Data analytics is all about looking at various factors to see how they impact certain situations and outcomes. When dealing with data that contains more than two variables, you’ll use multivariate analysis.

Multivariate analysis isn’t just one specific method—rather, it encompasses a whole range of statistical techniques. These techniques allow you to gain a deeper understanding of your data in relation to specific business or real-world scenarios.

(Link: https://careerfoundry.com/en/blog/data-analytics/multivariate-analysis/)

**1. What is multivariate analysis?**

In data analytics, we look at different variables (or factors) and how they might impact certain situations or outcomes.

For example, in marketing, you might look at how the variable “money spent on advertising” impacts the variable “number of sales.” In the healthcare sector, you might want to explore whether there’s a correlation between “weekly hours of exercise” and “cholesterol level.” This helps us to understand why certain outcomes occur, which in turn allows us to make informed predictions and decisions for the future.

There are three categories of analysis to be aware of:

* **Univariate analysis**, which looks at just one variable
* **Bivariate analysis**, which analyzes two variables
* **Multivariate analysis**, which looks at more than two variables

As you can see, multivariate analysis encompasses all statistical techniques that are used to analyze more than two variables at once. The aim is to find patterns and [correlations](https://careerfoundry.com/en/blog/data-analytics/covariance-vs-correlation/) between several variables simultaneously—allowing for a much deeper, more complex understanding of a given scenario than you’ll get with bivariate analysis.

## 1.1 What is correlation?

**Correlation** tells us both the strength and the direction of this relationship. Correlation is best used for multiple variables that express a linear relationship with one another.

When we assume a correlation between two variables, we are essentially deducing that a change in one variable impacts a change in another variable. Correlation helps us to determine whether or not, and how strongly, changes in various variables relate to each other.

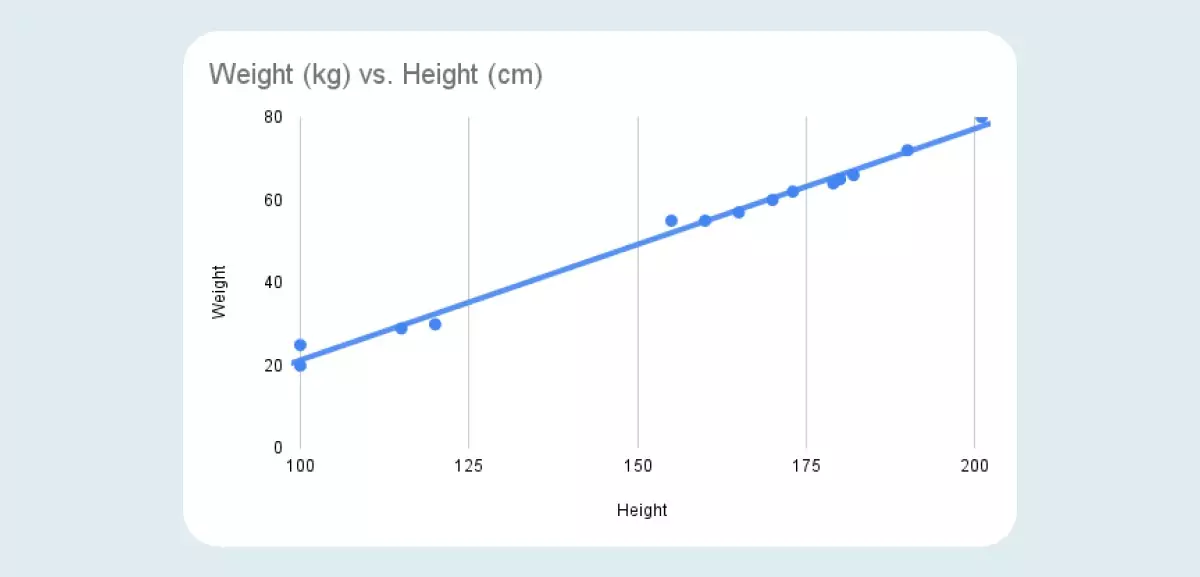
### Types of correlation

Correlation is classified into the following types based on diverse values: Positive correlation, negative correlation, and no correlation. Let’s explore those now.

#### Positive correlation

Two variables are considered to have a positive correlation if they are directly proportional. That is, if the value of one variable increases, then the value of the other variable will also increase.

A perfect positive correlation holds a value of “1”. On a graph, positive correlation appears as follows:



#### Negative correlation

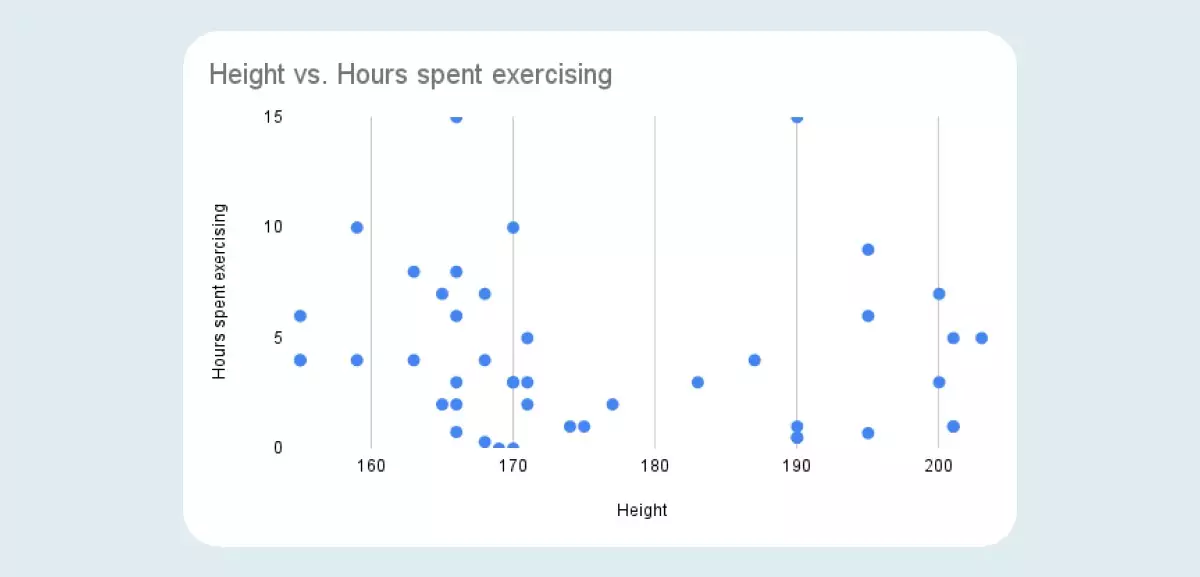
A perfect negative correlation holds a value of “-1” which means that, as the value of one variable increases, the value of the second variable decreases (and vice versa). In graph form, this is how negative correlation might look:



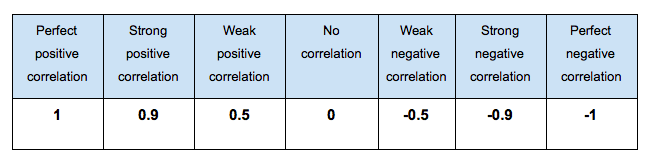
Zero or no correlation

The value “0” denotes that there is no correlation. It indicates that there is no relationship between the two variables, so an increase or decrease in one variable is unrelated to an increase or decrease in the other variable.

A graph showing zero correlation will follow a random distribution of data points, as opposed to a clear line:

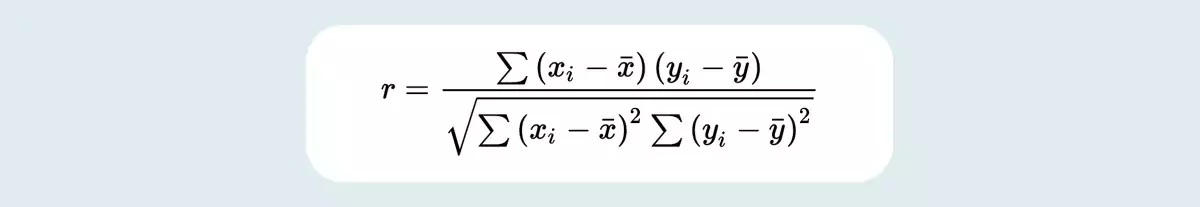


This table perfectly displays the varying degree of correlation between two values.



### What is the correlation coefficient?

Correlation is calculated using a method known as “Pearson’s Product-Moment Correlation” or simply “Correlation Coefficient.” Correlation is usually denoted by italic letter r. The following formula is normally used to find r for two variables X and Y.



Where:

* **r** represents the correlation coefficient
* **xi** represents the value of variable X in data sample
* **x** represents the mean (average) of values of X variable
* **yi** represents the value of variable Y in data sample
* **y** represents the mean (average) of Y variable

#### Alternative methods to calculate the correlation coefficient

Besides Pearson’s Product-Moment Correlation, some alternative techniques that are helpful in calculating correlation coefficient include:

* [**Spearman’s rank correlation**](https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient)
* [**Kendall rank correlation**](https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient)

### What is a correlation matrix?

We use correlation coefficients to determine the relationship between two variables, for example, to find the number of hours a student must spend working to complete a project within the desired timeline. But what if we want to evaluate the correlation among multiple pairs of variables? Then we use a correlation matrix.

A correlation matrix is essentially a table depicting the correlation coefficients for various variables. The rows and columns contain the value of the variables, and each cell shows the correlation coefficient.

#### Example:

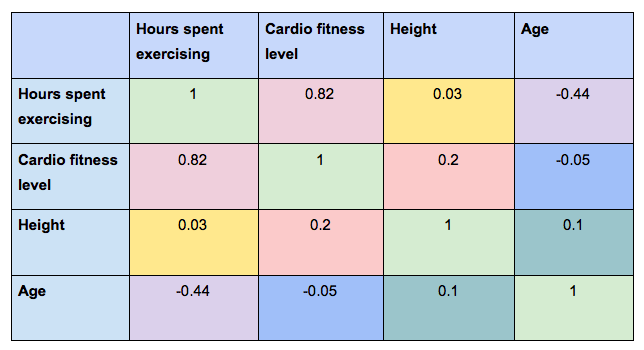
In the following correlation matrix, we can see the correlation coefficient for each possible combination of variables. In this example, we’re looking at:

* Hours spent exercising (per week)
* Cardio fitness level
* Height
* Age

If you consider the two variables “Hours spent exercising” and “cardio fitness level,” you’ll see that the correlation coefficient is 0.82.

This indicates a strong positive correlation between the two variables—which makes sense, right? The more you exercise, the more you’d expect your cardio fitness to increase.

At the same time, the correlation coefficient when comparing “Hours spent exercising” to “Height” is 0.03, indicating that there’s no association between how tall someone is and how often they exercise.



#### What is the correlation matrix used for?

Practically, the correlation matrix is used to analyze different data-driven problems. A few common use cases include:

* **To conveniently encapsulate datasets:** For large datasets containing thousands of rows, the correlation matrix is an effective way of summarizing the correlation among various variables of that dataset. The relationship between two variables can easily be interpreted by looking at raw data in the matrix.
* **To perform regression testing:** Multiple linear regression is difficult to interpret when two independent variables in the dataset are highly correlated. Two variables which are highly correlated can easily be located using a correlation matrix, as its convenient structure helps with quick and easy detection.
* **Input for various analyses:** Analysis methods such as structural equation models can use the correlation matrix as an input for their calculation process.

## 1.2 Covariance vs correlation: What is the difference?

Covariance measures whether a variation in one variable results in a variation in another variable; for example, looking at whether an increase in one variable results in an increase, decrease, or no change in the other variable. Correlation measures the direction as well as the strength of the relationship between two variables (i.e. how strongly these two variables are related to each other).

#### 1. Relationship constraints:

Covariance deals with the linear relationship of only two variables in the dataset, whereas correlation can involve two or multiple variables or data sets and their linear relationships.

#### 2. Value range

Although both correlation coefficient and covariance are measures of linear association, correlation coefficients are standardized, therefore displaying an absolute value within a definite range from -1 to 1. On the other hand, covariance values are not standardized and use an indefinite range from -∞ to +∞ , which makes the interpretation of covariance a bit tricky.

#### 3. Measurement units

Correlation is dimensionless, i.e. it is a unit-free measure of the relationship between variables. In contrast,  covariance is in units, which is formed by multiplying the unit of one variable by the unit of another variable.

#### 4. Change in scale

Covariance is affected by the change in scale, i.e. if all the values of one variable are multiplied by a constant and all the values of another variable are multiplied by a similar or different constant, then the covariance is changed. Conversely, correlation is not affected by the change in scale.

## 1.3 How are covariance and correlation relevant to data analytics?

Statistics forms the foundation of many [data analysis methods and techniques](https://careerfoundry.com/en/blog/data-analytics/data-analysis-techniques/). Some common use cases of covariance and correlation within the field of data analytics include:

* Comparing samples from two or more different populations. This is useful because it helps in analyzing common trends and patterns in different samples.
* In data-driven industries, covariance and correlation help in identifying multivariate data in order to process data and effectively perform analytical operations.
* Correlation is a key method for investigating relations between two variables before implementing statistical modeling.
* PCA (principal component analysis) is implemented using covariance and correlation in order to shrink dimensions of large datasets to enhance interpretability. Data scientists use PCA to carry out predictive analysis and [exploratory data analysis](https://careerfoundry.com/en/blog/data-analytics/exploratory-data-analysis/).
* Analytical processes such as [multivariate analysis](https://careerfoundry.com/en/blog/data-analytics/multivariate-analysis/) and feature selection are accomplished by employing covariance and correlation methods.

**2. Multivariate data analysis techniques and examples**

There are many different techniques for multivariate analysis, and they can be divided into two categories:

* Dependence techniques
* Interdependence techniques

#### 2.1 Dependence methods

Dependence methods are used when one or some of the variables are dependent on others. Dependence looks at cause and effect; in other words, can the values of two or more independent variables be used to explain, describe, or predict the value of another, dependent variable? To give a simple example, the dependent variable of “weight” might be predicted by independent variables such as “height” and “age.”

In machine learning, dependence techniques are used to build predictive models. The analyst enters input data into the model, specifying which variables are independent and which ones are dependent—in other words, which variables they want the model to predict, and which variables they want the model to use to make those predictions.

#### 2.2 Interdependence methods

Interdependence methods are used to understand the structural makeup and underlying patterns within a dataset. In this case, no variables are dependent on others, so you’re not looking for causal relationships. Rather, interdependence methods seek to give meaning to a set of variables or to group them together in meaningful ways.

So: One is about the effect of certain variables on others, while the other is all about the structure of the dataset.

With that in mind, let’s consider some useful multivariate analysis techniques. We’ll look at:

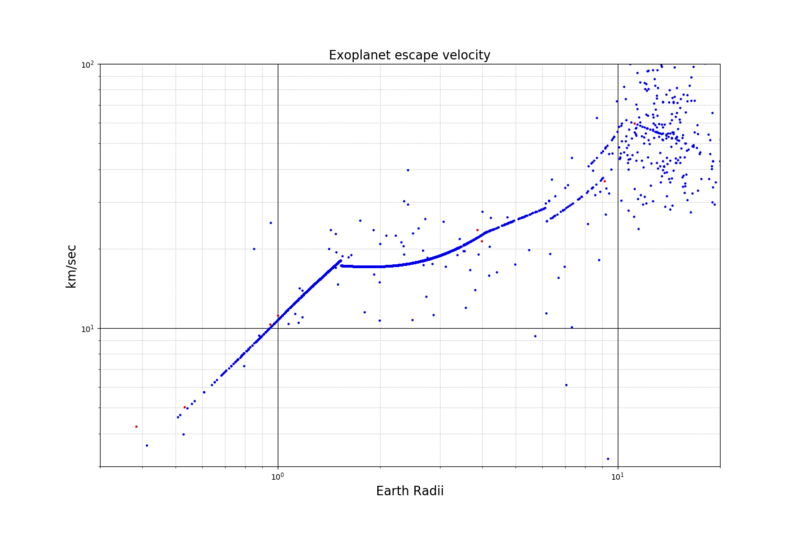
* Multiple linear regression
* Multiple logistic regression
* Multivariate analysis of variance (MANOVA)
* Factor analysis
* Cluster analysis

### 2.3 Multiple linear regression

Multiple linear regression is a dependence method which looks at the relationship between one dependent variable and two or more independent variables. A multiple regression model will tell you the extent to which each independent variable has a linear relationship with the dependent variable. This is useful as it helps you to understand which factors are likely to influence a certain outcome, allowing you to estimate future outcomes.

#### Example of multiple regression:

As a data analyst, you could use multiple regression to predict crop growth. In this example, crop growth is your dependent variable and you want to see how different factors affect it. Your independent variables could be rainfall, temperature, amount of sunlight, and amount of fertilizer added to the soil. A multiple regression model would show you the proportion of variance in crop growth that each independent variable accounts for.



### 2.4 Multiple logistic regression

Logistic regression analysis is used to calculate (and predict) the probability of a binary event occurring. A binary outcome is one where there are only two possible outcomes; either the event occurs (1) or it doesn’t (0). So, based on a set of independent variables, logistic regression can predict how likely it is that a certain scenario will arise. It is also used for classification.

#### Example of logistic regression:

Let’s imagine you work as an analyst within the insurance sector and you need to predict how likely it is that each potential customer will make a claim. You might enter a range of independent variables into your model, such as age, whether or not they have a serious health condition, their occupation, and so on. Using these variables, a logistic regression analysis will calculate the probability of the event (making a claim) occurring. Another oft-cited example is the filters used to classify email as “spam” or “not spam.”

### 2.5 Multivariate analysis of variance (MANOVA)

Multivariate analysis of variance (MANOVA) is used to measure the effect of multiple independent variables on two or more dependent variables. With MANOVA, it’s important to note that the independent variables are categorical, while the dependent variables are metric in nature. A categorical variable is a variable that belongs to a distinct category—for example, the variable “employment status” could be categorized into certain units, such as “employed full-time,” “employed part-time,” “unemployed,” and so on. A metric variable is measured quantitatively and takes on a numerical value.

In MANOVA analysis, you’re looking at various combinations of the independent variables to compare how they differ in their effects on the dependent variable.

#### Example of MANOVA:

Let’s imagine you work for an engineering company that is on a mission to build a super-fast, eco-friendly rocket. You could use MANOVA to measure the effect that various design combinations have on both the speed of the rocket and the amount of carbon dioxide it emits. In this scenario, your categorical independent variables could be:

* Engine type, categorized as E1, E2, or E3
* Material used for the rocket exterior, categorized as M1, M2, or M3
* Type of fuel used to power the rocket, categorized as F1, F2, or F3

Your metric dependent variables are speed in kilometers per hour, and carbon dioxide measured in parts per million. Using MANOVA, you’d test different combinations (e.g. E1, M1, and F1 vs. E1, M2, and F1, vs. E1, M3, and F1, and so on) to calculate the effect of all the independent variables. This should help you to find the optimal design solution for your rocket.

### 2.6 Factor analysis

Factor analysis is an interdependence technique which seeks to reduce the number of variables in a dataset. If you have too many variables, it can be difficult to find patterns in your data. At the same time, models created using datasets with too many variables are susceptible to overfitting. Overfitting is a modeling error that occurs when a model fits too closely and specifically to a certain dataset, making it less generalizable to future datasets, and thus potentially less accurate in the predictions it makes.

Factor analysis works by detecting sets of variables which correlate highly with each other. These variables may then be condensed into a single variable. Data analysts will often carry out factor analysis to prepare the data for subsequent analyses.

#### Factor analysis example:

Let’s imagine you have a dataset containing data pertaining to a person’s income, education level, and occupation. You might find a high degree of correlation among each of these variables, and thus reduce them to the single factor “socioeconomic status.” You might also have data on how happy they were with customer service, how much they like a certain product, and how likely they are to recommend the product to a friend. Each of these variables could be grouped into the single factor “customer satisfaction” (as long as they are found to correlate strongly with one another). Even though you’ve reduced several data points to just one factor, you’re not really losing any information—these factors adequately capture and represent the individual variables concerned. With your “streamlined” dataset, you’re now ready to carry out further analyses.

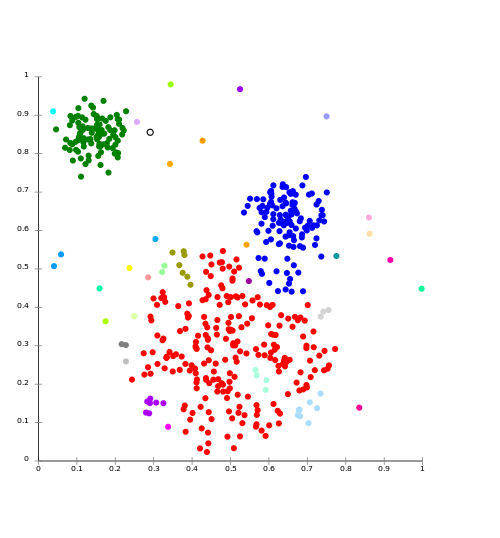
### 2.7 Cluster analysis

Another interdependence technique, cluster analysis is used to group similar items within a dataset into clusters.

When grouping data into clusters, the aim is for the variables in one cluster to be more similar to each other than they are to variables in other clusters. This is measured in terms of intra-cluster and inter-cluster distance. Intra-cluster distance looks at the distance between data points within one cluster. This should be small. Inter-cluster distance looks at the distance between data points in different clusters. This should ideally be large. Cluster analysis helps you to understand how data in your sample is distributed, and to find patterns.

#### Cluster analysis example:

A prime example of cluster analysis is audience segmentation. If you were working in marketing, you might use cluster analysis to define different customer groups which could benefit from more targeted campaigns. As a [healthcare analyst](https://careerfoundry.com/en/blog/data-analytics/what-is-a-healthcare-data-analyst/), you might use cluster analysis to explore whether certain lifestyle factors or geographical locations are associated with higher or lower cases of certain illnesses. Because it’s an interdependence technique, cluster analysis is often carried out in the early stages of data analysis.



This is just a handful of multivariate analysis techniques used by data analysts and data scientists to understand complex datasets. If you’re keen to explore further, check out

* discriminant analysis
* conjoint analysis
* canonical correlation analysis
* structural equation modeling
* multidimensional scaling.

## 3. What are the advantages of multivariate analysis?

The one major advantage of multivariate analysis is the depth of insight it provides. In exploring multiple variables, you’re painting a much more detailed picture of what’s occurring—and, as a result, the insights you uncover are much more applicable to the real world.

1. Univariate vs Multivariate Analysis: Most real-world problems involve multiple variables affecting the outcome. Univariate analysis considers one variable at a time, while multivariate analysis examines two or more variables simultaneously.
2. Complexity of Relationships: In real life, variables often interact with each other in complex ways. Multivariate analysis helps to uncover these interactions and understand the relationships between different variables.
3. Control for Confounding: Multivariate analysis allows us to control for confounding variables, which are external variables that may affect both the independent and dependent variables, leading to incorrect conclusions.
4. Increase in Predictive Power: By considering multiple variables, multivariate analysis often leads to models with better predictive power and accuracy compared to univariate models.
5. Reduction in Type I Error: Performing multiple univariate tests increases the risk of Type I error (false positives). Multivariate analysis considers all variables simultaneously, reducing this risk.
6. More Comprehensive Understanding: Multivariate analysis provides a more comprehensive understanding of the relationships between variables, which is essential for making informed decisions and creating effective strategies.
7. Real-world Applications: Multivariate analysis is widely used in various fields such as medicine, economics, psychology, and social sciences to study complex relationships and solve real-world problems.

In conclusion, while univariate analysis is useful for understanding individual variables, multivariate analysis is essential for a comprehensive understanding of the relationships between multiple variables and for creating more accurate and effective models.