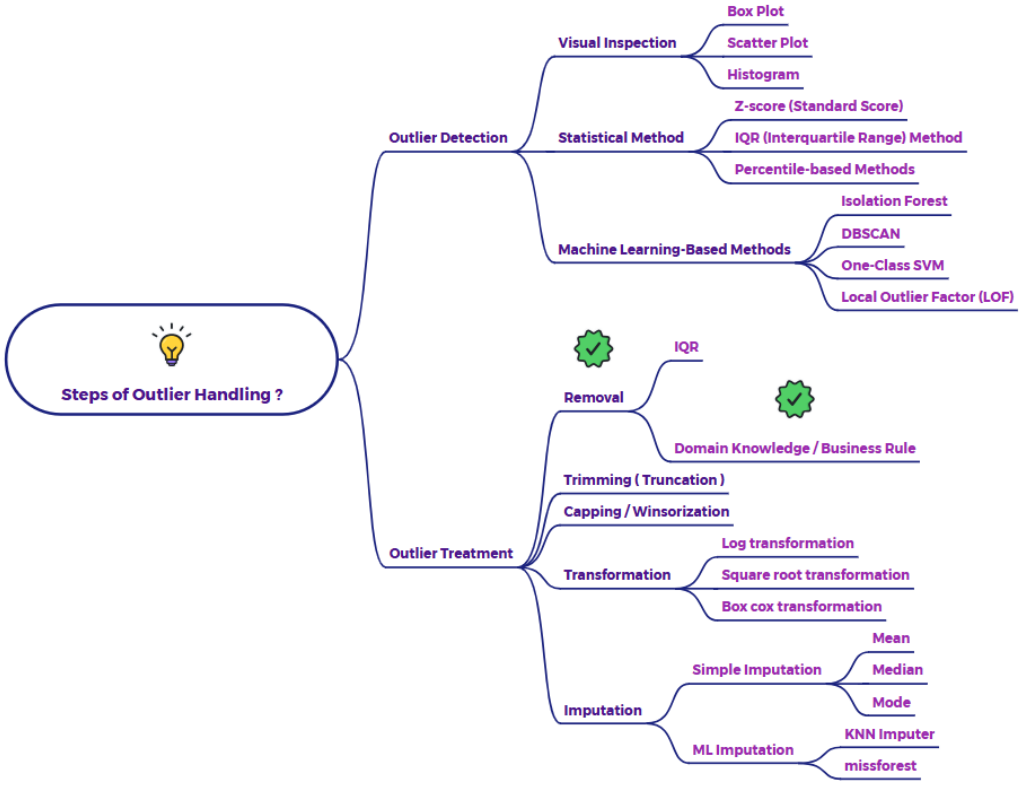
Explain Outlier treatment through removal (Domain knowledge / Business Rule)



Outlier Treatment: Removal (Domain Knowledge / Business Rule)

This approach involves removing outliers from a dataset based on an understanding of the data's context or predefined rules established by experts or business requirements.

Key Points:

- **Domain Knowledge:** This refers to expertise in the specific field or subject area to which the data pertains.
- **Business Rules:** These are specific constraints, policies, or guidelines that dictate how data should be handled.

Process:

1. **Acquire Knowledge/Rules:** Gather information from domain experts, business stakeholders, or relevant documentation to understand what values are plausible or acceptable within the given context.

2. **Define Criteria:** Establish clear criteria or rules for identifying outliers based on the acquired knowledge. These criteria could involve:
    - Specific value ranges
    - Logical inconsistencies
    - Impossible or improbable values
3. **Remove Outliers:** Eliminate data points that violate the defined criteria or rules.

Example:

Let's consider a dataset of customer records for a retail company, including the variable "Age."

- **Domain Knowledge:** We know that the company primarily serves adult customers.
- **Business Rule:** The company policy states that only individuals aged 18 and above can make purchases.

In this scenario, we can apply the following steps:

1. **Acquire Knowledge/Rules:** We have the business rule that the minimum age for a customer is 18.
2. **Define Criteria:** Any customer with an "Age" value less than 18 is considered an outlier.
3. **Remove Outliers:** We remove any customer records where the "Age" is less than 18 from the dataset.

Additional Examples:

- **Manufacturing:** In a dataset of product dimensions, any measurement that falls outside the tolerance limits specified by engineering specifications would be removed.
- **Finance:** In a dataset of stock prices, any price that is illogical (e.g., a negative value) or violates exchange regulations would be removed.
- **Healthcare:** In a dataset of patient vital signs, any value that is physiologically impossible (e.g., a heart rate of 300 bpm) would be removed.

Benefits:

- Provides a targeted and meaningful way to handle outliers.
- Ensures data quality and consistency with real-world constraints.

Cautions:

- Requires accurate and reliable domain knowledge or business rules.
- Can be subjective if the knowledge or rules are not well-defined.
- It should be used judiciously to avoid removing valid data.