Outliers

This lesson will focus on the different types of outliers and why they happen.

WE'LL COVER THE FOLLOWING

What are outliers?

Outliers are observations that are significantly distant from other observations. These do not follow the general trend of the data. Outliers can indicate variation or error in the data. Outliers in a single variable/column are called **univariate** while outliers in multiple variables/columns are called **multivariate**.

Sources of outliers

Outliers can be caused by a variety of reasons. Some common ones are:

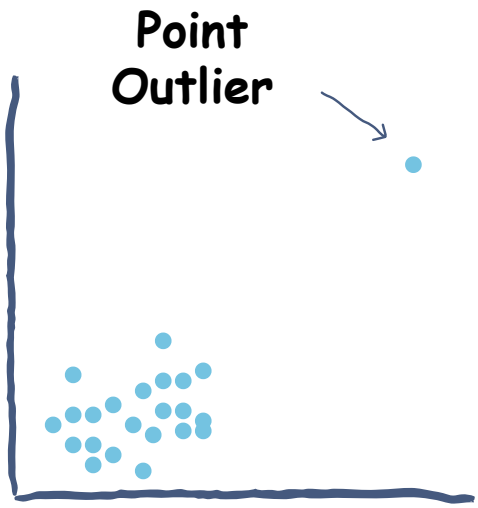
- Errors in entering data.
- Errors in measuring data, e.g., errors in the measuring instrument.
- Errors in collecting and merging data from multiple sources.
- Errors in processing data.
- Natural variance because of some unknown reason.

Types of outliers

Outliers can be classified into three broad categories:

1. Point or Global outliers

These are observations that deviate from all of the other observations, e.g., if the temperature is recorded as 100 degrees Celsius, or a person who usually spends $100 in a week spends $500 this week.



Point Outlier

2. Contextual or Conditional outliers

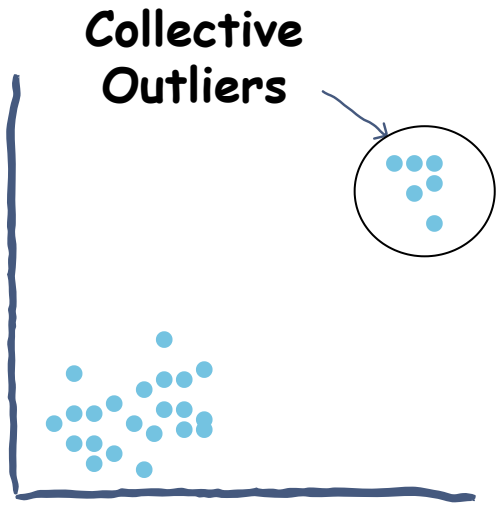
These are data points that are not outliers globally, but are outliers in their own context. If we look at a subset of the data, then we are looking in a context.

For instance, a sudden unusual temperature drop in the summer season is considered a contextual outlier where the context is the summer season.

Another example could be if the price of a good is $15 and its price falls below $10 during the Christmas period. If its price falls below $10 in July, then that would be a contextual outlier with the month of July being the context.

3. Collective outliers

These are a group of observations that are outliers globally from the rest of the observations but are not outliers within the group. An example could be a sudden increase in stock transactions of a particular company during a month or unusual delays in shipping orders over a period of three days.



Collective Outlier

Since we are now familiar with the existence of outliers, we will look at the detection and removal of outliers in the next lesson.