COSC 311 - Lab 3

Dr. Joe Anderson

Due: 29 October

1 Objectives

1. Practice efficiently manipulating data with Python
2. Use the `matplotlib`, `pandas` libraries
3. Gain familiarity with statistical tools

2 Tasks

1. You may submit this lab in groups of one or two.
2. Download the "Iris" dataset from the UCI Machine Learning data repository: `https://archive.ics.uci.edu/ml/datasets/Iris`. This dataset is record of some flowers, along with their sepal and petal length and width measurements. Each flower is "labeled" as to which type of Iris family it belongs to (Setosa, Versicolour, or Virginica).
3. In each of the following plots/visualization, be sure to:
   (a) Include titles for each plot.
   (b) Include legends to differentiate which plot elements correspond to which data.
   (c) Start your y and x index at logically consistent values (usually 0 for physical measurement-type data).
4. Among the sepal/petal length/width measurements, we can define six different pairwise comparisons (sepal length vs petal width, sepal length vs sepal width, etc.); show these two parameters together in (six) different scatter plots, where each class is shown by a different color and shape marker.
   (a) What is the correlation coefficient for each pair of measurements when class is disregarded?
   (b) What is the correlation coefficient for each pair of measurements when taking into account only measurements within the same class. I.e. what is the correlation between sepal length and width overall, and what is the correlation coefficient between sepal length and width among the Setosa class, Versicolour class, and Virginica class?
5. For each of the four numerical categories, compute the mean with a 95% confidence interval and show them in a bar chart.
6. Next, compute the mean of sepal and petal measurements with 95% confidence intervals, but this time separated by each of the three classes (so you will have a total of 12 bars and intervals).
7. From the two mean estimates above, draw at least one relationship conclusion (e.g. the mean of X is larger than the mean of Y) and find the p-value that shows the strength of that conclusion. Does this mean you can reject your hypothesis or not?

(a) Looking at the data (in the form of a histogram), how close is it to a Normal distribution?

(b) If it is not close, then what does this say about your p-value?

3 Submission

Zip your source files and upload them to the assignment page on MyClasses. Be sure to include all source files, properly documented, a `README` file to describe the program and how it works, along with answers to any above discussion questions.