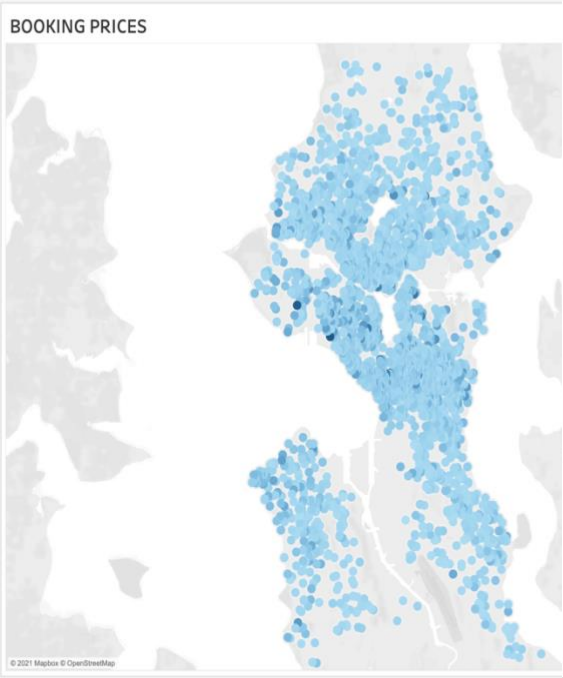
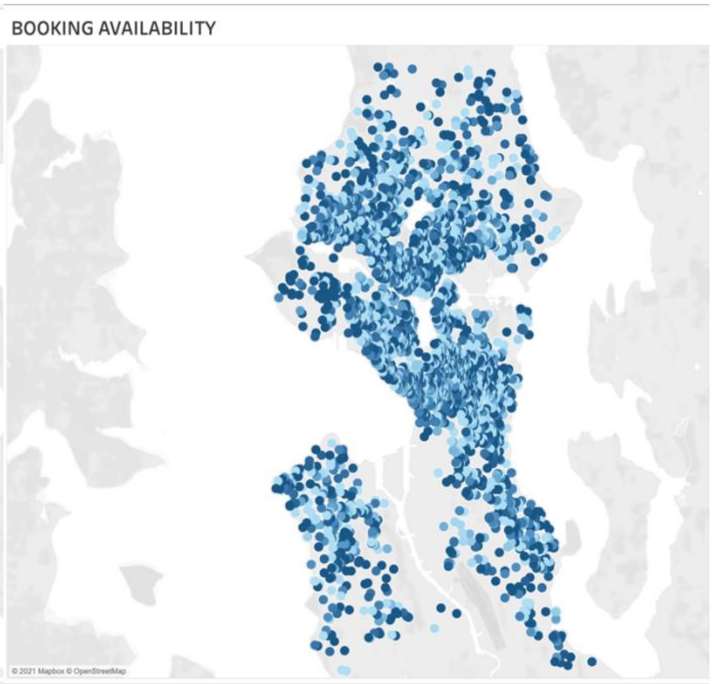
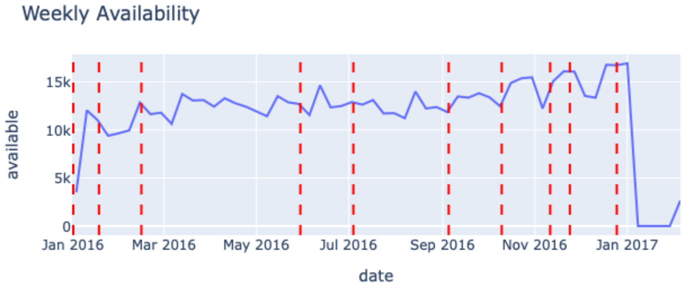
Capstone 3 Report

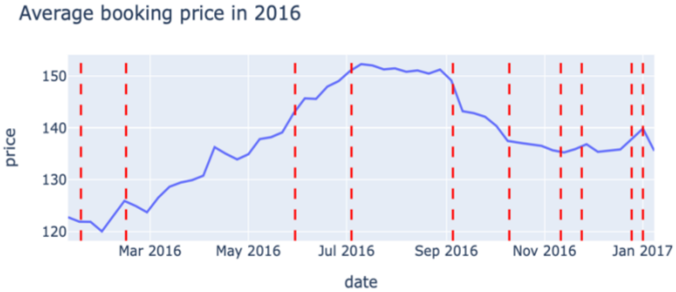
The client is moving to Seattle, and he is interested in the Airbnb market. He has some locations and prices in mind. However, he found some hosts adding additional charges like cleaning fees. He is curious about how the cleaning fee impacts the availability and the booking prices. He asks about important insight toward Airbnb markets. The Airbnb data obtained is from Seattle in 2016 and have 1,393,570 entry of listing id, date of the booking, the availability, and the prices.



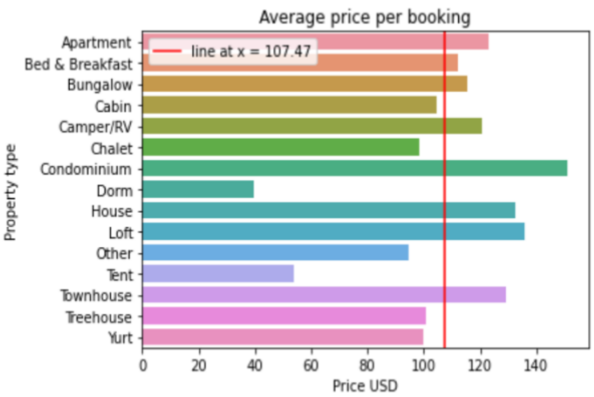


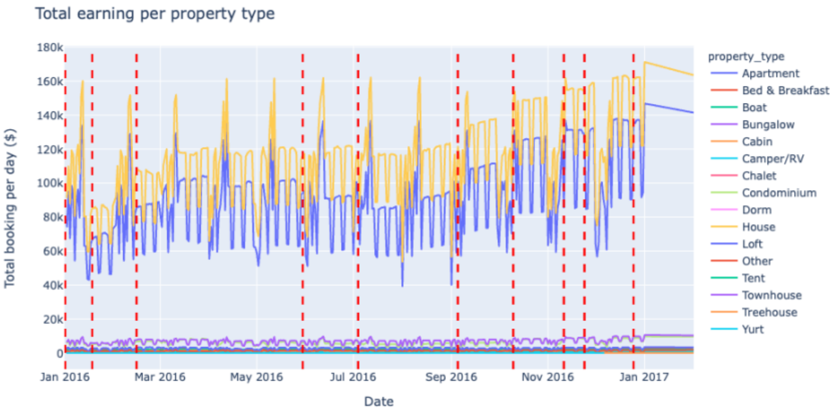
It is hard to detect an underlying pattern with the booking availability and booking prices. Interestingly, we can discern the busiest Airbnb could be those with lower average prices.



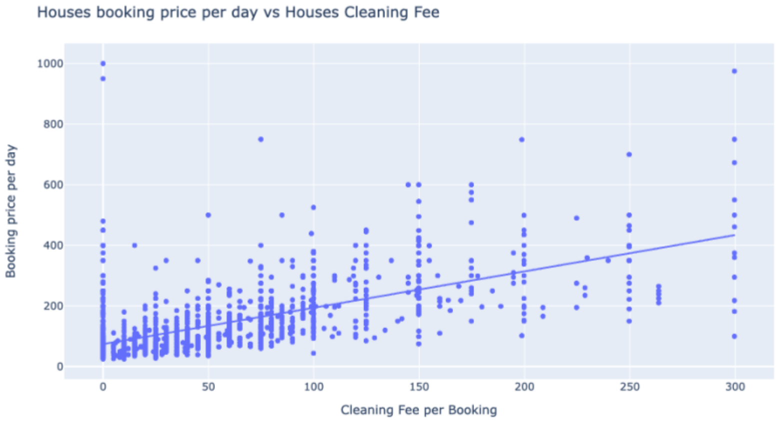


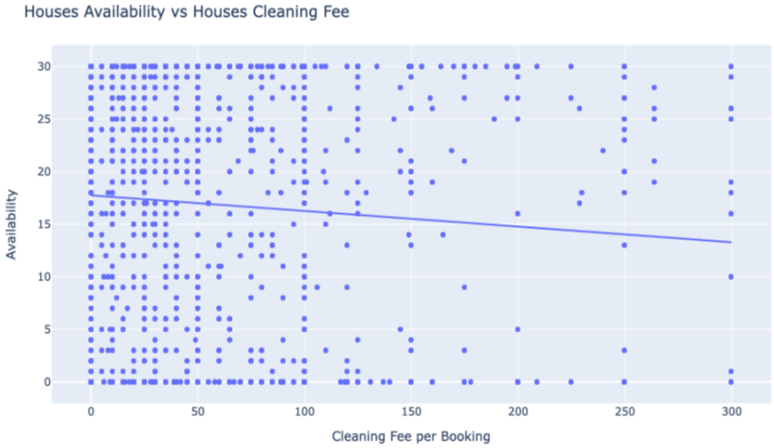
Throughout 2016 there is an underlying pattern of increasing the average price per booking between March 2016 and July 2016, and between July and September seems to be the highest average price season. This fact is crucial at the moment to select the booking price. The availability seems consistent through 2016.



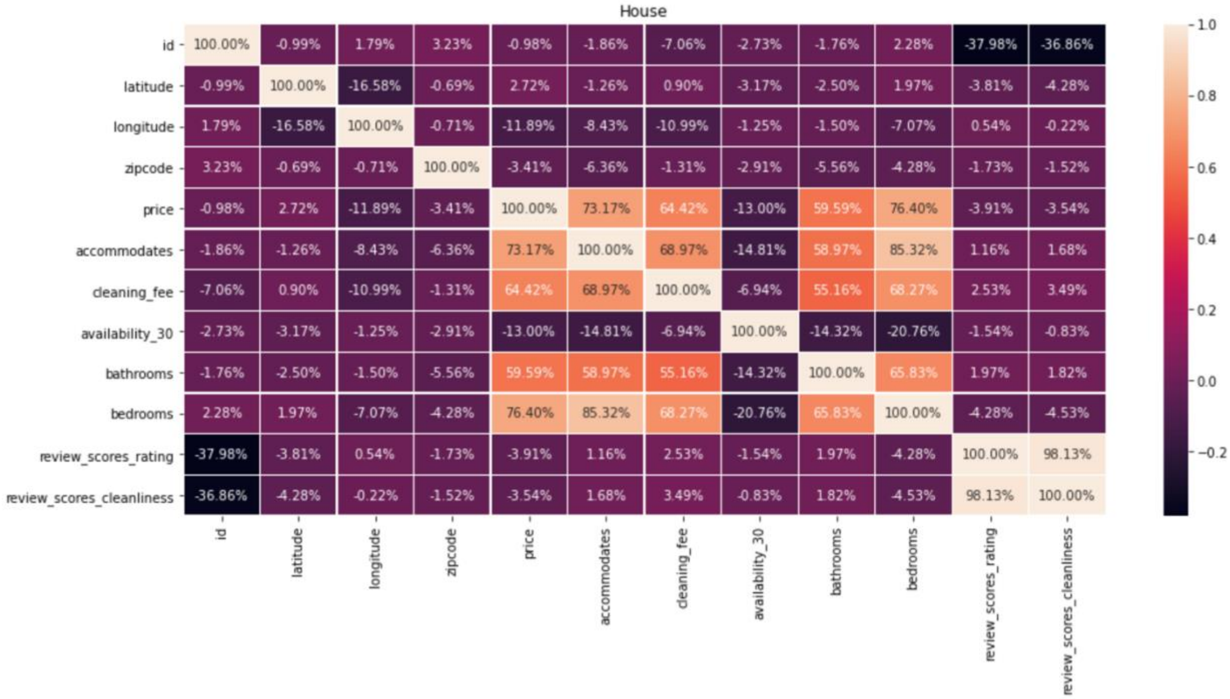


Airbnb market has a wide variety of property types like apartments, condominiums, lofts, houses, townhouses, etc. The average booking price is $ 107.47, and the most solicited are apartments and houses. I'm going to focus on these two most demanded units and study the correlation between features.





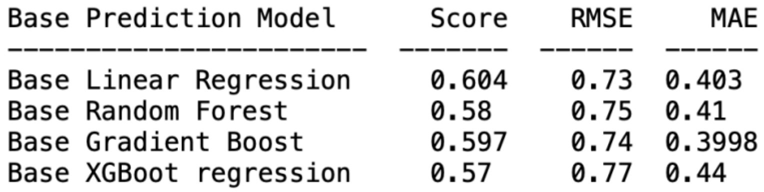
The regression plot between cleaning fee and booking price per day shows a trend indicating the highest the booking price is, the highest the cleaning fee is. Still, again, this is a moderate correlation, meaning is not always the case. On the other hand, the regression plot between cleaning fee and availability doesn't significantly correlate.



House

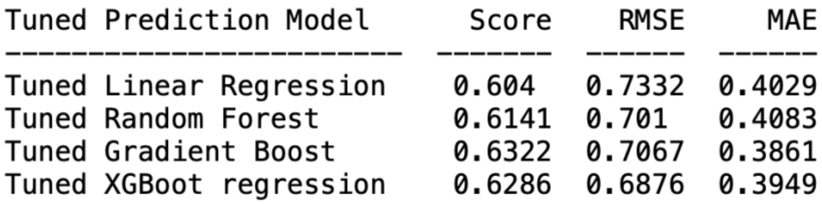
	id	latitude	longitude	zipcode	price	accommodates	cleaning_fee	availability_30	bathrooms	bedrooms	review_scores_rating	review_scores_cleanliness
id	100.00%	-0.99%	1.79%	3.23%	-0.98%	-1.86%	-7.06%	-2.73%	-1.76%	2.28%	-37.98%	-36.86%
latitude	-0.99%	100.00%	-16.58%	-0.69%	2.72%	-1.26%	0.90%	-3.17%	-2.50%	1.97%	-3.81%	-4.28%
longitude	1.79%	-16.58%	100.00%	-0.71%	-11.89%	-8.43%	-10.99%	-1.25%	-1.50%	-7.07%	0.54%	-0.22%
zipcode	3.23%	-0.69%	-0.71%	100.00%	-3.41%	-6.36%	-1.31%	-2.91%	-5.56%	-4.28%	-1.73%	-1.52%
price	-0.98%	2.72%	-11.89%	-3.41%	100.00%	73.17%	64.42%	-13.00%	59.59%	76.40%	-3.91%	-3.54%
accommodates	-1.86%	-1.26%	-8.43%	-6.36%	73.17%	100.00%	68.97%	-14.81%	58.97%	85.32%	1.16%	1.68%
cleaning_fee	-7.06%	0.90%	-10.99%	-1.31%	64.42%	68.97%	100.00%	-6.94%	55.16%	68.27%	2.53%	3.49%
availability_30	-2.73%	-3.17%	-1.25%	-2.91%	-13.00%	-14.81%	-6.94%	100.00%	-14.32%	-20.76%	-1.54%	-0.83%
bathrooms	-1.76%	-2.50%	-1.50%	-5.56%	59.59%	58.97%	55.16%	-14.32%	100.00%	65.83%	1.97%	1.82%
bedrooms	2.28%	1.97%	-7.07%	-4.28%	76.40%	85.32%	68.27%	-20.76%	65.83%	100.00%	-4.28%	-4.53%
review_scores_rating	-37.98%	-3.81%	0.54%	-1.73%	-3.91%	1.16%	2.53%	-1.54%	1.97%	-4.28%	100.00%	98.13%
review_scores_cleanliness	-36.86%	-4.28%	-0.22%	-1.52%	-3.54%	1.68%	3.49%	-0.83%	1.82%	-4.53%	98.13%	100.00%

It seems the house's data have more relationship with the booking prices. The apartments have accommodation with 46%, cleaning fees with 43%, while houses have 73% and 64% respectively. I will focus on house properties because they have stronger correlations between features, helping to have a more accurate prediction model. It is also one of the highest demands on the market. The booking prices are competitive.



```
Base Prediction Model      Score    RMSE     MAE
-----------------------  -------  ------  ------
Base Linear Regression     0.604    0.73  0.403
Base Random Forest         0.58     0.75  0.41
Base Gradient Boost        0.597    0.74  0.3998
Base XGBoot regression     0.57     0.77  0.44
```

I used four different prediction models Linear regression, Random Forest, Gradient boosting, and XGBoost. The best **base model** is the Linear Regression that scores 60% of accuracy. When it comes to the accuracy score metric, the score should not below 0.60 or 60%. If it is the case, the built model is insufficient for our data to solve the given issue. So, the ideal score should be between 0.60 and 1.0. After constructing the base model, I found the features "bedrooms," "distance," "cleaning fee," "bathrooms," "availability," and "accommodations" are the most crucial for the prediction model.



```
Tuned Prediction Model      Score    RMSE     MAE
-------------------------  -------  ------  ------
Tuned Linear Regression     0.604   0.7332  0.4029
Tuned Random Forest         0.6141  0.701   0.4083
Tuned Gradient Boost        0.6322  0.7067  0.3861
Tuned XGBoot regression     0.6286  0.6876  0.3949
```

After selecting the most important features and tuning the hyperparameter of each model, the Gradient Boost has the highest accuracy score with 63%. The prediction was close to the test dataset, having a low average prediction mistake. And because is above 60% the model is proper for solving real problems.