Predicting Fare Prices and Proportion of Trips for Green Taxis in NYC

Demands and Average Fare Prices in Different Boroughs

Park Chang Whan
Student ID: 1129623
Github repo with commit

August 21, 2022

1 Introduction

Due to the Covid-19 pandemic, the taxi industry has taken a big hit with all the restrictions that were set in place. Recovery from the pandemic-driven collapse has been significantly slow for taxis. The lowest recovery made was from green taxis, recovering only 42% of vehicles licensed to operate in the city [1]. In addition, the rise of app-based ride-hailing services in New York has further made it harder for green taxi drivers to earn an income, eventually forcing them to quit. Thus, green taxi drivers ideally need not be wasting time idling around NYC but rather drive around in locations where they are most likely to get customers to sustain a steady income.

In an attempt to help green taxi drivers maximize their profits, this report focuses on finding demands and fare prices in different boroughs of NYC (excluding Newark Airport) targeted towards drivers of green taxis post-pandemic. Two linear regression models were tested in order to find the best estimates for finding fare prices and proportion of total trips in NYC given borough location, day of week, shift and temperature. Using the linear regression model chosen, a driver will be able to find the average fare and demand they would expect in a trip given the temperature forecast for the day, their shift, day of the week and the borough chosen.

2 Data

The datasets chosen were green taxi trip records from 1st July 2021 to 30th April 2022. The range was chosen specifically for post-pandemic period which is defined as when all restrictions relating to the pandemic was lifted. As the New York City Mayor Bill de Blasio announced that New York is reopening in 1st July 2021, the chosen start range would be 1st July 2021 [2]. As the data was only released to the month of April in 2022, that will be the end range of data used. In all the datasets retrieved, only features relevant for analysis will be extracted from the them, explained in Section 4.

2.1 Green Taxi Data

Green taxi data was chosen as it is our main area of interest. Data was retrieved from The New York City Taxis and Limousine Commission(TLC) [3]. Data consists of individual trips recorded by green taxis which includes pick-up and drop-off (dates, time, locations), trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts.

2.2 Weather Data

Weather data was chosen in support of the green taxi data to further improve the analysis assuming people are more likely to take cabs in bad weather conditions for convenience. Data consists of daily weather information which includes station IDs, name of weather stations, date, precipitation, snowfall, snow depth and temperature.

3 Preprocessing

In this section, we transform all the data into a usable dataset for analysis, including feature selection, extracting all relevant features for further analysis.

3.1 Missing Values

- **Green Taxi Data**
  As tips did not include cash tips, tips will not be included inside the fare amount.
- **Weather Data**
  Missing values are present in average temperatures and observed temperatures, but we are only interested in weather data for specific stations (3 weather stations located near NYC) to determine the weather in NYC. They are JFK International Airport(KJFK), Newark Liberty International Airport(KEWR) and NY City Central Park(KNYC). As temperature data is not present in NY City Central Park, it will not be considered in our analysis.

3.2 Many Zero Values

- **Green Taxi Data**
  There seems to be a few columns containing many zeros in some of the itemized fares. Since we only care about the total fare, itemized fares will not be used but the total amount which includes all of the itemized fares (tips will be excluded) will be used.
- **Weather Data**
  Lots of zero values in snowfall, precipitation and snow depth, so they were removed.

3.3 Feature Engineering

- **Green Taxi Data**
  In order to have data for analysis on the day of the week and the shift the driver will be taking, features from the date and time of the pickups were extracted. Shifts of drivers were determined from an insider article explaining about the shifts of taxi drivers in NYC [4]. Morning shift is from 3am to 5pm and Night shift is from 6pm to 3am, but as the timings are not fixed, there would be drivers operating from 5pm to 6pm. This time frame would be added to morning shift as taxis would most likely still be driving while getting ready to stop work.
  Added features are:
  - shift - Morning shift from 3am to 6pm and Night shift from 6pm to 3am
  - day - day of the week of the trip
  - fare - total fare amount of the trip excluding tips
- **Weather Data**
  As mentioned in Section 3.1, 3 stations are used in determining weather conditions of NYC.

After dealing with missing values and many zero values, only temperature remains valid to be used for analysis. An average temperature is to be estimated to find the temperature of NYC. Temperature of NYC was computed by taking the weighted average of the individual contributions from each of the 3 stations seen in (Figure 1). Weights were proportional to the inverse of distance between NYC and each of the stations [5]. As NY City Central Park did not have information on temperature, the weightage of that station was distributed according to the weightage of the other two stations. So Newark contributed 67% while JFK contributed 33% to the estimation.
Added features are:

– WTAVG - weighted average of temperature between Newark and JFK stations

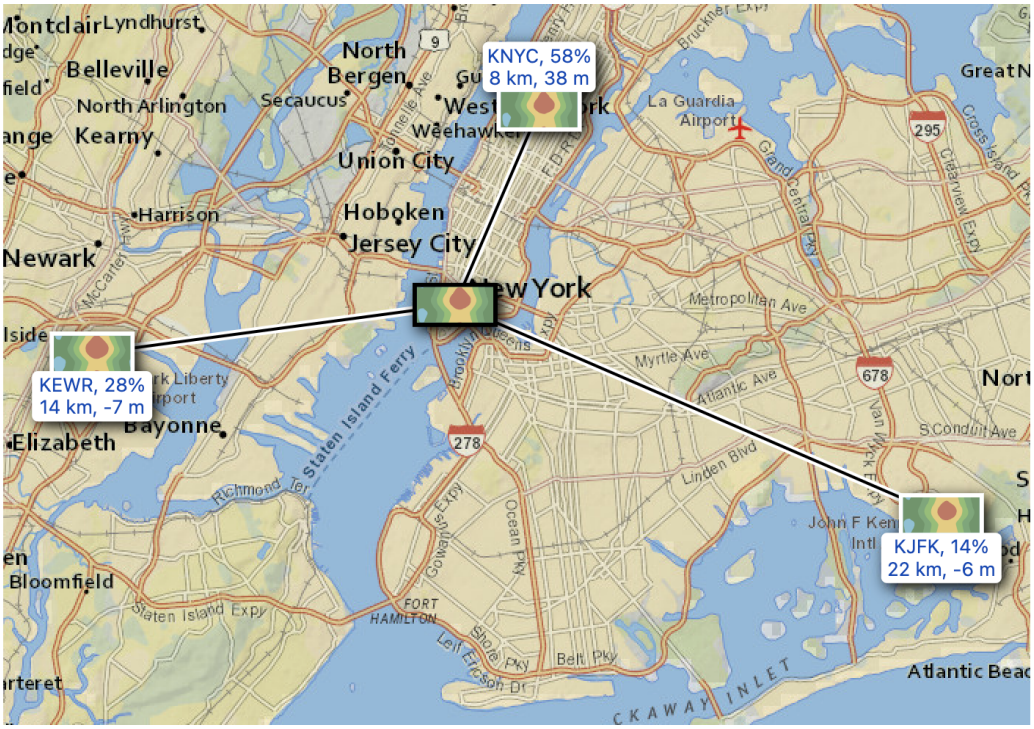


Figure 1: Stations' contributions to temperature estimate of NYC

3.4 Feature Scaling

- **Weather Data**
  As our analysis will include categorical variables, continuous variables with large values would affect the analysis, so temperature is normalized.

3.5 Feature Selection

Having unnecessary features would make the regression models not be accurate, so it was necessary to remove undesirable or unrelated features. This section will detail each feature's necessity and whether it is removed or used. Chosen features will be further checked for presence of relationship to the fare amount and demand for our analysis in Section 4.
Removed features are:

- Vendor's ID, drop-off location/time, passenger counts were removed as they had no relevance to our analysis

- Trip type was removed as if a taxi were to be dispatched, it would have to be in a close proximity to the customer, making the trip type not be useful
- With regards to time features, only shift and day will be included due to relevance to analysis

Features selected are:

- pick-up location
- day of the week
- shift
- temperature

3.6 Outlier Detection

This section summarizes the outliers present in the data.

- Fares below $2.50 USD were removed as the initial charge is $2.50 USD
- Pick-up Location IDs less than 1 or greater than 263 removed as the the 5 boroughs' location IDs range (2,263). Newark Airport is not considered part of the 5 boroughs according to the TLC's User Guide [6].
- Weather data on temperature did not have any outliers

4 Preliminary Analysis

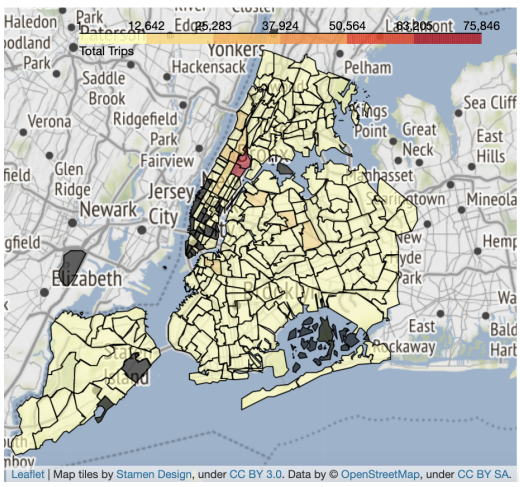
Features selected in Section 3.5 were further investigated to see if relationships existed between the features and fare/demand.

4.1 Pick-up Location

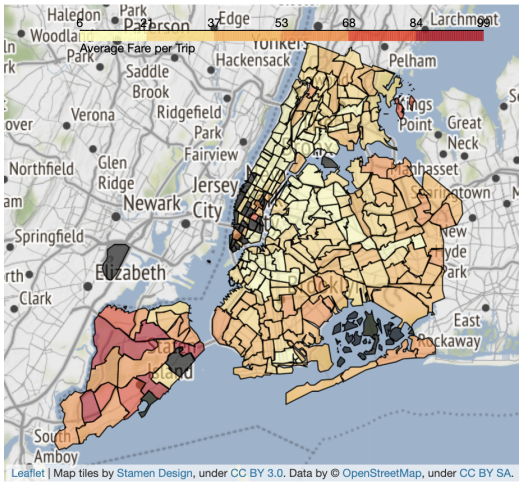
From (Figure 2a), the most profitable area so far based on total trips is right at the border where green taxis are allowed to pick up customers. As the green taxis are not allowed to operate south of West 110th St and East 96th St (aka South of Manhattan), they would wait right outside for potential customers. After further research, the reason why this would be the case is that the lower half portion of Manhattan is considered one of the word's foremost commercial, financial and cultural centres. Famous tourist attractions in New York are located in Manhattan like the statue of liberty, times square and empire state building. Not only that, many galleries, nightlights, museums and financial sectors like the Wall Street are located there as well. With that in mind, it is no surprise that the demand for taxis would be very high in Manhattan, which is why picking up a customer there would be very easy. But with this information,it shows how the pick up location ID is correlated to the demand and fare price of trips. But in (Figure 2b) we see that Staten Island actually has the highest average fare per trip. Most trips coming from Staten Island would be high in fares. With this information,it shows how the pick up location ID is correlated to the demand and fare price of trips.

An area of interest is that there seems to be a trade-off between fares and trips as can be seen in (Figure 2) where Staten Island has high average fares but low amount of trips while Manhattan has the highest amount of trips while having low average fares. This would need to be further investigated on which will be emphasized again in Section 6.

(Information on converting geometric objects to latitude/longitude to create chloropeth maps attained from Earth Lab [7])



(a) Total Trips

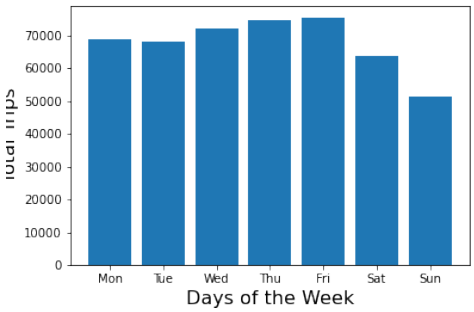


(b) Average Fare

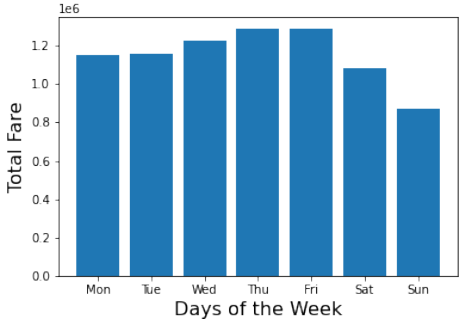
Figure 2: Total trips and Average Fare According to Location

4.2 Days of the Week

Looking at (Figure 3), there does exist a correlation in fares/demand to the days of the week. Thursday and Friday has the highest while the weekends, especially Sunday are pretty low. It is not surprising that Friday is the highest as it is the start of the weekends where people would go out and have fun after work or school. According to the day, we see that the fares/demand change accordingly as well as seen in the bar graph.



(a) Trips



(b) Fare

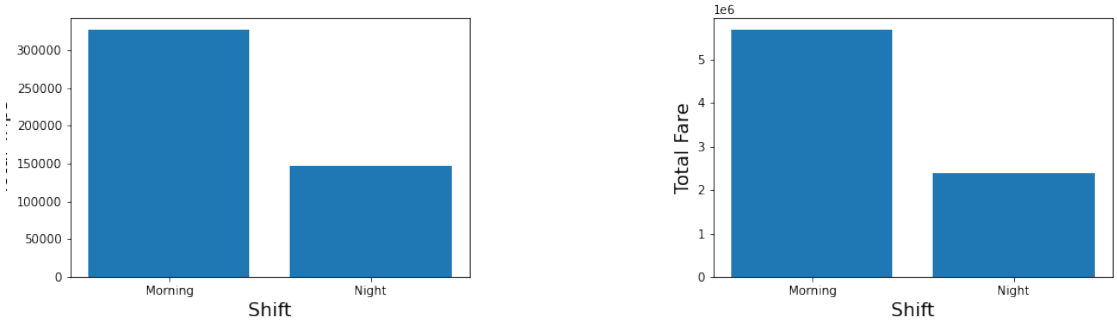
Figure 3: Trips and Fares According to Day

4.3 Shift

Looking at (Figure 4), shift has a clear relationship in the fares/demand made as well. It would make sense that during the day would be when most people would be out unlike in the night shift where people would be asleep.

4.4 Temperature

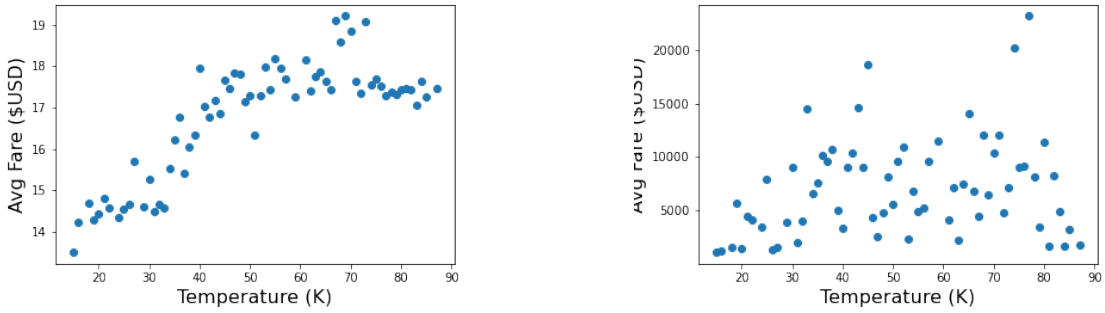
In (Figure 5a), it can be seen that the correlation is high correlation at 0.795. A possible explanation would be as the weather gets hotter, people are more likely to take cabs for longer distances. However,



(a) Trips (b) Fare

Figure 4: Trips and Fares According to Shift

the total trips does not reflect as well according to (Figure 5b). with the correlation of only 0.239, we see that trips and temperature do not have a good correlation.



(a) Average Fare (b) Total Trips

Figure 5: Total trips and Average Fare in Relation to Temperature

5 Modeling and Analysis

With 2 Linear Regression models, we would be able to estimate proportion of trips and average fare depending on the feature inputs. Linear Regression model assumes observations to be independent, homoskedasticity, normality and relationship between features and response variable is linear. As we are looking for demand and fare prices, we would use two linear regressions. Features considered are Borough(1), Day of Week(2), Shift(3), and Temperature(4). For training and test data, as we are dealing with time-series data, it would not make sense to use random train/test split. Instead the split will be a 80-20 ratio, where the first 8 months (1 July 2021- 28 Feb 2022) were used as training data while the last 2 months (1 March 2022 - 30 April 2022) were used as test data. The categorical variables were one-hot encoded before the regression. And one of the dummy variables from each categorical column were removed in the regression to prevent multi-collinearity. For evaluation, Mean Absolute Error(MAE) measuring average magnitude of errors in the set of predictions, and R-Squared for goodness of fit were used due to their high interpretability.

Regression on Average Fares	
Coefficient Name	Value
Intercept	67.51
Bronx	-44.41
Brooklyn	-50.39
Manhattan	-56.25
Queens	-51.54
Fri	-0.002
Mon	1.04
Sat	-0.50
Thu	-0.11
Sun	0.59
Tue	1.70
Morning	5.12
temp	1.81

Table 1: Regression Coefficients

Regression Error Analysis	
Evaluation Metric	Value
MAE	4.31
R-Squared	0.8619

Table 2: Evaluation of Model

5.1 Average Fare

In this Linear Regression model, all features mentioned above were used. The response variable is average fare.

With all the aggregation, the model was trained on 1713 instances. The coefficients and test results are seen in (Table 1 & 2). Any categorical values not listed in (Table 1) are accounted for in the intercept. Given the coefficients, our target audience (drivers of green taxis) can choose a borough, day of the week, shift, and temperature for the day to get estimated fare they would be expecting from a customer.

In (Table 2), the MAE of 4.31 can be interpreted as the model's forecast on fare of a trip's value from the true value is within $4.31. The R-Squared of 0.8619 states that 86.19% of the variance of fare is explained by the variance of the features in the model which is strong.

5.2 Proportion of Trips

In this Linear Regression model, all features mentioned above were used except temperature as we seen in (Figure 5b) which did not show much of a relationship. The response variable is the proportion of trips.

With all the aggregation, the model was trained on 70 instances. Although this might be small for training set, as all the variables are categorical, it would make sense to have 70 instances resulting to the maximum combination of all categorical variables. Redundancy of design variables would creates multicollinearity which violates assumptions of a linear regression. As we also only expect positive values in proportions, we will log the proportion before doing regression so as to not get any negative predictions.

The coefficients and test results are seen in (Table 3 & 4). Any categorical values not listed in (Table 3) are accounted for in the intercept. Given the coefficients, our target audience (drivers of green taxis) can choose a borough, day of the week, and shift for the day to get the estimated proportion of trips per total trips. Proportions alone might not be as useful and easy to interpret, so a possible way to get actual trip amounts could be to use the past week's sum of total trips which accounts for

Regression on Prop. of Trips	
Coefficient Name	Value
Intercept	-9.511499
Bronx	3.113775
Brooklyn	4.459772
Manhattan	5.715633
Queens	4.941597
Fri	0.108701
Mon	-0.025219
Sat	0.042180
Thu	0.068507
Sun	-0.125688
Tue	-0.052483
Morning	0.908181

Table 3: Regression Coefficients

Regression Error Analysis	
Evaluation Metric	Value
MAE	0.2693
R-Squared	0.9699

Table 4: Evaluation of Model

all the features in this model to estimate and multiply it by the proportion to get the rough estimate of trips to expect (make sure to revert back to original value).

In (Table 4), the MAE of 0.2693 can be interpreted that the model's estimation on proportion of trips from the true value is within 0.2693 (in log terms). The R-Squared of 0.9699 states that 96.99% of the variance of fare is explained by the variance of the features in the model which is strong.

6 Discussion and Recommendations

Using both models, demand(proportion of trips) and fare(average fare) can be forecasted with relatively good predictability power as mentioned in section 5 considering the little dataset that was available. Drivers are able to use these to gain information on where customers would most likely be and amount of fare they are likely to expect from the customers excluding tips. As good starting models, these models can be further enhanced.

Organizations overlooking these drivers can build on top of the models detailed in this report. As data is limited to only 10 months due to the post-pandemic period that started very recently, they can constantly update as time goes by and improve their models in the future.

As mentioned before in Section 4.1, organizations can also investigate deeper into the trade-off between average fares and total trips in boroughs of NYC.

In addition, only public datasets were available to the author. Organizations have the capability to retrieve much more external datasets to research on, finding useful features that can be added or in replacement of the weather data used in this report to further improve the models. Some areas to look for would be the number of taxis and for-hire vehicles, ride-hailing services like Uber or Lyft, geography of NYC including but not limited to population, resources, and political activities in NYC.

References

[1] Jose Martinez. *It's Not Easy Being Green: First Ever 'Boro Taxi' Driver Hits Brakes as Industry Tanks*. https://www.thecity.nyc/2022/3/1/22957301/its-not-easy-being-green-first-ever-boro-taxi-driver-hits-brakes-as-industry-tanks. Accessed: 2022-08-17.

[2] Devorah Lev-Tov. *New York State Lifts All COVID Restrictions*. https://web.archive.org/web/20210616232417/https://www.afar.com/magazine/new-york-city-set-to-fully-reopen-july-1. Accessed: 2022-08-14.

[3] The New York City Taxis and Limousine Commission. *TLC Trip Record Data*. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed: 2022-08-10.

[4] Madison Malone Kircher. *Here's the real reason why you can't get a New York cab at 5 p.m.* https://www.businessinsider.com/worst-time-to-hail-a-cab-in-nyc-2015-11. Accessed: 2022-08-17.

[5] Weather Spark. *Climate and Average Weather Year Round in New York City*. https://weatherspark.com/y/23912/Average-Weather-in-New-York-City-New-York-United-States-Year-Round#Sections-BestTime. Accessed: 2022-08-17.

[6] The New York City Taxis and Limousine Commission. *TLC Trip Records User Guide*. https://www1.nyc.gov/assets/tlc/downloads/pdf/trip_record_user_guide.pdf. Accessed: 2022-08-17.

[7] Leah Wasser. *Lesson 5. GIS in R: Understand EPSG, WKT and other CRS definition styles*. https://www.earthdatascience.org/courses/earth-analytics/spatial-data-r/understand-epsg-wkt-and-other-crs-definition-file-types/. Accessed: 2022-08-17.