SCRUBBING YOUR DATA CLEAN.

Scrubbing data, also known as cleaning data, is an important step in the data analysis process. It involves transforming raw, messy data into clean and usable data. The scrubbing process includes tasks like removing duplicates, formatting records, solving for missing values, and checking for mistakes or wrong values. By scrubbing data, you ensure that it is in good shape for further analysis, avoiding errors that could affect your conclusions.

Scrubbing data

- Removing duplicates
- Formatting records
- Solving for missing values
- Checking for wrong values

REMOVING DUPLICATE RECORDS

The first step in the data scrubbing process, which is removing duplicates. Duplicates are when there are multiple entries of the same data in a dataset. This can happen due to human error or machine errors. For example, when you scan a product at a self-checkout machine in a supermarket, the machine might accidentally read the barcode twice, creating a duplicate entry in the supermarket's database.

Removing duplicates is important because they can distort the analysis of the data. It's like counting the same thing multiple times, which can lead to incorrect conclusions. To remove duplicates, we first need to identify which records are duplicates. This can be done manually by inspecting the records one by one or using tools like Excel that can automatically find duplicates. Once identified, the duplicate entries can be deleted manually or programmatically using a script.

Before removing duplicates, it's important to make sure that the repeated data is actually a duplicate and not missing information. For example, if two records have the same first and last name, but including the middle name reveals that they are different, then they are not duplicates. Once the duplicates are removed, we can move on to the next step in the data scrubbing process.

Removing duplicates

1. Identify the duplicates
2. Remove the duplicates

Possible reasons for duplicates

- Entering the same info multiple times
- Accidentally copying data when saving or moving files
- Machine related errors, e.g. malfunctioning

FORMAT YOUR RECORDS

In data analytics, formatting your data consistently is important. This means making sure that all the data in your dataset follows the same format. If the data is not formatted consistently, it can lead to confusion and inaccurate results when analyzing the data.

Let's look at an example to understand this better. Imagine you have a table with different records of New York City, but the city name is recorded in different ways like NYC, New York, NY, or Manhattan NY. To accurately summarize the data and find the most common location, you would want to make sure that all these records have the same name. Otherwise, the data analytics tools may treat them as separate locations, leading to incorrect results.

Formatting your data consistently also applies to other aspects, like converting different currencies to a single currency or ensuring that data fields have the correct formatting rules for their type (text, numbers, dates, etc.). By maintaining consistency and using the right formatting, you can accurately analyze and explore your data to gain meaningful insights.

Formatting Records

- Ensure consistency
- Identify the data type

Handle Missing Values

In data analytics, one important step is to check if there are any missing values in the data. Missing values are quite common and can occur for various reasons, such as not knowing the value of a variable or errors in recording the data. It's important to find and address these missing values so they don't affect our analysis later on.

There are two options to deal with missing values. The first option is to fill in the missing values with an indicator that shows the value is not available. For example, if we don't have location data for some people, we can fill the missing values with "unknown" or "N/A". This way, we still keep the other data in those records that could be useful for analysis.

The second option is to delete the entire record that contains the missing value. However, this can result in losing valuable information and may introduce bias in the data. It's generally better to fill in the missing values rather than deleting records.

Taking care of missing values before starting the analysis is important. Later on, when working with spreadsheets, you'll learn how to detect and fill in missing values easily.

Dealing with missing values

- Fill in the missing value
- Delete the record with the missing value

Checking for wrong values

checking for obviously wrong values in our data. When we have a dataset, it's important to make sure that the values in it make sense and are within the expected range. For example, if we have a dataset of people's ages, we know that a person's age cannot realistically be over 150. So, any entries in the dataset that are higher than that would be flagged as wrong or invalid.

To identify obviously wrong values, we need to understand the context in which the data was collected and the expected range of values. Sometimes, what may seem like a wrong value at first glance may actually be correct when we consider the context. For example, in a dataset of sales from a clothing store, we might see negative values for price. Initially, we might think these are wrong because stores don't have items with negative prices. But if we know that the dataset also includes returns, it makes sense for the price to be negative in those cases.

Once we have identified obviously wrong values, we have two options. We can either replace the incorrect value with a word that indicates there was an error or that the value is not available, or we can delete the entire record that contains the wrong value. It's usually best to replace the incorrect value because it saves other valuable data points in the record. Deleting the record means losing all the other information in it.

By checking for obviously wrong values and treating them appropriately, we can ensure that our dataset is clean and ready for further analysis.

Treating wrong values

- Replace the wrong value
- Delete the record containing the wrong value

Summary: Scrubbing data

Scrubbing Checklist

The scrubbing stage is all about cleaning your data and getting your dataset ready for analysis. You can use this checklist to help you in the process.

1. Removing Duplicates

- ☐ **Identifying duplicate records**: *inspect records for duplicates and verify that they are actually a duplicate record.*
- ☐ **Remove duplicate records:** *remove the duplicate records from your dataset*

2. Formatting records

- ☐ **Ensure consistency**: *check all data follow a consistent format and adjust the format if necessary*
- ☐ **Identify the data type**: *make sure the data type is clear and identified*

3. Solving for missing values

- ☐ **Identify the missing values**: *Scan your data for any values that may be missing*
- ☐ **Solve for the missing values**: *Replace the missing values with text (e.g. NA) or delete the entire record with the missing value*

4. Checking for wrong values

- ☐ **Identify wrong values**: *Scan your data for any wrong values*
- ☐ **Solve for the wrong values**: *Replace the wrong values with the correct ones if you can or delete the entire record with the wrong values*