**NYC Taxi Driver Income Optimization**

The NYC 2017 June yellow taxi data contains 9656993 ride records with 17 features. The data should be pre-processed and cleaned to analyze the results.

**Data Pre-processing**:

The records with missing values, negative values and errors in the data should be cleaned. From the data, we can observe that there are lot of negative values in the dataset for features fare\_amount, extra, tip\_amount, improvement\_surcharge and total\_amount.

We can find the data errors by digging into each feature. Those records should be addressed and removed if needed.

* If the ride is recorded, the minimum number of passengers should be at least one. So, the records with passenger count zero should be removed.
* If the data is sorted by total amount feature, we can observe that there is huge fare recorded for small trips and trips that have distance zero. This can be considered as data error.
* If any ride has a total amount recorded with trip distance of zero, same pickup and drop-off location, then it should be removed
* The difference between Drop-off time and pick-up time is considered as trip duration. If the trip duration is zero with different pick-up and drop-off then it can be considered as an error
* As we want to analyze only rides from 2017, remove rides other than 2017 if exists

To analyze the data more efficiently, we need to add some features which would make our work easy.

* Calculate Trip duration which is the time difference between Drop-off time and Pick-up time
* Get Year, Month, Day of Week and Hour for calculating more insights into the data (For example, these features would be helpful to find the peak hours of each day in a week and which day has more rides)
* Look up into taxi zone sheet for Pick-up and Drop-off Zone based on Pick-up and Drop-off Location ID.

Search for null values or Nan and replace it with zero

I found that the feature VendorID is not useful for this problem. As we are trying to maximize the income of taxi driver, I think VendorID does not help in finding any insights into the data that are useful for the problem.

I would like to use Linear Optimization technique to maximize the income of NYC taxi driver.

**Linear Programming**: Linear Programming, also called linear optimization, involves maximizing or minimizing a linear objective function, subject to a set of linear inequality or equality constraints. It has a wide range of applications in the field of operations management, industrial design, planning etc.

In order to solve the problem using linear programming, we need to find the linear objective to maximize or minimize, set of decision variables and set of linear constraints on those variables.

* The linear objective for this problem is to maximize the income of taxi driver which is maximizing the income (total amount)
* The decision variables are number of rides in the dataset
* The constraints for the decision variables are: Driver can drive only up to 10 hours per week

My idea is to find the best combination of records (to maximize the income) which would add up to 10 hours of trip duration. This can be achieved by Linear Optimization.

The summary of my results is:

1. Friday has the highest number of rides followed by Sunday, Thursday, Monday, Saturday, Wednesday and Tuesday

A screenshot of a cell phone

Description automatically generated

1. All weekdays have peak hour at 4pm in the evening. Thursday and Friday also has peak hours at 12 am. On Saturday, lot of rides are recorded at 12am, 3am and 12pm - 4pm On Sunday, the peak hours are at 5pm

A close up of a map

Description automatically generated

1. Busiest routes are Manhattan -> Manhattan, Unknown -> Unknown, Queens -> Queens
2. Queens has the highest tip rate which is followed by Queens

A close up of text on a white background

Description automatically generated