International Journal of Application on Sciences, Technology and Engineering (IJASTE)
Volume 1, Issue 1, 2023. ISSN 2987-2499

K-Means Algorithm and Clustering Technique for A Recommender System

Sinnappan Glaret Shirley [1 a)] and Kodukula Subrahmanyam, Dusanapudi.Susrija, Palempati.Akhila [2,3,4 b)]

Author Affiliations
[1 a)] *Department of Information and Communication Teachnology*
Main Campus,Tengku Abdul Rahman University College
Jalan Genting Kelang, 53300 Kuala Lumpur, Wilayah Persekutuan Kuala Lumpur
[2, 3, 4 b)] *Department of Computer Science and Engineering*
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Guntur, India

Author Emails
[a)] glaret@tarc.edu.my
[b)] smkodukula@kluniversity.in
[c)] susrijadusanapudi@gmail.com
[d)] palempatiakhila@gmail.com

Submitted: November-December 2022, Revised: January 2023, Accepted: February 25, 2023

Abstract. Most of netizens are fond of traveling. Though they are interested to travel, majority of them are confused about where to go. Whenever the netizens planed for a travel, they spent hours searching for an interesting place that matches their point of interest. Therefore, there is a need for a recommender system which can recommend several interesting travel venues based on their preferences. Hence, information regarding different users prefers different travel locations are required. In this paper, Google review is used as a reference to divide the users into clusters of similar interests. The data has been divided into six scenarios which consist of the original data, scaled data, Principal Component Analysis (PCA) data, PCA scaled data, MinMax scaled data and Robust scaled data. The distance metric technique is performed in all these scenarios. Minkowski distance metric is used to measure the distance between the data points in the dataset and the elbow method to these scenarios to know the inertia values. This has revealed that the PCA with scaled data has lower distance and inertia value. The K-Means clustering is performed on this PCA-scaled data which divided the data into four different clusters. The results revealed are imperative to build any recommender system based on the users' preference.

INTRODUCTION

People often choose to travel in their free time to get out of the stress from day-to-day activities. Each person has their interests and choices. Not everyone likes everything. As part of their relaxation, some people like to visit holistic destinations while some people like to admire the beauty of nature and there also exits people who love to have tasty food (2)(3). Here comes the problem of choosing the holiday spot and plan. People feel hard deciding where they should visit. (1). They struggle to find places that match their lifestyle. It takes a lot of time finding whether such places exist and deciding among them. Spending too much time on this activity is a waste of time (3)(6).

Artificial Intelligence is a significant branch of Computer Science and Engineering that makes human work simple (4)(5). In Artificial Intelligence, we train a machine thus that it does all the average works that a man does. It helps each sector to enhance productiveness and proficiency. In this case, we need a recommendation system for our travel using Artificial Intelligence. It saves a lot of time of ours. Here is an idea to build a

recommendation system that can recommend several interesting places based on user preferences and choices (5) (13).

But before that, we should have accurate data to train the system. The main aim of this paper is to make an accurate data model from the raw dataset which helps the recommendation system. To make this we use Google review data of each user. Based on the rating that users give to each place, we categorize the users into different groups. Based on this grouped data, the system is trained and can suggest places based on user preferences (14)(15).

In Google maps, people can give a rating from 0 to 5 on a particular place and they can also comment about the place. We import that data and make clusters of users. To make clustering of data, we use the machine learning algorithm K- Means (13)(15). For the implementation of K-Means, we scale the original data for efficiency, and then using Minkowski distance we determine which data frame can be used for best accuracy. Using the elbow method, we find the number of clusters. This results in clustered data which helps to build a recommender system (15)(16).

RELATED WORK

This paper is totally related to travelers and the travel destinations. But travelers are very confused about their holiday destination in most of the time. They spent much of the time selecting a destination. Even though if they found any place, there is no guarantee that the place will match their point of interest. Therefore, there is a need can solve this problem.

Travel Research Review

The field of the travel industry had seen a drastic change in recent years with outstanding development in technology and research. There are many journals and publications in this field (16)(17)(18)(30). There was a huge increase in the records of the Encyclopedia of Tourism. Before 1980, there are fewer than 10 papers and now there around 290 papers where around 150 are distributed in English (30). In tourism field, research studies have left similar information to survey and research and to decide if they are getting more standard data (33). In the current years, the field of the travel industry has been identified as autonomous scholarly classification in the web sciences stream and reflecting the technological progression (31)(32). Henceforth, this research comes at an advantageous point, as the travel industry field is developing to scholastic maturity. Review studies left a strong perspective on the travel industry in the scholarly community, analyzing and finding new developments in tourism journals (26)(30)(34)(36). Furthermore, contrasting these patterns with different streams and fields (34)(36) will be an added value to tourisms research spectrum.

Clustering

One of the research papers titled "*Finding Best Possible Number of Clusters using K-means Algorithm*" which is published in Dec 2019 in the journal called International Journal of Engineering and Advanced Technology (IJEAT). In this paper, the author addressed a problem related to online shopping. Basically, online shopping these days is a very common site almost every person visits. It is very useful for each person to buy their product by staying in their safest place. A lot of different products, we can search on e-commerce sites. We can wonder how the data in these websites are arranged. They should analyze customer preferences, their shopping behaviors like what are their interests and their needs through different kinds of techniques. For organizing such data, we need proper classification. In this paper what they have done is, they have grouped the data of the customers with the same buying behavior on the features age and salary. The concepts that they have used are the K-Means algorithm, WSS, Distance Metric methods. The whole work of this paper is implemented in R software.

They have used the K-means algorithm to find similarities between the data points in the dataset. This is to group the data into clusters of data in this K-means algorithm. K-means algorithm is used to know the number of clusters the dataset is grouped into. It will select K clusters for the dataset selected and then it will assign the centroids randomly. K-means calculates the closest centroid for each data in the dataset and assigns that point to the cluster. Then it will set each cluster position to the mean of all data points assigned to the cluster. These two steps are repeated until we found there are no more changes.

Clustering can also be performed based on the distance metric measures. By calculating the distance measure, we can influence the shape of the cluster. We have different distance measures techniques: Euclidean distance is the

most used distance metric technique. It is calculated based on the Minkowski distance by setting p's value 2. Therefore, for calculating the Euclidean distance we should know the formula of Minkowski.

Formula for Minkowski distance metric:

$$\Sigma(|xi - yi|p)1/p$$

If we set the p value as 2, we can calculate the Euclidean Distance

$$\Sigma(|xi - yi|2)1/2$$

In this paper the Elbow method is also used to calculate the Within Sum of Squares (WSS). This calculation of WSS is used to figure out the right number of clusters. The minimum WSS value and minimum distance metric value is considered to form the clusters. The scale () function is also used to calculate the minimum WSS value. In addition, more scenarios and new techniques are added and implement them in the python language.

PROPOSED MODEL

There are various types of scaling like Standard scaling, Normalized scaling, Robust scaling, Minmax scaling, etc., We consider different types of scaling and store them into each data frame. Principal Component Analysis (PCA) is an unsupervised machine learning technique that decreases the dimensionality of the dataset. Using PCA, the unwanted data from the data set is removed and the data set is prevented from over fitting. So, we also implement PCA on the original data and scaled data and store them into different data frames.
The biggest part of clustering the data is finding the number of clusters. To find the number of clusters for the data we use the Elbow Method. Using the elbow method, we can find the optimal number of clusters that can be used for clustering the data. In the elbow method, we find to calculate the inertia value for each value in the set of an assumed number of clusters and plot a graph between the number of clusters and corresponding inertia value for all the data frames. The value from which there is a proportional decrease of the graph for the data frames is considered as the optimal value for the number of clusters. Once, the number clusters value is considered, we should find which data frame should be used to make the clustering model. We use the Distance Metrics concept for choosing the best data frame for implementing the model. In distance metrics, we find how similar the objects are. There are many distance metrics methods, among which Miskowski distance is one. In the Minkowski distance method, we consider two similar data frames and find the similarity between them that is the distance between each data point in the data frames is considered. The data frame which gives the least value is considered for clustering. To perform clustering, we create a model of the K-means algorithm and fit the data frame into it and check the accuracy of the model. After performing the clustering, we plot the data points in the data frame in a graph and then can observe that the points representing the users with similar preferences are together. Each cluster represents its highest or best-preferred categories. Using this grouped data recommendation systems are built. In a recommendation system, when a user gives his/her preference the models check with the already grouped data and shows the places related to the group's categories.

METHODOLOGY

Dataset

Dataset has been sourced from the Machine Learning Repository of the University of California. We have found this dataset on the Kaggle website. For this paper, we have used Google review as a reference. Basically, Google review is a platform where customers give their ratings to a specific place in Google maps. This dataset is generated by capturing user ratings from Google reviews. We have considered 24 categories of average reviews across Europe. The total number of entries in the dataset is 5456 and the total number of attributes is 24.

Data Distribution

As we mentioned above the dataset which we have considered consists of the ratings of the customers with different scales. So, by performing some operations we can find out how this data is distributed in the dataset. For this, we have used the melt () method and Seaborn library. By using the melt () function we can change the data frame format from wide to long. To create a specific format of the data frame object where one or more columns work as a single identifier. Melt is used for converting a set of columns into a single row. Seaborn is a package that was developed based on the Matplotlib library. It is used to create more attractive and as well as informative graphs. Through this, we can clearly understand how the data is distributed among all the attributes.

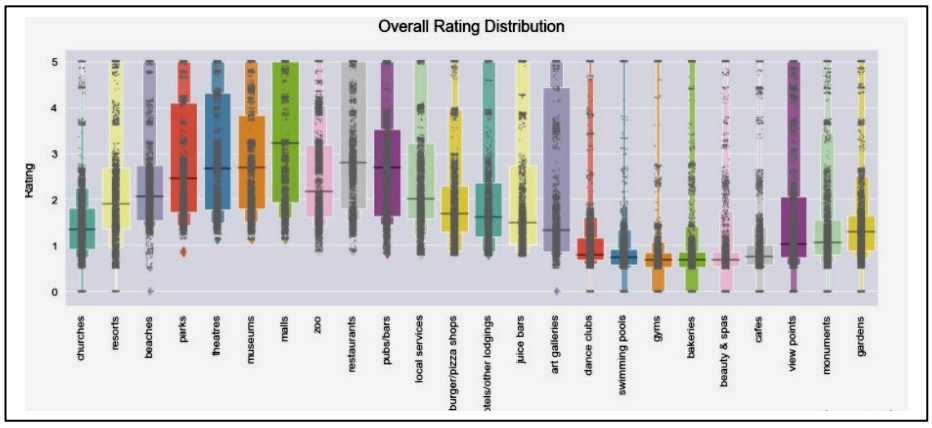

Figure 1: Overall rating Distribution

By observing the above graph, we can understand how the data is distributed in the dataset. The overall ratings are lying between 0.5 to 5. Some attributes like pubs/bars, restaurants which are common attraction points are having wide range of rating distribution. While considering the other attributes like gyms, bakeries, swimming pools have somewhat low ratings.

Data Frames

We have considered some scenarios that should be performed on the data set to get the desired data frames. The data sets are Original data, Scaled data, PCA without scaling, PCA with scaling, MinMax scaled data and Robust scaled data.

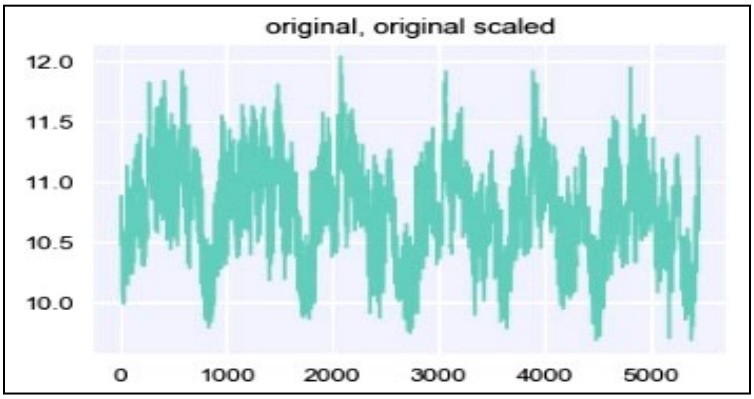

Figure 2: Original Data Scaled

Principle component analysis (PCA), generally called data reduction technique which tries to consider only the essential parts which have more variation of the data and remove the non-essential parts with a fewer variation. MinMax shows the minimum value in the dataset and the maximum value in the dataset. In this, the data is scaled to a fixed range. Robust scaled data tries to scale features using statics that is robust to outliers. This method removes the median and then scales the data. We have applied the distance metric technique and Elbow method to the above scenarios and compared the result.

Distance Metrics

The distance metric technique is used to know the input data patterns to make any data-based decision. The selection of a good distance metric helps us to improve the performance of classification, clustering, or any other algorithms. In this paper, we have chosen the Minkowski metric to measure the distance between the data points in the dataset. Minkowski distance is a generalized technique that can be used to manipulate the formula to calculate the distance between two data points in different ways.

The distance can be calculated using the formula:

$$\Sigma(|xi - yi|p)1/p$$

We can change the value of p and calculate the distance in three different ways which are P=1 Manhatten Distance, P=2 Euclidean Distance and P=3 Chebychev Distance. Observing the distance measures and the graphs, we can say that the distance is much lower in the case of PCA, PCA scaled.

Elbow Method

In the next step what we have done is, we have considered the elbow method to evaluate the inertia values. This inertia value is used to find out the error rate. The elbow method is the technique to determine k, the number of clusters.

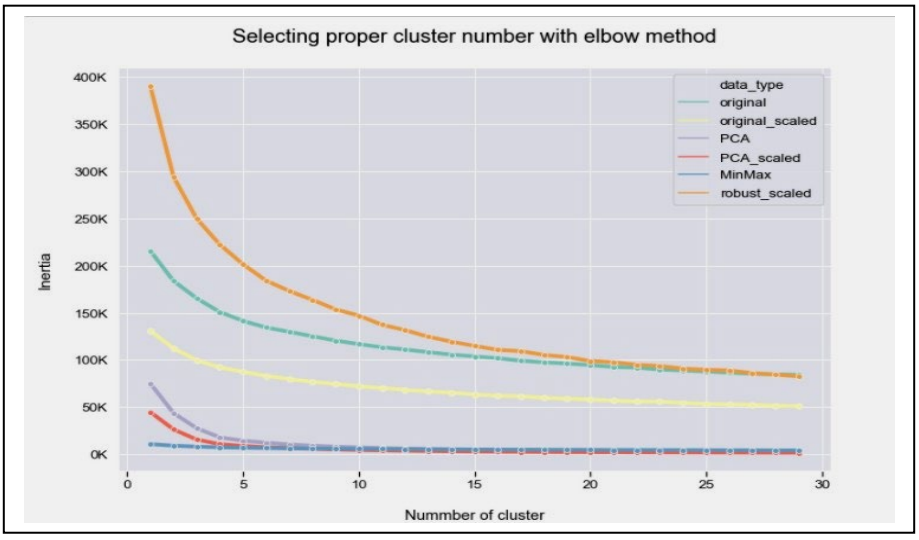

Figure 3: Cluster Selecting using elbow method

From this process what we came to understand is min-max data has a very low inertia value which is very negligible. So, we should consider only the PCA-scaled as the lower inertia value which is above the min-max plot. Therefore, by considering both the techniques which are the distance metric and the elbow method, we can observe that the PCA with scaled data has a lower value when compared to other scenarios.

Clustering

As we have mentioned that the PCA with scaled data has a lower value with respect to distance metric and inertia value, we have applied K-Means clustering analysis on this PCA scaled data. While applying K-Means clustering on PCA-scaled data, we used labels to predict which cluster each data point belongs to. To do this, we accessed the labels attribute from our model object using the dot operator. And then we have transformed the data that means papering the data into the coefficient variable. We also plotted the data before and after the PCA transform and color-coded each point using Eigen Vector representation. The Eigen Vector is an array with n entries where n is the number of attributes.

EXPERIMENTAL RESULTS

Below figures are the results after performing k-means clustering on PCA-scaled data. We can observe that in figure 4.7, the set of attractions in the dataset are scattered into different clusters and we can clearly see those using different colored dots in the graph. And in figure 4.8 we can clearly understand how the user reviews are distributed into four different clusters. A great level of understanding we can observe like how the data is scattered in every single cluster.

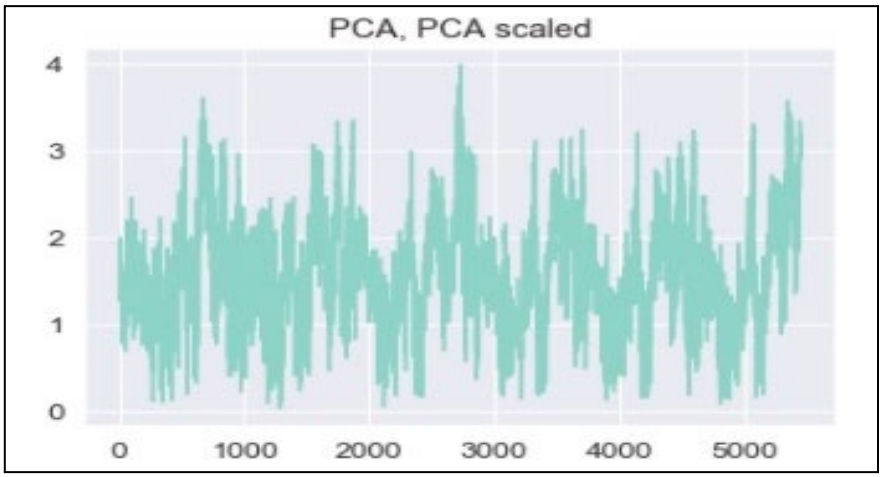

Figure 4: PCA Scaled

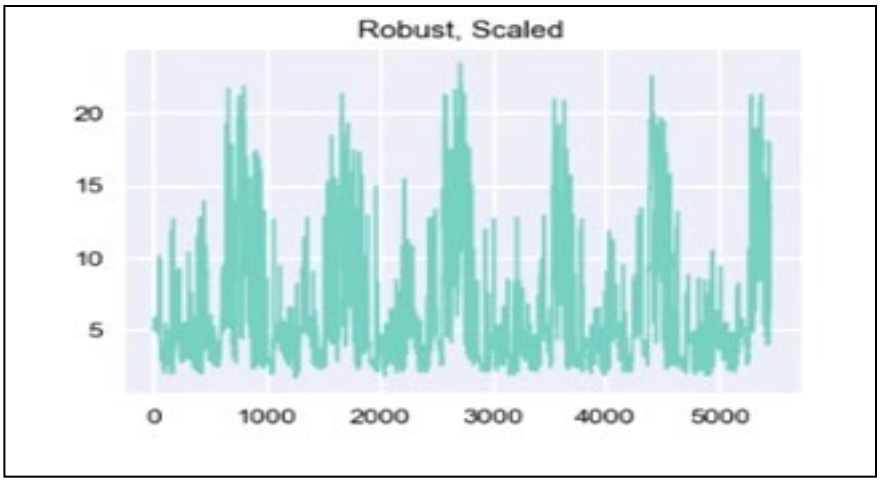

Figure 5: Robust Scaled

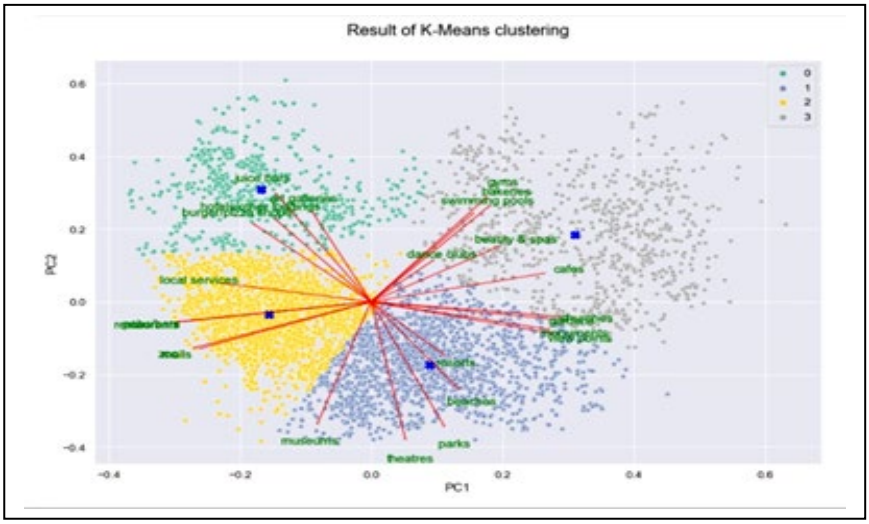

Figure 6: K-Means Clustering

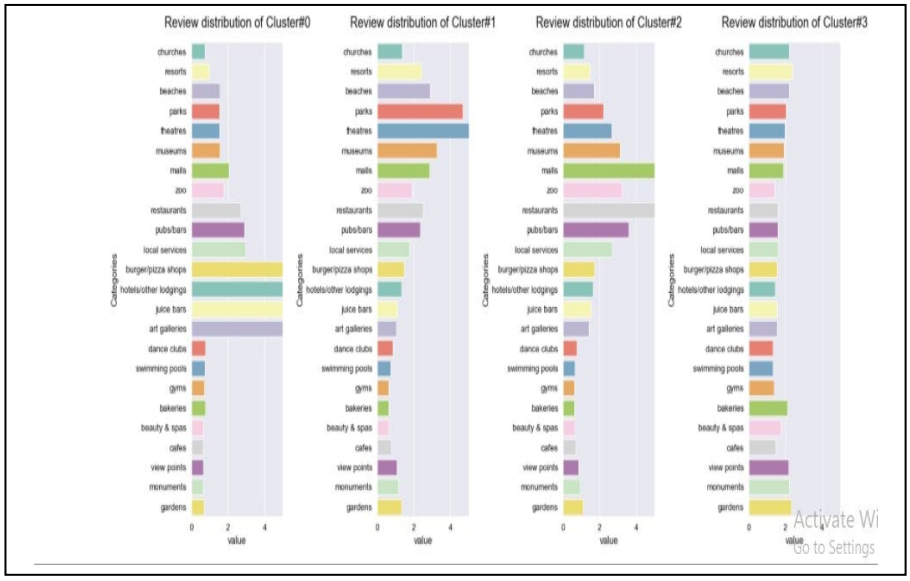

Figure 7: Review Distribution of Cluster

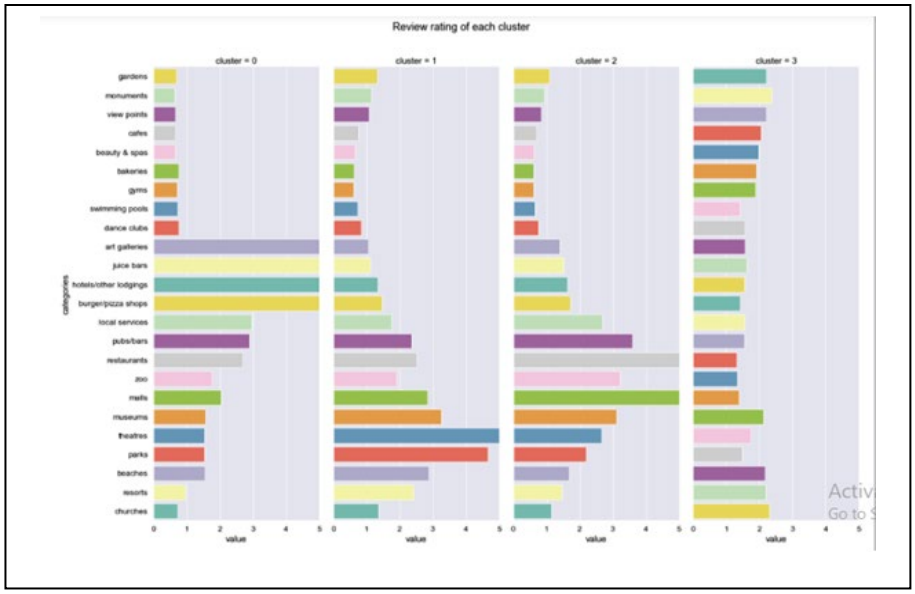

Figure 8: Review rating for each Cluster

K-means clustering result shows that the cluster is divided into 4 segments:

Cluster-0 (Green): In this cluster, we can observe that fast food lovers, pay attention to hotels and juice bars and have art in their hearts. In the total user prefers to love burger/pizza, hotel/other lodgings, juice bars and art galleries the most.

Cluster-1 (Yellow): We have identified that in this cluster the users love to go to local service, spend their time in the zoo, shopping in a mall, having dinner in fine restaurant and pubs/bars. Overall users love malls, zoos, restaurants, pubs/bars, local services the most.

Cluster-2 (Blue): In this cluster, we can observe that the user prefers nature like beaches and parks and loves to go to museums and theaters. By this cluster, we can say that users love beaches, parks, theaters, museums the most.

Cluster-3 (Violet): We can observe from the scatter plot that the member of this cluster is not loosely spread, they might or might not share a common interest. Although users in this group gave an overall rating of about 2 to all attraction but compared to the other 3 clusters, they seem to have more attention to dance clubs, swimming pools, gyms, bakeries, beauty and spas, cafes, viewpoints, monuments, and gardens than above 3 clusters. Therefore, cluster3 can be summarized as a healthy and sightseeing lover.

CONCLUSION

From this paper, we can conclude that if we want to build any recommended system, firstly we need to perform any clustering technique to clearly understand the relationship between the data in the dataset. To perform this clustering, we need to know that which scenario of the dataset is most precise to perform the clustering. Finally, we can conclude from this paper that we have clearly understood that k-means clustering is the best clustering algorithm to know how the data is distributed among the clusters and we have understood the relationship between the attractions in the dataset. This paper currently deals only with the clustering of data. Here we grouped people into different categories based on their preferences and reviews. This data can be used for building recommender systems. The clustered data makes the recommender system's job a little easy when working with the data set. Therefore, the data pre-processing for the recommender system becomes easy.

REFERENCES

1. Baloglu, S. & Assante, M. A content analysis of subject areas and research methods used in five hospitality management journals, Journal Hospitality and Tourism Research, (1999). 23 (1), pp. 53-70.
2. Barros, C.P. Evaluating the efficiency of a small hotel chain with a Malmquist productivity index, International Journal of Tourism Research, (2005). 7 (3), pp. 173-184. http://doi.org/10.1002/jtr.529.
3. Benavides-Velasco, C.A., Quintana-García, C. & Marchante-Lara, M. Total quality management, corporate social responsibility and performance in the hotel industry, International Journal of Hospitality Management, (2014). 41, pp. 77-87. http://doi.org/10.1016/j.ijhm.2014.05.003.
4. Benckendorff PJ, Xiang Z, Sheldon PJ (2019). Tourism information technology. CABI, Boston.
5. Buhalis D, Harwood T, Bogicevic V, Viglia G, Beldona S, Hofacker C Technological disruptions in services: lessons from tourism and hospitality. J Serv Manag (2019). 30:484–506.
6. Chan, W.W. & Ho, K. Hotels' environmental management systems (ISO 14001): creative financing strategy, International Journal of Contemporary Hospitality Management, (2006). 18 (4), pp. 302-316. http://doi.org/10.1108/09596110610665311. Problems and Perspectives in Management, Volume 14, Issue 4, 2016 89
7. Chan, W. & Wong, K. Towards a more comprehensive accounting framework for hotels in China, International Journal of Contemporary Hospitality Management, (2007). 19 (7), pp. 546-559. http://doi.org/10.1108/09596110710818293.
8. Claver-Cortés, E., Molina-Azorín, J.F., Pereira-Moliner, J. & López-Gamero, M.D. Environmental Strategies and Their Impact on Hotel Performance, Journal of Sustainable Tourism, (2007). 15 (6), pp. 663-679. http://doi.org/10.2167/jost640.0.
9. Crawford-Welch, S. & McCleary, K.W. An identification of the subject areas and research techniques used in five hospitality-related journals, International Journal of Hospitality Management, (1992). 11 (2), pp. 155-167.

10. Cvelbar, L.K. & Dwyer, L. An importance-performance analysis of sustainability factors for long-term strategy planning in Slovenian hotels, Journal of Sustainable Tourism, March 2015, (2012). pp. 1-18. http://doi.org/10.1080/09669582.2012.713965.
11. Dalci, I. & Kosan, L. Theory of Constraints Thinking-Process Tools Facilitate Goal Achievement for Hotel Management: A Case Study of Improving Customer Satisfaction, Journal of Hospitality Marketing & Management, (2012). 21 (5), pp. 541-568. http://doi.org/10.1080/19368623.2012.626751.
12. Espino-Rodríguez, T.F. & Padrón-Robaina, V. A resource-based view of outsourcing and its implications for organizational performance in the hotel sector, Tourism Management, (2005). 26 (5), pp. 707-721. http://doi.org/10.1016/j.tourman.2004.03.013.
13. Gavalas D, Konstantopoulos C, Mastakas K, Pantziou G Mobile recommender systems in tourism. J Netw Comput Appl (2014). 39:319–333.
14. Gretzel U Intelligent systems in tourism: a social science perspective. Ann Tour Res (2011). 38(3):757–779.
15. Gretzel U, Sigala M, Xiang Z, Koo C (2015). Smart tourism: foundations and developments. Electron Mark 25(3):179–188
16. Gunter U, Önder I Forecasting city arrivals with Google analytics. Ann Tour Res (2016). 61: 199–212.
17. Gursoy D Future of hospitality marketing and management research. Tour Manag Perspect (2018). 25:185–188.
18. Gursoy, D. & Sandstrom, J. K. An updated ranking of hospitality and tourism journals. Journal of Hospitality and Tourism Research, (2016). 40 (3), 3-18.
19. Kasim, A. Managerial attitudes towards environmental management among small and medium hotels in Kuala Lumpur, Journal of Sustainable Tourism, (2009). 17 (6), pp. 709-725. http://doi.org/10.1080/09669580902928468.
20. Law, R., Ye, Q., Chen, W., Leung, R., An analysis of the most influential articles published in tourism journals from 2000 to 2007: A Google Scholar approach. J. Travel Tour. Mark. 2009. 26 (7), 735–746. http://dx.doi.org/10.1080/ 10548400903284628.
21. Law, R., Qi, S., Buhalis, D., Progress in tourism management: a review of website C.S. Kim et al. International Journal of Hospitality Management 2010. 70 (2018) 49–58 57 evaluation in tourism research. Tour. Manage. 31 (3), 297–313.
22. Law, R., Leung, D., Cheung, C., A systematic review, analysis, and evaluation of research articles in the Cornell Hospitality Quarterly. Cornell Hosp. Q 2012. 53 (4), 365–381.
23. Law, R., Buhalis, D., Cobanoglu, C., Progress on information and communication technologies in hospitality and tourism. Int. J. Contemp. Hosp. Manage. 2014. 26 (5), 727–750.
24. Leung, D., Law, R., Van Hoof, H., Buhalis, D., Social media in tourism and hospitality: a literature review. J. Travel Tour. Mark. 2013. 30 (1–2), 3–22.
25. Leung, R., Au, N., Law, R., The recent asian wave in tourism research: the case of the journal of travel & tourism marketing. Asia Pac. J. Tour. Res. 2015a. 20 (1), 1–28. http:// dx.doi.org/10.1080/10941665.2014.881895.
26. Leung, X.Y., Xue, L., Bai, B., Internet marketing research in hospitality and tourism: a review and journal preferences. Int. J. Contemp. Hosp. Manage. 2015b, 27 (7), 1556–1572.
27. Lim, C., Review of international tourism demand models. Ann. Tour. Res. 1997. 24 (4), 835–849.
28. Lim, C., A meta-analytic review of international tourism demand. J. Travel Res. 1999. 37 (3), 273–284
29. Lee, M. & Jang, S. (Shawn) Market diversification and financial performance and stability: A study of hotel companies, International Journal of Hospitality Management, (2007). 26 (2), pp. 362-375. http://doi.org/10.1016/j.ijhm.2006.02.002.
30. McKercher, B., Tung, V., Publishing in tourism and hospitality journals: is the past a prelude to the future? Tour. Manage. 2015. 50, 306–315. http://dx.doi.org/10.1016/j. tourman.2015.03.008.
31. McKercher, B., Influence ratio: an alternate means to assess the relative influence of hospitality and tourism journals on research. Int. J. Hosp. Manage. 2012. 31 (3), 962–971.
32. McKercher, B., A citation analysis of tourism scholars. Tour. Manage. 29 (6), 1226–1232.
33. Scandura, T., & Williams, E.A. (2000). Research methodology in management: Current practices, trends, and implications for future research. Academy of Management Journal 2008. 43, 1248-1264.
34. Wu, B., Xiao, H., Dong, X., Wang, M., Xue, L., Tourism knowledge domains: a keyword analysis. Asia Pac. J. Tour. Res. 2012. 17 (4), 355–380. http://dx.doi.org/10. 1080/10941665.2011.628330.
35. Yoo, M., Lee, S., Bai, B., Hospitality marketing research from 2000 to 2009: topics, methods, and trends. Int. J. Contemp. Hosp. Manage. 2011. 23 (4), 517–532.

36. Yuan, Y., Gretzel, U., Tseng, Y.-H., Revealing the nature of contemporary tourism research: extracting common subject areas through bibliographic coupling. Int. J. Tour. Res. 2015. 17 (5), 417–431. http://dx.doi.org/10.1002/jtr.2004.
37. Zhang, H., Fu, X., Cai, L.A., Lu, L., Destination image and tourist loyalty: a metaanalysis. Tour. Manage. 2014. 40, 213–223. http://dx.doi.org/10.1016/j.tourman.2013.06. 006.