By OnlineInterviewQuestions.com

Python Pandas Interview Questions

Q1. What is Python pandas?

Pandas is a software library written for **Python** that is mainly used to analyze and manipulate data. It is an open-source, cross-platform library written **by Wes Mckinney** and **released in 2008**. This library offers data structures and operations for manipulating numerical and time-series data.

You can install Pandas using pip or with the Anaconda distribution. With this package, you can easily and quickly perform machine learning operations on the table data.

Q2. List major features of the Python pandas?

Some of the major features of Python Pandas are,

- Fast and efficient in handling the data with its DataFrame object.
- It provides tools for loading data into in-memory data objects from various file formats.
- It has high-performance in merging and joining data.
- It has Time Series functionality.
- It provides functions for Data set merging and joining.
- It has functionalities for label-based slicing, fancy indexing, and subsetting of large data sets.
- It provides functionalities for reshaping and pivoting of data sets.

Q3. Enlist different types of Data Structures available in Pandas?

Different types of data structures available in Pandas are,

Series - It is immutable in size and homogeneous one-dimensional array data structure.

DataFrame - It is a tabular data structure which comprises of rows and columns. Here, data and size are mutable.

Panel - It is a three-dimensional data structure to store the data heterogeneously.

Q4. What is Series in Pandas?

Series is a one-dimensional array data structure that is capable of holding data of any type. In can be explained as the column in an excel sheet that has a series of data of one type. It is the simplest of data structure in Pandas where the axis labels of the data are called the index.

Q5. What is use of Reindexing in pandas?

Reindexing is done to change the row and column labels of the DataFrame. It conforms to the data to match a given set of labels along a particular axis. It is also done to insert the missing value marker in the label locations where no data exists.

Q6. What is DataFrame in Pandas?

DataFrame is a data structure in Pandas to store data as two-dimensional size-mutable and heterogeneous tabular data with labeled rows and columns. It is aligned as a tabular form in rows and columns. With this structure, you can perform an arithmetic operation on rows and columns. Here, each column of data will have the same data type.

Q7. What is pylab?

Pylab is a module in the **Matplotlib library** that acts as a procedural interface to the Matplotlib. Matplotlib is an object-oriented plotting library. It combines the Matplotlib with the NumPy module for graphical plotting. This is not a separate module but is embedded inside Matplotlib to provide matplotlib like experience for the user.

Q8. What is use of GroupBy objects in Pandas?

GroupBy is used to **split the data into groups**. It groups the data based on some criteria. Grouping also provides a mapping of labels to the group names. It has a lot of variations that can be defined with the parameters and makes the task of splitting the data quick and easy.

Q9. What is Pandas NumPy?

Pandas Numpy is an open-source library developed for Python that is used to work with a large number of datasets. It contains a powerful N-dimensional array object and sophisticated mathematical functions for scientific computing with Python.

Some of the popular functionalities present with Numpy are Fourier transforms, linear algebra, and random number capabilities. It also has tools for integrating with C/C++ and Fortran code.

Q10. What is Matplotlib?

Matplotlib is the most popular data visualization library that is used to plot the data. This comprehensive library is used for creating a static, animated, and interactive visualization with the data. It **Developed by John D. Hunter**, this open-source library was first **released in 2003**. Matplotlib also provides various toolkits that extend the functionalities of it. Such toolkits are **Basemap**, **Cartopy**, **Excel tool**, **GTK tools**, and more.

Q11. What is dataframe.iterrows in Pandas?

dataframe.iterrows() is used to iterate over a pandas Data frame rows in the form of (index, series) pair such that it iterates over the data frame column and return a tuple with the column name and content in form of series.

Q12. What is Vectorization in Python pandas?

Vectorization is the process of running operations on the entire array. This is done to reduce the amount of iteration performed by the functions. Pandas have a number of vectorized functions like aggregations, and string functions that are optimized to operate specifically on series and DataFrames. So it is preferred to use the vectorized pandas functions to execute the operations quickly.

Q13. List some alternatives of Python Pandas?

Some of the alternatives to the Python Pandas are

- the NumPy,
- R language,
- Anaconda,
- SciPy,
- PySpark,
- Dask,
- Pentaho Data, and Panda.

Q14. How to convert a DataFrame to an array in Pandas?

The function to_numpy() is used to **convert the DataFrame** to a **NumPy array**.

```
//syntaxDataFrame.to_numpy(self, dtype=None, copy=False)
```

The dtype parameter defines the data type to pass to the array and the copy ensures the returned value is not a view on another array.

Q15. List some statistical functions in Python Pandas?

Some of the statistical functions in Python Pandas are,

sum() - it returns the sum of the values.

mean() - returns the mean that is the average of the values.

std() - returns the standard deviation of the numerical columns.

min() - returns the minimum value.

max() - returns the maximum value.

abs() - returns the absolute value.

prod() - returns the product of the values.