Pandas Interview Q&A

1. Different data structures in pandas.

A **Series** is a one-dimensional data structure in pandas, whereas the **DataFrame** is the two-dimensional data structure in pandas.

1. Define series

One-dimensional array that can store various data types.

The row labels of series are called the **index**.

By using a '**series**' method, we can easily convert the list, tuple, and dictionary into series.

A Series cannot contain multiple columns.

1. Define dataframe

A DataFrame is a widely used data structure of pandas and works with a two-dimensional array with labeled axes (rows and columns) DataFrame is defined as a standard way to store data and has two different indexes, i.e., row index and column index. It consists of the following properties:

The columns can be heterogeneous types like int and bool.

It can be seen as a dictionary of Series structure where both the rows and columns are indexed. It is denoted as "columns" in the case of columns and "index" in case of rows

1. Features of pandas

Memory Efficient

Data Alignment

Reshaping

Merge and join

Time Series

1. What is reindexing?

Reindexing is used to conform DataFrame to a new index with optional filling logic. It places NA/NaN in that location where the values are not present in the previous index. It returns a new object unless the new index is produced as equivalent to the current one, and the value of copy becomes False. It is used to change the index of the rows and columns of the DataFrame.

1. Pandas library tools used to create a scatter plot matrix?

Scatter\_matrix

1. Different ways of creating a pandas dataframe?

Lists, dict of ndarrays

1. What is categorical data?

Datatype that refers to statistically categorical data that has fixed number of possible values like gender, country, blood type etc.

1. Create a dataframe from 2 lists.

Pd.DataFrame(np.column\_stack([list1, list2]), columns=[‘Id’, ‘Name’])

1. Create a series from a dictionary?

Key becomes column, value becomes record.

import pandas as pd

import numpy as np

info = {'x' : 0., 'y' : 1., 'z' : 2.}

a = pd.Series(info)

print (a)

1. Difference between copy and deep copy.

Copy will do a shallow copy, so updating original series will update the contents of shallow copy as well, but deep will create a series with its own index and data not updating when original series is updated.

1. How to create an empty dataframe

import pandas as pd

info = pd.DataFrame()

print (info)

1. How can you remove duplicates from a dataframe?

Df.drop\_duplicates()

1. How can you get elements from one series not present in another series?

import pandas as pd

p1 = pd.Series([2, 4, 6, 8, 10])

p2 = pd.Series([8, 10, 12, 14, 16])

p1[~p1.isin(p2)]

1. How can you get elements not common to both series P1 and P2?

import pandas as pd

import numpy as np

p1 = pd.Series([2, 4, 6, 8, 10])

p2 = pd.Series([8, 10, 12, 14, 16])

p\_u = pd.Series(np.union1d(p1, p2)) # union

p\_i = pd.Series(np.intersect1d(p1, p2)) # intersect

p\_u[~p\_u.isin(p\_i)]

1. How to get the minimum, 25th percentile, median, 75th, and max of a numeric series?

import pandas as pd

import numpy as np

p = pd.Series(np.random.normal(14, 6, 22))

state = np.random.RandomState(120)

p = pd.Series(state.normal(14, 6, 22))

np.percentile(p, q=[0, 25, 50, 75, 100])