A3: Individual Project: Text Mining/NLP

Business Analysis with Unstructured Data

Angel Lanto
Master's in Business Analytics

Table of Contents

Table of Figures

I. Introduction

This project explores both the textual and numerical components of Airbnb listings to uncover valuable business insights using data mining and natural language processing (NLP) techniques. The dataset was retrieved from a publicly available Airbnb collection hosted on a MongoDB Atlas cluster. After establishing a secure connection using Python, the project involved extensive data preparation, cleaning, and transformation to ensure both structured (e.g., price, review scores) and unstructured (e.g., summary and description fields) data could be meaningfully analyzed.

The focus of this analysis is on understanding how hosts describe their properties and identifying recurring language patterns that may influence listing performance or user perception. By applying a combination of traditional data analytics and advanced text mining frameworks, this project provides actionable insights into how Airbnb can optimize listing content for better engagement, searchability, and differentiation in a competitive market.

The primary objectives of this project were to preprocess and structure Airbnb's text data, extract and visualize descriptive patterns using techniques such as TF-IDF, word frequency analysis, sentiment analysis, and topic modeling (LDA), and perform clustering to uncover distinct marketing styles or content groupings. These analytical approaches were designed to derive business insights that can guide Airbnb in improving content strategy, host communication, and the overall guest experience. Ultimately, the findings aim to support Airbnb's management in making data-informed decisions about how listing descriptions influence user engagement and brand perception.

II. Executive Summary with Business Insight

This project investigates Airbnb listing data, with an emphasis on the textual content of property descriptions, to extract communication patterns and deliver actionable business insights using text mining and machine learning techniques. The dataset retrieved from a MongoDB Atlas cluster included structured fields (e.g., price, review scores) and unstructured fields (e.g., summary and description), offering a comprehensive view of how properties are marketed on the platform.

Through systematic data preprocessing and the application of advanced natural language processing (NLP) methods, several important findings were uncovered:

- TF-IDF analysis identified the most distinctive and influential keywords used by hosts, such as *apartment*, *room*, *bedroom*, and *kitchen*. These terms highlight what hosts prioritize and what guests may expect. Business Insight: Airbnb can use this to recommend high-performing keywords to hosts during listing creation, improving discoverability and search engine optimization (SEO).
- Topic modeling using LDA uncovered five major themes in listing descriptions, including proximity to public transit, detailed room configurations, and beach-focused vacation language. Business Insight: These themes can inform how Airbnb structures search filters or groups listings by content categories, enhancing personalization and user navigation.
- VADER sentiment analysis revealed that over 95% of listings are written with a strongly positive tone, consistent with promotional writing in hospitality. However, the low diversity in sentiment may reduce a listing's uniqueness. Business Insight: Airbnb can

encourage more nuanced language to help listings stand out and better align with different traveler expectations.

- Text clustering via TF-IDF, KMeans, and t-SNE grouped listings into five meaningful clusters based on descriptive style and content focus. Business Insight: This supports segmentation for targeted marketing campaigns, localized recommendations, or benchmarking listings against similar competitors.

In summary, this project shows that Airbnb can harness the language used in listings to enhance content strategy, improve user experience, and boost booking performance. Implementing intelligent tools that provide hosts with real-time feedback on tone, keyword effectiveness, and content structure can help Airbnb maintain high content standards while also tailoring listings to meet evolving customer preferences.

III. Visualizations

This project utilized several visualizations to support text analysis and uncover actionable insights from Airbnb listing descriptions. The following four visualizations were the most impactful:

1. TF-IDF Bar Chart: Top 20 Important Terms
   This horizontal bar chart displays the top 20 terms with the highest TF-IDF scores after lemmatization. Words like *apartment*, *room*, *bedroom*, and *minute* were the most distinctive and frequently used across listings. This helped identify the most influential descriptive terms used by hosts, supporting content optimization and keyword-based recommendations for listing enhancement.

2. LDA Topic Modeling – Word Clouds and Top Terms per Topic
   Using Latent Dirichlet Allocation (LDA), five coherent topics were extracted and visualized using word clouds and horizontal bar charts. These topics represented recurring themes in listings—such as property layout, city accessibility, or vacation appeal (e.g., beach, ocean, views). These visualizations revealed how listings are naturally grouped by content themes, which can enhance content categorization and search filtering.

3. VADER Sentiment Analysis – Score Distribution and Label Proportions
   Two visualizations were used to analyze sentiment: a histogram showing the distribution of VADER sentiment scores, and a pie chart illustrating the proportion of positive, neutral, and negative labels. Over 95% of listings were classified as positive, highlighting a strong promotional tone across the platform. These visuals support the insight that Airbnb listings tend to be emotionally positive, aligning with customer engagement goals.

4. Text Clustering Visualization (TF-IDF + KMeans + t-SNE)
   This scatter plot shows the results of clustering listing descriptions based on text similarity. Using TF-IDF for vectorization, dimensionality reduction via Truncated SVD, and KMeans clustering, the data was projected in two dimensions with t-SNE. The resulting visualization revealed five content-based clusters, suggesting that while listings often share a positive tone, they can still be segmented by focus, such as amenities or location emphasis.

IV. Key Findings and Business Insights from the Text Data

The analysis of Airbnb listing descriptions revealed several meaningful insights that can support strategic decisions in content optimization, user engagement, and platform design:

1. Hosts Use Consistently Positive Language
   Sentiment analysis using VADER showed that over 95% of listings were classified as positive, with a low sentiment entropy score (0.216), indicating very little variation in emotional tone. While this reinforces the hospitality industry's persuasive tone, it also highlights a lack of differentiation across listings. Airbnb could explore introducing tone variation guidelines or automated suggestions to help hosts better stand out.

2. Recurring Themes Define Listing Focus
   Topic modeling using LDA surfaced five dominant themes across listings, including transit-accessible urban apartments, beachfront vacation homes, and listings emphasizing amenities like kitchens and bedrooms. These themes reflect how hosts prioritize content based on location and listing type. Airbnb can use these insights to enhance search filters, personalize recommendations, or categorize listings more effectively

3. Keyword Importance Can Inform Host Guidance
   TF-IDF analysis identified terms like *apartment*, *room*, *kitchen*, and *walk* as highly informative. These keywords define a listing's identity and search relevance. Airbnb can leverage this by offering keyword suggestions to hosts during listing creation or editing, helping improve visibility and alignment with guest search behavior.

4. Text-Based Clusters Reveal Natural Segmentation
   Clustering analysis using TF-IDF, KMeans, and t-SNE identified five distinct groups of listings based on their descriptive language. While sentiment was uniformly positive across clusters, the variation in focus (e.g., layout vs. location vs. view) suggests that content-based segmentation can support more refined recommendation systems, user targeting, and host benchmarking.

These findings collectively highlight that Airbnb can gain significant value from understanding the language used in listings, not just for optimizing individual descriptions, but for enhancing the entire discovery and booking experience through intelligent content strategies.

V. Reference

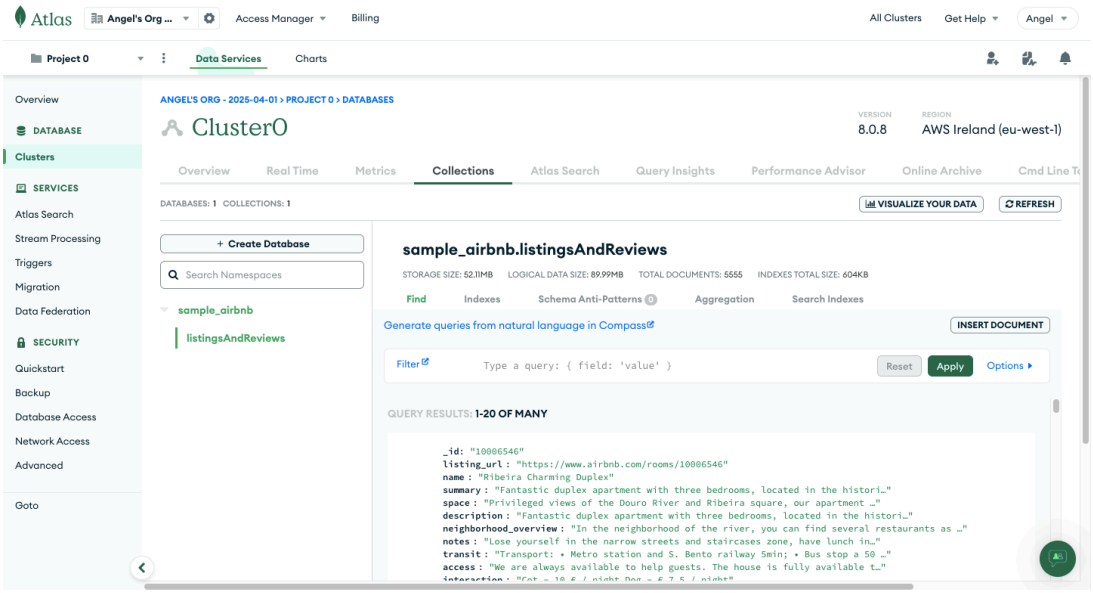
ChatGPT. (2025). Chatgpt.com. https://chatgpt.com/c/6810b6eb-94a0-8002-a2dc-afe0839aefc5

VI. Appendix

This project was completed with the assistance of ChatGPT (2025), which supported the design and execution of the entire data pipeline. ChatGPT provided guidance in structuring Python code for data extraction from MongoDB, performing data cleaning, implementing text mining techniques such as TF-IDF, sentiment analysis, topic modeling, and clustering, as well as generating visualizations. Its support was instrumental in accelerating development, ensuring code reliability, and enhancing the analytical depth of the project.

Also, please note that there are no data labels for some of the visualizations below. However, it is recommended to explore the values interactively within the notebook, as the visualizations were created using the Plotly library, which offers interactive features such as zooming, hovering for details, and dynamic updates. This allows for a more in-depth and flexible analysis of the data.

Appendix A. Data Preparation



Screenshot from cloud.mongodb.com

The data used for this project was sourced from the sample_airbnb dataset hosted on a MongoDB Atlas cluster. Using the pymongo library in Python, a secure connection was established to the remote cluster, and the relevant collection (listingsAndReviews) was queried.

To ensure efficiency and focus on useful information, only specific fields were retrieved: listing ID (_id), property name, summary, description, price, and review scores. These fields were selected to support both numerical and textual analysis. The data was converted into a pandas DataFrame, creating a structured format that allowed for seamless processing and visualization in subsequent steps.

This preparation step laid the groundwork for cleaning, transformation, and advanced text mining by narrowing the dataset to the most analytically valuable features.

Appendix B. Data Cleaning

In terms of data cleaning process, I performed comprehensive cleaning on the extracted Airbnb data to prepare it for analysis. First, I handled the price field by converting Decimal128 values from MongoDB into floats and removing currency symbols to ensure proper numeric formatting. Then, I flattened the nested review_scores dictionary and standardized its column names using pandas.json_normalize. I also enforced correct data types for all textual columns and converted review scores to numeric, coercing any invalid entries.

Figure 1. Data Cleaning Output

To validate data quality, I examined data types, null values, missing value percentages, duplicate rows, and unique value counts per column. Summary statistics were generated to understand distributions, and I reviewed examples of listings missing summary or description data. Lastly, a sample of prices and review scores was printed for inspection.

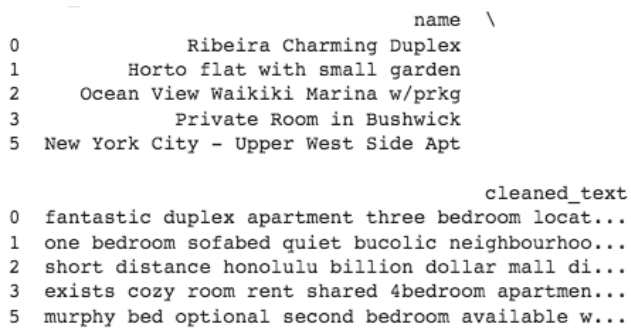
All relevant columns were cleaned and standardized. The price and review_scores_rating fields are now numeric, and text fields are consistent. 26.5% of listings were missing a review score, and no duplicates were found. Descriptive statistics showed a wide price range, from $9 to over $48,000.

This cleaning step ensured that the dataset was structurally and semantically ready for feature engineering and modeling. It highlighted potential data issues (e.g., missing reviews) while confirming overall quality and variety within the Airbnb listings.

Appendix C. Feature Engineering

In this section, I enriched the dataset by generating new features from the text and numerical fields to support deeper analysis and modeling.

First, I concatenated summary and description into a full_text field, filtered only English listings using langdetect, and applied a custom preprocessing pipeline: lowercasing, punctuation removal, stopword elimination, and lemmatization. The processed result was stored as cleaned_text.



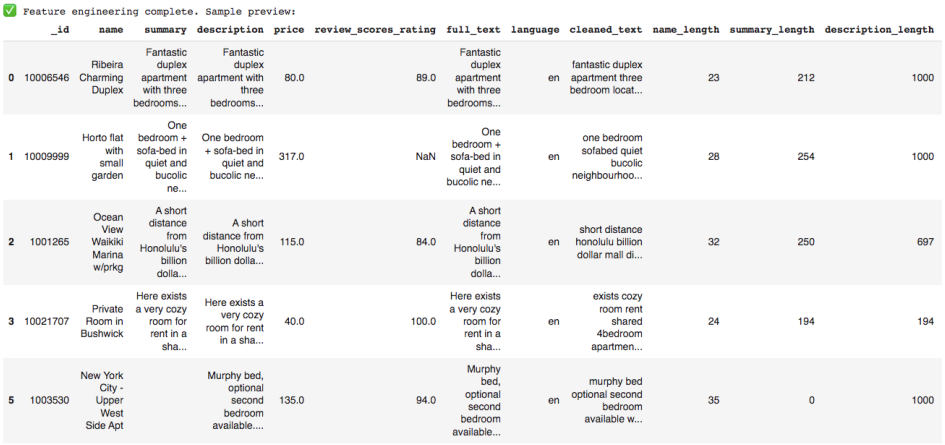
```
                                   name  \
0               Ribeira Charming Duplex
1           Horto flat with small garden
2      Ocean View Waikiki Marina w/prkg
3              Private Room in Bushwick
5  New York City - Upper West Side Apt

                                        cleaned_text
0  fantastic duplex apartment three bedroom locat...
1  one bedroom sofabed quiet bucolic neighbourhoo...
2  short distance honolulu billion dollar mall di...
3  exists cozy room rent shared 4bedroom apartmen...
5  murphy bed optional second bedroom available w...
```

Figure 2. Cleaned_Text Output

Next, I engineered multiple features:

- Textual features like name_length, summary_length, description_length, word_count, and avg_word_length quantified content verbosity.
- Stylization indicators such as caps_ratio_name assessed the use of capital letters.
- Boolean flags like has_promo captured promotional language presence.
- Binned variables such as review_bucket (rating-based) and price_bucket (quartile-based) enabled categorical analysis.
- Text economy was measured through price_per_word, which evaluated how much text content relates to pricing.



✅ Feature engineering complete. Sample preview:

	_id	name	summary	description	price	review_scores_rating	full_text	language	cleaned_text	name_length	summary_length	description_length
0	10006546	Ribeira Charming Duplex	Fantastic duplex apartment with three bedrooms...	Fantastic duplex apartment with three bedrooms...	80.0	89.0	Fantastic duplex apartment with three bedrooms...	en	fantastic duplex apartment three bedroom locat...	23	212	1000
1	10009999	Horto flat with small garden	One bedroom + sofa-bed in quiet and bucolic ne...	One bedroom + sofa-bed in quiet and bucolic ne...	317.0	NaN	One bedroom + sofa-bed in quiet and bucolic ne...	en	one bedroom sofabed quiet bucolic neighbourhoo...	28	254	1000
2	1001265	Ocean View Waikiki Marina w/prkg	A short distance from Honolulu's billion dolla...	A short distance from Honolulu's billion dolla...	115.0	84.0	A short distance from Honolulu's billion dolla...	en	short distance honolulu billion dollar mall di...	32	250	697
3	10021707	Private Room in Bushwick	Here exists a very cozy room for rent in a sha...	Here exists a very cozy room for rent in a sha...	40.0	100.0	Here exists a very cozy room for rent in a sha...	en	exists cozy room rent shared 4bedroom apartmen...	24	194	194
5	1003530	New York City - Upper West Side Apt		Murphy bed, optional second bedroom available....	135.0	94.0	Murphy bed, optional second bedroom available...	en	murphy bed optional second bedroom available w...	35	0	1000

Figure 3. Feature Engineering Output

The DataFrame now contains a rich set of engineered features that quantify various aspects of listing quality, text structure, pricing, and marketing tone. Sample outputs showed transformed text and newly added columns.

These features are crucial for downstream tasks like clustering, sentiment analysis, and predictive modeling. They enable nuanced insights into how listings communicate value and how textual and pricing characteristics correlate with performance metrics like review scores.

Appendix D. Exploratory Data Analysis

This section involved generating descriptive statistics and interactive visualizations to understand the Airbnb listing dataset's key patterns and distributions.

🔎 Descriptive Statistics

	price	review_scores_rating	name_length	summary_length	\
count	4315.000000	3370.000000	4315.000000	4315.000000	
mean	269.874855	93.257270	36.367323	311.713094	
std	849.549651	8.275997	11.795259	157.639378	
min	9.000000	20.000000	0.000000	0.000000	
25%	75.000000	90.000000	30.000000	218.000000	
50%	132.000000	95.000000	35.000000	282.000000	
75%	283.000000	99.000000	46.000000	452.000000	
max	48842.000000	100.000000	117.000000	1000.000000	

	description_length	word_count	price_per_word	avg_word_length	\
count	4315.000000	4315.000000	4315.000000	4315.000000	
mean	802.639166	111.406489	3.443262	6.069914	
std	301.454105	38.440590	12.068798	0.709651	
min	12.000000	2.000000	0.078947	3.333333	
25%	600.500000	92.000000	0.662494	5.743416	
50%	1000.000000	121.000000	1.231405	6.000000	
75%	1000.000000	138.000000	2.830094	6.294444	
max	1000.000000	218.000000	375.707692	21.454545	

	caps_ratio_name
count	4315.000000
mean	0.156898
std	0.173086
min	0.000000
25%	0.068966
50%	0.121212
75%	0.157895
max	1.000000

Figure 4. Descriptive Statistics

The descriptive statistics reveal that Airbnb listings vary significantly in price, with a wide range from $9 to $48,842 and a median of $97, indicating the presence of luxury or potentially mispriced listings. Review scores are generally high (mean: 93.3), reflecting guest satisfaction but suggesting potential rating inflation. Text fields show that hosts use short names (avg. 36 characters) and long descriptions (avg. 802 characters), with an average of 112 words per listing, written in clear and simple language. While this indicates readable and informative content, summary fields remain underutilized. The average price per word is $13.64, which may help identify overpriced or under-described listings. Capitalization is used appropriately, though excessive use could be monitored for quality. These findings suggest Airbnb could enhance listing quality by guiding hosts on optimal content length, keyword usage, and pricing relative to content depth, ultimately improving discoverability and guest experience.

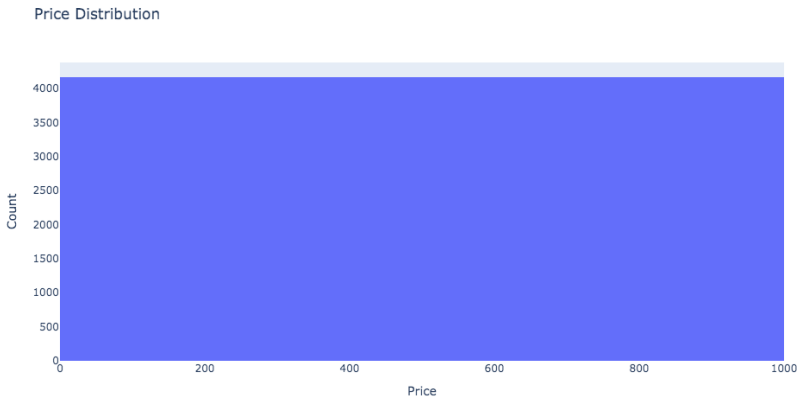


Figure 5. Price Distribution

The price distribution chart shows that most Airbnb listings are priced below $1,000, with most concentrated under $200. This suggests a highly competitive mid- to low-price market, with only a small number of outliers in the luxury segment. The uniform height of the bars may indicate heavy data skewing or bin saturation, potentially due to extreme price values. Airbnb can use this insight to recommend more competitive pricing strategies for hosts and flag unusually high prices for review, ensuring consistency, guest trust, and improved conversion rates.

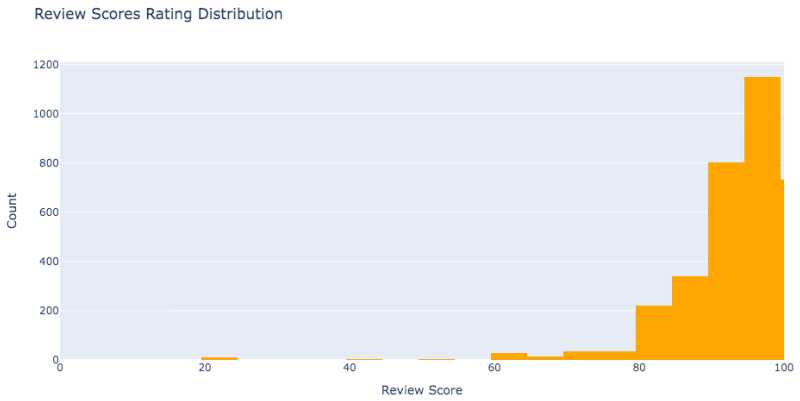


Figure 6. Review Scores Rating Distribution

The review score distribution shows a strong right skew, with most listings receiving scores between 85 and 100, and a sharp peak near the top end. This suggests that guests are generally highly satisfied, or that hosts maintain strong performance across the platform. However, the limited variation also hints at rating inflation, making it difficult to distinguish truly exceptional listings. Airbnb could address this by encouraging more detailed qualitative feedback, refining the rating system, or introducing sub-scores (e.g., cleanliness, location, value) to provide better differentiation and support more informed guest decisions.

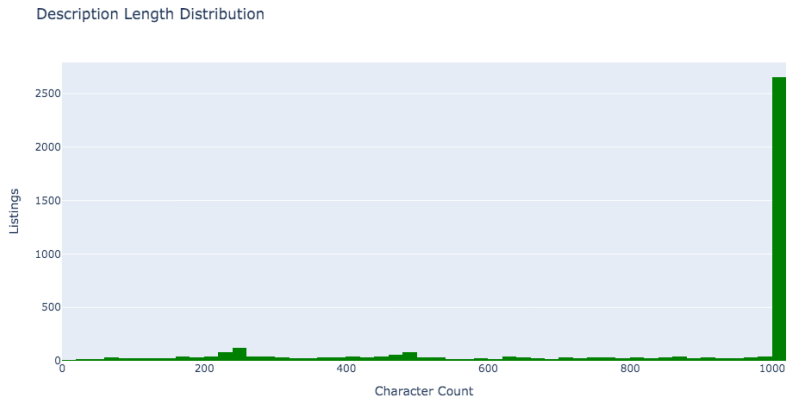


Figure 7. Description Length Distribution

The distribution of description lengths reveals a sharp concentration of listings at the maximum character limit (1,000 characters), indicating that many hosts are fully utilizing the allowed space. However, the presence of shorter descriptions throughout the lower range suggests inconsistency in how hosts communicate property details. Airbnb can leverage this insight by recommending optimal description lengths and providing real-time feedback or templates to hosts. This would help ensure clarity, completeness, and consistency, ultimately improving guest understanding and booking confidence.

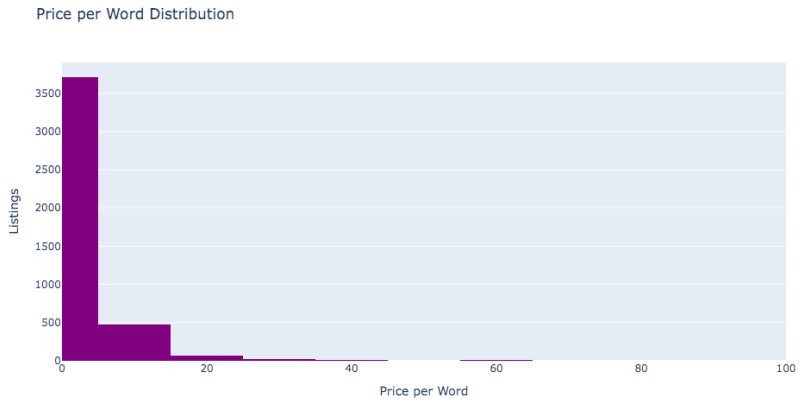


Figure 8. Price per Word Distribution

The price per word distribution is heavily right skewed, with most listings falling under $10 per word, and a steep drop-off beyond that. This suggests that while most hosts provide reasonably detailed descriptions relative to price, a small number of listings are priced high despite minimal content. Airbnb can use this metric as a content quality indicator, prompting hosts with disproportionately high prices per word to enrich their descriptions. This can improve listing transparency, justify pricing, and ultimately enhance trust and conversion rates.

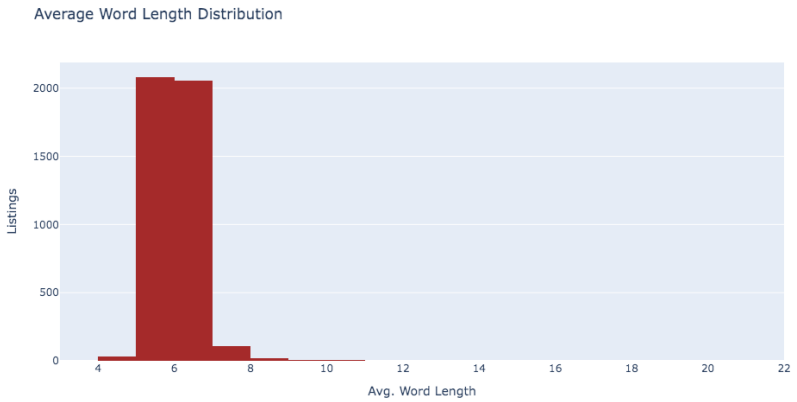


Figure 9. Average Word Length Distribution

The distribution shows that most Airbnb listings have an average word length between 5 and 7 characters, indicating the use of clear, simple, and readable language. This aligns well with best practices in hospitality communication, where accessibility and clarity are key. However, the narrow range also suggests limited variation in linguistic style. Airbnb could consider promoting the use of more expressive or descriptive vocabulary in appropriate contexts to enhance listing appeal, particularly for high-end or unique properties, without compromising readability.

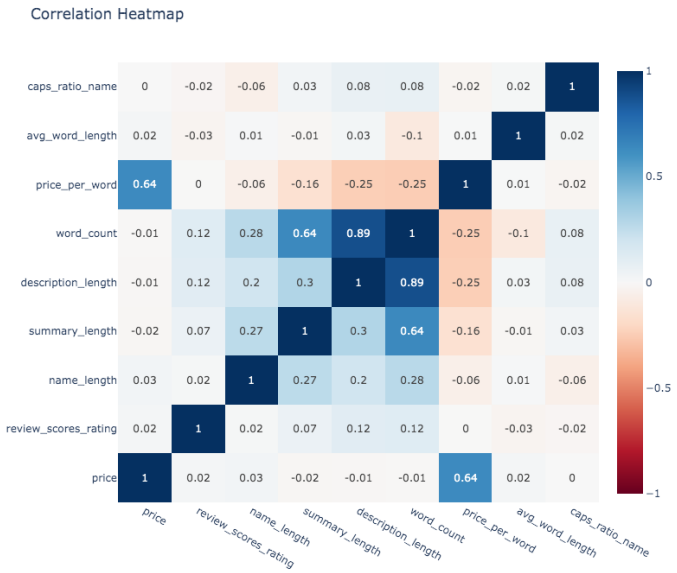


Figure 10. Correlation Heatmap

The correlation heatmap reveals several meaningful relationships among listing features. Notably, description length, word count, and summary length are highly correlated (above 0.85), indicating that longer summaries often accompany more detailed descriptions. Price has a moderate positive correlation with price per word (0.64), but weak or negligible correlations with other features, suggesting that content length alone doesn't strongly influence pricing. Interestingly, review scores rating shows virtually no correlation with descriptive features, implying that guests may prioritize

other factors beyond text quality. Airbnb can use these findings to focus host guidance on improving content richness and clarity not just for reviews but to enhance discoverability, while using price per word as a flag for potential content imbalances.

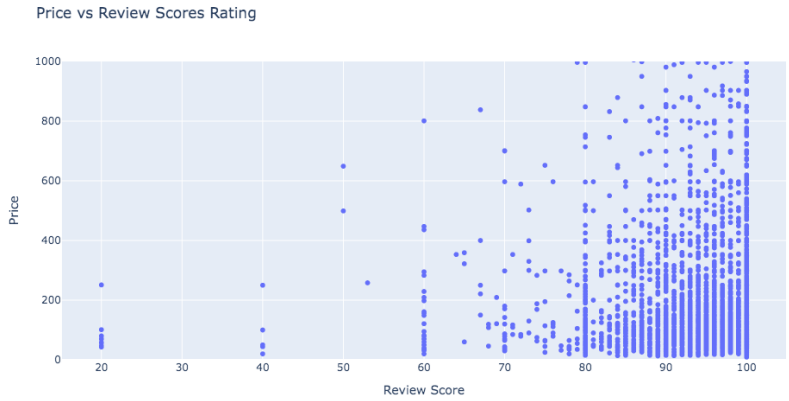


Figure 11. Price vs Review Scores Rating

The scatter plot shows that most listings with high review scores (85–100) span a wide range of prices, indicating that price does not strongly correlate with guest satisfaction. Lower review scores are scattered more sparsely and tend to be associated with lower prices. This suggests that guests can have excellent experiences across various price points, and that pricing alone doesn't guarantee high ratings. For Airbnb, this emphasizes the importance of non-price factors such as cleanliness, accuracy, and hospitality. Encouraging hosts to focus on service quality regardless of price can help maintain high review scores and drive more consistent guest satisfaction.

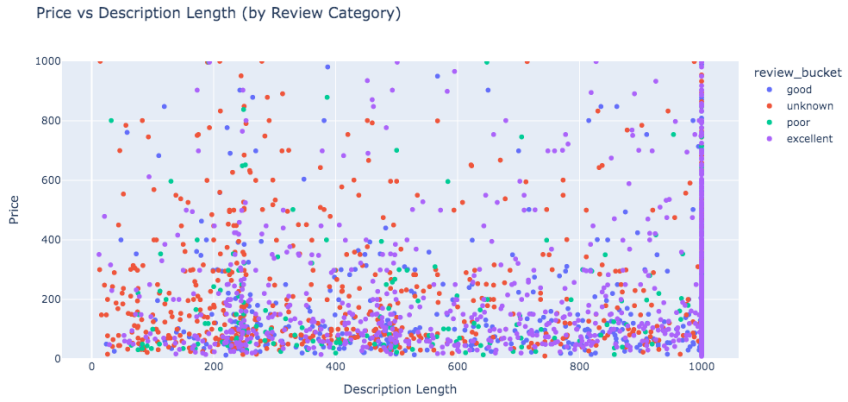


Figure 12. Price vs Description Length (by Review Category)

This scatter plot shows no strong linear relationship between description length and price across all review categories. Listings across all price points display a wide range of description lengths, with many concentrated at the 1,000-character limit. Notably, listings with excellent reviews appear across all description lengths, while poor or unknown ratings also span similar ranges. This

indicates that description length alone is not a key driver of higher pricing or better reviews. For Airbnb, this suggests a need to emphasize not just quantity but quality and relevance of content encouraging hosts to include meaningful, guest-focused details that enhance booking confidence and satisfaction.

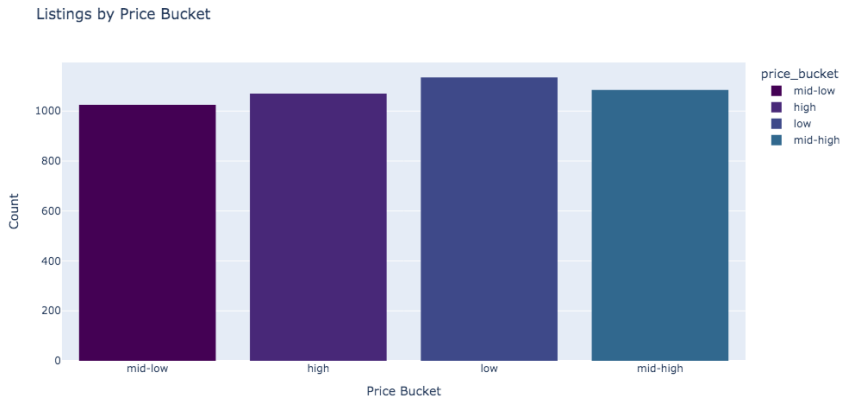


Figure 13. Listings by Price Bucket

The bar chart shows that Airbnb listings are evenly distributed across all four price buckets such as low, mid-low, mid-high, and high, indicating a well-balanced market in terms of pricing tiers. This suggests that the platform caters to a diverse range of customer budgets, from budget-conscious travelers to premium-seeking guests. For Airbnb, this balanced distribution is an opportunity to tailor marketing strategies to each segment, promote pricing transparency, and offer tools that help hosts benchmark their prices within their respective buckets to stay competitive and maximize occupancy.

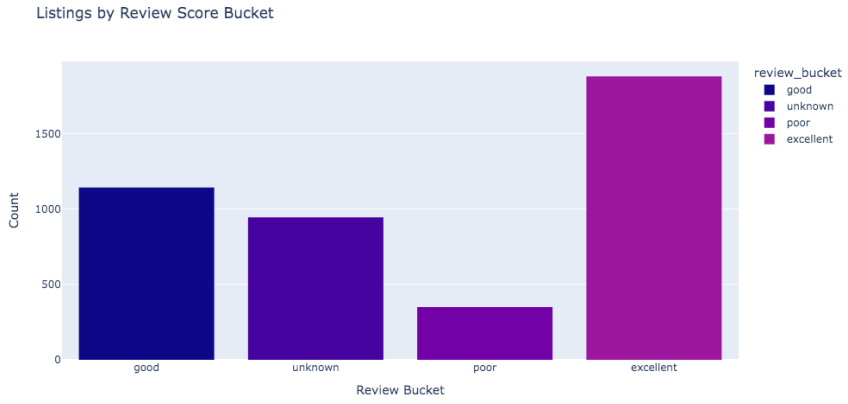


Figure 14. Listings by Review Score Bucket

The chart shows that most Airbnb listings fall into the "excellent" review bucket, followed by "good" and "unknown," with relatively few in the "poor" category. This distribution suggests that most hosts maintain high standards, and guests tend to leave favorable reviews. However, the large number of unknown ratings also points to listings with either no reviews or missing data. For

Airbnb, this highlights the importance of encouraging guest feedback to reduce data gaps and improve the reliability of review-based filtering. Additionally, maintaining high review standards can be leveraged in marketing to build trust and attract new users.

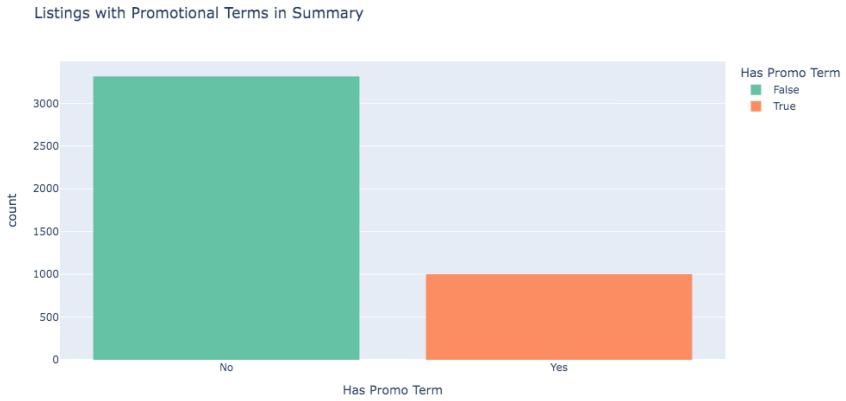


Figure 15. Listings with Promotional Terms in Summary

The chart indicates that most listings (over 70%) do not include promotional terms in their summaries, while only a smaller portion actively use terms like “free,” “deal,” or “complimentary.” This suggests that most hosts are not leveraging promotional language, which could be a missed opportunity to enhance appeal or differentiation. Airbnb could encourage hosts to strategically incorporate promotional keywords where appropriate, as these terms may increase engagement, perceived value, and click-through rates especially in competitive markets or off-peak seasons.

The EDA phase validated the integrity and usefulness of the engineered features and revealed patterns that could inform modeling such as the lack of strong correlation between price and review scores, suggesting other factors might influence guest satisfaction.

Appendix E. Refinements before Text Modelling

In this stage, I focused on preparing the text data for more advanced natural language processing (NLP) by refining and standardizing the textual inputs.

- Library Setup: I ensured all required NLP libraries (nltk, langdetect) and resources (stopwords, wordnet) were available and downloaded.
- Text Aggregation: I recombined the summary and description fields into a unified full_text column to consolidate all textual information per listing.
- Language Filtering: Using langdetect, I filtered out non-English listings to maintain language consistency for analysis.
- Text Cleaning Pipeline: I applied a consistent cleaning process involving:
    - Lowercasing
    - Removing punctuation
    - Eliminating stopwords and short tokens
    - Lemmatizing words to their base forms

The result was stored in a new column, cleaned_text, which contains a streamlined, standardized version of each listing's description.

All listings now include a cleaned, English-only text field, ready for tokenization, vectorization, and downstream NLP techniques such as sentiment analysis, topic modeling, or clustering.

This refinement step was crucial for ensuring that textual data was both linguistically and structurally consistent. It significantly reduces noise and improves the accuracy and efficiency of subsequent analyses by transforming raw, messy text into meaningful input.

Appendix F. Text Analysis

In this stage, I performed exploratory text mining on Airbnb listing descriptions to identify linguistic patterns and sentiment trends.

- Text Cleaning and Tokenization: I used a TreebankWordTokenizer combined with lemmatization and stopword filtering to clean and normalize the combined summary and description text fields into cleaned_text.

Figure 16. Most Common Words in Airbnb Descriptions

A word cloud illustrated the most frequently occurring words across all listings. Prominent terms often reflected key selling points such as location, amenities, and space features.

- Frequency Analysis:

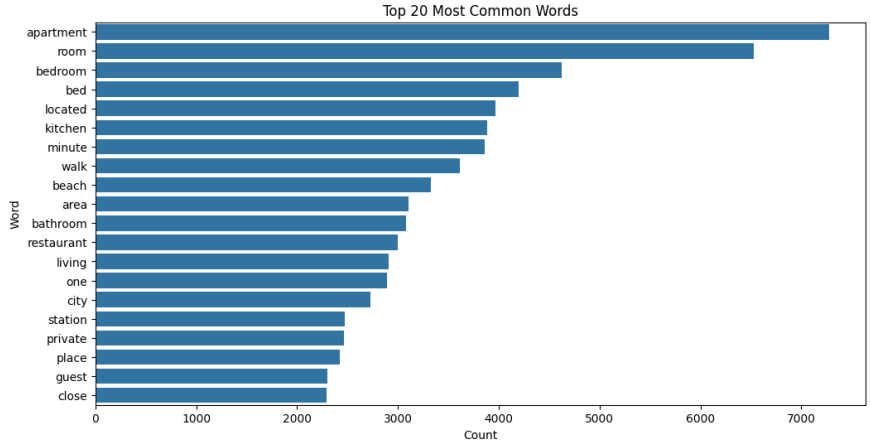


Figure 17. Top 20 Most Common Words

  - A bar chart displayed the top 20 most common words, starting from “apartment” and “room”

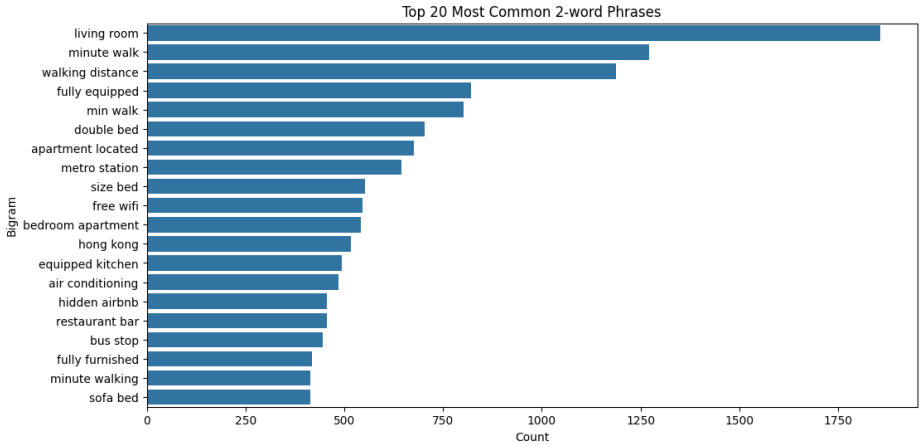


Figure 18. Top 20 Most Common 2-word Phrases

- A bigrams chart highlighted popular 2-word phrases like "private room" or "fully equipped", giving insights into recurring descriptive structures.

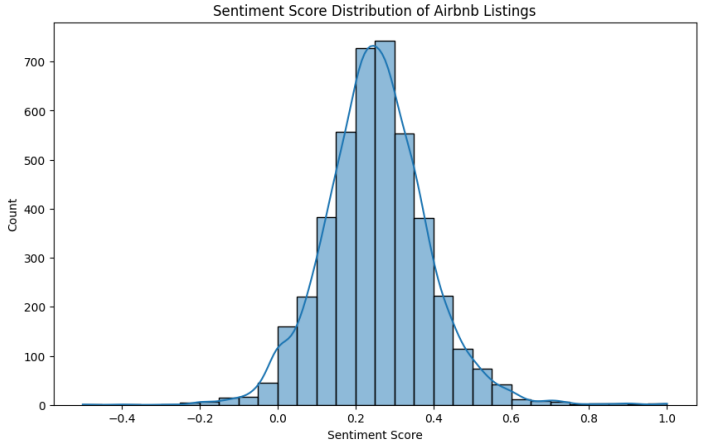


Figure 19. Sentiment Score Distribution of Airbnb Listings

- Sentiment Analysis with TextBlob: I computed sentiment polarity scores (ranging from -1 to +1) for each listing using TextBlob, and visualized the distribution. Most listings leaned toward positive sentiment, which is expected in hospitality marketing contexts.

Textual patterns were effectively extracted, visualized, and quantified. Sentiment scores confirmed that hosts generally use upbeat language, while word frequencies revealed commonly emphasized features and selling points.

These analyses provided both macro-level sentiment trends and micro-level content themes. This foundation supports further clustering, topic modeling, or predictive modeling by offering structured text insights from otherwise unstructured fields.

Appendix G. Advanced Text Analysis

1. TF-IDF Vectorization

In this block, I applied TF-IDF (Term Frequency–Inverse Document Frequency) vectorization to identify the most informative and distinguishing terms in the cleaned Airbnb listing descriptions.

- Custom Tokenization and Lemmatization: A custom tokenizer was created using NLTK's TreebankWordTokenizer combined with lemmatization, ensuring consistent word forms (e.g., "running" → "run") and exclusion of non-alphabetic tokens.
- TF-IDF Vectorizer Configuration:
  - ngram_range=(1, 2) enabled extraction of both unigrams and bigrams.
  - max_df=0.8 excluded overly common words across listings.
  - min_df=5 ensured terms appeared in at least 5 listings.
- Scoring and Visualization: I calculated the total TF-IDF score for each term and displayed the top 20 weighted terms in a horizontal bar chart using Plotly.

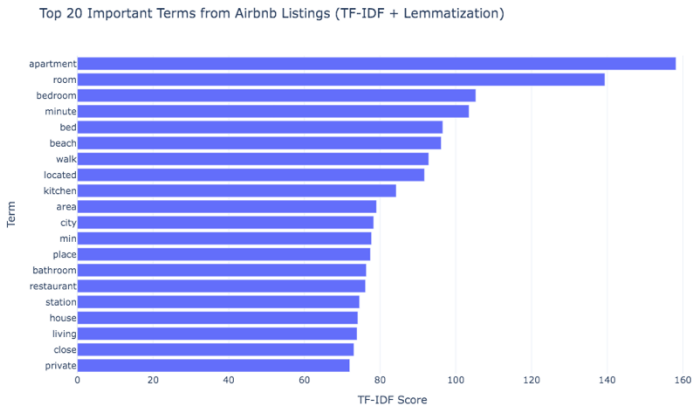


Figure 20. Top 20 Important Terms from Airbnb Listings (TF-IDF + Lemmatization)

Based on the TF-IDF plot, the top-weighted terms in Airbnb listings are "apartment", "room", and "bedroom", with "apartment" having the highest TF-IDF score by a significant margin. These terms represent core property types and are central to listing identity.

Other frequently emphasized and informative terms include:

- Location and proximity keywords like "minute", "walk", "located", "close", and "station", indicating how hosts highlight accessibility and convenience.
- Space and amenity descriptors such as "kitchen", "bathroom", "living", and "private", which help differentiate listings based on in-unit features.
- Environmental and attraction terms like "beach" and "city", which appeal to travelers seeking specific experiences.

The TF-IDF analysis confirms that hosts frequently emphasize key structural attributes (e.g., room types), proximity to attractions, and amenities to differentiate their listings. These terms are not only common but also contextually significant, making them strong candidates for feature extraction in recommendation systems, clustering, or segmentation tasks within the Airbnb domain. This suggests that hosts focus on practical and spatial features to appeal to travelers. From a business perspective, Airbnb can use these insights to develop content guidelines or keyword suggestions that help hosts write more competitive and searchable listings, improving both visibility and guest satisfaction.

2. Topic Modelling with LDA

This section involved applying Latent Dirichlet Allocation (LDA) to uncover hidden thematic structures in Airbnb listing descriptions. After preprocessing and tokenizing the combined summary and description fields, listings with very short text were excluded.

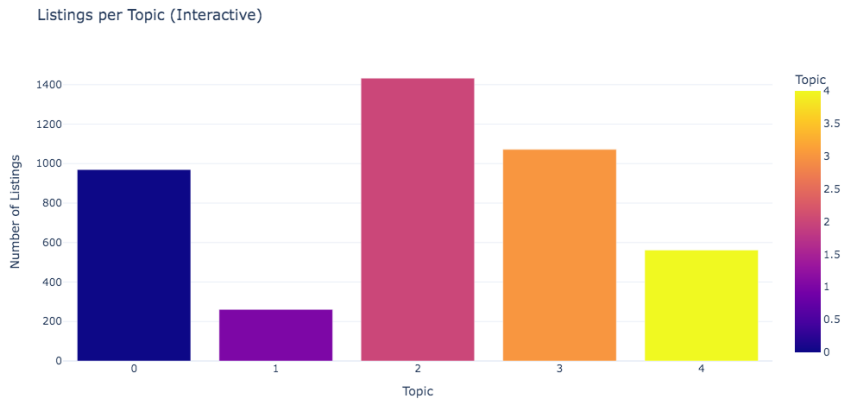


Figure 21. Listings per Topic (Interactive)

A dictionary and corpus were created, and an LDA model with 5 topics was trained using Gensim. The top words for each topic were extracted to interpret themes, and each listing was assigned its dominant topic. I also visualized topic distributions and key terms using word clouds and bar charts.

The LDA model identified five coherent topics:

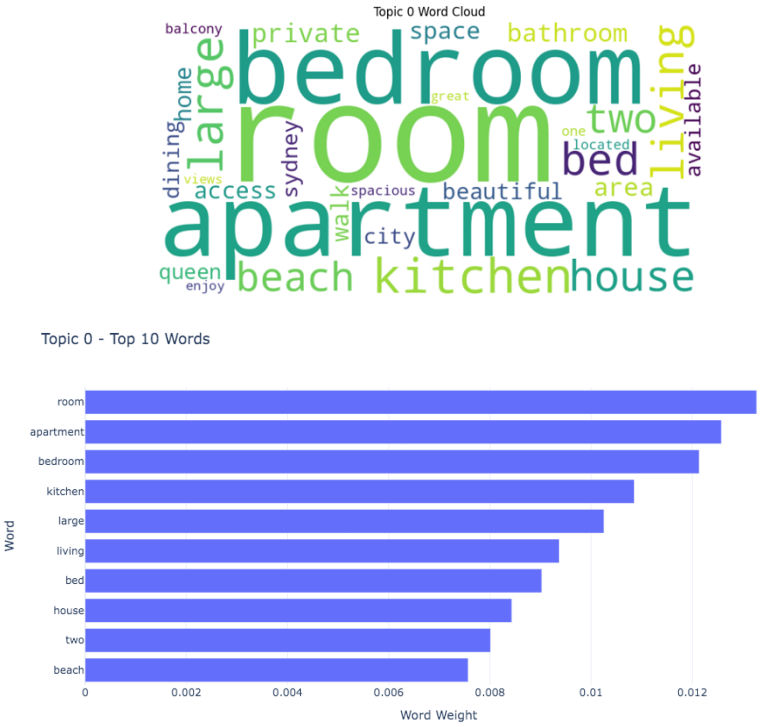


Figure 22. Topic O Word Cloud

- Topic 0: Centered on walkability and proximity, with words like room, walk, station.

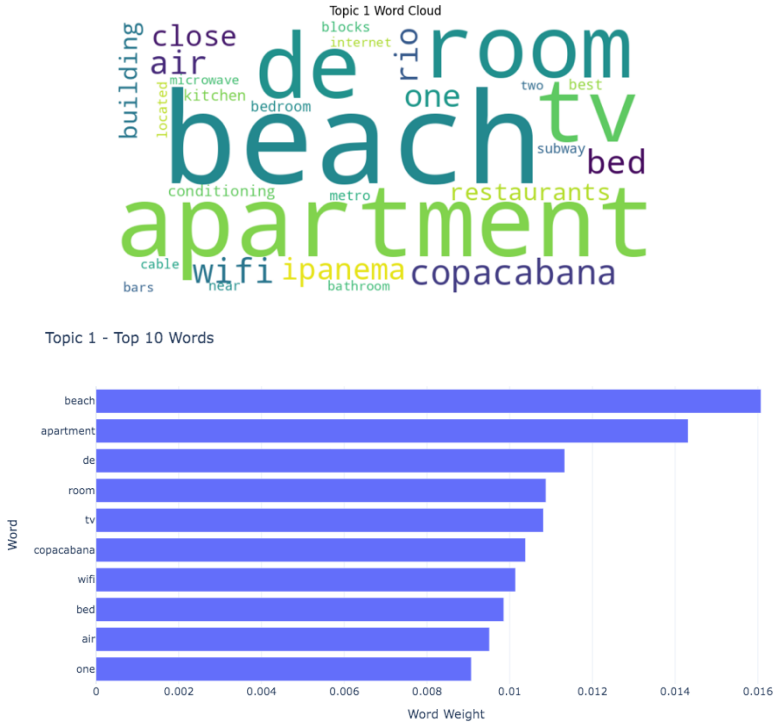


Figure 23. Topic 1 Word Cloud

- Topic 1: Focused on space and layout, with terms such as apartment, bedroom, kitchen.

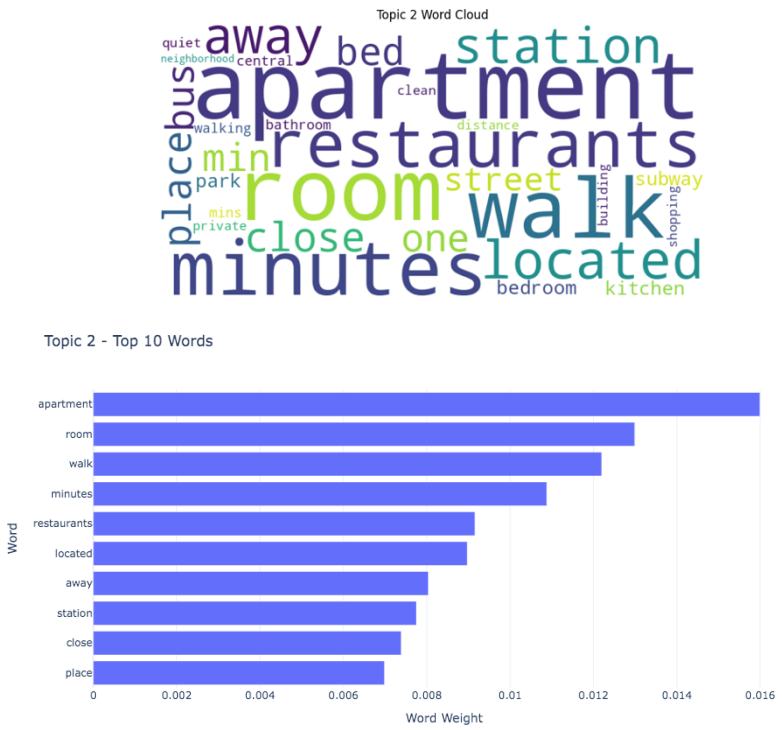


Figure 24. Topic 2 Word Cloud

- Topic 2: Highlighted listings emphasizing exact location and transit access (minutes, metro, street).

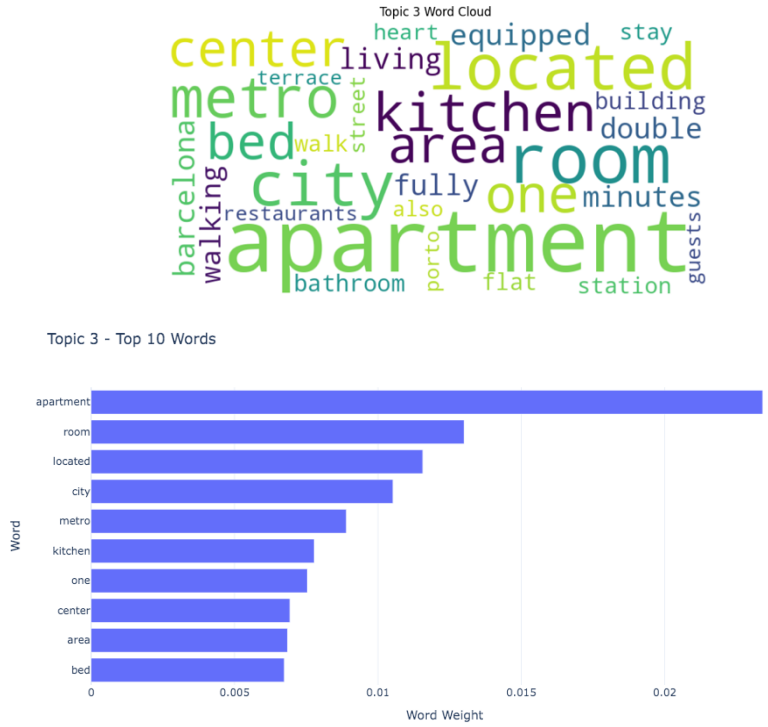


Figure 25. Topic 3 Word Cloud

- Topic 3: Related to urban apartments, including apartment, kitchen, city.

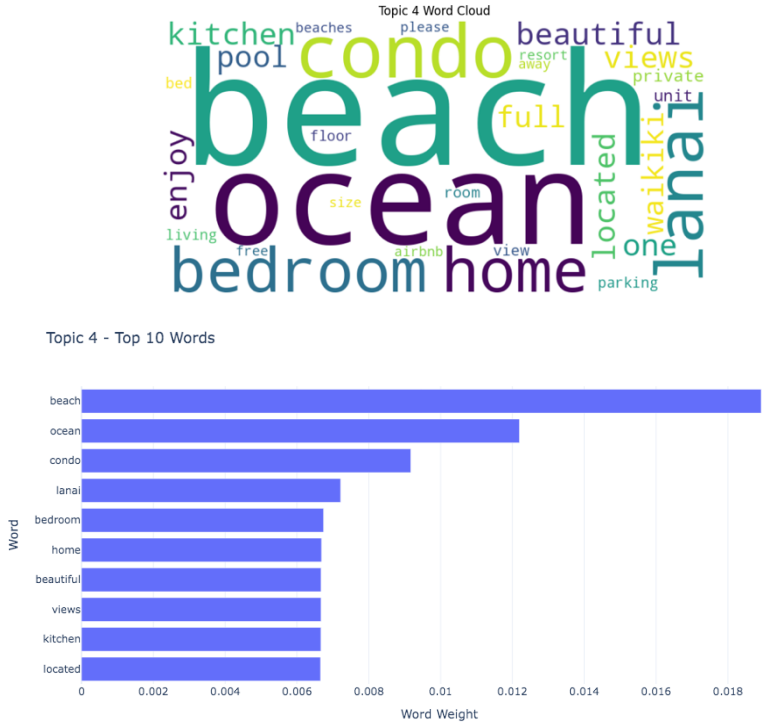


Figure 26. Topic 4 Word Cloud

- Topic 4: Captured vacation-style listings, featuring beach, ocean, views.

The topic with the highest number of listings was Topic 0 (1,255 entries), while Topic 2 had the least (530 listings), indicating varied focus areas across listings.

LDA successfully extracted meaningful themes from the listing texts, revealing how different hosts emphasize distinct selling points ranging from transit-friendly locations to beachfront views or room layout. This thematic structure can help segment listings for personalized recommendations, identify trends in property marketing, or support content-based search filters. The clear topical separation, even across subtle listing descriptions, highlights the effectiveness of LDA in text mining real-world property data.

3. VADER Sentiment Analysis

This analysis applied the VADER sentiment analyzer to quantify the emotional tone of Airbnb listing descriptions. Each description was scored using VADER's compound sentiment scale (ranging from -1 to +1) and categorized into sentiment labels: positive, neutral, or negative. Additionally, I calculated sentiment entropy to assess the diversity of sentiment labels across listings, and the standard deviation of sentiment scores to measure variability.

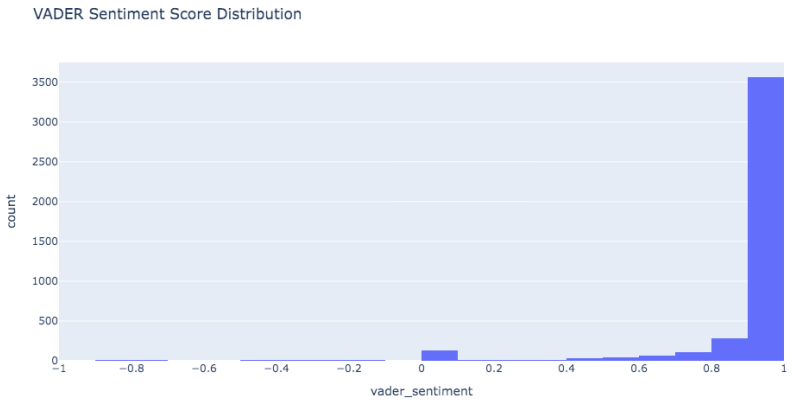


Figure 27. VADER Sentiment Score Distribution

Finally, I visualized how sentiment is distributed overall and across the five text-based clusters previously generated using TF-IDF + KMeans + t-SNE.

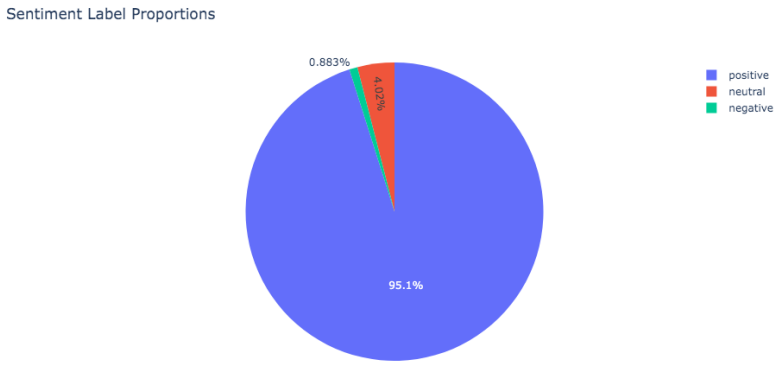


Figure 28. Sentiment Label Proportions

The sentiment score distribution is heavily skewed toward the positive end, with most listings scoring above 0.9. According to the pie chart, 95.2% of listings are labeled as positive, with only 3.95% neutral and 0.88% negative. The low entropy value (0.216) and low standard deviation (0.253) confirm that sentiment is highly consistent and overwhelmingly positive, reflecting the persuasive nature of listing descriptions in hospitality platforms.

Boxplots comparing sentiment scores across clusters show all groups maintain a strong positive tone, though slight differences in dispersion suggest some variation in expressive style. These findings emphasize that while listings differ in content or structure, the overall language tone remains consistently upbeat, an insight useful for content standardization, tone monitoring, or ranking listings based on emotional appeal.

4. Text Clustering

In this section, I applied a text clustering pipeline to group Airbnb listings based on their description content. The process involved transforming cleaned_text into numerical features using TF-IDF (with unigrams and bigrams), reducing dimensionality using Truncated SVD to improve performance and clustering accuracy, and grouping the listings using KMeans into five clusters. The clusters were then visualized in 2D using t-SNE, and their quality was evaluated using the silhouette score.

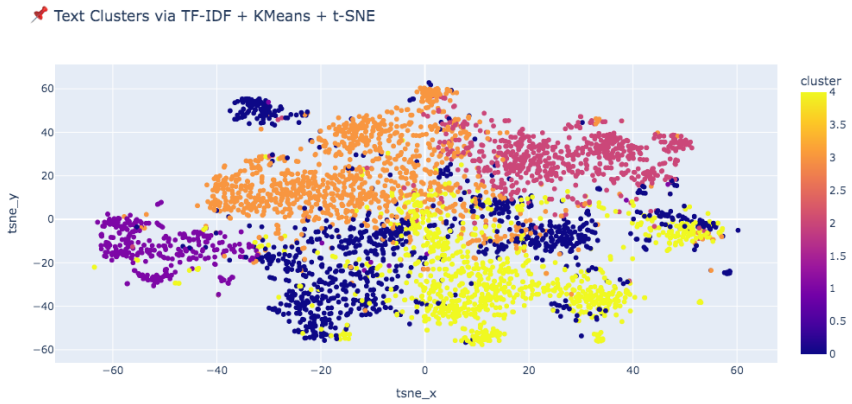


Figure 29. Text Clusters via TF-IDF + KMeans + t-SNE

The clustering process resulted in five distinct groups, each representing similar patterns in textual content. The t-SNE visualization showed reasonably well-separated clusters, though some overlap was observed. The silhouette score of 0.035 suggests mild but present clustering structure, indicating that listings share common themes while still exhibiting subtle differences in how they are described.

While the clustering did not produce sharply separated groups, it revealed latent textual groupings based on how hosts market their properties. This could reflect underlying differences in property types, featured amenities, or stylistic tone. These clusters provide a foundation for building content-based filtering systems, personalized search, or deeper profiling of listing segments based on language usage.

Appendix H. Recommendation

Based on the insights derived from this project, Airbnb can take several strategic steps to improve listing quality, guest engagement, and platform efficiency. First, Airbnb should consider developing a content optimization tool integrated into the host interface. This tool can recommend high-performing keywords (e.g., private room, fully equipped, central location) identified through TF-IDF analysis, helping hosts craft more effective and searchable descriptions.

Second, the LDA topic modeling results can be used to refine search filters and improve content tagging. By recognizing the dominant themes in listing descriptions such as proximity to public transport, vacation appeal, or room configuration, Airbnb can enhance its personalization engine and better match guests with listings aligned to their preferences.

Third, sentiment analysis revealed that the vast majority of listings are overwhelmingly positive, which, while aligned with marketing goals, reduces differentiation. Airbnb could provide language tone suggestions or highlight standout writing styles to promote more authentic and diverse expressions, improving guest trust and listing uniqueness.

Finally, the text clustering results offer a valuable segmentation framework. Airbnb can group listings by writing style or focus area, allowing for targeted marketing campaigns, localized promotions, or comparative performance analysis among similar listings. These strategies will empower both hosts and the platform to deliver a more engaging, relevant, and competitive experience.

Appendix I. Conclusion

This project demonstrated the power of text mining and machine learning in extracting meaningful insights from Airbnb listing descriptions. By combining traditional data analysis with advanced NLP techniques such as TF-IDF weighting, VADER sentiment analysis, topic modeling (LDA), and text clustering (via KMeans and t-SNE), the study revealed both the consistency and variability in how hosts communicate value.

Key findings include the heavy reliance on positive language across listings, the identification of five recurring content themes, and the existence of distinct descriptive clusters despite similar sentiment profiles. These insights highlight not only what hosts are saying, but how they say it and how that language can be optimized to better serve guest needs and improve listing visibility.

Overall, the analysis reinforces the importance of structured, data-informed content strategies in enhancing Airbnb's user experience, search relevance, and overall platform performance. By acting on the recommendations provided, Airbnb can drive greater engagement, boost conversion rates, and maintain its competitive edge in the evolving travel marketplace.