**DATA CLEANING**



**What is 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.

**Step 1: Remove duplicate 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.

**Step 2: 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 3: Filter outliers**

In simple terms, an outlier is**an extremely high or extremely low data point relative to the nearest data point and the rest of the neighboring co-existing values in a data graph or dataset you're working with.**

**Step 4: Data Type conversion**

The type conversion is an operation that takes a data object of one type and creates the equivalent data objects of multiple types.

**Step 5: Data Transformation**

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. This article focuses on the processes of cleaning that data.

**Data cleaning involved in my project**

1. Filter outliers
2. Type conversion
3. Data transformation

**Advantages and benefits of data cleaning**

Having clean data will ultimately increase overall productivity and allow for the highest quality information in your decision-making. Benefits include:

* Removal of errors when multiple sources of data are at play.
* Fewer errors make for happier clients and less-frustrated employees.
* Ability to map the different functions and what your data is intended to do.
* Monitoring errors and better reporting to see where errors are coming from, making it easier to fix incorrect or corrupt data for future applications.
* Using tools for data cleaning will make for more efficient business practices and quicker decision-making.