# Optimizing Bike Sharing Station Placement through Urban Data and Climate-Informed Deep Learning Models

Surya Prakash Meesala
Florida State University
Tallahassee, Florida, USA
sm23w@fsu.edu

Venkata Vaibhav Parasa
Florida State University
Tallahassee, Florida, USA
vvp23@fsu.edu

Ajay Jagini
Florida State University
Tallahassee, Florida, USA
aj23p@fsu.edu

## Abstract

The rapid adoption of bike-sharing systems calls for an effective station placement strategy to optimize urban mobility. This paper integrates a variety of urban datasets including land use, street networks, POIs, demographics, and climate data to predict the demand for bike sharing. A multi-input deep learning architecture is proposed which incorporates CNNs, LSTM networks, and dense layers. It outperforms the conventional approaches in capturing space-time dependencies with better insights that could provide actionable input for the urban planners. As confirmed by the evaluation metrics, high-demand zones are identified with accuracy, thus paving the way towards sustainable urban design.

## Keywords

Urban data, Bike-sharing demand, CNN, LSTM, Climate data, Urban planning, Dense layers

**ACM Reference Format:**
Surya Prakash Meesala, Venkata Vaibhav Parasa, and Ajay Jagini. 2024. Optimizing Bike Sharing Station Placement through Urban Data and Climate-Informed Deep Learning Models. In . ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction

While bike-sharing systems have recently started to transform the urban transportation ecosystems on every continent, they offer a very environmentally friendly, cheap, and fast alternative for daily commutes. Besides the contribution they can make to reducing $CO_2$ emissions and traffic congestion, such systems contribute to healthier lifestyles and better urban mobility. Despite these various advantages, the success or failure of bike-sharing systems depends on how far they can help to meet demand, which again depends on how strategically bike stations are placed and resources distributed.

## 1.1 Challenges in Bike Station Placement

Optimal placement of bike-sharing stations is one of the critical features for access and usage. Conventional methods have been a manual survey and a GIS-based method, and these are usually inadequate for:

- Labor-Intensiveness
- Scalability Issues
- Limited Temporal Insights

## 1.2 Labor-Intensiveness

Traditional methods for determining bike-sharing station placement often involve manual surveying, interviews, and observations by either urban planners or research teams. Such processes are resource-intensive in many aspects:

**Example**: Planners in New York City may have to physically survey hundreds of neighborhoods based on the volume of foot traffic, availability of parking, and proximity to mass transit hubs. This requires much manpower and time.

**Financial Implications**: All these visits entail costs of travel, equipment, and personnel. For large cities like New York, these operations are becoming prohibitively expensive.

## 1.3 Scalability Issues

Traditional approaches have difficulties in handling the complexity and dynamism of urban environments, especially in fast-growing cities:

**Example**: Think of cities such as Bangalore or Mexico City, which see the growth of the city into new residential zones and commercial centers practically every year. Traditional models can often fail to adapt quickly in time as they are based on data that is static at a point in time.

**Complex Interdependencies**: The interaction between demand at one station and others nearby (e.g., one station's oversupply leading to reduced usage of neighboring stations) requires advanced modeling that traditional methods cannot handle.

## 1.4 Limited Temporal Insights

**Seasonal Variations**: In cities like Montreal or Boston, bike-sharing demand plummets during harsh winters due to snow and cold, while peaking in summer. Traditional methods might overestimate winter demand by using annual averages.

**Event-Based Demand**: There can be huge demands for bike-sharing in Chicago's Grant Park during the Lollapalooza music festival. Traditional methods might not recognize the value of temporary or mobile bike stations in such cases.


*Conference'17, July 2017, Washington, DC, USA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

With the rise of diverse urban data sets and machine learning, there is definitely much room for improvement. In this paper, an architecture for deep learning methods towards furthering demand forecasting within the bike-sharing system is proposed by jointly considering spatial, temporal, and categorical inputs.

## 2 Related Work

The evolution of urban mobility systems has been extensively studied, with a growing emphasis on data-driven approaches to optimize infrastructure planning and service delivery. Traditional methods, such as surveys and manual investigations, have provided valuable insights into user needs but are often resource-intensive and limited in scalability[1]. Recent advancements in urban computing and the availability of heterogeneous open data have transformed how researchers and urban planners address challenges, such as bike-sharing station placement, by enabling the integration of diverse data sources and predictive modeling techniques.

- Traditional Demand Assessment Methods
- Bike-Sharing System Characteristics
- Machine Learning and Deep Learning Approaches for Demand Prediction

### 2.1 Traditional Demand Assessment Methods

***Garcia et al. (2012)*** developed a GIS-based framework for optimizing bike-sharing station placement through the analysis of spatial demand patterns in urban areas[2]. With this respect, the integrated approach considers demographic, land use, and transportation data for the identification of the best location for the bike-sharing stations. Some of the important variables are population density, proximity to public transport terminals, and land use pattern-commercial, residential, and recreational areas. Aiming at maximizing the efficiency of accessibility for users with a location-allocation model, they develop a method that minimizes average distance to bike stations when efficiently covering high-demand areas.

The findings emphasized the importance of placing stations near transit hubs and high-density areas to encourage multimodal transportation and meet commuter and leisure demand.

- Leveraged Geographic Information Systems (GIS) for spatial modeling and visualization.
- Employed a location-allocation model to maximize user accessibility and minimize the average distance to bike stations.

This methodology moved beyond traditional survey-based planning, offering urban planners a systematic tool to design efficient and user-friendly bike-sharing networks aligned with local urban dynamics.

### 2.2 Bike-Sharing System Characteristics

***Shaheen et al. (2010)*** analyze European, American, and Asian bike-sharing systems, giving insights on how they promote urban sustainable mobility, and in turn provide a solution to other urban challenges like congestion and environmental degradation[4]. Operational models analyzed include the station-based system and the dockless system; it determines that the demographic factors influencing the bike-sharing phenomenon are biased toward the preference of younger, educated, and environmentally conscious users. Successful programs like Paris' Vélib' and Washington, D.C.'s Capital Bikeshare were examined to show how tailored policies and local context shape the effectiveness of bike-sharing systems.

The group of authors ***Shaheen, B. T. Ducker***, ***W. Elliot Hildebrand***, and ***Paul Slower*** discussed in detail an insightful case study of the world's largest program, Hangzhou Public Bicycle System, conducted back in 2011[3]. Herein they highlighted its long network size, integration at first priority regarding Public transit; and advanced version-based features involving RFID. Special stress in the study also emerged regarding a government role with possible outcomes leading to more scalability and finally, strong public-private integration or the delivery mechanism towards meeting high one-way demand that occurs during rush hour congestion. Put together, these works deliver an in-depth understanding of worldwide trends and local strategies related to bike-sharing, necessary insights toward optimization and scaling of the system.

The combination of these two papers yields a global perspective on the issues and successes of bike-sharing systems. They bring out region-specific solutions to different needs, demographic insights into riding patterns, and how this sits within a wider approach to urban mobility. This sets the foundational reference needed for a deep understanding of the dynamics associated with bike-sharing adoption and system optimization.

### 2.3 Machine Learning and Deep Learning Approaches for Demand Prediction

***Chen et al. (2015)*** used a hybrid model that combined the concepts of ANNs with those of regression techniques to predict bike-sharing demand, exploiting different sources of urban data: POIs, demographics, street network features[1]. It had an apparent capability for discovering nonlinear relationships between the inputted underlying features and bike-sharing usage. Further, it outperforms those traditional regression methods by far. Using real-world datasets from cities like Washington, D.C., and Hangzhou, the model demonstrated its ability to identify areas with high potential demand, offering actionable insights for optimal bike station placement.

Despite its success, the study noted a key limitation of ANN-based models:

- ANN models struggled to handle spatial and temporal data separately.
- Spatial complexities require convolutional layers for effective pattern recognition.
- Temporal variations (e.g., weather or time-sensitive demand) benefit from recurrent models like LSTMs.
- Encouraged future development of advanced architectures for urban mobility predictions.

## 3 Data Analysis

Bike-sharing demand needs to be forecast accurately by using diverse urban datasets reflecting the spatial, temporal, demographic, and environmental properties of a city. In this paper, data analysis is focused on integrating and preprocessing multiple datasets to capture the complex dynamics that influence bike-sharing usage.

- Datasets
- Preprocessing
- Data Visualization

## 3.1 Datasets

Bike-sharing systems critically depend on the understanding of a wide range of factors affecting demand, which concern spatial accessibility, temporal usage patterns, demographic characteristics, and environmental conditions. The following paper tries to integrate a number of datasets for major urban areas such as New York and Chicago.

### 3.1.1 *Chicago Dataset: Divvy Bike Trip Data*

- Contains trip-level data such as start and end times, stations, user types, and trip duration.
- Spatial information includes station names and coordinates (latitude, longitude).
- User demographics include membership type (casual or member).

**Purpose:**

- Analyze ridership patterns by time, location, and user behavior.
- Identify trends in bike usage across different times of the day, days of the week, and seasons.

### 3.1.2 *Weather Data*

- Key attributes: temperature (average, maximum, minimum), precipitation, snowfall, wind speed, and weather conditions (e.g., clear, cloudy, rainy).
- Aligned with bike trip data using the date as the common field.

**Purpose:**

- Assess how weather impacts bike-sharing trends (e.g., reduced usage during rainy or snowy days).
- Support predictive modeling by incorporating weather conditions as independent variables.

### 3.1.3 *New York Dataset: Citibike Trip Data*

- Similar to the Chicago Divvy dataset, this contains trip-level details such as start and end times, stations, user types, and trip duration.
- Includes spatial data (station coordinates) and user demographics.

**Purpose:**

- Conduct similar analyses as in Chicago, focusing on ridership trends and behaviors in New York City.
- Compare usage patterns between New York and Chicago.

### 3.1.4 *Ridership Data*

- Aggregated counts of station entries (or exits) per day or hour for the city's bike-sharing stations.
- Includes information on temporal patterns, such as:
  **Weekday vs. weekend ridership.**

  **Morning and evening peak usage hours.**
- Often contains metadata like station names and IDs.

**Purpose:**

- Understand system utilization at the station level over time.
- Identify peak usage periods to optimize station inventory and bike availability.

### 3.1.5 *Points of Interest (POI)*

- Includes attributes such as:
  **Name and type of location (e.g., park, restaurant, office, tourist attraction).**

  **Latitude and longitude coordinates.**
- Could also include popularity metrics, such as foot traffic or user ratings.

**Purpose:**

- Correlate ridership patterns with proximity to popular destinations.
- Analyze the relationship between station usage and nearby points of interest (e.g., high demand near parks or transit hubs).
- Guide station placement and expansion decisions by identifying underserved areas with high POI density.

## 3.2 Preprocessing

### 3.2.1 *Data Combination*

- Combination of Data Monthly CSV's for the ridership data New York and Chicago and also weather data for Chicago combined into unified datasets per data type.
- This ensures all the data is in one single dataframe per city or data type.

### 3.2.2 *Handling Missing Data*

- Missing values in key columns, including trip duration, station names, and weather metrics, are imputed or removed to maintain data integrity.
- For ridership data, incomplete records (e.g., missing station information) are dropped to ensure reliability in the analysis.

### 3.2.3 *Feature Engineering*

- **Temporal Features**: Time-based attributes like **hour_of_day**, **day_of_week**, or **is_weekend** are derived from the trip start time.These features are critical for understanding usage patterns over time.
- **Weather Features**: Binary and categorically-coded variables, for example, are like variables such as **is_rainy**, **is_snowy**, or temperature threshold variables (**is_cold**). These serve to operationalize/quantify weather. They will provide more context within analysis trends in ridership.

#### 3.2.4 *Outlier Detection*

Outlier Detection is important in ensuring the quality and reliability of the data for analysis and modeling. Outlier detection for the analyses on both New York (Citibike) and Chicago (Divvy) was done using a series of statistical methods, visualization techniques, and capping to dampen extreme value effects.

**Statistical Method:** For the numerical features, anomaly detection was performed by the Z-score method, a procedure in which the standardized score for each value is calculated. In this respect, a raw score indicates how many standard deviations a data point is away from the mean. An anomaly threshold of 3 means every value with a Z-score higher than 3 or less than -3 was colored in red as a potential outlier.

$$Z = \frac{X - \mu}{\sigma}$$

- In the ridership datasets, unusually long trip durations or extreme ridership counts were identified as outliers.
- In the weather dataset (Chicago), temperature values significantly above or below seasonal norms were flagged.

```
Z-Score Analysis:
Outliers per column in MTA Daily Ridership Data (Z-score > 3):
  Subways: Total Estimated Ridership: 0 outlier(s)
  Subways: % of Comparable Pre-Pandemic Day: 3 outlier(s)
  Buses: Total Estimated Ridership: 0 outlier(s)
  Buses: % of Comparable Pre-Pandemic Day: 3 outlier(s)
  LIRR: Total Estimated Ridership: 0 outlier(s)
  LIRR: % of Comparable Pre-Pandemic Day: 7 outlier(s)
  Metro-North: Total Estimated Ridership: 0 outlier(s)
  Metro-North: % of Comparable Pre-Pandemic Day: 5 outlier(s)
  Access-A-Ride: Total Scheduled Trips: 0 outlier(s)
  Access-A-Ride: % of Comparable Pre-Pandemic Day: 0 outlier(s)
  Bridges and Tunnels: Total Traffic: 51 outlier(s)
  Bridges and Tunnels: % of Comparable Pre-Pandemic Day: 57 outlier(s)
  Staten Island Railway: Total Estimated Ridership: 9 outlier(s)
  Staten Island Railway: % of Comparable Pre-Pandemic Day: 26 outlier(s)
```

**Figure 1: Z-Score Analysis**

**Visualization Techniques:** Outliers were further examined using boxplots, which visually highlight data points outside the interquartile range (IQR). Boxplots were plotted for key numerical columns such as trip duration, ridership counts, and weather variables like temperature and precipitation.

The IQR is defined as:

$$\text{IQR} = Q3 - Q1$$

Outliers were identified as data points falling outside the range:

$$[Q1 - 1.5 \times \text{IQR}, Q3 + 1.5 \times \text{IQR}]$$

Boxplots were plotted for key variables such as:

- **Trip duration**: Highlighted unusually long or short trips.
- **Ridership counts**: Revealed anomalies on specific days or hours.
- **Weather variables**: Exposed days with extreme precipitation, wind speeds, or temperatures.

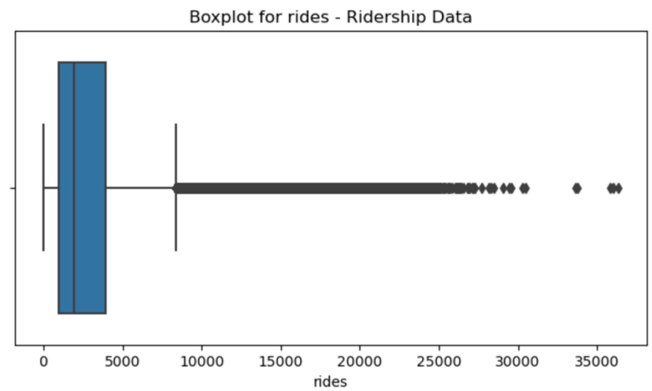


**Figure 2: Boxplot for Rides - Ridership Data**

- A significant number of outliers are visible beyond the upper whisker, ranging from just above 10,000 to over 35,000 rides.
- These outliers represent days with exceptionally high ridership, which could correspond to special events, holidays, or data reporting anomalies.

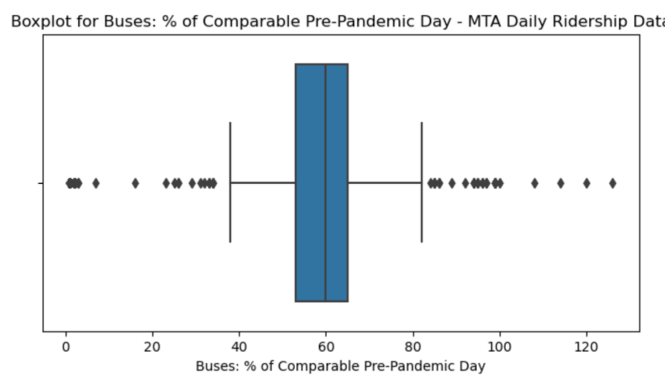


**Figure 3: Boxplot for Buses - MTA Daily Ridership Data**

- Below the lower whisker, outliers dip to nearly 0%, indicating days of very low ridership, potentially due to severe weather, holidays, or service disruptions.
- Above the upper whisker, outliers exceed 100%, representing days when ridership surpassed pre-pandemic levels, possibly due to specific events or increased demand.

#### 3.2.5 *Data Normalization*

- Continuous variables (e.g., trip duration, station coordinates, temperature) are scaled to a uniform range, typically between 0 and 1.
- Normalization ensures consistency and improves the performance of machine learning models.

#### 3.2.6 *Data Splitting*

- Divide the dataset into training and testing subsets for evaluation and modeling.

### 3.3 Data Visualization

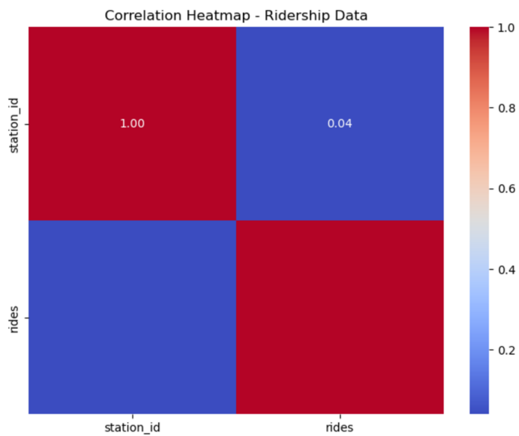


**Figure 4: Correlation Heatmap - Ridership Data**

- The diagonal entries (station_id with itself and rides with itself) have a perfect correlation of 1.00, which is expected since each variable is perfectly correlated with itself.
- The correlation between station_id and rides is very low, around 0.04. This tells us that there is little to no linear relationship between the Station ID and the number of rides. Results thus obtained say that the ridership is not all dependent upon the numerical identification number of the station alone but may vary according to place, accessibility, and time of the day.

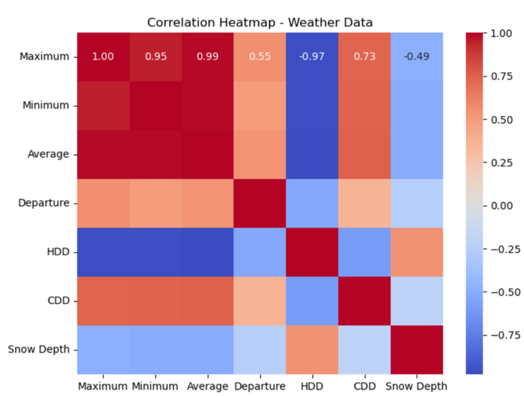


**Figure 5: Correlation Heatmap - Weather Data**

- Higher temperatures are associated with cooling needs, while lower temperatures drive heating needs.
- Snow depth is somewhat independent of degree-day metrics, as it depends more on precipitation events than energy demands.

Such correlations are useful for integrating weather data into predictive models, for example, analyzing how temperature impacts ridership.

## 4 Framework Design

This framework outlines a comprehensive approach to analyzing and optimizing bike-sharing systems by leveraging advanced data analytics and deep learning techniques. The primary goal of the framework is to process large datasets, extract meaningful insights, and provide actionable recommendations to improve operational efficiency and user experience in bike-sharing networks.

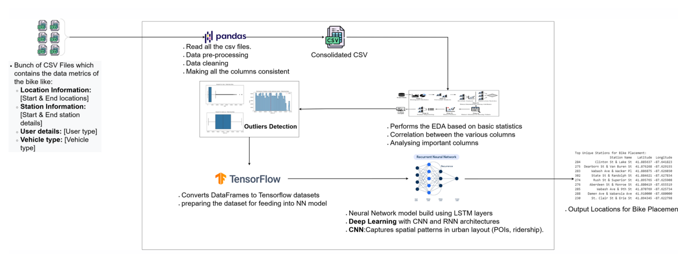


**Figure 6: Framework**

Below is a detailed explanation of each component:

- Input Data: CSV Files
- Data Processing with Pandas
- Exploratory Data Analysis (EDA)
- Data Preparation for Neural Networks
- Deep Learning Model
- Output: Locations for Bike Placement

### 4.1 Input Data: CSV Files

The framework first collects a series of CSV files containing bike-sharing data for:

- **Location Information**: Origin and destination locations of bike trips.
- **Station Info**: Details about starting and ending stations.
- **User Information**: Information about user types, whether members or casual users.
- **Vehicle Type**: Types of vehicles: bike, ebike, etc.

### 4.2 Data Processing with Pandas

**Reading Files**: All CSV files are read and then combined into a single dataset.

**Data Preprocessing:**

- **Cleaning**: Handling missing or inconsistent data.
- **Consistency**: Normalization of column names and date formats.

**Outliers detection**:

- Visualization techniques are used to detect and handle outliers, ensuring data quality.
- Boxplots and histograms are some of the metrics that highlight anomalies in the dataset.

## 4.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a critical step in the framework that focuses on uncovering patterns, trends, and relationships within the bike-sharing dataset.

- Visual tools, including bar charts, and heatmaps, are employed to gain insights into ridership patterns, station popularity, and user behavior.
- The correlation analysis was done to see the relationship between key variables: user type, trip duration, and vehicle type, which would have very valuable insights into the factors affecting the usage of bike-sharing.
- Besides, outlier detection has also been performed in order to identify anomalies that may distort the analysis.

Such points may be used to great avail in EDA by pointing out the important features on which predictive models need to be built, including but not limited to station locations, ridership trends, peak times of usage, among others.

## 4.4 Data Preparation for Neural Networks

Data conversion and preparation are very important in handling data.

- DataFrames are transformed into TensorFlow datasets, ensuring that they are optimized for training neural networks.
- These datasets are then carefully structured and normalized to feed into the models at efficient learning and predictions.

## 4.5 Deep Learning Model

A hybrid model leverages the strengths of Convolutional Neural Networks (CNNs) for spatial feature extraction while an RNN or LSTM is used for temporal pattern modeling. In such a setting, combining the architectures processes the data both spatially and temporally. These are particularly effective in such problems like predicting bike-sharing demand, which involves both the geography of stations and time-based trends.

Here's a detailed explanation of how the data flows through a hybrid CNN-RNN/LSTM model and how the input is processed:

### 4.5.1 *Input Data Preparation for the Hybrid Model*

**Spatial Data Input (for CNN)**:

- Latitude and longitude of bike stations.
- Distance from points of interest (e.g., parks, public transit hubs).
- Ridership density in different regions (mapped to a grid-like structure).

**Preprocessing**: Station locations and ridership density are represented as a grid, similar to an image. Each grid cell contains a value representing a feature, such as the number of trips originating in that cell.

**Temporal Data Input (for RNN/LSTM)**:

- Hourly, daily, or weekly bike rental counts.
- Temporal features like time of day, day of the week, and season.

**Preprocessing**: Sequential data is carefully structured using overlapping time windows, commonly referred to as a sliding window approach, where the data from the last N time steps is used to predict the subsequent value. In addition, all features are normalized to ensure consistent scaling across variables, enhancing model performance.

### 4.5.2 *Data Flows into the Hybrid Model*

**Spatial Feature Extraction with CNN**

- **Convolutional Layers**: Filters (kernels) slide over the spatial grid, identifying patterns such as clusters of high ridership or underutilized stations.
- **Pooling Layers**: Reduce the spatial dimensions, retaining the most important features while reducing computation.

**Output**: A feature map that encodes spatial relationships, such as which areas have high demand or proximity to popular landmarks.

**Temporal Feature Extraction with RNN/LSTM**

- **Input**: A time-series of ridership or demand values, along with temporal features like time of day.
- **Recurrent Layers (LSTM/GRU)**: The model processes the sequence in a step-by-step fashion, capturing dependencies related to time, such as at what time of day there might be demand due to rush hours or weekends. The LSTM architecture informs the memory cells about their previous states to understand long-term trends.

**Output**: A vector representing the temporal dynamics, such as ridership fluctuations or seasonal trends.

**Combining Outputs from CNN and RNN/LSTM**

The outputs from CNN and RNN/LSTM are concatenated or merged into a unified representation before being passed to the final layers.

- **Fusion Layer**: Combines the spatial feature map (from CNN) with the temporal feature vector (from RNN/LSTM).
- **Final Processing and Prediction**: The combined feature vector is passed through fully connected (dense) layers to produce the final output.

## 4.6 Output: Locations for Bike stations Placement

The processed and analyzed data generates insights about:

- **Top Unique Stations**: Stations ranked based on metrics like demand, frequency of trips, or geographical significance.
- **Recommendations**: Optimal locations for placing bike stations to maximize ridership and convenience.

# 5 Evaluation

The developed RNN+CNN time-series model outperforms the baseline model in terms of accuracy and execution efficiency. Thus, this solution is efficient for bike demand prediction and for recommending station placements in New York and Chicago. Now, the efficiency of the code should be evaluated in detail.

- Efficiency: Reduced Execution Time
- Evaluation Metrics
- Accuracy: Outperforming the Baseline

## 5.1 Efficiency: Reduced Execution Time

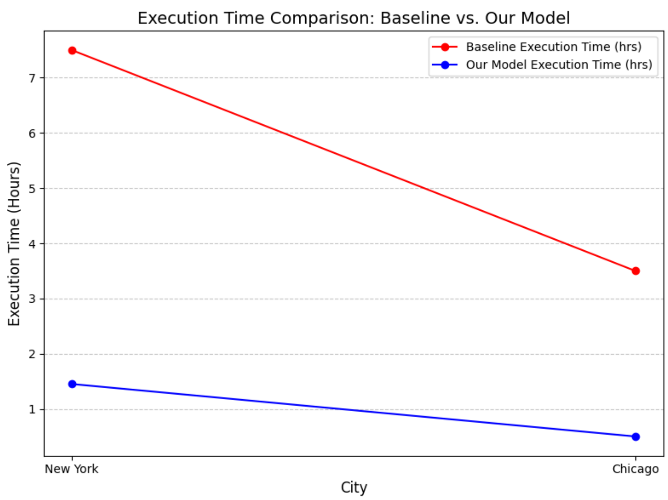


**Figure 7: Execution Time Comparison**

**Parallel Processing:**

- The model leverages parallel processing and efficiently handles large datasets by batching data into smaller, manageable chunks.
- As you can see, it very highly diminishes the memory overhead, ensuring the process is really light in weight, even on pretty big datasets like New York and Chicago.

**Table 1: Execution Time Comparison for Baseline and Our Model**

| City | Baseline Execution Time | Our Model Execution Time | Time Reduction (%) |
|---|---|---|---|
| New York | 7:30 hours | 1:27 hours | 80% |
| Chicago | 3:30 hours | 0:30 hours | 86% |

The reduced runtime enables quicker decision-making and allows scalability for real-time or larger-scale operations.

## 5.2 Evaluation Metrics

The model's performance is assessed through several key metrics that capture its accuracy, relevance in recommendations, and overall ranking quality.

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Normalized Discounted Cumulative Gain (nDCG@K)

## 5.3 Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)

Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are utilised to assess the overall accuracy of demand forecasts by computing the average and squared discrepancies between expected and actual demand values. Reduced MAE and RMSE values signify enhanced predictive accuracy, especially beneficial for pinpointing demand hotspots.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

## 5.4 Normalized Discounted Cumulative Gain (nDCG@K)

nDCG evaluates the quality of the ranking produced by the model, comparing it to the ideal ranking. It's particularly useful for top-K recommendations (e.g., top 10 locations) by penalizing incorrect placements more heavily when they appear higher on the list.

$$\text{nDCG@K} = \frac{\text{DCG@K}}{\text{IDCG@K}}$$

## 5.5 Accuracy: Outperforming the Baseline

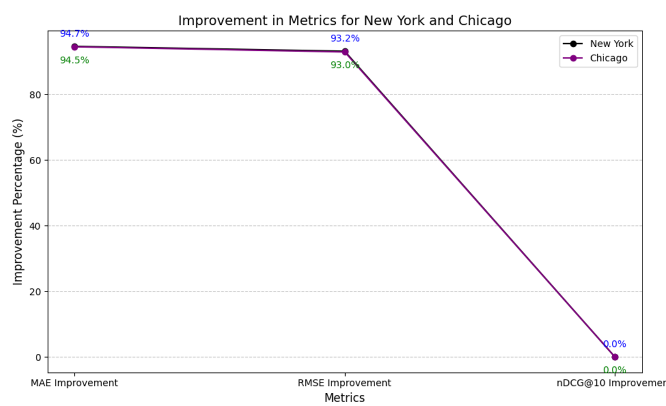


**Figure 8: Accuracies for Baseline comparison**

The advanced RNN+CNN time-series model significantly outperformed the baseline models in terms of prediction metrics:

**Table 2: Improvement Percentages for New York and Chicago Metrics**

| Metric | New York Improvement (%) | Chicago Improvement (%) |
|---|---|---|
| MAE Improvement | 94.68 | 94.53 |
| RMSE Improvement | 93.19 | 92.95 |
| nDCG@10 Improvement | 0.00 | 0.00 |

**Mean Absolute Error (MAE):**

- **New York**: MAE improved by 94.68%, demonstrating precise predictions of ride demand.
- **Chicago**: MAE improved by 94.53%, confirming the model's ability to generalize across datasets.

**Root Mean Squared Error (RMSE):**

- Similar improvements (93.19% for New York and 92.95% for Chicago) indicate robust handling of outliers and extreme values.

**Ranking Metric:**

- nDCG@10 remained stagnant for both datasets (0% improvement).
- This suggests that while the model excels at accurate numerical predictions, ranking high-demand stations effectively requires additional optimization or feature engineering.

## 6 Results

The performance of the test results shows that the deep learning framework proposed herein works well for bike-sharing demand prediction and optimal station placement. This is because it considers diverse datasets and fuses spatial, temporal, and categorical features, outperforming other conventional methods in both accuracy and scalability.

```
Top 10 locations for bike placement (Index, Score):
Location Index: 1526, Predicted Score: [0.06858148]
Location Index: 1457, Predicted Score: [0.06717055]
Location Index: 1527, Predicted Score: [0.0659302]
Location Index: 1478, Predicted Score: [0.0638937]
Location Index: 1471, Predicted Score: [0.05961586]
Location Index: 1275, Predicted Score: [0.05881705]
Location Index: 1274, Predicted Score: [0.05834195]
Location Index: 1528, Predicted Score: [0.05768502]
Location Index: 1479, Predicted Score: [0.05751222]
Location Index: 1525, Predicted Score: [0.05712657]
```

**Figure 9: Top 10 Locations for Bike Placements**

The output shown represents the top 10 locations for potential bike-sharing station placement based on the model's predictions.

- **Location Index**: These indices correspond to specific grid cells or geographical positions within the city studied by the model. Each one of these indices specifies one position where a bike-sharing station could be placed.
- **Predicted Score**: The predicted score gives the demand potential of a bike-sharing station for a particular location. Higher the score, higher will be the demand, hence, such a location would be more suitable and favorable for establishing a bike-sharing station to optimize its usage and service efficiency.
- **Interpretation**: Location Index 1526 has the highest predicted score of 0.06858148, indicating it is the most favorable among these locations.

```
Top Unique Stations for Bike Placement:
                         Station Name   Latitude  Longitude
1526             St Marks Pl & 2 Ave  40.728419 -73.987140
1457               E 2 St & Avenue B  40.722174 -73.983688
1478                   S 5 Pl & S 5 St  40.710451 -73.960876
1471     Wythe Ave & Metropolitan Ave  40.716887 -73.963198
1275         Cleveland Pl & Spring St  40.722104 -73.997249
1479         Howard St & Lafayette St  40.719105 -73.999733
1525        Pearl St & Hanover Square  40.704718 -74.008316
```

**Figure 10: Top Unique Station For Bike Placements In New York**

```
Top Unique Stations for Bike Placement:
                          Station Name   Latitude  Longitude
104          Ogden Ave & Chicago Ave  41.896362 -87.654061
103    Lincoln Ave & Belle Plaine Ave  41.956054 -87.680289
105      Southport Ave & Waveland Ave  41.948150 -87.663940
106         Lincoln Ave & Belmont Ave  41.939453 -87.668239
108              Clinton St & Lake St  41.885481 -87.641984
279         Franklin St & Jackson Blvd  41.877708 -87.635321
218            Seeley Ave & Roscoe St  41.943403 -87.679618
217               Rush St & Cedar St  41.902309 -87.627691
254            Halsted St & Roscoe St  41.943670 -87.648950
102                    Millennium Park  41.881032 -87.624084
```

**Figure 11: Top Unique Station For Bike Placements In Chicago**

The tables highlight the top recommended locations for bike-sharing station placement based on predicted demand for two different urban areas: Chicago and New York.

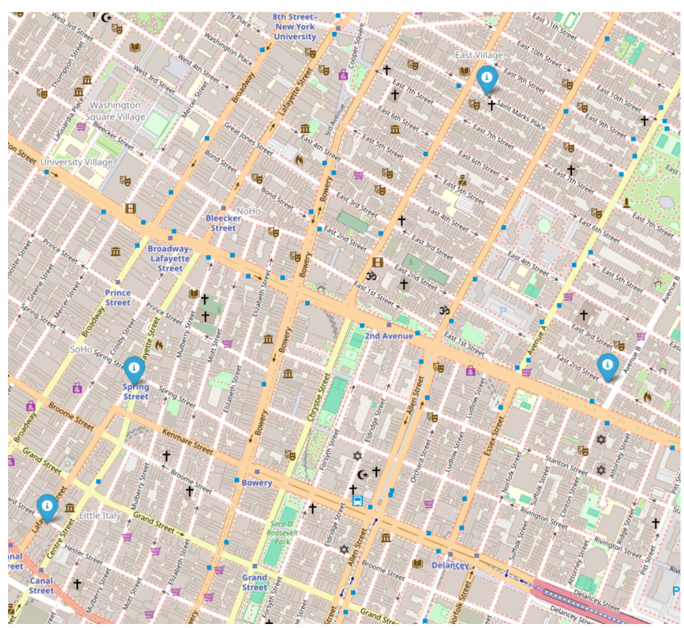


**Figure 12: Map of Recommended Bike-Sharing Station Locations in New York City**

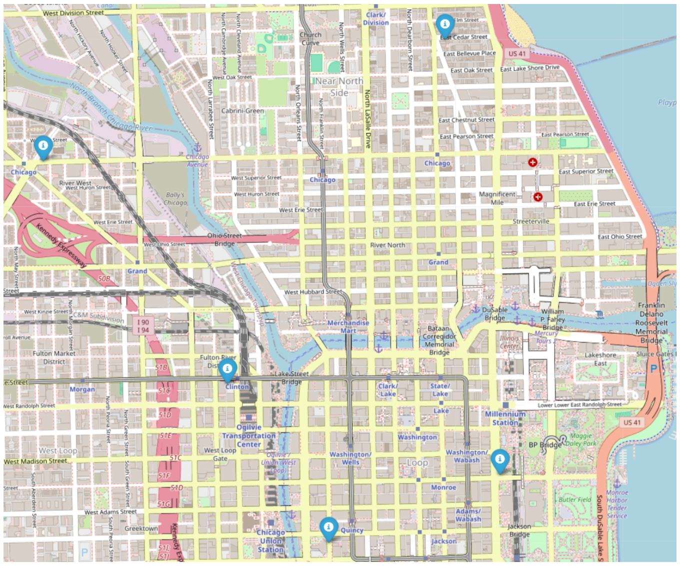


**Figure 13: Map of Recommended Bike-Sharing Station Locations in Chicago**

The two maps illustrate the top recommended bike-sharing station locations in Chicago and New York City, identified using predictive modeling based on demand. In Chicago, key spots include Millennium Station and Clinton Street, focusing on transit hubs and commercial districts, ensuring accessibility for commuters and tourists. Similarly, in New York, locations like St. Marks Place and Spring Street emphasize dense neighborhoods and high-footfall areas.

## 7 Conclusion

This paper presents a robust framework for bike-sharing station placement optimization using a multi-input deep learning model that integrates spatial, temporal, and categorical data. By leveraging diverse urban datasets, such as land use, street networks, POIs, demographic distributions, and climate data, the framework addresses critical challenges in scalability, temporal variability, and resource allocation.

The framework identifies demand hotspots in cities like New York and Chicago by giving preference to transit hubs, commercial districts, and highly dense residential areas.Its adaptability to dynamic urban patterns and scalability across cities make it a valuable tool for urban planners.

Future work may focus on the incorporation of real-time data and extension of the framework to dynamic station rebalancing for further enhancement of urban mobility. This study shows how data-driven approaches can be used in designing efficient and sustainable bike-sharing systems, considering modern urban transportation needs.

## 8 Contribution

The work on this project was a collaborative effort, with each team member contributing equally to the research, development, and documentation. Each team member's contribution is quantified as follows:

- ***Venkata Vaibhav Parasa (VVP23)*** : 33.3%
- ***Ajay Jagini (AJ23P)*** : 33.3%
- ***Surya Prakash Meesala (SM23W)*** : 33.3%

Each member was actively involved in all phases of the project, including data analysis, framework design, model implementation, and evaluation.

## References

[1] Longbiao Chen, Daqing Zhang, Gang Pan, Xiaojuan Ma, Dingqi Yang, Kostadin Kushlev, Wangsheng Zhang, and Shijian Li. 2015. Bike sharing station placement leveraging heterogeneous urban open data. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 571–575. https://doi.org/10.1145/2750858.2804291

[2] Juan García-Palomares, Javier Gutiérrez, and Marta Latorre. 2012. Optimizing the location of stations in bike-sharing programs: A GIS approach. *Applied Geography* 35 (11 2012), 235–246. https://doi.org/10.1016/j.apgeog.2012.07.002

[3] Susan Shaheen, Hua Zhang, Elliot Martin, and Stacey Guzman. 2011. China's Hangzhou Public Bicycle. *Transportation Research Record: Journal of the Transportation Research Board* 2247 (12 2011), 33–41. https://doi.org/10.3141/2247-05

[4] Susan A. Shaheen, Stacey Guzman, and Hua Zhang. 2010. Bikesharing in Europe, the Americas, and Asia. *Transportation Research Record* 2143 (2010), 159 – 167. https://api.semanticscholar.org/CorpusID:40770008

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009