Homework Exercise 8

1)

Use the **mysterious_samples.csv** file, which contains 4 datasets with X and Y values each.

A)

For each of the 4 datasets (independently), compute the Pearson correlation between X and Y.

What can you say about the relationship of these variables?

NOTE: if you calculated the correlations correctly, the r and p-values should be identical across all datasets.

B)

Draw a scatterplot visualizing each of the 4 datasets. Does this visualization change your answer to (A)?

C)

For each of the four datasets:

i) What do you think the true relationship is?
ii) Is it appropriate to use Pearson's correlation?
iii) How would you improve the analysis to reveal the true relationship between the two variables?

D)

What do you think of the statement: "numerical calculations are exact, but graphs are rough"?

2)

TiGER (http://bioinfo.wilmer.jhu.edu/tiger/) is a (rather old and not up-to-date) database for tissue-specific expression of genes in human, based on transcription factor binding sites. Download the raw data file of TiGER, showing the various tissues that appear to regulate genes (identified by RefSeq IDs) in a tissue-specific manner.

A)

Draw the number of genes reported in each tissue.

Which tissue is reported to have the highest number of tissue-specific regulated genes?

B)

Draw a heatmap showing the similarity of each pair of tissues mentioned in TiGER, given that similarity is defined by the number of common genes normalized by the total number of genes, according to the following formula:

$$s_{ij} = \log\left(\frac{N_{ij}}{\sqrt{N_i \cdot N_j}}\right)$$

where s_{ij} is the similarity of tissues i and j, N_{ij} is the number of reported genes shared by both tissues, N_i is the number of genes reported for tissue i, and N_j is the number of genes reported for tissue j. This measure gives a similarity score in the range $-\infty$ (no similarity at all) to 0 (complete similarity). By the way, the term $\frac{N_{ij}}{\sqrt{N_i \cdot N_j}}$ is closely related to Pearson's correlation, and is a commonly used similarity metric.

Because you haven't sorted the tissues by similarity groups (also known as "clusters"), expect the resulted heatmap to look somewhat messy. It is customary to cluster heatmap categories by similarity in order to obtain better-looking heatmaps, but this technique is beyond the scope of this course. If you are interested, you may read more about clustering algorithms and its uses (you can also find Python code examples online).

Which tissue seems to be the most isolated one (i.e. with the least number of similar tissues)?

Which tissues the heart seems the most similar to? Does this make sense?

Bonus Questions

1)

Learn how to create density plots and violin plots with Matplotlib. Repeat the analysis of question 1A in the lab exercise and plot the distribution of domain lengths using these plot types.

2)

Query for the following search term in UniProt: **synaptotagmin AND reviewed:yes**.

Create a 3D scatter plot in Matplotlib showing the following attributes of the queried proteins:

1. Frequency of acidic amino-acids
2. Frequency of polar amino-acids
3. Frequency of hydrophobic amino-acids

Pack in two more variables using size and color: let the length of the protein determine the size and the organism determine the color.

Make sure that you obtain an interactive 3D plot that allows you to change your “point of view” into the plot (abandon Jupyter if you need).