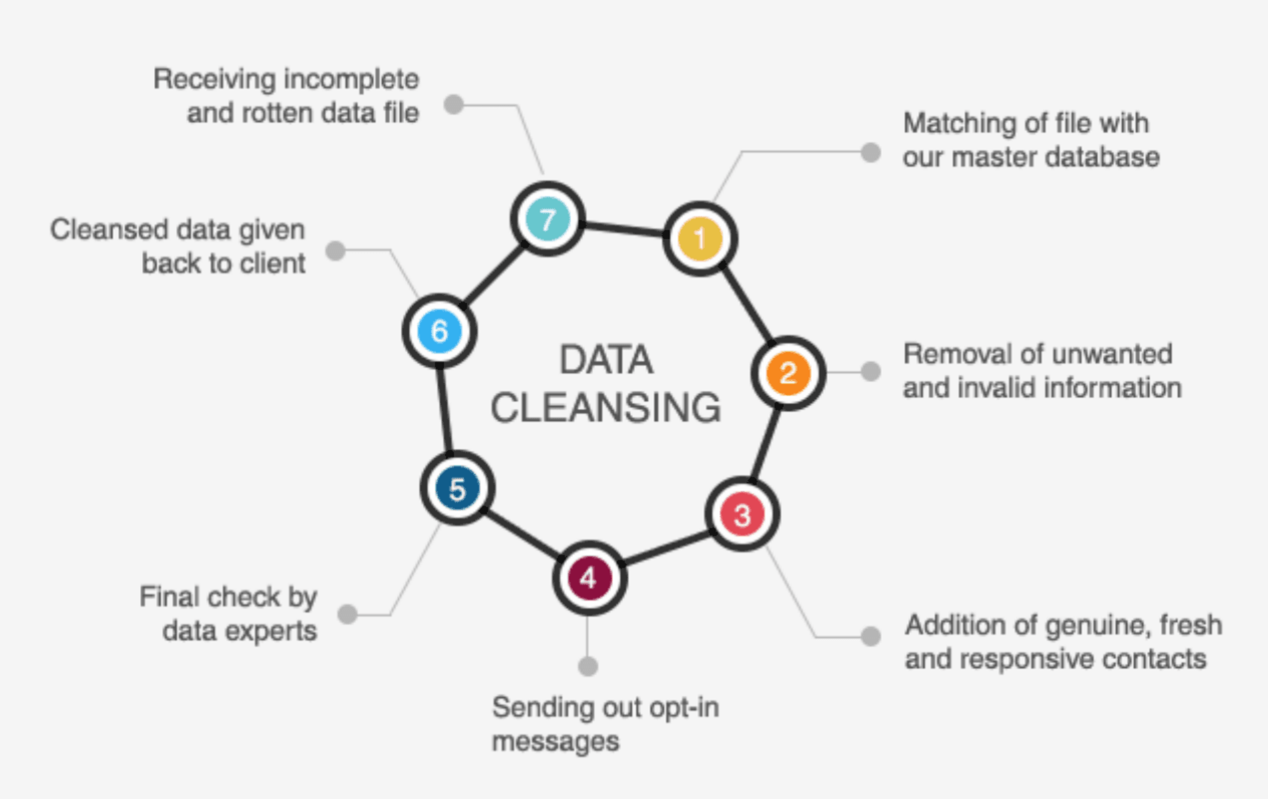
What is Data Cleaning in Data Science?

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. Some data pipeline systems also allow you to resume the pipeline from the middle, thus, saving time. In this article, we will look at 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

Why is Data Cleaning So Important?


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

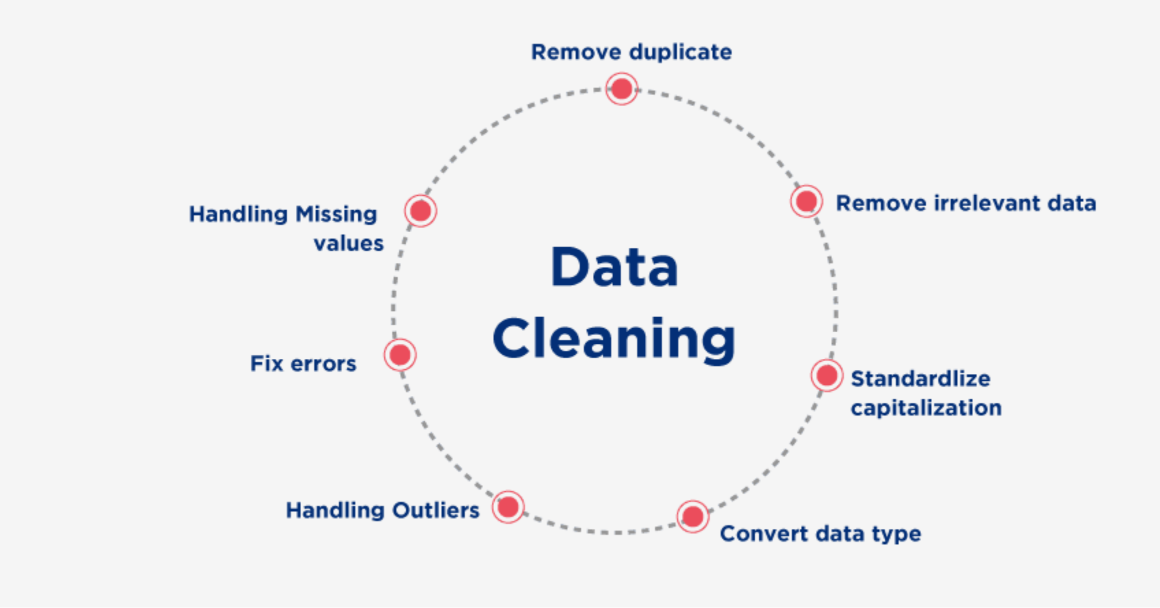
As an experienced **Data Scientist**, I have hardly seen any perfect data. Real-world data is noisy and contains a lot of errors. They are not in their best format. So, it becomes important that these data points need to be fixed.

It is estimated that data scientists spend between 80 to 90 percent of their time in data cleaning. Your workflow should start with data cleaning. You may likely duplicate or incorrectly classify data while working with large datasets and merging several data sources. Your algorithms and results will lose their accuracy if you have wrong or incomplete data.

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.

Consider another example where you are running a promotional campaign and you have collected data from different individuals. The data you collected contains the name of the individual, along with contact number, email, age, gender, etc. If you were to contact these individuals by mobile number or email, you must make sure that these are valid entries. For contact number, it should be a 10-digit numeric field and the email should follow a defined pattern. There can also be entries where you might not have either or both contact information. You would like to drop these entries since they are irrelevant and do not serve any purpose. This is again identified and fixed during data cleansing in data science before using it for our analysis or other purposes.

Data Cleaning Process [In 8 Steps]



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.

For example: while collecting names and email IDs through online forms, surveys, or other means, we can get inputs in assorted styles. We can fix them to avoid duplicate entries getting ignored. The email Id's 'myemail@hostname.com' and 'MYEMAIL@HOSTNAME.COM' can be interpreted as different email IDs, so it is better to make all the email ID values in the field lowercase. Similarly, for the email, we can follow the title case where all words are capitalized.

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.

Having a numeric column as an integer column will not allow you to perform this numerical analysis. Therefore, it becomes important to convert the data types in the required formats if they are not already.

The monthly sales figures of a store, the price of a product, units of electricity consumed, etc., are examples of a numeric column. However, it is worth noting that columns like a numeric ID or phone number should not be represented as numeric columns but instead as text columns. Though they represent numeric values, operations like minimum or average values on these columns do not provide any significant information. Therefore, these columns should be represented as text columns.

The data type if not identified correctly will end up being identified as a string or text column. In such cases, we need to explicitly define the data type of the column and the date format which is mentioned in the data. 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. Therefore, the observation with pizza price as INR 7500 is an outlier since it significantly deviates from the population. These outliers are usually identified using a box plot or scatter plot. These outliers result in skewed data. There are models which assume the data to follow a normal distribution, and the outliers can affect the model performance if the data is skewed thus, these outliers must be handled before the data is fed for model training. 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, box-cox, etc., to make the data values follow the normal or near-normal distribution.

You can learn about these methods and other data cleaning or wrangling skills with **Bootcamp for Data Science**. Develop your programming and analytical abilities as you gain confidence as a data scientist under the direction of expert professionals with six capstone projects and over 280 hours of on-demand self-paced learning.

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. Consider the following example cases.

- Removing the country code from the mobile field so that all the values are exactly 10 digits.
- Remove any unit mentioned in columns like weight, height, etc. to make it a numeric field.
- Identifying any incorrect data format like email address and then either fixing it or removing it.
- Making some validation checks like customer purchase date should be greater than the manufacturing date, the total amount should be equal to the sum of the other amounts, any punctuation or special characters found in a field that does not allow it, etc.

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.

A rule of thumb is that you can drop the missing values if they make up for less than five percent of the total number of records but however it depends on the analysis, the importance of the missing values, the size of the data, and the use case we are working on.

Consider a dataset that contains certain health parameters like glucose, blood pressure, insulin, BMI, age, diabetes, etc. The goal is to create a supervised classification model that predicts if the person is likely to have diabetes or not based on the health parameters. If the data has missing values for glucose and blood pressure columns for a few individuals, there is no way we can fill these values through any technique. And assuming these two columns are of high importance in predicting the presence of diabetes in an individual then we must look to drop these observations from our records.

Consider another dataset where we have information about the laborers working on a construction site. If the gender column in this dataset has around 30 percent missing values. We cannot drop 30 percent of data observations but on further digging, we found that among the rest 70 percent of observations and 90 percent of records are male. Therefore, we can choose to fill these missing values as the male gender. By doing this, we have made an assumption, but it can be a safe assumption because the laborers working on the construction site are male dominant and even the data suggests the same. We have used a measure of central tendency called Mode, in this case. There are also other ways of filling missing values in a numerical field by using Mean or Median values based on whether the field values follow a gaussian distribution or not.

Data Cleaning Tools

1. Excel

Excel's user-friendly design and range of features make it a popular tool for data cleaning and processing. It has several options that include data formatting and standardization, converting data types, data validation, text manipulation, removing duplicates, etc. Excel is the simplest tool that you can start with as most of us are aware of it. However, Excel can only handle simpler datasets due to which use of data cleaning in Excel is not generally preferred for larger datasets. Also, with increasing complexity, it is not the right choice to utilise Excel for your data cleaning needs. It is a recommended tool if you don't have a large amount of data and basic to intermediate complexities to resolve.

2. Python

Python and data analytic processes go hand in hand. Data cleaning in data science is mostly preferred using python if you are familiar with coding. Python provides a multitude of tools and libraries that cover many facets of data cleaning, transformation, and analysis, offering a broad variety of functionality to effectively clean and preprocess data. The particular requirements and type of data being processed are major factors in library selection. Libraries such as Pandas, NumPy, Seaborn, Matplotlib, Dask, Tabulate, Regex, etc. are some of the popular ones that help in data cleaning. You can also automate your data cleansing strategy using python for your application that will take care of any new incoming data.

3. SQL

The database programming language is called Structured Query Language (SQL). SQL queries can be used to retrieve filtered data from the databases. A Database Management System houses the data for the majority of applications (DBMS). As a result, it is a useful tool for source-level data management. Although it can carry out simple data-cleaning operations, it breaks down when dealing with complicated data. When it comes to data visualizations, it is ineffective since most database management systems cannot produce visualizations, which are an excellent method to address specific data issues. Some of the common data cleaning operation for which we can rely on SQL are data standardization, data type check, removing duplicates, transforming data, handling missing values, etc.

4. Tableau

With Tableau, a popular application for data visualization, you can make interactive dashboards for any kind of work. You can alter the charts, graphs, local and global filters, formulas, and more. We can carry out simple data-cleaning procedures before building the visuals. Tableau Prep Builder, a specialized program that handles data cleaning, is available if we choose to work with more intricate data cleaning procedures. The Tableau app can then utilize this cleansed data directly because of the seamless integration between the two applications. One of Tableau Prep Builder's advantages is that most cleaning chores can be completed with just a few clicks thanks to its user-friendly interface. Because of this, it's quite helpful to non-programmers. You can head over to **their website** where they have provided a comprehensive guide on data cleaning best practices.

5. OpenRefine

OpenRefine is an open-source data transformation
and data cleaning software that was originally known as Google Refine. Its purpose is to preprocess and clean up data that is not clean. It provides a broad range of features to clean, normalize, and transform datasets together with an intuitive user interface. Unlike other tools discussed so far, this tool is dedicatedly created for data cleaning operations. Hence, it can handle large and unorganised datasets. It can extract data from different sources, perform analysis, transform and clean the data. When working with very large datasets, OpenRefine's speed may suffer, using a lot of memory and slowing down processing for huge datasets. Despite of such limitations, it still remains a valuable tool to work with data cleaning tasks.

6. Trifacta

Recently, Trifacta, a commercial software solution, teamed with Alteryx to give its consumers advanced data analytic insights with the least amount of time and effort. Trifacta is one of the best data cleaning software when it comes to enterprise-grade solutions. This low-code/no-code platform's main goal is to give consumers access to cloud infrastructure so they can handle their big data analytics requirements. Trifacta supports collaboration by allowing users to share cleaning data pipelines and work together on the same dataset. Many of the leading analytics companies use the program for their analytical needs. They offer frequent upgrades to the application that include new beneficial features for its users and have a responsive support staff. You can use their free trial to check out the features of the application.

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.

Data Cleaning Tools & Software

There are several data cleaning tools and software available that can help streamline the process of cleaning and preparing data for analysis. Here are some popular options:

- **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.