Big Data & Artificial Intelligence Project

Project Title: Explaining High Bookings of Airbnb's Through the Lense of Price vs. Quality Tradeoff

Market: Los Angeles

“We, the undersigned, certify that the report submitted is our own original work; all authors participated in the work in a substantive way; all authors have seen and approved the report as submitted; the text, images, illustrations, and other items included in the manuscript do not carry infringement/plagiarism issue upon any existing copyrighted materials.”

Name	Signature
Adam Messina	Adam Messina
Sathish Komire	Sathish Komire
Naga Raju Nannapaneni	Naga Raju Nannapaneni
Siddharth Maredu	Siddharth Maredu
Sai Venkat Battula	Sai Venkat Battula

Table of Contents

Executive Summary

In our analysis, we have taken it upon ourselves to understand the Los Angeles market for Airbnb's to aid in insight to fuel our client's investment decisions. Our main findings surround what makes Airbnb's experience high bookings, which is a great indicator of our investor's profit-making potential. Our first approach was to create a general predictive model to assign Airbnb's as high booking or not for all markets. From this model we were able to deduce the important variables which guided our research into diving deeper into the Los Angeles market.

While exploring this market, we have found useful insights on pricing (including additional fees), size (the number of people an Airbnb can accommodate), as well as quality of service and management such as host response rate. We wanted to dive deeper into these topics to guide our investor on specific aspects to help them be as successful as possible while doing business with the portfolio of Airbnb's in the Los Angeles market. Our insights are important, because we can inform them on how to price their location, how large of a location they should book, and how attentive they must be in order to increase their probabilities of achieving high bookings.

Subsections / Main Focus and Questions

To explain what are the factors that cause high bookings. We have considered 3 factors that have been formulated into 2 sub questions.

- *Based on the number of people the Airbnb accommodates, how does price related factors such as price explain high booking in LA market?*
- *Based on the number of people the Airbnb accommodates, how does factors related to management/hospitality influence high booking in the LA market?*

We have chosen these factors not only based on our statistical tests, but on domain knowledge itself because in any Airbnb market, the consumer would be looking for the price, locality, number of people it accommodates and how well the host responds to communication.

Methodology

Note: these are the selected variables in our final model.

selected_columns <- c("extra_people", "host_response_rate", "accommodates", "price","requires_license","reviews_per_month", "high_booking", "zipcode")

While building our predictive model, we have put a strong emphasis on exploratory data analysis and data cleaning to create the most accurate model possible. Most of our time has been spent on data cleaning as the Airbnb dataset has a lot of missing values, handling duplicates and a combination of several different data types. At first, our approach was to determine how we can make the most out of the numerical data types. Some of these columns contained special characters which we needed to clean up before converting to numbers, such as dollar and percentage symbols. We did basic imputation such as mean and medium imputation to fill in the NA values in the data. After our group did our own data cleaning individually, we reconnected and made one dedicated data cleaning file comprising the basic transformations that have been done before we did our own experimentations. As far as modeling goes, we have focused our efforts on ensemble models (gradient boosted decision trees) as they are known to be the best for structured data problems.

With the progress and insights that we gained from building the predictive model, we took inspiration from the most important variables to help us further understand the causal relationships in the Los Angeles market.

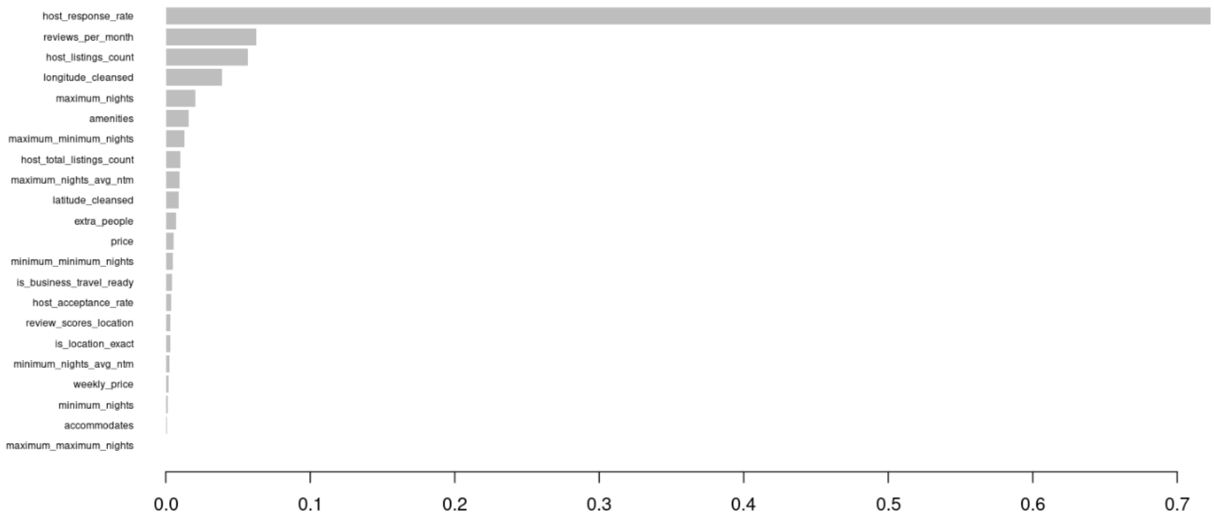

Fig 1

The predictive model gave us these outputs for importance scores of given features, see Fig 1. Based on this, we have decided to investigate further into these features based on domain knowledge in terms of answering our research questions. We re-ran this model, however this time filtered specific to the Los Angeles market. see Fig 2.

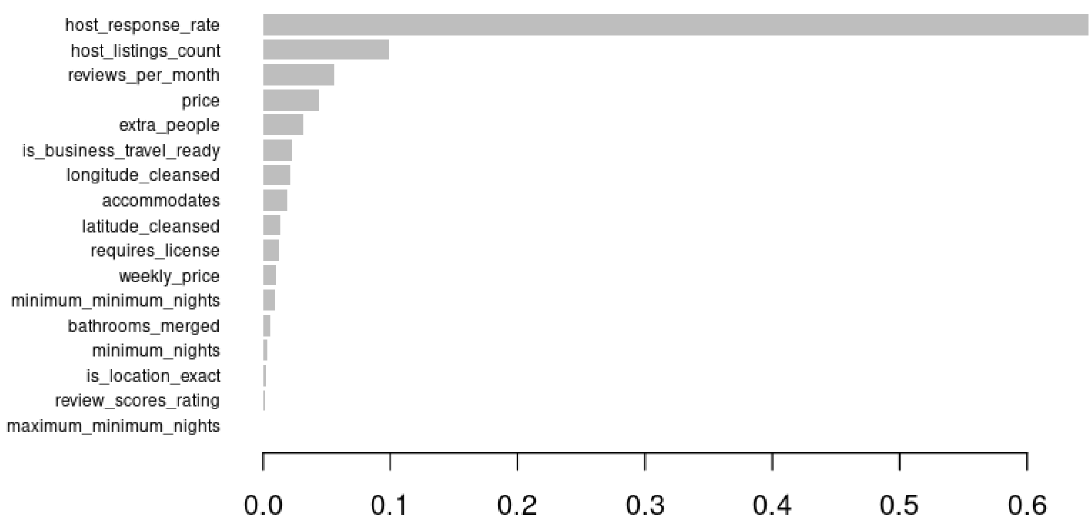

Fig 2

Now that we had the most important variables specific to the Los Angeles market, we wanted to place some of these top variables into a graph to show the causal relationships to help us guide our research. We did some background research and found that there are laws that prevented Airbnb hosts from having more than one listing, which informed us to disregard host_listings_count for our analysis. Some of the variables listed as most significant cannot be caused by anything, such as host_response_time, host_response_rate, price, and amenities, see Fig 3 for reference. To adjust for this when creating our model, we declared these variables as exogenous variables to ensure they were at the top of our causal graph. Furthermore, the main purpose of this analysis was to determine high booking, which resulted in us removing any features in our causal model that did not show causations for high booking. After many iterations we were able to achieve the following causal graph.

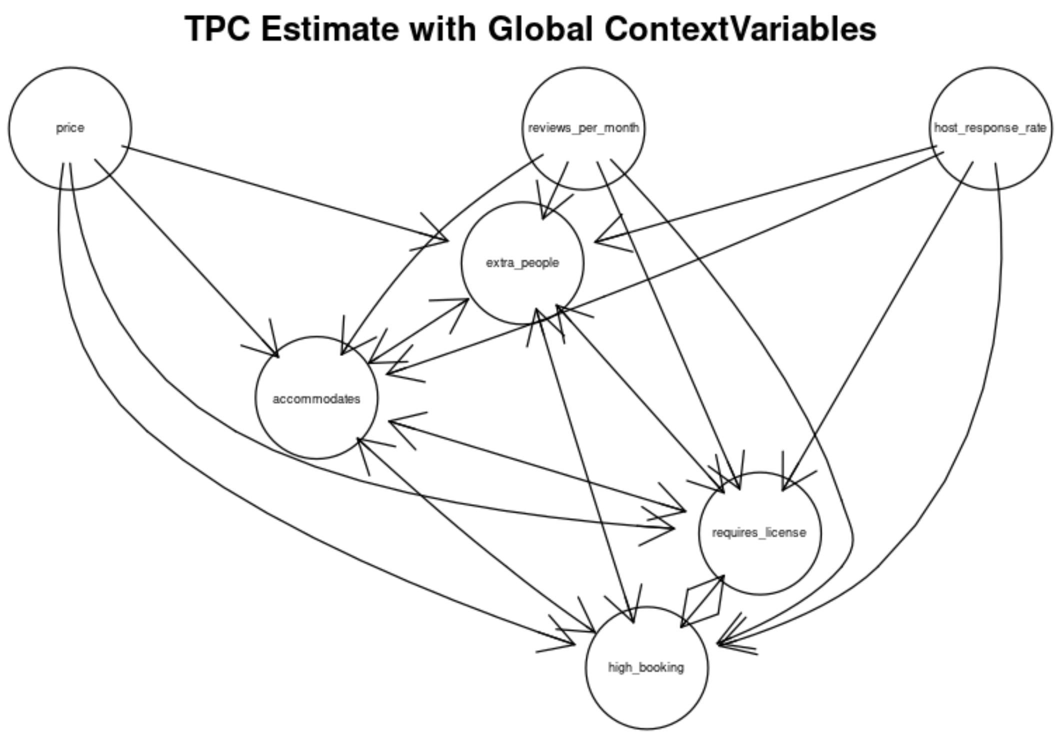

Fig 3

With this graph we were able to form and determine how to test our hypotheses, because we know what important variables causal relationships with high booking have, which helps us know what we need to control for.

Results and Findings

Fixed Effects of Accommodation Capacity, Price, and Host Response Rate on High Booking Rates: Controlling for Zip code

This study's main objective was to investigate the relationship between the number of accommodations on high booking of Airbnb. To be specific, we aimed to understand how different ranges of accommodation capacity affect the likelihood of achieving high booking. This analysis was conducted on Airbnb listings by segmenting the data based on accommodation capacity by controlling for zip code to see the fixed effects.
Prior to the analyses, we have visualized the distribution of number of people the Airbnb's accommodates, see Fig 4. We can see that most Airbnb's have fewer people to accommodate. This provides us with an insight into the type of Airbnb's the customers are looking for.

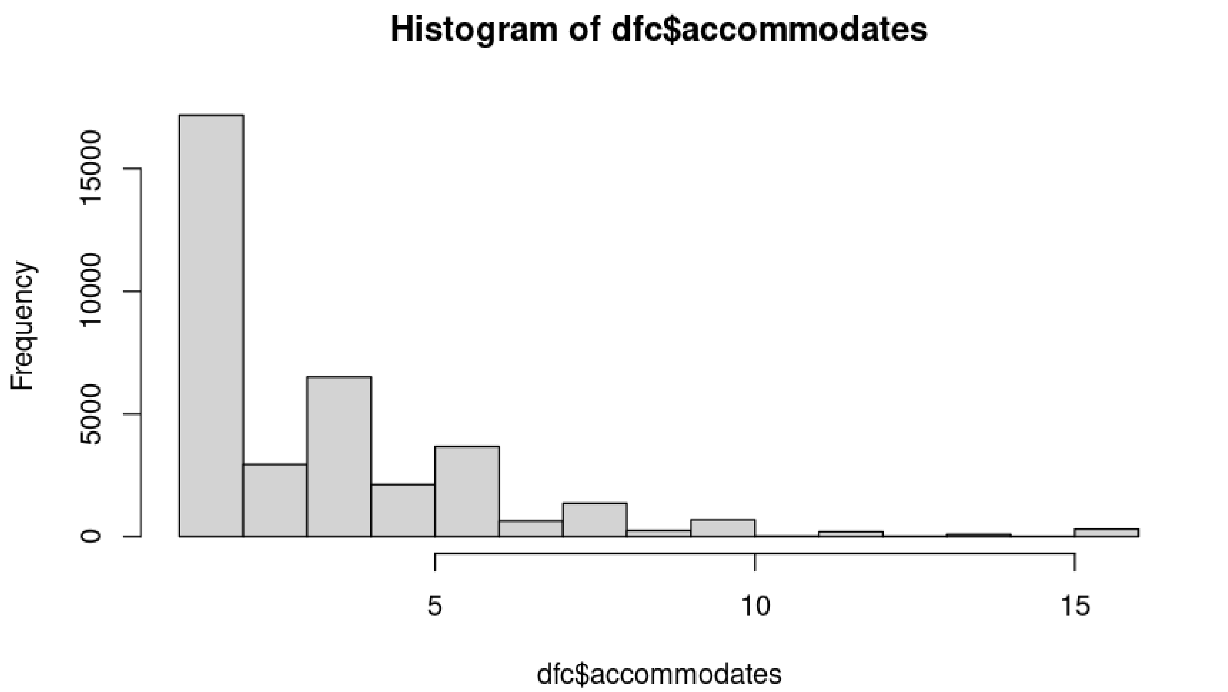

Fig 4

Controlling for zip code is crucial in this analysis to account for geographical variability. This control helps us to see isolated effect of accommodates on high booking without being influenced by zip code.

Data Preparation
Our data frame was divided into quartiles based on the number of guests the listings can accommodate. Four distinct categories were created:
1st Quartile (Lowest 25%): accommodates at most 2 people
2nd Quartile (25% to 50%): accommodates at most 3 people
3rd Quartile (50% to 75%): accommodates at most 5 people
4th Quartile (Top 25%): accommodates at most 16 people

Accommodates

Listings with the lowest accommodation capacity showed significance with negative effect on high booking rates when controlled for zip code. This suggests that smaller listings in their respective areas are more likely to achieve high booking.

From the chart, see Fig 5 and these quantiles, we can see that we want to invest in Q1 (accommodates for 1-2 people) and Q3 (accommodates for 4-5 people) to garner high booking because from the statistical results we can see that as we increase the number of people we accommodate we are less likely to get high booking.

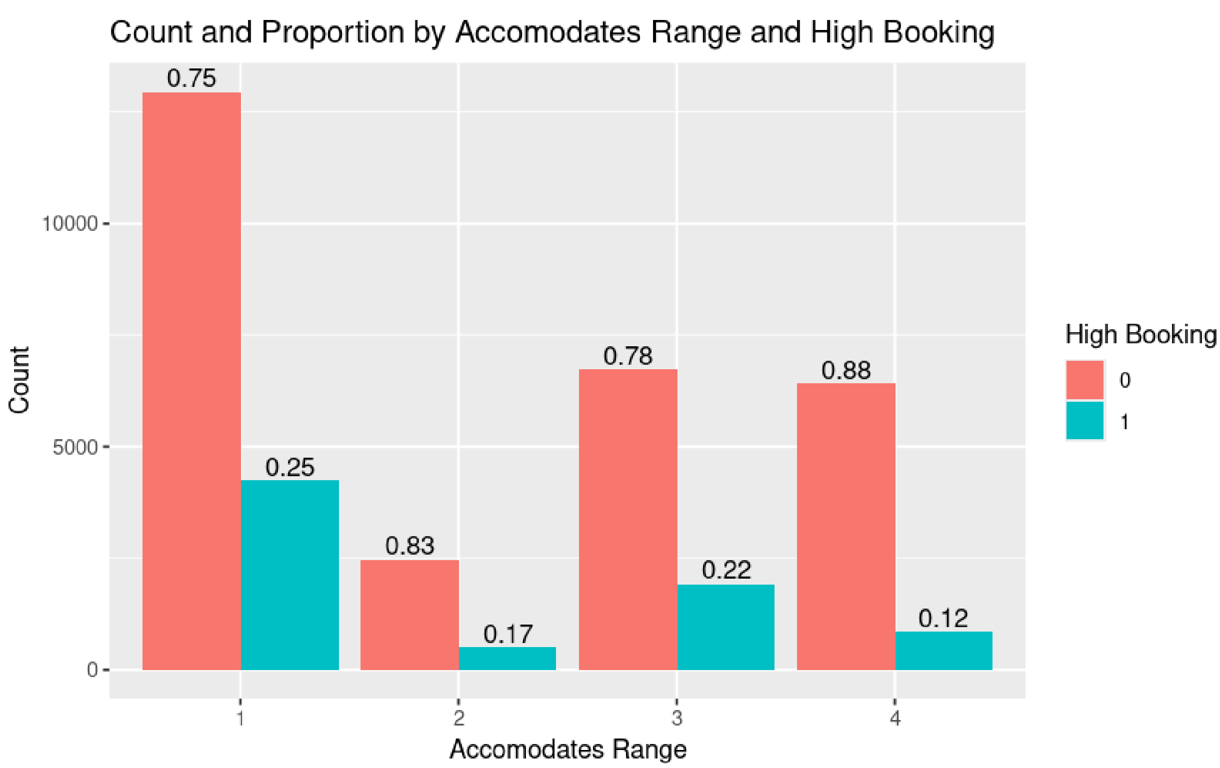

Fig 5

Fig 6 below shows us the distribution of the prices of Airbnb's that accommodates at most 2 people. We can see that they are majorly priced around 50-100 dollars. Since this is the significant section of Airbnb's that attracts high booking. The investor should consider investing in Airbnb’s that accommodate 1-2 people and try not to charge too much.

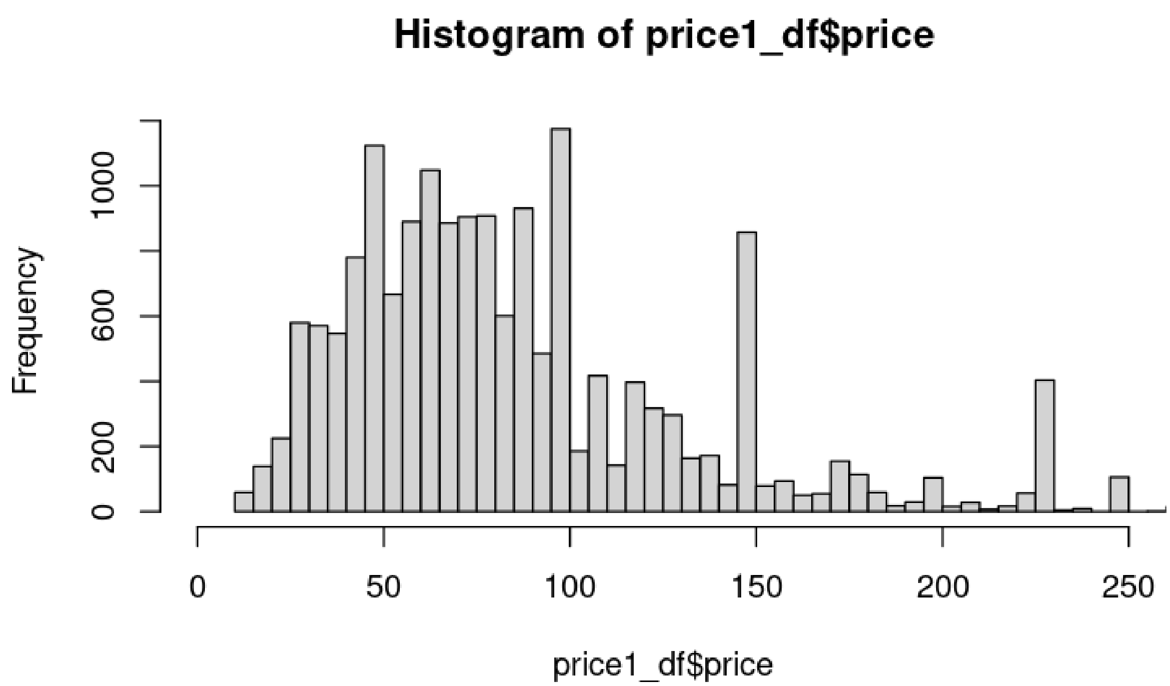

Fig 6

Price

To further strengthen our analysis, we have formulated a count and proportion of price range by high booking, see Fig 7. From the graph and the quantiles that is listed below, make sure not to charge over 110 dollars for Airbnb’s that accommodates 1-2 people.

```
```{r}
quantile(price1_df$price)
```

  0% 25% 50% 75% 100%
  10   55   78 110 5000
```

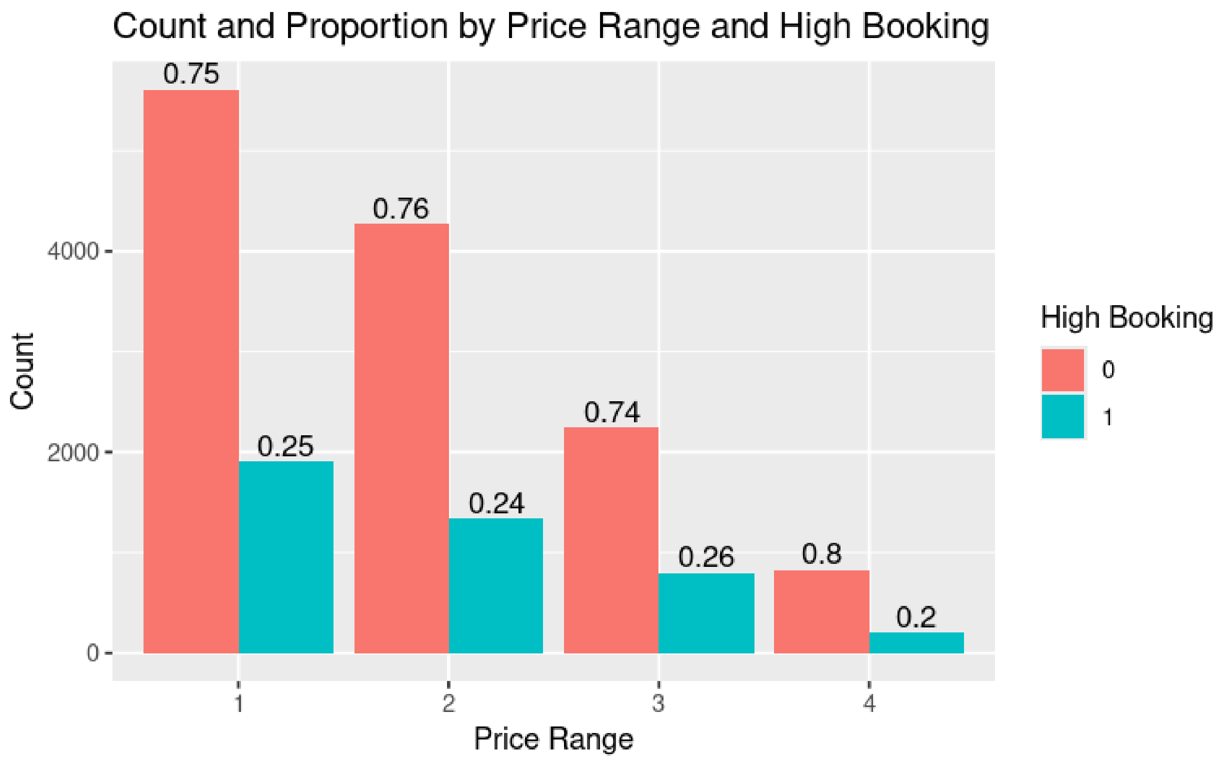

Fig 7

Host Response Rate

When we calculated the effect on high booking by host response rate, we found that it is almost a significant factor that an investor should consider before investing into Airbnb's in LA (based on our statistical results). Our analyses showed that it's positive, meaning as the host response rate increases the probability that the listing to have high bookings also increases.

If we see Fig 8. The plot describes count and proportion by price range and high booking. Above 75% response rate improves high booking for Airbnb’s that accommodates at most 2 people.

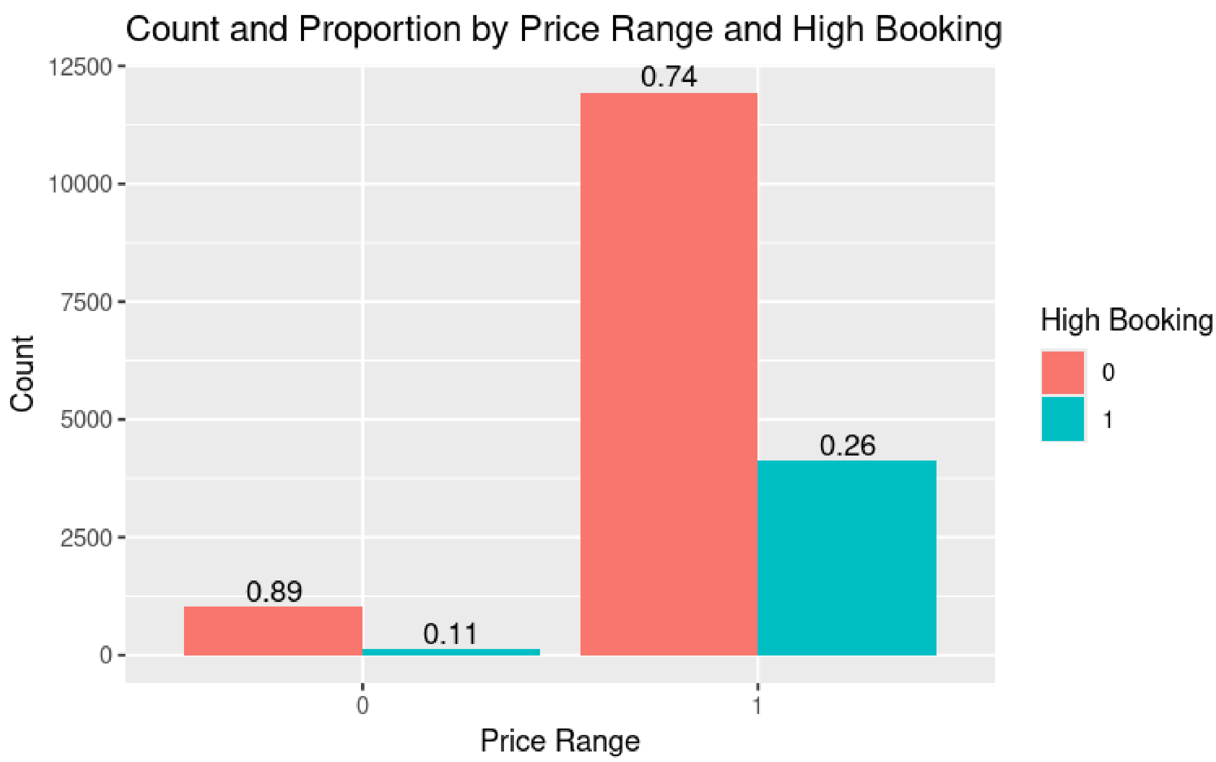

Fig 8

We should also take into consideration that for the airbnb's that accommodate at most 2 people, the host response rate should be at least 94%. It means that for Airbnb's in LA, the host should be proactive to customer requests and be able to give prompt responses to attract high booking. This can be explained by looking at Fig 9 where we can see, starting from 1st quartile itself the host response rate is more than 94%.

````
```{r}
quantile(host1_df$host_response_rate)
```

       0%       25%       50%       75%      100% 
0.0000000 0.9463407 1.0000000 1.0000000 1.0000000 
````

Fig 9

Conclusion and Discussion

Performance Evaluation

The models effectively supported our objectives by highlighting the varying impacts of accommodation capacity on high booking rates across different capacity ranges. The significant findings in the 1st and 3rd quartiles reveal crucial insights about the Airbnb market:

Positive Correlation in 1st Quartile: Smaller listings tend to achieve higher booking rates, which could be due to their affordability and suitability for solo travelers or couples.

Insignificance in 2nd and 3rd Quartiles: Mid-range listings may need to focus on other competitive factors, such as amenities or pricing strategies, to improve booking rates.

Negative Correlation in 4th Quartile: Larger listings might face challenges in maintaining high booking rates, potentially due to higher costs or niche demand

Recommendations

Invest in Smaller Listings: Focus on acquiring or optimizing smaller listings (1st quartile) as they have a higher likelihood of achieving high booking rates. These properties are often more affordable and cater to a broader market.

Optimize Mid-Range Listings: For listings in the 2nd and 3rd quartiles, consider enhancing competitive features such as unique amenities, excellent customer service, and competitive pricing. Since these listings did not show a significant impact from accommodation capacity alone, other factors can be leveraged to boost their attractiveness.

Strategic Pricing for Large Listings: For the largest listings (4th quartile), consider strategic pricing and targeted marketing towards groups or events that require large accommodations. Highlighting unique features and offering competitive packages may help improve booking rates. However, this cannot be a feasible solution because of the recent laws that have been passed which have been abided by Airbnb management that says, parties cannot be hosted in Airbnb's due to shooting incident that has happened in 2020.

So, our advice to the investor is, they should not bother investing in Airbnb's that accommodate more people.

References

Ozer, G. T., Greenwood, B. N., & Gopal, A. (2024). Noisebnb: An Empirical Analysis of Home-Sharing Platforms and Residential Noise Complaints. *Information Systems Research*. https://doi.org/10.1287/isre.2022.0070

Los Angeles Airbnb Regulation and Statistics

https://www.hostyapp.com/airbnb-statistics-laws/los-angeles

ChatGPT for technical assistance and interpretation