Introduction to data science

Patrick Shafto

Department of Math and Computer Science

Plan for today

- Projects
- HW 4
- Tidy data!
- HW for next week

Projects!

Pitches start Oct 3rd!

First round due Nov 18th by 11:59 pm

Will then get comments

Give presentation in class

Turn in final assignment Dec 14 11:59 pm

HW 4

- Write a short tutorial on pandas
- Cover basic functionality with worked examples
- Due Sunday by 11:59 pm
- Evaluations due Tuesday by 11:59pm
    - Use the rubric!
    - Justify your scores!
- Read http://vita.had.co.nz/papers/tidy-data.pdf

Feedback to Learner

Since panel is outdated and will be removed from future versions less coverage of the same would have sufficed. Apart from that good coverage of examples, have mentioned goals and references.... Nice work!!

Feedback to Learner

None

Feedback to Learner

I am satisfied with the content and examples you gave

Feedback to Learner

Very well documented tutorial covering the theoretical concepts along with sufficient number of examples. Formatting and overall presentation of the document could have been a little better.

Description for Pandas library is well structured. But very few basics have been covered in tutorials. There are some major functions needed to be implemented such as : importing the external dataset . As in further learning, we will use dataset and will perform function on that. So I find tutorial conceptually fine but missing lots of important functions.

Feedback to Learner

Good tutorial

Feedback to Learner

You have done a good work in this assignment. The examples used to explain the functions are good. Kindly add the sources of your material in the end of this tutorial.

Feedback to Learner

The tutorial is really good for the people who already know the basic concepts of pandas, it would have been better if at the start some basic concepts and syntax had been explained. Although the rest of the tutorial covers almost all topics and is complete. Do attach references.

Because there is a persistent problem with feedback, I will be going back and grading the evaluations.

You will not receive credit if you did not provide adequate feedback.

Tidy data?

- Wickham, 2014

Tidy data?

- It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data (Dasu and Johnson 2003).
- Data preparation is not just a first step, but must be repeated many over the course of analysis as new problems come to light or new data is collected.
- Despite the amount of time it takes, there has been surprisingly little research on how to clean data well.

Tidy data?

- Part of the challenge is the breadth of activities it encompasses:
  - from outlier checking,
  - to date parsing,
  - to missing value imputation.
- To get a handle on the problem, focus on a small, but important, aspect of data cleaning called data tidying: structuring datasets to facilitate analysis.

Tidy data?

- Two types of data:
  - Tidy data:
    - Variables on the columns, observations on the rows, each observational unit forms a table
  - Messy data:
    - Anything else!

HW for monday

- Replicate this analysis in your own jupiter notebook
- http://tomaugspurger.github.io/modern-5-tidy.html
- Add comments explaining what exactly the code is doing
- Stop at “Mini Project: Home Court Advantage?”