Big Data & Artificial Intelligence: DS 810.01

Professor Gorkem Turgut (G.T.)

Project title: Decoding Airbnb High Booking Rates

Market assigned to the team: San Diego

“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 any infringement/plagiarism issue upon any existing copyrighted materials.

|  | **Names of the signed team members** |
| --- | --- |
| Contact member | Ben Stafford |
| Team member 2 | Sree Vikas Sabbani |
| Team member 3 | Agastya Valambatla |
| Team member 4 | Samhitha Reddy Annem |
| Team member 5 | Srinivas Yedla |

**Table of Contents**

[**I. Executive Summary**](#_quwewak00ccq) **3**

[**II. Subsections / Main Focus**](#_biroqnzf7ogn) **4**

[**III. Methodology**](#_yp82az5mools) **6**

[**IV. Results and Findings**](#_lazruemg43j) **7**

**V**[**. Conclusion and Discussion**](#_dln476v3ae7h) **9**

**VI**[**. References**](#_illce1qw16hb) **10**

**VII**[**. Appendix**](#_z5tsyhshmiv7) **10**

**Executive Summary**

Our goal was to review moderately priced Airbnbs in order to tailor our research to be targeted towards the majority of “regular” consumers. We wanted to be able to act as a consultant of sorts, that current Airbnb hosts can hire to advise them on how to increase their booking rates. Based on the assumption that our investor wants to maximize profits which can be done by maximizing the probability of high booking rates, the decision was made that it was best to target the largest group of consumers that are in a middle budget tier. Through initial analysis of the data, we quickly realized that factors that you would intuitively think have a positive relationship on booking rates such as being a superhost, were showing the opposite when looking at the data as a whole. Considering what could be causing this anomaly, we thought more about the Airbnb domain. You will always have the upper end listings that are thousands of dollars per night which come with their own target customers and a much different set of criteria for attracting those customers. This can also be said for the low end listings that may be much lower in price but attractive to a consumer that may be in between houses and looking for a long term/affordable stay, or looking for the most inexpensive lodging option in a city without much care for anything else. However, what we felt was the most populated area of the Airbnb market, and therefore the potentially most lucrative to exploit, is the middle. This is the area where people may be visiting a city for a vacation, a wedding, family reunion, ect, and are in a position where they want to stay at a property that they can enjoy, but also need it to be relatively budget friendly.

We found this to be the sweet spot in our analysis. Overall, prices in our market ranged from $19 - $24,999 per night but 75% of all listings fell between $55 - $407 ([Figure 1](#_q5btr2qd42nz)) per night. As we have seen, you cannot treat vastly different priced listings as equals and compare the influence of variables because price plays such a key role in the customers decision. Therefore, by segmenting the data into this reasonably priced section, we are able to analyze a subset of the data that is all relatively similar in terms of the price variable, while still maintaining 75% of the total listings.

After analyzing this subset of data, we thought about what types of variables to focus on to achieve our goal.Variables such as bedrooms, bathrooms, square feet, and many other variables in this dataset are fixed in the short term. Therefore, with our goal being to advise current hosts, it would not do them any good to say “add an extra bathroom and another 1,000 square feet to your home”, this is not an attainable short term goal, not to mention very costly. We aimed to provide quick and easy suggestions to hosts that could help them drive up their booking rates and with this in mind we focused our efforts on analyzing variables that can be easily manipulated. This included variables that can be easily adjusted on the Airbnb platform such as, cancellation policy, minimum nights stay, and instantly bookable, as well as minor changes to the hosts home such as purchasing a barbeque grill, adding an outdoor fireplace, or turning an already existing room into a “workspace”. These material changes do not need to be expensive purchases, but adding these amenities to their home allows for them to include them in their listing description and that can ultimately make a consumer choose their listing over another one, especially in the market of San Diego.

**Subsections / Main Focus and Questions**

Question 1: Why can we not analyze each listing equally?

In the context of optimizing Airbnb listings in San Diego, it's critical to recognize that not all listings can be analyzed or treated equally due to several key factors. These factors include varying property characteristics, host attributes, and market dynamics, all of which can significantly influence booking rates.

Property Characteristics:

Different properties offer a wide range of features and amenities. For example, a beachfront property with luxury amenities such as a pool or hot tub will appeal to a different segment of travelers compared to a budget-friendly apartment located inland. Listings that emphasize proximity to popular attractions like beaches, parks, or downtown areas will naturally have different booking patterns and demand compared to those located in less central areas.

Host Attributes:

The status of the host also plays a significant role. Superhosts, for instance, are likely to have higher booking rates due to their proven track record of providing excellent guest experiences. This status not only enhances trust but can also justify higher pricing. The influence of the superhost status on booking rates is complex and requires careful analysis, considering it may intersect with other factors like price and location.

Market Dynamics:

San Diego's market dynamics, influenced by seasonal trends, local events, and economic factors, create variability in booking rates. For example, properties may see increased demand during peak tourist seasons or major local events, which would not be as pronounced during off-peak times. Additionally, external economic conditions can affect travel behavior and, consequently, the booking rates of different property types.

Therefore, it is essential to segment listings and analyze them within appropriate categories, considering these multifaceted influences. This approach ensures a more accurate and actionable understanding of the factors driving high booking rates and allows for tailored strategies to optimize each segment's performance.

Question 2: What variables in this dataset can be adjusted in the short term?  
  
The project we worked on is much related to this question: How small changes can improve the booking rate.

These variables can be modified without significant structural changes or long-term commitments. Here are the key short-term adjustable variables:

Amenities: Adding or highlighting specific amenities such as Wi-Fi, workspace, BBQ, grill, patio, pool, and fireplace can make the listing more attractive.

BBQ: Presence of BBQ amenity increases high booking odds by approximately 76%.

Workspace: Presence of Workspace amenity increases high booking odds by approximately 46%.

Host Response Time: Improving the speed at which hosts respond to inquiries can enhance guest experience and increase booking likelihood. Quick responses can be achieved by setting up alerts or using automated response tools.

Host Response Rate: Ensuring a high response rate by responding to all inquiries and messages promptly can improve the host's standing on the platform and attract more bookings.

Host is Superhost: While becoming a Superhost involves meeting certain criteria over time, hosts can focus on improving guest experience and collecting positive reviews to work towards this status.

Exact Location: Providing accurate and detailed location information can help guests make informed decisions, increasing the chances of bookings.

Cleanliness: Improving cleanliness scores can be achieved by ensuring thorough cleaning before each guest's arrival and considering professional cleaning services if necessary.

Cancellation Policy: Adjusting the cancellation policy to be more flexible can attract guests who prefer the option to cancel without significant penalties.

Question 3: What amenities are of high importance in San Diego?

In the competitive Airbnb market of San Diego, certain amenities are particularly valued by guests and can significantly influence booking rates. Given San Diego's unique climate and tourist appeal, the following amenities stand out as highly important:

1. BBQ and Grill: Outdoor cooking facilities such as BBQs and grills are popular amenities in San Diego. These features allow guests to enjoy outdoor dining experiences and make the most of the city's excellent weather. Properties with these amenities can attract groups and families looking for social and recreational activities.

2. Workspace: With the rise of remote work and digital nomadism, a dedicated workspace is increasingly important. Listings that provide comfortable and functional workspaces, equipped with reliable high-speed Wi-Fi, cater to the needs of business travelers and remote workers, enhancing the property's appeal.

3. Patio: A patio or outdoor seating area is a highly desirable feature. It provides a space for guests to relax and enjoy the climate, offering a private area for socializing or unwinding. Properties with well-maintained patios are often more appealing to travelers seeking outdoor leisure options.

4. Parking: Convenient parking is a top priority for many guests, especially those traveling by car. Properties that offer free and secure parking options are more attractive, as this can be a crucial factor for travelers looking to explore the city and its surroundings without the hassle of finding parking.

5. Pool: A swimming pool is one of the most sought-after amenities in San Diego. Given the city's warm climate, a pool provides a refreshing way to cool off and enjoy the outdoors. Properties with pools tend to attract more bookings, especially from families and groups looking for recreational activities.

6. Fireplace: An indoor or outdoor fireplace adds a cozy and inviting element to a property. It can be particularly appealing during cooler evenings or for guests looking to create a relaxing ambiance. Properties with fireplaces can offer a unique selling point, enhancing the overall guest experience.

By incorporating these key amenities, property owners can significantly increase the attractiveness of their listings, catering to the preferences and expectations of travelers visiting San Diego. These features not only enhance guest satisfaction but also contribute to higher booking rates and potentially increased revenue.

**Methodology**

Data Preparation Processing Steps:

In our project, we initially loaded two key datasets: the Airbnb and Crime datasets.

* Airbnb Dataset: This dataset was sourced from "airbnb-project-msba-kaggle-train.csv" and contains detailed information about Airbnb listings, including attributes such as location, amenities, pricing, and reviews.
* Crime Data: We obtained crime statistics from "crime\_data.csv", which provides crime frequencies detailed by zip code, encompassing various types of criminal activities.

Given our focus on San Diego for market analysis, we filtered the Airbnb training dataset accordingly. We also adjusted data types for compatibility across datasets; specifically, the 'zip code' columns in both datasets were converted to string format. This was essential to ensure seamless merging of the datasets based on zip code, avoiding any issues related to data type mismatches.

Subsequently, we performed an inner join on the 'zipcode' column to merge the Airbnb and crime data. This merge operation retained only those records that had matching zip codes in both datasets, integrating features from both sources. This integrated dataset facilitates comprehensive analysis by combining insights from Airbnb listing details with crime statistics.

To refine the dataset further for analysis, we removed irrelevant columns such as 'host\_id', 'host\_url', and 'host\_name'. This was based on either a high prevalence of missing values (over 50% blank cells in some columns) or the columns being extraneous for our analysis purposes. This step ensured that our dataset was clean, relevant, and primed for detailed analytical tasks.

We cleansed the dataset by removing unwanted symbols like $ and , from the "price", "cleaning\_fee", and "host\_response\_rate" columns. Additionally, we transformed binary text columns such as 'requires\_license', 'host\_is\_superhost', 'require\_guest\_phone\_verification', and 'require\_guest\_profile\_picture' by converting "true" to 1 and "false" to 0 for clearer quantitative analysis. We also categorized the 'bed type' column numerically: 'Real Bed' as 0, 'Pull-out Sofa' as 1, 'Futon' as 2, 'Airbed' as 3, and 'Couch' as 4. Finally, the 'room type' column was quantified based on the type of accommodation: “Entire home/apt" as 3, "Private room" as 1, "Shared room" as 0, and "Hotel room" as 2. These transformations streamlined the data for more effective analysis.

We removed the irrelevant symbols such as [, \, {, }, " and ] from the amenities column. Then we found the most significant values from the amenities column and applied one hot encoding on the amenities column. After performing the one hot encoding, we quantified the amenities

column by counting the number of amenities in each listing using the commas which are used to separate the amenities. We have also done the same thing as above with the ‘host\_verifications’ column except one hot encoding.

For the ‘property\_type’column, we first converted all entries to lowercase for standardization. It then simplifies the diverse property types into broader categories such as "entire", "apartment", "hotel", "house", "private room", "shared", and "other" by checking specific keywords within each property type entry. This categorization helps in reducing the complexity of the data. After reclassifying the property types, we applied one-hot encoding to the 'property\_type' column using pandas’ get\_dummies function, which transforms this categorical column into multiple binary columns for each category.

For the ‘cancellation\_policy’ column, we have different types of cancellation policies which are assigned a value of either 0 or 1. Policies classified as 'flexible', 'moderate', 'luxury\_moderate' are considered less strict and are assigned a value of 0. Conversely, more stringent policies like 'strict', 'super\_strict\_60', 'super\_strict\_30', 'luxury\_super\_strict\_95', and 'no\_refunds' are assigned a value of 1. This mapping effectively categorizes the cancellation policies into two groups: less strict and stricter. The map function is then used to apply this dictionary to the cancellation\_policy column in both datasets, replacing the textual descriptions of the policies with their corresponding numerical codes.

Lastly for the host\_response\_time column, we assigned numerical values to categorical data based on the response time of the host. This is facilitated through a dictionary called response\_time\_mapping, which maps various categories of response times to integers. Specifically, 'within an hour' is mapped to 0, 'Within a few hours' is given a value of 1, 'within a day' is rated as 2, and the slowest response, 'a few days or more', is assigned a value of 3. This mapping converts the textual descriptions into a more analytically useful ordinal scale, reflecting the increasing duration of response times.

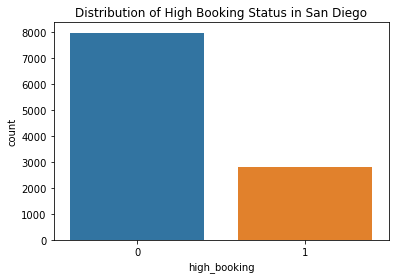
For missing data imputation, we converted the data frame into discrete and continuous based on the number of unique values they contain, followed by handling missing data in these columns. First, it sets a threshold of 15 unique values to distinguish between discrete and continuous columns. Columns with 15 or fewer unique values are categorized as discrete, while those with more are considered continuous. Specific columns are then explicitly listed as either discrete or continuous. For imputation of the missing values, we used different strategies based on the column type. Discrete columns are imputed with the mode (the most frequently occurring value in each column), ensuring that the filled values are the most representative of the column's data. For continuous columns, the KNN Imputer from the scikit-learn library is employed, which fills missing values using the mean of the nearest neighbors found in the dataset, maintaining a more nuanced approach to handling gaps in continuous data.

We checked the distribution of the class in the below exploratory data analysis and found our dataset is not balanced. So, we used the oversampling method called smote on our data. This technique generates synthetic samples from the minority class to create a balanced dataset where the minority and majority classes have an equal number of instances.

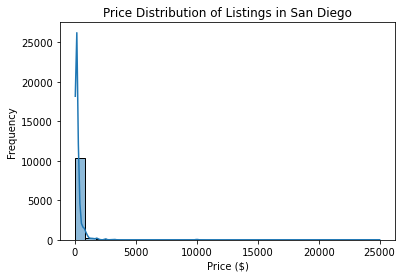
**Exploratory Data Analysis (EDA)** :

In our project, the exploratory data analysis was an essential phase where I visually and quantitatively analyzed the data to uncover underlying patterns, anomalies, and insights. This phase helped in understanding nuances of the dataset.

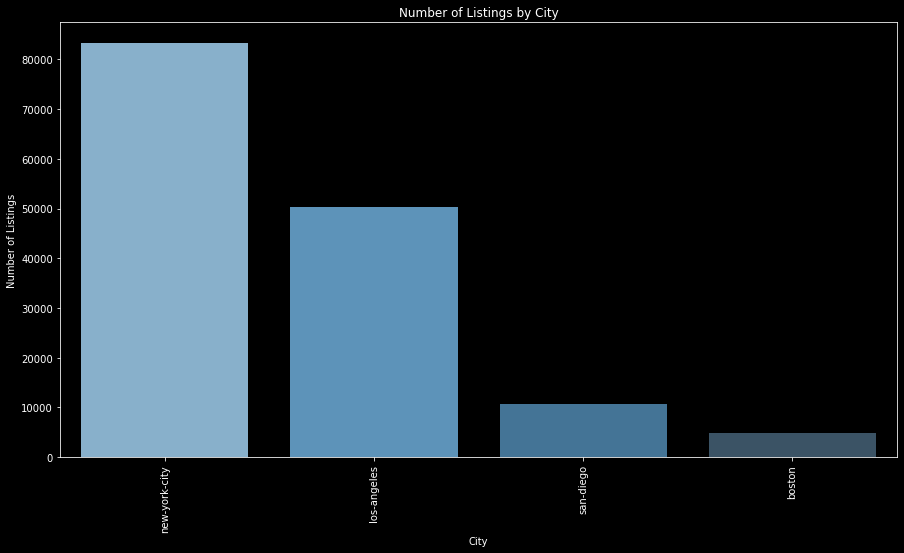
**Visualization Techniques:**

**Count Plots**: I used count plots to examine the distribution of the 'high\_booking' status across listings in San Diego. This helped determine how many listings are considered 'highly booked' and finding the class is imbalance.  
  


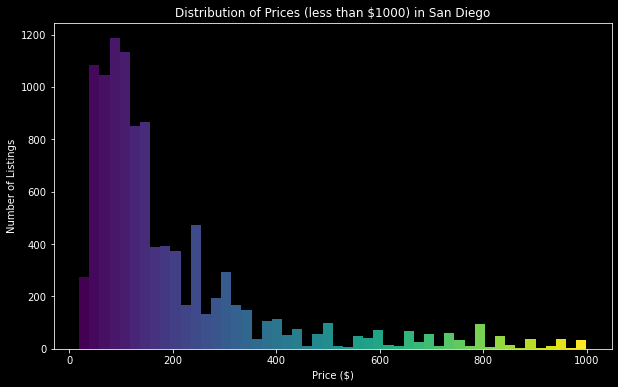
**Histograms**: I created histograms to analyze the price distribution of listings. This included examining prices across all listings and focusing on a subset of listings priced under $1000. These histograms revealed the frequency distribution of prices and highlighted the most common price points, which are critical for understanding pricing strategies in the market.



**Bar Plots**: To compare the number of listings across multiple cities, I plotted bar charts. This comparison was vital to identify which cities had more active listings, offering insights into market dynamics and potentially more lucrative markets.



**Styled Plots**: For the price distribution of listings priced under $1000, I customized the histogram with a dark background and a color gradient. This not only made the graph aesthetically pleasing but also facilitated easier identification of different ranges of prices through color-coding.



**Analytical Insights**:

Through these visualizations and analyses, I gained several insights:

The majority of listings in San Diego did not have a high booking status, which might indicate either a competitive market or a need for improvements in listings.

The price analysis showed a wide range of pricing, with a significant concentration of listings in the lower price range. This could inform pricing strategies for new listings or adjustments for existing ones to enhance competitiveness.

**List of Variables Used in Final Model:**

| **Variable Names** |
| --- |
| Zip code |
| Total Felonies (2020-2021) |
| host\_response\_time |
| host\_response\_rate |
| host\_is\_superhost |
| is\_location\_exact |
| room\_type |
| accommodates |
| bathrooms |
| bedrooms |
| beds |
| bed\_type |
| price |
| guests\_included |
| minimum\_nights |
| maximum\_nights |
| review\_scores\_rating |
| review\_scores\_accuracy |
| review\_scores\_cleanliness |
| review\_scores\_checkin |
| review\_scores\_communication |
| review\_scores\_location |
| review\_scores\_value |
| requires\_license |
| reviews\_per\_month |
| host\_verifications |
| host\_has\_profile\_pic |
| host\_identity\_verified |
| amenities |
| cleaning\_fee |
| instant\_bookable |
| cancellation\_policy |
| require\_guest\_profile\_picture |
| require\_guest\_phone\_verification |
| workspace |
| bbq |
| grill |
| patio |
| pool |
| fireplace |
| property\_type\_hotel |
| property\_type\_house |
| property\_type\_other |

**Implementation and Optimization of CatBoost model in Classification:**

We evaluated the performance of various machine learning models including XGBoost, Random Forest, and LightGBM to identify the most effective algorithm for our dataset. After rigorous testing and comparison, we observed that the CatBoost model consistently outperformed the others in terms of accuracy. This superior performance led us to select CatBoost for a more detailed exploration and presentation within this report. Our decision was based on the objective to utilize the most reliable and accurate model for our predictive tasks, ensuring that our results are both robust and highly applicable to real-world scenarios.

The CatBoost classification model, which is particularly effective with categorical input types, to classify the outcomes based on our training data. Here's a detailed explanation of the steps we took to ensure our model was well-tuned and effective:

We defined a dictionary named param\_grid containing different hyperparameters to be tested with the CatBoost classifier. The parameters included:

* iterations: Specifies the number of boosting iterations, helping to control how much the model fits to the data. More iterations can lead to better performance but might increase the risk of overfitting.
* learning\_rate: Determines the step size at each iteration while moving toward a minimum of a loss function. A smaller learning rate may require more iterations but can achieve more precise convergence.
* depth: Indicates the depth of each tree in the model. Deeper trees can model more complex patterns but may also capture noise in the data.
* bagging\_temperature: A parameter for Bayesian bootstrapping, influencing the selection and resampling of data for building trees. Higher values allow more diverse subsets, potentially improving model robustness.

Model Initialization and Grid Search: We initialized a CatBoostClassifier object and used GridSearchCV for hyperparameter tuning. GridSearchCV works through multiple combinations of parameter values, cross-validating as it goes to determine which tune gives the best performance based on the negative mean squared error. We set cv=10 to use 10-fold cross-validation, providing a good balance between training time and model validation accuracy.

Optimal Parameter Selection and Model Re-training: After the grid search, we extracted the best hyperparameters and used them to initialize a new CatBoostClassifier. This model was then trained on the entire training set to fully adapt it to our data.

Classification and Evaluation: Finally, the optimally tuned model was used to classify outcomes on the test dataset (X\_test). This step is critical for evaluating how well our model performs on unseen data, providing insights into its generalization capability.

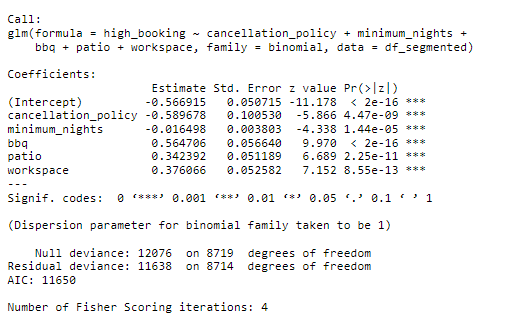
These steps contributed to building a robust predictive model, ensuring that we could confidently rely on its predictions for decision-making processes in our project. This systematic approach to model training and validation ensures that our findings are reliable.

**Results and Findings**

After many trials and various models, we returned results that both supported and refuted our initial hypothesis in relation to our research questions. In terms of treating all listings equally, we quickly found this produces inaccurate results, namely because price has such a heavy influence on the decision making of different consumer groups. We controlled for this by analyzing subsets of the data where we only compared and contrasted listings that were within the same price tiers. Next, we looked at what variables can be manipulated in the short term. Being in the position of a consultant it was our goal to provide clear and actionable recommendations to our clients (Airbnb hosts) and we targeted 4 variables to focus our attention on that can be easily implemented to increase high booking rates. These variables were, cancellation policy (0 if flexible or moderate, 1 if strict), minimum nights stay, bbq (represents the property having an outdoor bbq/grill area), patio and workspace (representing if the property has a room to do work in).

We found negative and significant relationships between cancellation policy and high bookings, and minimum nights stay and high bookings. Cancellation policy was denoted a higher number the stricter the penalty was for canceling and minimum nights stay represents the lowest number of nights a consumer can book the listing for. The host has full control over both of these variables and although reducing these variables brings on a higher level of risk to the host, we are confident that in the long term it will lead to higher booking rates overall. Many consumers want the peace of mind that they’re not going to lose out on their money if they have to cancel due to unforeseen circumstances and either consciously or subconsciously this always plays a part in decision making. Due to the nature of being “unforseen” it is certainly more likely than not that an event like this will not come up and therefore by offering a lenient cancellation policy you increase demand for your listing without adding too much real risk, even if the perceived risk feels high. The minimum nights stay variable has similar intuition in that the host may think they are earning more by forcing customers to stay for a certain number of nights. Again, from the hosts perspective they want to protect their investment by not having a lot of turnover, however this can reduce the visibility of their listing on Airbnb’s platform. By setting a low minimum nights stay number you open your listing up to more views and in turn, more conversions.

We then found positive and significant relationships between workspace, bbq, patio, and high bookings. These were variables that were pulled from the amenities and created as their own binary variable for each listing. In the post Covid world can and do work from anywhere and everywhere, if their job allows it. This has affected the hospitality industry in a way where you now need to support the remote work culture, and as an Airbnb listing this is an area to earn a competitive advantage ([R3](#_stjcl3hma5q)). As for bbq and patio, these were amenities that we identified as specifically relevant to the San Diego market. San Diego is a beautiful, coastal city, where the forecast is always 75 degrees and sunny. People visit for the beaches, culture, and predictably comfortable climate. With a city like this, the vast majority of visitors want to take advantage and be outside as much as possible. That is not to say that the house itself can be falling apart, people always want to stay in nice lodgings while traveling, especially with friends or family, but what can differentiate listings in San Diego is the offering of amenities that enhance the outdoor experience. Also, one does not need to go out and spend thousands of dollars remodeling a patio or purchasing an expensive barbeque set, but having some semblance of these amenities can make a big difference when it comes to booking rates. Even just a private area with some chairs are a small grill can give you an advantage over similar listings because this then allows you to list these as amenities. Of course, you don’t want to be disingenuous regarding the amenities you offer as this will have a negative effect on your reviews scores but one does not need to break the bank to offer these amenities and they can drive some positive results.



To interpret the results of our logistic regression model we looked at the odds ratios by calculating the exponential function for each coefficient. The odds ratios provide a clearer understanding of how each predictor variable influences the likelihood of high bookings. Interpreting these results shows us that going from a strict cancellation policy to a flexible/moderate policy increases the odds of high bookings by approximately 45%, suggesting that leniency in cancellation policies may enhance booking rates. Similarly, each additional minimum night required decreases the odds of high bookings by approximately 1.6%, indicating that properties with shorter minimum stay requirements might attract more bookings. Conversely, the presence of amenities such as BBQ, Patio, and Workspace significantly increases the odds of high bookings, with BBQ amenities boosting the odds by approximately 76%, Patio amenities by 41%, and Workspace amenities by 46%. These findings highlight the importance of amenity offerings in driving high booking rates, while also underscoring the impact of cancellation policies and minimum stay requirements on booking outcomes.

Our recommendations are to account for as many of these changes as you feel comfortable with. We understand each host has a different circumstance and may not be comfortable with some of the more risk-tolerant options such as having a lenient cancellation policy or reducing the minimum nights stay, but we are confident that these suggestions positively affect high booking rates in the long run. Even if you cannot reduce these variables to the lowest options, if you do have them set high to begin with please consider dropping them by a fraction that you are comfortable with, at least for a month or so, and see what the results yield. As for the amenities suggested, we propose creating a budget that you are comfortable with and allocate this to improvements on your amenities. Whether this is putting in a more comfortable outdoor area that has more seating, a grill, and perhaps some lawn games, or allocating this budget to enhancing your indoor workspace. If your listing is located closer to the beach and invites a more outdoorsy crowd you may want to lean more towards the first option, but if your listing is more inland or typically attracts people traveling for work it may be more resourceful to invest in the latter. However, in both cases we believe that these can be cost effective ways to improve your property, target more of the middle class group of consumers, and ultimately lead to high booking rates and more money in your pocket.

**Conclusion and Discussion**

The findings of our research suggest that adjusting certain flexible variables can significantly impact high booking rates for Airbnb hosts. We discovered that lenient cancellation policies and shorter minimum stay requirements increase the odds of high bookings, while amenities such as BBQ, Patio, and Workspace also positively influence booking rates. Our recommendations emphasize flexibility in policy adjustments and strategic investment in amenities tailored to your property's location and target audience. By implementing these suggestions, hosts can enhance their property's appeal, attract a broader consumer base, and ultimately increase booking rates and revenue.

Future developments could explore implementing a seasonality effect to account for fluctuations in demand throughout the year, allowing hosts to optimize pricing and availability. Additionally, understanding how to influence the Airbnb online platform algorithm to gain more views could provide valuable insights for hosts seeking to improve their listing visibility and attract more guests.

**References**

1. San Diego Police Department. "Crime Statistics." City of San Diego,<https://www.sandiego.gov/police/data-transparency/crime-statistics>.
2. San Diego. "Short-Term Rental Host Operating Requirements Checklist." City of San Diego,<https://www.sandiego.gov/sites/default/files/stro_host_operating_requirements_-_checklist_7.6.23.pdf>.
3. Airbnb. "How to Make Your Space Comfortable for Remote Workers." Airbnb Resources: Hosting Homes, Airbnb, n.d.,<https://www.airbnb.com/resources/hosting-homes/a/how-to-make-your-space-comfortable-for-remote-workers-236>.

**Appendix**

Figure 1

