Data Visualization Final Group Project Process Book

Group P members: Gerald Lee, Joellyn Heng, Kyung Suk Lee

Project Title

Airbnb PriceR: Airbnb Analytics

Data Description

Airbnb: Our dataset consists of Airbnb listings and reviews from 2009 to 2019. The dataset was obtained from Kaggle, with a pre-processed version that can be found on InsideAirbnb.com (http://insideairbnb.com/london/). The data on listings comprises over 90 columns (e.g. customer ratings, room price, location) and approximately 85,000 host listings in London, UK from 2009 to 2019. For the reviews dataset, which comprises plain-text reviews from Airbnb, we have applied text cleaning and preprocessing steps to the plain-text reviews. These steps include removing punctuations, removing numbers, making all text lowercase, removing stop words and stemming the words.

London Tube (London Metro System): Our data on the London tube stations was obtained from "Transport For London" (https://tfl.gov.uk/). The London Underground is a public rapid transit system serving the London region. The dataset comprises approximately 470 stations with geographic locations.

Purpose of Website

We decided to create this website as a tool that can be used by Airbnb hosts to better understand their potential competitors and customers. Although there are numerous analyses that are targeted towards Airbnb users, we thought information that aids existing hosts is rare. Thus, for hosts with established listings, by giving sense of how the price differs across various dimensions (e.g. location, room types, number of guests etc.), we hope that hosts can use this website for a pulse-check on whether they remain well-positioned for competition. Furthermore, for anyone who wishes to become a host and list their property for the first time, this tool would be very useful for a first-cut understanding of the competitive landscape for market entry and pricing strategy.

Our visualizations are broken down into the sections below, keeping in mind the needs of our target audience. Apart from being able to visualize their competitors' locations and pricing at a quick glance, they will also be able to visualize analytics on supply, prices, ratings etc. at the selected neighbourhood- and listing-specific levels. We also analysed cleaning fees as a component of total listing price, in order to better inform pricing strategy

of hosts. Finally, we also provide some sentiment analysis through ratings and reviews of customers in the neighbourhood, in order for hosts to have a sense of how they are faring vis-à-vis their nearest competitors. Airbnb users and travellers are also welcome to use the website for the same information.

Section 1: Listings

Here, we have an interactive map that allows users to view the available listings across London. Users can adjust the slider on the left to filter based on their neighborhood, number of guests, and type of lodging in order to produce matching results. Users can also choose to overlay subway stations on the map to find out how far the nearest subway station is to specific listings. By hovering on the dots, users will be able to see the price of each listing, and when they click on them, more details including rating and direct links to the listing are provided.

Section 2: Explore Listings

This section provides an alternative view of the listings based on the same filters that the user has chosen. Instead of a map, it contains a data table that allows users to observe the price, property type, review scores and the direct link to the listing.

Section 3: Neighborhood Analytics

In this section, users are able to see two sets of visualizations, one on the selected neighborhood and another on comparable listings within the neighborhood.

The first set comprises two graphs that take into account all the input filters chosen by the host, and provide analyses of close comparables within the neighborhood. The first graph is scatterplot that maps the ratings and prices of comparable listings. With the median of both axes in dotted lines, hosts are able to visualize and place themselves among their competitors, and inform them on price and rating strategies. Next, we chose to use a histogram as the second graph to provide a clear distribution of prices of closest comparables.

The next two graphs are aimed at providing hosts an overview of the current supply and price range across lodging types and sizes (i.e. bedrooms and maximum pax). We chose a stacked bar graph, so that hosts are able to see the total units per lodging size, as well as the proportions of types within each size. This would enable them to understand the supply situation within the neighborhood, and assess market gaps and saturation across listing types. Next, we chose to use boxplots to visualize prices, as it could efficiently show the median, quantiles as well as outliers. On the x-axis, we similarly used lodging size, and we

faceted it by lodging type. Hosts are then able to have a sense of the price differentials across these two dimensions.

Section 4: Cleaning Fees
In this section, users can analyze the breakdown between cleaning and room fee. We have chosen to visualize this, as cleaning fee is one other substantial component of total fees. Based on all the input filters, users can observe the breakdown between cleaning and the room fee per bed, and hosts can also be better informed on the average amounts that they should charge as room fee vs. cleaning fee. The second graph is a heat map that shows cleaning fees based on the number of guests, but compared across room types and all neighborhoods. This would provide hosts with an idea of the spread of cleaning fees across room types and all of London.

Section 5: Guest Ratings
Here, users can observe the distribution of customer ratings and polarity scores of reviews derived from sentiment analysis. In the first chart, a histogram highlights the distribution of numerical ratings. For the second chart, the distribution of sentiment scores is displayed. Both charts can be toggled and adjusted based on the neighbourhoods. The polarity scores were derived using the AFINN lexicon and uni-grams from the plain-text reviews. The final 2 charts show word clouds for the top 80 words from the positive and negative reviews. These visualizations can be important for hosts because it highlights where they stand in comparison to others and what are positive or negative attributes, especially from the word clouds they should account for.

Initial ideas and Refinements
One of the questions we were interested to explore was how a new Airbnb host can gain credibility as a newcomer. As an extension, we thought we could explore the behaviour and best practices of Airbnb hosts across multiple segments beyond newcomers, including listings that are popular and established, listings that are consistently rated badly, and listings that have been around but are unable to obtain sufficient reviews. Without the date that the listing was established, we used the date of first review as a proxy. However, we realized we were unable to capture certain important groups of interest, such as those with none or few ratings given the proxy we were using.

Another area we thought to explore was the importance of host indications (e.g. superhost, response rate, response time) on listing prices. The hypothesis was that positive host indications allowed them to price a premium on their listings, as it provided a sense of security to Airbnb travellers. One way we tested this was to plot price against stays per month (using reviews per month as a proxy) across superhost and non-superhosts. If the

negative correlation was weaker in the former, it would mean that prices did not deter stays as much among superhost listings. However, the negative correlation was stronger among superhosts, though not too significantly.

Another analysis we did to test for the price premium for super hosts was a bar chart that compares average fee across superhost and non-superhost (faceted by number of beds).
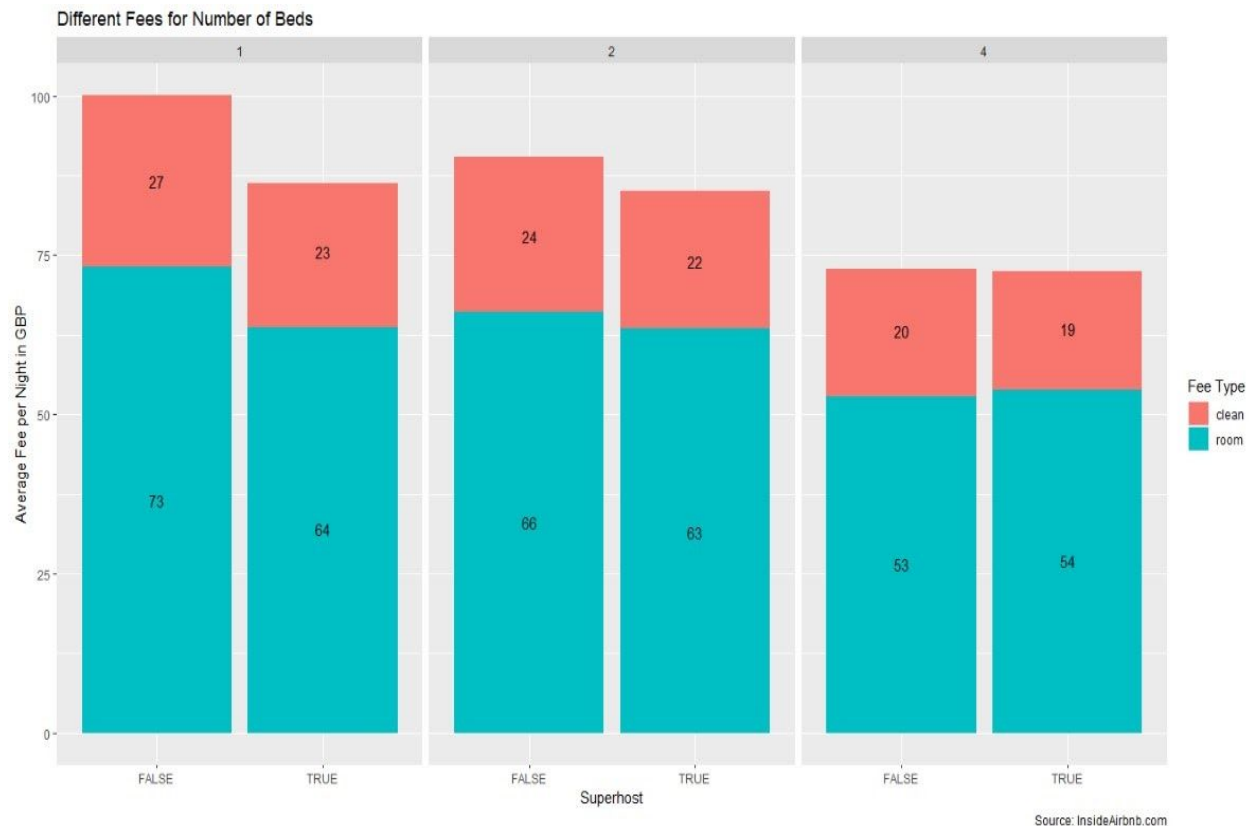

As seen from above, instead of pricing a premium, the average fee for superhosts were rather discounted. We however decided to exclude this in our final results, as we realized the graph did not take into account other features of the listings, such as property type (e.g. villa vs. hostel) and room type (e.g. shared room vs. entire apartment), which may affect prices more, but once we subset the data accordingly in order to compare price premiums, the dataset became too small to be representative.

As part of pricing analytics, one visualization that we attempted but decided to discard is a heatmap of London that shows the average price differences.
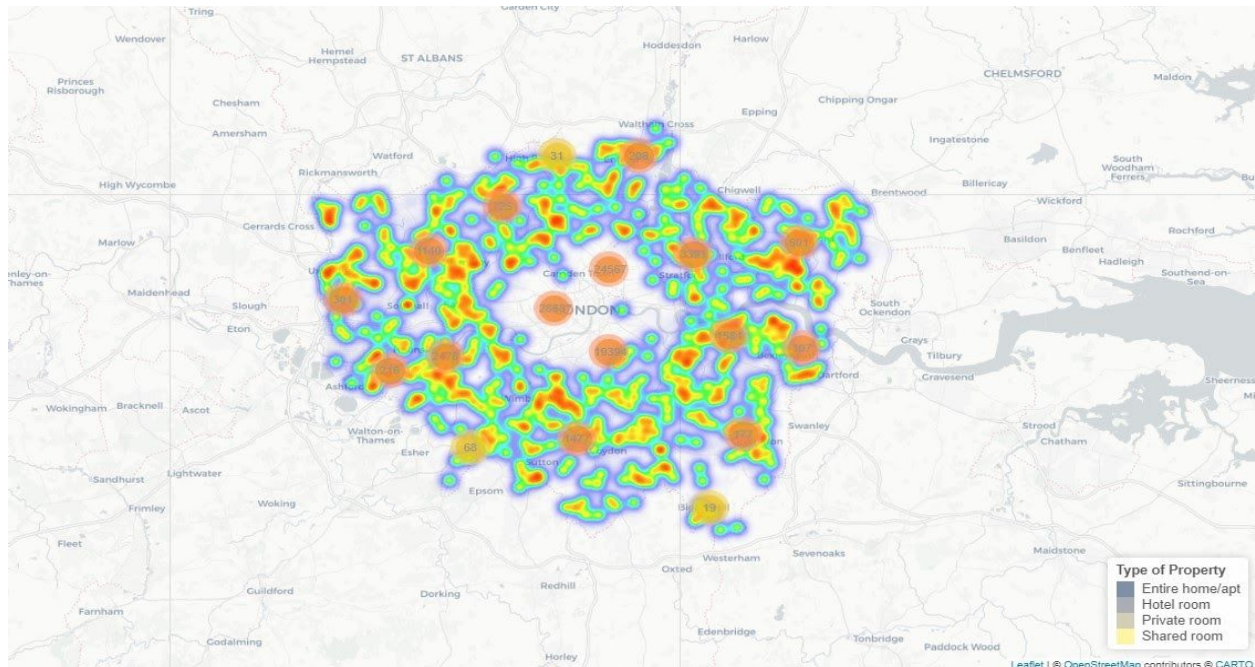
However, as the plots were scattered across the London region, there was no significant pattern observed. Thus, we decided to remove this analysis from the final output.

Also in line with pricing, we wanted to observe how the distance between the listings and the nearest tube station (London Metro System) affects the price of listings.
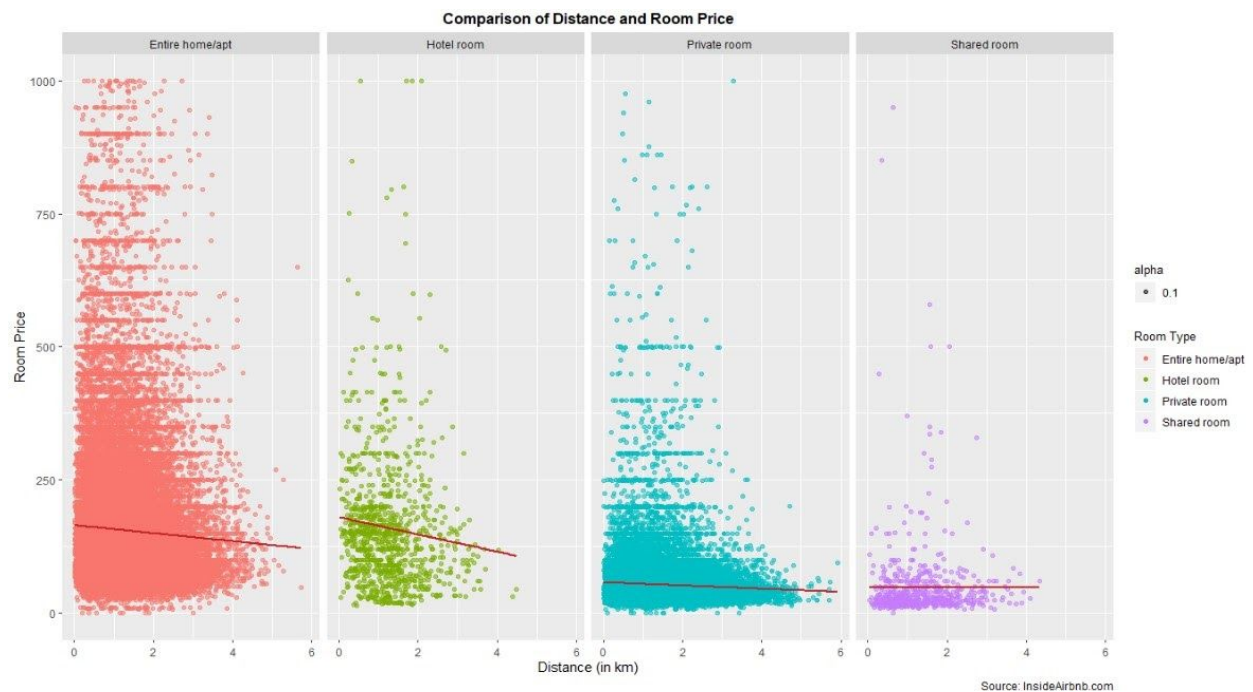
Above shows the average price difference with distance where the tube station is located, faceted by room types. However, the results were not significant nor coherent across different neighborhoods in London. While this insight was interesting to us, we have decided not to include it as it turned out not to be a significant price determinant.

In line with our interest in cleaning fee as a component, we also tried to plot room fee and cleaning fee to check for any observable patterns.
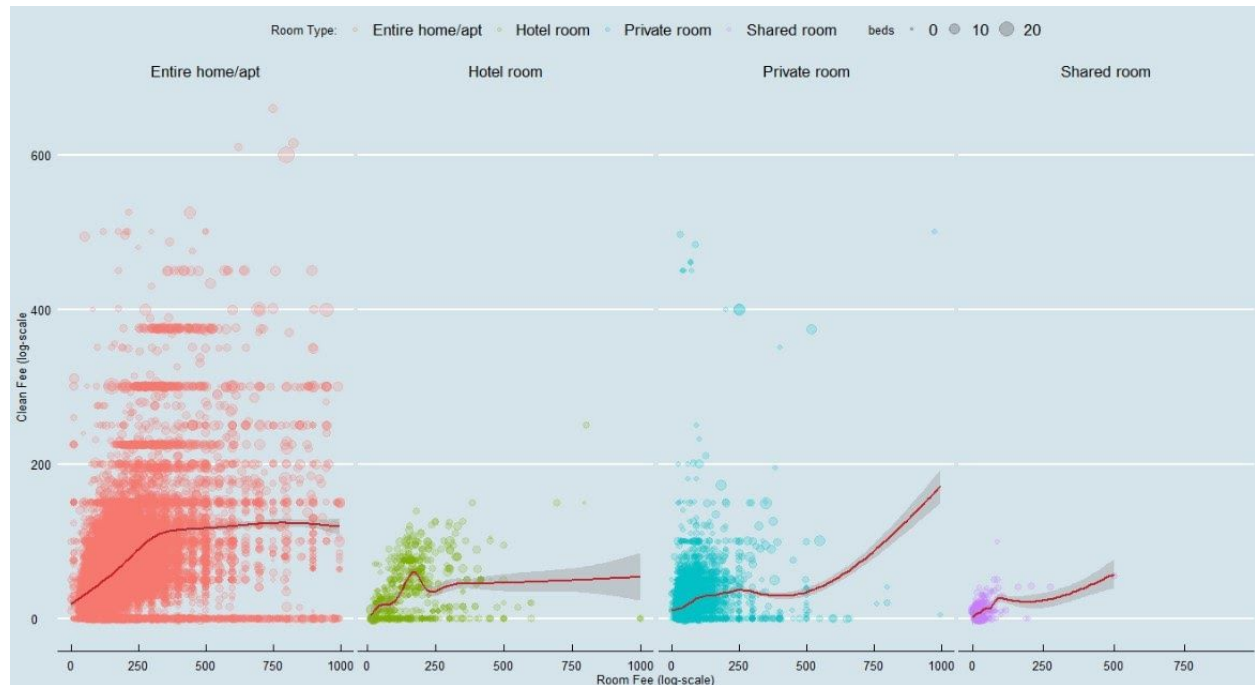
As you can see from the above, we were not able to find meaningful correlation among different room types. Thus, we have decided not to include it in the final visuzlatioan.

Another visualization related to cleaning fee we omitted is a density-barplot that shows the distribution of room fee and cleaning fee across all the listings, as seen below. After plotting the graphs below, we realized that showing the density of the different fees did not really make much contribution towards our hypotheses, and decided to leave them out.

Histogram of Cleaning-Fee per Night

Histogram of Room-Fee per Night
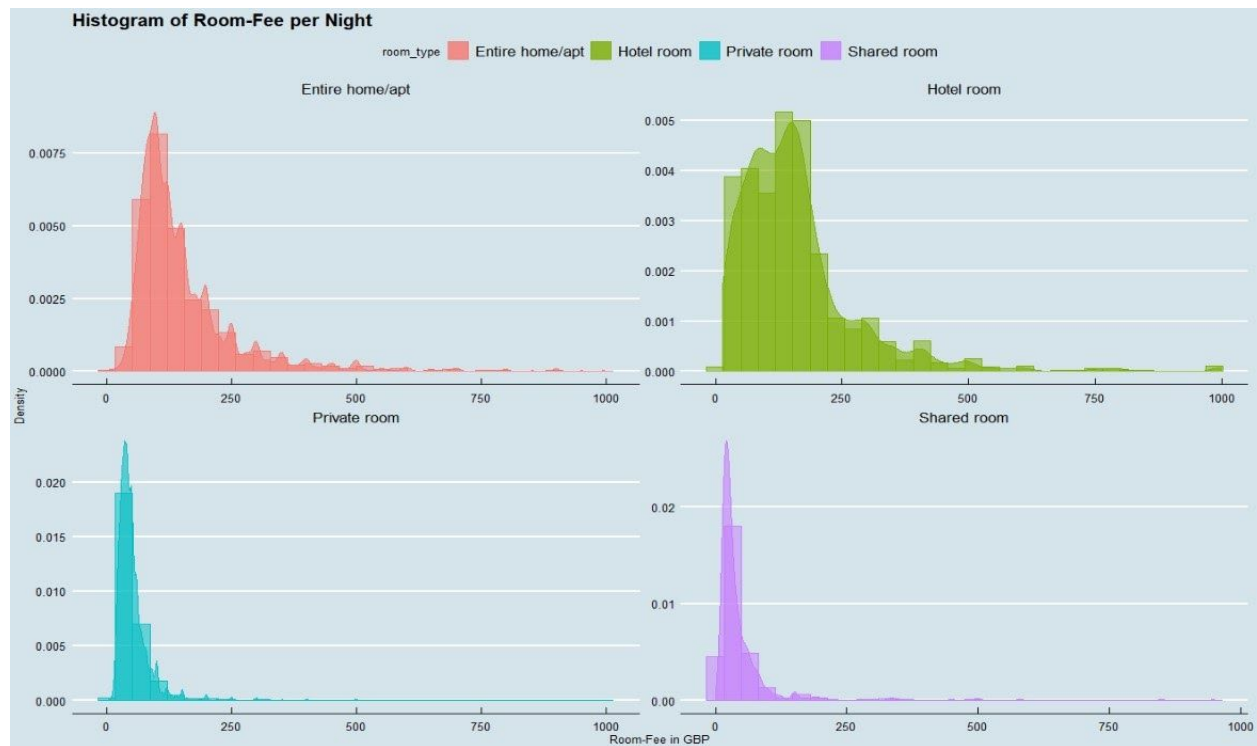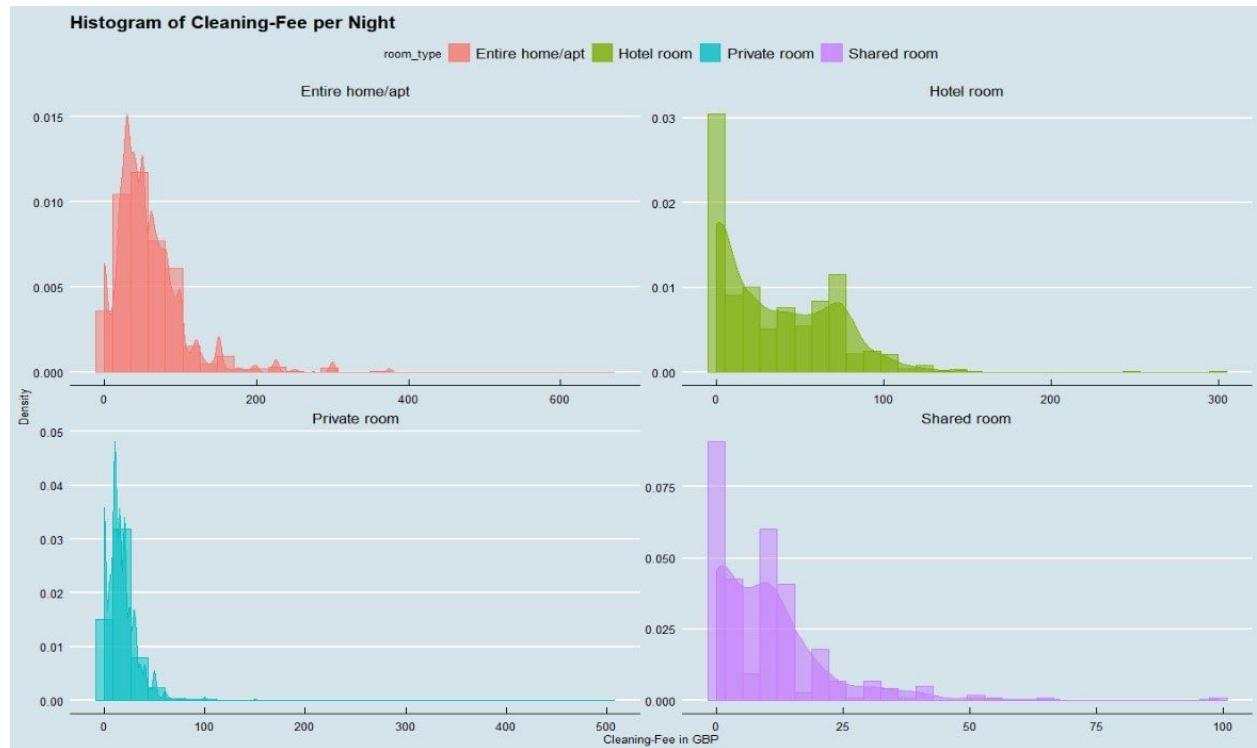
Final Remarks

The discarded visualizations mentioned above are just a few among many others that we tried. We then selected the most informative ones focused on aspects we believe are the most meaningful towards Airbnb hosts. By visualizing the most crucial features on our website, we hope to better equip Airbnb hosts to make data-driven decisions on pricing and other marketing strategies.