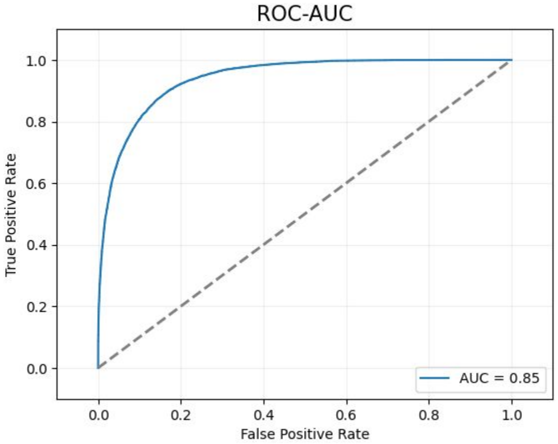


```
Confusion Matrix:
 [[23010  2341]
 [ 2849 10887]]
```

Model Evaluation



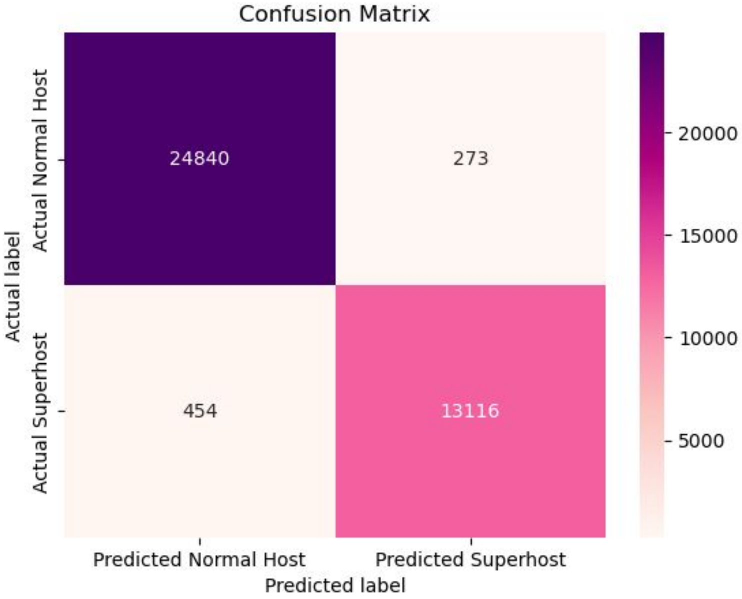
AUC score: 0.94



Stacking Models

Combine prediction from descriptive text and prediction from other features.

- **Independent features:**
  superhost rate predicted by NLP,
  superhost rate predicted by Catboost
- **Dependent feature:**
  actual superhost rate
- **Model:** Logistic Regression
- **Performance:** F1 0.97; Precision 0.98
- **Coefficient:** [14.86, 5.68]
  NLP prediction plays a key role in superhost detecting.



★★★★★

Business Application

Business Application - Criteria

Through our analysis we will be able to help Airbnb create a clear criteria for superhost and for host to align themselves better .

This can be achieved through:

- Improving host response rate – response in shorter time when guests have questions and requests
- Reviewing each request before accepting the booking
- Providing less listing homes/stays and make each of them perfect
- Design a user-friendly review page
    - Encourage positive reviews by highlighting benefits
    - Offer incentives to increase number of reviews

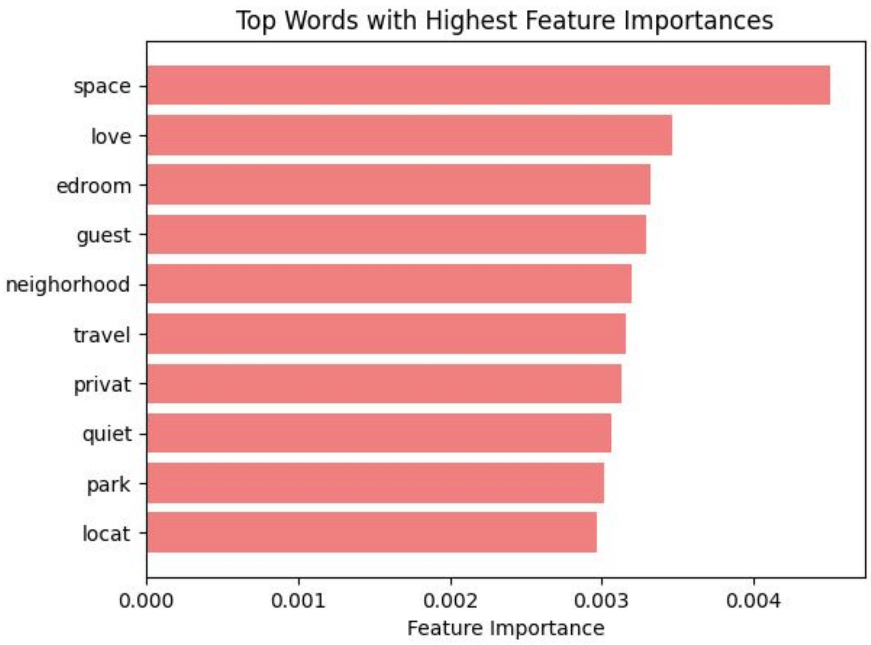
Airbnb can leverage the information and business strategies to maximize their superhost growth and revenue

Business Application - Description

Words that host should considers:

- Words about the location/city
- Words regarding space
- Words about the environment
- Words about the surrounding area
- Guest centric writing

Airbnb hosts can leverage the information to optimized the way they name and describe their property to increase the chances of being booked and becoming a superhost



Issues and Ethical Considerations

When deploying this predictive model, there are some risk and ethical consideration Airbnb should take into account of:

Data Privacy

- Model should not violate guest or host data privacy

Model Transparency

- Fair prediction with no use of identifying data (race, gender, religion) to reduce discrimination

Effect from the Model

- Do not unfairly penalized host without consulting with domain knowledge

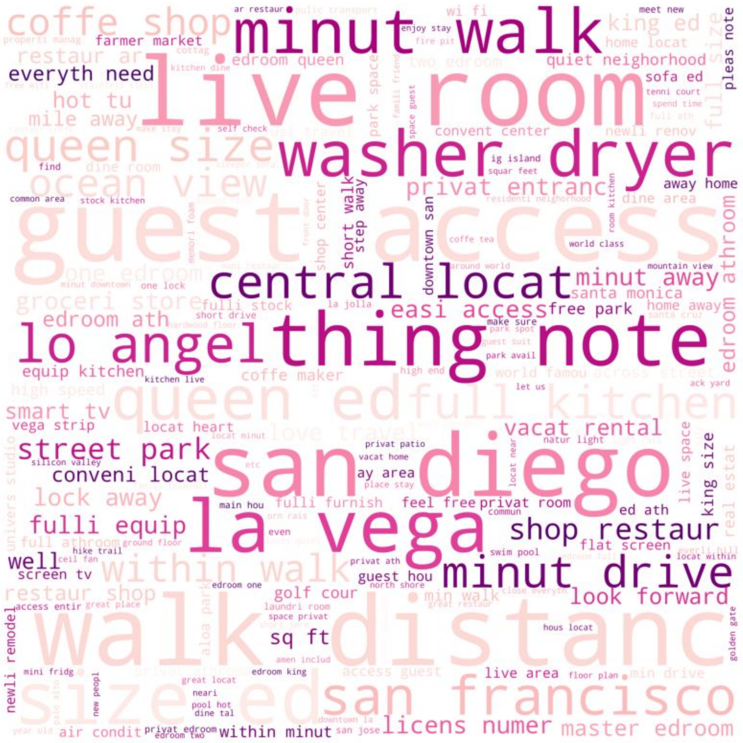
THANK YOU!

Any Question?

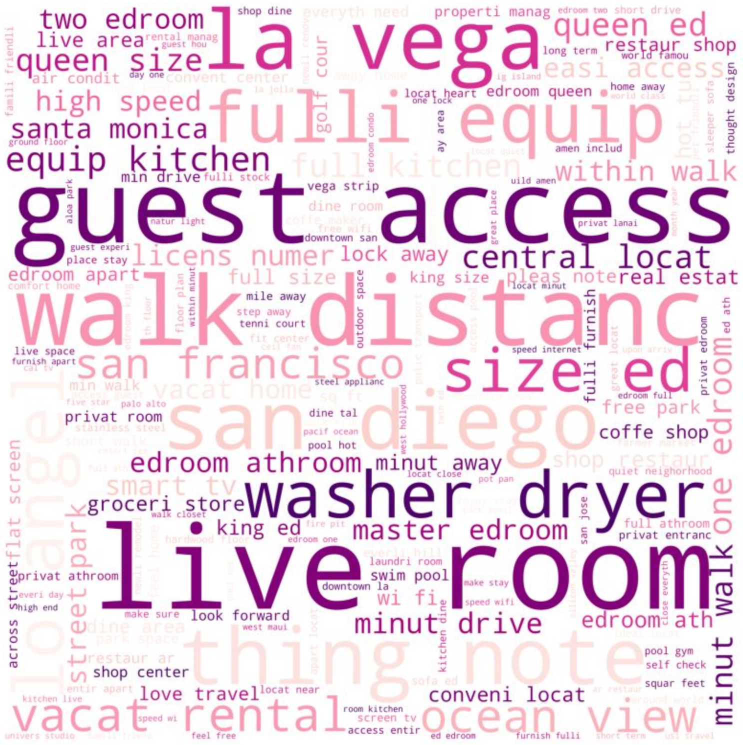
Appendix

Word Frequencies

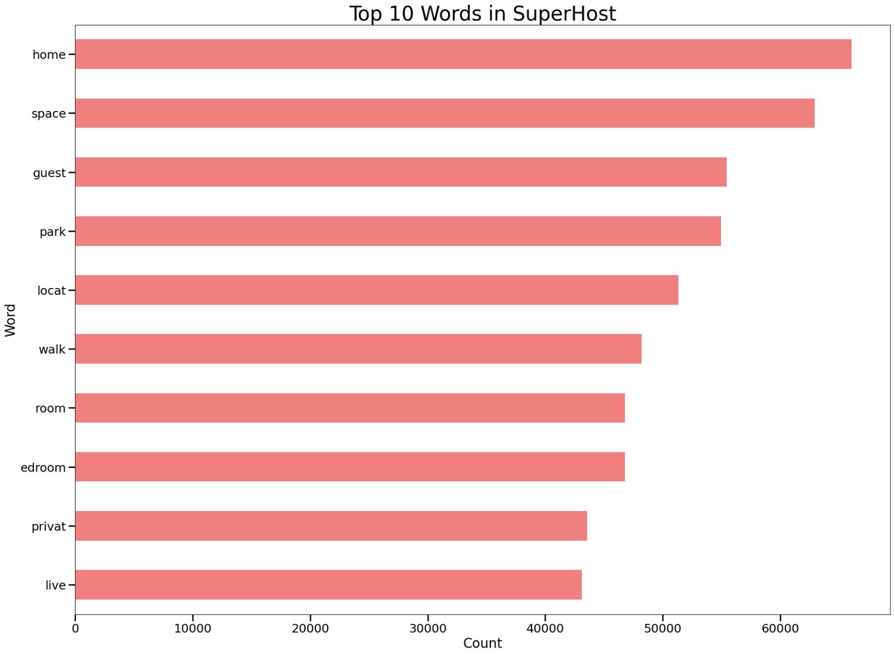
Superhost Description

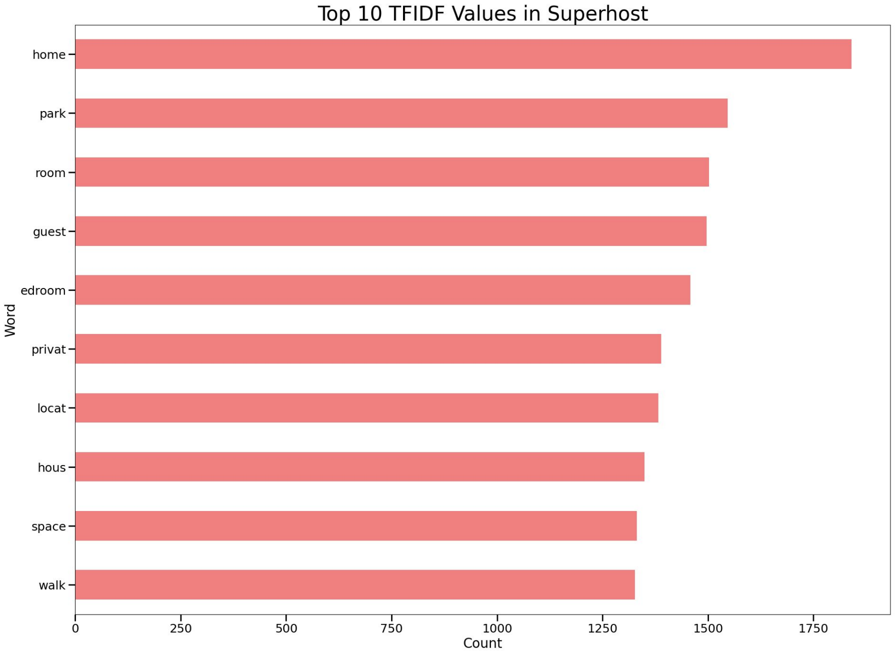


Normal Host Description



Top Words from Superhost


Top 10 Words in SuperHost


Top 10 TFIDF Values in Superhost

Future Improvement

Utilized other region and states data

- Include east coast data for comparison
- Explore other region trends

Explore Reviews data for NLP

- Guest reviews might have more impact on becoming a superhost
- Hotels can improve their amenities and services based on customer needs

Combine hotel booking information with customer information

- Explore the preference of customers in different segments
- Deliver more specific recommendation to hotel hosts (e.g. how to attract their targeted customers)