May 2023

An Analysis of New York City Taxicabs

Springboard Capstone Project

Alexandra Whiteford

Introduction

Why?

New York city cabs can be found driving all throughout the five boroughs, picking up anybody who raises their hand in the air and flags down a car for a ride. 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 growing competition from rideshare apps, fluctuating customer retention threatens to lead to a loss in sales and rides taken.

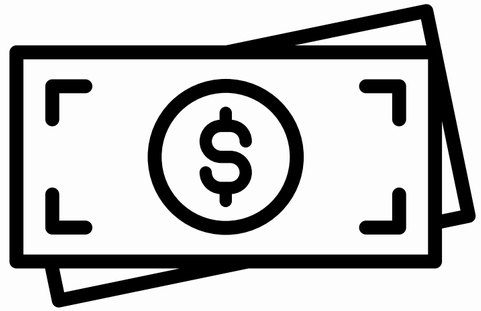
Audience

Officials at private New York taxicab companies such as:

- Gotham Yellow LLC
- NYC Yellow Cab Taxi
- NYC Perfect Transportation
- Quick Ride Corporation

Goal

Predicting the price and availability of taxicabs



By creating a model that can successfully predict the total fare of a ride, we can determine what features are most closely influencing the total fare. Furthermore, we can uncover what options taxicab businesses have to make themselves more attractive and readily available to customers.

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 Source

New York City Taxi and Limousine Commission; Kaggle: "Taxi Trip Data NYC", May 2022

This dataset comes from the TLC website, and includes the following columns:

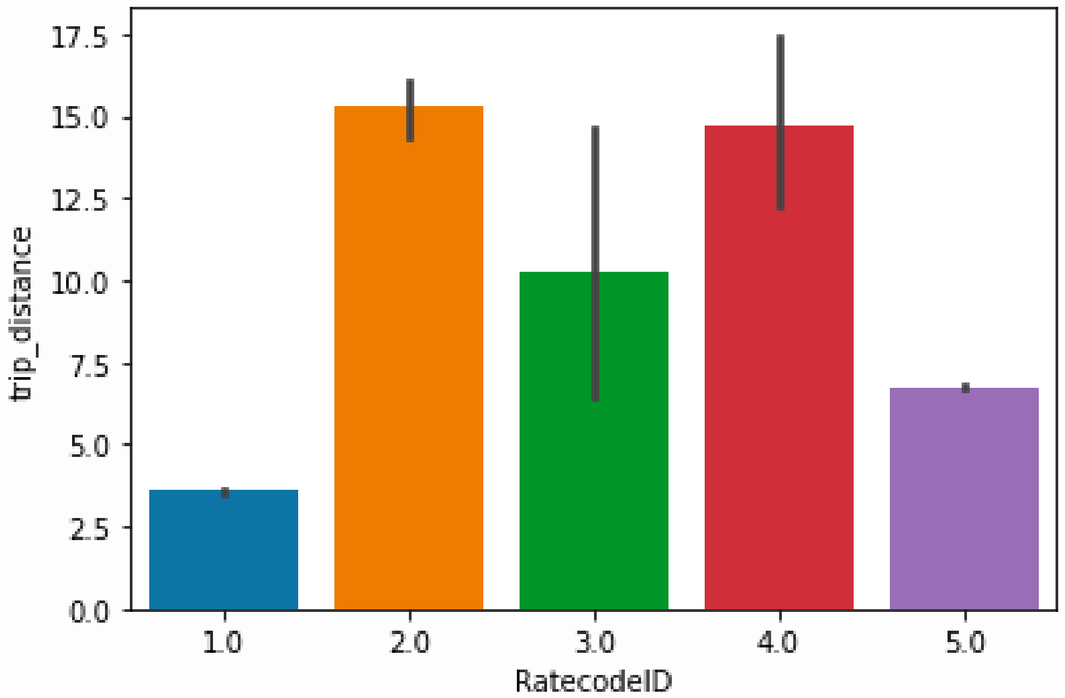
- VendorID - A code indicating the TPEP provider that provided the record.
- tpep_pickup_datetime - The date and time when the meter was engaged.
- tpep_dropoff_datetime - The date and time when the meter was disengaged.
- Passenger_count - Number of passengers in the vehicle; driver-entered value.
- PULocationID - TLC Taxi Zone in which the taximeter was engaged.
- DOLocationID - TLC Taxi Zone in which the taximeter was disengaged.
- RatecodeID - The final rate code in effect at the end of the trip; metered fare or pre-set airport fare.
- Store_and_fwd_flag - Indicates whether the trip record was held in vehicle memory before sending to the vendor, aka “store and forward,” because the vehicle did not have a connection to the server.
- Payment_type - A numeric code signifying how the passenger paid for the trip.
- Fare_amount - The time-and-distance fare calculated by the meter.
- Extra - Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.
- MTA_tax - $0.50 MTA tax that is automatically triggered based on the metered rate in use.
- Improvement_surcharge - $0.30 improvement surcharge assessed trips at the flag drop.
- Tip_amount - This field is automatically populated for credit card tips. Cash tips are not included.
- Tolls_amount - Total amount of all tolls paid in trip.
- Total_amount - The total amount charged to passengers. Does not include cash tips.
- Congestion_surcharge - Total amount collected in trip for NY congestion surcharge.
- Trip_distance - Total distance in miles from the pick-up location to the drop-off location.
- ehail_fee - Fee for hailing cab electronically; contains only null values, so this is irrelevant.

Data Cleaning and EDA

Issues encountered during data exploration and cleaning:

- Dropping unnecessary columns
  - ehail_fee - contains zero non-null values
  - trip_type - values in the data are undefined and impossible to interpret. Also redundant to RatecodeID
  - VendorID - not useful information
- Getting rid of null values
  - using fillna() and the forward fill method
- Converting implausible negative values to accurate positive counterparts
  - mta_tax, improvement_surcharge, congestion_surcharge, fare_amount, total_amount, tip_amount, extra
- Distance traveled unrealistically too high or too low
  - Dropping values above 50 miles; outliers, or likely entered incorrectly
  - By using RatecodeID categories to find the median distance traveled for each trip type, I replaced the zero mile distances with the median values.

Trip Distance and RatecodeID

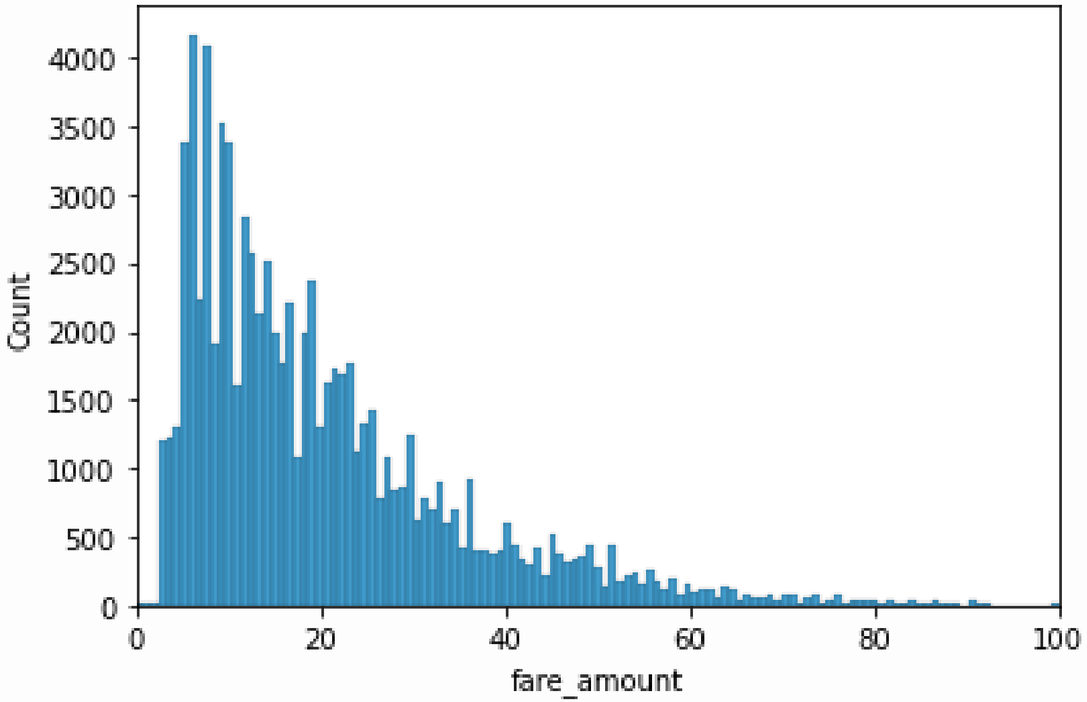


As I discussed in the previous slide, I used the RatecodeID values to find the median values of the distance traveled. I then replaced any distances of zero with the median values, and produced the bar graph with the updated median values, as seen here.

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.

Fare amount

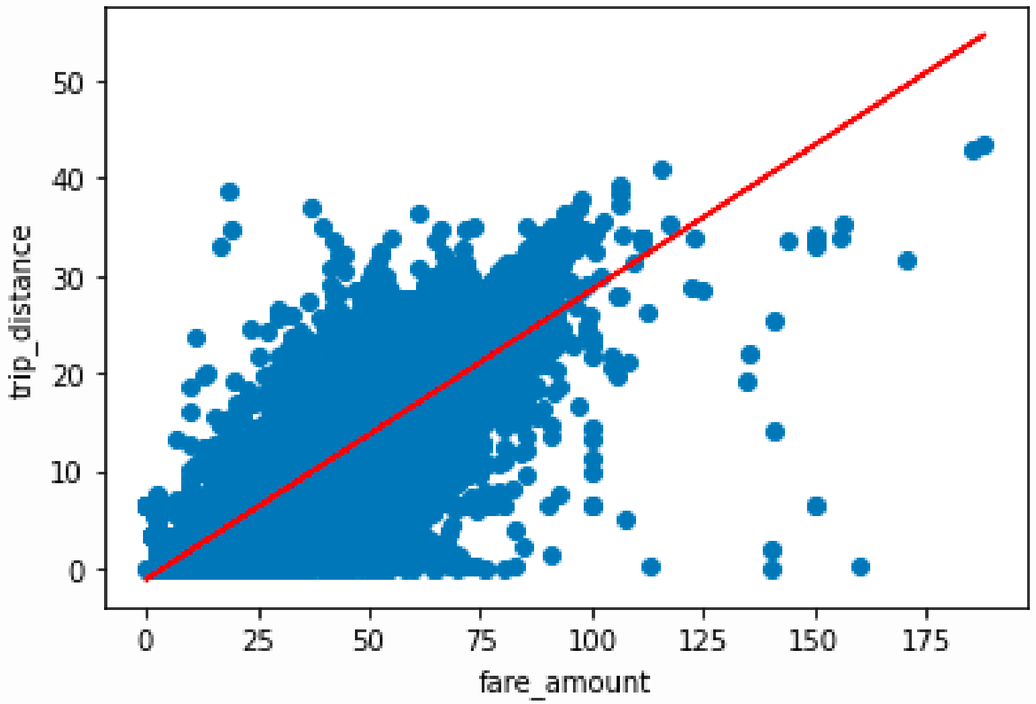


This is a histogram of fare_amount values in order to observe the most common fare prices.

- The vast majority of the fare amounts are less than $20
- Very rarely does a customer pay more than $80, but it all depends on the duration and mileage of the trip

Note that this value does not include any other travel costs aside from the metered fare calculated by time and distance.

Fare_amount and trip_distance



The scatterplot here shows us how common it is to see any given fare amount at the end of a trip.

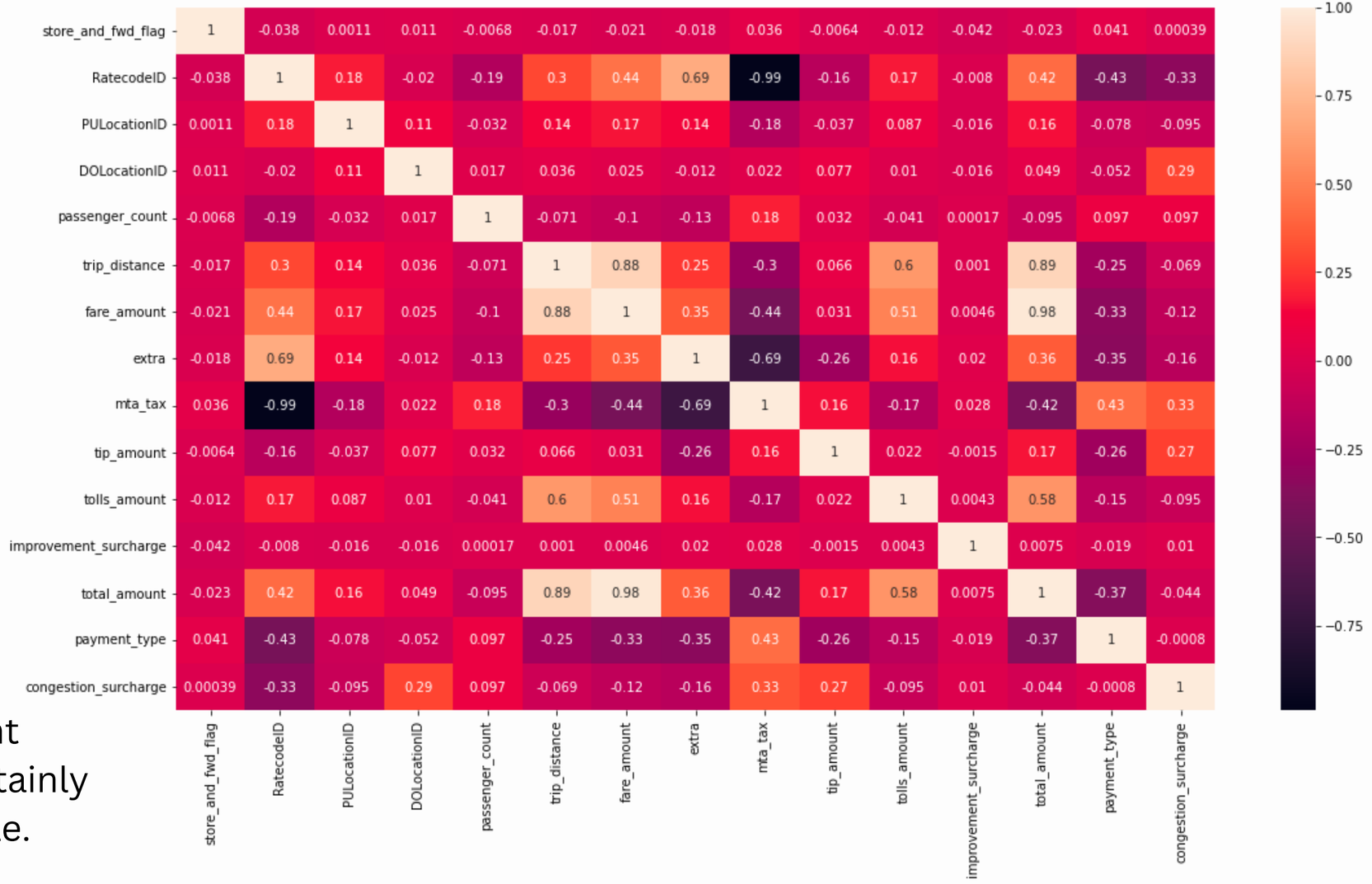
As we can see, the trend line shows a very strong positive correlation, as anticipated.

Correlation Heatmap

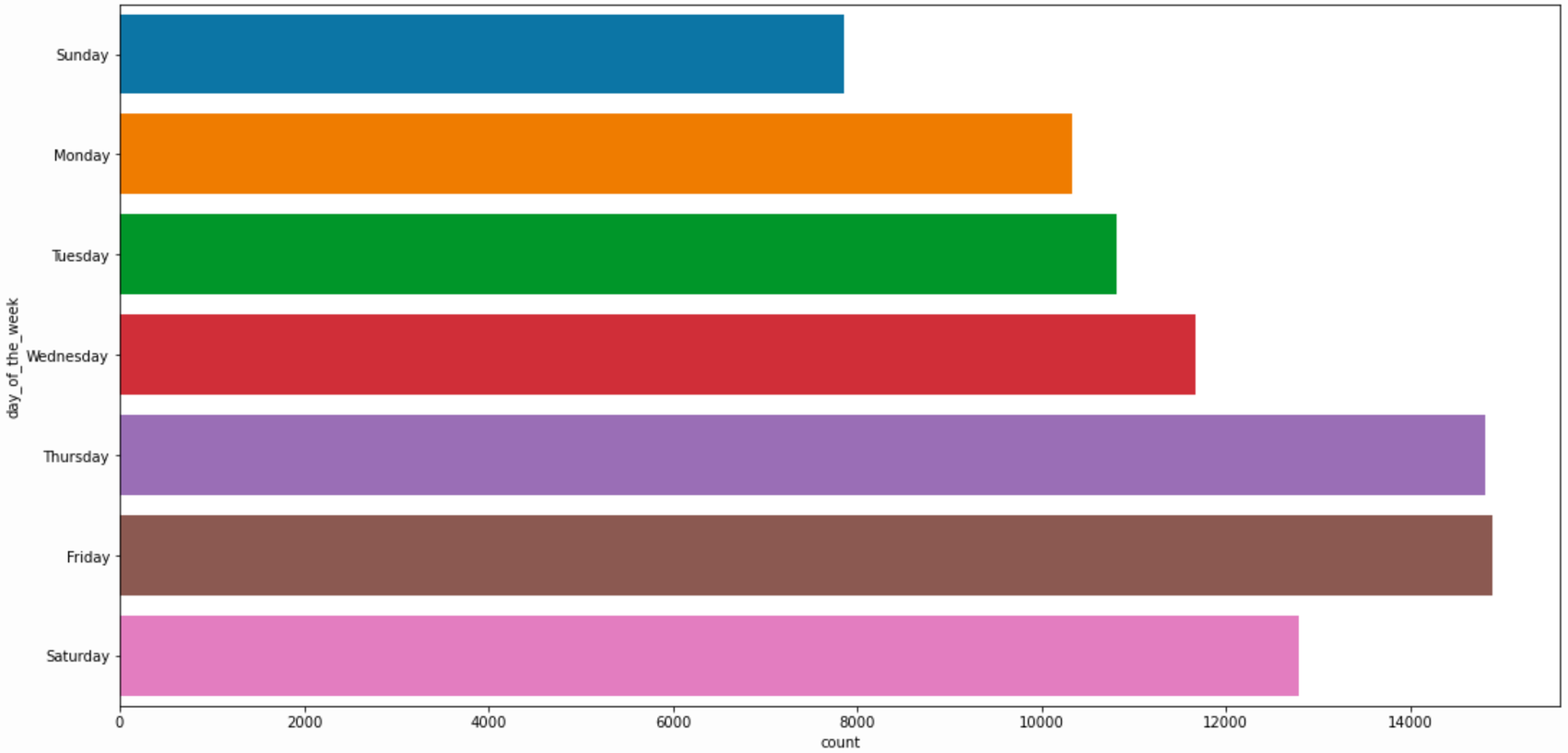
Notable correlations to consider:

- 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**

Total_amount appears to have slightly stronger correlations than fare_amount between other features, so we will certainly focus on this as our dependent variable.



Day of the Week



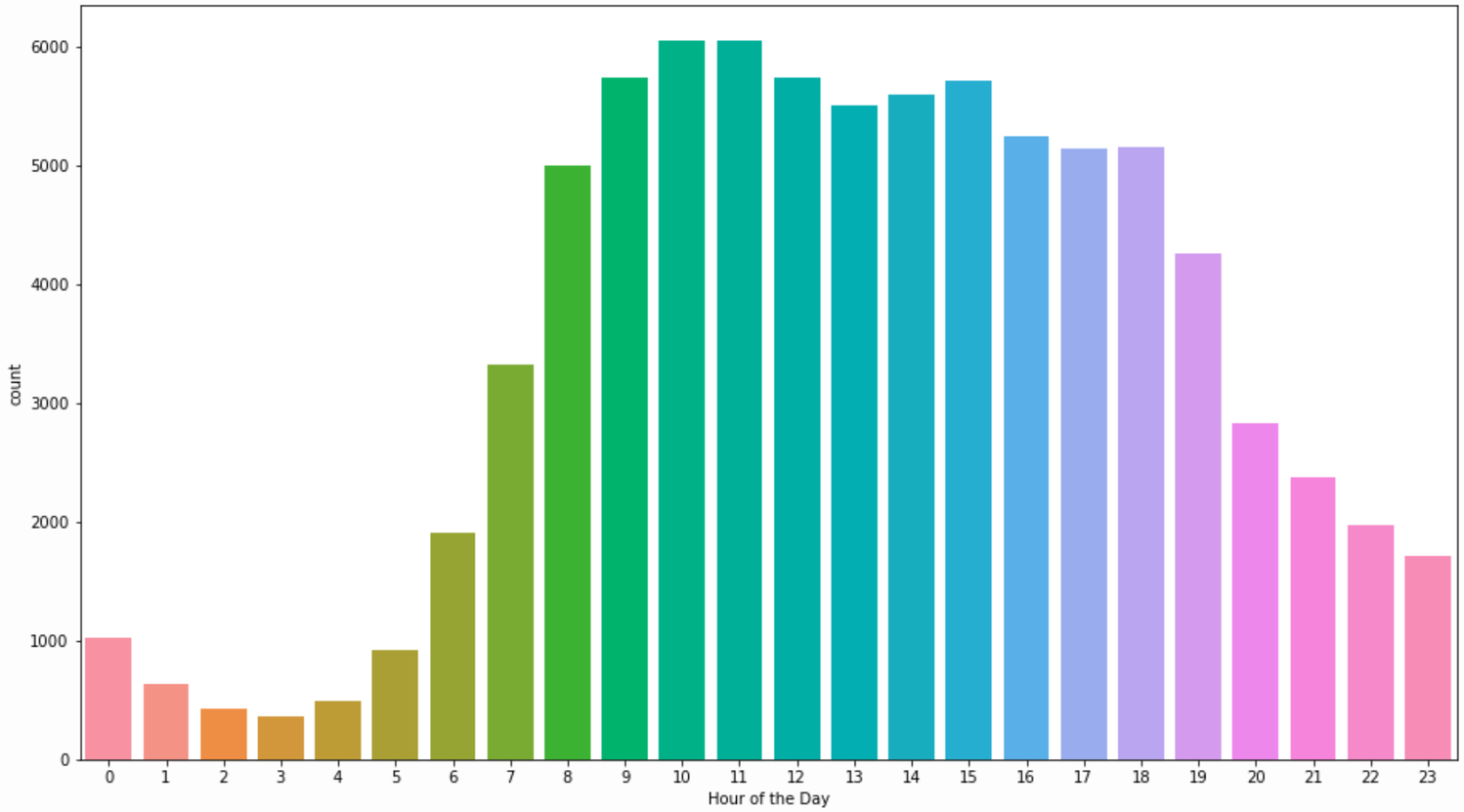
Thursday and Friday are the most busy days with Saturday as the next busiest day

It appears that the number of taxi rides slowly increase over the course of the week and then start to decline between Saturday and Sunday

Time of the Day

It appears that the most common time for taxi use is between the hours of 8:00 and 18:00, as these are the main hours of the day that the sun is out and when most people are out of their homes.

- Quick rise in demand from the hours of 6:00 - 9:00 represents morning rush hour travel
- a small increase in demand at 15:00, and then remains fairly constant from 16:00 - 18:00
- Overnight decline, few people needing rides



Modeling comparison table

Out of the 3 models tested, the Decision Tree Regressor was the most successful and reliable:

- Lowest RMSE value of 0.008
- Highest R2 value of 0.352

	RMSE	R2
LinearRegression	0.014723844045650638	0.00050020549713494
DecisionTreeRegressor	**0.00878704952524333**	**0.35217710077045294**
KNearestNeighbors	0.01611586702310151	-0.021194902701018

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.

Therefore, using the Decision Tree Regressor model provides us with confidence that the data is following a significant trend line representing the correlation and variation between the observed and the predicted data values.

Constraints

Two main constraints when working with the data:

I would have liked to use the Random Forest Regressor, but this too far too long to run and did not finish generating the RSME and R2 values after several hours of waiting.

Could have chosen the most recent dataset from the TLC website (January 2023) to reflect the updated fare prices put into effect in December of 2022. I chose this dataset because it was already in CSV format, and I did not realize at the start of the project that the fare prices had been updated so recently.

Recommendations

- Encourage drivers who split shifts and share a taxicab to make their shift change a few hours later, rather than from 4-5pm as usual in order to take advantage of the evening rush hour
  - this would lead to more rides completed, providing more customers with the travel they need at this time
  - Would also bring in more revenue for both the taxicab business and the driver
  - working through rush hours causes customers to pay marginally higher total fare amounts, leading to ideally more tip money for the driver.
- Create a mobile app that can locate and hail cabs from anywhere in the five boroughs
  - Run an A/B test using a small sample of drivers to compare total fares against control group of drivers. If successful, implement and advertise the mobile app across the boroughs.

In Summary

One of the main features that helps taxicabs make more revenue is accessibility to customers.

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.

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.