FINAL REPORT

ANALYSIS ON YELLOW TAXIS AND CONGESSIONS IN THE NEWYORK CITY

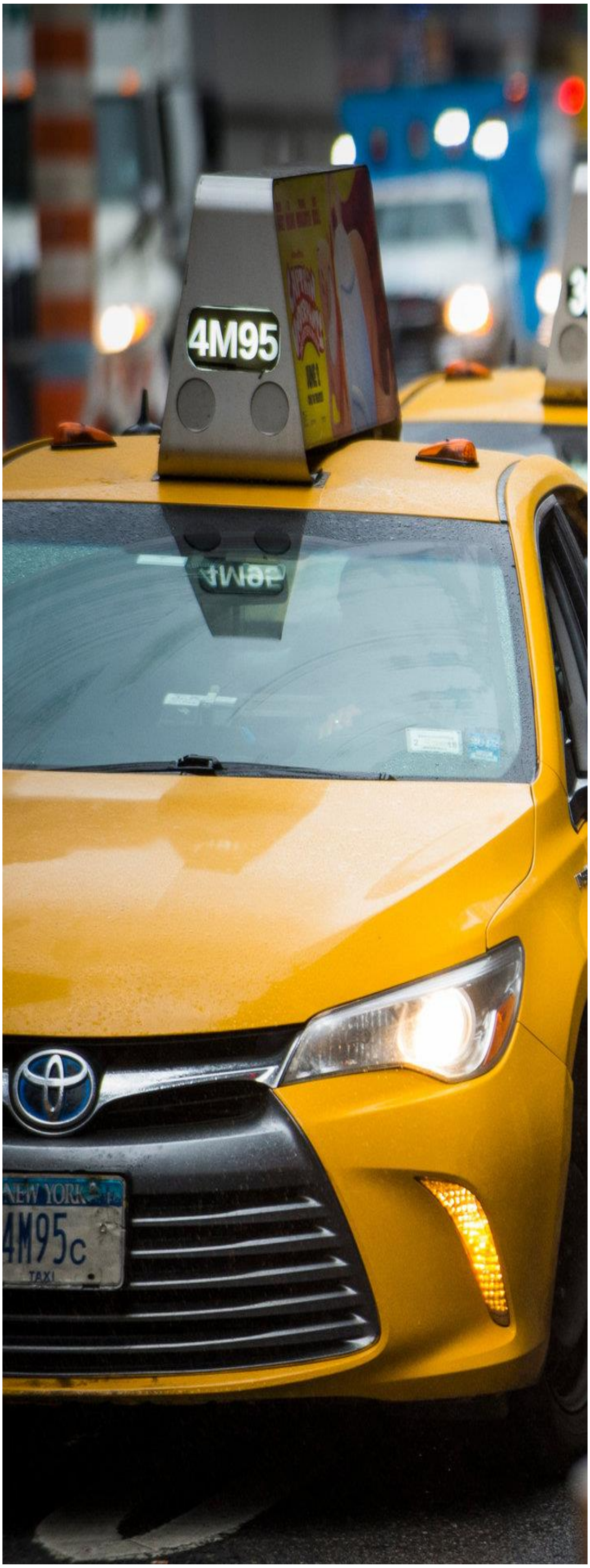
PRESENTED BY

Keren Melinda V, Lakshmi Priya, Liviya Sekar

Ritartha Chakraborty & Sriram S

Group-8

Mentored by: Mohit Sahu



The New York City taxi industry is an integral part of the city's transportation system, providing a convenient and accessible mode of travel for millions of residents and visitors every year. The industry is known for its iconic yellow taxi cabs, a symbol of New York City and it has a rich history dating back over a century.

Yellow cabs are regulated by the New York City Taxi & Limousine Commission (TLC) and are authorized to pick up passengers who hail them on the street. To operate a yellow taxi in New York City, drivers must obtain a special license called a "Taxi Medallion." These medallions act as permits, allowing drivers to legally pick up street hails. The medallion system was implemented in the 1930s as a way to control the number of taxis on the streets and maintain industry standards.

This traditional taxi industry is now facing increasing competition from ride-hailing services like Uber, Lyft, and Via.

On an average the industry sees 4 to 5 lakh trips every day. However, New York experiences severe traffic congestion during rush hours, typically in the morning and evening. Commuters from all five boroughs and neighboring areas flood the roadways, leading to slow-moving traffic and lengthy travel times.

To deal with this congestion, the city has taken several measures such as implementing congestion surcharges and rush hour surcharges for Taxis to disincentivize people, restricting taxis in several places, encouraging people to use public transportation etc.

Problem Statement

One of the major issues faced by the population of New York is the traffic congestion in the city during the rush hours which in week days falls between 07-10 hrs and 16-20hrs. The rush hours in weekends will vary based on major events happening in the city such concerts, major sporting events etc. As per the tom tom traffic index, due to the congestion average ride hour exceeds by approx. 12-17 minutes. This state might cause dissatisfaction to the passengers and taxi drivers. This major problem needs attention and right intervention.

Current solution to the problem

To handle the congestion is the city the government has taken the following measures.

1. Rush hour charges for yellow taxis – to disincentivize people to take taxis during the rush hours and to schedule a plan in the non-rush hour if possible.
2. Restriction of yellow taxis to certain areas/zones in the city.
3. High occupancy vehicles to encourage car pooling and reduce single occupancy vehicles during rush hours.
4. Encouraging the population to use alternate modes of transportation.

To alleviate the issue of dissatisfaction, the present system has facilities to give a predicted fare through the curb, a taxi hailing app, but it is important to keep the passengers informed of a fairly accurate fare prediction especially during the rush hours. and the total duration of the trip to alleviate the level of dissatisfaction when the rides happen directly by hailing.

Proposed solution to the problem

A fare prediction and trip duration prediction system specially designed for the rush hours will keep the passengers more informed of the situation reducing anxiety and tension.

A thorough analysis of the data pertaining only to the rush hours of the weekdays can help find the patterns which in turn will help the government to take informed decisions and prioritize improvements in the traffic regulation system. Also, this might benefit the taxi drivers in finding better rides and better revenue realization.

Project Outcomes

- A Rush hour fare and trip duration prediction system.
- Complete analysis of the present situation.
- Key findings of hidden patterns in the data.
- Key suggestions to address the congestions of the city.

DESCRIPTION:

DATA: New York Yellow taxi trip data (April 2023)

DATA DICTIONARY:

Field Name	Description
VendorID	A code indicating the TPEP provider that provided the record. **1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.**
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	The number of passengers in the vehicle. This is a driver-entered value.
Trip_distance	The elapsed trip distance in miles reported by the taximeter.
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. **1= Standard rate** **2=JFK (Trips between JFK airport and Manhattan)** **3=Trips to Newark airport** **4=Trips between New York and Nassau or Westchester** **5=Negotiated fare** **6=Group ride**
Store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka “store and forward,” because the vehicle (POS machine) did not have a connection to the server. **Y= store and forward trip** **N= not a store and forward trip**
Payment_type	A numeric code signifying how the passenger paid for the trip. **1= Credit card** **2= Cash** **3= No charge** **4= Dispute** **5= Unknown** **6= Voided trip**
Fare_amount	The time-and-distance fare calculated by the meter.
Extra	Miscellaneous extras and surcharges.
MTA_tax	$0.50 MTA tax that is automatically triggered based on the metered rate in use.
Improvement_surcharge	$1 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.
Tip_amount	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 NYS congestion surcharge.
Airport_fee	$1.75 for pick up only at LaGuardia and John F. Kennedy Airports

DATA DESCRIPTION:

The dataset consist of Taxi riders from 6 Boroughs of New York: Manhattan, Queens, Brooklyn, Bronx, EWR and Staten islands.

These Boroughs have been split in to 263 zones indicated as location ids in the dataset. The other major information to be known before proceeding is the fare structure of the taxi's being collected. The following the summary of the charging system in place

Yellow Taxi Rate chart								
Destination	**initial charge**	**further Charges**	**MTA**	**imp. Surcharge**	**Congession charge**	**Airport Fee**	**Rush hour surcharge**	**Other surcharges**
Within NY	3	0.7	0.5	1	2.5	-	2.5	-
Outside NY (except Westchester & Nassau)	Negotiated flat rate (decided between passenger and driver) **rate code-5**		0.5	1	2.5	-	2.5	-
Westchester & Nassau counties (**rate code 1**- till city limit, **4**- after crossing city)	3	0.7 (till city limit) 1.4 **(double after city limits)**	0.5	1	2.5	-		-
To and from La Gaurdia Airport	3	0.7	0.5	1	2.5	1.75 (only for pickup)	2.5	**5**
Between **JF Kenedy** Airport and **Manhattan** **Rate code -2**	70 (flat rate)		0.5	1	2.5	1.75 (only for pickup)	5	-

Destination	initial charge	further Charges	MTA	imp. Surcharge	Congession charge	Airport Fee	Rush hour surcharge	Other surcharges
Between **JF Kenedy** Airport and **Other city**	3	0.7	0.5	1	2.5	1.75 (only for pickup)	2.5	-
Newark Airport **rate code -3**	3	0.7	0.5	1	2.5	-	2.5	**20**

* Highlighted in Red were not explicitly mentioned in the taxi charges of the new York state.

Column	Description
Initial charge	Base fare amount of 3 dollars
Further charges	per 1/5 mile above 12mph or per 1 minute below 12mph or when stopped
MTA	Metropoliton Transport Authority surcharge: for all trips that **end in New York City or Nassau, Suffolk, Westchester, Rockland, Dutchess, Orange or Putnam Counties.**
Congestion Surcharge	for all trips that begin, end or pass through Manhattan south of 96th Street.
Rush hour surcharge	4-8 pm on working weekdays
Overnight surcharge	8pm to 6am (to check whether it has been charged for dropoff time exceeding 20:00)
Toll charges	Rides that touch: Westchester and Nassau Counties Trips over the Cross Bay Veterans and Marine Parkway-Gil Hodges Memorial Bridges Newark Airport (EWR)

ALTERNATE SOURCE OF DATA:

The Taxi dataset does not include the names of the location ids and to which borough it belongs to. This data is available as lookup table separately in the same Newyork website from where the main dataset has been taken.

Column	Description
Location ID	ID of the zone
Borough	Borough to which the zone belongs to
Zone	Name of the zone

Data Cleaning:

The New York yellow taxi rides dataset (April 2023) has been chosen, then as per the problem statement the data pertaining to the evening rush hours of New York (16:00 – 20:00 of non-holiday week days) has been filtered for further cleaning and exploration.

Shape of the dataset before cleaning: 143754,19

Null values:
3269 rows had null values for passenger count, Rate code ID, store_and_fwd_flag, congestion sur charge and Airport fee.

There were dropped as most of the values were missing/wrong in all of these rows.

Duplicates:
No duplicates were found.

Major discrepancies found in the dataset:

- **2 location ids 264 and 265:**
  - The locations were present in more than 2200 rows
  - Had unknown as values in the look up table. On further exploration it was seen that there was no specific pattern for these locations. Trips starting from the same location and ending at these places had very different trip distances.
  - This might be codes that the drivers use in their electronic device when the location they visit is unknown. Since there was no patterns present. Keeping this data might give us wrong inferences.
  - Therefore, these are dropped

- **Fare amount less than 3 (1293 rows) and 0 values in passenger count (2426 rows):**
  - All the other values in these rows were not proper.
  - Fare amount cannot be less than 3 as the base charge collected for the rides are three.
  - Therefore, these are dropped.

- **Ratecode id 99 (465 rows):**
  - There is no such rate code defined by the New York city for the taxi rides.
  - This might be a custom rate code used by the drivers for some other purposes other than the regular taxi rides.
  - Therefore, these are dropped.

- **Trip distance is 0 (992 rows):**
  - All the other values in these rows were not proper.
  - As the trip distance cannot be 0 these are dropped.

- **Trip Duration greater than 10 hours (96 rows):**
  - A driver is allowed to drive only for a period of 10 hours continuously as per rule. Hence these rows are removed.
  -

- **Trip pick up and drop off dates are different (15 rows):**
  - At no instance this can happen unless the ride starts at night 8 and goes above 12pm but this was not the case.Therefore, these are dropped.

- **Trip Duration of Less than 1 minute (344 rows):**
  - There were illogical wrong values present in the rows where the trip duration was less than a minute. Hence Dropped

Column wise anomaly imputations and removal in the dataset:

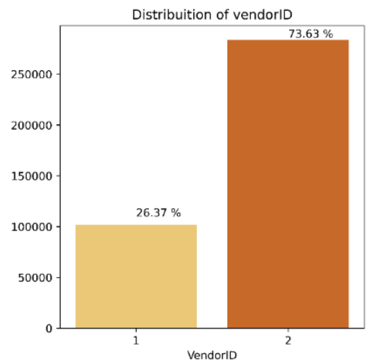
- **Extra:**
  - It was observed that the vendor id 1 and 2 had different ways of adding to the total amount.
  - The vendor id 2 added all the charges to come up with the total amount whereas the vendor id 1 added the congestion surcharge and Airport fee to the extras and then added the total amount.
  - To make this uniform, the congestion surcharge and Airport fee is removed from extras of vendor id 1.
  - Only the appropriate rows with extra values are kept and others are removed.

- **MTA Tax:**
  - MTA Tax should be 0.5 in all the cases so imputing 0 with 0.5 where ever present.
- **Fare amount:**
  - Rate code ID 2 is between JFK and Manhattan and the fare amount if flat 70. Imputing with 70 for 2 rows.
  - Dropping the rows with high fare amount but very low trip distance or trip duration.
  - Dropping the rows that start from JFK airport and had end location as JFK airport with a flat rate of 70 in fare amount. (one of the locations should have been Manhattan) since the destination cannot be determined these are dropped.
  - Congestion charge is only calculated if the trip touches 96th street of south Manhattan. It was noticed that even if the trip is within some other borough or different boroughs with less distance which cannot possibly touch the street had congestion charges. Therefore, they are appropriately imputed.
- **Improvement surcharge:**
  - Should be 1 for all. The rows with 0 had lot of other improper values. Dropping them.
  - The rows with wrong values of 0.3 – no specific pattern found, but the charge of 0.3 for improvement surcharge is charged before 2022 this can be considered as a mistake by driver and imputed with 1.
- **Tip amount:**
  - Tip amount has been mentioned for cash payment type in 2 rows which should not be the case as per data dictionary. Imputed with 0.
- **Airport Fee:**
  - Airport fee should be collected only for the trips that start from the JFK or LaGuardia airports and the amount is 1.75 but, in some occurrences, it has been collected from other irrelevant locations. Also, the old charge of 1.25 have been collected by the drivers in many occurrences.
  - Therefore, these are changed accordingly.
- **Total Amount:**
  - In few occurrences, the total amount was not equal to addition of all the other charges. They are also imputed with the right amount.

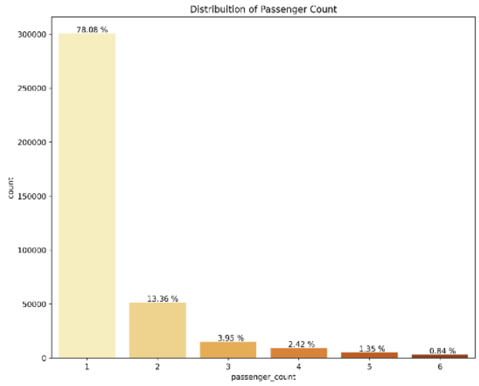
Shape of the dataset after cleaning: 132032 rows x 19 columns

Exploration for model building and finding patterns:

Univariate Analysis:



Vendor ID: The taxis which make 73.63% are being serviced by Vendor id 2 and the remaining are being serviced by the other vendor



Passenger count: Majority of the trips are taken by single passenger.

Ratecode ID: 95.5% of the rides are from rate code id 1, which is standard rate.

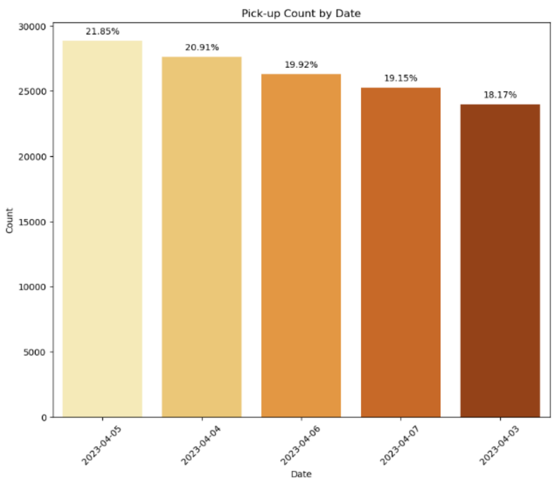
Store and Forward flag: During 99.51% of the times the vehicle had connection with the server to receive card payments from the passengers. However, there were connection issues during 0. 49% of rides, in which the payment details were stored and deducted later.

Payment type: A huge chunk of payments, around 82% are made through credit cards. Payment for around 16% of rides are completed through cash with some amount of no charge and disputed payments being the rest.

Trips - Pickup borough wise:

- 91.23% of the trips had the pickups from Manhattan
- 8.48% of the trips had the pickups from Queens
- 0.26% of the trips had the pickups from Brooklyn
- 0.02% of the trips had the pickups from Bronx
- Staten Island and EWR has close to 0 pickups

Number of trips each day:

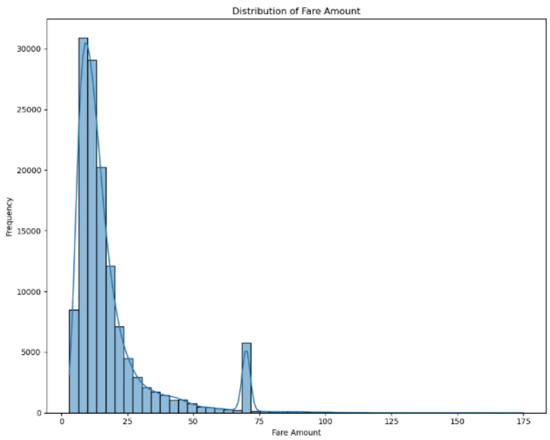


Congestion surcharge: Almost 95% of the trips pass through Manhattan south of 96th Street

Airport Trips: 7.86% of the trips were started from Laguardia and JFK airports and the trips were started from other locations.

Trip Duration: Most Trips are observed to be under an hour that is less than 5000 seconds

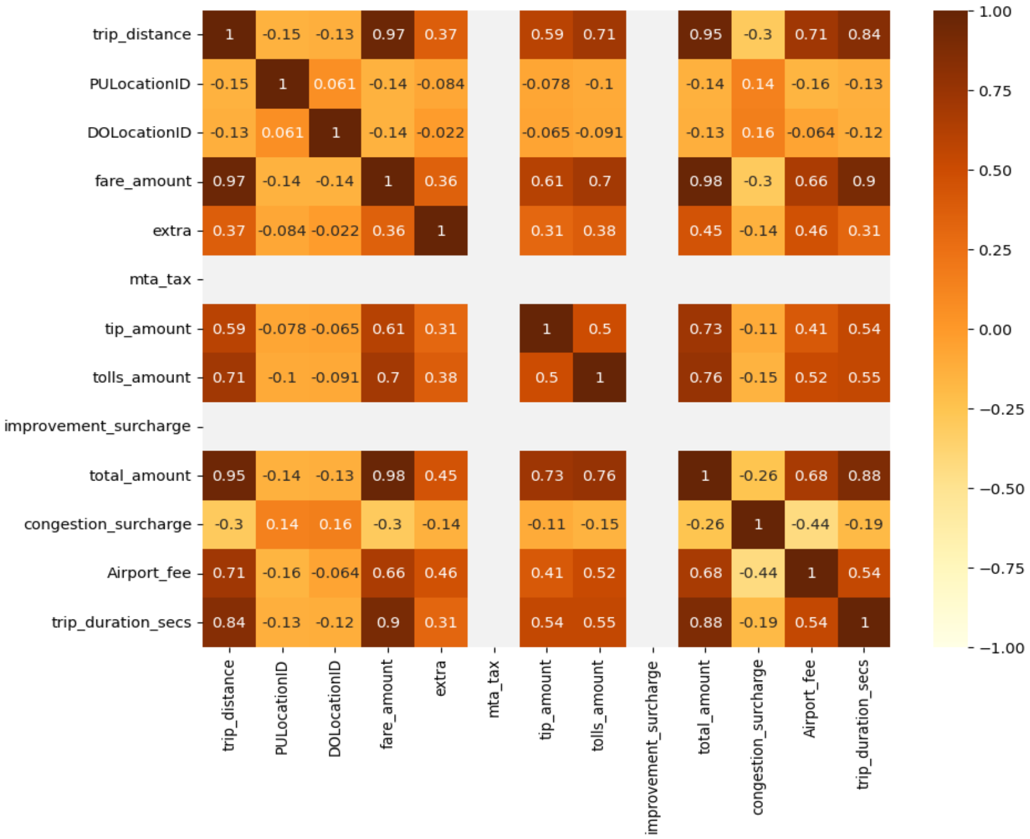
Fare Amount:



1) Most of the trips had fare amount charges below 20 dollars.

2) The graph showing a small peak at 70 indicating the flat rate charged for rate code ID 2 trips.

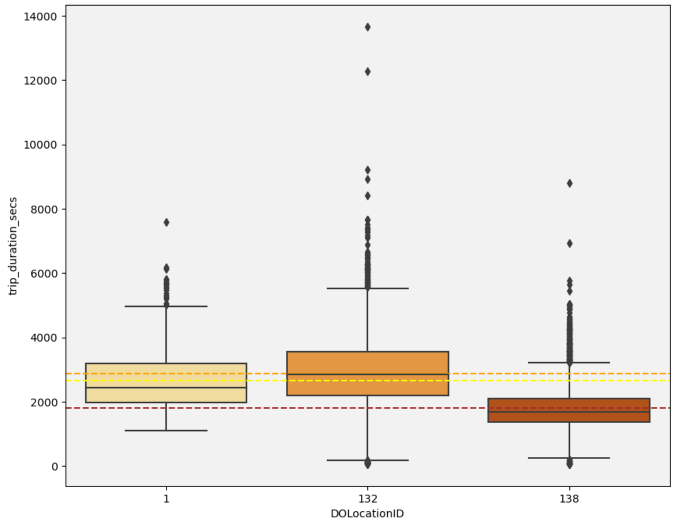
Bivariate Analysis:



As expected, the fare amount and trip duration have the highest correlations with the total amount. Also, the tolls amount and tip amount has a good amount of positive correlation with the total amount.

Multivariate analysis:

Airport Trips:



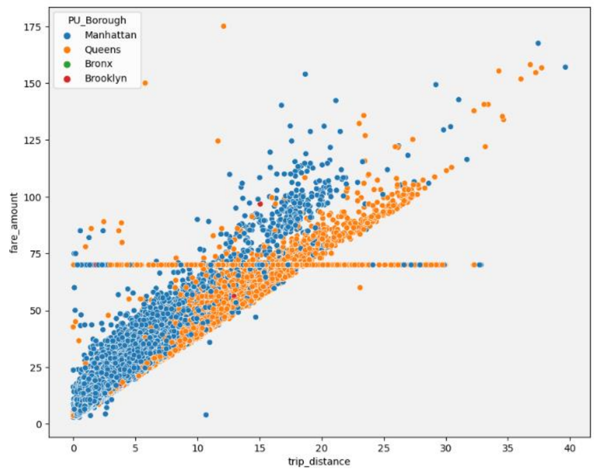
The average trip duration for drop offs in the JFK airport was the most.

The longest trip duration was to the JFK Airport.

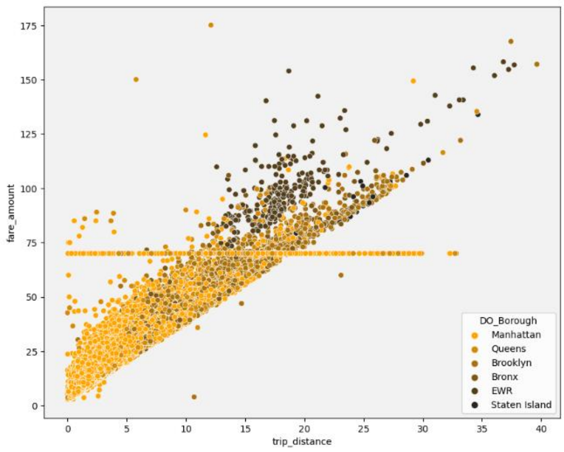
The JFK airport is far than the other airports therefore the durations are higher than the others.

This reveals that all of these routes have similar congestions.

Fare amount vs trip distance (with PU Borough):



For similar trip distances, most trips starting from Manhattan has more fare amounts than the other boroughs.



Fare amount vs trip distance (with DO Borough):

As the trip distances increases, the drop offs concentrated in the EWR borough.

The increase of fare amount with the trip distance remains more or less constant accross all the Boroughs.

Congested routes and high taxi trips during the congested hour:

Most congested locations are the ones where traffic regulations should be improved. This analysis gives out the most congested locations and routes to help the government prioritize any changes/regulations to these locations and routes.
During the peak hours, the following location ids are receiving the highest trips: 161, 162, 237, 132, 236,230, 142

Top 5 congested routes:

	pickup_day	PULocationID	DOLocationID	time_per_mile	time_per_mile_rank
3609	Monday	42	168	1022.62	3363.0
3679	Monday	45	66	961.39	3362.0
4446	Monday	114	209	936.71	3361.0
5903	Monday	211	144	891.85	3360.0
6237	Monday	234	65	880.27	3359.0

Top 5 less congested routes:

	pickup_day	PULocationID	DOLocationID	time_per_mile	time_per_mile_rank
5950	Monday	226	148	56.87	1.0
4078	Monday	82	170	75.49	2.0
6439	Monday	238	1	80.33	3.0
5929	Monday	216	132	81.25	4.0
5857	Monday	197	132	82.58	5.0

Congested short routes and better alternatives:

Taking short routes cannot be always the fast routes. Our analysis has found the short routes which take more time than its long route counterparts. This analysis has been done to suggest alternative routes for the drivers to take (when the passenger didn't ask for a specific route to take) so as to avoid congestion and to complete the rides faster. When properly executed the traffic may be split across different routes benefiting all the stakeholders.

better route:

	PULocationID	DOLocationID	trip_occurance	dist_lower_bound	dist_upper_bound	average_trip_dur	Rankfilter
256	13	163	6	6.0	8.0	1397	2.0

congested short distance:

	PULocationID	DOLocationID	trip_occurance	dist_lower_bound	dist_upper_bound	average_trip_dur	Rankfilter
255	13	163	19	4.0	6.0	1573	1.0

Hour wise top pick-up demand locations:

This analysis was made to give locations where more rides are happening to the drivers to benefit them from running more rides.

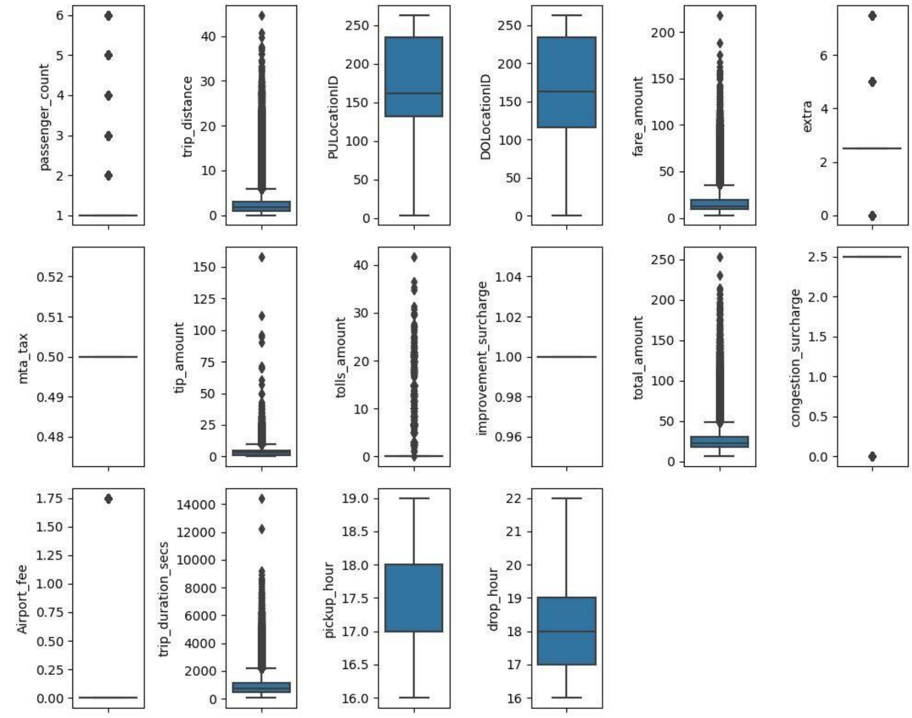
Top 5 Locations with most demand:

	pickup_hour	PULocationID	trip_occurence	trip_rank
62	16	161	1918	109.0
45	16	132	1596	108.0
94	16	237	1527	107.0
93	16	236	1368	106.0
63	16	162	1182	105.0

Top 5 Locations with least demand :

	pickup_hour	PULocationID	trip_occurence	trip_rank
15	16	51	1	16.0
16	16	52	1	16.0
17	16	56	1	16.0
29	16	89	1	16.0
31	16	91	1	16.0

Outliers and treatment:



The outliers present in few of the important features of the dataset are highly possible in the real world. The extreme outliers were removed and the remaining were chosen to be kept as it is considered important for our model to understand and work with all of these possible real-world scenarios and the number of rows were significant.

Overview:

Machine Learning Models built: 2 models

1. Total amount prediction

2. Trip duration prediction

Process Flow:

2 different kinds of models were built for each of the above models.

1st method:

- Take in to account all the features available in the dataset
- Feature engineering (if any)
- Scaling and proper encodings are performed.
- Use different methods such as no variance, correlation (heat map), statistical analysis and feature selection methods to select features for building different variants of Linear regression models
- For each variant the model is validated for assumptions
- As the assumptions are not met, a score card of other models such as KNN, Decision Tree, Random Forest, Ada boost, Gradient boost and XG boost regressors are built for both train and test sets.
- The best performing model will then be selected using R2 and other error values based on good scores, consistency and feature importance (which are not concentrated on 3/4 features).
- Grid search CV will then be implemented on the best algorithm and a potential final model is built.
- Cross validation is performed, if generalized well, the model is chosen as the final model.

2nd method:

- Take in to account only the logical features (the features that will be known before the start of the taxi ride in the real-world scenario), selected using domain knowledge.
- Scaling and proper encodings are performed.
- Linear regression models are skipped as assumptions are not met for any of the models in the 1st method
- A score card of other models such as KNN, Decision Tree, Random Forest, Ada boost, Gradient boost and XG boost regressors are built for both train and test sets.
- The best performing model will then be selected using R2 and other error values based on good scores, consistency and feature importance (which are not concentrated on 3/4 features).
- Grid search CV will then be implemented on the best algorithm and a potential final model is built.
- Cross validation is performed, if generalized well, the model is chosen as the final model.

1st method:

Features: All the features in the dataset are taken into account.

Feature engineering:

The Pick up and drop off date time columns have been split in to different categorical features such as **day, hour and minutes** for both pick up and drop off as the regression models cannot handle time series data directly.

By using the Pickup and drop off times, a **trip duration** column was created.

Statistical analysis:

After performing the general Exploratory data analysis, statistical analysis of checking the significance of the independent variables on the target variables are performed.

First, for each independent column and the target column, the assumptions of normality and equal variance of parametric tests are performed, when it is not satisfied then assumptions of non-parametric tests are made before proceeding with the tests.

The following parametric and non-parametric assumptions/tests were performed as the target column is numeric.

Type	Parametric tests	Non parametric tests	Assumptions of non-parametric tests
More than 2 categories vs numerical	One-way anova	Kruskal wallis H	Independence of observations and similar distribution of groups
2 categories vs numerical	T test independent	Mann-Whitney U	Independence of observations and similar distribution of groups
Numerical vs numerical	Pearson-R	Spearman R	Independence of observations and Monotonicity

Results of the statistical analysis: All of the independent variables had effect on the target variable. This might have also happened as we chose to keep the outliers found in the data and sometimes the results might have got influenced by the presence of it.

Scaling and Encodings:

Numerical columns had small values in several columns and large in others therefore **standard scaler** has been used to transform all the numeric columns.

One Hot encoding has been used for the categorical features where categories were a few.

Target encoding has been used for the categorical features with high number of categories for majority of the variants of the models ensuring no data leakage. The data leakage problem was avoided by encoding the train and test sets separately.

<u>Linear Regression model – Variant 1:</u>

Feature selection:

Columns with 0 standard deviations are removed – MTA tax and Improvement Surcharge.

Columns which gave redundant information are removed – PU Borough, DO Borough, Service zones, PU Zone, DO Zone, Drop off day.

All other features were used for the next step.

Encoding: OHE for categorical columns with few categories

Assumptions before model:

1. Numeric target column – satisfied

2. Multicollinearity – to reduce the multicollinearity present in the dataset Variance Inflation Factor method has been used with the threshold set to 5.

Features before VIF – 37

Features after VIF – 32

Train Test split: 70:30 train test split was performed on the dataset

OLS model:

With the obtained features an OLS model was built and fitted to the train set, following is the summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           total_amount   R-squared:                       0.959
Model:                            OLS   Adj. R-squared:                  0.959
Method:                 Least Squares   F-statistic:                 6.831e+04
Date:                Sun, 15 Oct 2023   Prob (F-statistic):               0.00
Time:                        15:50:34   Log-Likelihood:            -2.6288e+05
No. Observations:               92410   AIC:                         5.258e+05
Df Residuals:                   92377   BIC:                         5.261e+05
Df Model:                          32
Covariance Type:            nonrobust

==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                     27.4719      0.056    493.397      0.000      27.363      27.581
extra                      0.7891      0.022     36.170      0.000       0.746       0.832
tip_amount                 5.0013      0.020    244.365      0.000       4.961       5.041
tolls_amount               3.8576      0.024    162.126      0.000       3.811       3.904
congestion_surcharge      -0.3307      0.017    -19.837      0.000      -0.363      -0.298
Airport_fee                3.0504      0.024    125.976      0.000       3.003       3.098
trip_duration_secs        10.1679      0.021    492.733      0.000      10.127      10.208
VendorID_2                 0.6814      0.032     21.533      0.000       0.619       0.743
passenger_count_2          0.0866      0.039      2.214      0.027       0.010       0.163
passenger_count_3         -0.0303      0.066     -0.457      0.647      -0.160       0.100
passenger_count_4          0.1113      0.082      1.361      0.174      -0.049       0.272
passenger_count_5          0.0269      0.122      0.220      0.826      -0.213       0.267
passenger_count_6         -0.0511      0.155     -0.330      0.741      -0.354       0.252
RatecodeID_2.0            10.0206      0.099    101.355      0.000       9.827      10.214
RatecodeID_3.0            25.4919      0.372     68.477      0.000      24.762      26.222
RatecodeID_4.0            33.3013      1.157     28.772      0.000      31.033      35.570
RatecodeID_5.0            25.6211      0.593     43.198      0.000      24.459      26.784
store_and_fwd_flag_Y       0.0605      0.198      0.306      0.760      -0.327       0.448
payment_type_2             1.1407      0.044     25.770      0.000       1.054       1.227
payment_type_3             1.0349      0.257      4.024      0.000       0.531       1.539
payment_type_4             1.2588      0.183      6.864      0.000       0.899       1.618
pickup_hour_17             1.0892      0.050     21.810      0.000       0.991       1.187
pickup_hour_18             1.8266      0.052     35.379      0.000       1.725       1.928
pickup_hour_19             1.2020      0.046     26.058      0.000       1.112       1.292
drop_hour_17              -0.8924      0.050    -17.700      0.000      -0.991      -0.794
drop_hour_18              -1.2960      0.051    -25.555      0.000      -1.395      -1.197
drop_hour_20               2.3758      0.077     30.788      0.000       2.225       2.527
drop_hour_21             -20.0731      2.409     -8.332      0.000     -24.795     -15.351
pickup_day_Monday         -0.1521      0.045     -3.386      0.001      -0.240      -0.064
pickup_day_Thursday       -0.0662      0.044     -1.508      0.132      -0.152       0.020
pickup_day_Tuesday         0.0701      0.043      1.614      0.106      -0.015       0.155
pickup_day_Wednesday      -1.4078      0.043    -32.575      0.000      -1.492      -1.323
pickup_mins                0.0027      0.001      3.291      0.001       0.001       0.004
==============================================================================

------------------------------------------------------------------------------
Omnibus:                    48738.420   Durbin-Watson:                   2.013
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          6108376.702
Skew:                           1.556   Prob(JB):                         0.00
Kurtosis:                      42.708   Cond. No.                     5.96e+03
==============================================================================
```

The summary shows the presence of 7 insignificant features as per the p value of t test.

Though the whole model is statistically significant, other after model assumptions like normality of residuals etc. and multicollinearity did not meet.

Keeping these in mind, the next variant of the model was built.

<u>Linear Regression model – Variant 2:</u>

Transformation:

The dependent variable was transformed using Reciprocal transformation to make it the feature more normally distributed to try to meet the assumptions of linear regression model.

Train Test split: 70-30 split was made to the dataset

Encoding: along with the OHE was some features, Target encoding was done to PU minutes, Pick up location and dropped of location. TE was done after Train test split to avoid data leakage.

Assumptions before model:

1. Numeric target column – satisfied

2. Multicollinearity – to reduce the multicollinearity present in the dataset Variance Inflation Factor method has been used with the threshold set to 5.

Features before VIF – 37

Features after VIF – 31

OLS model:

With the obtained features an OLS model was built and fitted to the train set,

The results had 6 statistically insignificant variables in the t test. Though the entire model was statistically significant, the assumptions of Linear regression still did not meet.

<u>Linear Regression model – Variant 3:</u>

Feature selection:

RFE was performed to select features from the 37 features with which the other models were created. **18 features** were selected as best features. The chosen features were selected directly from the previously encoded train test split

OLS model: With the obtained features an OLS model was built and fitted to the train set

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           total_amount   R-squared:                       0.804
Model:                            OLS   Adj. R-squared:                  0.804
Method:                 Least Squares   F-statistic:                 2.101e+04
Date:                Sun, 15 Oct 2023   Prob (F-statistic):               0.00
Time:                        16:27:36   Log-Likelihood:             3.1568e+05
No. Observations:               92410   AIC:                        -6.313e+05
Df Residuals:                   92391   BIC:                        -6.311e+05
Df Model:                          18
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                     0.0265      0.003     10.275      0.000       0.021       0.032
PULocationID              0.1563      0.005     34.573      0.000       0.147       0.165
DOLocationID             -0.0150      0.005     -3.157      0.002      -0.024      -0.006
pickup_mins               0.2104      0.057      3.686      0.000       0.099       0.322
trip_distance             0.0072      0.000     54.848      0.000       0.007       0.007
fare_amount              -0.0195      0.000   -122.923      0.000      -0.020      -0.019
tip_amount               -0.0014   3.93e-05    -35.074      0.000      -0.001      -0.001
congestion_surcharge     -0.0022   3.18e-05    -70.021      0.000      -0.002      -0.002
trip_duration_secs       -0.0053   7.41e-05    -71.422      0.000      -0.005      -0.005
RatecodeID_2.0            0.0252      0.000    118.080      0.000       0.025       0.026
RatecodeID_3.0            0.0412      0.001     63.036      0.000       0.040       0.042
RatecodeID_4.0            0.0447      0.002     20.080      0.000       0.040       0.049
RatecodeID_5.0            0.0196      0.001     16.935      0.000       0.017       0.022
payment_type_2            0.0068   8.45e-05     79.969      0.000       0.007       0.007
payment_type_3            0.0108      0.000     22.087      0.000       0.010       0.012
payment_type_4            0.0085      0.000     24.169      0.000       0.008       0.009
drop_hour_20             -0.0030      0.000    -23.067      0.000      -0.003      -0.003
drop_hour_21              0.0248      0.005      5.379      0.000       0.016       0.034
pickup_day_Wednesday      0.0013    6.5e-05     19.652      0.000       0.001       0.001
==============================================================================
Omnibus:                    34224.874   Durbin-Watson:                   1.991
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           265511.615
Skew:                           1.580   Prob(JB):                         0.00
Kurtosis:                      10.679   Cond. No.                     4.00e+03
==============================================================================
```

All the selected features and the model was found to be significant. But the assumption of normality of residuals did not get satisfied and multicollinearity was still present. A model built with SFS was also found not performing well with insignificant features.

Linear Regression model – Variant 4:

For all the 37 features used in the variant 2, a Lasso regression model was built. The coefficients displayed usage of only 2 features out of 37 for the model. Therefore, a grid search CV was made to check if the model performed better.

Results of the GSCV: {'alpha': 0.001}

Another model was built with the selected parameters. Only 3 features were used Fare amount, Trip amount and Trip duration. Assumptions like Multicollinearity met. but both Adjusted R2 and R2 scores dropped significantly.

Also, assumptions of linear regression model did not meet. As a final measure, the selected independent features were transformed to see for improvement in assumptions of the model. Though the R2 stores significantly improved, the assumptions of the linear regression model such as normality of residuals, Homoskedasticity and linearity with fitted values assumptions of the model did not get satisfied.

As the linear regression assumptions are not satisfied after trying different variants of the model, other non-parametric regressor models such as K Nearest Regressor, Decision tree Regressor, Radom Forest Regressor, ADA boost Regressor, Gradient and XG Boost Regressors were built.

Other models:

A score card of R2, MSE, RMSE, MAE, MAPE for both train and test sets for all the models were built. Following is the result.

	MODEL	DATA	R2	MAE	MSE	RMSE	MAPE
0	K-NEAREST REGRESSOR	TRAIN	0.986768	1.341923	5.657011	2.378447	0.048605
1	K-NEAREST REGRESSOR	TEST	0.968097	2.069158	13.456162	3.668264	0.074284
2	DECISION TREE REGRESSOR	TRAIN	0.966044	2.282833	14.516541	3.810058	0.086593
3	DECISION TREE REGRESSOR	TEST	0.965031	2.299305	14.749414	3.840497	0.086637
4	RANDOM FOREST REGRESSOR	TRAIN	0.999076	0.159464	0.394966	0.628463	0.005256
5	RANDOM FOREST REGRESSOR	TEST	0.994183	0.646291	2.453594	1.566395	0.021906
6	ADA BOOST REGRESSOR	TRAIN	0.465186	13.933121	228.639566	15.120832	0.718141
7	ADA BOOST REGRESSOR	TEST	0.455730	13.969589	229.562306	15.151314	0.718143
8	GRADIENT BOOST REGRESSOR	TRAIN	0.994821	0.812273	2.213895	1.487916	0.028143
9	GRADIENT BOOST REGRESSOR	TEST	0.993897	0.846310	2.573965	1.604358	0.028860
10	XGBOOST REGRESSOR	TRAIN	0.998630	0.483223	0.585686	0.765301	0.018626
11	XGBOOST REGRESSOR	TEST	0.995602	0.627562	1.854805	1.361912	0.022142

From the features used in the first variant of the linear model, the features of fare amount, drop hour and vendor id were removed as it gave redundant information and feature concentration.

The best performing model was selected using R2 and other error values - based on good scores, consistency and feature importance. The Xgboost regressor had the best performance with good feature importance.

Feature importance:

trip_distance	0.724232
trip_duration_secs	0.077165
RatecodeID_2.0	0.061042
tip_amount	0.045186
tolls_amount	0.040914
RatecodeID_3.0	0.012944
RatecodeID_5.0	0.010112
extra	0.008833
RatecodeID_4.0	0.005039
congestion_surcharge	0.003700
Airport_fee	0.002040
DOLocationID	0.001487
PULocationID	0.001050
passenger_count_2	0.000831
pickup_mins	0.000678

Grid Search CV:

Before performing grid search CV, the columns which had insignificant importance in the previously selected model (payment type, passenger count and store & forward flag) were removed.

Best Parameters of Grid Search CV: {'gamma': 0, 'learning_rate': 0.4, 'max_depth': 4, 'n_estimators': 200}

The new model built after the Grid Search Performed similar:

	MODEL	DATA	R2	MAE	MSE	RMSE	MAPE
0	XGBOOST REGRESSOR	TRAIN	0.998658	0.473828	0.573842	0.757524	0.018191
1	XGBOOST REGRESSOR	TEST	0.995584	0.623557	1.862566	1.364758	0.021946
2	XGBOOST REGRESSOR WITH GCV	TRAIN	0.998410	0.511570	0.679957	0.824595	0.019375
3	XGBOOST REGRESSOR WITH GCV	TEST	0.995750	0.628803	1.792432	1.338817	0.022236

Cross Validation:

A separate 5 fold cross validation was performed by separately encoding train and test sets. The results were consistent.

results:

[0.9959468712697152, 0.9931604614226276, 0.9954494219631257, 0.9952487039027472, 0.9943398060405997]

pickup_day_Thursday	0.000613
pickup_hour_18	0.000607
pickup_day_Tuesday	0.000603
pickup_hour_19	0.000486
payment_type_2	0.000407
pickup_day_Wednesday	0.000315
pickup_day_Monday	0.000292
pickup_hour_17	0.000237
payment_type_3	0.000231
passenger_count_4	0.000227
passenger_count_3	0.000198
store_and_fwd_flag_Y	0.000189
payment_type_4	0.000135
passenger_count_6	0.000122
passenger_count_5	0.000084

Conclusion:

As the model performed well in train and test sets, also generalized well in the cross validation with score as high as 99, therefore the model was chosen as the final model.

2nd Method:

This method is performed considering the real-world scenario,

Feature selection:

The feature selection was performed based on the domain knowledge. All the features which will be known before the start of the trip and will not be relevant in predicting the total amount in the real-world scenario are chosen.

Feature engineering:

The Pick up and drop off date time columns have been split into different categorical features such as **day, hour and minutes** for both pick up and drop off as the regression models cannot handle time series data directly.

Scaling and Encodings:

Numerical columns had small values in several columns and large in others therefore **standard scaler** has been used to transform all the numeric columns.

One Hot encoding has been used for the categorical features where categories were a few.

Target encoding has been used for the categorical features with high number of categories for majority of the variants of the models ensuring no data leakage. The data leakage problem was avoided by encoding the train and test sets separately.

Model Building:

A score card of R2, MSE, RMSE, MAE, MAPE for both train and test sets for all the models were built. Following is the result.

	MODEL	DATA	R2	MAE	MSE	RMSE	MAPE
0	K-NEAREST REGRESSOR	TRAIN	0.962085	2.700069	16.209082	4.026050	0.101433
1	K-NEAREST REGRESSOR	TEST	0.905863	4.378675	39.705333	6.301217	0.167158
2	DECISION TREE REGRESSOR	TRAIN	0.944255	3.236258	23.831876	4.881790	0.118577
3	DECISION TREE REGRESSOR	TEST	0.945580	3.232409	22.953231	4.790953	0.117866
4	RANDOM FOREST REGRESSOR	TRAIN	0.994073	1.037813	2.533659	1.591747	0.037547
5	RANDOM FOREST REGRESSOR	TEST	0.957906	2.840016	17.754228	4.213577	0.102771
6	ADA BOOST REGRESSOR	TRAIN	0.686235	9.909049	134.138556	11.581820	0.463160
7	ADA BOOST REGRESSOR	TEST	0.683679	9.883858	133.417990	11.550671	0.460836
8	GRADIENT BOOST REGRESSOR	TRAIN	0.960541	2.764811	16.869270	4.107222	0.100009
9	GRADIENT BOOST REGRESSOR	TEST	0.959667	2.772439	17.011466	4.124496	0.099487
10	XGBOOST REGRESSOR	TRAIN	0.972378	2.414157	11.808938	3.436414	0.090889
11	XGBOOST REGRESSOR	TEST	0.959332	2.770775	17.152783	4.141592	0.098772

The best performing model was selected using R2 and other error values - based on good scores, consistency and feature importance.

The Xgboost regressor had the best performance with good feature importance.

Feature importance:

trip_distance	0.652795
RatecodeID_2.0	0.108833
tolls_amount	0.070926
RatecodeID_5.0	0.029349
congestion_surcharge	0.020446
extra	0.019491
pickup_day_Wednesday	0.017634
RatecodeID_3.0	0.016956
pickup_hour_19	0.015957
RatecodeID_4.0	0.011602
pickup_hour_18	0.009285
DOLocationID	0.005138
PULocationID	0.004756
pickup_day_Monday	0.004202
pickup_hour_17	0.003422
pickup_mins	0.002563
pickup_day_Tuesday	0.002384
Airport_fee	0.002353
pickup_day_Thursday	0.001909

Grid Search CV:

Grid search CV was performed to find the best parameters of the chosen model.

The following are the best parameters chosen for the XGBoost regressor model.

{'gamma': 0, 'learning_rate': 0.1, 'max_depth': 6, 'n_estimators': 200}

Performance of model after using the best parameters:

16	XGBOOST REGRESSOR WITH GCV	TRAIN	0.970228	2.489653	12.727945	3.567625	0.092501
17	XGBOOST REGRESSOR WITH GCV	TEST	0.960704	2.727109	16.574191	4.071141	0.097374

The performance of the model improved and became consistent after using the selected parameters. To check the generalization Cross validation was performed.

Cross Validation:

A separate 5 fold cross validation was performed by separately encoding train and test sets. The results were consistent.

results:

[0.9583076375511895, 0.9530383062238799, 0.9577956375110569, 0.9563562167506233, 0.9559003095113301]

Conclusion:

As the model performed well in train and test sets, also generalized well in the cross validation with score as high as 95, therefore the model was chosen as the final model.

1st Method:

Features: All the features in the dataset are taken into account.

Feature engineering:

The Pick up and drop off date time columns have been split in to different categorical features such as **day and hour** for both pick up and drop off as the regression models cannot handle time series data directly.

By using the Pickup and drop off times, a **trip duration in minutes** column was created as target variables.

Statistical analysis:

After performing the general Exploratory data analysis, statistical analysis of checking the significance of the independent variables on the target variables are performed.

First, for each independent column and the target column, the assumptions of normality and equal variance of parametric tests are performed, when it is not satisfied then assumptions of non-parametric tests are made before proceeding with the tests.

The following parametric and non-parametric assumptions/tests were performed as the target column is numeric.

Type	Parametric tests	Non parametric tests	Assumptions of non-parametric tests
More than 2 categories vs numerical	One-way anova	Kruskal wallis H	Independence of observations and similar distribution of groups
2 categories vs numerical	T test independent	Mann-Whitney U	Independence of observations and similar distribution of groups
Numerical vs numerical	Pearson-R	Spearman R	Independence of observations and Monotonicity

Results of the statistical analysis: All of the independent variables had effect on the target variable except vendor ID and Store & forward flag.

Transformation:

The target variable was transformed using Log transformation and the independent numerical variables were transformed using box-cox, square root and reciprocal transformations to meet the assumptions of linear regression model.

Scaling and Encodings:

Numerical columns had small values in several columns and large in others therefore **standard scaler** has been used to transform all the numeric columns.

One Hot encoding has been used for the categorical features where categories were a few.

Target encoding has been used for the categorical features with high number of categories for majority of the variants of the models ensuring no data leakage. The data leakage problem was avoided by encoding the train and test sets separately.

Linear Regression model – Variant 1:

Encoding: OHE for categorical columns with few categories

Feature selection:

Columns with 0 standard deviations are removed – MTA tax and Improvement Surcharge.

Columns which gave redundant information are removed – PU Borough, DO Borough, Service zones, PU Zone, DO Zone, Drop off day.

Dropped 15 features which showed multicollinearity in the heat map of correlation matrix. All other features were used for the next step.

Assumptions before model:

1. Numeric target column – satisfied

2. Multicollinearity – to reduce the multicollinearity present in the dataset Variance Inflation Factor method has been used with the threshold set to 5.

Features before VIF – 25

Features after VIF – 24

Train Test split: 80:20 train test split was performed on the dataset

Target encoding was performed for the DO location column

OLS model:

With the obtained features an OLS model was built and fitted to the train set, following is the summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     trip_duration_mins   R-squared:                       0.805
Model:                            OLS   Adj. R-squared:                  0.805
Method:                 Least Squares   F-statistic:                 1.812e+04
Date:                Sun, 15 Oct 2023   Prob (F-statistic):               0.00
Time:                        02:05:49   Log-Likelihood:                -28570.
No. Observations:              105612   AIC:                         5.719e+04
Df Residuals:                  105587   BIC:                         5.743e+04
Df Model:                          24
Covariance Type:            nonrobust

=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                     2.0093      0.010    191.525      0.000       1.989       2.030
trip_distance             0.6130      0.001    561.888      0.000       0.611       0.615
congestion_surcharge      0.0512      0.001     47.622      0.000       0.049       0.053
VendorID_2               -0.0423      0.002    -18.769      0.000      -0.047      -0.038
passenger_count_2         0.0208      0.003      7.488      0.000       0.015       0.026
passenger_count_3         0.0458      0.005      9.815      0.000       0.037       0.055
passenger_count_4         0.0705      0.006     12.154      0.000       0.059       0.082
passenger_count_5        -0.0122      0.009     -1.394      0.163      -0.029       0.005
passenger_count_6        -0.0071      0.011     -0.635      0.525      -0.029       0.015
RatecodeID_3.0           -0.0011      0.022     -0.049      0.961      -0.044       0.042
RatecodeID_4.0            0.0097      0.079      0.122      0.903      -0.146       0.165
RatecodeID_5.0            0.0970      0.040      2.419      0.016       0.018       0.176
store_and_fwd_flag_Y     -0.0181      0.014     -1.279      0.201      -0.046       0.010
payment_type_3            0.0495      0.018      2.731      0.006       0.014       0.085
payment_type_4            0.0488      0.013      3.699      0.000       0.023       0.075
pickup_hour_17           -0.0332      0.003    -12.083      0.000      -0.039      -0.028
pickup_hour_18           -0.1328      0.003    -48.574      0.000      -0.138      -0.127
pickup_hour_19           -0.2746      0.003    -89.941      0.000      -0.281      -0.269
drop_hour_20              0.0353      0.005      6.724      0.000       0.025       0.046
drop_hour_21              0.9257      0.224      4.126      0.000       0.486       1.365
pickup_day_Monday      9.088e-06      0.003      0.003      0.998      -0.006       0.006
pickup_day_Thursday       0.0114      0.003      3.647      0.000       0.005       0.018
pickup_day_Tuesday       -0.0213      0.003     -6.888      0.000      -0.027      -0.015
pickup_day_Wednesday      0.1918      0.003     62.603      0.000       0.186       0.198
DOLocationID              0.2274      0.004     56.887      0.000       0.220       0.235
==============================================================================
Omnibus:                    20060.469   Durbin-Watson:                   2.005
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           291162.271
Skew:                           0.499   Prob(JB):                         0.00
Kurtosis:                      11.073   Cond. No.                         656.
==============================================================================
```

The summary shows the presence of 6 insignificant features as per the p value of t test.

Though the whole model is statistically significant, all the other after model assumptions like normality of residuals except auto correlation did not satisfy and multicollinearity was moderate.

Keeping these in mind, the next variant of the model was built.

Linear Regression model – Variant 2:

Transformation: The dependent variable was transformed using Reciprocal transformation to make it the feature more normally distributed to try to meet the assumptions of linear regression model.

Encoding: OHE for categorical columns with few categories

Feature selection: SFS (forward selection method) using the best features as parameter

Features selected: 32 features

Assumptions before model:

1. Numeric target column – satisfied

2. Multicollinearity – to reduce the multicollinearity present in the dataset Variance Inflation Factor method has been used with the threshold set to 5.

Features before VIF – 32

Features after VIF – 26

Train Test split: 80:20 train test split was performed on the dataset

Target encoding was performed for the DO location column

OLS model: With the obtained features an OLS model was built and fitted to the train set, following is the summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      trip_duration_mins   R-squared:                       0.805
Model:                             OLS   Adj. R-squared:                  0.805
Method:                  Least Squares   F-statistic:                 1.680e+04
Date:                 Sun, 15 Oct 2023   Prob (F-statistic):               0.00
Time:                         02:31:15   Log-Likelihood:                -28387.
No. Observations:               105612   AIC:                         5.683e+04
Df Residuals:                   105585   BIC:                         5.709e+04
Df Model:                           26
Covariance Type:             nonrobust

==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                      2.5272      0.004    712.568      0.000       2.520       2.534
trip_distance              0.6267      0.001    482.101      0.000       0.624       0.629
extra                     -0.0103      0.001     -7.618      0.000      -0.013      -0.008
tip_amount                 0.0698      0.002     41.611      0.000       0.066       0.073
tolls_amount              -0.0323      0.002    -20.598      0.000      -0.035      -0.029
congestion_surcharge       0.0214      0.001     17.997      0.000       0.019       0.024
Airport_fee               -0.0257      0.002    -15.634      0.000      -0.029      -0.023
VendorID_2                -0.0460      0.002    -20.502      0.000      -0.050      -0.042
passenger_count_3          0.0457      0.005      9.863      0.000       0.037       0.055
passenger_count_4          0.0684      0.006     11.815      0.000       0.057       0.080
RatecodeID_2.0             0.1808      0.007     27.331      0.000       0.168       0.194
RatecodeID_3.0             0.1420      0.023      6.276      0.000       0.098       0.186
RatecodeID_4.0             0.1063      0.079      1.342      0.180      -0.049       0.262
RatecodeID_5.0             0.0648      0.041      1.597      0.110      -0.015       0.144
store_and_fwd_flag_Y      -0.0118      0.014     -0.835      0.403      -0.040       0.016
payment_type_2             0.1783      0.004     43.762      0.000       0.170       0.186
pickup_hour_17            -0.0681      0.004    -19.279      0.000      -0.075      -0.061
pickup_hour_18            -0.1689      0.004    -46.494      0.000      -0.176      -0.162
pickup_hour_19            -0.2540      0.003    -78.370      0.000      -0.260      -0.248
drop_hour_17               0.0424      0.004     11.822      0.000       0.035       0.049
drop_hour_18               0.0679      0.004     18.925      0.000       0.061       0.075
drop_hour_20               0.0360      0.005      6.858      0.000       0.026       0.046
drop_hour_21               0.9535      0.224      4.257      0.000       0.514       1.392
dropoff_day_Monday        -0.0055      0.003     -1.723      0.085      -0.012       0.001
dropoff_day_Thursday       0.0094      0.003      3.005      0.003       0.003       0.015
dropoff_day_Tuesday       -0.0247      0.003     -7.990      0.000      -0.031      -0.019
dropoff_day_Wednesday      0.1834      0.003     59.968      0.000       0.177       0.189
==============================================================================
Omnibus:                    23791.691   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           402122.970
Skew:                           0.636   Prob(JB):                         0.00
Kurtosis:                      12.474   Cond. No.                         395.
==============================================================================
```

The summary shows the presence of 4 insignificant features as per the p value of t test.

Though the whole model is statistically significant, all the other after model assumptions like normality of residuals except auto correlation did not satisfy and multicollinearity was moderate.

<u>Linear Regression model – Variant 3:</u>

For all the transformed, scaled and encoded features from the first variant, a Lasso regression model was built.

Though the R2 stores improved to 0.914, the assumptions of the linear regression model such as normality of residuals, Homoskedasticity and linearity with fitted values assumptions of the model did not get satisfied.

As the linear regression assumptions are not satisfied after trying different variants of the model, other non-parametric regressor models such as K Nearest Regressor, Decision tree Regressor, Radom Forest Regressor, ADA boost Regressor, Gradient and XG Boost Regressors were built.

<u>Other models:</u>

A score card of R2, MSE, RMSE, MAE, MAPE for both train and test sets for all the models were built. Following is the result.

All the features used in the first variant of the linear regression model were used.

	MODEL	DATA	R2	MAE	MSE	RMSE	MAPE
0	DECISION TREE REGRESSOR	TRAIN	0.897752	2.037141	16.206121	4.025683	0.131303
1	DECISION TREE REGRESSOR	TEST	0.895624	2.087275	15.982624	3.997827	0.137585
2	RANDOM FOREST REGRESSOR	TRAIN	0.994710	0.431909	0.838458	0.915674	0.029617
3	RANDOM FOREST REGRESSOR	TEST	0.959978	1.200258	6.128440	2.475569	0.082760
4	K-NEAREST REGRESSOR	TRAIN	0.936345	2.036423	10.089181	3.176347	0.168578
5	K-NEAREST REGRESSOR	TEST	0.870085	2.950406	19.893207	4.460180	0.255091
6	ADA BOOST REGRESSOR	TRAIN	0.072116	10.669815	147.067907	12.127156	1.400997
7	ADA BOOST REGRESSOR	TEST	0.029762	10.701725	148.567622	12.188832	1.418181
8	GRADIENT BOOST REGRESSOR	TRAIN	0.944301	1.547363	8.828217	2.971232	0.106837
9	GRADIENT BOOST REGRESSOR	TEST	0.942573	1.563357	8.793483	2.965381	0.109800
10	XGBOOST REGRESSOR	TRAIN	0.980388	0.977369	3.108488	1.763090	0.070103
11	XGBOOST REGRESSOR	TEST	0.955438	1.236070	6.823564	2.612195	0.081660

The best performing model was selected using R2 and other error values - based on good scores, consistency and feature importance. The Gradient boost regressor had the best performance with good feature importance, consistent R2 and error values.

Grid Search CV:

Grid Search CV was performed to select best parameters

Best Parameters of Grid Search CV: {'learning_rate': 0.09000000000000001, 'max_depth': 5, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 3, 'n_estimators': 200}

The new model built after the Grid Search Performed better and also consitent:

	MODEL	DATA	R2	MAE	MSE	RMSE	MAPE
0	GRADIENT BOOST REGRESSOR TUNED	TRAIN	0.973482	1.070488	4.203033	2.050130	0.074895
1	GRADIENT BOOST REGRESSOR TUNED	TEST	0.964127	1.177121	5.493110	2.343739	0.081772

Cross Validation:

A separate 5 fold cross validation was performed by separately encoding train and test sets. The results were consistent.

results: [0.9623408399038307, 0.9585086703975807, 0.9612383863402351, 0.9570192344392388, 0.9590085017583008]

Conclusion:

As the model performed well in train and test sets, also generalized well in the cross validation with score as high as 96, therefore the model was chosen as the final model.

2^{nd} Method:

This method is performed considering the real-world scenario,

Feature selection:

The feature selection was performed based on the domain knowledge. All the features which will be known before the start of the trip and will not be relevant in predicting the total amount in the real-world scenario are chosen.

Feature engineering:

The Pick up and drop off date time columns have been split into different categorical features such as **day and hour** for both pick up and drop off as the regression models cannot handle time series data directly.

Scaling and Encodings:

Numerical columns had small values in several columns and large in others therefore **standard scaler** has been used to transform all the numeric columns.

One Hot encoding has been used for the categorical features where categories were a few.

Target encoding has been used for the categorical features with high number of categories for majority of the variants of the models ensuring no data leakage. The data leakage problem was avoided by encoding the train and test sets separately.

Model Building:

A score card of R2, MSE, RMSE, MAE, MAPE for both train and test sets for all the models were built. Following is the result.

	MODEL	DATA	R2	MAE	MSE	RMSE	MAPE
0	DECISION TREE REGRESSOR	TRAIN	0.998759	0.075292	0.196626	0.443425	0.008792
1	DECISION TREE REGRESSOR	TEST	0.743272	4.205060	39.311501	6.269888	0.335115
2	RANDOM FOREST REGRESSOR	TRAIN	0.982706	1.096386	2.741026	1.655604	0.088747
3	RANDOM FOREST REGRESSOR	TEST	0.857668	3.150613	21.794565	4.668465	0.258034
4	K-NEAREST REGRESSOR	TRAIN	0.921963	2.280542	12.368702	3.516917	0.182130
5	K-NEAREST REGRESSOR	TEST	0.796773	3.789945	31.119091	5.578449	0.332803
6	ADA BOOST REGRESSOR	TRAIN	-0.629670	14.837466	258.299622	16.071703	1.916213
7	ADA BOOST REGRESSOR	TEST	-0.715428	14.948406	262.674852	16.207247	1.947141
8	GRADIENT BOOST REGRESSOR	TRAIN	0.865046	3.072830	21.390025	4.624935	0.246447
9	GRADIENT BOOST REGRESSOR	TEST	0.857025	3.135179	21.893027	4.678999	0.253729
10	XGBOOST REGRESSOR	TRAIN	0.916332	2.508926	13.261201	3.641593	0.203338
11	XGBOOST REGRESSOR	TEST	0.830937	3.439897	25.887728	5.087998	0.263061

The best performing model was selected using R2 and other error values - based on good scores, consistency and feature importance.

The Gradient boost regressor had the best performance with good feature importance.

Feature importance:

trip_distance	0.860406
pickup_day_Wednesday	0.033161
pickup_hour_19	0.031930
DOLocationID	0.025899
PULocationID	0.019849
pickup_hour_18	0.013257
pickup_day_Tuesday	0.004235
pickup_day_Monday	0.003282
congestion_surcharge	0.002393
tolls_amount	0.002107
pickup_hour_17	0.001320
RatecodeID_2.0	0.000971
pickup_day_Thursday	0.000633
Airport_fee	0.000264
RatecodeID_3.0	0.000198
RatecodeID_5.0	0.000089
RatecodeID_4.0	0.000008

Grid Search CV:

Grid search CV was performed to find the best parameters of the chosen model.

The following are the best parameters chosen for the Gradient Boost regressor model.

{'learning_rate': 0.09, 'max_depth': 5, 'max_features': 'auto', 'min_samples_leaf': 3, 'min_samples_split': 2, 'n_estimators': 200}

Performance of model after using the best parameters:

GRADIENT BOOST REGRESSOR	TRAIN	0.896927	2.744755	16.336841	4.041886	0.220560
GRADIENT BOOST REGRESSOR	TEST	0.867474	3.034169	20.293077	4.504784	0.240588

The performance of the were similar to before but the model has got higher scores in the train set than the test. To check the generalization of the model and to choose whether to rely on the model or not Cross validation was performed.

Cross Validation:

A separate 5 fold cross validation was performed by separately encoding train and test sets. The results were consistent.

results: [0.8684507916303713, 0.8656528555689679, 0.8637253052786742, 0.8670538163110549,0.8591113886343358]

Conclusion:

As the model generalized well in cross validation with score as 85 which is better in prediction than all the other models built, it was chosen as the final model.

Implications of the analysis and analytics:

- Currently the Taxis running in the New York do not have the embedded feature of predicting the fare of the trip and the trip duration. This feature is available only in the taxi hailing apps. This feature when embedded in the taxis hailed in the roads will highly be helpful in reducing the dissatisfaction of people at times of congestion.
- The Insights of congested and better routes with the past data can be used to save time and fuel costs. This can lead to increased profitability and improved customer satisfaction.
- Also, the congested routes insights will show the government the right places to improve the traffic regulation and prioritize investments so as to address the issue of congestion going on in the city.
- The Trip demand analysis will help the drivers in getting better rides, thus helping their revenue.
- A running display of fare and time taken to the major locations in the outer part of the taxi will help in getting better rides addressing the competitions of Uber and other private players.

Limitations:

- The prediction models have access only to the past taxi ride data and not to the real time traffic details. Access to the real time traffic details will enhance the prediction of fare and trip duration.
- The models built had only access to data of a week, access to large amounts of data in a highly equipped environment would have created a better model capturing more intricacies in a better manner.
- The models built are specially catered to the needs of the congested hours, the models cannot work on other hours of the day efficiently.

Conclusion:

In conclusion, the integration of machine learning models for trip duration prediction, total amount estimation, and route optimization provides a treasure trove of opportunities for taxi services and transportation-related businesses in New York. The success of these initiatives hinges on effective implementation, continuous model improvement, and a commitment to delivering safe and compliant services. Leveraging technology and data-driven insights is not just a competitive advantage; it's a pathway to reshaping urban transportation for the better, ensuring a smoother, more efficient, and user-centric experience for both drivers and passengers.