Taxicab Project

Jack Dzialo, Tayten Bennetsen, Rith Katikaneni, Griffin Ball

2024-12-05

Introduction

Urban mobility plays a critical role in shaping the daily lives of city dwellers, influencing everything from commute times to economic activity. As cities continue to grow, understanding the dynamics of transportation systems becomes increasingly important for improving efficiency, planning infrastructure, and enhancing the passenger experience. One key area of urban mobility is the taxicab industry, which has undergone significant changes over the years due to technological advancements, evolving passenger expectations, and the rise of ride-sharing services.

The traditional taxicab service, once the primary means of paid transportation in cities, now competes with app-based ride-sharing platforms like Uber and Lyft. This shift has led to increased competition, making it essential for traditional taxicab services to optimize their operations and improve customer satisfaction. To stay relevant in the rapidly changing landscape, taxicab companies must understand key factors such as fare pricing, trip characteristics, and passenger behavior. This knowledge is vital for adjusting pricing strategies, improving service quality, and identifying areas for growth.

This report aims to analyze a taxicab dataset to uncover insights about urban travel patterns, fare pricing, and tipping behavior. By examining the relationship between various factors—including trip distance, duration, payment type, passenger count, and location—the analysis seeks to answer the following key questions:

- How do fare amounts vary with trip distance and duration?
- What factors influence tipping behavior, and how do payment types affect the likelihood and amount of tips?
- Are there noticeable differences in fares and tips based on the pickup location?
- Does the number of passengers affect the cost of a trip or the tipping behavior?

To address these questions, the dataset will be explored using data cleaning techniques, descriptive statistics, and data visualization. The analysis will provide insights that can help taxicab companies make informed decisions, allowing them to better compete with ride-sharing services and cater to passenger needs. Furthermore, the findings could inform city planners and policymakers seeking to optimize transportation networks and improve urban mobility.

The dataset analyzed in this report includes trip-level details such as pickup and dropoff times, trip distance, fare amounts, tip amounts, passenger count, payment type, and location information. By delving into these aspects, the report will highlight patterns and trends that can be used to develop targeted strategies for optimizing pricing, encouraging tipping, and understanding travel behaviors across different areas of the city.

Data Description

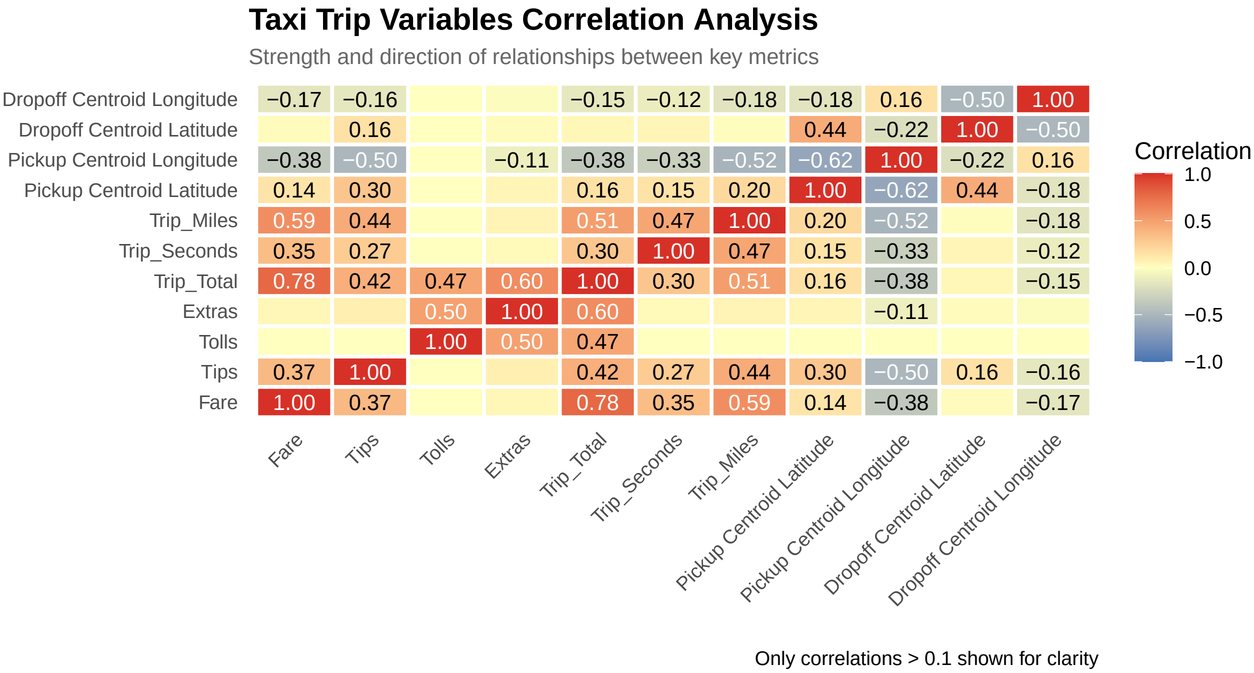
Our dataset contains relevant information on taxi trips in the city of Chicago, Illinois. It consists of two datasets with identical columns that have been combined—one with trips from 2013-2023 and another with trips from 2024. We compressed the data to include one full year of rides. In total, we have information on approximately six million rides that took place from October 1, 2023, to September 30, 2024. While not all taxi trips were reported, the city believes the dataset captures most of them, and we consider this sample size to be sufficiently large.

Each taxi has been assigned a unique taxi ID, allowing us to track individual rides for each vehicle. The taxi company under which each taxi operates has also been recorded. The start and end times of rides have been rounded to the nearest 15 minutes. The community area of each drop-off and pick-up location is included, making it possible to visualize the trips on a map of Chicago. Additionally, detailed cost information for each trip is provided, including tip amount, total fare, toll fees, and payment type. Overall, we were somewhat surprised by how long and expensive a few of the taxi trips turned out to be.

Plots

Before diving into specific factors that affect the rider experience, it's essential to identify the relationships between key variables. For example, do trip distance and fare amount exhibit a strong correlation, and if so, does this relationship hold consistently across different times of day or locations? We investigate these relationships by using a correlation heatmap.

Plot 1 - Heatmap


Taxi Trip Variables Correlation Analysis

Strength and direction of relationships between key metrics

Only correlations > 0.1 shown for clarity

The correlation heatmap reveals key relationships that shape the rider experience, providing a foundation for understanding which factors matter most in delivering value. For example, the strong positive correlation between `Fare` and `Trip_Total` suggests that fare transparency is crucial for a seamless rider experience, as passengers often equate higher fares with higher total costs. Similarly, the positive correlation between `Trip_Miles` and `Fare` shows that longer trips tend to result in higher fares, aligning with expectations for distance-based pricing models. However, the weaker correlation between `Trip_Seconds` (duration) and fare suggests that fare pricing may not account as heavily for time as for distance. Addressing these dynamics can help taxicab companies create more consistent pricing models, balancing time and distance to avoid penalizing riders during delays and ensuring transparent cost calculations.

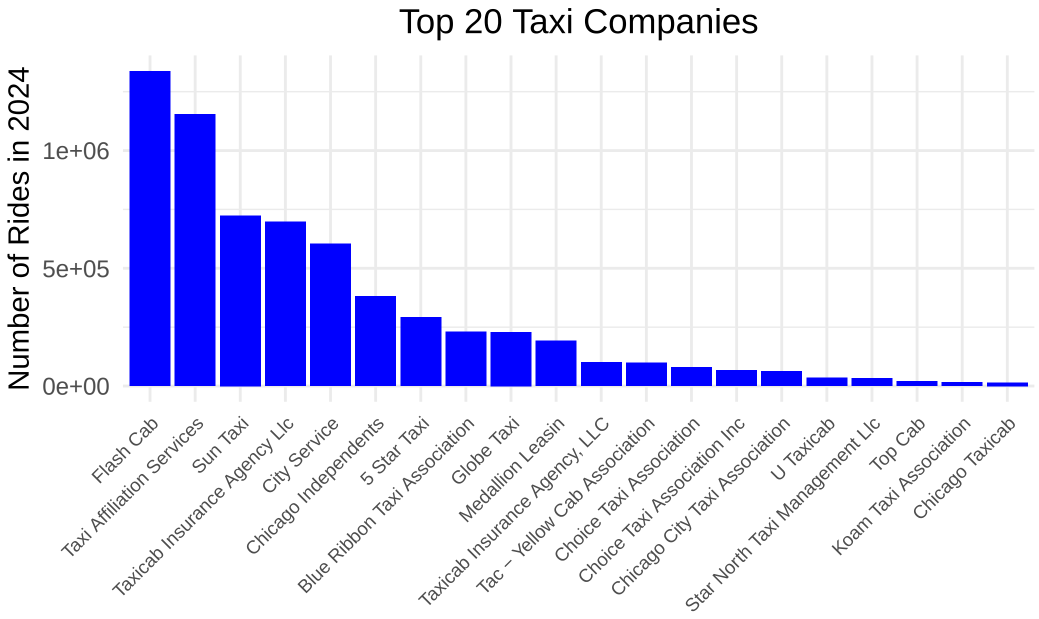
Additionally, the relationship between `Trip_Miles` and `Tip_Total` reveals intriguing insights into passenger tipping behavior. A noticeable correlation indicates that as the distance of a trip increases, so does the total tip amount. This may reflect a passenger's appreciation for longer services, increased driver effort, and/or time commitment. Thus, emphasizing this dynamic can guide companies in structuring fare and tipping suggestions, potentially enhancing overall service satisfaction and driver earnings.

With these insights, we can compare companies to identify those that excel at meeting rider expectations based on these key variables. Companies that perform well may demonstrate best practices, such as maintaining a close alignment between fare and trip distance while offering competitive rates during peak times or implementing dynamic incentives to encourage tipping. Understanding which companies achieve the best balance between fare predictability, trip duration, and distance traveled allows passengers to make informed choices while guiding companies on how to enhance their services to stay competitive in today's evolving transportation landscape.

An intriguing correlation identified is the -0.62 relationship between Pickup Centroid Latitude and Longitude, indicating that as the latitude of a pickup location increases, positioning more northward, the longitude decreases, moving westward. This observation aligns with the geographical layout of the Chicago area, which resembles a diagonal parallelogram extending from the northwest to the southeast. Consequently, we have elected to investigate the relationship between Trip Fare and the trip's duration or distance, or both, due to the relatively strong correlation of these variables with fare, aiming to assess the consistency of pricing across trips.

Plot 2 - Box Plot

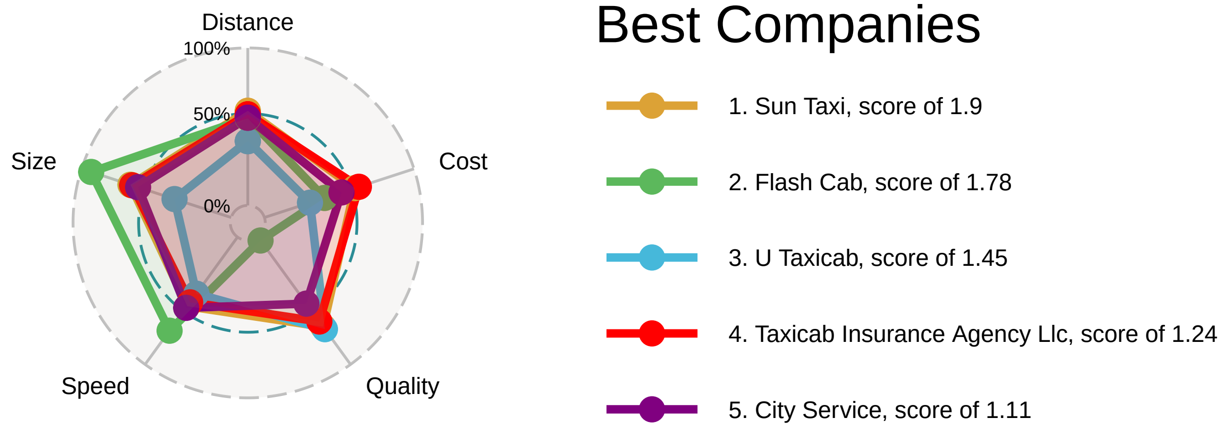
Before exploring the relationship between fare rate and time/distance, we wanted figure out the top 3 most popular taxi companies and then run an analysis on those companies, so that any insights made would apply to the most relevant companies, without any noise from other smaller companies. To do this, we decided to visualize each company's amount of rides in the year using a box plot.

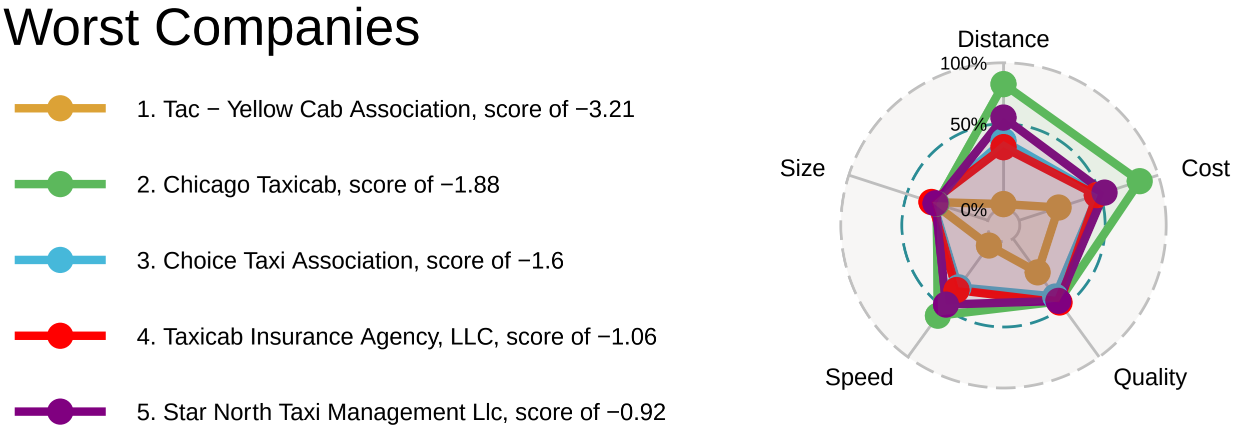


Chicago's taxi ecosystem in 2024 reveals a hierarchical structure dominated by key players, with Flash Cab leading at nearly 1 million rides and Taxi Affiliation Services following at 750,000 rides. For the average user, this concentration of service providers requires a strategic approach: maintain primary accounts with the top three providers (Flash Cab, Taxi Affiliation Services, and Sun Taxi) while keeping smaller companies as backups. This dual-provider strategy optimizes for both reliability and availability. Set up mobile apps for the major companies, as they typically offer shorter wait times and loyalty programs. However, don't discount smaller operators entirely – they often provide better rates during off-peak hours and can be crucial backups during high-demand periods. Consider using aggregator apps that connect to multiple services, effectively creating a personal dispatch system that maximizes your chances of finding a ride when needed.

Plot 3 - Spider Plot

We also wanted to see what companies had the best qualities that rides valued, so we decided to visualize each company's characteristics using a spider plot.





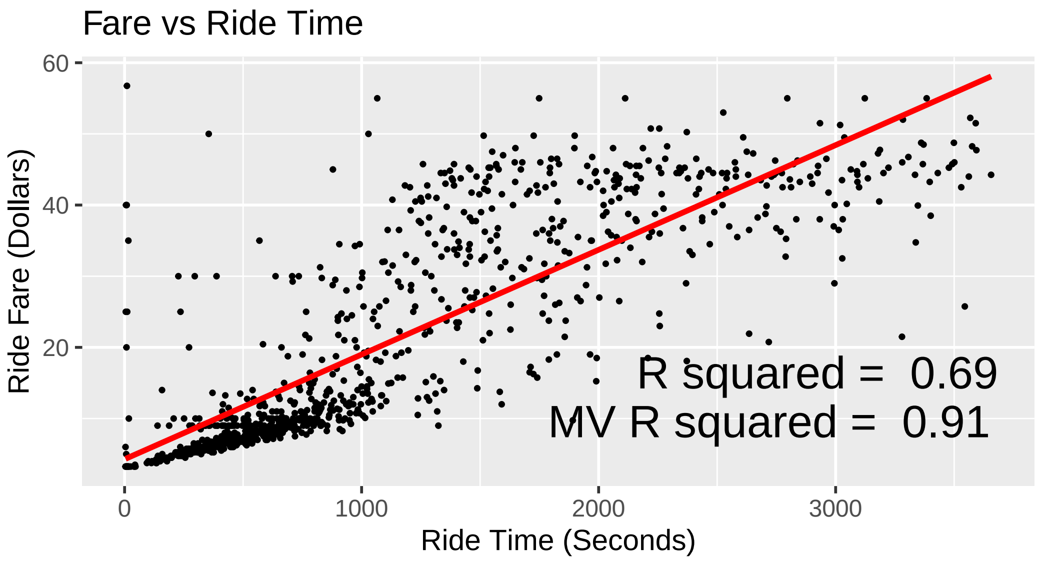
The attributes analyzed for each company include the average ride distance in miles, average ride cost in dollars, ride quality, average speed, and company size. Ride quality is computed as the average tip amount divided by the fare amount, indicating customer satisfaction. Average speed is calculated by dividing ride distance by ride time. Company size is evaluated by the total number of rides provided, reflecting driver activity or availability. The data is normalized using Z-Scores for consistent comparison across all attributes.

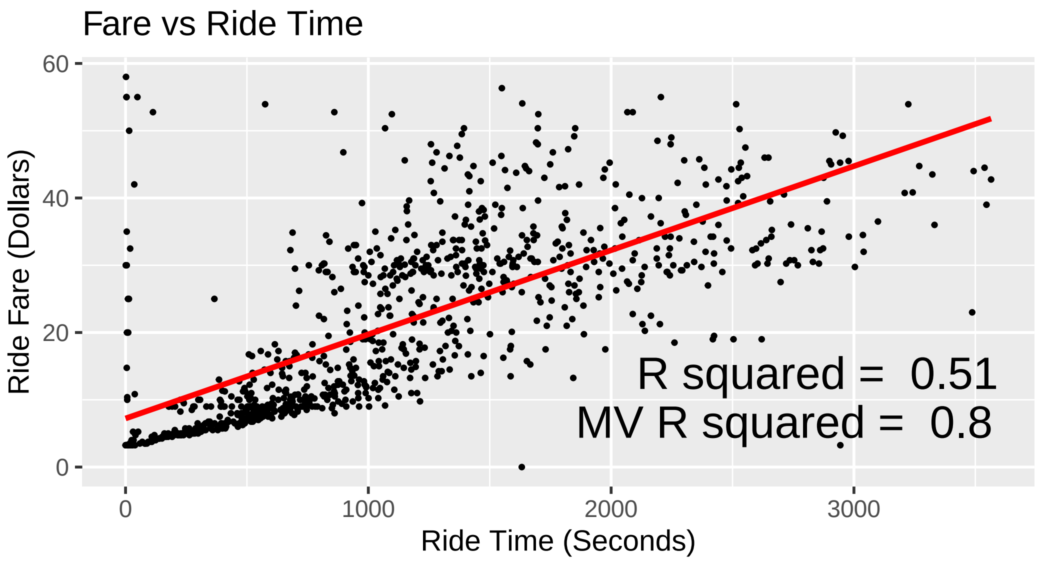
Plots are scaled from Z = -5 to Z = 5, covering nearly the entire value range, as values beyond this are negligible. The median line, labeled 50%, represents Z = 0, or the mean. The overall score for each company prioritizes speed, quality, and size, while minimizing cost, as these are key factors when selecting a taxi service. Ride distance is excluded as it doesn't impact perceived ride quality, being a customer choice.

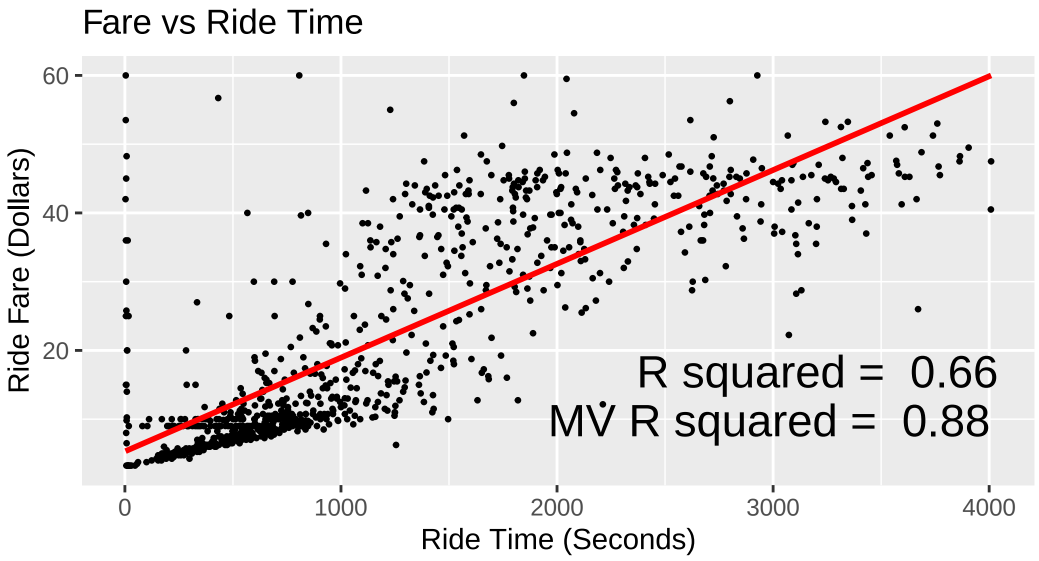
High-scoring companies generally exhibit slightly below-average cost and quality, and average speed and size. Notably, Flash Cab differentiates itself with high size, high speed, and low cost, targeting consumers who prioritize affordable, fast rides, even at the expense of quality. Conversely, companies with the lowest scores tend to have inferior attributes and average costs.

Plot 4 - Scatterplot and Analysis

After considering the last two plots, we decided to analyze Sun Taxi, Flash Cab, and Taxicab Insurance Agency due to their high positive characteristics and large size. For each company, we decided to run a regression analysis on Trip Fare vs Time due to the fact that most taxis are metered.


Fare vs Ride Time
Ride Fare (Dollars)
60
40
20
0
1000
2000
3000
Ride Time (Seconds)
R squared = 0.69
MV R squared = 0.91


Fare vs Ride Time
Ride Fare (Dollars)
60
40
20
0
0
1000
2000
3000
Ride Time (Seconds)
R squared = 0.51
MV R squared = 0.8

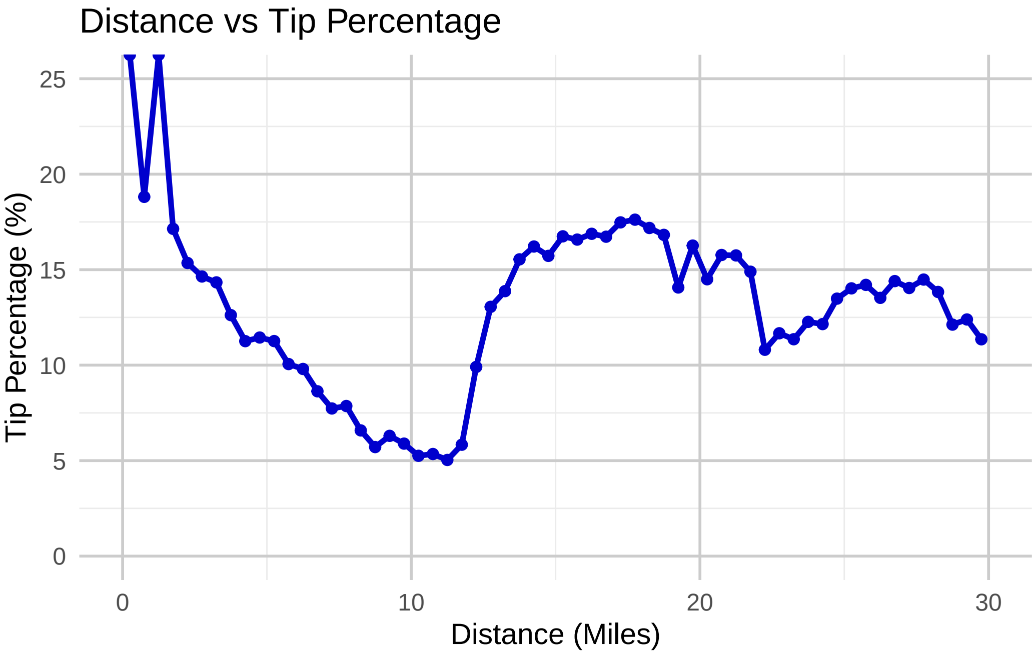

Fare vs Ride Time
Ride Fare (Dollars)
60
40
20
0
1000
2000
3000
4000
Ride Time (Seconds)
R squared = 0.66
MV R squared = 0.88

Since the p-values of the estimators of the regular model are so low (at least lower than 0.025), we can reject the null hypothesis that the estimators are equal to 0 at a significance level of a = 0.05. This signifies a strong connection between length of the ride with price of the ride, which makes sense, since most taxis use a meter and charge by the minute. The addition of distance into the model, making it multivariate leads to a higher R-Squared, signifying that it improves the model accuracy. We came up with the idea that because there may be a set fee for small distances, that it may be affecting the model that is solely based off time of the ride, and that incorporating distance could help account for this.

Another thing to note is that there is a dense collection of values that seem to follow a line for times under 1000 seconds, while the distribution of prices past that time is more random. We hypothesized that this is because higher times are more rare, and are prone to different prices due to the drivers pricing longer rides differently. The only concern with this plot is the outlier points with high fare prices but a zero second trip time. These are likely due to data collection errors or other anomalous circumstances.

Plot 5 - Line Plot

We also decided to investigate the relationship between trip distance and tip percentage, due to the correlation between `Trip Miles` and `Tips`.

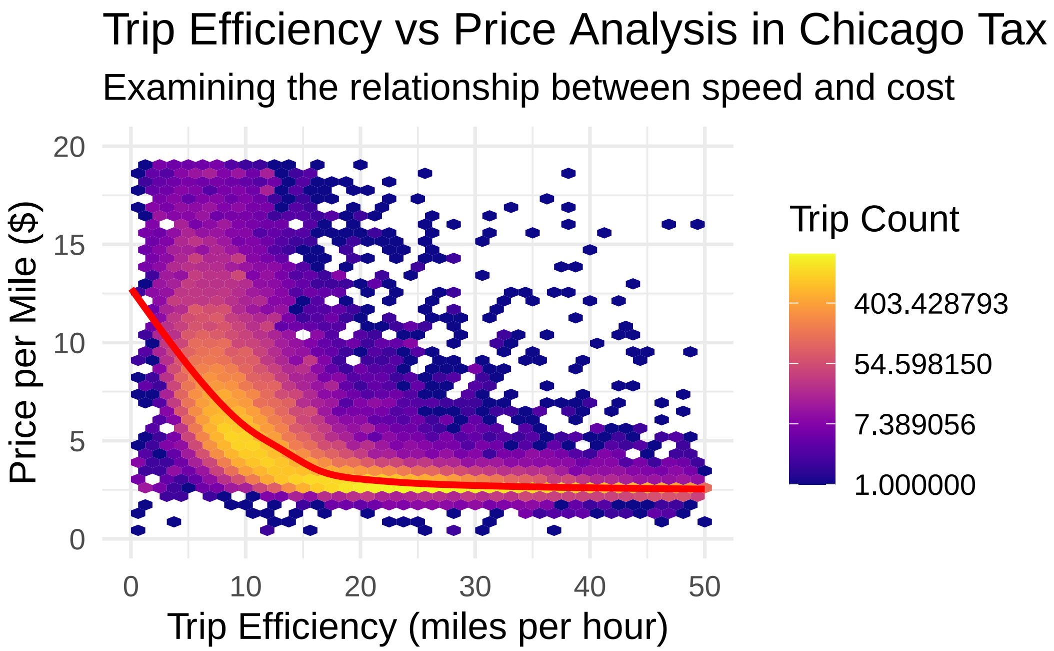


The relationship between trip distance and tipping behavior uncovers critical insights for cost-effective taxi usage. Short trips (under 10 miles) show higher percentage tips but lower absolute fares, making them ideal for downtown and inner-city travel. Medium-length journeys (11-15 miles) represent the sweet spot for both drivers and passengers, with optimal price-to-distance ratios and higher driver satisfaction through better tips. Beyond 15 miles, trips see declining tip percentages but higher absolute fares, suggesting the need for different strategies on longer journeys. For optimal results, consider scheduling longer trips during off-peak hours and negotiate flat rates in advance. For regular commutes, track your common routes and times to identify patterns in pricing and availability. Keep multiple payment methods available, and consider using company-specific apps that might offer lower rates or loyalty rewards for frequent users.

Plot 6

Hexagonal Plot

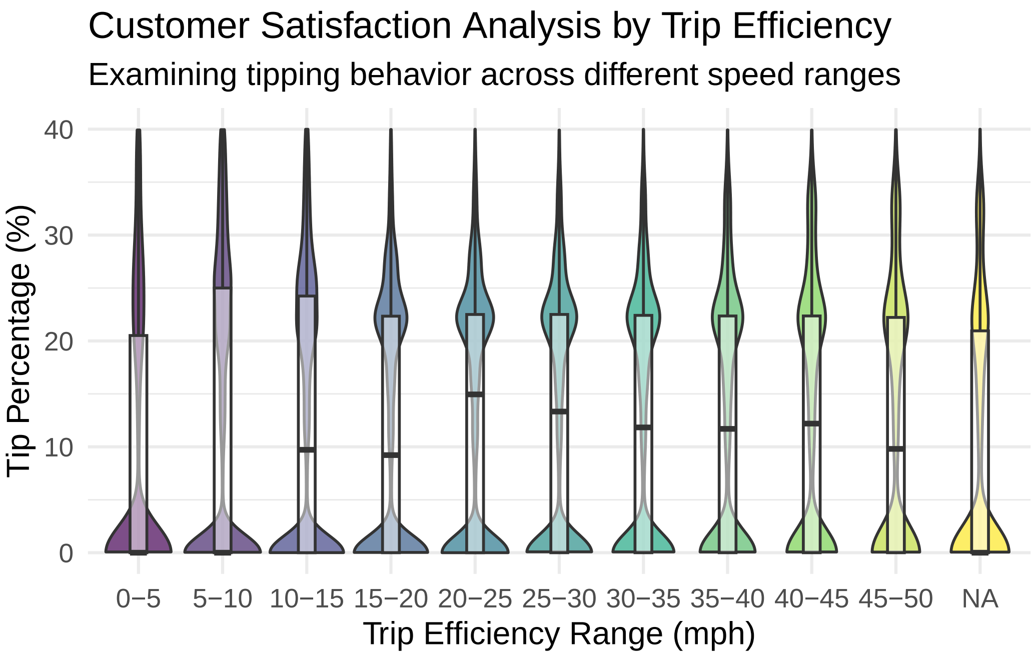
Trip Efficiency and Customer Value



The efficiency versus price analysis paints a compelling picture of how service speed affects customer perception of value. Most satisfying experiences cluster in what we might call the "sweet spot" - trips averaging 15-25 mph. This speed range represents an optimal balance between progress and safety, where customers feel they're getting good value for their money. When speeds drop below 10 mph, we see a concerning trend of higher prices per mile, likely reflecting frustrated passengers stuck in traffic or dealing with urban congestion. The hexagonal heatmap reveals that the majority of trips fall within this optimal range, suggesting that Chicago's taxi drivers have developed an intuitive understanding of this balance. However, the outliers tell their own story - very slow trips often correlate with lower customer satisfaction, while very high-speed trips show increased variability in both pricing and customer response.

Plot 7

Violin Plot



Perhaps the most revealing indicator of customer satisfaction comes from our analysis of tipping behavior. The violin plots across different speed ranges tell an eloquent story about passenger appreciation. Trips in the 15-25 mph range consistently earn higher tip percentages, suggesting that customers recognize and reward what they perceive as “good service.” This sweet spot represents a balance where passengers feel they’re making steady progress without experiencing the stress of either crawling traffic or excessive speed.

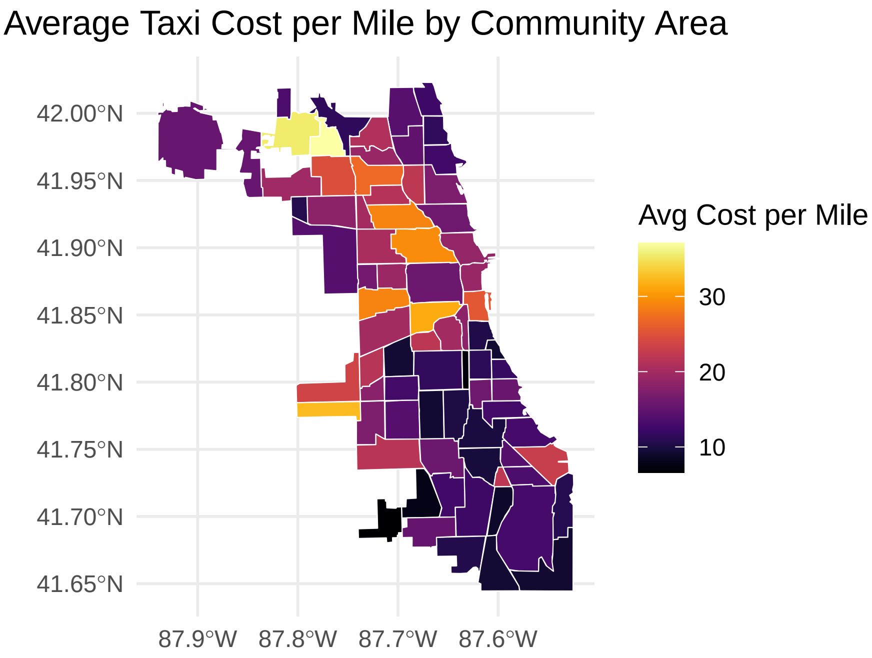
The wider spread of tip percentages during rush hours tells us about varying customer expectations during high-stress periods. Some passengers might be more generous, acknowledging the challenging conditions, while others express their frustration through reduced tips. This pattern suggests an opportunity for drivers to manage expectations better during peak times.

Plots 8, 9, 10, 11, 12, 13

Finally, we look at how specific situational factors impact the quality of a ride. The effects of location, time of day, and trip distance can all vary widely; for instance, trips from downtown locations might be more expensive due to higher demand, or certain times of day might see longer travel times due to traffic congestion.

Exploring these factors provides a nuanced view of the rider experience, allowing companies to implement dynamic pricing strategies or targeted incentives that address the unique challenges associated with different locations and times. This understanding ensures that passengers receive consistent value no matter when or where they travel.

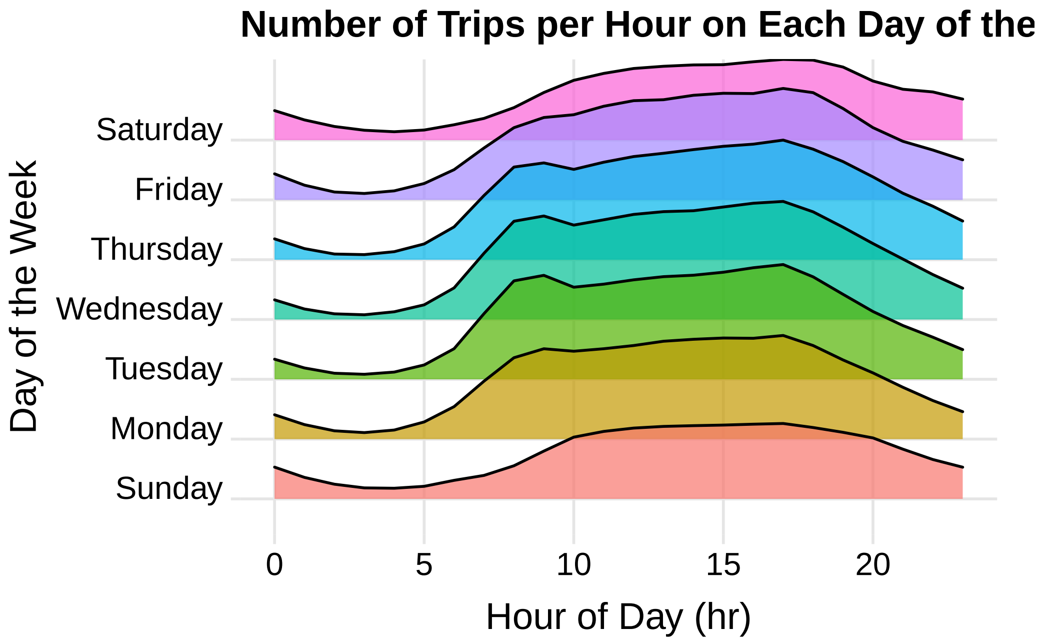
Geo Plot



Chicago's taxi pricing landscape shows distinct geographical patterns, with higher costs per mile in affluent neighborhoods and around major transit hubs like O'Hare and Midway airports. This spatial pricing variation demands strategic planning: for airport runs, consider booking in advance and asking for flat rates to avoid premium pricing. When traveling through high-cost areas (typically downtown and affluent neighborhoods), look for opportunities to combine trips or share rides. Lower-cost residential areas offer better value, especially for regular commuting routes. Develop a mental map of your most frequent destinations and their typical pricing tiers. Consider establishing relationships with specific drivers or companies that regularly service your common routes, as they may offer more consistent pricing. During peak hours in high-cost zones, compare prices across multiple services and consider alternative transportation options if available.

Ridgeline Plot

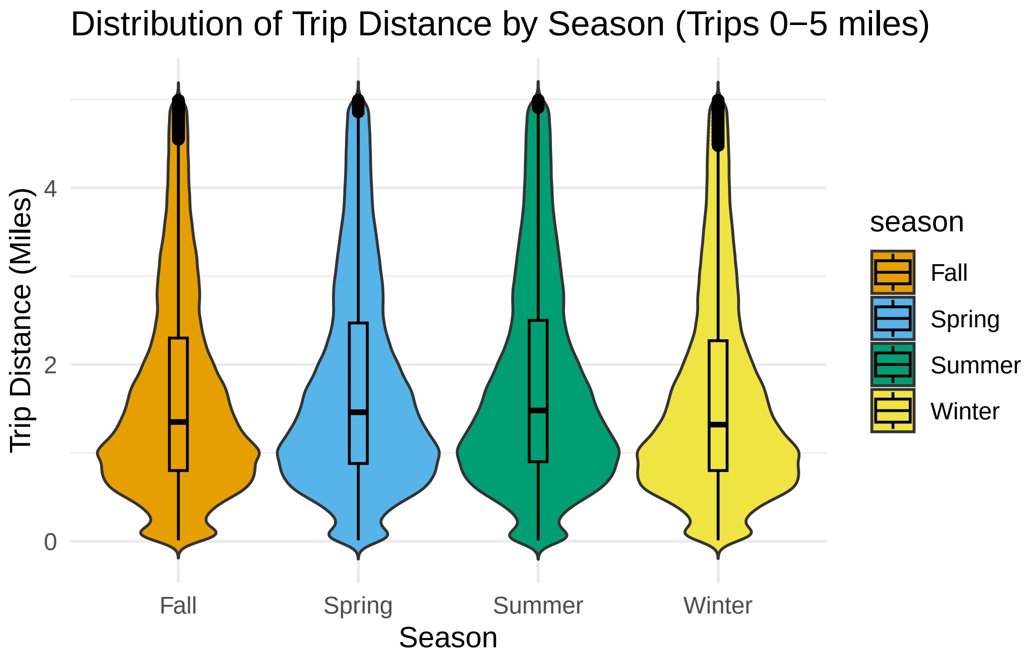
This plot shows the relationship between days of the week and frequency of trips.



This plot shows us that during the weekdays, there is a large spike in the mornings for the amount of trips that occur. I believe that this is due to the fact that many people go to work early in the morning and thus need taxis.

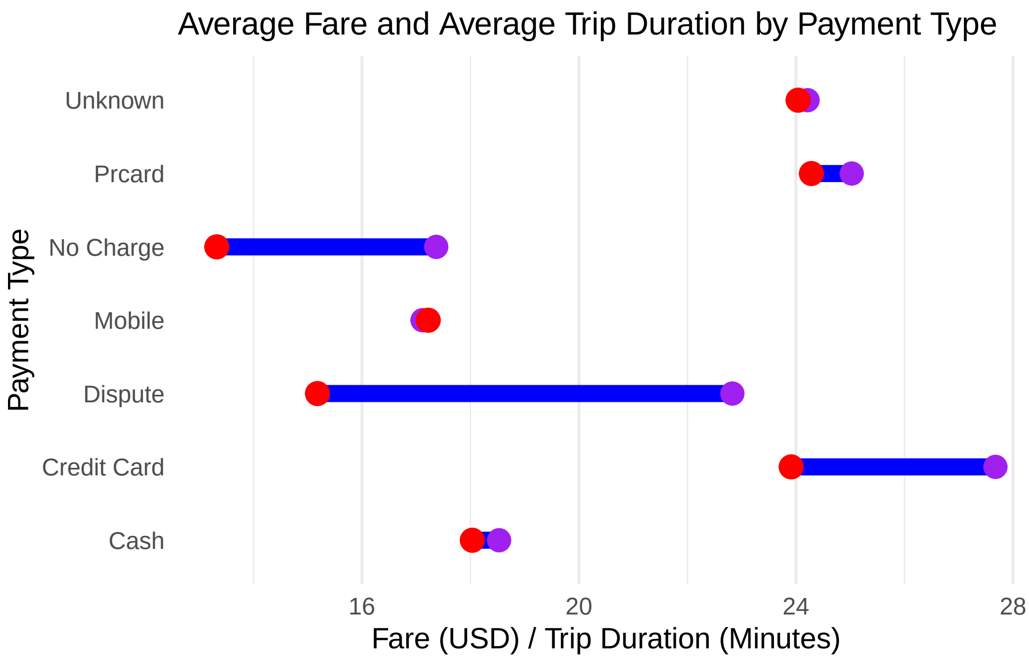
Violin Plot

This plot shows the relationship between trip distance and season of the year.



The seasons do not seem to have a significant impact on the trip distance, except for winter having a very slight increase in lower distance (0.1 - 1 mile) trips.

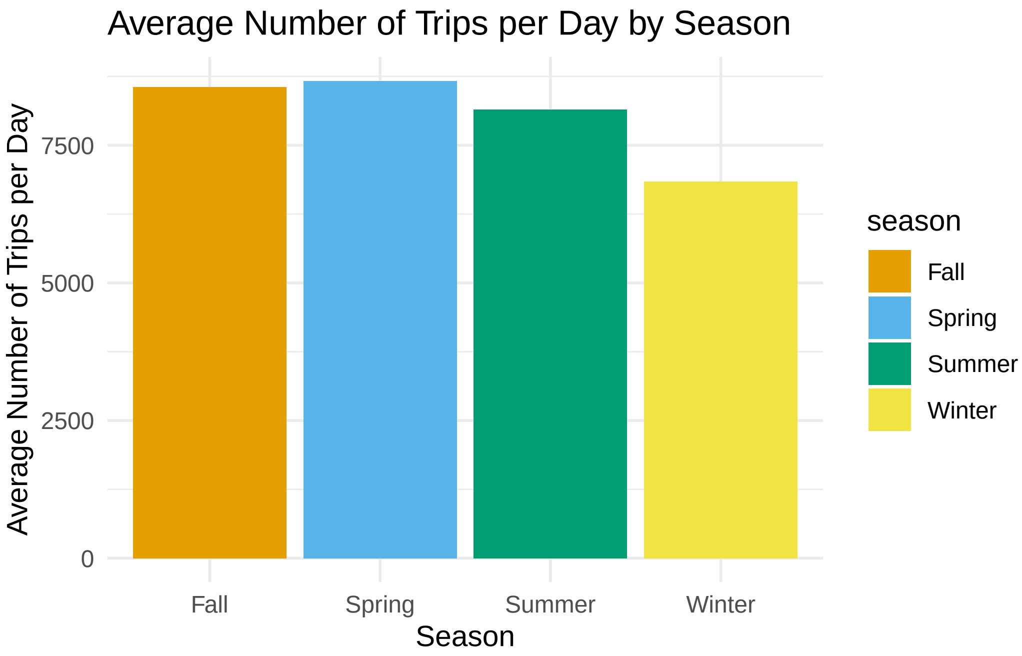
Dumbell Plot



This dumbbell plot visualizes the comparison between the average fare and average trip duration across various payment types in the dataset. The red dot represents the average fare, while the purple dot represents the average trip duration (in minutes). For payment types like No Charge and Dispute, the fare and trip duration are relatively similar, indicated by the shorter blue lines between the dots, suggesting that trips paid with these methods tend to have consistent costs and durations. In contrast, Unknown and Pcard show a larger gap between fare and trip duration, suggesting that trips using these payment methods are typically longer and more expensive. Notably, Mobile payments show relatively lower fares and trip durations, with the dots closely placed, indicating shorter trips. This plot offers valuable insights into how different payment methods correlate with trip costs and durations, potentially reflecting varying customer behaviors or trip purposes.

Box Plots

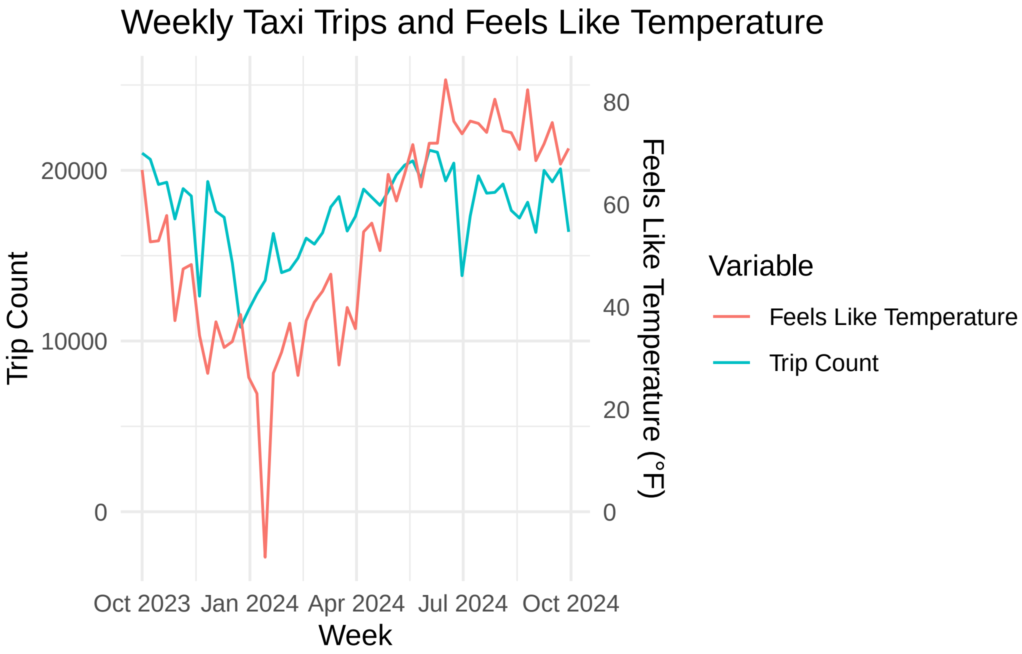
This plot shows the average number of trips per season.



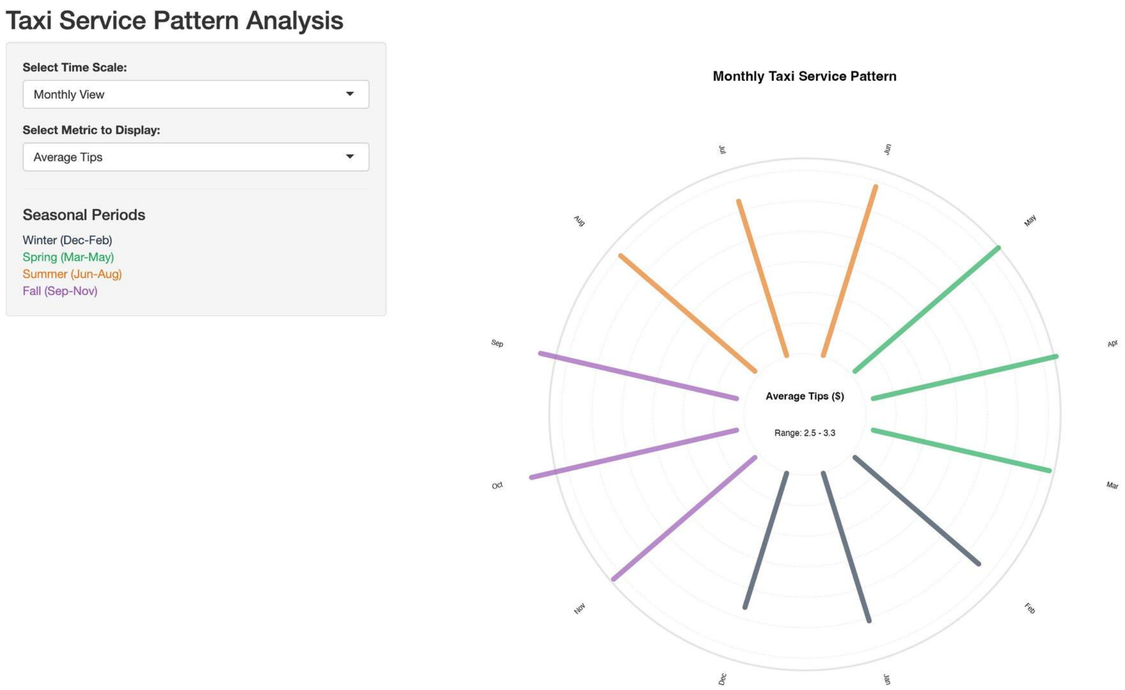
It seems as if Spring is the season with the most taxi trips, presumably due to the good weather, which leads to people wanting to travel around the city. Winter is the season with the least taxi trips, probably due to bad weather leading to people wanting to travel less.

Double Line Plot

This plot aims to show the how the temperature can affect the amount of taxi trips for any given week. The blue line is the number of trips, calculated on a weekly basis, and the red line is the average feels like temperature for that week. There is a direct correlation between temperature and trip count, meaning as temperature drops there are less and less taxi rides being taken. However, even in extremely cold conditions trip count does not drop below 10,000, suggesting that some trips must be taken regardless of weather condition.



Killer Plot Example



Taxi Service Pattern Analysis

Select Time Scale:

Monthly View

Select Metric to Display:

Average Tips

Seasonal Periods

Winter (Dec-Feb)
Spring (Mar-May)
Summer (Jun-Aug)
Fall (Sep-Nov)

This is an example of what is possible with our killer plot. The user can specify to view by month or hour of the day and depending on that input each line in the circle will represent an hour or month. The length of each line represents the metric chosen by the user, either trip count, trip miles, tips, or fare. In this example we have to chosen to show the average tips for each month of the year and it is clear that average tip peaks in September-November then falls off sharply in the winter months, possibly reflecting the slightly shorter rides in those months. The customization options allow users to easily gain different insights based on what they want to see and try many different combinations of metrics.