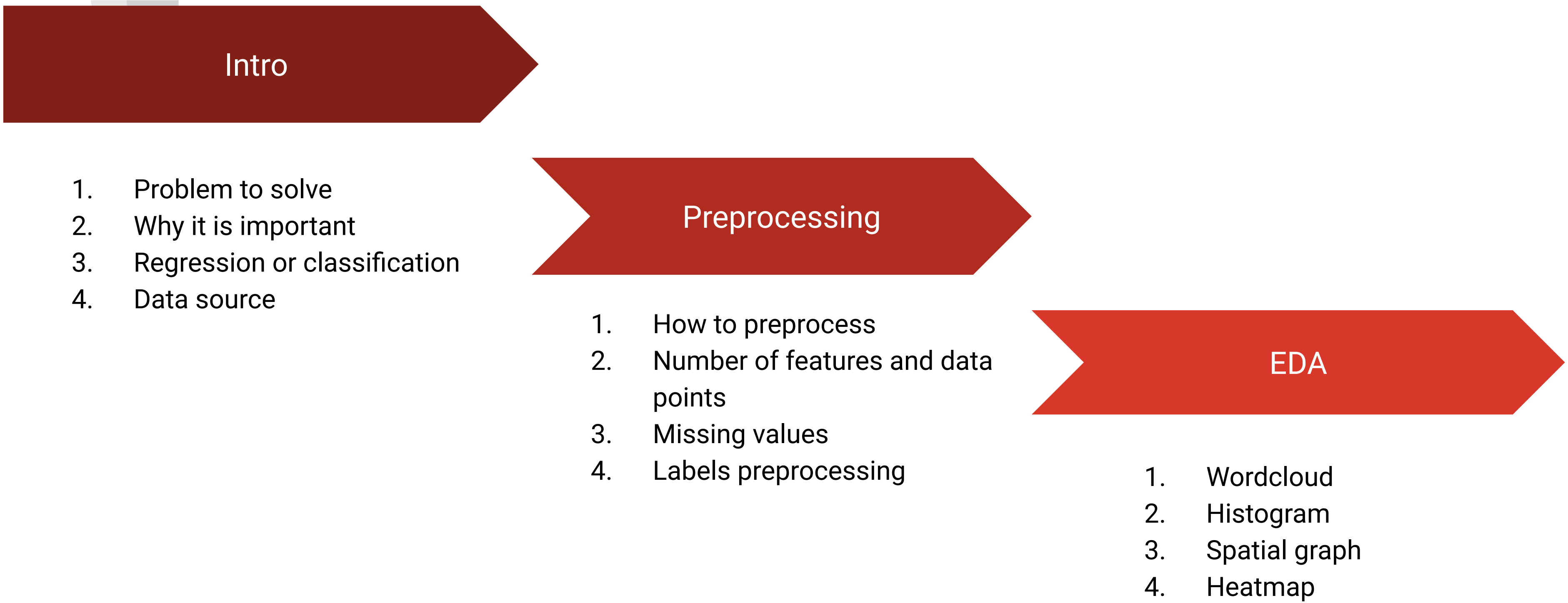
Airbnb in New York

Yuyang Li
Brown University
10/22/2019
github.com/mikelyy/Airbnb-in-New-York



Intro

1. Problem to solve
2. Why it is important
3. Regression or classification
4. Data source

Preprocessing

1. How to preprocess
2. Number of features and data points
3. Missing values
4. Labels preprocessing

EDA

1. Wordcloud
2. Histogram
3. Spatial graph
4. Heatmap

Intro

1. Build different machine learning models to predict the prices of Airbnb in New York.

2. It is important because nowadays more people use airbnb to expand on traveling possibilities. Staying in airbnb gives people more personalized experiences.

3. This is a regression task. The target variable is price.

4. The dataset is from Kaggle. https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/downloads/new-york-city-airbnb-open-data.zip/3#AB_NYC_2019.csv.

Preprocessing

How to Preprocess

1. Drop features id and host_id. For id, it represents listing id which is meaningless. For host_id, it is meaningful because one host could have more than one listing. However, calculated_host_listings_count represents the amount of listing per host. Since there is calculated_host_listings_count in the features, I will drop host_id.
2. Apply OneHotEncoder to categorical features neighbourhood_group, neighbourhood and room_type.
3. Apply MinMaxScaler to continuous features latitude, longitude and availability_365.
4. Apply StandardScaler to minimum_nights, number_of_reviews, reviews_per_month and calculated_host_listings_count, last_review and new features generated by applying NLTK and Ethnicolr.

Number of features and data points

There are in total 48895 rows and 240 features.

Missing values

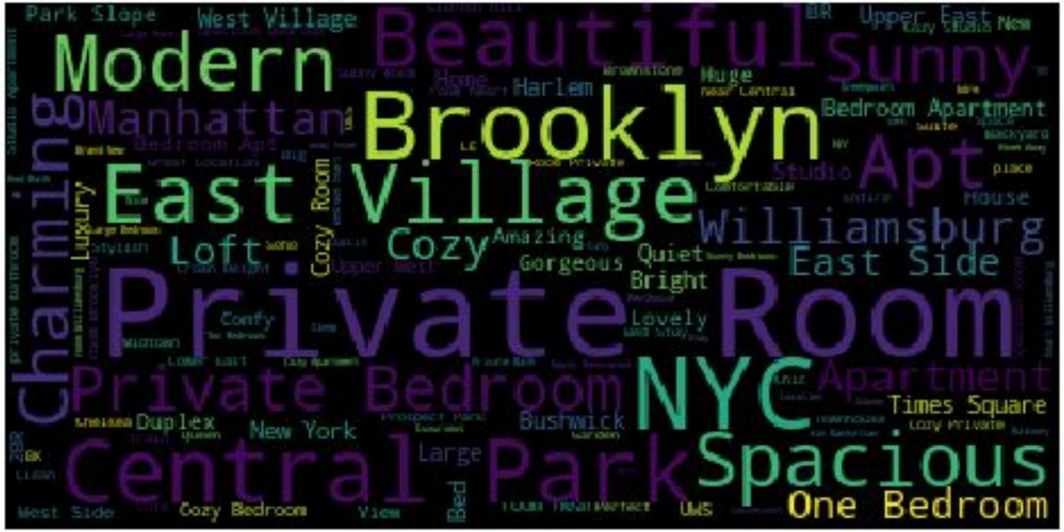
1. name, host_name, last_review and reviews_per_month are the four features which have missing values.
2. For reviews_per_month, I will replace the missing values by 0, because by intuition, if there is no reviews per month, it means that the reviews per month is 0. For last_review, if there is no last review, it means that the date for last review does not exist. I will convert last_review data into numerical value.
3. For name and host_name, I will firstly perform NLTK and Ethnicolr separately to generate new feature columns. After generating new feature columns, I will drop the original features name and host_name, and use the new feature columns instead.

Label preprocessing

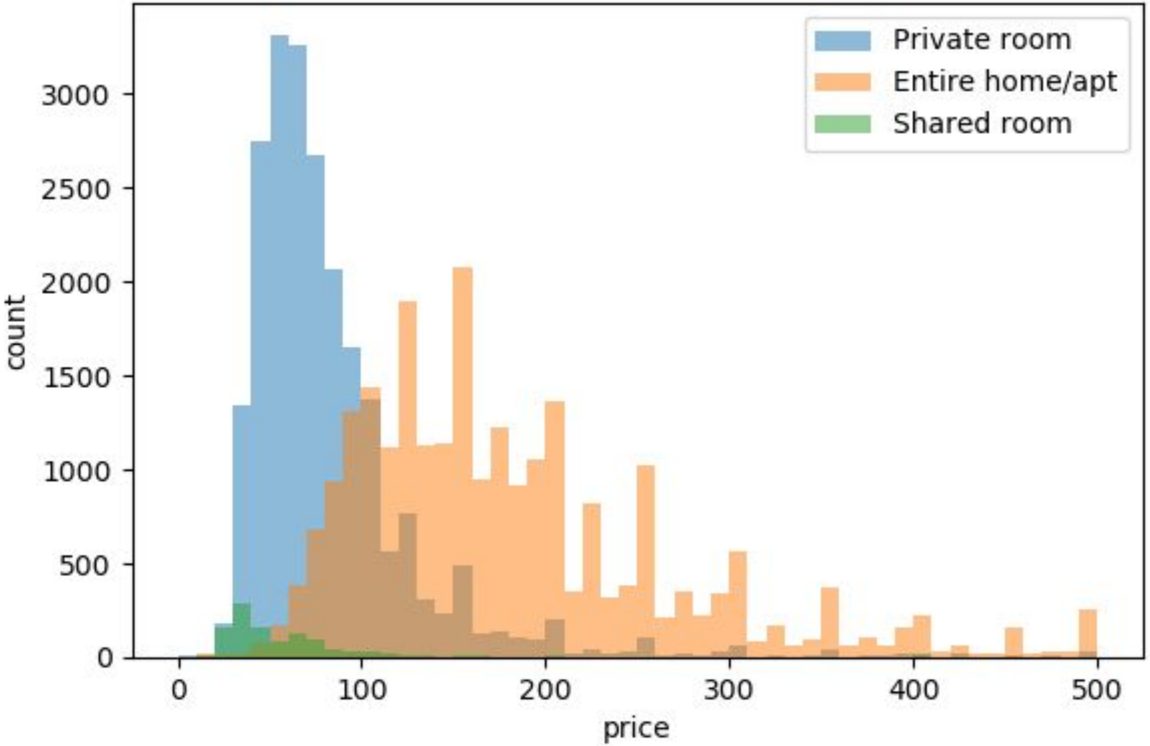
1. For the target variable price, no label processing is applied.

Exploratory Data Analysis

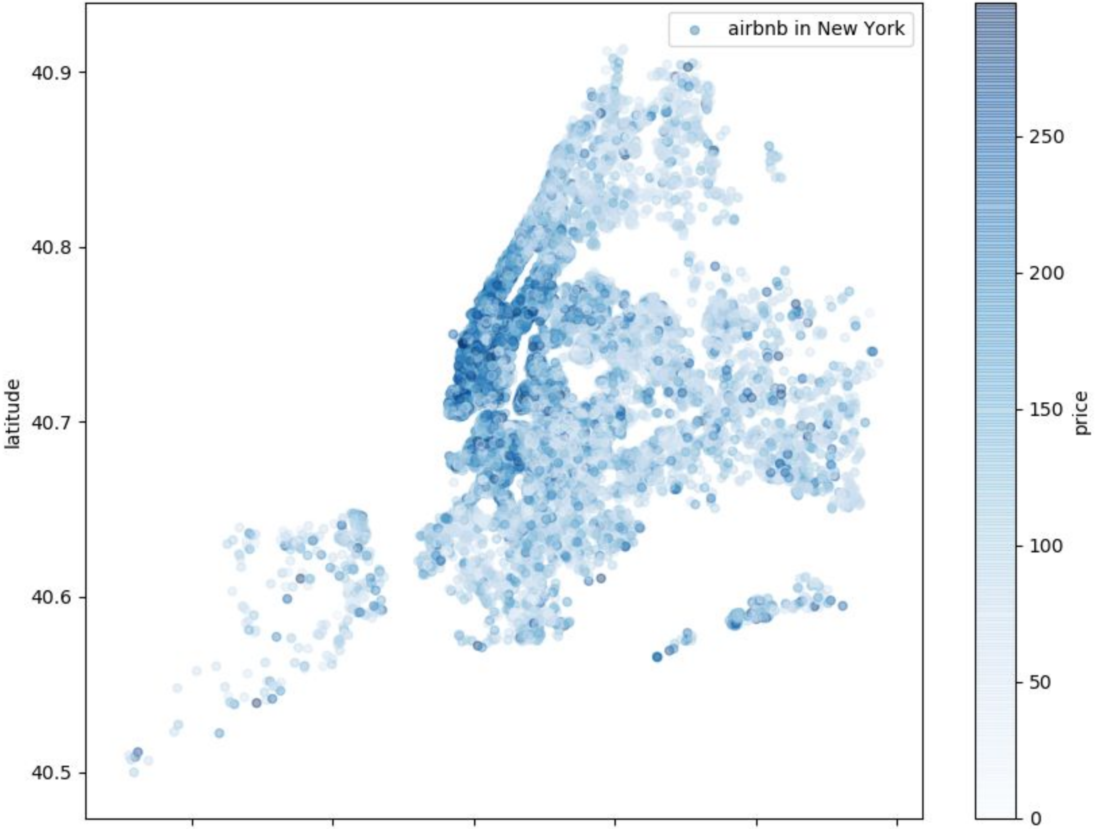
Wordcloud of frequent words appeared in name



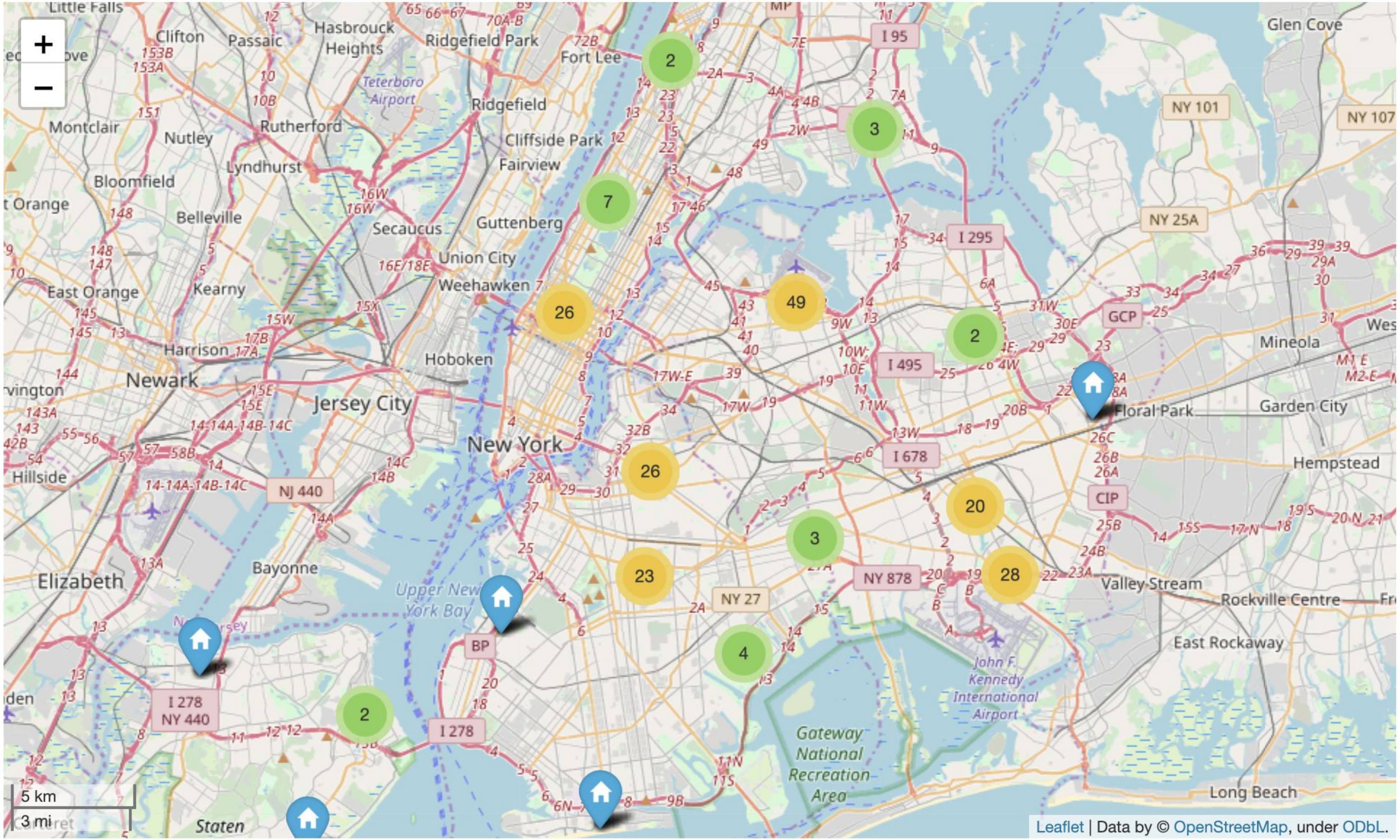
Relationship between different types of room and price



Spatial Graph



Heatmap to show airbnbs with top 200 reviews_per_month



Q and A