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 initial dataset contained over 600,000 records for this particular dataset. In order to tailor data for our analysis purposes and so that it smoothly runs in datanotebook.org, we limited the number of observations to 250,000. Our methodology to limit this will be discussed in a few cells. Most of the data comes from the following zip codes: 94107, 94105, 94133, 94103, and 94111, but there are thousands of distinct zip codes in the full dataset, which seems wrong when talking about county level data. After further investigation, many issues were discovered related to the zip\_code field including a mixture of 5 and 9 digit zip codes, whitespace, incomplete zip codes, and invalid zip codes for the San Francisco area. This is addressed when limiting the dataset to 250,000.

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 issue with this dataset is that there are only five distinct zip codes in the entire set. The weather zip codes were the following: 95113, 94041, 94107, 94301, and 94063. These codes are all seemingly valid and are not more than five digits. 94107 is also the most frequent zip code in the trips.csv discussed previously. Therefore, the first way the dataset was limited to 250,000 records was to filter to trips which contained the zip codes present in the weather data. Since there were some issues with data formatting discovered in the trips file, a new column was created called zip\_5, which removed the whitespace (since dashes present in this data converted the column to character) and then took a substring of the first five characters. Again, the first part of the data was the zip\_5 codes which matched the weather data zips. Next, the remaining data was filtered to valid, complete codes in the San Francisco Bay area and then randomly sampled to create a full dataset of 250,000 records. This data preparation was completed in R. Now we have more complete and correct records which will provide a valuable analysis. It will also be able to run through datanotebook.org without any issues. When analyzing the data, it will be necessary to keep in mind that the distributions of the data will be skewed to the zip codes that we have weather information for.

Weather: 95113, 94041, 94107, 94301, 94063

Top 5: 94107: 25,519, 94105: 18,193, 94133: 11,916, 94103: 9,166, 94111: 8,200

Delete crime data stuff

Initially, we observed a graphic created from the Kaggle competition which showed the amount of trips from one station to another. It seems that the busiest stations were Embarcadero at Sansome, Harry Bridges Plaza (Ferry Building), 2nd at Townsend, Towsend at 7th, and San Francisco Caltrain 2 (330 Townsend). This corresponds to what we saw previously in the results of the csvstat output for trips.csv.

After initial investigation, it seems that our data is in good shape to be uploaded into a database. The following steps loads SQL and calls PostgreSQL. Next, create the database. In order to ensure code is reproducible, drop existing database called “finalproject” before creating a new one. Create a new “finalproject” database and connect to it using the dbuser username.

The following steps creates a shell table for each of the datasets provided. The shell tables are called “station,” “trip,” and “weather.” Each column is given a format based on the csvstat output generated from the dataset based on the station.csv, trips250.csv, and weather.csv datasets respectively.