Learning Pandas - An Introduction

WE'LL COVER THE FOLLOWING

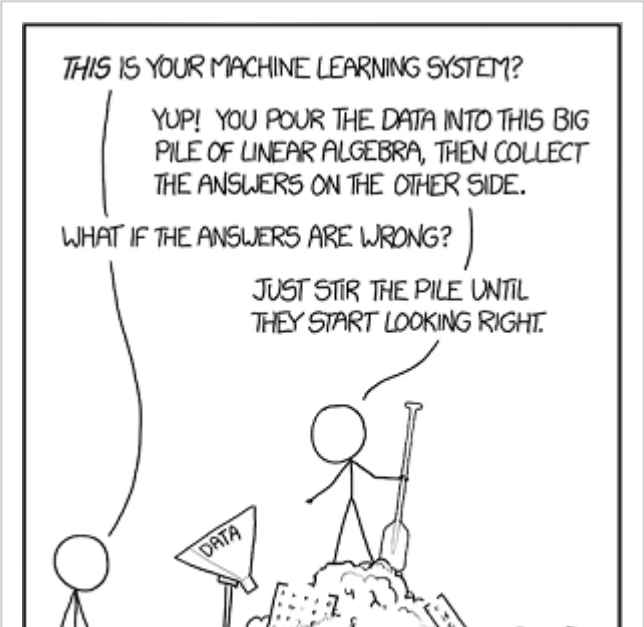
- Learning Data Manipulation With Pandas
- Lessons Overview

Learning Data Manipulation With Pandas

There are some ingredients, like salt, without which almost no dish is complete. Pandas is just like salt — some need more and some less, but almost every data science project needs it.

Pandas is a very powerful and popular package built on top of NumPy. It provides an efficient implementation of data objects built on NumPy arrays and many powerful data operations. These kind of operations are known as ***data wrangling*** — steps required to prepare the data so that it can actually be consumed for extracting insights and model building.

This might surprise you, but data preparation is what takes the longest in a data science project!



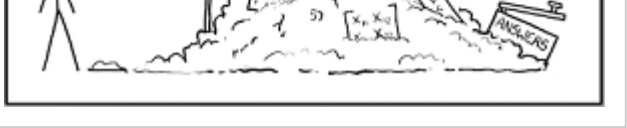


Image Source: https://xkcd.com/1838/

The two primary components of Pandas are the ***Series and DataFrame*** objects. A Series is essentially a column. And a DataFrame is a multi-dimensional table made up of a collection of Series; it can consist of heterogeneous data types and even contain missing data.

At the very basic level, Pandas objects can be thought of as enhanced versions of NumPy arrays in which the rows and columns are identified with ***labels*** instead of simple integer indices.

Lessons Overview

Pandas provides many useful tools and methods in addition to the basic data structures. These tools and methods require familiarity with the core data structures though, so we will start by understanding the nuts and bolts of Series and DataFrames. Then we will dive into all the good things that Pandas has to offer by analyzing some **real data** — be ready to explore the IMDB-movies dataset!