Sentiment Analysis of Boston's Airbnb Reviews

Group 11: Yifan Fan, Jinke Han, Xiang Liu, Qianyi Mo

Data Preparation :

Pre-processing: We started by uploading two CSV files, reviews and listings, and selected the necessary columns. After merging them on 'listing_id' and removing any missing values, we formatted the price from a string to a float and the date to datetime format.

Next, we cleaned up the dataset further by using the 'langdetect' function to filter out non-English comments, which were less than four percent. We then processed the reviews, making all text lowercase, stripping punctuation, special characters, and irrelevant numbers, and normalizing spaces using Regex. After tokenizing and vectorizing the text to improve data quality, we created a word cloud to visualize our findings.

EDA: In our exploratory data analysis, we checked how different variables relate to each other and found no strong correlations. We also looked at a paired plot to have a look whether there have significant trends or not.We then plotted the number of reviews against years and noticed a sharp decline in 2020, mainly due to the pandemic's impact on travel and hospitality. However, as the situation improved, the number of reviews began to recover, showing a gradual return to normal for the Airbnb market.We also analyzed hotel review scores, which mostly were high, indicating that guests generally had good experiences. Low scores were uncommon, suggesting that most accommodations met or exceeded expectations.

Additionally, our analysis showed that "entire house/apartment" listings were the most popular on Airbnb, far outpacing "private rooms" and even more so "shared rooms" and "hotel rooms," reflecting a clear preference among users for more private accommodations.

Finally, we used a word cloud to highlight frequent terms in the reviews, with positive words like "great" and "clean" and mentions of "Boston" being prominent. This suggests that guests' experiences in Boston were largely positive, setting a positive tone for our upcoming topic analysis and model evaluation.

Analysis Plan:

1. Topic Analysis: We conduct topic analysis on text data using the gensim library to extract key topics from user reviews, such as service quality, cleanliness, and location, to identify the core aspects that users care about. Initially, we create a dictionary mapping each unique word in our dataset to a specific index, and then build a 'corpus' where each word in our texts is represented by its dictionary index. We then train a LDA model on this corpus to discover various topics within the text. We specifically aim to analyze topics numbered 3, 6, 9, 12, and 15, extracting the top five words that define each of these topics to quickly grasp the main themes in our data. For model preparation, we employ the train_test_split function to divide our data into an 80% training set and a 20% testing set. We'll also define a function to print the top five words for each topic in the LDA model.

2. Model Evaluation: During model evaluation, we train models with varying numbers of topics and measure their coherence and perplexity to identify the optimal model configuration. We observe that perplexity decreases with more topics, while coherence peaks at three topics. However, coherence significantly drops beyond 15 topics, indicating a loss in topic distinction. Thus, we conclude that 15 topics offer an optimal balance of lower perplexity and maintained coherence, although we also need to qualitatively assess the topics to ensure their relevance and distinctiveness.

Preliminary Results:

Evaluating the Impact of Alpha and Eta in LDA models: After building and evaluating our LDA model. We use alpha and eta values which refer to the strength of LDA stems to see whether our findings are good enough. Alpha influences the density of topics per document, while eta controls keyword density within topics. Comparing coherence scores across three models gives us an idea about the best set of alpha and eta that generates coherent topics.The higher the coherence scores, the more semantically coherent and interpretable the model's generated topics are. Model 1 with symmetric alpha, Model 2 with asymmetric alpha and fixed eta, and Model 3 with auto-adjusting alpha and eta. These are respectively represented by 0.567, 0.508, and 0.590 for Models 1, 2, and 3. The optimization of two hyper-parameters: alpha and eta was addressed through a sensitivity analysis.This involved systematically varying these parameters within a plausible range and observing how the changes affected coherency scores.The aim is to detect which combination of alpha and eta would maximize the topic coherence score.This process optimizes LDA models so as to correctly capture data's thematic structure

Qualitative insights were gathered from the word clouds generated by each LDA model in addition to quantitative analysis. Thematic focus on aspects like location, comfort, cleanliness and guest service could be easily understood through these word clouds. Deeper understanding of how changes in alpha and eta parameters can affect the nature and clarity of topics derived from LDA models was provided by comparing these word clouds. This emphasizes the importance of proper selection of alpha and eta parameters in LDA models. The results indicate that an auto-adjusting approach to these parameters such as Model 3 might produce more cohesive topics which can be interpreted better. There is further scope for future works to include expanded sensitivity analysis for better understanding of these parameters. Moreover, a more exhaustive qualitative analysis of the word clouds would supplement the quantitative findings and help improve LDA models for improved topic modeling tasks.

Next Steps:

We've successfully developed a well-performing LDA model that has helped us understand the overall experience of guests using Airbnb in the Boston area from 2019 to 2023. We've also created attractive word cloud visuals to make our findings easier to see and understand. Next, we plan to dive deeper by examining different areas in Boston. We'll categorize our analysis by these areas to explore how guests feel about their stays in various parts of the city.

Based on this, we'll offer suggestions for improvements to both Airbnb hosts and management in the Boston area, aiming to enhance the development of the local Airbnb market and ensure guests have a great experience. If time allows, we'll factor in pricing along with other variables like location coordinates and room types to group Boston's Airbnb listings into clusters, using clustering methods like K-means. This will help hosts and Airbnb to come up with more reasonable pricing strategies, potentially increasing the rate at which guests book stays.

Coding Contribution:

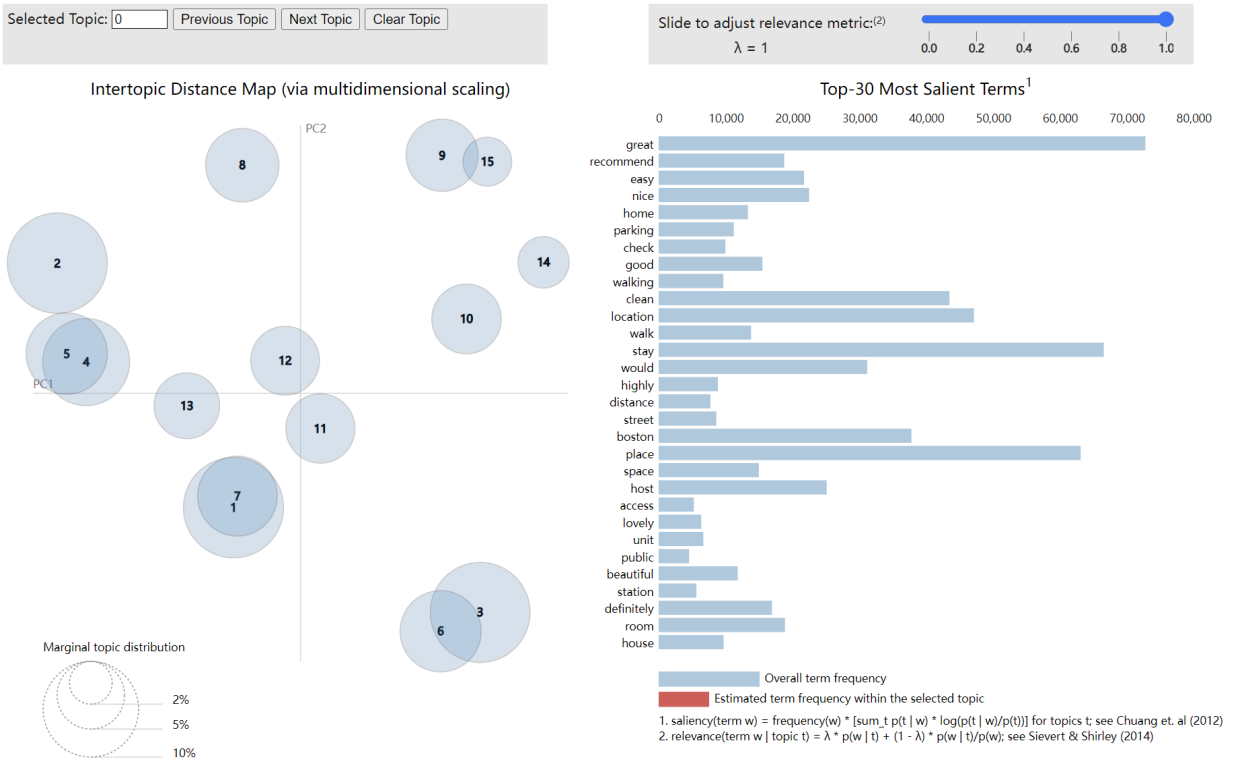
https://github.com/jinke7678/BA820_Project_Team11/blob/b609a2b151f33e234d1ed3083dfdc11289179b99/README.md

GitHub Project Link:https://github.com/jinke7678/BA820_Project_Team11.git

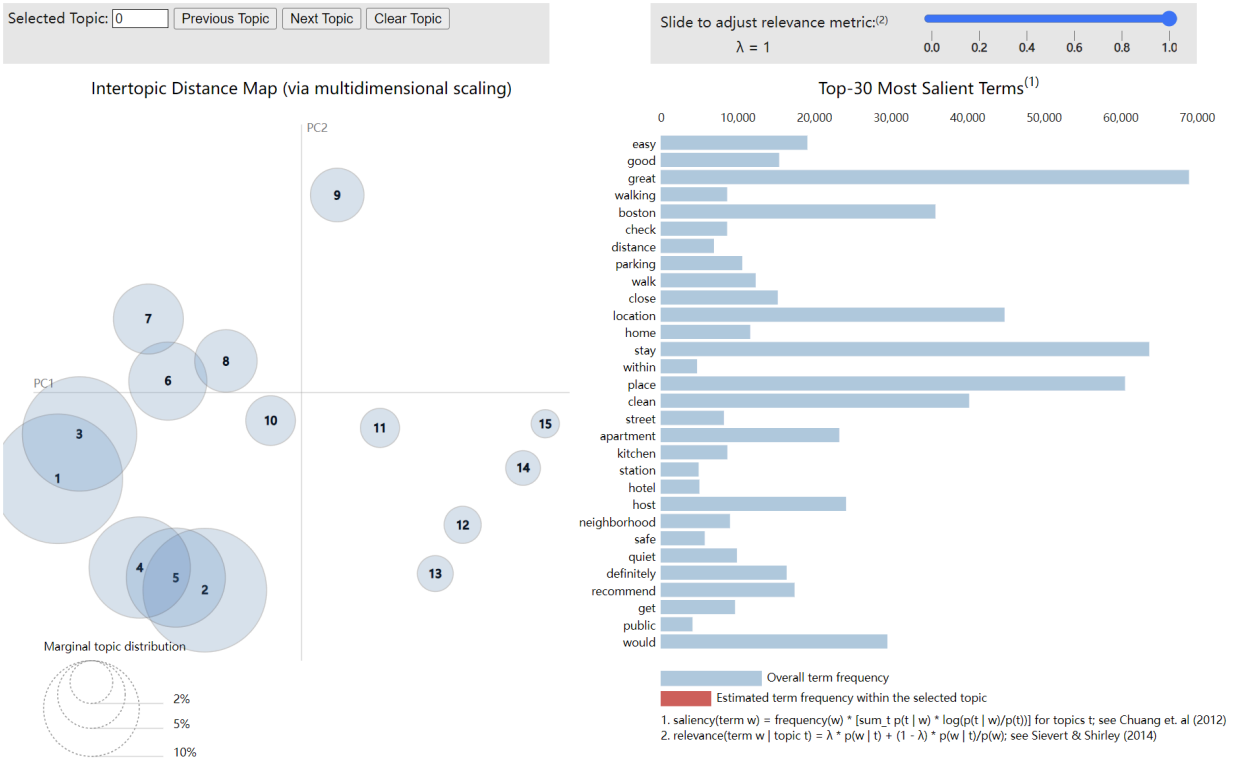
Use of AI: We used AI for grammar checking and content repetition to improve readability for this document.

Appendix

Model 1 results



Model 2 results



Model 3 results

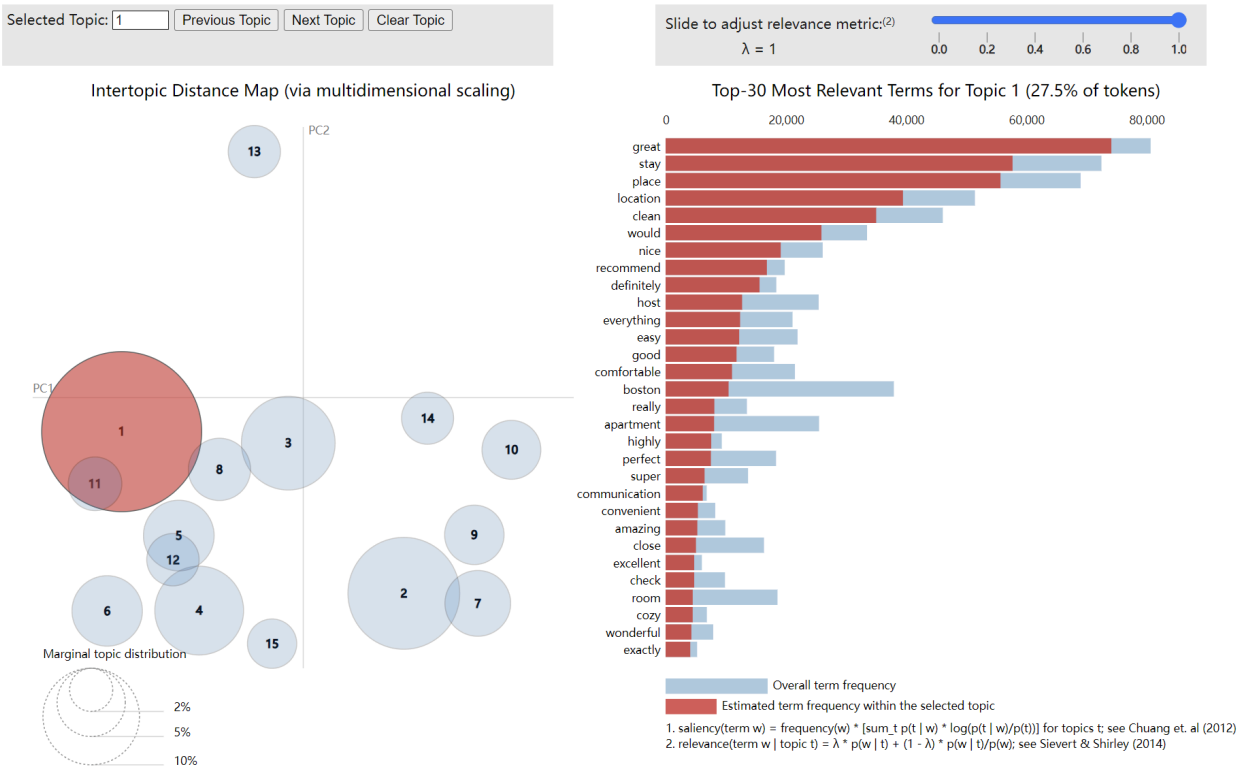


Figure 1

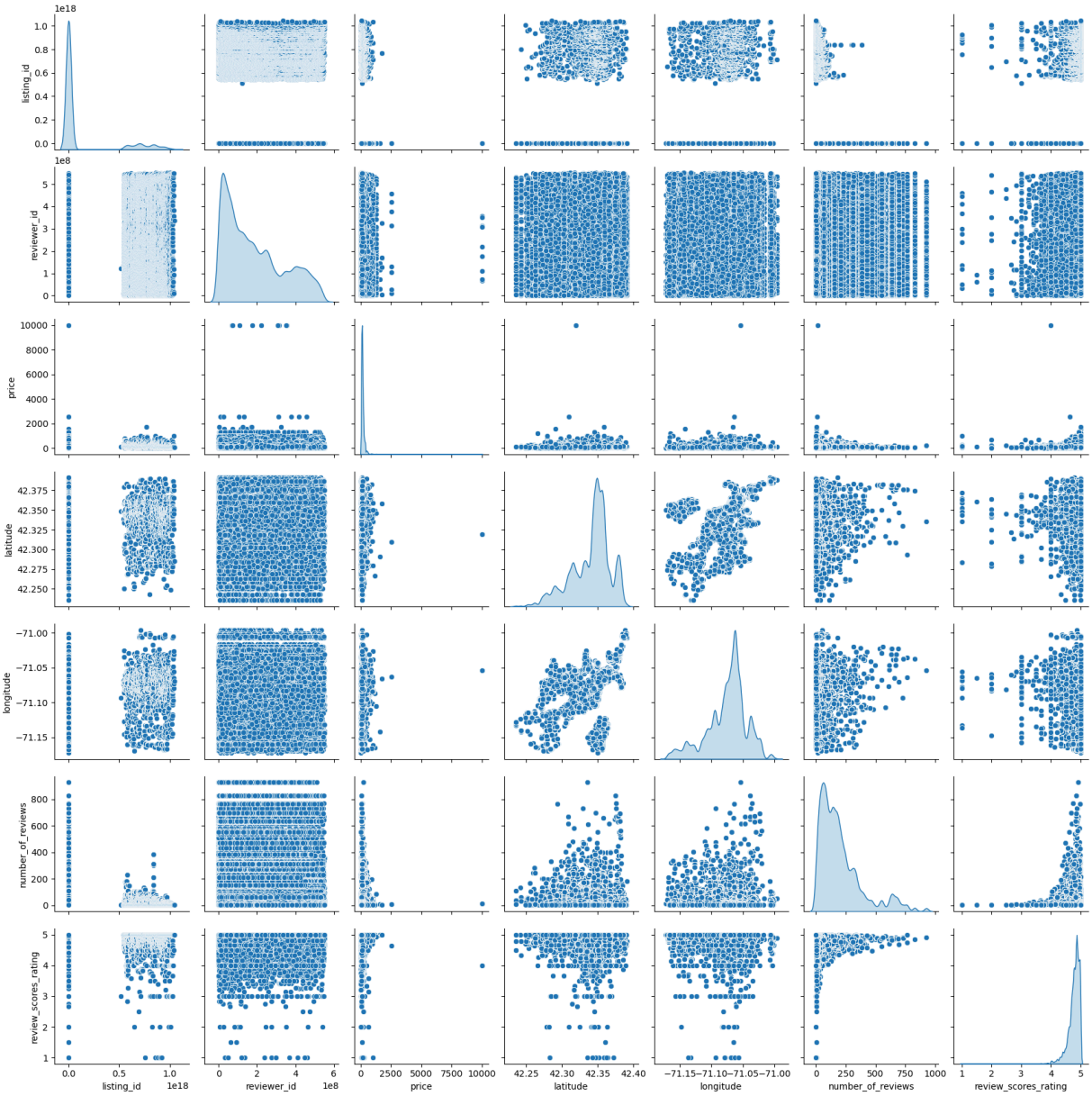

listing_id
reviewer_id
price
latitude
longitude
number_of_reviews
review_scores_rating
1e18
1e8

Figure 2

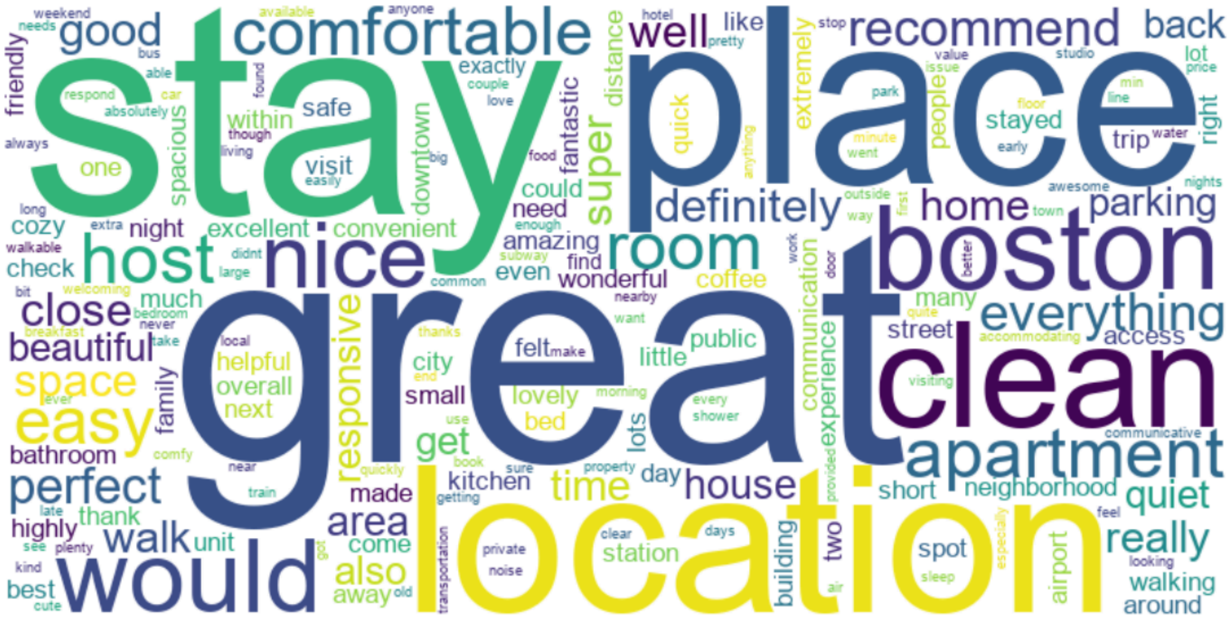


Figure 3

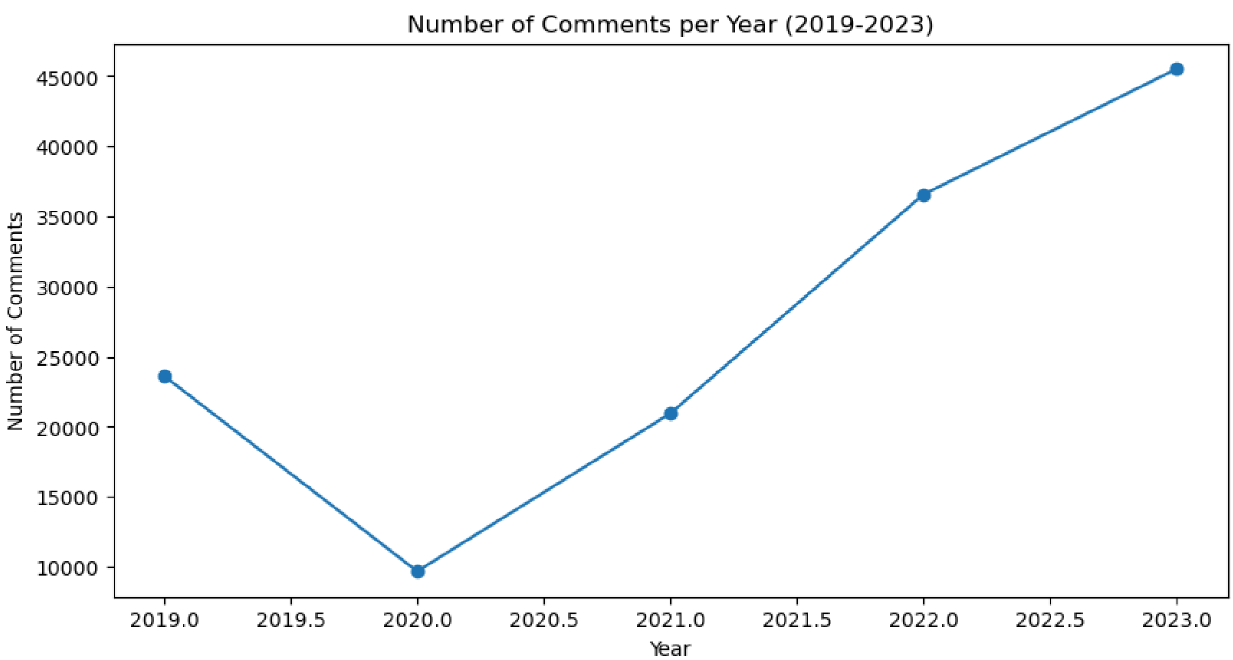


Figure 4

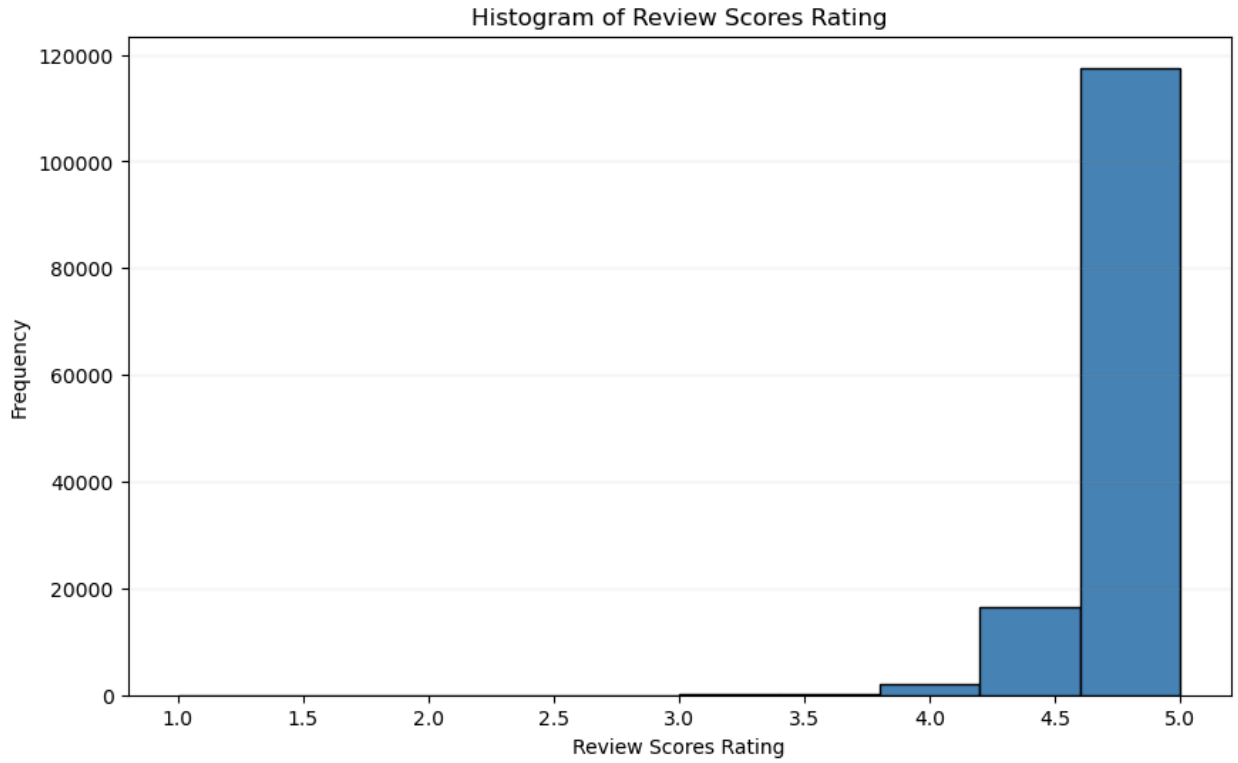


Figure 5