NYC AirBnb

Predicting Price

Photo by Andre Benz on Unsplash

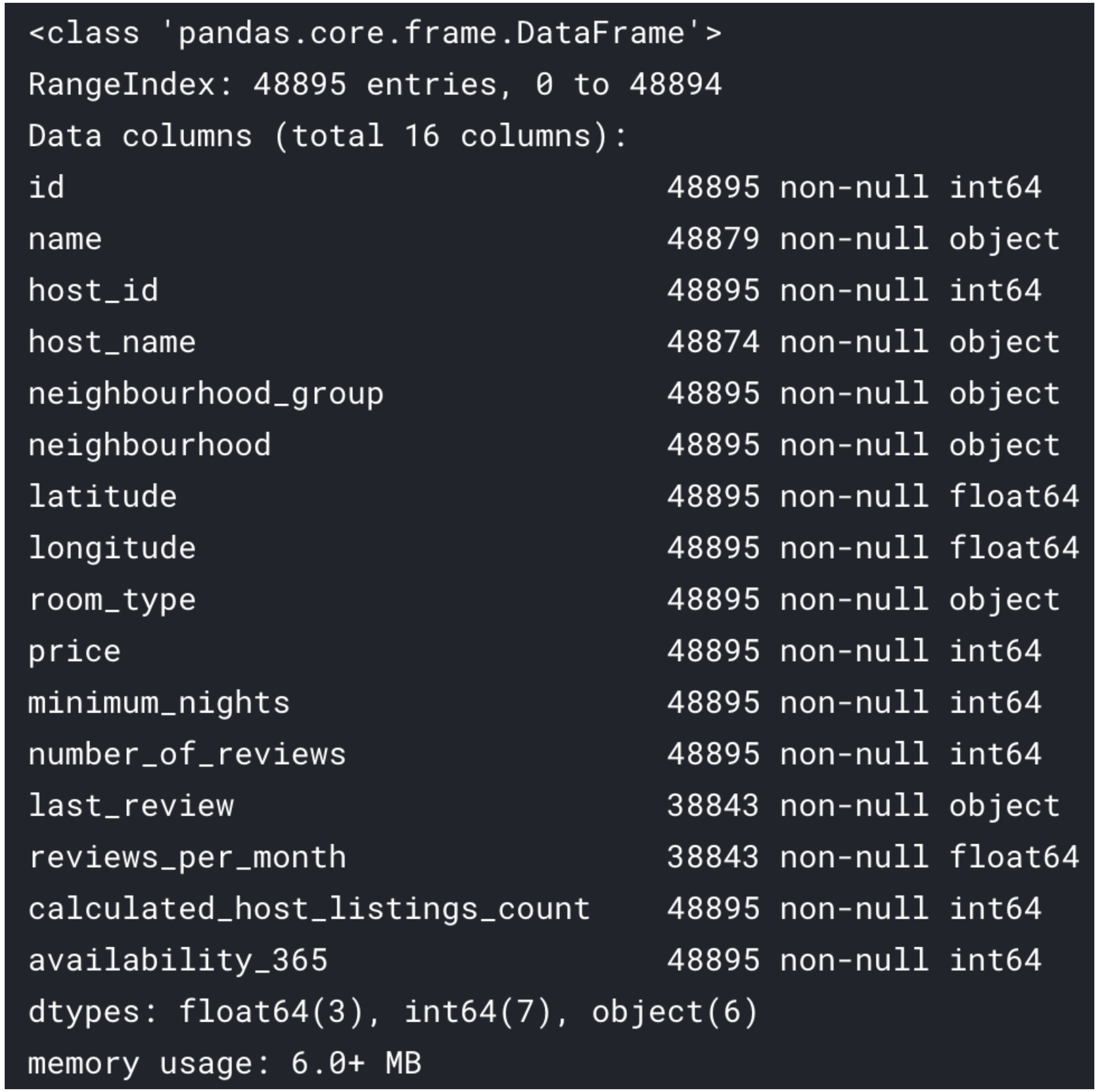
Exploratory Data Analysis

Dataset codebook

- **id**: listing ID
- **name**: name of the listing
- **host_id**: host ID
- **host_name**: name of the host
- **neighbourhood_group**: location
- **neighbourhood**: area
- **latitude**: latitude coordinates
- **longitude**: longitude coordinates
- **room_type**: listing space type
- **price**: price in dollars
- **minimum_nights**: amount of nights minimum
- **number_of_reviews**: number of reviews
- **last_review**: latest review
- **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

Exploratory Data Analysis

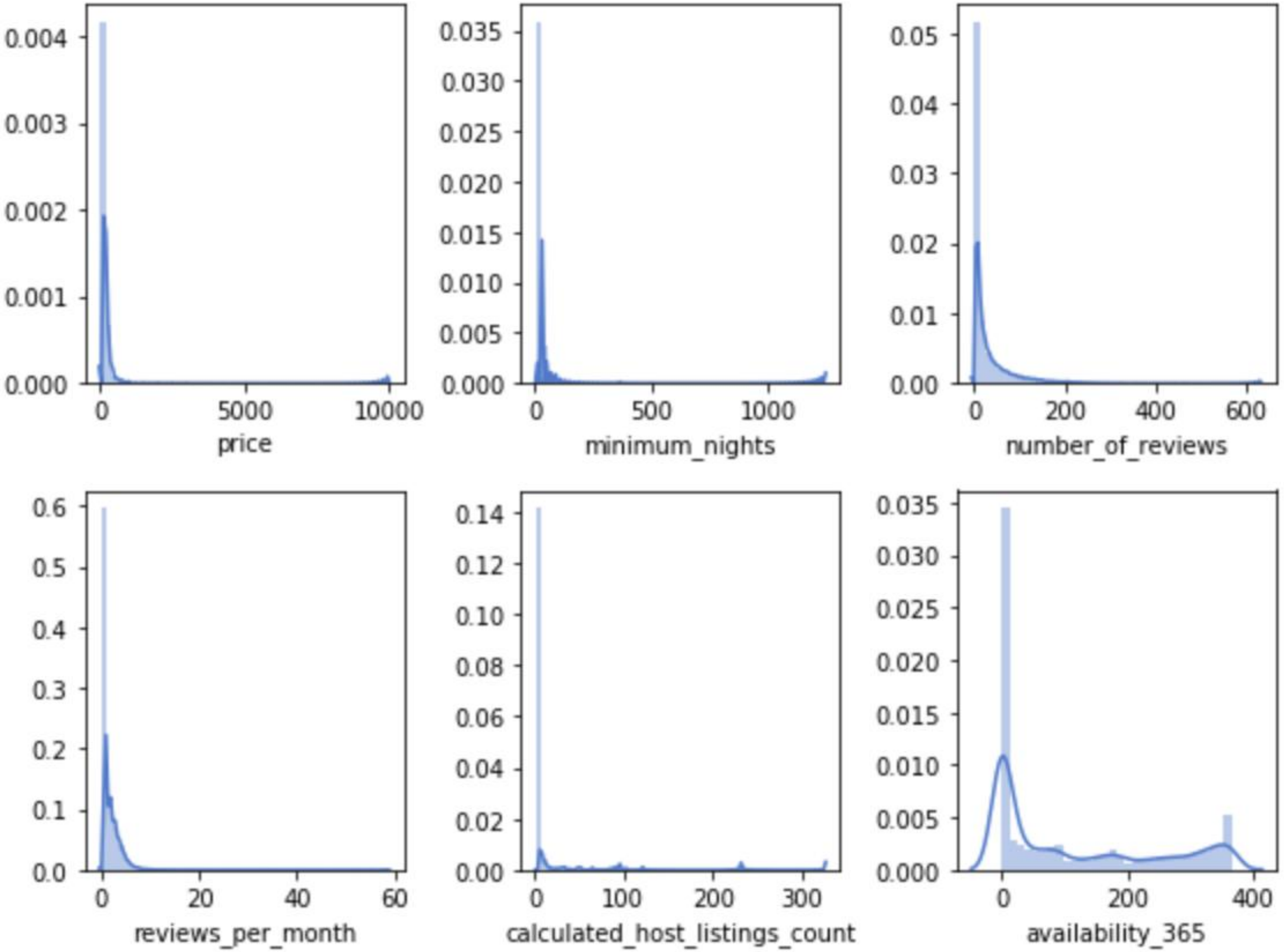
Visualize the shape of the distribution



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 16 columns):
id                                  48895 non-null int64
name                                48879 non-null object
host_id                             48895 non-null int64
host_name                           48874 non-null object
neighbourhood_group                 48895 non-null object
neighbourhood                       48895 non-null object
latitude                            48895 non-null float64
longitude                           48895 non-null float64
room_type                           48895 non-null object
price                               48895 non-null int64
minimum_nights                      48895 non-null int64
number_of_reviews                   48895 non-null int64
last_review                         38843 non-null object
reviews_per_month                   38843 non-null float64
calculated_host_listings_count      48895 non-null int64
availability_365                    48895 non-null int64
dtypes: float64(3), int64(7), object(6)
memory usage: 6.0+ MB
```

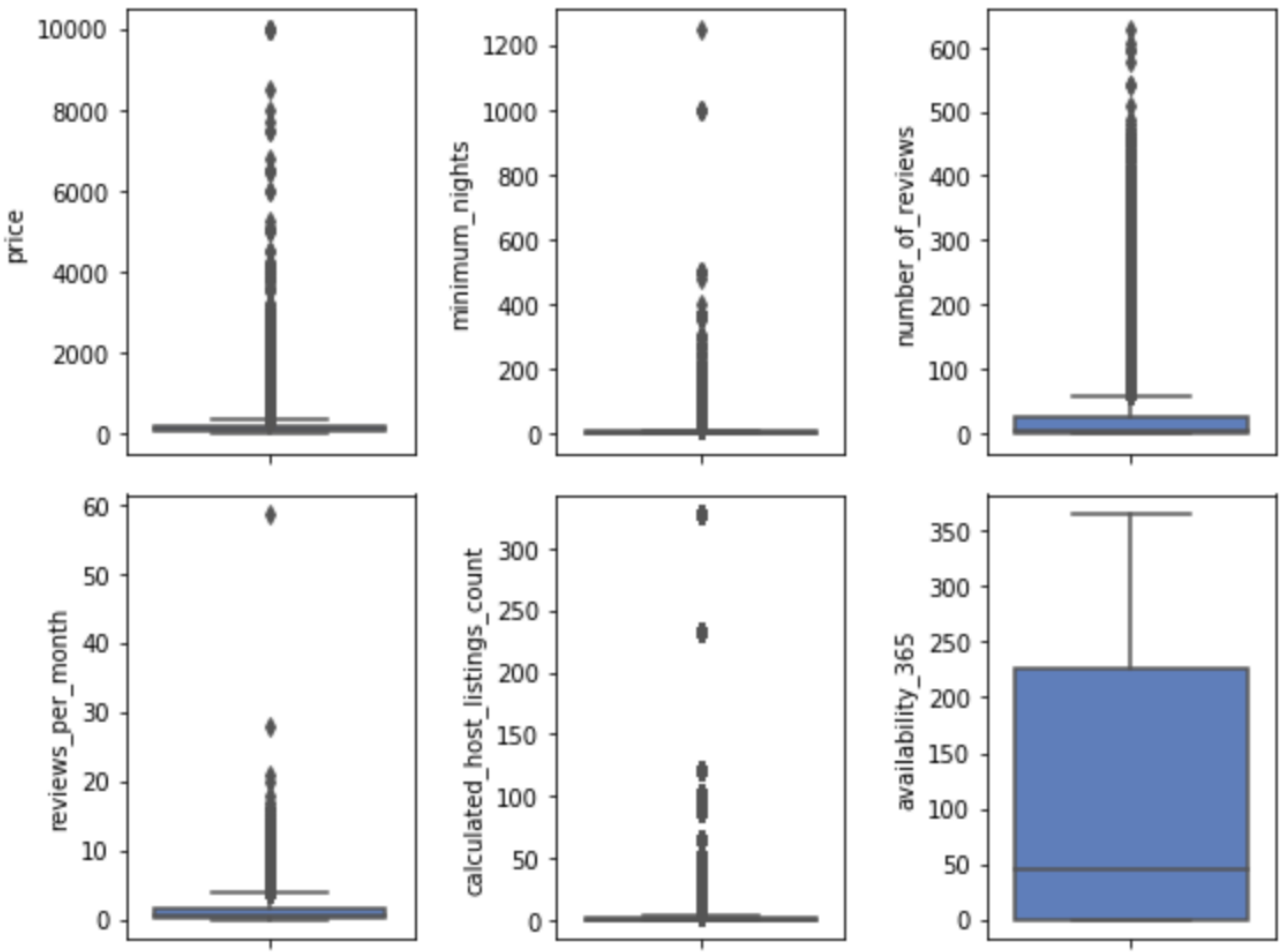
Exploratory Data Analysis

Visualize the shape of continuous data



Exploratory Data Analysis

Visualize the shape of continuous data



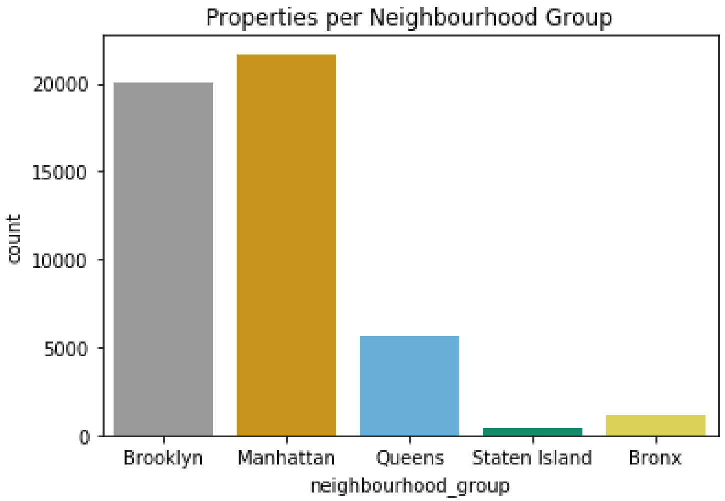
Exploratory Data Analysis

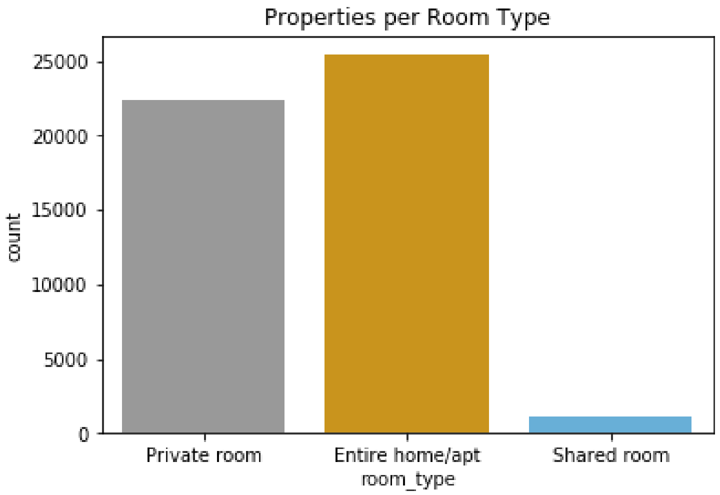
Visualize the shape of continuous data

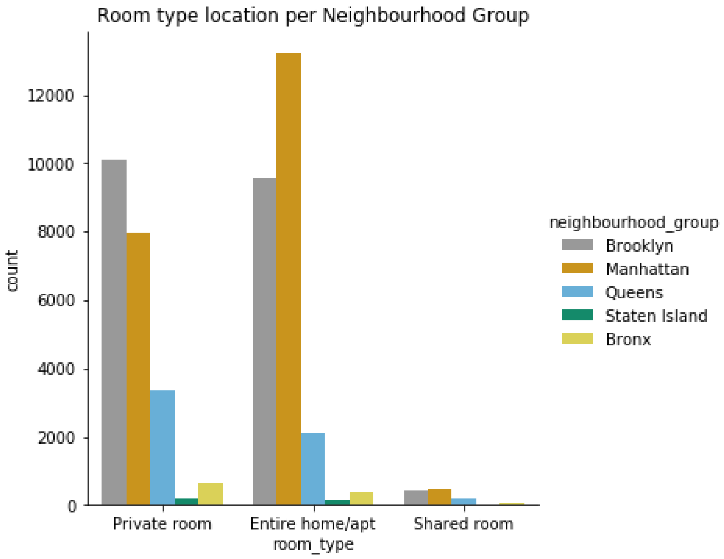
- Most features are left-skewed.
- I used a log transform before building the model.

Exploratory Data Analysis

Visualize the categorical features

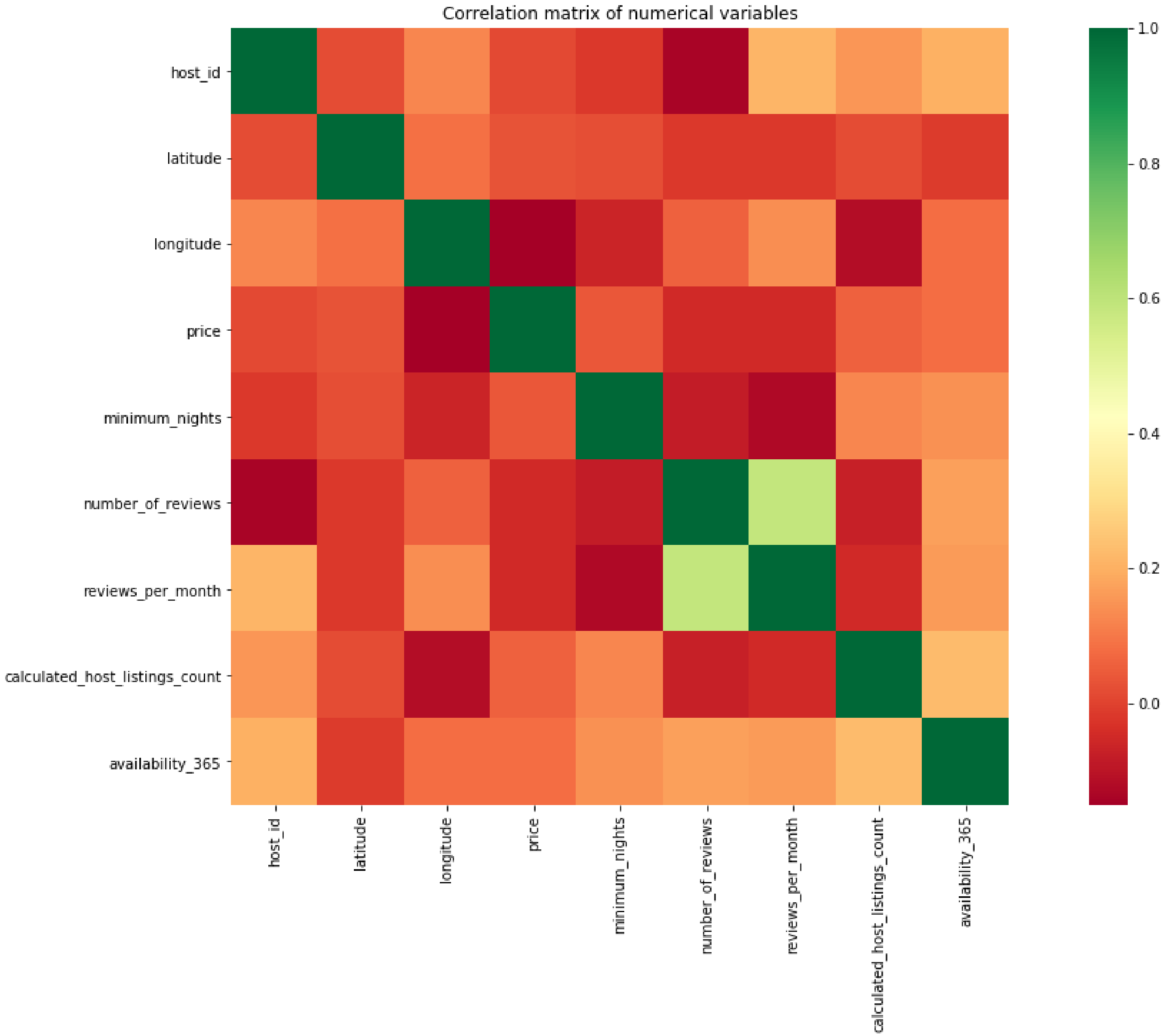

Properties per Neighbourhood Group


Properties per Room Type


Room type location per Neighbourhood Group

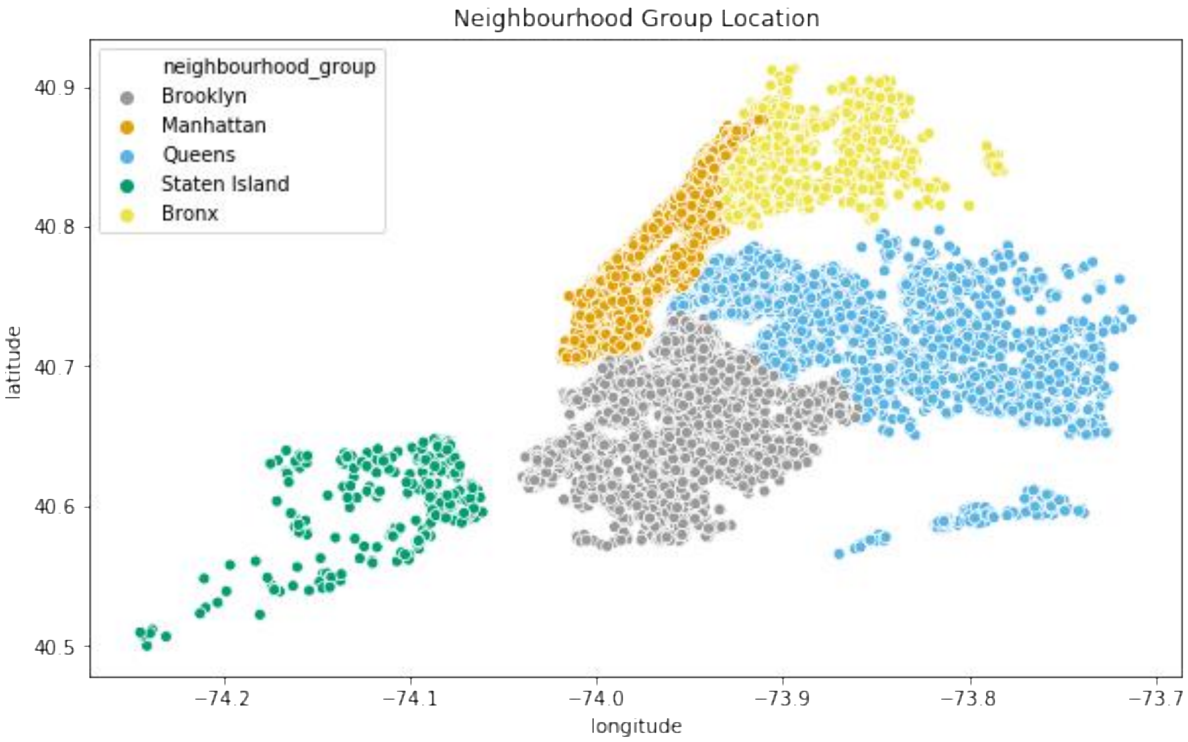
Data Visualization

Correlation between features

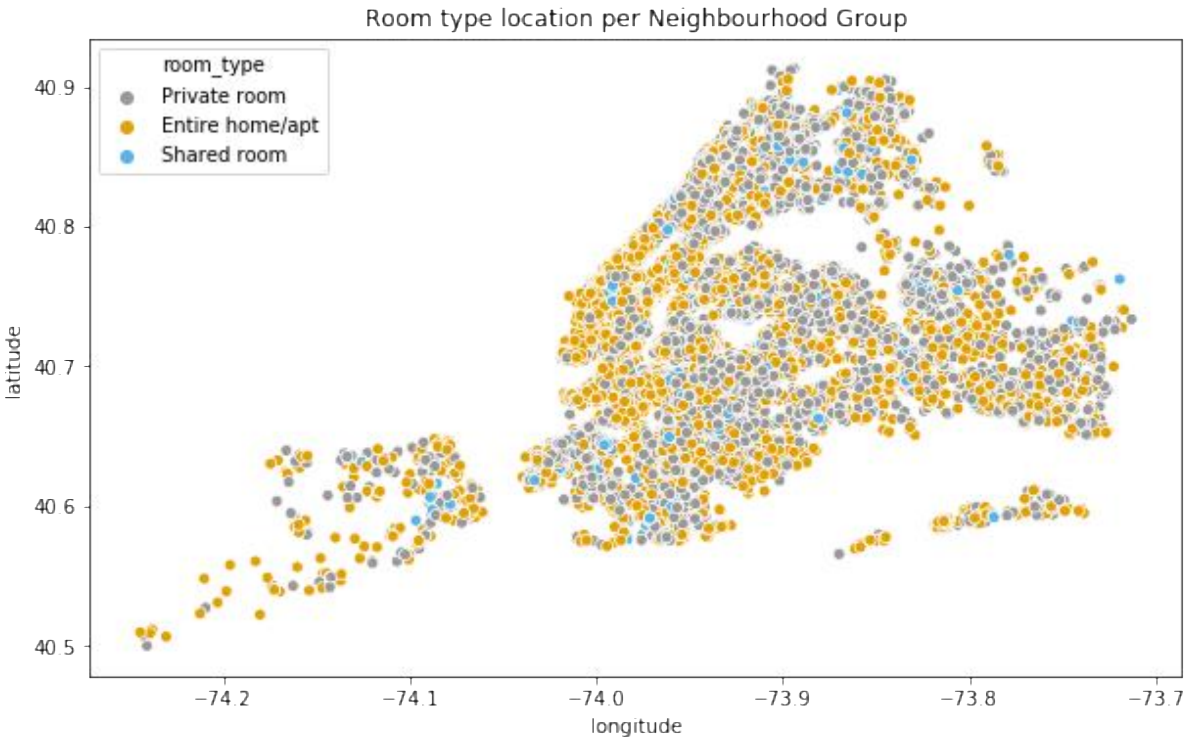


Data Visualization

Neighbourhood group and room type



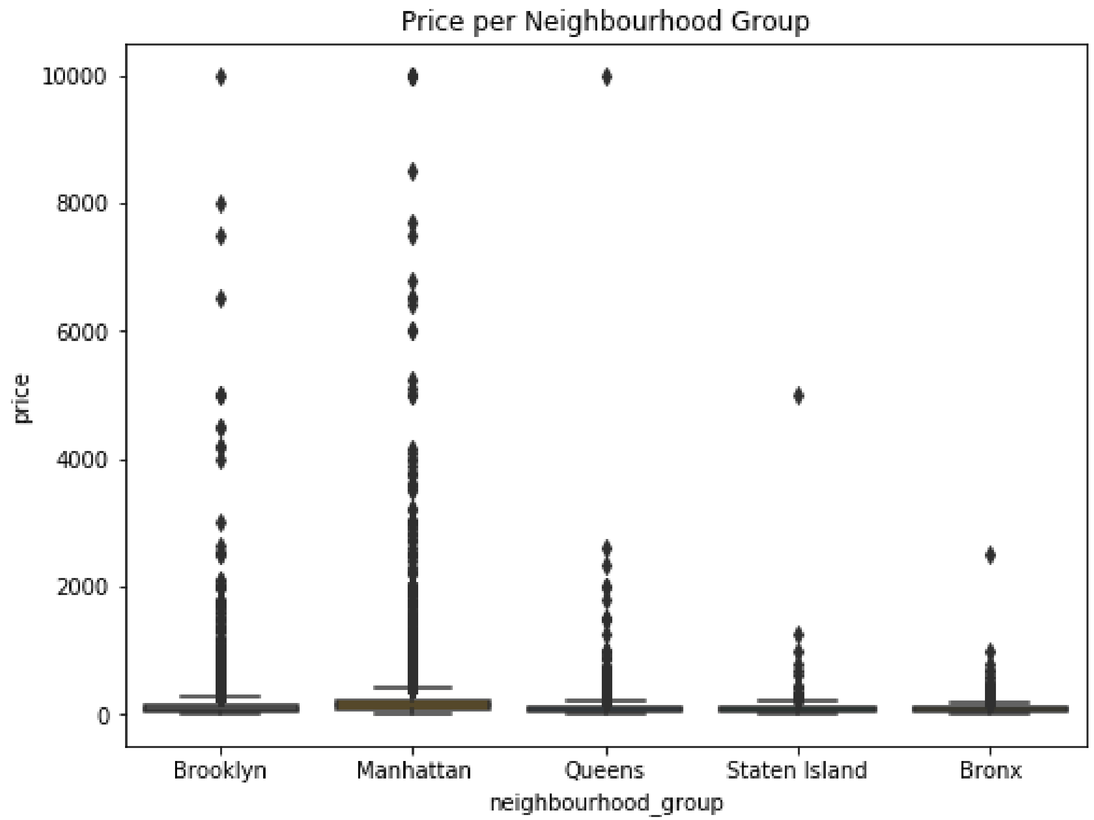
Neighbourhood Group Location

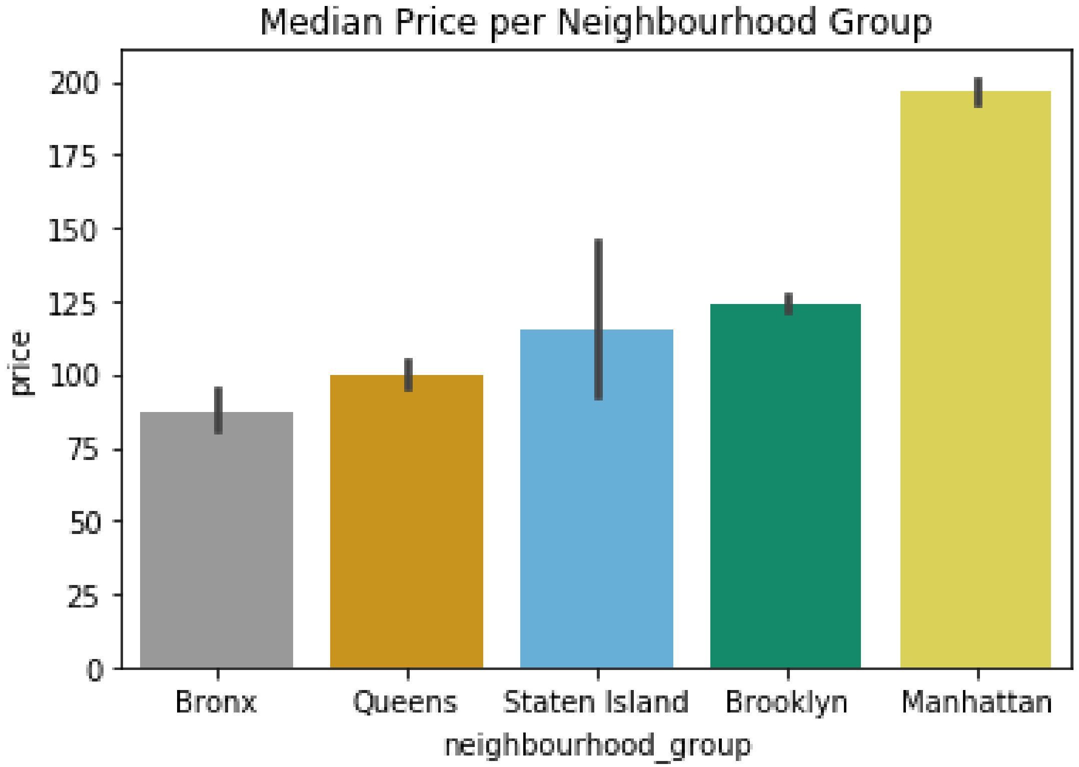


Room type location per Neighbourhood Group

Data Visualization

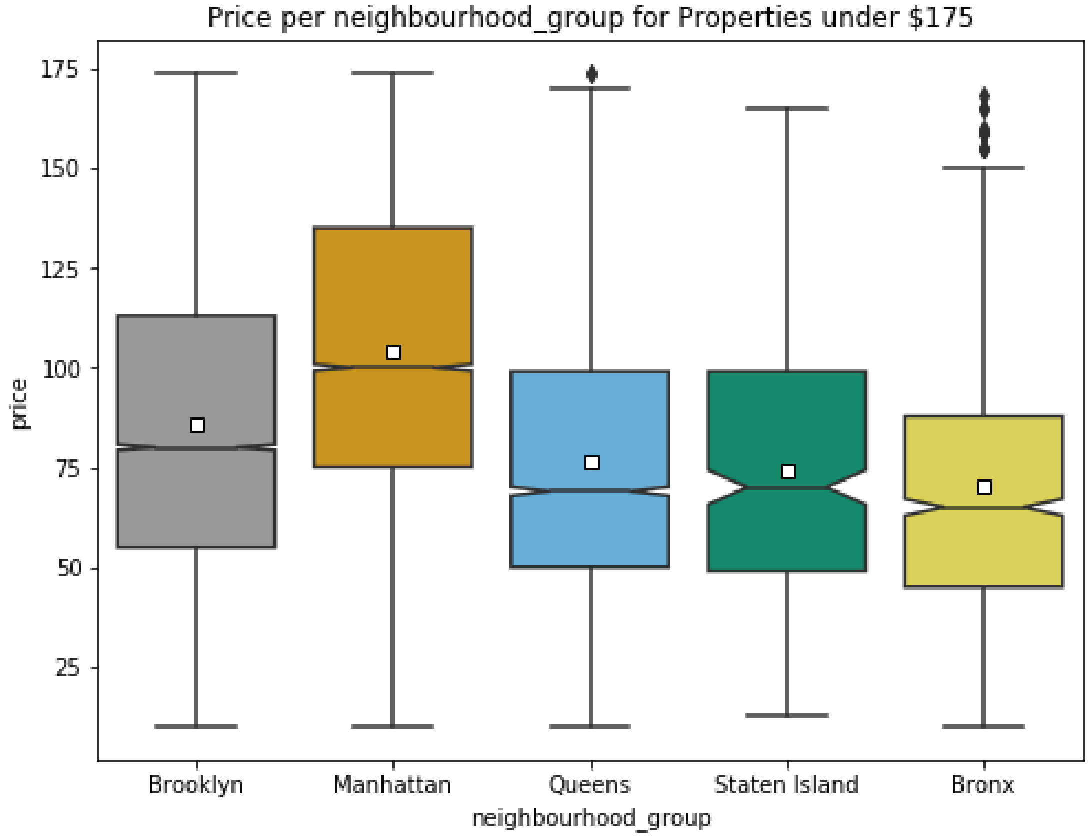
Neighbourhood impact on price

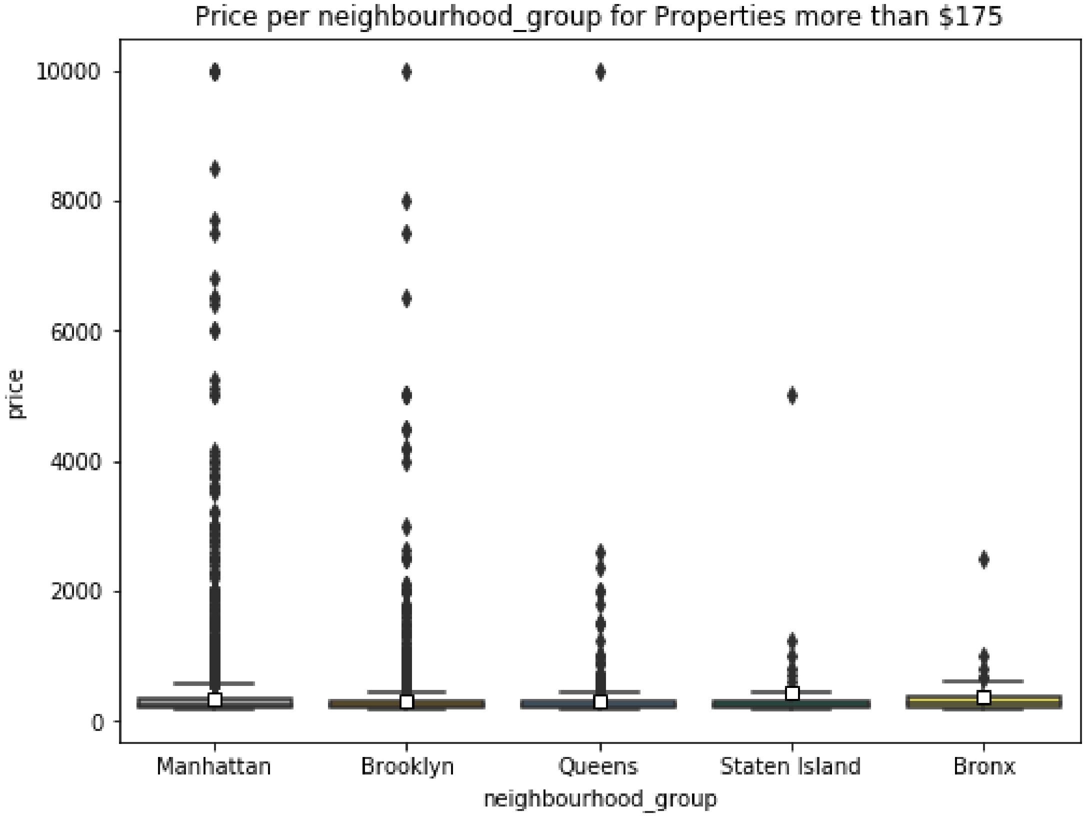




Data Visualization

Neighbourhood impact on price

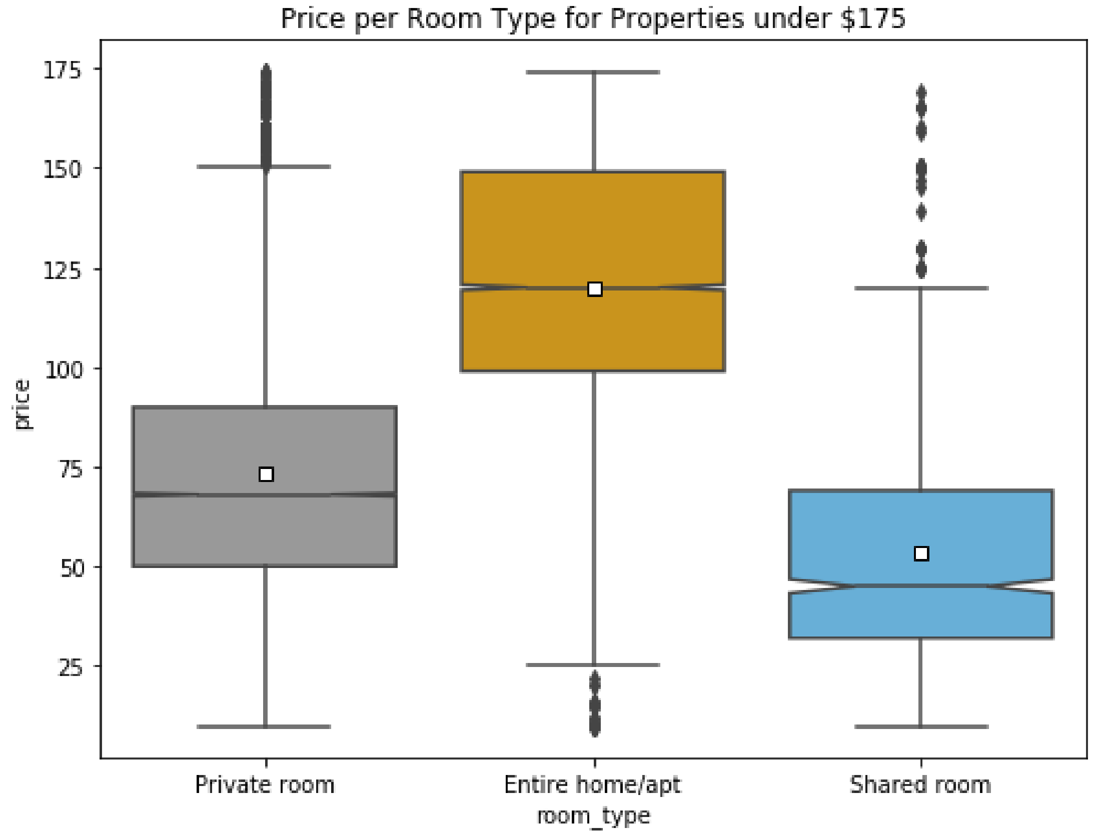


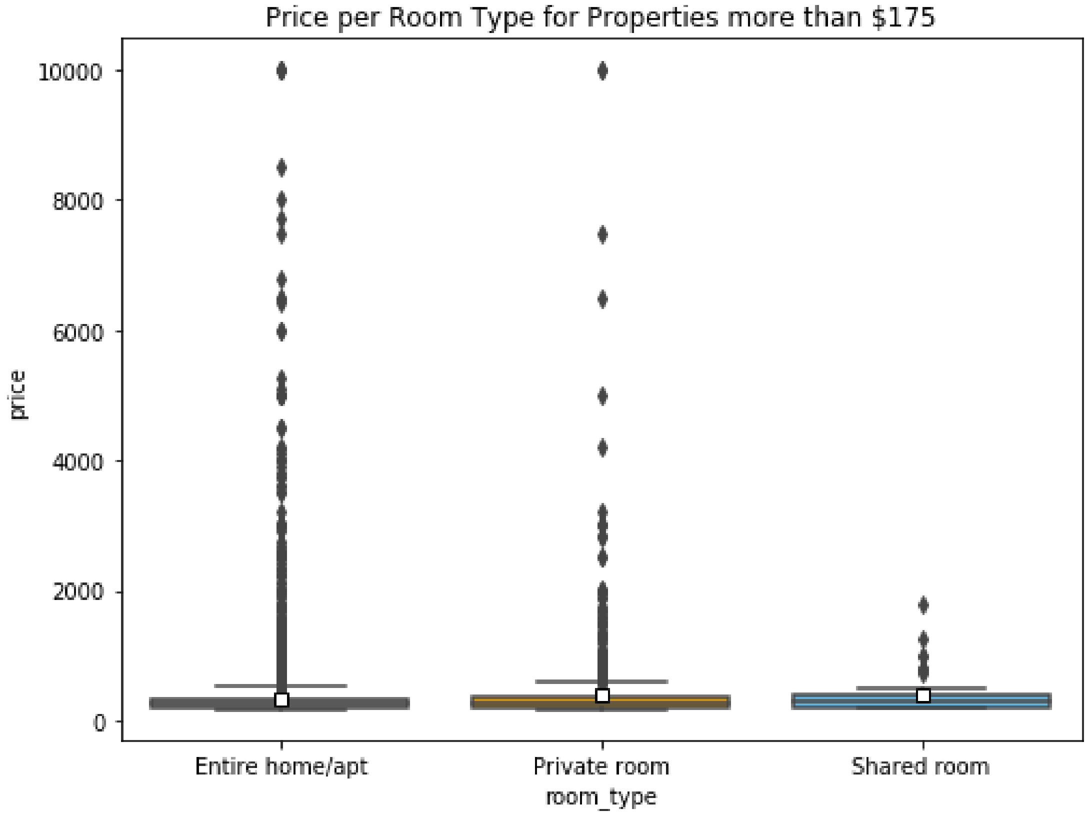


For the rest of the analysis, the dataset will be split between lows and higher prices

Data Visualization

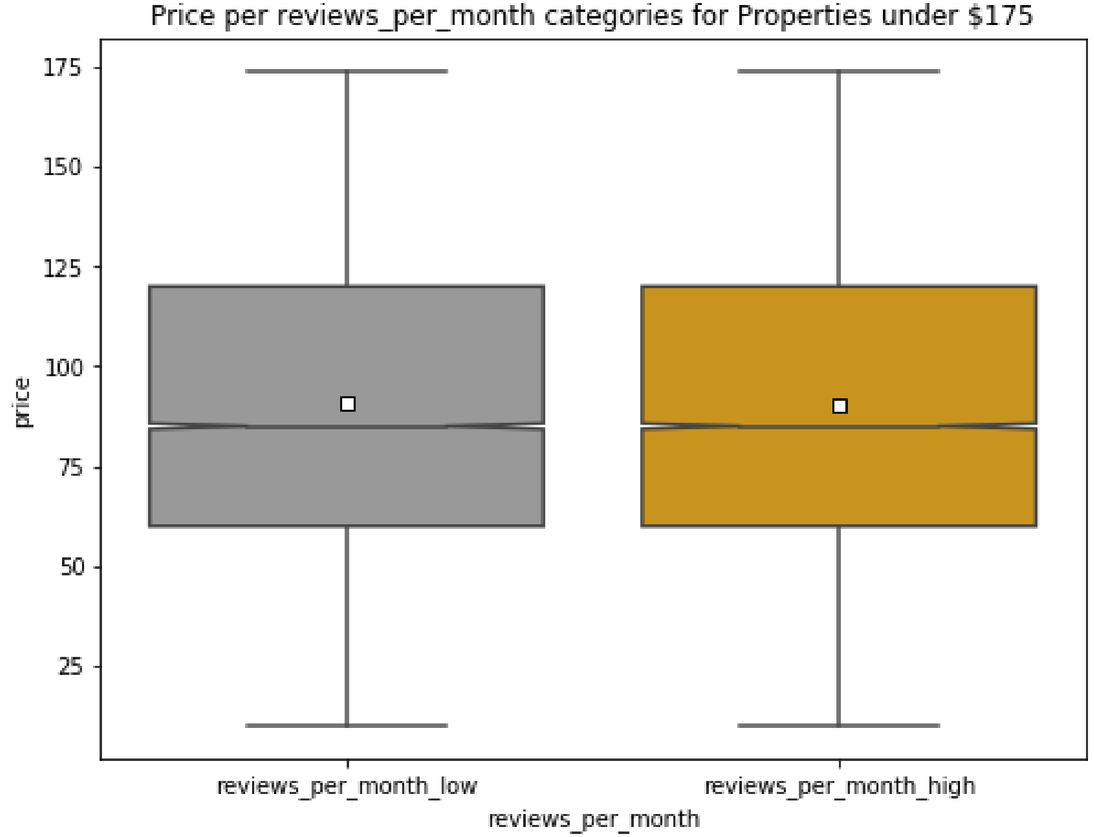
Room types impact on price

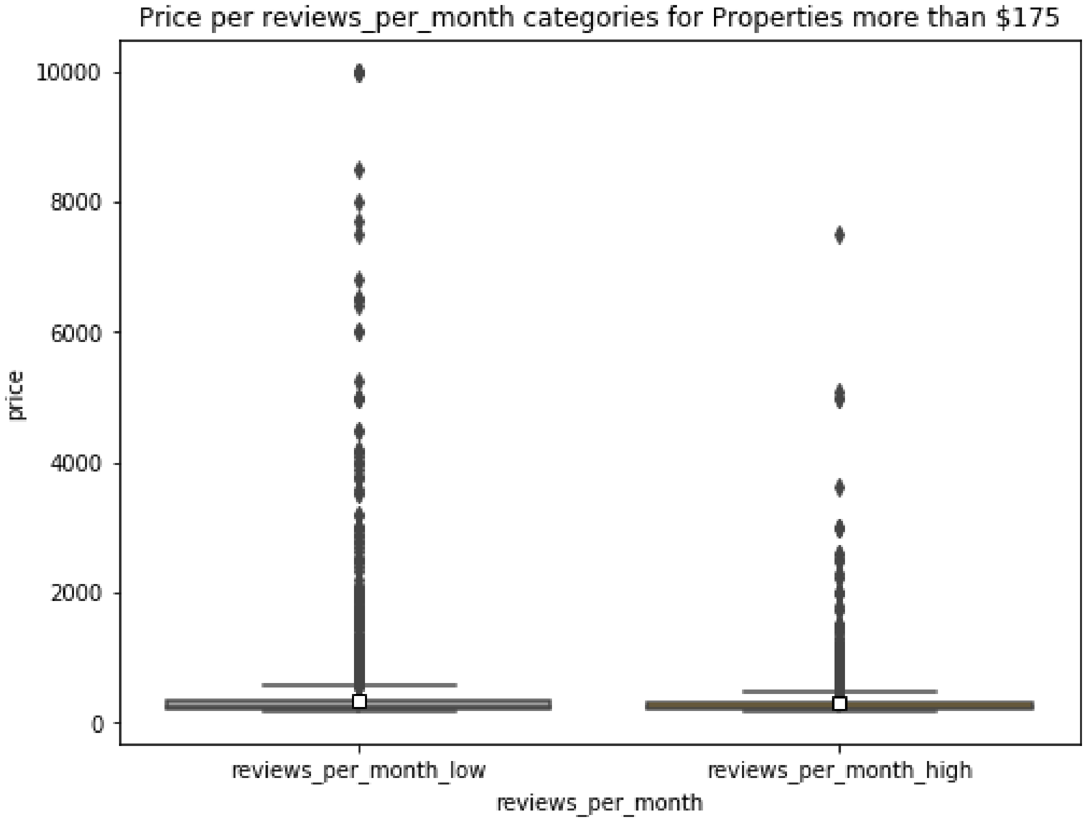




Data Visualization

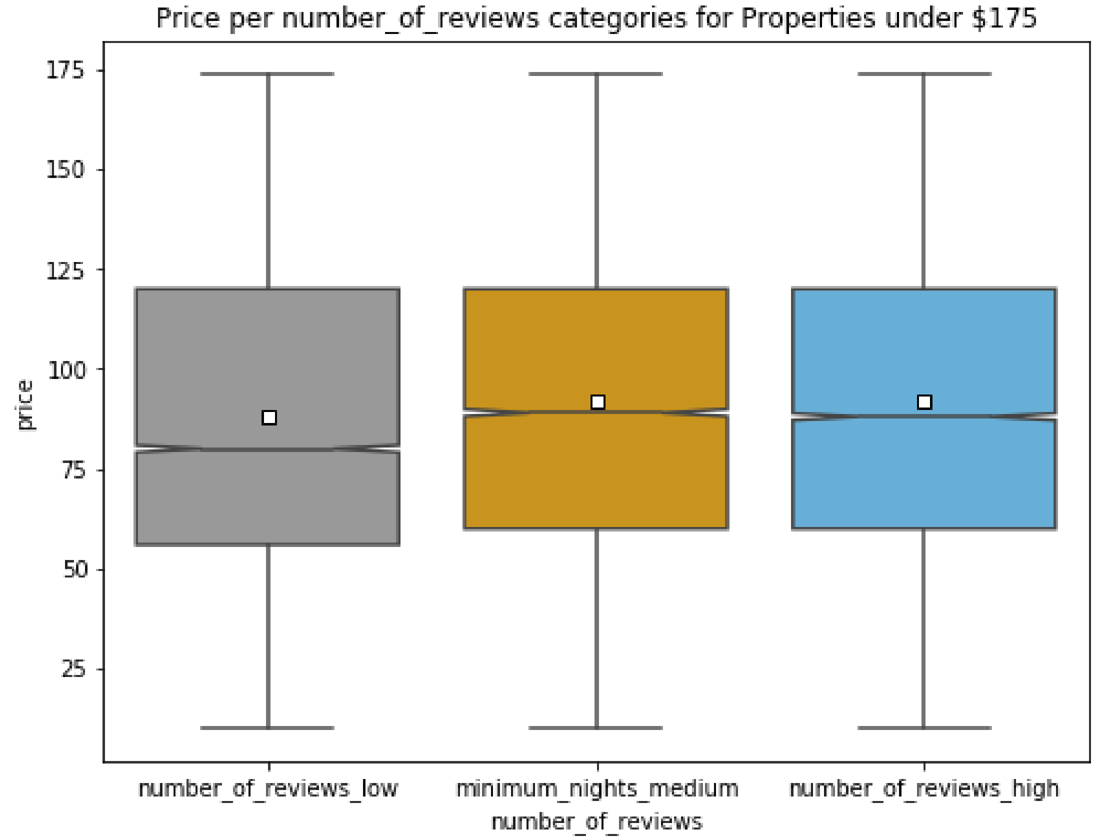
Reviews number impact on price

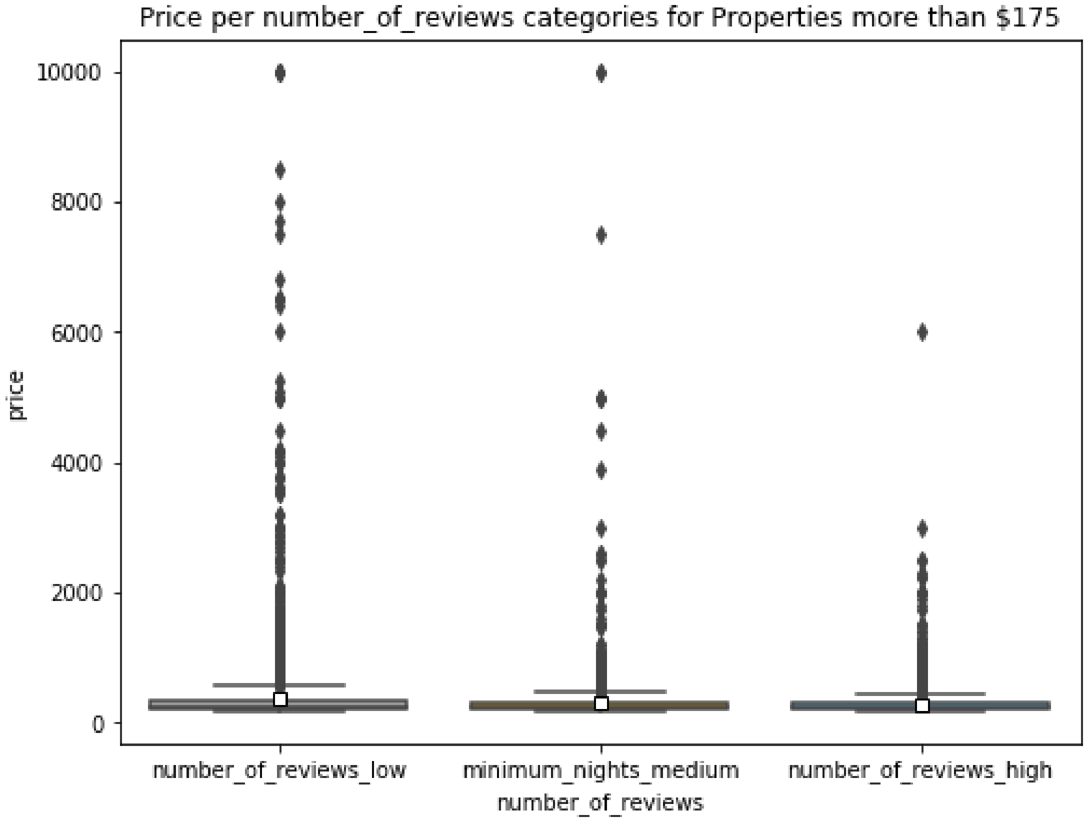




Data Visualization

Reviews number impact on price





Model

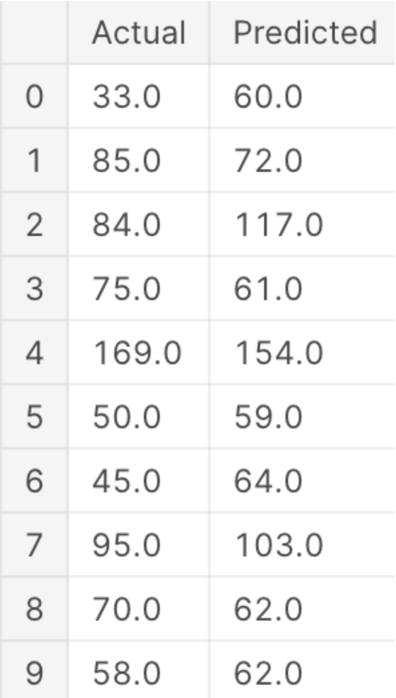
Prepare the datset

1. Log10 transform left skewed features
2. Split the dataset into low and high price
3. Evaluate the model

Model

Multiple linear regression

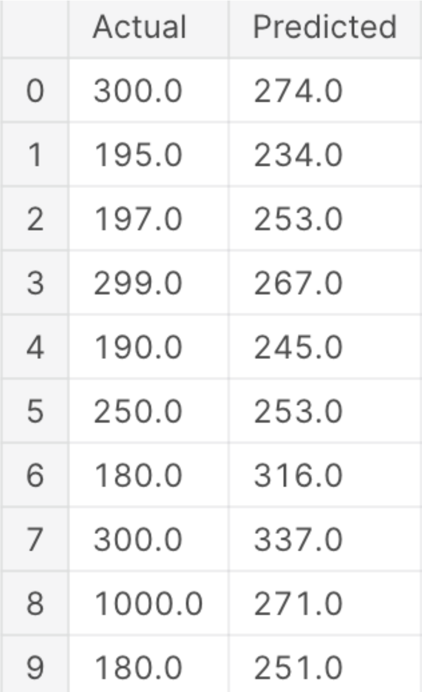
Low price dataset



	Actual	Predicted
0	33.0	60.0
1	85.0	72.0
2	84.0	117.0
3	75.0	61.0
4	169.0	154.0
5	50.0	59.0
6	45.0	64.0
7	95.0	103.0
8	70.0	62.0
9	58.0	62.0

- Price mean: 1.92
- Price std: 0.2
- RMSE: 0.14
- R2 score train: 0.54
- R2 score test: 0.52

High price dataset



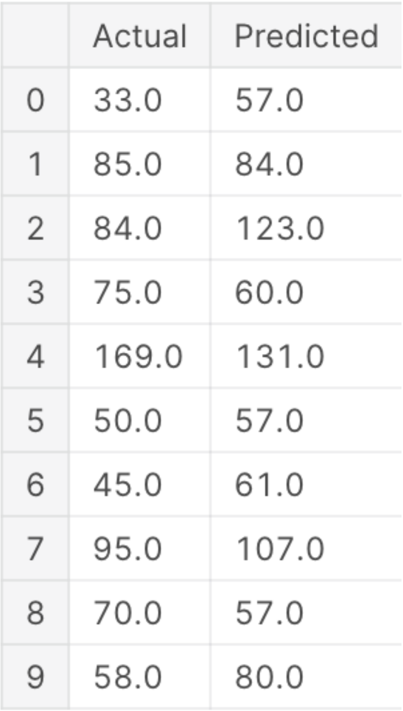
	Actual	Predicted
0	300.0	274.0
1	195.0	234.0
2	197.0	253.0
3	299.0	267.0
4	190.0	245.0
5	250.0	253.0
6	180.0	316.0
7	300.0	337.0
8	1000.0	271.0
9	180.0	251.0

- Price mean: 2.45
- Price std: 0.2
- RMSE: 0.2
- R2 score train: 0.09
- R2 score test: 0.05

Model

Random forest regression

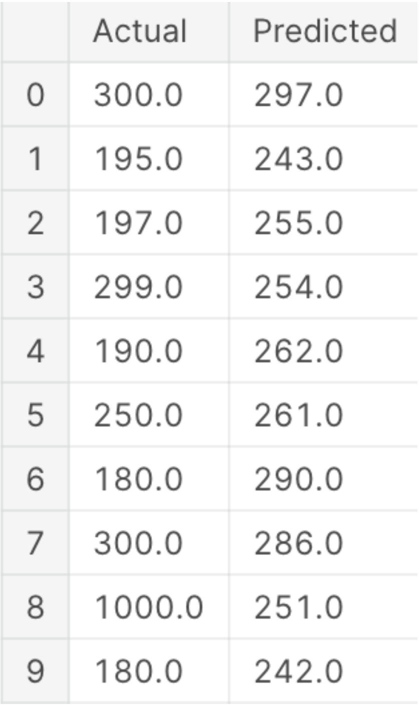
Low price dataset



	Actual	Predicted
0	33.0	57.0
1	85.0	84.0
2	84.0	123.0
3	75.0	60.0
4	169.0	131.0
5	50.0	57.0
6	45.0	61.0
7	95.0	107.0
8	70.0	57.0
9	58.0	80.0

- Price mean: 1.92
- Price std: 0.2
- RMSE: 0.13
- R2 score train: 0.62
- R2 score test: 0.55

High price dataset



	Actual	Predicted
0	300.0	297.0
1	195.0	243.0
2	197.0	255.0
3	299.0	254.0
4	190.0	262.0
5	250.0	261.0
6	180.0	290.0
7	300.0	286.0
8	1000.0	251.0
9	180.0	242.0

- Price mean: 2.45
- Price std: 0.2
- RMSE: 0.19
- R2 score train: 0.29
- R2 score test: 0.16