Budget trip to New York city?

Ng Yao Xuan

Can we predict the price of the Airbnb Listing?

Approach

1. Data collection
2. Data Cleansing
3. Data Exploring
4. Data Modelling

Data Collection

- New York City Airbnb Listings in 2019
- 48895 observations
- 16 features

Features

- Neighbourhood group
- Room type
- Minimum nights
- Number of reviews
- Last review date
- Number of days when listing is available for booking

Target

We want to predict

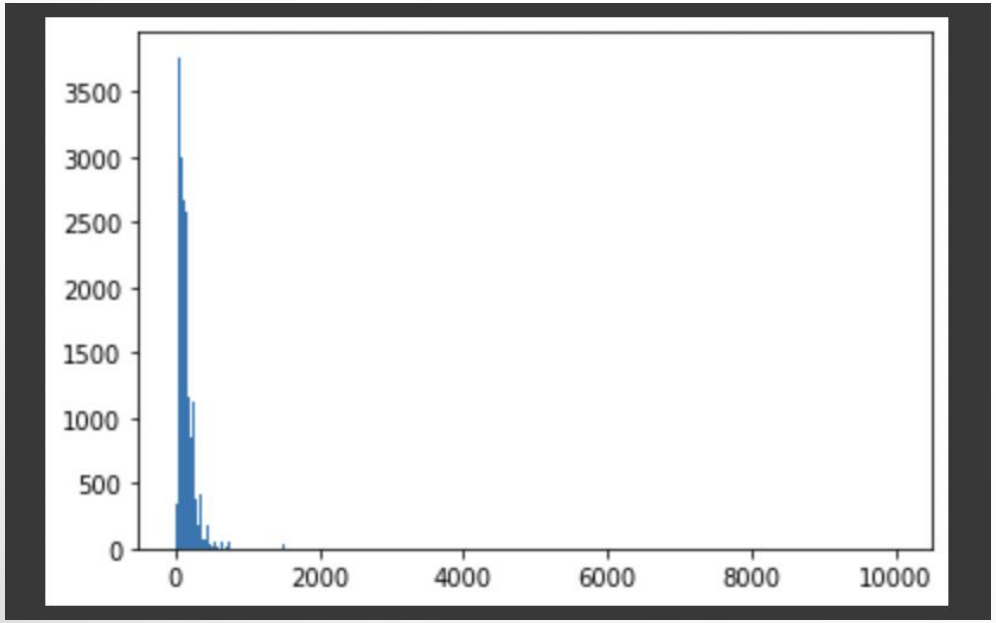
Price

of the Airbnb price listings

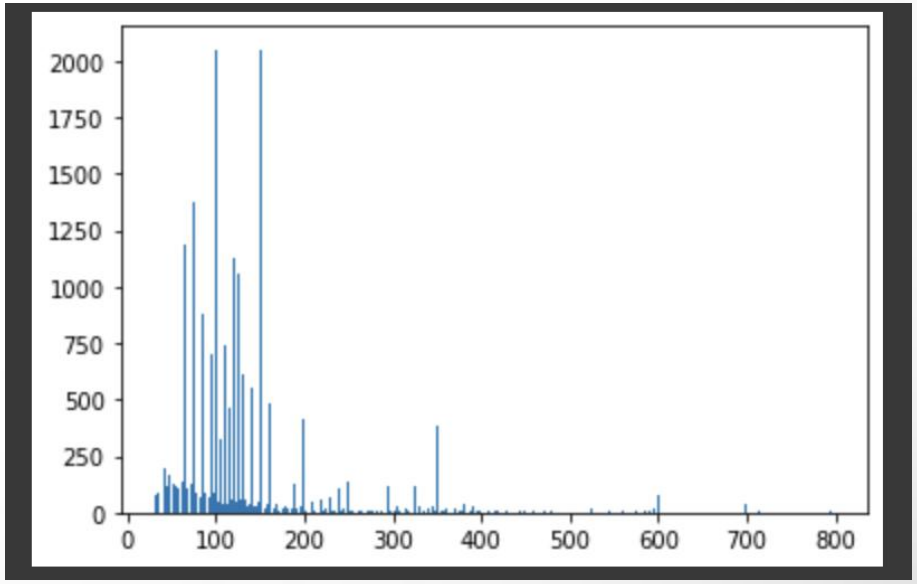
Data Cleansing

clean up price data as the original data is skewed

```
1  p1 = df.price.quantile(0.01)
2  p99 = df.price.quantile(0.99)
3  print(p1,p99)
4
5  df[(df.price >= p1) & (df.price <= p99)].describe()
6  df = df[(df.price >= p1) & (df.price <= p99)]
```



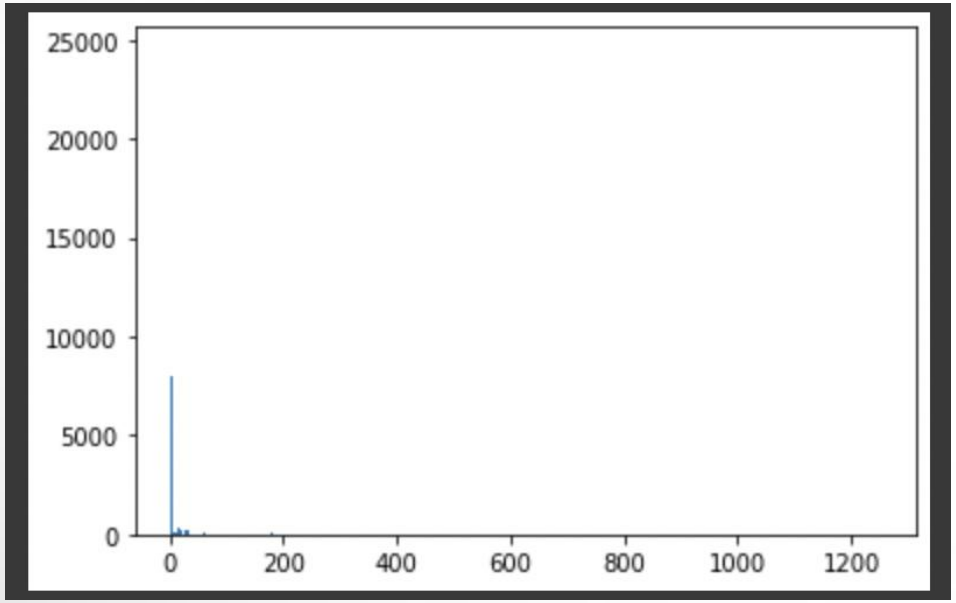
before

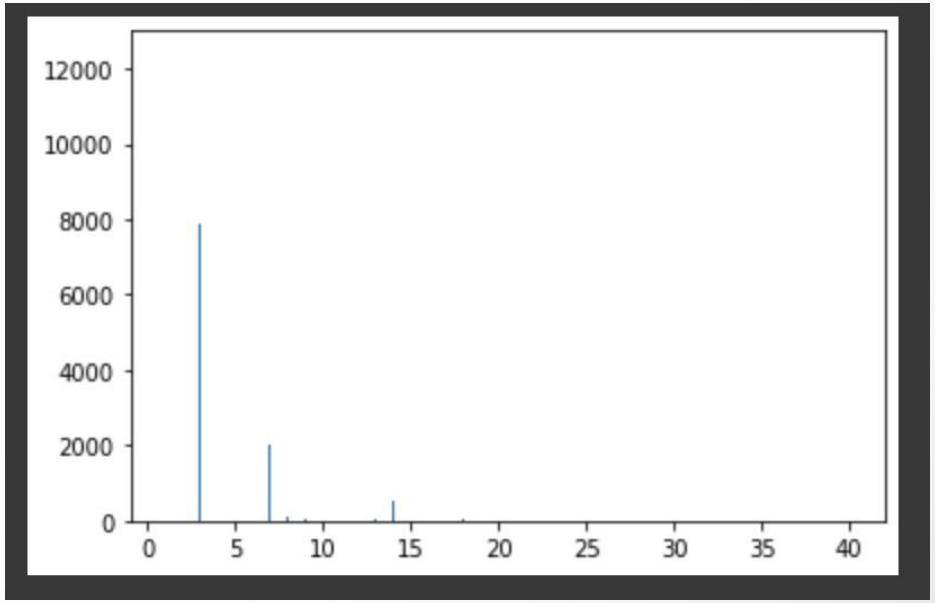


after

clean up minimum nights data as the original data is skewed

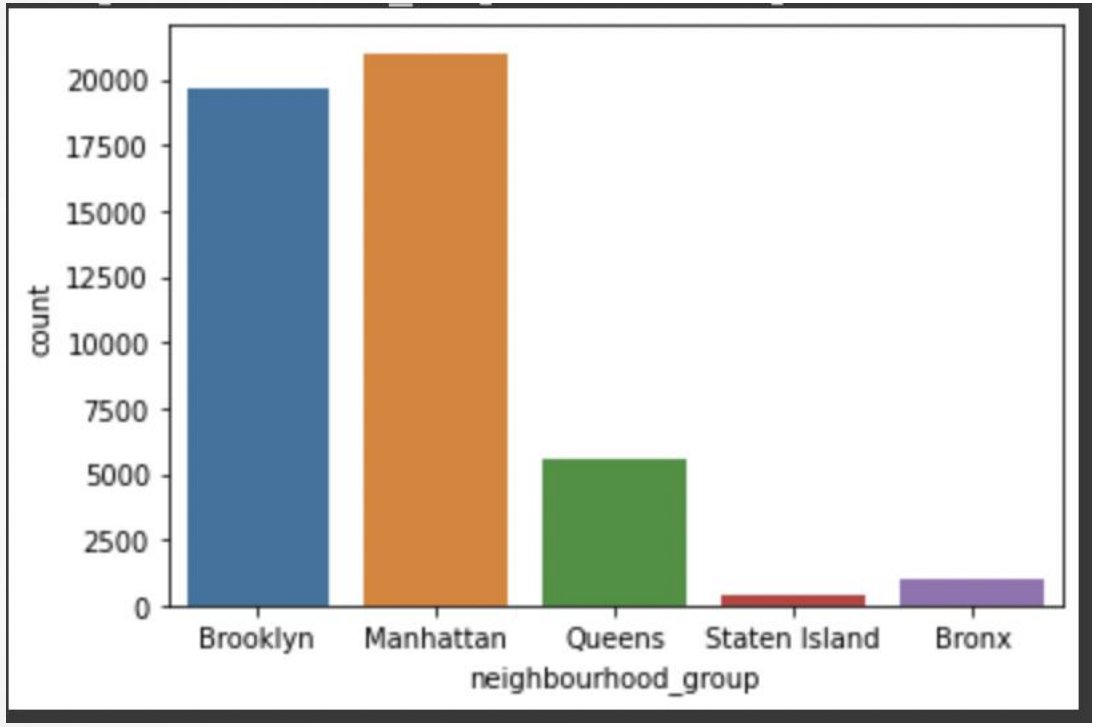
```
p1_min_night = df.minimum_nights.quantile(0.01)
p99_min_night = df.minimum_nights.quantile(0.99)
print(p1_min_night,p99_min_night )

df = df[(df.minimum_nights >= p1_min_night) & (df.minimum_nights <= p99_min_night)]
```



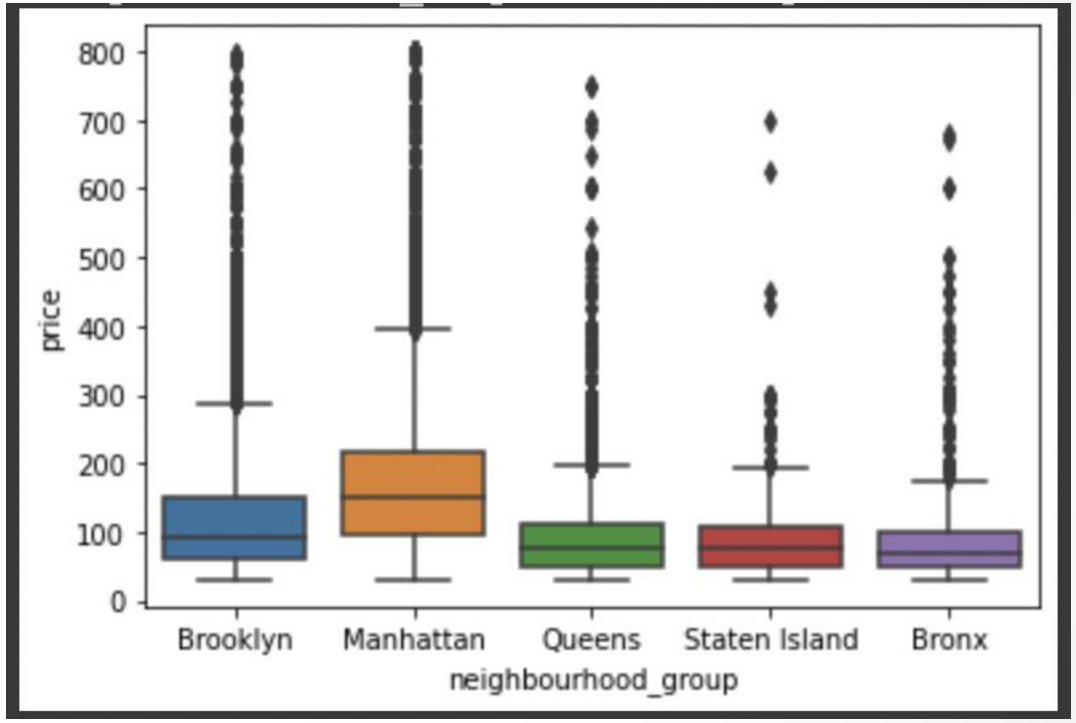


Data Exploring

Neighbourhood group



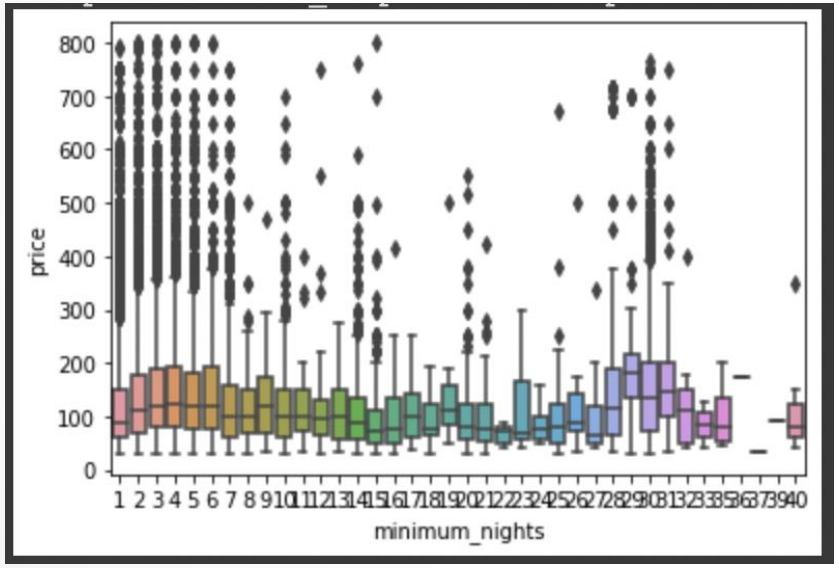
number of listings across neighbourhood group is skewed, staten island and bronx have way less number of listings compared to manhatten and brooklyn



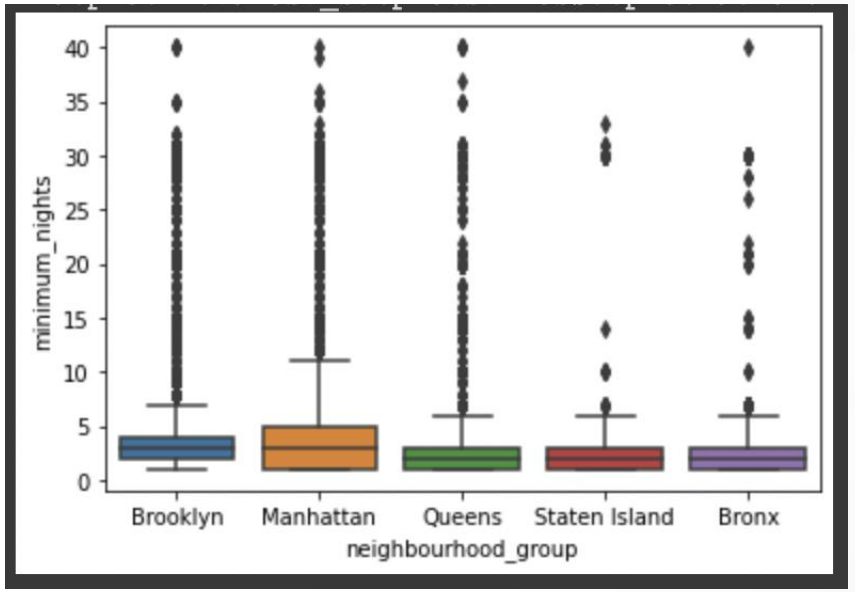
distribution of price across neighbourhood group, manhatten has the highest mean price

Data Exploring

Minimum number of nights



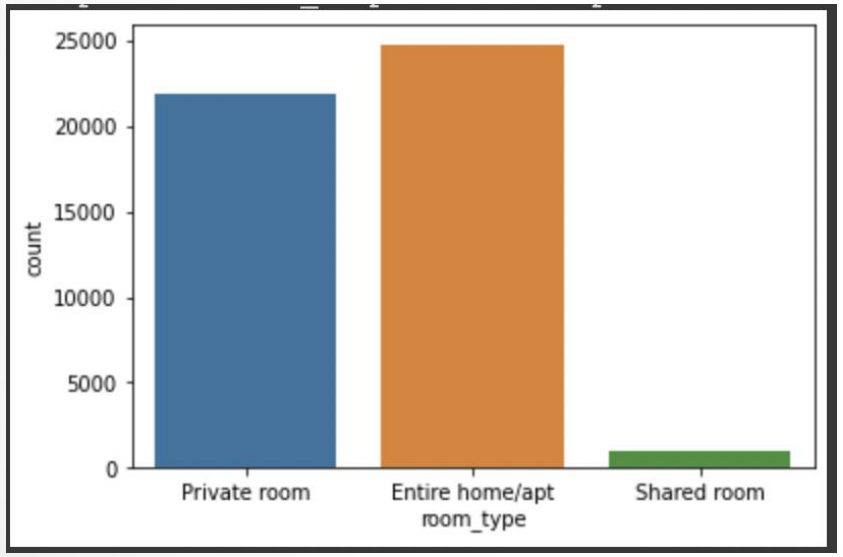
price generally on a downward trend with increase of minimum number of nights



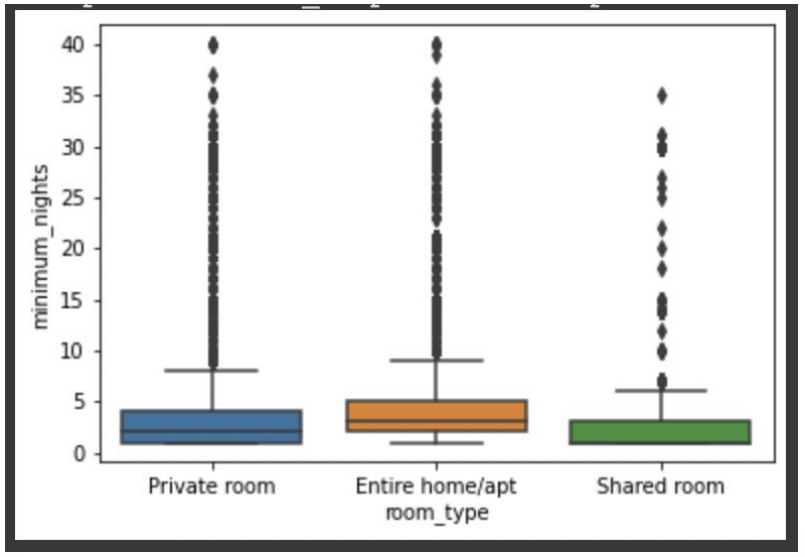
staten island has less minimum number of nights compared to other neighbourhood groups due to its remote location and more attractive tourist attractions in manhattan, brooklyn. queens, bronx

Data Exploring

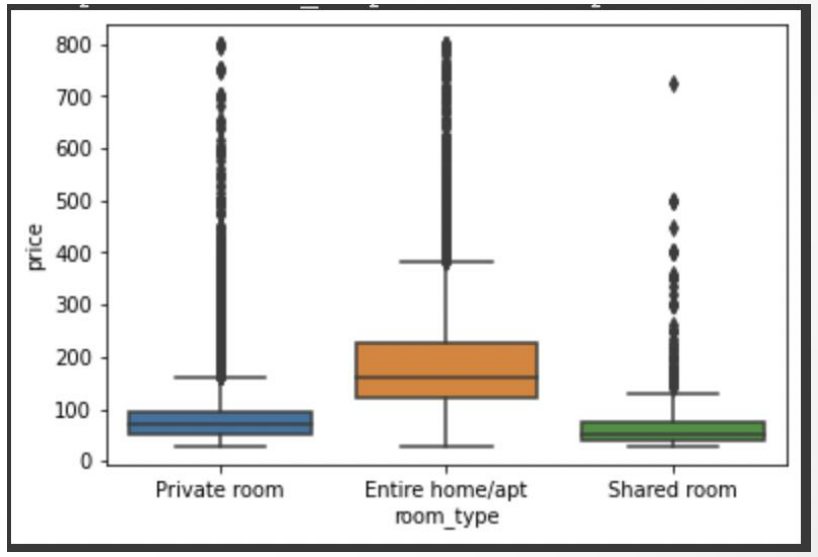
Room type



shared room has significantly less number of listings compared to private room or entire home



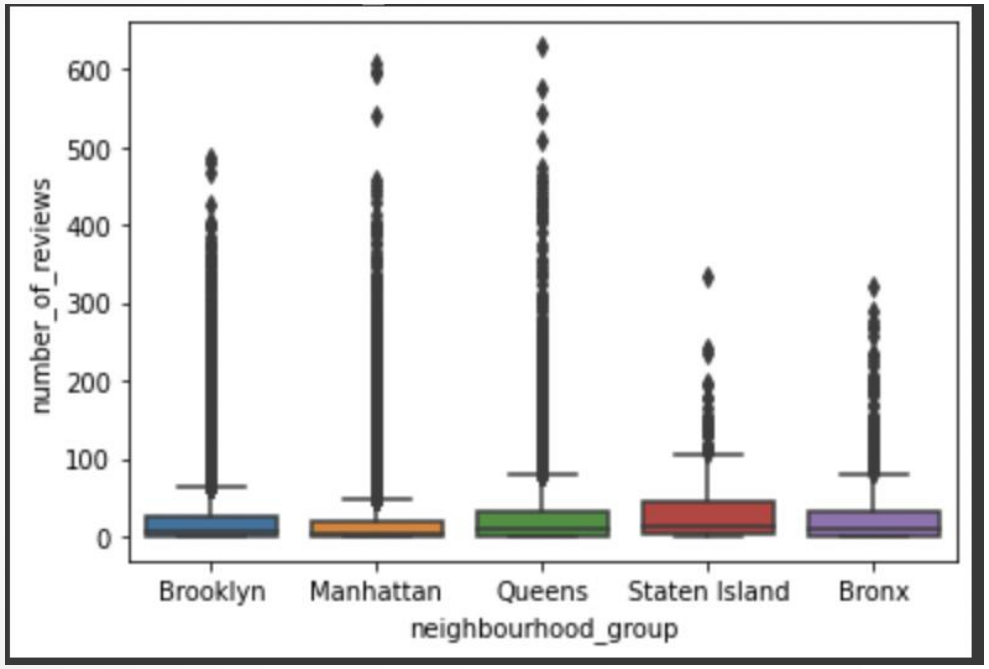
shared room generally has less minimum number of nights compared to private room or entire home

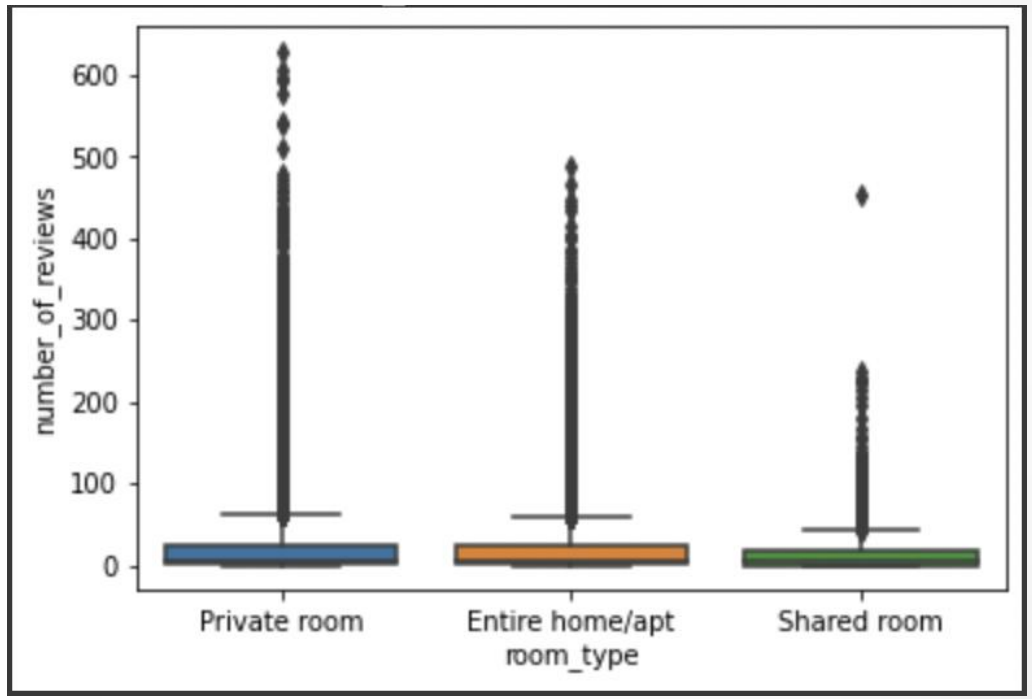


entire home is the most expensive, followed by private room and shared room

Data Exploring

Number of reviews

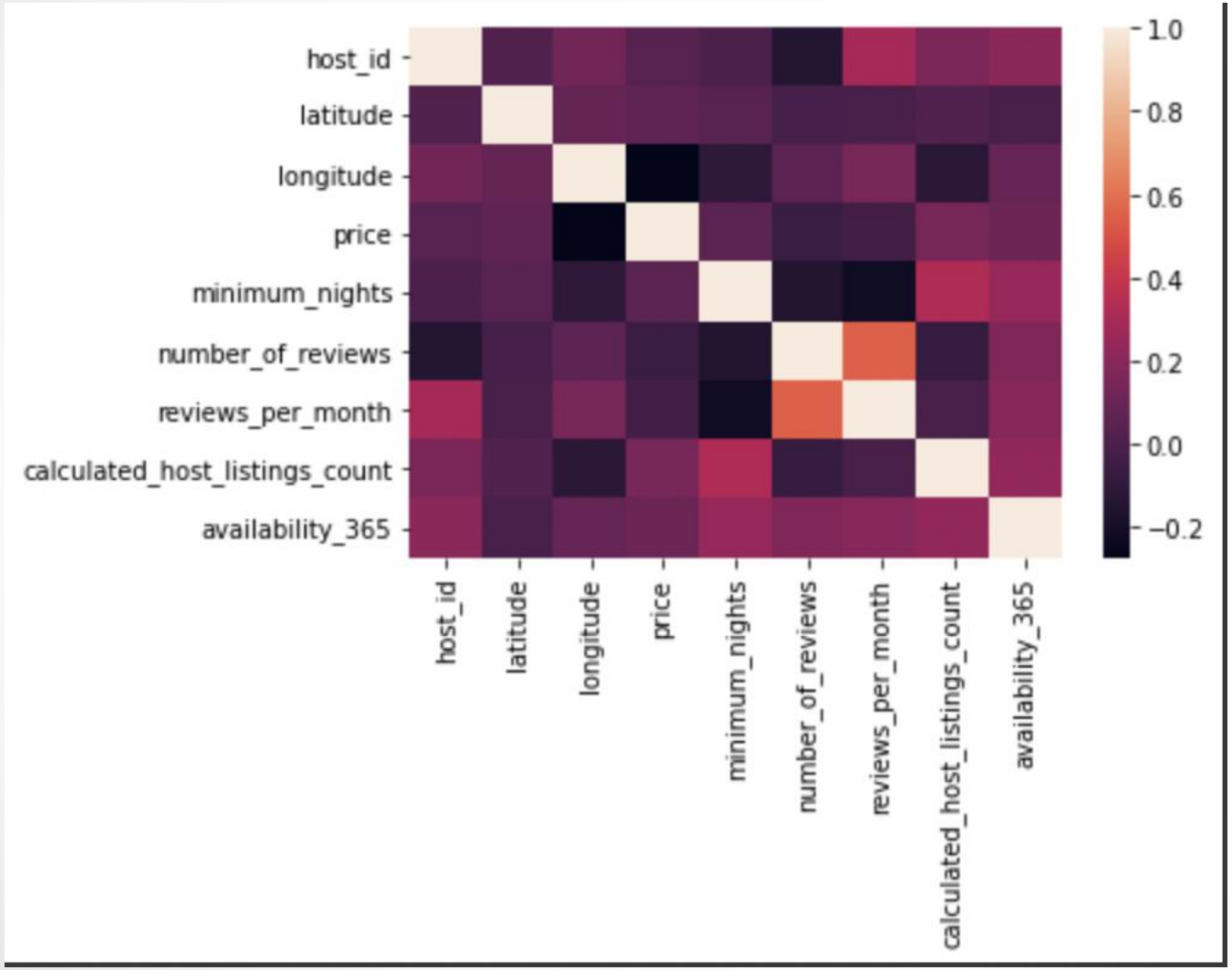




Queen has the highest number of reviews, followed by Manhattan. Staten island has the least due to low number of listings of airbnb

Private room has the highest number of reviews, followed by entire home and shared room

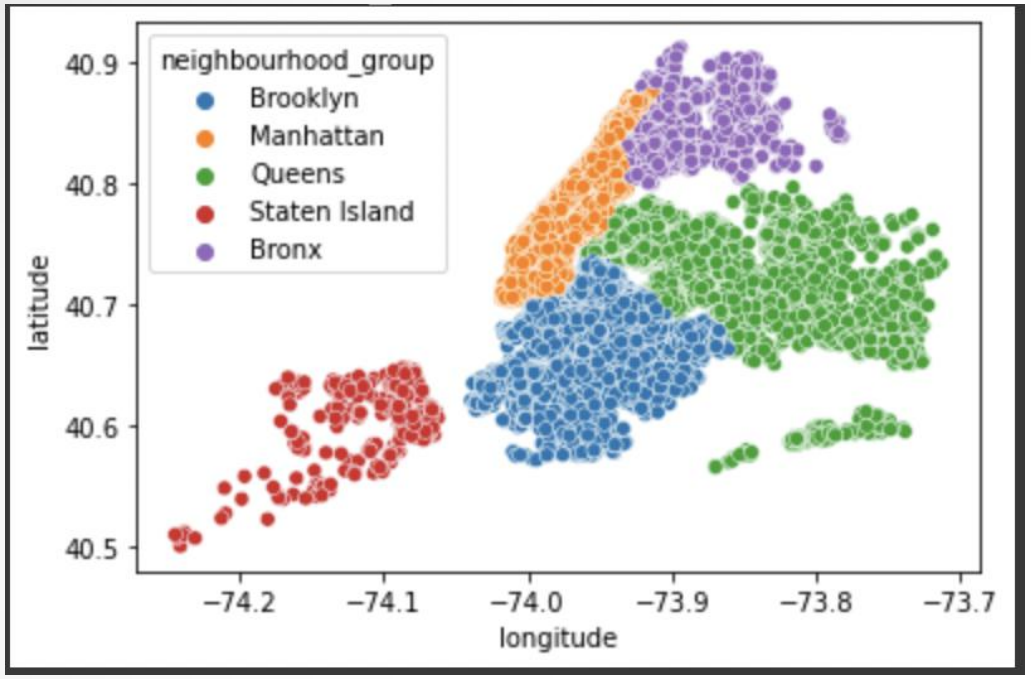
Data Exploring

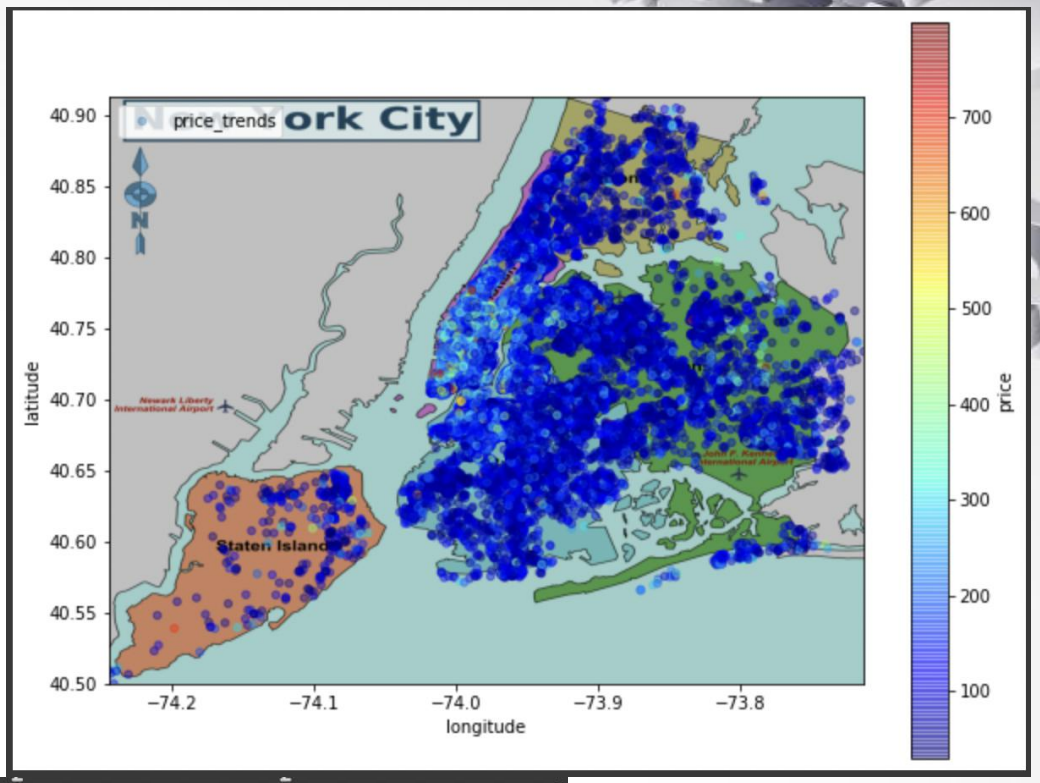


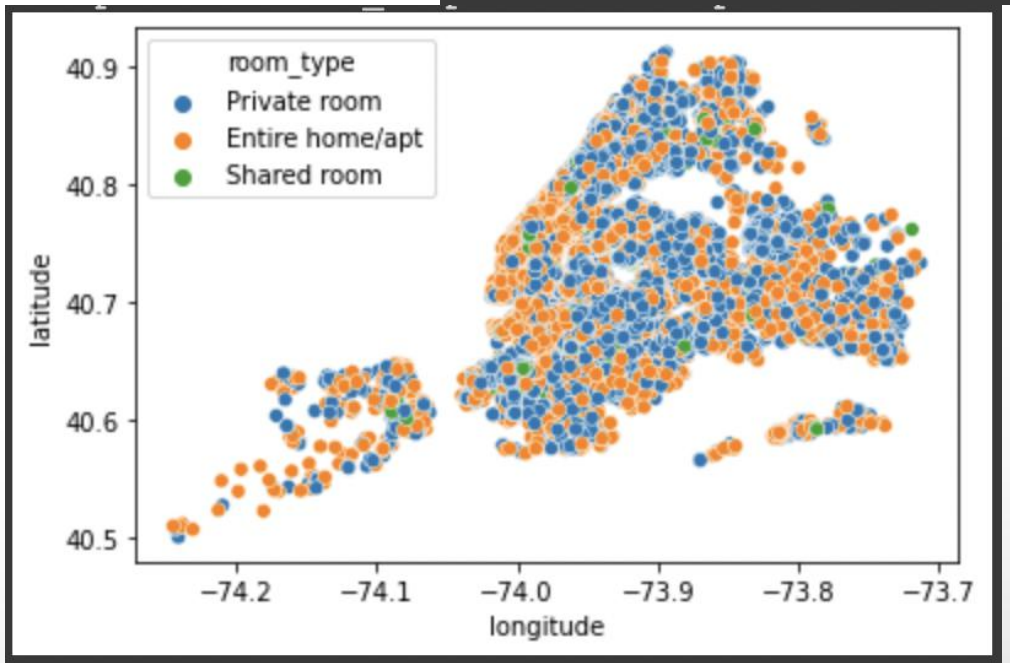
Correlation of price and other features

- no strong correlation of price with other features

Data Exploring







Data Modelling

Data pre-processing: z score

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
huber	Huber Regressor	40.2705	4348.3528	65.9101	0.4221	0.3898	0.2997	5.715
lightgbm	Light Gradient Boosting Machine	39.2662	3657.8809	60.4537	0.5137	0.3641	0.3151	0.429
rf	Random Forest Regressor	39.9117	3723.8836	60.9877	0.5050	0.3695	0.3223	32.748
ridge	Ridge Regression	42.6654	4096.8635	63.9820	0.4553	0.4223	0.3529	0.094
lr	Linear Regression	45.6774	26016.2705	134.5149	-2.4692	0.4344	0.3772	0.273
lasso	Lasso Regression	45.0090	4618.5216	67.9343	0.3860	0.4213	0.3814	0.086
dt	Decision Tree Regressor	54.1378	7276.4052	85.2879	0.0312	0.4973	0.4277	0.546

```
HuberRegressor(alpha=0.0001, epsilon=1.35, fit_intercept=True, max_iter=100,
               tol=1e-05, warm_start=False)
```

huber model is selected based on the lowest MAPE generated

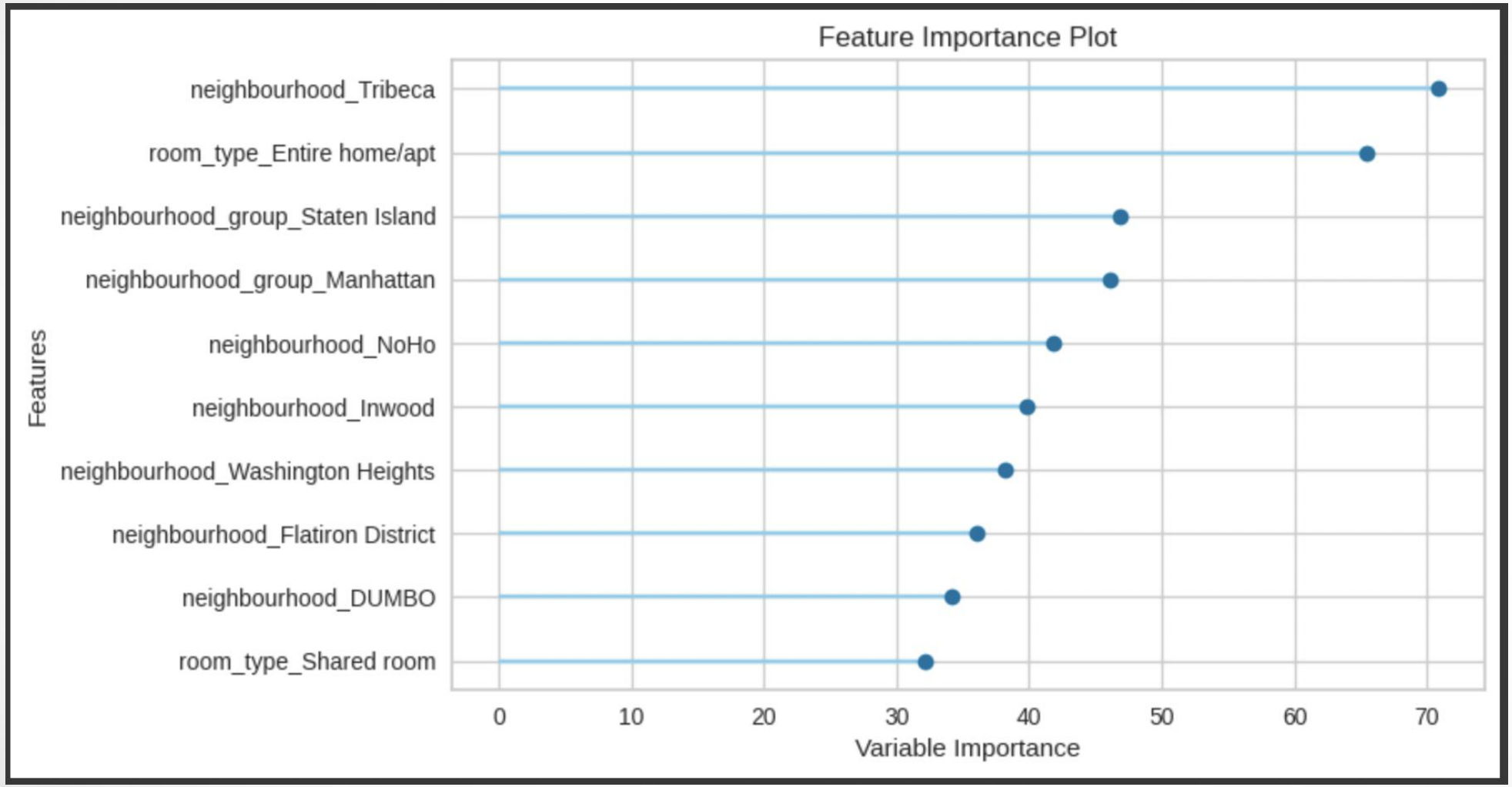
Data Modelling

```
[ ]  1  huber_zscore_tuned = tune_model(huber_zscore,
     2                          optimize='MAPE',
     3                          search_library='scikit-learn',
     4                          search_algorithm='random'
     5                          )
```

	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	44.2887	6735.3400	82.0691	0.3507	0.4069	0.3070
1	42.7686	5754.3020	75.8571	0.3771	0.3993	0.2994
2	42.9340	6160.4751	78.4887	0.3479	0.4068	0.2957
3	43.0407	6205.3800	78.7742	0.3590	0.4003	0.2943
4	42.0131	5879.4694	76.6777	0.3698	0.4004	0.2949
5	41.3756	5237.3275	72.3694	0.3868	0.3942	0.2939
6	44.2797	6516.1913	80.7229	0.3480	0.4147	0.3075
7	43.2525	6505.5954	80.6573	0.3458	0.4104	0.3026
8	42.6808	5855.0811	76.5185	0.3690	0.4054	0.3010
9	42.1657	5667.2722	75.2813	0.3717	0.4031	0.2968
Mean	42.8799	6051.6434	77.7416	0.3626	0.4042	0.2993
SD	0.8768	435.0105	2.8077	0.0135	0.0056	0.0048

Hyperparameter tuning reduces MAPE to 0.2993

Data Modelling



location and room type play the most important factors in determining the price

Data Modelling

```
[ ]  1  #check out the price here and compare
     2  user_request
```

```
neighbourhood_group                 Bronx
neighbourhood                   Mott Haven
latitude                           40.8079
longitude                          -73.924
room_type                  Entire home/apt
price                                  100
minimum_nights                           1
number_of_reviews                        2
last_review                     2019-07-07
reviews_per_month                        2
calculated_host_listings_count           1
availability_365                        40
year                                  2019
month                                   07
day                                     07
date_id                               1745
Name: 37368, dtype: object
```

actual price

the difference in price is around 10 dollar

predicted price

```
[ ]  1  predict_model(loaded_model, user_request)
```

nimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365	year	month	day	date_id	Label
1	2	2019-07-07	2	1	40	2019	07	07	1745	110.576903

Future Work

1. The number of listings in Staten Island and Bronx is too little compared to Queens, Brooklyn and Manhattan, this may cause inaccuracy in machine learning model

2. Gather Airbnb data after Covid-19 pandemic

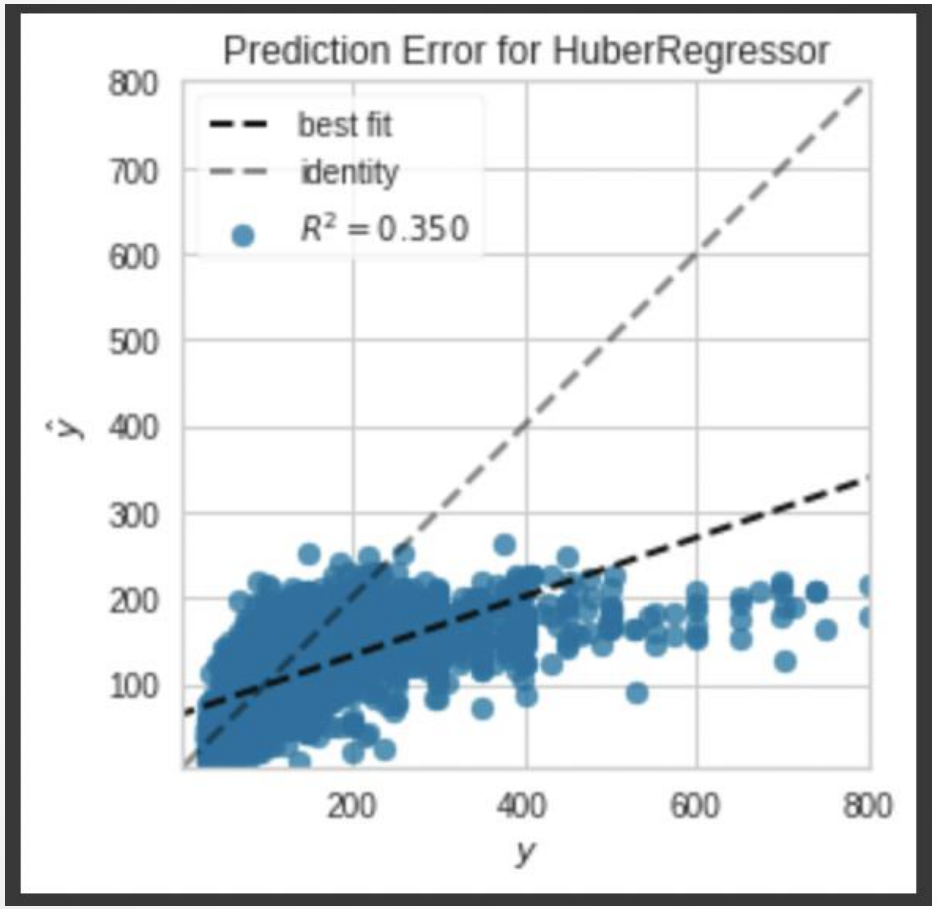
3. Entire home room type shuld provide information of number of rooms, otherwise, not fair to compare entire home with shared room and private room which are only 1 room.

4. Huber regression model provides the lowest MAPE, but MAPE of ~0.3 is stil considered high, this is due to skewed dataset with unbalanced counts of listings in different neighbourhood groups, as well as entire room which has more than 1 room being compared with shared room and private room

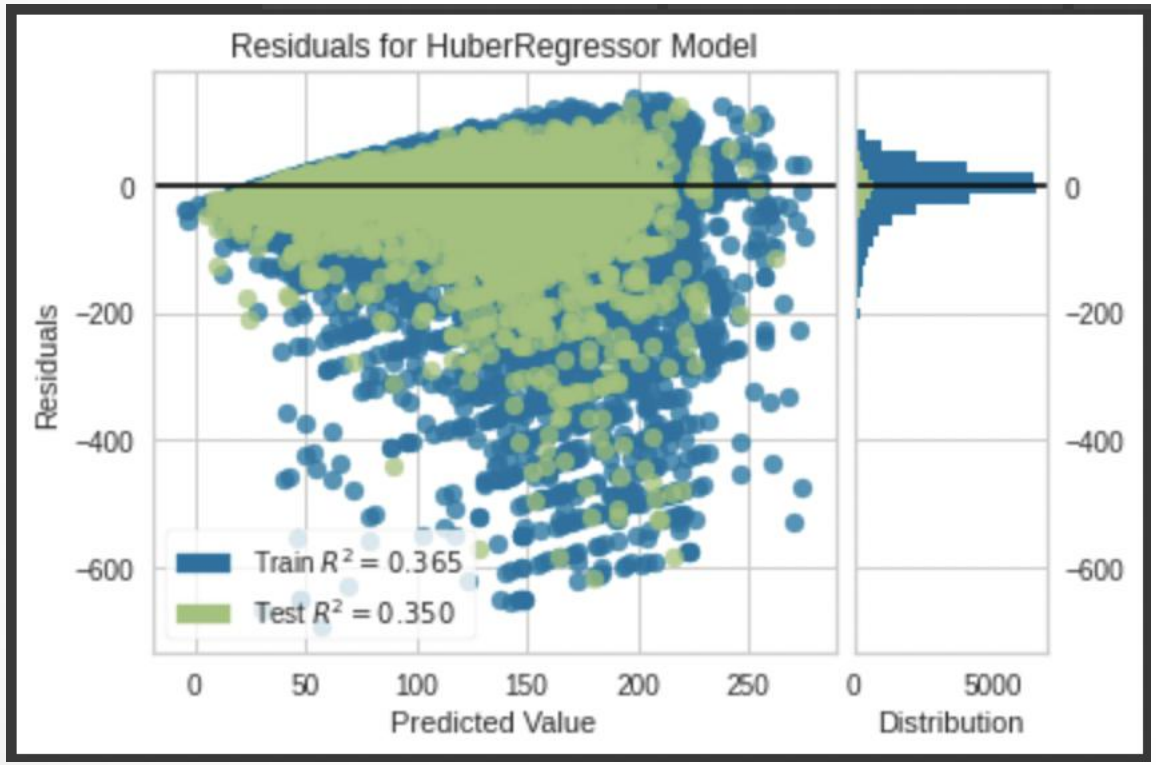
Thank you

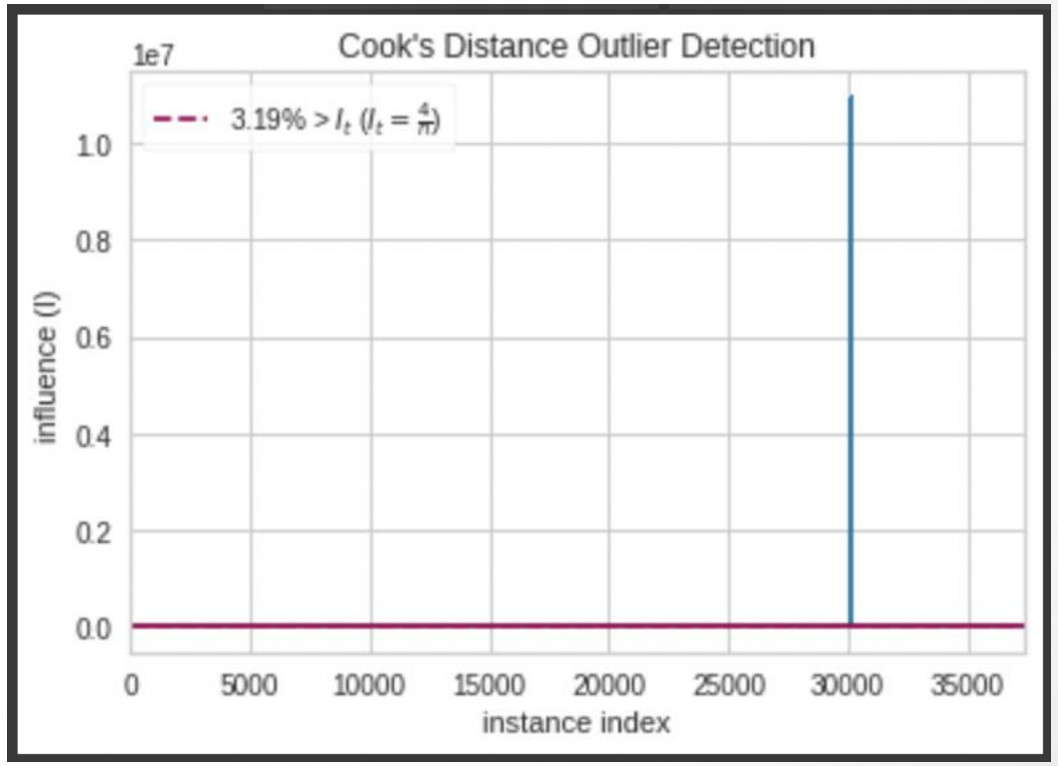
yaoxuan57@hotmail.com

Appendix


Prediction Error for HuberRegressor
best fit
identity
$R^2 = 0.350$
800
700
600
500
400
300
200
100
$\hat{y}$
200
400
600
800
y

Appendix


Residuals for HuberRegressor Model
Residuals
0
−200
−400
−600
Train $R^2 = 0.365$
Test $R^2 = 0.350$
0
50
100
150
200
250
Predicted Value
0
5000
Distribution


Cook's Distance Outlier Detection
1e7
3.19% > I_t ($I_t = \frac{4}{n}$)
influence (I)
1.0
0.8
0.6
0.4
0.2
0.0
0
5000
10000
15000
20000
25000
30000
35000
instance index