Understanding Airbnb listings in Australia

Louelle Teo Fengmin
Singapore Management University
louelle.teo.2020@mitb.smu.edu.sg

Jason Tey Shou Heng
Singapore Management University
jason.tey.2020@mitb.smu.edu.sg

Wong Kian Hoong
Singapore Management University
kh.wong.2020@mitb.smu.edu.sg

ABSTRACT

(abstract of not more than 300 words)

The abundance of Airbnb data provides great opportunity to conduct a variety of data analyses to understand the residential short-lease rental market. The dataset that has be scrapped on the Airbnb web and made publicly available by Inside Airbnb provides geospatial, textual, and quantitative data on each of the listings listed on the web. This project provides an analytics platform for interested parties (especially non-data specialists) to conduct exploratory spatial data, text, cluster, and regression analysis on the Australia Airbnb dataset using simple and user-friendly interactive dashboards that does not require programming knowledge.
(The actual research paper itself should not more than 6 pages excluding figures, tables, formula and references.)

1. INTRODUCTION

Airbnb is an online marketplace platform for accomodation rental. Founded in 2008 by Brian Chesky and Joe Gebbia who put an air mattress in their living room and offered bed & breakfast (thus "Airbnb"), the company has grown to be one of the most popular short-term accomodation rental platforms in multiple countries around the world.

With millions of listings in 220 countries and over 100,000 cities, Airbnb has a rich store of data from transactions between hosts and guests. Such data includes structured data like price, number of facilities (e.g. bedrooms, bathrooms), minimum and maximum number of nights' stay; and unstructured text data like description of the accommodation, and reviews by guests.

This assignment focuses on analysis of unstructured text data from Airbnb's online marketplace.

2. MOTIVATION

(Motivation of the application)

All of the studies discussed earlier focused on single dimensional aspect of the dataset - for example, Kwon (2018) on predicting ratings, Chen (2019) on visualizing the data, and Gedik (2020) on identifying common amenities potentially useful in predicting demand and price. The only attempt at merging the analyses into a centralised platform to be used for thematic analysis is Gupta's. Notwithstanding the fact that his Shiny apps are hosted on separate io pages, Gupta's analyses are also disjointed in content. Gupta started off providing descriptive analysis of Airbnb in New York, then transited to a Shiny for property locator, then more exploratory analysis on geospatial and categortical distribution, before transiting to temporal and text analysis - throughout his study, there is no one common theme that links the different methodologies and discussion together. A more concerted and systematic organization of the analysis and the tools used is needed.

Across all the studies examined, including Gupta's Shiny app, there are no provision of analytical tool that can allow users to change certain parameters to tweak the analyses to their own objective/preference. For example, in Kwon's predictive modeling, there is no way to change the methods or predictor that is included, or perhaps even to change the dependent variable of interest. In Chen's visualization, there is also no way to change the variable presented, say from her original price-as-size, to examine ratings using size or even colors on the viz. Similarly, Gedik does not offer any tools in this aspect. Gupta, while offering a wide range of analytical methods and even offering Shiny app, also surprisingly fell short in this area - for example, in his text analytics app, while user can change the word query, they are unable to, say, focus on reviews with more than a 4-star ratings to find out common words associated with higher ratings.

Another gap that is found across all study and one of the most major flaw considering the wealth of information provided by the datasets. None of the study offered options to, for example, zoom in on a particular sub-region, or to filter out whether a hosts is a superhost, or to subset data based on a range of prices. These are useful features that could tremendously enhance the usefulness of the analysis

#Review of past works *(Review and critique on past works)*
The data offered by Inside Airbnb are split into several different data sets - this assignment will be focusing on the listings file, which provides detailed data from individual listings put up on the AirBnb database. Information pro-

vided includes hosts details, response data, listing information (e.g. price, number of beds, property type, maximum nights), and review scores.

There are currently a number of analyses that has been conducted on the Airbnb dataset. While most of them are focused on one particular country, given that the datasets made available by Inside Airbnb are rather consistent across different geographical space, it is reasonable to assume that the analytical methodologies adopted by these other studies can be easily replicated and reproduced with data from other countries.

One of most prominent and recent study on Airbnb data is by Steve Deane from Stratos, who wrote a blogpost in January 2021 on the topic (which also inspired this project). In his post, Deane provides descriptive statistics (e.g. total number of hosts, most popular destination, number of guests) focusing on the global distribution with a slight inclination towards the United States of America. While some of the statistics were provided, Deane's analysis is heavily weighted towards the economic aspects of Airbnb (e.g. whether Airbnb affects property values and rents - "Airbnb effect," whether Airbnb is cheaper than hotels, and economic impact of Airbnb in each country). However, Deane does not provide any higher level data analysis beyond the descriptive ones. He did presented one segment explaining the factors that host should consider when purchasing a property to use as an Airbnb - reflecting some form of explanatory analysis (and perhaps even predictive) but the content was scarce with no elaboration on how these factors were derived. The major 'flaw,' though, remains the fact that Deane's blog post is heavy on qualitative write-up with minimal (or no) meaningful visualization of the statistics he has quoted.

Several blog posts on medium.com provided basic guides on data analysis on the Airbnb dataset. Kwon et. al. (2018), Chen (2019), and Gedik (2020) are some recent examples.

Kwon et. al. (2018) used the Inside Airbnb listing data from Austin, Texas with around 12,000 listings, and utilized the numerical features to apply Linear Discriminant Analysis, Outliers were detected and removed using the Cook's distance before a Box-Cox transformation was conducted to normalize the data, and the dependent variable (review score) binned to wrangle into categorical data type. Backward Elimination, Ordinal Logistic Regression, and LASSO Regression methods were employed to conduct explanatory analysis - with all three models turning number of bathrooms as significant predictor for customer ratings. Principal Component Analysis was then conducted with no good result after the model fail to meaningfully provide separation of classes (three classes). The Latent Discriminant Analysis was then conducted on each of the three classes resulted from the PCA. The study concluded that number of listings by hosts and having more bathrooms are crucial in securing higher review score. The advantage of this study is the depth, with detailed discussion on the methodology and results. However, the analysis focuses only on predictive/explanatory model, and is one-dimensional, studying only one dependent variable. The study does not offer interactive features to allow other forms of data exploration (e.g. other clustering methods, geographical influence) or changing of dependent variable (e.g. to examine factors affecting price of listing).

Chen (2019) on the other hand provides an analysis that emphasized on the geographical distribution of listings, and provided more data visualizations (playing to the advantage of using Tableau) that allow viewers to make quick noticeable trends. Chen analyzed the 2019 New York City listing data with the objective to predict future Airbnb performance in the city. However, while a predictive model was envisioned, there was little predictive analysis going on in her study - a greater emphasis was placed on qualitative inferences made via the visualization without proper data analytics methodologies. Chen was also limited in her methods of data visualization - for example, when presenting the geospatial distribution of areas by price, she had used a proportional dot map, without a convincing argument of its strength as compared to other visualization methods (such as a choropleth). In fact, the proportional size between the different dots are difficult to be differentiated based on her viz. There are also major flaws with other visualization choice (e.g. showing Average Price by Locations with equal-length bar differentiated by colors on a continuous scale), and the usefulness of some of the viz is also questionable (why might we be interested to know who are the Top 10 Busy Host, and indeed, what is the definition of 'business' here). All in all, while Chen presented some simple visualization to expose the potential of data viz using the Airbnb data, more is left desired from the analysis.

Comparatively, Gedik (2020) did a fair job in answering the questions set out by his objective. Using the Seattle listing dataset (via Kaggle), he aimed to find out the common amenities, and top features attracting guests and higher prices. In a Question-and-Answer format, Gedik presented visualizations to help answer each of the question he asked in a simple and succinct manner. In his third question, Gedik also presented the result table of a linear regression model he had ran, showing that property_type_Boat has the largest effect on price of listing. Gedik's post, however, suffers from the limitation of scarce discussion in the technical methodology aspect. Similar to earlier studies, there is also no interactivity offered to allow viewers autonomy in changing parameters.

Amongst all, Gupta (2019) provided the most well-rounded discussion and presentation using the Airbnb listing data. Gupta aimed to provide an exploratory analysis of Airbnb's data to understand the rental landscape in New York City. Gupta first employed descriptive time-series statistics to map out the increasing trends in number of listings and reviews in the city, before moving on to present an interactive Shiny App that provides information on individual listing based on sets of filters (e.g. max budget, number of people, minimum rating). While Gupta offered some form of interactivity, the Shiny does not provide any meaningful insights, and is essentially a replica of the user interface offered by Airbnb via the official website.

While non-interactive, Gupta's subsequent discussion provides a preliminary view of the analytical methodologies that could be employed using the Inside Airbnb listing data. Firstly, he mapped out a geographical spread of the ratings

and price by area using a Choropleth map (subsequent discussion in this paper will also conclude that it is indeed one of the better geospatial visualization option for this application). He also provided a quick bar-chart viz of the breakdown in terms of property types, before moving on to discuss the temporal nature and seasonality effect of price in New York City (using both a time series dot plot to analyse monthly patterns and a boxplot for each day of the week. He even employed a calendar heat map of occupancy within the city. After geospatial and temporal analysis, Gupta moved on to conduct text analytics on the reviews dataset to find words most commonly mentioned - an attempt to distil aspects that play important roles in shaping the Airbnb experience. A Shiny app was also developed to allow user to find similar word vectors based on a query word. Lastly, Gupta set out to find out whether there are any correlation between the host response rate, the average ratings, and whether the host is a Superhost. While no predictive model was used, the scatter plot presented provides a quick viz on the explanatory analysis he desired to conduct.

Within the industry, there exists interactive tools that allow analysts or potential hosts to analyze rental data using attributes and past performance. One such example is AIRDNA. Notwithstanding the fact that the platform only offers paid services, the results are also provided in a prescriptive manner with little analytical value-add.

3. DESIGN FRAMEWORK

(A detail description of the design principles used and data visualisation elements built (Refer to Section IV: Interface of this paper.)

4. DEMONSTRATION

(Use case)

#Discussion *(What has the audience learned from your work? What new insights or practices has your system enabled? A full blown user study is not expected, but informal observations of use that help evaluate your system are encouraged.)*

#Future Work *(A description of how your system could be extended or refined.)*

5. CONCLUSION

Duis nec purus sed neque porttitor tincidunt vitae quis augue. Donec porttitor aliquam ante, nec convallis nisl ornare eu. Morbi ut purus et justo commodo dignissim et nec nisl. Donec imperdiet tellus dolor, vel dignissim risus venenatis eu. Aliquam tempor imperdiet massa, nec fermentum tellus sollicitudin vulputate. Integer posuere porttitor pharetra. Praesent vehicula elementum diam a suscipit. Morbi viverra velit eget placerat pellentesque. Nunc congue augue non nisi ultrices tempor.

References

[1] Fenner, M. 2012. One-click science marketing. *Nature Materials*. 11, 4 (Mar. 2012), 261–263.

[2] Meier, R. 2012. *Professinal Android 4 Application Development*. John Wiley & Sons, Inc.