DATA EXPLORATION PROJECT

DATA EXPLORATION AND VISUALISATION

NIKITA BHARGAV

Better Airbnb Experience

1. Introduction:

Airbnb is becoming day by day popular for those who are interested in renting out their spare room, house, shared room as well to those who are looking to rent low price accommodation compared to the high market price of hotels and motels. Airbnb provides a family like environment and facilities while you travel. Now a day's hosts are earning some extra bucks with a choice of renting out the place according to their convenience dates and investors are investing more and more in this renting market. Airbnb is reflecting to be one of the future investment markets for investors of tomorrow.

In this report, I am going to visualise Airbnb dataset of Barwon South West, Victoria because I personally have experience with Airbnb stay and as a data scientist I would like to get insight into how this Airbnb is flourishing and how is it different from any other hospitality services. Also, why Victoria is because this is where I live and know some of the neighbourhoods help me in better analysis of rental landscape in Victoria.

The interactive visualisation and statistical analysis will help me to answers the questions mentioned below:

1. Which properties area will suit best for guest under the best price.
2. Which areas are more popular among guests.
3. How the hosts can improve their business from the feedback and reviews from customers.
4. Which locations and which part of the year is the best investment opportunity for investors.
5. Which type of properties and room types are famous in a different neighbourhood.
6. Sentimental Analysis of guests' expectations and satisfaction.
7. How popular is Airbnb throughout the years from the year it started in VIC.

To answer the above questions, this assignment used python and R for data pre-processing, combination of R and tableau graphs and charts for data wrangling and data validation, for visualisation. Tableau and R visualisation also helped in validating spatial data by plotting them on maps.

2. Data wrangling:

2.1. Data collection:

All the datasets have been collected from Inside Airbnb (http://insideairbnb.com/get-the-data.html) datasets for Barwon South West, Victoria.

Consisting of 3 different data sets:

1. Detail listings: This dataset comprises of 96 attributes for each rented property having, 4922 rows with unique listing. The main attributes used from this dataset are
   Continuous- latitude, longitude, price ($14-$10,000 per night), review_scores_location, number_of_reviews.
   Categorical- zipcode, neighbourhood_cleansed, property_type, room_type, bedrooms.
   Discrete – id, listing_id.
   Textual- amenities

2. Detailed Reviews: This dataset comprises of 6 attributes for reviews given by guests for the properties rented from 2012-2019 with more than 1 Million rows. The main attributes used are
   Textual- comment
   Datetime- date
   Discrete- listing_id

3. Detailed Calendar: This dataset comprises of all the bookings are done by guest till next year from Feb 2019- Feb 2020 with 1796530 rows. The main attributes used are
   Categorical- available
   Datetime- date
   Discrete- listing_id
   Continuous- price

2.2. Multiple data types: The datasets are in (csv, text, xlsx) forms and are pre-processed using python to get the desired dataset that can be visualised using R and tableau.

2.3. Merging Datasets: The listing dataset is merged with calendar dataset on Listing id column to get booking data for new calendar list (1529349 rows) comprises of those listing ids that are present in listing dataset. Python pandas are used to reformat dataset.

2.4. Splitting Datasets: The reviews dataset is used to create a subset of the comments for 50k random reviews given by a guest to create a word cloud. In Listing dataset amenities are column is processed to find individual amenities.

2.5. Missing data analysis: There were 428 state, 5 cities, 5 zip code, 5 market, 1 bedroom, 1 bathroom, 640 review_scores_location records were missing. The zip code, city and states were imputed using latitude and longitude and the remaining null rows were dropped as data were missing completely at random.

Visdat and naniar are plotted to find missingness in the data. Accordingly, imputation and deletion of data are performed. (Tierney, 2019)

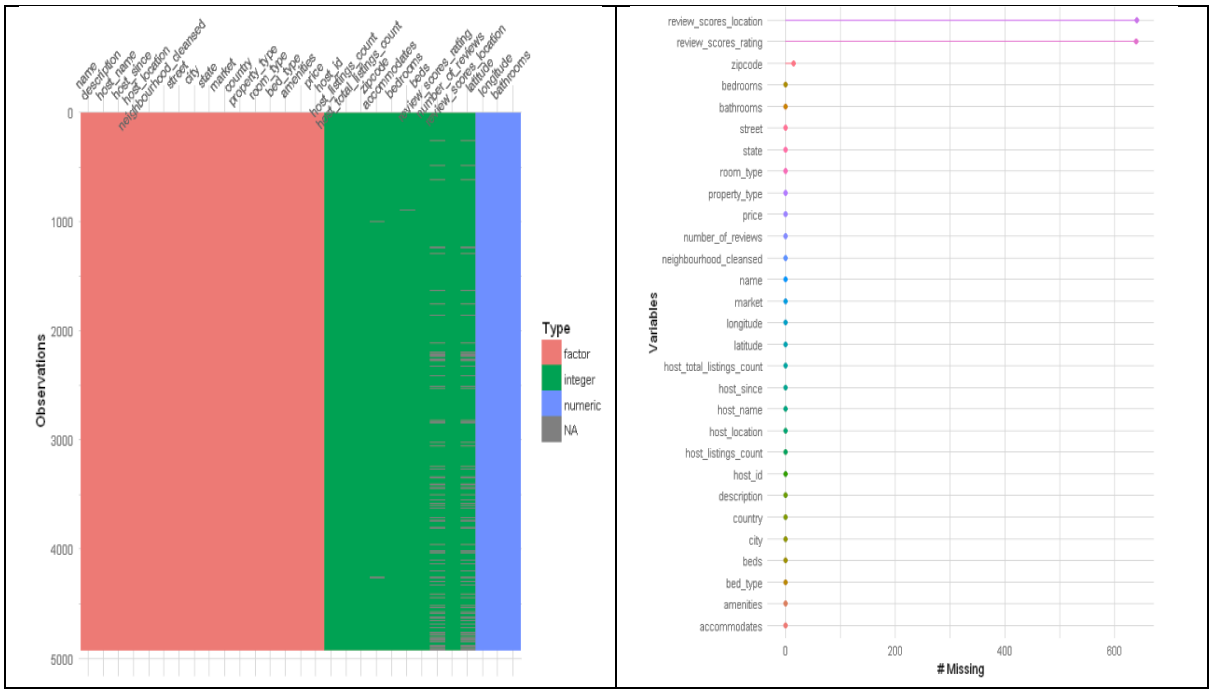


Figure 1 Information about data types and missing values in the dataset

3. Data checking:

For data validation tableau and python are used.

Listing Dataset:

While plotting dataset columns in tableau, the values in country column shows (Morocco) as an outlier with respect to the analysing country Australia data, so this value has been dropped.

state column is formatted to (VIC) to remove inconsistency,

plotting the zip code on the map shows zip code 3000, 3555 are wrongly mapped to latitude and longitude thus validating with city, neighbourhood, latitude and longitudes are fixed.

Amenities column is pre-processed using python NumPy and pandas to get the list of different amenities for all the properties. The amenities column consists of a list of different amenities with inconsistent string formats. The list processed by removing [{}] and “” and special characters from the amenities column. Only the meaningful amenities with consistent word names were considered, and a new dataset has been created.

Price column is converted to an integer from string values by removing ‘$’ and ‘,’ and the prices greater than $800 has been removed so that our data analysis not get affected by these outliers. Number of bedrooms more than 5 have been removed as outliers to get a general idea of normal rented size places.

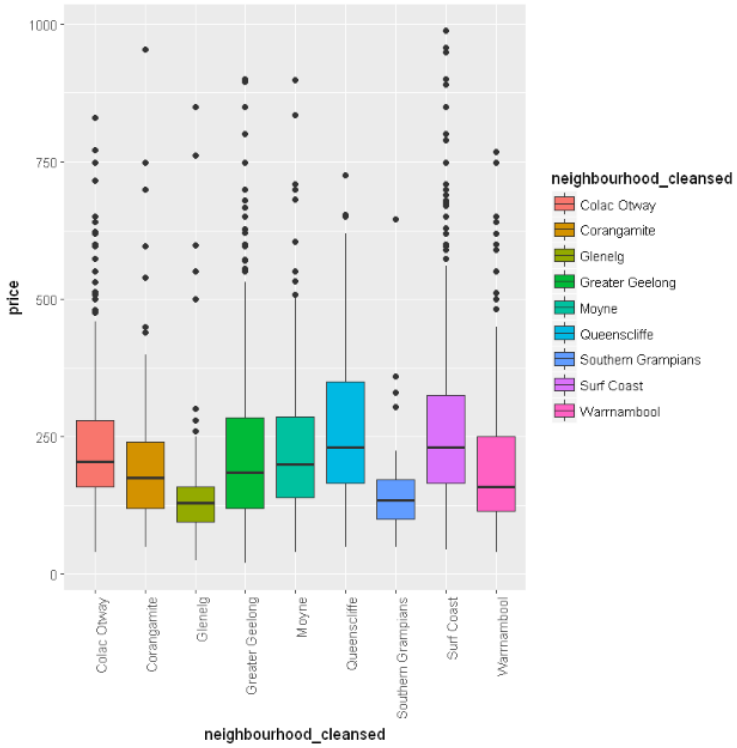


Figure 2 Outliers in Price values

For Outliers tableau and R are used.

Calendar Dataset: The Price is converted to integer, values greater than $800 has been removed. The format of dates is fixed to DD/MM/YYYY.

Review Dataset: The comment column consists of reviews in different languages other than English like Chinese and French. Thus, I did a high-level text filtering using regex in python to get only those reviews which are in English. The size of reviews is more than 1 Million, thus I pre-process some random 50,000 reviews to get a good idea of expectations by guests. All the special characters, meaningless words have been removed. Again, the processed text is converted to a list of words to be used as textual data for creating a word cloud.

Listing and calendar dataset after wrangling and correction is converted to tabular form, Amenities data is stored as Pandas data frame, Review data is also converted to tabular form. The comments after pre-processing are stored as text data to directly use for the word cloud.

All the dates in reviews, listing and calendar are converted to same format DD/MM/YYYY to have consistency throughout the datasets.

4. Data Exploration:

In this assessment, I am going to observe the cleaned and wrangled data for finding the trend and patterns using interactive visualisation and statistical plots and will conclude my findings.

4.1. Total listing Distribution in Barwon South West with respect to different room type and price.

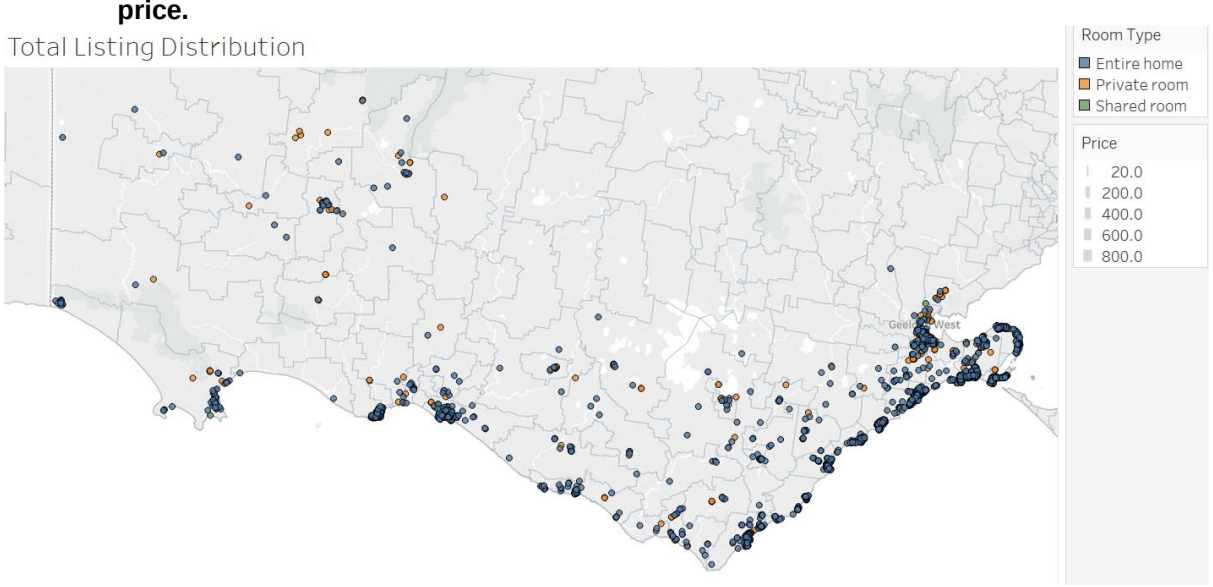


Figure 3 Distribution of entire house, a single and shared room in Barwon South West

The bubbles in the figure shows Room type where blue bubbles are Entire House/Property, green bubbles are shared room and orange bubbles are single room. The above map shows that the availability of the Entire House/Property is highest compared to the single or shared room. Map also shows that the shared rooms are least preferred by hosts, cloudiness of bubbles is more in Geelong West that shows a greater number of properties are listed in these areas. The map also shows that the places near beaches and shores are most popular among hosts as it attracts more guests. The size of the bubbles represents the price of properties and show comparatively whole properties are expensive than shared or single room per night.

4.2. The property area that will suit best for guest under the best price.

Property Price distribution throughout Barwon South:

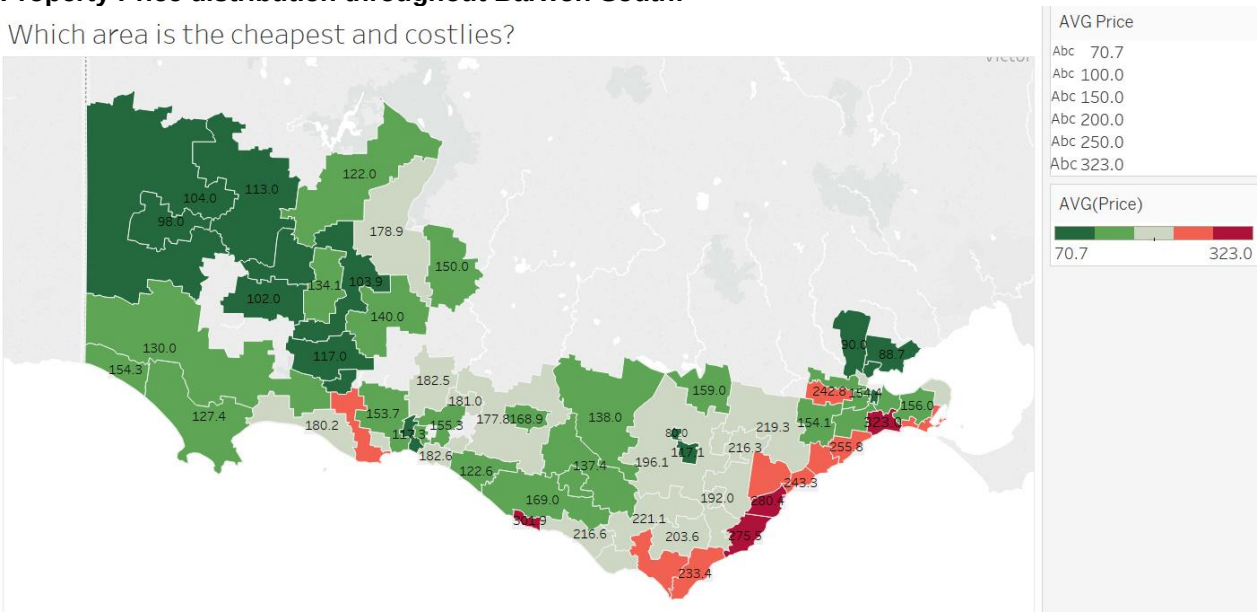


Figure 4 Price distribution among different regions of Barwon South West.

The map in figure 4 shows that the region in red shades is Barwon Heads, Peterborough, Lorne, Wongarra are the costliest properties with avg price more than $240 per night. On the other hand, the region in green shade regions are Corio, Norlane, Lara, Corooke and Anakie are the cheapest properties with prices less than $90 per night. Thus, keeping in mind their budget, guest can book properties in these different regions. (Gupta, n.d.)

Preferred Location by guests throughout Barwon South:

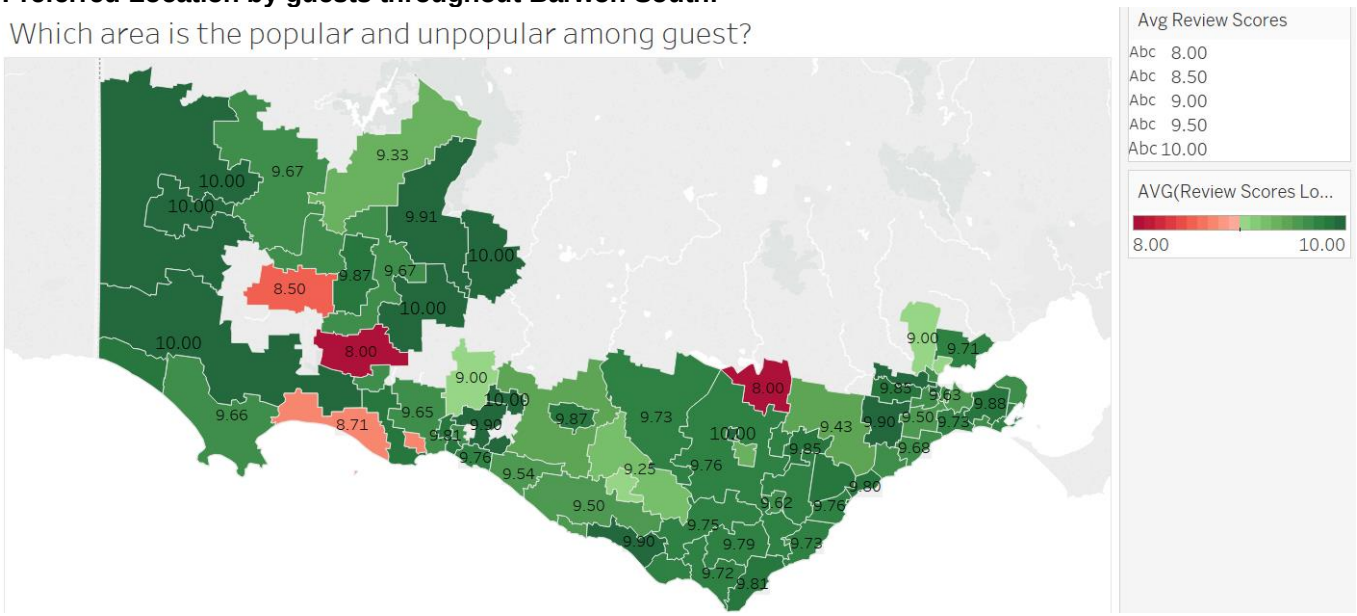
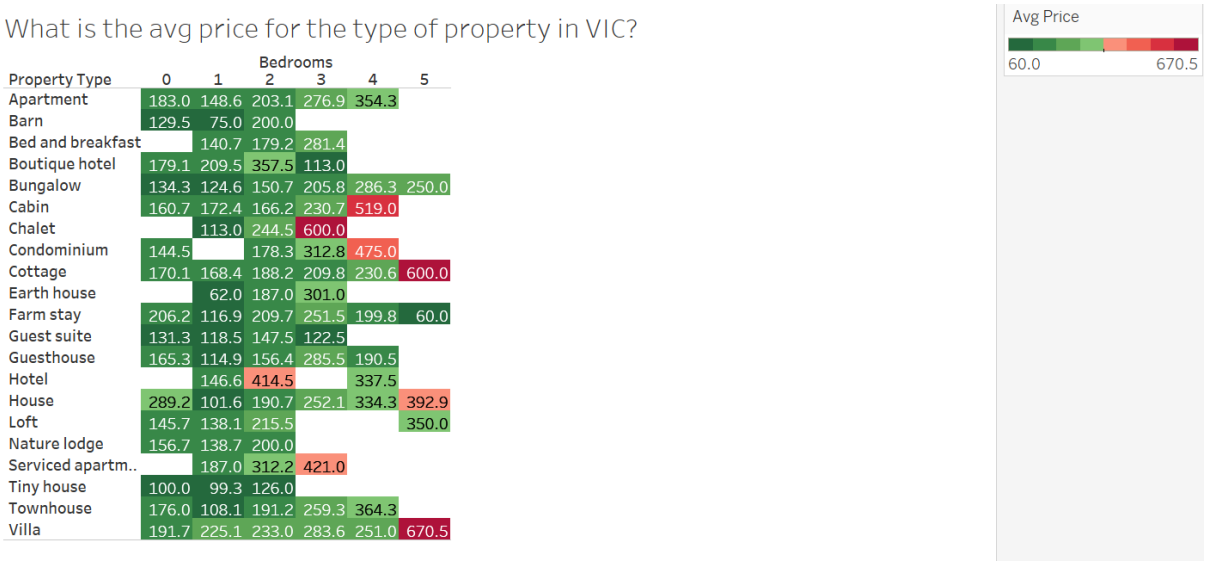


Figure 5 Popularity of different regions in Barwon South West.

The map above shows that regions like Carapook, Greenwald and Barongarook have dark green regions with a rating equal to 10 given by guest with respect to the location thus all the properties with a high rating are considered as most popular, while the red regions areas like Beeac, MacArthur, Branxholme, Narrawong are the least rated with ratings less than 8.8 thus are least popular among guests.

Price variation for different Property types with number of bedrooms:



	Bedrooms					
Property Type	0	1	2	3	4	5
Apartment	183.0	148.6	203.1	276.9	354.3	
Barn	129.5	75.0	200.0			
Bed and breakfast		140.7	179.2	281.4		
Boutique hotel	179.1	209.5	357.5	113.0		
Bungalow	134.3	124.6	150.7	205.8	286.3	250.0
Cabin	160.7	172.4	166.2	230.7	519.0	
Chalet		113.0	244.5	600.0		
Condominium	144.5		178.3	312.8	475.0	
Cottage	170.1	168.4	188.2	209.8	230.6	600.0
Earth house		62.0	187.0	301.0		
Farm stay	206.2	116.9	209.7	251.5	199.8	60.0
Guest suite	131.3	118.5	147.5	122.5		
Guesthouse	165.3	114.9	156.4	285.5	190.5	
Hotel		146.6	414.5		337.5	
House	289.2	101.6	190.7	252.1	334.3	392.9
Loft	145.7	138.1	215.5			350.0
Nature lodge	156.7	138.7	200.0			
Serviced apartm..		187.0	312.2	421.0		
Tiny house	100.0	99.3	126.0			
Townhouse	176.0	108.1	191.2	259.3	364.3	
Villa	191.7	225.1	233.0	283.6	251.0	670.5

Figure 6 Avg prices for different property type with 0-5 bedrooms

Figure 6 shows that houses are comparatively cheaper than villas, service apartments, hotels, condominium, bungalows and apartments with the same number of bedrooms. That means the guest should prefer the house and tiny house as compared to other types of properties. The heat plot also shows that red coloured hotels, chalet, cabins and villas are comparatively costliest among other property type.

4.3. Analysis to improve business from the feedback and reviews from customers.

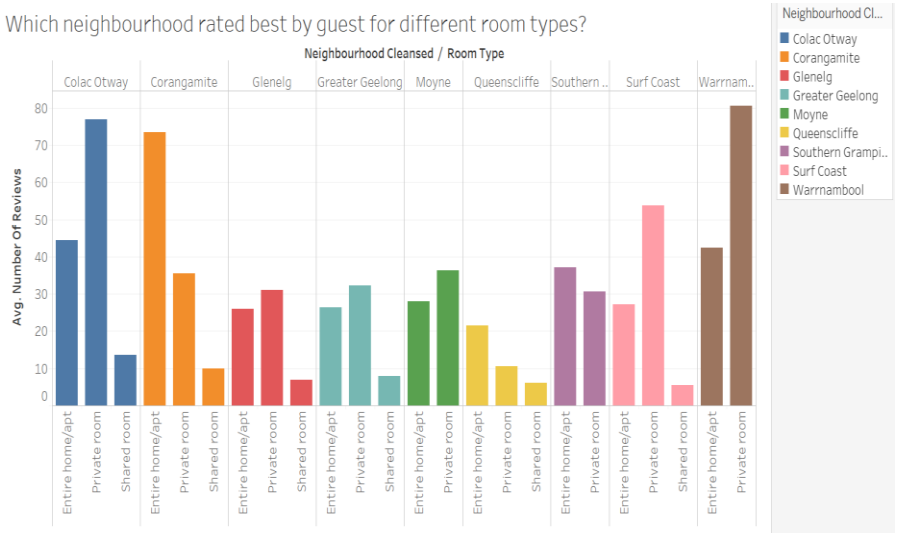


Figure 7 Number of reviews for the different neighbourhood for different room types

The bar graph shows that almost all the neighbourhood places have the highest reviews for private room type except Corangamite and Queenscliffe which means guest prefers a single room rather than shared or private homes. Thus, hosts should rent the single room as compared to the shared room or house which will increase their chances of renting the place. The graph also shows that the number of reviews is very less for shared room type thus less preferred by guests.

Airbnb Talk of the town:

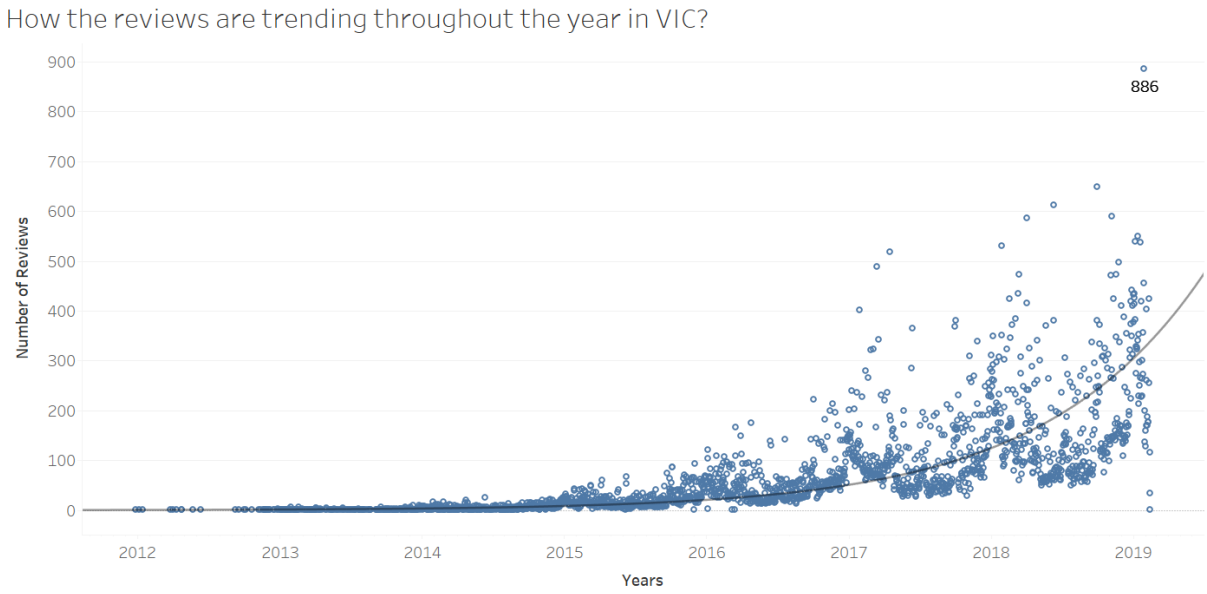


Figure 8 Reviews Trend Yearly

The above bubble chart shows that day by day Airbnb is becoming popular among guests and hosts as a greater number of hosts/guests are giving feedbacks and while comparing from year the 2012 the

trend has increased drastically till 2017 while there was a slight down slope in 2017 then again, the popularity has increased.

4.4. Feedback Sentimental Analysis:

Figure 9 Word Count of reviews and comments by reviewers for each listing

This word cloud is created from the comments given by guests for each listing they rented which shows the interest and expectation of the guest for the property they have rented and for how much extent these expectations are met. The word cloud above provides guest mindset that among different property type, house is preferred. The location word shows that the surrounding where guest stay is also important like beach, the ocean which are high frequency words. Different amenities like shops, kitchen, views from the property are also given significance. Which shows that guests want kitchens so that they can cook for themselves and require nearby shop that is easily accessed. Guest prefer clean, comfortable, spacious, nice, wonderful, loved surroundings. It also shows the hosts also plays a key role like host rating and friendly or helping nature etc. Guests prefer breakfast included in their rental offer. These are some to the findings from word cloud.

The most likely amenities provided by hosts:

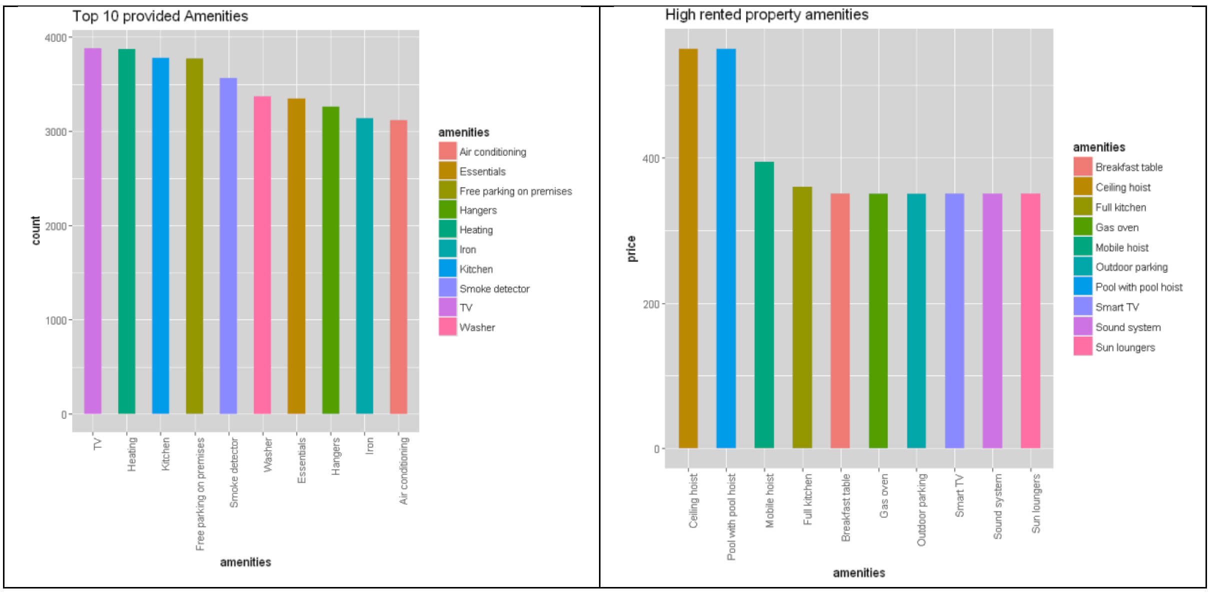


Figure 10 Top 10 price property amenities

The bar graph in the left of the figure 10 shows the decreasing order of the amenities provided by hosts regardless of Room type and its location. TV, Heating, free parking, a smoke detector is some of the most popular amenities the Airbnb properties have.

The bar graph in the right shows the high rented properties in decreasing trend have ceiling hoist, pool hoist, mobile hoist, full kitchen, breakfast table, gas oven, parking, TV for the guests. These amenities visualised above are almost expected once, as these are the general amenities a house should have.

It can be inferred that by providing these top amenities the chances of renting a property can be highly increased.

4.5. The type of properties and the season of the year would be best investment opportunity for investors and owners.

The Busiest duration of the Year from 2019-2020:

Figure 10 The trend of available listing throughout the year from 2019-20

The above trend shows that the busiest month was from Feb to Mar 2019 and the next busiest season will be Aug to 2019 to Jan 2020. While the busiest month will be summer Dec 2019 to Jan 2020 with the expectation of highest bookings in Jan 2020 which is around 74k. The flag false of newcalendar dataset has been used which shows the availability of listing for this time. Thus, investing in renting property from Oct to Jan is expected to be a better decision according to the trend. (Le, 2018)

The Yearly and Weekly price trend for the listed properties.

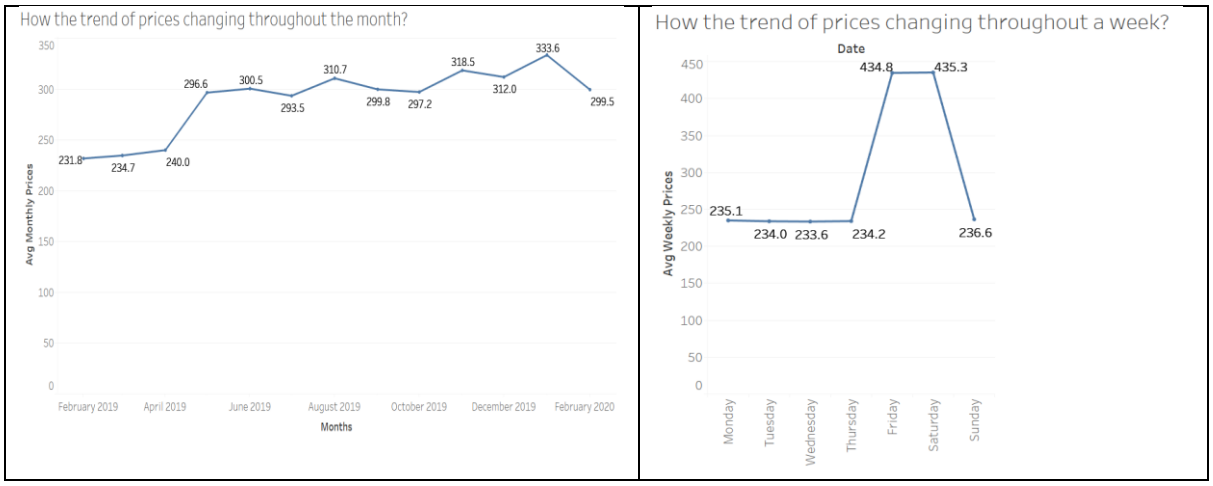


Figure 11 The trend of prices yearly and weekly from 2019-20.

The Yearly trend shows that from April to Mar 2019 the prices go up and keep on increasing slightly till Aug 2019 and will again rise from Oct 2019 to Jan 2020 as summer vacations starts. The weekly trend shows, that prices rise on Friday and Saturday as compared to any other weekdays. Which was already expected. As of Oct- Jan is the busiest season thus the prices are expected to rise, and the trend proves it.

Popularity of property type in different neighbourhoods.

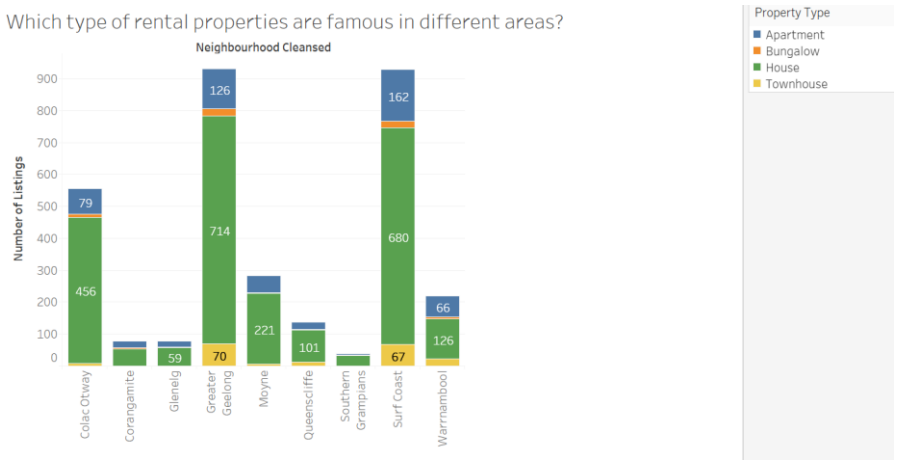


Figure 12 The popularity of the type of property in different neighbourhood

The stacked graph shows that for all the neighbourhoods, house property type is preferred by almost all the guests, thus more numbers of listings are of house property type because guests need a friendly and homely environment while travelling. Bungalows are the less preferred property type in all the neighbourhood with listings highly low. Bungalows are rarely preferred by guests.

5. Conclusion:

The data wrangling and data validation on listing and calendar datasets to create subset of datasets. The review file is pre-processed by text processing to provide the well formatted and cleaned data in the desired format required to be used as data source for R and Tableau. The visualisation performed in assessment provides insights of the data using the trend and plots.

These trends helped me to provide all the answers to the questions I am looking for.
4.1. Provides the overall distribution of different room type across the Barwon South West signify that single room type topped the list among different all the areas around Geelong west.
4.2 Price Map shows that Barwon Heads, Peterborough, Lorne are the costliest place while Corio, Norlane and Lara are highly affordable.
The Rating Map shows that Carapook and Greenwald are among the most popular rented properties among guests while Beeac and MacArthur among the least popular.
Heatmap helps to visualise that houses with 0-5 bedrooms are most affordable among other property types.
Section 4.2 gives answer to best property area for guest under the best prices.
4.3. Shows that the single room is mostly preferred by guests in the majority of neighbourhood.
The yearly trends show the increasing popularity of Airbnb from past 8 years.
4.4. Textual Analysis helped to analyse the expectations of guests.
Textual Analysis of amenities like TV, parking etc to get idea of highly available amenities for guests.
4.5. Helped the investors who are planning to invest in different properties of Airbnb. The yearly listing trends shows that Oct-Jan is the busiest season. While yearly price trends show that April to Mar and Oct-Jan are the costliest and Weekly price trend shows that Sat and Sun are the costliest.
The stacked graph analysed shows that investing in the house is always a good deal.
These help the investors to invest in the appropriate season to gain more profits.

6. Reflection:

By performing this Data Exploration assignment, I learned that different types of graphs have different purposes and should be used accordingly like trend and line graphs are for time series data. Choropleth and Maps are for representing geographical data. Bar graphs and Stacked Bar chart are used for comparison between different entities. I also learned that usage of proper colour combinations in graphs helps to understand the insight very easily. The textual analysis by using (Word Cloud) Network relationship diagram. While visualising data, it's important to understand the data thoroughly and the question that you are going to answer so it becomes easy to get an insight while finding the answer. In my dataset, there was a huge number of columns which can also be visualized apart from the some which I already explored, and which could have helped for better understanding of Airbnb data. With this project, I learned how to use many new packages of python like regex, Visdat and naniar, Some R packages for word cloud and ggplot2. I learned to create visualisations graphs using tableau. I also learned that tableau can also be used for data wrangling and data cleaning.

7. References:

Gupta, S. (n.d.). *Towards Data Science*. Retrieved from Medium: https://towardsdatascience.com/airbnb-rental-listings-dataset-mining-f972ed08ddec

Le, S. (2018, Nov 30). *Towards Data Science*. Retrieved from Medium: https://towardsdatascience.com/exploring-machine-learning-for-airbnb-listings-in-toronto-efdbdeba2644

Tierney, N. (2019, Feb 15). *Getting Started with naniar*. Retrieved from Cran: https://cran.r-project.org/web/packages/naniar/vignettes/getting-started-w-naniar.html