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: The dataset that we are using consists of data regarding Airbnb listings and reviews from 2009 to 2019. The dataset was obtained from InsideAirbnb.com (http://insideairbnb.com/london/) but a pre-processed version can be found on Kaggle. It comprises over 90 columns (e.g., customer ratings, room price, location etc.) 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.

Tube (London Metro System): For the data that we have used for metro information, we 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 that information that aids existing hosts is rare. Thus, for hosts with established listings, by giving sense of how the price is set for different aspects (e.g., location, room types, number of guests etc.), we hope that the hosts can use this website for a pulse-check on whether they remain well-placed 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 to gather information on how they should price their listing. New Airbnb hosts can use the website 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 details on each listing, and when they click on them, 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 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.

The next two graphs 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.

Section 4: Cleaning Fees

In this section, users can analyze how cleaning and room fee affects the total price that the customers would have to pay per night. Excluding deposits, cleaning fee is included in the final payment for customers. Thus, we wanted to provide some implications regarding the cleaning fee also. Here, users can observe how average fee (per one person) per night varies by number of beds. Based on the filtration (neighborhood, number of rooms, number of guests, and types of lodging) users can observe the visualized statistics of total room cost broken into cleaning and the room fee itself. For hosts, they could know how much is paid for the room itself and cleaning fee would be needed in average per person.

Section 5: Guest Ratings

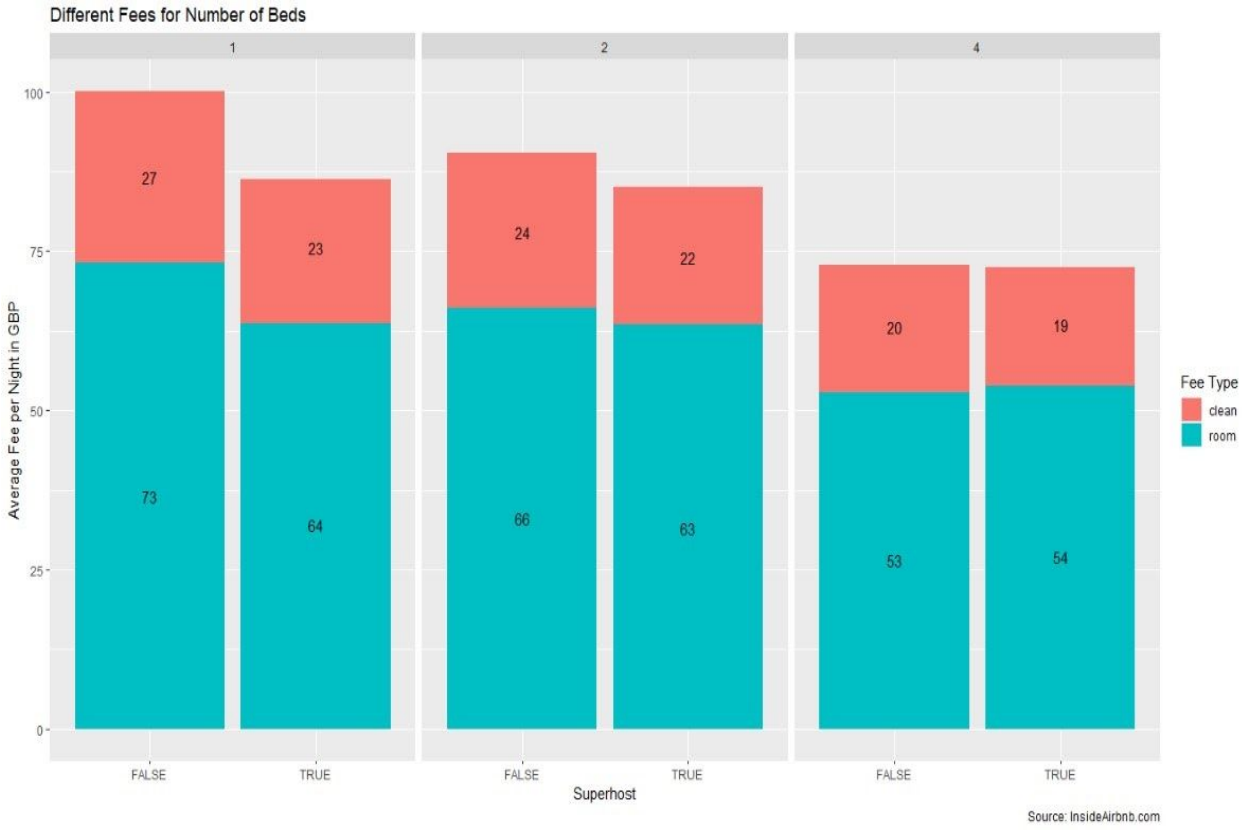
Here, users can observe the number of listings by customer ratings. Especially, for hosts, this could be an important indicator of how they are performing. Red dashed-line denotes the average review score rating.

Initial ideas and Refinements

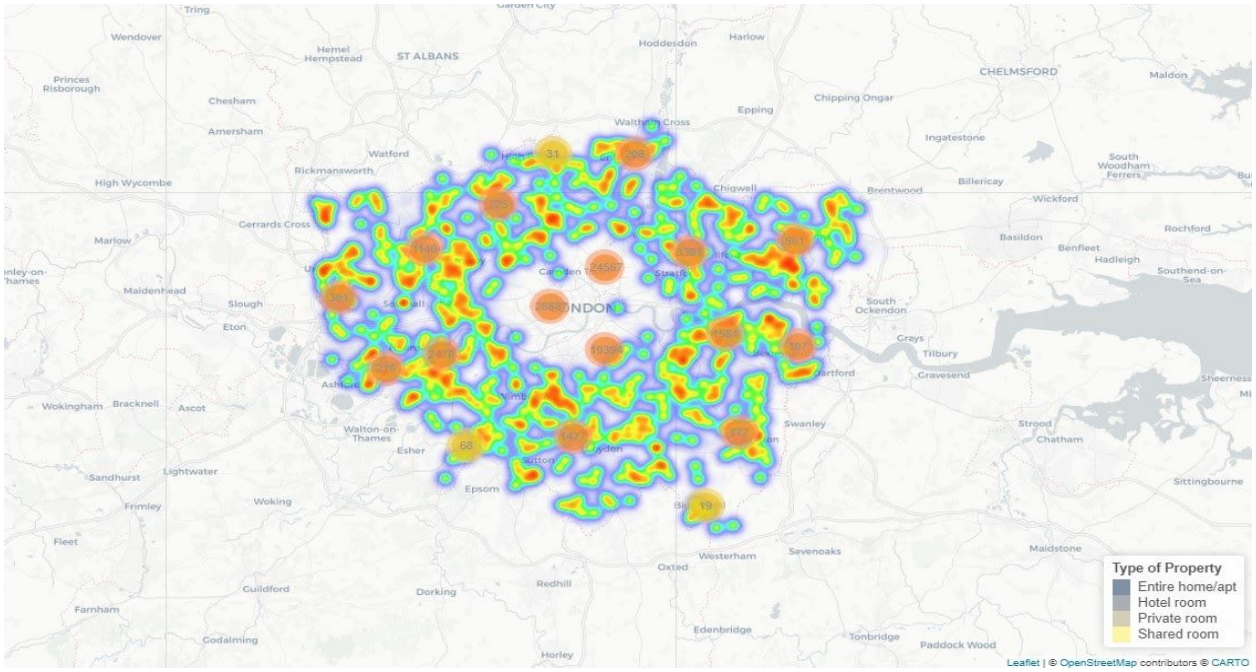
One of the questions we were interested to explore was how a new Airbnb host can gain credibility as a newcomer. As an extension, we thought we could explore the behaviour and best practices of Airbnb hosts across multiple segments beyond newcomers, including listings that are popular and established, 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.

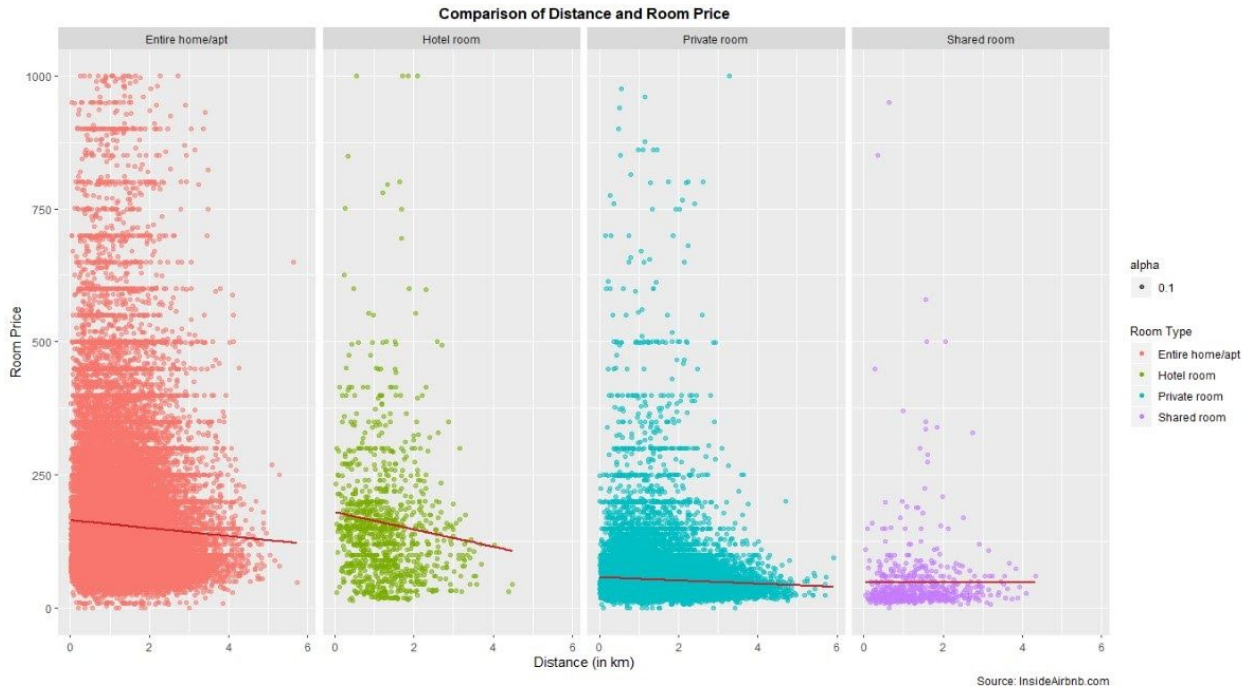
Furthermore, we wanted to observe whether there is a price premium for super hosts. For instance, below was one of the analyses we have decided not to include since the visualization itself was not too meaningful.



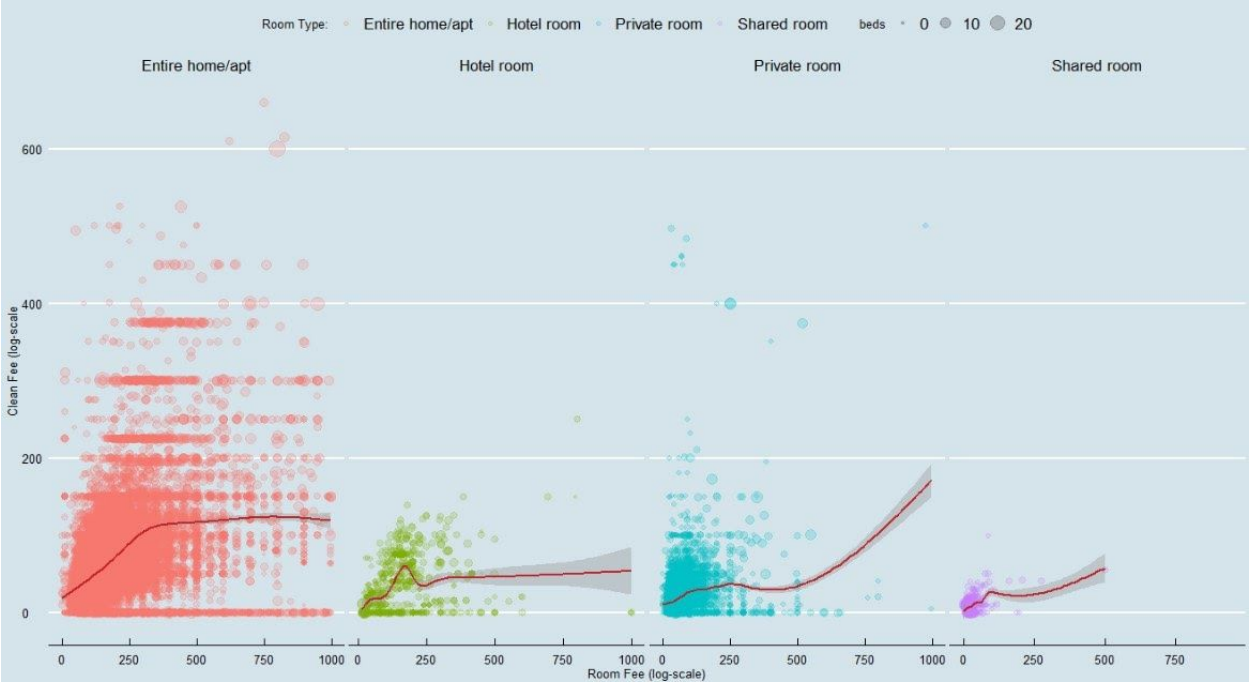
As you can see, rather than premiums, we see that the mean total fee for superhosts were rather discounted. However, the reason that we have not decided to include this in our final results was the price varies by different parameters we have incorporated in our final output and shown to have no distinct relationship. Second, we have demonstrated a heatmap of the London area in order to show the density of mean price differences.



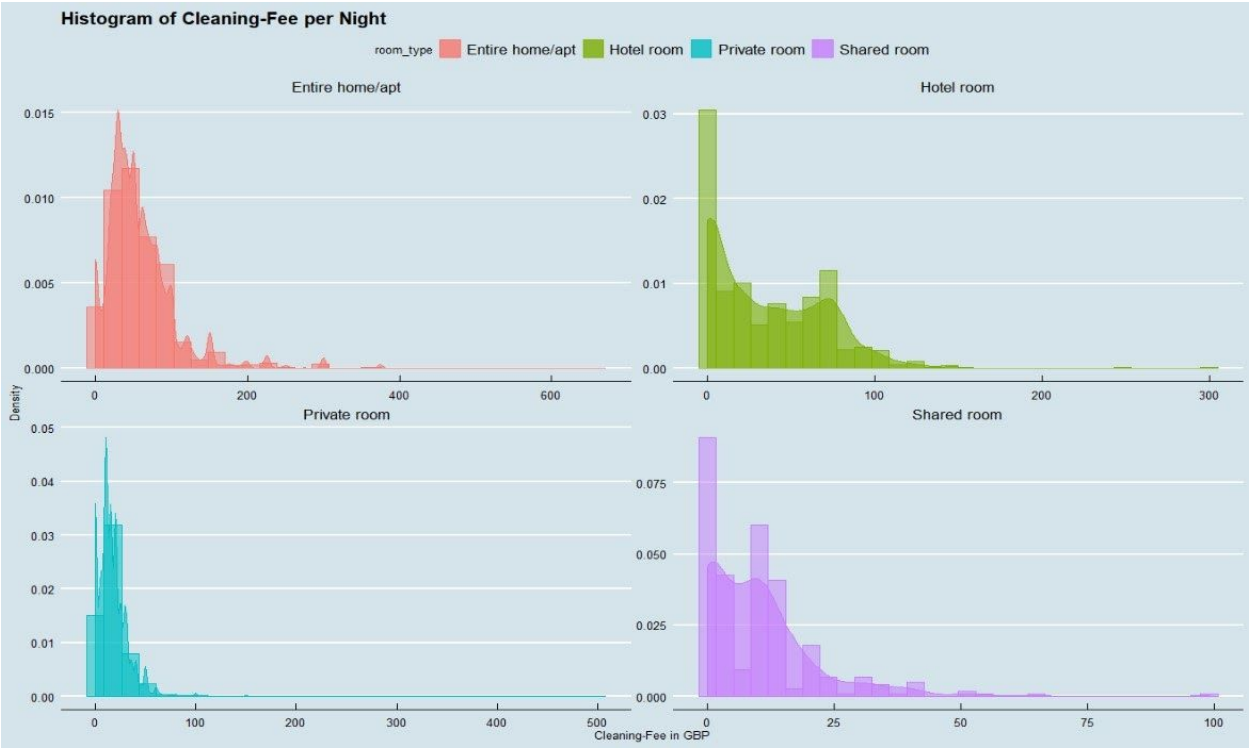
However, the plots were scattered across the London region, showing no trend of any kinds. Thus, we decided to remove this analysis from the final output. Third, we wanted to observe how the distance between the listings and Tube (London Metro System) affects the price of listings.



Above shows the average price difference with distance where the Tube is located by room types. However, by filtering out different parameters (e.g., number of rooms or neighborhood) the results were not coherent across different boroughs in London. Therefore, we have decided not to include this analysis. Fourth, we have observed whether there is a pattern between room fee and cleaning fee.



As you can see above, we were not able to find meaningful correlation among different room types. Furthermore, as discussed above, varying by different parameters also produced various results. Thus, we have decided not to include it in the final visuzlatioan. Lastly, we have made a density-barplot where it shows how the room-fee and cleaning-fee was distributed among total listings.



Nonetheless, this result also varied by the different parameters and we believed that just showing the density of how the different fees are outlined did not really make any

contribution towards our purpose of the project. Hence, we have removed these visualizations from the final submission also.

Final Remarks

With the visualizations in the website, we hope to equip Airbnb hosts with more information to make data-driven decisions on their pricing and other potential marketing strategies.