Chicago City’s Taxi Ride Analysis

- Final Project –

Main Insights From Chicago’s Taxi Data For 2023

Guy Ben Eli - Guybeneli@gmail.com
Shai Belfer - Shaibelfer@gmail.com
Miriam Feldman - Mirimir59@gmail.com
Karwan Khlaileh - Krwan.kh310@gmail.com
Baruch Shapira - Baruchshapira100m@gmail.com

Review

Chicago is one of the largest cities in the United States, with over 9.6 million people living in its metro area. Taxis are one of the main modes of transportation in the city, although they compete with public transportation including buses and the famous “L” trains, cycling paths, and private cars. The competition has become fiercer in the last decade with the entry of ride sharing services like Uber and Lyft.

Sources

Chicago Map - https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6

2022 TIGER/Line® Shapefiles: Census Tracts – https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2022&layergroup=Census+Tracts

Community Areas in Chicago - https://en.wikipedia.org/wiki/Community_areas_in_Chicago

Scope of Work

In this project we analyzed the data from taxi trips taken in the city of Chicago between January and October 2023.

We can not account for the competition to taxis shown by public transportation, Uber and Lyft, that could potentially reduce demand. We also cannot control for disruptions in those services (such as strikes) that would induce demand for taxis, or control for traffic events and weather that could affect the demand either way.

In addition, in the section analyzing veteran drivers, we could not check which drivers began work before 2023.

However, from the data we do have we can make conclusions regarding the individual companies and their relation to the areas of the city. We can also analyze the consumer's preferred payment modes, as well as the attributes of the individual drivers regarding their length of service and their revenue.

Research Questions

1. Citizens pay in different ways for taxis. What is the preferred payment method for the citizen? Our goal is to increase the use of taxis in Chicago and overall customer satisfaction.

2. The City of Chicago believes that a veteran driver provides better results. What is the average lifetime value of individual drivers? Is it affected by the driver's veterancy? Does it change between companies? What is the retention rate of drivers over the year?

3. Which taxi company in Chicago controls the most space in the city? What is their revenue? Additionally, are there specific companies that dominate particular areas within Chicago?

KPI & Measures

1. **Total_Space** — How many miles were travelled in total by each company.
2. **Total_Revenue** — How much each company made in total.
3. **Average_Rate** — The average customer rating of each company per area.
4. **Number_Of_Trips** — How many times each company made a pickup from that area.
5. **TOTAL_areas** — The number of areas each company operates in.
6. **Retention_rate** — The percentage of each company's drivers who continued to work for that company in the next month. This was calculated by a table that counted the **Number of workers** in each **company** at the **end of each month** (marked as A), the same for the previous month (marked as B), and the number of drivers that joined the company in that month (marked as C). The retention rate was the result of the calculation (A-C)/B, accounting for NULLs, zeros etc.
7. **Veterancy_in_Days** — the number of days the driver worked during the year (for any company), calculated as the difference in days between the earliest trip for that driver and the last trip for that driver.

Data List

Original File Name	SQL Table / Python Result File Name	Record Count	Record Count After Transformation
Taxi_Trips_2023_total.csv	Table_SQL.sql	5,502,739	5,487,880
Chicago Community areas.geojson		77	77
descriptive statistics.xlsx		22	22

Preparing Data

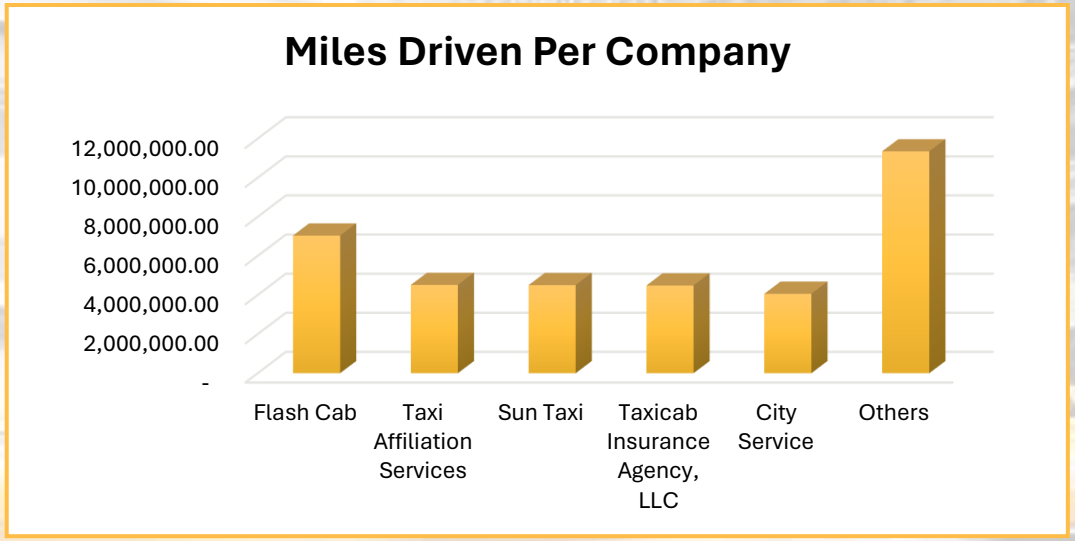
We have removed some of the entries in this data that we considered to be errors: rides where the number of seconds, miles, taxi_id or end timestamp was NULL or zero. We have also deleted entries where the amount paid by customers (trip_total) was “0”, but the payment type wasn’t marked as “No Charge”.

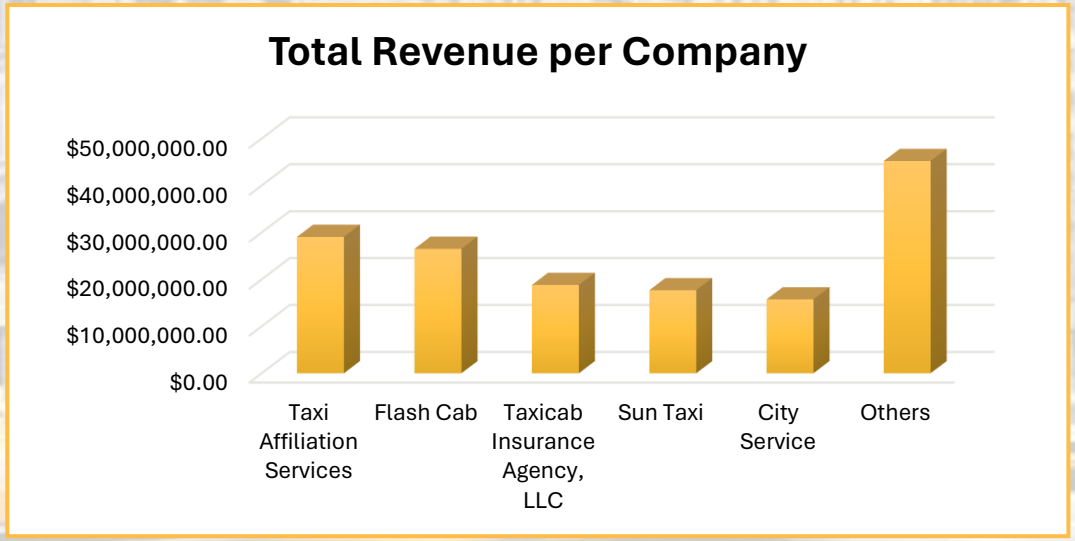
In addition, we have changed some of the company's names that we have safely assumed to be duplicates of other entries, such as “Taxicab Insurance Agency Llc” (as opposed to “Taxicab Insurance Agency, LLC”.)

```sql
--the rates of each company (simple for statistics)
SELECT Company,
       ROUND(AVG(customer_Rate), 2) AS Average_Rate
FROM (
    SELECT Company,
           AVG(customer_Rate) AS customer_Rate,
           ROW_NUMBER() OVER (PARTITION BY Company ORDER BY COUNT(*) DESC) AS rnk
    FROM [dbo].[taxi_trips_2023_total]
    WHERE Company IS NOT NULL AND customer_Rate IS NOT NULL
    GROUP BY Company
) AS ranked
WHERE rnk = 1
GROUP BY Company
ORDER BY Average_Rate DESC;
```

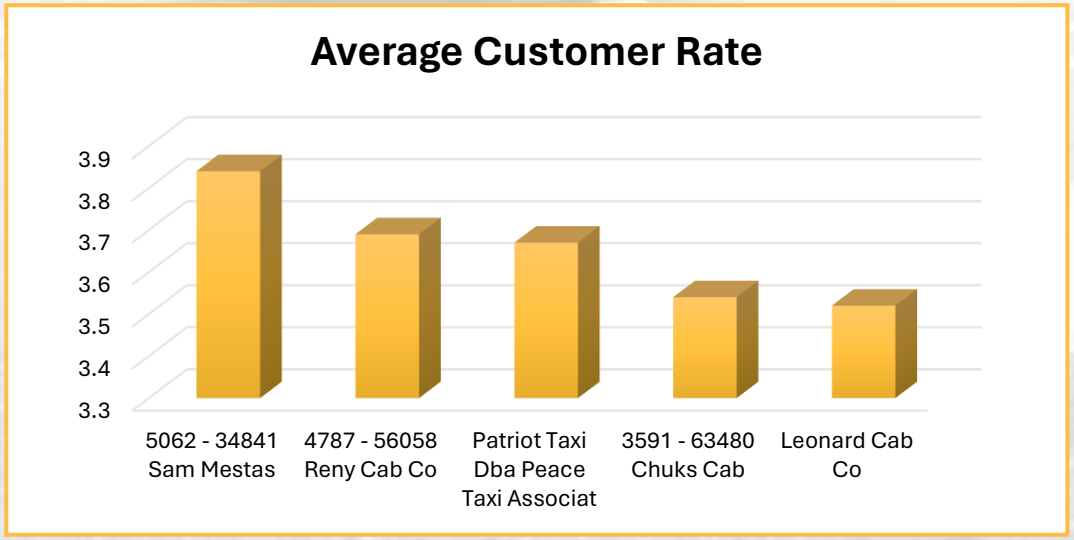
```sql
-- Delete column1
ALTER TABLE taxi_trips_2023_total
DROP COLUMN column1;

--Delete entries where the Taxi ID, Trip Total, Trip End Timestamp, Trip Miles or Trip Seconds are Null
DELETE FROM taxi_trips_2023_total
WHERE Taxi_ID IS NULL
   OR Trip_Total IS NULL
   OR Trip_End_Timestamp IS NULL
   OR Trip_Miles IS NULL
   OR Trip_Seconds IS NULL;

--Remove commas from the Trip Seconds column so it will clearly read as an INT
UPDATE taxi_trips_2023_total
SET "Trip_Seconds" = REPLACE("Trip_Seconds", ',', '')
WHERE "Trip_Seconds" LIKE '%,%';

-- Delete rows from taxi_trips_2023_total where Trip total is 0 and Payment time is not "no charge"
DELETE FROM taxi_trips_2023_total
WHERE
    Trip_Total = 0
    AND Payment_Type <> 'No Charge';
```

Descriptive Statistics





Average Company Revenue
4,227,270.31$
Average Miles Per Company
1,088,519.90
Average Rate Per Company
3.3557



Analysis of Preferred Payment Method

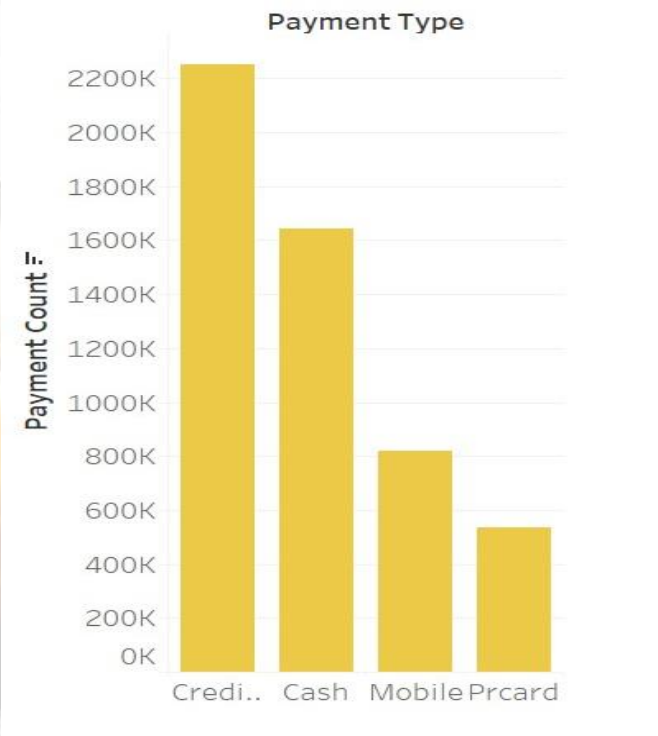
Preview:

1. What is the most common payment type?
2. Is there is any connection between payment type and customer rating?
3. Did the drivers with the highest income used more specific payment method than other?
4. How many times was each payment type used?

KPI & Measures:

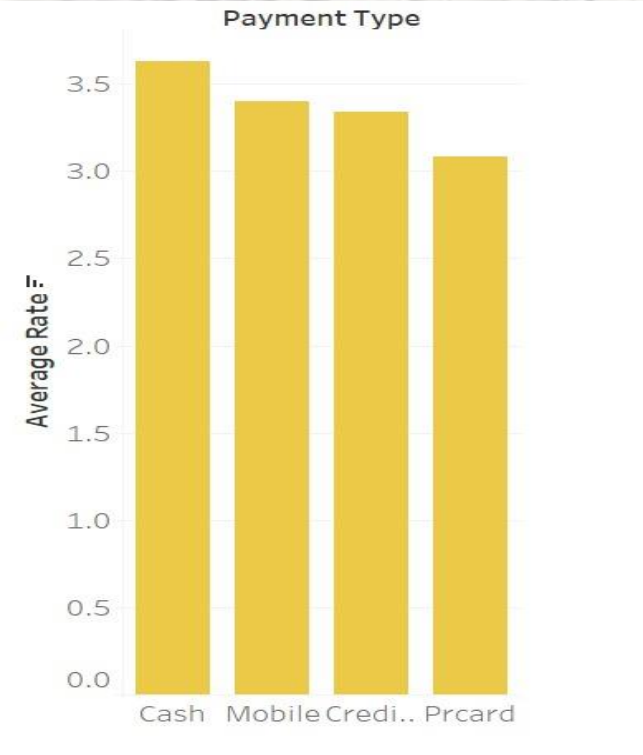
- The average customer rate for each payment type.
- The total income of the 20 drivers with the highest income.
- How many times each of those drivers used each individual payment method.

Payment Count For Each Payment Type



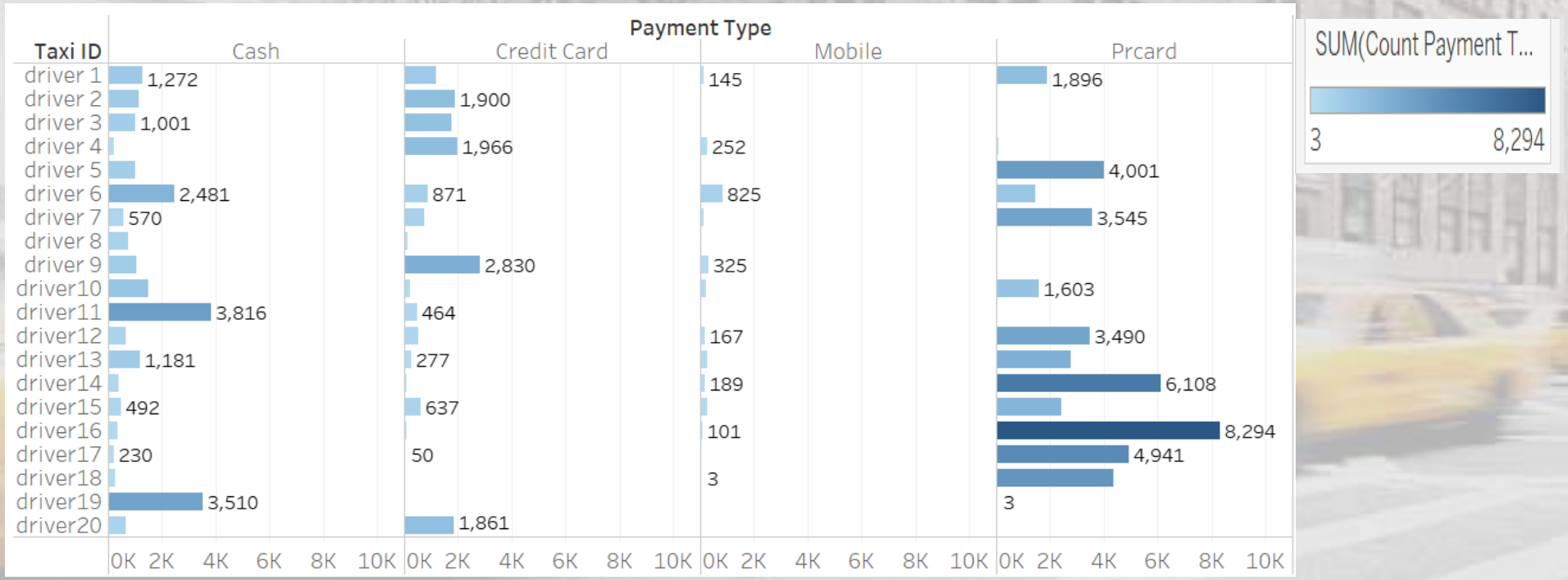
This graph shows that the most popular payment method for customers is “Credit Card”.

Average Rating Per Payment Method



This graph describes the average rate of customer satisfaction per payment type. It seems that all the methods are highly rated, with minor differences (all above 3.0), but the highest was “Cash”.

Total Revenue According To Payment Method Per Driver



After finding the top 20 paid drivers, the data shows the highest paid drivers used “Prcard” the most.

Query:

```sql
CREATE TABLE Taxi_Payment_Counts (
    Taxi_ID INT,
    Payment_Type VARCHAR(50),
    Count_Payment_Type INT
);

INSERT INTO Taxi_Payment_Counts (Taxi_ID, Payment_Type, Count_Payment_Type)
SELECT
    Taxi_ID,
    Payment_Type,
    COUNT(Payment_Type) AS Count_Payment_Type
FROM
    taxi_trips_2023_total
WHERE
    Payment_Type NOT IN ('Unknown', 'Dispute', 'No Charge') AND
    Taxi_ID IN (
        SELECT TOP 20 Taxi_ID
        FROM taxi_trips_2023_total
        WHERE Payment_Type NOT IN ('Unknown', 'Dispute', 'No Charge')
        GROUP BY Taxi_ID
        ORDER BY SUM(Trip_Total) DESC
    )
GROUP BY
    Taxi_ID,
    Payment_Type
ORDER BY Taxi_ID;
```

```sql
CREATE TABLE payment_summary (
    Payment_Type VARCHAR(50),
    payment_count INT
);

INSERT INTO payment_summary (Payment_Type, payment_count)
SELECT  Payment_Type, COUNT(*) AS payment_count
FROM taxi_trips_2023_total
WHERE Payment_Type NOT IN ('Unknown', 'Dispute', 'No Charge')
GROUP BY Payment_Type
ORDER BY payment_count DESC;
```

```sql
CREATE TABLE payment_type_average_rate (
    Payment_Type VARCHAR(50),
    Average_Rate DECIMAL(5, 2)
);

INSERT INTO payment_type_average_rate (Payment_Type, Average_Rate)
SELECT Payment_Type, AVG(customer_rate) AS Average_Rate
FROM dbo.taxi_trips_2023_total
WHERE Payment_Type NOT IN ('Unknown', 'Dispute', 'No Charge')
GROUP BY Payment_Type
ORDER BY AVG(customer_rate) DESC;
```

Conclusion

ALTHOUGH THE MOST USED PAYMENT METHOD WAS BY CREDIT CARD, IT APPEARS THAT THE HIGHEST REVENUE HAS BEEN OBTAINED FROM THE “PRCARD” METHOD. IT SEEMS THAT CUSTOMERS THAT HAVE “PRCARDS” USE TAXIS THE MOST.

FREQUENT CUSTOMERS PREFERRED PAYING BY CREDIT CARD, BUT IT WASN’T THE MOST CONVENIENT METHOD. CASH WAS HIGHLY RATED, WHICH MEANS IT IS VERY SUITABLE FOR VARIOUS TYPES OF CUSTOMERS.

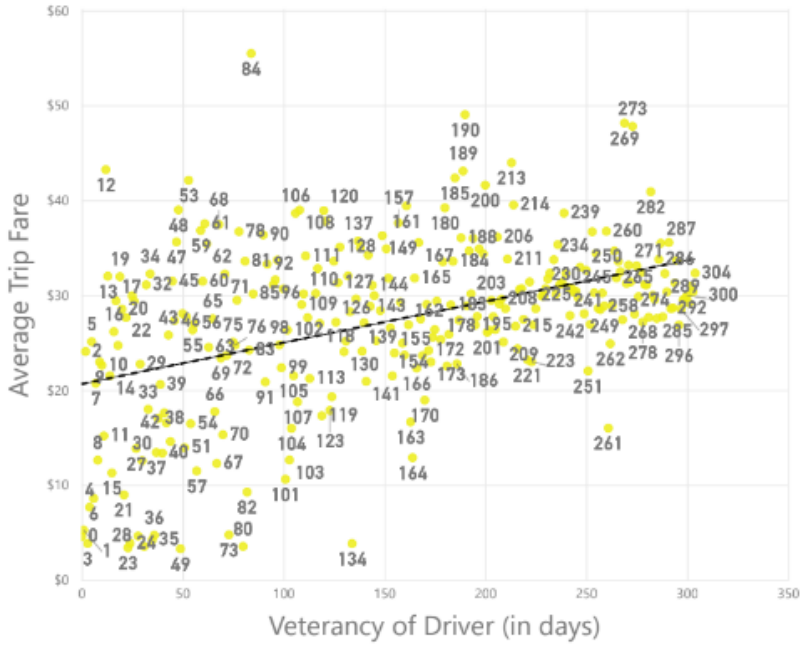
Driver Veterancy and Retention

The following graphs show a comparison between the total amount of days a driver has spent on the road and how much the same driver earned on average per trip.
There is a positive correlation, but not a large one.

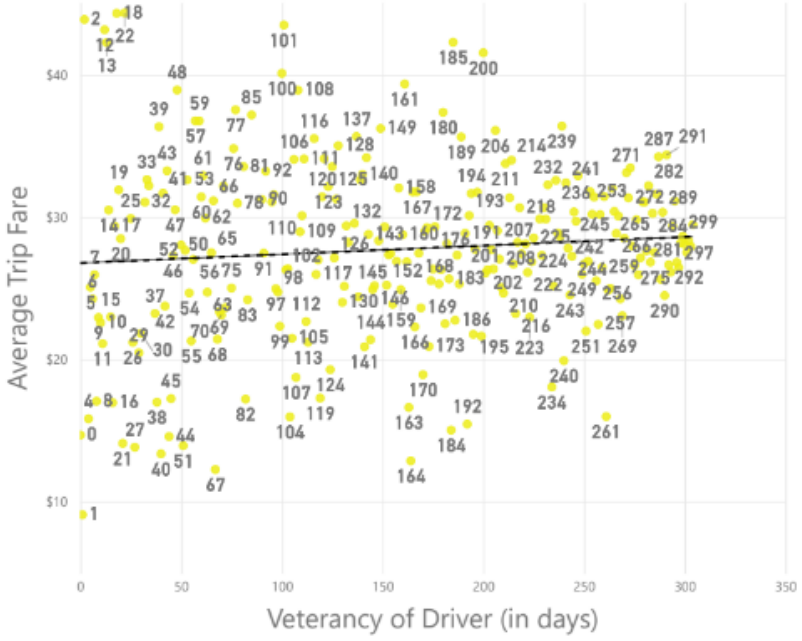
$45.742K

Average Driver Income Per year

All Drivers



Most Drivers (excluding top and bottom 10% of earnings)



- It has been suggested that veteran drivers — drivers who have much experience driving — will produce higher value, here expressed by their average earnings within their recorded activity during 2023.
- However, it seems that while this is true overall, the picture isn't as clear cut as it seems. There is only a modest positive correlation between the veterancy of the driver and his total earnings.
- This correlation becomes even less significant when the top and bottom 10% of earners are excluded.

Company	Amount of Taxis	Avg. Veterancy In Days	Avg Yearly Driver Income	Average Driver income per trip
Flash Cab	546	262.08	$57,753.01	$27.4
Taxi Affiliation Services	533	265.34	$55,116.17	$29.6
Taxicab Insurance Agency, LLC	509	220.89	$38,964.32	$27.4
Sun Taxi	367	242.95	$48,818.67	$29.2
City Service	326	247.18	$49,248.04	$28.6
Chicago Independents	269	224.87	$38,094.67	$28.3
5 Star Taxi	225	277.76	$56,306.79	$31.9
Blue Ribbon Taxi Association	193	185.92	$21,864.38	$19.8
Globe Taxi	149	218.54	$37,839.84	$24.7
Medallion Leasin	122	225.45	$44,248.08	$28.5
Choice Taxi Association	106	196.03	$34,322.75	$24.5

- However, companies with veteran drivers on average tend to have high income per driver on average. The correlation is high (0.71) and jumps to 0.96 when accounting for the largest companies (those who operate over 100 taxis). This can be easily explained — drivers who have higher veterancy have each also taken more passengers overall, so their total earnings will of course be higher.
- However, when the veterancy is correlated to how much each driver in each company made on average, we found a low correlation (0.27). When accounting for the largest companies, there is a much higher correlation (0.86).

Company	Total Taxis	Average Driver Retention Rate	February	March	April	May	June	July	August	September	October
Globe Taxi	149	96.20%	89.80%	84.40%	97.00%	102.06%	100.99%	100.97%	92.31%	97.25%	97.22%
5 Star Taxi	225	90.17%	85.38%	103.28%	104.51%	98.64%	100.00%	96.08%	101.34%	98.04%	103.29%
Sun Taxi	367	90.08%	92.55%	101.57%	99.63%	100.71%	98.63%	95.27%	100.69%	104.35%	96.94%
Taxi Affiliation Services	533	90.01%	94.20%	98.83%	100.45%	99.35%	98.52%	98.52%	99.57%	100.21%	99.79%
Flash Cab	546	89.65%	96.69%	100.81%	97.18%	99.24%	97.12%	96.84%	98.77%	100.00%	99.05%
City Service	326	89.29%	95.52%	96.31%	96.09%	100.78%	97.07%	98.53%	96.44%	99.64%	100.00%
Chicago Independents	269	88.95%	94.01%	100.61%	98.90%	102.16%	91.43%	97.49%	92.31%	102.05%	99.00%
Medallion Leasin	122	88.86%	96.00%	100.00%	98.85%	94.51%	95.88%	97.03%	99.00%	101.00%	94.23%
Taxicab Insurance Agency, LLC	509	88.43%	94.90%	93.81%	94.17%	95.45%	96.37%	95.42%	99.73%	98.68%	103.71%
Choice Taxi Association	106	87.66%	92.31%	83.61%	93.55%	100.00%	100.00%	93.15%	98.67%	93.51%	105.56%
Blue Ribbon Taxi Association	193	85.75%	84.38%	89.91%	97.32%	94.64%	95.80%	91.13%	93.13%	96.90%	99.21%
Average Per Month		**89.55%**	**92.34%**	**95.74%**	**97.97%**	**98.87%**	**97.44%**	**96.40%**	**97.45%**	**99.24%**	**99.82%**

- The following table shows each company's retention rate – how many employees who worked for the company in the previous month continued to work for it in the next month. It should be noted we couldn't give an accurate picture for January 2023, since we don't know the data for December 2022 or before.
- In some months, more new drivers joined the company in the previous month than quit during the current month – hence, in some months, the rate is higher than 100%.
- The leaders in this field, by far, are "Globe Taxi", who don't have a high average driver veterancy, and aren't notable in either yearly or per trip average fares. However, the top company in driver veterancy, yearly pay, and average pay, "5 Star Taxi", is second on the retention rate chart. "5 Star Taxi" also scored highly in the individual months of April and October compared to the average retention rate.

Retention rate code:

```sql
WITH MonthlyTaxiData AS (
    SELECT
        Company,
        DATEPART(MONTH, Trip_Start_Timestamp) AS Month,
        Taxi_ID,
        MIN(Trip_Start_Timestamp) OVER (PARTITION BY Taxi_ID) AS FirstTrip
    FROM
        test2.[dbo].[taxi_trips_2023_total]
    WHERE
        DATEPART(YEAR, Trip_Start_Timestamp) = 2023
),
UniqueTaxisPerMonth AS (
    SELECT
        Company,
        Month,
        COUNT(DISTINCT Taxi_ID) AS A,
        LAG(COUNT(DISTINCT Taxi_ID), 1, 0) OVER (PARTITION BY Company ORDER BY Month) AS B
    FROM
        MonthlyTaxiData
    GROUP BY
        Company, Month
),
NewTaxisPerMonth AS (
    SELECT
        Company,
        DATEPART(MONTH, FirstTrip) AS Month,
        COUNT(DISTINCT Taxi_ID) AS C
    FROM
        MonthlyTaxiData
    WHERE
        DATEPART(YEAR, FirstTrip) = 2023
    GROUP BY
        Company, DATEPART(MONTH, FirstTrip)
)
SELECT
    ut.Company,
    ut.Month,
    ut.A,
    ut.B,
    ISNULL(nt.C, 0) AS C,
    CAST((ut.A - ISNULL(nt.C, 0)) AS FLOAT) / NULLIF(ut.B, 0) AS Result
INTO
    RetentionRate
FROM
    UniqueTaxisPerMonth ut
LEFT JOIN
    NewTaxisPerMonth nt ON ut.Company = nt.Company AND ut.Month = nt.Month;
```

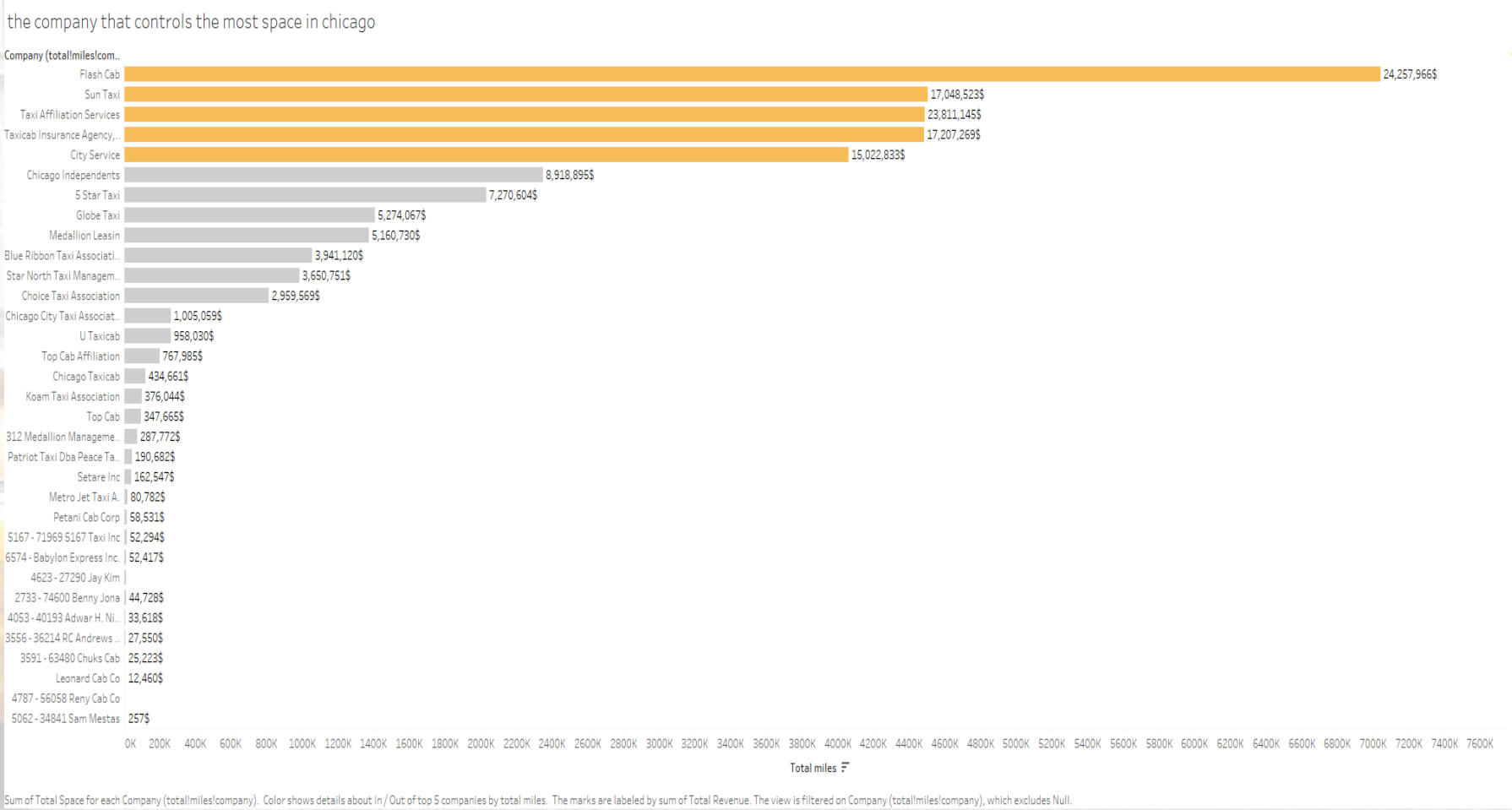
Conclusion

THERE APPEARS TO BE A POSITIVE LINK BETWEEN LONGER SERVING DRIVERS AND HIGHER EARNINGS, HOWEVER, THE CORRELATION TENDS TO BE MODEST.

COMPANIES WITH A HIGHER AVERAGE VETERANCY OF DRIVERS TEND TO HAVE DRIVERS WHO MAKE MORE ON AVERAGE — BUT THIS TREND HAS ONLY BEEN SEEN AMONG LARGE COMPANIES.

COMPANIES WITH A HIGH RETENTION RATE OF DRIVERS ALSO TENDED TO SCORE HIGHLY IN THE AVERAGE PAYMENT RATINGS. HENCE, COMPANIES THAT CAN RETAIN THEIR DRIVERS CAN EXPECT HIGHER OVERALL INCOME FROM THOSE DRIVERS.

Geography Analysis Per Company



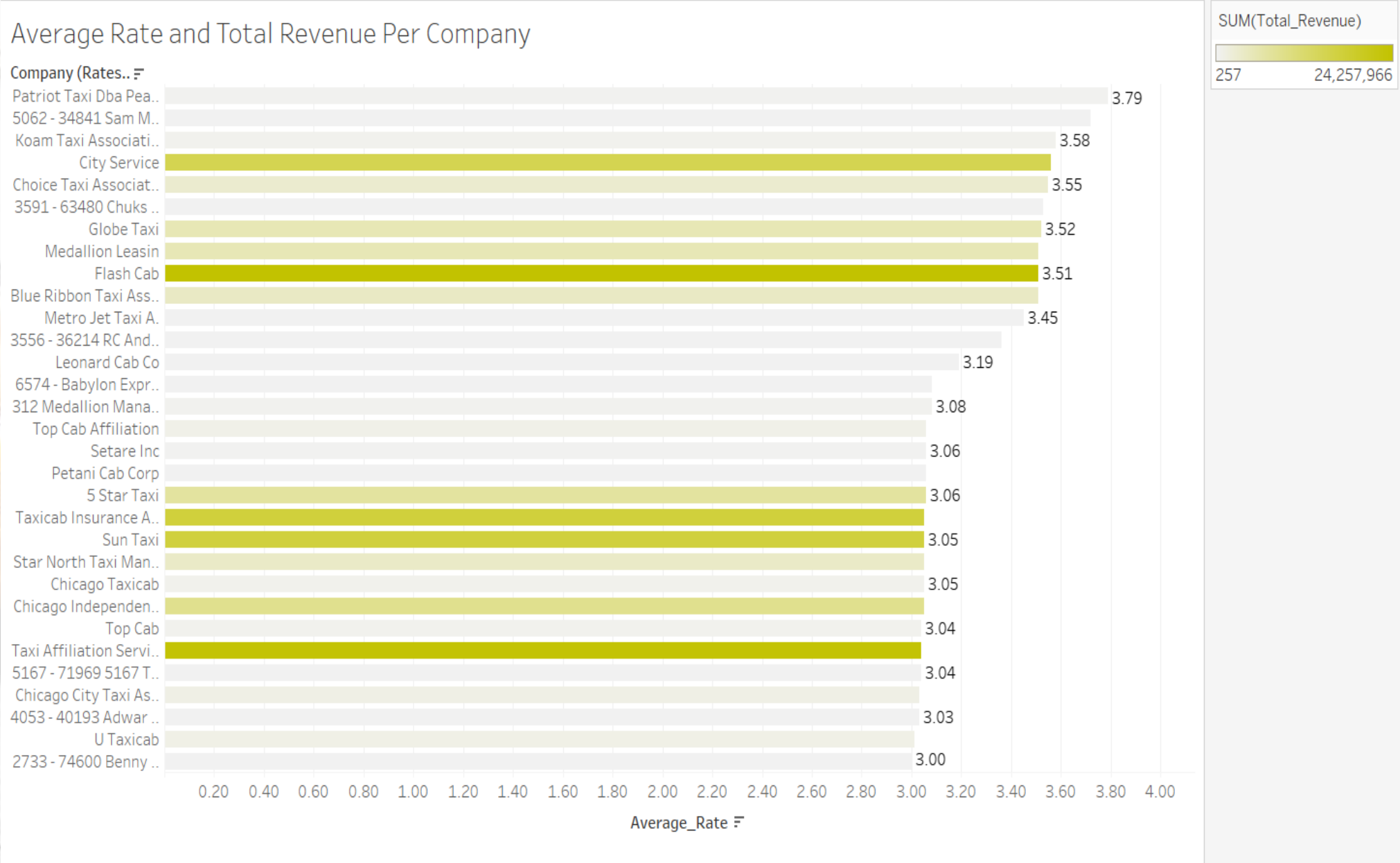
their revenue

top 5 companies by total ..	
Flash Cab	24,257,966$
Taxi Affiliation Services	23,811,145$
Taxicab Insurance Agency..	17,207,269$
Sun Taxi	17,048,523$
City Service	15,022,833$

Sum of Total Revenue broken down by top 5 companies by total miles. The view is filtered on top 5 companies by total miles, which keeps 5 members.

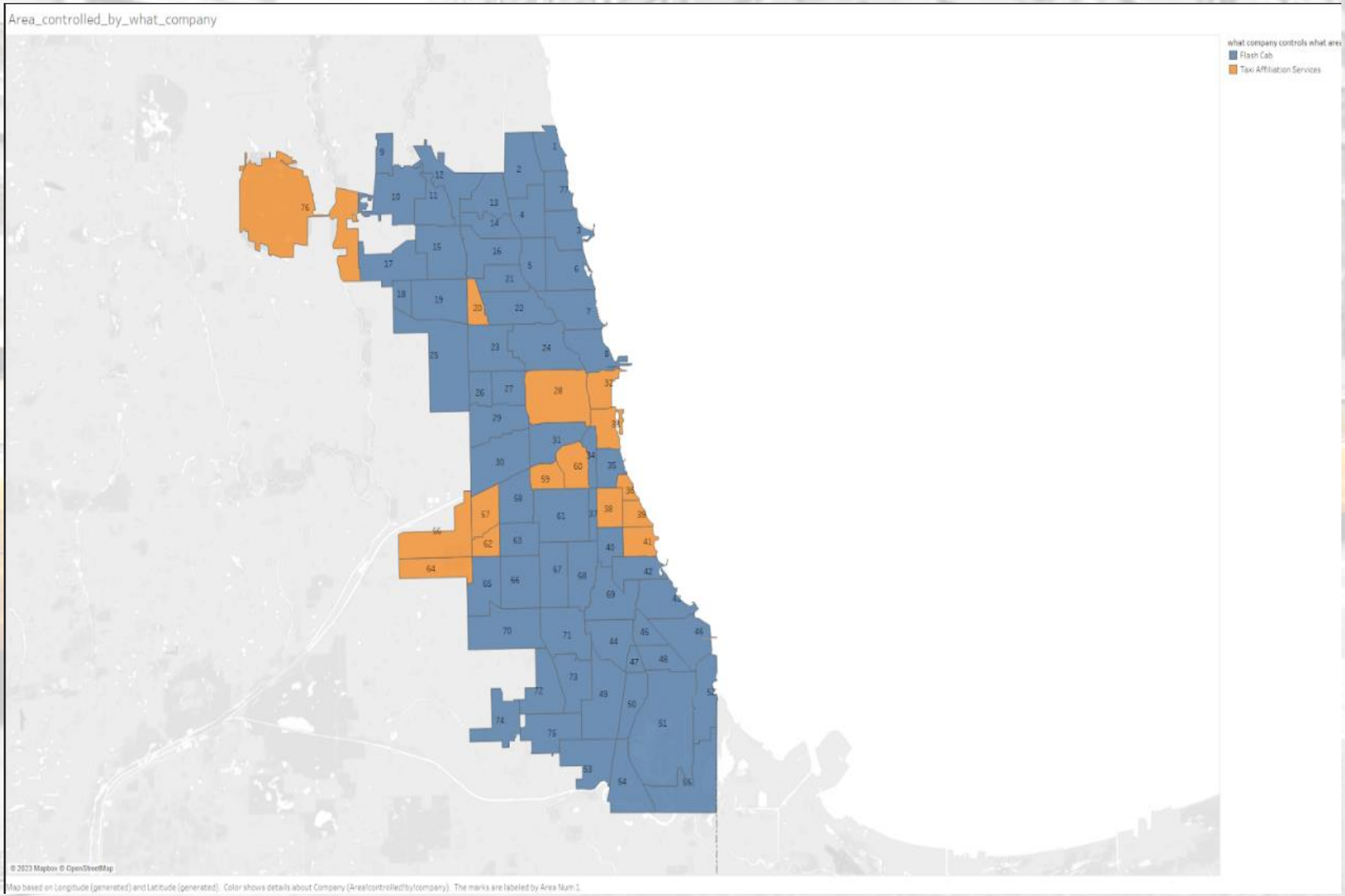
Revenue - This simple table shows the companies and their total revenue in a simpler way.

- **Top 5 miles graph -** This graph allows us to see which companies have the most control over the Chicago taxi industry, by showing the total miles that resulted from each company's rides. Yellow marks the top 5 companies, white the most miles. This graph also shows the total revenue for each company.

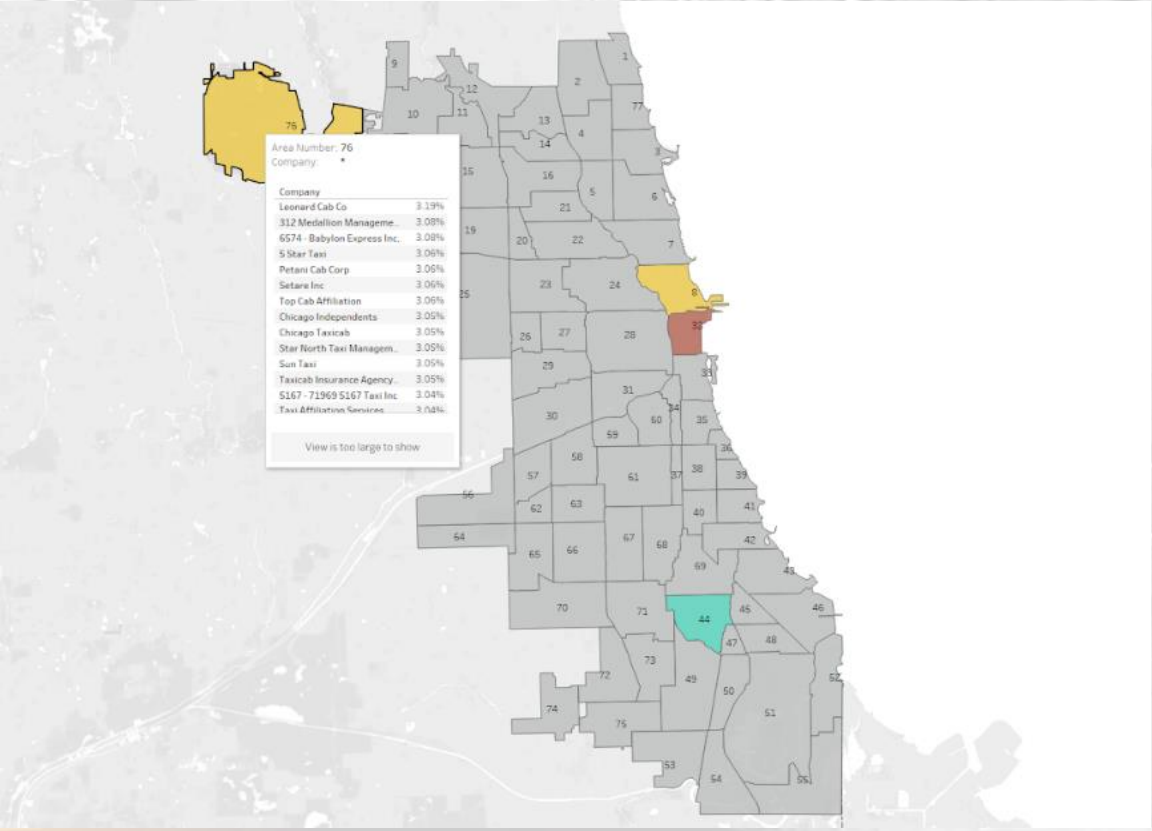


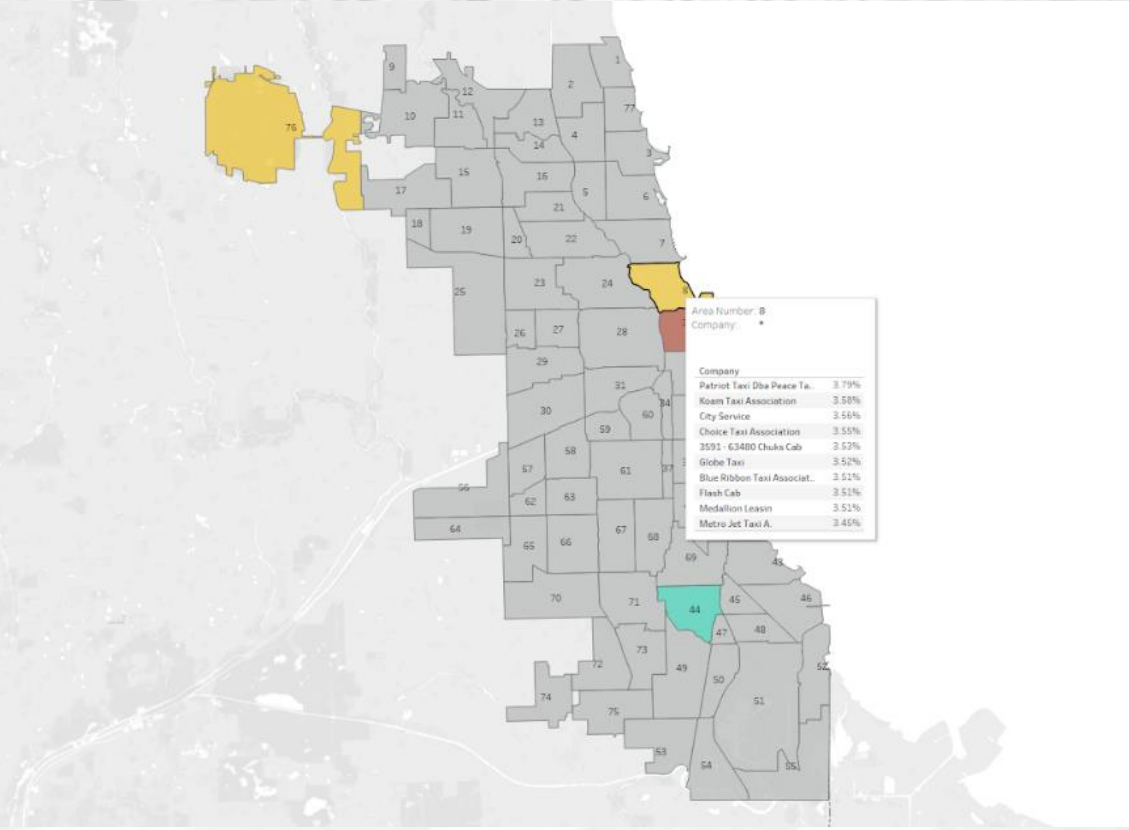
Average rate and total revenue per company

- This graph describes the relationship between the total revenue of the company and the average rate of customer satisfaction.
- It seems that the higher ratings do not affect the total revenue of the company.
- This can be explained due to the distribution of the company by area. Companies with the highest revenue in each area are those that also “control” it – they have the most pick-ups in each area.



Blue and Orange Map — shows which companies control each area. We checked this against the total pickups each company did in each area. There are only 2 companies that control all of Chicago: “Taxi Affiliation Services” and “Flash Cab” — the latter have more areas under their control.





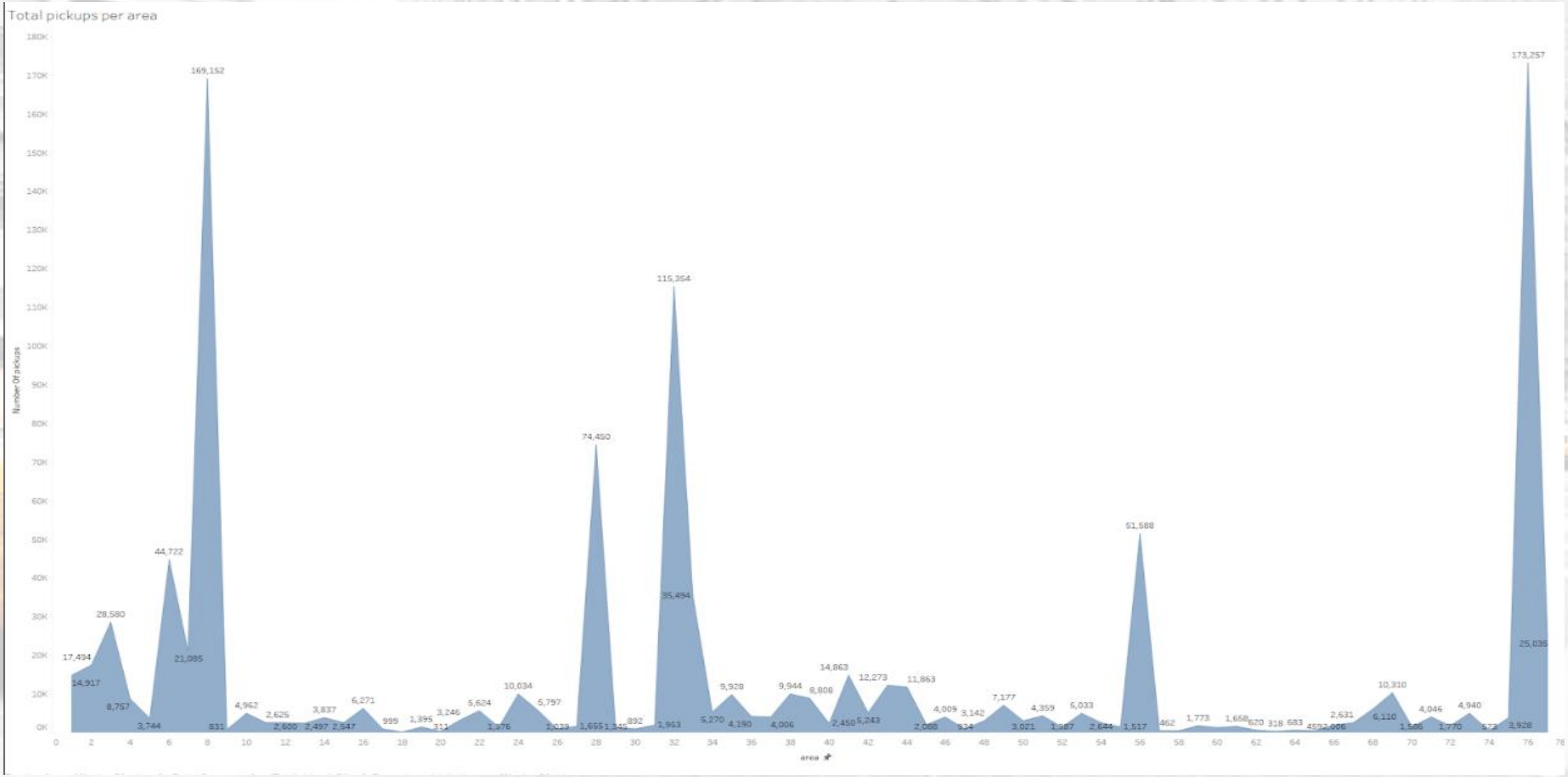
Colorful Map - Shows for each company where they did most of their pickups and therefore which areas they control. We can see that areas 76 and 8 are the most popular areas.

Area 76 - O'Hare Airport, Area 8 - City Center (North).

There are only 2 other companies that have more total pickups in different areas: 32 and 44.

Area 32 is another part of the city center.

Area 44 is a normal neighborhood. However, we can see that "3556 - 36214 RC Andrews Cab", the leading company in this area, is not a popular company. All the other areas in that city where that company has over 100 pickups border that area.



<u>Area Graph</u>- shows the total pickups for each area. As observed, area 8 and 76 have the most pickups.

Query

```sql
-- TotalSpace- the total meils each company did
SELECT Company, SUM(trip_Miles) AS Total_Space
FROM [dbo].[taxi_trips_2023_total]
GROUP BY Company
ORDER BY Total_Space desc;

-- TotalRevenue
SELECT Company, SUM(Fare + Tips + Tolls + Extras) AS Total_Revenue
FROM [dbo].[taxi_trips_2023_total]
GROUP BY Company
ORDER BY Total_Revenue DESC;

---TotalRevenue with area
SELECT Company, Pickup_Community_Area, SUM(Fare + Tips + Tolls + Extras) AS Total_Revenue
FROM [dbo].[taxi_trips_2023_total]
WHERE Pickup_Community_Area IS NOT NULL
GROUP BY Company, Pickup_Community_Area
ORDER BY Total_Revenue DESC;
```

```sql
-- Taxi Company and the controlled area
SELECT Company, Pickup_Community_Area
FROM (
SELECT Company, Pickup_Community_Area,
    ROW_NUMBER() OVER (PARTITION BY Company ORDER BY COUNT(*) DESC) AS rnk
    FROM [dbo].[taxi_trips_2023_total]
WHERE Pickup_Community_Area IS NOT NULL
GROUP BY Company, Pickup_Community_Area) AS ranked
WHERE rnk = 1;

-- Pickup community area and the company controller
SELECT Pickup_Community_Area, Company
FROM (
SELECT Pickup_Community_Area, Company,
    ROW_NUMBER() OVER (PARTITION BY Pickup_Community_Area ORDER BY COUNT(*) DESC) AS rnk
    FROM [dbo].[taxi_trips_2023_total]
WHERE Company IS NOT NULL AND Pickup_Community_Area IS NOT NULL
GROUP BY Pickup_Community_Area, Company) AS ranked
WHERE rnk = 1
ORDER BY Pickup_Community_Area;
```

Query

```sql
--with total Pickup_Community_Area per area and controler company
SELECT Pickup_Community_Area, Company, COUNT(*) as Number_Of_pickups
FROM taxi_trips_2023_total
GROUP BY Pickup_Community_Area, Company
HAVING COUNT(*) = (
    SELECT MAX(TripCount)
    FROM (
        SELECT Pickup_Community_Area, COUNT(*) as TripCount
        FROM taxi_trips_2023_total
        GROUP BY Pickup_Community_Area, Company
    ) as MaxTrips
    WHERE MaxTrips.Pickup_Community_Area = taxi_trips_2023_total.Pickup_Community_Area
) order by Pickup_Community_Area;
```

```sql
--with rates to Community Area
SELECT Pickup_Community_Area, Company,
ROUND(AVG(customer_Rate),2) AS Average_Rate
FROM (
    SELECT Pickup_Community_Area, Company, AVG(customer_Rate) AS customer_Rate,
    ROW_NUMBER() OVER (PARTITION BY Pickup_Community_Area ORDER BY COUNT(*) DESC) AS rnk
    FROM [dbo].[taxi_trips_2023_total]
    WHERE Company IS NOT NULL AND Pickup_Community_Area IS NOT NULL AND customer_Rate IS NOT NULL
    GROUP BY Pickup_Community_Area, Company) AS ranked
WHERE rnk = 1
GROUP BY Pickup_Community_Area, Company
ORDER BY Average_Rate DESC;

--with rates to company
SELECT Company, Pickup_Community_Area,
ROUND(AVG(customer_Rate),2) AS Average_Rate
FROM (
    SELECT Company, Pickup_Community_Area, AVG(customer_Rate) AS customer_Rate,
    ROW_NUMBER() OVER (PARTITION BY Company ORDER BY COUNT(*) DESC) AS rnk
    FROM [dbo].[taxi_trips_2023_total]
    WHERE Pickup_Community_Area IS NOT NULL AND Company IS NOT NULL AND customer_Rate IS NOT NULL
    GROUP BY Company, Pickup_Community_Area) AS ranked
WHERE rnk = 1
GROUP BY Company, Pickup_Community_Area
ORDER BY Average_Rate DESC;
```

Conclusion

IT APPEARS THAT ONLY TWO COMPANIES CONTROL MOST OF THE AREAS OF CHICAGO. THOSE COMPANIES ALSO HAVE THE HIGHEST TOTAL REVENUE.

TWO AREAS IN THE CITY HAVE MORE TRAFFIC THAN ALL THE REST, WHICH MAKES THEM THE MOST POPULAR AREAS FOR TAXI COMPANIES.

THERE IS NO LINK BETWEEN THE AVERAGE CUSTOMER RATING PER COMPANY AND THE COMPANY'S TOTAL REVENUE.