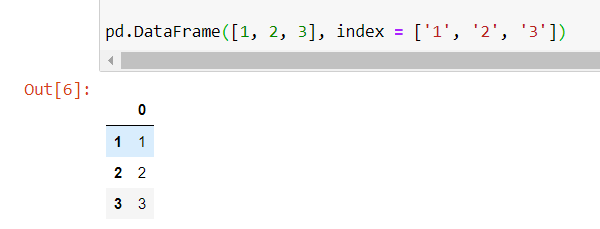
We know by now that data obtained “in the wild” will not always be neat and structured when encountered. We need to clean up said data before we can apply any data science techniques to it. One popular method of storing structured data is known as the data frame, and it is what spreadsheet programs (i.e., Microsoft Excel, Google Sheets) make easier to manipulate. I will be exploring what a data frame is, as well how they can be used.

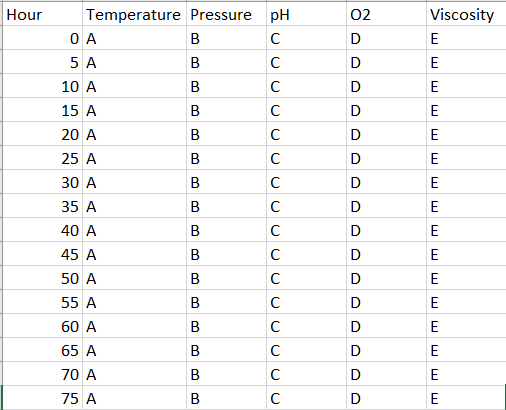
At its core, a data frame is a rectangular table of structured data. I like to imagine data frames as “data bookshelves.” Each series\*, or column, can contain a different data type and the overall frame can be indexed much like Python lists can. A data frame can be thought of as a dictionary of indices as keys and Pandasseries as values, and indeed, many manipulations applied to dictionaries and lists can be applied to data frames as well. We can add or remove series values, slice the frame, change the index names, or perform math. Going back to the bookshelf example, this would equate to adding/removing, separating, renaming, or rewriting the books, respectively.

Below is an example of a data frame in Python with indices assigned. This example only contains one series named “0”, but I could add more series and name them if I wanted to. If I wanted to use indexing to slice this data frame, I could use similar indexing as dictionaries to pull data: data\_frame[‘0’] would pull the ‘0’ series, for example, and data\_frame[‘0’][‘Index’] would pull the specified index from that series, just like Cartesian coordinates. Data frames also make graphing much quicker, as they allow for slicing, which frees up memory.



Those of us who frequently work with spreadsheet programs already know how data frames are used. Those who do not, however, need to know if they are to become successful data scientists. It is no exaggeration that data frames are a prerequisite to almost all data analyses out there, such as EDA or training a machine learning model. The reason? They reduce the amount of time spent in locating certain data while keeping scripts clear and concise, much like a bookshelf reduces the time to find a book. For example, if I had access to hospital patient data, I would use a data frame to organize the patient data by name or by another attribute. This would allow me to quickly filter through patients to pull a certain few, or plot visualizations with only a few lines of Python code. Without a data frame, these processes would take much longer.

At my job, I use data frames every day, although I hadn’t realized this until I began my Master of Science. I run bioreactors that display collected data as an Excel spreadsheet, one form of data frame. Data frames allow me to speed up my analysis by performing four steps: 1) Apply the necessary mathematical formulas, 2) Slice the data into its series, 3) Plot the sliced data over time, and 4) Analyze any trends. Below is an example of what my data frames might look like. If I wanted to get to step 4 without a data frame, I would have to put in way more time organizing all my points to be able to graph them, which would strain the rest of my time.



It is important to note that no matter what programming language is used to manipulate them, data frames remain an integral part of our data preparation step as data scientists. If there is an application where a data frame is not required, then chances are, it willmake any model construction or EDA that much easier.

I hope this has been an informative read about how data frames are applied in data science.

\*There is a slight difference between columns and series. Series are indexed columns generated by the Pandas library.

References:

Anthony, F. (2015). *Mastering Pandas.* Packt Publishing.

Sarkar, T. & Roychowdhury, S. (2019). *Data Wrangling with Python.* Packt Publishing.