# Can Yellow Taxis Survive? Analyzing the Battle for NYC's Streets

Quang Ngoc Dao
Student ID: 1338280
Github repo with commit

August 25, 2024

## 1 Introduction

The yellow cab is a New York symbol, immortalized through works of art like *Taxi Driver* (1976). At its height in 2012, the industry reached 12.5 million trips monthly[1]. The cost of a medallion (the taxi license) once peaked at an astonishing price of $1 million in 2013 [1]. Many aspiring drivers took on huge debts to earn it, seeing the job as a gateway to the American Dream.

However, the introduction of ride-hailing services like Uber disrupted the industry, introducing fierce competition. Many traditional drivers struggled to make ends meet, with some tragically choosing suicide [2]. The COVID-19 pandemic further exacerbated the situation, driving the industry to near extinction. Although the industry has somewhat recovered, with a current monthly count of around 3 million, it's still dwarfed by more than 18 million trips from its ride-hailing competitors[3].

In this paper, we analyze the proportion of pickups by yellow taxis in relation to the total combination with high-volume for-hire vehicles (HVFHV) across locations in every three-hour interval. We focus on pickups as yellow taxis rely more on street hailing. We don't focus on other competitors as they pose less threat. Green taxis for example is also a traditional street hailing based model which minimal month trips count(around 100 thousand)[3]. Other players like black cars and limousines have existed for long and play on niche market.

Our analysis employs two machine learning approaches (Lasso Regression and Random Forest Regression)[4], which we evaluate for their efficacy in predicting market trends. We hope that the findings from this study provide valuable insights taxi driver and firms as well as governing bodies like the Taxi and Limousine Commission (TLC). We hope to help them form critical decisions, such as whether to gradually phase out this traditional model or support it, depending on its performance and challenges. Ultimately, our analysis seeks to shed light on the evolving landscape of urban transportation in New York.

### 1.1 Dataset

For this analysis, we utilize datasets from TLC [5]. The first is the Yellow Taxi Trip Record Data, which includes detailed records of yellow taxi trips, such as pickup and drop-off times and locations and many trip-related metrics. The second dataset is the High-Volume For-Hire Vehicle (HVFHV) Trip Record Data, which captures similar trip details for ride-hailing services like Uber and Lyft. Given our focus on understanding the current and near-future market dynamics, we specifically use the most recent data available, covering the period from November 2023 to May 2024.

In addition to the trip record data, we incorporate external data sources to enrich our analysis. Weather data is obtained from Meteostat [6], a reliable source for historical weather data. Meteostat provides a Python API for accessing weather information, making it highly convenient for our analysis. We selected a central Manhattan coordinate for weather data collection, assuming that the city's weather patterns are relatively uniform across boroughs.

We also use the Primary Land Use Tax Lot Output (PLUTO) dataset [7], published by the NYC Department of City Planning. The PLUTO dataset provides detailed information about land use across the city, including the number of units, lot sizes, and building classifications for each parcel. We expect that weather patterns and land use data might have a connection to yellow taxis' pickup share.

| **Dataset** | **Instances** | **No. Features** |
|---|---|---|
| TLC Yellow Taxi Trip Record Data | 23,509,182 | 19 |
| HVFHV Trip Record Data | 140,526,989 | 20 |
| Primary Land Use Tax Lot Output | 859,012 | 91 |
| Meteostat data | 5,089 | 11 |

Table 1: Datasets Shape

# 2 Preprocessing

## 2.1 Yellow Taxi Dataset Preprocessing

Several preprocessing steps were necessary to ensure data consistency and relevance for our analysis due to the extensive nature of the Yellow Taxi dataset:

- **Date Range Filtering:** Filtered to include only those with pickup times ranging from November 1, 2023, to May 31, 2024, aligning with our research focus.
- **Removed records with negative fare, tips, fees-related features:** Ensured fare amounts were at least $3 as per TLC regulations.
- **Trip Duration and Distance Validation:** Discarded records with negative or shorter than 0.2 miles distances. Also, removed records with negative or under 1-minute trip durations. These are potentially erroneous and get removed to reflect rational customer behaviour
- **Keep those with rate code ID from 1 to 6** as defined on TLC dictionary
- **Geographical Filtering:** Retained only trips with pickup locations within valid NYC Location IDs (1-263), ensuring the focus remained on relevant urban areas.
- **Extreme Value Handling:** 99.98% of the records have tolls below $28, fare amounts below $220, and total amounts under $262, which are plausible. The remaining are unlikely large so we removed them. Additionally, we filtered out distances over 1,000 miles and durations exceeding 16 hours. These thresholds are enough to cover a round trip from Manhattan to Niagara Falls(an attraction lying on New York's western border). We assume tourists may opt for a long state tour at negotiated prices. Any values beyond these limits are unlikely and were removed. After all removal step above we are left with 20397304 rows.
- **Selecting relevant columns** (pickup datetime, Pick up LocationID) Other features are trips specific while we focus more on what would customers pick under similar circumstances prior

to a trip. However, we still processed them to ensure data validity. From the pick-up times, we extracted the date and hour, We grouped trips by locationID, date and 3-hour intervals (0-3, 3-6, etc.), then get the total trips. As many locations have 0 trips during certain hours. So this helps minimize cases in calculating pick up share such as 0/(0+0) when both taxis and HVFHV counts are zero. By now we are left with 439632 rows and only 4 features

### 2.2 High-Volume For-Hire Vehicle Dataset Preprocessing

The preprocessing steps for the HVFHV dataset were similar but required less work as there are less features and no extreme value found:

- **Date Range Filtering:** Included records with pickup times from November 1, 2023, to May 31, 2024, aligning with the Yellow Taxi dataset.
- **Invalid Records Removal:** Retained records where base passenger fare and driver pay were positive, with trip miles greater than 0.2 and trip time over 1 minute.(same as yellow taxis). By now the HVFHV data is left with 140386155 rows.
- **Retained relevant columns**: `pickup_datetime` and `PULocationID`.
- **Get total number of trips** for each combination of location, date, 3-hour interval.

We now join with the yellow taxi set to calculated percentage of pick up by yellow taxis for each location, date and time bucket. We now have 430757 rows and 4 columns in this aggregated set.

### 2.3 Weather Data Preprocessing

For the weather data sourced from Meteosat, we get data from a coordinate in central Manhattan, for the period from November 1, 2023, to May 31, 2024. We assume that weather patterns are relatively consistent throughout the city. The following features were retained: temperature, relative humidity, precipitation, wind speed, atmospheric, pressure. These features were selected as they represent key weather patterns and also have no null value.

To align this data with our analysis of taxi trips, we grouped the data into 3-hour buckets and in each group we got the mean for each feature. This is for later grouping with the taxi dataset and we also assume that the weather pattern would be similar during such a short interval.

### 2.4 Primary Land Use Tax Lot Output Data Preprocessing

The PLUTO dataset has many features but most are irrelevant (like owner name or height limit) or had significant missing data. We focused on features: building class, lot's assessed value. We also get latitude, longitude and lot's area. These are crucial for understanding the geographic and economic landscape which may affect taxi operations.

We link buildings to taxi zones, discarding those outside the zones. We extract building types from the bldgclass feature, group the data by LocationID, and calculate the proportion of area for each building type and the average land value in each zone. This adjustment is necessary because taxi zones vary in size, and using total building area alone could mislead our analysis of yellow taxi pickup share.

### 2.5 Data aggregation and final preprocessing

We aggregated the data by merging the taxi trip records with the weather dataset and the PLUTO dataset. To ensure that all numerical features were on a consistent scale, they were standardized.

This step is important for models like Lasso Regression, which is sensitive to the scale of the input features. We also create another column: week day, which is derived from date. This final dataset has 35 columns and 430757 rows.

## 3 Analysis and Geospatial Visualisation

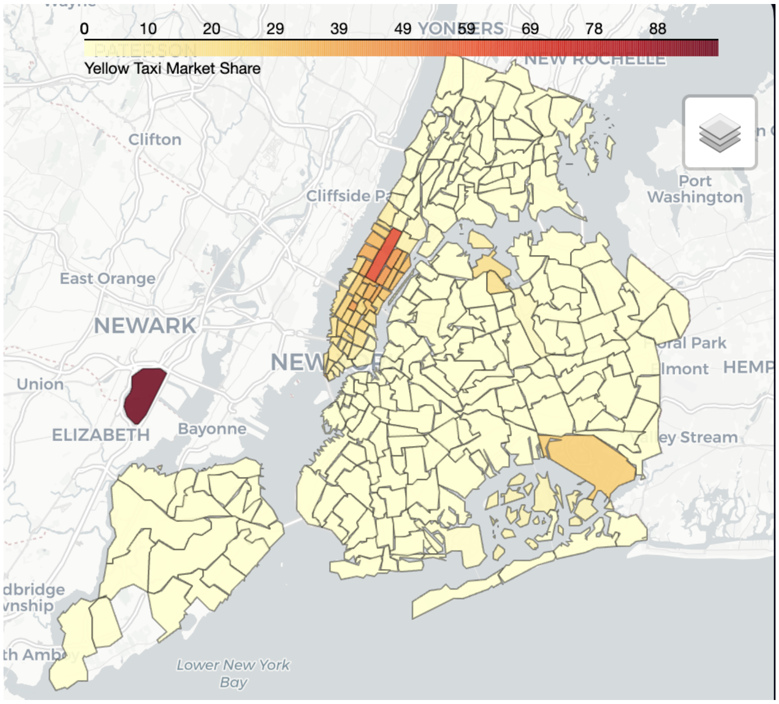

Figure 1: Distribution of yellow taxi's pick-up share in NYC.

### 3.1 Yellow Taxi Pickup Share by Location

Figure 1 clearly illustrates that the distribution of yellow taxi pickup share is heavily influenced by location. While the citywide average share is around 12%, yellow taxis perform significantly better in central Manhattan, with many locations showing a share of over 20% and even exceeding 60%. Yellow taxis also maintain a strong presence at the three airports in New York City (LaGuardia, John F. Kennedy, and Newark Liberty). Conversely, HVFHV services dominate the outer boroughs, where yellow taxi shares drop as low as 0% to 2%.

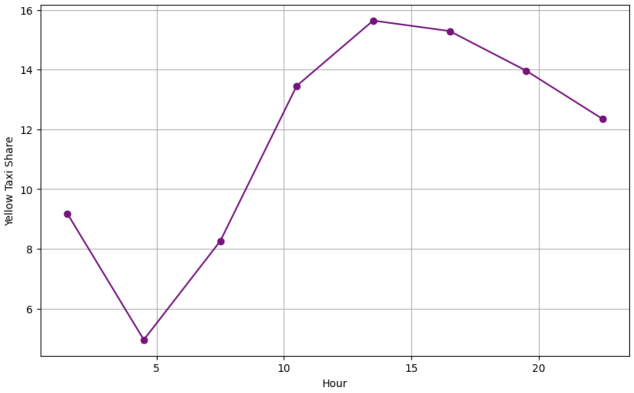

Figure 2: Yellow taxi pick-up share by hour.

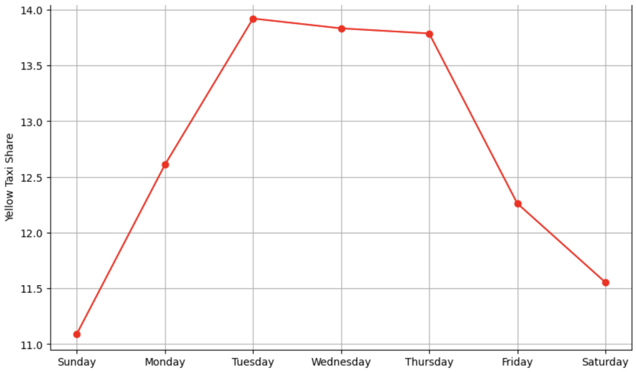

Figure 3: Yellow taxi pick-up share by weekday.

## 3.2 Yellow Taxi Pickup Share Through Time

Figure 2 shows that yellow taxi pick-up share fluctuates significantly throughout the day. It hits its lowest point of under 6% around 4 a.m., then rises to nearly 16% by noon. The share gradually declines to 12% by 11 p.m., before taking a sharp dive after midnight. Perhaps the app-based model of HVFHV allows drivers to connect with customers more easily during late-night hours when fewer people are on the streets.

Figure 3 reveals that yellow taxi share is lowest on weekends, at around 11.5%, and then gradually increases throughout the week, peaking at 14% on Tuesday and plateauing at that level on Wednesday and Thursday before declining toward the weekend. This trend may be due to the bustling activity during weekdays, particularly in areas like central Manhattan where people are concentrated in offices, giving yellow taxis an advantage. On weekends, people might shift to more relaxed activities in the outer boroughs or travel out of the city, making it harder for yellow taxis to attract customers.

### 3.3 Correlation Between Numerical Features of Interest

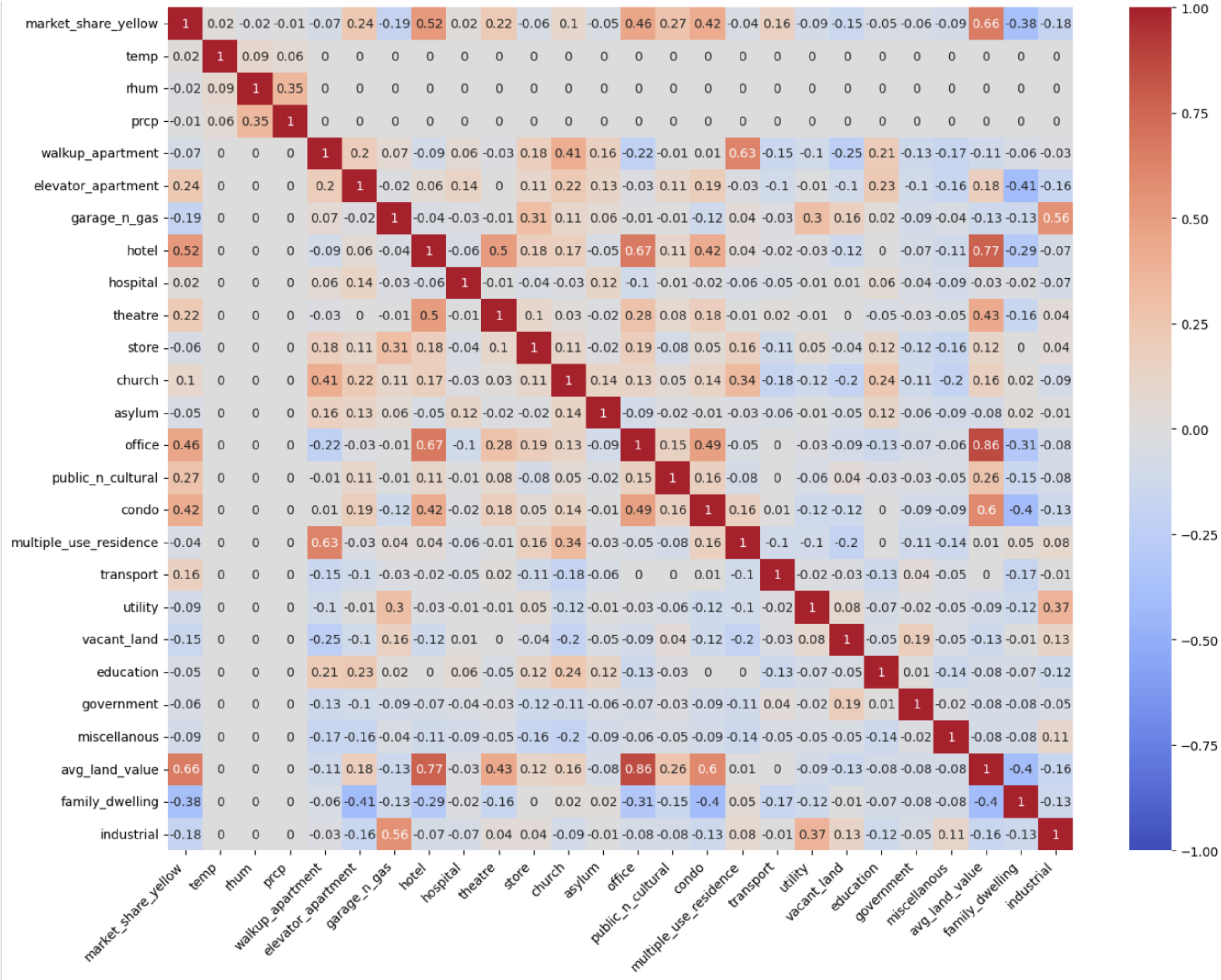

Figure 4: Correlation heatmap of numerical features.

From Figure 4, we clearly see that our target variable, market_share_yellow, has a strong positive linear relationship with the densities of elevator apartments, hotels, theaters, offices, public and cultural buildings, condominiums, and transportation facilities. It also has a strong positive relationship with average land value (0.64). These are typical characteristics of a central urban area with dense and bustling crowds of people, which attract yellow taxis significantly.

On the other hand, the densities of industrial buildings/factories, family dwellings, vacant land, garages, and gas stations seem to have a negative correlation. These are characteristics of the outer borough area with less density, where yellow taxis may struggle to find customers. Weather features seem to have a very small correlation with the target variable; wind speed and atmospheric pressure aren't shown here but have near 0 correlation, so they were removed for clearer visualization. Hospitals (0.01), education (-0.05), and government buildings (-0.06) also have relatively weak correlations, possibly because these facilities are more evenly distributed throughout the city and don't correlate much with yellow taxi share.

# 4 Modeling

We explore two different regression models: Lasso Regression (LR) and Random Forest Regression (RFR), to evaluate their performance in predicting the pick up share of yellow taxis in New York City. We assume that land use data is relatively constant throughout the period. We trained the models on the first 6 months of data(November 2023 to April 2024) and evaluated on May 2024 data. The target variable for our models is the market share of yellow taxis.

## 4.1 Lasso Regression

Lasso Regression is a linear model that is effective for feature selection by enforcing sparsity in the model coefficients through L1 regularization. This model assumes that the features in our model are independent of each other. Categorical features used, including PULocationID, day_of_week, and hour_bucket, were one-hot encoded to capture the non-ordinal nature of these features. We also trained the model on all the other numerical predictors, including weather and land use features.

## 4.2 Random Forest Regression

Random Forest Regression is an ensemble learning method that aggregates multiple decision trees to improve prediction accuracy and reduce variance. It is particularly effective for capturing non-linear relationships between the features and the target variable. It's also robust to feature interdependence, as evident from Figure 4. The model can handle both categorical and numerical features well. However, the Random Forest model is not as easy to interpret. For tuning the hyperparameters, we performed a 3-fold cross-validated grid search on a set of parameters on a subset of the training data. After identifying the best hyperparameters, we used them to train on the full training dataset.

# 5 Discussion

The Random Forest model outperformed the Lasso regression with a lower RMSE (2.96 vs. 4.21) and a higher R-squared (0.93 vs. 0.86), indicating greater accuracy. However, an R-squared over 0.86 still reflects a strong fit. Thus Both models show solid potential for predicting pickup share.

In the Lasso model, significant features include avg land value (35.05) and transport (13), while features like theater, hospital, and locationID are reduced to zero, likely due to intercorrelations with the more informative predictor. Weather-related features are suprisingly retained. This may be since they are independent with land use features and so can contribute independent predictive value. The Random Forest model, on the other hand, emphasizes hour bucket (0.456), walkup apartment (0.249), and weekday (0.065) as key predictors, while avg land value has minimal importance (0.0002). This difference underscores how Random Forest, which capable of capturing complex feature interactions, can have quite different approach to that of Lasso's.

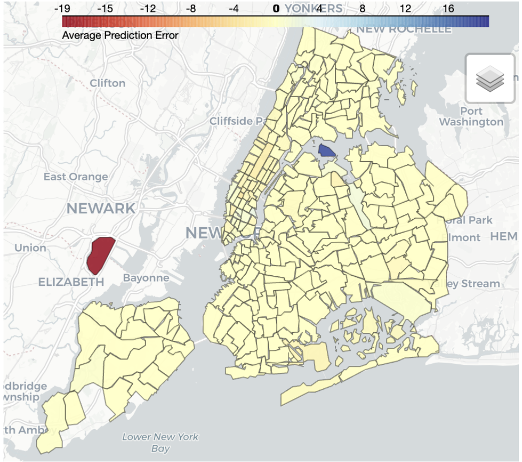

Figure 5: Random Forest model's average prediction error per region

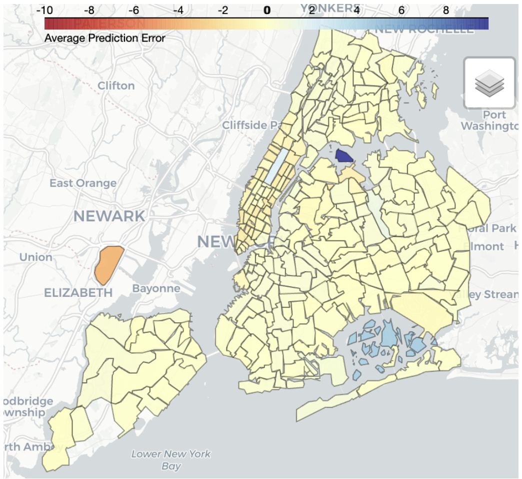

Figure 6: Lasso Regression model's average prediction error per region

Looking at figure 5 and 6, we can see that both models perform relatively well across locations, with average errors at many places close to 0. Random forest models seem to do a better job, while lasso regression seems to overestimate in places like Central Park and Jamaica bay(near JFK airport). Both models seem to underestimate share in Newark Liberty Airport, this may be understandable as this airport has an outstanding yellow cab share (over 80 percent). Both models also hugely overestimate in Riker island's (the notable blue area opposite to La Guardia airport). As the isolated island is notorious for housing NYC's largest jail complex [8]. This distinctive land use characteristic might confuse our models.

## 6 Recommendations and Conclusion

Yellow taxis, while facing significant challenges from HVFHV services, still possess niches where they can effectively compete.

We recommend taxi companies and drivers focus their efforts on areas such as central Manhattan and major airports, particularly during weekdays and midday hours. Our analysis indicates that yellow taxis maintain a competitive advantage in these zones.

Also given that the number of yellow taxis is controlled through the medallion system—unlike the more loosely regulated HVFHV services—there is a risk of increased congestion, reduced control, and diminished income per driver as more driver enter. The medallion system was originally established to prevent these very issues [9]. Therefore, we suggest that the TLC consider stricter regulation to ensure a more balanced competitive environment. Implementing dynamic caps or surcharge systems could help manage the number of HVFHV vehicles in areas where yellow taxis are competitive, while incentivizing their presence in locations where yellow taxis struggle . By expanding our predictive models, these could assist in identifying the specific times and locations where these measures would be most effective.

With strategic adjustments by taxi companies and targeted regulatory support from the TLC, yellow taxis can continue to play a vital role in New York City's transportation network.

# References

1. SpringerLink. ”Rise, Fall, and Implications of the New York City Medallion Market.” Accessed August 25, 2024. Available: https://link.springer.com/chapter/10.1007/978-3-319-95786-9_7

2. Entralgo, Rebekah. ”NYC Taxi Drivers Took on Predatory Lenders — And Won.” Inequality.org, November 12, 2021. Accessed August 25, 2024. Available: https://inequality.org/great-divide/nyc-taxi-drivers-hunger-strike/

3. Medium. ”Introducing the TLC Factbook: NYC TLC’s New Data Dashboard.” Accessed August 25, 2024. Available: https://medium.com/@NYCTLC/introducing-the-tlc-factbook-nyc-tlcs-new-data

4. Models studied from Machine Learning (COMP30027) University of Melbourne.

5. NYC.gov. ”TLC Trip Record Data.” Accessed August 25, 2024. Available: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

6. Meteostat. ”Global Weather and Climate Data.” Accessed August 25, 2024. Available: https://meteostat.net/en/

7. NYC Open Data. ”Primary Land Use Tax Lot Output (PLUTO).” Accessed August 25, 2024. Available: https://data.cityofnewyork.us/City-Government/Primary-Land-Use-Tax-Lot-Output-PLU
64uk-42ks/data

8. Untapped New York. ”The Top 10 Secrets of Rikers Island, NYC’s Main Jail Complex.” Accessed August 25, 2024. Available: https://untappedcities.com/2016/10/21/the-top-10-secrets-of-rikers-

9. Columbia Human Rights Law Review. ”Distressed Drivers: Solving the New York City Taxi Medallion Debt Crisis.” Accessed August 25, 2024. Available: https://hrlr.law.columbia.edu/hrlr-online/distressed-drivers-solving-the-new-york-city-taxi-medallion-debt-crisis/