Why Use Pandas?

The recent success of machine learning algorithms is partly due to the huge amounts of data that we have available to train our algorithms on. However, when it comes to data, quantity is not the only thing that matters, the quality of your data is just as important. It often happens that large datasets don't come ready to be fed into your learning algorithms. More often than not, large datasets will often have missing values, outliers, incorrect values, etc… Having data with a lot of missing or bad values, for example, is not going to allow your machine learning algorithms to perform well. Therefore, one very important step in machine learning is to look at your data first and make sure it is well suited for your training algorithm by doing some basic data analysis. This is where Pandas come in. Pandas Series and DataFrames are designed for fast data analysis and manipulation, as well as being flexible and easy to use. Below are just a few features that makes Pandas an excellent package for data analysis:

- Allows the use of labels for rows and columns
- Can calculate rolling statistics on time series data
- Easy handling of NaN values
- Is able to load data of different formats into DataFrames
- Can join and merge different datasets together
- It integrates with NumPy and Matplotlib

For these and other reasons, Pandas DataFrames have become one of the most commonly used Pandas object for data analysis in Python.