tabular data with python Documentation

Release 1.0.1

Matt Hanson

Jun 26, 2018

Using tabular data in python (need to have)

This course is aimed at people who have completed the Enviromental Scientist's Introduction to Python Course. If You have not completed this course, we recommend at least reviewing the course to make sure you have comparable background in basic python. The goal of this course is to introduce tabular data manipulation in python and to get you to the point that you no longer need to use excel for your data analysis.

CHAPTER 1

Introduction to Numpy and Pandas

1.1 What in the world is a Numpy and Why are Pandas relevant

#todo brief history/description #todo how to import

1.2 The Numpy Array

the numpy array # indexing, slicing, boolean indexing # mention the huge number of methods

1.3 The Pandas Series and Dataframe

todo define # indexing/slicing/boolean indexing # define dataframe, indexing slicing, columns # mention the huge number of methods # showcase head/tail, index, columns

1.4 Vectorised operations (or why pandas and numpy are cool)

just explain and show off vecorized math

1.5 Datatypes

different datatypes (int, float, bool, string, object) include a nod to datetime object, but hold for later # changing data types with astype

1.6 Vectorised boolean operations

explain bracketing + [~,|,&] # float problems (e.g. floating point errors) np.isclose pandas version # in1d, pd.Series.isin

1.7 Missing data in pandas and Numpy

show missing data types nan, NaN, NaT, None # np.isnan, isfinite # pd.isnull, .notnull # .fillna

CHAPTER 2

Reading and Writing Data with Pandas

talk about how cool pandas read/write functionality is # make a bunch of files that people can download and look at.. pass urls

2.1 Reading text files

start with read_csv # showcase

- header
- names
- index_col
- skip_rows / nrows

then explore read_table start with identical arguments as above, but a couple more useful ones # showcase

-sep -delim_whitespace

explain that there are many more arguments for both of these functions

2.2 Reading from excel

simple example # showcase

- sheetname

2.3 Writing text files

just show to_csv

2.4 Other supported formats

copy this table: https://pandas.pydata.org/pandas-docs/stable/io.html

2.5 Dataframe cleanup

rename, set_index, reset_index, dropping columns,

CHAPTER 3

Working with Tabular Data

3.1 Using a Dataframe Like an Excel Sheet

excel sheet like math e.g. new columns ect # math # using boolean indexing to do math

3.2 Merging Manipulating and Mapping data

merging data # concatenation of dataframes

transpose

replace method

CHAPTER 4

Summarising Data in Pandas

4.1 Dataframe and Series summaries

show .min() style functionality # including arguments # showcase as table (function as link to detail, function description, ipython example?) # showcase the different output with dataframe and series

-all -any -min -max -mean -std -median -mode -quantile -describe -sum -unique -value_counts -count

just a drop in the bucket, more methods here: https://pandas.pydata.org/pandas-docs/stable/generated/pandas.Series.html # and here https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.html

4.2 Using Groupby to Aggregate Data

groupby object # builtin methods # the aggregate method

CHAPTER 5

Time Series Analysis with Pandas

I have no plan for this yet, todo # testing webhook

CHAPTER 6

Complex Data Re-arrangement and the Multi-index

6.1 The Pandas Multi-index

#todo # introduce the multi index # creation # slicing ect

6.2 Data Structure Magic: Melt and Pivot

todo # melt method # pivot, pivot_table, unstack

todo sting functionality in pandas series/ str methods # no other plans yet # str methods in general # strip # replace (explain that assumes regular expressions and link to resource) # split # indexing with .str # direct to https://pandas.pydata.org/pandas-docs/stable/text.html

CHAPTER 7

Using Pandas for Quick Data Visualisation

todo basic plot structure # some examples of plots: # plot # scatter # hist # boxplot

list everything avalible and link to https://pandas.pydata.org/pandas-docs/stable/visualization.html

CHAPTER 8

Python build for this course

If you are not familar with using virtual python enviroments, we reccomend you review our lesson on installing python. For This course we recommend installing python as follows:

This installation includes basic python, the packages numpy and pandas, and the IDE spyder. It also installs matplotlib, which is a pandas dependency if you are using the plotting functionality.

1. **Install miniconda (if you haven't already)**
    (a) go to https://conda.io/miniconda.html and download the appropriate python 3.6 installer and accept all of the defaults
2. **Create a virtual environment for this course (tdip) for Tabular Data In Python**
    (a) open anaconda prompt and enter:

```
conda create -n tdip python=3.6 spyder numpy pandas matplotlib
```

2. When conda asks you to proceed, type `y`:

```
proceed ([y]/n)?
```

3. That's it python and spyder for this course should now be installed. To use python with spyder, in the start menu (under anaconda) you should see spyder (tdip). Open that up and get cracking!

CHAPTER 9

Practice Exercises

We have developed a set of practice exercises to give you a taste of doing your own scripting. These exercises are facillitated through Github Classroom. If you want to get on to the exercises, this link will create a new repository with a copy of the exercises for you to begin working.

exercise	prerequisites / associated lessons
exercise 1	• *Introduction to Numpy and Pandas*
exercise 2	• *Reading and Writing Data with Pandas*
exercise 3	• *Working with Tabular Data*
exercise 4	• *Summarising Data in Pandas*
exercise 5	• *Time Series Analysis with Pandas*
exercise 6	• *Complex Data Re-arrangement and the Multi-index*