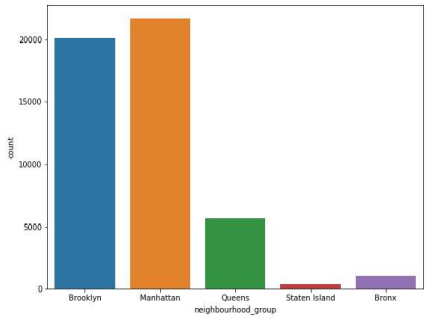
# New York Airbnb Data

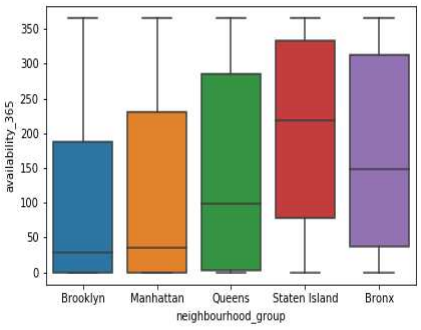
Team 33: Andrew Pesek, Chen Xu, Andrei Sebald, Boyuan Zheng

## Exploration:

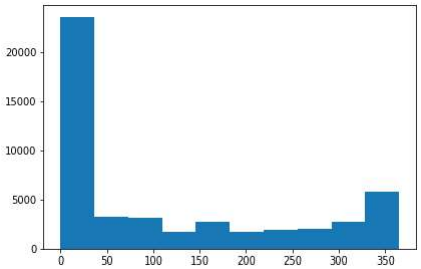
### Distribution of neighbourhood groups




### Box plot of availability across neigbourhood groups

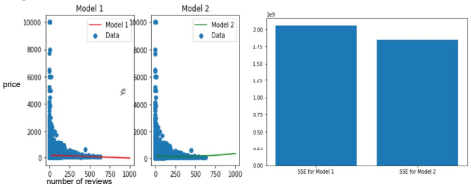



### Distribution of availability



## Linear Regression:

### Possible models of price prediction - 4 different linear regression models

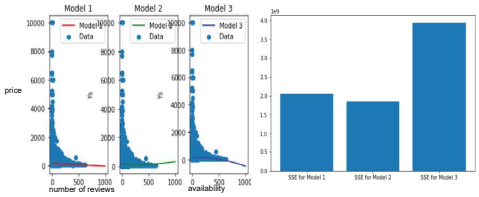



**Model 1: price = w0 + w1 x number_of_reviews**

**w0, w1 = 155.069, -0.197**

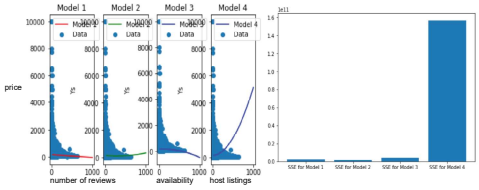
**Model 2: price = w0 + w1 x number_of_reviews + w2 x availability_365^2**

**w0, w1, w2 = 145.585, -0.267, 0.000439**




**Model 3: price = w0 + w1 x availability_365 + w2 x number_of_reviews^2**

**w0, w1, w2 = 138.847, 0.130, -0.000755**



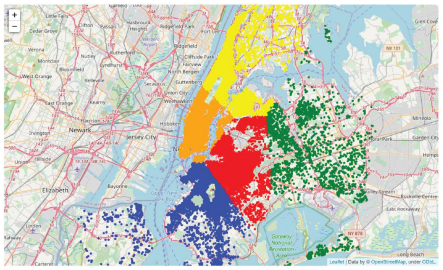

**Model 4: price = w0 + w1 x number_of_reviews + w2 x calculated_host_listings_count^2**

**w0, w1, w2 = 137.349, 0.107, 0.00490**

**Lowest SSE : Model 2**

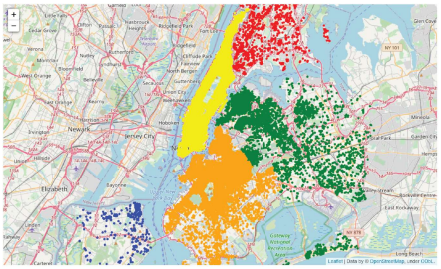
## Clustering:

### Method 1 - clustering by latitude and longitude



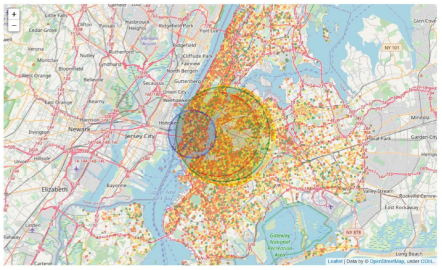
| | avg price | avg minimum nights | avg number of reviews | avg host listings | avg availability |
|---|---|---|---|---|---|
| 0: red | 114.75, (std=169.7) | 6.02, (std=16.5) | 24.38, (std=44.8) | 2.86, (std=8.4) | 103.75, (std=127.9) |
| 1: orange | 227.92, (std=327.6) ▲ | 9.28, (std=26.4) ▲ | 19.56, (std=41.1) ▼ | 17.05, (std=57.1) ▲ | 116.13, (std=134.9) |
| 2: yellow | 126.53, (std=210.2) | 6.56, (std=18.5) | 24.26, (std=46.2) | 3.29, (std=12.4) | 110.17, (std=129.8) |
| 3: green | 98.66, (std=117.4) ▼ | 3.69, (std=11.1) ▼ | 32.15, (std=56.0) | 3.56, (std=8.1) | 172.38, (std=133.7) ▲ |
| 4: blue | 130.81, (std=171.3) | 6.21, (std=18.7) | 24.19, (std=42.8) | 1.85, (std=2.6) | 106.70, (std=128.2) |

### Method 2 - samples labeled by neighborhood group



| | avg price | avg minimum nights | avg number of reviews | avg host listings | avg availability |
|---|---|---|---|---|---|
| 0: red | 87.58, (std=106.7) | 4.56, (std=15.6) | 25.98, (std=42.2) | 2.23, (std=2.4) | 165.79, (std=135.2) |
| 1: orange | 124.44, (std=186.9) | 6.06, (std=17.6) | 24.20, (std=44.3) | 2.28, (std=5.3) | 100.22, (std=126.3) |
| 2: yellow | 196.88, (std=291.4) ▲ | 8.58, (std=24.1) | 20.99, (std=42.6) | 12.79, (std=48.2) ▲ | 111.98, (std=132.7) |
| 3: green | 99.52, (std=167.1) | 5.18, (std=15.0) | 27.70, (std=52.0) | 4.06, (std=12.4) | 144.45, (std=135.5) |
| 4: blue | 114.81, (std=277.2) | 4.83, (std=19.7) | 30.94, (std=44.8) | 2.32, (std=1.9) | 199.68, (std=131.7) |

### Method 3 - clustering by price, minimum nights, number of reviews, host listings and availability



| | avg price | avg minimum nights | avg number of reviews | avg host listings | avg availability |
|---|---|---|---|---|---|
| 0: red | 136.73, (std=190.3) | 4.83, (std=13.2) | 12.28, (std=26.8) | 1.66, (std=6.5) | 5.20, (std=10.9) |
| 1: orange | 178.29, (std=311.3) | 11.84, (std=33.4) | 28.70, (std=49.9) | 9.73, (std=21.3) | 330.53, (std=29.8) |
| 2: yellow | 146.36, (std=251.1) | 6.10, (std=16.8) | 34.98, (std=51.2) | 2.54, (std=6.3) | 87.18, (std=26.9) |
| 3: green | 170.97, (std=271.2) | 7.95, (std=20.4) | 43.86, (std=64.6) | 5.50, (std=16.3) | 202.18, (std=36.1) |
| 4: blue | 273.57, (std=101.7) ▲ | 20.96, (std=16.0) ▲ | 2.39, (std=3.8) ▼ | 288.79, (std=46.6) ▲ | 287.76, (std=72.7) |