

02 - Feature engineering Python

File Edit View Run Help Last edit was 5 minutes ago Provide feedback

Run all cluster_01 Schedule Share

Cmd 16

There are several methods commonly used to remove outliers from a DataFrame. Here are a few of them:

1) Z-Score Method:

- Calculate the z-score for each value in the DataFrame.
- Remove rows where any column has a z-score greater than a predefined threshold (e.g., 3).
- This method assumes that the data follows a normal distribution.

2) IQR (Interquartile Range) Method:

- Calculate the IQR for each column in the DataFrame.
- Remove rows where any column value is below the first quartile minus a multiple of the IQR or above the third quartile plus a multiple of the IQR (e.g., 1.5 times the IQR).
- This method is robust to non-normal distributions.

3) Tukey's Fences Method:

- Calculate the lower and upper fences based on the first and third quartiles and the IQR.
- Remove rows where any column value is below the lower fence or above the upper fence (e.g., 1.5 times the IQR).
- Similar to the IQR method, this approach is robust to non-normal distributions.

4) Standard Deviation Method:

- Calculate the mean and standard deviation for each column in the DataFrame.
- Remove rows where any column value is above or below a certain number of standard deviations from the mean (e.g., 3 standard deviations).
- This method assumes a normal distribution of the data.

02 - Feature engineering Python

File Edit View Run Help Last edit was 7 minutes ago Provide feedback

Run all cluster_01 Schedule Share

Cmd 16

2) IQR (Interquartile Range) Method:

- Calculate the IQR for each column in the DataFrame.
- Remove rows where any column value is below the first quartile minus a multiple of the IQR or above the third quartile plus a multiple of the IQR (e.g., 1.5 times the IQR).
- This method is robust to non-normal distributions.

3) Tukey's Fences Method:

- Calculate the lower and upper fences based on the first and third quartiles and the IQR.
- Remove rows where any column value is below the lower fence or above the upper fence (e.g., 1.5 times the IQR).
- Similar to the IQR method, this approach is robust to non-normal distributions.

4) Standard Deviation Method:

- Calculate the mean and standard deviation for each column in the DataFrame.
- Remove rows where any column value is above or below a certain number of standard deviations from the mean (e.g., 3 standard deviations).
- This method assumes a normal distribution of the data.

5) Percentile Method:

- Calculate the lower and upper percentiles for each column in the DataFrame (e.g., 1st and 99th percentiles).
- Remove rows where any column value is below the lower percentile or above the upper percentile.
- This method is not distribution-specific and removes extreme values.

It's important to note that the choice of method depends on the characteristics of your data and the specific requirements of your analysis. You may need to experiment with different methods or use a combination of approaches to effectively remove outliers from your DataFrame.