JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 180-197
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.

Airport Clustering in Indonesia Based on Domestic Air Traffic Using the K-Means Algorithm

Mohammad Rohmad Nurkhoirofiq[1*], Rossi Passarella[2], Osvari Arsalan[3]
[1,2,3]Informatics Engineering, Sriwijaya University, Indonesia
[1*]morohmad27@gmail.com, [2]passarella.rossi@gmail.com, [3]osvari.arsalan@unsri.ac.id

Abstract

Indonesia, as an archipelagic country, relies heavily on air transportation to support interregional connectivity. Airports play an important role in supporting community mobility, economic activities, and equitable regional development. However, the level of air transportation activity at each airport in Indonesia varies and may change over time. This study aims to cluster airports in Indonesia based on domestic air transportation activity using the K-Means algorithm. The dataset used in this study consists of historical data from the period 2015–2023, including aircraft movements, passenger traffic, baggage volume, cargo volume, and mail shipments obtained from the HUBNET system of the Ministry of Transportation. The analysis process includes exploratory data analysis, data preprocessing, feature selection, and clustering modeling using the K-Means algorithm. The number of clusters is determined using the Elbow and Silhouette methods, resulting in candidate values of K = 2, 3, and 5. The evaluation results show that K = 2 has the best quality with a Silhouette Score of 0.6254, a Calinski Harabasz Index of 2635.59, and a Davies Bouldin Index of 0.5938. However, K = 3 is chosen as it is more representative, consisting of 260 low-activity airports, 1063 medium, and 627 high. Temporal analysis shows 233 cluster transitions across 123 airports, with the highest increase in 2016 (31 airports) and the highest decrease in 2018 (23 airports). The results indicate that airport activity is dynamic and that clustering can be used as a basis for data-driven airport planning and evaluation.

Keywords: Airport Clustering; K-Means; Air Transportation; Data Mining; Airport Activity

1. INTRODUCING

Airports function as gateways for economic activities that support growth, stability, and balanced development at both national and regional levels. In addition, airports enhance regional accessibility and play a significant role in supporting industrial, trade, and tourism sectors[1]. Airports also contribute to improving regional connectivity and facilitating economic interactions between regions. Their role in supporting mobility and strengthening transportation networks makes airports an important component of national infrastructure and regional development[2].

According to the *Air Transportation Statistics 2023*, there are 220 active airports operating under the management of PT Angkasa Pura I, PT Angkasa Pura II, Unit Penyelenggara Bandar Udara (UPBU), and Badan Usaha Bandar Udara (BUBU)[3]. The report indicates that during the period 2019–2023 the number of domestic aircraft movements decreased by an average of 4.88% per year, followed by a decline in passenger traffic by 4.75% per year and postal shipments by 43.73% per year, while cargo and baggage volumes increased by 0.93% and 3.03% per year, respectively[3]. These differences in activity trends indicate variations in the level of air transportation activity

Mohammad Rohmad Nurkhoirofiq: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 180-197
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.

among airports, highlighting the need for data-driven analysis to understand the characteristics and patterns of air transportation traffic at each airport in Indonesia.

Within the national air transportation system, airports have different roles and activity levels, functioning as hub airports, feeder airports, or regional airports, which are reflected in passenger volumes and flight movements[4]. These activities are influenced by several factors, including regional economic development, demand for air transportation, and macro-environmental conditions, causing airport roles and activity levels to change over time[5][6]. These changes indicate the importance of historical data-based analysis to objectively understand airport activity patterns.

In the transportation sector, several studies have applied data mining approaches, particularly the K-Means algorithm, to analyze marine transportation services by clustering ports as a basis for improving service planning[7]. Similar approaches have also been used to identify passenger density patterns at major airports in Indonesia and to group 21 airports based on operational and financial characteristics as a basis for adjusting organizational structures and airport resource management[8][9].

However, most studies in the transportation sector still employ a cross-sectional approach, which does not fully capture the patterns and characteristics of airports over time [10]. Meanwhile, the distribution of passenger flows and interregional connectivity patterns may change both spatially and temporally as air transportation systems evolve [11][12]. Therefore, further analysis is required to understand airport activity patterns in a more comprehensive manner.

This study performs clustering of airports in Indonesia based on domestic air transportation activity using the K-Means algorithm by utilizing historical data from 2015–2023, which include aircraft movements, passenger traffic, baggage volume, cargo volume, and postal shipments. The clustering results are expected to identify patterns of airport activity and the characteristics of each airport in Indonesia, which may serve as a basis for evaluation, mapping, and decision-making in airport management and planning.

2. METHODOLOGY

Figure 1 presents the research framework describing the stages of the airport clustering process in Indonesia using the K-Means algorithm.

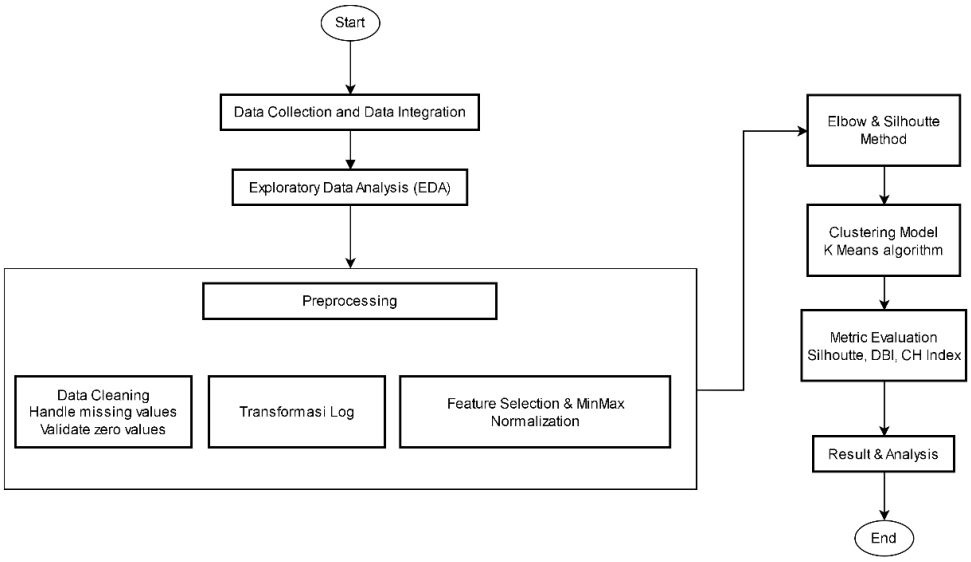


Figure 1. Research Framework

Mohammad Rohmad Nurkhoirofiq: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 180-197
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.

2. 1 Data Collection and Data Integration

The dataset used in this study was obtained from the HUBNET website, a transportation service integration system managed by the Ministry of Transportation of the Republic of Indonesia, which supports the digital transformation of the transportation sector[13]. The data consist of aircraft movements, passenger traffic, baggage volume, cargo volume, and postal shipments, which represent airport operational activities in domestic flights[13].

The data collection stage aims to obtain data or information that is relevant to the objectives of the study[14]. Initially, the data were available in separate files based on the type of air transportation traffic and the year of observation. Therefore, a data integration process was performed to combine these datasets in order to obtain more comprehensive information[15]. This process follows the structure of the air transportation statistics report, which categorizes air transportation traffic into aircraft movements, passenger traffic, baggage, cargo, and postal shipments[16]. After the integration process, the dataset consisted of eight variables, which are described as follows.

Table 1. Variable Description

Variabel	Data Type	Description
province	object	Province where the airport is located
airport_name	object	Name of the airport
Year	int64	Year of the observation
aircraft	float64	Number of domestic aircraft movement
passenger	float64	Number of domestic passengers
baggage	float64	Volume of domestic baggage transported
cargo	float64	Volume of domestic cargo transported
mail	float64	Volume of domestic mail and packages

2. 2 Exploratory Data Analysis (EDA)

At this stage, data exploration is conducted through statistical summaries and visualizations to identify patterns, relationships among variables, and the potential presence of extreme values or anomalies in the dataset. This process aims to provide an initial understanding of the data structure in order to determine appropriate data processing steps before further analysis is performed[17]. Skewness is a statistical measure used to indicate the degree of asymmetry in a data distribution relative to its mean. Meanwhile, kurtosis describes the thickness of the distribution tails, which is associated with the likelihood of extreme values occurring in the data[18].

2. 3 Preprocessing

Data preprocessing is an initial stage of data processing aimed at transforming raw data into data that are ready for modeling. This stage is performed to address issues such as inconsistencies, missing values, noise, and differences in scale among attributes[19]. This stage includes several processes such as data cleaning, log transformation, feature selection, data normalization, and adjustment of the data structure according to the requirements of the analysis algorithm Missing values, usually represented as NaN, null, or empty values, can reduce analysis quality and cause bias if not handled properly because many machine learning algorithms cannot process incomplete data[20]. Log transformation is used to stabilize data variance. The use of log(1+x) allows all data, including zero values, to be transformed without producing undefined values, making this method a commonly used approach in normalization and machine learning–based data analysis[21].

After the outlier handling process, feature selection was performed to identify the most relevant variables for clustering. Feature selection plays an important role in

Mohammad Rohmad Nurkhoirofiq: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 180-197
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.

producing models that are more robust, efficient, and have better generalization capability[22]. This is in line with the concept of sparsity, where only a small subset of features is truly relevant, while the remaining features can be ignored without significantly reducing model performance[23]. In addition, correlation analysis using the Pearson correlation coefficient was applied as part of the filter-based feature selection process[24]. Correlation analysis was conducted to identify relationships among features in the dataset. This statistical method measures the strength of linear relationships between variables using correlation coefficients ranging from -1 to 1 [25]. Although the dataset has no labels, correlation analysis remains relevant in unsupervised learning because the analysis is based on the intrinsic properties of the data, such as relationships and dependencies among features. Correlation can be used to identify associations between variables and detect potential information redundancy without requiring labeled data[26]. Data that have undergone preprocessing are then transformed into a more suitable format so that they can be used effectively in the subsequent analysis process[27].

2. 4 K-Means Clustering

K-Means is an unsupervised learning algorithm used to group data into several clusters based on feature similarity. The main objective of K-Means is to identify underlying structures within the data, reduce the complexity of the dataset, and produce meaningful segmentation, particularly for numerical features[28].

The distance between each data point and the centroid is calculated using the Euclidean distance as follows:

$$d(x_i, c_j) = \sqrt{\sum_{k=1}^{m} (x_{ik} - c_{jk})^2} \quad (1)$$

$d(x_i, c_j)$ = distance between the i-th data point and the j-th centroid
x_{ik} = value of the i-th data point on the k-th variable
c_{jk} = value of the j-th centroid on the k-th variable
m = number of variables

The cluster centroid is then updated by calculating the mean value of all data points within the cluster:

$$c_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i \quad (2)$$

c_j = centroid of cluster j
n_j = number of data points in cluster j
x_i = data point assigned to the cluster

2. 5 Elbow Method

The Elbow method is used to determine the optimal number of clusters in the K-Means algorithm by analyzing the value of the Sum of Squared Error (SSE) across different numbers of clusters. SSE is calculated from the squared distance between each data point and the centroid of its assigned cluster. The SSE value decreases as the number of clusters increases. However, at a certain point, the rate of decrease becomes less significant. This point is referred to as the elbow point and is considered an indication of the optimal number of clusters[29].

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} |x - c_i|^2 \quad (3)$$

k = number of clusters
x = data point

Mohammad Rohmad Nurkhoirofiq: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 180-197
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.

c_i = centroid of the i-th cluster
c_i = set of data points in the i-th cluster

2. 6 Silhouette

The Silhouette method evaluates clustering quality by comparing the similarity of data points within the same cluster with their dissimilarity to the nearest neighboring cluster. The Silhouette value ranges from −1 to 1, where a higher positive value indicates that the data point is well matched to its own cluster and poorly matched to neighboring clusters. Conversely, negative values indicate a potential misclassification of the data point. A higher average Silhouette value indicates a better clustering structure[30].

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (4)$$

$S(i)$ = silhouette value of the i-th data point
$a(i)$ = average distance between the i-th data point and other points within the same cluster
$b(i)$ = average distance between the i-th data point and the nearest neighboring cluster

2. 7 Davies-Bouldin Index

The Davies-Bouldin Index (DBI) is an internal evaluation method used to measure the quality of clustering results by considering the compactness within clusters (*intra-cluster similarity*) and the separation between clusters (*inter-cluster separation*). A lower DBI value indicates that the clusters have high internal similarity and clear separation from other clusters, resulting in better clustering quality[29].

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{S_i + S_j}{M_{ij}}\right) \quad (5)$$

k = number of clusters
S_i = average distance of all data points in cluster i to centroid i
S_j = average distance of all data points in cluster j to centroid j
M_{ij} = distance between the centroids of clusters i and j

2. 8 Calinski-Harabasz Index

The Calinski-Harabasz Index (CH) is used to evaluate the quality of clustering results by comparing the variation between clusters (*between-cluster dispersion*) and the variation within clusters (*within-cluster dispersion*). The CH value is calculated as the ratio between the Sum of Squares Between (SSB) and the Sum of Squares Within (SSW), normalized by the number of data points and the number of clusters. A higher CH value indicates better clustering quality, as it reflects greater separation between clusters and higher compactness within clusters[31].

$$CH = \frac{SSB}{SSW} \times \frac{n-k}{k-1} \quad (6)$$

SSB = between-cluster variation
SSW = within-cluster variation
n = number of data points
k = number of clusters

3. RESULT AND DISCUSSIONS

3. 1 Exploratory Data Analysis (EDA)

Based on the data structure inspection, the dataset consists of 1,950 observations and eight variables, including two categorical variables—province and airport name—and six numerical variables consisting of the observation year and five indicators of air transportation traffic. Several numerical variables contain missing values, specifically 73

Mohammad Rohmad Nurkhoirofiq: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 180-197
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.

values in the aircraft variable, 42 in passengers, 49 in baggage, 48 in cargo, and 39 in mail, resulting in a total of 251 missing values in the dataset. In addition to missing values, zero values were also identified in several air transportation traffic variables. The mail variable has the highest percentage of zero values at 77.33%, followed by cargo (31.85%), baggage (11.23%), passengers (7.90%), and aircraft movements (5.23%).

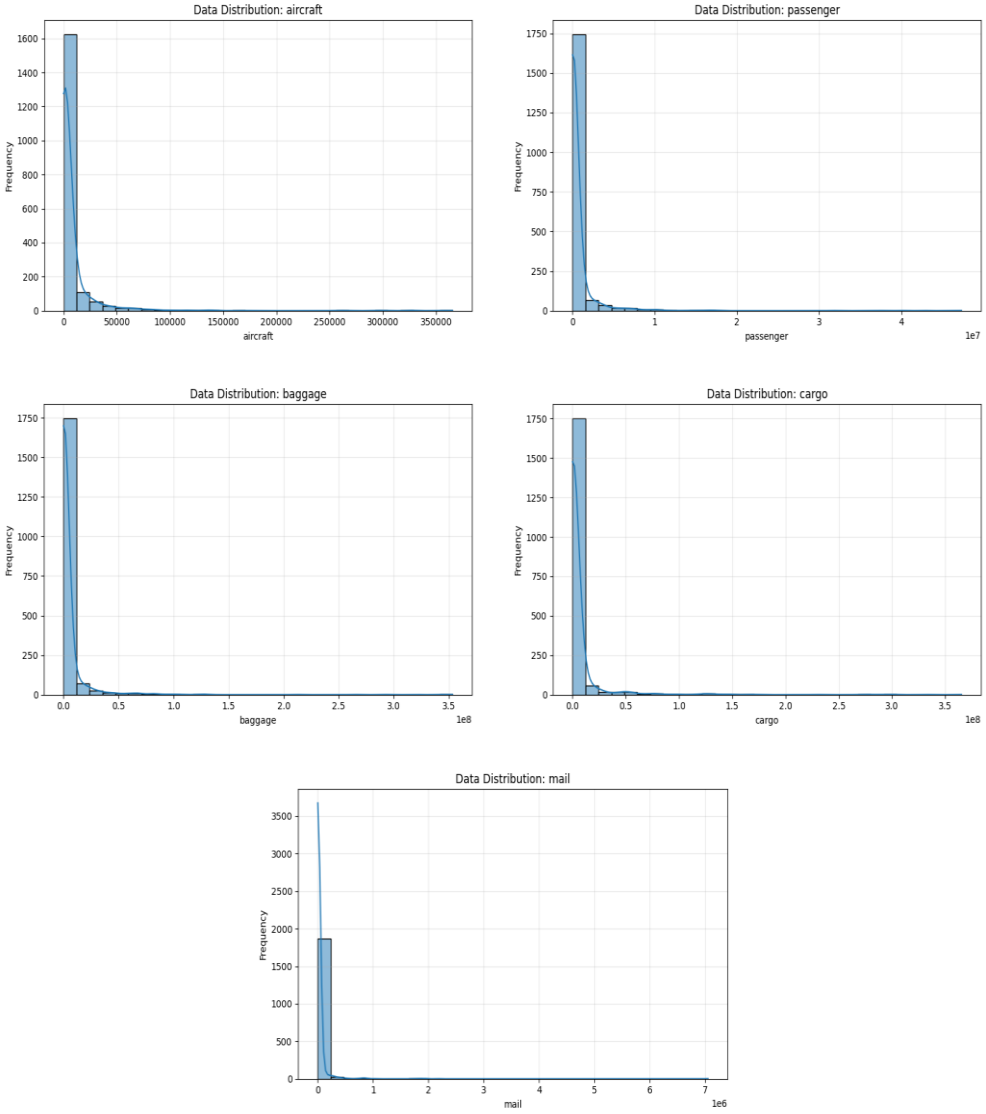


Figure 2. Distribution of Data Variables

The histogram results show that all variables tend to have right-skewed distributions. Most airports are concentrated within lower activity ranges, while only a few airports record very high activity levels, forming long tails on the right side of the distributions. The aircraft and passenger variables indicate that air transportation activity is unevenly distributed across airports. Similar patterns are also observed in the baggage and cargo variables, where the majority of airports have relatively low activity volumes compared to a small number of highly active airports.

In contrast, the mail variable exhibits the highest level of skewness among all variables. Most observations are concentrated near very low values, and many contain zero values, indicating that postal shipment activity is generally low and unevenly distributed

Mohammad Rohmad Nurkhoirofiq: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 180-197
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.

across airports. Overall, these distribution patterns reflect disparities in airport activity levels, highlighting the need for preprocessing before clustering in order to reduce the influence of extreme values and improve clustering stability.

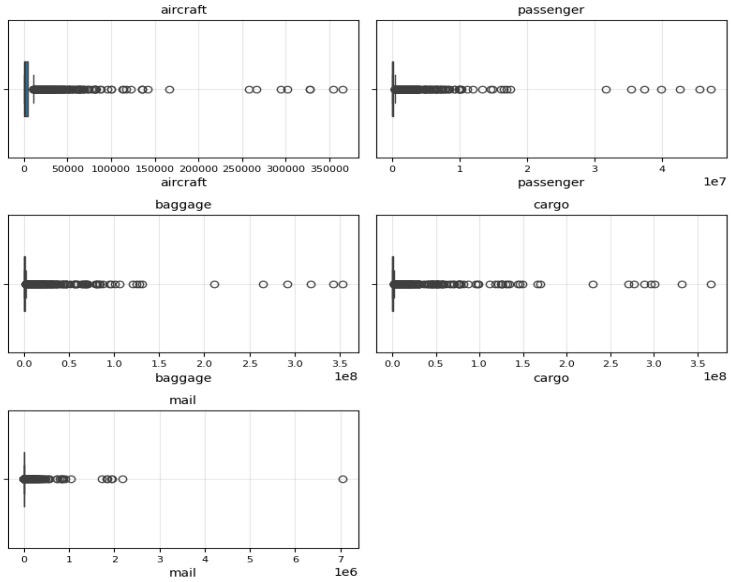


Figure 3. Boxplot Outlier Visualization

To identify the presence of extreme values in each variable, boxplot visualizations were performed. The results show that all air transportation traffic variables contain outliers on the right side of the distributions, indicating that several airports have substantially higher activity levels compared to most other airports. This pattern is observed in the aircraft, passenger, baggage, and cargo variables, while the mail variable exhibits the most extreme pattern because most observations are concentrated at very low or zero values, with only a few observations showing very high values. Outlier detection using the Interquartile Range (IQR) method indicates that the aircraft variable contains 271 outliers (13.90%), passengers 336 (17.23%), baggage 357 (18.31%), cargo 395 (20.26%), and mail 403 (20.67%). The high proportion of outliers suggests that the data distributions are uneven or skewed, which is a common characteristic of air transportation data due to significant differences in airport activity levels, meaning that the detected outliers do not necessarily represent data errors but rather reflect actual variations in airport operational activities.

Table 2. Skewness and Kurtosis

Variabel	*Skewness*	*Kurtosis*
aircraft	8.80	99.12
passenger	10.70	140.98
baggage	10.94	151.62
cargo	8.76	93.95
mail	22.98	704.19

The skewness and kurtosis calculations presented in Table 2 indicate that all variables have high positive skewness values, ranging from 8.76 to 22.98. These values indicate that the data distribution is not symmetrical and tends to be skewed to the right. In addition, the very high kurtosis values suggest the presence of extreme differences among data points, reflecting a significant imbalance in activity levels across airports.

Mohammad Rohmad Nurkhoirofiq: *Corresponding Author

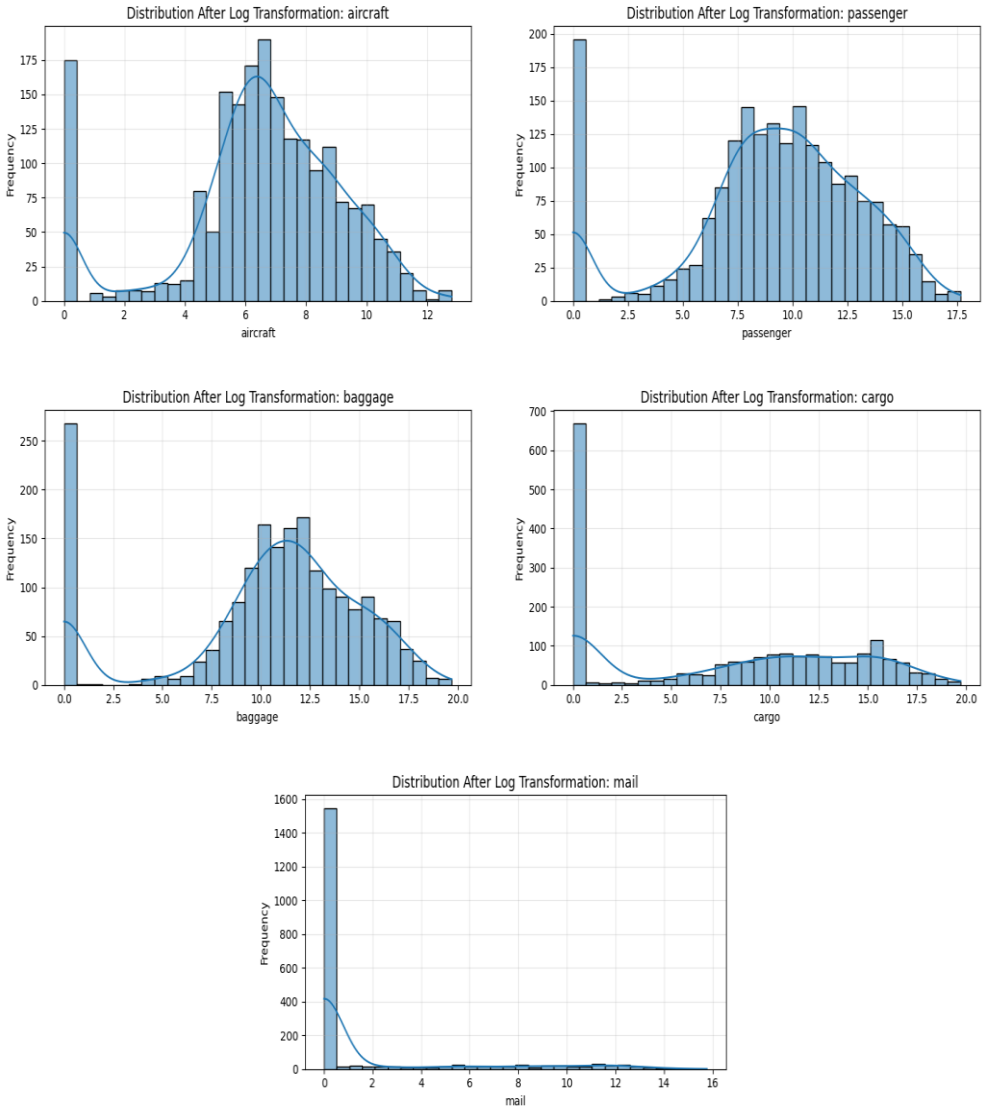
JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 180-197
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.

3. 2 Preprocessing

The preprocessing stage was conducted based on the results of exploratory data analysis (EDA), which revealed substantial differences in data scales and distributions that tended to be imbalanced or right-skewed. Therefore, several preprocessing steps were performed, including missing value handling, logarithmic transformation, feature selection, and data normalization using MinMaxScaler.

Missing values were handled contextually based on each airport and observation year. Missing values occurring before or after the operational period of an airport were filled with 0, representing periods with no recorded activity. Meanwhile, missing values located between observation periods were imputed using the median value of the corresponding variable for the same airport in order to preserve the characteristic patterns of each airport's data.

After the data imputation process, a log transformation was applied to the aircraft, passenger, baggage, cargo, and mail variables using the log(1 + x) function. This transformation aims to reduce the skewness of the data distribution and minimize the influence of extremely large values.



Mohammad Rohmad Nurkhoirofiq: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 180-197
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.

Figure 4. Distribution of Variables After Log Transformation

The histogram results after the log transformation show that the distributions of the aircraft, passenger, and baggage variables become more balanced compared to their distributions before transformation. In contrast, the cargo and mail variables still exhibit a concentration of values within the lower range, indicating that most observations remain clustered at relatively low activity levels.

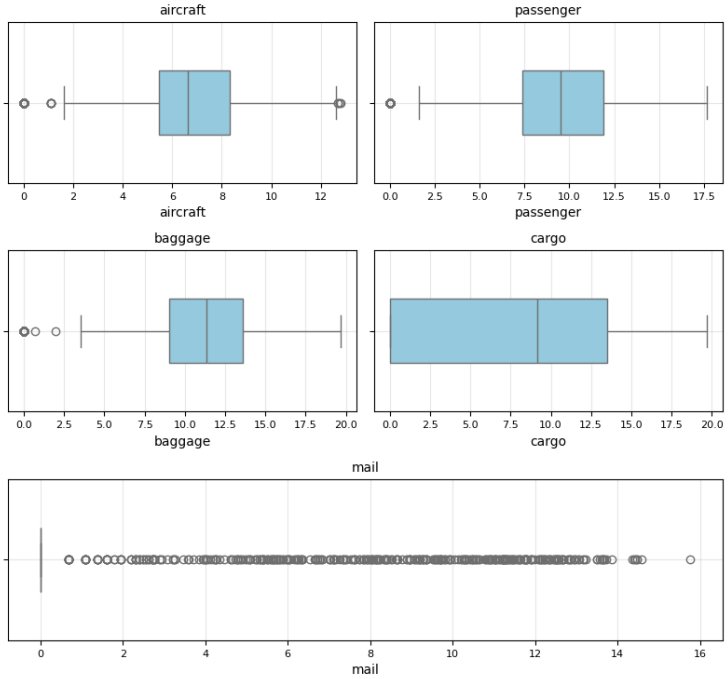


Figure 5. Boxplot Outlier After Log Transformation

The boxplot analysis results show that after the log transformation, outliers are still present but are no longer dominant, and the data distribution becomes more concentrated. This condition indicates that the log transformation is able to reduce the influence of extreme values without removing important data. This finding is consistent with previous studies stating that log transformation can compress large values so that they do not dominate the analysis [32]. Therefore, the outliers were not removed because they were considered to represent actual conditions, particularly for airports with high activity levels

Table 3 Skewness and Kurtosis After Log Transformation

Variabel	*Skewness*	*Kurtosis*
aircraft	-0.820	0.652
passenger	-0.737	0.237
baggage	-0.953	0.229
cargo	-0.059	-1.511
mail	2.151	3.215

Further distribution analysis was conducted using skewness and kurtosis values. The results show that the skewness values of the aircraft (-0.820), passenger (-0.737), baggage (-0.953), and cargo (-0.059) variables are close to zero, indicating that the data distributions become more balanced after the log transformation. In contrast, the mail variable still has a skewness value of 2.151, indicating that the distribution remains right-skewed.

Mohammad Rohmad Nurkhoirofiq: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 180-197
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.

The kurtosis values of the aircraft (0.652), passenger (0.237), and baggage (0.229) variables are also close to zero, suggesting a more stable distribution shape after the log transformation. The cargo variable has a kurtosis value of -1.511, indicating a flatter distribution peak, while the mail variable has a kurtosis value of 3.215, which suggests the presence of relatively extreme values. Overall, the log transformation helps improve the data distribution, making it more stable for further analysis.

The outlier detection results after applying the log(1+x) transformation show a reduction in the number of outliers for most variables, with aircraft decreasing to 185 data points (9.49%), passengers to 196 (10.05%), baggage to 270 (13.85%), and cargo to 0 (0.00%), while the mail variable remained at 403 outliers (20.67%). This condition indicates that the distribution of the mail variable is still highly uneven. Therefore, further treatment was performed using the Interquartile Range (IQR)-based winsorization method by limiting extreme values within the lower and upper boundaries defined by Q1 − 1.5×IQR and Q3 + 1.5×IQR without removing the data [33]. The winsorization results show that the number of outliers in all variables became 0 (0%), indicating that the extreme values were successfully controlled by adjusting them to the specified boundaries without eliminating important data

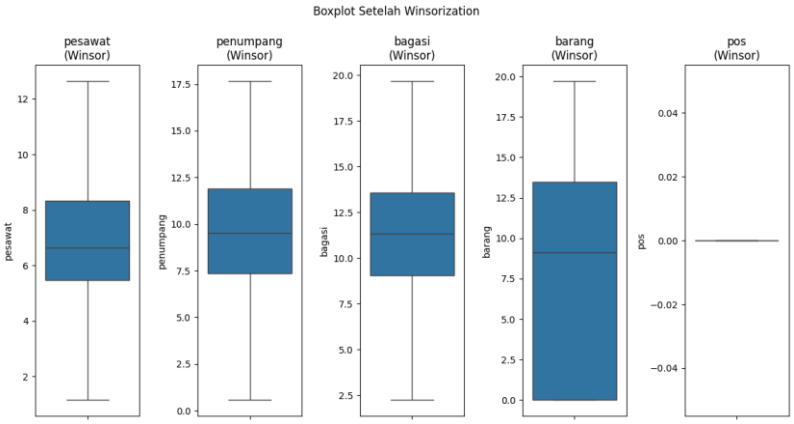


Figure 6. Winsorization Boxplot Visualization

The boxplot visualization after winsorization shows that no points remain outside the whiskers and the data distribution becomes more stable. However, the mail variable still exhibits a very narrow distribution because many extreme values were limited to the same range. These results indicate that winsorization is effective in reducing the influence of outliers without removing data.

In high dimensional data, features with high percentages of zero values may indicate sparsity characteristics that can reduce model effectiveness, making feature selection necessary [23]. The analysis was conducted by calculating the percentage of zero values in each variable. The results show that the aircraft variable contains 8.97% zero values, passengers 10.05%, and baggage 13.74%, indicating that these variables still have sufficient data variation. Meanwhile, the cargo variable contains 34.31% zero values and the mail variable 79.33%, indicating that most observations in these variables are zero. The high percentage of zero values reflects sparsity characteristics that may reduce the effectiveness of the clustering process.

Mohammad Rohmad Nurkhoirofiq: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 180-197
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.

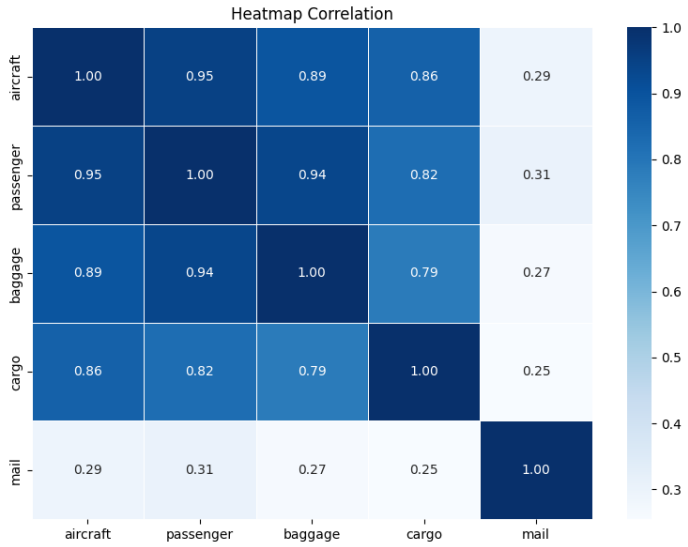


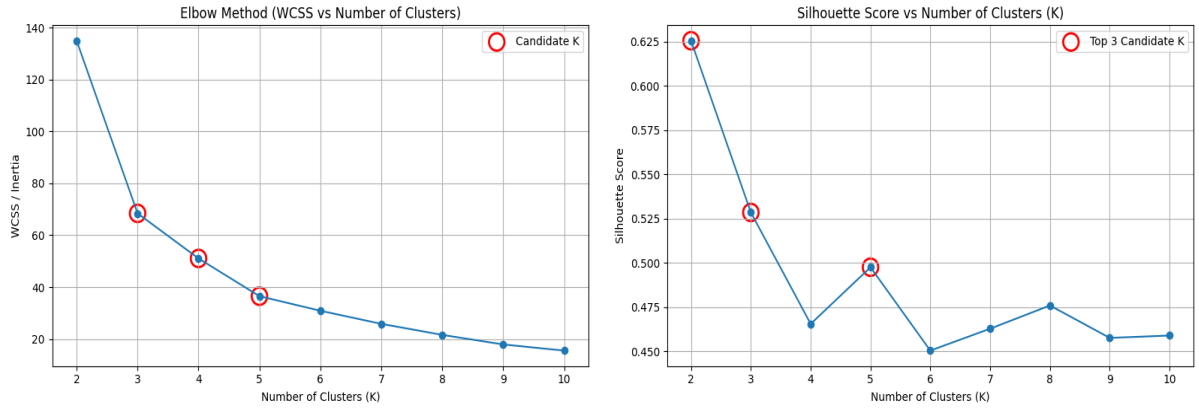
Figure 7. Heatmap Correlation

In addition to the zero-value analysis, Pearson correlation coefficient analysis was performed to examine relationships among variables. The correlation results indicate that most variables have strong positive relationships. The aircraft variable has correlations of 0.95 with passengers, 0.89 with baggage, and 0.86 with cargo. The passenger variable also shows high correlations with baggage (0.94) and cargo (0.82), while baggage and cargo have a correlation of 0.79. In contrast, the mail variable exhibits relatively low correlations with the other variables, ranging from 0.25 to 0.31, indicating that postal shipment activity is not strongly associated with other air transportation activities.

Based on the feature selection results, the cargo and mail variables were removed from the dataset because they contain high percentages of zero values, while the mail variable also shows relatively low correlations with the other variables. Therefore, both variables were considered less informative for the clustering process using the K-Means algorithm.

The normalization stage was performed using the Min-Max Normalization method through MinMaxScaler to transform the value range of each variable into a scale between 0 and 1. This process aims to minimize scale differences among variables so that each variable contributes more equally to the distance calculation in the K-Means algorithm and does not dominate the clustering process.

3. 3 Elbow and Silhouette Method



Mohammad Rohmad Nurkhoirofiq: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 180-197
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.

Figure 8. Elbow and Silhouette Values

Based on the Elbow Method graph in Figure 8, the Within-Cluster Sum of Squares (WCSS) value decreases sharply from K = 2 to K = 4, and the rate of decrease begins to level off after K = 5. This pattern indicates that increasing the number of clusters beyond this point does not result in a significant reduction in WCSS. Therefore, the values of K around this point can be considered as potential candidates for the optimal number of clusters.

Further evaluation was performed using the Silhouette Score to measure the quality of separation between clusters. The results show that the highest Silhouette value is obtained at K = 2 with a score of 0.625, followed by K = 3 with 0.528, K = 5 with 0.497, and K = 8 with 0.476.

Based on the combination of these two methods, K = 2, K = 3, and K = 5 are considered potential candidates for the optimal number of clusters, as they correspond to the points where the WCSS reduction is still significant and where the Silhouette values remain relatively higher compared to other values of K. The K-Means algorithm was then applied using these cluster values to obtain clustering results and evaluate the characteristics of the clusters formed.

3. 4 Metric Evaluation

The comparison of clustering quality across several values of K was conducted using three evaluation metrics: Silhouette Score, Davies–Bouldin Index (DBI), and Calinski–Harabasz Index (CHI). The evaluation results are presented in Table 4.

Table 4. Metric Evaluation

K	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
2	0.6254	0.5938	2635.59
3	0.5284	0.6061	3543.93
5	0.4975	0.7124	3742.02

Based on the evaluation results, K = 2 has the highest Silhouette Score (0.6254) and the lowest Davies–Bouldin Index (0.5938), indicating the best cluster separation. However, K = 3 still demonstrates good clustering quality with a Silhouette Score of 0.5284 and a Davies–Bouldin Index of 0.6061.

Although K = 2 provides more optimal metric values, this study uses K = 3 because it provides a more detailed grouping of airports, allowing variations in air transportation activity levels to be represented more clearly compared to using only two clusters.

3. 5 Cluster Result Analysis

Mohammad Rohmad Nurkhoirofiq: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 180-197
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.

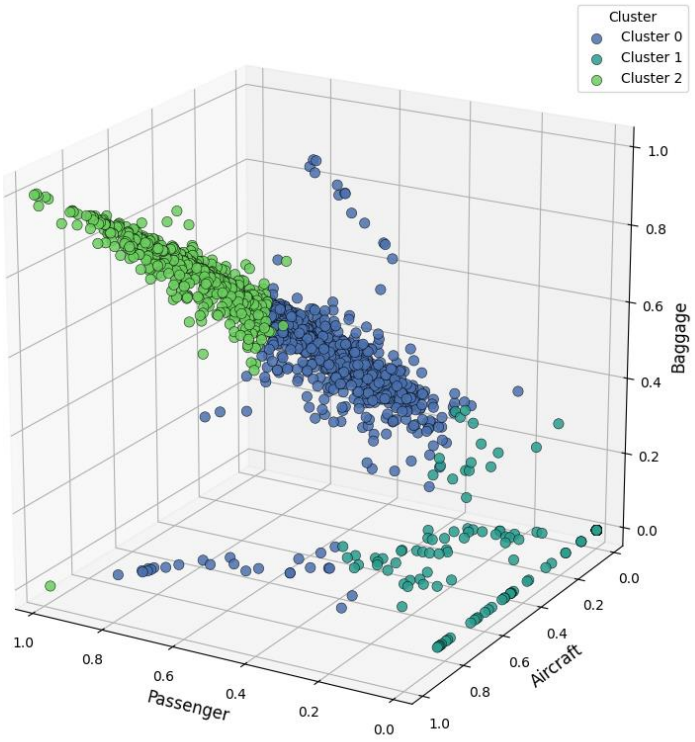


Figure 9. 3D PCA Visualization of K-Means Clustering

The PCA visualization in Figure 9 shows that the three clusters exhibit distinct distribution patterns within the three dimensional principal component space. Cluster 0 forms the largest group and is concentrated around the center of the distribution, while Cluster 1 appears more dispersed on the left side of the plot. In contrast, Cluster 2 forms a denser group on the right side of the plot. This pattern indicates that the K-Means algorithm successfully separates airport data into clusters with different activity characteristics.

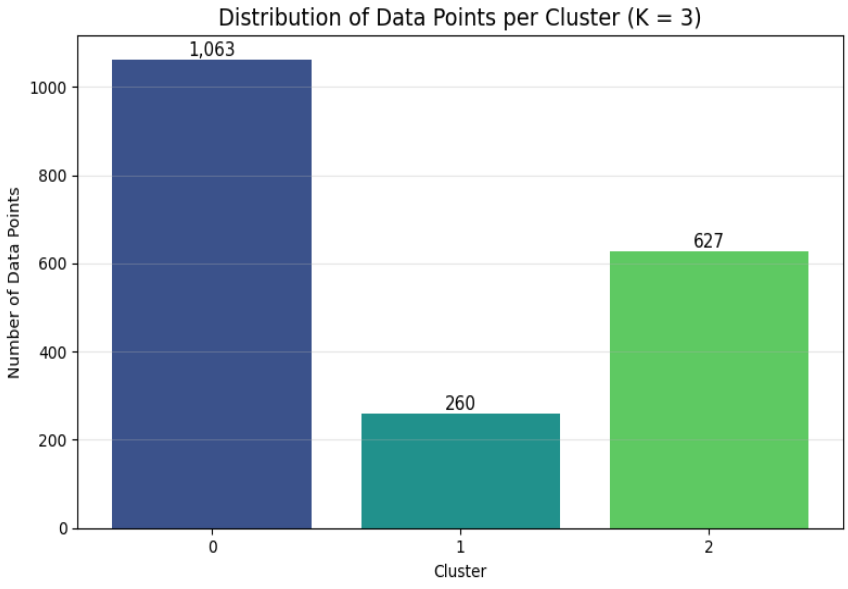


Figure 10. Cluster Data Distribution

Mohammad Rohmad Nurkhoirofiq: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 180-197
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.

The distribution of data across clusters shows that Cluster 0 contains the largest number of airports with 1,063 observations, followed by Cluster 2 with 627 observations, and Cluster 1 with 260 observations. These results indicate that the majority of airports share relatively similar activity characteristics and are grouped within Cluster 0.

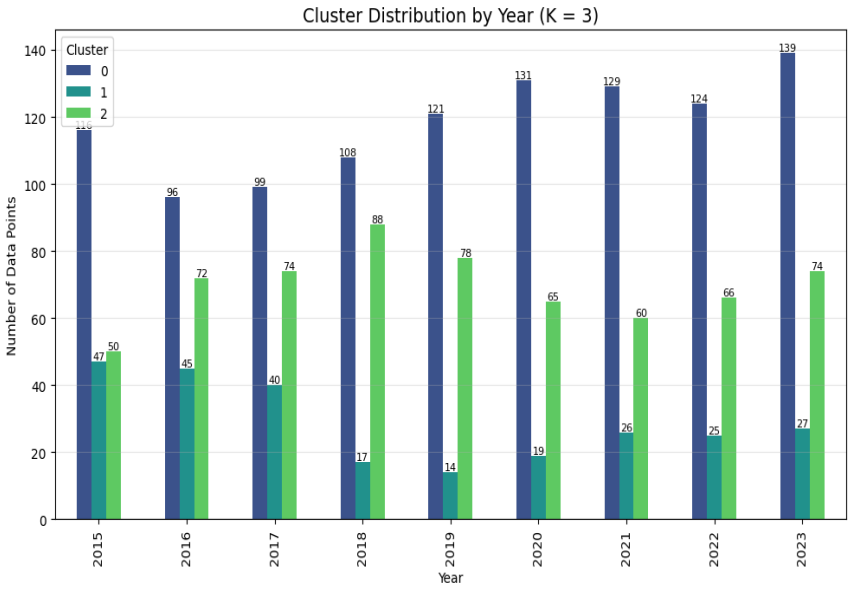


Figure 11. Cluster Distribution by Year

The yearly cluster distribution shows that Cluster 0 consistently contains the highest number of airports across all observation years. The highest number in this cluster occurs in 2023 with 139 airports, while the lowest occurs in 2016 with 96 airports. In Cluster 1, the number of airports is relatively smaller compared to the other clusters. The highest number occurs in 2015 with 47 airports, while the lowest occurs in 2019 with 14 airports. Meanwhile, Cluster 2 represents a moderate level of airport activity. The highest number in this cluster occurs in 2018 with 88 airports, while the lowest occurs in 2015 with 50 airports. These results indicate that airport activity levels vary across years, reflecting changes in operational patterns within the Indonesian air transportation system.

3. 6 Airport Cluster Pattern Analysis (2015–2023)

The K-Means clustering results with K = 3 indicate that airports in Indonesia can be grouped into three categories based on the level of air transportation activity: low, medium, and high activity levels. This classification is based on the average values of the aircraft, passenger, and baggage variables within each cluster.

Table 5. Cluster Characteristics Based on Average Values

Cluster	aircraft	passenger	baggage	Description
1	0.142	0.066	0.017	Low
0	0.478	0.492	0.517	Medium
2	0.711	0.742	0.761	High

The calculation results show that Cluster 1 has the lowest average values, with aircraft = 0.142, passengers = 0.066, and baggage = 0.017, indicating airports with low activity levels. Cluster 0 has average values of aircraft = 0.478, passengers = 0.492, and baggage = 0.517, representing airports with moderate air transportation activity.

Mohammad Rohmad Nurkhoirofiq: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 180-197
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.

Meanwhile, Cluster 2 has the highest average values, with aircraft = 0.711, passengers = 0.742, and baggage = 0.761, representing airports with high air transportation activity.

After identifying the characteristics of each cluster, further analysis was conducted to examine the airport cluster patterns from year to year during the period 2015–2023. The results indicate that several airports experienced cluster transitions, reflecting increases or decreases in air transportation activity levels.

Table 6. Airport Cluster Transitions During the Observation Period

Cluster Transition	Number of Airports
0 → 1	49
0 → 2	62
1 → 0	70
1 → 2	6
2 → 0	46

Overall, 233 cluster transitions were recorded involving 123 airports. The most frequent transition occurred from low activity to medium activity (1 → 0), involving 70 airports, followed by transitions from medium to high activity (0 → 2) involving 62 airports. On the other hand, 49 airports moved from medium activity to low activity (0 → 1), while 46 airports moved from high activity to medium activity (2 → 0). Direct transitions from low activity to high activity (1 → 2) occurred in 6 airports.

Table 7. Airport Cluster Changes per Year (2015–2023)

Year	Increase	Decrease	Total
2015	0	0	0
2016	31	9	40
2017	15	18	33
2018	18	23	41
2019	7	20	27
2020	9	19	28
2021	13	11	24
2022	10	5	15
2023	14	11	25

When examined annually, the cluster pattern shows varying dynamics. The highest increase in cluster level occurred in 2016, when 31 airports experienced an increase in activity level, while the largest decrease occurred in 2018, when 23 airports experienced a decline in activity level. Overall, these results indicate that airport activity levels in Indonesia change from year to year, which is reflected in cluster transitions among several airports.

4. CONCLUSION

This study aims to cluster airports in Indonesia based on domestic air transportation activity using the K-Means algorithm. The optimal number of clusters was evaluated using the Elbow and Silhouette methods, resulting in K = 2, K = 3, and K = 5 as potential cluster candidates. Although K = 2 produced the best evaluation metrics with the highest Silhouette Score (0.6254) and the lowest Davies–Bouldin Index (0.5938), K = 3 was selected because it provides a more interpretable representation of airport activity variations.

The clustering results indicate that airports in Indonesia can be grouped into three activity categories: low, medium, and high activity levels. Most airports are classified within the medium activity cluster, while the high activity cluster contains the fewest airports.

Mohammad Rohmad Nurkhoirofiq: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 180-197
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.

Temporal analysis during the 2015–2023 period also reveals dynamic changes in airport activity, reflected by 233 cluster transitions involving 123 airports. The highest increase in activity occurred in 2016, while the largest decrease occurred in 2018. These findings demonstrate that airport activity in Indonesia changes over time and highlight the importance of historical-data-based clustering in understanding airport activity patterns. The results of this study can serve as a reference for airport evaluation, transportation planning, and data-driven decision-making related to air transportation development in Indonesia.

For future research, it is recommended to incorporate additional variables related to airport operations and regional characteristics, as well as to utilize more recent datasets to obtain a more comprehensive analysis. Future studies may also compare different clustering approaches to achieve more optimal clustering results.

5. REFERENCES

[1] Kementerian Perhubungan Republik Indonesia, *Laporan Kinerja Kementerian Perhubungan*. 2023. Accessed: Nov. 10, 2025. [Online]. Available: https://ppid.dephub.go.id/fileupload/informasi-setiap-saat/20241021185227.LKIP_Kemenhub_Tahun_2023.pdf

[2] "Optimalisasi Pemanfaatan Transportasi Udara Bandara Aek Godang dan Bandara Jenderal Besar AH Nasution untuk di Sumatera Utara | Jurnal Informatika Ekonomi Bisnis." Accessed: Jan. 10, 2026. [Online]. Available: https://infeb.org/index.php/infeb/article/view/1080

[3] Badan Pusat Statistik, "Statistik Transportasi Udara 2023," 2023, Accessed: Nov. 10, 2025. [Online]. Available: https://www.bps.go.id/id/publication/2024/11/25/a8fac4f872ffc88d3f45d0d3/statistik-transportasi-udara-2023.html

[4] J. Pauwels, S. Buyle, and W. Dewulf, "Insights in diversity: A latent profile cluster analysis of regional airports in Western Europe," *Research in Transportation Business & Management*, vol. 63, p. 101477, Dec. 2025, doi: 10.1016/J.RTBM.2025.101477.

[5] L. Wan *et al.*, "Life Cycle Prediction of Airport Operation Based on System Dynamics," *Sustainability 2024, Vol. 16, Page 9596*, vol. 16, no. 21, p. 9596, Nov. 2024, doi: 10.3390/SU16219596.

[6] L. E. C. Rocha, "Dynamics of air transport networks: A review from a complex systems perspective," *Chinese Journal of Aeronautics*, vol. 30, no. 2, pp. 469–478, Apr. 2017, doi: 10.1016/J.CJA.2016.12.029.

[7] R. Passarella, R. Febrianti, M. Vindriani, and M. S. Adenan, "A Clustering Approach to Enhancing Marine Transportation Services in the Thousand Islands Regency, Indonesia," *International Journal of Transport Development and Integration*, vol. 9, no. 1, pp. 131–141, Mar. 2025, doi: 10.18280/ijtdi.090112.

[8] S. Siti Syaoqiyah, Y. Selvina Yulianti, and N. Ayu Rahmadenti, "Classification of Domestic Flight Passengers at Main Airports Using the K-Means Clustering Method," *International Journal of Information System & Technology Akreditasi*, vol. 8, no. 158, pp. 1–5, 2024, [Online]. Available: https://www.bps.go.id/

[9] D. L. Trenggonowati, M. Ulfah, R. Ekawati, and V. A. Yusuf, "Organization clustering airports using K-Means clustering algorithm," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 673, no. 1, Dec. 2019, doi: 10.1088/1757-899X/673/1/012081.

[10] Y. Gao, "What is the busiest time at an airport? Clustering U.S. hub airports based on passenger movements," *J. Transp. Geogr.*, vol. 90, p. 102931, Jan. 2021, doi: 10.1016/J.JTRANGEO.2020.102931.

[11] C. Luo, T. Ma, M. Wang, X. Hu, C. Liu, and X. Wang, "Analysis of the temporal and spatial evolution of China's air passenger transportation and the impact of policies

Mohammad Rohmad Nurkhoirofiq: *Corresponding Author

on passenger flows," *PLoS One*, vol. 20, no. 5, p. e0323482, May 2025, doi: 10.1371/JOURNAL.PONE.0323482.
[12] L. Siozos-Rousoulis, D. Robert, and W. Verbeke, "A study of the U.S. domestic air transportation network: temporal evolution of network topology and robustness from 2001 to 2016," *Journal of Transportation Security*, vol. 14, no. 1–2, p. 55, Jun. 2021, doi: 10.1007/S12198-020-00227-X.
[13] "HUBNET - KEMENHUB - Hubnet." Accessed: Jan. 15, 2026. [Online]. Available: https://hubnet.kemenhub.go.id/
[14] H. Rigneault *et al.*, "Data Collection Methods and Tools for Research; A Step-by-Step Guide to Choose Data Collection Technique for Academic and Business Research Projects," *International Journal of Academic Research in Management (IJARM)*, vol. 10, no. 1, pp. 10–38, Sep. 2021, doi: 10.34894/VQ1DJA.
[15] J. L. Ballard, Z. Wang, W. Li, L. Shen, and Q. Long, "Deep learning-based approaches for multi-omics data integration and analysis," *BioData Mining 2024 17:1*, vol. 17, no. 1, pp. 38-, Oct. 2024, doi: 10.1186/s13040-024-00391-z.
[16] "statistik-transportasi-udara-2023".
[17] S. Dhummad, "The Imperative of Exploratory Data Analysis in Machine Learning," *Scholars Journal of Engineering and Technology Abbreviated Key Title: Sch J Eng Tech*, vol. 13, no. 1, pp. 30–44, 2025, doi: 10.36347/sjet.2025.v13i01.005.
[18] J. H. T. Kim and H. Kim, "Estimating Skewness and Kurtosis for Asymmetric Heavy-Tailed Data: A Regression Approach," *Mathematics*, vol. 13, no. 16, p. 2694, Aug. 2025, doi: 10.3390/MATH13162694/S1.
[19] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, Jun. 2022, doi: 10.1016/j.gltp.2022.04.020.
[20] M. K. Hasan, M. A. Alam, S. Roy, A. Dutta, M. T. Jawad, and S. Das, "Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021)," *Inform. Med. Unlocked*, vol. 27, p. 100799, Jan. 2021, doi: 10.1016/J.IMU.2021.100799.
[21] A. Sina Booeshaghi and L. Pachter, "Normalization of single-cell RNA-seq counts by log(x + 1) or log(1 + x)," *Bioinformatics*, vol. 37, no. 15, pp. 2223–2224, Aug. 2021, doi: 10.1093/BIOINFORMATICS/BTAB085.
[22] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4, pp. 1060–1073, Apr. 2022, doi: 10.1016/J.JKSUCI.2019.06.012.
[23] M. Fira, L. Goras, and H. N. Costin, "Evaluating Sparse Feature Selection Methods: A Theoretical and Empirical Perspective," *Applied Sciences 2025, Vol. 15, Page 3752*, vol. 15, no. 7, p. 3752, Mar. 2025, doi: 10.3390/APP15073752.
[24] G. I. Hapsari, R. Munadi, B. Erfianto, and I. D. Irawati, "Feature Selection Using Pearson Correlation for Ultra-Wideband Ranging Classification," *Jurnal RESTI*, vol. 9, no. 2, pp. 209–217, Apr. 2025, doi: 10.29207/RESTI.V9I2.6281.
[25] F. V. P. Samosir, L. P. Mustamu, E. D. Anggara, A. I. Wiyogo, and A. Widjaja, "Exploratory Data Analysis terhadap Kepadatan Penumpang Kereta Rel Listrik," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 7, no. 2, pp. 449-467–449 – 467, Aug. 2021, doi: 10.28932/JUTISI.V7I2.3700.
[26] C. A. Rickert, M. Henkel, and O. Lieleg, "An efficiency-driven, correlation-based feature elimination strategy for small datasets," *APL Machine Learning*, vol. 1, no. 1, Mar. 2023, doi: 10.1063/5.0118207.
[27] V. Çetin and O. Yildiz, "A comprehensive review on data preprocessing techniques in data analysis," vol. 28, no. 2, pp. 299–312, 2022, doi: 10.5505/pajes.2021.62687.
[28] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

Mohammad Rohmad Nurkhoirofiq: *Corresponding Author

JURNAL TEKNOINFO
Volume 20, Number 2, July 2026, Page 180-197
E-ISSN 2615-224X P-ISSN 1693-0010
https://publikasi.teknokrat.ac.id/index.php/teknoinfo/index
DOI: https://doi.org/10.33365/teknoinfo.

[29] I. K. Khan *et al.*, "Determining the optimal number of clusters by Enhanced Gap Statistic in K-mean algorithm," *Egyptian Informatics Journal*, vol. 27, p. 100504, Sep. 2024, doi: 10.1016/j.eij.2024.100504.
[30] M. Shutaywi and N. N. Kachouie, "Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering," *Entropy 2021, Vol. 23, Page 759*, vol. 23, no. 6, p. 759, Jun. 2021, doi: 10.3390/E23060759.
[31] Z. Syahputri, S. Sutarman, and M. A. P. Siregar, "Determining The Optimal Number of K-Means Clusters Using The Calinski Harabasz Index and Krzanowski and Lai Index Methods for Groupsing Flood Prone Areas In North Sumatra," *Sinkron*, vol. 9, no. 1, pp. 571–580, Jan. 2024, doi: 10.33395/sinkron.v9i1.13246.
[32] Y. Zhang *et al.*, "Review and revamp of compositional data transformation: A new framework combining proportion conversion and contrast transformation," *Comput. Struct. Biotechnol. J.*, vol. 23, pp. 4088–4107, Dec. 2024, doi: 10.1016/J.CSBJ.2024.11.003.
[33] I. K. khan *et al.*, "Addressing limitations of the K-means clustering algorithm: outliers, non-spherical data, and optimal cluster selection," *AIMS Mathematics 2024 9:25070*, vol. 9, no. 9, pp. 25070–25097, 2024, doi: 10.3934/MATH.20241222.

Mohammad Rohmad Nurkhoirofiq: *Corresponding Author