Taxi & Limousine Commission

# New York City Taxi Rides: A Study on Green Taxi

This study utilizes the **NYC Green Taxi Trip dataset**, which records numerous taxi trips across the city, including timestamps, fares, distances, and locations.

# Introduction.

# New York City Taxi and Limousine Commission

created in 1971, is the agency responsible for licensing and regulating New York City's medallion (yellow) taxis, street hail livery (green) taxis, for-hire vehicles (FHVs), commuter vans, and paratransit vehicles. The TLC collects trip record information for each taxi and for-hire vehicle trip completed by our licensed drivers and vehicles.

TCL receive taxi trip data from the technology service providers (TSPs) that provide electronic metering in each cab, and FHV trip data from the app, community livery, black car, or luxury limousine company, or base, who dispatched the trip.

# Problem Statement

**To analyze trip demand patterns NYC taxi rides using temporal, geographical, and fare-based attributes.**

## Understanding NYC Taxi Trips:

### When and where are NYC taxis most in demand?

Covers peak hours, days, seasons, and taxi hotspots by location/borough.

### What factors influence taxi fares?

Explores fare components like trip distance, surcharge, zone comparison, and negotiated rates.

### How do passenger tipping behaviors vary by trip characteristics and location?

Examines how tips relate to trip length, time of day, and boroughs.

### Which pickup/dropoff patterns and routes dominates the NYC taxi system?

Includes most frequent routes, fare-per-mile insights, and trip duration by zone.

# About the Dataset.


# Dataset Description

This dataset contains green taxi trip records. Green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP).

In each trip record dataset, one row represents a single trip made by a TLC-licensed vehicle.

Dataset can be obtained from one of these two links:
NYC TLC - (PARQUET ONLY)
NYC OPEN DATA

# Overview of the Main Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 68211 entries, 0 to 68210
Data columns (total 20 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   VendorID               68211 non-null  int64
 1   lpep_pickup_datetime   68211 non-null  object
 2   lpep_dropoff_datetime  68211 non-null  object
 3   store_and_fwd_flag     63887 non-null  object
 4   RatecodeID             63887 non-null  float64
 5   PULocationID           68211 non-null  int64
 6   DOLocationID           68211 non-null  int64
 7   passenger_count        63887 non-null  float64
 8   trip_distance          68211 non-null  float64
 9   fare_amount            68211 non-null  float64
 10  extra                  68211 non-null  float64
 11  mta_tax                68211 non-null  float64
 12  tip_amount             68211 non-null  float64
 13  tolls_amount           68211 non-null  float64
 14  ehail_fee              0 non-null      float64
 15  improvement_surcharge  68211 non-null  float64
 16  total_amount           68211 non-null  float64
 17  payment_type           63887 non-null  float64
 18  trip_type              63877 non-null  float64
 19  congestion_surcharge   63887 non-null  float64
dtypes: float64(14), int64(3), object(3)
memory usage: 10.4+ MB
```

***# of unique values***

```
VendorID                    2
lpep_pickup_datetime    66575
lpep_dropoff_datetime   66519
store_and_fwd_flag          2
RatecodeID                  6
PULocationID              226
DOLocationID              249
passenger_count            10
trip_distance            1870
fare_amount              2553
extra                      16
mta_tax                     6
tip_amount               1492
tolls_amount               26
ehail_fee                   0
improvement_surcharge       5
total_amount             4670
payment_type                5
trip_type                   2
congestion_surcharge        4
dtype: int64
```

***# of missing values**** 

```
VendorID                    0
lpep_pickup_datetime        0
lpep_dropoff_datetime       0
store_and_fwd_flag       4324
RatecodeID               4324
PULocationID                0
DOLocationID                0
passenger_count          4324
trip_distance               0
fare_amount                 0
extra                       0
mta_tax                     0
tip_amount                  0
tolls_amount                0
ehail_fee               68211
improvement_surcharge       0
total_amount                0
payment_type             4324
trip_type                4334
congestion_surcharge     4324
dtype: int64
```

**initial*

# Categorical Features

| Name | Values |
|---|---|
| VendorID | 1 = Creative Mobile Technologies, LLC<br>2 = VeriFone Inc. |
| store_and_fwd_flag | Y = store and forward trip<br>N = not a store and forward trip |
| RatecodeID | 1 = Standard rate; 2 = JFK; 3 = Newark<br>4 = Nassau or Westchester<br>5 = Negotiated Fare; 6 = Group ride<br>99 = Null/unknown |
| payment_type | 1= Credit card; 2= Cash; 3= No charge<br>4= Dispute; 5= Unknown; 6= Voided trip |
| trip_type | 1 = Street-hail; 2 = Dispatch |
| PULocationID<br>DOLocationID | TLC Taxi Zone ID in which the taximeter was engaged/disengaged |

# Numerical Features

| Name | Description |
|---|---|
| lpep_pickup_datetime,<br>lpep_dropoff_datetime | The date & time when the meter was engaged/disengaged |
| passenger_count | # of passengers in the vehicle. |
| trip_distance | The elapsed trip distance in miles reported by the taximeter. |
| fare_amount, extra, mta_tax,<br>tip_amount, tolls_amount,<br>improvement_surcharge | Price breakdown details of each trip. |
| total_amount | Total amount charged to passengers. Does not include cash tips. |
| congestion_surcharge | Total amount collected in trip for NYS congestion surcharge. |

# Overview of the Supporting Dataset

To enrich the analysis, additional datasets were used alongside the main trip data:

## NYC Taxi Zone Lookup Table

To map **PULocationID** and **DOLocationID** to real-world zones and boroughs.

The dataset contains 265 entries of:

1. OBJECTID
2. Shape_Leng
3. Shape_Area
4. **zone**
5. **LocationID**
6. **borough**
7. **geometry**

Used For:

- Categorizing pickup and drop-off zones
- Visualizing geographic patterns
- Filtering by boroughs

## NYC Weather

To analyze taxi demand trends against weather (especially temperature).

Contains:

- AWND, FMTM, PGTM
- **PRCP**
- **SNOW**
- **SNWD**
- **TAVG/TMAX/TMIN**
- WDF2, WDF5, WSF2, WSF5
- WT**

Used For:

- Detecting how cold weather impacts demand
- Adding temperature data to trips via date merge

**These datasets were merged using LocationID and pickup_date respectively, allowing more contextual insights into trip patterns.**

# Data Preprocessing.

# The dataset underwent several preprocessing steps to ensure accuracy and consistency.

## Dropping uneccessary column(s)

Dropping column ***ehail_fee*** since the NYC TLC website stated that it is currently not used.

## Dropping Missing Values

- **RatecodeID = 99** and **trip_type = NaN**
  Value 99 in **RatecodeID** is set to be Null/unknown based on the dataset's documentation. There are 10 entries with this value.
- Missing values in **store_and_fwd_flag, RatecodeID passenger_count, payment_type, trip_type, congestion_surcharge**
  Missing values in these columns are present in the same trip entries.

## Dropping DateTime Outlier

There are a few entries where trips were taken other than January 2023. Since the percentage of these entries is small, dropping them would be better.

# The dataset underwent several preprocessing steps to ensure accuracy and consistency.

## Dropping Cancelled Trips

Trips that are cancelled on the spot are defined as following:

- No travel present after the trip is disengaged.
- The duration of the trip is not present.
- Dropoff is unknown (`DOLocationID` = 264) or same as Pickup (`PULocationID` = `DOLocationID`)
- Less than 3-minute waiting time (the longer it is, the more likely it is not a cancelled trip)
- No congestion surcharge means trips were not stuck in traffic.

## Dropping trips that are more than 10-hours

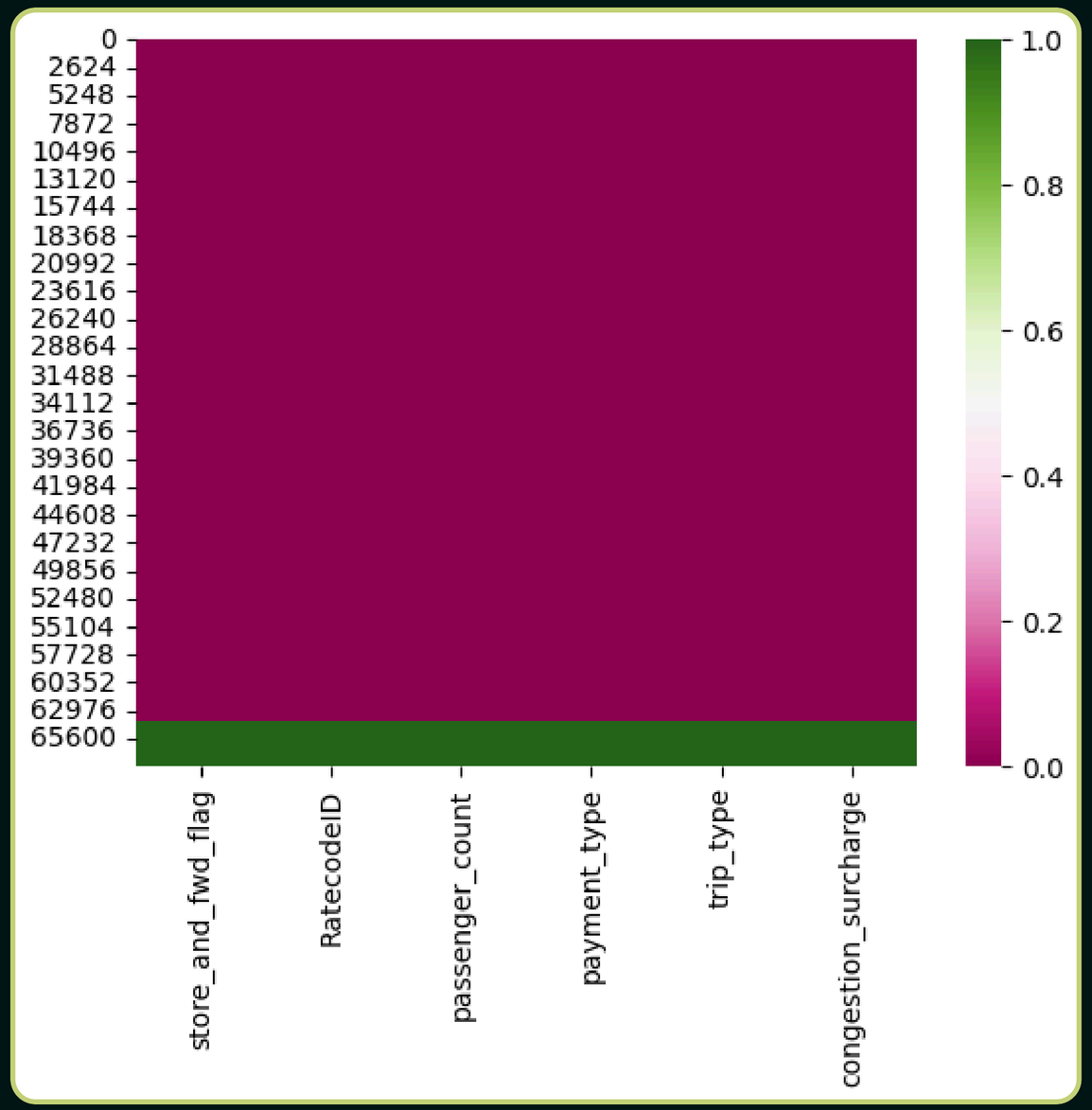
NYC Taxi drivers have limits on hours of driving. The daily driving limit for each driver is 10 hours. A driver cannot transport passengers for hire for more than 10 hours in total within 24 hours. Limit resets after an 8-hour break.

# The dataset underwent several preprocessing steps to ensure accuracy and consistency.

## Dropping Other Anomalous Trips

Anomalous trips are defined as follows:

- **Distance & Duration Anomalies**
  - Extremely short trips - *Near 0 miles but with a long duration (> 5 min) & long trip duration but very short distance.*
  - Extremely long trips - Above 99th percentile (outliers in long distances) & excessively long compared to the distance (e.g., 50+ miles in NYC, unless airport trips).
- **Speed-Based Anomalies**
  - Very slow speed - *speeds < 1 mph over a long distance*
  - Very fast speed - *speeds > 75 mph (highly unlikely in NYC traffic)*
- **Fare Anomalies**
  Unusual fares - *Fare too high for the distance (e.g. $300 for 1 mile)*




# Dropping Missing Values

```
RatecodeID              4324
passenger_count         4324
payment_type            4324
trip_type               4334
congestion_surcharge    4324
dtype: int64
```

trip_type has 10 more entries of null values. This is related to the 10 entries where **RatecodeID = 99**. These entries will be dropped as well.

Null values present in these columns are within the same trip entries. We can assume that these entries might be incomplete. It is possible that these rows were not fully recorded or were processed incorrectly. Thus, the conclusion is to drop the rows.

## Dropping Cancelled Trips

```
len(cancelled_trips)
✓ 0.0s
5375
```

**9.35%**

## Dropping Over 10-hour Trips

```
len(trips_long_600)
✓ 0.0s
204
```

**0.36%**

## Dropping Anomalous Trips

```
len(anomalies)
✓ 0.0s
834
```

**1.45%**

After filtering is applied, we know the number of entries that fall into the filters. Since these numbers are less than 10% of the total entries, dropping these entries would be better.

# Feature Engineering.

# Added Features

Several new features were added to the dataset. Some were extracted from existing columns (such as timestamps), while others were derived by combining multiple features to provide additional insights.

External sources are also merged to the main dataset, thus adding more features. Sources are from:

1. NYC Taxi Zone
2. NYC Weather (NOAA)

## Extracted (Time-based Features)

- **pickup_hour**, **dropoff_hour** - extracted from datetime
- **pickup_dayofweek**, **dropoff_dayofweek** - day of the week
- **pickup_month**, **dropoff_month** - for seasonal trends
- **pickup_period**, **dropoff_period** - time-of-day grouping (AM/PM)
- **is_weekend_pickup**, **is_weekend_dropoff** - weekend indicator

## Constructed (Trip Metrics)

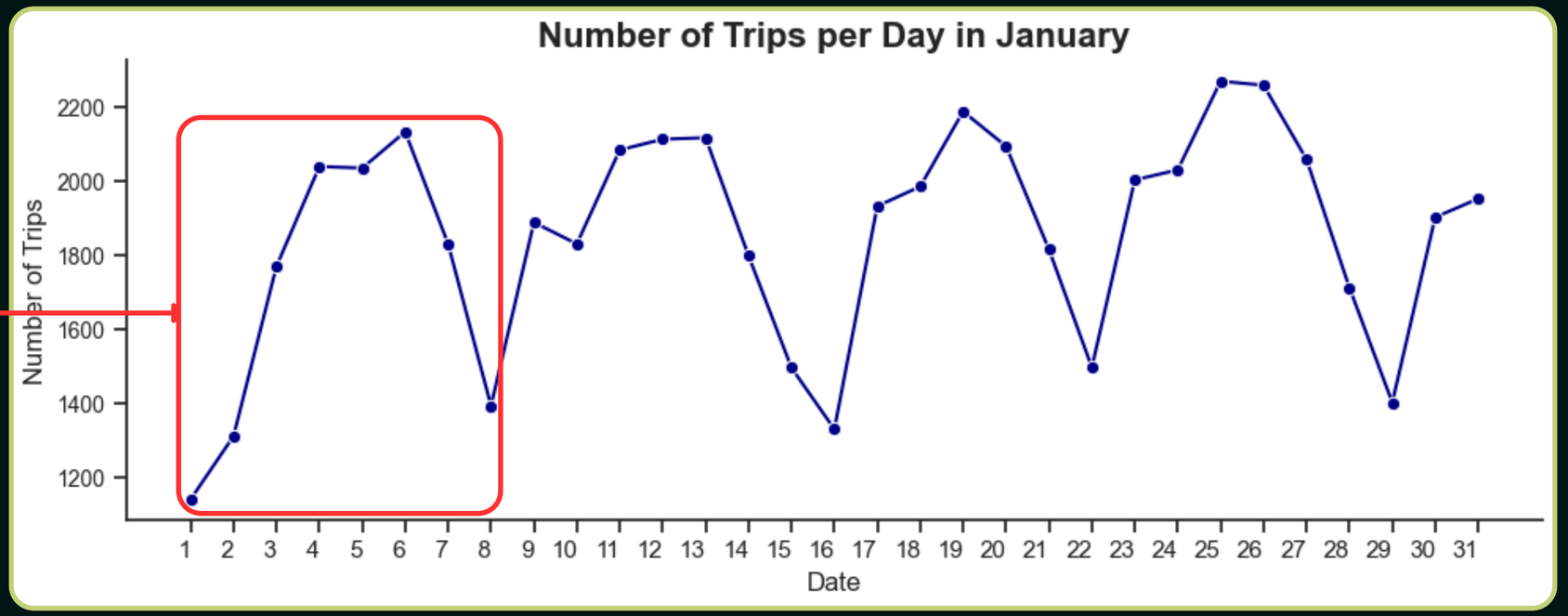
- **trip_duration_min** - time difference in minutes
- **speed_mph** - derived using trip_distance/trip_duration_min
- **fare_per_mile** - total fare normalized by trip distance

## External Sources

- **pickup_borough**, **dropoff_borough** - mapped using NYC Taxi Zone
- **pickup_zone, dropoff_zone**
- **pickup_latitude, pickup_longitude, dropoff_latitude, dropoff_longitude**
- **precipitation, snowfall, snowdepth, avg_temp, max_temp, min_temp** - mapped using NYC Weather

# Exploratory Summary.

# Is trend and seasonality present?

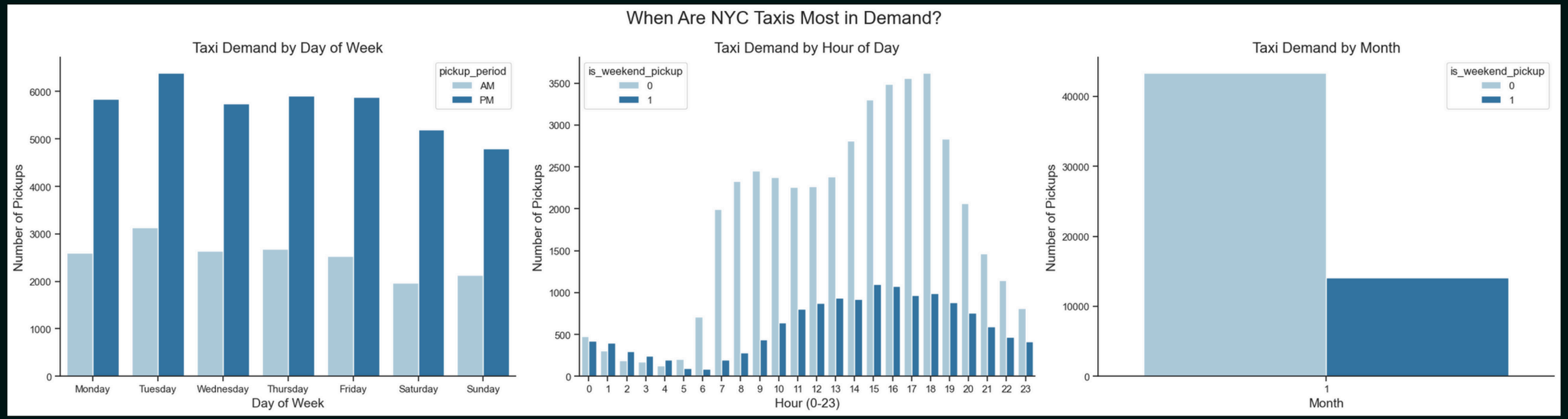



Weekly cycles are present highlighting predictable travel behavior.

Higher tip volumes are present on weekdays, especially mid-week, suggesting a strong commuting demand. Declines in trips are only present in weekends.

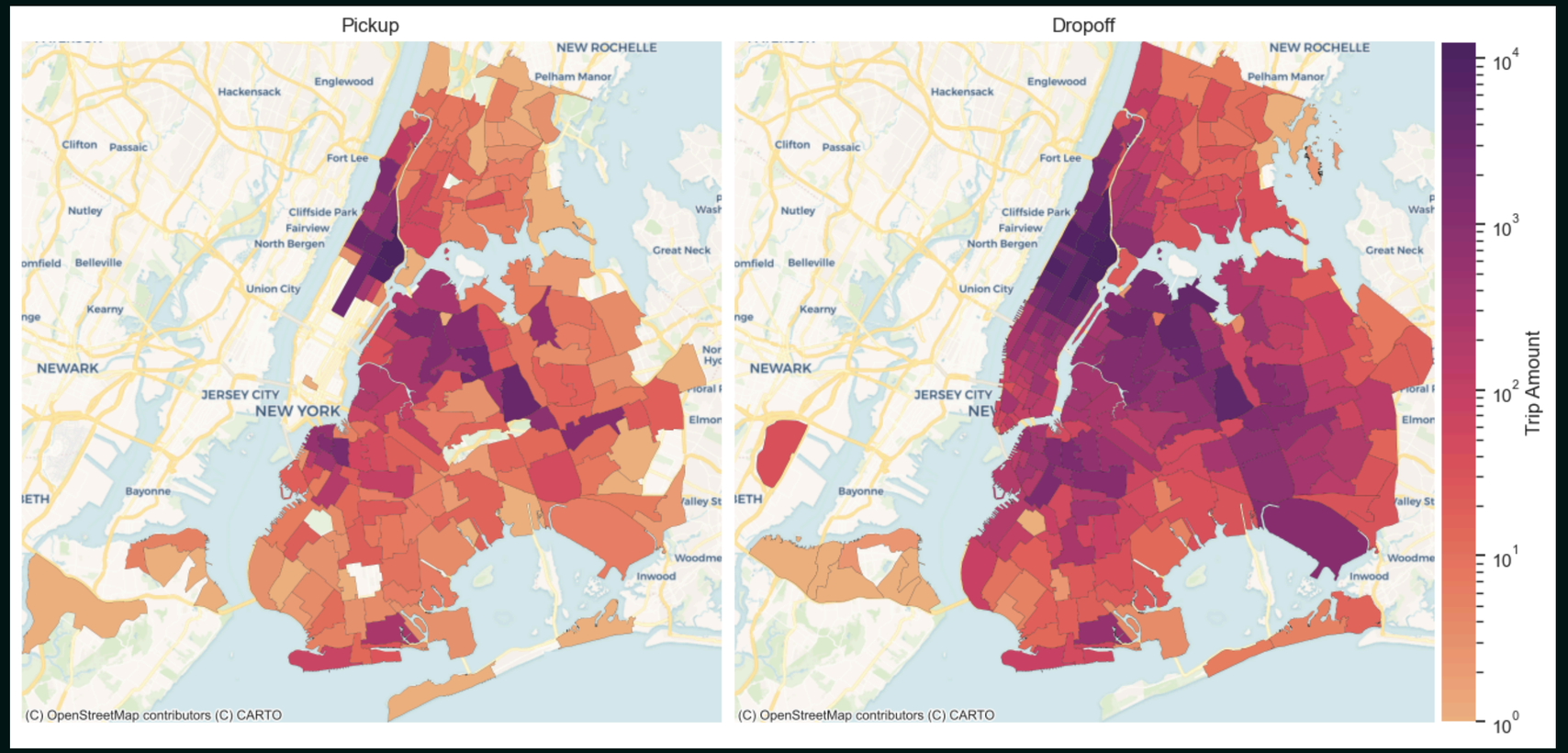
Increasing trend is also present, indicating resume of regular routines after New Year holiday.

# When and where are NYC taxis most in demand?



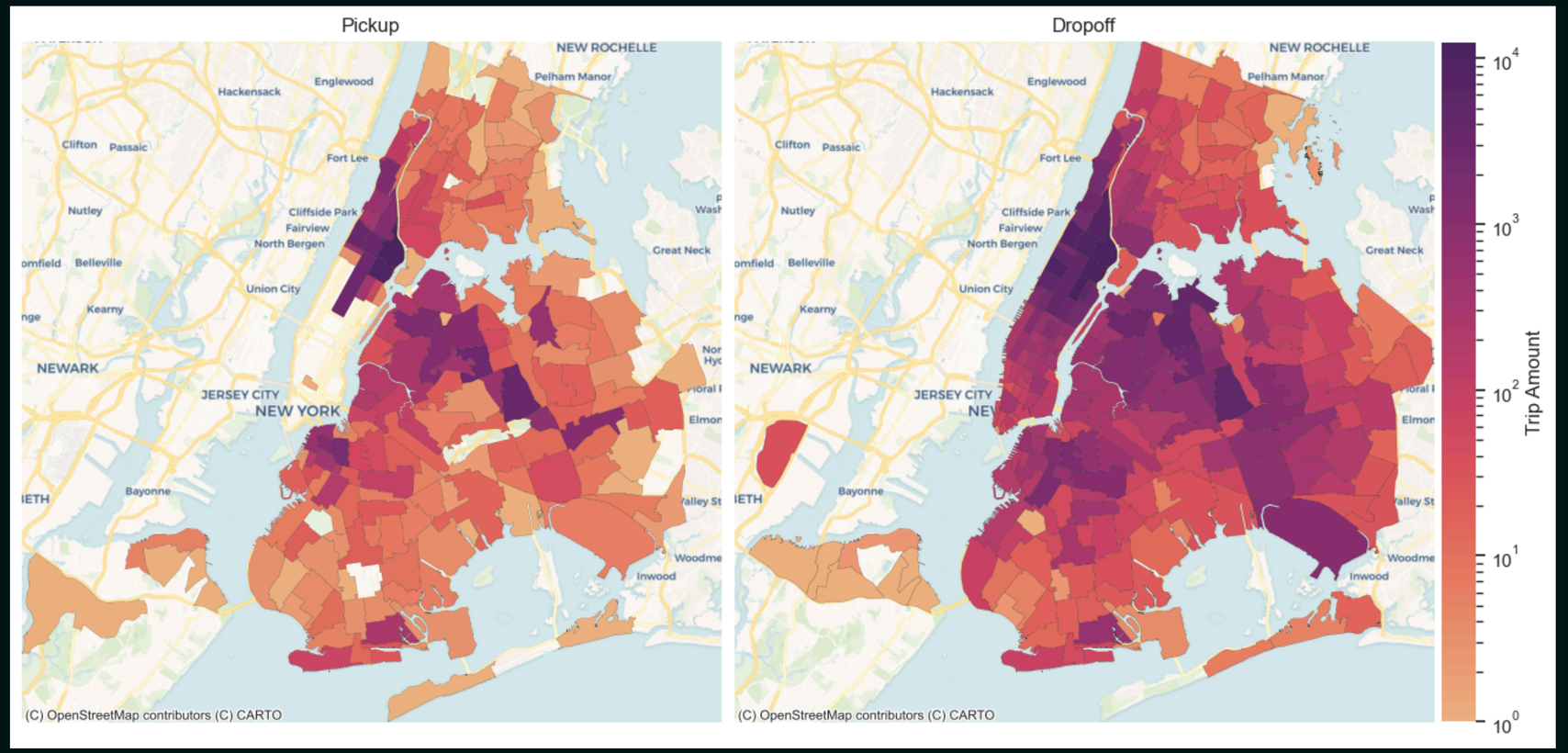

Taxi demand in NYC peaks on weekdays, especially midweek (Tuesday–Wednesday), and during traditional commuting hours (7–10 AM & 4–6 PM). Weekends show a significant drop in overall demand.

# When and where are NYC taxis most in demand?



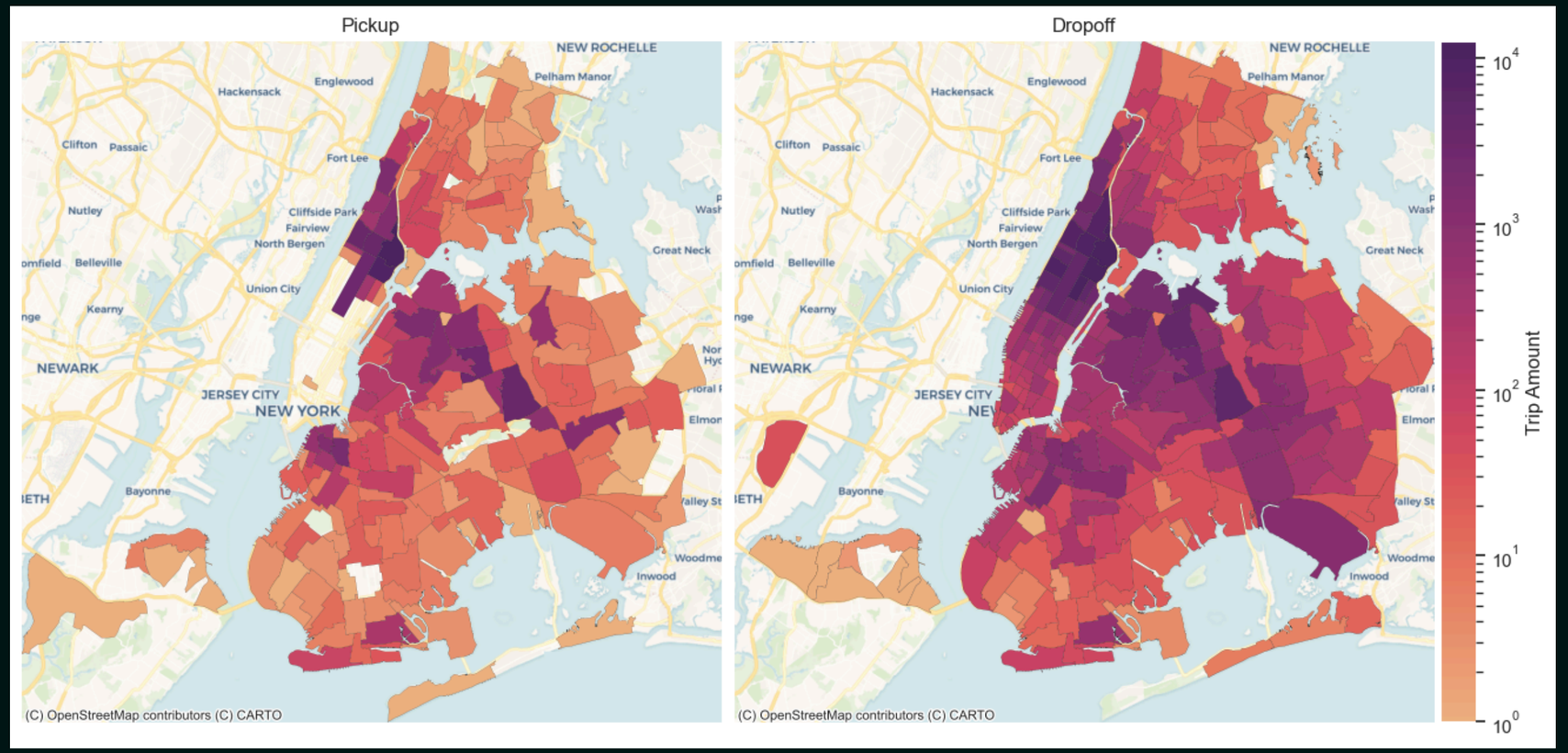

- Manhattan is top pickup hub, with peak activity in Midtown and the Upper East/West Sides.
- Brooklyn and Queens show moderate pickups while the Bronx and Staten Island have minimal activity.

# When and where are NYC taxis most in demand?



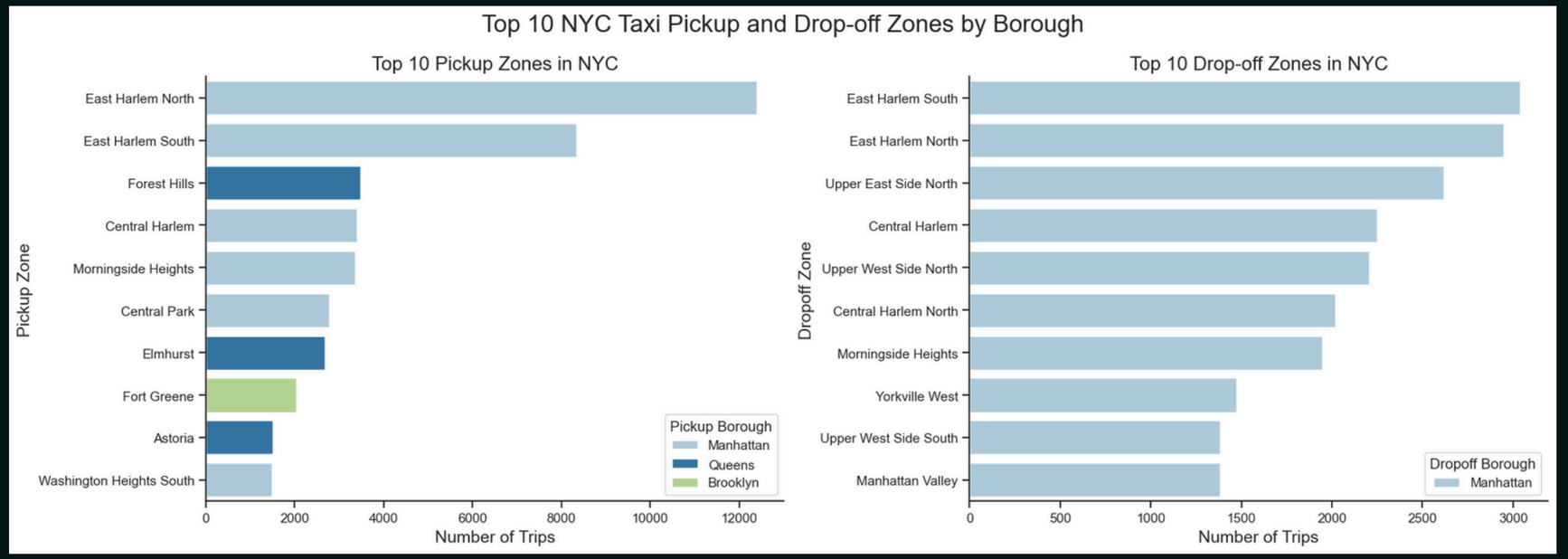

- Drop-offs are concentrated in Manhattan, particularly in Midtown, Lower Manhattan, and the Upper East/West Sides.
- Lower Manhattan is a significant drop-off zone, despite the pickup ban.
- Midtown sees a wide spread of drop-offs, reflecting its status as a hub for work, shopping, and entertainment.

# When and where are NYC taxis most in demand?



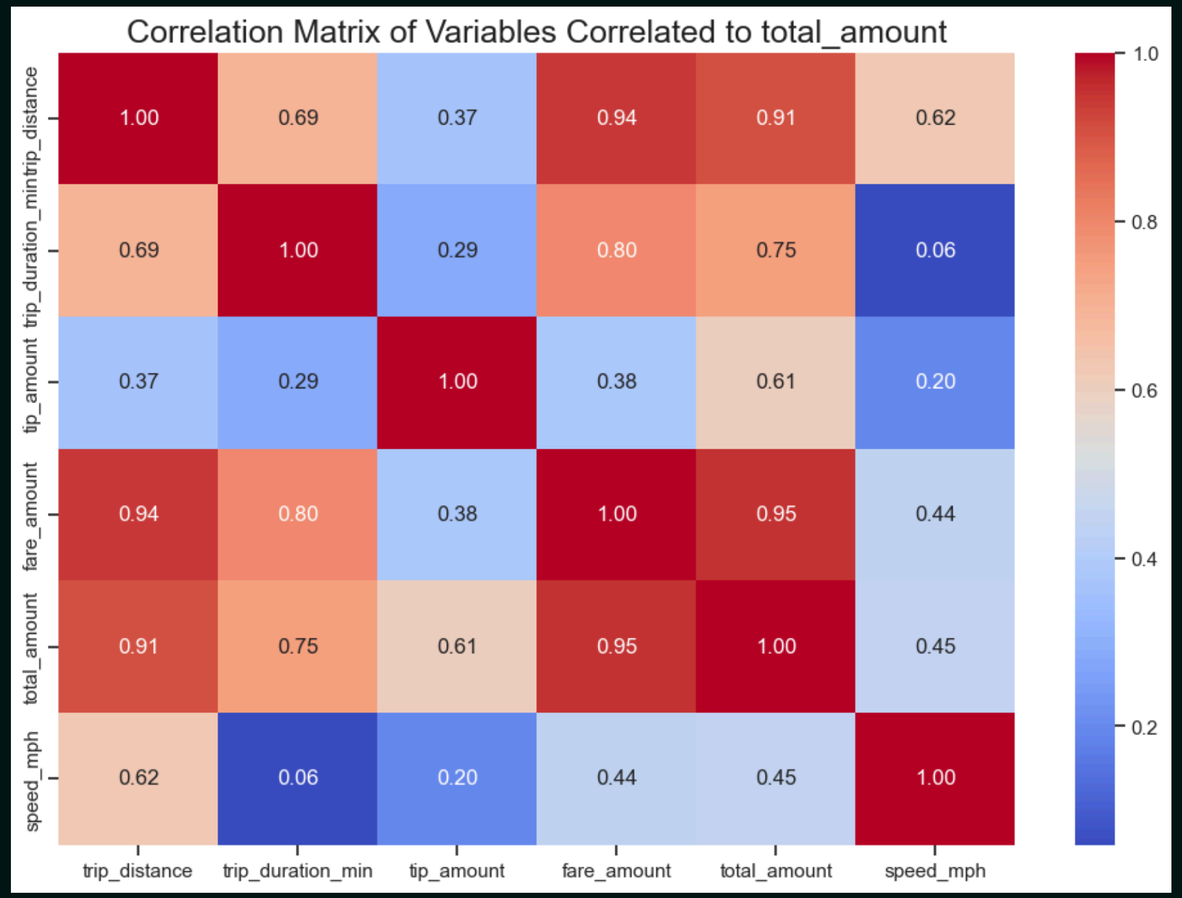

Sharp contrasts between Manhattan and outer boroughs underscore differences in accessibility and socioeconomic factors.

# When and where are NYC taxis most in demand?




Taxis in NYC are primarily used to get into and around Manhattan, especially from other boroughs like Queens and Brooklyn. East Harlem stands out as a key hub, and while pickups are more geographically diverse, drop-offs are largely concentrated in Manhattan’s residential and business corridors.

# What factors influence taxi fares?




**fare_amount (corr = 0.95)**

Fare is a major component of total_amount. This suggests that any changes in fare_amount will directly affect total_amount.

**trip_distance (corr = 0.91)**

This suggests that longer trips tend to gave higher total_amount. Longer trips incur higher fares due to the distance-based fare structure in NYC.

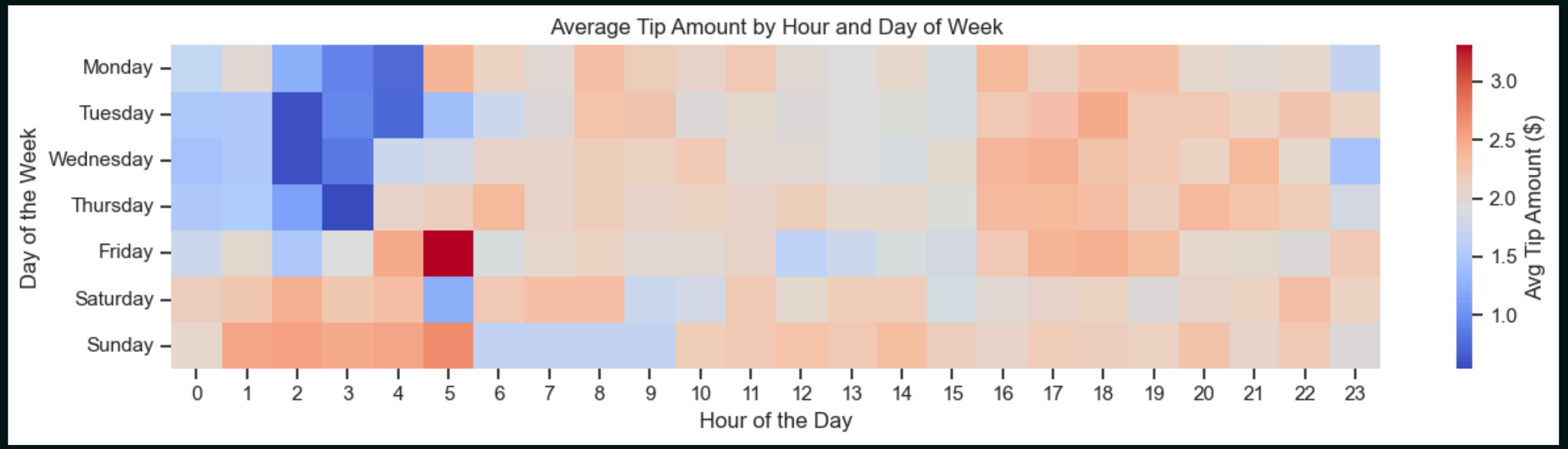
**trip_duration_min (corr = 0.61)**

Time spent on the trip is a contributing factor but not as strong as trip_distance or fare_amount.

**tip_amount (corr = 0.61)**

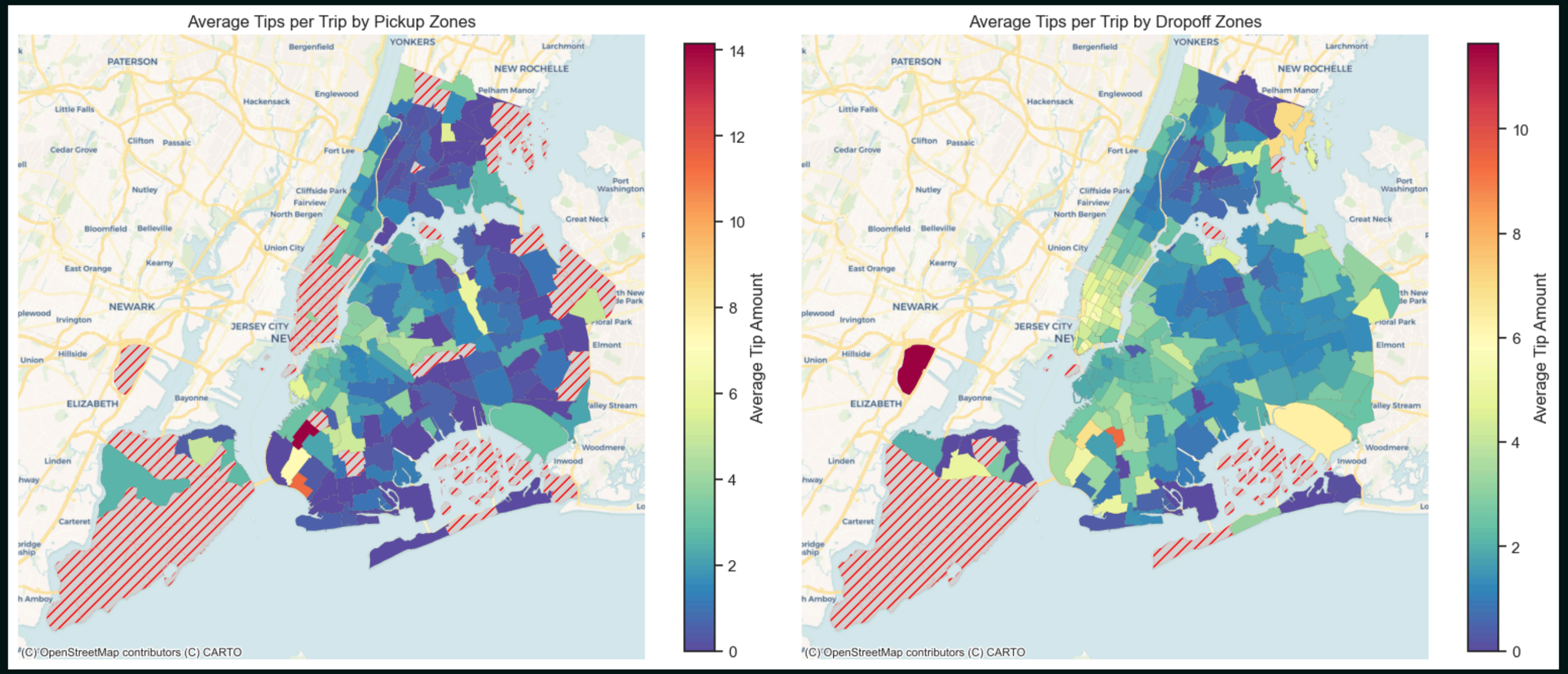
This somehow suggests that higher tip amounts are somewhat linked to higher total fares.

# How do passenger tipping behaviors vary by trip characteristics and location?

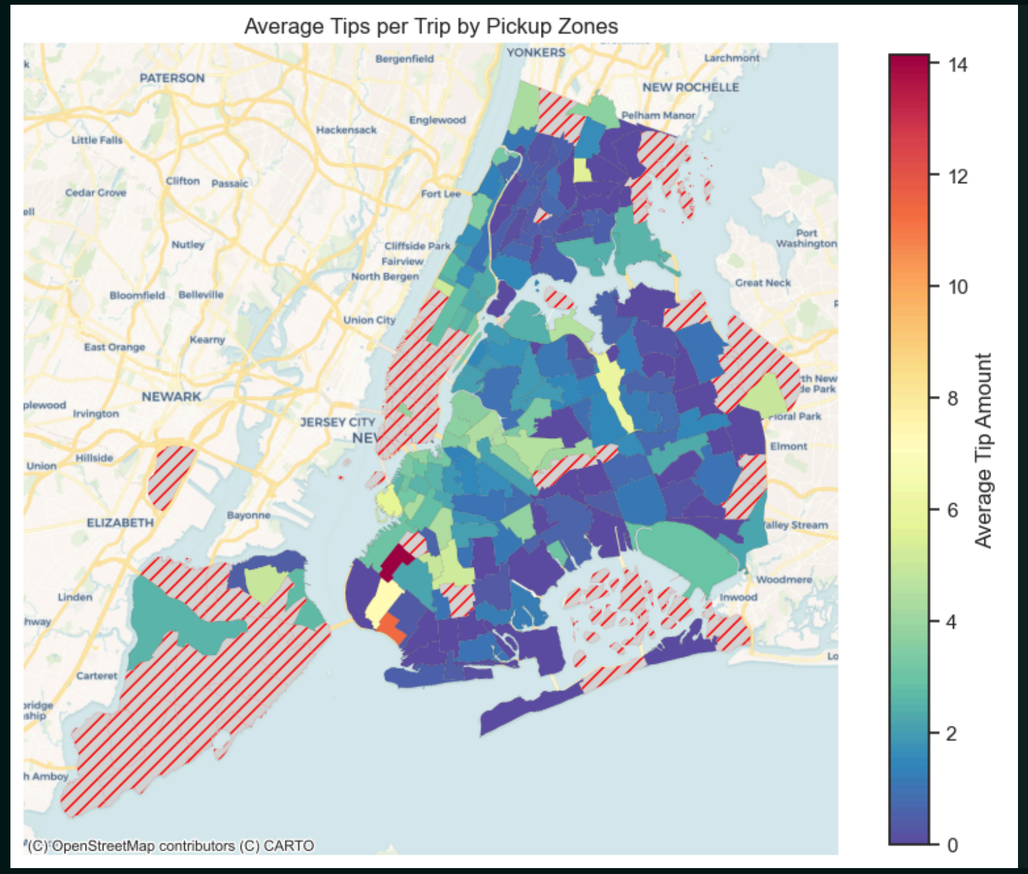


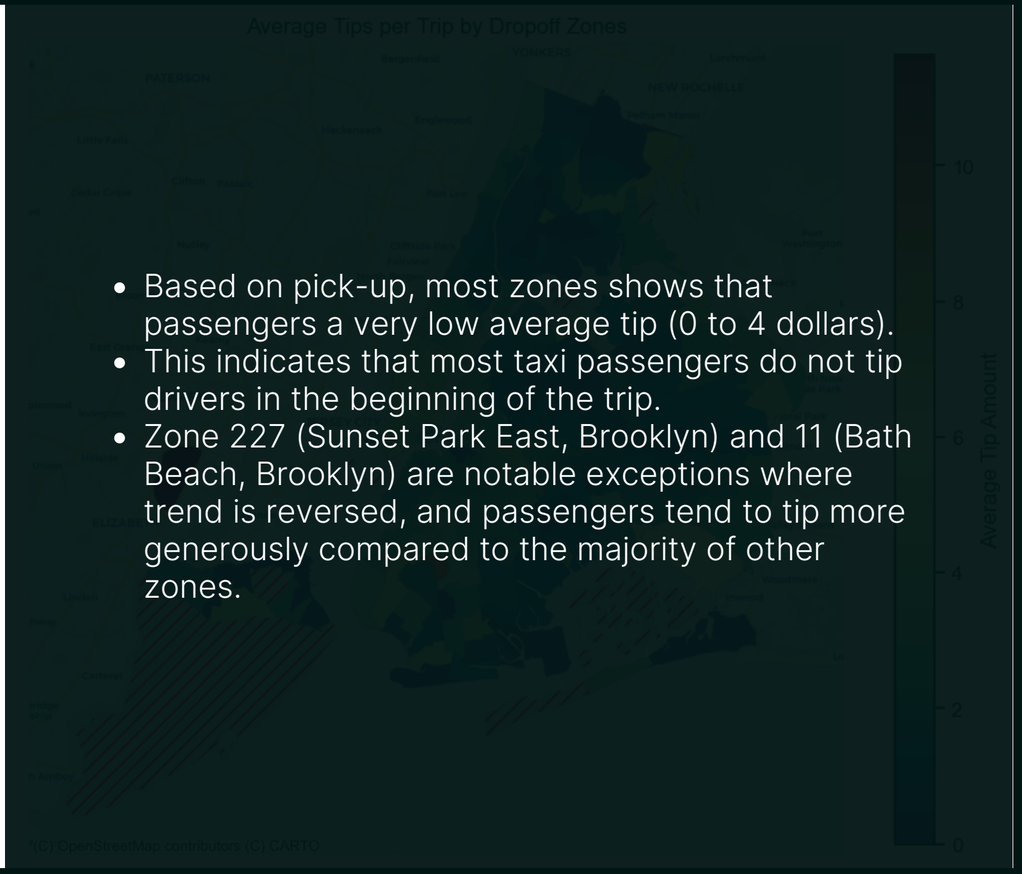

Presentations are communication tools that can be used as demonstrations, lectures, speeches, reports, and more. It is mostly presented before an audience. It serves a variety of purposes, making presentations powerful tools for convincing and teaching.Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

# How do passenger tipping behaviors vary by trip characteristics and location?



# How do passenger tipping behaviors vary by trip characteristics and location?

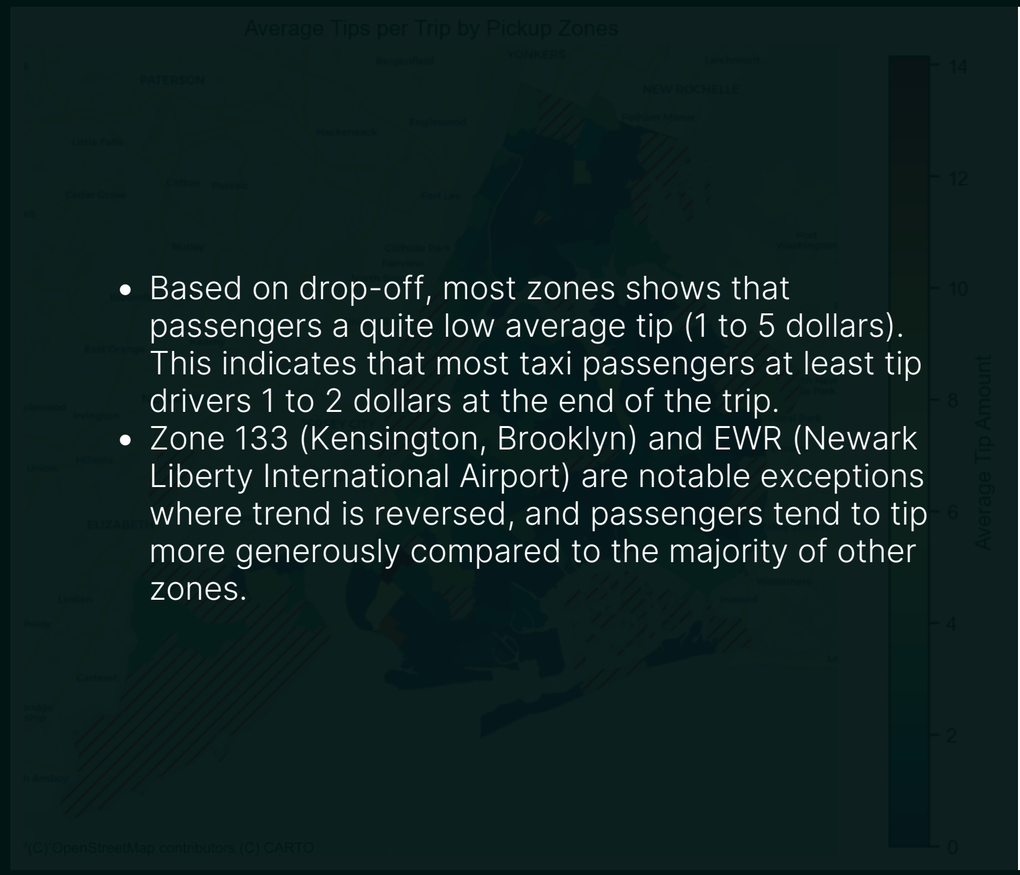


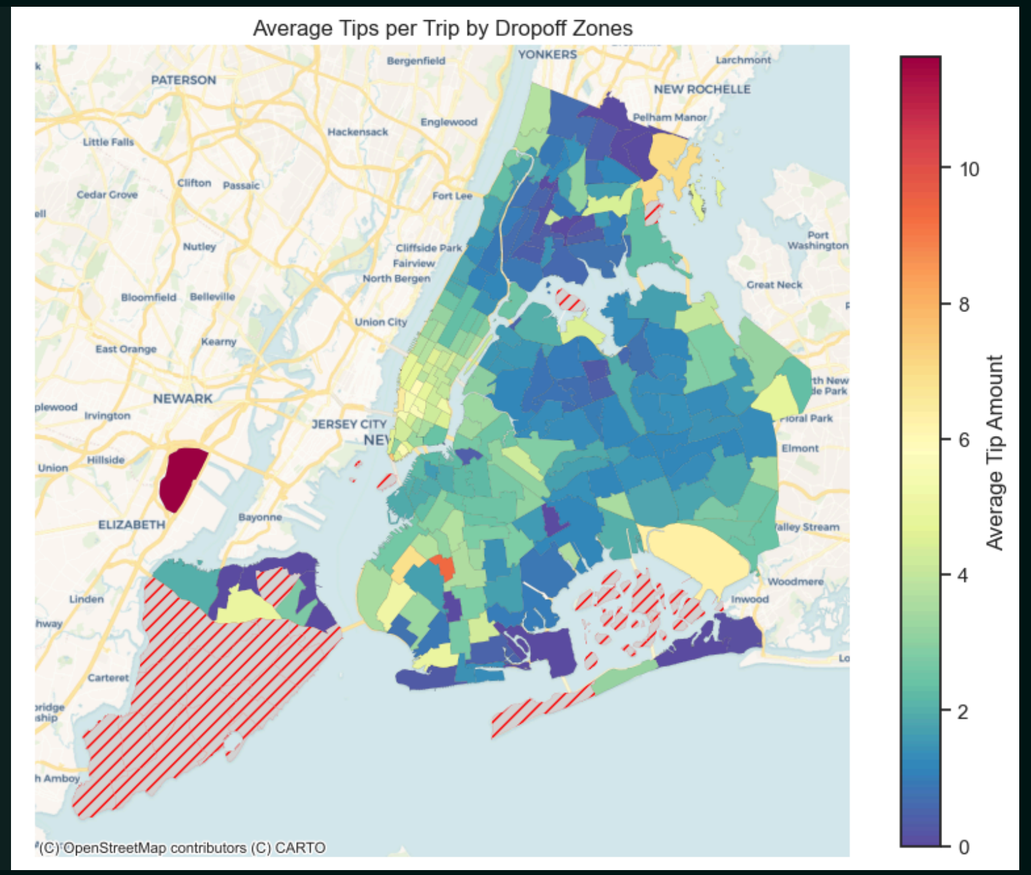




- Based on pick-up, most zones shows that passengers a very low average tip (0 to 4 dollars).
- This indicates that most taxi passengers do not tip drivers in the beginning of the trip.
- Zone 227 (Sunset Park East, Brooklyn) and 11 (Bath Beach, Brooklyn) are notable exceptions where trend is reversed, and passengers tend to tip more generously compared to the majority of other zones.

# How do passenger tipping behaviors vary by trip characteristics and location?

- Based on drop-off, most zones shows that passengers a quite low average tip (1 to 5 dollars). This indicates that most taxi passengers at least tip drivers 1 to 2 dollars at the end of the trip.
- Zone 133 (Kensington, Brooklyn) and EWR (Newark Liberty International Airport) are notable exceptions where trend is reversed, and passengers tend to tip more generously compared to the majority of other zones.



# Conclusion

## Trip Volume Trends

- Trips peaked on Fridays and Saturdays, especially during evening hours, reflecting strong weekend and nightlife activity.
- Daily trends showed a weekday-work pattern with moderate morning and evening peaks, indicating commuter behavior.

## Geospatial Hotspots

- Midtown and Downtown Manhattan are the dominant zones for both pickups and dropoffs.
- Zones with high commercial and tourist activity consistently record the highest trip counts.

## Fare Influencers

- Trip distance and fare amount have the highest correlation with the total fare, followed by trip duration and tip amount.
- These components are the most significant drivers of pricing variability.

## Tipping Behavior

- Higher tip amounts are seen during late evenings and weekends, and in zones frequented by tourists and wealthier residents.
- Tip percentages tend to be higher in central Manhattan and at pickup points near airports or high-end areas.

# Suggestions

## Optimize Driver Deployment

- Increase fleet presence during Friday–Saturday evenings, especially in Manhattan and nightlife hotspots.
- Assign more drivers to outer boroughs (e.g., Brooklyn, Queens) during late hours to accommodate longer trips.

## Dynamic Fare & Incentive Strategies

- Consider incentivizing drivers in low-demand zones to balance supply.
- Implement surge pricing adjustments during peak tipping hours (evenings, weekends) to improve profitability.

## Tipping Awareness Campaign

- Use app prompts or receipt messages to encourage tipping during off-peak times or low-tip zones.
- Educate passengers on average tipping behavior to promote fair tipping.

# Thank you.

Links:
1. To GitHub repository
2. To Tableau Dashboard