Data Mining Report

Nandini Agrawal,Sai Khurrana,Vir Jahangiani

Scrapping

We have scrapped our data from booking.com using selenium python. The script scrapped the first 20 hotels for each of the city with 100 reviews for each hotel in a city. The data for each city is then stored in two formats: csv(for nltk and mapbox purposes) and arff(for WEKA), so overall 6 files generated(2 for each city). Scraping data from websites like booking.com gave us the experience of working with real world data and taught us how we can deal with issues like data cleaning and integration, an important skill while solving actual problems and working with real world data .

The schema(columns) of the data with an example is as follows:

City	Hotel Name	Overall Rating	Comment	Comment Rating
Singapore	Resort Shangri-La's Rasa Sentosa Resort & Spa (SG Clean)	8.6	Exceptional	10

Visualisation Using Matplotlib-Python

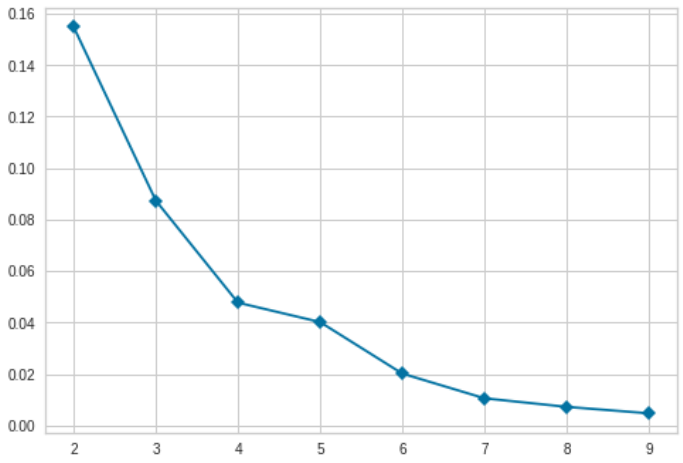
We used matplotlib to cluster hotels using k-means clustering with similar overall ratings for each city(also tried WEKA but quality of image was poor and compatibility issues with MacOS). Initially we thought of doing the visualisation on the combined data for all cities, but that would not be very insightful as you would want to know hotels that are similar(rating wise) for each city to decide your stay. Comparing hotels with similar ratings in different cities would not help with that decision.

For each city, we tried different things such as different initial centroids(randomised), different number of clusters, different seeds, etc. We chose to go with k=4 for Bangkok and Singapore and k=5 for Kuala Lumpur based on the elbow test using the distortion metric[1]. Please find the visualisations attached below for the optimal k(number of clusters) values. (Graphs for other trial k values can be seen in the appendix section as well as the code/demo)

Singapore:

Elbow Test Graph:

x axis -> indicates value of k(clusters)
y axis -> Distortion Score
Elbow at: 4

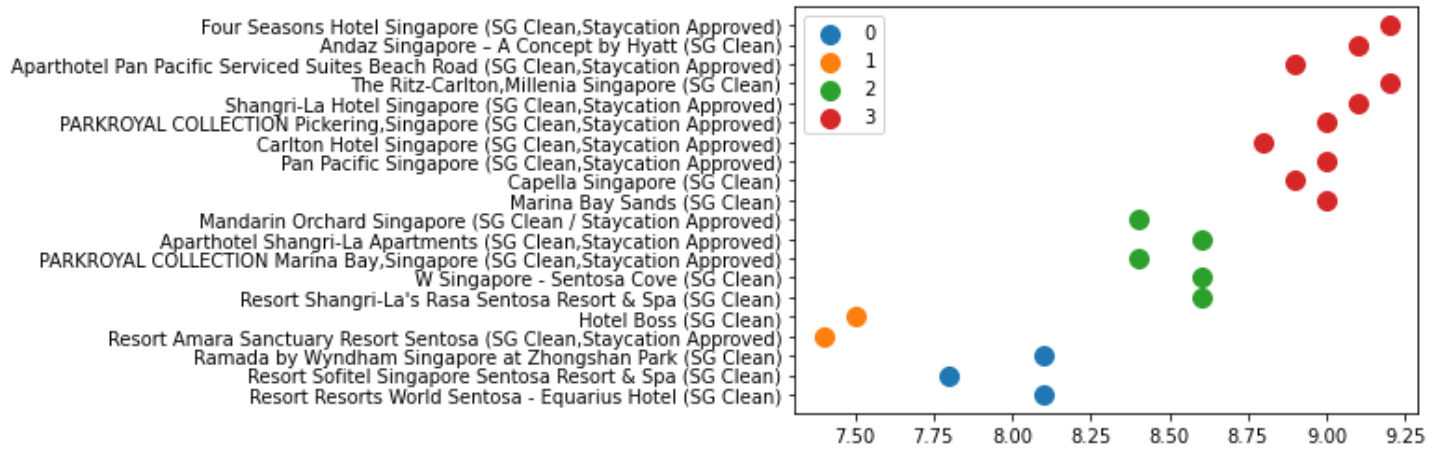


It makes sense to have four clusters here because having more would make each cluster very sparse giving us less hotels which are similar, and having smaller k would not give very accurate results when it comes to similarity.
[1]Distortion metric : Average of the squared distances(euclidean) from the cluster centers of the respective clusters.

k=4

Final cluster centroids:

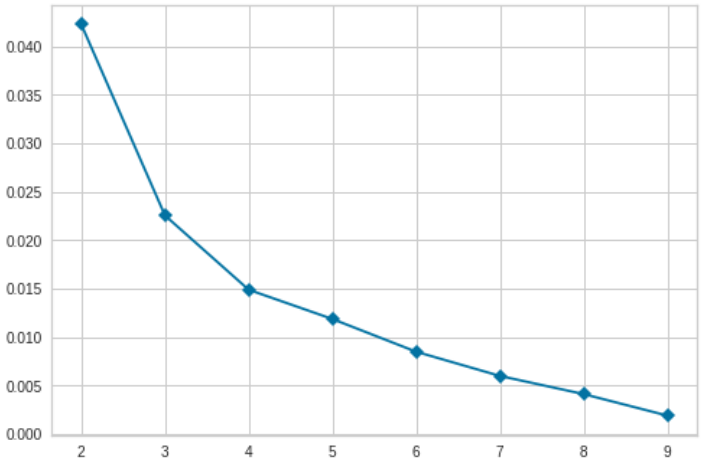
	Cluster#			
Attribute	0	1	2	3
hotel_rating	8	7.45	8.52	9.02



<u>Bangkok:</u>

Elbow Test Graph:

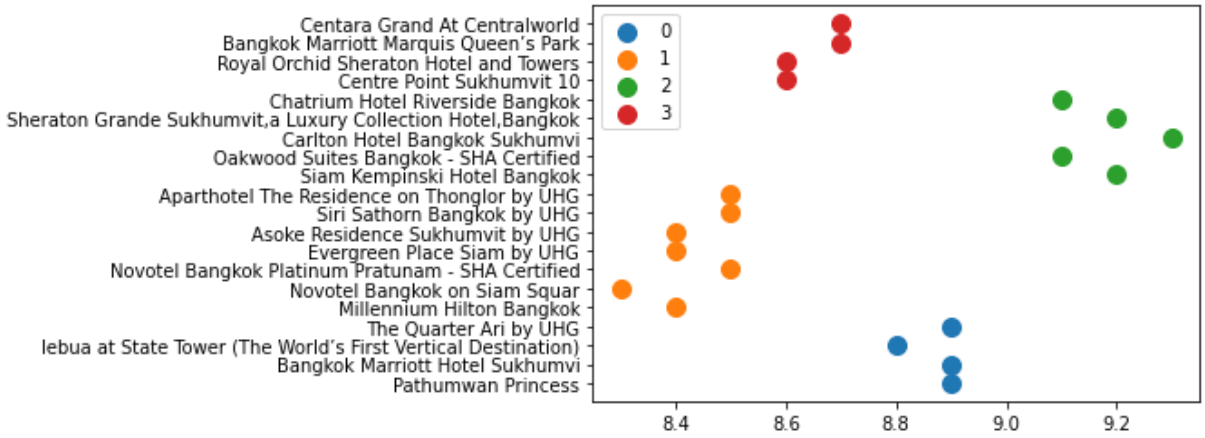
x axis -> indicates value of k(clusters)
y axis -> Distortion Score
Elbow at: 4
(Similar reason as above: Sufficient hotels for each cluster)



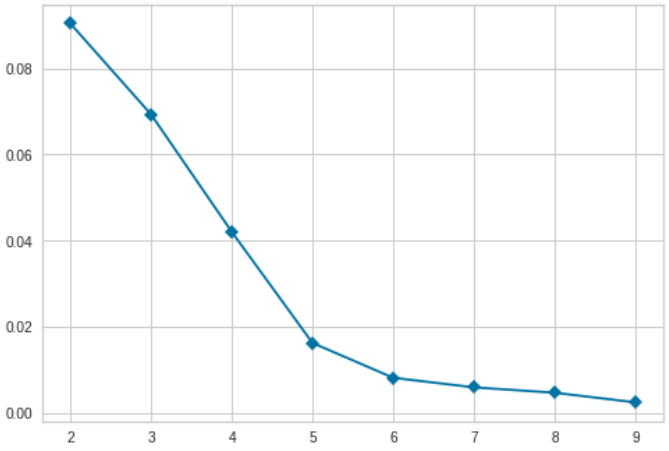
k=4

Final cluster centroids:

	Cluster#			
Attribute	0	1	2	3
hotel_rating	8.86	8.43	9.18	8.65



Kuala Lumpur:



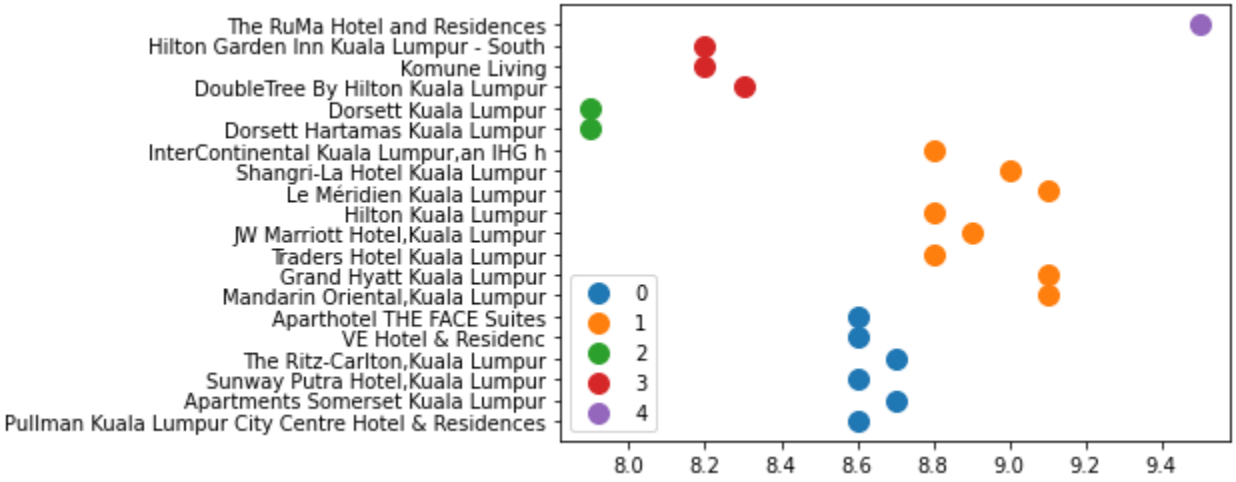
Elbow Test Graph:

x axis -> indicates value of k(clusters)
y axis -> Distortion Score
Elbow at: 5
Here we chose to have 5 clusters because it has a wider variety of overall ratings as compared to the other two cities, so lesser value of k would cause extremely high/low rated hotels to club with its nearest ratings, resulting into less accurate clustering.

k=5

Final cluster centroids:

	Cluster#				
Attribute	0	1	2	3	4
hotel_rating	8.63	8.95	7.9	8.23	9.5



Could have gone with k=4 here too but would end up clustering the extremely highly rated The RuMa hotel with its cluster 1 peers although it might have better services(as very highly rated)

Sentiment Analysis using NLTK

Library used for sentiment analysis: nltk.sentiment.SentimentIntensityAnalyzer

Library used for finding correlation: scipy.stats.pearsonr

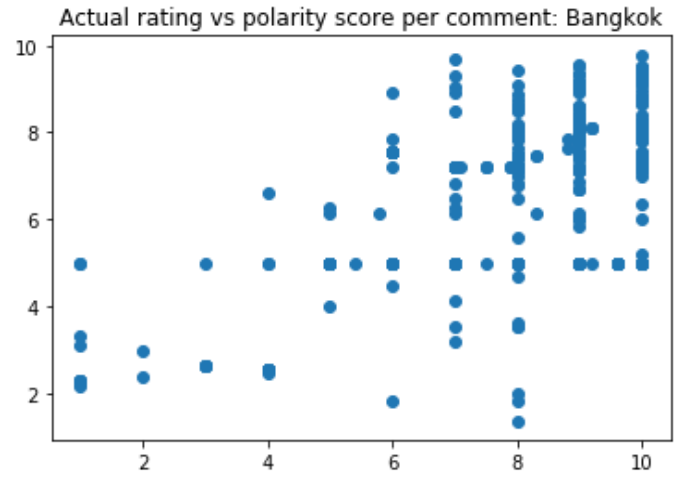
Singapore	**Kuala Lumpur**	**Bangkok**
Hotel name: Resort Shangri-La's Rasa Sentosa Resort & Spa (SG Clean) Hotel rating: 8.6 Average rating of comments: 8.457 Average rating of comment polarity: 6.9503949999999985	Hotel name: Mandarin Oriental,Kuala Lumpur Hotel rating: 9.1 Average rating of comments: 8.83 Average rating of comment polarity: 6.549734999999999	Hotel name: Millennium Hilton Bangkok Hotel rating: 8.4 Average rating of comments: 8.75 Average rating of comment polarity: 6.32794
Hotel name: Marina Bay Sands (SG Clean) Hotel rating: 9.0 Average rating of comments: 8.84 Average rating of comment polarity: 6.334004999999999	Hotel name: Grand Hyatt Kuala Lumpur Hotel rating: 9.1 Average rating of comments: 8.82 Average rating of comment polarity: 6.492504999999999	Hotel name: Centre Point Sukhumvit 10 Hotel rating: 8.6 Average rating of comments: 8.77 Average rating of comment polarity: 6.696799999999999
Hotel name: Resort Amara Sanctuary Resort Sentosa (SG Clean,Staycation Approved) Hotel rating: 7.4 Average rating of comments: 7.557 Average rating of comment polarity: 6.16014	Hotel name: Traders Hotel Kuala Lumpur Hotel rating: 8.8 Average rating of comments: 8.45 Average rating of comment polarity: 6.8199999999999985	Hotel name: Royal Orchid Sheraton Hotel and Towers Hotel rating: 8.6 Average rating of comments: 8.743 Average rating of comment polarity: 6.446574999999999
Hotel name: Capella Singapore (SG Clean) Hotel rating: 8.9 Average rating of comments: 9.01 Average rating of comment polarity: 6.704604999999999	Hotel name: The RuMa Hotel and Residences Hotel rating: 9.5 Average rating of comments: 9.126 Average rating of comment polarity: 6.586564999999999	Hotel name: Pathumwan Princess Hotel rating: 8.9 Average rating of comments: 9.048 Average rating of comment polarity: 6.35403
Hotel name: Pan Pacific Singapore (SG Clean,Staycation Approved) Hotel rating: 9.0 Average rating of comments: 8.687000000000001 Average rating of comment polarity: 6.6700149999999985	Hotel name: Pullman Kuala Lumpur City Centre Hotel & Residences Hotel rating: 8.6 Average rating of comments: 8.022 Average rating of comment polarity: 6.226479999999999	Hotel name: Bangkok Marriott Hotel Sukhumvi Hotel rating: 8.9 Average rating of comments: 8.963 Average rating of comment polarity: 6.175284999999999
Hotel name: Carlton Hotel Singapore (SG Clean,Staycation Approved)	Hotel name: DoubleTree By Hilton Kuala Lumpur	Hotel name: Siam Kempinski Hotel Bangkok

Hotel rating: 8.8 Average rating of comments: 8.62 Average rating of comment polarity: 6.947375000000001	Hotel rating: 8.3 Average rating of comments: 8.48 Average rating of comment polarity: 6.605535	Hotel rating: 9.2 Average rating of comments: 9.466000000000001 Average rating of comment polarity: 6.014594999999999
Hotel name: PARKROYAL COLLECTION Pickering,Singapore (SG Clean,Staycation Approved) Hotel rating: 9.0 Average rating of comments: 7.98 Average rating of comment polarity: 6.675459999999999	Hotel name: JW Marriott Hotel,Kuala Lumpur Hotel rating: 8.9 Average rating of comments: 8.36 Average rating of comment polarity: 6.310169999999999	Hotel name: Oakwood Suites Bangkok - SHA Certified Hotel rating: 9.1 Average rating of comments: 9.22 Average rating of comment polarity: 6.334214999999998
Hotel name: W Singapore - Sentosa Cove (SG Clean) Hotel rating: 8.6 Average rating of comments: 8.615 Average rating of comment polarity: 6.94281	Hotel name: Apartments Somerset Kuala Lumpur Hotel rating: 8.7 Average rating of comments: 8.419 Average rating of comment polarity: 6.68415	Hotel name: Novotel Bangkok on Siam Squar Hotel rating: 8.3 Average rating of comments: 8.332 Average rating of comment polarity: 6.853514999999999
Hotel name: Hotel Boss (SG Clean) Hotel rating: 7.5 Average rating of comments: 7.55 Average rating of comment polarity: 6.2710550000000005	Hotel name: Komune Living Hotel rating: 8.2 Average rating of comments: 6.75 Average rating of comment polarity: 5.8973	Hotel name: Carlton Hotel Bangkok Sukhumvi Hotel rating: 9.3 Average rating of comments: 9.25 Average rating of comment polarity: 6.221075
Hotel name: Resort Resorts World Sentosa - Equarius Hotel (SG Clean) Hotel rating: 8.1 Average rating of comments: 7.686 Average rating of comment polarity: 6.419564999999999	Hotel name: Hilton Kuala Lumpur Hotel rating: 8.8 Average rating of comments: 8.4 Average rating of comment polarity: 6.705164999999998	Hotel name: lebua at State Tower (The World’s First Vertical Destination) Hotel rating: 8.8 Average rating of comments: 9.01 Average rating of comment polarity: 6.23349
Hotel name: Shangri-La Hotel Singapore (SG Clean,Staycation Approved) Hotel rating: 9.1 Average rating of comments: 8.75 Average rating of comment polarity: 6.648594999999999	Hotel name: Dorsett Hartamas Kuala Lumpur Hotel rating: 7.9 Average rating of comments: 7.4510000000000005 Average rating of comment polarity: 6.395115	Hotel name: Sheraton Grande Sukhumvit,a Luxury Collection Hotel,Bangkok Hotel rating: 9.2 Average rating of comments: 9.315 Average rating of comment polarity: 6.285989999999999
Hotel name: PARKROYAL COLLECTION Marina Bay,Singapore (SG Clean,Staycation Approved)	Hotel name: Dorsett Kuala Lumpur Hotel rating: 7.9 Average rating of comments: 6.68	Hotel name: Novotel Bangkok Platinum Pratunam - SHA Certified Hotel rating: 8.5

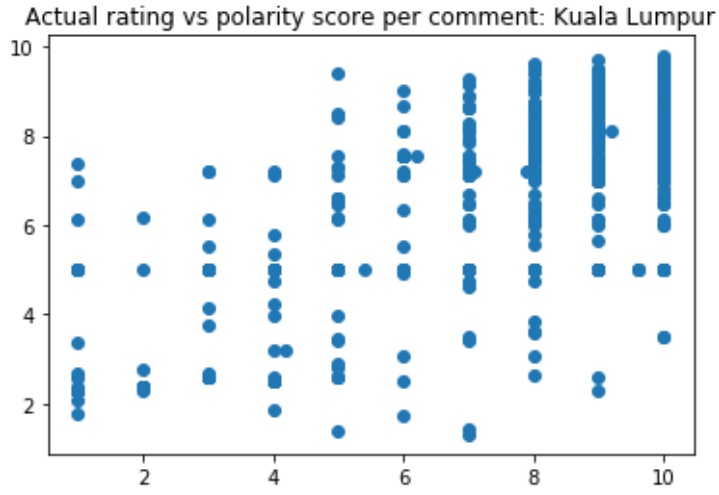
Hotel rating: 8.4 Average rating of comments: 8.73 Average rating of comment polarity: 6.71432	Average rating of comment polarity: 6.085875	Average rating of comments: 8.794 Average rating of comment polarity: 6.665109999999999
Hotel name: Resort Sofitel Singapore Sentosa Resort & Spa (SG Clean) Hotel rating: 7.8 Average rating of comments: 7.179000000000001 Average rating of comment polarity: 6.243759999999998	Hotel name: Le Méridien Kuala Lumpur Hotel rating: 9.1 Average rating of comments: 8.756 Average rating of comment polarity: 6.659414999999998	Hotel name: The Quarter Ari by UHG Hotel rating: 8.9 Average rating of comments: 8.91 Average rating of comment polarity: 6.2451599999999985
Hotel name: The Ritz-Carlton,Millenia Singapore (SG Clean) Hotel rating: 9.2 Average rating of comments: 9.05 Average rating of comment polarity: 6.668215	Hotel name: Hilton Garden Inn Kuala Lumpur - South Hotel rating: 8.2 Average rating of comments: 8.036 Average rating of comment polarity: 6.47057	Hotel name: Bangkok Marriott Marquis Queen's Park Hotel rating: 8.7 Average rating of comments: 8.777999999999999 Average rating of comment polarity: 6.465995
Hotel name: Aparthotel Shangri-La Apartments (SG Clean,Staycation Approved) Hotel rating: 8.6 Average rating of comments: 8.404000000000002 Average rating of comment polarity: 6.689309999999998	Hotel name: Sunway Putra Hotel,Kuala Lumpur Hotel rating: 8.6 Average rating of comments: 8.66 Average rating of comment polarity: 6.426984999999999	Hotel name: Evergreen Place Siam by UHG Hotel rating: 8.4 Average rating of comments: 8.75 Average rating of comment polarity: 6.4795549999999995
Hotel name: Aparthotel Pan Pacific Serviced Suites Beach Road (SG Clean,Staycation Approved) Hotel rating: 8.9 Average rating of comments: 8.822000000000001 Average rating of comment polarity: 6.45836	Hotel name: The Ritz-Carlton,Kuala Lumpur Hotel rating: 8.7 Average rating of comments: 8.155999999999999 Average rating of comment polarity: 6.302794999999999	Hotel name: Chatrium Hotel Riverside Bangkok Hotel rating: 9.1 Average rating of comments: 9.04 Average rating of comment polarity: 6.495234999999999
Hotel name: Andaz Singapore – A Concept by Hyatt (SG Clean) Hotel rating: 9.1 Average rating of comments: 8.64 Average rating of comment polarity: 6.745189999999999	Hotel name: Shangri-La Hotel Kuala Lumpur Hotel rating: 9.0 Average rating of comments: 8.78 Average rating of comment polarity: 6.840585	Hotel name: Asoke Residence Sukhumvit by UHG Hotel rating: 8.4 Average rating of comments: 8.189 Average rating of comment polarity: 6.4367849999999995

Hotel name: Ramada by Wyndham Singapore at Zhongshan Park (SG Clean) Hotel rating: 8.1 Average rating of comments: 8.47 Average rating of comment polarity: 6.877884999999999	Hotel name: InterContinental Kuala Lumpur,an IHG h Hotel rating: 8.8 Average rating of comments: 8.682 Average rating of comment polarity: 6.3205599999999995	Hotel name: Siri Sathorn Bangkok by UHG Hotel rating: 8.5 Average rating of comments: 8.6 Average rating of comment polarity: 6.49151
Hotel name: Mandarin Orchard Singapore (SG Clean / Staycation Approved) Hotel rating: 8.4 Average rating of comments: 8.779 Average rating of comment polarity: 6.622359999999999	Hotel name: VE Hotel & Residenc Hotel rating: 8.6 Average rating of comments: 8.64 Average rating of comment polarity: 6.655615000000001	Hotel name: Centara Grand At Centralworld Hotel rating: 8.7 Average rating of comments: 8.998 Average rating of comment polarity: 6.31476
Hotel name: Four Seasons Hotel Singapore (SG Clean,Staycation Approved) Hotel rating: 9.2 Average rating of comments: 9.256 Average rating of comment polarity: 7.187274999999999	Hotel name: Aparthotel THE FACE Suites Hotel rating: 8.6 Average rating of comments: 8.3 Average rating of comment polarity: 6.49598	Hotel name: Aparthotel The Residence on Thonglor by UHG Hotel rating: 8.5 Average rating of comments: 8.287 Average rating of comment polarity: 6.504879999999999
Pearson's Correlation Coeff : 57.439%	**Pearson's Correlation Coeff : 50.218%**	**Pearson's Correlation Coeff : 67.065%**

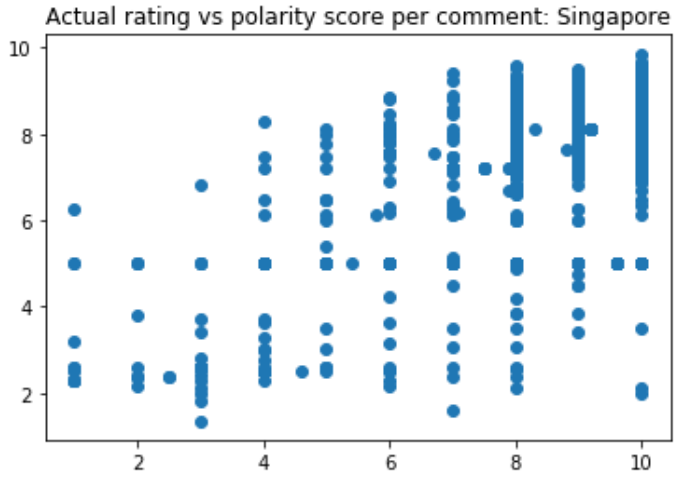
The following plots show the scatter plot of comment rating vs comment scaled polarity score(all hotels combined for a Country):




```
Average rating of al comments: 8.86065
Average polarity score of al comments: 6.402125
```



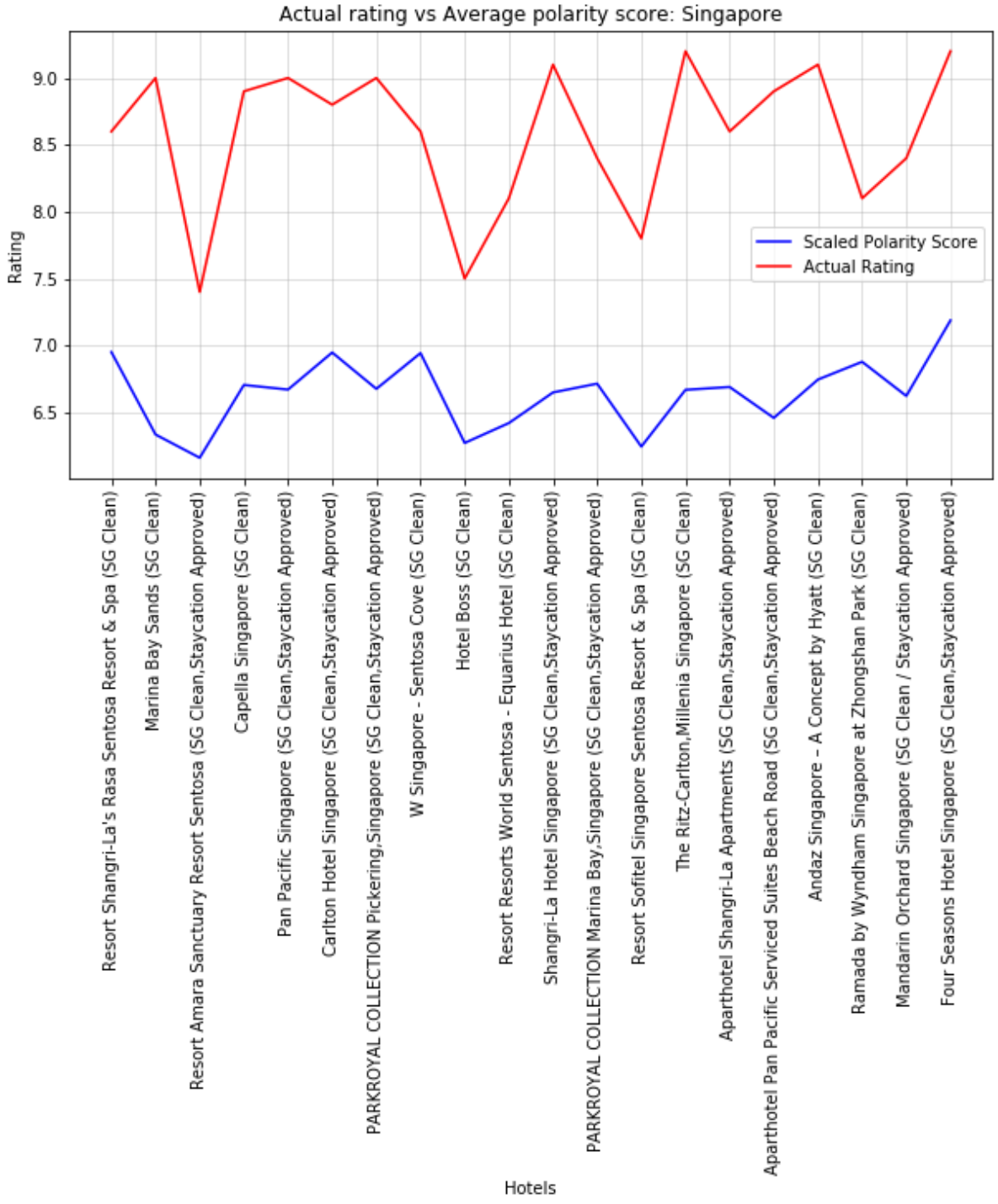

```
Average rating of al comments: 8.2899
Average polarity score of al comments: 6.476555
```



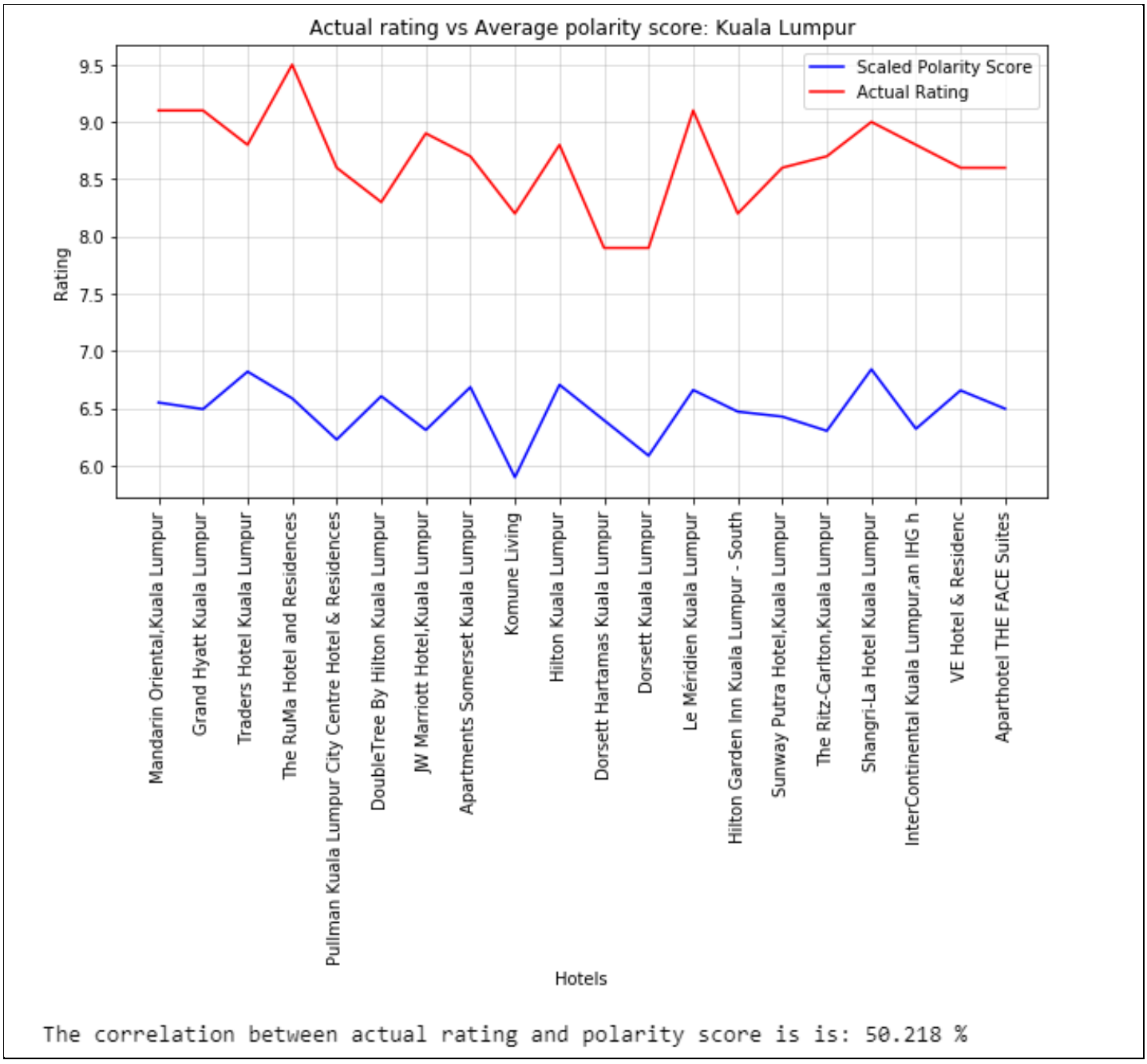

```
Average rating of al comments: 8.4541
Average polarity score of al comments: 6.64653475
```

These graphs show us how the polarity score prediction compares to the actual rating of each comment. If the polarity score perfectly predicted the rating given to a comment then the scatter plot graphs would look like a straight line at a 45 degree angle to the x axis. That means that if a comment rating is a 9/10 then even the polarity score should be a 9/10. But in these graphs we can observe that there are outliers, there are multiple points where the polarity score does not match the actual score for that specific comment.

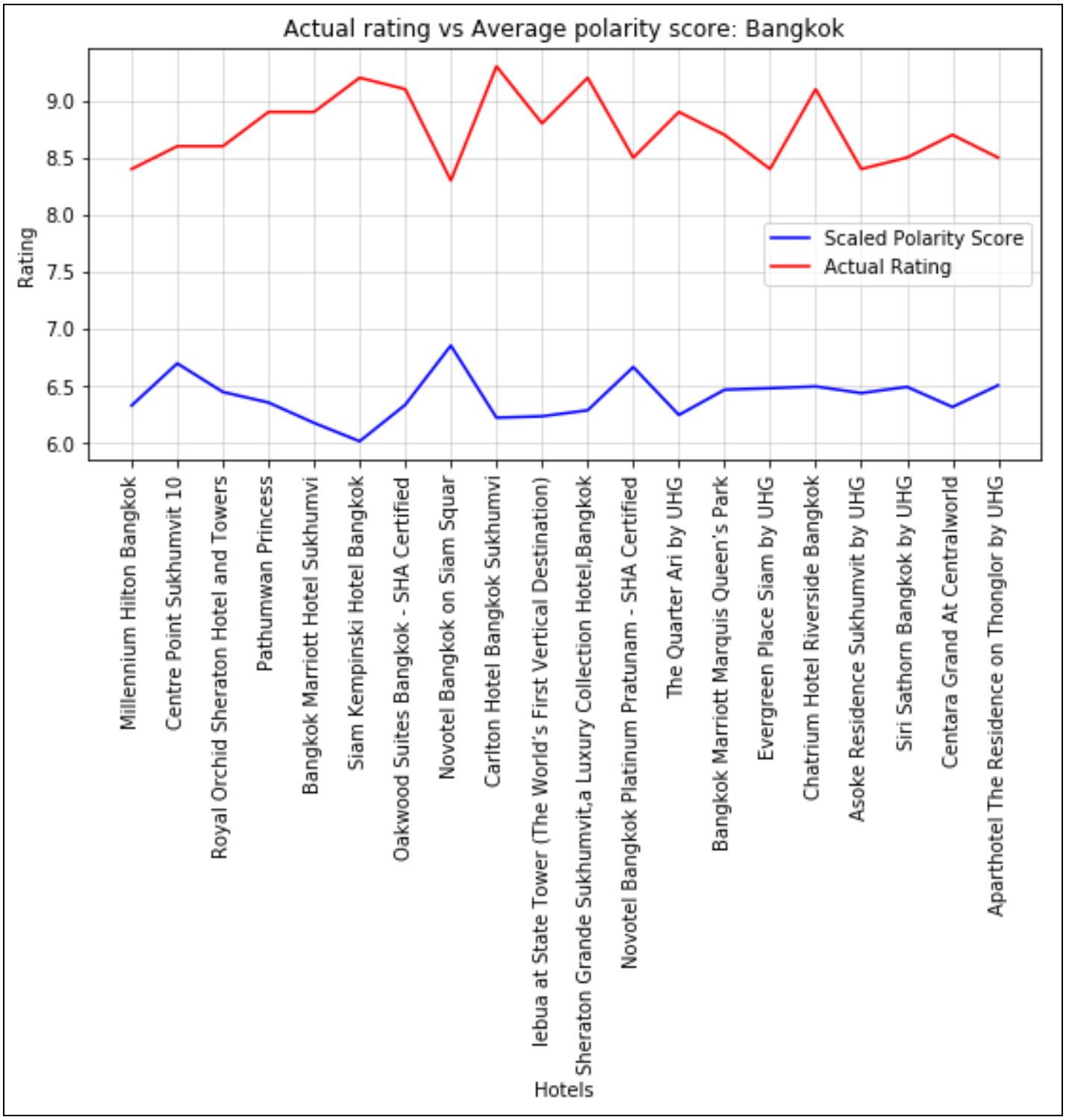
We also noticed that the average of the polarity score for all comments(overall and each hotel wise too) is lower than the average for the actual score of all comments. This shows us that there seems to be a slight error in the prediction of the polarity score. It is giving a sentence a lower score than it should be getting(based on how the compound score in nltk sentiment analyzer is calculated). One such example is the comment “Exceptional”. This comment has a 10/10 rating but the nltk semantics analyser considers this statement to be neutral so when its compound score is scaled, we get a polarity score of 5/10 for this. This behavior(lower avg) can be seen in the graphs below.



The correlation between actual rating and polarity score is is: 57.439 %



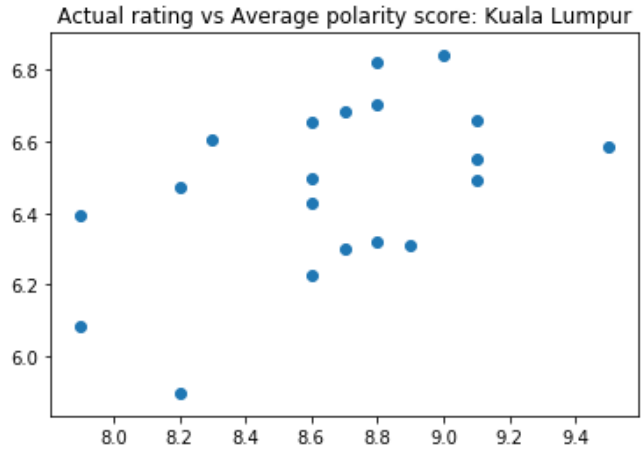
The correlation between actual rating and polarity score is is: 50.218 %

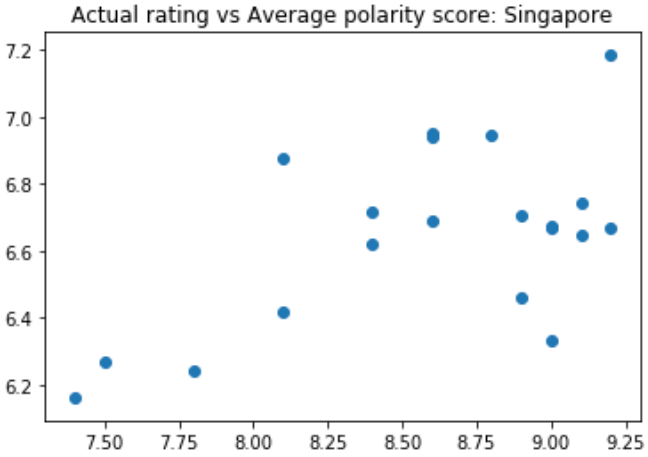


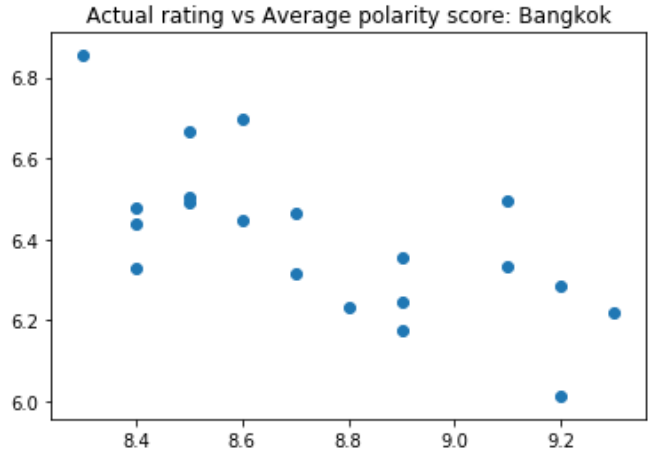
The correlation between actual rating and polarity score is is: 67.065 %

In each of these graphs, the graph of the scaled polarity is similar to the actual rating graph (the shape), it's just at a lower y axis space due to the average issue explained above.

In the following graphs we can observe the average polarity score per hotel vs actual overall rating given to that hotel. We can again observe that they aren't exactly accurate for all hotels. If the polarity scores were accurate then the graph would look like a straight line at a 45 degree angle with the x axis.







We can see a few anomalies here in the graph, eg for Bangkok, a hotel with a low overall rating has a very high average polarity score, whereas a hotel with high overall rating has the lowest avg polarity score.

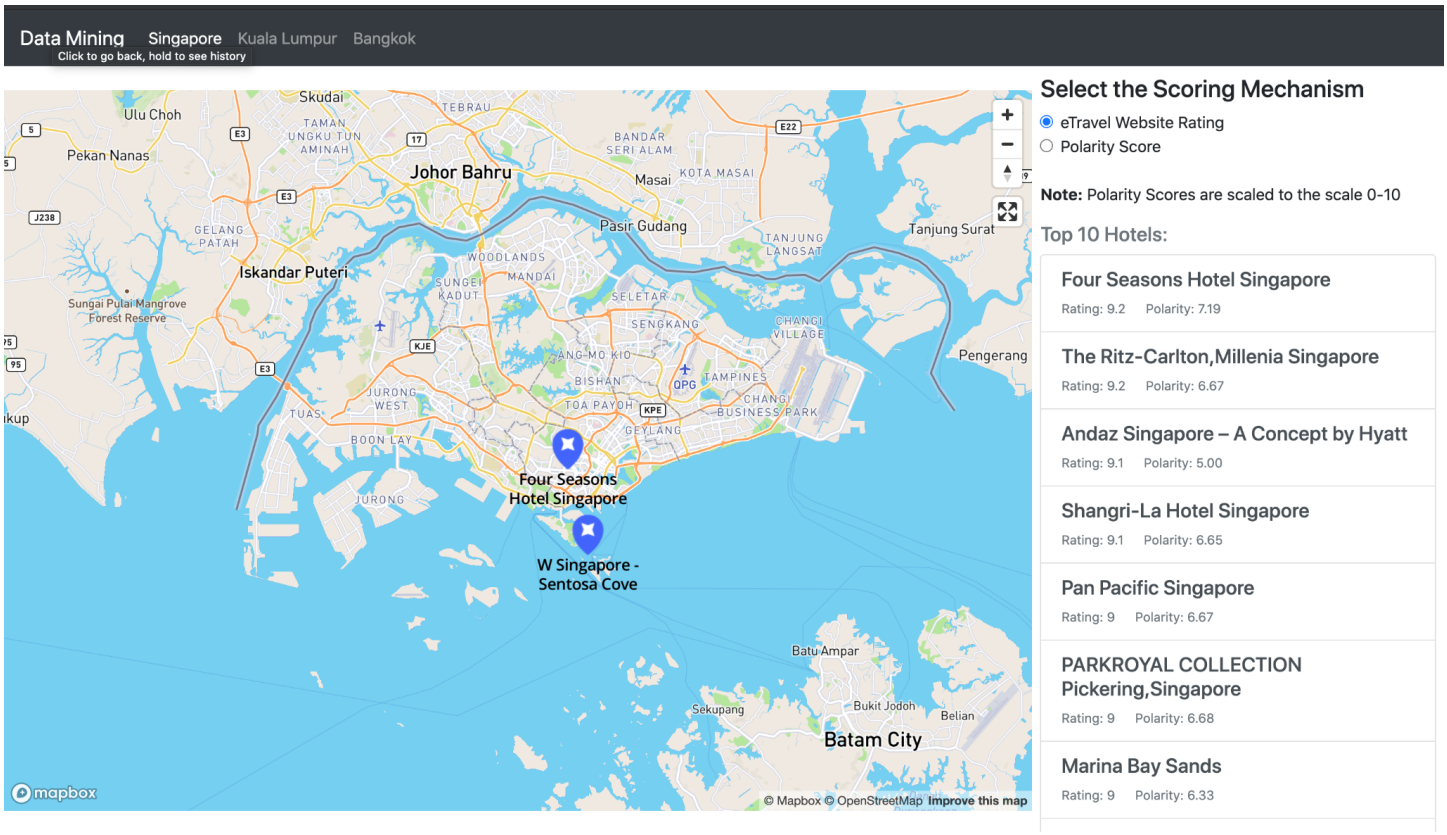
Geospatial Visualisation Using Mapbox

Visualisation of the top 10 hotels (overall rating and polarity wise respectively) can be found here: https://saikhurana98.github.io/DMDW-DecisionSupport/#

Tools used for developing this :
jQuery
Mapbox
HTML CSS
Bootstrap
Javascript WebAPI

The hotels and their overall ratings were scrapped and stored as a csv file by the scrapper, the polarity scores were calculated and added to the data while mapping these hotels. Also Mapbox requires coordinates(lat,long) for mapping the hotels so we had a separate json file for each country with all the hotels and their respective coordinates in the backend.

This is how the portal looks like:



Can toggle between countries, and chose the scoring mechanism to see the top 10 hotels in that country for that specific metric.

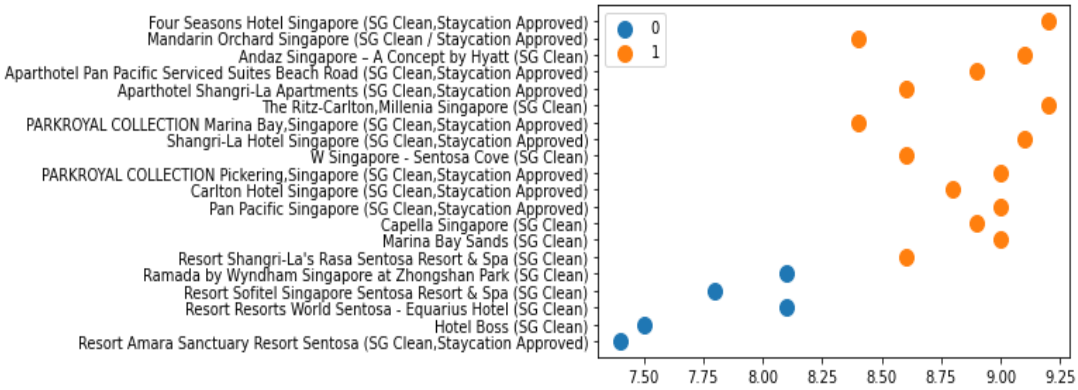
Appendix

Some graph for other values of k(this one is for Singapore, similar for other cities too)

k=2

Final cluster centroids:

	Cluster #	
Attribute	0	1
hotel_rating	7.78	8.85



k=3

Final cluster centroids:

	Cluster#		
Attribute	0	1	2
hotel_rating	7.78	9.04	8.5