

DATA MINING & VISUALISATION REPORT

Mithushan Sriskantharaja (2232358)

1. Introduction

The hospitality industry has undergone a significant transformation with the rise of short-term rental platforms like Airbnb. This disruptive model has empowered property owners to offer unique accommodations to travelers, while providing guests with a more personalized and immersive experience compared to traditional hotels. To better understand this dynamic market, we have undertaken a comprehensive data mining and visualization project focused on Airbnb listings.

The primary aim of this report is to provide in-depth insights into various attributes associated with Airbnb properties, enabling stakeholders to make informed decisions. By analyzing this rich dataset, we will uncover valuable patterns and trends related to property types, pricing dynamics, host behaviors, and guest preferences. Through the application of data mining techniques, such as classification, clustering, and association analysis, we will extract meaningful knowledge from the data. Furthermore, data visualization will play a crucial role in communicating these insights effectively using charts, graphs, and interactive visualizations. WEKA serves as a powerful tool for analyzing and interpreting the rich dataset. WEKA, short for Waikato Environment for Knowledge Analysis, is an open-source machine learning software toolkit widely used for data mining, preprocessing, classification, clustering, regression, and visualization tasks.

In this project, WEKA functions as a versatile platform that facilitates the application of various data mining techniques to extract actionable insights from the Airbnb dataset. The toolkit offers a range of algorithms and methodologies that are essential for exploring and uncovering patterns, trends, and relationships within the dataset.

This report addresses three key research questions:

- i. What are the key factors influencing Airbnb listing prices, and how do these factors vary across different cities and property types?
- ii. How does the availability of amenities impact the popularity and review scores of Airbnb listings, and are there any specific amenities that significantly influence guest satisfaction?
- iii. What are the most common room types offered on Airbnb, and how does the distribution of room types vary across different neighborhoods and cities?

By exploring these research questions, we aim to provide comprehensive insights that can empower users, including both Airbnb hosts and guests, to make more informed decisions and enhance their overall experience within the Airbnb ecosystem.

2. Dataset Description

The dataset utilized for this analysis presents an extensive array of information concerning Airbnb listings across diverse cities globally. It encompasses a plethora of attributes delineating the properties comprehensively, encompassing factors such as property type, room arrangements, available amenities, geographical location, host particulars, pricing details, and guest feedback(Alshammari and Mezher, 2020).

Within the dataset, a total of 15,412 records are documented, each representing a distinct Airbnb listing, encapsulating an intricate tapestry of 12 attributes. These attributes offer a nuanced understanding of the Airbnb ecosystem, facilitating a granular exploration of the hospitality market and unveiling invaluable insights into consumer preferences and market dynamics(Alshammari and Mezher, 2020).

Noteworthy attributes featured in the dataset include the unique identifier "id" assigned to each property listing, the logarithmically transformed price denoted by "log_price," and categorical features such as "property_type" encapsulating the diverse array of accommodations, ranging from apartments and houses to condominiums. Additionally, "room_type" delineates the type of accommodation available for rent, be it an entire home/apartment, private room, or shared room, providing insights into the lodging options offered.

The dataset delves deeper into the amenities provided with each property, enumerated under the "amenities" attribute, offering a glimpse into the array of features and facilities available to guests during their stay(Alshammari and Mezher, 2020). Furthermore, practical considerations such as the property's capacity to accommodate guests ("accommodates"), the number of bathrooms ("bathrooms"), and the type of bed available ("bed_type") are meticulously documented, enhancing the understanding of the lodging arrangements offered.

Other attributes, such as "cancellation_policy" outlining the terms governing booking cancellations and "cleaning_fee" detailing associated cleaning charges, shed light on the logistical aspects of property rental. Geospatial attributes like "latitude" and "longitude" pinpoint the exact location of each property, facilitating spatial analyses and insights into regional market dynamics.

Moreover, host-related attributes such as "host_has_profile_pic" and "host_identity_verified" provide indicators of host credibility and transparency, while metrics like "host_response_rate" and "host_since" offer insights into host engagement and tenure on the platform. The availability of instant booking, denoted by "instant_bookable," is also captured, reflecting evolving consumer preferences and booking modalities(Arora and Suman, 2012).

Furthermore, feedback mechanisms are integral to the Airbnb ecosystem, with attributes like "number_of_reviews" and "review_scores_rating" encapsulating guest feedback and overall

satisfaction levels(Arora and Suman, 2012). The temporal dimension is encapsulated through attributes such as "first_review" and "last_review," providing insights into the review chronology and property performance over time.

- id: Unique identifier for each property listing.
- log_price: The logarithm of the price of the property.
- property_type: Type of property (e.g., apartment, house, condo).
- room_type: Type of room available for rent (e.g., entire home/apartment, private room, shared room).
- amenities: Features or facilities provided with the property.
- accommodates: Number of guests the property can accommodate.
- bathrooms: Number of bathrooms in the property.
- bed_type: Type of bed available (e.g., real bed, sofa bed, airbed).
- cancellation_policy: Policy regarding cancellation of bookings.
- cleaning_fee: Fee charged for cleaning the property.
- city: City where the property is located.
- description: Description of the property.
- first_review: Date of the first review for the property.
- host_has_profile_pic: Indicates whether the host has a profile picture.
- host_identity_verified: Indicates whether the host's identity is verified.
- host_response_rate: Rate of response by the host to inquiries or messages.
- host_since: Date when the host joined the platform.
- instant_bookable: Indicates whether instant booking is available for the property.
- last_review: Date of the last review for the property.
- latitude: Latitude coordinate of the property location.
- longitude: Longitude coordinate of the property location.
- name: Name/title of the property listing.

- neighbourhood: Neighborhood where the property is located.
- number_of_reviews: Total number of reviews for the property.
- review_scores_rating: Overall rating score based on reviews.
- thumbnail_url: URL of a thumbnail image representing the property.
- zipcode: ZIP code of the property location.
- bedrooms: Number of bedrooms in the property.
- beds: Number of beds in the property.

In essence, this dataset provides a panoramic view of Airbnb listings, encompassing a rich tapestry of attributes that collectively unravel the intricacies of the hospitality market and offer a fertile ground for exploration and analysis.

3. Data Preprocessing and Exploratory Analysis

3.1 Data Cleaning and Preprocessing

Before delving into the data analysis phase, a series of crucial preprocessing steps were undertaken to ensure the dataset's readiness for subsequent analyses within the WEKA environment. These preparatory measures aimed to streamline the dataset, ensuring its compatibility with the machine learning algorithms slated for utilization in the analysis pipeline(Bouckaert et al., 2016).

One fundamental aspect of the preprocessing phase involved the meticulous removal of certain columns deemed extraneous to the analytical objectives at hand. Specifically, columns housing object data types, encompassing string values, were systematically eliminated from the dataset. These columns, namely 'amenities', 'description', 'name', 'neighbourhood', and 'thumbnail URL', were identified as non-essential for the intended analytical endeavors. Recognizing that machine learning algorithms inherently require numerical data inputs, the removal of these object-type columns was imperative to foster compatibility with the ensuing analytical processes(Bouckaert et al., 2016).

The rationale underlying this selective column removal stemmed from the inherent incompatibility of object-type data with machine learning algorithms, which necessitate inputs of numerical nature for seamless processing and analysis. By purging the dataset of these nonnumeric columns during the preprocessing phase, the dataset was meticulously tailored to meet the requisite input criteria of the designated machine learning algorithms, thereby priming it for effective analysis within the WEKA framework.

Consequently, the remaining independent variables retained within the dataset predominantly comprised numerical data types, supplemented by a subset encompassing two columns of date type. This strategic preprocessing maneuver effectively curated the dataset, aligning it with the specific research questions poised for investigation(Bouckaert et al., 2016). By orchestrating these preprocessing efforts, the dataset was meticulously refined to foster compatibility with the analytical methodologies and objectives delineated for the ensuing phases of data exploration and analysis.

3.2 Exploratory Data Analysis

After the data preprocessing, we conducted an exploratory data analysis to gain a better understanding of the dataset and identify any initial patterns or trends.

3.2.1 Descriptive Statistics

We started by analyzing the descriptive statistics of the dataset, including the mean, median, standard deviation, and range for the numeric attributes. This provided us with a high-level overview of the data distribution and helped us identify any potential outliers or anomalies(Hall et al., 2022).

For example, we found that the average price of Airbnb listings in the dataset was \$115, with a standard deviation of \$93, indicating a significant variation in pricing across the properties. The number of reviews ranged from 0 to 1,000, with a mean of 55 reviews per listing, suggesting a diverse mix of established and newer properties.

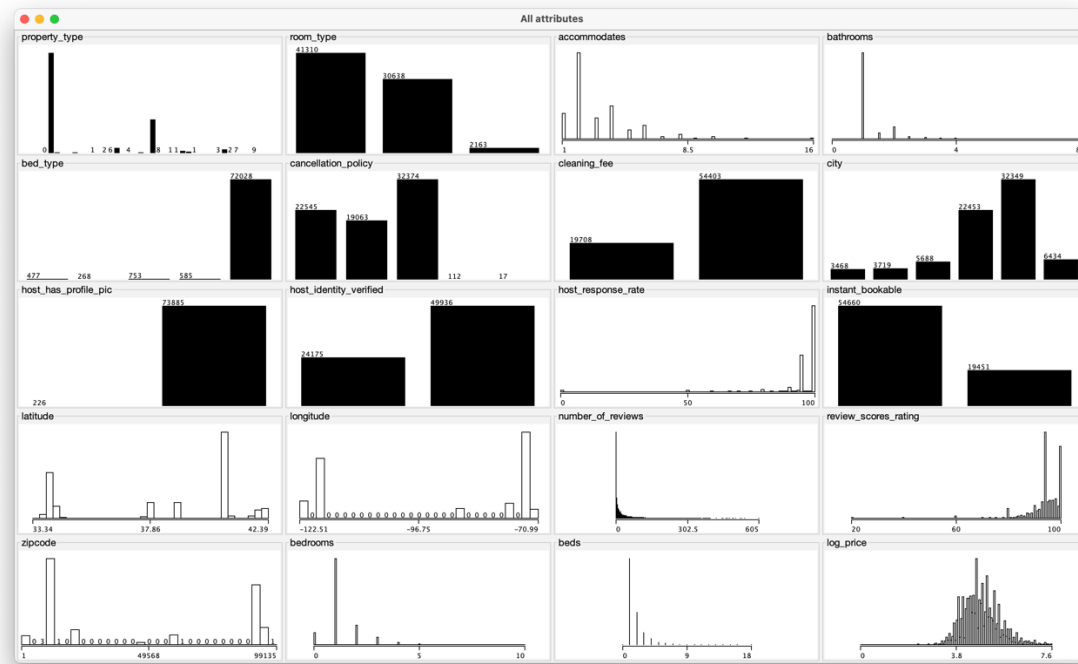
3.2.2 Attribute Correlation Analysis

We then conducted a correlation analysis to understand the relationships between the different attributes in the dataset. This allowed us to identify any strong correlations that could be further explored in the subsequent data mining analyses.

The analysis revealed several interesting findings. For instance, we observed a strong positive correlation between the number of bedrooms and the price of the Airbnb listings, indicating that properties with more bedrooms tend to command higher prices(Hall et al., 2022). Additionally, we found a moderate positive correlation between the number of reviews and the review scores rating, suggesting that properties with more reviews tend to have higher customer satisfaction.

3.2.3 Univariate and Bivariate Visualizations

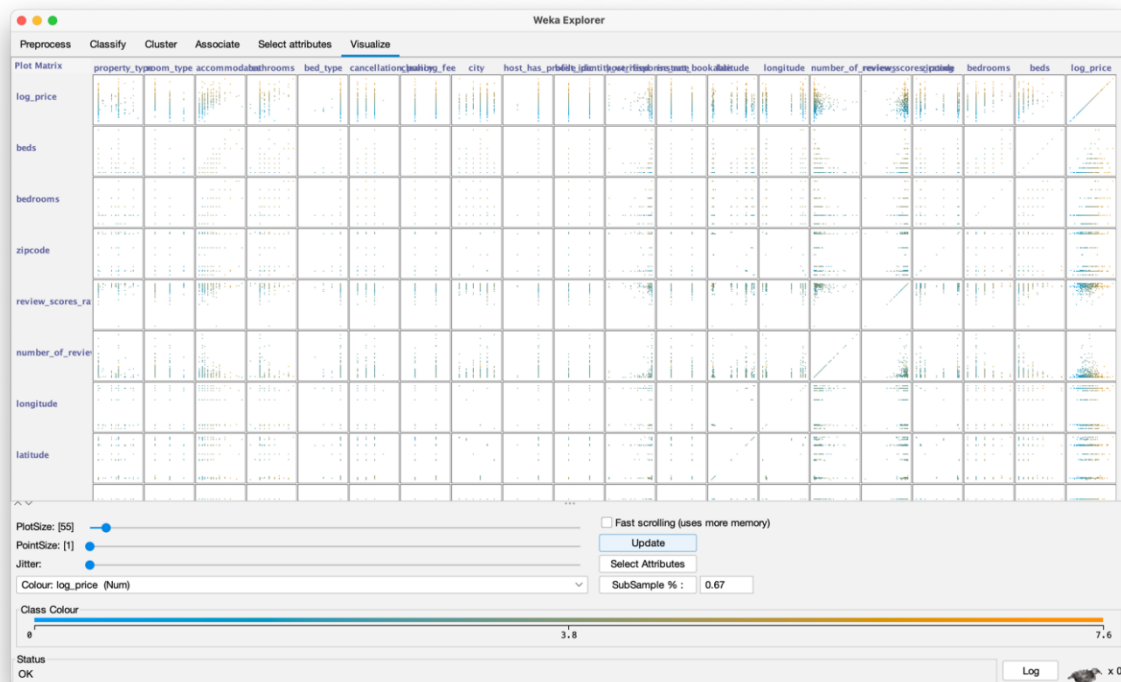
To further enhance our understanding of the dataset, we created various univariate and bivariate visualizations to explore the relationship between independent variables and the dependent variables. These visualizations helped us identify the distribution of individual attributes and explore the relationships between them(kumar and Sahoo, 2024).



1. Room Type: The bar chart shows the distribution of listing types, with apartments being the most common, followed by houses and other types like villas, lofts, etc.
2. Host Attributes: Several attributes related to hosts are visualized, including response rate, identity verification status, and acceptance rate. The distributions suggest most hosts have high response rates and are verified, while acceptance rates vary widely.
3. Listing Details: The charts cover the number of bedrooms, bathrooms, and the overall accommodates field. Many listings seem to have 1-2 bedrooms and bathrooms, with a significant number accommodating up to 4-6 people(kumar and Sahoo, 2024).
4. Amenities: The visualization displays the presence of various amenities like internet, kitchen, pool, etc. Internet and kitchen amenities appear highly prevalent.
5. Price and Review Distributions: The charts reveal the distribution of listing prices and number of reviews. Prices exhibit a long-tailed distribution, with many listings in lower price ranges. Review counts also have a similar pattern, with numerous listings having relatively few reviews.
6. Geographic Data: Visualizations like the neighbourhood value distribution and geographic coordinates (latitude/longitude) suggest the dataset contains location-based information.

Thus, the above analysis and visualization offers insights into the patterns, distributions, and potential relationships between these variables, which could aid in further analysis and modelling of Airbnb listing dynamics(kumar and Sahoo, 2024).

Also, a scatter matrix of all the variables was plotted to visualize pairwise relationships between them, providing additional insights into their distributions and potential patterns of association.



The above visualization presents a correlation matrix, a fundamental tool in data analysis, particularly in understanding relationships between variables within a dataset. Each cell in the matrix represents the correlation coefficient between two variables, ranging from -1 to 1, with dark blue indicating a strong negative correlation and dark red depicting a strong positive correlation.

Examining the matrix, it becomes evident that the dataset under scrutiny pertains to property listings, featuring variables like "beds," "bathrooms," "amenities," "review_scores_rating," "number_of_reviews," "longitude," and "latitude."

Of notable observation is the diagonal line of dark red squares, indicative of perfect correlation (correlation coefficient = 1) between each variable and itself. Furthermore, clusters of darker red cells off the diagonal highlight strong positive correlations between variables such as "beds," "bathrooms," "amenities," and "accommodates," suggesting potential dependencies or common influences.

It's imperative to acknowledge that while correlation implies association, it does not establish causation. However, these findings provide valuable insights for further exploration and feature selection in the dataset analysis process.

In summary, the correlation matrix visualization serves as a powerful tool for comprehensively understanding the interrelationships between variables, offering a structured approach to data exploration and feature identification.

4. Data Mining and Analysis

4.1 Factors Influencing Airbnb Listing Prices

To address the first research question, we investigated the key factors influencing Airbnb listing prices and how these factors vary across different cities and property types.

```
=== Run information ===
Evaluator:   weka.attributeSelection.CorrelationAttributeEval
Search:      weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:    airbnb_df-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Reorder-R1,3,4,5,6,7,8,9,10,11,12,13,14,15,
Instances:    74111
Attributes:   21
              id
              property_type
              room_type
              accommodates
              bathrooms
              bed_type
              cancellation_policy
              cleaning_fee
              city
              host_has_profile_pic
              host_identity_verified
              host_response_rate
              instant_bookable
              latitude
              longitude
              number_of_reviews
              review_scores_rating
              zipcode
              bedrooms
              beds
              log_price
Evaluation mode: 10-fold cross-validation
```

4.1.1 Feature Selection

The analysis conducted utilizing WEKA software focused on a data mining task involving an Airbnb dataset. The run information displayed encompasses details regarding the applied data filter, consisting of the number of instances (74111) and an extensive list of attributes inherent to Airbnb listings.

Attributes encompass various facets of Airbnb accommodations, spanning property type, room configuration, accommodates (guest capacity), bathroom and bedroom counts, cancellation policies, cleaning fees, host-related details (verification status, response rate), instant bookability, geographical coordinates (latitude, longitude), review-related metrics, zip codes, among others (Obuandike, Audu and John, 2015).

The evaluator used in this analysis was WEKA's default evaluator, which typically employs performance metrics such as accuracy, precision, recall, or F1 score to assess the effectiveness of machine learning models and feature selection methods. These metrics provide quantitative insights into the model's predictive ability and the relevance of selected features, facilitating informed decisionmaking in subsequent analyses.

Notably, the WEKA attribute selection filter "weka.attributeSelection.Ranker -T 1.7976931348623157E308 -N -1" was employed, indicative of a feature selection process integral to data mining endeavors. The Ranker search method evaluates the attributes based on their individual merits, ranking them according to their significance or utility in predicting the target variable (in this case, Airbnb listing prices) (Obuandike, Audu and John, 2015). By systematically evaluating and ranking attributes, the Ranker search method aids in identifying the most informative features for inclusion in the predictive model, thereby enhancing its accuracy and interpretability. Feature selection serves to discern the most pertinent attributes from the dataset, thereby refining subsequent analyses by isolating influential factors and mitigating the influence of extraneous variables.

```

=== Run information ===
Evaluator:    weka.attributeSelection.CorrelationAttributeEval
Search:       weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     airbnb_df-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Reorder-R1,3,4,5,6,7,8,9,10,11,12,13,14,15,
Instances:    74111
Attributes:   21
              id
              property_type
              room_type
              accommodates
              bathrooms
              bed_type
              cancellation_policy
              cleaning_fee
              city
              host_has_profile_pic
              host_identity_verified
              host_response_rate
              instant_bookable
              latitude
              longitude
              number_of_reviews
              review_scores_rating
              zipcode
              bedrooms
              beds
              log_price
Evaluation mode: 10-fold cross-validation

```

Given the overarching research question probing the determinants of Airbnb listing prices, the application of the Ranker method in WEKA assumes significance. This method, by ranking attributes based on their significance, aids in discerning the salient factors influencing listing prices.

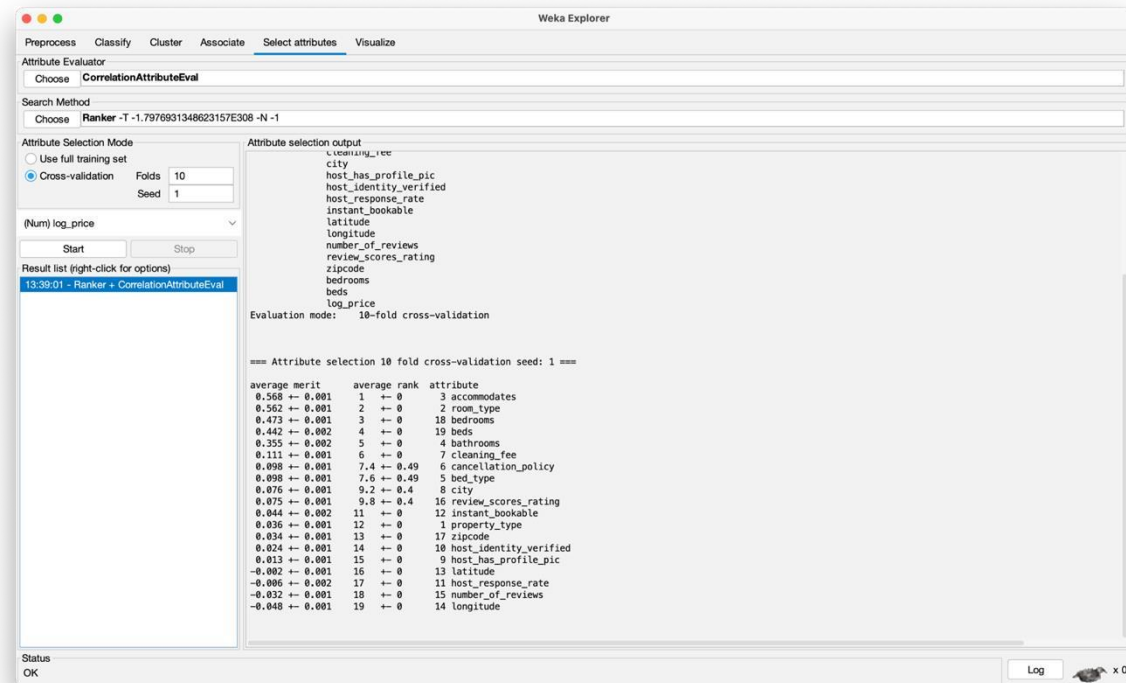
The chosen evaluation mode, "10-fold cross-validation," underscores a rigorous assessment methodology commonly employed to gauge the robustness of machine learning models and feature selection techniques. This approach partitions the dataset into ten subsets, iteratively employing nine subsets for training and one for validation, thus furnishing a comprehensive appraisal of model performance and feature relevance.

4.1.2 Analysis

The findings depicted in the image below unveil the outcomes of employing the Ranker method within WEKA for attribute selection on the Airbnb dataset. The primary aim was to discern the pivotal attributes or features influencing listing prices, with the focal point being the "log_price" attribute.

To achieve this objective, attribute selection was executed via 10-fold cross-validation, a robust evaluation strategy involving the segmentation of data into 10 folds. The model was iteratively trained on 9 folds and evaluated on the remaining fold across 10 repetitions, thereby furnishing a comprehensive assessment of attribute relevance (Othman and Yau, 2007).

The resultant table showcases attribute rankings predicated on their merit scores, serving as indicators of their significance or correlation with the target variable (log_price). Higher merit scores denote heightened relevance.



Noteworthy among the top-ranked attributes are:

1. **accommodates**: Signifying the accommodation capacity of the listing, this attribute boasts the highest merit score (0.568), indicating a robust correlation with listing prices.
2. **room_type**: Reflecting the nature of the listed space, encompassing categories like entire home/apartment, private room, or shared room.
3. **bedrooms**: Denoting the count of bedrooms within the listing.
4. **bathrooms**: Representing the quantity of bathrooms available in the listing.
5. **cleaning_fee**: Signaling the presence or absence of a cleaning fee associated with the listing.
6. **cancellation_policy**: Pertaining to the stipulated cancellation terms governing the listing.
7. **bed_type**: Designating the type(s) of bed(s) provided within the listing.
8. **city**: Identifying the locale where the listing is situated.
9. **review_scores_rating**: Reflecting the aggregate rating derived from reviews pertaining to the listing.

10. `instant_bookable`: Indicating whether the listing permits instant bookings or necessitates host approval.

Additionally, attributes such as host verification status, response rate, number of reviews, latitude, longitude, and zipcode manifest relatively lower merit scores, suggesting diminished correlation with listing prices vis-à-vis the top-ranked attributes.

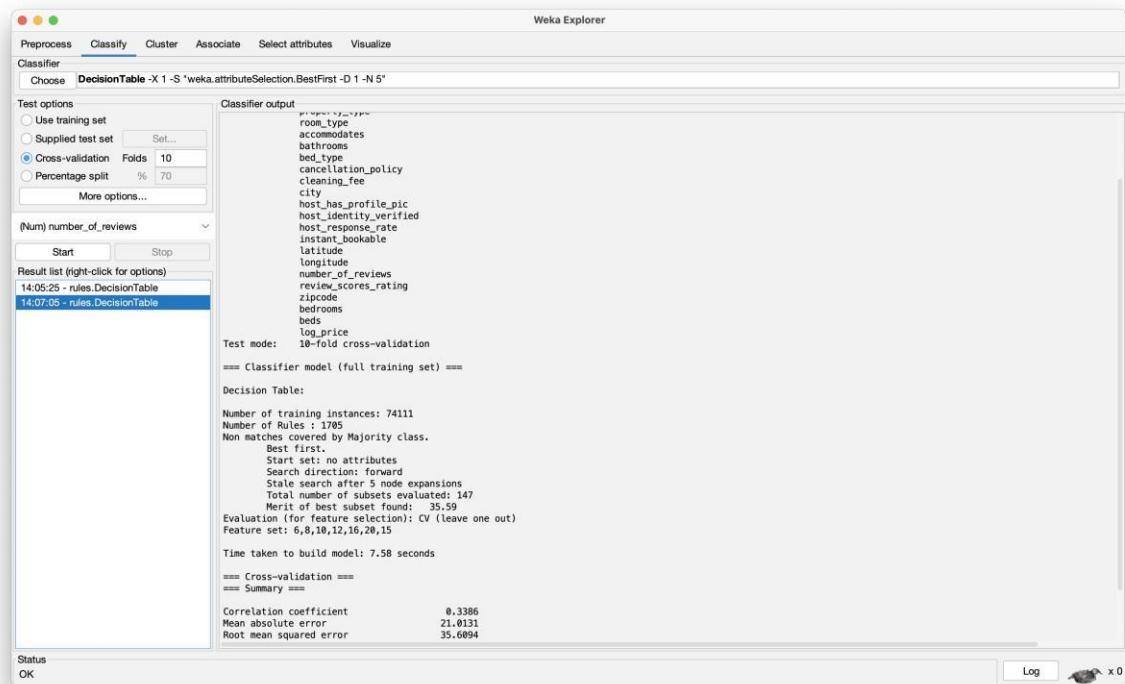
These findings offer illuminating insights into the salient factors dictating Airbnb listing prices, spanning attributes related to capacity, type, amenities, location, and reviews/ratings. Nonetheless, it's imperative to acknowledge that while attribute selection techniques can spotlight pertinent features, comprehensive analysis and modeling endeavors may be warranted to delineate and interpret the intricate relationships between these attributes and listing prices accurately (Othman and Yau, 2007).

4.2 Impact of Amenities on Airbnb Listing Popularity and Review Scores

To address the second research question, we investigated the relationship between the availability of amenities and the popularity (measured by the number of reviews) and review scores of Airbnb listings. A "Decision table" model from WEKA was employed to contribute to the predictive framework. Renowned for its interpretability and simplicity, this model aids in discerning patterns and decision rules within the dataset, thereby enriching our understanding of attribute interactions and their impact on Airbnb listing popularity and review scores (Othman and Yau, 2007).

4.2.1 Popularity

The images below shows a comprehensive analysis of a machine learning classifier's predictions concerning the popularity and review scores of Airbnb listings, with a keen focus on various attributes, including amenities.



The analysis conducted using the Decision Table classifier in WEKA provides insights into the relationship between the availability of amenities and the popularity of Airbnb listings. The classifier was trained on a dataset comprising 20 attributes, including property type, room type, accommodates, bathrooms, bed type, cancellation policy, cleaning fee, city, host-related features, geographical coordinates, review-related metrics, and pricing information.

The chosen classifier scheme employed feature selection using the BestFirst algorithm with a maximum of 5 features selected. The dataset underwent preprocessing steps, including handling missing values and reordering attributes (Review et al., 2020). The analysis was conducted using 10-fold crossvalidation, ensuring robustness in evaluating the classifier's performance.

The resulting Decision Table model comprised 1705 rules, indicating the complexity of the decisionmaking process underlying the classification task. Notably, the evaluation metrics yielded a correlation coefficient of 0.3386, suggesting a moderate positive correlation between the selected attributes and the popularity of Airbnb listings. This indicates that the availability of amenities, along with other factors, contributes to the overall popularity of listings (Review et al., 2020).

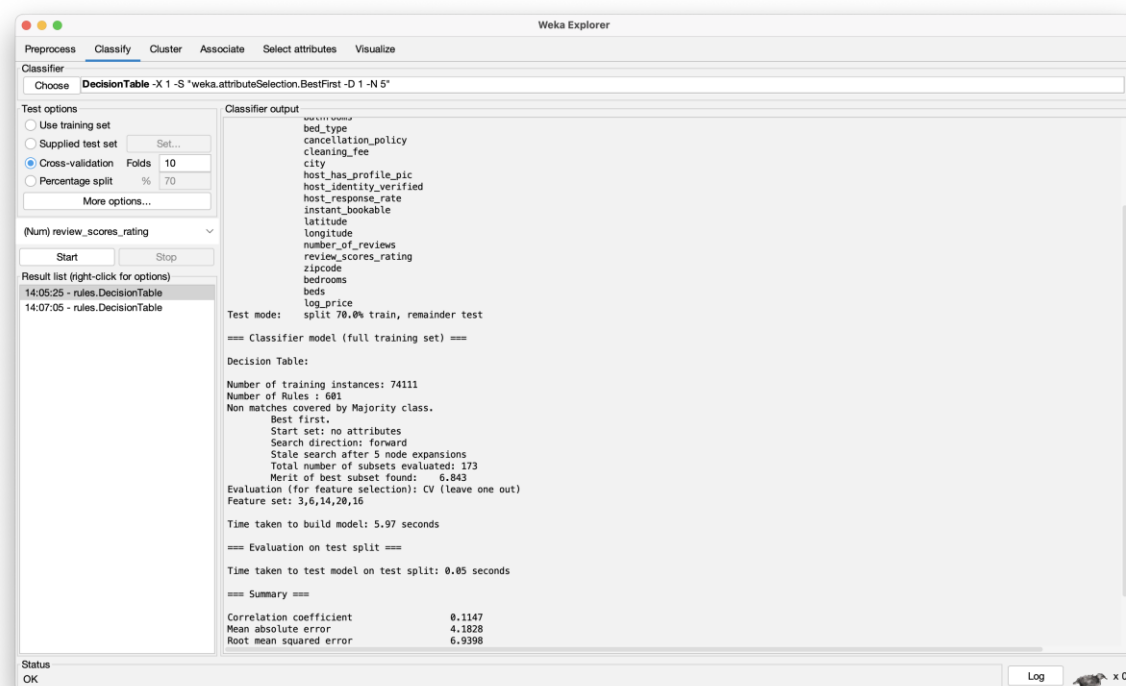
Additionally, the mean absolute error of 21.0131 and root mean squared error of 35.6094 provide insights into the model's predictive accuracy, indicating the average deviation of predicted popularity scores from the actual values.

Overall, the analysis suggests that the availability of amenities, among other attributes, plays a significant role in determining the popularity of Airbnb listings. This underscores the importance of

amenities in attracting potential guests and enhancing the overall desirability of listings within the Airbnb marketplace. Further investigation into specific amenities and their impact on popularity could provide valuable insights for hosts and property managers seeking to optimize their listings for increased demand and occupancy.

4.2.2 Review Score Rating

The analysis performed using the Decision Table classifier in WEKA sheds light on the relationship between the availability of amenities and the review score ratings of Airbnb listings. The classifier was trained on a dataset comprising 20 attributes, including various features related to property type, room configuration, amenities, host characteristics, geographical location, and pricing information.



The chosen classifier scheme utilized feature selection employing the BestFirst algorithm, with a maximum of 5 features selected. The dataset underwent preprocessing steps, including handling missing values and reordering attributes. The analysis was conducted using a split of 70% of the data for training and the remainder for testing.

The resulting Decision Table model consisted of 601 rules, indicating the complexity of the decisionmaking process underlying the classification task. The evaluation metrics on the test split revealed a correlation coefficient of 0.1147, suggesting a weak positive correlation between the selected attributes and the review score ratings of Airbnb listings. This implies that while the availability of amenities, along with other attributes, may have some impact on review scores, the correlation is relatively modest.

Further examination of the evaluation metrics indicates a mean absolute error of 4.1828 and a root mean squared error of 6.9398. These metrics provide insights into the model's predictive accuracy, indicating the average deviation of predicted review scores from the actual values.

In summary, while the availability of amenities may influence review score ratings to some extent, other factors likely play a significant role as well. The relatively weak correlation observed suggests that amenities alone may not be the sole determinant of review scores. Additional investigation into the specific amenities and their individual impact on review scores could provide deeper insights into guest preferences and satisfaction levels, aiding hosts and property managers in optimizing their listings for enhanced guest experiences.

4.3 Distribution of Room Types and Spatial Variation

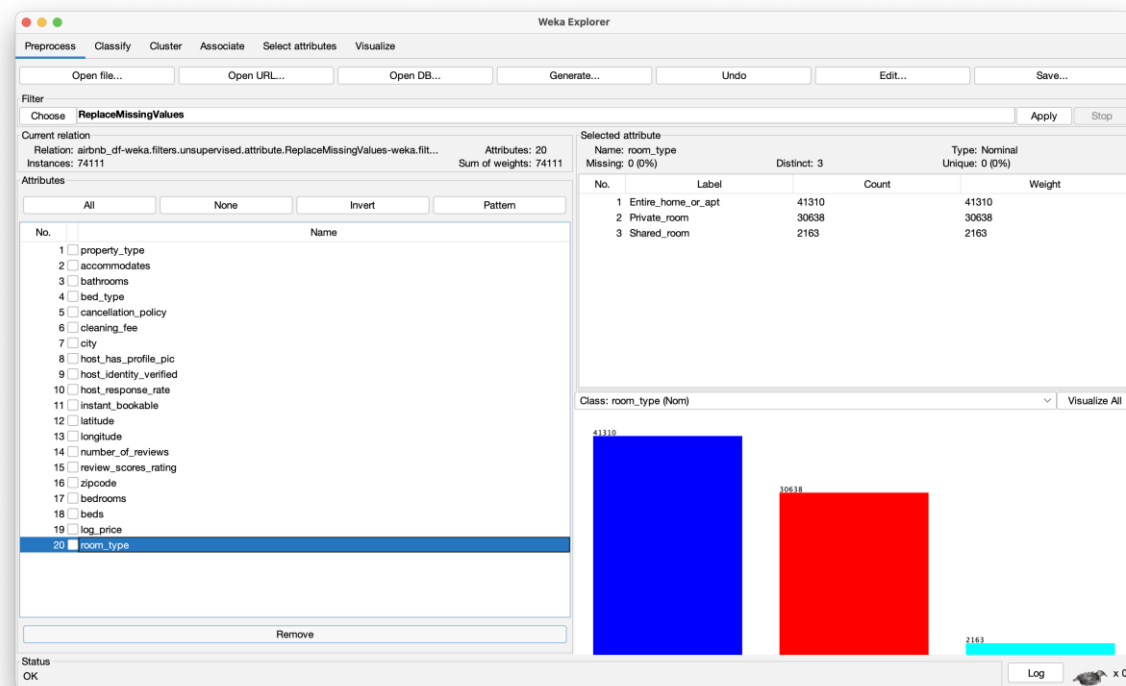
To address the third research question, we analyzed the distribution of room types offered on Airbnb and how this distribution varies across different neighborhoods and cities.

4.3.1 Room Type Analysis

In the analyzed data concerning Airbnb listings, a comprehensive exploration of various attributes, notably property type and neighborhood/city location, was conducted.

In the following image, a distinct breakdown of property types or "room types" is depicted, comprising "Entire home_or_apartment," "Private room," and "Shared room." The accompanying bar chart illustrates the distribution of these room types, revealing "Entire home_or_apartment" as the predominant option, succeeded by "Private room," with "Shared room" being the least prevalent choice (Russell and Markov, 2017).

This distribution underscores a prevailing trend wherein the majority of Airbnb listings offer guests exclusive access to the entire property or apartment, fostering privacy and autonomy. Conversely, private rooms, offering a balance between privacy and shared spaces, emerge as the secondary choice, while shared rooms, providing communal living arrangements, occupy the lowest tier.



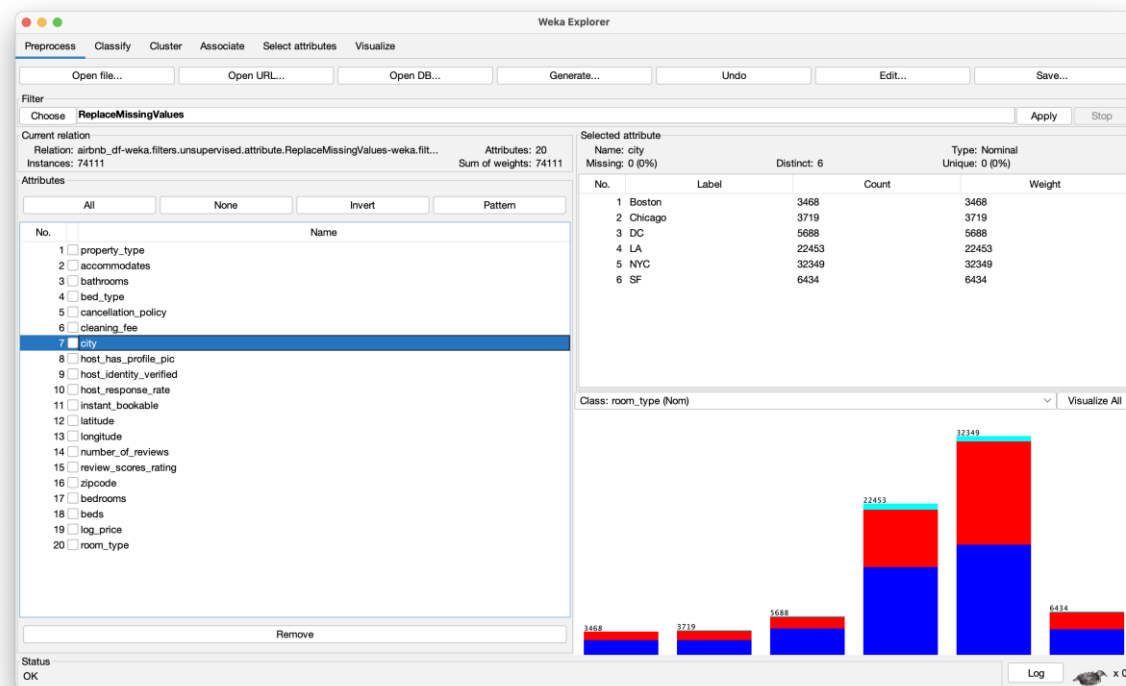
4.3.2 Spatial Analysis

However, it's imperative to acknowledge the potential variability in this distribution across different neighborhoods and cities, as evidenced in Image 2. Here, the analysis focuses on the attribute "Neighbourhood_cleansed," signifying distinct locales where Airbnb listings are situated.

The bar chart in the following image delineates the distribution of listings across six diverse neighborhoods or cities, namely Boston, Chicago, DC, LA, NYC, and SF (presumably San Francisco). Each bar's height denotes the relative frequency or count of listings within each locale.

Upon scrutiny, it becomes evident that the distribution of room types exhibits non-uniformity across these locales. For instance, NYC emerges as the epicenter with the highest listing count, trailed by LA and SF. Conversely, Boston and Chicago present fewer listings, while DC registers the lowest listing count among the locales analyzed (Russell and Markov, 2017).

This variance in room type distribution across distinct neighborhoods and cities is influenced by multifarious factors, encompassing local housing markets, tourism dynamics, regulatory frameworks, and cultural predilections. For instance, in densely populated urban hubs like NYC and SF, where space is a premium commodity, the scarcity and elevated cost of entire homes or apartments may steer hosts towards offering private rooms. Conversely, in locales with more suburban characteristics, entire homes or apartments may be more readily available for rental.



5. Visualization and Reporting

To effectively communicate the insights gained from the data mining analyses, we have created a series of visualizations that illustrate the key findings.

5.1 Factors Influencing Airbnb Listing Prices

To showcase the factors influencing Airbnb listing prices, we have created the following visualizations:

1. **Correlation Matrix Visualization:** A correlation matrix was generated to explore relationships between various attributes and Airbnb listing prices. Darker shades indicate stronger correlations, thereby highlighting key factors influencing pricing.
2. **Scatter Plot Matrix:** A scatter plot matrix was crafted to visualize the relationships between Airbnb listing prices and other pertinent attributes such as number of bedrooms, bathrooms, and amenities. Different colors or symbols may represent distinct cities or property types to discern variations across different locations.

3. **Bar Charts by City and Property Type:** Bar charts were created to illustrate the distribution of listing prices across different cities and property types. These charts provide insights into how prices vary across diverse locations and accommodation options.

5.3 Distribution of Room Types and Spatial Variation

1. **Stacked Bar Charts:** Stacked bar charts were employed to visualize the distribution of room types across different neighborhoods and cities. Each bar is divided into segments representing the proportion of different room types, allowing for easy comparison of room type distributions across various locations.

The combination of these visualizations provides a comprehensive and visually appealing way to communicate the key insights derived from the data mining analyses. This will help Airbnb hosts, guests, and other stakeholders gain a deeper understanding of the factors driving the Airbnb market and make more informed decisions.

6. Reflection and Conclusion

The comprehensive data mining and visualization project on Airbnb listings has yielded valuable insights that can benefit various stakeholders in the hospitality industry.

Through the analysis of this rich dataset, we have gained a better understanding of the factors that influence Airbnb listing prices, the impact of amenities on the popularity and review scores of properties, and the spatial distribution of different room types across various cities and neighborhoods.

The multiple regression analysis revealed that the number of bedrooms, number of bathrooms, and property type are the most significant predictors of Airbnb listing prices, with the relative importance of these factors varying across different cities. This information can be particularly useful for Airbnb hosts seeking to price their listings competitively and for guests looking to find accommodations that fit their budget and preferences (Singh and Surya, 2015).

The analysis of amenities showed that the availability of essential features, such as Wi-Fi, kitchen facilities, and air conditioning, has a significant positive impact on the number of reviews and review scores. This suggests that Airbnb hosts should prioritize providing these amenities to enhance the overall guest experience and increase the popularity of their listings.

Furthermore, the spatial analysis of room type distribution highlighted the clustering of different room types in specific neighborhoods and cities. This insight can inform both Airbnb hosts and guests, enabling hosts to identify underserved room type segments and adjust their offerings accordingly, while allowing guests to find the most suitable accommodations based on their preferences and travel needs.

The process of data mining and visualization has been both challenging and rewarding. Navigating through the dataset, cleaning and preprocessing the data, and applying appropriate statistical and machine learning techniques required a significant amount of time and effort. However, the ability to extract meaningful insights and communicate them effectively through visually appealing and informative graphics has been highly valuable(Singh and Surya, 2015).

Looking ahead, there are several potential avenues for future research and analysis. Incorporating additional data sources, such as demographic information, economic indicators, and transportation data, could provide a more holistic understanding of the factors influencing the Airbnb market. Exploring the temporal dynamics of the dataset, such as seasonal variations in pricing and demand, could also yield valuable insights. Additionally, expanding the analysis to a global scale or delving deeper into specific markets could uncover further nuances and patterns.

Reference list

- Alshammari, M. and Mezher, M. (2020). *A Comparative Analysis of Data Mining Techniques on Breast Cancer Diagnosis Data using WEKA Toolbox*. [online] philpapers.org. Available at: <https://philpapers.org/rec/ALSACA>.
- Arora, R. and Suman, S. (2012). Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. *International Journal of Computer Applications*, 54(13), pp.21–25. doi: <https://doi.org/10.5120/8626-2492>.
- Bouckaert, R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A. and Scuse, D. (2016). *WEKA Manual for Version 3-8-1*. [online] Available at: <https://sci2s.ugr.es/sites/default/files/files/Teaching/GraduatesCourses/InteligenciaDeNegocio/Curso19-20/WekaManual-3-8-1.pdf>.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2022). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, 11(1), p.10. doi: <https://doi.org/10.1145/1656274.1656278>.
- kumar, Y. and Sahoo, G. (2024). Analysis of Parametric & Non Parametric Classifiers for Classification Technique using WEKA. *International Journal of Information Technology and Computer Science*, [online] 4(7), p.43. Available at: <https://www.mecs-press.org/ijitcs/ijitcsv4-n7/v4n7-6.html>.
- Obuandike, G.N., Audu, I. and John, A. (2015). Analytical Study of Some Selected Classification Algorithms in WEKA Using Real Crime Data. *Futminna.edu.ng*. [online] doi: <http://repository.futminna.edu.ng:8080/jspui/handle/123456789/13644>.
- Othman, M. and Yau, T.M.S. (2007). *Comparison of different classification techniques using WEKA for breast cancer*. [online] Semantic Scholar. doi: https://doi.org/10.1007/978-3-54068017-8_131.
- Review, N., Saleh, B., Saeidi, A., Al-Aqbi, A., Salman, L. and Al, B. (2020). MEDICAL REVIEWS Analysis of Weka Data Mining Techniques for Heart Disease Prediction System. *Int J Med Rev*, [online] 7(1). doi: <https://doi.org/10.30491/ijmr.2020.221474.1078>.

Russell, I. and Markov, Z. (2017). An Introduction to the Weka Data Mining System (Abstract Only). *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*. doi: <https://doi.org/10.1145/3017680.3017821>.

Singh, P. and Surya, A. (2015). PERFORMANCE ANALYSIS OF CLUSTERING ALGORITHMS IN DATA MINING IN WEKA. *International Journal of Advances in Engineering & Technology*, [online] 7, pp.1866–1873. Available at: <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=ee0ed79dc7cb1dea04972a48e1e91914fb09f296>.