

Ch.-1 Pandas

1.1 Itroudction

1. Python library used to working with database.
2. Analyzing, Cleaning, Explaining, and manipulatig data.
3. Pandas
 - A. Panel Data
 - B. Python data analysis
4. Created by Wes McKinney in 2008.
5. Analyze big data and make conclusion based on stastical analysis.
6. Clean the messy data sets, make them readable and relevent.
7. Relevent data is important in data science.

1.2 Pandas Series

1. Series is like "Column in a table".
2. 1-D array holding the data of any type.

```
In [5]: import pandas as pd
a=[1,2,3,4]
myvar=pd.Series(a)
print(myvar)
print(myvar[2])
print(type(myvar))
myvar1=pd.Series(a,index=['w','x','y','z'])
print(myvar1)

0    1
1    2
2    3
3    4
dtype: int64
3
<class 'pandas.core.series.Series'>
w    1
x    2
y    3
z    4
dtype: int64
```

```
In [1]: import pandas as pd
a=[1.4,2,3,4]
myvar=pd.Series(a)
print(myvar)

0    1.4
1    2.0
2    3.0
3    4.0
dtype: float64
```

```
In [2]: import pandas as pd
sname=['ram','jenil','dhruv']
marks=[10,35,45]
myvar=pd.Series(sname,index=marks)
print(myvar)

10    ram
35   jenil
45   dhruv
dtype: object
```

```
In [4]: import pandas as pd
categories={'days1':400,'days2':250,'days3':350}
myvar=pd.Series(categories)
myvar

days1    400
days2    250
days3    350
dtype: int64
```

```
In [6]: import pandas as pd
categories={'days1':400,'days2':250,'days3':350}
myvar=pd.Series(categories,index=['days1','days2'])
myvar

days1    400
days2    250
dtype: int64
```

```
In [11]: import pandas as pd
marks={'Ram':100,'Shyam':80,'Radha':50}
m=pd.Series(marks,name='student result')
m

Ram      100
Shyam     80
Radha     50
Name: student result, dtype: int64
```

1.3 Series Attributes

```
In [12]: #size
import pandas as pd
marks={'Ram':100,'Shyam':80,'Radha':50}
m=pd.Series(marks,name='student result')
m.size

Out[12]: 3
```

```
In [13]: #Datatype
import pandas as pd
marks={'Ram':100,'Shyam':80,'Radha':50}
m=pd.Series(marks,name='student result')
m.dtype

Out[13]: dtype('int64')
```

```
In [14]: #Name
import pandas as pd
marks={'Ram':100,'Shyam':80,'Radha':50}
m=pd.Series(marks,name='student result')
m.name

Out[14]: 'student result'
```

```
In [15]: #Index
import pandas as pd
marks={'Ram':100,'Shyam':80,'Radha':50}
m=pd.Series(marks,name='student result')
m.index

Out[15]: Index(['Ram', 'Shyam', 'Radha'], dtype='object')
```

```
In [16]: #Values
import pandas as pd
marks={'Ram':100,'Shyam':80,'Radha':50}
m=pd.Series(marks,name='student result')
m.values

Out[16]: array([100,  80,  50], dtype=int64)
```

```
In [17]: #is_unique
import pandas as pd
marks={'Ram':100,'Shyam':80,'Radha':50}
m=pd.Series(marks,name='student result')
m.is_unique

Out[17]: True
```

```
In [18]: import pandas as pd
marks={'Ram':100,'Shyam':100,'Radha':50}
m=pd.Series(marks,name='student result')
m.is_unique

Out[18]: False
```

```
In [20]: import pandas as pd
subs=pd.read_csv("subs.csv")
subs
```

```
Out[20]:
```

Subscribers gained	
0	48
1	57
2	40
3	43
4	44
...	...
360	231
361	226
362	155
363	144
364	172

365 rows × 1 columns

```
In [21]: import pandas as pd
subs=pd.read_csv("subs.csv")
type(subs)

Out[21]: pandas.core.frame.DataFrame
```

```
In [22]: import pandas as pd
subs=pd.read_csv("subs.csv")
print(subs)

      Subscribers gained
0              48
1              57
2              40
3              43
4              44
..              ...
360            231
361            226
362            155
363            144
364            172

[365 rows x 1 columns]
```

```
In [23]: import pandas as pd
subs=pd.read_csv("subs.csv",squeeze=True)
print(subs)
print(type(subs))

0    48
1    57
2    40
3    43
4    44
...
360   231
361   226
362   155
363   144
364   172
Name: Subscribers gained, Length: 365, dtype: int64
<class 'pandas.core.series.Series'>
```

```
In [24]: import pandas as pd
subs=pd.read_csv("subs.csv")
subs.describe()
```

```
Out[24]:
```

Subscribers gained	
count	365.000000
mean	135.643836
std	62.675023
min	33.000000
25%	88.000000
50%	123.000000
75%	177.000000
max	396.000000

```
In [29]: import pandas as pd
subs=pd.read_csv("subs.csv",squeeze=True)
subs[subs>200]
```

```
Out[29]:
```

165	225
166	249
167	265
168	306
169	261
170	222
225	224
226	254
227	214
228	236
229	261
230	247
231	207
232	254
233	301
234	233
240	202
246	259
248	213
249	210
252	216
275	228
276	237
277	276
279	290
280	295
281	245
282	241
283	267
284	269
285	211
287	206
288	202
290	236
291	222
292	203
312	230
318	244
324	243
325	227
326	259
330	396
331	312
332	268
333	229
337	221
338	225
339	221
340	202
344	276
345	258
346	219
347	212
351	220
354	204
358	210
359	209
360	231
361	226

Name: Subscribers gained, dtype: int64

```
In [30]: import pandas as pd
subs=pd.read_csv("subs.csv",squeeze=True)
subs[subs>200].size

Out[30]: 59
```

```
In [34]: import pandas as pd
bolly=pd.read_csv("bollywood.csv",squeeze=True,index_col="movie")
bolly

Out[34]:
```

movie	
Uri: The Surgical Strike	Vicky Kaushal
Battalion 609	Vicky Ahuja
The Accidental Prime Minister (film)	Anupam Kher
Why Cheat India	Emraan Hashmi
Evening Shadows	Mona Ambegaonkar
	...
Hum Tumhare Hain Sanam	Shah Rukh Khan
Aankhen (2002 film)	Amitabh Bachchan
Saathiya (film)	Vivek Oberoi
Company (film)	Ajay Devgn
Awara Paagal Deewana	Akshay Kumar
Name: lead, Length: 1500, dtype: object	

```
In [36]: import pandas as pd
bolly=pd.read_csv("bollywood.csv",squeeze=True,index_col="movie")
bolly.head(10)
```

```
Out[36]:
```

movie	
Uri: The Surgical Strike	Vicky Kaushal
Battalion 609	Vicky Ahuja
The Accidental Prime Minister (film)	Anupam Kher
Why Cheat India	Emraan Hashmi
Evening Shadows	Mona Ambegaonkar
Soni (film)	Geetika Vidya Ohlyan
Fraud Saiyaan	Arshad Warsi
Bombairiya	Radhika Apte
Manikarnika: The Queen of Jhansi	Kangana Ranaut
Thackeray (film)	Nawazuddin Siddiqui
Name: lead, dtype: object	

```
In [38]: import pandas as pd
bolly=pd.read_csv("bollywood.csv",squeeze=True,index_col="movie")
bolly.tail()
bolly[0]
```

```
Out[38]: 'Vicky Kaushal'
```

```
In [ ]:
```