Project 1 MAST30034

On the ratio of Taxis and HVFHVs

Cassidy Spanger Student ID: 1356157
Github repository

September 7, 2025

Introduction

In the past few years, the factors that influence an individual's decision between hailing a taxi and ordering a ride through a mobile based service have been of interest to the fields that involve transportation and its realization. This varies from the individual drivers whose livelihoods rely on the side these decisions lean towards, to the government and concerns surrounding accessibility, congestion, safety and urban policy.

Traditional taxis operate under static pricing, with specific tolls applied to definitive situations to ensure fairness, transparency, and accessibility, while also allowing governments to collect predictable revenue through standardized taxes and surcharges. On the other hand, hailing services (such as Uber or Lyft) are subject to surge pricing during times of high demand or short supply. [1] In this report, the central hypothesis is that adverse weather conditions causing increased demand may impact choice between traditional street-hail taxis and app-based ride-hailing services in Manhattan. Considering whether higher prices during rain could influence rider behaviour.

Data Sources

New York City Taxi and Limousine Commission (TLC) Trip Record Data [2]

The TLC publishes monthly datasets of trip data for yellow and green NYC taxis and High-Volume For-Hire Vehicles (HVFHV) in Parquet format. Yellow taxis can respond to hailing in all NYC boroughs, while green taxis are only permitted to respond above W 110 St/E 96th St in Manhattan and in the boroughs. HVFHV's refer to for hire vehicles (FHV) operated by businesses that dispatch more than 10,000 FHV trips in New York City per day (Uber, Lyft).

The Open-Meteo Historical Weather API[3]

Open-Metro provides a comprehensive record of past weather conditions at an hourly granularity for locations based on global coordinates. They demarcate NYC as coordinates (40.712778, -74.006111).

Data choices and constraints

The following data was collected from the above sources

Pickup times of trips for Yellow Taxis and HVFHVs from TLC Trip record data from March 2024 to June 2025, excluding December–February. Yellow Taxis were chosen as the best baseline for comparison against HVFHVs because of static regulated pricing that directly addresses the research hypothesis and are concentrated in Manhattan. Including green taxis would add noise without directly strengthening the core hypothesis. Winter months were removed to avoid snow confounding rain effects.

Hourly rainfall and apparent temperature measurements from Open-Meteo from lower Manhattan for the same time period as the taxi data. Rainfall is the most direct application of the research hypothesis, as it is likely to discourage street hailing and could shift demand and therefore surge prices to app-based services. Temperature was also included since it is often the first condition people associate with weather when making decisions about being outside and has potential for significant interactions with rain as there are correlations between being wet and being cold.

Methodology

1. Downloaded TLC and weather datasets
2. Aggregated TLC counts by hour and computed the ratio of Yellow Taxis to HVFHV trips
3. Joined with the weather data and added extra time-based flags.
4. Conducted exploratory analysis, preliminary visualisations and outlier handling
5. Built statistical models:

   Ordinary Least Squares Regression to test linear effects of the variables against the ratio.

   Random Forest Regression to capture nonlinear relationships and interaction effects.
6. Compared model results and interpreted implications for target audience consisting of individual drivers whose income depends on demand, transportation regulatory bodies concerned with accessibility and fairness and ride hailing platforms interested in demand prediction for surge pricing and fleet allocation

Preprocessing

The three datasets [Yellow Taxi, HVFHV, Weather] were prepared for modeling with the following process. The datasets were in parquet format and processed using Pyspark.

The shape of the Yellow Taxi dataset was initially (19, 48559814).

The shape of the HVFHV dataset was initially (24, 260628583). The shape of the Weather dataset was initially (3, 9528).

1. The granularity of the pickup time fields for the transport datasets was truncated up to the hour, as to match with the available weather data.
2. The pickup times were aggregated by the hour, creating a new column of data that contained the number of pickups made by a transport service for each hour

   The shape of the new transport count datasets is now (2, 9528), containing a column of datetime information and counts of trips per hour.

3. The transportation datasets were outer joined on pickup hour, such that a new data frame called FinalData contained datetime information and count of transportation pickups per hour for both transport classifications with shape (3, 9528)
4. Another column was added for the ratio between Yellow Taxis and HVFHVs counts per hour. Ratio was chosen as the main response variable as by using the ratio instead of counts, the influence of demand fluctuations such as events where all transport is up or down is reduced, counts were kept for visualization purposes. Shape (4, 9526)
5. The data frame was then inner joined on hour with the weather data, shape (6, 9528)
6. Additional binary variables were created to mark if a particular hour was during rush hour, defined as 7am-9am and 4pm-6pm, or on a weekend. To be able to measure the potential interaction with the ratio. As well as a column containing just what hour of the day in 24 hour time to be used as model input. shape (9, 9528)
7. After considering data skew an additional binary flag was added to indicate if the rain amount was above 0.1 mm/h to classify raining and not raining. shape (10, 9528)

Issues and Outliers

- Rain measurements are heavily right skewed, with most hours having a value of 0.
- There were 6 hours in the count of HVFHV trips that were considered outliers (1.5*IQR). As this is a very small portion of the data and the correspoding ratio was not considered an outlier, they will be ignored.
- The data was checked for null values, there were none
- To ensure no division by 0, +1 has been added to HVFHV counts when calculating the ratio, due to consistent high counts this has a negligible effect.

Exploratory Data Analysis

This section presents visualisations the data and preliminary statistical testing to check modelling assumptions

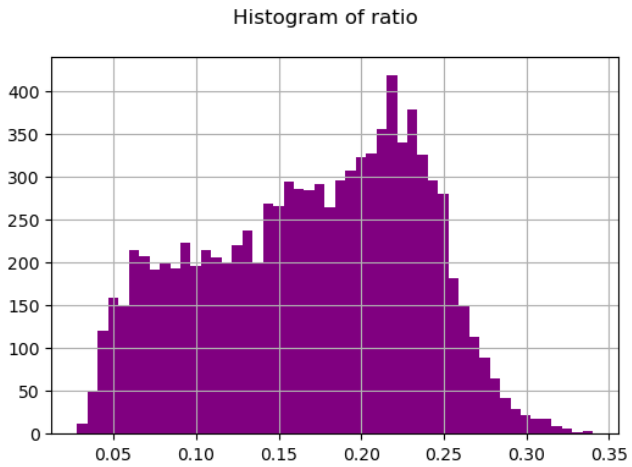

Figure 1: Histogram of the counts of the ratios between Yellow taxis and HVFHVs.

Distribution of Ratio

The distribution of ratio counts is explored in Figure 1. The shape is slightly unimodal and right skewed, with values all below 0.35 indicating that HVFHVs dominate the trips taken, with Yellow Taxi's still maintaining their share around 15-25% of the trips. Linear regression may fail to find significant coefficients for weather factors from the tight spread and regression assumptions may be violated.

Statistical Testing

Preliminary statistical tests were conducted to assess whether the assumptions of normality and homogeneity of variance were satisfied for the ratio under different rainfall thresholds. These tests compared the distribution of the ratio for "wet" and "dry" periods. Statistical tests were conducted at two rainfall thresholds: 0.1 mm/h (light rain) and 0.5 mm/h (moderate rain). The difference in distribution during rush hour when there are differences in demand and transport choices was also taken into account.

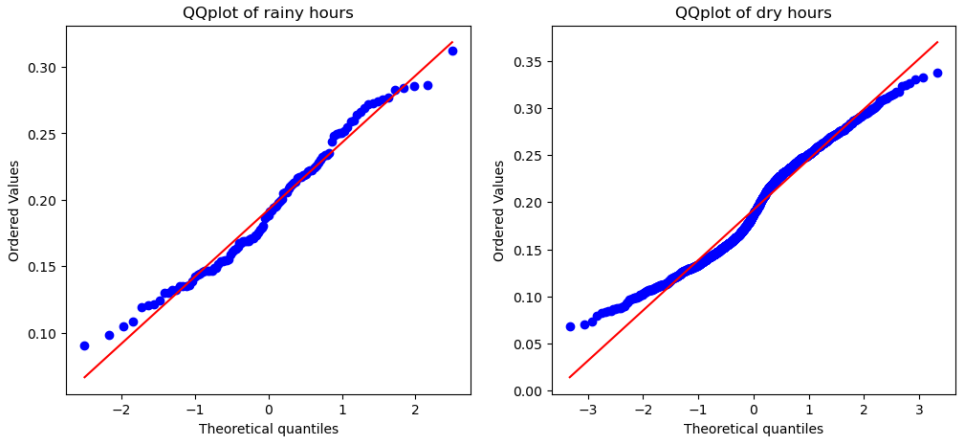

Figure 2: QQ plots Comparing the distribution of the Yellow/HVHFV ratio during rush hour under dry conditions ($\leq$ 0.5 mm/h) and wet conditions ($>$ 0.5 mm/h).

Quantile-Quantile (QQ) plots (Figure 2) showed that the ratio was approximately normal when restricted to rush hour periods under a rainfall threshold of 0.5 mm/h fig1. The ratio distribution for rainy hours approximately follows the normal line, with only mild tail deviations while the dry hours distribution exhibits heavier tails and modest skew, indicating mild violations of normality. Welch's t-test indicated no statistically significant difference in the mean ratio between rainy and dry conditions at this threshold ($p = 0.915$). These results imply that, when considering rainfall above 0.5 mm/h during rush hour, the ratios of Yellow to HVFHV trips are statistically indistinguishable from dry conditions.

However, at a lighter rainfall threshold of 0.1 mm/h during rush hour, the ratio was significantly different between rainy and dry conditions ($p = 0.034$). This suggests even light rain may alter transportation choices. It should be noted that rain below 0.1 mm/h exists very frequently in the data, meaning statistical tests have many observations to work with, which increases power and the chance of detecting subtle differences. In contrast, rainfall above 0.5 mm/h is relatively rare, leading to smaller sample sizes and lower power to detect effects. The lack of significance at the higher threshold therefore may be a result of limited data rather than a true absence of behavioural change.

When considering not just rush hour, assumptions were violated as QQ plots indicated deviations

from normality and Levene's test indicated heteroscedasticity for multiple thresholds. Although some t-tests were significant and the data provides a high sample size, violations of assumptions mean these findings have limited interpretability. It does suggest that rainfall has its linear effect clearest during busier periods of transport, which is the motivation for testing the relationship more systematically through linear regression.

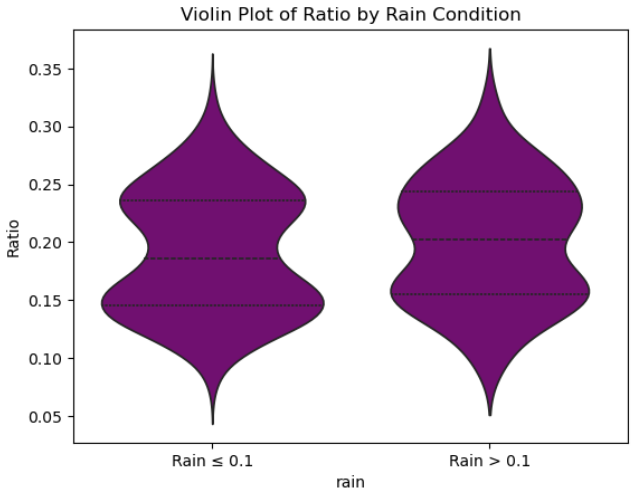

Figure 3: Violin plots of the distribution of the Yellow/HVHFV ratio during rush hour under dry ($\leq$ 0.1 mm/h) and wet ($>$ 0.1 mm/h) conditions

To further explore how rain alters the distribution of ratios, violin plots were produced (Figure 3). These plots provide a fuller view of the density of ratios in wet and dry conditions. Rainfall does not appear to drastically change the overall distribution of ratios. There is some indication that rainfall may correlate with less low ratios and more high ratios, which may be an indication of travellers preferring to take taxis in the rain to avoid surge pricing.

The approximate normality and variance equality observed under some conditions justify the use of OLS regression as a baseline descriptive model. However, violations at higher thresholds indicate that results should be carefully interpreted. Because assumptions do not universally hold flexible models such as Random Forest were implemented to capture nonlinearities and interactions without relying on assumptions.

Modelling

Two models were chosen to test the hypothesis that weather impacts the ratio of Yellow Taxi to HVFHV trips in Manhattan: Ordinary Least Squares Regression (OLSR) and Random Forest Regression (RF). These were selected as complementary approaches as OLSR provides a baseline linear and interpretable model, while RF can capture nonlinear relationships and interaction effects without requiring strong statistical assumptions. Both models were trained on the same dataset with the response variable of ratio of Yellow Taxi to HVFHV counts per hour.

Predictors:

Rain (mm/hr, continuous), Temperature (°C, continuous), Weekend (binary, 1 = weekend, 0 = weekday), Rushhour (binary, 1 = 7–9am, 4–6pm; 0 = otherwise). Hour (time of day, 0–23). In OLSR this was treated as categorical to capture cyclical nature of time, while in RF it was treated as numeric

since tree-based models can handle non-linearities. Interaction terms (rain*rushhour, rain*weekend) were included in OLSR to check if weather impacts choice more strongly during periods of increased demand.

Ordinary Least Squares Regression (OLSR)

OLSR was selected as the initial model as it is interpretable and explicitly tests linear relationships. The model assumes linearity, independence, homoscedasticity, and approximate normality of residuals. While QQ plots showed some mild deviations and Levene's test suggested heteroscedasticity in places, the large sample size (>9,500 rows) makes OLSR informative as a baseline.

The model explained approximately 72% of the variation in ratio ($R^2 = 0.717$). Hour of day was the dominant factor, with significantly lower ratios during the early morning hours (especially 4–5am), and a higher ratios during the middle of the day, peaking between 12–3pm (Figure 4.1). Rush hour increased the ratio by about 0.026, while weekends decreased it slightly by about 0.007.

Rainfall showed no significant effect, either directly or in the interaction terms, and temperature also had only a very small effect (Figure 4.2). The OLSR model suggests that weather conditions do not play a measurable role in explaining variation in the ratio.

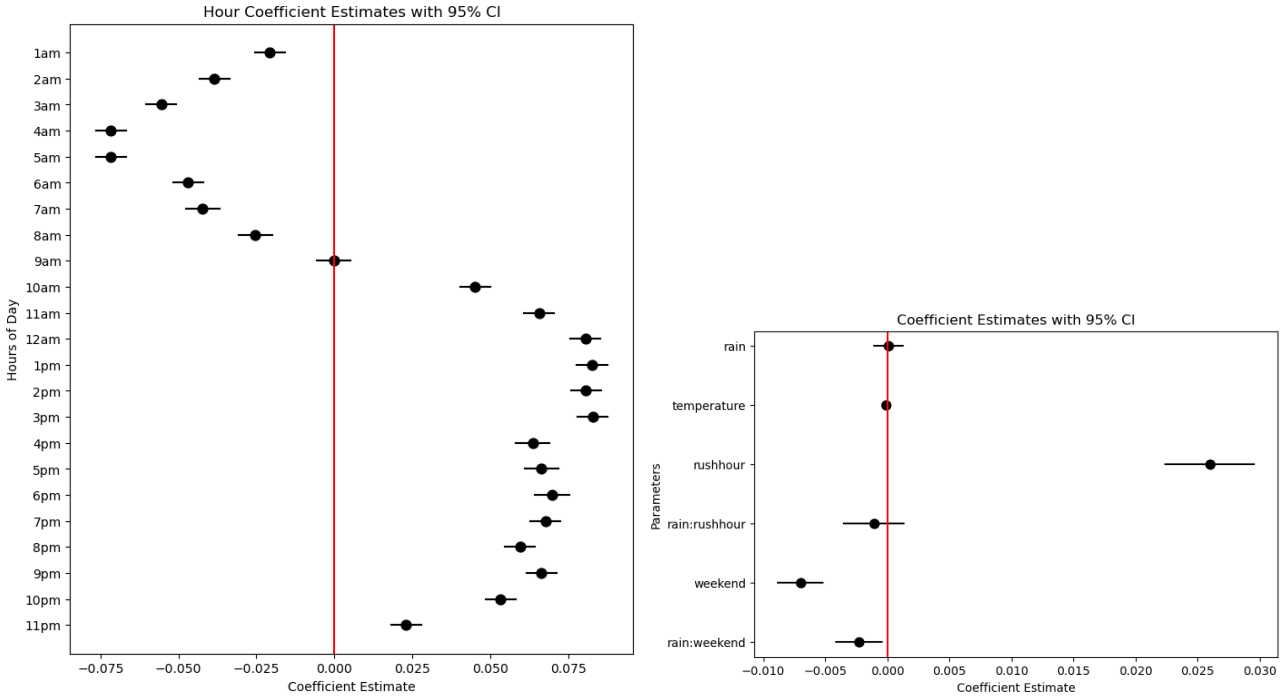

Figure 4: Regression coefficients of the effect of parameters on the Yellow/HVFHV ratio, with 95% confidence intervals

Random Forest Regression (RF)

RF regression was chosen as a complementary model to OLSR, as it is more flexible and can capture nonlinear or threshold effects of predictors. RF also does not require the same statistical assumptions as OLSR. RF is more equipped to handle dataset elements such as heavily right skewed rainfall and nonlinear time effects.

The RF model performed similarly to OLSR, explaining 72% of the variation in the ratio ($R^2 = 0.719$). Feature importance scores (Figure 6) showed that hour of the day was by far the most important predictor, followed by rush hour and weekend indicators. Rainfall and temperature contributed negligibly. This aligns with the regression results, which also found rain to be statistically insignificant, further strengthening the conclusion that weather is not a major driver of transportation choice.

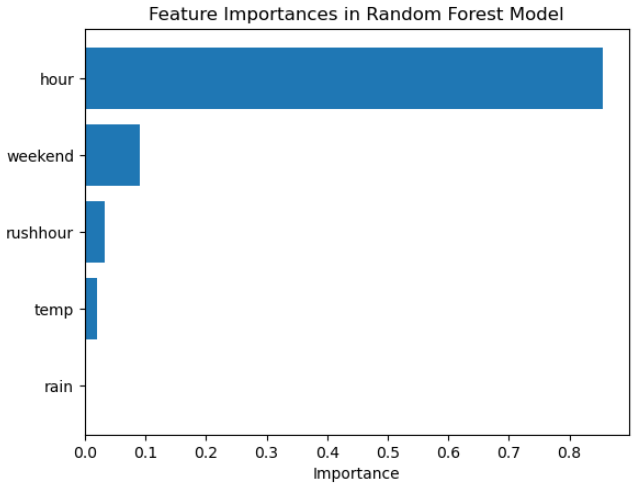

Figure 5: Feature importance scores from the Random Forest model, showing the relative contribution of each predictor to explaining variation in the Yellow/HVFHV ratio

Comparison of Models

Both models reached very similar explanatory power ($R^2 = 0.72$) and reached the same conclusion. Time of day drives the taxi/HVFHV ratio and weather does not. The models complemented each other by OLSR providing clear coefficients and confidence intervals that can be directly interpreted, while Random Forest allowed for the detection of nonlinear or threshold effects despite less interpretable variable importance ranking.

OLSR interpretability has more use for clear communication to policy makers and individuals when there are clear significant effects of parameters of interest. regression outputs can be fashioned into direct statements such as "weekends lower the taxi ratio by X%." Policy makers or city transport planners can easily understand which factors have the strongest and most consistent effects, without needing technical expertise in machine learning.

The benefit of RF comes from its capability of nonlinearities and skewed inputs, preventing model misspecification. While it does not have the usefulness of directionality, parameter ranking directly discriminates which features matter most in explaining variation in the ratio. This straightforward output is valuable for identifying what areas are worth further research or policy focus and what can be deprioritised. This allows for efficient resource allotment, for example, focusing on pricing policies for high demand hours rather than attempting to counteract weather effects on transportation. taxis and HVFHVs in Manhattan.

Recommendations

Support Yellow Taxis during off-peak hours

The ratio makes it clear that HVFHVs completely overshadow taxis at all hours, with Yellow having most trips compared to HVFHVs around midday. For government, this relates to a financial prerogative: if Yellow drivers cannot consistently find passengers outside these hours, their livelihoods are at risk. One way to address this would be to promote taxi usage during off-hours. This could be through advertising campaigns selling the benefit of standardised rates, fare incentives that make taxis cheaper than HVFHVs at certain times, or the introduction of designated taxi stands in consistently busy areas. These approaches would help Yellow drivers capture a greater share of the market and maintain a stable taxi sector that provides predictable tax revenue. This advice also applies to directly to drivers, who could improve their revenue by targeting off-peak hours where HVFHVs dominate, utilising quieter times into an opportunity to capture demand.

Don't expect adverse weather to change rider choice.

Both models showed rainfall and temperature had no real impact on the ratio. This means HVFHV surge pricing in the rain doesn't result in more use of Yellow Taxis, even though their fares stay fixed. For policy makers, this is evidence that fixed pricing might not provide the competitive benefit as hypothesised. Predictable pricing and street hail convenience is not enough, especially in result of bad weather, to encourage people to choose taxis over services like Uber or Lyft. Drivers should not rely on harsh weather to boost or decrease business in consideration of people seeking out static pricing.

Treat time as a vital parameter in planning and policy

Time effects were shown to be much stronger than all other factors, accounting for around 80% of the models' 72% explained variance (Figure 6), highlighting the dominant role of time in shaping demand. For regulators, this provides evidence for resource allocation and incentives in peak hours. For drivers, knowing that time plays a more important role than short term factors like weather is beneficial for knowing when is best to work (for taxis and HVFHVs). The feature importance should act as a guidance for directing future research and policy, showing that time is an important factor and should be further researched or compared to other factors not discussed in this report. Furthermore, predictors such as rain or temperature can be deprioritised, leaving room to investigate other influences such as special events or location choices.

Bibliography

Uber Technologies Inc.(2025). How surge pricing works. Retrieved from https://www.uber.com/au/en/drive/driver-app/how-surge-works/

New York City Taxi & Limousine Commission (TLC). TLC trip record data. Retrieved from https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Open-Meteo.com Weather API Retrieved from https://open-meteo.com/en/docs/historical-weather-api?latitude=40.7143&longitude=-74.006