## 2. What is NumPy and Pandas?

**What is NumPy and Pandas? What is the importance of these libraries for Data Analysis?**

﻿

NumPy is a python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant.

﻿

It is an open source project and you can use it freely. NumPy stands for Numerical Python.

﻿

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, load, manipulate, prepare, model, and analyze.

﻿

﻿**Key words:** data manipulation, algebra, arrays.

Pandas provides high level data manipulation tools built on top of NumPy.

NumPy by itself is a fairly low-level tool, similar to MATLAB.

Pandas on the other hand provides rich time series functionality, data alignment, NA-friendly statistics, groupby, merge and join methods, and lots of other conveniences. It has become very popular in recent years in financial applications.

Both of them are two of the most used libraries in Data Science, ML and AI.

Both of them are used to save n number of lines of Codes.

Key Differences:

* Pandas provides us with some powerful objects like DataFrames and Series which are very useful for working with and analyzing data whereas numpy library which provides objects for multi-dimensional arrays, Pandas provides in-memory 2d table object called Dataframe.
* Pandas works when data is in Tabular Format…whereas Numpy works really well when data is Numeric.
* Numpy consumes less memory as compared to Pandas.
* Numpy performs real good when there are 50k/less rows whereas pandas works really well when there are around 500k/more rows.

**What is a Python NumPy?**

NumPy is a Python package which stands for ‘Numerical Python’. It is the core library for scientific computing, which contains a powerful n-dimensional array object, provide tools for integrating C, C++ etc. It is also useful in linear algebra, random number capability etc. NumPy array can also be used as an efficient multi-dimensional container for generic data.

**NumPy Array:**Numpy array is a powerful N-dimensional array object which is in the form of rows and columns. We can initialize numpy arrays from nested Python lists and access it elements. In order to perform these numpy operations, the next question which will come in your mind is:

**How do I install NumPy?**

To install Python NumPy, go to your command prompt and type “pip install numpy”. Once the installation is completed, go to your IDE (For example: PyCharm) and simply import it by typing: “import numpy as np”

Moving ahead in python numpy tutorial, let us understand what exactly is a multi-dimensional numPy array.

**What is Python Pandas?**

Pandas is used for data manipulation, analysis and cleaning. Python pandas is well suited for different kinds of data, such as:

* Tabular data with heterogeneously-typed columns
* Ordered and unordered time series data
* Arbitrary matrix data with row & column labels
* Unlabelled data
* Any other form of observational or statistical data sets

**How to install Pandas?**

To install Python Pandas, go to your command line/ terminal and type “pip install pandas” or else, if you have anaconda installed in your system, just type in “conda install pandas”. Once the installation is completed, go to your IDE (Jupyter, PyCharm etc.) and simply import it by typing: “import pandas as pd”

Moving ahead in Python pandas tutorial, let’s take a look at some of its operations:

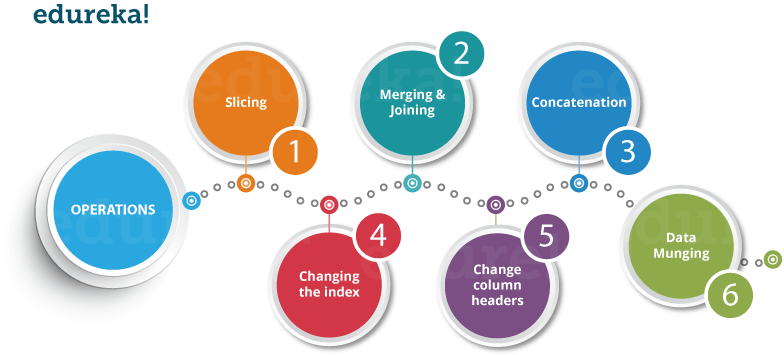
**Python Pandas Operations**

Using Python pandas, you can perform a lot of operations with series, data frames, missing data, group by etc. Some of the common operations for data manipulation are listed below:

What is Python Pandas?

Pandas is used for data manipulation, analysis and cleaning. Python pandas is well suited for different kinds of data, such as:

* Tabular data with heterogeneously-typed columns
* Ordered and unordered time series data
* Arbitrary matrix data with row & column labels
* Unlabelled data
* Any other form of observational or statistical data sets



**What is NumPy?**

NumPy stands for ‘Numerical Python’ or ‘Numeric Python’. It is an open source module of Python which provides fast mathematical computation on arrays and matrices. Since, arrays and matrices are an essential part of the Machine Learning ecosystem, NumPy along with Machine Learning modules like Scikit-learn, Pandas, Matplotlib, TensorFlow, etc. complete the Python Machine Learning Ecosystem.

NumPy provides the essential multi-dimensional array-oriented computing functionalities designed for high-level mathematical functions and scientific computation. Numpy can be imported into the notebook using

NumPy’s main object is the homogeneous multidimensional array. It is a table with same type elements, i.e, integers or string or characters (homogeneous), usually integers. In NumPy, dimensions are called axes. The number of axes is called the rank.

There are several ways to create an array in NumPy like np.array, np.zeros, no.ones, etc. Each of them provides some flexibility.

Some of the important attributes of a NumPy object are:

1. **Ndim:** displays the dimension of the array
2. **Shape:** returns a tuple of integers indicating the size of the array
3. **Size:** returns the total number of elements in the NumPy array
4. **Dtype**: returns the type of elements in the array, i.e., int64, character
5. **Itemsize:** returns the size in bytes of each item
6. **Reshape**: Reshapes the NumPy array

NumPy array elements can be accessed using indexing. Below are some of the useful examples:

* A[2:5] will print items 2 to 4. Index in NumPy arrays starts from 0
* A[2::2] will print items 2 to end skipping 2 items
* A[::-1] will print the array in the reverse order
* A[1:] will print from row 1 to end

**What is Pandas?**

Itis an open-source, BSD-licensed library written in *Python* Language. *Pandas* provide high performance, fast, easy to use data structures and data analysis tools for manipulating numeric data and time series. *Pandas* is built on the *numpy* library and written in languages like *Python*, *Cython*, and *C*. In pandas, we can import data from various file formats like *JSON, SQL, Microsoft Excel,* etc.

Similar to NumPy, Pandas is one of the most widely used python libraries in data science. It provides high-performance, easy to use structures and data analysis tools. Unlike NumPy library which provides objects for multi-dimensional arrays, Pandas provides in-memory 2d table object called Dataframe. It is like a spreadsheet with column names and row labels.

Some commonly used data structures in pandas are:

1. **Series objects**: 1D array, similar to a column in a spreadsheet
2. **DataFrame objects:** 2D table, similar to a spreadsheet
3. **Panel objects:** Dictionary of DataFrames, similar to sheet in MS Excel

Pandas Series object is created using pd.Series function. Each row is provided with an index and by defaults is assigned numerical values starting from 0. Like NumPy, Pandas also provide the basic mathematical functionalities like addition, subtraction and conditional operations and broadcasting.

Pandas dataframe object represents a spreadsheet with cell values, column names, and row index labels. Dataframe can be visualized as dictionaries of Series. Dataframe rows and columns are simple and intuitive to access. Pandas also provide SQL-like functionality to filter, sort rows based on conditions.

## What is matplotlib?

Matplotlib is a 2d plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments. Matplotlib can be used in Python scripts, Python and IPython shell, Jupyter Notebook, web application servers and GUI toolkits.

matplotlib.pyplot is a collection of functions that make matplotlib work like MATLAB. Majority of plotting commands in pyplot have MATLAB analogs with similar arguments.

Günümüzde en popüler ve sıklıkla bahsedilen konular; veri analizi nam-ı diğer ‘Data Analysis’, Veri yönetimi veya manipülasyonunu, Yapay zeka, makine öğrenmesi veya derin öğrenme dir. Bunlar ile ilgili uygulamalarda kullanabildiğimiz ön önemli programlama dili tabi ki Python ve onun eşsiz iki kütüphanesi NumPy ve Numpy’dan doğma Pandas…

NumPy, bilimsel hesaplamalarda ve işlemlerde kullanılıyor. Bu paket ile birlikte N-Dimensiyonel arrayleri manipüle edebildiğiniz ve hızlıca yaratabildiğimiz, linear cebirden, Fourier Transformasyona kadar tamamen matematiğin gücünü kullanabildiğimiz eşsiz kütüphane. Mesela arrayleri broadcast edebiliyor, böylece orijinal array üzerinde canlı bir şekilde, başka bir değişkene kopyalamaya gerek kalmadan değişiklik yapabiliyorsunuz. Bu da memory problemini ortadan kaldırıyor.

Verileri irdelediğimiz, manipüle ettiğimiz önemli bir kısım olan Pandas’ı Microsoft Excel olarak da düşünebiliriz. Burada ek olarak web sitelerinden html tagleri okuyarak veri çekebiliyor, bunu istediğimiz platforma (sql, excel, csv vb.) aktarabiliyoruz. Tabi ki ‘merge, join ve concatenation’ gibi global işlemleri de rahatlıkla pandas’in içerisinde bulunan fonksiyonlarla yapılabiliyor, verileri istediğimiz gibi yönetebiliyoruz.

### Pandas vs NumPy Comparison Table

Let’s discuss the top comparison between Pandas vs NumPy:

|  |  |  |
| --- | --- | --- |
| **Point of Comparison** | **Pandas** | **NumPy** |
| **Data Object/ Building Block** | Main data object in pandas is a series. Series is equivalent to one-dimensional array, whereas other data object Data Frame is equivalent to ndarray. | Main data object in numpy is ndarray (n-dimensional array). |
| **Popular Industry Usage** | Pandas is popularly used for data analysis and visualization. | NumPy is popularly used for numerical calculations. |
| **Type of Data Supported** | Pandas provide support for working with tabular data- CSV, Excel etc. | NumPy by default support data in the form of arrays and matrix. |
| **Usage in Deep Learning and Machine Learning** | Pandas series and dataframes cannot be directly fed as input in these toolkits. | Toolkits for machine/deep learning like Tensorflow,scikit can only be fed using numpy arrays. |
| **Performance** | Complex operations can make the overall process slow on pandas data objects. Pandas performs best with more amounts of data, say 500,000k or more rows. | Complex operations are faster on ndarrays. NumPy performs best with lesser amounts of data, say 50,000 or less rows. |
| **Indexing** | Data rows are by default indexed in pandas series and dataframes. | There is no default indexing of data rows in numpy arrays. |
| **Core Language** | Pandas uses R language as its reference language and hence provide many similar functions. | NumPy is written in the C programming language and hence uses multiple functionalities from it. |

NUMPY

Numpy, genel olarak bilimsel hesaplamalar için kullanılan bir python paketidir. Çok boyutlu dizileri kolayca ve hızlıca kullanmamızı sağlar. Daha çok makine öğrenmesi, derin öğrenme ve görüntü işleme gibi işlemlerde çok sık kullanılır. C++ tabanlı bir kütüphane olduğu içinde hız konusunda başarılıdır. Numpy ile oluşturulan diziler başlangıçta oluşan sabit boyutlu dizilerdir. Eğer daha sonra bu diziye bir şey eklemek istediğinizde eski diziyi silip yeni bir dizi oluşturur.

PANDAS

Pandas yüksek performanslı, kullanımı kolay veri yapılandırma ve veri analizi yapabileceğimiz bir kütüphanedir. Pandas ile excel, csv, json vs. dosyaları okuyup, yazabilirsiniz.

Pandas’daki dataframe ve seriler en önemli veri yapılarıdır. Dataframe nesnesi 2 boyutlu bir nesnedir. Index ve column olarak oluşur. Bunu aslında bir excel dosyası gibi düşünebilirsiniz. Index karşılığı satırlar, Column karşılığı ise sütunlardır.

Excel gibi düşünürsek burada nasıl filtreler kullanılıyorsa dataframelerde de kullanabiliriz.