

# CORRELATIONS<sub>IN</sub>

# R

A **ready-to-apply handbook** to perform correlation analyses using functions of **RSTATIX** package: Shapiro-Wilk normality test and Pearson and Spearman coefficients.

Roberto Delgado Castro  
San José, Costa Rica

---

519.5

D352v

Delgado Castro, Roberto

Correlations in R [recurso electrónico] / Roberto Delgado Castro. – primera edición – San José, Costa Rica : D. Castro R., 2023.

E-book : pdf ; 1,9 Mb – Colección y Serie (Correlations in R, no.1)

ISBN 978-9968-03-618-4

1. ESTADÍSTICA – ENSEÑANZA . 2. MATEMÁTICA ESTADÍSTICA. 3. PROBABILIDADES ESTADÍSTICAS. I. Título.

---

---

First edition 2023

Digital Edition

Layout: Aire Studio S.A.

The total or partial reproduction of the content, texts and images of this work by any means or procedure is prohibited, without the prior, express and written authorization of its author.

Any form of unauthorized use will be prosecuted in accordance with the Copyright Law.



# CORRELATIONS IN **R**

# Foreword

One of the most known and used statistical tools are correlations.

Correlations are performed to find out levels of relations between two or more objects, variables or situations, in order to extract insights that bring explanations about the reasons why such objects or variables are, or not, related between each other.

The usage of correlations, in general terms, brings along the development of quantitative coefficients. The most famous are the Pearson and Spearman coefficients. These metrics are used to quantify the correlation analysis.

The correlation analysis could be applied to find out relevant insights in daily-business tasks or projects. Furthermore, it could be applied to find explanations why determined business phenomemon occur. Bit more beyond, correlations could be used to solve problems in a wide variety of issues that appear daily in organizations, no matter their size, nature, type of operations, location or origin.

Therefore, correlation analysis is a fantastic tool that could be used in applied analytics. But, where to begin? The answer is responded within this book. A clear path or route is proposed to be followed in order to use this statistical technique to obtain the maximum benefit, especially, in common business issues.

What is the main purpose of this book? Put in your hands, respected reader, a ready-to-apply knowledge to put in practice in your daily-professional life.



**Roberto Delgado Castro**

# Acknowledgements

---

To God, who owns and governs everything.

My beautiful wife Ana... always fighting beside me without truce.

Felipe and Lucia, my wonderful kids... always giving magic to our lives.

# Contents

---

**08**

**Introduction**

**14**

**Why performing  
normality tests?**

**10**

**Correlation**

**18**

**Correlation coefficients:  
Pearson and Spearman**

**12**

**Normality  
tests**

**19**

**RSTATIX package**

# 22

## Procedure check list in R

Install RStatix package .....	23
Import input-data in R .....	23
Run Shapiro-Wilk normality test to variable 1 .....	24
Run Shapiro-Wilk normality test to variable 2 .....	25
Compare the results of Shapiro-Wilk normality tests .	27
Calculate correlation coefficients .....	28
Pearson Coefficient .....	30
Spearman Coefficient .....	31
Box plot construction: illustrate normality tests .....	32

# 34

## Closure

# 35

## References

# Introduction

---

For ages, one of the most common features of human beings is their capacity to compare objects and situations of real life.

At the very beginning, the ability to perform comparisons between objects has given humans the possibility to define dimensions of objects and levels of difficulty in problem solving tasks.

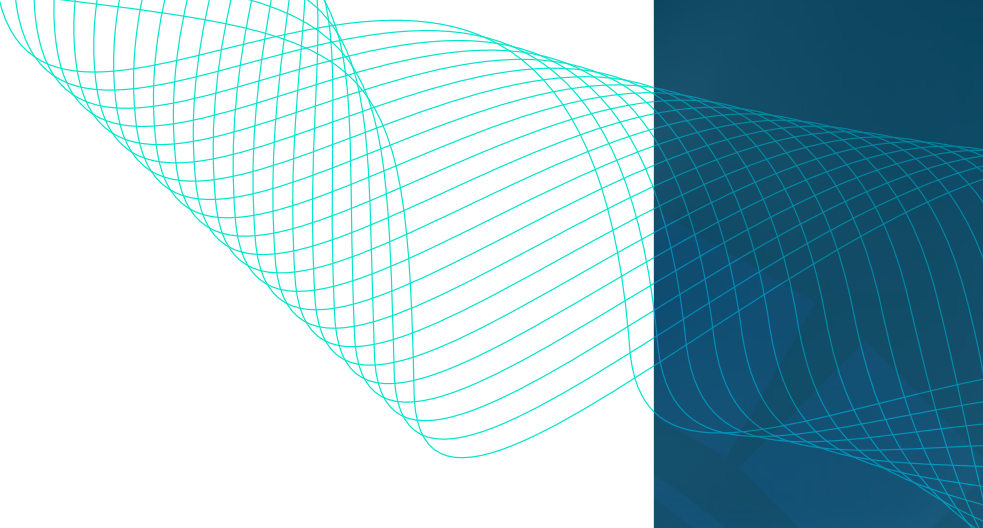
These comparisons, as time lasted, were used in quite simple trade negotiations. When humans were obligated to find objects that could help them increase their survival probability, they decided to exchange them with other individuals, based on the strategic usage and value they assigned to them. Precisely, they began to value things based on personal premises; those premises were based on comparisons. What objects have more value than other things? Their importance was defined by what object is more relevant than others.

Another important piece of analysis that has been performed for centuries, originated in correlations, is the famous cause-effect. Why certain events occur? What are the causes of determined events? Are there events that are closely related with another ones? Are there situations that occur based on another ones? These questions have been part of daily life of humans for a long time.

In addition, the mentioned cause-effect analysis is very common and is wide used in business and applied analytics. As organizations began to collect information in time to build time series data about operative processes, they discovered the necessity to explain potential relations between one serie and other ones, in order to find the reasons why determined business phenomenon occur. In other words, they needed to determine if a time series data of a certain variable is related with another one, to find out if one is the cause or the effect of the other.

As such analysis turned more and more complex, came up the necessity to transform such





examinations into numbers. And that originated the rise of correlation coefficients.

The mathematician Karl Pearson (1857) and prominent psychologist Charles Edward Spearman (1863) made stunning and impressive discoveries and developments in the field of correlations. Both of them worked with levels of relation between two variables; but much more than variables, I could mention that they developed the way to measure levels of association between two data sets or data series. Therefore, the calculation of quantitative metrics or coefficients to measure levels of association between two variables, must be a matter of attention in the field of Business Analytics. Why? Because these kind of analysis and developments bring along the possibility to open a wide world of new insights and discoveries for organizations. This handbook proposes a check list of systematic tasks intended to guide professionals within the performing of correlation analysis in R, specifically, the usage of the package RSTATIX in the following crucial functions:

1. Normality tests (Shapiro-Wilk).
2. Pearson coefficient.
3. Spearman coefficient.
4. Box plots.

Before going deep into such process, I will make a brief review of some crucial statistical concepts, in order to build a clear context.

# Correlation

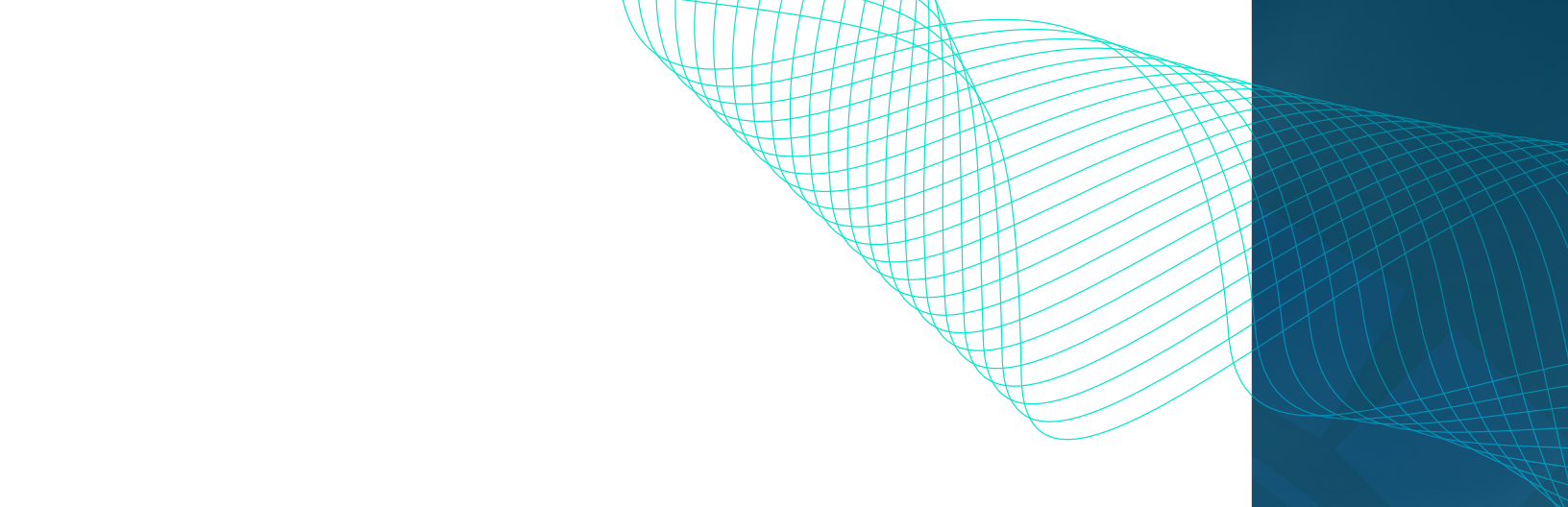
Shao, Shirvan and Alamer (2022) express that “understanding the association between theoretical constructs is at the heart of quantitative research. Researchers use correlation to understand how two or more variables are associated. The relationship between variables is usually obtained by assessing how measures/scales that represent the variables are correlated. Analysts rarely use single items to represent a complex phenomenon because single items cannot appropriately capture the complexity inherent in theoretical concepts” (1).

In addition, Prasad (2023) indicates that “correlation and regression are the techniques which are used to investigate if there is a relationship between two quantitative variables. Correlation answers three questions—is there a relationship, what is the strength of relationship and direction of the relationship? Regression expresses this relationship in a mathematical form so that the equation can be used for predicting other values. However, correlation does not deal with causation, that is even a high degree of correlation cannot be used to confirm which is the cause and which is the effect variable. However, this condition is clearly defined in “regression or connection between two or more things. A more formal definition says, correlation is a statistical method used to assess a possible association between two or more variables”.(2).

In these same ideas, Makowski, Ben-Shachar, Patil and Lüdecke (2020), point that “correlations tests are arguably one of the most commonly used statistical procedures, and are used as a basis in many applications such as exploratory data analysis, structural modelling, data engineering etc” (3).

Therefore, according with the mentioned authors, the heart of correlation is finding some levels of association between two or more variables. At this point, it is possible to beakdown its definition.

First, a statistical-quantitative method or procedure must be performed. Second, researchers must find two or more variables, that could be objects, data sets or any other phenomenon. Third, such association could be explained by the calculation of some



coefficients. Fourth, the variables under analysis, in the majority of cases, are data sets or time-series-data made of information that have been collected within time.

To support the mentioned ideas, Seeram (2019), expressed that “correlational research is a type of nonexperimental research that facilitates prediction and explanation of the relationship among variables. Researchers use a correlational research design to measure 2 or more variables to investigate the extent to which the variables are related” (4).

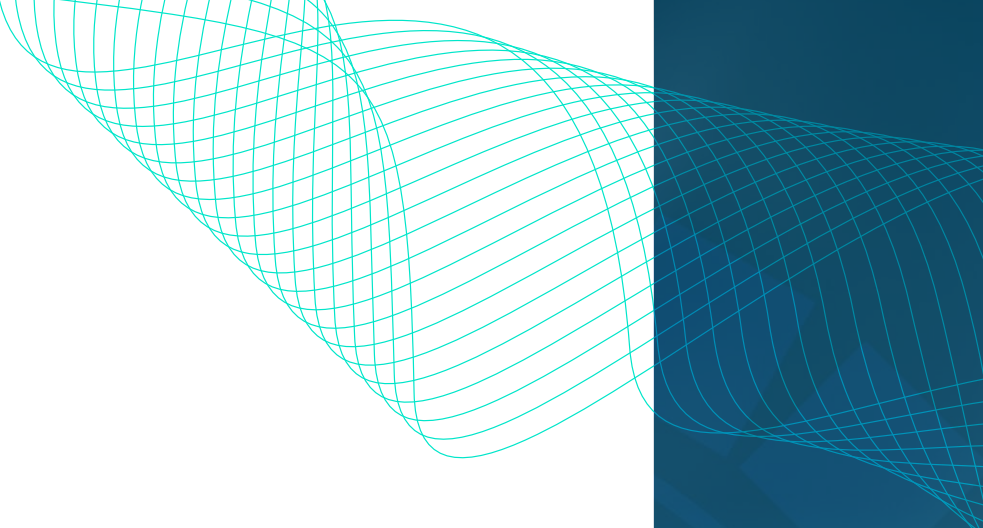
Thus, the usefulness of correlation analysis is absolutely relevant for organizations. Senthilnathan (2019), points that “many studies use correlation analysis to explore the degree association between study variables. Especially in social science research, linear correlation analysis is a tool for representing the closeness of one related variable to another. The linear correlation coefficient ( $r$  or  $R$ ) is such a measure providing information to the extent to which two variables have very close association. The purpose of carrying out correlation analysis is almost the same in every study and mostly, a correlation analysis becomes useful to explore the associative relationship between independent and dependent variables” (5).

In addition, Senthilnathan expresses that “correlation is meant for exploring the degree of relationship between two variables in consideration. Correlation coefficient is the measure to quantify such degree of relationship of the variables. Generally, two correlation coefficients are used in applications, namely: Pearson’s Product Moment Correlation Coefficient and Spearman’s Rank Correlation Coefficient” (5).

# Normality tests

Khatun (2021) expresses that “normality tests are used in different sectors” (6). Also, Das and Imon (2016), expresses that “in statistics it is conventional to assume that the observations are normal. The entire statistical framework is grounded on this assumption and if this assumption is violated the inference breaks down. For this reason it is essential to check or test this assumption before any statistical analysis of data” (7).

Also, Ahmad and Khan (2015), points that “standard statistical procedures often require data to be normally distributed and the results of these methods will be inappropriate when the assumption of normality is not satisfied. Therefore, the postulation of normality is strictly required before proceeding statistical analysis. Many parametric methods (like correlation, regression, t –test, analysis of variance etc) require normality assumption. The assumption of normality is one of the most important assumptions of parametric procedures because of its extensive range of practical applications”(8). In addition, Ahmand and Khan also pointed that “researchers developed many tests for the comparison of normality assumption in different years; some tests were modified for attaining better performance. Firstly, Pearson (1900) developed chi-square test for detection of non-normality. Kolmogrov and Smirnov (1933) suggested formal test for normality. The chi square test, based on acumulative distribution function, and can be used for any univariant distribution. After two decades, Anderson and Darling proposed their test for normality. Kuiper (1960) brought out the test of normality. Afterwards, Shapiro and Wilk (1965) suggested test of normality. In 1968, Ajne normality test was developed. After two years later, a modification of Kuiper and Ajne tests was proposed by Stephens (1970). D’Agostino (1972) introduced another test of normality. In the same year, modification of Kolmogrov Smirnov was proposed by Stephens. Four years later, Vasicek’s test of normality proposed by Vasicek (1976). Jarque and Bera designed test of normality in 1987” (8).



Therefore, Hernandez (2021) expresses that “ determining whether or not a data sample has been obtained from a normally-distributed population is a common practice in statistics and data analysis. Up to this date, several dozens of methods have been proposed in the scientific literature for testing normality” (9).

# Why performing normality tests?

According with the bibliography reviewed so far related with normality tests, from a pure statistical point of view, before applying any quantitative processing applied to a determined data set or variable, it is recommended to perform any kind of normality test, in order to determine whether or not such data set (variable) is normal or not.

Why?

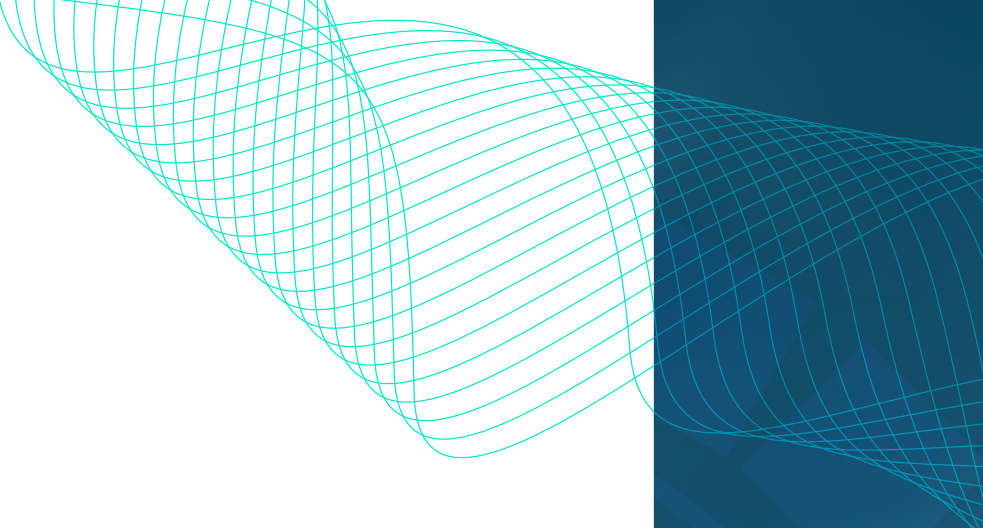
Dagnino (2014) states that “many times, when two variables that change together are measured, it is possible to determine which is independent and which is dependent. In these circumstances, it is only possible to describe the strength of association between them, since not causal predictions or estimates can be made” (17).

From this point on, the work of comparing and relating two variables constitutes one of the most used and common scientific works in the world of research and statistics. The collection of data for the construction of time series on particular variables, and the development of coefficients that quantify the association between them, are one of the cornerstones in what is known as Linear Regression.

Fiallos (2021) states that “the most common numerical index used to measure a correlation is the “Pearson Coefficient”. Pearson’s Coefficient (also called the product-momentum correlation coefficient), is represented by the symbol “r” and provides a numerical measure of the correlation between two quantitative variables. This coefficient has the following characteristics:

1. It tells us whether two variables are correlated or not.
2. It tells us whether the apparent relationship is positive or negative.
3. The coefficient sign denotes the strength or intensity of the correlation between the variables” (18).

It is particularly important to note that, if the data collected is parametric and numerical and follow a normal distribution trend, Pearson’s Coefficient should be used. On the other



hand, if you are working with data that is not parametric or follow a normal distribution trend, it is best to use, among others, the so-called Spearman's Coefficient. As Pinilla and Rico (2021) put it, "analyses based on Pearson's Coefficient are known as parametric, and those based on Spearman's Coefficient are known as non-parametric" (10). Likewise, Laguna (2014) states that "there are two correlation coefficients that are frequently used: Pearson's (parametric) and Spearman's (non-parametric, used in those cases where the variables examined do not meet normality criteria or when the variables are ordinal). Pearson's correlation coefficient specifically assesses the adequacy to the linear line that defines the relationship between two quantitative variables. Spearman's nonparametric coefficient measures any type of association, not necessarily linear" (11). In the same vein, Laguna indicates that there are certain conditions for applying both coefficients when calculating associations between variables:

**1. "Quantitative variables:** both variables examined must be quantitative. For ordinal variables, Spearman's coefficient can be used.

**2. Normality:** The normality of both variables is a requirement in the case of Pearson's correlation coefficient, but not in Spearman's.

**3. Independence:** Observations must be independent, i.e., there is only one observation of each variable for each individual” (11).

According to Palomar-Yarritu (2022), “Spearman’s correlation coefficient is a non-parametric coefficient alternative to Pearson’s correlation coefficient, when the variables, whose degree of association is to be studied, do not meet the hypothesis of normality or, simply, there is no certainty about it” (12).

Likewise, Zúñiga, Chambi, Carbajal, Meléndez, Figueroa, Viveros and Coaquira (2022), indicate that “in the Pearson [ $r$ ] (Pearson, 1896) and Spearman [ $r_s$ ] (Spearman, 1904) correlations, the procedures are related to the nature of the variables used. The first coefficient is a parametric statistic that requires prior fulfillment of several premises or assumptions, especially that of bivariate normality for the variables analyzed, while the second is a non-parametric estimator used in variables that do not necessarily meet the normality criteria” (13).

However, once the input data has been collected, as mentioned so far, it is necessary to run certain normality tests to determine whether the data follow a trend of a normal distribution, before knowing whether to calculate Pearson’s coefficient or Spearman’s coefficient for a given database or dataset.

According to Tapia and Cevallos (2021), “this test is used to contrast normality when the sample size is less than 50 observations and in large samples it is equivalent to the Kolmogorov-Smirnov test. The method consists of starting by ordering the sample from lowest to highest value, obtaining the new sample vector. Shapiro-Wilk, as a test of normality, was introduced considering that the normal probability plot that examines the fit of a sample dataset for the normal distribution is similar to the linear regression plot, the diagonal line of the graph is the perfect fit line, with the difference that this line is similar to the regression residuals” (14).

The quantitative result of the Shapiro Wilk test is given by the  $W$  statistic, which, according to Carmona and Carrión (2015), is between zero and one. “For small values of  $W$ , there is evidence of deviation from normal, i.e., normality is rejected; while the value of one indicates the normality of the data” (5). In addition, this interpretation is reinforced by Jaramillo, Pinos, Sarango, and Román (2023), who state, in a scientific work, that “as



in the Kolmogorov-Smirnov test, the p-value (0.598) is greater than a commonly used significance level, such as 0.05. This also indicates that there is not enough evidence to reject the null hypothesis of normality according to the Shapiro-Wilk test” (15).

Finally, and in addition to the previous paragraph, Sáenz, Balarezo and Yengle-Ruíz (2019), as part of the analysis of the results of a research carried out, indicate that “prior to the application of the various tests, the normality analysis was carried out, proposing the following hypotheses:  $H_0$  = The scores have a normal distribution;  $H_1$  = Scores are not normally distributed; Regarding the values compared to the significance, we will consider the following to accept or reject: If  $p < 0.05$   $H_0$  is rejected; because the scores do not have a normal distribution and if  $p > 0.05$  is not rejected,  $H_0$  i.e. the scores have a normal distribution” (16).

# Correlation coefficients: Pearson and Spearman

---

According with Schober, Boer and Schwarte (2018), “the Pearson correlation coefficient is typically used for jointly normally distributed data (data that follow a bivariate normal distribution). For nonnormally distributed continuous data, for ordinal data, or for data with relevant outliers, a Spearman rank correlation can be used as a measure of a monotonic association. Both correlation coefficients are scaled such that they range from  $-1$  to  $+1$ , where  $0$  indicates that there is no linear or monotonic association, and the relationship gets stronger and ultimately approaches a straight line (Pearson correlation) or a constantly increasing or decreasing curve (Spearman correlation) as the coefficient approaches an absolute value of  $1$ . Hypothesis tests and confidence intervals can be used to address the statistical significance of the results and to estimate the strength of the relationship in the population from which the data were sampled” (19).

Specifically in R, due to the functions of the package RSTATIX, Shapiro-Wilk normality tests and Pearson and Spearman coefficients could be calculated using a certain code structure. Within the next sections of this book, a proper code structure is proposed in order to perform such calculations.

# RSTATIX package

According with the formal prospect, “RSTATIX package is a pipe-friendly framework for basic statistical tests”. Provides a simple and intuitive pipe friendly framework, coherent with the ‘tidyverse’ design philosophy, for performing basic statistical tests, including t-test, Wilcoxon test, ANOVA, Kruskal- Wallis and correlation analyses. The output of each test is automatically transformed into a tidy data frame to facilitate visualization. The package contains helper functions for identifying univariate and multivariate outliers, assessing normality and homogeneity of variances”. (20).

RSTATIX package’s prospect states that it performs Shapiro-Wilk normality test, under the following technical criteria:

*Image 1*

*R-Statix package technical prospect*

*Shapiro-Wilk normality test*

---

<code>shapiro_test</code>	<i>Shapiro-Wilk Normality Test</i>
---------------------------	------------------------------------

---

## Description

Provides a pipe-friendly framework to performs Shapiro-Wilk test of normality. Support grouped data and multiple variables for multivariate normality tests. Wrapper around the R base function `shapiro.test()`. Can handle grouped data. Read more: [Normality Test in R](#).

## Usage

```
shapiro_test(data, ..., vars = NULL)
```

```
mshapiro_test(data)
```

## Arguments

<code>data</code>	a data frame. Columns are variables.
<code>...</code>	One or more unquoted expressions (or variable names) separated by commas. Used to select a variable of interest.
<code>vars</code>	optional character vector containing variable names. Ignored when dot vars are specified.

## Value

a data frame containing the value of the Shapiro-Wilk statistic and the corresponding p.value.

## Functions

- `shapiro_test()`: univariate Shapiro-Wilk normality test
- `mshapiro_test()`: multivariate Shapiro-Wilk normality test. This is a modified copy of the `mshapiro.test()` function of the package `mvnormtest`, for internal convenience.

## Examples

```
# Shapiro Wilk normality test for one variable
iris %>% shapiro_test(Sepal.Length)

# Shapiro Wilk normality test for two variables
iris %>% shapiro_test(Sepal.Length, Petal.Width)

# Multivariate normality test
mshapiro_test(iris[, 1:3])
```

*Source: R-STATIX package technical prospect.*

In addition, RSTATIX package also performs calculations for Pearson and Spearman coefficients, under the following technical criteria:

*Image 2*

*R-Statix package technical prospect  
Pearson and Spearman correlation coefficients*

---

`cor_test`

*Correlation Test*

---

## Description

Provides a pipe-friendly framework to perform correlation test between paired samples, using Pearson, Kendall or Spearman method. Wrapper around the function `cor.test()`.

Can also performs multiple pairwise correlation analyses between more than two variables or between two different vectors of variables. Using this function, you can also compute, for example, the correlation between one variable vs many.

## Usage

```
cor_test(  
  data,  
  ...,  
  vars = NULL,  
  vars2 = NULL,  
  alternative = "two.sided",  
  method = "pearson",  
  conf.level = 0.95,  
  use = "pairwise.complete.obs"  
)
```

*Source: R-STATIX package technical prospect.*

The values and functions are as follows:

*Image 3*

*R-Statix package technical prospect*

*Pearson and Spearman correlation coefficients: values and functions*

### **Value**

return a data frame with the following columns:

- `var1`, `var2`: the variables used in the correlation test.
- `cor`: the correlation coefficient.
- `statistic`: Test statistic used to compute the p-value.
- `p`: p-value.
- `conf.low`, `conf.high`: Lower and upper bounds on a confidence interval.
- `method`: the method used to compute the statistic.

### **Functions**

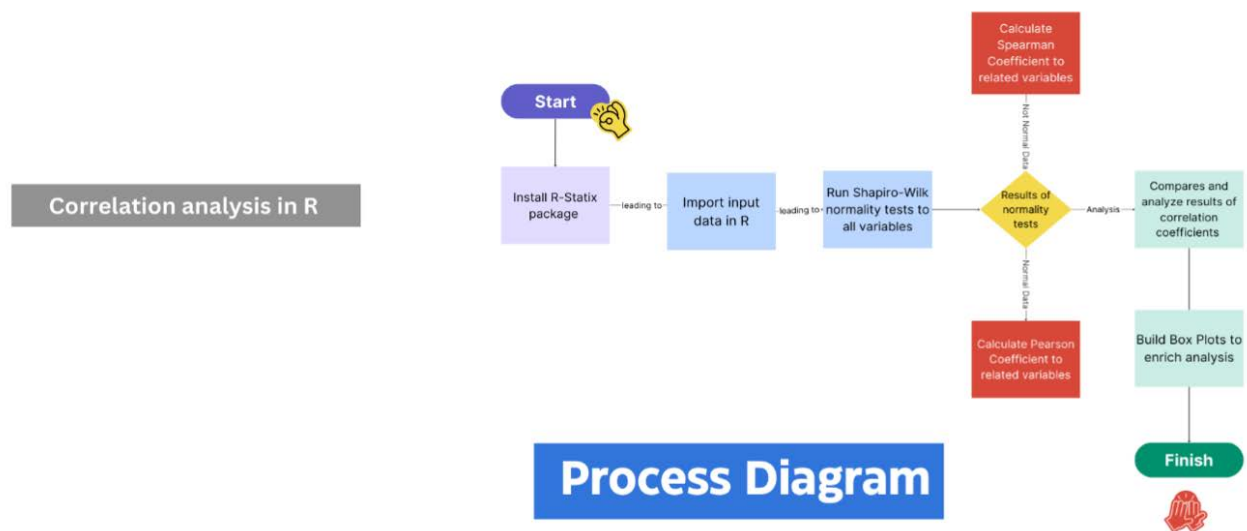
- `cor_test()`: correlation test between two or more variables.

*Source: R-STATIX package technical prospect.*

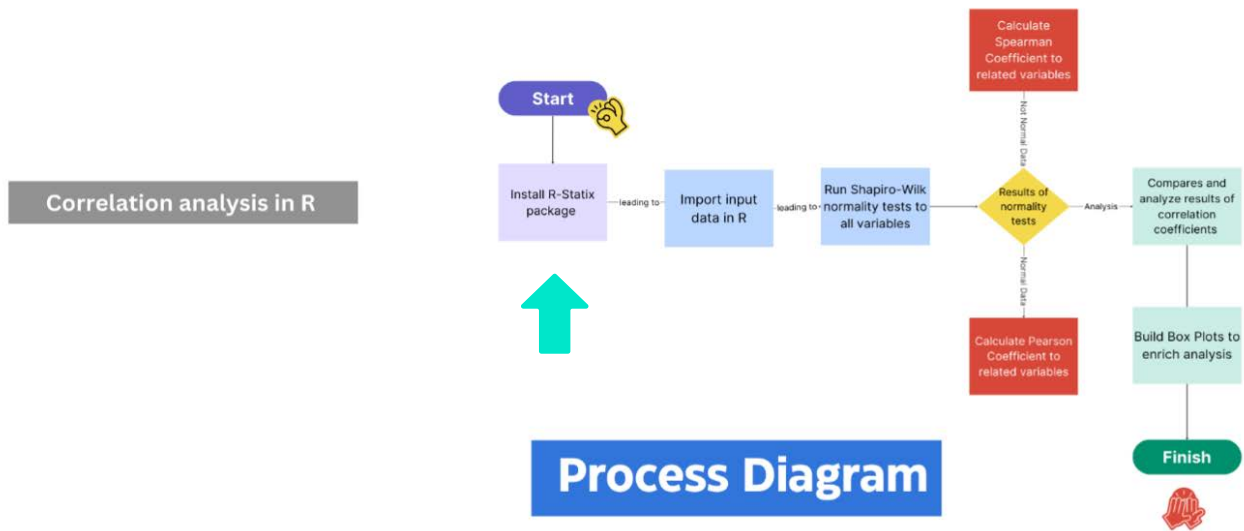
# Procedure check list in R

The following image shows the suggested process of correlation analysis:

Image 4  
Correlation analysis in R  
Recommended process diagram



## Install RStatix package

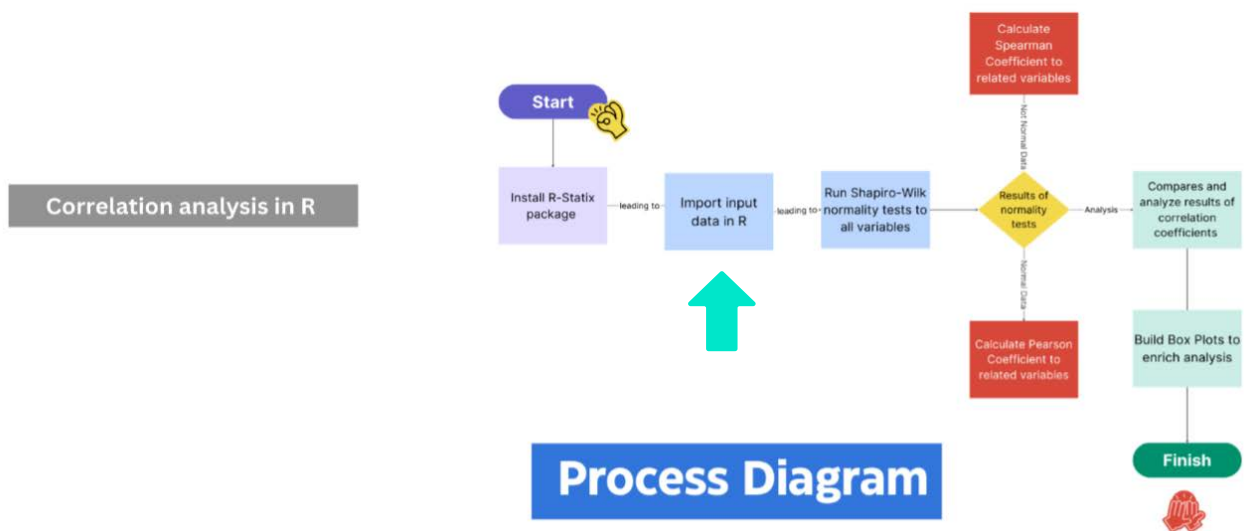


Using the LIBRARY function, the RStatix package must be installed in R prior to develop its functions.

```
library(rstatix)
```

Another way to install such a package, is due to the usage if the `install.packages()` function directly in the Console.

## Import input-data in R

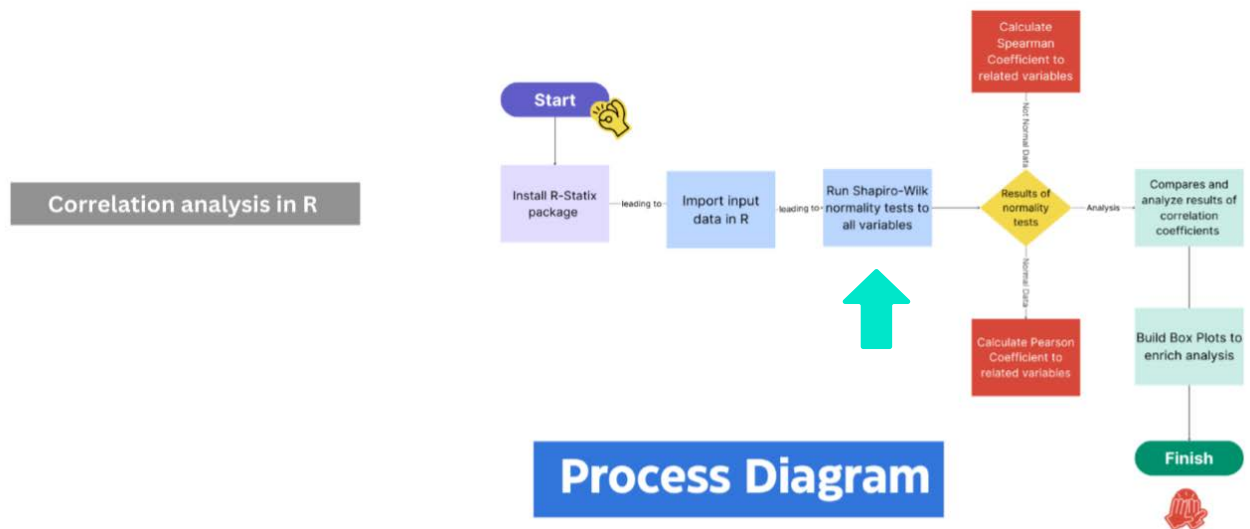


Using the **read\_delim** function, the input data set must be loaded to R:

```
## CARGA DE DATOS-INSUMO

```{r}
library(readr)
Coeficiente_Spearman <- read_delim("C:/Users/rdelgado/Desktop/ROBERTO/ROBERTO/ROBERTO LAPTOP/R
Projects/Coeficiente_Spearman/Coeficiente_Spearman.csv",
  delim = ";", escape_double = FALSE, trim_ws = TRUE)
view(Coeficiente_Spearman)
```
```

## Run Shapiro-Wilk normality test to variable 1



Using the **shapiro\_test** function, a Shapiro-Wilk normality test could be executed to variable 1 of the dataset:

```
## PRUEBA DE NORMALIDAD SHAPIRO-WILK PARA EGRESOS REALES

```{r}
Coeficiente_Spearman %>%
shapiro_test(Egresos_reales)
```
```

- In the first line of the chunk, it is necessary to invoke the name of the dataset.
- Note that the name of the variable must be typed inside the **shapiro\_test** function.



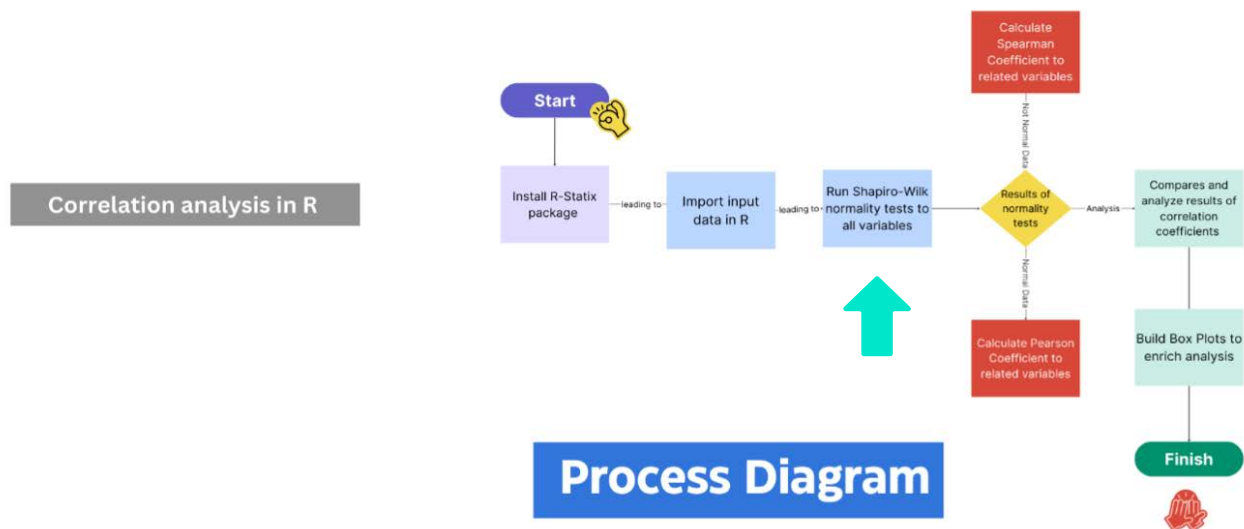
The values returned are the **statistic value** and the **p-value**:

| variable<br><chr> | statistic<br><dbl> | p<br><dbl> |
|-------------------|--------------------|------------|
| Egresos_reales    | 0.8933816          | 0.05269821 |

1 row

- The “p-value” refers to the result of the Shapiro-Wilk test.
- Generally, in order to evaluate the result of the Shapiro-Wilk test, a Level of Significance should be defined in 0.05. This value is usually called as “ $\alpha$  value”.
- If “p-value” is bigger than the “ $\alpha$  value”, then the data is considered as Normal and Parametric. Quite the opposite, if “p-value” is smaller than the “ $\alpha$  value”, then the data is considered as Not Normal or Not Parametric.

## Run Shapiro-Wilk normality test to variable 2



Using the **shapiro\_test** function, a Shapiro-Wilk normality test could be executed to variable 1 of the dataset:

```
## PRUEBA DE NORMALIDAD SHAPIRO-WILK PARA PRESUPUESTO NACIONAL
```{r}
Coeficiente_Spearman %>%
shapiro_test(Presupuesto_Nacional)
```
```

- In the first line of the chunk, it is necessary to invoke the name of the dataset.
- Note that the name of the variable must be typed inside the **shapiro\_test** function.

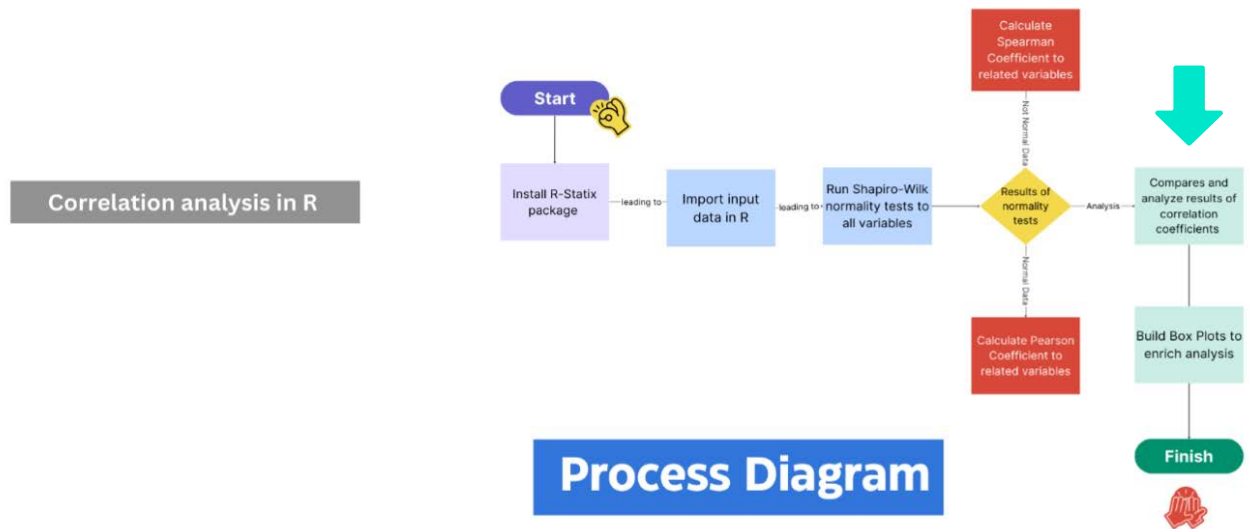
The values returned are the **statistic** value and the **p-value**:

| variable<br><chr>    | statistic<br><dbl> | p<br><dbl>   |
|----------------------|--------------------|--------------|
| Presupuesto_Nacional | 0.7664908          | 0.0007298568 |

1 row

- The “p-value” refers to the result of the Shapiro-Wilk test.
- Generally, in order to evaluate the result of the Shapiro-Wilk test, a Level of Significance should be defined in 0.05. This value is usually called as “ $\alpha$  value”.
- If “p-value” is bigger than the “ $\alpha$  value”, then the data is considered as Normal and Parametric. Quite the opposite, if “p-value” is smaller than the “ $\alpha$  value”, then the data is considered as Not Normal or Not Parametric.

# Compare the results of Shapiro-Wilk normality tests



In order to analyze the results of the Shapiro-Wilk normality tests for each variable, in order to determine which of the correlation coefficient (Pearson or Spearman) to calculate, it is strongly recommended to build a simple data frame, as follows:

Table 1  
Shapiro-Wilk tests results comparison analysis

| Variable | “p-value”<br>Shapiro-Wilk | “α value”<br>(level of<br>significance) | Assessment<br>of “p” with<br>respect to “α” | Shapiro-<br>Wilk <b>result<br/>criteria</b><br>(normal or not<br>normal) | Type of data<br>(parametric<br>or not<br>parametric) | Type of<br>correlation<br>coefficient to<br>calculate<br>(Pearson or<br>Spearman) |
|----------|---------------------------|---|---|--|--|---|
|          |                           |   |   |  |  |   |

## Calculate correlation coefficients

Once the researcher has obtained the results of Shapiro-Wilk normality tests (for each variable) and has assessed them with the level of significance defined previously, he knows clearly which of the correlation coefficients calculate: Pearson or Spearman. The criteria is as follows:

- If “p-value” is bigger than the “ $\alpha$  value”, then the data is considered as Normal and Parametric; the Pearson coefficient must be calculated to perform association analysis.
- If “p-value” is smaller than the “ $\alpha$  value”, then the data is considered as Not Normal or Not Parametric; the Spearman coefficient must be calculated to perform association analysis.



## Pearson Coefficient

The calculation of Pearson coefficient could be performed at the same time R executes a scatter plot. In this specific case, researcher could use the **cor.method** function to invoke the calculation of the coefficient, inside the code in which the plot is designed:

```
## COEFICEINTE DE PEARSON PARA VARIABLES DE EGRESOS REALES Y RECARGO DE PLANILLAS
```{r}
ggscatter(data = coeficiente_spearman, x = "Recargo_planillas", y = "Egresos_reales",
  add = "reg.line", conf.int = TRUE,
  cor.coef = TRUE, cor.method = "pearson",
  xlab = "Recargo_planillas", ylab = "Egresos_reales") +
  scale_y_continuous(labels = label_number()) +
  scale_x_continuous(labels = label_number())
```
```

- It is important to define the x and y variables. The x variable is the one that is independent; the y variable is the one that is dependant. Researcher must define clearly this structure, in order to let R execute the calculation of the coefficient properly.
- Researcher could change scales of the values of x and y axis. Due to the usage of **scale\_y\_continuous** and **scale\_x\_continuous** functions, the numbers of each labels could vary from scientific notation to common natural numbers.

Another way to calculate Pearson coefficient is through the **cor.test** function. In this case, variables x and y must be defined into a determined chunk as separate vectors:

```
```{r}
x <- c()
y <- c()

cor.test(x, y, method = "pearson", alternative = "greater")
```
```

The result of calculations is as follows:

```

Pearson's product-moment correlation

data:  x and y
t = 7.1812, df = 15, p-value = 1.585e-06
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
 0.7337707 1.0000000
sample estimates:
      cor
0.8801547
```

## Spearman Coefficient

The calculation of Spearman coefficient could be performed at the same time *R* executes a scatter plot. In this specific case, researcher could use the **cor.method** function to invoke the calculation of the coefficient, inside the code in which the plot is designed:

```
## COEFICIENTE DE SPEARMAN PARA LAS VARIABLES DE EGRESOS REALES Y PRESUPUESTO NACIONAL
```{r}
ggscatter(data = Coeficiente_Spearman, x = "Presupuesto_Nacional", y = "Egresos_reales",
  add = "reg.line", conf.int = TRUE,
  cor.coef = TRUE, cor.method = "spearman",
  xlab = "Presupuesto_Nacional", ylab = "Egresos_reales") +
  scale_y_continuous(labels = label_number()) +
  scale_x_continuous(labels = label_number())
```
```

- It is important to define the x and y variables. The x variable is the one that is independent; the y variable is the one that is dependant. Researcher must define clearly this structure, in order to let *R* execute the calculation of the coefficient properly.
- Researcher could change scales of the values of x and y axis. Due to the usage of **scale\_y\_continuous** and **scale\_x\_continuous** functions, the numbers of each labels could vary from scientific notation to common natural numbers.

Another way to calculate Spearman coefficient is through the **cor.test** function. In this case, variables x and y must be defined into a determined chunk as separate vectors:

```
```{r}
x <- c()
y <- c()

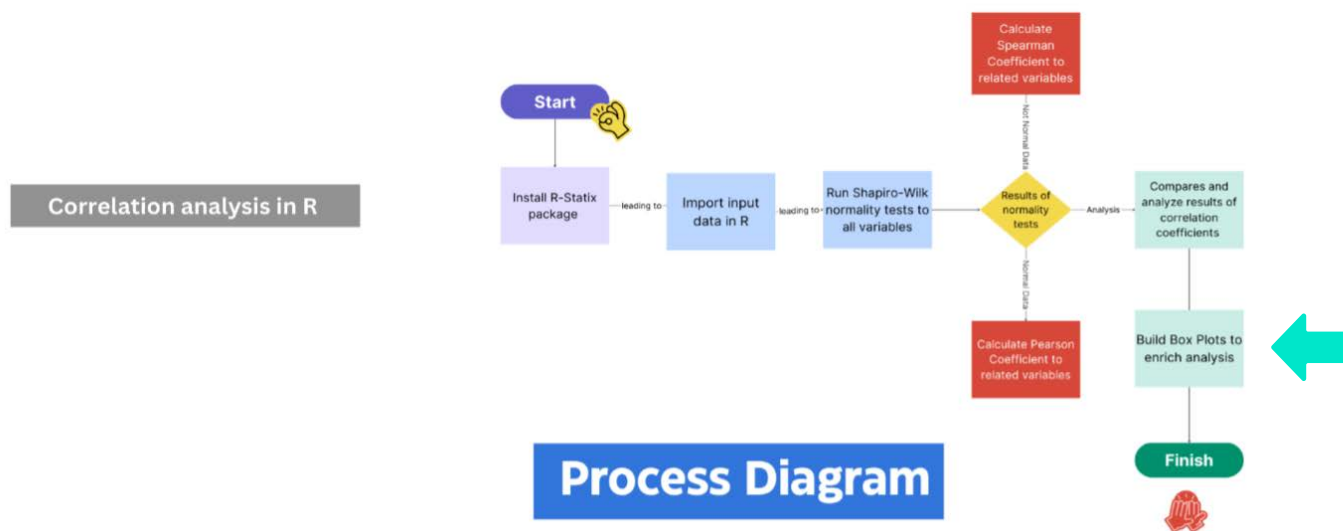
cor.test(x, y, method = "spearman", alternative = "greater")
```
```

The result of calculations is as follows:

```
Warning: Cannot compute exact p-value with ties
Spearman's rank correlation rho

data: x and y
S = 82.199, p-value = 4.561e-07
alternative hypothesis: true rho is greater than 0
sample estimates:
  rho
0.8992656
```

## Box plot construction: illustrate normality tests



Box plots are fantastic visualizations to show and illustrate the median and atypical data of a determined data set, through its quartiles.

Those plots are perfect visual tools to complement the analysis performed in the calculation of the Shapiro-Wilk normality tests.

The **box\_plot()** function takes into consideration all numbers of a vector, drawing a boxplot for each vector.

```
#GRAFICO DE CAJAS PARA LAS TRES VARIABLES JUNTAS
```{r}
x <- Coeficiente_spearman
boxplot(Coeficiente_spearman, col = rgb(0, 0.5, 1, alpha = 0.5, )) +
  scale_y_continuous(labels = label_number())
```
```





# Closure

---

The use of statistics provides real data on complex situations rather than making decisions based on assumptions. A manager must have the ability to look at data and make predictions about the future of the company or a particular department. Nowadays, every decision must be supported by concrete empirical data.

Correlations could be applied in business analytics. Some of its uses could be:

- Data exploratory analysis.
- Predictions and data modeling.

In addition, correlation is a powerful statistical tool that helps researchers to discover associations and relations between variables. When they examine the strength and the association between two or more variables, the correlation analysis let them obtain crucial information about behaviors, preferences, tendencies of processes and markets.

In order to decrease the probability of getting lost in correlation universe, with the purpose of focus on what I consider relevant and to improve the quality and impact of the research executed, the process explained in this book would help you, my appreciated reader, in your way to explore this technical topic and to obtain as much benefit as possible. The contents and knowledge of it is ready to be extracted and applied in daily business tasks. Cheers!!

# References

1. Shao, K., Elahi Shirvan, M., & Alamer, A. (2022). How accurate is your correlation? Different methods derive different results and different interpretations. *Frontiers in Psychology*, 13, 901412.
2. Prasad, S. (2023). Correlation and Regression. In *Elementary Statistical Methods* (pp. 241-279). Singapore: Springer Nature Singapore.
3. Makowski, D., Ben-Shachar, M. S., Patil, I., & Lüdtke, D. (2020). Methods and algorithms for correlation analysis in R. *Journal of Open Source Software*, 5(51), 2306.
4. Seeram, E. (2019). An overview of correlational research. *Radiologic technology*, 91(2), 176-179.
5. Senthilnathan, S. (2019). Usefulness of correlation analysis. Available at SSRN 3416918.
6. Khatun, N. (2021). Applications of normality test in statistical analysis. *Open Journal of Statistics*, 11(01), 113.
7. Das, K. R., & Imon, A. H. M. R. (2016). A brief review of tests for normality. *American Journal of Theoretical and Applied Statistics*, 5(1), 5-12.

8. Ahmad, F., & Khan, R. A. (2015). A power comparison of various normality tests. *Pakistan Journal of Statistics and Operation Research*, 331-345.
9. Hernandez, H. (2021). Testing for normality: What is the best method. *ForsChem Research Reports*, 6, 2021-05.
10. Pinilla, J. O., & Rico, A. F. O. (2021). ¿Pearson y Spearman, coeficientes intercambiables? *Comunicaciones en Estadística*, 14(1), 53-63.
11. Laguna, C. (2014). *Correlación y regresión lineal*. Instituto Aragonés de Ciencias de la Salud, 4.
12. Palomar Yarritu, I. (2022). *Diseño de gráficos de control no paramétricos para el coeficiente de correlación de Spearman*. Trabajo fin de Master. Universitat Politècnica de Valencia.
13. Apaza Zúñiga, E., Cazorla Chambi, S., Condori Carbajal, C., Arpasi Meléndez, F. R., Tumi Figueroa, I., Yana Viveros, W., & Quispe Coaquira, J. E. (2022). La Correlación de Pearson o de Spearman en caracteres físicos y textiles de la fibra de alpacas. *Revista de Investigaciones Veterinarias del Perú*, 33(3).
14. Tapia, c. e. f., & Cevallos, k. l. f. (2021). Pruebas para comprobar la normalidad de datos en procesos productivos: Anderson-Darling, Ryan-Joiner, Shapiro-Wilk y Kolmogórov-Smirnov. *Societas*, 23(2), 83-106.
15. Jaramillo, H. A. L., Pinos, C. A. E., Sarango, A. F. H., & Román, H. D. O. (2023). Histograma y distribución normal: Shapiro-Wilk y Kolmogorov Smirnov aplicado en SPSS: Histogram and normal distribution: Shapiro-Wilk and Kolmogorov Smirnov applied in SPSS. *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades*, 4(4), 596-607.

16. Tufinio Sáenz, B. B., Silva Balarezo, M. G., & Yengle Ruíz, C. (2019). Estrategia "lectura de objetos" para el desarrollo de competencia construye interpretaciones históricas. *Fides et Ratio-Revista de Difusión cultural y científica de la Universidad La Salle en Bolivia*, 17(17), 61-82.
17. Dagnino, J. (2014). Coeficiente de correlación lineal de Pearson. *Chil Anest*, 43(1), 150-153.
18. Fiallos, G. (2021). La Correlación de Pearson y el proceso de regresión por el Método de Mínimos Cuadrados. *Ciencia Latina Revista Científica Multidisciplinar*, 5(3), 2491-2509.
19. Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5), 1763-1768.
20. R-Statix package prospect. Updated February 1st, 2023. URL: <https://cran.r-project.org/web/packages/rstatix/index.html>

# CORRELATIONS

IN **R**

Roberto Delgado Castro  
San José, Costa Rica