

# ShinyPET: A Predictive, Exploratory and Text RShiny Application using Airbnb data

Ang Su Yiin  
Singapore Management University  
suyiin.ang.2020@mitb.smu.edu.sg

Joey Chua  
Singapore Management University  
joey.chua.2020@mitb.smu.edu.sg

Kevin Gunawan Albindo  
Singapore Management University  
kgalbindo.2019@mitb.smu.edu.sg

## ABSTRACT

The increasing availability of data has resulted in the increased demand for data driven decisions. Although there is an extensive range of commercial statistical tools, they are often subscription-based and demand good technical knowledge to mine and draw insights from. Therefore, it may not appeal to the average user. Using a collection of R packages available, ShinyPET, an R-Shiny application is developed for the average user to perform exploratory and confirmatory analysis, text mining and predictive analysis, as well as to formulate insights and make data-driven decisions. Airbnb data provides a baseline for this application as the data generated is rich in information, consisting of structured, unstructured, and location data. This paper discusses the design framework, use case and future works of the ShinyPET dashboard.

Keywords - Airbnb, Exploratory Analysis, Confirmatory Analysis, Text Mining, Predictive Analytics, Decision Making, R Shiny, Interactive Data Visualisation.

## 1. INTRODUCTION

With increasing affordable data storage and processing technologies, the demand for data-driven decision-making (DDDM)<sup>1</sup> has increased significantly. As Geoffrey Moore opines, "Without big data analytics, companies are blind and deaf, wandering out onto the Web like deer on a freeway." With the use of data driven decision making through analytics tools, firms performance would improve (Yasmin, M et al., 2020) [8]

Airbnb is an online vacation rental marketplace servicing a community of hosts and travellers. By 2020, Airbnb has millions of listings in over 220 countries and regions across 100,000 cities [1]. The data generated provides rich infor-

mation, including structured data e.g. price and location, as well as unstructured data e.g. reviews and listing descriptions. Thus, Airbnb provides a good use case and base case for exploratory and confirmatory analysis, text mining, and predictive modeling as presented in our ShinyPET dashboard.

## 2. MOTIVATION OF THE APPLICATION

The motivation of this project stems from two main issues - the proliferation of data and lack of user-friendly tools to make data-driven decision. According to Harris (2012) [2], data is impractical without the ability to analyse it. Although there is a wide range of commercial statistics and analytics tools, these tools are often subscription-based and require technical knowledge to mine and draw insights from. On the other hand, while open source tools as such Python and R allow for data visualisations, users would require extensive programming background to generate such insights.

Hence, this project aims to develop an interface which is concise, interactive, and user-friendly using R Shiny. With this interface, data-based decisions can be made from the interactive GUI. The R Shiny App will cover 3 modules: 1) Exploratory - users are able to draw interesting patterns based on selected variables, which are augmented by statistical tests based on the chosen variables. 2) Text - users are able to perform analysis on textual data such as reviews to generate quantitative insights.

3) Predictive - users are able to prepare and build a variety of prediction models without the need to have in-depth understanding of predictive models and their algorithms.

This application can be extended to Airbnb data from other countries, and also to other datasets.

## 3. REVIEW AND CRITIC ON PAST WORKS

Radiant application [6], an open-source platform-independent browser-based interface for business analytics in R, illustrates the robustness of Rshiny for web-based application. Developed to promote quick and reproducible data analytics, the application provides interactivity and flexibility in performing visualisation, statistical and predictive analysis. However, there are limitations to the application. First, in terms of exploratory data analysis, most of the plots produced are of static nature which can be enhanced by wrapping plotly around them. Secondly, for statistical testing, users are expected to have a basic understanding of statisti-

<sup>1</sup>DDDM refers to the systematic analysis, examination and integration of data to making strategic decisions, rather than based on intuition or observation alone (Mandinach, 2012) [5]

cal testing methods as they are first required to select their testing method, which can be further enhanced by automating testing methods based on inputs. In addition, newer packages such as visNetwork can be applied for interactive tree visualisation that in turn improves the assessment of decision tree model. Lastly, the statistical testing and charts are placed in separate tabs. In terms of visualisation, a single-page view would enhance the aesthetics and usability.

Text Mining with R Book [7] authored by Silge and Robinson presents a comprehensive approach to handle text. First, the book is content-heavy, which may not be appealing to the typical users. Second, tidytext used for data wrangling and visualisation is widely used, thus allowing users to apply such methods easily. However, these tools require technical skills from users and are not interactive. To allow easy usage and enhance interactivity, packages such as plotly and highchart can be used. Highcharter has various themes and features such as tooltips which greatly enhance visualisation.

Tidymodels [3] has gained interest by providing a framework for predictive modeling and machine learning. It is aligned with the tidyverse principles which leads to a tidier and consistent grammar in the predictive analytics process. Different models offered in Radiant package are also available for implementation in Tidymodels framework, which is why our application leverages Tidymodel as the main framework to conduct predictive analytics on Airbnb data.

Lu, Y., Garcia, R., Hansen, B. et al. (2017) [4] provides a comprehensive summary of research on Predictive Visual Analytics. The paper discusses how visual analytics systems are implemented to support predictive analytics process such as feature selection, incremental learning, model comparison and result exploration. The overall goal of visual analytics is to support explanation in each step of predictive analytics exercise which is also our motivation in developing this application.

## 4. DESIGN FRAMEWORK

The design of our shinyPET is based on ensuring a comprehensive data analysis coupled with aesthetics. Taking into account the user point of view, 3 main principles, namely user-friendliness, interactivity and ease of understanding are adopted.

To get started, the introduction page provides an overview of the application. This allow users to have an understanding of the case that he/she will be exploring.

In the exploratory page, data summary and their tabular form would be presented for user's understanding of data. In the explore sub tab, users are able to visualise provided data based on various variables. This user-friendliness and interactivity provides flexibility and ease of use without needing any technical knowledge.

In the text page, various text mining technique tools are presented in visualisation formats such as wordclouds and bar-charts. Concepts such as sentiment analysis are simplified and recreated into visualisations for ease of understanding.

In the predictive page, various predictive models are avail-

able for user selection. Each subtabs guides user through the process from data sampling until final model selection. Visualisation and interactivity are embedded through graphs and user input menu to provide an organised workflow of predictive analytics.

Aesthetically, the application's colour scheme should be based on the theme of the topic. Using our case of Airbnb, the official colours are Raush, Babu and Foggy (type of gray).

The combination of the 3 principles are consistently incorporated into various steps of the data analysis in the 3 modules, hence providing users an easy and comprehensive way to make data-driven decisions.

## 4.1 Exploratory module

The exploratory module enables users to perform Exploratory Data Analysis (EDA) and Confirmatory Data Analysis (CDA) on selected variables to identify interesting patterns. There are three sections in this module - observe, map and confirm & explore.

### 4.1.1 Observe submodule

As shown in Figure [1], the Observe section provides a summary of the data for users to quickly understand and form questions surrounding the data. Hence, this section was designed mainly based on ease of understanding principles.

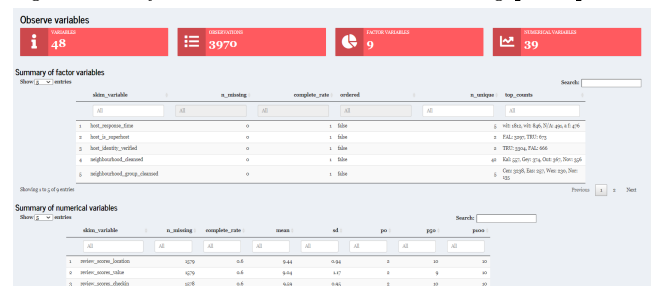
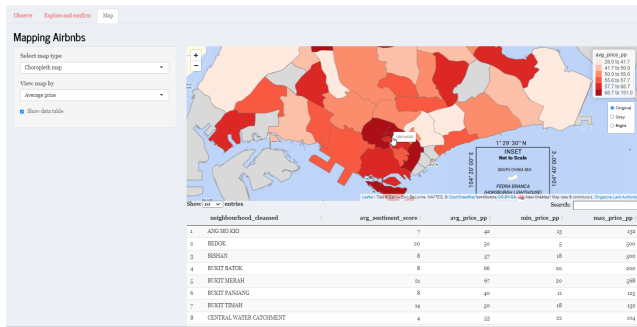


Figure 1: Interface and components of Observe section

There are two main components - first is the top 4 boxes that provide an overview of the data - number of variables, observations and data type. The second component is the tables below shows the summary of each variable by respective data type. The tables allow for some interactivity - search boxes allows user to filter the data accordingly, while the arrow icons next to the variables names allow users to sort the data according to their needs.

### 4.1.2 Map submodule

The Map section, Figure [2], the Map section allows user to explore the geographic patterns of Airbnb listings through thematic maps. Thus, this section was designed based on the three principles stated above and partially based on Shneiderman's interactive dynamics principle of "overview, zoom and filter, then details on demand," save for the 'zoom and filter' portion as it was not applicable to this data.



As such, there are 2 main components of this submodule - the map which provides a macro overview of the Airbnb listings by the selected variable. The second component is the table, which provides details of the map.

### 4.1.3 Explore and confirm submodule

The Explore and Confirm section, figure [3], enables user to explore and perform inferential statistics based on their exploration and questions generated from the previous two sections.

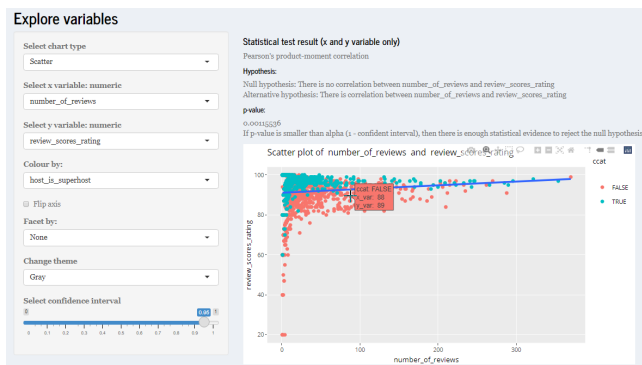


Figure 3: Interface and components of Explore and Confirm section

There are 3 main components - the selection input on the left, the statistical results and the chart.

The selection input was designed to be interactive and user-friendly, allowing users to customise charts based on the drop-down list provided. The application provides for 4 types of chart namely: distribution, mosaic, boxplot and scatter plot. In drop down menus will change according to the selected chart type, for example, if the 'Distribution' chart was selected, only the x-variable drop-down input will be shown.

The chart was designed according to Shneiderman’s interactive dynamics of highlight, filter or manipulate. This graph allows for users to manipulate views by selecting a single object in a plot, highlighting selected records and defining a region on the graph. Furthermore, the plotted chart can be downloaded for users to communicate their findings. See figure [4] for examples.

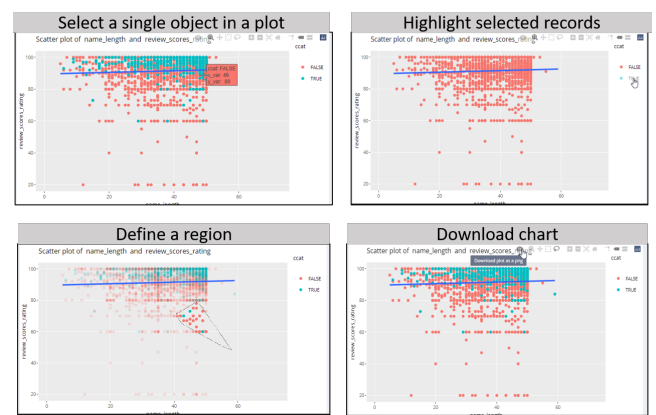


Figure 4: Graph's manipulation function of the Explore and Confirm section

Given that the application is tailored towards users that are not well versed in statistics, the statistical test was designed to be easy to understand, thus the test methods and results are automated based on the selected variables. An interactive slider is provided for user to easily adjust statistical test.

## 4.2 Text module

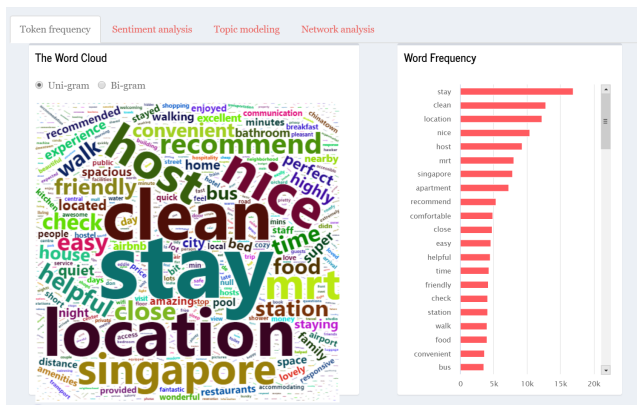
The text module utilise various text mining techniques to transform unstructured text i.e. reviews into structured format to identify patterns and bring about meaningful insights.

Prior to application of text mining techniques, text preprocessing has to be carried out. This involves the use of tokenisation, stemming and lemmatisation. Tokenisation is the process of splitting a column of reviews into tokens such that they are flattened into the format of one-token-per-row. Stemming is the process of separating the prefixes and suffixes from words to derive the root word form and meaning. Stemming algorithms work by cutting off the end or the beginning of the word, taking into account a list of common prefixes and suffixes that can be found in an inflected word. However, the stemming method changes words such as earlier to earli and checking to checkin as shown above. As such, this process was excluded. Lemmatization, on the other hand, takes into consideration the morphological analysis of the words.

#### 4.2.1 Token frequency submodule

To visualise token frequency, wordcloud is commonly used. Wordcloud provides an easy way to show how frequent a word appears in a corpus. In wordcloud, the size of a word indicates how frequent the word appears in a given text.

Other than considering words as individual units, “ngrams” are also used to tokenise pairs of adjacent words. ngrams provide context in sentiment analysis. For instance, while the word “happy” can be positive, in a sentence which contains the words “not happy” would mean otherwise. Hence, performing sentiment analysis on bigram allow us to examine sentiment-associated words.



There are two components: on the left is the wordcloud, and on the right is the bar chart the ranks the frequency of word in descending order. From the chart, it can be observed that the words “clean,” “stay,” “location,” and “nice” occurred most frequently. This could suggest that cleanliness of the environment/room is mentioned the most; it could mean that travellers generally prefer clean room. Additionally, location is mentioned the second-most, suggesting that a location nearer to city for convenience, or a quiet location is important as customers.

#### 4.2.2 Sentiment analysis submodule

In this submodeule, 3 dictionaries were used to plot word-cloud that shows both the frequency and sentiments. First, AFINN lexicon measures sentiment with a numeric score between -5 to 5. BING categorises words as either positive or negative. NRC categorise words into emotions.

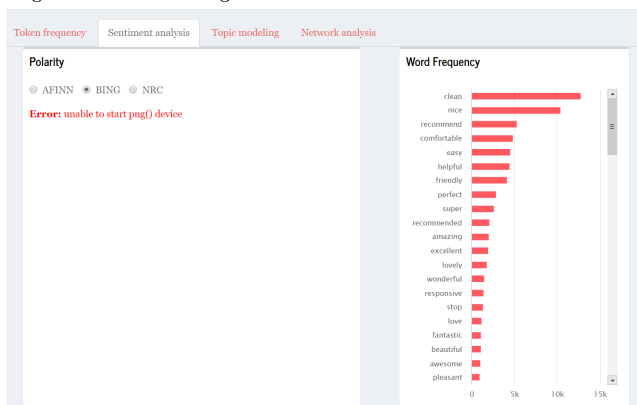


Figure 6: Interface and components of Explore and Confirm section

Users can select the various lexicons to view the wordcloud. To further value add to the wordcloud, a further visualisation is conducted. For AFINN, bar chart is plotted to show the spread and weightage of sentiments. For BING, the distribution of negative, positive and netural sentiments are shown. For NRC, a radial plot to show the tendency for customers to lean towards is shown. From the chart, it shows that sentiments generally lean towards positive and joy, while less on sadness and surprise.

### 4.2.3 Topic modelling submodule

Latent Dirichlet allocation is an example of topic modeling algorithm, based on 2 principles: 1. Every document is a mixture of topics. For example, document A is 90% topic on location and 10% on host's hospitality. Whereas, document B is 30% topic on location and 70% on host's hospitality. 2. Every topic is a mixture of words. For instance, based on Airbnb data, one topic can be cleanliness, and the other topic can be on amenities.

Figure 7: Interface and components of Explore and Confirm section

Users are able to select the number of topics they would like to understand further. For instance, by selecting 2 topics, we can observe that the most common word in topic 1 is "x," "y," "z." The most common words in topic 2 include "a," "b," "c." We can also observe that there are overlapping topics. An example of usage in topic modelling would be to improve search algorithms, suggestions for users. By analysing the topics and develop subtopics, the most relevant listings e.g. cleanest room, or closest location could be shown to the users.

#### 4.2.4 Correlation network submodule

Word occurrences and correlations are commonly used to identify family of words.

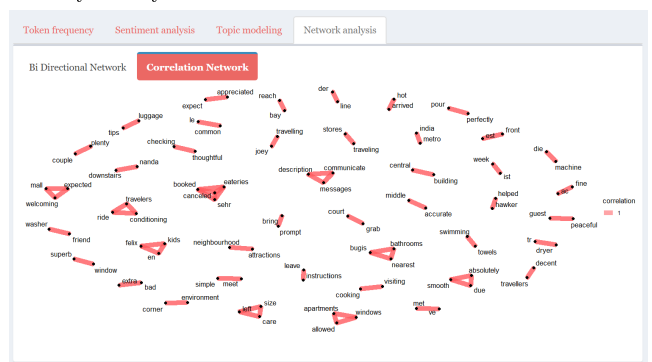


Figure 8: Interface and components of Explore and Confirm section

There are 2 options. First the is X graph. Second is the X graph. From the diagram, it can be seen that there is strong connection between words such as “X” and “X,” “Y” and “Y” and “Y.”

### 4.3 Predictive module

Our predictive module design framework follows Tidymodels framework for data pre-processing, model training, tuning, and validation. On top of that, feature selection are supported by other R packages such as ggcorplot (for correlation matrix), ranger and Boruta (for feature importance). The visualisation and interactivity are embedded in each step of predictive analytics as explained below.

#### 4.3.1 Data sampling submodule

In this submodule, selection of training-test split proportion provides user with flexibility in deciding how to spend data budget on the model development process. The distribution plot between training and test set are also displayed to highlight any potential bias in the training data set.

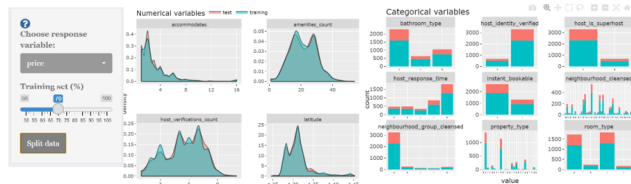


Figure 9: Data sampling and distribution plot

#### 4.3.2 Feature selection submodule

To support user with selection of variables, correlation matrix with customised correlation type and p-value criteria are provided. In addition, variable importance score from 2 different methods highlights useful predictors towards response variable.

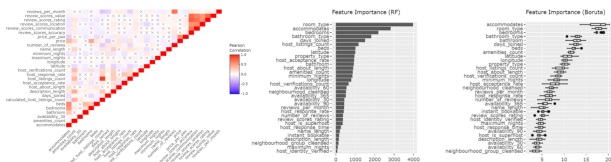


Figure 10: Correlation matrix and variable importance

#### 4.3.3 Data transformation submodule

Prior to training, transformation steps are performed using recipe package in Tidymodel. Following transformation steps, plot between pre and post processing step is added to increase user awareness on what transformation steps are performed and on which variables.

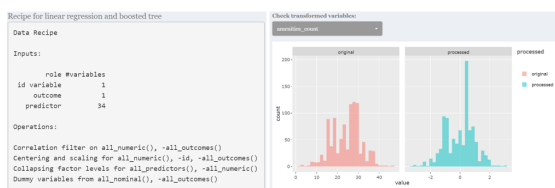


Figure 11: Data transformation steps

#### 4.3.4 Model training submodule

In this submodule, user will be able to select from 5 different types of predictive model for training. For linear regression model, coefficient estimates are shown with option to filter important variables based on p-value. For decision tree training result, visNetwork package is used for its decision tree plot to improve result evaluation.

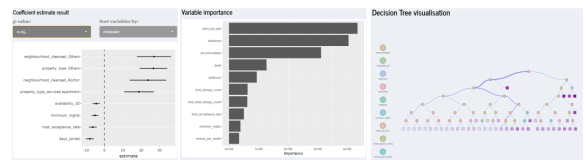


Figure 12: Training result evaluation

Trained model is then assessed using test set (unseen data) which is then assessed by plotting the actual and predicted value on an Rsquare plot. Table of metric performance such as root mean square, mean absolute error, and Rsquare value is also provided.

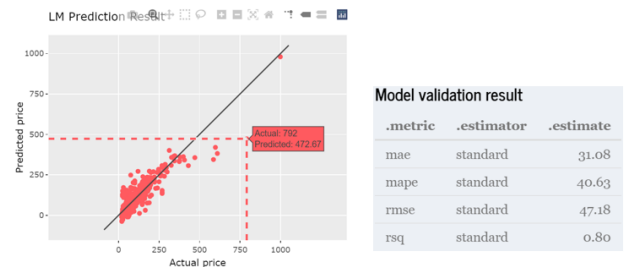


Figure 13: Validation result evaluation

For linear regression model, user will have an option to further evaluate cases with high prediction error. Training set distribution plot is overlapped with predicted values, which may highlight outlier in the predictors. User can also choose to show N number of top predictor based on p-value score to be displayed.



Figure 14: Prediction error assessment

Tree based model and generalised linear model use cross validation training set to tune the model's hyper-parameter. Plot of model performance using different hyper-parameters settings are available for user to understand the change in performance.

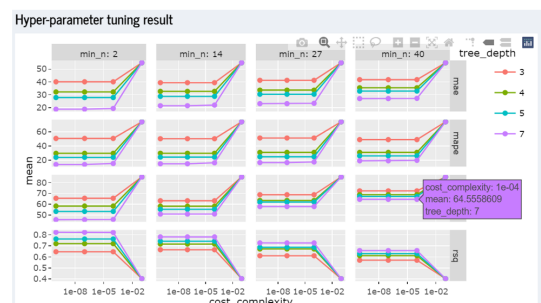


Figure 15: Hyper-parameter tuning result

#### 4.3.5 Model selection submodule

In the final submodule, all trained and validated models are gathered and their metrics are compared for user to choose the final model. Once selected, user will be able to provide new input variables to each predictor and get the response variable using the selected model.

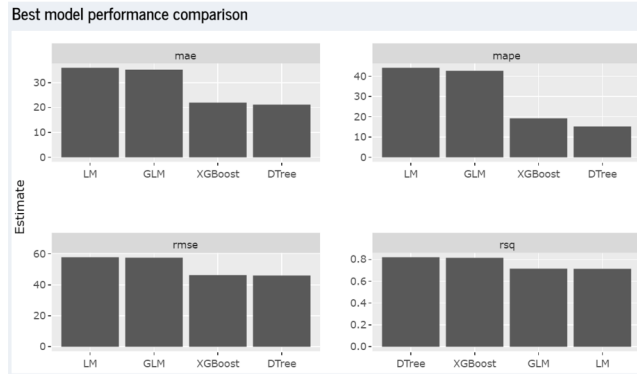


Figure 16: Models performance comparison

The combination of these three modules along with its interactivity and usability would empower users to make data driven decisions with the insights generated.

## 5. CASE STUDY : AIRBNB SINGAPORE

We selected Airbnbs in Singapore data as our case study due to its variety of data - structured, textual and location data.

Since founded in 2008, Airbnb has revolutionised the tourism industry and has become one of the largest sharing economy. In 2019, Airbnb generated USD4.7b in global revenue with close to 7 million listings on its website [1]. Airbnb plays an important role of connecting hosts and travelers who would not typically interact in absence of the platform. However, given the sheer number of Airbnb listings available, Airbnb hosts face stiff competition not only from other Airbnb hosts, but also from hotels and service apartments. Meanwhile, for Airbnb guests, going through each listings and their reviews on Airbnb website can be quite overwhelming and a tedious process. Hence, this application allows users to analyse their needs and compare across other listings.

The two main data used for analysis are the listings and reviews of Singapore Airbnb data, which were obtained from InsideAirbnb on the 27 January 2021. The listing dataset contains close to four thousand listings with 74 variables, while the review dataset consist of fifty thousand reviews and 6 variables.

In order to reduce the loading time of the application, the datasets were preprocess prior with only the cleaned dataset loaded. In addition to the usual pre-processing steps, redundant variables such as listing id, url id, were removed. Additionally, unstructured variables were converted into structured variables by counting the length of the text.

### 5.1 Geographical distribution of Airbnbs

The point symbol map reveals that Airbnbs are distributed throughout Singapore, with high concentration around town

center. The 4 distinct hotspots are (1) Geylang/Kallang, (2) Lavender/ Rochor / Bugis , (3) Orchard and (4) Chinatown. Area (3) and (4) are mainly tourist areas - (3) Orchard is the main shopping belt of Singapore, while (4) Chinatown retains significant historical and cultural landmarks. Areas (1) and (2) are popular due to its low price per person (see choropleth of figure x) while staying relatively close to town. Additionally, we noticed that Airbnbs tend to be located along the MRT station track. Given that Airbnb guests tend to be more budget conscious, they would most likely use public transportation.

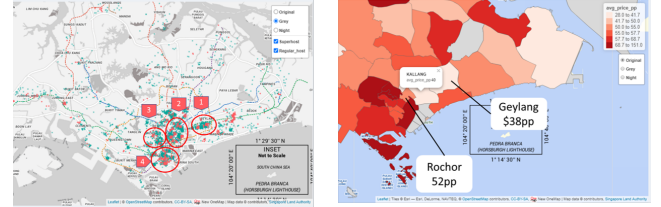


Figure 17: Point symbol map on the left, choropleth on the right

Through the use of these maps, potential Airbnb investor can identify areas that are highly saturated and the average price per person of that area to estimate their investment yield before committing to the investment.

### 5.2 Distribution of review score rating

The overall Airbnb listing review score - 'review\_scores\_rating', is capped at 100 and has a left skewed distribution. This seems to suggest that either most guests tend to score their reviews positively, or Airbnbs with low ratings tend to delist and exit the Airbnb market.

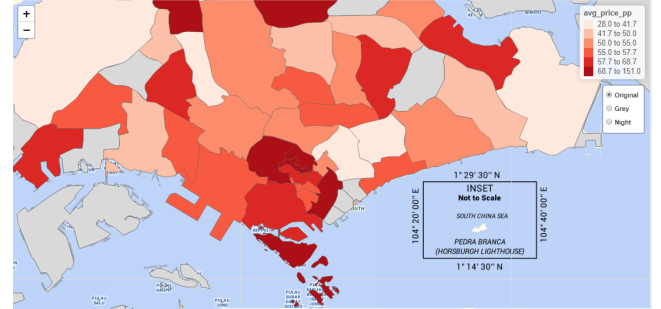


Figure 18: Distribution of review score rating

### 5.3 Confirmatory analysis of superhost and review scores hosting

Figure x suggests that listings with the superhost status tend to have higher review scores as indicated the statistical test results where p-value is less than alpha of 0.05.

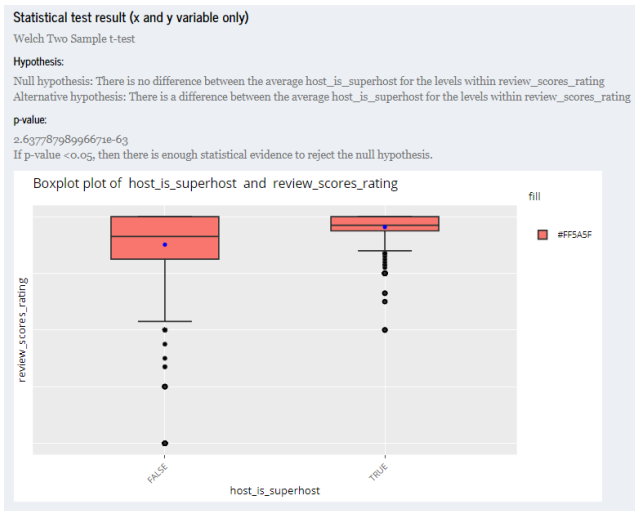


Figure 19: Exploratory and Confirmatory Analysis on superhost status and review scores

## 5.4 Observing correlation among variables

Data sets like Airbnb are rich with large numbers of variable. However, multicollinearity among variables are known to affect predictive model performance. Correlation matrix helps to identify multicollinearity by highlighting variables with high correlation value. In our example below, we observe correlations within rating score components, listing availability period, and review components. With this information, selection of variables can be done properly.

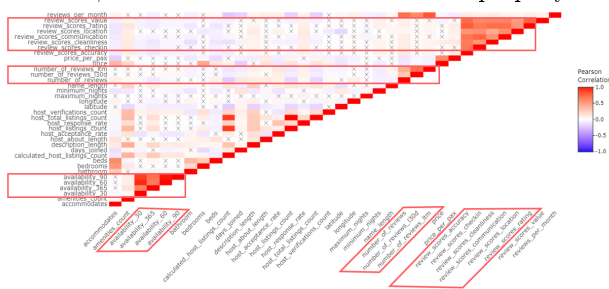


Figure 20: Correlation among variables

## 5.5 Model explanation

In predicting listing price using linear model, the plot of coefficient estimate helps to explain the trained model. In the example below, our interface allows sorting of variables based on p-value score where variables with lowest p-value is located on top. Property type which falls under “Others” category (those with counts of less than 5% in the data set) has the lowest p-value score and positive estimate, which may represent unique property type (e.g. boat, campsite, chalet, villa) where the listing price is above the average price of common property type like apartment and condominium (as shown in the boxplot from exploratory module). Amenities and beds are also in the top 5 predictor where it correlates positively with listing price. However, the error bar is wider for property type “Others” as compared to the amenities and beds, representing more uncertainty in the estimate value.

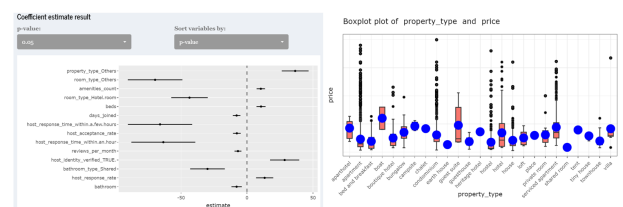


Figure 21: Coefficient estimate and boxplot from exploratory module

## 6. DISCUSSION

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

In this project, we have developed ShinyPET, an interactive web-based application to promote data exploratory, text analytics and predictive analytics on Airbnb data set. The interactivity and functionality of our application provides evidence on the robustness of R shiny as a framework to develop web application, along with the variety of available R packages that serve as building blocks for each module in our application. On top of the functionality, user interface is also carefully designed with the arrangement of submodules for each analytics task, to ease the usage for non-technical user. With the integration of main analytics task (exploratory, textual, and predictive), our application enables user to assess and make use of a dataset from different perspective in a single platform. The modularisation of analytics task allows user to quickly navigate between modules when necessary. Example of benefit from such case has been discussed previously where exploratory module can aid in the understanding of predictive model.

## 7. FUTURE WORK

Shiny PET was built with Singapore's Airbnb dataset as a usecase for using R Shiny to perform exploratory and confirmatory, text and predictive analytics without users needing extensive programming or statistical knowledge. Hence, the application could be further enhance by including a data load and wrangling function to accommodate different datasets.

Additionally, the current types of chart and statistical test are limited with only 4 types of charts and parametric statistical test for each chart type respectively. Other charts, such as violin and bar charts, can be incorporated further. Additional hypothesis testing methods can be included such as non-parametric test for median, statistical test by pairs and others. The current application only supports two types of map, other spatial maps such as kernel density map and navigation map can be included. Moreover, the explore and text module can be combined with views coordinated and linked to provide multiple dimensional exploration.

The current predictive module is limited to 5 types of predictive model. In future, more predictive models can be added to the list, such as neural network to provide user with wider model selection. In terms of hyper-parameter tuning, parameters can be made available for user input to provide more flexibility in developing predictive model. In-

depth statistical analysis in model training such as residual analysis are currently not available and this would be a good additional tool to improve our application.

## 8. ACKNOWLEDGEMENT

The authors wish to thank Professor Kam Tin Seong of Singapore Management University for his extensive guidance and support during this project.

## References

- [1] Curry, D. 2021. Airbnb revenue and usage statistics.
- [2] Harris, J. 2014. Data is useless without the skills to analyze it. Harvard Business Review.
- [3] Kuhn, M. and Wickham, H. 2020. Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles.
- [4] Lu, Y. et al. 2017. The state-of-the-art in predictive visual analytics. Computer Graphics Forum. 36, 3 (2017), 539–562.
- [5] Mandinach, E.B. 2012. A perfect time for data use: Using data-driven decision making to inform practice. Educational Psychologist. 47, 2 (2012), 71–85.
- [6] Nijs, V. 2019. Radiant – business analytics using r and shiny.
- [7] Robinson, J.S. and David Text mining with r: A tidy approach. Welcome to Text Mining with R | Text Mining with R.
- [8] Yasmin, M. et al. 2020. Big data analytics capabilities and firm performance: An integrated MCDM approach. Business Research. 114, (2020), 1–15.