Project 2: Tips in trips in the city of Chicago

MATH-516 Applied Statistics

Linda Mhalla

2025-03-03

1 Data

Transportation Network Providers (TNPs) such as Uber and Lyft operate in Chicago. As part of its licensing process, Chicago requires the companies to report on their activities monthly

- The dataset provides detailed trip-level information on reported trips
- This dataset covers trips during **October 2024**
- Source: City of Chicago Open Data Portal
- Key features:
  - Temporal trends
  - Spatial patterns
  - Fare and cost analysis
  - Shared ride behaviour

2 Data: Trip information

- `trip_id`: unique identifier for each trip
- `trip_start_timestamp`: timestamp when the trip started
- `trip_end_timestamp`: timestamp when the trip ended
- `trip_seconds`: total duration of the trip in seconds
- `trip_miles`: Total distance of the trip in miles
- `fare`: base fare of the trip
- `tip`: amount tipped by the passenger
- `additional_charges`: extra fees (e.g., service fees)
- `trip_total`: total cost of the trip (sum of fare, tip, and additional charges)
- `shared_trip_authorized`: whether the passenger opted for a shared ride
- `shared_trip_match`: whether the ride was actually matched with another passenger
- `trips_pooled`: number of passengers pooled in a shared trip

Note: times are rounded to the nearest 15 minutes, fares are rounded to the nearest 2.50, and tips are rounded to the nearest 1.00

3 The Goal

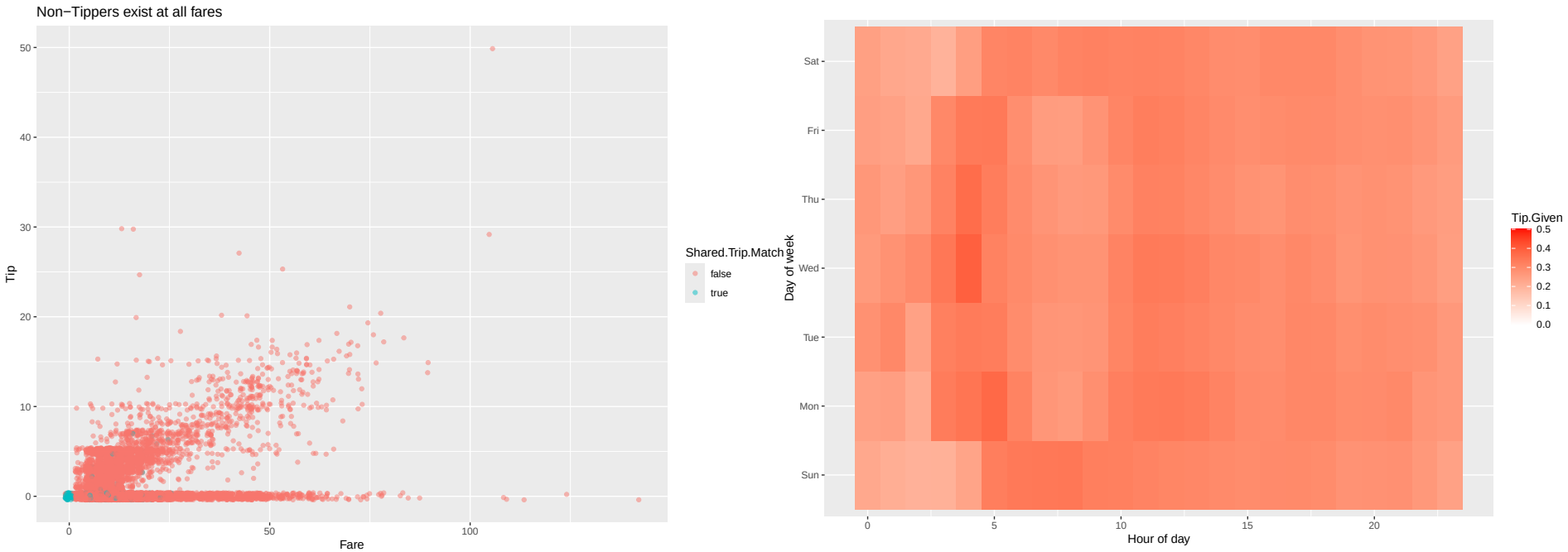
The dataset provides valuable insights into:

- temporal trends (peak ride-sharing hours and seasonal variations)
- spatial mobility (pickup/dropoff areas)
- fare & cost analysis (fare structures, tip patterns, and additional charges)
- ride-sharing trends (e.g., how many trips were pooled vs. individual rides)

We will focus on one specific aspect: **the tipping behaviour**

$\rightarrow$ Understanding tipping patterns can help companies adjust pricing models or **researchers study socio-economic factors influencing tips**

4 The Goal


Non−Tippers exist at all fares
Tip
Fare
Shared.Trip.Match
false
true
Day of week
Hour of day
Tip.Given

5 The research question: To tip or not to tip?

- Why do some passengers tip, while others don't tip at all?
- What factors influence tipping in ride-sharing services?
- Can we build predictive models to understand tipping behavior?

5.1 Key questions

1. How often do passengers tip in Chicago TNP rides?
2. Do trip duration, distance, fare, or time of day affect tipping?
3. Are certain pickup or drop-off locations associated with a higher probability of tipping?
4. Are certain time of the day or weekdays associated with a higher probability of tipping?
5. Do shared rides affect tipping behaviour?

6 Goals

- Understand the meaning of the different predictors
- Explore the data to understand the propensity to tip
- Build a logistic regression model
  - Should you include all predictors? Use LRTs to decide
  - How can you interpret the impact of the different predictors on the propensity to tip?
  - What type of predictors are the most relevant? Spatial, temporal?
- Evaluate prediction performance of your model, in terms of AUC, by cross-validation
- Use a different classification method and compare with logistic regression in terms of prediction

7 Cross-Validated AUC

```
library(pROC)
AUC_eval <- function(gmodel,Data, folds=5){
  set.seed(516)
  Folds <- matrix(sample(1:dim(Data)[1]), ncol=folds) # is this wise?
  AUC <- rep(0,folds)
  for(k in 1:folds){
    train <- Data[-Folds[,k],]
    test <- Data[Folds[,k],]
    my_gm <- glm(gmodel$formula, family="binomial", data=train)
    test_pred <- predict(my_gm, newdata = test, type="response")
    AUC[k]    <- auc(test$y,test_pred)
  }
  return(mean(AUC))
}
```