Data Cleaning

Dr : Noha Sakr

•••

Hazem Mamdouh
Bavly Adel
Menna Magdy
Andre Hany
Mariam Awad ALLah

Things you need to know

- Data Can be defined as a collection of individual facts like : names, phone numbers, marks, pictures and so on.....
- Data in the real world is noisy and incomplete.
- In the world of data processing, there is a wise saying that
- (garbage in garbage out), It means your insights are only as good as the data you're using to get them.
- Inaccurate data leads you to inaccurate conclusions and bad decisions, so we need to the Data Cleaning Process to produce clean data.

Properties of clean data:

1- Validity : The degree to which your data conforms to defined business rules or constraints.

2- Accuracy : Ensures that your data is close to the true values.

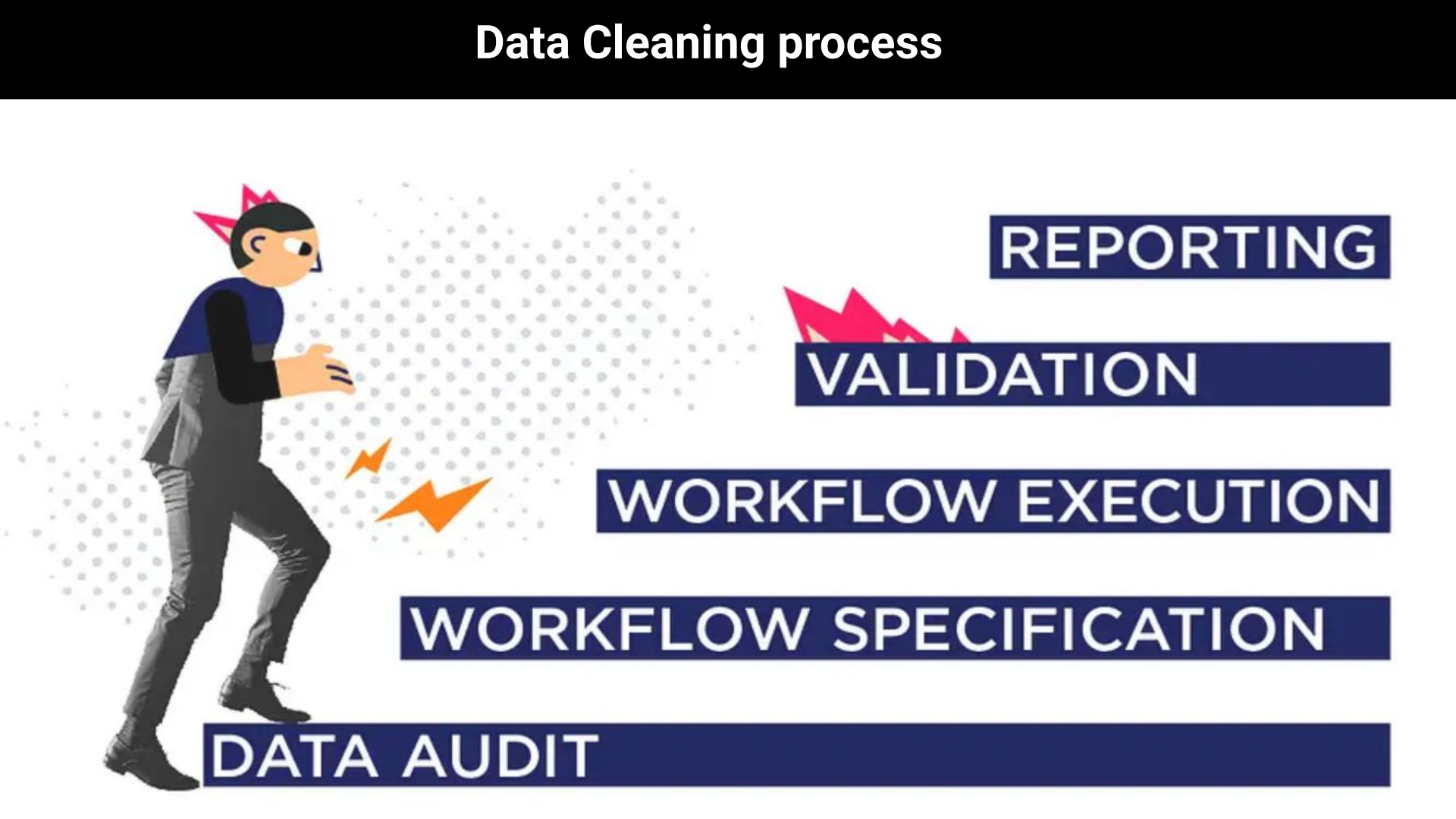
3- Completeness : The degree to which all required data is known.

4- Uniformity : Related to the consistency of the units of measure in all systems, e.g., data sets coming from the US and Germany might use different units of weight (pounds vs. kilos)

5- Timeliness : related to how up-to-date a piece of information is.

6- Relevancy : The degree of “usefulness” of data, determining how closely a piece of information is related to an issue you are researching.

Data Cleaning process



1. Data Audit

Any data cleaning process starts with taking a close look at your data. You have to determine what kind of errors your data set contains and where they're located.

We can do that through the use of statistical and database methods that help you detect anomalies and contradictions like:

Data profiling: is a technique that helps to examine the data and create general, informative reports about what's in the data set. This method might not be very in-depth but gives you a good initial idea of the types of data you're dealing with.

Software packages: allow you to set a particular type of constraint and then generate code that checks the dataset for errors based on the violation of those constraints. Software packages can also generate reports of what constraints have been violated, how many times, and create a visualization of those findings.

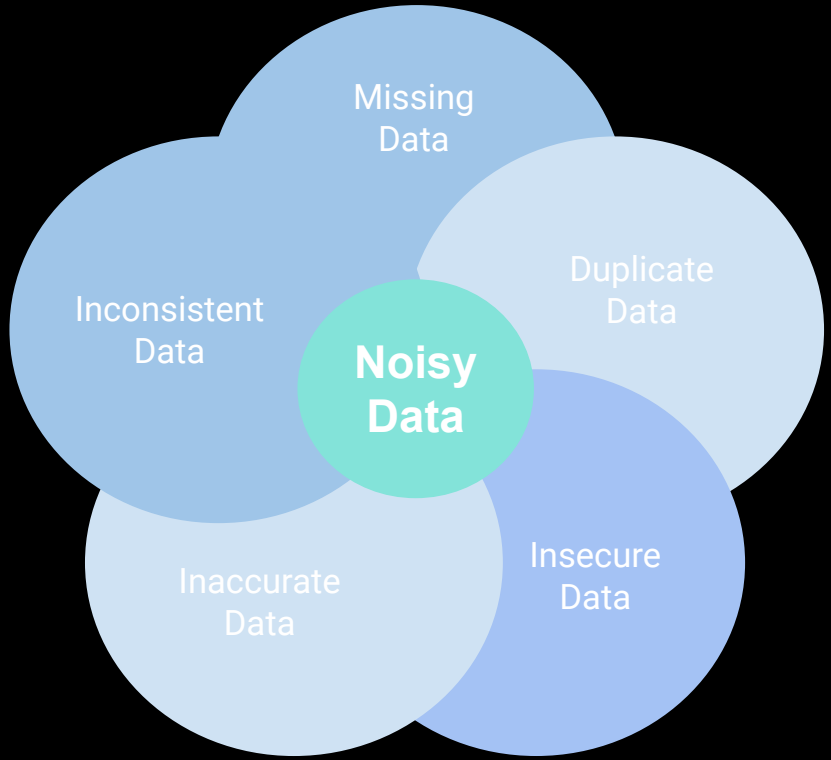
2. Workflow Execution:

This stage is where you specify what operations are a part of the sequence that cleans the data sets. That sequence is called the workflow.

3. Data cleaning

The cleaning stage is the execution of operations specified in the workflow.

The data cleaning process might feature different techniques relative to the project's nature and the data type. But the final objective is always the same to make our data clean, so let's take a look on some data cleaning cases.



Dealing with missing data

Missing data is just unavoidable. You're likely to find even whole rows and columns of missing values in your data sets during the data cleaning process.

Choosing the suitable technique to deal with missing values is so important step in data cleaning process to improve the accuracy of the process.

	Full name	Phone number
0	Noelani A. Gray	0017023975143
1	Myles Z. Gomez	0013294850540
2	Gil B. Silva	0011954922338
3	Prescott D. Hardin	0012979964904
4	Benedict G. Valdez	0019698203536
5	Reece M. Andrews	NaN
6	Hayfa E. Keith	0015361758444
7	Hedley I. Logan	0016815521823
8	Jack W. Carrillo	0019103235265
9	Lionel M. Davis	0011431199210

- **Impute:** This method involves calculating the missing values based on other observations when dealing with numerical data. Statistical techniques like median, mean, or linear regression are helpful if there aren't many missing values.
- **Flag:** Missing data can be informative, Null values can indicate something in your database, so you can replace it with a flag to mark that null value.
- **Drop:** When the missing values in a column are few and far between, the easiest way to handle them is to drop the missing data rows.

Dealing with incorrect data

Incorrect data is often easy to spot, as it's just illogical.

For example, when we have a dataset that includes the blood type of a person, we notice that row 5 has illogical value.

Action:

Depending on the data you work with.

	name	birthday	blood_type	
1	Beth	2019-10-20	B-	
2	Ignatius	2020-07-08	A-	
3	Paul	2019-08-12	O+	
4	Helen	2019-03-17	O-	
5	Jennifer	2019-12-17	Z+	<--
6	Kennedy	2020-04-27	A+	
7	Keith	2019-04-19	AB+	

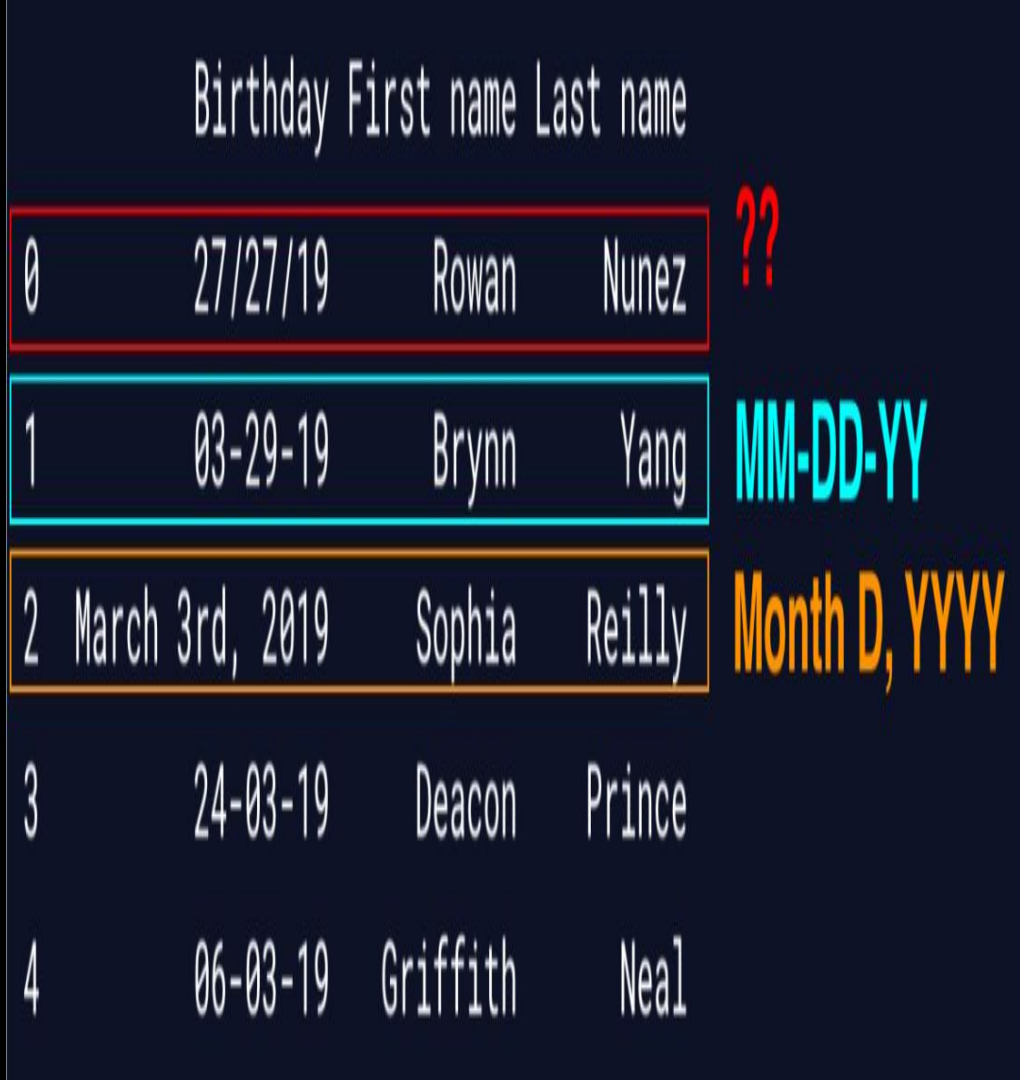
Dealing with inconsistent data

That happens when the dataset have several formats.

For example: in this dataset the Birthday column has several formats.

Action:

All the rows should have one format.



Dealing with Duplicate data

It happens when data is coming from different sources of users, for any reason, submit their entry more than once.

Action:

Remove one row of duplicated rows.

	first_name	last_name	address	height	weight
22	Cole	Palmer	8366 At, Street	178	91
102	Cole	Palmer	8366 At, Street	178	91
28	Desirae	Shannon	P.O. Box 643, 5251 Consectetuer, Rd.	195	83
103	Desirae	Shannon	P.O. Box 643, 5251 Consectetuer, Rd.	196	83
1	Ivor	Pierce	102-3364 Non Road	168	66
101	Ivor	Pierce	102-3364 Non Road	168	88
37	Mary	Colon	4674 Ut Rd.	179	75
100	Mary	Colon	4674 Ut Rd.	179	75

Dealing with outliers

For example: this database includes the rate of movies from (1 - 5), to know the outliers we should visualize the database.

Form the visualization of the database we indicate there is an outlier = 6.

Action:

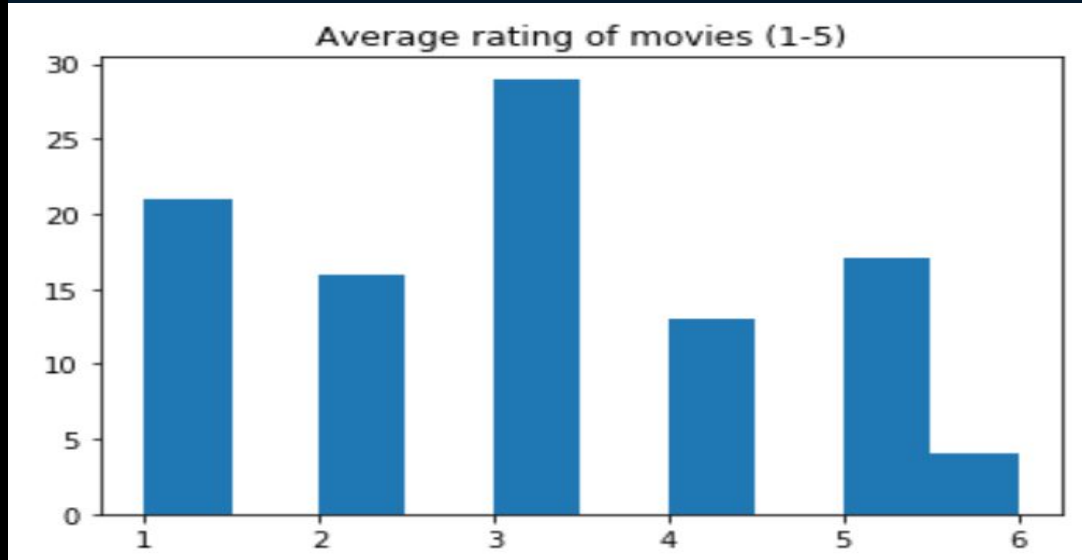
1- Set min and max values.

2- Treat as missing and impute

3- Setting custom value depending on business assumptions

4- Drop the row

	movie_name	avg_rating
0	The Godfather	5
1	Frozen 2	3
2	Shrek	4
...		



4. Validation

The next critical stage of the data cleaning process is quality assurance.

When you've finished the workflow execution, you should audit the data again and make sure all the rules and constraints were in fact executed.

To be sure that your data cleaning process is effective before reporting the final result.

Advantages of data cleaning:

Increased productivity: Effective data cleaning leads to consistent and highly functional databases. No errors mean faster, more effective workflows, which directly impacts productivity.

Better decision making: There is a direct correlation between clean, quality data and reliable business insights: the cleaner the former, the more abundant the latter.

Improved business results: Data cleaning is the key to a properly functioning data analytics solution. Whenever these two things occur, you can expect right conclusions.

Save time and money: Inaccurate data leads to business strategies based on false assumptions. Data cleaning saves your company from potentially wasting both time and money, developing an ineffective strategy.

Thanks