UNIVERSITY OF MALAYA

IVC on Astrostatistics and Machine Learning (2-13 August 2021)

Data in Python

Handling data arrays in Python

Norhasliza Yusof

https://github.com/lizayusof/IVC_Astrostat_ML

The notes is adapted from course content SIF2007 (Norhasliza Yusof) and SQA7006 (Hasan Abu Kassim)

Reading data arrays (tips and trick)

Handling data arrays as data frame using Pandas

- Pandas is a fast, powerful, flexible way and easy to use for data analysis and manipulations tools.
- Types of file format can be handled by Pandas in Python:
    - csv
    - txt
    - xlsx
    - hdf5
    - html
    - json
    - xml

Introduction to data structure

Getting started. Reading text file and visualising data

In [1]:
```
import numpy as np
import pandas as pd
```

We want to read data from text file

In [2]:
```
df = pd.read_fwf('data.txt') #reading text file. fwf = fixed-width text file
print(df)
```

```
    1   4
0   2   8
1   3  10
2   4  16
3   5  20
4   6  30
5   7  32
6   8  36
7   9  40
8  10  60
```

The data read without a header and pandas assign as first column as 1 and second column as 2 We can plot data from the file that we read using pandas. But we need to assign the header name

In [3]:
```
import matplotlib.pyplot as plt
```

In [4]:
```
df = pd.read_fwf('data.txt', sep='\s+', header=None, names=['x','y'])
print(df)
```

```
    x   y
0   1   4
1   2   8
2   3  10
3   4  16
4   5  20
5   6  30
6   7  32
7   8  36
8   9  40
9  10  60
```

Input Output (IO) tools

Pandas I/O is a set of top level reader functions accessed like pandas.read_csv() that generally return a pandas object

Selected format type commonly used in astronomy/astrophysics

Format Type	Data Description	Reader	Writer
text	csv	read_csv	to_csv
text	Fixed-width text file	read_fwf	
text	LateX		Styler.to_latex
binary	Xlsx (MS Excel)	read_excel	to_excel
binary	OpenDocument	read_excel	
binary	HDF5 Format	read_hdf	to_hdf

More info : pandas.pydata.org

Introduction to data structure

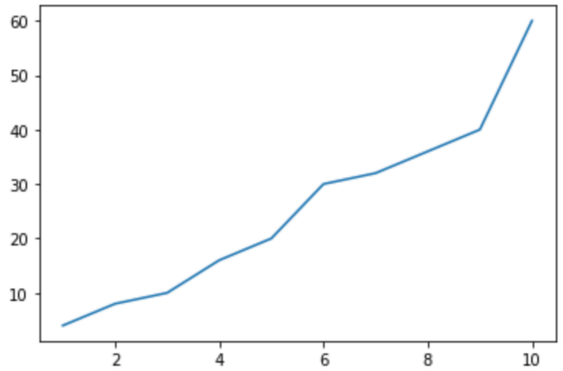
Continued from previous slide

```
In [3]: import matplotlib.pyplot as plt
```

```
In [4]: df = pd.read_fwf('data.txt', sep='\s+', header=None, names=['x','y'])
        print(df)
```

```
    x   y
0   1   4
1   2   8
2   3  10
3   4  16
4   5  20
5   6  30
6   7  32
7   8  36
8   9  40
9  10  60
```

```
In [5]: plt.plot(df['x'],df['y'])  #we assigned the header as x and x that is within the dataframe
Out[5]: [<matplotlib.lines.Line2D at 0x7ffbe3c17950>]
```



Introduction to data structure

Continued from previous slide

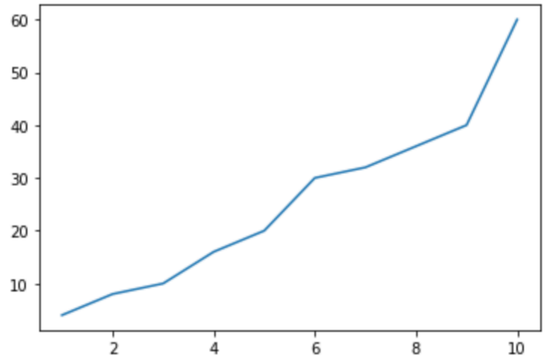
```
If your data contained header, we have to modify the read pandas statement and add header = 0
We assign new data frame to call new data. If you put the same name for data frame (df) as previous one, python
will overwrite the data frame.
```

```
In [7]: df1 = pd.read_fwf('data1.txt', sep='\s+', header=0)
        print(df1)
```

```
    x   y
0   1   4
1   2   8
2   3  10
3   4  16
4   5  20
5   6  30
6   7  32
7   8  36
8   9  40
9  10  60
```

```
Now pandas read file together with the header and note that we do not need to assign names in read pandas. Now we
plot again the data
```

```
In [9]: plt.plot(df1['x'],df1['y']) #remember the numbering for your new data frame!
Out[9]: [<matplotlib.lines.Line2D at 0x7ffbe3e7c350>]
```



Introduction to data structure

Exploring large dataset

Stars.csv - /Users/norhaslizayusof/Library/Mobile Documents/com~apple~CloudDocs/Docume...

File Edit Search Preferences Shell Macro Windows Help

```
Temperature,L,R,A_M,Color,Spectral_Class,Type
3068,0.0024,0.17,16.12,Red,M,0
3042,0.0005,0.1542,16.6,Red,M,0
2600,0.0003,0.102,18.7,Red,M,0
2800,0.0002,0.16,16.65,Red,M,0
1939,0.000138,0.103,20.06,Red,M,0
2840,0.00065,0.11,16.98,Red,M,0
2637,0.00073,0.127,17.22,Red,M,0
2600,0.0004,0.096,17.4,Red,M,0
2650,0.00069,0.11,17.45,Red,M,0
2700,0.00018,0.13,16.05,Red,M,0
3600,0.0029,0.51,10.69,Red,M,1
3129,0.0122,0.3761,11.79,Red,M,1
3134,0.0004,0.196,13.21,Red,M,1
3628,0.0055,0.393,10.48,Red,M,1
2650,0.0006,0.14,11.782,Red,M,1
3340,0.0038,0.24,13.07,Red,M,1
2799,0.0018,0.16,14.79,Red,M,1
3692,0.00367,0.47,10.8,Red,M,1
3192,0.00362,0.1967,13.53,Red,M,1
3441,0.039,0.351,11.18,Red,M,1
25000,0.056,0.0084,10.58,Blue White,B,2
7740,0.00049,0.01234,14.02,White,A,2
7220,0.00017,0.011,14.23,White,F,2
```

This is how the csv usually look like if you open using text editor/ vim/nedit

https://www.kaggle.com/brsdincer/star-type-classification

Stars

Temperature	L	R	A_M	Color	Spectral_Class	Type
3068	0.0024	0.17	16.12	Red	M	0
3042	0.0005	0.1542	16.6	Red	M	0
2600	0.0003	0.102	18.7	Red	M	0
2800	0.0002	0.16	16.65	Red	M	0
1939	0.000138	0.103	20.06	Red	M	0
2840	0.00065	0.11	16.98	Red	M	0
2637	0.00073	0.127	17.22	Red	M	0
2600	0.0004	0.096	17.4	Red	M	0
2650	0.00069	0.11	17.45	Red	M	0
2700	0.00018	0.13	16.05	Red	M	0
3600	0.0029	0.51	10.69	Red	M	1
3129	0.0122	0.3761	11.79	Red	M	1
3134	0.0004	0.196	13.21	Red	M	1
3628	0.0055	0.393	10.48	Red	M	1
2650	0.0006	0.14	11.782	Red	M	1
3340	0.0038	0.24	13.07	Red	M	1
2799	0.0018	0.16	14.79	Red	M	1
3692	0.00367	0.47	10.8	Red	M	1
3192	0.00362	0.1967	13.53	Red	M	1
3441	0.039	0.351	11.18	Red	M	1
25000	0.056	0.0084	10.58	Blue White	B	2
7740	0.00049	0.01234	14.02	White	A	2
7220	0.00017	0.011	14.23	White	F	2
8500	0.0005	0.01	14.5	White	A	2
16500	0.013	0.014	11.89	Blue White	B	2
12990	0.000085	0.00984	12.23	Yellowish White	F	2
8570	0.00081	0.0097	14.2	Blue white	A	2
7700	0.00011	0.0128	14.47	Yellowish White	F	2
11790	0.00015	0.011	12.59	Yellowish White	F	2
7230	0.00008	0.013	14.08	Pale yellow orange	F	2
39000	204000	10.6	-4.7	Blue	O	3
30000	28840	6.3	-4.2	Blue-white	B	3
15276	1136	7.2	-1.97	Blue-white	B	3
9700	74	2.89	0.16	Whitish	B	3

- If open file via Excel (in MS) or numbers (Mac OSX), the cvs file automatically converted in the column.
- Easier to see and check the data or column.
- Let's explore the data!
- Code for manipulating and visualisation of this data is available at https://github.com/lizayusof/IVC_Astrostat_ML/Intro_Python/dataset.ipynb