Airbnb

Dylan Abeyratne
Stephen Acurso
MinJae Song
CIS 4680

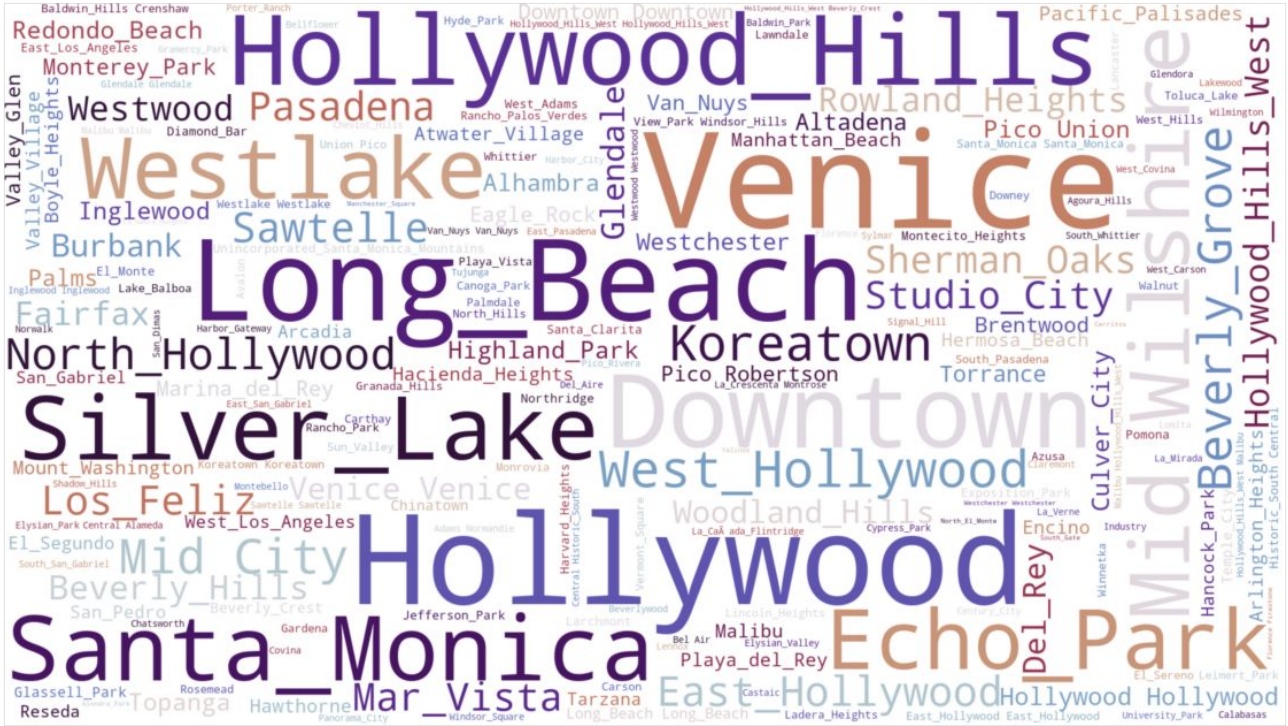
Intro

- Airbnb started August 2008, San Francisco, CA
- Online marketplace for arranging or offering lodging, primarily homestays, or tourism experiences.
- Does not actually own any of the listings instead acts as a broker.
- Goal of the project: Assist new listers raise their overall yield in the future based on several different factors and improve the overall rating of their property.

Data Collection and Variable Description

- Data was collected from http://insideairbnb.com/get-the-data.html
- The time span of the data was one year of listings in Los Angeles County
- Variables of interest included description, location, rating, price, host, neighborhood, room type, and property type

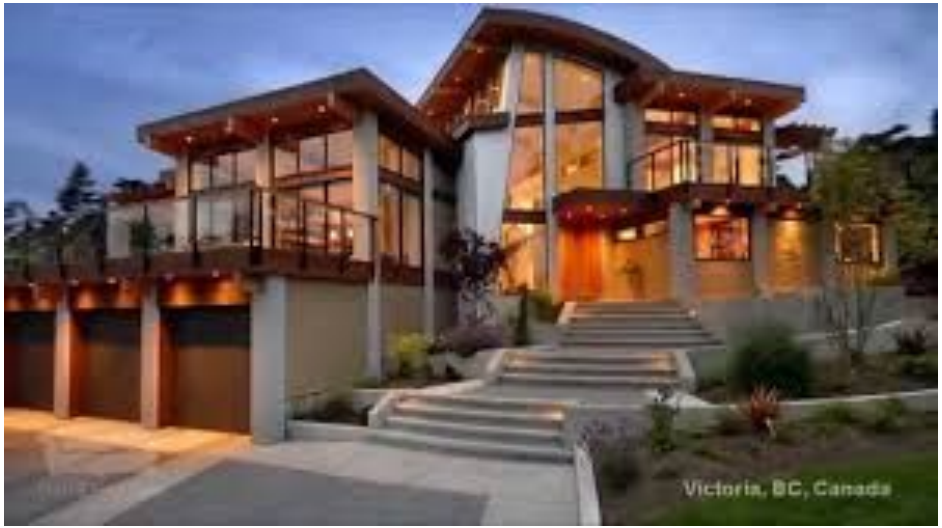
Word Cloud Showing Most Popular Locations



Data Analysis

In order to improve possible yield:

- We analyzed descriptions and pulled the 20 most frequent terms to make sure we include those in our listing
- Analyzed the word cloud to identify most popular locations for rentals



Descriptive Analysis

20 most popular used words in description of rental

- Room
- Private
- Bedroom
- House
- Beach
- Bed
- Kitchen
- Home
- Hollywood
- One

- Bathroom
- Large
- Located
- Full
- Parking
- LA
- Apartment
- Living
- Venice
- Restaurants

Data Analysis

To assist in keeping the model easy to manage and more effective we focused on the following features

- ‘Name’: listing header for text mining
- ‘Description’: listing description for text mining
- ‘Property_type’: apartment, house etc.
- 'room_type': Type of room e.g. private, shared etc
- 'accommodates': No. of people the listing can accommodate.
- 'bathrooms': No. of bathrooms.
- 'bedrooms': No. of bedrooms.
- 'beds': No. of beds.
- 'square_feet': Size of listing.
- 'price': Price of listing.
- 'cleaning_fee': Cleaning fee.
- 'guests_included': No. of guests included in the base price.
- 'extra_people': Price for extra guests not included in the base price.
- 'minimum_nights': Minimum no. of nights required for booking.
- 'availability_365': No. of nights the listing is available for booking.
- 'reviews_per_month': Average no. of reviews listing receives per month. Used to calculate yield.
- 'latitude': Latitude of listing.
- 'longitude': Longitude of listing.

Data Analysis

- Yield is a measure we used as our target variable to calculate a listing's future possible earnings
- We used the following variables to calculate yield
  - Average length of stay, Price, Number of reviews, Review rate
- After calculating yield, we needed to get rid of related variables
- Feature Engineering
  - NLP pipeline to create dictionary, matrix
  - LDA topic modelling for the descriptions
- Linear regression r-2 score 0.28
  - Intended for comparison
  - Low score made it a baseline model
- Decision Tree Regression r-2 score 0.32
  - Improvement
- Random Forest Regression r-2 score 0.39
  - Most complex model
  - Can make individual predictions

Summary of Findings

- The highest amount of listings were Hollywood, Long Beach, and Venice.
- Room Type, Longitude/Latitude, Minimum Nights, and Accommodates were the most important features in our predictive model.
- Random Forest Regression - highest r-2 score
- Linear Regression isn't the strongest model but serves as a good baseline model to improve on
    - Relationship between target and variables
- Most users don't focus on the description of the findings when booking, more on the filters and categories.
- Hosts should focus on the important features and check review sentiment analysis.

Implications

- The only legal problem we may come across is with gathering Airbnb Data to increase a particular individual or group of individuals yield.
- A way around this would be a partnership with Airbnb where our services could be a subscription or premium pay for use feature on the platform.
- We can help hosts optimize their properties based on what high rated listings have attributed to them in order for our hosts to maximize profits through higher yield.

Thank You

Questions?