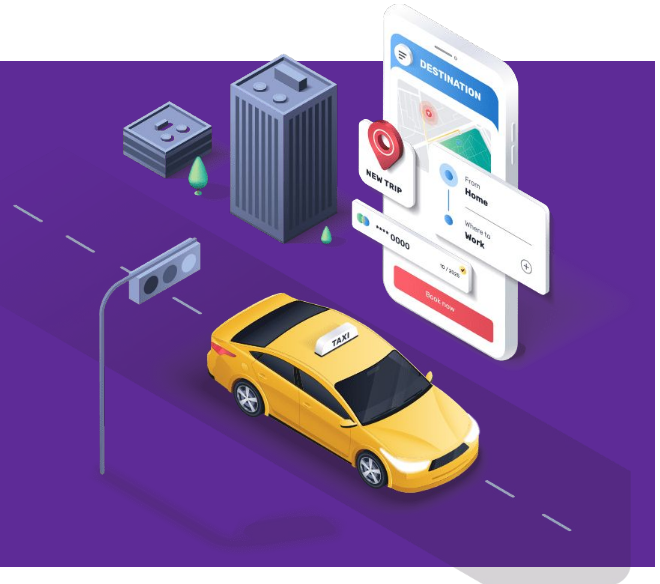
# Taxi Fleet Management
## using Clustering

Kahan Sheth, Nhat Pham, Vignesh Sankar



# Dataset

- **1.6M taxi trips** in Chicago (September-November 2023)
- Queried from GBQ, provided by the City of Chicago
- Features: pick-up & drop-off time, location, fare, payment
- Data prep: derive time attributes, outlier removal, meter error flags, area naming

| trip_id | start_time | end_time | period_start | duration | distance | pickup_area | dropoff_area | fare | tips | pickup_lat | pickup_lon | dropoff_lat | dropoff_lon | is_error |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| f8b933d75 | 2023-10-06 09:30:00 | 2023-10-06 09:30:00 | Morning Rush | 639.0 | 1.41 | Near North Side | Near North Side | 8.00 | 0.00 | 41.909496 | -87.630964 | 41.895033 | -87.619711 | False |
| 1f5599cc5 | 2023-10-20 11:30:00 | 2023-10-20 11:30:00 | Midday | 509.0 | 1.66 | Near North Side | Loop | 7.75 | 2.00 | 41.900221 | -87.629105 | 41.880994 | -87.632746 | False |

- Limitations: geospatial data restricted to centroids of neighborhoods, no data for areas outside city boundaries, no data on cash tips

# Problem Statement

As head of a taxi company in Chicago, we want to **optimize fleet distribution and taxi performance** to boost efficiency and profitability

**Challenge 1: Align fleet deployment with citywide demand**

➡ Cluster on spatial and temporal trip data to identify high-demand hotspots and patterns

➡ Enable strategic fleet positioning, reduce response times, enhance customer satisfaction

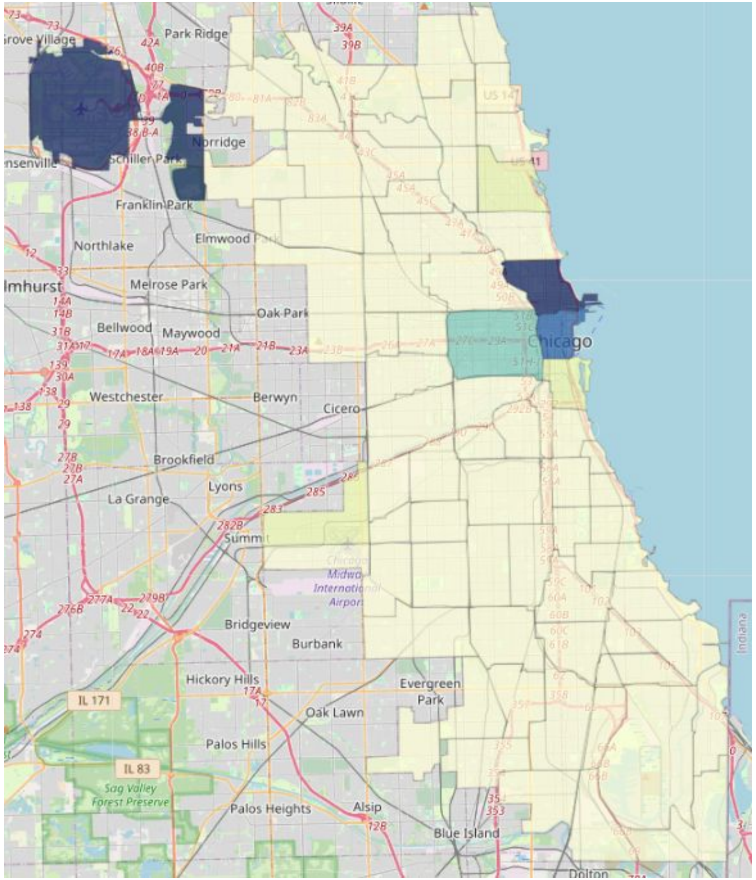
**Challenge 2: Evaluate and improve efficiency of the fleet**

➡ Cluster taxis based on performance metrics to distinguish efficiency profiles

➡ Enable interventions for low performers and adoption of best practices, enhance service quality and optimize costs

# EDA

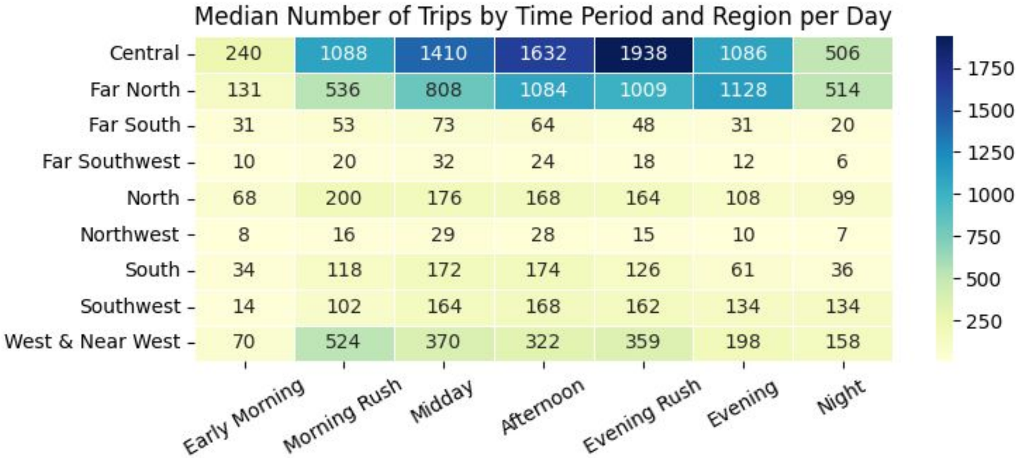
Two dominant hotspots: Downtown and O'Hare airport



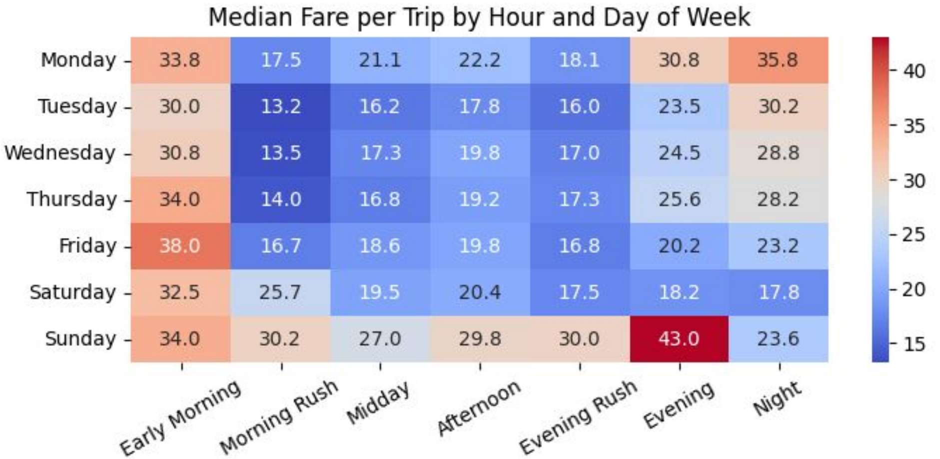

Total trip count by pickup area

# EDA

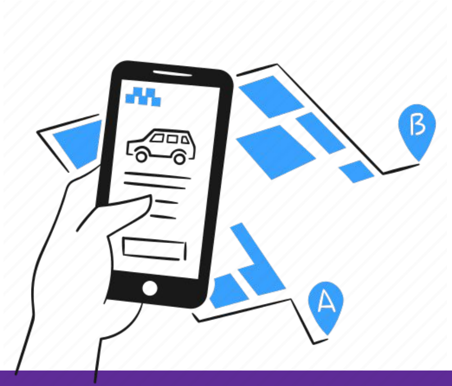
Central and Far North regions are busiest during business hours




Higher median fares in early mornings, late nights, and Sundays



# Trip Pattern Clustering



# Key Features

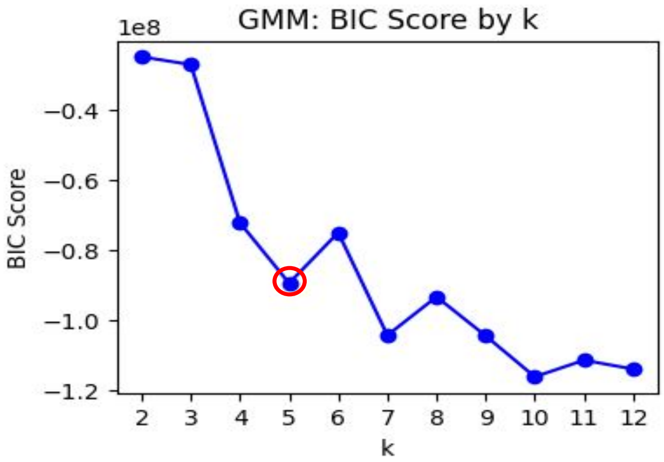
- **Pickup and dropoff coordinates:** capture the exact locations of service usage, crucial for understanding spatial patterns
- **Period Start:** provides insights into the temporal patterns of trips
- **Is Weekend:** distinguishes between weekdays and weekends, helping to understand variations in demand
- **Trip Total:** helps analyze fare-based patterns, which can be indicative of trip length, route popularity, or time of travel.

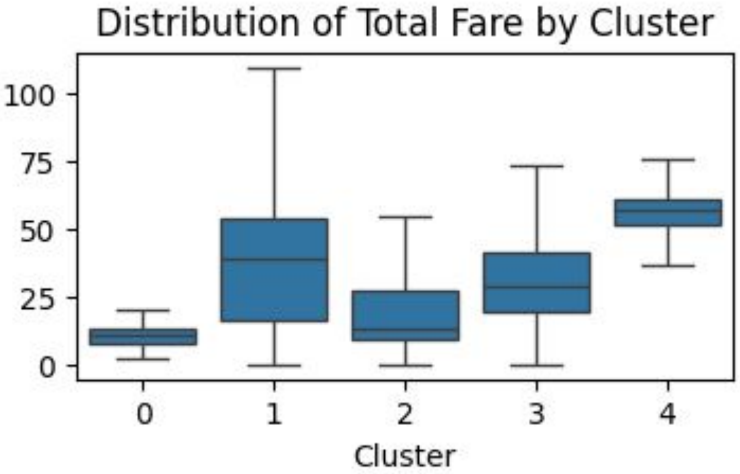
# Data Processing and Model Selection

- Geospatial data points were not scaled in order to maintain their original representation.
- Categorical features were one hot encoded
- Gaussian Mixture Model (GMM) and KMeans were chosen over HDBSCAN and hierarchical clustering due to their superior computational efficiency.
- Baseline model: K-means with k=3. However, this model resulted in indistinct clusters, unbalanced clusters.

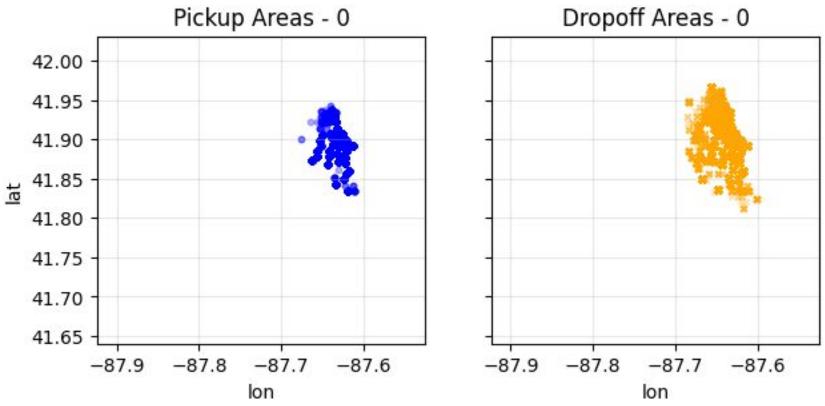
# Gaussian mixture model (GMM)

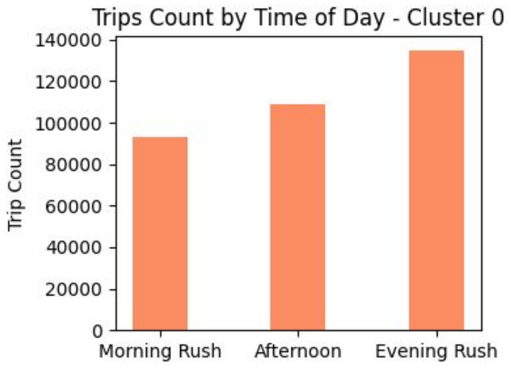
- Optimal k = 5 (BIC & testing), tested different initializations
- Noticeable variation of fare patterns can be seen across different clusters



# Cluster 0: Central City Routes

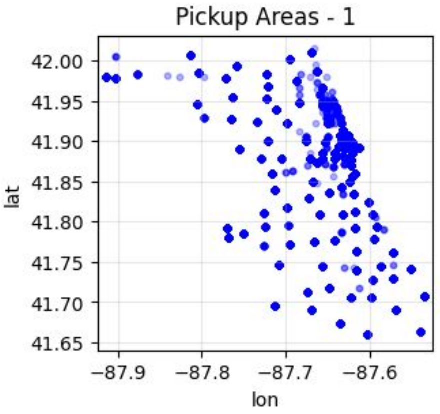


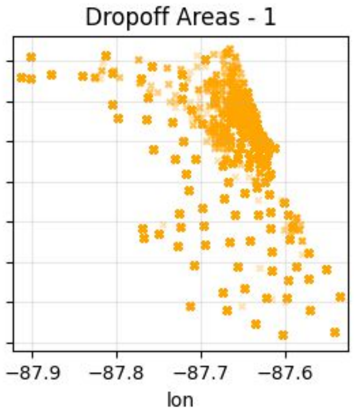



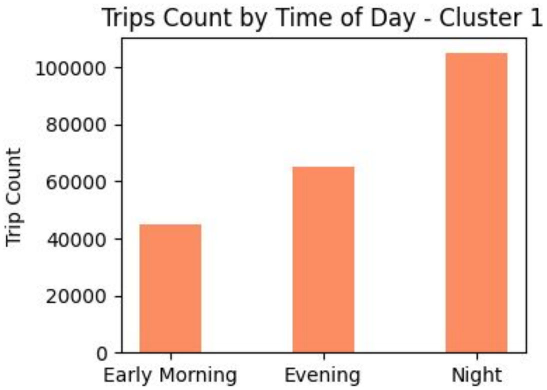

- Exclusive **downtown** pickups and dropoffs
- Peak activity during **office hours**
- Uniformly lower fare => **Short trips**

- Offer subscription services
- Optimize taxi availability
- Targeted advertising, special offers, and loyalty programs

# Cluster 1: Off-peak Urban and Airport Trips



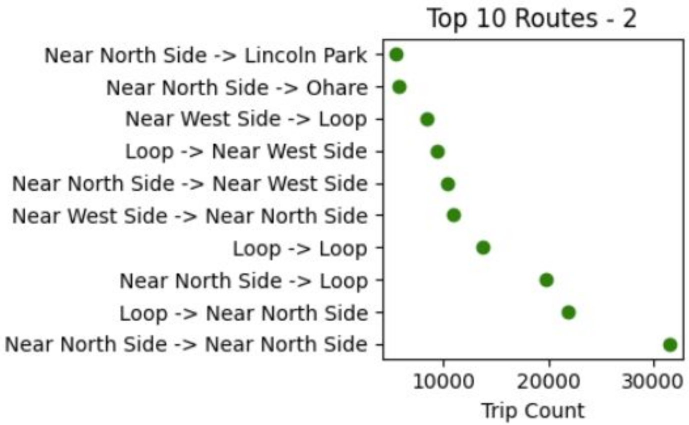




- **City-wide** coverage, urban routes and airport commutes
- **Off-peak hours**: early morning & late night
- Wide fare range, high median => **diverse trip lengths**

- Collaborate with businesses such as hotels, airlines, and night venues.
- Potential for dynamic pricing.

## Cluster 2: Non-commute Urban Travel

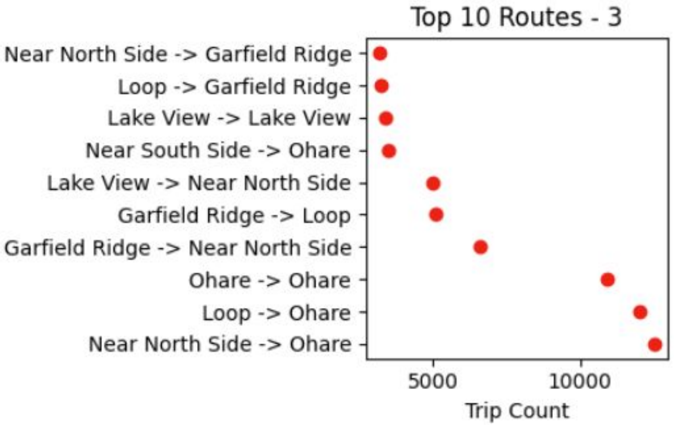



- Common routes **to and from downtown**, short trips (low fare)
- Midday and evening hours

⬇

- Likely **non-commute travel** (leisure or errands)
- Targeted ads for urban activities (shopping/ leisure outings)

## Cluster 3: Mixed Airport and Urban Trips

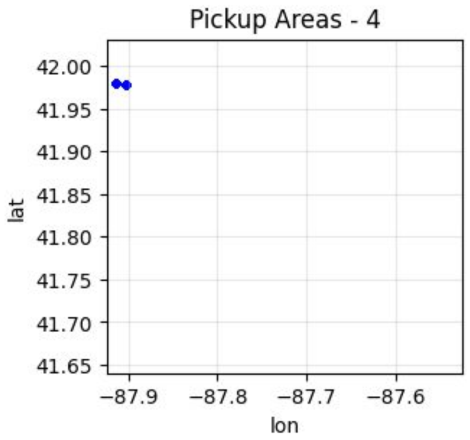


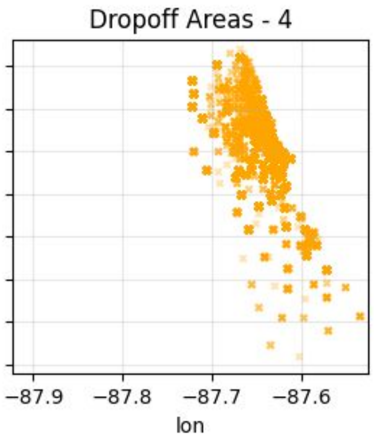

- Mixed distance trips
- Common routes between **downtown** & **airports (O'Hare, Garfield Ridge)**
- Afternoon and rush hours

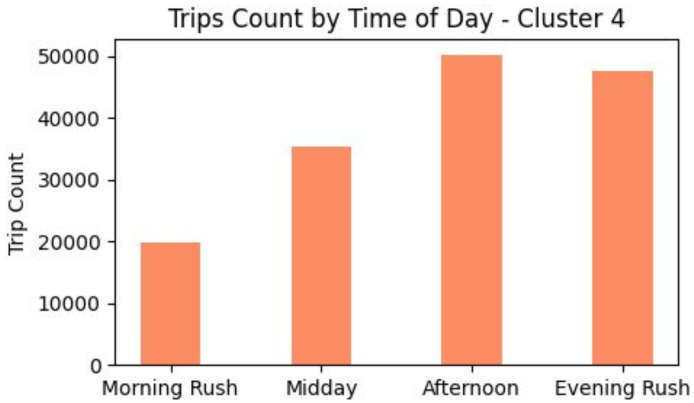
⬇

- Explore **Midway airport** as an underserved market
- Offer competitive pricing for airport trips

# Cluster 4: O’hare to City Trips




- Exclusive **Ohare pickups** to mainly downtown
- Higher fare => **Longer trips**
- High demand in afternoon and evening rush hour

- Ensure availability during peak times
- Offer competitive flat rates to compete with ride-sharing services

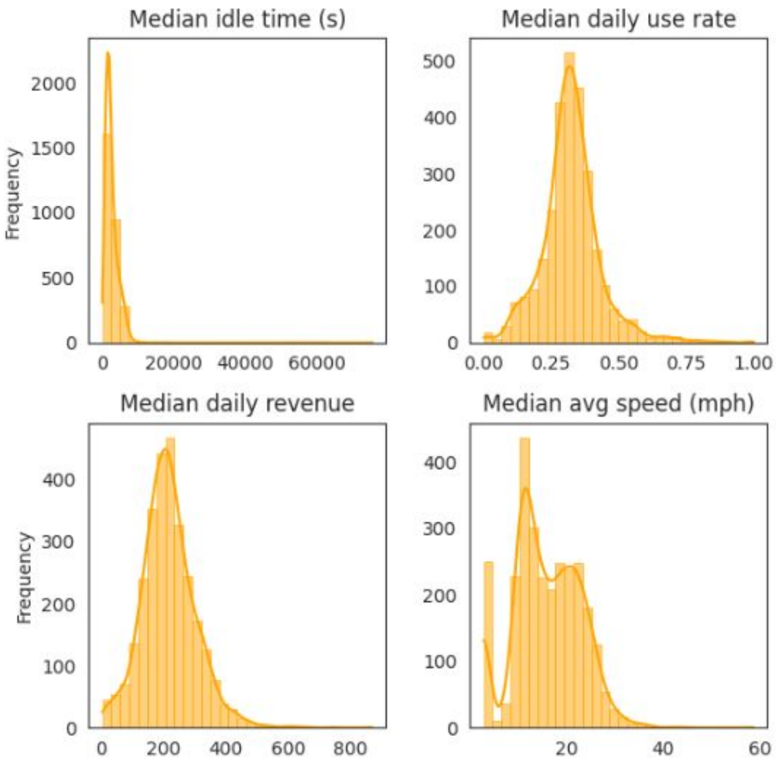
# Taxi Performance Clustering

# Data aggregation

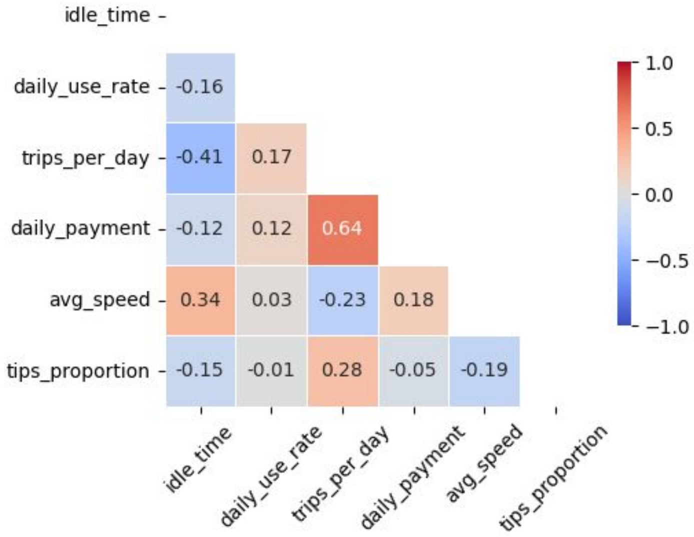
- Aggregated trip data into profiles for 2,864 taxis
- Derived taxi attributes
  - Performance: (median) **Idle Time, Daily Use Rate, Trips per Day, Average Speed**
  - Profitability: (median) **Daily Revenue**, **Tips Proportion**
  - Distribution of trips by pickup area, time of day, and payment type

| taxi_id | idle_time | daily_use_rate | trips_per_day | daily_payment | avg_speed | pay_cash | pay_mobile | morning_rush | midday | afternoon | pickup_north | pickup_central | ca_ohare |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 24da0b... | 4500.0 | 0.270 | 4.0 | 186.5 | 21.224 | 0.149 | 0.194 | 0.050 | 0.139 | 0.154 | 0.035 | 0.299 | 0.592 |
| c505f0a... | 2700.0 | 0.295 | 4.0 | 200.7 | 17.255 | 0.357 | 0.200 | 0.139 | 0.164 | 0.197 | 0.114 | 0.410 | 0.395 |
| 96a6dd... | 1800.0 | 0.348 | 2.5 | 68.7 | 10.595 | 0.399 | 0.246 | 0.162 | 0.105 | 0.162 | 0.026 | 0.579 | 0.123 |

# EDA



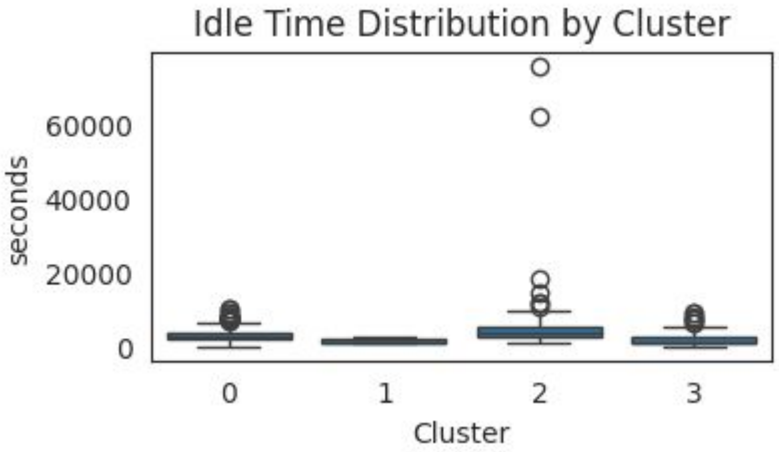
Median idle time (s)
Median daily use rate
Median trips/day
Median daily revenue
Median avg speed (mph)
Median tip proportion
Frequency

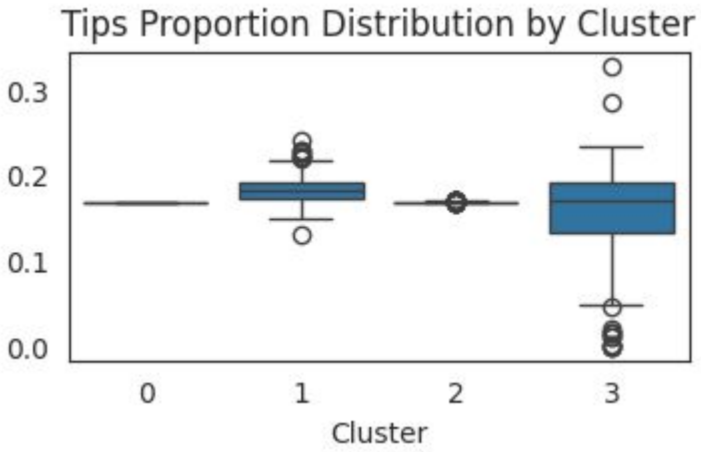
idle_time
daily_use_rate
trips_per_day
daily_payment
avg_speed
tips_proportion
-0.16
-0.41
0.17
-0.12
0.12
0.64
0.34
0.03
-0.23
0.18
-0.15
-0.01
0.28
-0.05
-0.19

# Baseline model

- RobustScaler() for feature scaling due to outliers
- GMM with k = 4 based on silhouette score, BIC score, and testing
- Challenges: imbalanced clusters, feature with little variance



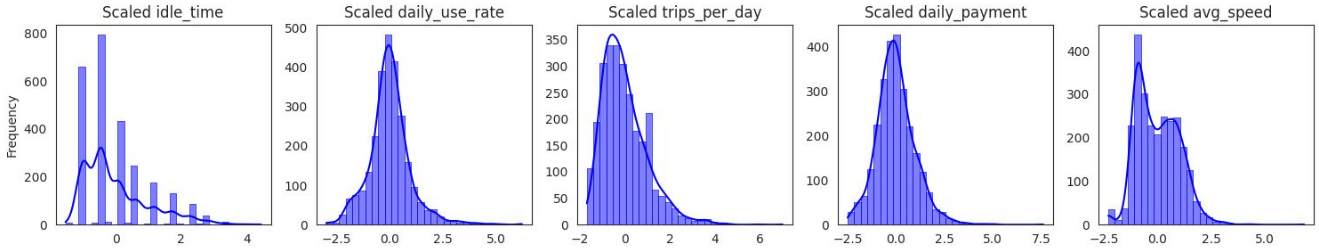

⇒ Remove outliers




⇒ Drop feature

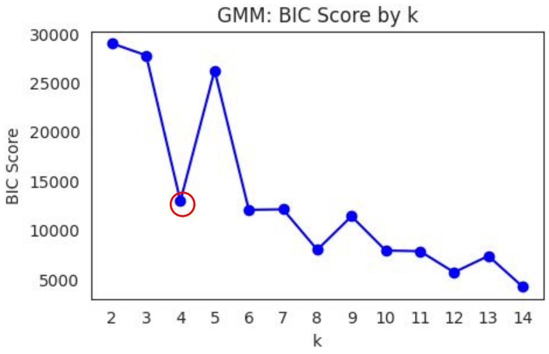
# Final model

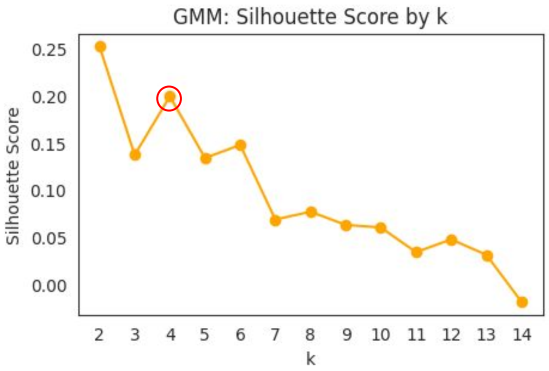
- Removed outliers, StandardScaler() for feature scaling



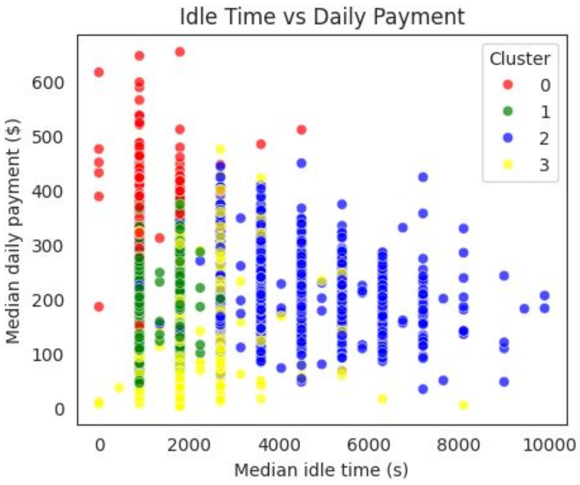
# Final model

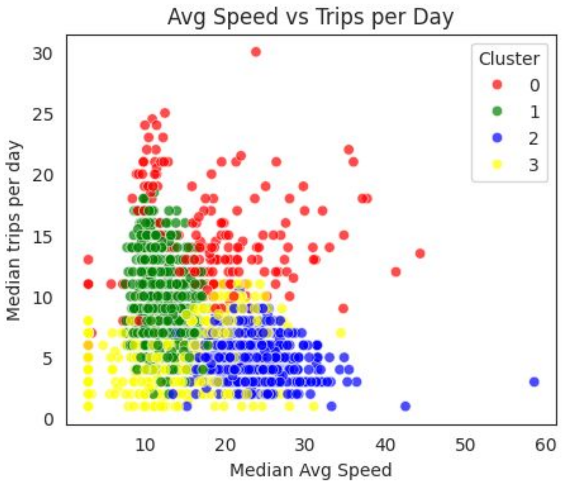
- Removed outliers, StandardScaler() for feature scaling
- Assess clustering methods
  - Test different parameters and initializations
  - Imbalance in cluster sizes with KMeans, HDBSCAN, hierarchical
  - GMM shows the most distinct clusters, optimal **k = 4**

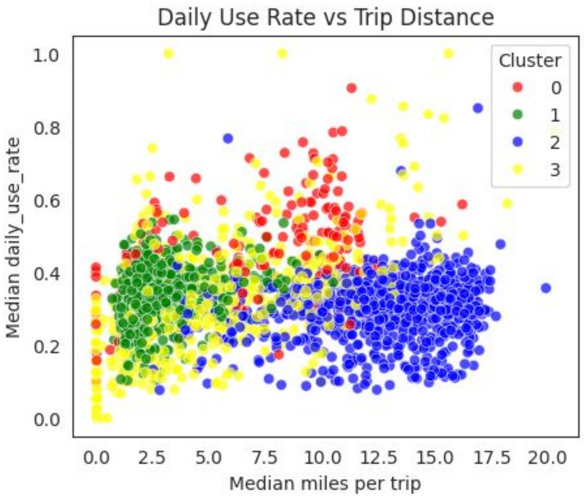


# Results




**C0:** 😊 **revenue** 😊 **idle time** 😐 speed 😊 **trips/ day** mix distance, 😊 use rate

**C1:** 😐 revenue 😊 idle time 😞 **speed** 😐 trips/ day short trips

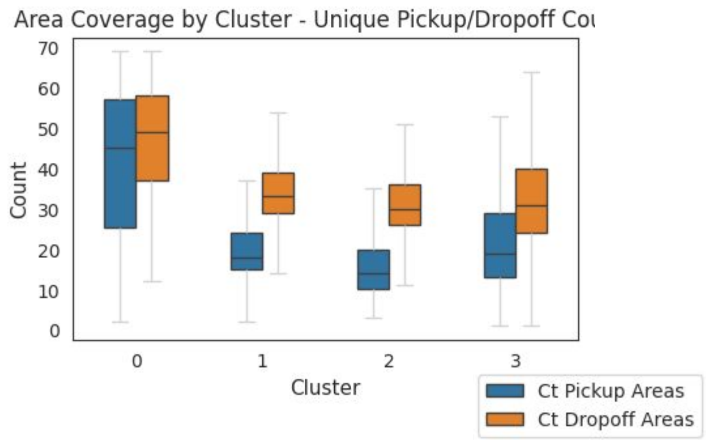
**C2:** 😐 revenue 😞 **idle time** 😊 **speed** 😞 trips/day long trips

**C3:** 😞 **revenue** 😞 idle time 😐 speed 😞 trips/day short and medium trips

# Results

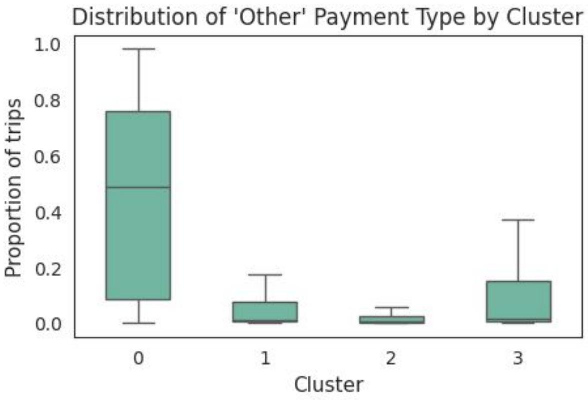
## Cluster 0 : The High Performers

- Top earner, high efficiency, steady demand
- Broad coverage beyond typical hotspots
- High usage of non-standard payment methods




## 💡 Suggestions

- Market wide coverage to attract diverse riders
- Explore 'other' payments for new revenue channels
- Adopt best practices to increase efficiency in other fleet

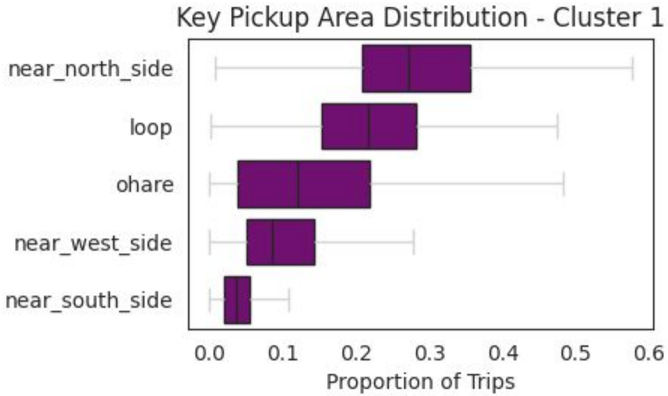


# Results

## Cluster 1 : Urban Cabs

- Short trips but slower speeds, steady service demand in downtown

- Active during rush/ business hours

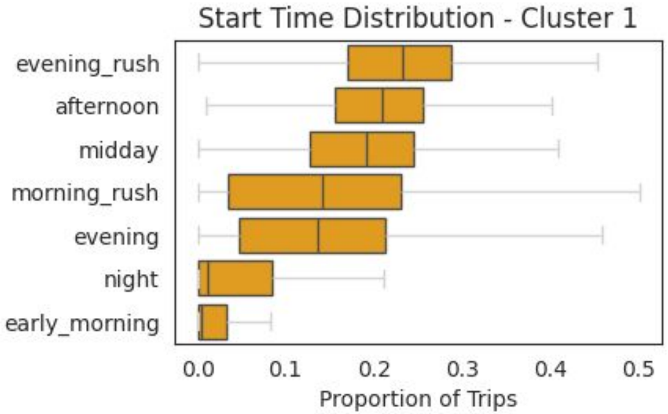



## 💡 Suggestions

- **Better route optimization** to avoid congested areas

- Offer perks for trips during less busy hours

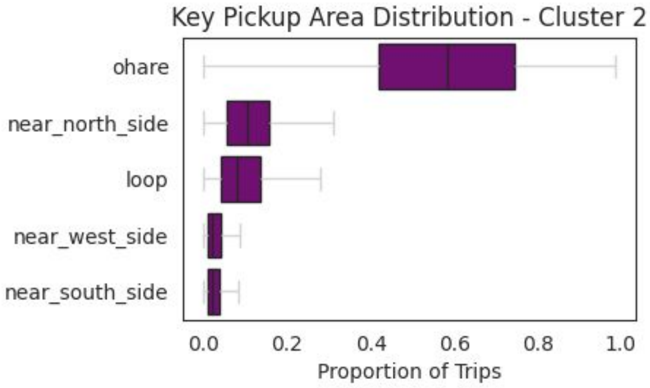
- **Expand coverage** to other high-demand city areas



# Results

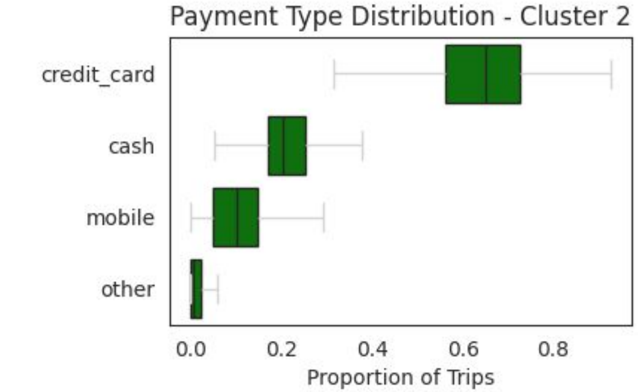
## Cluster 2 : Airport Cabs

- Primarily airport trips with peak evening activity, explains longer routes and fast speeds
- Low trip count, high idle time suggest downtime issues
- Preference for credit card payments




## 💡 Suggestions

- Adjust taxi scheduling to match flight arrival times to target peak airport demand
- Add courier services during downtime
- Partner with hotels/ airlines for airport pick-ups



# Results

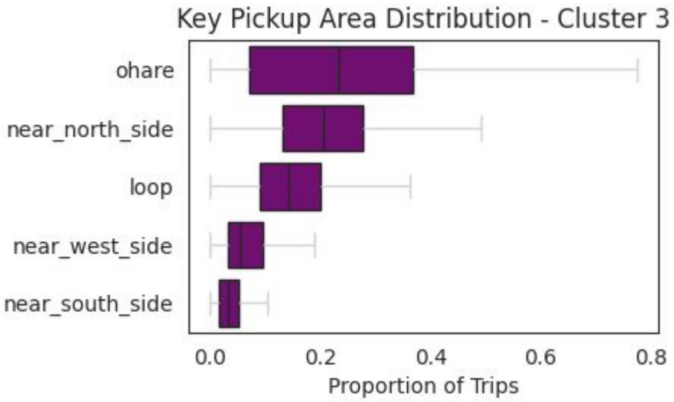
## Cluster 3 : Low Performers

- Balance airport and city trips but fail to optimize either

=> high idle time and low earnings

- Effective service but poor demand capture

=> infrequent trips




## 💡 Suggestions

- **Demand analysis:** understand reasons for low trip counts

- **Adopt high-performance practices** to reduce idle times, especially at O’Hare, while streamlining city operations

# Thank You!