Applied Statistics

Michał Wilkosz

Data: [New York City Airbnb Open Data]

1. About dataset

Description

- Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019.
- This data file includes all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions.

Goal

- Predict NYC Airbnb Rental Prices

2. Dataset overview

Data shape

- Dataset contains 48895 samples of 16 features
- Number of quantitative features: 10
- Number of qualitative features: 6

Columns in dataset

```
id                                int64
name                             object
host_id                           int64
host_name                        object
neighbourhood_group              object
neighbourhood                    object
latitude                        float64
longitude                       float64
room_type                        object
price                             int64
minimum_nights                    int64
number_of_reviews                 int64
last_review                      object
reviews_per_month               float64
calculated_host_listings_count    int64
availability_365                  int64
dtype: object
```

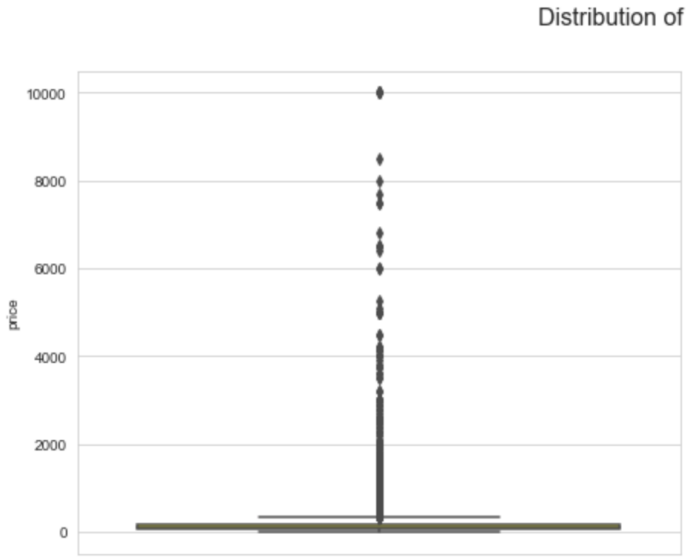
Quick preview of data

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	6	365
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	2	355
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	NaN	1	365
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.64	1	194
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.10	1	0

3. Target inspection

Description

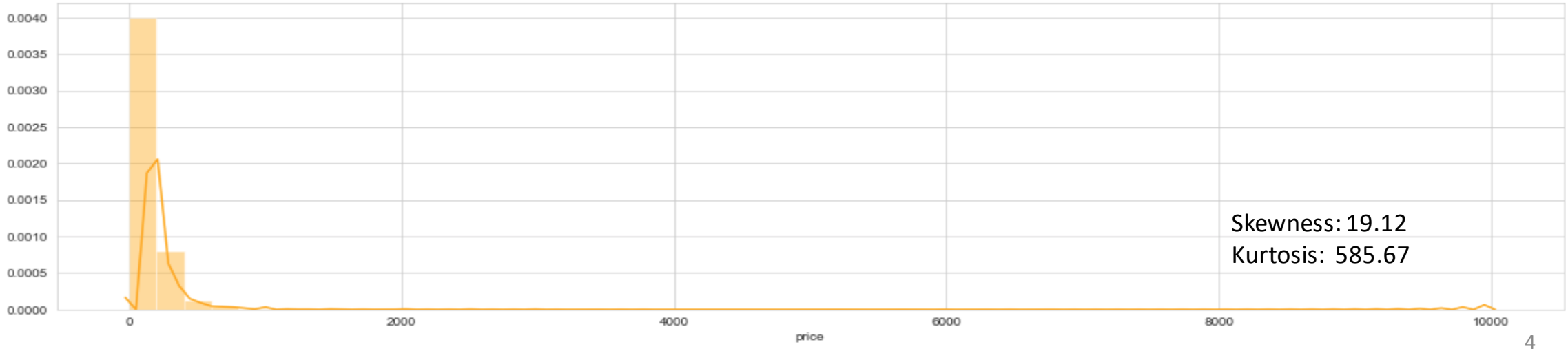
- The aim of project is to predict the prices of advertisements. Therefore, work on the project started with an inspection of the price distribution.



	price
count	48895.0
mean	152.72
std	240.15
min	0.0
25%	69.0
50%	106.0
75%	175.0
max	10000.0

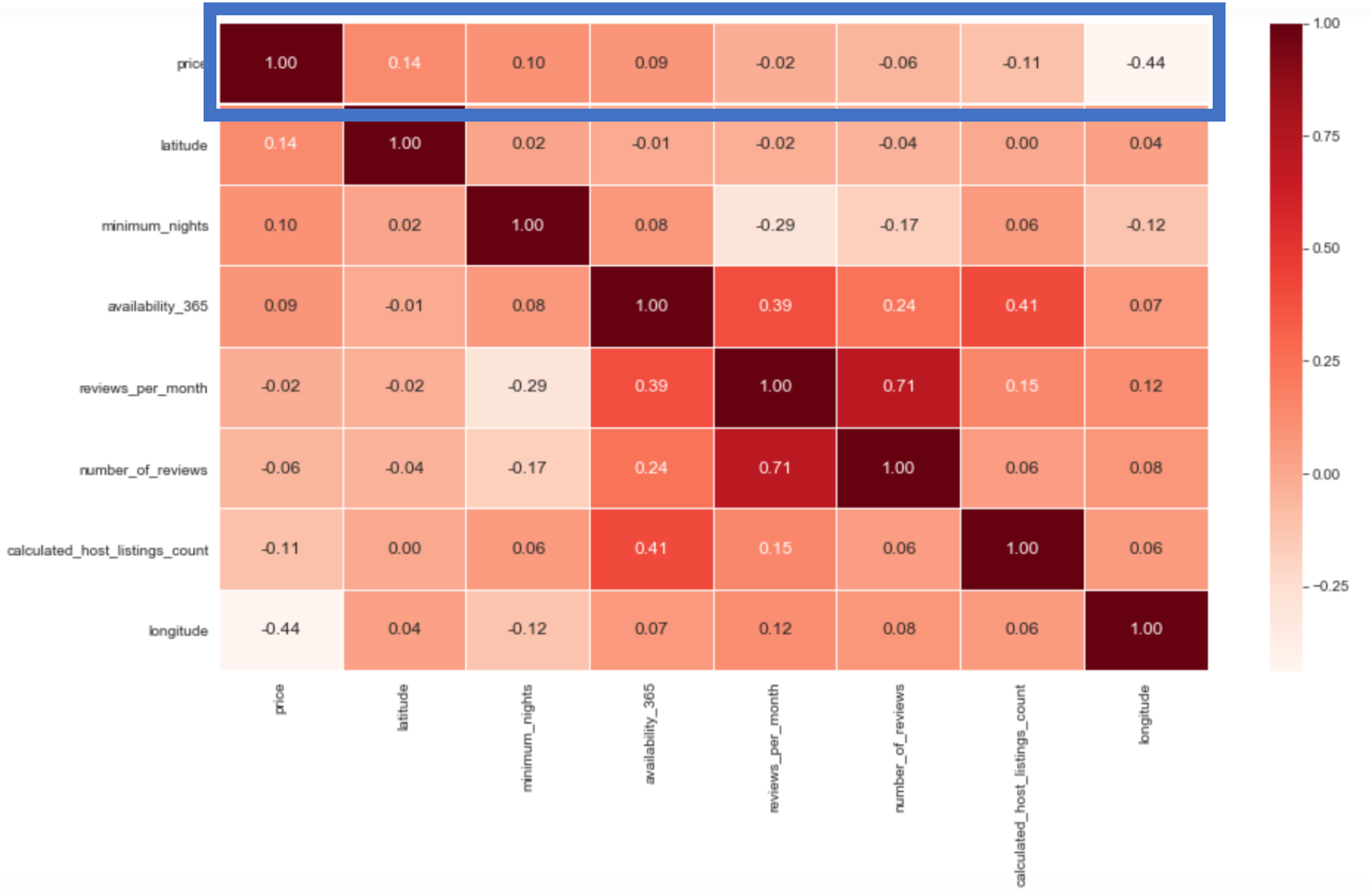
Distribution of rental price is:

- Deviating from normal distribution.
- Having appreciable positive skewness.
- Showing peakdness.

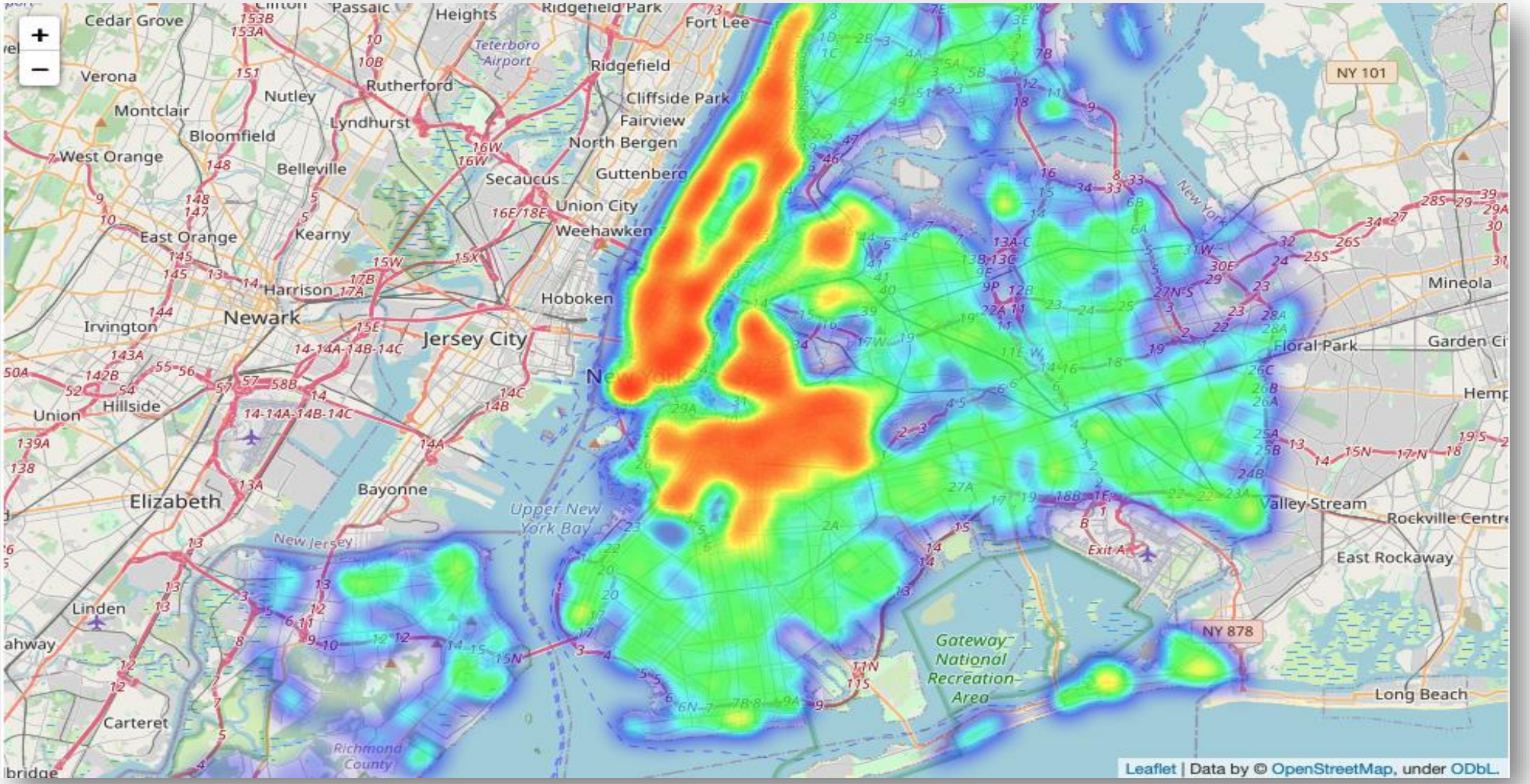


4. Correlation plot

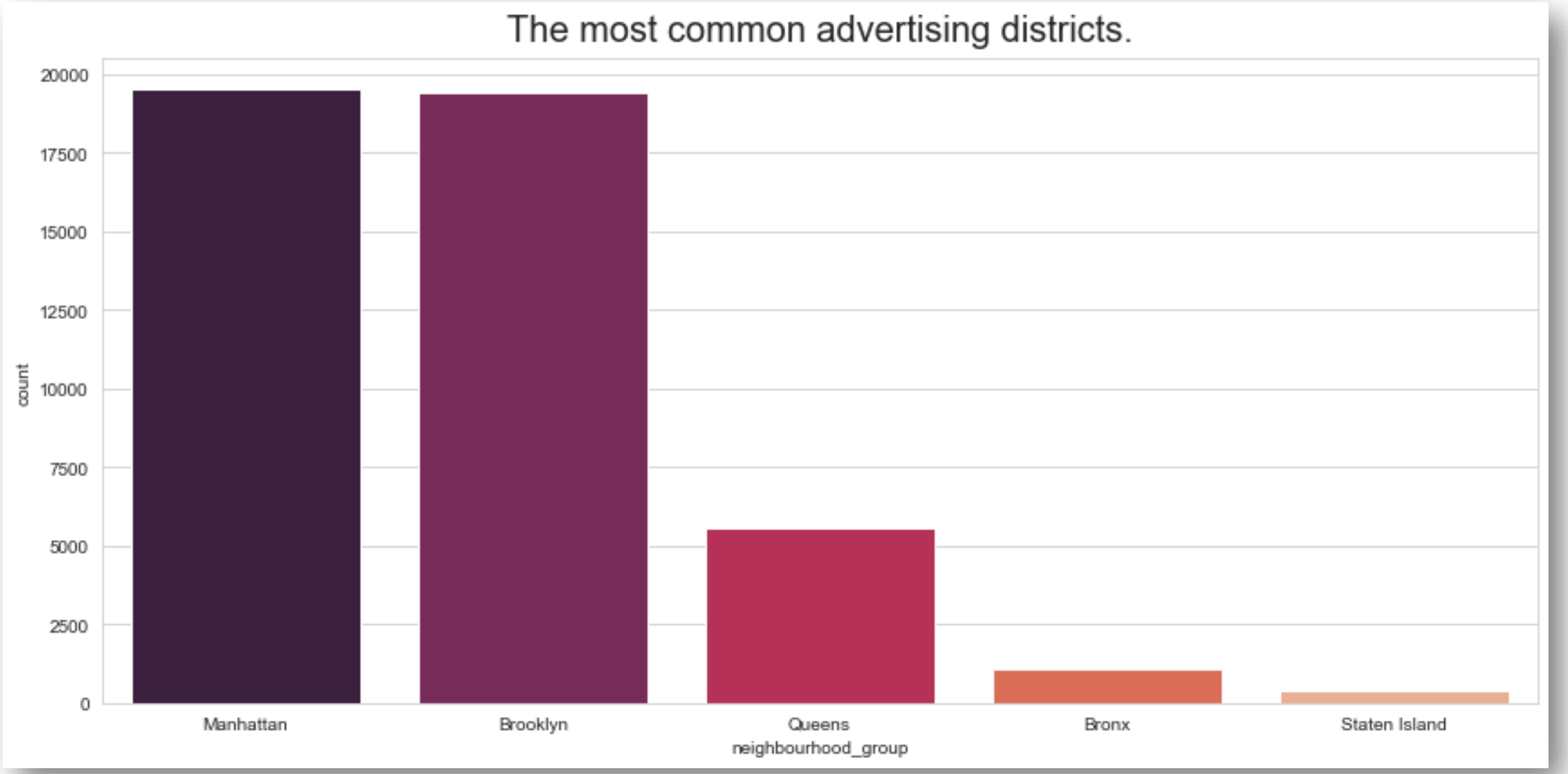
- Correlations(spearman) between quantitative features and target - 'price'.



5. Entry visualizations

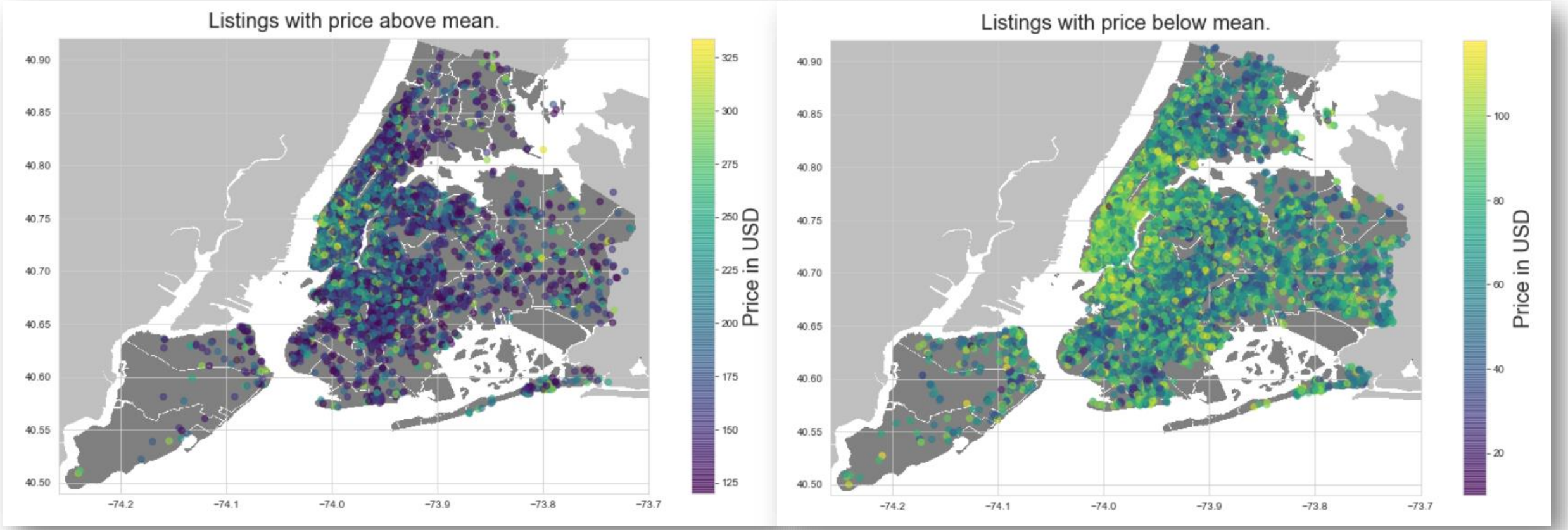


Visualization 1: Distribution of accommodations through heatmap



Visualization 2: The most common advertising districts

Mean price for whole city – 120 USD

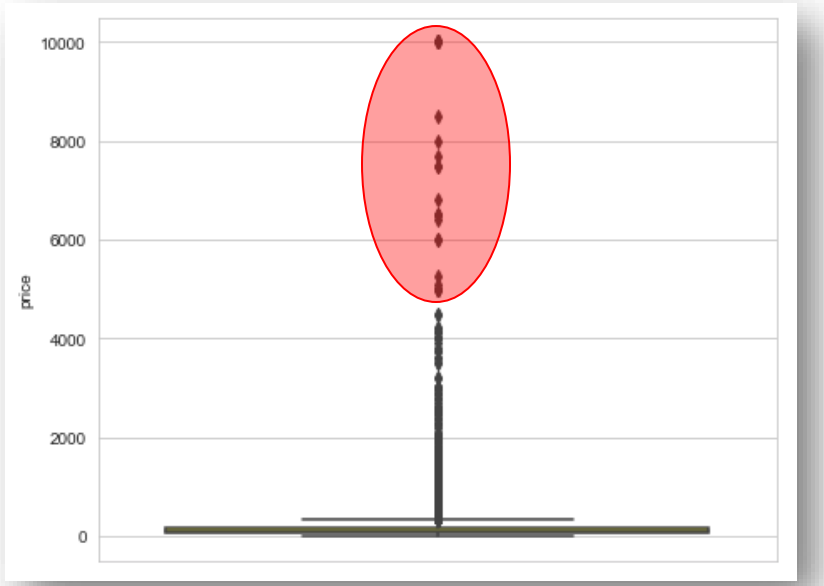


Due to the large amount of data, we have divided the distribution of prices below and over mean.

Visualization 3: Prices through districts

6. Cleaning data by target

Outliers



- Based on boxplot of distribution of rental price, it is clear to see that data contains outliers.
- To detect outliers on our target - 'price', Tukey rule will be implemented, which defines an interquartile range comprised between the 1st and 3rd quartile of the distribution values (IQR). Outliers are rows whose values are outside IQR.

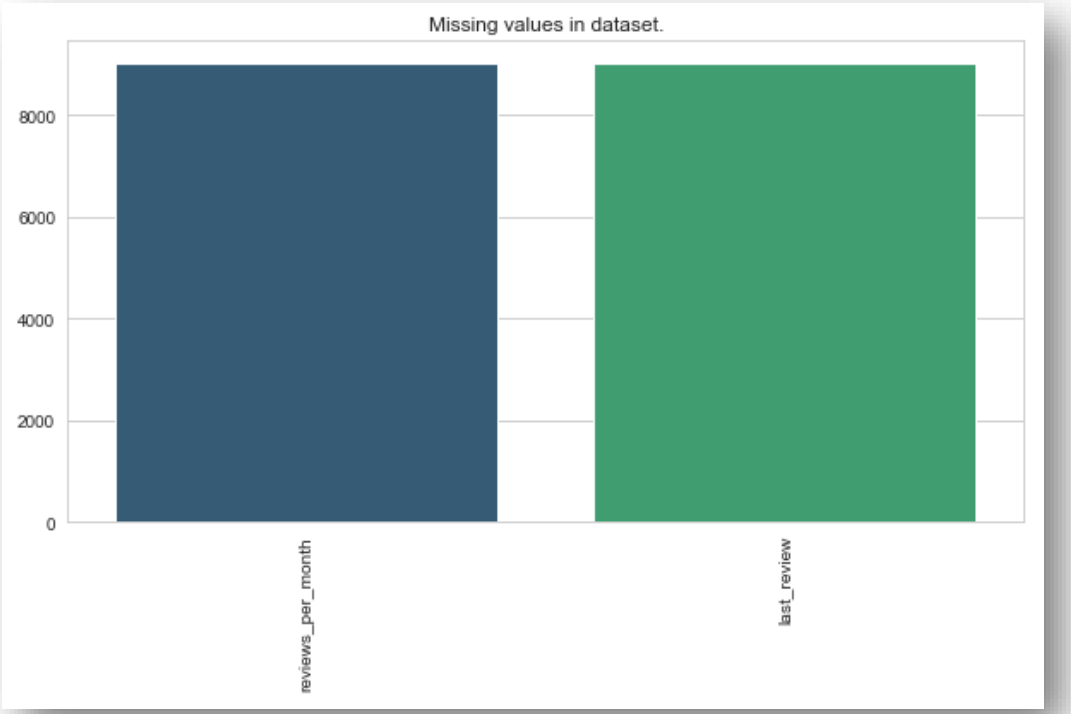
Zero values in target

- Because as we assume that, you cannot rent a room for free. 0 values in target price can affect our estimator as well as outliers, so we have to remove from our data samples whose price is 0.

	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price
21946	Brooklyn	Bedford-Stuyvesant	40.69023	-73.95428	Private room	0
24114	Bronx	East Morrisania	40.83296	-73.88668	Private room	0
24304	Brooklyn	Bushwick	40.69467	-73.92433	Private room	0
24420	Brooklyn	Greenpoint	40.72462	-73.94072	Private room	0
24443	Brooklyn	Williamsburg	40.70838	-73.94645	Entire home/apt	0

Missing values

Analysing the missing values, we've came to the conclusion that for:
•'reviews_per_month' : fill NaN values with 0
•'last_reviews': excluded feature
Commentary: 'last reviews' feature contributes
nothing that could improve our estimator
After data processing our dataset doesn't contain any of missing values.



Unnecessary features

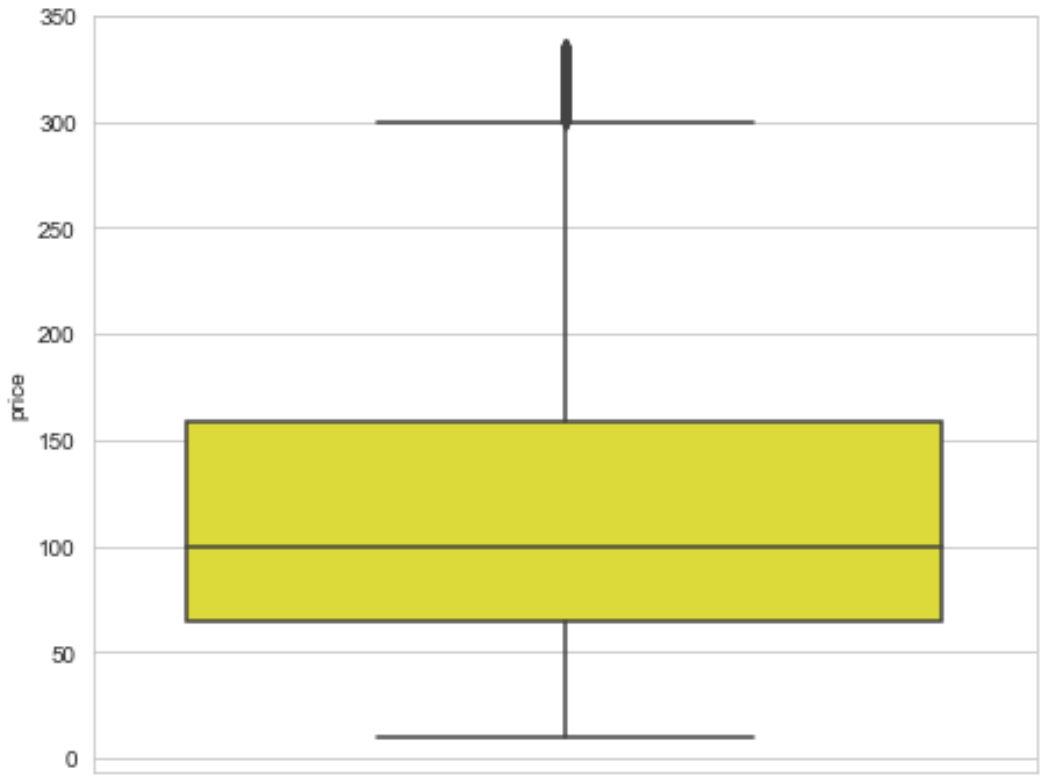
Features which have been excluded from dataset
- 'id'
- 'host_id'
- 'name'
- 'host_name'

	Missing values	Percentage
reviews_per_month	9011	0.5
last_review	9011	0.5

The reason for exclusion of these features out of dataset is the absence of any correlation with price.

7. Target distribution after data processing

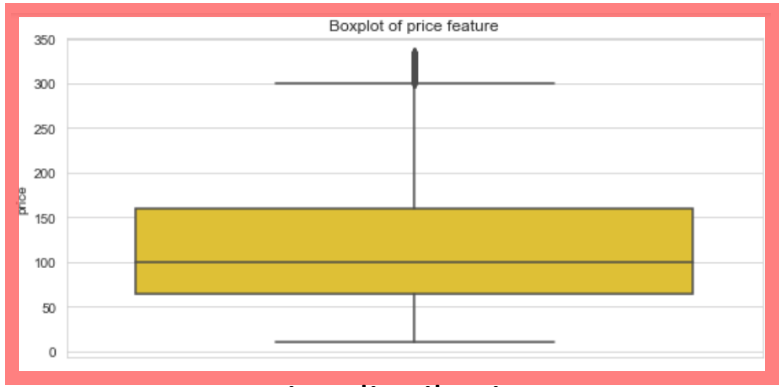
Distribution of price(after Tukey test) and without 0 values.



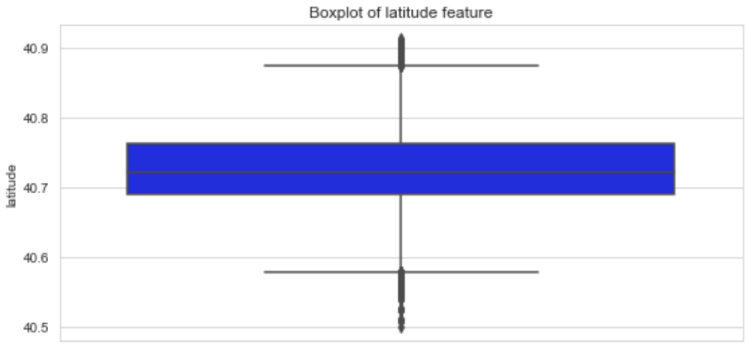
	price
count	45912.0
mean	120.0
std	68.13
min	10.0
25%	65.0
50%	100.0
75%	159.0
max	334.0

Commentary: Thanks to the Tukey test and by removing 0 values, distribution of target looks much better.

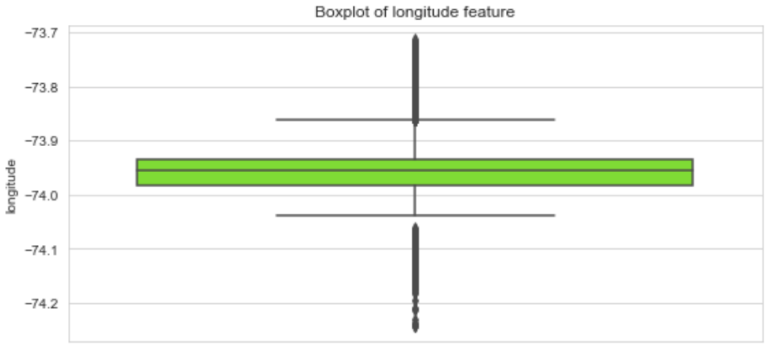
8. Distributions of all quantitative features



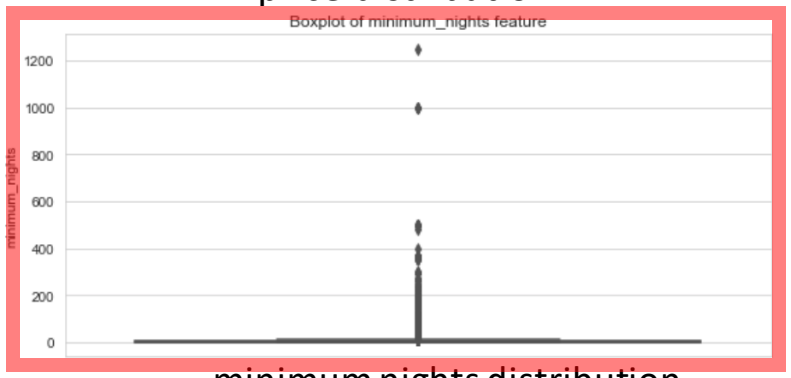
price distribution



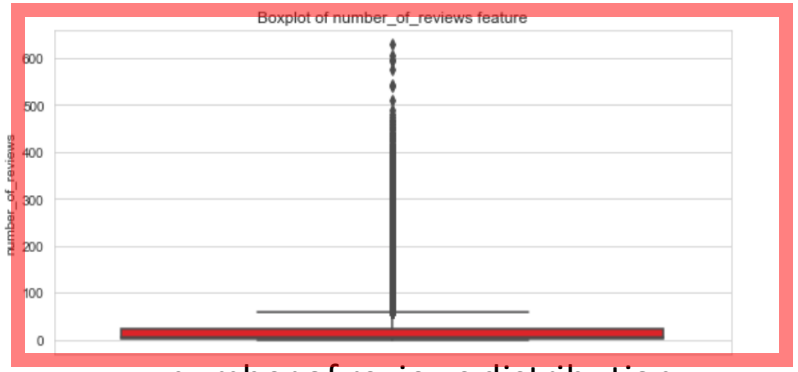
latitude distribution



longitude distribution

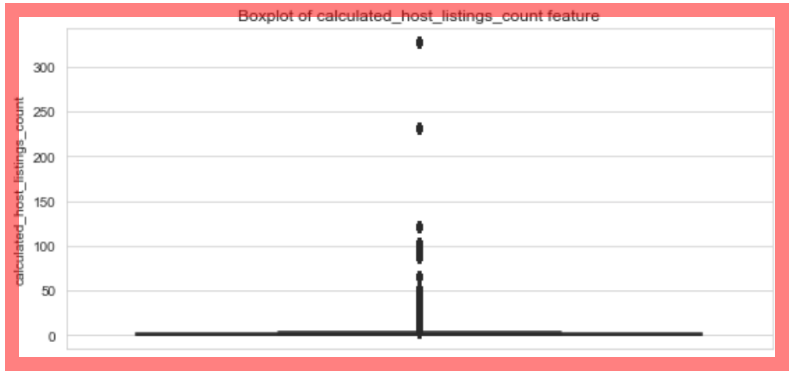


minimum nights distribution

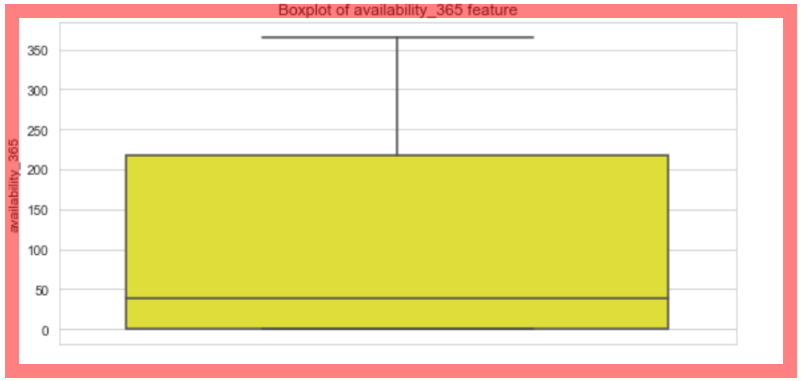


number of reviews distribution

reviews per month distribution

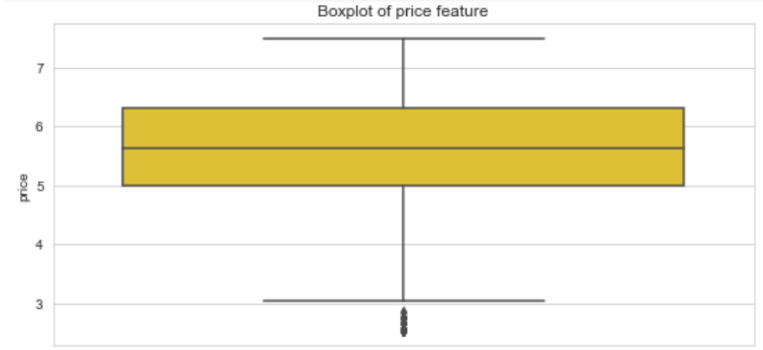


calculated host listings count distribution



availability 365 distribution

9. Solution?

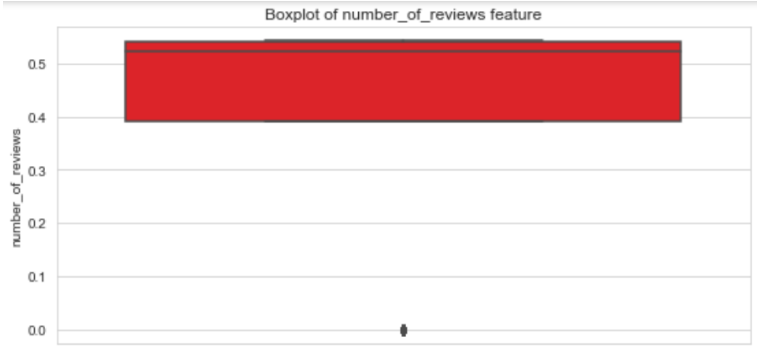


price distribution

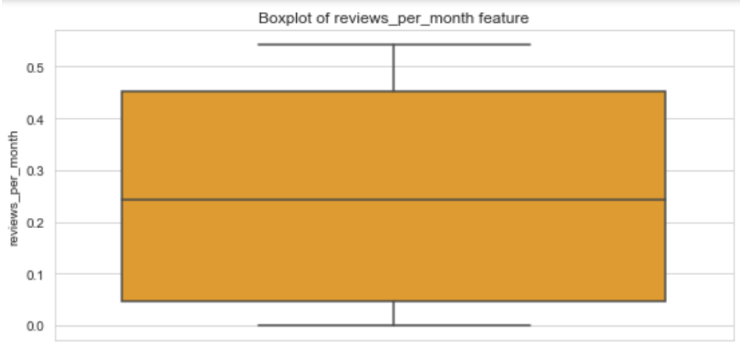
- Performing a box-cox transformation to all skewed features.

- The transformations carried out have not brought about much improvements

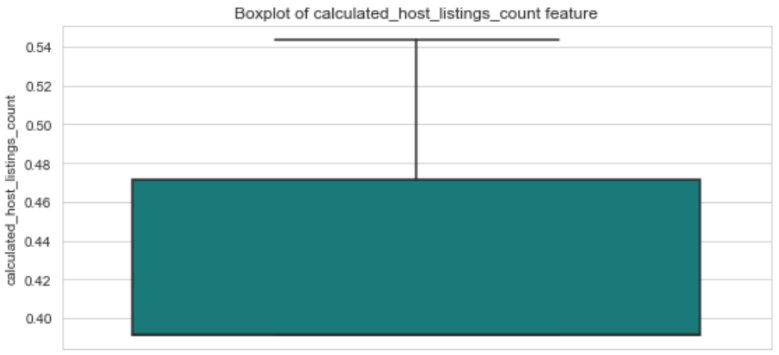
minimum nights distribution



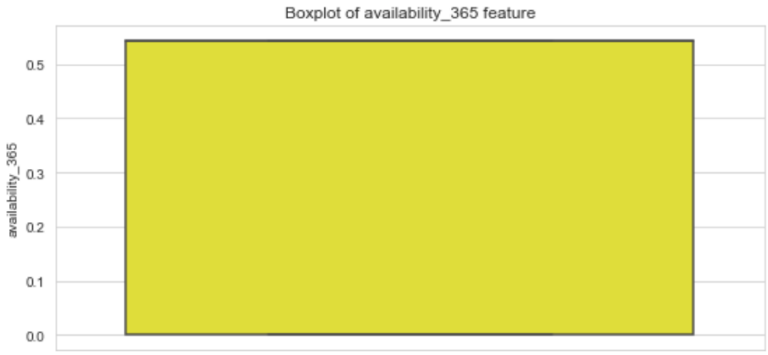
Number of reviews distribution



reviews per month distribution



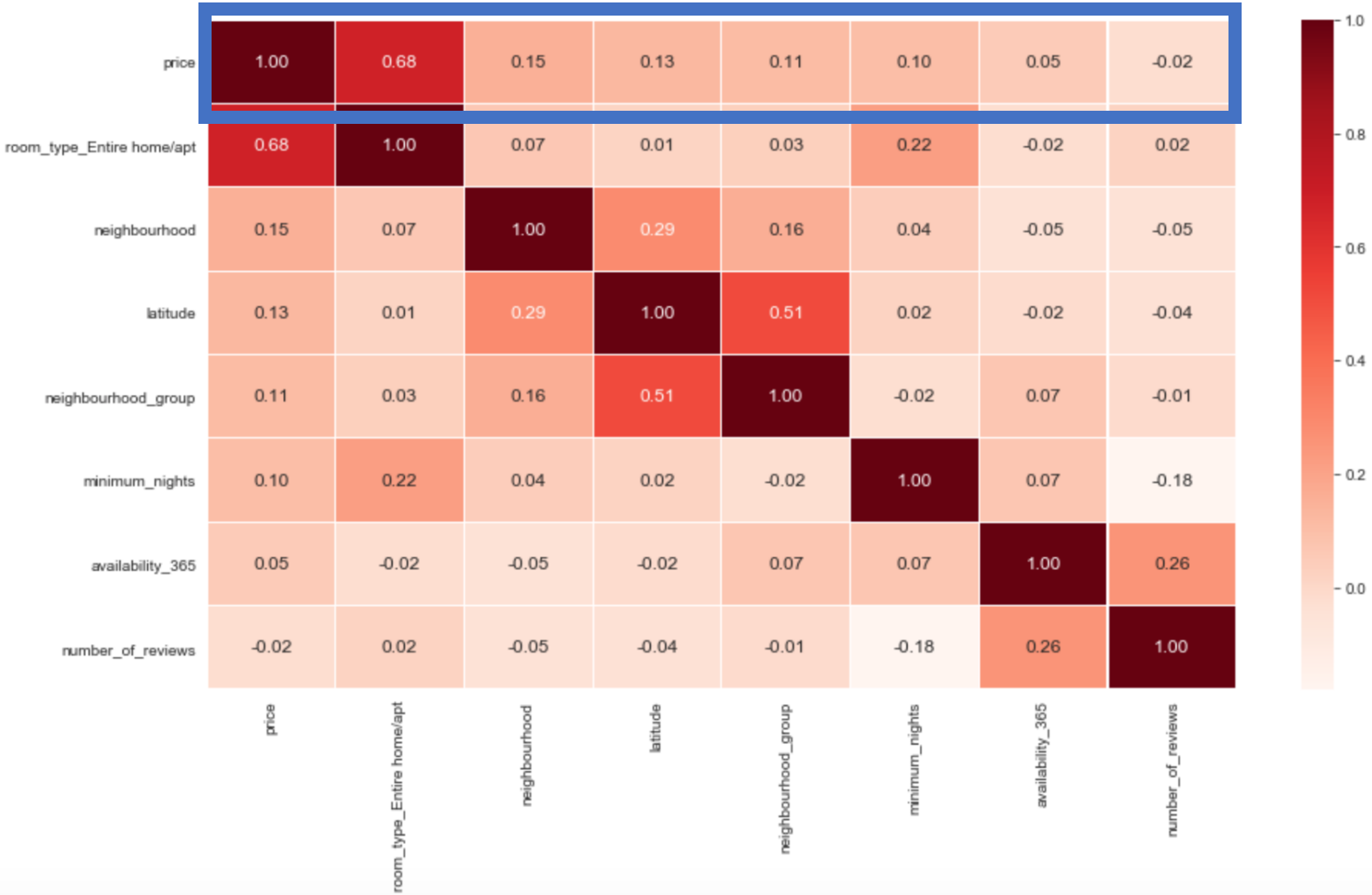
calculated host listings count distribution



availability 365 distribution

10. Correlation plot

- Correlations between quantitative features and target - 'price' after data transformations and with transformed qualitative features into quantitative ones

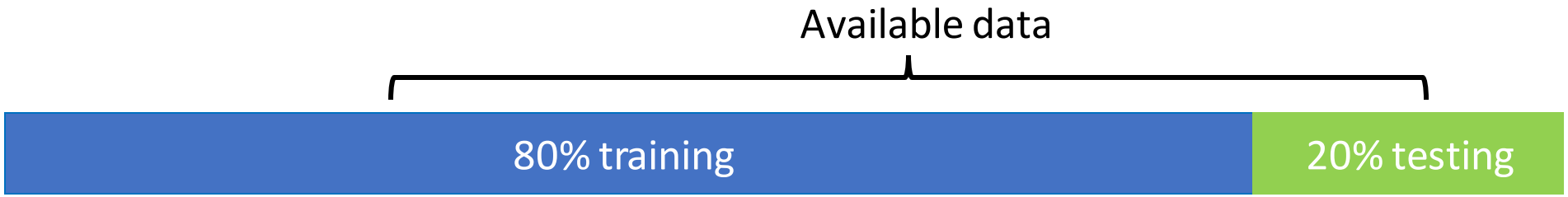


11. Preparing data to predictive models

- **Metric**

For score our models we've used RMSE, because it has the benefit of penalizing large errors.

- **Train test split**




```
Dimensions of the training feature matrix: (36729, 12)
Dimensions of the training target vector: (36729,)
Dimensions of the test feature matrix: (9183, 12)
Dimensions of the test target vector: (9183,)
```

12. Models

Simple linear regression between particular features and target price.

	MSE	RMSE	R2	AIC	BIC
latitude	0.6527	0.8079	-0.1703	-4.5305	-2.5390
longitude	0.5303	0.7282	0.0491	-8.6849	-6.6934
minimum_nights	0.6573	0.8107	-0.1786	-4.3901	-2.3986
number_of_reviews	0.6893	0.8302	-0.2360	-3.4390	-1.4476
neighbourhood_group	0.6612	0.8131	-0.1856	-4.2711	-2.2797
availability_365	0.6759	0.8221	-0.2118	-3.8339	-1.8425
neighbourhood	0.5446	0.7379	0.0235	-8.1528	-6.1613
calculated_host_listings count	0.6219	0.7886	-0.1151	-5.4973	-3.5058
room_type_Entire home/apt	0.3587	0.5989	0.3568	-16.5033	-14.5119
room_type_Private room	0.4483	0.6695	0.1962	-12.0456	-10.0541
room_type_Shared room	0.6279	0.7924	-0.1259	-5.3049	-3.3135

Linear regression

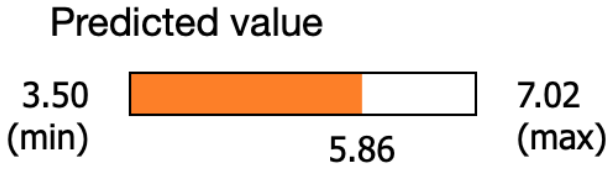
Scores:

	train	test
RMSE	0.5871	0.5879
MSE	0.3447	0.3456
R2	0.5270	0.5155
AIC	52710	82071
BIC	443490	409206

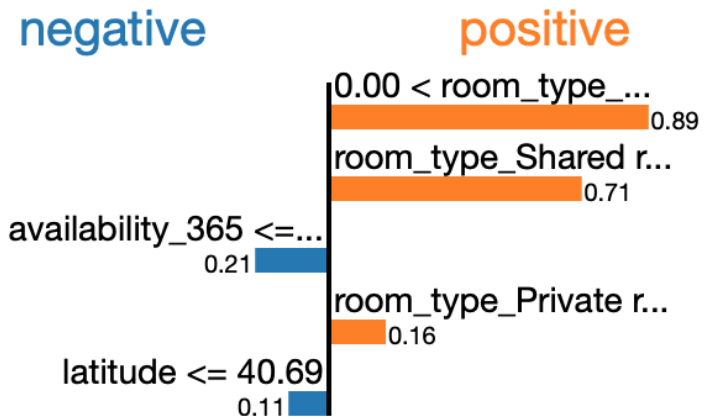
Model top features:

Weight[?]	Feature
+1.190	latitude
+0.893	room_type_Entire home/apt
+0.377	availability_365
+0.062	neighbourhood_group
+0.001	neighbourhood
-0.045	calculated_host_listings_count
-0.161	reviews_per_month
-0.163	minimum_nights
-0.168	room_type_Private room
-0.192	number_of_reviews
-0.724	room_type_Shared room
-4.512	longitude
-376.874	<BIAS>

Model interpretation for 500th data sample:




```
Intercept 4.5541278040747475
Prediction_local [5.99207647]
Right: 5.864529642731782
```



Feature	Value
room_type_Entire home/apt	1.00
room_type_Shared room	0.00
availability_365	0.00
room_type_Private room	0.00
latitude	40.69

Random forest

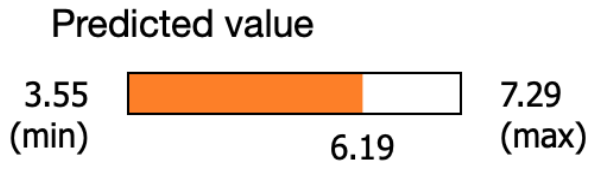
Scores:

	train	test
RMSE	0.2259	0.5341
MSE	0.0511	0.2853
R2	0.9299	0.6001
AIC	-17437	80309
BIC	373365	407444

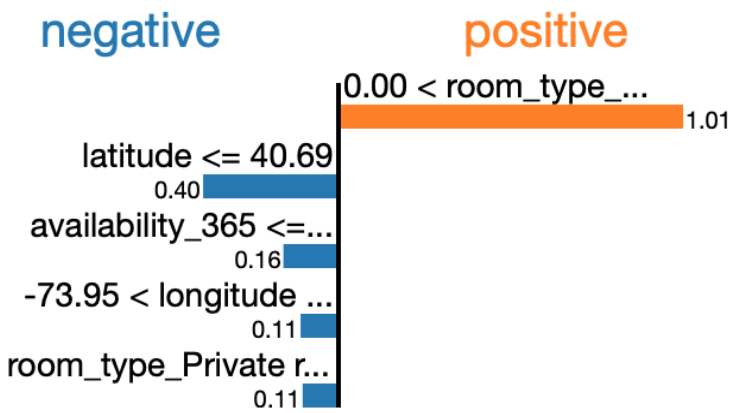
Model top features:

Weight	Feature
0.4369 ± 0.0043	room_type_Entire home/apt
0.1698 ± 0.0065	longitude
0.1380 ± 0.0063	latitude
0.0562 ± 0.0031	availability_365
0.0538 ± 0.0033	reviews_per_month
0.0433 ± 0.0018	minimum_nights
0.0397 ± 0.0022	number_of_reviews
0.0248 ± 0.0065	neighbourhood
0.0248 ± 0.0030	calculated_host_listings_count
0.0063 ± 0.0067	room_type_Private room
0.0039 ± 0.0056	room_type_Shared room
0.0025 ± 0.0018	neighbourhood_group

Model interpretation for 500th data sample:




```
Intercept 5.371827148152454
Prediction_local [5.60839617]
Right: 6.187036429193727
```



Feature	Value
room_type_Entire home/apt	1.00
latitude	40.69
availability_365	0.00
longitude	-73.95
room_type_Private room	0.00

XGBoost regressor

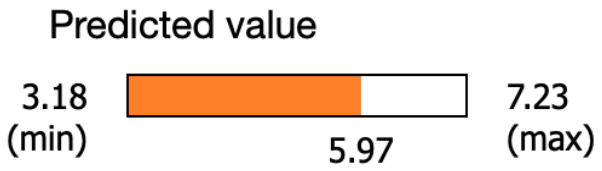
Scores:

	train	test
RMSE	0.4918	0.5311
MSE	0.2418	0.2821
R2	0.6681	0.6046
AIC	39694	80205
BIC	430475	407340

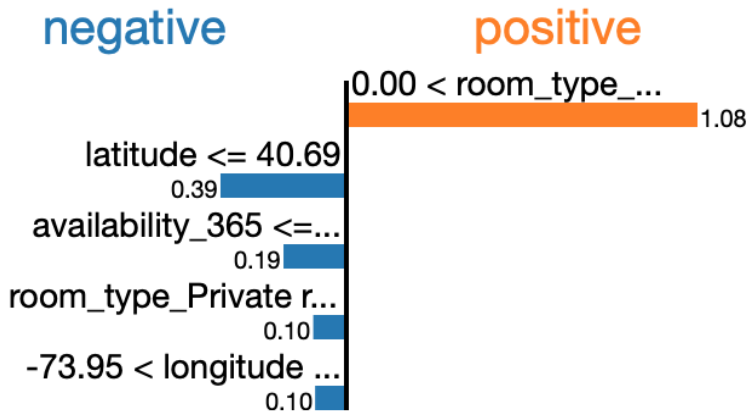
Model top features:

Weight	Feature
0.8753	room_type_Entire home/apt
0.0265	room_type_Private room
0.0190	room_type_Shared room
0.0175	longitude
0.0133	minimum_nights
0.0125	latitude
0.0120	availability_365
0.0068	calculated_host_listings_count
0.0055	neighbourhood
0.0041	reviews_per_month
0.0037	neighbourhood_group
0.0036	number_of_reviews

Model interpretation for 500th data sample:




```
Intercept 5.318123559942578
Prediction_local [5.61801047]
Right: 5.9693427
```



Feature	Value
room_type_Entire home/apt	1.00
latitude	40.69
availability_365	0.00
room_type_Private room	0.00
longitude	-73.95

13. Models for particular districts

- Basing on correlation plot, scores of linear regressions for features and model interpreters, we found that the most important features for models are:

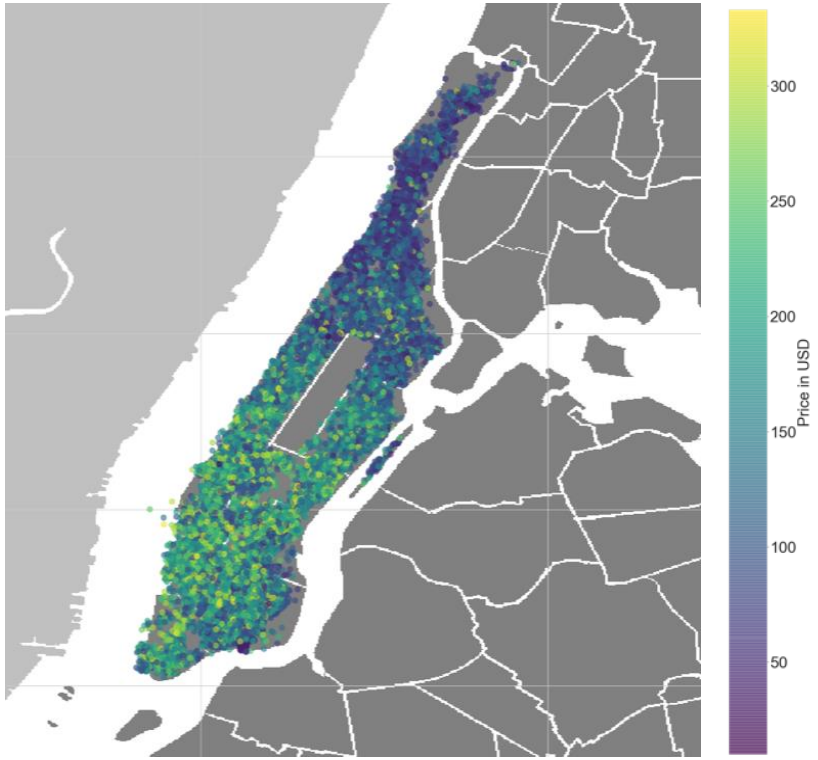
    1. Entire home/apt
    2. Private room
    3. Shared room
    4. Longitude
    5. Latitude
    6. Availability 365
    7. Neighbourhood

- Since the price results for the whole city did not yield very good results, we decided to divide our city by districts and try again using only most relevant features.

- We rejected the random forest model because of notable overfitting.

Manhattan

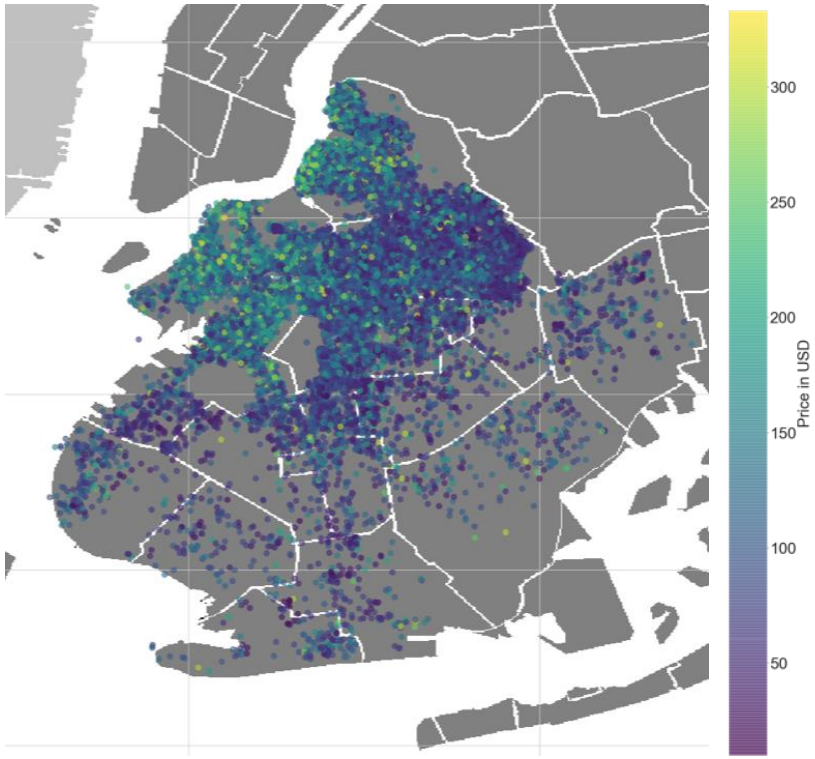
mean price: 145.96



	train	test
Linear regression		
RMSE	52.69	52.55
XGBoost regressor		
RMSE	47.37	50.97

Brooklyn

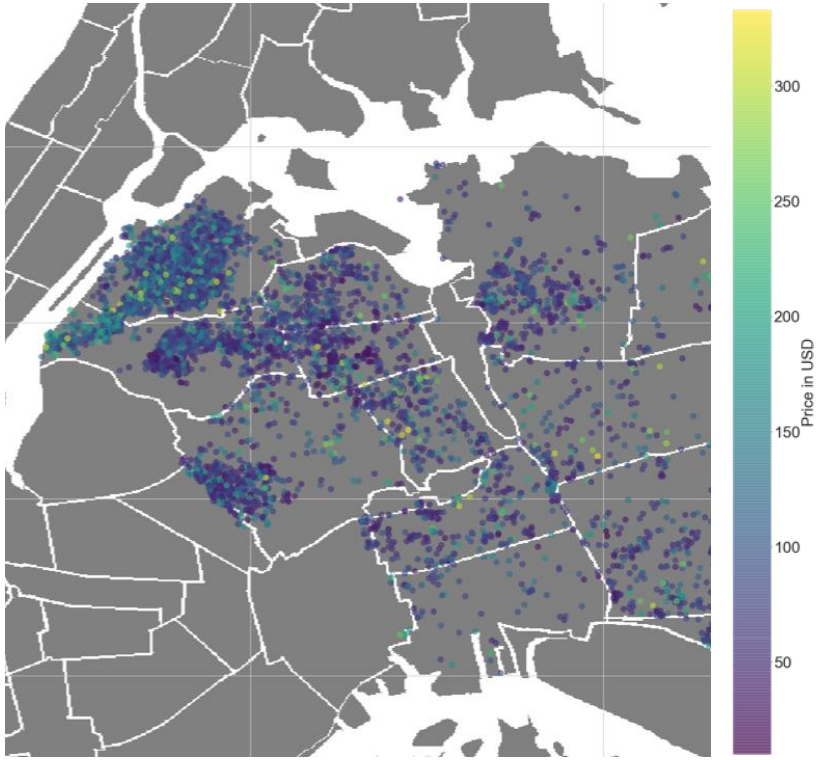
mean price: 105.74



	train	test
Linear regression		
RMSE	44.48	43.58
XGBoost regressor		
RMSE	39.67	42.08

Queens

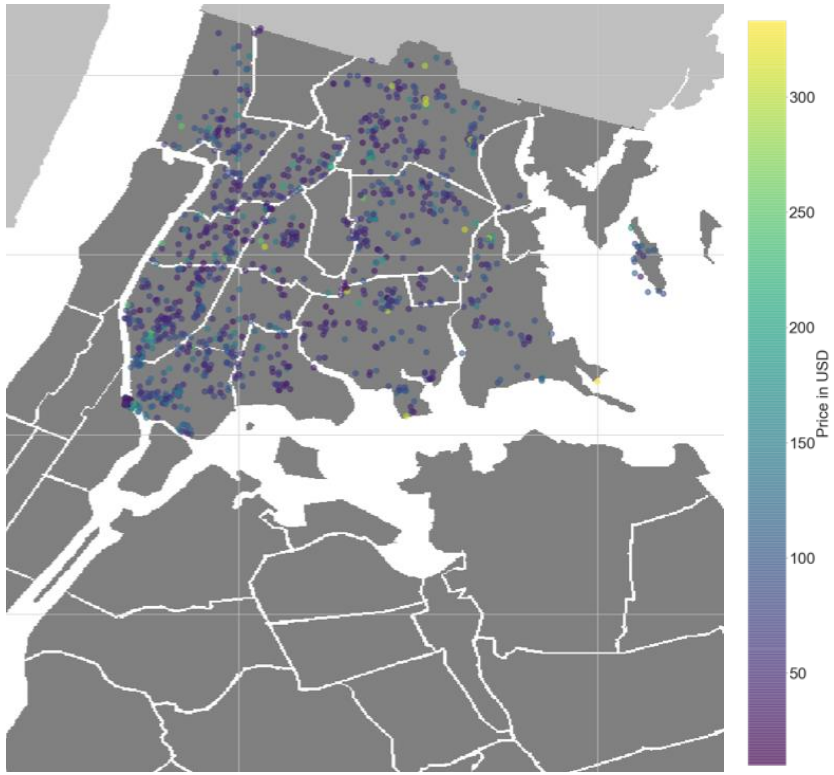
mean price: 88.90



	train	test
Linear regression		
RMSE	42.67	41.47
XGBoost regressor		
RMSE	32.51	41.18

Bronx

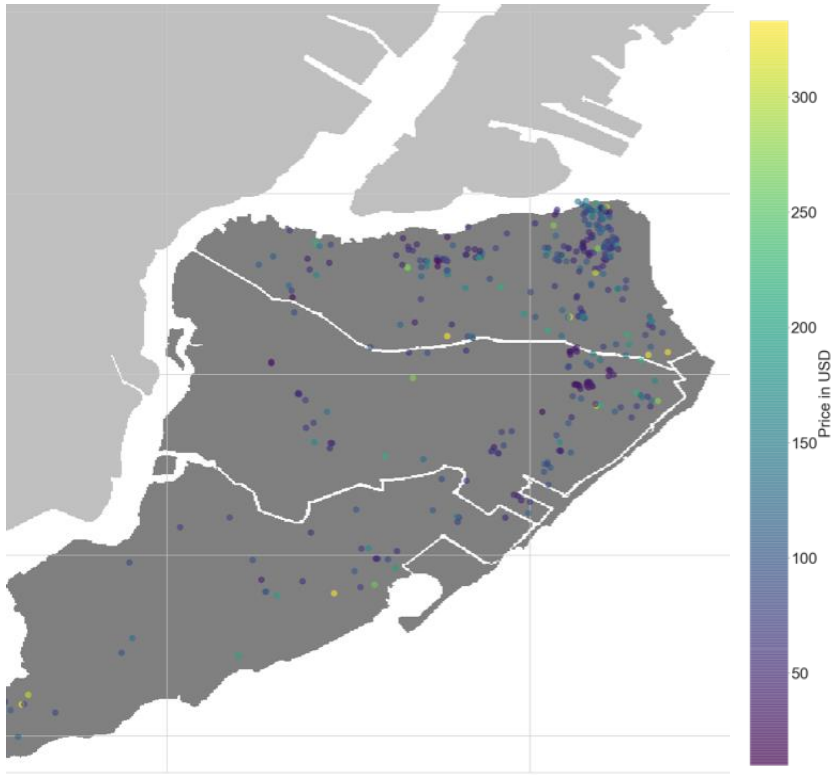
mean price: 77.43



	train	test
Linear regression		
RMSE	39.02	40.51
XGBoost regressor		
RMSE	14.95	46.06

Staten Island

mean price: 89.23



	train	test
Linear regression		
RMSE	46.27	56.60
XGBoost regressor		
RMSE	5.53	62.61

14. Conclusions

- Poor data quality and in particular uneven distributions of features made it difficult to build a good predictive model.
- Multiple operations performed on given dataset did not bring any significant improvement in the obtained prediction of prices.
- Deeper dive into given problem and extending dataset with more valuable feature manifesting better correlation with target could result in obtaining better predictions.