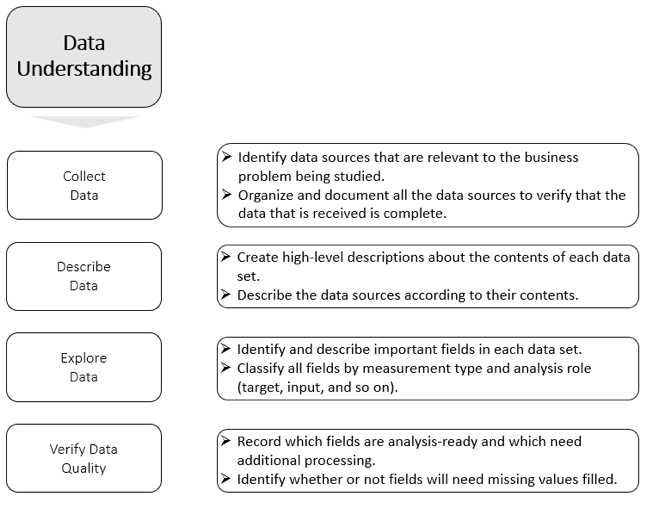
**Data Understanding main activities**

The following figure shows the four main activities for this phase. Each main activity is further divided into subtasks that explain the purpose of this activity.



**Collecting Data**

The data that you are looking for might not be available. In many cases, data scientists work with the available data sets that are collected during investigations and activities, which can place limitations on the questions that can be answered from these data sets, that is, the available data might not serve to solve the business problem.

Data scientists must consider the following items while collecting the data:

* Whether the data exists in the real world and if it is relevant to the business problem that is being studied.
* How this data can be collected in terms of efforts, time, privacy issues, and logistics.

**How the data is collected**

Here is the methodology that you can use to collect the data:

* Document and describe the data sources:
  1. Create a list of all the available data sources. The data might be enterprise data (customer-provided) or third-party like social media (Facebook, Twitter, and Instagram).
  2. Describe each data source volume (number of records) and its nature (structured, semi-structure, or unstructured).
* Verify data availability and document any restrictions.

Confirm whether the required data is available and whether it can be used. If some of the data is not available, you must address this issue and consider other alternatives, such as the following ones:

* 1. Use alternative data sources.
  2. Revisit the Business Understanding topic and minimize the scope of the project.

**Describing the Data**

Now that you have the data, you can deliver the data description report. This report usually has the following information:

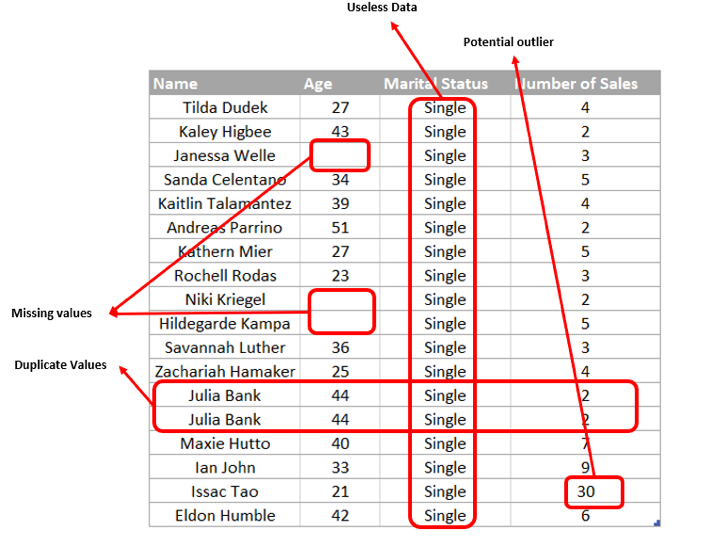
* **Business context**: Create a list of the most promising fields of all the data sources that are relevant to the business objectives.
* **Technical context**: The description of each data source in detail:
  1. Format of data (structured, semi-structured, or unstructured)
  2. Data dictionary, which contains the number of fields and their meaning to the business. For example:
     + Data set name: Customers, no of columns: 10.
     + Column: INCOME, Definition: The monthly income of each customer.
     + Column: Age, Definition: The Age of each customer.

**Exploring Data**

In this task, you use summary statistics, querying, visualization, and other items to better understand the data. You also can spot any data quality issues, such as:

* Potential outliers
* Missing data
* Useless data
* Duplicate records

The following table shows examples of data quality issues.



Here are descriptions of the data quality issues that are shown in the table:

* **Name:** The customer name or ID. It is text and may use any value. It is only an ID and does not affect the results of the modelling technique.
* **Age**: The age of the customer. There are missing values.
* **Marital Status**: Defines the marital status of the customer. It normally uses multiple values such as single, married, divorced, or widowed, but in this case the whole field accepts only one value. This situation is called zero variance, which makes this field useless.
* **Number of Sales**: The number of transactions for each customer. Most customers had fewer than 10 transactions, but there is one customer, “Issac Tao”, who had 30 transactions, which can be considered an outlier. The possible actions are to exclude the customer from the analysis or replace his number of transactions by the mean or median.

For the sake of illustration, we used a small data sample in the example, which was easy to review. However, what if the data set is large? With the aid of computers, data scientists can draw conclusions from large data sets and verify data issues by using code and plots.

**Verifying the data quality**

The deliverable for this task is the data quality report, which summarizes the data that you have, minor and major quality issues that you found, and possible remedies for quality problems or alternatives (such as using an alternative data resource). If you are facing any serious data quality issues and cannot identify an adequate solution, you might have to recommend reconsidering goals or plans that were defined earlier in the Business Understanding phase.

# **Data preparation**

The key objective of this phase is to construct a data set that is appropriate for modeling. According to IBM Big Data and Analytics HUB, Data Preparation is the most time-consuming phase of the overall project time. It can consume 70 - 90 % of the project time.

For more information and a detailed graph, see the following website:

<https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#56c1c7c76f63>

### ****Data Preparation main activities****

The following figure shows the four main activities for this phase. Each main activity is further divided into subtasks that explain the purpose of this activity.

### https://dw1.s81c.com/caas-storage/skillscollection/dna/africa-prod/data-understanding-and-preparation/en/_attachments/Data_Preparation_main_activities.png

**Selecting Data**

Machine learning works on a simple rule: If you put garbage in, you get garbage out (GIGO). By garbage, we mean noise in data or data quality issues.

GIGO becomes more obvious when the number of input variables (features) is large. You do not need to use every input feature. You can assist the algorithm by feeding it only those features that are relevant.

**Why feature selection is important**

Selecting features that are relevant to the algorithm produces the following outcomes:

### Enables the machine learning algorithm to train faster by reducing the number of input features.

### Reduces the complexity of a model and makes it easier to interpret.

### If the right subset is chosen, then it improves the accuracy of the model.

The selection may be based on its relevance to your goals, data quality, and technical issues, such as limits to the number of fields or rows that your tools can handle.

Features can be selected based on their scores in various statistical tests for their correlation or association with the target variable. For example, if you have a data set like the one that is shown in the following table and want to predict whether a car is stolen or not, then the field “Stolen” is the target, and all other variables are inputs (they are sometimes called predictors because they help in predicting the target variable).

### https://dw1.s81c.com/caas-storage/skillscollection/dna/africa-prod/data-understanding-and-preparation/en/_attachments/26-1.png

You might calculate association scores for the predictors in the above table with the target variable. Based on their scores, only relevant features are considered in your model. The higher the correlation score, the more relevant the predictor is to the target variable.

There are various statistical tests that can be used to determine the significance of a relationship between different types of variables. The following table summarizes the names of these tests.

### https://dw1.s81c.com/caas-storage/skillscollection/dna/africa-prod/data-understanding-and-preparation/en/_attachments/27-1.png

In this project, we rely on plotting (graphical) techniques to determine the significance of the relationships between the variables. We do not use any statistical tests.

**Cleaning Data**

The data that is selected to use in this case is not clean (that is, it still has quality issues). In this section, you address each quality issue. This section documents in detail the decisions and action you use to clean the data.

The section Exploring data addressed the data quality issues. Now, let us consider how we can resolve them:

* **Missing data**: How do you compensate for missing data? Should you remove the whole record? Fill it with zeros, or the most frequent or mean values from the same field? If the missing value is text, how do you handle it?

There are many options, and to decide on which one to use, you refer to the Business Understanding phase or engage the customer directly because the decision might shift the modelling results significantly.

### ****Zero variance****: The data uses a single value that does not vary with the samples. Therefore, these kinds of columns can be eliminated because they do not offer something new to the analysis.

### ****Duplicate values****: Duplicate records are useless and can be safely removed.

* **Outliers:** Some predictive models are more sensitive to outliers than others, that is, those models perform less accurately when the data set has many outliers in it. Should you, for example, remove or keep it, or replace it by using the population mean?

An outlier might indicate bad data. For example, the data might be coded incorrectly, or an experiment might not have been run correctly. If it can be determined that an outlying point is in fact erroneous, then the outlying value should be deleted from the analysis (or corrected if possible).

In some cases, it might not be possible to determine whether an outlier is bad data. Outliers might be due to random variation or indicate something scientifically interesting. In any event, you should not delete the outlier before performing a thorough investigation.

**Constructing data**

Construction of data here encompasses two things:

### Transformation of the values of existing fields

### Derivation of new fields that have not been seen before out of the available data.

We construct data to derive better features.

The features in your data are important to the predictive models that you use and influence the results that you are going to achieve. The quality and quantity of the features have great influence on whether the model is good.

The better the features are, the better the result is, but this is not entirely true because the results that are achieved also depend on the model and the data, and not only the chosen features. Choosing the correct features is still important. Better features can produce simpler and more flexible models, and they often yield better results.

**Derivation of new data**

Here are the types of new data that you can derive:

### ****Derived attributes:**** These are new attributes that are constructed from one or more existing attributes in the same record. For example, you might use the birth dates of customers to calculate their ages.

### ****Generated records****: Here you describe the creation of any new records. For example, you might need to create records for customers who made no purchases during the past year. There was no reason to have such records in the raw data, but for modelling purposes, it might make sense to explicitly represent the fact that some customers made zero purchases.

**Transformation of existing features**

Transformation is mapping the set of values of a feature to a new set of values to make the representation of the data more suitable or easier to process in the downstream analysis.

**Common feature transformation operations**

Here are some common feature transformation operations:

* **Scaling**: The process of transforming the set of values of a field to another specified range.

For example, if you have a field for the distance traveled that is measured in meters and you want to transform it into kilometers, divide all the values by 1000. Another example is where you can convert the distance traveled into a new scale of 0 - 1.

But why transform fields into a new scale? To answer that question, let us look at the following scenario.

Assume that you have the following table with Age and Income for a group of employees. The scale of Income is much larger than the Age. The income is in thousands and age is in tens, although they measure different quantities.

### https://dw1.s81c.com/caas-storage/skillscollection/dna/africa-prod/data-understanding-and-preparation/en/_attachments/28-1.png

These different scale formats might cause Income to dominate the results of the machine learning algorithm over the age. Therefore, you convert both scales into a common scale format so that the contributions of both are equal.

There are multiple approaches that can be used for scaling:

* 1. **Min-max scale:** This approach outputs values 0 – 1. The following formula calculates this scale:

### https://dw1.s81c.com/caas-storage/skillscollection/dna/africa-prod/data-understanding-and-preparation/en/_attachments/29-1.png

X is the input field value, and Xmin and Xmax are the input values range.

* 1. **Zero-normalization/Standardization**: This approach outputs values -1 - 1 with a zero mean and a variance equal to 1. The following formula calculates this scale:

### https://dw1.s81c.com/caas-storage/skillscollection/dna/africa-prod/data-understanding-and-preparation/en/_attachments/30-1.png

X is the input field value, µ is the mean of the input values, and σ is the standard deviation.

* **Binning and aggregation:** The process of grouping the values of features into groups to summarize data or reduce variability.

Each value within the same group can be approximated to the mean, median, or sum, or the highest or lowest of the samples within the same group.

You can perform binning and aggregation summaries for continuous or ordinal fields.

For example, the Age field in the following figure has values 1 - 25, which are divided into six groups. For example, all ages 1 - 5 are grouped and are approximated to the mean of all samples within this group and the rest of the groups.

### https://dw1.s81c.com/caas-storage/skillscollection/dna/africa-prod/data-understanding-and-preparation/en/_attachments/31-1.png

Although binning or aggregation wipes out some of the characteristics of the data by approximating a group of points to a common value, why do we apply it? Is it always necessary? Binning or Aggregation loses some of the information about data, but it can be beneficial in multiple areas.

For example, suppose that you want to describe something with high fluctuation that can be hard to explain, such as a stock price exchange. The stock prices fluctuate every minute, so it can be hard to trace or summarize the data. Binning and aggregation provide a summary that is descriptive and informative.

Another benefit of binning and aggregation is when a field contains missing values for a significant proportion of the population but might be highly predictive, and the missing values themselves bear predictive value about a distinct class within the data.

**Integrating Data**

During the Business Understanding phase, specifically in the Data Collection task, data might come from multiple sources. The data can be customer-provided (enterprise data) or from third-party sources like social media (Facebook, Twitter, Amazon, and others). In this section, you perform merging, appending, and aggregation operations on multiple sources.

Data integration encompasses two methods of combining data:

### Merging data sets (combining them as needed)

### Aggregation (generating new summaries from the merged data)

This activity is described in the Project section.

**Formatting Data**

Since the tasks for this activity depend on the chosen machine learning models, this activity is explained in the Classification Modeling course.