Pandas-Series

Saturday, October 7, 2023 9:58 AM

What is 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.

Pandas Series

A Pandas Series is like a column in a table. It is a 1-D array holding data of any type. (one data column)

Importing Pandas

```
import numpy as np
import pandas as pd
```

Series from lists

```
# string
country = ['India','Pakistan','USA','Nepal','Srilanka']
pd.Series(country)
```

→ It will convert into Series.

```
# integers
runs = [13,24,56,78,100]
runs_ser = pd.Series(runs)

# custom index
marks = [67,57,89,100]
subjects = ['maths','english','science','hindi']
pd.Series(marks,index=subjects)

# setting a name
marks = pd.Series(marks,index=subjects,name='Nabin Ko marks')
marks
```

Series from dict

```
marks = {
   'maths':67,
   'english':57,
   'science':89,
   'Nepali':100
}

marks_series = pd.Series(marks,name='nabin ko marks')
marks_series
```

Series Attributes

```
# size
marks_series.size

# dtype
marks_series.dtype

# name
marks_series.name

# is_unique
marks_series.is_unique
pd.Series([1,1,2,3,4,5]).is_unique

# index
marks_series.index
runs_ser.index

# values
marks_series.values
```

Series using read_csv

```
# with one col
subs = pd.read_csv('/content/subs.csv',squeeze=True)
subs

# with 2 cols
vk = pd.read_csv('/content/kohli_ipl.csv',index_col='match_no',squeeze=True)
vk

movies = pd.read_csv('/content/bollywood.csv',index_col='movie',squeeze=True)
movies
```

This parameter helps to convert into series.

Series methods

```
# head and tail
subs.head()
vk.tail(10)

# sample
movies.sample(5)

# value_counts -> movies
movies.value_counts()

# sort_values -> inplace
vk.sort_values(ascending=False).head(1).values[0]
vk.sort_values(ascending=False)

# sort_index -> inplace -> movies
movies.sort_index(ascending=False,inplace=True)
Movies

vk.sort_values(inplace=True)
vk
```

Series Maths Methods

```
# count
vk.count()

# sum -> product
subs.sum()

# mean -> median -> mode -> std -> var
subs.mean()
print(vk.median())
print(movies.mode())
print(subs.std())
print(vk.var())

# min/max
subs.max()

# describe
subs.describe()
```

Series Indexing

```
# integer indexing
x = pd.Series([12,13,14,35,46,57,58,79,9])
x

# negative indexing
x[-1]
marks_series[-1]

# slicing
vk[5:16]
```

is not possible but in string chan.

```
# negative slicing
vk[-5:]
movies[::2]

# fancy indexing
vk[[1,3,4,5]]

# indexing with labels -> fancy indexing
movies['2 States (2014 film)']
```

Editing Series

```
# using indexing
marks_series[1] = 100
marks_series

# what if an index does not exist
marks_series['evs'] = 100

marks_series

# slicing
runs_ser[2:4] = [100,100]
runs_ser

# fancy indexing
runs_ser[[0,3,4]] = [0,0,0]
runs_ser

# using index label
movies['2 States (2014 film)'] = 'Alia Bhatt'
movies
```

It will create new column

Series with Python Functionalities

```
# len/type/dir/sorted/max/min
print(len(subs))
print(type(subs))
print(dir(subs))
print(sorted(subs))
print(min(subs))
print(max(subs))

# type conversion
list(marks_series)
dict(marks_series)

# membership operator
'2 States (2014 film)' in movies
'Alia Bhatt' in movies.values
movies

# looping
for i in movies.index:
  print(i)

# Arithmetic Operators(Broadcasting)
100 + marks_series

# Relational Operators
vk >= 50
```

Boolean Indexing on Series

```
# Find no of 50's and 100's scored by kohli
vk[vk >= 50].size

# find number of ducks
vk[vk == 0].size
```

```
# Count number of day when I had more than 200 subs a day
subs[subs > 200].size

# find actors who have done more than 20 movies
num_movies = movies.value_counts()
num_movies[num_movies > 20]
```

Plotting Graphs on Series

```
subs.plot()
movies.value_counts().head(20).plot(kind='pie')
```

Some Important Series Methods

```
# astype
# between
# clip
# drop_duplicates
# isnull
# dropna
# fillna
# isin
# apply
# copy

import numpy as np
import pandas as pd

subs = pd.read_csv('/content/subs.csv',squeeze=True)
subs

vk = pd.read_csv('/content/kohli_ipl.csv',index_col='match_no',squeeze=True)
vk

movies = pd.read_csv('/content/bollywood.csv',index_col='movie',squeeze=True)
movies

# astype
import sys
sys.getsizeof(vk)
sys.getsizeof(vk.astype('int16'))

# between
vk[vk.between(51,99)].size

# clip
subs
subs.clip(100,200)

# drop_duplicates
temp = pd.Series([1,1,2,2,3,3,4,4])
temp

temp.drop_duplicates(keep='last')

temp.duplicated().sum()

vk.duplicated().sum()
movies.drop_duplicates()

temp = pd.Series([1,2,3,np.nan,5,6,np.nan,8,np.nan,10])
temp
temp.size
temp.count()

# isnull
temp.isnull().sum()

# dropna
temp.dropna()
```

→ less than 100 will be 100 & more than 200 will be 200 but value bet 100 & 200 will n't change.

 # fillna
```
temp.fillna(temp.mean())
```

 # isin
```
vk[(vk == 49) | (vk == 99)]
vk[vk.isin([49,99])]
```

 # apply
```
movies
movies.apply(lambda x:x.split()[0].upper())
subs
subs.apply(lambda x:'good day' if x > subs.mean() else 'bad day')
subs.mean()
```

 # copy

```
new = vk.copy()
```