# Capstone 3 Report

The client is moving to Seattle and is interested in the Airbnb market. He has some locations and prices in mind, but he realized that some hosts add additional charges like cleaning fees to their price. He is curious about how the cleaning fee impacts the availability and the booking prices. The Airbnb data obtained is from Seattle in 2016, and it has 1,393,570 entry points. The dataset includes listing id, date of the booking, availability, booking prices, and cleaning fees.

The average booking prices and availability of each listing were plotted in Figures 1 and 2 to see any data pattern. In the plot, the dark dot means high average booking price and for availability means the busiest. But with the plot is hard to tell if there is a specific area that has highest prices and also hard to tell if the availability change between areas.

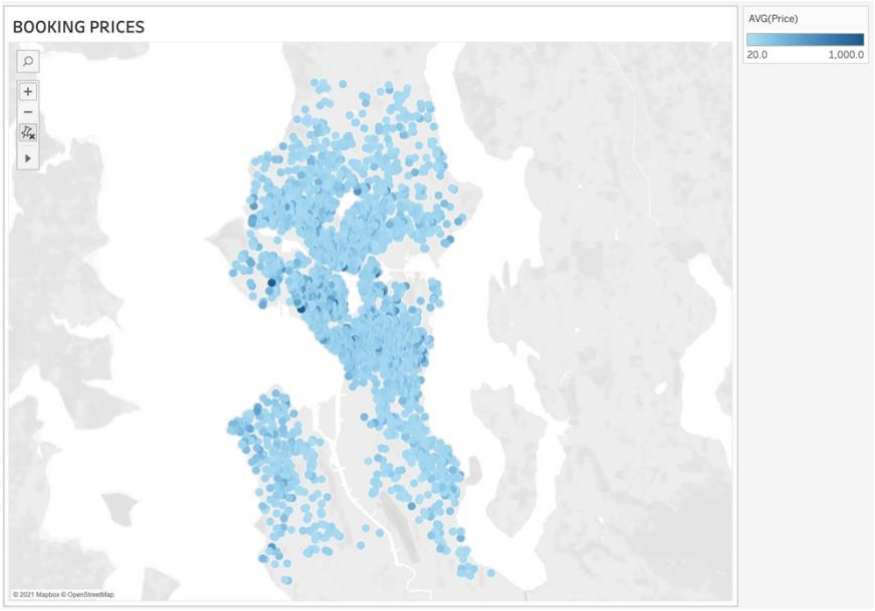


**Figure 1: Booking prices in Seattle (2016)**

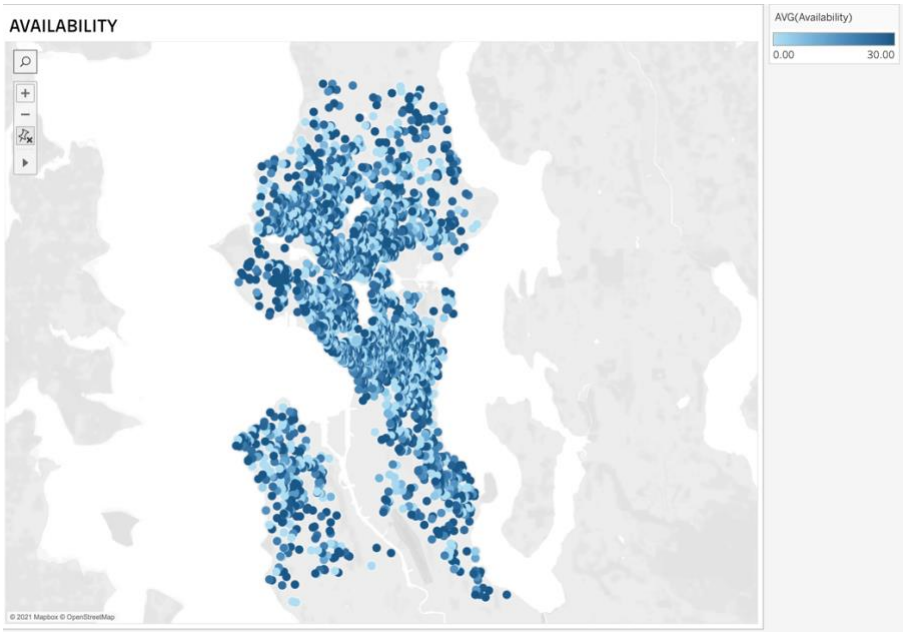


**Figure 2: Average availability in Seattle (2016)**

Figures 2 and 3 show the average booking price and availability over the year. The red lines are calendars holidays.

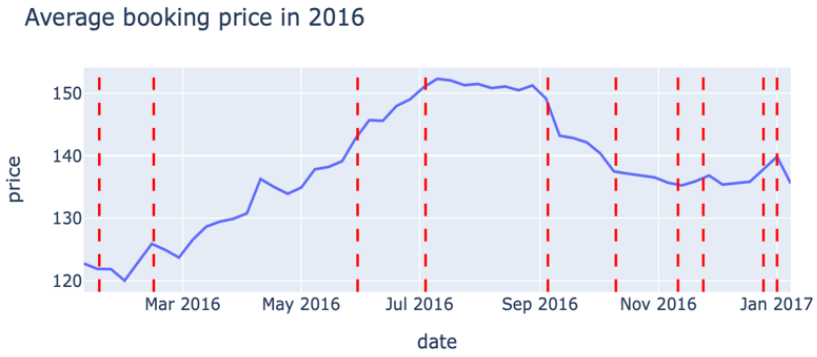


**Figure 3: Average booking price in Seattle (2016)**

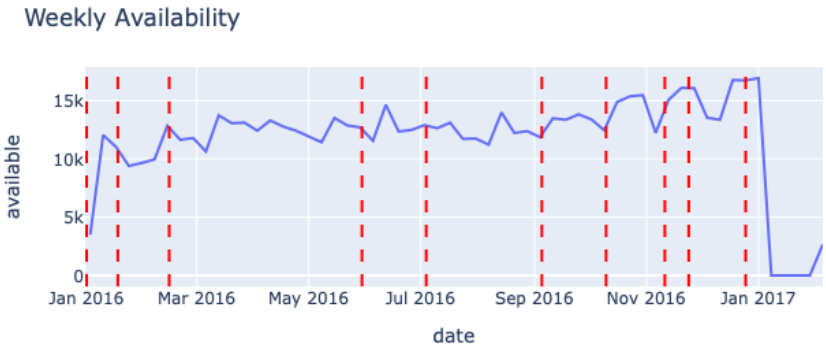


**Figure 4: Average availability in Seattle (2016)**

Figure 3 shows that the average price per booking slowly starts to increase in March. The highest prices are observed between July and September. This corresponds to the summer and early fall seasons. This information should be considered when booking a listing. The availability seems to get busier after thanksgiving but is consistent throughout the year.

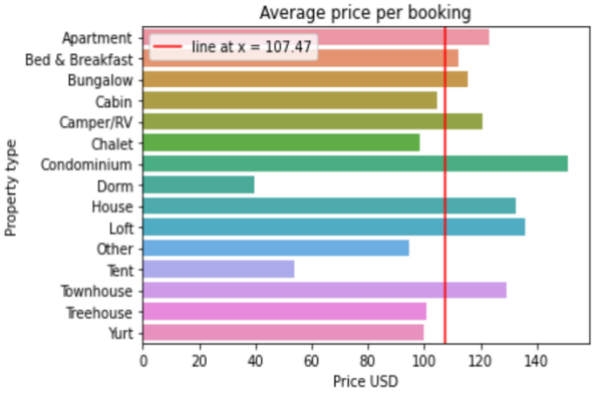


**Figure 5 Average price per booking**

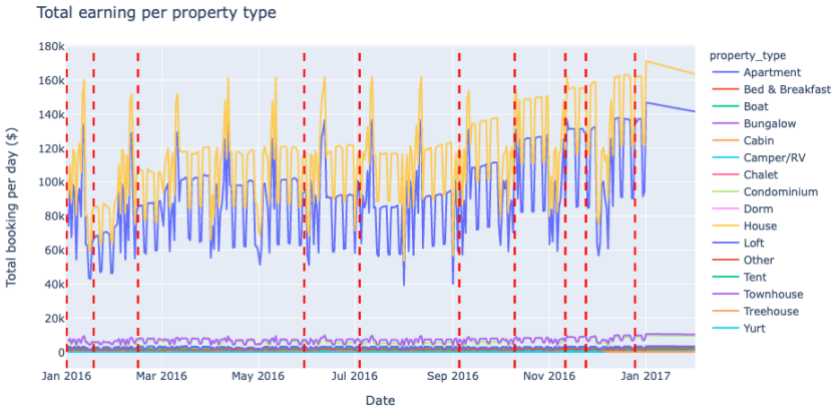


**Figure 6 Total earning per property type**

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.

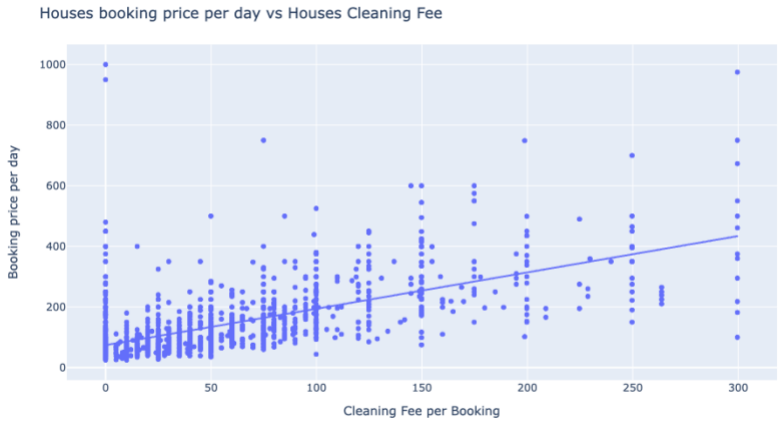


**Figure 7 House booking price per day vs House cleaning fee**

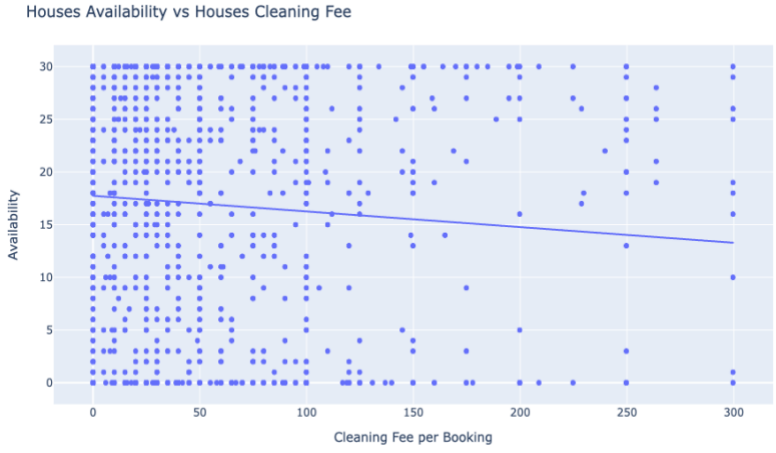


**Figure 8 House availability vs House Cleaning fee**

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.

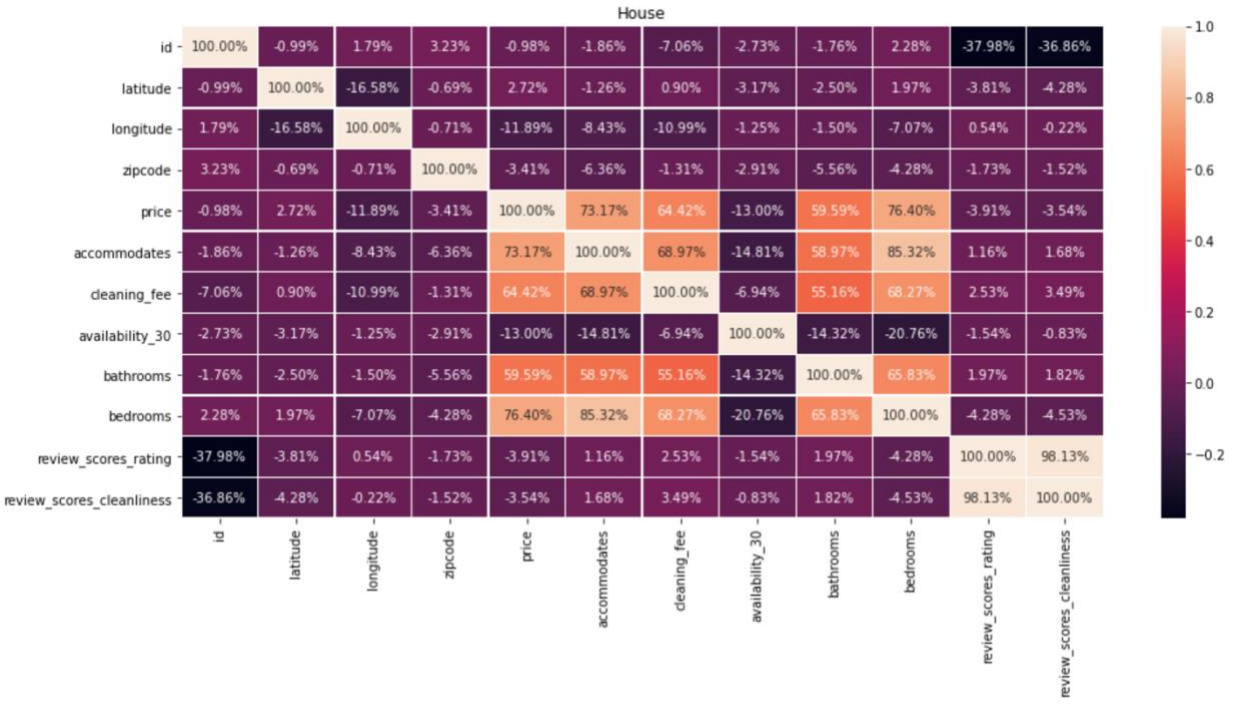


**Figure 9 House features Heatmap**

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 critical features and tuning each model's hyperparameter, 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 it is above 60%, the model is proper for solving real problems.