Airbnb New York 2021-2022

Prediction models

Julio Cesar Hernandez-Krol

Airbnb

Airbnb is an online marketplace that connects people who want to rent out their homes with people who are looking for accommodations in specific locales.

http://insideairbnb.com/get-the-data.html

Problem Statement

My project aims to help people set the best possible price for their Airbnb listing on New York.

- Analyzing various features such as location, rooms, number of accommodations, etc. based on price using graphs.
- Using NLP to create a sentimental analysis.
- Using different types of regression models.
- Plotting time series to analyze price and availability

Agenda

- Airbnb data
- Feature analysis (price)
- Sentiment analysis
- Time series
- Feature selection and correlations
- Regression models
- Conclusions and recommendations

Airbnb data

Listings, reservations and reviews dataframes

- Listing dataframe

	bedrooms	bathroom	beds	price	number_of_reviews	reviews_per_month
count	38149.000000	38149.000000	38149.000000	38149.000000	38149.000000	28989.000000
mean	1.290624	1.154159	1.567774	159.436027	23.789143	1.098895
std	0.673828	0.443529	1.063000	292.580710	50.998635	1.767994
min	1.000000	0.000000	1.000000	10.000000	0.000000	0.010000
25%	1.000000	1.000000	1.000000	68.000000	1.000000	0.120000
50%	1.000000	1.000000	1.000000	109.000000	4.000000	0.440000
75%	1.000000	1.000000	2.000000	175.000000	21.000000	1.530000
max	13.000000	8.000000	24.000000	10000.000000	1010.000000	92.920000

- Reservation dataframe

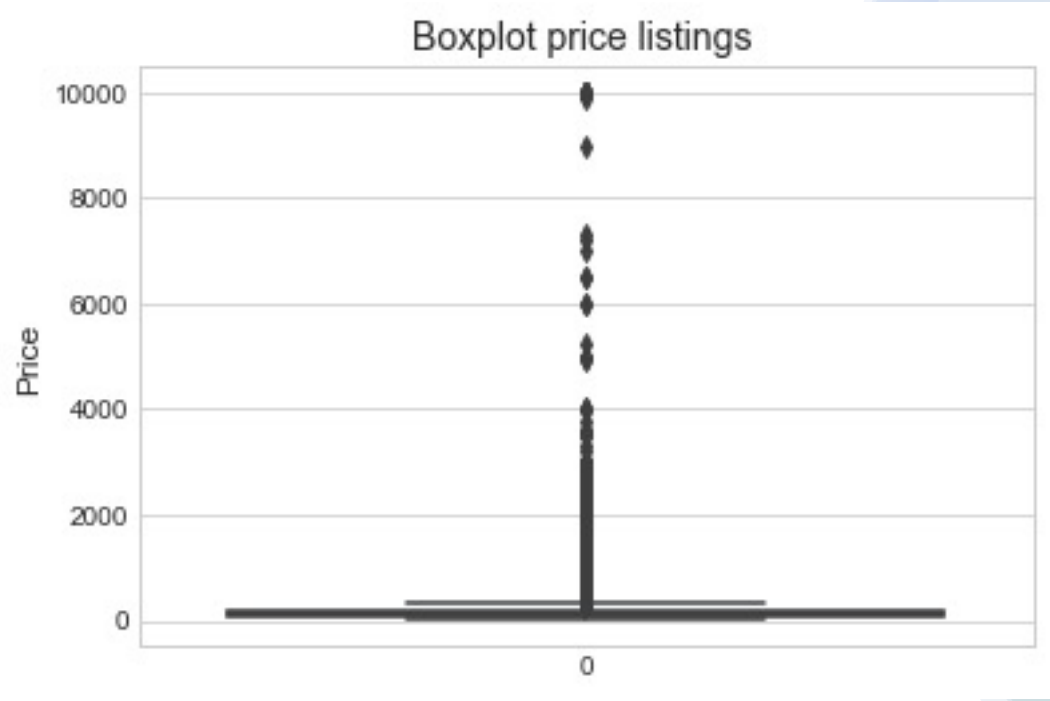
	listings_num_booked	listings_avg_price
count	695.000000	695.000000
mean	14283.725180	160.986956
std	1997.703427	10.551010
min	9204.000000	142.167703
25%	12289.500000	152.019107
50%	13682.000000	156.924221
75%	15766.500000	170.720531
max	18255.000000	184.568790

- This dataframe has the data of the listings and hosts. From the start of Airbnb in 2008 until the month of February 2022.

- This dataframe contain two dataframes, the first one is the reservations booked on February 2021 until 2022-01-31, and the second one from 2022-02 until 2023-01.

Feature analysis (price)

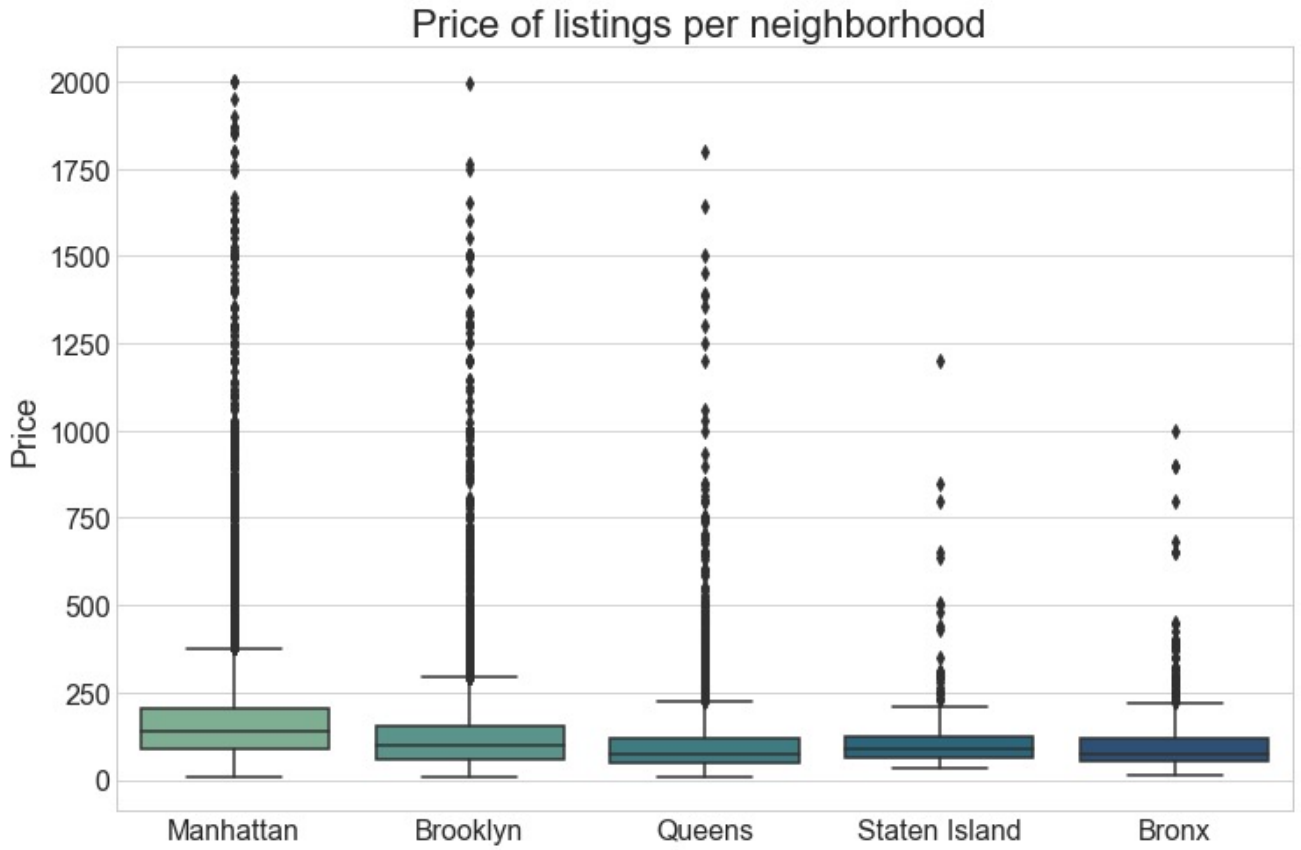
- Price analysis of Airbnb listings in New York



	price
count	38149.000000
mean	159.436027
std	292.580710
min	10.000000
25%	68.000000
50%	109.000000
75%	175.000000
max	10000.000000

- Most of the listings are in the 75% of the listing prices, for that reason the price in the graphic shrinking

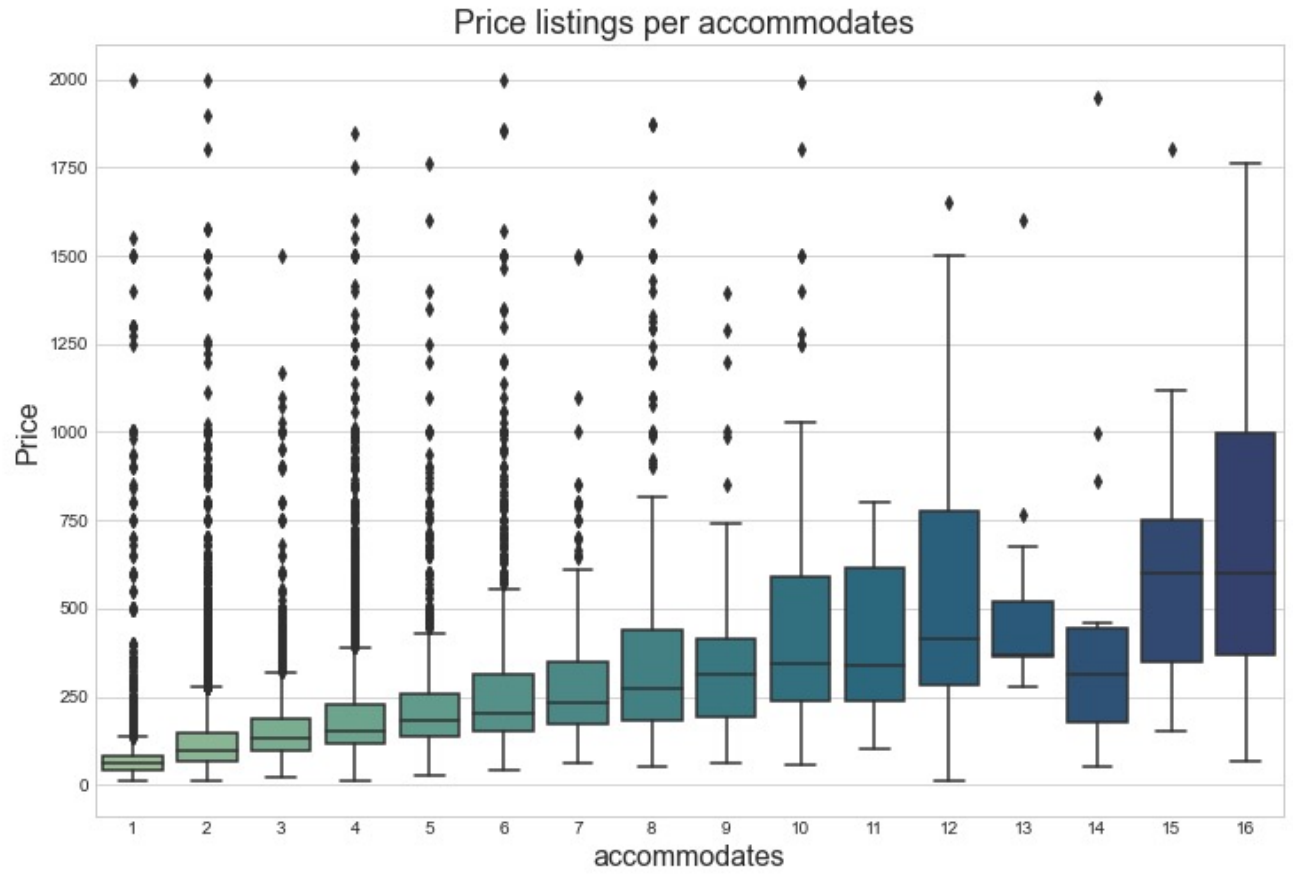
Feature analysis (price)



neighbourhood_group_cleansed	count	mean	max
Bronx	1125	101.655111	2000
Brooklyn	14678	132.591497	7184
Manhattan	16575	202.253152	10000
Queens	5431	115.961885	10000
Staten Island	340	116.617647	1200

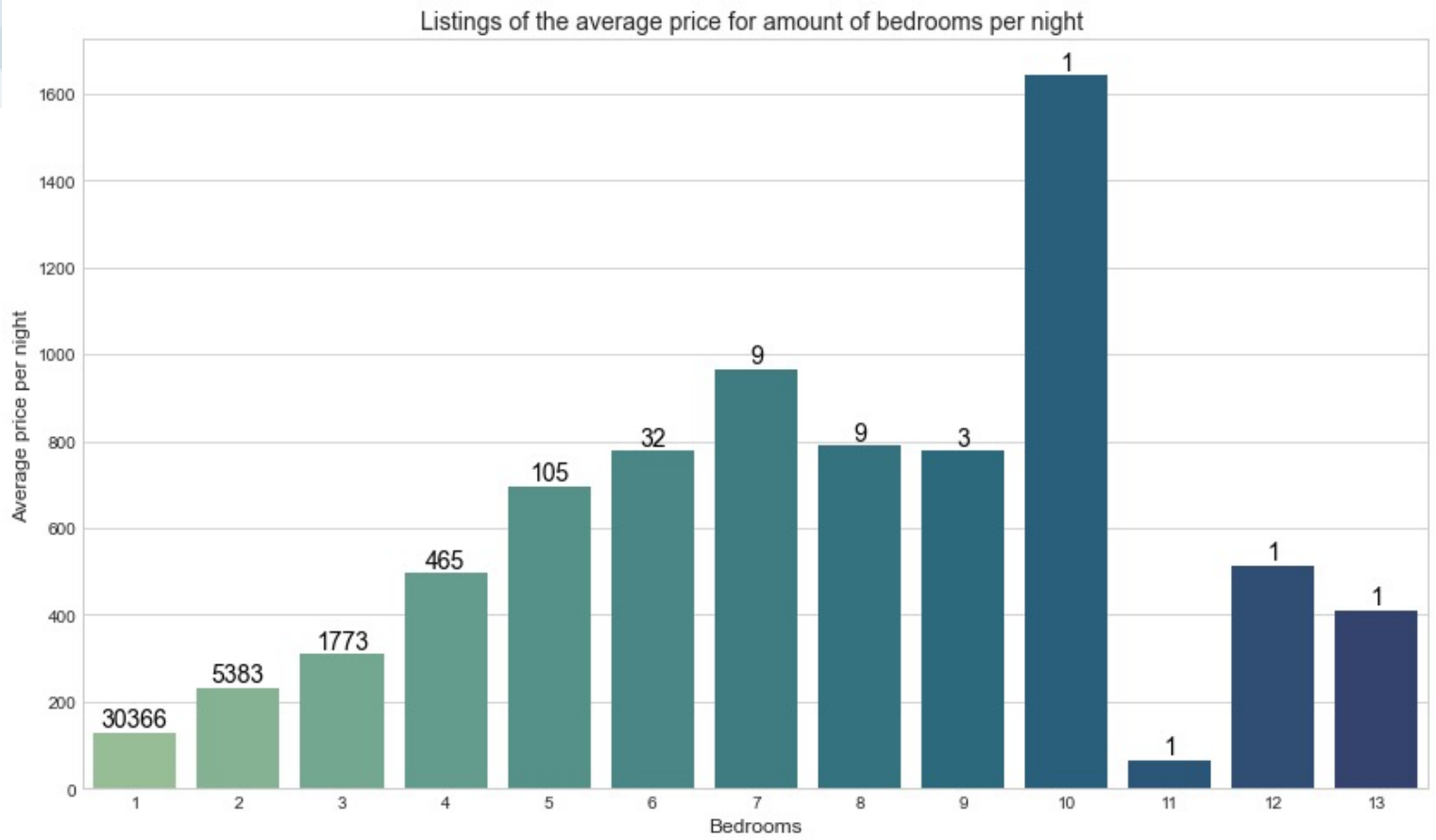
- Manhattan is the neighborhood that has most listings, followed by Brooklyn. The average price in Manhattan is 202 dollars and Brooklyn is 132. most of the prices on the boxplot are shrinking between 250 and 100 dollars. Even though Staten Island has the lowest number of listings, Bronx has the lowest average price with 101 dollars.

Feature analysis (price)



accommodates	count	mean	max
1	6545	83.099618	10000
2	17435	126.980327	10000
3	3698	158.765819	2750
4	5848	213.040869	10000
5	1516	230.389842	2695
6	1723	287.161346	6000
7	347	355.881844	6500
8	517	373.735010	2500
9	62	504.806452	3000
10	179	514.240223	3557
11	23	405.130435	799
12	92	542.467391	2000
13	14	521.714286	1600
14	17	548.352941	2175
15	13	647.384615	1800
16	120	1026.508333	5250

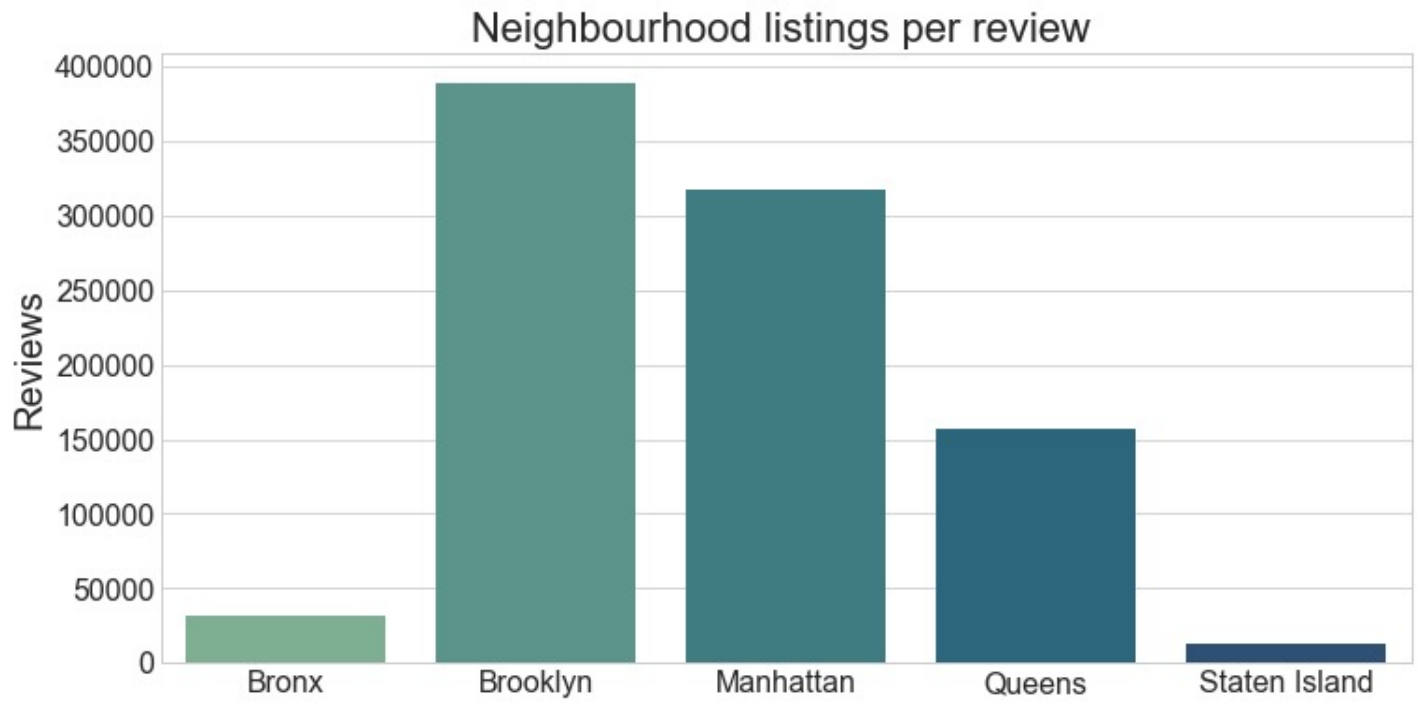
Feature analysis (price)



- Looking at the graph we can see the number of rooms per listing and the number of listings per room. most listings have only 1 room with a total of 30366. We see that as the number of rooms increases the number of listings decreases.

Sentiment analysis

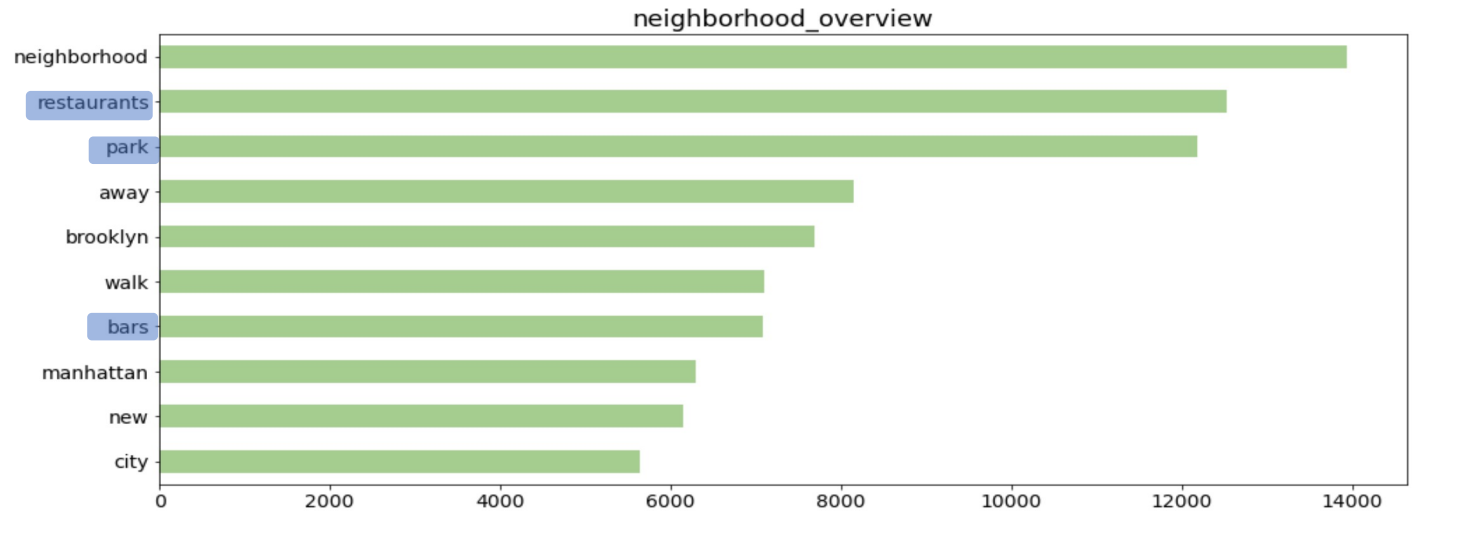
- Number of reviews of the 5 borough of New York



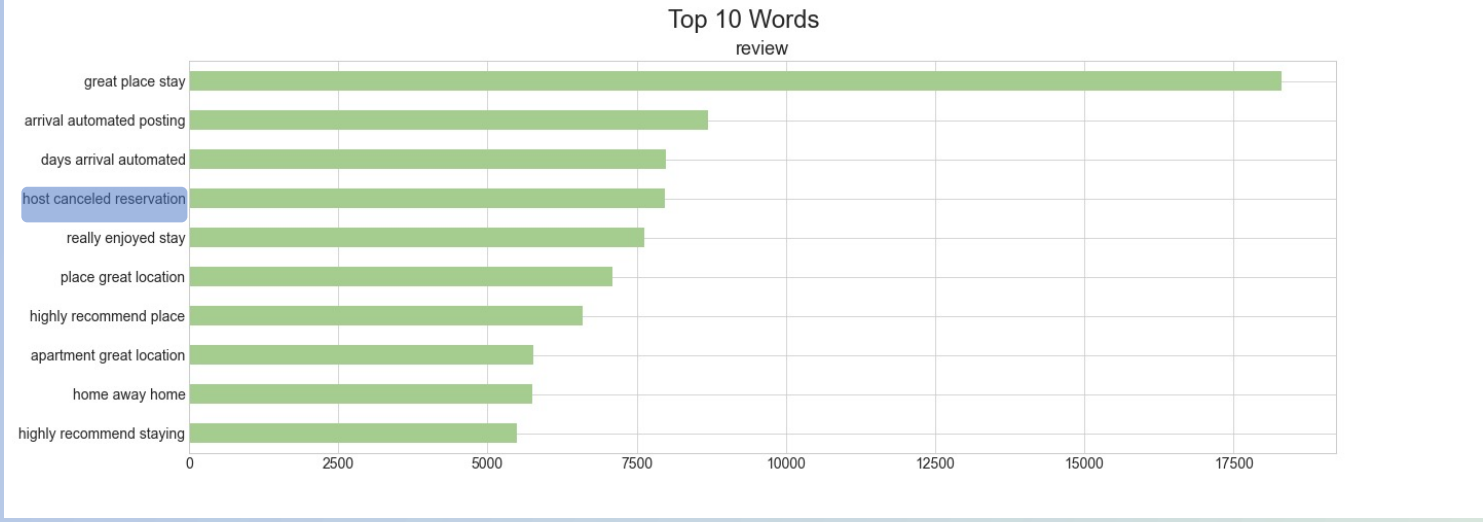
	number_of_reviews				price		
	count	sum	mean	max	mean	max	min
neighbourhood_group_cleansed							
Bronx	1125	31563	28.056000	445	101.655111	2000	11
Brooklyn	14678	389120	26.510424	678	132.591497	7184	10
Manhattan	16575	317654	19.164646	1010	202.253152	10000	10
Queens	5431	156337	28.786043	682	115.961885	10000	10
Staten Island	340	12858	37.817647	373	116.617647	1200	31

- The neighborhood with the highest number of reviews is Brooklyn with a number of 389120 reviews, although it is not the neighborhood with the highest number of listings.

Sentiment analysis CountVectorizer

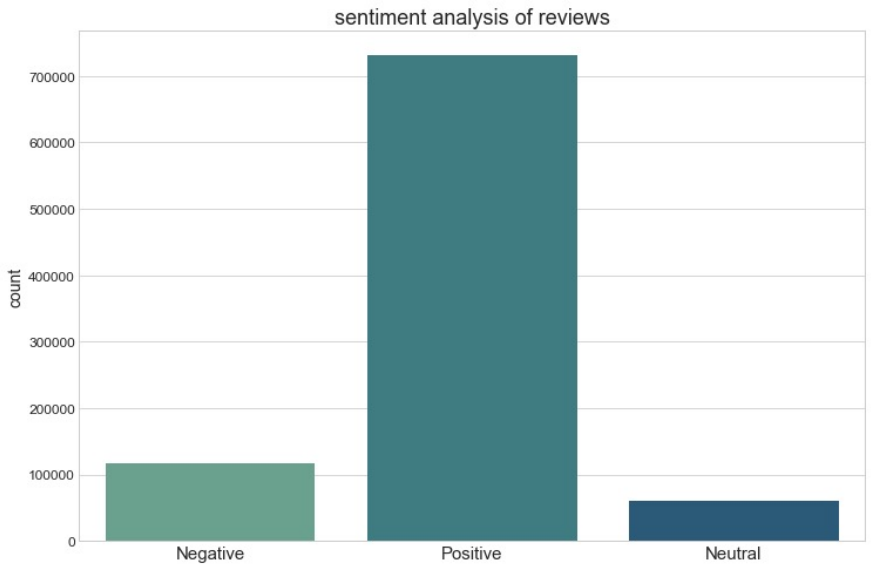


- Top 10 words most used by hosts to describe listings.



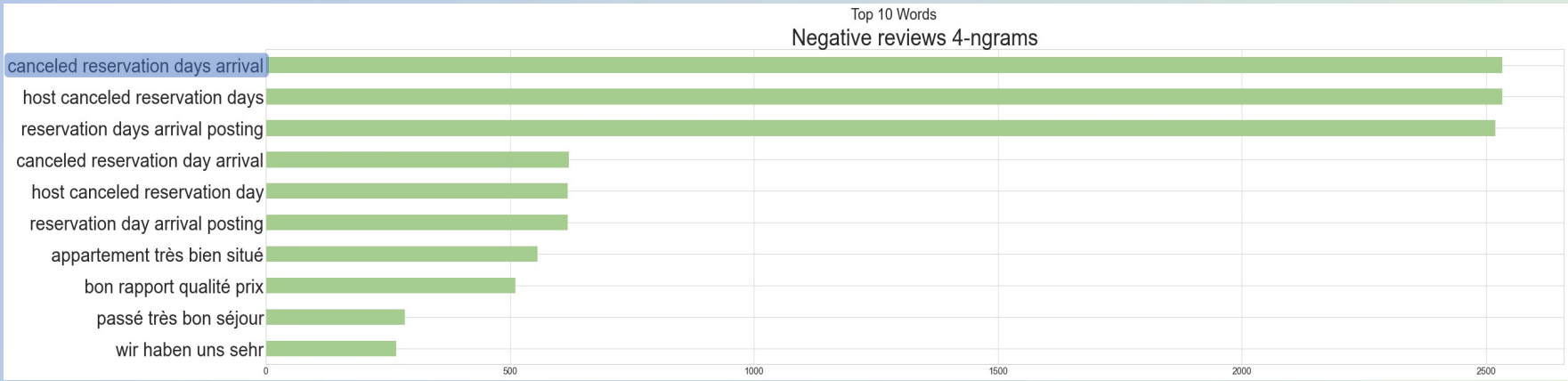
- Top 10 of the words most used by Airbnb customers

Sentiment analysis



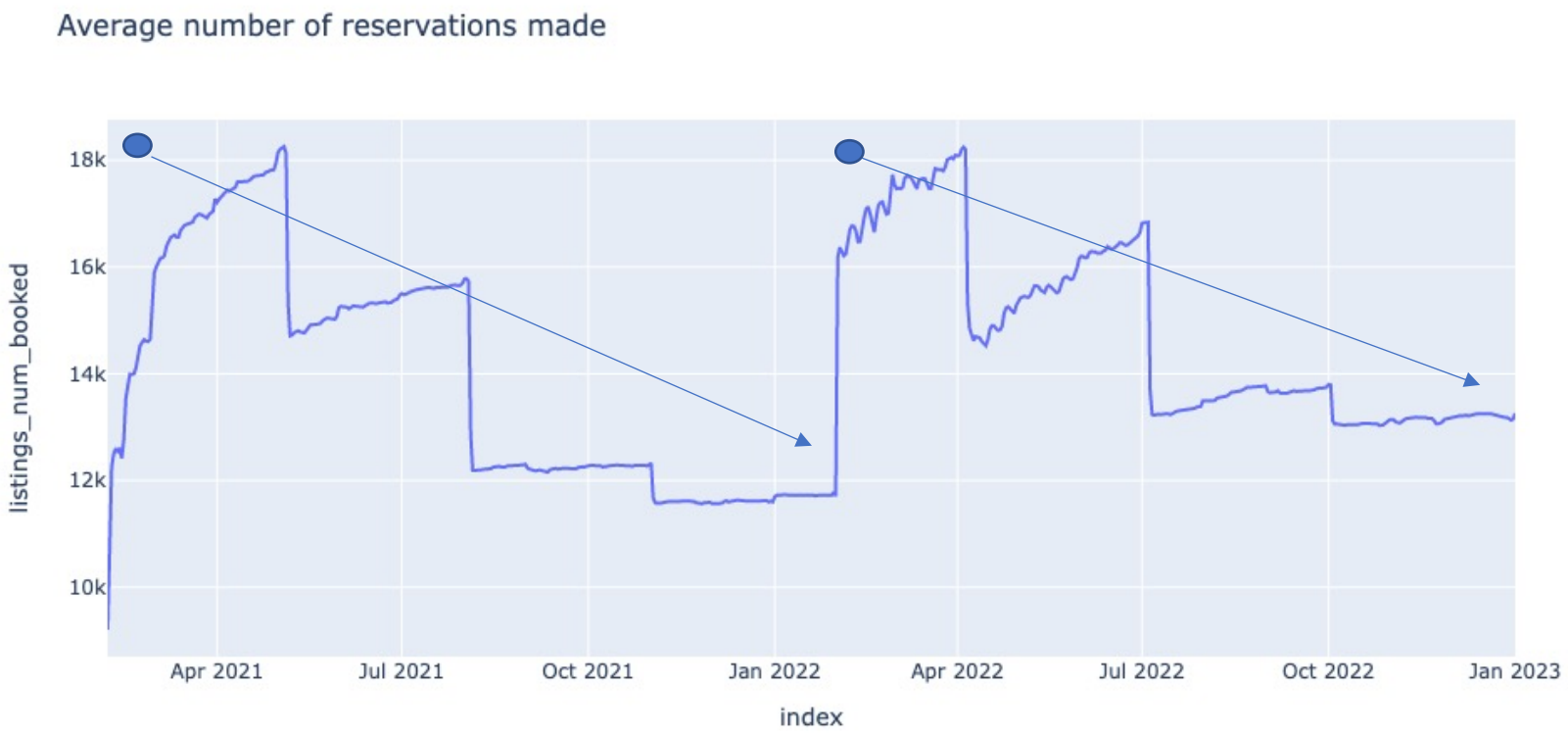
- As we saw before, most of the reviews are positive, almost 10% of the reviews are negative and a small percentage is neutral.

- When analyzing the words of the negative reviews we see a pattern of words such as cancellation of the listings, this shows that most complaints are due to the cancellation of the listing by the host.

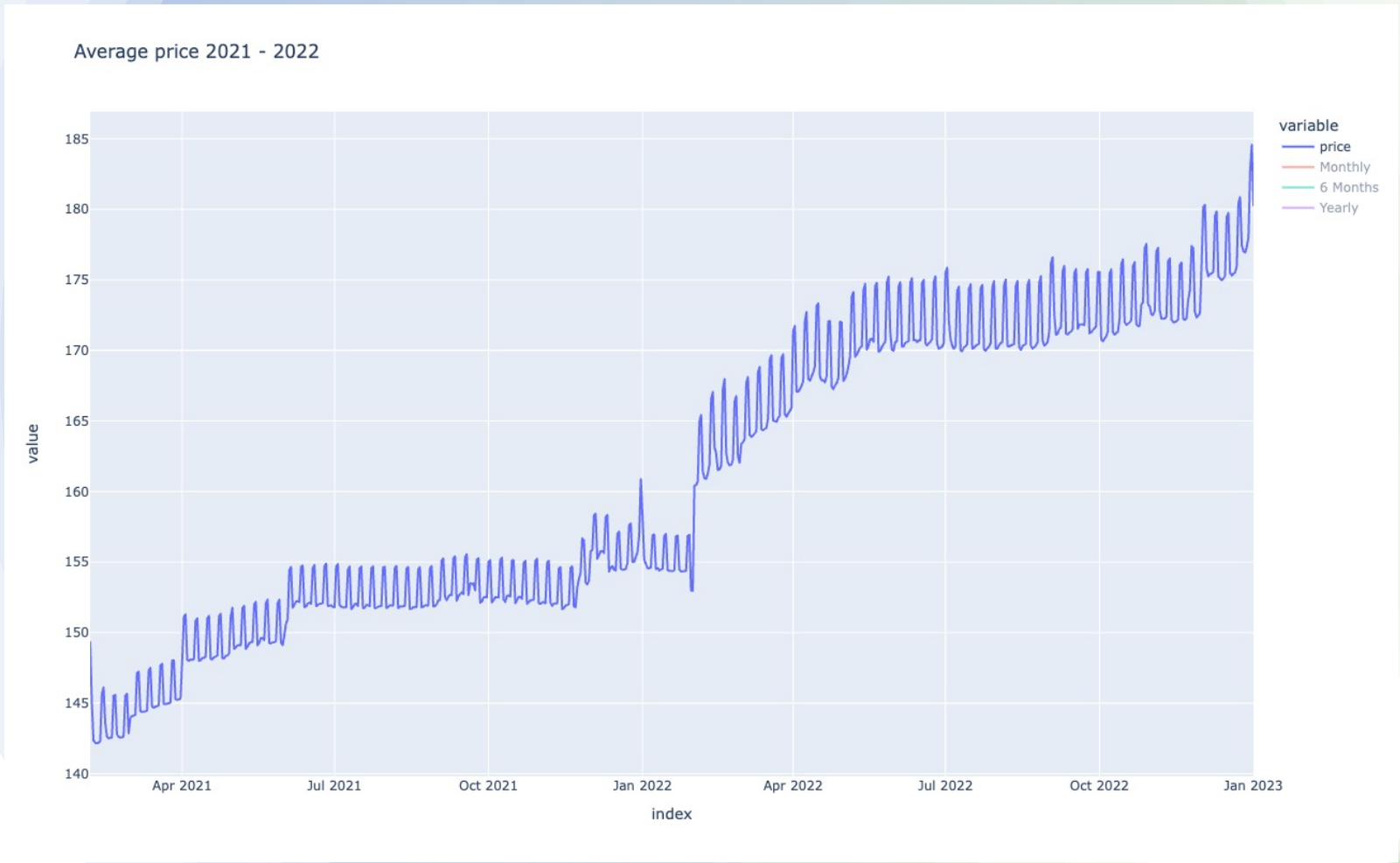


Time series (Reservations)

- When the reservations made is analyzed in a time series, a graph of descending scales is formed, this is because the reservation data has been taken on two dates (2021-02-05 to 2022-01-31 and 2022-02-01 to 2023-01-05



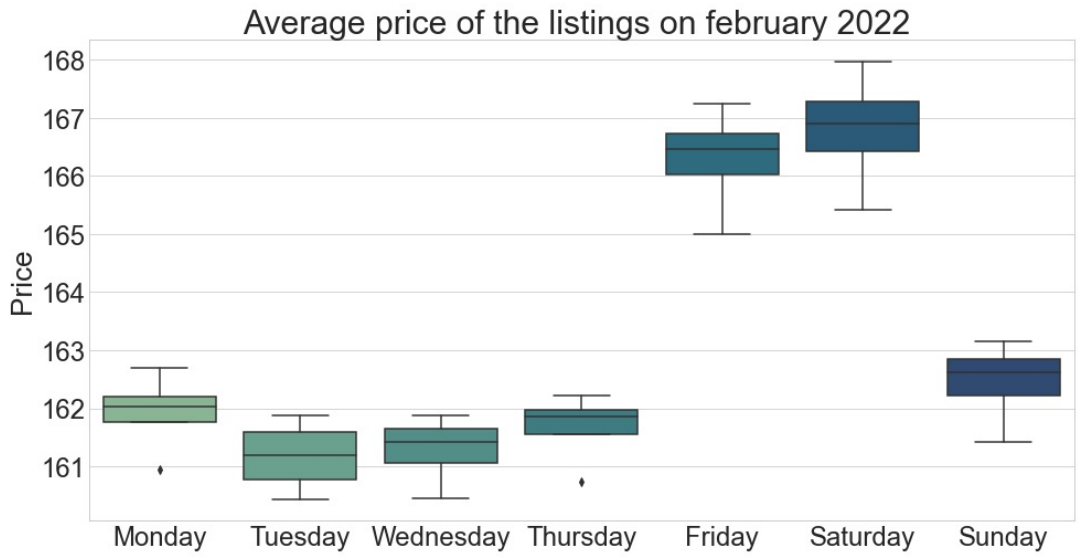
Time series (Price)

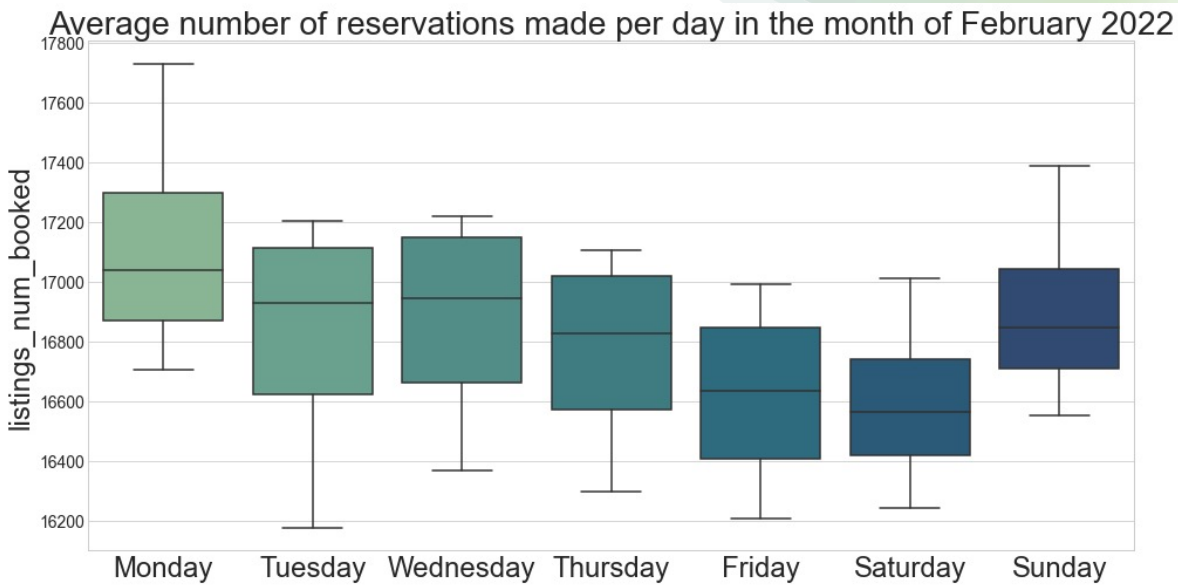


Time series

- Analysis of the price of listings and reservations made in the month of February 2022

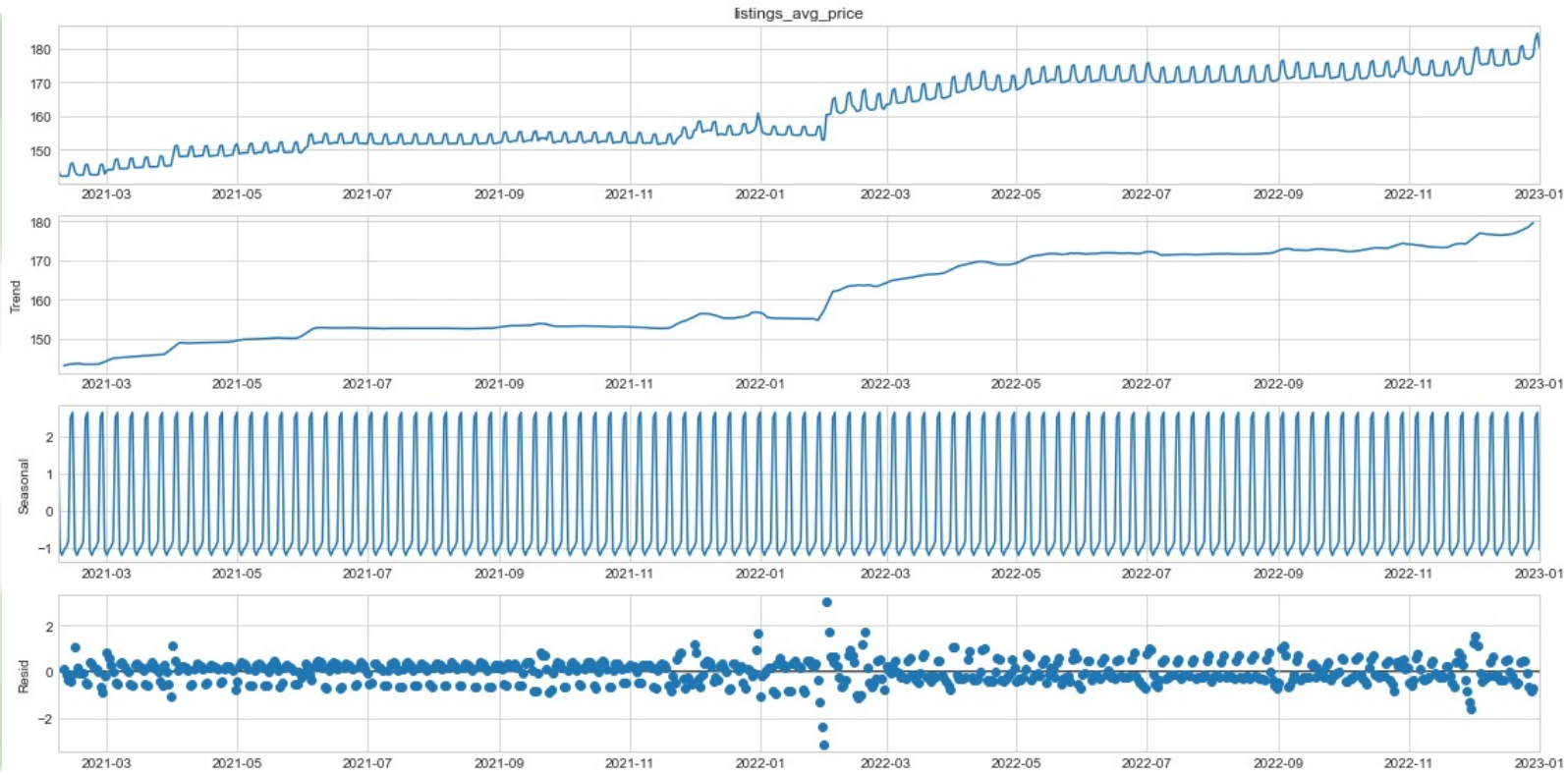
Comparing these two graphs we can see that on weekends the price of the listings increases and the reservations decrease.



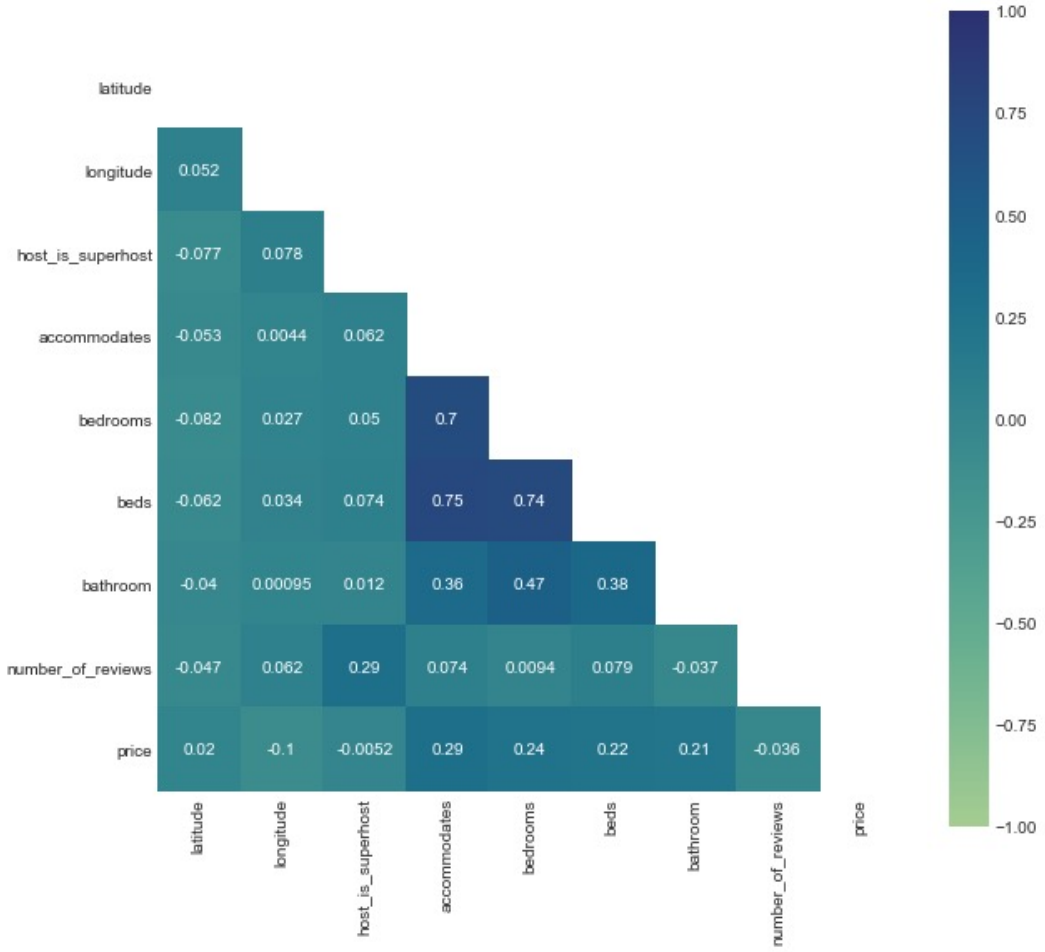


Time series (Price)

- The statsmodels library has a seasonal_() method that **lets you decompose a time series into trend, seasonality and noise in one line of code**.



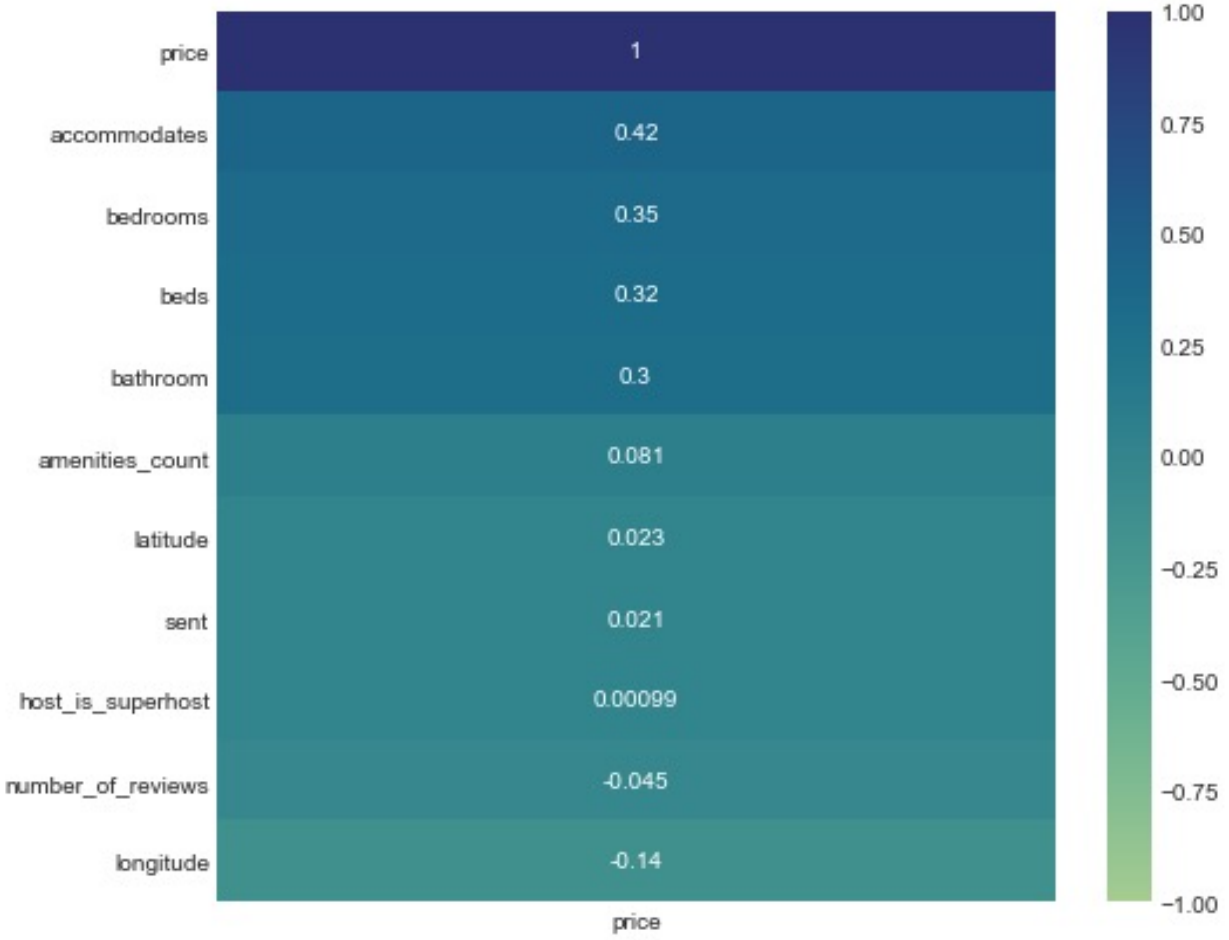
Feature selection and correlations




```
data.describe()
```

	latitude	longitude	host_is_superhost	accommodates	bedrooms	beds	bathroom	number_of_reviews	price
count	38149.000000	38149.000000	38149.000000	38149.000000	38149.000000	38149.000000	38149.000000	38149.000000	38149.000000
mean	40.729015	-73.948180	0.204094	2.794936	1.290624	1.567774	1.154159	23.789143	159.436027
std	0.055899	0.051352	0.403044	1.870302	0.673828	1.063000	0.443529	50.998635	292.580710
min	40.504560	-74.249840	0.000000	1.000000	1.000000	1.000000	0.000000	0.000000	10.000000
25%	40.688920	-73.983080	0.000000	2.000000	1.000000	1.000000	1.000000	1.000000	68.000000
50%	40.724770	-73.954320	0.000000	2.000000	1.000000	1.000000	1.000000	4.000000	109.000000
75%	40.762540	-73.929510	0.000000	4.000000	1.000000	2.000000	1.000000	21.000000	175.000000
max	40.914350	-73.710870	1.000000	16.000000	13.000000	24.000000	8.000000	1010.000000	10000.000000

Feature selection and correlations



- Dropping price values higher than 7000
- creating two new features (amenities_count and sent)
- Dummies for categorical features(room_type and neighbourhood_group)

Prediction models

- For the purpose of this project, the data will be divided into two parts, training with 75% and 25% for testing.
- Different types of prediction models for time series and regression will also be used.
- Metrics used for the models are going to be RMSE, MSE and R-SQUARED.

ARIMA

fbprophet

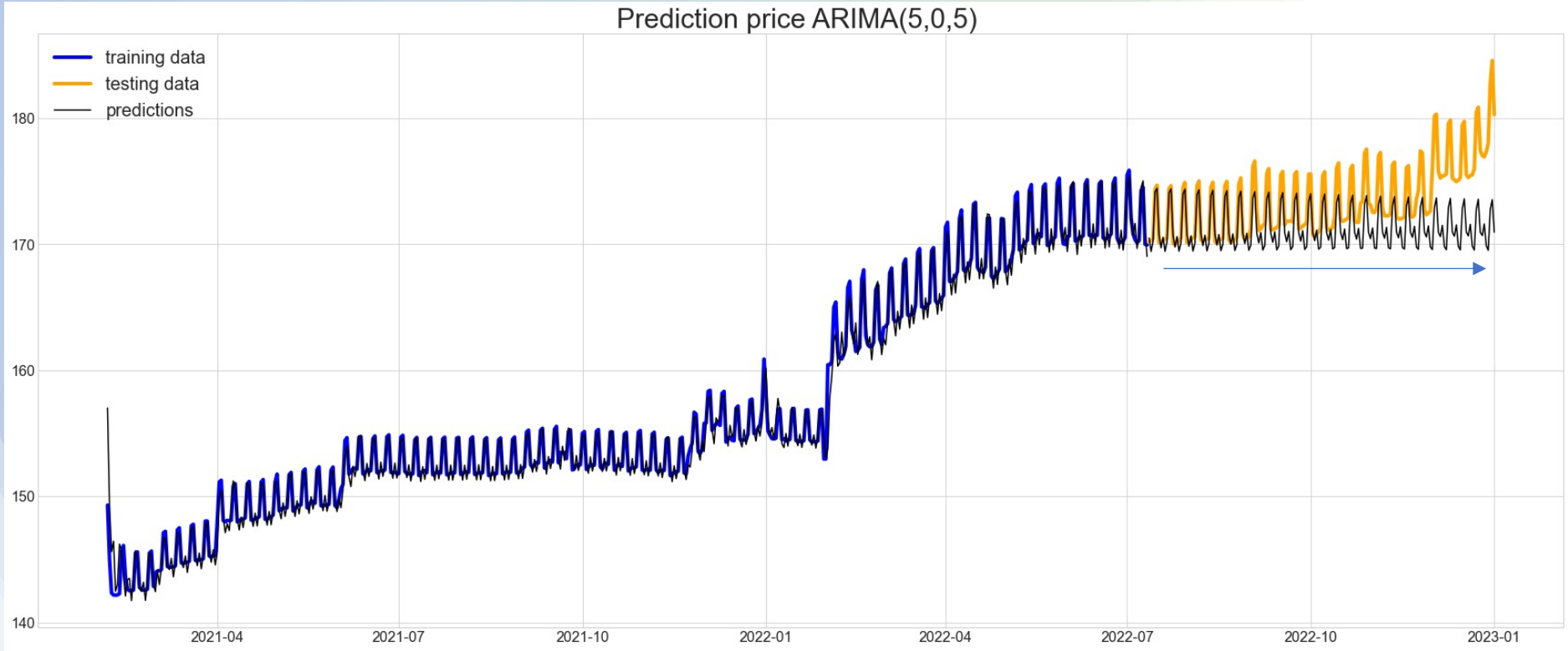
sklearn

Time series (Arima)

ARIMA (5,0,5)	TRAIN	TEST
AIC	1214.67	
MSE	0.76	10.19
RMSE	0.87	3.19

	price	predictions
2022-07-12	170.211308	170.498552
2022-07-13	170.289674	169.425841
2022-07-14	170.450913	170.230030
2022-07-15	174.369141	174.154001
2022-07-16	174.682605	174.440148
...	...	...
2022-12-28	177.325255	169.870177
2022-12-29	178.026148	169.520832
2022-12-30	182.766028	172.756131
2022-12-31	184.568790	173.534948
2023-01-01	180.275259	170.952623

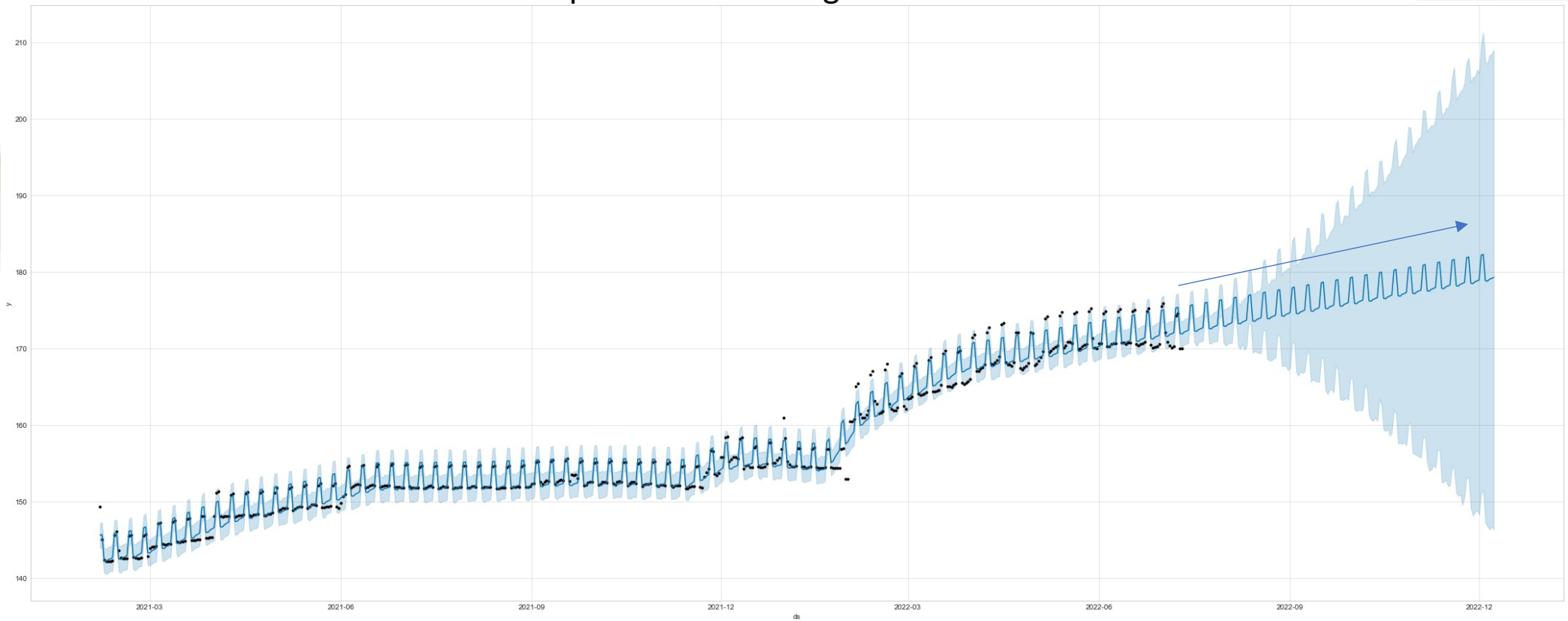
174 rows × 2 columns



Time series (Facebook prophet)

Train	Test
2021-02-05 **to** 2022-07-11	2022-07-12 **to** 2023-01-01

Forecasts prediction training data after 5 months



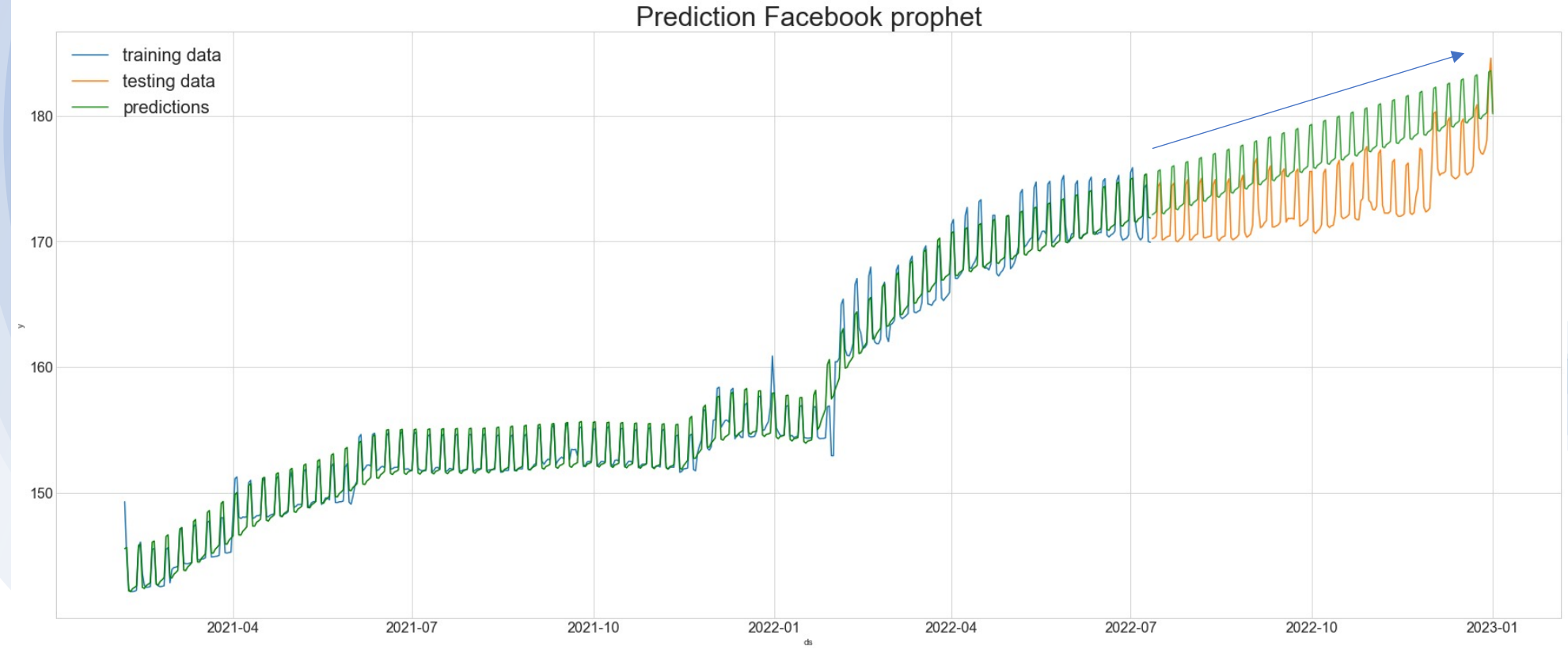
https://facebook.github.io/prophet/docs/quick_start.html

Time series (Facebook prophet)

	TRAIN	TEST
MSE	0.74	14.85
RMSE	0.86	3.85

	ds	y_test	yhat_test
522	2022-07-12	170.211308	172.122344
523	2022-07-13	170.289674	172.225491
524	2022-07-14	170.450913	172.360394
525	2022-07-15	174.369141	175.593593
526	2022-07-16	174.682605	175.691774
...	...	...	...
691	2022-12-28	177.325255	180.107501
692	2022-12-29	178.026148	180.242404
693	2022-12-30	182.766028	183.475603
694	2022-12-31	184.568790	183.573784
695	2023-01-01	180.275259	180.141905

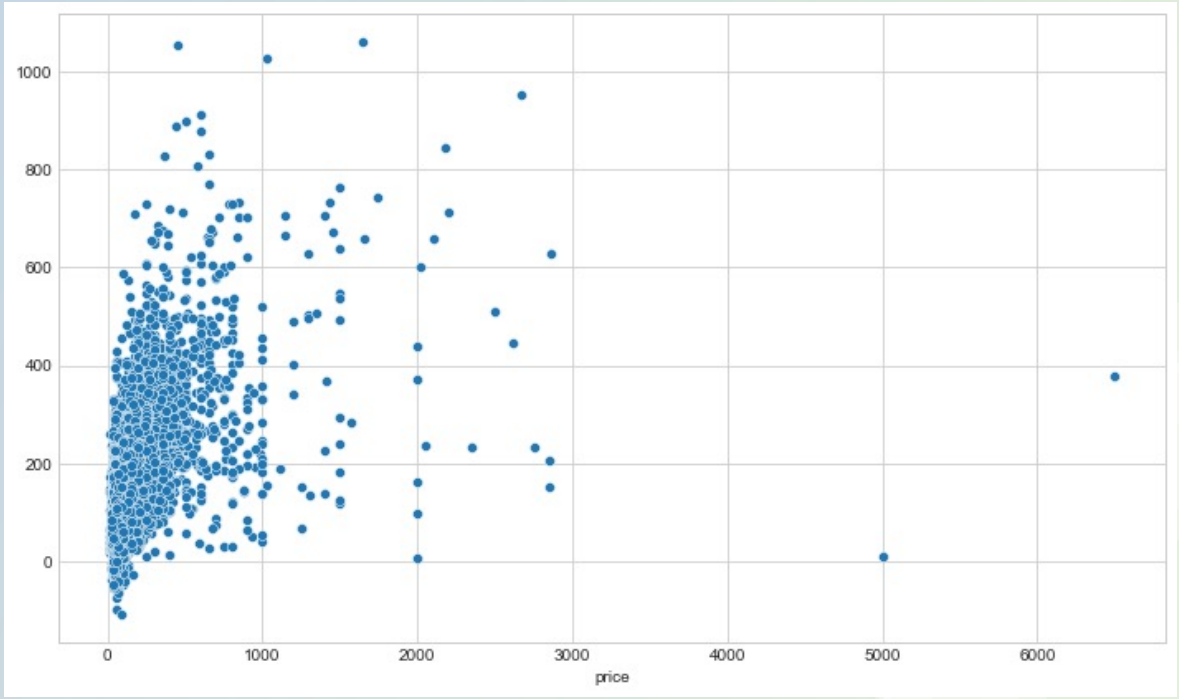
174 rows × 3 columns



Regression models (Linear Regression)

	Feature	Coefficient
3	accommodates	64.338171
6	bathroom	38.952058
4	bedrooms	12.476649
8	amenities_count	3.945961
2	host_is_superhost	1.552051
9	sent	-1.466077
7	number_of_reviews	-10.853057
5	beds	-11.117331
0	latitude	-16.381539
1	longitude	-23.784910

Linear Regression	TRAIN	TEST
Cross validation	0.26364	0.25279
R-Squared	0.26396	0.24863
RMSE	178.01	167.94



Regression models (Ridge and Lasso)

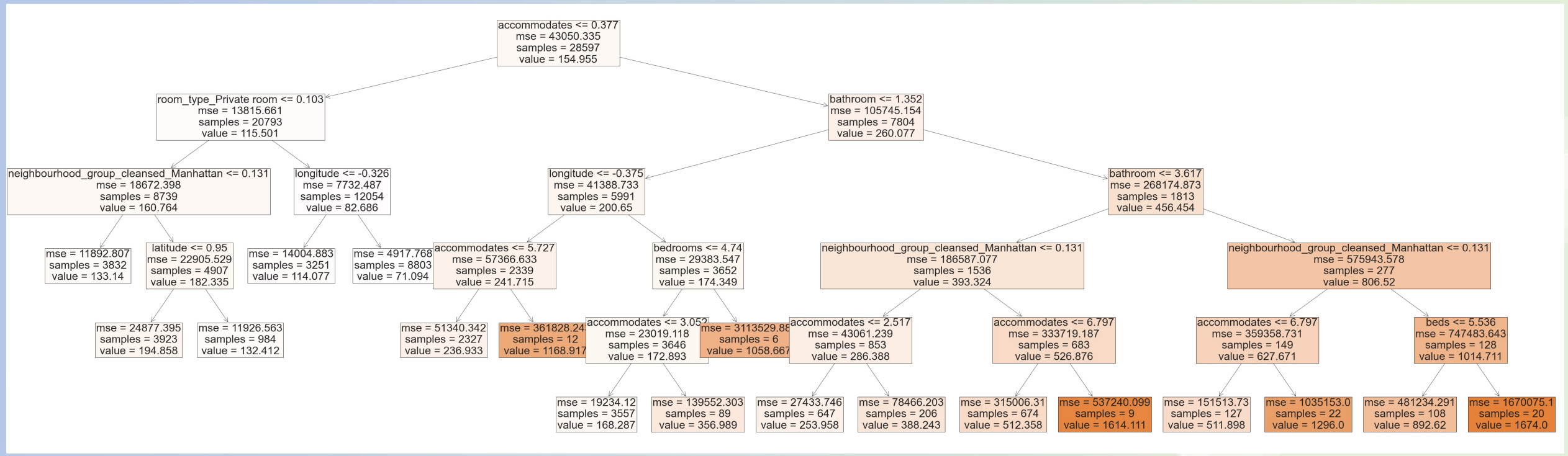
Ridge(alpha=10)	TRAIN	TEST
R-Squared	0.26396	0.24865
RMSE	178.01	167.94

Lasso(alpha=20)	TRAIN	TEST
R-Squared	0.22394	0.21828
RMSE	182.78	171.3

Regression models (Decision Tree)

parameters	
max_depth=5,	max_features=16,
max_leaf_nodes=18,	min_samples_leaf=6,
min_samples_split=20	

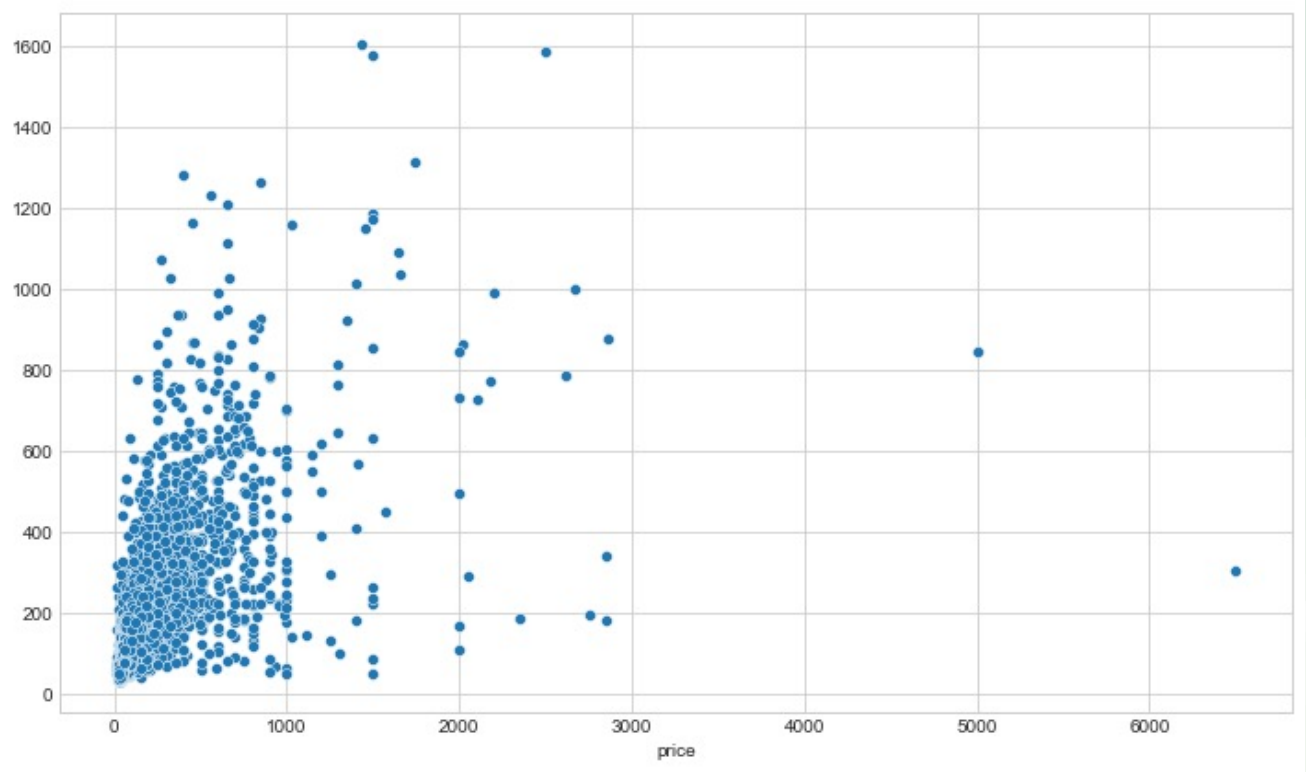
Decision tree	TRAIN	TEST
R-Squared	0.31322	0.25776
RMSE	171.95	166.92



Regression models (Random forest)

parameters	
max_depth=30, n_estimators=300	max_features='sqrt',

Random forest	TRAIN	TEST
R-Squared	0.91576	0.38404
RMSE	59.81	152.44



Regression models (Summary)

Lineal Regression	TRAIN	TEST
Cross validation	0.26364	0.25279
R-Squared	0.26396	0.24863
RMSE	178.01	167.94

Ridge(alpha=10)	TRAIN	TEST
R-Squared	0.26396	0.24865
RMSE	178.01	167.94

Lasso(alpha=20)	TRAIN	TEST
R-Squared	0.22394	0.21828
RMSE	182.78	171.3

Decision tree	TRAIN	TEST
R-Squared	0.31322	0.25776
RMSE	171.95	166.92

Random forest	TRAIN	TEST
R-Squared	0.91576	0.38404
RMSE	59.81	152.44

Recommendations and conclusions

To conclude with the analysis of the project, we can highlight a couple of important points:

- After analyzing some of the features of the listings we were able to see that the average price varies depending on the neighborhood or the number of accommodates. When the reviews were analyzed, I gathered that most reviews have a tendency to be positive and that the complaints are mostly due to the cancellation of the listings.
- It was observed that both price and availability have a positive trend, and that the price of listings increases on weekends, compared to availability, which decreases.
- ARIMA and Facebook prophet models performed well, but how we see these models are good just for short periods of time.
- After trying to organize the dataframes and trying various types of regression models, we can see that the models are overfitting some more than others. For example, the random forest model was one of the models that had the greatest overfitting.
- For this project I would choose the linear regression model due to the speed and the variety of options that I can have, for example being able to see the coefficients of each feature.

As recommendations, it would be to have more information about the listings to be able to make a better prediction model, such as the size of the listings.
Other valuable information would be to have the history of the listings, this in order to know how many times the listings have been rented.

https://github.com/julioh-k/airbnb