1. What are the key tasks that machine learning entails? What does data pre-processing imply?
2. Machine Learning is gaining some useful information from the data. Usually, Machine Learning is of two types Supervised Learning and Unsupervised Learning.  
     
   **Supervised Learning**  
       Classification and Regression are examples of Machine Learning. The task of classification is to predict what class an instance of data should fall into. Another task in machine learning is a regression which predicts a numeric value.  Classification deals with predicting discrete value like True/False, Male/Female, 1/2/3 Regression is used when the class to predict is of continuous value say from 0 to 100, -inf to +inf. The best example of regression is a best-fit line drawn through some data points to generalize the data points. This set of problems is called as supervised learning because here we tell the algorithm what to predict i.e. we know the label or target value.  
     
   **Unsupervised Learning**  
       In unsupervised learning, there's no label or target value given for the data. A task where similar items grouped together to form a cluster is known as clustering. Another task of unsupervised learning may be reducing the data from many features to a small number so that it becomes easier to visualize it in two or three dimensions.  
     
   The following table lists some common tasks in machine learning with algorithms used to solve these tasks.

|  |  |
| --- | --- |
| Supervise Learning tasks | |
| k-Nearest Neighbors | Linear |
| Naive Bayes | Locally weighted linear |
| Support Vector Machines | Ridge |
| Decision Trees | Lasso |
| Unsupervised Learning tasks | |
| k-means | Expectation minimization |
| DBSCAN | Parzen window |

**What is data preprocessing?**

Data preprocessing, a component of [data preparation](https://searchbusinessanalytics.techtarget.com/definition/data-preparation), describes any type of processing performed on [raw data](https://searchdatamanagement.techtarget.com/definition/raw-data) to prepare it for another data processing procedure. It has traditionally been an important preliminary step for the [data mining](https://searchbusinessanalytics.techtarget.com/definition/data-mining) process. More recently, data preprocessing techniques have been adapted for training machine learning models and AI models and for running inferences against them.

Data preprocessing transforms the data into a format that is more easily and effectively processed in data mining, machine learning and other data science tasks. The techniques are generally used at the earliest stages of the [machine learning](https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML) and AI development pipeline to ensure accurate results.

There are several different tools and methods used for preprocessing data, including the following:

* sampling, which selects a representative subset from a large population of data;
* transformation, which manipulates raw data to produce a single input;
* denoising, which removes [noise](https://whatis.techtarget.com/definition/noise) from data;
* imputation, which synthesizes statistically relevant data for missing values;
* [normalization](https://searchsqlserver.techtarget.com/definition/normalization), which organizes data for more efficient access; and
* feature extraction, which pulls out a relevant feature subset that is significant in a particular context.

These tools and methods can be used on a variety of data sources, including data stored in files or databases and streaming data.

**Why is data preprocessing important?**

Virtually any type of data analysis, [data science](https://www.techtarget.com/searchenterpriseai/definition/data-science) or AI development requires some type of data preprocessing to provide reliable, precise and robust results for enterprise applications.

Real-world data is messy and is often created, processed and stored by a variety of humans, business processes and applications. As a result, a data set may be missing individual fields, contain manual input errors, or have duplicate data or different names to describe the same thing. Humans can often identify and rectify these problems in the data they use in the line of business, but [data used to train machine learning](https://searchbusinessanalytics.techtarget.com/feature/Data-preparation-in-machine-learning-6-key-steps) or deep learning algorithms needs to be automatically preprocessed.

Machine learning and [deep learning](https://www.techtarget.com/searchenterpriseai/definition/deep-learning-deep-neural-network) algorithms work best when data is presented in a format that highlights the relevant aspects required to solve a problem. Feature engineering practices that involve data wrangling, [data transformation](https://searchdatamanagement.techtarget.com/definition/data-transformation), data reduction, feature selection and feature scaling help restructure raw data into a form suited for particular types of algorithms. This can significantly reduce the processing power and time required to train a new machine learning or AI algorithm or run an inference against it.

One caution that should be observed in preprocessing data: the potential for reencoding bias into the data set. Identifying and correcting bias is critical for applications that help make decisions that affect people, such as loan approvals. Although [data scientists](https://www.techtarget.com/searchenterpriseai/definition/data-scientist) may deliberately ignore variables like gender, race or religion, these traits may be correlated with other variables like zip codes or schools attended, generating biased results.

Most modern data science packages and services now include various preprocessing libraries that help to automate many of these tasks.

**What are the key steps in data preprocessing?**

The steps used in data preprocessing include the following:

**1. Data profiling**. [Data profiling](https://searchdatamanagement.techtarget.com/definition/data-profiling) is the process of examining, analyzing and reviewing data to collect statistics about its quality. It starts with a survey of existing data and its characteristics. Data scientists identify data sets that are pertinent to the problem at hand, inventory its significant attributes, and form a hypothesis of features that might be relevant for the proposed analytics or machine learning task. They also relate data sources to the relevant business concepts and consider which preprocessing libraries could be used.

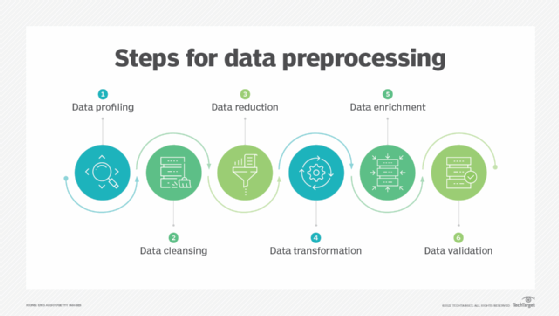
**2. Data cleansing**. The aim here is to find the easiest way to rectify quality issues, such as eliminating bad data, filling in missing data or otherwise ensuring the raw data is suitable for feature engineering.

**3. Data reduction.**Raw data sets often include redundant data that arise from characterizing phenomena in different ways or data that is not relevant to a particular ML, AI or analytics task. Data reduction uses techniques like principal component analysis to transform the raw data into a simpler form suitable for particular use cases.

**4. Data transformation**. Here, data scientists think about how different aspects of the data need to be organized to make the most sense for the goal. This could include things like structuring [unstructured data](https://searchbusinessanalytics.techtarget.com/definition/unstructured-data), combining salient variables when it makes sense or identifying important ranges to focus on.

**5. Data enrichment**. In this step, data scientists apply the various feature engineering libraries to the data to effect the desired transformations. The result should be a data set organized to achieve the optimal balance between the training time for a new model and the required compute.

**6. Data validation**. At this stage, the data is split into two sets. The first set is used to train a machine learning or deep learning model. The second set is the testing data that is used to gauge the accuracy and robustness of the resulting model. This second step helps identify any problems in the [hypothesis](https://whatis.techtarget.com/definition/hypothesis) used in the cleaning and feature engineering of the data. If the data scientists are satisfied with the results, they can push the preprocessing task to a [data engineer](https://searchdatamanagement.techtarget.com/definition/data-engineer) who figures out how to scale it for production. If not, the data scientists can go back and make changes to the way they implemented the data cleansing and feature engineering steps.

Data preprocessing typically includes these steps.

**Data preprocessing techniques**

There are two main categories of preprocessing -- [data cleansing](https://searchdatamanagement.techtarget.com/definition/data-scrubbing) and feature engineering. Each includes a variety of techniques, as detailed below.

**Data cleansing**

Techniques for cleaning up messy data include the following:

**Identify and sort out missing data.**There are a variety of reasons a data set might be missing individual fields of data. Data scientists need to decide whether it is better to discard records with missing fields, ignore them or fill them in with a probable value. For example, in an [IoT](https://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT) application that records temperature, adding in a missing average temperature between the previous and subsequent record might be a safe fix.

**Reduce**[**noisy data**](https://searchbusinessanalytics.techtarget.com/definition/noisy-data)**.** Real-world data is often noisy, which can distort an analytic or AI model. For example, a temperature sensor that consistently reported a temperature of 75 degrees Fahrenheit might erroneously report a temperature as 250 degrees. A variety of statistical approaches can be used to reduce the noise, including binning, regression and clustering.

**Identify and remove duplicates.** When two records seem to repeat, an algorithm needs to determine if the same measurement was recorded twice, or the records represent different events. In some cases, there may be slight differences in a record because one field was recorded incorrectly. In other cases, records that seem to be duplicates might indeed be different, as in a father and son with the same name who are living in the same house but should be represented as separate individuals. Techniques for identifying and removing or joining duplicates can help to automatically address these types of problems.

**Feature engineering**

Feature engineering, as noted, involves techniques used by data scientists to organize the data in ways that make it more efficient to train [data models](https://searchdatamanagement.techtarget.com/definition/data-modeling) and run inferences against them. These techniques include the following:

**Feature scaling or normalization.**Often, multiple variables change over different scales, or one will change linearly while another will change exponentially. For example, salary might be measured in thousands of dollars, while age is represented in double digits. Scaling helps to transform the data in a way that makes it easier for algorithms to tease apart a meaningful relationship between variables.

**Data reduction.**Data scientists often need to combine a variety of [data sources to create a new AI or analytics model](https://searchbusinessanalytics.techtarget.com/feature/6-data-preparation-best-practices-for-analytics-applications). Some of the variables may not be correlated with a given outcome and can be safely discarded. Other variables might be relevant, but only in terms of relationship -- such as the ratio of debt to credit in the case of a model predicting the likelihood of a loan repayment; they may be combined into a single variable. Techniques like principal component analysis play a key role in reducing the number of dimensions in the training data set into a more efficient representation.

**Discretization.**It's often useful to lump raw numbers into discrete intervals. For example, income might be broken into five ranges that are representative of people who typically apply for a given type of loan. This can reduce the overhead of training a model or running inferences against it.

**Feature encoding**. Another aspect of feature engineering involves organizing unstructured data into a structured format. Unstructured data formats can include text, audio and video. For example, the process of developing natural language processing algorithms typically starts by using data transformation algorithms like Word2vec to translate words into numerical vectors. This makes it easy to represent to the algorithm that words like "mail" and "parcel" are similar, while a word like "house" is completely different. Similarly, a [facial recognition](https://www.techtarget.com/searchenterpriseai/definition/facial-recognition) algorithm might reencode raw pixel data into vectors representing the distances between parts of the face.

These issues complicate the process of preparing data for BI and analytics applications.

**How is data preprocessing used?**

Data preprocessing plays a key role in earlier stages of machine learning and AI application development, as noted earlier. In an [AI](https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence) context, data preprocessing is used to improve the way data is cleansed, transformed and structured to improve the accuracy of a new model, while reducing the amount of compute required.

A good data preprocessing pipeline can create reusable components that make it easier to test out various ideas for streamlining business processes or improving customer satisfaction. For example, preprocessing can improve the way data is organized for a recommendation engine by improving the age ranges used for categorizing customers.

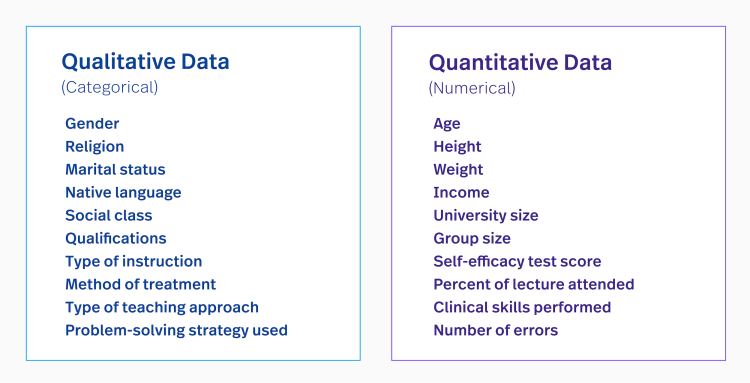
Preprocessing can also simplify the work of creating and modifying data for more accurate and targeted business intelligence insights. For example, customers of different sizes, categories or regions may exhibit different behaviors across regions. Preprocessing the data into the appropriate forms could help BI teams weave these insights into BI dashboards.

In a customer relationship management ([CRM](https://searchcustomerexperience.techtarget.com/definition/CRM-customer-relationship-management)) context, data preprocessing is a component of web mining. Web usage logs may be preprocessed to extract meaningful sets of data called user transactions, which consist of groups of URL references. User sessions may be tracked to identify the user, the websites requested and their order, and the length of time spent on each one. Once these have been pulled out of the raw data, they yield more useful information that can be applied, for example, to consumer research, marketing or [personalization](https://searchcustomerexperience.techtarget.com/definition/personalization).

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

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

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



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

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

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

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

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

**Advantages of quantitative data**

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

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

**Disadvantages of quantitative data**

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

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

**Advantages of qualitative data**

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

**Disadvantages of qualitative data**

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

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

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

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

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

**Quantitative data collection methods**

**Surveys**

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

**Polls**

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

**Experiments**

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

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

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

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

**Qualitative data collection methods**

**Interviews**

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

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

**Focus groups**

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

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

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

**Quantitative data example**

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

**Qualitative data example**

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

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

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

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

**Which type is better for data analysis?**

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

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

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

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

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

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

Common issues in Machine Learning

Although machine learning is being used in every industry and helps organizations make more informed and data-driven choices that are more effective than classical methodologies, it still has so many problems that cannot be ignored. Here are some common issues in Machine Learning that professionals face to inculcate ML skills and create an application from scratch.

* + Inadequate Training Data
* The major issue that comes while using machine learning algorithms is the lack of quality as well as quantity of data. Although data plays a vital role in the processing of machine learning algorithms, many data scientists claim that inadequate data, noisy data, and unclean data are extremely exhausting the machine learning algorithms. For example, a simple task requires thousands of sample data, and an advanced task such as speech or image recognition needs millions of sample data examples. Further, data quality is also important for the algorithms to work ideally, but the absence of data quality is also found in Machine Learning applications. Data quality can be affected by some factors as follows:
* **Noisy Data-** It is responsible for an inaccurate prediction that affects the decision as well as accuracy in classification tasks.
* **Incorrect data-** It is also responsible for faulty programming and results obtained in machine learning models. Hence, incorrect data may affect the accuracy of the results also.
* **Generalizing of output data-** Sometimes, it is also found that generalizing output data becomes complex, which results in comparatively poor future actions.
  + Poor quality of data
* As we have discussed above, data plays a significant role in machine learning, and it must be of good quality as well. Noisy data, incomplete data, inaccurate data, and unclean data lead to less accuracy in classification and low-quality results. Hence, data quality can also be considered as a major common problem while processing machine learning algorithms.
  + Non-representative training data
* To make sure our training model is generalized well or not, we have to ensure that sample training data must be representative of new cases that we need to generalize. The training data must cover all cases that are already occurred as well as occurring.
* Further, if we are using non-representative training data in the model, it results in less accurate predictions. A machine learning model is said to be ideal if it predicts well for generalized cases and provides accurate decisions. If there is less training data, then there will be a sampling noise in the model, called the non-representative training set. It won't be accurate in predictions. To overcome this, it will be biased against one class or a group.
* Hence, we should use representative data in training to protect against being biased and make accurate predictions without any drift.

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

**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 –

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

* Deleting Rows with missing values
* Impute missing values for continuous variable
* Impute missing values for categorical variable
* Other Imputation Methods
* Using Algorithms that support missing values
* Prediction of missing values
* Imputation using Deep Learning Library — Datawig

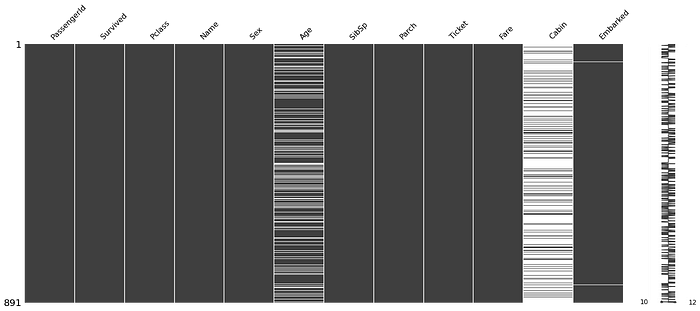
1. Describe the various methods for dealing with missing data values in depth.

This article covers 7 ways to handle missing values in the dataset:

1. Deleting Rows with missing values
2. Impute missing values for continuous variable
3. Impute missing values for categorical variable
4. Other Imputation Methods
5. Using Algorithms that support missing values
6. Prediction of missing values
7. Imputation using Deep Learning Library — Datawig

*Data used is*[*Titanic Dataset*](https://www.kaggle.com/c/titanic)*from Kaggle*

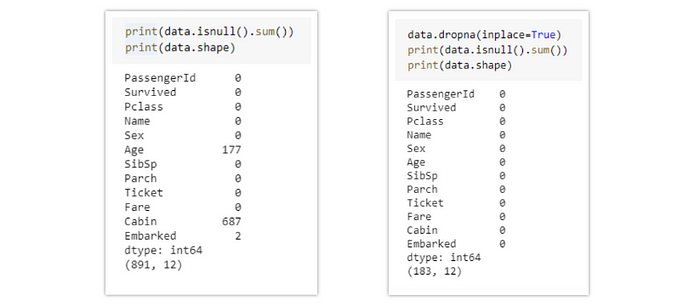
data = pd.read\_csv("train.csv")  
msno.matrix(data)



(Image by Author), Visualization of Missing Values: white lines denote the presence of missing value

**Delete Rows with Missing Values:**

Missing values can be handled by deleting the rows or columns having null values. If columns have more than half of the rows as null then the entire column can be dropped. The rows which are having one or more columns values as null can also be dropped.



(Image by Author) Left: Data with Null values, Right: Data after removal of Null values

**Pros:**

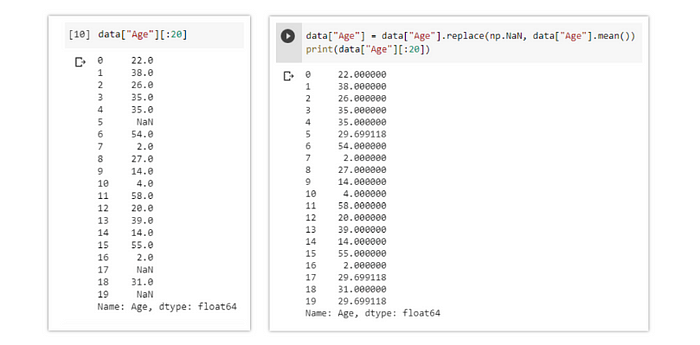
* A model trained with the removal of all missing values creates a robust model.

**Cons:**

* Loss of a lot of information.
* Works poorly if the percentage of missing values is excessive in comparison to the complete dataset.

**Impute missing values with Mean/Median:**

Columns in the dataset which are having numeric continuous values can be replaced with the mean, median, or mode of remaining values in the column. This method can prevent the loss of data compared to the earlier method. Replacing the above two approximations (mean, median) is a statistical approach to handle the missing values.



(Image by Author) Left: Age column before Imputation, Right: Age column after imputation by the mean value

The missing values are replaced by the mean value in the above example, in the same way, it can be replaced by the median value.

**Pros:**

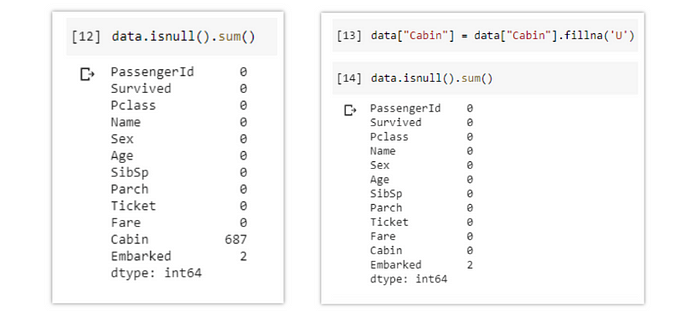
* Prevent data loss which results in deletion of rows or columns
* Works well with a small dataset and is easy to implement.

**Cons:**

* Works only with numerical continuous variables.
* Can cause data leakage
* Do not factor the covariance between features.

**Imputation method for categorical columns:**

When missing values is from categorical columns (string or numerical) then the missing values can be replaced with the most frequent category. If the number of missing values is very large then it can be replaced with a new category.



(Image by Author) Left: Data before Imputation, Right: Cabin column after imputation by ‘U’

**Pros:**

* Prevent data loss which results in deletion of rows or columns
* Works well with a small dataset and is easy to implement.
* Negates the loss of data by adding a unique category

**Cons:**

* Works only with categorical variables.
* Addition of new features to the model while encoding, which may result in poor performance

**Other Imputation Methods:**

Depending on the nature of the data or data type, some other imputation methods may be more appropriate to impute missing values.

For example, for the data variable having longitudinal behavior, it might make sense to use the last valid observation to fill the missing value. This is known as the Last observation carried forward (LOCF) method.

For the time-series dataset variable, it makes sense to use the interpolation of the variable before and after a timestamp for a missing value.

**Using Algorithms that support missing values:**

All the machine learning algorithms don’t support missing values but some ML algorithms are robust to missing values in the dataset. The k-NN algorithm can ignore a column from a distance measure when a value is missing. Naive Bayes can also support missing values when making a prediction. These algorithms can be used when the dataset contains null or missing values.

The sklearn implementations of naive Bayes and k-Nearest Neighbors in Python do not support the presence of the missing values.

Another algorithm that can be used here is RandomForest that works well on non-linear and categorical data. It adapts to the data structure taking into consideration the high variance or the bias, producing better results on large datasets.

**Pros:**

* No need to handle missing values in each column as ML algorithms will handle them efficiently.

**Cons:**

* No implementation of these ML algorithms in the scikit-learn library.

**Prediction of missing values:**

In the earlier methods to handle missing values, we do not use the correlation advantage of the variable containing the missing value and other variables. Using the other features which don’t have nulls can be used to predict missing values.

The regression or classification model can be used for the prediction of missing values depending on the nature (categorical or continuous) of the feature having missing value.

Here 'Age' column contains missing values so for prediction of null values the spliting of data will be,**y\_train**: rows from data["Age"] with non null values  
**y\_test**: rows from data["Age"] with null values  
**X\_train**: Dataset except data["Age"] features with non null values  
**X\_test**: Dataset except data["Age"] features with null values

(Code by Author)

**[Predict Missing Values in the Dataset](https://towardsdatascience.com/predict-missing-values-in-the-dataset-897912a54b7b" \t "_blank)**

[Understand how to predict missing values in the dataset using a Machine Learning model and its Implementation](https://towardsdatascience.com/predict-missing-values-in-the-dataset-897912a54b7b" \t "_blank)

[towardsdatascience.com](https://towardsdatascience.com/predict-missing-values-in-the-dataset-897912a54b7b" \t "_blank)

**Pros:**

* Gives a better result than earlier methods
* Takes into account the covariance between the missing value column and other columns.

**Cons:**

* Considered only as a proxy for the true values

**Imputation using Deep Learning Library — [Datawig](https://github.com/awslabs/datawig" \t "_blank)**

This method works very well with categorical, continuous, and non-numerical features. Datawig is a library that learns ML models using Deep Neural Networks to impute missing values in the datagram.

Install datawig library,  
**pip3 install datawig**

Datawig can take a data frame and fit an imputation model for each column with missing values, with all other columns as inputs.

Below is the code to impute missing values in the *Age* column

(Code by Author)

**Pros**:

* Quite accurate compared to other methods.
* It supports CPUs and GPUs.

**Cons:**

* Can be quite slow with large datasets.

**Conclusion:**

Every dataset has missing values that need to be handled intelligently to create a robust model. In this article, I have discussed 7 ways to handle missing values that can handle missing values in every type of column. There is no thump rule to handle missing values in a particular manner, the method which gets a robust model with the best performance. One can use various methods on different features depending on how and what the data is about. Having domain knowledge about the dataset is important, which can give an insight into how to preprocess the data and handle missing values.

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

Data preprocessing is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.

**Some common steps in data preprocessing include:**

Data preprocessing is an important step in the data mining process that involves cleaning and transforming raw data to make it suitable for analysis. Some common steps in data preprocessing include:

**Data Cleaning:**This involves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Various techniques can be used for data cleaning, such as imputation, removal, and transformation.

**Data Integration:**This involves combining data from multiple sources to create a unified dataset. Data integration can be challenging as it requires handling data with different formats, structures, and semantics. Techniques such as record linkage and data fusion can be used for data integration.

**Data Transformation:**This involves converting the data into a suitable format for analysis. Common techniques used in data transformation include normalization, standardization, and discretization. Normalization is used to scale the data to a common range, while standardization is used to transform the data to have zero mean and unit variance. Discretization is used to convert continuous data into discrete categories.

**Data Reduction:**This involves reducing the size of the dataset while preserving the important information. Data reduction can be achieved through techniques such as feature selection and feature extraction. Feature selection involves selecting a subset of relevant features from the dataset, while feature extraction involves transforming the data into a lower-dimensional space while preserving the important information.

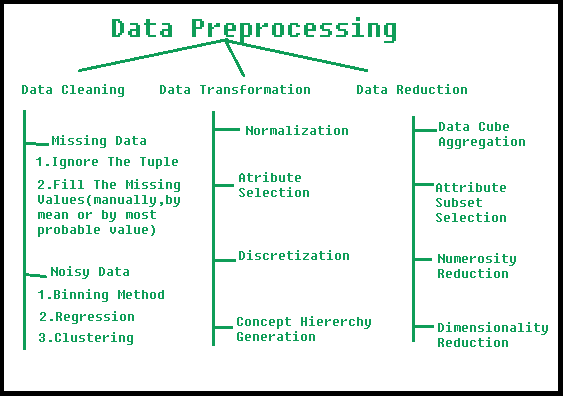
**Data Discretization:**This involves dividing continuous data into discrete categories or intervals. Discretization is often used in data mining and machine learning algorithms that require categorical data. Discretization can be achieved through techniques such as equal width binning, equal frequency binning, and clustering.

**Data Normalization:**This involves scaling the data to a common range, such as between 0 and 1 or -1 and 1. Normalization is often used to handle data with different units and scales. Common normalization techniques include min-max normalization, z-score normalization, and decimal scaling.

Data preprocessing plays a crucial role in ensuring the quality of data and the accuracy of the analysis results. The specific steps involved in data preprocessing may vary depending on the nature of the data and the analysis goals.

By performing these steps, the data mining process becomes more efficient and the results become more accurate.

**Preprocessing in Data Mining:**   
Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.



**Steps Involved in Data Preprocessing:**

**1. Data Cleaning:**   
The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc. 

* **(a). Missing Data:**   
  This situation arises when some data is missing in the data. It can be handled in various ways.   
  Some of them are:
  1. **Ignore the tuples:**   
     This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
  2. **Fill the Missing values:**   
     There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.
* **(b). Noisy Data:**   
  Noisy data is a meaningless data that can’t be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :
  1. **Binning Method:**   
     This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.
  2. **Regression:**   
     Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).
  3. **Clustering:**   
     This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

**2. Data Transformation:**   
This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

1. **Normalization:**   
   It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)
2. **Attribute Selection:**   
   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.
3. **Discretization:**   
   This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.
4. **Concept Hierarchy Generation:**   
   Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

**3. Data Reduction:**   
Data reduction is a crucial step in the data mining process that involves reducing the size of the dataset while preserving the important information. This is done to improve the efficiency of data analysis and to avoid overfitting of the model. Some common steps involved in data reduction are:

**Feature Selection:** This involves selecting a subset of relevant features from the dataset. Feature selection is often performed to remove irrelevant or redundant features from the dataset. It can be done using various techniques such as correlation analysis, mutual information, and principal component analysis (PCA).

**Feature Extraction:**This involves transforming the data into a lower-dimensional space while preserving the important information. Feature extraction is often used when the original features are high-dimensional and complex. It can be done using techniques such as PCA, linear discriminant analysis (LDA), and non-negative matrix factorization (NMF).

**Sampling:**This involves selecting a subset of data points from the dataset. Sampling is often used to reduce the size of the dataset while preserving the important information. It can be done using techniques such as random sampling, stratified sampling, and systematic sampling.

**Clustering:**This involves grouping similar data points together into clusters. Clustering is often used to reduce the size of the dataset by replacing similar data points with a representative centroid. It can be done using techniques such as k-means, hierarchical clustering, and density-based clustering.

**Compression:** This involves compressing the dataset while preserving the important information. Compression is often used to reduce the size of the dataset for storage and transmission purposes. It can be done using techniques such as wavelet compression, JPEG compression, and gzip compression.

9.

* + 1. What is the IQR? What criteria are used to assess it?

The interquartile range rule is useful in detecting the presence of outliers. [Outliers](https://www.thoughtco.com/what-is-an-outlier-3126227) are individual values that fall outside of the overall pattern of a data set. This definition is somewhat vague and subjective, so it is helpful to have a rule to apply when determining whether a data point is truly an outlier—this is where the interquartile range rule comes in.

What Is the Interquartile Range?

Any set of data can be described by its [five-number summary](https://www.thoughtco.com/what-is-the-five-number-summary-3126237). These five numbers, which give you the information you need to find patterns and outliers, consist of (in ascending order):

* The minimum or lowest value of the dataset
* The first quartile *Q*1, which represents a quarter of the way through the list of all data
* The [median](https://www.thoughtco.com/what-is-the-median-3126370) of the data set, which represents the midpoint of the whole list of data
* The third quartile *Q*3, which represents three-quarters of the way through the list of all data
* The maximum or highest value of the data set.

These five numbers tell a person more about their data than looking at the numbers all at once could, or at least make this much easier. For example, the [range](https://www.thoughtco.com/what-is-the-range-in-statistics-3126248), which is the minimum subtracted from the maximum, is one indicator of how spread out the data is in a set (note: the range is highly sensitive to outliers—if an outlier is also a minimum or maximum, the range will not be an accurate representation of the breadth of a data set).

Range would be difficult to extrapolate otherwise. Similar to the range but less sensitive to outliers is the interquartile range. The [interquartile range](https://www.thoughtco.com/what-is-the-interquartile-range-3126245) is calculated in much the same way as the range. All you do to find it is subtract the first quartile from the third quartile:

IQR = *Q*3 – *Q*1.

The interquartile range shows how the data is spread about the median. It is less susceptible than the range to outliers and can, therefore, be more helpful.

Using the Interquartile Rule to Find Outliers

Though it's not often affected much by them, the interquartile range can be used to detect outliers. This is done using these steps:

1. Calculate the interquartile range for the data.
2. Multiply the interquartile range (IQR) by 1.5 (a constant used to discern outliers).
3. Add 1.5 x (IQR) to the third quartile. Any number greater than this is a suspected outlier.
4. Subtract 1.5 x (IQR) from the first quartile. Any number less than this is a suspected outlier.

Remember that the interquartile rule is only a rule of thumb that generally holds but does not apply to every case. In general, you should always follow up your outlier analysis by studying the resulting outliers to see if they make sense. Any potential outlier obtained by the interquartile method should be examined in the context of the entire set of data.

Interquartile Rule Example Problem

See the interquartile range rule at work with an example. Suppose you have the following set of data: 1, 3, 4, 6, 7, 7, 8, 8, 10, 12, 17. The five-number summary for this data set is minimum = 1, first quartile = 4, median = 7, [third quartile](https://www.thoughtco.com/what-are-first-and-third-quartiles-3126235) = 10 and maximum = 17. You may look at the data and automatically say that 17 is an outlier, but what does the interquartile range rule say?

If you were to calculate the interquartile range for this data, you would find it to be:

*Q*3 – *Q*1 = 10 – 4 = 6

Now multiply your answer by 1.5 to get 1.5 x 6 = 9. Nine less than the first quartile is 4 – 9 = -5. No data is less than this. Nine more than the third quartile is 10 + 9 =19. No data is greater than this. Despite the maximum value being five more than the nearest data point, the interquartile range rule shows that it should probably not be considered an outlier for this data set.

* + 1. 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?
* A box and whisker plot—also called a box plot—displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum.
* In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median. The whiskers go from each quartile to the minimum or maximum.
* min�1*Q*1​Q, start subscript, 1, end subscriptmedian�3*Q*3​Q, start subscript, 3, end subscriptmax
* A box and whisker plot with the left end of the whisker labeled min, the right end of the whisker is labeled max. The beginning of the box is labeled Q 1. The end of the box is labeled Q 3. The line that divides the box is labeled median.

### Example: Finding the five-number summary

* A sample of 101010 boxes of raisins has these weights (in grams):
* 252525, 282828, 292929, 292929, 303030, 343434, 353535, 353535, 373737, 383838
* **Make a box plot of the data.**
* **Step 1:** Order the data from smallest to largest.
* Our data is already in order.
* 252525, 282828, 292929, 292929, 303030, 343434, 353535, 353535, 373737, 383838
* **Step 2:** Find the median.
* The median is the mean of the middle two numbers:
* 252525, 282828, 292929, 292929, 303030, 343434, 353535, 353535, 373737, 383838
* 30+342=32230+34​=32start fraction, 30, plus, 34, divided by, 2, end fraction, equals, 32
* The median is 323232.
* **Step 3:** Find the quartiles.
* The first quartile is the median of the data points to the left of the median.
* 252525, 282828, 292929, 292929, 303030
* �1=29*Q*1​=29Q, start subscript, 1, end subscript, equals, 29
* The third quartile is the median of the data points to the right of the median.
* 343434, 353535, 353535, 373737, 383838
* �3=35*Q*3​=35Q, start subscript, 3, end subscript, equals, 35
* **Step 4:** Complete the five-number summary by finding the min and the max.
* The min is the smallest data point, which is 252525.
* The max is the largest data point, which is 383838.
* The five-number summary is 252525, 292929, 323232, 353535, 383838.

### Example (continued): Making a box plot

* Let's make a box plot for the same dataset from above.
* **Step 1:** Scale and label an axis that fits the five-number summary.
* 252525282828313131343434373737404040Weight (grams)
* A number line labeled weight in grams. It is numbered from 25 to 40.
* **Step 2:** Draw a box from �1*Q*1​Q, start subscript, 1, end subscript to �3*Q*3​Q, start subscript, 3, end subscript with a vertical line through the median.
* Recall that �1=29*Q*1​=29Q, start subscript, 1, end subscript, equals, 29, the median is 323232, and �3=35.*Q*3​=35.Q, start subscript, 3, end subscript, equals, 35, point
* 252525282828313131343434373737404040�1*Q*1​Q, start subscript, 1, end subscriptmedian�3*Q*3​Q, start subscript, 3, end subscriptWeight (grams)
* The box of a box and whisker plot without the whiskers. The beginning of the box is labeled Q 1 at 29. The end of the box is labeled Q 3 at 35. The vertical line that divides the box is labeled median at 32.
* **Step 3:** Draw a whisker from �1*Q*1​Q, start subscript, 1, end subscript to the min and from �3*Q*3​Q, start subscript, 3, end subscript to the max.
* Recall that the min is 252525 and the max is 383838.
* 252525282828313131343434373737404040min�1*Q*1​Q, start subscript, 1, end subscriptmedian�3*Q*3​Q, start subscript, 3, end subscriptmaxWeight (grams)
* A box and whisker plot. The beginning of the box is labeled Q 1 at 29. The end of the box is labeled Q 3 at 35. The vertical line that divides the box is labeled median at 32. The left part of the whisker is labeled min at 25. The right part of the whisker is labeled max 38.
* We don't need the labels on the final product:
* 252525282828313131343434373737404040Weight (grams)
* A box and whisker plot. The beginning of the box is at 29. The end of the box is at 35. The vertical line that divides the box is at 32. The left part of the whisker is at 25. The right part of the whisker is at 38.
* Want to learn more about making box and whisker plots? Check out [*this video*](https://www.khanacademy.org/v/box-and-whisker-plot-exercise-example).
* Want to practice making box plots? Check out [*this exercise*](https://www.khanacademy.org/e/box-plots).

### Interpreting quartiles

* The five-number summary divides the data into sections that each contain approximately 25%25%25, percent of the data in that set.
* min25%25%25, percent�1*Q*1​Q, start subscript, 1, end subscript25%25%25, percentmedian�3*Q*3​Q, start subscript, 3, end subscript25%25%25, percentmax25%25%25, percent
* A box and whisker plot with the left end of the whisker labeled min, the right end of the whisker is labeled max. The beginning of the box is labeled Q 1. The end of the box is labeled Q 3. The line that divides the box is labeled median. The distance from the min to the Q 1 is twenty five percent. The distance from the Q 1 to the Q 2 is twenty five percent. The distance from the Q 2 to the Q 3 is twenty five percent. The distance from the Q 3 is Max is twenty five percent.

### Example: Interpreting quartiles

* **About what percent of the boxes of raisins weighed more than 292929 grams?**
* 252525282828313131343434373737404040Weight (grams)
* A box and whisker plot. The beginning of the box is at 29. The end of the box is at 35. The vertical line that divides the box is at 32. The left part of the whisker is at 25. The right part of the whisker is at 38.
* Since �1=29*Q*1​=29Q, start subscript, 1, end subscript, equals, 29, about 25%25%25, percent of data is lower than 292929 and about 75%75%75, percent is above is 292929.
* 252525282828313131343434373737404040�1*Q*1​Q, start subscript, 1, end subscript25%25%25, percent25%25%25, percent25%25%25, percentWeight (grams)
* A box and whisker plot. The beginning of the box is labeled Q 1 at 29. The end of the box is at 35. The vertical line that divides the box is at 32. The left part of the whisker is at 25. The right part of the whisker is at 38. The distance from the Q 1 to the dividing vertical line is twenty five percent. The distance from the vertical line to the end of the box is twenty five percent. The distance from the Q 3 is Max is twenty five percent.
* About 75%75%75, percent of the boxes of raisins weighed more than 292929 grams.

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

* 1. Data collected at regular intervals
* 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.

2. The gap between the quartiles

The median is considered the second quartile (Q2). The interquartile range is the difference between upper and lower quartiles. The semi-interquartile range is half the interquartile range.

3. Use a cross-tab

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.

* 1. Make a comparison between:

1. Data with nominal and ordinal values

Nominal vs. Ordinal Data

Ordinal data is a kind of qualitative data that groups variables into ordered categories. The categories have a natural order or rank based on some hierarchal scale.

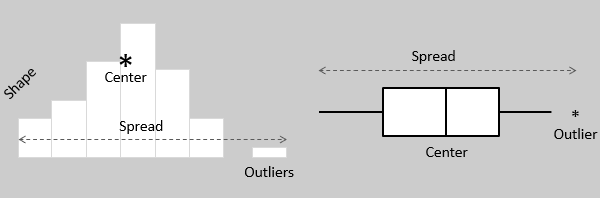
The main differences between Nominal Data and Ordinal Data are:

* While Nominal Data is classified without any intrinsic ordering or rank, Ordinal Data has some predetermined or natural order.
* Nominal data is qualitative or categorical data, while Ordinal data is considered “in-between” qualitative and quantitative data.
* Nominal data do not provide any quantitative value, and you cannot perform numeric operations with them or compare them with one another. However, Ordinal data provide sequence, and it is possible to assign numbers to the data. No numeric operations can be performed. But ordinal data makes it possible to compare one item with another in terms of ranking.
* Example of Nominal Data – Eye color, Gender; Example of Ordinal data – Customer Feedback, Economic Status

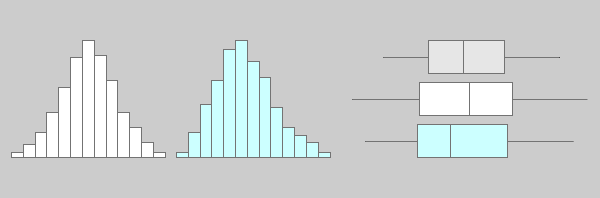
2. Histogram and box plot

[Histograms](https://citoolkit.com/articles/histogram/) and [box plots](https://citoolkit.com/articles/box-plot/) are graphical representations for the frequency of numeric data values. They aim to [describe](https://citoolkit.com/articles/descriptive-statistics/) the data and explore the central tendency and variability before using advanced statistical analysis techniques. In this article, we will further discuss the similarities and differences between these two tools.

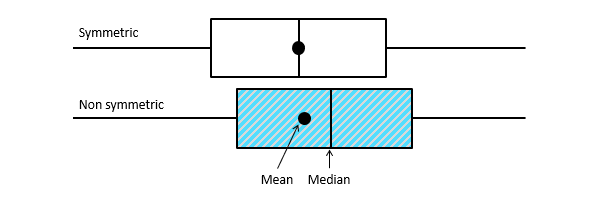
Both histograms and box plots allow to visually assess the central tendency, the amount of variation in the data as well as the presence of gaps, outliers or unusual data points.



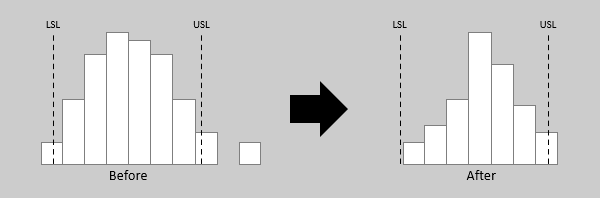
Both histograms and box plots are used to explore and present the data in an easy and understandable manner. Histograms are preferred to determine the underlying [probability distribution](https://citoolkit.com/articles/probability-distributions/) of a data. Box plots on the other hand are more useful when comparing between several data sets. They are less detailed than histograms and take up less space.



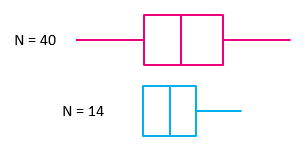
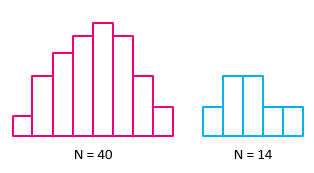
Although histograms are better in displaying the [distribution](https://citoolkit.com/articles/probability-distributions/) of data, you can use a box plot to tell if the distribution is symmetric or skewed. In a symmetric distribution, the mean and median are nearly the same, and the two whiskers has almost the same length.



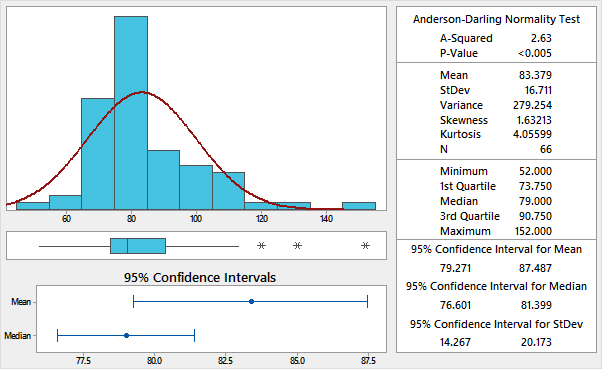
You can use histograms and box plots to verify whether an improvement has been achieved by exploring the data before and after the improvement initiative. Both tools can be helpful to identify whether variability is within specification limits, whether the process is capable, and whether there is a shift in the process over time.



Both histograms and box plots are ideal to represent moderate to large amount of data. They may not accurately display the distribution shape if the data size is too small. In practice, a sample size of at least 30 data values would be sufficient for both tools.



Many statistical applications allow the option of summarizing your data [graphically](https://citoolkit.com/articles/graphical-analysis/) (including plotting the data on histograms and box plots as shown below). This can reveal unusual observations in your data that should be investigated before performing detailed statistical analysis.



1. The average and median

### Average                                                       Median

|  |  |  |
| --- | --- | --- |
| **Definition** | The average is the arithmetic mean of a set of numbers. | The median is a numeric value that separates the higher half of a set from the lower half. |
| **When is it applicable?** | The mean is used for normal number distributions, which have a low amount of outliers. | The median is generally used to return the central tendency for skewed number distributions. |
| **How is it calculated?** | The average is calculated by adding up all the values and dividing the sum by the total number of values. | The median can be calculated by listing all numbers in ascending order and then locating the number in the centre of that distribution. |
| **Example: Normal distribution** | 2, 3, 3, 5, 8, 10, 11  (2+3+3+5+8+10+11)/7= 6  **AVG = 6** | 2, 3, 3, **5**, 8, 10, 11    **MED = 5** |
| **Example: Skewed distribution** | 2, 2, 3, 3, 5, 7, 8, 130  (2+2+3+3+5+7+8+130)/8= 20  **AVG = 20** | 2, 2, 3,**3**,**5**, 7, 8, 130  (3+5)/2=4  **MED = 4** |