IB120/201 - Lab 11

Data Wrangling

Due Date: April 10, 2020

University of California, Berkeley *GSI: Naveed Ziari*

Learning how to read data files and rearrange it for analysis is a crucial skill in doing any data analysis.

Background

Pandas

`Pandas` is a package in Python that provides functionality for fast and easy manipulation of large data structures. The best was to analogize this package is that it allows Python to work with data structures similar to data frames in R. In particular, it is useful at reading in high-throughput biological data either in the form of an `.xlsx, .csv,` or `.txt`. Remember that `pandas` is built on top of `numpy`, so both must be installed.

Questions

Please submit your assignment in the form of a iPython notebook similar to lecture. Each question has its own block.

1. Read in the file `"https://raw.githubusercontent.com/uiuc-cse/data-fa14/gh-pages/data/iris.csv"` into a data frame. This is the same dataset we worked with on R.
2. Create a new data frame only with the columns `Sepal.Length, Sepal.Width`. Write it into an Excel file.
3. Create 3 new data frames for each of the `species`.
4. Create a histogram of `Pedal.Width` for each of the 3 species. *hint: use numpy.hist()*
5. Create a new data frame where you perform some sort of transform on a column of numerical values in maximum three lines of code (e.g. multiplying by 2, taking the logarithm).
6. Create a list data structure from the column `Sepal.Length` and write a function with the input as that list which returns the mean of the column.
7. Create a dictionary with the keys being each column in the data frame (except for `species`) and the value as the mean of each column using the function you wrote above. *hint: you need to use the mean function in numpy since it is not a built-in keyword in Python*