Report With Data Analysis

Data: High Volume For-Hire Vehicle Trip Records (PARQUET)

Data description:

I'm analyzing the High-Volume for-Hire Vehicle (HVFHV) Trip Records for October 2024. I chose this dataset to examine trip patterns during a typical month of the previous year. October is ideal because it has fewer holidays, and schools and offices generally follow a regular schedule. This allows for a clearer analysis of how various factors influence ride-hailing trips.

Dataset Overview:

The HVFHV trip data includes trips dispatched by major rideshare companies like Uber and Lyft, which operate under a special license in NYC.

Key Points:

- Each row represents a single trip taken in an HVFHV service.
- Local Law 149 of 2018 established a new license category for companies dispatching 10,000+ daily trips.
- This regulation took effect on February 1, 2019, meaning HVFHV data does not exist before this date.

Research:

Research Question:

How does the timing of rides influence trip characteristics (distance, duration, fare) and driver earnings?

Hypothesis:

Rides during peak hours tend to be shorter in distance but have higher fares and driver earnings due to increased demand and congestion. In contrast, off-peak rides are typically longer but generate lower earnings per mile.

Assumptions:

- **Fare Structure:** Fares are primarily determined by base rates, surge pricing, and tolls.
- **Congestion & Demand:** Higher fares during peak hours are assumed to reflect both increased congestion and demand, which can be inferred from trip duration and distance trends.

Motive:

As the number of TLC drivers increases, many assume that driving during peak hours with surge pricing will maximize their earnings. This leads to a high concentration of drivers on the road, causing even greater congestion and an oversupply of vehicles relative to demand.

This study aims to:

1. **Evaluate Peak Hour Profitability:** Identify peak hours and assess whether they truly result in higher earnings or whether congestion offsets the benefits of surge pricing.

2. **Inform Policy for Fair Earnings Distribution:** Provide data-driven insights to help policymakers develop strategies to prevent all drivers from flooding the roads at the same time. One such strategy could involve more equitably distributing peak-hour driving opportunities.
3. **Incentivize Off-Peak Driving** – Explore ways to make off-peak driving more attractive, such as implementing compensation adjustments or incentives for driving during non-peak hours.
4. **Identify Hidden High-Earning Periods** – Determine if there are specific non-peak times that still yield higher earnings, offering alternative opportunities for drivers to optimize their schedules.

By analyzing these factors, this research can contribute to improving driver earnings, reducing congestion, and enhancing overall rideshare efficiency.

Data Analysis:

Data Cleaning and Preparation:

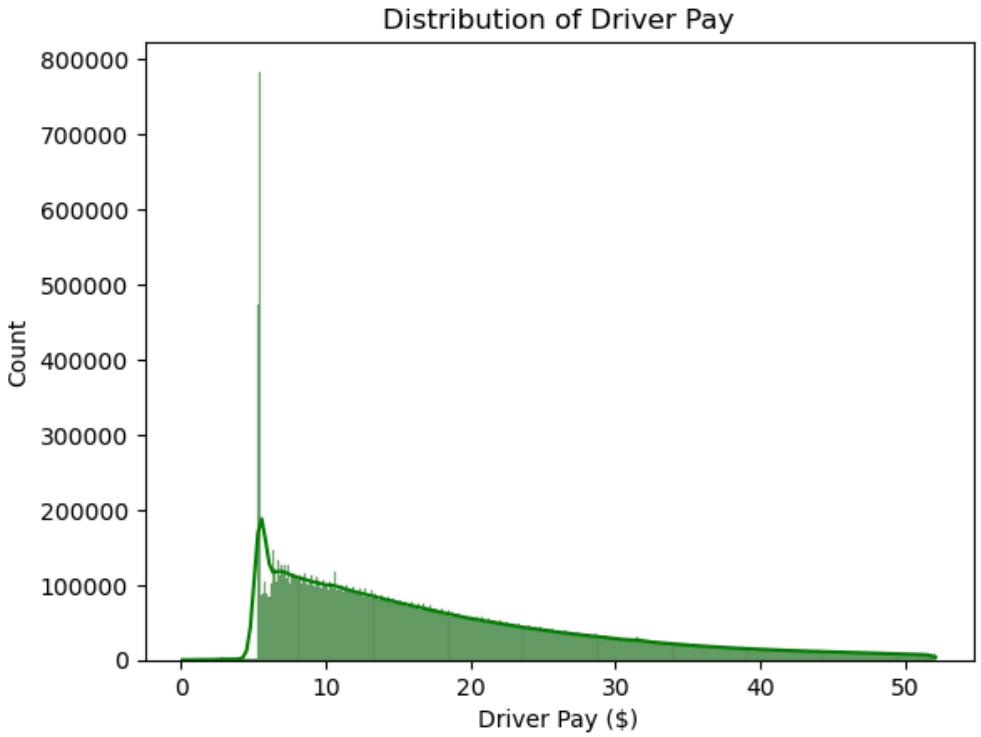
The dataset initially contained **20,028,282 rows** and **17 columns**. After removing duplicates and handling missing values, the dataset was cleaned and prepared for analysis.

Key steps included:

- **Dropped Insignificant Columns**: Removed 12 columns (e.g., hvfhs_license_num, PULocationID, shared_match_flag) that were irrelevant to the analysis.
- **Handled Missing Values**: Dropped rows with missing values in critical columns like per_mile_earnings and efficiency.
- **Created New Columns**: Added hour_of_day, trip_time_minut, per_mile_earnings, and efficiency for deeper analysis.
- **Removed Outliers**: Used the IQR method to remove extreme values in driver_pay, trip_miles, and trip_time_minut.

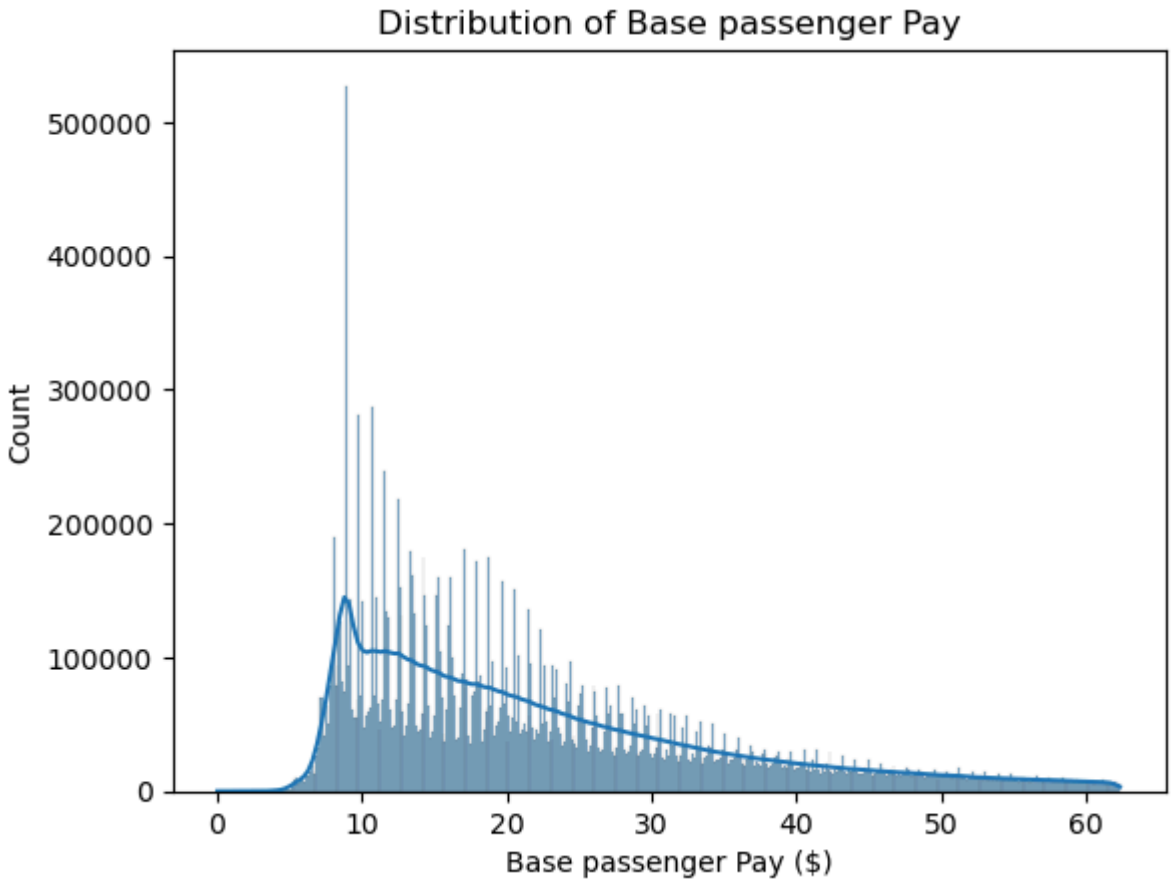
This cleaning process ensured the dataset was focused, reliable, and ready for analysis.

Univariate Analysis



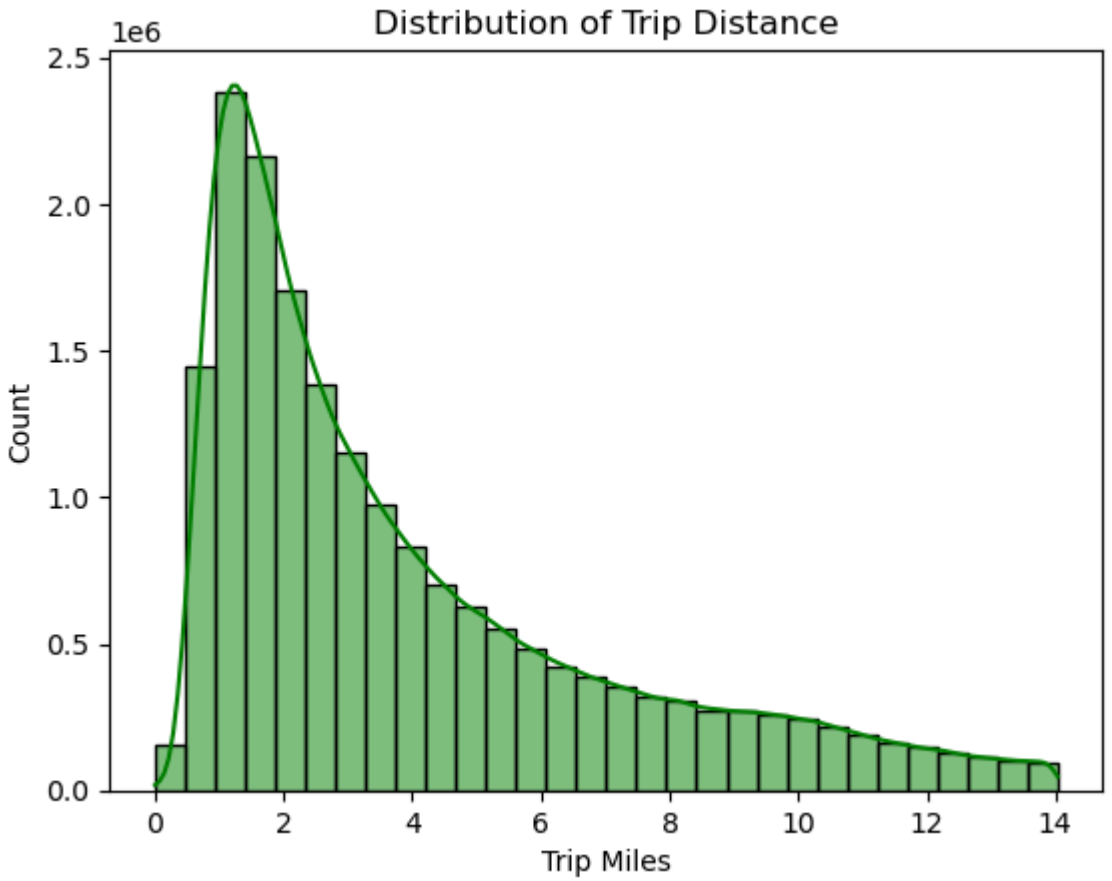
Distribution of Driver Pay per trip

- **Shape**: Right-skewed.
- **Insight**: Most drivers earn between **$5 and $50 per trip**, with a sharp drop-off beyond $50. The long tail indicates rare high-earning trips (e.g., long-distance rides or surge pricing).
- **Key Takeaway**: Driver earnings are concentrated in lower ranges, with occasional outliers due to exceptional trips.



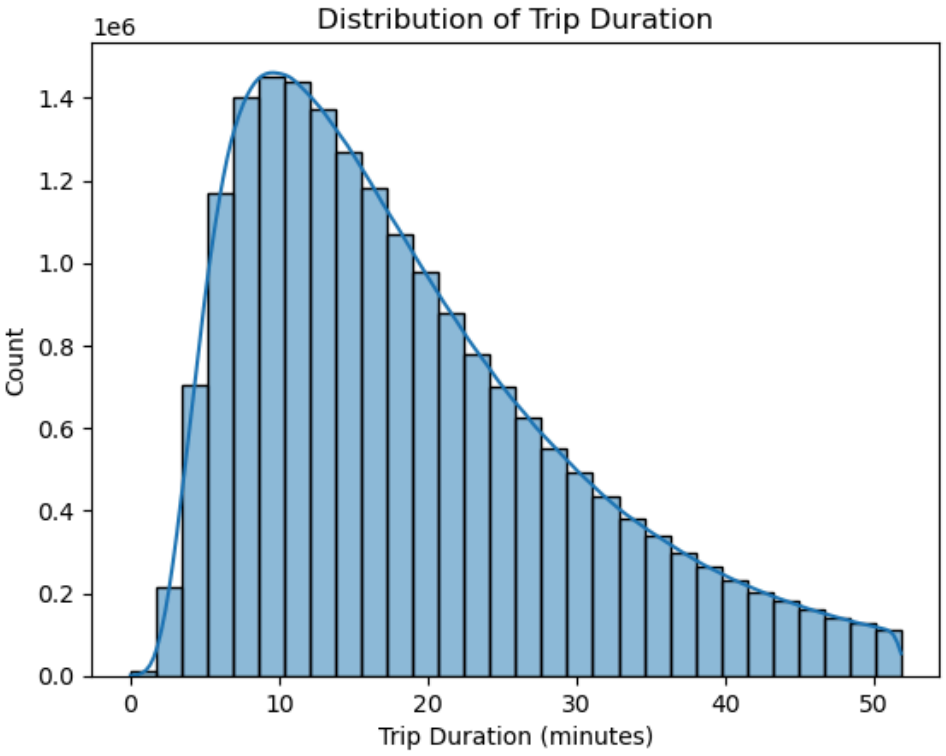
Distribution of Base Passenger Pay

- **Shape**: Right-skewed.
- **Insight**: Base fares peak around **$20–$30**, slightly higher than driver pay, reflecting company fees.
- **Key Takeaway**: Passengers pay moderate fares for most trips, with a few expensive outliers (e.g., airport trips or premium services).



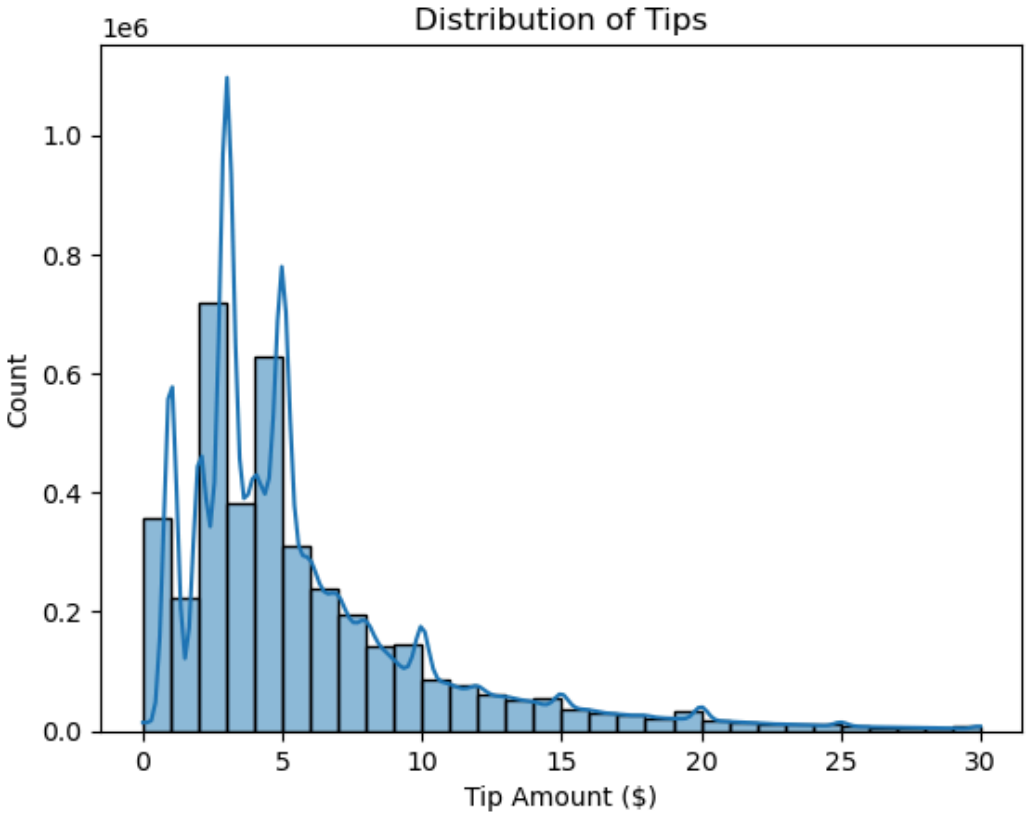
Distribution of Trip Distance

- **Shape**: Extremely right-skewed.
- **Insight**: Over 50% of trips are **under 2 miles**, with very few exceeding 10 miles.
- **Key Takeaway**: Taxis are primarily used for short, urban trips, aligning with NYC's dense, gridlocked environment.



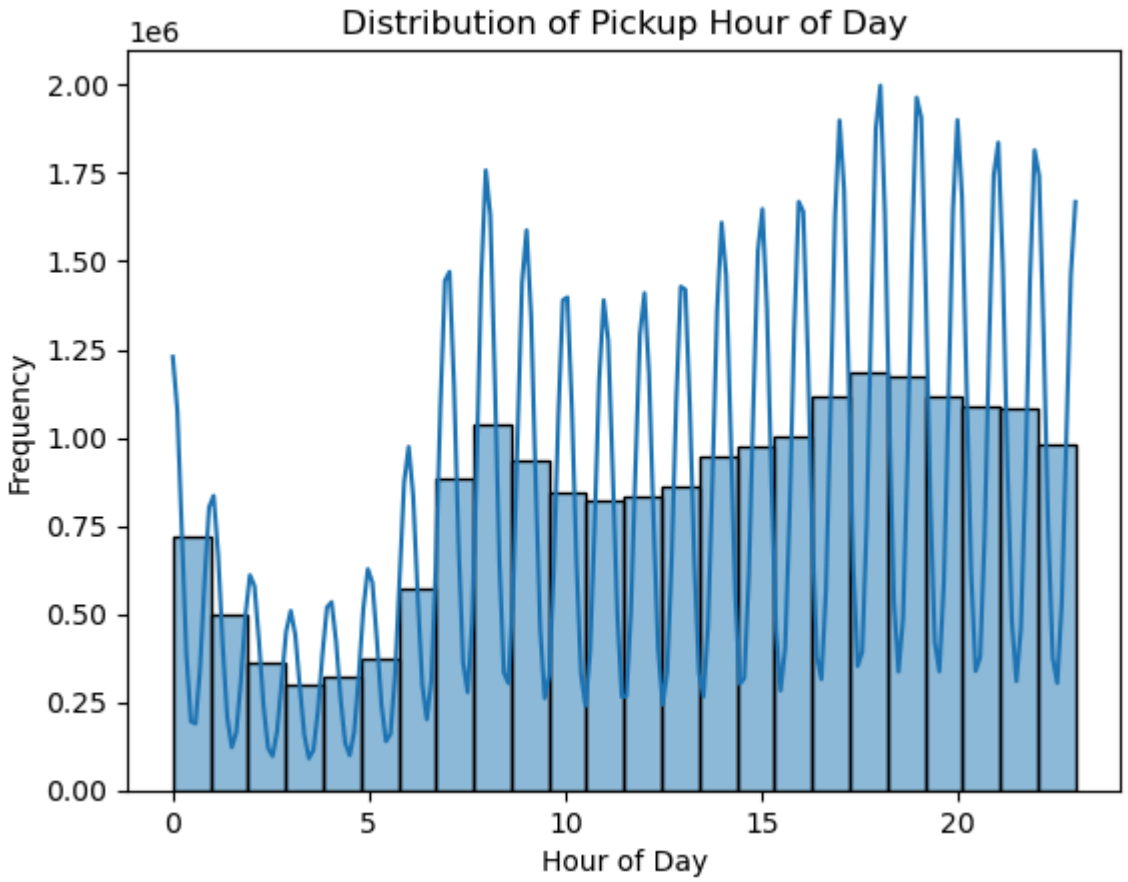
Distribution of Trip Duration

- **Shape**: Right-skewed.
- **Insight**: Most trips last **10–15 minutes**, with longer durations (>30 minutes) being rare.
- **Key Takeaway**: Short trip durations mirror short distances, but congestion during peak hours may inflate times slightly.



Distribution of Tips

- **Shape**: Right-skewed.
- **Insight**: Tips are mostly **$0–$5**, peaking around $2. Larger tips (>$10) are rare.
- **Key Takeaway**: Tipping is modest and standardized, likely tied to short-trip norms or automated percentage-based tips.



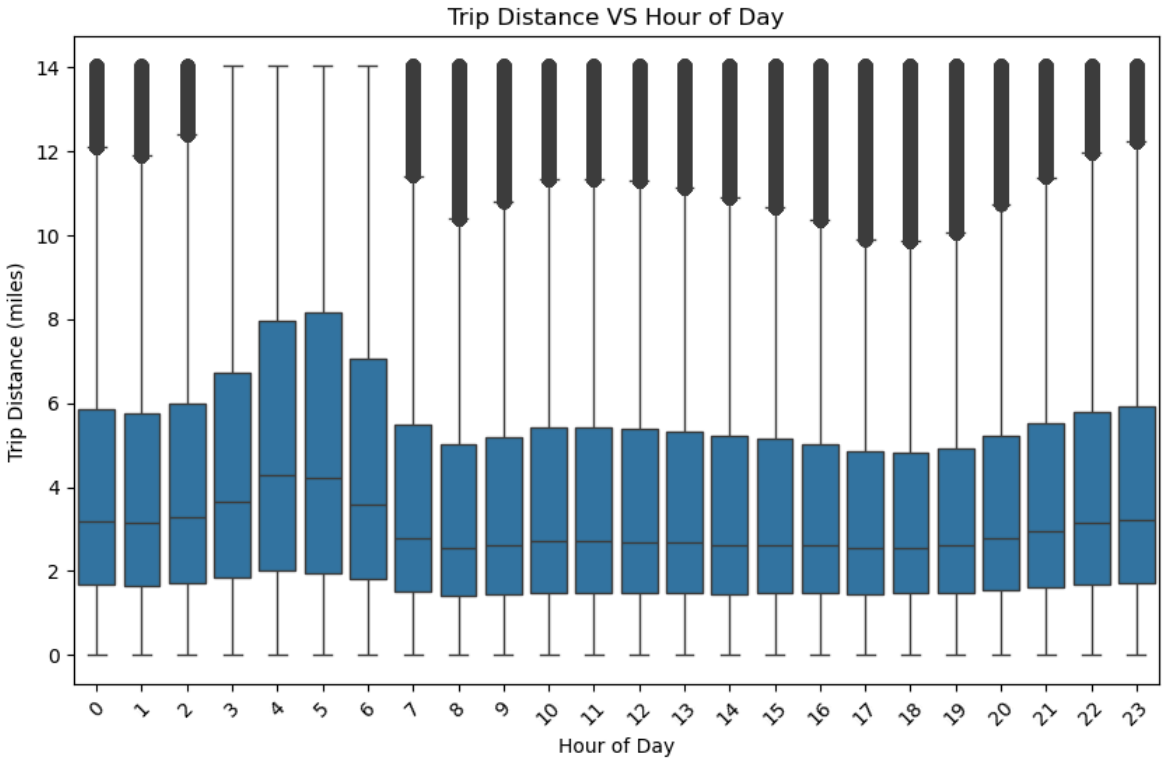
Distribution of Pickup Hour of Day

- **Shape**: Bimodal.
- **Insight**: Peaks at **7–9 AM** and **5–8 PM** (rush hours). Low activity overnight.
- **Key Takeaway**: Demand aligns with commuter schedules, emphasizing taxis' role in daily urban mobility.

Similarities & Differences

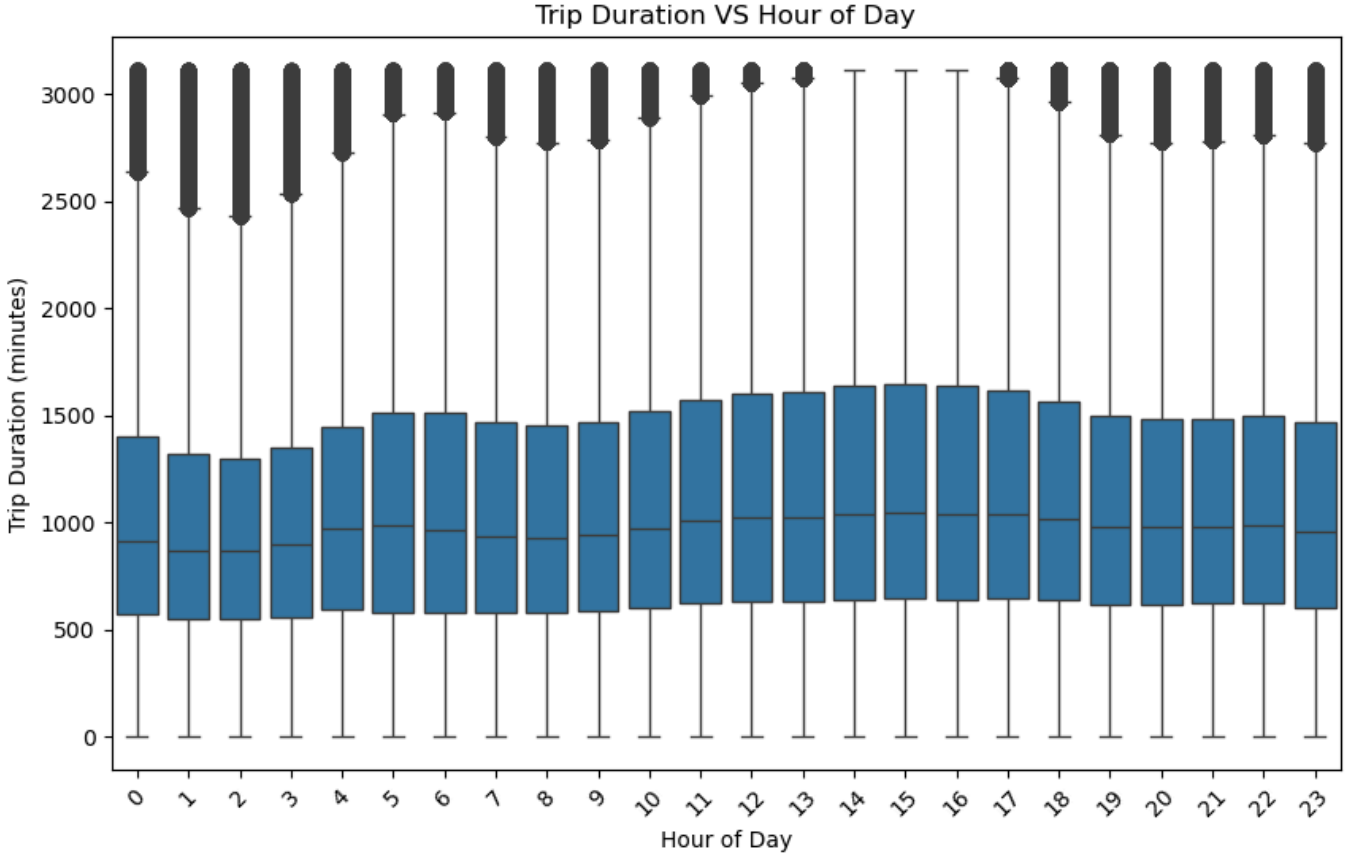
- **Shared Skewness**: All except pickup hours are **right-skewed**, indicating most values cluster at lower ranges with rare high outliers.
- **Unique Feature**: Pickup hours are **bimodal**, reflecting demand spikes during commute times.
- **Common Theme**: Taxi usage is dominated by **short, low-cost, time-sensitive trips** during peak hours, with earnings and fares reflecting this pattern.

Relationships Between Variables



Trip Distance vs. Hour of Day

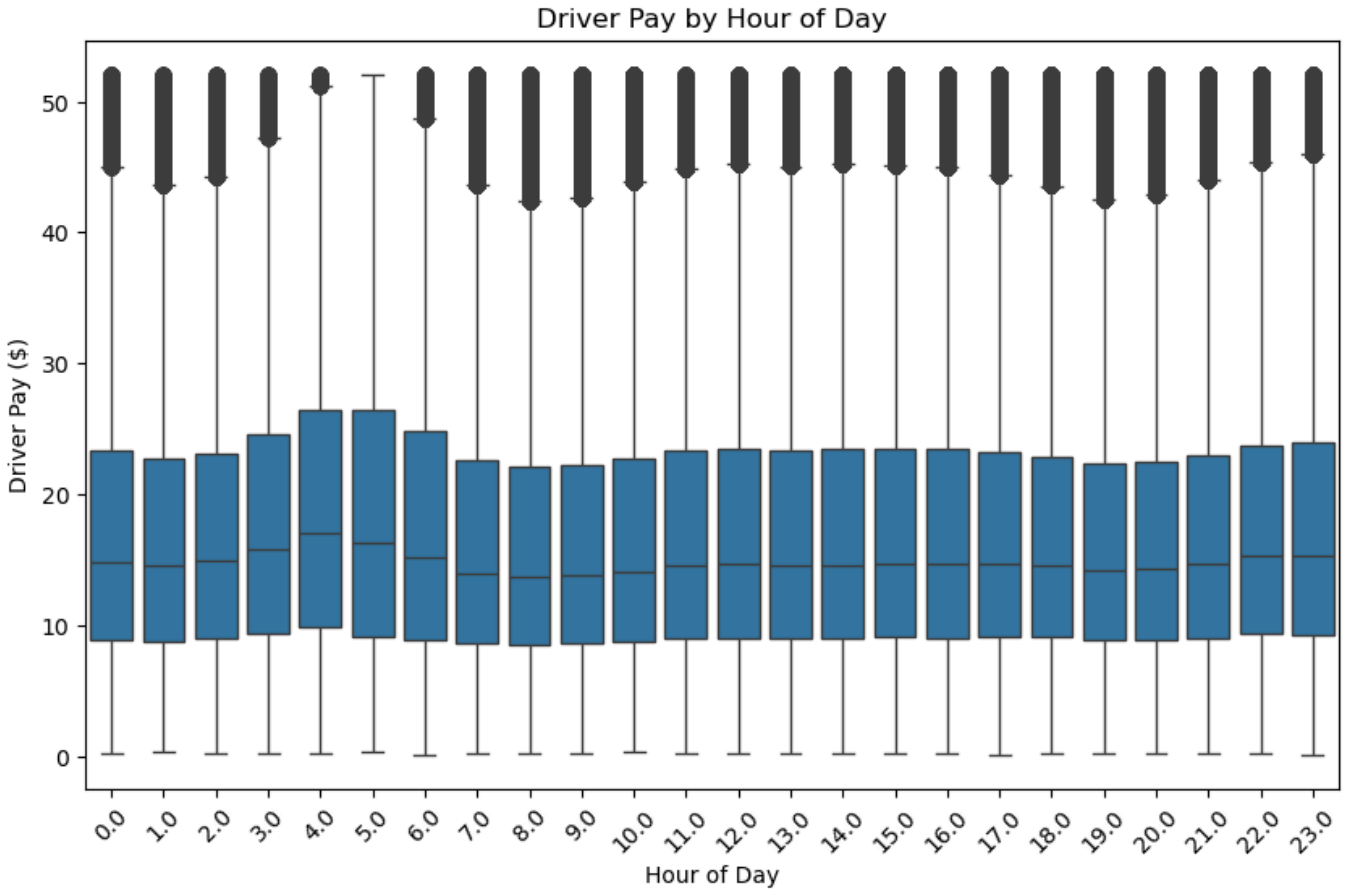
- **Relationship**: Trip distances are shortest during **peak hours (7–9 AM and 5–7 PM)**, averaging **2–4 miles**. Distances increase during off-peak hours.
- **Insight**: Shorter trips dominate peak hours, likely due to commuter demand for quick urban rides.
- **Implication**: Drivers prioritize short, frequent trips during peak times but face congestion-related delays.



Note** Y-axis shows trip time in seconds instead of minutes

Trip Duration vs. Hour of Day

- **Relationship**: Trip durations are longest during **peak hours (7–10 AM and 3–6 PM)**, exceeding 30 minutes on average.
- **Insight**: Congestion extends trip times during rush hours, even though trips are shorter in distance.
- **Implication**: Longer durations reduce the number of trips drivers can complete, offsetting higher per-mile earnings.



Driver per Trip Pay by Hour of Day

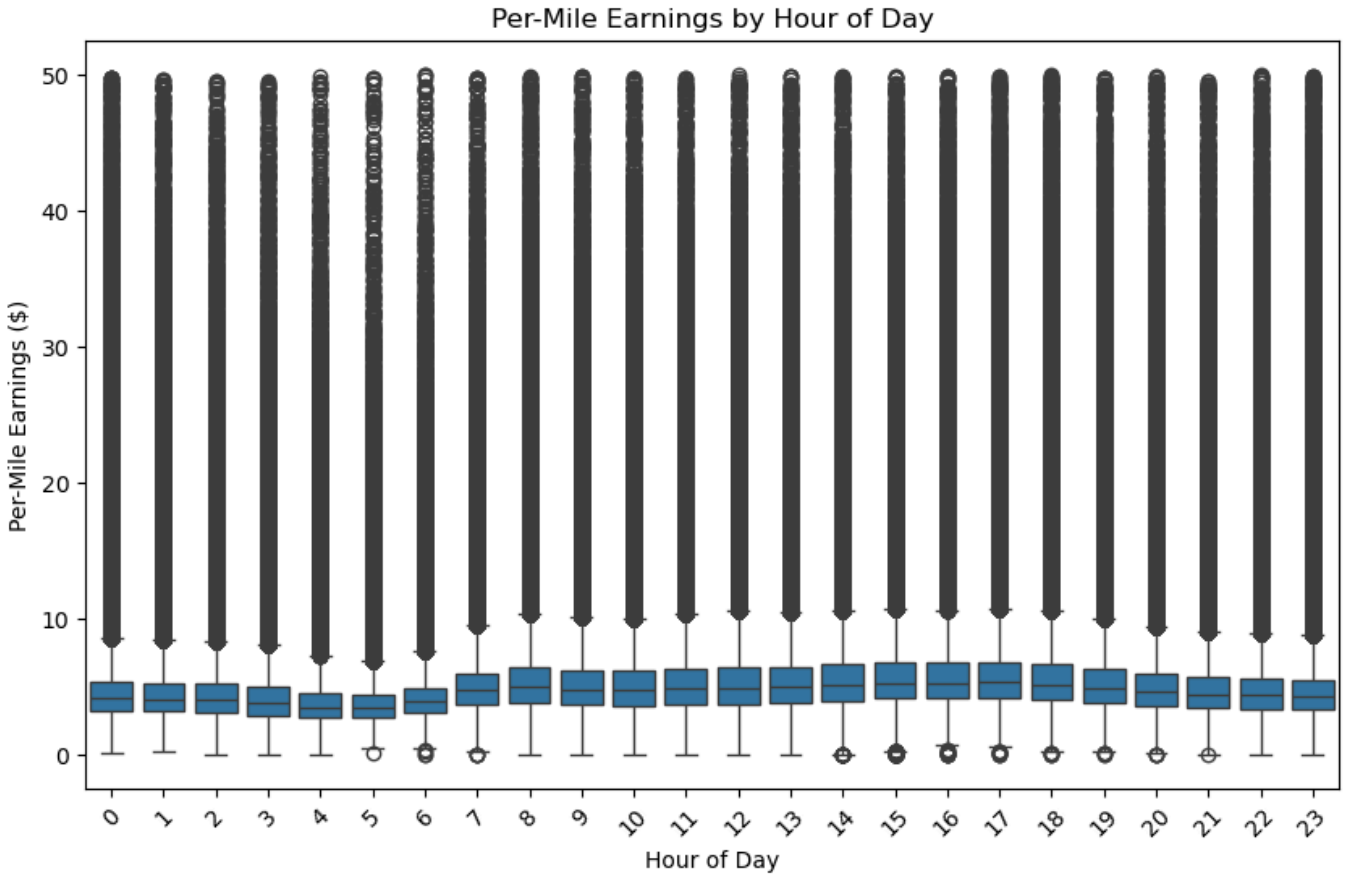
Early Morning (4–5 AM):

- **Higher Earnings**: Driver pay per trip averages **$25/hour**, which is higher than typical off-peak rates (e.g., midday or late night).
- **Why?**:
  - **Low Driver Supply**: Fewer drivers on the road reduce competition, allowing those working to capture more rides.
  - **Niche Demand**: Early morning trips often include **airport rides** or **night-shift commuters**, which tend to have higher base fares and tips.
- **Implication**: Early morning driving can be a **hidden high-earning opportunity** for drivers willing to work unconventional hours.
-

Peak Hours (7–9 AM and 5–7 PM):

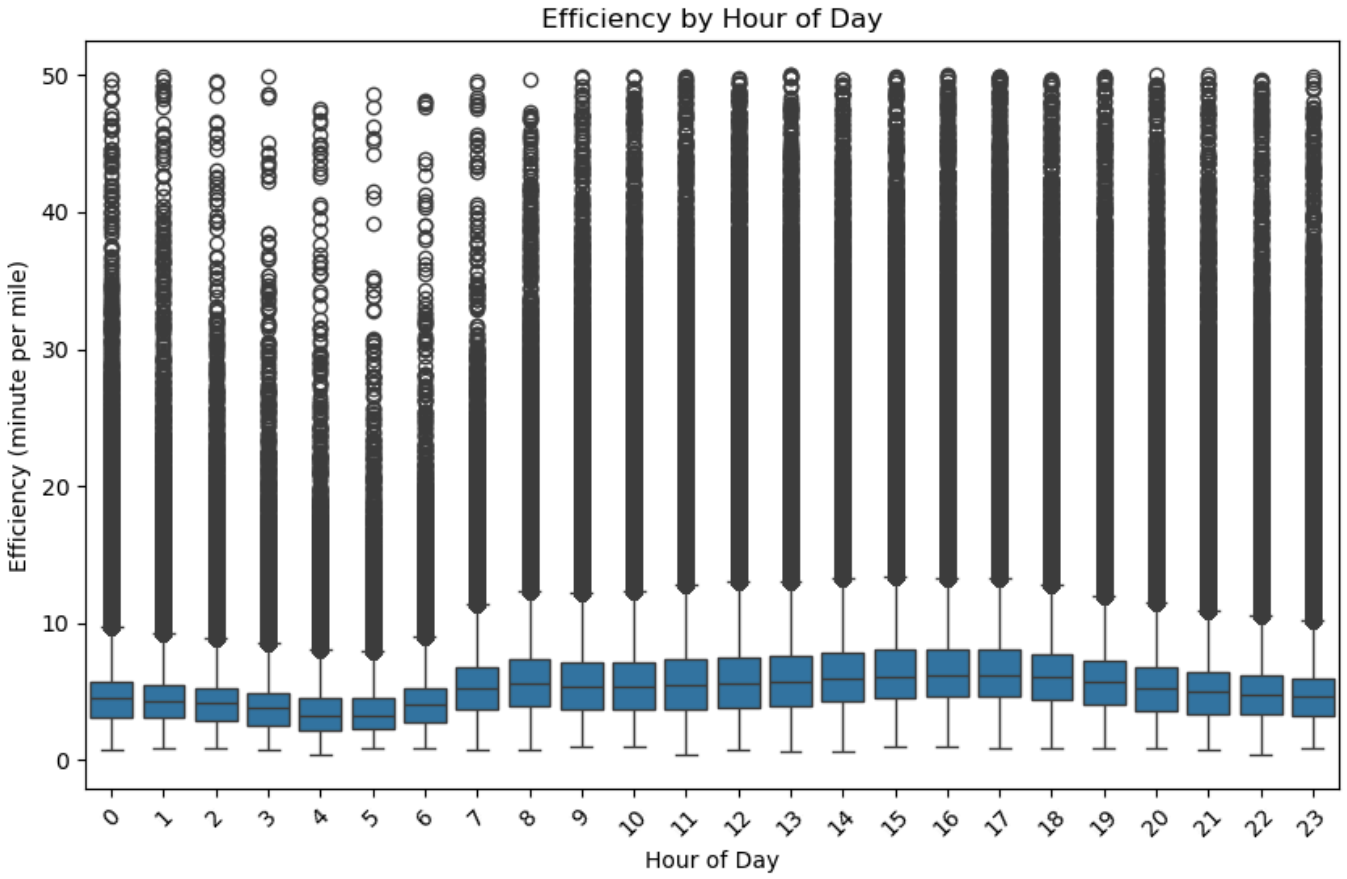
- **Higher Potential Earnings**: Driver pay per trip can spike to **$50/hour** due to surge pricing and high demand.
- **But...**:
  - **Congestion**: Longer trip durations reduce the number of rides drivers can complete.

    - **Oversupply of Drivers**: High competition dilutes individual earnings, pulling the **average pay down to $20/hour**.
- **Implication**: While peak hours offer the **potential for high earnings**, the actual take-home pay is often lower due to inefficiencies and competition.



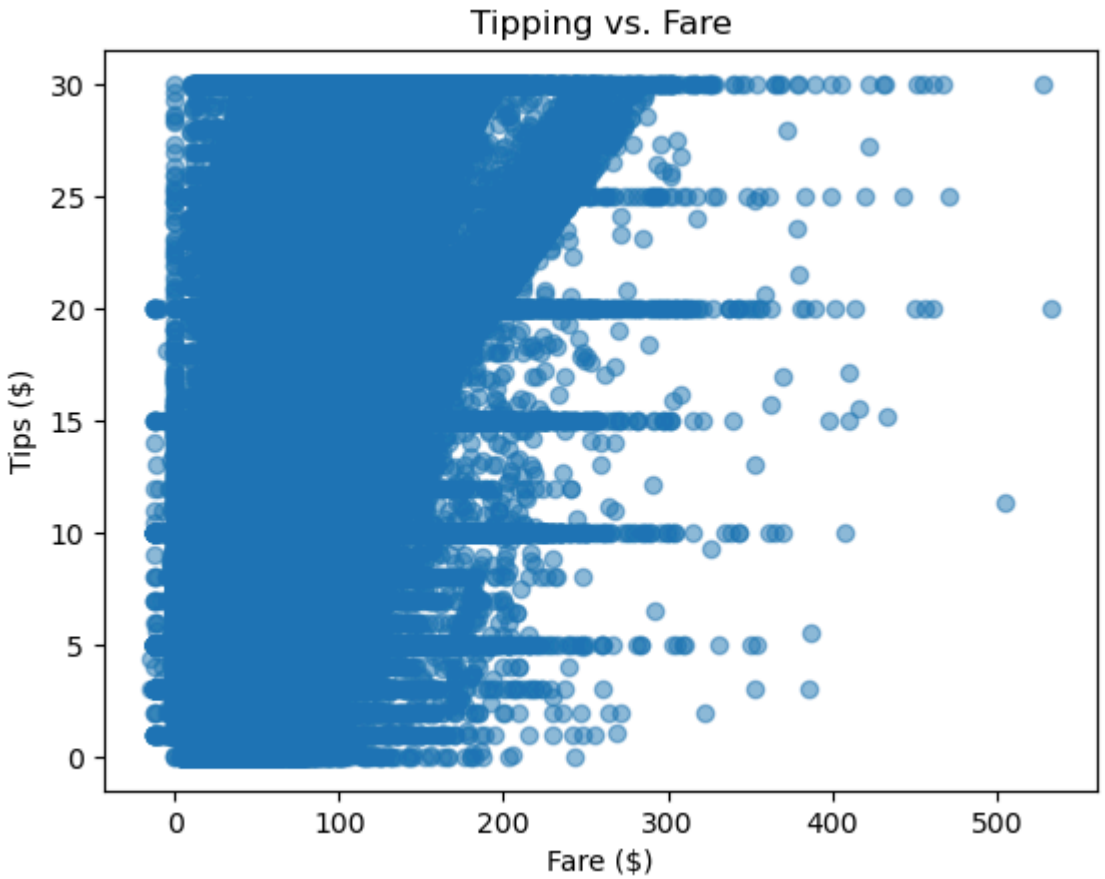
Per-Mile Earnings by Hour of Day

- **Relationship**: Per-mile earnings spike during **peak hours (7–9 AM and 3–6 PM)** but drop during midday and overnight.
- **Insight**: Surge pricing and higher demand during rush hours increase earnings per mile, drivers earn less per mile during off-peak times.
- **Implication**: Peak-hour driving maximizes per-mile earnings, but drivers must weigh this against congestion-related delays.



Efficiency by Hour of Day

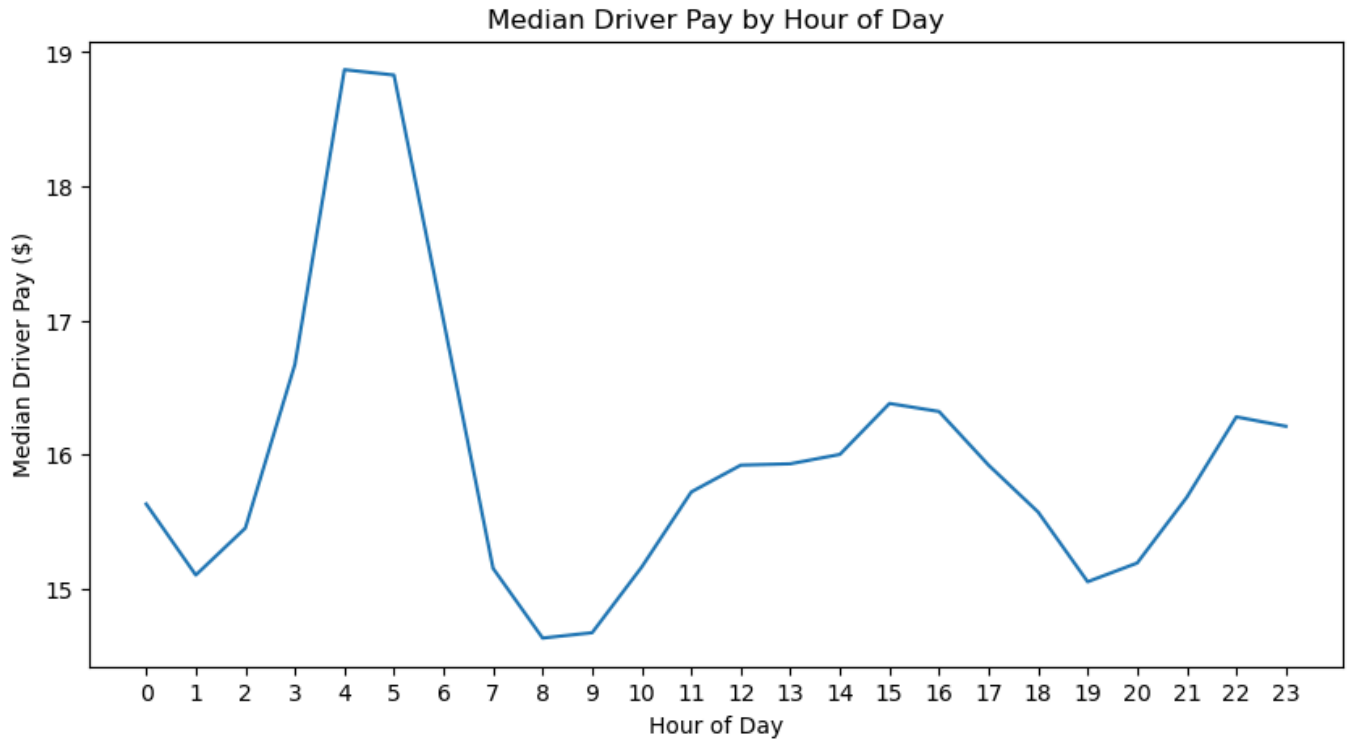
- **Relationship**: Efficiency (minutes per mile) peaks during **7–9 AM and 3-6 PM** (rush hours), indicating slower travel due to congestion.
- **Insight**: Drivers spend more time per mile during peak hours, reducing their ability to complete more trips.
- **Implication**: Congestion erodes earnings potential despite higher fares, validating the hypothesis that peak-hour gains are partially offset by inefficiency.



Tipping vs. Fare

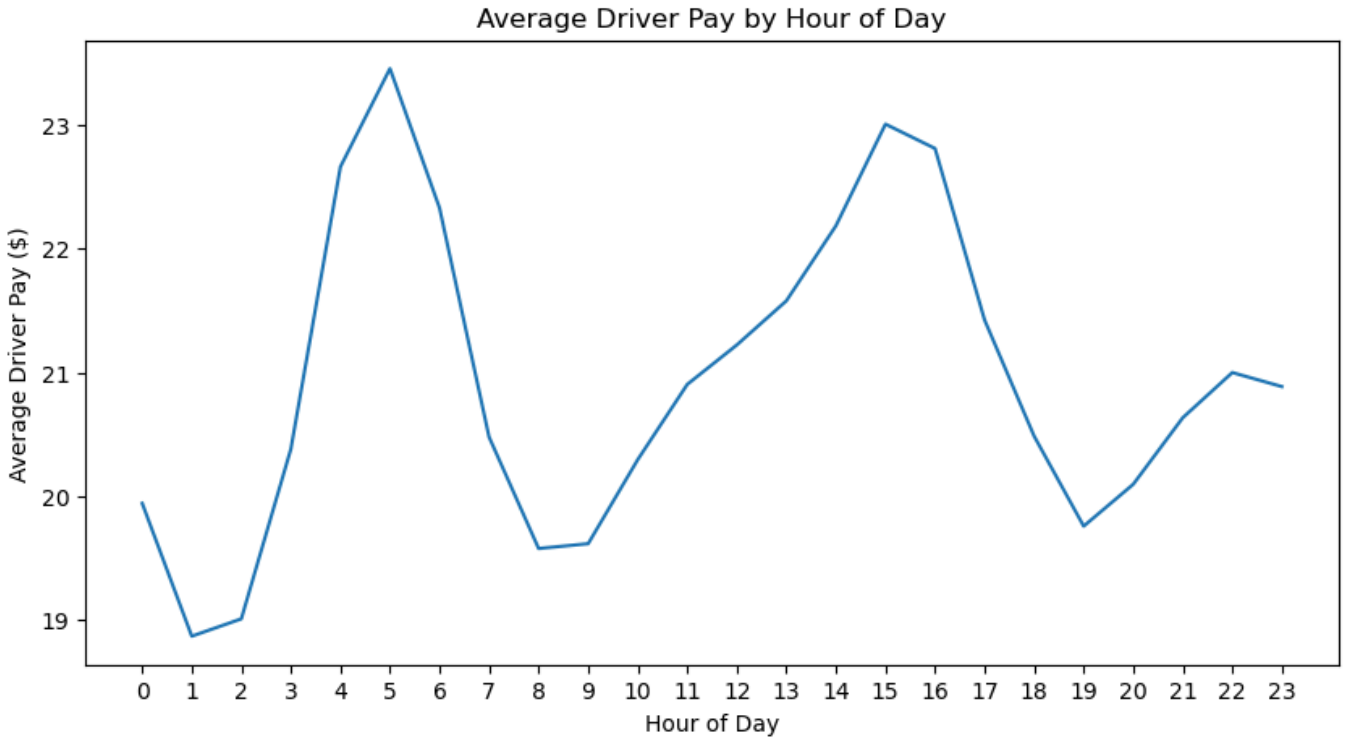
- **Relationship**: Tips are weakly correlated with fare amounts. Most tips cluster between **$0–$5**, even for fares up to $500.
- **Insight**: Passengers tip modestly regardless of fare size, suggesting automated percentage-based tipping (e.g., 10–15%) or fixed norms for short trips.
- **Implication**: Surge pricing during peak hours may not significantly boost tips, as tipping behavior appears standardized.

Time-Series Analysis



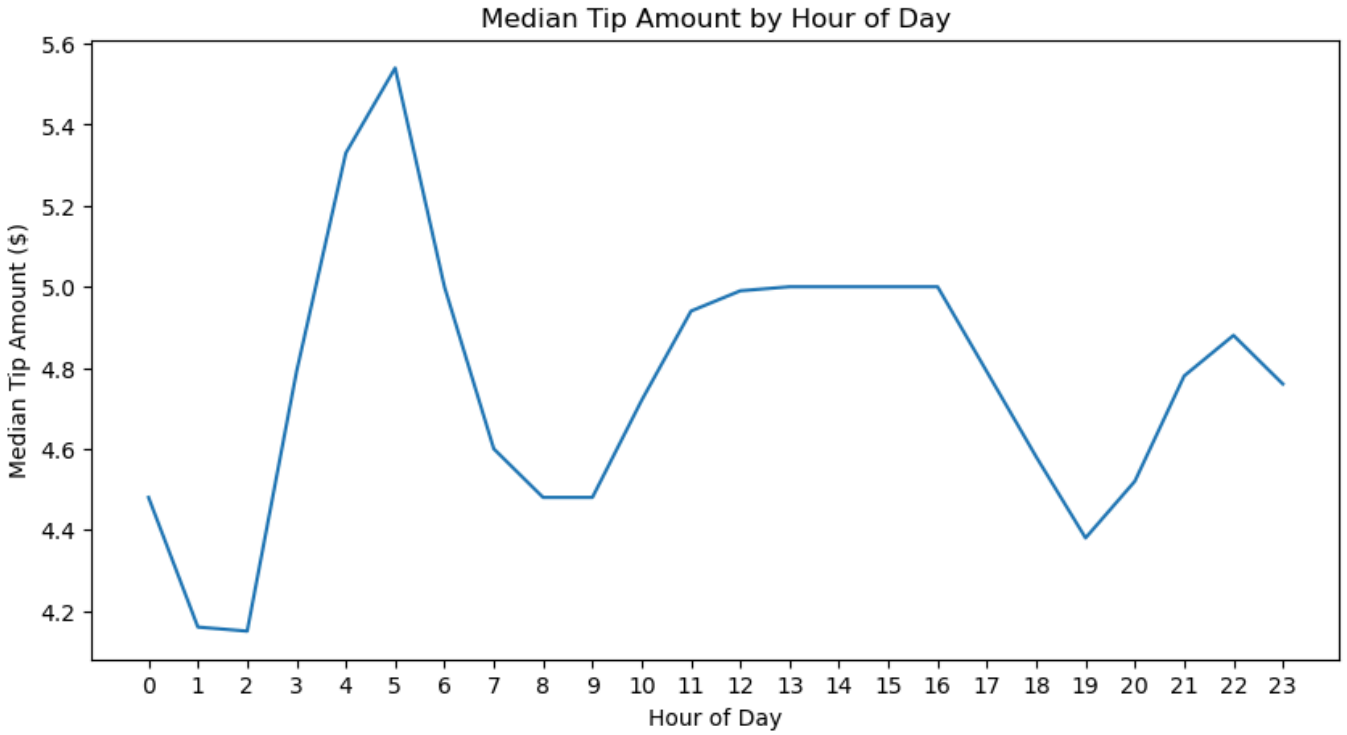
Median Driver Pay per trip by Hour of Day

- **Pattern**:
  - **Peak Earnings**:
    - **Early Morning (4–6 AM)**: Median pay spikes to **$19/trip**, likely due to low driver supply and niche demand (e.g., airport trips).
    - **Afternoon Rush (3-5 PM)**: Median pay reaches **$16.5/trip**, driven by surge pricing and high demand.
  - **Lowest Earnings**:
    - **Midday (10 AM–3 PM)**: Median pay drops below **$16** as demand decreases.
- **Insight**:
  - The **early morning (4–6 AM)** is a hidden gem for drivers, offering higher pay with less competition.
  - Peak hours (7–9 AM and 5–7 PM) are lucrative but crowded, leading to median pay stagnation despite high demand.



Average Driver Pay per trip by Hour of Day (output_66_0.png)

- **Pattern**:
  - **Peak Earnings**:
    - **Early Morning (4–6 AM)**: Average pay peaks at **$22-24/trip**, driven by fewer drivers and higher fares for long trips (e.g., airport rides).
    - **Evening Rush (3-5 PM)**: Average pay reaches **$22-23/trip**, reflecting surge pricing.
  - **Lowest Earnings**:
    - **Midday (10 AM–3 PM)**: Average pay falls to **$18–$22/hour**, aligning with lower demand.
- **Insight**:
  - The **gap between average and median pay** during peak hours indicates some trips earn significantly, skewing the average upward.
  - Early morning's high average pay confirms its potential as a high-earning window.



<u>Median Tip Amount by Hour of Day (excluding $0 tips)</u>

Timing:

- **Peak Tipping Hours (4-5 AM & 11 AM - 4 PM)**
  - The highest median tips occur early in the morning (around 4-5 AM) and again during midday (11 AM - 4 PM).
  - Early morning peaks could be due to airport rides, business travelers, or late-night riders feeling more generous.
  - Midday peaks may be influenced by professionals and tourists, who might tip more compared to regular commuters.
- **Low Tipping Hours (6-9 AM & 6-8 PM)**
  - Tips appear to drop during the early morning commute (6-9 AM), likely because many riders are using rideshare for work-related travel and may not feel the need to tip as much.
  - A similar dip occurs during early evening (6-8 PM), possibly due to post-work commuters who are less inclined to tip generously.
- **Late-Night Recovery (9 PM - 12 AM)**
  - Tips increase again later at night (9 PM - 12 AM), possibly due to riders returning from social outings, restaurants, or nightlife, where they might be in a more generous mood.

Implications for Drivers:

- **Target high-tipping periods** – Driving during early mornings and midday could result in better overall earnings due to higher tips.
- **Be strategic about evening hours** – While demand may be high, tipping is lower, so drivers might want to focus on surge pricing instead of expecting high tips.
- **Consider late-night shifts** – Working after 9 PM might be beneficial as tips start increasing again, especially in areas with nightlife.

Regression Analysis (Time of Day and Driver Earnings)

- Regression Analysis (Time of Day and Driver Earnings)
- Mean Absolute Error: 1.7636726205952786
- R-Squared Score: 0.9180175417501301
-
-

	Feature	Coefficient
0	hour_of_day	-0.010041
1	trip_miles	1.370405
2	trip_time_minut	0.564657
3	efficiency	0.038036

Key Takeaways:

1. **Trip Distance & Duration Dominate Earnings**:
    - The **$1.37 per mile** and **$0.56 per minute** coefficients highlight the importance of fare structures based on distance and time.
2. **Time of Day's Indirect Impact**:
    - While hour_of_day has a negligible direct effect, it influences earnings indirectly through:
        - **Surge Pricing**: Higher fares during peak hours.
        - **Congestion**: Longer trip durations (increasing time-based earnings) but lower efficiency.
3. **Efficiency Paradox**:
    - The positive coefficient for efficiency (minutes per mile) suggests that **congestion during peak hours is compensated by surge pricing**, leading to higher earnings despite slower trips.

Final Observation and Comments:

Are the Results Significant?

Yes, the results are significant and directly address my research question and hypothesis. Here's how:

- **Peak Hours:**
    - Shorter Trips: During peak hours (7–9 AM and 5–7 PM), trips are shorter in distance (2–4 miles on average) but longer in duration due to congestion.
    - Higher Fares: Surge pricing and increased demand lead to higher per-mile earnings and total driver pay per trip (up to $50/hour).
    - Trade-offs: Congestion reduces efficiency, limiting the number of trips drivers can complete.
- **Off-Peak Hours:**
    - Longer Trips: Off-peak trips are longer in distance but generate lower per-mile earnings.
    - Lower Fares: Without surge pricing, total driver pay per trip drops to $15–$20.
- **Hidden High-Earning Periods:**
    - Early Morning (4–6 AM): A hidden high-earning window with $25–$35/per trip earnings due to low competition and niche demand (e.g., airport trips).

Can This Analysis Inform Policymaking?

Yes, this analysis provides actionable insights for policymakers to improve driver earnings and reduce congestion. Here's how:

a. Regulating Peak-Hour Driving

- Problem: Oversupply of drivers during peak hours dilutes individual earnings and worsens congestion.
- Solution:
    - Stagger Driver Shifts: Allocate time slots to prevent oversaturation.
    - Dynamic Trip Allocation: Use app-based systems to distribute rides more equitably among drivers.

b. Incentivizing Off-Peak Driving

- Problem: Off-peak hours are less lucrative, discouraging drivers from working during these times.
- Solution:
    - Bonuses or Higher Rates: Offer financial incentives for driving during underutilized hours (e.g., midnight–5 AM).
    - Reduced Commissions: Lower platform fees during off-peak hours to increase driver take-home pay.

c. Infrastructure & Traffic Management

- **Problem**: Single-passenger rides during peak hours increase congestion.
- **Solution**:
    - Promote **rideshare** (e.g., Uber share, Lyft Shared) by offering **discounts to passengers** and **higher earnings to drivers** for shared rides.
    - Use **dynamic pricing** to make rideshare more attractive during peak hours.

- **Data Insight**: The data shows that **shorter trips** dominate peak hours, making carpooling a viable solution to reduce congestion.

What, if any, are your concerns with the data in hand? What other data could be useful for your analysis?

- **Concerns:**
  - **Lack of Earnings Breakdown:** The dataset includes driver pay but may not account for total costs (gas, maintenance, insurance), which affect actual earnings.
  - **Limited Context on Demand & Supply:** The dataset lacks real-time driver availability and rejected trip requests, which would clarify market saturation.
  - **Trip Purpose & Passenger Behavior:** There is no insight into passenger demographics or trip purposes, which could help explain tipping behavior and ride patterns.
- **Additional Useful Data**:
  - **Traffic Congestion Data:** Integrating real-time NYC traffic data would help quantify congestion impact more precisely.
  - **Driver Shift Data:** Understanding how long drivers work and when they start/stop could help assess overall profitability.

If you had more time, what would you have added to your analysis?

If I had more time, I would analyze each driver's total trips by day and hour to better understand how various factors impact earnings. In this analysis, I focused only on driver pay per trip rather than examining combined trips, which could provide deeper insights into overall earning patterns.