How to be a Data-Driven AirBnb Host

Utilizing natural language processing, Naive Bayes classification, and logistic regression to identify what makes an AirBnb successful

Evan A. Walker

Problem - What can I do to boost rentals and profitability as an AirBNB owner in Boston?

Hypothesis - A combination of different written and quantitative portions of an AirBnb listing are the best predictors for an AirBnb's success.

Assumptions - If a review is left it means someone paid to stay there, so the number of reviews is a good indicator for a listing’s success.

Goals - To identify the best text factor and quantitative factors that contribute to a high number of reviews per month to assist in the marketability of AirBNB

Risks & Limitations - There is not a specific value for profitability or total rentals. Therefore my assumption that rentals per month is a good indicator will be a huge part of this analysis.

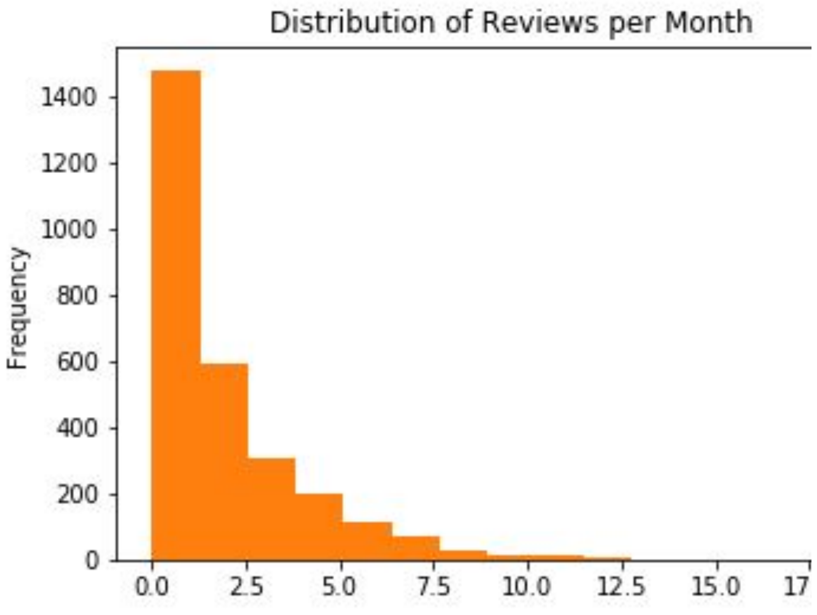
Data - List of **Boston** AirBnb listings that was pulled from AirBnb's "Get the Data" page on 9/7/2016. It can be found on Kaggle here. 95 columns and ~3900 entries.

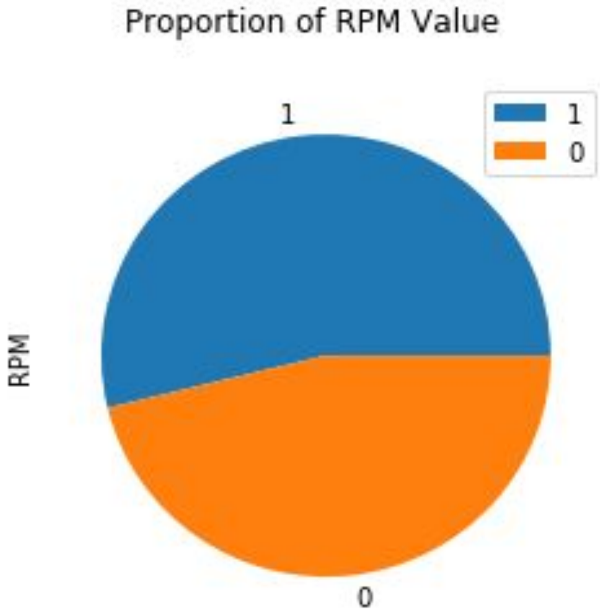
	host_total_listings_count	availability_90	price_bed	Host_Age	RPM
count	2829	2829	2829	2829	2829
mean	42.93	37.78	112.61	945.94	0.54
std	139.90	31.67	64.06	632.26	0.50
min	1	0	11	7	0
25%	1	2	65	474	0
50%	2	37	97.5	811	1
75%	6	65	150	1288	1
max	749	90	525	2857	1

Quantitative EDA

- Removed listings with null values for “reviews_per_month”
- Severe left skew in reviews per month column so I created a new boolean column called “RPM” where less than 1 review per month is a value of “0” and greater than 1 review per month is a value of “1”.
- ~54% True and ~46% False for RPM
- Calculated “Host_Age” column by converting “Host_Since” and “Last_Scraped” columns into datetime, to calculate the host age in days.
- Calculated “price per bed” field to normalize the price column because obviously bigger AirBnb’s are going to be more expensive
- Host total listings, accommodates, availability 90, price per bed, and host age for quantitative features

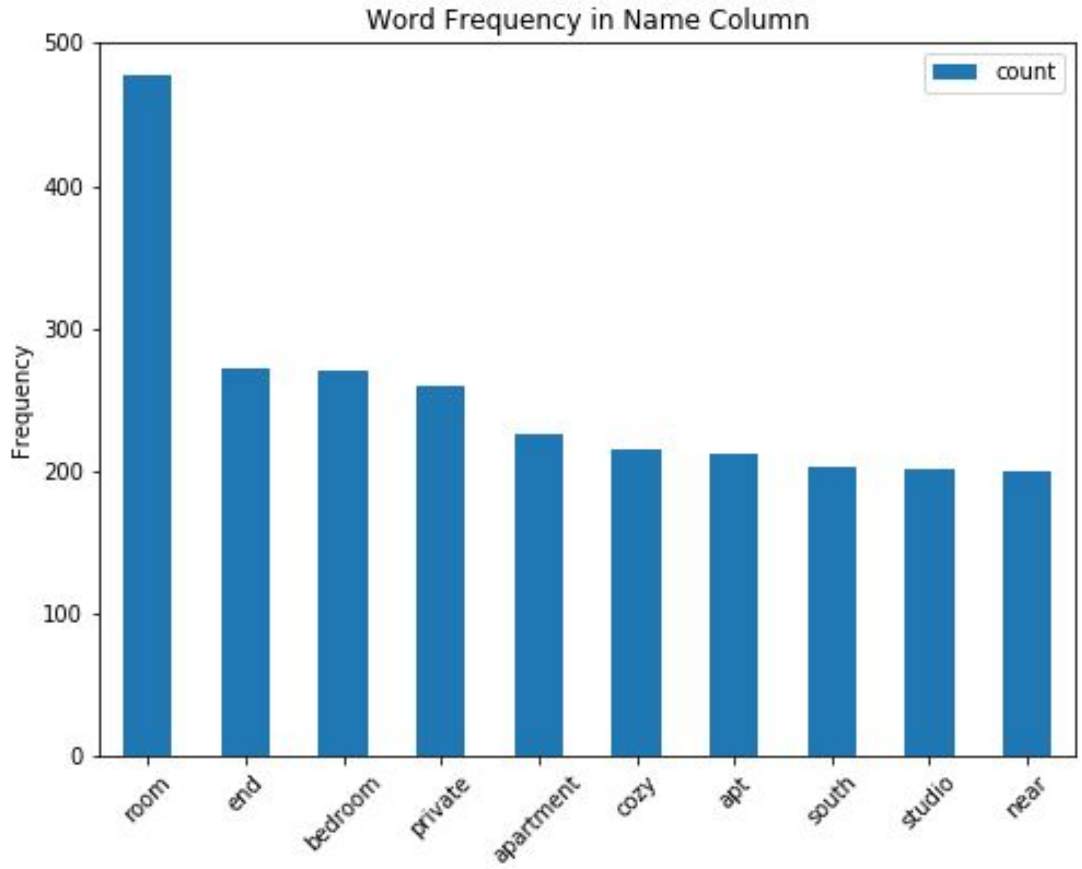
Quantitative EDA

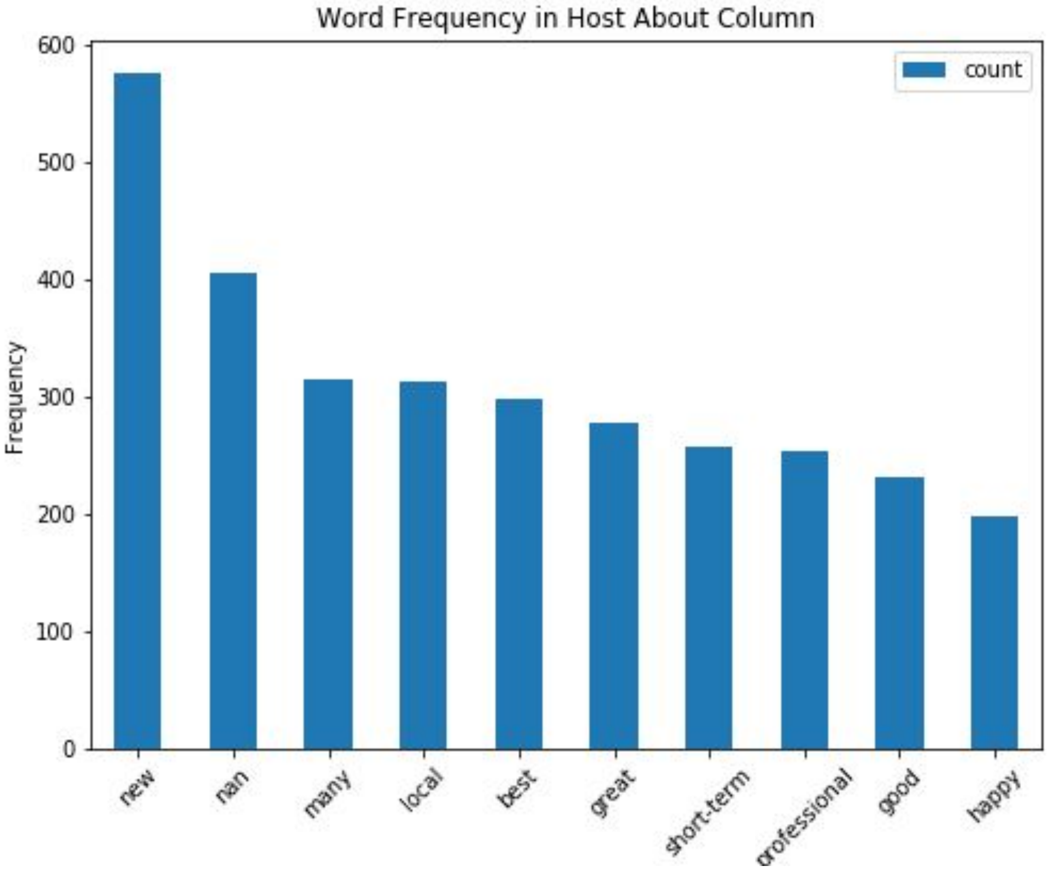


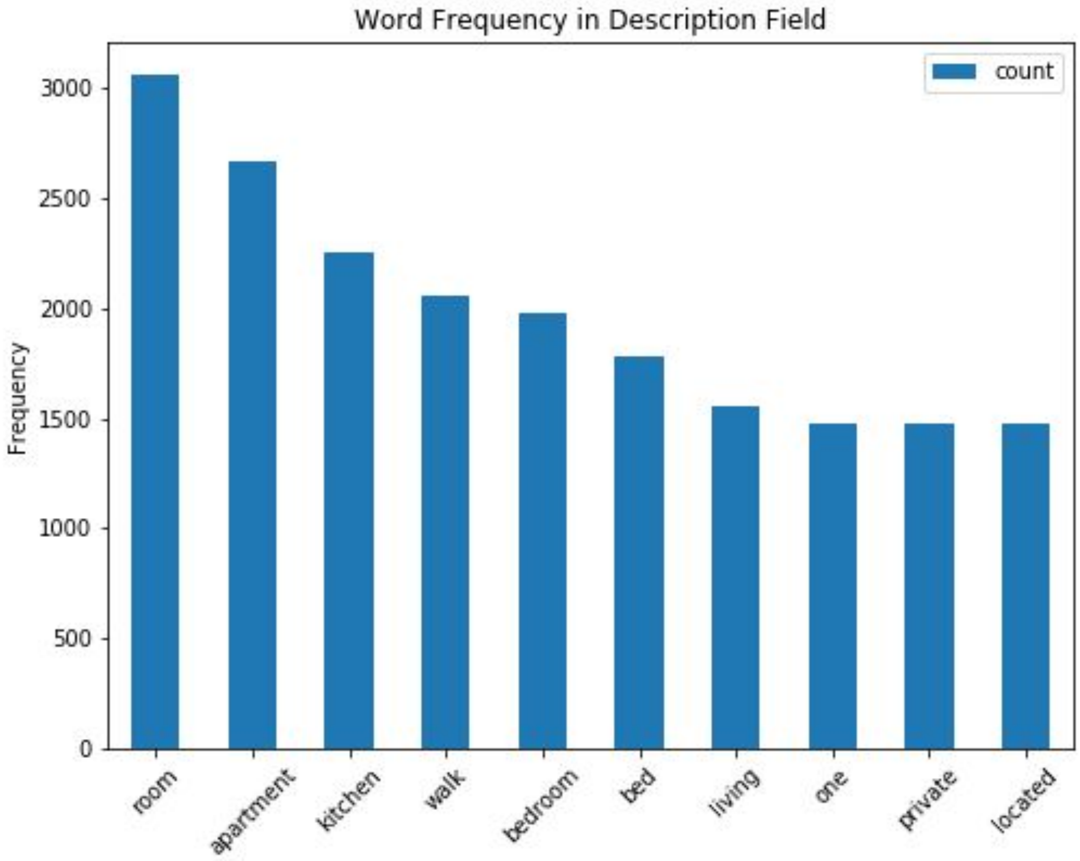


Text EDA

- “Name”, “Description”, and “Host About” were the text columns I chose.
- Used NLP to tokenize words in each text column
- Applied English stop words
- Identified part of speech for each token in the document
- Filtered all parts of speech that weren’t considered adjective, as well as very common words like “Boston”


Word Frequency in Name Column
Frequency
0
100
200
300
400
500
count
room
end
bedroom
private
apartment
cozy
apt
south
studio
near


Word Frequency in Host About Column
count
Frequency
0
100
200
300
400
500
600
new
nan
many
local
best
great
short-term
professional
good
happy


Word Frequency in Description Field
count
Frequency
0
500
1000
1500
2000
2500
3000
room
apartment
kitchen
walk
bedroom
bed
living
one
private
located

Modeling

- **Performed Logistic Regression with the selected quantitative variables.**
  - After several combinations dropping the “Accommodates” field resulted in the best Accuracy score
  - Achieved an accuracy score of **0.65**
  - This is ~12% better than the null model
  - I also ran the same model with Naive Bayes and the accuracy score decreased.
- **Naive Bayes classification with the text columns**
  - Started out with Count Vectorizer - N-grams: 1 & 2, and 10,000 max features
  - Switched to TF/IDF using same N-Gram and max feature parameters which boosted the accuracy score.
  - Using TF/IDF and the description field I achieved an accuracy score of **0.69** this is ~16% better than the null model
- **Combined continuous variables and text variables**
  - Since description was the highest rated text value, and removing accommodates from the continuous variables resulted in the highest accuracy score I used those for the combined model
  - Combined Logistic Regression model resulted in an accuracy score of **0.71**. This is ~18% higher than the null model

Conclusions & Continuation

Conclusions

- Created a model that was **18%** more efficient at predicting the success of AirBnb's in Boston
- The Description field in your AirBnb listing is the best for predicting success.
  - **Recommendation** - If you have a new listing focus on the description portion of the listing.
- Host age, price per bed, host's total listings, 90 day availability, and time being a host are the best continuous variables for predicting success
- There is not a single "buzz word" you can put in your description that will boost rentals.

Continuation & Shortcomings

- Review this study for newer AirBnb data, and post Covid-19 to see how the AirBnb industry has changed
- Try this for other cities to see if it changes
- Significant assumptions with the RPM field
- Nice to have data that has an actual profit per listing field or something similar