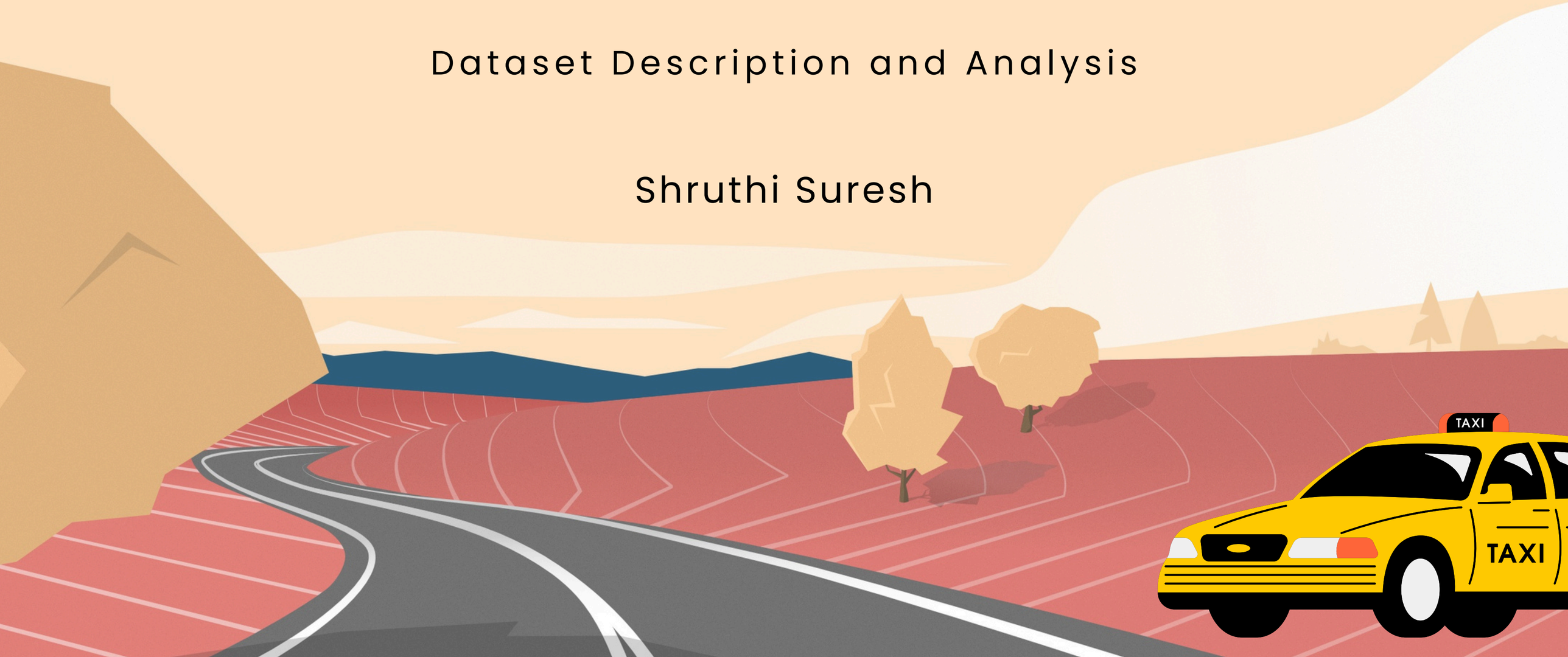
CITY OF CHICAGO TAXI TRIPS

Dataset Description and Analysis

Shruthi Suresh



PURPOSE OF DATA COLLECTION

Key Insights:

- **Operational Efficiency:** Trip durations, distances, routes, and peak demand.
- **Customer Behavior:** Payment preferences and tipping patterns.
- **Geographic Insights:** Popular pickup/drop-off locations and community-level trip analysis.
- **Revenue Optimization:** Anomalies in fare structures and additional charges.
- **Data Quality:** Addressing missing or anomalous data.

Objective:

1. Analyze and optimize taxi services in Chicago.
2. Understand transportation patterns, operational efficiencies, and passenger behaviors.

DATA COLLECTION

01 **Source**: Taxi meters integrated with GPS and point-of-sale systems.

02 **Sampling**: Census-style collection of all taxi trips during 2024.

03 **Timeframe**: Includes both AM and PM trips for the year 2024.

04 **Demographics:**

- Focus on Chicago (community areas, census tracts, geospatial data).
- Indirect passenger insights through payment modes and trip locations.

DATA CLEANING

01

Fill missing numeric values with the median

'Trip Seconds', 'Trip Miles', 'Trip Total', 'Extras', 'Tips', 'Tolls', 'Fare'

02

Fill missing categorical values with 'Unknown'

'Payment Type', 'Company'

03

Drop rows with missing critical location values (latitudes/longitudes)

'Pickup Centroid Latitude', 'Pickup Centroid Longitude', 'Dropoff Centroid Latitude', 'Dropoff Centroid Longitude'

OUTLIER REMOVAL USING IQR METHOD

01

Outlier Detection: Identified outliers in key numeric columns ('Trip Miles', 'Trip Seconds', 'Trip Total') using the Interquartile Range (IQR) method.

02

Methodology:

- Calculated the first (Q1) and third (Q3) quartiles.
- Defined outliers as values outside the range:
  - Lower Bound=Q1-1.5×IQR
  - Upper Bound=Q3+1.5×IQR
- Filtered out rows where values were below the lower bound or above the upper bound.

03

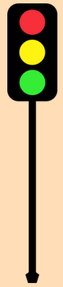
Result: Cleaned Dataset: Outliers were successfully removed, resulting in a more reliable dataset for analysis.

DESCRIPTIVE STATISTICS

	Trip Seconds	Trip Miles	Fare	Tips	Trip Total
count	647934.000000	648028.000000	646274.000000	646274.000000	646274.000000
mean	1272.107150	6.810736	23.037352	2.964785	28.440445
std	1618.118758	7.983354	30.334046	4.321618	36.354903
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	480.000000	1.000000	8.500000	0.000000	10.500000
50%	959.000000	3.300000	16.250000	1.000000	19.500000
75%	1743.000000	12.050000	35.000000	4.210000	44.100000
max	85829.000000	2166.390000	8050.000000	297.000000	8893.380000

1. **Trip Duration (Seconds)**: The average trip lasts about 21 minutes (1272 seconds), with a wide range, from 0 to over 23 hours.
2. **Trip Distance (Miles)**: The average trip covers 6.81 miles, with a range from 0 to over 2166 miles (likely due to data issues or outliers).
3. **Fare**: The average fare is $23.04, but there's significant variation (std = $30.33), with fares ranging from $0 to $8050.
4. **Tips**: The average tip is $2.96, with some trips receiving high tips up to $297, indicating potential outliers or special cases.
5. **Total Trip Cost**: The total cost, combining fare and tips, averages $28.44, with a broad range from $0 to $8893.38.
6. **Outliers**: There are significant outliers, especially in trip distance, fare, and total cost, suggesting data cleaning might be needed.

VARIABLE DESCRIPTION AND DATA TYPES




```
                                   Trip ID  \
0  0000184e7cd53cee95af32eba49c44e4d20adcd8
1  000072ee076c9038868e239ca54185eb43959db0
2  000074019d598c2b1d6e77fbae79e40b0461a2fc
3  00007572c5f92e2ff067e6f838a5ad74e83665d3
4  00007c3e7546e2c7d15168586943a9c22c3856cf

                                        Taxi ID    Trip Start Timestamp  \
0  f538e6b729d1aaad4230e9dcd9dc2fd9a168826ddadbd6...  01/19/2024 05:00:00 PM
1  e51e2c30caec952b40b8329a68b498e18ce8a1f40fa75c...  01/28/2024 02:30:00 PM
2  aeb280ef3be3e27e081eb6e76027615b0d40925b84d3eb...  01/05/2024 09:00:00 AM
3  7d21c2ca227db8f27dda96612bfe5520ab408fa9a462c8...  01/22/2024 08:45:00 AM
4  8ef1056519939d511d24008e394f83e925d2539d668a00...  01/18/2024 07:15:00 PM

      Trip End Timestamp  Trip Seconds  Trip Miles  Pickup Census Tract  \
0  01/19/2024 06:00:00 PM        4051.0       17.12         1.703198e+10
1  01/28/2024 03:00:00 PM        1749.0       12.70                  NaN
2  01/05/2024 09:00:00 AM         517.0        3.39                  NaN
3  01/22/2024 09:30:00 AM        2050.0       15.06                  NaN
4  01/18/2024 07:30:00 PM        1004.0        1.18         1.703184e+10

   Dropoff Census Tract  Pickup Community Area  Dropoff Community Area  ...  \
0         1.703132e+10                   76.0                    32.0  ...
1                  NaN                    6.0                     NaN  ...
2                  NaN                    6.0                     8.0  ...
3                  NaN                   76.0                     NaN  ...
4         1.703184e+10                   32.0                    32.0  ...

   Extras  Trip Total  Payment Type                     Company  \
0     4.0       60.00   Credit Card                  Flash Cab
1     0.0       33.75          Cash                  Flash Cab
2     1.0       14.69        Mobile  Taxicab Insurance Agency Llc
3     5.5       56.56   Credit Card                 Globe Taxi
4     0.0       19.66        Mobile                5 Star Taxi

   Pickup Centroid Latitude Pickup Centroid Longitude  \
0                41.979071                -87.903040
1                41.944227                -87.655998
2                41.944227                -87.655998
3                41.980264                -87.913625
4                41.880994                -87.632746

             Pickup Centroid Location  Dropoff Centroid Latitude  \
0  POINT (-87.9030396611 41.9790708201)                  41.884987
1  POINT (-87.6559981815 41.9442266014)                        NaN
2  POINT (-87.6559981815 41.9442266014)                  41.899602
3   POINT (-87.913624596 41.9802643146)                        NaN
4  POINT (-87.6327464887 41.8809944707)                  41.880994

   Dropoff Centroid Longitude           Dropoff Centroid  Location
0                -87.620993  POINT (-87.6209929134 41.8849871918)
1                       NaN                                   NaN
2                -87.633308   POINT (-87.6333080367 41.899602111)
3                       NaN                                   NaN
4                -87.632746  POINT (-87.6327464887 41.8809944707)

[5 rows x 23 columns]
```

VARIABLE DESCRIPTION AND DATA TYPES

Variable	Data Type	Type	Description
Trip ID	string	Categorical	Unique identifier for each trip
Taxi ID	string	Categorical	Unique identifier for each taxi
Trip Start Timestamp	datetime	Categorical	The start time of the trip
Trip End Timestamp	datetime	Categorical	The end time of the trip
Trip Seconds	float	Numerical	Duration of the trip in seconds
Trip Miles	float	Numerical	Distance covered in miles

VARIABLE DESCRIPTION AND DATA TYPES

Variable	Data Type	Type	Description
Pickup Census Tract	float	Numerical	Census identifier for the pickup location
Dropoff Census Tract	float	Numerical	Census identifier for the dropoff location
Pickup Community Area	float	Numerical	Community ID for the pickup area
Dropoff Community Area	float	Numerical	Community ID for the dropoff area
Fare	float	Numerical	Fare charged for the trip
Tips	float	Numerical	Tips given for the trip

VARIABLE DESCRIPTION AND DATA TYPES

Variable	Data Type	Type	Description
Tolls	float	Numerical	Tolls incurred during the trip
Extras	float	Numerical	Additional charges (e.g., waiting time)
Trip Total	float	Numerical	Total fare, including tips, tolls, and extras
Payment Type	string	Categorical	The mode of payment (e.g., cash, credit card)
Company	string	Categorical	The taxi company managing the trip
Pickup Centroid Latitude	float	Numerical	Latitude of the pickup location

VARIABLE DESCRIPTION AND DATA TYPES

Variable	Data Type	Type	Description
Pickup Centroid Longitude	float	Numerical	Longitude of the pickup location
Pickup Centroid Location	string	Categorical	Geospatial representation of the pickup location (e.g., "POINT (x y)")
Dropoff Centroid Latitude	float	Numerical	Latitude of the dropoff location
Dropoff Centroid Longitude	float	Numerical	Longitude of the dropoff location
Dropoff Centroid Location	string	Categorical	Geospatial representation of the dropoff location (e.g., "POINT (x y)")

EXPLANATION OF DATA TYPES

Categorical Variables:

- **Definition:** These variables describe categories or groups, often represented as strings or IDs
- **Examples:** Trip ID, Taxi ID, Payment Type, Company, Pickup Centroid Location, Dropoff Centroid Location

Numerical Variables:

- **Definition:** These variables contain measurable values, either continuous (e.g., Trip Miles, Fare) or discrete (e.g., Pickup Community Area if converted to integers)
- **Examples:** Trip Seconds, Trip Miles, Fare, Tips, Tolls, Extras, Trip Total, Pickup Centroid Latitude, Pickup Centroid Longitude

EXPLANATION OF DATA TYPES

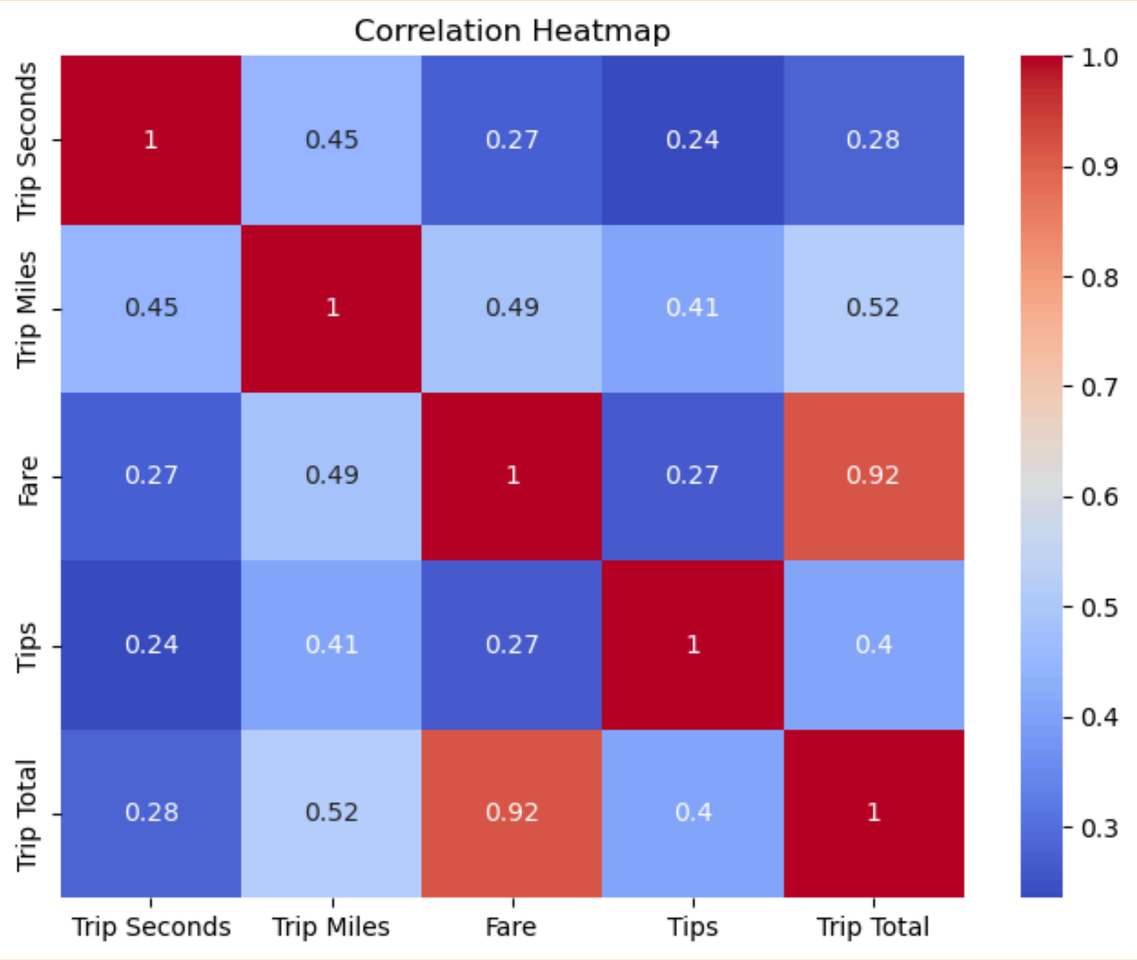
Datetime Variables:

- **Definition:** These variables represent date and time, useful for time-based operations
- **Examples:** Trip Start Timestamp, Trip End Timestamp

Special Cases:

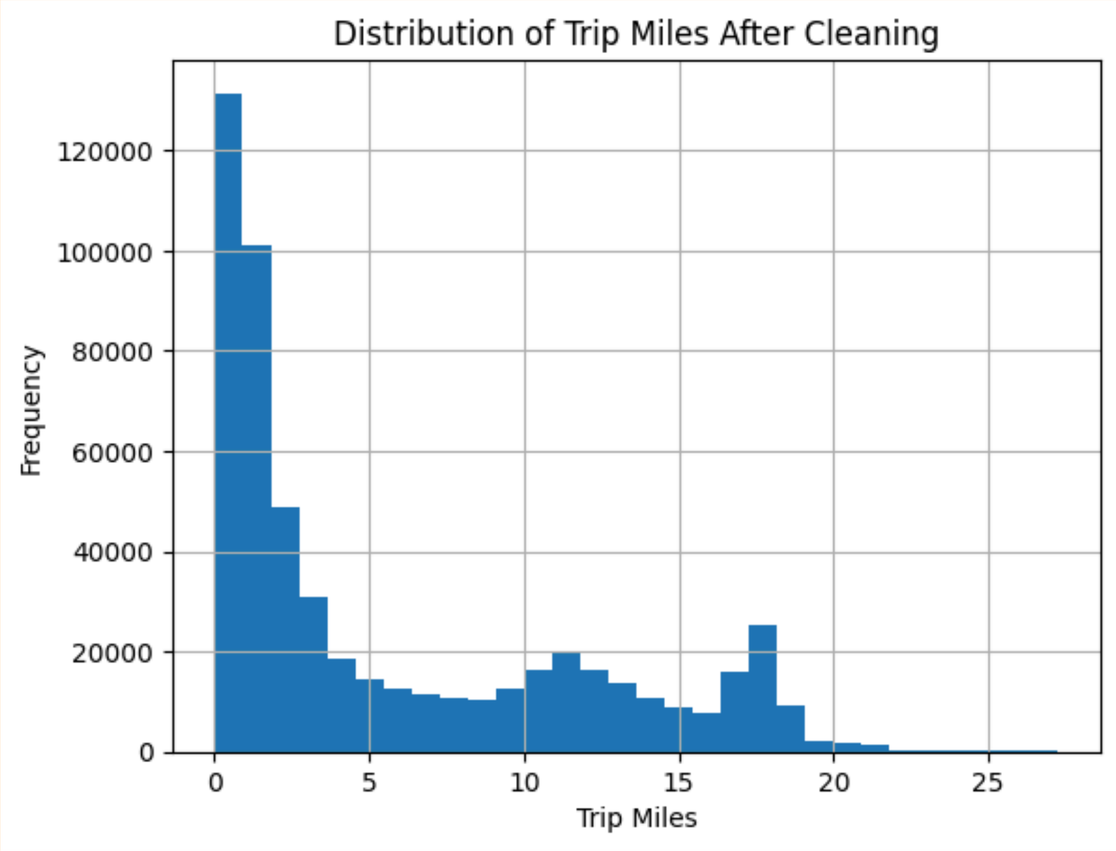
- **Community Area IDs:** These can be stored as integers (int) if there are no missing values, but remain float to account for potential gaps in the dataset.
- **Geospatial Locations:** While latitude and longitude are float, the geospatial points (Pickup Centroid Location, Dropoff Centroid Location) are stored as strings

Correlation Analysis

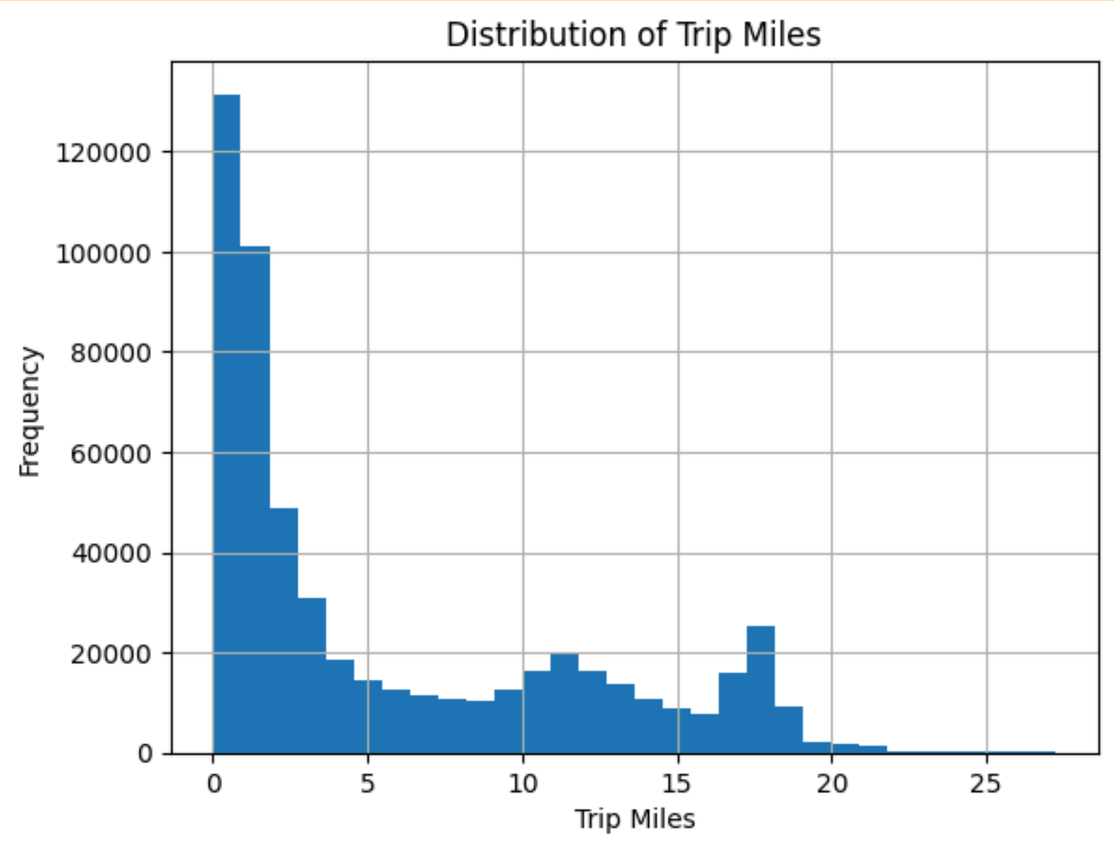


- **Fare and Trip Total (0.92)**: Strong correlation, as trip total is mainly based on fare, plus tips and tolls.
- **Trip Miles and Fare (0.49)**: Moderate correlation, indicating longer trips usually cost more, but other factors like base fare affect the price.
- **Trip Miles and Trip Total (0.52)**: Moderate correlation, with distance impacting total cost, though other factors (e.g., tolls) also play a role.
- **Trip Seconds and Trip Miles (0.45)**: Moderate correlation, suggesting longer trips usually take more time, but traffic can vary the relationship.
- **Tips and Other Variables (~0.4)**: Weak correlation, indicating tips are influenced more by service or personal factors than by fare or trip length.

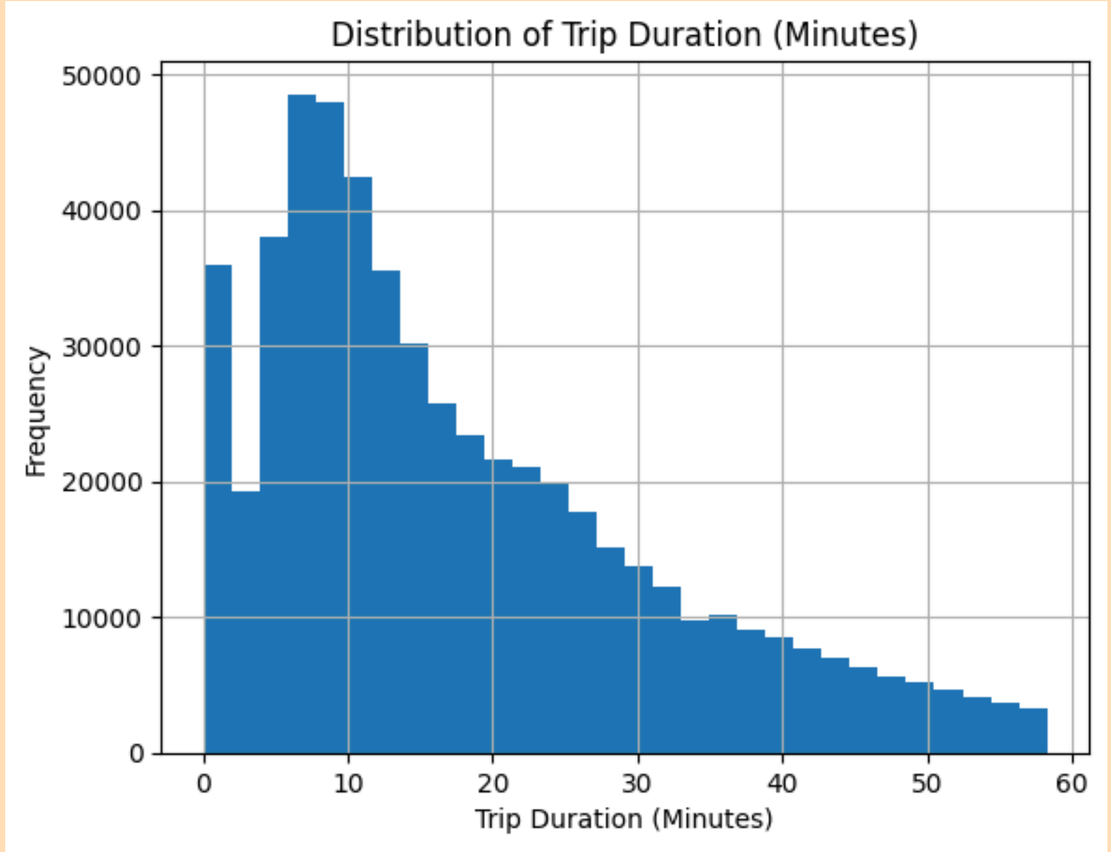
DATA VISUALIZATION



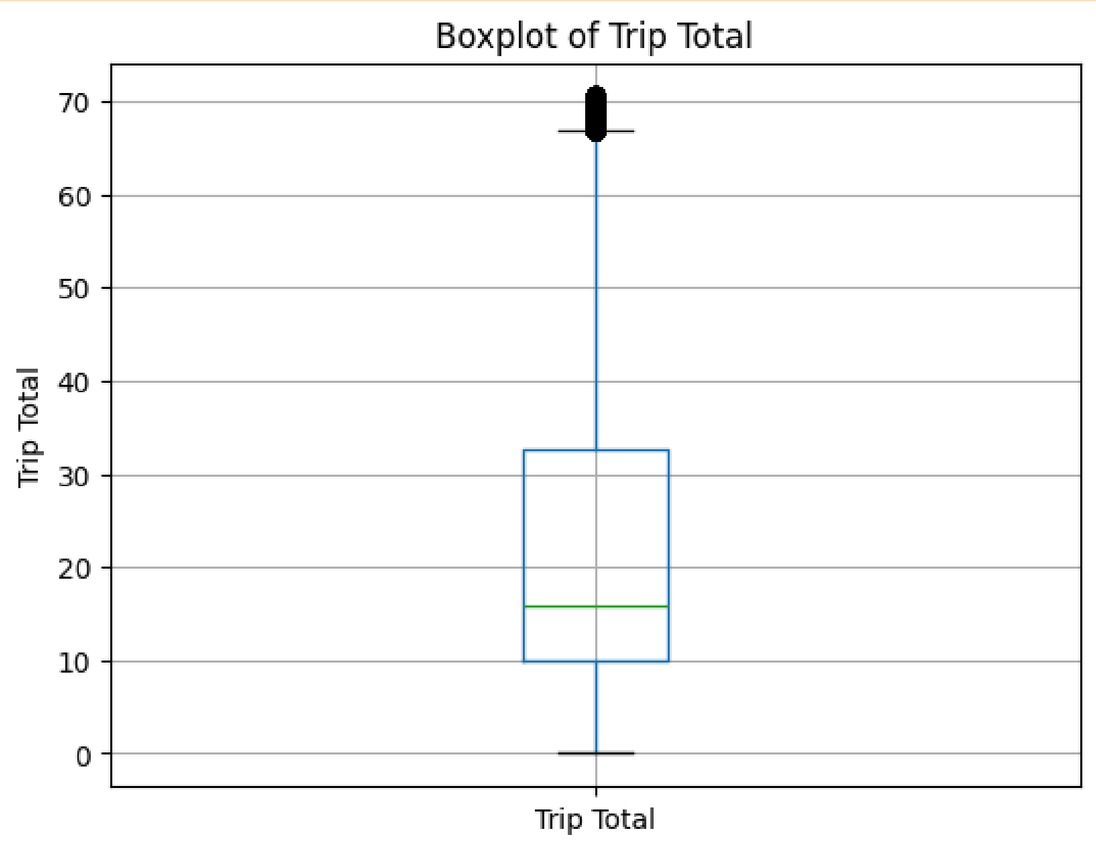
- Boxplot shows "Trip Miles" distribution after outlier removal, with median and IQR.
- No extreme values, indicating a cleaner dataset.
- Most trips fall between 0 and 5 miles, with fewer longer trips, reflecting typical city taxi patterns.
- Outliers (e.g., extremely long trips) were removed, retaining valid data up to 27.25 miles.



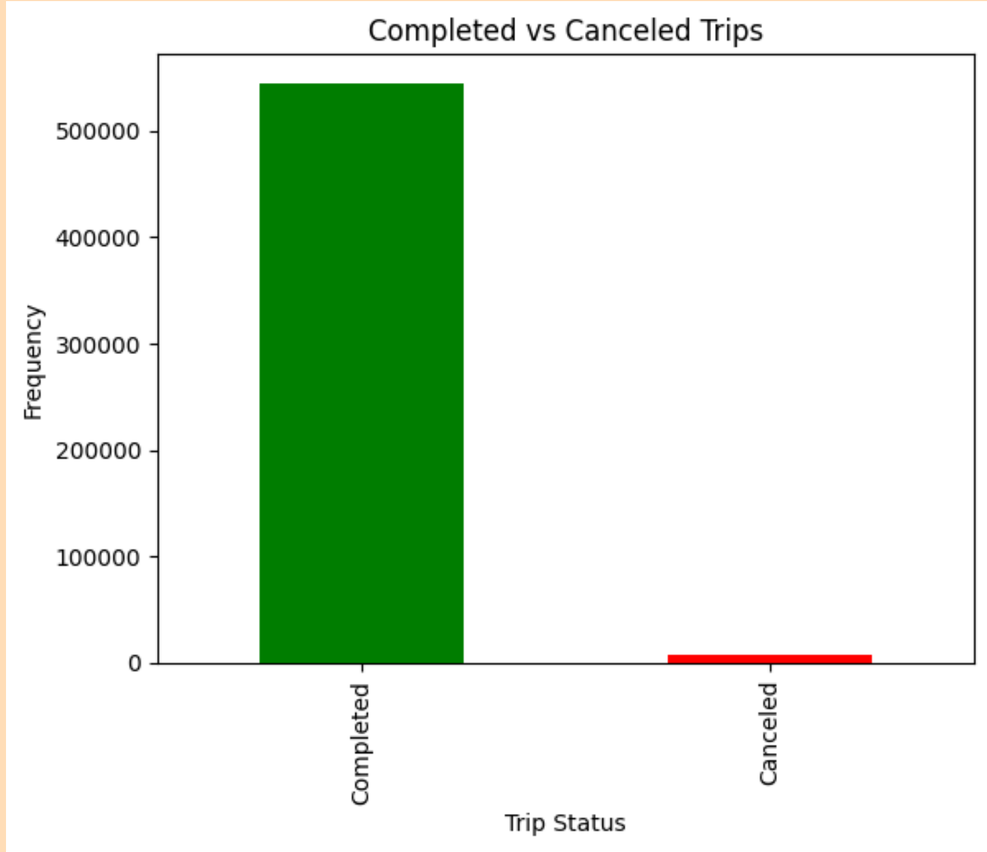
- **Most Trips Are Short**: The majority fall in the 0–5 miles range, with a peak around 1–2 miles, typical for city taxis.
- **Longer Trips Are Rare**: Trips over 5 miles are uncommon, with few exceeding 15 miles.
- **Right-Skewed Distribution**: The histogram shows a long tail, indicating a small number of longer trips.
- **Urban Taxi Behavior**: The data reflects typical urban travel, with short trips dominating and rare long trips like airport transfers.



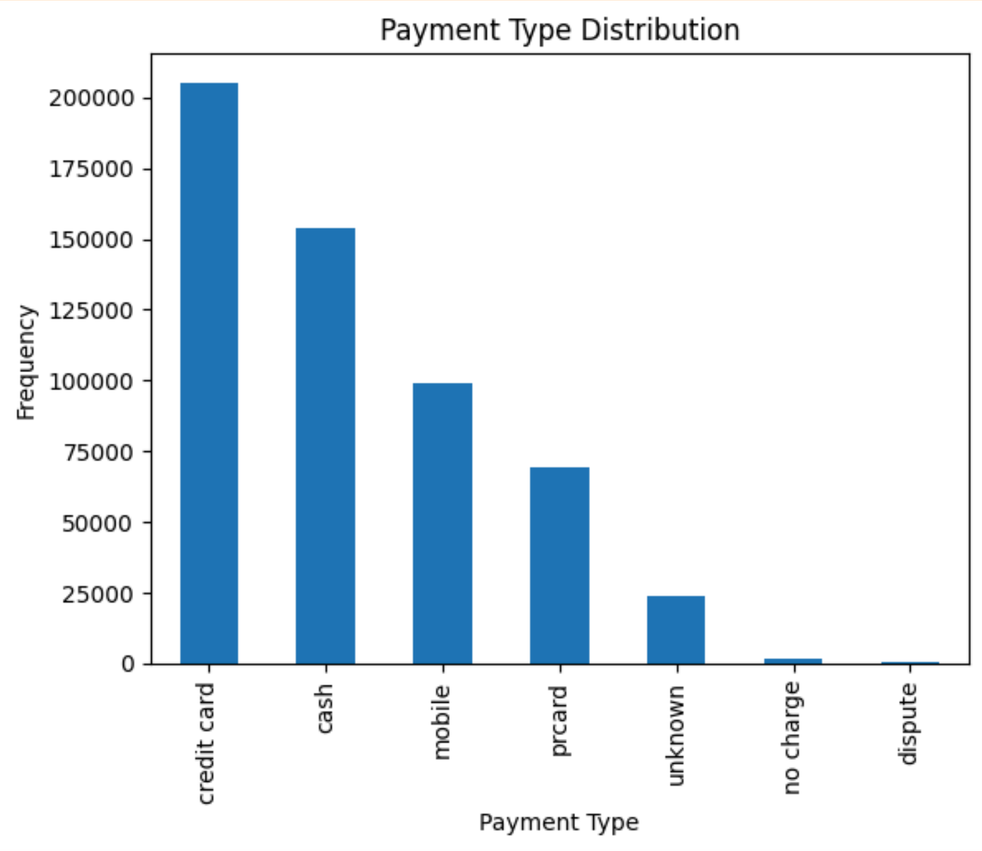
- **Most Trips Are Short**: The majority of trips last 5–20 minutes, peaking around 10 minutes, typical for urban taxi services.
- **Decline in Longer Durations**: Trips longer than 20 minutes become less frequent, with very few exceeding 40 minutes.
- **Right-Skewed Distribution**: The histogram shows a long tail, indicating rare longer trips.



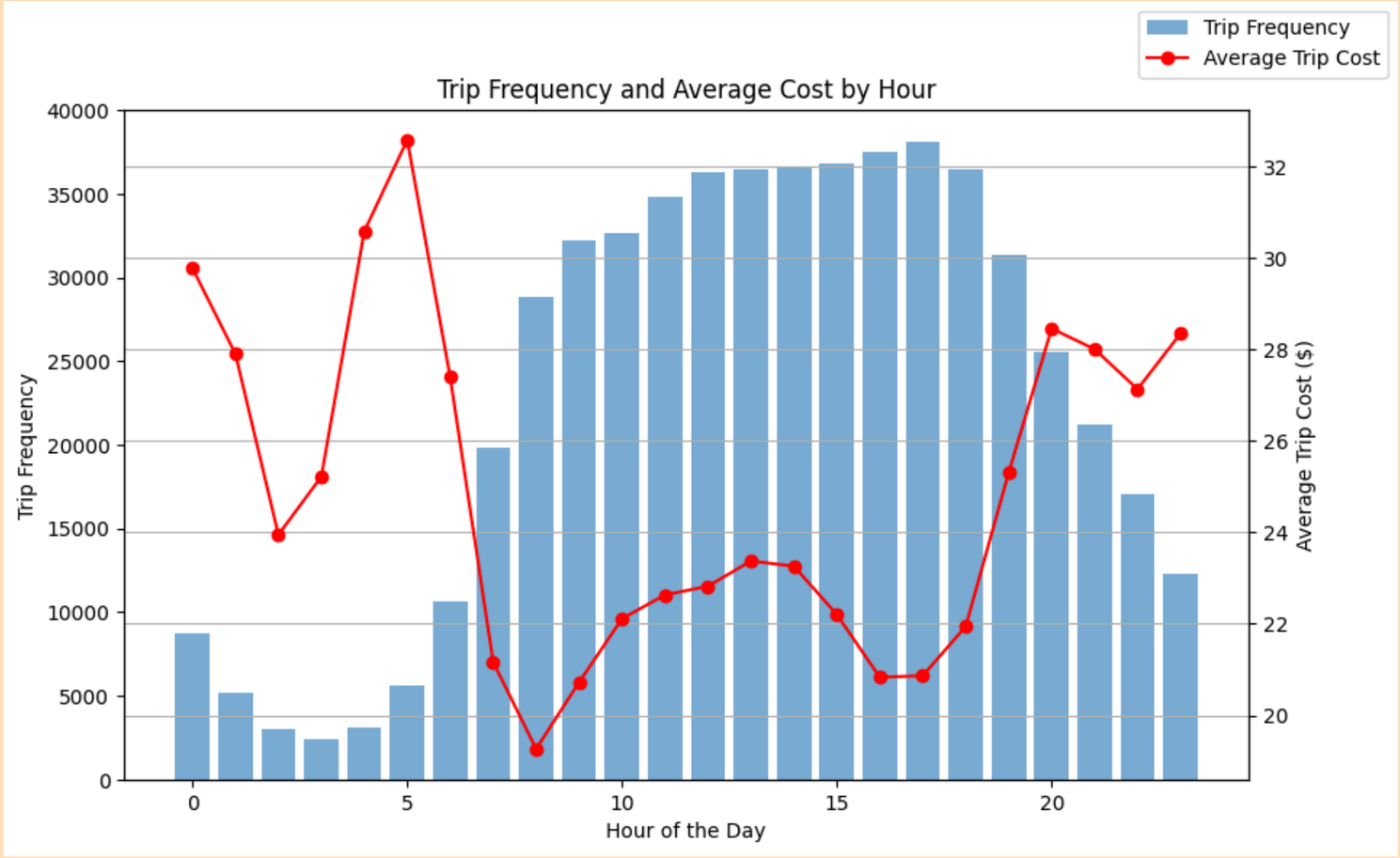
- **Median (Green Line)**: The median trip total is around $15–20, with half of the trips costing less.
- **Interquartile Range (Box)**: Most trip totals fall between $10 and $32.7, indicating reasonably priced trips.
- **Whiskers**: The whiskers extend from near $0 to just above $60, capturing the non-outlier range.
- **Outliers (Black Dots)**: A few trips exceed $60, likely due to long distances, high tips, or surcharges.



- **Most Trips Are Completed**: The green bar shows that most trips are completed, indicating efficient taxi operations and a low cancellation rate.
- **Few Canceled Trips**: The red bar represents a small number of canceled trips, aligning with the expectation that cancellations are rare in urban taxi services.



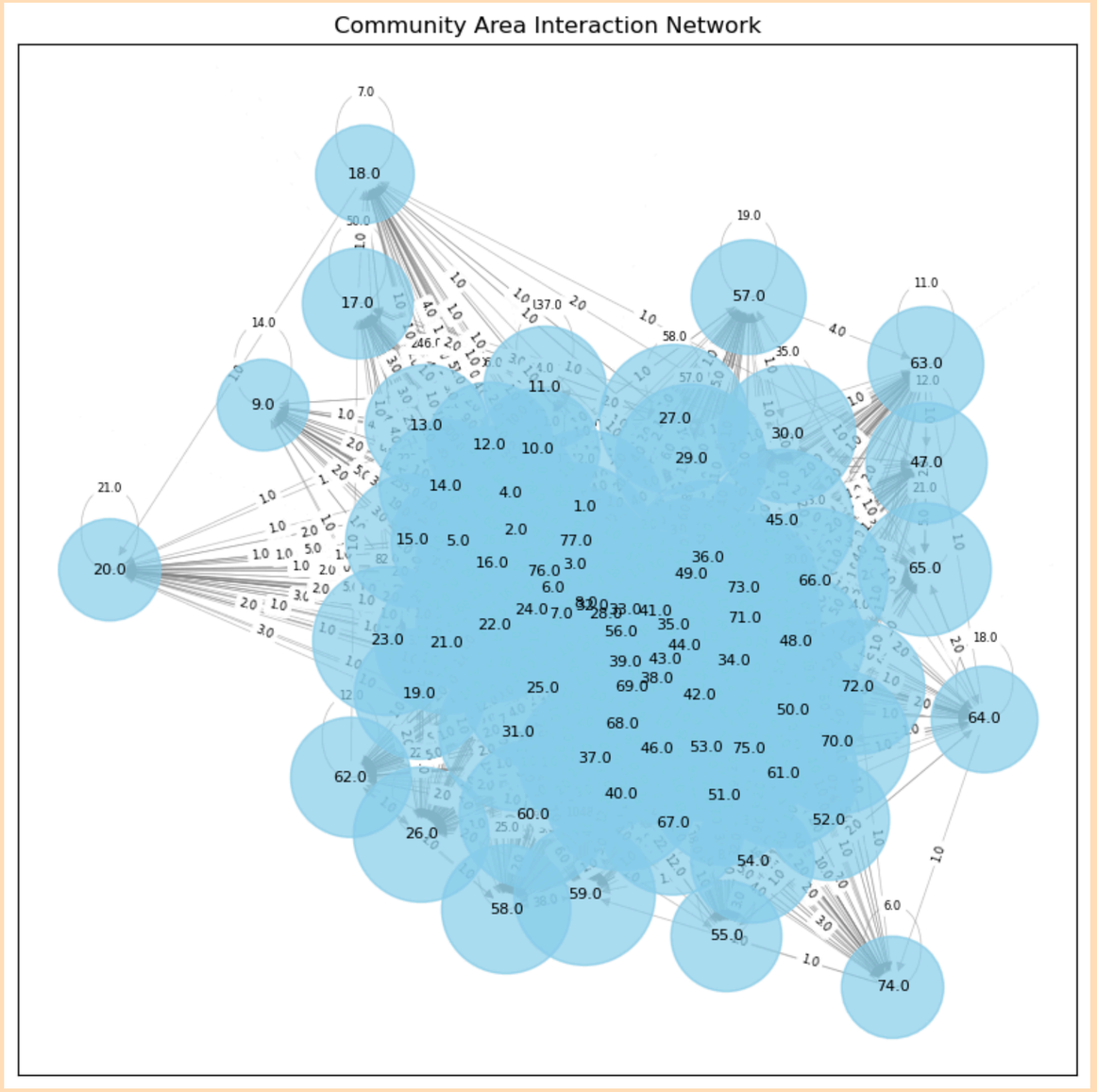
- **Credit Cards Dominate**: Most trips (over 175,000) are paid with credit cards, indicating a preference for electronic payments.
- **Cash as the Second Choice**: Cash is the second most common payment method, showing some passengers still use traditional payment.
- **Mobile Payments and Others**: Mobile payments are growing, while categories like "Prcard," "unknown," and "no charge" are less frequent, possibly due to errors or special cases.
- **Unusual Categories**: Rare categories like "dispute" and "no charge" suggest exceptional cases such as payment issues or promotional rides.

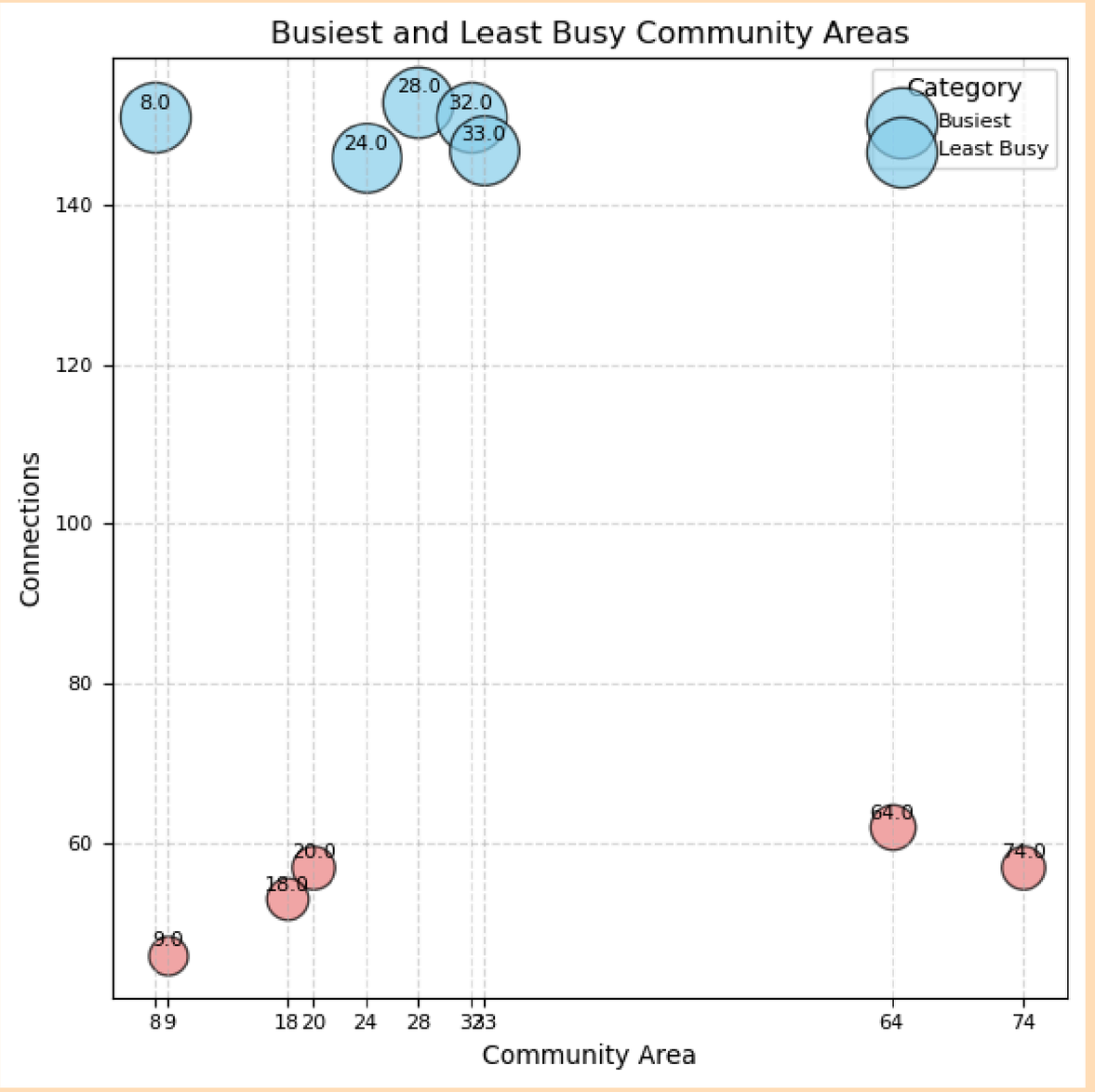


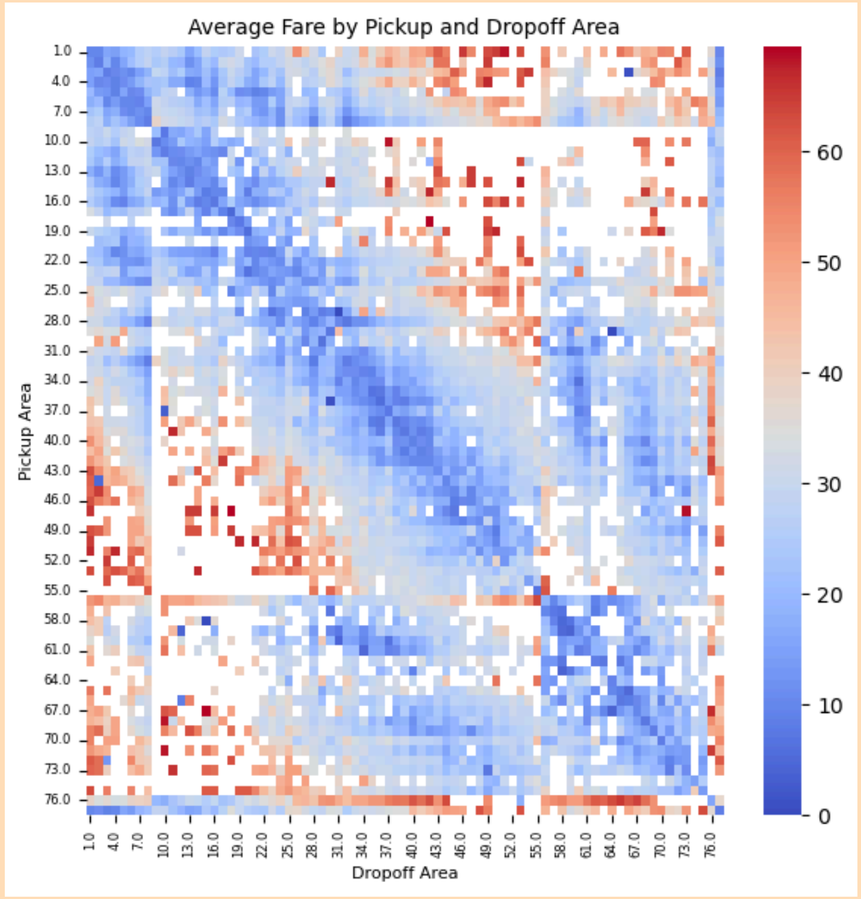
- **Demand Patterns**: Peak demand occurs from 7 AM–10 AM and 4 PM–8 PM, suggesting the need for more taxis during these hours.

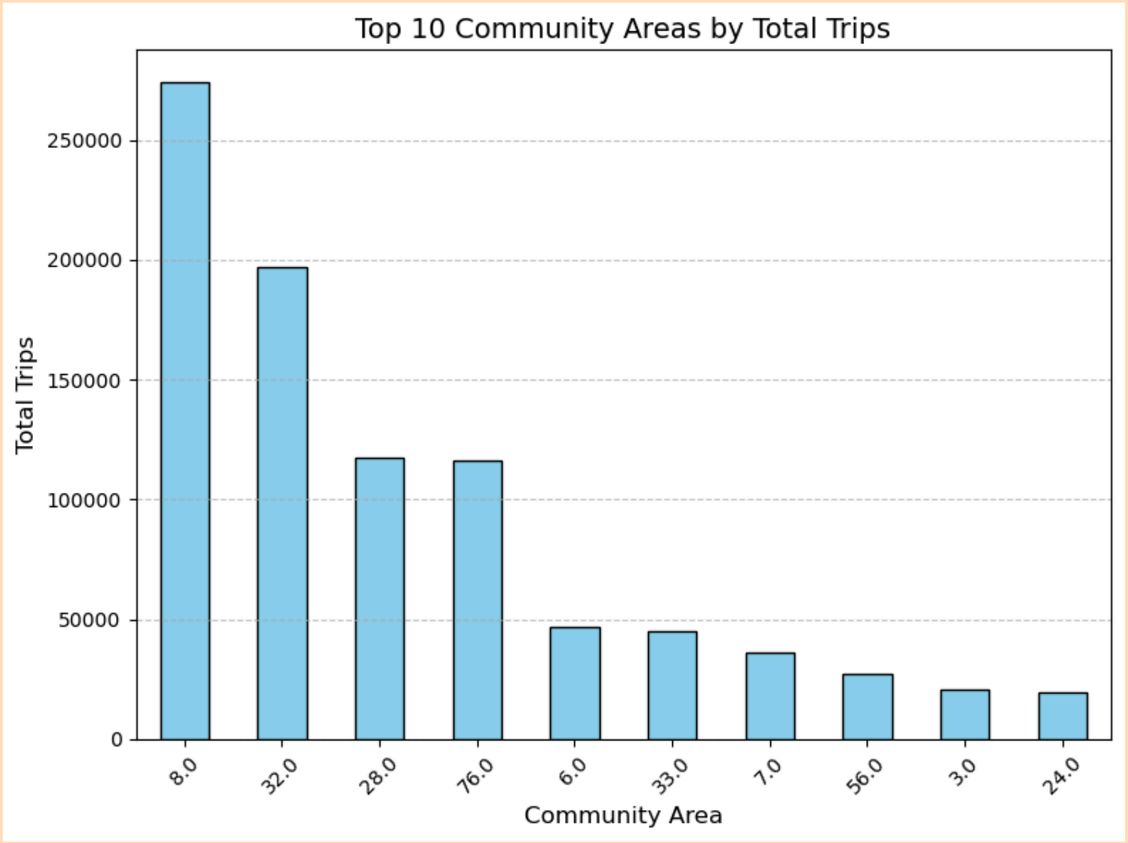
- **Revenue Opportunities**: The midnight to early morning period offers higher revenue per trip, indicating long-distance or premium rides.

- **Customer Behavior**: Commuters drive lower average costs during peak hours, while leisure and long-distance travelers increase costs late at night.


Community Area Interaction Network

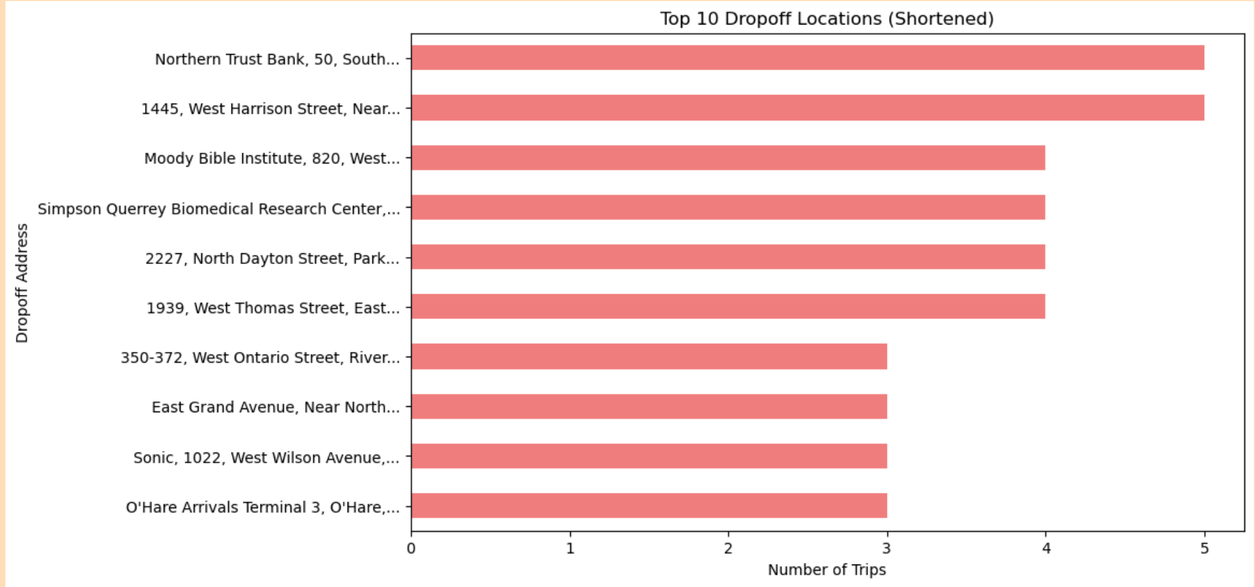

Busiest and Least Busy Community Areas
Category
Busiest
Least Busy
Connections
Community Area
140
120
100
80
60
8.0
28.0
32.0
33.0
24.0
64.0
74.0
20.0
18.0
9.0
8 9
18 20
24
28
32 33
64
74





- **High Fare Zones:** Red areas show high fares, likely from long trips or premium locations like airports.
- **Low Fare Zones:** Blue areas represent low fares from short-distance trips.
- **Sparse Data:** White spaces suggest fewer trips between certain areas.

- **High Trip Volume**: Community Area 8.0 leads with the highest number of trips, followed by 32.0 and 76.0, indicating high activity in these regions.
- **Moderate to Low Volume:** Community Areas like 7.0, 56.0, and 24.0 see fewer trips, likely reflecting less demand or smaller populations.



High-Demand Pickup Locations:

- O'Hare Arrivals Terminal 3 and O'Hare International Airport dominate the pickup locations, emphasizing the role of airports as major transportation hubs.
- Other key pickup locations, such as the Moody Bible Institute and Symphony Center, reflect demand from cultural, educational, and entertainment areas.

Geo-Coding Contribution:

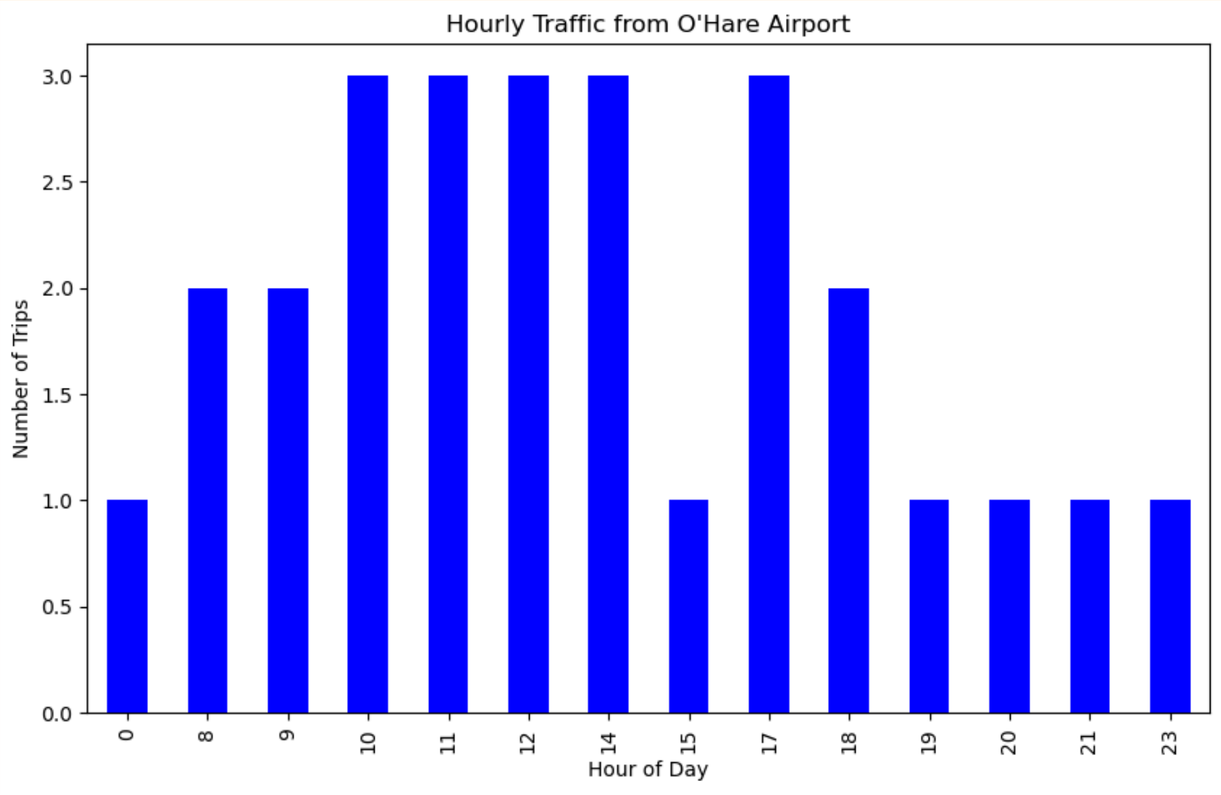
- Latitude and longitude data for pickup points were reverse-geocoded to extract readable addresses.
- These addresses were truncated to provide a concise summary while retaining meaningful location identifiers.

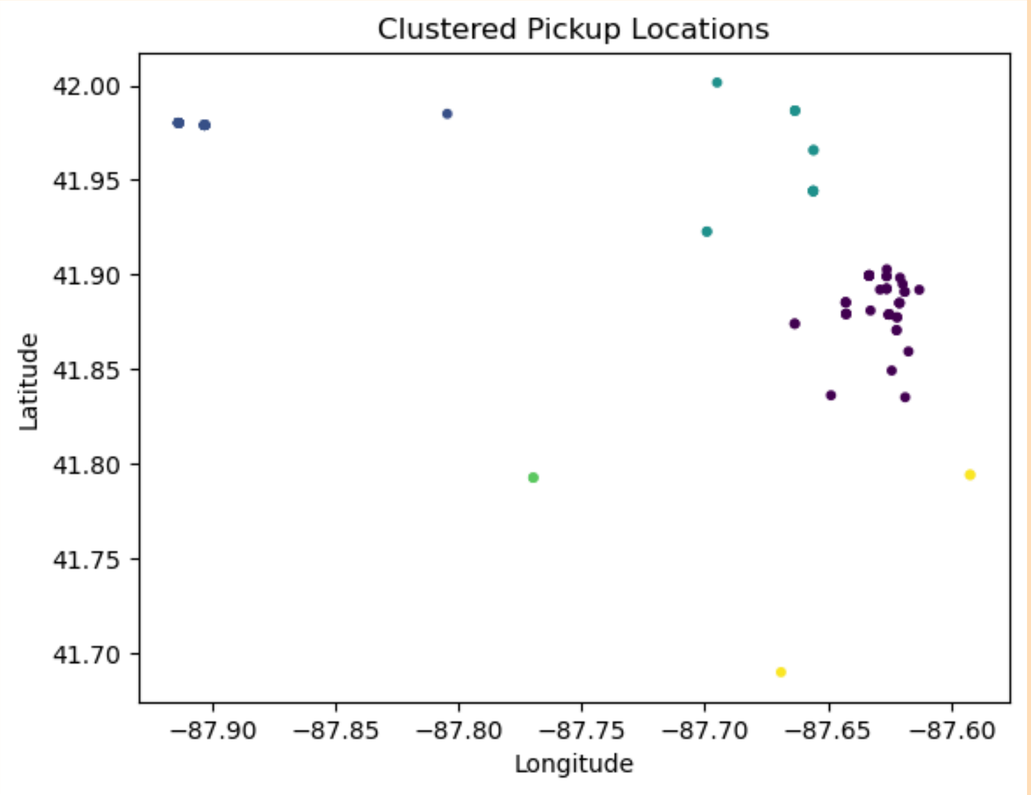
Frequent Dropoff Locations:

- Northern Trust Bank and 1445 West Harrison Street are prominent dropoff points, likely reflecting areas of economic activity and residential destinations.
- The Moody Bible Institute appears again as a key drop off point, showcasing its centrality in the taxi traffic network.

Geo-Coding Contribution:

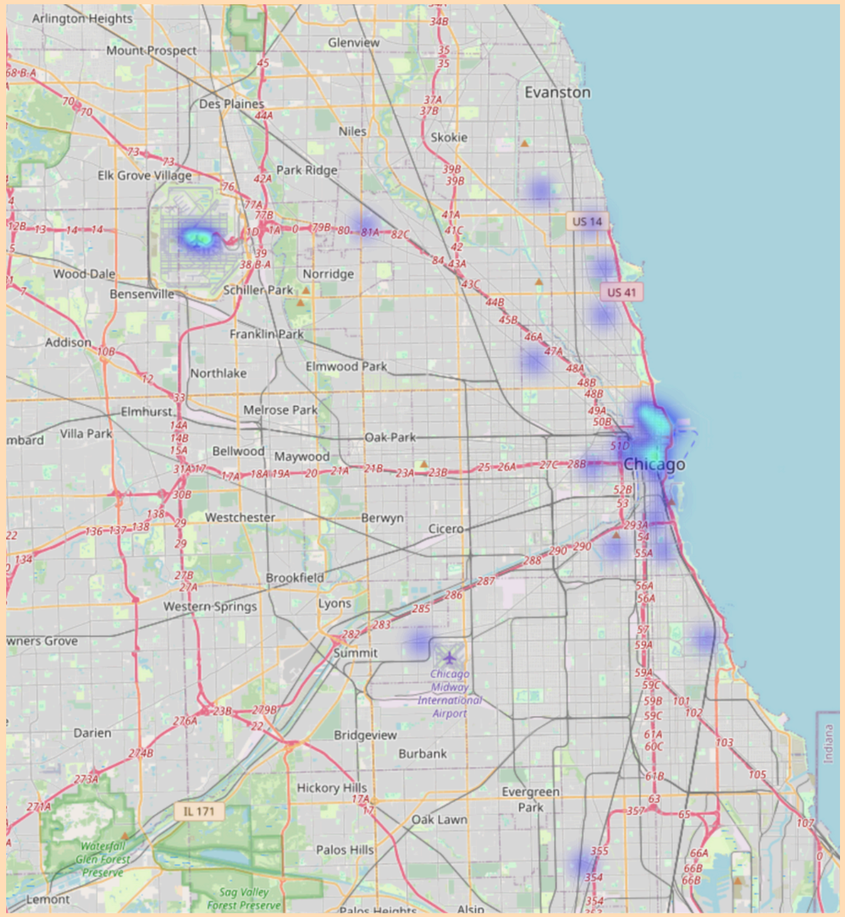
- Dropoff coordinates were reverse-geocoded into human-readable addresses, enabling meaningful interpretation of key destinations.
- Truncated addresses ensure the data remains clear and easy to visualize.



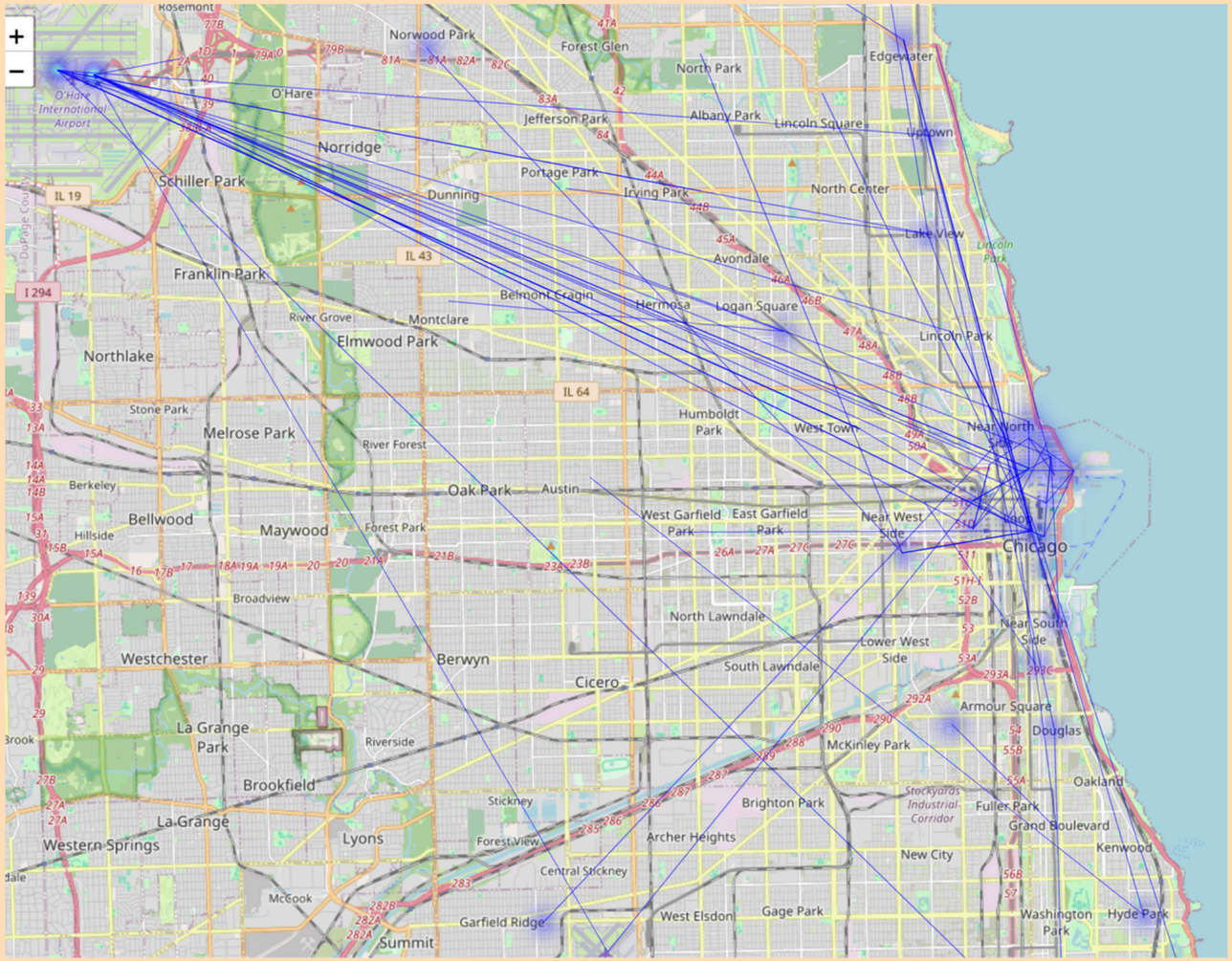


- **Peak Traffic Hours:** The highest taxi traffic occurs between 10 AM and 2 PM, likely due to morning flights and business activity.
- **Evening Spike:** Another noticeable increase in trips is observed around 6 PM, aligning with evening travel demand.
- **Low Traffic Periods:** Traffic drops significantly during the late-night hours, reflecting lower demand.

- **Geographical Concentration:** Pickup locations are densely clustered near central regions (indicated by the concentration of purple points).
- **Outlier Pickups:** A few points, like the yellow markers, represent distant or less frequent pickup locations, possibly serving remote areas or airports.
- **Spatial Patterns:** The clustering suggests high taxi demand in urban centers, while peripheral locations have sporadic pickups.



- **Traffic Corridors:** The flow map highlights major taxi trip corridors, especially from central Chicago and O'Hare Airport, indicating high interconnectivity between these regions.
- **Regional Dynamics**: Thicker lines represent higher traffic volumes, emphasizing significant demand between urban centers and surrounding suburbs.



- **Hotspots:** The heat map reveals concentrated taxi pickups in downtown Chicago and O'Hare Airport, reaffirming their importance as key transportation hubs.
- **Sparse Activity:** Peripheral areas show lower activity, indicating less frequent pickups, likely due to reduced demand or service availability in these regions.

BI QUESTIONS

- AVERAGE TRIP METRICS: WHAT ARE THE AVERAGE TRIP DISTANCE AND DURATION FOR COMPLETED RIDES?
- PEAK REVENUE HOURS: WHICH HOURS GENERATE THE HIGHEST REVENUE AND TRIP FREQUENCY?
- PAYMENT METHOD ANALYSIS: WHAT IS THE DISTRIBUTION OF PAYMENT TYPES, AND WHICH GENERATES THE MOST REVENUE?
- TRIP CANCELLATIONS: DO CANCELLATIONS OCCUR MORE FREQUENTLY FOR SPECIFIC TRIP LENGTHS OR DURATIONS?
- TIP PATTERNS: HOW DO TIP AMOUNTS VARY BASED ON DISTANCE, FARE, AND TIME OF DAY?

THANK YOU

ANY QUESTIONS?