CitiBike Usage Visualization

Yuvraj Maharia
The University of Texas at Dallas
yxm230002@utdallas.edu

Akhila Suggala
The University of Texas at Dallas
axs220338@utdallas.edu

Preethi Hena Dindigalla
The University of Texas at Dallas
pxd220005@utdallas.edu

***Abstract:** In recent years, the surge in popularity of bike sharing initiatives has been remarkable, as numerous cities worldwide adopt these programs as an alternative mode of transportation. Citi Bike, a prominent bike sharing program operating in New York City, stands out as one of the largest globally, boasting a network encompassing over 300 stations and 14,000 bikes spread across Manhattan, Brooklyn, Queens, and Jersey City. Despite the extensive presence of Citi Bike stations, users frequently encounter challenges in locating available bikes or empty docking stations to return their bicycles. This issue poses a source of frustration for riders, particularly when they are pressed for time to reach their destination or have limited opportunities to explore the city. Hence, comprehending the determinants influencing the demand and utilization of Citi Bike and its stations is of paramount importance.*

I. INTRODUCTION

In recent times, bike-sharing programs have emerged as a promising remedy for the transportation and environmental issues confronting urban areas globally. Bicycles play a pivotal role in public transportation owing to their eco-friendliness, affordability, and adaptability to diverse environments. Additionally, they serve as a popular recreational option. Presently, amidst the increasing emphasis on a healthier lifestyle, bicycles have gained even more traction. The advent of bike-sharing systems enables individuals to utilize bicycles conveniently and flexibly without the constraints of location or time [1].

In the 1960s, initial attempts to implement bike-sharing schemes encountered setbacks primarily due to technical limitations such as bike tracking and instant payment mechanisms. However, contemporary technological advancements, including bike tracking technologies, solar-powered sensors, mobile devices, widespread internet access, and online platforms, have played a pivotal role in turning the aspiration of bike-sharing into a tangible reality [2].

The Citi Bike initiative in New York City stands as a notable success story among similar programs, presenting an affordable, environmentally conscious, and convenient mode of transport. With a fleet surpassing 10,000 bikes and spanning across 700 stations, the program has become an integral component of the city's transportation network. Despite the widespread presence of Citi Bike stations, users have encountered challenges in securing available bikes or finding empty docks for returning their bicycles [3]. This predicament proves exasperating for riders, particularly when time is of the essence in reaching their destinations or when they have limited opportunities to explore the city. Hence, it becomes imperative to comprehend the factors influencing the demand and utilization of Citi Bike and its stations.

The program's escalating popularity has led to an increased demand for bicycles. In response, the City Bike initiative has expanded by augmenting its bicycle fleet and installing additional docking stations citywide. However, despite these expansions, various issues have emerged. These include disparities in the distribution of bicycles throughout the city, insufficient bike availability at high-traffic docking stations, and the absence of real-time visualization tools for users to monitor bicycle usage [8]. These challenges have resulted in user frustration and possess the potential to adversely affect the overall success of the program.

This paper aims to delve into the issue of uneven bicycle distribution in New York City and propose a resolution in the form of a visualization tool.

As prospective Citi Bike users, our curiosity led us to delve deeper into this issue. During our recent visit to New York City, we personally encountered the challenges associated with accessing available bikes and finding empty docking spaces, despite four stations being situated within a two-block radius of our hotel [4]. The limited bike availability surprised us, considering the extensive distribution of stations throughout the city. Our hypothesis was that multiple factors might influence the demand and utilization of Citi Bike stations, including nearby infrastructure, weather conditions, seasonal variations, user demographics, and borrowing and returning hours. Our aim was to analyze how these factors impact bike demand, the availability of docking spaces, and overall user satisfaction.

The issue plaguing the Citi Bike program in New York City is the inadequate supply of bicycles in areas experiencing high demand across the city. The current distribution of bicycles is suboptimal and fails to cater adequately to all user needs. Specific areas suffer from insufficient bicycle availability, while others have an excess. This issue holds significance as it directly impacts the overall success of the Citi Bike program [6]. If users encounter difficulties renting or returning bicycles, they may become frustrated and seek alternative modes of transportation. This potential shift could lead to reduced usage of Citi Bikes, resulting in adverse environmental and economic consequences. To ensure the success of such initiatives, studying the operations of existing public bicycle programs becomes imperative to identify features that enhance the effectiveness of bike-sharing implementations. For example, a comparison of bike-sharing schemes in China highlighted that government-backed investments, dedicated bicycle lanes, and advanced technological infrastructure significantly enhanced the performance of bike-sharing programs [7]. Making public bicycles accessible to non-registered users can boost trip numbers and introduce new traffic patterns between docking stations compared to systems exclusively reserved for subscribers.

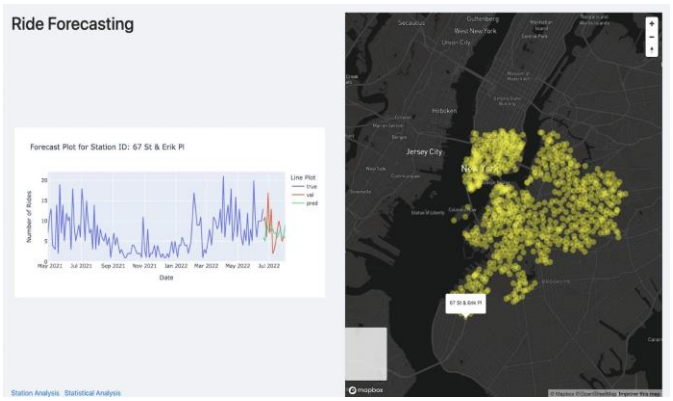

Fig. 1: Ride forecasting of a station selected on the map

To tackle this challenge, we have developed a dynamic visualization tool designed to offer users real-time insights into bicycle usage and analytics. This tool serves to furnish users with comprehensive information regarding the utilization of Citi Bike bicycles, presenting visual representations of analysis, statistical data, and future predictions concerning Citi Bike stations citywide. This data proves instrumental in comprehending the demand for Citi Bikes in diverse city zones, aiding in the optimization of station distribution. Such optimization enables users to plan their trips more effectively by steering clear of areas with limited bicycle availability or docking stations. Moreover, it assists the company's policymakers in strategic planning, allowing them to make informed decisions aimed at enhancing the overall ecosystem of the service.

II. RELATED WORK

Understanding the Usage of Public Bicycle-Sharing Systems by Chen et al. (2017) [9]: The research investigated the usage patterns within bike-sharing systems across various cities, encompassing New York City. Employing data mining techniques, the authors sought to gain insights into the factors influencing the demand for bike-sharing systems.

A Comparative Analysis of Bike Sharing Systems in Urban Ar- eas by Patsakis et al. (2018) [10]: The research conducted a comparative analysis of bike-sharing systems across multiple cities, including New York City. The authors scrutinized usage patterns, user satisfaction levels, and the influence of weather conditions on the utilization of the bike-sharing system.

Urban Mobility and Equity: A Case Study of Bicycle Sharing in Chicago by Smith et al. (2018) [11]: The research examined equity concerns associated with the utilization of bike-sharing systems within urban areas. The authors conducted an analysis of socio-economic factors impacting the usage of the bike-sharing system and put forward potential solutions to mitigate these equity issues.

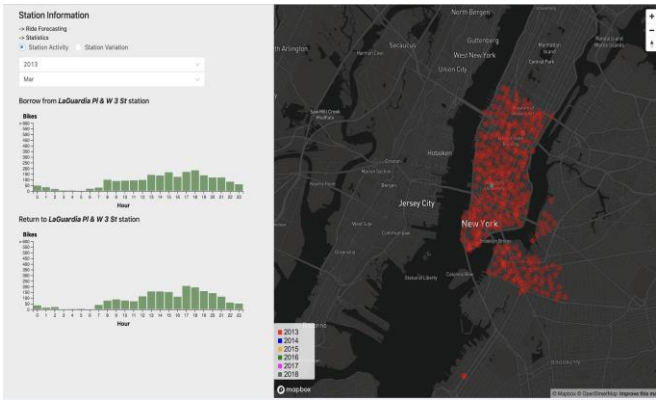

Fig. 2: Station Analysis of a station selected

Citi Bike Information Visualization by Jing Zhang, Tianyang Li, Ziwei Yuan, and Tong Lyu from University of Southern California (2018) [12]. The research paper introduces a visualization project designed to examine potential factors impacting bike-sharing orders. It leverages information visualization tools such as Angular, Bootstrap, d3, and Mapbox. The authors utilized data from sources such as Citi Bike Trip Histories and NYC Facilities to develop a comprehensive map illustrating all stations. This map includes station details such as a summary of bikes borrowed/returned and an overview of annual variations over a six-year period. Additionally, the study delved into analyzing the infrastructure's impact on the top 13 popular stations, the distribution of user age within the top six popular stations, and explored the relationship between precipitation levels and the volume of orders.

Our proposed research deviates from prior studies in several key aspects. Firstly, while previous research predominantly utilized data mining techniques, our proposed work aims to analyze the usage patterns of the Citi Bike sharing system by employing data visualization and analytics methodologies. This shift in methodology enables a different perspective in comprehending user behaviors within the system.

Secondly, our proposed research introduces modules such as Future Forecasting of Each Station, Subscription Analysis, Analyzing Top 20 Busy Stations, and Gender Analysis, which were not present in earlier studies. These additional modules contribute to a more comprehensive understanding of the Citi Bike system and its user dynamics.

Moreover, our research exclusively focuses on the Citi Bike sharing system within New York City, unlike prior studies that encompassed bike-sharing systems across various cities. This specificity allows for a more detailed and nuanced analysis tailored to the intricacies of New York City's bike-sharing landscape.

Additionally, our proposed research integrates newer technologies such as Deep Learning models and big data analytics to forecast future ride patterns within the Citi Bike system. This incorporation of cutting-edge technologies presents a novel approach that was not extensively utilized in previous research endeavors.

III. DATA DESCRIPTION

This project focusing on Citi Bike Usage Visualization and Analytics in New York City relies on the Citi Bike trip data, which spans daily trip records encompassing more than 100 million entries from 2013 to 2022. This dataset, continually updated and publicly accessible, encompasses various details

such as trip start and end points, trip durations, user demographic information, and other pertinent attributes. Additionally, it includes geographic information facilitating an analysis of Citi Bike station distribution across different neighborhoods in New York City, allowing insights into spatial usage patterns within the Citi Bike sharing system.

Through preprocessing and cleaning procedures, we transformed the original data and generated GeoJSON data spanning the last decade. Notable insights derived from this preprocessing include comprehensive details such as the total number of bikes borrowed and returned per hour at each station, segmented by month and year. Despite certain limitations like the absence of route specifics during trips (only covering trips taken within the Citi Bike sharing system), this dataset remains invaluable for analyzing usage patterns and user behaviors within New York City's Citi Bike system.

The dataset's extensive size and diverse attributes facilitate in-depth analyses of temporal and spatial usage patterns, identification of factors impacting demand and utilization, as well as the evaluation of equity and sustainability aspects within the system.

IV. RESEARCH

In our project, the primary goal is to comprehend the availability of Bike-Shares within New York City. We conducted an analysis encompassing over 100 million data entries spanning from 2013 to 2022. Our study involves creating multiple maps and plots to visually represent the distribution of bicycles. These visualizations offer valuable insights for policymaking and designing interventions, particularly for future forecasting.

Distinguishing our work from related studies, we covered a significantly longer timeframe—every month over a decade—compared to most other papers that typically focus on shorter or more specific periods. A notable disparity in prior research was its lack of specificity in analyzing New York City, which we specifically addressed in our study [13]. To accomplish this, we accessed and analyzed data unique to New York City sourced from the Citi Bike system's publicly available data feed. This dataset contained comprehensive ride details such as start and end station information, ride duration, and user-specific data.

Our approach involved the application of data visualization and analytics techniques, utilizing tools like plotly, d3 for data visualization, Python for data preprocessing, and Deep Learning algorithms for demand forecasting. Through these methodologies, we created interactive and dynamic visual representations, allowing for more efficient and effective data analysis.

Moreover, our study addressed prior research limitations by introducing new modules. For instance, the "Future Forecasting of Each Station" module was developed to predict future demand for individual Citi Bike stations. This was achieved by training machine learning algorithms on historical station data to forecast future demand. This module serves as a valuable tool for city planners, enabling them to strategize station expansions or reductions as necessary, ensuring an optimized system that aligns with user demand.

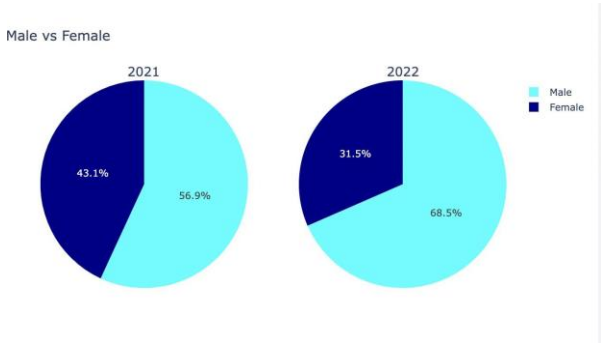

Fig. 3: Gender Analysis of Citi bike usage in years 2021 and 2022

One of the crucial modules we implemented in our study was the Gender Analysis, aimed at uncovering insights into the gender distribution within the Citi Bike system. To accomplish this, we conducted an analysis of user data, specifically identifying the gender of each user based on their ride IDs.

Our findings from this analysis revealed a notable discrepancy in usage between male and female riders within the Citi Bike system. The data showcased a higher frequency of use by male riders compared to female riders. This disparity suggests a potential opportunity for targeted marketing campaigns to encourage and increase female ridership within the system. By addressing this gender gap, initiatives can be developed to attract and engage more female users, thereby fostering a more inclusive and diverse ridership base.

A. Methods

1. **Station Analysis & Distribution Variation: [12]** This module is centered on the analysis of the distribution and accessibility of Citi Bikes throughout New York City. Through the collection and processing of data related to the geographical locations and availability of Citi Bike stations, our aim is to extract valuable insights regarding usage patterns and effectively optimize bike distribution to match demand.

   A pivotal metric under scrutiny in our analysis is the count of available bikes and docks at each station. By consistently monitoring this data over time, we can discern trends and patterns in bike usage. This analysis facilitates informed adjustments to the distribution network, ensuring a strategic allocation of bikes where they are most needed and when they are needed, thereby enhancing overall service efficiency.

2. **Future Forecasting of Each Station:** In this particular module, our approach involves employing predictive modeling techniques to forecast the anticipated future demand for each station, drawing from historical usage patterns. By leveraging this predictive modeling, we aim to provide valuable insights that can optimize the allocation of bikes across stations. This optimization ensures that stations are sufficiently stocked to meet the expected demand, thus enhancing the overall efficiency of the system.

3. **Analyzing Top 20 Busy Stations:** This module concentrates on the analysis of the top 20 busiest stations within New York City, aiming to uncover the factors contributing to their high popularity. The focus

involves exploring the usage patterns at these stations over time to discern trends and patterns. Through this analysis, we aim to extract insights into usage behaviors and identify potential areas for system enhancement and improvement.

4. **Demand of Citi Bikes Over the Years:** In this module, our objective is to analyze historical usage patterns of Citi Bikes to uncover trends in demand over the years. By discerning these trends, our goal is to derive insights that inform long-term planning strategies. This analysis aims to ensure that the Citi Bike program adapts to and fulfills the evolving needs of its users effectively.

Fig. 4: Subscription analysis of Citi bike in the years 2021 and 2022

5. **Gender Analysis and Subscription Analysis:** Ultimately, our study will encompass a gender analysis and subscription analysis to gain deeper insights into the demographics of Citi Bike users and their subscription behaviors. This examination aims to ensure that the program remains accessible and inclusive to all residents of New York City. By understanding the user demographics and subscription patterns, we aim to enhance the program's inclusivity and accessibility, catering to the diverse needs of the city's population.

B. Models

1. LSTM

Long Short-Term Memory (LSTM) models are adept at handling univariate time series forecasting tasks. As cited in reference [15], LSTM models are capable of learning a function that maps a sequence of past observations as input to predict an output observation. Initially, we employed an LSTM model for forecasting Citi Bike rides.

During this phase, we obtained Root Mean Square Error (RMSE) scores ranging between 10 to 18 for each station. These scores signify the level of error in the model's predictions compared to the actual observed values for the respective stations.

2. CNN-LSTM

In our effort to forecast demand for each station, we employed a hybrid model trained on daily ride start data [15]. Hybrid CNN-LSTM models are favored for time series forecasting due to their ability to capture both local and global temporal patterns while conducting timing analysis and extracting abstract features. This model's advantage lies in its capability to handle extended input sequences, where the CNN component interprets sub-sequences, later integrated by the LSTM model.

The dataset for each station underwent a 70/30 split for training and validation purposes. Approximately 750 stations were trained using this approach to predict future ride trends for each individual station, with a training duration of around 10 hours. Root Mean Square Error (RMSE) served as the evaluation metric, favoring smaller errors over larger ones. The validation RMSE for each station ranged from 2 to 8.

Figure 5 illustrates the Training loss versus Validation loss for both LSTM and CNN-LSTM models. The comparison demonstrates that the CNN-LSTM model outperforms the LSTM model, as evidenced by the rapid convergence of training and validation losses in the CNN-LSTM model compared to the LSTM model.

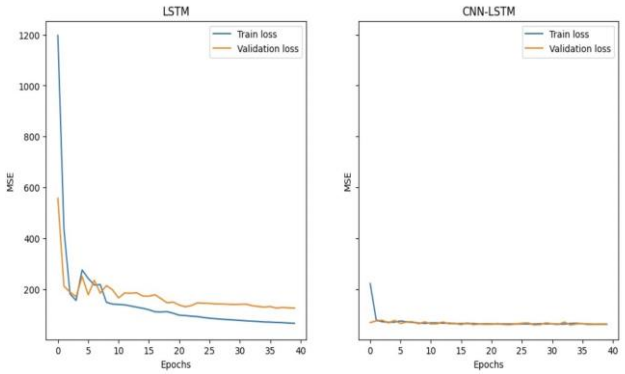

Fig. 5: Train loss vs Val loss for LSTM and CNN-LSTM

V. RESULTS

A. Station Analysis and Distribution Variation [12]:

Figure 2 showcases the station analysis module, a visualization module inspired by the University of Southern California's Citibike visualization. In this improved version, we have enhanced the dataset by incorporating data from every month, encompassing recent years from 2018 to 2022, extending the scope from the previous version. This updated version includes comprehensive monthly data. It allows users to click on individual stations, select specific years and months, and view bar charts detailing the total number of bikes borrowed and returned per selected station. Moreover, each station display provides hourly information concerning both returned and borrowed bikes, offering users detailed insights into usage patterns at various times throughout the day.

Fig. 6: Distribution variation on stations

The Distribution Variation module presents statistics derived from yearly data, showcasing the yearly fluctuations in bike ride counts across stations. Figure 6 specifically illustrates the variations between the years 2013 and 2018. A line plot is utilized to compare the number of rides on an hourly basis, providing a visual comparison of ride counts throughout the day. Below this plot, a bar chart showcases updated station information along with their respective ride counts. It's important to note that this updated version includes data from recent years, spanning from 2018 to 2022, extending beyond the data covered in previous editions.

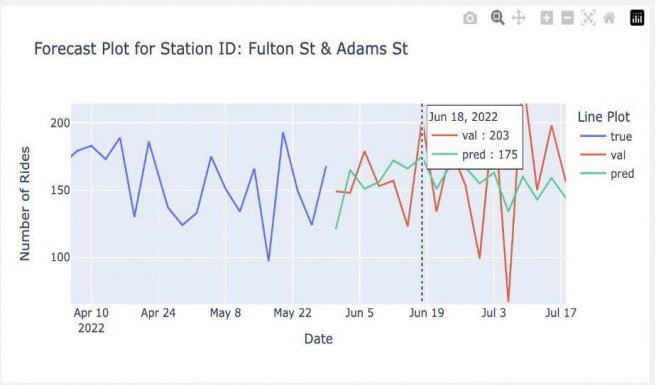

Fig. 7: Forecasting plot of a station

B. Ride Forecasting:

1. Figure 1 exhibits the future ride forecasts for a specific station when selected on the map. Each forecast plot showcases the previous trend, actual values, and predicted values of future rides.
2. The blue line represents previous ride trends, and the validation data is derived from a 3-month dataset. The X-axis displays the number of rides, while the Y-axis denotes the specific dates of the rides. The "Predicted vs True Values" plot illustrates the validation plot of the model.
3. Figure 7 provides a clear representation of previous data alongside the true versus predicted ride counts. All these plots are generated using the CNN-LSTM model, identified as the most suitable model for this particular use case. Notably, these plots are interactive and dynamic.
4. It's worth noting that these plots exhibit spikes due to the representation of daily data over the past year. A potential future improvement for this paper could involve smoothing these graphs to provide a clearer visualization of trends over time.

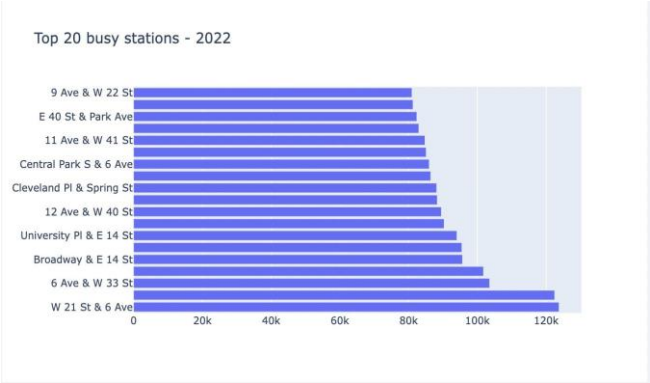

Fig. 8: Forecasting plot of a station

C. Top 20 busy stations analysis:

1. This section is part of the statistics page within our visualization dashboard. Here, we've conducted an analysis focusing on the top 20 stations exhibiting high demand for bikes specifically in the year 2022. This information serves as valuable insight for policymakers, aiding in decisions to augment the supply of Citi Bikes at these stations and potentially increase station capacity.
2. Figure 8 presents a horizontal bar chart where the X-axis denotes the number of rides, while the Y-axis represents the names of the stations. Based on our analysis, the station "W 21st & 6th avenue" recorded the highest number of rides at 123,745, followed by "6th avenue & W 33 st" with 122,460 rides. These statistics highlight the stations experiencing the highest demand, guiding policymakers in prioritizing necessary actions to meet this demand effectively.

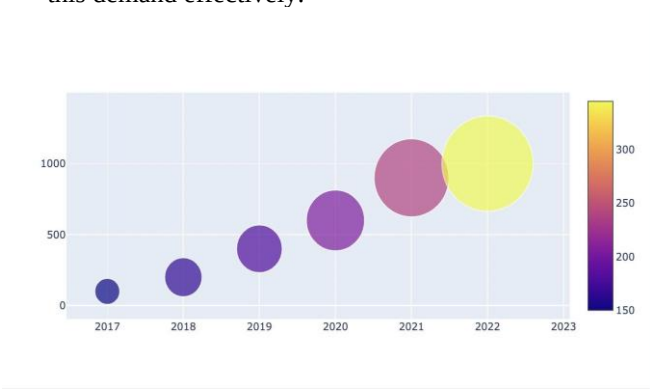

Fig. 9: Demand plot over the years

D. Ride Demand analysis:

1. In this analysis, we've examined the demand for stations using both the count of stations and the number of rides over several years, represented in a bubble scatter plot. This plot serves as a valuable tool for Citi Bike decision-makers to comprehend the demand for bikes and discern the evolving patterns over time.
2. Figure 9 displays the demand plot for Citi Bikes. The X-axis represents the total number of rides in a year (in thousands), the Y-axis indicates the respective years, and the size of the circle reflects the increasing count of stations.

3. This plot vividly illustrates the escalating demand for Citi Bikes over the years. The rapid increase in demand indicates the necessity for the organization to maintain and potentially enhance their operational standards to meet the burgeoning demand effectively.

E. Gender Analysis and Subscription Analysis:

1. Figure 3 illustrates the gender analysis of Citi Bike riders in the years 2021 and 2022, presented through a pie chart.
2. This analysis reveals that in 2022, male riders surpassed female riders by 30%. Additionally, the chart demonstrates a decline in female cyclists from 2021 to 2022. These findings highlight the necessity for the organization to focus on advertising efforts aimed at expanding female ridership significantly.
3. Figure 4 portrays the subscription analysis of Citi Bike users in the years 2021 and 2022, depicted through a donut chart.
4. This analysis showcases that members, as opposed to casual riders, demonstrate more frequent usage of the bikes. Consequently, there is an evident need for the organization to initiate campaigns and marketing strategies, promoting their application to facilitate easy access for casual riders and encourage increased utilization.

VI. FUTURE SCOPE AND CONCLUSION

There are numerous avenues for future enhancements to the dashboard, augmenting its functionality and usability. Firstly, improving the accuracy of the forecasting model by integrating more advanced algorithms and techniques stands as a key area for development. Secondly, enhancing the interactivity of visualizations will aid in user comprehension and exploration. Additionally, refining the dashboard's user interface (UI) will render it more user-friendly and accessible to a broader user base. Lastly, integrating the dashboard with other cities' bike-sharing ecosystems can offer users a comprehensive view of bike-sharing trends and patterns across multiple locations.

In summary, our team has successfully developed a dashboard for analyzing and exploring New York City's bike-sharing dataset. The dashboard offers a diverse array of visualizations and forecasting capabilities, establishing itself as a valuable tool for comprehending bike-sharing trends and patterns. Moving forward, our objectives revolve around refining the forecasting model's accuracy, enhancing visualization interactivity, improving the UI, and integrating the dashboard with other bike-sharing ecosystems. Ultimately, the dashboard holds the potential to yield valuable insights into bike-sharing trends, thereby facilitating informed decision-making within the bike-sharing industry.

VII. REFERENCES

[1] Fishman, Elliot. (2015). Bikeshare: A Review of Recent Literature. Transport Reviews. 36. 10.1080/01441647.2015.1033036. 1

[2] Leslie Lamport (1994) LATEX: a document preparation system, Addison Wesley, Massachusetts, 2nd ed. 1

[3] Faghih-imani, A., Eluru, N., El-geneidy, A. M., Rabbat, M., Haq, U. (2014). How land-use and urban form impact bicycle flows : evidence from the bicycle-sharing system (BIXI) in Mon- treal. Journal of Transport Geography, 41(August 2012), 306–314. https://doi.org/10.1016/j.jtrangeo.2014.01.013. 1

[4] Mátrai, T., Tóth, J. (2016). Comparative Assessment of Public Bike Sharing Systems. Transportation Research Procedia, 14, 2344–2351. https://doi.org/10.1016/j.trpro.2016.05.26. 1

[5] NYC Citibike Dataset, https://citibikenyc.com/system-data

[6] Volume, M. E., Me, I., Meng, O. (2011). Implementing bike-sharing systems, 164. https://doi.org/10.1680/muen.2011.164.2.8. 1

[7]]Zhao J, Wang J, Deng W. Exploring bikesharing travel time and trip chain by gender and day of the week. Transportation Research Part C: Emerging Technologies. 2015; 58, Part B:251–64. 1

[8] Goodman A, Cheshire J. Inequalities in the London bicycle sharing system revisited: impacts of extend- ing the scheme to poorer areas but then doubling prices. Journal of Transport Geography. 1

[9] Understanding the intention to use bike-sharing system: A case study in Xi'an, China.Xiaonan Zhang, Conceptualization, Methodology, Software, Writing – original draft, Writing – review editing,Jianjun Wang, Supervi- sion, Xueqin Long, Data curation, Validation and Weijia Li, Investigation . 2

[10] Operating Characteristics of Dockless Bike-Sharing Systems near Metro Stations: Case Study in Nanjing City, China by Yuan Li,Zhenjun Zhu andXiucheng Guo . 2

[11] Bikesharing, equity, and disadvantaged communities: A case study in Chicago August 2020 Transportation Research Part A Policy and Practice 140(3) DOI:10.1016/j.tra.2020.07.004. 2

[12] Citi Bike Information Visualization Jing Zhang, Tianyang Li, Ziwei Yuan, and Tong Lyu University of Southern California, Los Angeles CA 900007,USA 2, 3

[13] Cleaner Production, 2014. [19] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban computing: concepts, methodologies, and applications. ACM Transaction on Intelligent Systems and Technology (ACM TIST), 2014. 2

[14]X. Wang, G. Lindsey, J. E. Schoner, and A. Harrison. Modeling bike share station activity: The eFects of nearby businesses and jobs on trips to and from stations. Transportation Research Record, 43(44):45, 2012.

[15] LSTM, CNN-LSTM for time series forecasting https://machinelearningmastery.com/ 3