You are the newest member of the Atreides Lab. Their postdoc, Paul, had recently left and you are in charge of taking over the latest analysis.

Paul was collaborating with another lab and received clinical data from their team. He identified the bacteria present in the microbiome of 50 samples and calculated the taxonomic diversity for each sample.

His analyses outputted into two folders: one filled with diversity scores, and another with the final output of his distance matrices.

Your task:

1) Read in the data clinical file (clinical\_data.txt) as a dataframe and create two new columns called**averages** and **std**. You will read in each sample's diversity score list (found in **diversityScores**) and append the mean value and the standard deviation of the sample's diversity score. Then output the new clinical data (which includes the original clinical information from clinical\_data.txt along with the two new columns to a file called clinical\_data.stats.txt.)

This should be done using pandas and numpy python package.

2) Then find the animals (code names) that correspond with the two highest average diversity scores and one with the lowest average diversity score (three animals in total).  Find their respective txt file inside the **distanceFiles** directory and plot and save a scatter plot for each animal (three plots in total, and can be saved as png files). **Add a title** for each plot that has the animal's name in it.

Find max value of column to out put row

Idx max

**Extra credit.** Perform k-means on the three plots and color them by cluster. K-means is a very common algorithm to find patterns in a dataset. In our case we want to see if clustering our scatterplot reveals a biological meaning for our samples.

Use the elbow method to find the appropriate number of clusters (even though it may be very easy to see by eye).

K-means is a very common algorithm to find patterns in a dataset. In our case we want to see if clustering our scatterplot reveals a biological meaning for our samples (different species of bacteria in the microbiome).

You can use the following resources below to help you. The code for this should be in the same python script as the rest of the assignment. Feel free to make use of the scipy function for k-means (I'm not expecting you to write the code for k-means, just implement it).

What is Kmeans and when would we use it?

<https://www.datascience.com/blog/k-means-clustering>

[Code for k-means: https://mubaris.com/posts/kmeans-clustering/](https://mubaris.com/posts/kmeans-clustering/)

[Code for elbow plot: https://pythonprogramminglanguage.com/kmeans-elbow-method/](https://pythonprogramminglanguage.com/kmeans-elbow-method/)

**Requirements**

**Input files:**

You will find all necessary input files here: /home/rbif/week10/inputfiles

**Expected output:**

clinical\_data.stats.txt

3 plots with a title of the {animal}, and saved as {animal}.pdf

your python script

a readme

All of the above steps should be contained in a single python script.

You can use either matplotlib or seaborn to plot the data.

**Your code is expected to work independent of the data clinical file. For example, if I gave another file with a completely different set of animal names, your code should work without any issues.**

Please make use of the Questions and Answers forum if anything is unclear. If you need help with your code, **please provide the path on the AWS to your code.**

Assignment 5 due by Monday, midnight ET.