**New York City Taxi Trip Duration**

**Team Member (Team 5):**

1) Shanun Randev

2) Anand Raj

3) Mowzli Sre

4) Bala krishna Reddy

**Introduction:**

The research topic that our team has chosen is predicting the total ride duration of taxi trips in New York city.In cities like New york where the traffic is high and the distance between the destinations is short, everyone wants to reach their respective destinations as soon as possible.

The dependent variable in this research is “**trip\_duration**” which is the duration of the trip in seconds.There are 10 independent variables or features which we will use in our hypothesis to generate our predictions.

**Dataset Info:**

The independent variables are as follows :

1. Id - unique identifier for each trip
2. Vendor\_id - a code indicating the provider associated with the trip record
3. Pickup\_datetime - date and time when the meter was engaged
4. Dropoff\_datetime - date and time when the meter was disengaged
5. Passenger\_count - number of passengers in the vehicle
6. Pickup\_longitude - Longitude where the meter was engaged
7. Pickup\_latitude - Latitude where the meter was engaged
8. Dropoff\_latitude - latitude where the meter was disengaged
9. Dropoff\_longitude - longitude where the meter was disengaged
10. Store\_and\_fwd\_flag- This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y = store and forward; N = not a store and forward trip.

**Research Questions - S - Specific, M - Measurable, A - Achievable, R - Relevant, T - TimeBound**

**Smart Questions:**

1) can we predict the total distance using the pickup and dropoff latitude and longitude, then examine the relationship between the two variables: trip duration and distance

2) On what days and at what time during the day the trip duration is maximum

3) what are the top 5 locations with high demand of taxi in NYC?

4) can we specify the passenger count that corresponds to trip duration

5) can we check outliers for trip duration and figure out the impact with and without them?

**Gitub:**

[**Data mining Project Github Team -5**](https://github.com/shanunDS/Group5-project.git)