# Citi Bike Rental Problem

causaLens

Shubhani Jain, Ph.D. in Physics

# Project Overview:

In this problem, we are tasked with analysing data from Citi Bike rental company to provide insights and recommendations that will optimise their operations. The company aims to better understand how bikes should be distributed among various stations and determine the optimal stock levels required to meet demand efficiently.

The primary goal of this analysis is to:

1. Perform Exploratory Data Analysis (EDA): Understand the available data, identify patterns, and derive insights related to bike usage.

2. Predictive Modelling: Develop machine learning models to predict bike demand at different stations and times.

3. Optimisation: Propose strategies to optimise the distribution of bikes across stations and manage stock levels, ensuring that supply meets demand efficiently.

4. Recommendations: Provide actionable insights and recommendations that the company can implement to improve their operational efficiency and enhance user satisfaction.

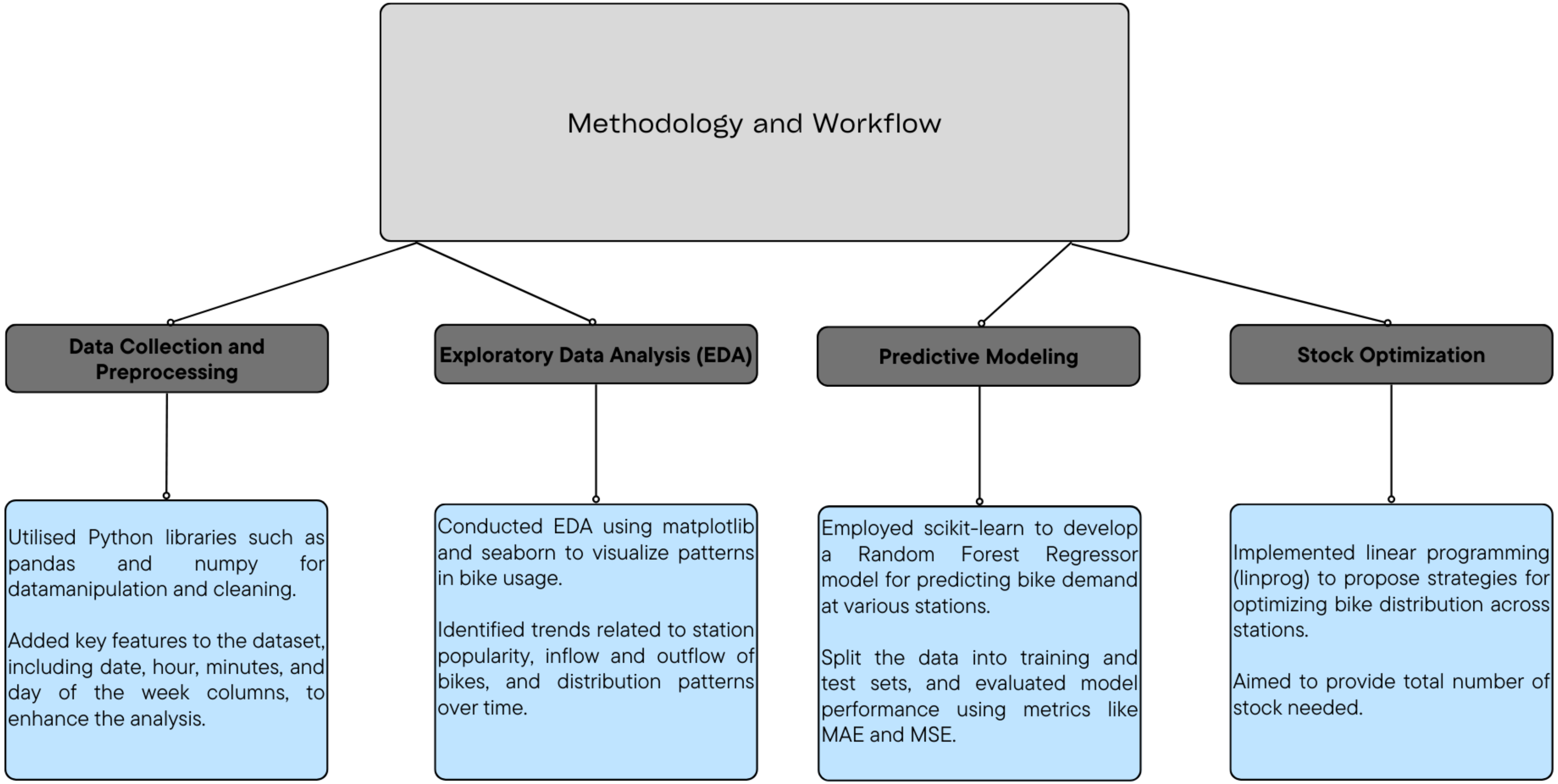
# Company Overview:

Citi Bike is a leading bike-sharing program that offers a flexible and reliable rental service in New York City, providing a convenient and eco-friendly transportation option for residents and visitors. Citi Bank first published it’s data for public in 2013, facilitating analysis, development, visualisation and other applications. The data for this project can be accessed through the following URL:  Citi Bike NYC System Data and Data Repository.

The data repository contains detailed trip data, including information on bike usage, station locations, and temporal patterns. For our analysis, we will use the data from May 2024, as it provides the most recent information on crucial trends and patterns.

citi bike®
provided by lyft

# Methodology and Workflow:



Methodology and Workflow

**Data Collection and Preprocessing**

Utilised Python libraries such as pandas and numpy for datamanipulation and cleaning.

Added key features to the dataset, including date, hour, minutes, and day of the week columns, to enhance the analysis.

**Exploratory Data Analysis (EDA)**

Conducted EDA using matplotlib and seaborn to visualize patterns in bike usage.

Identified trends related to station popularity, inflow and outflow of bikes, and distribution patterns over time.

**Predictive Modeling**

Employed scikit-learn to develop a Random Forest Regressor model for predicting bike demand at various stations.

Split the data into training and test sets, and evaluated model performance using metrics like MAE and MSE.

**Stock Optimization**

Implemented linear programming (linprog) to propose strategies for optimizing bike distribution across stations.

Aimed to provide total number of stock needed.

# Exploratory Data Analysis

causaLens

# Task:

Preference for Rideable Type

Member v/s Casual Riders

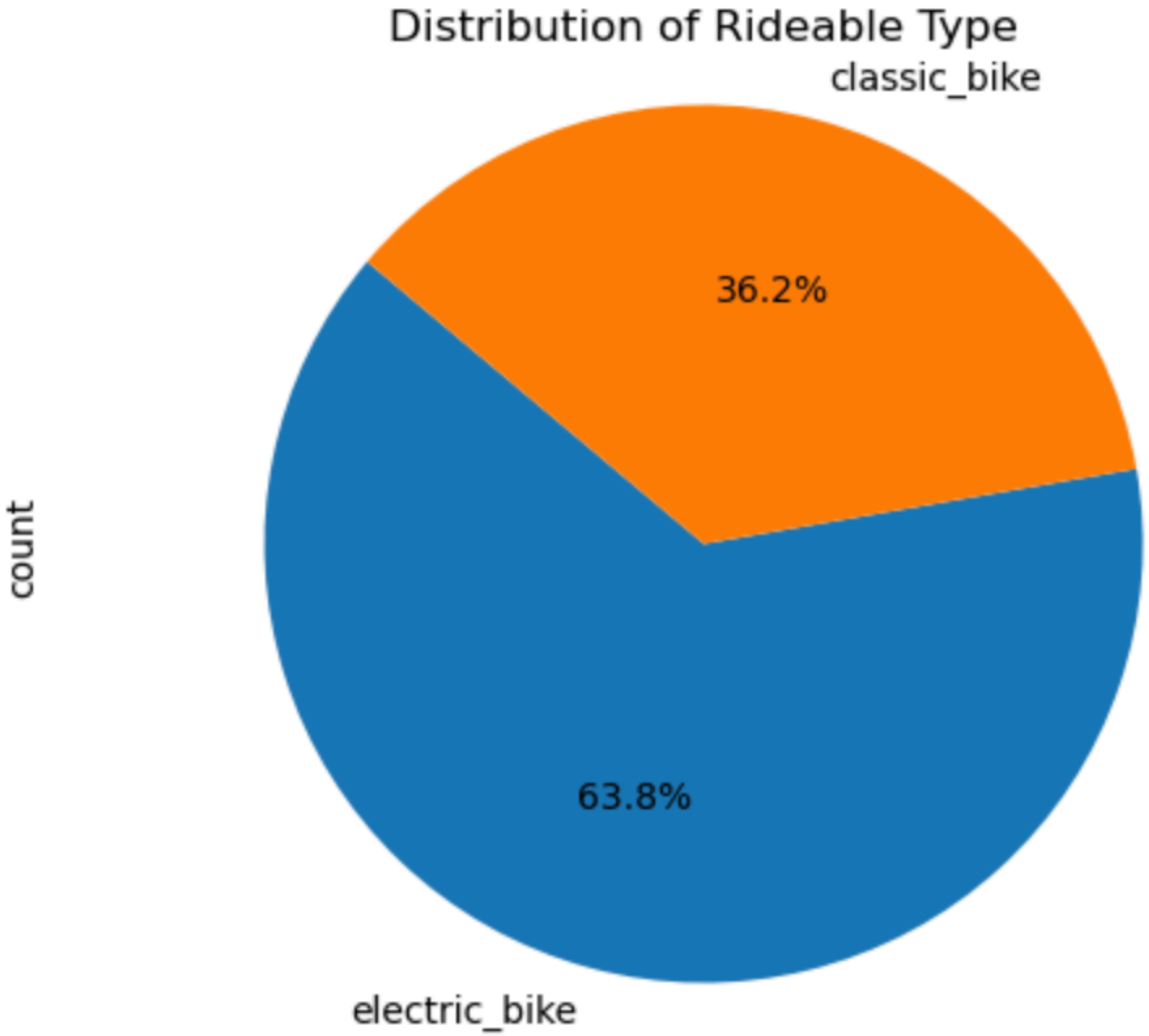
Trip Duration Analysis

Rides usage by day of week, per hour and per day

Station and Popular Routes Analysis

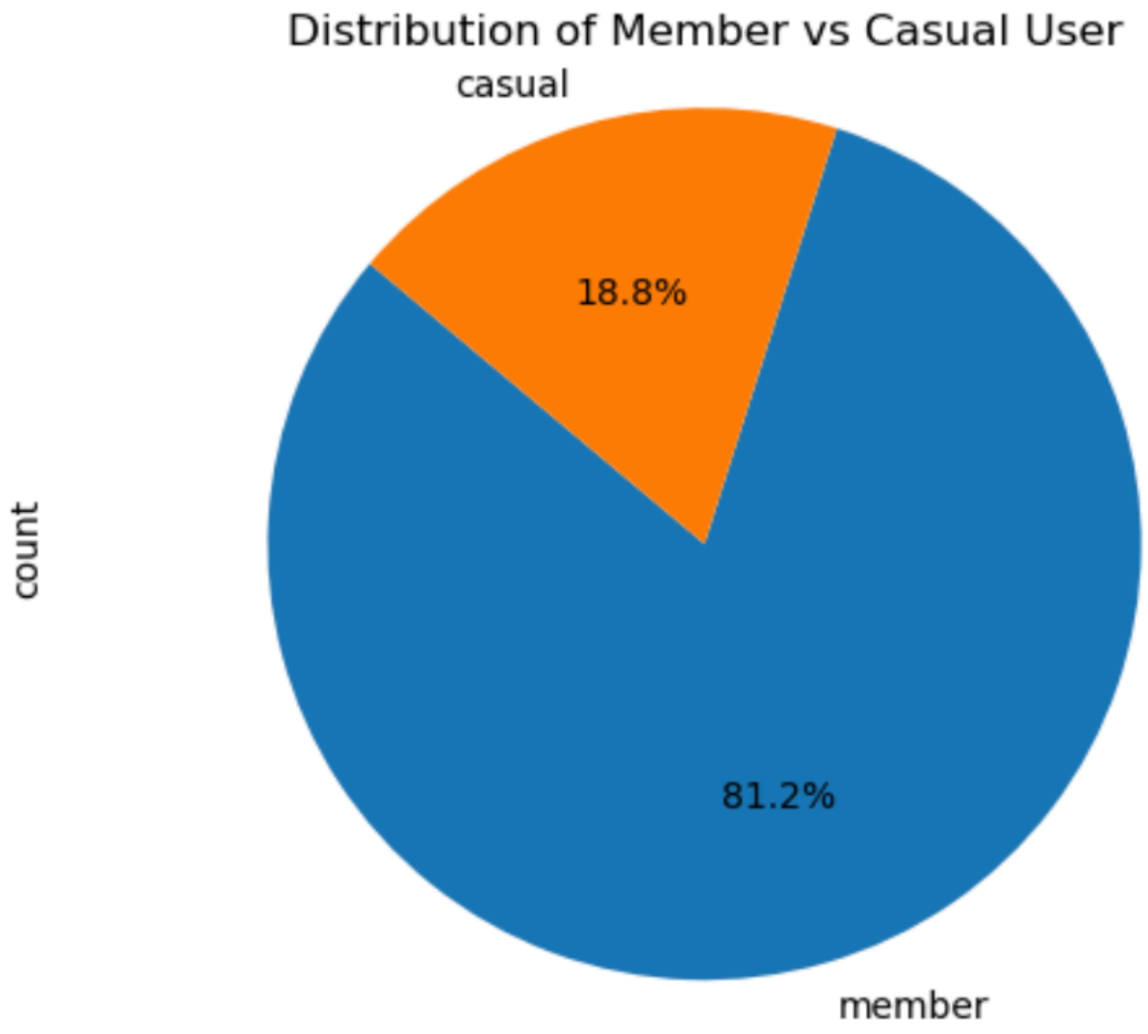
# Preference for Rideable type:

- This pie chart illustrate the distribution of different types of bikes used in the rental service.
- Electric bikes (63.8%) are significantly more popular among users than classic bikes (36.2%), accounting for nearly two-thirds of all rentals.
- The high demand for electric bikes indicate that they are preferred for their ease and efficiency, suggesting a need for increased investment and availability at high-demand stations.



# Member vs Casual Riders:

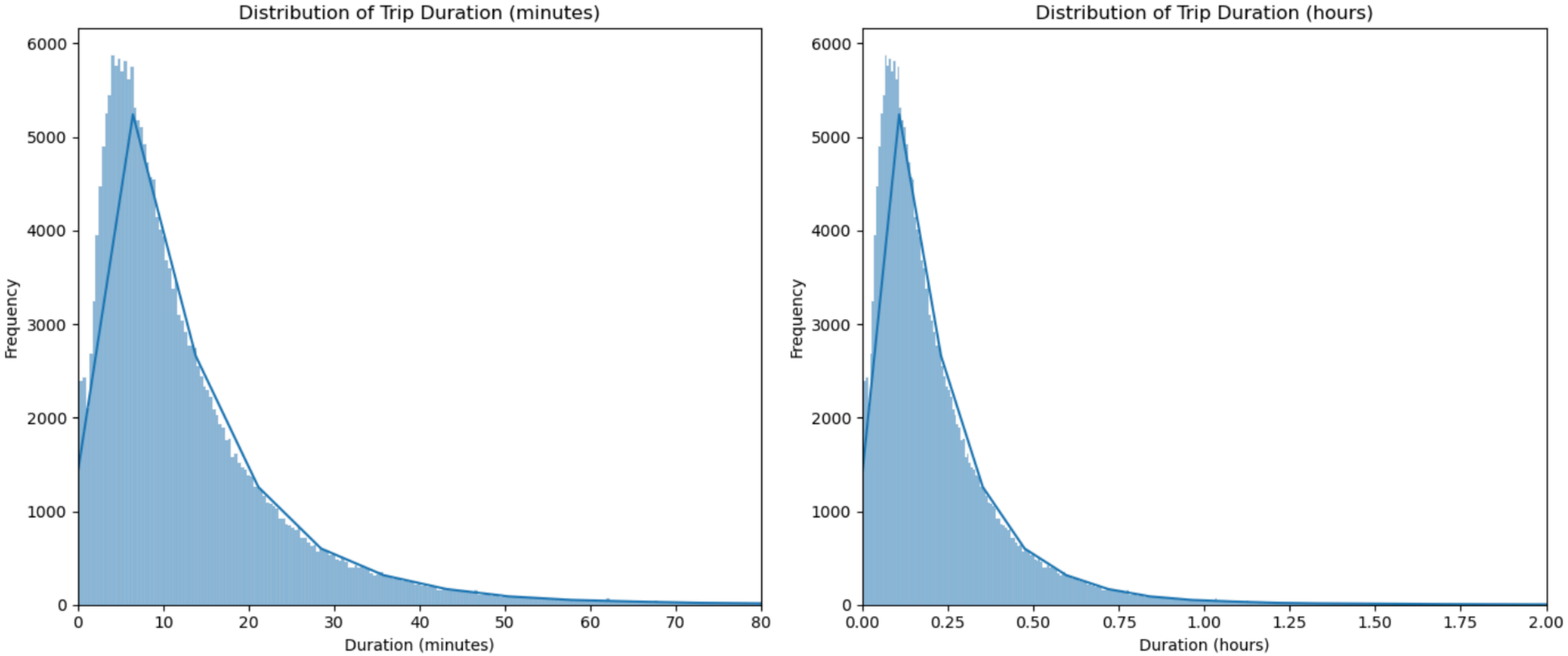
- This pie chart illustrates the distribution of members versus casual users.
- The majority of rentals (81.2%) come from members as compared to casual rentals (18.8%), indicating that the membership model is highly effective and should be relied on.
- The high percentage of member rentals suggest that increasing focus on member retention and attraction, and converting casual users through promotions, could boost rental frequency and revenue.



causaLens

# Trip Duration Analysis:

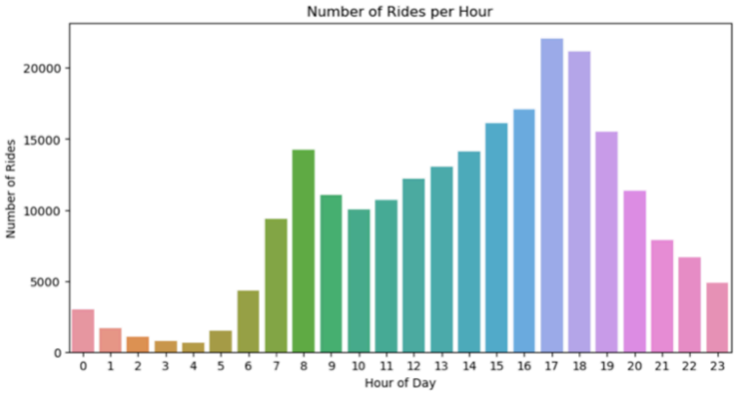
- These two plots present the distribution of trip durations, one in minutes and the other in hours.
- The distribution is right-skewed with the majority of trips lasting only a few minutes, peaking around 7 minutes, and extending into a long tail of durations over 80 minutes, indicating a significant portion of longer trips.
- The plots reveal a wide range of trip durations, with a dominance of short trips evident from peaks around 0.125 hour, while the long tails indicate a smaller but significant number of notably longer trips.

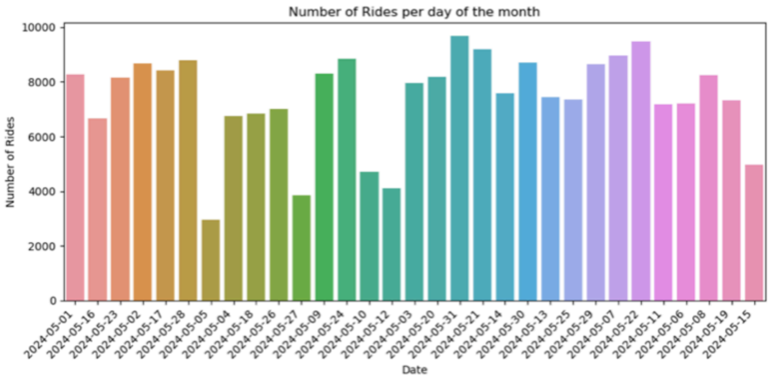


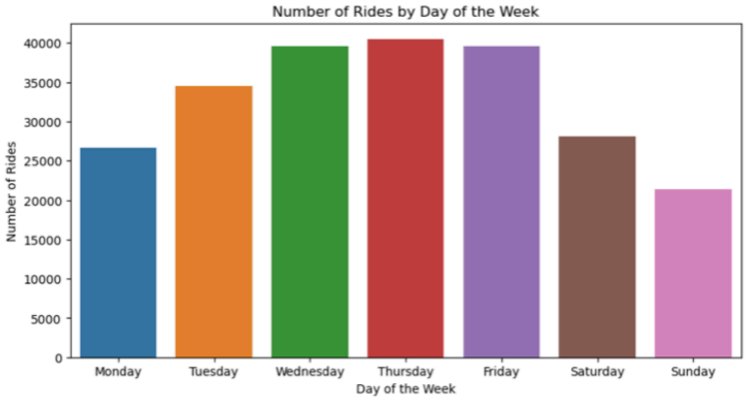
causaLens

# Rides usage by day of week, per hour and per day:

- The plot with number of rides per hour reveals bike ride peaks at 8 AM and 5-6 PM, indicating morning and evening commutes. This pattern can guide bike-sharing optimisation, station placement, and system planning.
- The plot with per day rides show significant fluctuations in bike rides throughout May 2024, with peak usage around mid-month and lower usage towards the end, indicating variability that can guide resource management and operational strategies.
- The plot shows that while ride patterns are consistent across weekdays and weekends, weekend peaks are less pronounced, suggesting opportunities for optimising bike-sharing systems, infrastructure, and targeted marketing.

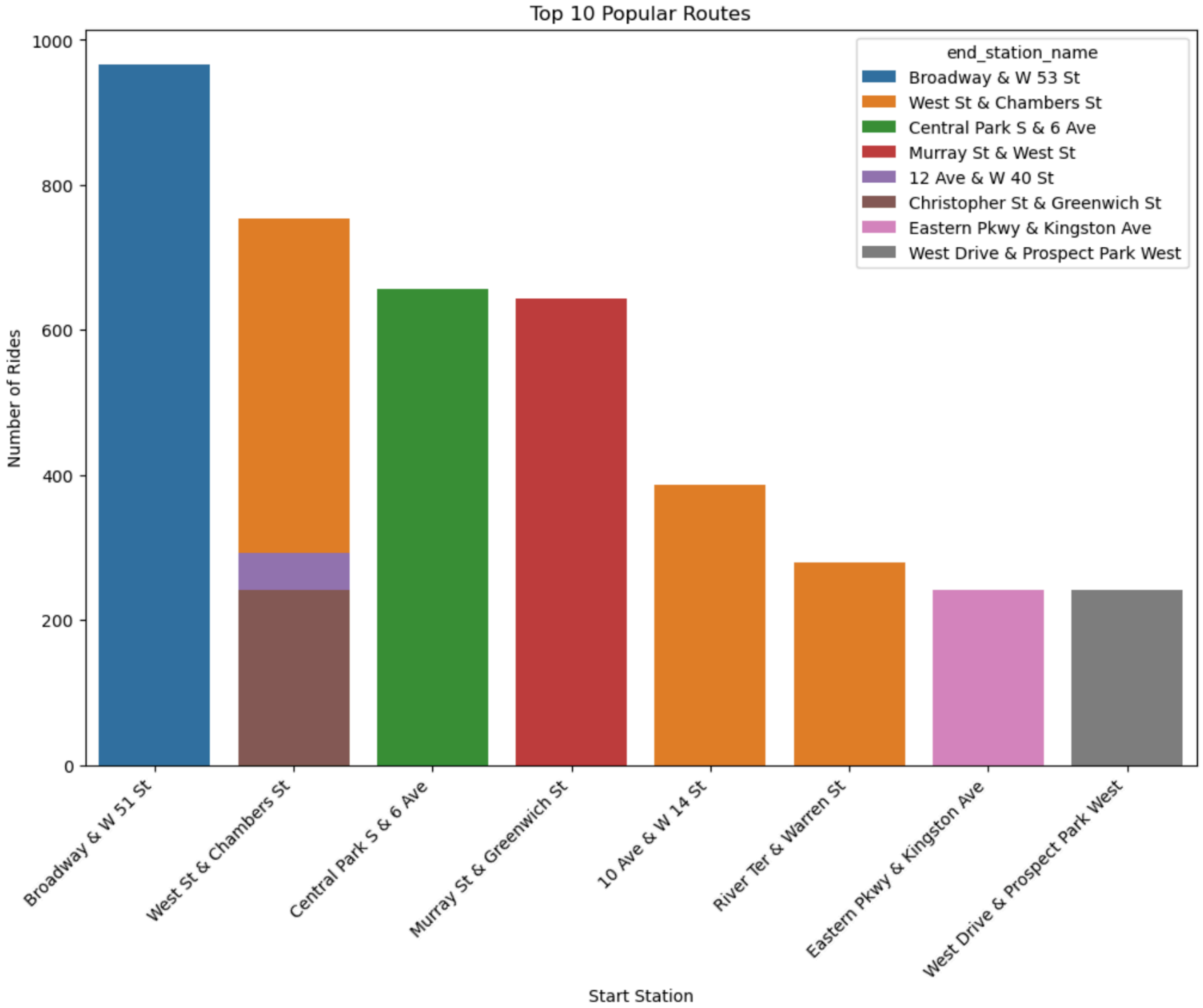






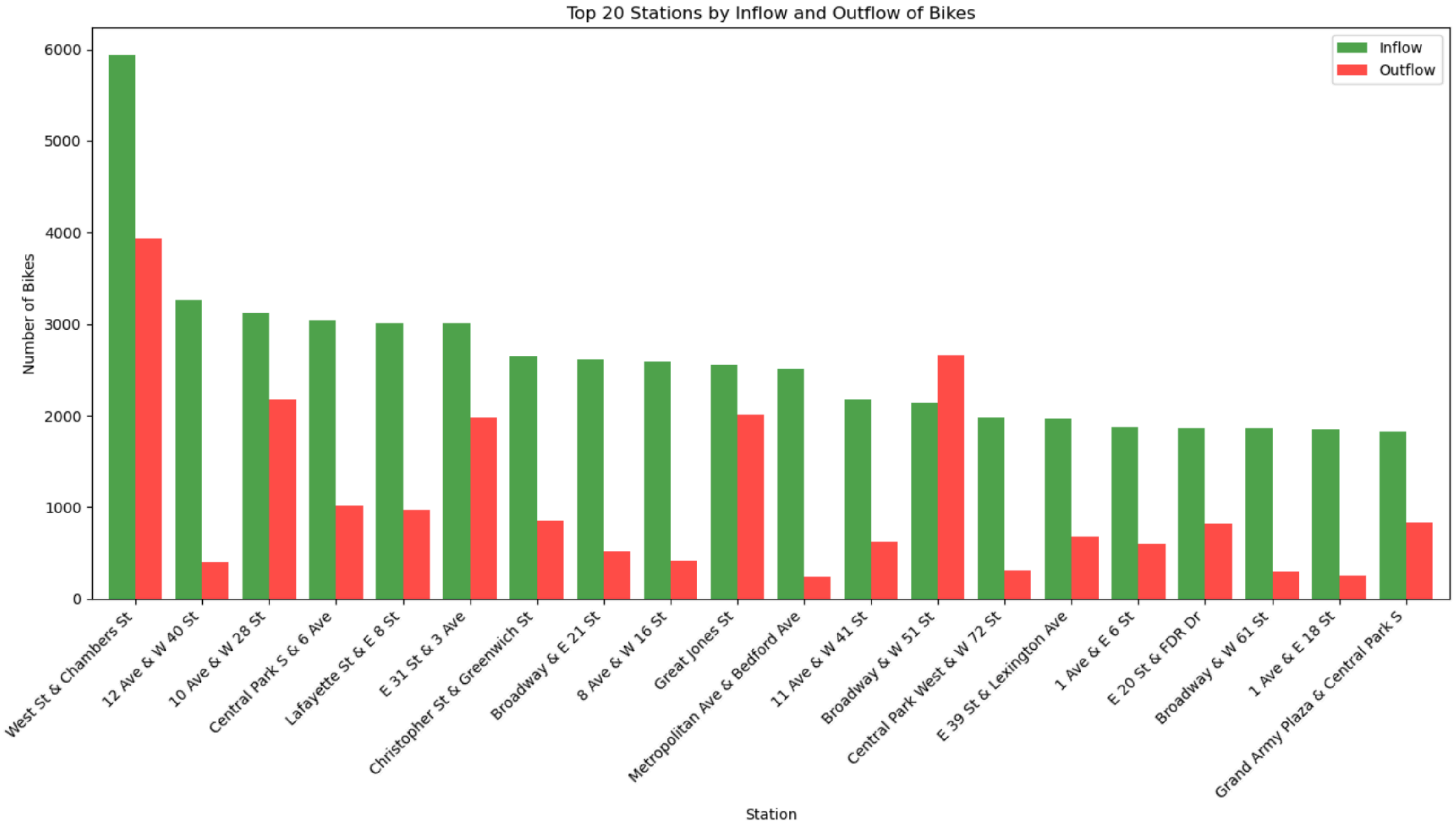
# Station and Popular Routes Analysis:

- The plot visualises the top 10 most popular routes based on the number of rides, considering both start and end stations.
- The route between “Broadway & W 21 St” and “Broadway & W 21 St” is the most popular, as evidenced by the tallest bar in the chart.
- The plot showcases a variety of popular routes, understanding the most popular routes can help optimise bike-sharing systems, such as station locations and bike distribution.



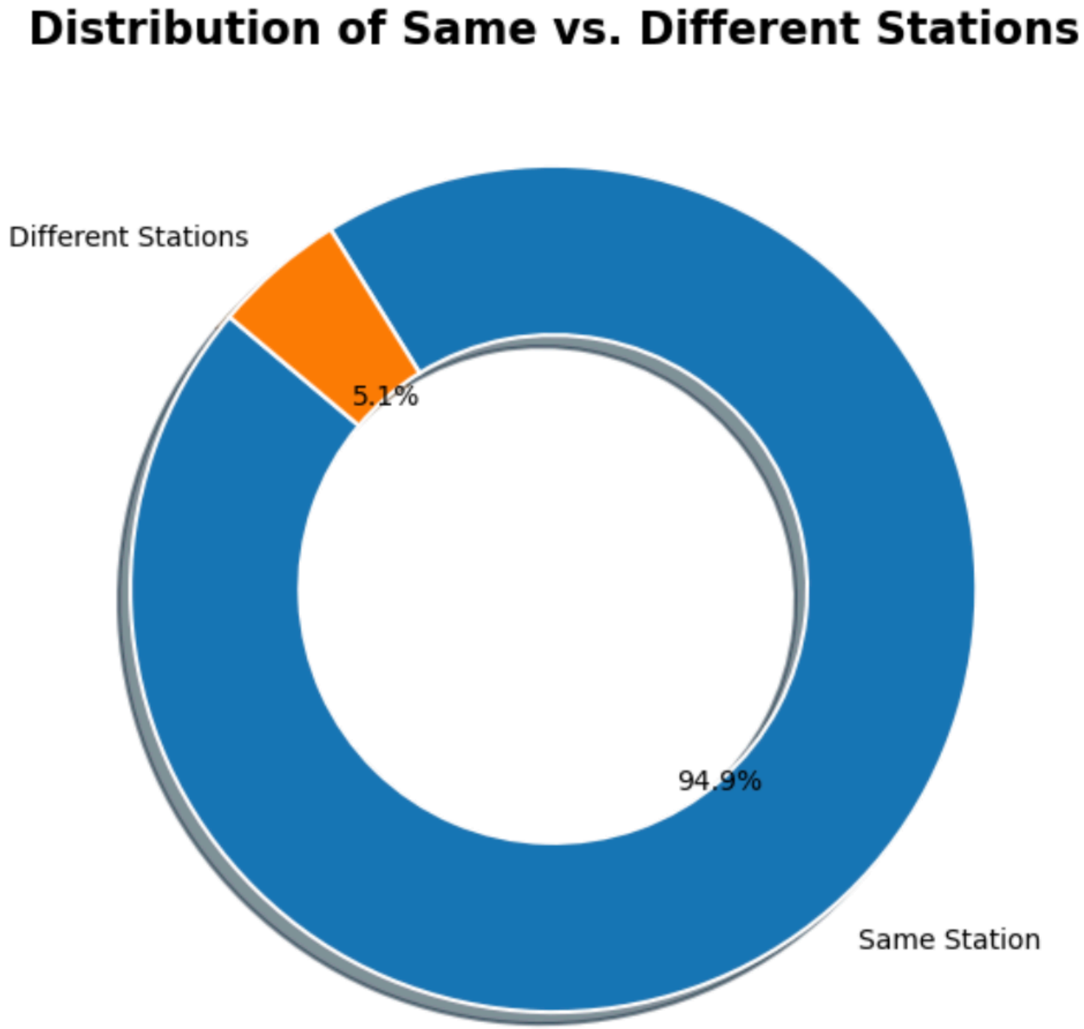
# Station and Popular Routes Analysis:

- The plot visualises the top 20 stations by inflow (green) and outflow (red) of bikes.
- The plot reveals a correlation between bike inflow and outflow at stations, with "West St & Chambers St" being notably the busiest, having the highest traffic for both.
- There are variations in the inflow-to-outflow ratio among stations, suggesting some are more commonly starting points and others are ending points for rides.



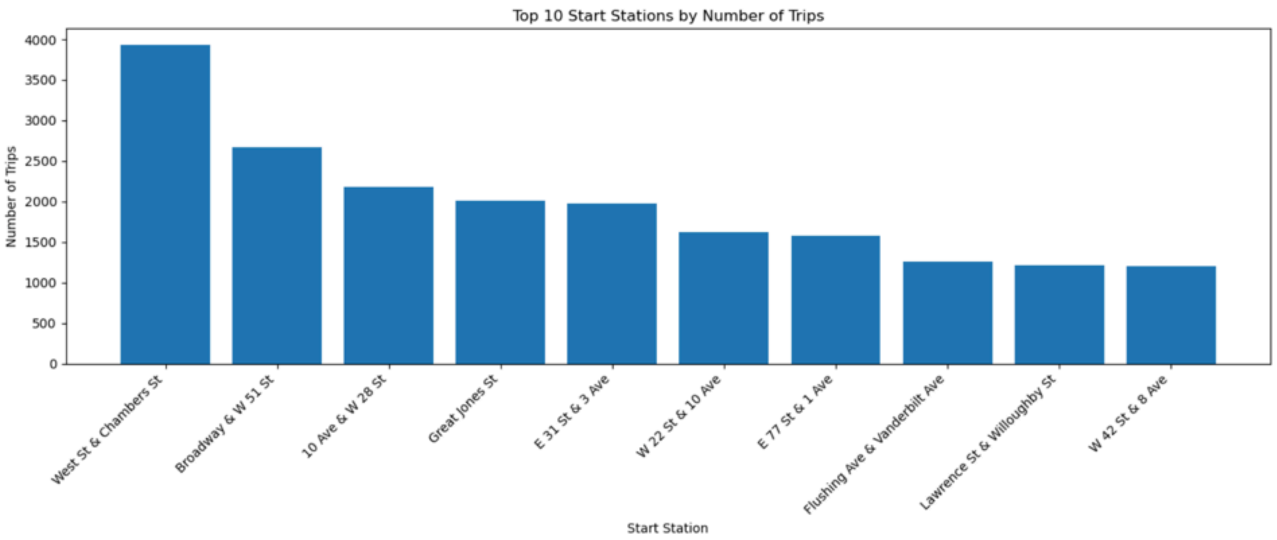
# Station and Popular Routes Analysis:

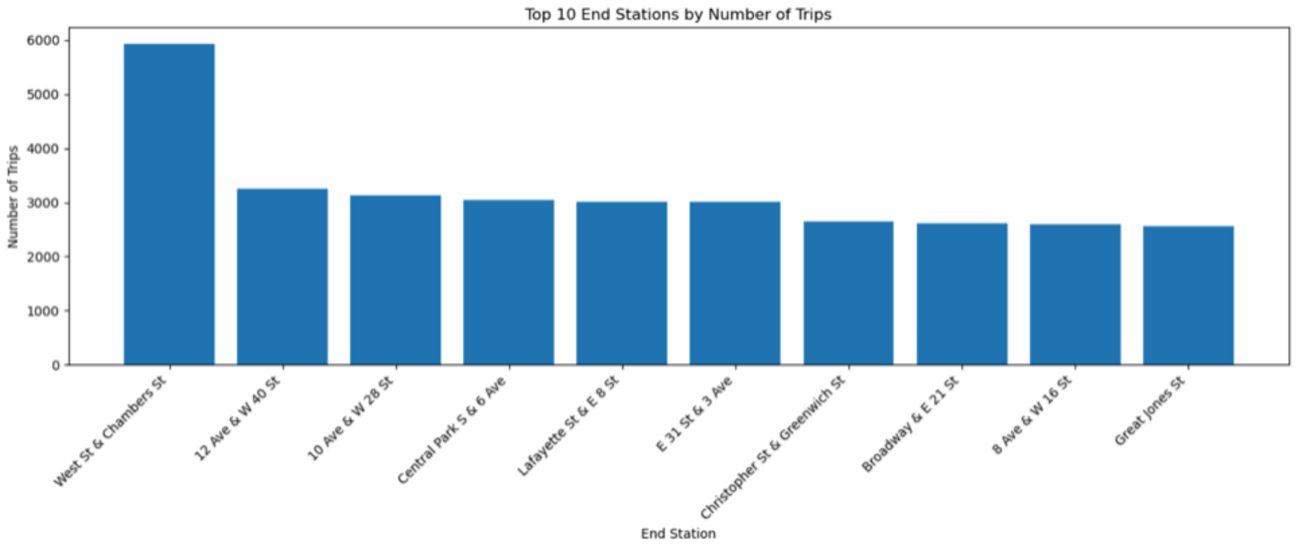
- With 94.9% of trips starting and ending at the same station, users demonstrate a strong preference for using familiar stations, likely due to convenience or proximity.
- Only 5.1% of trips involve different origin and different destination stations, suggesting limited inter-station travel.
- The dominance of same-station trips suggest opportunities to enhance services or infrastructure at key stations to better support the high volume of local travel.



# Station and Popular Routes Analysis:

- Both plots show the number of trips, highlighting the popularity of start and end stations.
- "West St & Chambers St" stand out as the most popular start and end station, with significantly higher trip counts compared to others.
- The stations are ranked by trip volume, revealing variation in usage levels among the top 10 stations for both starts and ends.
- Analysing popular start and end stations helps optimise bike-sharing systems by informing decisions on station placement, bike distribution, and understanding user behaviour.





# ML - Driven Solution

causaLens

# Demand Prediction Modelling:

Random Forest model was used to predict demand based on various features such as hours of the day, day of the week, and start station name. Linear Regression, sophisticated RNN and time series forecasting algorithms such as ARIMA and Facebook Prophet can also be used as an alternative.

**Workflow:**

**Data Preparation:**

- Aggregated historical data based on start_station_name, hour_of_day, and day_of_week and calculate demand.
- Applied one-hot encoding for categorical features and split data into training and test sets (80% train, 20% test).

**Model Training:**

- Utilised Random Forest Regressor and trained the model with progress tracking.
- Parameters used: n_estimators = 100, random_state = 42.

**Evaluation**:

- Assessed model performance using Mean Absolute Error (MAE) and Mean Squared Error (MSE) for accuracy.
- Plotted actual vs predicted demand.

# Model Evaluation Results:

**Mean Absolute Error (MAE):** 0.9952792146099442

- Indicates that, on average, the model's predictions are off by approximately 1 bike, reflecting a fairly accurate model.

**Mean Squared Error (MSE):** 4.18896689538689

- Shows that the average squared difference between predicted and actual values is 4.189, highlighting the presence of some larger errors.

**Some Insights:**

- The low MAE suggests that predictions are generally close to actual demand values, indicating good model accuracy.

- The MSE suggests that while most predictions are accurate, there are some larger discrepancies that may need further attention to improve overall model performance.

# Optimisation of Bike Distribution:

Once the demand predictions were ready, we optimised the distribution of bikes across stations. We used linear programming for this purpose.

**Workflow:**

**Data Preparation:**

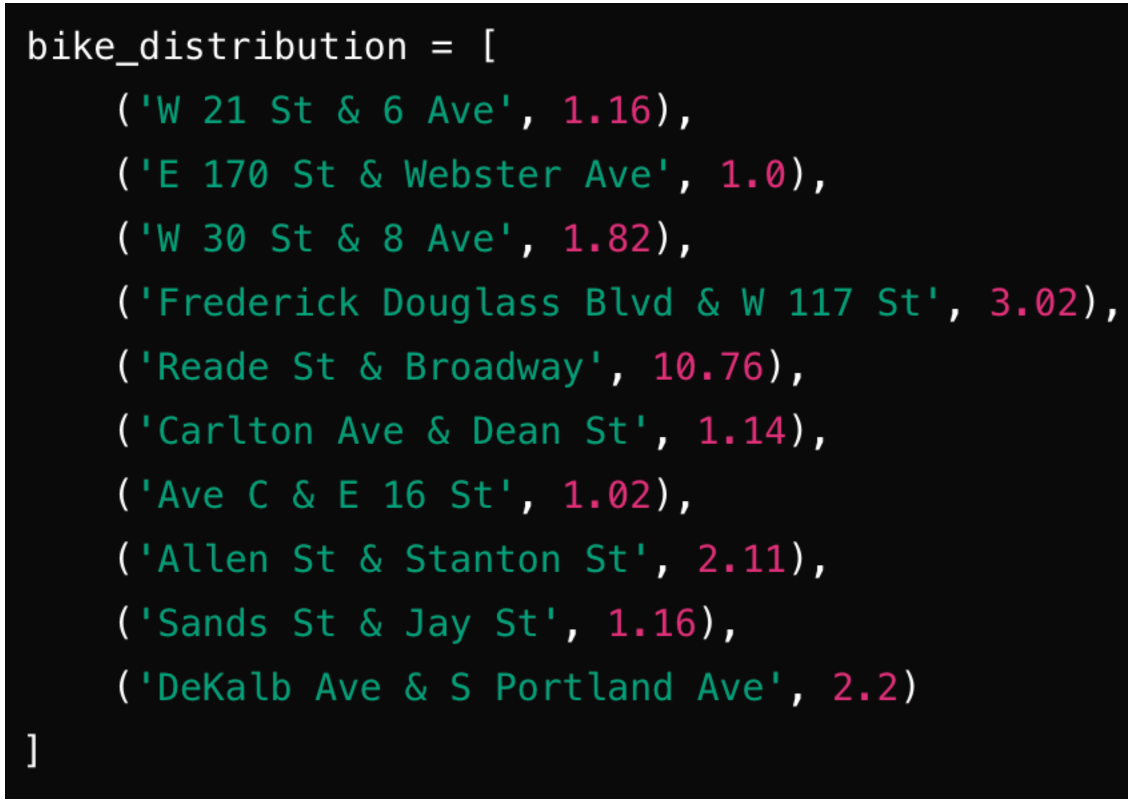
- Used predicted demand values (`y_pred`), and converted such values to a dictionary mapping station names to their respective demand.

**Optimisation Process:**

- Minimised the total number of bikes required.
- Ensured bike allocation met the predicted demand at each station.

**Results:**

- Computed the number of bikes to allocate to each station to meet forecasted demand.
- Provided a dictionary with station names and the corresponding optimal number of bikes.



```
bike_distribution = [
    ('W 21 St & 6 Ave', 1.16),
    ('E 170 St & Webster Ave', 1.0),
    ('W 30 St & 8 Ave', 1.82),
    ('Frederick Douglass Blvd & W 117 St', 3.02),
    ('Reade St & Broadway', 10.76),
    ('Carlton Ave & Dean St', 1.14),
    ('Ave C & E 16 St', 1.02),
    ('Allen St & Stanton St', 2.11),
    ('Sands St & Jay St', 1.16),
    ('DeKalb Ave & S Portland Ave', 2.2)
]
```

# Total Stock Required:

Finally, we provide the total number of bikes required to satisfy demand for Citi Bike Company:

Total stock required:

**3024**

This approach provides a robust framework for ensuring that bikes are distributed optimally across stations, meeting user demand while minimising costs. To further optimise the algorithms, fine tuning of Random Forest hyper parameters, feature engineering and use of more sophisticated Deep Learning model like RNN is suggested. Moreover, time series forecasting algorithms such as ARIMA or Prophet is advised for reduced computational cost.

# Recommendations:

- Optimise bike availability during peak times such as morning and evening commutes, to meet high demand and improve service reliability.
- Optimise bike distribution by allocating bikes based on predicted demand at each station to avoid shortages and ensure high availability where needed.
- Increase bike stock at high-demand stations to improve user satisfaction and reduce wait times.
- Regularly review demand predictions and adjust bike stock levels to account for fluctuations in usage patterns.
- Continuously refine predictive models and add new features to adapt to changing demand patterns.
- Consider implementing dynamic pricing strategies to manage demand and encourage bike usage during off-peak hours, while balancing overall bike availability.

# Recommendations:

- Upgrade infrastructure at key stations with high bike inflow and outflow to handle large volumes and increase operational efficiency.

- Offer incentives and improve service quality to retain and attract members, as they rent more frequently.

- Improve fleet management by regularly assessing and adjusting bike fleet size.

- Balance the cost of maintaining optimal stock levels with the benefits of reduced bike shortages and better service.

- Utilise current demand and stock analysis to plan for future expansions or adjustments, ensuring scalability and consistent service quality.

- Use advanced forecasting models like ARIMA or RNN to improve demand predictions and stock management.

# Thank you

causaLens

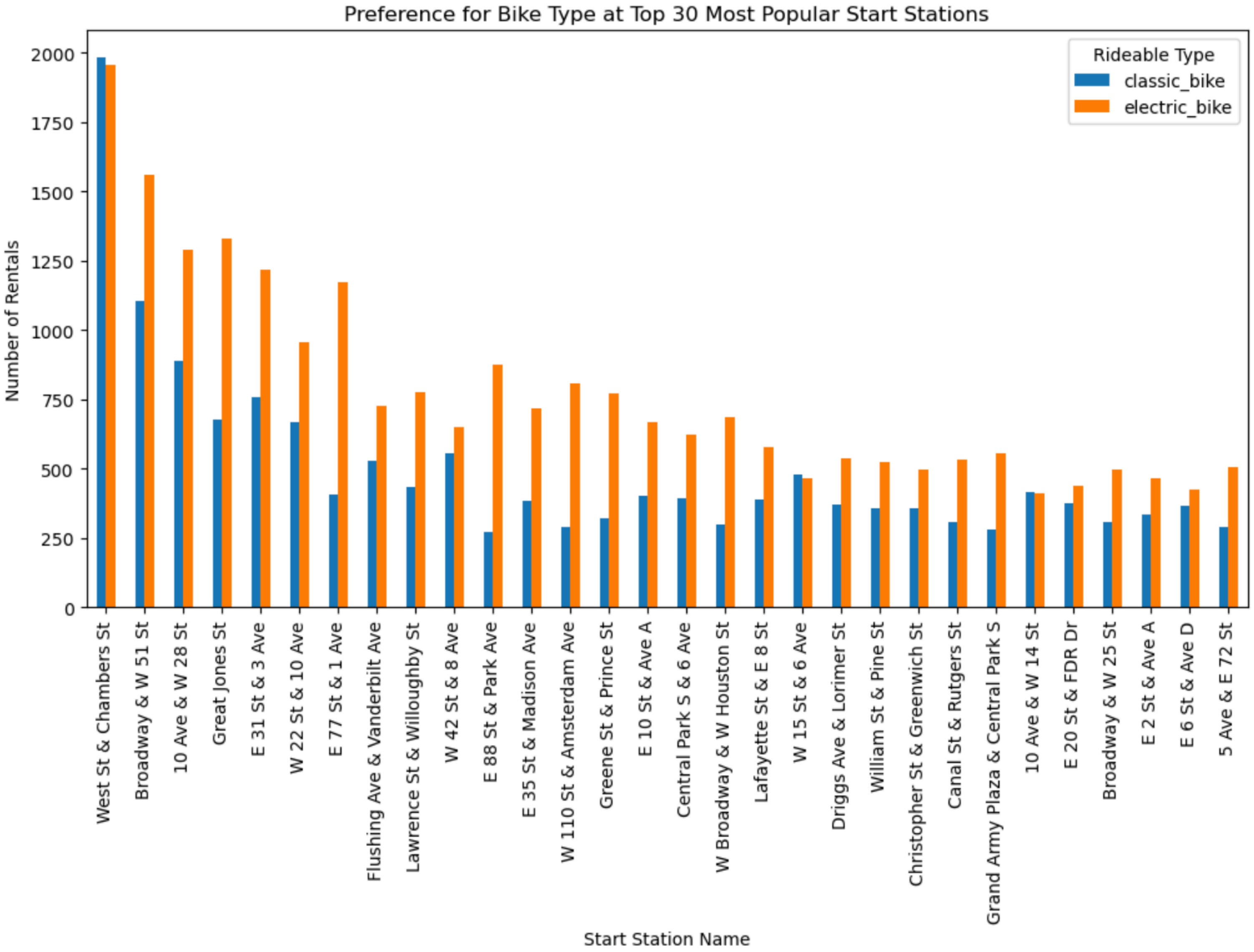
**info@causaLens.com**
www.causaLens.com

# Additional EDA Plots

causaLens

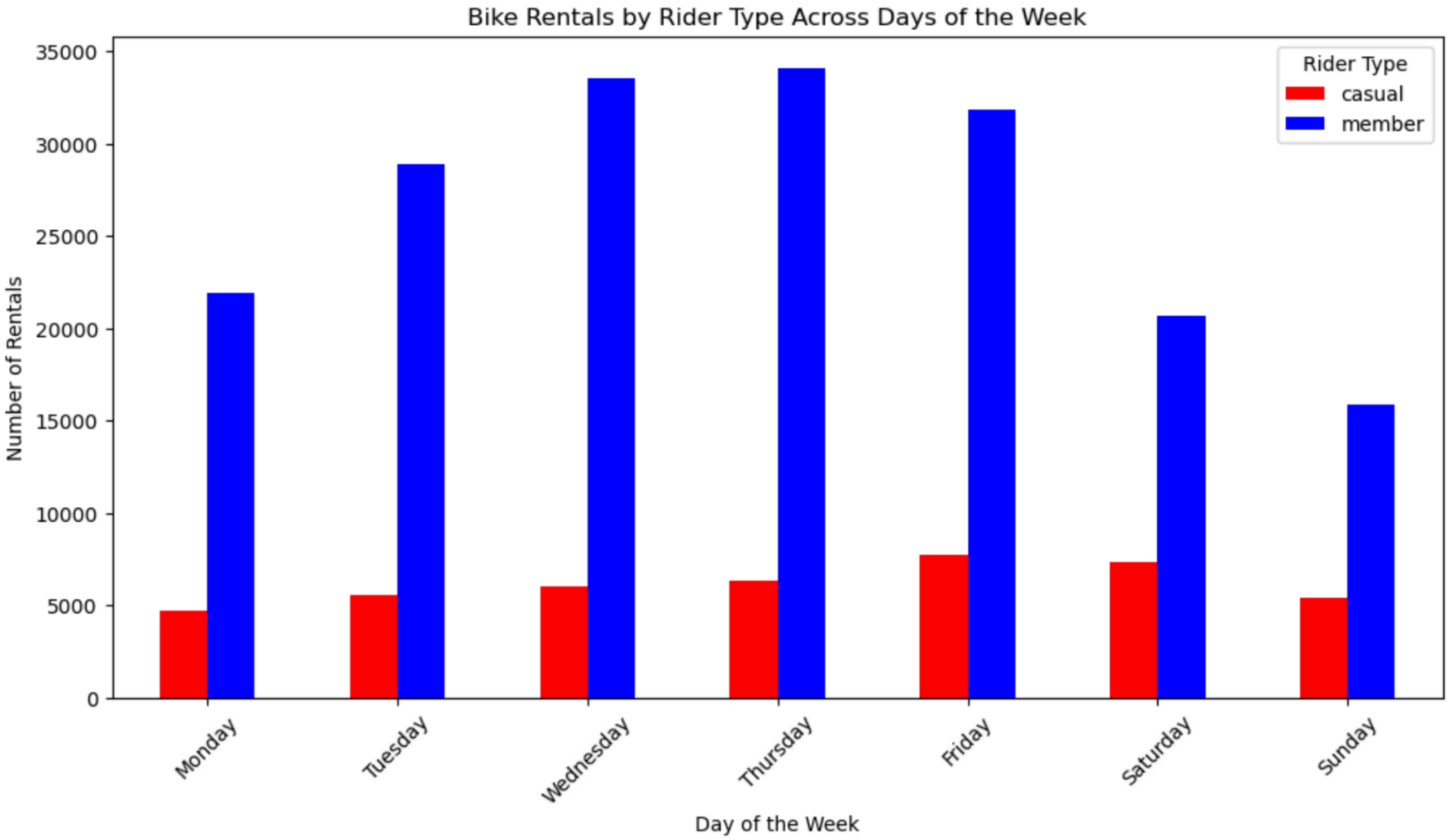
# Preference for Rideable type:

- This bar plot illustrates the distribution of different types of bikes used in the rental service across top 30 most popular start stations.
- Electric bikes are generally more popular than classic bikes across the top 30 start stations. For example: "West St & Chambers St" and “Broadway & W 51 St” are the busiest stations, with electric bikes leading in popularity.
- The preference for electric bikes differs between some stations, indicating potential factors like location or user demographics.



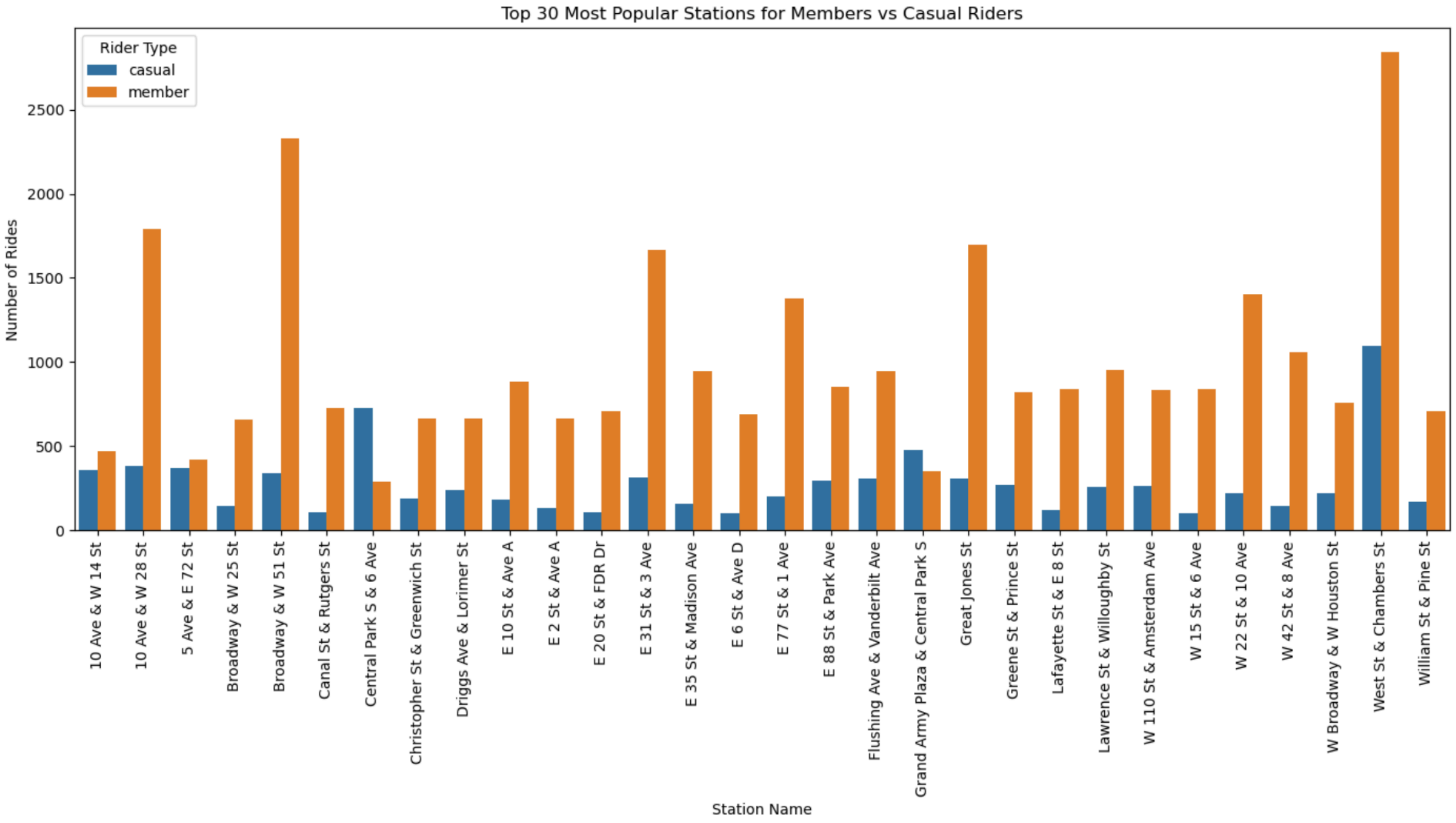
# Member vs Casual Riders:

- This bar plot illustrates the distribution of bike rentals by rider type across days of the week.
- Members tend to use bikes consistently throughout the week, suggesting bike sharing is often used for commuting or daily errands.
- Casual riders primarily use bikes on weekends, indicating leisure or recreational use.
- Overall bike rentals peak on weekdays due to member usage and on weekends due to casual riders.



# Member vs Casual Riders:

- This bar plot illustrate the distribution of top 30 most popular start stations for members vs casual riders, highlighting the demand patterns at different locations.
- There is a huge comparison of ride frequency between members and casual riders at each station, revealing preference trends and station-specific rider demographics. For example: ‘West St & Chamber St’ is more popular among member whereas ‘Central Park S & 6 Ave’ is more popular among casual member.
- Identifying stations with high demand for one rider type over the other helps in understanding usage patterns and planning bike distribution.


Top 30 Most Popular Stations for Members vs Casual Riders
Rider Type
casual
member
Number of Rides
10 Ave & W 14 St
10 Ave & W 28 St
5 Ave & E 72 St
Broadway & W 25 St
Broadway & W 51 St
Canal St & Rutgers St
Central Park S & 6 Ave
Christopher St & Greenwich St
Driggs Ave & Lorimer St
E 10 St & Ave A
E 2 St & Ave A
E 20 St & FDR Dr
E 31 St & 3 Ave
E 35 St & Madison Ave
E 6 St & Ave D
E 77 St & 1 Ave
E 88 St & Park Ave
Flushing Ave & Vanderbilt Ave
Grand Army Plaza & Central Park S
Great Jones St
Greene St & Prince St
Lafayette St & E 8 St
Lawrence St & Willoughby St
W 110 St & Amsterdam Ave
W 15 St & 6 Ave
W 22 St & 10 Ave
W 42 St & 8 Ave
W Broadway & W Houston St
West St & Chambers St
William St & Pine St
Station Name