Author: Christopher Tran

Goal: Dataframe was pulled from clinical\_data.txt. The objective is to add an additional 2 columns of data which is standard deviation and average. With those 2 columns, the two highest averages and the smallest average will be located and graphed into a scatter plot. In addition to the scatter plot graphs, an elbow method graph and a k means cluster graph was generated

Imported tools: import pandas as pd, numpy as pd, glob,numpy, matplotlib.pyplot as plt, seaborn as sns, os, from sklearn.cluster import KMeans, from scipy.spatiual.distance import cdist, fromcopy import deepcopy

File usage: clinical\_data.txt, distanceFiles/(name).distance.txt, diversityScores/(name).diversity.txt inputfiles

Files created: clinical\_data.stats.txt, (animal).PNG, (animal)\_elbow.PNG, (animal)\_cluster.PNG, readme.txt, datatable.py

Directories created : (animal)

Execution: python3 datatable.py

Instructions: datatable.py was created to locate the two highest averages and smallest average. The script will then locate the species and create a scatter plot for all three species. Below are instructions of how the script created and run.

1. Variables were set to pull txt.files and create dictionaries. file1 is set to pull all txt files in the diversitryScores directory. File is set to pull only the clinical\_data.txt file.

2 A loop was created to open the clinical\_data.txt. Using the Numpy, an array from the clinical\_data.txt data was created and converted into a float(np.array(open file, float)). This was set into a variable. From there the mean(np.mean(variable)) and the standard devation(np.std(variable)) were calculated with the respected function in parentheses. The values were appended into respected dictionaries that were set (‘Average’ and ‘STDev’). The columns were added to the table with the values using table[column title] = variable with values(dictionary). Once these columns were added a clinical\_data.stats.txt file was created and the data was appended to the file.

3. The next objective was to find the two highest averages and the lowest average. To find that, the function .nlargest was used. Paraments were set so 2 of the highest averages were found . The function should look like table.nlargest(2,[‘Averages’]). This located the 2 highest averages. A similar function was used to find the smallest which was .nsmallest. the function was used in a similar fashion as .nlargest.

4. Once this was located, the next objective waws to isolate the animal variable. To do so, the function of n.largest/n.smallest was still used but .values was added at the end. This was to locate which value and an index was able to be located to isolate the animal.

5. variables were set for the isolated animal, pull and open the corresponding distanceFiles/(animal).distance.txt files. Variables were also set to create path for (animal)

6. If and else statements were created to create (animal) directory. Print statements were there to show if directory was created or exists.

7. The next objected is to create a scatterplot with the respected code names that were located. Using np.loadtxt() the respected files and parameters were placed and looped to create all (animal).PNG scatterplots. This was set to be the X and Y access as a variable. sns.scatterplot(X,Y), plt.title(), plt.savefig(), and plt.clf was then used within the script. The sns.scatterplt is used to created the scatter plot while, plt.title is to create the title of the file and plt.show is to show the file. Plt.savefig() is to create the file title (animal).png. Plt.clf ()is to clear the plot to create another plot separately.

8. In the same loop to created the scatterplot, another graph was created after that. This graph was created to show the elbow method of the (animal) to find the k mean of each graph for the cluster k mean. X and Y variables were the same variables used from the scatterplot. An Array was set, and the .reshape and len function as this command was set in a variable. Colors and markers were also set in a dictionary and another dictionary was created. The K range was also set from 1-10. K means was found within a loop finding the KMeans function and appending it to the dictionary. Using the variables of K and the new set dictionary, the graph was plotted and labelled corrected with the functions plt.plot(), plt.xlabel(), plt.ylabel(),plt.title(), plt.savefig(), plt.clf(). Since this was in the same loop as the scatter plot, each (animal) was able to create its own elbow graph.

9. Once the K was found from the elbow method graph, the k means cluster graph will be created next. The graph was set with the values of X (different variable names per animal for both x and y) and Y were set using the same method as creating the scatter plot. Variables were set calculating the clusters using the animal K means from the elbow method. The K mean will needed to be adjusted for each new graph made from the elbow in order for the graph to be correctly made. While and for loops were set to constructed and color the clusters. Plots wre then titled, saved and clear using the functions plt.title(), plt.savefig(), and plt.clf. This is done per animal.

Results will be shown below.

Please direct yourself to the (animal) directory to see the results of each (animal) graph.

Below is the table with added columns named 'Averages' and 'STDev':

Discoverer Location Diamater (mm) Environment Status code\_name Averages STDev

0 Kerr New Jersey 68 River Sequenced addax 7.431083 4.214558

1 Schacher Sweden 44 Sewage Sequenced badger 7.520983 4.295290

2 No Florida 6 River Sequenced basilisk 7.345858 4.452870

3 Farrior New Jersey 75 River Sequenced bat 7.524125 4.284117

4 Look Florida 66 Sewage Sequenced bear 7.304550 4.297440

5 Acosta Staten Island 64 Swamp Not sequenced beetle 7.691817 4.358099

6 Corpus Staten Island 78 Swamp Sequenced budgerigar 7.487408 4.332675

7 Barrette Staten Island 19 River Sequenced buffalo 22.379342 4.271361

8 Mauricio Florida 57 River Sequenced bull 7.315000 4.308906

9 Kelleher New Jersey 26 Swamp Not sequenced bunny 7.510517 4.319311

10 Sambrano Sweden 72 Sewage Sequenced camel 7.482875 4.412080

11 Rivet Sweden 26 Swamp Sequenced chicken 7.369000 4.296681

12 Innis New Jersey 15 Sewage Not sequenced cow 7.436858 4.289867

13 Bock Staten Island 32 River Sequenced coyote 7.586125 4.362044

14 Noonan Florida 80 Sewage Sequenced deer 7.582300 4.325630

15 Notter Florida 41 Swamp Not sequenced dingo 7.566383 4.398453

16 Lalor Florida 19 Swamp Not sequenced dog 7.442350 4.226094

17 Forness New Jersey 30 Sewage Sequenced dromedary 7.456092 4.292788

18 Snuggs Sweden 39 Swamp Sequenced eland 7.496717 4.381462

19 Breton Staten Island 42 Swamp Sequenced elephant 7.608992 4.415699

20 Klump New Jersey 62 Swamp Not sequenced finch 7.439492 4.372534

21 Vanderburg Sweden 26 River Sequenced fox 25.099075 2.910538

22 Gorham New Jersey 24 Sewage Sequenced goat 7.383275 4.278916

23 Mccafferty Florida 21 Sewage Sequenced gopher 7.630517 4.370990

24 Landey Florida 30 River Sequenced gorilla 7.468758 4.311711

25 Gies New Jersey 47 River Sequenced groundhog 7.481942 4.409542

26 Sandoz Sweden 5 Swamp Not sequenced highland 7.535417 4.335513

27 Elridge New Jersey 55 Swamp Sequenced impala 7.635158 4.324884

28 Blanch Staten Island 48 Swamp Sequenced kangaroo 7.640967 4.301210

29 Gulino New Jersey 74 Swamp Sequenced kitten 7.497775 4.255692

30 Kesner New Jersey 28 Sewage Sequenced lamb 7.642958 4.321937

31 Labrie Florida 17 Sewage Not sequenced lion 7.240383 4.365681

32 Dorrough Staten Island 4 Sewage Sequenced lizard 7.516942 4.344718

33 Knighten New Jersey 25 River Not sequenced lovebird 7.565675 4.240565

34 Frew New Jersey 58 Swamp Not sequenced mongoose 7.394742 4.312361

35 Tilton Sweden 71 River Sequenced oryx 7.718458 4.294146

36 Pickron Sweden 55 Sewage Sequenced panther 7.601642 4.283868

37 Oommen Sweden 35 River Sequenced parrot 7.519575 4.356260

38 Twine Sweden 56 Sewage Sequenced peccary 7.358483 4.270564

39 Carstensen Staten Island 65 Sewage Sequenced pronghorn 7.265492 4.304024

40 Marchal Florida 80 Swamp Sequenced puma 7.638208 4.398481

41 Beer Staten Island 40 Sewage Not sequenced salamander 7.498358 4.379672

42 Urenda Sweden 78 Swamp Not sequenced sheep 7.618367 4.292565

43 Rule Sweden 46 River Sequenced snake 7.594992 4.327635

44 Mohamed Florida 5 Swamp Sequenced snowy 7.549550 4.273408

45 Hieber New Jersey 60 River Sequenced springbok 7.439258 4.280275

46 Bush Staten Island 20 River Not sequenced turtle 7.699617 4.283735

47 Sangster Sweden 14 River Sequenced walrus 7.661108 4.353376

48 Critchlow Staten Island 33 River Not sequenced warthog 7.581808 4.330992

49 Bonura Staten Island 66 Sewage Sequenced waterbuck 7.331375 4.389672

Below is the list of the highest averages withhin the data

Discoverer Location Diamater (mm) Environment Status code\_name Averages STDev

21 Vanderburg Sweden 26 River Sequenced fox 25.099075 2.910538

7 Barrette Staten Island 19 River Sequenced buffalo 22.379342 4.271361

Below is the list of the smallest averages withhin the data

Discoverer Location Diamater (mm) Environment Status code\_name Averages STDev

31 Labrie Florida 17 Sewage Not sequenced lion 7.240383 4.365681

fox has the highest average within the data

buffalo has the second highest average within the data

lion has the smallest average within the data