IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2025

DOI: 10.17148/IJARCCE.2025.145105

A Study of Pricing and Features with Considerations of Human Error

Ravikesh Kumar Singh[1], Bipin Kharwar[2], Aditya Gupta[3], Shubham Jaiswal[4], Mrs Namita Srivastava[5]

Department of Computer Science and Engineering, Goel Institue of technology and Management, Lucknow Uttar Pradesh, India[1-5]

Abstract: This research paper explores the application of data analytics to an Airbnb dataset containing 74,111 listings, focusing on variables such as room type, accommodates, bathrooms, cancellation policy, cleaning fee, instant bookability, review scores, bedrooms, beds, and log-transformed price. Using Python libraries including pandas, NumPy, and seaborn, we perform exploratory data analysis (EDA) to uncover trends and relationships within the data. The study highlights key statistical insights and identifies potential sources of human error that could impact data quality and analytical and implement KNN algorithm to treat outlier values . The findings provide a foundation for understanding pricing dynamics in the Airbnb market and underscore the importance of addressing human-induced inaccuracies in workflows.

I. INTRODUCTION

The process of making sense of something is called analysis, and the process of making sense of the data that is available is called data analytics. The management of data, which includes the gathering and storing of said data from a variety of sources, as well as the utilization of procedures, tools, and techniques to evaluate said data, is at the heart of this area. By analyzing data and making inferences from it, the purpose of data analytics is to derive correlations, obtain insights, and locate patterns. These actionable insights not only help firms with the decisionmaking process, but they also help with generating predictions and boosting efficiency . Amidst market uncertainties and various geopolitical crises including the Russia-Ukraine conflict, the need for resilience has expanded beyond organizations to encompass governments, citizens, armed forces, education, and other stakeholders. Data analytics and sciences are playing an essential role in a wide range of socio-economic and political initiatives, such as managing the displacement and rehabilitation of refugees, mitigating climate change, reducing food waste, enhancing aid programs' effectiveness, and more. As Web 3.0 and the metaverse continue to grow and gain adoption, organizations and other entities must consider incorporating them into their data analytics initiatives. Additionally, with humans and artificial intelligence collaborating and complementing each other in unprecedented ways, the next wave of data analytics is expected to provide optimal insights and decision-making capabilities. This will result in competitive advantages and enable businesses to adhere to their key performance indicators.

Steps in Data Analytics: Various steps included in data analytics are as follows;

a) Business Analysis: The first step is to comprehend the business issues, set organizational objectives, and plan a feasible solution. E-commerce businesses often face problems like predicting returns, recommending relevant products, identifying fraud, optimizing delivery routes, and more.

b) Data collection: The first step in the data analytics lifecycle is to collect data. To address the business problems, you'll need to gather transactional business data and customer-related info from the past few years. This data can include details such as the number of units sold, sales and profits, and order dates. This information is crucial in shaping the future of the business as it provides insights into past trends and helps identify patterns that can inform future decisions.

c) Data Cleaning: The second step in the data analytics lifecycle is to clean the data. The data collected can often be unstructured, disordered, and contain missing values. Therefore, it is necessary to clean the data by removing redundant, irrelevant, and missing values. This step ensures that the data is accurate, complete, and ready for analysis.

d) Analyzing the data: After collecting and processing the data, the next step is to conduct investigative data analysis using a variety of data mining, predictive analytics, and business intelligence tools, as well as data visualization methods. This enables the prediction and analysis of future outcomes. By examining the data, valuable insights can be uncovered, such as customer delivery timeframes, purchasing habits, returned items, and other important information.

e) Conclusion/Results: The final step in the data analytics lifecycle is to interpret the results. Once the data has been cleaned and analyzed, it's time to make sense of the findings. This step involves identifying hidden patterns, predicting future trends, and gaining insights that can support datadriven decision-making. It's important to validate whether the

DOI: 10.17148/IJARCCE.2025.145105

results meet your expectations and align with your business goals. This step enables organizations to make informed decisions and take actions based on data-driven insights

The rise of the sharing economy has transformed industries such as hospitality, with Airbnb emerging as a good player. Understanding the factors that influence pricing and customer satisfaction in Airbnb listings is critical for hosts, guests, and platform operators. Data analytics offers a powerful approach to uncovering these insights by leveraging large datasets to identify patterns and correlations.

This study analyzes an Airbnb dataset extracted from an Excel file ('Air_BNB.xlsx'), as processed in a Jupyter Notebook environment. The dataset includes 74,111 records with 10 variables after dropping an identifier column ('id'). The primary objective is to perform exploratory data analysis (EDA) to summarize the dataset's characteristics and explore potential relationships, while also considering the role of human error in data collection and preprocessing.

II. LITERATURE REVIEW

The rise of the sharing economy, particularly through platforms like Airbnb, has generated significant interest in understanding the factors influencing rental prices. This literature review synthesizes key studies that explore price prediction models for Airbnb listings, focusing on machine learning techniques, feature engineering, and the impact of various predictors on pricing.

Previous works

Study	Year	Key Focus	Methodology	Key Findings	Limitations	Future Directions
Quattrone et al.	2016	Determinants of Airbnb pricing in London	Statistical analysis	Property characteristics (room type, bedrooms) and location significantly affect prices	Limited to one city	Include multi-city analysis
Gibbs et al.	2018	Pricing factors in Canadian Airbnb markets	Hedonic pricing model	Accommodates, cancellation policies, and cleaning fees are key predictors	Limited feature set	Incorporate host reputation features
Zhang et al	2019	Machine learning for price prediction in NYC	Linear regression, random forest	Random forests outperform linear regression due to non-linear relationships	Single-city focus	Test models across multiple cities
Hong & Yoo	2020	Handling skewed price distributions	Robust regression	Log-transformation mitigates outliers	Limited feature exploration	Include external factors

III. METHODOLOGY

3.1 Data Source

The dataset, stored in `Air_BNB.xlsx`, contains Airbnb listing information with the following columns after preprocessing:

DOI: 10.17148/IJARCCE.2025.145105

room_type: Type of accommodation (e.g., Entire home/apt, Private room, Shared room),accommodates: Number of guests the listing can accommodate, bathrooms: Number of bathrooms, cancellation_policy: Policy type (e.g., strict, moderate, flexible), cleaning_fee: Binary indicator (1 = yes, 0 = no), instant_bookable: Binary indicator (t = true, f = false), review_scores_rating: Rating out of 100, bedrooms: Number of bedrooms, beds: Number of beds, log_price: Log-transformed price (target variable).

3.2 Tools and Techniques

The analysis was conducted using Python 3.12.3 in a Jupyter Notebook environment. Key libraries included:
pandas: For data manipulation and summary statistics.
NumPy: For numerical operations.
seaborn and matplotlib: For visualization .
sklearn.linear_model: for statistical analysis and creation regression model to find data accuracy

The methodology involved loading the dataset, dropping the `id` column, and performing EDA through descriptive statistics, unique value counts, and missing value analysis using KNN algorithm and visualization on power BI.

3.3 Data Preprocessing

The `id` column was removed to focus on meaningful features. No further preprocessing (e.g., handling missing values) was completed in the provided code, though missing values were identified.

IV. RESULTS AND ANALYSIS

4.1 Dataset Overview

The dataset comprises 74,111 rows and 10 columns. Descriptive statistics (Table 1) reveal the following:

Accommodates: Mean = 3.16, Range = 1–16, **Bathrooms:** Mean = 1.24, Range = 0–8, **Cleaning Fee:** 73% of listings have a cleaning fee (mean = 0.73)**, Review Scores:** Mean = 94.07, Range = 20–100 (with significant missing data), **Bedrooms:** Mean = 1.27, Range = 0–10, **Beds:** Mean = 1.71, Range = 0–18, **Log Price:** Mean = 4.78, Range = 0–7.6.

Table 1: Summary Statistics of Numerical Variables

variable	Count	Mean	Std	Min	25%	50%	75%	Max
Accommodate	74108	3.16	2.15	1.0	2.0	2.0	4.0	16.0
Bathroom	73908	1.24	0.58	0.0	1.0	1..0	1.0	8.0
Cleaning fee	74107	0.73	0.44	0.0	0.0	1.0	1.0	1.0
Review score rating	57389	94.07	7.84	20.0	92.0	96.0	100.0	100.0
Bedrooms	74109	1.27	0.85	0.0	1.0	1.0	1.0	10.0
Beds	73980	1.71	1.25	0.0	1.0	1.0	2.0	18.0
Log Price	74111	4.78	0.72	0.0	4.32	4.71	5.22	7.6

4.2 Categorical Variables

Room Type: 3 unique values (Entire home/apt: 41,308, Private room: 30,635, Shared room: 2,163).
Cancellation Policy: 3 unique values (strict: 32,500, moderate, flexible).
Instant Bookable: 2 unique values (f: 54,660, t: 19,451).

4.3 Missing Data

Missing values were identified as follows:
review_scores_rating: 10,215 missing (13.8% of total), bathroom: 195 missing, bedrooms: 92 missing, beds: 125 missing, Other columns: Minimal missingness (<10).

4.4 Observations

Entire home/apartment listings dominate (55.7%), suggesting a preference for private accommodations.
The high average review score (94.07) indicates general satisfaction, though missing ratings may skew this perception.
Log-transformed price suggests a skewed distribution, typical in pricing data, with a mean of 4.78 (approximately $117 when exponentiated, assuming base e).

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2025
DOI: 10.17148/IJARCCE.2025.145105

V. DISCUSSION

5.1 Insights from the Data

During exploratory data analysis we found that accommodations with more bedrooms, beds, and bathrooms tend to have higher log prices, consistent with expectations in hospitality pricing models. The prevalence of cleaning fees (73%) and strict cancellation policies (43.8%) may reflect host strategies to maximize revenue and minimize risk. The high proportion of non-instant-bookable listings (73.7%) suggests a preference for manual approval, possibly linked to trust or pricing control.

5.2 Human Error in Data Analytics

Human error can significantly impact data analytics, particularly in the following areas observed in this study:

1. Data Entry Errors

Listings with 0 bedrooms or beds (minimum values) may reflect input errors rather than actual offerings (e.g., a studio misreported). Similarly, a bathroom count of 0 seems implausible for functional accommodations.
Example: A listing with 0 beds but accommodating 2 guests is likely a typo or misclassification.

2. Missing Data Handling

The significant missingness in 'review_scores_rating' (13.8%) could result from human oversight during data collection or failure to prompt reviews. The lack of imputation or removal in the provided code suggests an incomplete preprocessing step, potentially biasing results if analyzed further.

3. Coding Errors

The import of 'sklearn.linear_model as LinearRegression' is incorrect (should be 'from sklearn.linear_model import LinearRegression'), indicating a potential oversight by the coder. While not executed here, such errors could derail subsequent modeling.
The loop printing 'room_type' value counts for all categorical variables (cell 25) is a logical error, limiting insight into 'cancellation_policy' and 'instant_bookable'.

4. Interpretation Errors

The dataset reduction from 74,111 to 54,117 rows (cell 30) is unexplained, possibly due to an accidental overwrite or filtering not documented in the code. This could lead to misinterpretation if not addressed.

5.3 Implications

Human errors in data entry, preprocessing, and coding can compromise the reliability of analytical outcomes. For instance, unaddressed missing values in 'review_scores_rating' could overestimate satisfaction, while erroneous feature values (e.g., 0 beds) could distort regression models predicting 'log_price'. Robust data validation and error-checking protocols are essential to mitigate these risks.

VI. CONCLUSION

This study demonstrates the utilization of data analytics in exploring Airbnb listing characteristics, revealing trends in pricing and accommodation features. However, the presence of human errors—ranging from data entry inaccuracies to coding oversights—underscores the need for rigorous quality control in data science workflows. By addressing human-induced errors, we can enhance the validity and applicability of findings in real-world contexts.

REFERENCES

[1] Baig, M. I., Shuib, L., & Yadegaridehkordi, E. (2020). Big data in education: A state of the art, limitations, and future research directions. International Journal of Educational Technology in Higher Education, 17(1), 44. https://doi.org/10.1186/s41239-020-00223-0

[2] Katkar, S. V., Kharade, S. K., Kharade, K. G., & Kamat, R. K. (2020). Integration of Technology for Advancement in Supply Chain Management. In New Paradigms in Business Management Practices (Vol. 3, pp. 116–123). Amazon Publication.

[3] Kharade, K. G., Kharade, S. K., Sonawane, V. R., Bhamre, S. S., Katkar, S. V., & Kamat, R. K. (2021). IoT Based Security Alerts for the Safety of Industrial Area. In M. Rajesh, K. Vengatesan, M. Gnanasekar, Sitharthan.R, A. B. Pawar, P. N. Kalvadekar, & P. Saiprasad (Eds.), Advances in Parallel Computing. IOS Press. https://doi.org/10.3233/APC210185

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2025

DOI: 10.17148/IJARCCE.2025.145105

[4] Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Defi nitive Guide to Dimensional Modeling (3rd ed.). Wiley.

[5] Maheshwari, A. (2014). Data Analytics Made Accessible.

[6] Sonavane, A. K. K. (2021). Study of Emerging Role of Data Science in Business Intelligence. Design Engineering, 6.

[7] Marchena Sekli, G. F., & De La Vega, I. (2021). Adoption of Big Data Analytics and Its Impact on Organizational Performance in Higher Education Mediated by Knowledge Management. Journal of Open Innovation: Technology, Market, and Complexity, 7(4), 221. https://doi.org/10.3390/joitmc7040221

[8] Maroufkhani, P., Wagner, R., Wan Ismail, W. K., Baroto, M. B., & Nourani, M. (2019). Big Data Analytics and Firm Performance: A Systematic Review. Information, 10(7), 226. https://doi.org/10.3390/info10070226

[9] Naikwadi, B. H., Kharade, K. G., Yuvaraj, S., & Vengatesan, K. (2021). A Systematic Review of Blockchain Technology and Its Applications. In Recent Trends in Intensive Computing (pp. 467–473). IOS Press.

[10] Patil, S., Mujawar, A., Kharade, K. G., Kharade, S. K., Katkar, S. V., & Kamat, R. K. (2022). Drowsy Driver Detection Using Opencv And Raspberry Pi3. Webology, 19(2), 6003–6010.

[11] Popovič, A., Hackney, R., Tassabehji, R., & Castelli, M. (2018). The impact of big data analytics on firms' high value business performance. Information Systems Frontiers, 20(2), 209–222. https://doi.org/10.1007/s10796-016-9720-4

[12] Sonavane, A. K. K. (2021). An In-Depth Study of Retail Sales Trend and Pattern based on Exploratory Data Analysis. Design Engineering, 6313-6327.

[13] Prathima, Ch., Muppalaneni, N. B., & Kharade, K. G. (2022). Deduplication of IoT Data in Cloud Storage. In Ch. Satyanarayana, X.-Z. Gao, C.-Y. Ting, & N. B. Muppalaneni (Eds.), Machine Learning and Internet of Things for Societal Issues (pp. 147–157). Springer Nature Singapore. https://doi.org/10.1007/978-981-16-5090-1_13

[14] Sarker, I. H. (2021). Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. SN Computer Science, 2(5), 377. https://doi.org/10.1007/s42979-021-00765-8

[15] Swami, A., Patil, A., & Kharade, K. G. (2019). Applications of IoT for Smart Agriculture or Farming. International Journal of Research and Analytical Reviews, 6(2), 537–540.

