J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

A Case Study: Unsupervised Approach for Tourist Profile Analysis by K-means Clustering in Turkey

Mustafa Eren Yildirim[1,2] Murat Kaya[3] Ibrahim FurkanInce[1,4*]

ABSTRACT

Data mining is the task of accessing useful information from a large capacity of data. It can also be referred to as searching for correlations that can provide clues about the future in large data warehouses by using computer algorithms. It has been used in the tourism field for marketing, analysis, and business improvement purposes. This study aims to analyze the tourist profile in Turkey through data mining methods. The reason relies behind the selection of Turkey is the fact that Turkey welcomes millions of tourist every year which can be a role model for other touristic countries. In this study, an anonymous and large-scale data set was used under the law on the protection of personal data. The dataset was taken from a leading tourism company that is still active in Turkey. By using the k-means clustering algorithm on this data, key parameters of profiles were obtained and people were clustered into groups according to their characteristics. According to the outcomes, distinguishing characteristics are gathered under three main titles. These are the age of the tourists, the frequency of their vacations and the period between the reservation and the vacation itself. The results obtained show that the frequency of tourist vacations, the time between bookings and vacations, and age are the most important and characteristic parameters for a tourist's profile. Finally, planning future investments, events and campaign packages can make tourism companies more competitive and improve quality of service. For both businesses and tourists, it is advantageous to prepare individual events and offers for the three major groups of tourists.

☞ keyword : Tourist Profile Analysis, Unsupervised Approach, K-means Clustering, Data Mining.

1. 서 론

Although it is essential to keep data for archives, it is even more important for companies to process and analyze this data to seek useful tips for future investments [1]. From this perspective, companies benefit from the developments in technology, and in this regard, they need a larger amount of data to be collected and processed to make fast and easy future predictions [2]. Data mining can be defined as the analysis and extraction of meaningful information from a set of unclear and large amounts of data [3]. It aims to represent large and messy data through logical pattern rules or visual models [4]. Data mining is used in industries such as tourism [5] electronics [6], education [7], credit-scoring [8], sports [9], library [10].

Tourism has an important portion of the Turkish economy [11]. According to the reports taken from the Association of Turkish Travel Agencies (TURSAB), income from tourism is 17.5% and 20.1% of the national export income in 2018 and 2019 respectively [11]. In 2019, the tourism-related income of Turkey was 34.5 billion USD [11]. This fact increases the competition between companies and makes them develop strategies to increase their market shares. One strategy is to analyze their customer's profiles and habits. With the help of current enhancements in information technology, data storage capacities, smart algorithms; companies can determine their target customer profiles, make proper marketing plans, decrease their advertisement costs and reach the most potential customers. By the use of data mining, customers can be known better and the customer services department can improve its services. A better knowledge of the customers allows more

1 Department of Electronics Engineering, Kyungsung University, Busan, SOUTH KOREA
2 Department of Electrical and Electronics Engineering, Bahcesehir University, Istanbul, TURKEY
3 ETS Tour, Istanbul, TURKEY
4 Department of Digital Game Design, Nisantasi University, Istanbul, TURKEY
* Corresponding author: furkanince@ks.ac.kr
[Received 17 Septebmer 2021, Reviewed 5 October 2021(R2 1 November 2021, R3 12 November 2021), Accepted 23 November 2021]

customized products, events, and services by the firms. Thus, companies can make empathy and earn the ability to think as their customers do.

In [12], the authors investigated used questionnaires and aimed to analyze the relationship between customer satisfaction and service they have been provided in Turkey. They conducted their study on 343 tourists. They confirmed that tourist satisfaction is directly proportional to parameters such as service quality, catering, etc. In [13], a study on the expectations and satisfaction of tourists in the Antalya region in Turkey with 10,393 tourists is presented. They also reported that 84.9% of the tourists preferred package tours and stated that the correlation between expectation and satisfaction was 0.724. The author in [14] presented a study that analysis the tourist profile in Turkey through document analysis. It is based on the periodically published reviews of TURSAB between 2000 and 2017. A case study [15] evaluated the place value of a UNESCO World Heritage Site in South Korea to classify visitors according to place value and create marketing strategizes. In another study [16] age group analysis of tourists from 29 European countries was conducted based on their domestic trips taken in 2016. It has been found that more than half of the tourists are represented by tourists from the age range 25-44. Along with the age group 45-54 years, these tourists hold almost 68.1% of the total number of resident tourists. Under the assumption that the domestic trips are connected to the tourists' willingness to spend their weekends and official holidays, then it is obvious that the tourists from these age groups are those more willing to travel in the country and represent the largest part of the domestic tourism demand.

This study aims to analyze very large-scale tourist data of Turkey which welcomes millions of tourists from all around the world every year by clustering and extracting tourist profiles depending on several features. Since the data is big data, it can be used as a reference study for other touristic countries as well. In the study, we defined 30 features of touristic attributes of the people, and the data used in this study is collected from an active tourism firm in Turkey which covers six years. K-means algorithm is used on the collected data for clustering because it is the best well-known automatic clustering algorithm among all other algorithms such as OPTICS [17] and DBSCAN [18]. If satisfactory results are taken with K-Means, it is highly expected that other algorithms such as OPTICS [17] and DBSCAN [18] will also achieve similar results. In other words, the purpose of the study is to discover whether an unsupervised approach will succeed in clustering tourist profiles or not. The remainder of the paper is laid out as follows: In Section 2, we explain the clustering method along with related phrases, and preprocessing of our dataset, after which, in Section 3, we deliver our dataset and results of k-means clustering. It is followed by the explanation of the dataset and steps of the applied methodology. We conclude with an overview of our findings and a discussion in Section 4.

2. Methodology

2.1 K-means algorithm

K-means algorithm is a widely used clustering algorithm that is a centroid-based clustering method and is generally used in data mining and pattern recognition problems [19]. The main goal of the algorithm is to put the members of a large dataset into smaller clusters that are different from each other while the members within clusters are similar to each other. At first, some samples are chosen to represent the initial cluster focal points. Next, the remaining samples are gathered to their focal points according to the distance requirements. This step is followed by the initial classification. If the classification is not well enough, previous steps are repeated until a reasonable cluster distribution is observed. This algorithm is very popular with problems in which unsupervised learning is required due to several advantages it has. Its easy implementation, applicability to very large datasets as well as certainty for convergence are among these advantages. On the other hand, there are a few drawbacks such as its dependency on initial samples, the decision of k value, and scale with the number of dimensions.

2.2 Preprocessing

The data collected in the tourism sector contains more noise and missing parts in comparison to the data in sectors such as banking or telecommunications. This problem occurs because most people enter only the required personal data which also are not as critical as in other sectors. Preprocessing helps to remove the noise from the data, fill the gaps, and set the dataset to the most appropriate form for a reasonable clustering scheme. It has been shown by many studies that preprocessing steps increase the recognition and classification performance of algorithms. Thus, before the application of k-means, some preprocessing steps are made on the collected data. Firstly, columns (features) that have more than 70% missing values are removed. In the remaining dataset, missing values are replaced with the mean of the column they belong to. In our dataset, there are categorical data as well as numerical data. These data are converted to numerical form by using a one-hot encoder method and got values of the set {0,1}.

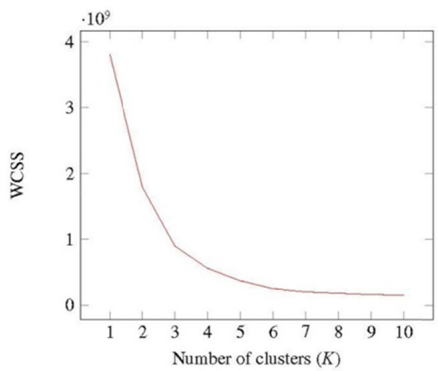
Another problem to be fixed in the dataset is the outliers. They are the data points that are very distant from the values of the same group. They may occur due to a mistake made during the recording or collection of the data. The outliers in a dataset can cause a corrupted distribution and also a low performance in further analysis. Thus, detection of the outliers before the analysis is essential. In our study, a z-score is used to detect the outliers in the dataset. The z-score aims to express any data point by finding its relationship with the standard deviation and mean of the group of data points [20]. It is finding the distribution of data where the mean is 0 and the standard deviation is 1. Points having a greater distance to zero more than a predefined threshold are labeled as outliers. We removed the detected outliers from the dataset. In the last stage of preprocessing, min-max normalization is conducted on each column separately and all values are brought to a range of [0,1].

Moreover, the correlation between the features is calculated. In case of a high correlation, only one column is left while others are removed. The final dataset for clustering consists of 1,000,000 samples and 30 distinct features.

2.3 Choosing k-value

The value of k may affect the performance of the clustering algorithm. It should be dramatically smaller than the number of samples but also large enough to express the characteristics of the dataset. Therefore, a single k value might be misleading or inappropriate for a given dataset. The heuristic approach can be used if the applicable k value range is not big. This approach can be accurate with a sufficient number of trials but it is also time-consuming. The Elbow method is a method to find the optimum value for k in unsupervised clustering problems [21].

The idea of the elbow method is to run the k-means algorithm on the data for a range of k and calculate the within-cluster-sum-of-squares (WCSS) after every run [22]. When the k is equal to the number of samples in the dataset, each sample would also be a cluster itself. Thus the distance between the data and its center would be zero. After obtaining the WCSS at every k value, a WCSS vs k graph is plotted. The result of the elbow method on our dataset is shown in Figure 1. It is seen that the elbow point occurs at k=3.



(Figure 1) Distribution of tourist population into three clusters

2.4 Validation of elbow method by silhouette score

The measure of silhouette score(SS) is used to validate the result of the elbow method. The silhouette score finds the optimum value of k for which the overall intra-cluster distance is minimized and inter-cluster distance is maximized. It is based on the formula given in (1).

$$SS = \frac{(x-y)}{\max(x,y)} \quad (1)$$

In (1), y is the average intra-cluster distance; average distance to the other samples of the same cluster. X is the average nearest cluster distance. SS varies between -1 and 1. A value close to 1 implies that the instance is close to its cluster is a part of the right cluster. Whereas, a value close to -1 means that the value is assigned to the wrong cluster. According to the results given in Table 1, the best score is obtained when k=3. This result validates the result obtained by the elbow method.

3. Test and Results

Dataset of this study consists of reservations and personal details of tourists for six years. All personal information such as name, contact details, etc is removed from the dataset following the law of protection of personal information of Turkey. Data is taken from the source by using Datameer.

(Table 1) Silhouette score results for different k values

Number of Clusters (k)	Silhouette Score (SS)
3	0.324222114436
4	0.307782351272
5	0.260950552594
6	0.256316559071
7	0.253742155711
8	0.257216372113
9	0.264428504859
10	0.265572065843

It is an extract-transform-load (ETL) and data processing software that operates on big data clusters. After

(Table 2) Features of the dataset with their importance coefficients according to the result of k-means clustering

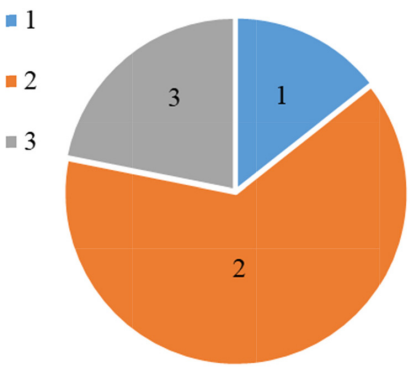
Feature	Definition	Importance Coefficient
months_since_last_res	Number of months since last reservation	0.30
total_days_between_res_and_arrival	Total time from reservation until vacation	0.26
average_days_between_res_and_arrival	Average time from reservation until vacation	0.25
max_days_between_res_and_arrival	Maximum time from reservation until vacation	0.25
active_year_count	Number of years a tourist had a reservation	0.22
total_res_count	Total number of reservations of a tourist	0.19
min_days_between_res_and_arrival	Minimum time from reservation until vacation	0.19
age	Age of tourist	0.19
total_adult	Total number of adults in a reservation	0.14
total_pax	Total number of tourists paid for in the reservation	0.14
resort_res_count	Number of resort reservations	0.13
total_accomodation	Total number of stayed nights of a tourist	0.13
2019_res_count	Reservations in 2019	0.11
avg_month_between_vocation_arrival_dates	Average months between vacations of tourist	0.11
q3	Reservations of a tourist in 3rdquartileoftheyear	0.10
total_amount	Total Money a tourist spent in all vacations	0.10
2017_res_count	Reservations in 2017	0.10
2018_res_count	Reservations in 2018	0.09
2015_res_count	Reservations in 2015	0.09
2016_res_count	Reservations in 2016	0.08
unit_price	Average money per night per person in a reservation	0.08
2014_res_count	Reservations in 2014	0.07
total_child	Total number of minors in a reservation	0.06
average_res_accomodation	The average number of days stayed per reservation	0.05
q2	Reservations of a tourist in 2ndquartileoftheyear	0.03
average_res_amount	Average money spent per reservation	0.03
q1	Reservations of a tourist in the 1stquartileoftheyear	0.02
q4	Reservations of a tourist in the 4thquartileoftheyear	0.01
cult_res_count	Number of cultural tours of a tourist	0.01
abroad_res_count	Number of abroad tours of a tourist	0.01

preprocessing steps mentioned in Section 2.2, the k-means clustering algorithm is applied to the dataset. The parameters of the test are set as k=3, max_iter=300, init= kmeans++. The clustering is done with Python and sklearn library. The parameters used in this study along with their resulting importance scores are shown in Table 2. According to the results, the top eight parameters are related to three main concepts. Parameters months_since_last_res, active_year_count, and total_res_count indicate the tourist's frequency of going on vacation with this company in Turkey. The second title is the period between the reservation and the vacation itself. This title includes the parameters as follows:

total_days_btw_res_and_arrival, avg_days_btw_res_arrival, min_days_btw_res_and_arrival, max_days_btw_res_and_arrival.

The reason for the second title is the discounts or events offered in early reservation packages.

The last title is the age of the tourists. Another observation is that the parameter q3 is more critical than the other quartile features q1, q2, and q4. This is because Turkey is mostly preferred for summer vacations due to its climate and geography. Figure 2 shows the cluster distribution with a pie chart, depending on three distinguishing features. The numbers on the chart are the population of tourists in that cluster.



(Figure 2) Distribution of tourist population into three clusters

Table 3 shows the characteristics of the clusters shown in Figure 2. According to the results, the first cluster refers to 14.3% of all tourists in the dataset. This cluster includes tourists with kids and made their reservations three to six months before the vacation. The second cluster, which occupies 63.7% of the dataset, consists of tourists who have made their reservations just a week before the vacation and also whose total vacation with this company to Turkey is less than three times. The last cluster is 21.8% of the dataset. The tourists of this group are older than 30 years old and spend more than 1000 TRY (160USD) per reservation.

(Table 3) Characteristics and amounts of tourists in three clusters

Clusters Index	Characteristics
1	Early reservation & Customer with kids
2	One week between reservation and vacation & Total reservation less than three
3	Customers older than 30 years old & Budget more than 1000TRY

4. Conclusions and Recommendations

This study aims to analyze the profile of Turkish tourists using data recovery methods. Turkey was chosen because it accepts millions of tourists each year and has the potential to become a role model for other tourist countries. This study used large, anonymous datasets within the scope of privacy law. The dataset was obtained from a major tourist company that is still operating in Turkey. The key parameters of the profile were obtained for this data using the k-means cluster algorithm and the individuals were grouped according to their characteristics. The distinctive features according to the output are grouped under three main headings. This is the age of the tourist. The frequency of their holidays and the time between bookings and the holidays themselves. The results show that the frequency of tourist holidays, the time between bookings and holidays, and age are the most important and characteristic

characteristics of a tourist's profile. Finally, planning future investments, events and campaign packages can make tourism companies more competitive and improve quality of service. For both businesses and tourists, it is advantageous to prepare individual events and offers for the three major tourist groups.

This paper gives a case study for tourist profile analysis in Turkey. The data used in this study contains information on 1,000,000 tourists for six years. First, preprocessing steps on the dataset were conducted. These preprocessing steps include removing the features that have a strong correlation with others, filling the empty values with the mean of that feature, removing features with more than 70% missing in the whole dataset. Outliers are removed by using a z-score and the remaining values are normalized. After obtaining the number of clusters as three with the elbow method, the k-means method is used to group the data. The silhouette score is calculated for validation of the cluster number. We obtained the highest score when the number of clusters was three. The importance coefficient of each feature is extracted.

According to the outcomes, the vacation frequency of a tourist, the period between reservation and vacation, and the age is found to be the most important and distinguishing parameters of the tourist profiles. Finally, planning the future investments, events, and promotion packages may help the tourism companies to be more competitive and increase their service quality. Preparing customized events and offers to target the tourists in three main groups would be more beneficial for companies and also tourists.

References

[1] Olmeda, I., & Sheldon, P. J., "Data mining techniques and applications for tourism internet marketing," Journal of Travel & Tourism Marketing, 11(2-3), 1-20, 2020. https://doi.org/10.1300/J073v11n02_01

[2] Law, R., Mok, H., & Goh, C., "Data mining in tourism demand analysis: A retrospective analysis," International Conference on Advanced Data Mining and Applications, pp. 508-515, Springer, Berlin, Heidelberg, 2007. https://doi.org/10.1007/978-3-540-73871-8_47

[3] Bose, I., Data Mining in Tourism. In Encyclopedia of Information Science and Technology, Second Edition, pp. 936-940, IGI Global, 2009.

[4] Li, Q., Li, S., Zhang, S., Hu, J., & Hu, J., "A review of text corpus-based tourism big data mining," Applied Sciences, 9(16), 3300, 2019. https://doi.org/10.3390/app9163300

[5] Cankurt, A., and Subasi, S., "Tourism demand modelling and forecasting using data mining techniques in multivariate time series: A case study in turkey," Turkish Journal of Electrical Engineering and Computer Sciences, 24, no. 5, 3388 – 3404, 2015.

[6] Lv, S., Kim, H. Y. and Zheng, H. And Jin, B., "A review of data mining with big data towards its applications in the electronics industry," Applied Sciences, 8, 582 – 66, 2018. https://doi.org/10.3390/app8040582

[7] Anjewierden, A., Kollöffel, B., and Hulshof, C., "Towards educational data mining: Using data mining methods for automated chat analysis to understand and support inquiry learning processes," International Workshop on Applying Data Mining in e-Learning, Crete, Grece, 27 – 36, 2007.

[8] Kirkos, E., Spathis, C. and Manolopoulos, Y., "Data mining techniques for the detection of fraudulent financial statements," Expert Syst. Appl., 32, no. 4, 995 – 1003, 2007.

[9] Soleimon, K. O., Data mining in sports: A research overview, a technical report, mis- masters project, 2006.

[10] Jadhav, S. and Kumbargoudar, P., "Multimedia data mining in digital libraries : Standards and features," READIT, 2007.

[11] Ekin, Y., "A Non-Profit Online Marketplace Platform of Travel Agencies in Turkey: TURSAB Rota," International Journal of Contemporary Economics and Administrative Sciences, 11(1), 247-262, 2021.

[12] Yüksel, A. and Yüksel, F., "Comparative performance analysis: Tourists' perceptions of turkey relative to other tourist destinations," Journal of Vacation Marketing, 7, no. 4, 333 – 355, 2001.

[13] Aksu, A., İçigen, E.T. and R. Ehtiyar, R., "A comparison of tourist expectations and satisfaction: A

case study from Antalya region of turkey," TURIZAM, 14, no. 2, 66 – 77, 2010.

[14] Baser, G., "Turkey's tourist profile: A document analysis for future implications," Journal of Tourism and Hospitality Management, 6, no. 5, 222 – 239, 2018.

[15] Song, H. and Kim, H., "Value-based profiles of visitors to a world heritage site: The case of suwon hwaseong fortress (in south korea)," Sustainability, 11, 132 – 151, 2018.

[16] Danut, J. I., "Using the factor analysis method to shape the tourist profile of several european countries by the age group of tourists," "Ovidius" University Annals, Economic Sciences Series, XVIII, no. 1, 303 – 308, 2018.

[17] Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J., "OPTICS: Ordering points to identify the clustering structure," ACM Sigmod record, 28(2), 49-60, 1999.

[18] Arlia, D., & Coppola, M., "Experiments in parallel clustering with DBSCAN," In European Conference on Parallel Processing, pp. 326-331, Springer, Berlin, Heidelberg, August 2001. Available from: http://ai.arizona.edu/hchen/chencourse/Osama-DM_in_Sports.pdf.

[19] MacQueen, J. B., "On the Asymptotic Behavior of k-means," CALIFORNIA UNIV LOS ANGELES WESTERN MANAGEMENT SCIENCE INST, 1965.

[20] Carter, D., "z scores," The Encyclopedia of Statistics in Behavioral Science, 4, 2131 – 2132, 2005.

[21] Sugar, C. and James, G., "Finding the number of clusters in a dataset: An information-theoretic approach," Journal of the American Statistical Association, 98, no. 463, 750 – 763, 2003.

[22] Edwards, A. W., & Cavalli-Sforza, L. L., "A method for cluster analysis," Biometrics, 362-375, 1965.

◐ Authors ◑

Mustafa Eren Yildirim
2008 BsC. in Electrical and Electronics Engineering, Bahcesehir University, Istanbul, Turkey
2010 M.S. in Electrical and Electronics Engineering, Kyungsung University, Busan, Korea
2014 PhD. in Electrical and Electronics Engineering, Kyungsung University, Busan, Korea
Research Interests: Image Processing, Signal Processing, Machine Learning.
E-mail: mustafaeren.yildirim@eng.bau.edu.tr

Murat Kaya
2011 BsC. in Mathematics, Akdeniz University, Antalya, Turkey
2017 MsC. in Information Technologies, Bahcesehir University, Istanbul, Turkey
Research Interests: Social/Semantic Network Data Analysis, IT Leadership, Future Media, Social Media.
E-mail : mkaya@etstur.com

Ibrahim Furkan Ince
2006 BsC. in Computer Education and Instructional Technologies, Bahcesehir University, Istanbul, Turkey
2008 M.S. in Computer Engineering, Bahcesehir University, Istanbul, Turkey
2010 PhD. in IT Convergence Design, Kyungsung University, Busan, Korea
Research Interests: Information Systems, Decision Support Systems, System Development, Image Understanding.
E-mail: ibrahim.ince@nisantasi.edu.tr