

# Forecasting Public Bicycle Utilization in New York City Utilizing the CitiBike Dataset

Xudong Wang<sup>1, a \*</sup>, Zile Xu<sup>2, b</sup>, and Yufen Zhang<sup>2, c</sup>

<sup>1</sup> School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, 200093, China;

<sup>2</sup> College of Science, University of Shanghai for Science and Technology, Shanghai.

<sup>a</sup> wxd\_7546@163.com, <sup>b</sup> math\_Xuzl@163.com, <sup>c</sup> zzyf5085@163.com

**Abstract.** The integration of public bicycles into urban transportation systems has gained popularity due to their potential to reduce traffic congestion and air pollution, while also addressing the “last kilometer” problem. This study examines the usage patterns and predictive analysis of public bicycles in New York City using the CitiBike dataset from 2013 to 2015. We employ the Spring algorithm for networkx visualization and analyze the loan and return data to understand spatial-temporal usage dynamics. Additionally, clustering techniques, including K-means and DBSCAN, are applied to the dataset to uncover usage trends and patterns. The findings provide valuable insights for the planning and optimization of public bicycle systems, such as CitiBike, and contribute to the promotion of sustainable and environmentally friendly urban transportation.

**Keywords:** Spring Algorithm; NetworkX Library; K-Means Clustering; DBSCAN.

## 1. Introduction

The escalation of urban populations and evolving societal demands have intensified concerns about traffic congestion and air pollution in cities. Consequently, there has been a global trend towards sustainable transportation solutions, with public bicycle systems emerging as a critical component in reducing pollution and addressing the “last-mile” challenge in public transit[1]. Despite their advantages, these systems encounter operational hurdles that impact user satisfaction and system performance.

The primary issues include uneven station distribution leading to imbalanced resource allocation, inefficient vehicle dispatch due to manual strategies, and difficulties in borrowing and returning bicycles during peak hours, resulting in the “bike crunch.” These challenges highlight the need for advanced data-driven analysis to optimize the placement and operation of public bicycle systems[2].

This study aims to address these challenges by utilizing the CitiBike dataset from New York City, spanning the years 2013 to 2015. By employing networkx analysis, clustering algorithms, and time series forecasting models, we aim to uncover insights into the usage patterns, peak demand periods, and operational efficiency of public bicycle systems. The findings are expected to contribute to the improvement of urban transportation planning and the optimization of public bicycle systems, ultimately enhancing the accessibility and sustainability of urban mobility.

## 2. Related Works

The integration of public bicycles into urban transportation systems has become a critical component of sustainable urban mobility. However, the challenges posed by the uneven spatial and temporal distribution of bicycle demand have led to the development of data-driven approaches for demand prediction and system optimization.

Sathishkumar V E et al[3]. (2020) explored a model for forecasting hourly bicycle rental demand using weather information and historical data. They employed feature filtering and trained various statistical regression models, with the gradient boosting machine showing superior performance. Liang Y et al[4]. (2022) proposed a graph-based deep learning method, B-MRGNN, to predict

bicycle demand using multi-modal historical data. This method captures cross-modal spatio-temporal dependencies and demonstrates superior performance compared to existing methods. Jiang W[5] (2022) conducted a comprehensive review of bike-sharing usage prediction using deep learning, classifying prediction problems and models, and discussing various applications. Zhou J et al[6]. (2022) used bibliometric methods to analyze shared bicycle research literature from 2010 to 2020, identifying key research themes and trends. Ma C, Liu T[7] (2024) studied the spatio-temporal characteristics of shared bicycles and proposed a CNN-LSTM-Attention algorithm for demand prediction, achieving a high prediction accuracy of 97.50%.

This paper adopts a multifaceted approach to analyze and forecast public bicycle usage patterns in New York City. The methodology begins with a network analysis of bicycle loan and return, constructing a matrix and developing a network graph to identify network metrics and local connectivity. This is followed by data preprocessing and clustering analysis, where the dataset undergoes cleaning and conversion, and clustering algorithms are employed to discern usage patterns and user segments. The most effective clustering method is selected based on quality indicators, and the results are scrutinized to gain insights into user behaviors and preferences. Ultimately, the study aims to provide actionable insights for the optimization of public bicycle systems and to bolster sustainable urban transportation planning.

### 3. Methods

#### 3.1 Network Diagram Layout Using the Spring Algorithm

The operator's role in public bicycle systems is pivotal for understanding user behavior and optimizing service. By analyzing historical data, the study aims to comprehend usage patterns of diverse user types and tailor promotional initiatives to boost bicycle usage[8]. Additionally, understanding the relationships between stations can enhance vehicle dispatching efficiency. The Spring algorithm is employed to arrange a network diagram, converting string data types to date formats for accurate analysis. The diagram consists of "start station" and "end station" data, with metrics like node count (representing stations), edge count (indicating borrowing/returning events), and network density (proportion of potential connections represented by actual edges) calculated. The diagram is visualized, as illustrated in Figure 1, to aid in interpreting the network's structure and dynamics.

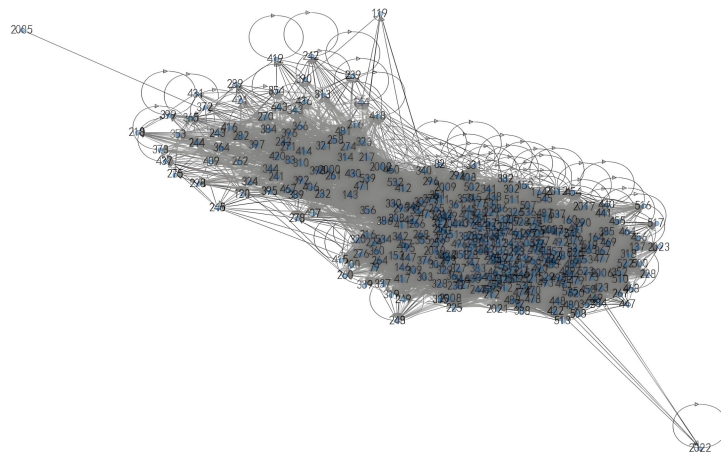


Fig. 1 2014-08-03 Public bicycle loan and return network diagram

Figure 1 visualizes the public bicycle network on August 3, 2014, illustrating a dense network of 322 stations and 4570 transactions, with a network density of 0.141, indicating a well-connected system. The study demonstrates how user behavior insights can inform targeted programs and scheduling strategies to enhance user satisfaction and system efficiency. By leveraging data to predict peak demand and popular routes, the system can strategically allocate resources, thereby improving availability and satisfaction. Moreover, by optimizing vehicle delivery and collection

based on user patterns and station proximity, the study explores methods to minimize congestion and idle time, further bolstering operational efficiency.

### 3.2 Data Preprocessing and Clustering Analysis for Bicycle Usage Patterns

#### 3.2.1 Networkx-based Subgraph Analysis for Average Path Lengths and Diameter

We begin by defining the geographical scope, encompassing the region with latitudes between 40.695 and 40.72, and longitudes ranging from -74.023 to -73.973. From the comprehensive bicycle rental and return network graph, we isolate the stations within this area to create a focused subgraph. Utilizing the capabilities of the networkx library, we then calculate the average shortest path length and network diameter for this subgraph, ensuring its strong connectivity. Our computations reveal an average shortest path length of 1.8849 and a network diameter of 4, providing insights into the interconnectivity and spatial distribution of bicycle rental and return stations within the localized network. Through meticulous analysis, we extract key data including the number of nodes, the count of edges, the network density, and the average shortest path length, as well as the network diameter. This information is invaluable for deciphering regional bicycle usage patterns and for informing urban transportation strategies and bicycle-sharing system enhancements.

#### 3.2.2 Data Preprocessing and Feature Selection

Prior to clustering bicycle data from July 1, 2013, to August 31, 2015, it is crucial to assess the size and quality of the dataset. The extensive volume of data may introduce various quality issues, such as missing values, outliers, and inconsistencies. Preprocessing is thus imperative to maintain the accuracy and reliability of the data. To handle missing values, the choice between direct removal of samples with missing values or applying an appropriate filling method is critical. While direct removal may reduce dataset size, it ensures data integrity given the substantial data volume.

In selecting features for cluster analysis, it is essential to identify those that are pertinent to bicycle data and influence clustering outcomes based on the specific context and objectives of the problem. Typically, features related to usage patterns are chosen, such as trip duration, start station latitude, and start station longitude. These features provide insights into the timing, geographic location, and duration of bike use, aiding in the discovery of underlying clustering patterns and groups.

#### 3.2.3 Construction of a K-means Clustering Model

This section explores the application of two clustering algorithms—K-means and hierarchical clustering—to a preprocessed dataset. We perform cluster analysis, scrutinize outcomes, and compare the results to select the most appropriate method for the problem. K-means requires a predefined number of clusters,  $K$ , and iteratively minimizes the objective function. The Elbow Method or Silhouette Score helps determine the optimal value for  $K$ , indicating the optimal number of clusters.

$$SSE = \sum_{i=1}^k \sum_{p \in C_i} |p - m_i|^2. \quad (1)$$

This analysis employs K-means clustering, an unsupervised learning algorithm, to partition the dataset into clusters, aiming to minimize the sum of squared errors between data points and their assigned cluster centers. An appropriate number of clusters, denoted as  $K$ , is selected using methods like the Elbow Method or Silhouette Score to find the balance between cluster number and homogeneity. Preprocessing includes handling missing values and outliers, standardizing features, and selecting relevant subsets like start and end station coordinates. K-means iteratively assigns data points to the nearest centroid until convergence, revealing data structure and similarities, aiding further analysis and decision-making.

#### 3.2.4 K-means Clustering Solution and Analysis

We employed programming packages, including ‘sklearn. cluster’, ‘matplotlib’, ‘seaborn’, and ‘pandas’, to conduct cluster analysis on the 26 datasets and generate result graphs through a for loop.

This analysis involved calculating the Sum of Squared Errors (SSE) and contour coefficient values. Figure 2 illustrates the K-means clustering results for the July 2013 dataset, demonstrating the determination of the optimal K value. The graph indicates that K=4 is the best value, as the change in distortion degree is significantly reduced at K=5 compared to K=4. When K=4, the relationship between SSE and K takes the form of an elbow, suggesting that the corresponding K value is the actual number of clusters in the data.

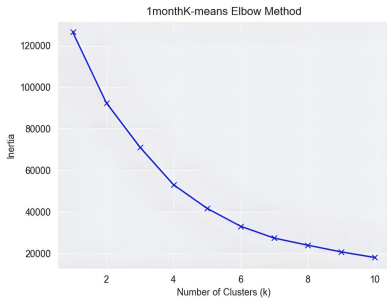


Fig. 2 July 2013 Dataset - Determining the Most Suitable Number of Clusters K Value Graph

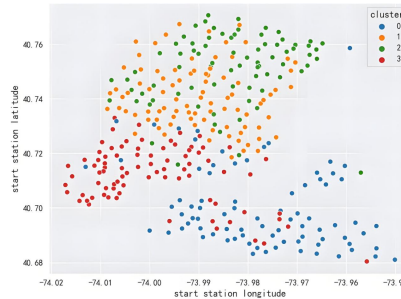


Fig. 3 July 2013 Dataset - K-means Clustering Results Graph

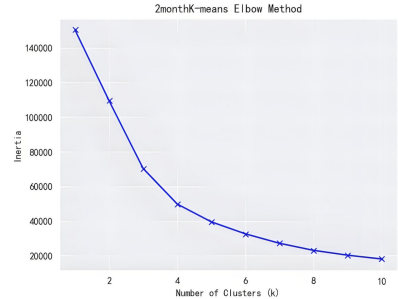


Fig. 4 August 2013 data set - Determine the most appropriate clustering number K-value plot

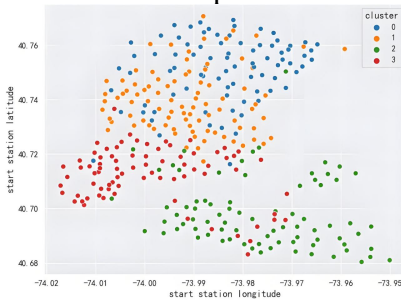


Fig. 5 Clustering results of data set -K means in August 2013

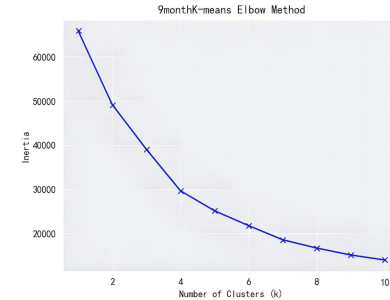


Fig. 6 March 2014 data set - Determine the most appropriate clustering number K-value plot

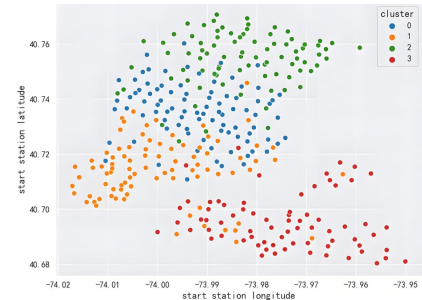


Fig. 7 Clustering results of data set K-means in March 2014

By setting K=4, we conducted clustering and obtained the cluster graph 3, which yielded the following results. Figure 4 reveals that the data points are effectively segmented into two distinct regions and four clusters, with an associated Sum of Squared Errors (SSE) of 21.37 and a contour coefficient of 0.2457. This pattern of clustering is consistent across other months' data. See Figure 5,6,7,8. Figure 8 shows a partial cluster map between July 2013 and August 2015.

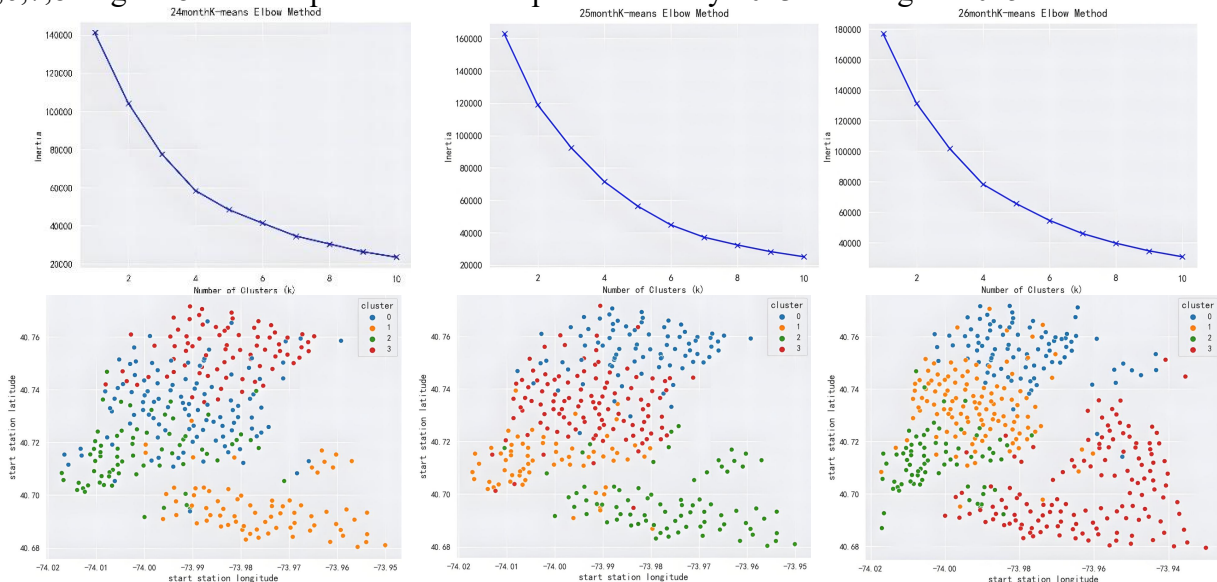


Fig. 8 The partial cluster map spanning from July 2013 to August 2015

K-means clustering of the bicycle dataset reveals two primary regions: the upper left and lower right quadrants. These areas exhibit higher bicycle borrowing intensity. Additionally, a distinct gap between the red cluster and others suggests potential spatial isolation due to geographical features like rivers or roads. These obstacles affect travel paths, leading to less frequent bike usage at certain stations. The larger blue, green, and orange clusters indicate a more extensive station network in these areas, where public bike usage is more prevalent and stations are busier. In contrast, the red area's development may be less robust, with users more concentrated in the upper left quadrant and less bike use in the red area, potentially related to factors like population density or transportation planning.

### 3.2.5 DBSCAN Clustering Model Solution and Analysis

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm that groups similar density data points without requiring the number of clusters in advance. It can identify clusters of arbitrary shapes, unlike K-means which assumes convex clusters. The algorithm efficiently clusters data, handles noise points, and identifies outliers during clustering.

We employed the pandas library for data processing, implemented DBSCAN clustering from the sklearn.cluster library, and utilized matplotlib.pyplot for plotting. StandardScaler from sklearn.preprocessing ensured feature standardization, while seaborn and sklearn.metrics.silhouette\_score facilitated visualization and contour coefficient calculation. We selected start and end station coordinates as feature subsets, standardized the data, and sampled 1/20 of the original dataset for clustering. Cluster analysis was performed on 26 datasets through a for loop, with the resulting graphs stored in the appendix. Figure 8 presents clustering outcomes for July-August 2013, March-April 2014, and July-August 2015 as illustrative examples.

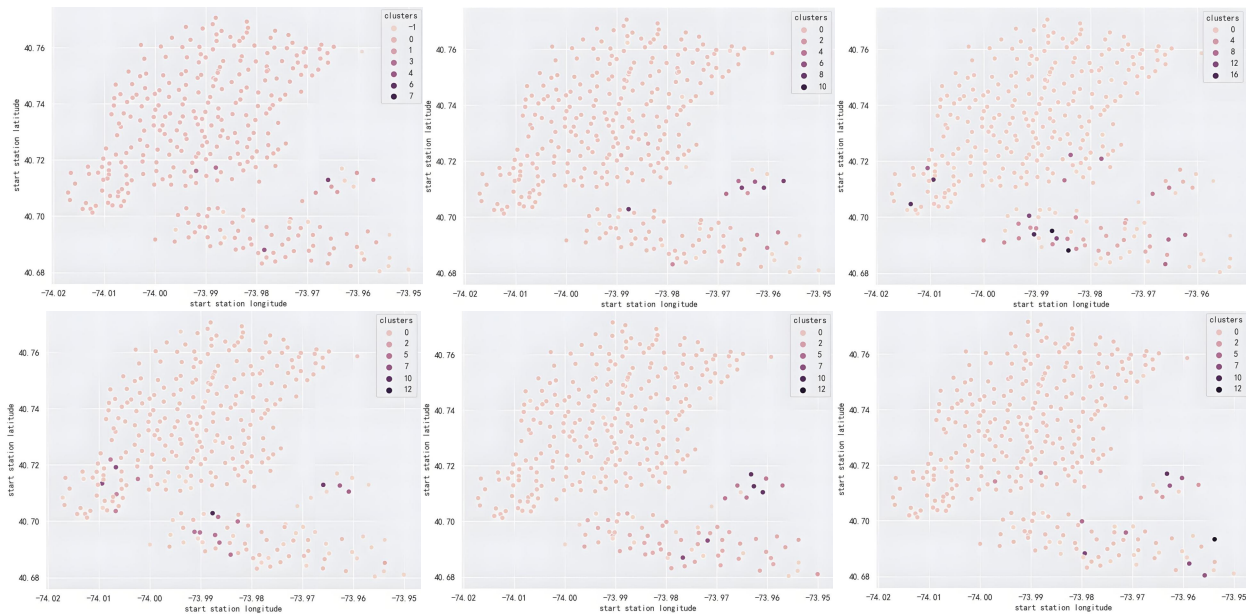


Fig. 9 The clustering results of July-August 2013, March-April 2014 and July-August 2015

Comparing K-means and DBSCAN clustering results reveals that user starting points are predominantly concentrated in two areas. K-means segments the data into four equal-sized clusters, while DBSCAN forms six clusters with varying sizes, including some containing only one or a few points[9]. This discrepancy is due to the distinct principles and parameter settings of each algorithm. K-means, based on distance metrics, is straightforward but susceptible to outliers and requires predefined cluster numbers. DBSCAN, a density-based method, identifies clusters based on the density around points, making it robust to outliers and capable of automatically determining cluster numbers. However, its results depend on parameter choices and may struggle with

high-dimensional data or irregular clusters. The choice of clustering algorithm should align with the specific problem's characteristics and requirements. For even data partitioning with known cluster numbers, K-means is suitable. DBSCAN is preferable for identifying regions of different densities and outliers, especially when the number of clusters is unknown.

#### **4. Assessment and Enhancement of the Model**

In this study, we utilized mathematical models and algorithms to predict public bicycle usage, thoroughly evaluating these methods. We used the Spring algorithm for network graph layout, reducing node overlap and enhancing readability. The networkx library was employed for network modeling and analysis, offering diverse functionalities. We applied both K-means and DBSCAN clustering algorithms, each suitable for different scenarios. These approaches not only supported public bicycle system optimization but also provided insights for other sectors like food delivery and logistics. Our research offers solutions for public bicycle systems and supports decision-making across various domains.

#### **5. Conclusions**

The integration of public bicycles into urban transportation systems has emerged as a popular strategy for mitigating traffic congestion and air pollution, as well as addressing the "last mile" problem. This study investigated the usage patterns and conducted predictive analyses of public bicycles in New York City using the CitiBike dataset from 2013 to 2015. We employed the Spring algorithm for networkx visualization and analyzed the loan and return data to elucidate spatial-temporal usage dynamics. Additionally, clustering techniques, including K-means and DBSCAN, were applied to the dataset to uncover usage trends and patterns. The findings provide valuable insights for the planning and optimization of public bicycle systems, such as CitiBike, and contribute to the promotion of sustainable and environmentally friendly urban transportation.

#### **References**

- [1] Li J, Ren C, Shao B, et al. A solution for reallocating public bike among bike stations[C]//Proceedings of 2012 9th IEEE International Conference on Networking, Sensing and Control. IEEE, 2012: 352-355.
- [2] Fang Y, Song Y, Chen D, et al. A location-routing problem for the public bike-sharing system with service level[C]//2019 International Conference on Industrial Engineering and Systems Management (IESM). IEEE, 2019: 1-5.
- [3] Sathishkumar V E, Park J, Cho Y. Using data mining techniques for bike sharing demand prediction in metropolitan city[J]. Computer Communications, 2020, 153: 353-366.
- [4] Liang Y, Huang G, Zhao Z. Bike sharing demand prediction based on knowledge sharing across modes: A graph-based deep learning approach[C]//2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2022: 857-862.
- [5] Jiang W. Bike sharing usage prediction with deep learning: a survey[J]. Neural Computing and Applications, 2022, 34(18): 15369-15385.
- [6] Zhou J, Guo Y, Sun J, et al. Review of bike-sharing system studies using bibliometrics method[J]. Journal of traffic and transportation engineering (English edition), 2022, 9(4): 608-630.
- [7] Ma C, Liu T. Demand forecasting of shared bicycles based on combined deep learning models[J]. Physica A: Statistical Mechanics and its Applications, 2024, 635: 129492.
- [8] Collini E, Nesi P, Pantaleo G. Deep learning for short-term prediction of available bikes on bike-sharing stations[J]. IEEE Access, 2021, 9: 124337-124347.
- [9] Yao X, Shen X, He T, et al. Demand estimation of public bike-sharing system based on temporal and spatial correlation[C]//2018 4th international conference on big data computing and communications (BIGCOM). IEEE, 2018: 60-65.