Brace Yourself - Data Cleanup Is Coming

Undoubtedly, data is the cornerstone of any data analysis. As for data, there are millions of things that can misfire. It can be the arrangement, additional spaces, data format problems, duplicates – the list goes on. Therefore, data analysis can become your personal nightmare. Just think about it: data specialists spend up to 80% of their time organizing and cleansing data, whereas other 20% are allocated to data analysis itself. It’s quite a counter effective ratio, don’t you think?

(There exists an alternative joke: Data scientists spend up to 80% of their time organizing and cleaning data and 20% of their time – whining about it. We feel you. Data cleanup is a wild-goose chase.)

As you can see, proper data analytics calls for various data cleansing techniques so that your data is all set for analysis.

Anyway, What Is Data Cleaning?

Essentially, data cleaning or cleansing refers to the process of pinpointing and fixing or deleting incorrect records from a database. It also presupposes identifying unfinished or non-relevant parts of the data and then replacing, altering, or deleting the dirty or coarse data.

Although it may sound intimidating, it is not that painful in reality. Master a few techniques and it can go smoothly and fast.

5 Steps to Do Your Cleanup

1. A little planning never hurts.

   And by little we mean thorough and profound planning. You didn't think it was that easy?

   Instead of focusing on the final objective at the very beginning, chart out an actual plan. It should include the necessary degree of precision, formatting, the relevance of data itself. If you are still not sure, go for a pilot study first. If you've outlined the phases of your study, you can foretell what result you are getting. (Remember that guy-tapping-head meme?)

2. Actually Clean Your Data

   You'd be surprised to know that data cleanup is not about cleaning. It's more about being organized. Here's how to become a guru of data organizing:

   - Create separate worksheets for Raw Data, Currently Cleaning, Cleansed Data and Ready Data.
   - Get rid of the Invisible Man. Extra spaces lingering in your dataset looking arrogant and self-satisfied. Dump them.
   - Remove duplicates.
   - Standardize the case of your text data.
   - Do everything it takes to fix structural errors.

3. Look for one-off outliers.

   If you spot an outlier that doesn't fit within the analyzed data, make sure you delete it. However, not all unwanted outliers are irrelevant, sometimes they help to prove a theory you are working on.

4. Get hold of the missing data.

   Most algorithms do not recognize missing values. Therefore, missing data will affect the performance of your data analysis. You have two options there: either skip observations that feature missing data or input missing values relying on other observations. Both options are not ideal, yet worth trying.

5. Do basic validation.

   Once your data cleanup is done, make sure you go over the following questions:

   - Is all your data relevant?
   - Does the data go by the rules necessary for its field?
   - Does it prove or invalidate your working hypothesis, or unravel any insight?

   Although these questions may seem plain as the nose on your face, most people don't stop to mull over them.

You Cut-Out-'N-Keep Summary

Data sparseness and formatting inconsistencies are the biggest challenges in data analysis. Having clean data will ultimately boost overall productivity and allow for the superior quality information in your decision-making. Cleanse your data and you won't have to wade through countless outdated documents ever again.