CA251: Data Mining and Big Data Lab.

1. Create the following NumPy arrays:
   a) A 1-D array called *zeros* having 10 elements and all the elements are set to zero.
   b) A 1-D array called *vowels* having the elements 'a', 'e', 'i', 'o' and 'u'.
   c) A 2-D array called ones having 2 rows and 5 columns and all the elements are set to 1 and *dtype* as int.
   d) Use nested Python lists to create a 2-D array called *myarray1* having 3 rows and 3 columns and store the following data:
      i. 2.7, -2, -19
      ii. 0, 3.4, 99.9
      iii. 10.6, 0, 13
   e) A 2-D array called *myarray2* using *arange()* having 3 rows and 5 columns with start value = 4, step size 4 and *dtype* as float.

   Using the arrays created in the above, write NumPy commands for the following:

   a) Find the dimensions, shape, size, data type of the items and itemsize of arrays *zeros, vowels, ones, myarray1 and myarray2*.
   b) Reshape the array ones to have all the 10 elements in a single row.
   c) Display the 2nd and 3rd element of the array *vowels*.
   d) Display all elements in the 2nd and 3rd row of the array *myarray1.*
   e) Display the elements in the 1st and 2nd column of the array *myarray1*.
   f) Display the elements in the 1st column of the 2nd and 3rd row of the array *myarray1*.
   g) Reverse the array of *vowels.*
   h) Divide all elements of array ones by 3.
   i) Add the arrays *myarray1* and *myarray2*.
   j) Subtract *myarray1* from *myarray2* and store the result in a new array.
   k) Multiply *myarray1* and *myarray2* element wise.
   l) Do the matrix multiplication of *myarray1* and *myarray2* and store the result in a new array *myarray3*.
   m) Divide *myarray1* by *myarray2*.
   n) Find the cube of all elements of *myarray1* and divide the resulting array by 2.
   o) Find the square root of all elements of myarray2 and divide the resulting array by 2. The result should be rounded to two places of decimals.
   p) Find the transpose of ones and myarray2.
   q) Sort the array vowels in reverse.
   r) Sort the array myarray1 such that it brings the lowest value of the column in the first row and so on.
   s) Use NumPy. split() to split the array myarray2 into 5 arrays column wise. Store your resulting arrays in *myarray2A, myarray2B, myarray2C, myarray2D* and *myarray2E*. Print the arrays *myarray2A, myarray2B, myarray2C, myarray2D* and *myarray2E*.
   t) Split the array *zeros* at array index 2, 5, 7, 8 and store the resulting arrays in *zerosA, zerosB, zerosC* and *zerosD* and print them.
   u) Concatenate the arrays *myarray2A*, *myarray2B* and *myarray2C* into an array having 3 rows and 3 columns.

v) Create a 2-D array called *myarray4* using *arange()* having 14 rows and 3 columns with start value = -1, step size 0.25 having. Split this array row wise into 3 equal parts and print the result.

w) Using the *myarray4* created in the above questions, write commands for the following:

i. Find the sum of all elements.
ii. Find the sum of all elements row wise.
iii. Find the sum of all elements column wise.
iv. Find the max of all elements.
v. Find the min of all elements in each row.
vi. Find the mean of all elements in each row.
vii. Find the standard deviation column wise.

2. Write a Python program to do the following operations(Using NumPy Library)
   a) Create multi-dimensional arrays and find its shape and dimension
   b) Create a matrix full of zeros and ones
   c) Reshape and flatten data in the array
   d) Append data vertically and horizontally
   e) Apply indexing and slicing on array
   f) Use statistical functions on array - Min, Max, Mean, Median and Standard Deviation.
3. Write a Python program to do the following operations(Using NumPy Library)
   a) Dot and matrix product of two arrays
   b) Compute the Eigen values of a matrix
   c) Solve a linear matrix equation such as $3 * x^0 + x^1 = 9$ and $x^0 + 2 * x^1 = 8$.
   d) Compute the multiplicative inverse of a matrix
   e) Compute the rank of a matrix
   f) Compute the determinant of an array.
4. Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

   Write a Python program for the following using the above data:

   a) Calculate measures of central tendency.
   b) To find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data.
   c) Give the five-number summary of the data.
   d) Show a boxplot of the data.
5. Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

   a) Calculate the mean, median, and standard deviation of age and %fat.
   b) Draw the boxplots for age and %fat.
   c) Draw a scatter plot and a q-q plot based on these two variables.
6. Write a Python program to perform the following using the three fundamental Pandas data structures: the Series, DataFrame, and Index.

a) Series as generalized NumPy array
b) Series as specialized dictionary
c) Constructing Series objects
d) DataFrame as a generalized NumPy array
e) DataFrame as specialized dictionary
f) Constructing DataFrame objects:
    i. From a single Series object.
    ii. From a list of dicts.
    iii. From a dictionary of Series objects.
    iv. From a two-dimensional NumPy array.
    v. From a NumPy structured array.
g) Index as immutable array.
h) Index as ordered set.
i) Data Selection in Series:
    i. Series as dictionary
    ii. Series as one-dimensional array
    iii. Indexers: loc, iloc, and ix
j) Data Selection in DataFrame
    i. DataFrame as a dictionary
    ii. DataFrame as two-dimensional array

7. Write a Python program to perform the following:
    a) Input as CSV File
    b) Reading a CSV File
    c) Reading Specific Rows
    d) Reading Specific Columns
    e) Reading Specific Columns and Rows
    f) Reading Specific Columns for a Range of Rows
    g) Identify the missing data
    h) Identify the outlier data
    i) Replace with mean or mode
    j) Remove Blank Rows
    k) Data Categories
    l) Data types
    m) Analyze the data
    n) Visualize the data
    o) Find correlation among all attributes
8. Write a python program to perform transformation of data using Discretization (Binning) and Normalization (MinMaxScaler or MaxAbsScaler) on given dataset.
9. Write a program to implement three frequent itemset mining algorithms:
    a) Apriori
    b) FP-growth
    c) Eclat
10. Write a program to implement Decision tree algorithm.
11. Write a program to implement Naïve Bayesian Classification.
12. Write a program to implement k-means clustering algorithm.