1. Data Analysis via coginitive labs using pandas and numpy

Software - IBM Cloud Watson Studio

NumPy stands for ‘Numerical Python’ or ‘Numeric Python’. It is an open source module of Python which provides fast mathematical computation on arrays and matrices.NumPy provides the essential multi-dimensional array-oriented computing functionalities designed for high-level mathematical functions and scientific computation.

Similar to NumPy, Pandas is one of the most widely used python libraries in data science. It provides high-performance, easy to use structures and data analysis tools. Unlike NumPy library which provides objects for multi-dimensional arrays, Pandas provides in-memory 2d table object called Dataframe. It is like a spreadsheet with column names and row labels.

Import Dataset

In [2]:

```
import types
import pandas as pd
from botocore.client import Config
import ibm_boto3

def __iter__(self): return 0

# @hidden_cell
# The following code accesses a file in your IBM Cloud Object Storage. It includes your
credentials.
# You might want to remove those credentials before you share the notebook.
client_38c83f8f236a4358a4f89d762c5e9aa4 = ibm_boto3.client(service_name='s3',
    ibm_api_key_id='RtNCGnw1LwvYR3MHbf5KmmIc8bFhdT5Bp8fKqi3ZP38z',
    ibm_auth_endpoint="https://iam.cloud.ibm.com/oidc/token",
    config=Config(signature_version='oauth'),
    endpoint_url='https://s3.eu-geo.objectstorage.service.networklayer.com')

body = client_38c83f8f236a4358a4f89d762c5e9aa4.get_object(Bucket='artificialintelligenc
epracticalfi-donotdelete-pr-guqjnqp0qaqgzi',Key='Ecommerce Customers.csv')['Body']
# add missing __iter__ method, so pandas accepts body as file-like object
if not hasattr(body, "__iter__"): body.__iter__ = types.MethodType( __iter__, body )

df_data= pd.read_csv(body)
df_data.head()
```

Out[2]:

	Email	Address	Avatar	Avg. Session Length	Time on App
0	mstephenson@fernandez.com	835 Frank Tunnel\nWrightmouth, MI 82180-9605	Violet	34.497268	12.655651
1	hduke@hotmail.com	4547 Archer Common\nDiazchester, CA 06566-8576	DarkGreen	31.926272	11.109461
2	pallen@yahoo.com	24645 Valerie Unions Suite 582\nCobbborough, D...	Bisque	33.000915	11.330278
3	riverarebecca@gmail.com	1414 David Throughway\nPort Jason, OH 22070-1220	SaddleBrown	34.305557	13.717514
4	mstephens@davidson-herman.com	14023 Rodriguez Passage\nPort Jacobville, PR 3...	MediumAquaMarine	33.330673	12.795189

Pandas - pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

In [14]:

```
df_data.columns
```

Out[14]:

```
Index(['Email', 'Address', 'Avatar', 'Avg. Session Length', 'Time on App',
       'Time on Website', 'Length of Membership', 'Yearly Amount Spent'],
      dtype='object')
```

In [3]:

```
df_data.head(5)
```

Out[3]:

	Email	Address	Avatar	Avg. Session Length	Time on App
0	mstephenson@fernandez.com	835 Frank Tunnel\nWrightmouth, MI 82180-9605	Violet	34.497268	12.655651
1	hduke@hotmail.com	4547 Archer Common\nDiazchester, CA 06566-8576	DarkGreen	31.926272	11.109461
2	pallen@yahoo.com	24645 Valerie Unions Suite 582\nCobbborough, D...	Bisque	33.000915	11.330278
3	riverarebecca@gmail.com	1414 David Throughway\nPort Jason, OH 22070-1220	SaddleBrown	34.305557	13.717514
4	mstephens@davidson-herman.com	14023 Rodriguez Passage\nPort Jacobville, PR 3...	MediumAquaMarine	33.330673	12.795189

In [4]:

```
df_data.tail(5)
```

Out[4]:

	Email	Address	Avatar	Avg. Session Length	Time on App	Ti W
495	lewisjessica@craig-evans.com	4483 Jones Motorway Suite 872\nLake Jamiefurt,...	Tan	33.237660	13.566160	36.4
496	katrina56@gmail.com	172 Owen Divide Suite 497\nWest Richard, CA 19320	PaleVioletRed	34.702529	11.695736	37.1
497	dale88@hotmail.com	0787 Andrews Ranch Apt. 633\nSouth Chadburgh, ...	Cornsilk	32.646777	11.499409	38.3
498	cwilson@hotmail.com	680 Jennifer Lodge Apt. 808\nBrendachester, TX...	Teal	33.322501	12.391423	36.8
499	hannahwilson@davidson.com	49791 Rachel Heights Apt. 898\nEast Drewboroug...	DarkMagenta	33.715981	12.418808	35.7

In [5]:

```
df_data.shape
```

Out[5]:

```
(500, 8)
```

In [6]:

```
df_data.describe()
```

Out[6]:

	Avg. Session Length	Time on App	Time on Website	Length of Membership	Yearly Amount Spent
count	500.000000	500.000000	500.000000	500.000000	500.000000
mean	33.053194	12.052488	37.060445	3.533462	499.314038
std	0.992563	0.994216	1.010489	0.999278	79.314782
min	29.532429	8.508152	33.913847	0.269901	256.670582
25%	32.341822	11.388153	36.349257	2.930450	445.038277
50%	33.082008	11.983231	37.069367	3.533975	498.887875
75%	33.711985	12.753850	37.716432	4.126502	549.313828
max	36.139662	15.126994	40.005182	6.922689	765.518462

In [7]:

```
df_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 8 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   Email                 500 non-null    object 
 1   Address               500 non-null    object 
 2   Avatar                500 non-null    object 
 3   Avg. Session Length   500 non-null    float64
 4   Time on App           500 non-null    float64
 5   Time on Website       500 non-null    float64
 6   Length of Membership  500 non-null    float64
 7   Yearly Amount Spent   500 non-null    float64
dtypes: float64(5), object(3)
memory usage: 31.4+ KB
```

In [8]:

```
df_data.dtypes
```

Out[8]:

```
Email                  object
Address                object
Avatar                 object
Avg. Session Length   float64
Time on App           float64
Time on Website       float64
Length of Membership  float64
Yearly Amount Spent   float64
dtype: object
```

In [9]:

```
df_data.sort_values("Avg. Session Length")
```

Out[9]:

	Email	Address	Avatar	Avg. Session Length	Time on App	Ti W
12	knelson@gmail.com	6705 Miller Orchard Suite 186\nLake Shanestad,...	RoyalBlue	29.532429	10.961298	37.4
312	douglasdunlap@boone-rose.com	093 Larson Ports\nWest Kathryn, OK 91243	Purple	30.393185	11.802986	36.3
299	morganorozco@hotmail.com	0001 Mack Mill\nNorth Jennifer, NE 42021-5936	LightPink	30.492537	11.562936	35.9
330	dbenson@simpson.net	732 Heather Place\nNorth Michael, VT 92527	DodgerBlue	30.574364	11.351049	37.0
15	jstark@anderson.com	49558 Ramirez Road Suite 399\nPhillipstad, OH ...	Peru	30.737720	12.636606	36.2
...	...	...	...	...	...	
257	maureenlopez@gmail.com	82537 Alice Centers\nGregland, OR 71749	SeaShell	35.530904	11.379257	36.6
488	zscott@wright.com	9909 Hoffman Ranch Suite 195\nScotthaven, SC 5...	PeachPuff	35.630854	12.125402	38.1
396	waltonkaren@gmail.com	355 Villegas Isle Apt. 070\nWest Jenniferview,...	Green	35.742670	10.889828	35.5
390	michaelcampbell@yahoo.com	96480 White Lane Suite 521\nPattersonhaven, OR...	Gray	35.860237	11.730661	36.8
154	nathan86@hotmail.com	748 Michael Plaza\nWest Billyside, UT 20799	MidnightBlue	36.139662	12.050267	36.9

500 rows × 8 columns

In [10]:

```
df_data.sort_values("Avatar")
```

Out[10]:

	Email	Address	Avatar	Avg. Session Length	Time on App	T V
209	wagnerbrian@hotmail.com	50593 Wells Roads Apt. 110\nSouth Amy, MI 0696...	AliceBlue	32.559493	11.797796	37.
448	flevine@gmail.com	5292 Melanie Crescent Apt. 064\nFischerborough...	AliceBlue	32.204655	12.480702	37.
376	sfarley@jones.com	0554 Powers Curve\nNathanchester, FL 06878-6336	AntiqueWhite	32.397422	12.055340	37.
302	lmalone@gmail.com	USS Beasley\nFPO AP 50556-7615	AntiqueWhite	32.975193	13.394452	37.
208	freemantina@cannon.org	870 Dennis Throughway\nWilsonport, PW 12658	AntiqueWhite	32.903454	10.542645	35.
...	...	...	...	...	...	
145	linda90@yoder.org	979 Alison Motorway Apt. 676\nNorth Frank, HI ...	WhiteSmoke	33.477190	12.488067	36.
207	lewiskendra@yahoo.com	988 Matthew Plaza\nLake Jacobshire, AL 37889	Yellow	33.324241	11.084584	36.
238	archeremily@baldwin.com	USCGC Hernandez\nFPO AE 53064	YellowGreen	31.260647	13.266760	36.
351	tluna@hotmail.com	01512 Hendricks Rue\nEast Pamela, PR 46481	YellowGreen	32.189845	11.386776	38.
469	kstafford@estes-nguyen.com	PSC 4856, Box 1297\nAPO AA 17032-7944	YellowGreen	31.169507	13.970181	36.

500 rows × 8 columns

In [12]:

```
df_data["Avatar"].value_counts()
```

Out[12]:

```
CadetBlue               7
Teal                    7
Cyan                    7
SlateBlue               7
GreenYellow             7
                      ..
PowderBlue              1
MediumPurple            1
Coral                   1
LightGoldenRodYellow    1
PaleGreen               1
Name: Avatar, Length: 138, dtype: int64
```

In [13]:

```
df_data.nunique()
```

Out[13]:

```
Email                   500
Address                 500
Avatar                  138
Avg. Session Length     500
Time on App             500
Time on Website         500
Length of Membership    500
Yearly Amount Spent     500
dtype: int64
```

In [15]:

```
df_data['Time on App'].mean()
```

Out[15]:

```
12.052487937166132
```

In [17]:

```
df_data[0:-1]
```

Out[17]:

	Email	Address	Avatar	Avg. Session Length	Time A
0	mstephenson@fernandez.com	835 Frank Tunnel\nWrightmouth, MI 82180-9605	Violet	34.497268	12.655
1	hduke@hotmail.com	4547 Archer Common\nDiazchester, CA 06566-8576	DarkGreen	31.926272	11.109
2	pallen@yahoo.com	24645 Valerie Unions Suite 582\nCobbborough, D...	Bisque	33.000915	11.330
3	riverarebecca@gmail.com	1414 David Throughway\nPort Jason, OH 22070-1220	SaddleBrown	34.305557	13.717
4	mstephens@davidson-herman.com	14023 Rodriguez Passage\nPort Jacobville, PR 3...	MediumAquaMarine	33.330673	12.795
...	...	...	...	...	
494	kellydeborah@chan.biz	354 Sanchez Wall Suite 884\nJuliabury, VI 39735	DarkTurquoise	33.431097	13.350
495	lewisjessica@craig-evans.com	4483 Jones Motorway Suite 872\nLake Jamiefurt,...	Tan	33.237660	13.566
496	katrina56@gmail.com	172 Owen Divide Suite 497\nWest Richard, CA 19320	PaleVioletRed	34.702529	11.695
497	dale88@hotmail.com	0787 Andrews Ranch Apt. 633\nSouth Chadburgh, ...	Cornsilk	32.646777	11.499
498	cwilson@hotmail.com	680 Jennifer Lodge Apt. 808\nBrendachester, TX...	Teal	33.322501	12.391

499 rows × 8 columns

Using Numpy

NumPy is a commonly used Python data analysis package. By using NumPy, you can speed up your workflow, and interface with other packages in the Python ecosystem, like scikit-learn, that use NumPy under the hood. NumPy was originally developed in the mid 2000s, and arose from an even older package called Numeric

In [25]:

```
import numpy as np
```

In [26]:

```
df_data = np.array(df_data)
```

In [27]:

```
df_data
```

Out[27]:

```
array([['mstephenson@fernandez.com',
        '835 Frank Tunnel\nWrightmouth, MI 82180-9605', 'Violet', ...,
        39.57766801952616, 4.0826206329529615, 587.9510539684005],
       ['hduke@hotmail.com',
        '4547 Archer Common\nDiazchester, CA 06566-8576', 'DarkGreen',
        ..., 37.268958868297744, 2.66403418213262, 392.2049334443264],
       ['pallen@yahoo.com',
        '24645 Valerie Unions Suite 582\nCobbborough, DC 99414-7564',
        'Bisque', ..., 37.110597442120856, 4.104543202376424,
        487.54750486747207],
       ...,
       ['dale88@hotmail.com',
        '0787 Andrews Ranch Apt. 633\nSouth Chadburgh, TN 56128',
        'Cornsilk', ..., 38.33257633196044, 4.958264472618699,
        551.6201454762477],
       ['cwilson@hotmail.com',
        '680 Jennifer Lodge Apt. 808\nBrendachester, TX 05000-5873',
        'Teal', ..., 36.840085729767004, 2.336484668112853,
        456.46951006629797],
       ['hannahwilson@davidson.com',
        '49791 Rachel Heights Apt. 898\nEast Drewborough, OR 55919-9528',
        'DarkMagenta', ..., 35.771016191612965, 2.7351595670822753,
        497.7786422156802]], dtype=object)
```

In [30]:

```
#indexing numpy arrays
df_data[2,3]
```

Out[30]:

```
33.000914755642675
```

In [31]:

```
#slicing numpy arrays
df_data[0:3,3]
```

Out[31]:

```
array([34.49726772511229, 31.92627202636016, 33.000914755642675],
      dtype=object)
```

In [33]:

```
df_data[:3]
```

Out[33]:

```
array([['mstephenson@fernandez.com',
        '835 Frank Tunnel\nWrightmouth, MI 82180-9605', 'Violet',
        34.49726772511229, 12.65565114916675, 39.57766801952616,
        4.0826206329529615, 587.9510539684005],
       ['hduke@hotmail.com',
        '4547 Archer Common\nDiazchester, CA 06566-8576', 'DarkGreen',
        31.92627202636016, 11.109460728682564, 37.268958868297744,
        2.66403418213262, 392.2049334443264],
       ['pallen@yahoo.com',
        '24645 Valerie Unions Suite 582\nCobbborough, DC 99414-7564',
        'Bisque', 33.000914755642675, 11.330278057777512,
        37.110597442120856, 4.104543202376424, 487.54750486747207]],
      dtype=object)
```

In [34]:

```
df_data[:,:]
```

Out[34]:

```
array([['mstephenson@fernandez.com',
        '835 Frank Tunnel\nWrightmouth, MI 82180-9605', 'Violet', ...,
        39.57766801952616, 4.0826206329529615, 587.9510539684005],
       ['hduke@hotmail.com',
        '4547 Archer Common\nDiazchester, CA 06566-8576', 'DarkGreen',
        ..., 37.268958868297744, 2.66403418213262, 392.2049334443264],
       ['pallen@yahoo.com',
        '24645 Valerie Unions Suite 582\nCobbborough, DC 99414-7564',
        'Bisque', ..., 37.110597442120856, 4.104543202376424,
        487.54750486747207],
       ...,
       ['dale88@hotmail.com',
        '0787 Andrews Ranch Apt. 633\nSouth Chadburgh, TN 56128',
        'Cornsilk', ..., 38.33257633196044, 4.958264472618699,
        551.6201454762477],
       ['cwilson@hotmail.com',
        '680 Jennifer Lodge Apt. 808\nBrendachester, TX 05000-5873',
        'Teal', ..., 36.840085729767004, 2.336484668112853,
        456.46951006629797],
       ['hannahwilson@davidson.com',
        '49791 Rachel Heights Apt. 898\nEast Drewborough, OR 55919-9528',
        'DarkMagenta', ..., 35.771016191612965, 2.7351595670822753,
        497.7786422156802]], dtype=object)
```

In [35]:

```
#to generate a random vector

np.random.rand(3)
```

Out[35]:

```
array([0.56241195, 0.81095058, 0.5500436 ])
```

In [ ]:

```

```