OVERVIEW

The organization in charge of granting permits and enforcing regulations over Medallion (Yellow) taxi cabs, green taxi for-hire vehicles (including commuter vans, black cars, and opulent limousines), community-based liveries, and paratransit vehicles is the New York City Taxi and Limousine Commission (TLC), which was founded in 1971. The organization is overseen by a paid Chair/Commissioner who preside over regularly scheduled public commission sessions and oversees a staff of about 600 TLC employees. Everyday there is more than 1,000,000 number of trips. The essential factor in the domain is:

- Ease of Access
- Privacy
- Availability
- Cost

PROBLEM STATEMENT

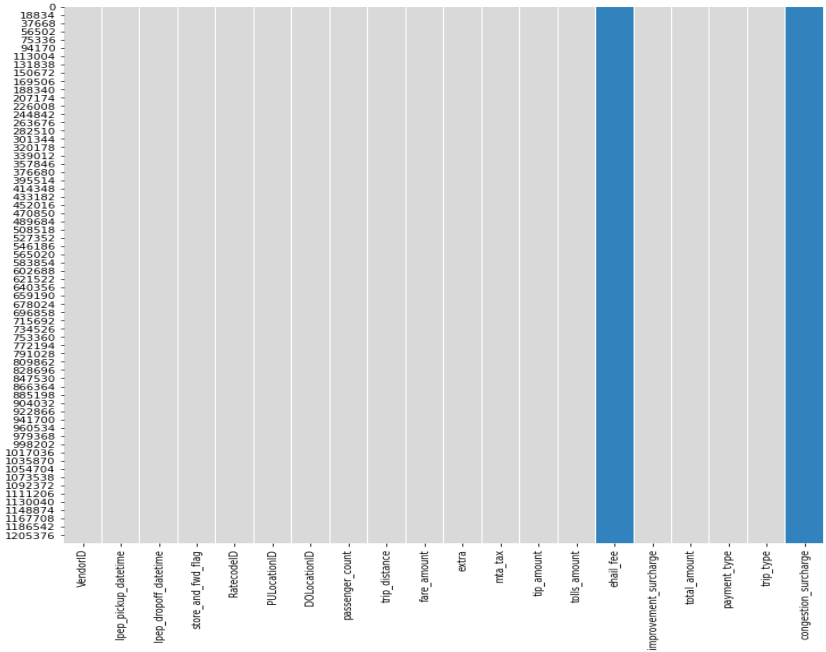
To analyze the given data set by summarizing them, finding insights by analysis and visualizing, finding the various anomalies in the dataset, and then building a predictive model to predict the tip amount for a cab driver based on different features from the dataset.

DATA:

There were more than 1,224,158 records with 20 attributes. The dataset records different trips for any duration with a parameters like trip amount, tax amount, other charges, payment type, trip type, etc.

DATA ISSUES/DATA QUALITY/ANOMALIES:

- It was found that the ehail_fee and congestion_surcharge was having the entire column as NULL so both columns were removed.



- There were no duplicate records in the dataset.

- Data characteristics were found using different techniques like summarizing, describing the data, and visualizing them.
- It was also found that some of the attribute's classes were having the very less values. Like payment type 6 – voided trip was having o records, 5 – Unknowns were 63, Rate Code ID – 6 was having 19 records.
- It was also seen that there was 0 passengers' trip with having other attributes values with it and not having 0 values.
- There are some invalid entries like for 1010408, 1011462 record the trip distance is 101 miles but still the total fare shown is 0. If the cab driver would have taken the charge in cash due to the connection issue or any other issue.
- The trip amount, fare amount, and the trip distance were found to have a negative value.
- The passenger count was found to have a zero passenger and there were meter readings associated with it, which might not be possible.
- The total amount was found to be zero for some records which might be due to people using coupons or other offers.
- Not all attributes follow the normal distribution.
- The extreme values are not actually an outlier this can be an actual value for any trips.

INSIGHTS:

- The total amount, fare amount, trip distance was found to have a negative value which were removed from the dataset assuming that they will not occur. Also, assuming passenger count can't be 0.
- The total amount and fare amount was having a linear relationship and a positive correlation of more than 98% meaning that fare amount was covering the 98 % of the data feature of the total amount.
- Passenger count can be more than 6 as well, as the taxi booked can be like minivan which can accommodate more than passengers so not considering this as outlier.
- Trip distance can't be considered as outlier as there might be chances that the passengers would have travelled this much distance. Hence these two points need to verify from the business stakeholders or business req.
- Assuming Fare amount and Total Amount can't be negative so remove the values below 0. Also, it was seen that even if the trip distance has been used but still the fare amount or Total Amount is 0 so ignoring these values as well. Taking only > 0.
- Considering all these attributes extreme values in the data. As these data might not be outliers and these might be the true values.
- None of the attributes follow gaussian or normal distribution, except tip amount, which is a dependent variable in the case.
- The maximum number of trips was found to be on 10^{th} of the month (Dec), and the maximum number of trips was found to be on 50^{th} Week of the year (Dec Month Data).
- The is relation between total amount and trip distance, fare amount and trip distance, fare amount and total amount.
- The tip amount varies irrespective of the trip distance and the total hours of the journey.
- The tip amount was maximum for the credit card payment type, followed by cash.

- The maximum number of tips given was using credit card payment.
- The maximum number of tips was given under the store and forward flag as No.
- There were instances where the trip distance was 0 but still the other attributes like fare amount, extra, surge charge, etc. were having the values associated with it.

HOURLY DATA INSIGHTS:

The data were grouped based on hourly to get the important hourly insights:

- The maximum number of trips were done for the 19th hour. It was also seen that the count of trips every hour has increased heavily after 12 afternoons till midnight. The analysis remains the same during the drop off times.
- The Average total passenger was again high in the evening time with a value between 60,000 to 118,394. The highest number was found during the 19th hour.
- It was seen that the maximum average trip distance was found to be during the 5th hour of the day with a value of 4.02 miles.
- The average tip amount remains same through out the day irrespective of the hours whereas the maximum average was found to be during the 6th hour of the day.
- The maximum average total amount was found to be during the 5th hour of the day.
- The average travel hours remain almost same throughout the day.
- ***It was seen that the 5th and 6th hour of the day was having better values as compared to the other time; like the average trip distance was maximum during these two, the average fare amount was maximum, the average total amount was maximum, the average total hours travelled was found to be maximum, and the average tip amount was maximum during these two times.***

	Total_trips	Total_Passenger	avg_trip_distance	avg_fare_amount	avg_tip_amount	avg_total_amount	avg_total_hours
count	24.000000	24.000000	24.000000	24.000000	24.000000	24.000000	24.000000
mean	50172.166667	68376.208333	2.811530	11.906343	1.160385	14.336764	0.349644
std	22297.220565	30600.725249	0.436602	0.962873	0.116909	1.070256	0.022682
min	11105.000000	15552.000000	2.349452	10.684695	1.039632	13.128225	0.318450
25%	35627.000000	49376.750000	2.571892	11.314305	1.092678	13.857118	0.332334
50%	47505.000000	64123.000000	2.687740	11.841229	1.128084	13.987320	0.344903
75%	67339.250000	92830.250000	2.867557	12.043120	1.163000	14.468845	0.364256
max	85837.000000	118394.000000	4.066593	14.559785	1.500952	17.615790	0.401613

MODEL BUILDING:

We must provide a way to let the driver know what kind of ride will yield them the better tips. Hence, we must predict the tips for the driver depending on the different attribute's values.

Target/Dependent Variable – Tip Amount

Independent Variable – All other remaining variables, except the time, PU Location ID and DO Location ID as both variables was having more than 250 classes under it.

All the data issues, preprocessing of the data was performed before model building. The Categorical variables were encoded.

Model Used – ***Multiple Regression***.

Train test ratio – 70:30

Below is the leader board for the model built:

	R2	Adj R2	RMSE	R
Model 1	0.982	0.982	0.279	0.99
Model 2	0.4487	0.4486	1.576	0.66
Model 3 A	0.5486	0.5487	1.42	0.74
Model 3 B	0.683	0.683	0.723	0.82

We always use Adj. R2 and RMSE for the model evaluation for the multiple regression. Below are the different models which were built:

***Model 1:** Where all attributes were used for the model building.

Outcomes:

- The R2 was found to be almost 98% which means that 98% of the variation on the target variable i.e., tip amount can be explained by the independent variables.
- The Adj. R2 was found to be same, which meant that that the variables were equally important to the target variable.
- The R value was found to be 0.99 which meant that the independent variables were 99% corelated to the target variable.
- The RMSE value was found to be least for this model, which meant that the error between the predicted and the actual value was minimum.

NOTE: The model accuracy is coming as almost 98%, which is not true in the real case. This might be due to the data leakage as the total amount contains the tip amount and there might be other aspects to it. Hence building the next model by removing the total amount from the independent variable list.

Model 2: By removing the fare amount from the independent variable list. This model was found to be least performing which was below 0.5 meant that it performed worse that the average one. Hence this model can't be considered.

Model 3: The accuracy was found to be best with a R2 value as 0.548 for all different combination of the variables chosen. This is not so good, but we fine tune this model for the better performance and add some more important feature to it so that it wors better.

Model 3B: Model 3 was enhanced for the better performance as the outliers were removed even though those might be the actual values but considering these extreme values to be not occurring regularly. Hence, we can build to model where we can have a model for the extreme values as well as it might be due to people taking the intra city taxi service which results in extreme values.

- The R2 was found to be almost 68.3% which means that 68.3% of the variation on the target variable i.e., tip amount can be explained by the independent variables for this case.
- The Adj. R2 was found to be same, which meant that that the variables were equally important to the target variable.
- The R value was found to be 0.82 which meant that the independent variables were 82% corelated to the target variable.
- The RMSE value was found to be 0.72 for this model, which meant that the error between the predicted and the actual value was a small value.

RECOMMENDATION: There is one important attribute which needs to be added in the data set which is success of the trip which will help to identify the records well and have better idea about those trips with zero passengers or the trip distance. Some other attributes i.e., cost can be added to perform the profit calculation for the trip for the company.

Some important attributes which need to be added for the better model building:

- **Trip satisfaction** that can include lot of other features.
- Cab/Taxi facility.

Both these attributes result in the tip amount for the driver hence this needs to be added in the dataset.

LIBRARIES REUQIRED:

!pip install fastparquet

!pip install Pyforest

!pip install scikit-learn

!pip install statistics