STATISTICAL METHODS FOR HANDLING OUTLIERS

There are basically different types of handling outliers. However, it is interesting and important to note that the right outlier removal technique to be used is dependent on the type of data being used.

1. **Winsorization (Flooring and Capping):** Here we simply remove the 10^{th} and the 95^{th} percentile of the data.
   Application
   Employed when certain similar distribution behavior is observed in both higher and lower percentile ranges.

2. **Trimming:** In this case, we eliminate outliers by setting our desired threshold
   Application
   - When you know that it is wrong with the data. Like what the actual data should be.
   - You have a lot of data and outliers are few
   - If the outlier creates a relationship where there isn't one otherwise

3. **IQR Score:** We simply remove outliers exceeding out IQR scores for outlier detection
   Application
   Used when data in question has defined distribution such as Normal or Gaussian distribution hence contains trends of varying points across percentile ranges.

4. **Log transformation:** Another common method to use is the by simply transforming the data either using logarithmic, square root or square transforms.
   Application
   It's often preferred when the response variable follows **exponential distribution or is right-skewed)** both pull in high numbers

5. **Replacing Outliers with median values:** In as much as this method me not be efficient however, it can be an alternative in cases where we already know our data pattern and hence can replace the outliers with the mean. Also applies to cases where the data is very few
   Application
   Applied whenever an outlier seems to be due to a mistake in your data, you try imputing a value such as mean, median etc.

All codes for the recommended methods above can be located here at the jupyter notebook

Note: Asides the above techniques, is equally a good practice to try

- feature scaling or normalization
- to use algorithms that are not easily affected by outliers such as tree-based models like Random Forest etc.
- ….

Credits: https://www.pluralsight.com/guides/cleaning-up-data-from-outliers