InsightBnB: Analyzing Descriptions of Airbnb Listings in the United States

The Project uses a comprehensive dataset from Airbnb retrieved from a MongoDB cluster, comprising web-scraped information on 5,555 listings including textual descriptions, pricing, features, location specifics, amenities, and host information. A focused analysis was conducted on descriptions of a subset of 1,218 U.S.-based listings, chosen based on the hosts' geographic location. This subset was further stratified by room type—Entire home/apt, private room, and shared room—and by property type. With more than 26 property types identified, our analysis centered on the top three categories—Apartments, Houses, and Condominiums—which collectively represent 83% of the listings.

Correlogram Analysis

As the first framework, the data was converted into a tidy format for each property type to analyze frequently associated terms with each property type in a correlogram. When words are close to the diagonal line, it suggests that the frequency of those words is similar across the two property types being compared. The focus is more on word clusters that indicate how hosts are positioning their properties on Airbnb that are most appealing to short-term renters: **Apartments:** mention of specific city locations such as "nyc," "brooklyn," "manhattan," and terms like "train," "bars," and "cozy" suggest that Airbnb guests are looking for apartments in the city with an inviting atmosphere that's suitable for short stays with convenience, accessibility to public transport, and the proximity to nightlife and city attractions. **Houses:** words like "beach," "family," "home," "ocean," and "sand" imply that Airbnb listings for houses promote a family-friendly environment. They're ideal for families seeking a home away from home, in a more natural setting with beach access. These listings could be in higher demand for extended stays or vacation periods where guests are looking to relax and unwind. **Condominiums**: The clustering of “pool," "resort," "lanai," "beachfront," "maui," and "oceanfront" words indicates that condos on Airbnb are often positioned as relaxing getaways, perhaps with an emphasis on leisure destinations and are targeted toward guests looking for a premium stay. "Lanai" and "Maui" point to a location that is common in the descriptions for condos.

Term Frequency – Inverse Document Frequency (TF- IDF) Framework

Subsequent analysis utilized the TF-IDF statistic to highlight the top 15 words critical to the identity of each property type within the collection and provide insight into the unique selling propositions for each type of listing. **Apartment:** Predominantly characterized by locational terms like "broadway," "soho," and "greenpoint," the listings emphasize distinctive urban traits. Words like "apt" and "roommate" suggest shared living spaces or rental units, likely resonating with temporary city explorers. **Condominium:** The chart for condos features several Hawaiian place names such as "wailea," "anapali," and "kapalua," along with words like "villa," "grills," and "yacht." These suggest a strong association with vacation spots and luxury, indicative of a preference for indulgent amenities and outdoor recreational options. **House**: Dominated by words like "cottage," "yard," "bungalow," and "acre," which convey a sense of space and privacy. Additionally, "volcano" and "rainy" could indicate unique environmental features or climates of the areas where these houses are located.

Visualizing a Bigram network by applying the TF-IDF framework

Focusing on the Apartment property type for this visual, the TF-IDF bigram network provides a textured view of apartment listings, highlighting unique and high-value features. Bigrams enhance semantic understanding by contextualizing word pairs, giving depth to property descriptions. For example, highlighting the proximity to city landmarks ("central" and "park"), well-known neighborhoods ("upper," "east," "lower," "village," and "west"), transportation ("subway" and "lines"), and specific desirable features of the apartment ("private bedroom", "newly renovated" and "hardwood floors") could be a strategy to attract guests looking for urban apartment stays.

Sentiment Analysis

In sentiment analysis, lexicons are dictionaries that associate words with emotional scores, often categorized as positive, negative, or neutral. For a balanced comparison, the NRC sentiment analysis has been filtered to only consider words with positive or negative sentiments, aligning with the type of sentiments AFINN and Bing et al. capture. All three lexicons suggest that the apartment listings have a

generally positive sentiment, implying that hosts are using positive language to describe their properties, potentially attracting more guests. The key takeaway here is that the sentiment scores are different across lexicons, which is due to the different number of sentiment words they recognize or the intensity of the sentiments they assign to each word. AFINN provides scores that range from -5 (very negative) to +5 (very positive), and suggests that the positive words have a higher intensity. Bing et al. categorize words simply as positive or negative implying more frequent positive word usage in the listings. NRC lexicon includes a range of emotions but here has been filtered to only consider positive and negative sentiments, indicating that positive sentiment words have stronger positive associations in the listings.

Business Insights

Hosts and Airbnb itself can leverage these insights to tailor their listings and search algorithms. For urban apartments, highlighting the cultural vibrancy and convenience of the location can attract guests looking for city breaks. For houses, emphasizing the suitability for families and the appeal of nature can attract guests planning leisurely vacations. For condos, focusing on luxury amenities and the vision of an exclusive experience can cater to those willing to pay more for comfort and leisure.

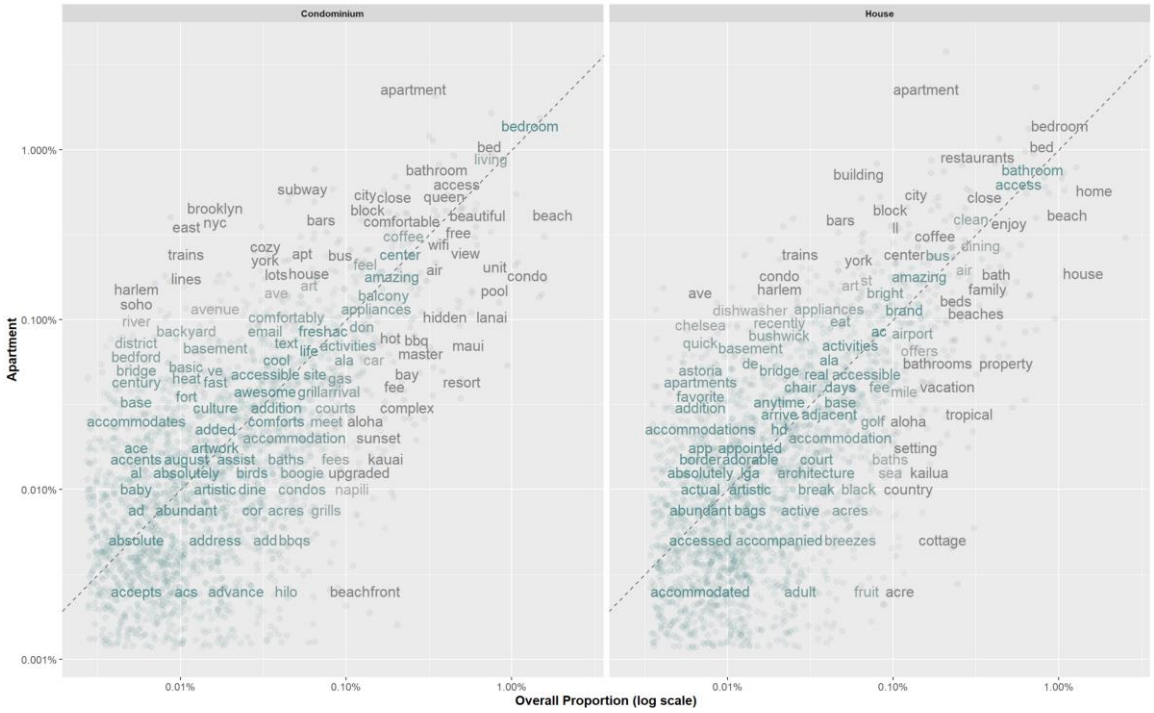
Dashboard

The dashboard designed for this analysis offers a user-driven experience, showcasing business insights derived from the textual content of property listings. Through interactive inputs, users can delve into a variety of visual data presentations. Correlograms provide a visual exploration of word frequency and relationships, enabling users to discern which terms are commonly associated with different types of properties and room categories. TF-IDF Analysis to pinpoint distinctive words that carry significant business value, filtering through the noise for different property and room types. Key Value boxes adjust dynamically based on user inputs, summarizing key information such as the count of listings, average price, and typical number of guests accommodated, within chosen categories. TF-IDF Bigram Graphs, reveal high-value features and amenities specific to property listings, enriching semantic understanding. Finally, sentiment analysis via three distinct lexicons, uncovers the overarching sentiment.

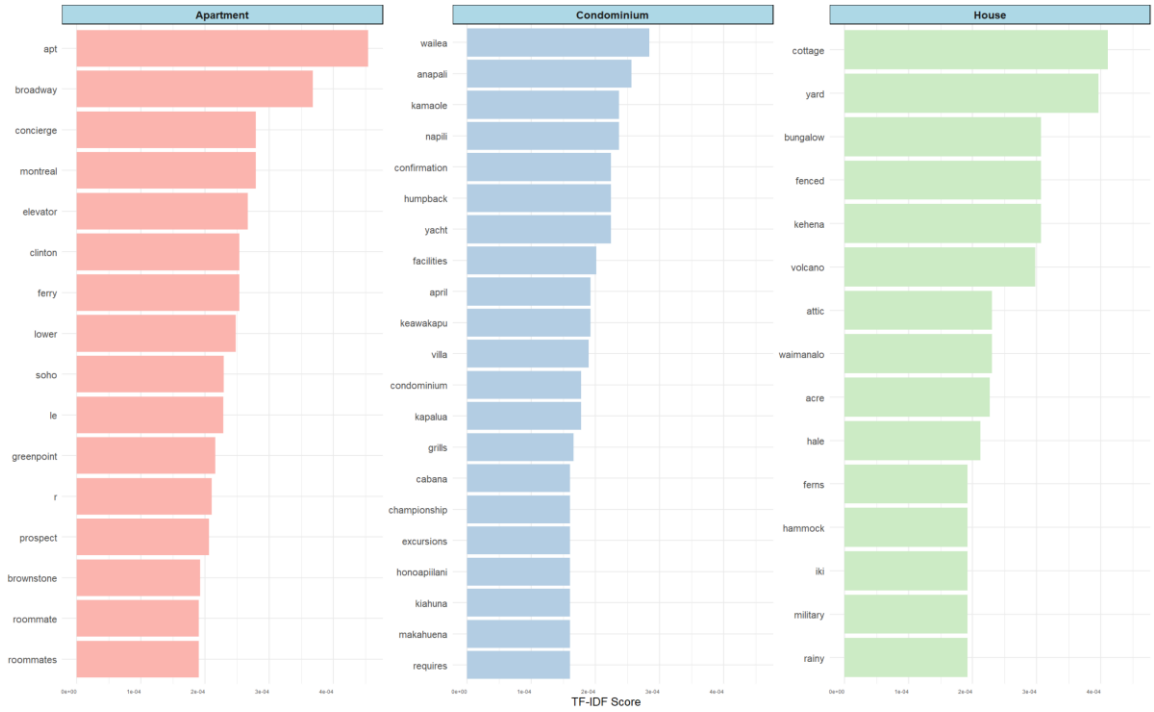
Appendix

A. Visuals

1. Correlogram Analysis Visual

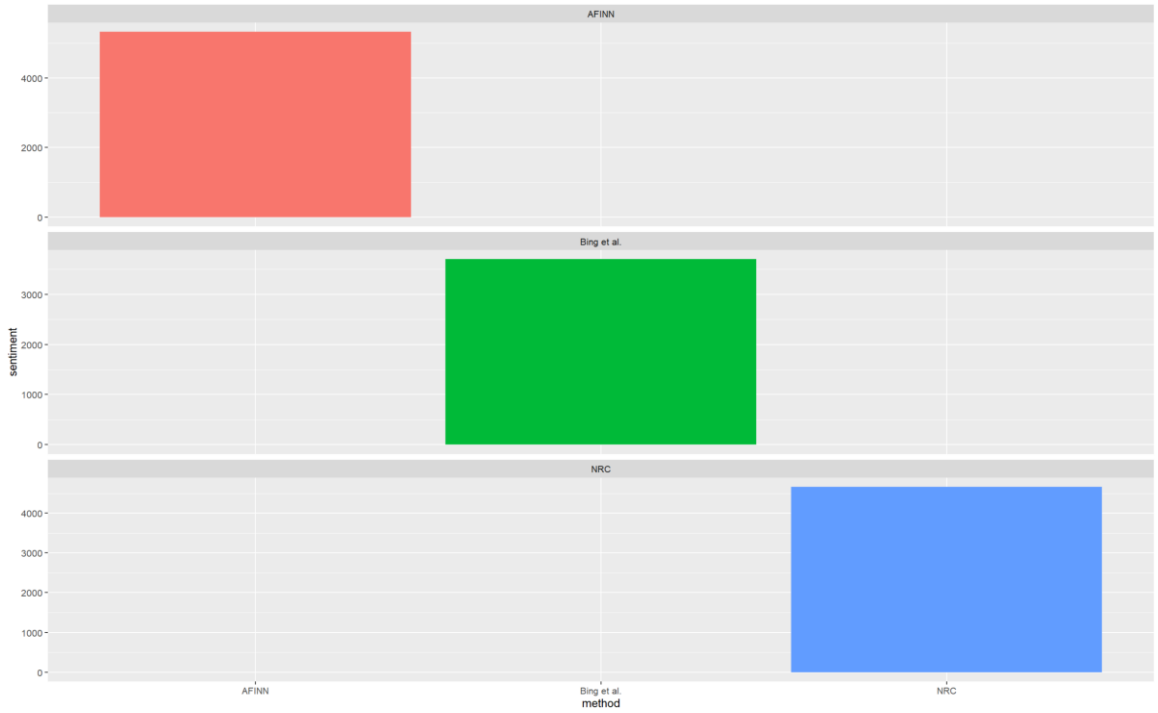


2. TF- IDF Framework



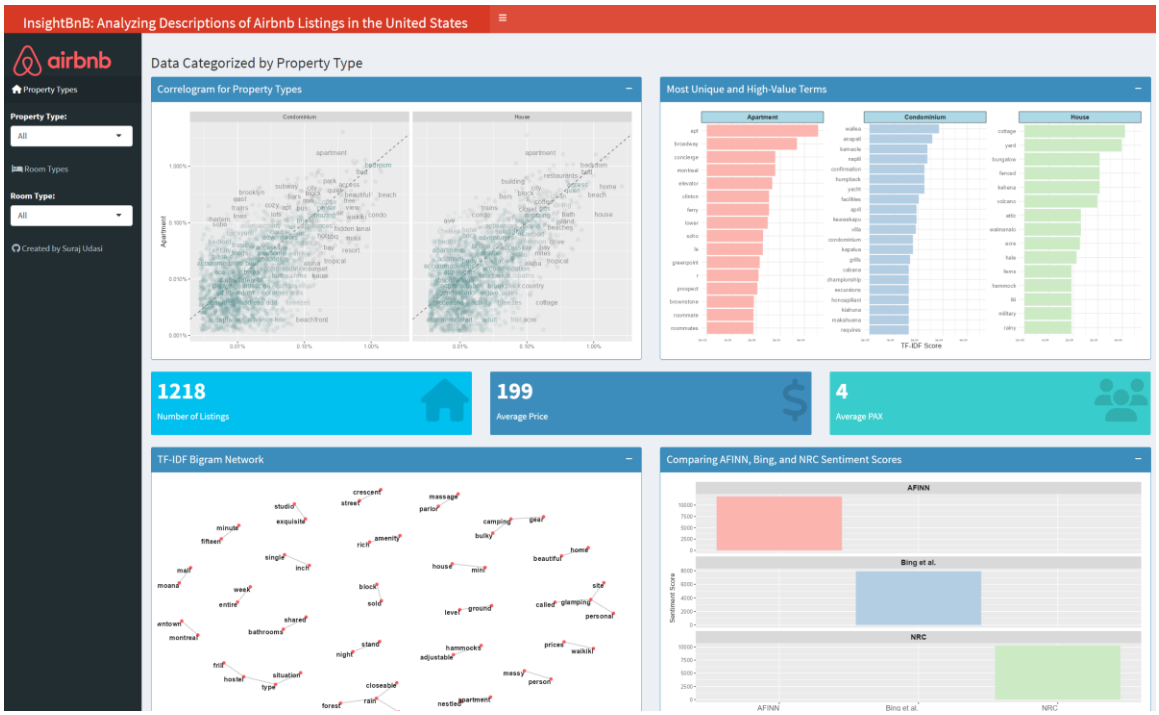
3. Visualizing a Bigram network by applying the tf_idf framework

4. Sentiment Analysis



B. Dashboard

1. Tab 1 – Property Type



2. Tab 2 – Room Type