Introduction to Data Cleaning

This lesson will focus on why data cleaning is necessary and go into some methods to clean data.

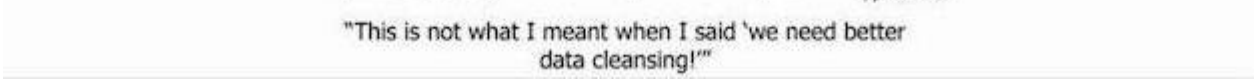
WE'LL COVER THE FOLLOWING

- Why clean data?
- Cleaning data

In this chapter, we will look at the third stage of the Data Science Lifecycle - *Data Cleaning*. But before we look at what steps are involved in Data Cleaning, a question arises; why do we need to clean data?

Why clean data?

The data that we receive and use is not perfect. Numerous factors such as data collection from multiple sources, or data corruption while storing or retrieving data, human errors in entering data, data loss while transferring data on some network, etc, can lead to incomplete, inconsistent, and incorrect data. If we use data as received in our analysis, then we will perform incorrect analysis and any conclusion drawn from the data will be wrong. Therefore, data cleaning is a necessary step before doing any analysis on the data.



Cartoon by Mark Anderson, www.andertoons.com.

Cleaning data

Data cleaning or **cleansing** is the process of detecting and correcting inconsistent, incorrect, and extraneous data. Data cleaning involves dealing with

- Missing data
- Duplicated data
- Outliers in the data
- Extra data that might not be needed
- Inconsistent data
- Converting data into a standard format so that it is easy to work on

We will look at all of these aspects in the upcoming lessons. But before that, we need to know *data types*. We will explore them in the next lesson.