**What is Lambda Function in Python?**

A lambda function is a small anonymous function.

A lambda function can take any number of arguments, but can only have one expression.

def myfunc(n):

return lambda a : a \* n

mydoubler = myfunc(2)

mytripler = myfunc(3)

print(mydoubler(11))

print(mytripler(11))

**Name some types of Hypothesis Testing used nowadays.**

Alternative Hypothesis

Alternative Hypothesis (H1) or the research hypothesis states that there is a relationship between two variables (where one variable affects the other). The alternative hypothesis is the main driving force for hypothesis testing.

It implies that the two variables are related to each other and the relationship that exists between them is not due to chance or coincidence.

When the process of hypothesis testing is carried out, the alternative hypothesis is the main subject of the testing process. The analyst intends to test the alternative hypothesis and verifies its plausibility.

Null Hypothesis

The Null Hypothesis (H0) aims to nullify the alternative hypothesis by implying that there exists no relation between two variables in statistics. It states that the effect of one variable on the other is solely due to chance and no empirical cause lies behind it.

The null hypothesis is established alongside the alternative hypothesis and is recognized as important as the latter. In hypothesis testing, the null hypothesis has a major role to play as it influences the testing against the alternative hypothesis.

Non-Directional Hypothesis

The Non-directional hypothesis states that the relation between two variables has no direction.

Simply put, it asserts that there exists a relation between two variables, but does not recognize the direction of effect, whether variable A affects variable B or vice versa.

Directional Hypothesis

The Directional hypothesis, on the other hand, asserts the direction of effect of the relationship that exists between two variables. Herein, the hypothesis clearly states that variable A affects variable B, or vice versa.

Statistical Hypothesis

A statistical hypothesis is a hypothesis that can be verified to be plausible on the basis of statistics. By using data sampling and statistical knowledge, one can determine the plausibility of a statistical hypothesis and find out if it stands true or not.

What is the difference between an error and a residual error?

**Errors** pertain to the true data-generating process (DGP), whereas **residuals** are what is left over after having estimated your model. In truth, assumptions like normality, homoscedasticity, and independence apply to the errors of the DGP, not your model's residuals. (For example, having fit p+1

parameters in your model, only N−(p+1)

residuals can be independent.) However, we only have access to the residuals, so that's what we work with.

The error term is the difference between the observed value for the dependent variable and its theoretical value, while a model is applied on overall population. We don't actually calculate it.

Residual is the practically calculated term during modeling exercise; It is the difference between the actual value in the sample and predicated value in the sample.

Underfitting, Overfitting from the original model

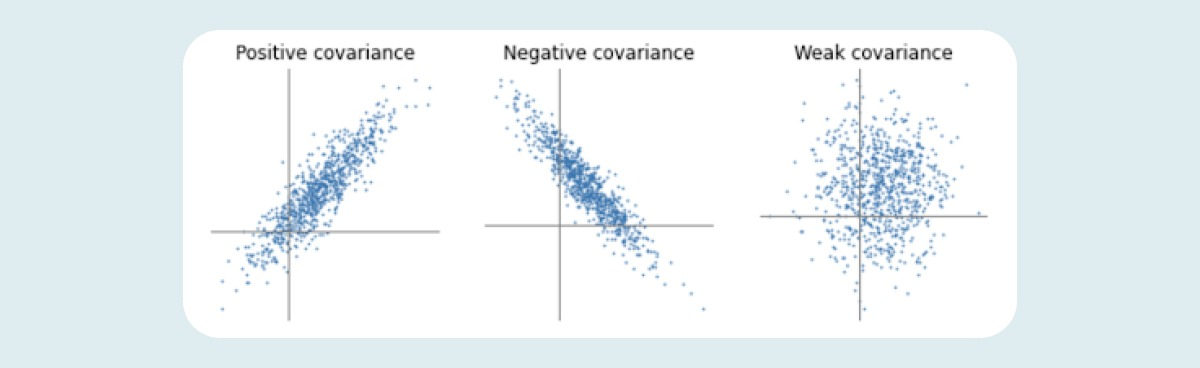
**How can you check if the Regression model fits the data well?**

Cross-validation (CV) is a technique used to assess a machine learning model and test its performance (or accuracy). It involves reserving a specific sample of a dataset on which the model isn't trained. Later on, the model is tested on this sample to evaluate it.

Cross-validation is used to protect a model from overfitting, especially if the amount of data available is limited. It's also known as rotation estimation or out-of-sample testing and is mainly used in settings where the model's target is prediction.

**What's the difference between Covariance and Correlation?**

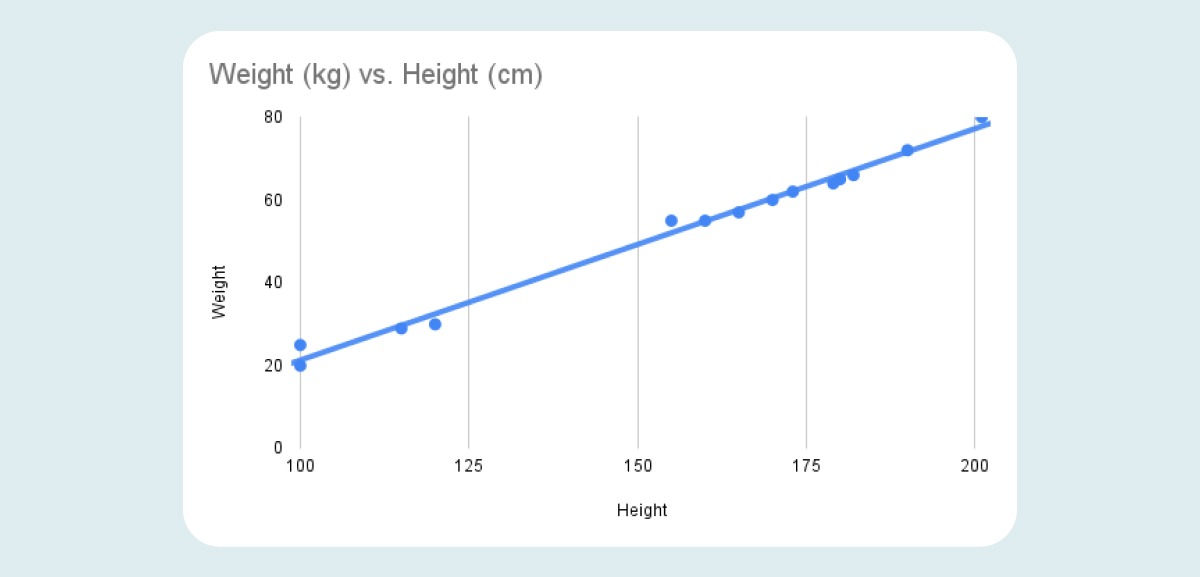
**Covariance** is a quantitative measure of the degree to which the deviation of one variable (X) from its mean is related to the deviation of another variable (Y) from its mean. To simplify, **covariance measures the joint variability of two random variables**. For example, if greater values of one variable tend to correspond with greater values of another variable, this suggests positive covariance. We’ll explore the different types of covariance shortly. First, let’s look at how covariance is calculated in mathematical terms.



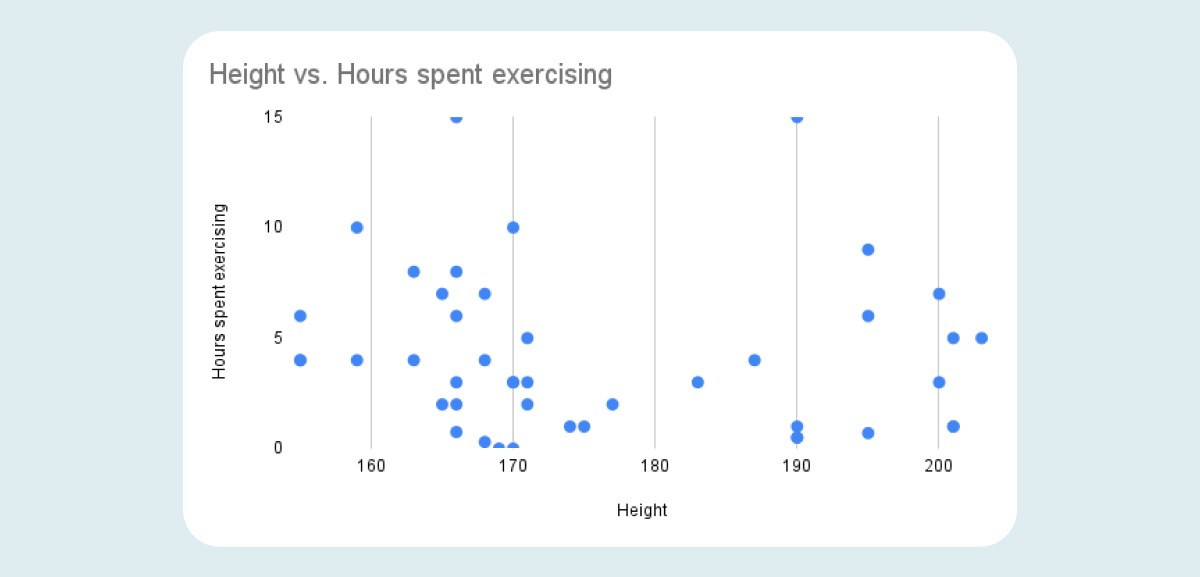
So far, we’ve established that covariance indicates the extent to which two random variables increase or decrease in tandem with each other. **Correlation** tells us both the strength and the direction of this relationship. Correlation is best used for multiple variables that express a linear relationship with one another. When we assume a correlation between two variables, we are essentially deducing that a change in one variable impacts a change in another variable. Correlation helps us to determine whether or not, and how strongly, changes in various variables relate to each other.



-1



1



0

**What's the difference between Feature Engineering vs. Feature Selection?**

**Feature Engineering** uses already modified features to create new ones, which will make it easier for any Machine Learning algorithm to understand and learn any pattern.

Refers to a process of selecting and transforming variables/features in your dataset when creating a predictive model using machine learning.

* Preparing the proper input dataset, compatible with the machine learning algorithm requirements.
* Improving the performance of machine learning models.

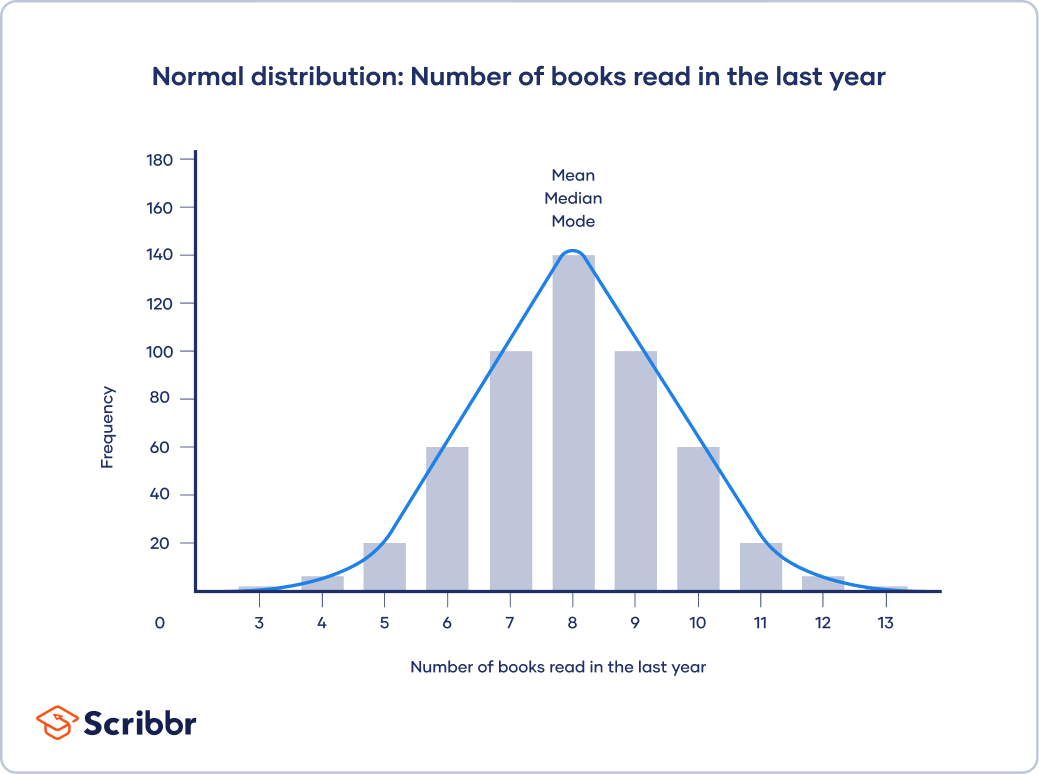
The **Feature Selection** learns the impact of each feature on your model and brings results. That way, you can create training, validation, and testing sets with the right resources. The process where you automatically or manually select the features that contribute the most to your prediction variable or output.

* It enables the machine learning algorithm to train faster.
* It reduces the complexity of a model and makes it easier to interpret.
* It improves the accuracy of a model if the right subset is chosen.
* It reduces overfitting.

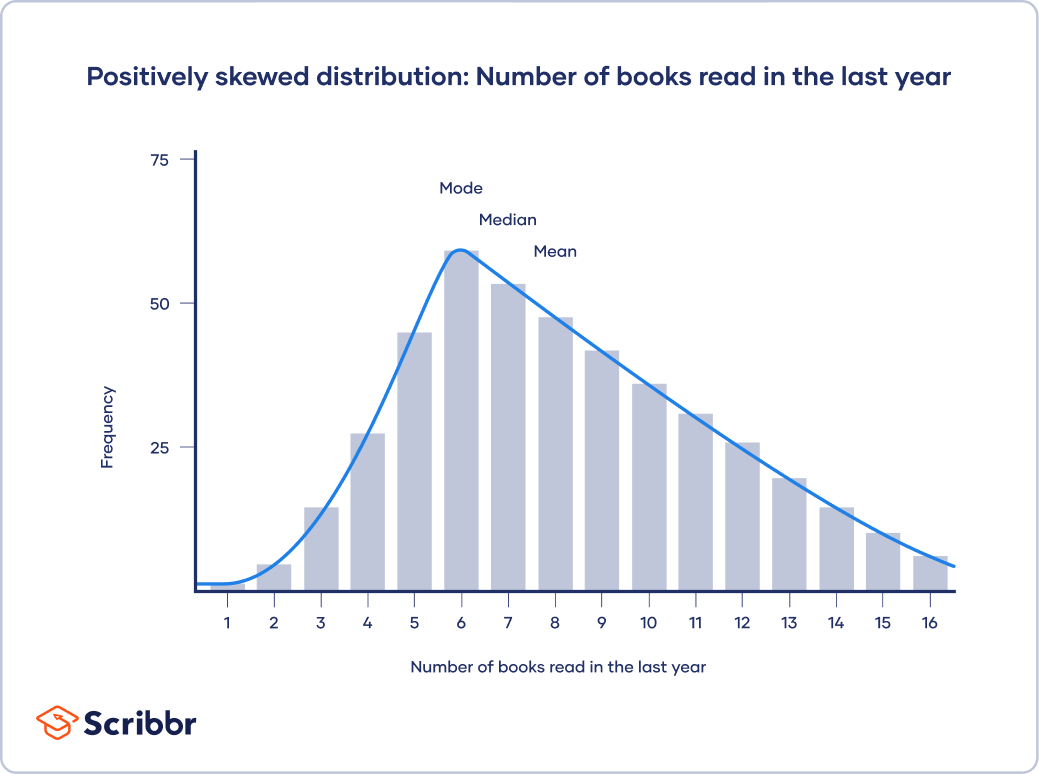
**What is Central Tendency?**

Measures of central tendency help you find the middle, or the average, of a dataset. The 3 most common measures of central tendency are the mode, median, and mean.

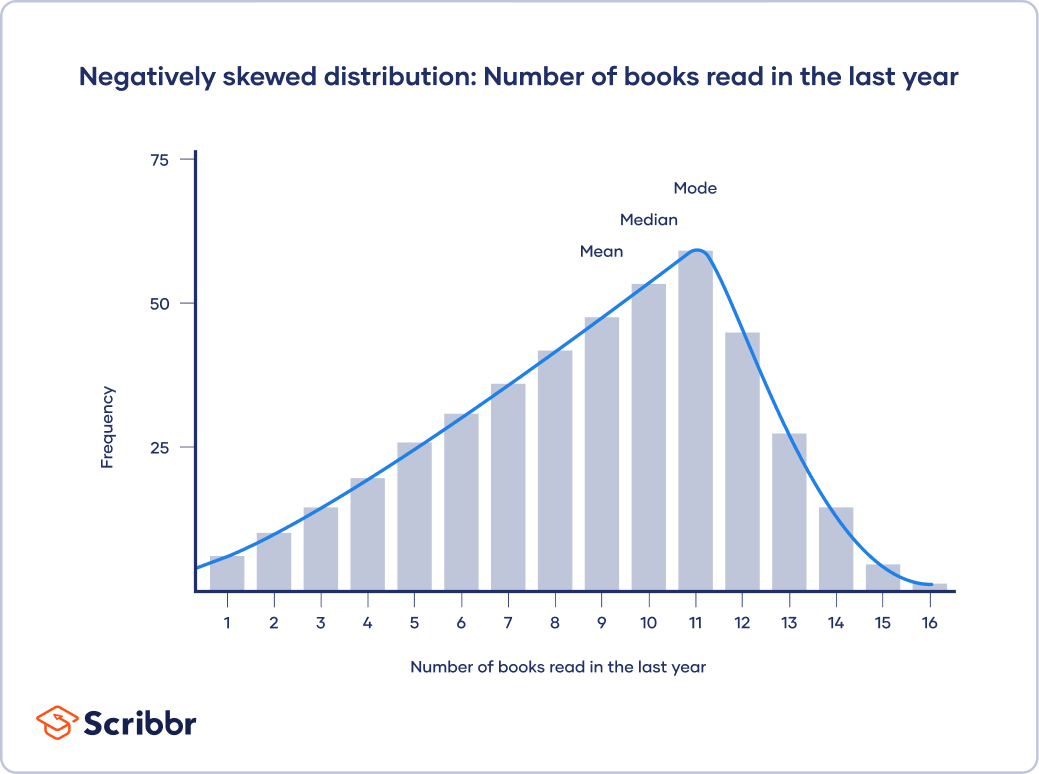
* Mode: the most frequent value.
* Median: the middle number in an ordered dataset.
* Mean: the sum of all values divided by the total number of values.



“The mean, median and mode are all equal; the central tendency of this dataset is 8.”



6



11

**How are outliers detected?**

**Sorting method**

You can sort quantitative variables from low to high and scan for extremely low or extremely high values. Flag any extreme values that you find.

This is a simple way to check whether you need to investigate certain data points before using more sophisticated methods.

**Using visualizations**

You can use software to visualize your data with a box plot, or a box-and-whisker plot, so you can see the data distribution at a glance. This type of chart highlights minimum and maximum values (the range), the median, and the interquartile range for your data.

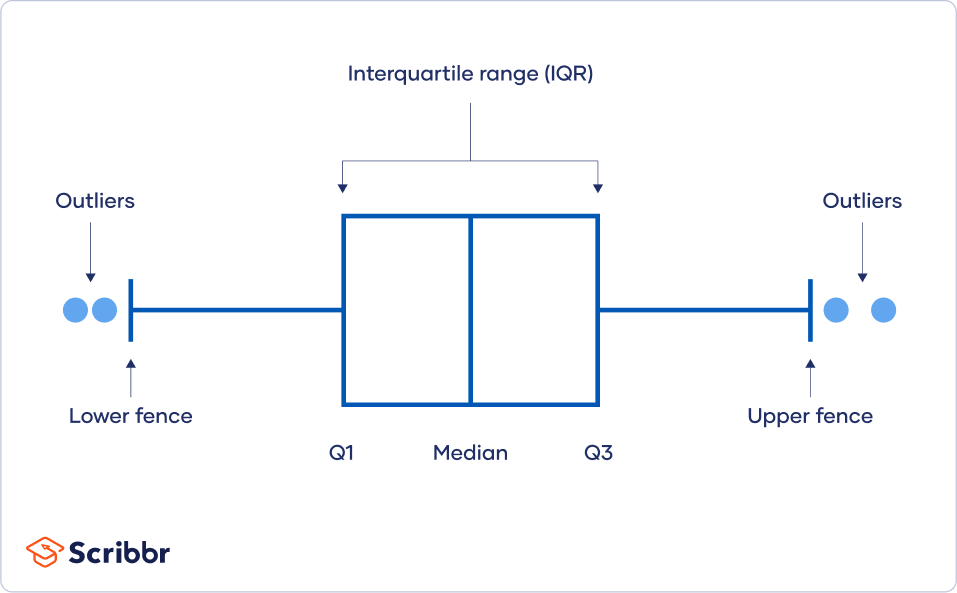
Many computer programs highlight an outlier on a chart with an asterisk, and these will lie outside the bounds of the graph.

**Statistical outlier detection**

Statistical outlier detection involves applying statistical tests or procedures to identify extreme values. You can convert extreme data points into z scores that tell you how many standard deviations away they are from the mean. If a value has a high enough or low enough z score, it can be considered an outlier. As a rule of thumb, values with a z score greater than 3 or less than –3 are often determined to be outliers.

**Using the interquartile range**

The interquartile range (IQR) tells you the range of the middle half of your dataset. You can use the IQR to create “fences” around your data and then define outliers as any values that fall outside those fences.



1. Sort your data from low to high

2. Identify the first quartile (Q1), the median, and the third quartile (Q3).

3. Calculate your IQR = Q3 – Q1

4. Calculate your upper fence = Q3 + (1.5 \* IQR)

5. Calculate your lower fence = Q1 – (1.5 \* IQR)

6. Use your fences to highlight any outliers, all values that fall outside your fences.

**What is data modelling?**

**Data modelling** is the process of creating a visual representation of either a whole information system or parts of it to communicate connections between data points and structures. The goal is to illustrate the types of data used and stored within the system, the relationships among these data types, the ways the data can be grouped and organized and its formats and attributes.

Data modelling makes it easier for developers, data architects, business analysts, and other stakeholders to view and understand relationships among the data in a database or data warehouse. In addition, it can:

Reduce errors in software and database development.

Increase consistency in documentation and system design across the enterprise.

Improve application and database performance.

Ease data mapping throughout the organization.

Improve communication between developers and business intelligence teams.

Ease and speed the process of database design at the conceptual, logical and physical levels.

**What is data visualization?**

**Data visualization** is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Additionally, it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion.

Some other advantages of data visualization include:

Easily sharing information.

Interactively explore opportunities.

Visualize patterns and relationships

Some other disadvantages include:

Biased or inaccurate information.

Correlation doesn’t always mean causation.

Core messages can get lost in translation.

**What is the difference between a heat map and a treemap?**

A **two-dimensional representation** of information with the help of colors is known as the **Heatmap**. These maps are used to visualize both simple and complex data. Heatmaps are frequently used in analyzing the patterns of consumer purchases

If you have a **large amount of highly structured data**, then the best option for visualization is a **Treemap**. The space in the visualization is split up into rectangles, sized and ordered according to the quantitative measures. The levels in the hierarchy are displayed as nested rectangles.

**When will you use a histogram and when will you use a bar chart? Explain with an example.**

A **bar chart**, or column chart, is a type of bar graph that's used to display categorical data. It allows you to easily compare variables across different qualitative categories. Double bar charts allow you to compare two similar data sets at the same time. You can use a bar chart to illustrate the following types of data:

Number of new employees per department

Number of clients you have per industry type

Sales numbers compared by quarter

A **histogram** is a type of frequency graph used to display statistical or quantitative data. It allows you to show the frequency or distribution of continuous data, like the length of store visits or the number of students involved in one or more extracurricular organizations.

Histograms include data ranges grouped into data bins, or intervals, on the x-axis, with frequency counts on the y-axis. You can use a histogram to illustrate the following types of data:

Amount of time visitors spend on your website

Average income by age range

Average number of properties available based on property value

**What is an outlier?**

An outlier is a value or point that differs substantially from the rest of the data.

Ex: extreme values that don’t fit the graph

**What type of data is box-plots usually used for? Why?**

A **Box Plot** is the visual representation of the statistical five-number summary of a given data set.

A Five Number Summary includes:

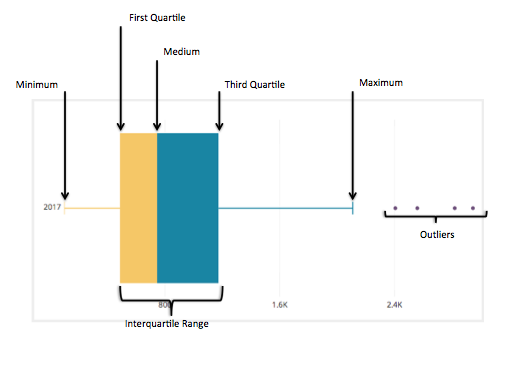
Minimum

First Quartile

Median (Second Quartile)

and Third Quartile

Maximum



A box plot allow us to quickly visualize the distribution of values in a dataset and see where the five-number summary values are located.

Side-by-side box plots allow us to visualize the differences between two or more distributions and compare the median values and the spread of values between distributions.

In box plots, outliers are typically represented by tiny circles that extend beyond either whisker. An observation is defined to be an outlier if it meets one of the following criteria:

**When do you use a boxplot and in what situation would you choose a boxplot over histograms?**

Histograms and box plots are graphical representations for the frequency of numeric data values. They aim to describe the data and explore the central tendency and variability before using advanced statistical analysis techniques. In this article, we will further discuss the similarities and differences between these two tools.

Although histograms are better in displaying the distribution of data, you can use a box plot to tell if the distribution is symmetric or skewed. In a symmetric distribution, the mean and median are nearly the same, and the two whiskers has almost the same length.

Histograms and box plots are very similar in that they both help to visualize and describe numeric data. Although histograms are better in determining the underlying distribution of the data, box plots allow you to compare multiple data sets better than histograms as they are less detailed and take up less space. It is recommended that you plot your data graphically before proceeding with further statistical analysis.