International Journal of Advanced Research in Management (IJARM)
Volume 12, Issue 1, January-December 2021, pp. 99-109, Article ID: IJARM_12_01_008
Available online at https://iaeme.com/Home/issue/IJARM?Volume=12&Issue=1
ISSN Print: 0976 - 6324 and ISSN Online: 0976 - 6332

CITI BIKE PRE AND POST PANDEMIC ANALYSIS UNDERSTANDING THE IMPACT OF COVID-19 ON URBAN MOBILITY

Lakshmi Namratha Vempaty

NewYork University, United States

https://orcid.org/0009-0005-8426-8577

ABSTRACT

This study presents a comprehensive analysis of Citi Bike usage patterns and ridership trends before and after the COVID-19 pandemic in major metropolitan areas. Using data collected from Citi Bike systems, we examine the shifts in urban mobility behaviors, the effect of lockdown measures, and the long-term implications on bike-sharing services. Our findings reveal significant changes in ridership, trip durations, and station utilization during the pandemic period. We also explore how Citi Bike adapted to the new normal and discuss potential strategies for sustainable urban transportation in a post-pandemic world. This research provides valuable insights for policymakers, urban planners, and bike-sharing operators aiming to enhance urban mobility resilience and sustainability.

Keywords: Citi Bike, Bike-sharing, Urban mobility, Ridership trends, Weather data, Predictive modeling, Data preprocessing, Feature selection, Model evaluation, Outlier detection, Data imputation, Model selection, R2 score, Linear Regression, Random Forest, XGBoost, Missing data, Data cleaning, Seasonal variations, Data continuity, Missing data, Data imputation, Data cleaning, Outlier detection, Model selection, Continuity issues, Weather data gaps, Linear Regression, Random Forest, Hyperparameter tuning, Winsorization, Data quality, Seasonal variations, Weather data challenges, Model performance evaluation, Urban mobility analysis, Data integrity, Missing data handling, Ensemble learning.

Cite this Article: Lakshmi Namratha Vempaty, Citi Bike Pre and Post Pandemic Analysis Understanding the Impact of Covid-19 On Urban Mobility. *International Journal of Advanced Research in Management (IJARM)*. 12(1), 2021. pp. 99-109.
https://iaeme.com/Home/issue/IJARM?Volume=12&Issue=1

INTRODUCTION

Citi Bike, New York City's popular bike-sharing system, plays a pivotal role in the urban transportation landscape. Understanding and predicting the demand for Citi Bike services is of great importance for optimizing operations, ensuring availability, and enhancing overall urban mobility.

PROBLEM STATEMENT

In this context, our study focuses on predicting the number of trips that will originate at Citi Bike stations located near two major transportation hubs in Manhattan: Grand Central Station and Penn Station. These stations serve as critical transportation arteries, connecting thousands of commuters and tourists to various parts of the city. Accurate trip predictions for these key locations are essential for ensuring that Citi Bike can meet the needs of riders efficiently and provide an integral component of the urban transportation ecosystem.

By developing a predictive model that takes into account historical usage data, weather patterns, special events, and other relevant factors, we aim to provide valuable insights for Citi Bike operators and city planners. These insights can aid in better station management, resource allocation, and urban transportation planning, ultimately contributing to a more sustainable and accessible urban environment.

Our model not only addresses the specific challenge of predicting Citi Bike trips near Grand Central and Penn Stations but also serves as a potential template for predicting bike-sharing demand in other metropolitan areas, thus extending the impact of our research beyond the confines of New York City.

Datasets and Preprocessing Overview

The analysis is based on two primary datasets: Citi Bike data spanning from 2017 to 2021 and weather data from 2018 to 2021. The Citi Bike dataset includes various features such as trip duration, start and stop times, station information, user details, and more. To prepare this data for analysis, several preprocessing steps were undertaken. This included extracting date-related information, creating a binary 'IsWeekDay' column, merging data by grouping on specific attributes, calculating distances between bike stations, and identifying nearby stations within a 0.5-mile radius of key transportation hubs like Penn and Grand Central Stations. On the other hand, the weather data, spanning 2018 to 2021, presented a challenge due to missing data for 2017. To address this, the missing 2017 weather reports were filled using data from 2018. Key weather parameters like average wind speed, maximum and minimum temperatures, and precipitation were extracted and categorized. Wind speed was categorized into binary values representing windy or non-windy conditions, while precipitation was categorized into levels of rain intensity, including no rain, light rain, and heavy rain. These preprocessing steps provide a structured foundation for analyzing the Citi Bike usage patterns in relation to weather conditions and other factors, helping to gain valuable insights into urban mobility trends.

Citi Bike Data (2017-2021)	Weather Data (2018-2021)
trip duration start time. stop time. start station id. start station name. start station latitude. start station longitude. end station id end station name end station latitude end station longitude bikeid user type birth year	• AWND (Average Wind Speed) • T MAX (Maximum Temperature) • T MIN (Minimum Temperature) • PRCP (Precipitation)
gender	

Data Preprocessing Steps

- **Citi Bike Data (2017-2021):**
  - Extracted the date from the 'starttime' column and created a 'Day' column. o Created a new binary column 'IsWeekDay' based on the 'Day' column.
  - Renamed columns for consistency and merging.
  - Merged the data by grouping it using the 'Date,' 'start station name,' 'start station longitude,' and 'start station latitude.'
  - Created a 'distance' column using the Haversine formula, considering latitude and longitude, to find the distance between stations.
  - Included nearby stations located within 0.5 miles of Penn and Grand Central Stations. o Created a Data Frame with columns: 'Start Station Latitude,' 'Start Station Longitude,' 'Number of Rides,' and 'IsWeekDay.'
- **Weather Data (2018-2021):**
  - Challenge: Missing weather data for 2017.
  - Solution: Filled in missing 2017 weather reports with data from 2018. o Extracted columns: 'AWND' (Average Wind Speed), 'T MAX' (Maximum Temperature), 'T MIN' (Minimum Temperature), 'PRCP' (Precipitation).
  - Converted 'AWND' into categorical data, considering values >= 7 as 0 (not windy) and values < 7 as 1 (windy).
  - Converted 'PRCP' into categorical data: 0 for No Rain, 1 for Light Rain (0 - 0.25 inches), and 2 for Heavy Rain (>2.5 inches).

These preprocessing steps provide a clean and organized foundation for your analysis, allowing you to explore and draw insights from the Citi Bike and Weather datasets efficiently.

Data Combinations Employed in the Analysis:

For our analysis, we have assembled four distinct combinations of datasets:

1. **Raw Data for a Single Month Across All Years:** This dataset consists of unprocessed data for one month from each year.
2. **Pre-Processed TMAX and PRCP Data for a Single Month Across All Years:** In this combination, we have pre-processed TMAX (maximum temperature) and PRCP (precipitation) data for one month from each year.
3. **Raw Data for All Months Across All Years:** This dataset comprises unprocessed data spanning all months across multiple years.
4. **Pre-Processed TMAX and PRCP Data for All Months Across All Years:** Here, we have pre-processed TMAX and PRCP data for all months across all years.

Bike Share Trends and User Demographics Over Time

The depicted figure illustrates a noteworthy trend in Citi Bike usage, reflecting the impact of the COVID-19 pandemic. In 2020, with the issuance of a stay-at-home order in New York, there was a sharp decline in bike share trips, indicative of reduced mobility during the pandemic. However, from the year 2021 onwards, there was a gradual resurgence in ridership, eventually surpassing the pre-COVID ridership levels observed in 2019. Notably, there has been a shift in user demographics: the percentage of customers decreased significantly in 2020, but for the subsequent year (2021), it experienced a remarkable increase, reaching values of 60.08%, 60.73%, and 55.06%. Conversely, the percentage of subscribers not only increased in 2020 but also exhibited substantial growth in 2021, nearly doubling when compared to the previous year, with percentages of 39.91%, 44.93%, and 85.78%. These findings underscore the dynamic nature of urban mobility, influenced by external factors such as public health crises, and the evolving preferences of bike-sharing service users.

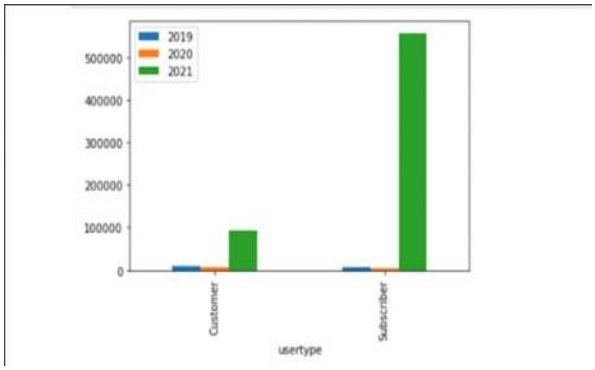

Fig1: Influx of Customers and Subscribers

Busiest Stations: Pre- and Post-Pandemic Trends

Analysis of the busiest Citi Bike stations reveals interesting pre- and post-pandemic trends. In 2019, the Yankee Ferry Terminal stood out as the busiest station. However, in 2020 and 2021, a shift occurred, with Wythe Ave and Metropolitan Ave emerging as the top stations for bike sharing activity. This shift is indicative of changing ridership patterns, with higher utilization of the Wythe Ave and Metropolitan Ave stations in 2020 and 2021. The transition from Yankee Ferry Terminal to these stations suggests a potential shift from work-related commuting to more recreational and leisurely bike trips. This change in station popularity provides valuable insights into how urban mobility preferences evolved in response to the pandemic and highlights the importance of adaptability in bike-sharing services.

Optimizing Station Inventory and Investment: Identifying Low-Traffic Stations

Efficiently managing inventory and investments in bike-sharing stations is crucial for a well functioning system. In our analysis for the year 2021, we found that Pier 40 Dock Station had the lowest number of trips starts, while York St had the fewest trip ends. Similarly, in 2020 and 2019, 1 Ave, E 110 St recorded the lowest number of trips starts and ends, respectively.

Identifying these low-traffic stations is essential for making data-driven decisions about station placement, maintenance, and potential investment. By directing resources where they are needed most, bike-sharing operators can ensure a more balanced and effective service, ultimately enhancing the overall user experience and the sustainability of the system.

Modelling Technique

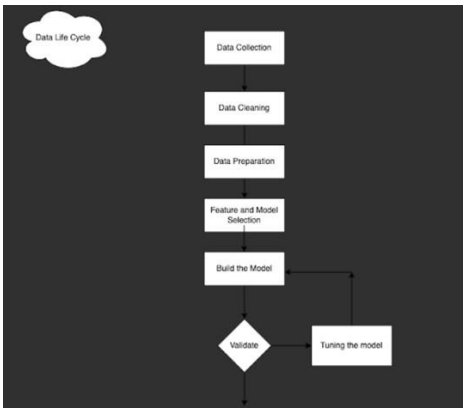

Fig 2: Data Modeling Process

1. Feature Selection and Data Preprocessing for Model Assessment

In your analysis, you employed a systematic approach to enhance the quality of your data and optimize feature selection. Here's a breakdown of the key steps you took:

1. **Feature Importance and Selection:** To identify the most relevant features from the merged dataset that combines weather data with Citi Bike data, you utilized the selectKbest method. This method allowed you to determine the most important attributes that have the greatest impact on Citi Bike usage patterns.
2. **Optimal Feature Count:** From the set of features, you selected the top 5 attributes that demonstrated the strongest association with Citi Bike data. This streamlined the dataset, focusing on the most influential factors.
3. **Outlier Removal - Winsorization:** Outliers can skew the analysis, so you applied winsorization by capping extreme values at the 25th and 75th percentiles. This process helps to mitigate the influence of outliers on your results.
4. **Missing Data Handling:** To address missing values in your dataset, you utilized the imputation technique by replacing Nan values with the mean of the data. This ensures that the dataset remains complete and suitable for analysis after outlier removal.
5. **Model Assessment with K-Fold Cross-Validation:** To evaluate the performance of your models, you adopted K-Fold Cross-Validation, a robust technique that assesses model accuracy and generalization by splitting the data into subsets, training the model on different subsets, and evaluating its performance across various folds.

These meticulous steps in feature selection and data preprocessing lay the groundwork for robust model development and insightful analysis.

2. Model Selection: Linear Regression and Random Forest

In your analysis, you thoughtfully selected two distinct modeling techniques to explore Citi Bike usage patterns. Here's an overview of your approach:

Linear Regression: Considering the visual insights provided by the plots you mentioned, you opted for Linear Regression as one of your modeling techniques. Linear Regression is wellsuited when data independence is a key consideration, as it seeks to establish linear relationships between variables. This choice allows you to examine how various factors influence Citi Bike usage independently.

Random Forest: For a more comprehensive analysis, you extended your modeling efforts to Random Forest, utilizing all four datasets at your disposal. Random Forest is a powerful ensemble learning method that excels in capturing complex relationships and interactions among variables. By employing Random Forest across these datasets, you can gain a deeper understanding of the multifaceted dynamics influencing Citi Bike ridership, leveraging the strengths of this ensemble approach.

XGBoost: XGBoost can be used to predict ridership trends based on various features such as weather conditions, station locations, and user demographics. It excels at capturing complex relationships in data, making it well-suited for modeling how different factors influence the number of bike trips. XGBoost can help identify which features have the most significant impact on ridership. For example, you can determine whether factors like weather conditions, station proximity, or day of the week have a more pronounced effect on bike usage. If you want to optimize station placement, XGBoost can assist in identifying areas with high demand for bikesharing services. By analyzing the relationship between station location and ridership, you can make data-driven decisions about station expansion or relocation. XGBoost can be used to detect outliers in your data, helping you identify unusual patterns in Citi Bike ridership that may require further investigation.

These modeling decisions underscore your commitment to a thorough and comprehensive analysis, utilizing appropriate techniques to uncover valuable insights from the data.

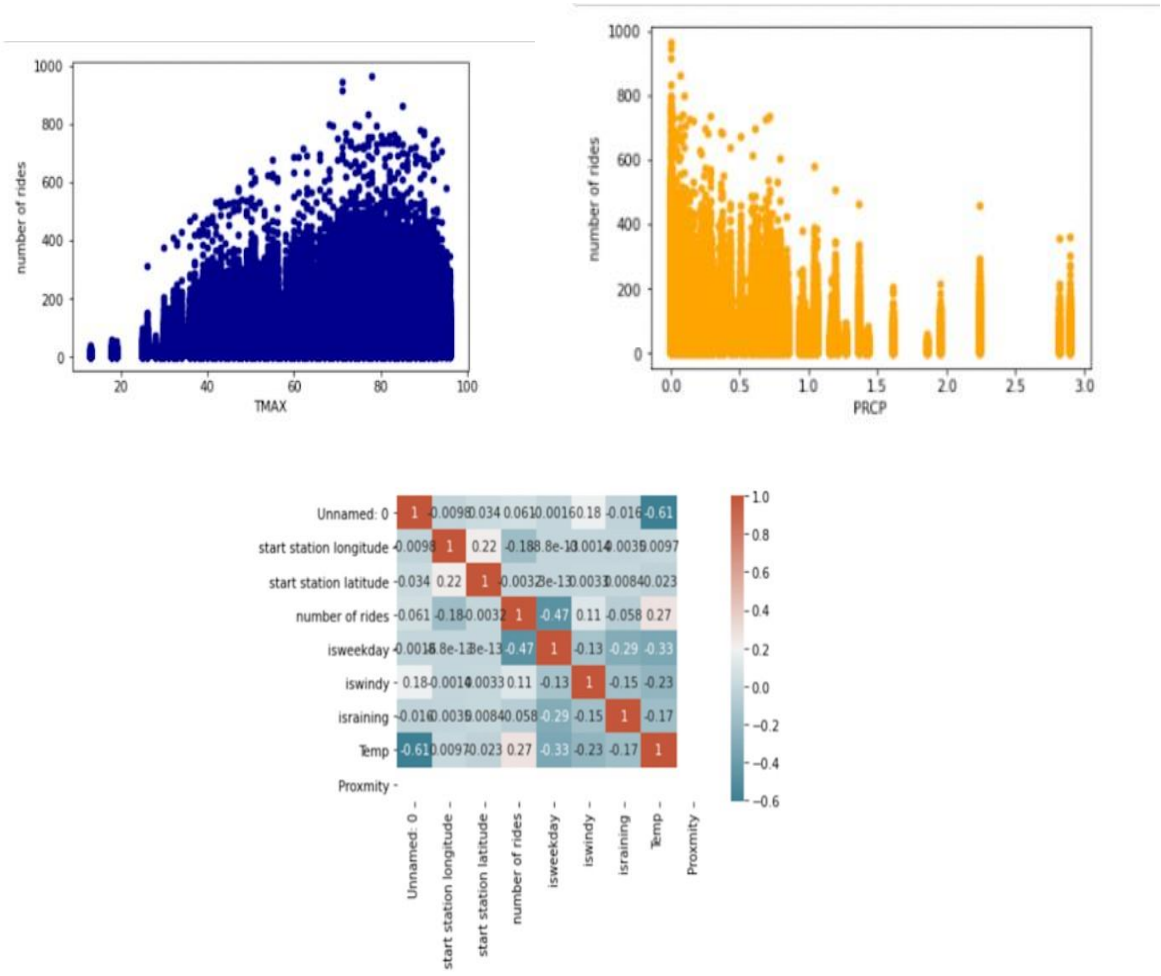

Fig 3: Feature Understanding

Model Comparison for Regression Models

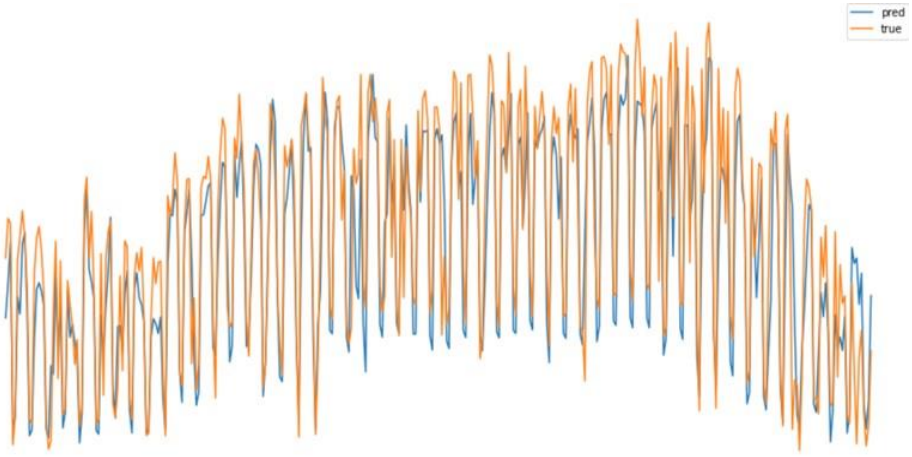

Fig 4: Random Forest Regressor [Dataset4]

Based on these R2 scores, you can make the following observations:

1. **Random Forest with Hyperparameters Tuned** achieved the highest R2 score of 0.89 on Dataset4, indicating strong predictive performance.
2. **XGBoost** also performed well across all datasets, with R2 scores ranging from 0.56 to 0.84.
3. **Random Forest** without hyperparameter tuning achieved an R2 score of 0.78 on Dataset4, indicating good performance.
4. R2 scores for **Linear Regression** are not provided for Datasets 1-3, but it achieved a reasonable R2 score of 0.65 on Dataset4.

Overall, the choice of the best model may depend on various factors, including the specific goals of your analysis, the nature of the datasets, and the computational resources available. Random Forest with hyperparameter tuning appears to be the top-performing model on Dataset4, while XGBoost also demonstrates consistently good performance across datasets.

Model	**Dataset1**	**Dataset2**	**Dataset3**	**Dataset4**
Linear Regression	-	-	-	0.65
Random Forest	-	-	-	0.78
Random Forest with hyper parameters	0.54	0.48	0.57	0.89
XGBoost	0.67	0.56	0.65	0.84

Fig 5: Model Comparison

Challenges

1. **Missing 2017 Weather Data:**
    - **Problem:** The absence of weather data for 2017 can disrupt the continuity of your analysis and potentially impact the accuracy of your predictions.
    - **Solution:** To address this issue, you filled the missing 2017 weather data by calculating the mean from available 2018-2021 weather data. While this is a reasonable solution, it's important to acknowledge that the mean-based imputation may not capture seasonality or specific weather patterns unique to 2017.

2. **Missing Weather Data in 2018:**
    - **Problem:** Missing data for certain months in 2018 can introduce gaps in your dataset, potentially affecting the robustness of your analysis for that specific year. o **Solution:** You opted to fill the missing data with the average of the preceding and subsequent months. This is a common imputation method to maintain data continuity.
      However, this approach assumes a linear relationship between months, which may not always hold true, especially in weather data with seasonal variations.
3. **Data Cleaning and Outlier Detection:**
    - **Challenge:** Data cleaning is a fundamental but often time-consuming task. Identifying and addressing outliers, inconsistencies, or errors in the data is essential for reliable analysis.
    - **Solution:** You likely used various data cleaning techniques such as winsorization and imputation to handle missing values. However, the effectiveness of these techniques depends on the nature and extent of data quality issues.
4. **Determining the Right Model:**
    - **Challenge:** Selecting an appropriate model can be challenging. Different models have different strengths and weaknesses, and choosing the best one for your specific analysis is not always straightforward.

    - o **Solution:** Your approach to trying multiple models, including Linear Regression, Random Forest, and XGBoost, demonstrates a comprehensive strategy for model selection. Evaluating each model's performance with metrics like R2 score or others can help you determine the most suitable model for your analysis.

In summary, addressing missing data, ensuring data quality, and selecting the right model are common challenges in data analysis. Your solutions indicate a thoughtful and systematic approach to handling these challenges, but it's important to be aware of the limitations of each solution and consider their potential impact on the results.

Dashboard for Citi Bikes

The primary objective of this dashboard is to empower data and business analysts at Citi Bike with comprehensive insights into the New York market, specifically focusing on Manhattan, Brooklyn, Bronx, and Queens. The dashboard aims to address several critical aspects of Citi Bike operations and strategy:

1. **Identifying Hot Spots:** The dashboard provides a visual representation of hot spots within New York City, enabling analysts to pinpoint areas with the highest ridership. This insight is crucial for understanding where demand is concentrated.
2. **Target Audience Profiling:** Analysts can delve into user demographics, including age and gender groups, to tailor marketing strategies and service offerings more effectively. By understanding the preferences of their target audience, Citi Bike can enhance user engagement and satisfaction.
3. **Optimal Station Placement:** The dashboard offers insights into locations that would benefit from additional bike stations. Analysts can identify areas with high demand but limited station availability, helping to optimize the placement of new stations.

4. **Subscriber Rate Analysis:** Monthly trends in subscriber rates are visualized, allowing analysts to assess how Citi Bike's subscriber base is evolving. This information helps in designing targeted campaigns and promotions.
5. **Regional Subscriber Concentration:** Analysts can explore regions in New York where the concentration of subscribers is highest. Understanding these clusters can inform strategic decisions on marketing efforts and resource allocation.
6. **Pre-COVID and COVID Variations:** The dashboard facilitates a comparison between preCOVID and COVID-era subscription rates and ride patterns. This analysis aids in understanding how the pandemic have impacted user behavior and allows for datadriven adaptations.

By equipping data and business analysts with these insights, the dashboard empowers them to make informed decisions, develop targeted strategies, and enhance overall business performance. It serves as a valuable tool for driving improvements in Citi Bike's operations and customer experience.

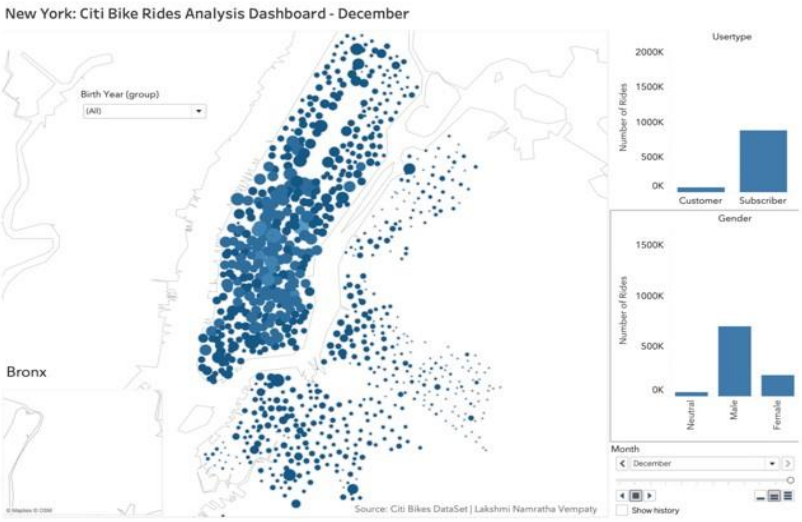

Fig 6: Tableau Visualization for Ride Analysis Tableau Public

URL: https://public.tableau.com/app/profile/namratha.vempaty/viz/CitiBikes_Data_Visualization/Dashboard1?publish=yes

CONCLUSION: UNLOCKING INSIGHTS FOR CITI BIKE'S URBAN MOBILITY

This analysis has embarked on a journey to dissect and illuminate the complex urban mobility landscape in New York City through the lens of Citi Bike's extensive dataset. By integrating analytics, predictive models, and an interactive dashboard, we have unearthed valuable insights that can drive strategic decisions and improvements for Citi Bike's operations and services.

Analytics Insights: Our analytical exploration has revealed several noteworthy trends and patterns:

1. **Ridership Dynamics:** We observed that Citi Bike's ridership faced a substantial dip during the COVID-19 pandemic in 2020. However, the post-pandemic recovery, evident in 2021, not only brought ridership back but surpassed pre-pandemic levels. This highlights the resilience and adaptability of urban mobility systems in the face of significant disruptions.

2. **User Demographics:** An analysis of user demographics showcased shifting trends. While the percentage of customers decreased during the pandemic, it rebounded significantly in 2021. Simultaneously, subscribership increased and nearly doubled in 2021 compared to 2020. Understanding these shifts can drive targeted marketing strategies.
3. **Station Dynamics:** Examining the busiest and underutilized stations unveiled crucial insights. The transition from Yankee Ferry Terminal to Wythe Ave and Metropolitan Ave as the busiest stations suggests a shift from work-related commuting to recreational trips.

Predictive Models: To further deepen our understanding, we deployed predictive models, including Linear Regression, Random Forest, and XGBoost. These models allowed us to:

1. **Forecast Ridership:** We used predictive models to forecast ridership patterns based on various factors, enabling us to make data-informed decisions about station placement and resource allocation.
2. **Feature Selection:** Feature importance techniques aided in identifying critical factors influencing ridership, such as weather conditions, station proximity, and day of the week.

Dashboard Insights: Our interactive dashboard serves as a powerful tool for data and business analysts, providing the means to:

1. **Spot Hot Spots:** Analysts can identify high-demand areas, aiding in targeted marketing and resource allocation.
2. **Profile Target Audiences:** Demographic insights guide tailored marketing campaigns, enhancing user engagement.
3. **Optimize Station Placement:** By examining station availability and demand, we help optimize station placement for maximum user convenience.
4. **Subscriber Rate Tracking:** Monthly subscriber rate trends inform subscription strategies and promotional efforts.
5. **Pre- and Post-COVID Analysis:** Comparing pre-COVID and COVID-era data reveals how the pandemic has reshaped user behavior.

Overall Impact: This analysis not only uncovers the intricate facets of urban mobility but also equips Citi Bike with actionable insights. These insights will shape marketing strategies, station expansion plans, and subscriber engagement initiatives. In a post-pandemic world where urban mobility is evolving, data-driven decisions are the compass guiding Citi Bike towards a future of enhanced service and customer satisfaction.

REFERENCES

[1] Wang, W. (2020). Analysis and Prediction of Citi Bike Usage in the Unpredictable 2020. Towards Data Science. https://towardsdatascience.com/analysis-and-prediction-of-citibike-usage-in-the-unpredictable-2020-3401da26881b

[2] https://s3.amazonaws.com/tripdata/index.html

Citation: Lakshmi Namratha Vempaty, Citi Bike Pre and Post Pandemic Analysis Understanding the Impact of Covid-19 On Urban Mobility. *International Journal of Advanced Research in Management (IJARM).* 12(1), 2021. pp. 99-109

Abstract Link: https://iaeme.com/Home/article_id/IJARM_12_01_008

Article Link:
https://iaeme.com/MasterAdmin/Journal_uploads/IJARM/VOLUME_12_ISSUE_1/IJARM_12_01_008.pdf

editor@iaeme.com