Ride Sharing and Tipping Behavior in Chicago 2018-2023

CS590 Data Science Project Report

Muhammad Imam Adli (ma510)
Alex Chao (ac330)
Layan Kaushik (ly220)
Minghui Zhu (mz223)

April 24, 2023

Abstract

Using ride-sharing data in Chicago from November 2018 to January 2023 supplemented with weather data and geographical community data, we explore three different stories surrounding ride-sharing and tipping behavior: the effects of the COVID pandemic, differences between income groups, and the effects of weather.

As it feels like people's general lifestyles have reverted back to very similar to prior to the pandemic, we expect that both the number of rides and tipping behavior will also have reverted. Instead, ride count has only returned to about $\frac{2}{3}$ of what it used to be, and the tipping percentage has actually increased. We are unsure if this is a cultural shift after the pandemic or if the numbers are still on the, but our regression and t-tests both allow us to conclude that these observations are statistically significant.

Across different income groups, the initial ride-sharing behavior is consistent which we expect due to the complete government shutdown. However, as restrictions started to loosen, the lowest income group had a sharper increase in rides when compared to different groups. This could be due to less flexibility in transportation or work options due to less access to resources. This shows that while on the surface it may seem like a blanket policy affects everyone equally, even this specific analysis of ride-sharing behavior highlights how socioeconomic status can affect people's lifestyles completely differently.

We expect bad weather will lead to an increase in both demand and tips for ride-sharing as people will be less likely to use alternative forms of transportation, and they will be appreciative of pickups during inclement weather. Instead, we found that while the demand increased significantly, there was almost no increase in tips. This may be due to the already increased fare lowering the willingness to tip or the shorter ride distance resulting in less opportunity for having an experience conducive to a higher tip. This information is helpful for drivers to weigh the tradeoffs of whether it's worth the slight increase in wages to drive in poor conditions.

1 Context: Brief Introduction

In recent years, the use of ride-sharing services such as Uber and Lyft has become increasingly popular, transforming the way people commute and navigate cities. As a result, the analysis of ride-sharing data has become an essential tool for policymakers, transportation planners, drivers, and business owners to make informed decisions about urban mobility.

The city of Chicago has been a hub for ride-sharing services, with millions of trips taken every year. By analyzing the data collected by ride-sharing companies operating in Chicago from 2018 to 2023, researchers and analysts can gain valuable insights into travel patterns, traffic congestion, and user behavior.

Such data can be used to inform transportation policies, including the allocation of resources to improve infrastructure such as roads, bridges, tunnels, and public transportation systems, and reduce traffic congestion. Additionally, businesses in the transportation and hospitality sectors can use this data to make strategic decisions about pricing, service offerings, and marketing strategies.

For drivers, the analysis of ride-sharing data can provide valuable insights into the demand for their services, optimal pick-up and drop-off locations, and pricing strategies. This information can help drivers maximize their earnings and improve their overall experience on the platform.

The target audience for this data analysis includes policymakers, urban planners, transportation officials, and business owners in Chicago and beyond. These stakeholders have a vested interest in understanding the dynamics of ride-sharing in the city, and how it affects urban mobility and economic growth. Therefore, the findings of this analysis could provide critical insights into improving transportation systems and fostering economic development in Chicago and other urban centers.

2 Data

2.1 Data Sources

We obtained ride-sharing data from the City of Chicago's website, comprising 300 million trips between November 2018 and January 2023. The data was publicly available from the Chicago city website, as the city requires the ride-sharing company to submit all the trip information anonymously. The data contain all relevant information about the trip information such as pick-up coordinates, drop-off coordinates, fare, tip, additional charge, duration of the trip, and trip distance.

Additionally, we collected hourly weather data using the Open-Metro Historical Weather API, and community area-level income data from CMAP's Community Data Snapshots (CDS). The datasets are introduced in the data merging subsection.

2.2 Data Descriptive Statistics

Below are the descriptive statistics for some of the main variables we use in this analysis. The descriptive statistics include trip information such as duration, distance, fare, and tip. Additionally, by combining Chicago weather and location data, we also obtained the weather condition during the trip and the neighborhood average income level from where the trip originated.

Variable	Description	Mean	Std. Dev.
Trip.Seconds	Duration of trip (seconds)	1069.147	760.77
Trip.Miles	Length of trip (miles)	6.539	7.308
Fare	Fare of the trip ($)	14.515	12.059
Tip	Tip given ($)	0.814	2.204
Additional.Charges	Additional charges ($)	3.626	3.088
Trip.Total	Total trip cost	18.954	14.529
Pickup.Centroid.Latitude	Pickup coordinate, latitude	41.89	0.066
Pickup.Centroid.Longitude	Pickup coordinate, longitude	-87.669	0.065
Dropoff.Centroid.Latitude	Dropoff coordinate, latitude	41.89	0.066
Dropoff.Centroid.Longitude	Dropoff coordinate, longitude	-87.671	0.069
Hour	Time, hour of the trip	10.723	5.728
precipitation..mm.	Precipitation during trip (mm)	0.129	0.556
rain..mm.	Rain during trip (mm)	0.115	0.545
snowfall..cm.	Snowfall during trip (cm)	0.01	0.077
PopPickup	Population at Pickup Community Area	62739.48	30538.12
IncomePickup	Income at Pickup Community Area	61223.54	30136
IncomeDropoff	Income at Dropoff Community Area	61376.15	29992.7
Tip_percent_PerRide	Tip as percentage of Fare	4.219	10.389

Table 0: Data Descriptive Statistics

2.3 Data Sampling

The ride-sharing data was over 76 gigabytes, so we used random sampling to obtain 3 million observations from the 300 million. The sampling method is done because the original 300 million original observation is too big to be loaded into a single pandas data frame, and 3 million datasets from 2018 - 2023 correspond to around 2.000 rides per day should suffice to draw a statistically significant conclusion.

One challenge we encounter during data sampling is how to random sample the 300 million data that cannot be loaded in standard Python packages and take too much time for standard Python software. We solve this by using chunks and loading 1 million datasets at a time, then saving each 1 million observations as separate .csv files. We then loop through all the chunk .csv files (300 files) and randomly sample 10.000 comments from each chunk to obtain the 3 million sample dataset. While these methods can be improved further by using 1) a data streaming technique or 2) a greater sample size from each chunk, we believe this current approach is sufficient to obtain a clear conclusion from the dataset.

2.4 Data Cleaning

To clean up the data, we dropped all records with missing data (i.e. the value is NA) in the subset (TripSeconds, Tip, Trip Miles, Fare, TipPercentPerRide,Trip Total). We also removed outliers, which were likely faulty observations: records with tip percent per ride > 1000%, time taken for the trip equal to 0 seconds, length of the trip <= 0.1 miles, average speed > 0.04 miles per second (144 mph). This comprised around 0.3% of total observations. To aid with our analysis, we added some new variables:

- TipPercentPerRide = (Tip/(Fare+additional charges)) * 100
- Tipped (boolean) = Tip > 0
- Weekday (boolean)
- Traffic Condition/Average Speed = trip miles/ trip seconds (miles per second)
- TipPercentInterval: bins=[-1, 0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, max(df[TipPercentPerRide])]
- Traffic Condition Interval2: Using Traffic Condition variable
  bins = [0, 0.0025, 0.005, 0.0085, 0.0140, max(df[Traffic Condition])]
  labels = ["Very high traffic", "high traffic", "high-avg traffic", "avg traffic", "light-no traffic"].

2.5 Data Merging

To acquire a thorough picture of transportation patterns, it is necessary to consider aspects other than simple trip information. Merging the transportation trips dataset with other datasets, such as community-level socioeconomic or weather data, can provide significant insights into how different factors may affect transportation patterns.

First, we combine Chicago trip data with fact data at the community-area level. We investigate the association between mobility patterns and socioeconomic characteristics using community area-level income data from CMAP's Community Data Snapshots (CDS).

At the community area level, the CDS data includes precise information on populations, household incomes, employment, and other important aspects. We combine this data with the transportation trips dataset based on pickup and dropoff locations. We use the average of the pickup and dropoff community areas' income as a proxy for the riders' income.[1]

Second, we incorporate the weather attributes since the weather is a crucial factor influencing travel behavior. We collect hourly weather data using the Open-Metro Historical Weather API, covering the timeline of the Chicago trips dataset. The hourly weather data includes temperature, precipitation (rain, snowfall), wind speed, and direction. We use each trip's starting timestamp to match it with the corresponding hourly weather data.[2] Since there is no missing data in the hourly weather data, we can be confident in the accuracy of our analysis. After merging the two datasets, we can conduct exploratory analysis and build statistical models to identify how weather conditions impact transportation patterns in Chicago. For example, we can examine how precipitation levels or temperature changes impact the demand for transportation services or the tipping levels.

3 Target Effect and Explanation

3.1 Story 1: From Crisis to Recovery: The Impact of COVID-19 on Economic Decline and Resurgence

We are interested in how the trips and tips change during and after the pandemic outbreak as it can help us understand how the industry was affected by the unprecedented circumstances. We would expect a decline in both the number of trips and tips during the pandemic, with a subsequent recovery after the pandemic. Furthermore, we would anticipate that these patterns would be in alignment, as they are both measures of demand for ride-sharing services.

We draw two graphs of the 7-day rolling average for the number of trips (Graph 1) and the tipping percentage in the fare (Graph 2), respectively. We use a 7 days rolling average due to very different trip numbers during weekdays and weekends, and hence a 7-day rolling average normalizes the days in a week variation.

[1] If one of the pickup and dropoff locations is missing, we use the existing one to represent the community area.

[2] We use the information for the beginning of the trip to simplify the matching process because there is no big difference between the starting point and ending point of the trip. Similarly, we fetch the data for the whole of Chicago rather than each pickup location since it's very time-inefficient.

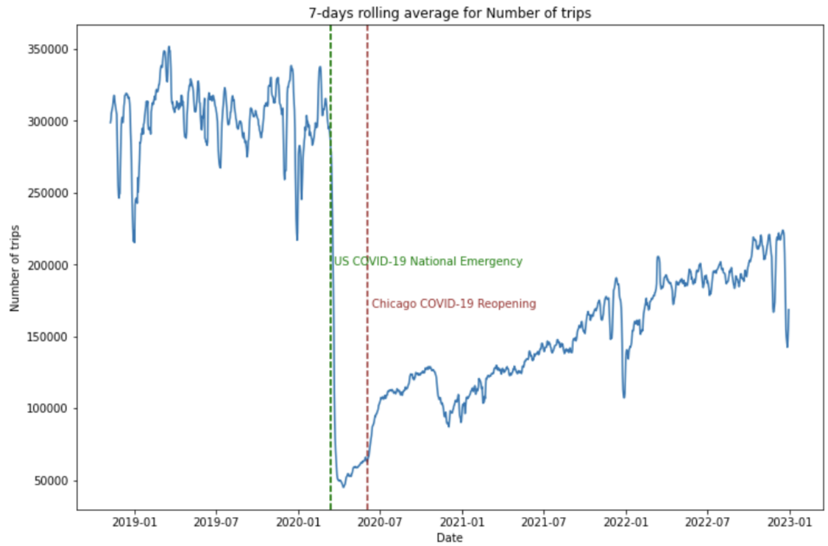


Figure 1: 7 days rolling average for the number of trips

It has been almost two years since the full reopening of Chicago, and one might expect that the ride-sharing numbers have bounced back towards or even beyond pre-pandemic levels already. In our daily lives, indeed, it feels like our habits and lifestyles have returned to the "old normal." It is even possible that people are more willing to go out to make up for the time lost in the past few years. However, our data suggests otherwise.

Firstly, we observed a substantial decline in the number of trips and tipping percentages during the pandemic outbreak. Prior to the pandemic, the average daily trips were approximately 300,000, with minor fluctuations, especially at Christmas and New Year. However, with the onset of the pandemic, the number of daily trips sharply decreased to around 60,000, representing a significant decline in demand for transportation services. Similarly, tipping percentages saw a sharp decrease, dropping from an average of 4% before the pandemic to 2.5% during the pandemic. This notable decrease in tipping percentage suggests that the pandemic's immediate impact on the transportation industry extended beyond a reduction in demand, with customers also changing their tipping behavior.

It is reassuring to observe a steady recovery in the number of trips and tipping percentages since 2021. Notably, tipping percentages have surpassed pre-pandemic levels, despite the number of trips still lagging behind. As of the end of 2022, daily trips have rebounded to about 200,000, representing approximately two-thirds of pre-pandemic levels. Meanwhile, tipping percentages have recovered and exceeded the pre-covid level, reaching an average of 5.5% at the end of 2022. Comparatively, the average level of the tip before the pandemic was around 4.7%. These observations in tandem are very surprising. We expected that the total ride-sharing numbers will be closer to pre-pandemic levels and that tipping percentages would be no more than pre-pandemic levels, but our

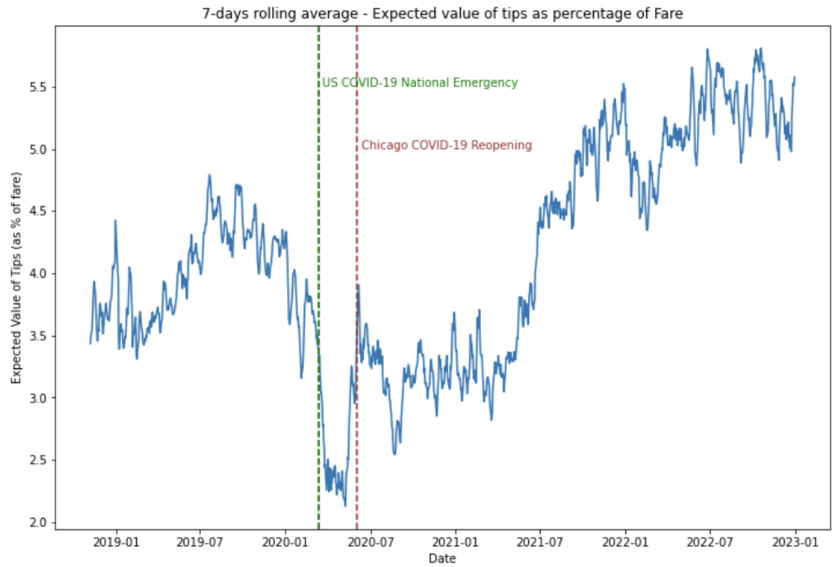


Figure 2: 7 days rolling average - expected value of tips as percentages of fare

initial analysis shows that both of these are false.

Technical Discussion

We conducted a series of regression analyses to determine whether the observed decrease and subsequent recovery (and even surpassing pre-covid levels for tipping percentages) were statistically significant. We used the tipping percentages and daily number of trips (adjusted for local community population) as dependent variables. We divided the data into three periods: pre-covid, peri-covid, and post-covid, and created dummy variables for each. Additionally, we included a time dummy variable called "recent" to account for the apparent significance of more recent periods. We added other independent variables to control for any potential confounding variables after reviewing the correlation matrix. We used an income fixed-effects model to assess whether differences were significant when applied to each income group separately to make the results more robust and persuasive.

Because our data was split into four income groups, and income was measured by location, we controlled for the impact of other location-based attributes of the location. For instance, we used the number of trips per capita (i.e., divided by the local population) as the Y variable to account for differences in population across locations. The results of our regression analyses are presented in Table 1.

Column (1) and (3) report the difference in tipping percentages and the number of trips per capita between pre-covid and peri-covid periods. The data shows that tipping

Table 1: Trend Changes in Tipping Percentage Per Ride/Daily Number of Trips Per Capita before and after Covid

	Dependent variable:			
	Tipping Percentage Per Ride		Daily Number of Trips Per Capita	
	(1)	(2)	(3)	(4)
Fare	0.025*	0.015	0.0001**	0.0001**
	(0.009)	(0.010)	(0.00003)	(0.00002)
Shared Trip Authorized	−1.841***	−1.870***	−0.001	−0.001
	(0.147)	(0.135)	(0.001)	(0.001)
Trips Pooled	−0.216**	−0.230***	−0.0001	−0.0001
	(0.045)	(0.038)	(0.0001)	(0.0001)
Hour	0.001	−0.012	−0.0001*	−0.0001*
	(0.011)	(0.016)	(0.00004)	(0.00003)
Temperature	0.017*	0.012	−0.00003*	−0.00002
	(0.006)	(0.005)	(0.00001)	(0.00001)
Relative Humidity	−0.005***	−0.005***	−0.00003**	−0.00003**
	(0.0002)	(0.001)	(0.00001)	(0.00001)
Rain	0.024	0.037	0.0003***	0.0003***
	(0.011)	(0.022)	(0.00003)	(0.00002)
Snowfall	0.265	0.182	0.001**	0.0001
	(0.145)	(0.134)	(0.0001)	(0.0002)
Traffic Condition	−42.931	−11.094	0.636	0.680
	(84.045)	(93.525)	(0.488)	(0.467)
Weekday	0.008	0.001	0.001*	0.0005*
	(0.015)	(0.022)	(0.0002)	(0.0002)
PeriCovid	−0.993**		−0.010***	
	(0.309)		(0.001)	
Recent		0.679		−0.007***
		(0.326)		(0.001)
Income Fixed effects	Y	Y	Y	Y
Observations	1,586,059	2,460,066	1,586,059	2,460,066
R^2	0.013	0.013	0.259	0.315
Adjusted R^2	0.013	0.013	0.259	0.315
Residual Std. Error	10.212 (df = 1586044)	10.492 (df = 2460051)	0.009 (df = 1586044)	0.009 (df = 2460051)

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

percentages and the number of trips per person have significantly decreased during the COVID-19 outbreak compared to before. The estimated effects are -99.3% and -1%, respectively.

Columns (2) and (4) compare the pre-COVID period to the recent period. The coefficient for column (4) is still negative, indicating that the number of trips per person is still lower than before COVID-19, by 0.7%. The coefficient for column (2) is positive but not significant, suggesting an average exceed of 67.9% in tipping percentages during the recent period compared to pre-COVID levels. However, this coefficient may be influenced by the fact that tipping percentages during most of the recent period were below pre-COVID levels. We can also use a T-test to compare the most recent time point to the one just before the COVID-19 outbreak.

A two-sample t-test was performed on the tipping percentages per ride between two periods, namely Dec 15, 2019, to Jan 15, 2020, and Dec 15, 2022, to Jan 15, 2023. The null hypothesis stated that the true difference in means between the two periods is zero, while the alternative hypothesis stated that the true difference in means is not equal to zero.

The mean tipping percentage per ride was 4.052305 in the pre-covid period and 5.219590 in the most recent period, with a t-statistic of -15.792 and a p-value less than 2.2e-16. Based on these results, the null hypothesis is rejected, and it can be concluded that the tipping percentage per ride has significantly increased beyond the pre-covid level.

Table 2: The Coefficients of Income Weather Attributes on Tipping Percentage Per Ride/Daily Number of Trips Per Capita

	Dependent variable:			
	Tip		Tipping Percentage Per Ride	
	All Trips	Tipped Trips	All Trips	Tipped Trips
	(1)	(2)	(3)	(4)
Hour	-0.007^{**}	-0.003	-0.009	0.102^{***}
	(0.002)	(0.001)	(0.011)	(0.015)
Temperature	0.003^{***}	0.001	0.026^{***}	-0.042^{***}
	(0.0005)	(0.001)	(0.002)	(0.008)
Trips Pooled	0.026^{***}	0.052^{***}	-0.168^{***}	-0.282
	(0.002)	(0.011)	(0.016)	(0.174)
Rain	0.019	-0.001	0.087^{*}	0.036
	(0.010)	(0.031)	(0.037)	(0.196)
Snowfall	0.023	0.038	0.086	-0.079
	(0.020)	(0.081)	(0.123)	(0.600)
Traffic Condition	4.804	39.608^{***}	-60.261^{***}	-514.205^{***}
	(3.991)	(4.957)	(16.218)	(20.416)
Temperature * Trips Pooled	-0.001^{***}	-0.002^{*}	-0.005^{***}	-0.001
	(0.0002)	(0.001)	(0.001)	(0.007)
Rain * Trips Pooled	-0.003	0.004	-0.027^{*}	0.009
	(0.002)	(0.011)	(0.012)	(0.105)
Temperature * Traffic Condition	0.153	1.098^{***}	-1.960^{***}	7.258^{***}
	(0.090)	(0.131)	(0.227)	(0.545)
Rain * Traffic Condition	-1.747	1.721	-3.508	-6.899
	(1.621)	(3.632)	(4.489)	(15.540)
Income Fixed effects	Y	Y	Y	Y
Traffic Fixed effects	Y	Y	Y	Y
Observations	2,991,753	600,496	2,991,753	600,496
R^2	0.117	0.529	0.009	0.051
Adjusted R^2	0.117	0.529	0.009	0.051
Residual Std. Error	2.069 (df = 2991732)	2.283 (df = 600475)	10.339 (df = 2991732)	13.242 (df = 600475)
Note:				*p<0.1; **p<0.05; ***p<0.01

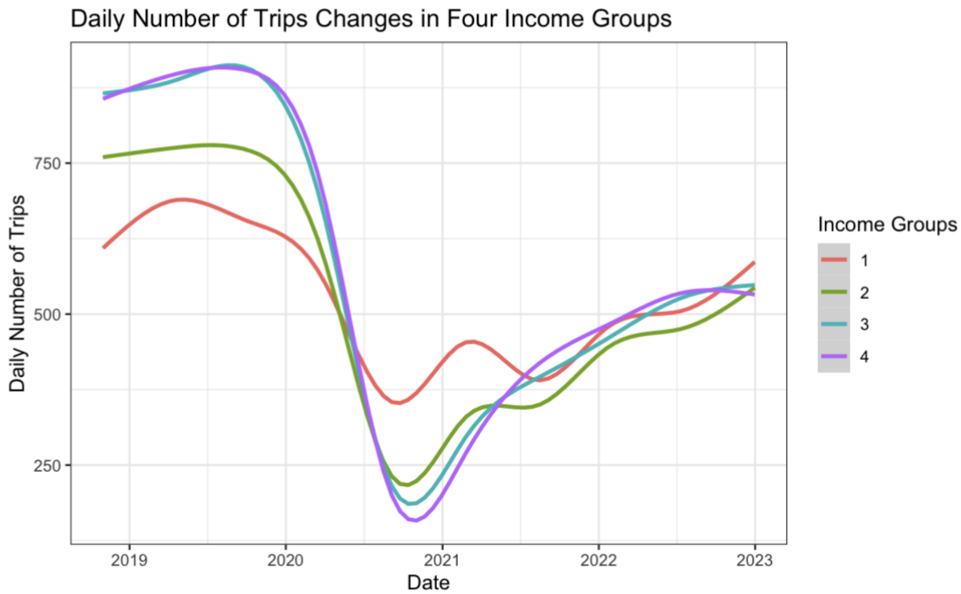


Figure 3: Daily Number of Trips Changes in Four Income Groups

3.2 Story 2: Shifting Transportation Habits: Exploring the Pandemic's Impact on Riding Behavior Across Income Groups

The impact of the pandemic on the behavior of various income groups has yielded unexpected and intriguing results. To quantify the income levels of riders, we have utilized the average income of the pickup and dropoff community areas and have categorized the riders into four distinct income groups, with Group 1 representing the poorest and Group 4 representing the richest. Based on this, we initially drew a graph to show how the daily number of trips in each income group changes.

However, after some preliminary analysis, we realized the accuracy of the graph depicting the daily number of trips in each income group may be biased as the classification of income groups also accounts for the local population in the respective areas. Specifically, certain areas that are considered "rich," such as downtown, experienced a high number of trips prior to the pandemic, followed by a sharp decrease during the pandemic, which may be attributed to the significant population residing in these areas and the forced closure of offices. As a result, we have refined our analysis to include a control variable based on the local population, resulting in the presentation of the daily number of trips per capita.

From this updated graph (Figure 4), we can find that after controlling the population, the poorer populations consistently have more trips compared to wealthier groups. This could be due to low-income individuals may have less access to private transportation.

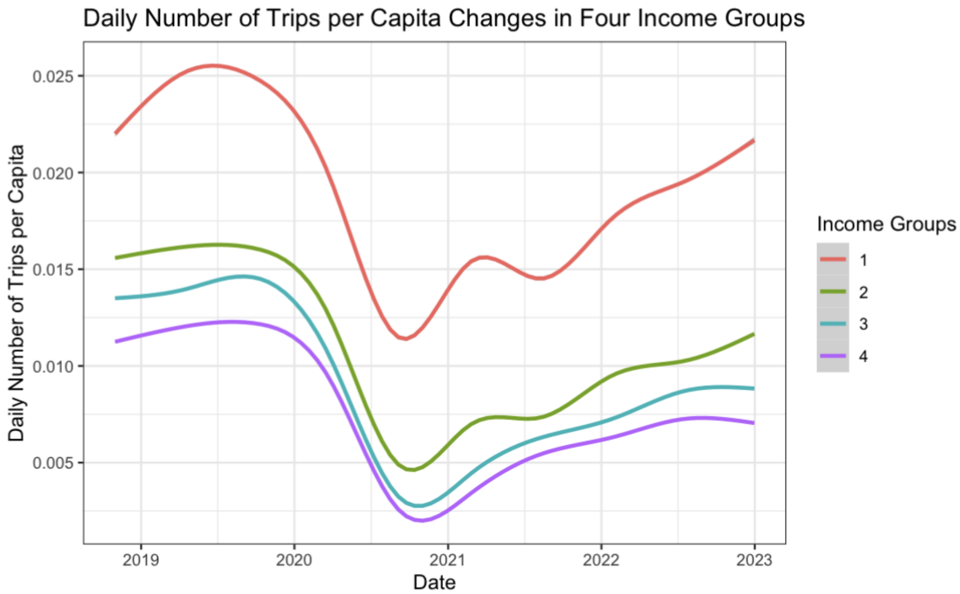


Figure 4: Daily Number of Trips per Capita Changes in Four Income Groups

During the pandemic, all income groups had an almost-same-level decrease in the figure. It makes sense that the pandemic had a widespread impact on all individuals regardless of their income levels. Government-imposed lockdowns, travel restrictions, and remote work policies affected people's daily routines, resulting in a significant reduction in the number of trips taken by all income groups.

After the pandemic, we observed a larger recovery for the low-income group which is unexpected as all income groups had decreased at a similar rate. To investigate it, we first assess whether the difference is statistically significant. We employed an interaction model of Income and Recent, where Recent is a time dummy variable. The coefficient of Income was found to be negative, indicating that higher income levels were associated with a lower number of trips per capita. The coefficient of Recent was positive, suggesting that all income groups experienced an increase in the daily number of trips per capita after the pandemic.

The coefficient of the intersection term (i.e., Income * Recent) is of particular interest, as it provides insight into whether the increase in the daily number of trips per capita after the pandemic differed significantly across income groups. The negative coefficient of the intersection term indicates that, compared to their counterparts during the pandemic, higher-income groups experienced a smaller increase in the daily number of trips per capita after the pandemic compared to the poorest group. The results of our statistical analysis confirm the observation made in the graph, indicating that there was indeed a larger recovery for the low-income group after the pandemic.

Table 3: Comparing Income's Effects During and After the Pandemic

	Dependent variable:
	Daily Number of Trips Per Capita
Income	-0.003^{***}
	(0.0003)
Recent	0.005^{***}
	(0.0003)
Income * Recent	-0.001^{***}
	(0.0001)
Additional Variables Included	Y
Weekday Fixed effects	Y
Hour Fixed effects	Y
Traffic Fixed effects	Y
Observations	1,074,821
R^2	0.493
Adjusted R^2	0.396
Residual Std. Error	0.007 (df = 903388)
Note:	*p<0.1; **p<0.05; ***p<0.01

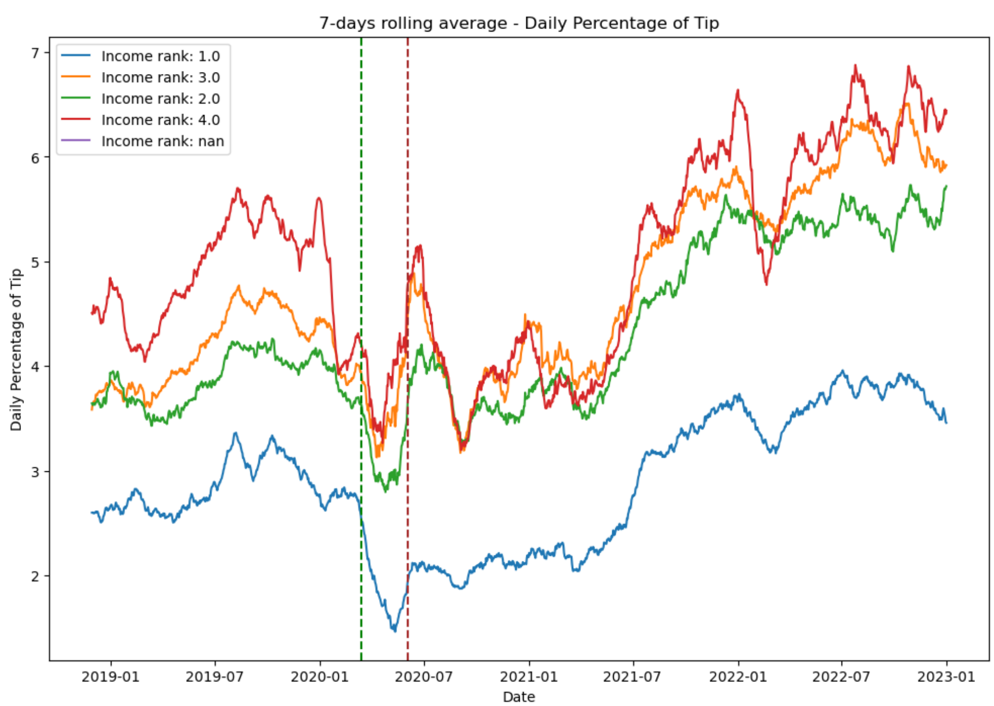


Figure 5: Daily Percentage of Tip with respect to fare across the 4 defined income groups

This could be explained by the fact that these populations have less resources and flexibility with their work. As an example, many "essential" workers could not work remotely and their working styles are fully in-person, while other income groups are more flexible with their work. Another explanation is that the richer group might have more health concerns about public transportation and turn to private vehicles. After the pandemic, the richer groups continue to use their personal vehicle, while the poor population may not be able to afford it and still need to use ride-sharing services. This highlights the importance of understanding how socioeconomic factors can impact individuals' ability to adapt to changes brought about by crises such as pandemics.

The data on tipping behavior among different income groups also shows some other interesting trends. Firstly, it is expected that higher-income areas have a larger percentage of riders who tip, given that they have more disposable income. This suggests that the ability to tip may be related to one's financial situation. However, when only considering rides that have tips, lower-income groups give both higher absolute and percentage tip amounts. This is surprising since higher-income groups are generally expected to tip more, but the data shows the opposite trend.

One potential explanation for this could be that some well-off riders in lower-income areas are particularly concerned about their safety during rides and are therefore more likely to give higher tips for a safe ride. This could be due to a perception that rides in lower-income areas are less safe or that there is a higher risk of crime. It is also possible that cultural or social factors play a role, such as a greater emphasis on community solidarity or reciprocity in lower-income areas.

To fully explain this unexpected result, additional qualitative data on tipping behaviors may be needed. For example, conducting interviews or surveys with riders from different income groups could help shed light on the reasons behind their tipping decisions. In addition, it is worth noting that the regression analysis in Table 3 confirms a positive correlation between higher income and tips for all rides, but a negative correlation for tip percentage and amount in tipped rides. This suggests that while higher-income riders may be more likely to tip in general, they may not necessarily give larger tips as a percentage of the ride cost.

3.3 Story 3: Impact of Bad Weather on Ride Demand and Tipping

We are interested in investigating the potential effects of bad weather conditions on the number of trips and tipping percentage. It can help companies better anticipate the market and prepare for an expected increased demand during inclement weather which can improve their overall service and customer satisfaction. Second, it can provide valuable insights into the behavior of consumers in response to weather conditions and how it may affect their decision-making regarding transportation options. This information can be helpful for drivers as they choose when to work.

We will test two hypotheses (target effects): compared to good weather conditions, the number of trips will increase on days with bad weather conditions and the tipping percentage will also increase on days with bad weather conditions.

Our hypothesis is based on the intuition (expected explanation) that under bad weather conditions, more people may choose to take a ride share rather than walking, biking, or taking public transportation in such conditions. Additionally, as tips are supposed to be positively correlated with customer satisfaction, we predict that the difficulty of getting anywhere during bad weather conditions may lead to an increase in tipping percentage, as riders may be more appreciative of the service.

The results from Table 4 show the coefficients of weather attributes on tipping percentages per ride and the daily number of trips per capita while controlling for income and traffic fixed effects. These regression results are applicable to all income groups and traffic conditions. In Column (2), it is demonstrated that during bad weather conditions, such as higher levels of rain, snow, and lower temperatures, the number of trips increases significantly, indicating a large demand for rides on such days. This finding supports our initial hypothesis. The estimated effects are found to be larger regarding snow compared to rain and temperature drop.

However, contrary to our intuition, we do not observe a significant increase in tipping percentages. Interestingly, as shown in table 4, we see a significant increase in the fare due to surge pricing. This may lower the motivation of riders to tip as they are already paying more for the same base ride.

Another possible explanation is that the trips taken during bad weather conditions may be shorter and primarily used to fill the gap between starting points and public transportation nodes, resulting in a lower ride time and, consequently, lower tipping amounts. The short duration of the trip may not be sufficient to motivate riders to tip the driver.

We can investigate the impact of trip length on tipping. In Table 5, we treat trip miles as the response variable and weather attributes as the explanatory variables. We find that during bad weather days, such as rainy or snowy days, the trip miles decrease.

The findings of this analysis can have several implications for ride-sharing drivers. While bad weather conditions may increase demand for rides, drivers may not receive higher tips as a result. However, they are still likely to get paid more than good weather days as the base fare has increased. This information is helpful for drivers to make the tradeoff of whether the wage increase is worth having to work in different types of inclement weather.

Table 4: The Coefficients of Income Weather Attributes on Tipping Percentage Per Ride/Daily Number of Trips Per Capita

	Dependent variable:			
	Tip		Tipping Percentage Per Ride	
	All Trips	Tipped Trips	All Trips	Tipped Trips
	(1)	(2)	(3)	(4)
Fare	0.060^{***}	0.162^{***}	0.023^{***}	-0.148^{***}
	(0.002)	(0.002)	(0.004)	(0.008)
Shared Trip Authorized	-0.320^{***}	-0.859^{***}	-1.950^{***}	0.554^{**}
	(0.005)	(0.027)	(0.036)	(0.196)
Hour	-0.007^{**}	-0.003	-0.009	0.102^{***}
	(0.002)	(0.001)	(0.011)	(0.015)
Temperature	0.003^{***}	0.001	0.026^{***}	-0.042^{***}
	(0.0005)	(0.001)	(0.002)	(0.008)
Trips Pooled	0.026^{***}	0.052^{***}	-0.168^{***}	-0.282
	(0.002)	(0.011)	(0.016)	(0.174)
Relative Humidity	-0.001^{*}	-0.002^{***}	-0.004	0.002
	(0.0004)	(0.001)	(0.002)	(0.003)
Rain	0.019	-0.001	0.087^{*}	0.036
	(0.010)	(0.031)	(0.037)	(0.196)
Snowfall	0.023	0.038	0.086	-0.079
	(0.020)	(0.081)	(0.123)	(0.600)
Income	0.183^{***}	-0.108^{***}	1.058^{***}	-0.842^{***}
	(0.021)	(0.011)	(0.118)	(0.088)
Traffic Condition	4.804	39.608^{***}	-60.261^{***}	-514.205^{***}
	(3.991)	(4.957)	(16.218)	(20.416)
Temperature * Trips Pooled	-0.001^{***}	-0.002^{*}	-0.005^{***}	-0.001
	(0.0002)	(0.001)	(0.001)	(0.007)
Rain * Trips Pooled	-0.003	0.004	-0.027^{*}	0.009
	(0.002)	(0.011)	(0.012)	(0.105)
Temperature * Traffic Condition	0.153	1.098^{***}	-1.960^{***}	7.258^{***}
	(0.090)	(0.131)	(0.227)	(0.545)
Rain * Traffic Condition	-1.747	1.721	-3.508	-6.899
	(1.621)	(3.632)	(4.489)	(15.540)
Income Fixed effects	Y	Y	Y	Y
Traffic Fixed effects	Y	Y	Y	Y
Observations	2,991,753	600,496	2,991,753	600,496
R^2	0.117	0.529	0.009	0.051
Adjusted R^2	0.117	0.529	0.009	0.051
Residual Std. Error	2.069 (df = 2991732)	2.283 (df = 600475)	10.339 (df = 2991732)	13.242 (df = 600475)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 5: The Coefficients of Weather and Attributes on Tipping Percentage Per Ride/Daily Number of Trips Per Capita/Fare

	Dependent variable:		
	Tipping Percentage Per Ride	Daily Number of Trips Per Capita	Fare
	(1)	(2)	(3)
Temperature	0.013	-0.0001^{***}	0.090^{***}
	(0.006)	(0.00001)	(0.015)
Relative Humidity	-0.004^{*}	0.00001	-0.031^{***}
	(0.002)	(0.00000)	(0.003)
Rain	0.032	0.0002^{***}	0.207^{***}
	(0.016)	(0.00001)	(0.014)
Snowfall	0.041	0.001^{***}	0.864^{**}
	(0.104)	(0.0001)	(0.254)
Additional variables included	Y	Y	Y
Income Fixed effects	Y	Y	Y
Traffic Fixed effects	Y	Y	Y
Observations	2,991,753	2,991,753	2,991,753
R^2	0.083	0.347	0.641
Adjusted R^2	-0.007	0.283	0.606
Residual Std. Error	10.425 (df = 2724830)	0.009 (df = 2724830)	7.560 (df = 2724831)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 6: The Coefficients of Weather Attributes on Trip Miles

	Dependent variable:
	Trip Miles
Fare	0.191***
	(0.001)
Shared Trip Authorized	0.525***
	(0.008)
Trips Pooled	0.644***
	(0.006)
Temperature	−0.011***
	(0.0002)
Relative Humidity	0.006***
	(0.0001)
Rain	−0.038***
	(0.003)
Snowfall	−0.095***
	(0.015)
Income	−0.091***
	(0.002)
Income Fixed effects	Y
Weekday Fixed effects	Y
Observations	2,991,753
R^2	0.913
Adjusted R^2	0.905
Residual Std. Error	2.252 (df = 2724828)
Note:	*p<0.1; **p<0.05; ***p<0.01

4 Conclusion, Summary and Impacts

4.1 Conclusion and Summary

It is important to note that although our dataset is a ride-sharing data from Chicago, we can still draw meaningful conclusions and stories that represent the general trends of the ride-sharing industry in metropolitan cities of the US. The data and analysis presented in this project provide valuable insights into the behavior of ride-sharing customers and drivers, as well as the impact of external factors such as the COVID-19 pandemic and weather conditions on the industry. These insights can inform future research and decision-making in the ride-sharing industry, and potentially be applied to other similar industries as well, with three of the prominent one narrated below.

Firstly, the COVID-19 pandemic had a significant impact on Chicago's ride-sharing industry, as seen in the sharp decline in both the number of trips and tipping percentages. However, there has been a steady recovery from 2021, with tipping percentages even surpassing pre-pandemic levels. As of the end of 2022, daily trips have rebounded to approximately two-thirds of pre-pandemic levels, while tipping percentages have reached an average of 5.5%, compared to the pre-pandemic average of 4.0%. These observations suggest that customer behavior in the ride-sharing industry has changed since the pandemic, and these changes may continue to affect the industry in the future.

Secondly, the pandemic has had a significant impact on ride-sharing behavior across Chicago's various income groups, with unexpected and intriguing results. The refinement of the analysis to include a control variable based on the local population revealed that low-income individuals had more trips compared to wealthier groups for all timelines. After the pandemic, we observed a slightly larger recovery for the low-income group, indicating that socioeconomic factors can impact individuals' ability to adapt to changes brought about by crises such as pandemics. The data on tipping behavior among different income groups also shows some unexpected trends, with lower-income groups giving higher absolute and percentage tip amounts, suggesting that cultural or social factors may play a role. Further qualitative data may be needed to fully explain these results.

Thirdly, bad weather conditions, such as rain, snow, and lower temperatures, significantly increase the number of trips, supporting the initial hypothesis. However, there was no significant increase in tipping percentage, possibly due to fare surcharge during bad weather days. These findings suggest that ride-sharing drivers may not receive higher tips on bad weather days, but should still consider working on those days to take advantage of the high demand. Instead, drivers may want to focus on providing quality service to build their reputation and potentially receive better tips in the long run.

Throughout the project, we learned how to handle large amounts of observation that would not be possible to work with standard dataset processing tools available in Python. We learned, by prompting chatGPT for a solution, to split the dataset into chunks and process each chunk into separate files. This enabled us to either random sample the data, or if necessary, work with streaming methods to obtain our estimate. This is an important toolbox in data science repertoire especially given limited computational and software access.

4.2 Impacts

The conclusions drawn from the above projects have important implications for several stakeholders, including ride-sharing drivers, the Chicago government, and ride-sharing companies. These stakeholders can use the findings to improve their services, better understand their customers' needs, and make informed decisions that impact their businesses.

For ride-sharing drivers, the analysis of bad weather conditions and its impact on trip demand and tipping behavior can help them anticipate potential increases in demand during inclement weather and adjust their expectations regarding tipping. By providing quality service, drivers can build their reputation and potentially receive better tips in the long run. Additionally, understanding how socioeconomic factors impact riders' behavior can help drivers cater to the needs of different income groups and improve their overall customer satisfaction.

The Chicago government can also benefit from the findings related to changing transportation habits during and after the pandemic. By understanding how different income groups were affected by the pandemic, the government can make informed decisions about public transportation and ride-sharing policies that benefit all members of the community. For instance, the government may consider subsidizing ride-sharing services for low-income individuals to help them access reliable transportation.

Ride-sharing companies can also use the conclusions drawn from these projects to improve their services and increase customer satisfaction. For example, by understanding how bad weather affects trip demand and tipping behavior, ride-sharing companies can prepare for potential increases in demand and incentivize drivers to work during these conditions. Additionally, by understanding how different income groups use ride-sharing services and how they tip, ride-sharing companies can tailor their services to cater to the needs of different income groups.

The statistical analyses conducted in these projects provide convincing evidence to support the conclusions drawn. The use of regression analysis allows for the identification of significant relationships between variables and the control of potential confounding variables. Furthermore, the use of interaction models allows for the assessment of how different factors interact with one another and impact outcomes. The results of these analyses are presented in tables and graphs, making them easy to interpret and understand. Overall, the conclusions drawn from the above projects have important implications for ride-sharing drivers, the Chicago government, and ride-sharing companies. By using the evidence provided by the statistical analyses and understanding the needs of different income groups, stakeholders can improve their services, make informed decisions, and increase customer satisfaction.

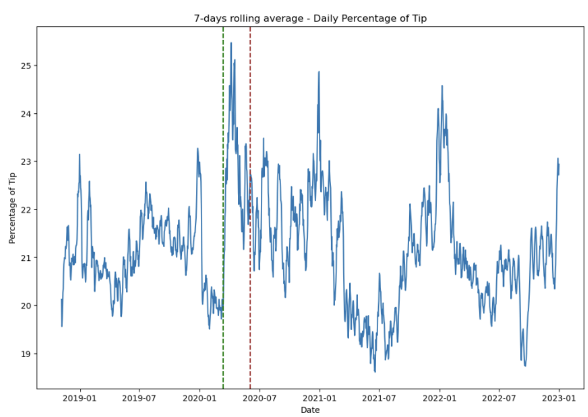
4.3 Related Links

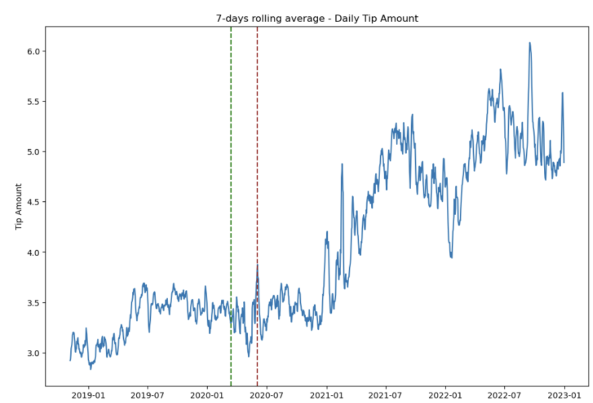
- Practical Impacts in GitHub

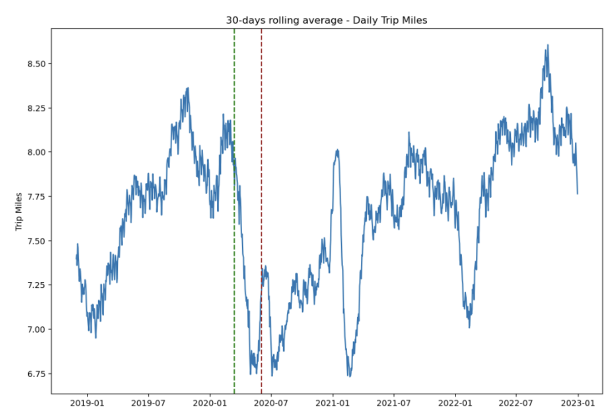
5 Appendix

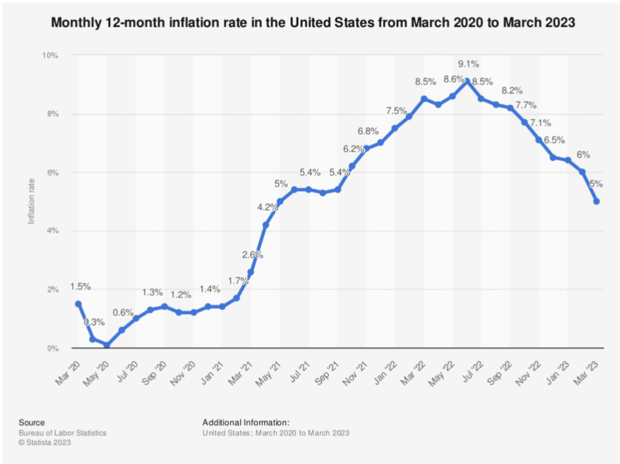
5.1 Analysis of Tipped Rides: Understanding Trends in Fare and Tipping Amounts and impact of inflation

After analyzing rides that received tips, we made some interesting observations. Firstly, we found that the distance covered during these rides remained relatively stable when compared to pre and post-Covid periods (statistically proved using proportional z-test; see code). However, we noticed a statistically significant increase in both the fare amount and the absolute tipping amount. This increase can be attributed to the rising inflation in the US since the onset of the Covid pandemic, as evidenced by Image 5. Therefore, while the distance of these rides remained consistent, the overall cost of the ride has increased, leading to a higher fare and tipping amount. These findings suggest that the impact of Covid on the ride-sharing industry goes beyond just changes in demand and supply, and has significant implications on the pricing dynamics of the industry.



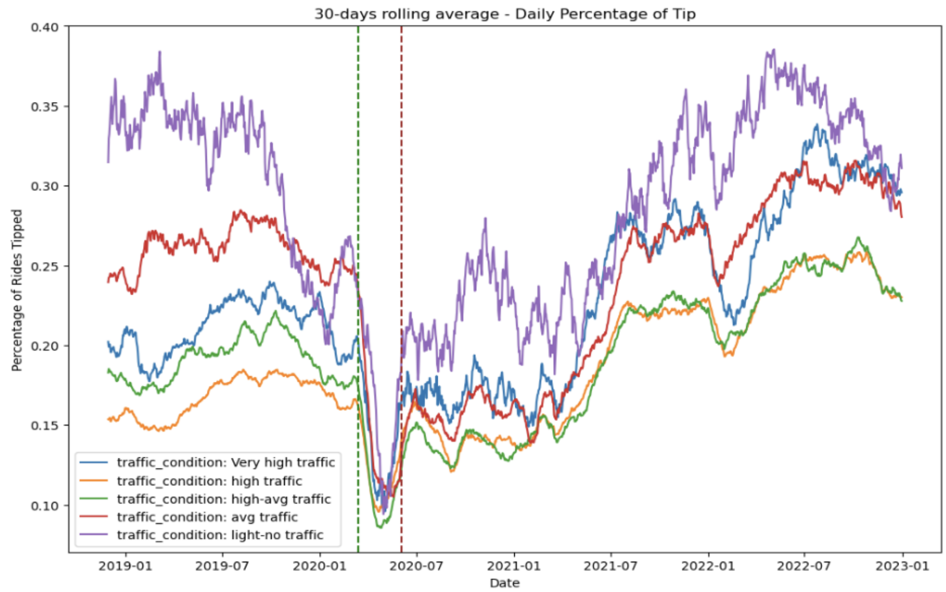


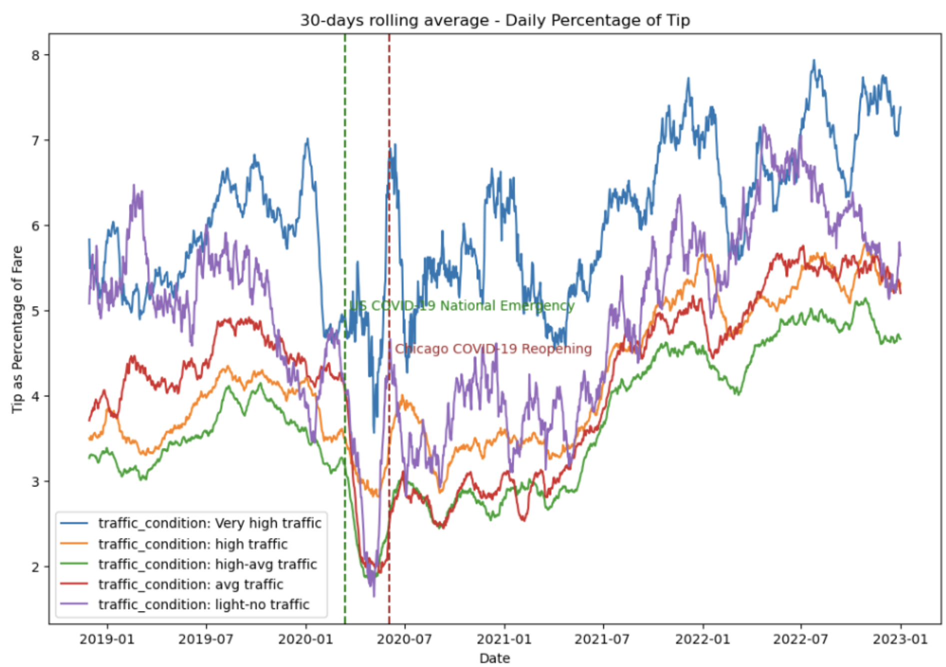




5.2 Tipping behavior based on Traffic Condition

In addition to analyzing tipping trends in rides, we also delved deeper into how tipping behavior varies across different traffic conditions. Our analysis revealed some interesting findings, including a statistically significant increase in the number of tipped rides and the percentage of tips compared to the fare in high-traffic rides during the post-Covid period as compared to the pre-Covid period. This increase suggests a change in the attitude of passengers towards tipping, especially during times of high traffic. One surprising observation was that in the pre-Covid period, passengers did not tip generously during high-traffic rides, indicating a lack of empathy for the driver's situation. However, in the post-Covid period, passengers seem to have become more generous with their tips, possibly due to the perceived risk associated with the pandemic and an increased appreciation for the drivers' efforts. These findings highlight the importance of understanding tipping behavior and its drivers in the ride-sharing industry, as it can have a significant impact on driver earnings and overall customer satisfaction.


30-days rolling average - Daily Percentage of Tip
Percentage of Rides Tipped
Date
traffic_condition: Very high traffic
traffic_condition: high traffic
traffic_condition: high-avg traffic
traffic_condition: avg traffic
traffic_condition: light-no traffic


30-days rolling average - Daily Percentage of Tip
Tip as Percentage of Fare
Date
US COVID-19 National Emergency
Chicago COVID-19 Reopening
traffic_condition: Very high traffic
traffic_condition: high traffic
traffic_condition: high-avg traffic
traffic_condition: avg traffic
traffic_condition: light-no traffic