Data Science Professional Development

INFS 5099

2021 Study Period 2

AirBnb Data Analysis Project

Project Members:

- Phouthun Nay (naypy003)
- Soriyavithya Neang (neasy005)
- Jinxi Luo (luojy025)
- Chandra Bahadur Gurung (gurcy003)
- Wai Cho Kwan (kwawy009)

25 June 2021

Table of Contents

Tables

Figures

1 Introduction

Founded in 2008, Airbnb has grown to become the largest online lodging marketplace today. Airbnb's website allows users to put parts of or entire properties on lease and allows guests who need a place to stay to have a short- or long-term rent. This online marketplace for lodging, therefore, allows anyone to become a renter and connects them with tenants worldwide. Renters can set their prices; however, guests can also leave reviews on properties and the renters themselves. Online lodging marketplaces such as Airbnb nowadays offer an alternative to travellers who prefer not to stay in a hotel, especially for short-term stays, and offer a convenient online option for quickly comparing properties on lease. Listings can be set up by anyone with a spare room or property, and the features on offer can vary significantly from listing to listing.

Prices are set by the hosts themselves, but as guests can easily compare prices, these factors such as amenities and extras (e.g., breakfast included and swimming pools) now become critical as they can affect the price significantly. Hosts who set unreasonably high prices or offer poor services will receive poor reviews, serving to warn other guests away from them. Hence, this investigation seeks to discover the factors that affect prices, rankings and reviews of the properties listed in Airbnb. The project will first focus on the Australian dataset, to analyse the data and understand the relationship between factors within Australian Airbnb listings. It is of interest to identify patterns of different factors within property listings and understand their relation to prices and review sentiment. By doing so, the results could help users find listings that are within their budget and according to their preferences. Moreover, it is possible to help property owners understand whether they are overpricing or not, and hence to improve the renting rate of their listings. This can provide a win-win situation for both guest users and the hosts. Lastly, via previous analysis results, it may be feasible to build a predictive model to help suggest prices for listings on Airbnb. A property's price or review rating based on different input factors such as rooms and residence location can be recommended as a reference.

With a better understanding of the data and descriptive knowledge from predictive models, results in the analysis may be extended to other countries. In general, this could improve the occupancy rates for all Airbnb properties, providing a benefit to both guests and hosts on the platform.

2 Literature Review

Amongst the literature reviewed, it was observed that the focus around Airbnb research is based around understanding trends and success factors. The dominant technique used is based around predictive modeling with a set of engineered features around location, housing information, reviews and renter information as factors. Common research goals were on prediction of rent price, review score and understanding spatial distribution. Research performed in this report takes inspiration from previous research to be applied to the Australian Airbnb dataset. Most notably, findings from the literature have

supported the location of an Airbnb listing to be an important factor. Using predictive modeling with distance of different locations with respect to Airbnb homes. One paper was able to find that successful Airbnb listings tend to be located around popular and essential public areas such as shopping malls, scenic spots and city centers (Sun, Zhang and Wang, 2021). These findings have an implication on the future success of future Airbnb popups as well as the general economic situation, as it has been found that an increase of an area's Airbnb housing also increases the number of traditional accommodations (Leick et al., 2021). The availability of Airbnb lodgings also promotes tourism which in turn drives the local economy (Leick et al., 2021) through the attraction of higher talent and creative social demographics (Quattrone et al., 2018). The adaptation of what renters consider to be profitable housing which has also seen evolution from small homes to larger multiple floor apartments (Sun, Zhang and Wang, 2021).

Sentiment analysis is another vital technique that has been applied previously with Airbnb data, while often used alongside predictive modeling (Kalehbasti, Nikolenko and Rezaei, 2019; Chiny et al., 2021), the significance of the analysis begins with appropriately categorizing texts. A research that has done this quite well as able to show that customers values factors such as, accuracy of information listed, cleanliness of homes, value of rent and location of homes when considering their final review score (Chiny et al., 2021).

Given the available data for this research, the use of reviews, sentiment, prices, location and property information will be of primary focus as those have been the variables that had been successful predictors in the literature (McNeil, 2020; Sun, Zhang and Wang, 2021; Chiny et al., 2021).

3 Data Summary

The data used for our analysis project was taken from the *Inside Airbnb* project, a publicly available collection of tools and data about Airbnb listings. Our project focused on data for Australian Airbnb listings only, available from http://insideairbnb.com/get-the-data.html, and was last collated and updated on 09 March 2021. The Australian dataset has multiple files available, with their description as follows:

- **listings.csv:** contains data regarding listings and hosts details.
- **calendar.csv:** contains data regarding the availability of listings.
- **reviews.csv:** contains reviews left by guests on listings.
- **neighborhoods.csv:** contains geo-positioning coordinates for listings.

Combined, the files contain enough information for us to perform analysis on a variety of factors regarding the listings, including their prices, ratings, locations and types of the room.

3.1 Data Exploration

Before performing analysis on our datasets, we first explore the Airbnb's listings distribution by their prices, room types, and State/Territory to gain some preliminary information regarding our data.

3.1.1 Listings prices distribution

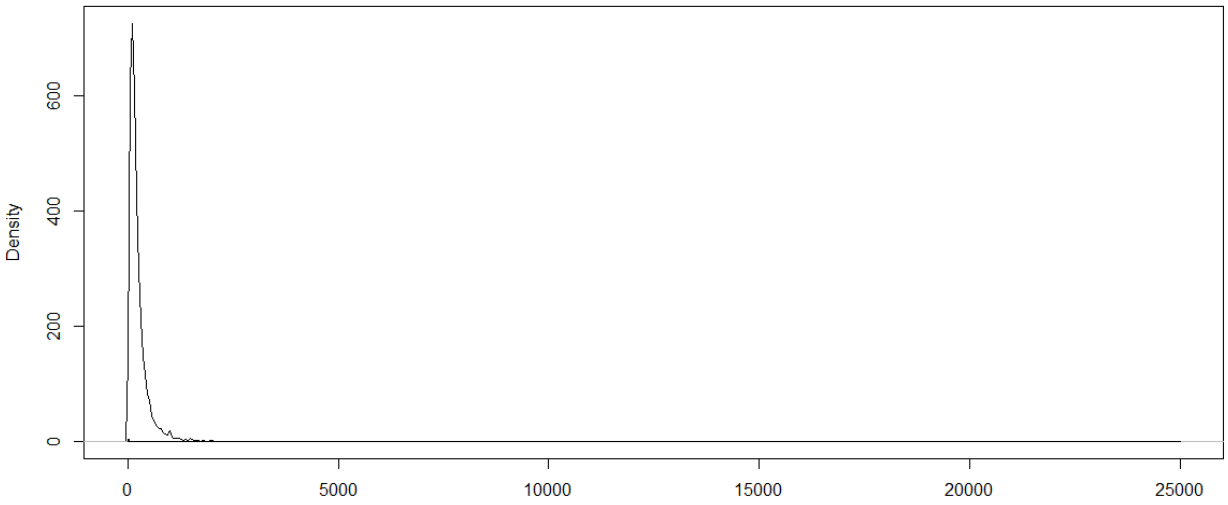

Figure 1. Listings Price Distribution

Looking at the distribution of listings' prices, most listings appear to be under $2500/night, with some extreme outliers. If we remove the outliers, we get a new price distribution and density graph as shown in the figure below.

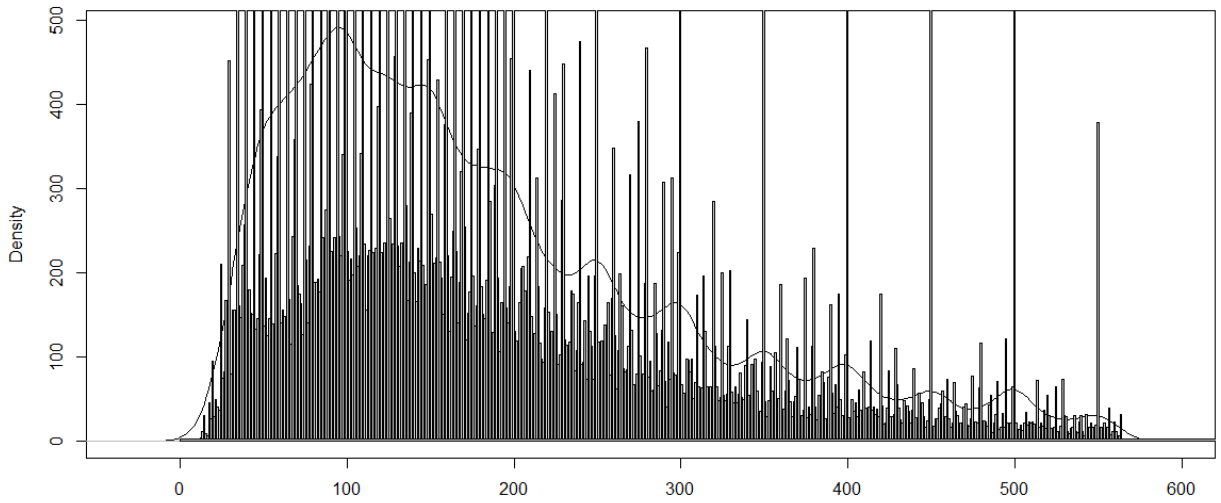

Figure 2. Listings Price Distribution without outliers

The most frequent price appears to be $100/night and there appears to be spikes in the number of listings at certain price points, in every $50 increment (e.g., $150, $200, $250 etc.), indicating a certain preference for hosts to set their prices at 'round' price points.

3.1.2 Listings by room type

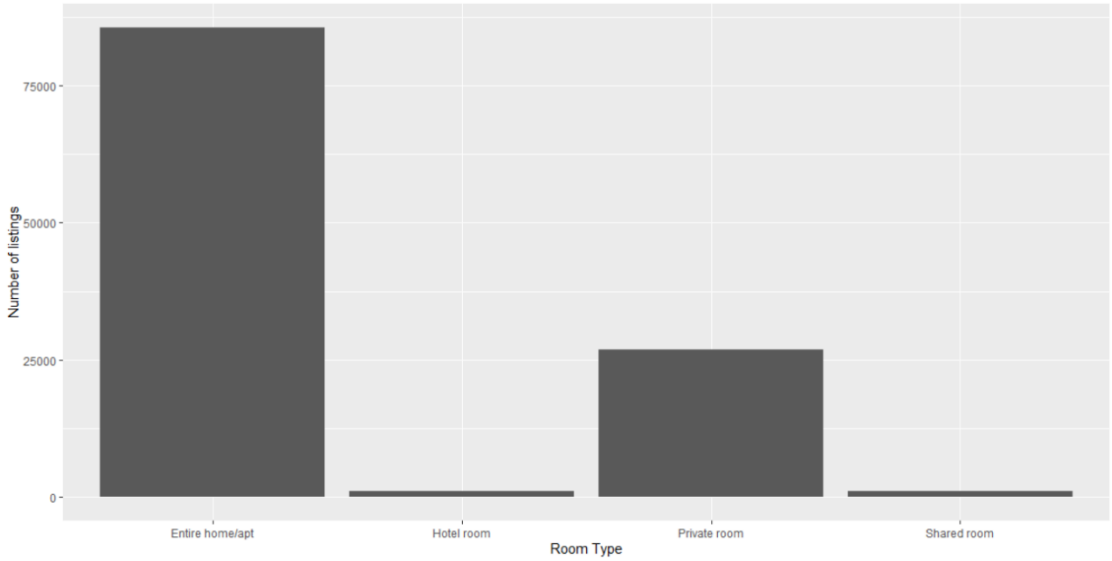

Figure 3. Listings per room type

Looking at the data we obtained, most listings are for an entire home or apartment, followed by private rooms. There are only a few listings for shared rooms or hotel room types available in Airbnb, suggesting there is not many demands for those room types.

3.1.3 Listings by State/Territory

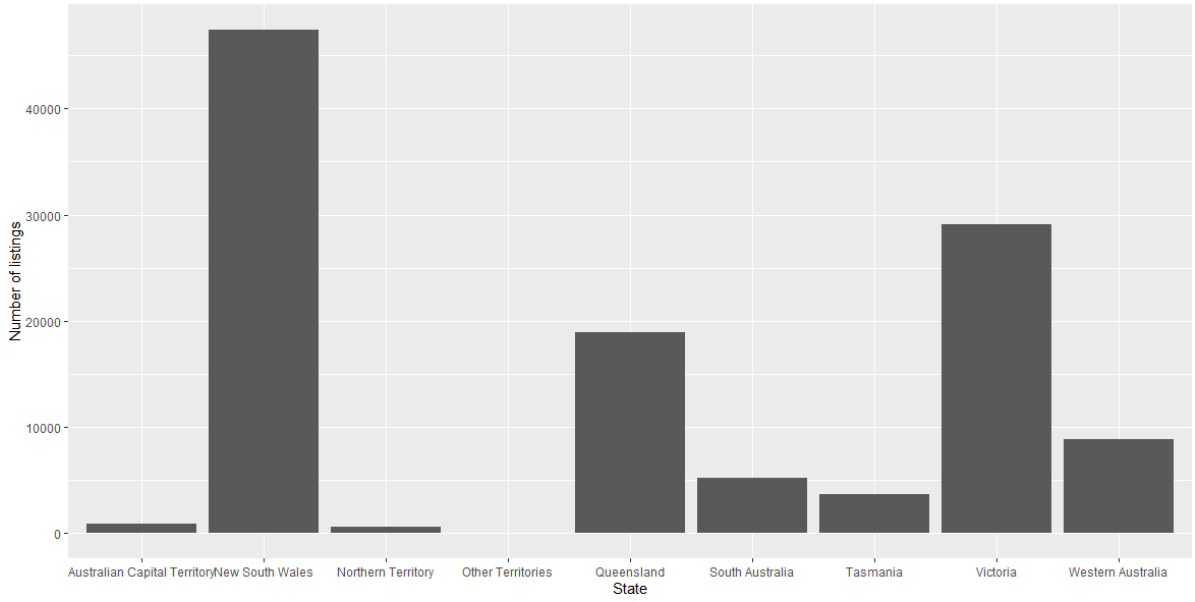

Figure 4. Listings per region

The data shows that NSW has by far the most listings, followed by VIC, QLD, WA, SA and TAS in that order. ACT and NT has very few listings available compared to the rest of the country. There is also a very small

number of listings in Other Territories. The number of listings per state appear to be influenced by the population of each state, as more populated states have more listings than less populated states.

4 Data Analysis

4.1 Prices

One of our tasks is to analyse one year of Australian Airbnb renting data, to identify which factors affects the listings' price in that year, then use this information to build a prediction model to predict the renting price for Airbnb properties. We will first investigate the details of different properties such as location, available facilities, and the property/room type, then compare with the renting/booking details, to see if there are any important relevant factors.

Two files from the dataset are used for this analysis task: *listings.csv* which shows all current registered Airbnb, their location, price, and all kinds of details about the property, and *calendar.csv* which shows booking information for the full year for all properties, which day is available and the booking price of the day etc.

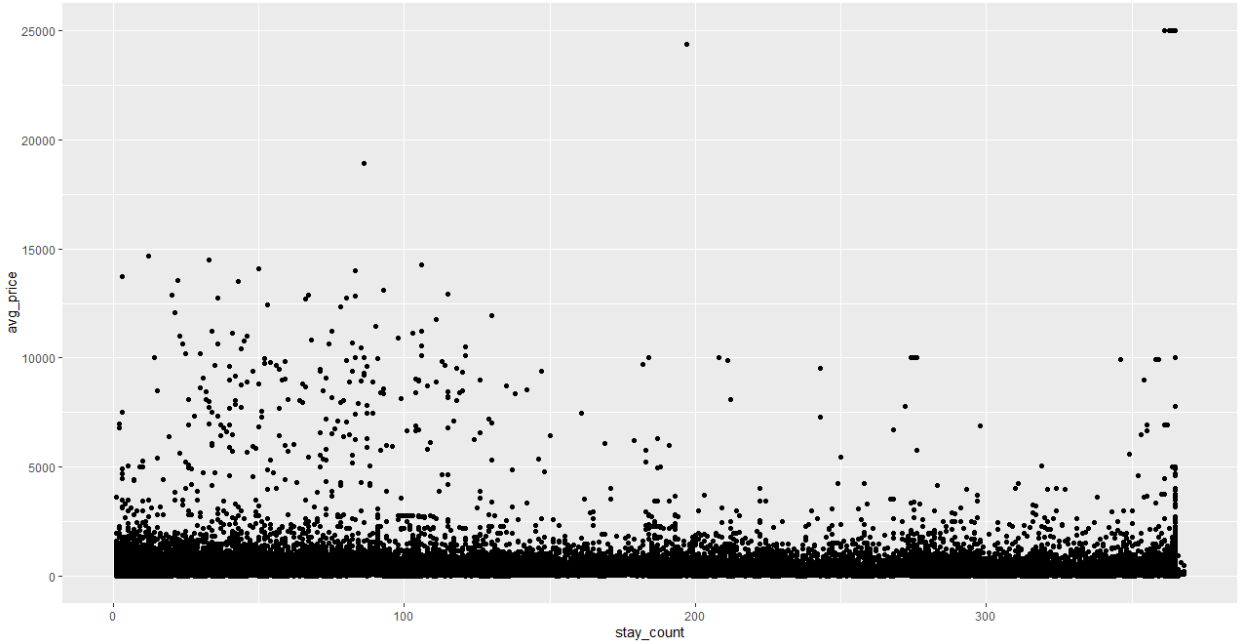

Figure 5. Average price by number of stays scatterplot

Calendar data includes more than 50 million rows of data, with only 7 factors. Listings' data has only slightly over 110 thousand rows of data, but with 80 factors. These 2 datasets have very differing structures, and we therefore need to perform pre-processing before we merge them together to train a prediction model.

First, we use data from the Calendar dataset and check the relation between price and booking status. We group data for every property to get an average daily price and count the total booking days for each property. In general, we assume that if a property has many bookings during the year, this means that it is

more popular, and the average price should be therefore higher. However, the above plot shows a different result. The graph above shows that the average price and booking days do not have a linear relationship, which means the price could be driven by other factors.

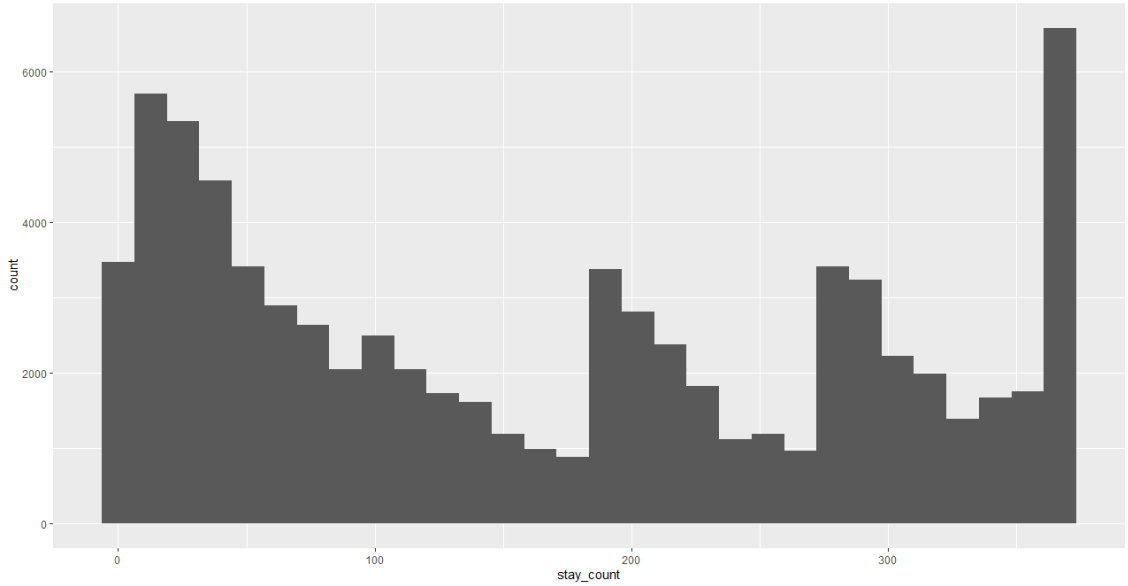

Figure 6. Listings' number of stays distribution

We then also look at historical data of booking days to see if there are any unusual data pattern. We also obtain some unexpected results. In general, we assume this data should be normally distributed, because in general most property should have similar bookings numbers. Based on city design patterns, residential building should normally be in the same area, hence the locations or the surrounding environments should also be similar. Therefore, most of the properties should have similar bookings, only a small number of properties in different areas should have an extremely high or low number bookings.

However, the graph above shows an unexpected result again. This result also indicates there could be some other factors which influences how people choose to stay in only some properties.

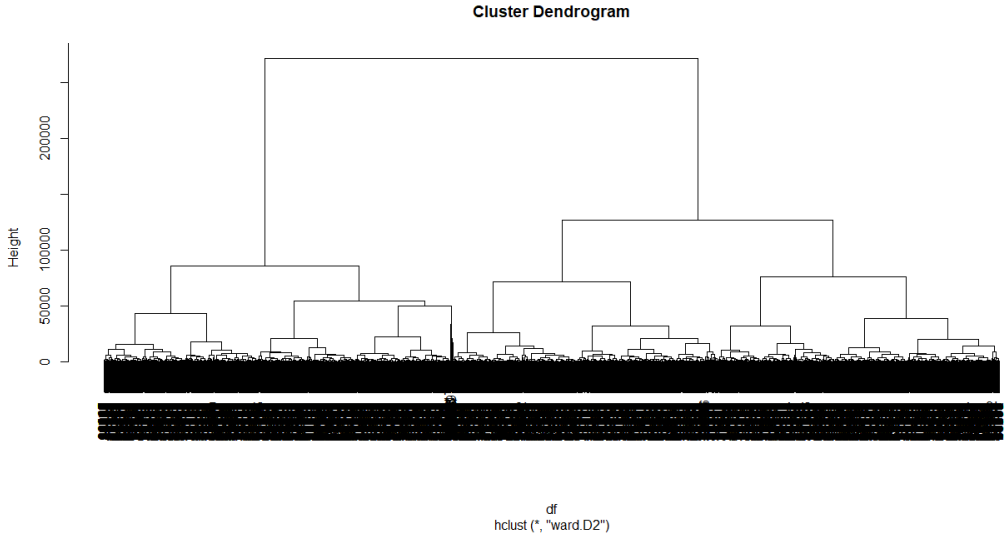

Figure 7. Cluster dendrogram for initial price analysis

Before we investigate each parameter in both datasets, we first look at the Listing dataset from an overall perspective. We first proceed with unsupervised clustering to examine if there are any similarities or correlation between different properties.

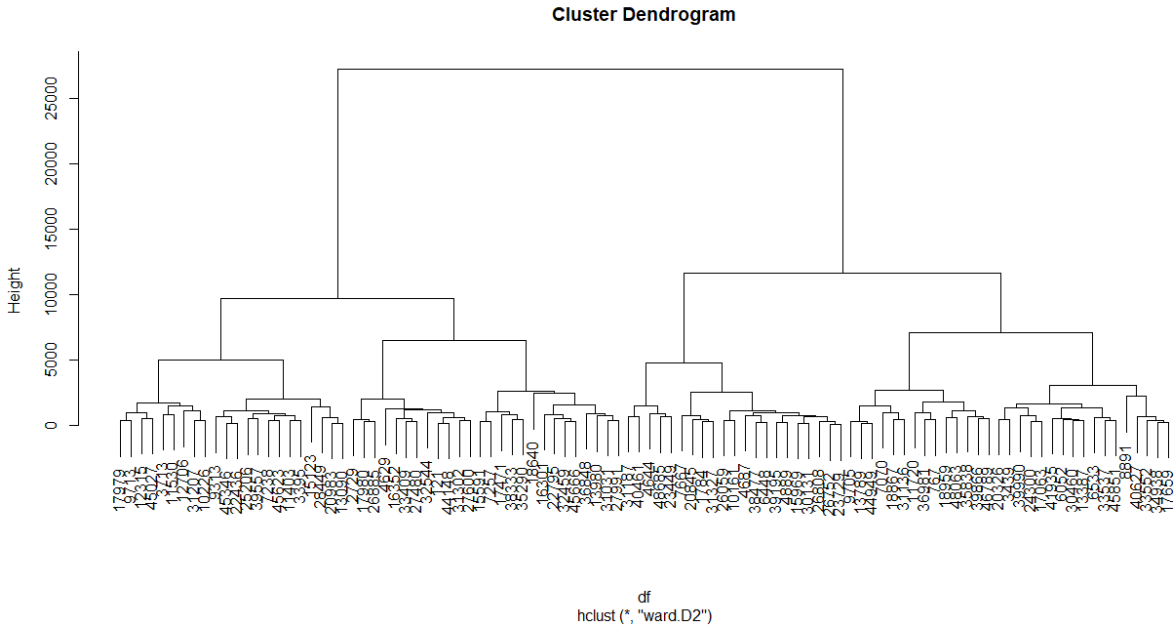

Figure 8. Cluster dendrogram with sampled data

However, the results obtained include too many rows of data, we therefore only take 1,000 random rows of sample data to make the result clearer. We can see there are mainly 4 clusters, but we cannot determine

from which height to separate the clusters, the result could therefore be 4 clusters, 8 clusters or more depending on the selected height.

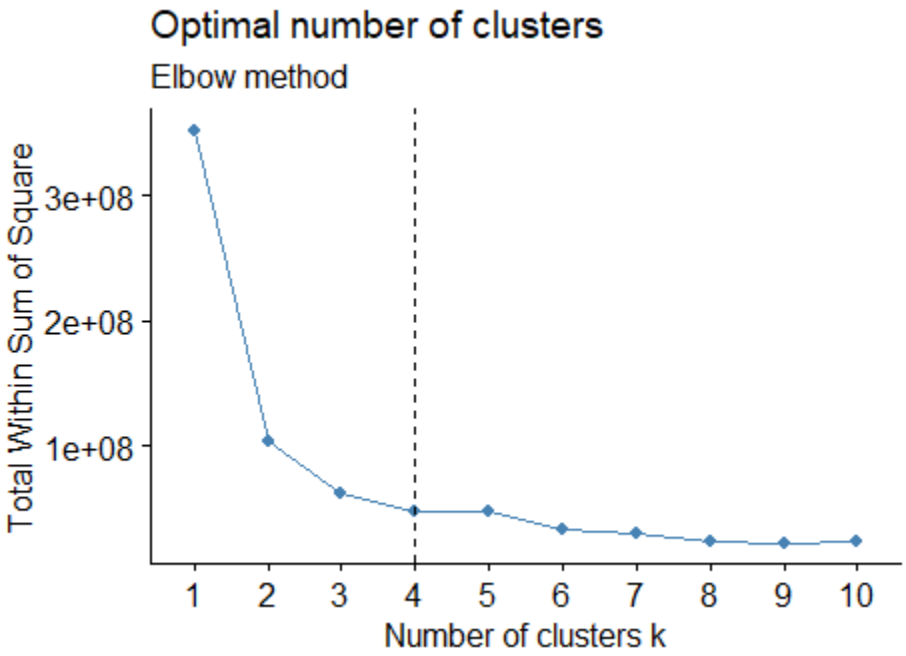

Figure 9. Elbow method results for prices' analysis

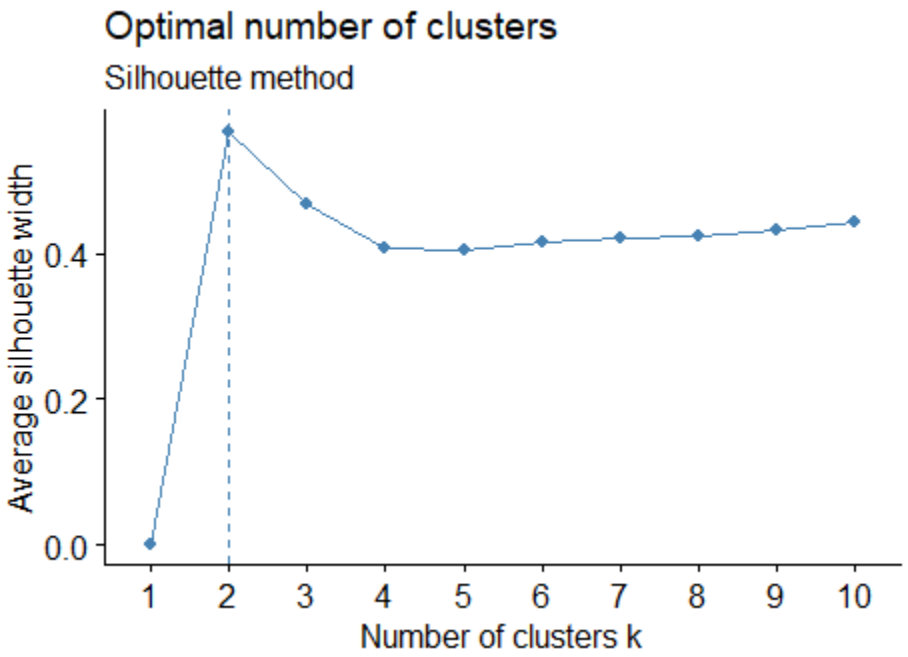

Figure 10. Silhouette method result for prices' analysis

Hence, we use the Elbow and Silhouette methods to determine the optimal number of clusters. However, these 2 methods returned different results, so we therefore try both suggested numbers to proceed with the k-means clustering.

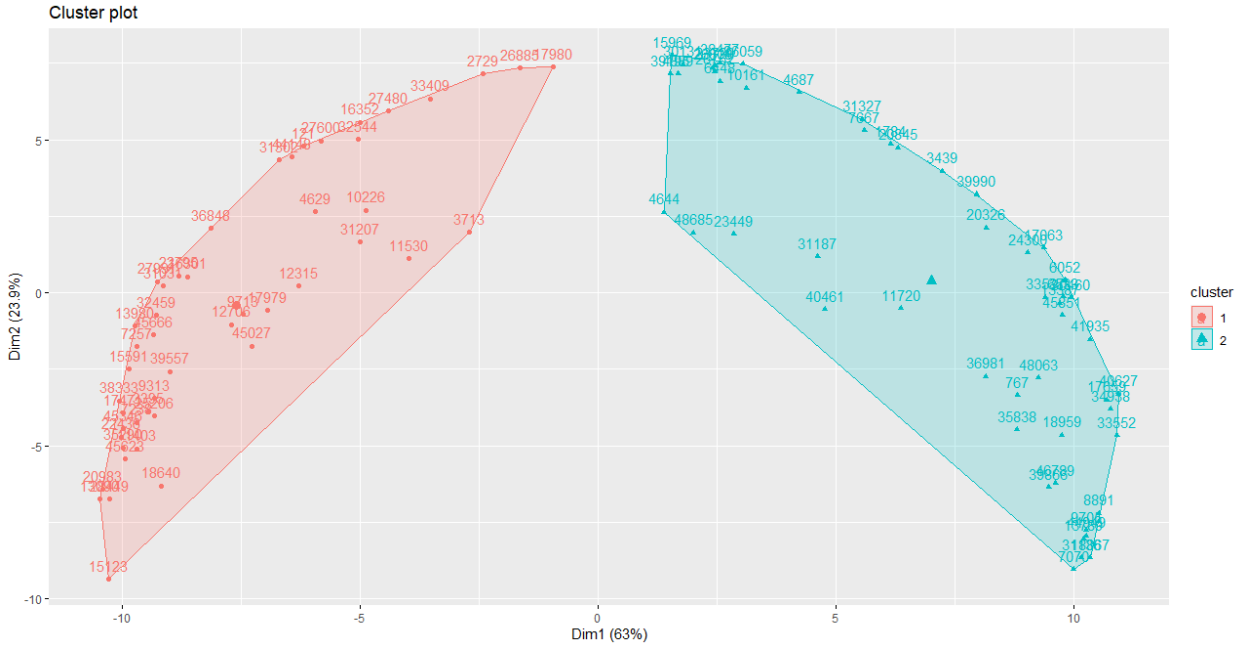

Figure 11. Dimensions scatter plot with K = 2

First, we proceed with k-means clustering with k = 2. Each datapoint corresponds to a property, with numbers being the property ID. We can see that these 2 clusters are clearly split, but there are some smaller cluster patterns within each cluster.

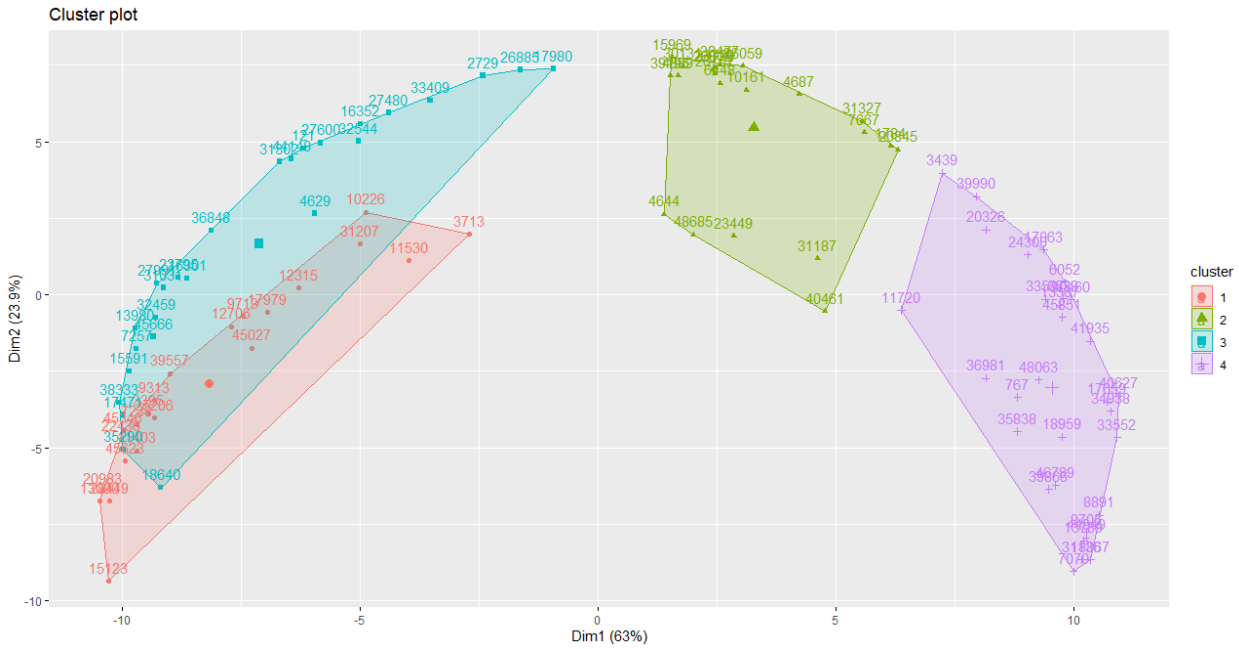

Figure 12. Dimensions scatter plot for K = 4

Therefore, we proceed with k-means with k = 4. Now the clusters on the right are improved and split to form 2 distinct clusters, however the clusters on the left are overlapping.

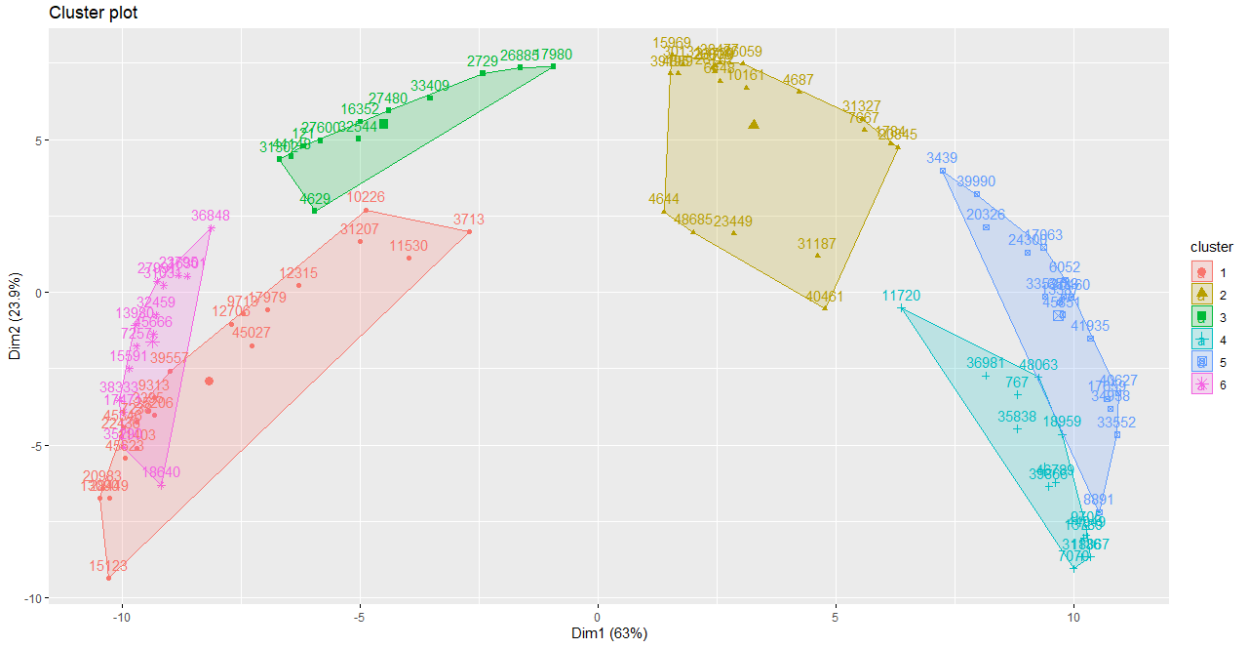

Figure 13. Dimensions scatter plot for K = 6

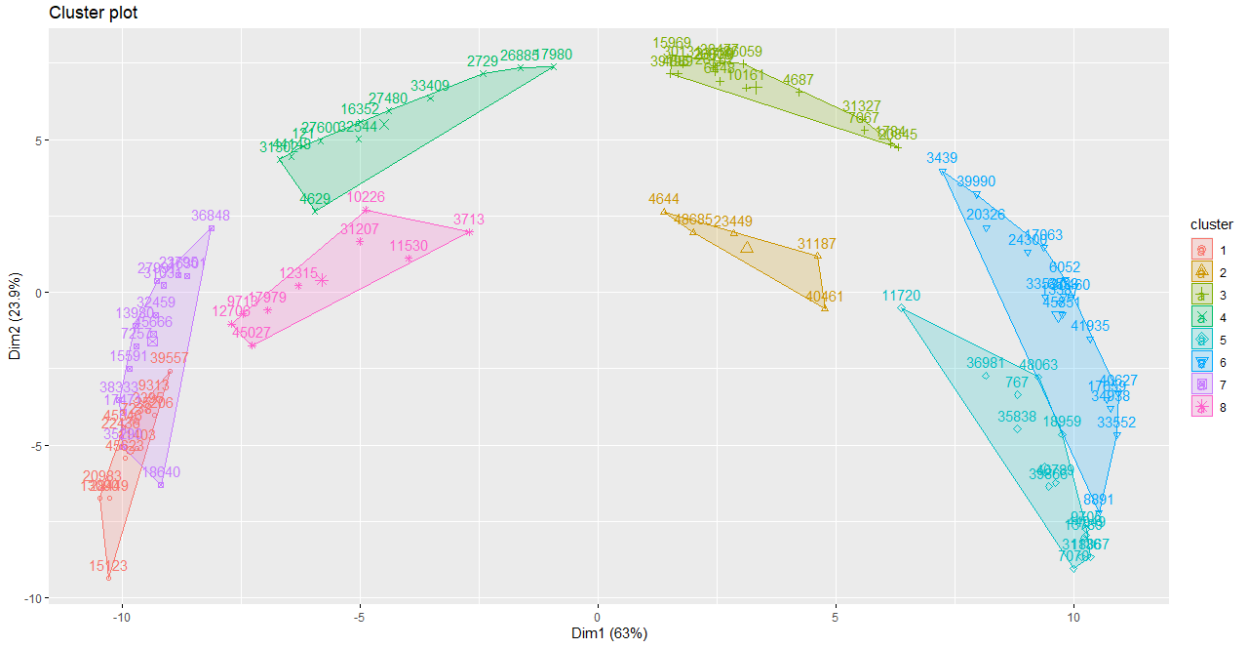

Figure 14. Dimensions scatter plot for K = 8

After going further with larger k for k-means clustering, we found k = 6 is the best option. We can see that these 6 clusters are well split, even though there is still some slight overlapping of the clusters. When compared with the results for with k = 8, however that overlap still exists, which shows that overlapping of clusters may not be easily removed.

After some analysis and clustering, we have an overall concept of the Airbnb dataset. First, we understand the renting price and the number of booked days have no direct correlation. This means the cheapest properties do not always have full booking across the year, while the expensive ones could have a lot of bookings instead. Second, the data pattern of the bookings also implies there could be other factors affecting the prices and the booking status.

Lastly, as the clustering shows, there may be 5 or 6 factors to differentiate properties into different categories of groups. Hence, we think there may be at least 5 to 6 factors could help us build a prediction model.

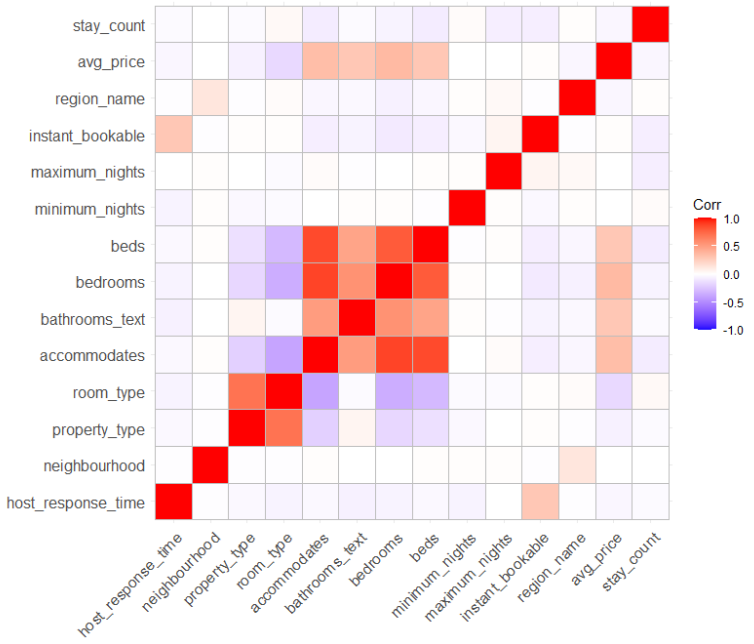

Figure 15. Correlation matrix of potential factors affecting price

Next, we go into details and check all parameters. As there are over 80 parameters in the listings dataset, we cannot describe them one by one. After some data cleaning and processing, we pick 12 parameters that we believe would have predictive values for our prediction models. We then check the correlation between them and price.

As we can see on the above graph, Price has a strong correlation with the number of beds, bedrooms, and bathrooms as well as the number of guests it can accommodate. It also has a slight to strong correlation with Room type and Property Type, which confirms our k-mean clustering (k = 6) results above.

After pre-processing, the final dataset has 76,935 rows of data remaining. The parameters we pick to build a prediction model are Neighbourhood, Property Type, Room Type, Accommodates, Bathroom's detail,

Bedrooms number, Beds number, Amenities, Region. Our prediction target is Renting Price. Even though Region and Neighbourhood may not be the most important factors, it may still be a good idea to include them for our models.

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	170.43772	11.34952	15.017	< 0.0000000000000002
neighborhood	-0.01253	0.00135	-9.287	< 0.0000000000000002
property_type	0.97440	0.25404	3.836	0.000125
room_type	-48.09163	4.88190	-9.851	< 0.0000000000000002
accommodates	26.25947	2.08474	12.596	< 0.0000000000000002
bathrooms_text	12.19056	0.65726	18.547	< 0.0000000000000002
bedrooms	61.86542	3.73039	16.584	< 0.0000000000000002
beds	-8.24511	2.07945	-3.965	0.0000735
amenities	-3.46076	0.23523	-14.712	< 0.0000000000000002
region_name	0.01156	0.02577	0.449	0.653752

Table 1. Results from predictive modelling

- Multiple R-squared: 0.09685, Adjusted R-squared: 0.0967
- F-statistic: 641.5 on 9 and 53844 DF, p-value: < 0.00000000000000022

As all the factors are either interval or ordinal, we can convert them all to numerical factors and proceed with multi-linear regression model. As we can see in the summary table above, most of input factor are highly significant judging by their p-value. However, the most important factor is bedroom numbers and the next one is room type. The least important factors are neighbourhood and property type, which means in effect that properties with more rooms command a higher renting price.

Below is some information regarding our model training steps:

- Step: AIC=686969
- Formula: avg_price ~ neighbourhood + property_type + room_type + accommodates + bathrooms_text + bedrooms + beds + amenities

	Sum of Sq	RSS	AIC
<none>		18663484682	686969
property_type	5105175	18668589857	686982
beds	5449654	18668934336	686983
neighbourhood	29832351	18693317033	687053
room_type	33569503	18697054185	687064
accommodates	54974447	18718459129	687125
amenities	75188635	18738673317	687184
bedrooms	95385899	18758870581	687242

bathrooms_text	119171083	18782655765	687310

Table 2. Model training step 1 results

- Start: AIC=686970.8
- Formula: avg_price ~ neighbourhood + property_type + room_type + accommodates + bathrooms_text + bedrooms + beds + amenities

	Sum of Sq	RSS	AIC
<none>		18663414941	686971
property_type	5099332	18668514272	686984
beds	5449410	18668864350	686985
neighbourhood	29896550	18693311491	687055
room_type	33636820	18697051761	687066
accommodates	54994654	18718409594	687127
amenities	75027879	18738442820	687185
bedrooms	95332719	18758747660	687243
bathrooms_text	119240628	18782655569	687312

Table 3. Model training step 2 results

The Sum of Square errors decreases for every step of training, showing that the input parameters for this model have a positive effect on improving accuracy.

We then use our trained model to predict listed properties' prices for part of our dataset.

Figure 16. Predicted vs Actual price scatter plot

	ME	*RMSE*	*MAE*	*MPE*	*MAPE*
Test set	-6.35963	668.0152	186.3661	-9.112506	78.36734

Table 4. Price prediction model validation metrics

In this training process, we used 50% of data to train the multi-regression model with the resulting prediction accuracy being relatively good. As you can see the prediction result and the actual value has a close to linear pattern. This shows that the accuracy is quite high, however we can still see there are quite a few data points which are far away from the main pattern and indicates that there could be some improvements in accuracy. This report also proves location has no effect on Australian Airbnb rental prices, the reason for that could be that the property itself matters far more than the specific location it is situated in, as guests could easily travel to their intended destinations provided the listing is in any large urban centre, which is the main case in Australia.

4.2 Reviews and Ratings

To understand what guests are looking for when selecting and renting out a listing on Airbnb, we can perform sentiment analysis on the text left by guests in their reviews of the listings as well as analyse which factors have the most effect on the star rating of listings.

Airbnb automatically reminds guests to leave a review and rating after they have checked out and encourages guests to leave a review by only allowing them to view the review that the host has made about them after they have written and posted their own review. This system therefore ensures that most listings have at least several reviews written about them, helping potential future guests to make their decision on which place to stay.

The views expressed in the reviews left by guests can therefore have a strong impact on how likely a listing is to be rented. By analysing the sentiments, both positive and negative that guests have expressed in their reviews, we can therefore gain an understanding regarding what guests liked and what they disliked during their stay.

Our sentiment analysis mainly uses data from the *reviews.csv* file, which contains the text reviews that guests have left on listings they have stayed at. The file contains 4,492,202 reviews for 126,525 different listings. We first perform text pre-processing on the review texts by removing all punctuation, numerical digits, and other special characters, then converting all of the text to lowercase characters. We also removed all common English stop words from the texts, using the default list of stop words in R's tm package for text mining (Feinerer and Hornik, 2020). We then compared the list individual words in the reviews to the words present in the sentiment lexicon by Hu & Liu (2004) and could therefore obtain an associated sentiment for some words, either positive or negative. We thereafter only retained words which had a sentiment associated with them, which was only a part of all the words that were used in the reviews, and from which we could then obtain a list of words associated with positive sentiment, and a list of words associated with negative sentiment. Some additional clean-up of the list of words was required however, as some of the words in each list were not indicative of a particular attribute of the listing e.g., "great" was the most common positive sentiment word but does not inform us what aspect of their stay the guest liked. Some words were also removed as the word itself did not necessarily express a sentiment but was

considered as such by the sentiment lexicon, e.g., “retreat” was one of the more common negative sentiment words but was likely meant in the sense of a break or vacation in the context of the Airbnb reviews, rather than a defeat or escape as the association with a negative sentiment would imply.

With the final list of both positive and negative sentiment words, we can obtain the following data visualisations:

Figure 17. Word cloud of overall most commonly expressed sentiments words

From the word cloud of the 150 most commonly expressed sentiment words, with positive sentiment words in blue and negative sentiment words in red, we can observe that the vast majority of sentiment words used in reviews by guests are positive, with only a few negative words appearing. This appears to indicate that overall, most guests are satisfied with their stay and leave positive reviews.

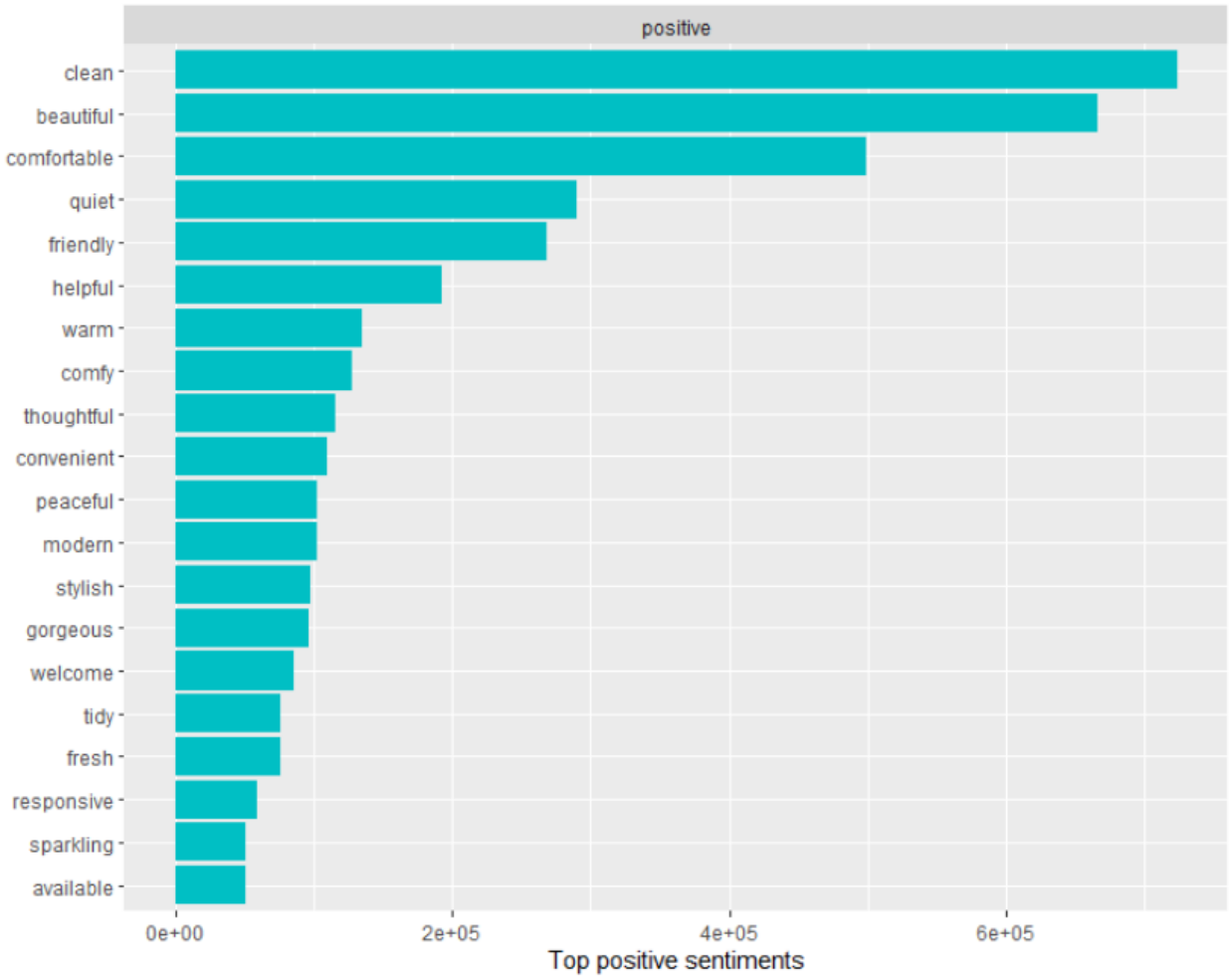

Figure 18. Most expressed positive sentiment words in reviews

From the list of the 20 most commonly expressed positive sentiment words across all reviews, 3 stand out in particular: clean, beautiful, and comfortable, indicating that guests valued these 3 qualities above all else during their stay. Other valued qualities that were less often cited include the quietness of the lodging (quiet, peaceful) and the accommodativeness of the host (friendly, helpful, thoughtful, responsive).

Figure 19. Word cloud of most commonly expressed positive sentiment words

The same observations can be made from the word cloud of the 150 most common positive sentiment words: guests favour cleanliness, beauty and comfort of the lodgings above all, with other positive sentiment words appearing much less frequently or expressing the same general qualities that guests seek out in different terms.

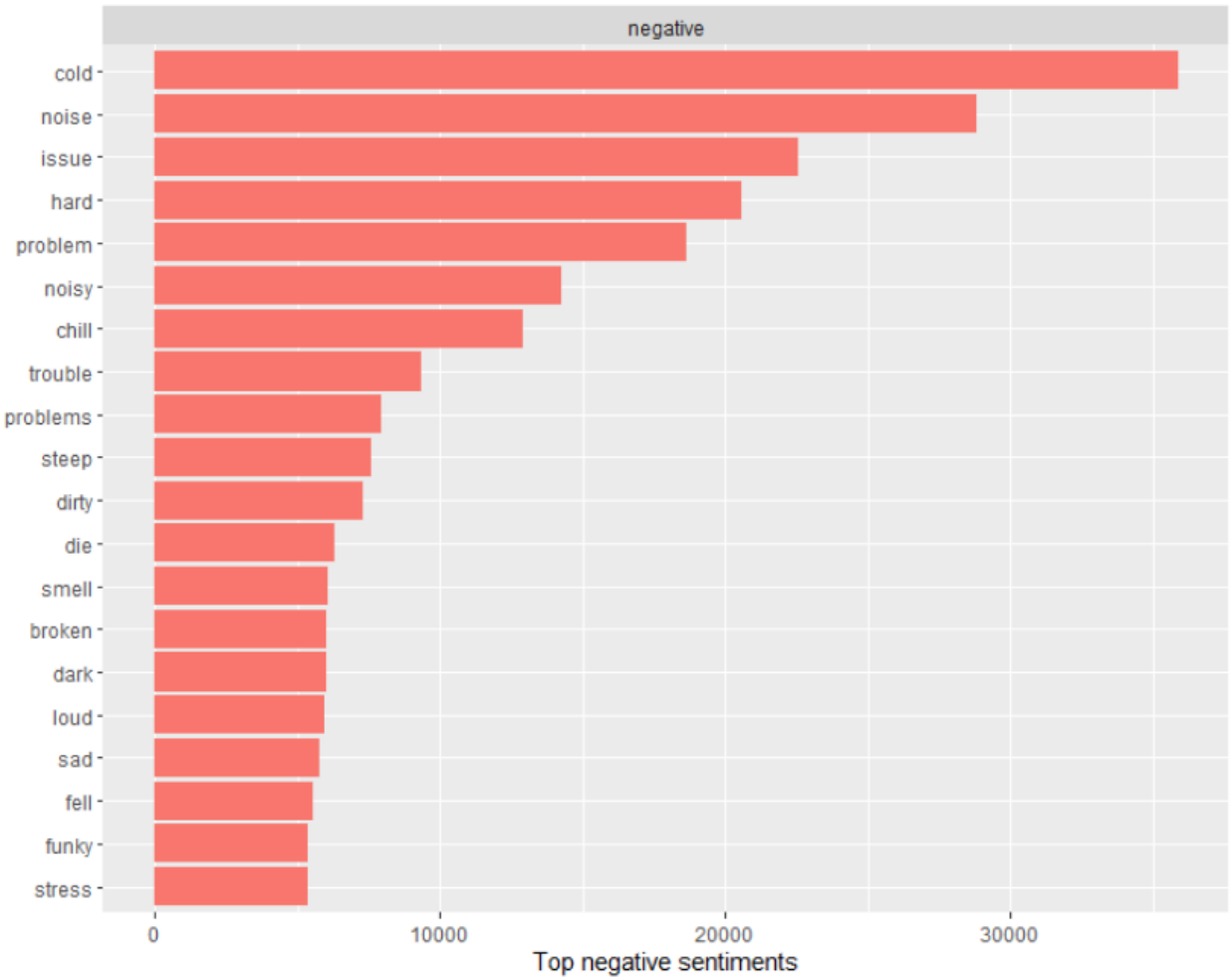

Figure 20. Most expressed negative sentiment words in reviews

Looking at the top 20 negative sentiment words, we can observe that negative sentiment words are firstly much less frequent than positive sentiment words. The most frequent negative sentiment word "cold" occurs around 35,000 times compared to the most frequent positive sentiment word "clean" which occurs about 70,000 times. The most common complaints judging by the use of negative sentiment words appear to be related to uncomfortably cold lodgings (cold, chill), noise (noisy, loud), bad smells (smelly, funky), dirtiness (dirty) and other problems associated with the state of the lodging (broken, dark, steep etc.).

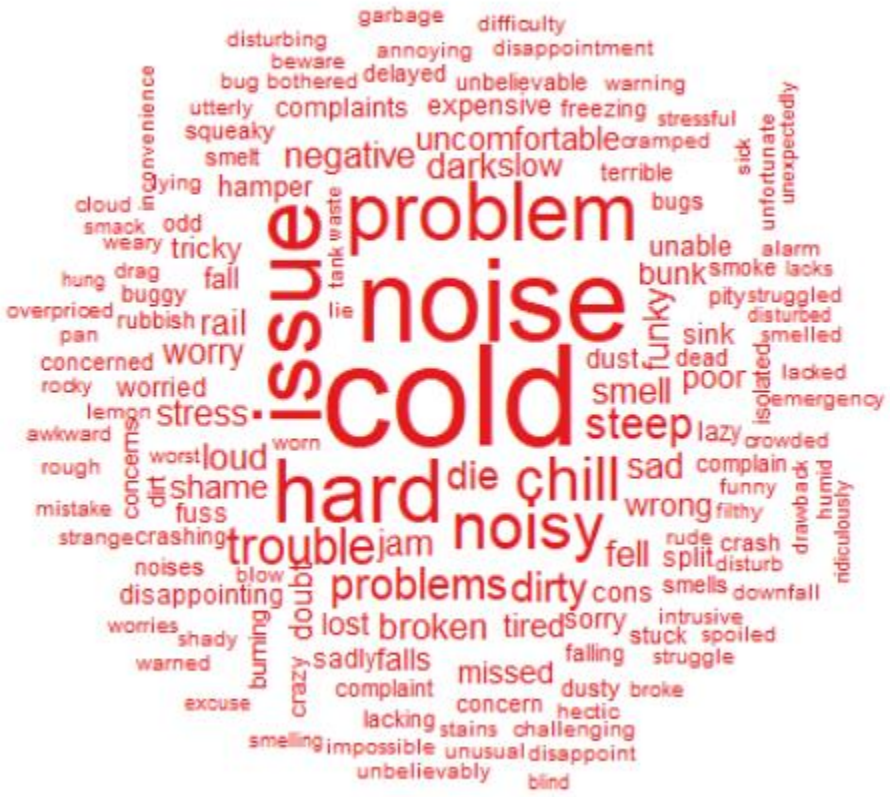

Figure 21. Word of most commonly expressed negative sentiment words

Looking at the word cloud of the top 150 negative sentiment words, in addition to the same previous observations where noise, cold and were the main negative factors expressed in user reviews, there also appears to be more variety in the words used to express negative sentiments compared to the words used to express positive sentiments, where only a few words dominated the word cloud. This perhaps indicate that when guests leave a negative review, they express their disappointment in a more detailed manner, listing specific aspects that they did not like as opposed to when they leave a positive review which is more generalised in its praise.

In addition to analysing the sentiment expressed by individual methods, we can also use syuzhet package in R (Jockers, 2015) to assign a sentiment score to each of the reviews in the *reviews.csv* file. The syuzhet package assigns a 'sentiment score' to text based on the appearance of sentiment words, with different words having different a different sentiment score value. Positive sentiment words have positive number values for the sentiment score whereas negative sentiment words have a negative number value. In addition to that, the stronger the sentiment expressed by the word, the greater the absolute value of the associated sentiment score. Once we obtain a sentiment score for each of the reviews in the data, which consists of the sum of the scores of the words in each review, we then calculate the average sentiment score for each available listing, as there are more often than not multiple reviews for each individual listing.

We can then obtain the following distribution plots for the sentiment score of the reviews:

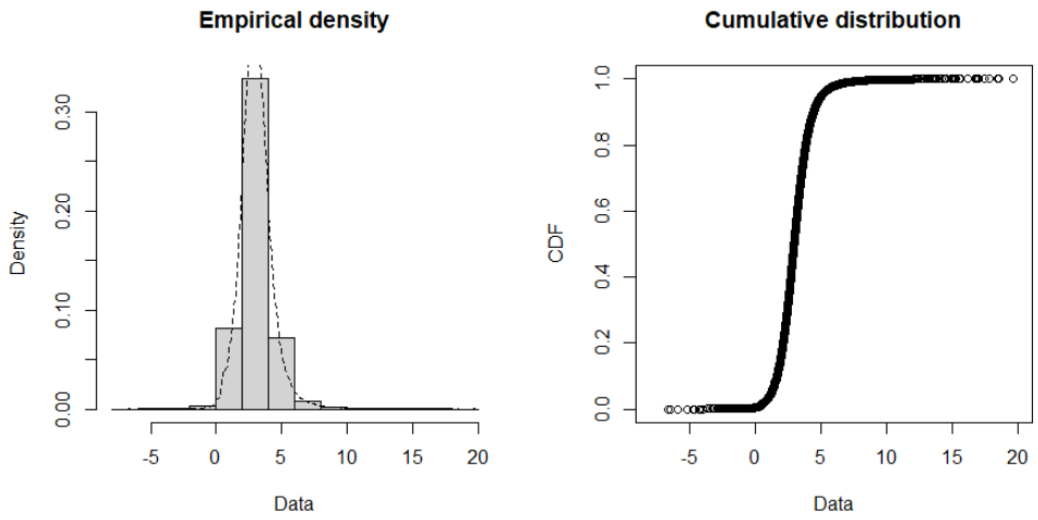

Figure 22. Distribution of review's average sentiment score

The distribution plot again confirms that the vast majority of reviews left by guests are positive, with very few reviews which receive a negative sentiment score. The majority of the reviews are between 0 and 5 sentiment scores, indicating that most reviews either are only slightly positive or that the reviews tend to be very short yet positive, therefore not having a very high score. Comparatively, there are very few reviews which have a sentiment score above 5.

Finally, we use SAS Enterprise Miner to determine the most significant correlations between a listing's final rating and the other attributes in a listing.

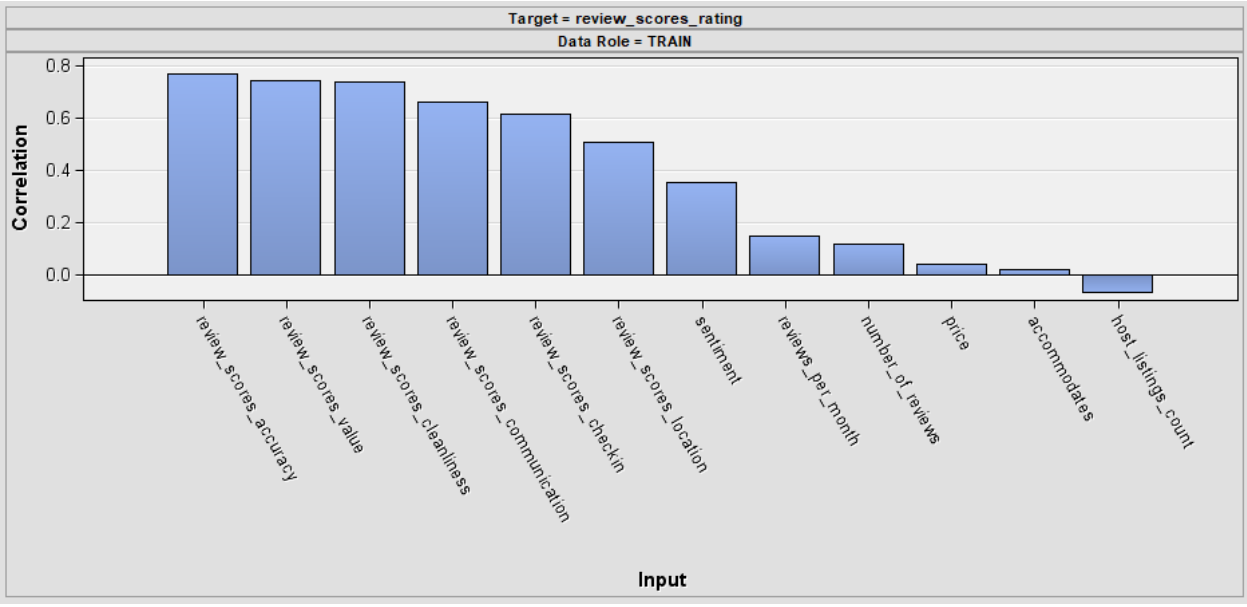

Figure 23. Correlation between listings rating and other listing details

The other review scores for a listing present in the dataset naturally have the highest correlation with the final rating score of a rating. The difference in the correlation magnitude between the different review scores however informs us that they do not all have the same importance towards the final rating. The accuracy of the listing's information, the value for money and the cleanliness of the lodging appears to be more important to guests than the communication by the host, how good the check-in was and the location of the lodging. The sentiment score of reviews left by guests for the listing also appears to have a high correlation with the final rating score. The number of reviews per month and total reviews also appear to have some correlation with the final rating, perhaps indicating that guests prefer listings with more reviews. Interestingly, price appears to have little bearing on the final rating of a listing, perhaps since Airbnb states all costs upfront, and therefore guests know exactly what they are paying for and how much. The total number of listings that a host has on the site is the only variable with a negative correlation to the rating score, indicating that hosts with many listings tend to receive lower ratings, perhaps due to them not being able to personally attend to all their listings, and therefore it is likely that those lodgings are of poorer quality.

4.3 Prices and Location

The initial interest of this section of research focuses on studying Airbnb properties (nightly) rent prices and location. The research sets out to identify any trends that may be seen within different geographic regions to help investors and renters make appropriate decisions based on their budget. Beginning with using clustering to search for any obvious relationships, further investigations are then carried out.

Basic data cleaning is performed to remove rows that have missing latitude, longitude, state name or price as the dataset is largely complete within these fields, resulting in 116,694 observations. Exploratory analysis with a boxplot as seen below on the distribution of price and its location can be seen to show that while price has a lot of outliers for extremely expensive homes. They are not removed as it makes sense contextually that expensive homes to rent are typically significantly more expensive.

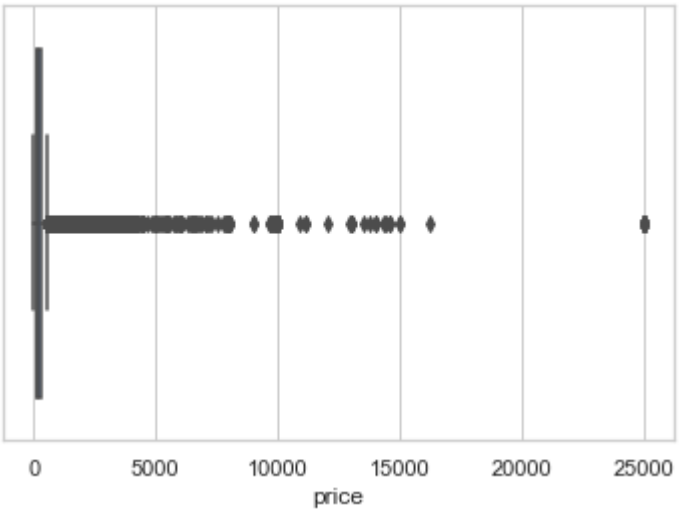

Figure 24. Boxplot of rent prices

The data shows that the typical home costs under $300 a night and on average $255 with a median of $163. And there being a lot more homes for rent and expensive ones situated around the biggest cities of Canberra, Sydney (New South Wales), Brisbane (Queensland), and Melbourne (Victoria). As shown in the boxplot for each region. Where the standard of living is often more expensive than the more isolated and less populated areas of Australia, which agrees with results found in literature. See heatmap below for details: blue to green to yellow indicating increasing Airbnb density.

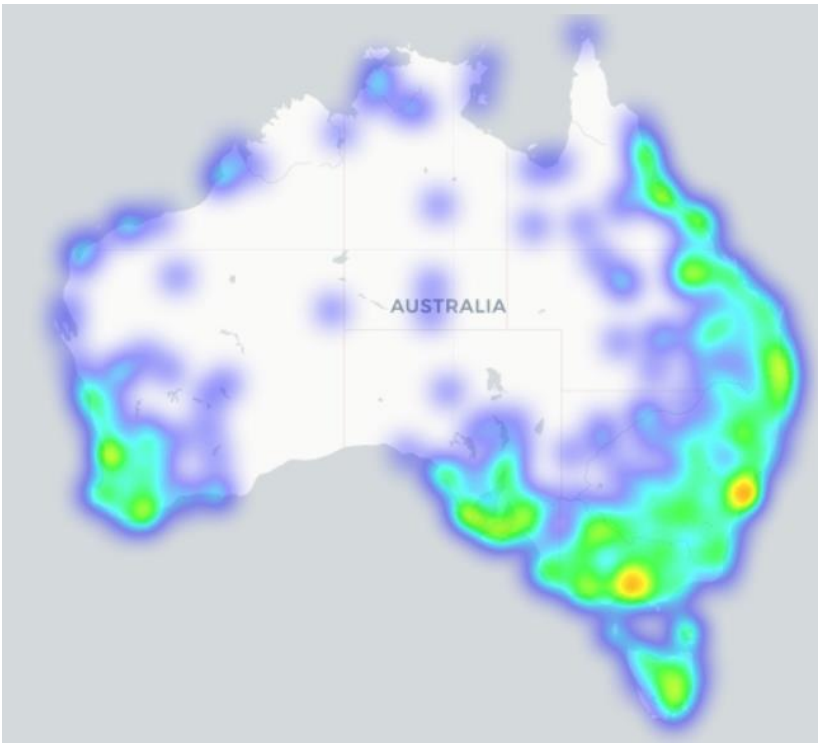

Figure 25. Heatmap of Airbnb listings in Australia

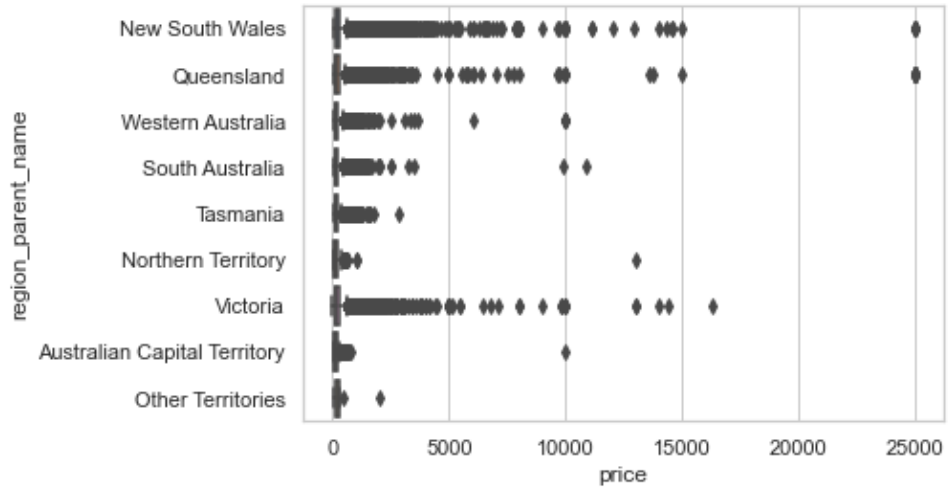

Figure 26. Boxplot of Airbnb prices in each region

For clustering, the K-means algorithm will be used for the analysis as it is simple and works well with numerical data types of geographical location and prices. Standardization is not performed to retain the context of the coordinate systems. Initial results show that the optimal number of clusters is around 5 clusters with the elbow method varying the number of clusters from 2 to 10. It was difficult to decide which solution to use so a comparison of the median prices using 4-6 groups is tested in the dimension scatter plots.

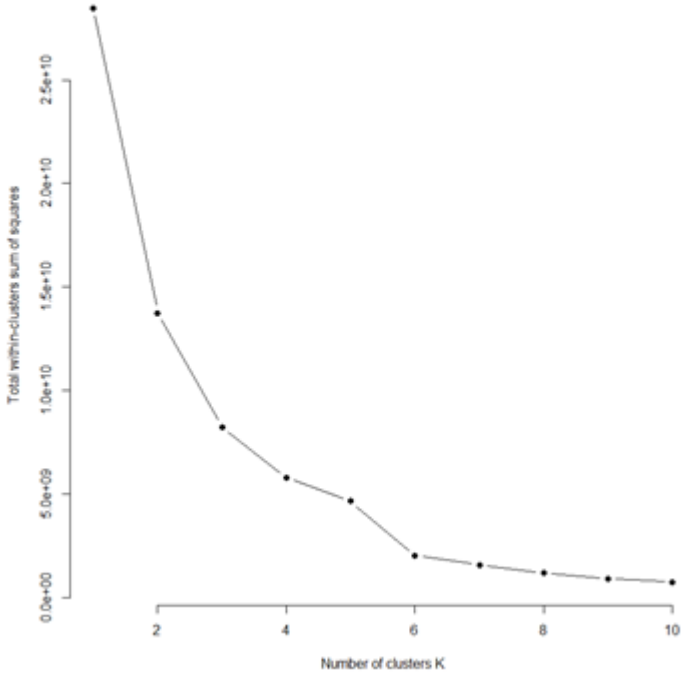

Figure 27. Elbow method for K-means clustering

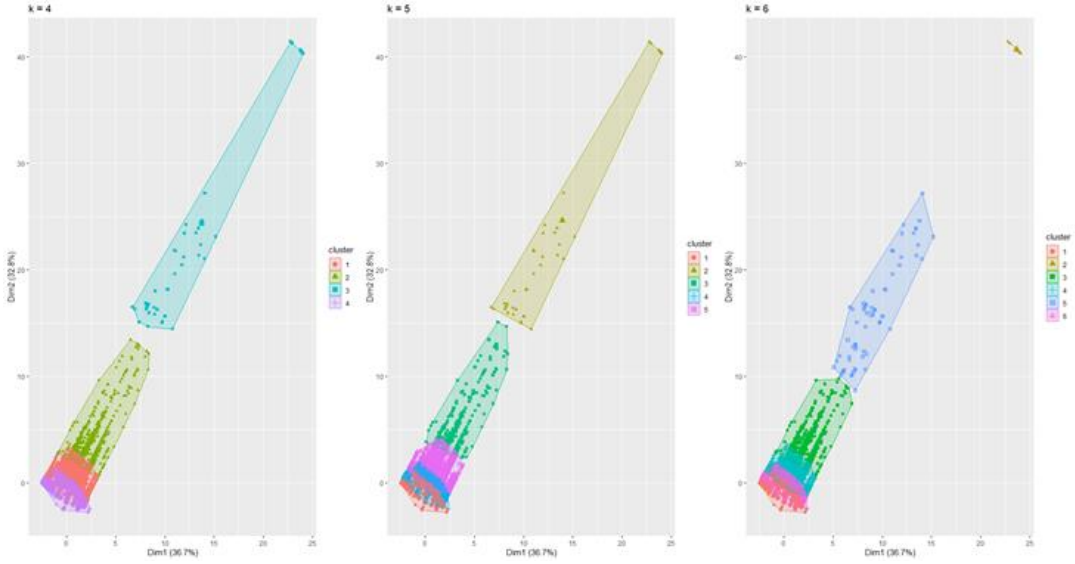

Figure 28. Dimensions scatter plot for K=4 to K=6

It then becomes clearer that K = 5 is the best choice given the most optimal division of houses in each cluster and it is easier to label each cluster based on price as seen in table below. Here the median price is used as to limit the impact of different prices' skewness within the clusters as is visually seen in dimensions scatter plot above.

	Cluster #	*Median Price $*	*Number of Houses*
K = 4	1	148	102,024
	2	610	13,879
	3	2,273	724
	4	12,000	67
K = 5	1	129	87,480
	2	400	24,408
	3	1,000	4,415
	4	3,052	326
	5	12,951	65
K = 6	1	125	84,689
	2	371	25,577
	3	900	5,674
	4	2,286	659
	5	9,999	77
	6	24,999	18

Table 5. Median price and number of houses when clustering using K = 4, 5 and 6

Using the groupings with K = 5, the median price within each cluster has been ordered from lowest to highest and assigned an appropriate label. These labels may be understood as the budget in which either the renter or investor is willing to pay to stay or purchase for a given property as shown below.

Budget	*Median Price $*	*Min Price $*	*Max Price $*
Low	*129*	0	285
Medium	*400*	286	788
High	*1,000*	789	2,420
Wealthy	*3,052*	2421	9,623
Super Wealthy	*12,951*	9,624	25,000

Table 6. Median price of listings when clustering using K = 5

To simplify analysis, 0.7% of properties that were in the Australian Capital Territory are included within the New South Wales Airbnb listings. The 0.15% of houses in Other Territory were reverse geocoded and

included into one of the seven Australian states given the following distribution of homes in Australia. The distribution of number of Airbnb can be seen in a pie chart.

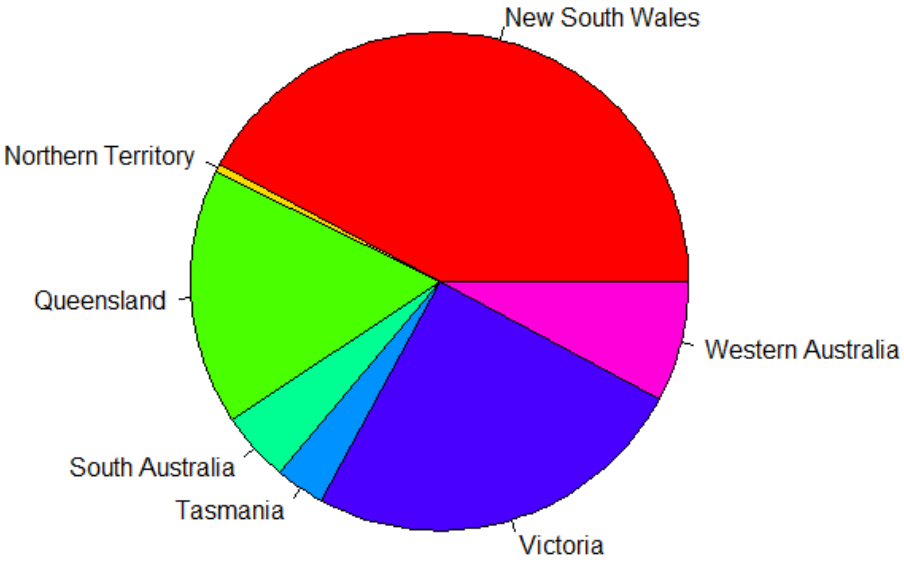

Figure 29 Pie chart of the number of Airbnb listings in each state.

Now it can be identified that the majority of Airbnb homes fit into the low budget bracket range as observed in the percentage stacked bar graph, which are widely scatter around suburban and other suburban areas. It is expected to see that most of the most expensive homes are situated in the biggest cities. High and super wealthy homes are mostly underrepresented on the map. There may be holiday homes for the rich that were built to escape the populace and that preferably while wealthy homes are rare on their own right for being expensive, high budget homes may not also be the best option financially.

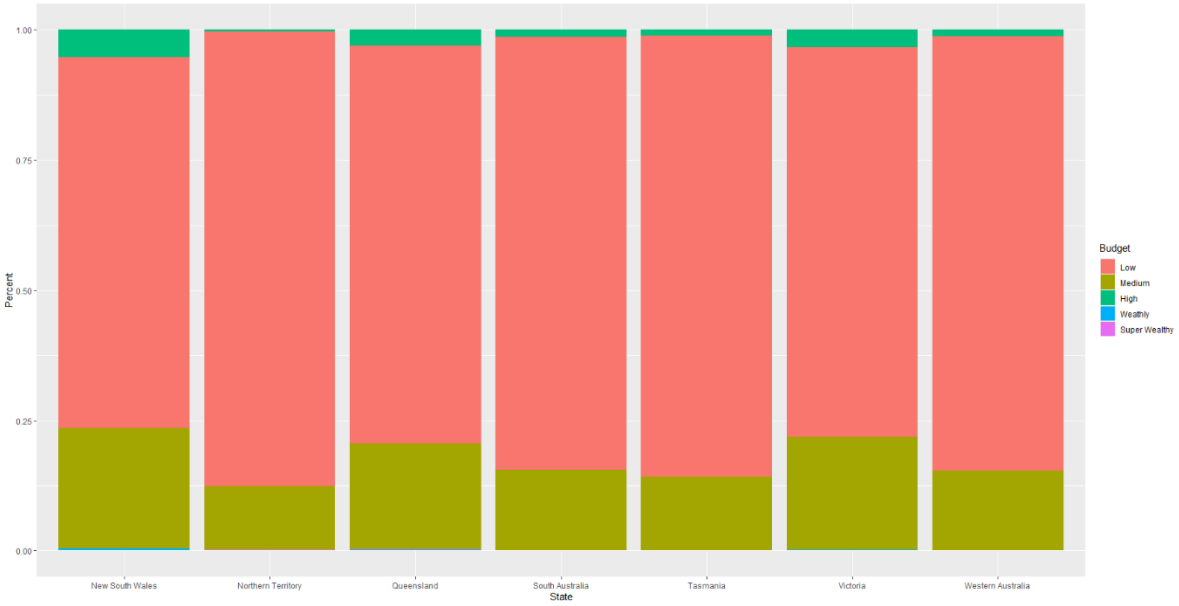

Figure 30. Stacked percentage bar chart of properties in each of the budget ranges per state

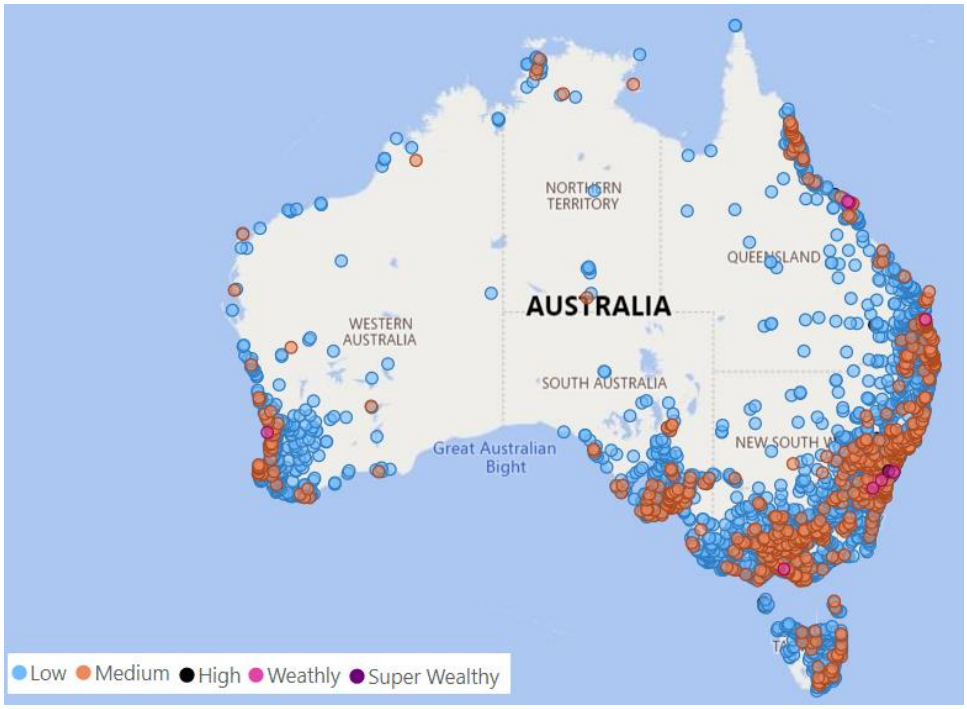

Figure 31. Map of the distribution of Airbnb by budget range in Australia

4.4 Price and Types of room

This portion of the report mainly focuses on the types of room or property and rent per days prices of different properties. The research will help to determine the prices of different properties accordingly, for the further research to predict investors the correct type of property which would give them base for future investments.

For processing of data, the more relevant fields are selected while avoiding others. Therefore, room type, average price, accommodates, bathroom, bedrooms and beds are kept, and others are removed. By looking at the new dataset, a simple box plot for price and room type can be created as shown in the figure below which would help us to show the different nightly renting prices per room type.

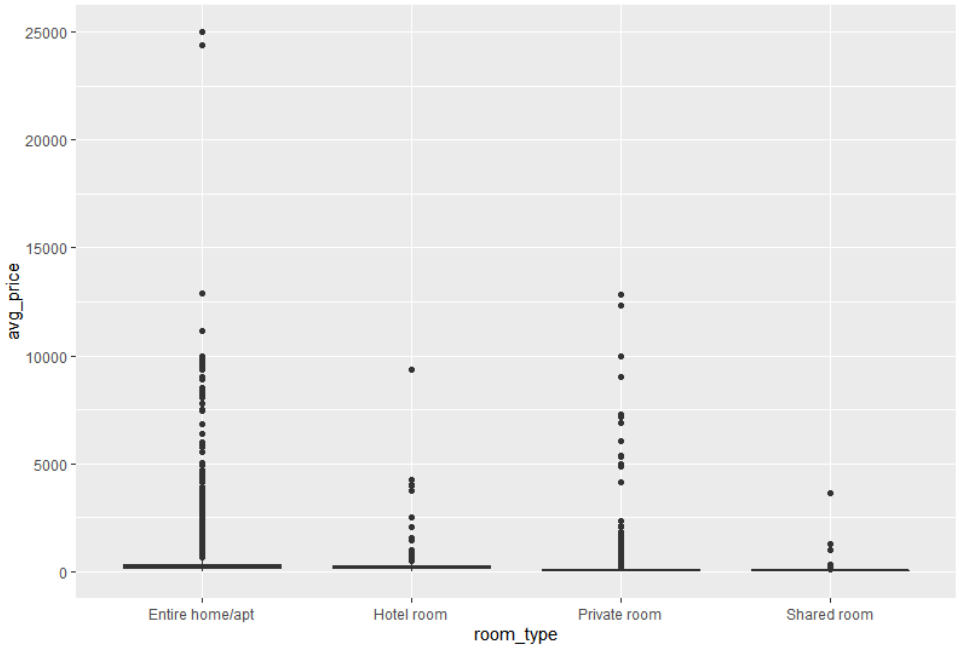

Figure 32. Box plot of prices in all room types

From the above box plot, it is easy to notice that average price of entire home and private room is usually higher than shared rooms. This corresponds to the distribution of listings by room type we have covered earlier in previous sections.

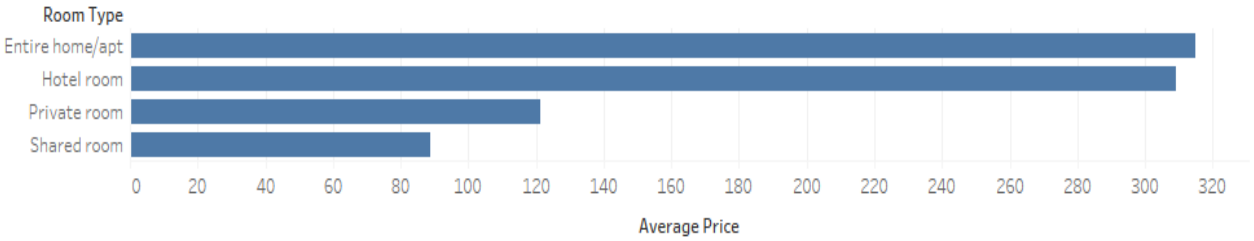

Figure 33. Average price for each Room type

This order of price range versus room type would help investors to understand what kind of property are more profitable and allows them to set the price of their properties they own to a reasonable margin.

For clustering all the variables, k-mean clustering algorithm can used to do different levels of clustering. Determining the k-means that can be done by using the wssplot() function in R as shown in figure below which helps us understand that the k value can be 8, 9 or 10.

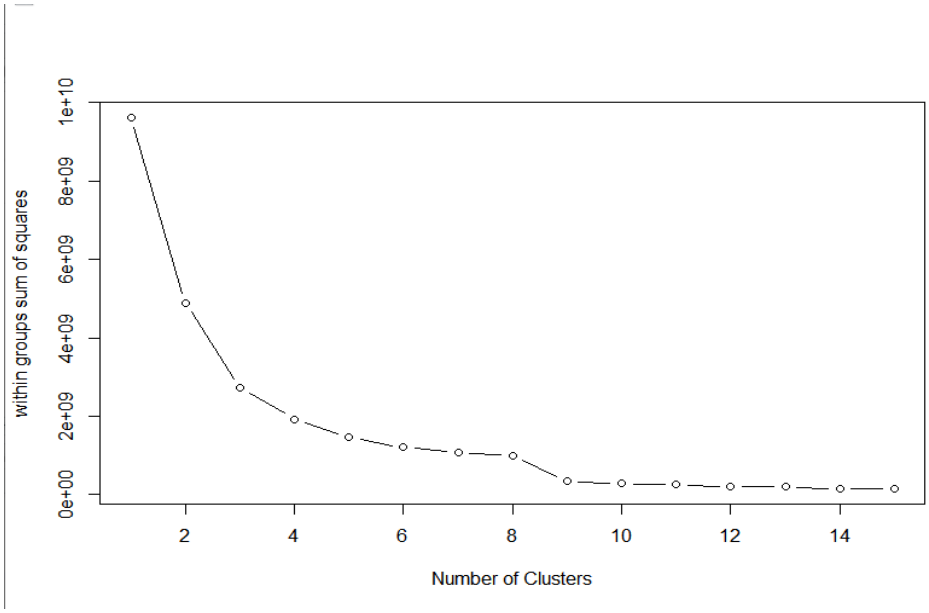

Figure 34. Elbow method for k-mean clustering of listings by room types.

Therefore, we try these three values of k different clusters and plot the results as shown in the figures below.

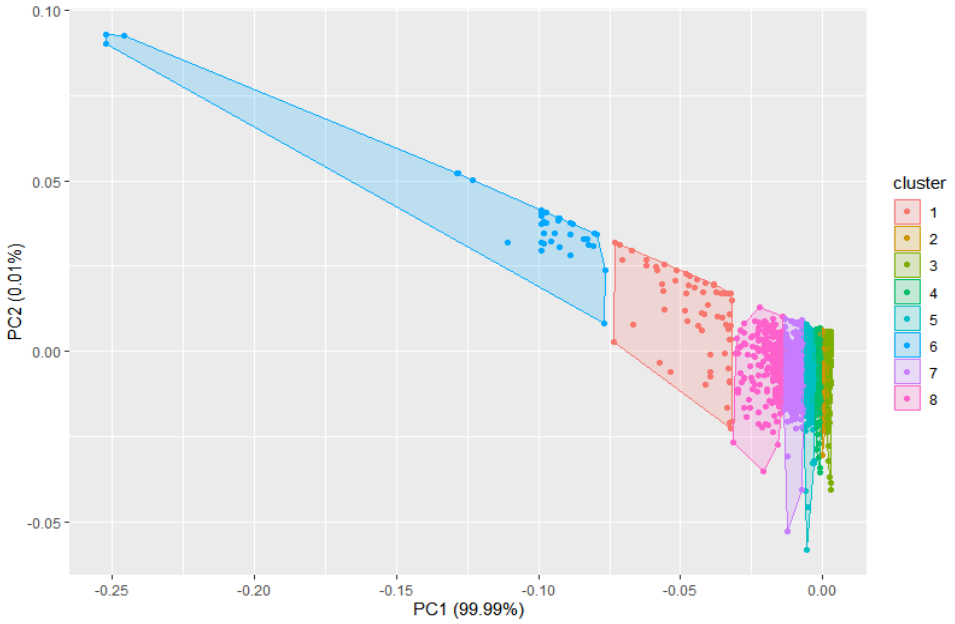

Figure 35. Dimensions scatterplot for k=8

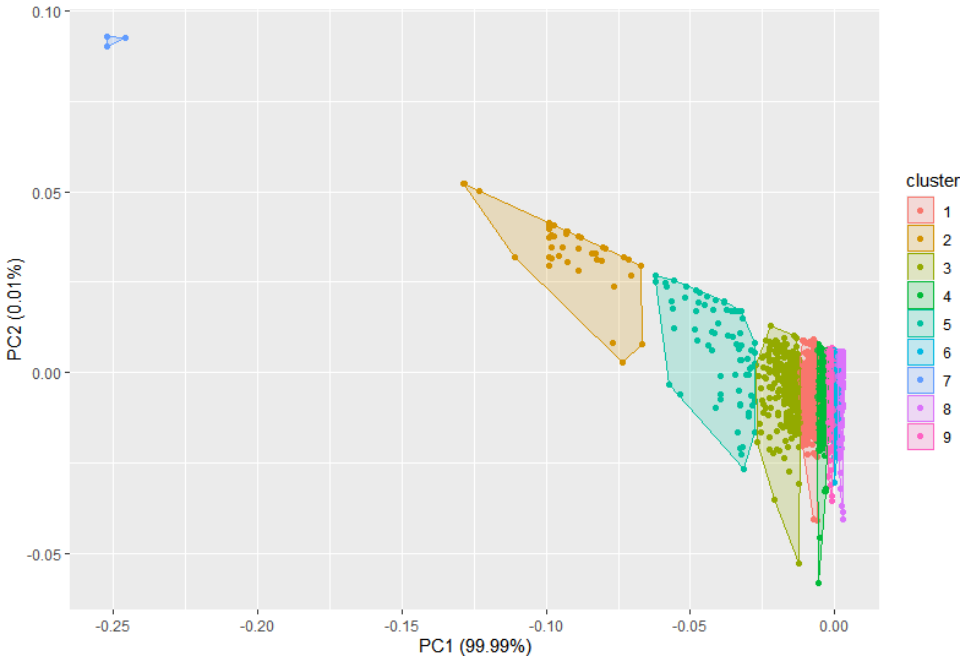

Figure 36. Dimensions scatterplot for k=9

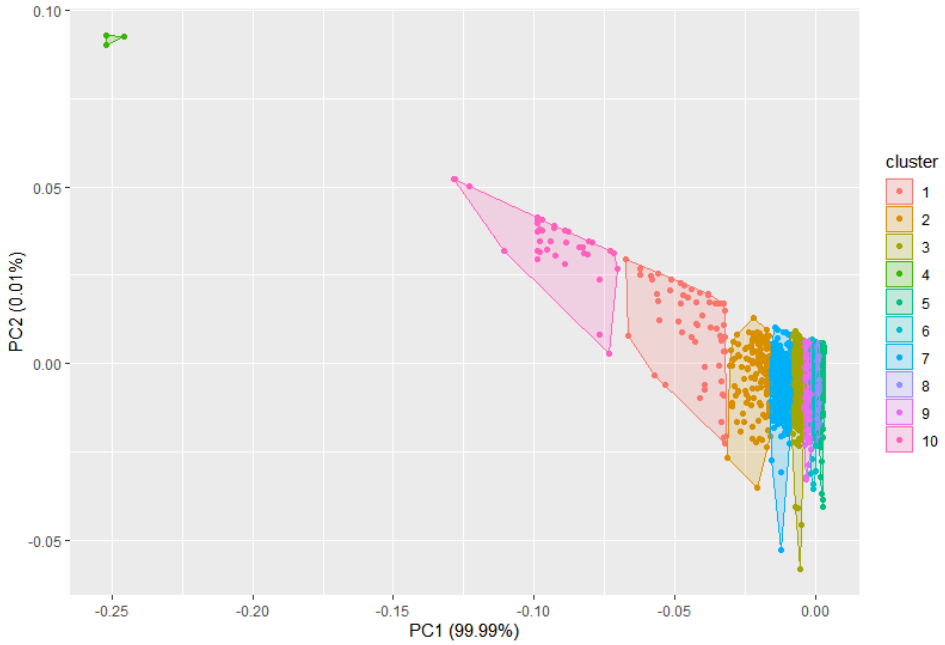

Figure 37. Dimensions scatter plot for k=10

According to elbow method we determine 9 as our value for k. The table below shows different values for variables to be checked.

#	Accommodates	Bathrooms	Bedrooms	Beds	Average price	No. of properties
1	3.3	1.2	1.5	2	110	21441
2	5.6	1	2.7	2.7	24790	3
3	5	2	2.5	2.8	9281	42
4	5.1	1.5	2.3	3.2	235	15515
5	6.8	2	3	4	412	7031
6	8.3	2.4	3.7	5.3	676	3083
7	9.4	2.9	4.2	6	1083	1234
8	10.5	3.6	5	7.3	1855	379
9	8.8	3.7	4.6	7.6	4110	74

Table 7. Mean variables for clustering with k=9.

The values for number of properties are comparatively very low clusters 2, 3 and 9; therefore, these clusters are removed from consideration. The sorted table below shows the different levels of the variables, accommodates, bathrooms, bedrooms, and beds, which can be helpful to standardise the size of property which can be compared with price and number of properties available.

Levels	Accommodates	Bathroom	Bedrooms	Beds	Average price	No. of properties
1	3.3	1.2	1.5	2	110	21441
2	5.1	1.5	2.3	3.2	235	15515
3	6.8	2	3	4	412	7031
4	8.3	2.4	3.7	5.3	676	3083

5	9.4	2.9	4.2	6	1083	1234
6	10.5	3.6	5	7.3	1855	379

Table 8. Sorted table with lower iterations.

Those variables from the previous table can be further condensed into three levels of size of property as shown in the table below.

Size of properties	Average price	Price range	No. of properties
Small	162.48	0 - 329	36956
Medium	492.47	330 - 879	10114
Big	1264.39	880+	1613

Table 9. Price range for 3 levels of property and number of properties.

The findings show different property sizes with range of price and the number of properties available. The findings can help investor understand the margin they can get, or they can set for their property for rent as per trend. While these findings have some obvious answers as big property should have more price and small properties should have less price but settling on a margin understanding the market will be a huge help for investors to make sure the price is reasonable enough so that customers will be interested to spend their nights.

5 Summary of Results

Multiple analyses have been conducted to examine the correlation of Airbnb's properties to other factors, most notably, the price of listings. The price of listings appears to be associated with some main factors such as the location of the listing and types of room the hosts offered publicly on the Airbnb website.

In accordance with the distributions of listings by prices, the data analysis shows that the best k-mean clustering for the data is 6 clusters, associated with bed number, bedrooms number, bathroom details, accommodates details, room type and property type. As the k-mean clustering was based on the combined dataset with location, price and other details of the properties, the results further implied the location of the properties do not affect the price of listings.

The sentiment analysis of reviews and ratings left by guests after their stays shows majority of them are satisfied during their stay; thus, leaving more positive reviews as compared to negative reviews. The most positive words often left by guests including clean, beautiful, and comfortable; implying that these are the qualities that guests favour the most. Some minor negative reviews often tend to be cold, noise and general cleanliness issues which are partly due to the surroundings of the listings and the condition of the property itself. In addition, the ratings of satisfaction by guests during their stay show that the accuracy of the listing's information, the value for money and cleanliness of the lodging contributed the most to the positive ratings; whereas the negative ratings after the stay often correspond to host which provided a lot of listings and are unlikely to personally attend to all of their listings, resulting in those lodgings to be of less well taken care of and of poorer quality.

The data analysis of price and location shows that the price of a typical lodging costs under $300 per night and a lot more properties for rent are situated around larger cities such as Melbourne and Sydney. The

best k-mean clustering for this analysis was 5 clusters, identifying most Airbnb homes fit into the low budget bracket range which are widely scattered around suburban areas. It also shows that the most expensive lodgings are often located in cities with high costs of living, for instance, Canberra and Sydney. Moreover, the analysis of price and types of room gives insight of how room types listed in Airbnb could affect the price of properties. Depending on the size of properties, the analysis shows that small size properties cost an average $162.48 per night, medium size properties cost around $492.47 per night, and big size properties account for $1264.39 on an average per night, with the size of the property being determined mostly by the number of rooms it includes.

6 Conclusion

It is seen that some of the most expensive Airbnb homes are in New South Wales, Queensland, and Victoria. These states also contain the biggest and growing cities with most medium budget and above homes, these results correspond with that found in literature and makes sense as the standard cost of living usually increases closer to the city. Thus, from a business perspective and following what works and is already present, one recommendation for renters and investors is to focus on low to medium homes which range from $0 to $285 and $286 to $788 a night. Regarding which state would be best, it is up to debate as research is missing on the actual profits gained from Airbnb properties within these budgets. But investing in properties in the most expensive states could set up more future investments at and above the high budget range, however, one would often need a higher income for the investment and even to rent there. So, the other four less expensive and relatively less developed states provide less risk for beginner investors and lower income renters. Making it also a suitable choice as it may be easier for investors to strategically position future investments as the regions develop.

As a precaution for Airbnb hosts, sentiment analysis suggests that the opportunists who are looking to expand their number of investments in any area or just starting out should try and keep the number of properties under management to a minimum and low number. Given this is the only factor that had negatively impacted a rating score. This is a reflection on the quality of homes-maintained honesty and interpersonal services that a host can provide to each guest that may diminish with additional responsibilities. While most reviews for each listing are found to be positive, it is important to keep in mind that guest value accurate listings, cleanliness, price, and location of the lodgings. Which reinforce the importance of hosts to not have more properties to take care of then they are comfortable with and adds to the importance of investing in having low and medium budget homes within suitable areas. These homes have been found to be often in the range of small properties fit for a few people to larger Airbnb that can accommodate six or more at and past the medium budget. Hosts that are unsure of how well they are doing in the market can simply check the reviews with low ratings as those two variables are found correlated providing a means to understand areas of importance in lacking and things that needs improvement.

For those interested in staying at an Airbnb with a budget, investigations had shown that only a few factors have largely any dollar impact on the change of price. These factors largely pertain to how much a place can accommodate, number of bath and bedrooms and property type. Which makes sense considering that these factors can easily represent more expensive places with heaps of rooms and better quality, such as a

large apartment and big house. But number of beds, type of room the Airbnb is, and amenities included seem to negatively influence price. Suggesting that these factors are more indicative of lower quality Airbnb that may be more suitable for backpacking accommodation, low social economic accommodation and rooms designed to fit many people into a smaller space. The exact reasons on why these factors have a significant impact on price is still up to speculation but serves to provide a good indicator to those looking for a place to stay, given how many are going to be with them and the quality of livings standards they prefer. For investors and hosts however, this sets up an opportunist situation that they can use to gain a competitive advantage and as consideration in future cost benefit analysis. Focusing on the factors that tend to raise and lower accommodation prices they can try and provide the same quality and services offered at a lower price. This could be at the cost of location that the property is in but if the hosts still maintain good quality rooms, accurate listings. The decrease in listings price at the cost of a less desirable location could be still have a favourable outcome.

7 References

Chiny, M, Bencharef, O, Hadi, M & Chihab, Y 2021, 'A Client-Centric Evaluation System to Evaluate Guest's Satisfaction on Airbnb Using Machine Learning and NLP', *Applied Computational Intelligence and Soft Computing*, vol. 2021, pp. 1–14.

Feinerer, I & Hornik, K 2020, 'tm : Text Mining Package', *R package version 0.7-8*, viewed 17 June 2021, <https://CRAN.R-project.org/package=tm>.

Hu, M & Liu, B 2004, 'Mining and summarizing customer reviews', *Proceedings of the ACMN SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004),* Seattle, Washington, USA.

Jockers, ML 2015, 'Syuzhet: Extract Sentiment and Plot Arcs from Text', viewed 20 June 2021, <https://github.com/mjockers/syuzhet>.

Kalehbasti, P, Nikolenko, L & Rezaei, H 2019, 'Airbnb Price Prediction Using Machine Learning and Sentiment Analysis', *Cornell University*.

Leick, B, Kivedal, B, Eklund, M & Vinogradov, E 2021, 'Exploring the relationship between Airbnb and traditional accommodation for regional variations of tourism markets', Tourism Economics : the Business and Finance of Tourism and Recreation, p. 135481662199017.

McNeil, B 2020, 'Price Prediction in the Sharing Economy: A Case Study with Airbnb data', *University of New Hampshire*.

Rpubs.com 2021, "RPubs - k-means clustering", viewed 08 June 2021, <https://rpubs.com/violetgirl/201598>.

Sun, S, Zhang, S & Wang, X 2021, 'Characteristics and influencing factors of Airbnb spatial distribution in China's rapid urbanization process: A case study of Nanjing', *PloS One*, vol. 16, no. 3, pp. e0248647–e0248647.

Quattrone, G, Greatorex, A, Quercia, D, Capra, L & Musolesi, M 2018, 'Analyzing and predicting the spatial penetration of Airbnb in U.S. cities', *EPJ Data Science*, vol. 7, no. 1, pp. 1–24.