**Data Quality:**

Duplicate Data:

* occurs when data representing the same transaction is accidentally duplicated within a system.

Redundant Data:

* redundant data happens when the same data elements exist in multiple places within a system.
* data redundancy is a function of integrating multiple systems.

Missing Values:

* Missing values occur when you expect an attribute to contain data but nothing is there.
* Missing values are also known as **null values**. A null value is the absence of a value.
* A null is not a space, blank, or other character.
* **Note : Null values poses an error when working with R and Python programing try using SQL. Equivalents functions in these languages do not Handle Null values.**

Invalid Data

* Invalid data are values outside the valid range for a given attribute.
* An invalid value violates a business rule instead of having an incorrect data type.

Nonparametric Data:

* Nonparametric statistics are not based on assumptions, that is, the data can be collected from a sample that does not follow a specific distribution.

Data Outliers:

* A data outlier is a value that differs significantly from other observations in a dataset.

Specification mismatch:

* A specification describes the target value for a component.
* A specification mismatch occurs when an individual component's characteristics are beyond the range of acceptable values.

Data type Validation:

* Data type validation ensures that values in a dataset have a consistent data type.

**DATA MANIPULATION:**

Recoding Data

* Recoding is helpful when you have numeric data you want to analyze by category.
* Recoding data is a technique you can use to map original values for a variable into new values to facilitate analysis.
* Recoding groups data into multiple categories, creating a categorical variable. A categorical variable is either **nominal** or **ordinal**.
* **Nominal** variables are any variable with two or more categories where there is no natural order of the categories, like hair color or eye color.
* **Ordinal** variables are categories with an inherent rank. For example, T-shirt size is an example of an ordinal variable, as sizes come in small, medium, large, and extra-large.

Derived Variables:

* A derived variable is a new variable resulting from a calculation on an existing variable.

Data Merge:

* A data merge uses a common variable to combine multiple datasets with different structures into a single dataset.
* Merging data improves data quality by adding new variables to your existing data

Data Blending:

* Data blending combines multiple sources of data into a single dataset at the reporting layer.
* While data blending is conceptually similar to the extract, transform, and load process, there is a crucial difference.
* Data blending differs from ETL in that it allows an analyst to combine datasets in an ad hoc manner without saving the blended dataset in a relational database.
* Instead of the blended dataset persisting over time, it exists only at the reporting layer, not in the source databases.

Concatenation:

* Concatenation is the merging of separate variables into a single variable.
* highly effective technique when dealing with a source system that stores components of a single variable in multiple columns.
* frequently occurs when dealing with date and time data.
* useful when generating address information.

Data Append:

* A data append combines multiple data sources with the same structure, resulting in a new dataset containing all the rows from the original datasets.
* When appending data, you save the result as a new dataset for ongoing analysis.

Imputation:

* Imputation is a technique for dealing with missing values by replacing them with substitutes.
* When merging multiple data sources, you may end up with a dataset with many nulls in a given column.
* **Example :** If you are collecting sensor data, it is possible to have missing values due to collection or transmission issues.

Approach for Imputing Data:

* Remove Missing Data
* Replace with zero
* Replace with overall average
* Replace with most frequent (mode)
* Closest value average

Reduction:

* Reduction is the process of shrinking an extensive dataset without negatively impacting its analytical value.
* Dimensionality reduction and numerosity reduction are two techniques for data reduction.

Dimensionality reduction:

* dimensionality reduction removes attributes from a dataset.
* Removing attributes reduces the dataset's overall size.

Numerosity reduction:

* numerosity reduction reduces the overall volume of data.
* As data volumes grow, numerosity reduction can improve the efficiency of your analysis.

Sampling:

* technique that selects a subset of individual records from the initial dataset.
* **Aggregation** - Data aggregation is the summarization of raw data for analysis.
* Aggregation is also a means of controlling privacy.
* **Transposition** - Transposing data is when you want to turn rows into columns or columns into rows to facilitate analysis.
* **Normalization** - normalizing data converts data from different scales to the same scale.
* If you want to compare columns whose measurements use different units, you want to normalize the data.
* After normalization is complete, the dataset is ready for statistical analysis.
* **Min-Max Normalization** - makes sure all elements lie within zero and one. It is useful to normalize our data, given that the distribution of data is unknown.
* Formula : Xnormalized = X – Xmin / Xmax– Xmin
* **Parsing\string manipulation** - Raw data can contain columns with composite or distributed structural issues.
* A composite issue is when a raw data source has multiple, distinct values combined within a single character column.
* When this happens, each value in a composite column has data that represents more than one attribute.
* Composite columns need to be split into their component parts to aid analysis.

**Managing Data Quality:**

Circumstances to check for Quality:

* Data Acquisition
* Data transformation and conversion
* Data Manipulation
* Final product preparation

**Automated Validation:**

Data Quality dimensions

* It is essential to consider multiple attributes of data when considering its quality.
* Six dimensions to take into account when assessing data quality are :
* Accuracy
* Completeness
* Consistency
* Timeliness
* Uniqueness
* validity.

Data quality rules and metrics:

* conformity or nonconformity of data
* If source data does not match the destination data type size and format, you have nonconformity.
* validate data conformity issues is to confirm how many rows pass successfully to the target environment and how many fail.

Methods to validate Quality:

* Reasonable expectation
* determine whether or not the data in your analytics environment meets your reasonable expectations.
* Data profiling
  + - profile your data. Data profiling uses statistical measures to check for data discrepancies, including values that are missing, that occur either infrequently or too frequently, or that should be eliminated.
    - Profiling can also identify irregular patterns within your data.
* Data audits
* Data audits look at your data and help you understand whether or not you have the data you need to operate your business
* Sampling
* Sampling is a statistical technique in which you use a subset of your data to inform conclusions about your overall data.
* Cross-validation
* Cross-validation is a statistical technique that evaluates how well predictive models perform.
* Cross-validation works by dividing data into two subsets.
* The first subset is the training set, and the second is the testing, or validation, set.