EDA Case Study

Mitnala Karthik

Problem statement

Objective

Perform EDA on the “Yellow Taxi Rides in New York City” to understand the patterns leading to a consumer default

Problem Statement

Analyse patterns in the data that can inform strategic decisions to improve service efficiency, maximise revenue, and enhance passenger experience.

Overall approach for data analysis

- Data Understanding, Cleansing & Manipulation
- Dealing With Columns Having Missing Data, Outlier Analysis & Treatment
- Univariate Analysis: Understanding the various columns & drawing insights
- Bivariate Analysis: Understanding the relationship between multiple variables to see if there is a pattern or a trend - Categorical Variables & Numerical Variables
- Merging With Previous Data – Final Analysis
- Recommendations

Note: Only Few Snippets have been shown for each category; to save on no of slides. Full analysis available in jupyter notebook

Data UNderstanding cleaning and manipulation

- Understanding the various columns & drawing insights

- Outlier Analysis

Dealing With Columns Having Negative Values

Application Data

• Imputed the negative values with the mode of the column as these columns have finite set of values (3 or 4 max) that are possible - it made sense to replace with mode

- Extra
- Mta_tax
- Improvement_surcharge

Removed the values which were negative in the below rows

- Total_amount - removed the rows as they were very less (11) in number compared to total, since this is a calculated value it is better to remove than impute

Other two rows were taken care automatically

- Airport_fee
- Congestion_charge

  Both got handled automatically

COLUMNS HAVING MISSING VALUES

```
missing_data[missing_data['% missing']>0]
```

	column_name	% missing
3	passenger_count	3.326478
5	RatecodeID	3.326478
6	store_and_fwd_flag	3.326478
17	congestion_surcharge	3.326478
18	Airport_fee	3.326478

DEALING WITH MISSING/INCORRECT VALUES

All of the columns have skewness towards one value and hence stuck to one methodology and hence i have used the mode as the mode for imputation.

DEALING WITH OUTLIERS

- Vendor ID can be 1 and 2 according to data dictionary,

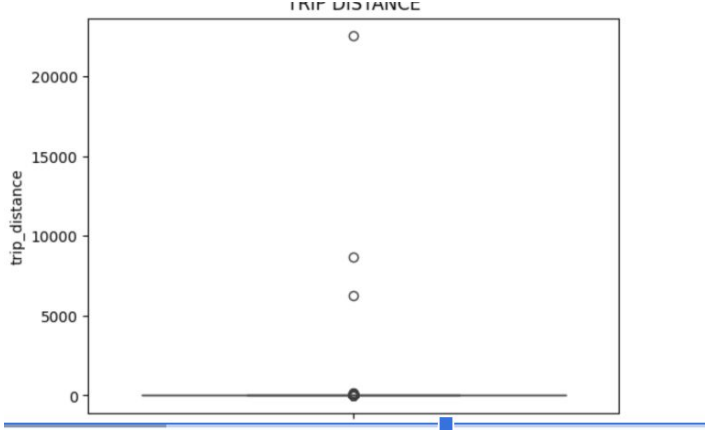
Removed the values with vendor ID 6

- Passenger count greater than 6 are less in number removed them
- Rate code cannot have 99 as value as it has fixed values as given in dictionary removing it (This might be case that it is missing as values are less than 0.005% removed them)
- Removed payment_type = 0 as it is invalid

Dealing with outliers

- Trip distance has clear outliers that are way above all data

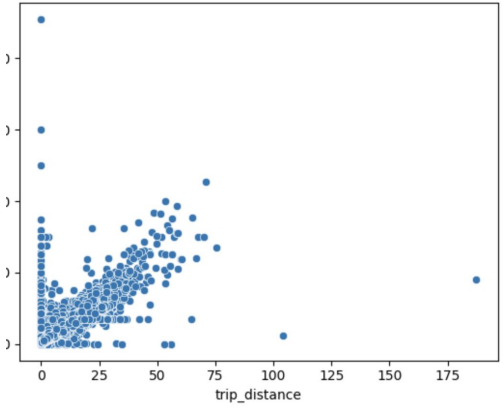
  Three such values are there and can be deleted



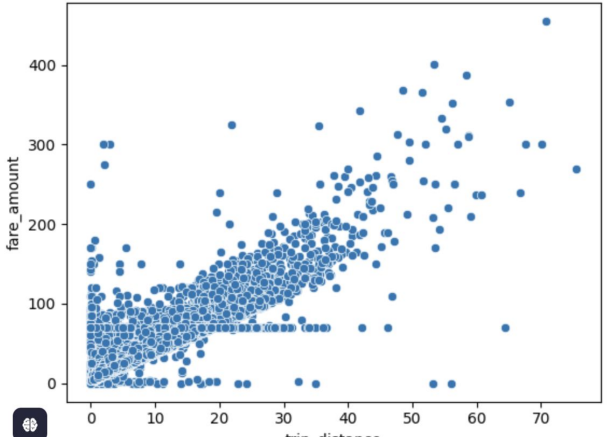
Dealing with outliers

- Plotted trip distance vs trip fare and saw that 0 distance trip have large fares.
- Removed 0 fare , 0 distance when in different zones
- Removed 0 distance all together - as it doesn't make sense
- Removed almost 0 distance and high fare > 250 as well

Initial



Final



Variable categories

```
variable_categories = {
    'Numerical': [
        'trip_distance',      # Continuous distance in miles
        'trip_duration',      # Continuous time measurement
    ],
        'Ordinal/Discrete': [
        'passenger_count',    # Count of passengers (discrete but ordered)
        'pickup_hour',        # Hour of day (0-23, discrete but cyclical)
         ]
'Categorical': [
```

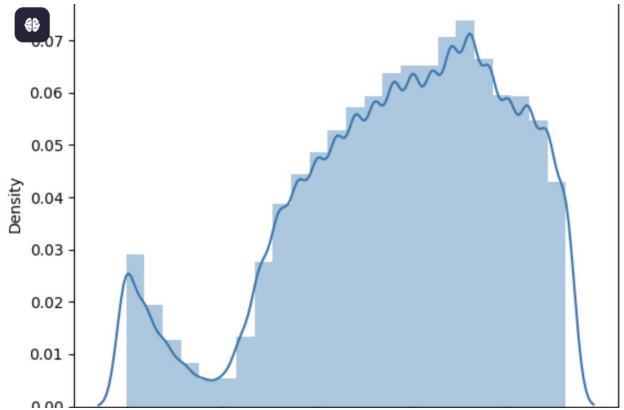
- 'VendorID', # Discrete identifier (1 or 2)

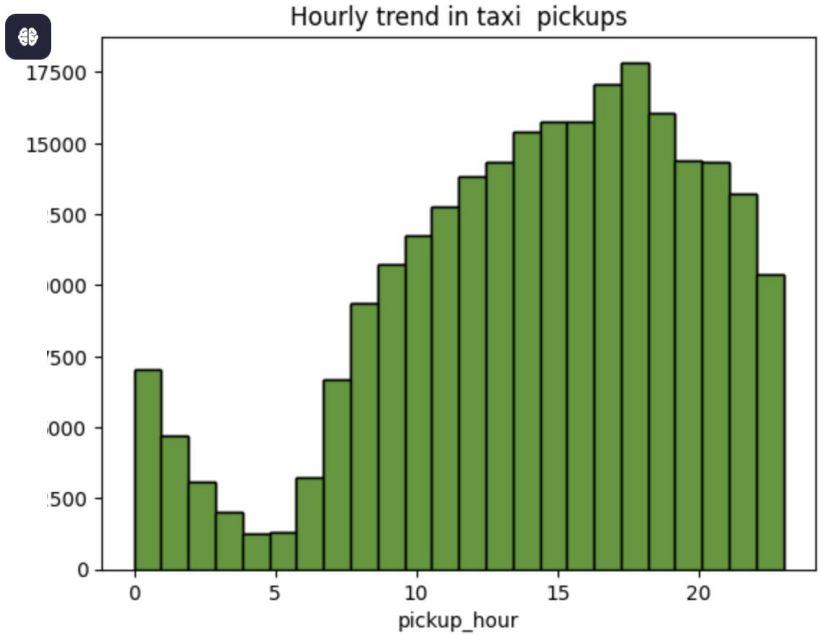
```
        'RatecodeID',        # Discrete code for rate type
        'PULocationID',      # Discrete location identifier
        'DOLocationID',      # Discrete location identifier
        'payment_type',      # Discrete payment method code
    ]
```

Note: Only Few Snippets have been shown for each category; to save on no of slides. Full analysis available in jupyter notebook

Hourly trend in taxi pickups

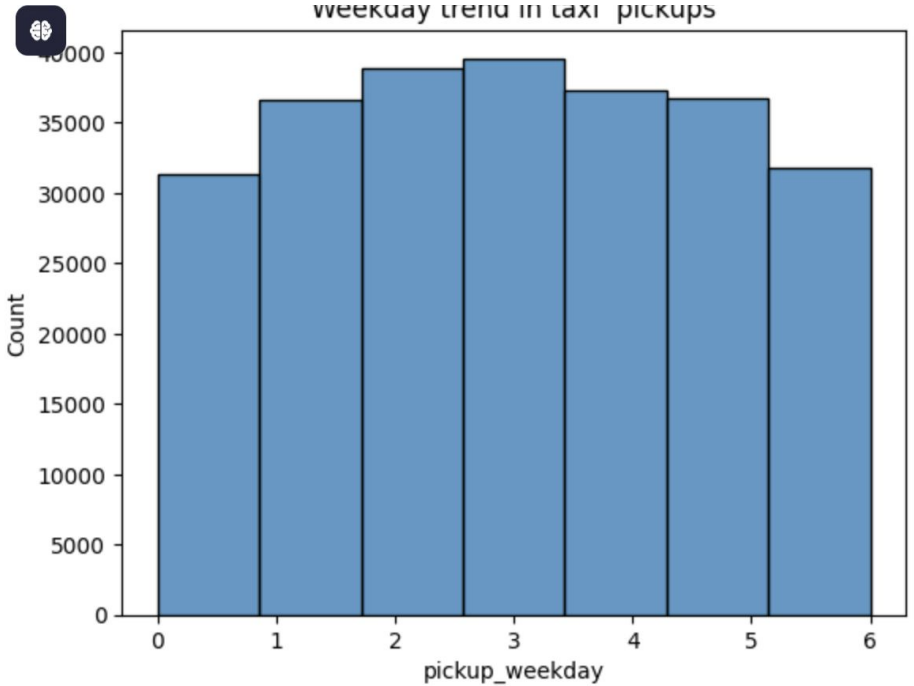
- Pickups high from afternoon to night





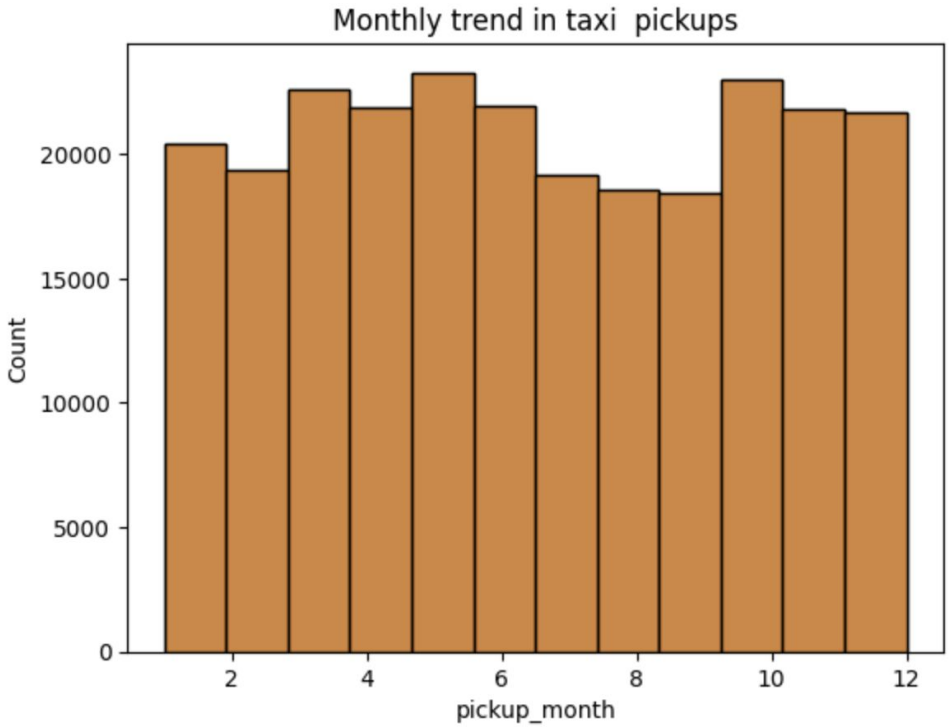
Day wise trend in taxi pickup

- Similar by the days



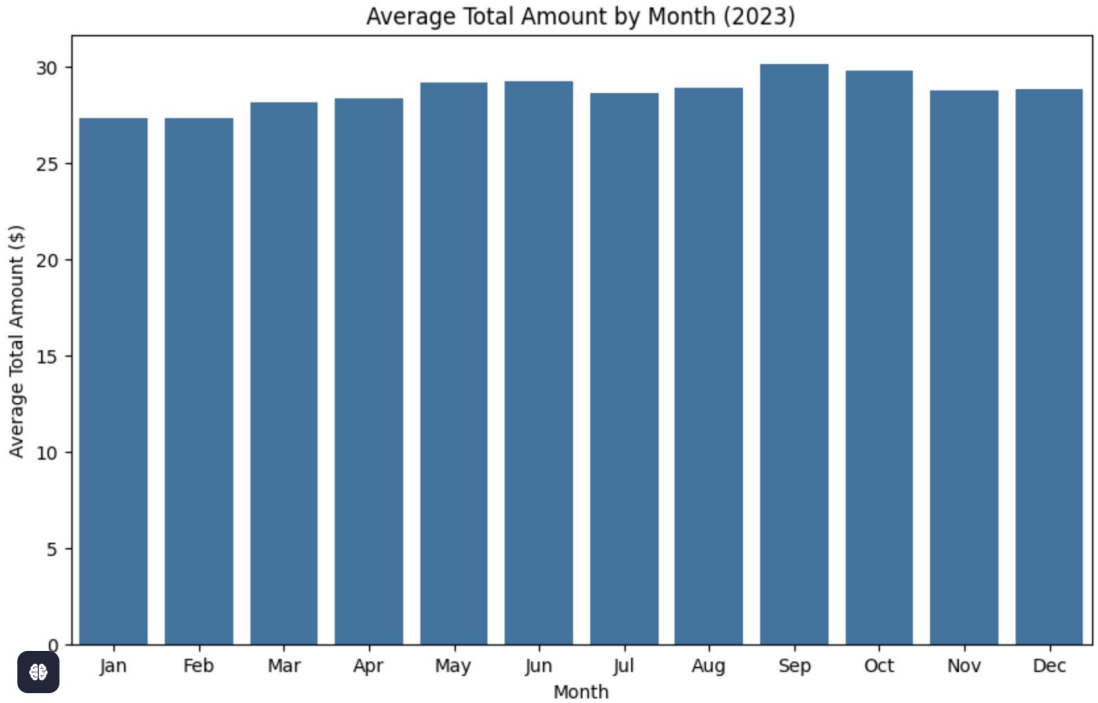
MONTHLY TREND IN TAXI PICKUPS

- Largely similar
- Lower demand in (Jul, Aug ,sep)



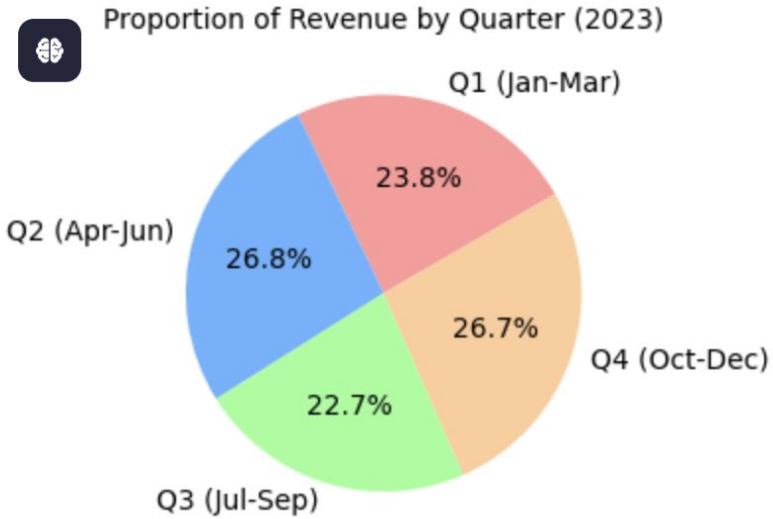
Month wise Revenue

- Similar Avg's across months



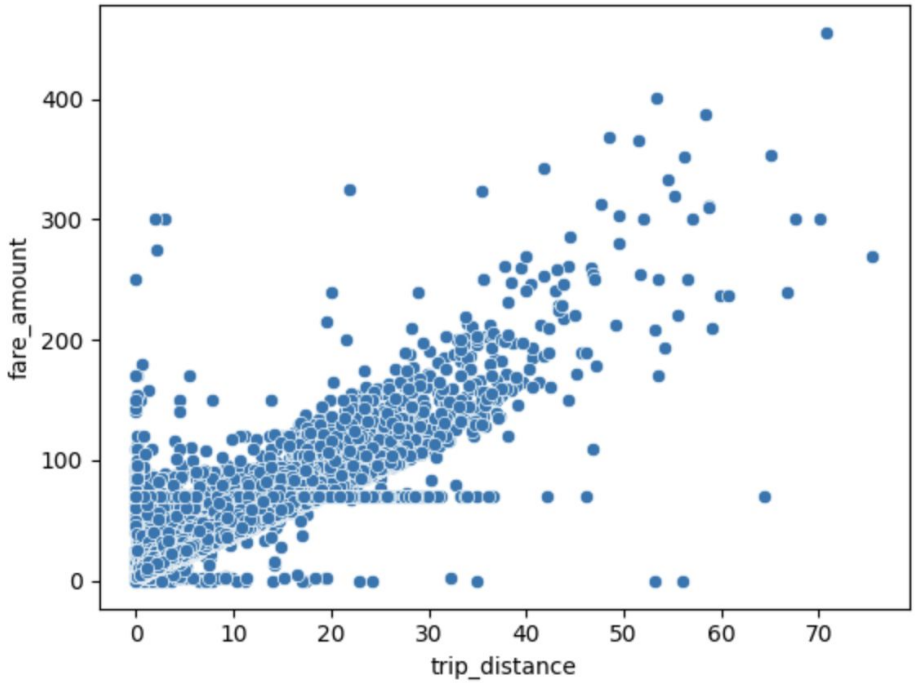
Proportion of revenue by quarter

- Similar proportion by quarter
- Q2 and Q 4 have larger contributions



Trip Distance and Fare Amount

Shows a linear relationship

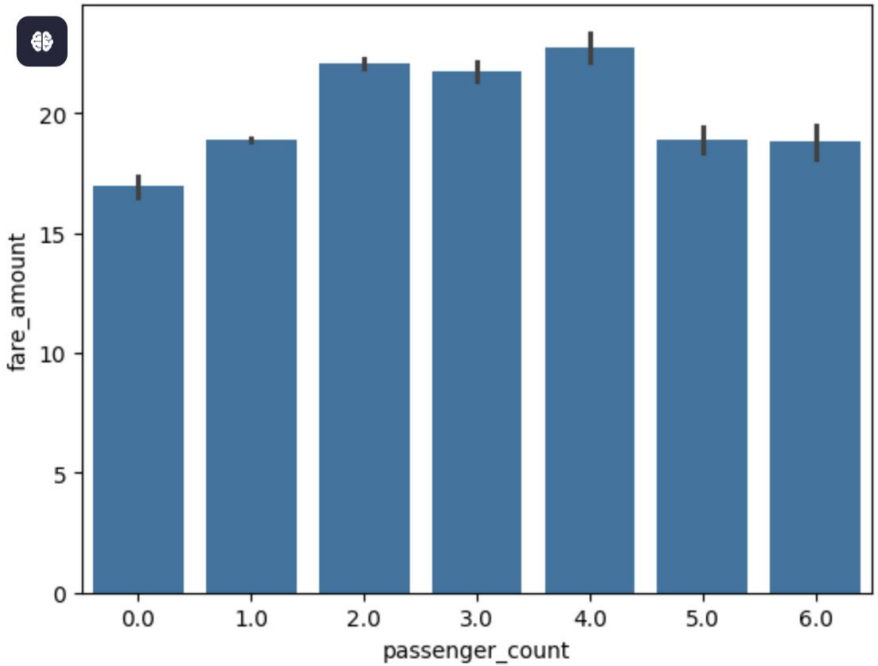


Passenger count vs fare amount

Ideal passenger count

for higher

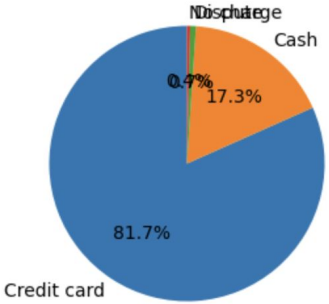
prices are around 3 and 4

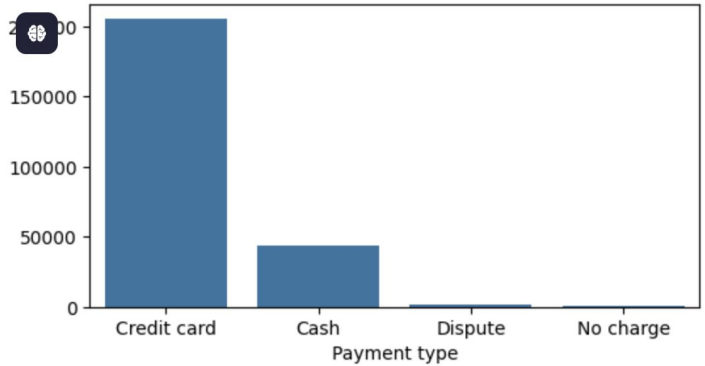


Cash vs Credit customer

Credit card is the most used

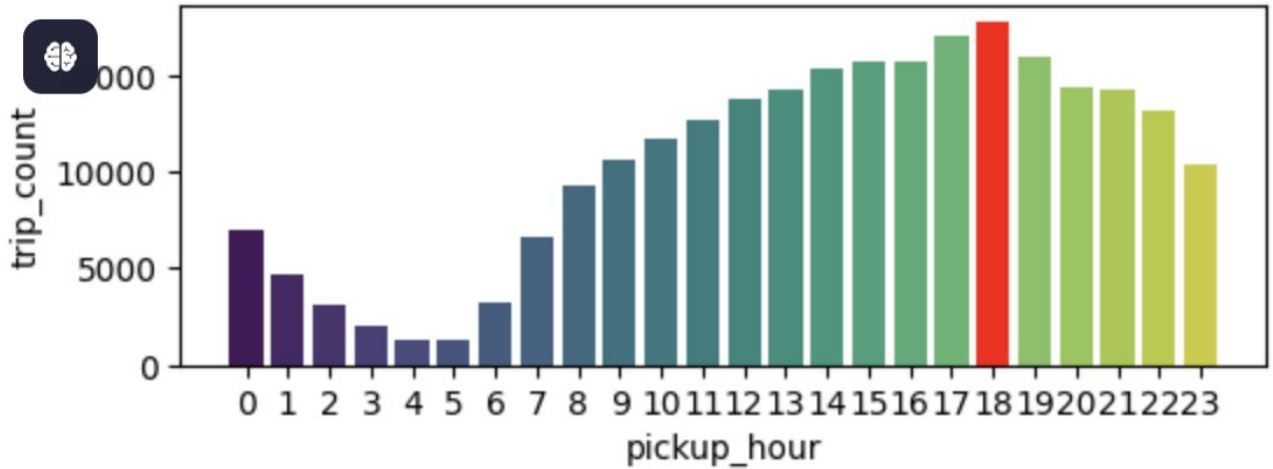
payment method





Busiest hour and pickups by hours

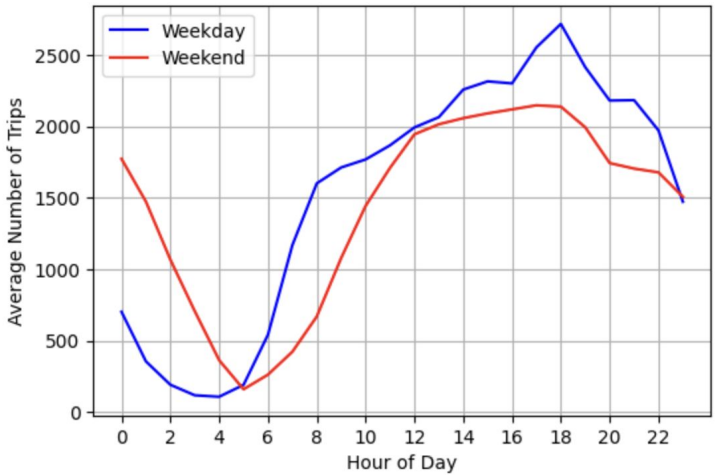
```
Busiest hour: 18:00 - 18:59
Number of trips: 17850
```



WEEKDAY VS WEEKEND

Weekday patterns typically show two distinct peak periods:

- Morning rush hour (around 8-9 AM) when people commute to work
- Evening rush hour (around 5-7 PM) when people return home



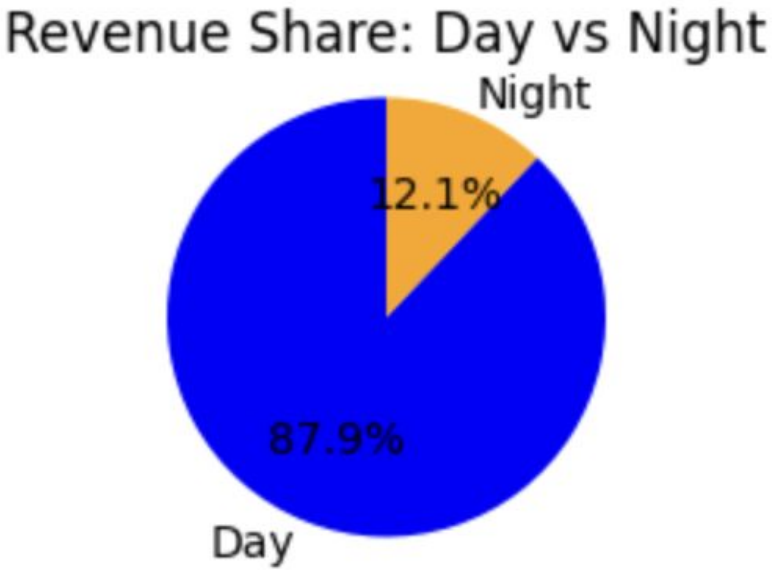
Weekend patterns usually show:

- A later morning rise in activity (people sleep in)
- A more gradual increase throughout the day
- Higher evening/night activity (dining, entertainment)
- A peak that often occurs later than weekday peaks

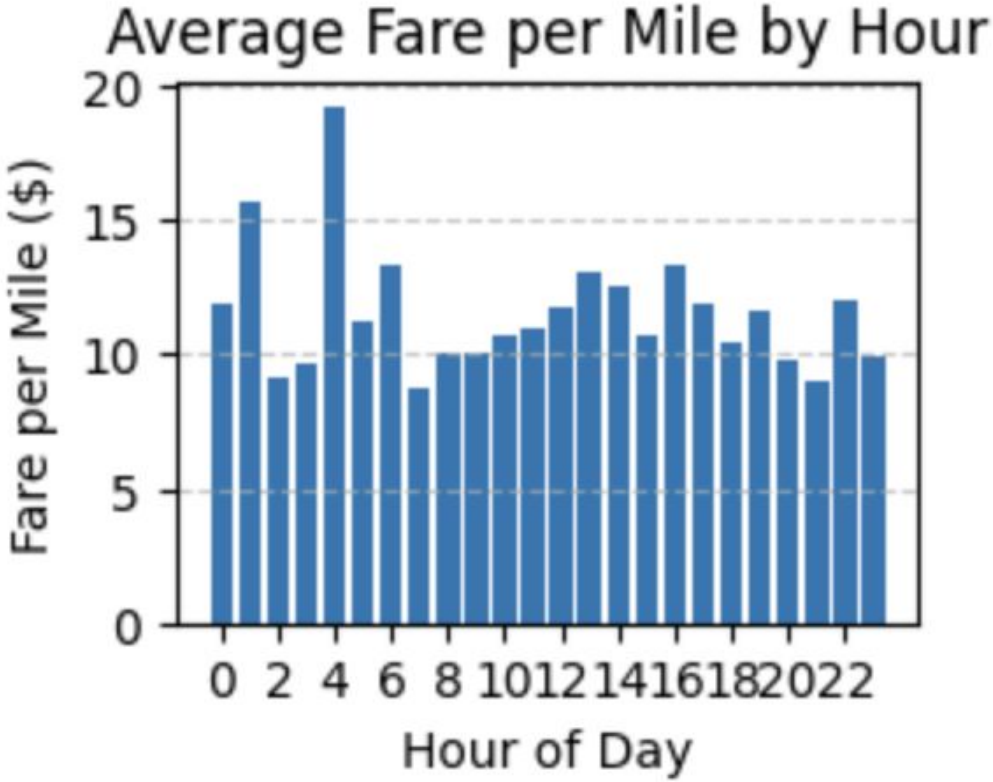
Key business implications:

- Driver allocation should be adjusted based on day of week
- More drivers needed during weekday rush hours
- On weekends, coverage should be stronger in afternoons and evenings
- Early morning hours (2-5 AM) typically show minimal demand on both weekdays and weekends

Revenue is more in the day than night

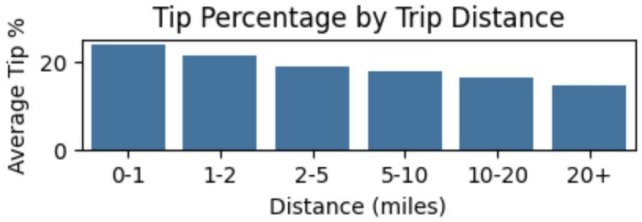


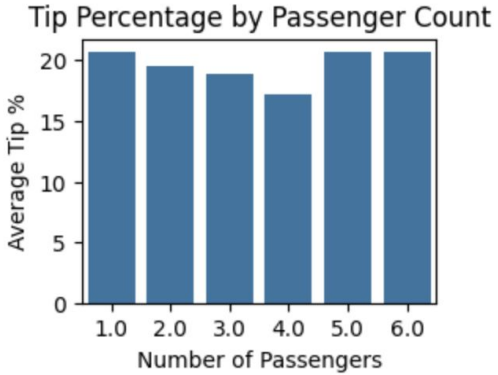
Average fare is high in the early morning hours

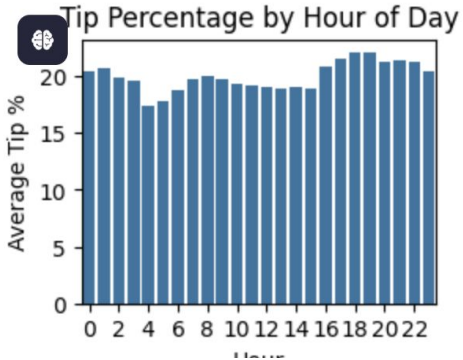


TIP BY DISTANCE

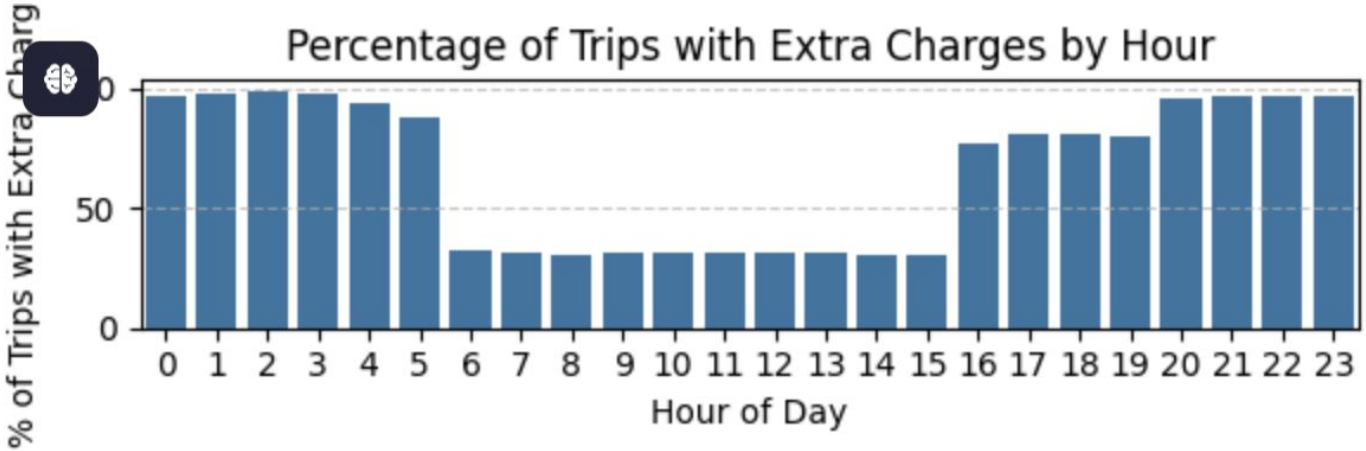
- Very short trips (passengers may feel less inclined to tip for minimal service)
- Very long trips (the fare is already high, so percentage may be lower)
- Trips with higher passenger counts (cost per person is lower, possibly leading to lower tips)
- Early morning hours (passengers may be tired or using taxis out of necessity rather than convenience)
- Late night/early morning hours when passengers might be intoxicated or less attentive







EXTRA CHARGES BY HOUR



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

Strategic Dispatch Improvements

Predictive Demand Forecasting

- Implement a predictive algorithm that forecasts demand by zone and hour based on historical patterns
- Pre-position vehicles in areas with anticipated high demand 15-30 minutes before peak times

Dynamic Vehicle Allocation

- Adjust the number of available taxis by time of day based on the identified demand patterns
- Increase fleet availability by 15-20% during weekday morning (7-9AM) and evening (5-7PM) rush hours
- Reduce fleet size during consistently low-demand periods (e.g., 2-5AM weekdays) to minimize idle time

Zone-Based Balancing

- Implement incentives for drivers to relocate to high pickup/low dropoff ratio areas
- Establish a "pickup probability score" that helps drivers identify areas with high likelihood of getting fares
- Create a rebalancing bonus during off-peak hours to maintain coverage in areas with irregular demand

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

Route Optimization Strategies

Congestion-Aware Routing

- Develop alternative route suggestions for the identified slow routes during peak hours
- Incorporate real-time traffic data to dynamically adjust recommended routes
- Create a database of historically slow intersections and areas to avoid during specific hours

Time-Based Route Selection

- Implement different routing algorithms based on time of day
- Prioritize speed during off-peak hours and reliability during rush hours

Multi-Passenger Efficiency

- Optimize routing for vehicles with multiple passengers(upto 4) to minimize overall trip time
- Develop zone-specific routing strategies based on typical passenger counts in those areas

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

Time-Based Positioning

Rush Hour Deployment (7-9AM, 5-7PM Weekdays)

- Deploy vehicles to residential areas with high morning outflow
- Pre-position cabs in office-dense areas 30 minutes before evening rush

Mid-Day Strategy (10AM-4PM)

- Redistribute 20% of vehicles to shopping and tourist areas
- Rotate vehicles between high-turnover zones to maximize trip volume

Night Operations (11PM-5AM)

- Concentrate 40% of night fleet around entertainment districts and nightlife hotspots
- Maintain strong presence near major hotels for airport runs
- Position vehicles strategically near late-night restaurant zones
- Reduce coverage in predominantly residential areas

- **Day-of-Week Adjustments**
-
- **Weekday Focus**
-

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

Day-of-Week Adjustments

Weekday Focus

- Prioritize business districts and commuter routes
- Position vehicles to capture predictable commuting patterns

Weekend Deployment

- Shift 25% of fleet from business districts to entertainment/shopping zones
- Increase coverage in tourist areas and parks during daylight hours
- Enhance late-night coverage around nightlife centers
- Maintain strong airport connection coverage throughout weekends

Seasonal Adjustments

Month-to-Month Planning

- Adjust zone coverage based on identified monthly variations
- Increase airport service during peak tourist seasons
- Develop specialized holiday deployment strategies
- Scale back service in typically lower-demand months in specific zones
-

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

Zone-Specific Strategies

High Pickup/Low Dropoff Zones

- These are "origination hubs" - ensure consistent coverage
- Implement a rotation system to prevent oversaturation
- Schedule regular vehicle replenishment during peak demand hours

Low Pickup/High Dropoff Zones

- These are "destination sinks" - implement incentives for drivers to remain
- Create "reverse flow" bonus payments during key times
- Develop multi-zone circuits to keep drivers in profitable routes

Balanced Zones

- Maintain steady coverage in zones with balanced pickup/dropoff ratios
- Use these areas as "neutral rebalancing zones" between demand spikes
- Develop specialized micro-positioning within larger balanced zones

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

Tiered Distance Pricing

Short Trips (0-2 miles)

- Implement a slightly higher per-mile rate for short trips to compensate for fixed costs
- Set base fare at $3.00 with a per-mile rate of $3.50 for the first 2 miles
- This addresses the finding that shorter trips have higher operational costs per mile

Medium Trips (2-5 miles)

- Offer a moderate per-mile rate to remain competitive with ridesharing services
- Set per-mile rate at $2.75 for miles 2-5
- This price point balances profitability with competitiveness in the most common trip distance range

Long Trips (5+ miles)

- Provide a volume discount for longer trips
- Set per-mile rate at $2.25 for miles beyond 5
- This encourages longer trips while remaining profitable due to amortized fixed costs

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

Time-Based Pricing Adjustments

Peak Hour Surcharges

- Implement a $1.50 surcharge during identified peak hours (7-9AM, 5-7PM weekdays)
- Analysis shows these times have highest demand and lowest price sensitivity
- Adjust surcharge based on day of week (higher on Friday, lower on Monday)

Night Premium

- Apply a $1.00 premium for trips between 11PM-5AM
- Data shows night trips have different passenger profiles with less price sensitivity
- Compensates drivers for working less desirable hours while capturing revenue from nightlife activity

Dynamic Multiplier

- Implement a subtle multiplier (1.1x - 1.3x) during periods of exceptionally high demand
- Cap increases to prevent customer alienation
- Use historical data to predict and pre-announce likely surge periods

Zone-Based Adjustments

High-Demand Zone Premium

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

Zone-Based Adjustments

High-Demand Zone Premium

- Apply a modest $0.75 pickup premium in the top 10 highest-demand zones
- Research shows passengers in these zones prioritize availability over small price differences
- Graduate the premium based on demand intensity by zone

Congestion Zone Pricing

- Implement a $1.00 surcharge for pickups/dropoffs in identified congestion zones
- Compensates for longer wait times and slower speeds in these areas
- Aligns with city congestion reduction goals while capturing appropriate revenue

Airport Specialized Rates

- Maintain competitive flat rates to/from airports
- Implement special rates during peak flight arrival/departure times
- Offer multi-passenger discounts to compete with shared shuttle services

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

Passenger-Count Optimization

Group Ride Incentives

- Offer a 10% discount for trips with 3+ passengers
- Data shows larger groups typically require less price incentive per person
- Increases vehicle utilization and total revenue per trip

Solo Passenger Premium

- Consider a subtle $0.50 premium for single-passenger rides during peak hours
- This encourages carpooling and captures appropriate revenue from those requiring exclusive service