Airbnb Demand and Revenue Analysis

Lulude Sun LULUDESUN@BERKELEY.EDU
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720, USA

Editor: Lulude Sun

Abstract

Airbnb, the emergence of a new online marketplace, transformed the hospitality industry. This new platform provided new investment insights for the company and potential hosts. Airbrnb becomes increasingly popular among the millennials with growing demand for unique listing and unforgettable experience, which generates greater opportunities for more potential hosts. With an increasing demand for airbnb and a robust growth of emerging hosts, the company needs to identify important factors that influence customers' demand and measure the amount of impact each variable (e.g price, location, etc.) make on customers' demand. To create more investment opportunities, attract more hosts, and help existing hosts to maximize revenue, we will analyze revenue spatially and seasonally to find the best places to growth their business and create more unique experiences for customers.

1. Problem Description

One of the highlights that demonstrated Airbnb's business growth is when the city council of Los Angeles voted to legalize home sharing which created a pathway to the long-standing vacation rental market. As one of the first cities that integrated Airbnb culture, Los Angeles essentially sees an increasing demand among customers. Understanding the demand of Airbnb will brought forth more business opportunities and create much more needed places for customers. We are going to analyze the demand of short term renting of Airbnb in Los Angeles to provide more insights for the company and hosts. Airbnb is mostly preferred by travelers, therefore we will only focus on short-term renting and disregard the demand of those who seek long-term renting that only occupies a minute proportion of the population of customers. Besides analyzing demand, it is also important to predict revenue based on several factors including spatial location and neighborhood. We will build models to predict demand and revenue to bring more insights to the Airbnb company.

1.2 Data

We obtained Airbnb datasets from *InsideAirbnb*. Since *InsideAirbnb* archives past datasets periodically, only past year's datasets are made available to the public. We obtained Airbnb listings and reviews of March, June, September, and December, which perfectly represent the four seasons of a year. We will mainly use listings datasets for analyzing the factors that influence demand and building models to predict revenue. The reviews datasets will be used for natural language process (NLP) to give us more insights on how reviews also

affect demand, which ultimately impact revenue. We also obtained housing price data from *Zillow*, which will be mainly used to calculate ROI.

1.3 Questions of Interest and Motivation

Some of the interesting questions that we are seeking to answer are what do hosts need to improve if they want to increase return-on-investment? Does location have an impact on revenue? How does price elasticity take place in measuring demand? What factors influence demand the most and to what extend do they impact the prediction of revenue? Where should hosts invest in Los Angles? Our motivation is to help generate the best investment ideas for existing and future hosts to maximize profit by analyzing the best places to invest. This also help Airbnb company to provide the best traveler experience for customers by having more rentals where demand is the highest. Our project is highly relevant to several major scientific disciplines including statistics, economics, and social sciences.

2. Data Selection

2.1 Origin of Data

For our listing and review datasets, *InsideAirbnb* originally collected them to provide data that quantifies the impact of short-term rentals on housing and residential communities; and also provides a platform to support advocacy for policies to protect our cities from the impacts of short-term rentals. The data has been analyzed, cleansed and aggregated where appropriate to faciliate public discussion.

Besides airbnb listing and review data, we also obtained Los Angeles housing price data from *Zillow*, an American tech real-estate marketplace company. They collected housing prices data for to provide both buyers and sellers more insights on the current market price and the trend. *Zillow* has access to data from local listing services and are able to provide sellers and buyers the earliest and most reliable data on the state of the housing market.

2.2 Data Background

For the airbnb listing datasets there are 74 columns and more than 10,000 rows, meaning there are 74 variables in total and more than 10,000 listings in the year of 2021. We will not be using every variable to build of model as it will probably lead to over-fitting which will increase bias. Some of the most important variables from the listings datasets are neighborhood (cleansed), latitude and longitude, room type, price, availability for future 30/60/90/365 days, accomodations, and number of bedrooms.

2.3 Useful Variables

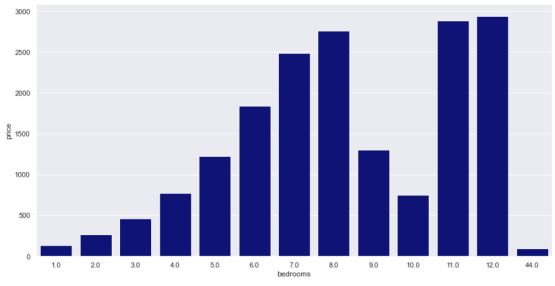
Someo of the most important variables are price and location (longitutde and latitude). With the price variable along with the geographical variables (neighborhoods, latitude and longitude), we can analyze how price changes with respect to spatial location. We are

suspecting that price would be higher in richer neighborhoods and lower in less developed neighborhoods. We also think that demand would be higher in urban neighborhoods than suburban neighborhoods since urban areas would potentially have more entertainment activities that would attract tourism. Neighborhoods where housing price is high would essentially cause airbnb rental price to be high as well. With higher rental price, revenue would potentially be higher for hosts which means more investment opportunities for even more hosts.

Factors like room type, review scores, number of bedrooms, accommodations, and future availability are also important factors that we need to consider when predicting demand. There are four main types of airbnb: private room, hotel room, entire home, and shared room. Clearly, as the rental place gets larger from private rooms to entire home, price also increase significantly. Instead of looking at price and revenue of a single airbnb directly, we are going to examine the price and revenue per person to adjust for size. We hypothesize that with lower occupancy rate, demand would be lower and less revenue.

2.4 Visualization of Data

Figure 1: Average Price vs. Number of Bedrooms



As *Figure 1* shows, when the number of bedrooms increases, price generally also increase. In relation to revenue, hosts might think that having more bedrooms will allow them to set rental price higher which will generate them more revenue. However more bedrooms means larger rental place which leads to higher rental price. This increasing trend might be misleading as it does not adjust for area.

Given that the majority of airbnb listings are of lesser bedrooms, meaning there more smaller airbnbs and majority of airbnbs accepts less guests, the occupancy rate is still consistently high for smaller airbnbs. From *Figure 2*, it is evident that smaller airbnbs on average has an above average occupancy rate than large airbnbs. From *Figure 3*, we see that airbnbs that accepts less guests have higher demand than airbnbs that accommodate more guests. Therefore, even if larger airbnbs tend to have higher prices, the demand is evidently lower than smaller airbnbs. Because we cannot predict revenue only based on rental price, demand/occupancy rate is also an important factor that we need to consider.

As *Figure 4* shows, we adjusted our spatial analysis on annual return per person instead of only annual return. The reason is that annual return will always be higher given your airbnb has a large size and will have a higher rental price, however larger airbnbs also have

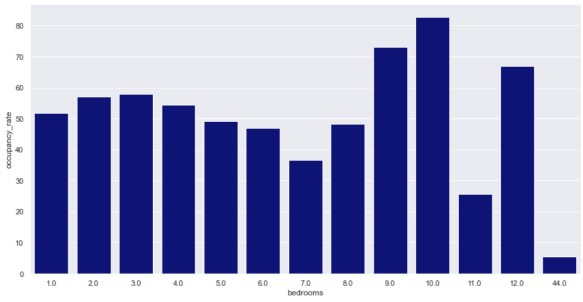


Figure 2: Demand vs. Number of Bedrooms

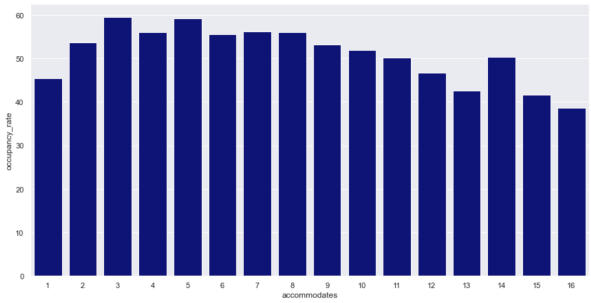
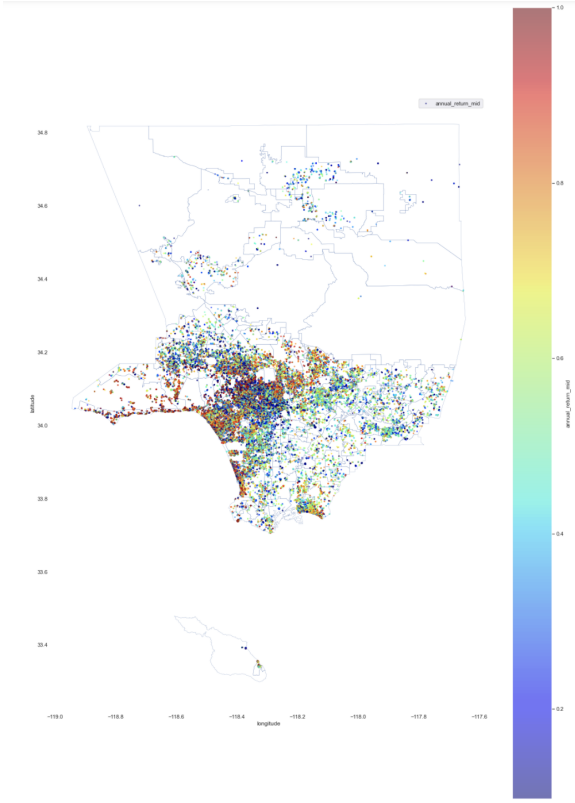


Figure 3: Demand vs. Number of Guests

very high capital cost. It is evident that airbnbs along the coastline have very high annual revenue, probably due to its proximity to the beach. Beverley Hills also have high annual revenue as it is also a tourist attraction. Spillover effect is also evident where airbnbs close to Beverley Hills also have above average annual revenue and gradually decreases as they get further away from Beverley Hills.

Figure 4: Spatial Analysis of Annual Return per Person



3. Methods

3.1 Notations

The goal of our project is to analyze and predict demand as well as revenue. The relevant variables that we will be using are occupancy rate and revenue. We will define demand as customers' willingness to book an airbnb and we will define revenue as the return in investment for the hosts.

Definition 1 (Availability) *Let $A_i(t)$ be a discrete random variable that defines the availability of Airbnb i on date t where $A_i(t) = 1$ means the Airbnb i is occupied/unavailable on date t and $A_i(t) = 0$ means Airbnb i is unoccupied/available on date t.*

Here, we only observe the availability of an airbnb in future date t at the time when we record the data. Therefore the data is subject to change meaning if there exist a customer who will book the airbnb after the recorded date, then $A_i(t)$ will no longer be the same.

Definition 2 (Occupancy Rate) *Let OR_i be a continuous random variable that defines the occupancy rate of Airbnb i where $OR_i = E[A_i(t)]$ and $0 \leq OR_i \leq 1$.*

We define occupancy rate as the expectation of being occupied or not. Let t be the date when we observed the future availability data. $\hat{OR}_i(t) = \frac{1}{30}\sum_{t'=t}^{t+30} A_i(t')$. For example, on 1/1/2022 we observed that there is 12 days available in the month of March, then our occupancy rate for March would be $\frac{12}{30} = 0.4$. Because we only have four data point, each for four seasons. We will adjust the seasonal effect of our occupancy rate variable where

$$OR_i = \frac{\hat{OR}_i(3/2021) + \hat{OR}_i(6/2021) + \hat{OR}_i(9/2021) + \hat{OR}_i(12/2021)}{4}$$

Definition 3 (Return on Investment) *We will define revenue as ROI_i for each host i will get for airbnb i where $ROI_i = \frac{OR_i \times Price}{C_i}$ where C_i is the capital cost which is the price the hosts paid to buy the property.*

As mentioned before, predicting demand and revenue solely based on price will incur an overestimation of the ROI. Therefore by using our defined ROI_i formula, we will adjust for price and demand when predicting revenue.

3.2 Statistical Modeling

We are implementing and comparing XGBoost, Random Forest, ridge regression, and support vector regression models for predicting demand and estimating revenue. Before we built our model, we used MCA (multiple correspondonce analysis) to perform dimension reduction on variables such as amenities and room type to reduce multi-collinearity and over-fitting issues. Some of the advantage of XGBoost models are that it uses a earning method that combines a set of weak learners into a strong learner to minimize training errors, which is a primary advantage over gradient boosting machines. At each step, it will corrects the previous steps and errors of the previous models are correct by the next predictor. It is also commonly used in industry for its strong predictive power and accuracy.

The trade-offs of XGBoost's high accuracy is that it does not perform as well when data is not well-balanced and are very sensitive to outfliers, which we take into account when cleaning data. Some of the advantage of random forest models are the efficiency when handling large datasets and it provides a high level of accuracy when predicting outcomes over decision tree models. It can perform both regression and classification tasks which is useful for our dataset. With its ensemble learning technique and bagging algorithm, it combines the output of all trees and reduces the variance. In comparison to XGBoost, random forest is very robust to outliers. The main disadvantage of random forest is the time complexity given it uses multiple trees in the algorithm. Some of the advantange of ridge regression is that it is able to fix over-fitting as it will recognize important features, unlike regular least squared regression. It is also very useful when there are multi-colinearity between variables which does exist in our data (e.g price and room type). The disadvantage for ridge regression model is that it uses all features in the dataset to improve test accuracy. Given the large amount of variables in our dataset, it is not very efficient to use every single feature. It also reduces the coefficients θ to very low values if the feature is not important but will not make them zero therefore it will still consider every feature in the model. Some advantage of support vector regression model are its robustness to outliers and its excellent generalization capability with high prediction accuracy. Some of the drawbacks are its lack of ability to perform well when the datasets has more noise and for large datasets.

3.3 Model Comparison

We performed 10-fold cross validation to select hyperparameter for our training set and build models to predict Return on Investment. We then used our model on the testing test to evaluate the performance of each model and select the more accurate model. *Figure 5* shows the distribution of the residuals of the gradient boost model and *Figure 6* shows the distribution of the residuals of the ridge regression model. Clearly, both residual distribution is normal. It is evident that the standard deviation of the gradient boost model residual is less than the ridge regression model residual distribution. Therefore, we can say that there is less variance in the estimates of ROI for the gradient boost model than the ridge regression model and we can conclude that the XGBoost model is the better model.

We also evaluated each model's performance on the test set by calculating the MSE. *Table 1* allows us to see a clear comparison of the error of the four models. Evidently, XGBoost has the lowest MSE and Ridge regression has the highest MSE. Again, we can conclude that the XGBoost the best algorithm to use to predict ROI.

Algorithm	Mean Squared Error
XGBoost	0.03666
Random Forest	0.03791
Ridge Regression	0.04357
Support Vector Regression	0.03983

Table 1: Mean Squared Error for Models on Test Set

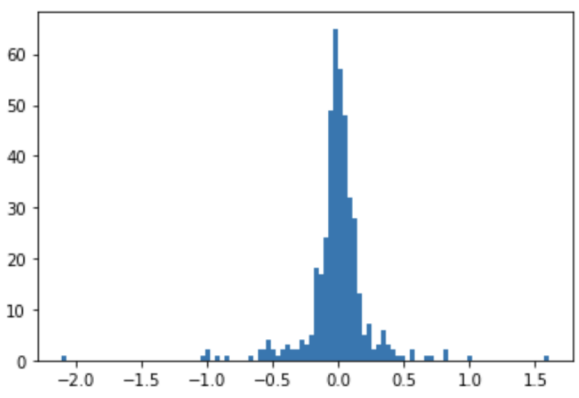


Figure 5: Residual Distribution of Gradient Boost Model

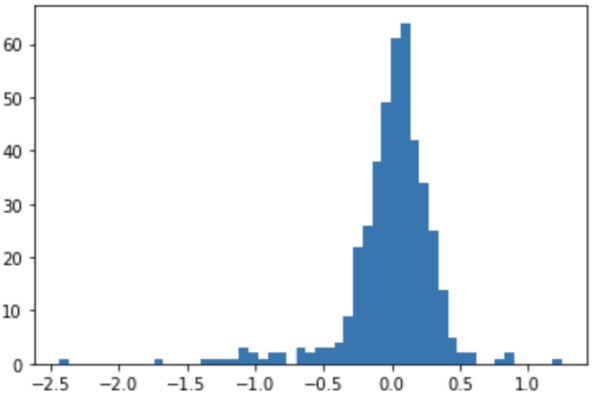
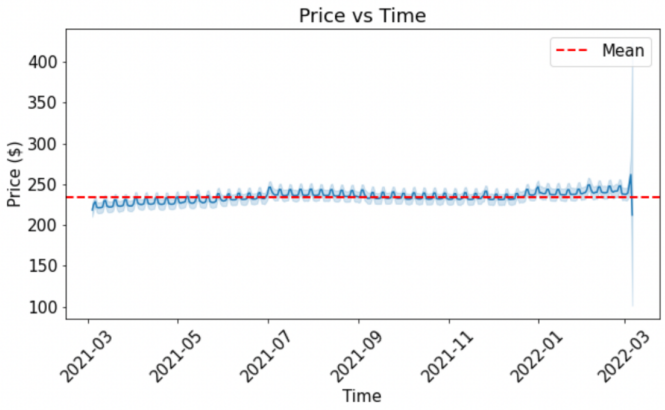


Figure 6: Residual Distribution of Ridge Regression Model

3.4 Assumption Justification

Our model makes the assumption that price will be fixed and to justify our assumption, we performed time series analysis on price. As *Figure 7* shows, price remains consistent and steady over time; meaning that Airbnb hosts tend to keep their price constant over time. Therefore, we concluded that it is reasonable to assume price is fixed for our model.

Figure 7: Time Series Analysis of Price



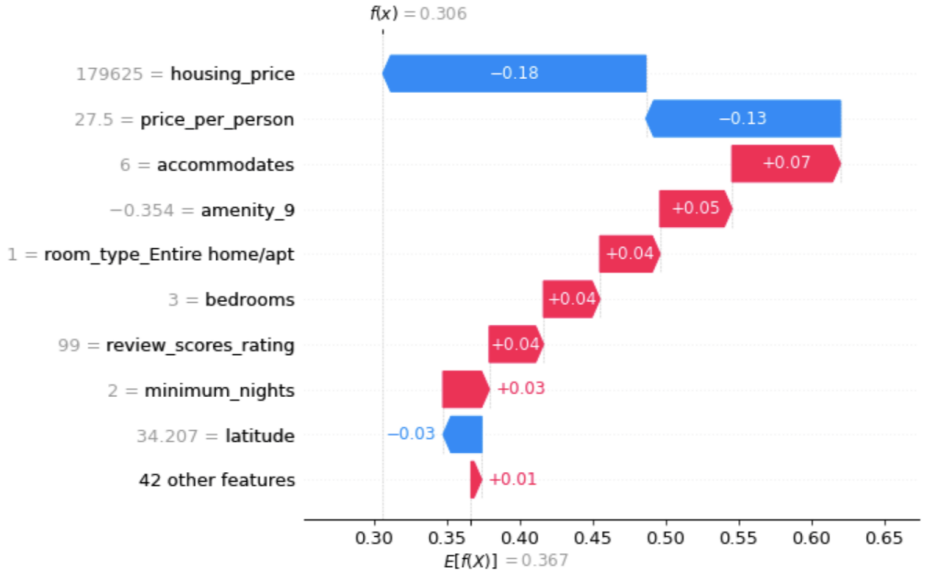
4. Results

4.1 Key Features

We used SHAP to determine the feature importance for each variable. As *Figure 8* shows, housing price, Airbnb price, number of accommodates, amenities, room type, and review scores are some of the top important features for our model. Both housing price and price

per person have a negative impact on revenue. When housing price is high meaning Airbnb hosts are paying a high capital cost (house sale price), revenue would decrease. According to our ROI formula (*Definition 3*) capital cost is in the denominator therefore as capital cost increase, ROI will decrease. Hosts will need more years to break even compared to hosts who paid a lower capital cost. The number of accomodates appear to have a positive impact on revenue. It can be explained as travelers besides businessmen are unlikely to travel alone therefore Airbnbs that accommodates for greater number of guests are more likely to have higher demand thus higher revenue.

Figure 8: Impact of Key Features



Price have an negative effect on ROI because Airbnbs with a higher renting price would lead to lower demand (occupancy rate) which would ultimately lead to lower revenue. As *Figure 9* shows, there is clearly a decreasing trend for ROI as price increases. This can be explained by the fact that people look for the most cost-effective Airbnb, meaning travelers are unlikely to pay for a very high price for an Airbnb. As price directly relates to demand, it would be unreasonable for hosts to place their price too high. In contrast, review scores have a positive impact on ROI because people rely on reviews to make decisions. As *Figure 10* shows, when the number of reviews increase and review scores improve, we see an increasing trend for ROI. Review scores are affected a number of factors (host respond time and rate). We can conclude that that hosts should focus on bringing a positive and enjoyable experience for their guests in order to improve review scores and ultimately increase revenue.

As *figure 11* shows, revenue has an increasing trend as the number of bedrooms approaches 3 and starts to decrease. The distribution of bedrooms are right skewed given the lack of datapoints on Airbnbs with more than 6 bedrooms and it is very rare to Airbnbs with more than 6 bedrooms. Thus we will consider the increasing trend after 6 bedrooms as implausible and will not draw conclusions based on that. Clearly, revenue is maximized when there are 3 bedrooms and we would suggest hosts to keep their rental place at 3 bedrooms and no more than 5 bedrooms.

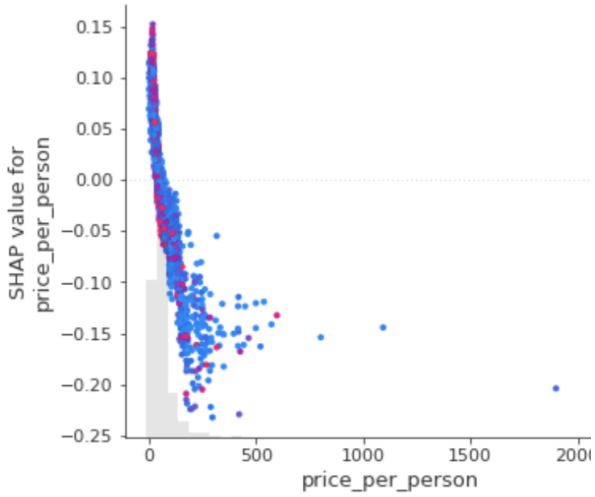


Figure 9: Effect of Price on ROI

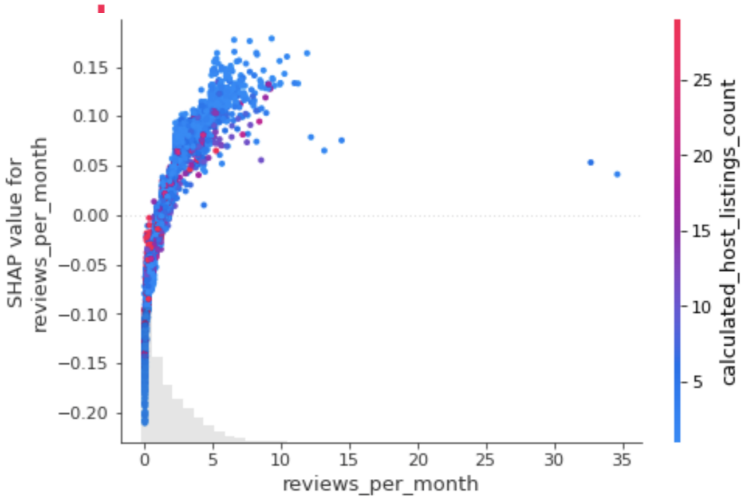
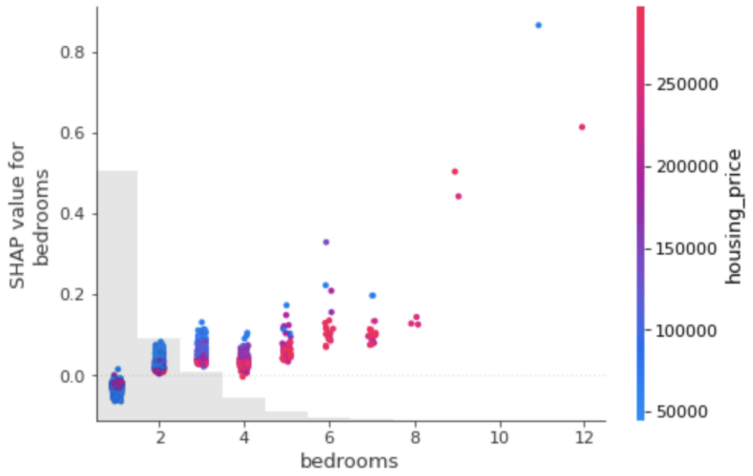


Figure 10: Effect of Reviews on ROI

Figure 11: Effect of the number of Bedrooms on ROI

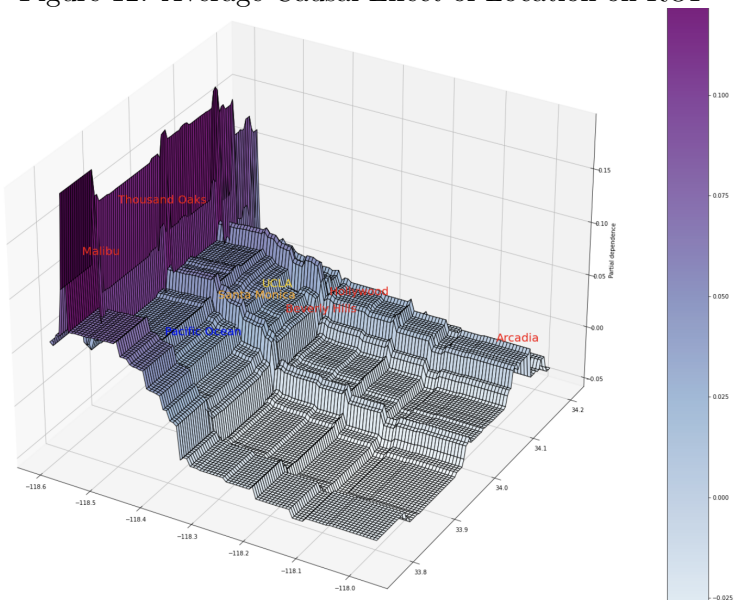


4.2 Average Location Effect on Revenue

To test the effect of location on revenue, we used partial dependence plot. By keeping all the other variables (review scores, price, etc.) fixed, we can draw causal inference on the impact of location on revenue. *Figure 12* shows the partial dependence plot with longitude on the x-axis, latitude on the y-axis, and the location impact score on the z-axis. As expected, central LA consisting of UCLA, Beverly Hills, and Hollywood that attracts the most tourists in LA County have higher ROI than East LA where there is less tourism hot spots. It is also evident that areas along the coastlines such as Malibu and Santa Monica also have greater ROI with Malibu having the highest ROI in LA County. Note Thousands

Oaks and Pacific Ocean is also on our plot due to the nature of the plot and information is from extrapolation. We will not consider those areas also due to the lack of datapoint and their likeliness of being outliers. In general, we can conclude that West LA generates better investment opportunity and returns a higher revenue for hosts than East LA. This is essentially due to the laid-off nature and well-known entertainment industry of West LA where it also consists of more than half of the best neighborhoods in LA.

Figure 12: Average Causal Effect of Location on ROI



5. Conclusion

In conclusion, our prediction model can help Airbnb hosts to find new investment opportunity, maximize revenue, and forecast their return on investment. We can conclude that in order for hosts to increase return on investment, they should focus on providing customers a better experience by improving key features including amenities, response rate, and response time which affect review scores directly. They should also rent out room types that match more closely with customer's needs which would be having more entire-home Airbnbs with ideally three bedrooms. Based on our price elasticity model, we can conclude that as price increases, demand generally decreases which ultimately leads to a decrease in revenue. Therefore we would recommend hosts not to set prices too high and use our price elasticity model to determine the optimal price for the goal of maximizing ROI. For existing Airbnb hosts, our model is beneficial to set price, predict their occupancy rate and forecast their ROI. For future Airbnb hosts who are looking to enter the market, we would recommend them to purchase entire home with three bedrooms and ideally with a lower than average housing price on the market. Spatially, we would suggest new host to invest in West LA around central LA (Bevelly Hills and Hollywood area) and along the coastline (Santa Monica), ideally in Malibu. Our model would be beneficial as it can help hosts to predict demand and forecast ROI in places where they are looking forward to invest.