# New York City Airbnb Data Analysis and Price Predictions

Jihan Liang
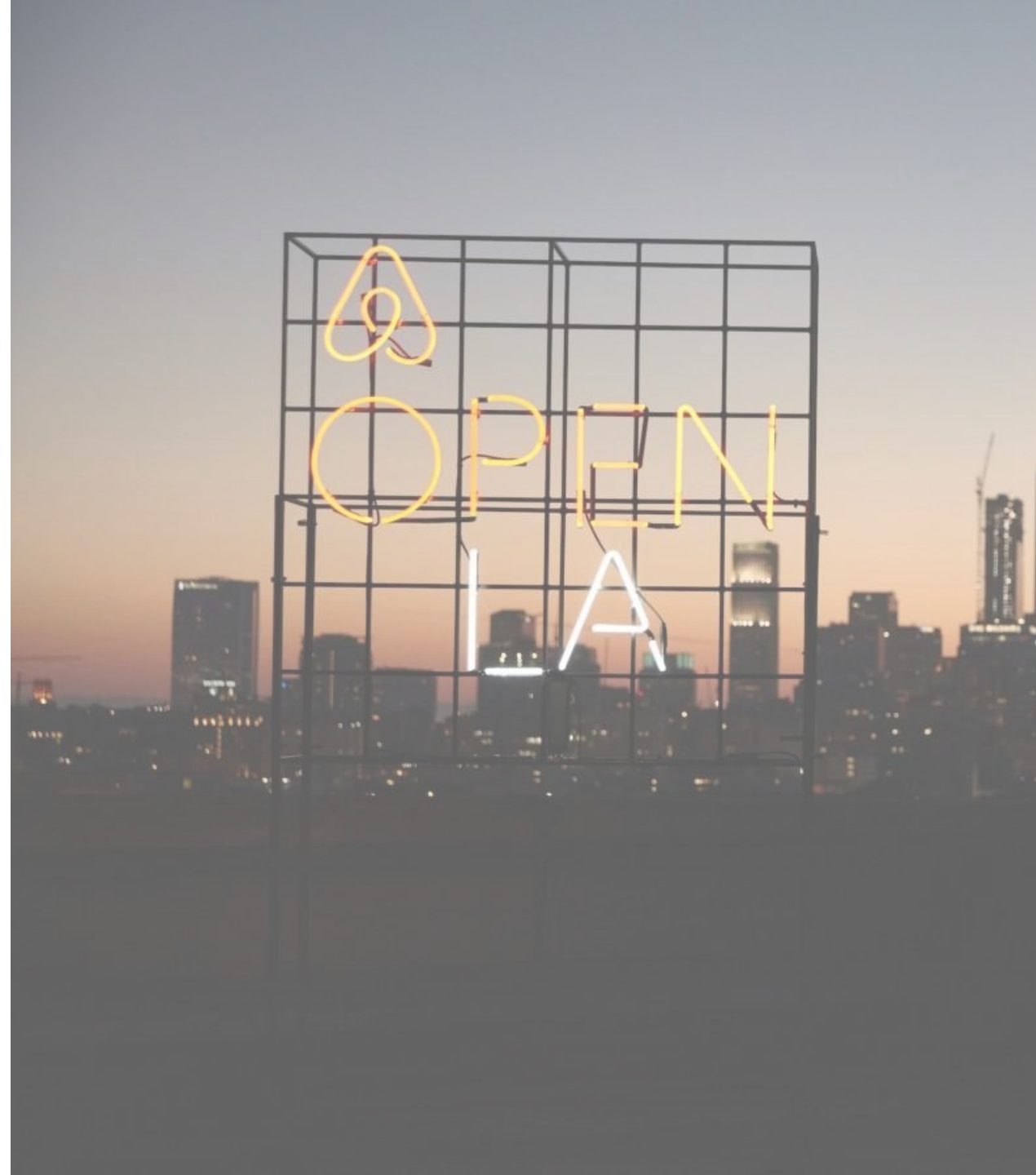Tingjun Man
DS-GA1007: Programming for Data Science

# Overview

1. Data Cleaning

2. Patterns of Interests

3. Preprocessing

4. Modeling

# Data Cleaning

**Data Type**

Change data type

**Nonsense Data**

Drop data that doesn't make sense

**Transform Data**

Transform data into usable forms

**Missing Data**

Tailored filling value for each variable

dataset size after cleaning: (27620, 29)

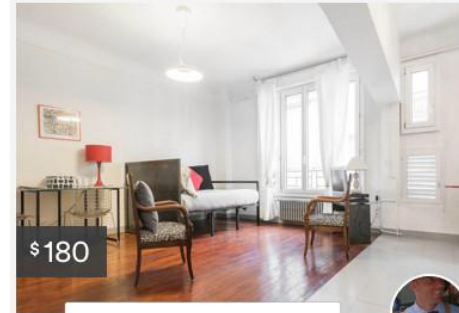# Patterns of Interests

1. Price
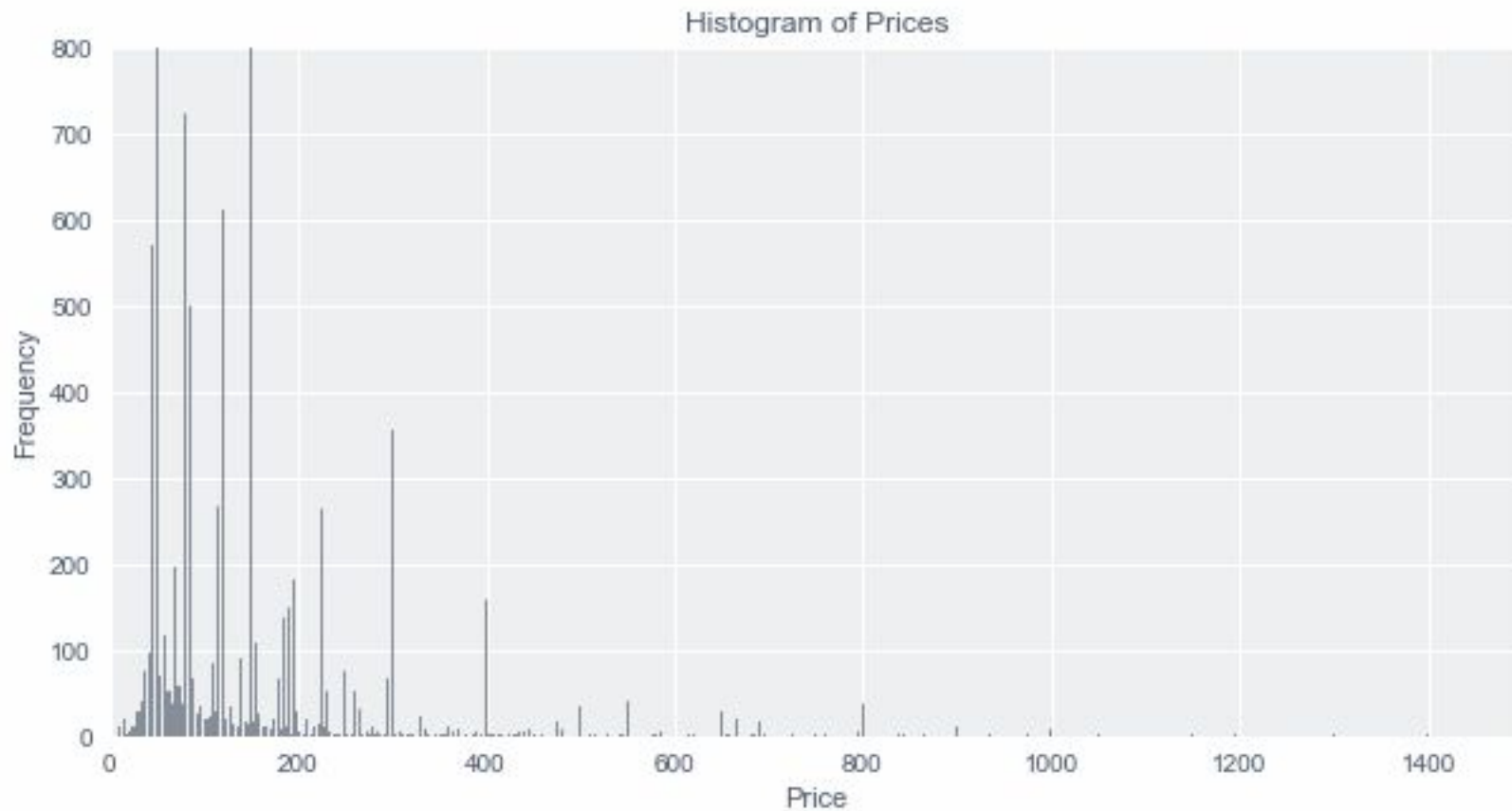
2. Neighbourhoods

3. Seasonality
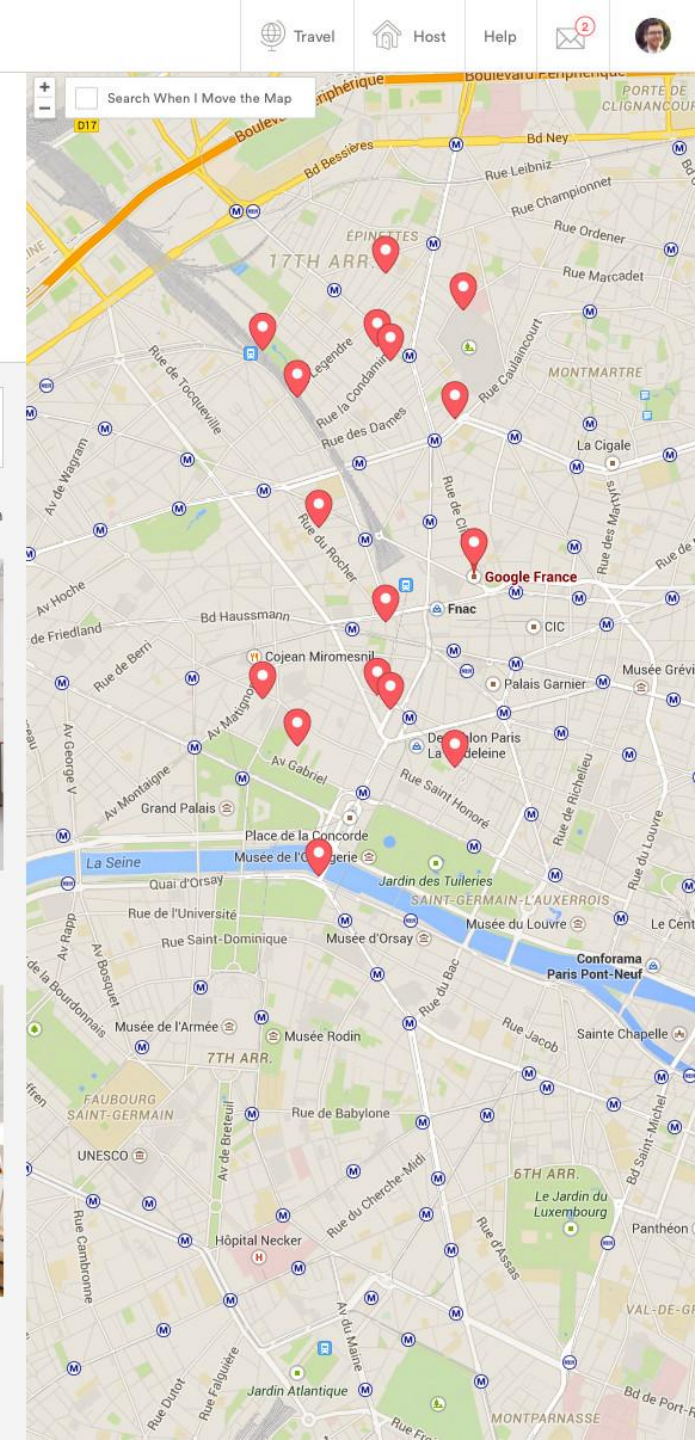
# NYC Price

Mean: $155.07
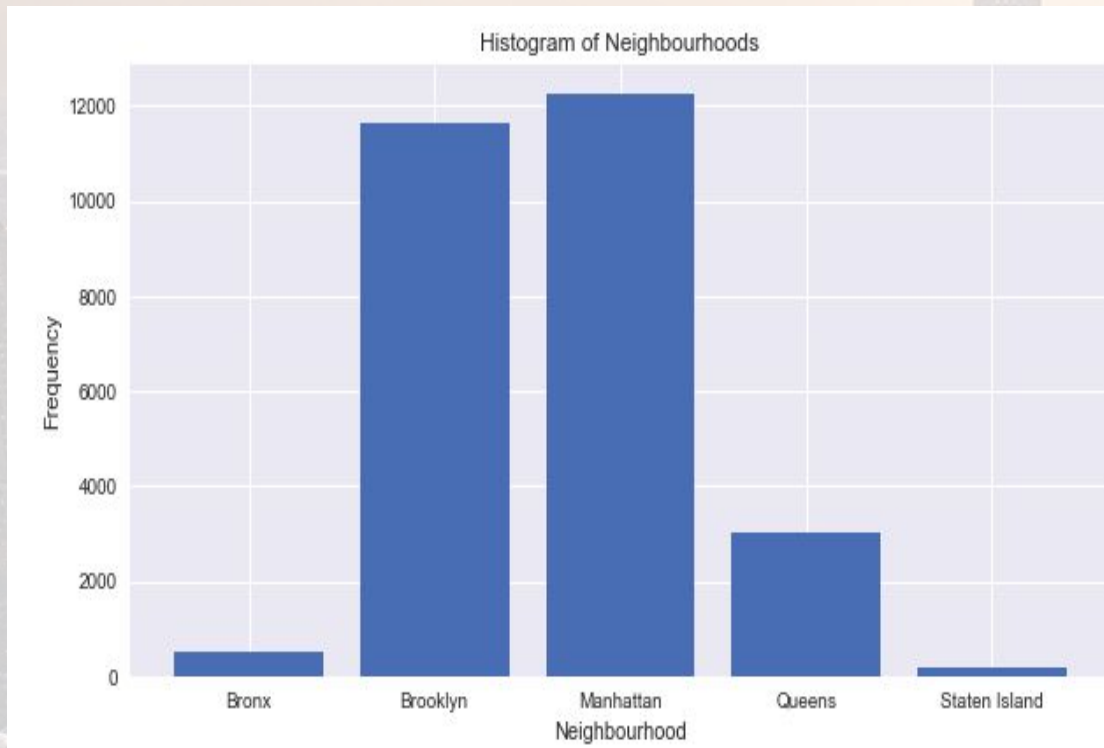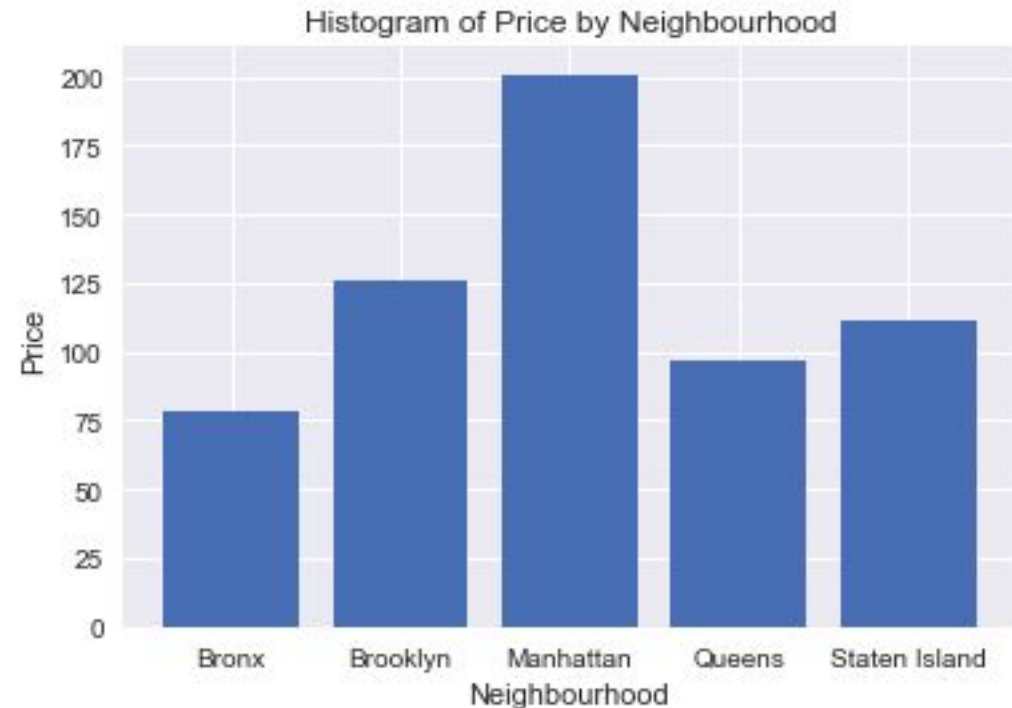
Median: $105

Range: [$10,$10000]

The majority of New York City Airbnb prices fall into the interval between $50/night to $300/night



Histogram of Prices
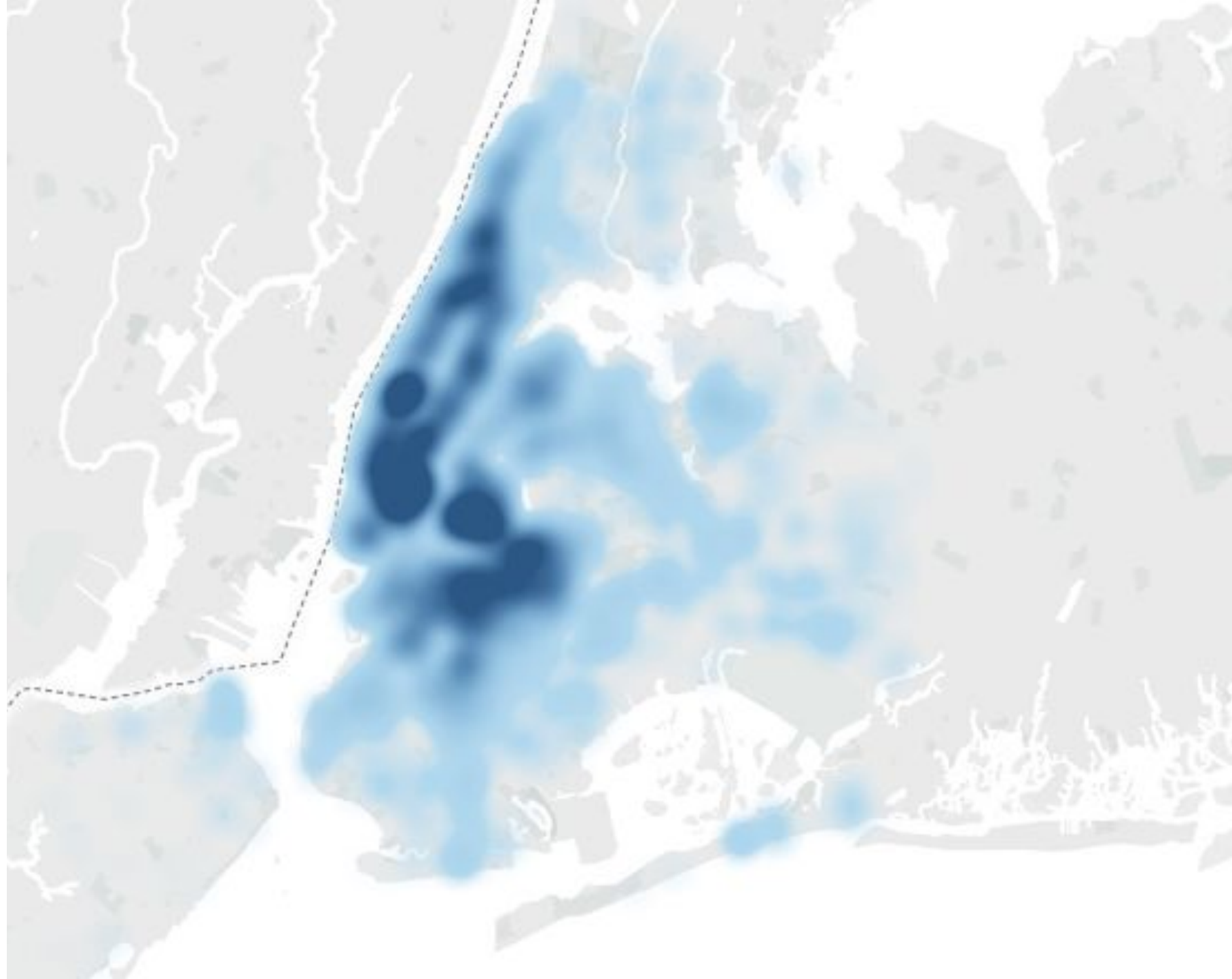
# Neighborhoods and Price-by-Neighborhood



Brooklyn and Manhattan have the most Airbnb listings.

Manhattan has the highest average listing price among all five neighborhoods.
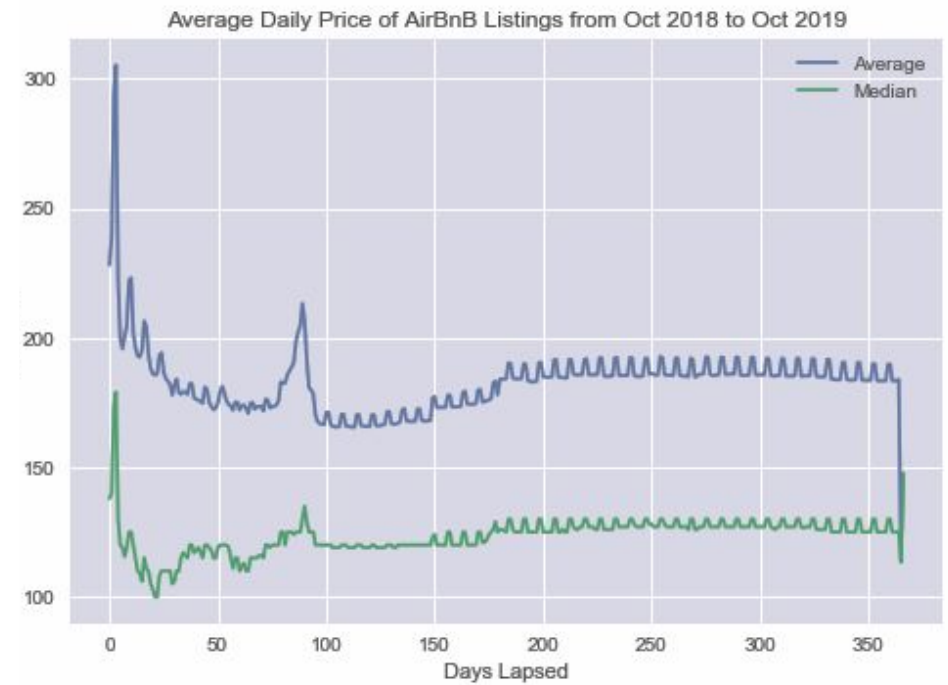
# Density Map

- Airbnb listing is more dense in **Manhattan** and **Brooklyn,** which is aligned with previous descriptive results.

- Specifically, **Midtown** and **Lower Manhattan** and **Downtown Brooklyn** have high density

Seasonal Trend
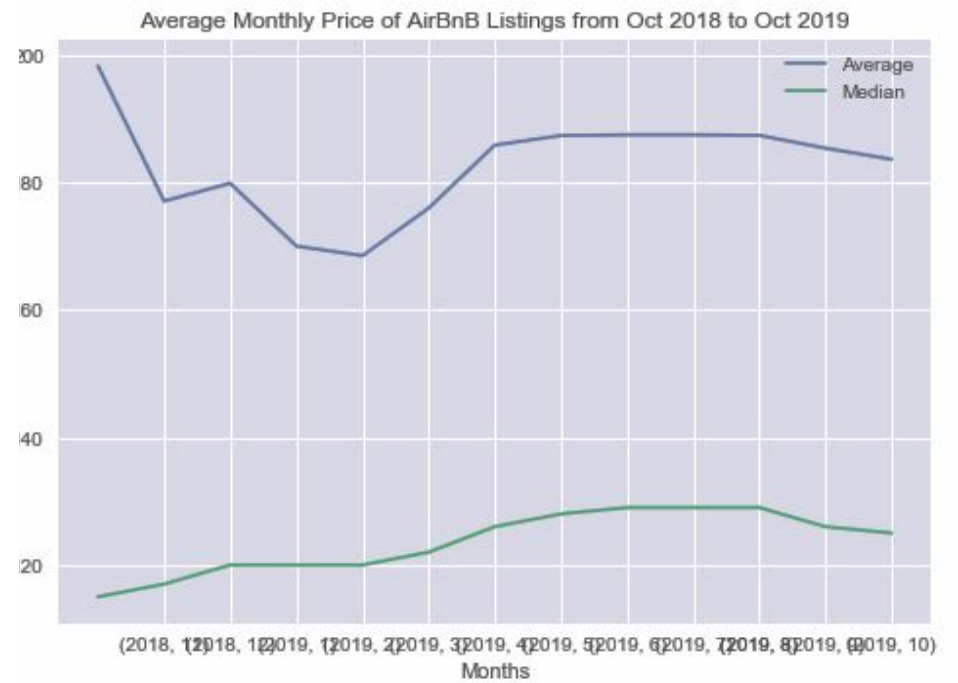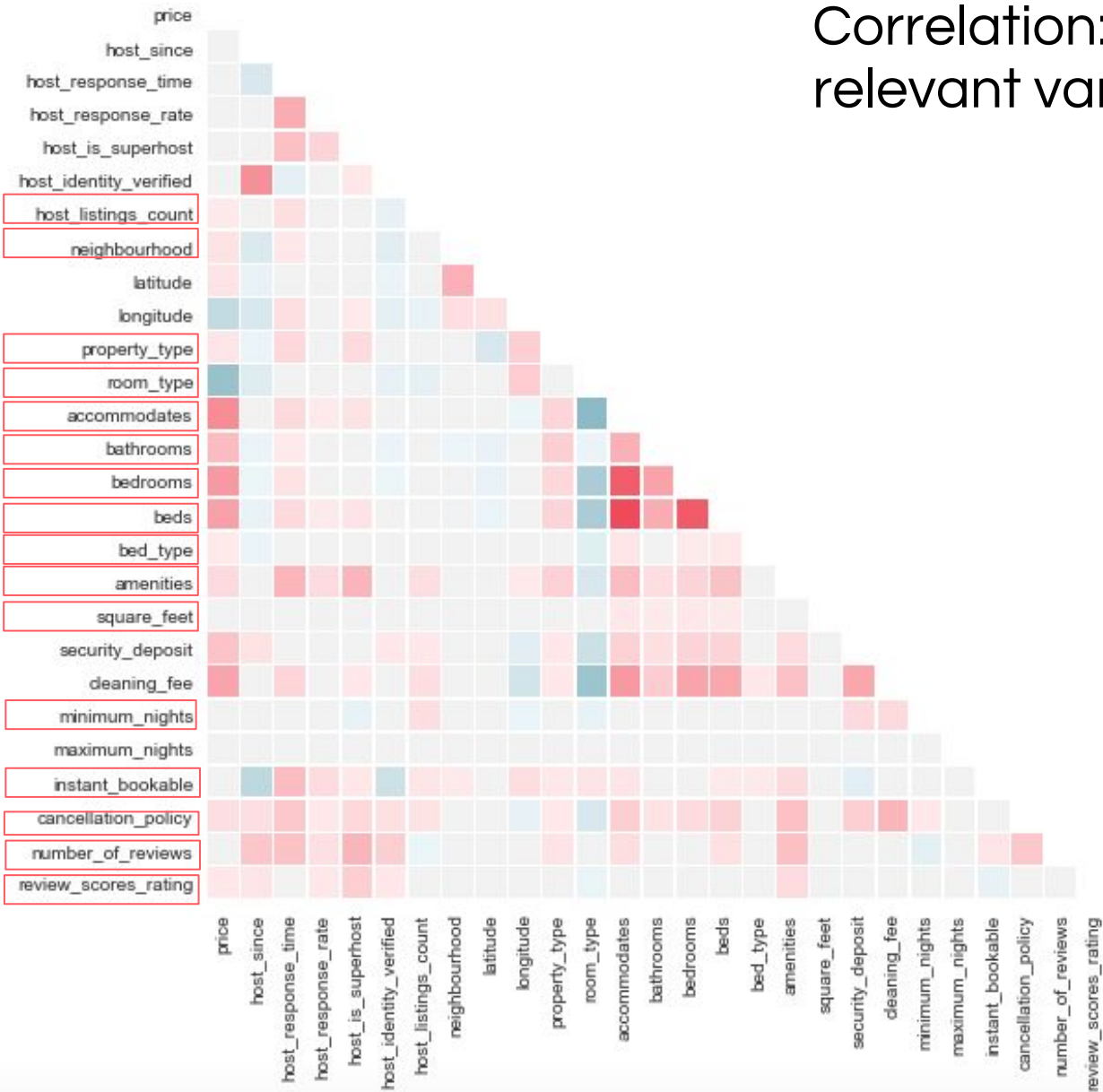
Daily average price



Average Daily Price of AirBnB Listings from Oct 2018 to Oct 2019

Monthly average price



Average Monthly Price of AirBnB Listings from Oct 2018 to Oct 2019

# Preprocessing

Correlation: to identify relevant variables
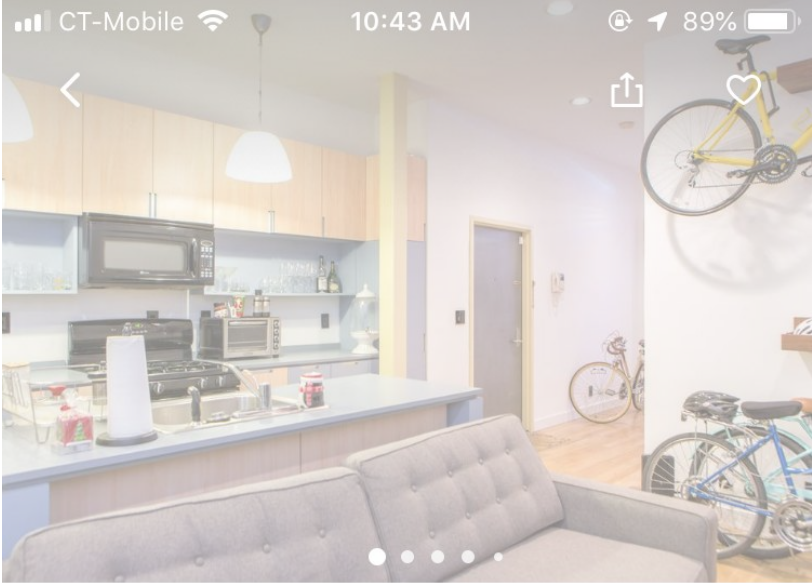
# Preprocessing

## DBSCAN: to detect and remove noise



Estimated number of clusters: 11
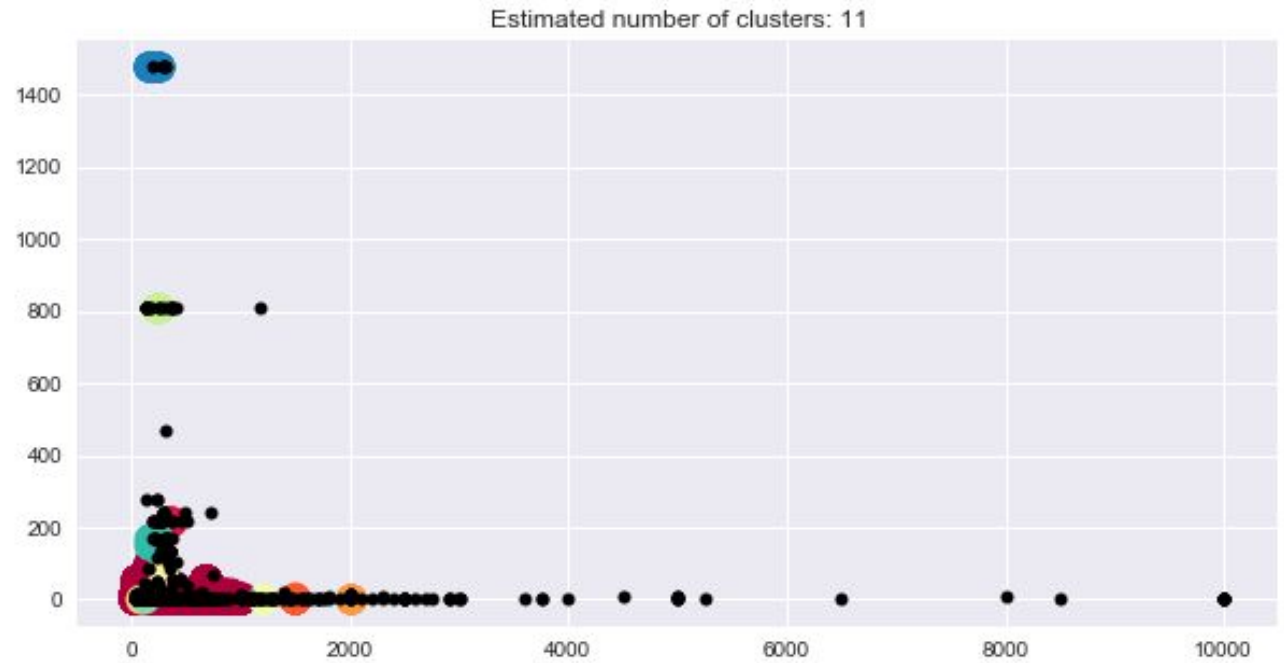
dataset size after preprocessing: (26844, 17)

# Modeling – with seasonality

- Use Monday price as the base value, and calculate a multiplier for each other weekday

- Predict Monday price with the model, then multiply the prediction by each multiplier to get prices for the whole week

- Compare the set of true prices for each weekday with the set of predicted prices

Median-Absolute-Error

```
Ordinary Least Squares
OLS(train): 71.86255395139138
OLS(test): 67.92803575470674
-----
Decision Tree
Decision Tree (train): 60.70415356751309
Decision Tree (test): 58.94940107637346
-----
Random Forest
Random Forest (train): 62.75918152821175
Random Forest (test): 61.88386849951431
```

# Modeling-Without Seasonality

## Median-Absolute-Error



```
Ordinary Least Squares
OLS(train): 34.0089250842336
OLS(test): 34.33885959126813

-----

Decision Tree
Decision Tree (train): 24.941295546558706
Decision Tree (test): 26.27906976744186

-----

Random Forest
Random Forest (train): 9.80000000000011
Random Forest (test): 26.90000000000006
```

### Scatterplot of Error