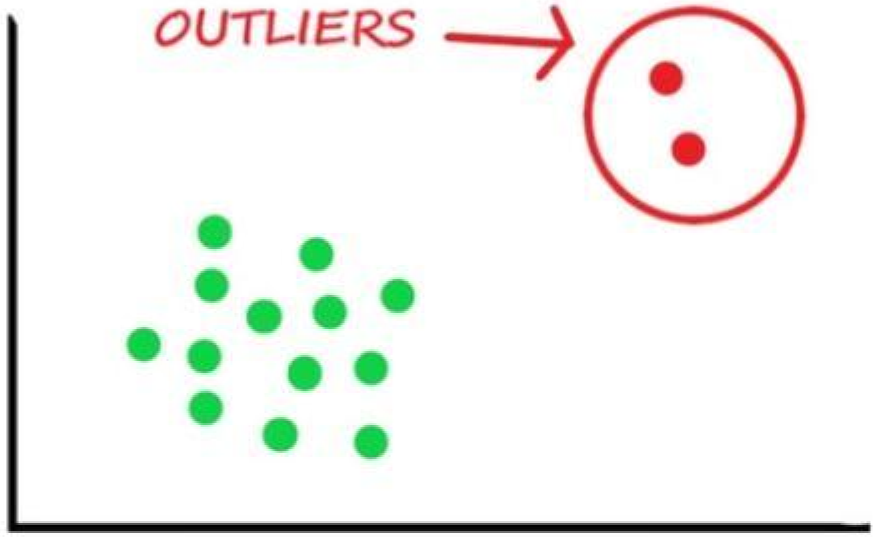
# On Outliers

## What is an Outlier?

An **outlier,** in the field of statistics**,** is a point of data, or an observation that differs significantly from other observations, or from the average observation in the dataset. Because outliers, when they occur, are the most extreme of the data values, or the most extreme in distance from the mean value, the sample maximum, minimum, or both may be outliers, if the sample is found to contain outliers.



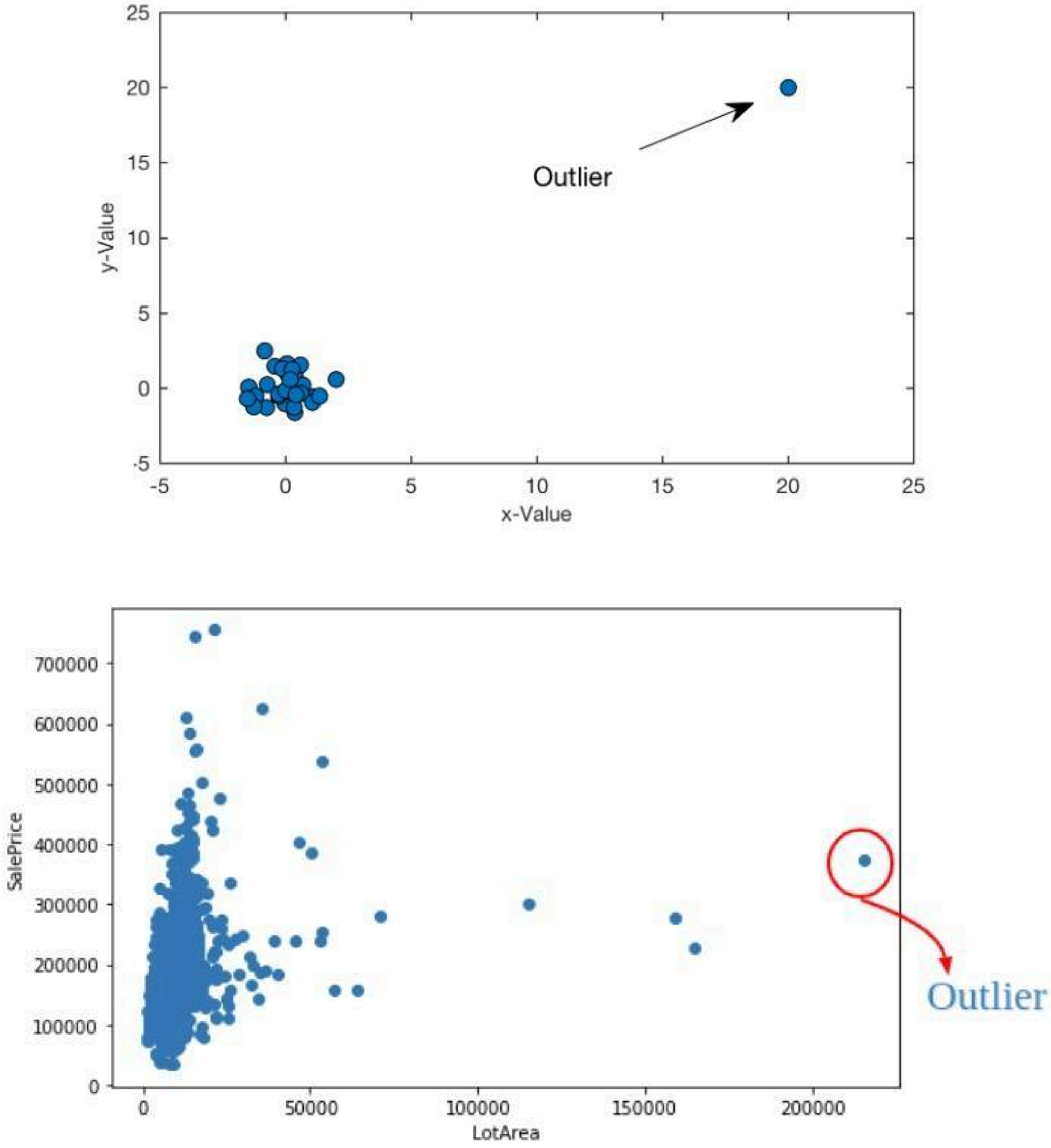


**Figure 1: Three examples of outliers in datasets.**

The presence of outliers in the data may be due to wide variations in the measurement, it may be the result of experimental error, or simply bad data entry. Outliers can be harmless, but they can also cause serious problems in statistical analyses, depending on the underlying distribution.

## Why do Outliers Matter?

Consider the  data set below that gives the household income for 5 households in a certain area.

| Household | Income (thousands of dollars/ month) |
|---|---|
| 1 | 3.5 |
| 2 | 4.2 |
| 3 | 6.4 |
| 4 | 6.1 |
| 5 | 18.5 |

Clearly the income for Household 5 is an outlier, although it is correct.
The average household income for the five households is:

$$Mean = \frac{\Sigma X_i}{5} = \frac{3.5+4.2+6.4+6.1+18.5}{5} = \frac{38.7}{5} = 7.7$$

The calculation for the mean gives 7.7 thousand dollars/month . As the average, this figure is supposed to be representative of the incomes of the 5 households. However, since four out of the five households show incomes far lower than 7.7, this mean is not very representative, and any analysis done using the mean here as the central tendency measure will be erroneous. The calculated mean gave a large value owing to the presence of the outlier, which tended to pull the mean up.  Similarly, low-end outliers will tend to pull the mean down.
This is a simple example of why detecting and treating outliers are important aspects of data-cleaning and pre-analysis data preparation.

# How Do You Deal with Outliers?

Sometimes outliers may be errors in the data and should be removed. In other cases, these points are correct readings yet they are so different from the other points that they appear to be incorrect.

The simplest way to deal with outliers is to remove those points entirely from the dataset. However, as a rule of thumb, if the percentage of data points that are outliers is greater than 5%, then this method becomes inadvisable and other methods should be looked into.

When deciding whether to eliminate an outlier, the cause has to be considered. If the outliers' origin can be attributed to an experimental error, or if it can be otherwise determined that their presence is due to bad data entry, it is generally recommended to eliminate them. However, if it is possible to correct the erroneous value, then this might be the best path of action, for in many cases, outliers contain valuable information about the process under study. Before considering the possible elimination of these points from the data, one should try to understand why they appeared and whether it is likely similar values will continue to appear.

## A Few of the Simpler ways to deal with outliers after identifying (flagging) them:

1. Set up a Validation to exclude extreme numbers beyond or below certain limits, i.e., use fences. For example, we could agree that all values at a distance of greater than 3 standard deviations from the mean will be removed. We will be seeing this method up close in a subsequent section.
2. Change the value of outliers using the mean value. In other words, replace all outliers in the group by the group mean value.
3. Use statistics that are impervious to outliers. If we need a central tendency measure, then instead of using the mean, as it is heavily affected by outliers, use the *median*, as it is impervious to outliers. Statistics like the median that are not affected by outliers are called '**robust**' measures.
4. Drop the outliers entirely.