DataFrame Validation In Python

A Practical Introduction

Yotam Perkal • 04.06.18
Sounds Familiar?

https://www.youtube.com/watch?v=UXd0EDy7aTY
About Me
Motivation
Why do we need data validation?
Data Quality Dimensions

- Valid
- Accurate
- Complete
- Consistent
- Uniform
- Unique
It can happen to all of us
Perfect World

WELCOME TO UTOPIA

ENJOY THE RIDE
Perfect World

WELCOME TO UTOPIA
WE ARE CLOSED!
Model Deterioration
3 Accidental Discovery
Ignorance Is Bliss
Let’s See Some Tools?
Voluptuous

Voluptuous is a Python data validation library.

Benefits:
• Simplicity.
• Support for complex data structures.
• Useful error messages.

Similar to Json Schema but works on the data itself.

https://github.com/alecthomas/voluptuous
Engarde

Explicitly **state your assumptions** about the data and check that they're **actually** true.

Benefits:

- Great for flat files like csv
- Supports two methods of execution:
  - As decorators, which are most useful in .py scripts
  - Interactively at the interpreter

https://github.com/TomAugspurger/engarde
Test Driven Data Analysis

Applying Test Driven Development (TDD) principals to data analysis.

Benefits:

- Correctness
- Regression detection
- Specification, Design and Documentation
- Refactoring
- Portability

https://github.com/tdda/tdda
TDDA Engarde

Voluptuous
Credit - Practical Data Cleaning with Python

KATHARINE JARMUL

http://kjamistan.com/
“Quality is never an accident; it is always the result of intelligent effort.”

— John Ruskin