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

**Step 1: Collect Data**

Given the problem you want to solve, you will have to investigate and obtain data that you will use to feed your machine. The quality and quantity of information you get are very important since it will directly impact how well or badly your model will work. You may have the information in an existing database or you must create it from scratch. If it is a small project you can create a spreadsheet that will later be easily exported as a CSV file. It is also common to use the web scraping technique to automatically collect information from various sources such as APIs.

**Step 2: Prepare the data**

This is a good time to [visualize your data](https://analyticsindiamag.com/how-to-get-started-with-visual-ai-the-new-automl-solution-by-datarobot/) and check if there are correlations between the different characteristics that we obtained. It will be necessary to make a selection of characteristics since the ones you choose will directly impact the execution times and the results. You can also reduce dimensions by applying PCA if necessary.

Additionally, you must balance the amount of data we have for each result -class- so that it is significant as the learning may be biased towards a type of response and when your model tries to generalize knowledge it will fail.

You must also separate the data into two groups: one for training and the other for model evaluation which can be divided approximately in a ratio of 80/20 but it can vary depending on the case and the volume of data we have.

At this stage, you can also pre-process your data by normalizing, eliminating duplicates, and making error corrections.

**Step 3: Choose the model**

There are several models that you can choose according to the objective that you might have: you will use algorithms of [classification](https://analyticsindiamag.com/transfer-learning-for-multi-class-image-classification-using-deep-convolutional-neural-network/), prediction, [linear regression](https://analyticsindiamag.com/ann-with-linear-regression/), [clustering](https://analyticsindiamag.com/comparison-of-k-means-hierarchical-clustering-in-customer-segmentation/), i.e. [k-means](https://analyticsindiamag.com/comparison-of-k-means-hierarchical-clustering-in-customer-segmentation/)or K-Nearest Neighbor, Deep Learning, i.e Neural Networks, [Bayesian](https://analyticsindiamag.com/deepmind-researchers-develop-tools-to-visualise-unfairness-using-causal-bayesian-networks/), etc.

1. There are various models to be used depending on the data you are going to process such as images, sound, text, and numerical values. In the following table, we will see some models and their applications that you can apply in your projects:

**Step 4 Train your machine model**

You will need to train the datasets to run smoothly and see an incremental improvement in the prediction rate. Remember to initialize the weights of your model randomly -the weights are the values that multiply or affect the relationships between the inputs and outputs- which will be automatically adjusted by the selected algorithm the more you train them.

**Step 5: Evaluation**

You will have to check the machine created against your evaluation data set that contains inputs that the model does not know and verify the precision of your already trained model. If the accuracy is less than or equal to 50%, that model will not be useful since it would be like tossing a coin to make decisions. If you reach 90% or more, you can have good confidence in the results that the model gives you.

**Step 6: Parameter Tuning**

If during the evaluation you did not obtain good predictions and your precision is not the minimum desired, it is possible that you have overfitting -or underfitting problems and you must return to the training step before making a new configuration of parameters in your model. You can increase the number of times you iterate your training data- termed epochs. Another important parameter is the one known as the “learning rate”, which is usually a value that multiplies the gradient to gradually bring it closer to the global -or local- minimum to minimize the cost of the function.

Increasing your values by 0.1 units from 0.001 is not the same as this can significantly affect the model execution time. You can also indicate the maximum error allowed for your model. You can go from taking a few minutes to hours, and even days, to train your machine. These parameters are often called Hyperparameters. This “tuning” is still more of an art than a science and will improve as you experiment. There are usually many parameters to adjust and when combined they can trigger all your options. Each algorithm has its own parameters to adjust. To name a few more, in Artificial Neural Networks (ANNs) you must define in its architecture the number of hidden layers it will have and gradually test with more or less and with how many neurons each layer. This will be a work of great effort and patience to give good results.

**Step 7: Prediction or Inference**

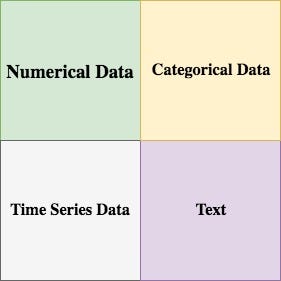
You are now ready to use your Machine Learning model inferring results in real-life scenarios.



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

Almost anything can be turned into DATA. Building a deep understanding of the different data types is a crucial prerequisite for doing Exploratory Data Analysis (EDA) and Feature Engineering for Machine Learning models. You also need to convert data types of some variables in order to make appropriate choices for visual encodings in data visualization and storytelling.

Most data can be categorized into 4 basic types from a Machine Learning perspective: numerical data, categorical data, time-series data, and text.

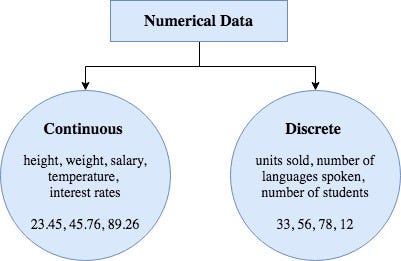


Data Types From A Machine Learning Perspective

**Numerical Data**

Numerical data is any data where data points are exact numbers. Statisticians also might call numerical data, quantitative data. This data has meaning as a **measurement** such as house prices or as a count, such as a number of residential properties in Los Angeles or how many houses sold in the past year.

Numerical data can be characterized by continuous or discrete data. Continuous data can assume any value within a range whereas discrete data has distinct values.



Numerical Data

For example, the number of students taking Python class would be a discrete data set. You can only have discrete whole number values like 10, 25, or 33. A class cannot have 12.75 students enrolled. A student either join a class or he doesn’t. On the other hand, continuous data are numbers that can fall anywhere within a range. Like a student could have an average score of 88.25 which falls between 0 and 100.

The takeaway here is that numerical data is not ordered in time. They are just numbers that we have collected.

**Categorical Data**

Categorical data represents characteristics, such as a hockey player’s position, team, hometown. Categorical data can take numerical values. For example, maybe we would use 1 for the colour red and 2 for blue. But these numbers don’t have a mathematical meaning. That is, we can’t add them together or take the average.

In the context of super classification, categorical data would be the class label. This would also be something like if a person is a man or woman, or property is residential or commercial.

There is also something called ordinal data, which in some sense is a mix of numerical and categorical data. In ordinal data, the data still falls into categories, but those categories are ordered or ranked in some particular way. An example would be class difficulty, such as beginner, intermediate, and advanced. Those three types of classes would be a way that we could label the classes, and they have a natural order in increasing difficulty.

Another example is that we just take quantitative data, and splitting it into groups, so we have bins or categories of other types of data.



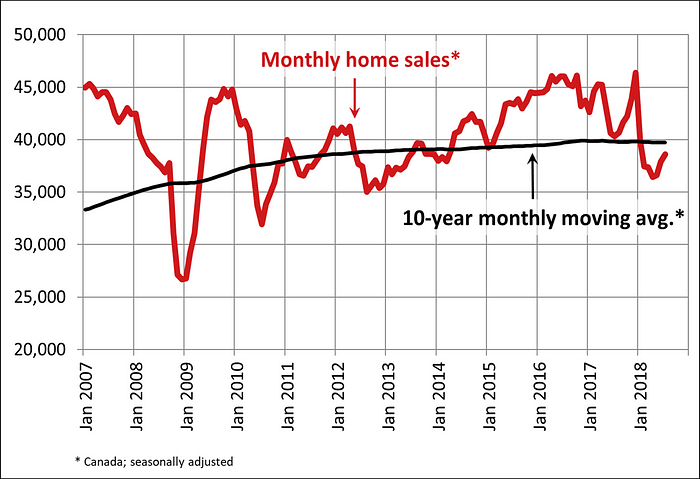
Ordinal Data

For plotting purposes, ordinal data is treated much in the same way as categorical data. But groups are usually ordered from lowest to highest so that we can preserve this ordering.

**Time Series Data**

Time series data is a sequence of numbers collected at regular intervals over some period of time. It is very important, especially in particular fields like finance. Time series data has a temporal value attached to it, so this would be something like a date or a timestamp that you can look for trends in time.

For example, we might measure the average number of home sales for many years. The difference of time series data and numerical data is that rather than having a bunch of numerical values that don’t have any time ordering, time-series data does have some implied ordering. There is a first data point collected and the last data point collected.



CREA

**Text**

Text data is basically just words. A lot of the time the first thing that you do with text is you turn it into numbers using some interesting functions like the bag of words formulation.

These are four types of data from a Machine Learning perspective. Depending on exactly the type of data, this might have some repercussions for the type of algorithms that you can use for feature engineering and modeling, or the type of questions that you can ask of it.

3. Distinguish:

1. Numeric vs. categorical attributes

## **What is categorical data?**

Categorical data can be put in groups or categories using names or labels. This grouping is typically generated using a matching procedure based on data attributes and similarities between these qualities.

Each piece of a categorical dataset, also known as **qualitative data**, may be assigned to only one category based on its qualities, and each category is mutually exclusive.

There are two primary categories of categorical data:

* **Nominal data:** This is the data category that names or labels its categories. It has features resembling a noun and is occasionally referred to as naming data.
* **Ordinary data:** Elements with rankings, orders, or rating scales are included in this category of categorical data. [Nominal data](https://www.questionpro.com/blog/nominal-data/) can be ordered and counted but not measured.

Data expressed in numerical terms rather than in natural language descriptions are called numerical data. It can only be gathered in numerical form, keeping its name. This numerical data type also referred to as quantitative data can be used to measure a person’s height, weight, IQ, etc.

Numerical data can be of two types:

* **Discrete Data:** Countable numerical data are discrete data. They are mapped one-to-one to natural numbers, in other words. Age, the number of students in a class, the number of candidates in an election, etc., are a few examples of discrete data in general.
* **Continuous Data:** This is an uncountable data type for numbers. A series of intervals on a natural number line is used to depict them. Student CGPA, height, and other continuous data types are a few examples.

1. Feature selection vs. dimensionality reduction

# Feature Selection vs Dimensionality Reduction

Often, feature selection and dimensionality reduction are grouped together (like here in this article). While both methods are used for reducing the number of features in a dataset, there is an important difference.

Feature selection is simply selecting and excluding given features **without changing** them.

Dimensionality reduction **transforms** features into a lower dimension.

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

* 1. The histogram

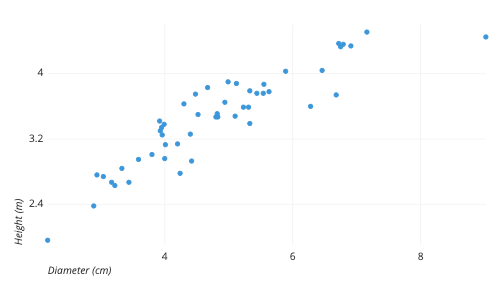
A histogram is a graphical representation of data points organized into user-specified ranges. Similar in appearance to a [bar graph](https://www.investopedia.com/terms/b/bar-graph.asp), the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.

### **KEY TAKEAWAYS**

* A histogram is a bar graph-like representation of data that buckets a range of classes into columns along the horizontal x-axis.
* The vertical y-axis represents the number count or percentage of occurrences in the data for each column
* Columns can be used to visualize patterns of data distributions.
* In trading, the MACD histogram is used by technical analysts to indicate changes in momentum.
* The MACD histogram columns can give earlier buy and sell signals than the accompanying MACD and signal lines.

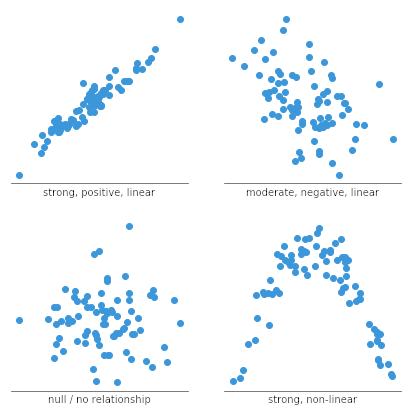
1. Use a scatter plot

A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.

The example scatter plot above shows the diameters and heights for a sample of fictional trees. Each dot represents a single tree; each point’s horizontal position indicates that tree’s diameter (in centimeters) and the vertical position indicates that tree’s height (in meters). From the plot, we can see a generally tight positive correlation between a tree’s diameter and its height. We can also observe an outlier point, a tree that has a much larger diameter than the others. This tree appears fairly short for its girth, which might warrant further investigation.

Scatter plots’ primary uses are to observe and show relationships between two numeric variables. The dots in a scatter plot not only report the values of individual data points, but also patterns when the data are taken as a whole.

Identification of correlational relationships are common with scatter plots. In these cases, we want to know, if we were given a particular horizontal value, what a good prediction would be for the vertical value. You will often see the variable on the horizontal axis denoted an independent variable, and the variable on the vertical axis the dependent variable. Relationships between variables can be described in many ways: positive or negative, strong or weak, linear or nonlinear.



A scatter plot can also be useful for identifying other patterns in data. We can divide data points into groups based on how closely sets of points cluster together. Scatter plots can also show if there are any unexpected gaps in the data and if there are any outlier points. This can be useful if we want to segment the data into different parts, like in the development of user personas.

3.PCA (Personal Computer Aid)

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

At its simplest, data can be broken down into two different categories: [quantitative data](https://www.fullstory.com/quantitative-data/) and qualitative data. But what’s the difference between the two? And when should you use them? And how can you use them together?So let’s demystify the complexities by thoroughly explaining the similarities and differences between qualitative and quantitative data and how they are both crucial to the success of any data research and analysis. Knowing both approaches can help you in understanding your data better—and ultimately understand your customers better.

**Key takeaways:**

* Quantitative data refers to any information that can be quantified, counted or measured, and given a numerical value. Qualitative data is descriptive in nature, expressed in terms of language rather than numerical values.
* Quantitative research is based on numeric data. Qualitative research focuses on the qualities of users—the 'why' behind the numbers.
* It's hard to conduct a successful data analysis without qualitative and quantitative data. They both have their advantages and disadvantages and often complement each other.

**What is quantitative data?**

Qualitative and differ in their approach and the type of data they collect.

Quantitative data refers to any information that can be quantified — that is, numbers. If it can be counted or measured, and given a numerical value, it's quantitative in nature. Think of it as a measuring stick.

Quantitative variables can tell you "how many," "how much," or "how often."

**Some examples of quantitative data**:

* How many people attended last week's webinar?
* How much revenue did our company make last year?
* How often does a customer [rage click](https://www.fullstory.com/blog/rage-clicks-turn-analytics-into-actionable-insights/) on this app?

To analyze these research questions and make sense of this quantitative data, you’d normally use a form of [statistical analysis](https://www.sas.com/en_us/insights/analytics/statistical-analysis.html)—collecting, evaluating, and presenting large amounts of data to discover patterns and trends. Quantitative data is conducive to this type of analysis because it’s numeric and easier to analyze mathematically.

Computers now rule statistical analytics, even though traditional methods have been used for years. But today’s data volumes make statistics more valuable and useful than ever. When you think of statistical analysis now, you think of powerful computers and algorithms that fuel many of the software tools you use today.

Popular quantitative data collection methods are surveys, experiments, polls, and more.

**What is quantitative research?**

It’s all about the numbers. Quantitative research is based on the collection and interpretation of numeric data. It focuses on measuring (using [inferential statistics](https://www.myaccountingcourse.com/accounting-dictionary/inferential-statistics)) and generalizing results.

In terms of digital experience data, it puts everything in terms of numbers (or [discrete data](https://www.isixsigma.com/dictionary/discrete-data/))—like the number of users clicking a button, [bounce rates](https://www.fullstory.com/bounce-rate/), time on site, and more.

**Some examples of quantitative research:**

* What is the amount of money invested into this service?
* What is the average number of times a button was [dead clicked](https://www.fullstory.com/blog/list-your-top-rage-clicks-and-dead-clicks/)?
* How many customers are actually clicking this button?

Essentially, quantitative research is an easy way to see what’s going on at a 20,000-foot view.

Each data set (or customer action, if we’re still talking digital experience) has a numerical value associated with it and is quantifiable information that can be used for calculating statistical analysis so that decisions can be made.

You can use [statistical operations](https://www.statisticshowto.com/operational-statistics-definition/#:~:text=In%20general%2C%20the%20term%20%E2%80%9COperational,example%2C%20a%20production%20line).) to discover feedback patterns (with any representative sample size) in the data under examination. The results can be used to make predictions, find averages, test causes and effects, and generalize results to larger measurable data pools.

Unlike qualitative methodology, quantitative research offers more objective findings as they are based on more reliable numeric data.

**What is qualitative data?**

Unlike quantitative data, qualitative data is descriptive, expressed in terms of language rather than numerical values.

Qualitative data analysis describes information and cannot be measured or counted. It refers to the words or labels used to describe certain characteristics or traits.

Think of qualitative data as the type of data you’d get if you were to ask someone why they did something. Popular data collection methods are in-depth interviews, focus groups, or observation.

### What is qualitative research?

Qualitative research does not simply help to collect data. It gives a chance to understand the trends and meanings of natural actions. It’s flexible and iterative.

Qualitative research focuses on the qualities of users—the actions that drive the numbers. It's descriptive research. The qualitative approach is subjective, too.

It focuses on describing an action, rather than measuring it.

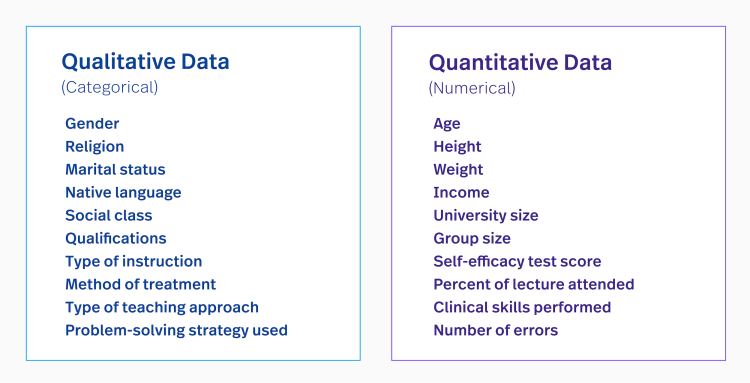
**Some examples of qualitative research:**

* The sunflowers had a fresh smell that filled the office.
* All the bagels with bites taken out of them had cream cheese.
* The man had blonde hair with a blue hat.

## What are the differences between qualitative vs. quantitative data?

When it comes to conducting data research, you’ll need different collection, hypotheses and analysis methods, so it’s important to understand the key differences between quantitative and qualitative data:

* **Quantitative data** is numbers-based, countable, or measurable. **Qualitative data** is interpretation-based, descriptive, and relating to language.
* **Quantitative data** tells us how many, how much, or how often in calculations. **Qualitative data** can help us to understand why, how, or what happened behind certain behaviors.
* **Quantitative data** is fixed and universal. **Qualitative data** is subjective and unique.
* **Quantitative research** methods are measuring and counting. **Qualitative research** methods are interviewing and observing.
* **Quantitative data** is analyzed using statistical analysis. **Qualitative data** is analyzed by grouping the data into categories and themes.



As you can see, both provide immense value for any data collection and are key to truly finding answers and patterns.

### Quantitative Data 101: What is quantitative data?

Take a deeper dive into what quantitative data is, how it works, how to analyze it, collect it, use it, and more.

## What are the advantages and disadvantages of quantitative data?

Each type of data set has its own pros and cons.

### Advantages of quantitative data

* It’s relatively quick and easy to collect and it’s easier to draw conclusions from.
* When you collect quantitative data, the type of results will tell you which statistical tests are appropriate to use.
* As a result, interpreting your data and presenting those findings is straightforward and less open to error and subjectivity.

Another advantage is that you can replicate it. Replicating a study is possible because your data collection is measurable and tangible for further applications.

### Disadvantages of quantitative data

* Quantitative data doesn’t always tell you the full story (no matter what the perspective).
* With choppy information, it can be inconclusive.
* Quantitative research can be limited, which can lead to overlooking broader themes and relationships.
* By focusing solely on numbers, there is a risk of missing larger focus information that can be beneficial.

## What are the advantages and disadvantages of qualitative data?

### Advantages of qualitative data

* Qualitative data offers rich, in-depth insights and allows you to explore context.
* It’s great for exploratory purposes.
* Qualitative research delivers a predictive element for continuous data.

### Disadvantages of qualitative data

* It’s not a statistically representative form of data collection because it relies upon the experience of the host (who can lose data).
* It can also require multiple data sessions, which can lead to misleading conclusions.

The takeaway is that it’s tough to conduct a successful data analysis without both. They both have their advantages and disadvantages and, in a way, they complement each other.

## What are the collection methods of both quantitative and qualitative data?

In order to analyze both types of data, you’ve got to collect the information first, of course.

Qualitative research methods are more flexible and utilize open-ended questions. Quantitative data collection methods focus on highly controlled approaches and numerical information.

### Quantitative data collection methods

#### Surveys

A [survey](https://www.qualtrics.com/experience-management/research/survey-basics/) is one of the most common research methods with quantitative data that involves questioning a large group of people. Questions are usually closed-ended and are the same for all participants. An unclear questionnaire can lead to distorted research outcomes.

#### Polls

Similar to surveys, [polls](https://en.wikipedia.org/wiki/Opinion_poll) yield quantitative data. That is, you poll a number of people and apply a numeric value to how many people responded with each answer.

#### Experiments

An experiment is another common method that usually involves a [control group](https://www.thoughtco.com/what-is-a-control-group-606107) and an [experimental group](https://www.thoughtco.com/what-is-an-experimental-group-606109). The experiment is controlled and the conditions can be manipulated accordingly. You can examine any type of records involved if they pertain to the experiment, so the data is extensive.

Or you can mix it up — use mixed methods of both to combine qualitative and quantitative data.

The best practices of each help to look at the information under a broader lens to get a unique perspective. Using both methods is helpful because they collect rich and reliable data, which can be further tested and replicated.

Controlled experiments, [A/B tests](https://www.fullstory.com/ab-testing/), [blind experiments](https://psychology.wikia.org/wiki/Blind_experiment), and many others fall under this category.

### Qualitative data collection methods

#### Interviews

An [interview](https://guides.lib.vt.edu/researchmethods/interviews#:~:text=Interviews%20are%20most%20effective%20for,depth%20information%20will%20be%20collected.) is the most common qualitative research method. This method involves personal interaction (either in real life or virtually) with a participant. It’s mostly used for exploring attitudes and opinions regarding certain issues.

Interviews are very [popular methods for collecting data in product design](https://www.fullstory.com/blog/product-design-interview-questions/).

#### Focus groups

Data analysis by [focus group](https://www.b2binternational.com/research/methods/faq/what-is-a-focus-group/) is another method where participants are guided by a host to collect data. Within a group (either in person or online), each member shares their opinion and

## What’s an example of the difference between quantitative and qualitative data?

You’ve most likely run into quantitative and qualitative data today, alone. For the visual learner, here are some examples of both quantitative and qualitative data:

**Quantitative data example**

* The customer has clicked on the button 13 times.
* The engineer has resolved 34 support tickets today.
* The team has completed 7 upgrades this month.
* 14 cartons of eggs were purchased this month.

**Qualitative data example**

* My manager has curly brown hair and blue eyes.
* My coworker is funny, loud, and a good listener.
* The customer has a very friendly face and a contagious laugh.
* The eggs were delicious.

The fundamental difference is that one type of data answers primal basics and one answers descriptively.

What does this mean for [data quality](https://www.fullstory.com/blog/data-quality-and-dxi/) and analysis? If you just analyzed quantitative data, you’d be missing core reasons behind what makes a data collection meaningful. You need both in order to truly learn from data—and truly learn from your customers.

### Digital Leadership Webinar: Accelerating Growth with Quantitative Data and Analytics

## Which type is better for data analysis?

So how do you determine which type is better for [data analysis](https://www.coursera.org/articles/what-is-data-analysis-with-examples)?

Quantitative data is structured and accountable. This type of data is formatted in a way so it can be organized, arranged, and searchable. Think about this data as numbers and values found in spreadsheets—after all, you would trust an Excel formula.

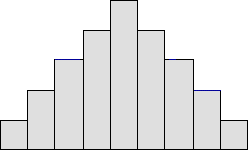
Qualitative data is considered unstructured. This type of data is formatted (and known for) being subjective, individualized, and personalized. Anything goes. Because of this, qualitative data is inferior if it’s the only data in the study. However, it’s still valuable.

Because quantitative data is more concrete, it’s generally preferred for data analysis. Numbers don’t lie. But for complete statistical analysis, using both qualitative and quantitative yields the best results.

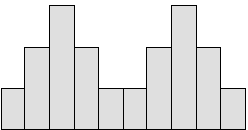
1. What are the various histogram shapes? What exactly are ‘bins'?

A histogram can be created using software such as [SQCpack](https://www.pqsystems.com/quality-solutions/statistical-process-control/SQCpack/?WhereFrom=QualityAdvisor). How would you describe the shape of the histogram?

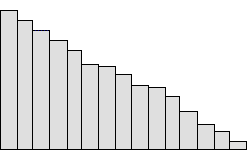
**Bell-shaped:**A bell-shaped picture, shown below, usually presents a normal distribution.



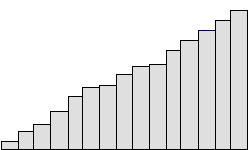
**Bimodal:** A bimodal shape, shown below, has two peaks. This shape may show that the data has come from two different systems. If this shape occurs, the two sources should be separated and analyzed separately.



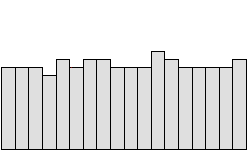
**Skewed right:**Some histograms will show a skewed distribution to the right, as shown below. A distribution skewed to the right is said to be positively skewed. This kind of distribution has a large number of occurrences in the lower value cells (left side) and few in the upper value cells (right side). A skewed distribution can result when data is gathered from a system with has a boundary such as zero. In other words, all the collected data has values greater than zero.



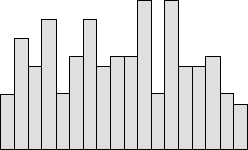
**Skewed left:**Some histograms will show a skewed distribution to the left, as shown below. A distribution skewed to the left is said to be negatively skewed. This kind of distribution has a large number of occurrences in the upper value cells (right side) and few in the lower value cells (left side). A skewed distribution can result when data is gathered from a system with a boundary such as 100. In other words, all the collected data has values less than 100.



**Uniform:**A uniform distribution, as shown below, provides little information about the system. An example would be a state lottery, in which each class has about the same number of elements. It may describe a distribution which has several modes (peaks). If your histogram has this shape, check to see if several sources of variation have been combined. If so, analyze them separately. If multiple sources of variation do not seem to be the cause of this pattern, different groupings can be tried to see if a more useful pattern results. This could be as simple as changing the starting and ending points of the cells, or changing the number of cells. A uniform distribution often means that the number of classes is too small.



**Random:**A random distribution, as shown below, has no apparent pattern. Like the uniform distribution, it may describe a distribution that has several modes (peaks). If your histogram has this shape, check to see if several sources of variation have been combined. If so, analyze them separately. If multiple sources of variation do not seem to be the cause of this pattern, different groupings can be tried to see if a more useful pattern results. This could be as simple as changing the starting and ending points of the cells, or changing the number of cells. A random distribution often means there are too many classes.



A histogram displays numerical data by grouping data into "bins" of equal width. Each bin is plotted as a bar whose height corresponds to how many data points are in that bin.

Bins are also sometimes called "intervals", "classes", or "buckets".

1. How do we deal with data outliers?

When you feel that you have outliers in your data, the most important is to identify the reasons behind these values. This can vary in each case. For instance, in my last job I worked with particle matter (PM) sensors, which simply can estimate the concentration of PM in the air. In this field it was challenging to idenfity outliers because we could have mistakes by the instruments or in some cases a particular event such as a wildfire increased the concentration of PM in the air from a moment to other. In the first case, we recognise these mistakes by other features with problems recorded by the sensors. In conclusion, you should analyse why these "outliers" appears.

As mention before other users, there are different methods to remove outliers. The most commons are the use of the **mean +/- 2 or 3 standard deviation (SD)** and **Q1 − 1.5 IQR or above Q3 + 1.5 IQR (interquartile range**). The first is used when you have data with normal distribution. In other cases, it is recommended to use the IQR method. In fact, this is the typical method used to create boxplot. It is important to know that in statistic and ML several methods are based on the assumption of normality, which is not always meet.

With respect to the remotion of outliers, this should be carefully due to you can delete important data unnecessary. Sometimes, when you have outliers you can replace with a more appropriate value, for instance, considering the values of the nearest neighbors. [KNN](https://towardsdatascience.com/missing-value-imputation-with-python-and-k-nearest-neighbors-308e7abd273d) is an attractive aproach for this. However, this depends on the nature of the feature. For instance, a feature/variable with a expected behaviour as temperature can be easier to imputate using values of neighbors (spatially or temporally neighbors), but for other variables can be a more complex situation and imputation may add noise and bias. In conclusion, I recommend that you review cases applied the field that your are analysing.

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

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics. The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode.

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

## **Mean (Arithmetic)**

The mean (or average) is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data (see our [Types of Variable](https://statistics.laerd.com/statistical-guides/types-of-variable.php) guide for data types). The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have � values in a data set and they have values �1,�2, …,��, the sample mean, usually denoted by �― (pronounced "x bar"), is:

�―=�1+�2+⋯+���

This formula is usually written in a slightly different manner using the Greek capitol letter, ∑, pronounced "sigma", which means "sum of...":

�―=∑��

You may have noticed that the above formula refers to the sample mean. So, why have we called it a sample mean? This is because, in statistics, samples and populations have very different meanings and these differences are very important, even if, in the case of the mean, they are calculated in the same way. To acknowledge that we are calculating the population mean and not the sample mean, we use the Greek lower case letter "mu", denoted as �:

�=∑��

The mean is essentially a model of your data set. It is the value that is most common. You will notice, however, that the mean is not often one of the actual values that you have observed in your data set. However, one of its important properties is that it minimises error in the prediction of any one value in your data set. That is, it is the value that produces the lowest amount of error from all other values in the data set.

An important property of the mean is that it includes every value in your data set as part of the calculation. In addition, the mean is the only measure of central tendency where the sum of the deviations of each value from the mean is always zero.

### When not to use the mean

The mean has one main disadvantage: it is particularly susceptible to the influence of outliers. These are values that are unusual compared to the rest of the data set by being especially small or large in numerical value. For example, consider the wages of staff at a factory below:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Staff | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Salary | 15k | 18k | 16k | 14k | 15k | 15k | 12k | 17k | 90k | 95k |

The mean salary for these ten staff is $30.7k. However, inspecting the raw data suggests that this mean value might not be the best way to accurately reflect the typical salary of a worker, as most workers have salaries in the $12k to 18k range. The mean is being skewed by the two large salaries. Therefore, in this situation, we would like to have a better measure of central tendency. As we will find out later, taking the median would be a better measure of central tendency in this situation.

Another time when we usually prefer the median over the mean (or mode) is when our data is skewed (i.e., the frequency distribution for our data is skewed). If we consider the normal distribution - as this is the most frequently assessed in statistics - when the data is perfectly normal, the mean, median and mode are identical. Moreover, they all represent the most typical value in the data set. However, as the data becomes skewed the mean loses its ability to provide the best central location for the data because the skewed data is dragging it away from the typical value. However, the median best retains this position and is not as strongly influenced by the skewed values. This is explained in more detail in the skewed distribution section later in this guide.

## **Median**

The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data. In order to calculate the median, suppose we have the data below:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 65 | 55 | 89 | 56 | 35 | 14 | 56 | 55 | 87 | 45 | 92 |

We first need to rearrange that data into order of magnitude (smallest first):

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14 | 35 | 45 | 55 | 55 | **56** | 56 | 65 | 87 | 89 | 92 |

Our median mark is the middle mark - in this case, 56 (highlighted in bold). It is the middle mark because there are 5 scores before it and 5 scores after it. This works fine when you have an odd number of scores, but what happens when you have an even number of scores? What if you had only 10 scores? Well, you simply have to take the middle two scores and average the result. So, if we look at the example below:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 65 | 55 | 89 | 56 | 35 | 14 | 56 | 55 | 87 | 45 |

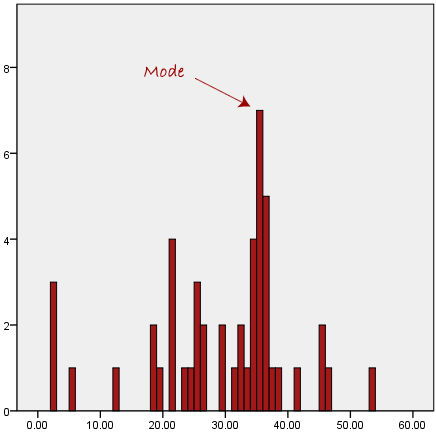
We again rearrange that data into order of magnitude (smallest first):

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14 | 35 | 45 | 55 | **55** | **56** | 56 | 65 | 87 | 89 |

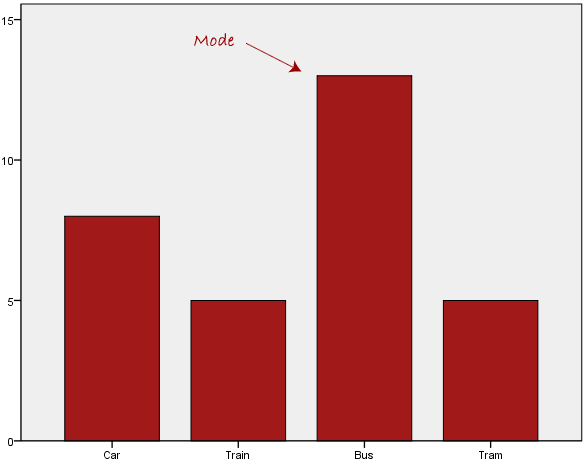
Only now we have to take the 5th and 6th score in our data set and average them to get a median of 55.5.

## **Mode**

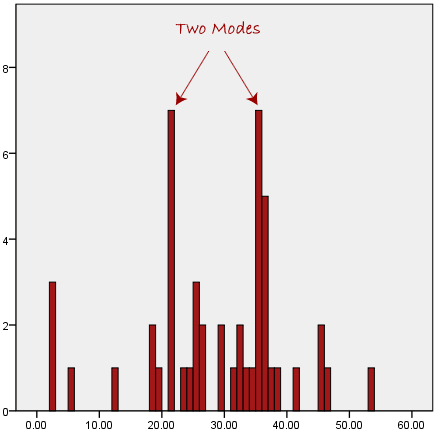
The mode is the most frequent score in our data set. On a histogram it represents the highest bar in a bar chart or histogram. You can, therefore, sometimes consider the mode as being the most popular option. An example of a mode is presented below:



Normally, the mode is used for categorical data where we wish to know which is the most common category, as illustrated below:

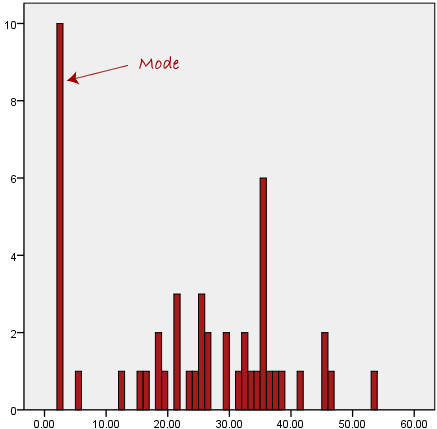


We can see above that the most common form of transport, in this particular data set, is the bus. However, one of the problems with the mode is that it is not unique, so it leaves us with problems when we have two or more values that share the highest frequency, such as below:



We are now stuck as to which mode best describes the central tendency of the data. This is particularly problematic when we have continuous data because we are more likely not to have any one value that is more frequent than the other. For example, consider measuring 30 peoples' weight (to the nearest 0.1 kg). How likely is it that we will find two or more people with **exactly** the same weight (e.g., 67.4 kg)? The answer, is probably very unlikely - many people might be close, but with such a small sample (30 people) and a large range of possible weights, you are unlikely to find two people with exactly the same weight; that is, to the nearest 0.1 kg. This is why the mode is very rarely used with continuous data.

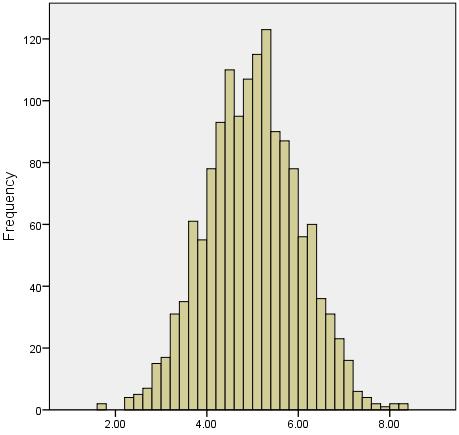
Another problem with the mode is that it will not provide us with a very good measure of central tendency when the most common mark is far away from the rest of the data in the data set, as depicted in the diagram below:



In the above diagram the mode has a value of 2. We can clearly see, however, that the mode is not representative of the data, which is mostly concentrated around the 20 to 30 value range. To use the mode to describe the central tendency of this data set would be misleading.

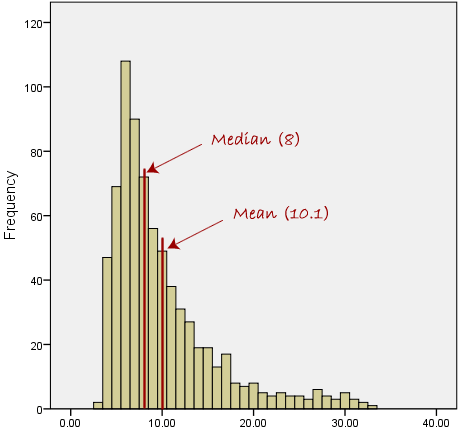
## **Skewed Distributions and the Mean and Median**

We often test whether our data is normally distributed because this is a common assumption underlying many statistical tests. An example of a normally distributed set of data is presented below:



When you have a normally distributed sample you can legitimately use both the mean or the median as your measure of central tendency. In fact, in any symmetrical distribution the mean, median and mode are equal. However, in this situation, the mean is widely preferred as the best measure of central tendency because it is the measure that includes all the values in the data set for its calculation, and any change in any of the scores will affect the value of the mean. This is not the case with the median or mode.

However, when our data is skewed, for example, as with the right-skewed data set below:



We find that the mean is being dragged in the direct of the skew. In these situations, the median is generally considered to be the best representative of the central location of the data. The more skewed the distribution, the greater the difference between the median and mean, and the greater emphasis should be placed on using the median as opposed to the mean. A classic example of the above right-skewed distribution is income (salary), where higher-earners provide a false representation of the typical income if expressed as a mean and not a median.

If dealing with a normal distribution, and tests of normality show that the data is non-normal, it is customary to use the median instead of the mean. However, this is more a rule of thumb than a strict guideline. Sometimes, researchers wish to report the mean of a skewed distribution if the median and mean are not appreciably different (a subjective assessment), and if it allows easier comparisons to previous research to be made.

## **Summary of when to use the mean, median and mode**

Please use the following summary table to know what the best measure of central tendency is with respect to the different [types of variable](https://statistics.laerd.com/statistical-guides/types-of-variable.php).

|  |  |
| --- | --- |
| **Type of Variable** | **Best measure of central tendency** |
| Nominal | Mode |
| Ordinal | Median |
| Interval/Ratio (not skewed) | Mean |
| Interval/Ratio (skewed) | Median |

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

## What is a scatterplot?

A scatterplot is a type of data display that shows the relationship between two numerical variables. Each member of the dataset gets plotted as a point whose (�,�)(*x*,*y*)left parenthesis, x, comma, y, right parenthesis coordinates relates to its values for the two variables.

For example, here is a scatterplot that shows the shoe sizes and quiz scores for students in a class:

111222333444555666777888999101010111111101010202020303030404040505050606060707070808080909090**ScoreShoe size**

A graph plots Score on the y-axis, versus Shoe size on the x-axis. Approximately 2 dozen points are scattered sporadically between x = 5.5 and x = 11, and between y = 52 and y = 87. All values estimated.

Each data point is a student whose �*x*x-coordinate gives their shoe size and �*y*y-coordinate gives their quiz score.

Want to learn more about constructing scatterplots? Check out [*this video*](https://www.khanacademy.org/v/constructing-scatter-plot).

## What is correlation?

We often see patterns or relationships in scatterplots.

When the �*y*y variable tends to increase as the �*x*x variable increases, we say there is a **positive correlation** between the variables.

Positive correlationPositive correlationstart color #1fab54, start text, P, o, s, i, t, i, v, e, space, c, o, r, r, e, l, a, t, i, o, n, end text, end color #1fab54

111222333444555666777888999111222333444555666777888999�*y*y�*x*x

A scatterplot plots points x y axis. Approximately 2 dozen points are rise diagonally in a relatively narrow patterm between (1 half, 1 half) and (9, 7 and 1 half). All values estimated.

When the �*y*y variable tends to decrease as the �*x*x variable increases, we say there is a **negative correlation** between the variables.

Negative correlationNegative correlationstart color #ca337c, start text, N, e, g, a, t, i, v, e, space, c, o, r, r, e, l, a, t, i, o, n, end text, end color #ca337c

111222333444555666777888999111222333444555666777888999�*y*y�*x*x

A scatterplot plots points x y axis. Approximately 2 dozen points are fall diagonally in a relatively narrow patterm between (1 half, 6) and (8, 2). All values estimated.

When there is no clear relationship between the two variables, we say there is **no correlation** between the two variables.

No correlationNo correlationstart color #e07d10, start text, N, o, space, c, o, r, r, e, l, a, t, i, o, n, end text, end color #e07d10

111222333444555666777888999111222333444555666777888999�*y*y�*x*x

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

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.

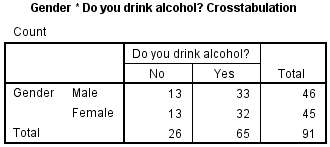
This type of table is also known as a:

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

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.

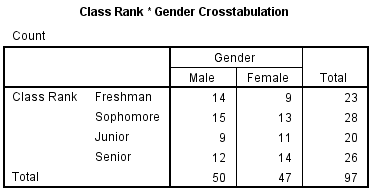
Additionally, a "square" crosstab is one in which the row and column variables have the same number of categories. Tables of dimensions 2x2, 3x3, 4x4, etc. are all square crosstabs.

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



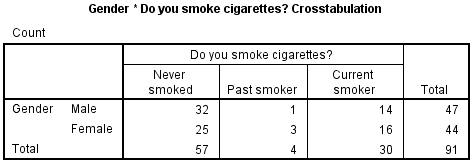
* **Row variable**: Gender (2 categories: male, female)
* **Column variable**: Alcohol (2 categories: no, yes)
* **Table dimension**: 2x2 (square)

#### **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