00:00-19:26

Outlier Detection and Removal Techniques in Data Analysis

Introduction

Outliers are data points that deviate significantly from the majority of the data in a dataset. They can be caused by data errors or represent extreme variations in the data. In this article, we will explore different techniques for detecting and removing outliers to ensure the integrity and accuracy of our dataset.

Standard Deviation Method

One commonly used technique for outlier detection is the standard deviation method. By calculating the standard deviation of a feature, we can identify data points that lie beyond a certain threshold from the mean. These points can be considered outliers and removed from the dataset.

Domain Knowledge Approach

Another approach for outlier detection is using domain knowledge. In certain domains, there are specific rules or expectations that can help identify outliers. For example, in the real estate domain, a two-bedroom apartment would typically have a minimum square footage threshold. If any data points fall below this threshold, they can be considered outliers and removed from the dataset.

Data Cleanup Process

To illustrate the outlier removal process, let's consider a scenario where we have a dataset of real estate properties. We will go through the steps of outlier detection and removal using various techniques.

Step 1: Detecting Square Footage Outliers

We consult with a real estate expert to determine the typical square footage per bedroom ratio. Based on their input, we identify data points where the square footage per bedroom falls below the expected threshold. These data points are considered outliers and need to be removed from the dataset.

Step 2: Removing Square Footage Outliers

We create a new dataframe and filter out the identified outliers based on the square footage per bedroom criteria. After removing these outliers, we verify the number of remaining data points.

Step 3: Detecting Price per Square Foot Outliers

Next, we examine the price per square foot feature. We look for properties with abnormally high or low price per square foot values, which can be considered outliers. By using statistical measures such as mean and standard deviation, we can filter out the extreme cases.

Step 4: Removing Price per Square Foot Outliers

We develop a function that removes outliers based on the standard deviation approach. This function groups the data by location, calculates mean and standard deviation values, and filters out data points beyond one standard deviation from the mean. By applying this function to the dataframe, we remove the price per square foot outliers.

Step 5: Analyzing Bedroom vs. Price

We compare the property prices for two-bedroom and three-bedroom apartments with the same square footage. If we observe that the two-bedroom apartments have higher prices, it could indicate outliers or anomalies. We create scatter plots to visualize these cases and identify any outliers that need to be removed.

Step 6: Removing Bedroom vs. Price Outliers

We develop a function that removes outliers based on the comparison of bedroom and price values. By calculating the mean of one-bedroom apartments and filtering out two-bedroom apartments with prices below the mean, we can remove these outliers.

Step 7: Analyzing Bathroom Feature

We examine the bathroom feature and notice that some properties have an unusually high number of bathrooms. We consult with our business manager and establish a criterion for identifying bathroom outliers. In this case, we remove properties with a number of bathrooms greater than the number of bedrooms plus two.

Step 8: Data Cleanup and Preparation

After removing all the identified outliers, we drop unnecessary features such as price per square foot and size from the dataset. This step helps prepare the dataset for machine learning training.

Conclusion

Outlier detection and removal are crucial steps in data analysis to ensure the quality and reliability of the dataset. By applying techniques such as standard deviation, domain knowledge, and statistical analysis, we can identify and eliminate outliers that could potentially affect our analysis or machine learning models. Through careful data cleanup and preparation, we can create a clean and accurate dataset for further analysis or model training.