Alexandra Whiteford

May 2023

# New York City Taxicab Analysis

Business Problem:

In New York City, yellow cabs can be found driving all throughout the five boroughs (Manhattan, Queens, Brooklyn, The Bronx, and Staten Island) offering rides to whoever may wave their hand in the air and flag one down for a ride across town. The city is known as the home of the yellow taxi cabs, as these cabs are heavily relied on for travel by locals and tourists alike. However, with the rise of rideshare companies such as Uber and Lyft growing in popularity since around 2015, the demand for yellow taxis have been slowly declining. Unlike taxicabs, these rideshare companies make themselves more appealing to customers by allowing riders to hail a car to their location from a mobile app. The app then provides the customer with pre-arrival access to the driver's name, history, and reviews, as well as the trip's anticipated price, route, and estimated pick-up and drop-off time. However, unlike rideshare drivers, taxis are set to a consistent, standard fare rate, as well as set toll and surge prices, so customers can feel comfortable The yellow taxicabs are owned and operated by several private companies, such as Gotham Yellow LLC, Queens Medallion Leasing, and several others, and are licensed by the New York City Taxi and Limousine Commision (TLC). This organization oversees approximately 40,000 for-hire vehicles including taxicabs, specialty black cars, commuter vans, limousines, and ambulettes (NYC TLC, 2022).

While learning about the differences between taxicabs and rideshare cars helps us understand the context of how taxicab businesses are operating in New York City, this report does not focus on the different features of rideshare businesses. However, it is important to consider some key points regarding the format by which taxicabs calculate their trips and how they are able to remain successful from a business perspective with competition on the streets. For example, both taxicabs and rideshare cars charge fares based on a combination of distance and time. However, Uber/Lyft do not factor in cruising or stop-and-go traffic speeds, while taxis charge different rates based on speed and distance rather than only time. Also, Uber/Lyft hike up the price of trips during periods of high demand called "surge pricing" while taxis do not, but rather apply a smaller rush hour fee or "congestion surcharge" during peak hours when the traffic is heavy. However, during peak hours of demand, customers may have to wait for longer times to hail an available cab. We must also consider the advantage taxis have by not needing a mobile app. If a customer does not have a smartphone that is up to date, or if their phone has died, they are unable to hail a rideshare car as New York City regulations prohibit street hails for private ride services (NYC TLC, 2022).

While the multitude of travel options is wonderful for New Yorkers, yellow cab owners often struggle to compete with the rise of rideshare cars as more citizens choose to call for a car mobily rather than hail a classic taxi in person. The demand for yellow taxis in New York have dropped rather quickly over the last decade, and continue to decline with each passing year as fewer people choose to ride by taxicab. There are several elements that make riding with rideshare cars more appealing, as I have discussed, and if the yellow taxicab businesses wish to increase their customer retention, these factors must be addressed and made aware to business owners and stakeholders in order to make the proper adjustments.

The goal of this project is to use observations about the data and how the different features of taxi rides impact the total fare in order to help New York taxicab businesses make predictions about how to make taxicabs more desirable to customers. By creating a model that can successfully predict the total fare of a ride based on the time and distance traveled as well as several other factors, we can determine what features are most closely influencing the total fare. From there, we can then establish a business plan that incorporates the most desirable and lucrative features that drive up revenue and potentially increase customer engagement. One way that I will approach this issue is by determining the peak time periods of taxicab use, which can be used to assume the peak times of customer demand, and implement a plan that will take advantage of these busy periods and, hopefully, drive up sales and customer retention.

## Data Wrangling:

The dataset that I am using comes from the New York City Taxi and Limousine Commission website where they keep datasets from each month of each year dating from January 2023 to January 2009. I decided to work on the dataset from May 2022 because I had discovered the link to the dataset through Kaggle, under the title “Taxi trip data NYC”, and worked with this public dataset because it was already in CSV format, unlike the other datasets on the TLC website in parquet format. The dataset was generated within the last 12 months, so I felt confident that it would be sufficient for data exploration. The website also included a dictionary key with the dataset containing brief descriptions of each of the column names, seen here: https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf There are 20 columns in the dataset initially, but first some cleaning must be done in order for us to properly assess the relationships between the different features. Since one of the major features to be examined is total fare amount, I have produced a list of the charges added to the total in addition to the time-and-distance-based fare. 2022 fares have since been updated and slightly increased, and comparisons have been made across several decades (Woodhouse, 2022).

Metered fare: **$2.50** initial fee, **$0.50** per 1/5 mile when traveling above 12mph or per 1 minute in slow or stopped traffic.
MTA tax: **$0.50**, Metropolitan Transportation Authority State surcharge for all trips that end in New York City, Nassau, Suffolk, Westchester, Rockland, Dutchess, Orange or Putnam Counties.
Improvement surcharge: **$0.30** for accessibility upgrades and other service improvements
Congestion surcharge: Total amount collected
Extra: Overnight surcharge: **$0.50** from 8 p.m. to 6 a.m. & Rush hour surcharge: **$1.00** from 4 p.m. to 8 p.m. on weekdays, however the data suggests that most customers pay around $2.75 as seen in the two categorical values produced.
Tolls: The cost of any tolls passed through during trip
Airport fees: **$1.25** when picked up from either JFK or LaGuardia Airports.

- If the trip is between JFK and Manhattan, in either direction, the flat rate is **$52**, including an increased congestion surcharge, the MTA tax, and any tolls.
- If a trip is from LaGuardia into Manhattan, the standard fare is paid.
- If a trip is between Newark Airport and Manhattan, a **$17.50** surcharge is added onto the standard metered fare.

Tip amount: The amount tipped to the driver by the passenger. The dataset only includes tips paid by credit or debit card, so it is unknown how many tips were paid in cash and how many did not pay any tip at all.

Methods for problem solving:

In order to predict the total fare amount, there are several steps that must first be done in order to ensure the data is tidy and trustworthy. Several components that make up the total fare amount have been listed above, and are represented in the columns on the dataset. However, there are some features of the dataset that are unnecessary or null, so these columns must be handled accordingly during the data cleaning process for correlation interpretation.

Predicting the fare amount based on travel time and miles traveled will help us create a regression model that can accurately predict what the total fare amount will be based on potential future taxi travel outcomes. The features in the dataset will help us define the relationship between each independent variable and the total fare, our dependent variable.

Success of the model is defined by the strength of the R2 and RMSE values which provide us with confidence for how accurate our predicted values are, which can be compared across several different models. The R-Squared (R2) value describes the error metrics for our regression model to help us evaluate the efficiency and accuracy of the model based on the data sets we are providing it with. This value is also known as the coefficient of determination, and it describes the variation of our model-produced data based on the independent variables we are providing it with. How well these independent variables can predict the outcome of the results

The Root Mean Squared Error (RMSE) provides us with an absolute number that represents how much our predicted results deviate from the actual results. Therefore, the smaller the error value, the better the regression model.

Data Cleaning & Exploratory Data Analysis:

One of the first major steps that was taken was to identify and get rid of any null values in the data. Upon loading the data for the first time and observing the tail values at the end of the 80,000+ data points, it is clear to see that many columns contain null values. Then, after displaying more detailed information using taxi.info(), I was able to see the total non-null values for each column. These columns were VendorID, store_and_fwd_flag, RatecodeID, passenger_count, ehail_fee, payment_type, trip_type, and congestion_surcharge. Immediately, we can see that the ehail_fee column has zero non-null values, so this column can be dropped.

Before entirely dropping or replacing the rest of these values, I wanted to do a bit more EDA to understand these features better.

After doing value counts on each of these partially null columns, I determined that the entire trip_type column can be dropped as well. There is no way to interpret the 1.0 or 2.0 values because it is not listed in the dataset's feature dictionary. Additionally, this information can be provided with the RatecodeID column instead to determine the type of trip taken, which is also more detailed with five categories rather than only two. I also decided to drop the VendorID column because it categorically describes the mobile technology system that the taxicab was using to log the data, which was not useful to the research being done.

Once I had dropped the few columns that were deemed unnecessary for the research, it was time to treat the remainder of the null and misentered values. I used the 'fillna' command using the forward fill method to replace any null values with those that are similar rather than dropping them entirely. In the mta_tax column, there are three value counts when only two make sense: Customers either pay the MTA tax ($0.50) or do not pay the tax ($0.00) however there are over 100 values that are listed as a negative tax value (-$0.50) which is not possible and the data were likely misentered into the system by the driver. The taxicab tax in the Metropolitan Commuter Transportation District states that a tax of $0.50 per taxicab must be paid if the trip:

1. Starts and ends in New York City (meaning New York (Manhattan), Kings (Brooklyn), Queens, Richmond (Staten Island), and Bronx counties)
2. Starts in New York City and ends in any of the following counties: Dutchess, Nassau, Orange, Putnam, Rockland, Suffolk, and Westchester.

Therefore, we can make the assumption that the negative -$0.50 values in the mta_tax column were accidentally entered as negative values, and so we will reassign them as the positive values we assume them to be. This will output a percentage of 59% total rides charged with the MTA tax in this dataset.

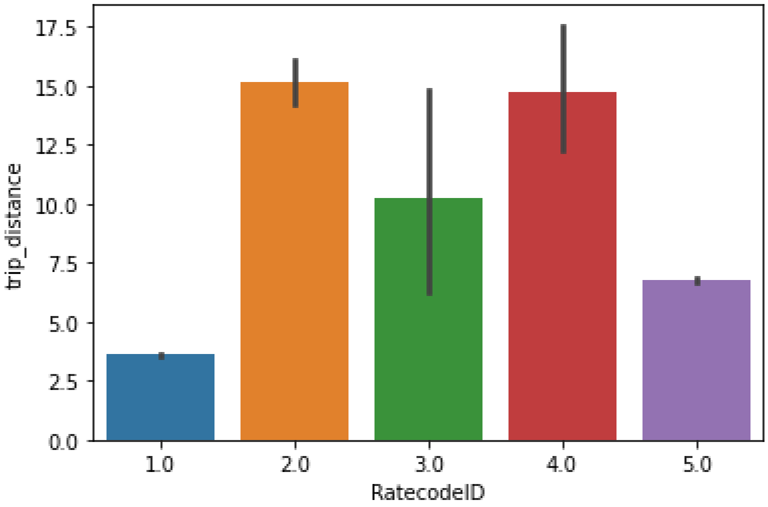
Similarly to the mta_tax column, there is a small portion of the data that appears to have been misentered in the improvement_surcharge column. In 2015, the Taxi and Limousine Commission initiated the improvement surcharge of $0.30 for every ride in addition to the rest of their fare. This tax is used to help taxi companies pay for accessibility upgrades and other improvements on their services, as well as payments to keep the proper number of active taxis on the streets as per New York City mandate. When we run a value count on this column, we can see that there are two common categorical values, as well as one negative with over 100 data points that we want to eliminate with replacement. Therefore, we can make the assumption that the negative values in the improvement_surcharge column were accidentally entered as negative and we will reassign them as their positive counterpart. Only 0.4% of rides were not charged with this surcharge. The same had to be done to the congestion_surcharge column by replacing the three lone negative values of -$2.75 with $2.75. However, my work of replacing negative values was far from over.

One major constraint that I had to work through was replacing these negative values with their positive counterparts in the columns fare_amount, total_amount, tip_amount, and extra. I ran a for loop on each of these columns to identify any values that were below zero. Since these columns would not realistically allow for negative values due to the fact that it is very rare for taxicabs to offer refunds, we can assume that these negative values are all misentered data points that we must now correct. In the fare_amount column, over two dozen negative value counts needed to be replaced one by one, which was quite tedious. In the extra column, there were only five negative values that needed replacing. In the tip_amount column, six negative values needed to be replaced. Finally, I corrected the total_amount column by replacing about 30 negative values with the corresponding positive counterparts. While I could have simply just eliminated these entries all together, I wanted to try and maintain as many data points as I could for later evaluation. Furthermore, as frustrating as this whole process was, there still seemed to be negative values remaining when I later ran a count on values that were less than zero. My solution was to simply eliminate those negative values altogether, which I determined shouldn’t be a concern for the strength of the data since this only eliminated 299 values.

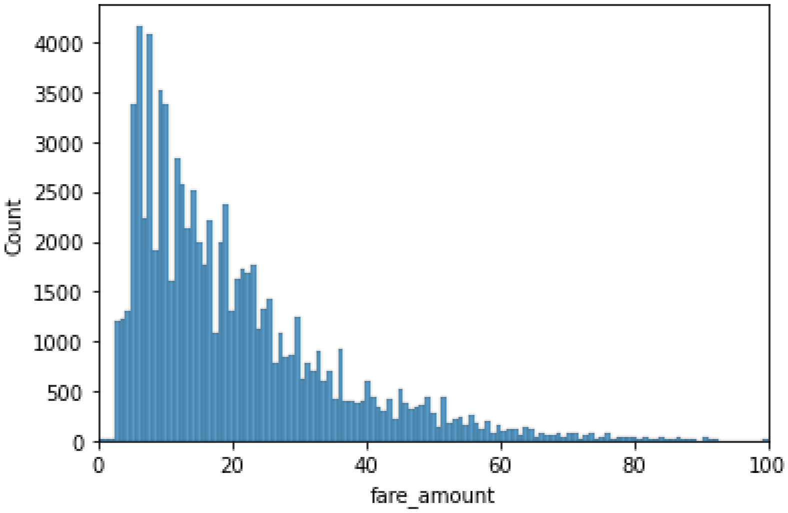
Another cleaning step that I did was replacing many values in the trip_distance column, as there were hundreds of values that didn’t make sense logistically. For example, the maximum distance traveled for one trip was listed as 260,517 miles, which is not possible in the slightest. This led me to question how likely it would be for a taxi driver to make a trip outside of the city, and after calculating the average distance to each of the major city and county destinations outside of NYC, I determined that it would be very unlikely for a driver to take someone more than 50 miles in one trip. This uncovered 235 trips with distances listed as over 50 miles, which I determined could be removed from the dataset due to possible inaccuracy. Furthermore, the minimum value listed for trip_distance was zero, which is also not likely to occur. After checking for the number of values less than or equal to zero, over 3000 values were identified. Rather than simply dropping these values from the dataset, I decided it would be a better idea to replace them with similar values found based on the type of trip that was taken. To do this, I used the five categorical values of RatecodeID: 1-Standard Rate, 2-JFK, 3-Newark, 4-Nassau or Westchester, 5-Negotiated Fare. Negotiated fare describes an agreed upon price negotiated between the driver and the customer when traveling outside of the city. I grouped the data by RatecodeID and found the median trip_distance values for each grouped trip type. The approximate distances between dropoff locations from central Manhattan are: Nassau county - 37 miles, Westchester - 30 miles, JFK airport - 15 miles, Newark - 16 miles. However, the pick-up location could be anywhere in the greater NYC area, so these values are generalized. Then, using lambda, I was able to replace each trip_distance value of zero with the corresponding median values based on the RatecodeID category. These compiled median values are shown below.

| | Original Median (miles) | Updated Median (miles) |
|---|---|---|
| 1 - Standard Rate | 2.00 | 2.20 |
| 2 - JFK | 17.32 | 17.32 |
| 3 - Newark | 0.11 | 7.955 |
| 4 - Nassau or Westchester | 11.9 | 11.9 |
| 5 - Negotiated Fare | 4.77 | 5.24 |

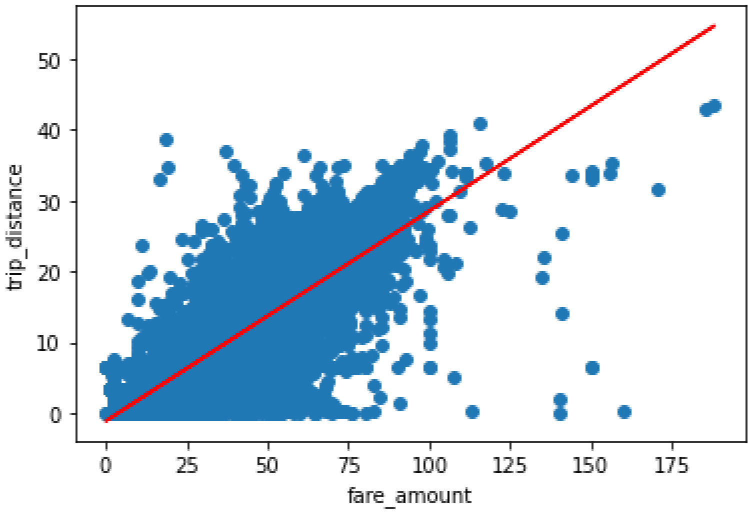
Due to the fact that distance traveled is one of the main features that will determine the total fare at the end of the trip, it is crucial to clean, verify, and retain as many data points as possible to help us build a functioning model later on. The updated median values are now more accurate for evaluation. Below, I have also included a bar graph of this for better visualization.



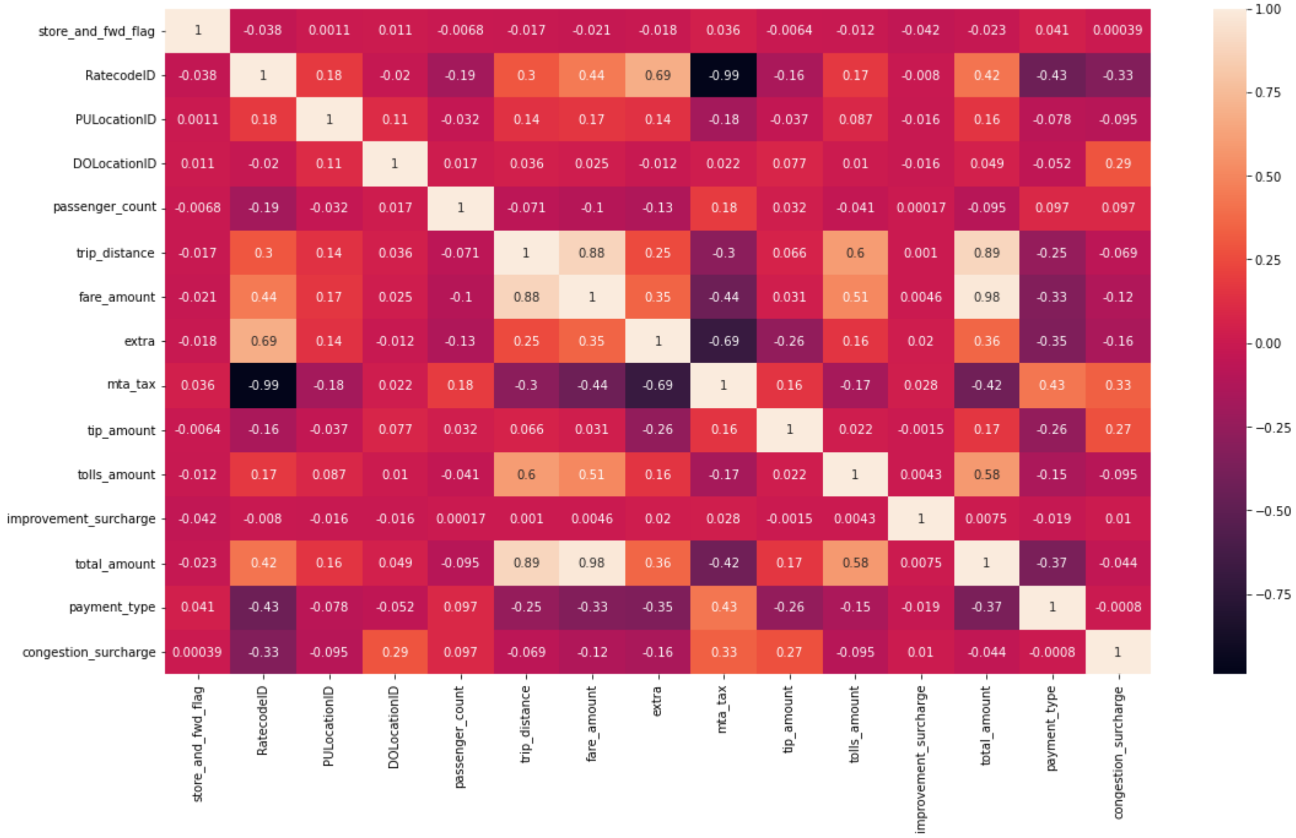
Continuing with visualization of the data, let's observe some trends within the dataset. I created a histogram of fare_amount values in order to observe the most common fare prices. Note that this value does not include any other travel costs aside from the metered fare calculated by time and distance. The vast majority of the fare amounts are less than $20, and very rarely does a customer pay more than $80, but it all depends on the duration and mileage of the trip.



Based on this visualization, I wanted to explore fare_amount further by graphing it alongside trip_distance to see if my assumption about the positive relationship between the two is correct. The scatterplot below shows us how common it is to see any given fare amount at the end of a trip. As we can see, the trendline shows a very strong positive correlation, as I expected.



Below is a correlation heatmap between all of the remaining features in the dataset. The lighter the color block and the closer the value is to 1.00, the correlation is positive and strong. While some correlations fit within this range, I have determined the most notable and relevant comparisons below in regards to the business problem.



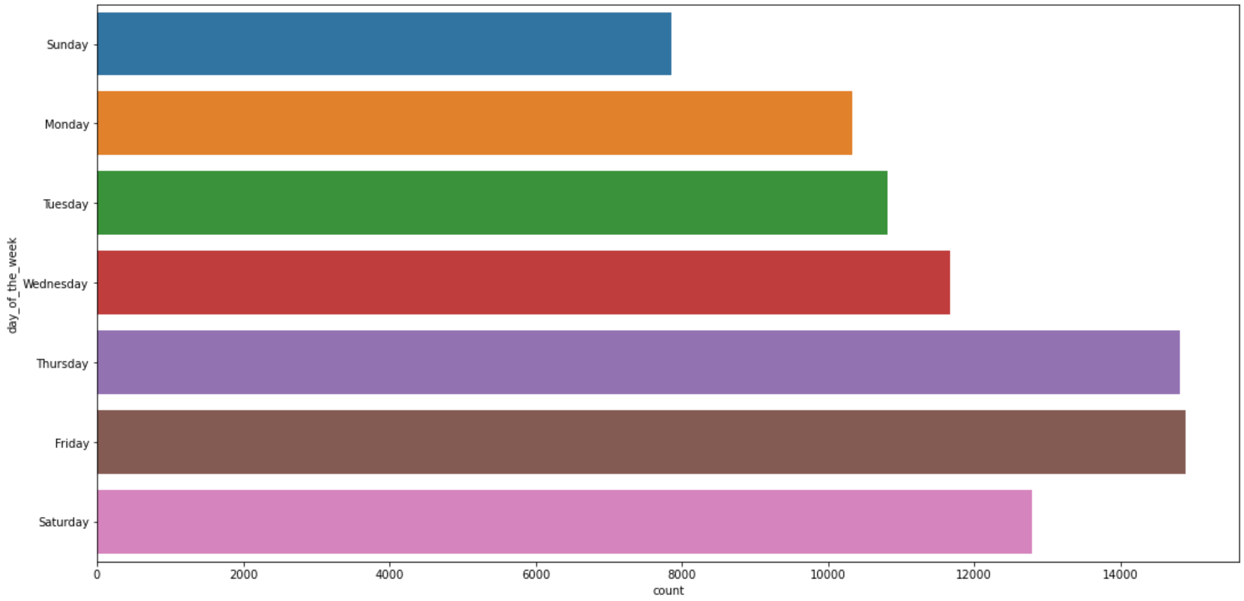
Notable correlations:

- total_amount & trip_distance = **0.89**
- fare_amount & trip_distance = **0.88**
- RatecodeID & extra = **0.69**
- tolls_amount & trip_distance = **0.6**
- tolls_amount & total_amount = **0.58**
- Ratecode ID & fare_amount = **0.44**
- RatecodeID & total_amount = **0.42**
- Extra & fare_amount = **0.35**
- Extra & total_amount = **0.36**

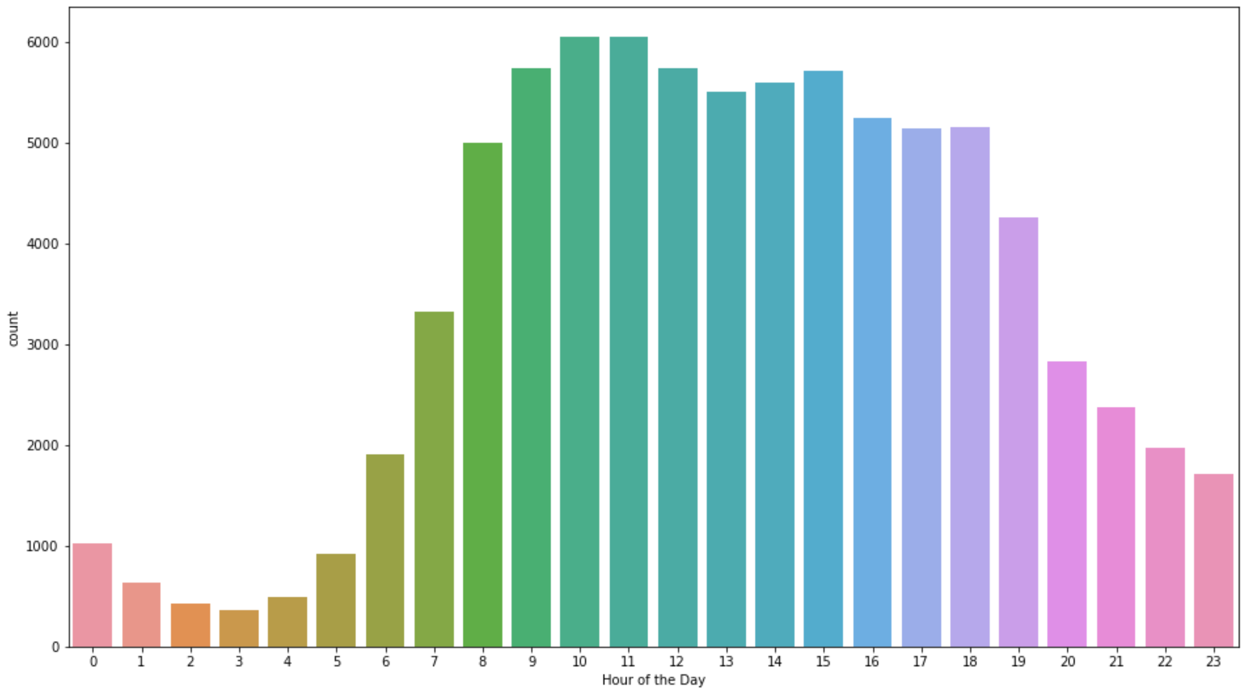
When comparing the correlation metrics between trip_distance against both fare_amount and total_amount, it appears that total_amount is a slightly higher value. With a stronger correlation metric, we will continue with our use of total_amount as the dependent variable. The next highest correlation metric is between RatecodeID and extra, which was surprising to me. The values in the extra column describe the additional charges to the ride that relate to time, such as overnight surcharges and rush hour surcharges, while RatecodeID values describe distance-related prices. Since we are observing the overall relationship between time and distance on price, this is definitely an interesting metric to take into account, however it is difficult to determine whether this has a direct impact on the total amount. Therefore, we can further investigate the relationship between both of these features on total_amount as well as fare_amount for comparison. We can see in the table above that the correlation between extra and total_amount is 0.36, and it makes sense that the relationship with fare_amount is slightly less strong as the additional surcharges seen in the extra column go towards the final total amount rather than the metered fare amount. Conversely, we can see that the RatecodeID relationship with fare_amount is slightly stronger than with total_amount, based on the direct charges to the fare_amount that are seen with the categorized, pre-set fares that certain RatecodeID trip types may include.

Several of the correlation metrics above are moderately strong and positive as anticipated. For example, it was expected to see a relationship of 0.6 between tolls_amount and trip_distance due to the idea that the farther one travels, the more likely it is expected that a toll will have to be paid. For very similar reasons, this also contributes to the metric of 0.58 between tolls_amount and total_amount.

The next process in the data visualization step is to observe the datetime metrics to determine the most popular, busy times of taxi travel. First we will observe the amount of taxi rides taken on each day of the week, where we can see that Thursday and Friday are the most busy days with Saturday as the next busiest day. Based on this visualization, it would appear that the number of taxi rides slowly increase over the course of the week and then start to decline between Saturday and Sunday. This graph can be seen below.



To further investigate peak times of travel, let's visualize the average amount of taxi rides taken during each hour of the day. It would appear that the most common time for taxi use is between the hours of 8:00 and 18:00, which makes plenty of sense considering that these are the main hours of the day that the sun is out and when most people are out of their homes. Upon closer examination of these peak hours, it is clear that the quick rise in demand from the hours of 6:00 - 9:00 represents morning rush hour travel. There also appears to be a spike in demand at 15:00, and then remains fairly constant from 16:00 - 18:00, however I anticipated more of an increase in total rides during this time. Due to the evening rush hour when people are leaving work and heading home or stepping out for a dinner event, it seems like there would be more of a spike in demand. This led me to do further research in order to investigate the hours that taxicabs are most active on the street and what role the drivers play in influencing these metrics.



During my research on the TLC website and database, I learned that taxi drivers are prohibited from working more than 12 hours in a 24 hour period, with a required 8 hour break after working those 12 hours. It is typical of most cab drivers to schedule their hours either in the morning or the evening to ensure they work through at least one rush hour period. Additionally, most taxicabs are shared between two drivers in a 24 hour period, each splitting their shifts into 12 hours. The drivers aim to schedule their shift change for the early morning and the middle of the afternoon in order to ensure that both shifts are equally attractive for both drivers based on demand (NYC TLC, 2022). In New York City, the morning rush hour is typically from the hours of 7:30 a.m. to 9:30 a.m., and the evening rush hour is usually from 5 p.m. to 7 p.m.. While 5 p.m. is a convenient time of day for drivers to either end or begin their shift, this is often an hour that customers seem to require a taxi the most. According to past taxi GPS records, between the hours of 4 and 5 p.m. are when the number of active taxicabs on the street decreases by 20% as this is a convenient hour for cabs to change shifts (Grynbaum, 2011). Many New Yorkers finish their workday and often struggle to find a taxicab to flag down, leaving them wondering where all the taxis went.

In 2004, the city of New York took action to attempt to halt the decline in active cabs on the street at 5 p.m. by implementing a congestion surcharge of $1 added to the fare for riders between 4 and 8 p.m. on weekdays. This was done to encourage taxi drivers with an incentive to stay out on the streets during the afternoon rush hour, as congestion surcharges would increase the total fare amount and further increase the potential tip amount. TLC chairman Yassky commented, “We believe that the 20 percent dip would be even worse if it weren’t for the surcharge,” (Grynbaum, 2011). In 2019, a new congestion surcharge was implemented by the city from $1 up to $2.50, which was likely due to inflation. It is also important to note that this surcharge is only applied when driving within the “congestion zone” which is classified as the main area of Manhattan south of 96th street (NYC DTF, 2022).

The issue of the afternoon shift change has clearly been a notable topic of discussion among the taxi businesses of New York for several years now. Unless taxicabs are privately owned by the drivers or do not require two drivers to share one car, it would be very difficult to change this commonly used practice. This shift change cannot be done earlier in the afternoon at, say, 14:00, because then the next shift change would have to be done in the extreme early hours of the morning at 2:00, which is not favorable to either driver. Thus, the resulting 5 p.m. (17:00) compromise.

## Preprocessing and Modeling:

Now that we have cleaned, explored, and highlighted several notable relationships within the dataset, it is time to move onto preprocessing and modeling. The dataset was defined with X and y values: X represented the independent variables, excluding total_amount, and the y value consisted of only the total_amount column. The dataset was put into numeric form and formatted with get_dummies, and then split into testing and training data with a train size of 0.75. The data was then scaled using StandardScaler and fit_transform on the X_train data.

I decided to work with three different regression models to test on the data: LinearRegression, DecisionTreeRegressor, and KNeighborsRegressor. From there, the most reasonable metrics to calculate in order to test the success of these models were R-Squared and Root Mean Squared Error. I have produced a table of the modeling results below.

|  | **RMSE** | **R2** |
|---|---|---|
| **LinearRegression** | 0.014723844045650638 | 0.0005002054971349405 |
| **DecisionTreeRegressor** | 0.00878704952524333 | 0.35217710077045294 |
| **KNearestNeighbors** | 0.01611586702310151 | -0.021194902701018 |

*Linear Regression*

The Linear Regression model is one of the simplest regression methods, searching for and calculating the relationships among the variables in the dataset in order to create predictions. The RMSE is approximately 0.0147 which is very low; this tells us that the residuals and variance were very similar to what was anticipated. This confirms that our predicted values align closely with the observed values, however the R2 value is approximately 0.0005 which is extremely low and not ideal. This tells us that while the data may be following a significant trendline, the data is also noisy and has high variability. This indicates that although the predicted values fall further from the regression line in this case, the predicted values are still able to give us information about the trendline.

*Decision Tree Regressor*

The Decision Tree Regressor is used to create a model that can predict the value of a target variable through the use of simple decisions, forming what are known as decision trees, into smaller and smaller subsets. This can be used to tell us how much our predictions deviate from the original observed values. The RMSE is approximately 0.0088 which is very low and a good sign; this confirms that there is little variance between the predicted values and the observed values. The R2 value is approximately 0.3522, which is a bit lower than the range we were hoping for, however it is still significant and can be used to make assumptions about the predicted data. However, this regression can often create unnecessarily complex trees that can often not generalize the data very well if there are too many compounded features. This is called overfitting, which is possible to have occurred here.

*K-Nearest Neighbors*

The K-Nearest Neighbors model works to classify the proximity of a predicted value to others so that they may be grouped or classified together. The RMSE value once again tells us the difference between the observed values and the predicted values, and that value is about 0.0161 which is very low and thus very good for our correlation between our variables and the predicted values. However, The R2 value is -0.0212 which is not good at all, as a negative R2 value means that the model's predictions are bad and less accurate than with our linear regression model. Using the K-Neighbors Classifier was not successful in this case, even though the RMSE value was low as we were hoping.

Overall, the model I am most confident in is the Decision Tree Regressor. With such a low RMSE value and a reasonable R2 value, I decided that this model best represents the predicted data. With an R2 value of 0.35, this indicates that 35 percent of the variation in the data can be explained through predicting the outcome based on the dependent variable's relationship with the other features in the dataset.

Constraints within this dataset:

I had initially included the Random Forest Regressor as the fourth model to test, however this took far too long to run and still did not produce results after several hours of waiting for the code to run. Therefore, I decided that the Decision Tree Regressor would be the best remaining option to pick based on success.

Ideally, I would have also chosen a dataset that was from the most recent date, which was January 2023. This way, I would be working with a dataset that included all of the newest miscellaneous taxes and surcharges that were updated on December 19, 2023. Instead, I had to determine what the price of each additional charge was prior to this updated list of charges, and it would have made the project even more relevant. However, as I stated earlier, the dataset that I decided to use was chosen based on the easy access of the CSV file format I found it in.

Recommendations for taxicab businesses:

There are a few recommendations I can make for yellow taxicab companies operating under the New York City Taxi and Limousine Commission to make efforts to increase customer retention and sales. One of the main recommendations I will make is to encourage more drivers to work through the evening shift change by scheduling the shift changes either earlier or later in the evening. My suggestion would be to for higher management to incentivize the drivers to rescheduling the shift changes by earnestly reminding them that if they were to take part in a temporary shift change, say, two hours later in the day, they have a higher chance of making more money in tips that would be brought in from working through the evening rush hour period. Not only would more rides be taken based on the significant demand for more active taxis at this time, but these additional rides would include congestion surcharges. A more favorable time to change shifts would be between 9:00 - 10:00, and then later again at around 21:00 - 22:00. With this implementation of a new shift change between two drivers sharing one taxicab, both drivers would still have the chance to work through one rush hour shift and still change shifts at reasonable hours in the day when there has proven to be less demand. However, due to the fact that these shift schedules have been in effect for several decades now, it would definitely be a challenge to convince many drivers to make the switch based on the long-established schedules and routines these drivers have been following for decades. Therefore, one long-term solution to this stubbornness would be to run an A/B test comparing the total fare amounts between drivers scheduled on the new shifts versus the standard shifts. The hypothesis being tested with this new implementation is that drivers working during the updated shifts will make more money and complete more rides by offering customers a higher percentage of active taxicabs on the streets during the evening rush hour.

An additional recommendation for taxicab businesses would be to use the funds collected from the improvement surcharge fee to fund tracking devices that would be placed in a small sample group of taxicabs, along with a simple mobile app that can help customers locate available cars. From here, we could run an A/B test for a several week-long trial to compare the control group of cabs, operating normally, with the group of cabs operating with the aid of a mobile app. Ideally, the app would also have the ability to call for a ride using the customers pick-up location and can provide the customer with the driver's ETA, and an estimated total fare price and travel time based on the results of our regression model. I strongly believe that the development of a mobile app such as this would make taxicabs more desirable as well as more accessible, which provides both the customer with quicker service and the driver with more available rides. The app would need to be advertised to customers as well, and the taxicab company could potentially advertise it on the empty ad spaces on the sides of the cab or on billboards throughout the five boroughs. Additionally, in either possible form of the mobile app, the driver would potentially save money on gas by stationing the cab in locations of high-traffic. While this business recommendation of mobile app development may be more expensive than a simple change in shifts, beta versions of the app can be tested in small groups and improved over time.

## Summary

One of the main features that helps taxicabs make more revenue is accessibility to customers. During my research on the Taxi and Limousine Commission and analysis of this 2022 dataset, it is clear that one of the main appeals to New Yorkers is the stable presence and dependability of taxicabs. The consistent pricing and availability of taxis are crucial components of reliable travel in a major metropolitan city such as New York, however the growing competition between taxicabs and rideshare cars along with rising economic inflation means that taxicab companies will soon require changes to their operations in order to remain in-demand financially successful. Therefore, I would strongly advise private taxicab businesses to implement new shift changes accompanying an A/B test in order to offer more availability to customers during peak hours of travel and furthermore compare total fare amounts and revenue between drivers in both test groups. By ensuring that more active taxicabs are made available, not only will more customers be provided with the travel that they need, but taxicab businesses will be able to increase profits for both the drivers and the company overall.

Sources:

Grynbaum, Michael. "Where Do All the Cabs Go in the Late Afternoon?" *New York Times*, 11 Jan. 2011.

NYC DTF. "Congestion Surcharge." *New York State Department of Taxation and Finance*, 12 Aug. 2022, www.tax.ny.gov/bus/cs/csidx.htm.

NYC Taxi and Limousine Commission. "Taxi Fare." *New York City Government*, 19 Dec. 2022, www.nyc.gov/site/tlc/passengers/taxi-fare.page#:~:text=There%20is%20a%2050%20cents,Manhattan%20south%20of%2096th%20Street.

Woodhouse, Skylar. "Taxi Fares Are Going Up 23% in New York City." *Bloomberg*, 15 Nov. 2022, www.bloomberg.com/news/articles/2022-11-15/nyc-taxi-cab-fares-to-rise-23-in-first-increase-since-2012.