If you have messy data with errors or legitimate outliers, it's important to remove those outliers before building your model. To do this, you first need to calculate the mean and standard deviation. For example, if the mean height is 66 inches with a standard deviation of 3.84 inches, you can use plus and minus 3 standard deviations to remove the outliers.

- Any number between 54.82 and 77.91 inches is a valid number.
- If a height is less than 54.82 or greater than 77.91 inches, then it's an outlier.
- You can create a new data frame with the outliers removed to use for building your machine learning model.

Z-Score

Z-Score is a measure of how many standard deviations away a data point is from the mean. You can calculate the Z-Score for every single data point in your data set by subtracting the mean and dividing by the standard deviation. Once you have the Z-Score for each data point, it's even easier to remove the outliers.

Exercise

Complete the exercise provided in the link in the video description to practice using Z-Score and outlier removal. Working on exercises like this is important for understanding and retaining the concepts.

Understanding Logarithmic Function and Its Significance in Data Science and Machine Learning

Consider a dataset of property prices where we want to detect an outlier. An outlier is a value that is much higher or lower than the other values in the dataset. In this case, 300k is an outlier. Depending on the situation, you may want to treat this value as a normal value or an outlier. If you use the mean as your basis for detecting outliers, it may not detect this value as an outlier because the mean is affected by this high value. Instead, you can use the median and median absolute deviation (MAD) to detect outliers.

Calculating MAD and Modified Zscore

MAD is similar to standard deviation but uses the median instead of the mean as its basis point. The usual criteria for MAD is 3.5, which is used in modified zscore. The formula for modified zscore is (x - median(x)) / (MAD(x) * 0.6745), where x is a single data point.

Excel Demo

We provide an excel demo to demonstrate the calculation of modified zscore. In the demo, we calculate zscore and modified zscore for a dataset of people's heights and show how to use them to detect outliers.

Z-Score and Modified Z-Score for Outlier Detection

To detect outliers in a dataset, a common method is to use z-scores. Any value that has a z-score of 3.5 or more is considered an outlier. However, when the sample size is smaller, outliers can have a bigger impact on the average and standard deviation. In such cases, using modified z-scores may be more effective.