Introduction to Pandas

"Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language"

https://pandas.pydata.org

How to get started?

In [ ]:

In [ ]:

Make sure you have the file "hour.csv" in the current directory. If not copy it here or go work there!

In [ ]:

Now let's read the data from the CSV file into a dataframe:

In [ ]:

What is a dataframe?

A dataframe is like an Excel spreadsheet within Python:

In [ ]:

It is a two-dimensional set of data, where the rows and columns can have labels. We can retrieve the data using these labels:

In [ ]:

Note that a colum of a dataframe is returned as a pandas series:

In [ ]:

A Pandas series is a one-dimensional data object with row labels.

When you import from a CSV file, the column labels are imported, but the row labels are just the numbers of the data rows:

In [ ]:

It is often convenient to use the values in one of the columns as the labels of the rows. We call these the *index* for the rows:

In [ ]:

Note that that is actually returning a new dataframe and the original dataframe is unchanged:

In [ ]:

If we wish to work with the original one, we have to replace it

In [ ]:

This makes indexing much easier

In [ ]:

Note that the row labels carry over to the Pandas series that is returned by indexing a particular column of the dataframe:

In [ ]:

In [ ]:

If all we want is the numerical values in the data series, we can convert it to a `numpy` array:

In [ ]:

Creating new Dataframes

From existent ones

Suppose that we want to create a dataframe with the columns: "temp", "atemp", "hum", "windspeed", "casual", "registered" and "cnt". We can create it this way:

In [ ]:

In [ ]:

In [ ]:

From numerical values

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

Visualization

`pandas` offers a wide range of plotting functions provided by the `matplotlib` library.

For example, to plot the feature "temp", you can:

In [ ]:

Alternatively, you can pass it directly to `matplotlib` functions:

In [ ]:

```
import matplotlib.pyplot as plt
%matplotlib inline
plt.style.use('seaborn-colorblind')
```

In [ ]:

`pandas` also includes a plotting module:

In [ ]:

In [ ]:

Summary Statistics

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

This covers some basics of working with Pandas dataframes and series, we can begin to work with real data in the next class.

More Resources

- Read chapter 3 "Data Manipulation with Pandas" from the book *Python Data Science Handbook* by Jake VanderPlas.
- Watch the video "pandas in 10 minutes" from the pandas *getting started* website
- Read "10 minutes to pandas" tutorial series provided in the User Guide documentation website
- Pandas cheat sheet: https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf

In [ ]: