**Starter Kits**

* Introduction to libraries
* Important/Popular libraries for data science
  + Data manipulation
  + Visualization
  + Machine Learning
* Documentation
* Working with Numpy Arrays
* Comparing Numpy Arrays with Python lists. Advantages of Numpy Arrays.

1) Why numpy array uses less memory then python lists? 10 marks

2) Write code to check blank numpy array? 2 marks

There is size attribute available for same.

>>> arr = numpy.zeros(1,0)

>>>arr.size

0

3) write code to get occurrence of value(lets say marks) in pandas dataframe when you have below dataframe? 3 marks

|  |  |  |
| --- | --- | --- |
|  | marks | Name |
| 0 | 98 | Shubham |
| 1 | 98 | Kapil |
| 2 | 96 | Sachin |
| 3 | 95 | Yogesh |

Answer :- df.groupby('marks').size()

4) Write code to extract all even from numpy array? 2 marks

Given numpy arry

array([ 1, 2, 3, 4, 5, 6, 7, 8, 9, 10])

Solution :-

>>> arr[arr%2==0]

array([ 2, 4, 6, 8, 10])

5)Write code to replace all even numbers in arr with 999 without changing array ? 3 marks

array([ 1, 2, 3, 4, 5, 6, 7, 8, 9, 10])

solution

>>> out = numpy.where(narray % 2 == 0, 999, narray)

>>> print(arr)

[ 1 2 3 4 5 6 7 8 9 10]

>>> print(out)

[ 1 999 3 999 5 999 7 999 9 999]

6) Write code to get common items from numpy array ? 3 marks

>>> x = numpy.array([11,12,13,12])

>>> y = numpy.array([17,12,10,12])

solution :-

>>> numpy.intersect1d(x,y)

array([12])

7) Write code to get values of specified indexes in numpy array? 2 marks

Given array is

array([ 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]) and index of(3,1,2)

solution:-

>>> indexes = numpy.array([3, 1, 2 ])

>>> newarr = arr[indexes]

>>> newarr

array([4, 2, 3])

8) Write code to stack two array(output should be in desired format and student have to think about horizontal or vertically stacking)? 3 marks

Note :- use function of numpy library

array a

array([[0, 1, 2, 3, 4],

[5, 6, 7, 8, 9]])

array b

array([[1, 1, 1, 1, 1],

[1, 1, 1, 1, 1]])

Desired output :-

array([[0, 1, 2, 3, 4],

[5, 6, 7, 8, 9],

[1, 1, 1, 1, 1],

[1, 1, 1, 1, 1]])

**Solution:-**

numpy.vstack([a, b])

9) Write code to split array(output should be in desired format and student have to think about horizontal or vertically splitting)? 3 marks

Note :- use function of numpy library

Given array :-

array([[0, 1, 2, 3, 4],

[5, 6, 7, 8, 9],

[1, 1, 1, 1, 1],

[1, 1, 1, 1, 1]])

Desired output :-

First array a

array([[0, 1, 2, 3, 4],

[5, 6, 7, 8, 9]])

Second array b

array([[1, 1, 1, 1, 1],

[1, 1, 1, 1, 1]])

**Solution:-**

numpy.vsplit(array,2)

10) Write code to read tsv file using python from these url <https://www.ebi.ac.uk/arrayexpress/files/E-MEXP-3682/E-MEXP-3682.sdrf.txt> 10 marks

**Solution :-**

import json

a1 = {}

a2 = open("Untitled.txt", "r");

for lin in a2.readlines():

filds = lin.rstrip().split("\t")

a1[filds[0]] = "\"" + filds[1]+"\" " +"\" " + filds[2]+"\" " +"\"" + filds[3]+"\" " +"\" " + filds[4]+"\" " +"\" " + filds[5]+"\" " +"\" " + filds[6]+"\" " +"\" " + filds[7]+"\" " +"\" " + filds[8]+"\" "

print(a1)

a2.close()

11) Write code to convert a numpy array to a dataframe of given shape?

ser = pd.Series(np.random.randint(1, 10, 35))

# Solution

df = pd.DataFrame(ser.values.reshape(7,5))

print(df)

0 1 2 3 4

0 1 2 1 2 5

1 1 2 4 5 2

2 1 3 3 2 8

3 8 6 4 9 6

4 2 1 1 8 5

5 3 2 8 5 6

6 1 5 5 4 6

12) Below exercise is related to pandas with matplotlib

**Exercise : Show Total turnover with respect to associated month in line graph**

In these exercise you will have to plot line graph with mentioned condition.

Attached data is total turnover data with specified month of different units of organization.

Generated line plot must include the following properties: –

X label name = Month

Y label name = Total turnover

**Solution :-**

import pandas as pd

import matplotlib.pyplot as plt

df = pd.read\_csv("turnover\_data.csv")

turnoverList = df ['Total\_turnover'].tolist()

months = df ['Month'].tolist()

plt.plot(months, turnoverList, label = 'Month-wise Turnover data of last year')

plt.xlabel('Month')

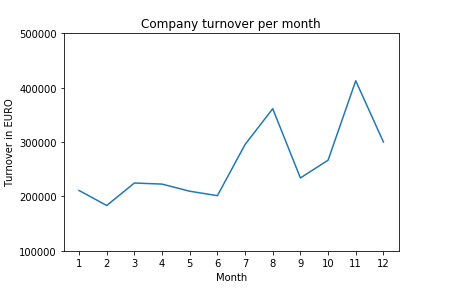
plt.ylabel('Turnover in EURO')

plt.xticks(months)

plt.title('Company turnover per month')

plt.yticks([100000, 200000, 300000, 400000, 500000])

plt.show()



13) With above data now we will do one more exercise

**Exercise: Show Total turnover with respect to associated month in line graph with below specifications.**

**Specifications for plots:-**

Line Style dashed and Line-color should be green

X label name = Month Number

Y label name = Total profit number

Add a point marker.

Line width 5

**Solution :-**

import pandas as pd

import matplotlib.pyplot as plt

df = pd.read\_csv("turnover\_data.csv")

turnoverList = df ['Total\_turnover'].tolist()

months = df ['Month'].tolist()

plt.plot(months, turnoverList, label = 'Turnover data',

color='g', marker='.',

linestyle='dashed', linewidth=5)

plt.xlabel('Month')

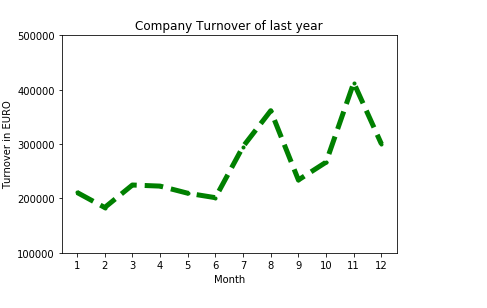
plt.ylabel('Turnover in EURO')

plt.title('Company Turnover of last year')

plt.xticks(months)

plt.yticks([100000, 200000, 300000, 400000, 500000])

plt.show()



**Memory :-**

* Introduction to Pandas Library?
* What are the Different Data Structures in Pandas?
* Summary Statistics

**DAY-1**

**2 hrs**

* Indexing and slicing numpy arrays
* Working with 2 dimensional arrays (slicing, indexing, comparison operations)
* Arithmetic operations on 2 dimensional arrays

**2 hrs**

* Iterating 2 dimensional arrays using for loops
* Operations on 2 dimensional arrays (stacking and splitting , vstack, hstack, vsplit and hsplit operations)

Lab Exercises (2 hrs)

**Day - 2**

**2 hrs**

* Explain the Data Structures with examples?
* How to create, manipulate the data frames?

**2 hrs**

* Reading data from various sources.
* Indexing, sorting, rank.

Lab Exercises (2 hours)

**Day -3**

**2 hrs**

* Merge, join, concatenate.
* Reshaping,pivoting,duplicating,mapping,replacing,groupby
* Summary statistics (Mean, Median, Mode, Skewness, Kurtosis)

Lab Exercises (2 hours)

**Day-4**

**2 hrs**

* What are the Visualization libraries?
* Overview on seaborn and matplotlib packages
* What are the Various plot using this visualization libraries?
* Distribution plots – Histogram, frequency polygon

**2 hrs**

* Representing data using charts – bar chart, pie chart
* Checking for data anomalies and outliers – box plots
* Association between variables – correlation heatmap, scatter plots, pairplots

Lab Exercises (2 hours)