NYC TLC TAXI DATA ANALYSIS: PROFIT OPTIMIZATION AND OPERATIONAL EFFICIENCY

By: Tengku Arika Hazera

NYC Taxi & Limousine Commission

Table of Content

Background

What is TLC?

The New York City Taxi and Limousine Commission (TLC) is the city government agency responsible for regulating and licensing taxis, for-hire vehicles, limousines, commuter vans, and paratransit vehicles in New York City.

NYC's taxi system follows strict service zones:

- **Yellow taxis** can pick up passengers anywhere
- **Green Boro taxis** are limited to street hails in the outer boroughs and northern Manhattan.

Background

TLC Divisions

- Uniformed Service Bureau
- Finance and Administration
- IT (Information Technology)
- Legal Affairs
- Licensing and Standards
- Policy and Community Affairs
- Communications

Who are the stakeholders?

Problem Statement

- TLC operates across all five boroughs, but lacks a data-driven understanding of:
  - customer preferences
  - which zones and what time range generate the highest revenue
  - which areas contribute to operational inefficiencies.

- This raises the question: **How can TLC improve profit and operational efficiency by analysing the revenue, demand patterns, and customer preferences?**

Objectives

- To analyze NYC taxi trip data to understand how revenue, demand patterns, and customer preferences vary across time and geography.
- To identify opportunities for profit improvement and operational efficiency.

Executive Summary

Key findings:

1. Strong revenue concentration in Manhattan (revenue by borough)
2. Peak demand and high revenue hours → 4 PM - 6 PM
3. Zones show significant revenue inequality
4. Distance does not strongly determine revenue
5. Tip behavior varies by time
6. Statistical tests confirm significant relationships

NYC Taxi & Limousine Commission

Data Understanding

(Dataset 1: Main dataset), (Dataset 2: Complementary)

Data Cleaning

Data Transformation

Data Cleaning

Taxi & Limousine Commission

DATASET 1: VENDOR IDENTIFIER

- VendorID: A code indicating the LPEP provider that provided the record.

DATASET 1: LOCATION IDENTIFIER

- PULocationID: TLC Taxi Zone in which the taximeter was engaged.
- DOLocationID: TLC Taxi Zone in which the taximeter was disengaged.

DATASET 1: TIME IDENTIFIER

- lpep_pickup_datetime: The date and time when the meter was engaged.
- lpep_dropoff_datetime: The date and time when the meter was disengaged.

DATASET 1: TRIP IDENTIFIER

- passenger_count: The number of passengers in the vehicle.
- trip_distance: The elapsed trip distance in miles was reported by the taximeter.
- store_and_fwd_flag: This flag indicates whether the trip record was held in the vehicle memory before sending to the vendor
- trip_type: A code indicating whether the trip was a street hailor a dispatch.
- RateCodeID: The final rate code is in effect at the end of the trip.

Data Cleaning

Taxi & Limousine Commission

DATASET 1: PAYMENT IDENTIFIER

- payment_type: A numeric code signifying how the passenger paid for the trip.
- fare_amount: The time-and-distance fare is calculated by the meter.
- extra: Miscellaneous extras and surcharges.
- mta_tax: $0.50 MTA tax that is automatically triggered based on the metered rate in use.
- improvement_surcharge: $0.30 improvement surcharge assessed on hailed trips at the flag drop.
- tip_amount: This field is automatically populated for credit card tips.
- tolls_amount: The total amount of all tolls paid in the trip.
- ehail_fee: fees charged by the e-hail app provider.
- total_amount: The total amount charged to passengers.
- congestion_surcharge: A fee drivers charge to passengers because they need to pay to enter a specific, congested area.

DATASET 2 IDENTIFIER

- LocationID: The same ID used in PULocationID and DOLocationID.
- Borough: The borough which the LocationID is located in.
- Zone: The specific zone area that the taxis covered within the specific borough.
- service_zone: The type of taxis or vehicle that are available in a specific zone.

Data Cleaning

Taxi & Limousine Commission

- Dataset 1 has a total of 68,211 data with 93.1% null values. Meanwhile, dataset 2* has a total of 265 data with 99.62% null values.
- Based on the heatmap analysis, seven columns high correlations. However, the handling of the missing values differs.

DATA GROUP	COLUMN	FILLNA METHOD
Trip identifier	store_and_fwd_flag	Fillna with mode value
	RatecodeID	Fillna with mode value
	passenger_count	Fillna with median value
	trip_type	Fillna with mode value
Payment identifier	payment_type	Fillna with mode value
	congestion_charge	Fillna with median value
	ehail_fee	Dropna

*Data cleaning is not demonstrated here.

Data Transformation

Taxi & Limousine Commission

COLUMN NAME	TRANSFORMATION FORMULATION	TRANSFORMED DATA
lpep_pickup_datetime, lpep_dropoff_datetime	The date of trip → convert to day name The time of trip → convert to time range	day_name hour
tip_amount, fare_amount	Calculating the tip percentage per trip	tip_percent
Borough	Joining data LocationID to PULocationID and DOLocationID to generate borough column	PU_borough, DO_Borough
Zone	Joining data LocationID to PULocationID and DOLocationID to generate zone column	PU_Zone, DO_Zone

Final Dataset

Taxi & Limousine Commission

VENDOR IDENTIFIER

- VendorID:

LOCATION IDENTIFIER

- PULocationID
- ***PU_Borough***
- ***PU_Zone***
- DOLocationID
- ***DO_Borough***
- ***DO_Zone***

TIME IDENTIFIER

- lpep_pickup_datetime
- lpep_dropoff_datetime
- ***day_name***
- ***hour***

TRIP IDENTIFIER

- passenger_count
- trip_distance
- store_and_fwd_flag
- trip_type
- RateCodeID

PAYMENT IDENTIFIER

- payment_type
- fare_amount
- extra
- mta_tax
- improvement_surcharge
- tip_amount
- tolls_amount
- total_amount
- congestion_surcharge
- ***tip_percent***

Data Analysis

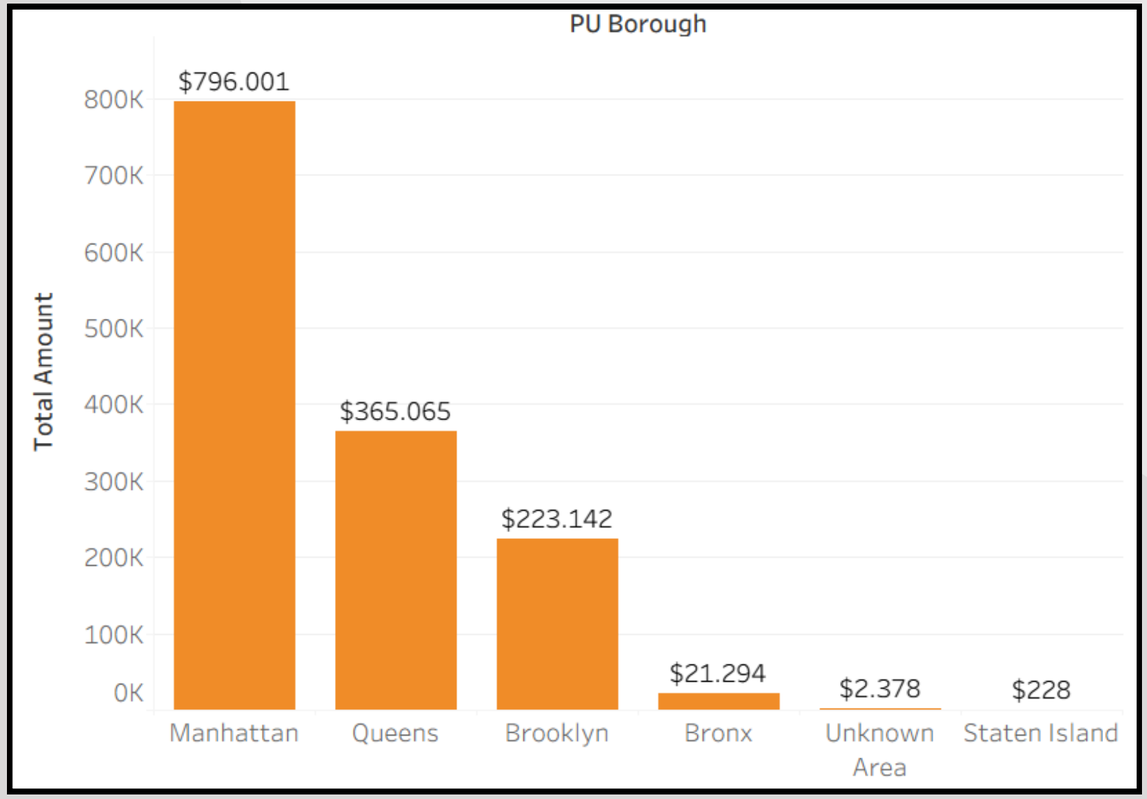
NYC Taxi & Limousine Commission

KPI Overview

Total Revenue	Total Trips	Average Revenue Per Trip	Average Trip Distance
$1.408.108	64.698	$21,8	2,81 Miles

- The analysis shows that the dataset contains **64,698 total trips**, generating a total revenue of approximately **$1.41 million**.
- On average, each trip contributes about **$21.8** in revenue, indicating strong earning efficiency per ride.
- The average trip distance is **2.81 miles**, suggesting that most trips are relatively short, typical of dense urban travel patterns.

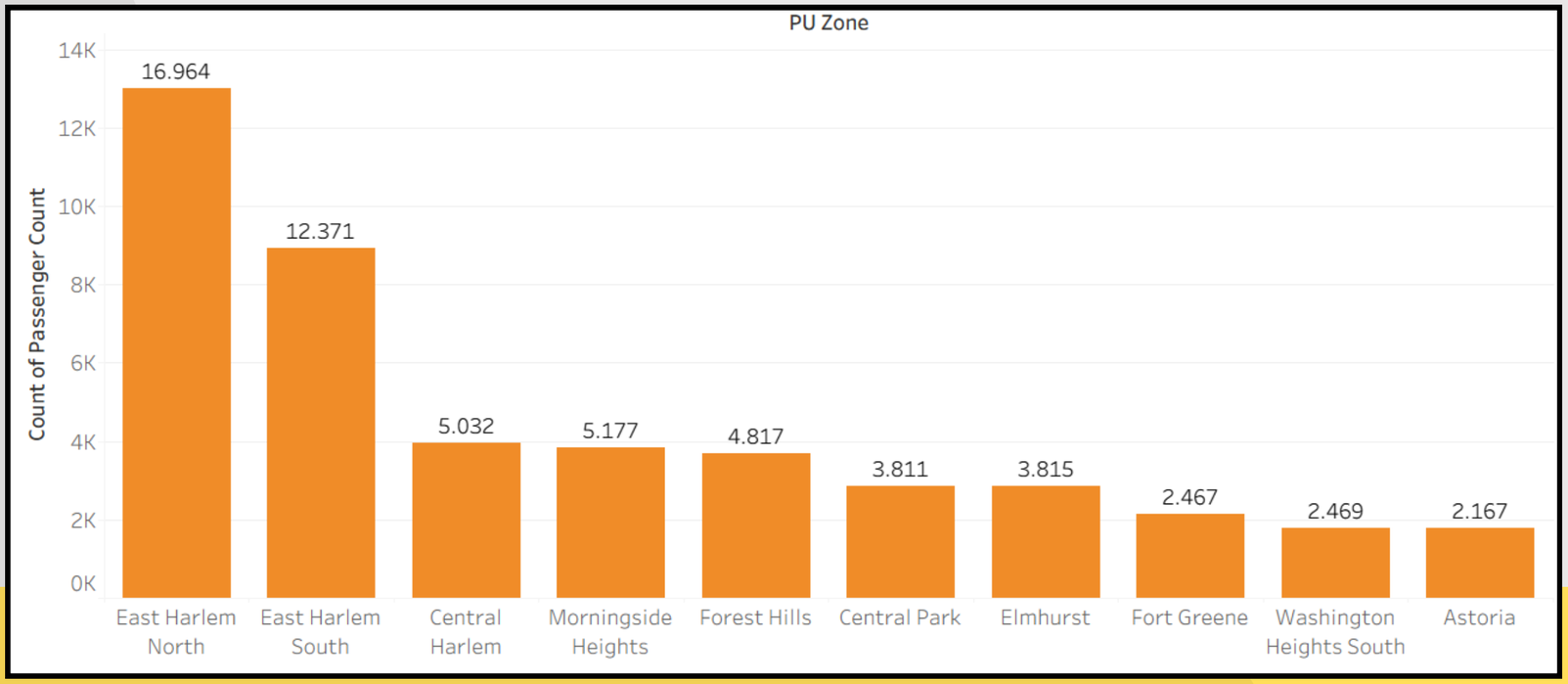
Revenue by Borough



Insights:

- Manhattan produces the highest revenue by a large margin
  - Two times higher than the revenue in Queens
- Revenue from “Unknown Area” includes the zone from outside of NYC

Top 10 Taxi Demand Zone

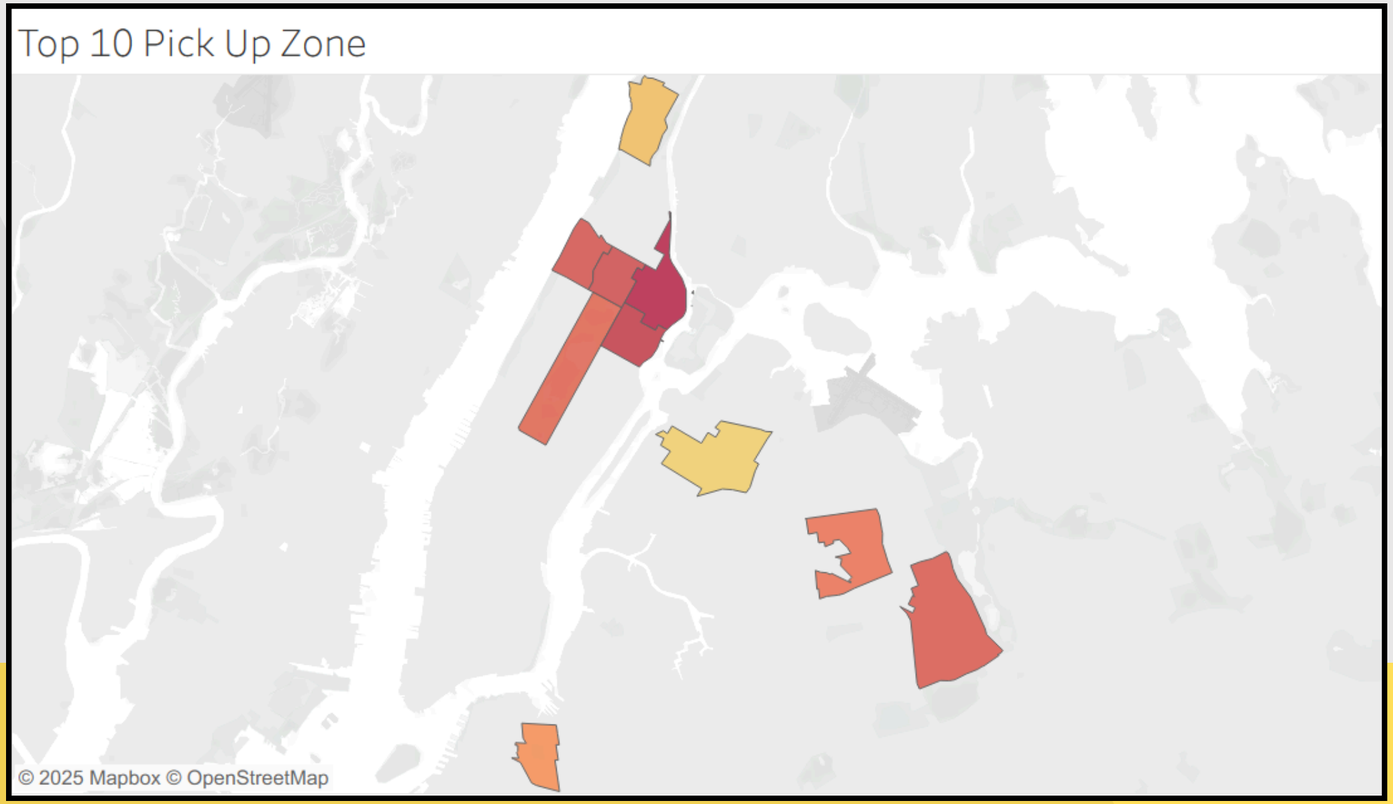


Insights:

- Six of these zones are located in Manhattan:
  - East Harlem North, East Harlem South, Central Harlem, Morningside Heights, Central Park, and Washington Heights South

Top 10 Taxi Demand Zone

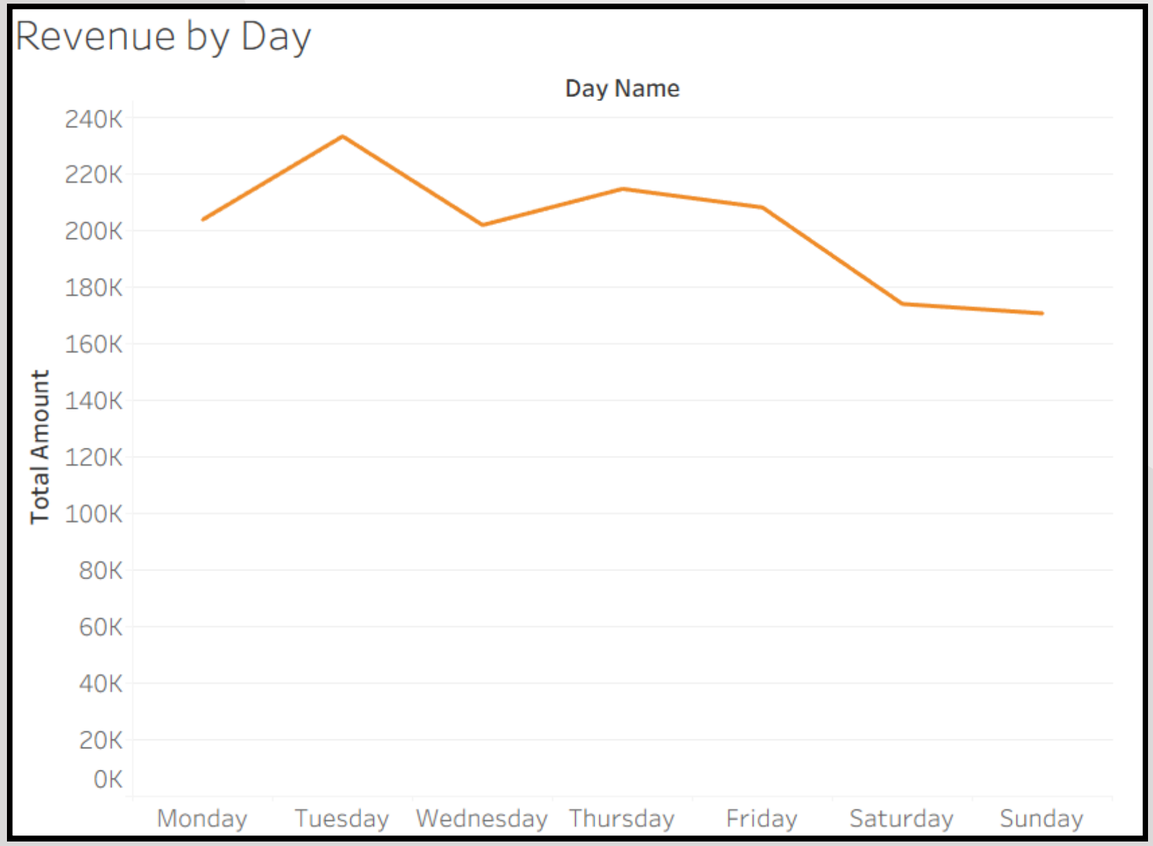
Taxi & Limousine Commission



Insights:

- Six of these zones are located in Manhattan:
  - East Harlem North, East Harlem South, Central Harlem, Morningside Heights, Central Park, and Washington Heights South

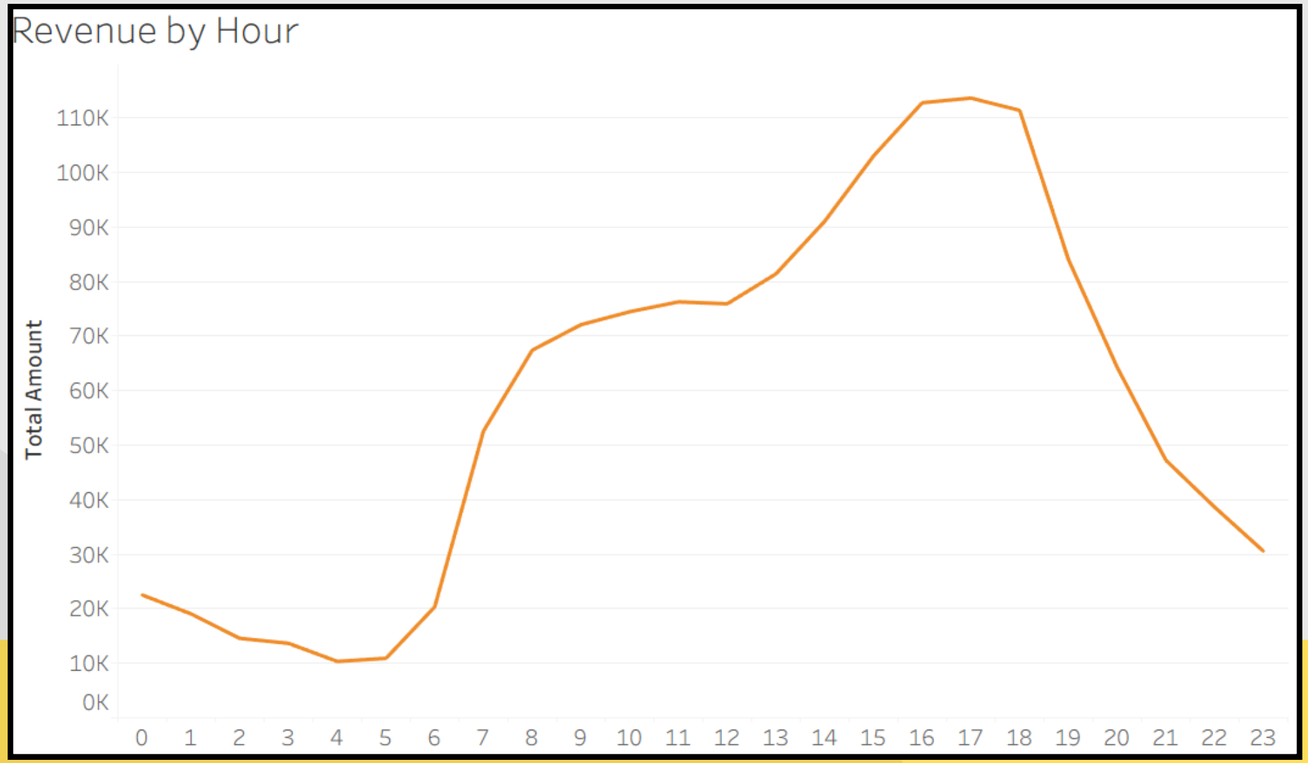
Daily Revenue Trend



Insights:

- The highest revenue is generated on Tuesday
- The revenue on weekends is surprisingly lower than on weekdays

Hourly Revenue Trend

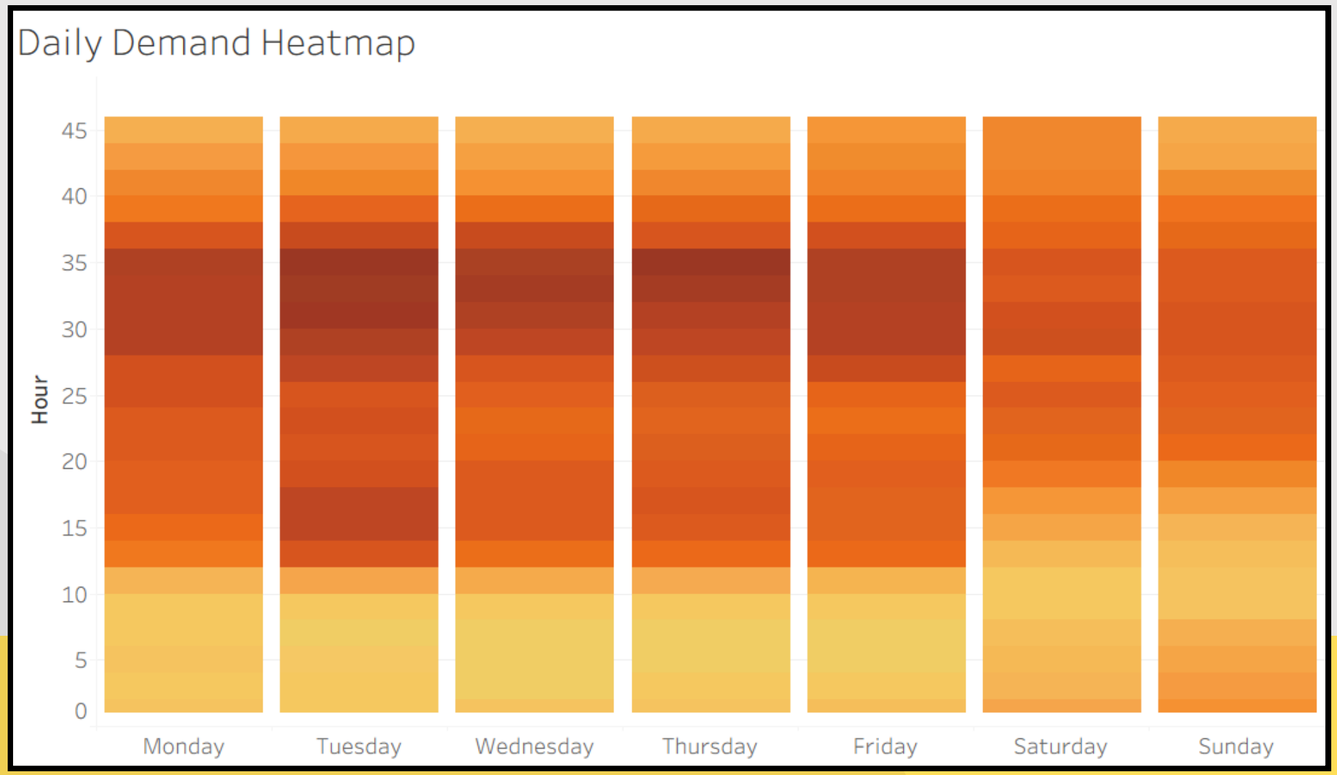


Insights:

- High revenue hours: **4 PM – 6 PM** (evening rush)
- Lower revenue after 12 AM
- This insight is useful for operational shifts and pricing signals

Daily Demand Heatmap

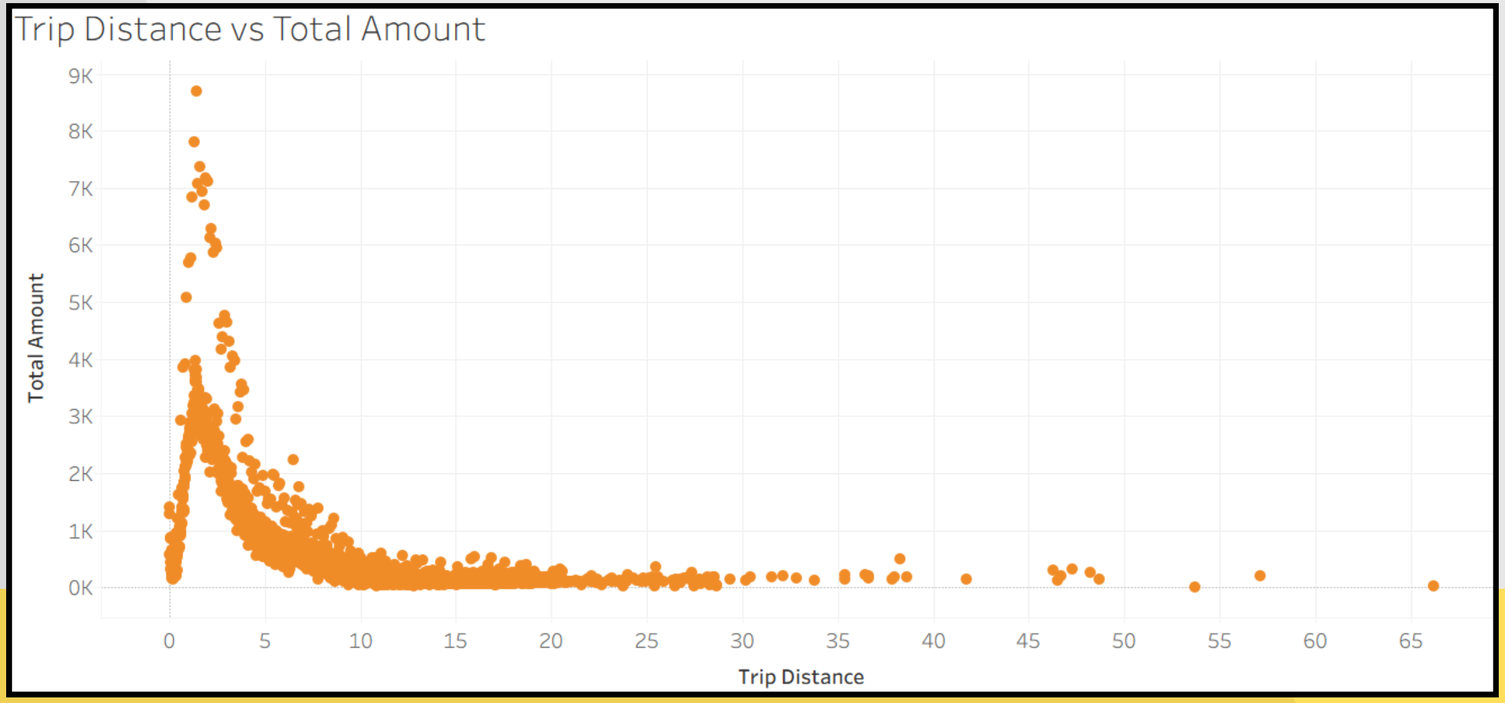
Taxi & Limousine Commission



Insights:

- Evening peak consistently shows that the highest taxi demand
- This insight can be used to optimize taxi allocation based on the time

Trip Distance vs Total Revenue



Insights:

- There is a **weak** correlation between distance and total amount
- However, it illustrates how the majority of profitable trips are **short-to-medium** distances

Recommendations

Recommendations

1

Prioritize resource allocation to high-revenue boroughs and zones

What TLC can do:

- Increase driver availability and dispatch efficiency in Manhattan and key Queens zones
- use borough-level revenue metrics to guide fleet distribution
- develop zone-specific policies to support high-demand areas.

Recommendations

Implement hour-based operational optimization

What TLC can do:

- Reinforce driver supply during peak revenue hours (rush hours and evenings)
- Consider incentive programs (driver bonuses, reduced idle time) for peak periods
- Reduce operational costs during low-demand hours (for example, shift adjustments)

Recommendations

3

Enhance driver efficiency through data-informed routing

What TLC can do:

- Provide drivers with real-time information about high-demand zones
- Reduce empty miles by strategically positioning available drivers before peaks
- Explore AI-driven dispatch routing to minimize fuel and time costs

Thank You