Pandas is very popular for data science. It has been one of the most popular and favorite data science tools used for data wrangling and analysis.

Data is unavoidably messy in real world, and Pandas is seriously a game changer when it comes to cleaning, transforming, manipulating and analyzing data. Pandas helps to clean the mess.

Pandas is an open source Python library providing high-performance, easy-to-use data structures and data analysis tools. Pandas runs on top of NumPy.

Pandas provides …

- High level data structure (data frames).
- More streamlines handling of tabular data, and rich time series functionality.
- Data alignment, missing-data friendly statistics, groupby, merge and join methods.
- We can use Pandas data structures and freely draw on NumPy and SciPy functions to manipulate them.

Differences between the Python data science modules Pandas and Numpy

The *Pandas* module is used for working with tabular data. It allows us to work with data in table form, such as in CSV or SQL database formats. We can also create tables of our own, and edit or add columns or rows to tables. Pandas provides us with some powerful objects like DataFrames and Series which are very useful for working with and analyzing data.

The *Numpy* module is mainly used for working with numerical data. It provides us with a powerful object known as an Array. With Arrays, we can perform mathematical operations on multiple values in the Arrays at the same time, and also perform operations between different Arrays, similar to matrix operations.

When to use Pandas and Numpy

Pandas in general is used for financial time series data/economics data (it has a lot of built in helpers to handle financial data).

Numpy is a fast way to handle large arrays multidimensional arrays for scientific computing (scipy also helps). It also has easy handling for what are called sparse arrays (large arrays with very little data in them).

First, you need to install pandas package

On MacOS or Linux
pip3 install pandas

On Windows
pip install pandas

Quick Introduction on Pandas

```
Last login: Fri Jan 17 11:19:20 on ttys000
Sam MacOS: ipython
Python 3.7.2 (v3.7.2:9a3ffc0492, Dec 24 2018, 02:44:43)
Type 'copyright', 'credits' or 'license' for more information
IPython 7.11.1 -- An enhanced Interactive Python. Type '?' for help.

In [1]: import pandas as pd

In [2]: ds = pd.Series([2,4,6,8,10])

In [3]: print(ds)
0     2
1     4
2     6
3     8
4    10
dtype: int64

In [4]: print(ds.describe())
count     5.000000
mean      6.000000
std       3.162278
min       2.000000
25%       4.000000
50%       6.000000
75%       8.000000
max      10.000000
dtype: float64

In [5]: data = [1,2,3,4,5]

In [6]: df = pd.DataFrame(data)
```

```
In [7]: print(df)
   0
0  1
1  2
2  3
3  4
4  5

In [8]: data = [['Ana',21],['Bob',22],['Clarke',23]]

In [9]: df = pd.DataFrame(data,columns=['Name','Age'])

In [10]: print(df)
Name       Age
0     Ana   21
1     Bob   22
2  Clarke   23

In [11]: print(df.describe())
            Age
count   3.0
mean   22.0
std     1.0
min    21.0
25%    21.5
50%    22.0
75%    22.5
max    23.0

In [12]:
```

Pandas Documents

Panda Series: https://dzone.com/articles/python-pandas-tutorial-series-methods

Panda DataFrames: https://www.datacamp.com/community/tutorials/pandas-tutorial-dataframe-python