Table of Content

Market Segmentation

Market segmentation is essential to both derive insights and act upon them systematically. Based on our deep investigations of the trip datasets, we understood that temporal and spatial segmentation of the trips would provide the most actionable insights for the operating taxi companies we serve. While temporal patterns of *day of the week* and *month of the year* are briefly tapped on, our emphasis in this report is on *hour-of-day* temporal patterns and locational patterns.

Other segmentation patterns that could be insightful are:

1. Fare segmentation: segment customers based on the fare they paid

2. Distance segmentation: segment customers based on the distance they traveled

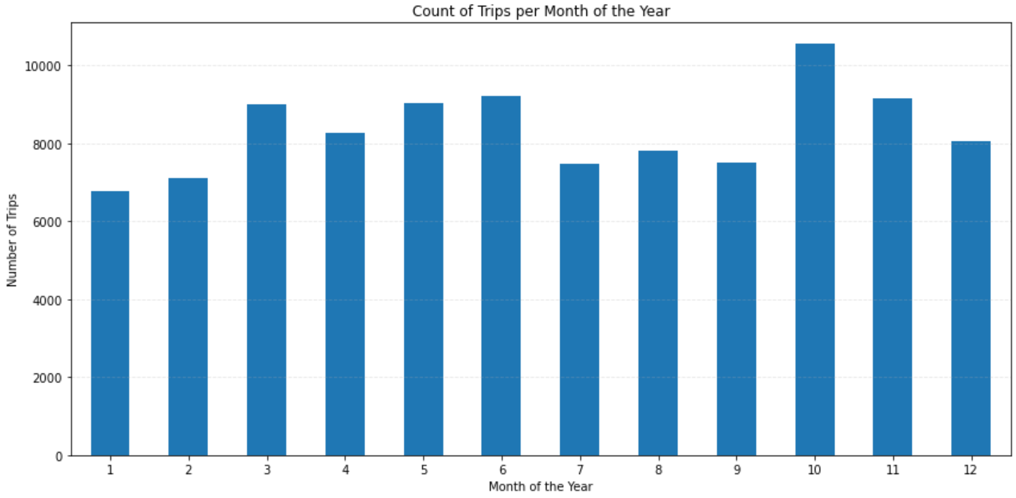
3. Payment type segmentation: segment customers based on the payment type they used

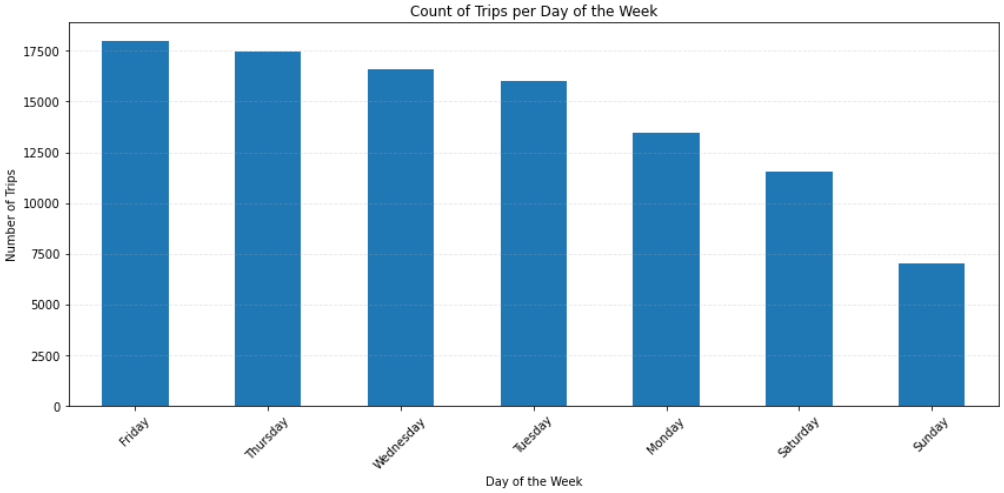
4. K means clustering: segmenting customers into statistical clusters

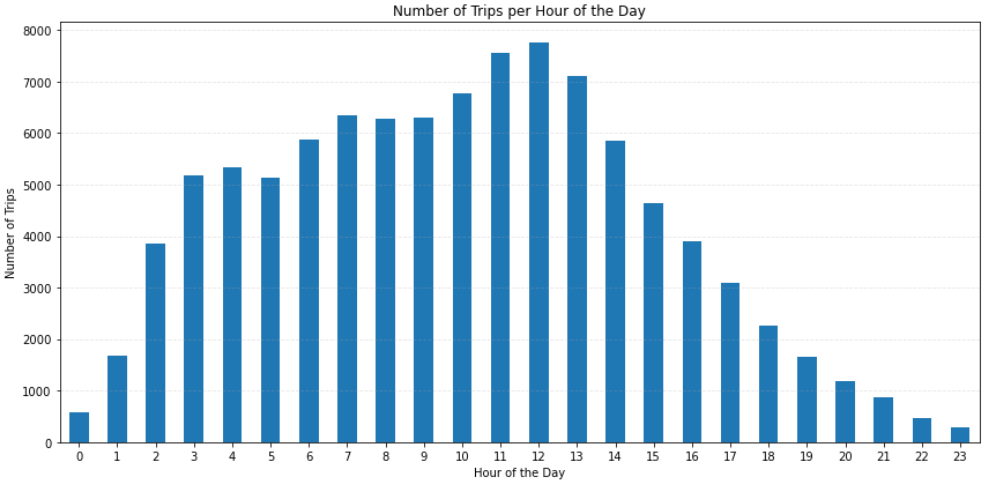
For the sake of this analysis, a sample of 100,000 travel data points has been collected between Oct 1st 2017 to 2019 and Oct 1st 2022 to 2023. This sampling period excludes the pandemic times and ensures sampling from the most recent times.

The spatial and temporal dynamics of the trips

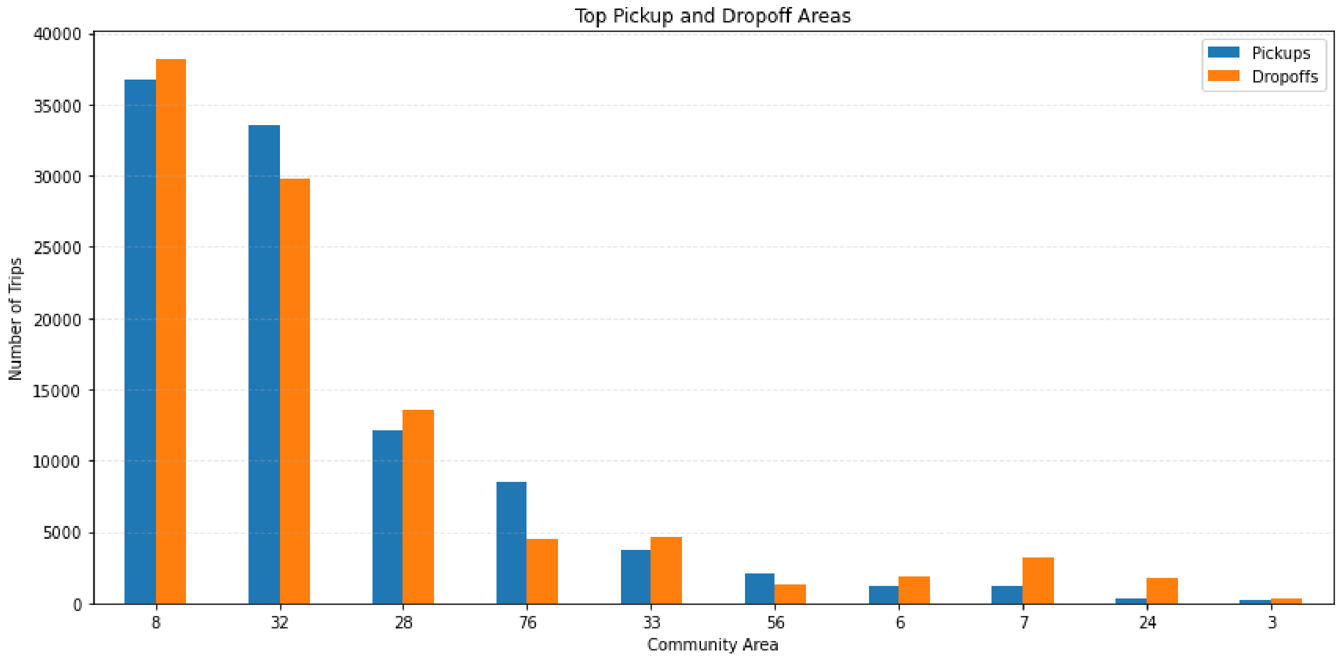
First of all, the number of trips in each month of the year, day of the week, and hour of the day are represented below.




Count of Trips per Day of the Week
Number of Trips
17500
15000
12500
10000
7500
5000
2500
0
Friday
Thursday
Wednesday
Tuesday
Monday
Saturday
Sunday
Day of the Week


Number of Trips per Hour of the Day
Number of Trips
8000
7000
6000
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour of the Day

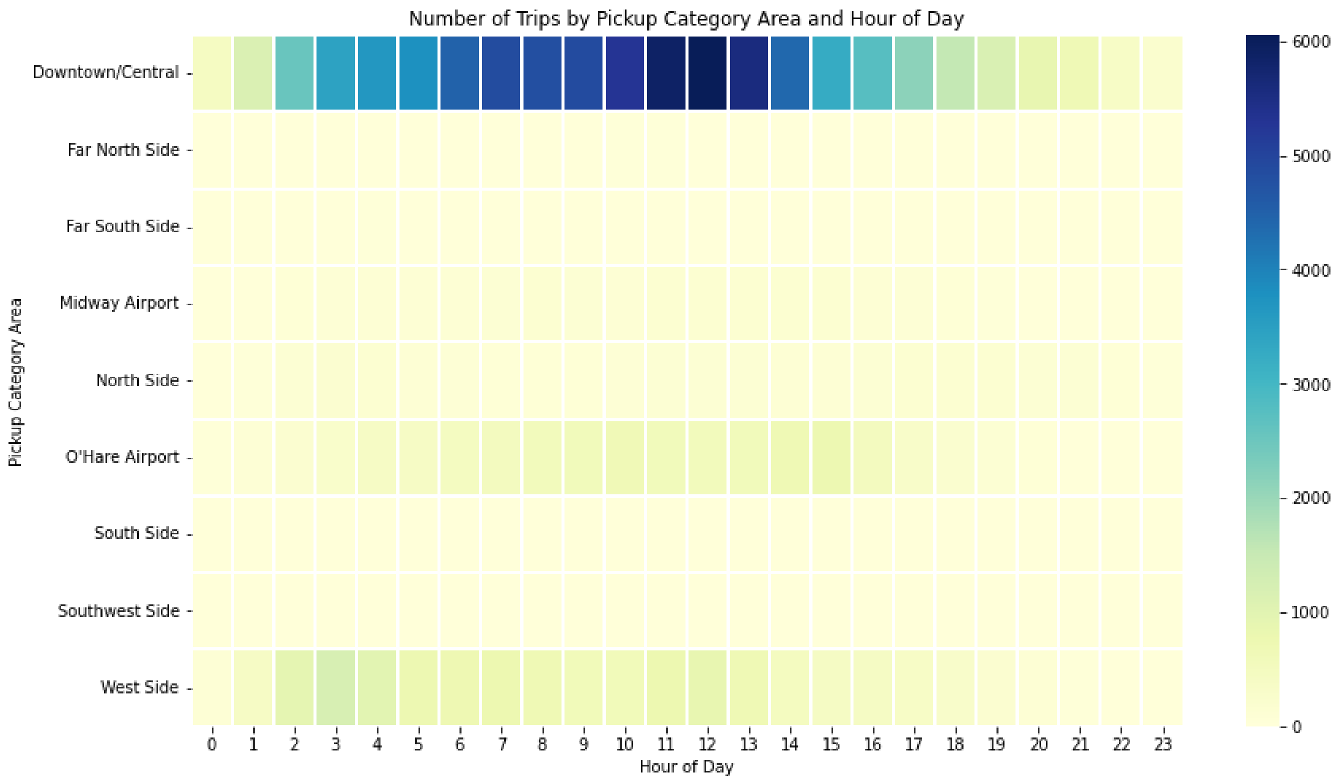
To have a better understanding of how the trips are distributed among different community areas, below are the top 10 pickup and drop-off community areas:



From now on in this report, we use a mapping of the community areas to 7 greater category areas of Chicago city defined as below:

```
'Downtown/Central': [8, 32, 33],
'North Side': [5, 6, 7, 21, 22],
'West Side': [23, 24, 25, 26, 27, 28, 29, 30, 31],
'Northwest Side': [15, 16, 17, 18, 19, 20],
"O'Hare Airport": [76],
'Midway Airport': [56],
'South Side': [1, 9, 13, 36, 40, 44, 48, 52, 57, 61, 65, 69, 73],
'Far North Side': [2, 10, 14, 37, 41, 45, 49, 53, 58, 62, 66, 70, 74],
'Far South Side': [3, 11, 34, 38, 42, 46, 50, 54, 59, 63, 67, 71, 75],
'Southwest Side': [4, 12, 35, 39, 43, 47, 51, 55, 60, 64, 68, 72, 77]
```

For that mapping, the trip distribution for times and greater category areas could be depicted as below.



Fare and tips patterns

To understand the variables that impact the fare of the taxi trips, the impact of numerical and categorical variables is investigated. Starting with numerical variables, the Pearson correlation of the numeric values is investigated. It can be seen that fare has the highest correlation with **independent** parameters gc_distance[1], trip_miles, and trip_seconds, respectively.

Variable	Pearson Correlation with ‘fare’
trip_seconds	0.513428
trip_miles	0.664957
fare	1
Miles/seconds	0.039874
gc_distance	0.674956

To understand other potential underlying dynamics of fare and tips, the normalized fare per mile and seconds of trip is analyzed.

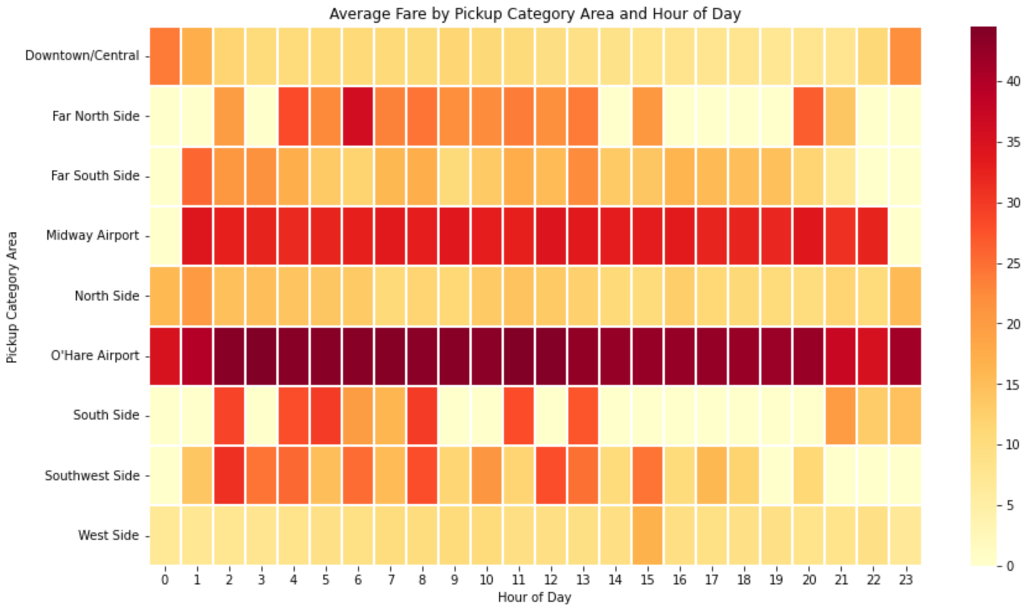
[1] gc_distance is the greater circle distance between pickup and drop-off locations. It is the length of the line between pickup and drop-off locations (calculated using pickup and drop-off coordinates).

Variable	Pearson Correlation with 'fare/miles of trip'	Pearson Correlation with 'fare/seconds of trip'
trip_seconds	-0.0106	-0.02354
trip_miles	-0.03377	0.01394
fare	0.601104	0.14174
tips	-0.00295	0.048001
tolls	-0.0003	-4.7E-05
extras	-0.00667	0.064953
trip_total	0.529854	0.137745
miles/seconds	-0.00321	0.378284
gc_distance	-0.01557	-0.00434
fare/miles	1	0.425098
fare/seconds	0.425098	1

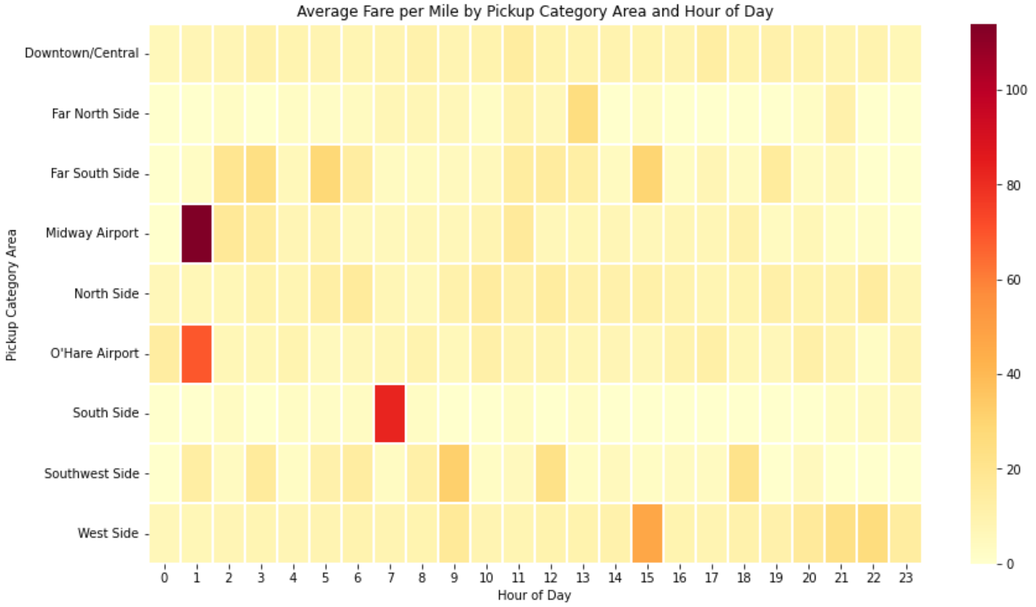
It is evident that there's no correlation between **independent** parameters and fare/miles and fare/seconds values.

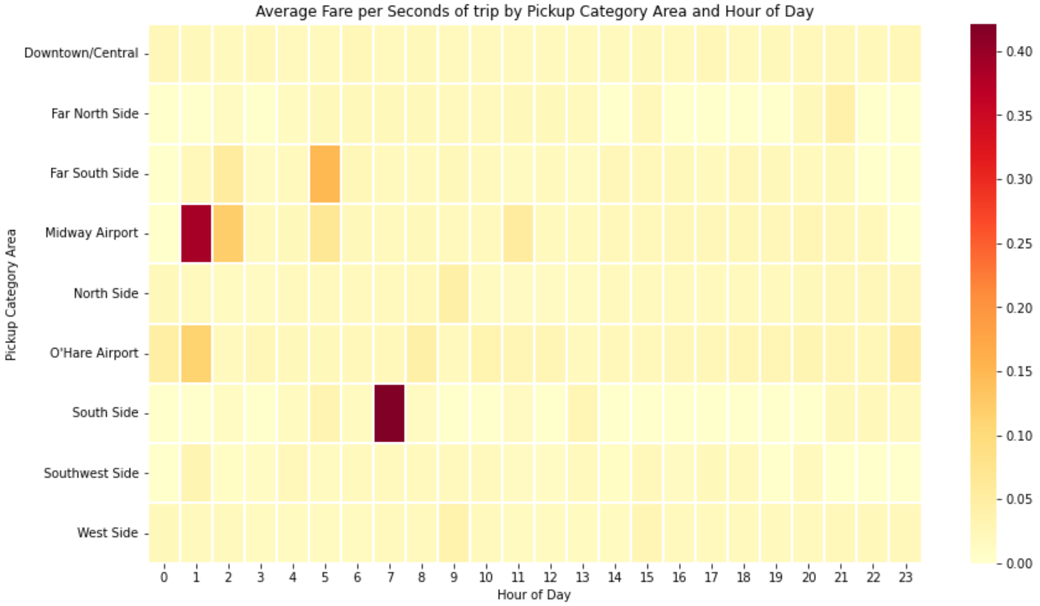
Now, let's analyze the impact of non-numerical values on the fare and tip amounts.

Below is the average fares by pickup category area and hour of day.

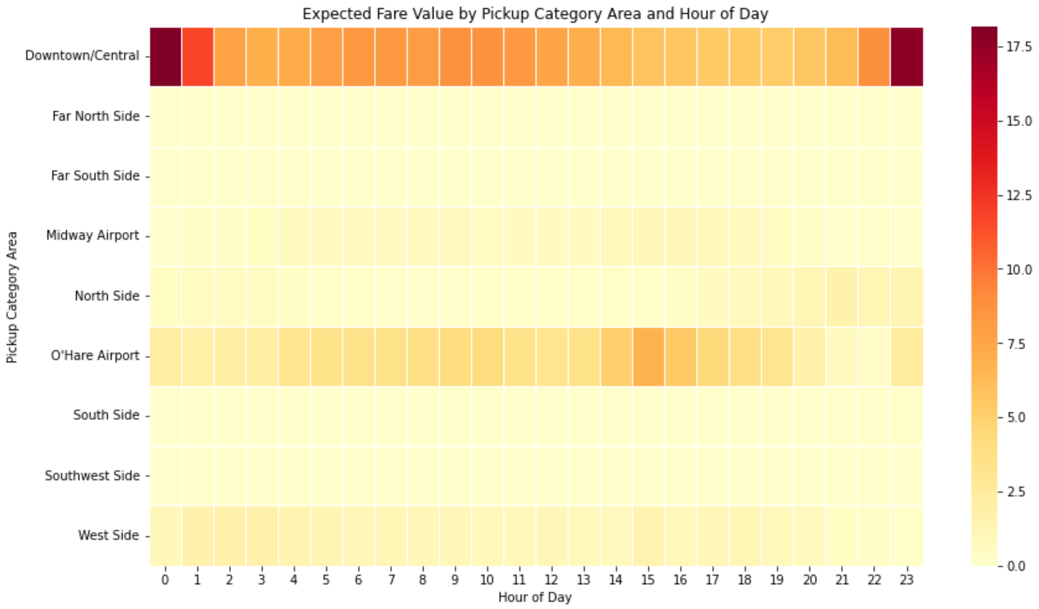


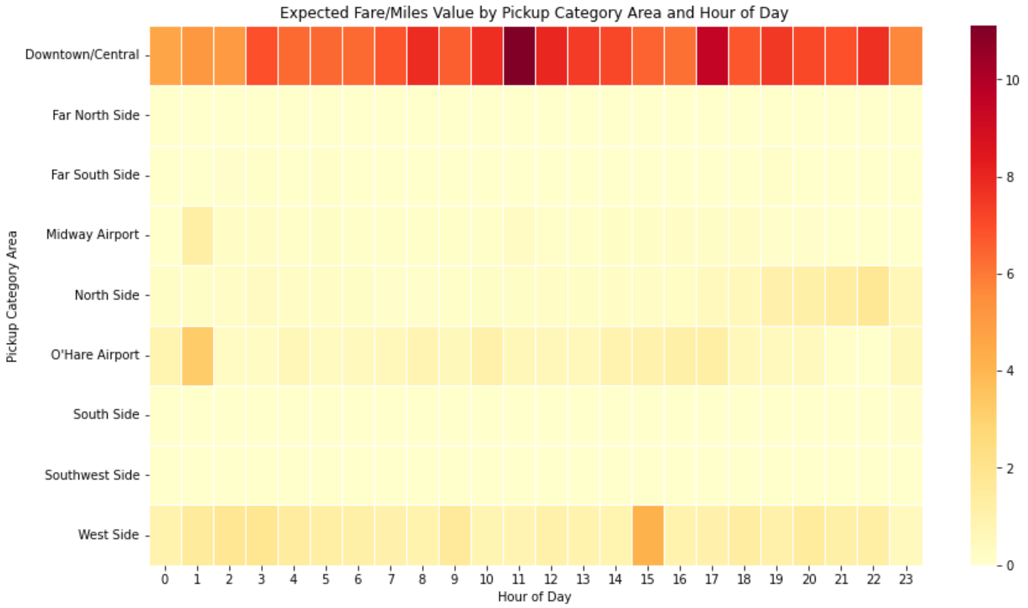
It is evident that the fare values very much reflect the miles (and seconds) of the trip. The two figures below show the normalized fare prices per trip miles and seconds.

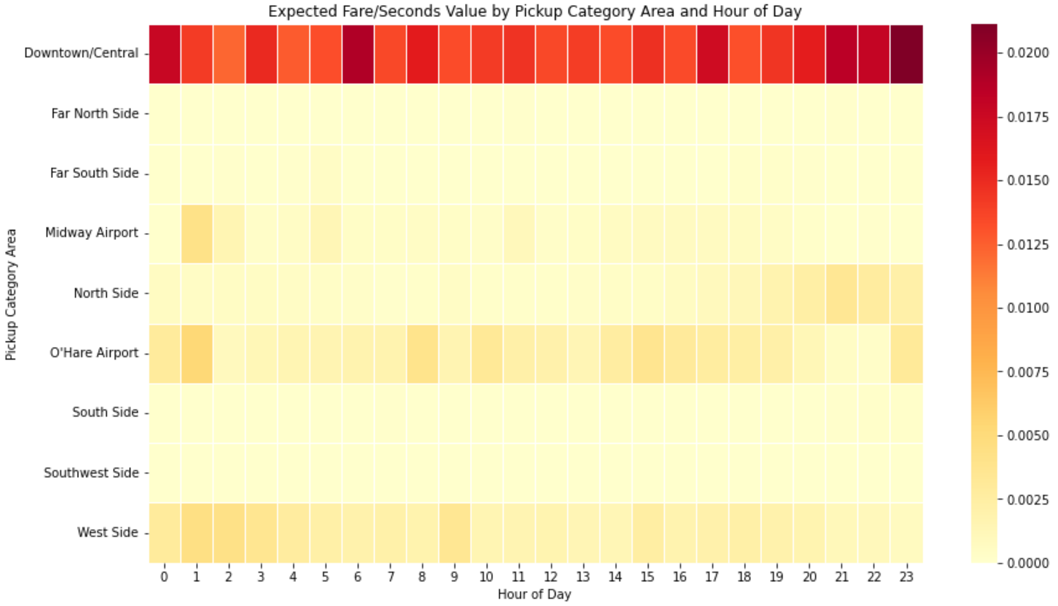




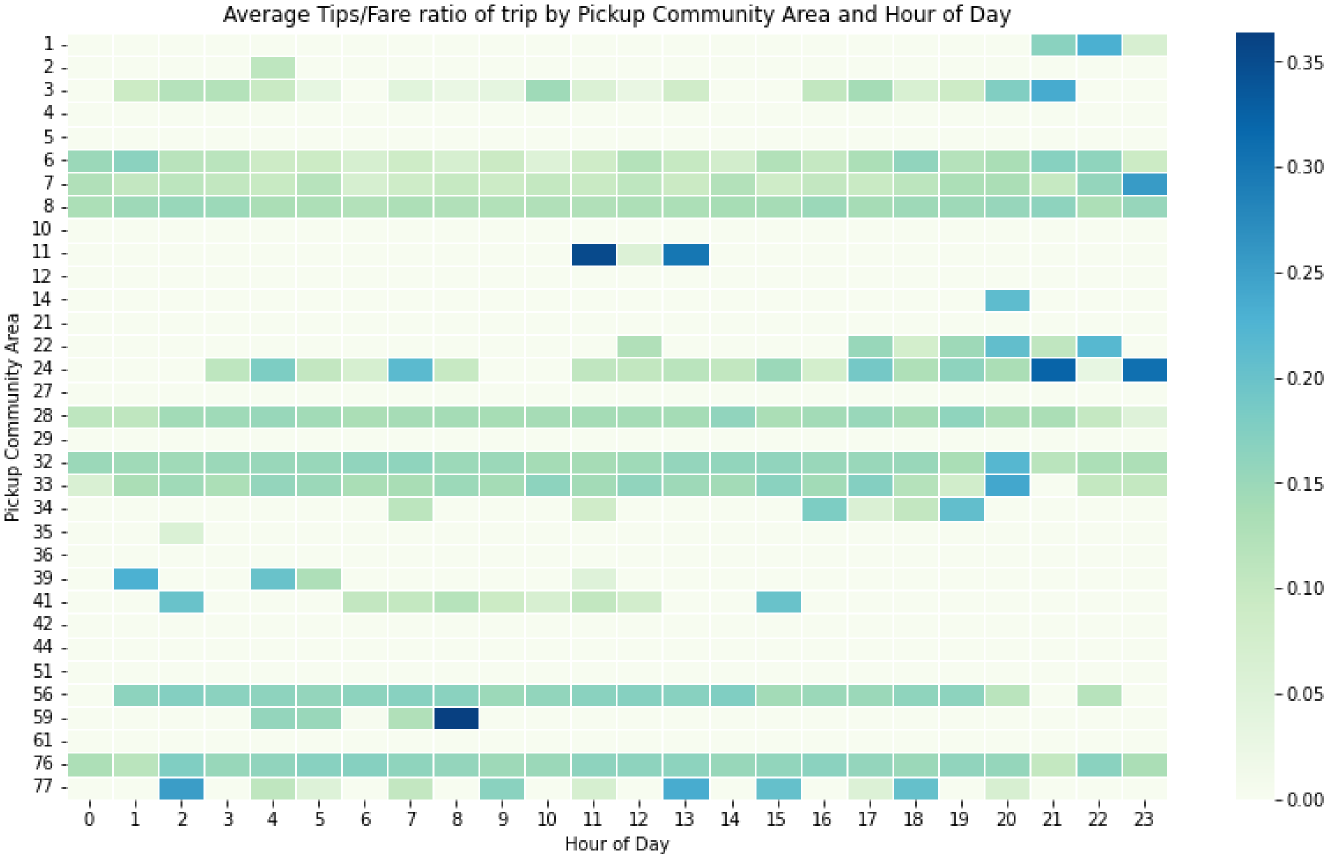
But fares and normalized fares on their own might not indicate the realized profit. The probability of having a trip is also worth considering when deciding where to put the taxis. Below are the expected fare, expected fare/miles, and expected fare/seconds of trip for category community areas at different hours of the day.


Expected Fare Value by Pickup Category Area and Hour of Day
Pickup Category Area
Downtown/Central
Far North Side
Far South Side
Midway Airport
North Side
O'Hare Airport
South Side
Southwest Side
West Side
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour of Day
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0


Expected Fare/Miles Value by Pickup Category Area and Hour of Day
Pickup Category Area
Downtown/Central
Far North Side
Far South Side
Midway Airport
North Side
O'Hare Airport
South Side
Southwest Side
West Side
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour of Day
10
8
6
4
2
0

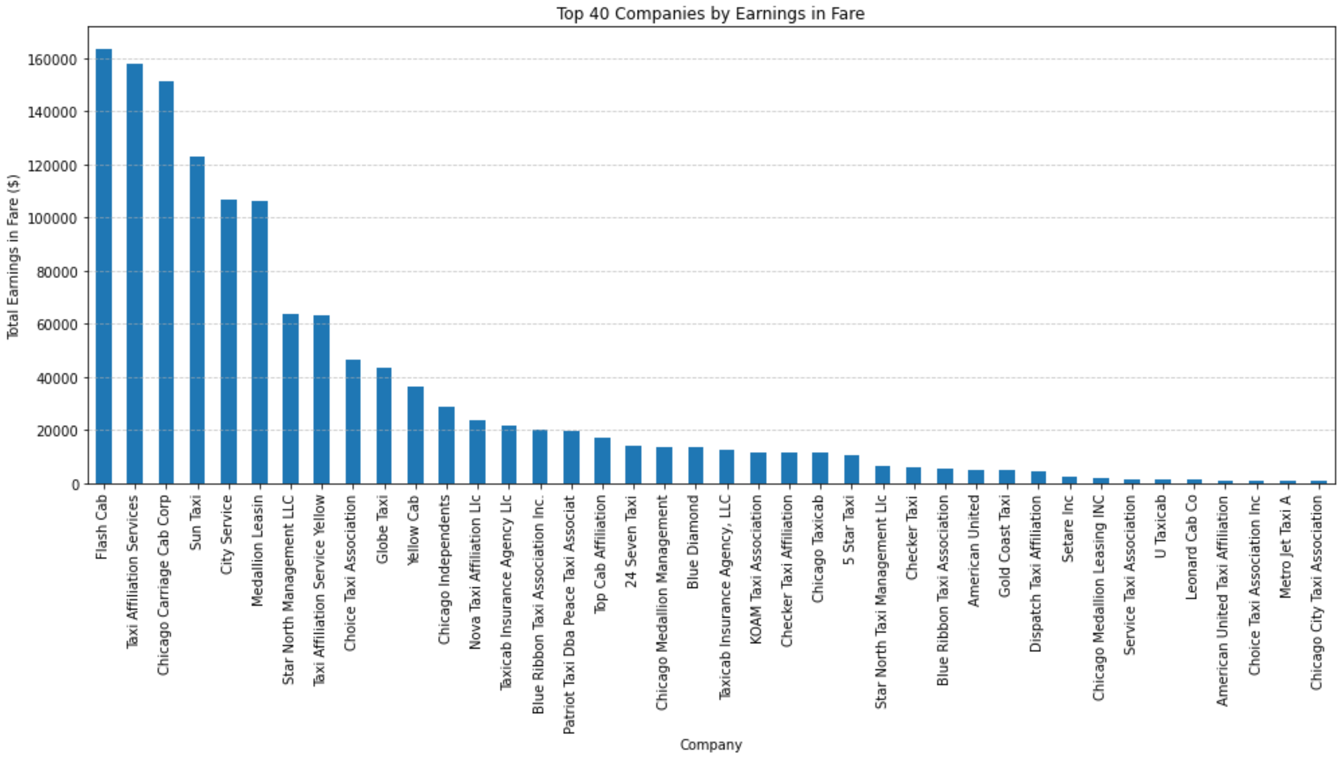


Finally, to have more detailed information on the tipping behavior of commuters, the below figure is presented.



Comparison of companies

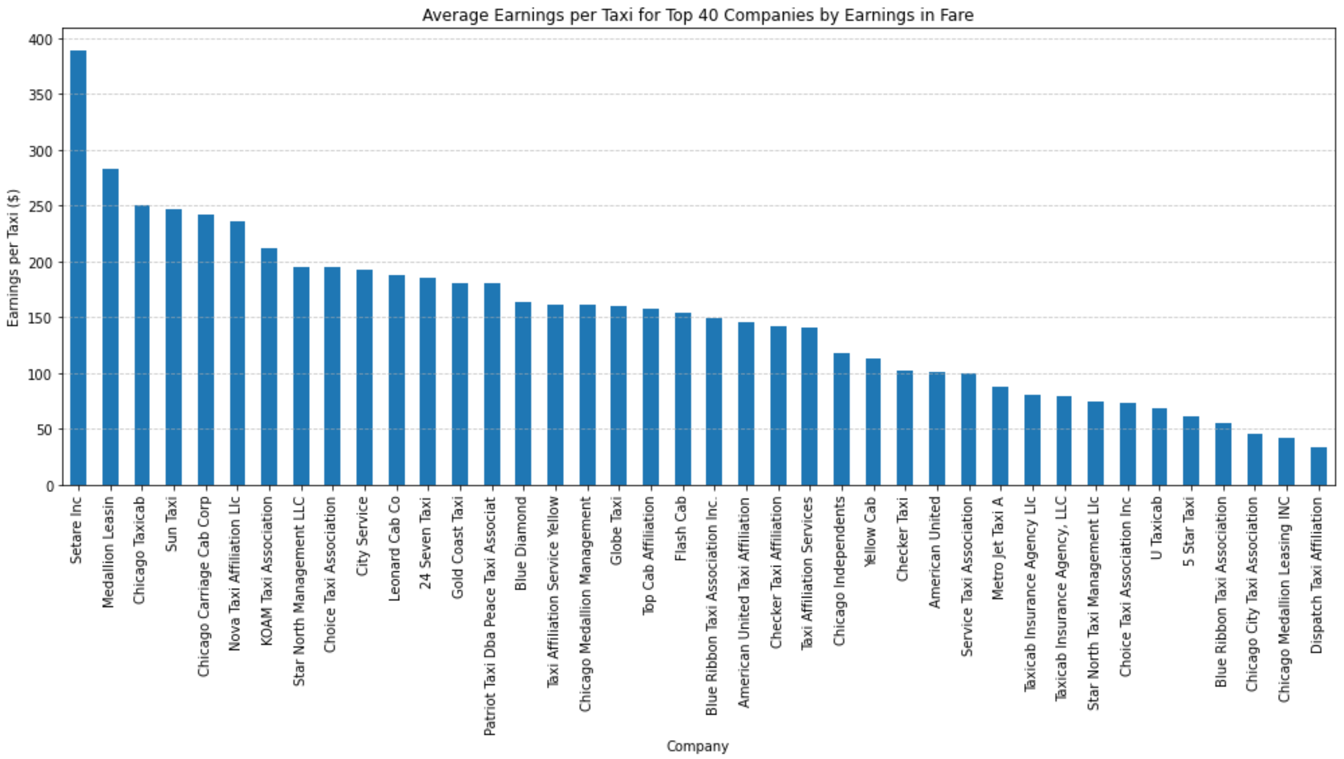
Here, we focus on the top 40 companies in terms of total earnings. The figure below shows those top 40 companies in terms of earnings by fare.

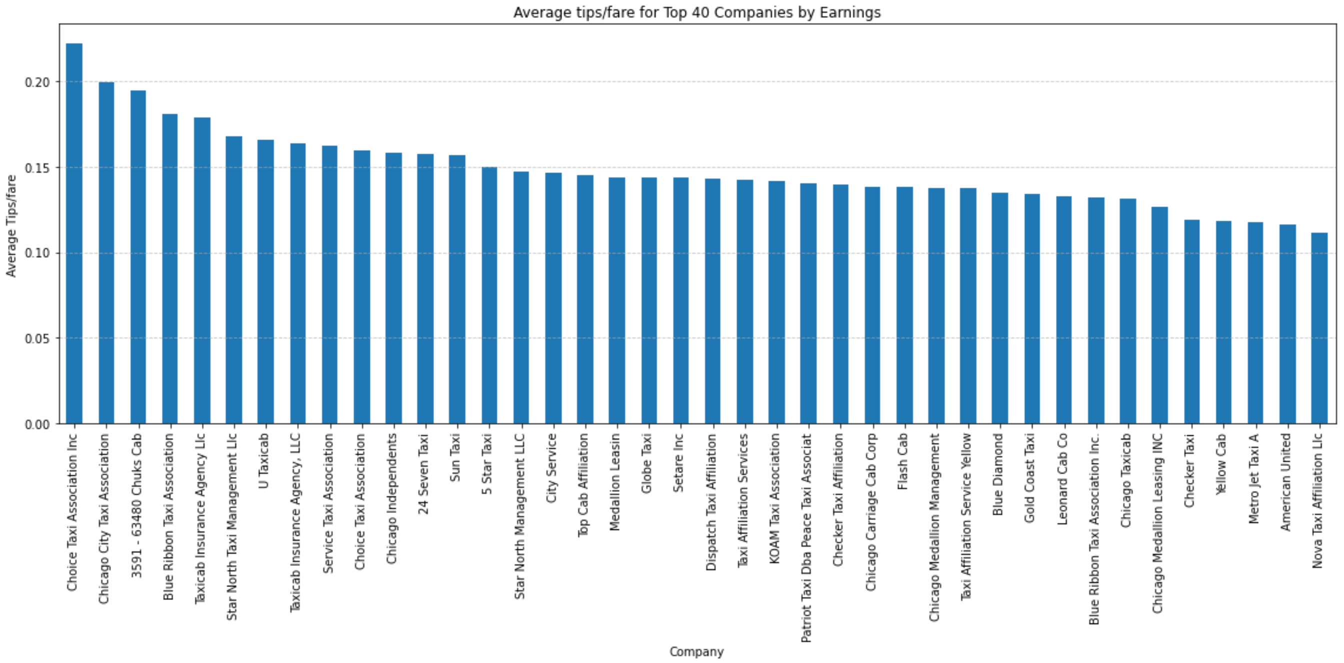


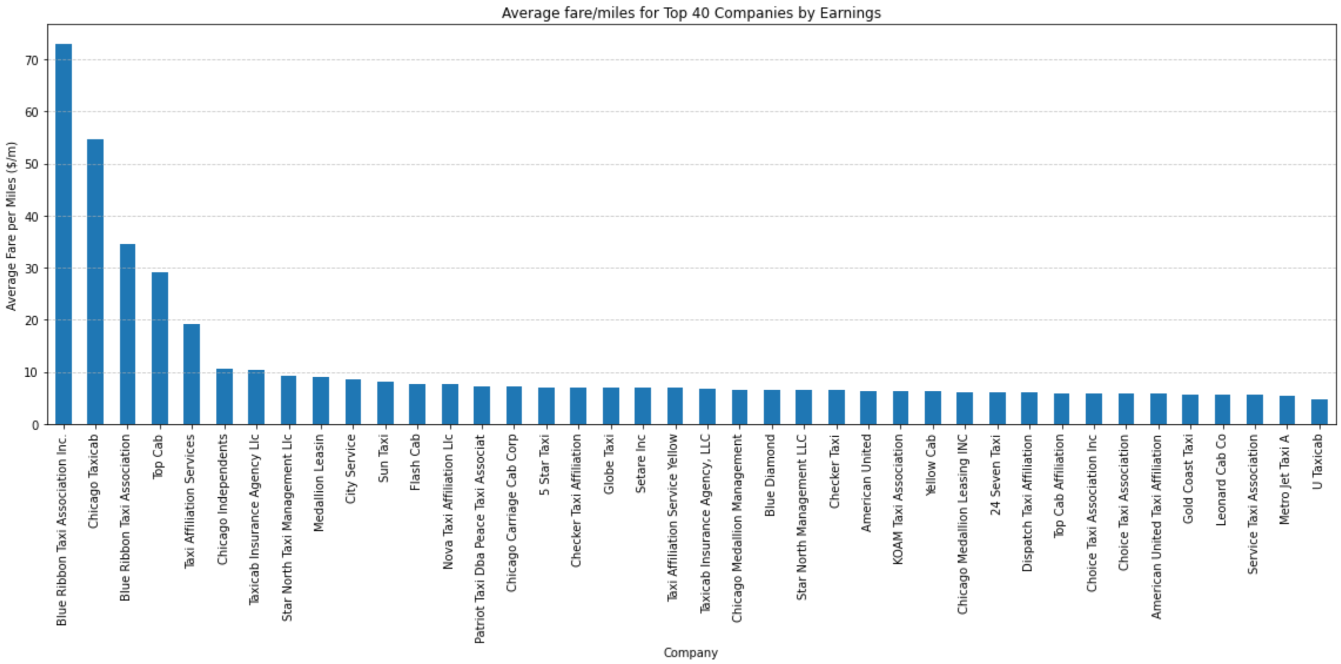
With those 40 companies as the basis of comparison, 4 different KPIs are defined for comparing the taxi companies:

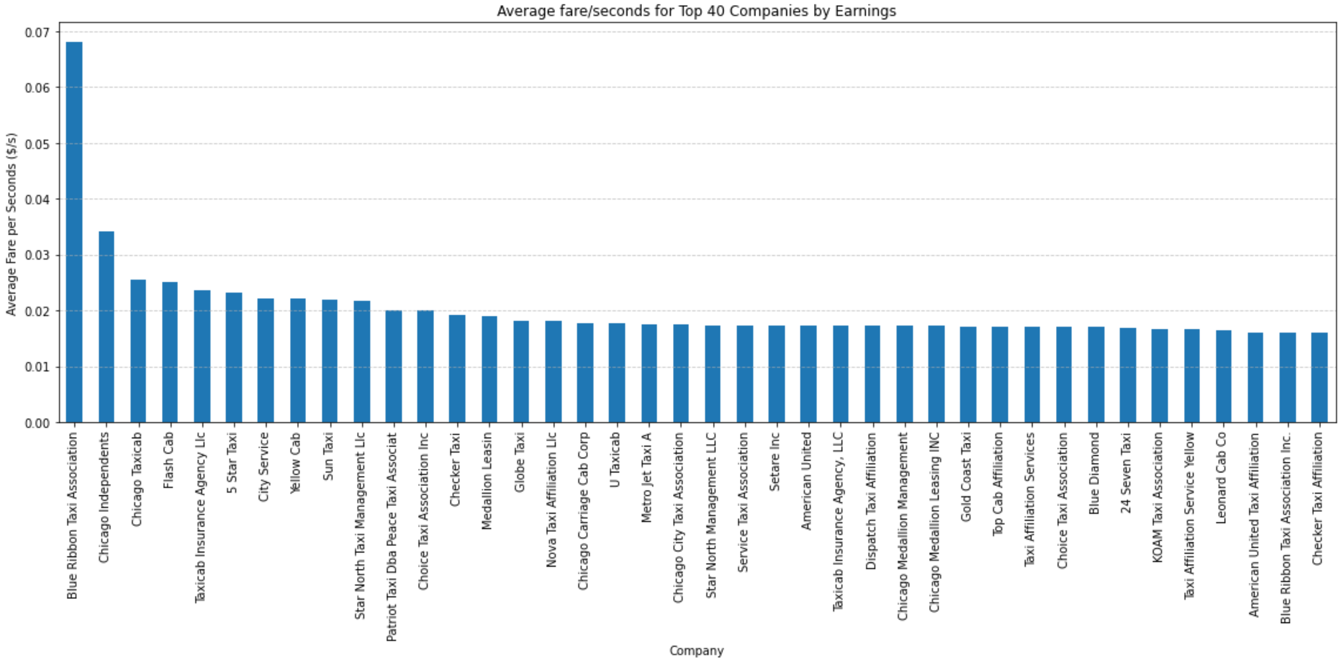
1- Company earnings per number of taxis they have (as an indicator of effective deployment of resources)
2- Average tip/fare their taxis had (as an indication of customer satisfaction)
3- Company earnings per mile of trip their taxis rode commuters (as an indicator of revenue per cost of operating)
4- Company earnings per second of their taxis were riding commuters (as an indicator of revenue per cost of operating #2)
5- Average fare per trip their taxis had (as an indicator of revenue per the chance of getting a ride)

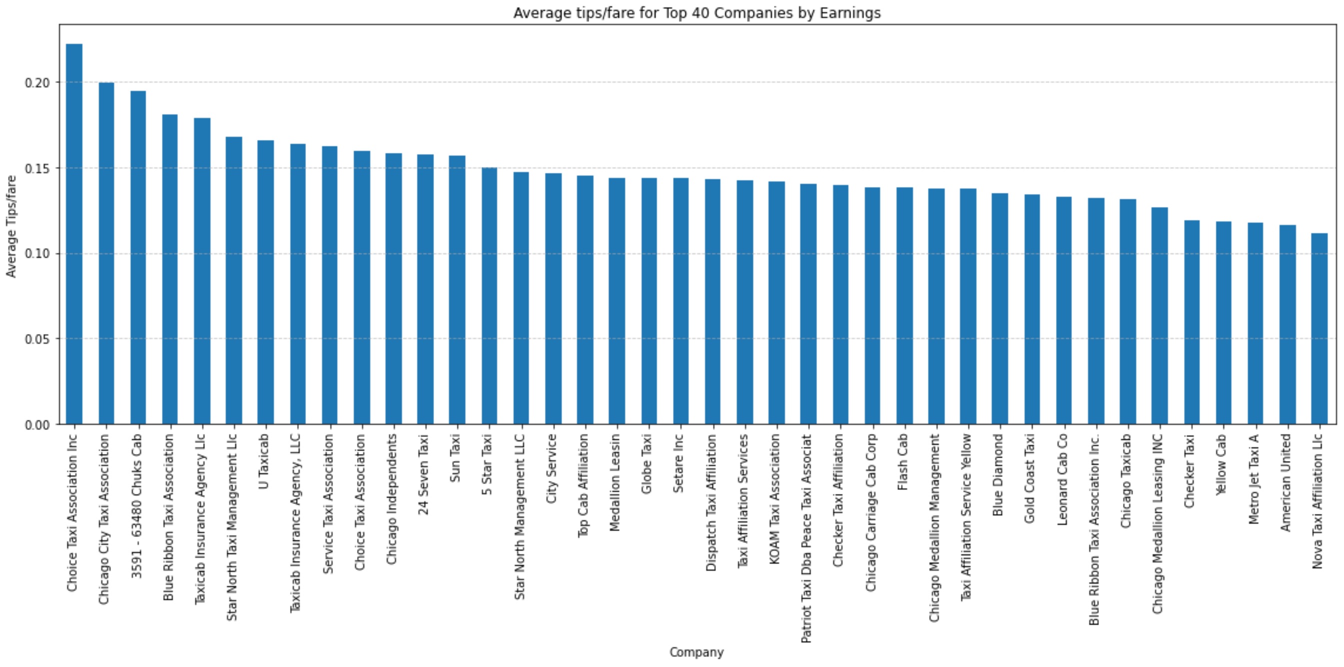
The following figures indicate those KPIs for the top 40 earning companies.


Average Earnings per Taxi for Top 40 Companies by Earnings in Fare
Earnings per Taxi ($)
400
350
300
250
200
150
100
50
0
Setare Inc
Medallion Leasin
Chicago Taxicab
Sun Taxi
Chicago Carriage Cab Corp
Nova Taxi Affiliation Llc
KOAM Taxi Association
Star North Management LLC
Choice Taxi Association
City Service
Leonard Cab Co
24 Seven Taxi
Gold Coast Taxi
Patriot Taxi Dba Peace Taxi Associat
Blue Diamond
Taxi Affiliation Service Yellow
Chicago Medallion Management
Globe Taxi
Top Cab Affiliation
Flash Cab
Blue Ribbon Taxi Association Inc.
American United Taxi Affiliation
Checker Taxi Affiliation
Taxi Affiliation Services
Chicago Independents
Yellow Cab
Checker Taxi
American United
Service Taxi Association
Metro Jet Taxi A
Taxicab Insurance Agency Llc
Taxicab Insurance Agency, LLC
Star North Taxi Management Llc
Choice Taxi Association Inc
U Taxicab
5 Star Taxi
Blue Ribbon Taxi Association
Chicago City Taxi Association
Chicago Medallion Leasing INC
Dispatch Taxi Affiliation
Company


Average tips/fare for Top 40 Companies by Earnings
Average Tips/fare
0.20
0.15
0.10
0.05
0.00
Choice Taxi Association Inc
Chicago City Taxi Association
3591 - 63480 Chuks Cab
Blue Ribbon Taxi Association
Taxicab Insurance Agency Llc
Star North Taxi Management Llc
U Taxicab
Taxicab Insurance Agency, LLC
Service Taxi Association
Choice Taxi Association
Chicago Independents
24 Seven Taxi
Sun Taxi
5 Star Taxi
Star North Management LLC
City Service
Top Cab Affiliation
Medallion Leasin
Globe Taxi
Setare Inc
Dispatch Taxi Affiliation
Taxi Affiliation Services
KOAM Taxi Association
Patriot Taxi Dba Peace Taxi Associat
Checker Taxi Affiliation
Chicago Carriage Cab Corp
Flash Cab
Chicago Medallion Management
Taxi Affiliation Service Yellow
Blue Diamond
Gold Coast Taxi
Leonard Cab Co
Blue Ribbon Taxi Association Inc.
Chicago Taxicab
Chicago Medallion Leasing INC
Checker Taxi
Yellow Cab
Metro Jet Taxi A
American United
Nova Taxi Affiliation Llc
Company


Average fare/miles for Top 40 Companies by Earnings
Average Fare per Miles ($/m)
0
10
20
30
40
50
60
70
Company
Blue Ribbon Taxi Association Inc.
Chicago Taxicab
Blue Ribbon Taxi Association
Top Cab
Taxi Affiliation Services
Chicago Independents
Taxicab Insurance Agency Llc
Star North Taxi Management Llc
Medallion Leasin
City Service
Sun Taxi
Flash Cab
Nova Taxi Affiliation Llc
Patriot Taxi Dba Peace Taxi Associat
Chicago Carriage Cab Corp
5 Star Taxi
Checker Taxi Affiliation
Globe Taxi
Setare Inc
Taxi Affiliation Service Yellow
Taxicab Insurance Agency, LLC
Chicago Medallion Management
Blue Diamond
Star North Management LLC
Checker Taxi
American United
KOAM Taxi Association
Yellow Cab
Chicago Medallion Leasing INC
24 Seven Taxi
Dispatch Taxi Affiliation
Top Cab Affiliation
Choice Taxi Association Inc
Choice Taxi Association
American United Taxi Affiliation
Gold Coast Taxi
Leonard Cab Co
Service Taxi Association
Metro Jet Taxi A
U Taxicab


Average fare/seconds for Top 40 Companies by Earnings
Average Fare per Seconds ($/s)
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Company
Blue Ribbon Taxi Association
Chicago Independents
Chicago Taxicab
Flash Cab
Taxicab Insurance Agency Llc
5 Star Taxi
City Service
Yellow Cab
Sun Taxi
Star North Taxi Management Llc
Patriot Taxi Dba Peace Taxi Associat
Choice Taxi Association Inc
Checker Taxi
Medallion Leasin
Globe Taxi
Nova Taxi Affiliation Llc
Chicago Carriage Cab Corp
U Taxicab
Metro Jet Taxi A
Chicago City Taxi Association
Star North Management LLC
Service Taxi Association
Setare Inc
American United
Taxicab Insurance Agency, LLC
Dispatch Taxi Affiliation
Chicago Medallion Management
Chicago Medallion Leasing INC
Gold Coast Taxi
Top Cab Affiliation
Taxi Affiliation Services
Choice Taxi Association
Blue Diamond
24 Seven Taxi
KOAM Taxi Association
Taxi Affiliation Service Yellow
Leonard Cab Co
American United Taxi Affiliation
Blue Ribbon Taxi Association Inc.
Checker Taxi Affiliation



Recommendations

The recommendations for the taxi companies are as follows:

1- Based on your company goals, set an objective for your dispatch strategy. As a taxi company, you might want to maximize the number of trips revenue per trip or a mix of both. Also, you might want your taxis to have trips with higher earnings per mile or seconds they were operating to minimize your cost of revenue.

2- If maximizing the number of trips is concerned, follow the guidelines[2] for dispatching your taxis to the areas with the highest number of trips in each hour of the day.

3- If maximizing the revenue is concerned, follow the guidelines for dispatching your taxis to the areas with the highest fare prices in each hour of the day.

4- If a mix of both is considered, follow the guidelines for dispatching your taxis to the areas with the highest expected fare values in each hour of the day.

5- If you want to minimize the cost of revenue, follow the guidelines for dispatching your taxis to the areas with the highest fare/miles (of fare/seconds).

6- For a mix of all conditions, consider dispatching your taxis to the areas with the highest expected fare/miles or fare/seconds.

[2] The heatmaps presented earlier in this report are considered as the guidelines for temporal and spatial dispatch.

Note 1: Always design your dispatches so that your taxis remain/move to the areas with the highest expected fare[3].

Note 2: The tipping behavior of the commuters is illustrated in the heatmap of 'Average Tips/Fare by pickup community area and hour of day'. That's a good guide for sharing with your taxi drivers to incentivize them.

Note 3: It is a good idea to follow the dispatch patterns of the companies with KPI values that are desirable to your company.

Note 4: Upon request, the guidelines with the resolution of each community area or even census tract could be provided.

The machine learning model for predicting fare prices

Based on the problem statement, there could be two different approaches to predicting the fare prices:

1- **Predicting the prices after commute**: for this price estimation procedure, all features available on the dataset are known.

2- **Predicting the prices before the commute**: this could be the subject of price estimation on the TaxiTech platform, where commuters request a taxi trip and want to have an estimation of the prices. In this case, *trip_miles* and *trip_seconds* features are not available, and it is suggested to have a two-step model where the trip_miles and trip_seconds are estimated[4], and then a machine learning model is used to estimate the price.

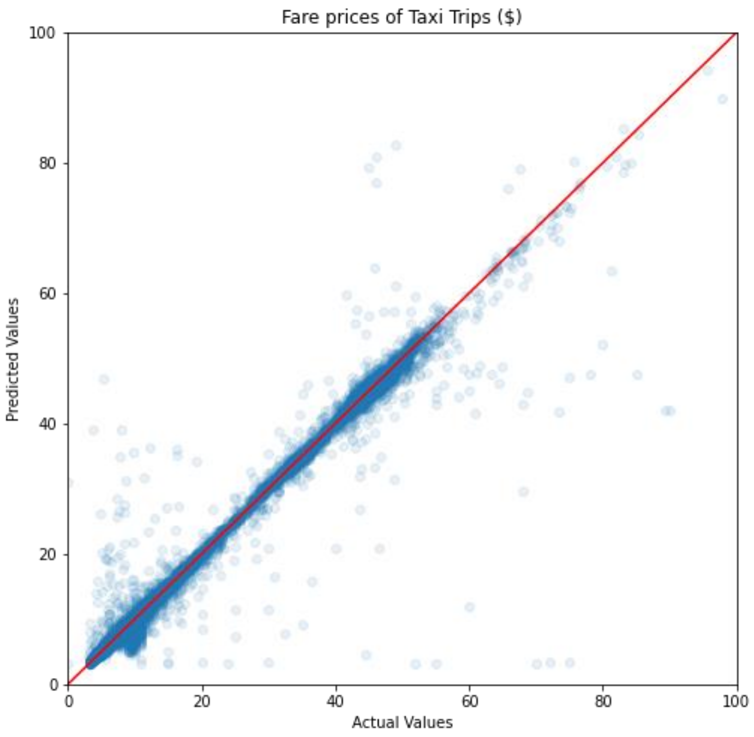
Since we don't have access to a system for *trip_miles* and *trip_seconds* estimation, given the pickup and drop-off locations, we will assume those two features given and use them in our machine learning model.

Given the amount of data available and the robustness of deep learning models on data shift, the deep learning model is used for fare price prediction here. Embedding and one-hot encoding are used to transform the categorical values. Details of the model training are available on the attached Jupyter Notebook.

The actual vs. predicted values of fare prices for a test set of 3000 data points are provided below:

[3] or any other objective your company may have

[4] This could be done using an API from Google Maps



The model has a mean absolute percentage error (MAPE) of 14% and a mean absolute error (MAE) of 0.5 $. To put the value of MAE into context, the fare price statistics are provided below.

count	30000
mean	13.60749
std	21.91856
min	0.01
25%	6
50%	7.75
75%	11.5
max	3000.75

Appendix A: Investigative SQL Queries

Below are some queries I ran on BigQuery to get a better understanding of the trip dynamics. In case of minor results, I have put the results there as well. The queries are time confined to the period of Oct 1st 2017-2019 and 2022-2022.

Total Number of Trips:

45627704

```
SELECT COUNT(*) as total_trips
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE (trip_start_timestamp >= '2017-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2019-10-01
00:00:00 UTC')
OR (trip_start_timestamp >= '2022-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2023-10-01
00:00:00 UTC');
```

Trips with coordinates (with company name):

40853037

```
SELECT COUNT(*) AS trip_count_with_coordinates
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE ((trip_start_timestamp >= '2017-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2019-10-01
00:00:00 UTC')
   OR (trip_start_timestamp >= '2022-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2023-10-01
00:00:00 UTC'))
AND pickup_latitude IS NOT NULL
AND pickup_longitude IS NOT NULL
AND dropoff_latitude IS NOT NULL
AND dropoff_longitude IS NOT NULL;
(AND company IS NOT NULL;)
```

Trips with either seconds, miles, or fare as zero:

5849027

```
SELECT COUNT(*) AS zero_metric_trips
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE ((trip_start_timestamp >= '2017-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2019-10-01
00:00:00 UTC')
   OR (trip_start_timestamp >= '2022-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2023-10-01
00:00:00 UTC'))
AND (trip_seconds = 0 OR trip_miles = 0 OR fare = 0);
```

Trip with non-zero metrics:

39764565

```
SELECT COUNT(*) AS trip_count_with_full_data

FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`

WHERE ((trip_start_timestamp >= '2017-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2019-10-01 00:00:00 UTC')

   OR (trip_start_timestamp >= '2022-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2023-10-01 00:00:00 UTC'))

AND (trip_seconds > 0 AND trip_miles > 0 AND fare > 0)
```

Trips with coordinates with non-zero metrics:

35749411

```
SELECT COUNT(*) AS trip_count_with_full_data
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE ((trip_start_timestamp >= '2017-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2019-10-01 00:00:00 UTC')
   OR (trip_start_timestamp >= '2022-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2023-10-01 00:00:00 UTC'))
AND (trip_seconds > 0 AND trip_miles > 0 AND fare > 0)
AND pickup_latitude IS NOT NULL
AND pickup_longitude IS NOT NULL
AND dropoff_latitude IS NOT NULL
AND dropoff_longitude IS NOT NULL;
```

Trips with recorded census tract and community area + nonzero metrics:

25097974

```
SELECT COUNT(*) AS trip_count_with_full_data
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE ((trip_start_timestamp >= '2017-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2019-10-01 00:00:00 UTC')
   OR (trip_start_timestamp >= '2022-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2023-10-01 00:00:00 UTC'))
AND (trip_seconds > 0 AND trip_miles > 0 AND fare > 0)

AND pickup_census_tract IS NOT NULL
AND dropoff_census_tract IS NOT NULL
AND pickup_community_area IS NOT NULL
AND dropoff_community_area IS NOT NULL;
```

Unique Taxi IDs:

7427

```
SELECT COUNT(DISTINCT taxi_id) as unique_taxi_ids
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE (trip_start_timestamp >= '2017-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2019-10-01
00:00:00 UTC')
OR (trip_start_timestamp >= '2022-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2023-10-01
00:00:00 UTC');
```

Taxi Companies:
87

```
SELECT COUNT(DISTINCT company) AS unique_company_count
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE ((trip_start_timestamp >= '2017-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2019-10-01
00:00:00 UTC')
   OR (trip_start_timestamp >= '2022-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2023-10-01
00:00:00 UTC'));
```

Taxi Companies with all records available:
82

```
SELECT COUNT(DISTINCT company) AS total_unique_companies
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE ((trip_start_timestamp >= '2017-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2019-10-01
00:00:00 UTC')
   OR (trip_start_timestamp >= '2022-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2023-10-01
00:00:00 UTC'))
AND (trip_seconds > 0 AND trip_miles > 0 AND fare > 0)
AND pickup_latitude IS NOT NULL
AND pickup_longitude IS NOT NULL
AND dropoff_latitude IS NOT NULL
AND dropoff_longitude IS NOT NULL
AND pickup_census_tract IS NOT NULL
AND dropoff_census_tract IS NOT NULL
AND pickup_community_area IS NOT NULL
AND dropoff_community_area IS NOT NULL
AND company IS NOT NULL;
```

Taxis with more than one company associated with them:

(3543 taxis)- taxi_more_than_one_company.csv

```
SELECT taxi_id, COUNT(DISTINCT company) AS num_companies
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE ((trip_start_timestamp >= '2017-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2019-10-01
00:00:00 UTC')
   OR (trip_start_timestamp >= '2022-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2023-10-01
00:00:00 UTC'))
GROUP BY taxi_id
HAVING num_companies > 1;
```

Number of companies who have bought or sold taxi medallions:

75

```
SELECT COUNT(DISTINCT company) AS num_unique_companies
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE taxi_id IN (
    SELECT taxi_id
    FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
    WHERE ((trip_start_timestamp >= '2017-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2019-10-01
00:00:00 UTC')
       OR (trip_start_timestamp >= '2022-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2023-10-01
00:00:00 UTC'))
    GROUP BY taxi_id
    HAVING COUNT(DISTINCT company) > 1
)
AND ((trip_start_timestamp >= '2017-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2019-10-01
00:00:00 UTC')
OR (trip_start_timestamp >= '2022-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2023-10-01
00:00:00 UTC'));
```

Number of unique pickup and drop off census tracts:

unique_pickup_census_tracts	unique_dropoff_census_tracts
743	1053

```
SELECT COUNT(DISTINCT pickup_census_tract) AS unique_pickup_census_tracts,
       COUNT(DISTINCT dropoff_census_tract) AS unique_dropoff_census_tracts
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE ((trip_start_timestamp >= '2017-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2019-10-01
00:00:00 UTC')
   OR (trip_start_timestamp >= '2022-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2023-10-01
00:00:00 UTC'))
AND (trip_seconds > 0 AND trip_miles > 0 AND fare > 0);
```

Number of unique pickup and drop off community areas:

unique_pickup_community_areas	unique_dropoff_community_areas
77	77

```
SELECT COUNT(DISTINCT pickup_community_area) AS unique_pickup_community_areas,
       COUNT(DISTINCT dropoff_community_area) AS unique_dropoff_community_areas
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE ((trip_start_timestamp >= '2017-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2019-10-01
00:00:00 UTC')
   OR (trip_start_timestamp >= '2022-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2023-10-01
00:00:00 UTC'))
AND (trip_seconds > 0 AND trip_miles > 0 AND fare > 0);
```

Appendix B: Sampling SQL Queries

For the sake of this task, I just queried the clean data with the least possible missing data.

A random Sample

```
SELECT *
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE ((trip_start_timestamp >= '2017-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2019-10-01 00:00:00 UTC')
   OR (trip_start_timestamp >= '2022-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2023-10-01 00:00:00 UTC'))
AND (trip_seconds > 0 AND trip_miles > 0 AND fare > 0)
AND pickup_latitude IS NOT NULL
AND pickup_longitude IS NOT NULL
AND dropoff_latitude IS NOT NULL
AND dropoff_longitude IS NOT NULL
AND pickup_census_tract IS NOT NULL
AND dropoff_census_tract IS NOT NULL
AND pickup_community_area IS NOT NULL
AND dropoff_community_area IS NOT NULL
AND company IS NOT NULL
ORDER BY RAND()
LIMIT 100000;
```

Stratified for Time-based sampling

```
WITH ranked_data AS (
  SELECT *,
         ROW_NUMBER() OVER (PARTITION BY DATE(trip_start_timestamp) ORDER BY RAND()) AS rn
  FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
  WHERE ((trip_start_timestamp >= '2017-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2019-10-01 00:00:00 UTC')
     OR (trip_start_timestamp >= '2022-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2023-10-01 00:00:00 UTC'))
  AND (trip_seconds > 0 AND trip_miles > 0 AND fare > 0)
  AND pickup_latitude IS NOT NULL
  AND pickup_longitude IS NOT NULL
  AND dropoff_latitude IS NOT NULL
  AND dropoff_longitude IS NOT NULL
  AND pickup_census_tract IS NOT NULL
  AND dropoff_census_tract IS NOT NULL
  AND pickup_community_area IS NOT NULL
  AND dropoff_community_area IS NOT NULL
  AND company IS NOT NULL
)
SELECT *
FROM ranked_data
WHERE rn <= 100
ORDER BY trip_start_timestamp;
```

Sampling for balanced number of companies

```sql
WITH ranked_data AS (
  SELECT *,
         ROW_NUMBER() OVER (PARTITION BY company ORDER BY RAND()) AS rn
  FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
  WHERE ((trip_start_timestamp >= '2017-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2019-10-01
00:00:00 UTC')
     OR (trip_start_timestamp >= '2022-10-01 00:00:00 UTC' AND trip_start_timestamp <= '2023-10-01
00:00:00 UTC'))
  AND (trip_seconds > 0 AND trip_miles > 0 AND fare > 0)
  AND pickup_latitude IS NOT NULL
  AND pickup_longitude IS NOT NULL
  AND dropoff_latitude IS NOT NULL
  AND dropoff_longitude IS NOT NULL
  AND pickup_census_tract IS NOT NULL
  AND dropoff_census_tract IS NOT NULL
  AND pickup_community_area IS NOT NULL
  AND dropoff_community_area IS NOT NULL
  AND company IS NOT NULL
)
SELECT *
FROM ranked_data
WHERE rn <= 1220  -- This is coming from 100,000 (number of required samples) / 82 (number of
companies)
ORDER BY company, rn;
```