Bi 410/510: Introduction to Programming for Biologists

Spring 2018

Project 8: `pandas`

This project is an introduction to data analysis using `pandas`, a popular data science library for Python. To work on the project you need to have the library installed on your computer. If you installed Python using Anaconda/miniconda simply open a terminal window and type these commands:

```
$ conda install numpy
$ conda install pandas
```

(if you did not install Anaconda see me if you need help installing the libraries)

Your job for this project is to write a Python program that will create a DataFrame object (a data type defined in `pandas`) and then use methods of that object to answer questions about the data. An optional part of the project is to use `pyplot` to create figures that visualize the data.

Download the data file and save it in the folder you are using for this project:

```
$ curl pages.uoregon.edu/conery/Bi410/oregon_football.csv > oregon_football.csv
```

DataFrame

The data is a CSV file were each line describes a football game played by the University of Oregon. Here is an example input line:

```
1917-01-01,1916,Pennsylvania,14,0
```

The first field is the date the game was played. The second is the season the game belonged to; in this example, the game was played on January 1, 1917, but that was the end of the 1916 football season so the second field is `"1916"`.

The third field is the name of the opposing team, and the last two fields are the number of points scored by the Ducks and the number of points scored by the opponent. The data on the line above shows the Ducks won this game, 14-0.

The first step is to load the data into a `pandas` DataFrame object. The easiest way to do this is to call the `read_csv` function to create the frame. If you print the first 5 lines of the frame this is what you should see:

```
         date  season          opponent  points_scored  points_allowed
0  1916-10-07    1916        Willamette             97               0
1  1916-10-14    1916    Multnomah A.C.             28               0
2  1916-10-21    1916        California             39              14
3  1916-11-04    1916        Washington              0               0
4  1916-11-11    1916  Washington State             12               3
```

Data Analysis

Write a Python program named `score_trends.py` that will use the DataFrame to answer various questions about the data. The only command line argument is the CSV file name, so this is how you'll run the program:

```
$ python score_trends.py oregon_football.csv
```

Each time the program runs it should print the answers to the questions listed below and (optionally) generate the figures that visualize the data.

When you are done simply upload your program to Canvas. We will run your program, check the output to make sure it's correct, and look at your visualizations to see if they are accurate.

IMPORTANT: Make sure all group members' names are included in the file!

Note: We will look for the correct results for each part of the project, but we will not check to see if you print results in the exact same format.

Questions

1. Figure out how many games were played and the total win-loss record for the Ducks. Since ties were allowed in college football until the 1980s you will find several games where the points scored by each team are the same.

   The output for this section should look something like this:

   ```
   Results from 1055 games: 570-451-34
   ```

   (this is a standard format in sports pages; it means the team won 570 games, lost 451, and there were 34 ties).

2. What is the most points scored by the Ducks? By an opponent? When you display this information, print the entire line that corresponds to the game(s) with the highest points. Here's the expected output:

   ```
   Most points scored:
   0  1916-10-07    1916  Willamette             97               0

   Most points allowed
   222  1941-12-06    1941    Texas               7              71
   ```

3. What are the most points scored by each team in a tie game? The fewest points? Here is part of the output you should see:

   ```
   Most points scored in a tie game
   493  1970-11-14    1970     Army             22              22

   Fewest points scored in a tie game
   3    1916-11-04    1916        Washington             0               0
   32   1920-11-20    1920      Oregon State             0               0
   40   1921-11-19    1921      Oregon State             0               0
   54   1923-10-27    1923             Idaho             0               0
   ```

4. Make a new data frame that has the average (mean) number of points scored by the Ducks and the average number of points allowed in each season. Print the new frame. It should start like this:

```
        points_scored  points_allowed
season
1916        30.500000        2.125000
1917        10.428571       10.571429
1918        13.500000        5.833333
...
```

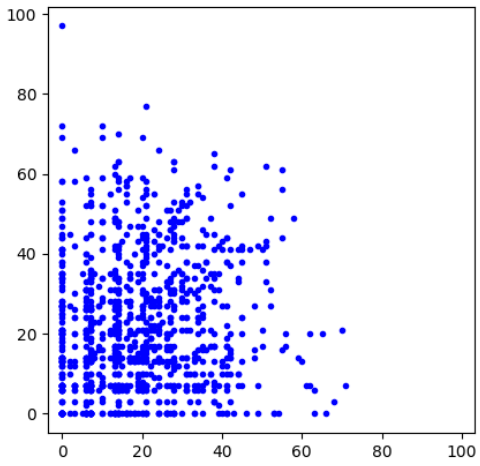
Visualization (Optional)

If you want to create visualizations of the data we suggest using PyPlot. Install the library with this shell command:

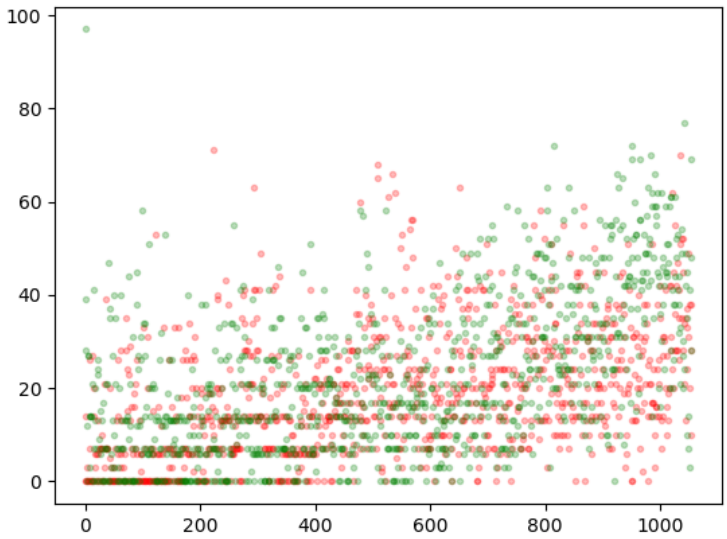
```
$ conda install matplotlib
```

You can pass columns from a data frame directly to the plotting functions when you make dot plots and line plots.

1. Create a dot plot (*aka* "scatter plot") comparing the number of points allowed (y-axis) *vs.* number of points scored (x-axis) for every game. Your plot should look something like this:



2. Create another dot plot, this time with the game number along the x-axis, *i.e.* games played earlier are on the left, and most recent games are on the right. Include two sets of points: plot the number of points scored by the Ducks in green, and the number scored by the opponent in red.



3. Plot the average number of points scored by the Ducks for each season as a green line, and the average number of points scored by the opponent as a red line: