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 availability means the busiest. The plot is hard to tell if there is a specific area with the highest prices and hard to know if the availability change between rooms.

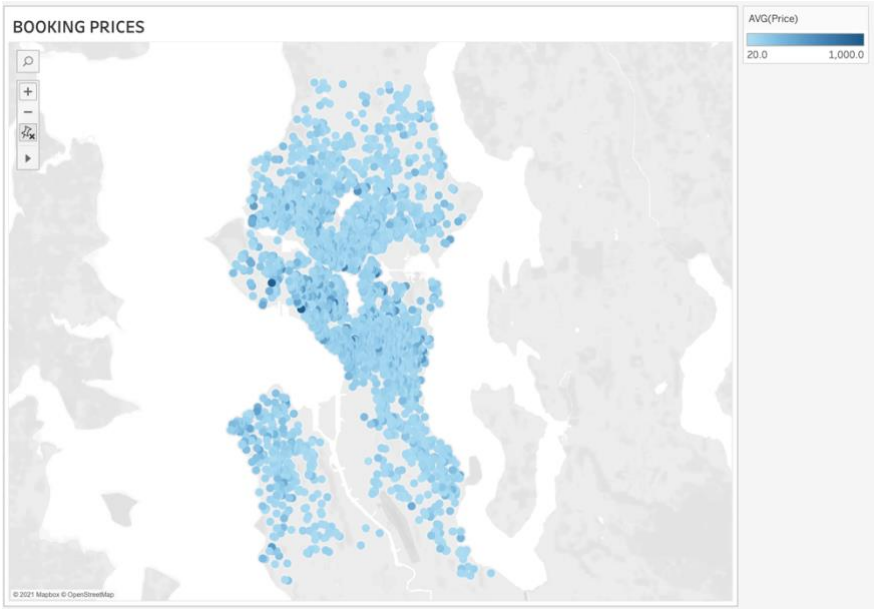

Figure 1: Booking prices in Seattle (2016)

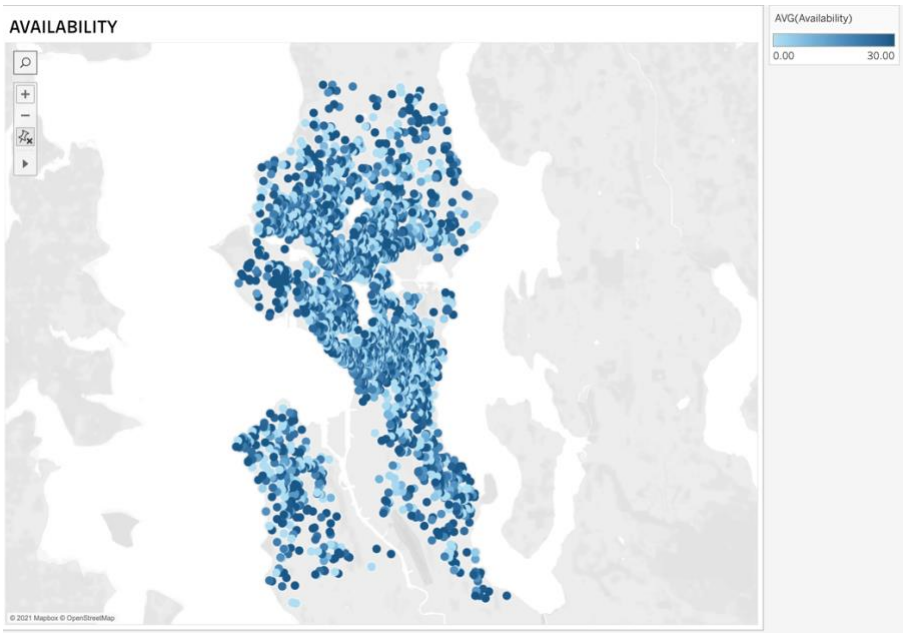

Figure 2: Average availability in Seattle (2016)

When I investigate the average prices and average availability by zip code, in figure 3, the average prices are higher in the center of Seattle than in the north and south of Seattle. The zip code 98134 has the highest average price per day at $206.6.

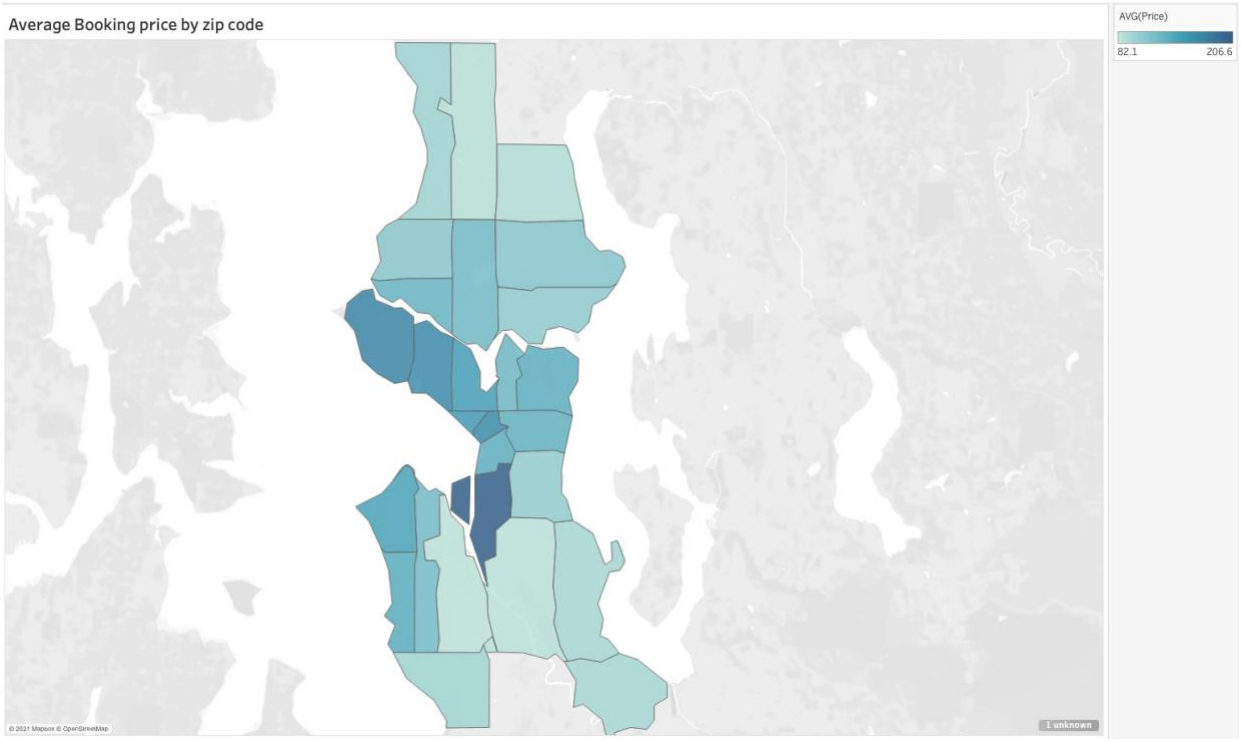

Figure 3 Average Booking Price by Zip Code

Figure 4 shows the average monthly availability; the darker the blue, the higher availability per month. And the zip code 98134 has an average of 24 days available per month. This could be because this area has the highest booking price per day.

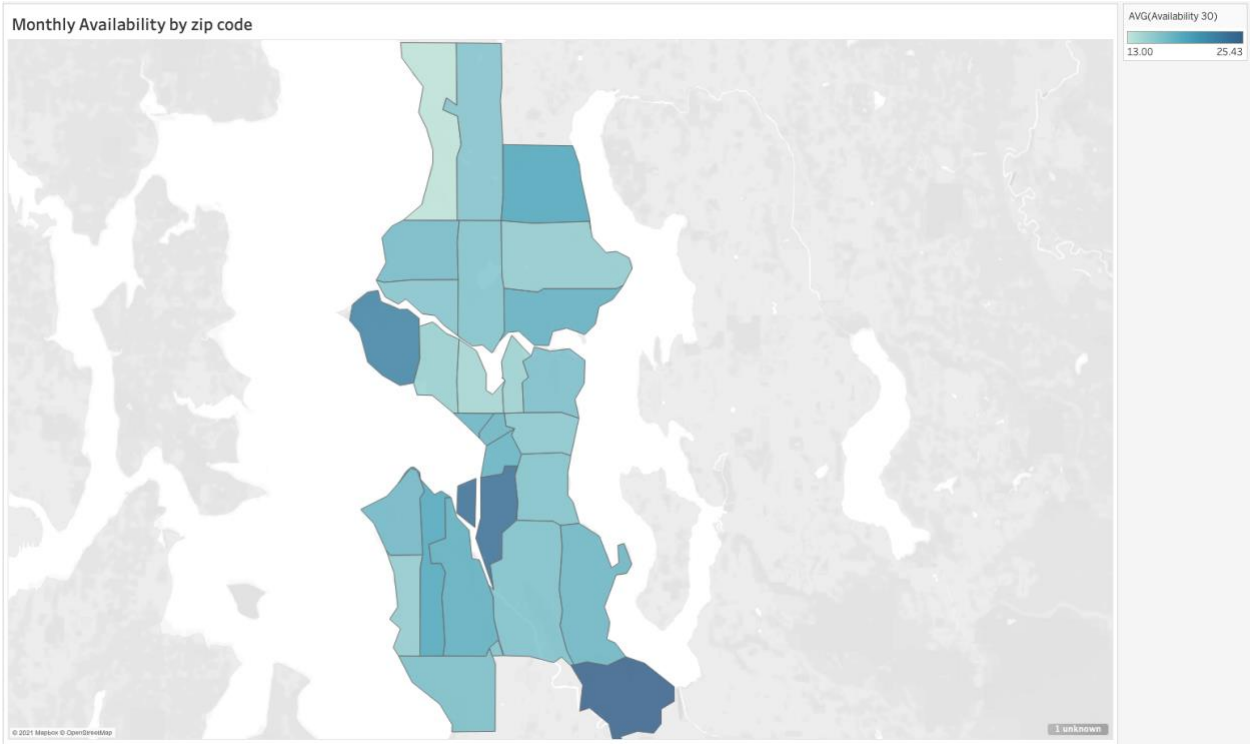

Figure 4 Average Availability per month

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

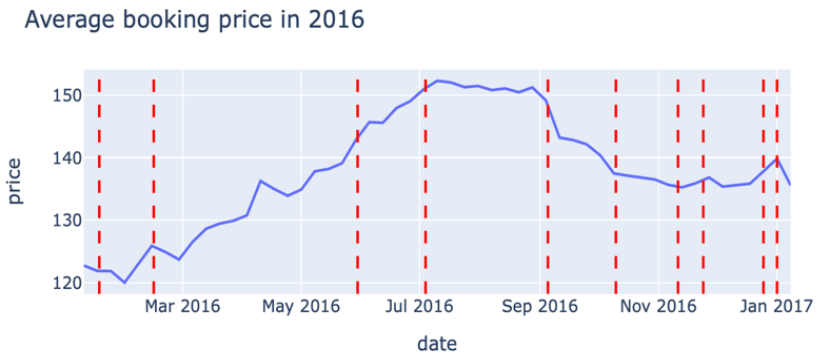

Figure 5: Average booking price in Seattle (2016)

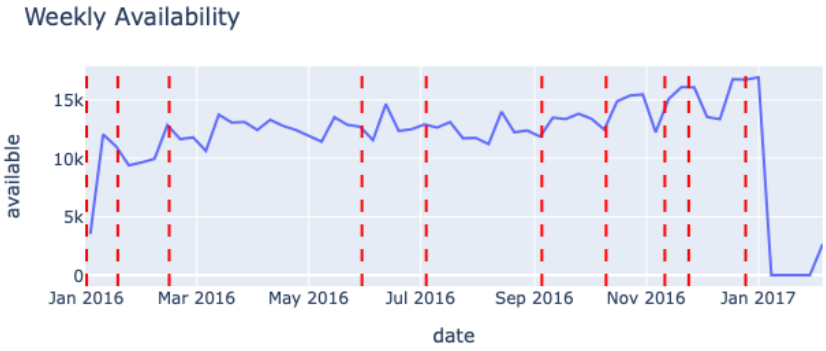

Figure 6: Average availability in Seattle (2016)

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.

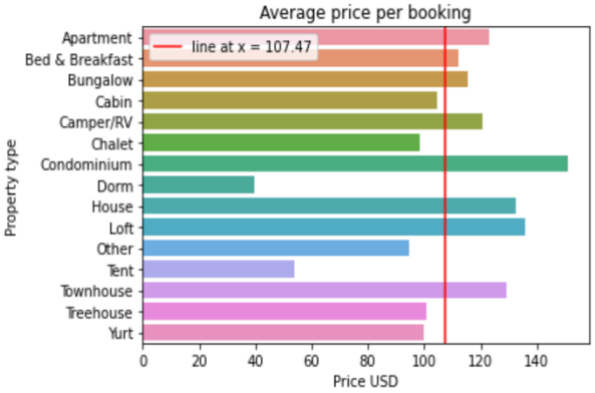

Figure 7 Average price per booking

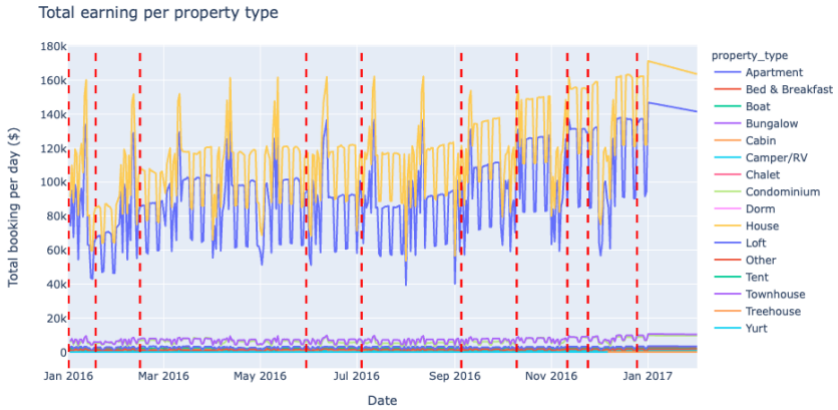

Figure 8 Total earning per property type

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.

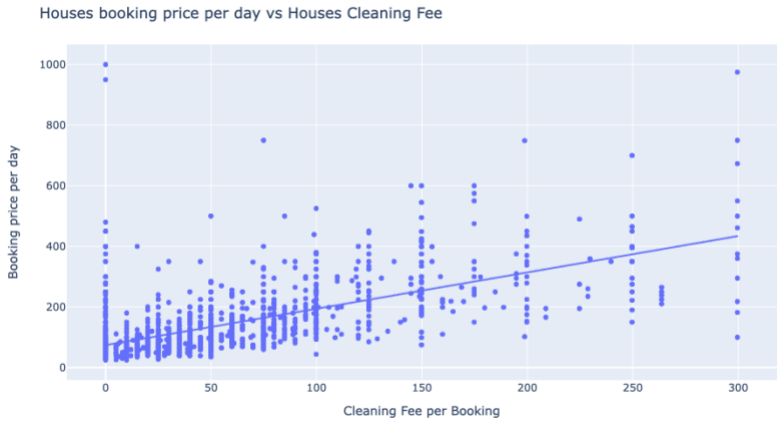

Figure 9 Booking price per day vs House cleaning fee

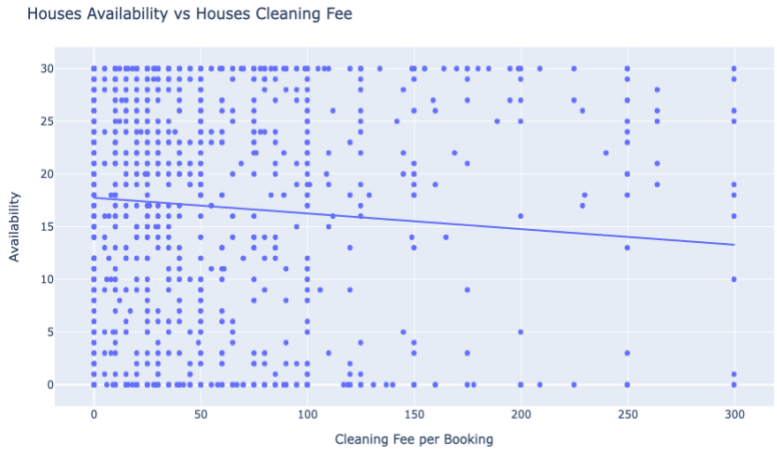

Figure 10 Availability vs House Cleaning fee

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.

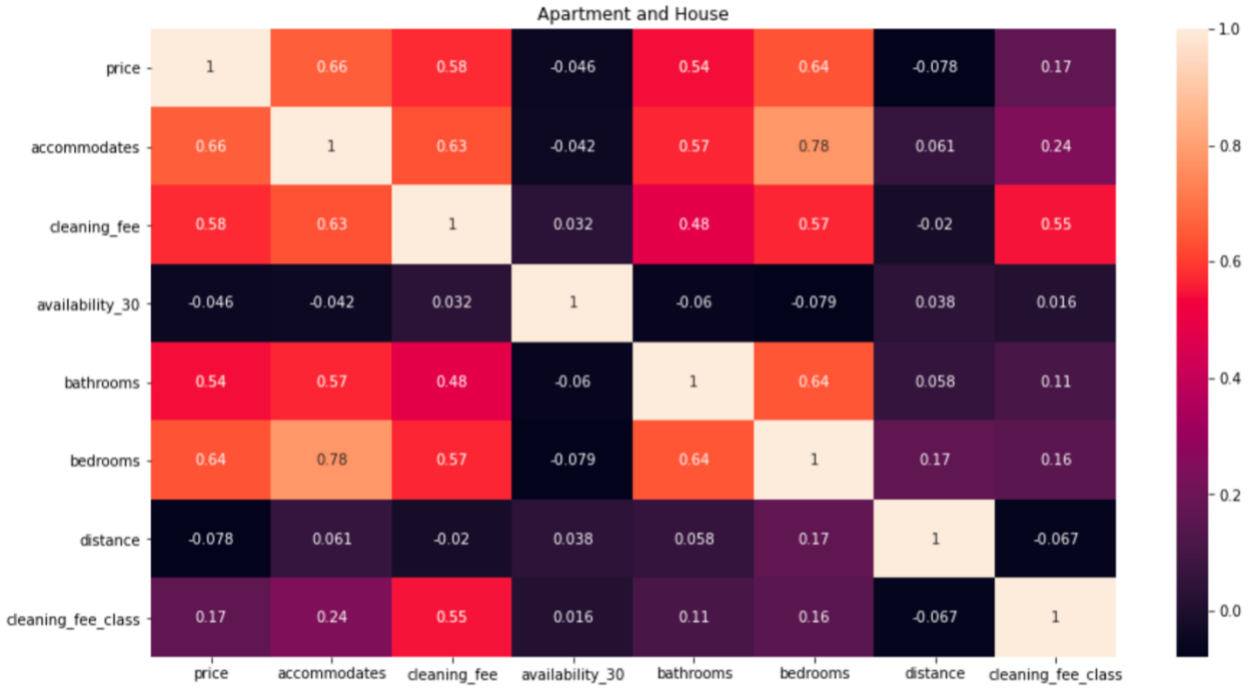

Figure 11 House features Heatmap

It seems the apartments and house data have more relationship with the booking prices. The cleaning fee classification does seem to have a strong correlation be with any other feature. This could mean that the client doesn't consider the cleaning fee to make a booking. I will focus on house and apartment properties because they are the highest requested throughout 2016.

I used four different prediction models Linear regression, Random Forest, Gradient boosting, and XGBoost. The best **base model** is the Gradient boost that scores 66% 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.

Model	Test Score	MAE	RMSE
Base Linear Regression	0.633	34.3829	48.8098
Base Random Forest	0.6412	31.972	48.2641
Base Gradient Boost	0.6615	31.689	46.8783
Base XGB	0.582	34.0866	52.0957

After selecting the most critical features and tuning each model's hyperparameter, the Gradient Boost has the highest accuracy score with 68%. 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.

Tune Model	Test Score	MAE	RMSE
Tune Linear Regression	0.6331	34.3821	48.8098
Tune Random Forest	0.6443	31.6446	47.2283
Tune Gradient Boost	0.6814	31.0009	45.4787
Tune XGB	0.6054	32.9597	49.7412

This model will help the client choose the ideal booking price per day when he decides to buy a unit and put it on the Airbnb market. The model will tell the ideal price depending on where the unit is located, how many bathrooms and bedrooms the units have, how many people will be staying, how many days, and season.