Text Mining Analysis Report

Executive Summary

The analysis explores the Airbnb listings descriptions to identify customer segments, key resonating words, and price indicators. By tokenizing and structuring the text data, three main customer segments are identified:

1. **Couples** – Prefer vacation homes with luxurious features like swimming pools, valuing great views, and appreciating privacy.
2. **Solo Adventurers** – Prioritize destination exploration over luxury, choosing practical stays close to sites and spots of interest.
3. **Business Travelers** – Value transportation convenience, staying close to subway stations, airport, or important business hubs.

Generally, the most frequent words across all listings emphasize convenience and practicality, including “walking distance”, “washer & dryer”, and “air conditioning”. A correlation analysis between the description and price uncovered drivers of listing prices:

- **Renovation & Modernity** – Words like “recently renovated” and “newly furnished” are related to higher prices listings
- **Hot Water & Heating** – Especially value-adding for Solo Adventurers
- **Scenic Views** – Listings highlighting “great views” influence pricing, resonating with Couples.
- **Transport Accessibility** – Keywords like “station exit”, “bus stop”, “subway station” are linked to higher prices, aligning with priorities of Business Travelers.

Hosts can target their accommodations to customer segments by specifically tailoring the wording. Airbnb on the other hand can refine marketing efforts, promoting listings based on segment-specific priorities. Further analysis of the customer reviews could generate more insights.

Report

Before conducting any analysis, the description column in the Airbnb dataset is filtered for English words using the GradyAugmented.

Sentiment Analysis

The overall sentiment of the words used in the description of a listing is predominantly positive, described with terms like “joy”, “trust”, and “anticipation”. The average sentiment score (AFINN score = 1.677) confirms that descriptions tend to use optimistic language. This aligns with the business object of making listing appealing and engaging for users.

Customer Segmentation, Preferences, and Commonalities

The analyzed listings target three main customer segments: Couples, Solo Adventurers, and Business Travelers. The types of accommodations and destinations differ between these segments. This insight is identified using quadrograms (four-word sequence), while bigrams (two-word combinations) provide further analytical insights and context.

Couples

- Prefer vacation-oriented stays, oftentimes close to a beaches or vibrant cities like Hawaii or New York City.
- Enjoy luxurious features like great views, pools, grills, and appreciate privacy.
- Value short distance to restaurants and bars, prioritizing well-maintained amenities.

Solo Adventurers

- Put their emphasis on the destination rather than luxurious features.
- Prioritize flexibility while being ready to pay a premium for features like “hot water”.

Business Travelers

- Prioritize convenience of transportation, staying close to subways, bus station, or airports.
- Most commonly travel to major cities like New York City.

Beyond segmentation, certain words appear frequently across all listings, emphasizing general selling points. “Walking distance” is the most used English bigram in the description of accommodations. Furthermore, other high frequency words mostly describe appliances and amenities like “washer & dryer”, “air conditioning”, and “iron board”. Overall, more general descriptions like the bed size, the parking situation, and shopping opportunities are mentioned in most cases, seen on the high frequencies in bigrams and quadrograms.

Price Correlation

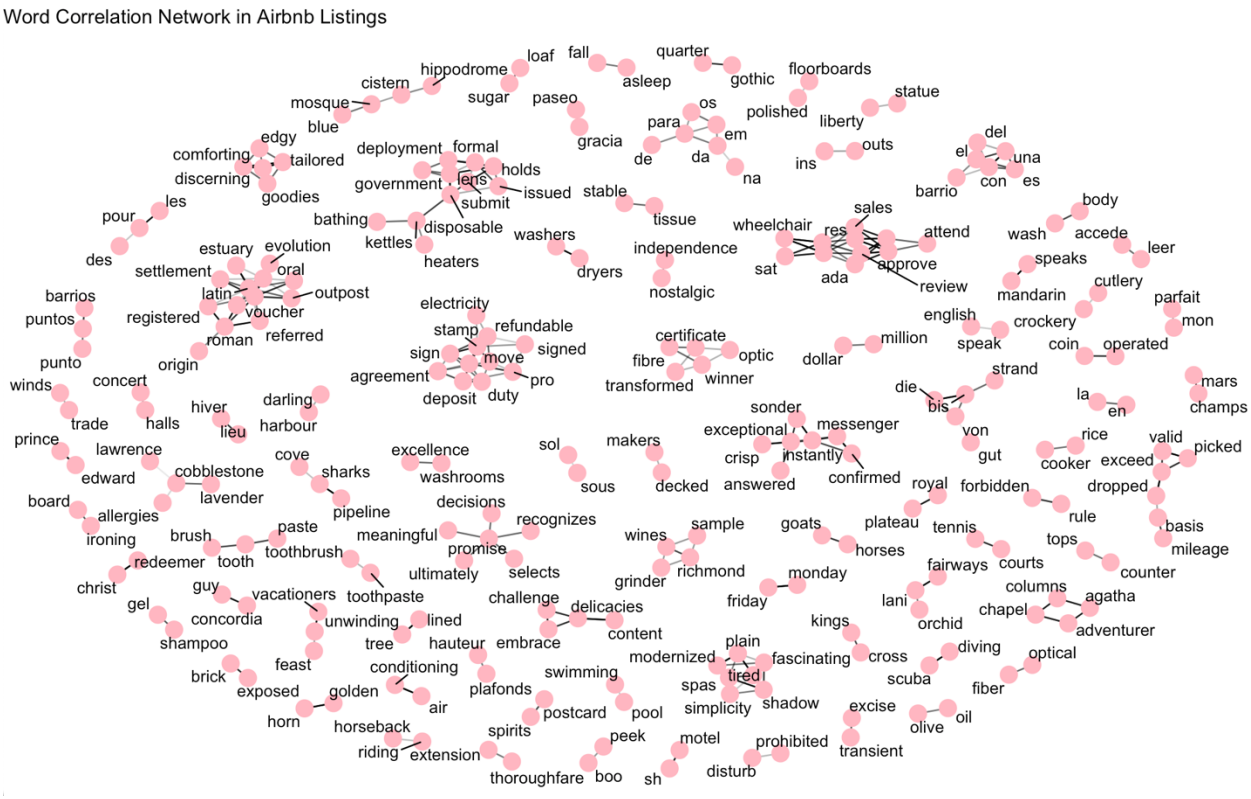
The analysis examines the correlation between words in listing descriptions and its price to identify key drivers. For this part , bigrams are used again to provide some contextual meaning while not losing the analytical power due to chaining too many words, e.g. with quadrograms.

Several key themes are identified:

- **Renovation & Modernity**: Words like “recently renovated” or “newly furnished” indicate that an accommodation tends to have updated appliances and interior.
- **Essential Amenities**: Terms like “hot water” and “water heater” drive the price, presumably because this feature is especially relevant for the Solo Adventurer segment who are ready to pay a premium for the convenience of having access to hot water.
- **Scenic Views**: Highlighting good views, regardless of whether it is in a city or at a beach, drives the price, appealing to travelers looking for aesthetic value.
- **Transport Accessibility**: Words such as “station exit” or “bus stop” show a positive correlation with price, emphasizing the importance of transportation convenience – especially for Business Travelers.

Appendix 1

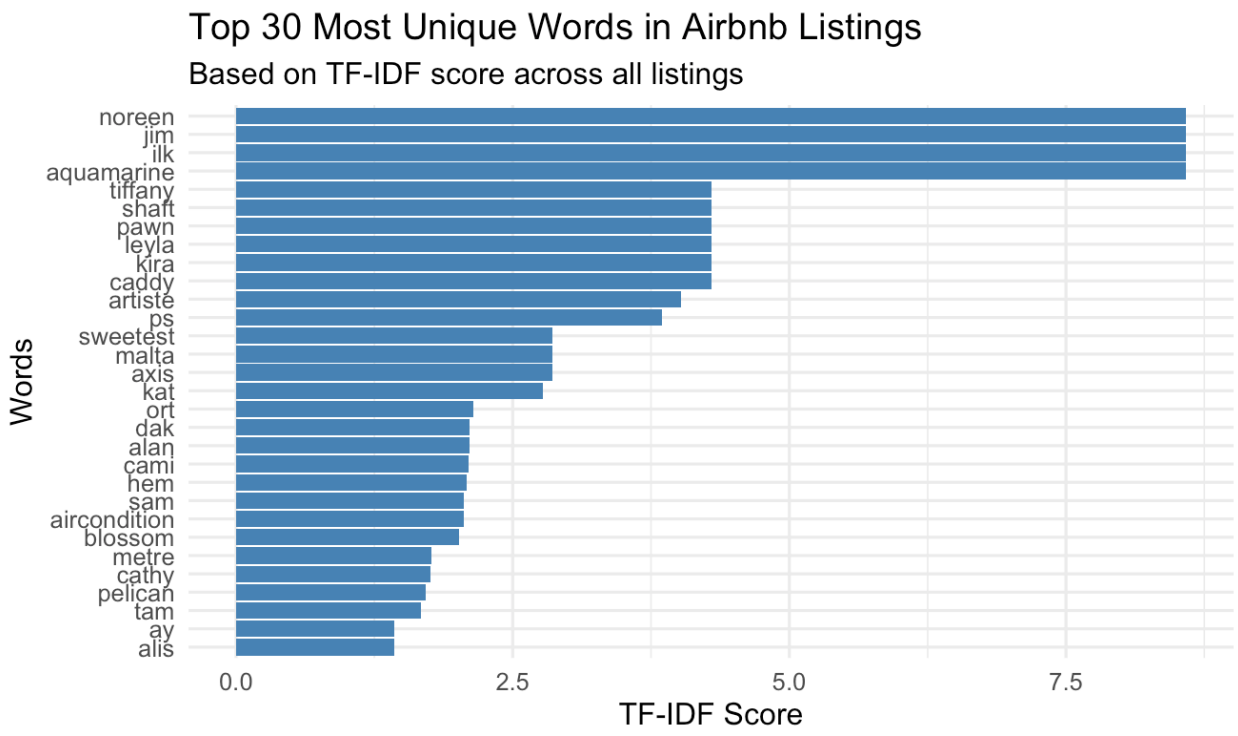
Correlation Network



The correlation network shows existing positive correlations between individual tokens. To keep computational efforts efficient, the subset used for this visualization considers the 200 tokens with the strongest correlations. The network shows tokens that occur oftentimes together. Some of these are common sense, e.g. air+conditioning, scuba+diving, horseback+riding, body+wash, or swimming+pool. Furthermore, house rules and compliance related words oftentimes appear together, including disturb+prohibited, forbidden+rule, and speaks+mandarin. Though, some relationships are revealed that are not so obvious from the beginning, e.g. independence+nostalgic, quarter+gothic, redeemer+christ, or darling+harbor. These relationships can hint to certain destinations or specific style of accommodation for a niche customer segment.

Appendix 2

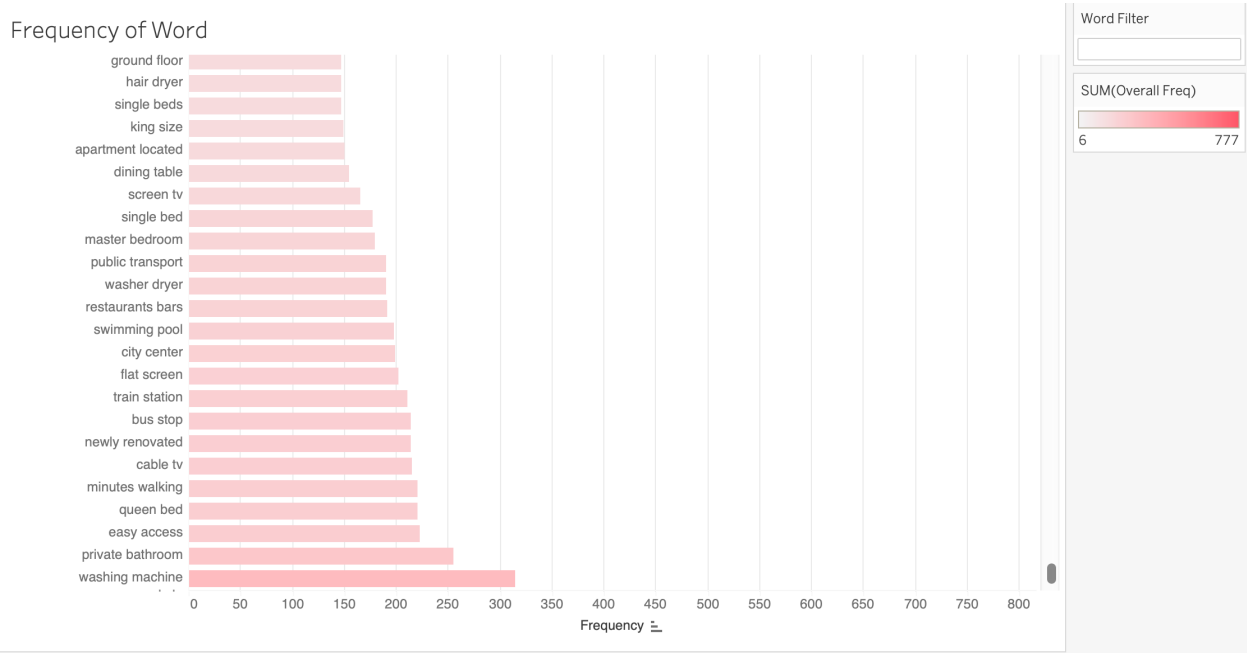
Term Frequency – Inverse Document Frequency



Opposite to the frequency bar chart, the TD-IDF expresses the uniqueness of a word. It shows that the most unique words are words like names, or words in a different language while being mentioned in an English description. For the case of presence and searchability of listings, this methodology can bring insights when being applied to a specific, predefined subset of listings under a certain theme. In the context of this analysis, which takes a more general perspective on the dataset, the TF-IDF does not drive beneficial insights.

Appendix 3

Word Frequency Bar Chart

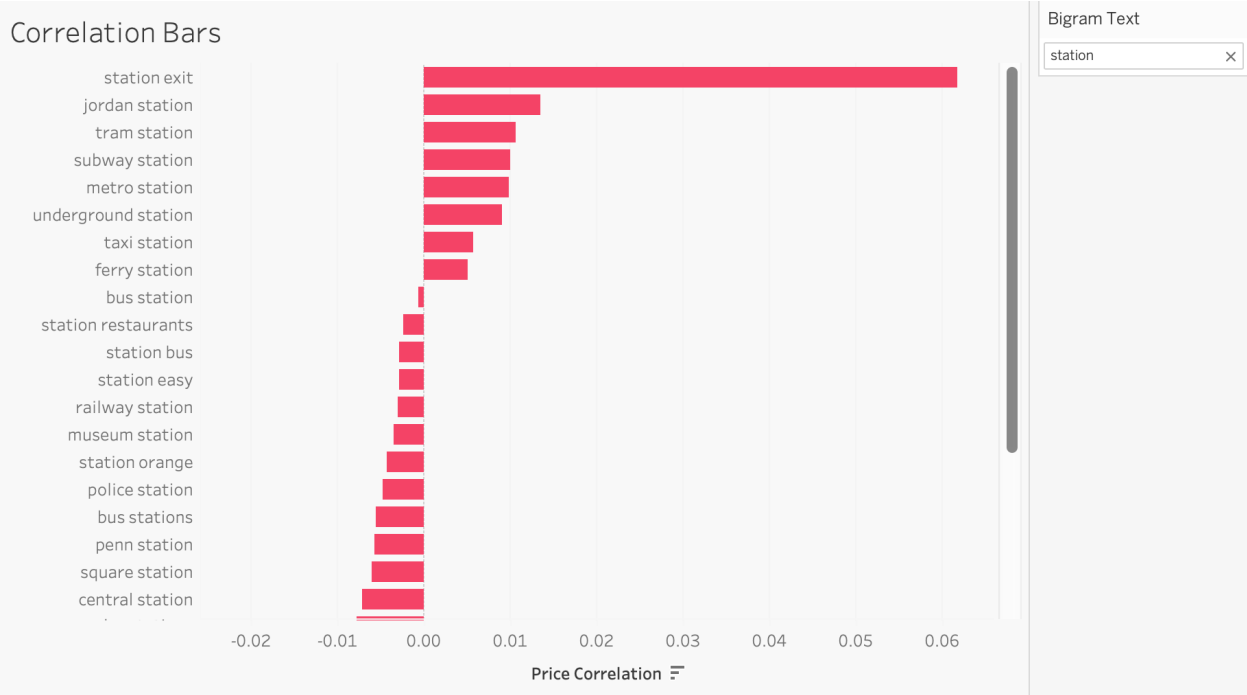


This bar chart is displaying how often a bigrams appear in listing descriptions. It is sorted ascending to show the least frequent – or the most unique – word combination first. Using this approach, deeper insights on less frequent used bigrams can be driven.

In a next step, when combining this visualization to others in a dashboard, insights on the relationship between the frequency of a word and its correlation to a listing price can gained.

Appendix 4

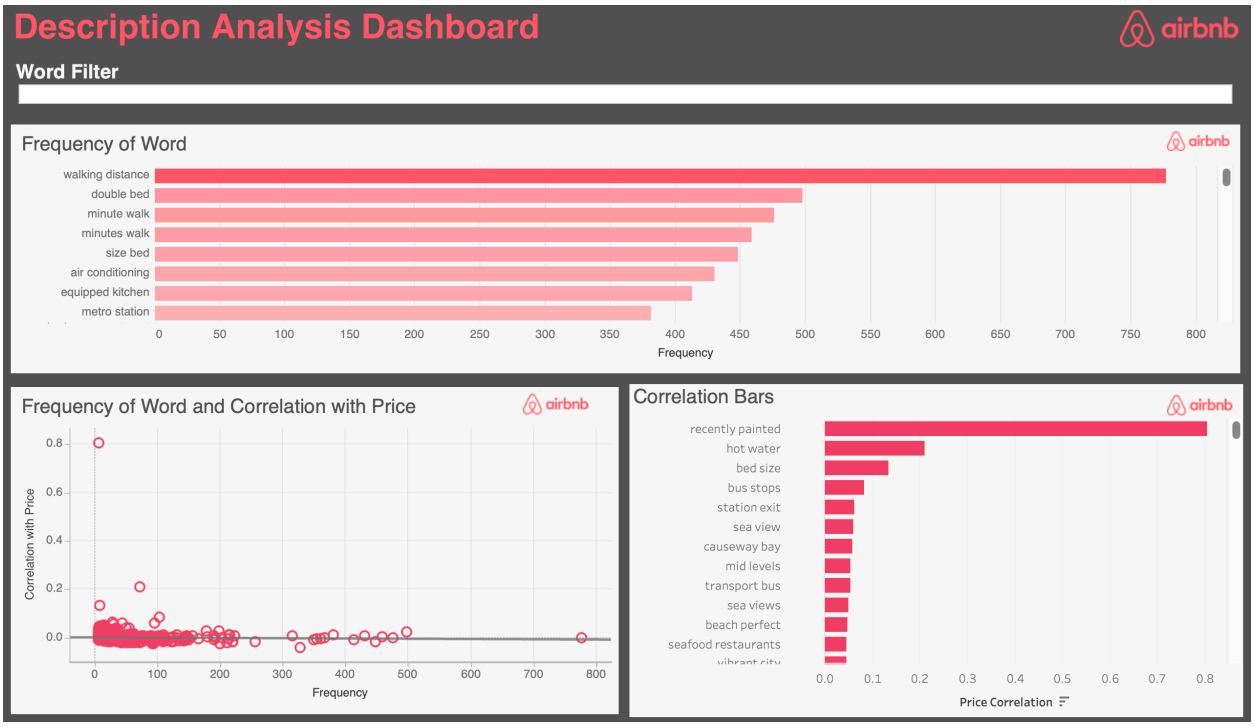
Correlation Bars



The correlation bar chart shows the strength of a correlation per bigram. In this case, it is filters to the word “station”. It shows that “station exit” is positively correlated with the price of a listing while “police stations” are negatively correlated. Even though these correlations are very weak, they give an indication on what semantics drive potentially drive the price of a listing.

Appendix 4

Dashboard

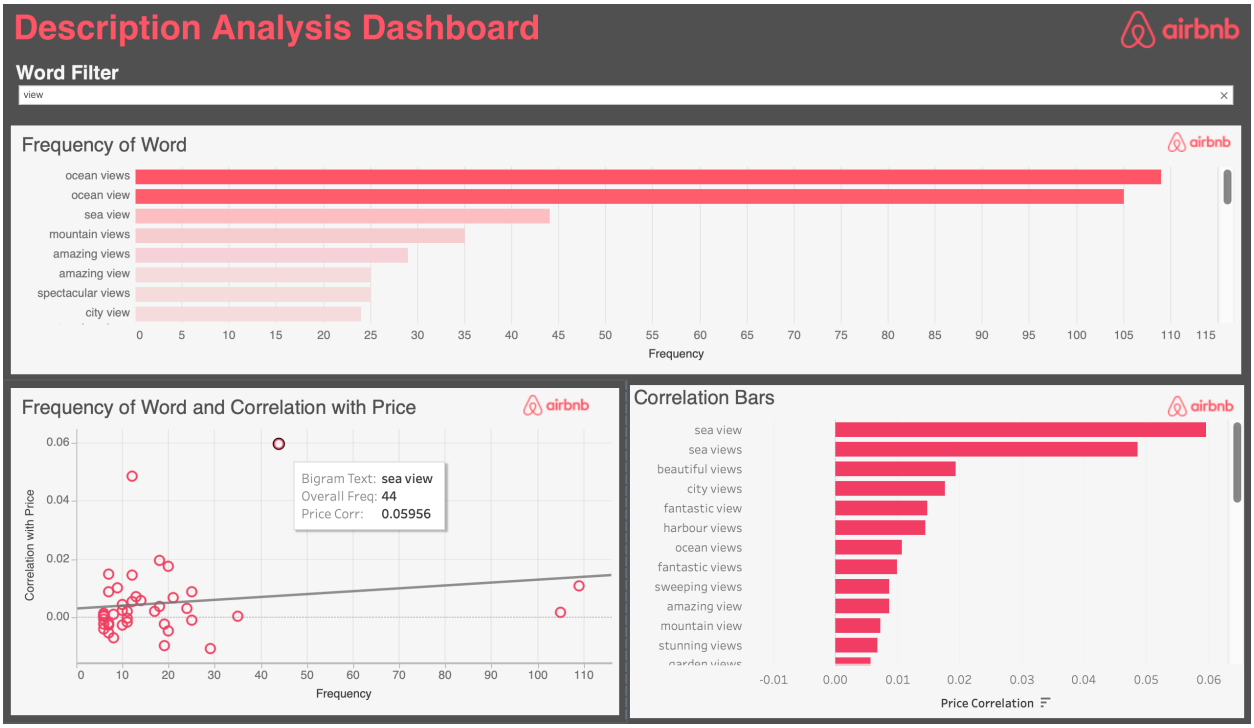


This Description Analysis Dashboard shows three visualizations.

1. A bar chart of the frequency of the bigrams a specific word is occurring in
2. A scatterplot showing the relation between frequency of the bigrams a specific word is occurring in and its correlation with the price of a listing
3. A bar chart to showcase which bigram has the strongest correlation with price

On top, the “Word Filter” field lets the user filter for a specific word. The search is a wildcard match, meaning the word is included in the bigram in some way. With this method it is ensured that all bigrams this word occurs in are shown and that plurals of words are included.

In this example, the dashboard was filtered for the word “view”. All visualizations automatically now filter to all bigrams that include the word “view”.



This gives the user a good overview in what constellation the word “view” is used the most and which bigram is the strongest driver of the price of a listing. In this case, “ocean views” is the most frequent bigram while “sea view” has the highest correlation with price. The scatterplot shows that not necessarily all listings including “view” are positively correlated with price. Therefore, specifically tailoring the words for the customer segment is important to be able to achieve higher pricing.

Since Airbnb is operating on revenue per booking based on the listing price, this insight can be used to make suggestions to users that are related with higher priced accommodations.