Airbnb Data Analysis in Los Angeles County: Insights for Hosts Using SAP Analytics Cloud (SAC)

Adalberto Delgado, Tremon Everett, Angel Garcia, Nicholas Perez, Jessica Salazar, Jocelyn Salazar Andrade
Dr. Jongwook Woo
California State University, Los Angeles | BUS 5100-93: Introduction To Business Analytics
{adelga126, teveret2, agarci772, nperez125, jherna364, jsalaz18, jwoo5}@calstatela.edu

Abstract: The purpose of this paper is to analyze the Airbnb dataset for Los Angeles County using SAP Analytics Cloud (SAC). The dataset encompasses essential details such as property location, type, capacity, amenities, pricing, booking frequency and guest ratings. Airbnb has revolutionized travel by offering diverse accommodations to suit every traveler's preference and budget. By examining various types of listings and identifying factors contributing to popularity or pricing, this analysis aims to provide insights and recommendations for hosts to set competitive prices while enhancing the appeal of their listings to guests. This story generated by SAC uses simple charts and comparisons to evaluate the factors influencing Airbnb pricing, thereby assisting hosts in optimizing earnings and ensuring guests perceive their stays as excellent value for money.

1. Introduction

This paper utilizes SAP Analytics Cloud to dissect the Airbnb dataset for Los Angeles County. Our dataset was retrieved from Inside Airbnb, a mission-driven project that provides data and advocacy about Airbnb’s impact on residential communities. We have chosen this dataset because Airbnb has transformed the travel landscape, offering diverse accommodations to suit every traveler's preference and budget. Leveraging simple charts and comparisons, this study seeks to highlight the factors influencing Airbnb pricing.

2. Related Work

The significance of analyzing accommodation data, particularly from platforms like Airbnb, has garnered attention over the years. Prior research has explored multiple aspects of Airbnb datasets to highlight factors influencing pricing dynamics, booking trends, and guest satisfaction.

For instance, studies such as the Airbnb Project conducted by UCLA have dived into the correlation between accommodation pricing and key features such as property location, type, and amenities. By employing statistical methods and data visualization techniques, these studies have identified significant predictors of pricing variations, enabling hosts to make informed decisions in setting competitive rates [1].

Research like Mapping Airbnb in Los Angeles conducted by Max Belasco also focused on understanding the impact of Airbnb on local communities and housing markets. Through empirical analyses, he examined the relationship between Airbnb activity and housing affordability, neighborhood dynamics, and regulatory challenges. Such studies provide valuable insights for policymakers and urban planners grappling with the implications of the sharing economy on residential communities [2].

In addition to academic research, industry reports, and analyses from organizations like Inside Airbnb have contributed to our understanding of the platform's impact and operating dynamics. These reports offer insights into accommodation availability, demand trends, and user behavior. This aids in gaining a deeper insight into the data, helping hosts and stakeholders make informed decisions.

Building upon the existing body of work, this paper aims to contribute to the conversation on Airbnb pricing dynamics by leveraging SAP Analytics Cloud to analyze the dataset for Los Angeles County. By synthesizing insights from previous studies and incorporating our analysis of the dataset, we seek to provide recommendations for hosts to optimize their earnings and enhance the guest experience, ultimately ensuring that stays are perceived as excellent value for money.

3. Specifications

Our dataset was retrieved from Inside Airbnb, a project dedicated to providing data and advocacy on Airbnb's impact on residential communities [3]. This source offers valuable insights into Airbnb's operations and its effects on local neighborhoods. Inside Airbnb scrapes data quarterly and covers data from various sources, providing a full understanding of Airbnb accommodations in Los Angeles County. With its straightforward integration and comprehensive coverage, this dataset serves as a reliable foundation for analyzing Airbnb pricing, booking trends, and guest satisfaction.

Table 1 Data Specifications

Data Set:	Size
Airbnb Listings in Los Angeles County – Scraped December 2023	86.4 MB

4. Implementation Flowchart

The dataset sourced from Inside Airbnb encompasses a wide array of information, including Airbnb prices, accommodations, and popular locations within Los Angeles County. Inside Airbnb collects this data from publicly available sources on the Airbnb website, compiling listings and reviews for each property. The data manipulation process involves four key phases: retrieving, cleaning, analyzing, and storytelling. After retrieving the dataset, a few data cleaning steps were performed in Excel, as SAP does not have capabilities for those specific actions. Once completed, the data was uploaded to SAP Analytics cloud, cleaned further, and used to create a story and models. These visualizations were then exported for integration into a PowerPoint presentation.

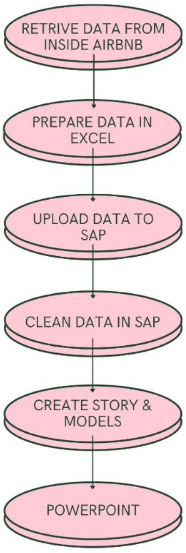


Figure 1 Implementation Flowchart

5. Data Cleaning

After downloading the dataset from Inside Airbnb, we opened the CSV file in Excel to perform actions that cannot be completed in SAP Analytics Cloud. In excel, we toggled off the “Show Formulas” view, performed a count of the number of amenities listed in each cell, and deleted the “host_about” column as it exceeds character limits in SAP and will trigger an error when uploading. After this preparation, the dataset is uploaded to SAP Analytics Cloud directly into a new story. Once uploaded, we further cleaned the data by deleting columns and ensuring that both measures and dimensions were assigned and detailed. After thoroughly cleaning, the data was ready to create a comprehensive story in SAP.

6. Analysis and Visualization

Following data cleaning and identification of the most relevant data for our analysis, we constructed a story and predictive model within SAP Analytics Cloud. This dataset offered visual depictions of key metrics such as the average price per neighborhood, average nightly prices for the number of people accommodated per neighborhood, the density of beds per listing, as well as host response rate v. profile pic v. amount of reviews. Additionally, a time series was created, as well as a predictive model identifying the top three factors influencing bookings.

6.1 Average Listing Price per Neighborhood

The first visualization (Figure 2), a bar chart, was created in SAP Analytics Cloud, and shows the average listing price per neighborhood. The data is sorted from highest to lowest, displaying only the top 10 values. It provides insight into the comparative revenue potential across different neighborhoods based on average listing prices. This information can be useful for current or prospective hosts seeking to make informed decisions regarding the selection of location and pricing strategy for their listings.

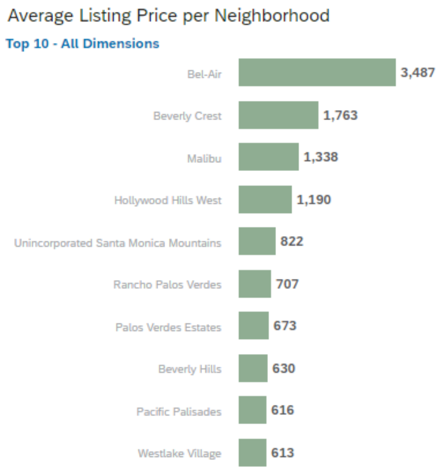


Figure 2 Average Listing Price per Neighborhood

6.2 Distribution of Listings per Room Type

The pie chart (Figure 3), created utilizing SAP Analytics Cloud, displays the proportion of different types of accommodations available on Airbnb for Los Angeles County. It provides a breakdown of the types of rooms hosts currently offer, such as private rooms, entire homes/apartments, shared rooms, and hotel rooms. This information demonstrates the diversity of options available to guests within a particular market or location. This pie chart shows that most listings are for an entire home or apartment.

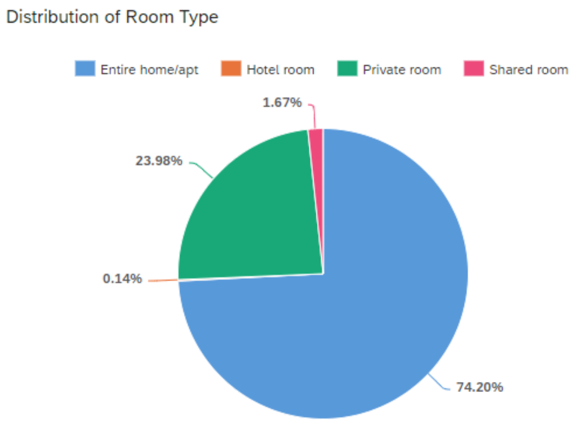


Figure 3 Distribution of Listings per Room Type

6.5 Los Angeles County Airbnb Listings

The geo-map (Figure 4) has multiple layers detailing different data measures from our dataset. The heat layer displays the number of beds per listing and offers valuable insights into the distribution of accommodation sizes across different geographic locations. By identifying regions with a higher concentration of larger accommodations, potential preferences or trends can be identified. This information is important for both hosts and guests: hosts can tailor their listings to match prevailing accommodation sizes in specific areas, while guests can utilize it to find accommodations that meet their bed capacity requirements. Also, this map can unveil potential opportunities or market gaps for hosts interested in offering accommodations with specific bed configurations. The geo-map's bubble layer depicts the number of listings within LA County and their density. The density of listings we display in our geo-map are consistent with the findings of the research conducted by Max Belasco, where he compared Los Angeles Airbnb listings to Los Angeles evictions. [2]. Additionally, the graph incorporates the number of amenities through bubble size and the range of prices through bubble color. Ultimately, this comprehensive visualization provides an overall understanding of the accommodation landscape, facilitating informed decision-making for both hosts and guests.

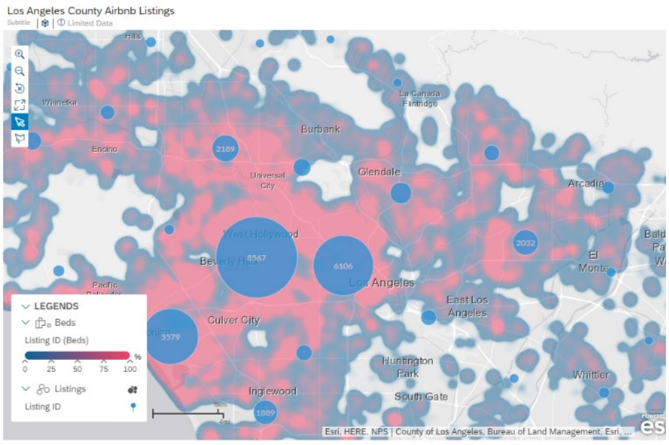


Figure 4 Los Angeles County Airbnb Listings

6.3 Host Response Rate v. Profile Pic v. Number of Reviews

The scatterplot (Figure 5), also generated in SAP Analytics Cloud, depicts how host behaviors influence listing success. The scatterplot indicates that hosts who respond more frequently tend to receive more reviews, which aligns with expectations. However, there are deviations from this trend influenced by the appeal of the listing. One additional dimension represented in this graph is whether the listing host has a profile picture. This is depicted with yellow and blue dots, where yellow indicates "True" and blue indicates "False." Interestingly, the graph reveals not only the correlation between host response rate and reviews but also highlights a significant decrease in reviews when the host does not have a profile picture, regardless of their response rate. Our findings are consistent with the findings of the study conducted by UCLA [1].

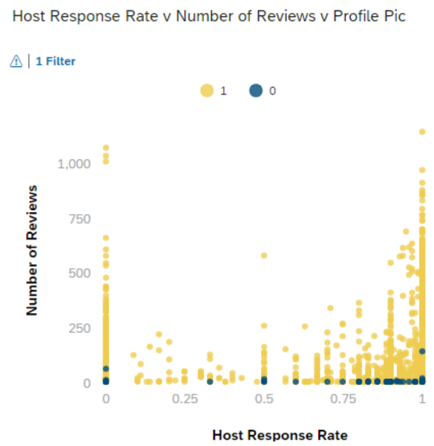


Figure 5 Host Response Rate v. Profile Pic v. Number of Reviews

6.4 New Hosts per Year

The time series (Figure 6) tracks the number of new hosts per year and offers a glimpse into future projections. The data reflects a peak in new hosts during 2016, followed by a dip in 2020 coinciding with the pandemic. Since then, there has been a notable increase in new host registrations. This information provides valuable insights into the growth and dynamics of the Airbnb platform. This data allows stakeholders, including Airbnb itself, investors, policymakers, and researchers, to gauge the platform's expansion over time. Investors can use this data to evaluate the platform's potential for revenue growth and market penetration. Ultimately, tracking new hosts per year serves as a key performance indicator for Airbnb's growth trajectory and its broader implications for the economy and society.

Figure 6 New Hosts per Year

6.6 Average Price, Accommodates, and Amenities per Neighborhood

The bubble chart (Figure 7) shows the average price per listing per neighborhood against the average number of guests the listing can accommodate. Additionally, the bubble size represents the average number of amenities per listing within the neighborhood. This data offers insights into the pricing dynamics of Airbnb listings, revealing how prices fluctuate depending on accommodations, amenities, and location. It aids hosts in comprehending the market demand for accommodations of varying sizes within specific neighborhoods, enabling them to adjust their pricing strategies accordingly. Research conducted by

UCLA supports these findings, indicating a correlation between accommodation size, amenities, and pricing in Airbnb listings [1]. This bubble chart suggests that listings become pricier as they accommodate more people and have more amenities. However, it's essential to note that this trend may vary depending on the listing's location. For example, Westlake Village's average listing accommodates 5 people, 50 amenities and is priced at $613. Then we have a neighborhood like Walnut Park where the average listing accommodates 4.75 people, has 60 amenities and is priced at $114.25. We have two neighborhoods where the accommodations and amenities are practically the same but there is a huge disparity in price. This disparity comes from the ability to charge a higher price on a listing that is in a nicer, favorable geographic area.

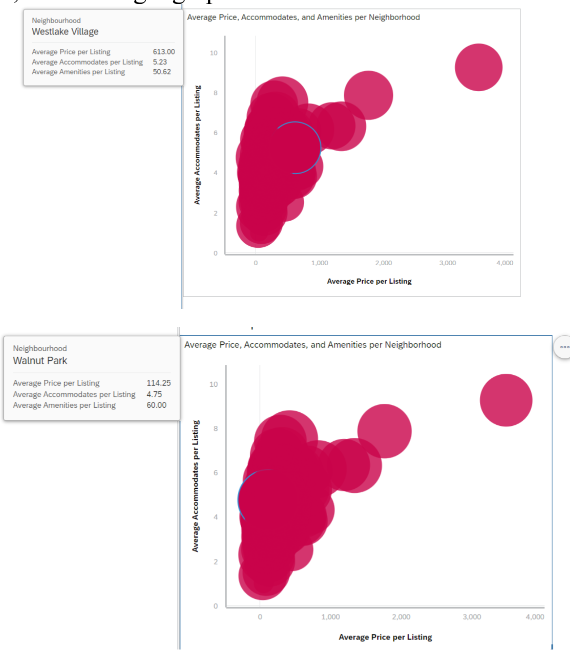


Figure 7 Average Price, Accommodates, and Amenities per Neighborhood

6.6 The Regression Analysis

Understanding what influences rental bookings when hosting an Airbnb is an important part of managing the listing. We created a regression analysis (Figure 8) to help us predict the influencer contributions that most impact bookings. For this analysis we decided to use "number of reviews' as the target predictive goal. We decided on this as the target since the data we used did not specifically have "number of bookings." The number of reviews is a good indication of bookings since the more reviews a listing has the more likely that they were booked. For this analysis we chose five influencers; accommodates, neighborhood, instant bookable, review score rating, and review score cleanliness. The overview report will show how the influencers impact the target and a point of note is that only contributive influencers are displayed in the report. This means that any variables with no contribution are hidden, in our analysis there was one influencer that did not provide meaningful contribution and that was "instant bookable."

Our analysis overview report provided the following results:

Global Performance Indicators

A Root Mean Squared Error (RMSE) of 58.08, this measures the average difference between the values predicted by the model and the actual values. It provides an estimation of how accurate the model will predict the target. After reviewing our score and influencers we decided this was a good score, if necessary, we could have added more influencer variables to lower the score.

The Prediction Confidence score for our analysis was 99.74%, this indicated that our predictive model could achieve the same accuracy if and when applied to any new data set utilizing the same influencers.

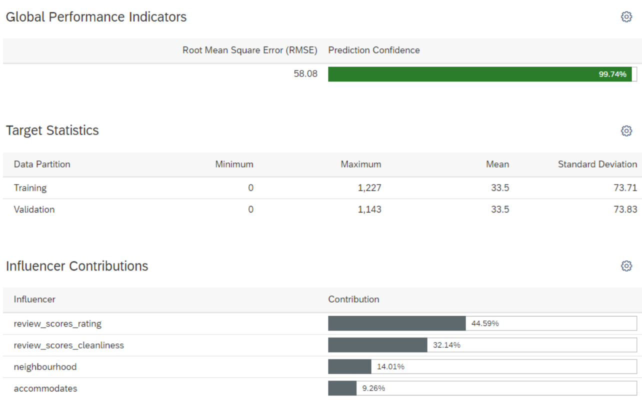
Influencer Contributions

Here we show the importance of each influencer used in our model and sorted by importance to the target. The top two influencers in our model were review scores rating and review score cleanliness, these two contributed ~ 77% of the contribution. This will be extremely helpful when managing Airbnb listings and will provide guidance on what focus areas the host can leverage in order to increase rentals.

Predicted vs. Actual

Our regression analysis displayed a validation and perfect model curve that closely matched. This helps determine if our predictive model is accurate, in order to validate we checked our predictive confidence indicators which we described above as meeting our expectations and we trust the predictive model and the predictions.

We reviewed a comparative analysis on www.medium.com [4] that conducted a similar on Airbnb with the target of "number of bookings". The analysis utilized similar influencers and concluded that it also found review scores rating as their highest contributor to the target. While this result was similar, we also found that instant bookable did contribute to their model whereas it did not for ours.



Global Performance Indicators

Root Mean Square Error (RMSE)	Prediction Confidence
58.08	99.74%

Target Statistics

Data Partition	Minimum	Maximum	Mean	Standard Deviation
Training	0	1,227	33.5	73.71
Validation	0	1,143	33.5	73.83

Influencer Contributions

Influencer	Contribution
review_scores_rating	44.59%
review_scores_cleanliness	32.14%
neighbourhood	14.01%
accommodates	9.26%

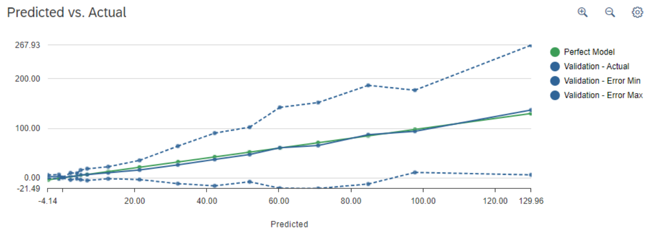


Figure 8 Regression Analysis: Influence on Bookings

7. Conclusion

Finally, we can conclude that based on all the above work:

i. Hosts with higher response rates and profile pictures tend to receive more reviews, indicating a positive correlation between response rate, profile picture presence, and review count. However, hosts with a high response rate but no profile picture receive fewer reviews.
ii. Through the distribution of listings per room type, we observe that the majority of listings offered on Airbnb in Los Angeles County are entire homes or apartments.
iii. Listings generally increase in price with more guests and amenities, although this trend may vary by location.
iv. Hosts must consider factors like amenities, location proximity, and market trends when setting prices, as different neighborhoods offer varying price points.
v. The three primary factors influencing bookings are review scores, cleanliness scores, and neighborhood.

References

[1] Son, M. (n.d.). Air BNB Project. Airbnbproject.Humspace.Ucla.edu. Retrieved March 31, 2024, from https://airbnbproject.humspace.ucla.edu/

[2] Belasco, M. H. (2021, March 23). *Mapping airbnb in Los Angeles*. eScholarship, University of California. https://escholarship.org/uc/item/0nn168dp

[3] Cox, M., Morris, J., & Higgins, T. (2023, December 3). *Get The Data*. Inside Airbnb. Retrieved March 9, 2024, from https://insideairbnb.com/get-the-data

[4] Augustinus, N. (2018, September 10). *This is how you get more bookings and higher ratings as Airbnb host*. Medium. Retrieved March 20, 2024, from https://medium.com/@agustinus.thehub/this-is-how-you-get-more-bookings-and-higher-ratings-as-airbnb-host-ff2846834b3e

[5] https://github.com/salazarjp/airbnbgroup4