1. What are the key tasks involved in getting ready to work with machine learning modelling?

1. Gathering Data:

Data Gathering is the first step of the machine learning life cycle. The goal of this step is to identify and obtain all data-related problems.

In this step, we need to identify the different data sources, as data can be collected from various sources such as **files**, **database**, **internet**, or **mobile devices**. It is one of the most important steps of the life cycle. The quantity and quality of the collected data will determine the efficiency of the output. The more will be the data, the more accurate will be the prediction.

This step includes the below tasks:

* **Identify various data sources**
* **Collect data**
* **Integrate the data obtained from different sources**

By performing the above task, we get a coherent set of data, also called as a **dataset**. It will be used in further steps.

2. Data preparation

After collecting the data, we need to prepare it for further steps. Data preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training.

In this step, first, we put all data together, and then randomize the ordering of data.

This step can be further divided into two processes:

* **Data exploration:**  
  It is used to understand the nature of data that we have to work with. We need to understand the characteristics, format, and quality of data.  
  A better understanding of data leads to an effective outcome. In this, we find Correlations, general trends, and outliers.
* **Data pre-processing:**  
  Now the next step is preprocessing of data for its analysis.

3. Data Wrangling

Data wrangling is the process of cleaning and converting raw data into a useable format. It is the process of cleaning the data, selecting the variable to use, and transforming the data in a proper format to make it more suitable for analysis in the next step. It is one of the most important steps of the complete process. Cleaning of data is required to address the quality issues.

It is not necessary that data we have collected is always of our use as some of the data may not be useful. In real-world applications, collected data may have various issues, including:

* **Missing Values**
* **Duplicate data**
* **Invalid data**
* **Noise**

So, we use various filtering techniques to clean the data.

It is mandatory to detect and remove the above issues because it can negatively affect the quality of the outcome.

4. Data Analysis

Now the cleaned and prepared data is passed on to the analysis step. This step involves:

* **Selection of analytical techniques**
* **Building models**
* **Review the result**

The aim of this step is to build a machine learning model to analyze the data using various analytical techniques and review the outcome. It starts with the determination of the type of the problems, where we select the machine learning techniques such as **Classification**, **Regression**, **Cluster analysis**, **Association**, etc. then build the model using prepared data, and evaluate the model.

Hence, in this step, we take the data and use machine learning algorithms to build the model.

5. Train Model

Now the next step is to train the model, in this step we train our model to improve its performance for better outcome of the problem.

We use datasets to train the model using various machine learning algorithms. Training a model is required so that it can understand the various patterns, rules, and, features.

6. Test Model

Once our machine learning model has been trained on a given dataset, then we test the model. In this step, we check for the accuracy of our model by providing a test dataset to it.

Testing the model determines the percentage accuracy of the model as per the requirement of project or problem.

7. Deployment

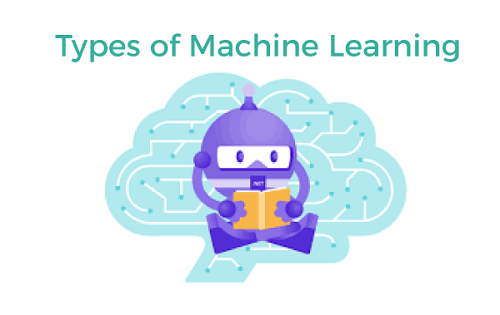
The last step of machine learning life cycle is deployment, where we deploy the model in the real-world system.

If the above-prepared model is producing an accurate result as per our requirement with acceptable speed, then we deploy the model in the real system. But before deploying the project, we will check whether it is improving its performance using available data or not. The deployment phase is similar to making the final report for a project.

2. What are the different forms of data used in machine learning? Give a specific example for each of them.

# Types of Machine Learning

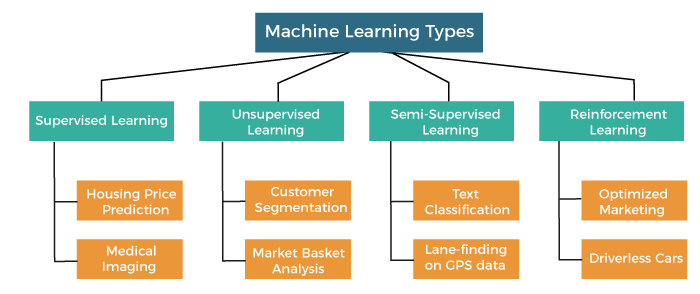
**Machine learning is a subset of AI, which enables the machine to automatically learn from data, improve performance from past experiences, and make predictions**. Machine learning contains a set of algorithms that work on a huge amount of data. Data is fed to these algorithms to train them, and on the basis of training, they build the model & perform a specific task.



These ML algorithms help to solve different business problems like Regression, Classification, Forecasting, Clustering, and Associations, etc.

Based on the methods and way of learning, machine learning is divided into mainly four types, which are:

1. Supervised Machine Learning
2. Unsupervised Machine Learning
3. Semi-Supervised Machine Learning
4. Reinforcement Learning



In this topic, we will provide a detailed description of the types of Machine Learning along with their respective algorithms:

## 1. Supervised Machine Learning

As its name suggests, [Supervised machine learning](https://www.javatpoint.com/supervised-machine-learning) is based on supervision. It means in the supervised learning technique, we train the machines using the "labelled" dataset, and based on the training, the machine predicts the output. Here, the labelled data specifies that some of the inputs are already mapped to the output. More preciously, we can say; first, we train the machine with the input and corresponding output, and then we ask the machine to predict the output using the test dataset.

Let's understand supervised learning with an example. Suppose we have an input dataset of cats and dog images. So, first, we will provide the training to the machine to understand the images, such as the **shape & size of the tail of cat and dog, Shape of eyes, colour, height (dogs are taller, cats are smaller), etc.** After completion of training, we input the picture of a cat and ask the machine to identify the object and predict the output. Now, the machine is well trained, so it will check all the features of the object, such as height, shape, colour, eyes, ears, tail, etc., and find that it's a cat. So, it will put it in the Cat category. This is the process of how the machine identifies the objects in Supervised Learning.

**The main goal of the supervised learning technique is to map the input variable(x) with the output variable(y).** Some real-world applications of supervised learning are **Risk Assessment, Fraud Detection, Spam filtering,** etc.

### Categories of Supervised Machine Learning

Supervised machine learning can be classified into two types of problems, which are given below:

* **Classification**
* **Regression**

### a) Classification

Classification algorithms are used to solve the classification problems in which the output variable is categorical, such as "**Yes" or No, Male or Female, Red or Blue, etc**. The classification algorithms predict the categories present in the dataset. Some real-world examples of classification algorithms are **Spam Detection, Email filtering, etc.**

Some popular classification algorithms are given below:

* **Random Forest Algorithm**
* **Decision Tree Algorithm**
* **Logistic Regression Algorithm**
* **Support Vector Machine Algorithm**

### b) Regression

Regression algorithms are used to solve regression problems in which there is a linear relationship between input and output variables. These are used to predict continuous output variables, such as market trends, weather prediction, etc.

Some popular Regression algorithms are given below:

* **Simple Linear Regression Algorithm**
* **Multivariate Regression Algorithm**
* **Decision Tree Algorithm**
* **Lasso Regression**

### Advantages and Disadvantages of Supervised Learning

**Advantages:**

* Since supervised learning work with the labelled dataset so we can have an exact idea about the classes of objects.
* These algorithms are helpful in predicting the output on the basis of prior experience.

**Disadvantages:**

* These algorithms are not able to solve complex tasks.
* It may predict the wrong output if the test data is different from the training data.
* It requires lots of computational time to train the algorithm.

### Applications of Supervised Learning

Some common applications of Supervised Learning are given below:

* **Image Segmentation:**  
  Supervised Learning algorithms are used in image segmentation. In this process, image classification is performed on different image data with pre-defined labels.
* **Medical Diagnosis:**  
  Supervised algorithms are also used in the medical field for diagnosis purposes. It is done by using medical images and past labelled data with labels for disease conditions. With such a process, the machine can identify a disease for the new patients.
* **Fraud Detection -** Supervised Learning classification algorithms are used for identifying fraud transactions, fraud customers, etc. It is done by using historic data to identify the patterns that can lead to possible fraud.
* **Spam detection -** In spam detection & filtering, classification algorithms are used. These algorithms classify an email as spam or not spam. The spam emails are sent to the spam folder.
* **Speech Recognition -** Supervised learning algorithms are also used in speech recognition. The algorithm is trained with voice data, and various identifications can be done using the same, such as voice-activated passwords, voice commands, etc.

## 2. Unsupervised Machine Learning

[Unsupervised learnin](https://www.javatpoint.com/unsupervised-machine-learning)g is different from the Supervised learning technique; as its name suggests, there is no need for supervision. It means, in unsupervised machine learning, the machine is trained using the unlabeled dataset, and the machine predicts the output without any supervision.

In unsupervised learning, the models are trained with the data that is neither classified nor labelled, and the model acts on that data without any supervision.

**The main aim of the unsupervised learning algorithm is to group or categories the unsorted dataset according to the similarities, patterns, and differences.** Machines are instructed to find the hidden patterns from the input dataset.

Let's take an example to understand it more preciously; suppose there is a basket of fruit images, and we input it into the machine learning model. The images are totally unknown to the model, and the task of the machine is to find the patterns and categories of the objects.

So, now the machine will discover its patterns and differences, such as colour difference, shape difference, and predict the output when it is tested with the test dataset.

### Categories of Unsupervised Machine Learning

Unsupervised Learning can be further classified into two types, which are given below:

* **Clustering**
* **Association**

### 1) Clustering

The clustering technique is used when we want to find the inherent groups from the data. It is a way to group the objects into a cluster such that the objects with the most similarities remain in one group and have fewer or no similarities with the objects of other groups. An example of the clustering algorithm is grouping the customers by their purchasing behaviour.

Some of the popular clustering algorithms are given below:

* **K-Means Clustering algorithm**
* **Mean-shift algorithm**
* **DBSCAN Algorithm**
* **Principal Component Analysis**
* **Independent Component Analysis**

### 2) Association

Association rule learning is an unsupervised learning technique, which finds interesting relations among variables within a large dataset. The main aim of this learning algorithm is to find the dependency of one data item on another data item and map those variables accordingly so that it can generate maximum profit. This algorithm is mainly applied in **Market Basket analysis, Web usage mining, continuous production**, etc.

Some popular algorithms of Association rule learning are **Apriori Algorithm, Eclat, FP-growth algorithm.**

### Advantages and Disadvantages of Unsupervised Learning Algorithm

**Advantages:**

* These algorithms can be used for complicated tasks compared to the supervised ones because these algorithms work on the unlabeled dataset.
* Unsupervised algorithms are preferable for various tasks as getting the unlabeled dataset is easier as compared to the labelled dataset.

**Disadvantages:**

* The output of an unsupervised algorithm can be less accurate as the dataset is not labelled, and algorithms are not trained with the exact output in prior.
* Working with Unsupervised learning is more difficult as it works with the unlabelled dataset that does not map with the output.

### Applications of Unsupervised Learning

* **Network Analysis:** Unsupervised learning is used for identifying plagiarism and copyright in document network analysis of text data for scholarly articles.
* **Recommendation Systems:** Recommendation systems widely use unsupervised learning techniques for building recommendation applications for different web applications and e-commerce websites.
* **Anomaly Detection:** Anomaly detection is a popular application of unsupervised learning, which can identify unusual data points within the dataset. It is used to discover fraudulent transactions.
* **Singular Value Decomposition:** Singular Value Decomposition or SVD is used to extract particular information from the database. For example, extracting information of each user located at a particular location.

## 3. Semi-Supervised Learning

**Semi-Supervised learning is a type of Machine Learning algorithm that lies between Supervised and Unsupervised machine learning**. It represents the intermediate ground between Supervised (With Labelled training data) and Unsupervised learning (with no labelled training data) algorithms and uses the combination of labelled and unlabeled datasets during the training period.

**A**lthough Semi-supervised learning is the middle ground between supervised and unsupervised learning and operates on the data that consists of a few labels, it mostly consists of unlabeled data. As labels are costly, but for corporate purposes, they may have few labels. It is completely different from supervised and unsupervised learning as they are based on the presence & absence of labels.

**To overcome the drawbacks of supervised learning and unsupervised learning algorithms, the concept of Semi-supervised learning is introduced**. The main aim of [semi-supervised learning](https://www.javatpoint.com/semi-supervised-learning) is to effectively use all the available data, rather than only labelled data like in supervised learning. Initially, similar data is clustered along with an unsupervised learning algorithm, and further, it helps to label the unlabeled data into labelled data. It is because labelled data is a comparatively more expensive acquisition than unlabeled data.

We can imagine these algorithms with an example. Supervised learning is where a student is under the supervision of an instructor at home and college. Further, if that student is self-analysing the same concept without any help from the instructor, it comes under unsupervised learning. Under semi-supervised learning, the student has to revise himself after analyzing the same concept under the guidance of an instructor at college.

### Advantages and disadvantages of Semi-supervised Learning

**Advantages:**

* It is simple and easy to understand the algorithm.
* It is highly efficient.
* It is used to solve drawbacks of Supervised and Unsupervised Learning algorithms.

**Disadvantages:**

* Iterations results may not be stable.
* We cannot apply these algorithms to network-level data.
* Accuracy is low.

## 4. Reinforcement Learning

**Reinforcement learning works on a feedback-based process, in which an AI agent (A software component) automatically explore its surrounding by hitting & trail, taking action, learning from experiences, and improving its performance.** Agent gets rewarded for each good action and get punished for each bad action; hence the goal of reinforcement learning agent is to maximize the rewards.

In reinforcement learning, there is no labelled data like supervised learning, and agents learn from their experiences only.

The [reinforcement learning](https://www.javatpoint.com/reinforcement-learning) process is similar to a human being; for example, a child learns various things by experiences in his day-to-day life. An example of reinforcement learning is to play a game, where the Game is the environment, moves of an agent at each step define states, and the goal of the agent is to get a high score. Agent receives feedback in terms of punishment and rewards.

Due to its way of working, reinforcement learning is employed in different fields such as **Game theory, Operation Research, Information theory, multi-agent systems.**

A reinforcement learning problem can be formalized using **Markov Decision Process(MDP).** In MDP, the agent constantly interacts with the environment and performs actions; at each action, the environment responds and generates a new state.

### Categories of Reinforcement Learning

Reinforcement learning is categorized mainly into two types of methods/algorithms:

* **Positive Reinforcement Learning:** Positive reinforcement learning specifies increasing the tendency that the required behaviour would occur again by adding something. It enhances the strength of the behaviour of the agent and positively impacts it.
* **Negative Reinforcement Learning:** Negative reinforcement learning works exactly opposite to the positive RL. It increases the tendency that the specific behaviour would occur again by avoiding the negative condition.

### Real-world Use cases of Reinforcement Learning

* **Video Games:**  
  RL algorithms are much popular in gaming applications. It is used to gain super-human performance. Some popular games that use RL algorithms are **AlphaGO** and **AlphaGO Zero**.
* **Resource Management:**  
  The "Resource Management with Deep Reinforcement Learning" paper showed that how to use RL in computer to automatically learn and schedule resources to wait for different jobs in order to minimize average job slowdown.
* **Robotics:**  
  RL is widely being used in Robotics applications. Robots are used in the industrial and manufacturing area, and these robots are made more powerful with reinforcement learning. There are different industries that have their vision of building intelligent robots using AI and Machine learning technology.
* **Text Mining**  
  Text-mining, one of the great applications of NLP, is now being implemented with the help of Reinforcement Learning by Salesforce company.

### Advantages and Disadvantages of Reinforcement Learning

**Advantages**

* It helps in solving complex real-world problems which are difficult to be solved by general techniques.
* The learning model of RL is similar to the learning of human beings; hence most accurate results can be found.
* Helps in achieving long term results.

**Disadvantage**

* RL algorithms are not preferred for simple problems.
* RL algorithms require huge data and computations.
* Too much reinforcement learning can lead to an overload of states which can weaken the results.

The curse of dimensionality limits reinforcement learning for real physical systems.

3. Distinguish:

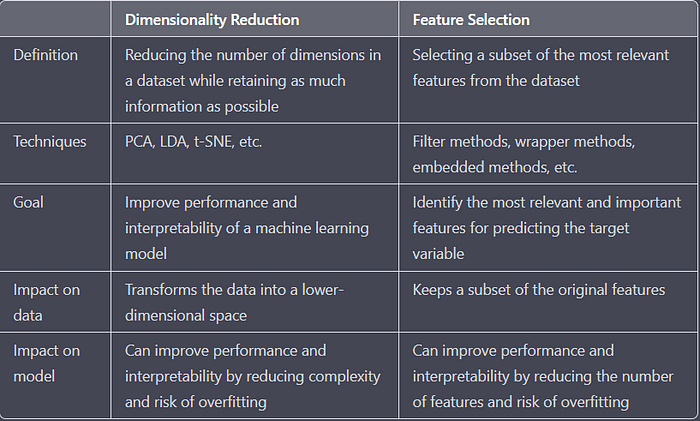
1. Numeric vs. categorical attributes

The difference between categorical data vs numerical data

Many things are different between these 2 types of data. Let’s find out what and how they are different below:

|  |  |  |  |
| --- | --- | --- | --- |
| No | Features | Categorical data | Numerical data |
| 1 | Definition | Categorical data can be stored and identified by names or labels. | Numerical data are numbers, not words or descriptions. |
| 2 | Alias | Because it qualifies data before categorizing it, it is sometimes referred to as qualitative data. | Quantitative data represents numerical values for arithmetic processes. |
| 3 | Examples | Define gender.   * Male * Female * Other | Test score out of 20?   * Below 5 * 5-10 * 10-15 * 15-20 * 20 |
| 4 | Types | Nominal data and Ordinal data. | Discrete data and Continuous data. |
| 5 | Characteristics | * There is no scale of orders. * Natural language description * Can take numerical numbers but has qualitative characteristics * Can be represented graphically using pie charts and bar charts. | * Possesses a logical scale * Using a description that is not in natural language * Takes numbers and numbers as inputs. * Bar charts and pie charts can be used to visualize this |
| 6 | User-friendly design | Long surveys are a possibility and may turn off responders. | Survey interaction is quick and short, reducing abandonment. |
| 7 | Data collection method | * Nominal data: open-ended questions * Ordinal data: multiple-choice questions | Mostly multiple-choice, sometimes open-ended questions. |
| 8 | Data collection tools | Questionnaires, surveys, and interviews | Questionnaires, surveys, interviews, focus groups and observations |
| 9 | Uses | Used when a survey demands respondents personal information, opinions, and experiences. Used in business research | Statistical calculations based on arithmetic performance. |
| 10 | Compatibility | It’s incompatible with most statistical approaches. Thus researchers avoid it. | It supports most statistical calculations. |
| 11 | Visualization | Bar and pie charts alone. | Bar graphs, pie charts, and scatter plots can be used. |
| 12 | Structure | Unstructured data Like Google, Bing, etc., it can index data. | Since it is structured, it can be easily arranged and understood. |

2. Feature selection vs. dimensionality reduction



4. Make quick notes on any two of the following:

1. The histogram

**Histograms** helps visualizing and comprehending the data distribution. The article aims to provide comprehensive overview of **histogram** and its **interpretation**.

## What is Histogram?

[Histograms](https://www.geeksforgeeks.org/histograms-in-r-language/) are graphical representations of data distributions. They consist of bars, each representing the frequency or count of observations falling within specific intervals, known as bins. We can also say a histogram is a variation of a bar chart in which data values are grouped together and put into different classes. This grouping enables you to see how frequently data in each class occur in the dataset.

The histogram graphically shows the following:

* Frequency of different data points in the dataset.
* Location of the center of data.
* The spread of dataset.
* [Skewness](https://www.geeksforgeeks.org/difference-between-skewness-and-kurtosis/)/[variance](https://www.geeksforgeeks.org/mathematics-mean-variance-and-standard-deviation/) of dataset.
* Presence of [outliers](https://www.geeksforgeeks.org/machine-learning-outlier/) in the dataset.

The features provide a strong indication of the proper distributional model in the data. The probability plot or a [goodness-of-fit test](https://www.geeksforgeeks.org/how-to-perform-a-chi-square-goodness-of-fit-test-in-python/) can be used to verify the distributional model.

The histogram contains the following axes:

* **Vertical Axis**: Frequency/count of each bin.
* **Horizontal Axis**: List of bins/categories.

## How histogram works?

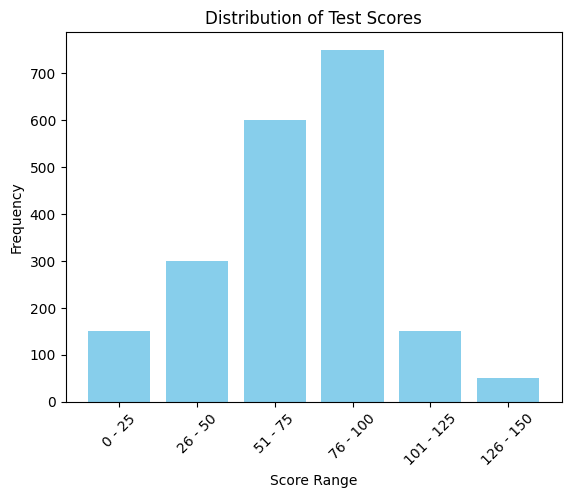
The histogram works by organizing and visualizing the distribution of data into intervals or bins along a continuous scale.

* The range of data values is divided into intervals called “bins.” The number of bins and their widths can be predefined or determined algorithmically based on the range and distribution of the data.
* Each data point in the dataset is assigned to a corresponding bin based on its value. As data points are assigned to bins, the frequency or count of data points falling within each bin is calculated.
* The histogram is constructed by plotting the bins along the x-axis and the frequencies (or densities) along the y-axis. Each bin is represented by a bar, and the height of the bar corresponds to the frequency of data points in that bin.

By examining the histogram, you can gain insights into the distribution of the data. You can identify patterns, trends, central tendencies, variability, outliers, and other characteristics of the dataset. For example, a **symmetric bell-shaped histogram** suggests a [normal distribution](https://www.geeksforgeeks.org/normal-distribution/), while skewed histograms indicate asymmetry in the data.

Suppose you’re analyzing the distribution of scores on a standardized test. You have data for 2000 students, and you want to visualize how many students scored within different score ranges. For this you can create a histogram using the following data.

| **Score Range** | **Frequency** |
| --- | --- |
| **0-25** | 150 |
| **26-50** | 300 |
| **51-75** | 600 |
| **76-100** | 750 |
| **101-125** | 150 |
| **126-150** | 50 |



*Histogram*

The histogram show that the data is normally distributed, and the students have mostly score between 76-100. This histogram displays the frequency of students falling within different score ranges on the standardized test. Each bar represents a score range, and the height of the bar represents the frequency of students in that range. By customizing the x-axis intervals and the labels, you can effectively visualize the distribution of test scores. Additionally, you can further customize the histogram by changing the y-axis to display percentages or density if needed.

## Histogram and its Interpretation

### Normal Histogram

Normal histogram is a classical bell-shaped histogram with most of the frequency counts focused on the middle with diminishing tails and there is symmetry with respect to the median. Since the normal distribution is most commonly observed in real-world scenarios, you are most likely to find these. In Normally distributed histogram mean is almost equal to median.

We have plotted a normal distribution graph.

* The peak of the curve represents the mean of the dataset.
* The normal distribution graph is symmetric.

### Non-normal Short-tailed/ long-tailed histogram

In short-tailed distribution tail approaches 0 very fast, as we move from the median of data, In the long-tailed histogram, the tail approaches 0 slowly as we move far from the [median](https://www.geeksforgeeks.org/problems/find-median-in-a-stream-1587115620/1). Here, we refer tail as the extreme regions in the histogram where most of the data is not concentrated and this is on both sides of the peak.

### Bimodal Histogram

A[mode](https://www.geeksforgeeks.org/mode/)of data represents the most common values in the histogram (i.e. peak of the histogram. A bimodal histogram represents that there are two peaks in the histogram. The histogram can be used to test the unimodality of data. The bimodality (or for instance non-unimodality) in the dataset represents that there is something wrong with the process. Bimodal histogram many one or both of two characters: Bimodal normal distribution and symmetric distribution.

### Skewed Left/Right Histogram

[Skewed histogram](https://www.geeksforgeeks.org/skewness-and-kurtosis-in-r-programming/) is those where the one-side tail is quite clearly longer than the other-side tail. A right-skewed histogram means that the right-sided tail of the peak is more stretched than its left and vice-versa for the left-sided. In a left-skewed histogram, the mean is always lesser than the median, while in a right-skewed histogram mean is greater than the histogram.

### Uniform Histogram

In uniform histogram, each bin contains approximately the same number of counts (frequency). The example of uniform histogram is such as a die is rolled n (n>>30) number of times and record the frequency of different outcomes

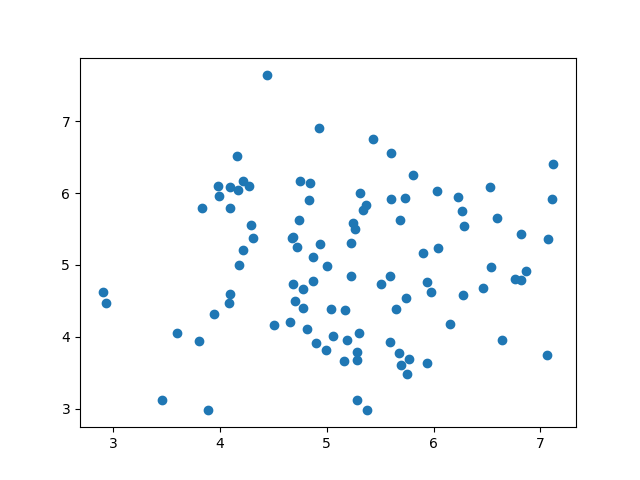
### ****Normal Distribution with an Outlier****

This histogram is similar to normal histogram except it contains an outlier where the count/ probability of outcome is substantive. This is mostly due to some system errors in process, which led to faulty generation of products etc.

2. Use a scatter plot

## Scatter Plot

A scatter plot is a diagram where each value in the data set is represented by a dot.



The Matplotlib module has a method for drawing scatter plots, it needs two arrays of the same length, one for the values of the x-axis, and one for the values of the y-axis:

x = [5,7,8,7,2,17,2,9,4,11,12,9,6]

y = [99,86,87,88,111,86,103,87,94,78,77,85,86]

The x array represents the age of each car.

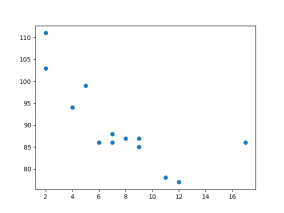
The y array represents the speed of each car.

### Example[Get your own Python Server](https://www.w3schools.com/python/python_server.asp)

Use the scatter() method to draw a scatter plot diagram:

import matplotlib.pyplot as plt  
  
x = [5,7,8,7,2,17,2,9,4,11,12,9,6]  
y = [99,86,87,88,111,86,103,87,94,78,77,85,86]  
  
plt.scatter(x, y)  
plt.show()

### Result:



[Run example »](https://www.w3schools.com/python/trypython.asp?filename=demo_ml_scatterplot)

### Scatter Plot Explained

The x-axis represents ages, and the y-axis represents speeds.

What we can read from the diagram is that the two fastest cars were both 2 years old, and the slowest car was 12 years old.

**Note:** It seems that the newer the car, the faster it drives, but that could be a coincidence, after all we only registered 13 cars.

## Random Data Distributions

In Machine Learning the data sets can contain thousands-, or even millions, of values.

You might not have real world data when you are testing an algorithm, you might have to use randomly generated values.

As we have learned in the previous chapter, the NumPy module can help us with that!

Let us create two arrays that are both filled with 1000 random numbers from a normal data distribution.

The first array will have the mean set to 5.0 with a standard deviation of 1.0.

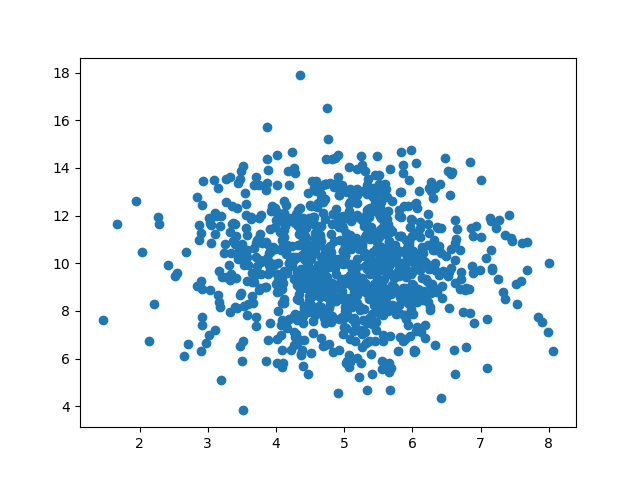
The second array will have the mean set to 10.0 with a standard deviation of 2.0:

### Example

A scatter plot with 1000 dots:

import numpy  
import matplotlib.pyplot as plt  
  
x = numpy.random.normal(5.0, 1.0, 1000)  
y = numpy.random.normal(10.0, 2.0, 1000)  
  
plt.scatter(x, y)  
plt.show()

### Result:



[Run example »](https://www.w3schools.com/python/trypython.asp?filename=demo_ml_scatterplot2)

### Scatter Plot Explained

We can see that the dots are concentrated around the value 5 on the x-axis, and 10 on the y-axis.

We can also see that the spread is wider on the y-axis than on the x-axis.

3.PCA (Personal Computer Aid)

# Principal Component Analysis(PCA)

**Last Updated :**06 Dec, 2023

As the number of features or dimensions in a dataset increases, the amount of data required to obtain a statistically significant result increases exponentially. This can lead to issues such as overfitting, increased computation time, and reduced accuracy of machine learning models this is known as the curse of dimensionality problems that arise while working with high-dimensional data.

As the number of dimensions increases, the number of possible combinations of features increases exponentially, which makes it computationally difficult to obtain a representative sample of the data and it becomes expensive to perform tasks such as clustering or classification because it becomes. Additionally, some [machine learning](https://www.geeksforgeeks.org/machine-learning/) algorithms can be sensitive to the number of dimensions, requiring more data to achieve the same level of accuracy as lower-dimensional data.

To address the [curse of dimensionality](https://www.geeksforgeeks.org/videos/curse-of-dimensionality-in-machine-learning/), [Feature engineering](https://www.geeksforgeeks.org/what-is-feature-engineering/)techniques are used which include feature selection and feature extraction. [Dimensionality reduction](https://www.geeksforgeeks.org/dimensionality-reduction/) is a type of feature extraction technique that aims to reduce the number of input features while retaining as much of the original information as possible.

In this article, we will discuss one of the most popular dimensionality reduction techniques i.e. Principal Component Analysis(PCA).

## What is Principal Component Analysis(PCA)?

[Principal Component Analysis](https://www.geeksforgeeks.org/principal-component-analysis-with-python/)(PCA) technique was introduced by the mathematician **Karl Pearson** in 1901**.** It works on the condition that while the data in a higher dimensional space is mapped to data in a lower dimension space, the variance of the data in the lower dimensional space should be maximum.

* **Principal Component Analysis (PCA)**is a statistical procedure that uses an orthogonal transformation that converts a set of correlated variables to a set of uncorrelated variables.PCA is the most widely used tool in exploratory data analysis and in machine learning for predictive models. Moreover,
* Principal Component Analysis (PCA) is an [unsupervised learning](https://www.geeksforgeeks.org/supervised-unsupervised-learning/) algorithm technique used to examine the interrelations among a set of variables. It is also known as a general factor analysis where regression determines a line of best fit.
* The main goal of Principal Component Analysis (PCA) is to reduce the dimensionality of a dataset while preserving the most important patterns or relationships between the variables without any prior knowledge of the target variables.

Principal Component Analysis (PCA) is used to reduce the dimensionality of a data set by finding a new set of variables, smaller than the original set of variables, retaining most of the sample’s information, and useful for the [regression and classification](https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning/) of data.

1. Principal Component Analysis (PCA) is a technique for dimensionality reduction that identifies a set of orthogonal axes, called principal components, that capture the maximum variance in the data. The principal components are linear combinations of the original variables in the dataset and are ordered in decreasing order of importance. The total variance captured by all the principal components is equal to the total variance in the original dataset.
2. The first principal component captures the most variation in the data, but the second principal component captures the maximum [variance](https://www.geeksforgeeks.org/python-statistics-variance/) that is [orthogonal](https://www.geeksforgeeks.org/orthogonal-and-orthonormal-vectors-in-linear-algebra/) to the first principal component, and so on.
3. Principal Component Analysis can be used for a variety of purposes, including data visualization, feature selection, and data compression. In data visualization, PCA can be used to plot high-dimensional data in two or three dimensions, making it easier to interpret. In feature selection, PCA can be used to identify the most important variables in a dataset. In data compression, PCA can be used to reduce the size of a dataset without losing important information.
4. In Principal Component Analysis, it is assumed that the information is carried in the variance of the features, that is, the higher the variation in a feature, the more information that features carries.

Overall, PCA is a powerful tool for data analysis and can help to simplify complex datasets, making them easier to understand and work with.

## Step-By-Step Explanation of PCA (Principal Component Analysis)

### Step 1: Standardization

First, we need to [standardize](https://www.geeksforgeeks.org/normalization-vs-standardization/) our dataset to ensure that each variable has a mean of 0 and a standard deviation of 1.

Here,

* is the mean of independent features
* is the [standard deviation](https://www.geeksforgeeks.org/mathematics-mean-variance-and-standard-deviation/) of independent features

### Step2: Covariance Matrix Computation

[Covariance](https://www.geeksforgeeks.org/mathematics-covariance-and-correlation/) measures the strength of joint variability between two or more variables, indicating how much they change in relation to each other. To find the covariance we can use the formula:

The value of covariance can be positive, negative, or zeros.

* Positive: As the x1 increases x2 also increases.
* Negative: As the x1 increases x2 also decreases.
* Zeros: No direct relation

### Step 3: Compute Eigenvalues and Eigenvectors of Covariance Matrix to Identify Principal Components

Let A be a square nXn matrix and X be a non-zero vector for which

for some scalar values . then  is known as the [eigenvalue](https://www.geeksforgeeks.org/how-to-compute-the-eigenvalues-and-right-eigenvectors-of-a-given-square-array-using-numpy/) of matrix A and X is known as the [eigenvector](https://www.geeksforgeeks.org/applications-of-eigenvalues-and-eigenvectors/) of matrix A for the corresponding eigenvalue.

It can also be written as :

where I am the identity matrix of the same shape as matrix A. And the above conditions will be true only if  will be non-invertible (i.e. singular matrix). That means,

From the above equation, we can find the eigenvalues \lambda, and therefore corresponding eigenvector can be found using the equation .

5. Why is it necessary to investigate data? Is there a discrepancy in how qualitative and quantitative data are explored?

### Understanding Discrepancies between Quantitative and Qualitative Data in Surveys

* [**SURVEY**](javascript:void(0))
* [**TIPS**](javascript:void(0))

There may be times when the quantitative results yield a different outcome than the qualitative results. For example, in a 360 survey, the competency communication may have a positive score, yet the open-ended comments for that same competency may have negative undertone. A survey that includes both quantitative and qualitative questions can yield different results due to the inherent differences in the types of data collected and the analysis methods used for each.

Quantitative questions typically involve closed-ended responses where respondents choose from pre-determined options or provide numerical ratings. These questions focus on measurable aspects and generate numerical data. The analysis of quantitative data involves statistical techniques such as averages, percentages, correlations, and regression analysis. This type of analysis aims to identify patterns, relationships, and trends within the data.

On the other hand, qualitative questions involve open-ended responses, allowing respondents to provide detailed and subjective information. These questions delve into individuals' experiences, opinions, beliefs, and perceptions. The analysis of qualitative data involves a process of coding and categorizing the responses, looking for recurring themes and patterns. It often involves techniques like content analysis, thematic analysis, or discourse analysis.

### The differences in results arise from several factors:

**Data representation:**Quantitative data is typically summarized using numerical values and statistical measures, such as means, standard deviations, or percentages. This allows for easy comparison and generalization. In contrast, qualitative data is represented through descriptive narratives or quotes, providing richer and contextualized insights but making it challenging to summarize or generalize.

**Different data collection methods:** Quantitative questions typically employ structured surveys with closed-ended response options, while qualitative questions often involve open-ended responses. These methods capture distinct types of information and may elicit different perspectives from respondents (Creswell, 2013).

**Analysis techniques:**Quantitative data is often analyzed using statistical methods to identify patterns, correlations, and significant relationships (Bryman, 2016). On the other hand, qualitative data requires thematic or content analysis to identify recurring themes, codes, and categories (Creswell, 2013). The differences in analysis techniques can lead to distinct interpretations and findings.

**Perceptual differences:** Quantitative responses in a 360 survey often involve rating scales or numerical assessments that provide a quantitative measure of leadership competencies. These ratings can be influenced by various factors, such as rater bias, differences in perception, or personal relationships between the leader and raters. On the other hand, qualitative responses in the form of open-ended comments allow raters to provide subjective insights, examples, or anecdotes about the leader's behaviors or performance. These qualitative responses can capture more nuanced and contextualized information, highlighting specific incidents or patterns that may not be apparent in numerical ratings.

Due to these differences, quantitative and qualitative questions may generate contrasting findings. For instance, quantitative data might reveal statistical differences between groups or highlight numerical trends, while qualitative data may provide rich contextual details, individual perspectives, or unexpected insights that cannot be captured through numbers alone.

### What you can do if you see differences:

* Dig deeper. Interview survey participants to gain a better understanding of why the difference in scores may have happened.
* Understand that there may be survey takers that give the highest ratings across all questions that could cause that discrepancy between the quantitative and qualitative questions.
* Focus on the big picture. For example, look for overall patterns.

#### ****About EDA, Inc.****

EDA, Inc., a world-class human capital firm with powerful Employee Experience platform, providing SaaS and AI solutions, premium executive development services and leading industry research. EDA’s services include.

* C-Suite & Executive Coaching,
* Custom Executive Development,
* Succession Planning,
* Assessment & Selection Services, and
* High-Potential Development.

It is our pleasure and passion to help organizational leaders create conducive workforce environments so that they can attract, retain, and develop top talent and advance their mission and strategy.

6. What are the various histogram shapes? What exactly are ‘bins'?

## ****What is a Histogram?****

A histogram represents values in a group of data that appear as vertical bars side-by-side. The height of each bar tells how many times (frequency or count) these values in that range occur in the data. It is a visual summary of a dataset's distribution and frequency of values.

Components of a histogram include

* **Bins or Intervals**- Bins in histogram represent the intervals or ranges into which the data is divided. Each bar in the histogram defines the frequency of data points within a particular bin.
* **Bars** - The bars are the vertical columns representing the frequency or count of data points falling within each bin. The height of each bar corresponds to the frequency of data points in that bin.
* **X-axis and Y-axis** -  The x-axis represents the intervals or bins, while the y-axis represents the frequency or count of data points within each interval.
* **Title and Labels** - A histogram should have a title that describes the data being represented. Additionally, it should have labels for the x-axis and y-axis to provide context and clarity.

## ****Basic Histogram Example****

Given this basic histogram definition, let’s look at one example.

Let's say you have a list of exam scores for a class of students as a dataset.

70, 85, 92, 78, 89, 92, 95, 78, 85, 90

Each number (70, 85, 92, etc.) is a value. When constructing a histogram for these exam scores, you might group them into intervals like 70-79, 80-89, and 90-99. The histogram would then show how many students scored in each of these score ranges.

**Stay updated with the latest blogs on online courses and skills**

Enter Mobile Number

Register Now

So, the histogram helps you visualise the distribution of scores and how frequently different score ranges occur in the dataset.

## ****How to Create a Histogram Step-by-Step****

How would you make this histogram of exam scores for a class of students?

Using the example above, let’s create a histogram in Google Spreadsheets. You could similarly follow the same steps in [MS Excel](https://www.shiksha.com/it-software/ms-excel-chp).

**Step 1**: Open Google Sheets

Open Google Sheets and create a new sheet or use an existing one.

**Step 2**: Enter Data

In a column, enter the exam scores. Let's assume you enter them in column A starting from A1.

|  |
| --- |
| Exam Scores |
| 70 |
| 85 |
| 92 |
| 78 |
| 89 |
| 92 |
| 95 |
| 78 |
| 85 |
| 90 |

**Step 3**: Create Bins

In another column, create bins (score ranges). For example, you can use the following bins.

|  |
| --- |
| Score Ranges |
| 70-79 |
| 80-89 |
| 90-99 |

**Step 4**: Count Scores in Bins

In the adjacent column, use the COUNTIFS function to count how many scores fall into each bin. Assuming you put the bins in column B and the scores in column A.

Type or paste the following in each row.

=COUNTIFS(A:A,">=70",A:A,"<=79")

=COUNTIFS(A:A,">=80",A:A,"<=89")

=COUNTIFS(A:A,">=90",A:A,"<=99")

Results will show this way, corresponding to each row.

|  |
| --- |
| 3 |
| 3 |
| 4 |

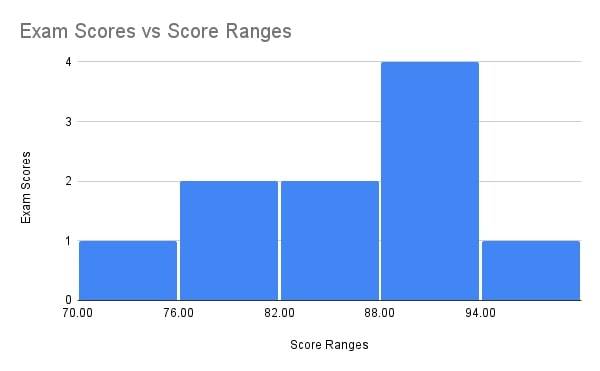
You could additionally look into our guide to [COUNTIF in Excel](https://www.shiksha.com/online-courses/articles/countif-in-excel/)

**Step 5**: Create a Bar Chart

Select the data in the bins and the corresponding counts. Then, go to Insert > Chart. Choose 'Chart type' as 'Bar chart'.

Adjust the chart settings if needed, such as giving it a title, labelling axes, etc.

Now, you've successfully created a histogram in Google Sheets showing how many students scored in each score range.



As you can see, the exam scores on the y-axis show the frequency or count of the values that fall into each bar/interval/bin. The score ranges on the x-axis are the type of data you want to measure, where 70 to 76 is one bin, and 88 to 94 is another.

## ****Histogram in Statistics for Analysis****

A histogram in statistics is a visual representation of the frequency distribution of a dataset. It is particularly useful for displaying the central tendency and the spread of continuous data. The horizontal axis of the histogram represents the continuous data values grouped into the specified bins. The vertical axis indicates the frequency of occurrence for each bin.

By examining the shape and pattern of the histogram, one can easily identify the central tendency of the dataset, observe its distribution, and discern any patterns or trends present.

Now, let’s focus on learning the key areas mentioned with regard to statistics (and the same example of exam scores in a class from above).

### ****Frequency Distribution****

Frequency distribution refers to how often the data appears in the dataset.

The frequency distribution of the exam scores in this very histogram will show that

* 3 students got marks between 70 and less than 79
* 3 students scored between the range of 80 and 89
* 4 students scored above 90 and less than 99

The histogram displays the distribution of exam scores, showing how many students achieved scores within specific ranges (bins or intervals). It indicates that more students fall into certain score ranges than others.

To illustrate the concept further, the main components of frequency distribution would be  
**Data Points**: These are the individual pieces of information in a dataset. For instance, the data points are 70, 85, 92, 78, 89, 92, 95, 78, 85, 90. In the context of frequency distribution, these values are grouped or categorised.

**Frequency**: This is the count of how many times each value or range of values occurs in the dataset. It represents how often each category appears. For instance, the score 78 appears twice, and the score 85 appears twice.

**Bins or Intervals**: In the context of continuous data, values are often grouped into intervals or bins. The frequency distribution then shows how many values fall into each interval. For instance, the bins could be defined as 70-79, 80-89, and 90-99.

If you want to learn more about it, you may as well take the free [Frequency Distribution course](https://www.shiksha.com/online-courses/frequency-distribution-course-grlel456) from Great Learning. This short course covers the basics of the topic, all within an hour.

### ****Central Tendency****

Central tendency is a statistical measure, representing the central or typical value of a dataset.

The main measures of central tendency are the mean, median, and mode.

Let's use the example dataset of exam scores along with its frequency distribution:

Example Dataset: 70,85,92,78,89,92,95,78,85,90

#### ****Mean (Average)****

The mean is calculated by adding up all the values and dividing by the total number of values.

**Example**.

Formula for Mean

(70+85+92+78+89+92+95+78+85+90)/10 = 85.4

Go check out our blog on the [Mean Formula](https://www.shiksha.com/online-courses/articles/mean-statistics-formula/) too!

#### ****Median****

The median is the middle value when the data is arranged in ascending or descending order.

If there's an even number of values, the median is the average of the two middle values.

**Example.**

Arranging the scores in ascending order-

70,78,78,85,85,89,90,92,92,95. The median is 87

#### ****Mode****

The mode is the value that appears most frequently in the dataset.

Example: The mode of this dataset is 92 as it appears twice, more than any other score.

Central tendency measures are often used in statistical models. For example, the mean is a key parameter in many statistical models, and understanding its value helps in making predictions based on the model.

You may check courses like [Measures of Central Tendency](https://www.shiksha.com/online-courses/measures-of-central-tendency-course-grlel205), another useful and free course from Great Learning.

## ****Types of Histograms****

Moving on, let’s help elaborate the types of histogram.

The shape and characteristics of histograms can vary based on the distribution of the data. Different types of distributions result in different types of histograms.

### ****Normal Distribution****

When the data is symmetrically distributed around the mean, the histogram will have a bell-shaped curve with the highest frequency at the mean and symmetrical tails on both sides.

### ****Skewed Distribution****

In a skewed distribution, the data is not symmetric and tends to have a longer tail on one side. A positively skewed distribution has a tail on the right side, while a negatively skewed distribution has a tail on the left side.

### ****Bimodal Distribution****

Bimodal distributions have two distinct peaks, indicating the presence of two different groups or conditions within the data.

### ****Uniform Distribution****

In a uniform distribution, all values have approximately the same frequency, resulting in a rectangular-shaped histogram.

### ****Exponential Distribution****

An exponential distribution often results in a histogram with a rapidly decreasing frequency as the values increase, creating a skewed, right-tailed shape.

Understanding the type of distribution in the dataset is essential for selecting the appropriate analysis techniques and drawing meaningful conclusions from the histogram. Each type of distribution provides valuable information about the underlying data and can offer insights into the nature of the variables being studied.

When interpreting histograms, it's important to consider the specific characteristics of the distribution and how they impact the shape and appearance of the histogram. This understanding allows for a more comprehensive analysis of the dataset.

7. How do we deal with data outliers?

# Detecting and Treating Outliers | Treating the odd one out!

## Introduction

One of the most important steps as part of [data preprocessing i](https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/)s detecting and treating the outliers as they can negatively affect the statistical analysis and the training process of a [machine learning algorithm](https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/)resulting in lower accuracy. In this article, we will be discussing how to handle outliers!

Wow, these are lovely! Wait, where does this yellow Tulip come from?

#### Learning Objectives:

* Understanding the definition and characteristics of outliers in data
* Recognizing why outliers occur and their potential sources (e.g., variability, experimental errors)
* Grasping the impact of outliers on statistical measures (mean, median, mode) and model performance

This article was published as a part of the [Data Science Blogathon.](https://datahack.analyticsvidhya.com/blogathon/)

## Table of contents

## What are Outliers?

We all have heard of the idiom ‘odd one out which means something unusual in comparison to the others in a group. Similarly, an [Outlier](https://www.analyticsvidhya.com/blog/2021/07/how-to-treat-outliers-in-a-data-set/) is an observation in a given dataset that lies far from the rest of the observations. That means an outlier treatment is vastly larger or smaller than the remaining values in the set.

## Why Do they Occur?

An outlier may occur due to the variability in the data, or due to experimental error/human error.

Sharpen your knowledge of Detecting and Treating Outliers!   
They may indicate an experimental error or heavy skewness in the data(heavy-tailed distribution).

## What Do They Affect?

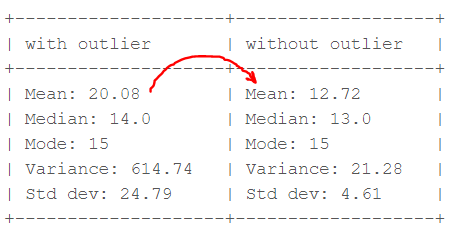
In statistics, we have three measures of central tendency namely Mean, Median, and Mode. They help us describe the data.

* Mean is the accurate measure to describe the data when we do not have any outliers present.
* Median is used if there is an outlier in the dataset.
* Mode is used if there is an outlier AND about ½ or more of the data is the same.

Mean’ is the only measure of central tendency that is affected by the outlier treatment which in turn impacts Standard deviation.

### Example

Consider a small dataset, sample= [15, 101, 18, 7, 13, 16, 11, 21, 5, 15, 10, 9]. By looking at it, one can quickly say ‘101’ is an outlier that is much larger than the other values.

Computation with and without outlier (Image by author)

From the above calculations, we can clearly say the Mean is more affected than the Median.

## Detecting Outliers

If our dataset is small, we can detect the [outlier](https://www.analyticsvidhya.com/blog/2022/09/dealing-with-outliers-using-the-iqr-method/)by just looking at the dataset. But what if we have a huge dataset, how do we identify the outliers then? We need to use visualization and mathematical techniques.

Below are some of the techniques of detecting outliers

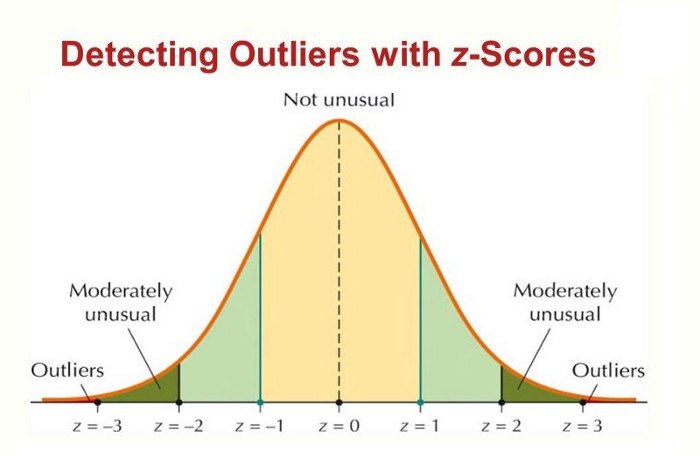
* Boxplots
* Z-score
* Inter Quantile Range(IQR)

### Detecting Outliers Using Boxplot

Python code for boxplot is:

### ****Detecting Outliers using the Z-scores****

**Criteria:**any data point whose Z-score falls out of 3rd standard deviation is an outlier treatment.

Detecting Outliers with Z-scores

### ****Steps****

* loop through all the data points and compute the Z-score using the formula (Xi-mean)/std.
* define a threshold value of 3 and mark the datapoints whose absolute value of Z-score is greater than the threshold as outliers.

**import** numpy **as** np

outliers = []

**def** **detect\_outliers\_zscore**(data):

thres = 3

mean = np.mean(data)

std = np.std(data)

# print(mean, std)

**for** i **in** data:

z\_score = (i-mean)/std

**if** (np.abs(z\_score) > thres):

outliers.append(i)

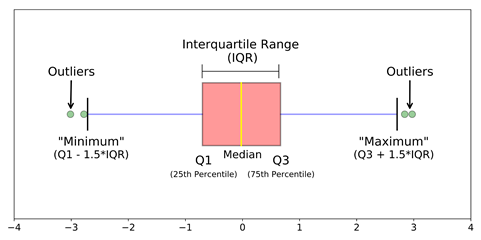
**return** outliers# Driver code

sample\_outliers = detect\_outliers\_zscore(sample)

print("Outliers from Z-scores method: ", sample\_outliers)

The above code outputs: **Outliers from Z-scores method: [101]**

### ****Detecting Outliers using the Inter Quantile Range(IQR)****

IQR to detect Outliners

**Criteria:** data points that lie 1.5 times of IQR above Q3 and below Q1 are outliers. This shows in detail about outlier treatment in Python.

#### ****Steps****

* Sort the dataset in ascending order
* calculate the 1st and 3rd quartiles(Q1, Q3)
* compute IQR=Q3-Q1
* compute lower bound = (Q1–1.5\*IQR), upper bound = (Q3+1.5\*IQR)
* loop through the values of the dataset and check for those who fall below the lower bound and above the upper bound and mark them as outlier treatment in python

#### Python Code

outliers = []

def detect\_outliers\_iqr(data):

data = sorted(data)

q1 = np.percentile(data, 25)

q3 = np.percentile(data, 75)

# print(q1, q3)

IQR = q3-q1

lwr\_bound = q1-(1.5\*IQR)

upr\_bound = q3+(1.5\*IQR)

# print(lwr\_bound, upr\_bound)

**for** i in data:

**if** (i<lwr\_bound **or** i>upr\_bound):

outliers.append(i)

**return** outliers# Driver code

sample\_outliers = detect\_outliers\_iqr(sample)

**print**("Outliers from IQR method: ", sample\_outliers)

The above code outputs: **Outliers from IQR method: [101]**

## How to Handle Outliers?

Till now we learned about detecting the outliers handling. The main question is how to deal with outliers?  
  
Below are some of the methods of treating the outliers:

#### Step 1: Trimming/Remove the outliers

In this technique, we remove the outliers from the dataset. Although it is not a good practice to follow.  
  
Python code to delete the outlier treatment and copy the rest of the elements to another array.

# Trimming for **i** in sample\_outliers: a = np.delete(sample, np.where(sample==i)) print(a) # print(len(sample), len(a))

The outlier ‘101’ is deleted and the rest of the data points are copied to another array ‘a’.

#### Step 2: Quantile Based Flooring and Capping

In this technique, the outlier is capped at a certain value above the 90th percentile value or floored at a factor below the 10th percentile value. Python code to delete the outlier and copy the rest of the elements to another array.

# Computing 10th, 90th percentiles and replacing the outlier treatment in python

tenth\_percentile = np.percentile(sample, 10)

ninetieth\_percentile = np.percentile(sample, 90)

# print(tenth\_percentile, ninetieth\_percentile)b =

np.where(sample<tenth\_percentile, tenth\_percentile, sample)

b = np.where(b>ninetieth\_percentile, ninetieth\_percentile, b)

# print("Sample:", sample)

print("New array:",b)

The above code outputs: **New array:**[15, 20.7, 18, 7.2, 13, 16, 11, 20.7, 7.2, 15, 10, 9]

The data points that are lesser than the 10th percentile are replaced with the 10th percentile value and the data points that are greater than the 90th percentile are replaced with 90th percentile value.

#### Step 3: Mean/Median Imputation

As the mean value is highly influenced by the outlier treatment, it is advised to replace the outliers with the median value.

**Python Code:**

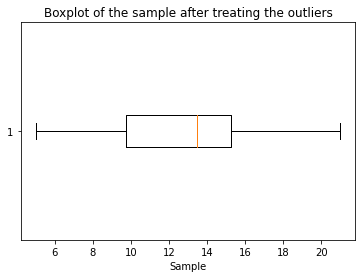
median = np.median(sample)# Replace with median **for** i in sample\_outliers: c = np.where(sample==i, 14, sample) print("Sample: ", sample) print("New array: ",c) # print(x.dtype)

#### Step 5: Visualizing the Data after Treating the Outlier

plt.boxplot(c, vert=False)

plt.title("Boxplot of the sample after treating the outliers")

plt.xlabel("Sample")

Data after treating Outliner

## Conclusion

In conclusion, identifying and addressing [outliers handling](https://www.analyticsvidhya.com/blog/2021/05/feature-engineering-how-to-detect-and-remove-outliers-with-python-code/) is paramount in data analysis. These data anomalies can skew results, leading to inaccurate insights and decisions. By employing robust detection techniques and thoughtful treatment strategies, we can enhance the integrity of our analyses and unlock hidden patterns within our data. How to Handle Outlier treatment in python, in this article once understood and managed, become valuable sources of information, ultimately contributing to more informed and reliable decision-making processes.

#### Key Takeaways:

* Learning techniques to detect outliers: boxplots, Z-score method, interquartile range (IQR) method
* Strategies to handle outliers: trimming/removing, quantile-based flooring and capping, mean/median imputation
* Visualizing and evaluating the data after treating outliers for improved analysis and decision-making

8. What are the various central inclination measures? Why does mean vary too much from median in certain data sets?

# Central Tendency | Understanding the Mean, Median & Mode

**Published on July 30, 2020 by [Pritha Bhandari](https://www.scribbr.com/author/pritha/" \o "All articles by Pritha Bhandari). Revised on June 21, 2023.**

**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**](https://www.scribbr.com/statistics/mode/)**:** the most frequent value.
* [**Median**](https://www.scribbr.com/statistics/median/)**:** the middle number in an ordered dataset.
* [**Mean**](https://www.scribbr.com/statistics/mean/)**:** the sum of all values divided by the total number of values.

In addition to central tendency, the variability and distribution of your dataset is important to understand when performing [descriptive statistics](https://www.scribbr.com/statistics/descriptive-statistics/).

## Distributions and central tendency

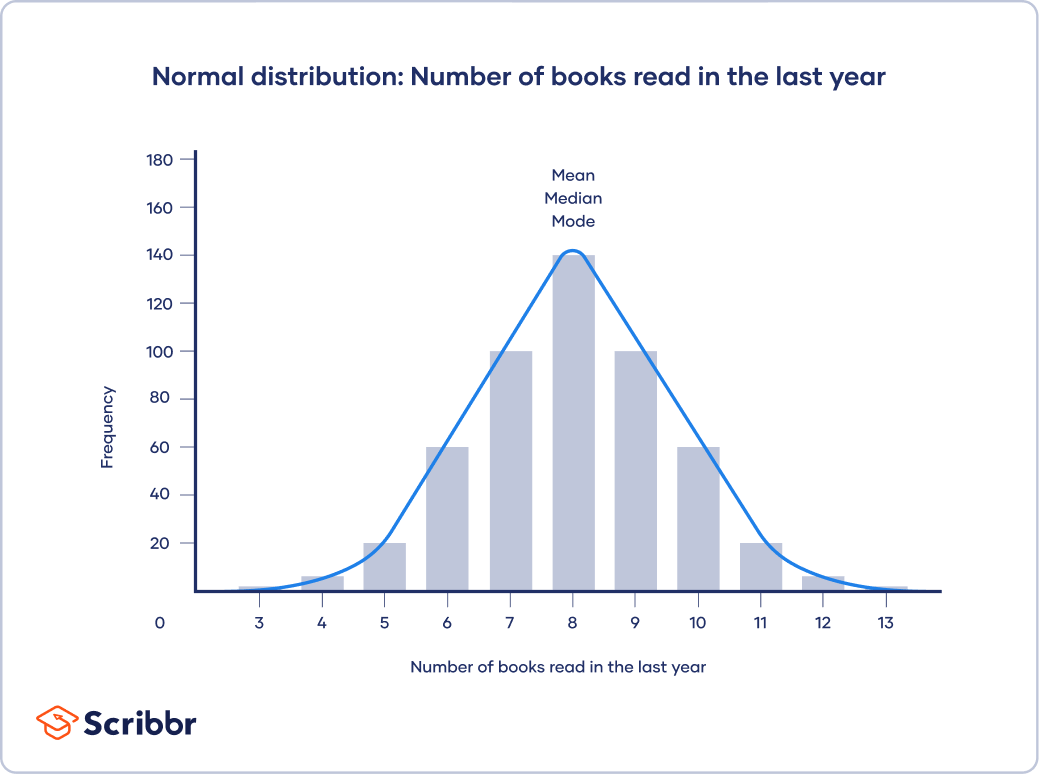
A dataset is a **distribution** of n number of scores or values.

### Normal distribution

In a [normal distribution](https://www.scribbr.com/statistics/normal-distribution/), data is symmetrically distributed with no [skew](https://www.scribbr.com/statistics/skewness/). Most values cluster around a central region, with values tapering off as they go further away from the center. The mean, mode and median are exactly the same in a normal distribution.

Example: Normal distributionYou [survey](https://www.scribbr.com/methodology/survey-research/) a [sample](https://www.scribbr.com/methodology/sampling-methods/) in your local community on the number of books they read in the last year.

A histogram of your data shows the frequency of responses for each possible number of books. From looking at the chart, you see that there is a normal distribution.



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

### Skewed distributions

In [skewed](https://www.scribbr.com/statistics/skewness/) distributions, more values fall on one side of the center than the other, and the mean, median and mode all differ from each other. One side has a more spread out and longer tail with fewer scores at one end than the other. The direction of this tail tells you the side of the skew

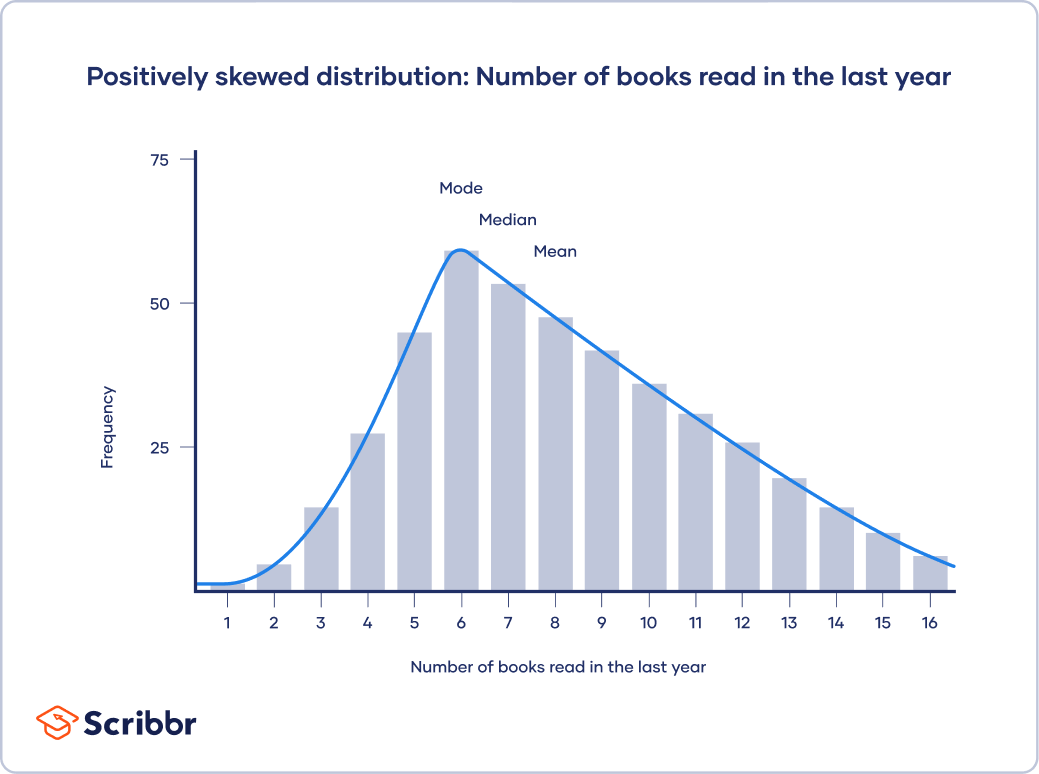
In a positively skewed distribution, there’s a cluster of lower scores and a spread out tail on the right. In a negatively skewed distribution, there’s a cluster of higher scores and a spread out tail on the left.

* [**Positively skewed distribution**](https://www.scribbr.com/statistics/central-tendency/)

* [**Negatively skewed distribution**](https://www.scribbr.com/statistics/central-tendency/)

In this histogram, your distribution is skewed to the right, and the central tendency of your dataset is on the lower end of possible scores.

In a positively skewed distribution, mode < median < mean.



## Here's why students love Scribbr's proofreading services

[**Discover proofreading & editing**](https://www.scribbr.com/proofreading-editing/paper-editing-service/?scr_source=knowledgebase&scr_medium=in-text&scr_campaign=proofreading&src_content=trustpilot)

## Mode

The [**mode**](https://www.scribbr.com/statistics/mode/) is the most frequently occurring value in the dataset. It’s possible to have no mode, one mode, or more than one mode.

To find the mode, sort your dataset numerically or [categorically](https://www.scribbr.com/methodology/types-of-variables/#quantitative-vs-categorical) and select the response that occurs most frequently.

Example: Finding the modeIn a survey, you ask 9 participants whether they identify as conservative, moderate, or liberal.

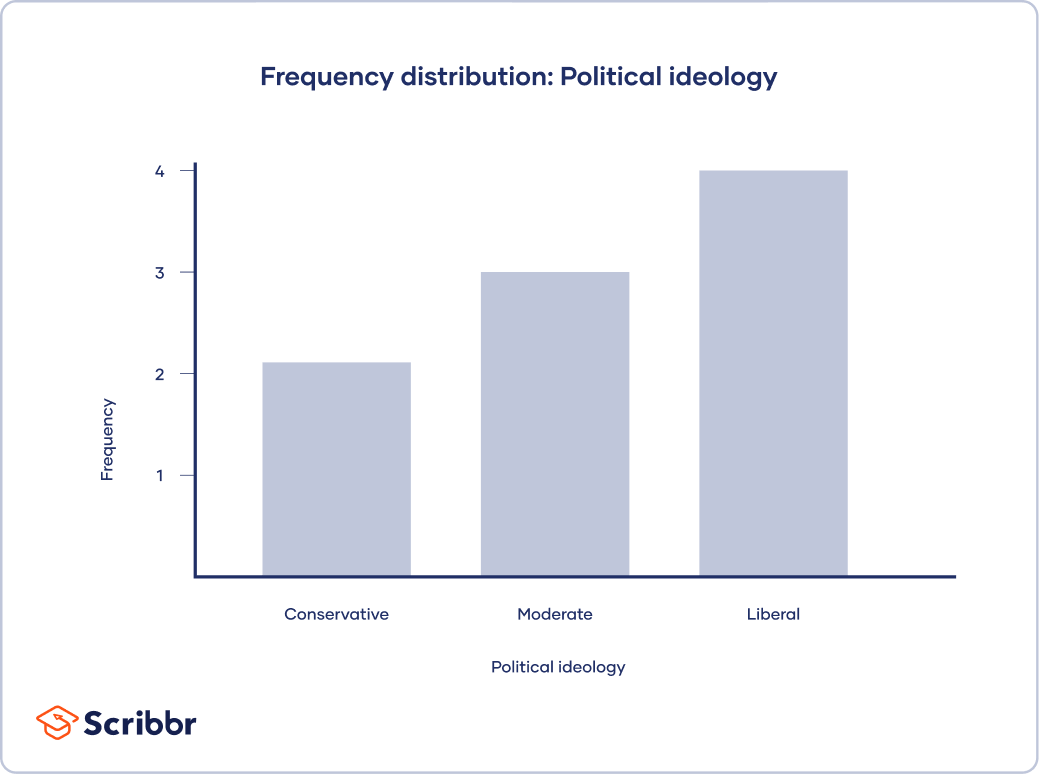
To find the mode, sort your data by category and find which response was chosen most frequently.

To make it easier, you can create a [frequency table](https://www.scribbr.com/statistics/frequency-distributions/#how-to) to count up the values for each category.

| **Political ideology** | **Frequency** |
| --- | --- |
| **Conservative** | 2 |
| **Moderate** | 3 |
| **Liberal** | 4 |

**Mode: Liberal**

The mode is easily seen in a bar graph because it is the value with the highest bar.



### When to use the mode

The mode is most applicable to data from a nominal level of measurement. [Nominal data](https://www.scribbr.com/statistics/nominal-data/) is classified into mutually exclusive categories, so the mode tells you the most popular category.

For continuous variables or [ratio levels](https://www.scribbr.com/statistics/ratio-data/) of measurement, the mode may not be a helpful measure of central tendency. That’s because there are many more possible values than there are in a nominal or [ordinal](https://www.scribbr.com/statistics/ordinal-data/) level of measurement. It’s unlikely for a value to repeat in a ratio [level of measurement](https://www.scribbr.com/statistics/levels-of-measurement/).

Example: Ratio data with no modeYou [collect data](https://www.scribbr.com/methodology/data-collection/) on reaction times in a computer task, and your dataset contains values that are all different from each other.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Participant** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| **Reaction time (milliseconds)** | 267 | 345 | 421 | 324 | 401 | 312 | 382 | 298 | 303 |

In this dataset, there is no mode, because each value occurs only once.

## Median

The[median](https://www.scribbr.com/statistics/median/) of a dataset is the value that’s exactly in the middle when it is ordered from low to high.

Example: Finding the medianYou measure the reaction times of 7 participants on a computer task and categorize them into 3 groups: slow, medium or fast.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Participant** | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| **Speed** | Medium | Slow | Fast | Fast | Medium | Fast | Slow |

To find the median, you first order all values from low to high. Then, you find the value in the middle of the ordered dataset—in this case, the value in the 4th position.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Ordered dataset** | Slow | Slow | Medium | Medium | Fast | Fast | Fast |

**Median: Medium**

In larger datasets, it’s easier to use simple formulas to figure out the position of the middle value in the distribution. You use different methods to find the median of a dataset depending on whether the total number of values is even or odd.

### Median of an odd-numbered dataset

For an odd-numbered dataset, find the value that lies at the \dfrac{(n+1)}{2} position, where n is the number of values in the dataset.

ExampleYou measure the reaction times in milliseconds of 5 participants and order the dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Reaction time (milliseconds)** | 287 | 298 | 345 | 365 | 380 |

The middle position is calculated using \dfrac{(n+1)}{2}, where n = 5.

\begin{equation*}\dfrac{(5+1)}{2}=3$\end{equation*}

That means the median is the 3rd value in your ordered dataset.

**Median: 345 milliseconds**

### Median of an even-numbered dataset

For an even-numbered dataset, find the two values in the middle of the dataset: the values at the \dfrac{n}{2} and (\dfrac{n}{2})+1 positions. Then, find their mean.

ExampleYou measure the reaction times of 6 participants and order the dataset.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Reaction time (milliseconds)** | 287 | 298 | 345 | 357 | 365 | 380 |

The middle positions are calculated using \dfrac{n}{2} and (\dfrac{n}{2})+1, where n = 6.

\begin{equation*}\dfrac{6}{2}=3\end{equation*}

\begin{equation*}(\dfrac{6}{2})+1=4\end{equation*}

That means the middle values are the 3rd value, which is **345**, and the 4th value, which is **357**.

To get the median, take the mean of the 2 middle values by adding them together and dividing by 2.

\begin{equation*}\dfrac{(345+357)}{2}=351\end{equation*}

**Median: 351 milliseconds**

## Mean

The [arithmetic mean](https://www.scribbr.com/statistics/mean/) of a dataset (which is different from the [geometric mean](https://www.scribbr.com/statistics/geometric-mean/)) is the sum of all values divided by the total number of values. It’s the most commonly used measure of central tendency because all values are used in the calculation.

Example: Finding the mean

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Participant** | 1 | 2 | 3 | 4 | 5 |
| **Reaction time (milliseconds)** | 287 | 345 | 365 | 298 | 380 |

First you add up the sum of all values:

\begin{equation*}\sum{x}=287+345+365+298+380=1\,675\end{equation*}

[Then](https://www.scribbr.com/commonly-confused-words/then-vs-than/) you calculate the mean using the formula

\begin{equation*} \frac {\sum{x}}{n} \end{equation*}

There are 5 values in the dataset, so n = 5.

\begin{equation*}\bar{x}=\dfrac{1\,675}{5}=335\end{equation*}

**Mean (x̄): 335 milliseconds**

### Outlier effect on the mean

[Outliers](https://www.scribbr.com/statistics/outliers/) can significantly increase or decrease the mean when they are included in the calculation. Since all values are used to calculate the mean, it can be affected by extreme outliers. An outlier is a value that differs significantly from the others in a dataset.

Example: Mean with an outlierIn this dataset, we swap out one value with an extreme outlier.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Participant** | 1 | 2 | 3 | 4 | 5 |
| **Reaction time (milliseconds)** | 832 | 345 | 365 | 298 | 380 |

\begin{equation*}\sum{x}=832+345+365+298+380=2\,220\end{equation*}

\begin{equation*}\bar{x}=\dfrac{\sum{x}}{n}=\dfrac{2\,220}{5}=444\end{equation*}

Due to the outlier, the mean (\bar{x}) becomes much higher, even though all the other numbers in the dataset stay the same.

**Mean: 444 milliseconds**

### Population versus sample mean

A dataset contains values from a [sample or a population](https://www.scribbr.com/methodology/population-vs-sample/). A population is the entire group that you are interested in researching, while a sample is only a subset of that population.

While data from a sample can help you make estimates about a population, only full population data can give you the complete picture.

In statistics, the notation of a sample mean and a population mean and their formulas are different. But the procedures for calculating the population and sample means are the same.

Sample mean formulaThe sample mean is written as M or x̄ (pronounced x-bar). For calculating the mean of a sample, use this formula:

\begin{equation*}\bar{x}=\dfrac{\sum{x}}{n}\end{equation*}

* x̄:  sample mean
* \sum{x}: sum of all values in the sample dataset
* n: number of values in the sample dataset

Population mean formulaThe population mean is written as μ (Greek term mu). For calculating the mean of a population, use this formula:

\begin{equation*}\mu=\dfrac{\sum{X}}{N}\end{equation*}

* μ: population mean
* \sum{X}: sum of all values in the population dataset
* N: number of values in the population dataset

### Receive feedback on language, structure, and formatting

Professional editors proofread and edit your paper by focusing on:

* Academic style
* Vague sentences
* Grammar
* Style consistency

[**See an example**](https://www.scribbr.com/proofreading-editing/paper-editing-service/?scr_source=knowledgebase&scr_medium=in-text&scr_campaign=proofreading&src_content=feedback+language+structure+layout)

## When should you use the mean, median or mode?

The 3 main measures of central tendency are best used in combination with each other because they have complementary strengths and limitations. But sometimes only 1 or 2 of them are applicable to your dataset, depending on the level of measurement of the variable.

* The mode can be used for any level of measurement, but it’s most meaningful for nominal and ordinal levels.
* The median can only be used on data that can be ordered – that is, from ordinal, interval and ratio levels of measurement.
* The mean can only be used on interval and ratio levels of measurement because it requires equal spacing between adjacent values or scores in the scale.

| **Levels of measurement** | **Examples** | **Measure of central tendency** |
| --- | --- | --- |
| [**Nominal**](https://www.scribbr.com/statistics/nominal-data/) | * Ethnicity * Political ideology | * Mode |
| [**Ordinal**](https://www.scribbr.com/statistics/ordinal-data/) | * Level of anxiety * Income bracket | * Mode * Median |
| [**Interval**](https://www.scribbr.com/statistics/interval-data/)**and**[**ratio**](https://www.scribbr.com/statistics/ratio-data/) | * Reaction time * Test score * Temperature | * Mode * Median * Mean |

To decide which measures of central tendency to use, you should also consider the distribution of your dataset.

For [normally distributed](https://www.scribbr.com/statistics/normal-distribution/) data, all three measures of central tendency will give you the same answer so they can all be used.

In [skewed](https://www.scribbr.com/statistics/skewness/) distributions, the median is the best measure because it is unaffected by extreme [outliers](https://www.scribbr.com/statistics/outliers/) or non-symmetric distributions of scores. The mean and mode can vary in skewed distributions.

.

9. Describe how a scatter plot can be used to investigate bivariate relationships. Is it possible to find outliers using a scatter plot?

## What are outliers in scatter plots?

Scatter plots often have a pattern. We call a data point an **outlier** if it doesn't fit the pattern.

consider the scatter plot above, which shows data for students on a backpacking trip. (Each point represents a student.)

Notice how two of the points don't fit the pattern very well. These points have been labeled Brad and Sharon, which are the names of the students they represent.

Sharon could be considered an outlier because she is carrying a much heavier backpack than the pattern predicts.

Brad could be considered an outlier because he is carrying a much lighter backpack than the pattern predicts.

**Key idea:** There is no special rule that tells us whether or not a point is an outlier in a scatter plot. When doing more advanced statistics, it may become helpful to invent a precise definition of "outlier", but we don't need that yet.

## Practice problems

To fully wrap our minds around why certain data points might be considered outliers, let's try a couple of practice problems.

### Problem 1: Computer shopping

**Quality ratingPrice (dollars)DBAC**

Michelle was researching different computers to buy for college. She looked up the prices and quality ratings for a sample of computers. Her data is shown in the scatter plot to the right, where each point is a computer.

Michele wants to buy a computer whose quality rating is far higher than the pattern would predict based on its price.

**Which of the labeled points represents a computer that Michele wants to buy?**

Choose 1 answer:

### Problem 2: Test scores

**Average math scoreParticipation (% taking SAT)ABC**

Some high school students in the U.S. take a test called the SAT before applying to colleges. The scatter plot to the right shows what percent of each state's college-bound graduates took the SAT in  , along with that state's average score on the math section.

10. Describe how cross-tabs can be used to figure out how two variables are related.

## Crosstabs

To describe a single categorical variable, we use [**frequency tables**](https://libguides.library.kent.edu/SPSS/FrequenciesCategorical). To describe the relationship between two categorical variables, we use a special type of table called a cross-tabulation (or "crosstab" for short). In a cross-tabulation, the categories of one variable determine the rows of the table, and the categories of the other variable determine the columns. The cells of the table contain the number of times that a particular combination of categories occurred. The "edges" (or "margins") of the table typically contain the total number of observations for that category.

This type of table is also known as a:

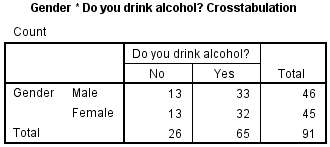
* Crosstab.
* Two-way table.
* Contingency table.

## Describing a Crosstab

The dimensions of the crosstab refer to the number of rows and columns in the table. (The "total" row/column are not included.) The table dimensions are reported as as RxC, where R is the number of categories for the row variable, and C is the number of categories for the column variable.

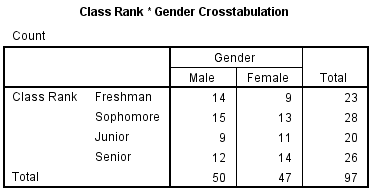
Additionally, a "square" crosstab is one in which the row and column variables have the same number of categories. Tables of dimensions 2x2, 3x3, 4x4, etc. are all square crosstabs.

#### Example 1: A "square" table (2x2)



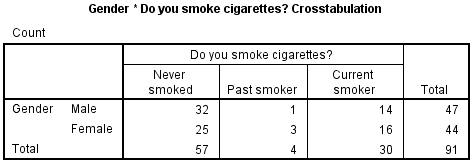
* **Row variable**: Gender (2 categories: male, female)
* **Column variable**: Alcohol (2 categories: no, yes)
* **Table dimension**: 2x2 (square)

#### Example 2: A "long" table (4x2)



* **Row variable**: Class Rank (4 categories: freshman, sophomore, junior, senior)
* **Column variable**: Gender (2 categories: male, female)
* **Table dimension**: 4x2

#### Example 3: A "wide" table (2x3)



* **Row variable**: Gender (2 categories: male, female)
* **Column variable**: Smoking (3 categories: never smoked, past smoker, current smoker)
* **Table dimension**: 2x3

## Understanding Row, Column, and Total Percents

A typical 2x2 crosstab has the following construction:

|  | **Column 1** | **Column 2** | **Row totals** |
| --- | --- | --- | --- |
| **Row 1** | a | b | a + b |
| **Row 2** | c | d | c + d |
| **Column totals** | a + c | b + d | a + b + c + d |

The letters a, b, c, and d represent what are called cell counts.

* a is the number of observations corresponding to Row 1 AND Column 1.
* b is the number of observations corresponding to Row 1 AND Column 2.
* c is the number of observations corresponding to Row 2 AND Column 1.
* d is the number of observations corresponding to Row 2 AND Column 2.

By adding a, b, c, and d, we can determine the total number of observations in each category, and in the table overall.

* Row sum of row 1 (i.e., total number of observations in Row 1): a + b
* Row sum of row 2 (i.e., total number of observations in Row 2): c + d
* Column sum of column 1 (i.e., total number of observations in Column 1): a + c
* Column sum of column 2 (i.e., total number of observations in Column 2): b + d
* Total sum (i.e., total number of observations in the table): n = a + b + c + d

The row sums and column sums are sometimes referred to as marginal frequencies. Note that if you were to make frequency tables for your row variable and your column variable, the frequency table should match the values for the row totals and column totals, respectively.

When you are describing the composition of your sample, it is often useful to refer to the proportion of the row or column that fell within a particular category. This can be achieved by computing the row percentages or column percentages.

| Formulas for computing row percentages | | | |
| --- | --- | --- | --- |
|  | **Column 1** | **Column 2** | **Row totals** |
| **Row 1**  **Row 1 %** | a  a / (a + b) | b  b / (a + b) | a + b  (a + b) / (a+b) = 100% |
| **Row 2**  **Row 2 %** | c  c / (c + d) | d  d / (c + d) | c + d  (c + d) / (c + d) = 100% |
| **Column totals**  **% of total** | a + c  (a + c) / (a + b + c + d) | b + d  (b + d) / (a + b + c + d) | a + b + c + d  (a + b + c + d) / (a + b + c + d) = 100% |

Notice that when computing row percentages, the denominators for cells a, b, c, d are determined by the row sums (here, a + b and c + d). This implies that the percentages in the "row totals" column must equal 100%.

| Formulas for computing column percentages | | | |
| --- | --- | --- | --- |
|  | **Column 1** | **Column 2** | **Row totals** |
| **Row 1**  **Column 1 %** | a  a / (a + c) | b  b / (b + d) | a + b  (a + b) / (a + b + c + d) |
| **Row 2**  **Column 2 %** | c  c / (a + c) | d  d / (b + d) | c + d  (c + d) / (a + b + c + d) |
| **Column totals**  **Percentage %** | a + c  (a + c) / (a + c) = 100% | b + d  (b + d) / (b + d) = 100% | a + b + c + d  (a + b + c + d) / (a + b + c + d) = 100% |

Notice that when computing column percentages, the denominators for cells a, b, c, d are determined by the column sums (here, a + c and b + d). This implies that the percentages in the "column totals" row must equal 100%.

| Formulas for computing total percentages | | | |
| --- | --- | --- | --- |
|  | **Column 1** | **Column 2** | **Row totals** |
| **Row 1**  **% of total** | a  a / (a + b + c + d) | b  b / (a + b + c + d) | a + b  (a + b) / (a + b + c + d) |
| **Row 2**  **% of total** | c  c / (a + b + c + d) | d  d / (a + b + c + d) | c + d  (c + d) / (a + b + c + d) |
| **Column totals**  **% of total** | a + c  (a + c) / (a + b + c + d) | b + d  (b + d) / (a + b + c + d) | a + b + c + d  (a + b + c + d) / (a + b + c + d) = 100% |

Notice that when total percentages are computed, the denominators for all of the computations are equal to the total number of observations in the table, i.e. a + b + c + d.

## Data Set-Up and Requirements

### DATA REQUIREMENTS

Your data must meet the following requirements:

1. Two categorical variables.
2. Two or more categories (groups) for each variable.

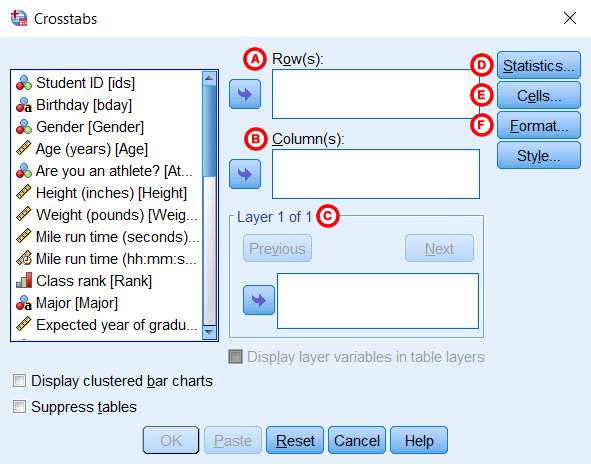
### DATA SET-UP

The categorical variables in your SPSS dataset can be numeric or string, and their measurement level can be defined as nominal, ordinal, or scale. However, crosstabs should only be used when there are a limited number of categories.

Note that in most cases, the row and column variables in a crosstab can be used interchangeably. The choice of row/column variable is usually dictated by space requirements or interpretation of the results.

## Create a Crosstab in SPSS

To create a crosstab, click **Analyze**> **Descriptive Statistics** > **Crosstabs**.



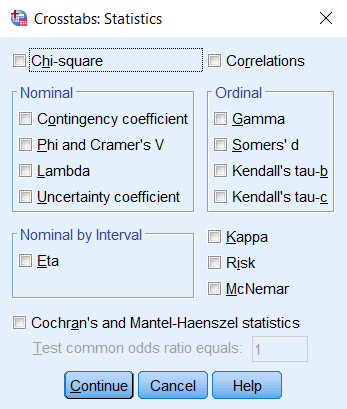
**A** **Row(s):** One or more variables to use in the rows of the crosstab(s). You must enter at least one Row variable.

**B** **Column(s):** One or more variables to use in the columns of the crosstab(s). You must enter at least one Column variable.

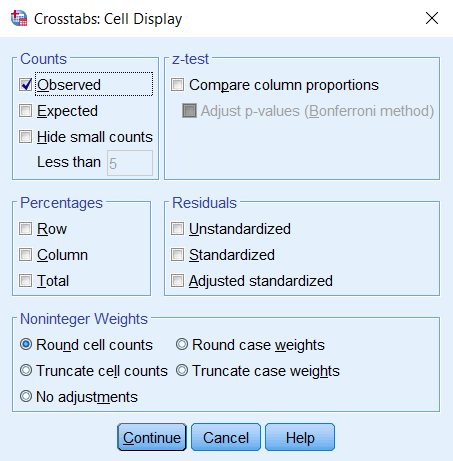
Also note that if you specify one row variable and two or more column variables, SPSS will print crosstabs for each pairing of the row variable with the column variables. The same is true if you have one column variable and two or more row variables, or if you have multiple row and column variables.

**C** **Layer:** An optional "stratification" variable. When a layer variable is specified, the crosstab between the Row and Column variable(s) will be created at each level of the layer variable. You can have multiple layers of variables by specifying the first layer variable and then clicking **Next**to specify the second layer variable. Alternatively, you can try out multiple variables as single layers at a time by putting them all in the Layer 1 of 1 box.

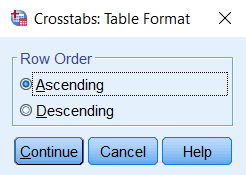
**D** **Statistics:**Opens the Crosstabs: Statistics window, which contains fifteen different inferential statistics for comparing categorical variables. (These statistics will be covered in detail in a later tutorial.)



**E** **Cells:** Opens the Crosstabs: Cell Display window, which controls which output is displayed in each cell of the crosstab.



**F** **Format:** Opens the Crosstabs: Table Format window, which specifies how the rows of the table are sorted.



## Example: Summarizing the Relationships of Three Categorical Variables

### PROBLEM STATEMENT

Some universities in the United States require that freshmen live in the on-campus dormitories during their first year, with exceptions for students whose families live within a certain radius of campus. That is, certain freshmen whose families live close enough to campus are permitted to live off-campus. After completing their first or second year of school, students living in the dorms may choose to move into an off-campus apartment. How prevalent is this pattern?

In the sample dataset, there are several variables relating to this question:

* Rank - Class rank (Freshmen, Sophomore, Junior, Senior)
  + RankUpperUnder - Class rank recoded into Underclassman/Upperclassman (see the [**Recode into Different Variables**](https://libguides.library.kent.edu/SPSS/RecodeVariables) tutorial)
* LiveOnCampus - Do you live on campus? (Yes/No)
* State - Are you an in-state or out-of-state student? (In State, Out of state)
  + State\_Residency - State residency, converted from string to numeric so that missing values are correctly identified (See the [**Automatic Recode**](https://libguides.library.kent.edu/SPSS/AutomaticRecode) tutorial)

Let's use different aspects of the Crosstabs procedure to investigate the relationship between class rank and living on campus.

### PART 1 - SIMPLE CROSSTABS

Using the sample data, let's make crosstab of the variables Rank and LiveOnCampus. Let the row variable be Rank, and the column variable be LiveOnCampus.

#### Running the Procedure

##### **Using the Crosstabs Dialog Window**

1. Open the Crosstabs window (**Analyze > Descriptive Statistics > Crosstabs**).
2. Select Rank as the row variable, and LiveOnCampus as the column variable.
3. Click **OK**.

##### **Using Syntax**

**CROSSTABS**

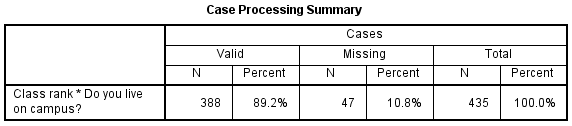
**/TABLES=Rank BY LiveOnCampus**

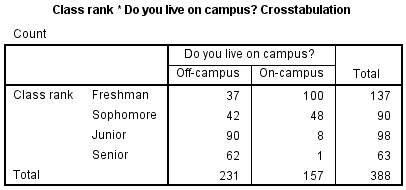
**/FORMAT=AVALUE TABLES**

**/CELLS=COUNT**

**/COUNT ROUND CELL.**

#### Output





The **Case Processing Summary**tells us what proportion of the observations had nonmissing values for both Rank and LiveOnCampus. In this sample, there were 47 cases that had a missing value for Rank, LiveOnCampus, or for both Rank and LiveOnCampus.

The second table (here, **Class Rank \* Do you live on campus? Crosstabulation**) contains the crosstab. We can quickly observe information about the interaction of these two variables:

* Many more freshmen lived on-campus (100) than off-campus (37)
* About an equal number of sophomores lived off-campus (42) versus on-campus (48)
* Far more juniors lived off-campus (90) than on-campus (8)
* Only one (1) senior lived on campus; the rest lived off-campus (62)

Note the margins of the crosstab (i.e., the "total" row and column) give us the same information that we would get from frequency tables of Rank and LiveOnCampus, respectively:

* The sample had 137 freshmen, 90 sophomores, 98 juniors, and 63 seniors
* There were 231 individuals who lived off-campus, and 157 individuals lived on-campus

### PART 2 - ROW, COLUMN, AND TOTAL PERCENTAGES

Let's build on the table shown in Example 1 by adding row, column, and total percentages. For simplicity's sake, let's switch out the variable Rank (which has four categories) with the variable RankUpperUnder (which has two categories).

#### Running the Procedure

##### **Using the Crosstabs Dialog Window**

1. Reopen the Crosstabs window (**Analyze > Descriptive Statistics > Crosstabs**).
2. In the Row box, replace variable Rank with RankUpperUnder.
3. Click **Cells**. In the Percentages area, check off Row, Column, and Total percentages. (In the following examples, we will be showing each of these one at a time for ease of reading.) Click **Continue**.
4. Click **OK**to run.

##### **Using Syntax**

**CROSSTABS**

**/TABLES=RankUpperUnder BY LiveOnCampus**

**/FORMAT=AVALUE TABLES**

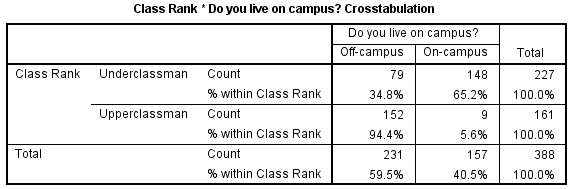
**/CELLS=COUNT ROW COLUMN TOTAL**

**/COUNT ROUND CELL.**

#### Output

##### **Row percents**

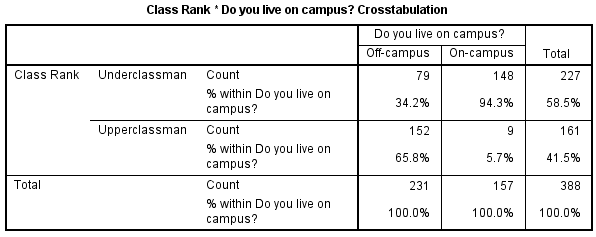
If the row variable is RankUpperUnder and the column variable is LiveOnCampus, then the row percentages will tell us what percentage of the upperclassmen or what percentage of the underclassmen live on campus. That is, variable RankUpperUnder will determine the denominator of the percentage computations.



* The proportion of underclassmen who live off campus is 34.8%, or 79/227.
* The proportion of underclassmen who live on campus is 65.2%, or 148/226.
* The proportion of upperclassmen who live off campus is 94.4%, or 152/161.
* The proportion of upperclassmen who live on campus is 5.6%, or 9/161.

##### **Column percents**

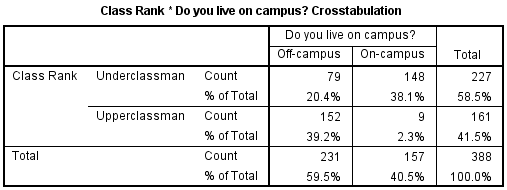
If the row variable is RankUpperUnder and the column variable is LiveOnCampus, then the column percentages will tell us what percentage of the individuals who live on campus are upper or underclassmen. That is, variable LiveOnCampus will determine the denominator of the percentage computations.



* The proportion of individuals living off campus who are underclassmen is 34.2%, or 79/231.
* The proportion of individuals living off campus who are upperclassmen is 65.8%, or 152/231.
* The proportion of individuals living on campus who are underclassmen is 94.3%, or 148/157.
* The proportion of individuals living on campus who are upperclassmen is 5.7%, or 9/157.

##### **Total percents**

If the row variable is RankUpperUnder and the column variable is LiveOnCampus, then the total percentage tells us what proportion of the total is within each combination of RankUpperUnder and LiveOnCampus. That is, the overall table size determines the denominator of the percentage computations.



* Underclassmen living off campus make up 20.4% of the sample (79/388).
* Underclassmen living on campus make up 38.1% of the sample (148/388).
* Upperclassmen living off campus make up 39.2% of the sample (152/388).
* Upperclassmen living on campus make up 2.3% of the sample (9/388).

### PART 3 - CROSSTABS WITH LAYER VARIABLE

Let's modify our analysis slightly by taking into account the students' state of residence (in-state or out-of-state). Here, we will be working with three categorical variables: RankUpperUnder, LiveOnCampus, and State\_Residency.

In this example, we want to create a crosstab of RankUpperUnder by LiveOnCampus, with variable State\_Residency acting as a strata, or layer variable.

#### Running the Procedure

##### **Using the Crosstabs Dialog Window**

1. Open the Crosstabs dialog (**Analyze > Descriptive Statistics > Crosstabs**).
2. Select RankUpperUnder as the row variable, and LiveOnCampus as the column variable.
3. Select State\_Residency as the layer variable.
4. You may want to go back to the **Cells** options and turn off the row, column, and total percentages if you have just run the previous example.
5. Click **OK**.

##### **Syntax**

**CROSSTABS**

**/TABLES=RankUpperUnder BY LiveOnCampus BY State\_Residency**

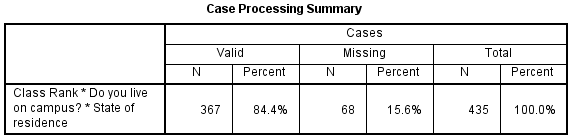
**/FORMAT=AVALUE TABLES**

**/CELLS=COUNT**

**/COUNT ROUND CELL.**

#### Output

Again, the Crosstabs output includes the boxes **Case Processing Summary** and the crosstabulation itself.



Notice that after including the layer variable State Residency, the number of valid cases we have to work with has dropped from 388 to 367. This is because the cr