Report: Optimising NYC Taxi Operations

Include your visualisations, analysis, results, insights, and outcomes. Explain your methodology and approach to the tasks. Add your conclusions to the sections.

1. Data Preparation

1.1. Loading the dataset

1.1.1. Sample the data and combine the files

The monthly Parquet files were processed in sequence. For each day and each hour, 5% of the records were randomly selected. This helped keep the time-based patterns while reducing the size of the data. All the selected records were then combined into a single Dataframe for the entire year.

2. Data Cleaning

2.1. Fixing Columns

2.1.1. Fix the index

The index was reset to ensure a clean, continuous index after concatenation.

2.1.2. Combine the two airport_fee columns

The dataset contained two columns, Airport_fee and airport_fee, representing the same variable with inconsistent naming. An exploratory check showed that values were mutually exclusive across rows. These columns were safely merged using a null-preserving strategy (combine_first) into a single airport_fee column, after which the redundant columns were removed.

2.1.3. Fixing negative values

Negative values were identified in several surcharge-related columns and in trip duration. Since fare components such as taxes and surcharges are non-negative by definition, negative values were corrected to zero. Records with

negative total fare or negative trip duration were removed, as they represent invalid trips. The number of affected records was minimal and did not materially impact the dataset.

2.2. Handling Missing Values

2.2.1. Find the proportion of missing values in each column

The proportion of missing values for each column was computed as the ratio of null entries to total records. Most columns had little or no missing data. A few columns had some missing values, so they are reviewed separately to see if they are important to keep.

2.2.2. Handling missing values in passenger_count

Rows containing missing values were identified using row-wise null checks. Missing values in passenger_count are imputed using the median, which is considered robust to skewness and outliers and appropriate for discrete count data

Trips with passenger_count equal to zero are treated as data entry errors. These values are converted to missing and imputed using the median passenger count to preserve realistic trip characteristics without discarding data.

2.2.3. Handle missing values in RatecodeID

RatecodeID is a categorical identifier representing fare rules. Missing values are imputed using the mode (most frequent category), which corresponds to the standard rate and preserves the categorical distribution without introducing artificial values.

2.2.4. Impute NaN in congestion_surcharge

congestion_surcharge applies only to trips entering designated congestion zones. Missing values therefore indicate that the surcharge was not applicable. These values are imputed as 0.0 to preserve fare consistency.”

2.2.5.

store_and_fwd_flag indicates whether trip data was temporarily stored on the meter and forwarded later. Missing values were imputed as 'N', assuming no store-and-forward event occurred, which aligns with domain conventions and preserves categorical consistency.”

2.3. Handling Outliers and Standardising Values

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

Trips with passenger_count greater than 6 were removed, as standard NYC yellow taxis have a maximum passenger capacity of six. These records were treated as invalid entries.

2.3.2.

A total of 35 trips exhibited near-zero recorded distance with unusually high fares. Given the extremely small proportion of such records and their potential to represent data recording or fare rule anomalies, these entries were retained but explicitly flagged for downstream analysis.

2.3.3.

Trips with zero recorded distance and zero fare despite differing pickup and dropoff zones were identified as invalid records. These entries were removed as they violate fundamental trip consistency rules.

2.3.4.

Extreme values in fare, total amount, and tip amount were identified using predefined thresholds (fare and total amount > $500, tip amount > $100). These records represent a very small proportion of the dataset and are likely due to rare edge cases or data recording anomalies. Rather than removing these observations, they were flagged to preserve transparency and enable separate analysis without distorting overall distributions.

2.3.5.

Trips with unusually long distances or large pickup–dropoff time gaps were flagged as potential anomalies and retained for separate analysis. Invalid RatecodeID values (99) were treated as missing and imputed using the mode to preserve categorical consistency without removing records.

3. Exploratory Data Analysis

3.1. General EDA: Finding Patterns and Trends

3.1.1. Classify variables into categorical and numerical

*** `VendorID`: Categorical**

*** `tpep_pickup_datetime`: DateTime**

*** `tpep_dropoff_datetime`: DateTime**

*** `passenger_count`: Numerical**

*** `trip_distance`: Numerical**

*** `RatecodeID`: Categorical**

*** `PULocationID`: Categorical**

*** `DOLocationID`: Categorical**

*** `payment_type`: Categorical**

*** `pickup_hour`: Numerical**

*** `trip_duration`: Numerical**

All the below are numerical variables

*** `fare_amount`**

*** `extra`**

*** `mta_tax`**

*** `tip_amount`**

*** `tolls_amount`**

*** `improvement_surcharge`**

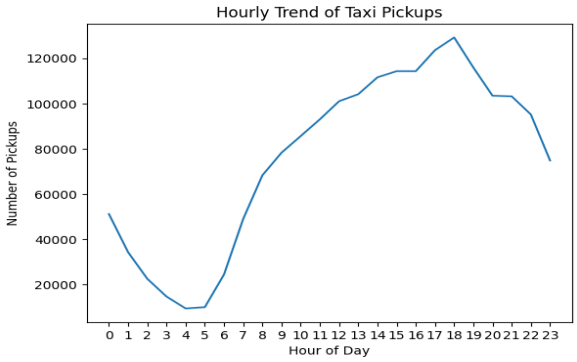
*** `total_amount`**

*** `congestion_surcharge`**

*** `airport_fee`**

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

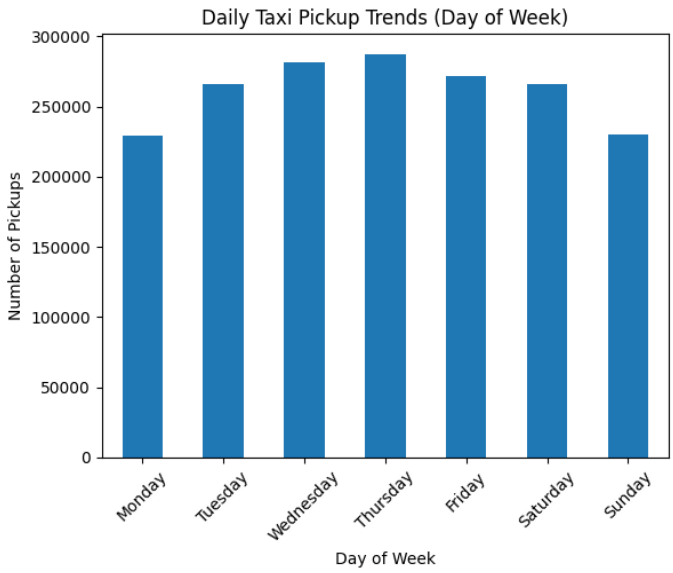
1.1.1. Analysis of the hourly trend of taxi pickups:



Taxi pickups show a **clear daily cycle** with predictable peaks and troughs.

Lowest demand: Early morning (4–5 AM)
Highest demand: Evening (6–7 PM)
Two demand surges: Morning commute and evening commute, with the evening peak being stronger.

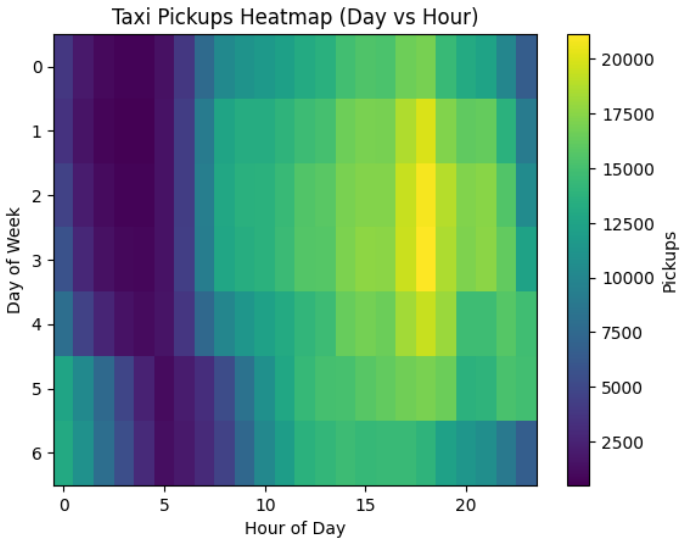
Analysis of the daily trends of taxi pickups



Taxi pickups increase from Monday through Thursday, peaking on Thursday, indicating strong weekday commuter demand.

Pickups decline slightly on Friday and remain steady on Saturday, reflecting a shift from work-related to leisure travel.

Sunday shows the lowest pickup volume, suggesting reduced overall travel activity compared to the rest of the week.

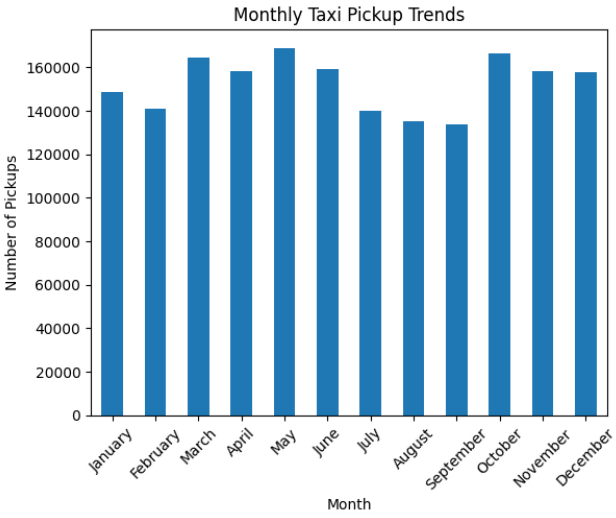


Taxi pickups are lowest during late night and early morning hours (midnight–5 AM) across all days of the week.

Weekdays show clear demand peaks in the afternoon and early evening, especially around 4–7 PM, reflecting commuter travel.

Weekend patterns are flatter, with relatively higher activity from late morning to evening, indicating more leisure-driven trips.

Analysis in monthly trends of taxi pickups



Taxi pickups show seasonal variation across the year, with higher activity during spring and early summer, peaking around May.

A noticeable dip occurs in mid to late summer (July–September), indicating reduced travel demand during these months.

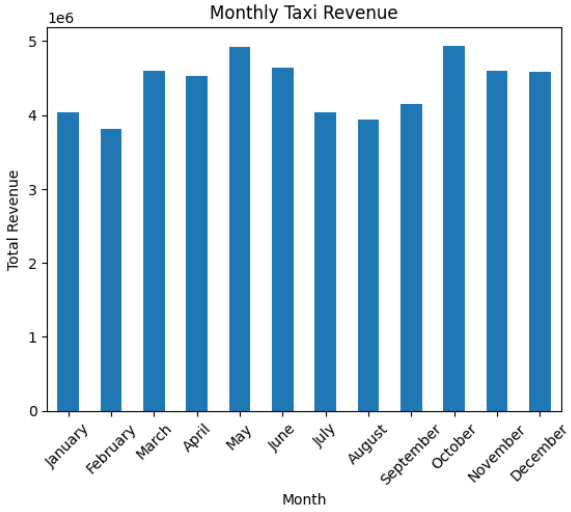
Pickups rise again in autumn, with October showing strong activity, and remain relatively stable toward the end of the year.

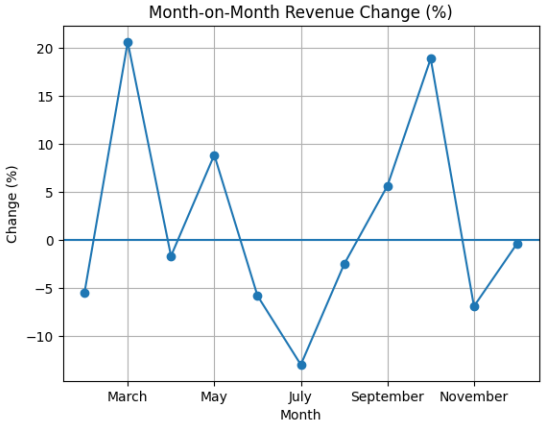
1.1.2. Filter out the zero/negative values in fares, distance and tips

An examination of key financial and distance variables revealed the presence of zero and negative values. While zero values in tip amount are expected, zero or negative values in fare amount, total amount, and trip distance indicate data quality issues and were flagged or removed as appropriate

Trips with zero recorded distance were not removed indiscriminately, as zero distance can occur for very short trips within the same pickup and drop-off zone. Only zero-distance trips with differing pickup and drop-off zones were treated as invalid and removed.

1.1.3. Analyse the monthly revenue trends



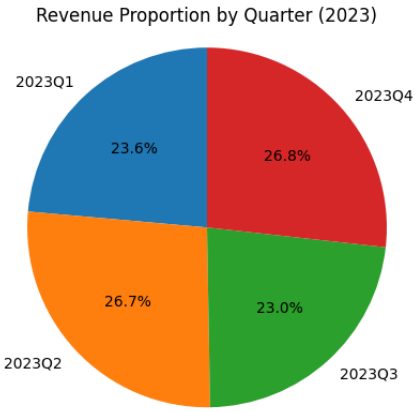


Monthly taxi revenue shows clear seasonal variation, increasing from February through May, with May recording the highest revenue.

Revenue declines during mid-year (July–August), indicating reduced travel demand in the summer months.

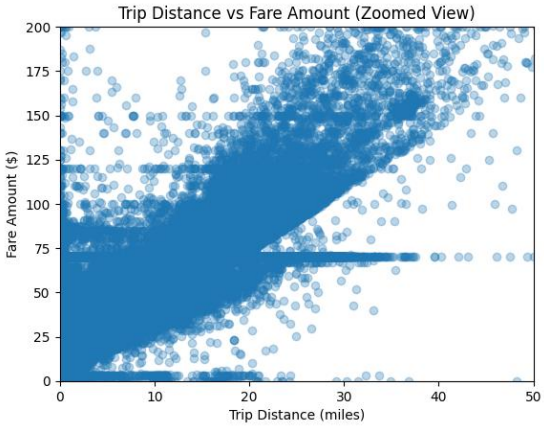
Earnings recover strongly in autumn, with October showing another peak, and remain relatively stable toward the end of the year.

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



The pie chart shows that revenue is fairly evenly distributed across all four quarters in 2023, with Q2 and Q4 contributing slightly larger shares. This indicates moderate seasonality, with stronger performance during spring and autumn.

1.1.5. Analyse and visualise the relationship between distance and fare amount



The scatter plot shows a clear positive relationship between trip distance and fare amount, indicating that fares generally increase as trip distance increases.

The widening spread at higher distances reflects fare variability caused by traffic conditions, surcharges, and differing fare rules.

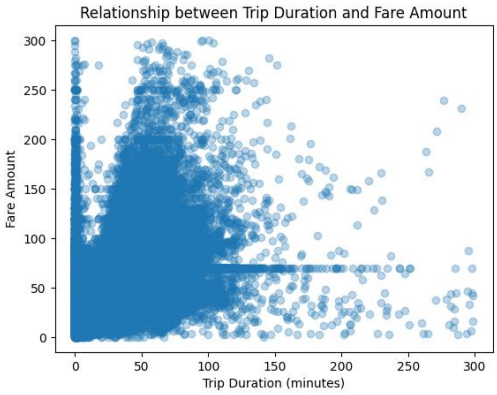
Horizontal bands in the plot suggest the presence of flat fares or capped pricing, especially for certain distance ranges.

Unfiltered Pearson r = 0.01

Filtered Pearson r ≈ 0.95

The unfiltered Pearson correlation is near zero (r ≈ 0.01), indicating that extreme outliers and anomalous records obscure the true relationship. After filtering realistic trip ranges, the Pearson correlation increases to approximately 0.95, revealing a strong positive relationship between trip distance and fare amount.

1.1.6. Analyse the relationship between fare/tips and trips/passengers

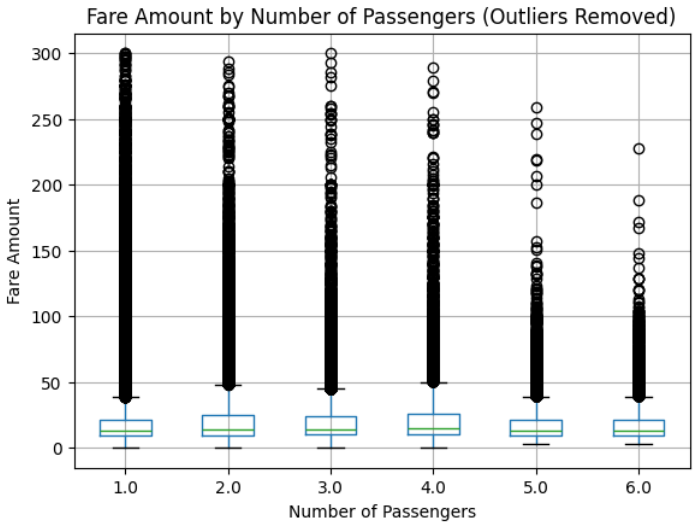


Correlation value = 0.82

The scatter plot shows a general positive relationship between trip duration and fare amount, with fares tending to increase as trip duration increases.

The wide vertical spread at similar durations indicates substantial fare variability, driven by factors such as distance traveled, traffic conditions, surcharges, and flat-fare rules.

The presence of horizontal bands suggests fixed or capped fares for certain trip types, while long-duration trips with relatively low fares highlight congestion or waiting-time effects.

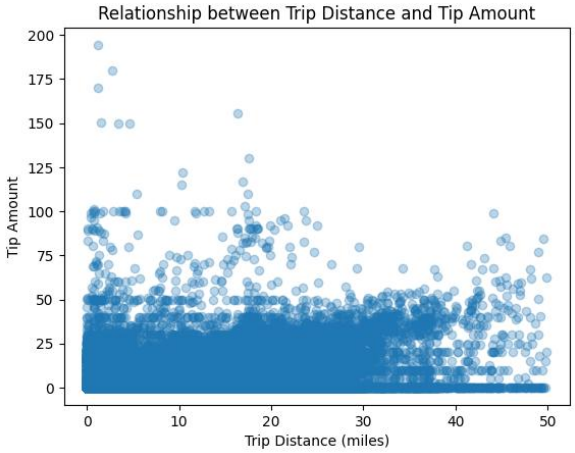


Correlation value = 0.01

The box plots show that median fare amounts are broadly similar across all passenger counts, with only minor variation as the number of passengers increases.

The spread and presence of outliers within each passenger group indicate that fare variability is driven more by trip distance and duration than by the number of passengers.

Overall, the plot confirms that fare amount is not strongly dependent on passenger count, supporting the expectation that taxi fares are primarily distance- and time-based.



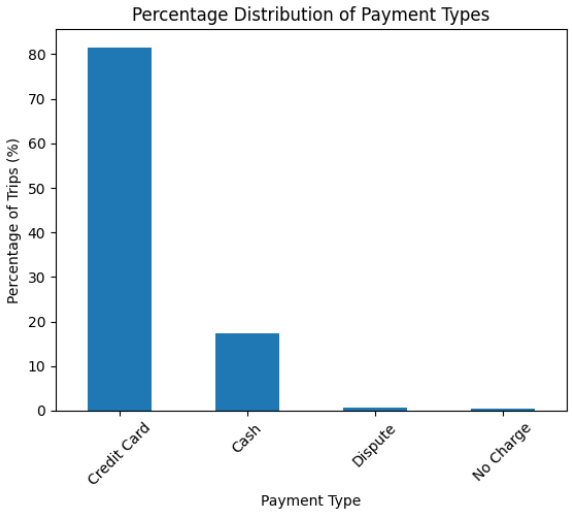
Correlation value = 0.59

The scatter plot shows a weak positive relationship between trip distance and tip amount, with tips tending to increase slightly for longer trips.

A large concentration of points near zero tip across all distances highlights high variability in tipping behaviour and the prevalence of no-tip trips.

Overall, the plot indicates that trip distance alone is not a strong determinant of tip amount, with factors such as payment method and rider preference playing a larger role.

1.1.7. Analyse the distribution of different payment types



The bar chart shows that credit card payments dominate taxi trips, accounting for the vast majority of transactions (around four-fifths of all trips).

Cash payments form a much smaller share, indicating a clear shift toward digital payment methods.

Dispute and no-charge transactions are negligible, suggesting that fare waivers and payment issues are rare in the dataset.

1.1.8. Load the taxi zones shapefile and display it

	OBJECTID	Shape_Leng	Shape_Area	zone	LocationID	borough	geometry
0	1	0.116357	0.000782	Newark Airport	1	EWR	POLYGON ((933100.918 192536.086, 933091.011 19...
1	2	0.433470	0.004866	Jamaica Bay	2	Queens	MULTIPOLYGON (((1033269.244 172126.008, 103343...
2	3	0.084341	0.000314	Allerton/Pelham Gardens	3	Bronx	POLYGON ((1026308.77 256767.698, 1026495.593 2...
3	4	0.043567	0.000112	Alphabet City	4	Manhattan	POLYGON ((992073.467 203714.076, 992068.667 20...
4	5	0.092146	0.000498	Arden Heights	5	Staten Island	POLYGON ((935843.31 144283.336, 936046.565 144...

1.1.9. Merge the zone data with trips data

```
df_merged[['PULocationID', 'LocationID', 'zone', 'borough']].head(
```

	PULocationID	LocationID	zone	borough
0	142	142.0	Lincoln Square East	Manhattan
1	132	132.0	JFK Airport	Queens
2	249	249.0	West Village	Manhattan
3	144	144.0	Little Italy/NoLiTa	Manhattan
4	79	79.0	East Village	Manhattan

```
#missing values
```

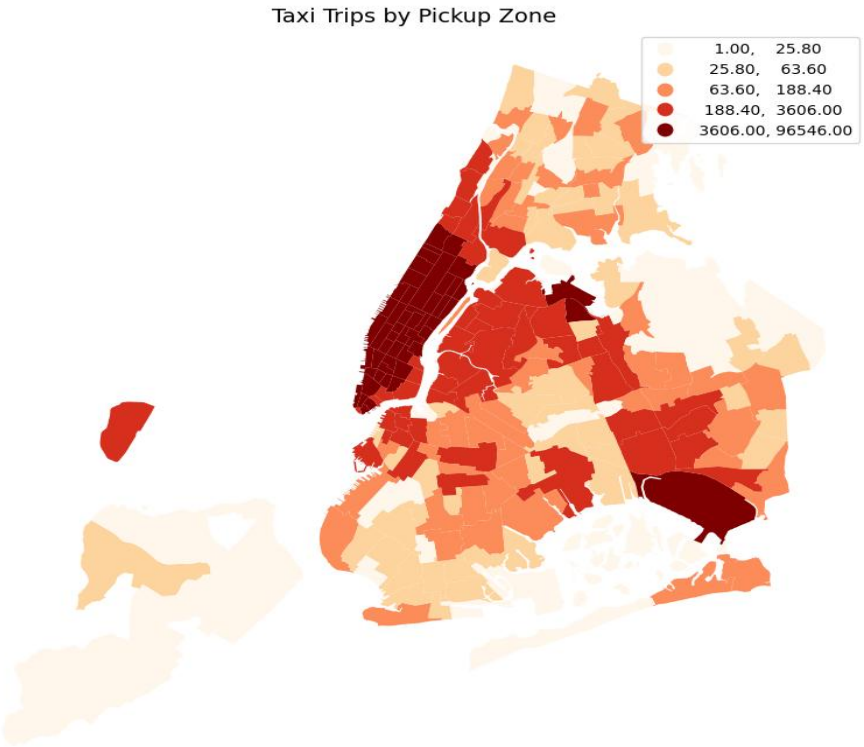
1.1.10. Find the number of trips for each zone/location ID

	PULocationID	Total_Trips
0	1	200
1	2	2
2	3	34
3	4	1843
4	5	10

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

	LocationID	zone	borough	Total_Trips
0	1	Newark Airport	EWR	200.0
1	2	Jamaica Bay	Queens	2.0
2	3	Allerton/Pelham Gardens	Bronx	34.0
3	4	Alphabet City	Manhattan	1843.0
4	5	Arden Heights	Staten Island	10.0

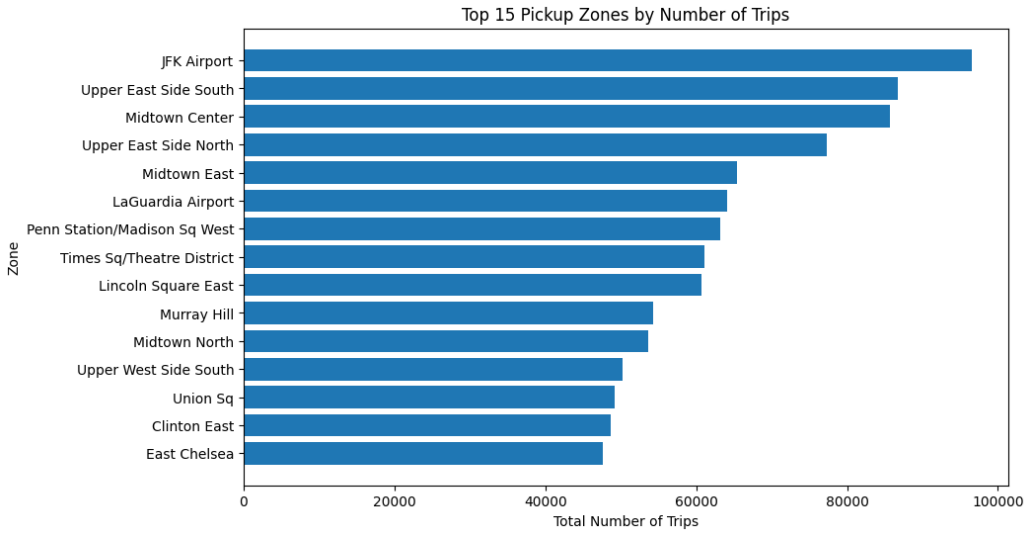
1.1.12. Plot a map of the zones showing number of trips



The choropleth map shows clear spatial concentration of taxi pickups, with darkest zones representing the highest trip volumes.

Central Manhattan and major transit/airport-adjacent areas stand out as high-demand pickup zones, reflecting dense commercial activity and transportation hubs.

Peripheral and residential zones appear in lighter shades, indicating lower pickup frequency and more localized travel demand.



The bar chart shows that JFK Airport is the busiest pickup zone, indicating a very high volume of taxi demand associated with air travel.

Central Manhattan zones such as Midtown Center, Upper East Side, Times Square, and Penn Station areas dominate the top rankings, reflecting dense commercial activity, tourism, and major transit hubs.

Overall, the chart highlights a strong spatial concentration of taxi pickups, where a small number of high-activity zones account for a disproportionately large share of total trips.

1.1.13. Conclude with results

Taxi demand shows clear temporal patterns, peaking during weekday afternoons and evenings, with Tuesday–Thursday being the busiest days and spring and autumn months recording the highest activity.

Revenue trends closely mirror trip volumes, with strong monthly seasonality and slightly higher contributions in Q2 and Q4, indicating balanced but seasonally influenced earnings.

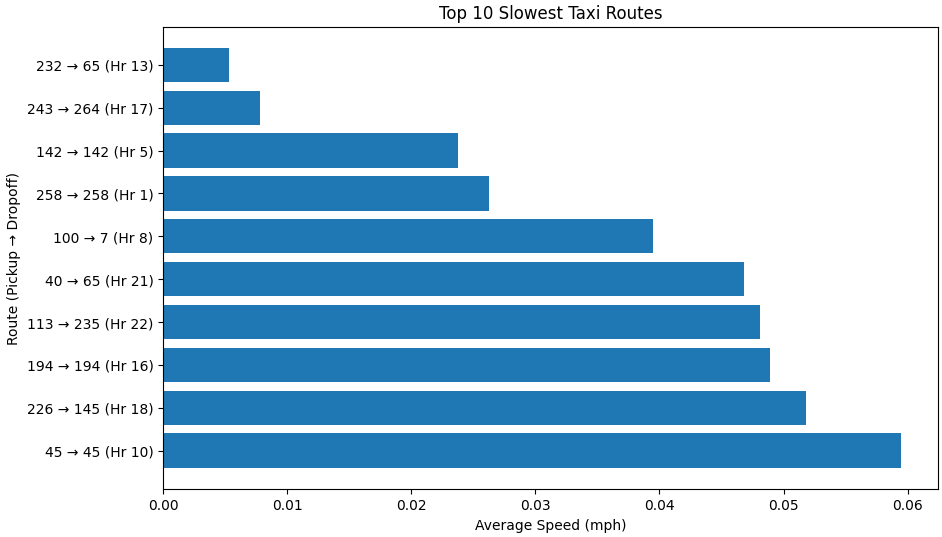
Fare amount is strongly driven by trip distance and moderately by trip duration, while passenger count has minimal impact, confirming distance- and time-based pricing.

Tip amounts exhibit high variability and show only a weak positive relationship with trip distance, suggesting that tipping is more influenced by rider behavior and payment method.

Spatial analysis highlights central Manhattan and major transit or airport zones as the busiest pickup locations, with significantly higher trip volumes than peripheral residential areas.

1.2. Detailed EDA: Insights and Strategies

1.2.1. Identify slow routes by comparing average speeds on different routes

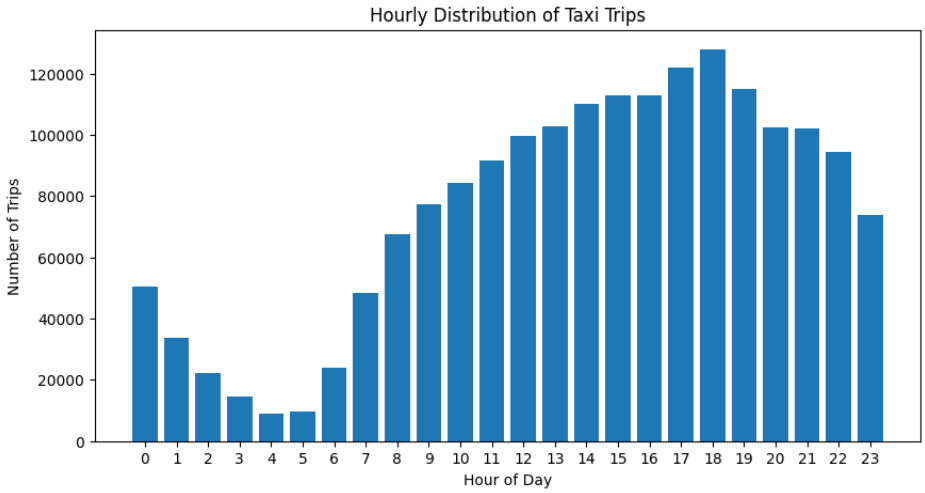


The bar chart highlights the top 10 slowest taxi routes, showing extremely low average speeds, which indicates severe congestion or frequent stop-and-go conditions.

Many of the slowest routes occur during peak traffic hours (morning and evening), suggesting that congestion rather than route length is the primary cause of delay.

Routes with the same pickup and drop-off zone IDs appear among the slowest, likely reflecting short intra-zone trips with high idle or waiting time, such as dense commercial or downtown areas.

1.2.2. Calculate the hourly number of trips and identify the busy hours

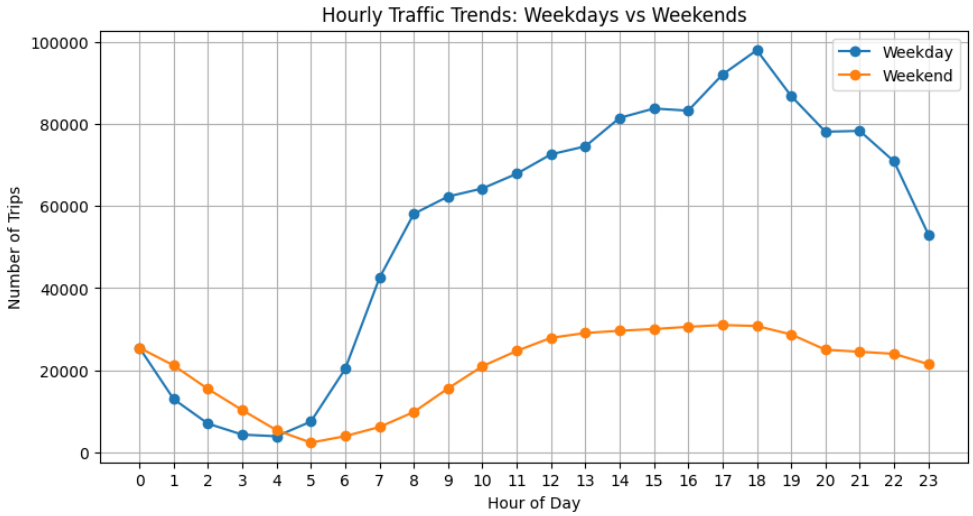


Busiest hour: 18:00 – 19:00 with 127,873 trips

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

	Hour_Range	Estimated_Total_Trips
18	18:00–19:00	2557460
17	17:00–18:00	2441360
19	19:00–20:00	2296620
15	15:00–16:00	2256900
16	16:00–17:00	2254740

1.2.4. Compare hourly traffic on weekdays and weekends

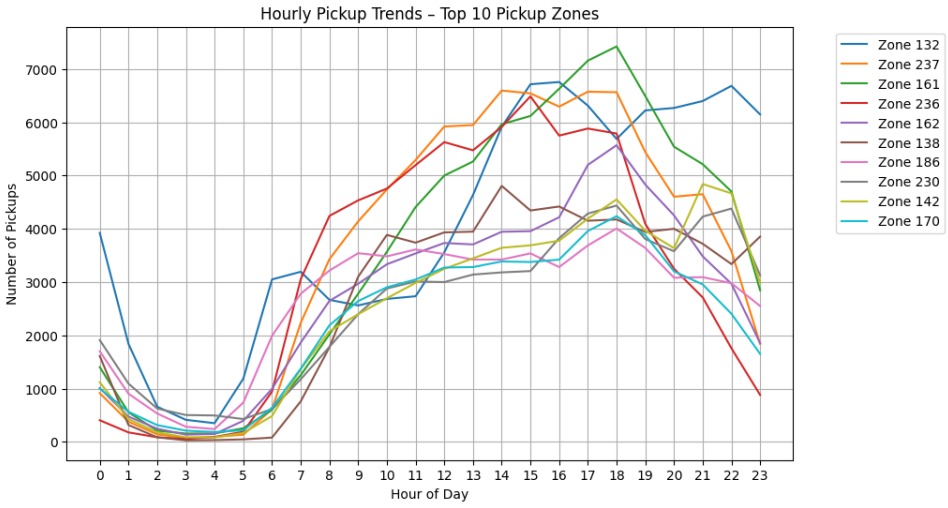


Weekdays show strong commuting peaks, with a sharp rise in trips during the morning (7–9 AM) and a higher peak in the late afternoon to early evening (5–7 PM), reflecting work-related travel.

Weekends follow a flatter pattern, with lower early-morning demand and gradually increasing activity from late morning, peaking in the afternoon and evening, driven by leisure and social travel.

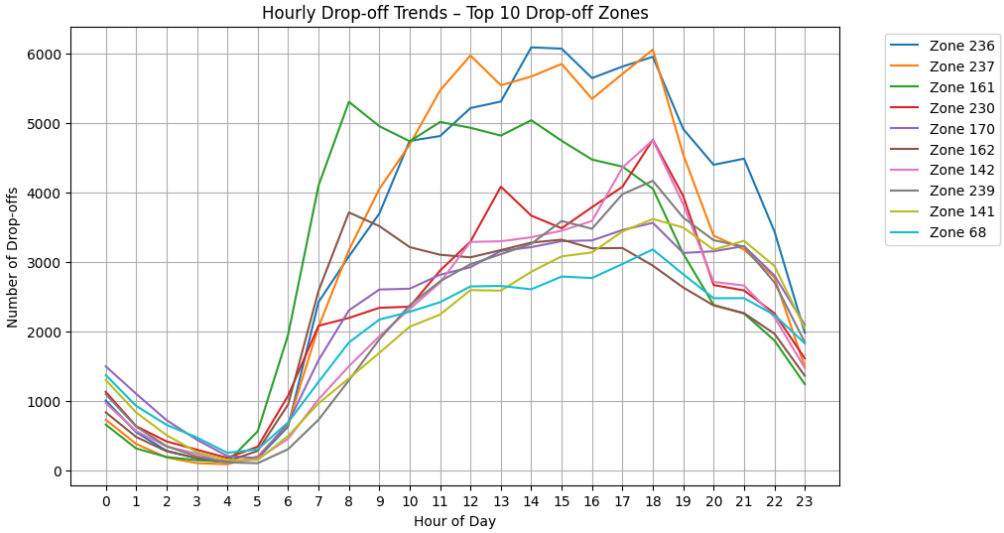
Overall, weekday traffic volumes are significantly higher than weekend volumes, especially during traditional commuting hours.

1.2.5. Identify the top 10 zones with high hourly pickups and drops



The line chart shows that all top 10 pickup zones follow a similar daily pattern, with very low pickup activity during late-night and early-morning hours and a sharp rise starting around 6–8 A

Peak pickup volumes occur between early afternoon and early evening (approximately 2–6 PM), indicating strong commuter and commercial activity in these high-demand zones.



The chart shows that drop-off activity in the top 10 zones rises sharply from early morning, with a pronounced increase starting around 6–8 AM, reflecting inbound commuter travel.

Peak drop-offs occur slightly earlier or around mid-day to early evening (approximately 12–6 PM), suggesting these zones act as key destination hubs such as business districts, transit nodes, or commercial areas.

Compared to pick-up trends, drop-off volumes in these zones peak earlier and decline more steadily in the evening, highlighting directional travel patterns where people arrive at these zones earlier and depart later.

1.2.6. Find the ratio of pickups and dropoffs in each zone

Top 10 highest pickup/dropoff ratio

	LocationID	pickup_count	dropoff_count	pickup_dropoff_ratio
69	70.0	8302.0	976.0	8.506148
127	132.0	96546.0	20914.0	4.616334
133	138.0	64077.0	22170.0	2.890257
181	186.0	63236.0	40009.0	1.580544
42	43.0	30661.0	22298.0	1.375056
109	114.0	24059.0	17507.0	1.37425
244	249.0	40324.0	30387.0	1.327015
157	162.0	65457.0	52102.0	1.256324
156	161.0	85701.0	71445.0	1.199538
99	100.0	30087.0	25261.0	1.191045

Top 10 lowest pickup/dropoff ratio

	LocationID	pickup_count	dropoff_count	pickup_dropoff_ratio
171	176.0	0.0	12.0	0.0
98	99.0	0.0	3.0	0.0
29	30.0	0.0	18.0	0.0
240	245.0	0.0	30.0	0.0
26	27.0	1.0	38.0	0.026316
216	221.0	1.0	34.0	0.029412
252	257.0	23.0	751.0	0.030626
0	1.0	200.0	5312.0	0.037651
193	198.0	44.0	981.0	0.044852
110	115.0	1.0	22.0	0.045455

The pickup-to-dropoff ratio highlights directional travel behavior across zones. Zones with high ratios act primarily as trip origins, while zones with low ratios function as major destinations.

1.2.7. Identify the top zones with high traffic during night hours

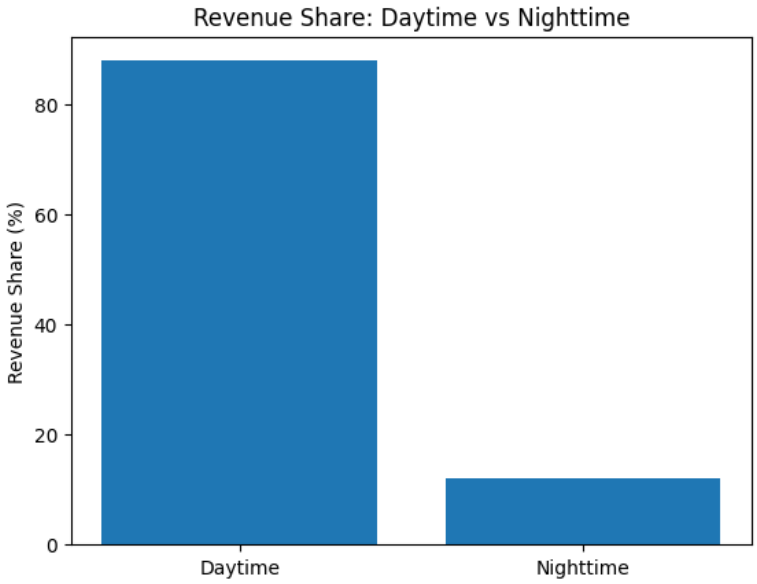
Top 10 Night Pickup Zones

	zone	borough	Night_Pickups
0	East Village	Manhattan	15503
1	JFK Airport	Queens	14517
2	West Village	Manhattan	12445
3	Clinton East	Manhattan	10432
4	Lower East Side	Manhattan	9615
5	Greenwich Village South	Manhattan	8737
6	Times Sq/Theatre District	Manhattan	8180
7	Penn Station/Madison Sq West	Manhattan	6943
8	Midtown South	Manhattan	6130
9	East Chelsea	Manhattan	6016

Top 10 Night Dropoff Zones

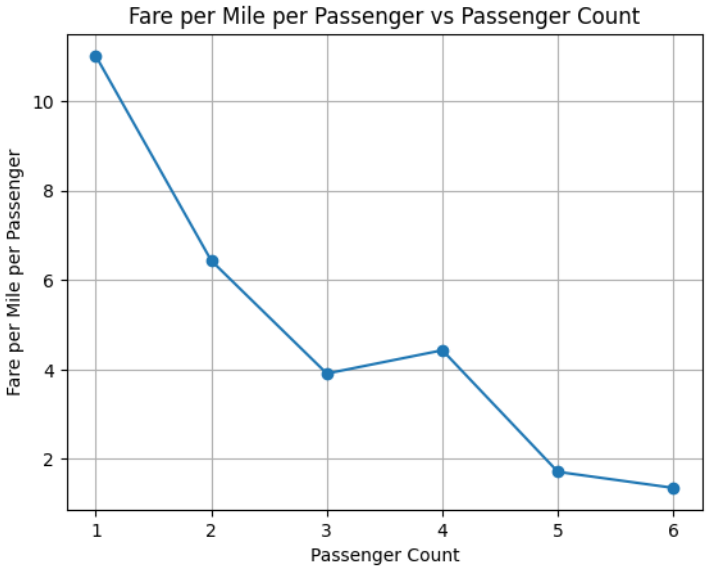
	zone	borough	Night_Dropoffs
0	East Village	Manhattan	8295
1	Clinton East	Manhattan	6864
2	Murray Hill	Manhattan	6253
3	East Chelsea	Manhattan	5824
4	Gramercy	Manhattan	5734
5	Lenox Hill West	Manhattan	5259
6	Yorkville West	Manhattan	4968
7	West Village	Manhattan	4935
8	Times Sq/Theatre District	Manhattan	4624
9	Sutton Place/Turtle Bay North	Manhattan	4371

1.2.8. Find the revenue share for nighttime and daytime hours



	time_period	Total_Revenue	Revenue_Share_%
0	Daytime	46406220.20	87.94
1	Nighttime	6365147.18	12.06

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

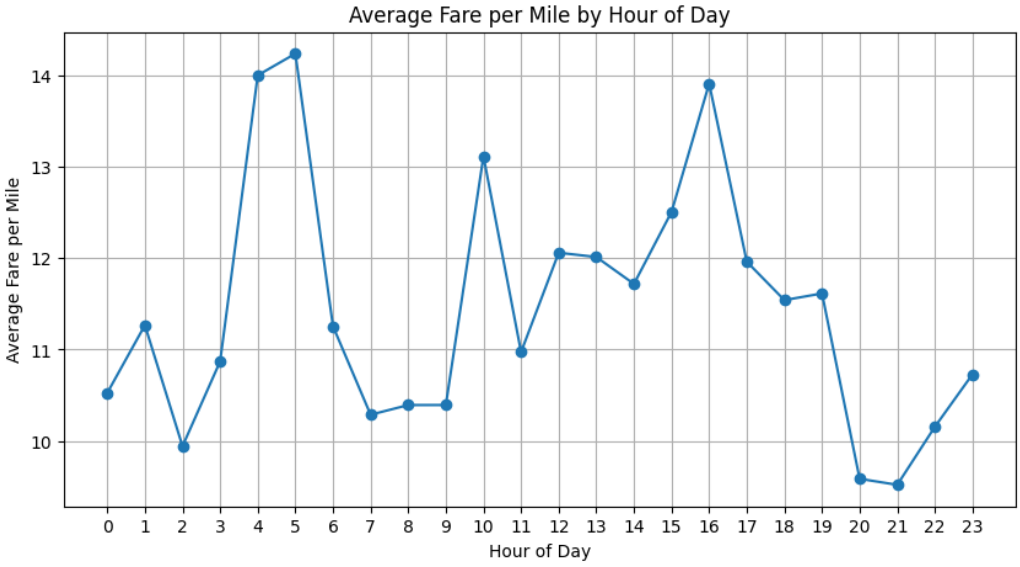


The plot shows a clear decreasing trend in fare per mile per passenger as the number of passengers increases, indicating strong cost-sharing benefits for group travel.

Single-passenger trips have the highest per-passenger cost, while trips with 5–6 passengers have the lowest cost per passenger per mile.

The slight deviation around 4 passengers reflects sample variability rather than a change in pricing structure, confirming that taxi fares are charged per trip rather than per passenger.

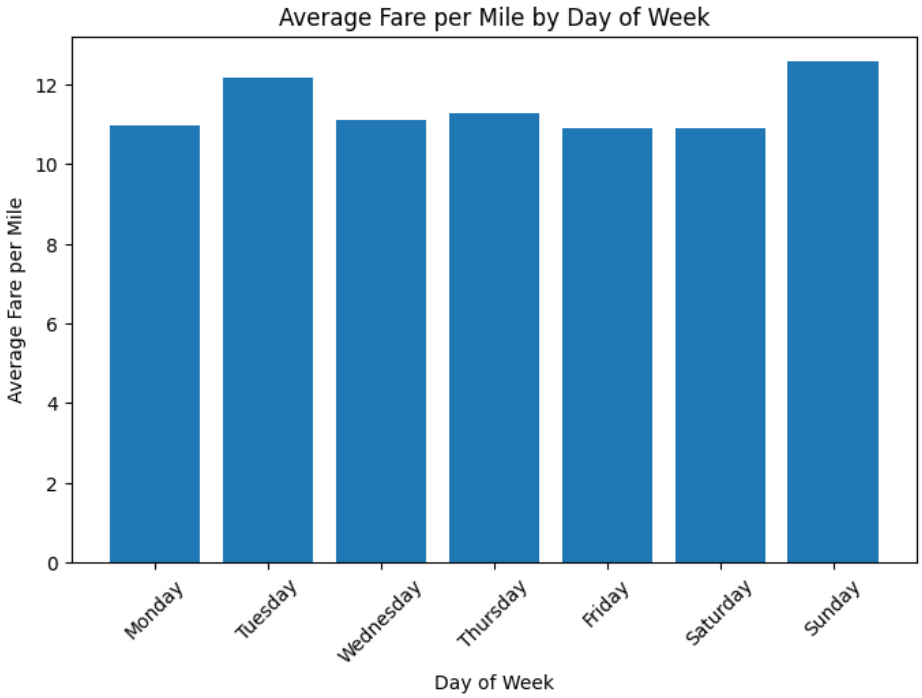
1.2.10. Find the average fare per mile by hours of the day and by days of the week



The plot shows that average fare per mile varies noticeably across the day, with higher values during early-morning and late-afternoon hours, reflecting shorter trips, congestion effects, and fare minimums.

Lower fare-per-mile values during late evening and night hours indicate longer or smoother trips with less stop-and-go traffic.

Overall, the pattern highlights how time-of-day traffic conditions and trip characteristics influence fare efficiency rather than changes in pricing policy.

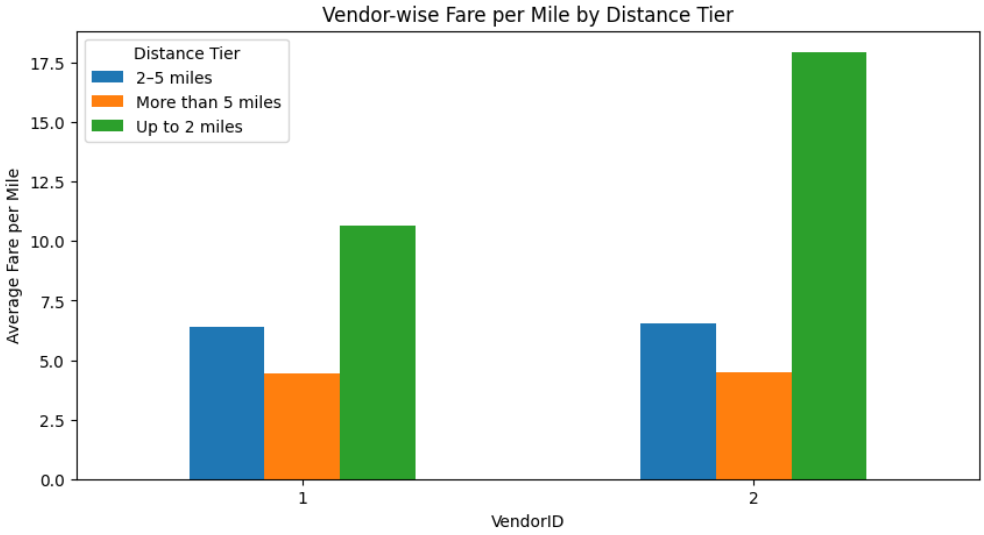


The bar chart shows clear variation in average fare per mile across the week, with Sunday recording the highest fare per mile, followed by Tuesday.

Weekdays exhibit relatively stable values, reflecting consistent commuting patterns and longer, more efficient trips.

The higher fare per mile on weekends, especially Sunday, likely results from shorter leisure trips, increased congestion around entertainment areas, and minimum fare effects, rather than changes in pricing.

Analyse the average fare per mile for the different vendors



The chart shows that average fare per mile is highest for short trips (up to 2 miles) for both vendors, reflecting minimum fare effects and short-trip inefficiency.

As trip distance increases, fare per mile decreases and stabilizes, with the lowest values observed for trips longer than 5 miles, indicating better distance efficiency on longer journeys.

Vendor 2 consistently exhibits a higher fare per mile than Vendor 1 for short trips, while fare differences between vendors narrow for medium (2–5 miles) and long trips, suggesting broadly similar pricing structures at larger distances.

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

Short trips (up to 2 miles) show the highest fare per mile for both vendors, driven by minimum fare rules and fixed base charges. Vendor 2 charges noticeably higher per mile than Vendor 1 in this tier, indicating stronger short-trip pricing effects.

For medium-distance trips (2–5 miles), fare per mile drops substantially for both vendors and the pricing gap narrows, suggesting similar metered rate structures beyond the base fare.

On long trips (more than 5 miles), fare per mile is the lowest and nearly identical across vendors, reflecting distance efficiency and convergence of pricing models.

Vendor differences are most pronounced for short trips due to base fare effects, while distance increasingly dominates pricing behavior as trip length increases, leading to comparable fare efficiency across vendors on longer journeys.

1.2.11. Analyse the tip percentages

payment_method	Low Tip (<10%)	High Tip (>25%)
Cash	0.66	0.0
Credit Card	0.30	1.0
Dispute	0.03	0.0
No Charge	0.01	0.0

The table shows a strong dependence of tipping behavior on payment method.

Low-tip trips (<10%) are dominated by cash payments (66%), indicating that cash transactions often result in little or unrecorded tipping.

In contrast, 100% of high-tip trips (>25%) are paid by credit card, clearly demonstrating that digital payments strongly encourage proportional tipping.

Dispute and No Charge trips contribute negligibly to tipping behavior and are largely irrelevant for tip analysis.

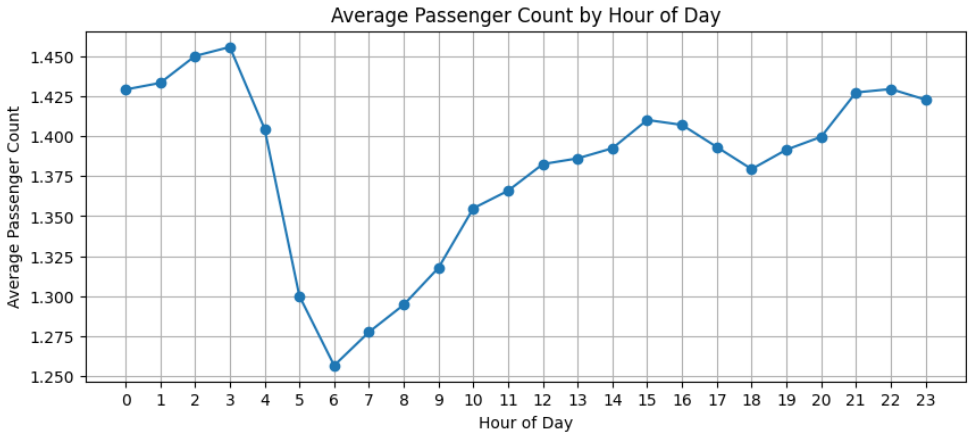
time_of_day	Low Tip (<10%)	High Tip (>25%)
Afternoon (12–5 PM)	152003	222668
Evening (5–9 PM)	110292	235337
Late Morning (9–12 PM)	70345	105649
Night (9–12 AM)	61955	129160
Morning (5–9 AM)	40502	61938
Late Night / Early Morning (12–5 AM)	35989	58650

The table shows that high-tip trips (>25%) exceed low-tip trips (<10%) across all time-of-day categories, indicating generally positive tipping behavior throughout the day.

Evening (5–9 PM) and Afternoon (12–5 PM) have the highest counts of high-tip trips, aligning with peak travel demand, credit-card usage, and service-oriented trips.

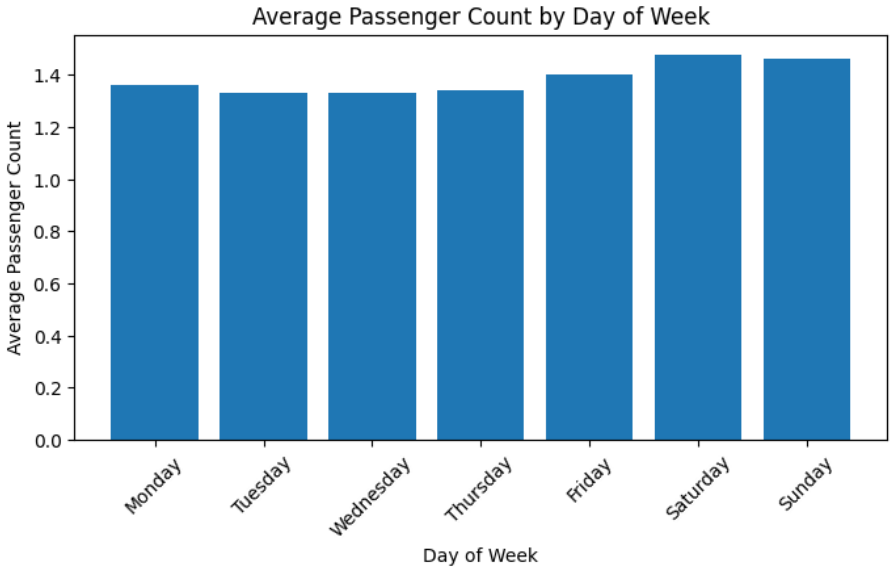
Late night and early morning hours (12–5 AM) have the lowest overall volumes but still show a higher proportion of high-tip trips than low-tip trips, suggesting appreciation for late-hour service.

1.2.12. Analyse the trends in passenger count



By hour of day, average passenger count is highest during late-night and early-morning hours (around 12–3 AM) and again during late evening (9–11 PM), reflecting social, leisure, and group travel.

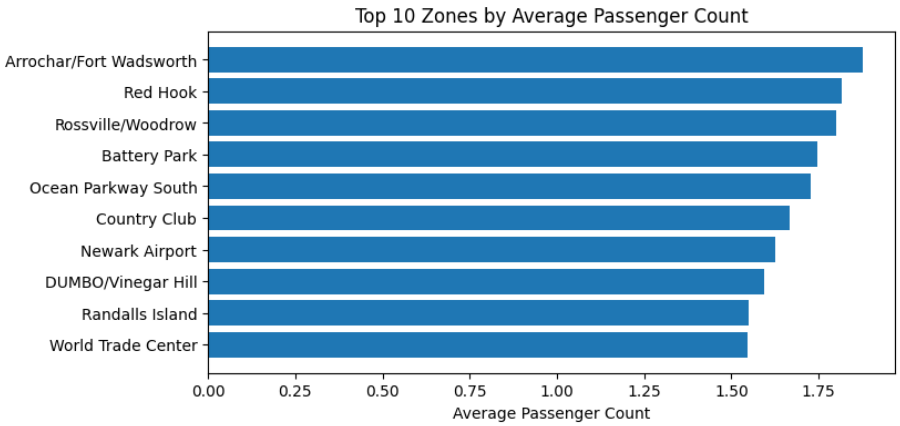
It drops to its lowest levels during early morning commute hours (5–7 AM), when solo travel dominates

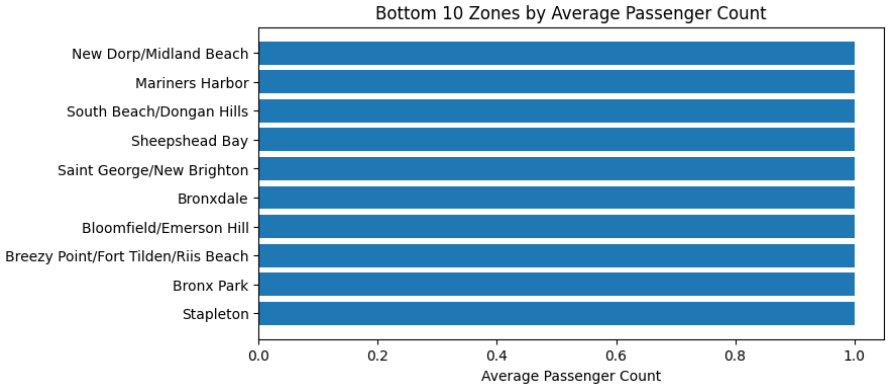


By day of week, passenger counts are lower and more stable on weekdays, consistent with routine work-related trips, while weekends (Saturday and Sunday) show higher average occupancy, indicating increased group and family travel.

Overall, the patterns suggest that trip purpose strongly influences vehicle occupancy, shifting from individual commuting on weekdays and mornings to shared, leisure-oriented travel during evenings and weekends.

1.2.13. Analyse the variation of passenger counts across zones

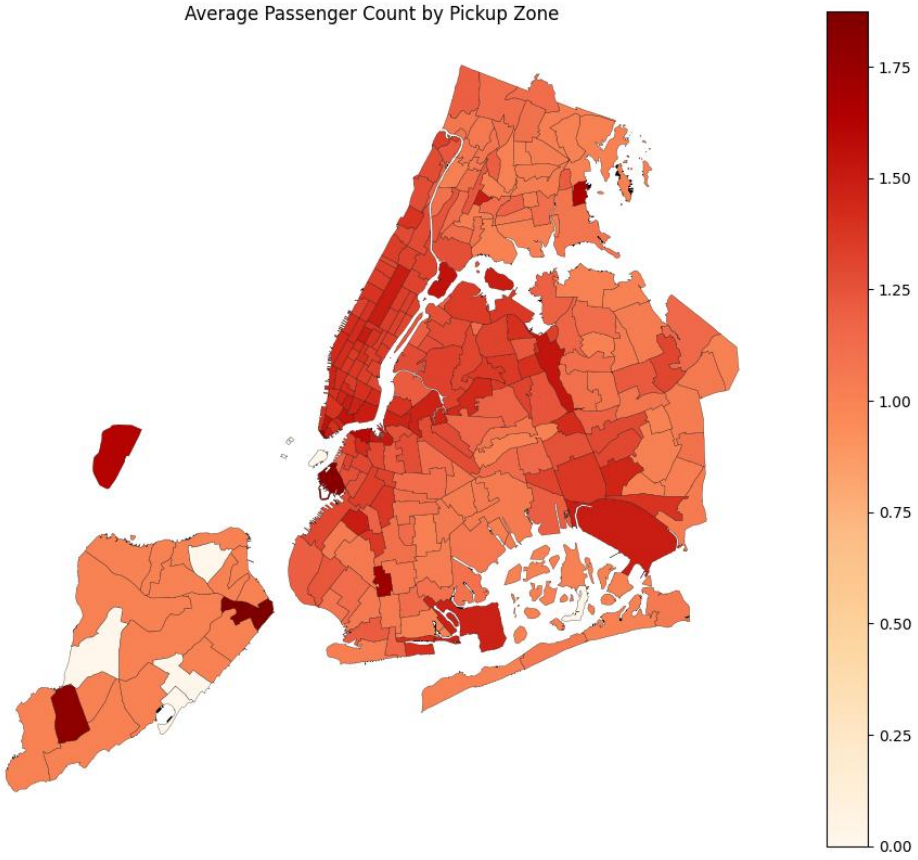




The top 10 zones by average passenger count are dominated by tourist areas, waterfront locations, airports, and major attractions (e.g., Battery Park, World Trade Center, Newark Airport), where group and family travel is more common. These zones consistently show average occupancies well above 1.5 passengers per trip.

In contrast, the bottom 10 zones exhibit average passenger counts close to 1, indicating that trips from these areas are primarily solo journeys, often associated with residential neighborhoods or routine commuting.

The clear separation between the two groups highlights how trip purpose and land-use characteristics strongly influence vehicle occupancy, with leisure- and travel-oriented zones encouraging shared rides while residential and commuter zones remain individual-centric.

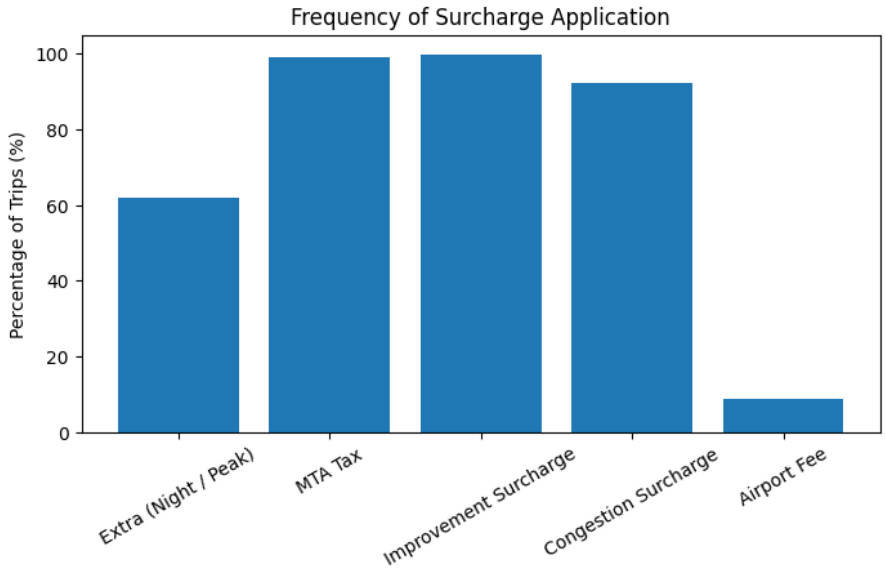


The choropleth map shows clear spatial variation in average passenger count across pickup zones, with darker zones indicating higher occupancy per trip.

Airport-adjacent zones, tourist areas, and waterfront/leisure districts stand out with higher average passenger counts, reflecting group travel, family trips, and luggage-heavy journeys.

Business districts and inner residential zones generally show lighter shades, indicating predominantly solo or two-passenger trips, consistent with commuter-driven demand.

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



The chart shows that mandatory surcharges—MTA Tax and Improvement Surcharge—are applied to almost all trips, confirming their role as standard regulatory fees.

The Congestion Surcharge is applied to a large majority of trips, reflecting frequent travel through congestion-prone central zones.

Extra charges (night/peak) affect a substantial but smaller share of trips, aligning with off-hour and peak-period travel patterns.

The Airport Fee is applied to relatively few trips, as expected since only airport-related journeys incur this charge.

2. Conclusions

2.1. Final Insights and Recommendations

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

Time- and zone-aware dispatching: Pre-position drivers in high-demand zones during peak afternoon/evening hours and weekends, and reduce supply in low-demand early-morning periods except near airports and hospitals.

Congestion-aware routing: Avoid assigning trips through historically slow routes during peak hours; dynamically reroute based on known low-speed corridors to reduce delays and fuel inefficiency.

Pickup–dropoff imbalance correction: Use pickup/dropoff ratios to proactively rebalance vehicles toward zones that consistently generate outbound demand.

Nighttime optimization: Increase driver availability in safe, high-demand night zones where surcharge frequency and tipping rates are higher, improving driver earnings and service availability.

Occupancy-based matching: Deploy larger vehicles to zones and times with higher average passenger counts (airports, tourist areas, weekends) to improve utilization.

Predictive planning: Leverage historical hourly and zone-level trends to forecast demand spikes and adjust dispatch density in advance rather than reacting to congestion.

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

Peak-hour, high-demand zones: Position more cabs in central business districts, transit hubs, and commercial areas during weekday mornings and evenings, when pickup volumes peak due to commuting activity.

Evening and weekend hotspots: Increase cab availability in entertainment, tourist, and waterfront zones during evenings and weekends, which show higher trip volumes, passenger counts, and tipping rates.

Night-time strategy: Deploy cabs selectively in safe, high-demand night zones (e.g., airports, hospitals, nightlife districts) during late-night and early-morning hours, when overall volume is lower but revenue per trip is higher due to surcharges.

Seasonal and monthly adjustments: Adjust fleet distribution based on monthly trends, scaling up availability during consistently high-demand months and reallocating cabs from low-demand periods to maintenance or driver rest.

Zone rebalancing using pickup–dropoff ratios: Continuously rebalance cabs toward zones with high pickup-to-dropoff ratios, ensuring vehicles are positioned where outbound demand is strongest and minimizing idle time.

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

Refine short-trip pricing: Since fare per mile is highest for trips under 2 miles and differs across vendors, consider slightly lowering base fares or offering micro-discounts for short trips in high-competition zones to attract demand while preserving margins through higher trip volumes.

Time-of-day pricing optimization: Use historical demand and congestion data to fine-tune peak and night surcharges, ensuring they align with actual congestion and willingness-to-pay rather than being uniformly applied across all zones.

Distance-tier smoothing: As fare efficiency converges for longer trips, maintain competitive, stable per-mile rates for trips over 5 miles to encourage high-value, long-distance journeys and airport runs.

Zone-sensitive pricing: Apply location-based pricing adjustments in zones with consistently high demand, high pickup-dropoff imbalance, or frequent congestion, while offering competitive rates in low-demand zones to stimulate usage.

Vendor benchmarking: Continuously compare vendor-wise fare per mile across distance tiers to avoid overpricing in short trips, where vendor differences are most noticeable and customer sensitivity is highest.

Payment-linked incentives: Encourage card-based payments through small incentives, as these trips are associated with higher recorded tips and smoother fare realization, improving total effective revenue without raising base fares.