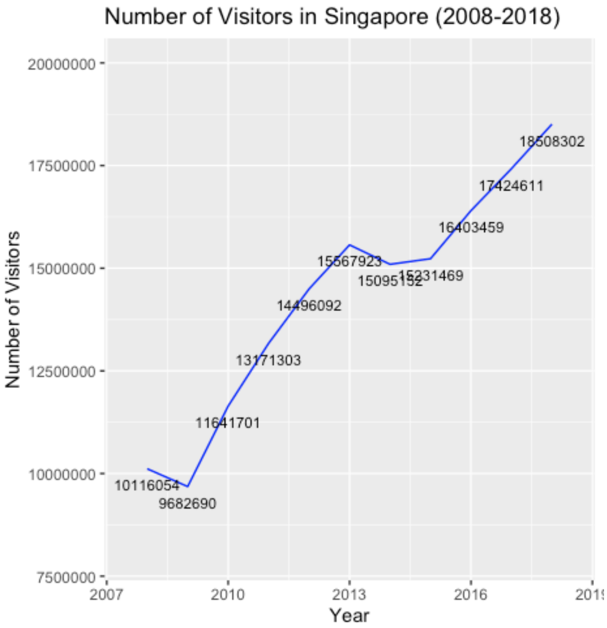
Clustering Airbnb Data In Singapore Using K-Means and CLARA

===

Adinda Rizky Herawati - 425148

1. Introduction

In last 10 years, trend to travel to Singapore is increase, it clearly describe from data source, Singapore Tourism Board [1], that in 2018 the number of tourism arrivals is about 18.5 Million which is twice larger than in 2008. Moreover, it clearly happened because based on news from Channel News Asia [2], many international events held in Singapore such as ASEAN chairmanship in 2018 and ASEAN related events. Following that, also there were some promote Singapore as destination by movie such as Crazy Rich Asians.



Graph 1. Number of visitors in Singapore (2008-2018)

Increasing number of visitors in Singapore give significant effect to revenues in Hotel Industries. In 2018 based on Singapore Government Agency website [3], Singapore's government got 188.2 M SGD from Hotel Industry Sectors. Tourism sector, specifically in Hotel Industries are being great opportunity for residents in Singapore who have available private room/apartment to be rented using airbnb. Airbnb is one of popular home-sharing platform in the world and now there are 150 millions users airbnb in the world. Based on muchneeded.com [4] said that 49% of airbnb users in 2016 prefer stay at airbnb's room as an alternative to hotels.

From airbnb data, In Singapore there are thousands of listing rooms to be rented and by this paper, I would like to give illustration to residents who are interested in housing business to choose which locations in Singapore neighbourhood which have potential rented rooms/apartment to be rented by visitors. Also informed them price of rented room in airbnb applications. In this paper, I am willing to cluster the airbnb data using 2 methods, there are K-Means and CLARA. The reasons, I choose K-Means and CLARA to find out which clustering methods will give the best results for airbnb data. Also in this paper, I would like to know the best number of clusters using silhouette plot.

2. Structure Data

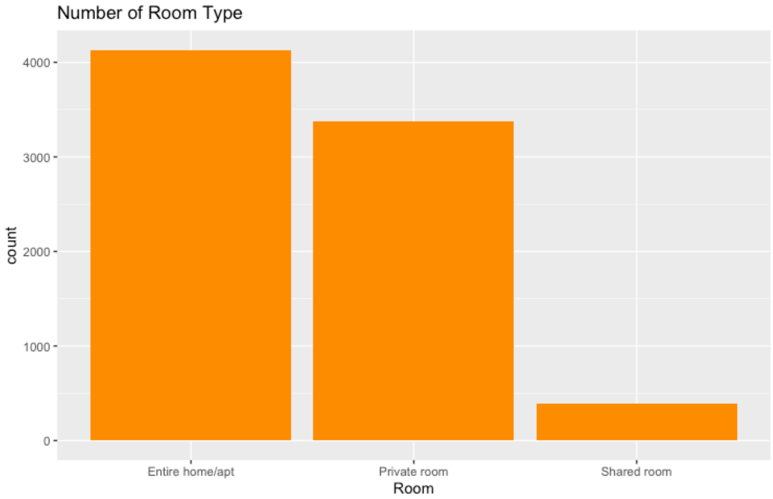
Data for this first project sourced from airbnb website [5]. There are 7907 observations which started from 21st October 2013 to 27th August 2019. This data consists of 15 variables and it shows in Table 2 below.

Table 1. Structure Data

No	Variables	Meaning
1	Id	Listing ID
2	Name	Listing Name
3	Host Id	Host ID
4	Host Name	Host Name
5	Neighbourhood Group	Singapore Regions
6	Neighbourhood	Specific Place in Singapore
7	Latitude	Listing Latitude
8	Longitude	Listing Longitude
9	Room Type	Type of Listing (Entire Home/Apt, Private Room, Shared Room)
10	Price	Listing price in Singapore Dollar Per Night
11	Minimum Nights	Required Minimum Nights Stay
12	Number of Reviews	Total number of Reviews
13	Review Per Month	Average Number of Reviews Per Month
14	Calculated Host Listings Count	Total Number of Listings For This Host
15	Availability 365	Number of Days Listing is Available out of 365

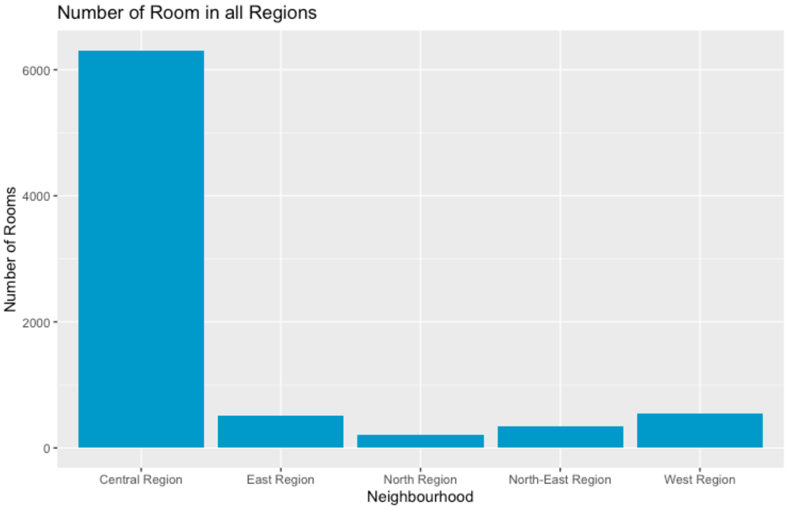
3. Analysis Results

Before starting to analyse data using clustering method. It will be better to know the whole data using descriptive statistics.



Graph 2. Number of Room Type

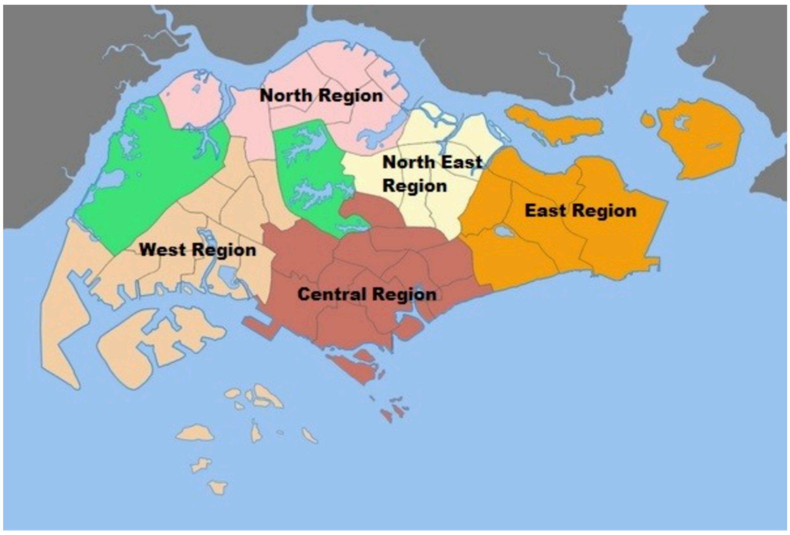
We can see from graph 2, we can see that the highest type of airbnb in Singapore is Entire Home/Apt, it is about 4132 rooms. Following that, the private room in airbnb is being the second largest rooms in Singapore about 3381 rooms and there are only 394 shared room in airbnb Singapore.



Graph 3. Number of Room Type

The highest room number is in Central Region about 6309 rooms, it is clearly because in that region is well-know place for foreigners. Central Region is central of business and shopping centre in Singapore such as Orchard, Chinatown, Bugis, Little India, Marina Bay and etc. Second highest airbnb rooms is in West Region about 540 rooms because that region is quite popular also for tourism such as Singapore botanic garden, Jurong bird park and Science Centre. Following that, in East Region There are about 508 rooms in this region have popular place for tourism like Changi Airport because now Changi Airport become World's Best Airport in 2019 in the world and it attractive for tourism to visit

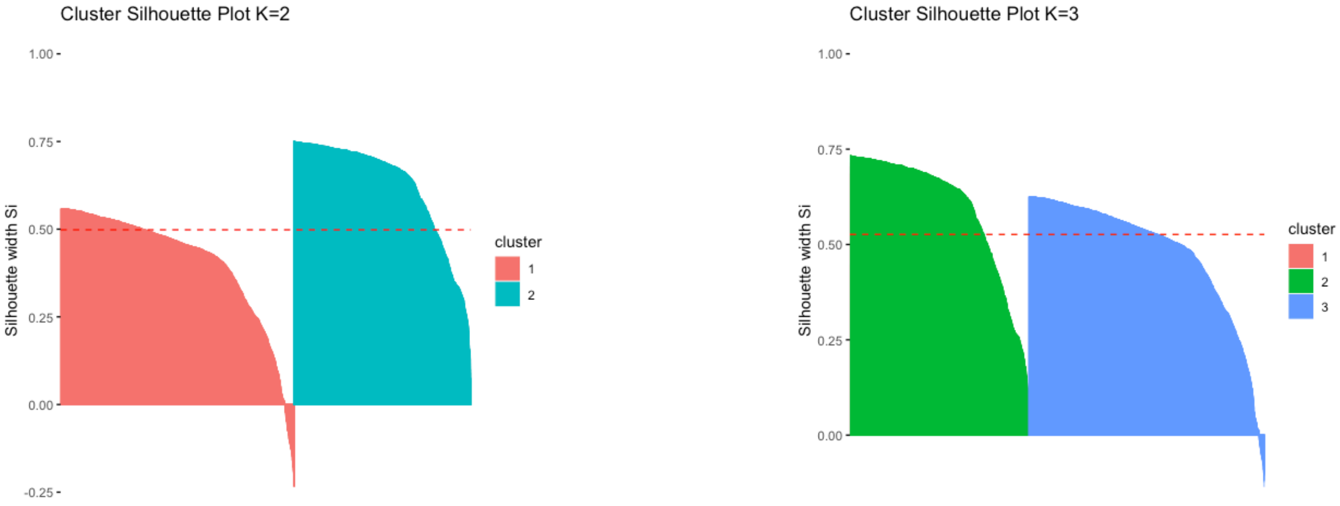
Changi. Next there are 346 rooms to be rented including private room/shared room/ apartments in North East Region and last for the north region have 204 rooms for rented in airbnb because mostly they have natural tourism and it quite attract visitors to stay at North Region.

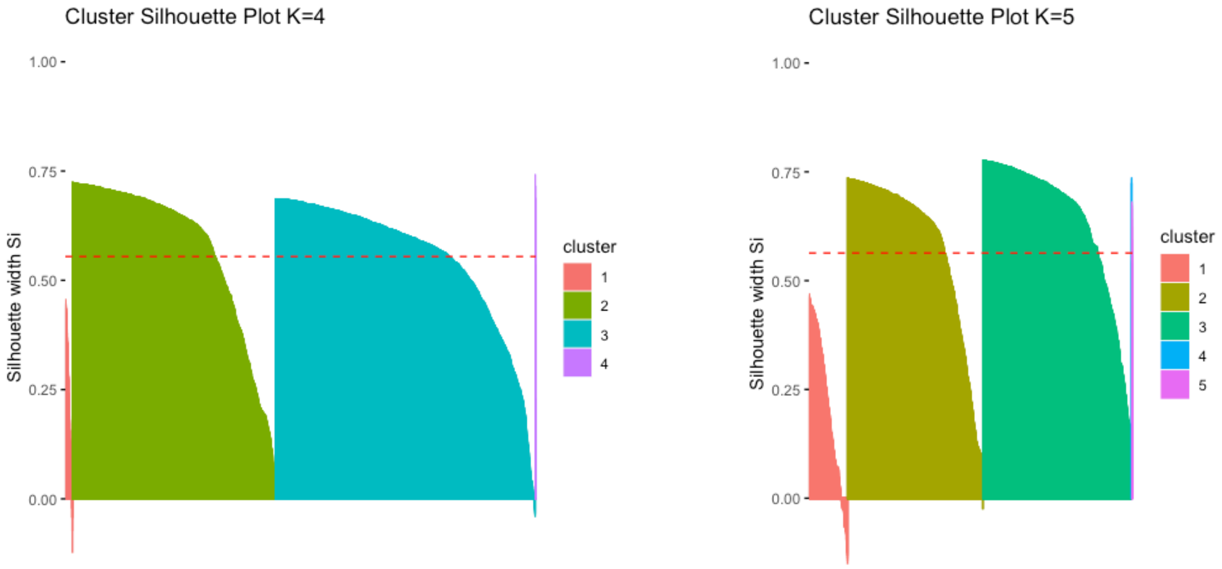


Graph 4. Maps of Singapore

In this paper, I choose two variables for clustering, there are Price and Availability of the rooms out of 365 days because between these two variables there are positive correlation about 0.09800419.

Before explaining in clusters results, first trying to select best number of clusters based on Silhouette Plot. To select the best cluster, the clusters should be above mean or close to +1 and the thickness of the silhouette plot should be quite similar. In Graph 5, The silhouette plot shows almost in all clusters (Cluster Silhouette Plot K=2, K=3, K=4 and K=5) are above mean (red line) and also they have clusters which headed to -1. Clusters who are close to -1, it means that it might have been that those clusters assign to wrong clusters. In Cluster Silhouette Plot K=2, K=4 and K=5, they have different width of the plot, in opposite in Silhouette Plot K=3, they have almost have similar width. This assume that, Clustering using K-Means in K=3 will be the best cluster number.





Graph 5. Clusters Silhouette Plot (K-Means)

Furthermore, in Table 2 below shows that the Cluster 3 has the highest average silhouette score. This results support the silhouette plot before in Graph 5 that Clustering with 3 Clusters will be the best clusters in K-Means Method.

Table 2. Average Silhouette Score

Cluster	Average Sillhouette Score
K=2	0,515
K=3	0,580
K=4	0,490
K=5	0,502

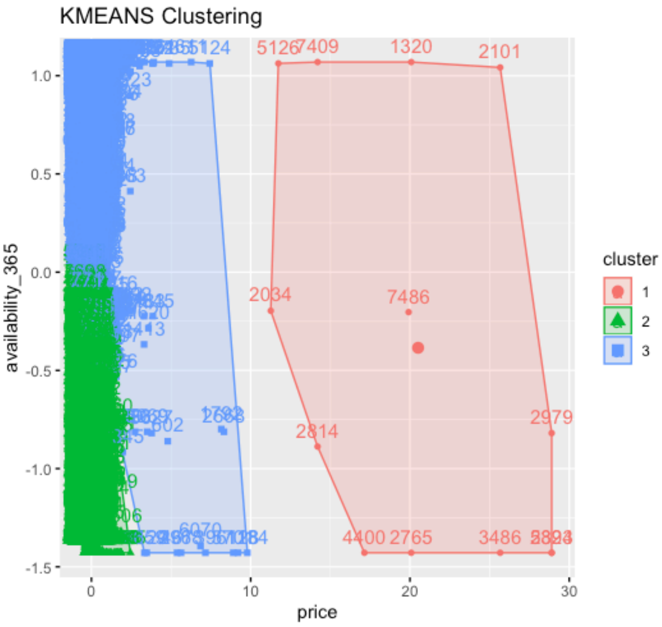
For creating cluster in K-Means, they are using Means as their parameter to determine each cluster. In Table 3 shows, the cluster means of K-Means (K=3). For creating cluster in Cluster 1, the observations must be around mean price 7146.6154 SGD and their availability about 152.46154 ~ 152 days. However in cluster 2, they need the observations for Price around mean 127.6240 ~ 128 SGD and in availability about 53.45876 ~ 53 days. Last for Cluster 3, they need observations for Price around mean 180.7878 ~ 181 SGD and in availability, they need availability around 326.78471 ~ 327 days.

Table 3. Cluster Means of K-Means

Cluster	Price	Availability_365
1	7146.6154	152.46154
2	127.6240	53.45876

3	180.7878	326.78471

After choosing the best cluster, it will analysis further about Clustering K-Means with 3 Clusters. In Graph 6 below, shows the clustering plot between price and availability room in 365 days. However, there are many outliers in cluster 2 and cluster 3, it might happens because K-means method is very sensitive to outliers.



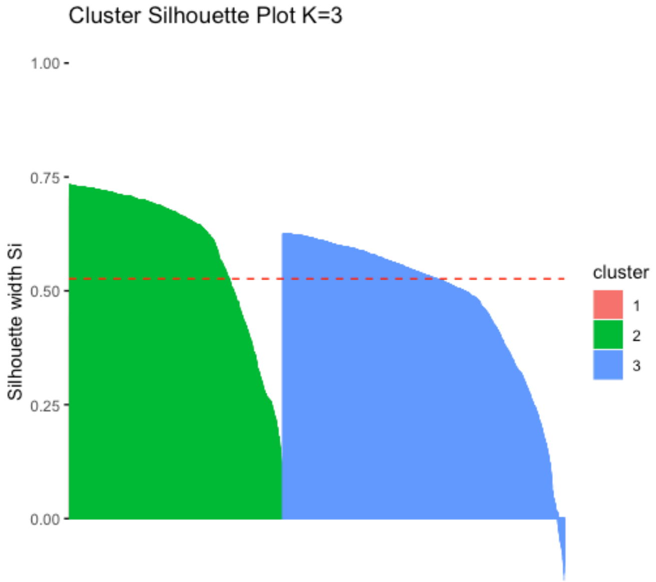
Graph 6. K-Means Clustering

However in K-Means Clustering (K=3) all the cluster size are shown in Table 4 below. In first cluster there are 13 listing name airbnb, following that there are 3407 listing name airbnb in second cluster. Third cluster is the biggest cluster size because they have 4487 listing name airbnb in that cluster.

Table 4. Cluster Size K-Means (K=3)

Cluster	Size	Average Silhouette Width
1	13	0,57
2	3407	0,59
3	4487	0,48

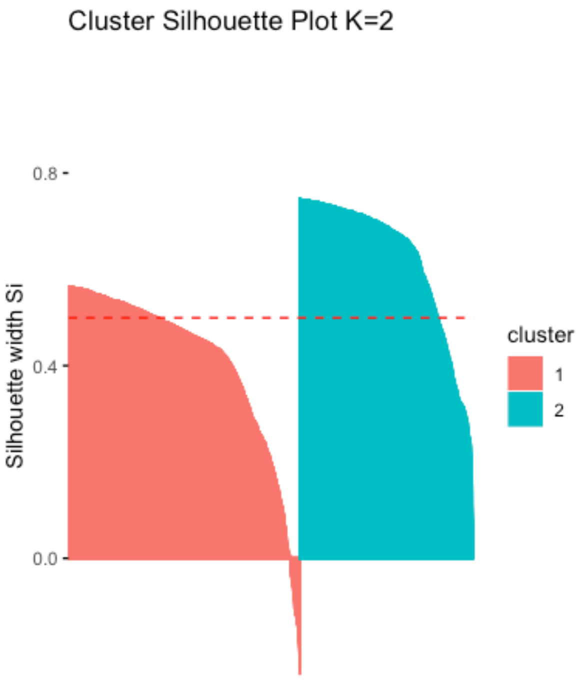
In Graph 7 below, The biggest average silhouette width is in cluster 2, it is about 0.59. Moreover, in cluster 1 and 3 the average silhouette width about 0.57 and 0.48 separately. That is the reason why in Graph 5, cluster 2 being the widest plot than the others two cluster. We can say also that in K-Means Clustering (K=3) that these clusters are imbalance because they have difference cluster size because there is the differences average silhouette width besides look at the number of cluster size.

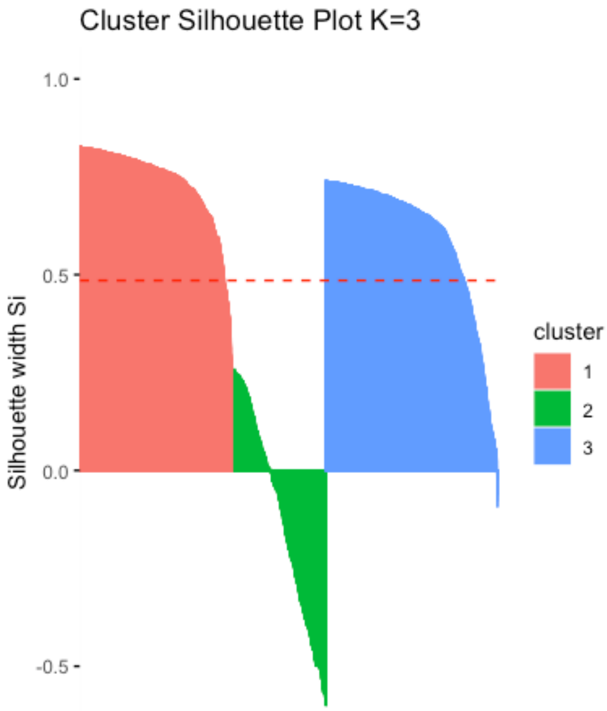


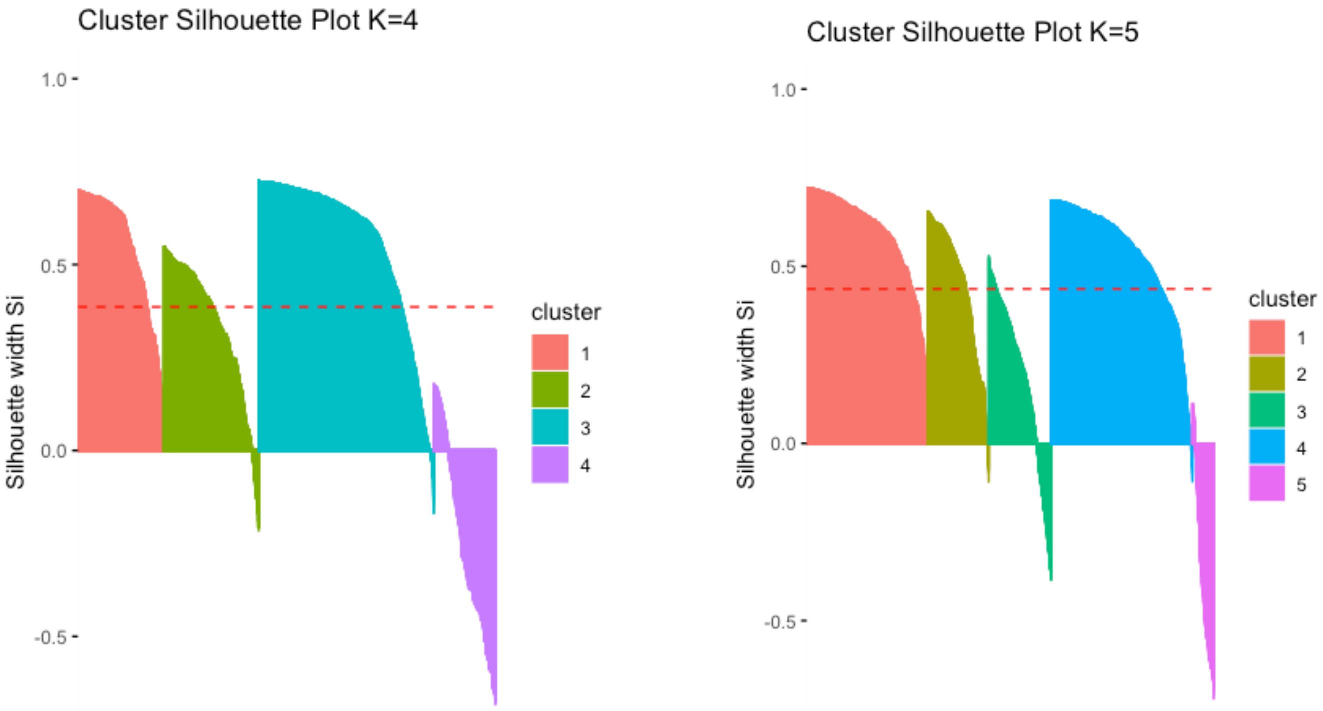
Graph 7. Optimum Cluster Numbers (K-Means, k=3)

Eventhough we got the optimum cluster number using K-Means method, but K-Means still have problem with inaccurately clusters like in Graph 7 above that in Cluster 3 some of observations got the wrong cluster and we can not see the cluster 1 because of the silhouette width is too small. Therefore in this paper, we tried to use another method which will be more fit with large database and robust to outliers by using CLARA.

With the same similar data in K-Means, it will be clustered start from 2 to 5 clusters with Clara. Before starting the analysis in clustering, I tried to choose the best cluster using silhouette plot first. To select the best cluster, the clusters should be above mean or close to +1 and the thickness of the silhouette plot should be quite similar. As we can see in Graph 8 below,

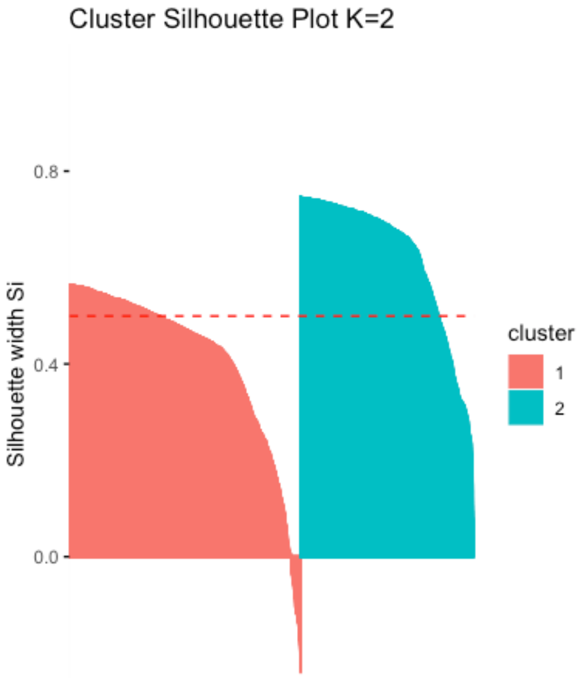






Graph 8. Clusters Silhouette Plot (CLARA)

that same with the K-Means method that in CLARA we still found that in all clusters which headed to -1 it means that in those clusters are false clusters. In cluster silhouette plot K=3, K=4 and K=5, all those clusters have clusters which not pass average silhouette (red line), on the other hand only in cluster 2, all cluster have clustered by passing the average (red line).



Graph 9. Optimum Cluster Silhouette Plot (CLARA)

Then based on silhouette plot, we choose Cluster 2 (See Graph 9 above) because they fulfil the requirement to select the best cluster. Besides all their clusters are headed to +1, they have similar width of cluster silhouette plot. This results also supported by average silhouette score that they have the higher results (Please see Table 5 below) for average Silhouette score among others clusters.

Table 5. Average Silhouette Score (CLARA)

Cluster	Average Silhouette Score
K=2	0,515
K=3	0,48
K=4	0,425
K=5	0,422

Relate with Graph 9, that Table 6 shows the average silhouette width for cluster 1 and cluster 2 about 0.41 and 0.62 successively. That is one of the reason why Cluster Silhouette Plot in K=2 almost have similar width and it can say that they have balance data in both clusters.

Table 6. Cluster Size CLARA (K=2)

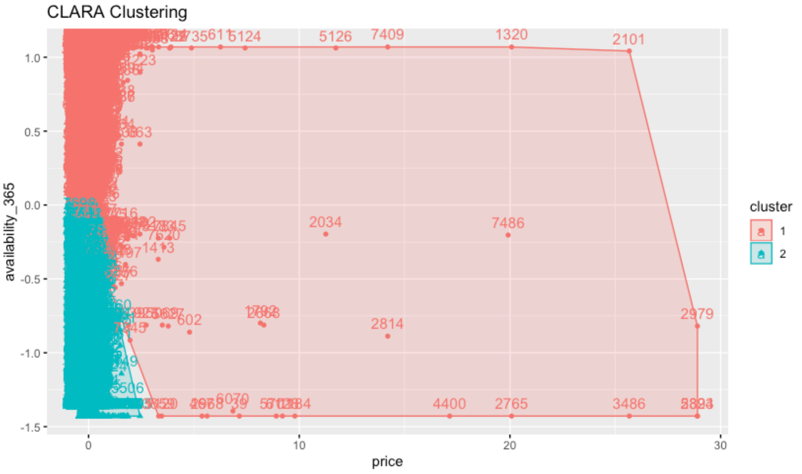
Cluster	Size	Average Silhouette Width
1	4519	0,41
2	3388	0,62

CLARA is similar with PAM, they are using medoids as their parameter to choose cluster. as we can see Table 7 below. In Cluster 1, they are using parameter or medoids in 140 SGD and availability about 338 days. On the other hand in cluster 2, their medoids are 90 SGD in Price's Variable) and 45 days in Availability's variable

Table 7. Medoids of CLARA

Cluster	Price	Availability_365
1	140	338
2	90	45

Graph 7 is clustering results between Price and Availability of rooms in 365 days using CLARA. Although in CLARA clustering result still found some outliers but it is better than clustering using K-Mean.



Graph 7. Optimum Cluster Numbers (CLARA, k=2)

From this results, I would say that to people who are interested have business in housing specifically using airbnb, you can check in my table 8 below in which locations your competitor now and you start your business in housing likes rented room/apartment/ shared room in airbnb which may have high opportunity based on this clustering methods.

Table 8. Clustering Results Using CLARA (K=2)

No	Name	Neighbourhood Group	Neighbourhood	Price	Availability	New Cluster
1	COZICOMFORT LONG TERM STAY ROOM 2	North Region	Woodlands	83	365	1
2	Pleasant Room along Bukit Timah	Central Region	Bukit Timah	81	365	1
3	COZICOMFORT	North Region	Woodlands	69	365	1
4	Ensuite Room (Room 1 & 2) near EXPO	East Region	Tampines	206	353	1
5	B&B Room 1 near Airport & EXPO	East Region	Tampines	94	355	1
6	Room 2-near Airport & EXPO	East Region	Tampines	104	346	1
7	3rd level Jumbo room 5 near EXPO	East Region	Tampines	208	172	2

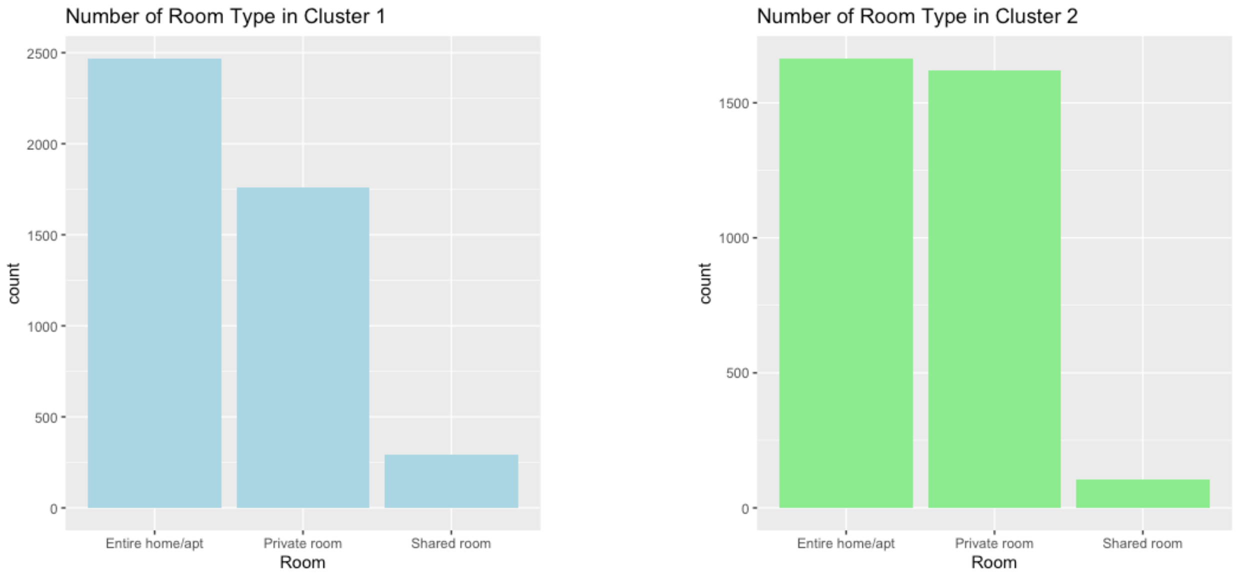
8	Long stay at The Breezy East "Leopard"	East Region	Bedok	50	59	2
9	Long stay at The Breezy East "Plumeria"	East Region	Bedok	54	133	2
10	Long stay at The Breezy East "Red Palm"	East Region	Bedok	42	147	2
…	…	…	…	…	…	…
7907	Amazing room with private bathroom walk to Orchard	Central Region	River Valley	65	365	1

Also based on statistics descriptive in Table 9, it shows that range of rented room/apartment and share room in cluster 1 is around 15-10.000 SGD per days with minimum 5 days and maximum 365 days to rent. However, in Cluster the price is quite cheaper than in cluster 1 because the maximum price in cluster only 1.000 SGD but their availability around 1 until 200 days only.

Table 9. Descriptive Statistics in New Cluster

Cluster	Price (SGD)		Availability (Days)	
	Minimum	Maximum	Minimum	Maximum
1	15	10.000	5	365
2	14	1.000	1	200

Also in these two clusters in Graph 8 Below shows the difference number of Entire home/Apartment/ Private Room/Shared Room. All type of rooms in Cluster 1 have higher number of room than in cluster 2. Nevertheless in both of Clusters, Entire home/apartment being the highest number among other categories. In cluster 1, Entire Home/apt have 2468 entire home/apt but in cluster 2, they only have 1664 rooms. However in Private Room, in Cluster 1 there are 1760 private rooms on the other hand in cluster 2 they only have 1621 private rooms. Last for shared room category, in Cluster 1 area, they have higher number of rooms about 291 shared rooms but then in cluster 2 they only have 103 shared rooms.



Graph 8. Type of Room in Cluster 1 and Cluster 2

4. Conclusions

After all the analysis data, we can conclude that:

1. K-Means was very sensitive with outliers and CLARA perform better to this airbnb data. In K-Means got 3 Clusters as best optimum clusters using silhouette as their parameter.
2. Difference from K-Means, in CLARA we got the best optimum cluster, in 2 cluster. Also in CLARA, they don't have many outliers as in K-Means. We got the results that in cluster 1 mostly their price for renting private room/apartment/shared room is higher than in cluster 2.
3. Based on CLARA results, in airbnb data we can say that in last 10 years most of airbnb's user prefer to stay in entire home/apartment than the other two type of rooms and my suggestion to new occupants who want to join in this home-sharing business like this, perhaps you can provide entire home/apartment to be rented also. However in cluster 1, the price in all type of rooms are in between 15 and 10,000 SGD but in Cluster 2, the price are between 14 and 1,000 SGD.

References

[1] https://www.stb.gov.sg/content/stb/en/statistics-and-market-insights/tourism-statistics/international-visitorarrivals.html

[2] https://www.stb.gov.sg/content/stb/en/statistics-and-market-insights/tourism-statistics/hotel-statistics.html

[3] https://www.channelnewsasia.com/news/singapore/visitor-arrivals-to-singapore-rise-6-2-to-hit-new-high-in-2018-11237564

[4] https://muchneeded.com/airbnb-statistics/

[5] http://insideairbnb.com/get-the-data.html.