New York Airbnb Price Prediction

Hanjun Wei
Data Science Initiative

Oct. 19, 2022

GitHub Repo

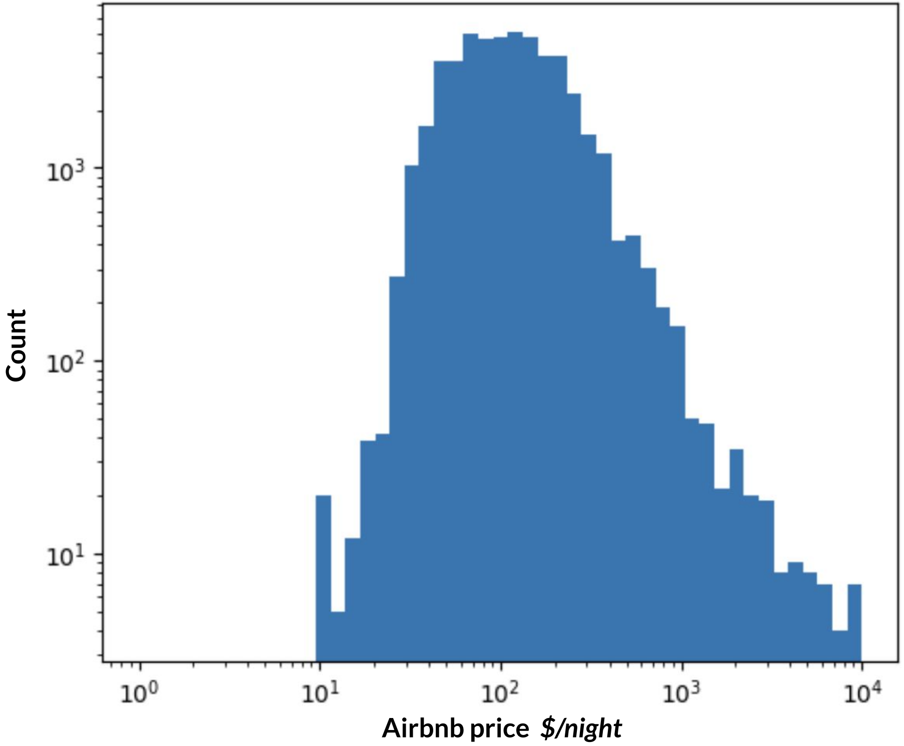
Introduction

- Individual Airbnb hosts in New York might not have efficient pricing strategies due to a lack of market information
- Incorrect pricing might potentially decrease hosts' revenue
- Regression Task: Developing a pricing technique based on New York Airbnb market data
- Data used: New York City Airbnb Open Data in 2019
    - Sourced from *Inside Airbnb*
    - **48,895** Observations
    - **16** Variables

EDA - Price

- Target Variable: *Price*
- Right Skewed
  - Listing price can be very expensive
    - Maximum price: $**10,000**
  - Listing price can not be below $**0**
    - Minimum price: $**0**
    - Drop rows with price = $**0**
- Percentage of extreme outlier: **2.7%**

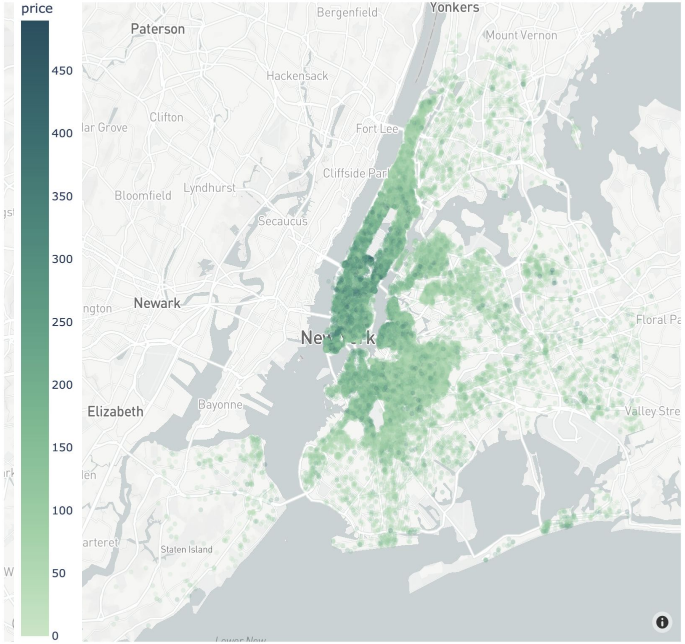
New York City Airbnb Price Distribution



EDA - Price and Location

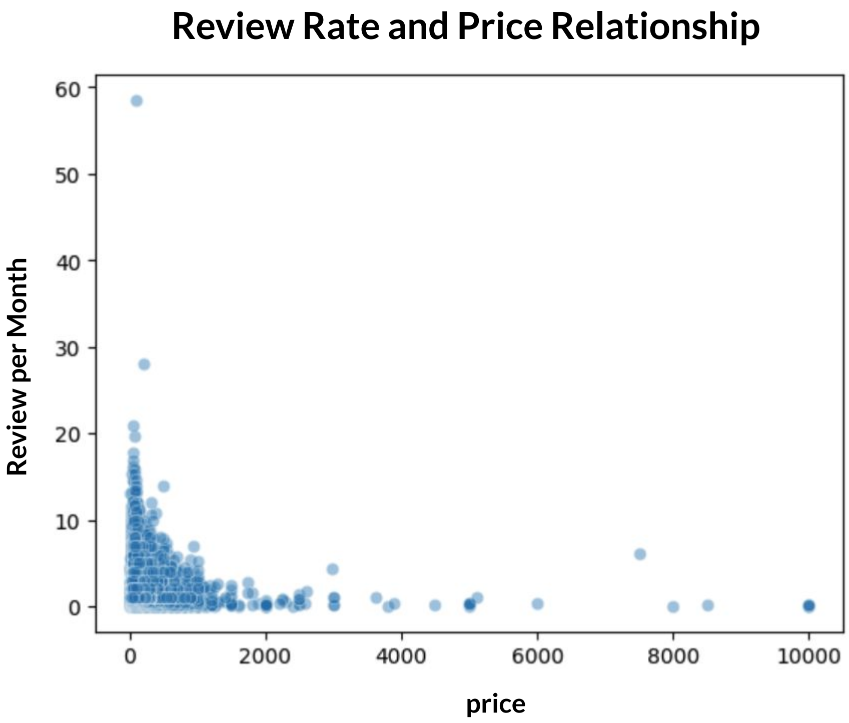
- Ignore the extreme outlier effect
- Darker Shade indicates a higher price
  - High price Airbnb centered at **Manhattan** and **North Brooklyn**
  - Price decreases as we moved away from the center
- Dots (Airbnb) are more crowded in downtown region

Relationship between Price and Location



EDA - Price and Review Rate

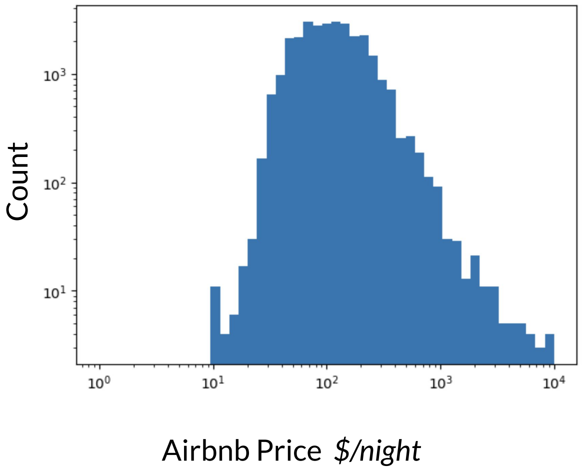
- Most dots (Airbnb) are located in bottom left corner
- Price increases, review rate decreases



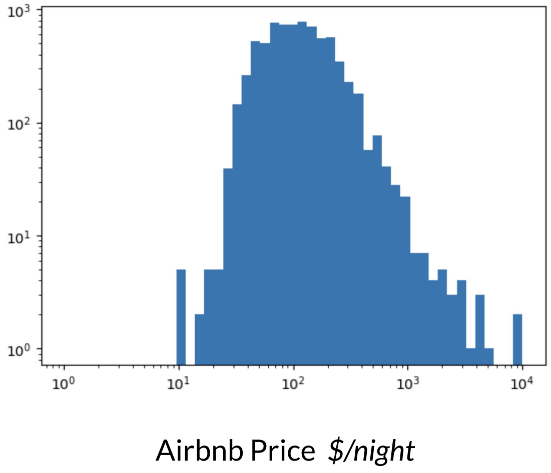
Data Splitting

- Stratification on right skewed y data
  - Train: **70%**, Validation **20%**, Test **10%**
- Overall distribution captured

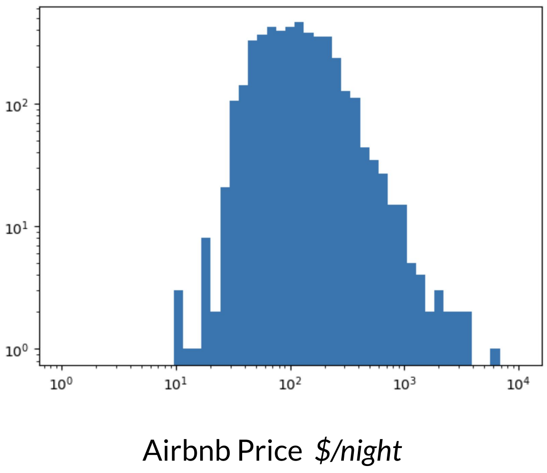
Training Set Price Distribution



Validation Set Price Distribution



Test Set Price Distribution



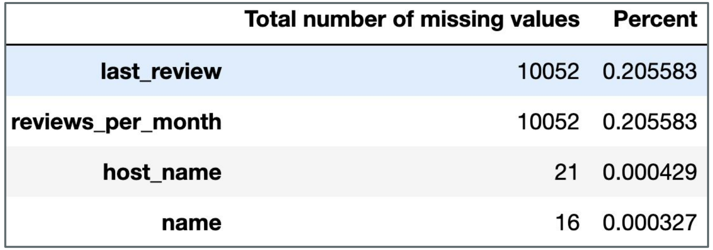
Preprocessing

- 4 types of preprocessor were used:
  - **Ordinal encoder**
    - e.g. room type
      - share room, private room, entire house
  - **One-hot encoder**
    - e.g. neighborhood
      - Manhattan, Brooklyn
  - **Minmax scaler**
    - e.g. availability
      - 0 - 365
  - **Standard scaler**
    - e.g. number of reviews
      - 0 - infinity
- Feature number change for ***X***:
  - Before: **15** columns
  - After: **231** columns

Missing value

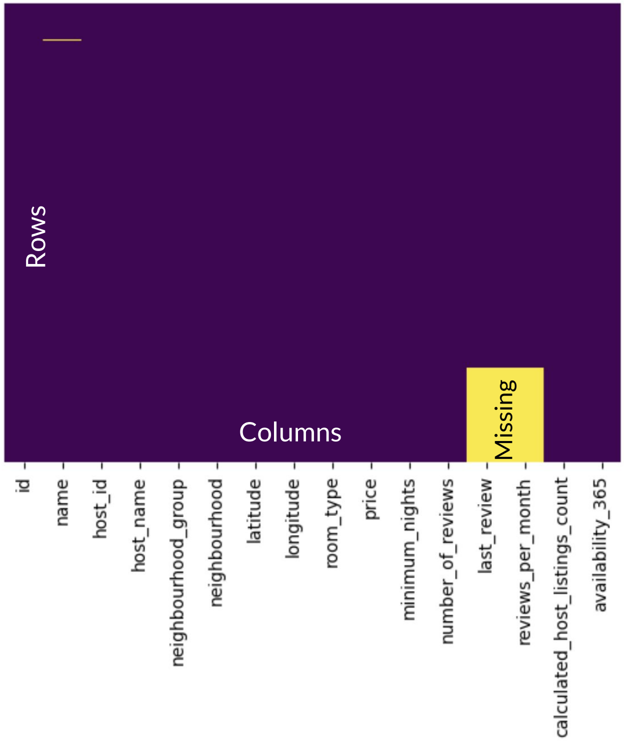
- 4 variables contain missing values:
- ***name*** and ***host_name***
  - ○ Require NLP strategies to extract potential value
  - ○ Drop
- ***last_review*** and ***reviews_per_month***
  - ○ Missing at the same time
  - ○ More advanced method required (after midterm)
  - ○ Keep

Missing Table



	Total number of missing values	Percent
last_review	10052	0.205583
reviews_per_month	10052	0.205583
host_name	21	0.000429
name	16	0.000327

Missing Pattern



Thank You for Your Time

Any questions?