Predicting Airbnb Listing Prices in New York City Using Regression Models

Midterm Presentation

Peter Huang

Brown University

10/22/2019

https://github.com/huang960404/data1030-project

Introduction

- Goal: predict the price of future Airbnb listings as accurately as possible
- Why: Airbnb is becoming more and more popular these days, and customers want to understand the pattern of pricing and what characteristics impact the pricing decision the most.
- Data set: New York City Airbnb Open Data
  - 48,895 observations, 16 variables, 782,320 data points
  - 11 numeric variables, 5 categorical variables
- Source: Kaggle
  - https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data
- Type: regression model
  - Target variable: price

Preprocessing - Missing Values

- id, host_id: dropped

number_of_reviews	last_review	reviews_per_month
0		

- reviews_per_month: set missing values to 0

- last_review
  - Days between dates (reference date: 7/9/2019)
  - Set missing values to 0

Preprocessing - Feature Engineering

- name
  - Length
  - “!” : 1 (contains “!”) or 0 (does not contain “!”)

- host_name
  - Length: 1 (>= 3) or 0 (<= 2)

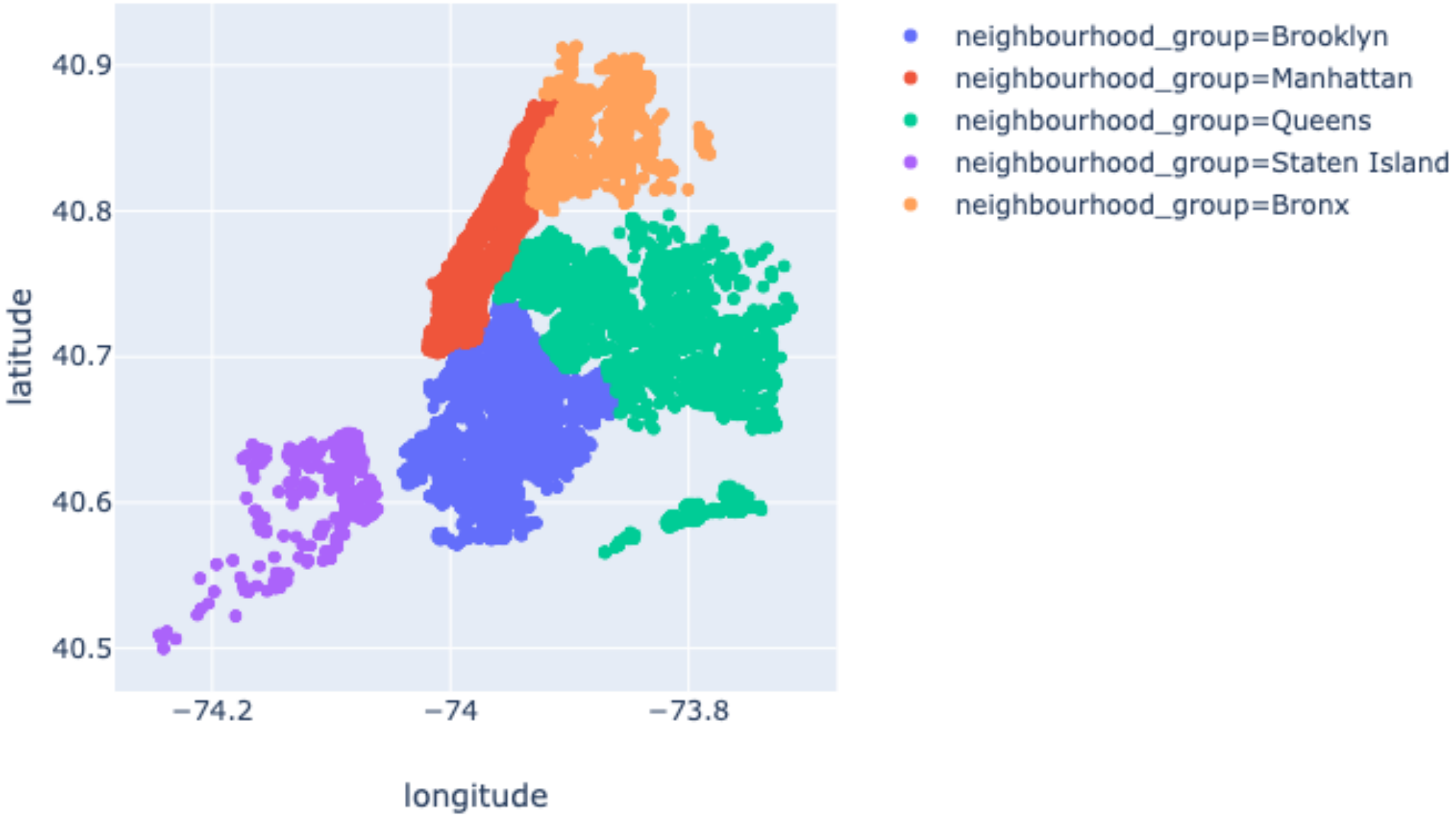
Preprocessing - Encoders

- OneHotEncoder
  - neighbourhood_group: Bronx, Brooklyn, Manhattan, Queens, Staten Island
  - neighbourhood: Kensington, Midtown, Harlem, Hell's Kitchen, Chinatown, etc.
- OrdinalEncoder
  - room_type: 0 (Shared room), 1 (Private room), 2 (Entire home/apt)

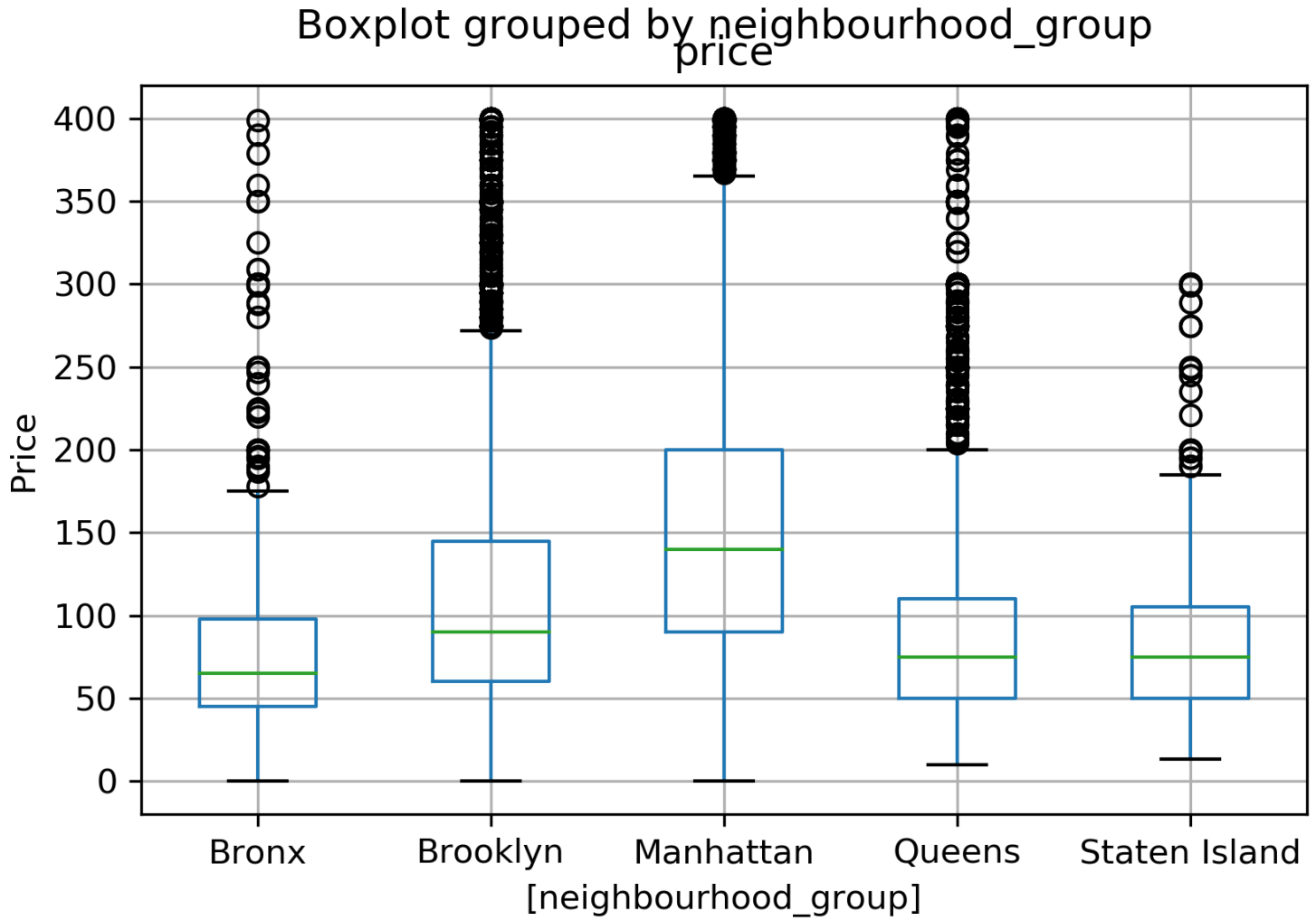
Preprocessing - Scalers

- MinMaxScaler
  - latitude, longitude, availability_365
- StandardScaler
  - name, minimum_nights, number_of_reviews, last_review, reviews_per_month, calculated_host_listings_count
- 239 features, 11,685,905 data points

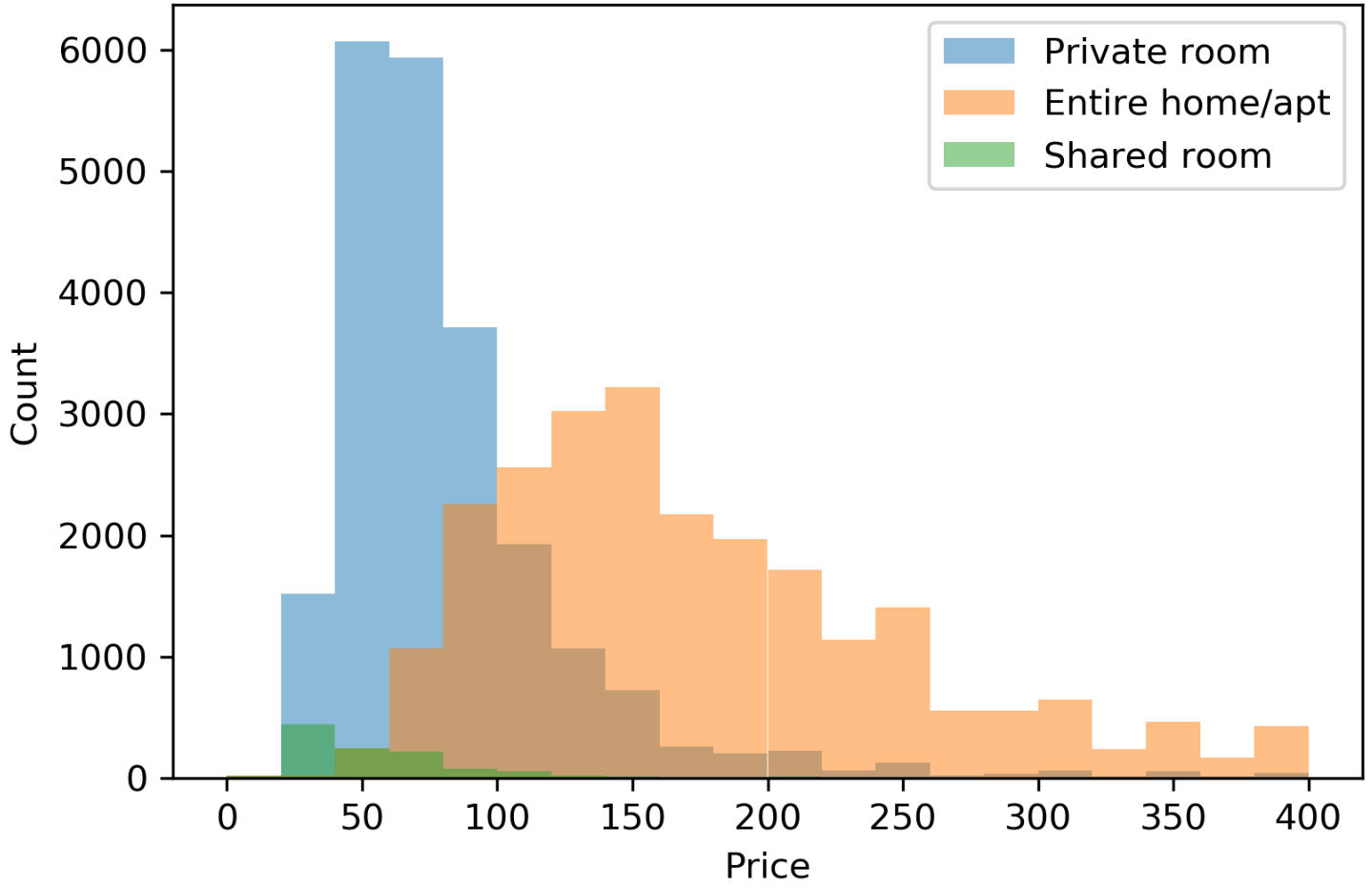
Exploratory Data Analysis - Map



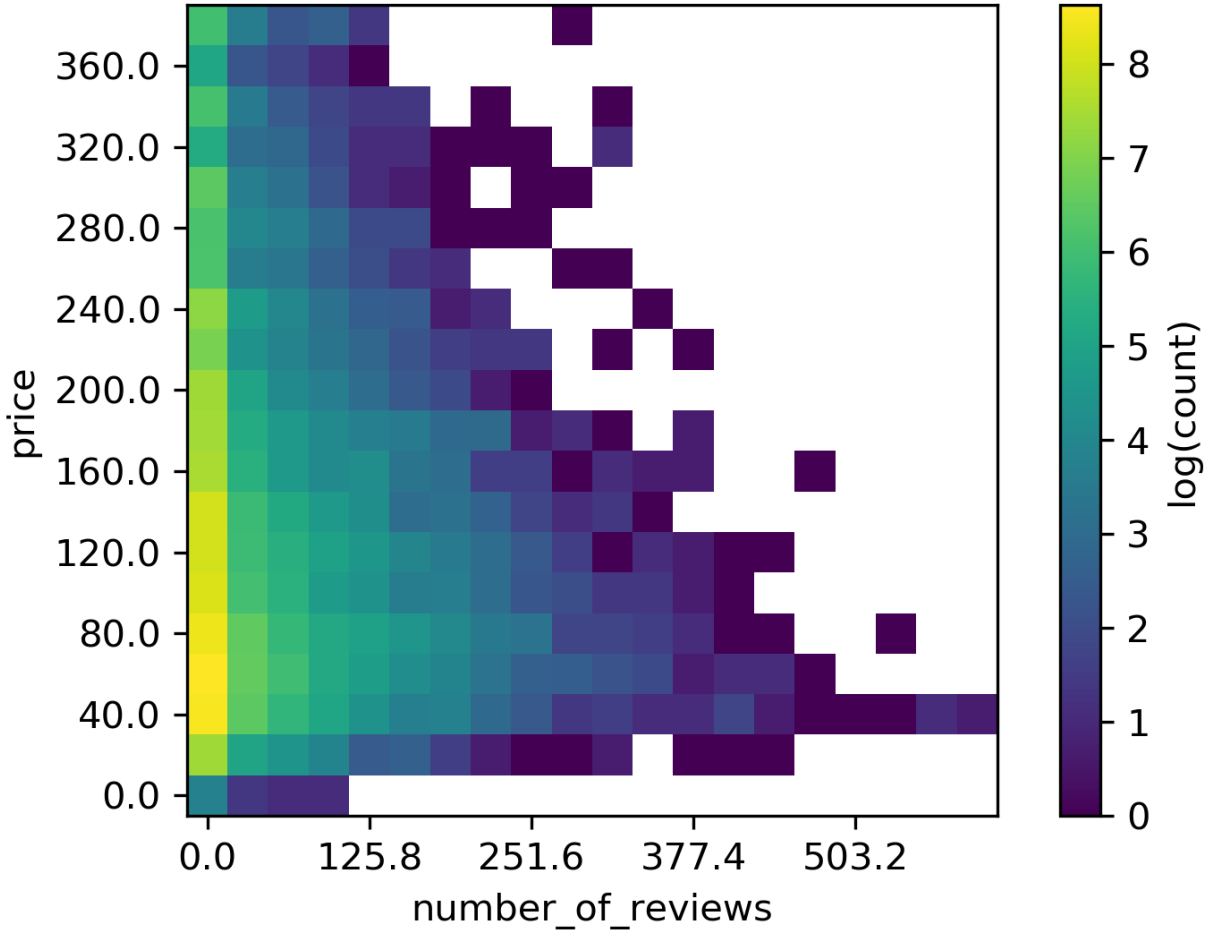
EDA - Price vs. Region



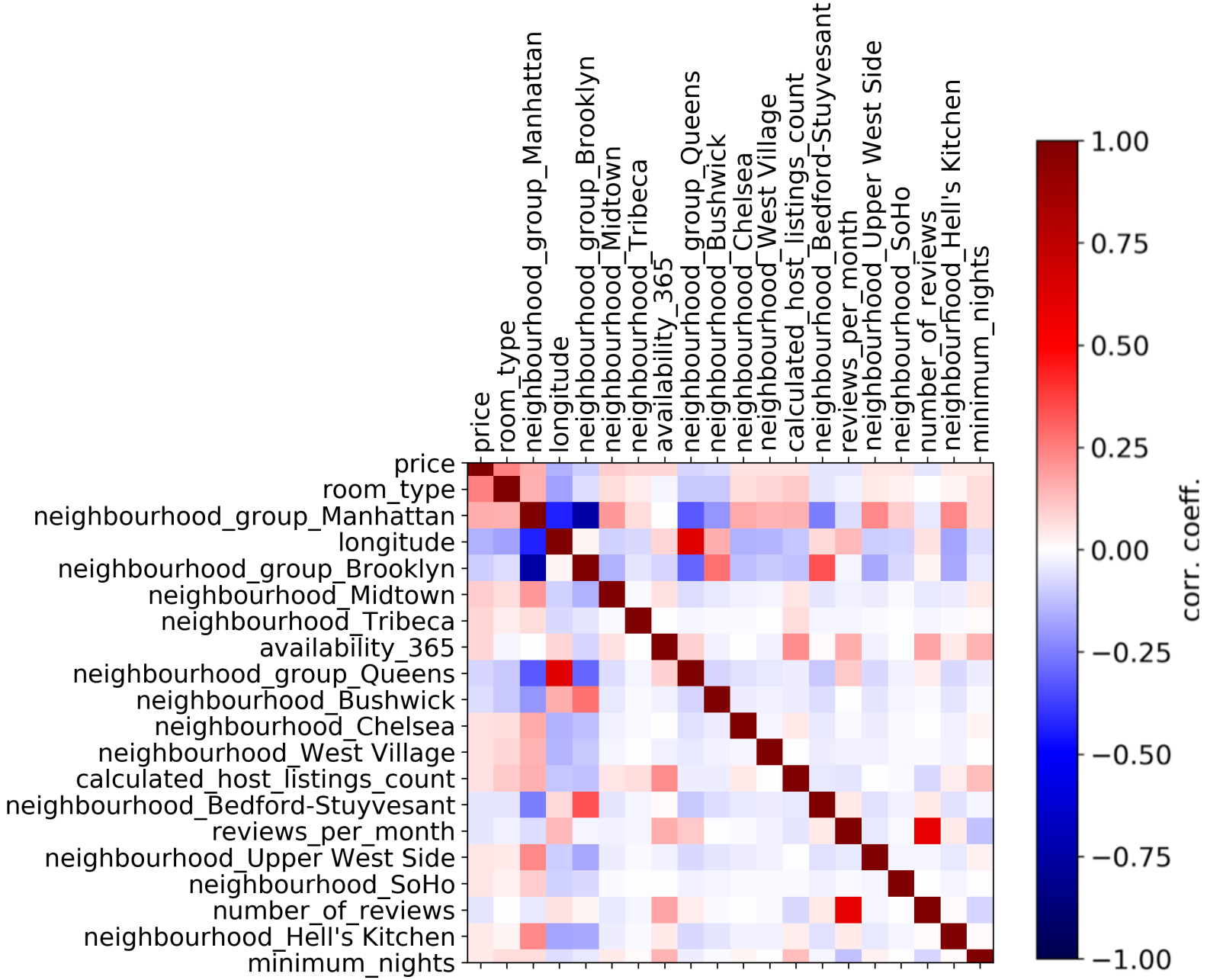
EDA - Price vs. Room Type



EDA - Price vs. Number of Reviews



EDA - The Correlation Matrix



Thank you!

Q & A