## AirBNB Pricing Prediction

**For AirBnB, there are two main options from where the features will be extracted:**

1. Text features
2. Image features

**Use a classifier method to group:**

1. Neighborhood
2. Listing price based on Neighborhood

**Determining Listing price:**

**Basic features from the neighborhood:**

1. Prices
2. Amenities
3. Area of the house with # of beds/baths
4. Local Transit nearby
5. Host bio ??
6. Image thumbnail

**Apart from the basic ones, more significant features could be:**

1. Number of reviews
2. Host response rate
3. Number of references

**Neighborhood Classification:**

1. Crime rate
2. Population
3. Culture, night-life??
4. Facebook/Uber check-ins to see

**Dynamic pricing tool consideration??**

1. Already a tool called Aerosolve. 🡪 Can we come up with a substitute and not a complicated one

**Dataset:**

1. <http://data.beta.nyc/dataset/inside-airbnb-data/resource/9d64399b-36d6-40a9-b0bb-f26ae0d9c53f?view_id=33b9a800-4ed6-4d41-8f87-494c6c8582eb>
2. For considering holiday seasons

http://data.beta.nyc/dataset/insideairbnbdata/resource/ce0cbf4683f9414a8a1d7fd5321d83ca

1. For text analysis: <http://data.beta.nyc/dataset/insideairbnbdata/resource/8115833e8a0e4af68aed4d96a0ae0b73>

**Feature Extraction:**

*For pricing prediction, can we use only #1 or #2 can be helpful?*

*For neighborhood prediction, we might use 2,3,4 below:*

1. **Listing info** 🡪Take the data and fill the missing values if needed. This info could be whether the house is apartment, condo, dorm, # of bed, charge/guest/night
2. **Bag of words** 🡪 Info like Summary of listing, space, description, experiences offered 🡪 Which technique can be used here?

Word-class: In paper, 9-word class were chosen people, nightlife, activities, style, accessibility, culture, nature, amenities, and comfort – might not be needed

1. **Text sentiment features** 🡪 TextBlob package, which calculates the polarity of a segment of text by averaging the polarity of each word in the text included in the package’s lexicon
2. **Visual features** 🡪 download all listing images and extract visual features.

Speeded Up Robust Features (SURF) descriptors from the 100 images using OpenCV

**Analysis and Prediction**:

1. KNN classifier can be used to neighborhood detection. We can remove few neighborhoods with fewer listings. Determining optimal value of K will be important. We have always run K-classfier and checked its R2 value with various value of K. According to paper, sklearn’s Recursive Feature Elimination (RFE) was used
2. Regression Technique like Lasso or Ridge will be useful to find significant features while also on bag of words to see which are good enough for deductions.
3. Train, Test using any of the CV method 🡪 Bootstrapping for better estimation of BETAs.

Also is R2, AIC, BIC only the measure for good fit?

1. As of now, polynomial or in general Linear regression technique but if any better algorithm is taught in the class then we can use that.
2. Correlation between neighborhood and pricing will depend on the area of the house located. How will we consider the intersection?