**Sentiment Analysis of Boston's Airbnb Reviews**

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

**Executive Summary:**

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.
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.

**Data Preparation :**

1.Pre-processing and Text Mining: 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.
2.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 Process:**

1.T-SNE: T-SNE is mainly used to display high-dimensional figures in a lower-dimensional format. In this case, we are displaying words, which have a dimension of 200, in a 2-D figure.
2.Sentiment Analysis: Our sentiment analysis workflow begins by using the VADER tool from NLTK library, which is good at measuring the sentiment expressed in text. We calculate a composite sentiment score for Airbnb reviews, classifying them as positive, neutral, or negative based on a preset threshold of 0.1. Our dataset is heavily biased toward positive sentiment, as shown in the histogram and confirmed by the word cloud, which highlights popular terms in positive reviews, indicating that Airbnb users had a good experience. Furthermore, we utilize logistic regression models trained on Word2Vec embeddings to classify sentiments. This rigorous

process not only quantifies sentiment distribution but also double-checks the classification accuracy of our models, ensuring a comprehensive and trustworthy sentiment analysis.
3. Topic Analysis using LDA: 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.
4. Model Evaluation: We train models with varying numbers of topics and measure their coherence and perplexity to identify the optimal model. 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.

**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 0.574,0.5256 and 0.59833 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.

**Conclusions:**
The comprehensive analysis of Boston Airbnb reviews provides us with valuable insights into guest preferences, highlighting key factors contributing to their satisfaction and pinpointing opportunities for hosts to improve their service. Our findings underscore the importance of cleanliness, quietness, and hospitality as the cornerstones of a positive guest experience. These elements not only enhance satisfaction but also encourage repeat visits, contributing to the growth and reputation of Boston's Airbnb market.
Based on our analysis, we strongly advise Airbnb hosts to focus on maintaining impeccable standards of cleanliness, actively manage noise levels, and create a welcoming and friendly atmosphere. Ensuring that the property descriptions accurately reflect the reality can help manage guest expectations and prevent potential disappointments.
To further elevate the guest experience, hosts could consider personal touches that make a stay memorable, such as offering local recommendations or small amenities. Consistently delivering on these aspects can lead to better reviews, higher ratings.

**Challenges**:
1.Data Cleaning and Preprocessing Challenges: A major obstacle was ensuring the cleanliness and readiness of data for analysis. The reviews are important to sentiment analysis. They were fraught with linguistic challenges such as slang, spelling errors, and diverse language usage. This complexity made the extraction of accurate sentiments from the text a complicated task.
2.Navigating Sentiment Ambiguities: Sentiment analysis was further complicated by the presence of mixed emotions in reviews. Guests often mentioned both positive and negative aspects in one review, presenting a challenge in accurately capturing the overarching sentiment.
3.Interactive Mapping Limitation: Our original plan included creating a geographical interactive map. However, the sheer volume of data made this task time-consuming and unmanageable within our time constraints. Consequently, we had to forgo this feature.
4.Sentiment Analysis Vocabulary Issue: In our sentiment analysis, we initially relied on a GitHub-sourced list of positive and negative words. However, this approach was flawed as we discovered inaccuracies. Such as negative words intermixed within the positive list. A more reliable solution was found in using nltk's in-built vocabulary classification function, which offered a more accurate and systematic approach to sentiment analysis.
5.Strategic Methodological Revisions: Initially, our project included association and clustering analyses. However, these methods proved impractical due to computational constraints and the sparsity of our large dataset. Our large dataset made the association analysis computationally intensive, and the results did not yield statistically significant insights. Clustering analysis was hindered by missing data and outliers. To navigate these issues, we focused on more feasible and insightful methods. We enhanced our data preprocessing and embarked on exploratory data analysis. Our efforts then centered around sentiment analysis and topic modeling using LDA, which aligned better with our research objectives and provided clear, actionable insights.1.

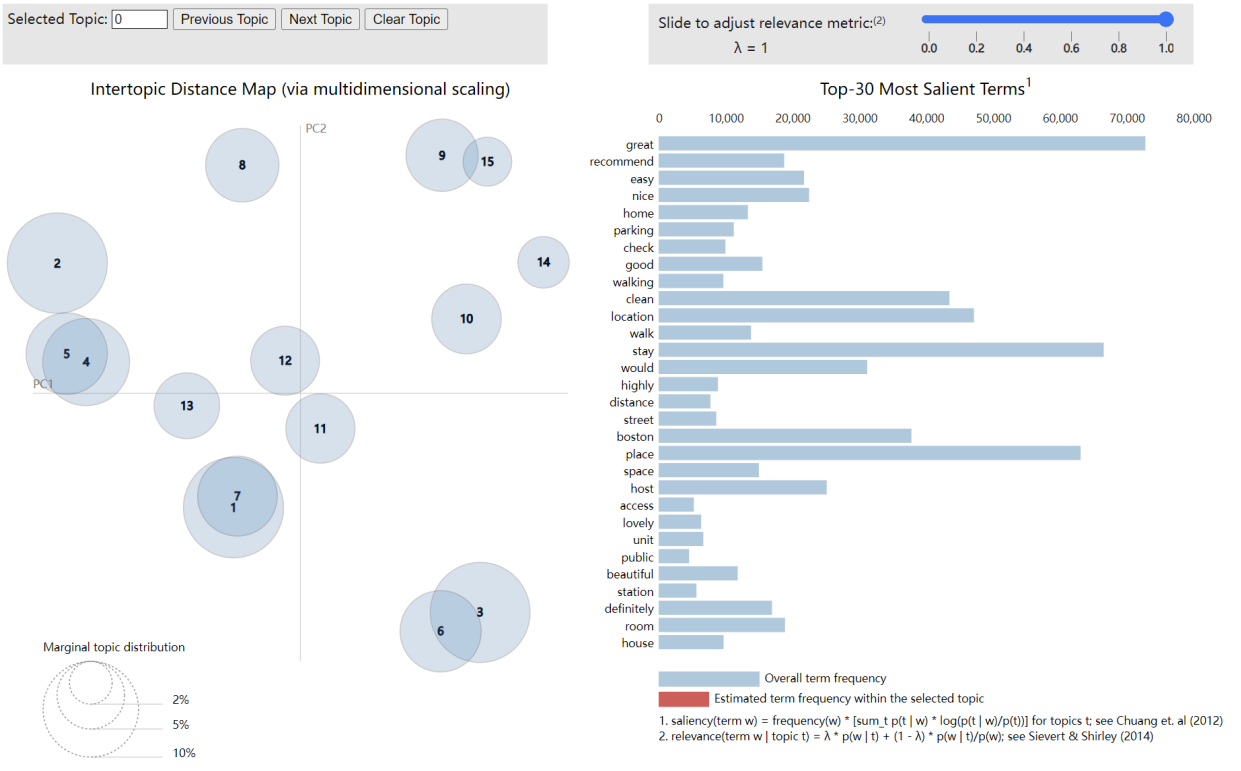
**Coding Contribution:**
https://github.com/jinke7678/BA820_Project_Team11/blob/6a6c6f5767a8efe8b0bb30ed073786f5cca9ec6d/Project%20Contribution.md
**GitHub Project Link:**https://github.com/jinke7678/BA820_Project_Team11.git
**Use of AI:** We used AI for grammar checking and content repetition to improve readability for this document. Also, used AI to explore further machine learning tools to help us analyze data.

**Coding Contribution**

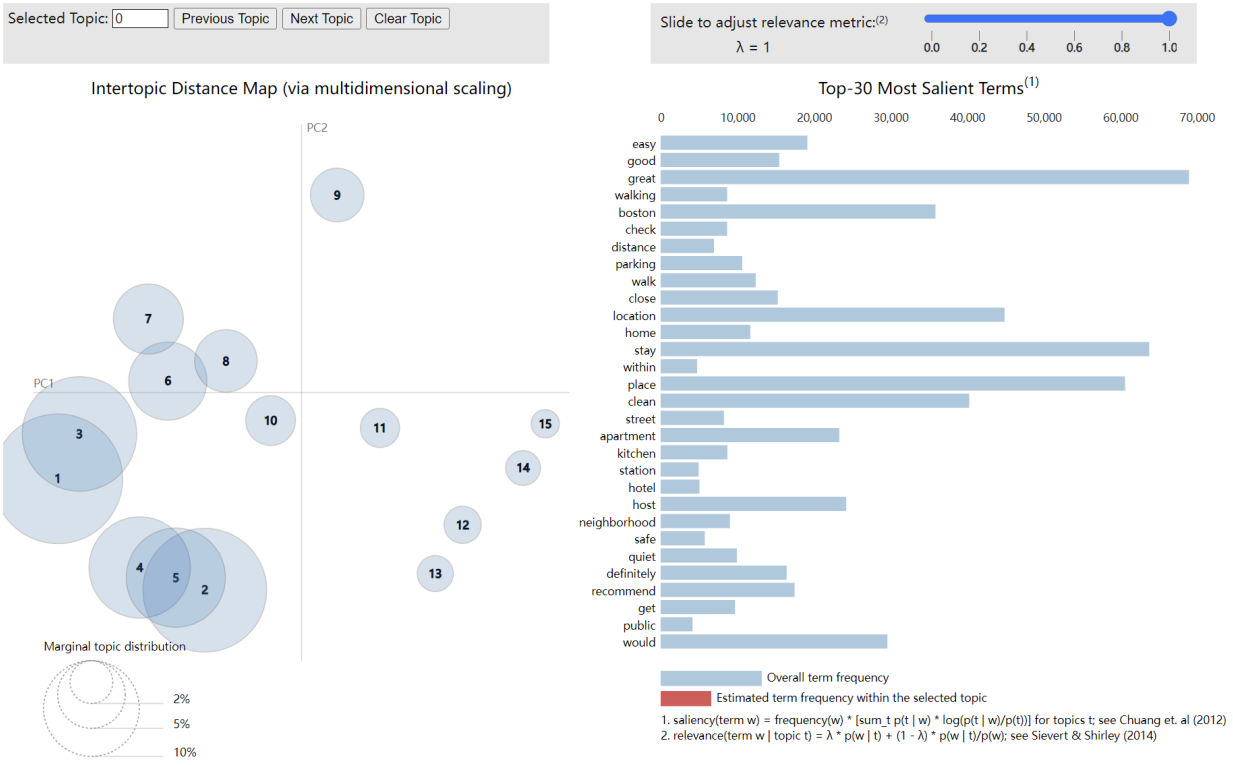
| Name | Workload | Description |
|---|---|---|
| Xiang Liu | 20% | Data cleaning and EDA part with some pictures. Made some simple graphs in some parts. Xiang undertook the critical task of cleaning and preparing the data for analysis, ensuring that our data sets were free of inconsistencies and ready for in-depth inspection. In addition to data cleaning, she began an EDA phase, where created insightful charts and performed analyzes to reveal trends and patterns in the data. Her work helps determine the impact of the pandemic on review frequency and guest preferences for accommodation types. She introduced the data preparation process in detail and contributed to the report, laying a solid foundation for our research. |
| Yifan Fan | 27% | Texting Mining by Regex and Tokenization, Sentiment analysis to find positive, negative, and neutral words. Yifan specializes in text mining, focusing on processing reviews to extract meaningful insights. This involves converting all review text to lowercase, removing punctuation, special characters and irrelevant numbers, and then tagging the cleaned text. Her expertise was crucial in vectorizing the data, a step that significantly improved the quality of the dataset we used for further analysis. She also creatively uses word cloud visualizations to represent our findings, making the data more accessible and understandable. These word clouds highlight common terms found in reviews, revealing the positive aspects of the Boston guest experience. |
| Jinke Han | 27% | T-SNE, Topic Analysis using LDA, and Model Evaluation to find the best model. Jinke Han's responsibilities focus on conducting topic analysis using LDA models and evaluating the model's performance. By applying the gensim library, Jinke extracted key topics such as service quality and location from user reviews to gain in-depth insights into what guests value most. She carefully trained the LDA model, tuning it to find the optimal number of topics that balances coherence and complexity. Jinke's analytical skills were critical in interpreting the model's results, making a significant contribution to our understanding of thematic structure in Airbnb reviews. |
| Qianyi Mo | 26% | Alpha and Eta in LDA by dividing 3 models, LDA visualization with word clouds and intertropical distance Map. She focused on optimizing the LDA model by adjusting the alpha/eta parameters, which affect the distribution of topics and words respectively. Her work involves conducting sensitivity analysis to find the best combination of these parameters to maximize thematic coherence. Also, she leveraged Jupyter Notebook for advanced visualization, creating engaging and informative graphics to illustrate our findings. Qianyi was responsible for drafting the preliminary results section of our report, concisely summarizing the impact of these parameters on model performance and clearly describing our analysis process. |

# Appendix

## Model 1 results

Model 2 results

# Model 3 results

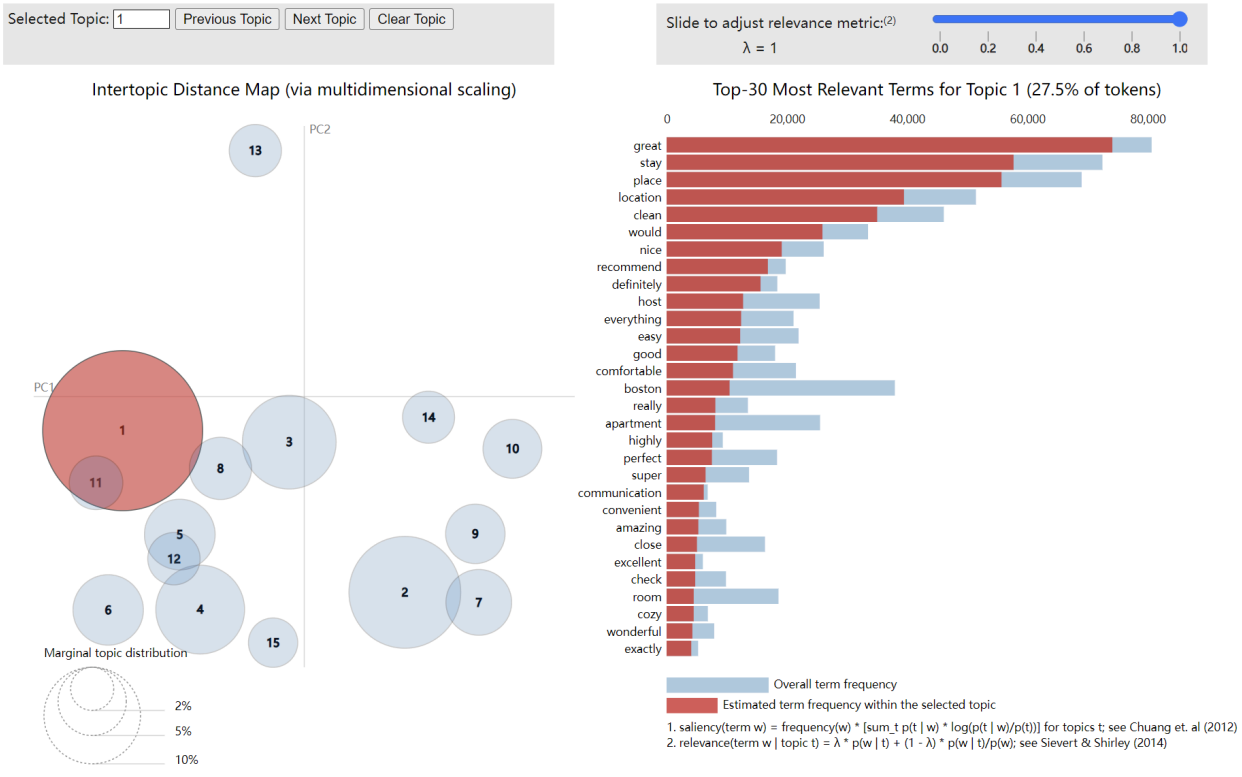



Figure 1

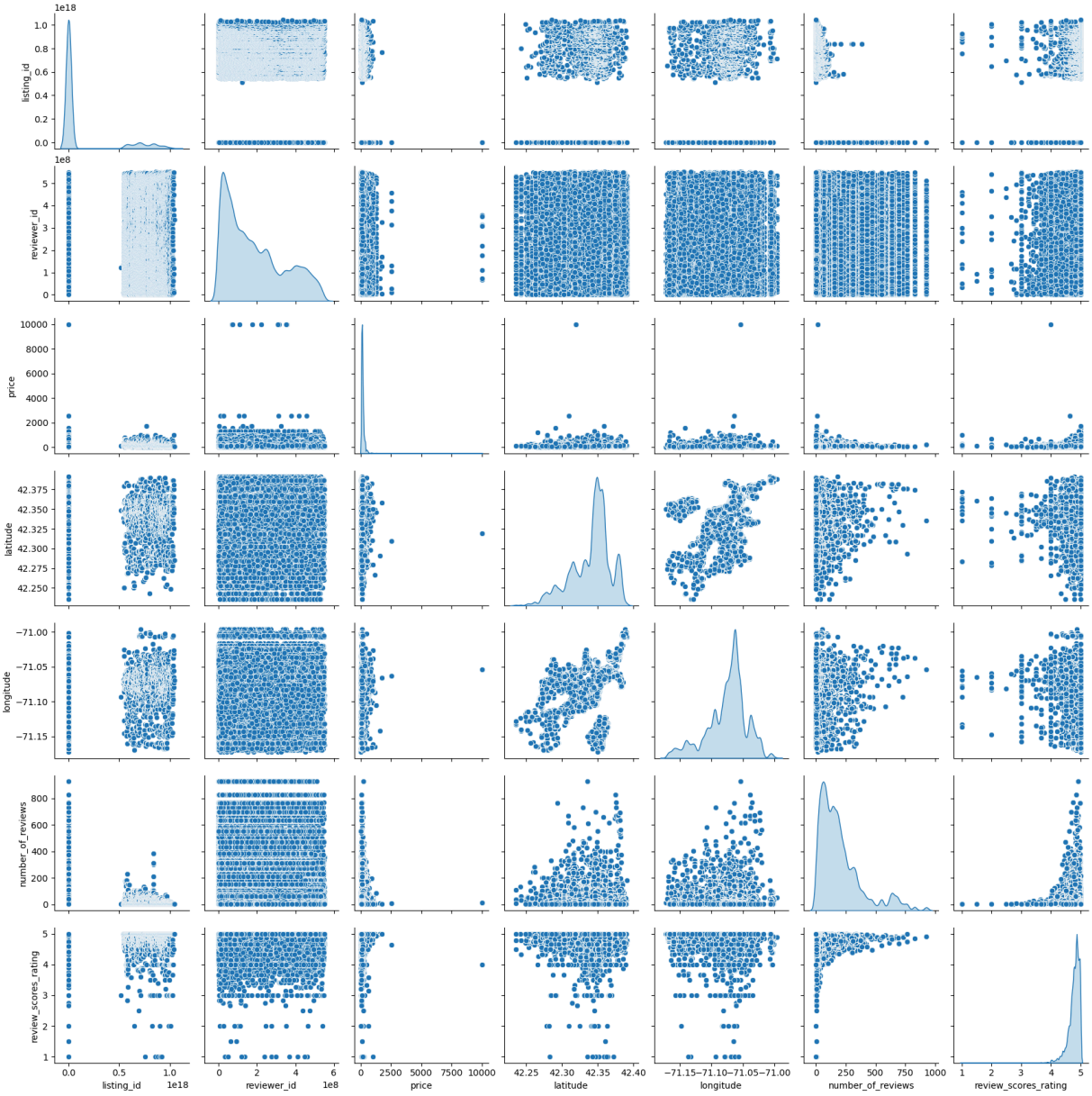

listing_id
reviewer_id
price
latitude
longitude
number_of_reviews
review_scores_rating
1e18
1e8

Figure 2

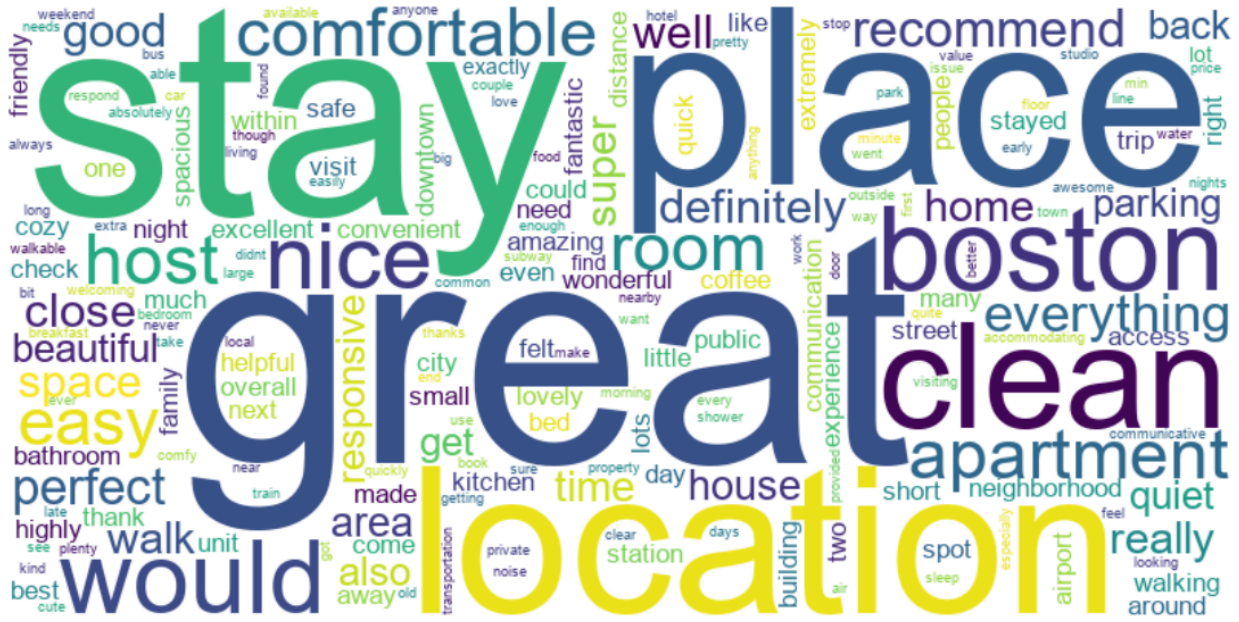



Figure 3

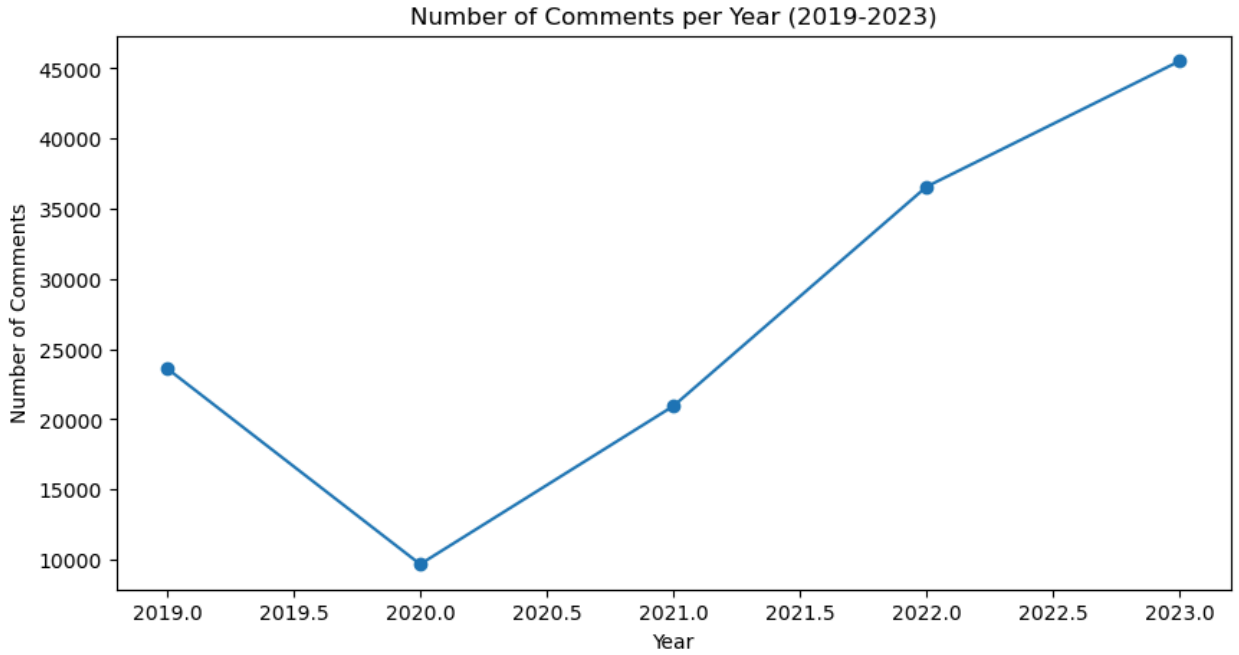


Figure 4

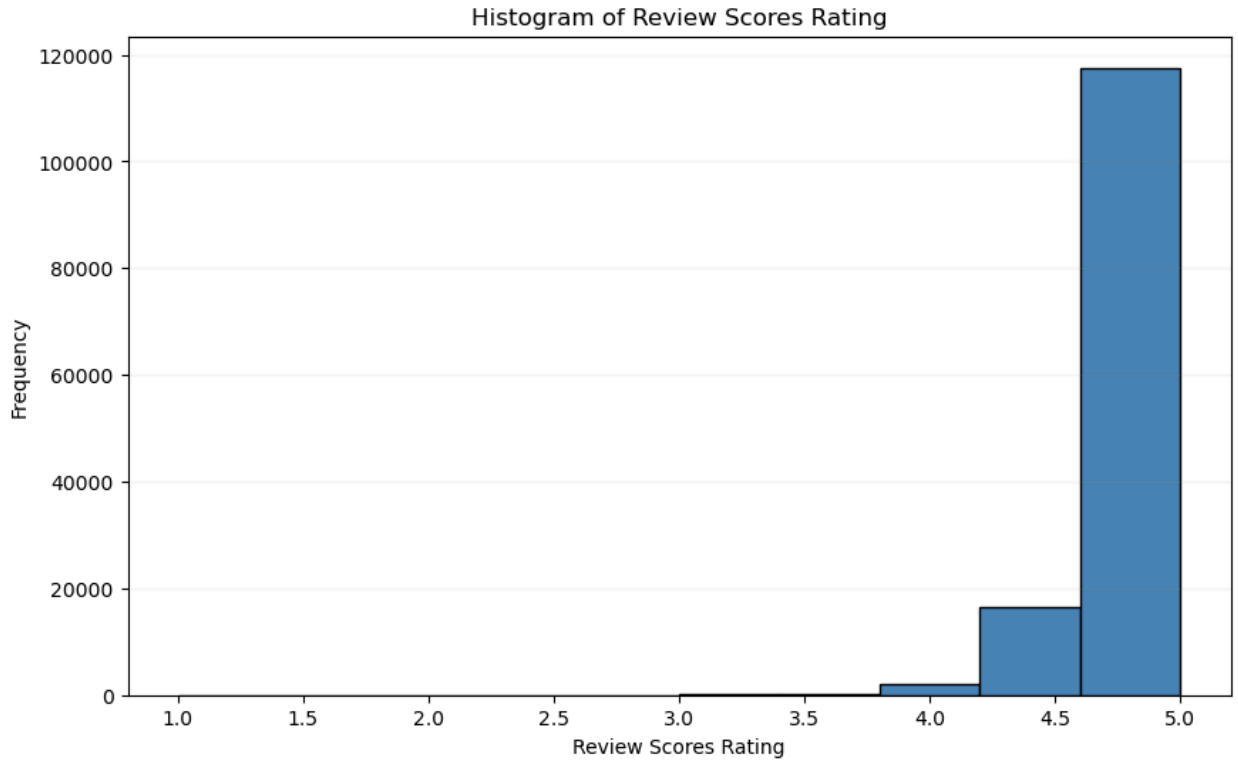



Figure 5

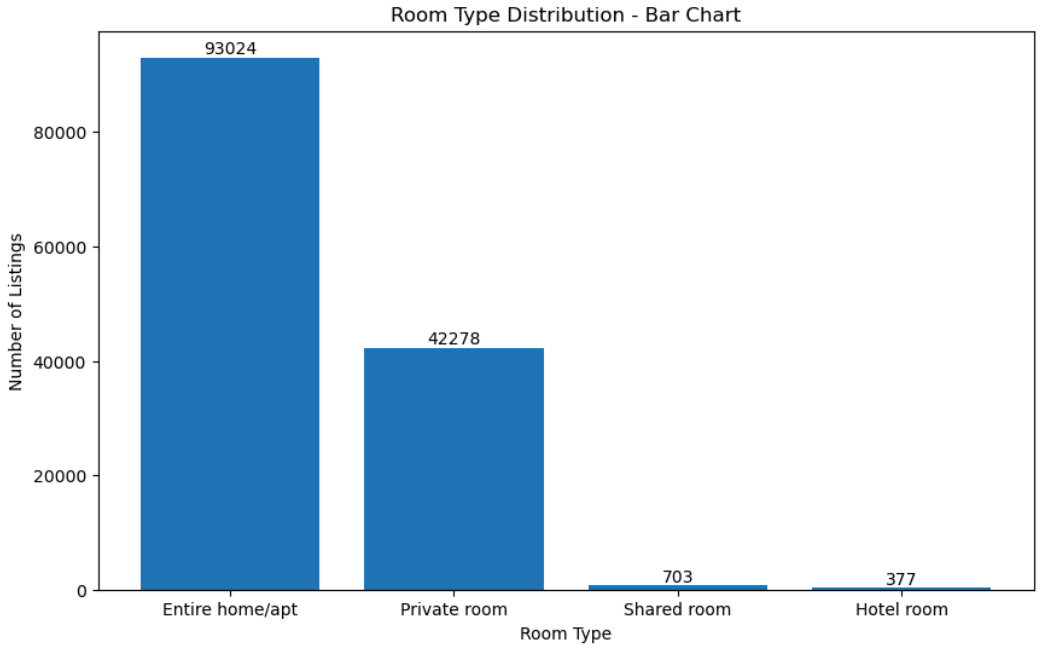
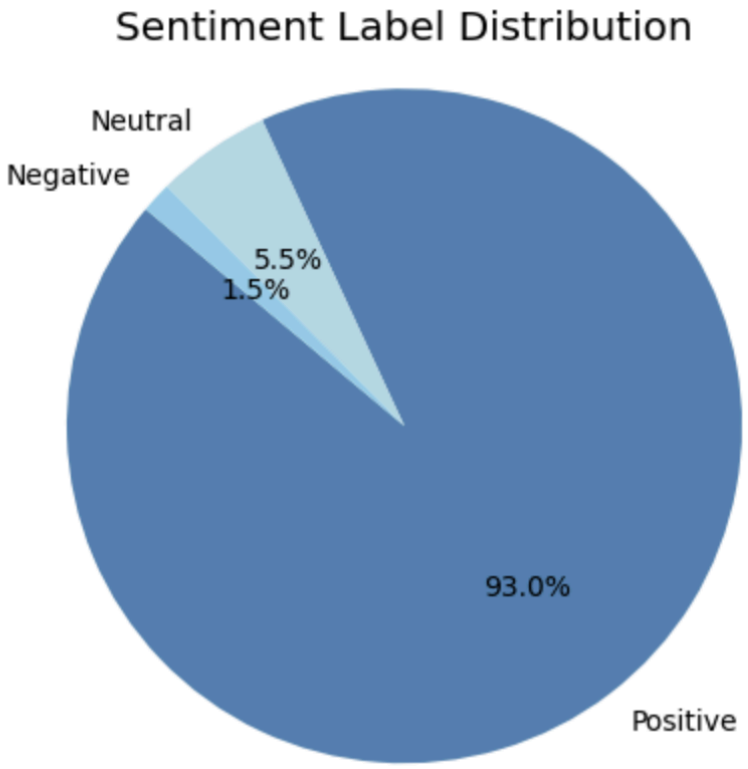


Figure 6


Sentiment Label Distribution
Neutral
Negative
5.5%
1.5%
93.0%
Positive