Homework 7: `pandas`

Due March 14, 11:59 pm

Worth 10 points

Read this first. A few things to bring to your attention:

1. **Important:** If you have not already done so, please request a Flux Hadoop account. Instructions for doing this can be found on Canvas.
2. Start early! If you run into trouble installing things or importing packages, it's best to find those problems well in advance, not the night before your assignment is due when we cannot help you!
3. **Make sure you back up your work!** I recommend, at a minimum, doing your work in a Dropbox folder or, better yet, using `git`, which is well worth your time and effort to learn.

Instructions on writing and submitting your homework.

Failure to follow these instructions will result in lost points. Your homework should be written in a jupyter notebook file. I have made a template available on Canvas, and on the course website at `http://www-personal.umich.edu/~klevin/teaching/Winter2018/STATS701/hw_template.ipynb`. You will submit, via Canvas, a `.zip` file called `yourUniqueName_hwX.zip`, where `X` is the homework number. So, if I were to hand in a file for homework 1, it would be called `klevin_hw1.zip`. Contact the instructor or your GSI if you have trouble creating such a file.

When I extract your compressed file, the result should be a directory, also called `yourUniqueName_hwX`. In that directory, at a minimum, should be a jupyter notebook file, called `yourUniqueName.hwX.ipynb`, where again `X` is the number of the current homework. You should feel free to define supplementary functions in other Python scripts, which you should include in your compressed directory. So, for example, if the code in your notebook file imports a function from a Python file called `supplementary.py`, then the file `supplementary.py` should be included in your submission. In short, I should be able to extract your archived file and run your notebook file on my own machine. Please include all of your code for all problems in the homework in a single Python notebook unless instructed otherwise, and please include in your notebook file a list of any and all people with whom you discussed this homework assignment. Please also include an estimate of how many hours you spent on each of the three sections of this homework assignment.

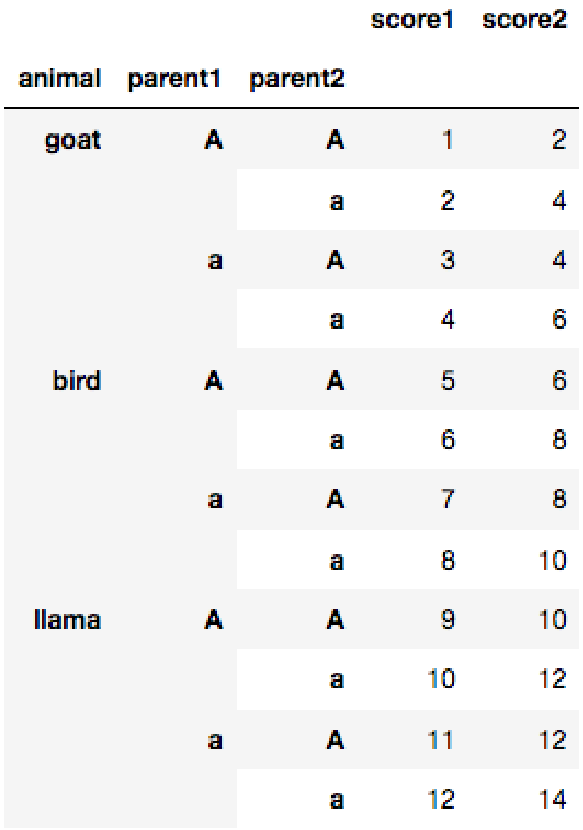
These instructions can also be found on the course webpage at `http://www-personal.umich.edu/~klevin/teaching/Winter2018/STATS701/hw_instructions.html`. Please direct any questions to either the instructor or your GSI.

1 Warmup: constructing `pandas` objects (2 points)

In this problem, you will create two simple `pandas` objects.

1. Create a `pandas` Series object with indices given by the first 10 letters of the English alphabet and values given by the first 10 primes.

2. Below is a table that might arise in a genetics experiment. Reconstruct this as a `pandas` DataFrame.



			score1	score2
animal	parent1	parent2		
goat	A	A	1	2
		a	2	4
	a	A	3	4
		a	4	6
bird	A	A	5	6
		a	6	8
	a	A	7	8
		a	8	10
llama	A	A	9	10
		a	10	12
	a	A	11	12
		a	12	14

2 Working with `pandas` DataFrames (4 points)

In this problem, you'll get practice working with `pandas` DataFrames, reading them into and out of memory, changing their contents and performing aggregation operations. For this problem, you'll need to download the celebrated iris data set, available as a .csv file from my website: `www-personal.umich.edu/~klevin/teaching/Winter2018/STATS701/iris.csv` **Note:** for the sake of consistency, please use this version of the CSV, and not one from elsewhere.

1. Download the iris data set from the link above. Please include this file in your submission. Read `iris.csv` into Python as a `pandas` DataFrame. Note that the CSV file includes column headers. How many data points are there in this data set? What are the data types of the columns? What are the column names? The column names correspond to flower species names, as well as four basic measurements one can make of a flower: the width and length of its petals and the width and length

of its sepal (the part of the pant that supports and protects the flower itself). How many species of flower are included in the data?

2. The data that I uploaded to my website, which you have downloaded, is based on the data initially uploaded to the UC Irvine machine learning repository. It is now known that this data contains errors in two of its rows (see the documentation at `https://archive.ics.uci.edu/ml/datasets/Iris`). Using 1-indexing, these errors are in the 35th and 38th rows. The 35th row should read 4.9,3.1,1.5,0.2,”Iris-setosa”, where the fourth feature is incorrect as it appears in the file, and the 38th row should read 4.9,3.6,1.4,0.1,”Iris-setosa”, where the second and third features are incorrect as they appear in the file. Correct these entries of your DataFrame.

3. The iris dataset is commonly used in machine learning as a proving ground for clustering and classification algorithms. Some researchers have found it useful to use two additional features, called *Petal ratio* and *Sepal ratio*, defined as the ratio of the petal length to petal width and the ratio of the sepal length to sepal width, respectively. Add two columns to you DataFrame corresponding to these two new features. Name these columns `Petal.Ratio` and `Sepal.Ratio`, respectively.

4. Save your corrected and extended iris DataFrame to a csv file called `iris_corrected.csv`. Please include this file in your submission.

5. Use a `pandas` aggregate operation to determine the mean, median, minimum, maximum and standard deviation of the petal and sepal ratio for each of the three species in the data set. **Note:** you should be able to get all of these numbers in a single table (indeed, in a single line of code) using a well-chosen group-by or aggregate operation.

3 Plotting `pandas` DataFrames (4 points)

Note: This problem makes use of the iris data set and depends upon your having completed the previous problem, so please do that first.

1. Use the built-in `pandas` plotting tools to make a box-and-whisker plot showing the distribution of petal ratio and sepal ratio for each of the three species. Your plot should have two subplots, one for petal ratio and one for sepal ratio. You may choose the details of your plots (i.e., how to handle outliers, displaying mean vs median, etc) however you think is best. Please include labels on your x- and y-axes and give an appropriate title to your plot.

2. Use the built-in `pandas` plotting tools to make a scatter matrix plot for the four original features (petal width, petal length, sepal width and sepal length). Each point in the scatter plot should be colored according to its species. **Hint:** see the documentation at `https://pandas.pydata.org/pandas-docs/stable/visualization.html#scatter-matrix-plot` to get started.