1. What are the key tasks that machine learning entails? What does data pre-processing imply?

Data preprocessing for [machine learning](https://www.purestorage.com/uk/knowledge/what-is-deep-learning.html) (ML) refers to the preparation and transformation of raw data into a format suitable for training ML models. It’s an essential step in an ML (or AI) [pipeline](https://blog.purestorage.com/perspectives/bytes-ai-data-lifecycle/) because it directly impacts the performance and accuracy of the models.

Data preprocessing involves several techniques such as cleaning the data to handle missing values, removing outliers, scaling features, encoding categorical variables, and splitting the data into training and testing sets. These techniques are key for ensuring the data is in a consistent and usable format for the ML algorithms.

This article covers everything you need to know about data preprocessing for machine learning, including what it is, its benefits, steps, and examples.

**What Is Data Preprocessing?**

Data preprocessing is the transformation of raw data into a format that is more suitable and meaningful for analysis and model training. Data preprocessing plays a vital role in enhancing the quality and [efficiency](https://www.purestorage.com/uk/solutions/analytics-and-ai/artificial-intelligence.html) of ML models by addressing issues such as missing values, noise, inconsistencies, and outliers in the data.

**Benefits of Data Preprocessing for Machine Learning**

Data preprocessing for machine learning has many benefits, and these benefits are the same as the steps involved in data preprocessing. Let’s have a look.

**1. Data Cleaning**

Data cleaning is an essential part of the data preprocessing pipeline in machine learning. It involves identifying and correcting errors or inconsistencies in the data set to ensure that the data is of high quality and suitable for analysis or model training.

Data cleaning typically includes:

**Handling Missing Values**

Missing values are a common issue in real-world data sets and can adversely affect the performance of ML models. To identify and deal with missing values:

* Use descriptive statistics or visualizations to identify columns/features with missing values. Common indicators of missing values include NaN (Not a Number) or NULL values.
* Determine the impact of missing values on your analysis or model. Consider the percentage of missing values in each column and their importance to the overall data set.
* If the percentage of missing values is small and those rows or columns are not critical, you can choose to remove them using methods like dropna() in pandas or similar functions in other tools.
* For numerical features, you can impute missing values using techniques like mean, median, or mode imputation (fillna() method in pandas). For categorical features, you can impute with the most frequent category.

You can also consider more advanced imputation methods such as regression imputation, k-nearest neighbors imputation, or using ML models to predict missing values based on other features.

**Handling Outliers**

Outliers are data points that significantly differ from other observations in the data set and can skew statistical analysis or machine learning models.

To detect and handle outliers:

* Use box plots, histograms, or scatter plots to visualize the distribution of numerical features and identify potential outliers visually.
* Calculate summary statistics like mean, standard deviation, quartiles, and interquartile range (IQR). Outliers are often defined as data points that fall below Q1 - 1.5 \* IQR or above Q3 + 1.5 \* IQR.
* In some cases, removing outliers can be appropriate, especially if they’re due to data entry errors or anomalies. Use filtering techniques based on statistical thresholds to remove outliers.
* Apply transformations like log transformation, square root transformation, or Box-Cox transformation to make the data more normally distributed and reduce the impact of outliers.
* Consider using robust machine learning models that are less sensitive to outliers, such as support vector machines (SVM), Random Forests, or ensemble methods.

**Handling Duplicates**

Duplicate records can skew analysis and model training by inflating certain patterns or biases.

To detect and handle duplicates:

* Use functions like duplicated() in pandas to identify duplicate rows based on specific columns or the entire row.
* If duplicate records are redundant and provide no additional information, you can remove them using the drop\_duplicates() function in pandas or similar methods in other tools.
* In some cases, duplicates may occur due to multiple entries but have unique identifiers. Ensure that you retain unique identifiers or key columns that differentiate between duplicate records.

By following these steps and using appropriate techniques, you can effectively clean and preprocess your data for machine learning tasks, improving the quality and reliability of your models' predictions.

**2. Data Normalization**

Normalization is a data preprocessing technique used to scale and standardize the values of features within a data set. The main goal of normalization is to bring all feature values into a similar range without distorting differences in the ranges of values. This is important because many machine learning algorithms perform better or converge faster when the input features are on a similar scale and have a similar distribution.

Normalization benefits include:

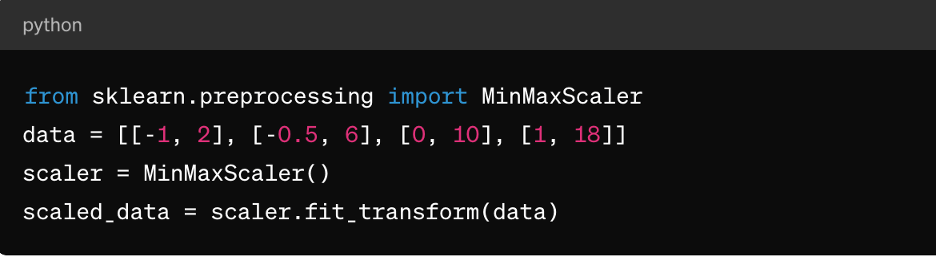
* Helping prevent features with large scales from dominating those with smaller scales during model training.
* Algorithms like gradient descent converge faster when features are normalized, leading to quicker training times.
* Reduction of the impact of outliers by bringing all values within a bounded range. Normalized data can be easier to interpret and compare across different features.

**Normalization Techniques**

Min-max Scaling

* Formula:Xnorm​=Xmax​−Xmin​/Xmax​−Xmin​​
* Range: Transforms values to a range between 0 and 1.

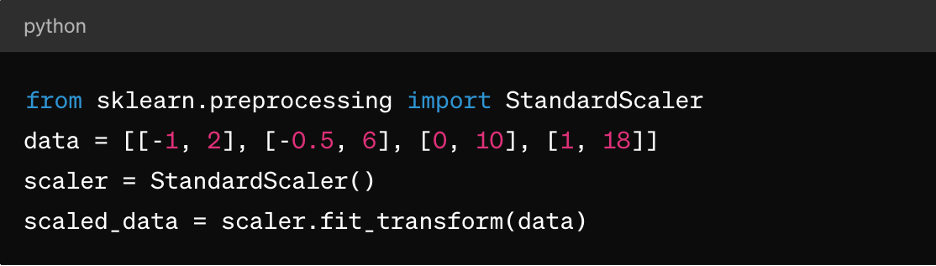
Example:



Z-score Normalization (Standardization):

* ​Formula: Xstd​=σX/μ​
* Range: Transforms values to have a mean of 0 and standard deviation of 1.

Example:



**Guidelines for Applying Normalization**

**Min-max scaling**: Min-max scaling is suitable for algorithms that require input features to be within a specific range, such as neural networks and support vector machines. Make sure outliers are handled appropriately as they can affect the scaling.

**Z-score normalization**: This is suitable for algorithms like k-means clustering, linear regression, and logistic regression. It results in a distribution centered around 0 with a standard deviation of 1, making it ideal for algorithms that assume normally distributed data.

**Sparse data**: For sparse data sets (where most values are zero), consider using techniques like MaxAbsScaler or RobustScaler for normalization.

**Categorical data**: For categorical features, consider techniques like one-hot encoding before normalization to ensure meaningful scaling.

It's important to note that the choice of normalization technique depends on the specific characteristics of your data and the requirements of the machine learning algorithm you plan to use. Experimentation and understanding the impact on model performance are key aspects of applying normalization effectively.

**3. Feature Scaling**

Feature scaling is a data preprocessing technique used to standardize the range of independent variables or features of a data set. The goal of feature scaling is to bring all features to a similar scale or range to avoid one feature dominating over others during model training or analysis. Feature scaling can improve the convergence speed of optimisation algorithms and prevent certain features from having undue influence on the model.

**Role of Feature Scaling in Data Preprocessing**

Scaling features ensures ML algorithms treat all features equally, preventing bias toward features with larger scales. It also enhances convergences, as many optimisation algorithms (e.g., gradient descent) converge faster when features are scaled, leading to quicker model training. It can also prevent numerical instability issues that may arise due to large differences in feature magnitudes. And finally, scaling can make it easier to interpret the impact of features on the model's predictions.

**Feature Scaling Methods**

In addition to the above-described min-max scaling and Z-score normalization, there is also:

**MaxAbsScaler**: This scales each feature by its maximum absolute value, so the resulting values range between -1 and 1. It’s suitable for sparse data where preserving zero entries is important, such as in text classification or recommendation systems.

**RobustScaler**: This uses statistics that are robust to outliers, such as the median and interquartile range (IQR), to scale features. It’s suitable for data sets containing outliers or skewed distributions.

**Guidelines for Applying Feature Scaling**

To apply feature scaling:

* Apply standardization (Z-score normalization) when the data follows a normal distribution or when using algorithms like linear regression, logistic regression, or k-means clustering.
* Apply normalization (min-max scaling) when you need the data to be within a specific range, such as neural networks or support vector machines.
* Use MaxAbsScaler when dealing with sparse data, such as text data or high-dimensional sparse features.
* Use RobustScaler when dealing with data sets containing outliers or non-normally distributed features.

Keep in mind that categorical features may need encoding (e.g., one-hot encoding) before applying feature scaling, especially if they’re nominal (unordered categories).

**4. Handling Categorical Data**

Categorical variables represent groups or categories and are often non-numeric in nature, posing challenges during model training, including:

* **Non-numeric representation**: Categorical variables are typically represented using strings or labels, which most machine learning algorithms cannot directly process. Algorithms require numeric inputs for training and predictions.
* **Ordinal vs. nominal variables**: Categorical variables can be either ordinal (with a meaningful order) or nominal (without a specific order). Treating ordinal variables as nominal or vice versa can lead to incorrect model interpretations or biased predictions.
* **Curse of dimensionality**: One-hot encoding, a common technique for handling categorical data, can lead to an increase in the dimensionality of the data set, especially with a large number of unique categories. This can impact model performance and increase computational complexity.

**Techniques for Encoding Categorical Variables**

Techniques for encoding categorical variables include:

**Label encoding**: Label encoding assigns a unique numerical label to each category in a categorical variable. It’s suitable for ordinal variables where there is a meaningful order among categories.

Here’s an example using Python's scikit-learn:

from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()

encoded\_labels = le.fit\_transform(['cat', 'dog', 'rabbit', 'dog'])

**One-hot encoding**: One-hot encoding creates binary columns for each category in a categorical variable, where each column indicates the presence or absence of that category. It’s suitable for nominal variables without a specific order among categories.

Here’s an example using pandas:

import pandas as pd

df = pd.DataFrame({'category': ['A', 'B', 'C', 'A']})

one\_hot\_encoded = pd.get\_dummies(df['category'], prefix='category')

**Dummy encoding**: Dummy encoding is similar to one-hot encoding but drops one of the binary columns to avoid multicollinearity issues in linear models. It’s commonly used in regression models where one category serves as a reference category.

Here’s an example using pandas:

dummy\_encoded = pd.get\_dummies(df['category'], prefix='category', drop\_first=True)

**Guidelines for Handling Categorical Data**

To correctly handle categorical data, you should:

**Understand variable types**: Determine whether categorical variables are ordinal or nominal to choose the appropriate encoding technique.

**Avoid ordinal misinterpretation**: Be cautious when using label encoding for nominal variables, as it can introduce unintended ordinality in the data.

**Deal with high cardinality**: For categorical variables with a large number of unique categories, consider techniques like frequency encoding, target encoding, or dimensionality reduction techniques such as PCA.

This is all in addition to the already-mentioned handling of missing values and normalizing numerical data.

**5. Dealing with Imbalanced Data**

Dealing with imbalanced data is a common challenge in machine learning, especially in classification tasks where the number of instances in one class (minority class) is significantly lower than in the other classes (majority classes). Imbalanced data can have a profound impact on model training and evaluation, leading to biased models that favor the majority class and perform poorly on minority classes.

Here are some key points regarding imbalanced data and techniques for handling it:

**Impact of Imbalanced Data on Model Performance**

Models trained on imbalanced data tend to prioritize accuracy on the majority class while neglecting the minority class. This can lead to poor performance on minority class predictions. Also, metrics like accuracy can be misleading in imbalanced data sets, as a high accuracy may result from correctly predicting the majority class while ignoring the minority class. Evaluation metrics like precision, recall, F1-score, and area under the ROC curve (AUC-ROC) are more informative for imbalanced data sets compared to accuracy alone.

**Techniques for Handling Imbalanced Data**

The most common techniques for handling imbalanced data are oversampling and undersampling. Oversampling involves increasing the number of instances in the minority class to balance it with the majority class. Undersampling involves reducing the number of instances in the majority class to balance it with the minority class. You can also take a hybrid approach by combining oversampling and undersampling.

There’s also class weighting, where you adjust class weights during model training to penalize errors on the minority class more than errors on the majority class. This is only useful for algorithms that support class weighting, such as logistic regression or support vector machines.

**Guidelines for Handling Imbalanced Data**

To handle imbalanced data, you should:

**Understand data distribution**: Analyse the class distribution in your data set to determine the imbalance severity.

**Choose the appropriate technique**: Select the oversampling, undersampling, or hybrid technique based on your data set size, imbalance ratio, and computational resources.

**Evaluate metrics**: Use appropriate evaluation metrics like precision, recall, F1-score, or AUC-ROC curve to assess model performance on both classes.

**Cross-validate**: Apply techniques within cross-validation folds to avoid data leakage and obtain reliable model performance estimates.

1. Describe quantitative and qualitative data in depth. Make a distinction between the two.

## What is Quantitative Analytics?

How many coins do you have in your pocket? How old is your car? What’s your height and weight? Those are simple examples of using quantitative traits. Each can be expressed as a number.

In business, quantitative analytics uses such traits to create datasets that managers can consider when making strategic decisions. Examples of this could include the following:

* Using web traffic data to identify what areas of a website are most frequently visited.
* Using services such as Google Analytics to determine what search terms most commonly lead customers to your website.
* Using website traffic data to determine when potential customers are abandoning the conversion funnel and not completing the transaction, then using this data to make changes to improve conversion numbers.

In all the above cases, quantitative analysis is used to come up with hard data that leads to better decisions.

Because quantitative analysis strips all issues down to facts and figures, all ambiguity of language, interpretation, and emotion is removed. If done properly using strict rules, smaller datasets can be extrapolated to analyze and [**predict the behavior of larger groups**](https://www.michiganstateuniversityonline.com/resources/business-analytics/using-analytics-to-enhance-marketing-and-sales/).

For example, if 700 out of 1000 potential customers abruptly leave your website from the same page, it’s reasonable to assume this is happening at a similar ratio with larger amounts of customers.

## What is Qualitative Analytics?

Businesses use qualitative analytics to assess situations where hard numbers are impossible. Where quantitative analytics is objective and deductive in assessing a situation, qualitative is subjective and inductive.

Simplified versions of qualitative analysis would include answering questions about the softness of a blanket, the aesthetic value of a new piece of clothing or the impact of a Claude Monet painting. None of these can be expressed in numbers, but it’s possible to have a strong and clear – if subjective – opinion.

In business, gathering information on this type of subjective material is part of qualitative analytics. It could involve:

* Asking customers for information on why they chose your product, which provides a new perspective on your company’s competitive advantage.
* Gathering opinions from customers on the value they see in the product or service offered. Often this perception differs from what was assumed.
* Using written customer feedback on websites to guide future marketing campaigns, as it gives a clearer picture of why customers like your product.

Clearly, these are areas where opinion counts. However, a drawback of qualitative analysis is that findings cannot be applied to larger populations. Just because 50% of one set of customers prefer your product in the color red, that doesn’t mean a similar percentage of a larger group will feel the same way.

Some of the most common methods of gathering qualitative data include:

* Focus group discussions
* In-depth interviews
* Searching for the dominant opinion in chat rooms and online forums
* Seeking the opinion of an online community
* Web trend monitoring

Additionally, most traditional forms of qualitative research employ trained moderators. This helps keep interviewer bias from creeping into the process.

With qualitative analytics, it’s more difficult to generate information that is definitively factual. However, it is the only way to [**create useful data**](https://www.michiganstateuniversityonline.com/resources/business-analytics/actionable-tips-to-analyze-unstructured-data/) in areas that cannot be reduced to numbers.

## ****The Differences Between Quantitative and Qualitative Analytics****

This chart below gives an overview of the differences between quantitative and qualitative analytics.

|  |  |  |
| --- | --- | --- |
|  | Quantitative Analytics | Qualitative Analytics |
| How is data collected? | Close-ended questions with multiple-choice format, surveys, polls or questionnaires. | Open-ended questions with interviews and observations. |
| How is data analyzed? | Mathematical and statistical analysis communicated with numbers, graphs and charts. | Verbal communication and analysis of summarizations, categorizations and interpretations. |
| Advantages | Impartiality, fast and reliable data collection methodology, larger sample sizes. | More detailed insights, methodology encourages deeper discussion. |
| Disadvantages | Unable to learn more context in answers, abnormal research environment, limited answers for data collection and insights. | Smaller sample sizes, more risk of biasness, requires highly skilled moderator. |
| Common industries | Finance, accounting, consulting. | Healthcare, health sciences, social sciences, legal, e-commerce, marketing. |

3. Create a basic data collection that includes some sample records. Have at least one attribute from each of the machine learning data types.

## 0. How to collect data for machine learning if you don’t have any

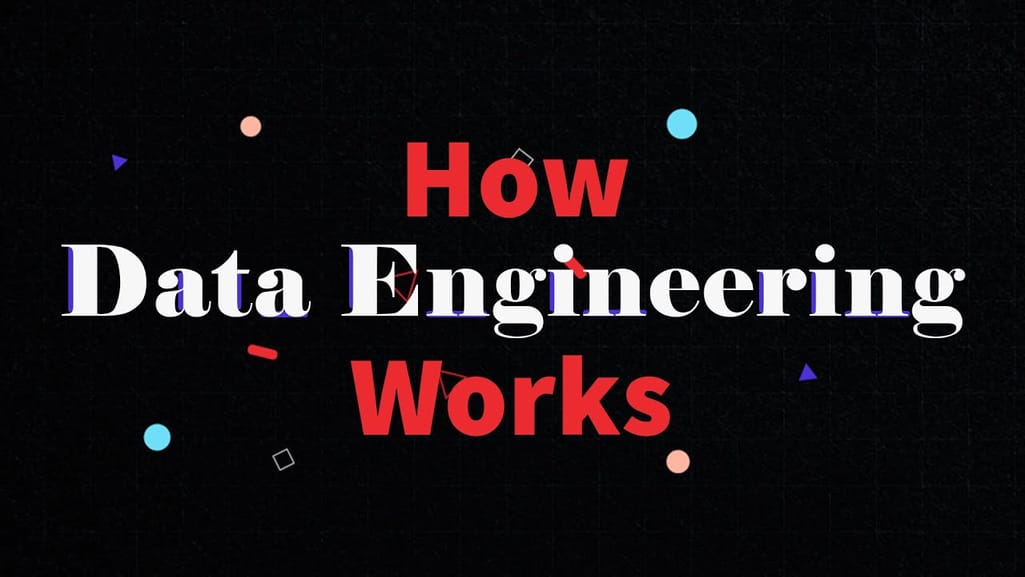
The line dividing those who can play with ML and those who can’t is drawn by years of collecting information. Some organizations have been hoarding records for decades with such great success that now [they need trucks to move it](https://aws.amazon.com/snowmobile/) to the cloud as conventional broadband is just not broad enough.  
  
For those who’ve just come on the scene, lack of data is expected, but fortunately, there are ways to turn that minus into a plus.  
  
First, rely on open source datasets to initiate ML execution. There are mountains of data for machine learning around and some companies (like Google) are ready to give it away. We’ll talk about public dataset opportunities a bit later. While those opportunities exist, usually the real value comes from internally collected golden data nuggets mined from the business decisions and activities of your own company.  
  
Second – and not surprisingly – now you have a chance to collect data the right way. The companies that started data collection with paper ledgers and ended with .xlsx and .csv files will likely have a harder time with data preparation than those who have a small but proud ML-friendly dataset. If you know the tasks that machine learning should solve, you can tailor a data-gathering mechanism in advance.  
  
What about big data? It’s so buzzed, it seems like the thing everyone should be doing. Aiming at [big data](https://www.altexsoft.com/blog/big-data-engineer/) from the start is a good mindset, but big data isn’t about petabytes. It’s all about the ability to process them the right way. The larger your dataset, the harder it gets to make the right use of it and yield insights. Having tons of lumber doesn’t necessarily mean you can convert it to a warehouse full of chairs and tables. So, the general recommendation for beginners is to start small and reduce the complexity of their dat

**1. Articulate the problem early**

Knowing what you want to predict will help you decide which data may be more valuable to collect. When formulating the problem, conduct data exploration and try to think in the categories of classification, clustering, regression, and ranking that we talked about in our whitepaper on [business application of machine learning](https://www.altexsoft.com/whitepapers/machine-learning-bridging-between-business-and-data-science/). In layman’s terms, these tasks are differentiated in the following way:  
  
**Classification.** You want an algorithm to answer binary yes-or-no questions (cats or dogs, good or bad, sheep or goats, you get the idea) or you want to make a multiclass classification (grass, trees, or bushes; cats, dogs, or birds etc.) [You also need the right answers labeled](https://www.altexsoft.com/blog/data-labeling/), so an algorithm can learn from them. Check our guide on how to tackle [data labeling in an organization](https://www.altexsoft.com/blog/datascience/how-to-organize-data-labeling-for-machine-learning-approaches-and-tools/).  
  
**Clustering.** You want an algorithm to find the rules of classification and the number of classes. The main difference from classification tasks is that you don’t actually know what the groups and the principles of their division are. For instance, this usually happens when you need to segment your customers and tailor a specific approach to each segment depending on its qualities.  
  
**Regression.** You want an algorithm to yield some numeric value. For example, if you spend too much time coming up with the right price for your product since it depends on many factors, regression algorithms can aid in estimating this value.  
  
**Ranking.** Some machine learning algorithms just rank objects by a number of features. Ranking is actively used to recommend movies in video streaming services or show the products that a customer might purchase with a high probability based on his or her previous search and purchase activities.  
  
It’s likely that your business problem can be solved within this simple segmentation and you may start adapting a dataset accordingly. The rule of thumb on this stage is to avoid over-complicated problems.

**2. Establish data collection mechanisms**

Creating a data-driven culture in an organization is perhaps the hardest part of the entire initiative. We briefly covered this point in our story on [machine learning strategy](https://www.altexsoft.com/blog/datascience/machine-learning-strategy-7-steps/). If you aim to use ML for predictive analytics, the first thing to do is combat data fragmentation.  
  
For instance, if you look at [travel tech](https://www.altexsoft.com/travel-technology/) – one of AltexSoft’s key areas of expertise – data fragmentation is one of the top analytics problems here. In hotel businesses, the departments that are in charge of physical property get into pretty intimate details about their guests. Hotels know guests’ credit card numbers, types of amenities they choose, sometimes home addresses, room service use, and even drinks and meals ordered during a stay. The website where people book these rooms, however, may treat them as complete strangers.  
  
This data gets siloed in different departments and even different tracking points within a department. Marketers may have access to a CRM but the customers there aren’t associated with web analytics. It’s not always possible to converge all data streams into a centralized storage if you have many channels of engagement, acquisition, and retention, but in most cases it’s manageable.  
  
Usually, collecting data is the work of a [data engineer](https://www.altexsoft.com/blog/datascience/what-is-data-engineering-explaining-data-pipeline-data-warehouse-and-data-engineer-role/), a specialist responsible for creating data infrastructures. But in the early stages, you can engage a software engineer who has some database experience.



**Data engineering, explained**

There are two major types of data collection mechanisms.

### Data Warehouses and ETL

The first one is depositing data in [warehouses](https://www.altexsoft.com/enterprise-data-warehouse-concepts/). These storages are usually created for [structured (or SQL) records](https://www.altexsoft.com/structured-unstructured-data/), meaning they fit into standard table formats. It’s safe to say that all your sales records, payrolls, and CRM data fall into this category. Another traditional attribute of dealing with warehouses is transforming data before loading it there. We’ll talk more about data transformation techniques in this article. But generally it means that you know which data you need and how it must look, so you do all the processing before storing. This approach is called Extract, Transform, and Load (ETL).  
  
The problem with this approach is that you don’t always know in advance which data will be useful and which won’t. So, warehouses are normally used to access data via [business intelligence interfaces](https://www.altexsoft.com/business/complete-guide-to-business-intelligence-and-analytics-strategy-steps-processes-and-tools/) to visualize the metrics we know we need to track. And there’s another way.

### Data Lakes and ELT

Data lakes are storages capable of keeping both structured and unstructured data, including images, videos, sounds records, PDF files... you get the idea. But even if data is structured, it’s not transformed before storing. You would load data there as is and decide how to use and process it later, on demand. This approach is called Extract, Load, and -- then when you need -- Transform.  
  
More on the [difference between ETL and ELT](https://www.altexsoft.com/etl-vs-elt/) you can find in our article. So, what should you choose? Generally, both. Data lakes are  considered a better fit for machine learning. But if you’re confident in at least some data, it’s worth keeping it prepared as you can use it for analytics before you even start any data science initiative.  
  
And keep in mind that modern [cloud data warehouse providers](https://www.altexsoft.com/snowflake-redshift-bigquery-data-warehouse-tools/) support both approaches.

### Handling human factor

Another point here is the human factor. Data collection may be a tedious task that burdens your employees and overwhelms them with instructions. If people must constantly and manually make records, the chances are they will consider these tasks as yet another bureaucratic whim and let the job slide. For instance, Salesforce provides a decent toolset to track and analyze salespeople activities but manual data entry and activity logging alienates salespeople.  
  
This can be solved using [robotic process automation](https://www.altexsoft.com/robotic-process-automation/) systems. RPA algorithms are simple, rule-based bots that can do tedious and repetitive tasks.

## 3. Check your data quality

The first question you should ask -- do you trust your data? Even the most sophisticated machine learning algorithms can’t work with poor data. We’ve talked in detail about [data quality](https://www.altexsoft.com/data-quality-management-and-tools/) in a separate article, but generally you should look at several key things.  
  
**How tangible is human error?**If your data is collected or labeled by humans, check a subset of data and estimate how often mistakes happen.  
  
**Were there any technical problems when transferring data?**For instance, the same records can be duplicated because of server error, or you had a storage crash, or maybe you experienced a cyberattack. Evaluate how these events impacted your data.  
  
**How many omitted values does your data have?**While there are ways to handle omitted records, which we discuss below, estimate whether their number is critical.  
  
**Is your data adequate to your task?**If you’ve been selling home appliances in the US and now plan on branching into Europe, can you use the same data to predict stock and demand?  
  
**Is your data imbalanced?**Imagine that you’re trying to mitigate supply chain risks and filter out those suppliers that you consider unreliable and you use a number of [metadata](https://www.altexsoft.com/metadata-management/) attributes (e.g., location, size, rating, etc.). If your labeled dataset has 1,500 entries labeled as reliable and only 30 that you consider unreliable, the model won’t have enough samples to learn about the unreliable ones.

**4. Format data to make it consistent**

Data formatting is sometimes referred to as the file format you’re using. And this isn’t much of a problem to convert a dataset into a file format that fits your machine learning system best.  
  
We’re talking about format consistency of records themselves. If you’re aggregating data from different sources or your dataset has been manually updated by different people, it’s worth making sure that all variables within a given attribute are consistently written. These may be date formats, sums of money (4.03 or $4.03, or even 4 dollars 3 cents), addresses, etc. The input format should be the same across the entire dataset.  
  
And there are other aspects of data consistency. For instance, if you have a set numeric range in an attribute from 0.0 to 5.0, ensure that there are no 5.5s in your set.

**5. Reduce data**

It’s tempting to include as much data as possible, because of… well, big data! That’s wrong-headed. Yes, you definitely want to collect all data possible. But if you’re preparing a dataset with particular tasks in mind, it’s better to reduce data.  
  
Since you know what the target attribute (what value you want to predict) is, common sense will guide you further. You can assume which values are critical and which are going to add more dimensions and complexity to your dataset without any forecasting contribution.  
  
This approach is called **attribute sampling**.  
  
For example, you want to predict which customers are prone to make large purchases in your online store. The age of your customers, their location, and gender can be better predictors than their credit card numbers. But this also works another way. Consider which other values you may need to collect to uncover more dependencies. For instance, adding bounce rates may increase accuracy in predicting conversion.  
  
That’s the point where domain expertise plays a big role. Returning to our beginning story, not all data scientists know that asthma can cause pneumonia complications. The same works with reducing large datasets. If you haven’t employed a unicorn who has one foot in healthcare basics and the other in data science, it’s likely that a data scientist might have a hard time understanding which values are of real significance to a dataset.  
  
Another approach is called **record sampling**. This implies that you simply remove records (objects) with missing, erroneous, or less representative values to make prediction more accurate. The technique can also be used in the later stages when you need a model prototype to understand whether a chosen machine learning method yields expected results and [estimate ROI of your ML initiative](https://www.altexsoft.com/business/how-to-estimate-roi-and-costs-for-machine-learning-and-data-science-projects/).  
  
You can also reduce data by **aggregating** it into broader records by dividing the entire attribute data into multiple groups and drawing the number for each group. Instead of exploring the most purchased products of a given day through five years of online store existence, aggregate them to weekly or monthly scores. This will help reduce data size and computing time without tangible prediction losses.

**6. Complete data cleaning**

Since missing values can tangibly reduce prediction accuracy, make this issue a priority. In terms of machine learning, assumed or approximated values are “more right” for an algorithm than just missing ones. Even if you don’t know the exact value, methods exist to better “assume” which value is missing or bypass the issue. [How to сlean data](https://www.altexsoft.com/blog/data-cleaning/)? Choosing the right approach also heavily depends on data and the domain you have:

* Substitute missing values with dummy values, e.g., n/a for categorical or 0 for numerical values
* Substitute the missing numerical values with mean figures
* For categorical values, you can also use the most frequent items to fill in.

If you use some ML as a service platform, data cleaning can be automated. For instance, Azure Machine Learning allows you to choose among available techniques, while Amazon ML will do it without your involvement at all. Have a look at our [MLaaS systems comparison](https://www.altexsoft.com/datascience/comparing-machine-learning-as-a-service-amazon-microsoft-azure-google-cloud-ai-ibm-watson/) to get a better idea about systems available on the market.

**7. Create new features out of existing ones**

Some values in your data set can be complex and decomposing them into multiple parts will help in capturing more specific relationships. This process is actually the opposite to reducing data as you have to add new attributes based on the existing ones.  
  
For example, if your sales performance varies depending on the day of a week, segregating the day as a separate categorical value from the date (Mon; 06.19.2017) may provide the algorithm with more relevant information.

**8. Join transactional and attribute data**

Transactional data consists of events that snapshot specific moments, e.g. what was the price of the boots and the time when a user with this IP clicked on the Buy now button?  
  
Attribute data is more static, like user demographics or age and doesn’t directly relate to specific events.  
  
You may have several data sources or logs where these types of data reside. Both types can enhance each other to achieve greater predictive power. For instance, if you’re tracking machinery sensor readings to enable [predictive maintenance](https://www.altexsoft.com/predictive-maintenance/), most likely you’re generating logs of transactional data, but you can add such qualities as the equipment model, the batch, or its location to look for dependencies between equipment behavior and its attributes.  
  
Also you can aggregate transactional data into attributes. Say, you gather website session logs to assign different attributes to different users, e.g., researcher (visits 30 pages on average, rarely buys something), reviews reader (explores the reviews page from top to bottom), instant buyer, etc., then you can use this data to, for example, optimize your retargeting campaigns or predict customer lifetime value.

**9. Rescale data**

Data rescaling belongs to a group of **data normalization** procedures that aim at improving the quality of a dataset by reducing dimensions and avoiding the situation when some of the values overweight others. What does this mean?  
  
Imagine that you run a chain of car dealerships and most of the attributes in your dataset are either categorical to depict models and body styles (sedan, hatchback, van, etc.) or have 1-2 digit numbers, for instance, for years of use. But the prices are 4-5 digit numbers ($10000 or $8000) and you want to predict the average time for the car to be sold based on its characteristics (model, years of previous use, body style, price, condition, etc.) While the price is an important criterion, you don’t want it to overweight the other ones with a larger number.  
  
In this case, **min-max normalization** can be used. It entails transforming numerical values to ranges, e.g., from 0.0 to 1.0 where 0.0 represents the minimal and 1.0 the maximum values to even out the weight of the price attribute with other attributes in a dataset.  
  
A bit simpler approach is **decimal scaling**. It entails scaling data by moving a decimal point in either direction for the same purposes.

**10. Discretize data**

Sometimes you can be more effective in your predictions if you turn numerical values into categorical values. This can be achieved, for example, by dividing the entire range of values into a number of groups.  
  
If you track customer age figures, there isn’t a big difference between the age of 13 and 14 or 26 and 27. So these can be converted into relevant age groups. Making the values categorical, you simplify the work for an algorithm and essentially make prediction more relevant.

## Public datasets

Your private datasets capture the specifics of your unique business and potentially have all relevant attributes that you might need for predictions. But when can you use public datasets?  
  
Public datasets come from organizations and businesses that are open enough to share. The sets usually contain information about general processes in a wide range of life areas like healthcare records, historical weather records, transportation measurements, text and translation collections, records of hardware use, etc. Though these won’t help capture data dependencies in your own business, they can yield great insight into your industry and its niche, and, sometimes, your customer segments.  
  
To learn more about open data sources, consider checking our article about the [best public datasets](https://www.altexsoft.com/datascience/best-public-machine-learning-datasets/) and resources that store this data.  
  
Another use case for public datasets comes from startups and businesses that use machine learning techniques to ship ML-based products to their customers. If you recommend city attractions and restaurants based on user-generated content, you don’t have to label thousands of pictures to train an image recognition algorithm that will sort through photos sent by users. There’s an [Open Images dataset](https://github.com/openimages/dataset) from Google. Similar datasets exist for [speech](https://catalog.ldc.upenn.edu/LDC2002T43) and [text](https://drive.google.com/drive/u/0/folders/0Bz8a_Dbh9Qhbfll6bVpmNUtUcFdjYmF2SEpmZUZUcVNiMUw1TWN6RDV3a0JHT3kxLVhVR2M) recognition. You can also find a [public datasets compilation on GitHub](https://github.com/caesar0301/awesome-public-datasets). Some of the public [datasets are commercial](https://cloud.google.com/commercial-datasets/) and will cost you money.  
  
So, even if you haven’t been collecting data for years, go ahead and search. There may be sets that you can use right away.

4. What are the various causes of machine learning data issues? What are the ramifications?

## ****Common Issues in Machine Learning****

Machine Learning (ML) has undoubtedly transformed industries by enabling data-driven decision-making. However, it's crucial to acknowledge the practical challenges that professionals face while honing ML skills and developing applications from scratch. In this discussion, we'll delve into common issues encountered in the realm of Machine Learning, offering a pragmatic viewpoint without embellishing the complexities.



**1. Inadequate Training Data**

The backbone of any [ML algorithm](https://iabac.org/blog/types-of-machine-learning-algorithms) is the data it is trained on. The challenge arises when there is a shortage of both quality and quantity in the training dataset. Noisy, incorrect, or unclean data can significantly impact the effectiveness of ML algorithms. Addressing issues such as noisy data, inaccuracies, and difficulties in generalizing output data becomes paramount for accurate predictions.

**2. Poor Quality of Data**

Data quality is a recurring issue, with noisy, incomplete, and inaccurate data undermining the accuracy of classification and overall results. Achieving high-quality data is essential for the success of ML models, necessitating a meticulous approach to data preparation.

**3. Non-representative Training Data**

The representativeness of training data directly influences the generalization capability of ML models. If training data fails to cover all relevant cases, the model may produce less accurate predictions, leading to bias against specific classes or groups. Using representative data in training mitigates biases and enhances prediction accuracy.

**4. Overfitting and Underfitting**

Overfitting occurs when a model captures noise and inaccuracies from a large dataset, adversely affecting its performance. This can be mitigated by employing linear and parametric algorithms, increasing training data, or reducing model complexity. Conversely, underfitting arises from a model being too simple for the data, resulting in incomplete and inaccurate predictions. Methods to address underfitting include increasing model complexity, using better features, and adjusting constraints.

**5. Monitoring and Maintenance**

Regular monitoring and maintenance are essential to ensure the continued effectiveness of ML models. Changes in data or user expectations may necessitate code adjustments and resource updates, emphasizing the need for ongoing vigilance.

**6. Getting Bad Recommendations**

ML models operating in a specific context may provide outdated or irrelevant recommendations, known as data drift. Regularly updating and monitoring data helps mitigate this issue, ensuring recommendations align with current user expectations.

**7. Lack of Skilled Resources**

The shortage of skilled professionals with in-depth knowledge of mathematics, science, and technology poses a challenge in the ML industry. Addressing this gap requires investing in training and education to cultivate a workforce equipped to handle the intricacies of ML.

**8. Customer Segmentation**

Accurate customer segmentation is crucial for effective ML algorithms. Developing algorithms that recognize customer behavior and trigger relevant recommendations based on past experiences is essential for personalized user interactions.

**9. Process Complexity of Machine Learning**

The complexity of the ML process, marked by experimental phases and continuous changes, presents a challenge for engineers and data scientists. The evolving nature of ML and the multitude of experiments contribute to a higher probability of errors, making the process intricate and demanding.

**10. Data Bias**

Data bias introduces errors when certain elements in the dataset are given disproportionate weight. Detecting and mitigating bias requires careful examination of the dataset, regular analysis, and implementing strategies to ensure data diversity.

While machine learning has revolutionized industries, it grapples with challenges such as inadequate training data, data quality issues, and algorithmic biases. These practical hurdles require a pragmatic approach, emphasizing the importance of high-quality, representative data, and ongoing model monitoring. Addressing these issues fosters the responsible development and deployment of machine learning applications, ensuring they contribute positively to diverse sectors while mitigating ethical and operational concerns.

5. Demonstrate various approaches to categorical data exploration with appropriate examples.

# Exploring Categorical Data

**Last Updated :**25 Jul, 2022

**Categorical Variable/Data (or Nominal variable):**Such variables take on a fixed and limited number of possible values. For example – grades, gender, blood group type, etc. Also, in the case of categorical variables, the logical order is not the same as categorical data e.g. “one”, “two”, “three”. But the sorting of these variables uses logical order. For example, gender is a categorical variable and has categories – male and female and there is no intrinsic ordering to the categories. A purely categorical variable is one that simply allows you to assign categories, but you cannot clearly order the variables. **Terms related to Variability Metrics :**

* **Mode :**Most frequently occurring value in the given data **Example-**

Data = ["Car", "Bat", "Bat", "Car", "Bat", "Bat", "Bat", "Bike"]

Mode = "Bat"

* **Expected Value :**When working in machine learning, categories have to be associated with a numeric value, so as to give understanding to the machine. This gives an average value based on a category’s probability of occurrence i.e. Expected Value. It is calculated by –

-> Multiply each outcome by its probability of occurring.

-> Sum these values

* So, it is the sum of values times their probability of occurrence often used to sum up factor variable levels.
* **Bar Charts :**Frequency of each category plotted as bars. Loading Libraries –
* Python3

|  |
| --- |
| import matplotlib.pyplot as plt  import numpy as np |

* Data –
* Python3

|  |
| --- |
| label = ['Car', 'Bike', 'Truck', 'Cycle', 'Jeeps', 'Ambulance']  no\_vehicle = [941, 854, 4595, 2125, 942, 509] |

* Indexing Data –
* Python3

|  |
| --- |
| index = np.arange(len(label))    print ("Total Labels : ", len(label))  print ("Indexing : ", index) |

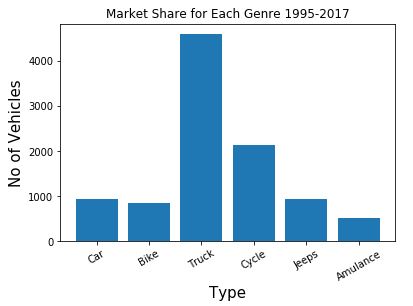
* **Output:**

Total Labels : 6

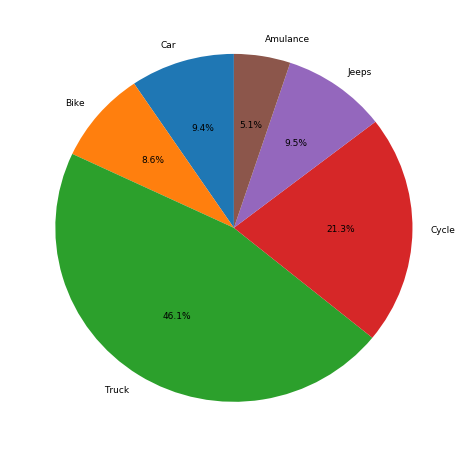
Indexing : [0 1 2 3 4 5]

* Bar Graph –
* Python3

|  |
| --- |
| plt.bar(index, no\_vehicle)  plt.xlabel('Type', fontsize = 15)  plt.ylabel('No of Vehicles', fontsize = 15)  plt.xticks(index, label, fontsize = 10, rotation = 30)  plt.title('Market Share for Each Genre 1995-2017')    plt.show() |

* **Output:**
* **Pie Charts :**Frequency of each category plotted as pie or wedges. It is a circular graph, where the arc length of each slice is proportional to the quantity it represents.
* Python3

|  |
| --- |
| plt.figure(figsize =(8, 8))  plt.pie(no\_vehicle, labels = label,          startangle = 90, autopct ='%.1f %%')  plt.show() |

* **Output:**

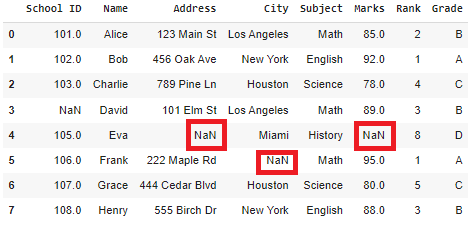
6. How would the learning activity be affected if certain variables have missing values? Having said that, what can be done about it?

# ML | Handling Missing Values

Missing values are a common issue in machine learning. This occurs when a particular variable lacks data points, resulting in incomplete information and potentially harming the accuracy and dependability of your models. It is essential to address missing values efficiently to ensure strong and impartial results in your machine-learning projects. In this article, we will see How to Handle Missing Values in Datasets in [Machine Learning](https://www.geeksforgeeks.org/machine-learning/).

## What is a Missing Value?

Missing values are data points that are absent for a specific variable in a dataset. They can be represented in various ways, such as blank cells, null values, or special symbols like “NA” or “unknown.” These missing data points pose a significant challenge in data analysis and can lead to inaccurate or biased results.



*Missing Values*

Missing values can pose a significant challenge in data analysis, as they can:

* **Reduce the sample size:** This can decrease the accuracy and reliability of your analysis.
* **Introduce bias:** If the missing data is not handled properly, it can bias the results of your analysis.
* **Make it difficult to perform certain analyses:** Some statistical techniques require complete data for all variables, making them inapplicable when missing values are present

### Why Is Data Missing From the Dataset?

Data can be missing for many reasons like technical issues, human errors, privacy concerns, data processing issues, or the nature of the variable itself. Understanding the cause of missing data helps choose appropriate handling strategies and ensure the quality of your analysis.

**It’s important to understand the reasons behind missing data:**

* **Identifying the type of missing data:** Is it Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR)?
* **Evaluating the impact of missing data:** Is the missingness causing bias or affecting the analysis?
* **Choosing appropriate handling strategies:** Different techniques are suitable for different types of missing data.

### Types of Missing Values

There are three main types of missing values:

1. **Missing Completely at Random (MCAR):**MCAR is a specific type of missing data in which the probability of a data point being missing is entirely random and independent of any other variable in the dataset. In simpler terms, whether a value is missing or not has nothing to do with the values of other variables or the characteristics of the data point itself.
2. **Missing at Random (MAR):**MAR is a type of missing data where the probability of a data point missing depends on the values of other variables in the dataset, but not on the missing variable itself. This means that the missingness mechanism is not entirely random, but it can be predicted based on the available information.
3. **Missing Not at Random (MNAR):**MNAR is the most challenging type of missing data to deal with. It occurs when the probability of a data point being missing is related to the missing value itself. This means that the reason for the missing data is informative and directly associated with the variable that is missing.

### Methods for Identifying Missing Data

Locating and understanding patterns of missingness in the dataset is an important step in addressing its impact on analysis. [Working with Missing Data in Pandas](https://www.geeksforgeeks.org/working-with-missing-data-in-pandas/) there are several useful functions for detecting, removing, and replacing null values in Pandas DataFrame.

| **Functions** | **Descriptions** |
| --- | --- |
| **.isnull()** | Identifies missing values in a Series or DataFrame. |
| **.notnull()** | check for missing values in a pandas Series or DataFrame. It returns a boolean Series or DataFrame, where True indicates non-missing values and False indicates missing values. |
| **.info()** | Displays information about the DataFrame, including data types, memory usage, and presence of missing values. |
| **.isna()** | similar to notnull() but returns True for missing values and False for non-missing values. |
| **dropna()** | Drops rows or columns containing missing values based on custom criteria. |
| **fillna()** | Fills missing values with specific values, means, medians, or other calculated values. |
| **replace()** | Replaces specific values with other values, facilitating data correction and standardization. |
| **drop\_duplicates()** | Removes duplicate rows based on specified columns. |
| **unique()** | Finds unique values in a Series or DataFrame. |

### How Is a Missing Value Represented in a Dataset?

Missing values can be represented by blank cells, specific values like “NA”, or codes. It’s important to use consistent and documented representation to ensure transparency and facilitate data handling.

**Common Representations**

1. **Blank cells:** Empty cells in spreadsheets or databases often signify missing data.
2. **Specific values:** Special values like “NULL”, “NA”, or “-999” are used to represent missing data explicitly.
3. **Codes or flags:** Non-numeric codes or flags can be used to indicate different types of missing values.

7. Describe the various methods for dealing with missing data values in depth.

## Imputation vs. Removing Data

When dealing with missing data, [data scientists](https://www.mastersindatascience.org/careers/data-scientist/) can use two primary methods to solve the error: imputation or data removal.

The imputation method substitutes reasonable guesses for missing data. It’s most useful when the percentage of missing data is low. If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

The other option is to remove data. When dealing with data that is missing at random, the entire data point that is missing information can be deleted to help reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. In some situations, observation of specific events or factors may be required, even if incomplete.

Before deciding which approach to employ, it helps to understand why the data is missing.

### Missing at Random (MAR)

[Missing at Random (MAR) means the probability of data being missing is relative to a variable where there is complete informationExternal link:open\_in\_new](https://stefvanbuuren.name/fimd/sec-MCAR.html). The data is not missing across all observations but only within sub-samples of the data. The missing data can be predicted based on the complete observed data. For example, in a survey of the general population, we might be missing data from certain populations because of a known property, such as responsiveness among men.

### Missing Completely at Random (MCAR)

When data is [MCAR, the data is missing across all observations regardless of the expected value or other variablesExternal link:open\_in\_new](https://www.ncbi.nlm.nih.gov/books/NBK493614/). Data scientists can compare two sets of data, one with missing observations and one without. Using a t-test, if there is no difference between the two data sets, the data is characterized as MCAR.

Data may be missing due to test design, failure in the observations or failure in recording observations. This type of data is seen as MCAR because the reasons for its absence are external and not related to the value of the observation.

It is typically safe to remove MCAR data because the results will be unbiased. The test may not be as powerful, but the results can still be reliable.

### Missing Not at Random (MNAR)

The [MNAR category applies when the probability that data is missing seems to be dependent on the unobserved or missing values themselvesExternal link:open\_in\_new](https://www.bookdown.org/rwnahhas/RMPH/mi-mechanisms.html). For example, if people with weaker opinions are less likely to answer, then that data is MNAR. Like MAR, the data cannot be determined by the observed data, because the missing information is unknown. Data scientists must model the missing data to develop an unbiased estimate. Simply removing observations with missing data could result in a model with bias.

## Deletion

There are three primary methods for deleting data when dealing with missing data: listwise, pairwise and dropping variables.

### Listwise

In this method, all data for an observation that has one or more missing values are deleted. The analysis is run only on observations that have a complete set of data. If the data set is small, it may be the most efficient method to eliminate those cases from the analysis. However, in many cases, the data are not missing completely at random (MCAR). Deleting the instances with missing observations can result in biased parameters and estimates and reduce the statistical power of the analysis.

### Pairwise

[Pairwise deletion assumes data are missing completely at random (MCAR) and the statistical analysis uses all cases with data, even if some is missingExternal link:open\_in\_new](https://www.ibm.com/support/pages/pairwise-vs-listwise-deletion-what-are-they-and-when-should-i-use-them). Pairwise deletion allows data scientists to use more of the data. However, the resulting statistics may vary because they are based on different data sets. The results may be impossible to duplicate with a complete set of data.

### Dropping Variables

If data is missing for a large proportion of the observations, it may be best to discard the variable entirely if it is insignificant.

## SPONSORED SCHOOLS

## The London School of Economics and Political Science

info

### LSE Applied Data Analysis

Discover how to harness the enormous potential of data with the London School of Economics and Political Science (LSE) Applied Data Visualisation and Analysis for Business online certificate course.

* Become proficient in Tableau and Microsoft Excel
* 6 weeks, excluding 1 week orientation
* 4–7 hours of self-paced learning per week, entirely online

[Learn More from the London School of Economics and Political ScienceExternal link:](https://onlinecertificatecourses.lse.ac.uk/presentations/lp/lse-data-science-text-analysis-using-r-online-certificate-course/?lsrc=mastersdatasciencesite&l=how-to-deal-with-missing-data)

## Rice University

info

### Rice Data Analysis and Visualization

Reduce decision-making uncertainty with a data analysis and visualization toolkit.

[Learn more from Rice University.External link:](https://online-short-courses.rice.edu/presentations/lp/rice-data-analysis-and-visualization-online-short-course/?utm_source=minisites&utm_medium=www.mastersindatascience.org&utm_campaign=rice_dav&utm_content=how-to-deal-with-missing-data)

infoSPONSORED

## Imputation

When data is missing, it may make sense to delete data, as mentioned above. However, that may not be the most effective option. For example, if too much information is discarded, it may not be possible to complete a reliable analysis. Or there may be insufficient data to generate a reliable prediction for observations that have missing data.

Instead of deletion, data scientists have multiple solutions to impute the value of missing data. Depending why the data are missing, imputation methods can deliver reasonably reliable results. These are examples of single imputation methods for replacing missing data.

### Mean, Median and Mode

This is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations, data scientists can [calculate the mean or median of the existing observations and insert them in place of the missing observationsExternal link:open\_in\_new](https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4). However, when there are many missing variables, mean or median results can result in a loss of variation in the data. This method does not use time-series characteristics or depend on the relationship between the variables.

### Time-Series Specific Methods

Another option is to use time-series specific methods when appropriate to impute data.

The time series methods of imputation assume the adjacent observations will be like the missing data. These methods work well when that assumption is valid. However, these methods won’t always produce reasonable results, particularly in the case of strong seasonality.

### Last Observation Carried Forward (LOCF) & Next Observation Carried Backward (NOCB)

These options are used to analyze longitudinal repeated measures data, in which follow-up observations may be missing. In these methods, every missing value is replaced with either the last observed value or the next one. Longitudinal data track the same instance at different points along a timeline. This method is easy to understand and implement. However, this method may introduce bias when data has a visible trend. It assumes the value is unchanged by the missing data.

### Linear Interpolation

Linear interpolation is often used to approximate a value of some function by using two known values of that function at other points. This formula can also be understood as a weighted average. The weights are inversely related to the distance from the end points to the unknown point. The closer point has more influence than the farther point.

When dealing with missing data, you might use this method in a time series that exhibits a trend line, but it’s not appropriate for seasonal data.

### Seasonal Adjustment with Linear Interpolation

When dealing with data that exhibits both trend and seasonality characteristics, consider using seasonal adjustment with linear interpolation. First, you would perform the seasonal adjustment by computing a centered moving average or taking the average of multiple averages—for example, two one-year averages—that are offset by one period relative to another. You can then complete data smoothing with linear interpolation as discussed above.

## Multiple Imputation

Multiple imputation is considered a good approach for data sets with a large amount of missing data. Instead of substituting a single value for each missing data point, the [missing values are exchanged for values that](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/)encompass the natural variability and uncertainty of the right values. Using the imputed data, the process is repeated to make multiple imputed data sets. Each set is then analyzed using the standard analytical procedures and the results are combined to produce an overall result.

The various imputations incorporate natural variability into the missing values, which creates a valid statistical inference. Multiple imputations can produce statistically valid results even when there is a small sample size or a large amount of missing data.

## K-Nearest Neighbors

In this method, data scientists determine a data point’s nearest neighbors and approximate an estimate based on the values of points closest and other variables. The data scientist must select the number of nearest neighbors and the distance metric. KNN can identify the most frequent value among the neighbors and the mean among the nearest neighbors.

## Learn More About Data Science

When working as a data scientist, you often will be faced with imperfect data sets. Analyzing data with missing information is an important part of work as a data scientist. Advancing your career in data science can help you learn how to tackle these issues and more.

8. What are the various data pre-processing techniques? Explain dimensionality reduction and function selection in a few words.

# Introduction to Dimensionality Reduction

**Last Updated :**06 May, 2023

**Machine Learning:**As discussed in this [article](https://www.geeksforgeeks.org/demystifying-machine-learning/), machine learning is nothing but a field of study which allows computers to “learn” like humans without any need of explicit programming.

**What is Predictive Modeling:**Predictive modeling is a probabilistic process that allows us to forecast outcomes, on the basis of some predictors. These predictors are basically features that come into play when deciding the final result, i.e. the outcome of the model.

Dimensionality reduction is the process of reducing the number of features (or dimensions) in a dataset while retaining as much information as possible. This can be done for a variety of reasons, such as to reduce the complexity of a model, to improve the performance of a learning algorithm, or to make it easier to visualize the data. There are several techniques for dimensionality reduction, including principal component analysis (PCA), singular value decomposition (SVD), and linear discriminant analysis (LDA). Each technique uses a different method to project the data onto a lower-dimensional space while preserving important information.

**What is Dimensionality Reduction?**

Dimensionality reduction is a technique used to reduce the number of features in a dataset while retaining as much of the important information as possible. In other words, it is a process of transforming high-dimensional data into a lower-dimensional space that still preserves the essence of the original data.

In machine learning, high-dimensional data refers to data with a large number of features or variables. The curse of dimensionality is a common problem in machine learning, where the performance of the model deteriorates as the number of features increases. This is because the complexity of the model increases with the number of features, and it becomes more difficult to find a good solution. In addition, high-dimensional data can also lead to overfitting, where the model fits the training data too closely and does not generalize well to new data.

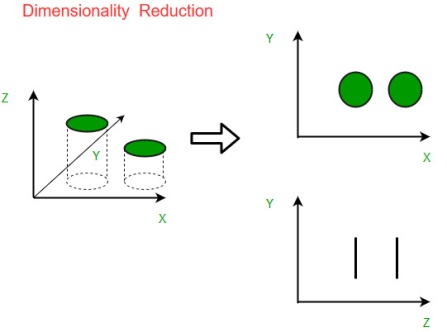
Dimensionality reduction can help to mitigate these problems by reducing the complexity of the model and improving its generalization performance. There are two main approaches to dimensionality reduction: feature selection and feature extraction.

Feature Selection:  
Feature selection involves selecting a subset of the original features that are most relevant to the problem at hand. The goal is to reduce the dimensionality of the dataset while retaining the most important features. There are several methods for feature selection, including filter methods, wrapper methods, and embedded methods. Filter methods rank the features based on their relevance to the target variable, wrapper methods use the model performance as the criteria for selecting features, and embedded methods combine feature selection with the model training process.

Feature Extraction:  
Feature extraction involves creating new features by combining or transforming the original features. The goal is to create a set of features that captures the essence of the original data in a lower-dimensional space. There are several methods for feature extraction, including principal component analysis (PCA), linear discriminant analysis (LDA), and t-distributed stochastic neighbor embedding (t-SNE). PCA is a popular technique that projects the original features onto a lower-dimensional space while preserving as much of the variance as possible.

**Why is Dimensionality Reduction important in Machine Learning and Predictive Modeling?**

An intuitive example of dimensionality reduction can be discussed through a simple e-mail classification problem, where we need to classify whether the e-mail is spam or not. This can involve a large number of features, such as whether or not the e-mail has a generic title, the content of the e-mail, whether the e-mail uses a template, etc. However, some of these features may overlap. In another condition, a classification problem that relies on both humidity and rainfall can be collapsed into just one underlying feature, since both of the aforementioned are correlated to a high degree. Hence, we can reduce the number of features in such problems. A 3-D classification problem can be hard to visualize, whereas a 2-D one can be mapped to a simple 2-dimensional space, and a 1-D problem to a simple line. The below figure illustrates this concept, where a 3-D feature space is split into two 2-D feature spaces, and later, if found to be correlated, the number of features can be reduced even further.



**Components of Dimensionality Reduction**

There are two components of dimensionality reduction:

* **Feature selection:** In this, we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem. It usually involves three ways:
  1. Filter
  2. Wrapper
  3. Embedded
* **Feature extraction:** This reduces the data in a high dimensional space to a lower dimension space, i.e. a space with lesser no. of dimensions.

**Methods of Dimensionality Reduction**

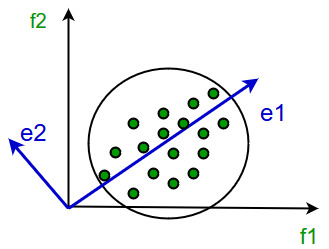
The various methods used for dimensionality reduction include:

* Principal Component Analysis (PCA)
* Linear Discriminant Analysis (LDA)
* Generalized Discriminant Analysis (GDA)

Dimensionality reduction may be both linear and non-linear, depending upon the method used. The prime linear method, called Principal Component Analysis, or PCA, is discussed below.

**Principal Component Analysis**

This method was introduced by Karl Pearson. 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.



It involves the following steps:

* Construct the covariance matrix of the data.
* Compute the eigenvectors of this matrix.
* Eigenvectors corresponding to the largest eigenvalues are used to reconstruct a large fraction of variance of the original data.

Hence, we are left with a lesser number of eigenvectors, and there might have been some data loss in the process. But, the most important variances should be retained by the remaining eigenvectors.

**Advantages of Dimensionality Reduction**

* It helps in data compression, and hence reduced storage space.
* It reduces computation time.
* It also helps remove redundant features, if any.
* Improved Visualization: High dimensional data is difficult to visualize, and dimensionality reduction techniques can help in visualizing the data in 2D or 3D, which can help in better understanding and analysis.
* Overfitting Prevention: High dimensional data may lead to overfitting in machine learning models, which can lead to poor generalization performance. Dimensionality reduction can help in reducing the complexity of the data, and hence prevent overfitting.
* Feature Extraction: Dimensionality reduction can help in extracting important features from high dimensional data, which can be useful in feature selection for machine learning models.
* Data Preprocessing: Dimensionality reduction can be used as a preprocessing step before applying machine learning algorithms to reduce the dimensionality of the data and hence improve the performance of the model.
* Improved Performance: Dimensionality reduction can help in improving the performance of machine learning models by reducing the complexity of the data, and hence reducing the noise and irrelevant information in the data.

**Disadvantages of Dimensionality Reduction**

* It may lead to some amount of data loss.
* PCA tends to find linear correlations between variables, which is sometimes undesirable.
* PCA fails in cases where mean and covariance are not enough to define datasets.
* We may not know how many principal components to keep- in practice, some thumb rules are applied.
* Interpretability: The reduced dimensions may not be easily interpretable, and it may be difficult to understand the relationship between the original features and the reduced dimensions.
* Overfitting: In some cases, dimensionality reduction may lead to overfitting, especially when the number of components is chosen based on the training data.
* Sensitivity to outliers: Some dimensionality reduction techniques are sensitive to outliers, which can result in a biased representation of the data.
* Computational complexity: Some dimensionality reduction techniques, such as manifold learning, can be computationally intensive, especially when dealing with large datasets.

### Important points:

* Dimensionality reduction is the process of reducing the number of features in a dataset while retaining as much information as possible.  
  This can be done to reduce the complexity of a model, improve the performance of a learning algorithm, or make it easier to visualize the data.
* Techniques for dimensionality reduction include: principal component analysis (PCA), singular value decomposition (SVD), and linear discriminant analysis (LDA).
* Each technique projects the data onto a lower-dimensional space while preserving important information.
* Dimensionality reduction is performed during pre-processing stage before building a model to improve the performance
* It is important to note that dimensionality reduction can also discard useful information, so care must be taken when applying these techniques.

9.

i. What is the IQR? What criteria are used to assess it?

# Interquartile Range to Detect Outliers in Data

An observation that differs from an overall pattern on a sample dataset is called an outlier.

### ****Outliers****

The outliers may suggest experimental errors, variability in a measurement, or an anomaly. The age of a person may wrongly be recorded as 200 rather than 20 Years. Such an outlier should definitely be discarded from the dataset. However, not all outliers are bad. Some outliers signify that data is significantly different from others. For example, it may indicate an anomaly like bank fraud or a rare disease.

### ****Significance of outliers:****

* Outliers badly affect the mean and standard deviation of the dataset. These may statistically give erroneous results.
* Most machine learning algorithms do not work well in the presence of outliers. So it is desirable to detect and remove outliers.
* Outliers are highly useful in anomaly detection like fraud detection where the fraud transactions are very different from normal transactions.

### ****What is Interquartile Range IQR?****

IQR is used to **measure variability** by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts.

* Q1 represents the 25th percentile of the data.
* Q2 represents the 50th percentile of the data.
* Q3 represents the 75th percentile of the data.

If a dataset has 2n or 2n+1 data points, then  
Q2 = median of the dataset.  
Q1 = median of n smallest data points.  
Q3 = median of n highest data points.

IQR is the range between the first and the third quartiles namely Q1 and Q3: IQR = Q3 – Q1.

The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers.

ii. Describe the various components of a box plot in detail? When will the lower whisker surpass the upper whisker in length? How can box plots be used to identify outliers?

# Box Plot

Box Plot is a graphical method to visualize data distribution for gaining insights and making informed decisions. Box plot is a type of chart that depicts a group of numerical data through their quartiles.

In this article, we are going to discuss **components of a box plot, how to create a box plot, uses of a Box Plot, and how to compare box plots.**

**Table of Content**

* [What is a Box Plot?](https://www.geeksforgeeks.org/box-plot/#what-is-a-box-plot)
* [How to create a box plots?](https://www.geeksforgeeks.org/box-plot/#how-to-create-a-box-plots)
* [Uses of a Box Plot](https://www.geeksforgeeks.org/box-plot/#uses-of-a-box-plot)
* [How to compare box plots?](https://www.geeksforgeeks.org/box-plot/#how-to-compare-box-plots)

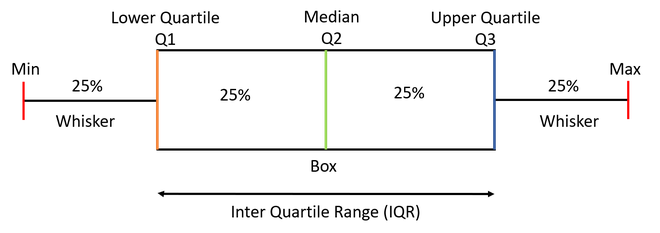
## What is a Box Plot?

The idea of box plot was presented by John Tukey in 1970. He wrote about it in his book “Exploratory Data Analysis” in 1977. Box plot is also known as a whisker plot, box-and-whisker plot, or simply a box-and whisker diagram. Box plot is a graphical representation of the distribution of a dataset. It displays key summary statistics such as the [median](https://www.geeksforgeeks.org/median/), [quartiles,](https://www.geeksforgeeks.org/quartile-formula/) and potential [outliers](https://www.geeksforgeeks.org/machine-learning-outlier/)in a concise and visual manner. By using Box plot you can provide a summary of the distribution, identify potential and compare different datasets in a compact and visual manner.

### ****Elements of Box Plot****

A box plot gives a five-number summary of a set of data which is-

* **Minimum** – It is the minimum value in the dataset excluding the outliers.
* **First Quartile (Q1)** – 25% of the data lies below the First (lower) Quartile.
* **Median (Q2)**– It is the mid-point of the dataset. Half of the values lie below it and half above.
* **Third Quartile (Q3)**– 75% of the data lies below the Third (Upper) Quartile.
* **Maximum**– It is the maximum value in the dataset excluding the outliers.



***Note:****The box plot shown in the above diagram is a perfect plot with no skewness. The plots can have skewness and the median might not be at the center of the box.*

The area inside the box (50% of the data) is known as the [**Inter Quartile Range**](https://www.geeksforgeeks.org/interquartile-range-and-quartile-deviation-using-numpy-and-scipy/)**.**The **IQR**is calculated as –

IQR = Q3-Q1

**Outlies**are the data points **below and above** the**lower and upper limit**. The lower and upper limit is calculated as –

Lower Limit = Q1 - 1.5\*IQR

Upper Limit = Q3 + 1.5\*IQR

The values below and above these limits are considered outliers and the minimum and maximum values are calculated from the points which lie under the lower and upper limit.

## ****How to create a box plots?****

Let us take a sample data to understand how to create a box plot.

Here are the runs scored by a cricket team in a league of 12 matches –***100, 120, 110, 150, 110, 140, 130, 170, 120, 220, 140, 110.***

To draw a box plot for the given data first we need to arrange the data in ascending order and then find the minimum, first quartile, median, third quartile and the maximum.

**Ascending Order**

100, 110, 110, 110, 120, 120, 130, 140, 140, 150, 170, 220

**Median (Q2)** = (120+130)/2 = **125;** Since there were even values

To find the First Quartile we take the first six values and find their median.

**Q1** = (110+110)/2 = **110**

For the Third Quartile, we take the next six and find their median.

**Q3** = (140+150)/2 = **145**

**Note:**If the total number of values is odd then we exclude the Median while calculating Q1 and Q3. Here since there were two central values we included them. Now, we need to calculate the Inter Quartile Range.

**IQR** = Q3-Q1 = 145-110 = **35**

We can now calculate the Upper and Lower Limits to find the minimum and maximum values and also the outliers if any.

**Lower Limit** = Q1-1.5\*IQR = 110-1.5\*35 = **57.5**

**Upper Limit** = Q3+1.5\*IQR = 145+1.5\*35 = **197.5**

So, the minimum and maximum between the range [57.5,197.5] for our given data are –

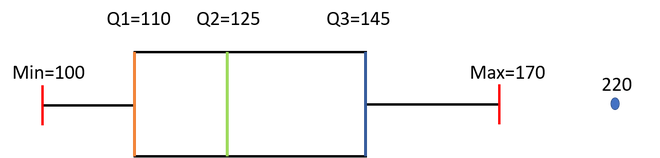
**Minimum = 100**

**Maximum = 170**

The outliers which are outside this range are –

**Outliers = 220**

Now we have all the information, so we can draw the box plot which is as below-



We can see from the diagram that the Median is not exactly at the center of the box and one whisker is longer than the other. We also have one Outlier.

## ****Use-Cases of Box Plot****

* Box plots provide a visual summary of the data with which we can quickly identify the average value of the data, how dispersed the data is, whether the data is skewed or not (skewness).
* The Median gives you the average value of the data.
* Box Plots shows Skewness of the data-

**a)** If the Median is at the **center** of the Box and the **whiskers** are almost the

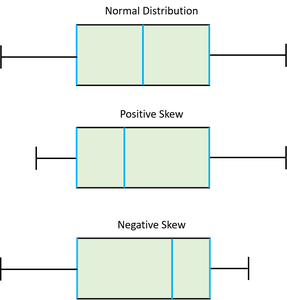
**same on both the ends** then the data is **Normally Distributed**.

**b)** If the Median lies **closer to the First Quartile** and if the **whisker at the lower**

**end is shorter** (as in the above example) then it has a **Positive Skew (Right Skew)**.

**c)** If the Median lies **closer to the Third Quartile** and if the **whisker at the**

**upper end is shorter** than it has a **Negative Skew (Left Skew).**

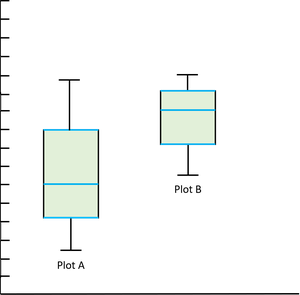


* The dispersion or spread of data can be visualized by the minimum and maximum values which are found at the end of the whiskers.
* The Box plot gives us the idea of about the Outliers which are the points which are numerically distant from the rest of the data.

## ****How to compare box plots?****

As we have discussed at the beginning of the article that box plots make comparing characteristics of data between categories very easy. Let us have a look at how we can compare different box plots and derive statistical conclusions from them.

Let us take the below two plots as an example: –



* **Compare the Medians —**If the median line of a box plot lies outside the box of the other box plot with which it is being compared, then we can say that there is likely to be a difference between the two groups. Here the Median line of the plot B lies outside the box of Plot A.
* **Compare the Dispersion or Spread of data —**The Inter Quartile range (length of the box) gives us an idea about how dispersed the data is. Here Plot A has a longer length than Plot B which means that the dispersion of data is more in plot A as compared to plot B. The length of whiskers also gives an idea of the overall spread of data. The extreme values (minimum &maximum) give the range of data distribution. Larger the range more scattered the data. Here Plot A has a larger range than Plot B.
* **Comparing Outliers —**The outliers give the idea of unusual data values which are distant from the rest of the data. More number of Outliers means the prediction will be more uncertain. We can be more confident while predicting the values for a plot which has less or no outliers.
* **Compare Skewness —**[Skewness](https://www.geeksforgeeks.org/skewness-measures-and-interpretation/) gives us the direction and the magnitude of the lack of symmetry. We have discussed above how to identify skewness. Here Plot A is Positive or Right Skewed and Plot B is Negative or Left Skewed.

10. Make brief notes on any two of the following:

1. Data collected at regular intervals

[Home](https://www.questionpro.com/blog/)  [Market Research](https://www.questionpro.com/blog/category/market-research/)

# Interval Data: Definition, Characteristics and Examples

## Interval Data: Definition?

Interval data, also called an integer, is defined as a data type which is measured along a scale, in which each point is placed at equal distance from one another. Interval data always appears in the form of numbers or numerical values where the distance between the two points is standardized and equal.

Interval data cannot be multiplied or divided, however, it can be added or subtracted. Interval data is measured on an [interval scale](https://www.questionpro.com/blog/interval-scale/). A simple example of interval data: The difference between 100 degrees Fahrenheit and 90 degrees Fahrenheit is the same as 60 degrees Fahrenheit and 70 degrees Fahrenheit.

In [market research](https://www.questionpro.com/blog/what-is-market-research/) or in any other forms of social, economic or business research interval data plays a pivotal role. What makes interval data so popular and in-demand is because interval data supports almost all statistical test and transformations in obtaining [quantitative data](https://www.questionpro.com/blog/quantitative-data/).

**LEARN ABOUT:**[Test Market Demand](https://www.questionpro.com/blog/how-to-test-market-demand-for-a-new-service/)

Interval data has very distinctive attributes that make it distinct in comparison to [nominal data](https://www.questionpro.com/blog/nominal-data/), ordinal data or even ratio data. Interval data doesn’t have a defined absolute zero point which is present in ratio data. The lack of absolute point zero makes comparisons of direct magnitudes impossible. For example, Object A is twice as large as Object B is not a possibility in interval data.

Learn more: [Variable Measurement Scales- Nominal, Ordinal, Interval and Ratio.](https://www.questionpro.com/blog/nominal-ordinal-interval-ratio/)

### Interval Data Analysis

Since interval data is [quantitative analysis](https://www.questionpro.com/blog/quantitative-analysis/) data type almost all the methods used to analyze quantitative can be used. Here are a few examples:

**1. Trend analysis**

[Trend analysis](https://www.questionpro.com/features/trend-analysis.html) is a popular interval data analysis technique, used to draw trends and insights by capturing survey data over a certain period of time. In other words, a trend analysis on interval data is conducted by capturing data using an interval scale survey in multiple iterations, using the same question.

**2. SWOT Analysis**

Analysis conducted to evaluate an organization’s strengths, weaknesses, opportunities, and threats is called [SWOT analysis](https://www.questionpro.com/blog/swot-analysis-example/) and is widely used to evaluate interval data. Strengths and weaknesses are internal aspects of an organization while opportunities and threat are external to an organization. An organization can measure interval data to evaluate market competition as well as plan future marketing activities using the SWOT analysis results.

***LEARN ABOUT:*** [*Level of Analysis*](https://www.questionpro.com/blog/level-of-analysis/)

**3. Conjoint Analysis**

[Conjoint Analysis](https://www.questionpro.com/blog/what-is-conjoint-analysis/) is an advanced level market research technique usually implemented to analyze how individuals make complicated decisions in an interval scale. Which factors are important for customers before they make decisions where they have multiple options available at their disposal.

**4. TURF Analysis**

[TURF analysis](https://www.questionpro.com/article/turf-analysis.html) stands for Totally Unduplicated Reach and Frequency analysis- is a method that allows a marketer to analyze the potential of market research for a combination of products and services. It evaluates the interval data of customers reached by a particular source of communication and its frequency. This analysis technique is used by researchers to understand whether a new product or service will be well-received in the target market or not. This [unit of analysis](https://www.questionpro.com/blog/unit-of-analysis/) method was primarily used for designing media campaigns but has expanded to being used in product distribution and line analysis.

**Collect and Analyze Interval Data with Surveys**

[Get Your Free Account Now](https://www.questionpro.com/a/showEntry.do)

### Key Characteristics of Interval Data

Here are a few characteristics of Interval data:

* **Measurement:**Interval data is measured using an interval scale, which not only shows the order and direction but also shows the exact difference in the value. For example, the markings on a thermometer or a ruler are equidistant, in simpler words they measure the same distance between the two markings.
* **Interval Difference:**The distances between each value on interval data is equal. For example, the difference between 10 cm and 20 cms is the same as 20 cms and 30 cms.
* **Calculation:** In interval data, one can add or subtract values but cannot divide or multiply. Almost all [statistical analysis](https://www.questionpro.com/blog/statistical-analysis-methods/) are applicable when calculating interval data, mean, mode, median etc.
* **Point Zero:** Absolute zero point is arbitrary, which means a variable can be measured even if it has a negative value like temperature can be -10 below zero but height cannot be below zero.

### Interval Data Examples

1. One can measure time during the day using a 12-hour clock, this is a good example of interval data. Time in a 12-hour format is a rotational measure that keeps restarting from zero at set periodicity. These numbers are on an interval scale as the distance between them is measurable and comparable. For example, the difference between 5 minutes and 10 minutes is the same as 15 minutes and 20 minutes in a 12-hour clock.

2. The temperature measured in Fahrenheit and Celsius but not in Kelvin. If you measure temperature in Fahrenheit and Celsius then it will be considered interval data as 0 is arbitrary. But in Kelvin, 0 is absolute. There can’t be a temperature below zero degrees in Kelvin.

3. When you calculate intelligence score in an IQ test. There is no zero point for IQ. According to psychological studies, a person cannot have zero intelligence, therefore in this example, zero is arbitrary. IQ is numeric data expressed in intervals using a fixed measurement scale.

4. Test scores of examination like SAT. Scores in SAT test are in the range of 200-800. The numbers from 0 to 200 are not used when they scale the raw score (number of questions answered correctly) to the section score. The reference point is not an absolute zero, thus, it qualifies to become interval data.

5. Age is also a variable that can be measured on an interval scale. For example if A is 15 years old and B is 20 years old, it not only clear than B is older than A, but B is elder to A by 5 years.

**Interval data** is one of the most used data types. Survey tools offer several ways to capture interval data. When a [survey](https://www.questionpro.com/blog/surveys/) is deployed to a respondent, with a certain [demographic question](https://www.questionpro.com/blog/demographic-survey-questions/) that asks respondents to state their income, these figures can range from zero to infinity!

For example:

**Please state your annual income**

* Below $40,000
* $40,000- $60,000
* $60,000- $80,000
* $80,000- $100,000
* Above $100,000

Numerical data collected in this manner can be can be categorized into groups, in the above mentioned examples groups can based on the respondents annual income. People falling under same income category.

2. The gap between the quartiles

# Quartile Formula

Quartiles are the set of values that divide the data points into four identical values using three individual data points. Thus, a quartile is a very important topic in Statistics that helps us to study large amounts of data, they are used to divide the large data values into four equal quarters. These quartiles show the data that is near the middle points of the large data set.

In this article, we will learn about the quartiles as well as the formulas for the first quartile, second quartile, and third quartile and also provide a step-by-step guide to help you easily calculate quartiles. So, let’s start with the definition of quartile first.

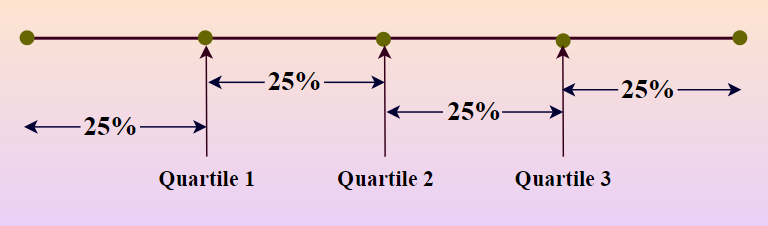
## Quartiles Definition

Quartiles are the values from the dataset which divide the dataset into four equal parts where each part of the dataset contains an equal number of observations. There are three quartiles such as,

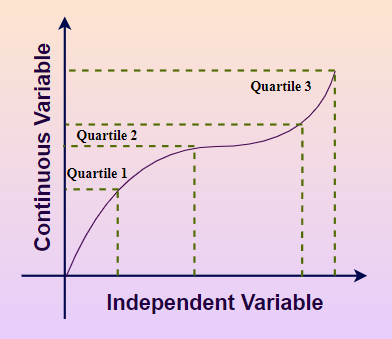
* First or Lower Quartile
* Second Quartile or [Median](https://www.geeksforgeeks.org/mean-median-mode/#:~:text=)
* Third or Upper Quartile

## What is the Quartile Formula?

As mentioned above Quartile divides the data into 4 equal parts. This can be represented visually by the below figure.



* Quartile 1 lies between the starting term and the middle term.
* Quartile 2 lies between the starting terms and the last terms i.e., the Middle term.
* Quartile 3 lies between quartile 2 and the last term.



There is a separate formula for finding each quartile value. And in order to find these quartile values first, sort the given number series data into ascending order.

The steps to obtain the quartile formula are as shown below as follows:

* **Step 1:** Sort the given data in ascending order.
* **Step 2:**Find respective quartile values/terms as per need from the below formulae.
* ***First Quartile = ({n + 1}/{4})th term***
* ***Second Quartile = ({n + 1}/{2})th term***
* ***Third Quartile = ({3(n + 1)}/{4})th term***

*Where****n****is the total count of numbers in the given data.*

## Quartiles in Statistics

We know that the Median divides the data into two equal parts, in the same way, the quartile divides the data into four parts. Similar to the median which divides the data into half so that 50% of the estimation lies below the median and 50% lies above it, the quartile splits the data into i.e.,

* **First Part of Data:** From smallest to largest of numbers 25% of the value, comes under this part and also this part lies below the first quartile.
* **Second of Data:** Value between 25% and 50% of the data comes under this part and this part lies between the first and second quartile (Median).
* **Third of Data:** Value between 50% and 75% of the data comes under this part and this part lies between the second and third quartile.
* **Fourth of Data:** Greatest 25% of all values in the data comes under the fourth part and this part lies above the fourth quartile.

## Generalized Formula for Quartile

The generalized formula for the quartile is,

*Quartile𝑟=𝑙1+(𝑖⋅𝑛4–𝑐𝑓)⋅(𝑙2−𝑙1)𝑓Quartile****r****​=****l1****​+(****i****⋅****4n****​****–cf****​)⋅****f****(****l2****​−****l1****​)​*

*Where,*

* ***Quartiler****indicates****rth****quartile.*
* ***l1, l2****are lower and upper limit value that contains ith quartile,*
* ***f****is the frequency count.*
* ***cf****is the cumulative frequency of class preceding the quartile class.*

Using this generalized formula, the first and third quartiles can be calculated as:

* *Q1=𝑙1+(𝑛4–𝑐𝑓)⋅(𝑙2−𝑙1)𝑓Q****1****​=****l1****​+(****4n****​****–cf****​)⋅****f****(****l2****​−****l1****​)​*
* *Q3=𝑙1+(3𝑛4–𝑐𝑓)⋅(𝑙2−𝑙1)𝑓Q****3****​=****l1****​+(****43n****​****–cf****​)⋅****f****(****l2****​−****l1****​)​*

### ****Interquartile Range****

Interquartile Range is the distance between the first quartile and the third quartile. It is also known as a mid-spread. It helps us to calculate variation for the data which is divided into quartiles. The formula for calculating the Interquartile range is given by,

***Interquartile Range (IQR) = Q3 – Q1***

*Where,*

* ***Q3****is third/upper quartile, and*
* ***Q1****is first/lower quartile.*

### ****Quartile Deviation****

Quartile Deviation is defined as half of the distance between the first quartile and the third quartile. It is also known as Semi Interquartile Range. The formula for quartile deviation is given by,

***Quartile Deviation =******(Q3 – Q2)/2***

## Quartile vs Percentile

The key differences between Quartile and Percentile are given as follows:

| **Aspect** | **Quartile** | **Percentile** |
| --- | --- | --- |
| **Definition** | A quartile is a type of quantile that divides a data set into four equal parts | A percentile is a type of quantile that divides a data set into 100 equal parts |
| **Range** | Quartiles divide a dataset into four parts:  Q1 = 25th Percentile  Q2 = 50th Percentile or Median  Q3 = 75th Percentile | Percentiles divide a dataset into 100 parts, with each percentile representing 1% of the data. |
| **Calculation** | Quartiles are calculated by dividing the data set into four equal parts, with each part containing 25% of the data | Percentiles are calculated by dividing the data set into 100 equal parts, with each part containing 1% of the data |
| **Represented by** | Quartiles are often represented as Q, Q2, and Q3. | Percentiles are often represented as P1, P2, P3, and so on up to P99 |
| **Usefulness** | Quartiles are useful for identifying the spread and distribution of data, particularly in box plots and histograms | Percentiles are useful for comparing an individual the data point to the rest of the data set, and for identifying extreme values or outliers |

3. Use a cross-tab

## What is cross tabulation?

Cross tabulation (crosstab) is a useful analysis tool commonly used to compare the results for one or more variables with the results of another variable. It is used with data on a nominal scale, where variables are named or labeled with no specific order.

[Crosstabs](https://www.surveymonkey.com/curiosity/using-cross-tabulation-to-understand-respondents/) are basically data tables that present the results from a full group of survey respondents as well as subgroups. They allow you to examine relationships within the data that might not be obvious when simply looking at total survey responses.

### Benefits of cross tabulation

With cross tabulation, you can examine your data in a variety of ways to achieve a [deeper understanding of groups](https://www.surveymonkey.com/mp/tour/crosstabfilter/) within your respondents.

#### Reduce confusion when analyzing data

Analyzing large datasets can be difficult. Finding relevant, actionable insights within large amounts of data is even more complicated. Crosstabs simplify and divide data into subgroups for ease of interpretation—they show percentages and frequencies that may change when contrasted with variables in other categories. By making datasets more manageable at scale, fewer errors will result.

#### More granular data points

Using crosstabs, you can examine the relationships between one or more variables, which leads to insights on a more granular level. These insights could go unnoticed without crosstabs, lost in a sea of data, or require additional work to reveal. Use multiple filters to dig even deeper into data to uncover more details.

#### Actionable insights

Using crosstabs simplifies datasets so that you can make quick comparisons between them. This means faster insights for creating new marketing strategies guided by the data. You are also able to watch for global trends across survey responses and take action accordingly.

#### Clarity of interpretation

When you use crosstabs, datasets are simplified and divided into subgroups. The resulting clean data is in a more digestible format and easily viewed and used by research professionals and team members without analytics training.

## When to use cross-tabulation

Cross tabulation is typically used when you have information that can be divided into mutually exclusive groups, also known as categorical variables. It allows you to examine relationships within the data that may not be readily apparent. A [crosstab report](https://www.surveymonkey.com/curiosity/cross-tab-survey-analysis/) can show the connection between two or more [survey questions](https://www.surveymonkey.com/mp/survey-question-types/) from the study in market research studies.

### Who uses cross tabulation?

Crosstabs are used across multiple industries and job functions. Examples of departments that benefit from crosstab analysis include:

**HR departments**can use cross tabulation to examine employee survey data about company culture, managerial guidance, employee engagement, and more. Using crosstabs will assist HR with determining departments that have particular problems or needs that they can address.

**Market research teams**can take raw data and make it more digestible for management decision-making with crosstabs.

**Customer support teams**can use crosstabs to evaluate customer satisfaction levels between long-term and new customers.

**School administration**can use instructor evaluation data from students and cross-tabulate it with the class subject, time of class, and other data to help improve the educational experience for students.

## Cross tabulation and Chi-square

A Chi-square test is used to test data in a cross-tabulation table to determine whether or not the data is statistically significant. The results are statistically significant if the two categorical variables are independent (unrelated). Basically, the Chi-square test is a correlation test for categorical variables.

## Crosstab vocabulary

Here are some vocabulary words that can be used when it comes to crosstabs:

* **Banners (Cuts):**the headers at the top of each column that name the categories of the data
* **Categories:**how variables are grouped
* **Chi-square test:**analysis to determine the statistical significance of cross-tabulation by determining if the compared variables are independent. A measure of how actual data compares to expected data
* **Columns:**cells that display data vertically
* **Column-percentage:**view of data that calculates the column data belonging to a particular row
* **Count (Frequency):** total number of responses that fall into a row or column
* **Crosstabs, cross tabulation or contingency table:**table used to analyze categorical data
* **CSV:**comma-separated values files are used to export complex data from one application and import into another, such as a notes file or spreadsheet
* **Filters:** used to focus your crosstab on a particular view of the data
* **Percentage:**percentage of responses that fall into rows or columns
* **Pivot table:** If you use Excel, pivot tables can be used to create crosstabs
* **Rows:**cells displaying data horizontally
* **Row-percentage:**data view that calculates the row data that belongs to a particular column
* **Statistically significant:** determination that results of data are not explainable by chance
* **Stubs:**headers that name the categories of data displayed by rows

1. Make a comparison between:

1. Data with nominal and ordinal values

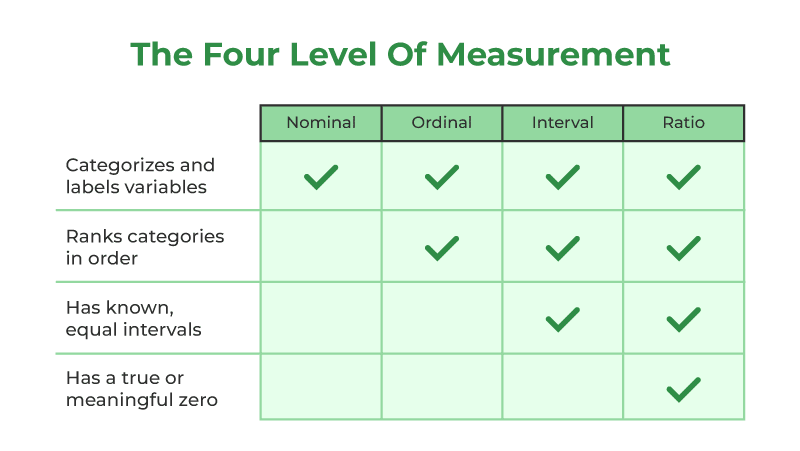
# Nominal vs Ordinal Data

Data science revolves around the processing and analysis of data utilizing a range of tools and techniques. In today’s data-driven world, we come across types of data each requiring handling and interpretation. It is important to understand different types of data for proper data analysis and statistical interpretation. The type of data determines the proper statistical methods and operations that should be used. Various data types need different analysis and interpretation methods to draw significant conclusions. In this article we will explore the concept of data, and its significance provide real-world examples, and guide you through ways to work with it.

## Levels of Measurement

Before analyzing a dataset, it is crucial to identify the type of data it contains. Luckily, all data can be grouped into one of four categories: nominal, ordinal, interval, or ratio data. Although these are often referred to as “data types,” they are actually different levels of measurement. The level of measurement reflects the accuracy with which a variable has been quantified, and it determines the methods that can be used to extract insights from the data.

The four categories of data are not always straightforward to distinguish and instead belong to a hierarchy, with each level building on the preceding one.

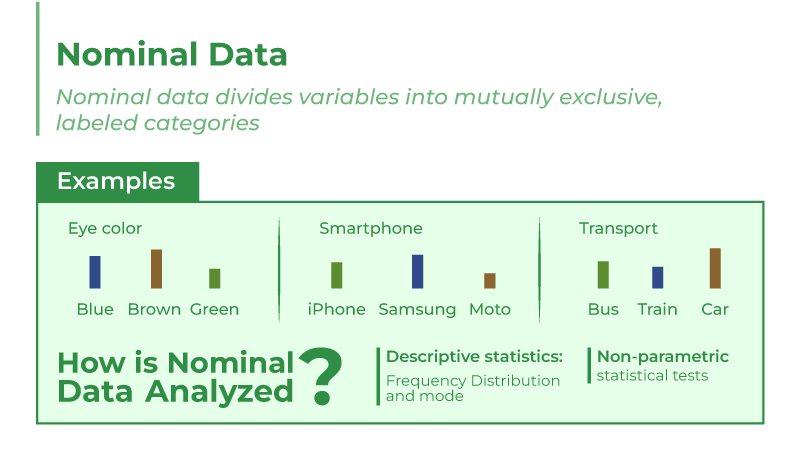


There are four types of data: categorical, which can be further divided into nominal and ordinal, and numerical, which can be further divided into interval and ratio. The nominal and ordinal scales are relatively imprecise, which makes them easier to analyze, but they offer less accurate insights. On the other hand, the interval and ratio scales are more complex and difficult to analyze, but they have the potential to provide much richer insights.

* **Nominal Data** – Nominal data is a basic data type that categorizes data by labeling or naming values such as Gender, hair color, or types of animal. It does not have any hierarchy.
* **Ordinal Data** – Ordinal data involves classifying data based on rank, such as social status in categories like ‘wealthy’, ‘middle income’, or ‘poor’. However, there are no set intervals between these categories.
* **Interval Data**– Interval data is a way of organizing and comparing data that includes measured intervals. Temperature scales, like Celsius or Fahrenheit, are good examples of interval data. However, interval data doesn’t have a true zero, meaning that a measurement of “zero” can still represent a quantifiable measure (like zero degrees Celsius, which is just another point on the scale and doesn’t actually mean there is no temperature present).
* **Ratio Data** – The most intricate level of measurement is ratio data. Similar to interval data, it categorizes and arranges data, utilizing measured intervals. But, unlike interval data, ratio data includes a genuine zero. When a variable is zero, there is no presence of that variable. A prime illustration of ratio data is height measurement, which cannot be negative.

## What is Nominal Data?

Categorical data, also known as nominal data, is a crucial type of information utilized in diverse fields such as research, statistics, and data analysis. It comprises of categories or labels that help in classifying and arranging data. The essential feature of categorical data is that it does not possess any inherent order or ranking among its categories. Instead, these categories are separate, distinct, and mutually exclusive.



For example, Nominal data is used to classify information into distinct labels or categories without any natural order or ranking. These labels or categories are represented using names or terms, and there is no natural order or ranking among them. Nominal data is useful for qualitative classification and organization of information, enabling researchers and analysts to group data points based on specific attributes or characteristics without implying any numerical relationships.

* Eye color categories like “blue” or “green” represent nominal data. Each category is distinct, with no order or ranking.
* Smartphone brands like “iPhone” or “Samsung” are nominal data. There’s no hierarchy among brands.
* Transportation modes like “car” or “bicycle” are nominal data. They are discrete categories without inherent order.

### Characteristics of Nominal Data

* Data that is classified as nominal is comprised of categories that are completely separate and distinct from one another.
* Data that falls under the nominal category is distinguished by descriptive labels rather than any numeric or quantitative value
* Nominal data cannot be ranked or ordered hierarchically, as no category is superior or inferior to another.

### Example

Here are a few examples of how nominal data is used to classify and categorize information into distinct and non-ordered categories:

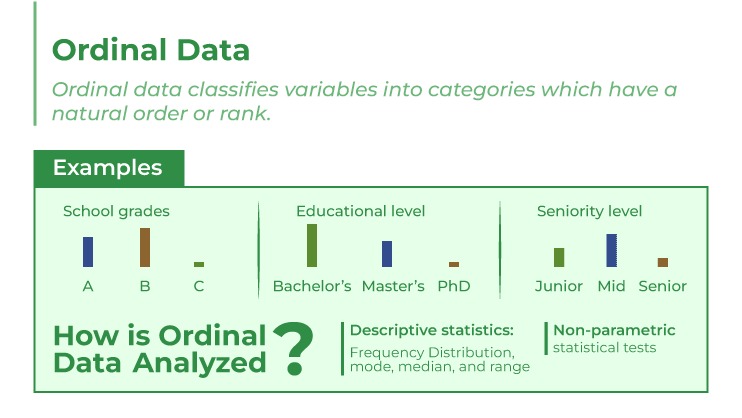
1. **Colors of Car:**Car colors are nominal data, with clear categories but no inherent order or ranking. Each car falls under one color category, without any logical or numerical connection between colors.

2. **Types of Fruits:** Fruit categories in a basket are nominal. Each fruit belongs to a specific category with no hierarchy or order. All categories are distinct and discrete.

3. **Movie Genres:** Movie genres are nominal data since there’s no ranking among categories like “action” or “comedy.” Each genre is unique, but we can’t say if one is better than another based on this data alone.

## What is Ordinal Data?

Ordinal data is a form of qualitative data that classifies variables into descriptive categories. It is characterized by the fact that the categories it employs are ranked on some sort of hierarchical scale, such as from high to low. Ordinal data is the second most complicated type of measurement, following nominal data. Although it is more intricate than nominal data, which lacks any inherent order, it is still relatively simplistic.



For example, Ordinal data is a type of data used to categorize items with a meaningful hierarchy or order. These categories help us to compare and rank different achievements, positions, or performance of students, even if the intervals between them are not equal. Ordinal data is useful for understanding ordered choices or preferences and for assessing relative differences.

* School Grades: Grades like A, B, C are ordinal data, ranked by achievement, but intervals between them vary.
* Education Level: Levels like high school, bachelor’s, master’s are ordinal data, ordered by education, but gaps between levels differ.
* Seniority Level: Job levels like entry, mid, senior are ordinal data, indicating hierarchy, but the gap varies by job and industry.

### Characteristics of Ordinal Data

* Ordinal data falls under the category of non-numeric and categorical data, but it can still make use of numerical values as labels.
* Ordinal data are always ranked in a hierarchy (hence the name ‘ordinal’).
* Ordinal data may be ranked, but their values are not evenly distributed.
* With ordinal data, you can calculate frequency distribution, mode, median, and range of variables.

### Example

Here are a few examples of how ordinal data is used in fields and domains:

**1. Educational Levels:** Ordinal data is commonly used to represent education levels, such, as ” school,” “bachelors degree,” “masters degree,” and “Ph.D.” These levels have an order.

**2. Customer Satisfaction Ratings:**Another application of data is in customer satisfaction surveys. These surveys often ask respondents to rate their experience on a scale, from “poor” to “excellent.”

**3. Economic Classes:**classes including ” class ” “middle class,” and “upper class ” can be classified as ordinal data based on their ranking.

These examples demonstrate the ways in which ordinal data is utilized across fields and domains.

## Nominal Vs Ordinary Data

| **Characteristics** | **Nominal data** | **Ordinal Data** |
| --- | --- | --- |
| **Nature of Categories** | Distinct and Discrete | Discrete and Distinct |
| **Order/Ranking** | No inherent order | Has a clear order or ranking |
| **Numerical Values** | No meaningful numerical values | No meaningful numerical values |
| **Analysis Techniques** | Frequency counts, percentages, bar charts | Ranking, median, non-parametric tests, ordered bar charts, ordinal regression |
| **Example** | Colors, gender, types of animals | School grades, education level, seniority level |
| **Interpretation** | Used for classification and grouping based on category | Used for assessing ordered preferences, hierarchy, or rankings |

2. Histogram and box plot

## Histograms

### Overview

A histogram is a graphical display of data using bars (also called buckets or bins) of different height, where each bar groups numbers into ranges. Histograms reveal a lot of useful information about numerical data with a single explanatory variable. Histograms are used for getting a sense about the distribution of data, its median, and skewness.

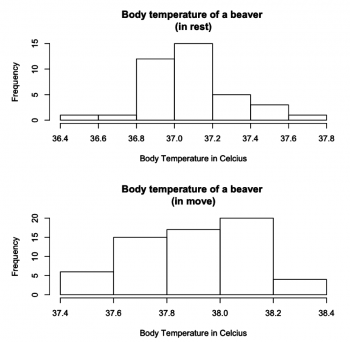
[](https://sustainabilitymethods.org/index.php/File:Beaverbodytemperatures.png)

Fig.3

### Plotting in R

#Fig.3

par(mfrow = c(2,1))

hist(beaver2$temp[1:38], main = "Body temperature of a beaver (in rest)", xlab = "Body Temperature in Celcius", breaks = 5)

hist(beaver2$temp[39:100], main = "Body temperature of a beaver (in move)", xlab = "Body Temperature in Celcius", breaks = 5)

The two histograms are plotted from the “beaver2” dataset and illustrate how a beaver’s body temperature changes when it starts moving. Both histograms resemble the bell-curved shape of normal distribution. We can see a change in the beaver’s body temperature from approximately 37 degrees to 38 degrees.

#### Identifying and interpreting histograms

**Histograms Vs. Bar charts** Histograms are different than bar charts, and one should not confuse them. A histogram does not have gaps between the bars, but a bar chart does. Histograms have the response variable on the X-axis, and the Y-axis shows the frequency (or the probability density). In contrast, the Y-axis in a bar chart shows the frequency and the X-axis shows the response variable, which however represents nominal data.

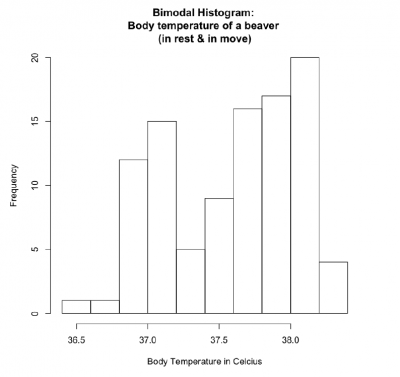
[](https://sustainabilitymethods.org/index.php/File:Bimodalhistrogram.png)

Fig.4

**Patterns** Histograms display well how data is distributed. For instance, a the symmetric, unimodal pattern of a histogram represents a normal distribution. Likewise, skewed right and left patterns in histograms display skewness of data - asymmetry of the distribution of data around the mean. See “[Histogram: Study the Shape](https://www.pqsystems.com/qualityadvisor/DataAnalysisTools/interpretation/histogram_shape.php)” to learn more about histogram patterns.

#Fig.4

hist(beaver2$temp, main = "Bimodal Histogram: Body temperature of a beaver (in rest & in move)", xlab = "Body Temperature in Celcius", breaks = 12)

If the beaver2 dataset plotted into one histogram, it takes bimodal pattern and represents binomial distribution as there are two means of sample points - the temperature of a beaver in rest and in the move.

**Number and width of the bars (bins)** Histograms can become confusing depending on how the bin margin is put. As it is said in The [R Book, p231](https://www.wiley.com/en-us/The+R+Book%2C+2nd+Edition-p-9781118448960) - “Wide bins produce one picture, narrow bins produce a different picture, unequal bins produce confusion.” Choice of number and width of bins techniques can heavily influence a histogram’s appearance, and choice of bandwidth can heavily influence the appearance of a kernel density estimate. Therefore, it is suggested that the bins stay in the same width and that the number of the bins is selected carefully to best display pattern in data.

1. The average and median

The median and the average (or mean) are both measures of central tendency used in statistics to describe the center of a data set, but they are calculated in different ways and can give different insights into the data.

1. Median:  
   - The median is the middle value in a data set when the values are arranged in numerical order.  
   - To find the median, you first order the data from smallest to largest, and then find the middle value. If the data set has an odd number of values, the median is the middle value. If the data set has an even number of values, the median is the average of the two middle values.  
   - The median is not affected by extreme values or outliers in the data set, making it a robust measure of central tendency in the presence of skewed data.
2. Average (Mean):  
   - The average, also known as the mean, is calculated by summing up all the values in a data set and dividing by the total number of values.  
   - The mean is sensitive to extreme values or outliers in the data set because it takes into account the magnitude of all values when calculating the average.  
   - The mean is affected by extreme values, so it may not always be a good representation of the central tendency of a data set, especially if the data is skewed.

In summary, the median is the middle value in a data set, while the average (mean) is the sum of all values divided by the number of values. The median is often preferred when dealing with skewed data or data containing outliers, while the mean is a useful measure of central tendency when the data is normally distributed and there are no extreme values influencing the average.