
NYC Taxis and Bike data
NYC

OPIM-5512-Data Science using Python

Under the guidance and teaching of Professor Ramesh Shankar

So we are here to present our final project for Course

Team member
1. Manmeet Kaur
2. Pavan Bolla
3. Shubham Yedekar

Before diving into the details, let's take a brief look at the history to understand the significance of this dataset and our reasons for selecting it.

History of NYC Bike Share Program

Introduction: NYC's bike share program, known as Citi Bike, launched in May 2013 as a partnership between NYC and a private company to promote eco-friendly transportation.

Growth & Expansion: Initially limited to lower Manhattan and parts of Brooklyn, Citi Bike quickly expanded due to popularity. By 2015, it was one of the largest bike-sharing programs in the U.S.

Current Status: As of today, Citi Bike has over 24,000 bikes and 1,500 stations across Manhattan, Brooklyn, Queens, the Bronx, and Jersey City.

Ridership: Citi Bike continues to set records, with over 100 million rides taken since its launch.

History of NYC Yellow Taxis

Introduction: NYC yellow taxis, an iconic part of the city's identity, began in the early 20th century. The first yellow cabs appeared in 1907, with the standardization of the yellow color occurring in the 1960s.

Peak and Decline: Yellow taxis reached peak numbers in the early 2000s but faced challenges as ride-hailing services like Uber and Lyft entered the market around 2011.

Current Status: There are approximately 13,500 yellow taxis, significantly fewer than in previous years due to competition from ride-hailing services.

Usage & Challenges: Yellow taxis still serve Manhattan heavily, especially for hailing on the street. However, they face challenges as customers increasingly prefer app-based services.

History of NYC Green Taxis

Introduction: NYC's green taxis, officially called Boro Taxis, were introduced in 2013 to serve outer boroughs and areas of upper Manhattan underserved by yellow taxis.

Purpose & Reach: Green taxis were created to legally pick up street hails in neighborhoods outside Manhattan's core, meeting a demand that yellow taxis traditionally avoided.

Current Status:There are roughly 4,000 active green taxis, down from an initial peak due to competition with app-based ride services.

Relevance: Despite reduced numbers, green taxis are still a key transport option in underserved areas of NYC.

Why we selected this Data set ???

Multi-Modal Analysis for Optimal Choice

Peak Hour Demand Insights

Location-Based Recommendations

Environmental Impact Evaluation

So, lets dive into our data set

- We analyzed both the Taxi (Yellow & Green) and Bike datasets to uncover valuable insights and patterns within urban transportation usage.
- Our goal was to identify trends, compare usage dynamics, and assess the role each mode plays in New York City's transportation landscape.
- By combining these datasets, we were able to reveal comprehensive insights into ride preferences, peak usage times, geographic distribution, and the potential impacts of each mode on traffic flow and environmental sustainability.

Origin of our data set

TAXIS:- https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

BIKES:- https://citibikenyc.com/system-data

Details of our data set

Green Taxi Data - 51,837rows , 20 columns

Yellow Taxi Data - 3,076,903 rows , 19 columns

Citi Bike Data - 112,443 rows , 13 columns

DATA CLEANING

- Datetime Conversion
- Hour Extraction for Demand Analysis
- Grouping and Aggregating by Location and Hour
- Removing Duplicates
- Removing Unnecessary Columns

Details of our data set

Green Taxi Data - 51,837rows , 20 columns

Green Taxi Data - 46827 rows , 20 columns

Yellow Taxi Data - 3,076,903 rows , 19 columns

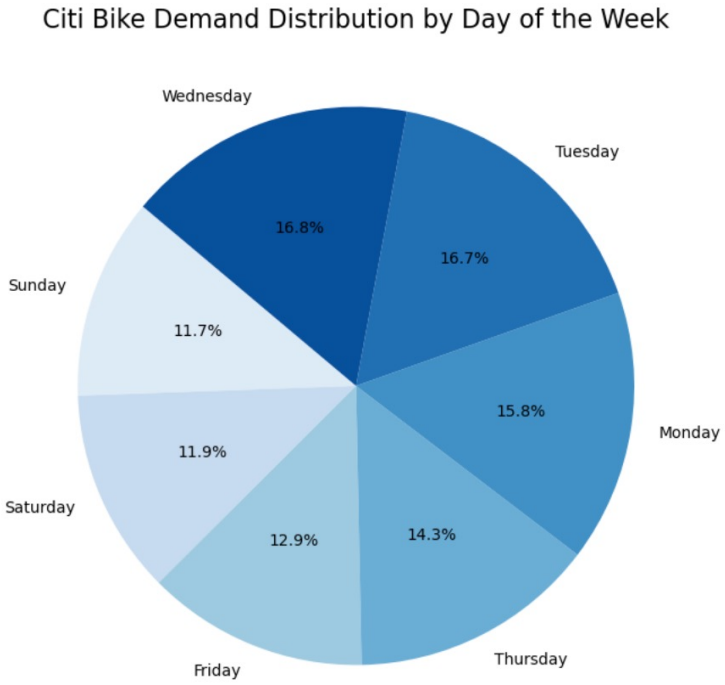
Yellow Taxi Data – 2729770 rows, 19 columns

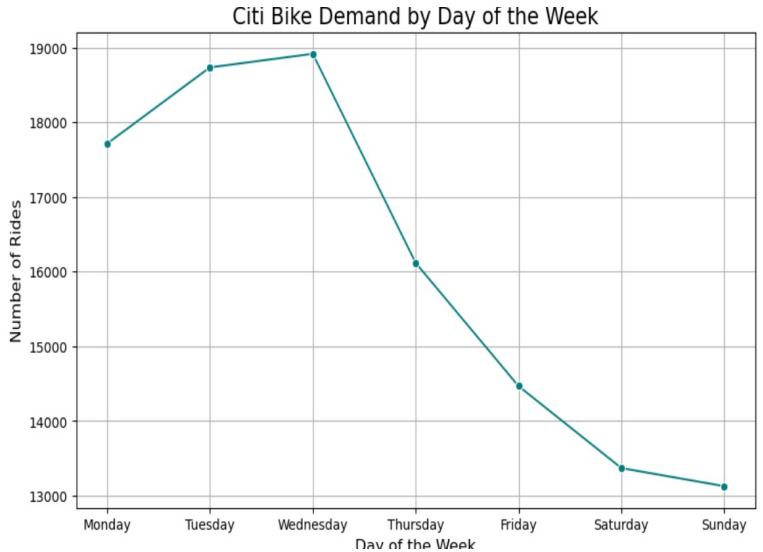
Citi Bike Data - 112,443 rows , 13 columns

Citi Bike Data - 112,443 rows , 13 columns

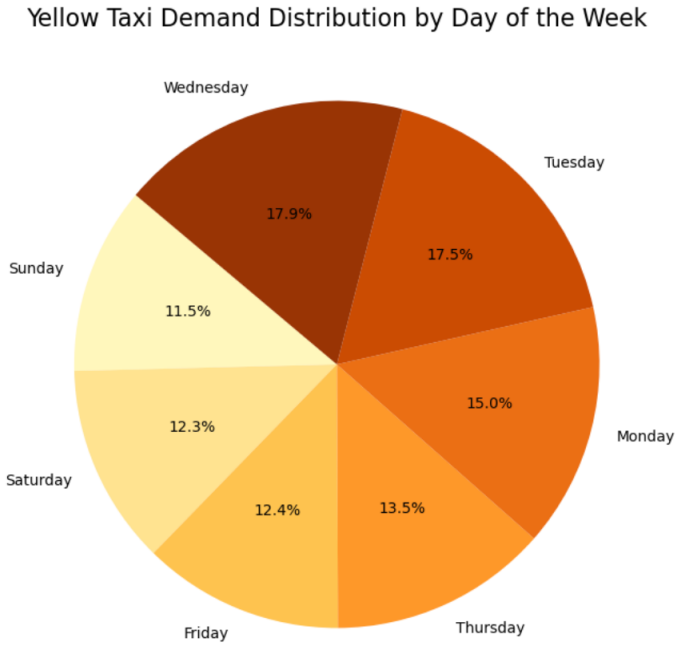
Exploratory Data Analysis

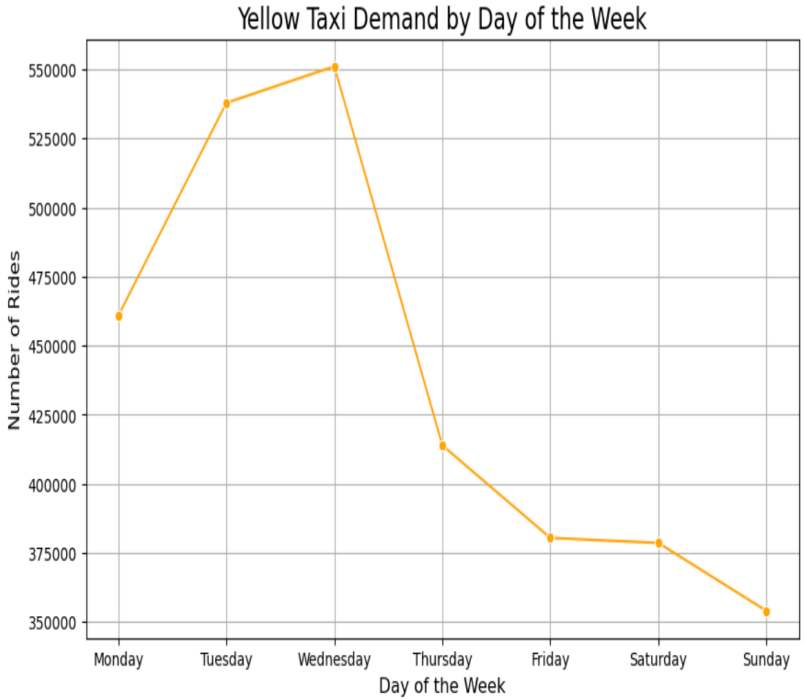
Citi Bike


Citi Bike Demand Distribution by Day of the Week
Wednesday
Tuesday
16.8%
16.7%
Sunday
11.7%
15.8%
Monday
11.9%
Saturday
12.9%
14.3%
Friday
Thursday

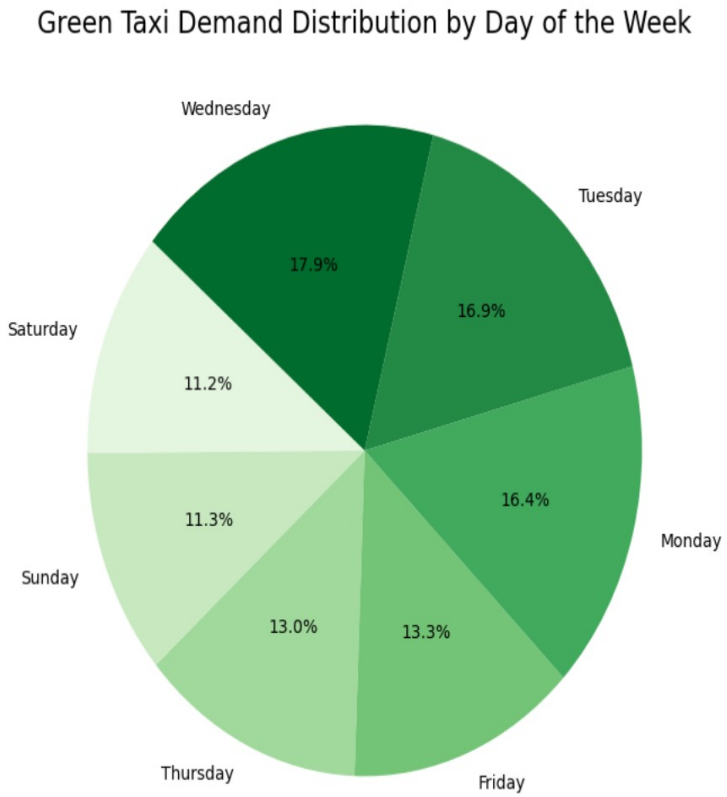

Citi Bike Demand by Day of the Week
Number of Rides
19000
18000
17000
16000
15000
14000
13000
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of the Week

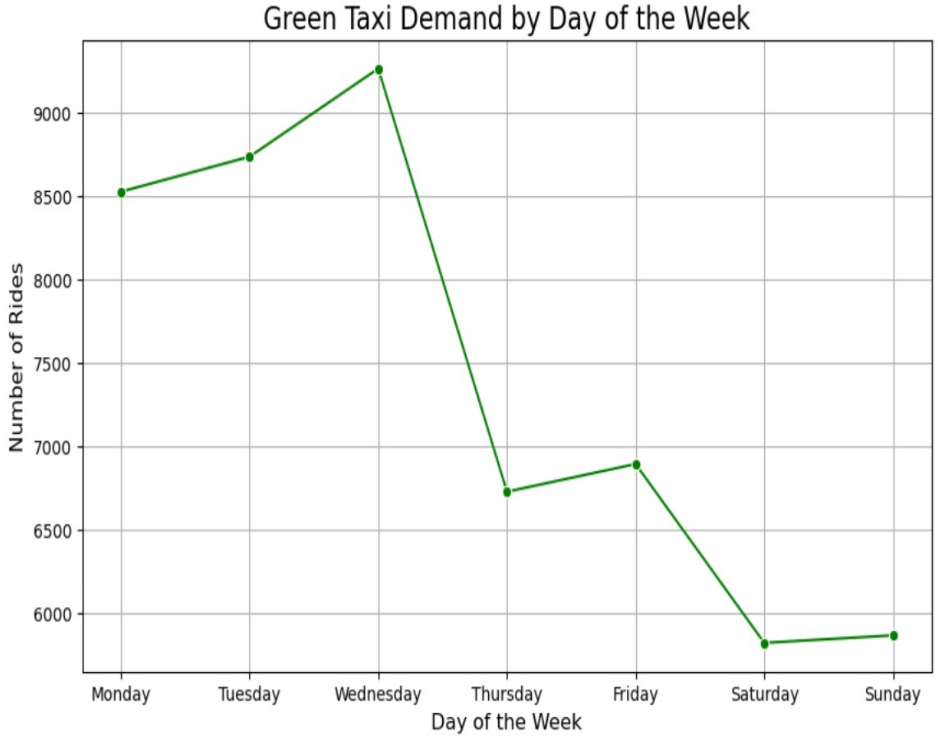
Yellow Taxi


Yellow Taxi Demand Distribution by Day of the Week
Wednesday
Tuesday
Sunday
Monday
Saturday
Friday
Thursday
17.9%
17.5%
11.5%
15.0%
12.3%
12.4%
13.5%


Yellow Taxi Demand by Day of the Week
Number of Rides
550000
525000
500000
475000
450000
425000
400000
375000
350000
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of the Week

Green Taxi


Green Taxi Demand Distribution by Day of the Week
Wednesday
Tuesday
Saturday
Monday
Sunday
Thursday
Friday
17.9%
16.9%
11.2%
16.4%
11.3%
13.0%
13.3%


Green Taxi Demand by Day of the Week
Number of Rides
9000
8500
8000
7500
7000
6500
6000
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of the Week

The analysis of Citi Bike, Yellow Taxi, and Green Taxi demand by day of the week, we can conclude:

PEAK DEMAND DAYS

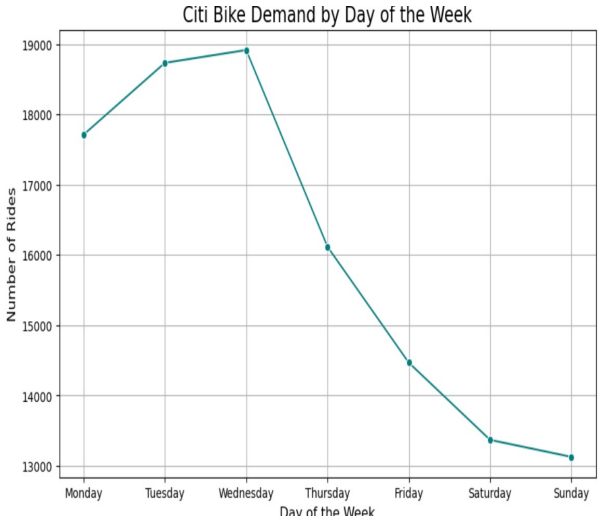
SERVICE-SPECIFIC PATTERNS

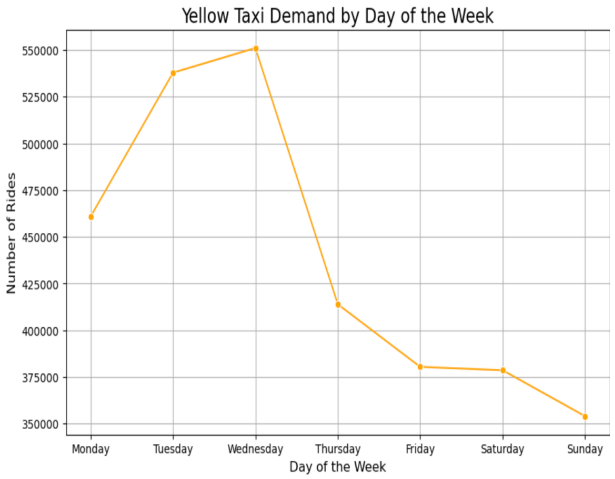
WEEKEND USAGE

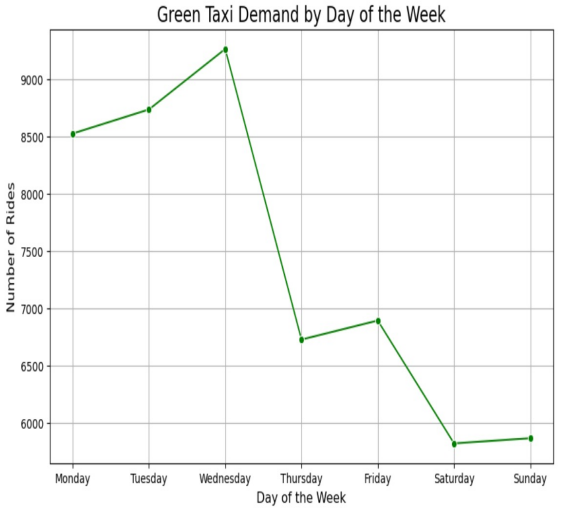
IMPLICATIONS FOR TRANSPORTATION PLANNING

PEAK DEMAND DAYS

•All three modes of transportation show peak demand on weekdays, with Wednesday being particularly high across services.
•Weekends (Saturday and Sunday) show lower demand, which could indicate reduced commuter traffic and more leisure-oriented travel.





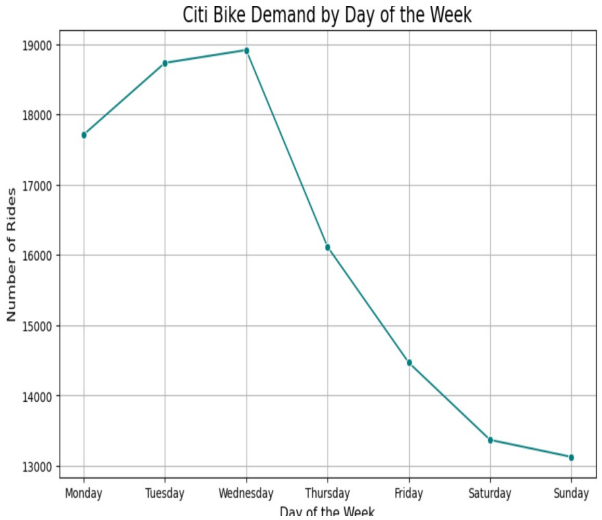


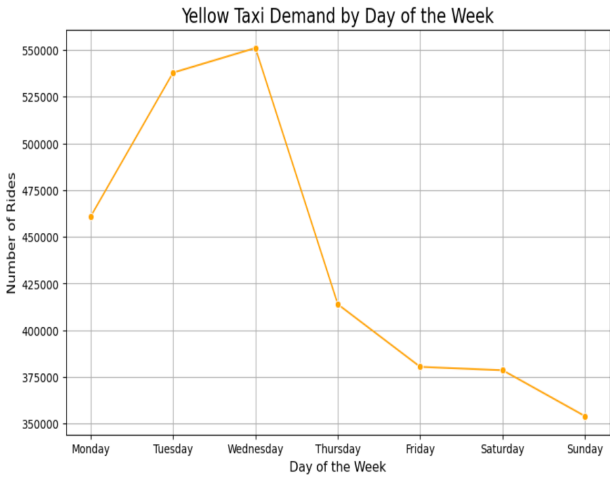
SERVICE-SPECIFIC PATTERNS

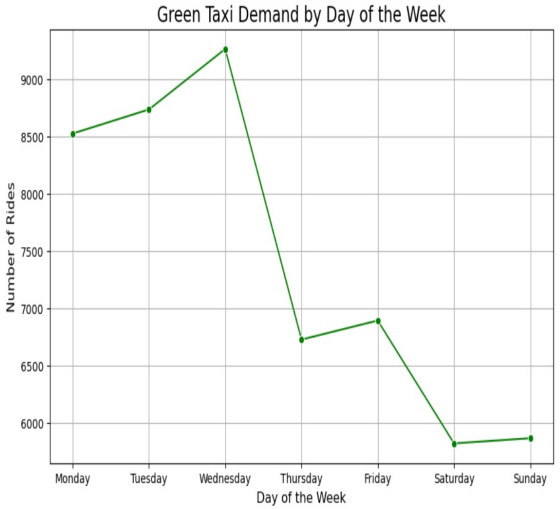
•**Citi Bike**: Has more evenly distributed usage throughout the week, but a noticeable drop in demand towards the weekend. This could reflect commuter reliance on bikes during the work week.

•**Yellow Taxi**: Maintains a high level of usage on weekdays, with a significant peak on Wednesday, possibly due to business travel. Yellow taxis are in heavy demand mid-week and taper off towards the weekend.

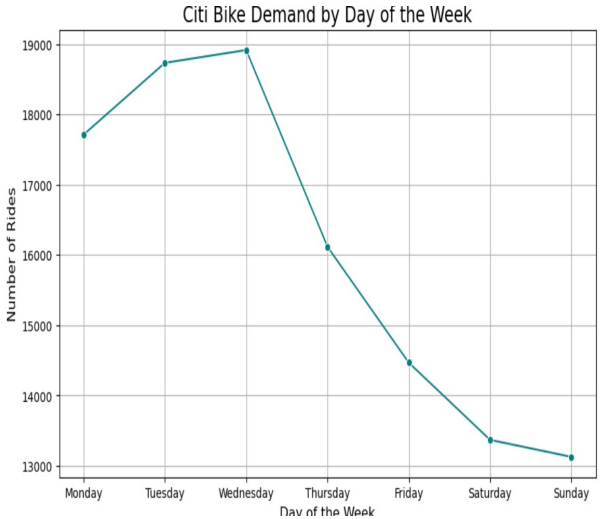
•**Green Taxi**: Has similar demand trends to Yellow Taxis, with a mid-week peak, but slightly more distributed demand across the week due to its coverage of outer boroughs.

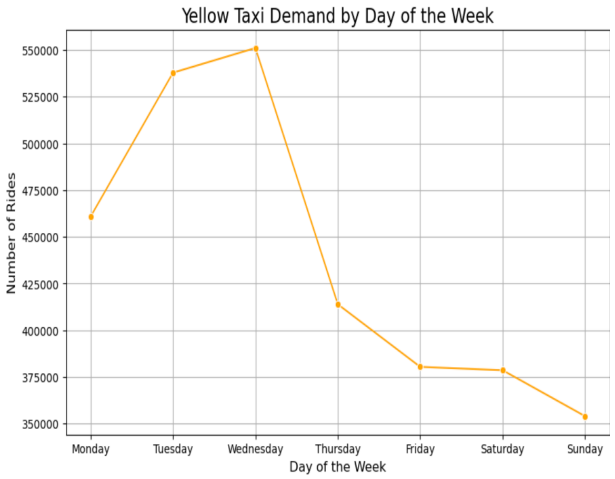


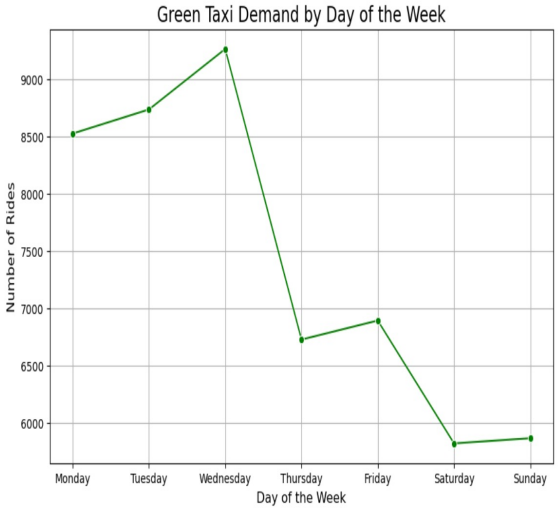




Weekend Usage: Citi Bike maintains more weekend usage than both Yellow and Green taxis, indicating it may be favored for recreational use in addition to weekday commuting.

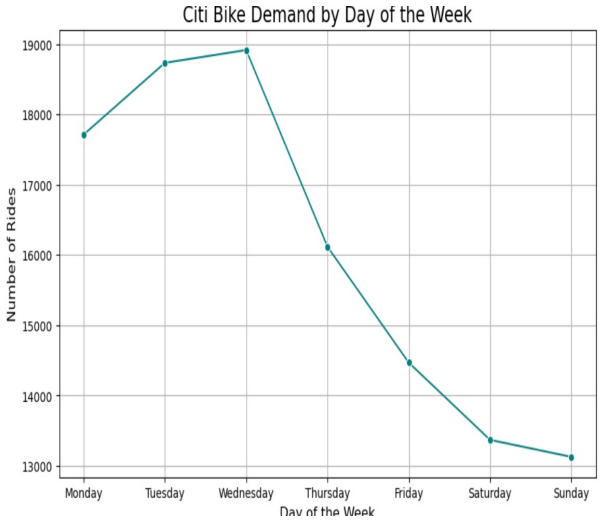


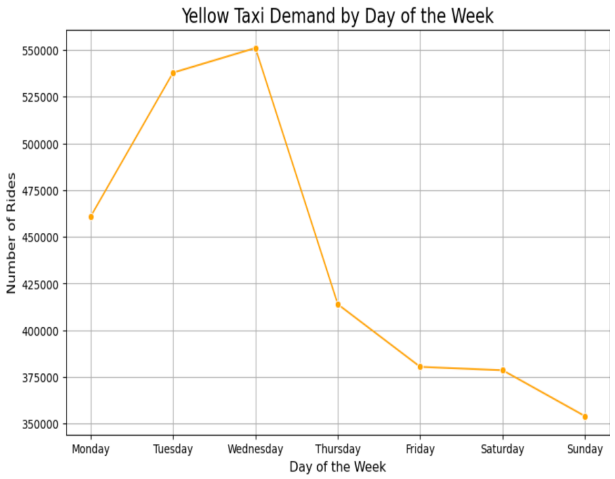


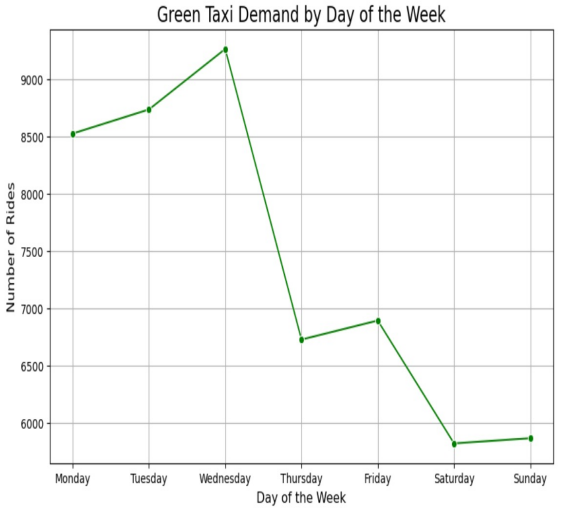


IMPLICATIONS FOR TRANSPORTATION PLANNING

- The distinct peaks and troughs in demand highlight opportunities for optimizing fleet distribution and staffing, especially for weekdays versus weekends.
- Insights from peak demand days could guide potential partnerships, promotional pricing, or other strategies to balance demand throughout the week.







Identifying Key Business Questions for Analysis

As we dive into this dataset, we aim to explore and address important business questions, which will help uncover actionable insights and support data-driven decisions.

Understanding the Top locations of each Transport modes

Top Locations by Demand

Top Green Taxi Locations

Rank	PULocationID	Demand
1	74	12177
2	75	7229
3	43	2841
4	82	2399
5	41	2225

Top Yellow Taxi Locations

Rank	PULocationID	Demand
1	132	182054
2	161	156024
3	237	128579
4	162	113486
5	236	109672

Top CitiBike Stations

Rank	Start Station ID	Demand
1	HB102	5430
2	JC115	4951
3	HB105	3035
4	JC066	2663
5	HB103	2624

But as we see the location is present in terms of id

We found out two files
taxi_zones.csv
citibike_stations

who can help us to find the exact location of the transportation individually.....

And the result after using this.....

Top Green Taxi Locations with Names and Boroughs:

Rank	Zone	Borough	Demand
1	East Harlem North	Manhattan	12,177
2	East Harlem South	Manhattan	7,229
3	Central Park	Manhattan	2,841
4	Elmhurst	Queens	2,399
5	Central Harlem	Manhattan	2,225

Top Yellow Taxi locations:

Rank	Zone	Borough	Demand
1	JFK Airport	Queens	182,054
2	Midtown Center	Manhattan	156,024
3	Upper East Side South	Manhattan	128,579
4	Midtown East	Manhattan	113,486
5	Upper East Side North	Manhattan	109,672

Top CitiBike stations by demand:

Rank	Start Station Name	Demand
1	Hoboken Terminal - River St & Hudson Pl	5,430
2	Grove St PATH	4,951
3	City Hall - Washington St & 1 St	3,035
4	Newport PATH	2,663
5	South Waterfront Walkway - Sinatra Dr & 1 St	2,624

Predictive Modeling for Demand Forecasting

Question: Can we predict peak demand times for taxis and bikes in different locations using historical data?

Application: Using regression techniques (such as linear regression or time series analysis), forecast demand based on historical hourly and daily trends.

Business Value: Stakeholders can use demand forecasts to adjust fleet allocations, staffing, and bike/taxis availability, potentially increasing availability and reducing waiting times.

Feature Engineering:
We add new time-based features:

- **pickup_hour**: The hour of the day the ride started.
- **day_of_week**: The day of the week the ride started.
- **is_weekend**: A binary flag indicating if the ride took place on a weekend (Saturday/Sunday).
- **is_holiday**: A binary flag indicating if the ride took place on a public holiday (based on U.S. holidays).

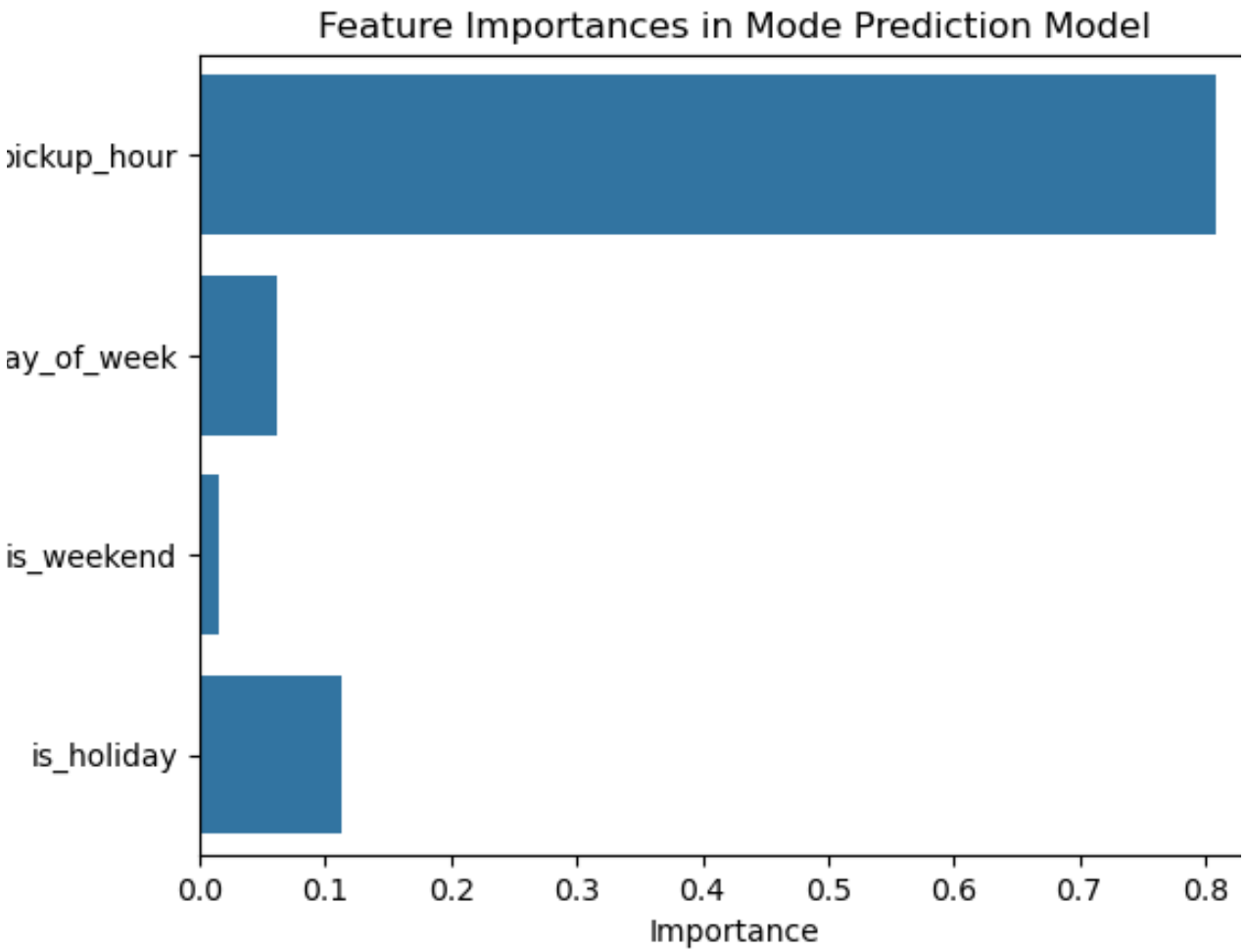
Combining Datasets:
The green taxi, yellow taxi, and CitiBike datasets are standardized and combined into one dataset with the same columns. The mode column is encoded as binary, where 0 represents taxis and 1 represents bikes, for use in the predictive model.

Modeling for Demand Prediction:
The combined data is split into **features** (pickup_hour, day_of_week, is_weekend, is_holiday) and **target** (mode) variables.
The features are scaled to ensure that they are on the same scale, improving model performance.

Train-Test Split:
The dataset is split into training and test sets (70% training, 30% testing), so the model can be evaluated on unseen data.

Model Training:
A **RandomForestClassifier** model is trained on the data to predict the "mode" (whether the ride is likely to be a bike or a taxi) based on time-based features.

Model Evaluation:
The classification report and **accuracy score**, shows us how well the model distinguishes between taxis and bikes. accuracy is (97%)



This analysis highlights that:

- The time of day (pickup_hour) is the most important factor in determining demand for taxis vs. bikes.
- Weekdays and weekends show demand variation, likely due to commuter behavior.
- This model can help stakeholders forecast demand patterns and adjust resources accordingly.

How can we help travelers determine the most economical transport option for their journey, and how can this inform our service offerings?

Define Constants for Each Mode of Transport:
"This analysis defines key parameters—average speeds, base fares, and costs—for Yellow Taxis, Green Taxis, and CitiBike, forming the basis for evaluating each mode's efficiency and affordability."

Prepare Taxi Data:
The helper function (prepare_data_for_model) preprocesses taxi data by extracting features like trip distance, passenger count, hour, and day of the week, preparing it for a fare prediction model.

Train a Fare Prediction Model for Taxis:
The data is cleaned and used to train a Regression model to predict taxi fares based on trip distance and other factors, enabling accurate fare estimation.

Calculate Trip Cost and Time for Each Mode
The `calculate_trip_cost_and_time` function estimates fare and travel time for Yellow Taxi, Green Taxi, or CitiBike trip
Surge Factor: A 1.5x surge applies during peak hours (7-10 AM, 4-7 PM).
- **Yellow Taxi**: Fare includes base fare, per-mile rate, and surge factor; travel time is based on average speed.
- **Green Taxi**: Similar to Yellow Taxi, with adjusted base fare and per-mile rate.
- **CitiBike**: The base fare covers 30 minutes, with per-minute charges beyond that.
The function returns both fare and estimated travel time.

Recommend the Best Transport Option:
The recommend_transport_option function calculates the cost and time for each mode for a given distance and pickup hour. It loops through each mode, calculates the fare and time using calculate_trip_cost_and_time, and stores the results in a list called options

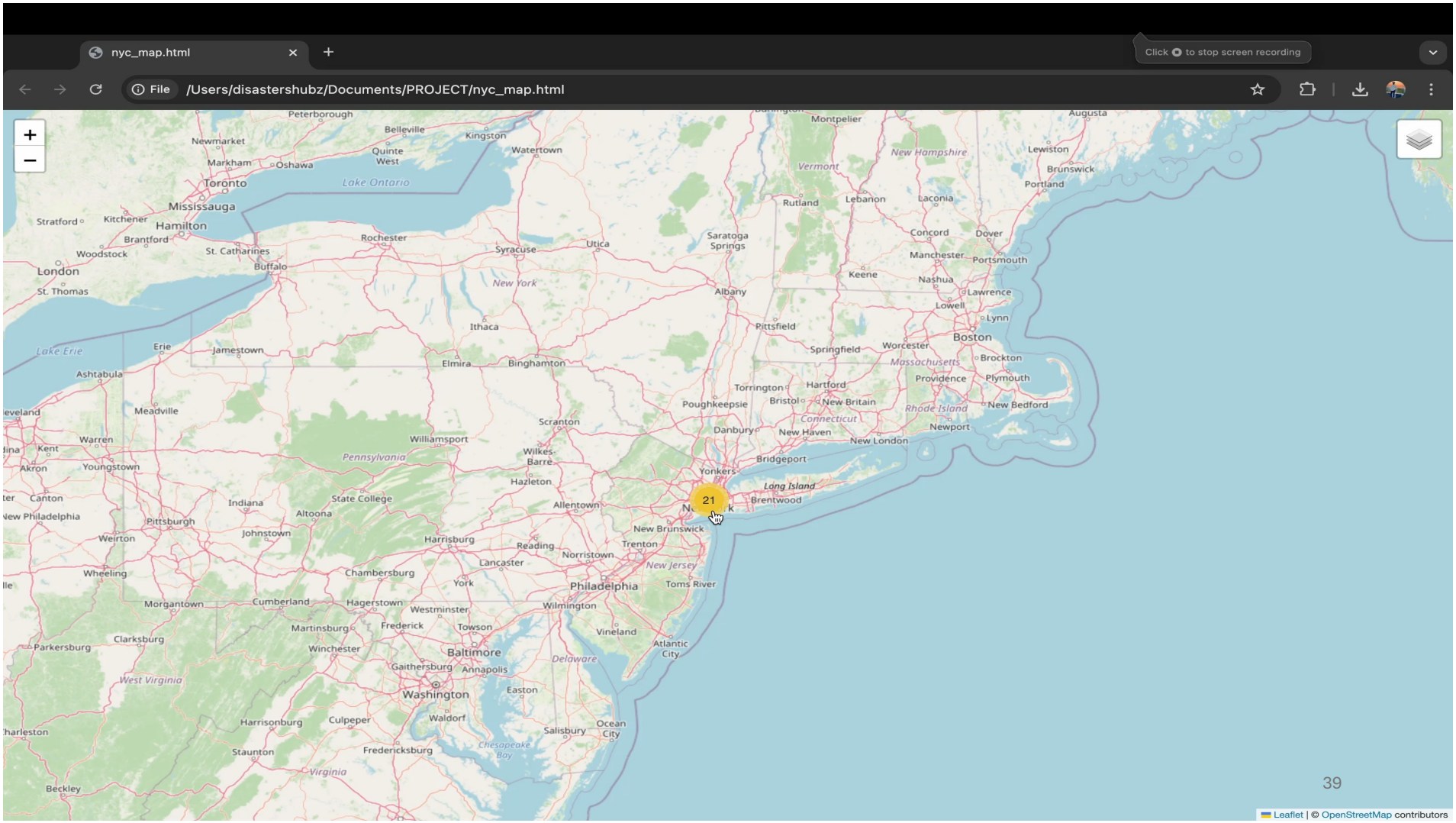
Finding the Best Option:
The cheapest option is identified by finding the mode with the lowest fare.
The fastest option is identified by finding the mode with the shortest travel time.

- **Options for a 4.2-mile trip at 20:00: -**
- Yellow taxi: Cost = $13.00, Time = 16.80 mins
- Green taxi: Cost = $14.30, Time = 16.80 mins
- Citi bike: Cost = $4.79, Time = 25.20 mins

- Cheapest option:
- Citi bike - Cost: $4.79 , Time: 16.80 mins

- Fastest option:
- yellow taxi : - $13.00, Time: 16.80 mins

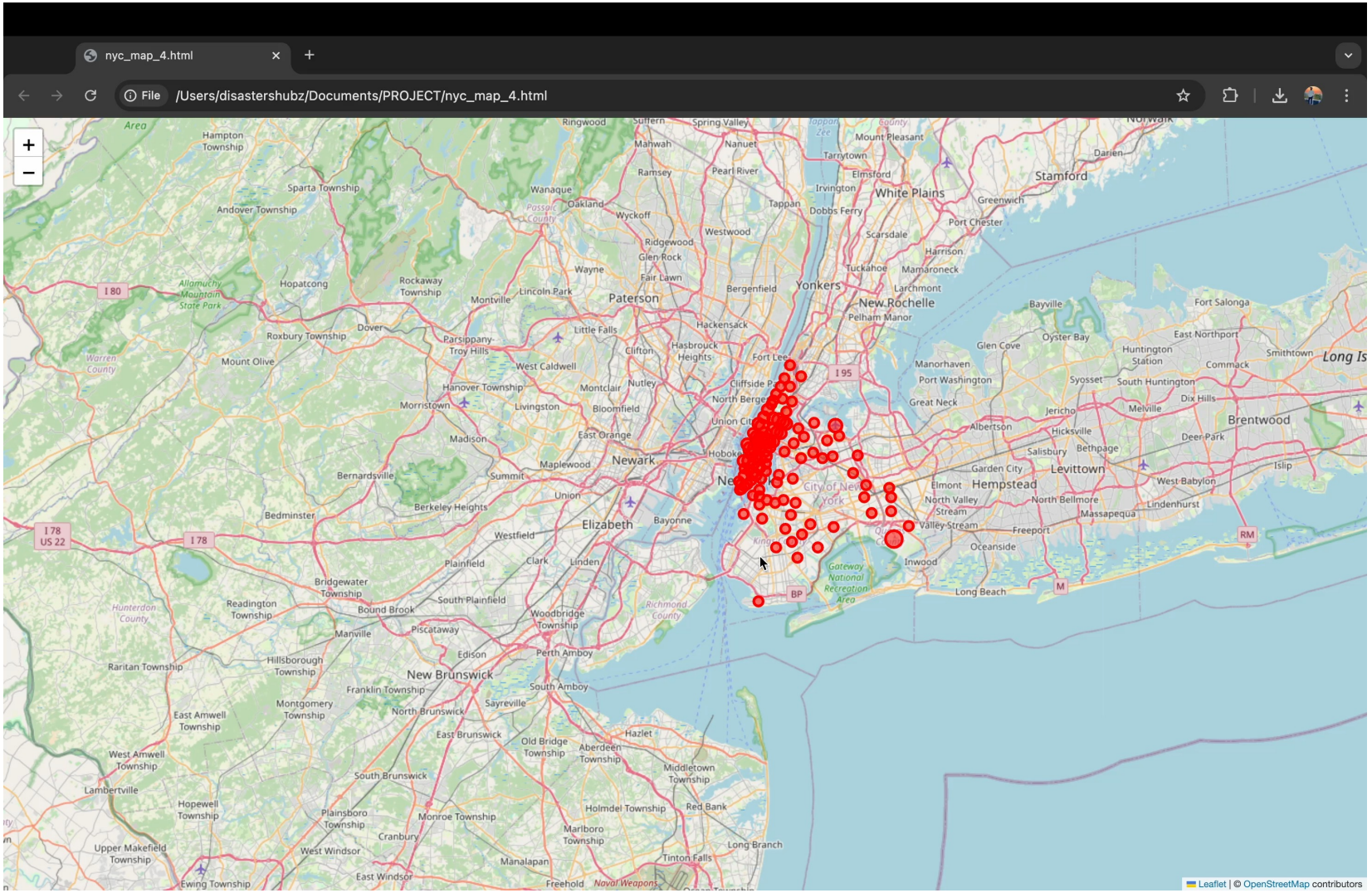

Location-Based Recommendations: For high-density pickup/drop-off zones, can Citi bike stations help relieve taxi congestion?


nyc_map.html
Click to stop screen recording
File
/Users/disastershubz/Documents/PROJECT/nyc_map.html
21
39
Leaflet | © OpenStreetMap contributors

Environmental Agencies:

Track environmental impact by promoting bike usage over taxis in high-congestion areas.

Zone	Borough	Total Trip Count	Estimated CO_2 Emission	CitiBike Station Capacity
JFK Airport	Queens	177,077	88,538.5	0
Midtown Center	Manhattan	142,023	71,011.5	0
Upper East Side South	Manhattan	120,256	60,128	0
Midtown East	Manhattan	105,872	52,936	0
Penn Station/Madison Sq West	Manhattan	101,484	50,742	0
Richmond Hill	Queens	521	260.5	0
South Jamaica	Queens	519	259.5	0
Old Astoria	Queens	515	257.5	0
Clinton Hill	Brooklyn	507	253.5	0
Springfield Gardens	Queens	501	250.5	0


nyc_map_4.html
File
/Users/disastershubz/Documents/PROJECT/nyc_map_4.html
Leaflet | © OpenStreetMap contributors

"Over time, there are numerous opportunities for continuous improvement in our services and systems. By leveraging data insights and feedback, we can optimize operations, enhance customer satisfaction, and foster sustainable growth.

Key Takeaway: Our commitment to ongoing enhancement ensures we remain adaptable and forward-thinking, ultimately driving better outcomes for our stakeholders and the communities we serve."

Our Potential Stakeholder we kept in mind while dealing this project

City Transport Planners:

New York City Department of Transportation (NYC DOT): Oversees transportation planning and infrastructure in NYC.

Metropolitan Transportation Authority (MTA): Manages public transit systems, including buses and subways.

Ride-Hailing Companies:

Uber NYC: Significant presence in NYC for taxis and ridesharing.

Lyft NYC: Competitor to Uber with bike-sharing services via Citibike.

Citibike Operations:

Citibike NYC (operated by Lyft): The largest bike-sharing program in the city, directly involved in expanding bike infrastructure.

Environmental Agencies:

NYC Mayor's Office of Sustainability: Promotes initiatives to reduce congestion and environmental impacts.

New York State Department of Environmental Conservation (NYS DEC): Works on state-wide and city-specific environmental goals.

Environmental Defense Fund (EDF): Collaborates with the city on climate-friendly transportation policies.

Thank you