My Philosophy of ‘Clean Data’

Ramesh Kattumenu

Data is always a ‘mess’.

The story being told all over is that 80% of the work of a data scientist is collecting, cleaning and organizing data and that only 20% is the real specialty. If you are looking to take up data science as a career, don’t be bogged down by the 80% statistic. You might be looking through a dirty lens. Though data cleaning is part of the process, it is the insights and predictions that can be made with the data that is the icing on the cake. One without the other is boring and mundane.

Data scientists have to take the time to examine the data before making predictions or building a model else they will find themselves redoing most of the work as bad data fields are discovered that need to be transformed before modeling. In the worst case, the model will be built that returns incorrect predictions which will make no sense. Data scientists are not wasting their time combing and cleaning through massive amounts of data.

A data scientist starts a project by:

1. Understanding the problem to be solved
2. Getting a high level overview of the data.

In the early stages, the data scientist makes guesses on the relationship between the data and pursues a direction to explore the data. Data is cleaned, variables are formed to validate the initial hypothesis. Sufficient knowledge about the data will be gained during this process. This triggers the next question to explore the data. The hypothesis will be reworked and the data will be modified to prove the research question.[[1]](#endnote-1),[[2]](#endnote-2) Cleaning data actually refers to filtering or transforming the data to suit the question we are trying to answer. Hence the majority of work for a Data Scientist is spent on cleaning the data by fixing issues related to missing data, misinterpreted data, incorrect data and inconsistent data.

The initial work of a data scientist follows a famous quote by Abraham Lincoln:

*“Give me six hours to chop down a tree and I will spend the first four sharpening the axe.”*

What exactly do we mean when we talk about cleaning data: There are a few important steps to make sure the data is good to answer the research question: 1) Loading data, 2) shaping data, 3) checking variable types, 4) managing bad values, 5) dealing with missing values, 6) creating variables to answer the research question, 7) formatting variables, 8) creating data visualizations.

In the words of Josh Wills, Director of Data Engineering at Slack, as told to Technology Review, Data scientists are ‘Data Janitors’. It is a little baffling but very flattering.

1. http://www.ruthstevens.com/wp-content/uploads/2014/05/WP\_Data\_Hygiene\_Pro\_3-06.pdf [↑](#endnote-ref-1)
2. http://winvector.github.io/DataPrep/EN-CNTNT-Whitepaper-Data-Prep-Using-R.pdf [↑](#endnote-ref-2)