The University of Texas at Austin
Department of Economics
College of Liberal Arts

Singapore Airbnb Market Segmentation

Using K-means and PCA

Department of Economics, the University of Texas at Austin

ECO 395M: Data Mining and Statistical Learning

Professor: James G. Scott

Yangxi Yu, ShuyanYue, Chen Tang

Abstract

This project leverages machine learning algorithms to perform market segmentation based on Airbnb datasets of Singapore[1], so to help Airbnb expand the Singapore market by designing market strategies or advertisements for different target customers. After cleaning and processing datasets containing information, we successfully perform initial EDA (exploratory data analysis) and then implement K-means clustering and PCA(Principal Component Analysis) to identify several meaningful clusters of Airbnb listing: *popular and quick budget-friendly, high-end listings, affordable longer getaway listings, extended stay listings, and mid-range highly accessible.* Finally, we create several advanced visualizations using both geo-spatial data and the results of our analyses.

Introduction

Founded in 2008 and based in San Francisco, California, Airbnb is a marketplace that provides a platform for hosts to accommodate guests with short-term lodging and tourism-related activities. Guests can search for specific types of homes, such as bed and breakfasts, unique homes, and vacation homes etc. Meanwhile, Asia is a vibrant tourist destination, and Singapore now is one of the top 3 fast-growing markets for Airbnb. As tourism is one of the main drivers of Singapore's GDP growth, Singaporean citizens will continue to leverage the opportunity offered by Airbnb to supplement their incomes. Thus, analyzing the Airbnb in Singapore market will be helpful to its company development and citizens to make business strategies.

The goals of the project are to make an overview of the Airbnb Singapore and to segment the existing listings in these areas based on several shared features including things like price, availability, and popularity that might interest different groups of customers. For a user's point of view, the project could be useful to have information about existing houses.

[1] Data source: Inside Airbnb: Get the Data

Strategy and Method

General Logic

1. Gather, clean, and process Airbnb datasets that contain information about the specific Airbnb listings in Singapore, and then make a grasp of the Singapore Airbnb market.
2. Perform market segmentation based on the data by employing an unsupervised machine learning algorithm--*K-means clustering* and *Principal Component Analysis* (i.e., PCA).
3. Identify any meaningful segments that will help achieve the overall goal which is to help Airbnb expand the Singapore market by designing market strategies and advertisements for different target customers.

Detailed Logic

Our project will proceed as follows.

(1) Import libraries and data.

Preliminary data exploration analysis and visualization by box plots and histograms. And then use the interactive map to show insightful and interesting findings.

(2) Clean and process the data for the application of the K-means clustering algorithm.

Firstly, we drop irrelevant columns that are not necessary for this project, such as name, host_id, and host_name. Next, we identify and handle missing values in the dataset. Although the data frame is free of missing values, there was a small subset of listings that had a price of 0, and intuitively this is neither typical nor realistic given that renting a property is rarely free, so we will replace those numbers with the median value. In the following steps, we remove the duplicate data and convert integers and categorical variables into numbers for consistency by using the dummy_cols method to implement one-hot encoding. Moreover, given the primary aim of the project, we're going to temporarily drop some of the columns and create a new, simpler data frame that focuses more on price, reviews, availability, and minimum number of nights, we therefore temporarily drop all columns except for *Price, Minimum Nights, Number of Reviews, Reviews per Month, and Availability (365)*. Lastly, we scale the data to implement the machine learning algorithms.

(3). Determine the optimal number of clusters using the elbow method.

To determine the optimal k value for clusterings, we will take elbow graph and CH graph to make sure the best clustering that fits the data with simplicity

(4). Train and implement the K-means clustering algorithm.

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. To put it in simple, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. We will use K-means to make partition of dataset.

(5) Implement the Principal Component Analysis algorithm to perform dimensionality reduction.

PCA, refers to principal Component Analysis, focuses on reducing the feature space, allowing most of the information or variability in the data set to be explained using fewer features. In this part, the main purposes for implementing PCA to validate the outcome of Cluster and better analyze the data.

(6) Create advanced visualizations using the clusters and geo-spatial data.

Using ggmap, we visualized the clusters with different properties in Singapore's map

Exploration, Visual Illustration and Cleaning of Dataset

Exploratory Data Analysis

1. Summary Statistics

(1) Singapore has an average price of around 209 dollars

(2) One peculiar result is that Singapore lists properties that list the price as being 0.

(3) Singapore has the most expensive property at 10,286 dollars.

(4) Singapore has 1 as their minimum value for minimum number of nights, average for minimum_nights is around 41. Singapore has a property where the minimum number of nights is 1000 (that's nearly 2.7 years).

2 Drill Down

Let's take a closer look at some of the interesting features we have discovered.

(1). Singapore has 1 hotel room listed for 0 dollars. This property is hotel room, and the low value is likely due to some promotion.

(2). Singapore's most expensive property is 10,286 dollars. The listing includes the entire home, which located in West Region. And it is a penthouse condo unit.

(3). Singapore has 5 properties that requires 1,000 minimum nights (around 2.7 years).

3 Analyzing Unique Property Types

The majority of the listings for Singapore is private room, which is 1693. Entire home/apt is almost the same as private room, which is 1670. Moreover, shared room is the scarcest type of property.

4 Data Visualizations

We'll now visualize the data. Constructing visualizations can help detect interesting trends and patterns.

- ***Box Plots***

We'll create a few boxplots that visualize the ranges of prices associated with different types of rooms, and it shows that private room and entire home / apartments in Singapore have a wider range of prices with many outliers above the upper quartile.

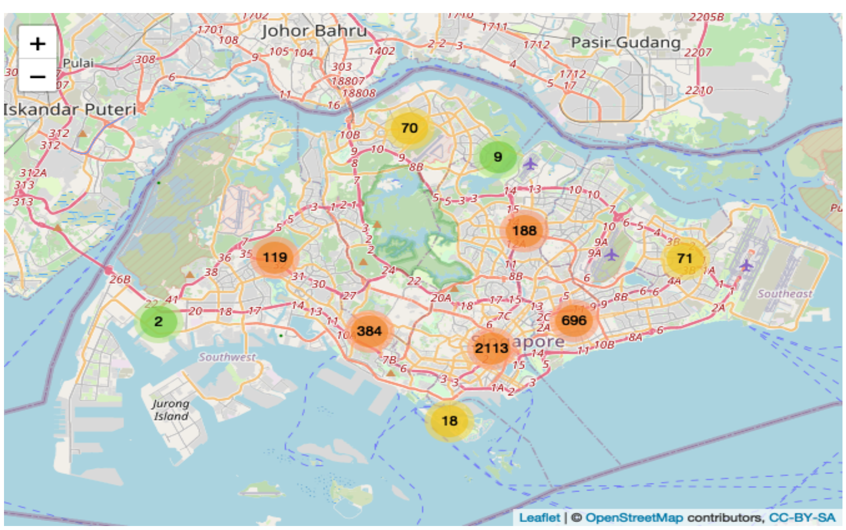
- ***Histograms***

We'll now visualize the distribution of values for the features, minimum nights and price. In order to show more clearly, we select the listings with minimum nights less than 200. And according to Airbnb website, we set the minimum nights less than 30 as short-term rental. The histograms are skewed right. We see that lower prices in Singapore have the highest frequency. As noted earlier, the average price in Singapore is $209. Similarly, the lower values for minimum nights have the highest frequency in Singapore. The average minimum number of nights in Singapore is 41.

Map Visualization

1. Listing Distribution on Map

Take a look at map below showing the number of properties in Singapore, and we conclude that the most listings are in Central Region.



2. Room Type on Map

The entire home/apt represented by the green dot and the hotel room represented by the orange dot are more concentrated and mainly distributed in the Central Region while the private room and shared room are more scattered.

3. Price on Map

Because of the number of properties with higher than 700 only 100, only 2%. In order to show the distribution of property price more clearly, we select the data of property price below 700.

The first map shows that the redder the dot the higher the price of the property in the area. According to the first map we can observe that there are more high priced properties in Central Region. By checking with the second map, which shows the average price in different neighborhood, we can see that the highest average price is truly in Central Region, which is more than $200.

4. Minimum Nights on Map

Because the number of properties with higher than 200 only 42, only 1%. To show the distribution of minimum nights more clearly, we select the data of property price below 200, and we conclude that the distribution of minimum nights is random and dispersed.

Data Cleansing

At this point, we’re going to clean and process the data to be used for analysis.

1. Dropping Irrelevant Columns

There are several columns that are not necessary for this project. These columns include name, host_id, and host_name. So we drop these columns.

2. Missing Values

In this section, we will identify and handle missing values in the dataset. As a result, all missing and anomalous values have been corrected.

There are missing values under the reviews_per_month column. Missing values make up approximately 44% of the values in reviews_per_month. For our purpose, we will simply impute the missing values with the mean value for that column.

Remark: The data frame is free of missing values. However, it was previously noted that there was a small subset of listings that had a price of 0. Our intuition is that this is neither typical nor realistic given that renting a property is rarely free. It is likely that the price of 0 reflects some promotion or deal that is applied to the overall price after certain conditions are met (e.g., renting a certain number of nights). Considering that we only have a small number of these properties, we will replace those numbers with the median value. We will use the median value considering that the price column contains several extreme outliers.

3. Duplicate Data

We'll now check to see if there are any duplicate rows in the data frame. If there are, then we will proceed to remove the duplicated data.

Remark: We are good to go.

4. Check Data Types

Let's examine the types of data and fix any errors by converting data into the proper type. For the sake of consistency, we convert integers into numeric.

5. One-hot Encoding

The data frame contains categorical variables. To apply the machine learning algorithms, we need to convert these variables into a form of numerical format. We will use the dummy_cols method to implement one-hot encoding.

6. Feature Selection

Given the primary aim of the project, we're going to temporarily drop some of the columns and create a new, simpler data frame that focuses on price, reviews, availability, and minimum number of nights. We will therefore temporarily drop all columns except for Price, Minimum Nights, Number of Reviews, Reviews per Month, and Availability (365). The resulting data frame will be ideal for the K-means clustering algorithm.

Remark: The data frame is now ready to undergo the data normalization process

7. Scale Data

At this point, we need to scale the data will help us successfully implement the machine learning algorithms.

Results – Statistical Models

Unsupervise Machine Learning wiht k-means and PCA

Unsupervised algorithm is one of the most effective methodology that making inferences from datasets using only input vectors without referring to known, or labeled outcomes. K-means clustering is one of the simplest and most popular unsupervised machine learning algorithms. To put it simply, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. In this section, we will use k-means to make clear market segmentation to better illustrate the airbnb market in Singapore.

1. Optimal Cluster

Initially, we need to ensure the number of clustering for the dataset on the basis of statistical and empirical knowledge. On the one hand, from the elbow graph, we can eyeball that SSE(the sum of the squared distance between the centroid and each member of the cluster) exponentially plummets before 5 clusters and then gradually decreases. Besides, CH graph shows cluster 5 is a locally optimal point conveying that setting 5 clusters best balances the fitness and simplicity. To sum up, Combined Elbow-plot and CH-plot, we are inclined to choose 5 clusters. On the other hand, The strategy for Airbnb is to build 5 categories of rentals that satisfies all types of demand and purposes for customers, which includes **Budget-friendly listings**, **High-end listings**, **Popular listings**, **Extended stay listings** and **Accessible listings**. In conclusion, we will make a trial for 5 clusters for make segmentation initially.

2. Applied K-means

The next step is applying K-means to the selected feature and artificially label the clusters when observing the properties of these centroids.

3. Identifying and Interpreting Important Clusters

Table 3.1 : Summary of Cluster-center

price	minimum_nights	number_of_reviews	reviews_per_month	availability_365	CLUSTER
125.3636	25.40559	137.776224	2.7310490	199.6853	popular and quick budget-friendly
5259.2667	26.60000	1.000000	0.4236086	224.3333	high-end listings
164.3652	30.74201	9.904070	0.4338533	63.6468	affordable longer getaway listings
144.0818	252.70440	5.622642	0.4161385	200.7547	Extended Stay Listings
214.0121	32.55230	4.547751	0.4051042	341.3562	mid-range highly accessible

(1) For **Cluster 1**, we assigned index 1 as **popular and quick budget-friendly** because the price is cheapest and the requirement of minimum nights is not really demanding.

(2) For **Cluster 2**, we assigned index 2 as **high-end listings** because the price of the center in cluster 2 is about 5259.

(3) For **Cluster 3**, we assigned index 3 as **affordable longer getaway listings** because the price is fairly cheap but not cheapest and the requirement of minimum_nights are relatively inclusive.

(4) For **Cluster 4**, we are assigned index 4 as **Extended Stay Listings** because the requirement of minimum nights is about 252, which means the guests must stay for the longest amount of time compared with others.

(5) For **Cluster 5**, we are assigned index 5 as **mid-range highly accessible** because the available day is the most even though the price is relatively high.

4. Apply Principal Component Analysis

PCA, which refers to principal Component Analysis, focuses on reducing the feature space, allowing most of the information or variability in the data set to be explained using fewer features. In this part, the main purpose of implementing PCA is to validate the outcome of Cluster and better analyze the data.

Table 4.1 : Summary of PCA Variance

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.324926	1.055977	0.9818177	0.9541128	0.5051622
Proportion of Variance	0.351090	0.223020	0.1927900	0.1820700	0.0510400
Cumulative Proportion	0.351090	0.574100	0.7669000	0.9489600	1.0000000

Table 4.2 : Variables in PCA

Variables	PC1	PC2
price	0.0741258	0.6480182
minimum_nights	0.0699482	-0.5690210
number_of_reviews	-0.7028424	-0.0116170
reviews_per_month	-0.6950741	0.1032787
availability_365	0.1117908	0.4954655

Table 4.1 illustrates that variance cumulation of PCA variance. The proportion of Variance means How much variance has been explained in this dimension. Cumulative Proportion means how much variance has been explained in cumulation. In this table, we will reduce the dimension from 5 to 2 because PC1 and PC2 have explained 57.4% of PCA Variance.

Table 4.2 illustrates provides some insightful explanations for PCA. For example, the higher the value of PCA, the higher the price, minimum nights and availability.

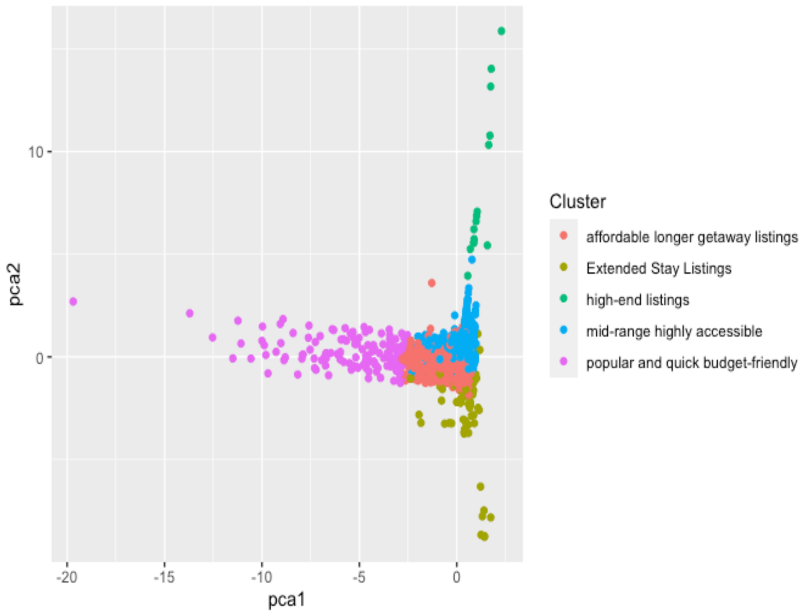


Figure 4.1 Cluster in PCA

Figure 4.1 Cluster in PCA shows the 5 different clusters in pca1 and pca2. Using pca1 and pca2, we can extract all information in only 2 dimensions and better explain the properties of different clusters. As shown in this figure, 5 clusters have been clearly partitioned mutually and exclusively.

(1) From the PC1 dimension, Cluster **popular and quick budget-friendly**, **affordable longer getaway listings**, **mid-range highly accessible**, **Extended Stay Listings** and **high-end listings** lie in pc1 dimension in order from bottom to top. This dimension vividly illustrates the variation of price from low to high. Especially, the **popular and quick budget-friendly** lies far from the origin negatively, which reflects the situation for the low price of this cluster.

(2) From the PC2 dimension, Cluster **Extended Stay Listings**, **affordable longer getaway listings**, **popular and quick budget-friendly**, **mid-range highly accessible** and **high-end listings** lie in pc2 dimension in order from bottom to top. This dimension vividly illustrates the variation of **price**, **minimum nights** and **availability_365**. For **price** level, high-end listings lie far from the origin positively, which shows the property of the highest price.

For **minimum nights** level, Cluster **Extended Stay Listings** lies far from the origin negatively, which shows the demanding requirement for long **minimum nights**.

For **availability_365**level, Cluster **mid-range highly accessible** slightly lies above **affordable longer getaway listings** and **popular and quick budget-friendly**, which accurately reflects that **mid-range highly accessible** shares the good property of accessibility.

5. Visualizing the Difference Between Property Types in Different Singapore Areas

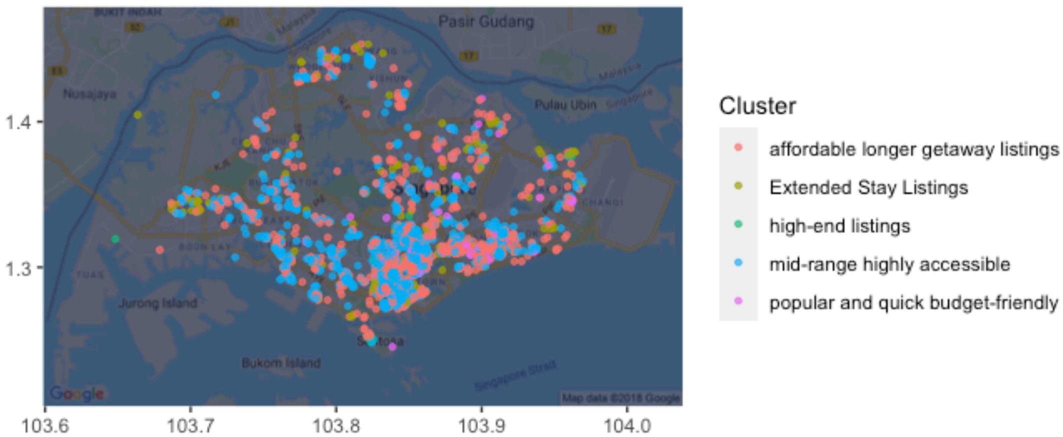


Figure 5.1 Cluster in Singapore

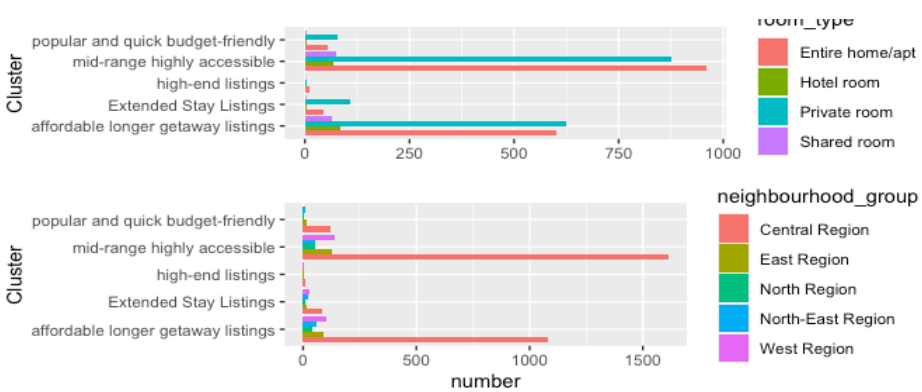


Figure 5.2 Room Type and Neighborhood in Clusters

5.1 Room Types

For Figure 5.1, we find that **private rooms** and **entire room** are the main types for these 5 clusters. For **mid-range highly accessible** and **affordable longer getaway listings**, they include **hotel** and **shared_room**.

5.2 Neighborhood

For Figure 5.1, we find that **Central Region** and **West Region** are the main regions for these different cluster. For the Figure 5.3, we take a new look at the distribution of Cluster in **Central Region**. Here are popular locations for different Clusters. **Downtown Core**, **Geylang**, **Orchard** and **Outram** are the main location for **popular and quick budget-friendly**, **mid-range highly accessible** and **affordable longer getaway listings**. **Rochor** is the main location for **high-end listings**.

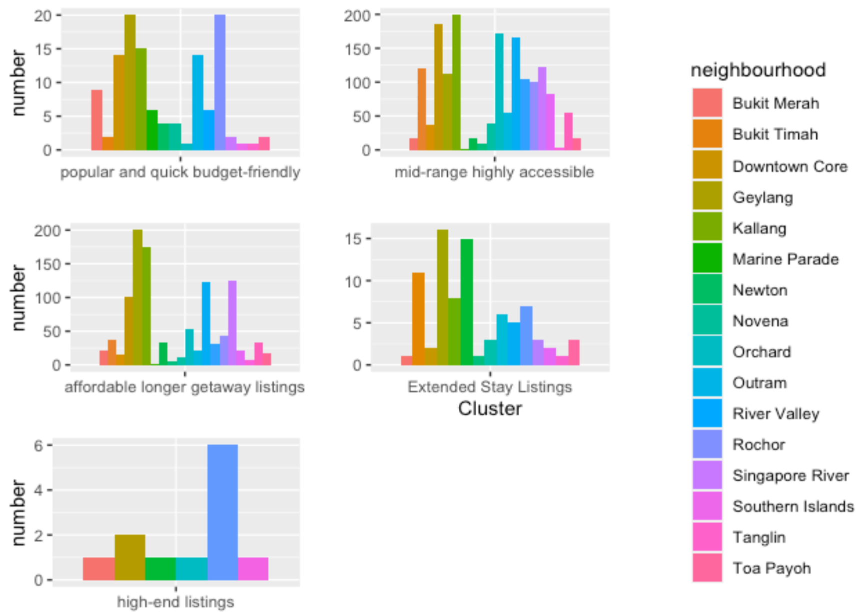


Figure 5.3 Cluster in Different Neighborhood in Center Region

Conclusion

In this project, we implement machine learning algorithms to achieve market segmentation in Singapore. After cleansing several Airbnb datasets, we conducted **EDA (exploratory data analysis)**. We then employed **K-means Clustering** and **Principal Component Analysis** to identify several practical and clear segments with shared features. These shared features included things like price, availability, popularity, and minimum number of nights. After grouping individual listings into their appropriate clusters, we constructed several advanced geospatial visualizations and plots that illuminated interesting aspects of the data.

The project identified the following key segments:

1. Extended Stay Listings

2. High-End Listings

3. Popular, Quick, Budget-Friendly Listings

4. Affordable Longer Getaway Listings

5. Mid-Range Accessible Listings

Admittedly, the project still has several limitations: (1) The result of clustering is not balanced. For example, Affordable Longer Getaway Listings and Mid-Range Accessible Listings take almost 70% of the whole dataset while High-End Listings take only 10 listings. (2) Lack of time-variation scale. The dataset is the newest information on the

Singapore market. It would be better if we could make the contrast between the existing and the past.

This project is beneficial for several reasons. It successfully addresses the primary business problem and achieves the overall business aim. The key segments that we identified can help Airbnb craft use full marketing strategies and promotions that effectively target customers. Last but not least, the purpose for this project is to help our classmate *Jipeng Cheng*, who is a candidate PhD student at Singapore Management University and provide insightful instructions with details selflessly for graduate studying, better seek for an apartment in Singapore.

Appendix

Table A-1 E/V/C - EDA 1. Summary of Statistics

average_price	209.50054
min_price	0.00000
max_price	10286.00000
average_minimum_nights	41.10403
min_minimum_nights	1.00000
max_minimum_nights	1000.00000

Table A-2 E/V/C - EDA 2. Drill Down

Airbnb Properties with the Lowest Prices - $0

	id	neighbourhood_group	neighbourhood	room_type	price	minimum_nights
3120	47790035	Central Region	Museum	Hotel room	0	1

Airbnb Properties with the Highest Prices - $10286

	id	neighbourhood_group	neighbourhood	room_type	price	minimum_nights
982	20791161	West Region	Tuas	Entire home/apt	10286	2

Airbnb Properties with the Highest Minimum Number of Nights - 1000 Nights

	id	neighbourhood_group	neighbourhood	room_type	price	minimum_nights
1365	25505227	Central Region	Kallang	Private room	30	1000
1366	25516668	Central Region	Kallang	Private room	80	1000
1435	27140686	Central Region	Geylang	Private room	50	1000
2317	40113996	Central Region	Kallang	Private room	30	1000
2326	40202857	Central Region	Kallang	Private room	40	1000

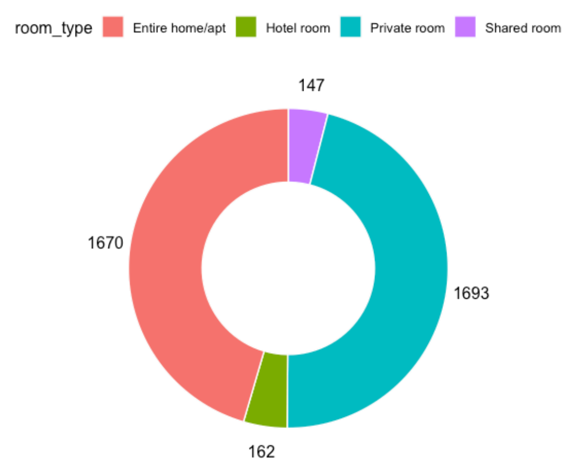


Figure A-1 E/V/C - EDA 3. Analyzing Unique Property Types

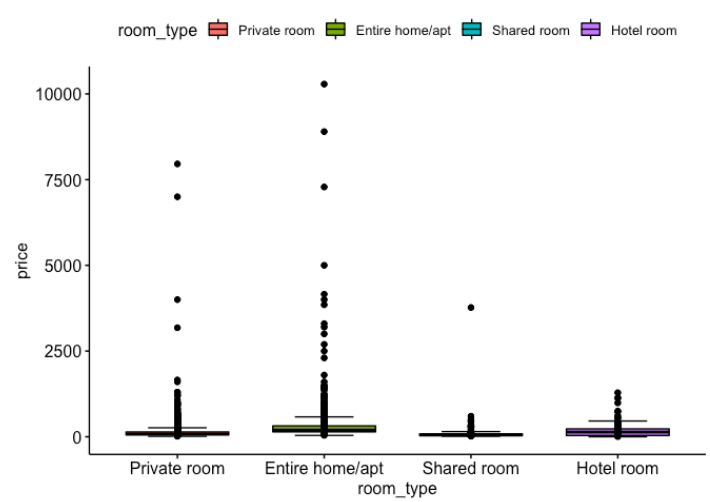


Figure A-2 E/V/C - EDA 4. Box Plots

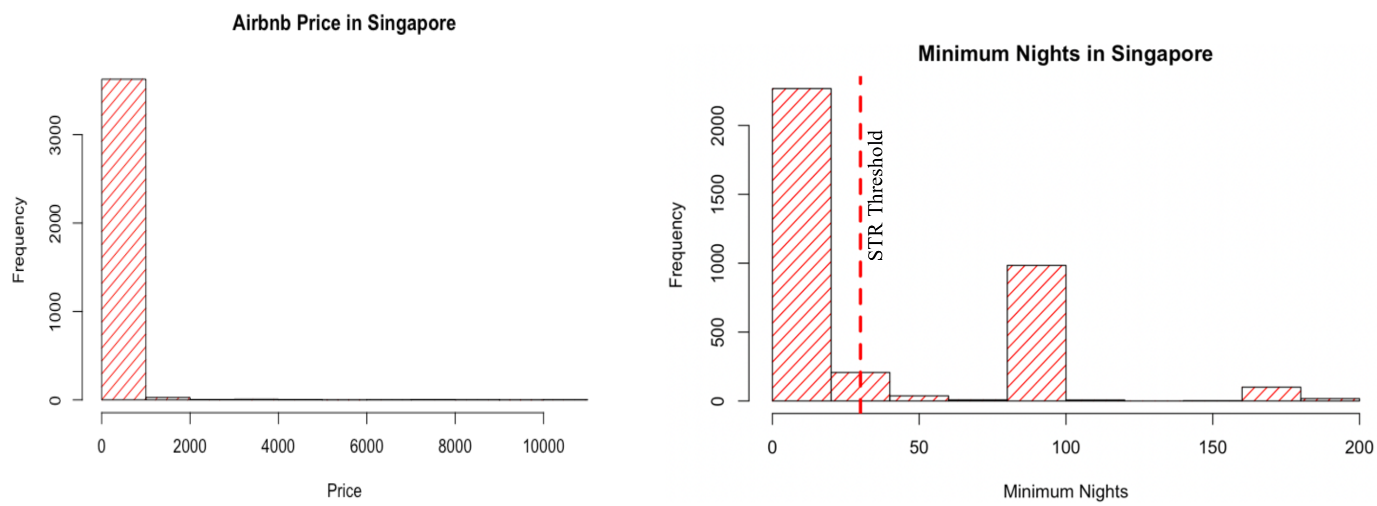
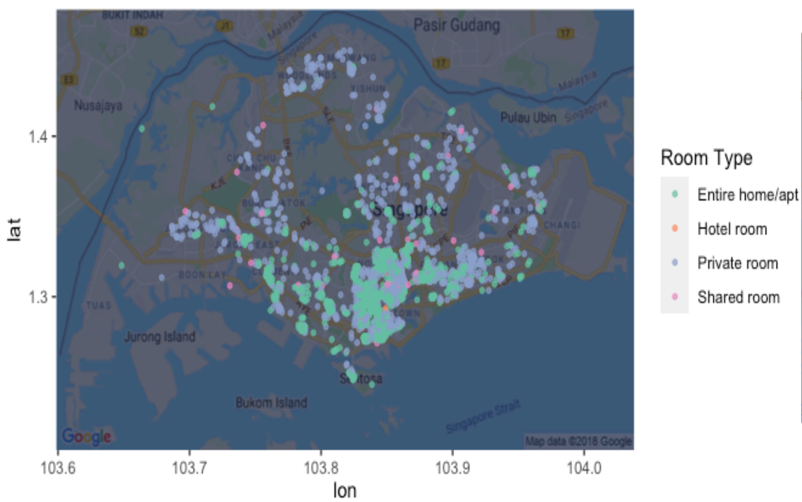


Figure A-3 E/V/C - EDA 4. Histograms



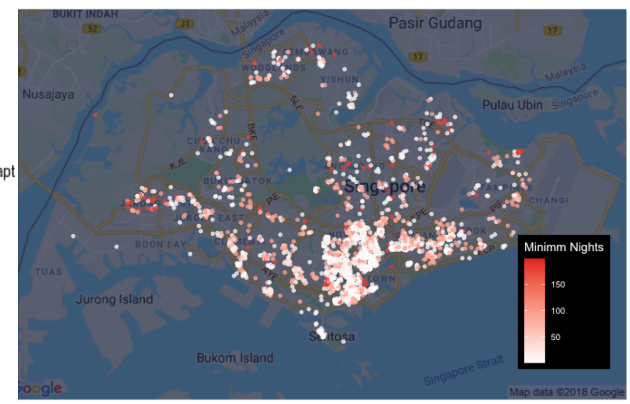


Figure A-4 E/V/C - MP 2. Room Type on Map

Figure A-5 E/V/C - MP 4. Minimum nights on Map

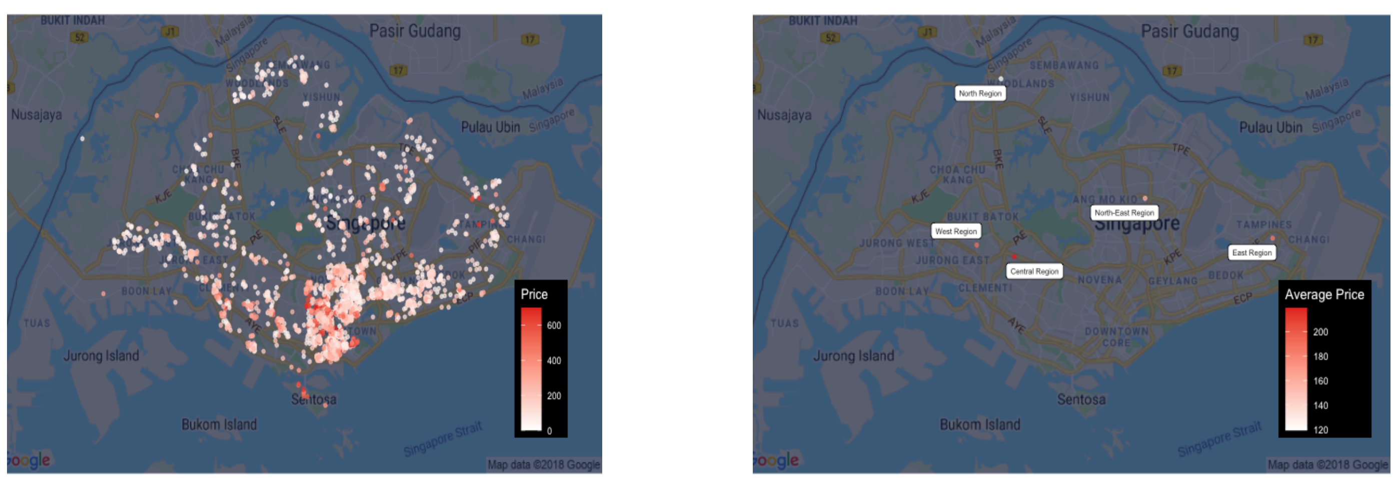


Figure A-6 E/V/C - MP 3. Price on Map

```
## Missings per variable:
##                     Variable Count
##                           id     0
##          neighbourhood_group     0
##                neighbourhood     0
##                     latitude     0
##                    longitude     0
##                    room_type     0
##                        price     0
##               minimum_nights     0
##            number_of_reviews     0
##                  last_review     0
##            reviews_per_month  1628
## calculated_host_listings_count     0
##             availability_365     0
##        number_of_reviews_ltm     0
##                      license     0
##
## Missings in combinations of variables:
##                Combinations Count  Percent
## 0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0  2044 55.66449
## 0:0:0:0:0:0:0:0:0:0:1:0:0:0:0:0  1628 44.33551
```

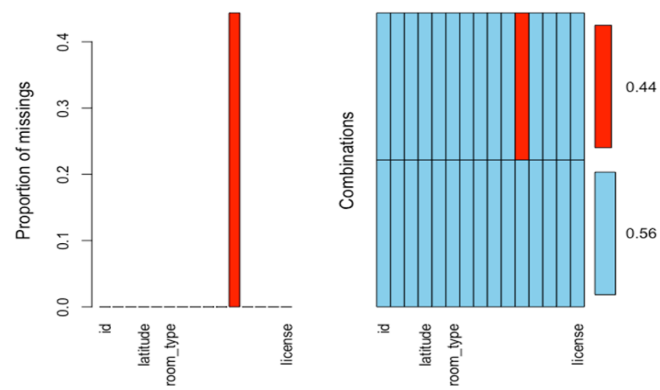


Figure A-7 E/V/C - DC 2. Missing Values

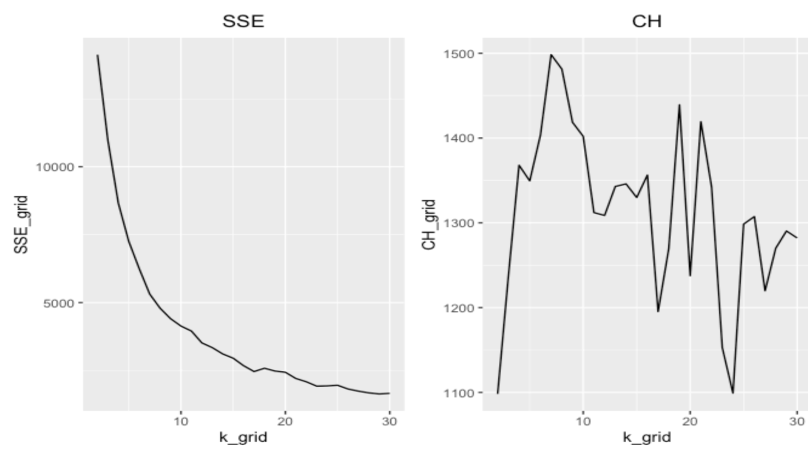


Figure A-8 Results - UML 1. Optimal Cluster - Elbow Plot + CH index

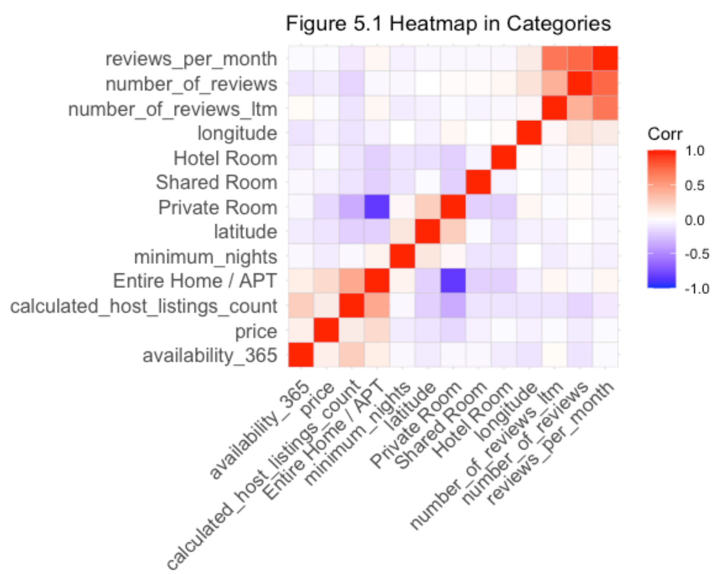


Figure A-9 Results - UML 4. Apply Principal Component Analysis