Q & A HW4

Q: Is the dendrogram plotted on page 15 corresponding to the data?

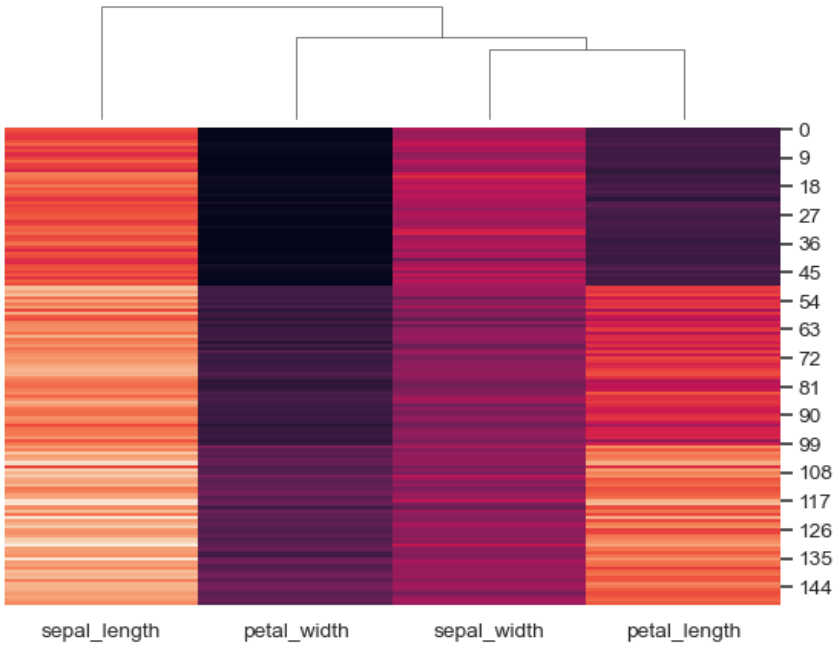
A: No. Any illustration given in lecture has been and will be only for "illustration". There are many code samples on the web and adapting them is what is being required for the homework exercises.

Q: The implementation of degreesOfFreedom(X, Y) seems incorrect.

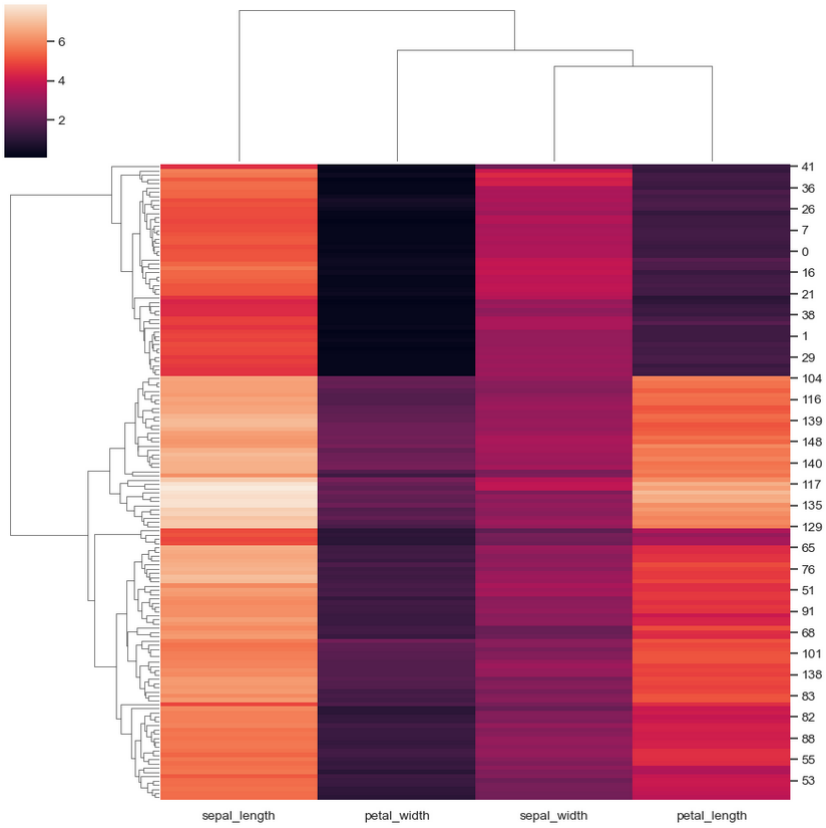
A: The DF implementation illustrated had an error in calculating stdev(X). It should have been len(X) -1, not len(X). It has been revised in the new version. The rest of DF implementation has been rewritten to make the calculation precisely follow the DF calculation given in lecture.

Q: Would you clarify further how to do gene-wise or sample-wise clustering?

A: Additional URLs are newly given in the revised HW4 description. The following images are directly from https://seaborn.pydata.org/generated/seaborn.clustermap.html. The answer for gene-wise (Part II: Step 3) would be like the one given below, considering columns in the original .csv are genes. Here column names at the bottom should be gene names and row ids should be sample ids if the clustering is to be shown in this style. Note that sns.clustermap provides a way to control which side should be used for clustering, e.g., row_cluster = False. There are other packages to create clustermap.



The answer for Part II: Step 4 should be something similar to the following style.



Q: Using 400 points results in fractional dot values - e.g., 4.7, 7.2, etc. How should we depict this?

A: Rounding down or up should do it. For visualization we will not worry of fraction.

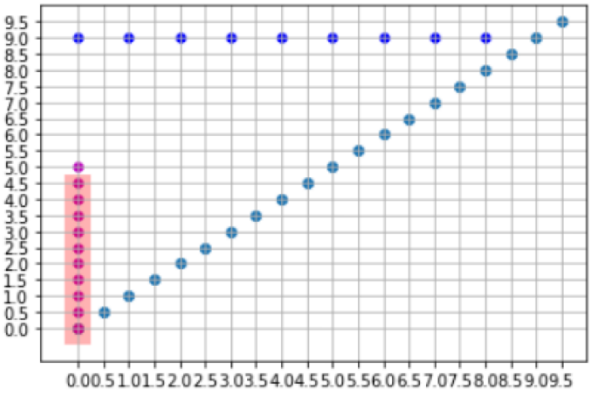
Q: There are only 9 samples from group 2, although the gene expression is a 40 x 54 matrix with 10 samples each from four groups.

A: One less sample for group 2 was an error and new versions (both .xlsx and .csv) have been uploaded.

Q: There are only 9 samples from group 2, although the gene expression is a 40 x 54 matrix with 10 samples each from four groups.

A: One less sample for group 2 was an error and new versions (both .xlsx and .csv) have been uploaded.

Q: We are supposed to use 400 samples instead of 1000, and the red and blue dots are people with drug use and without drug use, respectively. Then we box the red dots to identify the number of people who are drug users and tested positive. We concentrate the dots into one row or column to better show the percentage, and the dots lie on the anti-diagonal line represents all the 400 people we tested. I am not sure if I correctly understand the figure.



A: Note that the code given is to help you get started with "placing colored dots" on 2D grid using python; same for red-filled coloring – to suggest how to box dots. You should place dots following the convention that was used in "Frequency box visualization" of Example 2 – Drug testing slide in lecture. That is, flushed to left for users and flushed to top for non-users tested Positive (i.e., FP). In this way, it would be easier for you to color the dots properly. Do not use diagonal as given in the sample code – again this is just to illustrate you that you can place dots using an equation.

Some may think boxing dots that can show how different colored dots could be bundled be tricky depending on your python coding skill – doing it is NOT required.