The data which was used for this assignment was Bike Share data from the Bay Area Bike Share. This data is at the trip level and contains information about trip duration, starting and ending location, and rider information. The full dataset contains data from 600,000 trips, with 42 columns of information total. The data can be found on Kaggle in the SF Bay Area Bike Share competition. The original source data can be found from the following link to the Bay Area Bike Share website, which allows this data to be accessed openly (<http://www.bayareabikeshare.com/open-data>). This data interested us because we have been working with bikeshare data in examples frequently over the course of the semester. It interested us to find some similarities and differences in the DC Bike Share data versus the SF Bike Share data, and also to delve into some of the analysis that was discussed in class but not actually performed. A concern that we have with the data is that the full set is too large to use directly in datanotebook.org. Therefore, it is necessary to use PostgreSQL on a local machine for it to work correctly. Additionally, the columns in the trips.csv data are somewhat limited, so we will have to use the additional datasets provided. The data is broken out into three different datasets for the Kaggle competition, so some adjustments may be made to create the star schema and fact table in the correct way. Finally, since the data is so similar to the DC Capital Bike Share set used for analysis in class, our group will face the challenge of creating original and insightful analysis from this data which is different from the questions we have looked at throughout the semester. The weather data provided for the Kaggle competition was pretty extensive, weather data was provided, so we will be joining the information from that set later in the analysis.

From the initial descriptive statistics, it is clear that there is a unique identifier for each of the three main datasets, therefore, it may be possible to keep the existing structure in constructing the star schema. For the station dataset, the ID is representative of the unique station. Location information, name, number of docks, and date of installation are also provided in this set. There are 70 different stations in the data provided. The most recent installation date was 4/9/2014.

The Trip dataset is the main dataset for analysis. This data contains the most granular information at the trip level. There is an ID which identifies each unique trip, and also provides information about the dates and duration of the trip, start station and ending station, bike information, subscription information, and zip code.

The weather dataset contains detailed weather data by date and zip code for the San Francisco area. The columns provided are temperature, humidity, sea level pressure, visibility, wind speed, inches of precipitation, cloud cover, events, and degree of wind direction. There are null values present for some of the date and zip code combinations.

An additional dataset provided contains weather information for each service area. This information will be joined to the existing dataset in