Microsoft: DAT210x Programming with Python for Data Science

Bookmarks

4. Transforming Data > Lecture: Data Cleansing > Intro

Bookmark

Data Cleansing

In data *wrangling*, irrelevant, incomplete, and missing data is either defaulted to a specific value or removed entirely. NaNs are stripped out, typographical errors are patched, and perhaps even some data normalization occurs. The goal of data *cleansing* is to take wrangling a step further by rectifying inaccurate and inconsistent data to standardize it. Inconsistent data can lead to false intelligence being produced by your machine learning algorithms, or no intelligence at all.

Simple data cleansing tasks might be automated and applied out of the box. More occupation specific tasks require you fully understand the working environment that generated your raw data. Knowledge of the range of values you expect to see for a particular feature will help you find any anomalies that need attention.

A classical example of when cleansing is necessary is when data comes from multiple sources. If, on average, a specific source consistently reports figures offset from others, identifying the source of the error, be it a faulty sensor, or bad reporting, etc., and then making calculated adjustments is a way to improve your overall data accuracy. But without carefully balancing keeping your data as close as possible to its raw from and making these error corrections, you might get accused of cooking your data. After all, it's always possible that there is no error at all.

Dive Deeper

- 5. Data Modeling