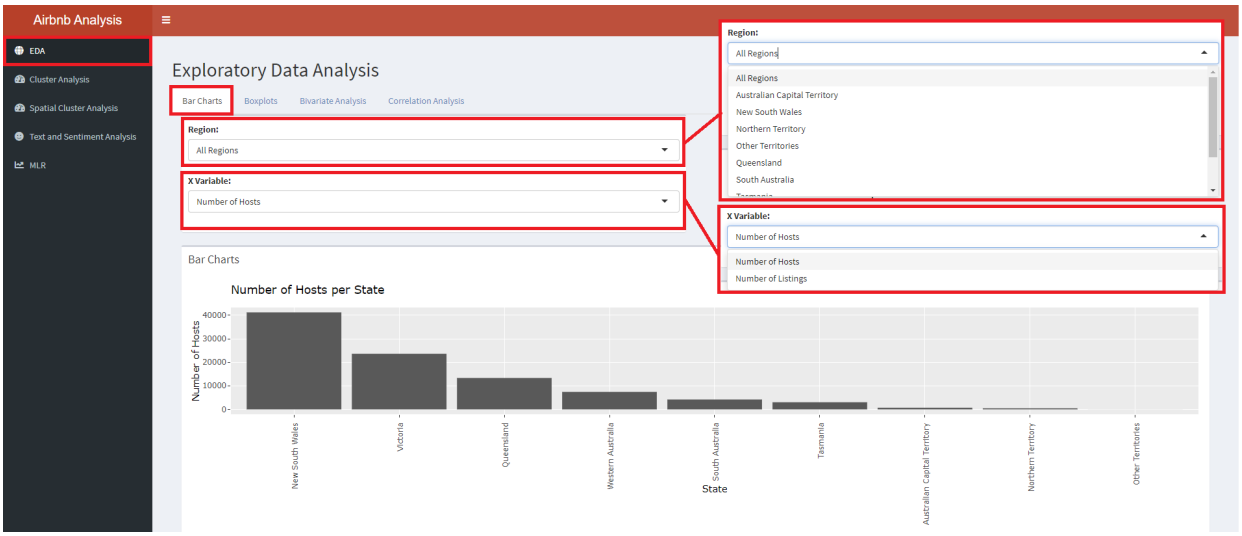
User Guide - Understanding Airbnb listings in Australia

true true true

Exploratory Data Analysis (EDA)

Exploratory Data Analysis allows you to perform initial investigation on the data, so that you may be able to discover patterns, and explore the different variables in the data set. It will allow us to formulate hypothesis and explore different statistics models that could be developed after.

Bar Charts

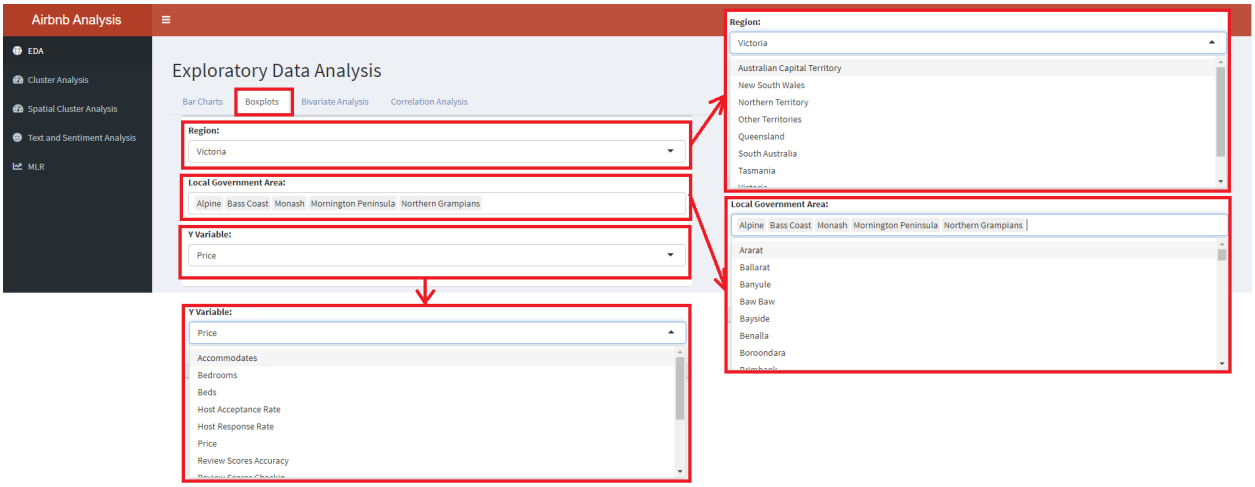


- Click on *"Region:"* to explore the different states in Australia
- Click on *"X Variable:"* to explore *"Number of Hosts"* and *"Number of Listings"* in various different States.

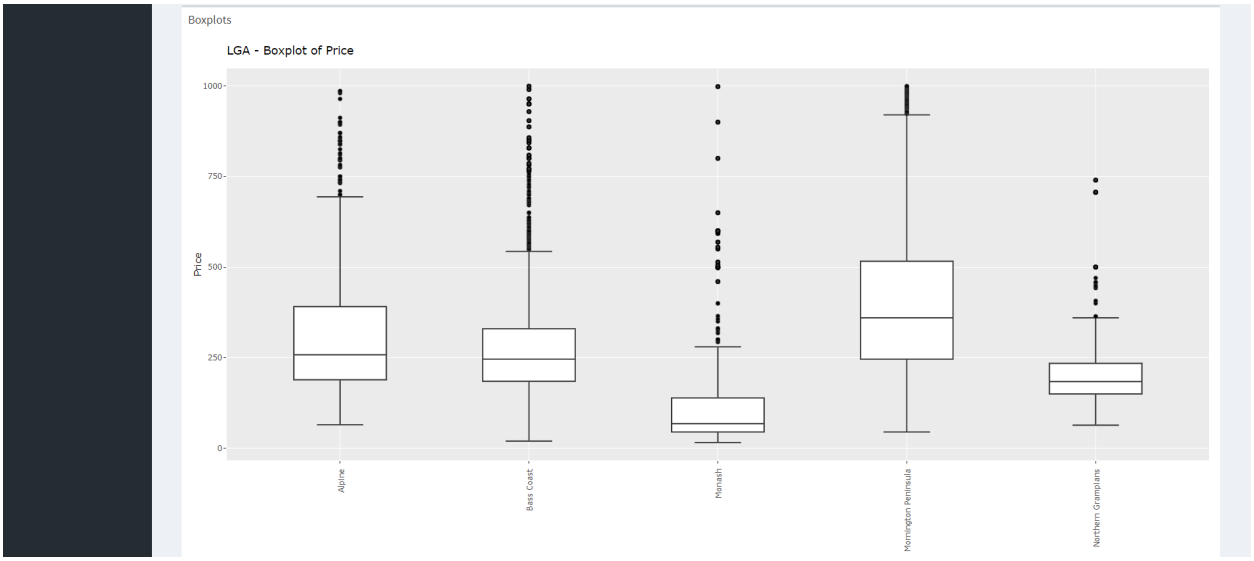
The top barchart showcases the *"X Variable"* per State.

The bottom barchart showcases the *"X Variable"* per top 10 Local Government Area per State.

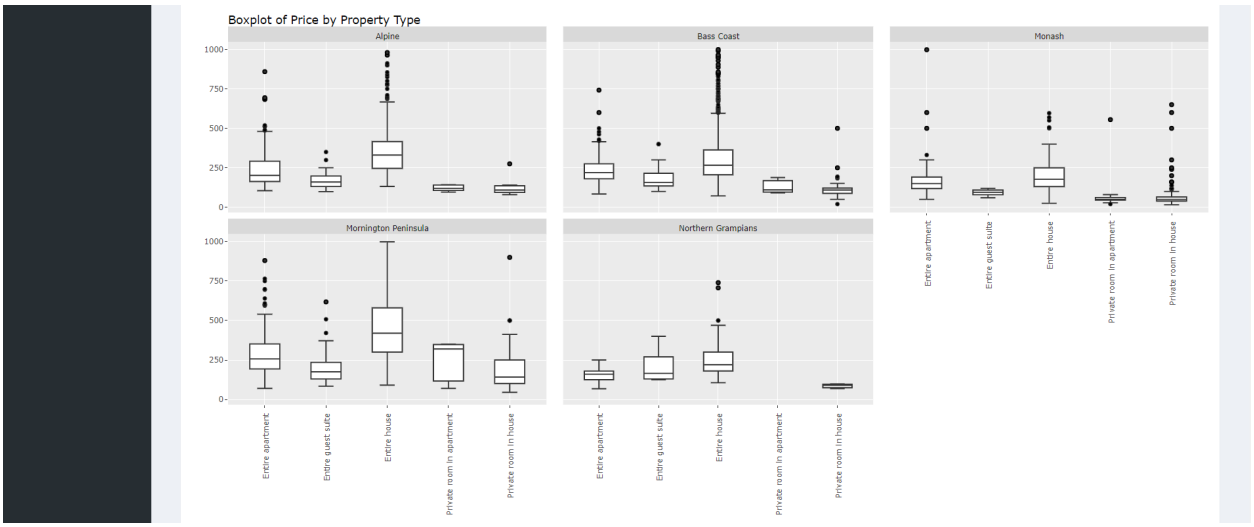
Boxplots



- Click on *"Region:"* to explore the different states in Australia.
- Click on *"Local Government Area"* to explore the different cities in the State. You are able to choose multiple cities to view at a single time through clicking the chosen city in the drop down bar.
- Click on *"Y Variable:"* to view the boxplots of different variables such as *Price, Review Scores Ratings, Bedrooms, Beds* etc.

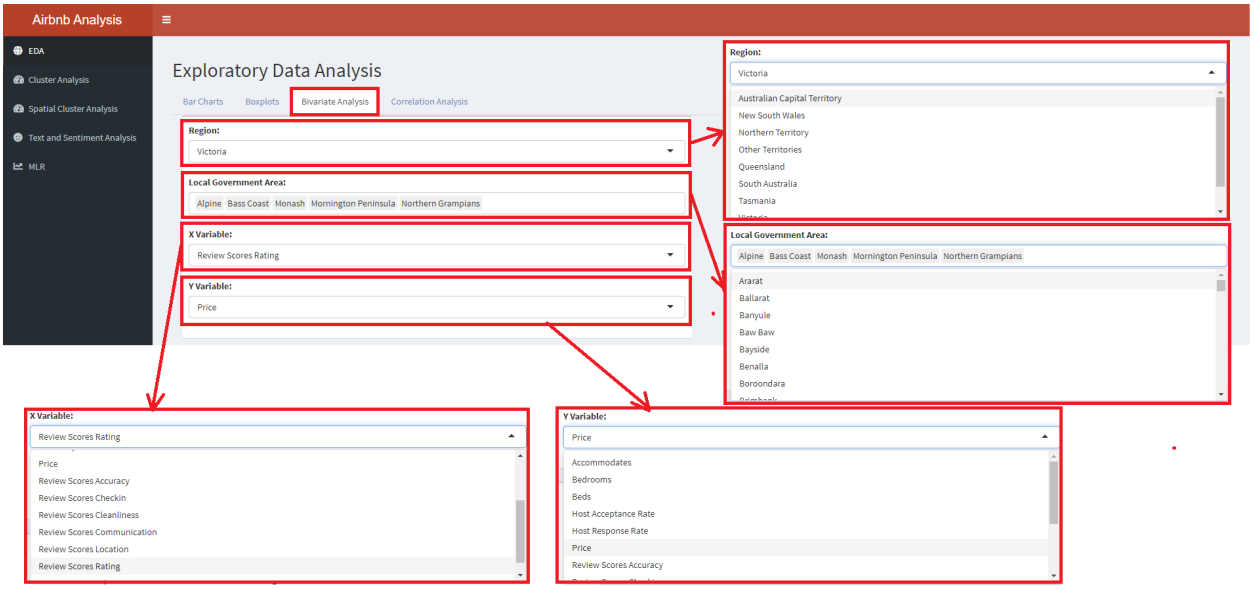


The top boxplot showcases Boxplots of *"Y Variable"* against the Local Government Area.

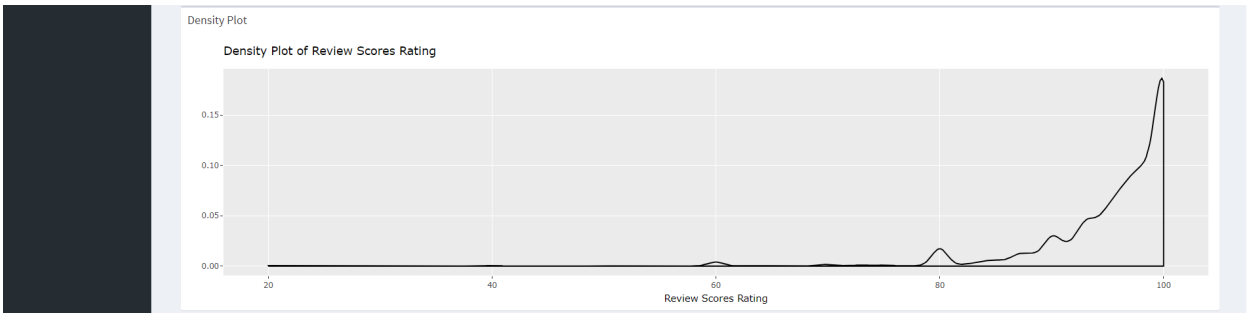


The bottom boxplot showcases *"Price"* vs *"Property Type"* with a Facet Wrap of Local Government Area.

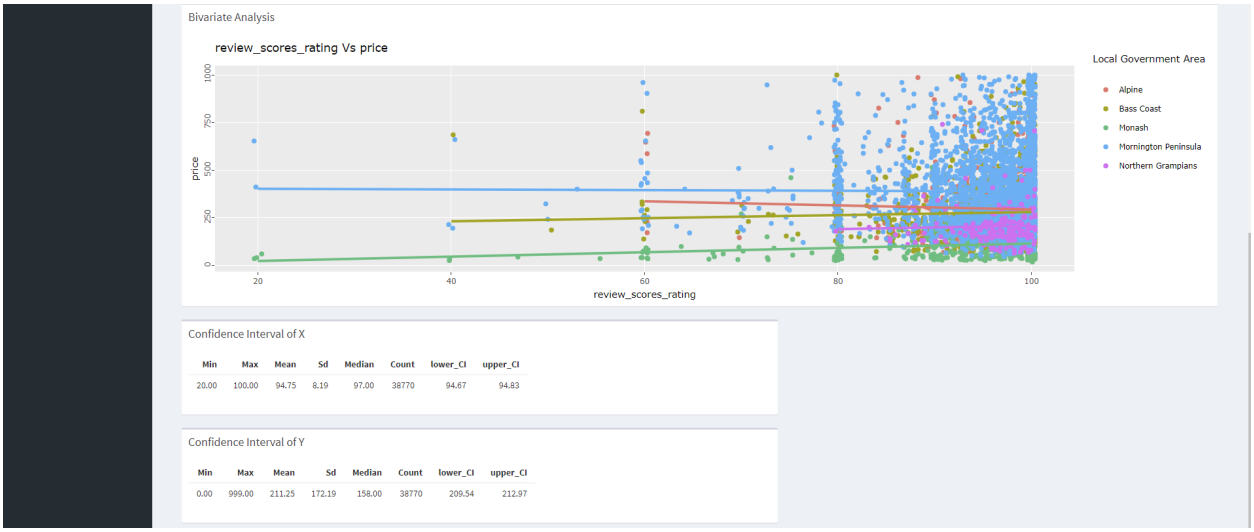
Bivariate Analysis



- Click on *"Region:"* to explore the different states in Australia.
- Click on *"Local Government Area"* to explore the different cities in the State. You are able to choose multiple cities to view at a single time through clicking the chosen city in the drop down bar.
- Click on *"X Variable:"* and *"Y Variable:"* to view the bivariate plot with different variables such as *Price, Review Scores Ratings, Bedrooms, Beds* etc. This will allow us to explore how 2 variables interact with each other



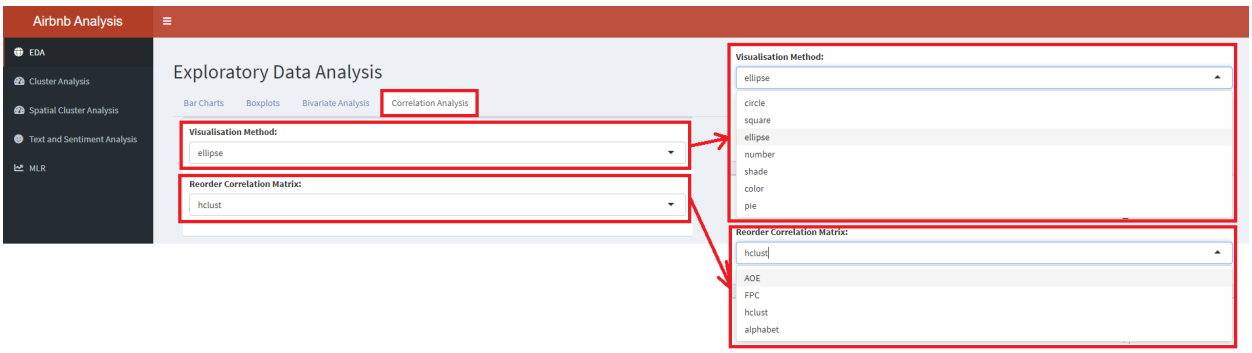
The top plot is a density plot to showcase the density of the *"X Variable"*.



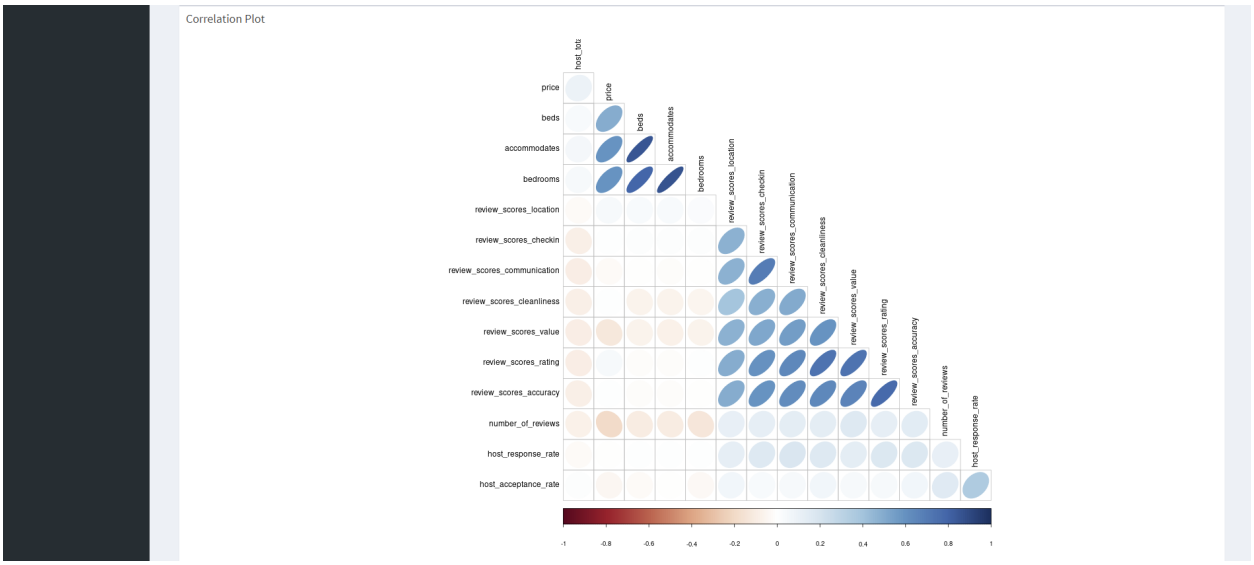
The bottom plot is a bivariate plot to explore the relationship between 2 variables.

The two tables below the bivarate plots showcase the confidence interval of the *"X Variable"* and *"Y Variable:"*.

Correlation Analysis

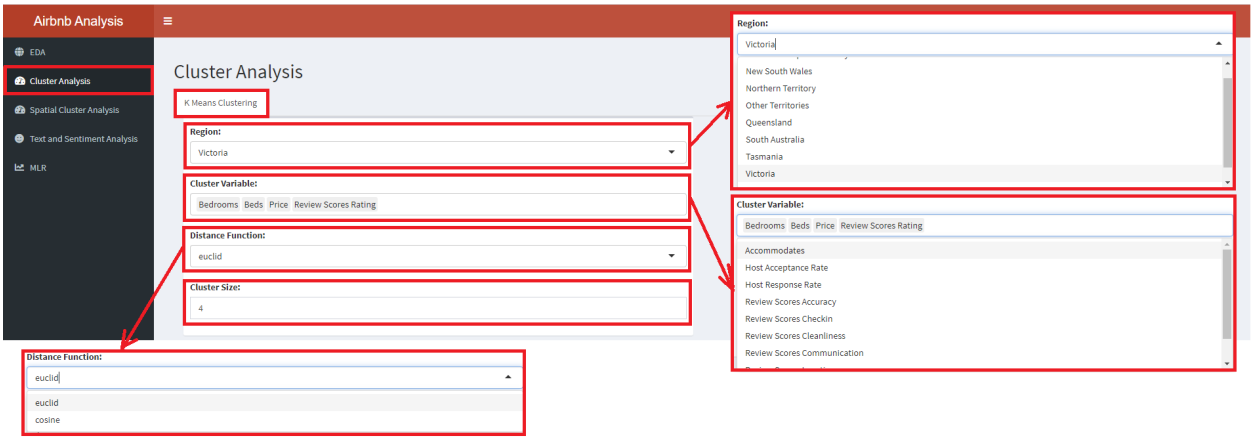


- Click on *"Visualisation Method:"* to view the correlation plots in different methods such as *"circle, ellipse and number"*.
- Click on *"Reorder Correlation Matrix:"* to view the correlation plots in different orders such as *"hclust, alphabet and AOE"*.

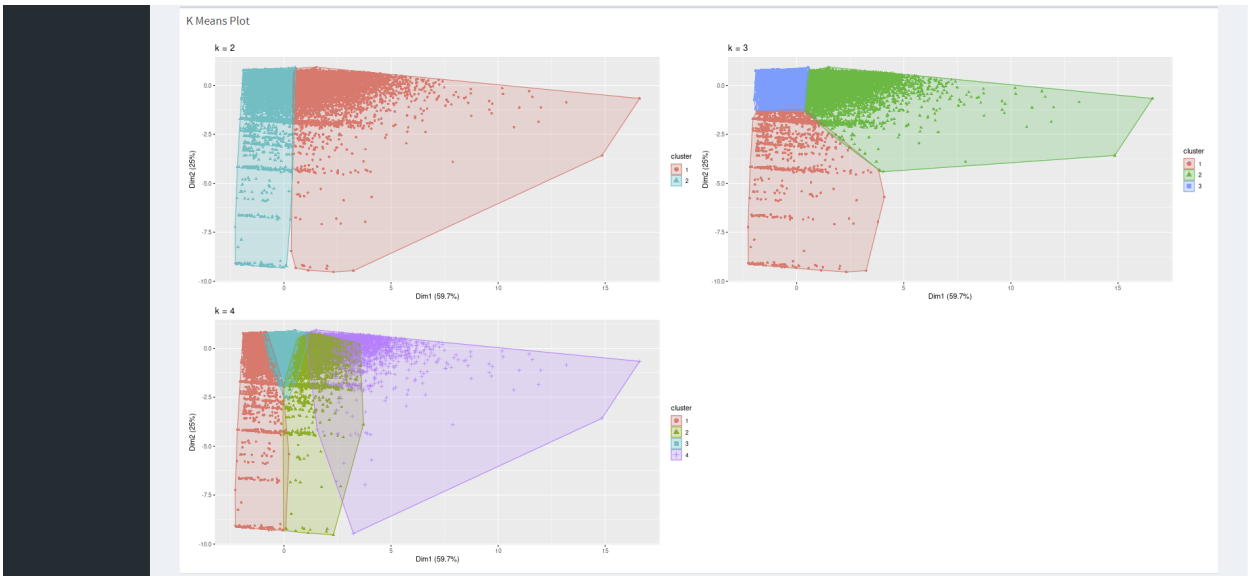


The plot represents the correlation plot of all the variables in the Airbnb data set.

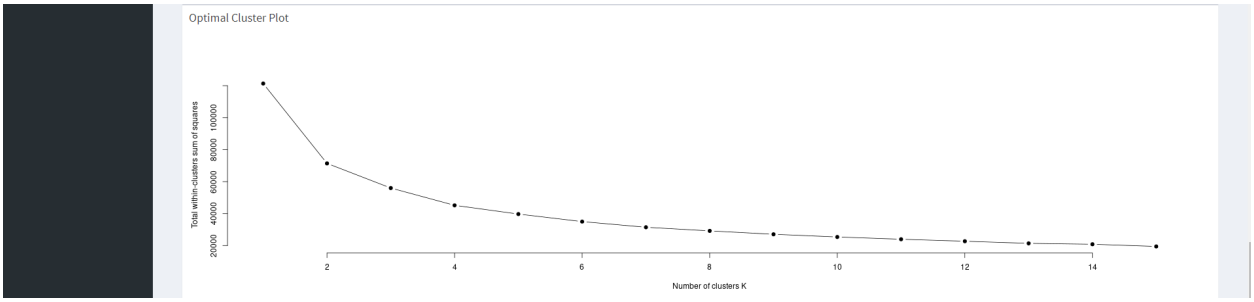
Cluster Analysis



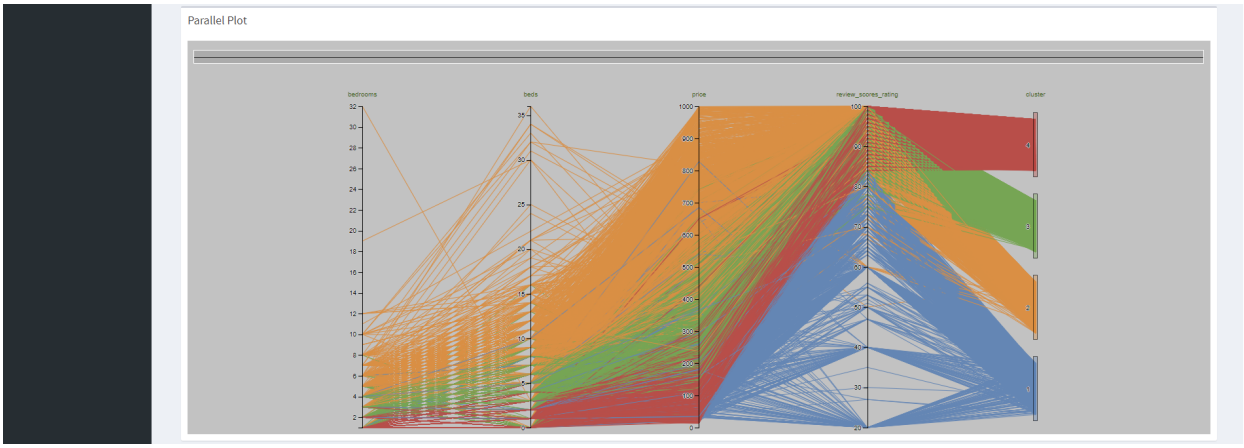
- Click on *"Region:"* to explore the different states in Australia.
- Click on *"Cluster Variable:"* to explore the different variables you would like to cluster. Types of variables are *"Price, Bedrooms, Beds, Review Score Ratings"* etc. You are also able to click to choose more variables that would be used for clustering.
- Click on *"Distance Function:"* to choose between *"euclid"* and *"cosine"* distance for the clusters.
- Toggle *"Cluster Size"* between the sizes 2 and 15, to see different cluster sizes in the data set.



The first plot is the kmeans plot. it showcases the different cluster sizes.



The second plot is an Optimal Cluster Plot, that gives you an indication the number of clusters that will be best for the data set used.



The third plot is a parallel plot that showcases the relationship of the variables and the clusters. It gives an indication of the characteristics of the cluster that was formed.

Exploratory Spatial Data Visualization and Analysis

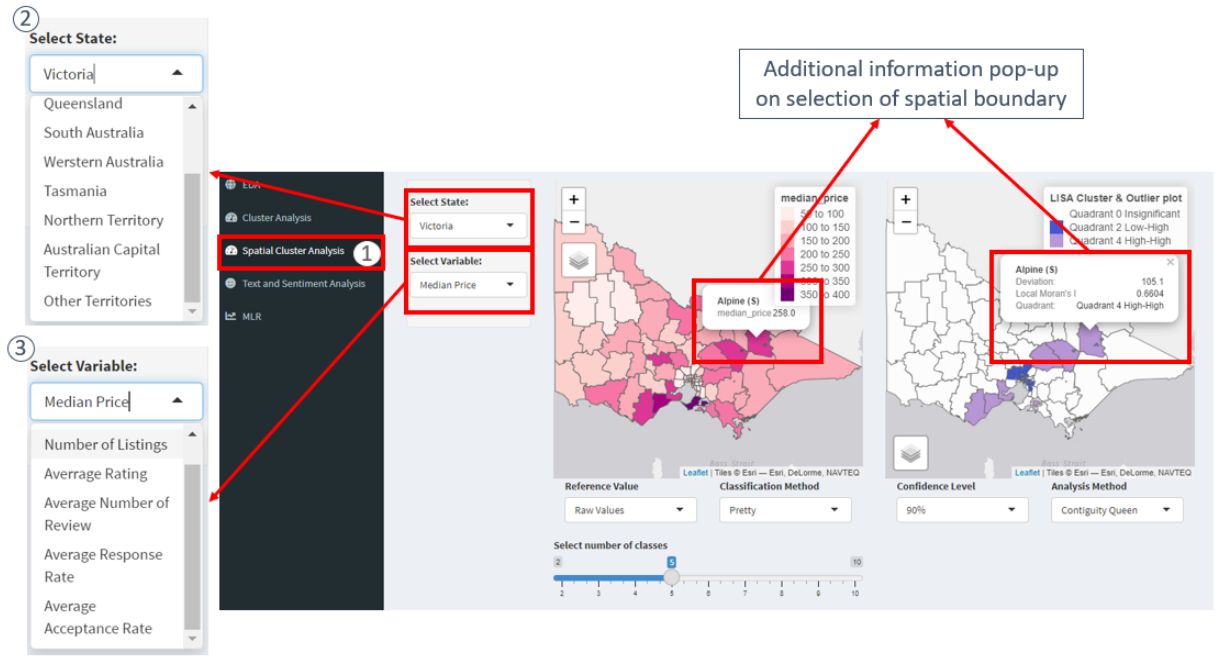
This tab can be used for exploratory spatial data visualization and analysis on the Local Government Area (LGA) level for each of the States of Australia.

Step 1) by clicking on the Spatial Cluster Analysis main left-hand panel, we enter this tab, with two maps coming into view (Note: it might take awhile for the viz to load). The left-hand viz provides an overview of the distribution of either the raw values (default) or clustering statistics for the LGAs within a selected area (State) of Australia. The right-hand viz displays the spatial cluster analysis results - showing areas of high/low clusters, and outliers

Step 2) clicking on the drop down list on the "Select State:" tab allows you to select the state of interest from a list of nine different states in Australia (*Note: while the visualization allows for user to select the "All" state option, it is highly discouraged as some of the spatial analyses do not have meaningful interpretations for areas with considerable missing data - for which at the whole-of-Australia level, data are more sparsed. When analyzing the data, users are also advised to conduct due diligence in determining whether the selected data range of interest has sufficiently satisfied analysis requirements.*)

Step 3) allows you to select the variable of interest from a list of six..

The resultant maps also allows you to click on the area within each contiguous boundary to have more information, such as the name of the LGA, the status (e.g. (C) is a city, (A) an area, (T) towns, (M) municipalities, and (S) shire), and the value of the variable of interest on the first geographical visualization, and the clustering analysis statistics on the second geo-viz on the right.



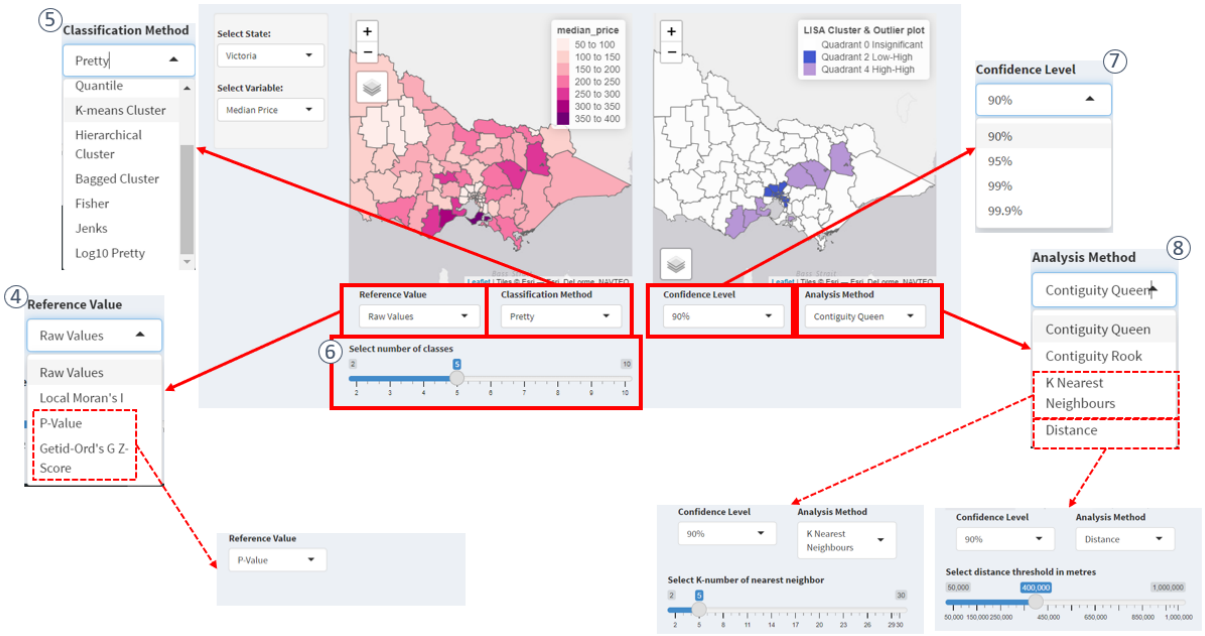
Step 4) we go on to the options for individual geoviz. For the left-hand viz, user can first select whether you want the distribution of raw values displayed, or that of the local indicators of spatial association's statistics displayed. Note that if the Moran's I's p-value or the Getis-Ord's G z-scores are selected, selection for other options (see step 4 and 5 below) will be disabled as these statistical results are to be interpreted on pre-determined intervals.

Step 5) allows you to select the binning methods - in other words, how should the values be split up, be it by equal intervals, by natural breaks (Jenks' and Fisher's), standard deviation bins, etc. A total of 10 options are available under the dropdown.

Step 6) allows user to select the number of intervals to have - i.e. how many different classes should the data be split into. User can shift the slider manually to select the desired class. Note that the selection is very critical (compulsory) for classification methods such as the K-Mean selection; practical for methods such as Equal, Pretty, Jenks and Fisher; but not as useful for methods such as log10 Pretty or even standard deviation.

Step 7) on the right-hand viz, user can select variables to alter the clustering parameters. Firstly, the level of confidence - i.e. how much confidence do we need to reject the null hypothesis that there are no spatial clustering/outlier at the areas identified to have such cluster/outlier.

Step 8) in geospatial analysis, there are also different methods to select the definition of 'neighbor' - this can be selected by changing the "Analysis methods:" panel options. Note that if "K Nearest Neighbors" is selected, an additional slider requesting user to input the number of nearest neighbors to be considered will appear. Same thing, if "Distance" is selected, a slider to select the distance radius to be considered neighbor will appear as well.



Sentiment Analysis

The sentiment analysis module of this Shiny app includes two analysis models: a word cloud and a topic model.

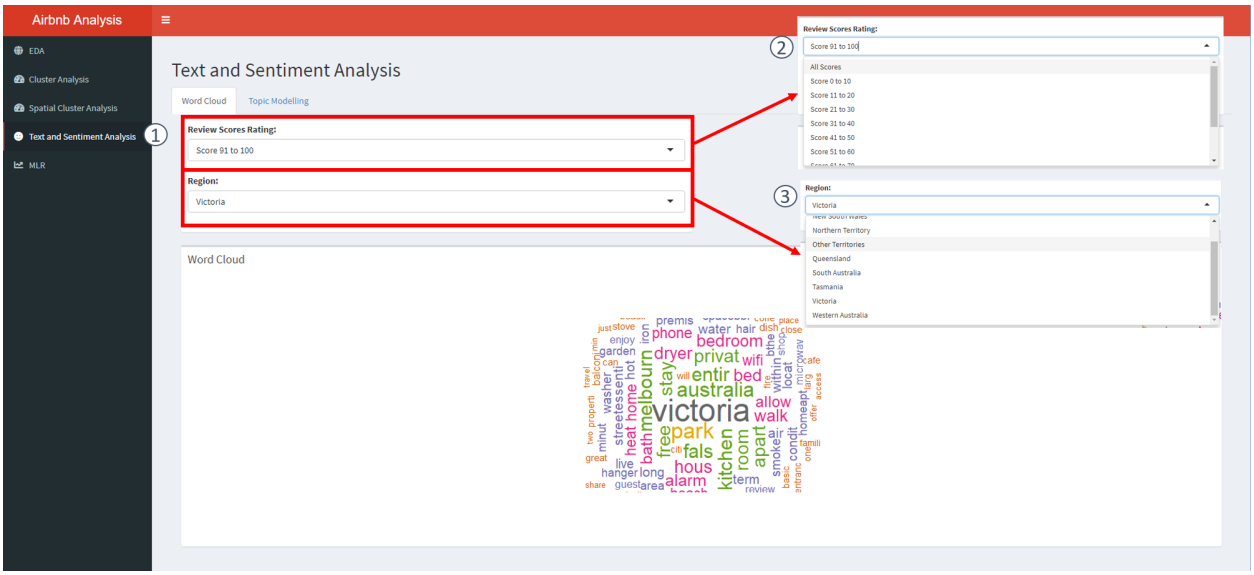
Word Cloud

Step 1) Click on Text and Sentiment Analysis to open the module.

Step 2) Select the range of review score ratings that you want to focus on. By default, all scores will be included.

Step 3) Select the region of focus. In this case, it is Victroria.

The word cloud will be generated. The size of the font represents the relative frequency of the word in the description section of Airbnb listings that fall in the category of the filter settings above (the word in the largest font is the most common word).



Topic Modelling

Topic modeling is an unsupervised machine learning technique that detects word and phrase patterns in documents and clusters them into groups known as topics.

Latent Dirichlet Allocation (LDA) is one common topic modeling technique. The basic assumption of LDA is that similar topics make use of similar words (i.e. distributional hypothesis). The purpose of LDA is to map the corpus to topics covering a significant number of words in the documents in the corpus.

LDA assigns topics to arrangements of words for example, the best word for a topic related to accommodation. This is based on the assumption that documents are written with a certain arrangement of words and that those arrangements will determine the topics. 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.

The navigation is similar to the Word Cloud pane.

Step 1) Click on Text and Sentiment Analysis to open the module.

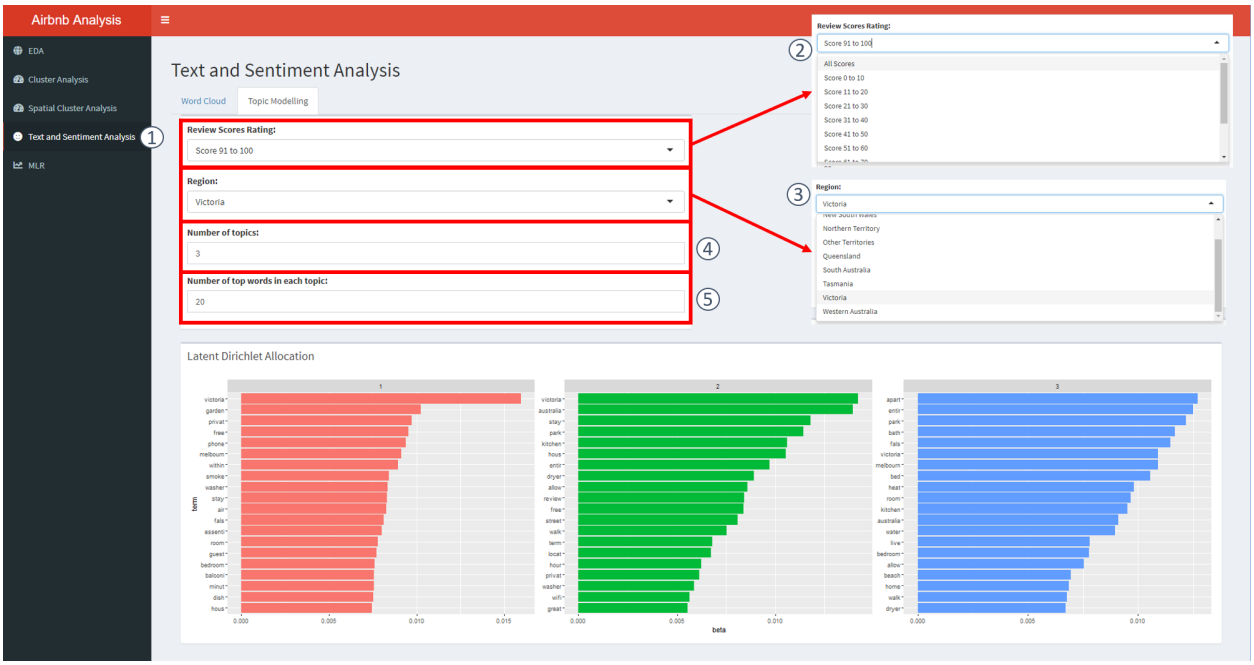
Step 2) Select the range of review score ratings that you want to focus on. By default, all scores will be included.

Step 3) Select the region of focus. In this case, it is Victoria.

Step 4) Enter the number of topics you want identified. The default it three topics.

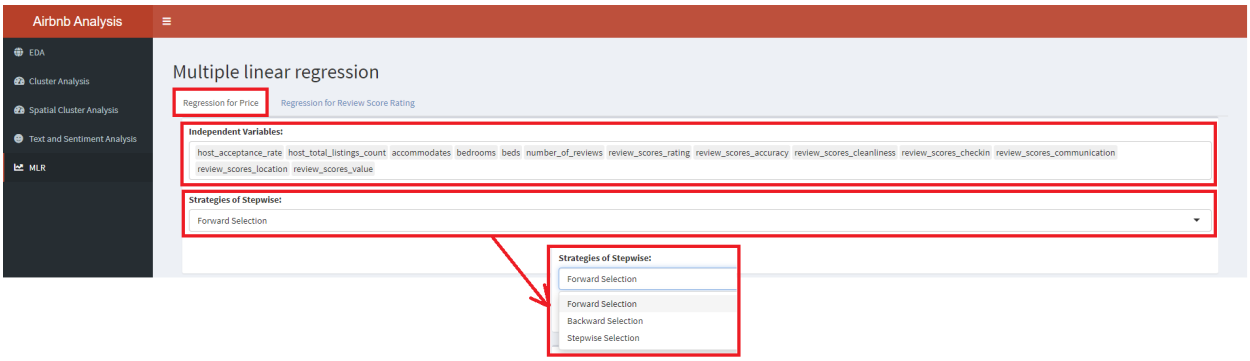
Step 5) Enter the number of words you would like to be displayed for each topic. The default is 10 words.

The LDA topic model will be generated. 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.

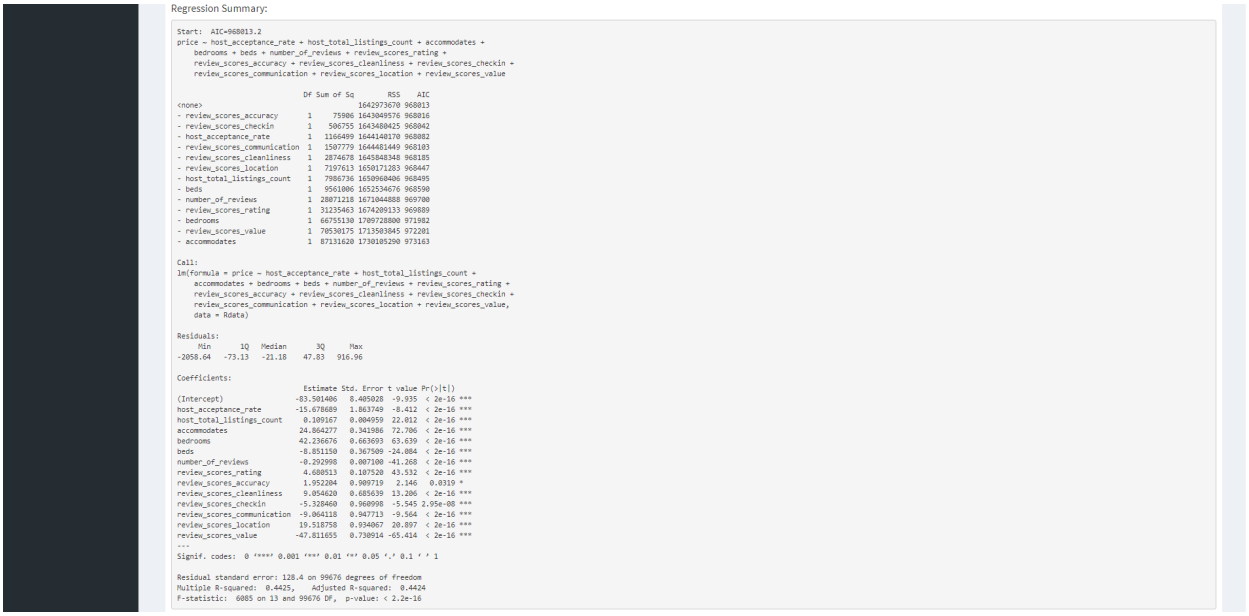


Multiple Linear Regression

Regression for Price

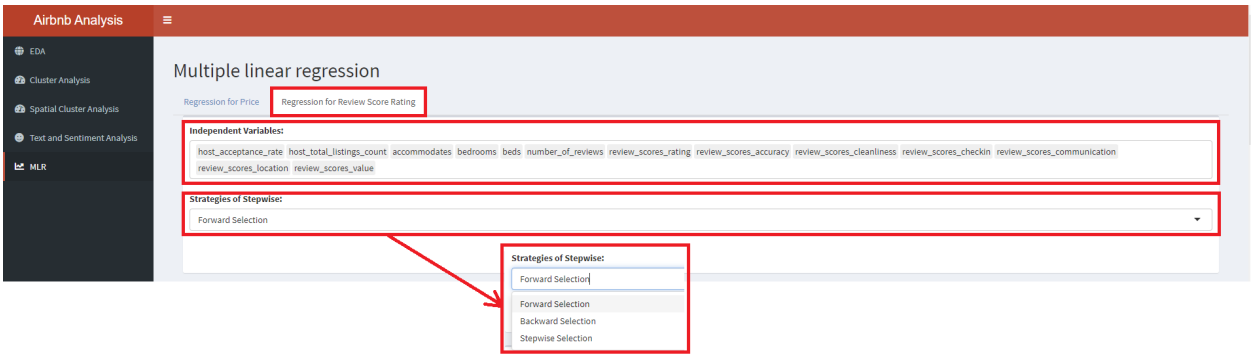


- Click on *"Independent Variables:"* to choose the multiple variables that you would like to apply in the Multiple Linear Regression (MLR). The choices of variables could be influenced through the correlation plot found in the EDA. It is better to choose variables with low correlation with each other.
- Click on *"Strategies of Stepwise:"* to choose whether the MLR will move in a *"Forward"*, *"Backward"*, or *"Stepwise"* direction.



The *"Regression Summary"* showcases the MLR calculation and gives us an indication of the significance of each variable in the regression with respect to *"Price"*.

Regression for Review Score Ratings



- Click on *"Independent Variables:"* to choose the multiple variables that you would like to apply in the Multiple Linear Regression (MLR). The choices of variables could be influenced through the correlation plot found in the EDA. It is better to choose variables with low correlation with each other.
- Click on *"Strategies of Stepwise:"* to choose whether the MLR will move in a *"Forward"*, *"Backward"*, or *"Stepwise"* direction.

The *"Regression Summary"* showcases the MLR calculation and gives us an indication of the significance of each variable in the regression with respect to *"Review Score Ratings"*.