NYC Taxi Analysis Report

Introduction

At the heart of the New York City transportation ecosystem is the iconic yellow taxi, which has served millions of residents and visitors for over a century. However, in recent years, the landscape has evolved dramatically with the rise of ride-hailing services such as Uber and Lyft, challenging the traditional yellow taxi industry.

In 2020 alone, the New York City Taxi and Limousine Commission (TLC) reported approximately 71 million yellow taxi trips, a significant decrease from the 175 million trips in 2013, before the introduction of Uber and Lyft. Meanwhile, the number of trips completed by ride-hailing services has skyrocketed, with Uber and Lyft collectively accounting for over 500 million trips in the city in the same year.

To understand the factors driving these changes and identify strategies for yellow taxis to remain competitive, we have conducted a comprehensive analysis of trip records for yellow taxis, Uber, and Lyft. Our analysis focuses on key aspects such as fares, service coverage & availability, and operational efficiency.

We will also provide data-driven recommendations for the yellow taxi industry by presenting our findings, highlighting the areas where yellow taxis can leverage their unique strengths and address potential weaknesses.

Data Sources

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

In order to conduct a comprehensive analysis of the yellow taxi industry in New York City, we utilized the following data sources:

TLC Trip Record Data - 2022 November (Yellow Taxi only):
The primary source of data for this analysis is the 2022 November TLC Trip Record Data for yellow taxis, provided by the New York City Taxi and Limousine Commission (TLC), it contains detailed trip-level data.

Taxi Zone Lookup Table:
To supplement the TLC Trip Record Data, we used the Taxi Zone Lookup Table, which provides information about the LocationID, Borough, Zone, and service zone for each taxi trip. This lookup table allows us to better understand the geographical distribution of taxi trips across New York City .

High-Volume For-Hire Vehicle Records (Includes Uber/Lyft data):
In order to compare the performance of yellow taxis with that of their ride-hailing competitors, we utilized the High-Volume For-Hire Vehicle Records dataset, which includes trip-level data for Uber and Lyft in New York City.

Data Preparation and Approach

- We began by extracting the November 2022 trip records data (Parquet file) for yellow taxis
- Next, we joined this dataset with the location lookup data to obtain pick-up and drop-off location information, including Borough, Zone, and Service Zone.
- We then incorporated the High-Volume For-Hire Vehicle (FHV) data, which contains Uber and Lyft trip records for November 2022
- To distinguish between the two types of services, we created a flag for Uber/Lyft (1) and yellow taxi (0)
- We extracted the hour from the pick-up time to analyze hourly traffic patterns for FHV and yellow taxi trips
- Similarly, we derived the day from the pick-up date-time to examine daily traffic patterns for both FHV and yellow taxi trips

The dataset used in this analysis consists of variables such as 'pickup_datetime', 'dropoff_datetime', 'Pick up LocationID', 'Drop off LocationID', 'trip_miles', 'base_passenger_fare', 'Taxi or uber/lyft', 'LocationID', 'PUBorough', 'PUZone', 'PUservice_zone', 'hour', 'day', and 'month'. This rich set of variables enables us to derive valuable insights into the research problem by examining the following aspects:

Pricing strategy: The 'base_passenger_fare' and 'trip_miles' variables in the dataset allow us to compare the fares of yellow taxis, Uber, and Lyft for similar distances and locations. By analyzing the fare data, we can examine pricing trends over time and identify peak and off-peak fare differences between the services. This information can be crucial in helping yellow taxis develop a more competitive pricing strategy.

Service coverage and availability: The 'pickup_datetime', 'dropoff_datetime', 'Pick up LocationID', 'Drop off LocationID', 'PUBorough', 'PUZone', and 'PUservice_zone' variables enable us to analyze location and trip data to determine high-demand areas and times where ride-hailing services dominate, as well as areas with limited service from both yellow taxis and ride-hailing companies. By identifying these gaps, help us identify patterns in service availability during specific timeframes, which can be valuable for yellow taxis to optimize their operations and better serve areas with unmet demand.

Data Analysis

Yellow taxis primarily operate in Manhattan and Queens, with approximately 2 million and 300,000 rides respectively. In contrast, Uber and Lyft function across almost all boroughs, with Manhattan and Brooklyn accounting for the highest number of rides (7,586,146 and 4,602,459 respectively).**(Figure 1)**

1. **Zone-wise percentage of trips for yellow cabs: (Figure 2)**
   Yellow cabs constitute a significant portion of total rides in certain zones. In Central Park, they account for 66.10% of all rides, while in Upper East Side North and Upper East Side South, the percentages are 48.65% and 48.16% respectively.

2. **Hour and zone-wise percentage of trips for yellow cabs and Uber/Lyft: (Figure3)**
Between 7 am and 4 pm, yellow taxis experience peak demand across the top zones, compared to the total rides (including Uber and Lyft). In Central Park, yellow cabs have the highest number of rides compared to FHV throughout the day, peaking at 11 am with a 72.70% share. Interestingly, at Penn Station, the percentage of rides by yellow taxis peaks to 64% of total rides in the area at 6 am.

3. **Day-wise analysis of top zones for yellow cabs and FHV (Figure 4)**
Central Park accounts for the maximum ride percentage across all days, with Mondays being the highest at approximately 70%. In Central Park, yellow cabs have the most number of rides compared to FHV across all hours of the day, peaking at 11 am with a 72.70% share. Another interesting observation is that in Penn Station, the percentage of rides by yellow taxis peaks to 64% of total rides in the area at 6 am.

4. **Fare comparisons between yellow taxis and ride-hailing services: (Figure 5)**
We examined zones where yellow taxis charge significantly less than Uber and Lyft. Rides in Penn Station by yellow taxis cost $16.91 less than FHV rides, while in Central Park, yellow cabs charge $14.28 less than FHV. Upper East Side has the second-highest percentage of rides by yellow taxis.

In summary, our analysis reveals that yellow taxis continue to be a significant mode of transportation in certain areas of Manhattan and Queens. They also provide a more affordable option in specific zones when compared to ride-hailing services. By understanding these patterns and focusing on areas with higher demand, yellow taxi companies can optimize their services and better compete with Uber and Lyft.

Logistic Regression Model

To better understand the drivers of taxi rides and address our business problem, we employed a supervised machine learning technique for a comprehensive analysis of trip records involving yellow taxis, Uber, and Lyft. We focused on key aspects such as fares, service coverage and availability, and operational efficiency. The regression model consisted of independent variables like "trip_miles," "fare_amount," and dummies for identifying the time of day, the PU Borough Zone, and the day of the week. The dependent variable was "HighVolume," with a value of 1 for Uber/Lyft rides and 0 for Yellow Taxi rides. (Figure 6,7,8)

Results:

- The likelihood of a ride being a high-volume service (Uber/Lyft) versus a Yellow Taxi is not significantly influenced by trip distance or base passenger fare.
- Rides originating from Manhattan and Queens are less likely to be from a high-volume service and more likely to be from a Yellow Taxi.
- In Manhattan, significant zones are Alphabet City, Central Harlem, Central Harlem North, East Harlem North, Highbridge Park, Inwood, and Inwood Hill Park (p-values < 0.05).
- In Queens, statistically significant variables are JFK Airport, LaGuardia Airport, and East Elmhurst, with the F1 score of the model at 0.96.
- Yellow Taxi companies could potentially improve demand by focusing on providing better services and convenience for passengers traveling to and from these locations.

Recommendations

Based on our analysis, we recommend the following strategies for yellow taxi companies to improve their services and better compete with ride-hailing services like Uber and Lyft:

1. **Optimize Service Coverage and Availability:**

- Efficiently allocate cabs in areas with high demand, such as Central Park, Upper East Side North, and Upper East Side South, to better serve passengers and capitalize on the high percentage of yellow taxi rides in these zones.
- Focus on peak hours (7 am to 4 pm) in top zones, ensuring adequate availability of yellow taxis during these times to meet demand.
- Capitalize on the observation that yellow taxis have a 64% share of total rides at Penn Station at 6 am by ensuring a consistent supply of cabs during this time, thus monopolizing this early morning market.

2. **Dynamic Pricing Strategy:**

- In zones where yellow taxis are charging less, such as Penn Station and Central Park, consider allocating more taxis to meet the potential increase in demand as passengers become aware of the lower prices.
- Alternatively, yellow taxis could match their prices to Uber and Lyft in these areas to earn more, while maintaining a competitive edge by offering better service quality and convenience.

3. **Target Key Boroughs and Zones:**

- · Prioritize service in Manhattan and Queens, where rides are more likely to be from a yellow taxi rather than a high-volume service.
- · Enhance service quality and convenience for passengers traveling to and from statistically significant zones like Alphabet City, Central Harlem, Central Harlem North, East Harlem North, Highbridge Park, Inwood, and Inwood Hill Park in Manhattan, as well as JFK Airport, LaGuardia Airport, and East Elmhurst in Queens, since these are areas that Uber/Lyft dominate.

4. **Data-driven Resource Allocation:**

- · Use the analysis of day-wise and hour-wise demand patterns to allocate resources more efficiently, ensuring that yellow taxis are available in high-demand areas during peak times and days.
- · Continuously monitor and analyze trip data to identify emerging trends and adjust resource allocation accordingly to maintain a competitive edge.

By implementing these recommendations, yellow taxi companies can improve their competitiveness, better serve passengers, and ultimately capture a larger share of the market in key areas where demand for their services is high.

Data that can help with better analysis:

1. **Speed and Coordinates:**
   Analyze ride speed and exact coordinates to optimize routing and dispatching systems, improving service efficiency and attracting riders.

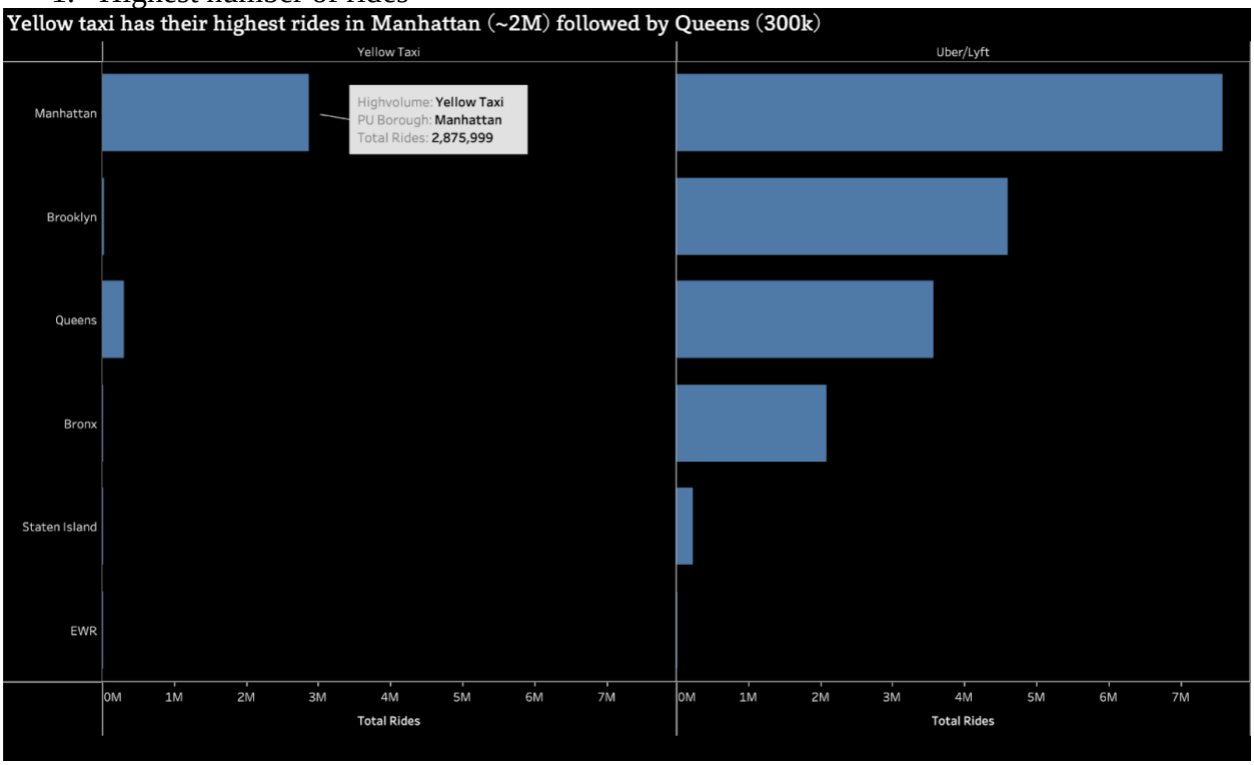
2. **Customer Feedback:**
   Use customer feedback and ratings to identify areas of improvement for yellow taxis and enhance their competitiveness with ride-hailing services.

3. **Socioeconomic Data:**
   Incorporate income levels and population density to understand target markets and tailor services to attract more customers.

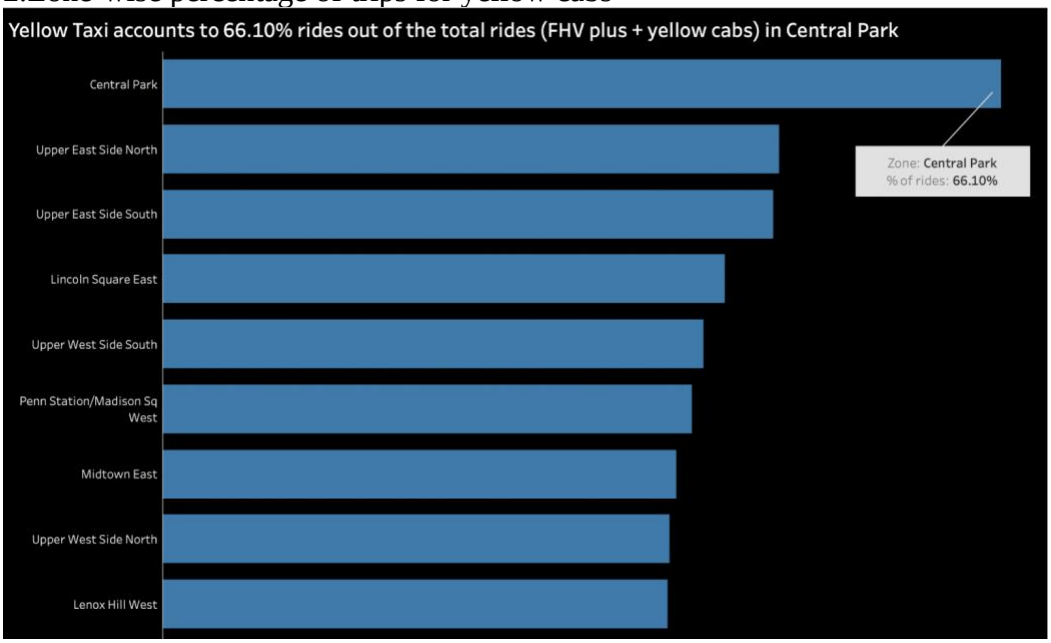
4. **Recurring Events:**
   Consider events like concerts and meetings to anticipate increased demand, enabling better resource allocation and maximizing business potential during these occasions.

Appendix

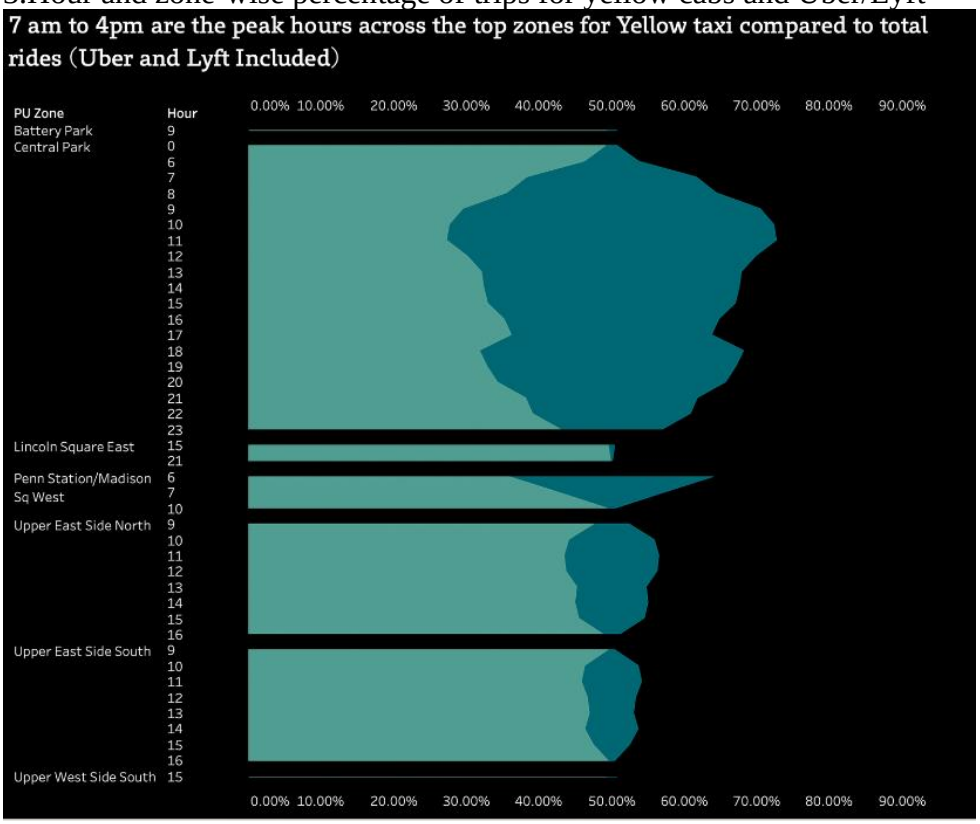
1. Highest number of rides



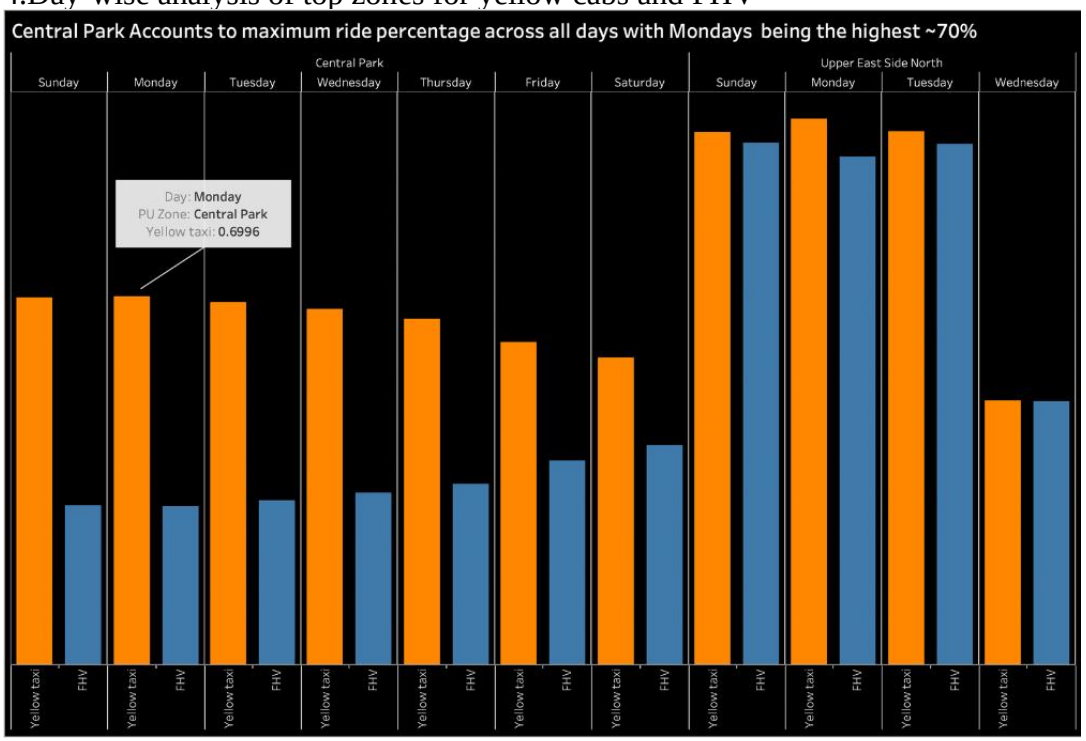
2.Zone-wise percentage of trips for yellow cabs



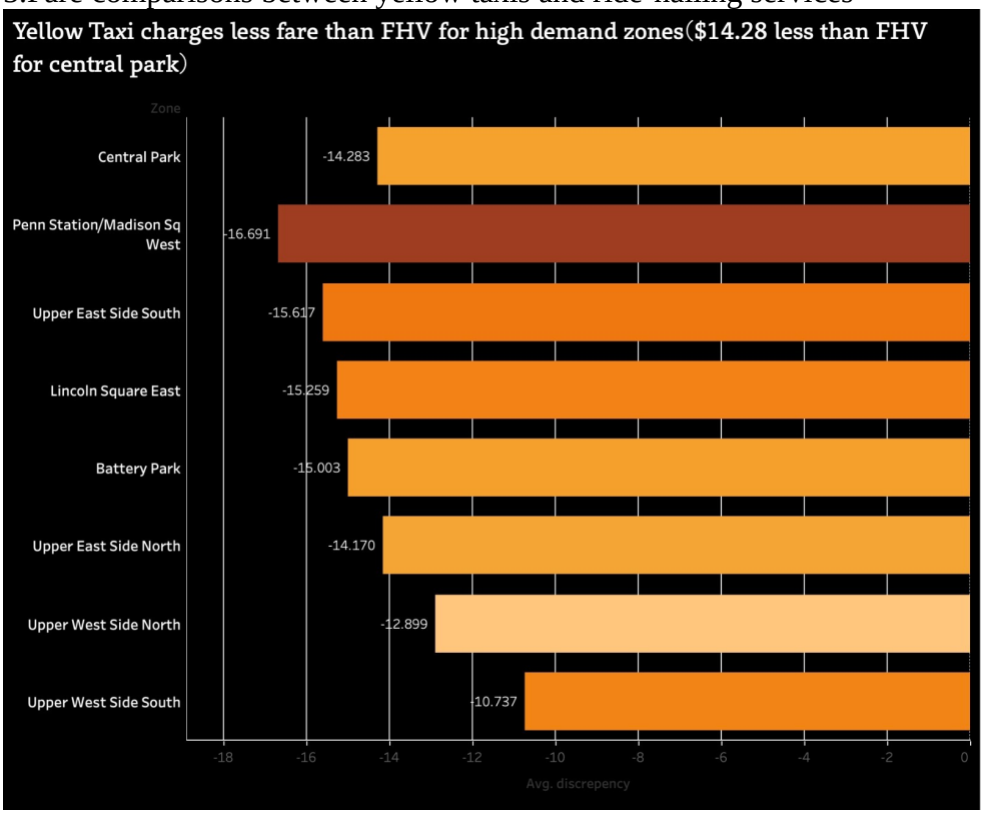
3.Hour and zone-wise percentage of trips for yellow cabs and Uber/Lyft



4.Day-wise analysis of top zones for yellow cabs and FHV



5.Fare comparisons between yellow taxis and ride-hailing services



6. Logistic Regression Model - Output

Feature	Coefficient	p-value
trip_miles	-0.484	0.629
base_passenger_fare	0.199	0.842
Morning	0.72	0.471
Midday	-0.38	0.704
Evening	-0.234	0.815
Night	0.276	0.783
Midnight	0.751	0.453
Monday	-0.196	0.845
Tuesday	-0.161	0.872
Wednesday	-0.127	0.899
Thursday	0.022	0.983
Friday	0.118	0.906
Saturday	0.142	0.887
Bronx	0.45	0.653
Brooklyn	-0.654	0.513
EWR	-18.944	0.000*
Manhattan	-5.43	0.000*
Queens	-3.437	0.001*

7.Logistic Regression Model: Output Manhattan (Significant Zones)

Feature	Coefficient	p-value
trip_miles:	-0.649	0.517
base_passenger_fare:	0.287	0.774
Morning:	0.011	0.991
Mid_morning:	-0.743	0.458
Midday:	-1.159	0.246
Night:	-0.355	0.723
Evening:	-0.97	0.332
Monday:	-0.311	0.756
Tuesday:	-0.305	0.76
Wednesday:	-0.325	0.745
Thursday:	-0.126	0.9
Friday:	-0.048	0.962
Saturday:	0.072	0.943
Alphabet City	2.116	0.034*
Central Harlem	2.317	0.021*
Central Harlem North	3.908	0.000*
East Harlem	2.605	0.009*
Hamilton Heights	3.123	0.002*
Highbridge Park	2.198	0.028*
Inwood	5.311	0.000*
Inwood Hill Park	3.105	0.002*
Manhattanville:	2.658	0.008*
Marble Hill	4.622	0.000*
Roosevelt Island	3.268	0.001*
Two Bridges/Seward Parl	2.473	0.013*
Washington Height North	4.979	0.000*
Washington Height South	3.962	0.000*

8.Logistic Regression Model: Output Queens (Significant Zones)

Feature	Coefficient	p-value
trip_miles	-0.105	0.916
base_passenger_fare	0.049	0.961
Morning	-0.447	0.655
Mid_morning	-0.643	0.52
Midday	-0.879	0.379
Night	-0.441	0.659
Evening	-0.765	0.445
Monday	-0.057	0.955
Tuesday	-0.04	0.968
Wednesday	0.041	0.967
Thursday	0.107	0.914
Friday	0.137	0.891
Saturday	0.147	0.883
East Elmhurst	-4.065	0.000*
JFK Airport	-4.781	0.000*
LaGuardia Airport	-4.108	0.000*
Flushing Meadows-Corona Park	-3.048	0.002*
Queensbridge/Ravenswood	-2.886	0.004*
Saint Michaels Cemetery/Woodside	-2.769	0.006*