# NYC Citi bike Data Analysis

## About the dataset

- Citi Bike is New York City's bike sharing system. Intended to provide New Yorkers and visitors with an additional transportation option for getting around the city, bike sharing is fun, efficient and convenient. Citi Bikes are available 24/7, 365 days a year. There are 507 stations across New York and Jersey City
- Station locations are based on population and transit needs
- We are using dataset for the year 2014 till 2015
- Below are the parameters of the dataset:
  - Trip duration
  - Start and Stop
    - Time
    - Date
    - Station Name
    - Station Id
    - Latitude and Longitude
  - Bike ID
  - User
    - Type (Customer/ Subscriber)
    - Gender
    - Year of Birth

## External data sets: Year 2014 till 2015

Weather

- SNWD (Snow Depth in mm)
- SNOW (Snow in mm)
- PRCP (Precipitation in 10’s of mm)
- Temp (Avg. Temperature in 10’s of degree Celsius)

Holiday

- Weekday/ Weekend – TRUE/ FALSE
- National Holiday – TRUE/ FALSE

INFO 7250

Neha Firodiya (firodiya.n)
Smit Shah (shah.smi)

## Tools and Technologies

- Hadoop 2.7.2
- Hive 2.0.0
- AWS – EMR (4.6.0)
- AWS – S3
- RStudio
- Eclipse IDE
- Tableau

## Analysis of data:

Cleaning of data was one of the major tasks. We cleaned the data using MapReduce job. We then analyzed the data based on below criteria:

1. Number of Bike rides: HIVE queries
    a. Gender
    b. User Type – by Year and Months
    c. Age group – by Year and Months
    d. Holidays – Rides by national holidays
    e. Weekdays/ Weekend
    f. Daily
    g. Time
        i. Pick up
        ii. Drop Off

2. Average Trip Duration: HIVE queries
    a. Age group
    b. Gender
    c. User Type

3. Location: Source to Destination (Total count/ frequency: Custom MapReduce jobs
    a. Top pick up locations
    b. Top drop off locations
    c. Top Routes

    We ran the jobs on EMR and visualized the data on Tableau

4. Page Rank
    Page Rank implementation of destination location using custom MapReduce code based on bikes coming from a specific source (which is counted as a vote)

5. R
    a. Multivariate Regression
        i. Trips based on weather conditions
    b. Prediction
        i. Arima Time series model
        ii. Neural Networks