Student Name: Carlos Nakagomi

Instructor: Jonathan Agyeman

Langara College

Clustering NYC Taxi Trips to Identify Travel Patterns and High-Demand Zones

Abstract

This project applies clustering techniques to analyze New York City taxi trip data from the Taxi & Limousine Commission (TLC) to uncover meaningful travel patterns and demand zones. Using a cleaned and preprocessed dataset from early 2016, the study employs Principal Component Analysis (PCA) to reduce dimensionality and assesses clustering tendency through the Hopkins statistic and distance matrix. K-Means clustering is identified as the most effective algorithm based on a comprehensive comparison of internal (e.g., Davies-Bouldin Index, Calinski-Harabasz Index), external (e.g., Adjusted Rand Index, Variation of Information), and stability validation metrics. The optimal number of clusters (k = 4) is determined using the Elbow and Silhouette methods. Key findings reveal distinct demand hotspots around Manhattan and JFK Airport, temporal trip trends with peak activity between 12:00 PM and 5:00 PM, and fare-based customer segmentation that highlights the dominance of very low and very high fare rides. These insights can inform operational decisions for taxi services and support urban mobility planning.

1. Introduction

Understanding urban mobility patterns is essential for improving transportation efficiency, reducing congestion, and enhancing the quality of public services in large cities.

This project leverages the New York City Taxi and Limousine Commission (TLC) dataset to uncover hidden structures within taxi trip data using unsupervised machine learning. By applying clustering techniques to numerical, spatial, and temporal attributes, the goal is to identify meaningful patterns, such as demand hotspots, time-based behaviors, and fare-related customer segmentation.

Through a combination of preprocessing, dimensionality reduction, and cluster validation methods, the project aims to derive actionable insights that can support urban planning, optimize taxi services, and guide future transportation policies.

1. Objective

The goal of this project is to apply clustering techniques to the New York City Taxi & Limousine Commission (TLC) dataset to identify patterns in taxi trips based on geospatial, temporal, and fare-related attributes. This analysis will provide insights into demand hotspots, fare-based customer segmentation, and time-based trip behaviors.

1. Exploratory Data Analysis

After downloading the zipped file containing all the datasets from the Kaggle website, four datasets were found, but only three of them (from the year 2016) were used. The first step after loading the datasets was to check if the variable names were consistent across them, to ensure they could be merged properly.

* 1. Description of Variables:

The dataset consists of the following key attributes:

Categorical Variables:

VendorID: Identifies the taxi service provider.

RateCodeID: Indicates the rate category for the trip.

Store\_and\_fwd\_flag: Indicates whether the trip was stored before transmission.

Payment\_type: Specifies how the passenger paid for the trip (e.g., Credit Card, Cash)

Numerical Variables:

Passenger\_count: Number of passengers in the taxi.

Trip\_distance: Distance of the trip in miles.

Fare\_amount: Cost of the trip before additional fees.

Extra, MTA\_tax, Improvement\_surcharge: Additional charges applied.

Tip\_amount: Tip provided by passengers (for credit card payments).

Tolls\_amount: Amount paid for tolls.

Total\_amount: Total fare paid, including fees and tolls.

Datetime Variables:

tpep\_pickup\_datetime: Start time of the trip.

tpep\_dropo \_datetime: End time of the trip.

Coordinates:

Pickup\_longitude and Pickup\_latitude: GPS coordinates of trip start.

Dropoff \_longitude and Dropoff \_latitude: GPS coordinates of trip end.

1. Stratified Sampling

A stratified sampling was performed, resulting in a subset of 3,449,985 observations. This approach helps maintain class balance across payment\_type categories, preventing the model from being biased in favor of more frequent or dominant trip types and ensuring fair representation of less common but potentially important patterns. Stratified sampling was performed using the payment\_type as the stratifying variable to proportionally represent each class.

1. Data Cleaning

No missing values or duplicated rows were detected in the stratified dataset, but several outliers were found, as shown in the output below. The outliers were handled by individual analysis basis on real scenarios valid ranges.

|  |  |
| --- | --- |
| **Variable** | **Outliers** |
| passenger\_count | 365676 |
| trip\_distance | 364111 |
| pickup\_longitude | 241218 |
| pickup\_latitude | 167490 |
| RatecodeID | 87250 |
| dropoff\_longitude | 216949 |
| dropoff\_latitude | 202849 |
| payment\_type | 4471 |
| fare\_amount | 309408 |
| extra | 8725 |
| mta\_tax | 14936 |
| tip\_amount | 179660 |
| tolls\_amount | 172689 |
| improvement\_surcharge | 1945 |
| total\_amount | 308301 |

Figure 1

1. Handling Outliers

In order to clean the outliers, only values within a valid range were used in the analysis. The goal was to preserve the most realistic scenarios that this dataset can represent. Handling the outliers with the mean, median or mode wouldn't be appropriate because these statistical imputation methods could distort the actual distribution of the data, especially in a context like taxi trip records where extreme values may correspond to valid but rare events, such as long-distance airport rides or high-tip fares.

* 1. Valid Range

A screenshot of a computer

AI-generated content may be incorrect.

Figure 2

VendorID: There is no valid range as it is index.

Tpep\_pickup\_datetime: There is no valid range as it is date and time.

Tpep\_dropoff\_datetime: There is no valid range as it is date and time.

passenger\_count: This range does not fit the possible number of passengers inside of a taxi, the maximum value for this variable will be adjusted to 3 by dropping all the observations with passenger\_count higher than 3.

trip\_distance: This range does fit a reasonable distance for a taxi trip. Take as reference, the size of the land boundary of New York City (perimeter of the city), stretches about 305 miles. A threshold of 35 miles at most will be determined as the maximum range.

pickup\_longitute: This range is not within the real geographic boundaries of New York City. Only observations within the valid New York City longitude range of -74.26 to -73.70 will be kept.

pickup\_latitue: This range is outside the valid geographic boundaries of New York City. Only observations within the valid NYC latitude range of 40.49 to 40.92 will be kept.

RatecodeID: There is no valid range as it is index.

store\_and\_fwd\_flag: Refers to whether the trip record was stored temporarily and forwarded later to the central system, meaning that there is no valid range.

dropoff\_longitute: This variable will be filtered to retain only observations within the actual geographic boundaries of New York City. Valid values must fall within the range of -74.26 to -73.70.

dropoff\_latitude: This variable will be filtered to retain only observations within the actual geographic boundaries of New York City. Valid values must fall within the range of 40.49 and 40.92.

Payment\_type: This variable is based on criteria set by the New York City Taxi & Limousine Commission, so it is not possible to determine the valid range.

fare\_amount: Outliers were detected as a fare\_amount of 2020.37 is unrealistic for a taxi trip. Less or equally 250 will be used as a threshold for this variable. Additionally, there are negative values, which are not possible when it comes to payments.

extra: Outliers were detected as there are negative values in the variable range, which is not possible when it comes to payments.

mta\_tax: This tax helps fund mass transit projects (subway, buses, etc.) in NYC. Outliers were detected as there is negative values, which are not possible when it comes to payments. The threshold for this variable was defined as less than 0.5 and higher than 0, which means that this variable is outside of the valid range.

tips\_amount: This value is outside the valid range, as the maximum observed is unusually high for a typical taxi tip. Outliers were detected as there is negative values, which are not possible when it comes to payments. The threshold for this variable was defined as less than 50 and higher than 0, which means that this variable is outside of the valid range.

tolls\_amount: The tolls\_amount variable represents the amount paid for tolls during a taxi ride, typically for crossing bridges, tunnels, or toll roads. Outliers were detected as there is negative values, which are not possible when it comes to payments. The threshold for this variable was defined as less than 0.5 and higher than 0, which means that this variable is outside of the valid range.

improvement\_surchange: The improvement\_surcharge is a fixed surcharge added to each taxi fare to fund improvements to New York City's taxi industry. The threshold for this variable was defined as less or equal to 0.3, meaning that this variable is inside of the valid range.

total\_amount: This variable is out of the valid range as its maximum value does not reflect the reality of a taxi trip. Outliers were detected as there is negative values, which are not possible when it comes to payments. The threshold for this variable was defined as less than 300 and higher than 0, which means that this variable is also outside of the valid range.

* 1. Dropping Columns

Irrelevant columns that do not contribute to the exploratory data analysis were dropped, as they will only be used for interpretative purposes later. These columns are VendorID, tpep\_pickup\_datetime, tpep\_dropoff\_datetime, pickup\_longitude, pickup\_latitude, RatecodeID, store\_and\_fwd\_flag, dropoff\_longitude, and dropoff\_latitude.

1. Assessing Clustering Tendency

To understand better if distinct patterns or groupings existed within the trip data, the data was scaled, and a Principal Component Analysis (PCA) was performed to reduce dimensionality and reveal underlying structure.

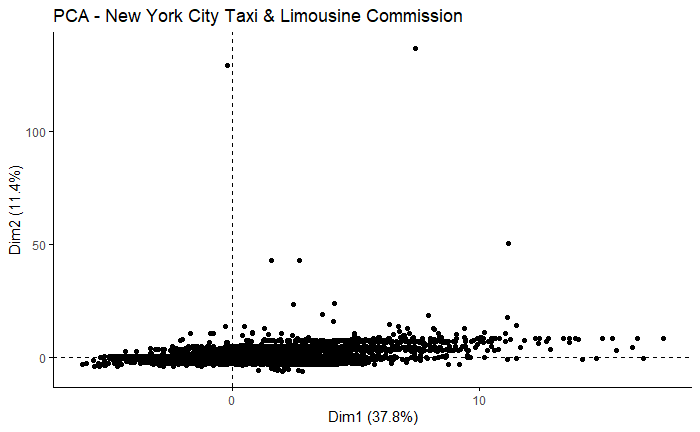


Figure 3

The PCA scatter plot (Figure 3) illustrates the NYC taxi dataset projected onto the first two principal components, Dim1 and Dim2, which account for 37.8% and 11.4% of the total variance, respectively.

A large concentration of points near the origin indicates that most taxi trips exhibit similar characteristics across the key numerical features. In contrast, several data points are dispersed along Dim1, highlighting trips that deviate notably from the norm. These outliers likely correspond to rides with atypical fare amounts, trip distances, or other uncommon attributes. While the variation along Dim2 is less pronounced, it still reveals meaningful differences among a smaller subset of trips. This demonstrates that PCA effectively reduces dimensionality while preserving critical data structure, making it a valuable tool for detecting both common patterns and anomalies within the dataset.

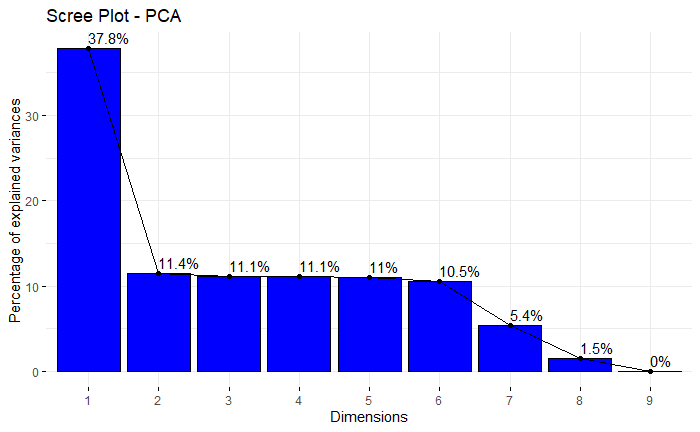


Figure 4

The scree plot (Figure 4) shows the percentage of variance explained by each principal component, offering insight into how much information each dimension retains from the original dataset.

Principal Component 1 (PC1) accounts for the largest portion of variance at approximately 37.8%, followed by PC2 at 11.4%. From PC3 onward, the explained variance gradually decreases, with each subsequent component contributing around or below 11%. This pattern indicates that the most significant variance is captured by the first few components, meaning that dimensionality reduction can be effectively performed by retaining only these components, as they preserve most of the dataset’s variability while filtering out redundant or less informative features.

* 1. Hopkins Statistics

The Hopkins statistic measures the spatial distribution of data points in a vector space using Euclidean distance. To compute this statistic, only numerical variables from the pre-scaled dataset were used. A simple random sample (0.1% of the data) was taken to reduce computational load and processing time. A Hopkins value close to 1 indicates a strong tendency for clustering. In this case, the Hopkins statistic is 0.8497, suggesting that the dataset exhibits a very high tendency to form distinct and well-defined clusters.

* 1. Distance Matrix

A distance matrix is a graphical representation (heatmap) which shows the distances between observations in a dataset after applying any metric technique. In this case it also was performed a Simple Random Sampling with 1000 observations of the scaled dataset to plot the distance matrix.

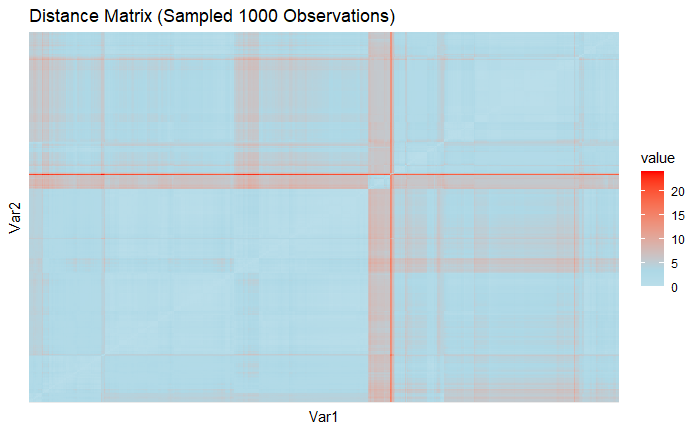


Figure 5

Figure 5 displays the pairwise Euclidean distances between 1,000 randomly sampled NYC taxi trips based on their numerical features. The matrix is predominantly shaded in light blue, indicating that most trips are relatively similar in terms of variables such as fare amount, trip distance, tips, and tolls. However, the appearance of distinct red vertical and horizontal lines reveals that certain trips stand out as significantly different from the rest. These deviations may correspond to unique trip profiles resulting from extreme values or unusual feature combinations.

1. Determining the Optimal Number of Clusters
   1. Elbow Method

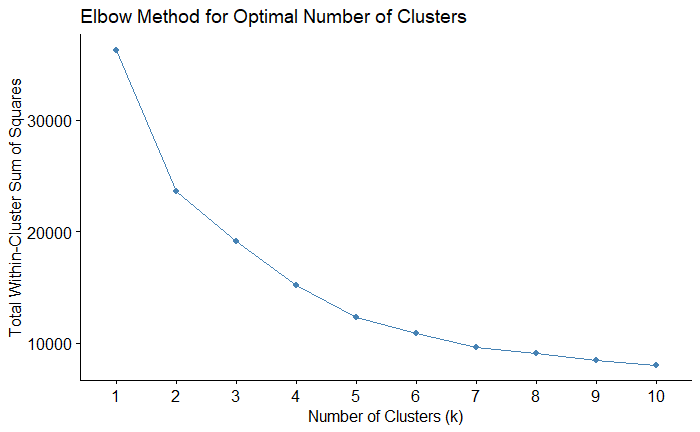


Figure 6

The Elbow Method plot illustrates the total within-cluster sum of squares (WSS) across a range of cluster values from k = 1 to k = 10. A steep decline in WSS is observed from k = 1 to k = 4, indicating that increasing the number of clusters in this range greatly enhances the compactness of the clustering. The “elbow” of the curve is visible at k = 4, which represents the point where the benefit of adding more clusters begins to level off. This supports the selection of 4 clusters as an appropriate choice, balancing clustering performance with model simplicity.

* 1. Average Silhouette Method

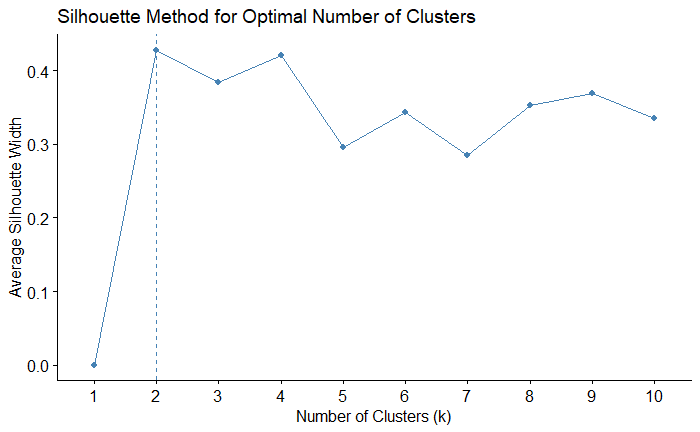


Figure 7

The Silhouette Method plot displays the average silhouette width across different cluster counts (k), ranging from 1 to 10. This score reflects how well each observation fits within its assigned cluster—higher values indicate more distinct and well-separated groupings. While the peak silhouette value is observed at k = 2, relatively high scores are also evident for k = 3 and k = 4. Given the overall structure and interpretability of the clusters.

* 1. “K” Winner (Elbow Method x Average Silhouette Method)

k = 4 was selected as it offers a good balance between cohesion within clusters and separation between them, while still capturing more nuanced patterns in the data.

1. Cluster Validation Statistics

To assess how well a clustering algorithm has performed, six cluster validation measures will be used. These measures help to ensure that the identified patterns are meaningful and not simply the result of random noise.

* 1. Internal measures for cluster validation

Internal cluster validation evaluates the quality of a clustering structure using only the information inherent to the dataset and clustering results.

* + 1. Silhouette Coefficient for Hierarchical Clustering

The silhouette coefficient measures how similar an observation is to its own cluster compared to other clusters. It ranges from -1 to +1 (+1 = well-clustered, 0 = borderline/mixed assignment, -1 = likely assigned to the wrong cluster). The result value of 0.443 is moderately good, meaning that there is some structure in the clusters.

* + 1. A Davies-Bouldin Index

The Davies-Bouldin Index (DBI) results suggest that K-Means achieved the best clustering performance among the three methods, with the lowest DBI score of 0.9075. PAM (K-Medoids) followed with a higher score of 1.1754, indicating slightly less compact and well-separated clusters. CLARA had the highest DBI value at 1.3266, suggesting its clusters were the least distinct and more dispersed. Since lower DBI values indicate better-defined clusters with less intra-cluster variance and greater inter-cluster separation, K-Means appears to be the most effective clustering algorithm for this dataset based on this evaluation.

* + 1. Calinski-Harabasz Index

The Calinski-Harabasz Index results show that K-Means achieved the highest score of 1305.96, indicating the best overall clustering performance in terms of between-cluster dispersion and within-cluster compactness. PAM (K-Medoids) had a slightly lower score of 1272.01, while CLARA scored the lowest at 1232.16. Since a higher Calinski-Harabasz Index reflects better-defined clusters with greater separation, K-Means once again outperformed the other methods, followed closely by PAM, with CLARA showing the weakest clustering structure among the three.

* + 1. Dunn Index

The Dunn Index results indicate that K-Means achieved the highest score at 0.0012, followed by CLARA at 0.0009 and PAM (K-Medoids) at 0.0005. Therefore, based on the Dunn Index, K-Means performed the best, with CLARA showing moderate performance and PAM trailing behind.

* 1. External measures for clustering validation

External cluster validation involves comparing the results of a clustering analysis to known external classifications, such as predefined class labels. It evaluates how well the generated cluster assignments align with these actual labels.

* + 1. Rand Index

The Adjusted Rand Index (ARI) results reveal a strong agreement between K-Means and CLARA, with a high ARI of 0.9286, indicating that these two methods produced very similar cluster assignments. In contrast, the agreement between K-Means and PAM was lower at 0.623, and the similarity between PAM and CLARA was even lower at 0.599. These results suggest that K-Means and CLARA were most consistent with each other in their clustering structure, while PAM deviated more from both, particularly from CLARA.

* + 1. Meila’s variation index

Meila’s Variation of Information (VI) results further support the clustering agreement observed with the Adjusted Rand Index. The lowest VI value was between K-Means and CLARA at 0.3404, indicating a high degree of similarity between their cluster assignments. In contrast, K-Means and PAM had a higher VI of 1.0613, while PAM and CLARA showed the greatest divergence with a VI of 1.1039. Since lower VI values represent greater similarity between clustering outcomes, these results confirm that K-Means and CLARA produced the most consistent clustering structures, while PAM differed substantially from both.

* 1. Stability Validation Metrics

The Average Proportion of Non-overlap (APN) score for K-Means with 2 clusters is 0.0098, indicating very high clustering stability. Since lower APN values signify better stability, this result strongly supports the robustness of the K-Means clustering. The Average Distance (AD) for PAM with 6 clusters is 1.6808, which is relatively high-suggesting that the PAM clustering configuration is less stable. The Average Distance between Means (ADM) for K-Means with 2 clusters is 0.0600, a low value that reflects good separation and consistency between clusters. Lastly, the Figure of Merit (FOM) for PAM with 6 clusters is 0.5880. Although not as low as APN or ADM, it still indicates moderate internal consistency.

A screenshot of a graph

AI-generated content may be incorrect.

Figure 8

1. Choosing the Best Clustering Algorithms

Based on a thorough evaluation using internal, external, and stability validation metrics, K-Means emerges as the most effective and consistent clustering algorithm for this NYC Taxi dataset.

From an internal validation perspective, K-Means demonstrated superior clustering quality. It achieved the lowest Davies-Bouldin Index (0.9075) and the highest Calinski-Harabasz Index (1305.96), indicating that its clusters were both compact and well-separated. Although CLARA had a slightly better Dunn Index (0.0009) than PAM (0.0005), K-Means still outperformed both with a Dunn score of 0.0012, reinforcing its ability to form meaningful clusters across all three internal metrics.

In terms of external validation, K-Means exhibited strong consistency with CLARA. The Adjusted Rand Index (ARI) between K-Means and CLARA was 0.9286, the highest among all comparisons, and their Variation of Information (VI) score was also the lowest at 0.3404, reflecting a high level of agreement in their cluster assignments. PAM, by contrast, showed significantly lower alignment with both K-Means (ARI = 0.623; VI = 1.0613) and CLARA (ARI = 0.599; VI = 1.1039), suggesting less consistent clustering behavior.

Regarding stability validation, K-Means once again showed robust performance. It recorded the lowest APN (0.0098) and ADM (0.0600) scores, highlighting its stability and consistent cluster formation across resampling. PAM, on the other hand, exhibited the highest AD (1.6808) and FOM (0.5880), indicating weaker performance in terms of cluster reliability and internal consistency.

Taken together, these findings clearly support K-Means as the optimal clustering method for this analysis. It consistently outperformed PAM and CLARA across internal cohesion, external agreement, and cluster stability, making it the most suitable approach for uncovering meaningful patterns in the NYC Taxi dataset.

1. Key Findings & Insights

In this section, the K-Means algorithm is applied to the cleaned NYC taxi dataset using **k = 4** clusters. It also incorporates geospatial, temporal, and fare-based analyses to uncover meaningful patterns in passenger behavior, trip characteristics, and demand hotspots. The resulting clusters are interpreted to provide actionable insights into how different types of taxi trips are distributed across time, location, and fare amount.

* 1. Demand Hotspots

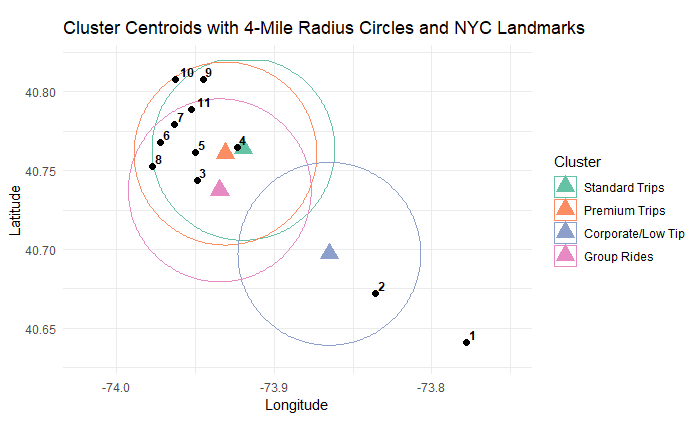


Figure 9

|  |  |
| --- | --- |
| 1 | JFK Airport |
| 2 | Resorts World Casino |
| 3 | Long Island City |
| 4 | Astoria |
| 5 | Roosevelt Island |
| 6 | Central Park |
| 7 | The Met (Museum) |
| 8 | Grand Central Terminal |
| 9 | Harlem (125th St) |
| 10 | Columbia University |
| 11 | Mount Sinai Hospital |

The scatter plot of pickup locations by cluster reveals distinct spatial patterns across trip types in New York City. The "Corporate/Low Tip" cluster (in blue) is notably concentrated around JFK Airport and other peripheral areas, suggesting these trips may involve longer distances or airport-related services. In contrast, "Standard Trips" (green) and "Premium Trips" (orange) are densely clustered in the central Manhattan region, indicating high activity and demand in the commercial core of the city. The "Group Rides" cluster (pink) appears slightly more dispersed, suggesting variability in pickup zones, possibly due to ride-sharing or multi-passenger arrangements.

* 1. Time-Based Trip Behaviors

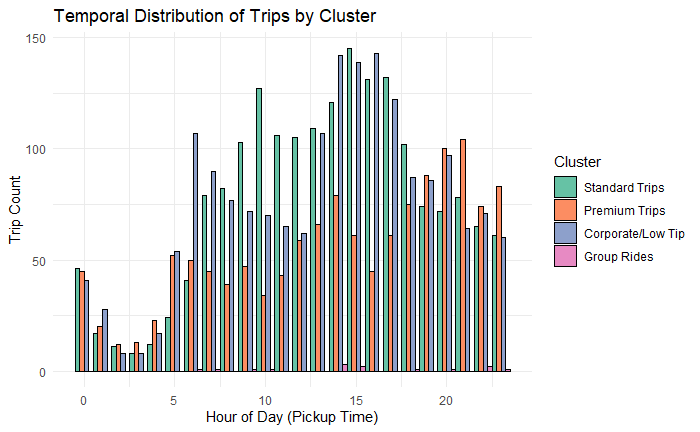


Figure 10

The bar chart shows that taxi demand peaks between 12:00 PM and 5:00 PM, with all clusters experiencing higher activity during these hours. Standard Trips and Corporate/Low Tip trips dominate the afternoon peak, suggesting a strong presence of work-related or routine personal travel. Premium Trips exhibit more even distribution throughout the day, with noticeable activity in the late evening, potentially reflecting nightlife or higher-end service usage. Group Rides show a relatively consistent pattern but slightly lower volume overall, with minor peaks aligning with afternoon and early evening hours.

* 1. Fare-Based Customer Segmentation

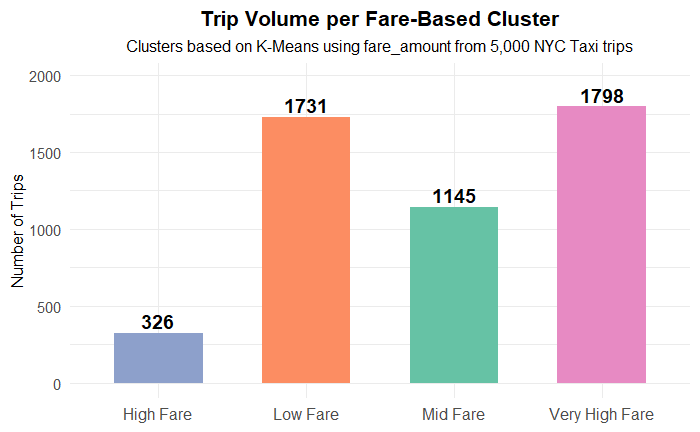


Figure 11

The bar chart illustrates the distribution of NYC taxi trips based on fare amount using K-Means clustering. The majority of trips fall into the Very High Fare (1,798 trips) and Low Fare (1,731 trips) clusters, suggesting that passengers are either taking short, low-cost rides or longer, more expensive trips-possibly to or from airports or across boroughs. The Mid Fare cluster accounts for 1,145 trips, representing moderate-distance rides typical of intra-city travel. The High Fare cluster is the smallest (326 trips), indicating that this fare range may capture outliers or less common trip types.

1. Conclusion

This project applied unsupervised clustering techniques to the NYC Taxi & Limousine Commission dataset to uncover patterns in taxi usage based on spatial, temporal, and fare-related variables.

After rigorous data cleaning and dimensionality reduction, K-Means clustering was identified as the most effective algorithm, outperforming PAM and CLARA across internal, external, and stability validation metrics.

Using K = 4 clusters, the analysis revealed meaningful insights: distinct pickup zones such as JFK Airport and central Manhattan emerged as high-demand hotspots; time-based patterns showed peak taxi usage in the afternoon; and fare-based segmentation highlighted clear differences in customer behavior and trip types.

Overall, the findings highlight the practical value of clustering in revealing hidden structures within complex transportation data. By identifying distinct trip patterns and demand trends, this approach can support data-driven decision-making for improving service allocation, enhancing passenger experience, and guiding future urban transportation planning initiatives.