International Journal of Research in Engineering, Science and Management
Volume 8, Issue 6, June 2025
https://www.ijresm.com | ISSN (Online): 2581-5792

Clustering Tourism Hotspots in Indonesia Using Hotel Guest Data and the K-Means Algorithm

Muhammad Yusuf Suhardiman*
Student, Department of Management and Industrial Engineering, Faculty of Engineering, Diponegoro University, Semarang, Indonesia

***Abstract*: Tourism plays a strategic role in increasing foreign exchange earnings and driving economic growth in Indonesia. This study aims to cluster the potential for tourist visits based on the distribution of hotel guests across all provinces in Indonesia using the K-Means Clustering method. The data used in this research is secondary data sourced from the 2021 publication of the Statistics Indonesia, which includes the number of both foreign and domestic guests at star-rated and non-star-rated hotels. The analysis was conducted using SPSS software version 25. The results indicate that all provinces can be grouped into three clusters: those with the highest, moderate, and lowest number of hotel guests. Provinces such as DKI Jakarta, West Java, and Bali are included in the cluster with the highest number of guests, while several provinces in Eastern Indonesia tend to fall into the cluster with fewer hotel guests. These findings are expected to provide insights to support the development of the tourism sector, particularly in the planning of promotional strategies and the enhancement of tourism infrastructure in Indonesia.**

***Keywords*: Clustering, Hotel Guests, Indonesian Tourism, K-Means Algorithm, Tourist Distribution.**

1. Introduction

Tourism is a vital sector that contributes significantly to the Indonesian economy. According to data from Badan Pusat Statistik Indonesia (BPS), in 2019, this sector contributed USD 15.16 billion to the national foreign exchange earnings, making it one of the main sectors in Indonesia, alongside oil and gas [3]. Additionally, the growth of the tourism sector has also stimulated the development of supporting industries such as hospitality, transportation, and food services, which, directly and indirectly, create job opportunities and improve the welfare of local communities.

However, despite the substantial potential of Indonesian tourism, there are notable disparities in the level of tourist visits across various provinces. Regions such as DKI Jakarta, Bali, and West Java have become favored destinations for both domestic and international tourists. In contrast, several provinces in Eastern Indonesia, such as North Maluku and Papua, experience lower levels of visitation [4]. This disparity raises questions regarding the distribution of tourism potential and the effectiveness of tourism destination management across different regions.

The distribution of tourists is also influenced by the characteristics of the available accommodation. Research by Zahara et al. (2018) indicates that the presence of adequate star-rated hotels can attract more international tourists, while nonstar-rated hotels are more commonly used by domestic travelers. This phenomenon suggests that the type and quality of accommodation have a significant impact on the level of tourist visits [13].

To understand the tourism potential across all provinces in Indonesia, analysis using regional clustering methods, such as K-Means Clustering, can be highly beneficial. By applying this method, regions can be grouped based on characteristics such as the number of star-rated and non-star-rated hotel guests, including both domestic and international tourists. This provides a more in-depth overview of the distribution of tourist visits throughout Indonesia.

The aim of this study is to offer a clear depiction of the tourism potential in each province using the K-Means Clustering method. It is hoped that the results of this research will serve as a reference for making strategic decisions in tourism development, particularly in formulating policies that support the equitable development of infrastructure and the promotion of tourist destinations across Indonesia.

2. Material and Method

A. Data Collection

The documentation method is used to collect data on various subjects or variables in the form of records, transcripts, books, newspapers, agendas, or other similar sources. Compared to other methods, this approach is relatively less complicated, as any errors in the data source can be easily traced and corrected, as the source itself remains unchanged. The data collection method used in this study is the documentation method, which involves secondary data. This method is employed to gather data and information through the collection of data from the publication of the Central Statistics Agency, specifically the 2021 data on the number of hotel guests by province [7].

B. Research Variables

Research variables are essentially anything in various forms that the researcher defines for study, in order to obtain information about the subject, from which conclusions are then drawn [11]. The variables used in this study are as follows:

1. Provinces across Indonesia (Y1)

*Corresponding author: yusuf.suhardiman97@gmail.com

2. Number of Foreign Guests at Star-Rated Hotels (X1)
3. Number of Domestic Guests at Non-Star-Rated Hotels (X2)
4. Number of Foreign Guests at Non-Star-Rated Hotels (X3)
5. Number of Domestic Guests at Star-Rated Hotels (X4)

C. Research Process

The data analysis diagram used in this study is as follows:

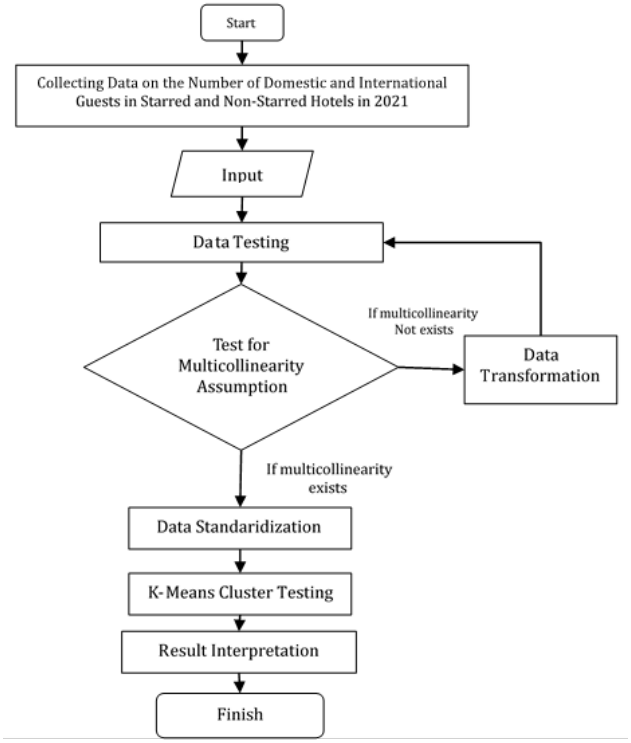

Fig. 1. Steps of K-Means clustering analysis

3. Results and Discussion

A. Research Result

1) Hotel Guest Data for 2021

Prior to the analysis, all relevant data were collected from the Central Bureau of Statistics. Once the data had been compiled, the analysis proceeded using SPSS version 25, applying a clustering method to group provinces based on four variables: the number of foreign guests at star-rated hotels (X1), the number of domestic guests at non-star hotels (X2), the number of foreign guests at non-star hotels (X3), and the number of domestic guests at star-rated hotels (X4).

2) Descriptive Analysis

Prior to the analysis, all relevant data were collected from the Central Bureau of Statistics. Once the data had been compiled, the analysis proceeded using SPSS version 25, applying a clustering method to group provinces based on four variables: the number of foreign guests at star-rated hotels (X1), the number of domestic guests at non-star hotels (X2), the number of foreign guests at non-star hotels (X3), and the number of domestic guests at star-rated hotels (X4).

Based on the table above, there are four variables analyzed, including the number of data points (N), minimum, maximum, mean, and standard deviation values. For the variable X1, the data consists of 34 entries, with a minimum value of 10, a maximum of 283,010, a mean of 19,092.3529, and a standard deviation of 51,438.61926. For variable X2, the data includes 34 entries, with a minimum value of 143,630, a maximum of 4,295,985, a mean of 1,000,028.6176, and a standard deviation of 1,133,129.48111. For variable X3, there are 34 data points, with a minimum value of 0, a maximum of 88,294, a mean of 3,769.9706, and a standard deviation of 15,023.39042. For variable X4, the data consists of 34 entries, with a minimum value of 17,130, a maximum of 8,504,510, a mean of 1,367,930.2941, and a standard deviation of 1,935,334.57340.

3) Cluster Analysis with K-Means

In this study, cluster analysis using the K-Means method was performed to classify hotel guest data from 2021 into three clusters: regions with the lowest, moderate, and highest number of hotel guests.

Table 3
Initial cluster centers

	Cluster		
	1	2	3
X1	10.00	56810.00	15120.00
X2	196100.00	4295985.00	3771242.00
X3	.00	6011.00	789.00
X4	17130.00	8504510.00	4525920.00

The initial cluster centers table above represents the first clustering process before the data undergoes iteration, which is part of the process to form three clusters. Therefore, the table does not require further analysis at this stage.

Table 4
Iteration history

Iteration	Change in Cluster Centers		
	1	2	3
1	700890.468	.000	1001984.230
2	.000	.000	.000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is .000. The current iteration is 2. The minimum distance between initial centers is 4013265.370.

The table above illustrates the iterative process in cluster grouping from the initial table, resulting in two iterations. In the first iteration, the centroid changes were not significant, and significant centroids were only reached in the second iteration. Thus, all clusters were formed, and the iteration stopped at iteration 2, with a minimum distance of 4,013,265.370.

Table 5
Final cluster centers

	Cluster		
	1	2	3
X1	8863.93	56810.00	68828.00
X2	550774.68	4295985.00	2856659.40
X3	3998.61	6011.00	2041.40
X4	621579.29	8504510.00	4120180.00

Table above shows the final results of the clustering process, which formed three clusters for each of the education variables. The variables in the final cluster centers table represent the standardized values. Numbers with a negative sign (-) indicate

Table 1
Hotel guest data for 2021

No	Province	Foreign Guest Numbers Star-Rated Hotel (X1)	Domestic Guest Numbers Non-Star Rated Hotel (X2)	Foreign Guest Numbers Non-Star-Rated Hotel (X3)	Domestic Guest Numbers Star-Rated Hotel (X4)
1	Aceh	120	760758	834	195250
2	Sumatera Utara	4930	3104121	4486	1937910
3	Sumatera Barat	2370	627289	1310	1175550
4	Riau	5410	1173258	389	1341380
5	Jambi	810	461091	74	386670
6	Sumatera Selatan	3230	936355	352	1502830
7	Bengkulu	30	299329	69	197240
8	Lampung	570	846967	421	642800
9	KEP. Bangka Belitung	580	160737	1490	312460
10	KEP. Riau	30300	746304	2178	824020
11	DKI Jakarta	283010	1703088	2333	5997870
12	Jawa Barat	56810	4295985	6011	8504510
13	Jawa Tengah	15120	3771242	789	4525920
14	DI Yogyakarta	14200	1498108	515	3132160
15	Jawa Timur	26880	4206738	2084	5007040
16	Banten	66480	460186	619	1785860
17	Bali	97830	598234	88294	1839420
18	Nusa Tenggara Barat	7570	508162	5840	439090
19	Nusa Tenggara Timur	6100	286607	4165	322020
20	Kalimantan Barat	2360	1122672	975	783980
21	Kalimantan Tengah	1120	737948	1113	275480
22	Kalimantan Selatan	1050	618299	50	831010
23	Kalimantan Timur	6500	745114	192	1307090
24	Kalimantan Utara	430	217331	57	81180
25	Sulawesi Utara	10190	341187	151	446000
26	Sulawesi Tengah	300	541316	116	114010
27	Sulawesi Selatan	2130	1265458	178	1604120
28	Sulawesi Tenggara	340	554939	1909	279090
29	Gorontalo	300	143630	25	95600
30	Sulawesi Barat	10	196100	0	17130
31	Maluku	810	366861	0	86290
32	Maluku Utara	130	224058	43	88410
33	Papua Barat	970	162052	1006	194340
34	Papua	150	319449	111	235900

Table 2
Descriptive statistics

	N	Minimum	Maximum	Mean	Std. Deviation
X1	34	10.00	283010.00	19092.3529	51438.61926
X2	34	143630.00	4295985.00	1000028.6176	1133129.48111
X3	34	.00	88294.00	3769.9706	15023.39042
X4	34	17130.00	8504510.00	1367930.2941	1935334.57340
Valid N (listwise)	34				

that the data is below the overall mean, while numbers with a positive sign (+) indicate that the data is above the overall mean.

Next, to understand the effect of the variables, we will perform calculations using the formula: $X=\mu+z.\sigma$

The average data or sample = population mean + standardized value × standard deviation

X = sample or data mean

μ = population mean

z = standardized value

σ = standard deviation

The following is a chart illustrating the average values of each variable within their respective clusters.

Based on the calculations from each variable, the characteristics of each cluster are identified. The following is the explanation for each cluster:

1. Cluster 1, The characteristic of cluster 1 consists of provinces where the number of hotel guests is below the population mean.
2. Cluster 2, The characteristic of cluster 2 consists of provinces where the number of hotel guests is at the population mean.
3. Cluster 3, The characteristic of cluster 3 consists of

Table 6
The effect of the variables

	Cluster 1	Cluster 2	Cluster 3
X1	45595022965258	292222986939589	354041919566257
X2	62409912735978800	486790735390920000	323696508363287000
X3	6007305616438	9030597681168	3066912620045
X4	120296402683945000	1645907236961910000	797392693942424000

Table 7
ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
X1	8360086231.954	2	2277276411.221	31	3.671	.037
X2	16874966316987.360	2	278112492194.300	31	60.677	.000
X3	10712842.546	2	239572544.770	31	.045	.956
X4	52201138646455.670	2	619350960341.475	31	84.284	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

provinces where the number of hotel guests is above the population mean.

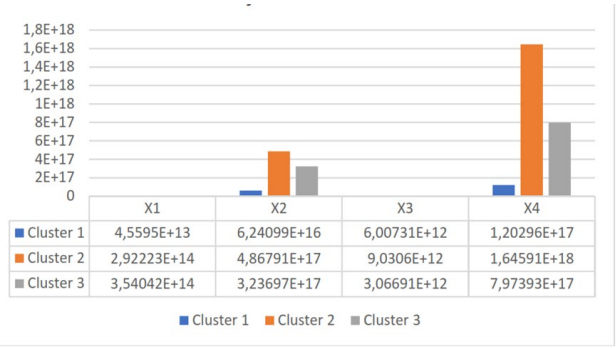

Fig. 2. Average cluster for each variable of 'number of hotel guests'

4) Significance Between Clusters

To test the level of significance between clusters and identify the differences within each cluster, an ANOVA test is required. The F value in cluster analysis indicates that the higher the F value, the greater the difference between the three formed clusters (Bastian, 2018). The following are the results of the ANOVA test:

Based on the obtained table, the "Cluster" column represents the between-cluster means, and the "Error" column represents the within-cluster means. Meanwhile, the F column is derived from the following formula:

$$F = \frac{between\ cluster\ mean}{within\ cluster\ mean}$$

Hypothesis:

H0: The three clusters do not show significant differences.

H1: The three clusters show significant differences.

If the significance value > 0.05; H_0 is accepted and H_1 is rejected. If the significance value < 0.05; H_0 is rejected and H_1 is accepted

In the ANOVA table, all variables have significance values greater than 0.05. Therefore, the three clusters do not have significant differences. The variables with the least differentiation between clusters are X2 and X4, as they have the largest F values among the other variables, namely 84.284, with a significance of 0.003.

5) Cluster Composition Profilling

After determining the number of members in each cluster, the cluster membership results (QCL_5) are entered into the clustering table, and the results are as follow:

Based on the cluster composition profiling, the data were subsequently classified into three distinct clusters.

B. Discussion

In this analysis, there are five background education variables used to group new job applicants in the Department of Industry and Manpower of Pekalongan City in 2021, Number of Foreign Guests at Star-Rated Hotels (X1), Number of Domestic Guests at Non-Star Hotels (X2), Number of Foreign Guests at Non-Star Hotels (X3), and Number of Domestic Guests at Star-Rated Hotels (X4). Before conducting the K-Means Cluster test, the data must meet the classical assumption in clustering, i.e., the multicollinearity test, which serves to examine whether there are independent variables that show similarities among other independent variables. Based on the results of the K-Means Cluster analysis, we can conclude which cluster contains provinces with the highest, moderate, and lowest number of hotel guests based on the characteristics of each cluster.

In Clusters 1, 2, and 3, based on the average values from the cluster table for the number of hotel guests, we can conclude that Domestic Guests at Star-Rated Hotels (X4) occupy the highest position in terms of the number of hotel guests in 2021 compared to other types of guests. The second-highest position in terms of average guest count is Domestic Guests at Non-Star-Rated Hotels (X2), followed by Foreign Guests at Star-Rated Hotels (X1) and Foreign Guests at Non-StarRated Hotels (X3).

4. Conclusion

In Clusters 1, 2, and 3, based on the average values from the cluster table for the number of hotel guests, we can conclude that Domestic Guests at Star-Rated Hotels (X4) occupy the highest position in terms of the number of hotel guests in 2021 compared to other types of guests. The second-highest position in terms of average guest count is Domestic Guests at Non-Star-Rated Hotels (X2), followed by Foreign Guests at Star-Rated Hotels (X1) and Foreign Guests at Non-StarRated Hotels (X3). analysis was continued. This analysis grouped the data over 12 months into 3 registration clusters: highest, moderate, and lowest guest count months.

a. *Cluster 1*: Represents the category of provinces with the lowest number of hotel guests. The provinces

Table 8
Cluster composition profiling

No	Province	Foreign Guest Numbers Star-Rated Hotel (Person)	Domestic Guest Numbers Non-StarRated Hotel (Person)	Foreign Guest Numbers Non-Star-Rated Hotel (Person)	Domestic Guest Numbers Star-Rated Hotel (Person)	QCL_1	QCL_2
1	Aceh	120	760758	834	195250	1	47.532.738.840
2	Sumatera Utara	4930	3104121	4486	1937910	3	219.718.650.252
3	Sumatera Barat	2370	627289	1310	1175550	1	55.927.398.776
4	Riau	5410	1173258	389	1341380	1	95.164.253.406
5	Jambi	810	461091	74	386670	1	25.160.644.472
6	Sumatera Selatan	3230	936355	352	1502830	1	96.193.557.196
7	Bengkulu	30	299329	69	197240	1	49.333.785.460
8	Lampung	570	846967	421	642800	1	29.708.887.309
9	KEP. Bangka Belitung	580	160737	1490	312460	1	49.775.399.555
10	KEP. Riau	30300	746304	2178	824020	1	28.227.075.319
11	Dki Jakarta	283010	1703088	2333	5997870	3	221.411.849.844
12	Jawa Barat	56810	4295985	6011	8504510	2	.00000
13	Jawa Tengah	15120	3771242	789	4525920	3	100.198.423.021
14	DI Yogyakarta	14200	1498108	515	3132160	3	168.072.364.627
15	Jawa Timur	26880	4206738	2084	5007040	3	161.585.659.088
16	Banten	66480	460186	619	1785860	1	116.922.492.437
17	Bali	97830	598234	88294	1839420	1	122.491.186.119
18	Nusa Tenggara Barat	7570	508162	5840	439090	1	18.741.196.534
19	Nusa Tenggara Timur	6100	286607	4165	322020	1	39.940.955.803
20	Kalimantan Barat	2360	1122672	975	783980	1	59.455.191.661
21	Kalimantan Tengah	1120	737948	1113	275480	1	39.355.668.332
22	Kalimantan Selatan	1050	618299	50	831010	1	22.022.126.838
23	Kalimantan Timur	6500	745114	192	1307090	1	71.253.967.586
24	Kalimantan Utara	430	217331	57	81180	1	63.506.121.138
25	Sulawesi Utara	10190	341187	151	446000	1	27.344.404.021
26	Sulawesi Tengah	300	541316	116	114010	1	50.774.448.487
27	Sulawesi Selatan	2130	1265458	178	1604120	1	121.499.730.370
28	Sulawesi Tenggara	340	554939	1909	279090	1	34.262.702.202
29	Gorontalo	300	143630	25	95600	1	66.521.434.795
30	Sulawesi Barat	10	196100	0	17130	1	70.089.046.757
31	Maluku	810	366861	0	86290	1	56.607.394.852
32	Maluku Utara	130	224058	43	88410	1	62.538.404.491
33	Papua Barat	970	162052	1006	194340	1	57.767.637.817
34	Papua	150	319449	111	235900	1	44.983.455.518

included in this category are:Aceh, West Sumatra, Riau, Jambi, South Sumatra, Bengkulu, Lampung, Bangka Belitung Islands, Riau Islands, West Java, Banten, Bali, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, Central Kalimantan, South Kalimantan, East Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, and Papua.

b. *Cluster 2*: Represents the category of provinces with a moderate number of hotel guests. The province in this category is West Java

c. *Cluster 3*: Represents the category of provinces with the highest number of hotel guests. The provinces in this category are:North Sumatra, DKI Jakarta, Central Java, Yogyakarta, and East Java.

Table 9
Distribution into 3 clusters

Cluster 1		Cluster 2	Cluster 3
Aceh	Kalimantan Tengah	Jawa Barat	Sumatera Utara
Sumatera Barat	Kalimantan Selatan		Dki Jakarta
Riau	Kalimantan Timur		Jawa Tengah
Jambi	Kalimantan Utara		Di Yogyakarta
Sumatera Selatan	Sulawesi Utara		Jawa Timur
Bengkulu	Sulawesi Tengah		
Lampung	Sulawesi Selatan		
Kep. Bangka Belitung	Sulawesi Tenggara		
Kep. Riau	Gorontalo		
Jawa Barat	Sulawesi Barat		
Banten	Maluku		
Bali	Maluku Utara		
Nusa Tenggara Barat	Papua Barat		
Nusa Tenggara Timur	Papua		
Kalimantan Barat			

References

[1] Akramunnisa, "K-MEANS Clustering Analysis Pada Persebaran Tingkat Pengangguran Kabupaten/Kota di Sulawesi Selatan," J. Varian, vol. 3, no. 2, 2020.

[2] Badan Pusat Statistik, Statistik Pariwisata Indonesia 2019, Jakarta: BPS, 2020.

[3] Ministry of Tourism and Creative Economy, Annual Tourism Report 2020, Jakarta: Kemenparekraf, 2021.

[4] J. F. Hair, Multivariate Data Analysis, 7th ed., Upper Saddle River, NJ: Prentice Hall, 2009.

[5] T. Hidayat, "Analisis Karakteristik Konsumen Hotel 'X' dengan Menggunakan Metode K-Means Clustering," J. Media Teknik & Sistem Industri, vol. 4, no. 2, pp. 53–59, 2020.

[6] A. Kunto, Riset Pemasaran dan Perilaku Konsumen, 2010.

[7] U. Narimawati, Metodologi Penelitian Kualitatif dan Kuantitatif, Teori dan Aplikasi, Bandung: Agung Media, 2008.

[8] S. Santoso, Statistika Ekonomi Plus Aplikasi SPSS, Universitas Muhammadiyah Ponorogo Press, 2013.

[9] R. Sari, "Data Mining: Algoritma K-Means pada Pengelompokkan Wisata Asing ke Indonesia Menurut Provinsi," in Seminar Nasional Sains & Teknologi Informasi (SENSASI), 2018.

[10] Sugiyono, Metode Penelitian Kuantitatif Kualitatif dan R&D, Bandung: Alfabeta, 2011.

[11] Y. L. Sukestiyarno and P. D. MS, Statistika Dasar, Universitas Negeri Semarang, 2012.

[12] P. Zahara, M. Agustina, and Y. Zainuddin, "Pengelompokan Objek Daya Tarik Wisata (ODTW) di Kota Banda Aceh Berdasarkan Cluster Analysis," J. Arsitektur dan Perencanaan, vol. 11, no. 1, 2018.