Pandas

Pandas is a Python module that is designed for user-friendly data analysis. It runs on top of a key Python module viz. Numpy. Pandas has become very popular among data scientists.

A quick comment on Pandas vs. Numpy (both are essentially data handling and analysis libraries). The core data structure in Numpy is a multidimensional homogeneous array - think of an n-dimensional matrix of real numbers. Numpy comes with a wide range of mathematical tools to use on such arrays. The main data structure in Pandas is the DataFrame which is tabular i.e. akin to the familiar spreadsheet. Pandas is optimised to handle and manipulate tabular data. It also meshes well with date & time variables making it convenient for time series operations. One can access various type of data files (e.g. csv, xlsx, txt, SQL etc) using Pandas, organise the data for analysis, and then use Numpy and Scipy functions to carry out the analyses. For the latest on Pandas and to read detailed documentation, go to https://pandas.pydata.org (https://pandas.pydata.org) .

Please read through a quick and helpful guide for Pandas beginners: https://pandas.pydata.org/docs/user_guide/10min.html (https://pandas.pydata.org/docs/user_guide/10min.html)

To get access to many useful Python modules, such as Numpy and Pandas, you have to tell Python to import these. The import X as Y statement instructs the current Python environment to seach for module X and provide access to its code by giving it a name, Y, in the local scope. This later allows you to reference the imported module.

import numpy as np

import pandas as pd

Import of Pandas

In [4]:
```
import pandas as pd
import numpy as np # and also load numpy
```

Create a Series with numbers 50, 10 and 20

In [5]:
```
series1 = pd.Series([50, 10, 20])
```

In [6]:
```
series1
```

Out[6]:
```
0    50
1    10
2    20
dtype: int64
```

Create another Series with colors

In [7]:
```
series2 = pd.Series(['red', 'purple', 'green', 'black'])
```

In [8]:
```
series2
```

Out[8]:
```
0       red
1    purple
2     green
3     black
dtype: object
```

Create a DataFrame

In [9]:
```
dataf = pd.DataFrame()
```

In [10]:
```
dataf
```

Out[10]:

—

In [11]:
```
dataf['Numbers'] = series1
```

In [12]:
```
dataf
```

Out[12]:

	Numbers
0	50
1	10
2	20

In [13]:
```
dataf['Colours'] = series2
```

In [14]:
```
dataf
```

Out[14]:

	Numbers	Colours
0	50	red
1	10	purple
2	20	green

In [15]:
```
dataf[:2]
```

Out[15]:

	Numbers	Colours
0	50	red
1	10	purple

In [18]:
```
dataf[1:]
```

Out[18]:

	Numbers	Colours
1	10	purple
2	20	green

In [17]:
```
dataf[1:3]
```

Out[17]:

	Numbers	Colours
1	10	purple
2	20	green

In [19]:
```
dataf[:]
```

Out[19]:

	Numbers	Colours
0	50	red
1	10	purple
2	20	green

Reading in a DataFrame

In [20]:
```
#import pandas as pd
#import numpy as np
```

In [22]:
```
df = pd.read_csv(r'C:\Documents\Beer.csv')
```

In [23]:
```
print(df)
```

```
    place     pop2023  growthRate      area                country cca3 cca2
\
0     203    10495295     0.00013     78865         Czech Republic  CZE   CZ
1      40     8958960     0.00216     83871                Austria  AUT   AT
2     616    41026067     0.02933    312679                 Poland  POL   PL
3     642    19892812     0.01188    238391                Romania  ROU   RO
4     276    83294633    -0.00090    357114                Germany  DEU   DE
5     233     1322765    -0.00249     45227                Estonia  EST   EE
6     440     2718352    -0.01153     65300              Lithuania  LTU   LT
7     516     2604172     0.01448    825615                Namibia  NAM  NaN
8     703     5795199     0.02689     49037               Slovakia  SVK   SK
9     724    47519628    -0.00082    505992                  Spain  ESP   ES
10    372     5056935     0.00673     70273                Ireland  IRL   IE
11    266     2436566     0.01991    267668                  Gabon  GAB   GA
12    178     6106869     0.02285    342000  Republic of the Congo  COG   CG
13    246     5545475     0.00085    338424                Finland  FIN   FI
14    100     6687717    -0.01390    110879               Bulgaria  BGR   BG
15    840   339996563     0.00505   9372610          United States  USA   US
16    191     4008617    -0.00539     56594                Croatia  HRV   HR
```

In [24]:
```
df.head()
```

Out[24]:

	place	pop2023	growthRate	area	country	cca3	cca2	ccn3	region	subregion	landAr
0	203	10495295	0.00013	78865	Czech Republic	CZE	CZ	203	Europe	Eastern Europe	77
1	40	8958960	0.00216	83871	Austria	AUT	AT	40	Europe	Western Europe	82
2	616	41026067	0.02933	312679	Poland	POL	PL	616	Europe	Eastern Europe	306
3	642	19892812	0.01188	238391	Romania	ROU	RO	642	Europe	Eastern Europe	230
4	276	83294633	-0.00090	357114	Germany	DEU	DE	276	Europe	Western Europe	349

In [25]:
```
df.head(10)
```

Out[25]:

	place	pop2023	growthRate	area	country	cca3	cca2	ccn3	region	subregion	landAr
0	203	10495295	0.00013	78865	Czech Republic	CZE	CZ	203	Europe	Eastern Europe	77
1	40	8958960	0.00216	83871	Austria	AUT	AT	40	Europe	Western Europe	82
2	616	41026067	0.02933	312679	Poland	POL	PL	616	Europe	Eastern Europe	306
3	642	19892812	0.01188	238391	Romania	ROU	RO	642	Europe	Eastern Europe	230
4	276	83294633	-0.00090	357114	Germany	DEU	DE	276	Europe	Western Europe	349
5	233	1322765	-0.00249	45227	Estonia	EST	EE	233	Europe	Northern Europe	42
6	440	2718352	-0.01153	65300	Lithuania	LTU	LT	440	Europe	Northern Europe	62
7	516	2604172	0.01448	825615	Namibia	NAM	NaN	516	Africa	Sub-Saharan Africa	823
8	703	5795199	0.02689	49037	Slovakia	SVK	SK	703	Europe	Eastern Europe	49
9	724	47519628	-0.00082	505992	Spain	ESP	ES	724	Europe	Southern Europe	499

In [26]:
```
df.tail()
```

Out[26]:

	place	pop2023	growthRate	area	country	cca3	cca2	ccn3	region	subregion	l
42	250	64756584	0.00201	551695	France	FRA	FR	250	Europe	Western Europe	
43	156	1425671352	-0.00015	9706961	China	CHN	CN	156	Asia	Eastern Asia	
44	764	71801279	0.00145	513120	Thailand	THA	TH	764	Asia	South-Eastern Asia	
45	608	117337368	0.01539	342353	Philippines	PHL	PH	608	Asia	South-Eastern Asia	
46	356	1428627663	0.00808	3287590	India	IND	IN	356	Asia	South Central Asia	

In [27]:
```
print(df.head())
```

```
   place   pop2023  growthRate    area         country cca3 cca2  ccn3  \
0    203  10495295     0.00013   78865  Czech Republic  CZE   CZ   203
1     40   8958960     0.00216   83871         Austria  AUT   AT    40
2    616  41026067     0.02933  312679          Poland  POL   PL   616
3    642  19892812     0.01188  238391         Romania  ROU   RO   642
4    276  83294633    -0.00090  357114         Germany  DEU   DE   276

   region       subregion  landAreaKm   density  densityMi  Rank  consmPerCap
\
0  Europe  Eastern Europe     77198.5  135.9521   352.1158    89        181.7
1  Europe  Western Europe     82520.0  108.5671   281.1889   100         96.8
2  Europe  Eastern Europe    306130.0  134.0152   347.0993    37         96.0
3  Europe  Eastern Europe    230080.0   86.4604   223.9325    64         95.0
4  Europe  Western Europe    349390.0  238.4002   617.4564    19         92.5

   consm  consmGals   pop2020  rank
0   1946     514079  10708981     1
1    872     230358   9006398     2
2   3633     959737  37846611     3
3   1828     482907  19237691     4
4   7746    2046277  83783942     5
```

In [28]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 47 entries, 0 to 46
Data columns (total 19 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   place        47 non-null     int64
 1   pop2023      47 non-null     int64
 2   growthRate   47 non-null     float64
 3   area         47 non-null     int64
 4   country      47 non-null     object
 5   cca3         47 non-null     object
 6   cca2         46 non-null     object
 7   ccn3         47 non-null     int64
 8   region       47 non-null     object
 9   subregion    47 non-null     object
 10  landAreaKm   47 non-null     float64
 11  density      47 non-null     float64
 12  densityMi    47 non-null     float64
 13  Rank         47 non-null     int64
 14  consmPerCap  47 non-null     float64
 15  consm        47 non-null     int64
 16  consmGals    47 non-null     int64
 17  pop2020      47 non-null     int64
 18  rank         47 non-null     int64
dtypes: float64(5), int64(9), object(5)
memory usage: 7.1+ KB
```

In [29]:
```
df.shape
```

Out[29]:
```
(47, 19)
```

In [30]:
```
df.dtypes
```

Out[30]:
```
place            int64
pop2023          int64
growthRate     float64
area             int64
country         object
cca3            object
cca2            object
ccn3             int64
region          object
subregion       object
landAreaKm     float64
density        float64
densityMi      float64
Rank             int64
consmPerCap    float64
consm            int64
consmGals        int64
pop2020          int64
rank             int64
dtype: object
```

In [ ]:
```

```

In [31]:
```
df.consmGals.sum()
```

Out[31]:
```
40907836
```

In [32]:
```
Var1 = df['consmPerCap']
```

In [33]:
```
Var1
```

Out[33]:
```
0     181.7
1      96.8
2      96.0
3      95.0
4      92.5
5      84.4
6      83.4
7      83.4
8      82.2
9      81.6
10     81.0
11     80.0
12     78.8
13     77.2
14     75.6
15     72.8
16     72.1
17     71.6
18     69.7
19     68.3
20     67.1
21     65.1
22     64.3
23     63.7
24     62.2
25     60.2
26     60.1
27     59.2
28     58.0
29     55.4
30     53.1
31     52.4
32     52.3
33     51.9
34     50.6
35     44.1
36     43.0
37     39.8
38     39.5
39     37.8
40     34.9
41     31.2
42     30.4
43     25.1
44     24.1
45     13.3
46      1.2
Name: consmPerCap, dtype: float64
```

In [34]:
```
Var1_v2 = df.consmPerCap
Var1_v2
```

Out[34]:
```
0     181.7
1      96.8
2      96.0
3      95.0
4      92.5
5      84.4
6      83.4
7      83.4
8      82.2
9      81.6
10     81.0
11     80.0
12     78.8
13     77.2
14     75.6
15     72.8
16     72.1
17     71.6
18     69.7
19     68.3
20     67.1
21     65.1
22     64.3
23     63.7
24     62.2
25     60.2
26     60.1
27     59.2
28     58.0
29     55.4
30     53.1
31     52.4
32     52.3
33     51.9
34     50.6
35     44.1
36     43.0
37     39.8
38     39.5
39     37.8
40     34.9
41     31.2
42     30.4
43     25.1
44     24.1
45     13.3
46      1.2
Name: consmPerCap, dtype: float64
```

In [36]:
```
df[df.consmPerCap > 90]
```

Out[36]:

	place	pop2023	growthRate	area	country	cca3	cca2	ccn3	region	subregion	landAr
0	203	10495295	0.00013	78865	Czech Republic	CZE	CZ	203	Europe	Eastern Europe	77
1	40	8958960	0.00216	83871	Austria	AUT	AT	40	Europe	Western Europe	82
2	616	41026067	0.02933	312679	Poland	POL	PL	616	Europe	Eastern Europe	306
3	642	19892812	0.01188	238391	Romania	ROU	RO	642	Europe	Eastern Europe	230
4	276	83294633	-0.00090	357114	Germany	DEU	DE	276	Europe	Western Europe	349

In [37]:
```
df = df.set_index('rank')
```

In [38]:
```
df
```

Out[38]:

	place	pop2023	growthRate	area	country	cca3	cca2	ccn3	region	subreg
rank										
1	203	10495295	0.00013	78865	Czech Republic	CZE	CZ	203	Europe	East Eur
2	40	8958960	0.00216	83871	Austria	AUT	AT	40	Europe	West Eur
3	616	41026067	0.02933	312679	Poland	POL	PL	616	Europe	East Eur
4	642	19892812	0.01188	238391	Romania	ROU	RO	642	Europe	East Eur
5	276	83294633	-0.00090	357114	Germany	DEU	DE	276	Europe	West Eur
6	233	1322765	-0.00249	45227	Estonia	EST	EE	233	Europe	North Eur
7	440	2718352	-0.01153	65300	Lithuania	LTU	LT	440	Europe	North

In [39]:
```
df = df.reset_index()
```

In [40]:
```
df
```

Out[40]:

	rank	place	pop2023	growthRate	area	country	cca3	cca2	ccn3	region	su
0	1	203	10495295	0.00013	78865	Czech Republic	CZE	CZ	203	Europe	
1	2	40	8958960	0.00216	83871	Austria	AUT	AT	40	Europe	
2	3	616	41026067	0.02933	312679	Poland	POL	PL	616	Europe	
3	4	642	19892812	0.01188	238391	Romania	ROU	RO	642	Europe	
4	5	276	83294633	-0.00090	357114	Germany	DEU	DE	276	Europe	
5	6	233	1322765	-0.00249	45227	Estonia	EST	EE	233	Europe	I
6	7	440	2718352	-0.01153	65300	Lithuania	LTU	LT	440	Europe	I

In [41]:
```
df.sort_index(ascending=False)
```

Out[41]:

	rank	place	pop2023	growthRate	area	country	cca3	cca2	ccn3	region	su
46	47	356	1428627663	0.00808	3287590	India	IND	IN	356	Asia	
45	46	608	117337368	0.01539	342353	Philippines	PHL	PH	608	Asia	
44	45	764	71801279	0.00145	513120	Thailand	THA	TH	764	Asia	
43	44	156	1425671352	-0.00015	9706961	China	CHN	CN	156	Asia	
42	43	250	64756584	0.00201	551695	France	FRA	FR	250	Europe	
41	42	380	58870762	-0.00282	301336	Italy	ITA	IT	380	Europe	S