Homework 2: Munging and Joining Data

Often in the "real world", you will have to join-together data from multiple sources. Here we are going to build on the homework exercise from last week, adding-in another dataset featuring weather data.

Another "real world" research challenge is that you do not always know exactly what function/method to use in order to arrange the data the way you want. This assignment will mention certain operations that we want to do on the data, with the expectation that you will figure out how to do these operations by reading the documentation or searching the web. Doing this efficiently takes practice, and starts with thinking hard about *what* terms to search for.

Problem

Download the weather data (download link) and put it in the `data` directory of your project from last week (this weather is from the National Climatic Data Center, and you can find the documentation for the data here).

Create a new IPython notebook for this week's analysis.

First we will load and visualize the new weather data:

1. Load the weather data into a Pandas dataframe. You might notice that the date column is in the form `YYYYMMDD`: while you could extract the year, month, and day yourself, Pandas has the ability to automatically *parse dates* when reading a CSV. Re-read the CSV file, specifying that pandas should parse these dates.
2. Now that we have this data loaded and the dates parsed, we would like to *set the index* of the dataframe to be the date rather than the line number in the file. Set the index of your dataframe to the date.
3. Create plots showing the temperature and precipitation as a function of date (because the date is the index, this should be a very easy `plot()` command!)

Next we will load and resample the bicycle data we explored last week

1. Load the Fremont Bridge data into a dataframe – use what you learned above to automatically parse the date column, and make this date the index for the data.
2. This data is reported hourly, but we would prefer daily totals. Find the pandas operation which lets you *resample* the time series resulting in a *daily sum*.
3. Plot the total ride count as a function of the date (again, with the date as the index this should be very easy!)

Finally, we want to join the two datasets together so that we can compare them.

1. Look up how to use the pandas `join()` command, and use it to join the resampled bicycle counts and weather data into a single dataframe, indexed by the date.

2. Create scatter plots of the number of rides as a function of temperature, precipitation, wind speed, or other quantities you think might be interesting (you'll probably have to look at the data documentation link above to learn what the column names mean). Can you find any interesting patterns? Are there trends that reflect how weather influences the number of riders
3. Thinking more broadly, what other factors might you expect to influence the number of bicyclists crossing the Fremont bridge? Write down a few ideas that you think would be worth exploring in the future.

Hints:

The final result should require only a few lines of code for each of the above steps. The challenge here is to use the resources available to you (Google, function documentation, etc.) to figure out how to accomplish each of these tasks using Pandas! As a hint, usually doing a web search for "Pandas" plus a couple keywords about what you would like to do (look for the *itallic* terms above) is enough to get you to a documentation page or StackOverflow question that will help you.