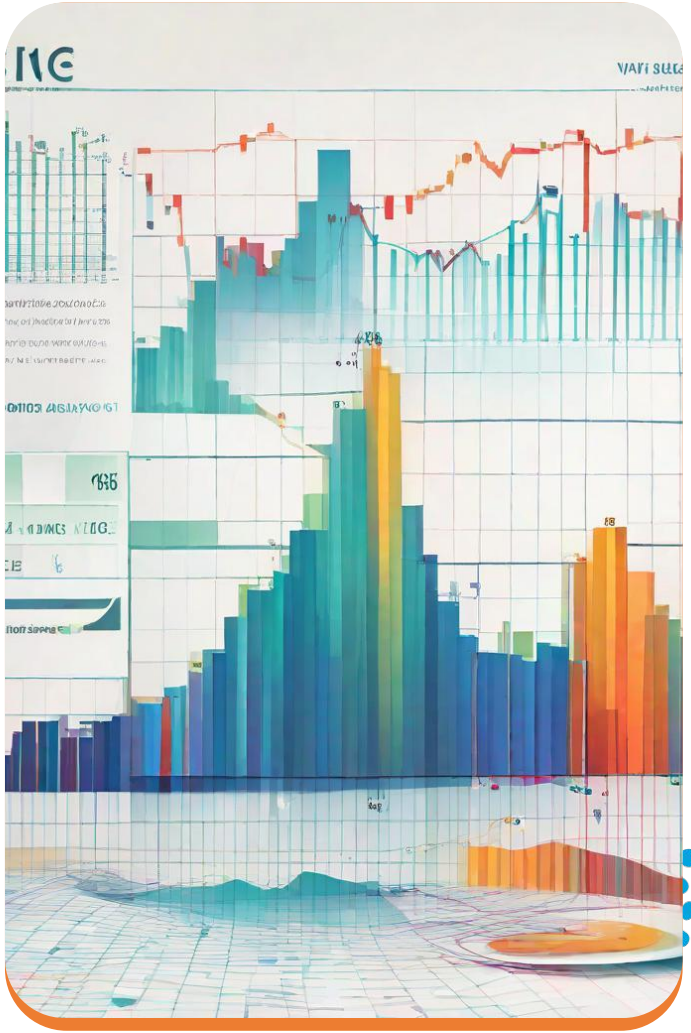
Various Aspects of Data Cleaning

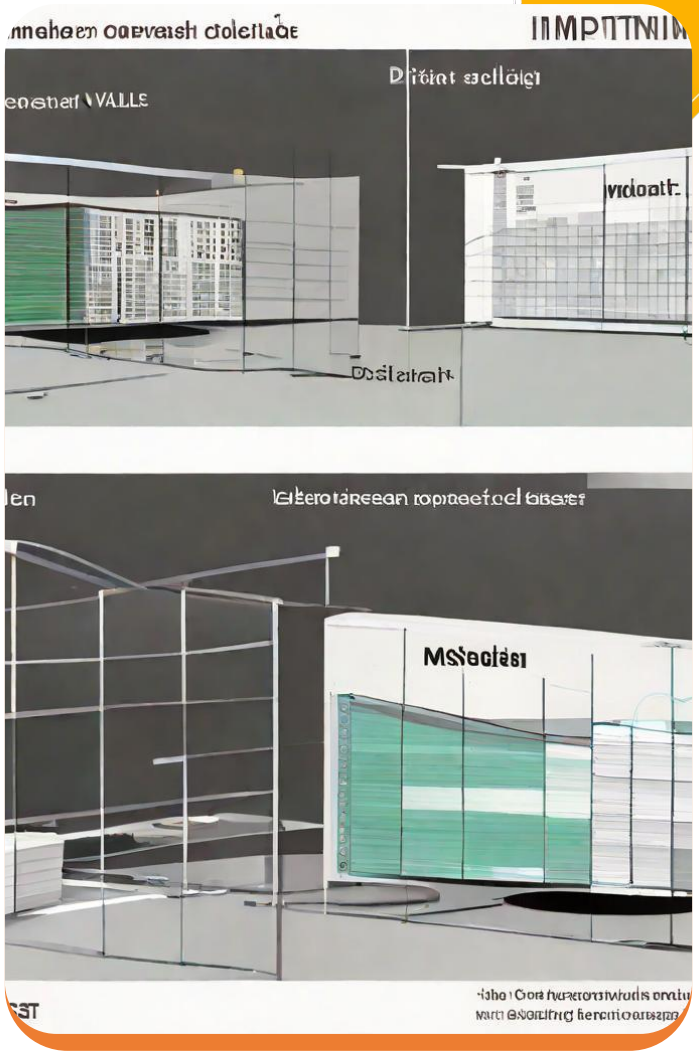
Data cleaning is nothing but a step-by-step process. You must work sequentially to find the inaccuracies and deal with them one-by-one. This lesson will develop your theoretical advancements on each part of the data cleaning process. Let’s understand various aspects of data cleaning and how to deal with them properly.



Dealing with Missing Values

In data analysis, missing values are variables or observations without data. Missing values might result from data collection errors, incomplete survey replies, equipment failure, or unrecorded data. Missing values in data analysis can skew statistical analyses and lead to erroneous findings if not handled properly.

A	B	C	D	E	F
	Income				
	33604				
	26736				
	26732				
	32835				
	14763				
	14041				
	28833				
	Nan				
	42020				
	17669				
	19669				



Strategies for Handling Missing Values

01 Imputation: Imputation estimates missing values from data. Mean, median, and mode imputation are common methods for replacing missing values with the variable's mean, median, or most frequent value;

02 Deletion: Missing observations or variables can be removed from the dataset. This basic method can lose information and bias if the missing data is not random.

Dealing with Incorrect Values

Data points or entries that do not match the anticipated format, range, or pattern in a dataset are inconsistent or inaccurate. Inconsistencies can result from data entry errors, missing numbers, corruption, or data source discrepancies.

A	B	C	D	E	F
	Age				
	30				
	35				
	30				
	49				
	XXX				
	50				
	44				
	49				
	37				
	32				
	50				
	40				

Strategies for Handling Inconsistent values

- Delete observations with incorrect values from the dataset to ensure accuracy and reliability.

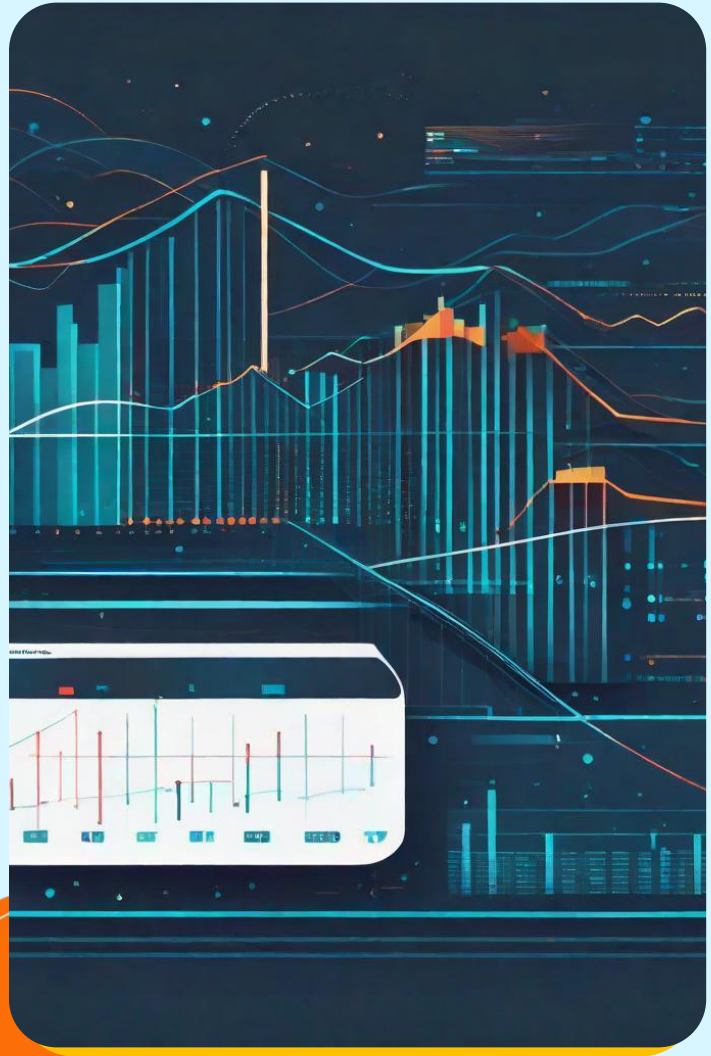
Several Data Types in Data Analysis

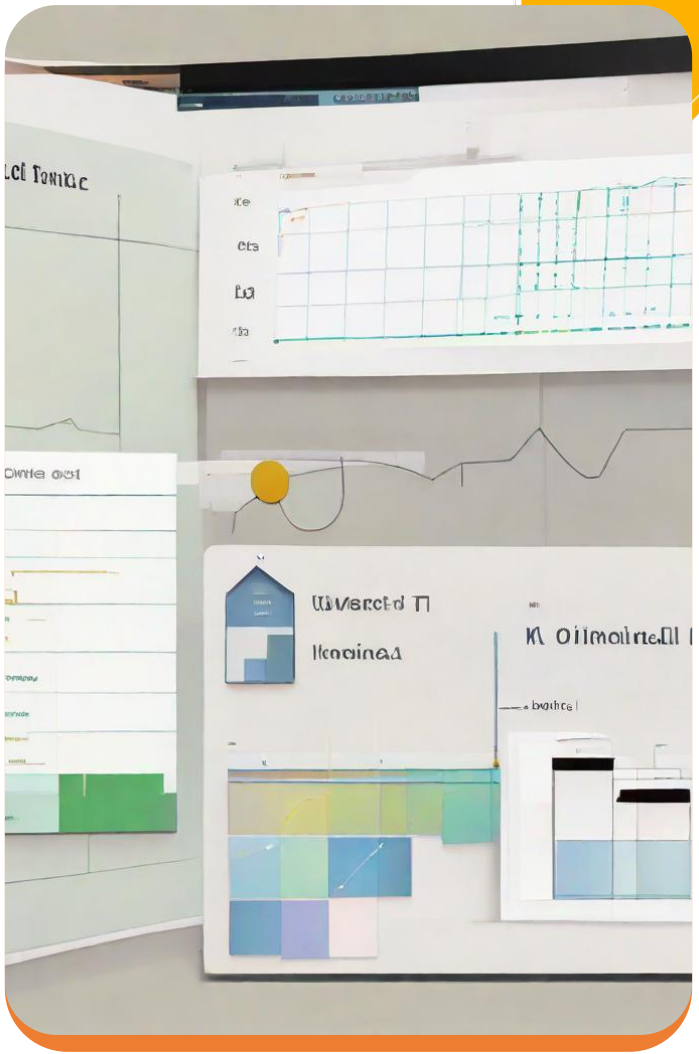
In data analysis, data types refer to the different categories or classifications of data that can be encountered in a dataset. Understanding the types of data you're working with is crucial because it informs the kind of analysis you can perform and the appropriate methods for handling and interpreting the data.

Numeric (Quantitative) Data

01 Discrete: Data that can only take specific, separate values.

02 Continuous: Data that can take any value within a range.





Categorical (Qualitative) Data

01 Ordinal: Data that represents categories with a specific order or ranking.

02 Nominal: Data that represents categories with no inherent order or ranking.

Integers

- Represents whole numbers without any fractional component.
- Can be positive, negative, or zero.
- Typically represented using a fixed amount of memory in programming languages and data analysis tools.
- Examples include counts (e.g., number of items sold), identifiers (e.g., customer IDs), and indices (e.g., array indices in programming).

Float

- Represents numbers with decimal points or fractions.
- Used for storing real numbers, including rational and irrational numbers.
- Examples include temperature readings, financial data, averages, percentages, and continuous data not restricted to whole numbers.

Object

- More generic and flexible compared to integer and float.
- Often used to represent strings (sequences of characters).
- Can also represent complex data structures or mixed data types within a single column.
- Examples include textual data (e.g., names, addresses), categorical data (e.g., labels, categories), or any data not neatly fitting into numerical representations.

Boolean

- Can only take two values: true/false, yes/no, or 1/0.
- Commonly used for representing binary decisions or states.
- Example: whether a customer made a purchase (yes/no).

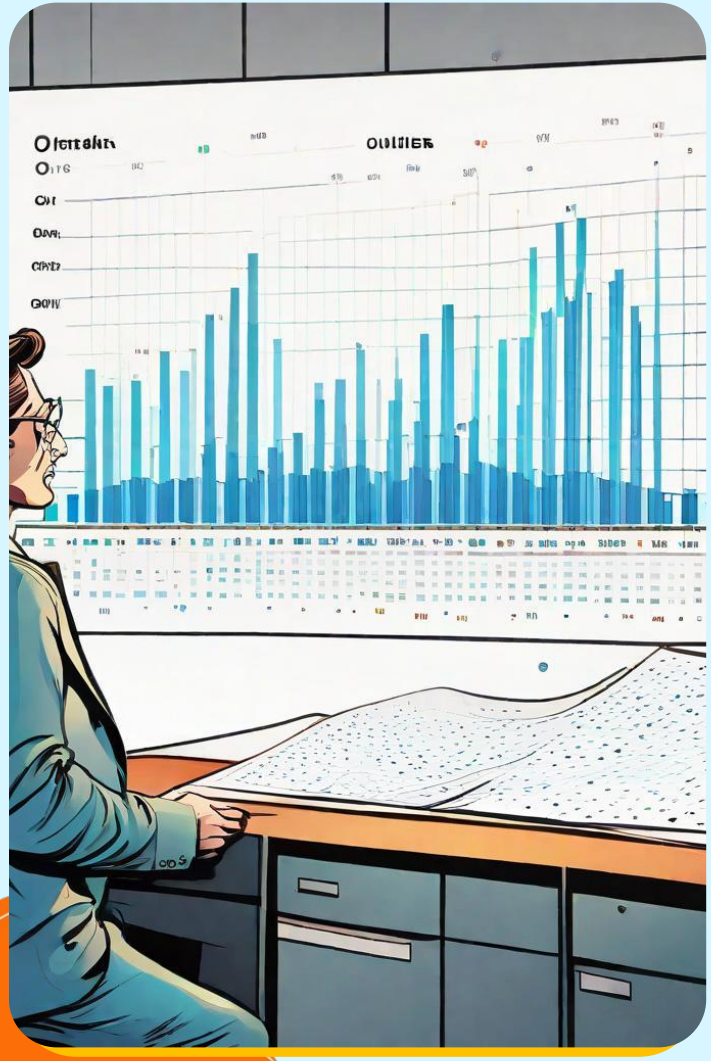
Strategies for Handling Incorrect Data Types

- Converting: Sometimes, data types for specific variables are incorrectly identified and need to be converted to their correct data type.

Dealing with Outliers

01 Outliers can have a significant impact on statistical analyses and machine learning models, potentially skewing results and leading to erroneous conclusions.

02 In data analysis, outliers refer to data points that significantly differ from the rest of the observations in a dataset.



A	B	C	D	E	F
	Age				
	18				
	28				
	1000				
	19				
	20				
	29				
	5				
	19				
	27				
	29				
	29				
	28				

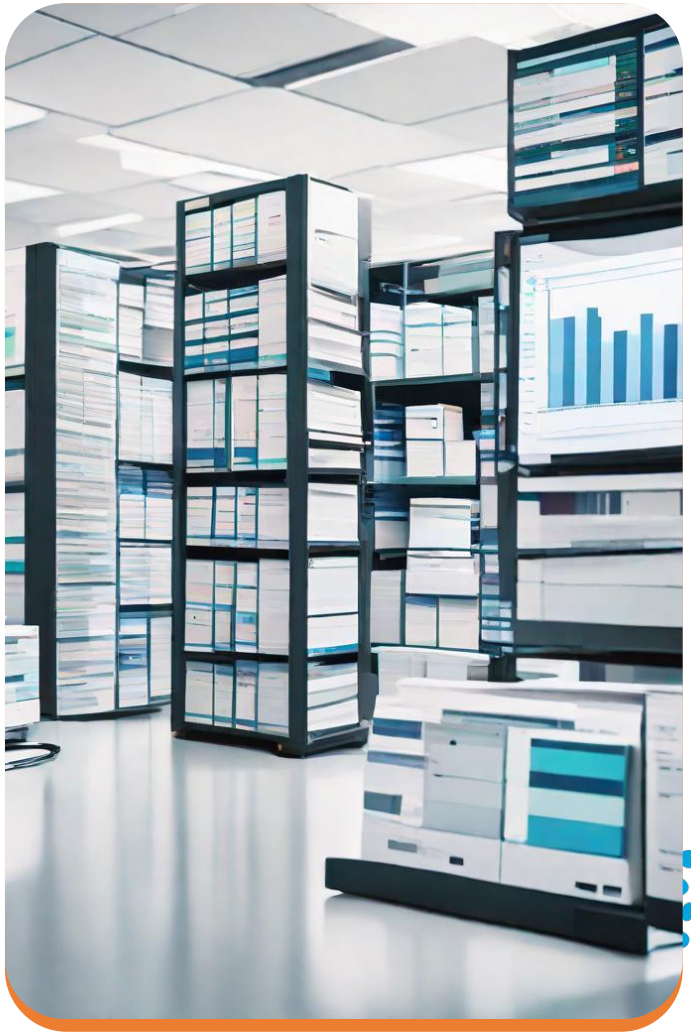
Strategies for Handling Outliers

01 Data Transformation: Transform the data to reduce the impact of outliers.

02 Data Inspection: Visually inspect the data using plots to identify potential outliers.

03 Z-Score: Calculate the z-score for each data point to identify outliers.

04 Winsorization: Replace extreme values with less extreme values.



Dealing with Duplicate Values

- In data analysis, "duplicates" refer to the occurrence of identical records or observations within a dataset.
- Handling duplicates is important to ensure data integrity and accuracy.

Strategies for Handling Duplicate Values

- Delete duplicates within the dataset to avoid biases in analysis and modeling.

Thank you for your time 😊