The Bike Sharing Data Set from [Capital Bikeshare](https://www.capitalbikeshare.com/system-data) contains information regarding customers’ bike use in New York since the 4th quarter of 2010. The data is published every quarter and contains the following fields:

* Duration – Duration of trip
* Start Date – Includes start date and time
* End Date – Includes end date and time
* Start Station – Includes starting station name and number
* End Station – Includes ending station name and number
* Bike Number – Includes ID number of bike used for the trip
* Member Type – Indicates whether user was a "registered" member (Annual Member, 30-Day Member or Day Key Member) or a "casual" rider (Single Trip, 24-Hour Pass, 3-Day Pass or 5-Day Pass)

Capital Bikeshare states: “This data has been processed to remove trips that are taken by staff as they service and inspect the system, trips that are taken to/from any of our “test” stations at our warehouses and any trips lasting less than 60 seconds (potentially false starts or users trying to re-dock a bike to ensure it's secure).” Further, the 3-Day Membership passes were replaced by the 5-Day Membership passes in Fall 2011.

The original data set as described above lends itself to a classification problem to determine the “Member Type” of any particular user, but I will be aiming to solve a regression problem to determine how many bikes are at each station in the future (PROBLEM: How do I get starting numbers and come up with a running total for the current hour? The data does not seem to capture current bike rack numbers). This is known as a rebalancing issue that bike share platforms have due to the nature of commuting patterns. Bikes tend to pile up in one area in the morning and in another in the afternoon, and so customers cannot begin trips at an empty station or end trips at a full station. To address this problem, bikeshare operators drive trucks to reallocate bikes from full stations to empty ones. This solution is not optimized because it is reactive in nature – operators can only move bikes between stations that they know to be full or empty. By predicting the number of bikes at any given station in the future, the reallocation of bikes becomes a proactive solution and drivers can be dispatched in a much more efficient manner.

In addition to the original data set, I plan to use public data from the city in which this data set is recorded. This external data could include weather, financial, taxi, and subway data as well as other available data. I plan to either acquire the data straight from public websites or use APIs to gather the data I need from public databases. I plan to aggregate all data and determine how much data I need to solve the problem in a practical manner since bikeshare data is published quarterly there might not be a need to use all available data. Next, I will need to explore the data to determine any cleaning that needs to be done with Python’s Pandas, Matplotlib, and Seaborn libraries. I will then clean the data and choose a predictive algorithm to implement using Python’s Scikit-Learn library.

In terms of final deliverables, I plan to provide the following:

* All code used to solve to above the problem.
* A paper outlining the data story.
* Any visualizations that might aid in telling the data story.