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

The abundance of Airbnb data provides great opportunity to conduct a variety of data analyses to understand the residential short-lease rental market. Most publicly available analysis tools of Airbnb data do not have much interactive features and are therefore, by design, narrow in their application and scope. This Shiny application provides an analytics platform for interested parties such as analysts, especially those who are not familiar with coding and programming languages, to conduct exploratory spatial data analysis, text and sentiment analysis, cluster analysis, and predictive analysis on the Australia Airbnb dataset. No prior programming knowledge is required to use this interactive dashboard, while the environment allows users to customise their scope of analysis. This paper details the design framework, data preparation and use case demonstration using the simple and user-friendly interactive dashboards.

1. INTRODUCTION

Airbnb is an online marketplace platform for accommodation rental. Founded in 2008 by Brian Chesky and Joe Gebbia who had the idea to put an air mattress in their living room and offered bed & breakfast[5] (thus "Airbnb"), the company has grown to be one of the most popular short-term accommodation rental platforms in multiple countries around the world.

With millions of listings in 220 countries and over 100,000 cities[1], 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 such as description of the accommodation, and reviews by guests. This paper utilizes such data scraped by Inside Airbnb[10] via the Airbnb website to create an interactive dashboard that allows analyst or anyone interested in the AirBnb landscape to conduct the statistical analyses without the need of any prior coding or programming knowledge. Analytic tools included in this dashboard are Exploratory Data Analysis, Cluster Analysis, Exploratory Spatial Analysis, Text and Sentimental Analysis, and Predictive Analysis.

2. MOTIVATION

Airbnb is one of the pioneers and largest players in the industry. It is a rapidly-growing company in the increasingly important sharing economy of the world, given rich multitude of data made available through its website, there are huge potential for a free-to-use analytic dashboard that housing market analyst, sharing economy professionals, property agents, research institutes, and would-be Airbnb hosts to utilize and better understand this business environment.

However, analysts, hosts, and professionals who are keen to delve into such data sets might not have the requisite knowledge or programmatic skills to develop their own analytic tools. This application aims to bridge this gap by providing a publicly available analytic dashboard that provides various analytic techniques to enable interested parties to tap on the rich data provided to derive meaningful insights for their professional use. The target users of this application are analysts with some knowledge of statistical techniques but lack the coding skills to develop their own tools.

While this application focuses on the Australia Airbnb listings data set, the data processing, methodologies, and application development process can be easily replicated and reproduced with datasets of other geographical areas that bear the same data structure.

3. REVIEW OF PAST WORKS

There are currently a number of analyses that has been conducted on the Airbnb dataset.

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. In his post, Deane provides descriptive statistics. While some of the statistics were provided, Deane's analysis is heavily weighted towards the economic aspects of Airbnb. Deane does not provide any higher level data analysis beyond the descriptive ones, with limited statistical content and lacking in elaboration on how factors used were derived. The major 'flaw,' though, remains the fact that Deane's blog post is heavy on qualitative write-up with minimal visualization - much less interaactivity - of the statistics he quoted.

Several blog posts on provided basic guides on data analysis on the Airbnb dataset. Kwon et. al.[11], Chen[6], and Gedik[8] are some recent examples.

Kwon et. al.[11] used the Inside Airbnb listing data 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. The study concluded that number of listings by hosts and having more bathrooms are crucial in securing higher review score. While the study provided a well-rounded discussion on the statistical methodoly, it does not offer interactive features to allow other forms of data exploration or parameter adjustments.

Chen[6] on the other hand provides an analysis that emphasized on the geographical distribution of listings, and provided more data visualizations that allow viewers to observe quick noticeable trends. However, Chen was limited with her visualization choice (e.g. showing Average Price by Locations with equal-length bar differentiated by colors on a continuous scale), and similarly does not provide interactivity to conduct analysis that cater to needs of individual analysts.

Comparatively, Gedik[8] did a fairer job in terms of data analysis. However, similar to earlier studies, there is also no interactivity offered to allow viewers autonomy in changing parameters.

Amongst all, Gupta[9] 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. He 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.

Within the industry, there exist interactive tools that allow analysts and 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, and limited customisation for users to tweak the analysis to their individual requirements.

4. DESIGN FRAMEWORK

The application makes use of the R stastical language which is open-source and offers many tried-and-tested packages for the type of analysis that the application will feature. The design considerations are as follows:

1. Reproducibility of results by performing calculations within the application itself.

2. Adoption of common R packages in the Comprehensive R Archive Network[13] (CRAN) for supportability.

3. Use of the R Shiny package for interactivity and easy deployment.

4. Interactive features for easy use.

4.1 Data Preparation

All data preparation was performing using R in the RStudio IDE. This included dropping columns from the data set that were irrelevant for the scope of our research, converting columns into the appropriate data types, and imputing NA values where appropriate.

4.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach to data sets to summarise their main characteristics. With EDA, we can:

1. Propose hypotheses to explain the observed characteristics of the data

2. Assess the data sets and observe any anomalies and outliers

3. Support the selection of appropriate statistical tools and techniques

4. Provide a basis for further data collection through surveys and experiments

4.2.1 Barcharts

Barcharts in the application will showcase the number of Airbnb listings or hosts per State and Local Government Area (LGA). This allows users to understand the market conditions of the area they are interested in.

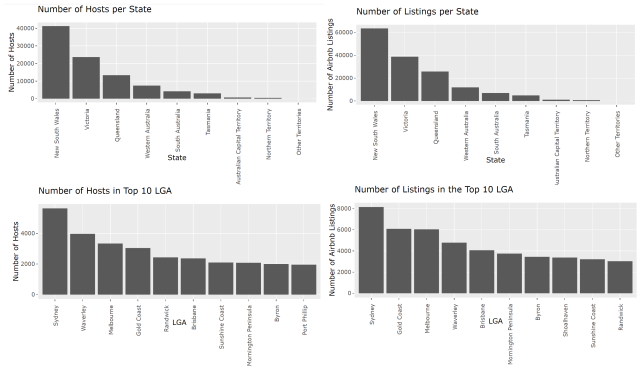

Figure 1: Barcharts of Airbnb listings and hosts per state and LGA

The above visualization is obtained by counting the unique Host ID and Listing ID against a State or LGA. It is noted that **New South Wales** have the highest number of Hosts and Listings, with the LGA, **Sydney** taking the top spot.

4.3 Boxplots

Boxplots in the application allow us to explore statistical data of the different variables in each LGA, or per *Property Type*. It is a convenient way of depicting groups of numerical data through their 5-number summaries.
- Smallest Observation
- Lower Quartile
- Median
- Upper Quartile
- Largest Deviation

The application also allows users to toggle between different continuous data to analyze the boxplots in different LGAs.

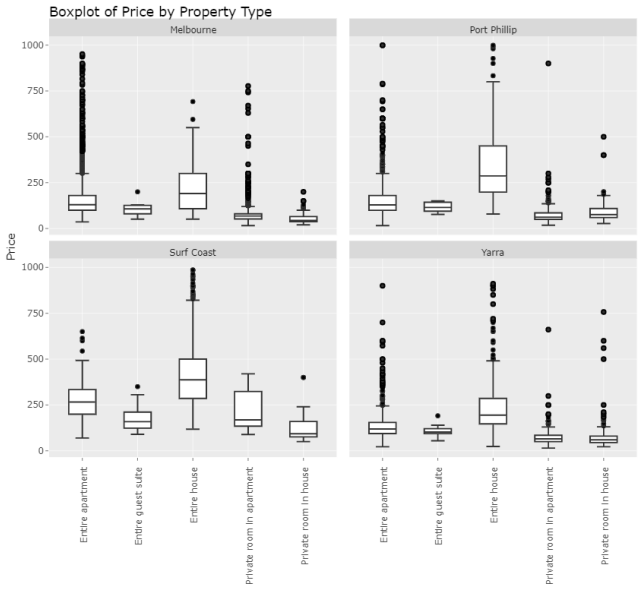

Figure 2: Boxplots of price by property type

Taking a look at another State **Victoria**, we can analyze the distribution of the pricing or other variables through a boxplot. It gives us an insight into the pricing range for the type of housing.

4.4 Density & Bivariate Analysis

The Density plot of variables allows us to view and observe any skewness in the distribution.

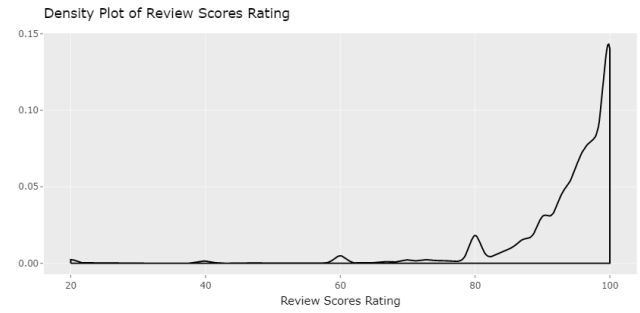

Figure 3: Density plot of review scores rating

For example, in this density plot of the State of **Victoria**, the *Review Scores Ratings* are generally between 80 to 100, and peaks can be seen at multiples of 20.

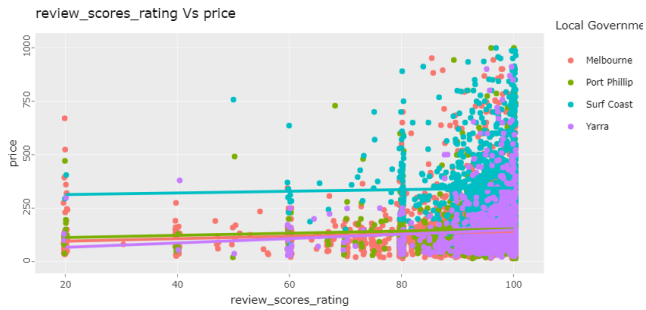

Figure 4: Scatterplot of review scores rating vs price

In the Bivariate analysis, we can review how the X Variable of interest (eg, *Review Score Rating*) correlates to another variable. In the plot above, we found the four LGAs to have a low correlation between *Price* and *Review Scores Rating*.

4.4.1 Correlation Matrix

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables.

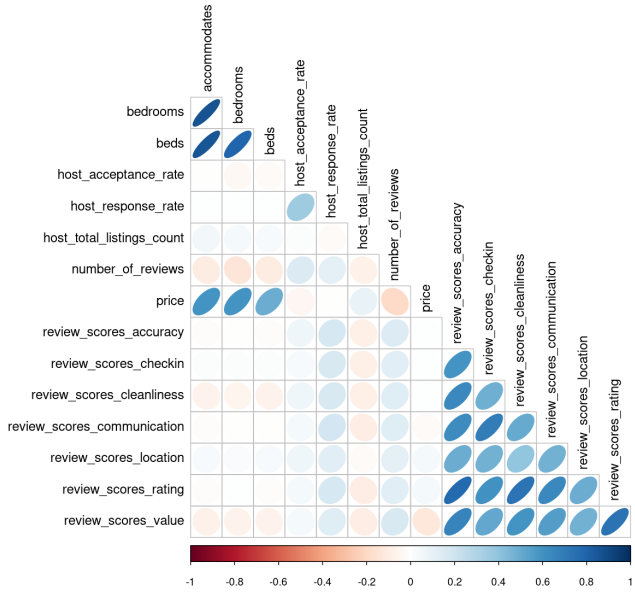

Figure 5: Correlation matrix of different variables

Highly correlation variables should be treated carefully. For example, with linear regression, a high amount of correlations suggests that the linear regression estimates will be unreliable.

4.5 Cluster Analysis

K Means clustering is a partitioning clustering approach. Each cluster is associated with a centroid, where the data is assigned to the cluster based on distance. The number of clusters, K must be specified, and the basic algorithm will repeatedly reassign cases to clusters. Clustering is an unsupervised learning technique that help users to identify data driven patterns that may warrant further investigations.

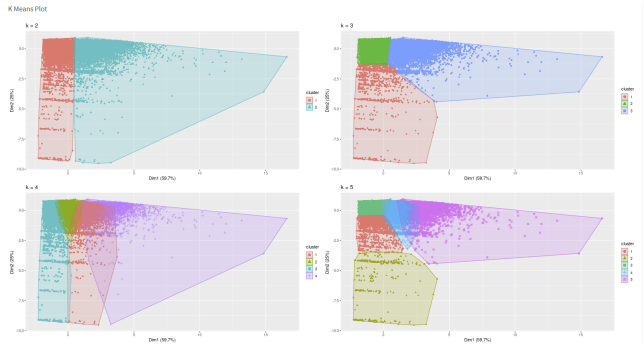

Figure 6: K Means clustering

The above visualisation is achieved by using package *biganalytics*. It clusters the chosen variables using euclid distance in the region of Victoria.

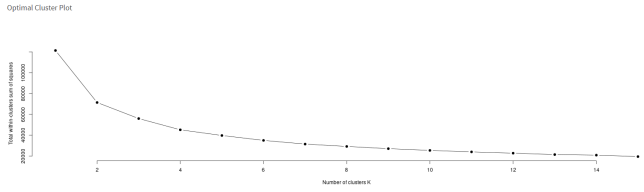

Figure 7: Optimal clustering plot

An optimal cluster size is plotted using the Elbow Method, to find out the number of clusters that is optimized for our dataset. The cluster size of four would be chosen, and a parallel plot is created to further visualise the characteristics of the dataset.

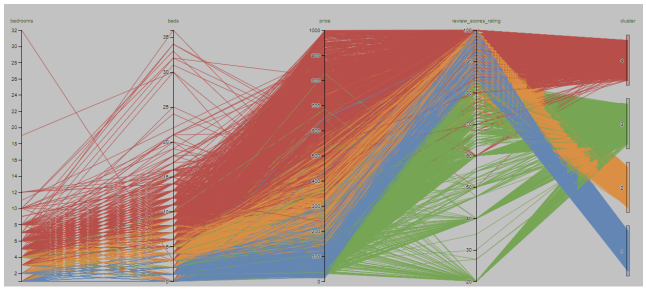

Figure 8: Parallel plot of four clusters

From the parallel plot above, cluster 4 appears to be listings with higher number of beds and bedrooms that are listed at a higher price with varying review scores rating. This cluster of listings likely caters to large groups.

4.6 Spatial Data Analysis

The spatial data analysis provides an analytical overview of the geographical distribution of different variables of interest across the different Local Government Areas (LGAs) of Australia. For example, the choropleth below presents the median price of listing is in each of the LGAs in the Victoria State - the darker the color, the higher the median price for that LGA, and we see the darkest range of median price in the area around the tip of the Melbourne bay. This way, we can easily compare whether median prices are within the same range, higher, or lower across the different LGAs.

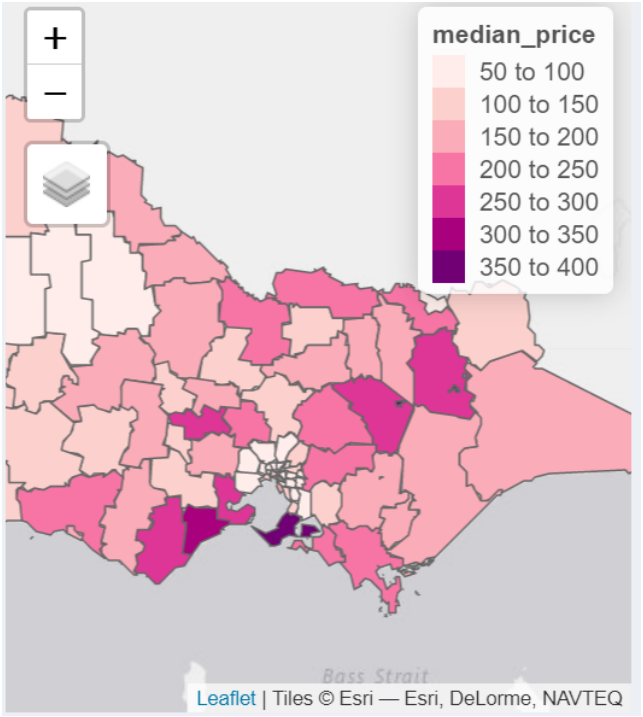

Figure 9: Choropleth of median price of Airbnb listings

The above visualization is achieved by merging the Airbnb listings data scrapped by Inside Airbnb (including the coordinates of each listing's geospatial position) with the LGA digital boundaries in ESRI Shaprefile format made available by the Australian Bureau of Statistics[4]. Firstly, six variables from the raw listing data (i.e.. price, count, ratings, number of reviews, acceptance and response rates) are summarized and grouped using the base R *summarise()* and *group_by()* modules into their LGAs to obtain aggregated variables.

Next, we import the shapefile using the *st_read* function from the *sf* package. We then left-join the summarized aggregated values onto the spatial dataframe.

Finally, we remove rows that have missing values from our data, filter data down to the state we are interested in (Victoria in this case) and then plot the viz using functions under the *tmap* package.
Beyond the selection of state, the application also allows user to tweak the parameters, such as the variable of interest (6 in total) and binning methods.

However, we also observed that there are LGAs of lower median price north of the bay Melbourne Bay area) and clusters of higher median prices (south-west of the bay). While good for overall distribution, the choropleth does not provide an objective view of how the variable (e.g. median price) of one LGA compares with the other LGAs around ir or across Australia. We hence aim to identify clusters based on variable of interest to identify "hot/cold spots" and outlier LGAs. To do this, we employ a Local Indicators for Spatial Autocorrelation (LISA) analysis, which helps to reveal clusters of LGAs/outliers based on their attributes.

The first step to determining the local spatial autocorrelation is to determine what constitutes a neigboring area to consider and hence the spatial weight to be assigned for analyzing association. There are several ways to do this, categorized mainly into two types: (i) contiguity-based[2] and (ii) distance-based[3]. (i) is further split into Queen's contiguity (with shared vertex - no shared border needed) and Rook's contiguity (smaller distance threshold; only shared borders), analogous to a chess board; (ii) is also further split into K's Nearest Neighbor (i.e. the K LGAs nearest to subject, centroid-to-centroid) and Distance-band (any area with centroid within a distance band radius from subject centroid). The neighbor spatial weights are derive using the respective modules within the *spdep* package.

The resultant spatial weight and spatial dataframe is then used to derive the LISA - in this case, the Local Moran's I (note: Getis-Ord's G Z-score is also available; discussion and application is available via this and this link respectively; the latter was used as a reference to develop this sub-module). We then use the *localmoran* function from the same *spdep* package to calculate the Local Moran's I statistics for each of the LGAs. The underlying formula is given as follows:

$$I_i = \frac{(x_i - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2/(n-1)} \sum_{j=1}^{n} w_{ij}(x_j - \bar{x})$$

Figure 10: Formula for Local Moran

where x(i) is the variable of interest for LGA i, w(ij) the spatial weight with neighbor j, and x-bar the mean. The statistic is then tested for statistical significance - LGAs with statistically significant statistic rejects null hypothesis that they are no spatial associated with their neighbors - these LGAs are then further categorized into four classes based on their deviation from the mean. If an LGA's variable of interest is higher than its mean, it has a comparatively high value; if its local Moran's is higher than the statistic's mean, the neighbors also have high values - this thus presents a High-High cluster. A summary of the four clusters is presented below:

For LGAs with statistically significant local Moran's I	Subject LGA has comparatively **larger** LISA statistics *(local Moran larger than mean)*	Subject LGA has comparatively **smaller** LISA statistics *(local Moran smaller than mean)*
Subject LGA has comparatively **high** value for variable of interest *(Variable larger than variable mean)*	High-High cluster LGA is part of a high value cluster, having high value itself and neighbours that also have high value	High-Low outlier LGA is high-value outlier with neighbours that have low value
Subject LGA has comparatively **low** value for variable of interest *(Variable smaller than variable mean)*	Low-High outlier LGA is low-value outlier with neighbours that have high value	Low-Low cluster LGA is part of a low value cluster, having low value itself and neighbours that also have low value

Table 1: Interpreting Local Moran

The resultant LISA analysis for median price in LGAs in the State of Victoria is as shown below. The areas shaded (regardless of which cluster/outlier the LGA belongs to) on the right-hand viz according to the confidence level selected (90%, the default, in the viz below) should correspond to the p-value of the left-hand viz (all the colored area has p-value corresponding to 90% confidence level and above). The methods of selecting neighbor and the confidence level for the statistical analysis can be varied according to user preference.

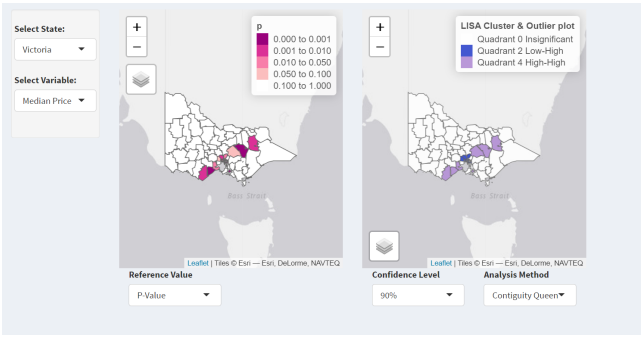

Figure 11: LISA analysis for median price in LGAs

4.7 Sentiment Analysis

4.7.1 Word Cloud

Word clouds are an effective way to represent the frequency that a word appears in a particular corpus (a collection of texts). For this application, the size of the word represents the relative frequency of the word in the description section of Airbnb listings. The word in the largest font is the most common word.

The first step to clustering is to create the corpus. This is done by scraping all the text from the desired column in the Airbnb data set - in this case, the description field. Once the corpus is created, tokenisation is done to clean up the words or terms and create a tibble of single terms in a column. Tokenisation includes removing extra white spaces from between terms, removing numbers (number do not have any meaning in a textual word cloud), punctuation symbols, and removing *stop words* (which refer to words that have no meaning like "I," "a," "this," "and," etc), and stemming words (reducing words to their root like "walking" to "walk"). This cleaning process ensures that only valuable terms are included.

After tokenisation is performed, the corpus of words are transformed into a Document Term Matrix (DTM), which is a matrix of the frequency of each term in the corpus. Based on the DTM, the word cloud will generate the words according to their frequency.

Figure 12: Word Cloud of words for a region

4.7.2 Topic modelling using Latent Dirichlet Allocation

Topic modelling is a statisical model to identify topics in textual data. It is an unsupervised machine learning technique that detects word and phrase patterns in documents and clusters them into groups known as topics. The concept behind topic modelling is that different topics would have certain words appear together frequently. Based on this, topics can be discovered based on the words that appear frequently together.

The two common approaches to topic modelling are Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA)[7].

As described by Cvitanic, et al. [7], LSA is a text analysis method that makes use of a semantic space (representations of natural language meant to identify meaning in language) to calculate or compute the similarity between words, phrases, sentences, paragraphs, or even whole documents. Similar to the word cloud, LSA makes use of a document term matrix to weight the words in terms of the frequency that they appear in the corpus.

LDA, according to Cvitanic, et al. [7] was a topic modelling technique that evolved from LSA. The basic idea of LDA is that "documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words."[7] To be able to identify the topics in a corpus, a "generative process whereby the documents are created" is done so as to infer the topics[12]. This inference is done by imagining documents being random mixtures of terms with latent topics underlying them, where each topic has a unique distribution across all the words in the corpus[12]. LDA requires a defined number of topics as an input parameter.

In other words, LDA assumes that all words in the document can be assigned a probability of belonging to a topic. As such, the goal of LDA is to determine the mixture of topics that a document contains.

LDA was chosen at the topic modelling technique to implement in this application.

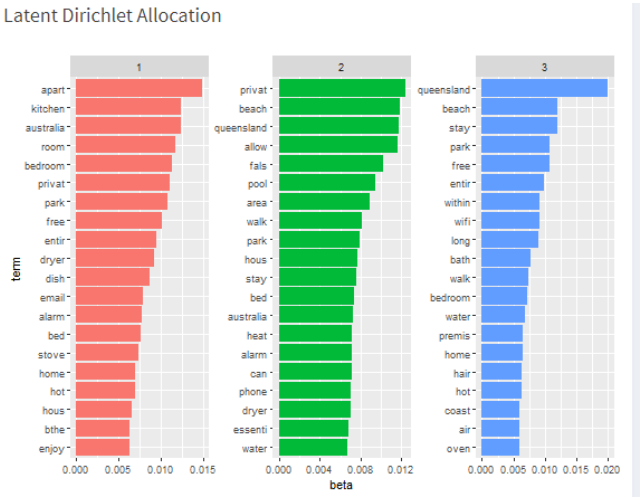

Figure 13: Topic Modelling using LDA for a region

4.8 Multi-linear Regression Analysis

Multi Linear Regression (MLR) is a statistical technique that uses multiple variables to predict the outcome of a target variable.

Regression Summary:

```
Call:
lm(formula = price ~ number_of_reviews + review_scores_rating +
    host_total_listings_count + host_acceptance_rate, data = Rdata)

Residuals:
     Min       1Q   Median       3Q      Max
-1161.36  -109.35   -43.35    58.62   870.00

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               81.656237   6.632017  12.312  < 2e-16 ***
number_of_reviews         -0.512785   0.008677 -59.097  < 2e-16 ***
review_scores_rating       1.809511   0.067091  26.971  < 2e-16 ***
host_total_listings_count  0.139778   0.005378  25.989  < 2e-16 ***
host_acceptance_rate     -11.570028   2.357183  -4.908  9.2e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 166.1 on 104802 degrees of freedom
Multiple R-squared:  0.0443,    Adjusted R-squared:  0.04426
F-statistic:  1214 on 4 and 104802 DF,  p-value: < 2.2e-16
```

Figure 14: Multi-Linear Regression Analysis

Users will be able to explore a Multi Linear Regression of either *Price* or *Review Score Ratings* to understand the variables that will directly affect the target variables. With this information, they can therefore understand the effect an individual predictor can cause.

Users are able to select multiple variables, and toggle through different analysis method such as "Forward," "Backward" and "Stepwise" regression.

The regression above returns four variables that are of statistical significance in predicting *Price*. However, the model is only able to explain 4% of the variation in price. (Adjusted R-squared = 0.044)

5. DEMONSTRATION

5.1 Spatial Data Analysis

This demonstration focuses on the LISA analysis for median price of listings in each LGAs within the regions of Victoria, Australia. Based on the LISA plot below (result of 90%

confidence level, Queen's Continuity for spatial weight), we conclude with 90% confidence that the LGAs shaded has spatial association with the LGAs around them.

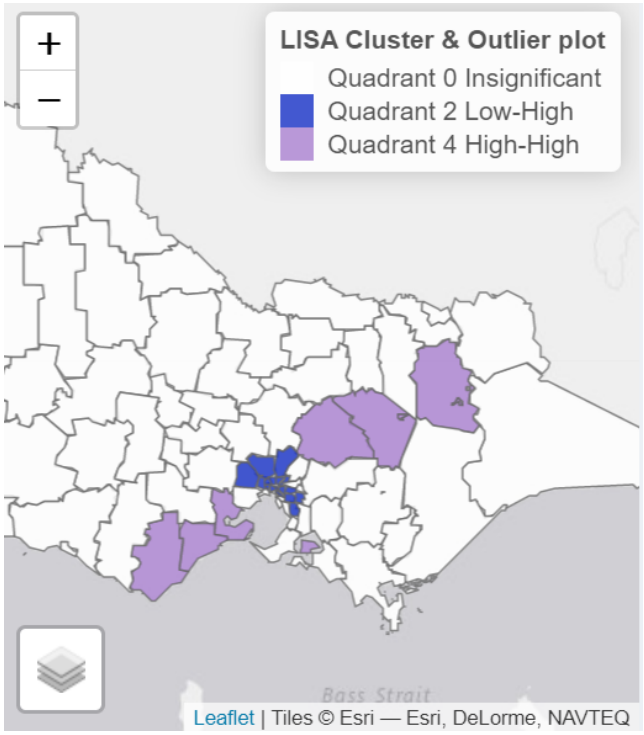

Figure 15: LISA analysis

We observed that there is an area of Low-High outliers (dark blue) in the north of the Melbourne Bay Area - these are outlier LGAs with lower median price as compared to their surrounding neighbors, which have higher median price for the Airbnb listings within. This is potentially driven by the largely tourist hotspot in the Melbourne area that drives Airbnb listing prices higher around the Low-High region we observed.

On the other hand, we also notice two distinct High-High clusters in purple. These area reflects clusters or congregation of LGAs with higher median Airbnb listing prices. The cluster on the southwestern end of the bay area corresponds to the start point of the famous Great Ocean Road and is hence not a surprise to have wide-spread area of high accommodation prices. The cluster to the East, though, comprises of mainly smaller and more remote tourist spots such as the Swifts Creek - the fact that they are detected as a High-High cluster is a little unexpected. However, once we shift to take a look at the raw median price (see viz in Section 5.3), we notice that this is mainly due to the fact that surrounding LGAs typically have lower median price (since it is a remote area) - any minor hotspot that increases the median listing price slightly (though statistically significant) in LGAs of close proximity (each with one small attraction) would result in a "High-High" cluster. This is confirmed when we notice the raw median price of this cluster is of lower price range (ligher pink) than those in the Great Ocean Road area.

5.2 Sentiment Analysis

Sentiment analysis with this app consists of a word cloud and topic modelling using Latent Dirichlet Allocation (LDA). For this demonstration, the region of Victoria, Australia, is selected. The review scores range of 91 to 100 is also selected.

5.2.1 *Word Cloud*

With the above filter selections, the word cloud created is shown below.

Figure 16: Word Cloud for Victoria

It is not unexpected that the word "victoria" is the most common word in this word cloud, since the region in question here is Victoria. Some other common words are "kitchen," "park," "privat" (private), "bed,"bedroom", and"walk".

This would suggest that these words are common in the descriptions that hosts write about their listings, and these listings garner review score ratings of 91 and above, which is the highest decile of review scores. Thus for an Airbnb host who aims to obtain review score ratings within that range, they should include these words into their descriptions. Of course, the accommodation itself should actually offer these facilities so as not to dissapoint the guest, which would expectedly lead to lower rating scores.

5.2.2 *Topic Modelling*

Similar to the word cloud,the region of Victoria, Australia, is selected with the review scores range of 91 to 100. The number of topics to be identified is three, with the top ten words of each topic presented in the visualisation.

With the above filter selections, the topic model is shown below.

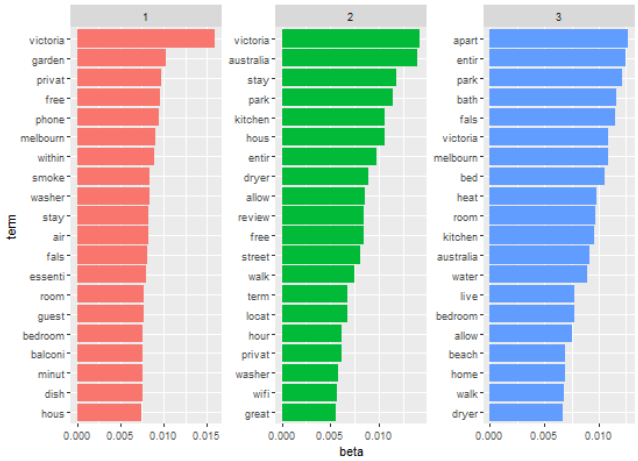

Figure 17: Topic Modelling using LDA for Victoria

The bar beside the corresponding word represents the probability (beta) of that word appearing in that particular topic. The longer the bar, the higher the probability.

Based on the top words in each topic we can identify the following topics:

- Topic 1: The word "garden" is very common in this topic (other than the word "victoria" which is the region in question). Along with "garden," other words like "phone," "washer," "essenti" (essentials), "room," "bedroom," "balconi" (balconies), "minut" (minutes) feature in the top 20 most common words in this cluster. This suggests that listing descriptions with these words tend to appeal to guests who stay in the accommodation more often, and enjoy facilities such as a garden, washer, balconies, etc. Such guests may be older persons who enjoy staying in the comforts of the house.

- Topic 2: Words common to topic 2 are "park," "kitchen," "dryer," "street," "walk" "washer," "wifi." These are some of the words in the list of top 20 most common words in this topic. This type of accommodation description may appeal more to younger guests who want Wi-Fi access in the accommodation, and also who might enjoy going outdoors to the park for a walk.

- Topic 3: The common words in this cluster are "park," "bath," "bed,"heat","bedroom","beach","walk","dryer". These terms may appeal to those who are looking to visit the beach during their stay.

6. DISCUSSION

The geospatial analysis allows audience to have a quick overview on the geographical distribution of a variable of interest (say, median price) across a selected level of detail (whole-of-Australia or individual LGA). The LISA analysis further adds on to this overview to allow identification of hotspots/coldspots (clusters of high/low value) and outliers (LGAs with abnormally high/low median price vis-a-vis its neighbors). User can form hypothesis about the spatial correlation between different variable of interest (e.g. price hotspots in Victoria corresponds to rating hotspots), where further confirmatory analysis can be conducted as an extension based on observations from this system. In fact, with the clusters and outliers identified, users can conduct deeper analysis on these areas using other tools available on this system (e.g. sentiment analysis, predictive analysis) to understand why these patterns formed.

From the sentiment analysis, it was observed that most guests only gave a review score if they has a positive experience staying at the accommodation. As such, most of the review scores were in the range of 91 - 100. This caused a lack of data that could have highlighted what words in a listing description might cause garner low review score (claims of accommodation facilities that were not actually available, for example), or what topics might receive the same (these topics might reveal unpleasant sentiments for guests). However, it was still valuable to know what terms and topics were common in listings that garnered high review scores. Such insight would advise a potential host on what kind of words or accommodation facilities would appeal more to guests. In addition, different regions in Australia yield different results. For example, the word cloud for Queensland yielded the terms "Beach" and "park" as the 2nd and 3rd most common terms, while the word cloud for Queensland yielded only "park." This indicates that Queensland listings that garner high review scores are likely near a beach and made that explicit in the listing description, whereas listings in Victoria does not seem to have that, but still can highlight their proximity to parks.

7. FUTURE WORK

Based on the exploratory spatial data analysis, we noted that prices are highly driven by location-based factors. This is further confirmed when the MLR fails to account for a large portion of the variation in price. A Geographically-Weighted Regression (GWR) can be introduced in future editions to enhance the performance of the predictive model.

Other sentiment analysis techniques such as LSA could also be included, to provide the analyst with a more comprehensive suite of statistical analysis tools. Complex statistical analysis such as hierarchical clustering can also be included, with the support of adequate computing hardware.

Deeper analysis could also be done on the profile of hosts as well as guests, for example the type of hosts that are more likely to have listings with good reviews, or the type of guest are more likely to give good reviews, for a particular listing type.

On a whole-of-Australia scale, there are numerous LGAs with missing data. In order support a more robust and holistic geospatial association analysis, these information gaps will need to be plucked.

Future versions of such an application could include other countries to provide more comprehensive analysis of different regions and markets. FOr example, Airbnb listings in South Africa may provide different insight into what hosts provide and what guests want, compared to Australia or other countries.

References

[1] Airbnb 2020. 2020 Airbnb Update.

[2] Anselin, L. 2020. Contiguity-Based Spatial Weight.

[3] Anselin, L. 2018. Distance-Based Spatial Weight.

[4] Australian Bureau of Statistics 2020. Australian Statistical Geography Standard (ASGS): Volume 3 - Non ABS Structure.

[5] Brand Education 2019. Airbnb Inc.

[6] Chen, S. 2019. How to Analyze Airbnb Performance Data in the Right Way.

[7] Cvitanic, T. et al. 2016. LDA v. LSA: A Comparison of Two Computational Text Analysis Tools for the Functional Categorization of Patents.

[8] Gedik, E.B. 2020. Seattle Airbnb Listings Analysis.

[9] Gupta, S. 2019. Airbnb Rental Listings Dataset Mining.

[10] Inside Airbnb 2021. Inside Airbnb.

[11] Kwon, J. et al. 2020. Airbnb Listings Data Analysis with R (Analyses).

[12] Latent Dirichlet allocation: 2021. *https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation*.

[13] The Comprehensive R Archive Network 2021. The Comprehensive R Archive Network.