**Outliers:**

* If the data, or feature of interest is normally distributed, you may use standard deviation and z-score to label points that are farther than three standard deviations away from the mean as outliers.
* If the data is not normally distributed, you can use the interquartile range or percentage methods to detect outliers.

Outliers are those data points that are significantly different from the rest of the dataset. They are often abnormal observations that skew the data distribution, and arise due to inconsistent data entry, or erroneous observations.

To ensure that the trained model generalizes well to the valid range of test inputs, it’s important to detect and remove outliers.

In the machine learning pipeline, data cleaning and preprocessing is an important step as it helps you better understand the data. During this step, you deal with missing values, detect outliers, and more.

As outliers are very different values—abnormally low or abnormally high—their presence can often skew the results of statistical analyses on the dataset. This could lead to less effective and less useful models.

The goal of outlier detection is to remove the points—which are truly outliers—so you can build a model that performs well on unseen test data. We’ll go over a few techniques that’ll help us detect outliers in data.

**When the data, or certain features in the dataset, follow a normal distribution, you can use the standard deviation of the data, or the equivalent z-score to detect outliers.**

In statistics, standard deviation measures the spread of data around the mean, and in essence, it captures how far away from the mean the data points are.

Let’s denote the standard deviation of the distribution by σ, and the mean by μ.

One approach to outlier detection is to set the lower limit to three standard deviations below the mean (μ - 3\*σ), and the upper limit to three standard deviations above the mean (μ + 3\*σ). Any data point that falls outside this range is detected as an outlier.

You can use Seaborn’s displot() function to visualize the data distribution. In this case, the dataset follows a normal distribution, as seen in the figure below.

sns.set\_theme()

sns.displot(data=scores\_data).set(title="Distribution f Scores", xlabel="Scores")

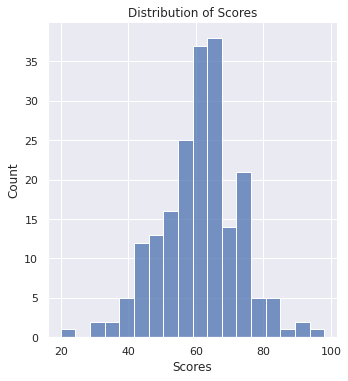


Figure 1: Normal Distribution of Scores

To obtain the mean and standard deviation of the data in the dataframe df\_scores, you can use the .mean() and the .std() methods, respectively.

**Code for Outlier Detection Using Interquartile Range (IQR)**

In statistics, interquartile range or IQR is a quantity that measures the difference between the first and the third quartiles in a given dataset.

* The first quartile is also called the one-fourth quartile, or the 25% quartile.
* If q25 is the first quartile, it means 25% of the points in the dataset have values less than q25.
* The third quartile is also called the three-fourth, or the 75% quartile.
* If q75 is the three-fourth quartile, 75% of the points have values less than q75.
* Using the above notations, IQR = q75 - q25.

You can use the box plot, or the box and whisker plot, to explore the dataset and visualize the presence of outliers. The points that lie beyond the whiskers are detected as outliers.

You can generate box plots in Seaborn using the boxplot function.

## How to Detect Outliers Using Percentile

the interquartile range works by dropping all points that are outside the range [q25 - 1.5\*IQR, q75 + 1.5\*IQR] as outliers. But removing outliers this way may not be the most optimal choice when your observations have a wide distribution. And you may be discarding more points—than you actually should—as outliers.

Depending on the domain, you may want to widen the range of permissible values to estimate the outliers better. Next, let’s revisit the scores dataset, and use percentile to detect outliers.