**Box plot**

* A graphical representation of the distribution of a dataset.
* Displays the median, quartiles, and potential outliers.
* The box represents the interquartile range (IQR), which contains the middle 50% of the data.
* The line inside the box represents the median.
* Whiskers extend from the box to indicate the range of the data within the fences.
* Outliers are plotted as individual points beyond the fences.

**Fence**

* A cutoff value used in box plots to identify potential outliers.
* Calculated as 1.5 times the interquartile range (IQR) above the upper quartile and below the lower quartile.
* Data points outside the fences are considered potential outliers.

**Whiskers**

* Lines extending from the box in a box plot.
* Indicate the range of the data within the fences.
* The upper whisker extends to the largest data point within the upper fence.
* The lower whisker extends to the smallest data point within the lower fence.

**IQR (Interquartile Range)**

* The range between the first quartile (Q1) and the third quartile (Q3).
* Contains the middle 50% of the data.
* Calculated as Q3 - Q1.

**Quartiles**

* Values that divide a dataset into four equal parts.
* Q1: The 25th percentile.
* Q2: The median (50th percentile).
* Q3: The 75th percentile.

**Quantiles**

* A generalization of quartiles that divide a dataset into any number of equal parts.
* Quartiles are a specific type of quantile.
* Percentiles are another common type of quantile.

**Outliers**

* Data points that are significantly different from other observations in the dataset.
* Can be identified using various methods, including box plots and z-scores.
* May indicate errors, unusual events, or simply the natural variability of the data.

**Cardinality**

* In the context of data, cardinality refers to the number of unique values in a dataset or a particular column (feature) of a dataset.
* High cardinality means many unique values (e.g., customer IDs, product names).
* Low cardinality means few unique values (e.g., binary categories like gender, boolean values).
* Important for understanding data types, storage efficiency, and potential issues like high-cardinality categorical features in machine learning.

**Features with Constant and Semi-Constant Values**

* **Constant features**: Have only one unique value in the entire dataset.
* **Semi-constant features**: Have a dominant unique value and very few occurrences of other values.
* These features provide little to no information for predictive modeling or analysis.
* Often removed during feature selection to improve model efficiency and prevent overfitting.

**MCAR, MNAR, MAR**

These are types of missing data mechanisms:

* **MCAR (Missing Completely at Random)**: The probability of data being missing is unrelated to any observed or unobserved variables. Essentially, missingness occurs entirely by chance.
* **MNAR (Missing Not at Random)**: The probability of missingness depends on the missing values themselves. This is the most complex type of missing data, as the missingness is related to the information we don't have.
* **MAR (Missing at Random)**: The probability of missingness depends on observed variables but not on the missing values themselves. The missingness can be explained by other information in the data.

**Complete Case Analysis or Listwise Deletion**

* A simple method for handling missing data.
* Involves removing any rows (observations) from the dataset that have at least one missing value.
* Can lead to loss of information and biased results if the missing data is not MCAR.
* Suitable when the proportion of missing data is small and MCAR can be reasonably assumed.

**Use of** any **and** all **in a Pandas DataFrame**

* any: Returns True if at least one value in a Series or DataFrame meets a certain condition.
* all: Returns True only if all values in a Series or DataFrame meet a certain condition.
* Useful for filtering data, checking for missing values, and performing logical operations on data.