What is data cleansing in data science

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.

Data Cleaning Steps & Techniques

1. Step 1: Remove irrelevant data.
2. Step 2: Deduplicate your data.
3. Step 3: Fix structural errors.
4. Step 4: Deal with missing data.
5. Step 5: Filter out data outliers.
6. Step 6: Validate your data.

Step 1: Remove duplicate or irrelevant observations

Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations. Duplicate observations will happen most often during data collection. When you combine data sets from multiple places, scrape data, or receive data from clients or multiple departments, there are opportunities to create duplicate data. De-duplication is one of the largest areas to be considered in this process. Irrelevant observations are when you notice observations that do not fit into the specific problem you are trying to analyze. For example, if you want to analyze data regarding millennial customers, but your dataset includes older generations, you might remove those irrelevant observations. This can make analysis more efficient and minimize distraction from your primary target—as well as creating a more manageable and more performant dataset.

Step 2: Fix structural errors

Structural errors are when you measure or transfer data and notice strange naming conventions, typos, or incorrect capitalization. These inconsistencies can cause mislabeled categories or classes. For example, you may find “N/A” and “Not Applicable” both appear, but they should be analyzed as the same category.

Step 3: Filter unwanted outliers

Often, there will be one-off observations where, at a glance, they do not appear to fit within the data you are analyzing. If you have a legitimate reason to remove an outlier, like improper data-entry, doing so will help the performance of the data you are working with. However, sometimes it is the appearance of an outlier that will prove a theory you are working on. Remember: just because an outlier exists, doesn’t mean it is incorrect. This step is needed to determine the validity of that number. If an outlier proves to be irrelevant for analysis or is a mistake, consider removing it.

Step 4: Handle missing data

You can’t ignore missing data because many algorithms will not accept missing values. There are a couple of ways to deal with missing data. Neither is optimal, but both can be considered.

1. As a first option, you can drop observations that have missing values, but doing this will drop or lose information, so be mindful of this before you remove it.
2. As a second option, you can input missing values based on other observations; again, there is an opportunity to lose integrity of the data because you may be operating from assumptions and not actual observations.
3. As a third option, you might alter the way the data is used to effectively navigate null values.

Step 5: Validate and QA

At the end of the data cleaning process, you should be able to answer these questions as a part of basic validation:

- Does the data make sense?
- Does the data follow the appropriate rules for its field?
- Does it prove or disprove your working theory, or bring any insight to light?
- Can you find trends in the data to help you form your next theory?
- If not, is that because of a data quality issue?

False conclusions because of incorrect or “dirty” data can inform poor business strategy and decision-making. False conclusions can lead to an embarrassing moment in a reporting meeting when you realize your data doesn’t stand up to scrutiny. Before you get there, it is important to create a culture of quality data in your organization. To do this, you should document the tools you might use to create this culture and what data quality means to you.

What is data integration :

Data integration is the process of combining data from multiple sources into a cohesive and consistent view. This process involves identifying and accessing the different data sources, mapping the data to a common format, and reconciling any inconsistencies or discrepancies between the sources. The goal of data integration is to make it easier to access and analyze data that is spread across multiple systems or platforms, in order to gain a more complete and accurate understanding of the data.

Data integration can be challenging due to the variety of data formats, structures, and semantics used by different data sources. Different data sources may use different data types, naming conventions, and schemas, making it difficult to combine the data into a single view. Data integration typically involves a combination of manual and automated processes, including data profiling, data mapping, data transformation, and data reconciliation.

Data integration is used in a wide range of applications, such as business intelligence, data warehousing, master data management, and analytics. Data integration can be critical to the success of these applications, as it enables organizations to access and analyze data that is spread across different systems, departments,

and lines of business, in order to make better decisions, improve operational efficiency, and gain a competitive advantage.

There are mainly 2 major approaches for data integration – one is the “tight coupling approach” and another is the “loose coupling approach”.

Tight Coupling:

This approach involves creating a centralized repository or data warehouse to store the integrated data. The data is extracted from various sources, transformed and loaded into a data warehouse. Data is integrated in a tightly coupled manner, meaning that the data is integrated at a high level, such as at the level of the entire dataset or schema. This approach is also known as data warehousing, and it enables data consistency and integrity, but it can be inflexible and difficult to change or update.

- Here, a data warehouse is treated as an information retrieval component.
- In this coupling, data is combined from different sources into a single physical location through the process of ETL – Extraction, Transformation, and Loading.

Loose Coupling:

This approach involves integrating data at the lowest level, such as at the level of individual data elements or records. Data is integrated in a loosely coupled manner, meaning that the data is integrated at a low level, and it allows data to be integrated without having to create a central repository or data warehouse. This approach is also known as data federation, and it enables data flexibility and easy updates, but it can be difficult to maintain consistency and integrity across multiple data sources.

What is Data Transformation?

Data transformation is the mutation of data characteristics to improve access or storage. Transformation may occur on the format, structure, or values of data. With regard to data analytics, transformation usually occurs after data is *extracted* or *loaded* (E**T**L/EL**T**).

Data transformation increases the efficiency of analytic processes and enables data-driven decisions. Raw data is often difficult to analyze and too vast in quantity to derive meaningful insight, hence the need for clean, usable data.

During the transformation process, an analyst or engineer will determine the data structure. The most common types of data transformation are:

- **Constructive:** The data transformation process adds, copies, or replicates data.
- **Destructive:** The system deletes fields or records.
- **Aesthetic:** The transformation standardizes the data to meet requirements or parameters.
- **Structural:** The database is reorganized by renaming, moving, or combining columns.

What Is Data Science?

Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. Data science uses complex machine learning algorithms to build predictive models.

The data used for analysis can come from many different sources and presented in various formats.

Now that you know what data science is, let’s see why data science is essential to today’s IT landscape.

What Does a Data Scientist Do?

You know what is data science, and you must be wondering what exactly is this job role like - here's the answer. A data scientist analyzes business data to extract meaningful insights. In other words, a data scientist solves business problems through a series of steps, including:

- Before tackling the data collection and analysis, the data scientist determines the problem by asking the right questions and gaining understanding.
- The data scientist then determines the correct set of variables and data sets.
- The data scientist gathers structured and unstructured data from many disparate sources—enterprise data, public data, etc.
- Once the data is collected, the data scientist processes the raw data and converts it into a format suitable for analysis. This involves cleaning and validating the data to guarantee uniformity, completeness, and accuracy.
- After the data has been rendered into a usable form, it’s fed into the analytic system—ML algorithm or a statistical model. This is where the data scientists analyze and identify patterns and trends.

- When the data has been completely rendered, the data scientist interprets the data to find opportunities and solutions.
- The data scientists finish the task by preparing the results and insights to share with the appropriate stakeholders and communicating the results.

Now we should be aware of some machine learning algorithms which are beneficial in understanding data science clearly.