Chauvenet’s Criterion

Chauvenet's criterion is a specific method for detecting outliers in data. It is based on the idea that, for a given dataset, the probability of an outlier occurring is relatively low. Chauvenet's criterion can be useful in certain situations, such as when you want to identify outliers in data that follows a normal distribution. However, it is not necessarily the best method for detecting outliers in all cases, and there are other methods that may be more appropriate in certain situations.

According to Chauvenet’s criterion we reject a measurement (outlier) from a dataset of size N when it’s probability of observation is less than `1/2N`. A generalization is to replace the value 2 with a parameter `C`.

Source — Hoogendoorn, M., & Funk, B. (2018). Machine learning for the quantified self. *On the art of learning from sensory data*.

Function

```
def mark_outliers_chauvenet(dataset, col, C=2):
    """Finds outliers in the specified column of datatable and a
    the same name extended with '_outlier' that expresses the re

    Taken from: https://github.com/mhoogen/ML4QS/blob/master/Pyt

    Args:
        dataset (pd.DataFrame): The dataset
        col (string): The column you want apply outlier detectio
        C (int, optional): Degree of certainty for the identific
                           of a normal distribution, typicaly be

    Returns:
        pd.DataFrame: The original dataframe with an extra boole
        indicating whether the value is an outlier or not.
    """

    dataset = dataset.copy()
    # Compute the mean and standard deviation.
    mean = dataset[col].mean()
    std = dataset[col].std()
    N = len(dataset.index)
    criterion = 1.0 / (C * N)

    # Consider the deviation for the data points.
    deviation = abs(dataset[col] - mean) / std

    # Express the upper and lower bounds.
    low = -deviation / math.sqrt(C)
    high = deviation / math.sqrt(C)
    prob = []
    mask = []

    # Pass all rows in the dataset.
    for i in range(0, len(dataset.index)):
        # Determine the probability of observing the point
```

Normal distribution

It's important to note that Chauvenet's criterion is only applicable to datasets that are normally distributed. If your dataset is not normally distributed, this method may not be suitable for identifying outliers.

- Histogram — Do you see a bell shaped curve?
- Boxplot — Is the box symmetrical?