# Title: San Francisco Bay Area Bike Share Analysis

First Name: Shashank Bharatlal, Last Name: Khede (A20386301)

**Type of Your Projects:** Multiple linear regression

1. **Introduction**

This application has interesting collection of data containing information about people who rented bikes throughout the bay area in San Francisco from August 2013 to August 2015. Basically, the Bay Area Bike Share is the region for sharing the bike across 70 stations and 700 bikes. The bikes can be rented from any station and can be returned to any other station across the region making it easier for people to commute. The system is available for 24\*7 use and all year round. The collected data has all the details about when the bike was rented and when was it returned. It also has the name of the start and end station, zip code and other details. We can use this data for making predictions that what will be the peak time for the bike rental according to specific area or we can analyze the duration of the trips. Some more analysis can be done on the average duration of the trips, does the duration of the bike ride vary according to different cities, time, or day of the week, do people prefer taking bike rides on weekends more than on the weekdays.

1. **Data Sets**

We collected the data from [www.kaggle.com](http://www.kaggle.com). The data is based on users in San Francisco. We will be studying the effect of different variables on the Bicycle share program for example how weather influences usage of Bicycle Share program.

Our data contains Four table namely; Station, Status, Trip, and Weather.

The table Station contains 70 rows and 7 columns, columns names are as follows:

* **id, name, lat, long, dock\_count, city, installation\_date**

Table Status contains 71984434 rows and 4 columns:

* **station\_id, bikes\_available, docks\_available, time**

**Trip table contains** 669959 rows and 11 columns.

* id, duration, start\_date, start\_station\_name, start\_station\_id, end\_date, end\_station\_name, end\_station\_id, bike\_id, subscription\_type, zip\_code

**Weather table has** 3665 rows and 24 columns:

* date, max\_temperature\_f, mean\_temperature\_f, min\_temperature\_f, max\_dew\_point\_f, mean\_dew\_point\_f, min\_dew\_point\_f, max\_humidity, mean\_humidity, min\_humidity, max\_sea\_level\_pressure\_inches, mean\_sea\_level\_pressure\_inches, min\_sea\_level\_pressure\_inches, max\_visibility\_miles, mean\_visibility\_miles, min\_visibility\_miles, max\_wind\_Speed\_mph, mean\_wind\_speed\_mph, max\_gust\_speed\_mph, precipitation\_inches, cloud\_cover, events, wind\_dir\_degrees, zip\_code

1. **Research Problems**

Below are the research problems which we are trying to solve using the multiple regression model:

* What is the average duration of the bike trips and does it vary according to the cities.
* What is the busiest time of the week for the bike trips, the weekends or the weekdays.
* Can we predict that what role the weather plays in bike share trip?
* Do the bike trip patterns vary with the time and day of the week?

1. **Potential Solutions**

For the above research problems, we will be performing Multiple Linear Regression Analysis and will decide the dependent and independent variables according to the problem set. Based on that we predict the values of dependent variable based on the value of independent variables. The analysis will help us explain the impact of changes in an independent variable on the dependent variable.

* We will start by plotting a scatter plot between the different variables (Y and x-variables) and observe if there is a linear relationship or not
* If the relationship is not linear then per that we perform transformation on Y.
* After this we will build our Multiple linear regression model.
* Then will perform Parameter estimates
* We will then see how well our fitted model can be used to explain the observation by performing Goodness of fit test.
* After the goodness of fit test, we will now perform Residual analysis to see if there is some important variation in Y that is not explained by the regression model.
  + - * 1. Check linearity assumptions & constant variance for the errors by plotting Residual vs predicted values.
        2. Check linearity assumptions for Y and x-variables.
        3. Draw normal probability plot of residuals; to check normality assumptions for the error terms.
* We will then be using this model for prediction.

The analysis of residuals is useful to detect possible problems and anomalies in our regression model.

1. **Evaluations**

We will divide our data in training data and test data. We will learn about the data using the training data and test using test data to assess accuracy. Apply the selected model to unseen data. We will be applying Hold-out evaluation on our dataset as it is large enough and we have enough observations in the sample for splitting.

* **Estimate model using training set**
* **Compute predictions on testing set.**

**We try to avoid two problems in Multiple linear regression model i.e Multicollinearity and influential points. They impact the performance of our MLR model.**

* Check for x-variables that are strongly correlated. Examine the Pearson correlation matrix and the scatter plots for each pair of x-variables. Correlation values larger than 0.9 or so indicate a serious collinearity problem.
* Influential points are the outliers that affect the fitted model. Influential points are observations (typically outliers) that have a strong influence on the fitted model. If removed, the parameter estimates change.

**We will also select the best set of x-variables to be included in the model.**

**The following criteria is used to define the best model.**

* Optimize coefficient of determination R2 -adj
* Optimize Mallows’ Cp Statistics
* Minimize PRESS statistic
* Backward elimination, best subset, stepwise.

1. **Expected Outcomes**

With our best fitted model, we will be able to predict below things:

1. The average duration of the bike trips and can be able to provide answer if it varies per the cities.
2. We can predict the busiest time for the week if weekends are busier than the weekdays.
3. We can predict the role of weather in bike share trips that if its rainy day, the people renting the bikes are lesser or not.
4. We can even predict how the bike trip patterns vary with the time and day of the week that what is the peak time of the day like mornings when people go to office and evenings when people come back from office.