Report: Optimising NYC Taxi Operations

(Sirish Kumar)

Table of Contents

Data Preparation

Loading the dataset

Sample the data and combine the files

The full dataset contained over 30 million records across 12 monthly Parquet files, making full-scale processing computationally expensive. Even a single month exceeded 3 million rows.

To address this, a structured sampling strategy was applied by randomly selecting 5% of trips for every hour of each day across all months in 2023. Sampling was performed at the date and hour level to preserve daily and hourly demand patterns.

Took 5% of trips from each hour of every day throughout 2023 Looped through all 12 monthly files systematically For each month → each date → each hour, I randomly sampled 5% of the trips This preserved the patterns and trends in the data while making it manageable

This approach reduced the dataset to approximately 1896400 records, which were combined into a single DataFrame and saved as a single Parquet file. The resulting dataset is representative of full-year trends while remaining efficient for analysis.

Data Cleaning

Fixing Columns

Fix the index

sampled_data = pd.concat([sampled_data, sampled_hour], **ignore_index=True**)

Combine the two airport_fee columns

Checked unique values in both columns to understand their content. Counted missing values in each column to determine data completeness. Found that Airport_fee had fewer null values than airport_fee

Used the Airport_fee column as the primary column since it had better data quality. Filled any missing values in Airport_fee with corresponding values from airport_fee. Verified by checking remaining null counts

Dropped the redundant airport_fee column

The two airport fee columns were successfully merged into a single Airport_fee column, eliminating duplicate information and improving data consistency. This consolidation is important because having duplicate columns can cause confusion during analysis and waste storage space.

Handling the negative values

Checked each monetary column for negative values and converted negative values to their absolute (positive) equivalents using the .abs() method. Here it is assumed that negative monetary values are inserted by human error so negative is converted into positive.

Follow-up Verification is performed. A second check to confirm all negative values have been addressed. Final conversion of any remaining negative values to absolute values across all numeric columns
Finally, all monetary columns are cleaned to contain only zero or positive values, ensuring the data is logically consistent for financial analysis. The cell notes an observation about the RateCodeID column potentially having different values in records with negative amounts, suggesting a possible pattern worth investigating, probably purging such rows containing the negative values.

Handling Missing Values

Find the proportion of missing values in each column

Column 'passenger_count': Missing Values = 64874, Total cound = 1882715, Percentage = 3.45%
Column 'RatecodeID': Missing Values = 64874, Total cound = 1882715, Percentage = 3.45%
Column 'store_and_fwd_flag': Missing Values = 64874, Total cound = 1882715, Percentage = 3.45%
Column 'congestion_surcharge': Missing Values = 64874, Total cound = 1882715, Percentage = 3.45%
Column 'Airport_fee': Missing Values = 64874, Total cound = 1882715, Percentage = 3.45%

Handling missing values in passenger_count

Calculate the mean passenger count across all trips and round the mean to the nearest whole number (since partial passengers don't make sense) and fill all NaN values with this rounded mean
After imputation, the code checks for any negative passenger counts, which would be logically impossible. This ensures data integrity and identifies any potential data quality issues that might have been introduced or overlooked.

Number of zero passenger count with non-zero trip distance before removal: 28655. All such rows were dropped. Number of zero passenger count with non-zero trip distance after removal is 1854060

Handle missing values in RatecodeID

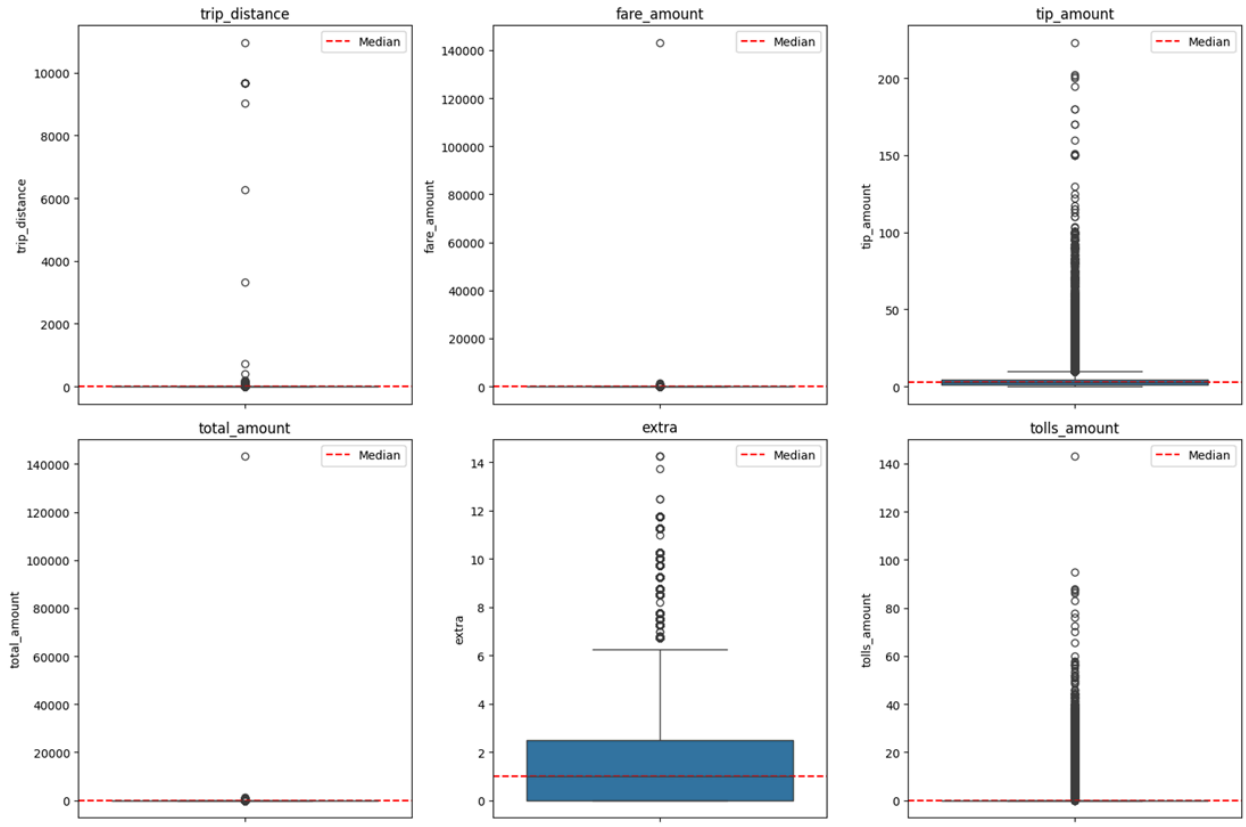
Number of null/empty rows in RatecodeID before imputation: 64874. All these rows were dropped. After dropping, number of null/empty rows in RatecodeID: 2948119.0

Impute NaN in congestion_surcharge

No rows were found with empty or NaN in the congestion_surcharge column.

Handling Outliers and Standardising Values

Check outliers in payment type, trip distance and tip amount columns



There are outliers in trip_distance, fare_amount, total_amount and tolls_amount.

Used the Quartile method to exclude the outliers.

Q1 = 0.05

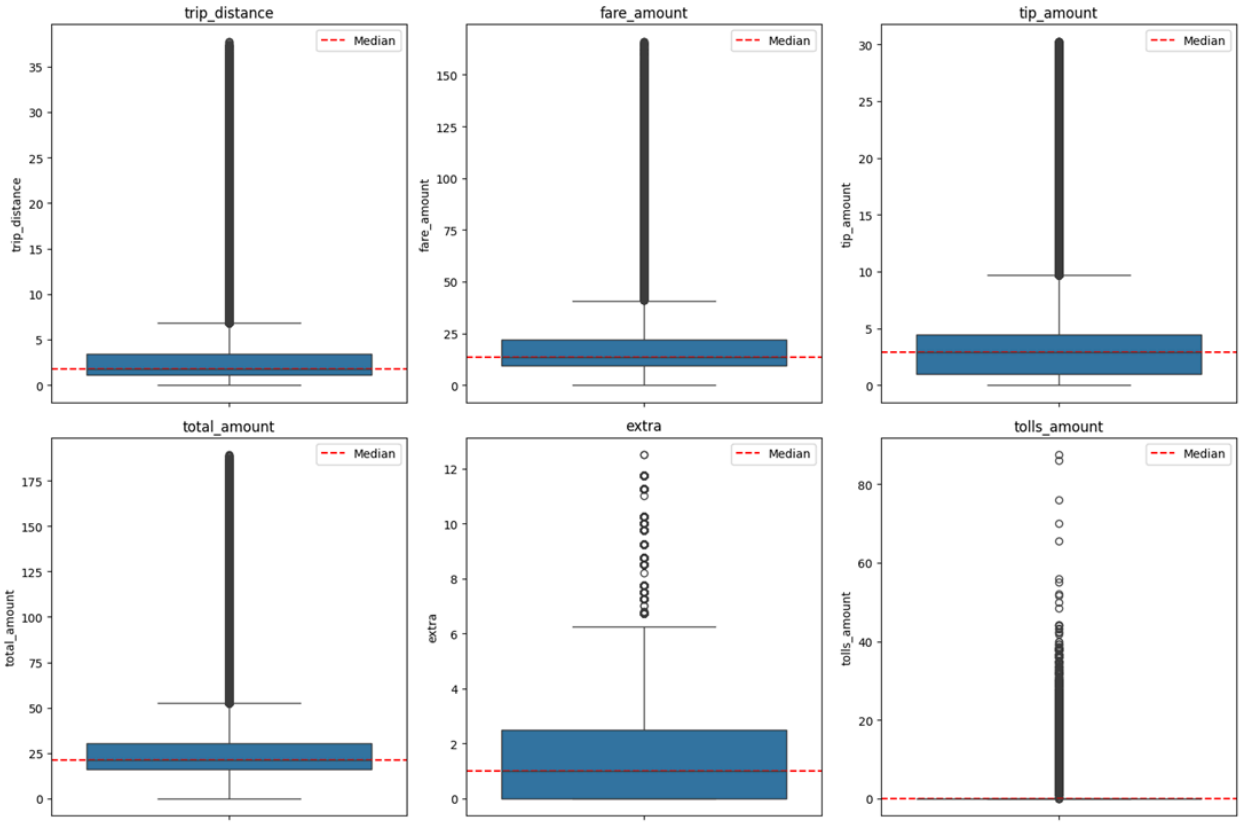
Q2 = 0.95

And removed following

- Entries where `trip_distance` is nearly 0 and `fare_amount` is more than 300
- Entries where `trip_distance` and `fare_amount` are 0 but the pickup and dropoff zones are different (both distance and fare should not be zero for different zones)
- Entries where `trip_distance` is more than 250 miles.

- Entries where `payment_type` is 0 (there is no payment_type 0 defined in the data dictionary)
- Entries with invalid payment_type and invalid RateCodeID

After the outliers removal, following is the visualization for the outlier. This seems fine there are no significant outliers


trip_distance
fare_amount
tip_amount
total_amount
extra
tolls_amount
Median

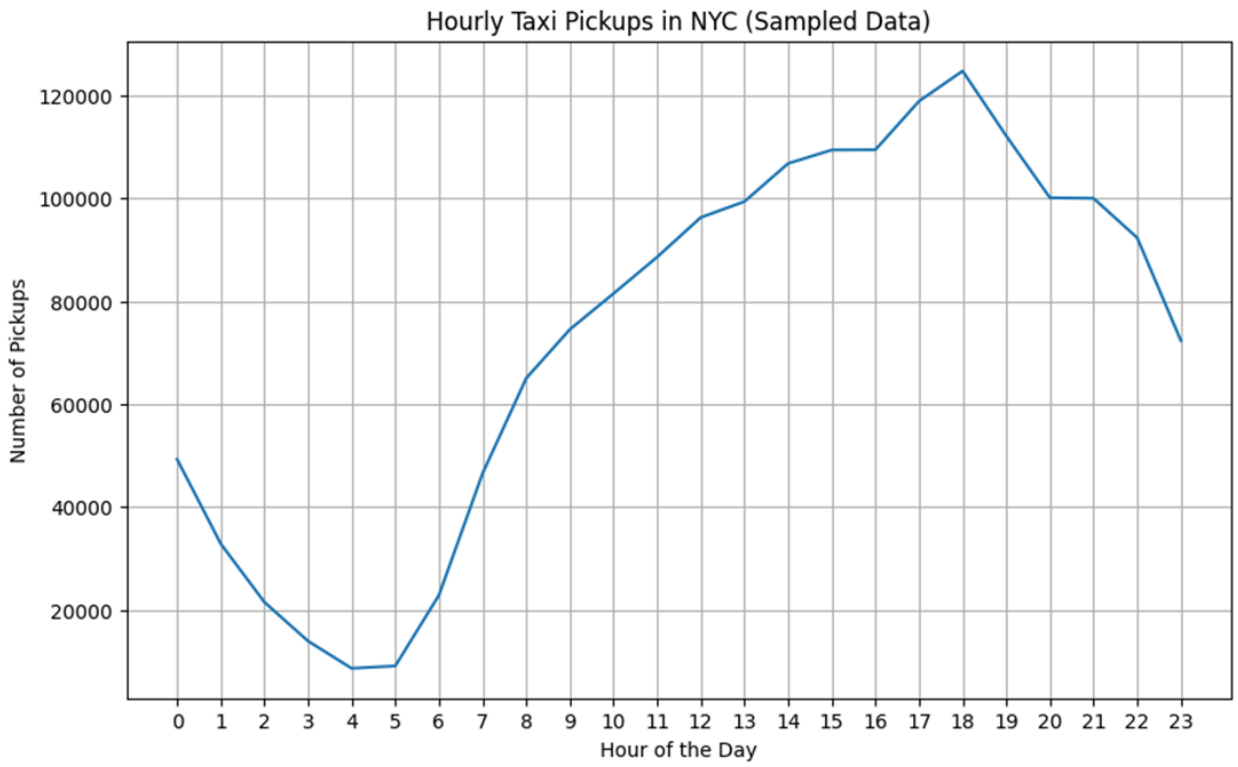
Exploratory Data Analysis

General EDA: Finding Patterns and Trends

Classify variables into categorical and numerical

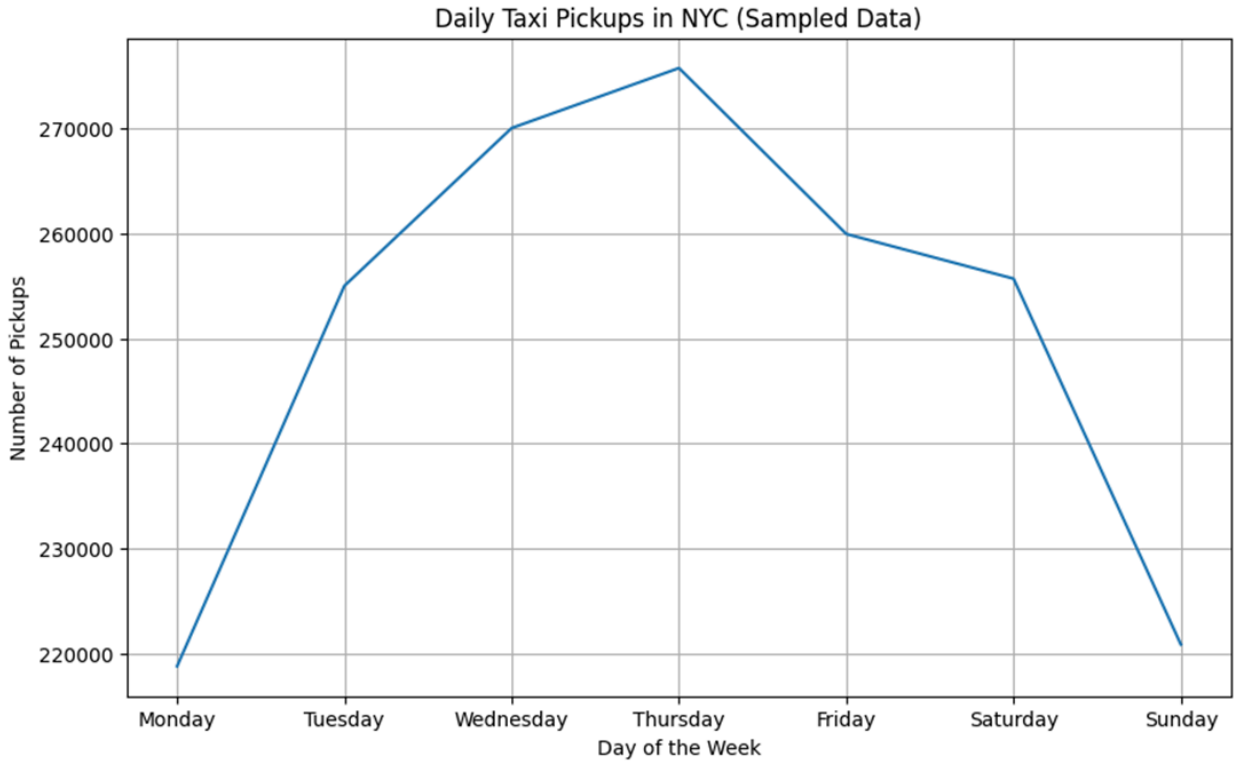
```
Data columns (total 19 columns):
 #   Column                 Dtype
---  ------                 -----
 0   VendorID               int64
 1   tpep_pickup_datetime   datetime64[us]
 2   tpep_dropoff_datetime  datetime64[us]
 3   passenger_count        float64
 4   trip_distance          float64
 5   RatecodeID             float64
 6   store_and_fwd_flag     object
 7   PULocationID           int64
 8   DOLocationID           int64
 9   payment_type           int64
 10  fare_amount            float64
 11  extra                  float64
 12  mta_tax                float64
 13  tip_amount             float64
 14  tolls_amount           float64
 15  improvement_surcharge  float64
 16  total_amount           float64
 17  congestion_surcharge   float64
 18  airport_fee            float64
dtypes: datetime64[us](2), float64(12), int64(4), object(1)
```

Analyse the distribution of taxi pickups by hours, days of the week, and months



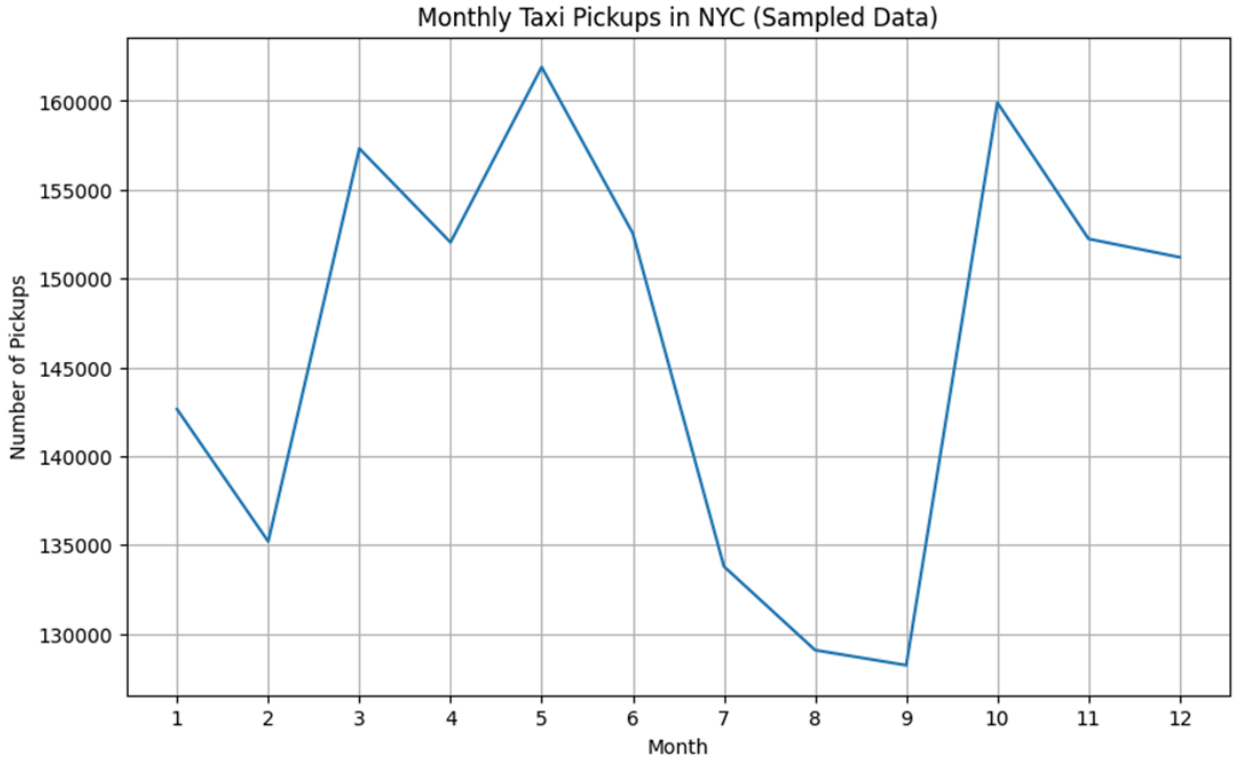
Peak: 6-8 PM (evening rush), 8-9 AM (morning commute)

Low: 3-6 AM (night hours)



Weekdays: Higher demand Tuesday-Friday

Weekends: lower overall demand



High Demand: During months March – June, Oct-December

Low Demand: Jan, Feb, July, August, September (probably due to rainy weathers or summers)

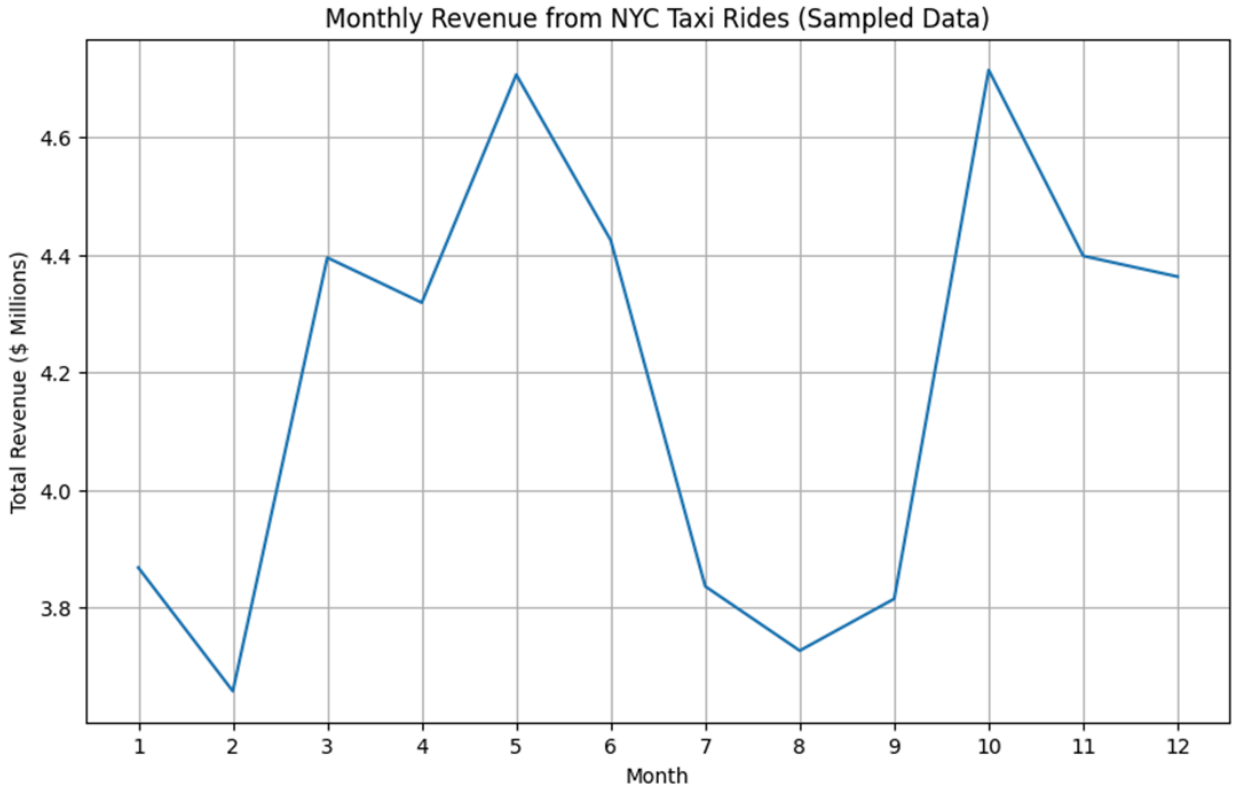
Final Analysis (Distribution of Taxi): Deploy more cabs during weekday rush hours (7-9 AM, 5-8 PM) in business districts.

Filter out the zero/negative values in fares, distance and tips

Removed zero fare and total amount rows.Filtered out trips where fare_amount or total_amount is zero.This is appropriate as legitimate trips should have some fare.Converted negative values to absolute.Changed negative values in financial columns to positive. This assumes data entry errors rather than refunds.

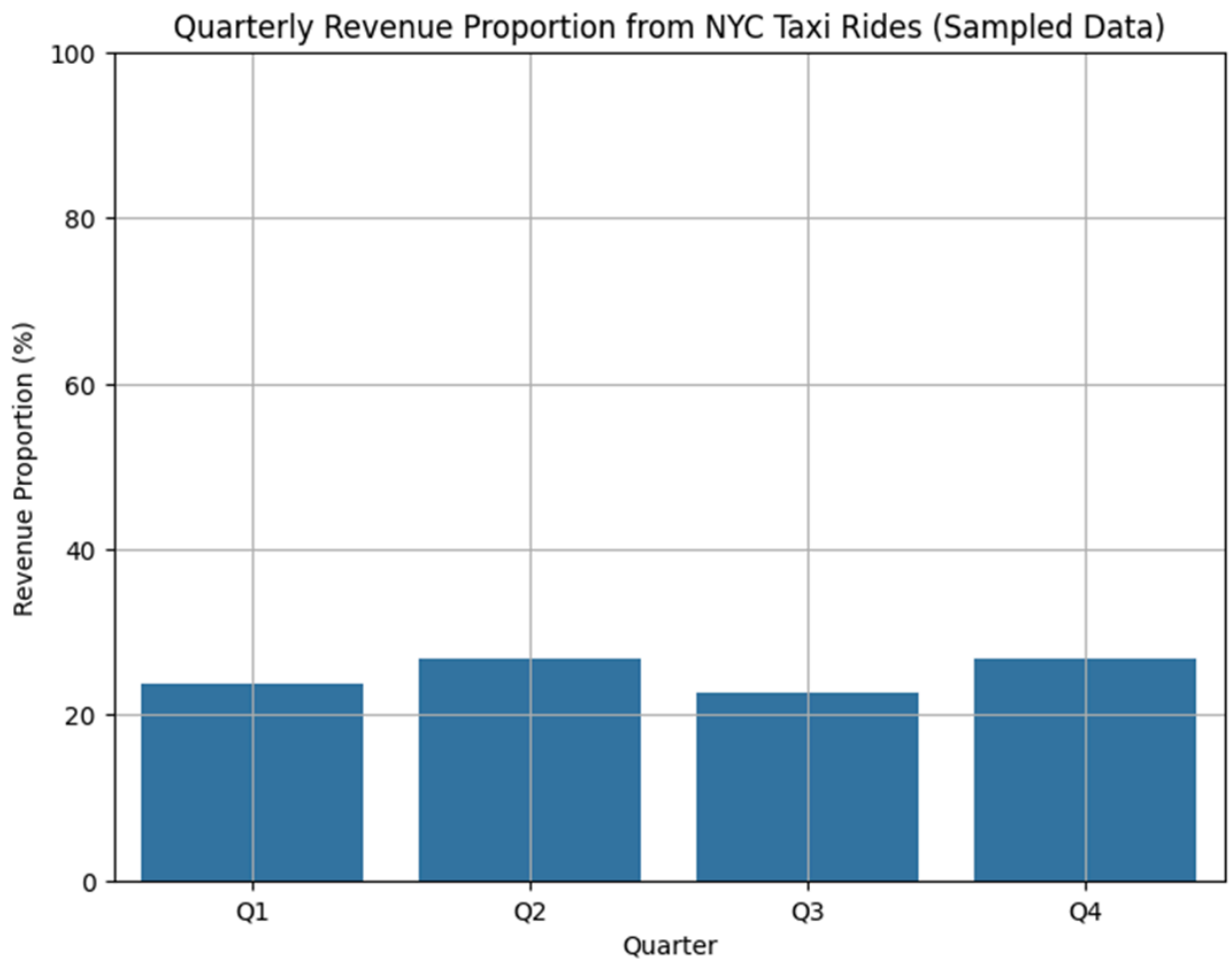
Zero Tips: Mostly from cash payments, which is expected as cash tips aren't recorded
Zero Distance: Can be legitimate when pickup/dropoff are in the same zone, but zero distance with high fare is suspicious
Data Quality: Converting negatives to absolute values is reasonable for small errors, but large negative values might indicate refunds or disputes that should be handled differently

Analyse the monthly revenue trends



Analysis: Relatively flat trend across all 12 months (3.8 to 4.6 million USD) however, March, April, May and Oct, November, December has relatively greater revenue

Find the proportion of each quarter's revenue in the yearly revenue

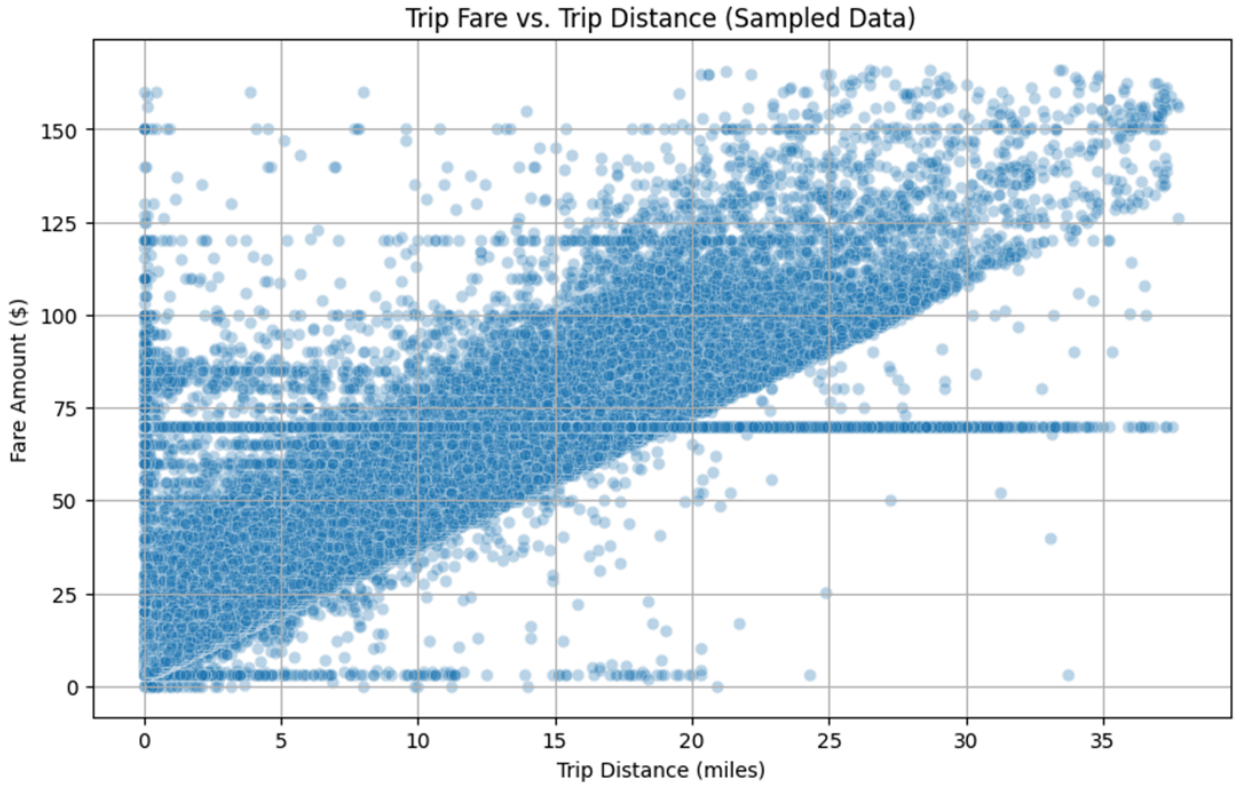


Analysis

- Q2 & Q3 slightly higher
- Q1 & Q4 slightly lower

Revenue can be predicted across the months and quarters and operations could be optimized based on that.

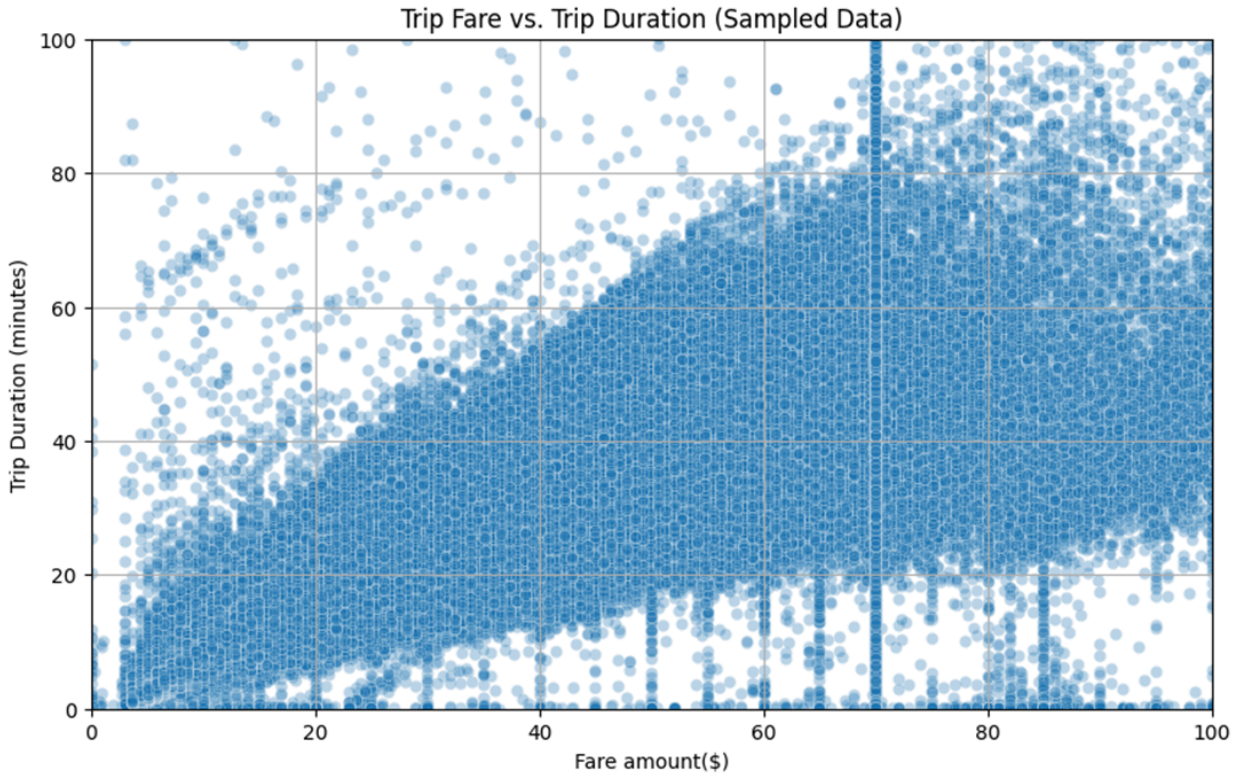
Analyse and visualise the relationship between distance and fare amount



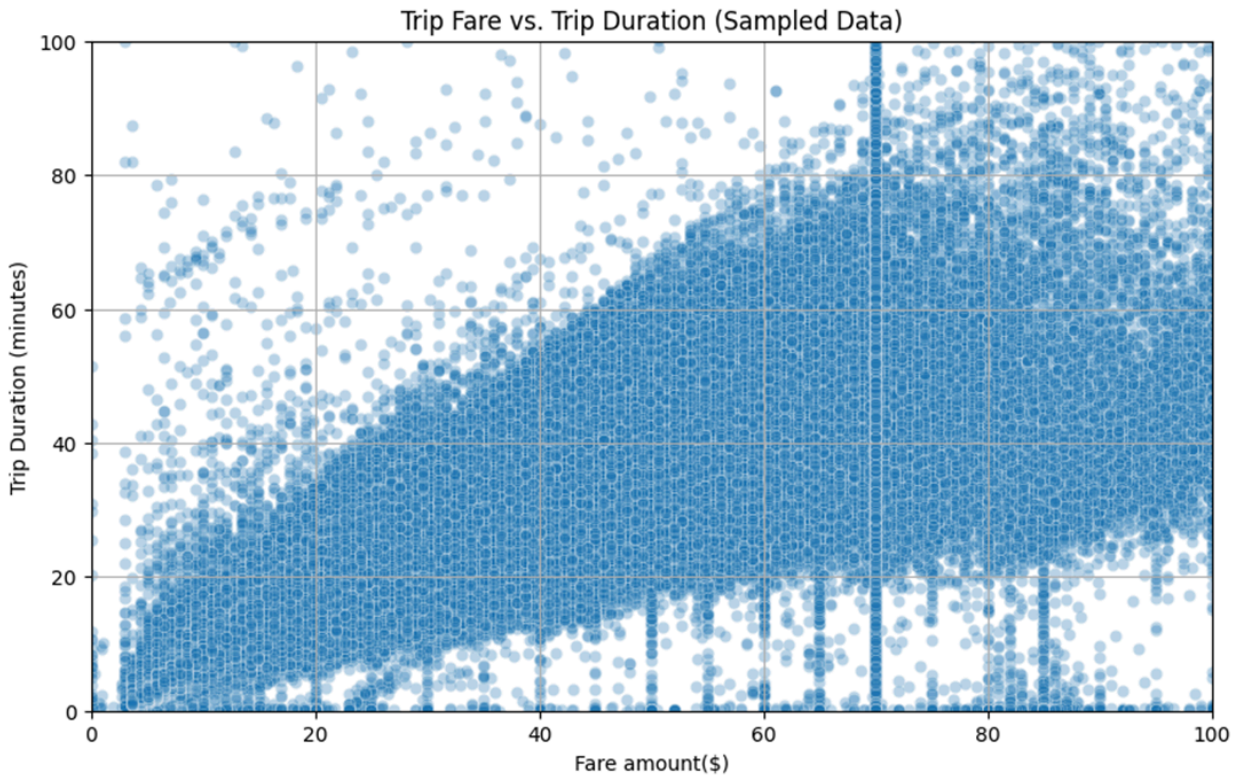
Analysis:

Correlation between trip distance and fare amount: 0.95. **Very strong positive linear relationship**

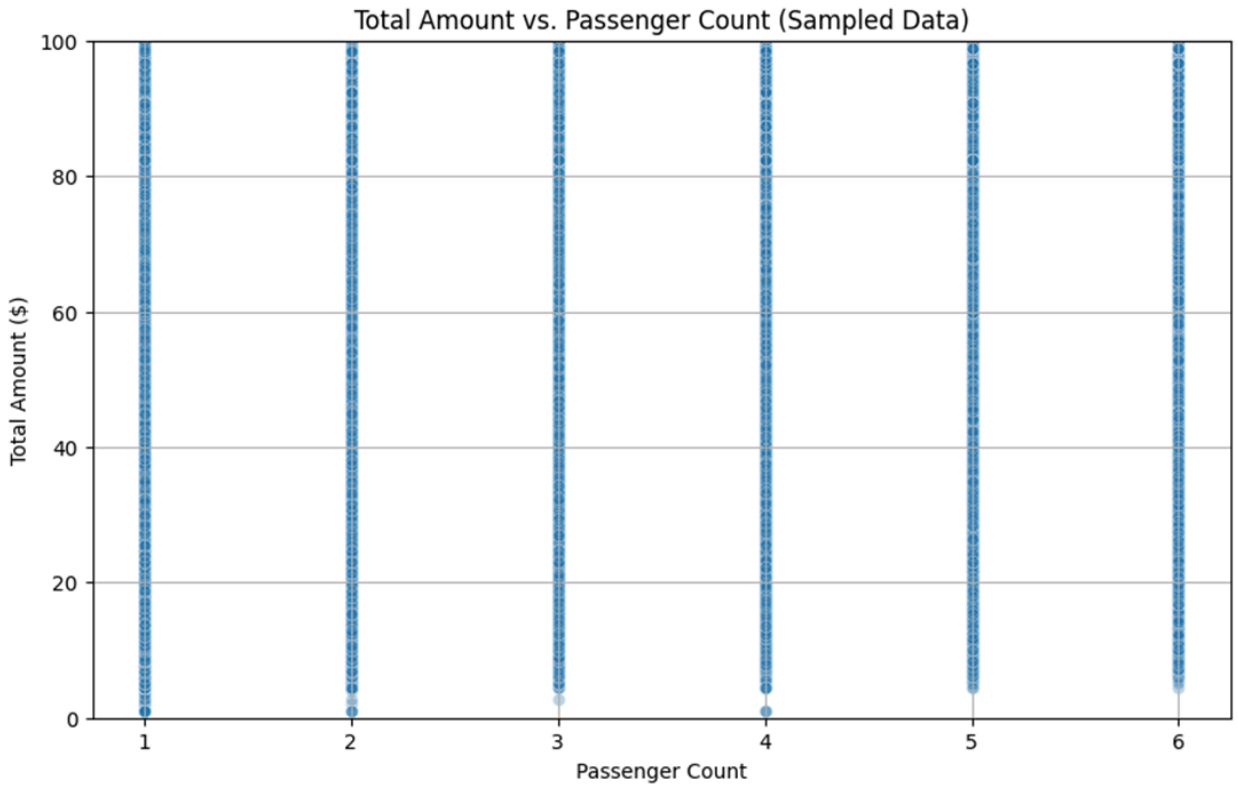
Analyse the relationship between fare/tips and trips/passengers



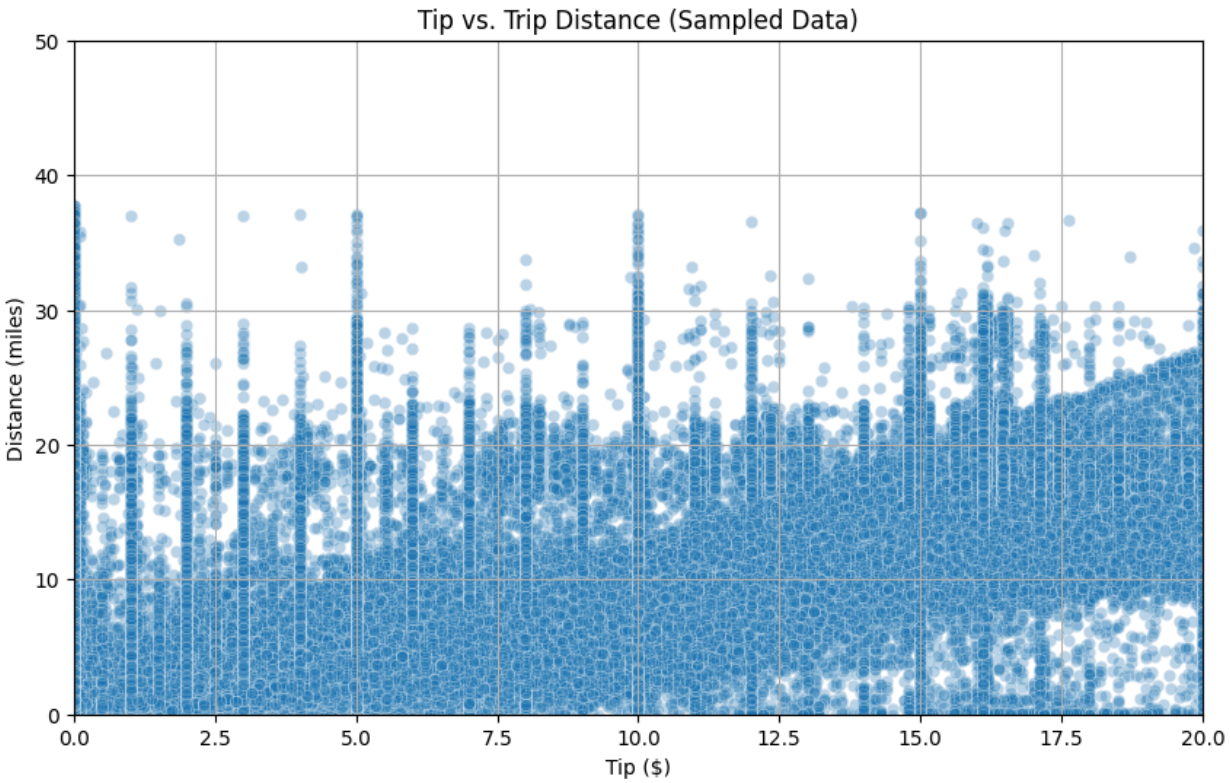
Analysis: Correlation between trip duration and fare amount: 0.27 but it could be concluded that fare amount increases (not strict linear) by trip duration



Analysis: Correlation between trip duration and fare amount: 0.27 but it could be concluded that fare amount increases (not strict linear) by trip duration

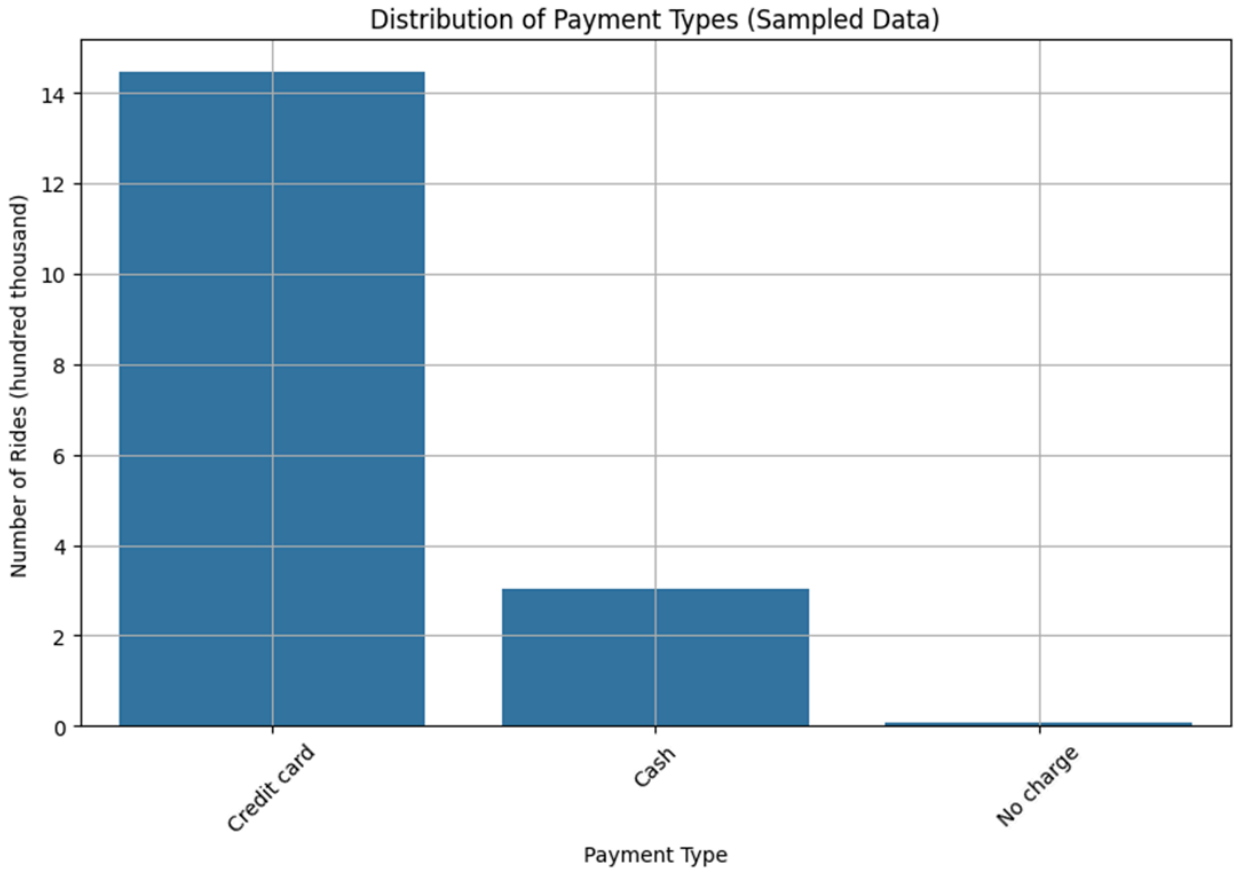


Analysis: Correlation between passenger count and total amount: 0.04 so, passenger count has no impact on the total amount.



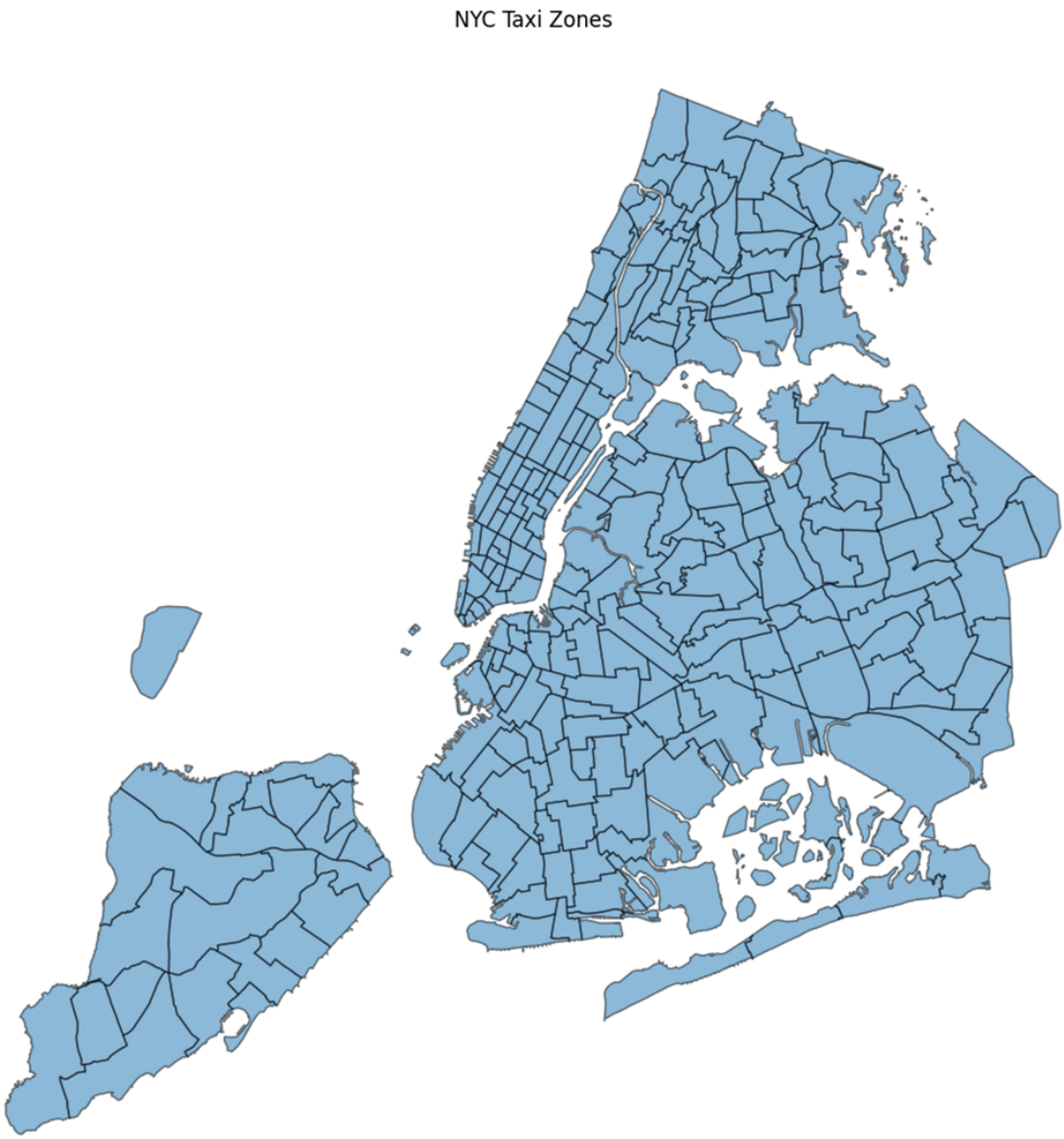
Analysis: Correlation between trip distance and tip amount: 0.62. It does seem that longer the trip distance the better the tip is.

Analyse the distribution of different payment types



Analysis: Passengers prefer using the credit cards.

Load the taxi zones shapefile and display it



Merge the zone data with trips data

```
zones_with_trips = pd.merge(zones, trip_counts, left_on='LocationID',
right_on='PULocationID', how='left')
```

	OBJECTID	Shape_Leng	Shape_Area	zone	LocationID	borough	geometry	PULocationID	number of trips
0	1	0.116357	0.000782	Newark Airport	1	EWR	POLYGON ((933100.918 192536.086, 933091.011 19...	1.0	44.0
1	2	0.433470	0.004866	Jamaica Bay	2	Queens	MULTIPOLYGON (((1033269.244 172126.008, 103343...	2.0	2.0
2	3	0.084341	0.000314	Allerton/Pelham Gardens	3	Bronx	POLYGON ((1026308.77 256767.698, 1026495.593 2...	NaN	NaN
3	4	0.043567	0.000112	Alphabet City	4	Manhattan	POLYGON ((992073.467 203714.076, 992068.667 20...	4.0	1731.0
							POLYGON		

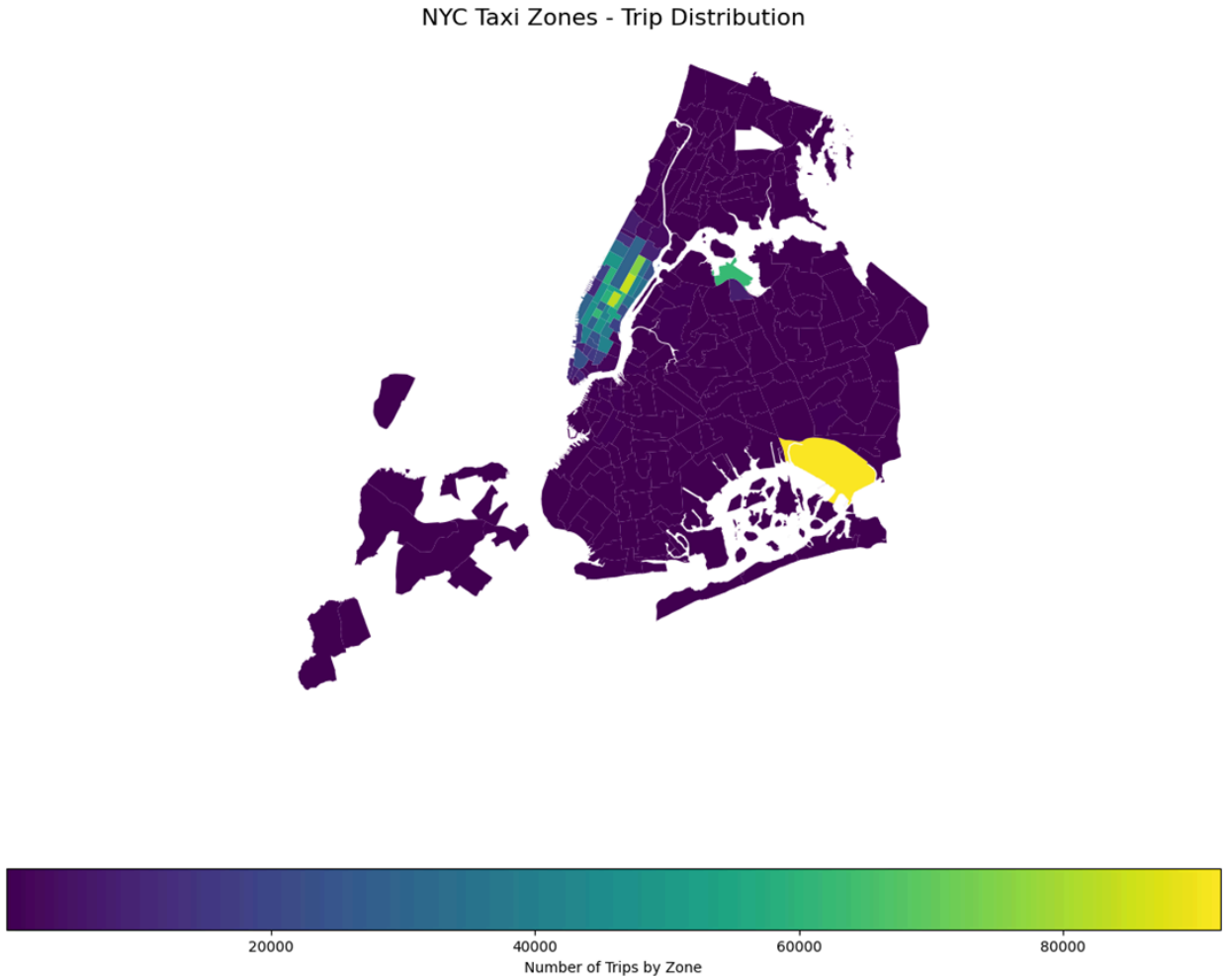
Find the number of trips for each zone/location ID

Zone	Number of Trips
JFK Airport	91,920
Upper East Side South	84,453
Midtown Center	83,350
Upper East Side North	75,327
Midtown East	63,716
LaGuardia Airport	62,331
Penn Station / Madison Sq West	61,406
Times Sq / Theatre District	59,223
Lincoln Square East	59,057
Murray Hill	52,795
Midtown North	52,098
Upper West Side South	48,781
Union Sq	47,989
Clinton East	47,136
East Chelsea	46,287
Lenox Hill West	42,144
East Village	41,628
Midtown South	41,366
West Village	39,335
Gramercy	37,272

Add the number of trips for each zone to the zones dataframe

	OBJECTID	Shape_Leng	Shape_Area	zone	LocationID	borough	geometry	PULocationID	number of trips
0	1	0.116357	0.000782	Newark Airport	1	EWR	POLYGON ((933100.918 192536.086, 933091.011 19...	1.0	44.0
1	2	0.433470	0.004866	Jamaica Bay	2	Queens	MULTIPOLYGON (((1033269.244 172126.008, 103343...	2.0	2.0
2	3	0.084341	0.000314	Allerton/Pelham Gardens	3	Bronx	POLYGON ((1026308.77 256767.698, 1026495.593 2...	NaN	NaN
3	4	0.043567	0.000112	Alphabet City	4	Manhattan	POLYGON ((992073.467 203714.076, 992068.667 20...	4.0	1731.0
							POLYGON		

Plot a map of the zones showing number of trips



Conclude with results

Analysis:

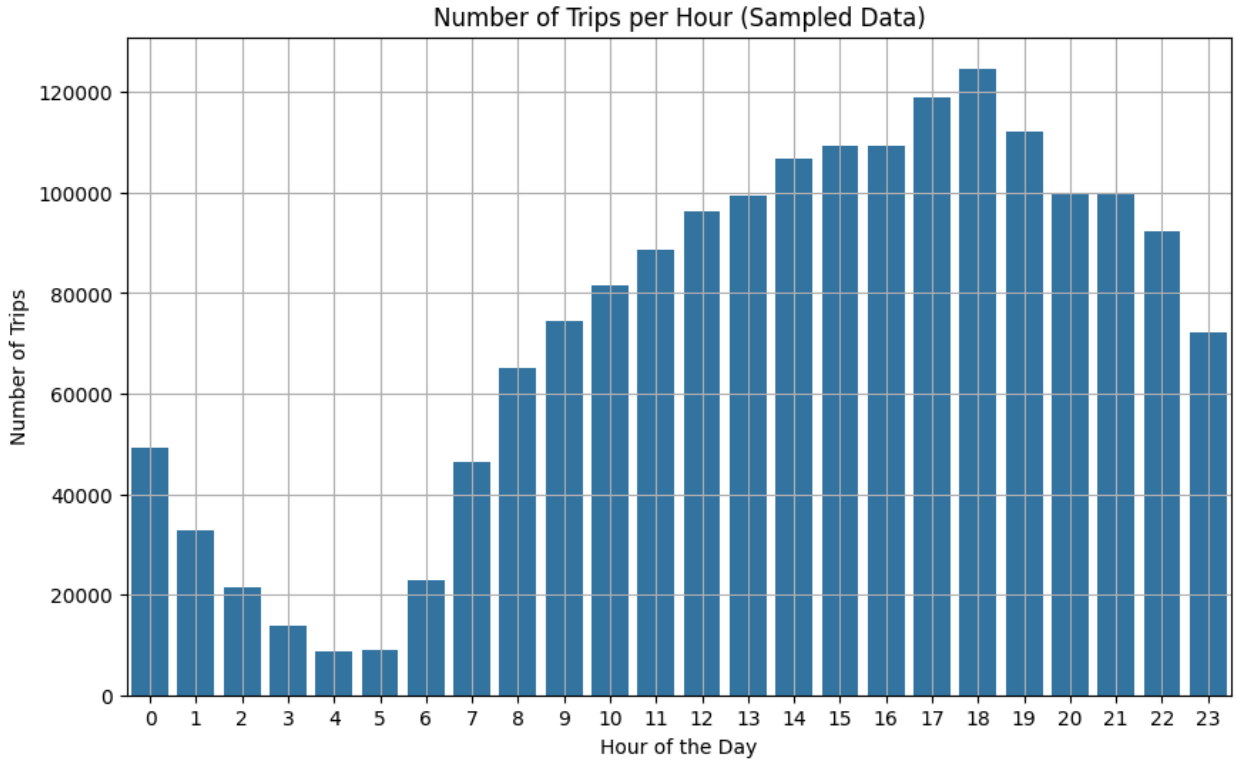
- Manhattan captures 80%+ of all taxi trips
- Airport zones (JFK, LaGuardia) high pickups/drops

Detailed EDA: Insights and Strategies

Identify slow routes by comparing average speeds on different routes

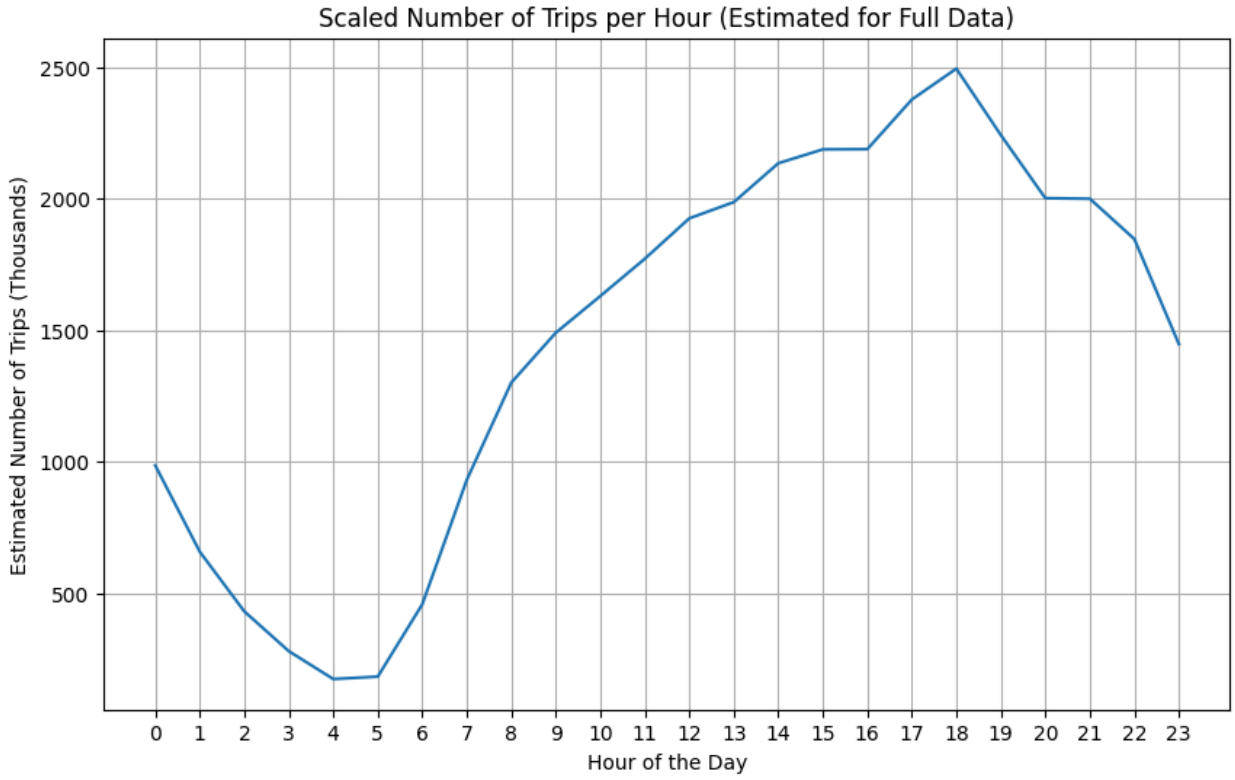
	PULocationID	DOLocationID	hour	speed_mph
94533	232	65	13	0.005324
106750	243	264	17	0.007772
56372	142	142	5	0.023770
29708	100	7	8	0.039470
5311	40	65	21	0.046848
35630	113	235	22	0.048105
82833	194	194	16	0.048913
8353	45	45	10	0.059484
117854	264	168	23	0.071573
35732	113	244	9	0.084739
18090	70	138	6	0.085750
1549	12	45	9	0.090760
86026	215	215	23	0.110837
39853	128	128	7	0.115385
102950	237	238	4	0.120376
59418	144	100	2	0.121962
10805	48	184	1	0.127919
85520	211	230	4	0.132710
84572	211	52	18	0.132920
94220	231	247	15	0.142979

Calculate the hourly number of trips and identify the busy hours



Busy Hours: 9am till Midnight.

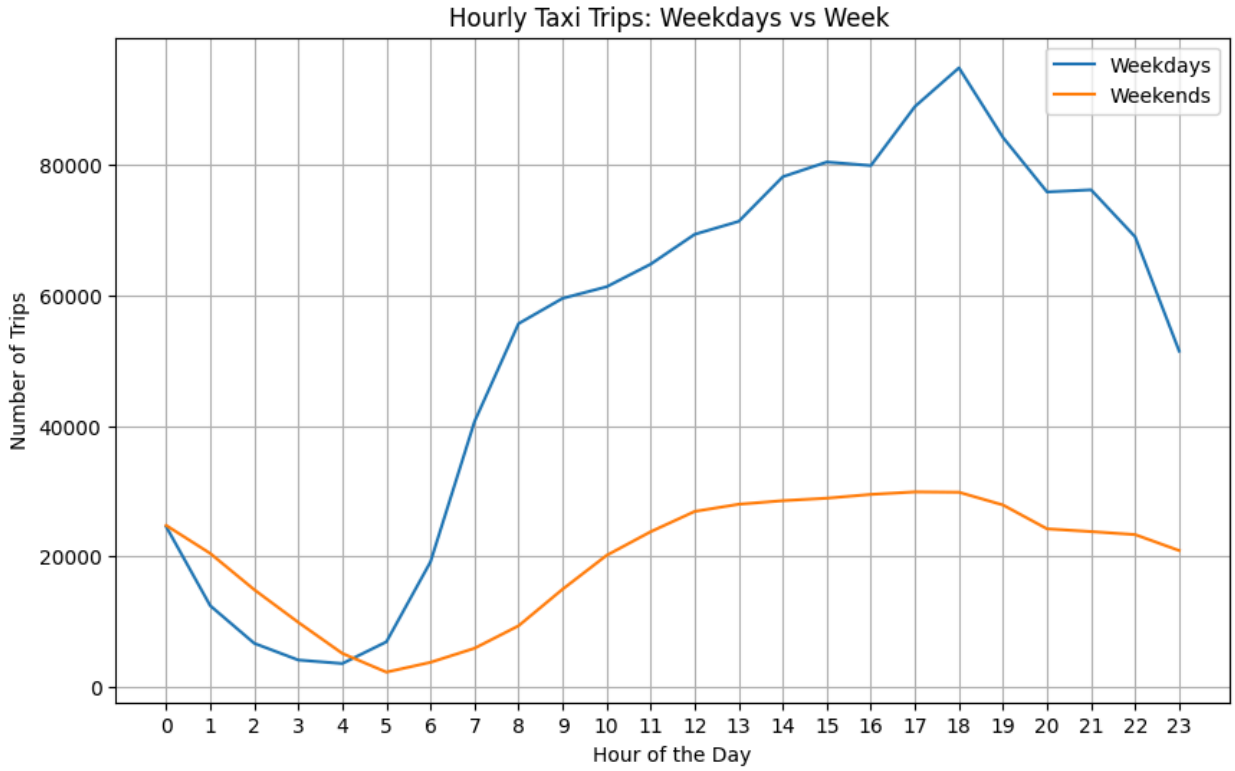
Scale up the number of trips from above to find the actual number of trips



Top 5 Busiest Hours:

Hour 18:00 - 2,494,360 trips (2494.4K)
Hour 17:00 - 2,376,780 trips (2376.8K)
Hour 19:00 - 2,241,660 trips (2241.7K)
Hour 16:00 - 2,187,980 trips (2188.0K)
Hour 15:00 - 2,187,240 trips (2187.2K)

Compare hourly traffic on weekdays and weekends



Analysis:

On weekdays, we see the usual double spike: a rush at 8 AM and an even bigger surge around 6 PM. Weekends are much less busier in the morning, peaking instead late in the evening. Boost the fleet by 50% for the weekday evening rush

Identify the top 10 zones with high hourly pickups and drops

```
Top 10 Pickup Zones:
                           zone  pickup_count
0                   JFK Airport         91920
1         Upper East Side South         84453
2                Midtown Center         83350
3         Upper East Side North         75327
4                  Midtown East         63716
5             LaGuardia Airport         62331
6  Penn Station/Madison Sq West         61406
7     Times Sq/Theatre District         59223
8           Lincoln Square East         59057
9                   Murray Hill         52795

Top 10 Dropoff Zones:
                         zone  dropoff_count
0      Upper East Side North          79047
1      Upper East Side South          75316
2             Midtown Center          69427
3  Times Sq/Theatre District          54413
4                Murray Hill          52663
5               Midtown East          50628
6        Lincoln Square East          49975
7      Upper West Side South          49849
8            Lenox Hill West          47144
9               East Chelsea          44840
```

Find the ratio of pickups and dropoffs in each zone

```
Top 10 Zones by Pickup/Dropoff Ratio:
                            zone  pickup_dropoff_ratio
195                Rikers Island                   inf
70                 East Elmhurst              9.699147
128                  JFK Airport              4.893005
134            LaGuardia Airport              2.928125
182  Penn Station/Madison Sq West              1.585244
110       Greenwich Village South              1.377358
42                  Central Park              1.373667
245                 West Village              1.328661
158                 Midtown East              1.258513
157                Midtown Center              1.200542
```

Identify the top zones with high traffic during night hours

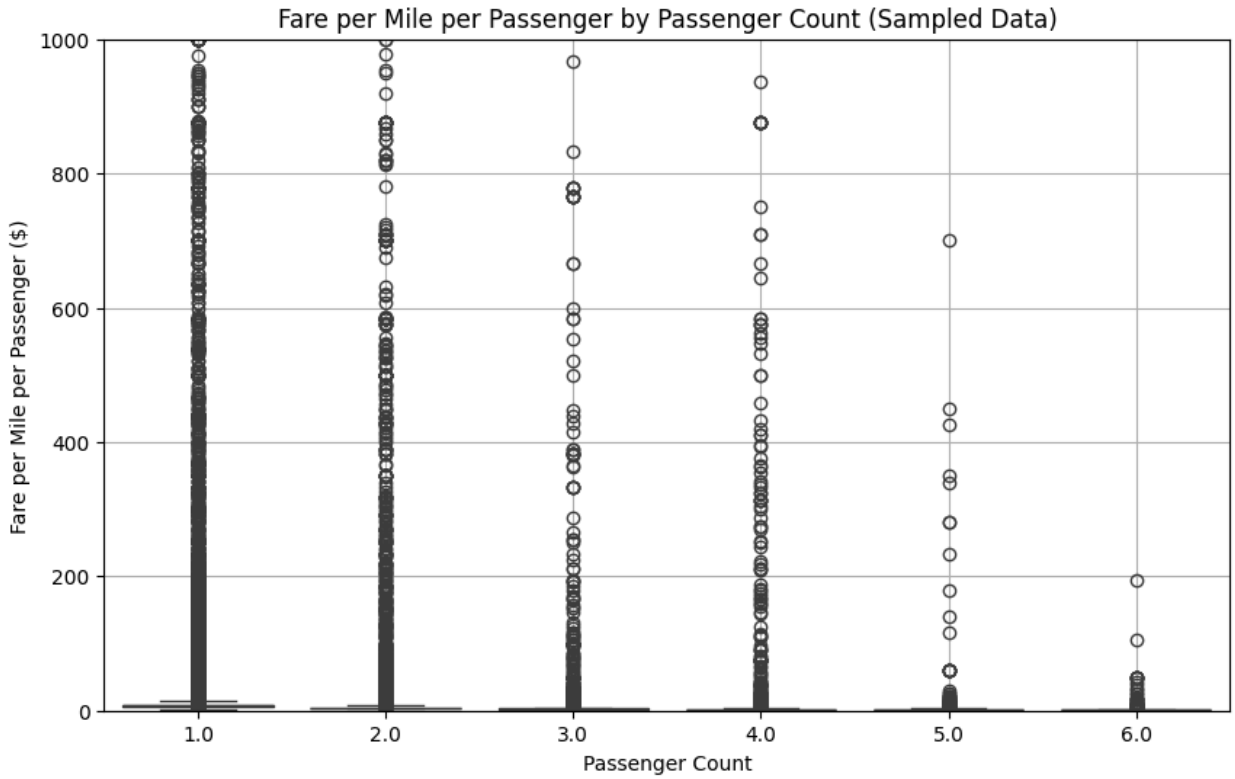
```
Top 10 Pickup Zones during Night Hours:
                            zone  pickup_count
2                   East Village       15060.0
5                    JFK Airport       13867.0
13                  West Village       12128.0
0                   Clinton East       10134.0
7                Lower East Side        9357.0
4         Greenwich Village South        8503.0
11     Times Sq/Theatre District        7917.0
10  Penn Station/Madison Sq West        6733.0
8                   Midtown South        5944.0
1                    East Chelsea        5806.0
3                        Gramercy           NaN
6                Lenox Hill West           NaN
9                    Murray Hill           NaN
12         Upper West Side South           NaN
14                Yorkville West           NaN

Top 10 Dropoff Zones during Night Hours:
                            zone  dropoff_count
2                   East Village         8045.0
0                   Clinton East         6638.0
9                    Murray Hill         6044.0
1                    East Chelsea         5631.0
3                        Gramercy         5575.0
6                Lenox Hill West         5143.0
14                Yorkville West         4851.0
13                  West Village         4770.0
11     Times Sq/Theatre District         4417.0
12         Upper West Side South         4254.0
```

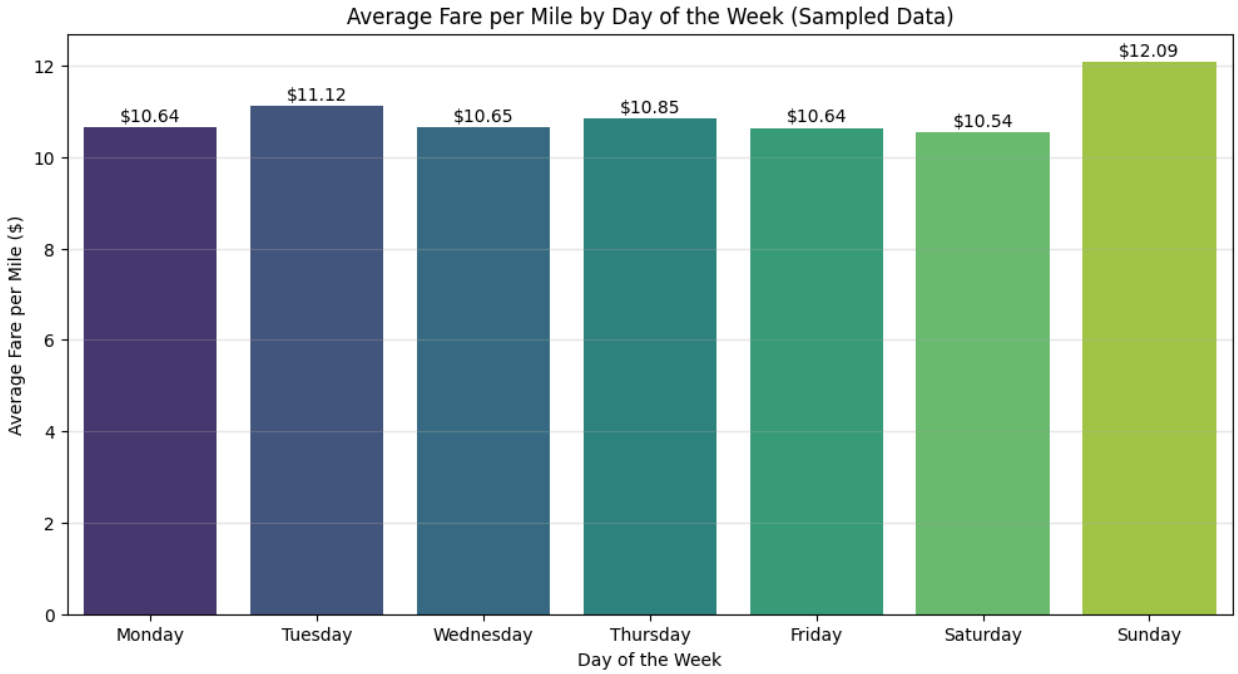
Find the revenue share for nighttime and daytime hours

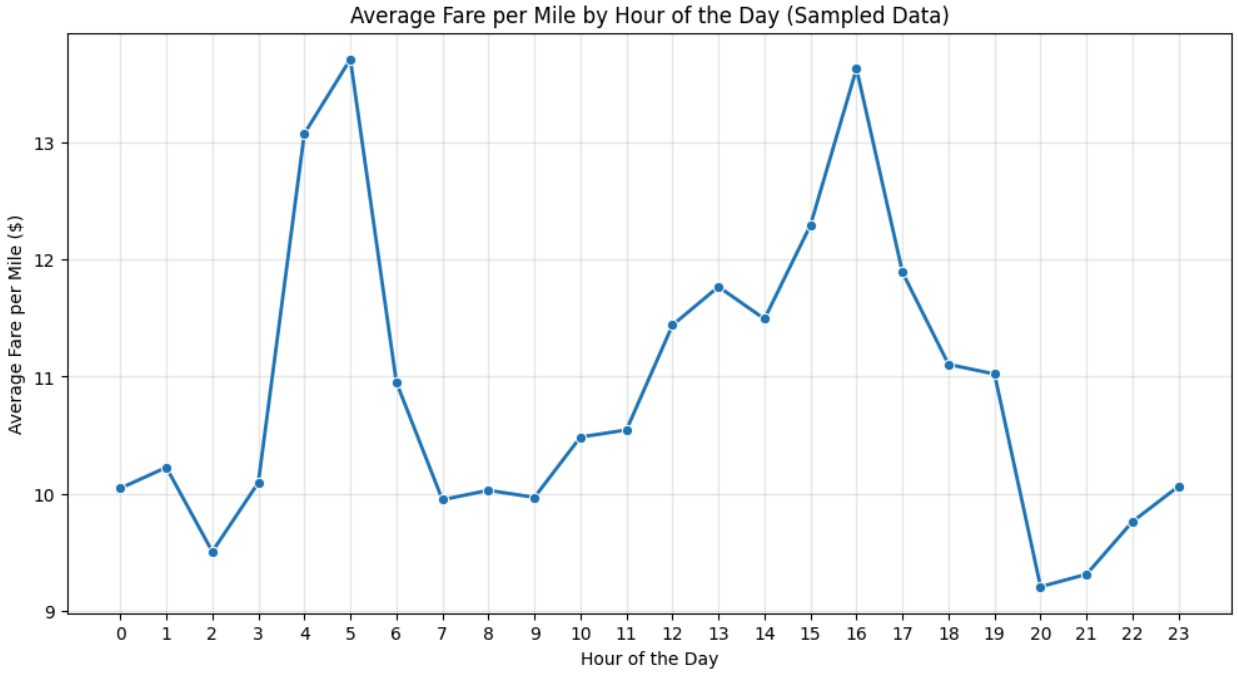
Total Revenue during Night Hours (11 PM to 5 AM): $6,024,441.89
Total Revenue during Day Hours (6 AM to 10 PM): $44,196,326.47

For the different passenger counts, find the average fare per mile per passenger

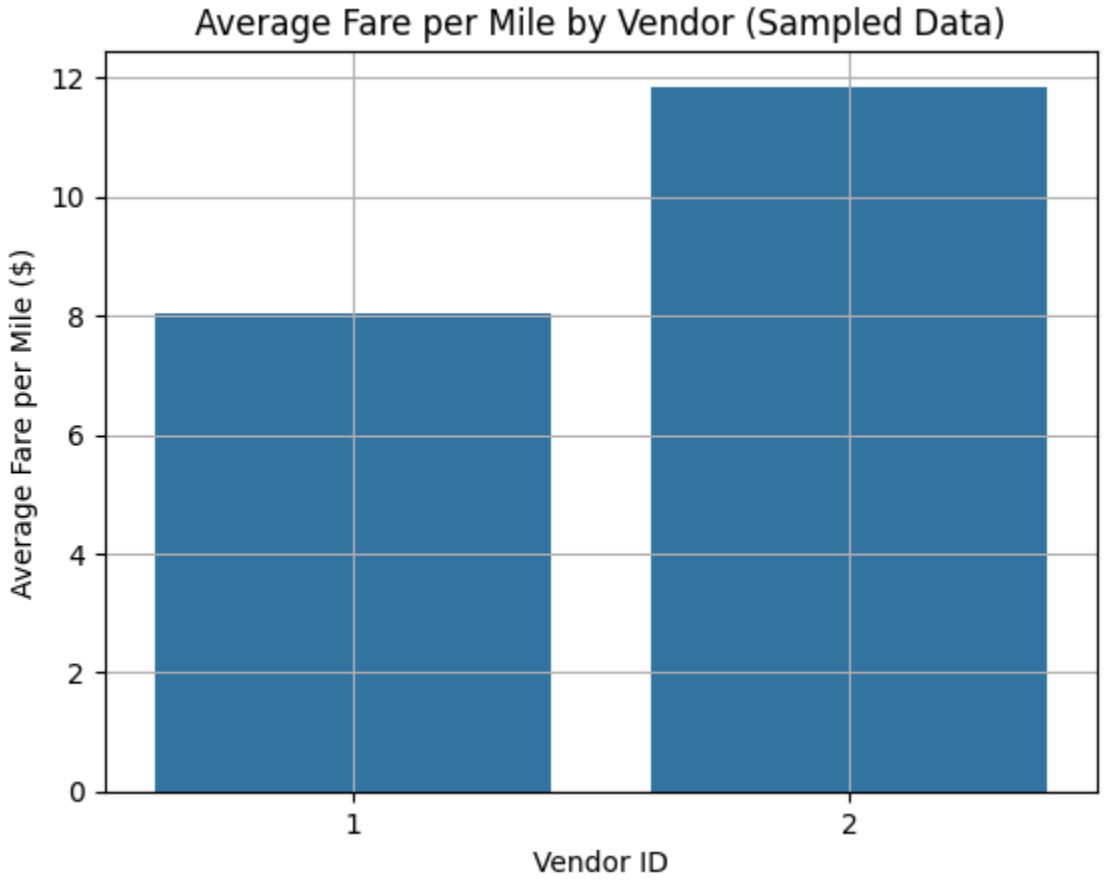


Find the average fare per mile by hours of the day and by days of the week

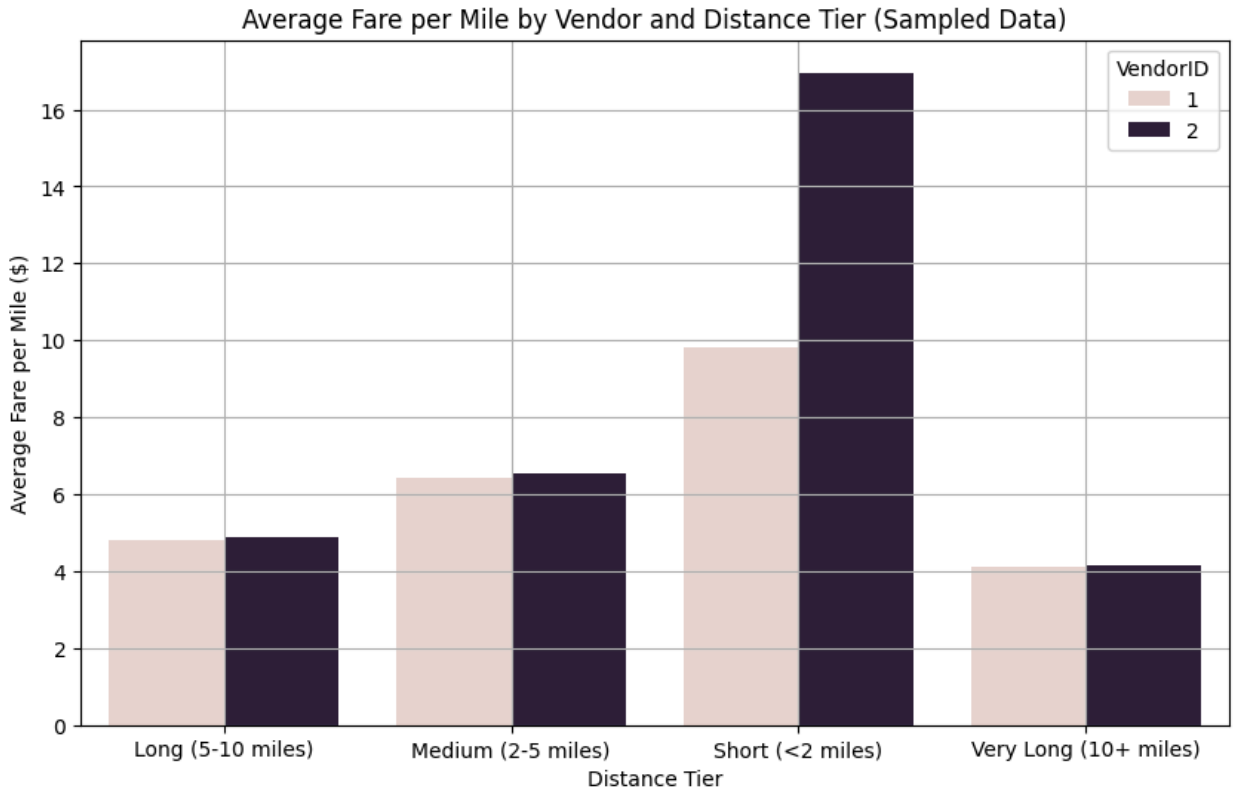




Analyse the average fare per mile for the different vendors



Compare the fare rates of different vendors in a distance-tiered fashion



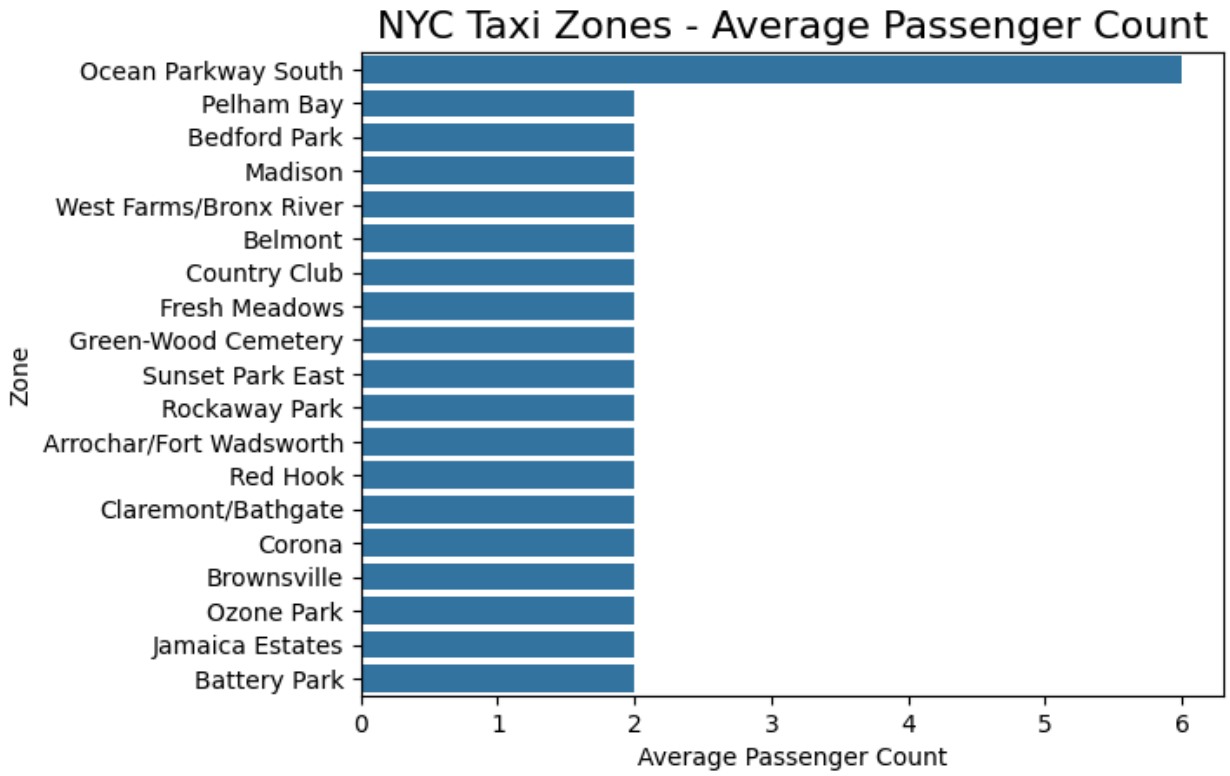
Analyse the tip percentages

distance_tier	tip_percentage
Very Long (10+ miles)	16.126536
Long (5-10 miles)	17.571591
Medium (2-5 miles)	18.954700
Short (<2 miles)	22.607886

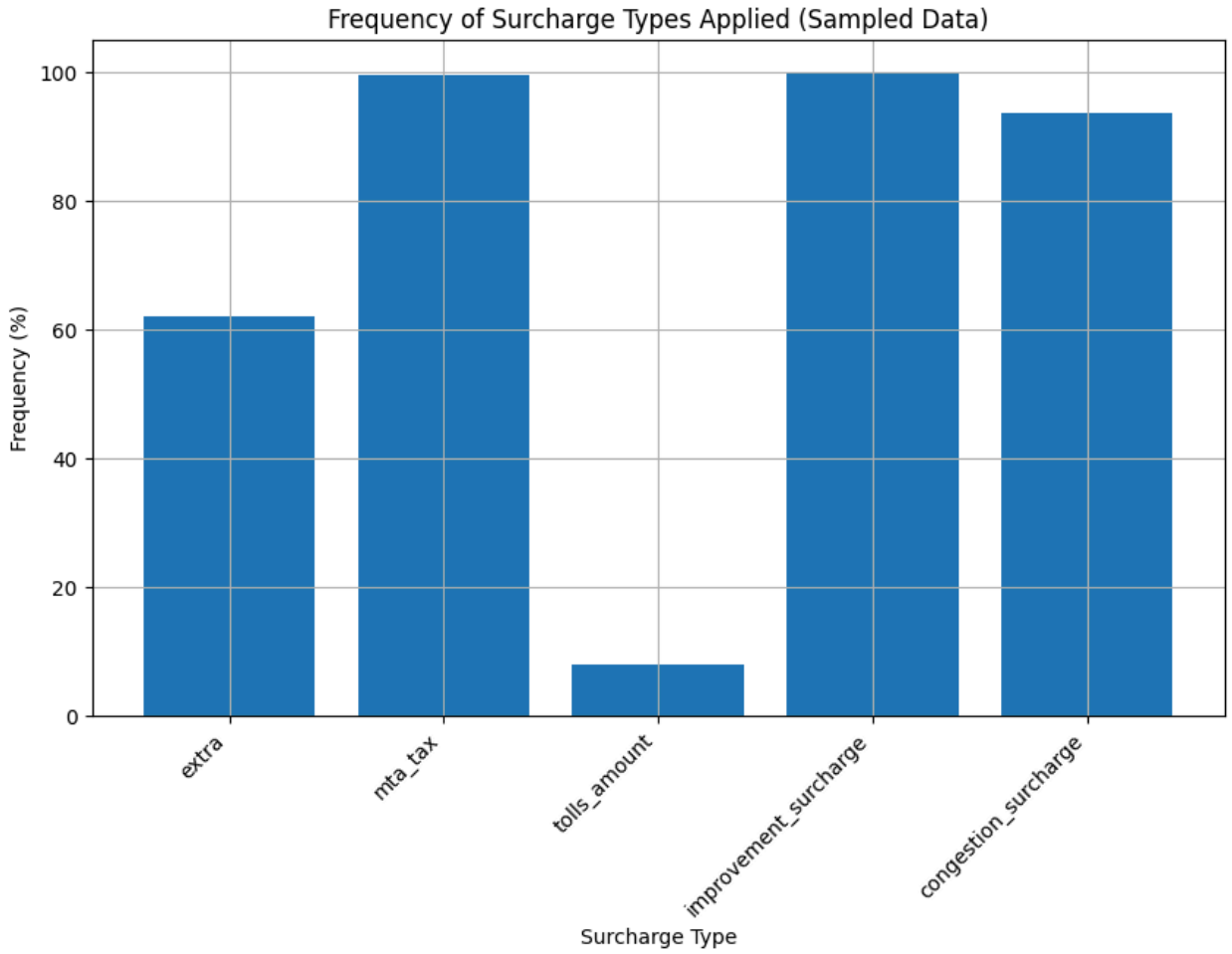
Analyse the trends in passenger count

	passenger_count	tip_percentage
3	4.0	17.598668
2	3.0	19.087426
1	2.0	19.816373
5	6.0	20.456117
4	5.0	20.503994
0	1.0	20.920909

Analyse the variation of passenger counts across zones



Analyse the pickup/dropoff zones or times when extra charges are applied more frequently.



Conclusions

Final Insights and Recommendations

When and where people ride The busiest time by far is the evening (between 6 PM and 8 PM), which actually sees more action than the morning commute. During the week, we see those typical "work rush" hours, but on weekends, demand is spread more evenly throughout the day. Also, most of our business is coming from just a handful of specific zones.

More revenue comes from the longer trips. Having more people in the car doesn't really change the fare since taxis charge by the trip, not by the person.

Some areas get tons of pickups but very few drop-offs (and vice versa). This means drivers often end up stuck in some spots or have to drive back empty to find their next fare. We need a better plan to get those cars back to where the people actually are.

Recommendations to optimize routing and dispatching based on demand patterns and operational inefficiencies.

Optimize the deployments

Focus most of our available drivers and cars during the busiest times, especially in the evening when demand is highest. On weekdays, keep a strong presence during the usual morning and evening rush hours. For weekends, ensure reliable coverage throughout the afternoon and evening, instead of just targeting short, intense peaks.

Maintain Vehicle Supply Between Zones

Our data shows that certain areas end up with too many cars, while others consistently need more vehicles. To fix this, we can create incentives for drivers to take rides that start in zones with excess cars and head toward areas where demand is high

Predictable Trends

Revenue stays fairly stable throughout the year, with no major ups and downs. This means we can plan staffing, maintenance, and resources more consistently, without having to ramp up or

down for certain times of the year. we should direct our attention to improving how we operate day-to-day and hour-by-hour.

Suggestions on strategically positioning cabs across different zones to make best use of insights uncovered by analysing trip trends across time, days and months.

We've pinpointed the zones that generate the most trips. We should always have a minimum number of cars in these key areas and add more during the peak times. We have seen—a few specific locations are responsible for a large share of our total demand.

Adjust Strategy for Weekdays vs. Weekends

- **Weekdays:** Position vehicles to catch the morning rush, then actively move them to prepare for the evening peak.
- **Weekends:** We can start later and focus our efforts on zones that stay busy from lunchtime through the late evening.

Anticipate the Hourly Flow

Demand changes sharply by the hour. To stay ahead:

- **Before a Surge:** Move cars into position 15-30 minutes before a known busy period starts.
- **During Lulls:** When things are slower, concentrate the fleet in the zones where pickups are still most likely. This keeps wait times short for customers and ensures we catch every possible trip.

Propose data-driven adjustments to the pricing strategy to maximize revenue while maintaining competitive rates with other vendors.

Fine-Tune Pricing for Different Trip Lengths

Current model of charging by distance is working well, as fare and distance are closely connected. However, we've found that short trips are actually more profitable per mile. We should introduce or adjust a minimum fare to better reward drivers for these quick rides, while keeping our rates attractive for longer trips.

Use Smart Pricing During Busy Times

Evening rush hours are our busiest and show the strongest demand, giving us a chance to introduce higher "peak pricing" during these times. The data confirms demand is concentrated in specific windows, which suggests customers are willing to pay a little extra when they need a ride most.

Make Fees Clear and Upfront

Our current trips include several types of fees. To build trust and improve the booking experience, we should show customers the full price—including any applicable surcharges—before they confirm their ride. This reduces surprises at the end of the trip and helps maintain satisfaction, while protecting our revenue.