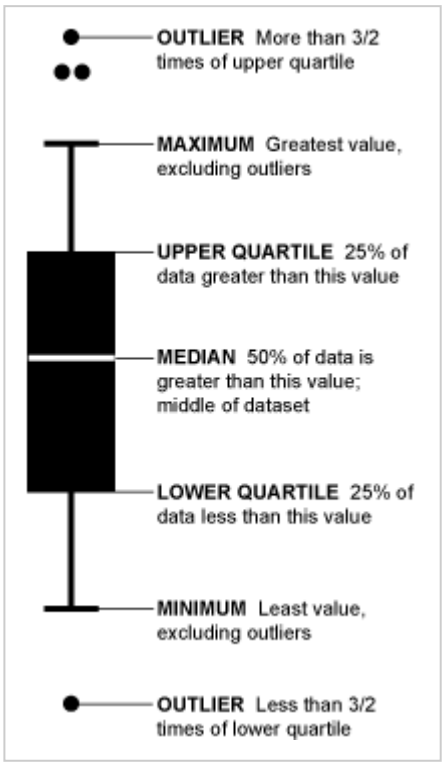
Statistical Features - Working With Box Plots

WE'LL COVER THE FOLLOWING

Another important statistical concept is that of percentile, so let's get a good understanding of this essential feature. Also, let's learn to interpret statistical features from plots.

Anatomy of a Box Plot

Do you remember the box plots from the lessons on data visualization? As we saw earlier, we can write some very simple code using Matplotlib's `boxplot()` method to obtain statistical features in the form of box plots:



How to read a box-plot (Image credits: flowingdata.com)

A boxplot is basically a graph that presents information from a **five-number summary**. If we look at the diagram above, we can see that in a box plot:

- The ends of the box are the first (lower) and third (upper) quartiles — the box spans the so-called interquartile range. The first quartile basically represents the 25th percentile, meaning that 25% of the data points fall below the first quartile. The third quartile is the 75th percentile, meaning that 75% of the points in the data fall below the third quartile.
- The median, marked by a horizontal line inside the box, is the middle value of the dataset, the 50th percentile. Median is used instead of mean because it is more robust to outlier values *(we will talk about this again later and understand why)*.
- The whiskers are the two lines outside the box that extend to the highest and lowest (or min/max) observations in our data.

Five-Number Summary

To recap, a five-number summary is made up of these five values: the maximum value, the minimum value, the lower quartile, the upper quartile, and the median.

These values are presented together and ordered from lowest to highest:

- **Minimum value**
- **Lower quartile (Q1/25th Percentile)**
- **Median value (Q2/50th Percentile)**
- **Upper quartile (Q3/75th Percentile)**
- **Maximum value**

These five numbers give us a summary of the data as each value describes a specific part of a dataset: the median identifies the center of a dataset; the upper and lower quartiles span the middle half of a data set; and the highest and lowest observations give us insights into the actual dispersion of the data. The five-number summary is a **useful measure of spread in the dataset**.

Interpreting A Box Plot

- A short box plot tells us that many of our data points are similar, we have

many values in a small range. On the other hand, a tall box plot implies that much of the data points are quite different, we have values that are spread over a wide range.

- A median value that is closer to the bottom tells us that most of our data points have lower values. While a median value closer to the top tell us that most of our data has higher values. Basically, a median line that is not in the middle of the box is an indication of skewed data.
- *What about the length of those whiskers?*
  Long whiskers tell us that our data has a high standard deviation and variance, i.e., the values are spread out and vary a lot. If there are long whiskers on one side of the box, but not the other, then it's an indication that our data varies, but only in one direction.

Isn't this a lot of useful information from a few simple statistical features that are easy to calculate? Remember to make use of them while doing a **preliminary investigation of a large dataset**, when comparing two or more datasets, and when you need a **descriptive analysis including data skewedness or outliers** of your data.