**Standard Deviation method of identifying the Outliers**

If a value is a certain number of standard deviations away from the mean, that data point is identified as an outlier. The specified number of standard deviations is called the threshold. The default value is 3. This method can fail to detect outliers because the outliers increase the standard deviation.

If we know that the distribution of values in the sample is Gaussian or Gaussian-like, we can use the standard deviation of the sample as a cut-off for identifying outliers.

The [Gaussian distribution](https://machinelearningmastery.com/continuous-probability-distributions-for-machine-learning/) has the property that the standard deviation from the mean can be used to reliably summarize the percentage of values in the sample.

For example, within one standard deviation of the mean will cover 68% of the data.

So, if the mean is 50 and the standard deviation is 5, as in the test dataset above, then all data in the sample between 45 and 55 will account for about 68% of the data sample. We can cover more of the data sample if we expand the range as follows:

* 1 Standard Deviation from the Mean: 68%
* 2 Standard Deviations from the Mean: 95%
* 3 Standard Deviations from the Mean: 99.7%

A value that falls outside of 3 standard deviations is part of the distribution, but it is an unlikely or rare event at approximately 1 in 370 samples.

Three standard deviations from the mean is a common cut-off in practice for identifying outliers in a Gaussian or Gaussian-like distribution. For smaller samples of data, perhaps a value of 2 standard deviations (95%) can be used, and for larger samples, perhaps a value of 4 standard deviations (99.9%) can be used.

Why is prop table used in R?

The prop. table() function in R **can be used to calculate the value of each cell in a table as a proportion of all values**

What %>% means in R?

%>% is called the **forward pipe operator** in R. It provides a mechanism for chaining commands with a new forward-pipe operator, %>%. This operator will forward a value, or the result of an expression, into the next function call/expression.

What is **stat=”identity**” in geom\_bar function in R?

If you want the heights of the bars to represent values in the data, use stat=”identity” and map a value to the y aesthetic.

What does Ldply do in R?

ldply: **Split list, apply function, and return results in a data frame**.

What is the complete function in R?

complete. cases() function in R Language is **used to return a logical vector with cases which are complete**, i.e., no missing value.

options(scipen = 999) The following command extends the number of **lines of printing** your results in the console.

"So, in essence this value determines how likely it is that scientific notation will be triggered. So, to prevent scientific notation, simply use a large positive value like 999"

scale\_x\_continuous() and scale\_y\_continuous() are the default scales for continuous x and y aesthetics.

What does Facet\_wrap mean in R?

facet\_wrap() : “wraps” **a 1d ribbon of panels into 2d**.

facet\_grid() : produces a 2d grid of panels defined by variables which form the rows and columns.

When should you use Facet\_wrap?  
  
This is useful if **you have a single variable with many levels and want to arrange the plots in a more space efficient manner**. You can control how the ribbon is wrapped into a grid with ncol , nrow , as. table and dir . ncol and nrow control how many columns and rows (you only need to set one).

facet\_wrap() : "wraps" a **1d** ribbon of panels into 2d.

facet\_grid() : produces a 2d grid of panels defined by variables which form the rows and columns.