FREQUENTLY ASKED QUESTIONS:

PYTHON: PANDAS

PART 1

FAQ

1) What is Pandas/Python pandas?

Pandas is defined as an open-source library that provides high-performance data manipulation in Python. The name of Pandas is derived from the word Panel Data, which means an Econometrics from Multidimensional data. It can be used for data analysis in Python and developed by Wes McKinney in 2008. It can perform five significant steps that are required for processing and analysis of data irrespective of the origin of the data, i.e., load, manipulate, prepare, model, and analyze.

2) What are the most important features of the pandas Library?

- Memory Efficient
- Data Alignment
- Reshaping
- Merge and join
- Time Series

3) Explain Categorical data in Pandas.

A Categorical data is defined as a Pandas data type that corresponds to a categorical variable in statistics. A categorical variable is generally used to take a limited and usually fixed number of possible values. Examples: gender, country affiliation, blood type, social class, observation time, or rating via Likert scales. All values of categorical data are either in categories or np.nan.

This data type is useful in the following cases:

It is useful for a string variable that consists of only a few different values. If we want to save some memory, we can convert a string variable to a categorical variable.
It is useful for the lexical order of a variable that is not the same as the logical order (?one?, ?two?, ?three?) By converting into a categorical and specify an order on the categories, sorting and min/max is responsible for using the logical order instead of the lexical order.

It is useful as a signal to other Python libraries because this column should be treated as a categorical variable.

FAQ

4) How could someone iterate over a Pandas DataFrame?

You can iterate over the rows of the DataFrame by using for loop in combination with an iterrows() call on the DataFrame.

5) Could we sort the DataFrame?

We can efficiently perform sorting in the DataFrame through different kinds:

- By label
- By Actual value

- ### By label

The DataFrame can be sorted by using the sort_index() method. It can be done by passing the axis arguments and the order of sorting. The sorting is done on row labels in ascending order by default.

- ### By Actual Value

It is another kind through which sorting can be performed in the DataFrame. Like index sorting, sort_values() is a method for sorting the values.

It also provides a feature in which we can specify the column name of the DataFrame with which values are to be sorted. It is done by passing the 'by' argument.

6) What are the different types of Data Structures in Pandas?

Pandas provide two data structures, which are supported by the pandas library, Series, and DataFrames. Both of these data structures are built on top of the NumPy.

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

FAQ

7) Reindexing in pandas?

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.

8) What is Data Aggregation?

The main task of Data Aggregation is to apply some aggregation to one or more columns. It uses the following:

- sum: It is used to return the sum of the values for the requested axis.
- min: It is used to return a minimum of the values for the requested axis.
- max: It is used to return a maximum values for the requested axis.

9) Explain Time Series in Pandas.

The Time series data is an essential source for information that provides a strategy that is used in various businesses. From a conventional finance industry to the education industry, it consists of a lot of details about the time.

Time series forecasting is the machine learning modeling that deals with the Time Series data for predicting future values through Time Series modeling.

10) Explain Time Offset and Time Periods.

- The offset specifies a set of dates that conform to the DateOffset. We can create the DateOffsets to move the dates forward to valid dates.
- The Time Periods represent the time span, e.g., days, years, quarter or month, etc. It is defined as a class that allows us to convert the frequency to the periods.

FAQ

11) What is a Series in Pandas?

A Series is defined as a one-dimensional array that is capable of storing 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.

12) What is the name of Pandas library tools used to create a scatter plot matrix?

Scatter_matrix

13) Explain DataFrame in Pandas.

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.

14) How to add an Index, row, or column to a Pandas DataFrame?

- Adding an Index to a DataFrame

Pandas allow adding the inputs to the index argument if you create a DataFrame. It will make sure that you have the desired index. If you don?t specify inputs, the DataFrame contains, by default, a numerically valued index that starts with 0 and ends on the last row of the DataFrame.

FAQ

- Adding Rows to a DataFrame

We can use .loc, iloc, and ix to insert the rows in the DataFrame.

- The loc basically works for the labels of our index. It can be understood as if we insert in loc[4], which means we are looking for that values of DataFrame that have an index labeled 4.

- The iloc basically works for the positions in the index. It can be understood as if we insert in iloc[4], which means we are looking for the values of DataFrame that are present at index '4`.

- The ix is a complex case because if the index is integer-based, we pass a label to ix. The ix[4] means that we are looking in the DataFrame for those values that have an index labeled 4. However, if the index is not only integer-based, ix will deal with the positions as iloc.

- Adding Columns to a DataFrame

If we want to add the column to the DataFrame, we can easily follow the same procedure as adding an index to the DataFrame by using loc or iloc.

15) What is Pandas NumPy array?

Numerical Python (Numpy) is defined as a Python package used for performing the various numerical computations and processing of the multidimensional and single-dimensional array elements. The calculations using Numpy arrays are faster than the normal Python array.

16) What is Pandas Index?

Pandas Index is defined as a vital tool that selects particular rows and columns of data from a DataFrame. Its task is to organize the data and to provide fast accessing of data. It can also be called a Subset Selection.

FAQ

17) How can you Reset the index?

The Reset index of the DataFrame is used to reset the index by using the 'reset_index' command. If the DataFrame has a MultiIndex, this method can remove one or more levels.

18) How do we convert DataFrame into an excel file?

We can export the DataFrame to the excel file by using the to_excel() function.

To write a single object to the excel file, we have to specify the target file name. If we want to write to multiple sheets, we need to create an ExcelWriter object with target filename and also need to specify the sheet in the file in which we have to write.

19) How do you Rename the Index or Columns of a Pandas DataFrame?

You can use the .rename method to give different values to the columns or the index values of DataFrame.

20) Describe Data Operations in Pandas?

In Pandas, there are different useful data operations for DataFrame, which are as follows:

- Row and column selection

We can select any row and column of the DataFrame by passing the name of the rows and columns. When you select it from the DataFrame, it becomes one-dimensional and considered as Series.

- Filter Data

We can filter the data by providing some of the boolean expressions in DataFrame.

FAQ

- Null values

A Null value occurs when no data is provided to the items. The various columns may contain no values, which are usually represented as NaN.

FAQ