Airbnb Descriptions

Using Natural Language Processing on listings

## Is my Airbnb description attracting enough guests?

- Airbnb has listings in over 100,000 cities worldwide

- There are 660,000 listings in the United States alone

- Most Airbnb hosts are normal people who might not know how to write a good listing description

Data

- Source of the data: *Inside Airbnb*
  - An independent project that aggregates and analyzes publicly available Airbnb data

- Only used Seattle data for this model

- Utilized current listings to train a model that compares the words used in the description to the performance of the listing

Target Variable (Listing Success)
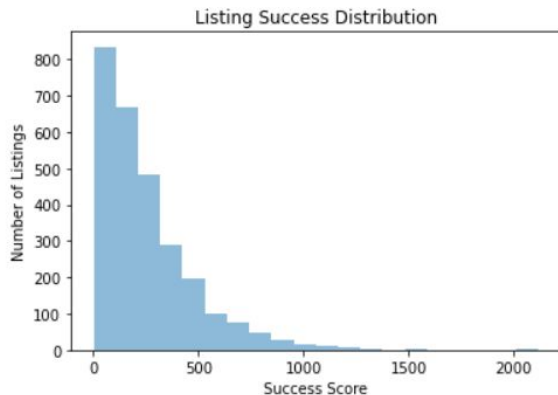
Listing Success:

- Represents the amount of revenue that each listing makes

- Price per night x Reviews per month

Model Preparation

What data is left out of the model?

- Luxury listings (anything over $400/night) **$$$**

- Abandoned listings (no reviews from the past year)

- Hotel listings (some hotels list on Airbnb as well)

So what data is left?

- Still ~3500 rows of data remain

- Transformed some data from strings to floats

- Performed text analysis on the descriptions

Text Analysis

Step 1: Clean and remove punctuation

```
3490      Located in South Park Seattle our AirBNB suite...
3491      Looking for an urban treehouse in Seattle Welc...
```

Step 2: Tokenization

```
3490      [located, in, south, park, seattle, our, airbn...
3491      [looking, for, an, urban, treehouse, in, seatt...
```

Step 3: Remove Stop Words

```
3490      [located, south, park, seattle, airbnb, suite,...
3491      [looking, urban, treehouse, seattle, welcome, ...
```

Step 4: Lemmatization

```
3490      locate   south park seattle airbnb suite featur...
3491      look     urban treehouse seattle welcome trend ...
```

Model Results

Using word2vec we can use the lemmatized words to learn word associations and match them to high performing or low performing listings

$ *Exciting description = profit*
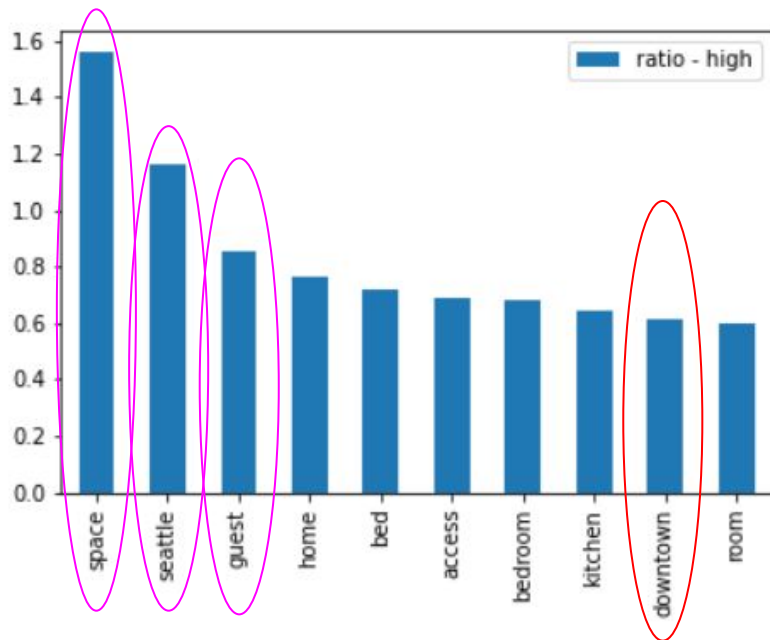
Boring description = empty room $

```
---------------------Gradient Boosting Scores--------------------
Training set score: 0.8071127185051236

Test set score: 0.7497741644083108
```
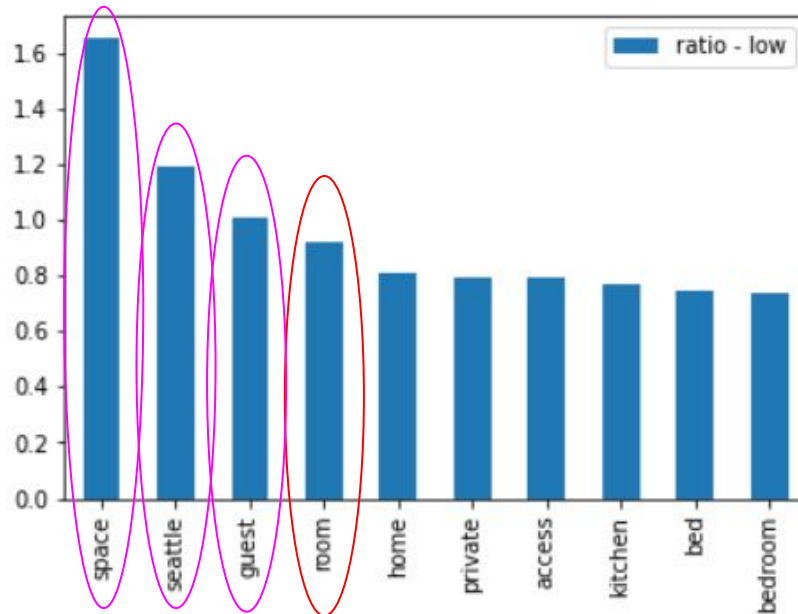
Takeaways

Best Words

Worst Words
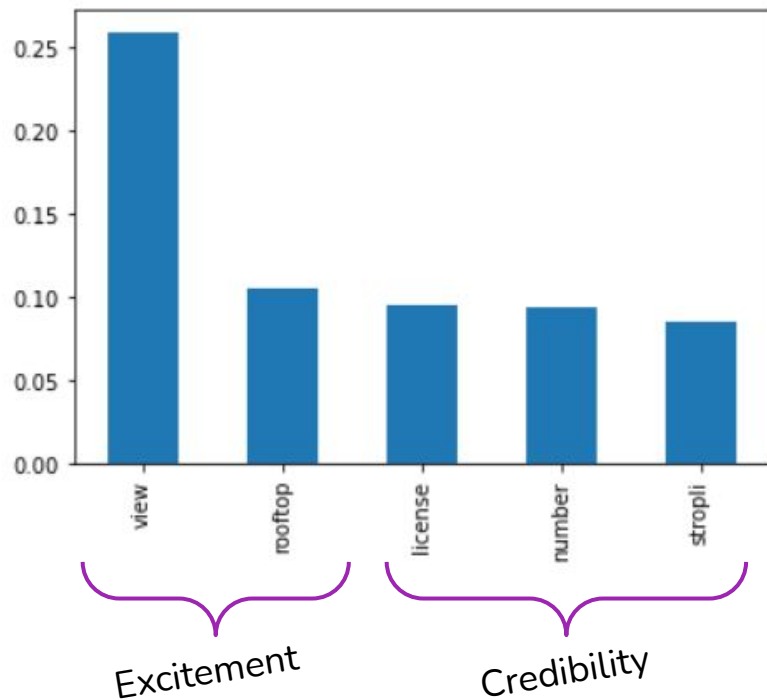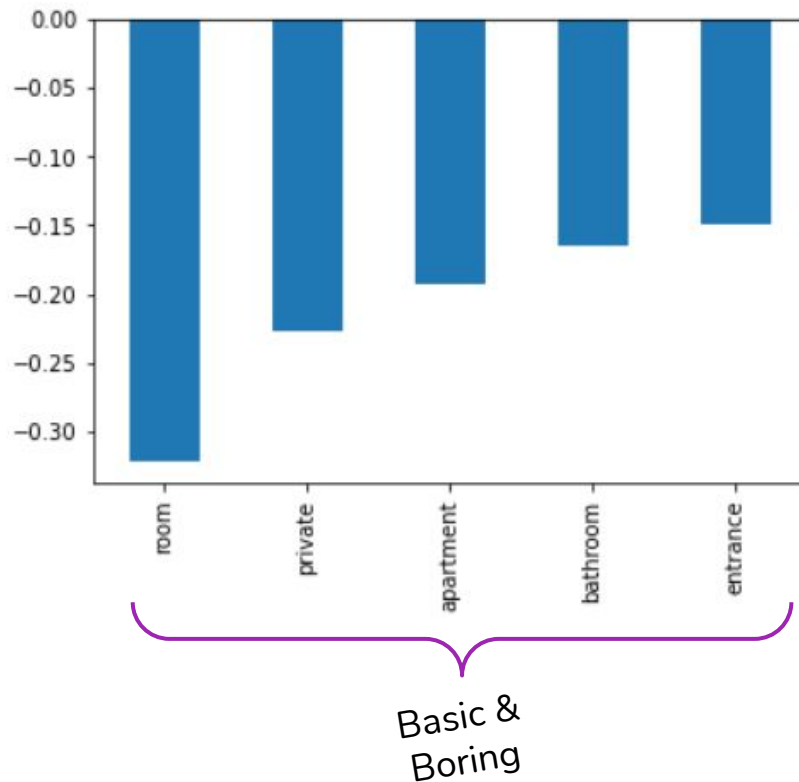
Some interesting takeaways here

Also some very not-so-interesting takeaways

Takeaways - Take 2

Best Words

Worst Words

Next Steps

How can the model be improved?

- Add in data from other cities

- Analyze other sources of text from the listings (Name, Amenities, Reviews)

- Perform sentiment analysis instead of relying on individual words

Questions