Clean (Cleansed) data
is a piece of information that meets the
requirements of quality data
and able to contribute to uncovering valuable business insights
(i.e. **satisfy the business requirements)**.

Validity. The degree to which data in question conforms to defined rules or constraints, e.g., information containing dates, should fall within a specific numerical range.

Relevancy. The degree of "usefulness" of data, determining how closely a piece of information is related to an issue you're researching.

Timeliness. Most data loses relevance pretty quickly, so this parameter is related to how "fresh" and up-to-date a piece of information is.

Accuracy. Measured by a degree of compatibility between the data in question and an outside data source. It is our "true value" or a "golden standard."

Completeness. The degree to which all required measures are known

Consistency. It relates to the degree of compatibility of the databases across the whole system.

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)

List of possible errors

- Irrelevant data
- Duplicates
- Incorrect data (calculation)
- Structural errors: inconsistancy in naming, data types, format, capitalization, differet words/ numbers for the same category, misspellings, etc.
- Missing data
- Outliers

Data Cleansing / Data Cleaning activities