Homework 7: `pandas`

Due March 22, 11:59 pm

Worth 10 points

Instructions on writing and submitting your homework can be found on the course webpage at `http://pages.stat.wisc.edu/~kdlevin/teaching/Spring2024/STAT606/hw_instructions.html`. *Failure to follow these instructions will result in lost points.* Please direct any questions the instructor.

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 in lower-case and values given by the first 10 primes. Store it in a variable called `alphabet`.
2. Below is a table that might arise in a genetics experiment. Reconstruct this as a `pandas` DataFrame. Store it in a variable called `simple_table`. **Note:** be sure that your table has a hierarchical row index, just like this table does.

			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 (3 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: `http://pages.stat.wisc.edu/~kdlevin/teaching/Spring2024/STAT606/iris.csv` **Note:** for the sake of consistency, please use this version of the CSV, and not one from elsewhere (e.g., downloaded from the UC Irvine repository).

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 plant 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,"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,"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 your 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. Store the results of your aggregation operation in a variable called `species_agg`. The rows of your table should be indexed by the function names `mean`, `median`, `min`, `max` and `std`, and the columns should be the same as the species names from the `iris` data frame. **Note:** you should be able to get all of these numbers in a single table using a single line of code.

6. Create a scatterplot of the iris specimens, with sepal ratio on the x-axis and petal ratio on the y-axis, and with observations colored according to species. Save your plot in a file called `iris_scatter.pdf`. Please include this file in your submission.

3 Plotting Dataframes: Major League Baseball (5 points)

In this problem, you'll get more practice working with `pandas` data frames and performing basic plotting. We'll work with a data set consisting of all the baseball games from the 2023 Major League Baseball (MLB) regular season, compiled by `retrosheet.org`. Don't worry– you don't need to know anything about baseball to complete this assignment! You can download the relevant CSV file (in a zipped archive) either from the course web page at `http://pages.stat.wisc.edu/~kdlevin/teaching/Spring2024/STAT606/mlb2023.zip` or directly from the original source at `https://www.retrosheet.org/gamelogs/gl2023.zip`. **Note:** even though the zipped file is named as a `.txt` file, it is in fact a CSV file, which `pandas` will still be able to read.

Requisite legal boilerplate: The information used here was obtained free of charge from and is copyrighted by Retrosheet. Interested parties may contact Retrosheet at "www.retrosheet.org".

1. Read the data into a table called `mlb_df`. Each row of the table represents the outcome of a single game from the 2023 MLB season. Take note that the file does not have columns names; see the `header` keyword to the `pandas.read_csv` function. The columns are explained in a .txt file which you can download from `https://www.retrosheet.org/gamelogs/glfields.txt`, but we will only make use of a few of them in this problem. The 10-th and 11-th columns (using 1-indexing) are the scores of the visiting and home teams, respectively. Rename these columns `v_score` and `h_score`, respectively. MLB comprises two leagues, the American League and the National League, encoded as `AL` and `NL` in the table. The 5-th and 8-th columns (also 1-indexed) are the league affiliations of the visiting and home team, respectively. Rename these columns `v_league` and `h_league`.

2. Create a plot with two subplots, placed side-by-side. Each subplot should be a scatter plot in which the x- and y-axes correspond to the home and visitor scores, respectively, and in which each point corresponds to a game from the season. In the left-hand plot, include all games in which both teams were in the NL, and in the right-hand plot, include all games in which both teams were in the AL. Games in which the teams were from different leagues should be ignored. Specify the transparency (cf. the `alpha` parameter in the `matplotlib` documentation) so that scores that occur more often will be shaded darker than rare scores. Color the points in the scatter plot according to the league affiliation of the two teams as follows: games between two teams both in the AL should be rendered as red points in the scatter plot. Games between two teams both in the NL should be rendered as blue points in the scatter plot. Label your axes and provide an appropriate title for your plot as well as its subplots. Save your plot in a file called `home_away.pdf` and include it in your submission. **Note:** you may find it useful to create an extra column in the data frame encoding whether a given game is AL vs AL, NL vs NL or mixed.

3. The Skellam distribution (`https://en.wikipedia.org/wiki/Skellam_distribution`) is the distribution that results from taking the difference between two Poisson random variables. It is often suggested as a model for the difference between scores in sports games, particularly baseball. Add a new column to the data frame called `score_diff`, given by the home score minus the away score. Make a histogram of this score difference and give the plot an appropriate title. Save your histogram in a file `score_diff.pdf` and include it in your submission.

4. Read the documentation about the `scipy` implementation of the Skellam distribution at `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.skellam.html`. If λ_H and λ_V are the means of two independent Poisson random variables K_H and K_V, respectively, then the Skellam distribution that describes the difference $K_H - K_V$ has parameters λ_H and λ_V. We will assume (perhaps incorrectly) that the location parameter of the Skellam distribution (`loc` in the `scipy` documentation) is 0. To fit a Skellam distribution to the data, we will first fit Poisson distributions to the home and away teams.

   Estimate parameters $\hat{\lambda}_H$ and $\hat{\lambda}_V$ as the means of the home and visitor scores, respectively. There are several possible ways to assess how well a Skellam distribution with parameters $(\mu_1, \mu_2) = (\hat{\lambda}_H, \hat{\lambda}_V)$ and location parameter 0 describe the data. The most obvious would be to use a chi-square test[1] or the Cressie-Read test[2] (though admittedly, I am suggesting these primarily because they are already implemented in `scipy`). One problem with both of these is that they require either a finite number of categories or that, heuristically, the expected count be at least five for all categories. Both of these requirements are failed by the Skellam distribution, since it has infinite support, and it puts miniscule probabilities on large values. There are several possible ways to fix this problem. Indeed, if you look at the data, you will likely observe several other problems, most notably that baseball games cannot end in ties (aside from a few exceptional circumstances), and therefore our data will include few if any zeros, in contradiction with the prediction of the Skellam distribution. Taking these various concerns (and any others you may find) into account, perform an analysis to assess whether or not the Skellam distribution (perhaps af-

[1] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html#scipy.stats.chisquare
[2] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.power_divergence.html`

ter some modification, such as removing the zero outcome) adequately explains the data. **Note:** this problem is purposefully open-ended, and there is no right answer, per se. The point of this problem is to get you doing some exploratory data analysis on your own. Explain your decisions clearly and you will receive full credit (so long as said decisions are reasonable!).