2025 6th International Conference on Data Analytics for Business and Industry (ICDABI) | 979-8-3315-6982-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICDABI67967.2025.11547485

Comparative Characteristics of Foreign Tourists Inflow to Indonesia: A Clustering Perspective

Gunawan
Faculty of Engineering
University of Surabaya
Surabaya, Indonesia
gunawan@staff.ubaya.ac.id

***Abstract*—Foreign tourists play a vital role in driving a country's economic development by contributing significantly to foreign exchange earnings, job creation, and regional economic growth. Indonesia, as an archipelagic country, offers numerous tourist attractions and destinations. The objective of this study is to compare and contrast the characteristics of foreign tourist inflow to Indonesia based on their countries of origin. The official data on tourist inflow from 53 countries and regions are analyzed using the CRISP-DM methodology and KNIME as a computational tool. Two clusters, named distant and nearby clusters, are identified based on tourist occupation, visit purpose, tourism activity, and expenditure per visit. Tourists from the distant cluster are more likely to have a professional occupation and are more recreational in their visit purposes. Tourists from nearby countries are more likely to have a business purpose. The result supports the gravity model of tourism. This study contributes to the literature on analyzing foreign tourist inflow and applying data analytics to official tourism data.**

***Index Terms*—CRISP-DM, gravity model, k-means, k-medoids**

I. Introduction

The tourism industry stimulates local economies by increasing demand for services and products, creating employment opportunities, and generating income. Research on the BRICS nations demonstrates that international tourism is a crucial component of foreign trade [1]. Additionally, a study concerning Indonesia indicates that international tourists drive local economic growth [2]. In general, foreign tourists play a vital role in driving a country's economic development by contributing significantly to foreign exchange earnings, job creation, and regional economic growth. Therefore, countries are striving to develop their tourism industries to attract more foreign tourists.

Understanding foreign tourist behavior is crucial in formulating marketing and operational strategies. For instance, research conducted in Indonesia found that visitors from the Middle East tend to favor Arabic dining options, primarily due to their general dislike for local eateries [3]. Another study found that tourists heading to Japan were more likely to be high-income travelers, while tourists to Thailand were budget-conscious [4]. Segmenting foreign tourists based on their characteristics is a valuable approach for effective marketing and tourism management. Geographic segmentation was the most effective approach, whereas motivation segmentation was found to be the least reliable [5].

Destinations and tourist attractions must be continually promoted to potential tourists in countries that are identified as target markets. In marketing, potential target markets could be identified by grouping them based on their attributes. Understanding these groups enables more efficient and effective promotional and marketing activities. Research on grouping or clustering tourist destinations or objects is common. Some studies are about clustering countries of foreign tourist origin [6], clustering attractions based on tourist movement in Hong Kong [7], and clustering local tourist objects [8].

One of the theories used in tourism research is the gravity model. The gravity model is inspired by Newton's law of gravity, which states that interaction is directly proportional to the economic size of the units and inversely proportional to the distance between them. Tourism's gravity model proposes that travel between two sites increases with their economic size (such as population or GDP) and decreases with the distance between them. A review article on the gravity model for tourism confirms that GDP, population, and distance are the most determining variables for tourist inflow [9].

Indonesia, as an archipelagic country, offers numerous tourist attractions and destinations. As one of the world's most diverse and culturally rich destinations, Indonesia attracts millions of international visitors each year, generating substantial revenue for various economic sectors, including hospitality, transportation, food and beverages, and the creative industries. Following the COVID-19 pandemic, the number of international tourists increased from 5.9 million in 2022 to 11.7 million in 2023 and then to 13.9 million in 2024 [10]. To facilitate international flights to Indonesia, the Indonesian Ministry of Transportation designated 17 international airports in April 2024. Additionally, nine international ports welcome visitors from overseas. These facilitate the growth of foreign visitors to Indonesia.

The marketing question is whether foreign tourists from various countries share similar or distinct characteristics that are useful for formulating marketing strategies. Grouping the countries with similar characteristics will simplify the understanding of foreign tourists. While secondary data on foreign tourists is available, the grouping is not readily apparent. The objective of this study is to compare and contrast the characteristics of foreign tourist inflow to Indonesia based on their country of origin. The specific aims are (1) to group

tourists' countries of origin based on some characteristics, and (2) to contrast the characteristics of tourists' countries of origin.

II. METHODS

This study could be classified as exploratory and secondary research. This study adopted a data mining approach, with the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology [11]. This methodology consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The computation is conducted through the KNIME Analytics Platform, and visualization with Tableau.

A. Data Understanding

The Phase of Business Understanding refers to the objective of the data analysis, which is aligned with the research objective. The aim is to understand the classification and characteristics of foreign tourists, which will guide further managerial decisions.

Furthermore, the phase of data understanding refers to identifying and selecting data that is suitable to meet the analysis purpose. Secondary data for analysis are sourced from the BPS_Statistics Indonesia (National Statistics Office) in the published report titled "International Visitor Arrivals Statistics 2024" [10] and "International Visitors Expenditure Statistics 2024" [12]. Data are reported on the tourism characteristics for each country or region of origin. Therefore, the unit of analysis is a country or region. The total number of objects is 53 countries and regions, as shown in Table I. Here, regions represent a collection of countries, such as Other Eastern Europe, Other Middle East, and Other Africa.

Variables for analysis were extracted and selected from the available data. Though this study is exploratory, a conceptual framework, shown in Fig. 1, is relevant as a guide for analysis. The framework describes four variables: visitor occupation, visit purpose, tourism activity/destination, and expenditure per visit. The association among these variables can be loosely defined, not as a cause-and-effect relationship, as follows. The visitor occupation can be associated with their visit purpose, which in turn is linked to the tourism activity. Next, their activities determine the average expenditure per visit.

The initial number of measures for each variable is as follows: five for visitor occupation, three for visit purpose, 11 for tourism activities, and one for expenditure (Fig. 1). Data were cleaned by deleting measures with missing values exceeding 15%. After data cleaning, the number of measures and their names for each variable are as follows. First, visitor occupation includes two measures: Manager and Professional. Second, the purpose of the visit covers two: Business and Recreation. Then, tourism activities consist of seven measures: Adventure Tourism, Art and Culinary Tourism, Eco Tourism, Heritage and Religious Tourism, Marine Tourism, Rural Tourism, and Urban Tourism. Finally, expenditure is measured by the average spending per visit.

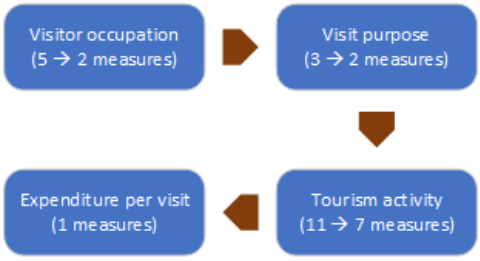

Fig. 1. Analysis framework.

TABLE I
LIST OF COUNTRIES AND REGIONS

Country	Code	Country	Code
Brunei Darussalam	BN	Netherlands	NL
Philippines	PH	Belgium	BE
Cambodia	KH	Denmark	DK
Lao People's D.R.	LA	Russian Federation	RU
Malaysia	MY	Finland	FI
Myanmar	MM	United Kingdom	GB
Singapore	SG	Italy	IT
Thailand	TH	Germany	DE
Viet Nam	VN	Norway	NO
Bangladesh	BD	France	FR
China	CN	Portugal	PT
Hong Kong	HK	Spain	ES
India	IN	Sweden	SE
Japan	JP	Switzerland	CH
South Korea	KR	Other Eastern Europe	OEE
Pakistan	PK	Other Western Europe	OWE
Sri Lanka	LK	USA	US
Taiwan	TW	Canada	CA
Timor-Leste	TL	Central America	CAm
Other Asia	OAs	South America	OSA
Saudi Arabia	SA	Other America	OAm
Kuwait	KW	Australia	AU
Egypt	EG	New Zealand	NZ
Qatar	QA	Other Oceania	OOc
U.A.E	AE	South Africa	ZA
Other Middle East	OME	Other Africa	OAF
Austria	AT		

Table II presents the descriptive statistics of the measures. All measures in the visitor occupation, visit purpose, and tourism activity are in percentages. For example, the recreation indicates the portion of visitors in a particular country who have recreation as their primary purpose. The expenditure is presented as the average expenditure per visitor per visit in USD. Therefore, all measures are relative, making them comparable across countries.

B. Analysis Method

From a machine learning perspective, all measures have no label, which means there is no dependent variable. Therefore, the analysis can be classified as unsupervised learning, and clustering analysis is the most appropriate approach. An initial examination of the object distribution was conducted, revealing no discernible pattern. Therefore, a centroid-based clustering algorithm was chosen. Centroid-based clustering is a partitioning method in machine learning that groups data points into clusters based on their proximity to cluster

TABLE II
DESCRIPTIVE STATISTICS OF MEASURES

Measure	min	mean	max
Visitor occupation			
1 Manager %	0	19	43
1 Professional %	24	45	60
Visit purpose			
2 Business %	1	9	24
2 Recreation %	15	78	93
Tourism activity			
3 Adventure Tourism %	0	42	59
3 Art and Culinary %	0	37	52
3 Eco Tourism %	0	30	52
3 Heritage and Religious Tourism %	0	39	63
3 Marine Tourism %	12	44	79
3 Rural Tourism %	0	19	50
3 Urban Tourism %	38	49	90
Expenditure			
4. Average expenditure per visit USD	445	1607	2194

centroids. In this type, the prominent technique is k-means, and the less popular one is k-medoids. In both techniques, the number of clusters (k) is specified beforehand. The main difference is that the centroids in k-means are the mean of the points, while k-medoids are one of the points [14]. The analysis compares both techniques with variations of the k number.

C. Computational Tool

The data analysis was performed using the KNIME Analytical Platform as a non-code data analytics application. KNIME is a data analytics tool that enables the building of analysis workflows through drag-and-drop programming nodes. The k-means and k-medoids clustering workflows composed of squared KNIME nodes are presented in Fig. 2. Expenses in USD were scaled, with the Normalizer node, to 0 – 100.

III. RESULTS AND DISCUSSION

This section first presents the results of the cluster analysis. The results are then interpreted using the gravity model for tourism. Lastly, the recommendation is drawn.

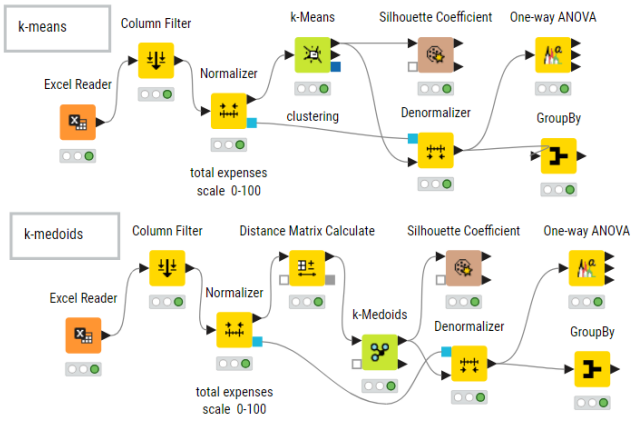

Fig. 2. Basic k-means and k-medoids workflows.

A. Result of Cluster Analysis

The clustering analysis was performed using KNIME for k-means and k-medoids with k values of 2, 3, and 4. Table III presents the comparison results of k-means and k-medoids clustering algorithms. The goodness of a cluster is measured by the silhouette coefficient (SC), which is a metric used to evaluate the quality of a clustering result. It measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The mean silhouette coefficient is the average of all individual silhouette scores for each data point. The range is between -1 and +1. A higher mean silhouette coefficient indicates better-defined and more separated clusters. Table III shows that silhouette coefficients for k-medoids are higher than for k-means. It means that clusters formed by k-medoids are more divided (better) than by k-means. This result is supported by a study demonstrating that k-medoids outperforms k-means, including faster execution times, noise reduction, and robustness to outliers [13].

In Table III, the k-means part indicates that the highest silhouette coefficient is achieved at k = 2, with a mean score of 0.469, which is the same as that in k-medoids. The cluster size for both k-means and k-medoids with k = 2 is 20 and 33 members. Therefore, both techniques yield the same result for k = 2. The first cluster, named cluster_0, contains 33 countries and regions as shown in Table IV. Furthermore, the second cluster, named cluster_1, covers 20 countries and regions as shown in Table V.

Fig. 3 presents the geographical distribution of the two clusters. Nine non-specific country regions, such as other Asia, Central America, and other Eastern Europe, were excluded from the map, because the map uses ISO country codes. The map indicates that countries in cluster_1 are predominantly Asian, while those in cluster_0 are located in Europe, America, Africa, and Australia. Therefore, cluster_1 is referred to as the

TABLE III
SILHOUETTE COEFFICIENTS AND CLUSTER SIZE

k	k-means		k-medoids	
	SC	*size*	*SC*	*size*
k = 2	0.469	20, 33	0.469	20, 33
k = 3	0.280	9, 20, 24	0.422	5, 16, 32
k = 4	0.303	5, 15, 10, 23	0.315	5, 9, 15, 24

TABLE IV
MEMBER OF CLUSTER_0 (DISTANT CLUSTER)

Country/Region	Country/Region	Country/Region
Other Asia	Russian Federation	Other East. Europe
Saudi Arabia	Finland	Other West. Europe
Kuwait	UK	USA
Egypt	Italy	Canada
Qatar	Germany	Central America
UEA	Norway	South America
Other Middle East	France	Other America
Austria	Portugal	Australia
Netherlands	Spain	New Zealand
Belgium	Sweden	South Africa
Denmark	Switzerland	Other Africa

TABLE V
MEMBER OF CLUSTER_1 (NEARBY CLUSTER)

Country/Region	Country/Region	Country/Region
Brunei	Pakistan	Myanmar
Philippines	Sri Lanka	Other Oceania
China	Taiwan	Singapore
Hong Kong	Timor Leste	Thailand
India	Cambodia	Vietnam
Japan	Laos	Bangladesh
Korea Rep	Malaysia	

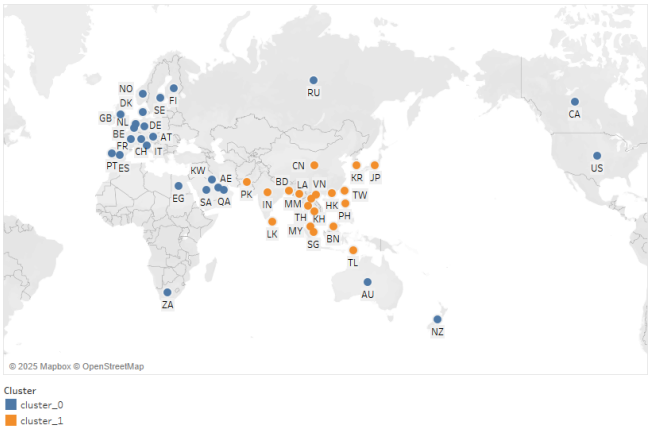

Fig. 3. Geographic map of clusters.

nearby cluster, and cluster_0 is referred to as the distant cluster.

Furthermore, the characteristics of each cluster were investigated. The first step is to identify which measures are significantly different between the two clusters by applying the ANOVA test. The p-values from ANOVA are the same as the p-values from the t-test assuming equal variances. The measure of total visitors per country was included in the analysis. The second step is to perform cross-tabulation using KNIME's Groupby node to reveal the mean score for each measure. The results are presented in Table VI, which shows the significance level and the mean score of each measure. Ten out of 13 measures significantly differentiate between the two clusters. Three measures do not significantly discriminate between the two clusters: occupation as a manager, tourism activity in art and culinary, and urban tourism.

A comparison of mean values between the two clusters reveals that the distant cluster has higher scores for eight significant measures and lower scores for two measures compared to the nearby cluster. Tourists from the distant cluster are more likely to be professionals and recreational visitors. As the purpose is more recreational, their tourist destinations are more focused on adventure tourism, eco-tourism, heritage and religious tourism, marine tourism, and rural tourism. They spend more money per visit than tourists from nearby countries.

Conversely, tourists from nearby countries are more likely to have a business purpose for their visit and a higher number of visitors. Although the result is slightly non-significant (p = 0.062), the nearby cluster indicates a higher mean score for

TABLE VI
COMPARISON OF DISTANT AND NEARBY CLUSTERS

Cluster	p-value	Distant cluster	Nearby cluster
1 Manager %	0.613	20	19
1 Professional %	0.000	48	40
2 Business %	0.002	8	12
2 Recreation %	0.000	83	69
3 Adventure Tourism %	0.000	50	30
3 Art and Culinary %	0.568	59	58
3 Eco Tourism %	0.000	35	21
3 Heritage and Religious Tourism %	0.037	41	36
3 Marine Tourism %	0.000	52	32
3 Rural Tourism %	0.000	22	13
3 Urban Tourism %	0.062	47	32
4 Ave exp per visit USD	0.000	1,939	1,058
5 Number of visitors	0.028	158,432	433,709

urban tourism than the distant cluster. This urban destination supports that the business purpose for visitors from nearby countries is higher than for those from distant countries.

Furthermore, the average expenditure per visit for the distant cluster is about 1.8 times higher than that of the nearby cluster. However, the number of visitors from countries in the nearby cluster is 2.7 times higher than that from the distant cluster. Table VII presents three measures and their multiplication to estimate the total expenditure of tourists for each cluster. The total tourist expenditure from the distant cluster is about 10% higher than that from the nearby cluster.

The characteristic difference between clusters could be expressed through two dimensions of measures. For example, Fig. 4 presents the graph of two measures: marine tourism vs. expenditure per visit. The plot reveals that countries in the nearby cluster have lower average expenditure per visit and are less likely to choose marine tourism than those in the distant cluster.

B. Interpretation based on the Gravity Model

The results of the cluster analysis, which grouped foreign tourist markets into two main clusters, can be interpreted using the gravity model of tourism. In this study, the nearby cluster primarily consists of Asian countries, whereas the distant cluster includes countries from Europe, North America, Australia, and Africa. Previous research found that the distance from the tourist's origin is related to behavioral patterns [15]. Tourists from the nearby cluster tend to have a higher volume of visits but lower average expenditure per visit. This segment could be described as short-haul, repeat visitors

TABLE VII
TOURIST SPENDING PER CLUSTER

Measure	Distant cluster	Nearby cluster
Average expenditure per visit USD	1,939	1,058
Number of visitors	158,432	433,709
Countries and regions	33	20
Total expenditure USD	10,137,588,384	9,177,282,440

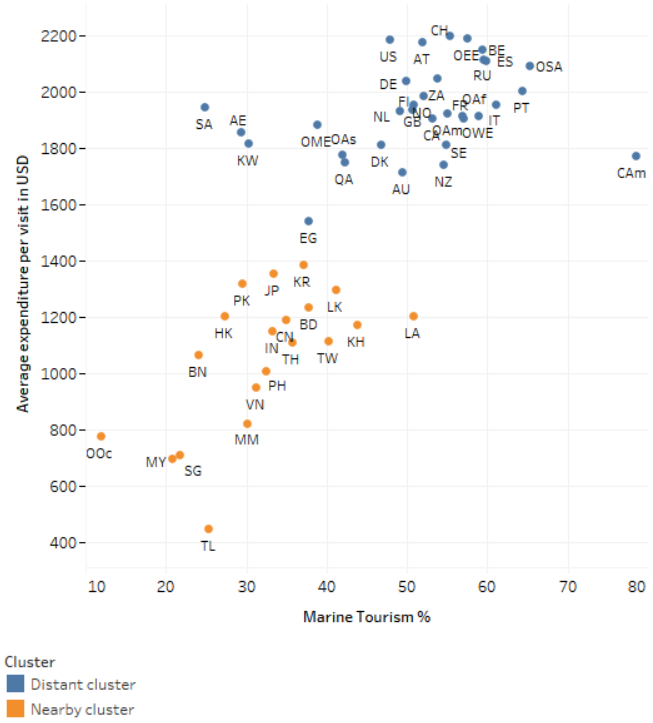

Fig. 4. Geographical mapping of two clusters.

prioritizing convenience and time. They take short trips with lower spending per visit but substantial cumulative value via repeat travel, visiting friends and relatives, and quick breaks. In contrast, those from the distant cluster exhibit lower visit numbers but higher spending per visit. This segment could be described as long-haul, purpose-driven visitors who prioritize experience depth, take fewer trips but stay longer, and prefer higher-priced accommodation, guided activities, and typical experiences.

These patterns align with the expectations of the gravity model: proximity reduces travel costs and time, facilitating higher tourist flows from nearby Asian countries, many of which share regional, cultural, or linguistic similarities with Indonesia. For example, Malaysia, Singapore, and Brunei share historical ties and a common linguistic root with Indonesia through the Malay language, making communication easier for travelers. Additionally, common religious practices, culinary preferences, and cultural values foster a sense of familiarity and comfort among tourists from neighboring ASEAN countries.

Conversely, tourists from more distant, higher-income countries, such as those in Europe or North America, face greater travel costs and time demands. This results in fewer trips but typically longer stays and higher per-visit expenditures. They usually choose Bali as their primary destination because of its strong international brand as a tropical paradise, offering beaches, resorts, cultural heritage, and wellness tourism. Thus, the gravity model helps explain both the spatial distribution and economic impact of foreign tourists to Indonesia, highlighting the importance of both market size and distance in shaping international tourism flows.

C. Recommendations

The Indonesian government should tailor its tourism marketing strategies to target each cluster more effectively. For the nearby cluster (Asian countries), campaigns should emphasize affordability, proximity, and ease of travel, with a focus on short stays and cultural experiences. In contrast, for the distant cluster (Europe, America, Australia, and Africa), marketing efforts should highlight luxury, extended stays, and high-value experiences, showcasing destinations like Bali, which attracts longer-stay, higher-spending tourists. The government could also focus on niche markets such as wellness tourism or ecotourism for European and Australian visitors, who are often drawn to sustainable and unique travel experiences.

As tourists from the distant cluster tend to spend more per visit, the government should develop and promote sustainable tourism initiatives that align with the preferences of high-spending, environmentally-conscious visitors. The promotion could include highlighting Indonesia's natural beauty and offering eco-friendly travel packages. Specific programs could target luxury and adventure travelers, such as yacht tourism, volcano treks, or cultural heritage tours. Promoting Bali's success as a leading sustainable destination could serve as a model for other regions to follow, ensuring that the influx of higher-income tourists does not come at the expense of the environment or local communities.

IV. Conclusion

This study addressed whether foreign tourists from various countries share similar or distinct characteristics that are useful for formulating marketing strategies. The study examined the inflow of foreign tourists to Indonesia by clustering countries of origin based on tourist occupation, visit purpose, activity, and expenditure per visit. The findings indicate two clusters of countries, named the nearby cluster and the distant cluster, each with distinct characteristics. The work contributes to the gravity model of tourism. The results suggest that the Indonesian government should customize its tourism marketing strategies to more effectively target each cluster based on its characteristics. These findings should be considered in the context of the methodology employed, which involved a secondary data analysis of a single year's information. Further study should consider analyzing data over multiple years to achieve more comprehensive insights.

References

[1] R. Garidzirai, "The role of international tourism on foreign trade in the BRICS nations," Cogent Soc. Sci., vol. 8, no. 1, 2022.
[2] M. Azizurrohman, R. B. Hartarto, Y.-M. Lin, and F. H. Nahar, "The Role of Foreign Tourists in Economic Growth: Evidence from Indonesia," J. Ekon. Stud. Pembang., vol. 22, no. 2, pp. 313–322, 2021.
[3] T. Hidayat, J. Damanik, Nopirin, and J. Soeprihanto, "Characteristics and Behaviors of Tourists: Case of Middle East Tourists in Puncak Cianjur, Indonesia, from Tour Guides' Perspective," in Advances in Social Science, Education and Humanities Research, 2019, pp. 326–329.
[4] C. Thongma and C. Chang, "Annals of Tourism Research Empirical Insights Drivers of multi-destination tourism in APEC (2008 – 2023)," Ann. Tour. Res. Empir. Insights, vol. 6, no. 2, p. 100194, 2025.

[5] B. McKercher, D. Tolkach, N. M. E. Mahadewi, and D. G. N. Byomantara, "Choosing the Optimal Segmentation Technique to Understand Tourist Behaviour," J. Vacat. Mark., vol. 29, no. 1, pp. 71–83, 2023.
[6] H. Hasanah, N. A. Sudibyo, and R. M. Galih, "Data Mining Using K-Means Clustering Algorithm for Grouping Countries of Origin of Foreign Tourist," in Basic and Applied Science Conference (BASC) 2021. NST Proceedings., 2021, pp. 88–94.
[7] X. Zhou and Z. Chen, "Destination attraction clustering: segmenting tourist movement patterns with geotagged information," Tour. Geogr., 2021.
[8] A. Jauhari, D. R. Anamisa, and F. A. Mufarroha, "Analysis of Clusters Number Effect Based on K-Means Method for Tourist Attractions Segmentation," J. Phys. Conf. Ser., vol. 2406, no. 1, pp. 0–7, 2022.
[9] J. Rosselló Nadal and M. Santana Gallego, "Gravity models for tourism demand modeling: Empirical review and outlook," J. Econ. Surv., vol. 36, no. 5, pp. 1358–1409, 2022.
[10] BPS-Statistics Indonesia, "International Visitor Arrivals Statistics 2024, Volume 16. BPS-Statistics Indonesia, 2025.
[11] F. Martinez-Plumed et al., "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," IEEE Trans. Knowl. Data Eng., vol. 33, no. 8, pp. 3048–3061, 2019.
[12] BPS-Statistics Indonesia, "International Visitors Expenditure Statistics 2024, Volume 6. BPS-Statistics Indonesia, 2025.
[13] S. Nirmal, "Comparative study between k-means and k-medoids clustering algorithms," Int. Res. J. Eng. Technol., vol. 839, pp. 839–844, 2019.
[14] P. Arora, Deepali, and S. Varshney, "Analysis of K-Means and K-Medoids Algorithm for Big Data," Procedia Comput. Sci., vol. 78, pp. 507–512, 2016.
[15] L. Xue and Y. Zhang, "The effect of distance on tourist behavior: A study based on social media data," Ann. Tour. Res., vol. 82, no. 102916, 2020.