Enterprise Cloud Computing and Big Data (BUDT737)

Project Report

Section 0501 - Team 13

I. Project Title

Enhancing Airbnb Experience: Predicting Review Scores for Rental Properties

Team Members:

1. Bharath Sreekumar
2. Sneha Murali
3. Zeba Karkhanawala

Original Work Statement:

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

	Typed Name	**Signature**
1.	Bharath Sreekumar	B.S
2.	Sneha Murali	S.M
3.	Zeba Karkhanawala	Z.K

II. Executive Summary

In this project, we aim to leverage the Airbnb Listings dataset, obtained through the Open Datasoft API, to enhance the rental experience for users. By analyzing various features such as locality, price, amenities, and other factors, we seek to develop a predictive model that can accurately forecast the review scores given by users who have rented a property. This predictive model will enable hosts to better understand the factors influencing user satisfaction and make informed decisions to improve their rental properties, ultimately enhancing the overall Airbnb experience for both hosts and guests.

III. Data Description

The dataset titled Airbnb-listings has been obtained from the internet using Open Datasoft API (Updated till February 2024). The dataset has details related to Airbnb rental properties and has 85 columns. The data from the API has around 500k observations and after filtering it down we are using a subset of the dataset having 10,000 observations. The data has numerous numerical as well as categorical observations in the form of Strings or even Array Strings.

Here is a brief overview of the columns:

- access: Description of access to the property.
- accommodates: Number of guests the property can accommodate.
- amenities: List of amenities provided, such as TV, Internet, etc.
- availability_30, availability_60, availability_90, availability_365: Number of available nights in the next 30, 60, 90, and 365 days.
- bathrooms, bedrooms, beds: Number of bathrooms, bedrooms, and beds in the property.
- bed_type: Type of bed (e.g., Real Bed).
- calculated_host_listings_count: Number of listings the host has.
- cancellation_policy: Cancellation policy of the property.
- city, country, country_code: Location details.
- description: Description of the property.
- experiences_offered: Type of experiences offered by the host.
- extra_people: Additional cost for extra people.
- features: Features offered by the host.
- first_review, last_review: Dates of the first and last review.
- geolocation: Geographical coordinates of the property.
- guests_included: Number of guests included in the base price.
- host_about, host_acceptance_rate, host_id, host_listings_count, host_location, host_name, host_neighbourhood, host_picture_url, host_response_rate, host_response_time, host_since, host_thumbnail_url, host_total_listings_count, host_url, host_verifications: Host-related information.
- house_rules, interaction, jurisdiction_names: Rules and interaction instructions.

- last_scraped, license, listing_url, market, maximum_nights, medium_url, minimum_nights, monthly_price, name, neighborhood_overview, neighbourhood, neighbourhood_cleansed, neighbourhood_group_cleansed, notes: Various details and URLs related to the listing.
- number_of_reviews: Total number of reviews.
- price: Listing price per night.
- property_type: Type of property (e.g., Apartment).
- review_scores_accuracy, review_scores_checkin, review_scores_cleanliness, review_scores_communication, review_scores_location, review_scores_rating, review_scores_value: Scores given by guests in different aspects.
- reviews_per_month: Number of reviews per month.
- room_type: Type of room (e.g., Entire home/apt).
- scrape_id: ID of the data scrape.
- security_deposit: Security deposit amount.
- smart_location, space, square_feet, state, street, summary, thumbnail_url, transit, weekly_price, xl_picture_url, zipcode: More details about the property.

This dataset is interesting as it provides a comprehensive view of Airbnb listings, encompassing diverse details about properties, hosts, pricing, and availability. Analyzing this data could unveil insights into factors influencing rental trends, host behaviors, and the varying features that contribute to the popularity of listings in different locations analyses related to Airbnb listings, such as pricing trends, popularity, and host behavior.

Link to the data source

https://public.opendatasoft.com/explore/dataset/airbnb-listings/api/?flg=en-us&disjunctive.host_verifications&disjunctive.amenities&disjunctive.features&dataChart=eyJxdWVyaWVzIjpbeyJjaGFydHMiOlt7InR5cGUiOiJjb2x1bW4iLCJmdW5jIjoiQ09VTlQiLCJ5QXhpcyI6Imhvc3RfbGlzdGluZ3NfY291bnQiLCJzY2llbnRpZmljRGlzcGxheSI6dHJ1ZSwiY29sb3IiOiJyYW5nZS1jdXN0b20ifV0sInhBeGlzIjoiY2l0eSIsIm1heHBvaW50cyI6IiIsInRpbWVzY2FsZSI6IiIsInNvcnQiOiIiLCJzZXJpZXNCcmVha2Rvd24iOiJyb29tX3R5cGUiLCJjb25maWciOnsiZGF0YXNldCI6ImFpcmJuYi1saXN0aW5ncyIsIm9wdGlvbnMiOnsiZmxxIjoiZW4tdXMiLCJkaXNqdW5jdGl2ZS5ob3N0X3ZlcmlmaWNhdGlvbnMiOnRydWUsImRpc2p1bmN0aXZlLmFtZW5pdGllcyI6dHJ1ZSwiZGlzanVuY3RpdmUuZmVhdHVyZXMiOnRydWV9fX1dLCJ0aW1lc2NhbGUiOiIiLCJkaXNwbGF5TGVnZW5kIjp0cnVlLCJhbGlnbk1vbnRoIjp0cnVlfQ%3D%3D

IV. Research Questions

From our analytical knowledge, we aim to utilize this extensive dataset to answer the following questions to the best of our capabilities:

1. Do price and location have an impact on the listing of properties on Airbnb?
2. Does the larger number of reviews and review scores have anything to do with the property's security deposit and designated price?

3. Does the property owner list multiple properties in a preferred locality?
4. What is the impact of characteristics such as Property Type, Room Type, Cancellation Policy, Host Response Rate, Host Listings Count, Accommodates, Bathrooms, Bedrooms, Beds, and Price on the final review score?
5. How can we predict the pricing category (high or low) of Airbnb listings based on their characteristics and what implications does this have for hosts aiming to competitively price their listings?

V. Methodology

- **Kafka Producer:** This component is responsible for fetching data from the opendatasoft Airbnb-Listings dataset. The producer sends data in chunks (messages) to Kafka topics.
- **Kafka Consumer:** This component subscribes to Kafka topics, receives the messages produced by the Kafka producer, and processes them. In our case, it is the initial step for streaming the data.
- **MongoDB:** The Kafka Consumer reads messages from Kafka topics and writes them into MongoDB. This step is often used for real-time data processing and storage. Mongo is a NoSQL database that stores data in BSON (binary JSON) format. It's suitable for handling large volumes of unstructured or semi-structured data, making it a good fit for our diverse Airbnb property information.
- **Graphframes:** GraphFrames can be used for easy visualizations, providing a clear representation of the network of hosts and neighborhoods. Visualization aids in better understanding the structure of the data, making it easier to communicate findings and insights to stakeholders. Geomaps also aid in a better and clear representation of the data.
- **Linear regression:** Linear regression allows for the identification of feature importance. By examining the coefficients, one can assess which features contribute more or less to the prediction of the Review Score. This information can guide feature selection and model refinement. If the underlying assumptions are met, it can provide accurate predictions, especially if there are linear patterns in the data.
- **Logistic Regression:** Logistic Regression is apt for classifying Airbnb properties into 'high' and 'low' price categories based on features like Accommodates, Bathrooms, Bedrooms, etc. The binary categorization is determined by comparing listing prices to the mean. With an 80-20 train-test split, the model's interpretability and simplicity make it suitable for this task, providing insights into the factors influencing price categories. Evaluation metrics such as accuracy and precision can gauge the model's performance on unseen data.
- **K-Means Clustering (with Elbow Method):** K-Means Clustering was utilized to categorize Airbnb properties based on diverse features such as Country, Price, Review Scores Value, Security Deposit, and Number of Reviews. The application of the Elbow Method to determine the optimal number of clusters revealed a k value of 4, allowing for a meaningful segmentation of properties. This technique is valuable for identifying inherent patterns and similarities within the dataset, providing a structured approach to

grouping properties, and gaining insights into potential underlying trends or distinctions among Airbnb listings.

VI. Results and Finding

- The linear regression model produced an **RMSE of 0.83** and an **MAE of 0.57**, indicating that on average, the model's predictions of review scores deviate from the actual scores by less than one point on a 10-point scale. This performance suggests a reasonably accurate fit of the model to the data, capturing the essential trends and relationships between the predictors and the outcome variable, which in this case, is the review score.

review_scores_value	prediction
9	9.331778297518312
9	9.149975241143668
9	9.144292468879089
9	9.37180917259551
9	9.271245407043423
9	9.217965378546907
10	9.049471516805285
9	8.852775661561127
9	8.978817680876091
9	8.857756333804916
9	8.831544966396878
9	8.840923110768086
10	8.90723106356027
9	8.82046962046358
9	8.980244775029448
10	8.924014044722611
4	8.934993451010683
9	8.856737027452724
9	9.004280439258602
10	8.964430858344812

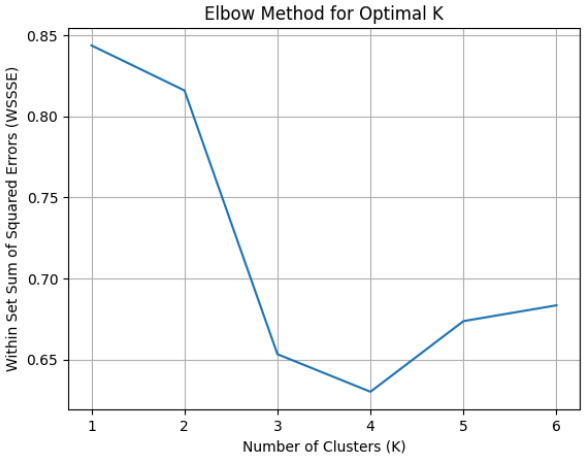
This study's linear regression analysis provides insightful findings on factors influencing Airbnb review scores. The quantified model performance (RMSE and MAE) suggests that the selected predictors effectively capture the variability in review scores.

- The logistic regression model used for classification achieved an **accuracy of 0.80**, indicating that 80% of the time, the model correctly predicts the pricing category of an Airbnb listing as either high or low based on the training data provided. This high level of accuracy suggests that the features selected for the model—such as accommodates, bathrooms, bedrooms, beds, security deposit, and others—are good predictors of a listing's pricing category.

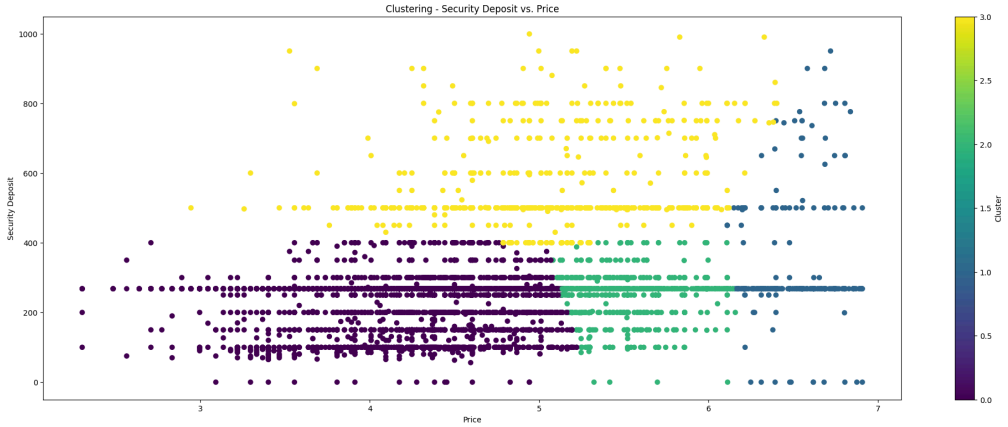
price_category	new_price_category	probability	prediction
high	1.0	[0.46425283853764...]	1.0
high	1.0	[0.17576475029292...]	1.0
low	0.0	[0.81517805444616...]	0.0
high	1.0	[0.12478739722453...]	1.0
low	0.0	[0.58449883785313...]	0.0
low	0.0	[0.48084715041049...]	1.0
low	0.0	[0.68267013735722...]	0.0
high	1.0	[0.40630642563511...]	1.0
low	0.0	[0.74532263260747...]	0.0
low	0.0	[0.84821445644159...]	0.0
low	0.0	[0.87956994655389...]	0.0
low	0.0	[0.69121904096245...]	0.0
low	0.0	[0.90743028441874...]	0.0
low	0.0	[0.84573095531833...]	0.0
high	1.0	[0.02717703458870...]	1.0
high	1.0	[0.09600742258766...]	1.0
low	0.0	[0.90582300054999...]	0.0
high	1.0	[0.00334889633477...]	1.0
low	0.0	[0.78780965907859...]	0.0
low	0.0	[0.64423748883250...]	0.0

This analysis, through logistic regression, demonstrates that it is possible to accurately predict the pricing category of Airbnb listings based on a set of descriptive features. For hosts, this insight offers a valuable tool for strategically pricing their listings to align with market expectations while maximizing their competitive edge.

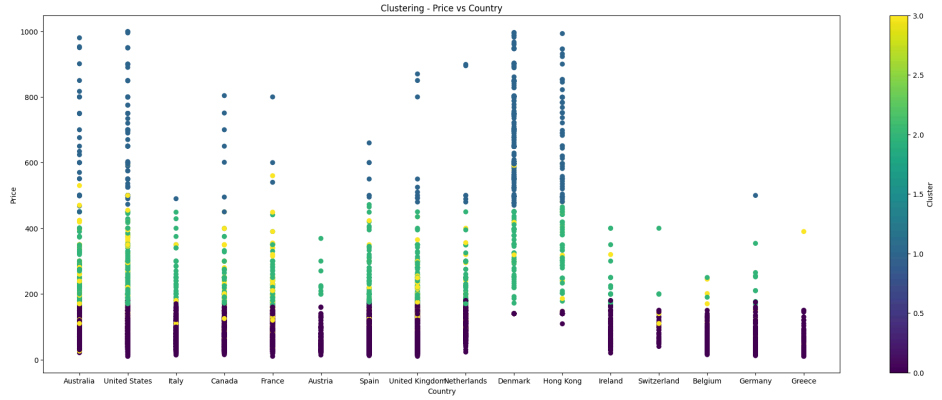
- Before diving directly into clustering and training the model, we used the Elbow method to determine the optimal value of K which we tested across the values 2 to 7. It works by plotting the within-cluster sum of squares (WCSS) for different values of k, and then selecting the value of k where the rate of decrease in WCSS slows down abruptly, creating an "elbow" shape in the plot. The optimal value of k that came out from this analysis was **4.**



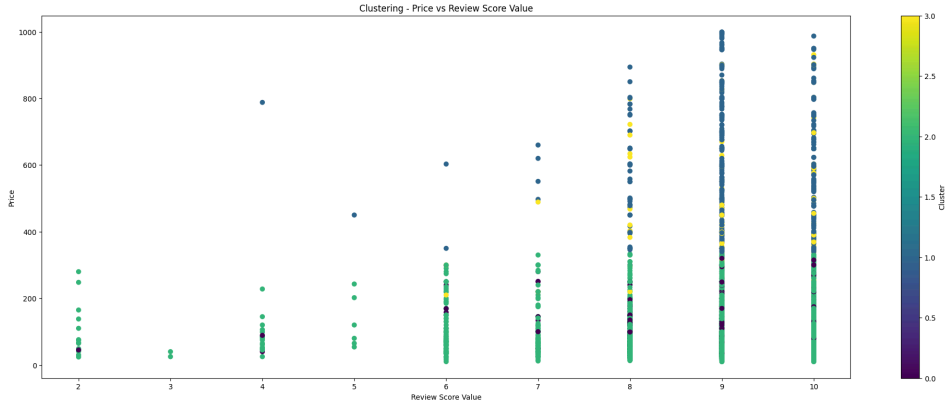
- Using K-means clustering, we were able to cluster the properties into neat clusters to observe the underlying patterns. As seen below, when we plotted the clusters between Security Deposit and Price, we found that there is a clear market segmentation of properties such as one cluster that might represent luxury properties with high prices and high-security deposits. In contrast, another cluster might represent budget properties with lower prices and lower security deposits.



- We also plotted the made clusters between Price and Country and found out that even within countries, we can see that there are subgroups or segments of properties that differ in terms of their pricing. This could represent different tiers of properties within each country.



- When we plotted the clusters between Price and Review Score values, we noticed the pattern that for properties with higher review scores, the price of properties ran from very low to very high. And for properties with lower review scores, the price of properties was generally low.



VII. Conclusion

Our analysis of the Airbnb rental landscape yields actionable insights beneficial to both businesses and hosts. By accurately predicting review scores, hosts gain valuable feedback to enhance their properties and provide a more satisfying experience for guests. Positive reviews not only improve reputation but also foster customer loyalty, ultimately driving higher occupancy rates and revenue. Additionally, our classification model for pricing categories empowers hosts to competitively price their listings based on key property attributes, maximizing revenue potential and market competitiveness.

Furthermore, our clustering analysis identifies distinct property segments, enabling businesses to tailor marketing and pricing strategies to specific customer segments. Targeted campaigns and promotions resonate more effectively with different audiences, driving higher booking rates and customer engagement. Overall, our project equips hosts and businesses with data-driven tools and insights to optimize pricing strategies, improve property quality, and drive greater success in the dynamic Airbnb marketplace.

For future scope, we plan on refining all our models to improve their performance. We are also planning on creating a Recommendation System which based on a few fields, can recommend you an Airbnb Listing that can help Airbnb users in selecting a property. We are also planning on creating another Linear Regression model that can predict the rough price estimate based on different fields to help Hosts estimate the price to offer for their respective properties.