NYC Taxi Trip Data – EDA Cleaning & Business Insights Report

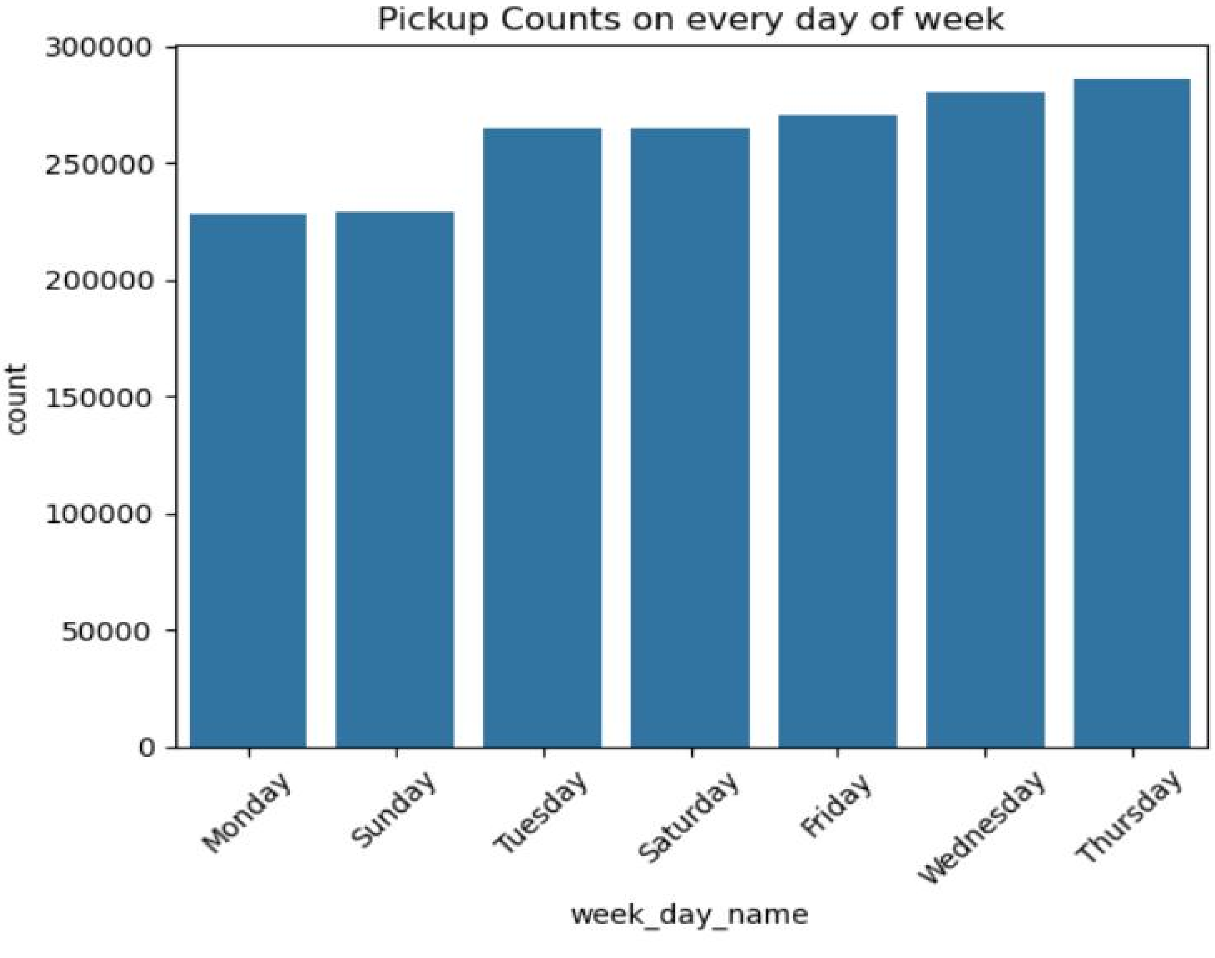
By Nishant A Bilagi

1. Data Cleaning & Handling Process

Initial data inspection revealed missing, invalid, and inconsistent records.
Columns such as fare_amount, trip_distance, and tips_amount were validated.
Trips with zero or negative fare and distance values were removed.
This ensured that only realistic taxi trips were retained.
Outliers in fare and trip distance were identified using statistical thresholds.
Extreme values were removed to reduce skewness.
A cleaned dataset was used for all further analysis.
This improved reliability of trends and insights.

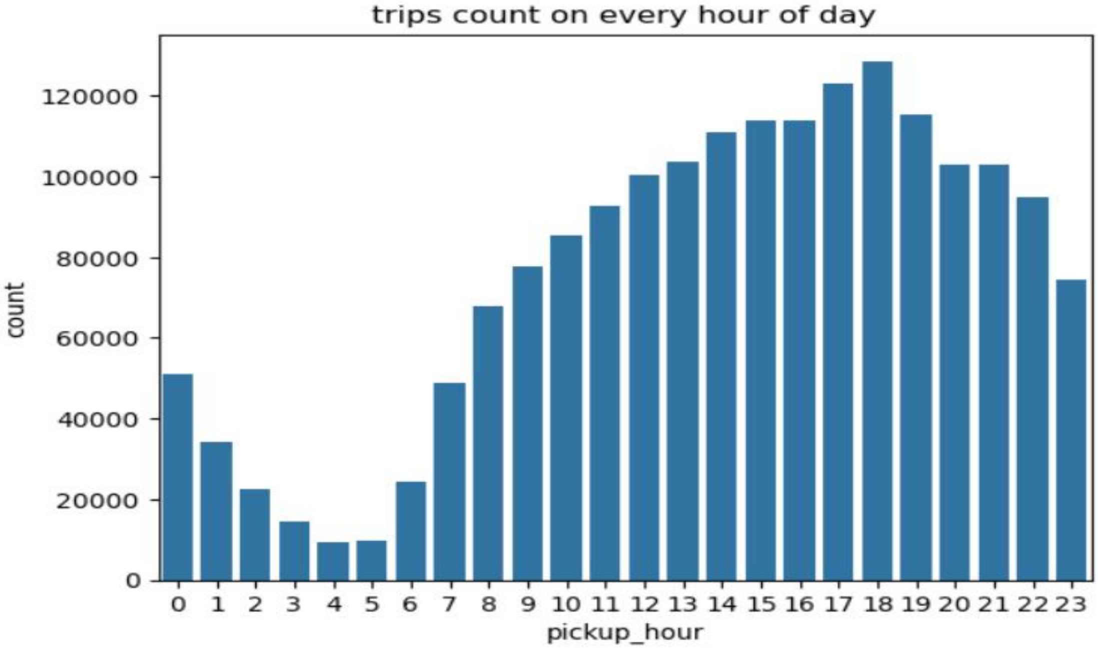
2. Business Insights from EDA Visualizations

Pickup counts on every day of week



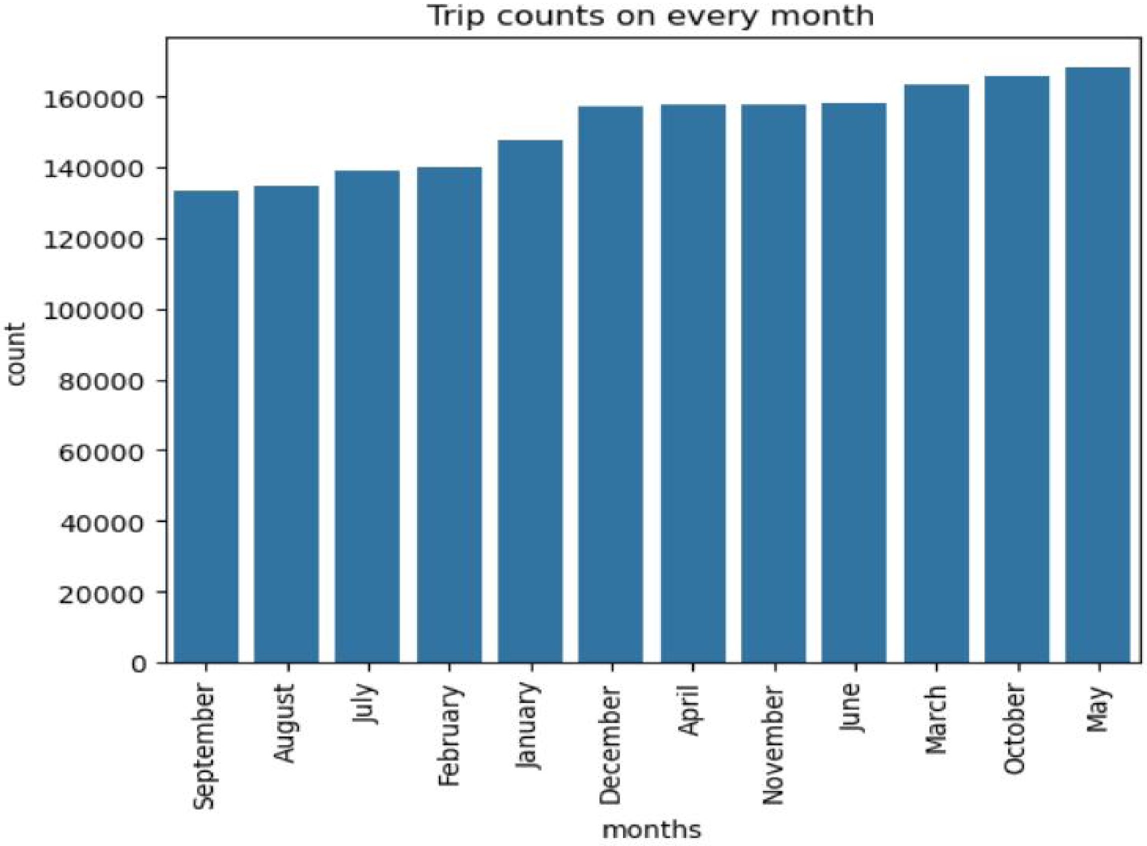
Maximum pickup trips happens on Thursday, Wednesday, and lest pickup on Monday and Sunday

Trips count on every hour of day



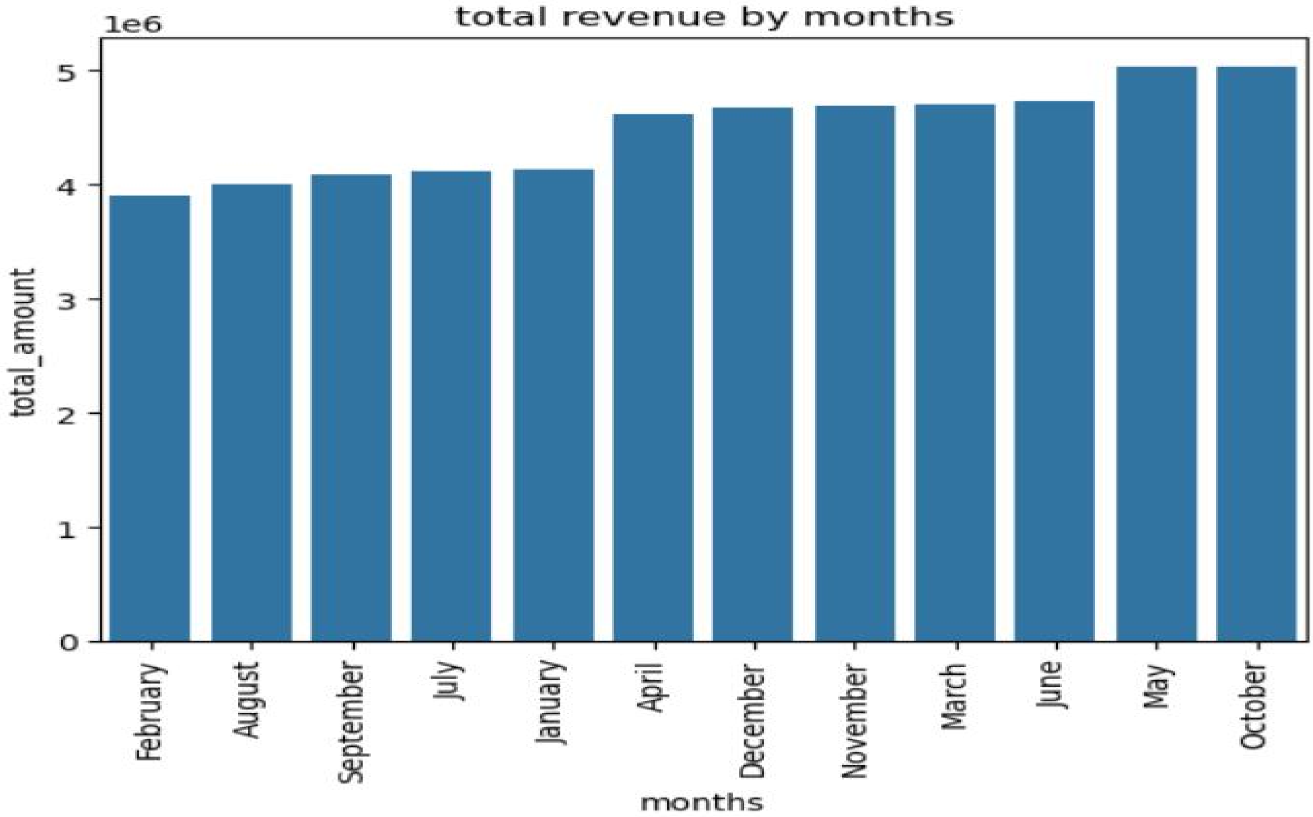
On 17th and 18th hour of the day there maximum trip counts and on 4th and 5th hour trip counts are the least.

Trip counts on every month of the year



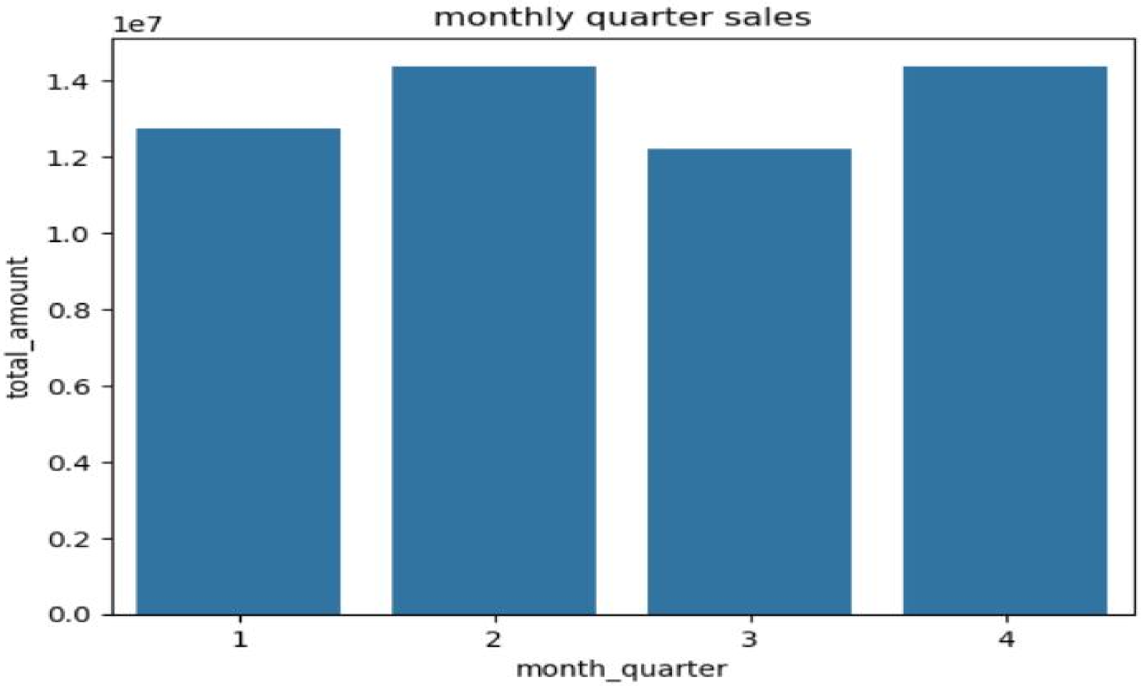
May and October there is maximum trips and on September and august are having least no of trips

Total amount revenue by months



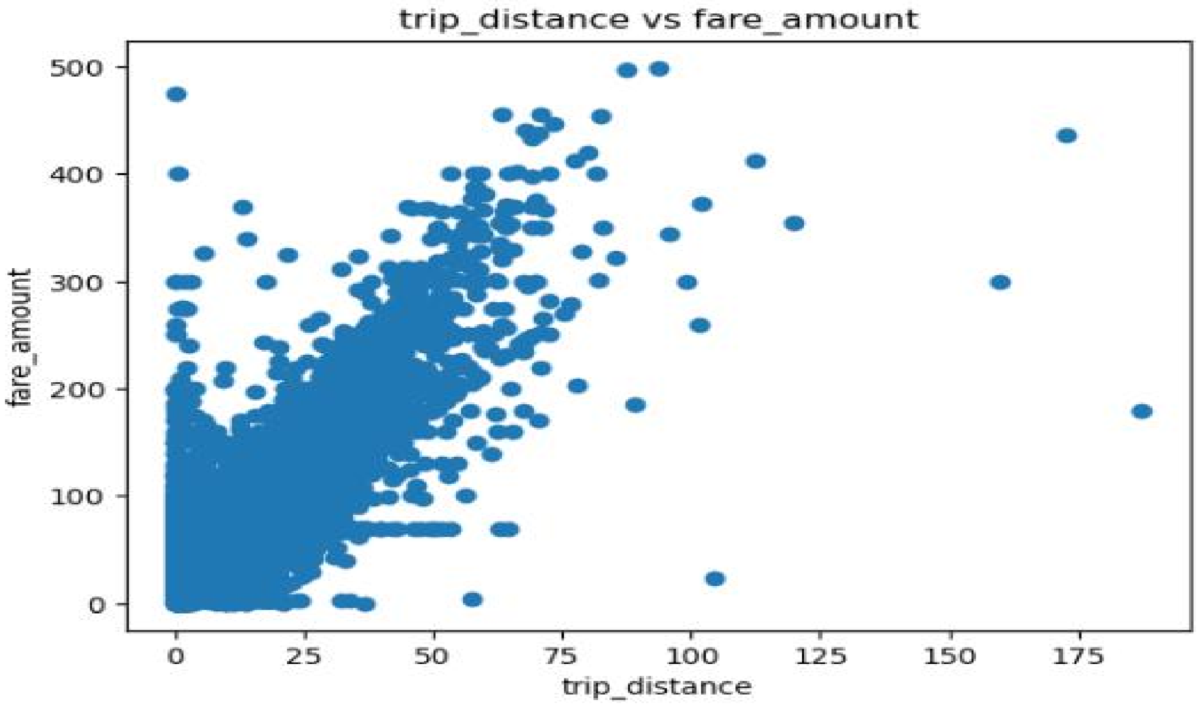
May and October months with maximum revenue and February and August are one with least amount of revenue.

Monthly quarter sales



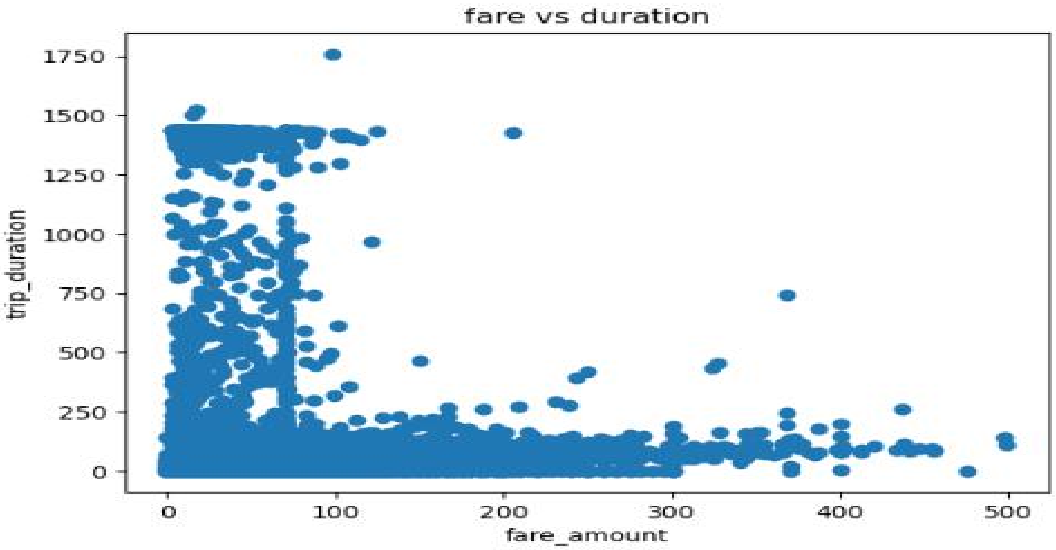
1st and 4th quarter is having maximum sales and 2nd and 3rd are have lowest sales

Trip Distance vs Fare Amount



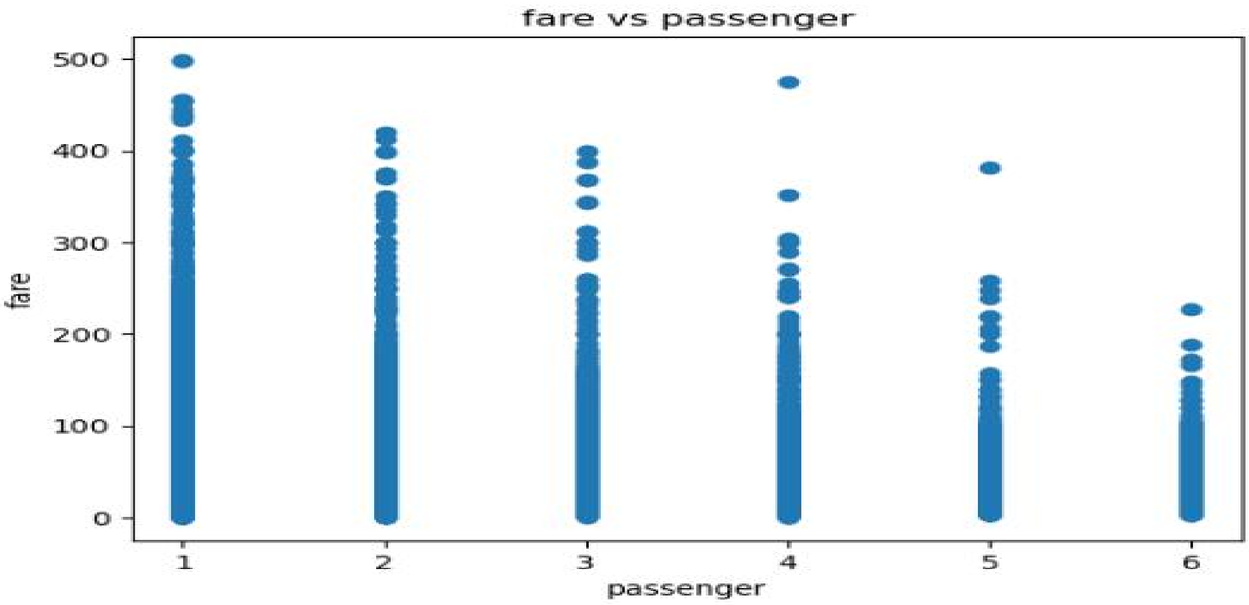
This scatter plot shows a strong positive relationship between trip distance and fare.
Most trips are short-distance with lower fares.
Long-distance trips are fewer but generate higher revenue.
Distance is the primary driver of fare and revenue.

Fare vs Trip Duration



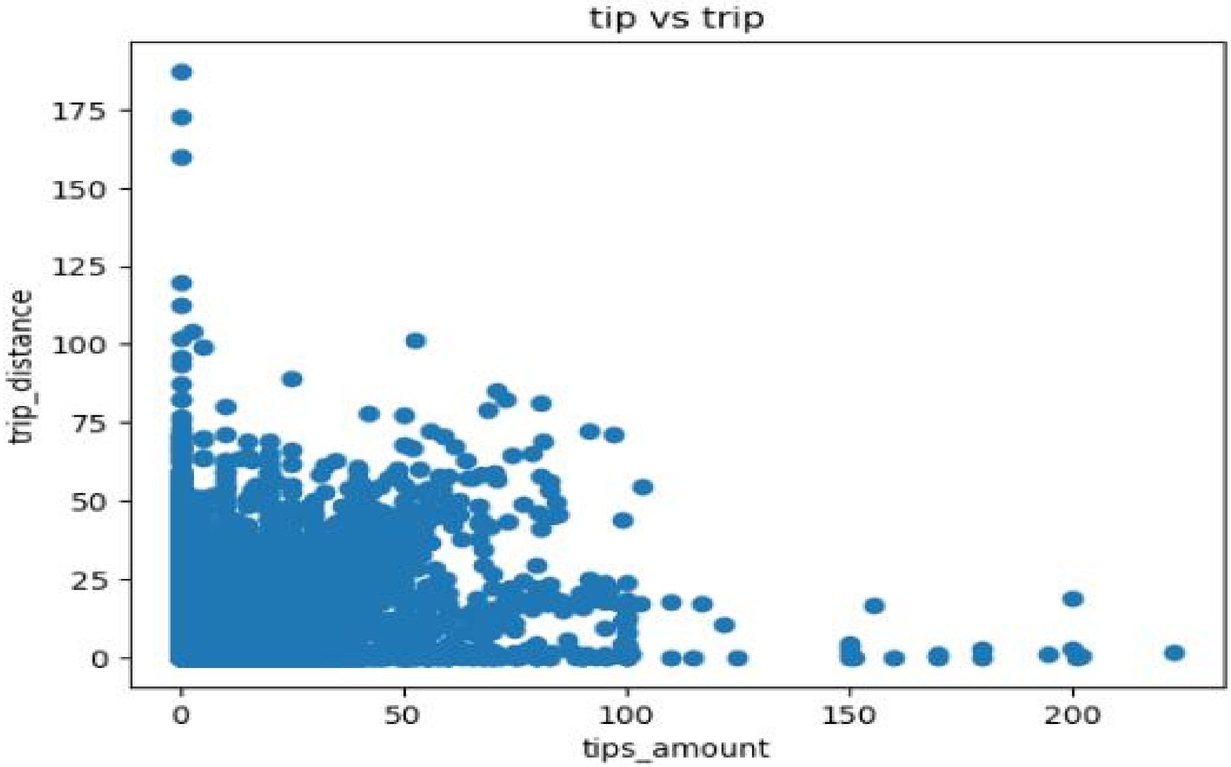
This plot compares fare amount with trip duration.
Longer duration does not always result in higher fare.
Traffic delays increase duration without proportional revenue.
Distance matters more than time for pricing.

Fare vs Passenger Count



Fare values remain similar across different passenger counts.
No clear upward trend is visible with more passengers.
Passenger count has limited impact on fare.
Pricing models should prioritize distance.

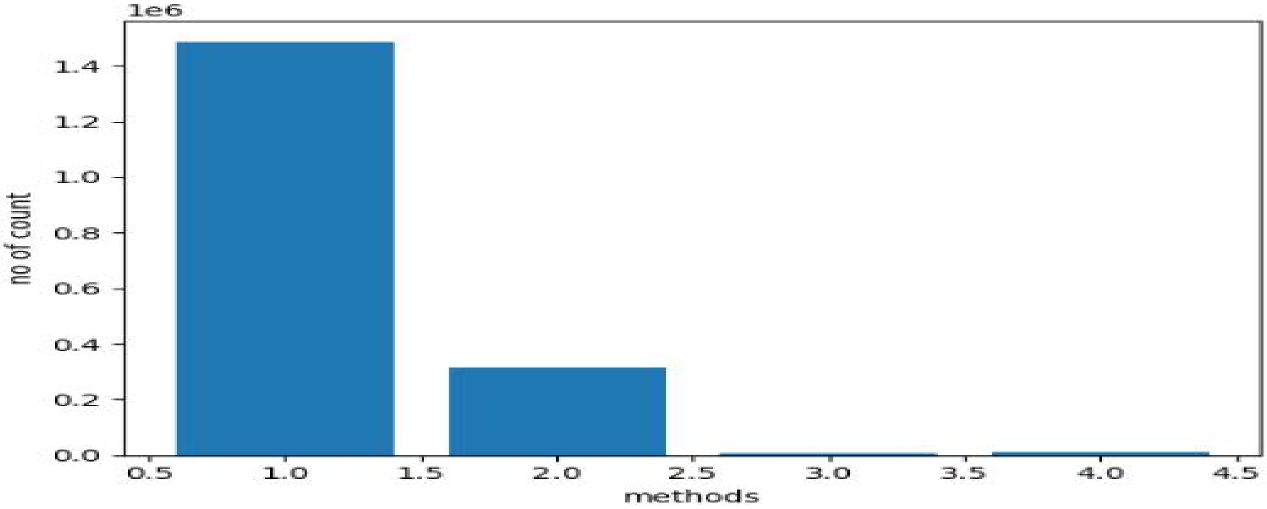
Tip Amount vs Trip Distance



Tips remain low for most trips regardless of distance.
High tips are rare and scattered.
Trip distance does not strongly influence tips.

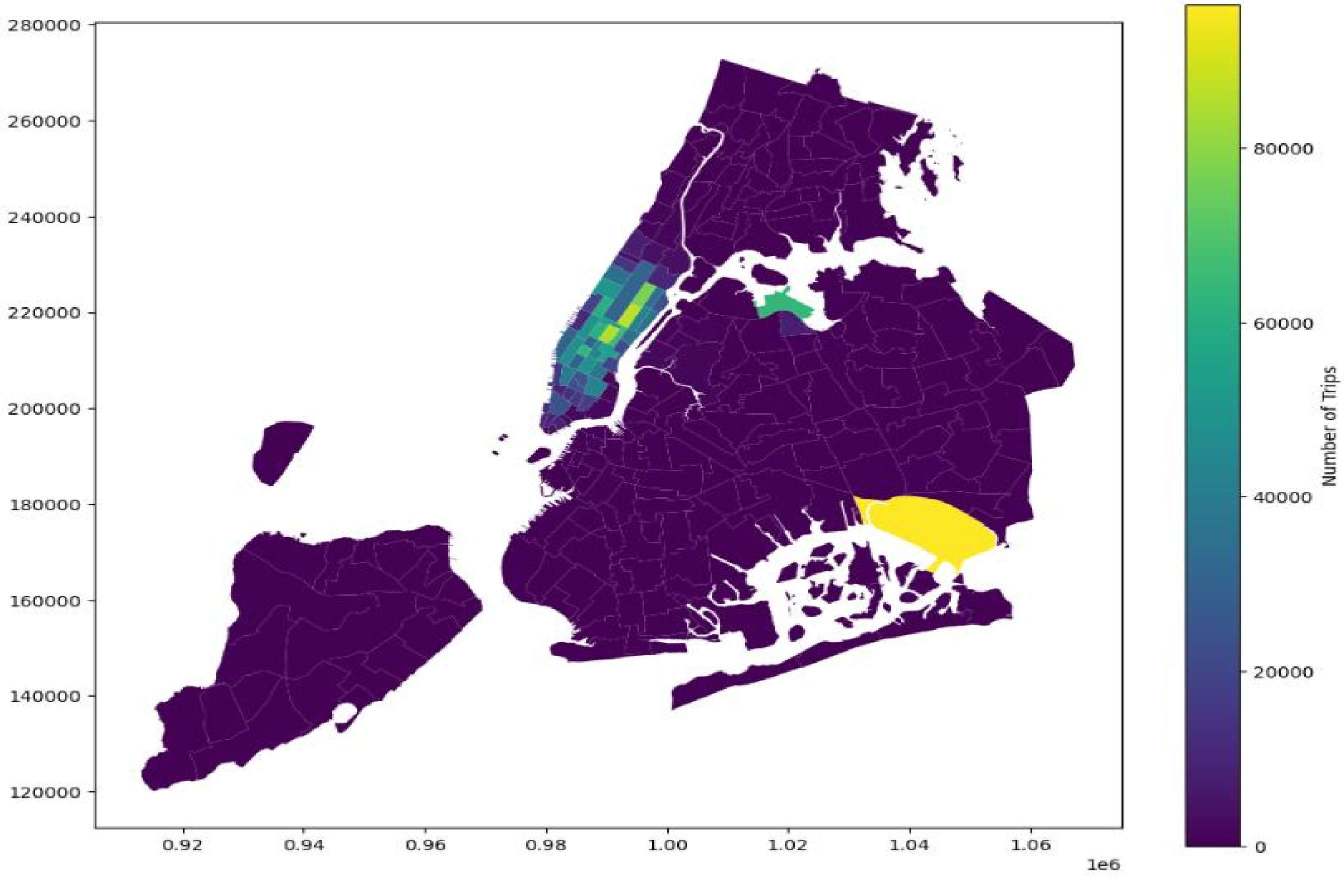
Service quality likely plays a larger role.

Payment Method Distribution



One payment method dominates total trips.
Other payment modes are used far less frequently.
Digital payments drive most transactions.
System reliability for dominant methods is critical.
(1= Credit card, 2= Cash, 3= No charge, 4= Dispute)

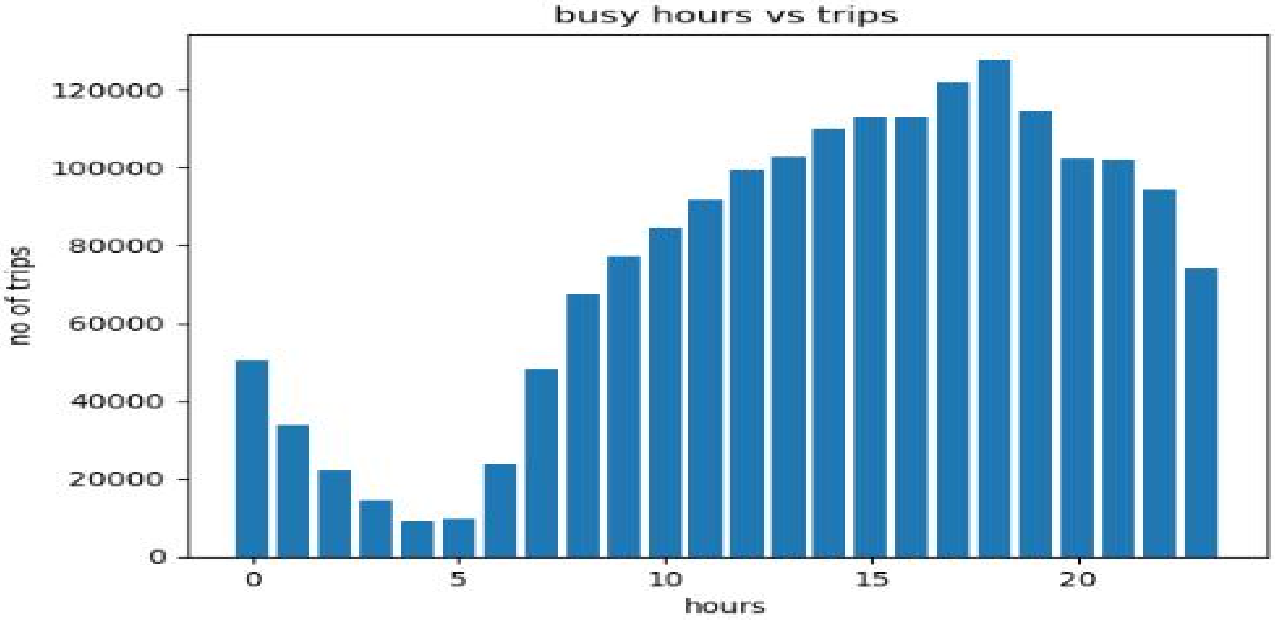
Geographical Demand Heatmap



Trip demand is concentrated in central NYC regions.
Outer regions show significantly lower demand.

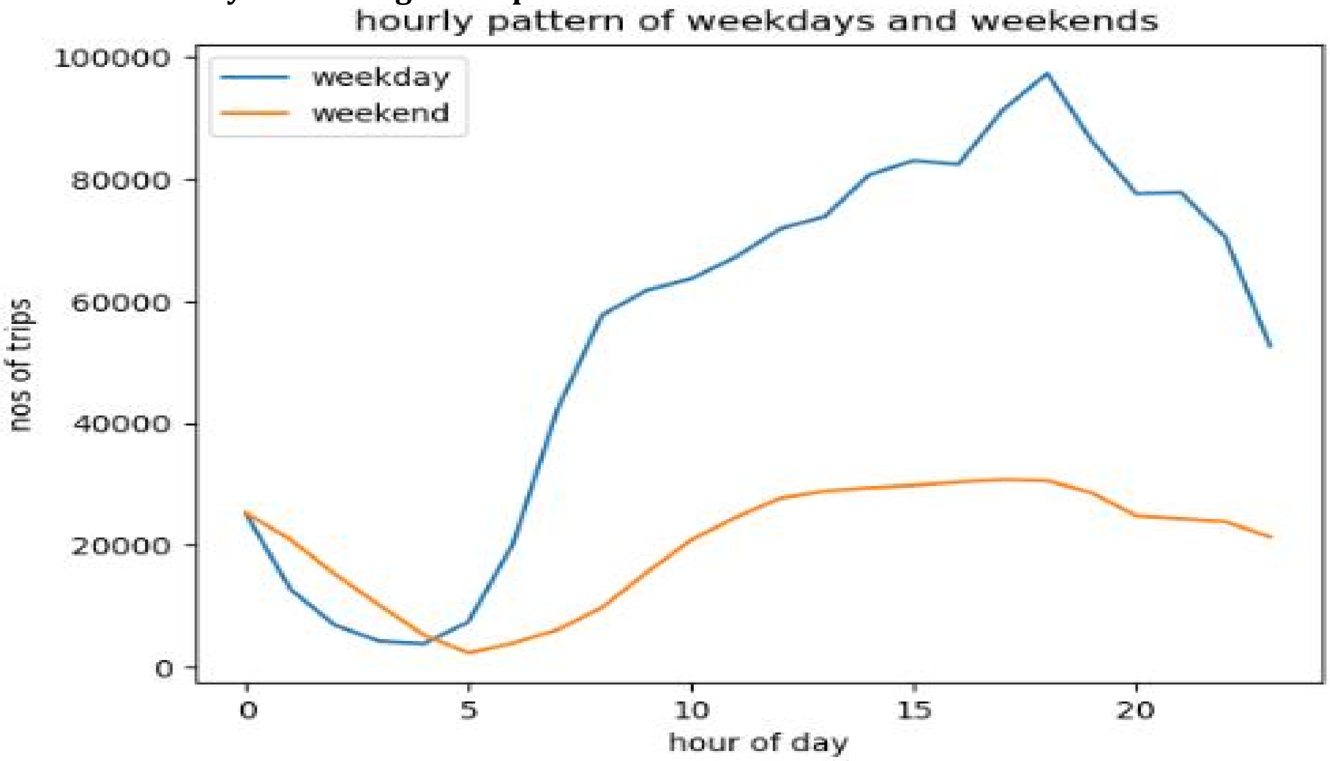
Demand is geographically clustered.
Fleet allocation should prioritize high-density zones.

Busy Hours vs Number of Trips––––



Trip volume peaks during daytime and evening hours.
Early morning hours show low demand.
Clear peak and off-peak periods exist.

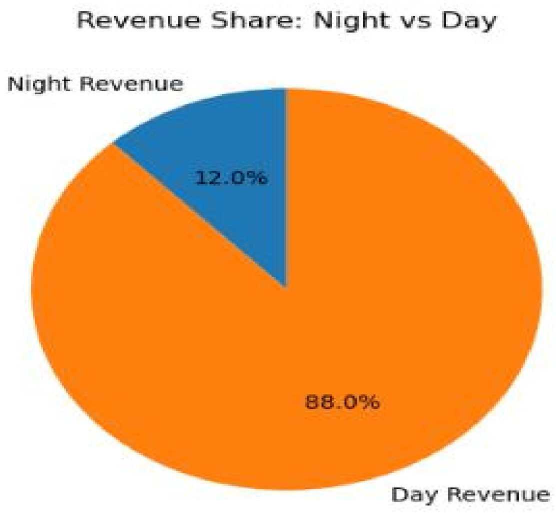
Fleet availability should align with peak hours.



Weekday vs Weekend Hourly Patterns
Weekdays show strong commute-hour peaks.
Weekend demand is flatter and more evenly spread.

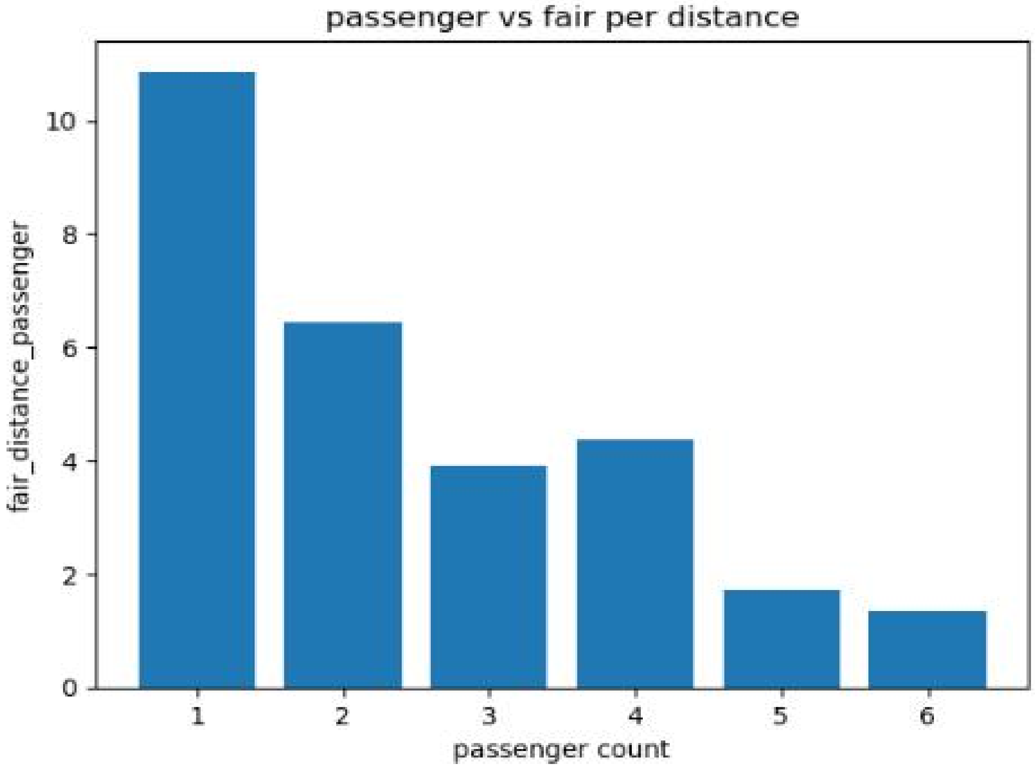
Travel purpose differs by day type.
Staffing and fleet strategy should adjust accordingly.

Revenue Share: Day vs Night



Daytime trips contribute most of the revenue.
Night trips form a smaller revenue share.
Revenue concentration is time-dependent.
Day operations are more profitable.

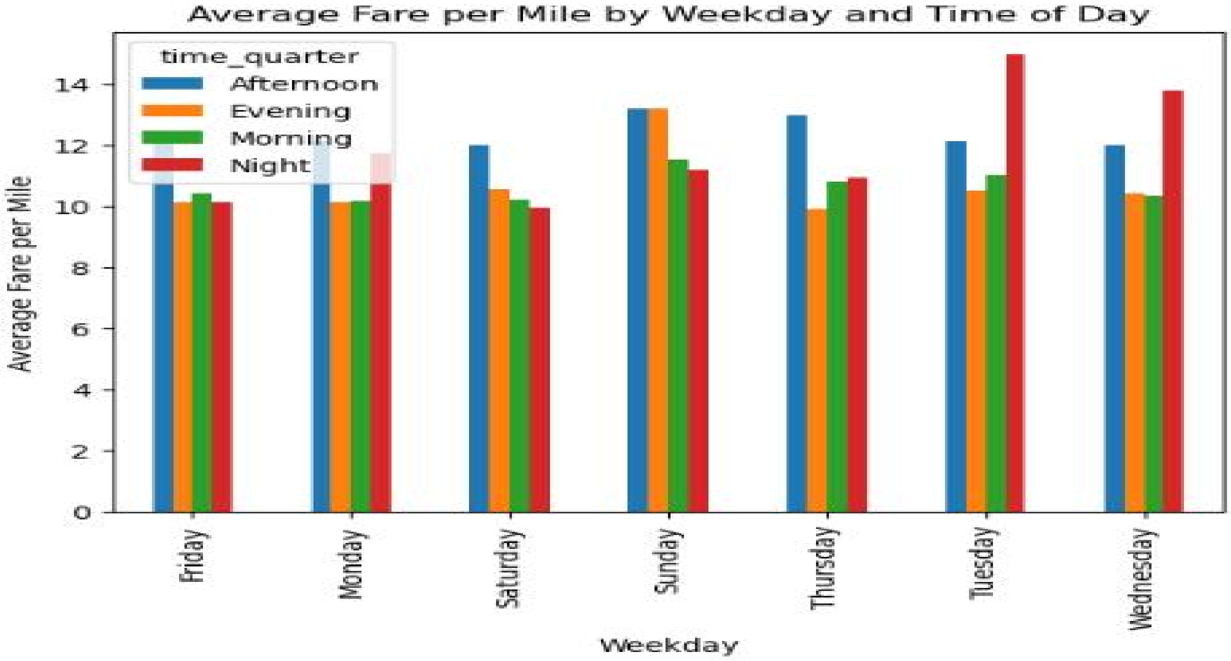
Passenger vs Fare per Distance



Fare per distance decreases as passenger count increases.
Single-passenger trips generate higher revenue per mile.

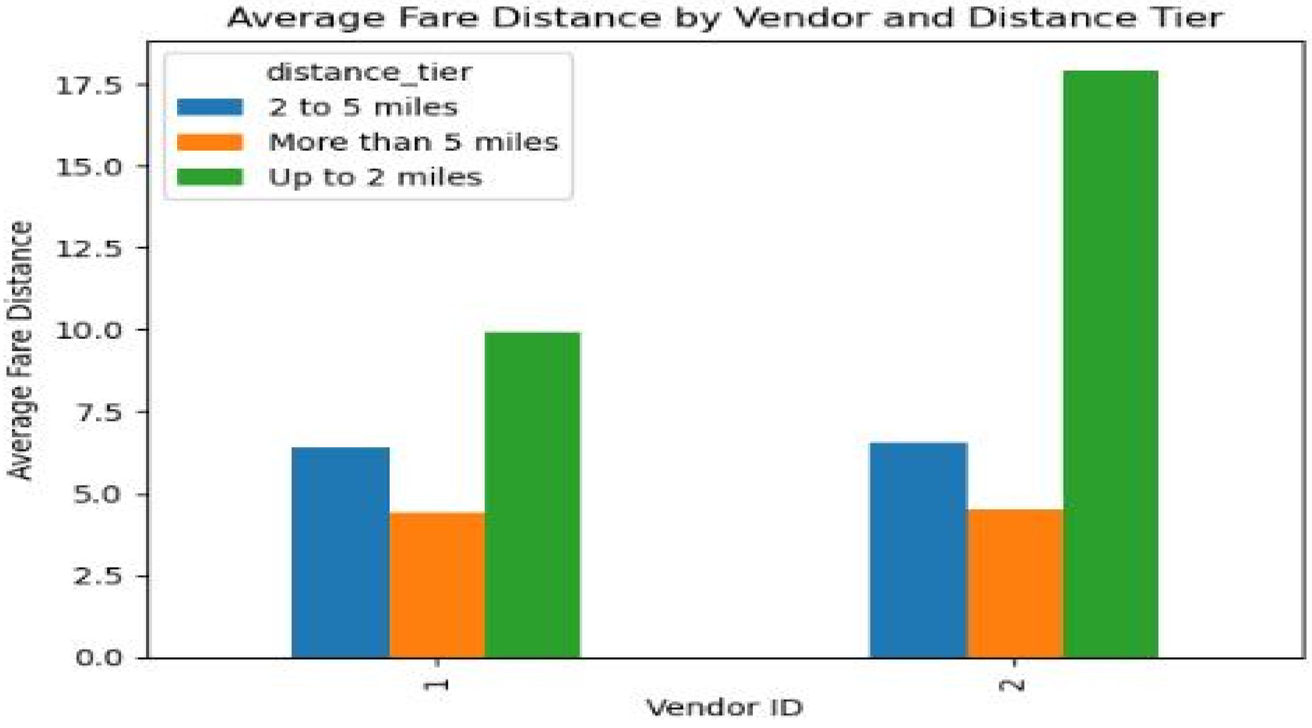
Shared rides reduce revenue efficiency.
Single-passenger demand is more profitable.

Average Fare Distance by Vendor and Distance Tier



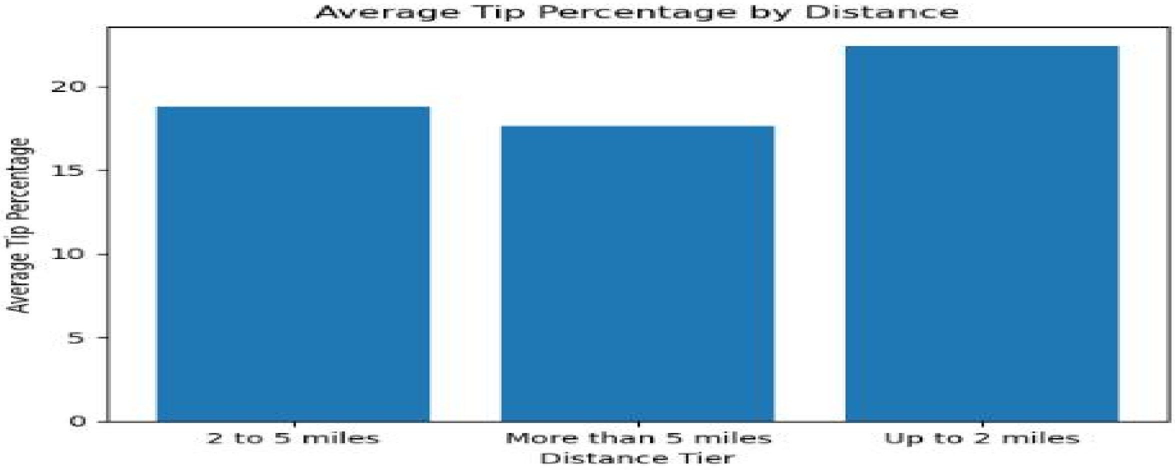
Vendor performance varies across distance tiers.
Short-distance trips dominate across vendors.
Vendor 2 shows higher average distance in short trips.
Vendor-level optimization opportunities exist.

Average Fare Distance by Vendor and Distance Tier



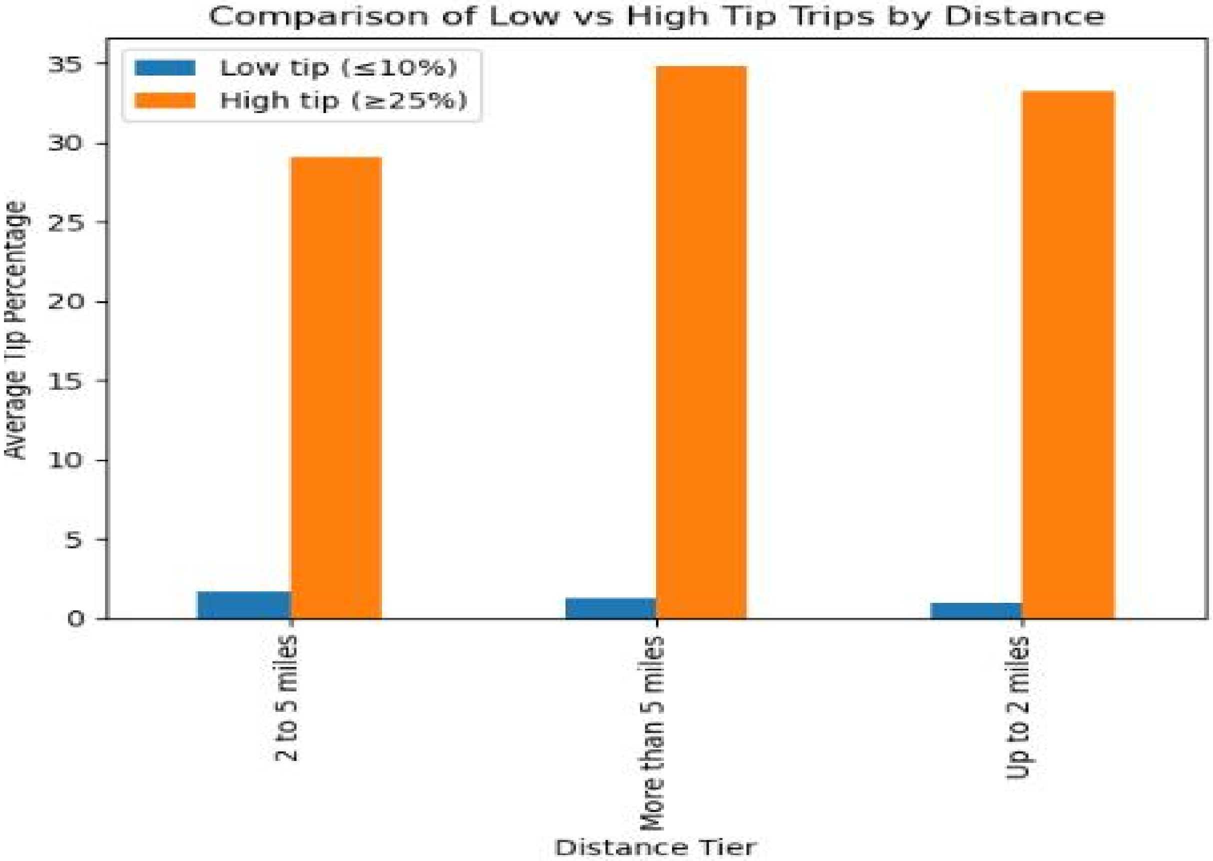
The chart compares average trip distance across vendors and distance tiers.
Trips are grouped into up to 2 miles, 2–5 miles, and more than 5 miles.
Both vendors show higher averages for short-distance trips.

Average Tip Percentage by Distance Tier



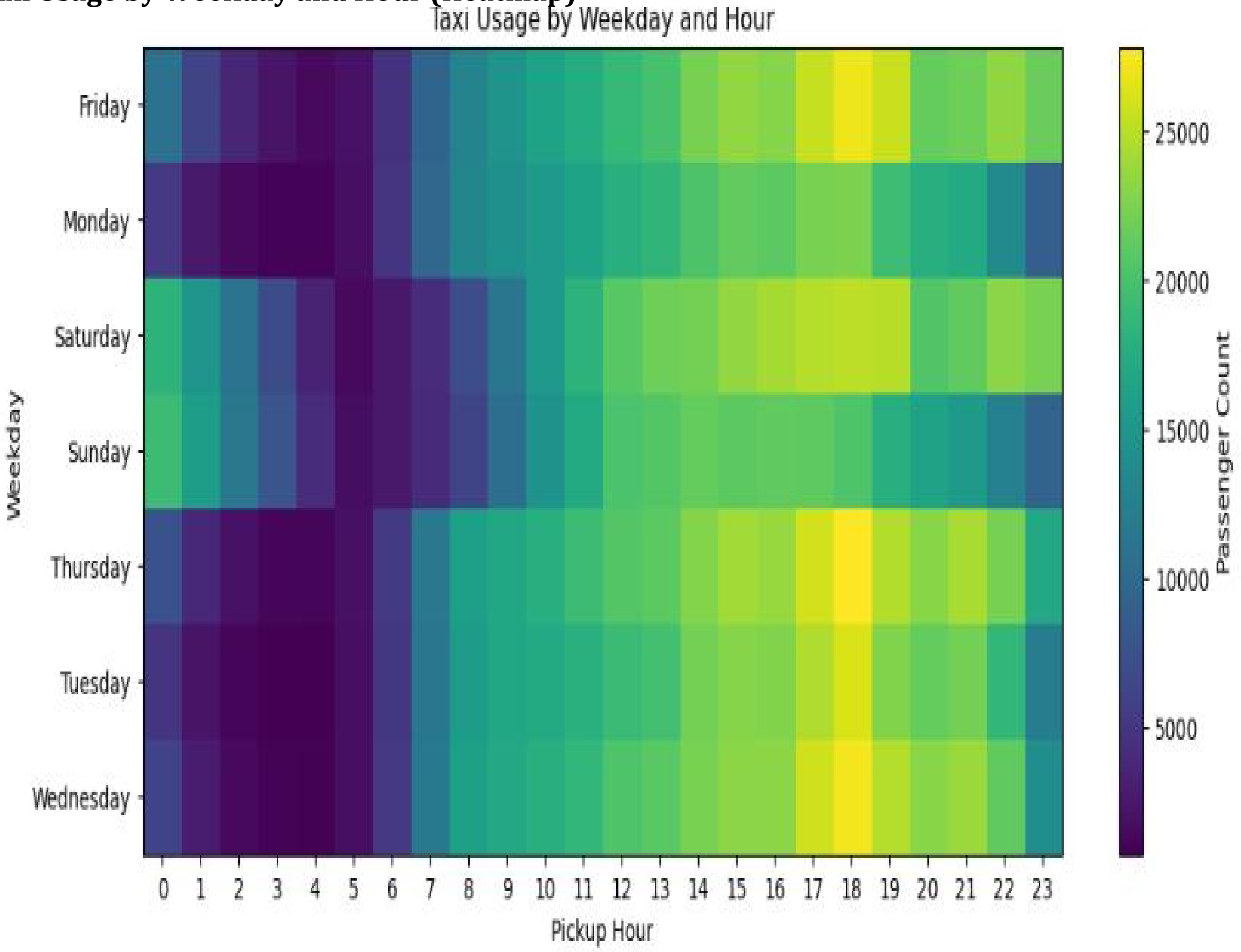
Short-distance trips show higher average tip percentage.
Longer trips do not guarantee higher tips.
Tipping behavior is not distance-driven.
Customer behavior impacts tips more than trip length.

Low vs High Tip Trips by Distance Tier



High-tip trips are more common in short and medium distances.
Low-tip trips remain consistently low across tiers.
Tipping behavior is polarized.
Service quality is a key differentiator.

Taxi Usage by Weekday and Hour (Heatmap)



Clear hourly demand patterns are visible across weekdays.
Morning and evening peaks dominate weekdays.
Late-night hours show lower demand.
Time-based fleet optimization is essential.

Conclusion

4.1 Final Insights and Recommendations

The analysis of NYC taxi trip data revealed clear demand patterns across time, location, and trip characteristics.
Data cleaning was critical to remove invalid fares, distances, and durations, ensuring reliable insights.
Trip distance emerged as the strongest driver of fare revenue, while passenger count and trip duration had limited impact.
Demand varied significantly by time of day, day of week, and pickup location.
High trip volumes were observed during daytime and evening hours, especially on weekdays.
Certain pickup zones consistently showed higher demand, indicating geographic clustering of trips.

Overall, combining time-based and location-based insights enables better demand forecasting.
These insights can be used to optimize routing, dispatching, fleet positioning, and pricing strategies.

4.1.1 Recommendations to Optimize Routing and Dispatching *(5 marks)*

The analysis of top pickup routes by time of day shows that specific locations are used more frequently during certain periods such as morning, afternoon, evening, and night.
During afternoon and evening hours, a small set of pickup locations accounts for a large share of trips.
This indicates concentrated demand in business and commercial zones during peak hours.
Recommendation:
Taxis should be pre-positioned near high-demand pickup locations during corresponding time periods.
Dynamic routing strategies should prioritize these zones to reduce passenger wait times.
By aligning dispatch decisions with time-specific demand patterns, idle time can be reduced.
This improves fleet utilization and operational efficiency.

4.1.2 Strategic Positioning of Cabs Across Zones *(5 marks)*

The analysis of top pickup locations by month, weekday, and hour highlights strong spatio-temporal demand patterns.
Certain zones consistently appear as high-demand areas during specific hours and days.
For example, late-night and early-morning demand is concentrated in fewer zones, while daytime demand is more widespread.
Weekend patterns differ from weekdays, showing more leisure-driven travel.
Recommendation:
Cabs should be strategically positioned in high-demand zones based on historical patterns of hour, day, and month.
This reduces idle driving and improves response time.
Such demand-aware positioning minimizes passenger wait times and maximizes trip opportunities.
It also improves overall fleet utilization and driver earnings.

4.1.3 Data-Driven Pricing Strategy Adjustments *(5 marks)*

Zone-level analysis shows that some pickup locations have both high trip volumes and higher average fares.
These zones indicate strong demand and higher willingness to pay.
Conversely, zones with low trip volume and lower average fares show price sensitivity.
Uniform pricing across all zones may not maximize revenue.
Recommendation:
Zones with consistently high demand and higher average fares can support slight fare premiums during peak hours.
Low-demand zones should use competitive pricing to attract more riders.
This zone-based pricing strategy balances revenue maximization with customer retention.
It enables more efficient demand management while remaining competitive.