Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

June 2026. Vol. 5. No. 3; e-ISSN: 2808-4519

Application of K-Means Clustering for Urban Transportation Pattern Analysis Using Big Data Trip Dataset

Tegas Ramadhan[1]*, Hafizh Ariiq[2], Muhammad Dzaki Arjun[3] , Muhammad Ridho Ananda Aditya[4]

[1]*Ilmu Komputer, FMIPA, Universitas Negeri Medan, Indonesia*
kuy410zml@gmail.com[1*], hafizhriiq12@gmaill.com[2], mdzakiarjunzaki@gmail.com[3], ridhoaditya.iskandar@gmail.com[4]

Abstract

The rapid growth of urban transportation systems has led to the generation of massive amounts of data, commonly referred to as big data. This study aims to analyze transportation patterns using large-scale data obtained from the NYC Taxi Trip Records. The dataset exhibits key big data characteristics, including volume, velocity, and variety. This research applies the K-Means clustering algorithm to group taxi trip data based on features such as trip distance, fare amount, and trip duration. Several preprocessing techniques are performed, including data cleaning, feature engineering, sampling, and normalization. The optimal number of clusters is determined using the Elbow Method and Silhouette Score. The results show that the dataset can be effectively grouped into three clusters representing distinct transportation patterns. These findings demonstrate the capability of clustering techniques in extracting meaningful insights from large-scale datasets and highlight their potential application in urban transportation planning.

Keywords: *Big Data; Clustering; K-Means; Data Mining; Transportation Analysis*

1. Introduction

The rapid development of information technology has significantly impacted urban transportation systems, leading to the generation of large-scale data. Taxi services, in particular, continuously produce data containing information such as trip distance, fare amount, and travel duration. This type of data is commonly referred to as big data due to its volume, velocity, and variety. Although transportation data holds valuable insights, analyzing such large and complex datasets remains challenging. Traditional analysis methods are often insufficient, making data mining techniques essential for extracting meaningful patterns. One of the most effective approaches in data mining is clustering, which groups data based on similarity without requiring labeled data.

This study utilizes data from the NYC Taxi Trip Records to analyze urban transportation patterns. The K-Means clustering algorithm is applied to group taxi trips based on key features, including trip distance, fare amount, and trip duration. The objective of this research is to identify meaningful patterns in transportation behavior using clustering techniques. The results are expected to provide insights that can support data-driven decision-making in urban transportation systems.

2. Theoretical Review

2.1. Big Data Fundamentals

2.1.1. Definition of Big Data

Big data refers to datasets that are extremely large, complex, and continuously generated, making them difficult to process using traditional data processing techniques. The concept of big data has become increasingly important in various fields, including transportation, healthcare, and finance, where large volumes of data are generated in real time. In the context of transportation systems, data is produced from multiple sources such as GPS devices, sensors, and transaction records. These datasets contain valuable information that can be used to analyze patterns and improve decision-making processes.

2.1.2. Characteristics of Big Data (5V)

Big data is commonly described using five main characteristics, known as the 5V:

1. Volume refers to the massive amount of data generated.
2. Velocity describes the speed at which data is produced and processed.
3. Variety indicates the different types of data formats, including structured and unstructured data.
4. Veracity relates to the quality and reliability of data.
5. Value represents the usefulness of data in generating insights.

The NYC Taxi Trip Records satisfies these characteristics, as it contains millions of records generated continuously with diverse attributes.

2.2. Data Mining Concepts

2.2.1 Definition of Data Mining

Data mining is the process of discovering meaningful patterns, correlations, and knowledge from large datasets. It involves the application of statistical, mathematical, and machine learning techniques to analyze data and extract useful information. Data mining plays a crucial role in transforming raw data into actionable insights, particularly in the era of big data.

2.2.2 Data Mining Techniques

Several techniques are commonly used in data mining, including:

1. Classification**,** which assigns data to predefined categories
2. Regression**,** which predicts continuous values
3. Clustering**,** which groups data based on similarity
4. Association Rule Mining**,** which identifies relationships between variables

Among these techniques, clustering is particularly useful when dealing with unlabeled data.

2.3 Clustering in Data Mining

2.3.1. Definition of Clustering

Clustering is an unsupervised learning technique that groups data points based on similarity. Unlike supervised learning, clustering does not rely on labeled data, making it suitable for exploratory data analysis. The goal of clustering is to ensure that data points within the same cluster are highly similar, while those in different clusters are significantly different.

2.3.2. Types of Clustering

Clustering methods can be categorized into several types, including:

1. Partition-based clustering (e.g., K-Means)
2. Hierarchical clustering
3. Density-based clustering (e.g., DBSCAN)

This study focuses on partition-based clustering using the K-Means algorithm.

2.4. K-Means Clustering Algorithm

2.4.1. Basic Concept

K-Means is a partition-based clustering algorithm that divides a dataset into K clusters. Each cluster is represented by a centroid, which is the mean value of all data points in the cluster. The algorithm aims to minimize the distance between data points and their respective cluster centroids.

2.4.2. Algorithm Steps

The K-Means algorithm follows an iterative process:

1. Initialize K centroids randomly
2. Assign each data point to the nearest centroid
3. Recalculate the centroid of each cluster

4. Repeat the process until convergence is achieved

2.4.3. Advantages and Limitations

Advantages:

1. Simple and easy to implement
2. Efficient for large datasets
3. Scalable

Limitations:

1. Requires predefined number of clusters (K)
2. Sensitive to outliers
3. Results depend on initial centroid selection

2.5. Cluster Evalution Methods

2.5.1. Elbow Method

The Elbow Method is used to determine the optimal number of clusters by analyzing the Within-Cluster Sum of Squares (WCSS). The method identifies the point where the decrease in WCSS becomes less significant, forming an “elbow” shape.

2.5.2. Silhoutte Score

The Silhouette Score measures how well each data point fits within its assigned cluster. It considers both cohesion (similarity within clusters) and separation (difference between clusters). A higher Silhouette Score indicates better clustering performance.

3. Methodology

3.1. Research Design

This study adopts a quantitative approach by applying data mining techniques to analyze large-scale transportation data. The main objective is to identify patterns in taxi trip behavior using clustering methods. ,The research process is structured into several stages, including data acquisition, data preprocessing, feature selection, clustering, and evaluation. Each stage is designed systematically to ensure that the results are valid, reproducible, and scientifically reliable.

3.2. Data Acquisition

3.2.1. Dataset Source

The dataset used in this study is obtained from the NYC Taxi Trip Records, which contains large-scale records of taxi trips. This dataset is widely used in transportation research due to its completeness and representation of real-world conditions.

3.2.2. Data Loading Process

```
import pandas as pd

df = pd.read_csv("taxi_data.csv")
print(df.head(10))
```

Fig. 1 : Data Loading Process

The dataset is loaded using the pandas library, which provides efficient data structures for handling large datasets. The read_csv() function reads the dataset file and converts it into a DataFrame format, allowing structured data manipulation. The head(10) function displays the first ten rows of the dataset. This step is essential for verifying that the dataset has been correctly loaded and for understanding its structure, including column names and sample values. This initial inspection helps identify potential issues such as missing values, incorrect data types, or unexpected formats.

3.3. Data Preprocessing

Data preprocessing is a critical step in data mining, as raw data often contains inconsistencies, missing values, and irrelevant features. Proper preprocessing ensures that the dataset is clean and suitable for analysis.

3.3.1. Feature Engineering (Trip Duration Calculation)

```
df['trip_time_in_secs'] = (
    pd.to_datetime(df['tpep_dropoff_datetime']) -
    pd.to_datetime(df['tpep_pickup_datetime'])
).dt.total_seconds()
```

Fig. 2: Feature Engineering (Trip Duration Calculation)

This step calculates the duration of each taxi trip by subtracting the pickup time from the drop-off time. The result is converted into seconds to ensure numerical consistency. Trip duration is an important feature because it captures the temporal aspect of transportation behavior. Without this feature, the analysis would only consider distance and fare, which may not fully represent trip characteristics.

3.3.2. Feature Selection

```
df = df[['trip_distance','fare_amount','trip_time_in_secs']].dropna()
```

Fig. 3: Feature Selection

Feature selection is performed to reduce the dimensionality of the dataset by selecting only relevant attributes. In this study, three features are chosen:

1. Trip Distance
2. Fare Amount
3. Trip Duration

The dropna() function removes rows with missing values, ensuring that the dataset used for clustering is complete and reliable. This step is crucial because missing values can negatively affect clustering results.

3.3.3. Sampling Strategy

```
df = df.sample(10000)
```

Fig.4: Sampling Strategy

Due to the large size of the dataset, a sampling technique is applied to reduce computational complexity. The sample() function randomly selects a subset of the data. This approach allows efficient processing while maintaining the overall characteristics of the dataset. It ensures that the clustering results remain representative of the original data.

3.3.4. Data Normalization

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
scaled_data = scaler.fit_transform(df)
```

Fig. 5: Data Normalization

Normalization is performed using the StandardScaler method, which standardizes features by removing the mean and scaling to unit variance. This step is essential because the K-Means algorithm relies on distance calculations. Without normalization, features with larger numerical values (e.g., fare amount) may dominate the clustering process, leading to biased results.

3.4. Cluster Optimization

Determining the optimal number of clusters is a crucial step in clustering analysis. In this study, two methods are used: the Elbow Method and the Silhouette Score.

3.4.1. Elbow Method Implementation

```
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

wcss = []

for i in range(1, 10):
    kmeans = KMeans(n_clusters=i, random_state=42, n_init=10)
    kmeans.fit(scaled_data)
    wcss.append(kmeans.inertia_)

plt.plot(range(1,10), wcss, marker='o')
plt.title("Elbow Method")
plt.xlabel("Number of Clusters")
plt.ylabel("WCSS")
plt.show()
```

Fig. 6 : Elbow Method Implementation

This code calculates the Within-Cluster Sum of Squares (WCSS) for different values of K. WCSS measures the compactness of clusters, with lower values indicating tighter clusters. The resulting graph is used to identify the optimal number of clusters by observing the point where the rate of decrease in WCSS slows down significantly. This point is commonly referred to as the “elbow.”

3.4.2. Silhouette Score Evaluation

```
from sklearn.metrics import silhouette_score

for i in range(2, 10):
    kmeans = KMeans(n_clusters=i, random_state=42, n_init=10)
    labels = kmeans.fit_predict(scaled_data)
    score = silhouette_score(scaled_data, labels)
    print(f"K={i}, Score={score}")
```

Fig. 7: Silhouette Score Evaluation

The Silhouette Score evaluates clustering quality by measuring how similar each data point is to its own cluster compared to other clusters. A higher score indicates better-defined clusters. This method complements the Elbow Method by providing a numerical evaluation of clustering performance.

3.5. Clustering Implementation

3.5.1. Final Model Construction

```
kmeans = KMeans(n_clusters=3, random_state=42, n_init=10)
df['cluster'] = kmeans.fit_predict(scaled_data)
```

Fig. 8 : Final Model Construction

After determining the optimal number of clusters, the K-Means algorithm is applied to group the data into three clusters. Each data point is assigned a cluster label, which is stored in a new column called cluster. This label represents the group to which each trip belongs.

3.5.2. Output Preparation for Analysis

At this stage, the dataset is fully prepared for analysis. Each data point has been assigned to a cluster, allowing further examination of cluster characteristics in the results section.

4. Results and Discussion

4.1. Dataset Overview

The dataset used in this study has been successfully preprocessed by selecting relevant features, namely trip distance, fare amount, and trip duration. After the preprocessing stage, the dataset is reduced through sampling to improve computational efficiency while maintaining its representative characteristics.

The resulting dataset is clean, consistent, and free from missing values, making it suitable for clustering analysis. The addition of the cluster label also indicates that the clustering process has been successfully applied.

4.2. Determination of Optimal Number of Clusters

4.2.1. Elbow Method Analysis

The Elbow Method is used to determine the optimal number of clusters by analyzing the relationship between the number of clusters and the Within-Cluster Sum of Squares (WCSS).

The result of the Elbow Method is presented in Figure 9.

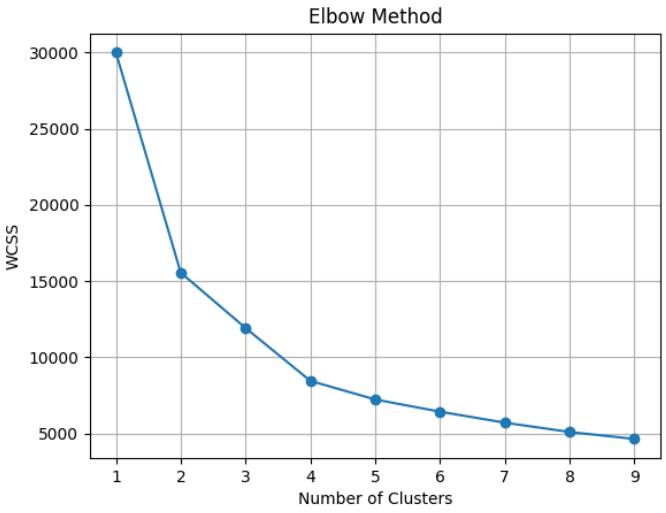

Fig. 9 : Elbow Method Analysis

Based on Figure 9, the graph shows a significant decrease in WCSS values from K = 1 to K = 3. After this point, the decrease becomes more gradual, forming an “elbow” shape. This indicates that increasing the number of clusters beyond K = 3 does not significantly improve the compactness of the clusters. Therefore, K = 3 is considered the optimal number of clusters for this dataset.

4.2.2. Silhoutte Score Analysis

To further validate the selection of the optimal number of clusters, the Silhouette Score is calculated for different values of K.

The results show that the Silhouette Score reaches a relatively high value at K = 3 compared to other values. This indicates that the clustering structure at K = 3 provides a good balance between cohesion and separation. The consistency between the Elbow Method and Silhouette Score strengthens the decision to use three clusters in this study

```
=== SILHOUETTE SCORE ===
K = 2, Score = 0.7502115653505829
K = 3, Score = 0.7477856349373807
K = 4, Score = 0.5887645194690613
K = 5, Score = 0.4660838792271443
K = 6, Score = 0.4745827383388814
K = 7, Score = 0.47863844627877616
K = 8, Score = 0.47283396422040513
K = 9, Score = 0.4012462367727421
```

Fig. 10: Silhoutte Score Analysis

4.3. Clustering Results

After determining the optimal number of clusters, the K-Means algorithm is applied to group the dataset into three clusters. The results show that each data point has been successfully assigned to a cluster. The distribution of data across clusters indicates that the clustering process is effective, as no single cluster overwhelmingly dominates the dataset.

This suggests that the dataset contains distinct patterns that can be meaningfully grouped.

```
=== HASIL CLUSTERING ===
         trip_distance  fare_amount  trip_time_in_secs  cluster
1661588           0.97          8.6              437.0        0
2269859           0.60          5.8              306.0        0
1876274           0.80          9.3              571.0        0
976447            1.93         10.7              517.0        0
2962359           4.44         26.3              531.0        0
2618157           1.10          7.9              425.0        0
2800941           1.33         10.7              641.0        0
2075873           1.25         11.4              669.0        0
2052243           1.70         14.9              954.0        0
1728339           1.20         11.4              656.0        0

=== DISTRIBUSI CLUSTER ===
cluster
0    9084
1     915
2       1
Name: count, dtype: int64
```

Fig. 11: Dataset contains distinct

4.4. Visualization of Clustering Results

To better understand the clustering results, a visualization is created using a scatter plot that shows the relationship between trip distance and fare amount.

The clustering visualization is presented in Figure 12.

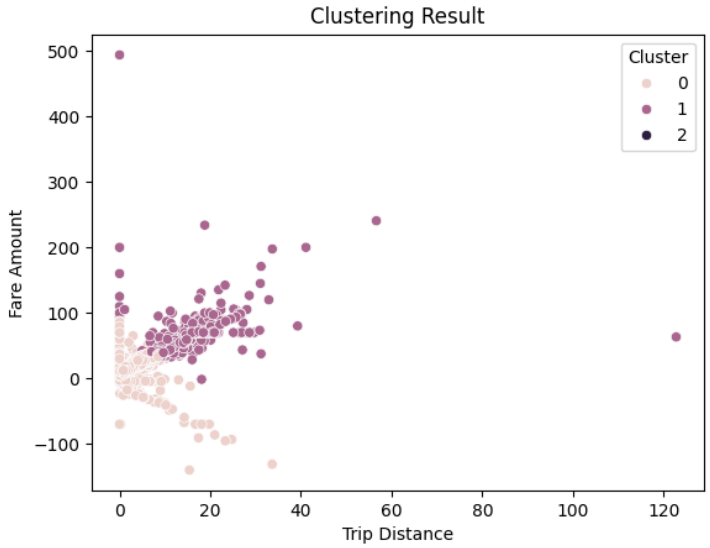

Fig. 12: Clustering visualization

As shown in Figure 12, the data points are grouped into three distinct clusters, each represented by a different color. The clusters appear to be well separated, indicating that the K-Means algorithm has successfully identified meaningful patterns in the dataset.

The visualization also shows that data points within the same cluster are relatively close to each other, suggesting strong similarity within clusters.

4.5 Cluster Characteristics Analytics

To further analyze the clustering results, the average values of each feature are calculated for each cluster. The results are presented in Figure 13.

```
=== RINGKASAN CLUSTER ===
         trip_distance  fare_amount  trip_time_in_secs
cluster
0             2.012378    12.709562         730.827389
1            14.162579    59.279847        2331.744262
2             3.670000    20.500000       57688.000000
```

Fig 13 : Each feature are calculated for each cluster

Based on Figure 13, each cluster exhibits distinct characteristics:

1. One cluster is characterized by low trip distance, low fare amount, and short duration, representing short-distance trips.
2. Another cluster shows moderate values, representing medium-distance trips.
3. The third cluster has high values for all features, representing long-distance trips with higher fares and longer durations.

These differences indicate that the clustering process successfully captures variations in transportation behavior.

4.6. Discussion

The results of this study demonstrate that the K-Means clustering algorithm is effective in analyzing large-scale transportation data. The Elbow Method and Silhouette Score consistently indicate that K = 3 is the optimal number of clusters. The clustering results reveal clear distinctions between different types of trips, supported by both visualization and statistical analysis.The findings show that most trips fall into distinct categories such as short, medium, and long-distance trips. This pattern is consistent with real-world urban transportation behavior. Overall, this study highlights the potential of clustering techniques in extracting meaningful insights from big data and supporting data-driven decision-making in urban transportation systems.

5. Conclusion

This study aims to analyze urban transportation patterns using clustering techniques on large-scale taxi trip data. The results show that the application of the K-Means algorithm is effective in grouping the dataset into meaningful clusters based on trip distance, fare amount, and trip duration.

Based on the Elbow Method and Silhouette Score analysis, the optimal number of clusters is determined to be three. The clustering results reveal distinct patterns of transportation behavior, consisting of short-distance trips, medium-distance trips, and long-distance trips. These patterns are clearly supported by both visualization and statistical analysis.

The findings indicate that clustering techniques can successfully extract valuable insights from large datasets, particularly in the context of transportation analysis. This demonstrates the potential of data mining methods in supporting data-driven decision-making processes.

For future research, it is recommended to incorporate additional features such as location data and time-based variables to obtain more comprehensive insights. Furthermore, other clustering algorithms may be explored to compare performance and improve the accuracy of the results.

References

[1] M. R. Alfahri, M. Z. Alkautsar, N. Khoiriah, A. P. H. Simbolon, and F. Ramadhani, "Analisis Pola dan Optimalisasi Rute Perjalanan Taksi Menggunakan K-Means Clustering," *Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)*, vol. 8, no. 2, 2025.

[2] X. Zhang and X. Zhao, "A Clustering-aided Ensemble Method for Predicting Ridesourcing Demand in Chicago," *arXiv preprint arXiv:2109.03433*, 2021.

[3] J. Lang, Z. Yang, Y. Zhou, C. Wen, and X. Cheng, "Four-dimensional aircraft emission inventory dataset of the landing-and-takeoff cycle in China (2019-2023)," *Earth System Science Data*, vol. 17, pp. 2489–2506, 2025, doi: 10.5194/essd-17-2489-2025.

[4] M. B. Hasan and M. Sarker, "Unraveling Urban Traffic Congestion Patterns in Bangladesh," in *Proceedings of the 11th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2025)*, 2025, pp. 319–325, doi: 10.5220/0013193600003941.

[5] E. A. Prasetio, D. Novizayanti, and A. N. A. Putri, "Cluster analysis of potential autonomous vehicle (AV) adopters in Indonesia's new capital," *Transportation Research Interdisciplinary Perspectives*, vol. 29, 2025, doi: 10.1016/j.trip.2024.101318.

[6] A. L. D. Loureiro, V. L. Miguéis, Á. Costa, and M. Ferreira, "Improving customer retention in taxi industry using travel data analytics: A churn prediction study," *Journal of Retailing and Consumer Services*, vol. 85, 2025, doi: 10.1016/j.jretconser.2024.104288.

[7] X. Li, J. Mango, J. Song, and D. Zhang, "XStar: a software system for handling taxi trajectory big data," *Computational Urban Science*, vol. 1, no. 17, 2021, doi: 10.1007/s43762-021-00015-w.

[8] D. Tzika-Kostopoulou, E. Nathanail, and K. Kokkinos, "Big data in transportation: a systematic literature analysis and topic classification," *Knowledge and Information Systems*, vol. 66, pp. 5021–5046, 2024, doi: 10.1007/s10115-024-02112-8.

[9] E. Sebti and Y. Chen, "Mining Hidden Ridesharing Patterns: A Data-Driven Gap Analysis of Chicago TNC Trips," *Data Science for Transportation*, vol. 8, no. 7, 2026, doi: 10.1007/s42421-026-00149-5.

[10] W. Jiang, "Data-driven Analysis of Taxi and Ride-hailing Services: Case Study in Chengdu, China," *Computer and Decision Making - An International Journal*, vol. 2, pp. 357–373, 2025.

[11] S. Alam, M. S. Ayub, H. Cui, and M. A. Khan, "A comparative study of machine learning models for taxi-demand prediction using a big data framework," *Public Transport*, vol. 17, pp. 803–833, 2025, doi: 10.1007/s12469-025-00401-1.