Data Cleaning

Content

- What is Data Cleaning in Data Science?
- Why is Data Cleaning So Important?
- Data Cleaning Process [In 8 Steps]
- Data Cleaning Tools

What is Data Cleaning in Data Science?

- Data cleaning is the process of identifying and fixing incorrect data. It can be in incorrect format, duplicates, corrupt, inaccurate, incomplete, or irrelevant.

- Various fixes can be made to the data values representing incorrectness in the data.

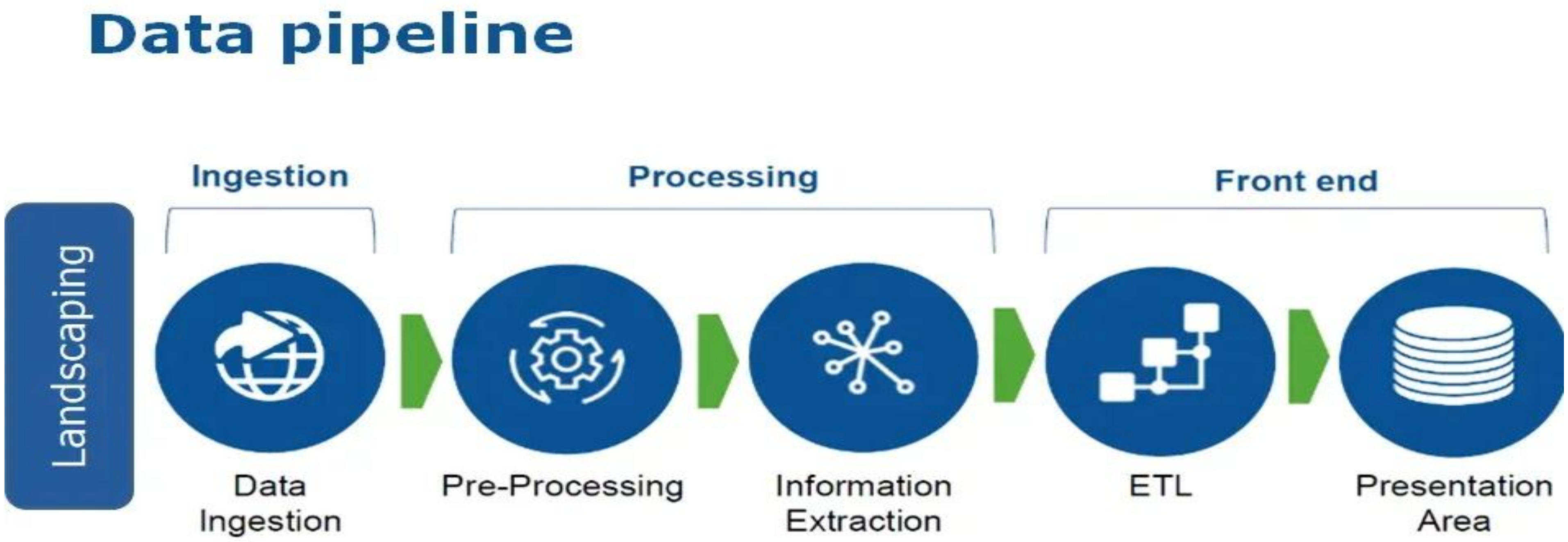
- The data cleaning and validation steps undertaken for any data science project are implemented using a data pipeline.

- Each stage in a data pipeline consumes input and produces output.

- The main advantage of the data pipeline is that each step is small, self-contained, and easier to check.

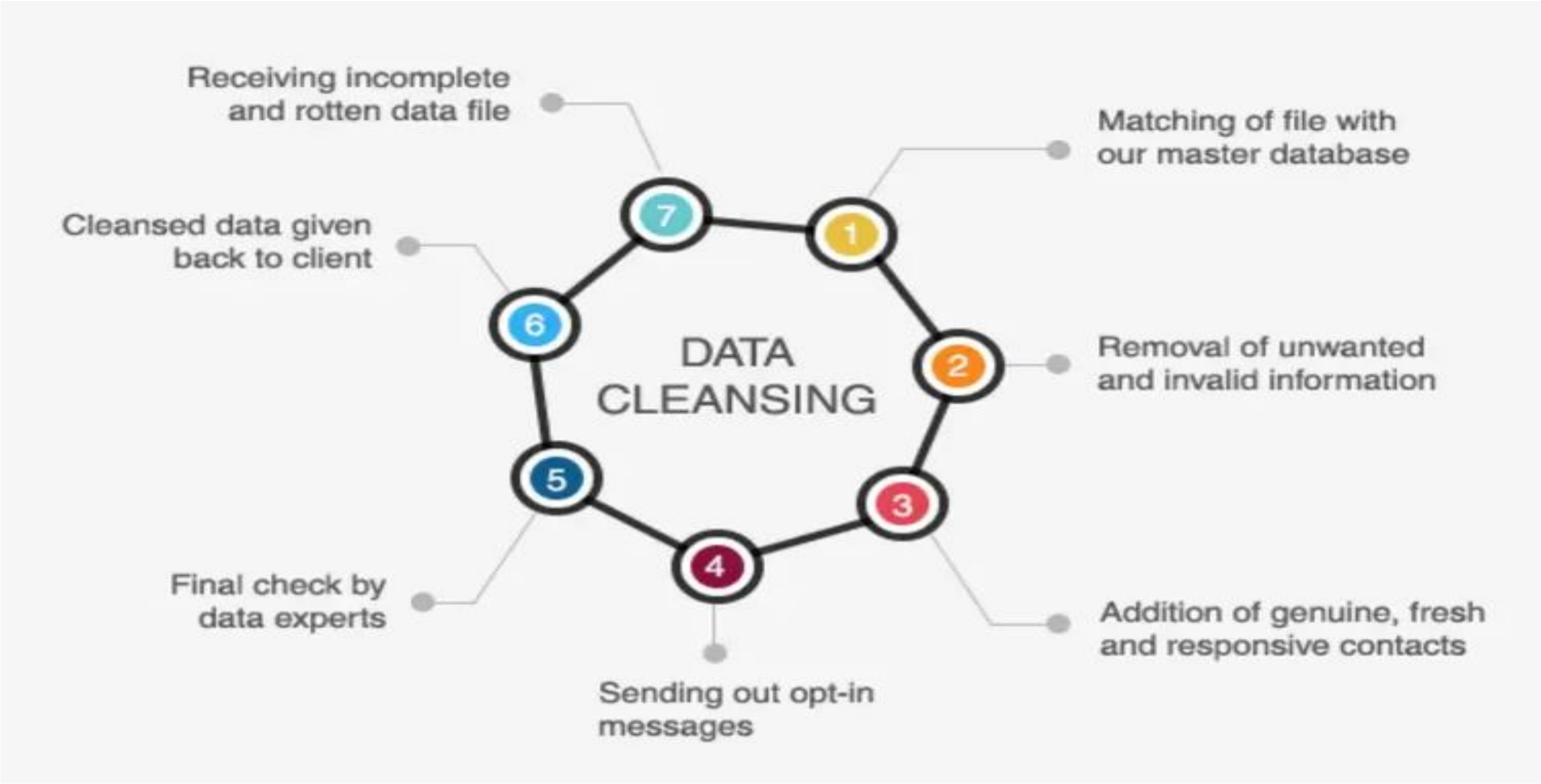
Data pipeline

A data pipeline is a method in which raw data is ingested from various data sources, transformed and then ported to a data store, such as a data lake or data warehouse, for analysis.



Why is Data Cleaning So Important?

Data cleaning is important because it ensures consistency within your data set and helps you achieve reliable results from any analysis you perform on it. Additionally, regularly checking for inconsistencies allows you to identify problems in your data sets before they become bigger issues down the line.


DATA CLEANSING
1
Matching of file with our master database
2
Removal of unwanted and invalid information
3
Addition of genuine, fresh and responsive contacts
4
Sending out opt-in messages
5
Final check by data experts
6
Cleansed data given back to client
7
Receiving incomplete and rotten data file

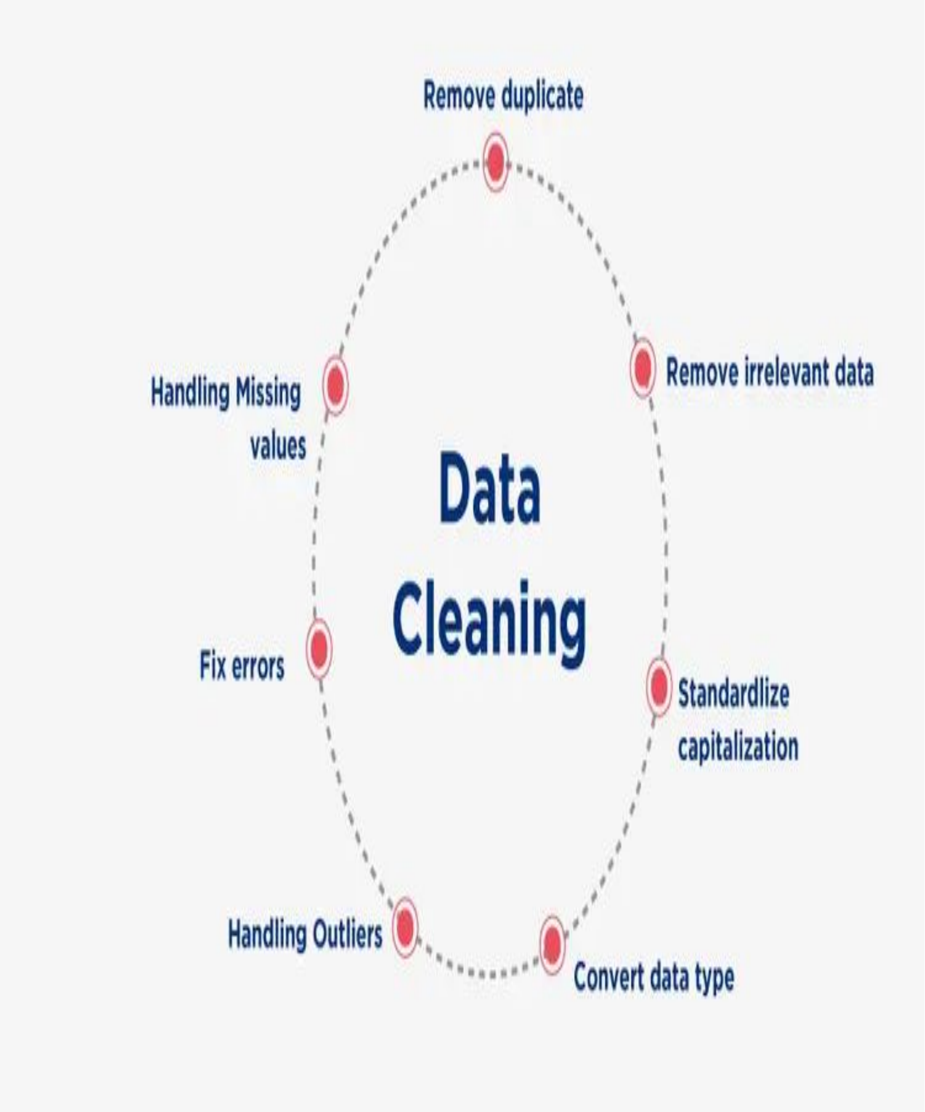
Data Cleaning Example:

- Consider data where we have the gender column. If the data is being filled manually, then there is a chance that the data column can contain records of ‘male’ ‘female’, ‘M’, ‘F’, ‘Male’, ‘Female’, ‘MALE’, ‘FEMALE’, etc. In such cases, while we perform analysis on the columns, all these values will be considered distinct. But in reality, ‘Male’, ‘M’, ‘male’, and ‘MALE’ refer to the same information. The data cleaning step will identify such incorrect formats and fix them.

Data Cleaning Process [In 8 Steps]

Eight common steps in the data cleaning process, as mentioned below:

1. Removing duplicates
2. Remove irrelevant data
3. Standardize capitalization
4. Convert data type
5. Handling outliers
6. Fix errors
7. Language Translation
8. Handle missing values



Step 1: Remove Duplicates

- When you are working with large datasets, working across multiple data sources, or have not implemented any quality checks before adding an entry, your data will likely show duplicated values.
- These duplicated values add redundancy to your data and can make your calculations go wrong. Duplicate serial numbers of products in a dataset will give you a higher count of products than the actual numbers.
- Duplicate email IDs or mobile numbers might cause your communication to look more like spam. We take care of these duplicate records by keeping just one occurrence of any unique observation in our data.

Step 2: Remove Irrelevant Data

- Consider you are analyzing the after-sales service of a product. You get data that contains various fields like service request date, unique service request number, product serial number, product type, product purchase date, etc.
- While these fields seem to be relevant, the data may also contain other fields like attended by (name of the person who initiated the service request), location of the service center, customer contact details, etc., which might not serve our purpose if we were to analyze the expected period for a product to undergo servicing. In such cases, we remove those fields irrelevant to our scope of work. This is the column-level check we perform initially.
- Next comes the row-level checks. Assume the customer visited the service center and was asked to visit again after 3 days to collect the serviced product. In this case, let us also assume that there are two different records in the data representing the same service number.
- For the first record, the service type is ‘first visit’ and the service type is ‘pickup’ for the second record. Since both records represent the same service request number so we will likely drop one of them. For our problem statement, we need the first occurrence of the record or the ones which correspond to the service type as ‘first visit’.
- To remove irrelevant data while cleaning data for effective data science, we must understand the data and the problem statement.

Step 3: Standardize capitalization

- You must ensure that the text in your data is consistent. If your capitalization is inconsistent, it could result in the creation of many false categories.

- **For example:** having column name as “Total_Sales” and “total_sales” is different (most programming languages are case-sensitive).

- To avoid confusion and maintain uniformity among the column names, we should follow a standardized way of providing the column names. The most preferred code case is the snake case or cobra case.

- Cobra case is a writing style in which the first letter of each word is written in uppercase, and each space is substituted by the underscore (_) character. While, in the snake case, the first letter of each word is written in lowercase and each space is substituted by the underscore. Therefore, the column name “Total Sales” can be written as “Total_Sales” in the cobra case and “total_sales” in the snake case. Along with the column names, the capitalization of the data points should also be fixed.

Step 4: Convert data type

- When working with CSV data in python, pandas will attempt to guess the types for us; for the most part, it succeeds, but occasionally we'll need to provide a little assistance.
- The most common data types that we find in the data are text, numeric, and date data types. The text data types can accept any kind of mixed values including alphabets, digits, or even special characters. A person’s name, type of product, store location, email ID, password, etc., are some examples of text data types.
- Numeric data types contain integer values or decimal point numbers, also called float. Having a numeric data type column means you can perform mathematical computations like finding the minimum, maximum, average, and median, or analyzing the distribution using histogram, box plot, q-q plot, etc.
- Ex: The date column can be represented in different formats:
  - October 02, 2023
  - 02-10-2023
  - 2023/10/02
  - 2-Oct-2023

Step 5: Handling Outliers

- An outlier is a data point in statistics that dramatically deviates from other observations. An outlier may reflect measurement variability, or it may point to an experimental error; the latter is occasionally removed from the data set.
- **For example:** let us consider pizza prices in a region. The pizza prizes vary between INR 100 to INR 7500 in the region after surveying around 500 restaurants. But after analysis, we found that there is just one record in the dataset with the pizza price as INR 7500, while the rest of the other pizza prices are between INR 100 to INR 1500.
- There are two common ways to deal with these outliers.
  - Remove the observations that consist of outlier values.
  - Apply transformations like a log, square root, etc., to make the data values follow the normal or near-normal distribution.

Step 6: Fix errors

- Errors in your data can lead you to miss out on the key findings. This needs to be avoided by fixing the errors that your data might have.
- Systems that manually input data without any provision for data checks are almost always going to contain errors.
- To fix them, we need to first get the data understanding. Post that, we can define logic or check the data and accordingly get the data errors fixed.

Step 7: Language Translation

- Datasets for machine translation are frequently combined from several sources, which can result in linguistic discrepancies. Software used to evaluate data typically uses monolingual Natural Language Processing (NLP) models, which are unable to process more than one language. Therefore, you must translate everything into a single language. There are a few language translational AI models that we can use for the task.

Step 8: Handle missing values

- During cleaning and munging in data science, handling missing values is one of the most common tasks. The real-life data might contain missing values which need a fix before the data can be used for analysis.
- We can handle missing values by:
  - Either removing the records that have missing values or
  - Filling the missing values using some statistical technique or by gathering data understanding.

Data Cleaning Tools & Software

- **Sigma AI - Input Tables:** Sigma's AI product, Input Tables, can clean, classify, extract, and autofill table data effortlessly. This tool is particularly useful for those already using Sigma's platform for data analysis.

- **OpenRefine:** Formerly known as Google Refine, OpenRefine is a powerful open-source tool for working with messy data. It allows users to clean, transform, and extend data with web services and external data sources.

- **WinPure:** WinPure is an affordable data cleaning tool that can handle large datasets, remove duplicates, as well as correct and standardize data. It supports various data sources, including databases, spreadsheets, and CRMs.

- **Melissa Clean Suite:** Melissa Clean Suite is a data cleaning solution that enhances data quality in CRM and ERP platforms like Oracle CRM, Salesforce, Oracle ERP, and Microsoft Dynamics CRM. It offers features such as data deduplication, data verification, contact autocompletion, data enrichment, and real-time and batch processing.

- **Trifacta Wrangler:** Trifacta Wrangler is a data cleaning tool that helps users explore, clean, and prepare data for analysis. It offers features like data profiling, transformation, and validation, making it easier to work with messy data

Benefits of Data Cleaning in Data Science

Your analysis will be reliable and free of bias if you have a clean and correct data collection. We have looked at eight steps for data cleansing in data science. Let us discuss some of the benefits of cleaning data science.

- **Avoiding mistakes:** Your analysis results will be accurate and consistent if data cleansing techniques are effective.
- **Improving productivity:** Maintaining data quality and enabling more precise analytics that support the overall decision-making process are made possible by cleaning the data.
- **Avoiding unnecessary costs and errors:** Correcting faulty or mistaken data in the future is made easier by keeping track of errors and improving reporting to determine where errors originate.
- **Staying organized**
- **Improved mapping**

Data Cleaning vs Data Transformation

Data Cleaning	Data Transformation
Data cleaning refers to the process of identifying errors, anomalies, and inconsistencies in the dataset.	Data transformation refers to formatting, restructuring, and modifying the original data to a more suitable or recommended format.
Data cleaning aims to improve the quality of the data.	Data transformation aims to transform the data that is suitable for analysis.
Certain kinds of inconsistencies in the data cannot be handled by all data tools.	Data transformation can be done with almost any data tool.
Inadequate data cleansing results in skewed and inconsistent analysis and insights from the data.	Untransformed data may be hard to understand or visualize, which makes it difficult for stakeholders to draw conclusions that are significant.
Imputing missing value or removing duplicates are some examples of common data cleaning strategies.	Modifying data types, pivoting, or reshaping data for analysis are some examples of common data transformation operations.