New York City: Taxi or Rideshare

Team46

Benjamin Falby, Zhantong Mao, Ning Kang, David Weon, Jair Pisani

Abstract: This project's goal is to aid budget-conscious New York City travelers with the choice between taxi or ride-sharing services. Consumers are not typically privy to the information regarding how certain factors may influence the total price of a taxi or rideshare. By discovering and analyzing these impacting factors, the project aims to provide budget-conscious travelers with a clearer understanding of their trip cost. In doing so, travelers can make better-informed decisions on how much they pay for a ride.

1. Overview of Project

1.1 Background Information

Over the last decade, popular ride-sharing companies have caused quite a disruption to the taxi market. In fact, ever since the introduction of ride sharing apps, there has been a steady decline of New York City taxis. In addition, the published article, "The Success of Ridesharing," attributes the meteoric rise of companies such as Uber and Lyft to the success and innovation of technology integration into their apps. Another article titled "The Geography of Ridesharing" says that ridesharing is particularly popular in New York given the city's widespread attraction for travelers. The increasing popularity of ride sharing seems to suggest the end of the New York City taxi. For the budget-conscious traveler in New York City, this may become a problem.

1.2 Business Justification

Rideshare companies are known to implement a system called surge pricing. According to Uber's website, "in these cases of very high demand, prices may increase to help ensure that those who need a ride can get one." Given the recent market dominance and popularity of rideshares in New York City, consumers are more likely to face surge prices when selecting a rideshare due to increased demand; this would present a financial stress on the budget-conscious traveler. The New York City taxis pose as a cheaper alternative since they do not have a surge pricing system – but is there a way to definitively tell through the available data? What about when there is no surge pricing? Are there other factors that play an important part in the total price of a ride? Without a level of supporting information, the budget-conscious traveler would be taking a random guess as to which is cheaper between a New York City taxi or rideshare. Our goal is to focus on these cost-conscious consumers and provide them with a better understanding of their trip prices. We want to provide users with this information and enable them to make the best choice financially. We want our model to satisfy a market segment that requires greater cost transparency as they consider a decision between their options.

1.3 Analysis Approach - Hypothesis

The goal of this analysis is to discover how comparable New York City taxis and rideshares are in terms of trip price. In order to determine comparable trip prices, key factors that impact a ride price will be identified and implemented into a regression model. Our initial hypothesis states that New York City taxis will be the cheaper option given the trip occurs in an area of high traffic due to surge pricing. When the ride occurs outside of surge pricing, we anticipate the trip prices will be generally the same.

2. Data

2.1 Data Sets

Four separate datasets were used for the project. First among our datasets was the New York City taxi trip data that included rides for both Yellow cabs and Green cabs – Green cabs serve areas typically not served by Yellow cabs (i.e. outside Manhattan). The second was the For-Hire-Vehicle (FHV) trip data, which had the rideshare trip information for those tracked by the Taxi & Limousine Commission of New York City. Third was the New York City zone data, which mapped the location IDs, as seen in the New York City taxi and FHV datasets, to the names of the corresponding neighborhood or district in New York City. Fourth was the weather data in New York City, which included measurements for temperature, precipitation, cloud coverage, and windspeed per day.

Initially, we narrowed the scope to July 2022 due to the size of the trip datasets. Our discussions led us to believe that one of the summer months would have the highest volume of rides compared to the rest of the year due to several factors, including tourism and better weather for traveling. However, as we continued with our analysis, this second reason was discarded. Although we only kept the July 2022 data for this project's analysis, any other month could have been selected, keeping in mind the file size of the trip datasets for 2022.

2.2 Data Cleaning, Key Variables, and Feature Engineering

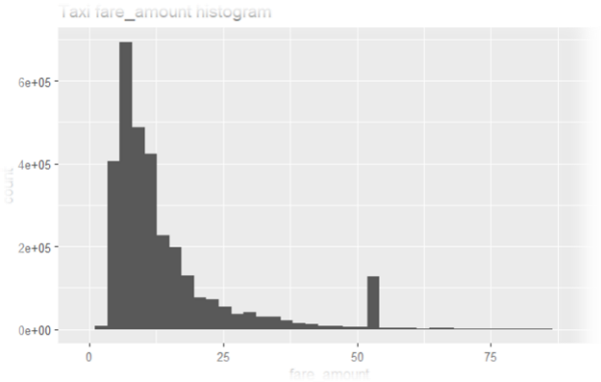
With the July sets selected, we standardized the dates and times for both taxis and For-Hire Vehicles (FHV), created new factors like day of the week and trip duration, and renamed others like trip distance and fare amount so that predictor names were consistent between the two types of rides. We narrowed in on what we expected would be our chief independent variables impacting price from the ride service datasets: trip distance, trip duration, day of the week, pick up and drop off hour and pick up and drop off location. Some residual cleaning included filtering out N/A values and restricting trips to those with positive distances and fare amounts.

From the weather data set, we homed in on temperature, rain, and wind as our potentially impactful independent variables. We ended up abandoning wind due to the number of N/A values present in the data set.

The complexity of our dependent variable, price, lead to unexpected discoveries. When we first encountered what we thought would be our final results, the taxi appeared to be the superior price option for nearly every set of circumstances considered, 99.13% of cases on a test set. Surprised by the results, we revisited the cleaning of our data and an exploration of what elements were contributing to the fares for both taxis and rideshares. Simple averages of the originally named fare_amount predictor for the taxis and base_passenger_fare for the For-Hire-Vehicles hinted that our taxi prices were missing costs. Our first finished models reinforced our concerns that the fare comparisons needed refining at the data cleaning level. Reconsidering the variety of options for fare comparison, the originally named total_amount feature for taxis captured prices that had simply been rolled into the comparable FHV base_passenger_fare. We made the adjustment to the total_amount feature as our new basis for fare comparison, taking care to subtract the included tip_amount feature so that we would have an apples-to-apples comparison.

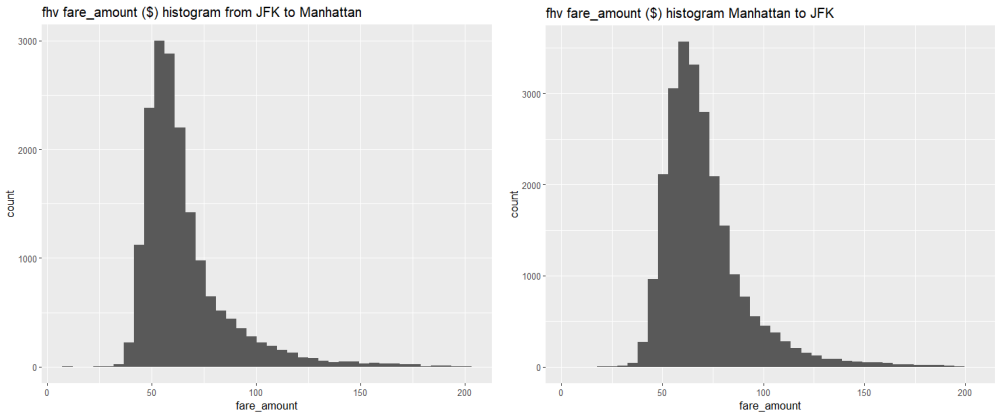
2.3 Additional Exploratory Data Analysis Insights

We found a remarkable uptick in price when a trip's Destination ID did not exist. We created an additional factor to account for unknown locations if we needed to filter them out or address them otherwise. Upon further exploration, we discovered that this designation indicated a trip that ended outside of New York City, and 4.2% of trips had an unknown drop off location. A much larger 20% of trips, those with the specific JFK airport pickup location, had drop off locations outside NYC. In general, JFK trips posed several unique challenges.



One such idiosyncrasy showed up in a histogram for the fare_amount feature. We encountered an unexpected spike at $52. This discovery warranted further domain exploration. A quick perusal of the data suggested the $52 fares were tethered to locationID 132. In fact, 94% of the $52 fares involved that precise location. Research revealed that NYC taxis have a flat rate from or to JFK airport (locationID 132). This flat rate was exactly $52 in July of 2022 (It has since been raised to $70). We verified this domain discovery in our data and removed these trips (1.8% of the total). Were these trips considered, the assumptions of independence of errors and homoscedasticity (constant variance) of errors would not have held. The $52 flat rate would have created a variance in our regression model that would have been unlike any other variation associated with other fare amounts.

An interesting side note: In July of 2022, a taxi was often discernibly cheaper than the For-Hire-Vehicle (FHV) option when the departure or destination location was JFK airport. The histograms below generally illustrate the volume of FHV trips that exceeded $52. While we removed these observations in anticipation of our linear regression, our exploratory data analysis revealed a situational answer to our primary question. For this location, the taxi was the cheaper option for a budget-conscious traveler. Of course, the aforementioned price-hike occurred within 6 months of this observation.

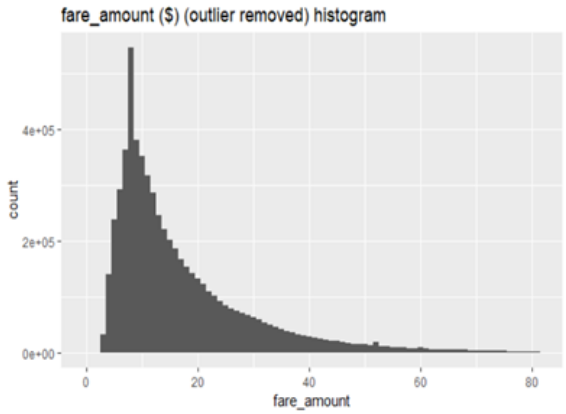


2.4 Outlier Analysis

Outlier analysis was run on combined Taxi and FHV data. Each type of ride had been randomly sampled to 3 million trips for the combined set. We investigated the distribution for the count of a variety of key features, among them our dependent variable: price and two of the independent variables: trip distance and trip duration.

For each of these variables, we explored removing outliers by Tukey Fences and quantiles, considering the relative effects of each approach on skewness. With little information available about the most extreme values, we opted to use the 99th percentile as our upper criteria. We leaned towards removing proportionally fewer observations.

After JFK airport flat rate trips were deleted, fare_amount presented a positive skew (5.14), a 99th percentile of $81.46 and a maximum value $1,267. This suggested the presence of outliers in our data. The fare_amount histogram with outliers removed appears below.



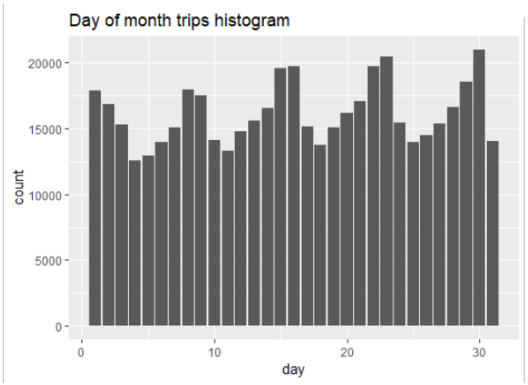
Trip distance also presented a positive skew (329.28), a 99th percentile of 22.5 miles and a maximum value of 290,250 miles. These results dramatically indicated the presence of outliers and possible errors in our data. A histogram illustrating trip distance with outliers removed appears above to the right. Again, the 99th percentile is the upper criteria.

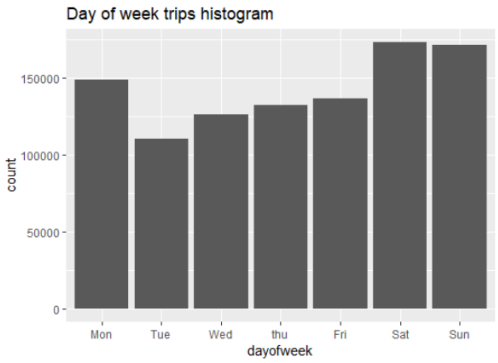
Finally, trip duration presented a positive skew as well (35.14), a 99th percentile of 60.5 minutes and a maximum value of 5,634 minutes, about four days. This, too, indicated the presence of outliers in our data. The image below illustrates the histogram after outliers were removed.

In all, we removed outliers for fare amount, trip distance and trip duration using the 99th percentile as our upper outlier criteria. The removed data represented 1.9% of our dataset. After outlier removal, the skewness of fare amount, trip distance and trip duration changed to a far more acceptable 1.75, 2.04 and 1.31, respectively. We also removed 980 trips with zero fare amount and 480 trips with zero trip duration. These represented 0.02% of the total trips in our combined taxi/for hire vehicle dataset

2.5 Peak Usage Exploratory Analysis & Additional Engineered Features

We expected peak usage would play a role in our final model and did preliminary work to explore its impact on price. To that end, we separated our dataset (all_cab_fhv) into groups that only included taxi rides and only included rideshares. While doing so, we reduced our dataset sizes for practical reasons, taking a random sample of 500,000 observations for each. We then recombined our data to do this additional data exploration. We wanted to see if there were any seasonal elements to demand.





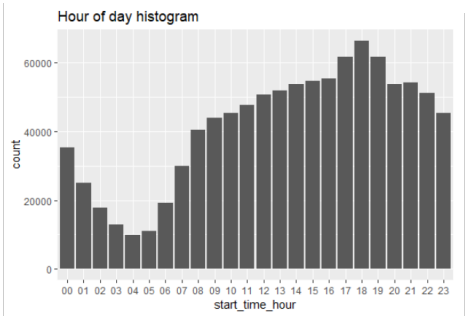
Weekly seasonality was observed in the number of trips. As this could impact our model, we conducted exploratory analysis to develop new insights for our data. We regressed fare_amount on the day of the week.

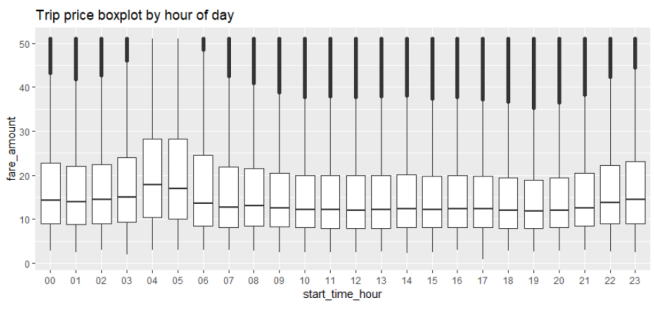
```
Call:
lm(formula = fare_amount ~ dayofweek, data = x)

Residuals:
    Min      1Q  Median      3Q     Max
-15.724  -7.591  -3.254   4.811  35.437

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.59428    0.02602  637.78   <2e-16 ***
dayofweekTue -0.75530    0.03989  -18.93   <2e-16 ***
dayofweekWed -1.00359    0.03840  -26.14   <2e-16 ***
dayofweekthu -1.03149    0.03791  -27.21   <2e-16 ***
dayofweekFri -0.83038    0.03761  -22.08   <2e-16 ***
dayofweekSat -0.66067    0.03548  -18.62   <2e-16 ***
dayofweekSun -0.34036    0.03557   -9.57   <2e-16 ***
```

```
Call:
lm(formula = fare_amount ~ isMonSun, data = all_cab_fhv)

Residuals:
    Min      1Q  Median      3Q     Max
-15.542  -7.598  -3.248   4.772  35.252

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.74788    0.01218 1292.53   <2e-16 ***
isMonSunTRUE  0.66427    0.02152   30.86   <2e-16 ***
```

Monday was, on average, the most expensive day. We chose to include a new dummy variable in our dataset indicating if the trip took place on the highest volume day. After another simple linear regression, we saw that the trips, on average, contributed $0.66 to the fare. Attuned to the possibility of correlation with other important variables, we expected to complete further VIF analysis to determine whether the engineered feature would remain in our final model.

Searching for more time-related price variations, we plotted the hour of day (0:00 to 24:00 format) in a histogram too.

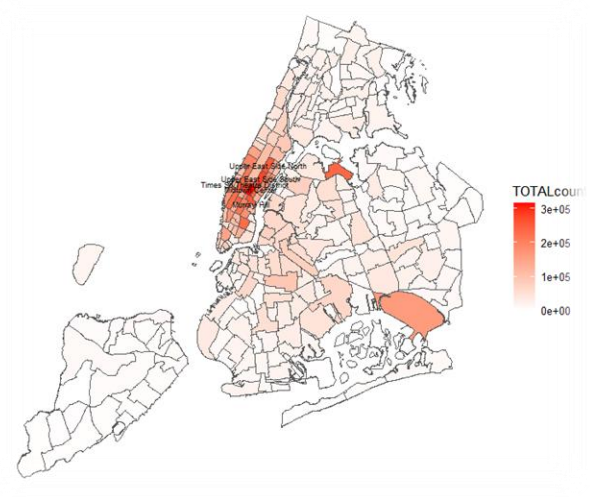




We ran a simple linear regression to identify the most expensive hours on average. There was a clear relationship between the trip price and hour of day, but the boxplot showed us that the relationship was neither linear nor loglinear. We recognized through the boxplot and linear regression that trips from 10pm to 6am were, on average, more expensive. Therefore, we included the new dummy variable (isFrom10pmTo6am) in our dataset, indicating if the trip was between 10pm and 6 am. After another simple linear regression, we saw that trips at these hours

were, on average, $1.92 more expensive. As with the day of week analysis, further VIF analysis would be conducted before using this new variable on our final model.

Our final peak-usage consideration surrounded locations with the highest demand. We grouped and counted trips by pick-up location, drop-off location and calculated their combined total. The map below shows locationIDs by demand, a visual representation of ride density. We experimented with the creation of both an indicator variable marking membership among the top trip locationIDs, those that counted for more than 1% of the total rides, and a numeric variable representing each location as a percentage of the total rides. Ultimately, VIF values in later analysis suggested that we select only one of the two variables for our final model. We opted for the percentages indicating the portion of rides represented by a given location.



3. Modeling

3.1 Model Selection and Tuning

As our goal has been to find the cheapest price among different providers, we created two separate linear regression models that allowed us to predict the price of taxi trips and rideshare trips in the city. We again split the combined data set into two frames, one for taxis and one for rideshares, and subdivided those frames into training and test sets.

While we eventually settled on fare amount as our dependent variable (mentioned above), it is worth a short digression to mention our consideration of other potential, price-related dependent variables, namely fare amount per distance traveled and fare amount per trip duration. We entertained the possibility that those unit measurements might smooth out variations between different trip lengths and durations. It quickly became clear that engineering that dependent variable reduced the interpretability of our feature coefficients. We returned to the simpler fare amount as dependent variable and committed to greater interpretability.

While linear regression was certainly our primary means for predicting pricing, we also performed log-linear transformations to explore nonlinear effects and search for superior representations of the data. Among the For-Hire-Vehicle (FHV) models, the highest performers were evaluated by adjusted R-squared. Of those results, a linear model had an adjusted R-squared of 0.8399 while the best log linear model had an adjusted R-squared 0.7963. Both the linear and the best log-linear models yielded features with high significance.

```
lm(formula = fare_amount ~ isFrom10pmTo6am + isMon + trip_duration +
    trip_distance + shared_match_flag + access_a_ride_flag +
    wav_match_flag + temperature + rain + Percentage, data = fhv_train)

Residuals:
    Min      1Q  Median      3Q     Max
-22.670  -2.719  -0.825   1.413  57.639

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.197803   0.074460   29.52   <2e-16 ***
isFrom10pmTo6amTRUE  1.380844   0.028114   49.12   <2e-16 ***
isMonTRUE           0.253071   0.022923   11.04   <2e-16 ***
trip_duration       0.504020   0.001381  365.06   <2e-16 ***
trip_distance       1.807222   0.003566  506.75   <2e-16 ***
shared_match_flagY  -5.618146   0.234275  -23.98   <2e-16 ***
access_a_ride_flagN -1.977845   0.018893 -104.69   <2e-16 ***
wav_match_flagY     -0.487793   0.035093  -13.90   <2e-16 ***
temperature          0.041089   0.002700   15.22   <2e-16 ***
rain                -0.767550   0.027630  -27.78   <2e-16 ***
Percentage           2.135040   0.012654  168.72   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.342 on 400460 degrees of freedom
Multiple R-squared:  0.8399,    Adjusted R-squared:  0.8399
F-statistic: 2.101e+05 on 10 and 400460 DF,  p-value: < 2.2e-16
```

```
lm(formula = ln_fare_amount ~ isFrom10pmTo6am + isMon + trip_duration +
    trip_distance + shared_match_flag + access_a_ride_flag +
    wav_match_flag + temperature + rain + Percentage, data = fhv_train)

Residuals:
     Min       1Q   Median       3Q      Max
-1.50953 -0.17940 -0.00291  0.14960  1.53033

Coefficients:
                     Estimate Std. Error  t value Pr(>|t|)
(Intercept)         2.044e+00  3.645e-03  560.736   <2e-16 ***
isFrom10pmTo6amTRUE  7.572e-02  1.376e-03   55.027   <2e-16 ***
isMonTRUE           2.436e-02  1.122e-03   21.710   <2e-16 ***
trip_duration       2.831e-02  6.758e-05  418.853   <2e-16 ***
trip_distance       5.775e-02  1.746e-04  330.845   <2e-16 ***
shared_match_flagY  -2.014e-01  1.147e-02  -17.562   <2e-16 ***
access_a_ride_flagN -9.979e-02  9.248e-04 -107.907   <2e-16 ***
wav_match_flagY     -1.561e-02  1.718e-03   -9.088   <2e-16 ***
temperature          1.526e-03  1.322e-04   11.542   <2e-16 ***
rain                -3.768e-02  1.352e-03  -27.859   <2e-16 ***
Percentage           9.366e-02  6.194e-04  151.203   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2615 on 400460 degrees of freedom
Multiple R-squared:  0.7963,    Adjusted R-squared:  0.7963
F-statistic: 1.565e+05 on 10 and 400460 DF,  p-value: < 2.2e-16
```

Furthermore, our pre-tuned linear model showed a Cook’s distance that yielded no additional outliers.

```
> cook_dist <- cooks.distance((fhv_model))
> influentials <- cook_dist[cook_dist>1]
> influentials
named numeric(0)
```

Variance Inflation Factors for all predictors were less than five, indicating desirably low correlation between the independent variables.

```
> vif(fhv_model)
  isFrom10pmTo6am            isMon      trip_duration      trip_distance  shared_match_flag access_a_ride_flag     wav_match_flag        temperature
         1.050793           1.012723           2.942006           2.918646           1.000843           1.003178           1.006038           1.050506
             rain         Percentage
         1.007080           1.028376
```

Considering the superior adjusted R-squared, the presence of no additional outliers and the low VIF values, we opted to pursue a tuned linear model as our final means for price prediction.

For the final models, we selected the following features: isFrom10pmTo6am, isMon, trip_duration, trip_distance, temperature, rain, and Percentage. Those were significant common factors for the FHV and Taxi models. We decided to remove the various ride flag fields unique to the ride share data. They had included reductions in price for shared rides where a requesting consumer was matched with another rider (shared_match_flagY), an increase in price if the ride was requested by the Metropolitan Transit Authority (note: the access_a_ride_flagN presented as a decrease if the situation was not true, a double negative!), and a decrease in price when a wheelchair accessible vehicle was matched (wav_match_flagY). While significant and impactful, we preferred the transparency and interpretability of exclusively common factors. We also introduced indicator variables to both models so that we could draw distinctions between companies and cab colors in the final results. For the FHV model, an Ind_Uber indicator marked a trip as Uber-provided or not (otherwise Lyft provided). For the Taxi, a cab_color factor marked a trip as yellow cab or not (otherwise green cabs, concentrated in outer boroughs).

3.2 Model Performance

At this point, we had completed our model tuning and training. The For-Hire-Vehicle (FHV) model produced an adjusted R-squared of 0.8401, a slight improvement after trimming the flag features from our collection of predictors. The taxi model came in with a superior adjusted R-squared of 0.9214, in line with our expectation of greater consistency for its price schedule. The initial summaries suggested both were more than suitable for achieving our aim.

```
lm(formula = fare_amount ~ isFrom10pmTo6am + isMon + trip_duration +
    trip_distance + temperature + rain + Percentage + Ind_Uber,
    data = fhv_train)

Residuals:
    Min      1Q  Median      3Q     Max
-23.159  -2.716  -0.821   1.412  57.673

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.187312   0.075451   2.483    0.013 *
isFrom10pmTo6amTRUE  1.380393   0.028125  49.080   <2e-16 ***
isMonTRUE         0.255123   0.022931  11.125   <2e-16 ***
trip_duration     0.504297   0.001381 365.137   <2e-16 ***
trip_distance     1.808972   0.003563 507.711   <2e-16 ***
temperature       0.040638   0.002701  15.047   <2e-16 ***
rain             -0.769150   0.027636 -27.832   <2e-16 ***
Percentage        2.125679   0.012642 168.150   <2e-16 ***
Ind_Uber          1.989772   0.018893 105.320   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.34 on 399941 degrees of freedom
Multiple R-squared:  0.8401,    Adjusted R-squared:  0.8401
F-statistic: 2.627e+05 on 8 and 399941 DF,  p-value: < 2.2e-16
```

```
lm(formula = fare_amount ~ isFrom10pmTo6am + isMon + trip_duration +
    trip_distance + temperature + rain + Percentage + cab_color,
    data = allcab_train)

Residuals:
    Min      1Q  Median      3Q     Max
-26.675  -1.311   0.106   1.009  58.377

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.9255214  0.0540029  54.173  < 2e-16 ***
isFrom10pmTo6amTRUE -0.1158059  0.0171704  -6.744 1.54e-11 ***
isMonTRUE         -0.3301827  0.0142081 -23.239  < 2e-16 ***
trip_duration      0.3661769  0.0008736 419.170  < 2e-16 ***
trip_distance      2.5324116  0.0026047 972.241  < 2e-16 ***
temperature        0.0472996  0.0015171  31.177  < 2e-16 ***
rain               0.0874740  0.0152865   5.722 1.05e-08 ***
Percentage         0.3554949  0.0082353  43.167  < 2e-16 ***
cab_coloryellow    1.5353994  0.0356445  43.075  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.027 on 400149 degrees of freedom
Multiple R-squared:  0.9214,    Adjusted R-squared:  0.9214
F-statistic: 5.862e+05 on 8 and 400149 DF,  p-value: < 2.2e-16
```

After training with tuning, we fed the test data to our models with common columns plus their respective indicators for company and color. We used our models to predict prices we could expect from each ride type given the values of our test observations. The head of that predicted output is shown below. Note that for each observation, we identify which service provides the minimum price.

	isFrom10pmTo6am	isMon	trip_duration	trip_distance	temperature	rain	Percentage	Uber_price	Lyft_price	Taxi_Yellow_price	Taxi_Green_price	Min_Price	MIN_Price_Provider
1	FALSE	FALSE	16.63333	6.07	24.0	0	0.4842611	23.55038	21.56060	27.23074	27.23074	21.56060	Lyft_price
2	FALSE	FALSE	10.68333	1.47	23.7	0	1.4764315	14.32539	12.33561	13.74142	13.74142	12.33561	Lyft_price
3	FALSE	FALSE	12.65000	1.57	24.6	0	0.5007317	13.46062	11.47084	14.41052	14.41052	11.47084	Lyft_price
4	FALSE	FALSE	14.00000	5.62	24.6	0	0.8215801	22.14977	20.16000	25.27519	25.27519	20.16000	Lyft_price
5	FALSE	FALSE	43.66667	4.25	23.0	0	2.3038879	37.71817	35.72840	33.12031	33.12031	33.12031	Taxi_Yellow_price
6	TRUE	FALSE	16.68333	2.57	26.9	0	0.6274844	19.04688	17.05711	18.45789	18.45789	17.05711	Lyft_price

4. Model Analysis, Evaluation and Conclusion

4.1 Model Analysis and Statistical Values

Our final models only included statistically significant coefficients, and all our factors (save for the intercept of the For-Hire-Vehicle (FHV) model) were significant to a 99.9% confidence level.

Diving into the differences between these significant coefficients, the “hotspot” factor, designed to capture the percentage of rides involving a particular location, contributed appreciably more to the price of the FHV rides than the Taxi rides. Trip distance contributed more to the Taxi price as distance rose than to the FHV price, but both models showed their dependent variables deriving much of their value from the feature. Where distance highly impacted taxi price, there was less of a positive price-impact from trip duration. The FHV model showed a higher duration coefficient, indicating a larger contribution, effectively pushing the prices closer together again. Distance and duration were related but VIF-distinct features for our models.

During the lower trip-volume times from 10am to 6pm, price increased in the FHV model and decreased for the Taxi model. With the low volume and increased price, there was an insinuation that FHV supply marginally tightened compared to demand during the overnight hours. Taxi prices appeared to be more stable. During our

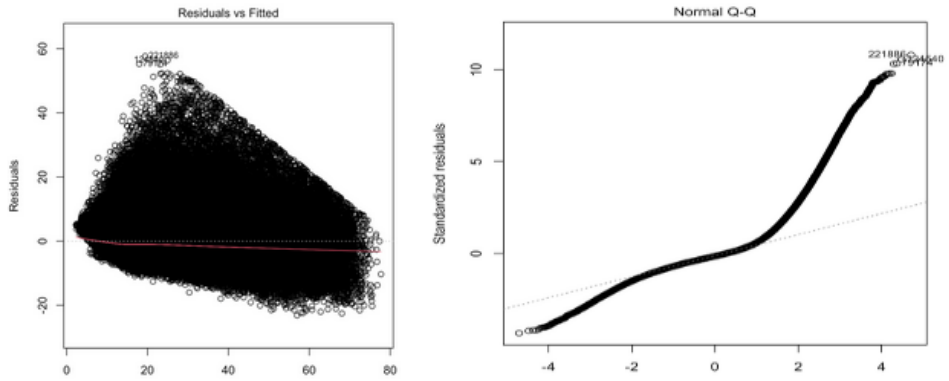
highest trip-volume day, the FHV model had a positive coefficient (isMonTrue) while our Taxi model had a negative one.

Weather-related factors had limited effect on the price for each model save for rain's impact on the FHV price. In this case, it decreased the ride price by $0.77. As for the company and color variables. Uber contributed more to the model than Lyft. Yellow taxis contributed more than Green.

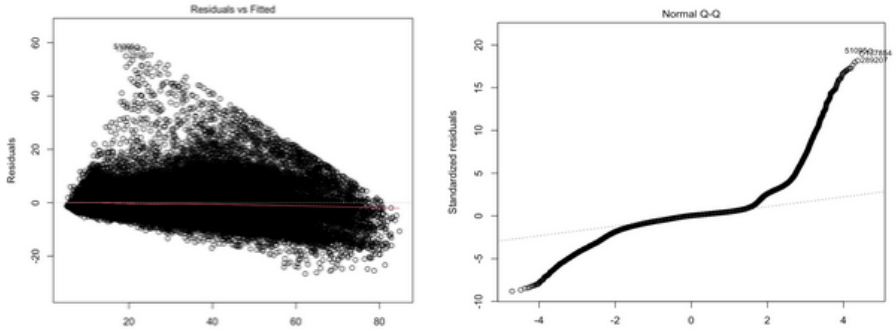
4.2 Evaluation of Model

We extensively discussed residual and QQ plots for our final models. While the residual plots adhered to our "formless cloud" expectations, the QQ plots of the residuals did not present a normal distribution.

For-Hire Vehicle residual and QQ-plots:



Taxi Residual and QQ-plots:



Here, we leaned heavily into the size of our dataset. Our sample was comfortably above four hundred thousand trips, thereby mitigating the centrality of the normality assumption. We were not as concerned with the non-normal distribution of residuals because other linear regressions assumptions were holding.

As articulated in the paper, "Assumptions of Multiple Regression: Correcting Two Misconceptions", "There is a common misconception of the necessity of the assumption of Normal distribution of residuals for linear regression. In fact, if the assumptions of zero mean of residuals, independence of residuals and homoscedasticity of residuals holds, then the linear regression coefficient estimates will be unbiased, consistent, and efficient, even though the residuals are not normally distributed." Williams, Grajales and Kurkiewicz go on to say, "Normally distributed errors are not required for regression coefficients to be unbiased, consistent, and

efficient (at least in the sense of being best linear unbiased estimates) but this assumption is required for trustworthy significance tests and confidence intervals in small samples (Cohen et al., 2003). The larger the sample, the lesser the importance of this assumption."

Seeking a variety of ways to assess the validity of our factor-based models, we curiously decided to compare our taxi model to advertised fare information provided by the New York Taxi & Limousine Commission. We were pleased to see that our $2.92 intercept coincided nicely with the advertised $3.00 initial charge. The per mile calculation that the Commission uses is slightly more complicated than a direct comparison permits because their $.70 price is assessed per 1/5 mile when taxi speeds exceed 12 miles per hour and per 60 seconds when speeds fall below that threshold. Our taxi per mile price was also a cocktail of distance, duration, and other factors. Still, a gross comparison revealed that the Commission sets the taxi price at roughly $3.50 per mile. And our per mile worked out to $2.53 plus the charges associated with trip duration and other factors.

This comparison with the Commission's fare information provided a validation of the model's accuracy in capturing the base fare and per mile pricing components of taxis, supporting the validity of the factor-based modeling approach used in the project. It also reinforced the reliability of the model in providing insights into the cost of taxi rides, further bolstering confidence in the project's findings and conclusions.

4.3 Conclusion and Key Takeaways

We originally hypothesized that taxis may indeed be the superior option for a budget-conscious traveler under defined circumstances. Specifically, we suspected that New York City taxis could be the cheaper option in areas of high demand due to For-Hire-Vehicle use of "surge pricing." Our "hot spot" exploration reinforced our suspicion. Our location demand feature, percentage, made a comparably high contribution to total trip cost for the ride share trips. Controlling for other factors, taxis can be the cheaper option when a person is seeking a ride service in a high demand area.

This came with a significant caveat. Trip distance played an outsize role in determining ride price for both of our models. The taxi per mile contribution was recognizably higher. Again, controlling for other factors, the longer the trip, the more often the For-Hire Vehicle option would provide the superior price. As noted, hotspots matter when anticipating a ride price, but they matter less as trips get longer.

Given each model's descriptive strengths, notably demonstrated by their adjusted R-squared values, we are confident the created models can identify the cheaper ride when provided a set of circumstances.

5. Future Considerations

There were several avenues we could pursue to deepen the discoveries and utility afforded by our initial model. The success of our modeling effort for July of 2022 suggests a straightforward extension to other months. In turn, this extension to additional months would present opportunities to consider additional factors (snow) or a broadening range for existing factors (cold temperatures).

Sorting out the peculiarities of pricing for JFK airport yielded us a wealth of new domain expertise. Because of the effects its inclusion would have had on the homoskedasticity of our models, we removed it. Creating companion models that explored JFK would be another way to expand the project's application.

An aspirational deliverable: we also thought the project lent itself well to the creation of a simple phone app that could suggest the cheaper option given value inputs for our various predictors.

References

1. Disruptive Change in the Taxi Business: The Case of Uber. J. Cramer, A. B. Krueger. NBER Working Paper No. 22083, 2016

2.Taxi Demand Prediction Based on a Combination Forecasting Model in Hotspots: Journal of Advanced Transportation Volume 2020, Article ID 1302586

3. The Geography of Ridesharing: A Case Study on New York City, C. T. Lam, et al. Information Economics and Policy, 57, 2021

4. TLC (Taxi & Limousine Commission) Trip Record Data, NYC Taxi & Limousine Commission, City of New York, https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

5. NYC Taxi Zones, Data Catalog City of New York, NYC Open Data, https://catalog.data.gov/dataset/nyc-taxi-zones

6.NYC Weather – 2016 to 2022, Copernicus Climate Change Service, Kaggle, https://www.kaggle.com/datasets/aadimator/nyc-weather-2016-to-2022?select=NYC_Weather_2016_2022.csv

7. New York City Approves Taxi Fare Hike, Raising Average Fare 23%, Ana Ley, New York Times 11/17/22, https://www.nytimes.com/2022/11/17/nyregion/taxi-fare-hike-nyc.html

8. How Surge Pricing Works, Uber, https://www.uber.com/us/en/drive/driver-app/how-surge-works/#:~:text=Prices%20go%20up,to%20be%20a%20reliable%20choice

9. Taxi and Ridehailing Usage in New York City, Todd Schneider, https://toddwschneider.com/dashboards/nyc-taxi-ridehailing-uber-lyft-data/

10. Williams, Matt N.; Grajales, Carlos Alberto Gomez; and Kurkiewicz, Dason (2013) "Assumptions of Multiple Regression: Correcting Two Misconceptions," Practical Assessment, Research, and Evaluation: Vol. 18, Article 11. DOI: https://doi.org/10.7275/55hn-wk47 Available at: https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1308&context=pare

11. NYC Taxi & Limousine Commission Taxi Fare descriptions, https://www.nyc.gov/site/tlc/passengers/taxi-fare.page