# Exploratory Data Analysis
# &
# Smart Price Prediction

Citina Liang
sl4671

# Data Overview

- 48,895 observations, 16 variables (before cleaning)
- 47,709 observations, 12 variables
- Numerical Variables:

  **latitude:** latitude coordinates, **longitude:** longitude coordinates, **price:** price in dollars, **minimum_nights:** amount of nights minimum, **number_of_reviews:** number of reviews, **reviews_per_month:** number of reviews per month, **calculated_host_listings_count:** amount of listing per host, **availability_365:** number of days when listing is available for booking

- Categorical Variables:

  **name:** name of the listing, **neighbourhood_group:** location, **neighbourhood:** area, **room_type:** listing space type

# Data Overview

| name | neighbourhood_group | neighbourhood | latitude | longitude | room_type | price | minimum_nights | number_of_reviews | reviews_per_month | calculated_host_listings_count | availability_365 |
|------|---------------------|---------------|----------|-----------|-----------|-------|----------------|-------------------|-------------------|-------------------------------|------------------|
| Clean & quiet apt home by the park | Brooklyn | Kensington | 40.64749 | -73.97237 | Private room | 149 | 1 | 9 | 0.21 | 6 | 365 |
| Skylit Midtown Castle | Manhattan | Midtown | 40.75362 | -73.98377 | Entire home/apt | 225 | 1 | 45 | 0.38 | 2 | 355 |
| THE VILLAGE OF HARLEM....NEW YORK ! | Manhattan | Harlem | 40.80902 | -73.94190 | Private room | 150 | 3 | 0 | 0.00 | 1 | 365 |
| Cozy Entire Floor of Brownstone | Brooklyn | Clinton Hill | 40.68514 | -73.95976 | Entire home/apt | 89 | 1 | 270 | 4.64 | 1 | 194 |
| Entire Apt: Spacious Studio/Loft by central park | Manhattan | East Harlem | 40.79851 | -73.94399 | Entire home/apt | 80 | 10 | 9 | 0.10 | 1 | 0 |
| Large Cozy 1 BR Apartment In Midtown East | Manhattan | Murray Hill | 40.74767 | -73.97500 | Entire home/apt | 200 | 3 | 74 | 0.59 | 1 | 129 |
| Large Furnished Room Near B'way | Manhattan | Hell's Kitchen | 40.76489 | -73.98493 | Private room | 79 | 2 | 430 | 3.47 | 1 | 220 |
| Cozy Clean Guest Room - Family Apt | Manhattan | Upper West Side | 40.80178 | -73.96723 | Private room | 79 | 2 | 118 | 0.99 | 1 | 0 |
| Cute & Cozy Lower East Side 1 bdrm | Manhattan | Chinatown | 40.71344 | -73.99037 | Entire home/apt | 150 | 1 | 160 | 1.33 | 4 | 188 |
| Beautiful 1br on Upper West Side | Manhattan | Upper West Side | 40.80316 | -73.96545 | Entire home/apt | 135 | 5 | 53 | 0.43 | 1 | 6 |

# Number of Listings in 5 Neighbourhood Groups
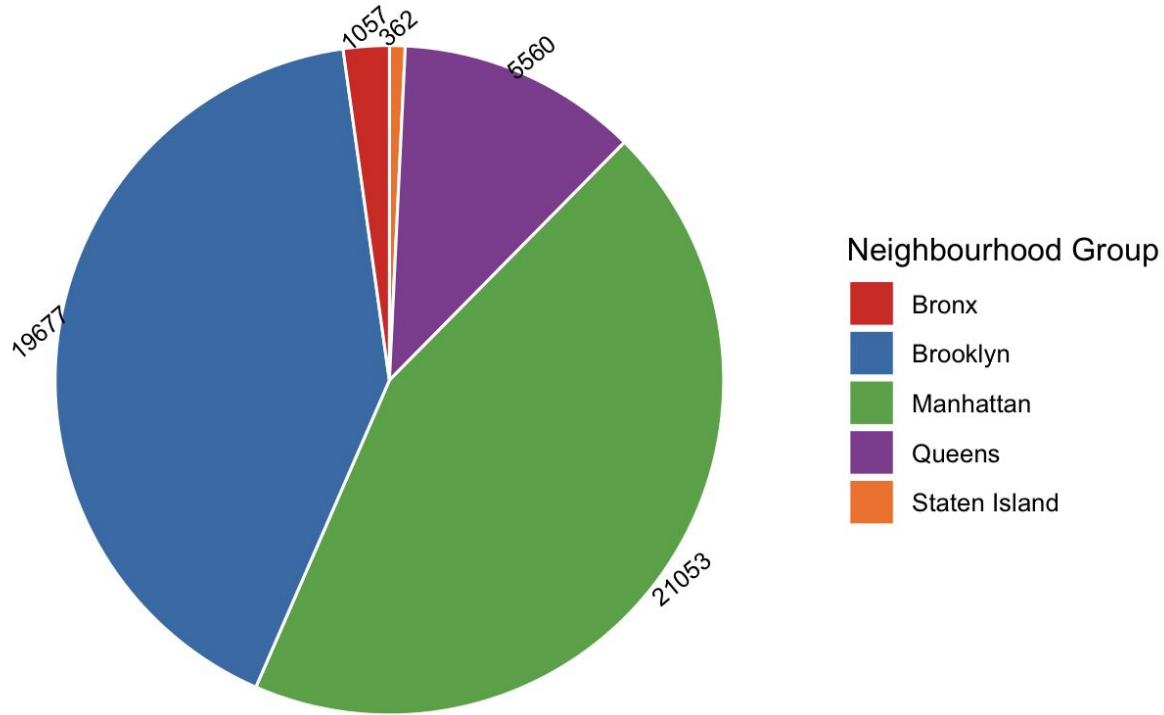
# Average Price Per Night by Neighbourhood Group

Price
Heatmap

300

200

100
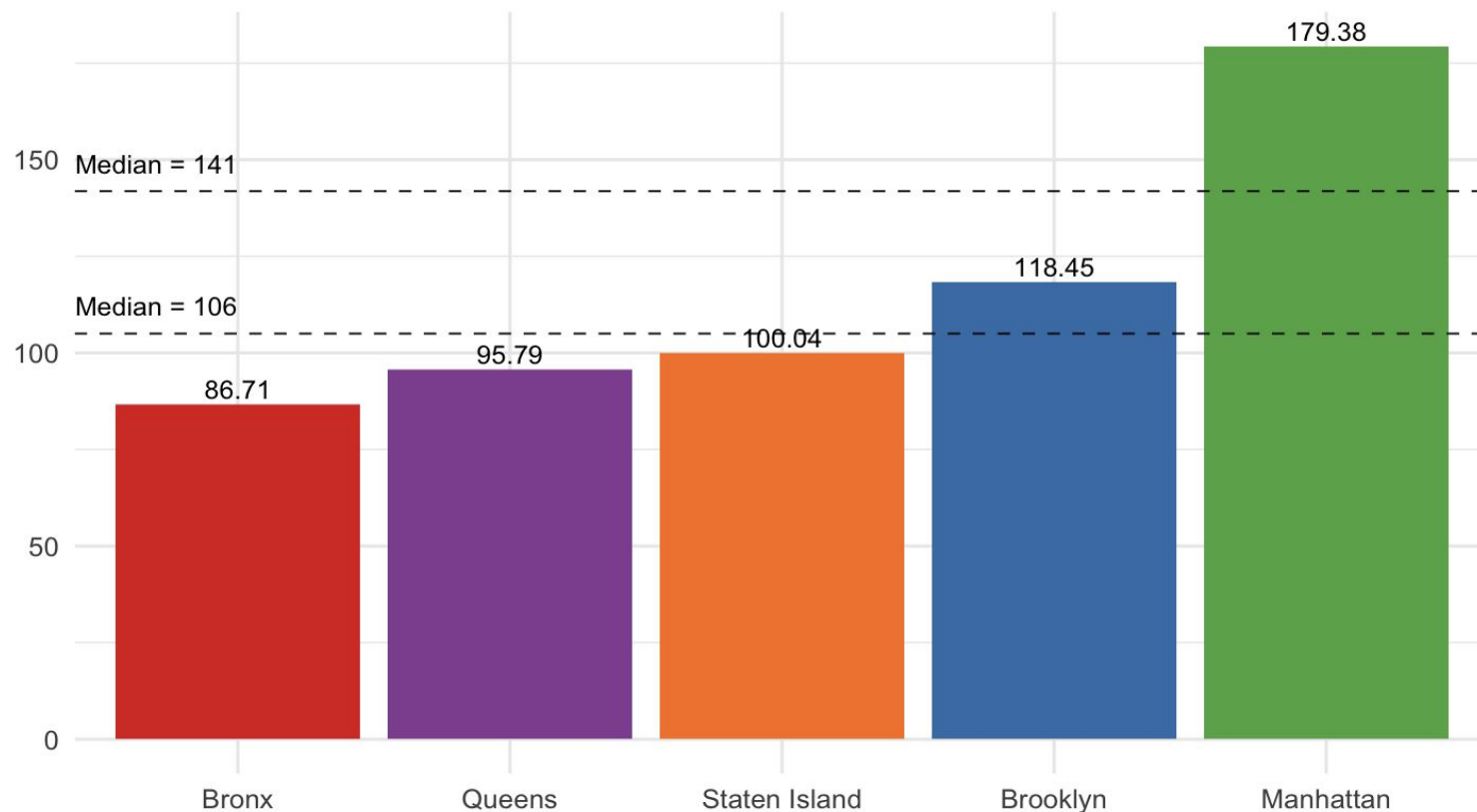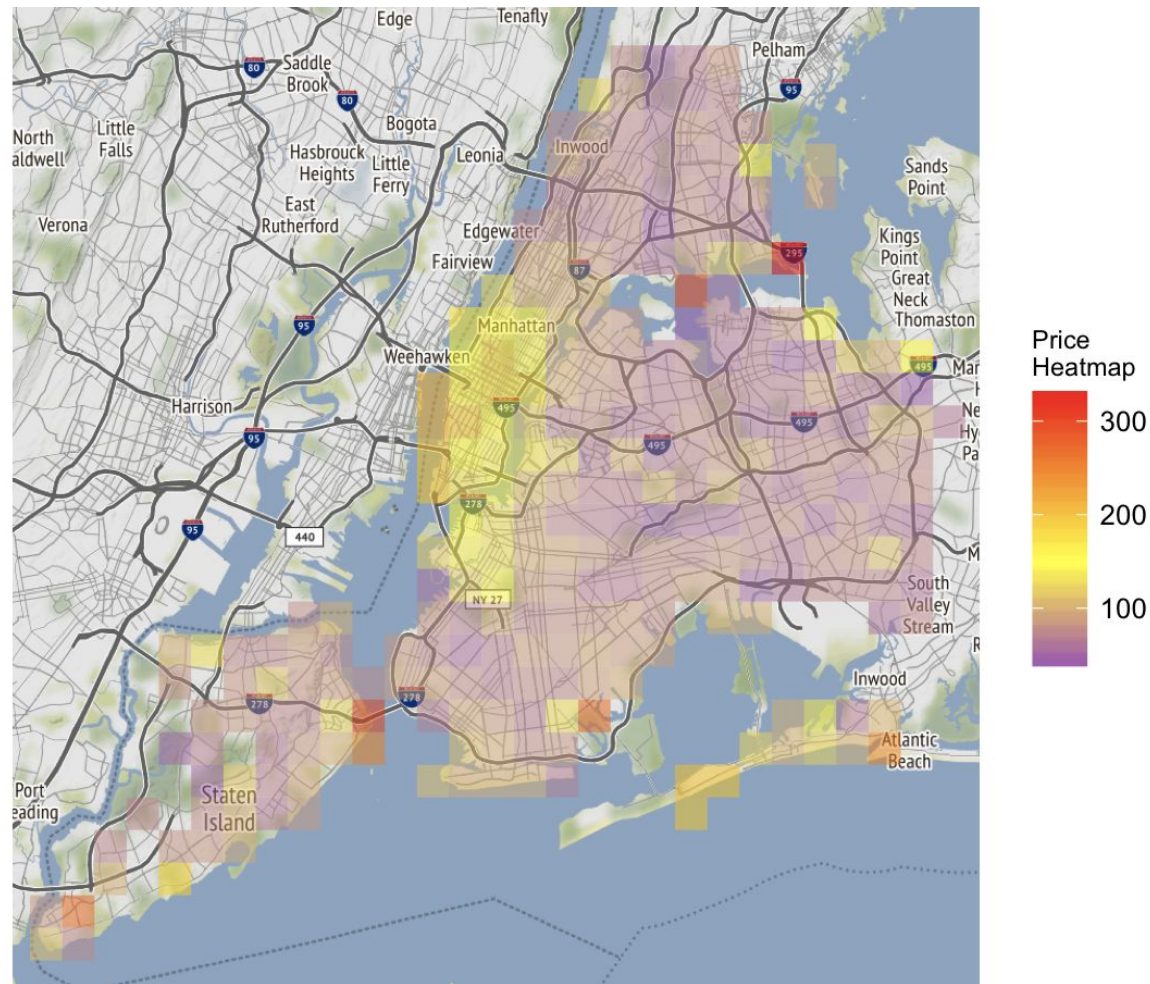
- Listings near Manhattan Midtown, west village area tend to be more expensive.

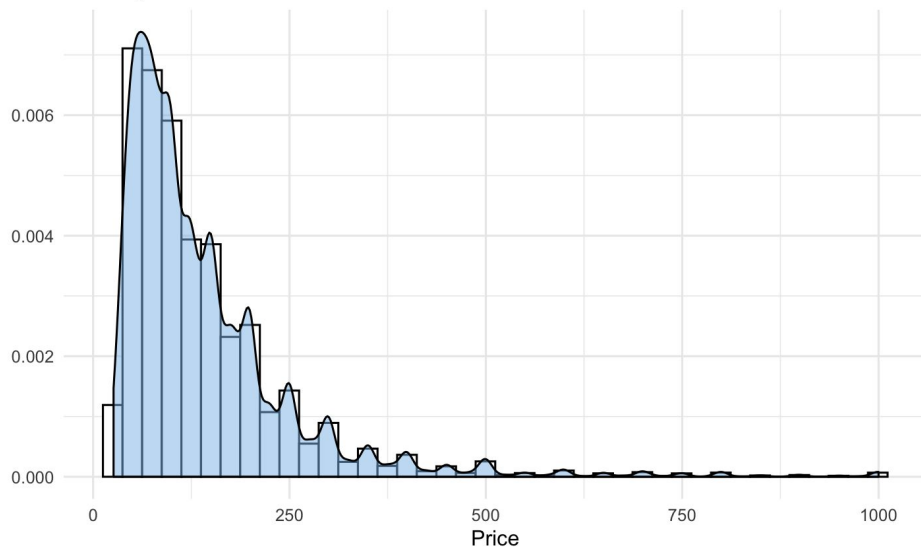- Other areas have some stays more expensive than usual price, and most of them are near the gulf area
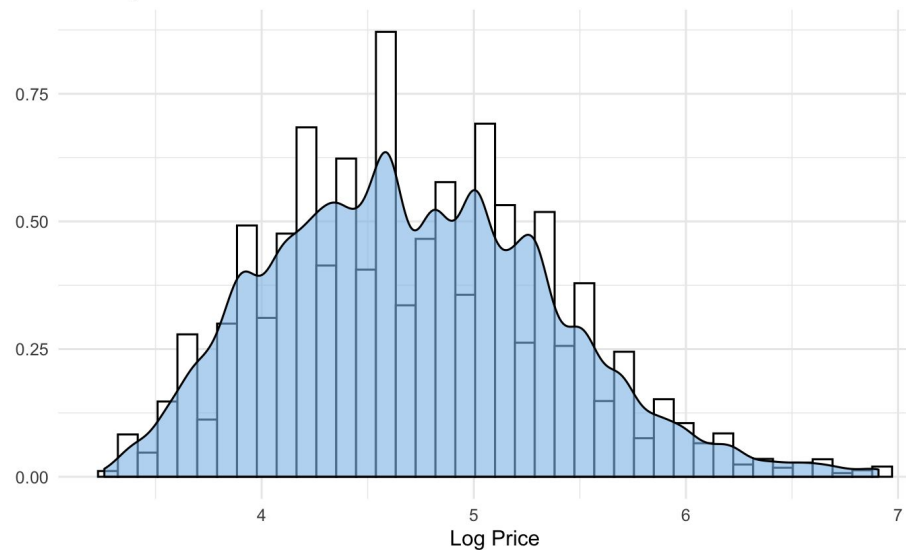
# Average Price by Neighbourhood and Room Type
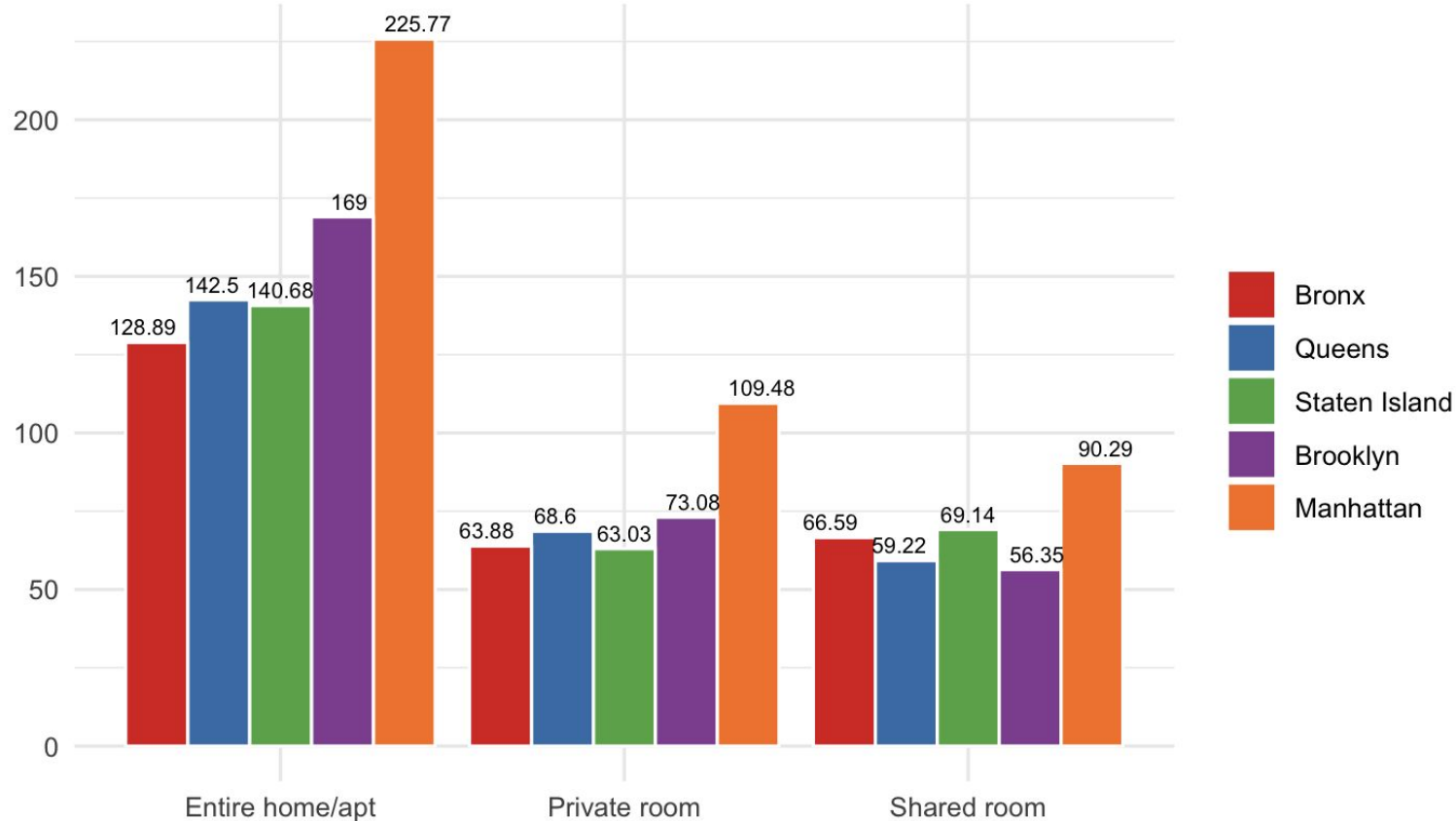
# Multiple Linear Regression for Price Prediction

● Log transformation on predicted variable (price)



Density Plot of Price

Density Plot of Price after Transformation

# Multiple Linear Regression for Price Prediction

- Model building & variables selection

1. Split data into training (80%) and testing (20%)
2. Build a full model with training data:
   lm(log_price ~ latitude + longitude + room_type + minimum_nights + availability_365 + number_of_reviews + reviews_per_month + calculated_host_listings_count + neighbourhood_group, data = train)
3. Stepwise model selection from both directions, the result gives back the full model
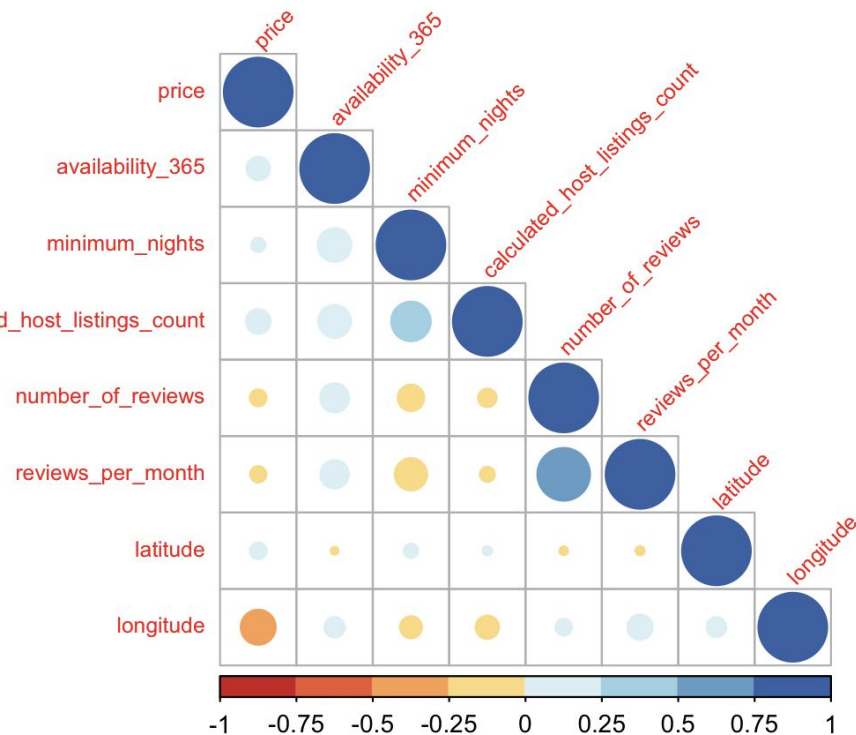
# Diagnostic Plots and Model Prediction Result
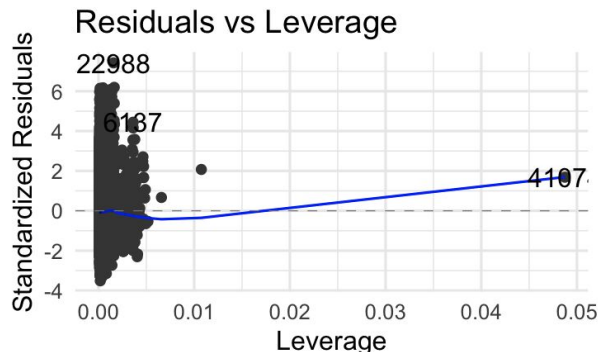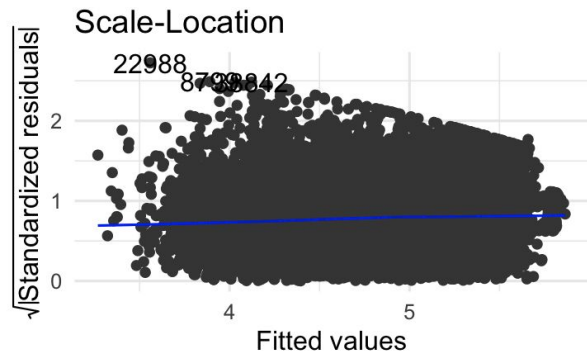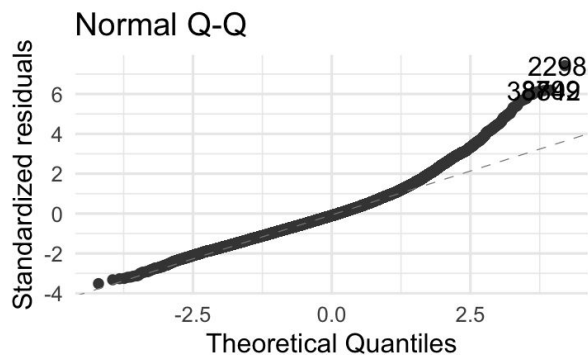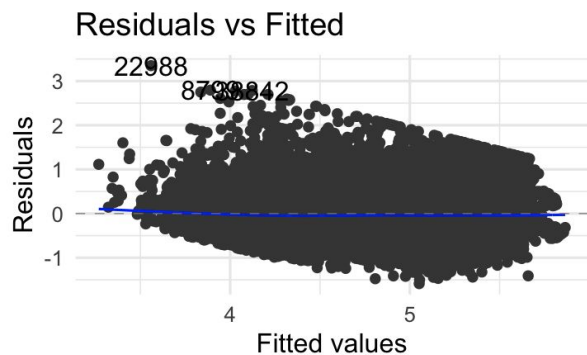


1. Diagnostic plots look fine
2. Adjusted R² of training data is 0.5328
3. Adjusted R² of testing data is 0.5316
4. All predictors are significant expect
   neighbourhood_groupBrooklyn

```
Coefficients:
                                   Estimate Std. Error  t value Pr(>|t|)
(Intercept)                      -1.912e+02  7.173e+00  -26.653  < 2e-16 ***
latitude                         -5.261e-01  7.029e-02   -7.484 7.34e-14 ***
longitude                        -2.942e+00  8.054e-02  -36.532  < 2e-16 ***
room_typePrivate room            -7.643e-01  4.849e-03 -157.601  < 2e-16 ***
room_typeShared room             -1.092e+00  1.606e-02  -67.960  < 2e-16 ***
minimum_nights                   -1.174e-02  3.182e-04  -36.886  < 2e-16 ***
availability_365                  8.437e-04  1.946e-05   43.350  < 2e-16 ***
number_of_reviews                -6.540e-04  6.446e-05  -10.146  < 2e-16 ***
reviews_per_month                -1.467e-02  1.820e-03   -8.057 8.04e-16 ***
calculated_host_listings_count    4.490e-04  7.795e-05    5.760 8.48e-09 ***
neighbourhood_groupBrooklyn      -6.317e-03  1.978e-02   -0.319    0.749
neighbourhood_groupManhattan      2.907e-01  1.794e-02   16.203  < 2e-16 ***
neighbourhood_groupQueens         1.170e-01  1.903e-02    6.148 7.93e-10 ***
neighbourhood_groupStaten Island -7.686e-01  3.742e-02  -20.544  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4515 on 38153 degrees of freedom
Multiple R-squared:  0.533,    Adjusted R-squared:  0.5328
F-statistic:  3350 on 13 and 38153 DF,  p-value: < 2.2e-16
```

# Limitations and Future Work

1. Consider interaction terms
2. Try more models (ridge, lasso, knn, etc.)
3. Try dimension reduction
4. Cross validation
5. Can do a time series analysis with more data from more years
6. Can do a text mining analysis (for example, topic modeling to create new predictors)
7. Can access external data (area criminal rates, transportation, etc.) to better estimate the price
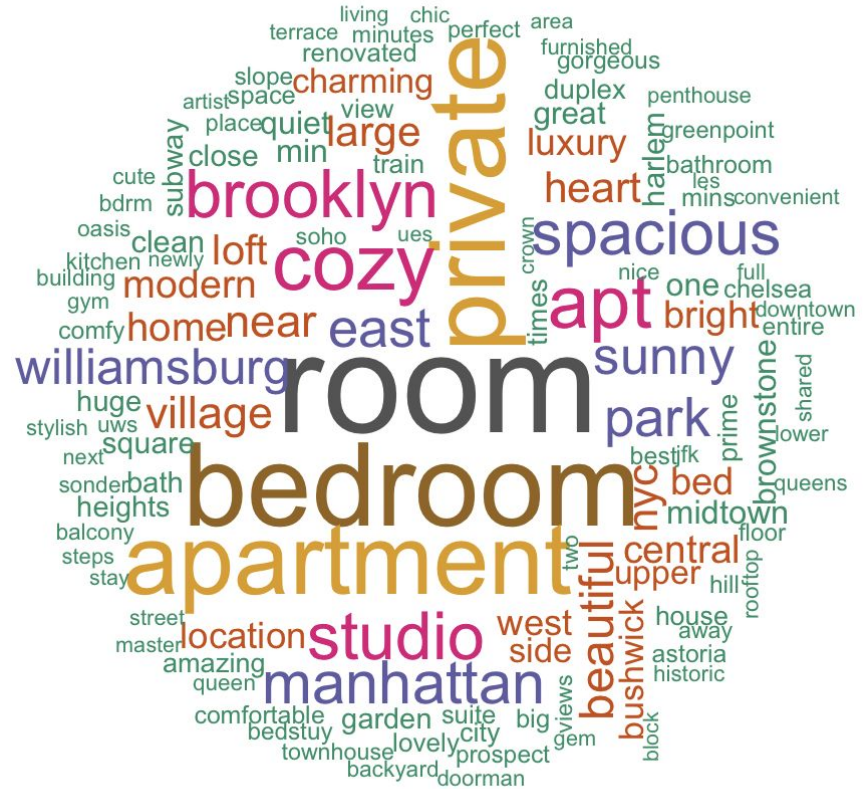
# Thank you!
# &
# Happy Wednesday!

# Data Cleaning

- 48,895 observations, 16 variables
- Procedures:
  - Deal with missing variables (replace or delete the observation)
  - Remove uninformative variables (id, host_id, host_name)
  - Remove outliers (price out of 99.5% interval)
  - Change categorical variables into factors