PANDAS

Data analysis: It is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, formation conclusion, and supporting decision-making.

Python libraries for data analysis: NUMPY,STATSMODEL,MATPLOTLIB,SCIPY,SCIKIT LEARN,PANDAS,SEABORN.

What is Pandas?

* The name “pandas” has a reference to both “panel data”, and “python data analysis” and was created by wes mckinney in 2008
* Pandas is a python library used for working with datasets
* It has functions for analyzing, cleaning, exporting, and manipulating data.
* Read and write data structures and different formats : CSV,T XML ,JSON,ZIP etc.

Importance of Pandas in Python

* Pandas allows us to analyze big data and make conclusion based on statistical theories.
* Pandas can clean messy data sets, and make them readable and relevant.
* Easy handling of missing data in floating as well as non-floating point data.
* Size mutability: Columns can be inserted and deleted from DataFrame and higher dimensional objects.
* Data set merging and joining, Flexing reshaping and pivoting of data sets providing time-series functionality.

Data structures in python Pandas

* The pandas provide two data structures for processing the data.

Series and DataFrame and panel

\*\*Series: It is defined as a one-dimensional array that is capable of storing various data types.

#syntex:

Import pandas as pd

x=[3,4,5,6,7,8]

var=pd.Series(x)

print(var,type(var))

print(var[2])

output:

0 3

1 4

2 5

3 6

4 7

5 8

dtype: int64 <class 'pandas.core.series.Series'>

5

#CODE

x=[1,2,3,4]

var=pd.Series(x,index=['a','s','d','f'],dtype="float",name="python")

print(var)

output:

a 1.0

s 2.0

d 3.0

f 4.0

Name: python, dtype: float64

#CODE

dic={"name":['pyhon','c','c++','java'],"por":[12,13,14,15],"rank":[1,4,3,2]}

var=pd.Series(dic)

print(var)

output:

name [pyhon, c, c++, java]

por [12, 13, 14, 15]

rank [1, 4, 3, 2]

dtype: object

#CODE

svar=pd.Series(12,index=[1,2,3,4,5])

print(svar)

output:

1 12

2 12

3 12

4 12

5 12

dtype: int64

DataFrame in python pandas

A Pandas DataFrame is **a 2 dimensional data structure**, like a 2 dimensional array, or a table with rows and columns.

#CODE

l=[1,2,3,4]

var=pd.DataFrame(l)

print(var)

print(type(var))

output:

0

0 1

1 2

2 3

3 4

<class 'pandas.core.frame.DataFrame'>

#CODE

d={"a":[1,2,3,4,5],"s":[1,2,3,4,5]}

var=pd.DataFrame(d)

print(type(var))

print(var)

output:

<class 'pandas.core.frame.DataFrame'>

a s

0 1 1

1 2 2

2 3 3

3 4 4

4 5 5

NOTE: Here size of a and s should be same

#code

d={"a":[1,2,3,4,5],"s":[1,2,3,4,5],"d":[1,2,3,4,5],1:[1,2,3,4,5]}

var=pd.DataFrame(d,columns=["a",1],index=["a","s","d","f","g"])

print(type(var))

print(var)

output:

<class 'pandas.core.frame.DataFrame'>

a 1

a 1 1

s 2 2

d 3 3

f 4 4

g 5 5

#CODE  
d={"a":[1,2,3,14,5],"s":[1,2,3,4,5],"d":[1,2,3,4,5],1:[1,2,3,4,5]}

var=pd.DataFrame(d)

print(type(var))

print(var)

print()

print(var["a"][3])

output:

<class 'pandas.core.frame.DataFrame'>

a s d 1

0 1 1 1 1

1 2 2 2 2

2 3 3 3 3

3 14 4 4 4

4 5 5 5 5

14

DataFrame using series:

#Code

d={"s":pd.Series([1,2,3,4]),"f":pd.Series([1,2,3,4])}

var=pd.DataFrame(d)

print(var)

output:

s f

0 1 1

1 2 2

2 3 3

3 4 4

#Arithmetic operations in Python Pandas

\*Addition/subtraction/multiply/divide

#CODE

var=pd.DataFrame({"A":[1,2,3,4],"B":[5,6,7,8]})

print(var)

output:

A B

0 1 5

1 2 6

2 3 7

3 4 8

var["C"]=var["A"]+var["B"]

print(var)

output:

A B C

0 1 5 6

1 2 6 8

2 3 7 10

3 4 8 12

var["C"]=var["A"] \* var["B"]

print(var)

output:

A B C

0 1 5 5

1 2 6 12

2 3 7 21

3 4 8 32

#Code to check the values in each columns of DataFrame

var["Check\_A"]=var["A"] <=3

var["Check\_B"]=var["B"] >=7

var["Check\_C"]=var["C"] >=15

print(var)

OUTPUT:

A B C Check\_A Check\_B Check\_C

0 1 5 5 True False False

1 2 6 12 True False False

2 3 7 21 True True True

3 4 8 32 False True True

**#Delete and Insert Data in Pandas**

* Insert