Data Cleaning

Tushar B. Kute,
http://tusharkute.com

Data Cleaning

- When using data, most people agree that your insights and analysis are only as good as the data you are using. Essentially, garbage data in is garbage analysis out.
- Data cleaning, also referred to as data cleansing and data scrubbing, is one of the most important steps for your organization if you want to create a culture around quality data decision-making.

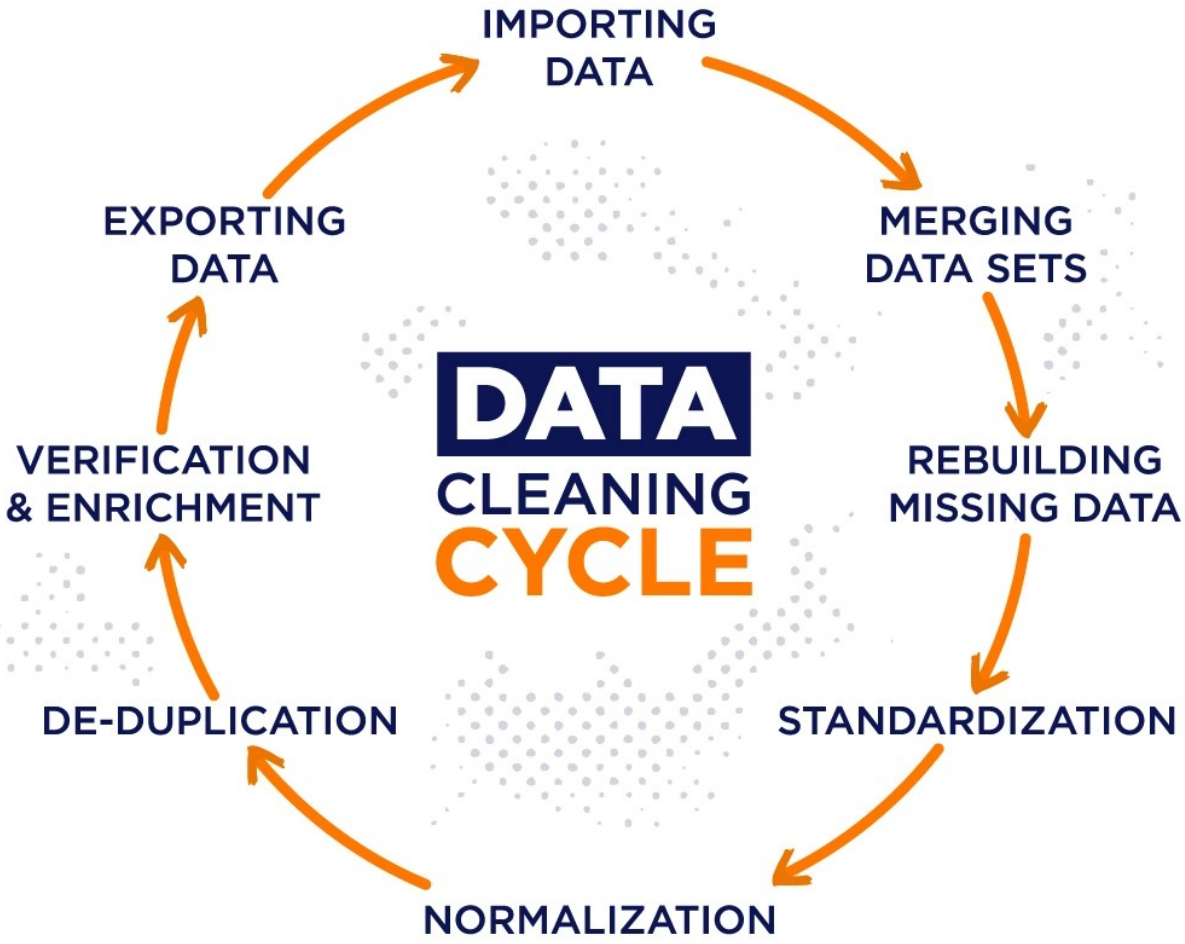
Data Cleaning

- Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.
- When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.
- If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct.
- There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset.
- But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

Data Cleaning vs. Transformation

- Data cleaning is the process that removes data that does not belong in your dataset. Data transformation is the process of converting data from one format or structure into another.
- Transformation processes can also be referred to as data wrangling, or data munging, transforming and mapping data from one "raw" data form into another format for warehousing and analyzing.

Steps


IMPORTING DATA
MERGING DATA SETS
REBUILDING MISSING DATA
STANDARDIZATION
NORMALIZATION
DE-DUPLICATION
VERIFICATION & ENRICHMENT
EXPORTING DATA
DATA CLEANING CYCLE

Steps

- While the techniques used for data cleaning may vary according to the types of data your company stores, you can follow these basic steps to map out a framework for your organization.
    - Step 1: Remove irrelevant data
    - Step 2: Deduplicate your data
    - Step 3: Fix structural errors
    - Step 4: Deal with missing data
    - Step 5: Filter out data outliers
    - Step 6: Validate your data

Remove irrelevant data

- First, you need to figure out what analyses you'll be running and what are your downstream needs.
- What questions do you want to answer or problems do you want to solve?
- Take a good look at your data and get an idea of what is relevant and what you may not need. Filter out data or observations that aren't relevant to your downstream needs.

Remove irrelevant data

- If you’re doing an analysis of SUV owners, for example, but your data set contains data on Sedan owners, this information is irrelevant to your needs and would only skew your results.
- You should also consider removing things like hashtags, URLs, emojis, HTML tags, etc., unless they are necessarily a part of your analysis.

Deduplicate your data

- If you’re collecting data from multiple sources or multiple departments, use scraped data for analysis, or have received multiple survey or client responses, you will often end up with data duplicates.
- Duplicate records slow down analysis and require more storage.
- Even more importantly, however, if you train a machine learning model on a dataset with duplicate results, the model will likely give more weight to the duplicates, depending on how many times they’ve been duplicated. So they need to be removed for well-balanced results.

Fix structural errors

- Structural errors include things like misspellings, incongruent naming conventions, improper capitalization, incorrect word use, etc.
- These can affect analysis because, while they may be obvious to humans, most machine learning applications wouldn’t recognize the mistakes and your analyses would be skewed.

Fix structural errors

- For example, if you're running an analysis on different data sets – one with a 'women' column and another with a 'female' column, you would have to standardize the title.
- Similarly things like dates, addresses, phone numbers, etc. need to be standardized, so that computers can understand them.

Deal with missing data

- Scan your data or run it through a cleaning program to locate missing cells, blank spaces in text, unanswered survey responses, etc.
- This could be due to incomplete data or human error. You'll need to determine whether everything connected to this missing data – an entire column or row, a whole survey, etc. – should be completely discarded, individual cells entered manually, or left as is.

Deal with missing data

- The best course of action to deal with missing data will depend on the analysis you want to do and how you plan to preprocess your data.
- Sometimes you can even restructure your data, so the missing values won't affect your analysis.

Filter out data outliers

- Outliers are data points that fall far outside of the norm and may skew your analysis too far in a certain direction.
- For example, if you're averaging a class's test scores and one student refuses to answer any of the questions, his/her 0% would have a big impact on the overall average.
- In this case, you should consider deleting this data point, altogether. This may give results that are "actually" much closer to the average.

Filter out data outliers

- However, just because a number is much smaller or larger than the other numbers you're analyzing, doesn't mean that the ultimate analysis will be inaccurate.
- Just because an outlier exists, doesn't mean that it shouldn't be considered.
- You'll have to consider what kind of analysis you're running and what effect removing or keeping an outlier will have on your results.

Validate your data

- Data validation is the final data cleaning technique used to authenticate your data and confirm that it’s high quality, consistent, and properly formatted for downstream processes.
  - Do you have enough data for your needs?
  - Is it uniformly formatted in a design or language that your analysis tools can work with?
  - Does your clean data immediately prove or disprove your theory before analysis?

Validate your data

- Validate that your data is regularly structured and sufficiently clean for your needs. Cross check corresponding data points and make sure nothing is missing or inaccurate.
- Machine learning and AI tools can be used to verify that your data is valid and ready to be put to use.
- And once you've gone through the proper data cleaning steps, you can use data wrangling techniques and tools to help automate the process.

Data Cleaning Tips

- Create the right process and use it consistently
  - Set up a data cleaning process that's right for your data, your needs, and the tools you'll use for analysis.
  - This is an iterative process, so once you have your specific steps and techniques in place, you'll need to follow them religiously for all subsequent data and analyses.

Data Cleaning Tips

- Use tools
  - There are a number of helpful data cleaning tools you can put to use to help the process – from free and basic, to advanced and machine learning augmented.
  - Do some research and find out what data cleaning tools are best for you.
  - If you know how to code, you can build models for your specific needs, but there are great tools even for non-coders.

Data Cleaning Tips

- Pay attention to errors and track where dirty data comes from
  - Track and annotate common errors and trends in your data, so you'll know what kinds of cleaning techniques you need to use on data from different sources.
  - This will save huge amounts of time and make your data even cleaner – especially when integrating with analysis tools you use regularly.

Summary

- It’s clear that data cleaning is a necessary, if slightly annoying, process when running any kind of data analysis.
- Follow the steps above and you’re well on your way to having data that’s fully prepped and ready for downstream processes.
- Remember to keep your processes consistent and don’t cut corners on data cleaning, so you’ll end up with accurate, real-world, immediately actionable results.

References

- www.mitu.co.in
- https://monkeylearn.com

Thank you

This presentation is created using LibreOffice Impress 5.1.6.2, can be used freely as per GNU General Public License

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com