New York City Airbnb (Mini Project 2)

8 Feb 2022

By Derek Tan

Agenda

- Business Problem
- Data Science Problem
- Statistics for NYC Airbnb dataset
- Exploratory data analysis
- Machine Learning Model
- Machine Learning Model Evaluation
- Conclusion
- Future work
- Deployment

Business Problem

What is the recommended rental price to charge per day in New York City?

Data Science Problem

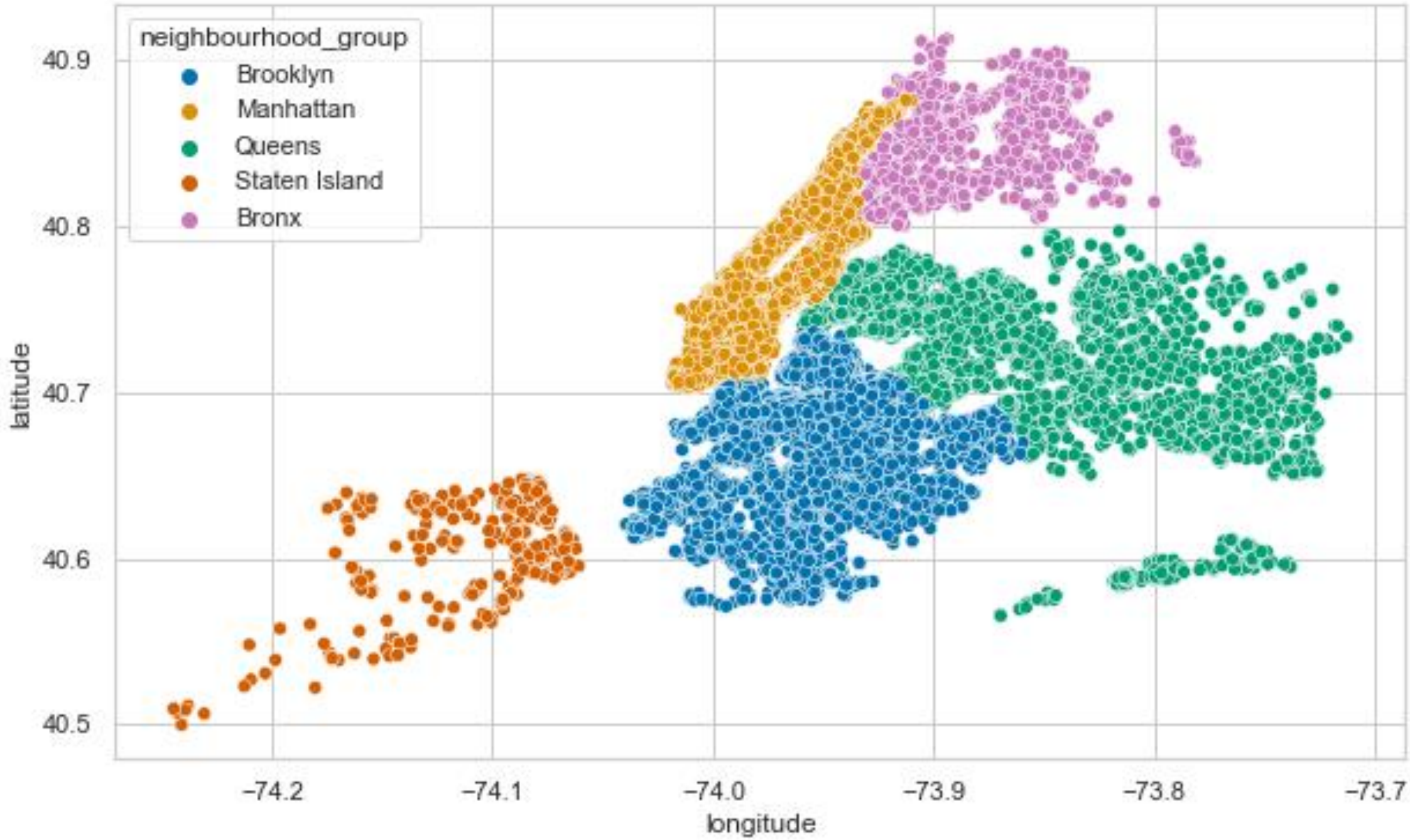
Price prediction using machine learning with Airbnb New York City Dataset

Statistics for NYC Airbnb dataset (Price)

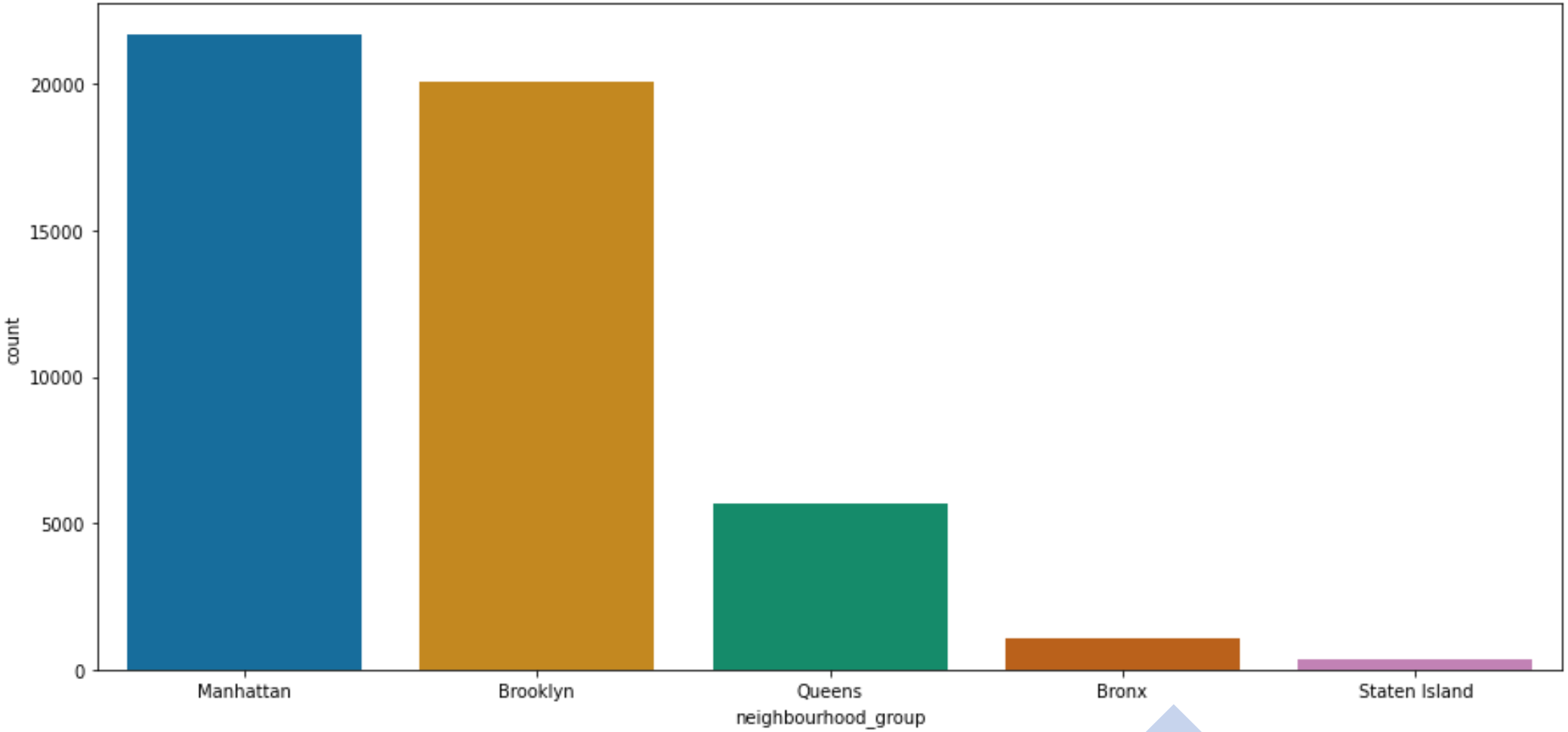
Statistics for NYC Airbnb dataset(Price)

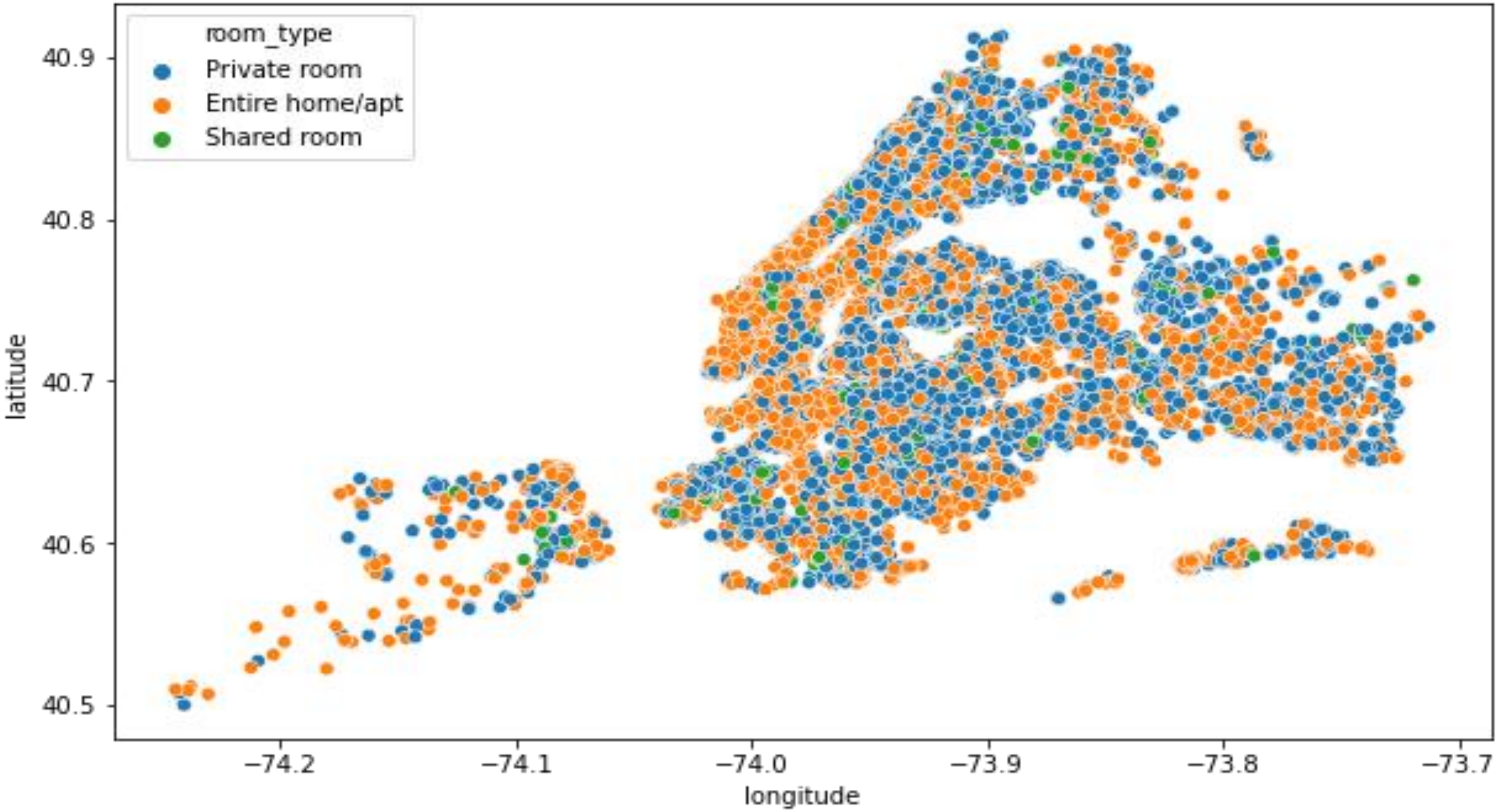
- Minimum price: $10.00
- Maximum price: $10,000.00
- Mean price: $152.76
- Median price $106.00
- Standard deviation of prices: $240.14
- First quartile of prices: $69.00
- Second quartile of prices: $175.00
- Interquartile (IQR) of prices: $106.0

Let’s Explore!

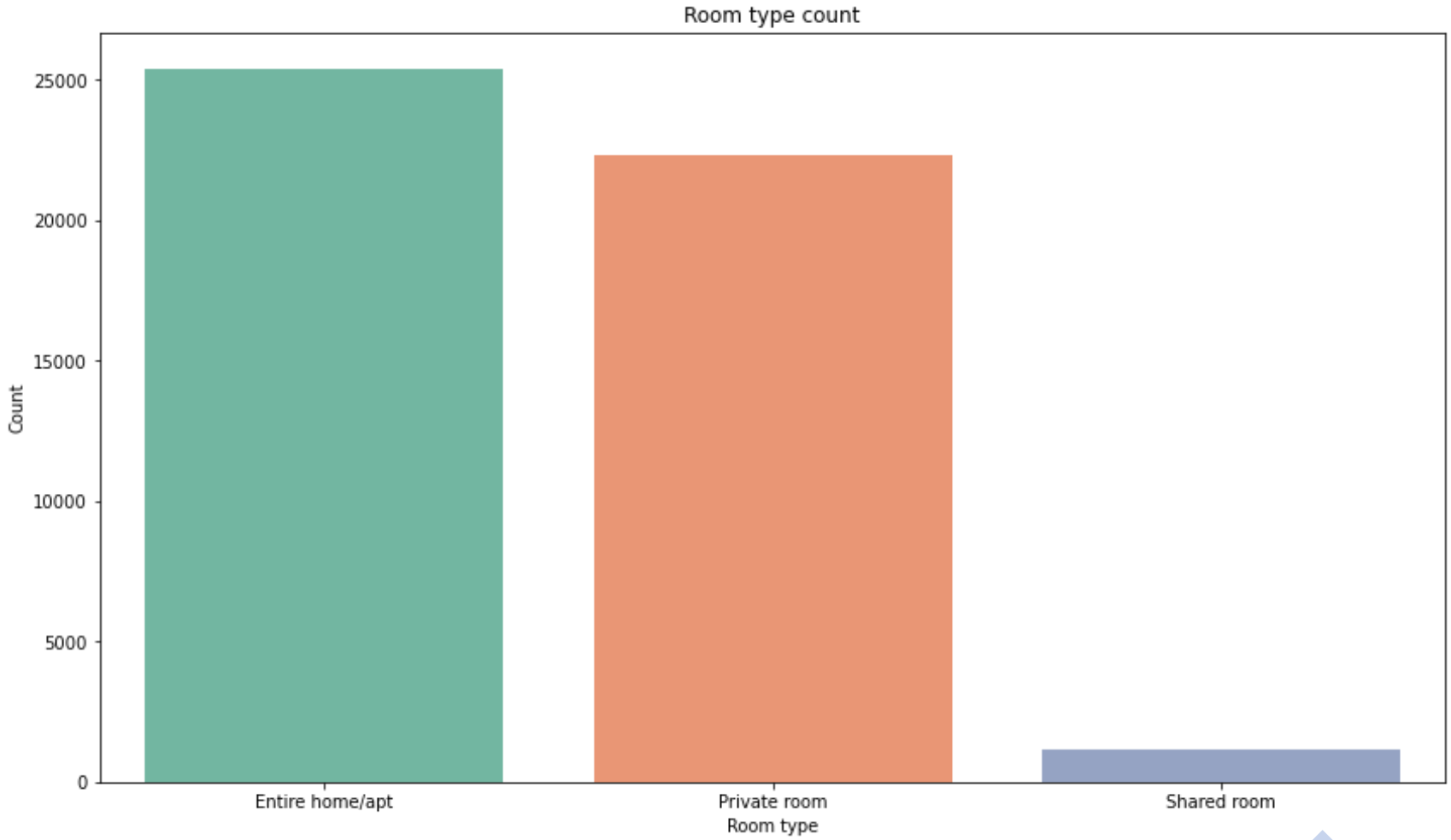

neighbourhood_group
Brooklyn
Manhattan
Queens
Staten Island
Bronx
40.9
40.8
40.7
40.6
40.5
latitude
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
longitude

New York City



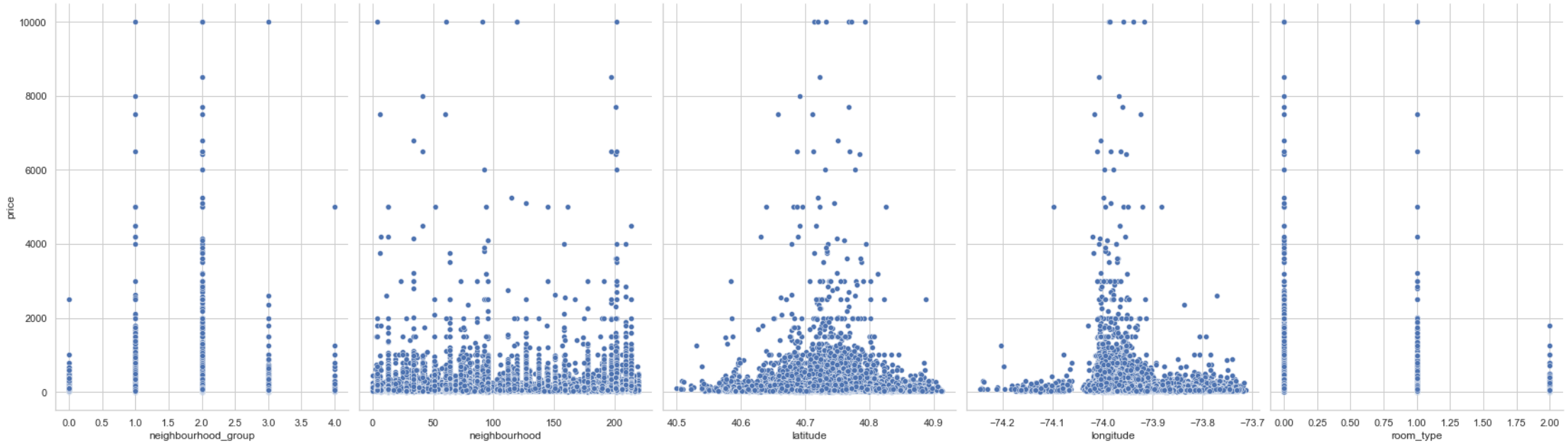

room_type
Private room
Entire home/apt
Shared room
40.9
40.8
40.7
40.6
40.5
latitude
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
longitude

New York City(Room type)

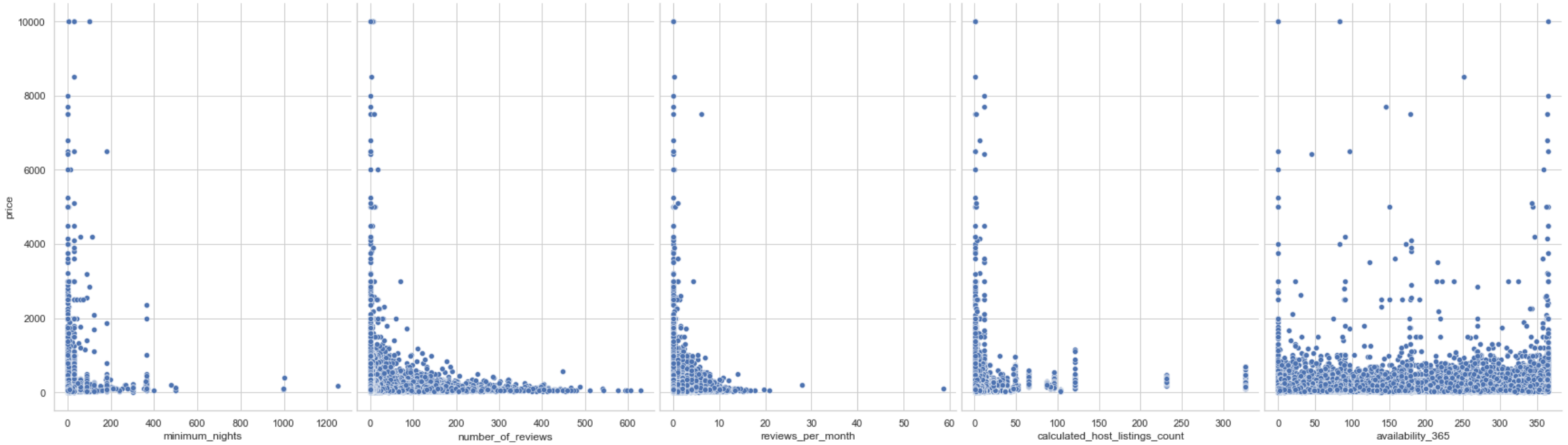


Feature Observations

Feature Observations



Feature Observations



Preparing The Data

Preparing The Data: Independent Variable

Latitude – Within New York City

Longitude – Within New York City

Room type – Entire home/apt, Private Room or Shared Room

Minimum Nights – Minimum nights renter needs to rent

Availability 365 – How many days available for rent.

Preparing The Data: Independent Variable

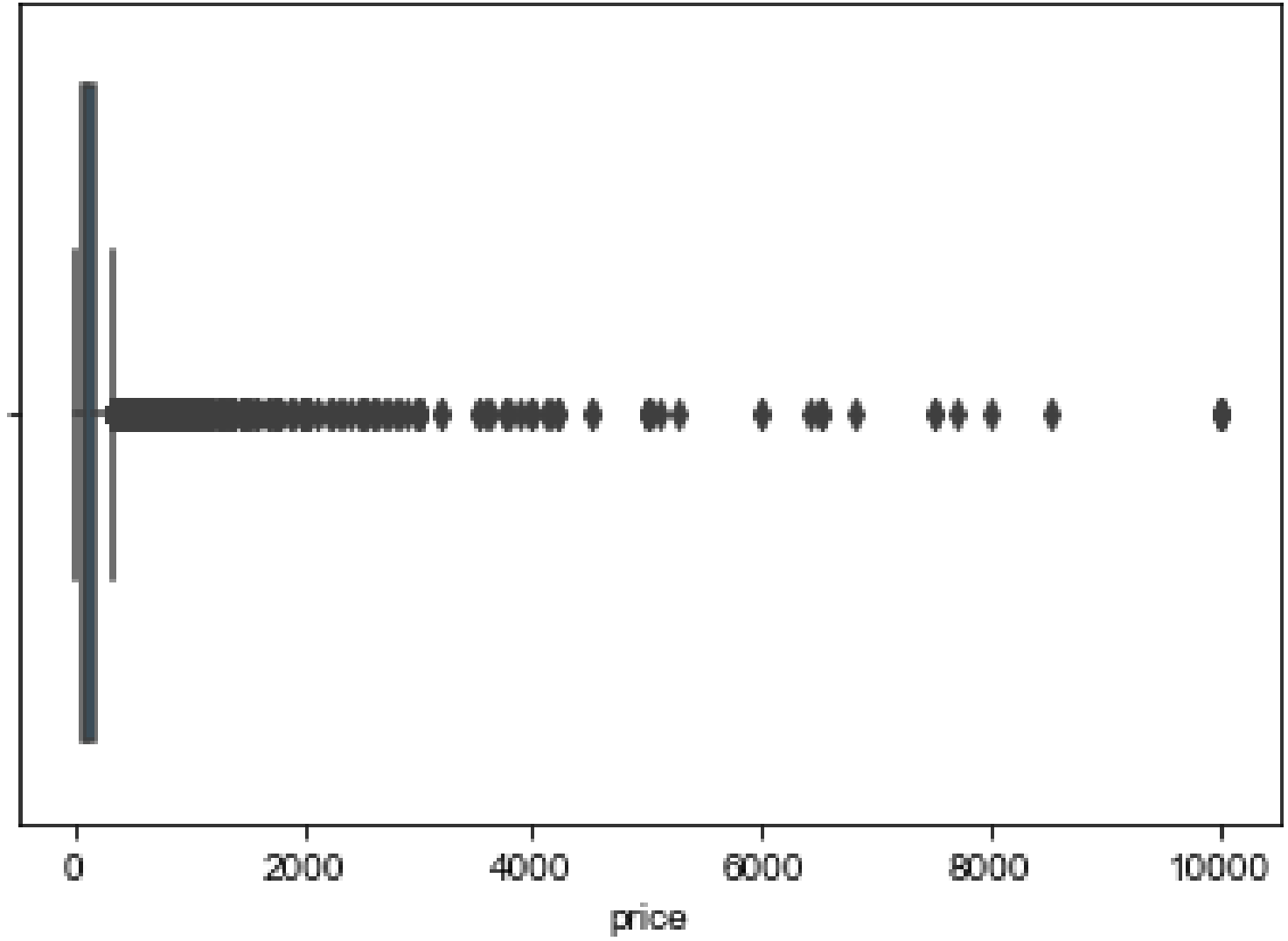
Dep. Variable:	price	**R-squared:**	0.085
Model:	OLS	**Adj. R-squared:**	0.085
Method:	Least Squares	**F-statistic:**	905.4
Date:	Mon, 07 Feb 2022	**Prob (F-statistic):**	0.00
Time:	21:02:35	**Log-Likelihood:**	-3.3522e+05
No. Observations:	48895	**AIC:**	6.704e+05
Df Residuals:	48889	**BIC:**	6.705e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-5.407e+04	1940.384	-27.863	0.000	-5.79e+04	-5.03e+04
latitude	204.6043	19.138	10.691	0.000	167.094	242.115
longitude	-620.8811	23.114	-26.862	0.000	-666.185	-575.578
room_type	-101.0187	1.942	-52.014	0.000	-104.825	-97.212
minimum_nights	0.0490	0.051	0.951	0.341	-0.052	0.150
availability_365	0.1766	0.008	22.030	0.000	0.161	0.192

Omnibus:	109722.610	**Durbin-Watson:**	1.845
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	949954485.480
Skew:	21.148	**Prob(JB):**	0.00
Kurtosis:	684.538	**Cond. No.**	3.42e+05

Dependent Variable Analysis: price

Dependent Variable Analysis: price

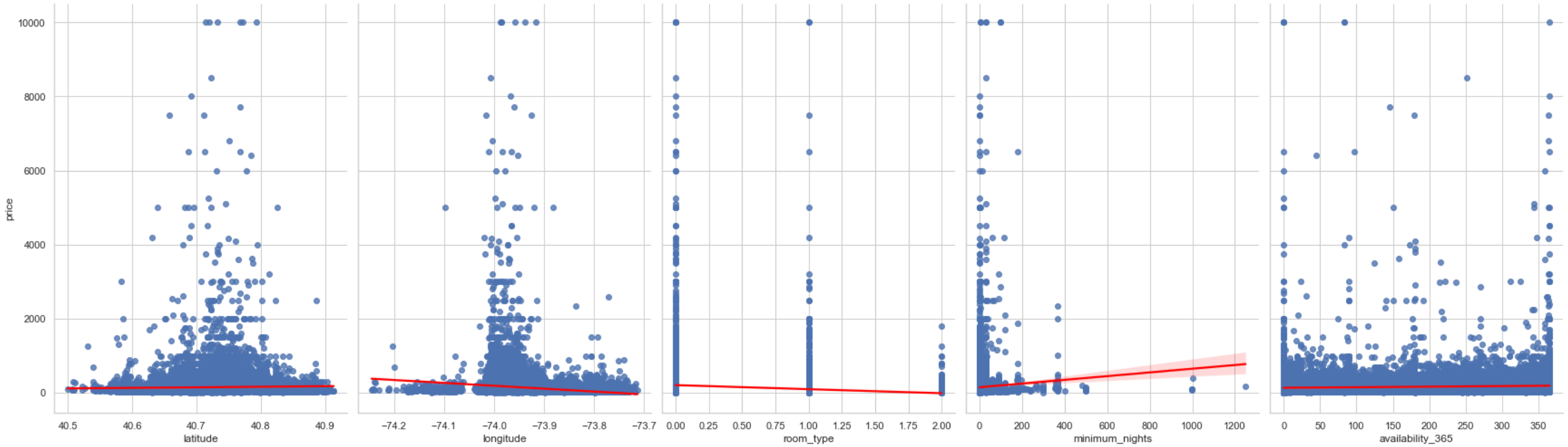


Model

Linear Regression

Model Evaluation

Model Evaluation



Model Evaluation

```
Model Score 0.124

Test Score with CV = [0.14674162 0.09000172 0.18408251 0.1600858  0.09915317]
Train Score with CV = [0.0663044  0.14350955 0.09537528 0.06429441 0.06497465]
```

Conclusion

Prices have huge differences and lots of Outliers

Underfitting data model, having a model score of 0.124

Future work

Research	Increase	Analyze
Research other relevant datasets to increase the number of features. Perform feature engineering	Increase the complexity of the model	Analyze alternative model such as PCA, decision tree, random forest, etc

Deployment: Prediction Demo using Flask

Thank You