Chapter 12
Correlation Tests in R: Pearson Cor, Kendall's Tau, and Spearman's Rho

12.1 Introduction

Correlation (Cor) is a statistical procedure or method used by researchers or the data analysts to evaluate the *strength* or *degree* of relationship between two variables (continuous or categorical) (Privitera, 2023; Schober & Schwarte, 2018). Statistically, the correlation test can be defined as a "bivariate analysis" that measures the strength of association or relationship between two variables or datasets and the direction of the relationship (see Chap. 6, Sect. 6.2.9). The result of the test (usually for linearity or strength of association) between the datasets or data points (depending on the type of correlation method being used or applied and usually determined through the *p*-values: where $p \leq 0.05$) means that a high correlation statistics indicates that the variables or data being measured have a strong relationship between each other. On the other hand, a weak correlation ($p > 0.05$) signifies that the variables are barely (insignificantly) related or associated.

Thus, with correlated datasets, it is assumed that a change in the magnitude of one variable is statistically associated with a change in the magnitude of another variable that it is being measured against, be it in the same direction (positive correlation) or in the opposite direction (negative correlation) (Akoglu, 2018; Privitera, 2023; Schober & Schwarte, 2018).

According to Akoglu (2018), the correlation (relationship, association) between the two specified variables is denoted by the letter *r* and quantified through a number, that varies between -1 and $+1$ (denoting the negative and positive correlations, respectively). Whereby, a value of zero (0) implies that there is no correlation between the variables, and a value of one (1) denotes an absolute (perfect) correlation. Therefore, whereas *r* represents the direction of the correlation, a positive *r* signifies that the measured variables are *certainly* (positively) related, while a negative *r* signifies that the measured variables are *inversely* (negatively) related. Statistically, the strength of the correlation increases both from 0 to $+1$, and from 0 to -1, respectively (Akoglu, 2018).

K. Okoye and S. Hosseini, *R Programming*,
https://doi.org/10.1007/978-981-97-3385-9_12

There are three main types of correlation analysis commonly applied by the researchers, in theory. These are (i) Pearson product–moment correlation, (ii) Kendall's tau correlation, and (iii) Spearman's rho correlation (Akoglu, 2018; Brossart et al., 2018; Hauke & Kossowski, 2011; Puth et al., 2014; Schober & Schwarte, 2018; Wang et al., 2019; Zar, 2014).

Pearson correlation (also known as Pearson product–moment correlation coefficient) is described as a parametric test that measures the strength of linear association (linear trend) that exists between two continuous variables. Statistically, the method (Pearson correlation, denoted by r) draws a "line of best fit" through the two datasets or variables by establishing how far away the two data points are to the drawn line (model) of best fit.

Mathematically, to apply the Pearson's statistics by measuring the two quantities or variables X and Y on each of N individuals in order to produce a data set of X_1, Y_1, …, X_N, Y_N (Puth et al., 2014), the formula to calculate the correlation coefficient is given as:

$$\text{Cor}\,(r) = \frac{N\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[N\sum x^2 - \left(\sum x\right)^2\right]\left[N\sum y^2 - \left(\sum y\right)^2\right]}}$$

whereby

N	the number of pairs of scores
Σxy	the sum of the products of paired scores
Σx	the sum of x scores
Σy	the sum of y scores
Σx^2	the sum of squared x scores
Σy^2	the sum of squared y scores

Just like many of the other existing types of *parametric* procedures or statistical methods (see Chap. 4), the Pearson's product–moment correlation coefficient requires the assumption that the relationship between the variables is *linear* and is measured on an interval (continuous) scale. Thus, the researchers or data analysts must check that the following below assumptions are met before applying or using the Pearson correlation.

Pearson's Correlation Assumptions

- Independence: the drawn dataset or sample must be independent to each other.
- Linearity: the two tested variables should be linearly related to each other, e.g., when plotted in a graph should result in a moderately straight line.
- Normality: the dataset must be normally distributed, i.e., should produce a bell-shaped graph when the means of the samples are plotted.
- Homoscedasticity or equality of variances must be present.

Furthermore, on the other hand, *Kendall's tau* correlation (also known as Kendall rank correlation coefficient) is a non-parametric test (i.e., an alternative to Pearson's

correlation) mainly used by the researchers to measure the *strength of dependence* between two *categorical* or *ordinal* variables. According to Couso et al. (2018), the method (Kendall's tau) can be applied as an efficient and robust way of identifying monotone relationships between two data sequences, although when applied to digital data (e.g., discrete or discontinuous format), the high number of ties could produce inconsistent results due to quantization.

Theoretically, the Kendall's tau (τ) statistics symbolizes the degree of agreement between two specified "ordinal" variables by indicating how similarly the two variables order a set of individuals or data points (Brossart et al., 2018). Thus, mathematically, the following formula is used to calculate the value of Kendall's tau statistics or rank correlation coefficient:

$$\text{Kendall's tau } (\tau) = \frac{C - D}{C + D} \text{ or } \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

whereby

n_c number of concordant, i.e., ordered in the same way.
n_d Number of discordant, i.e., ordered differently.

With the Kendall's tau statistic, commonly calculated through pairwise comparison; a value of $\tau_{(X,Y)} = +1$ means that the data points for the two (ordinal) variables (X and Y) are ordered in exactly the same way, i.e., occupies the same rank position. While on the other hand, a value of $\tau_{(X,Y)} = -1$ implies that the data points for the two variables are ordered in exactly the opposite way, with one data point occupying the first rank in one variable and the last rank in the other variable. Accordingly, a value of $\tau_{(X,Y)} = 0$ indicates that there is no relationship in the way or order that the two variables are ranked considering the data points, thus, are independent (Brossart et al., 2018).

In the same vein or similar manner, just like the Kendall's tau correlation, *Spearman's rho* correlation (also known as Spearman rank correlation coefficient) is another type of non-parametric (i.e., alternative to Pearson correlation) test used by the researchers to measure the degree of association between two (ordinal) variables. The method can also be applied for interval or ratio datasets provided the datasets are found to be distribution-free. Mathematically, the following formula is used to calculate the value of the Spearman's rho statistics or rank correlation coefficient:

$$\text{Spearman's rho } (\rho) = 1 - \frac{6\sum(d_i^2)}{n(n^2-1)}$$

whereby

n number of data points of the two variables (x and y).
d_i rank difference of element "n", i.e., difference between the corresponding statistics of order of $x - y$.

The only difference between the *Spearman's rho* versus *Kendall's tau* method is that while the Spearman's rho (ρ) statistics or results are calculated through the "ordinary least squares", the Kendall's tau (τ) statistics is calculated through the "pairwise comparison" of all the data points (Brossart et al., 2018). Thus, whilst the Kendall's tau (τ) statistics are based on "concordant and discordant pairs", the Spearman's rho (ρ) statistics are based on "deviations".

It is also noteworthy to mention that Spearman's rho (ρ) method is much more sensitive to error and handling discrepancies in data samples than the Kendall's tau (τ) method, which, on the other hand, are more accurate with smaller sample sizes than the Spearman's rho (ρ).

In any case, a lot of the time the interpretations of the two methods (Kendall's tau and Spearman's rho) are very similar, thus, tend to invariably lead to the same inferences or statistical results.

Also, unlike Pearson correlation, both methods (Kendall's tau and Spearman's rho) do not require the available data or sample to meet the assumption that the relationship between the considered variables is linear (i.e., when plotted does not necessarily need to result in a moderately straight line), or normally distributed (i.e., distribution-free), nor does it require the measurement scale of the variables to be represented on a continuous or interval scale.

Table 12.1 is a summary of the differences and similarities between the Pearson cor, Kendall's tau, and Spearman's rho Correlation tests including the conditions that are required to perform the different tests, which the authors will be demonstrating using R in the next sections (Sect. 12.2 and 12.3) of this chapter.

In the next sections of this chapter (Sects. 12.2 and 12.3), the authors will be demonstrating to the readers how to conduct the Pearson cor, Kendall's tau, and Spearman's rho correlation tests in R, harmoniously. We will illustrate the different steps to performing the three types of tests in R using the following steps outlined in Fig. 12.1.

12.2 Pearson Correlation Test in R

Pearson correlation measures the strength of linear association (correlation) that exists between two "continuous" variables. Thus, it calculates the effect of change (be it positive or negative) in one variable when the other variable changes.

By default, the hypothesis for testing whether there is a *correlation* (measure of linearity or association) between the two given set of (continuous) variables is; *IF* the p-value of the test is less than or equal to 0.05 ($p \leq 0.05$), *THEN* we assume that there is a statistically significant strong relationship between the two analyzed variables and that this is not by chance (H_1). *ELSE IF* the p-value is greater than 0.05 ($p > 0.05$) *THEN* we can conclude that there is no significant relationship between the two variables, and any observed association could only have occurred by chance (H_0).

Table 12.1 Differences and similarities between the Pearsoncor, Kendall's tau, and Spearman's rho correlation tests and assumptions

Pearson	Kendall's tau	Spearman's rho
Data sample should be independently drawn from the population	Data sample should be independently drawn from the population	Data sample should be independently drawn from the population
Used for continuous (intervalor ratio) datasets	Used for categorical (ranked or ordinal) datasets. Although can also be applied to intervalor ratio datasets	Used for categorical (ranked or ordinal) datasets. Although can also be applied to intervalor ratio datasets
Data sample or observations must be normally distributed, i.e., bell-shaped	Data samples are distribution-free, thus, are not normally distributed, i.e., skewed	Data samples are distribution-free, thus, are not normally distributed, i.e., skewed
Calculated by measuring the "average weight" of the two variables (i.e., covariance of the two variables divided by the product of their standard deviations)	Calculated through the "pairwise comparison" of the data points based on concordant and discordant pairs	Calculated through "ordinary least squares" based on deviations
Described as parametric test for linearity or relationship between two variables	Non-parametric test for strength of dependence between two variables	Non-parametric test to measure the degree of association between two variables

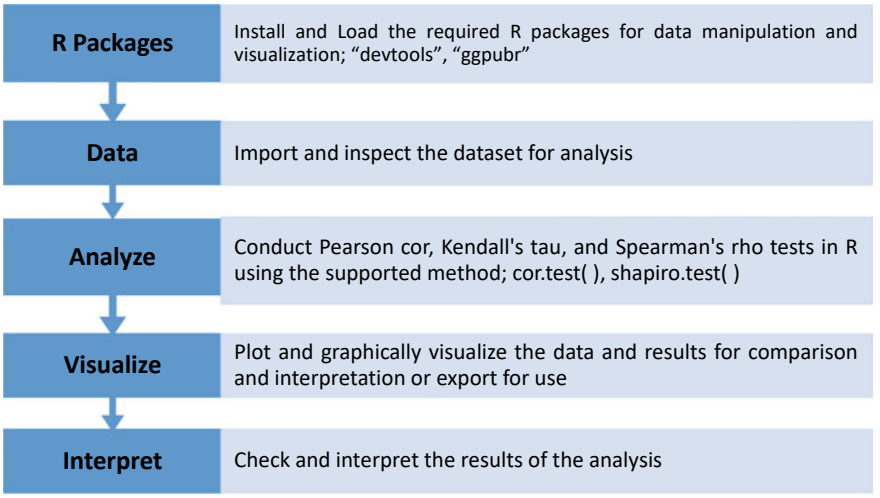

Fig. 12.1 Steps to conducting the Pearson cor, Kendall's tau, and Spearman's rho correlation tests in R

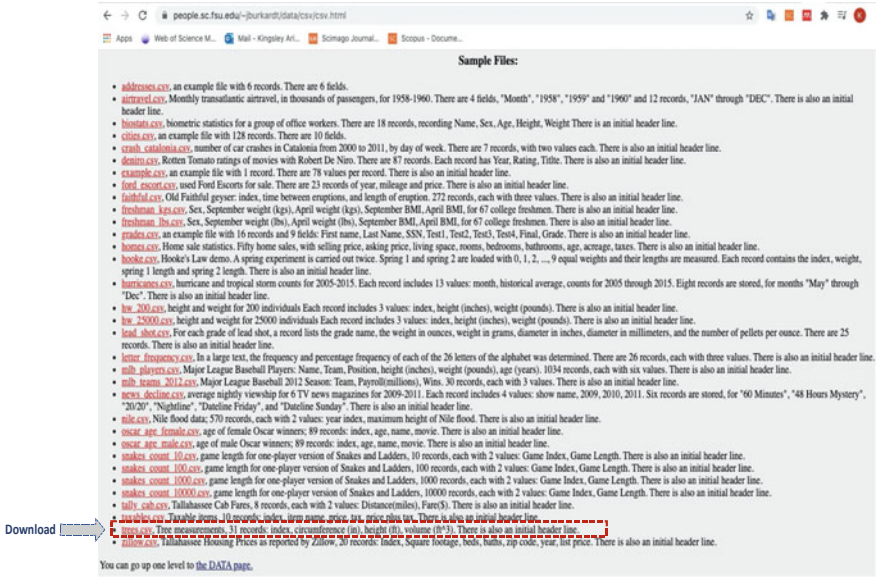

Fig. 12.2 Example of CSV file download (*Source* https://people.sc.fsu.edu/~jburkardt/data/csv/csv.html)

Here, the authors will demonstrate to the readers how to conduct the Pearson correlation test in R using the **cor.test( )** function in R. We will do this using the steps outlined in Fig. 12.1.

To begin, **Open RStudio** and **Create a new or Open an existing project**. Once the user has the RStudio and an R Project opened, **Create a new R Script** and name it "**PearsonCorrDemo**" or any name the user may preferentially choose (see Chap. 1 and 2 if the user needs to refresh on how to do this step).

Now, we are going to download an example file or dataset that we will use to demonstrate the Pearson correlation test (the users are welcome to use any dataset or format if they wish to do so).

As shown in Fig. 12.2, download the example CSV dataset named "**trees.csv**" via the following source: https://people.sc.fsu.edu/~jburkardt/data/csv/csv.html and save the file on the users' local machine or computer. *** The users can also access the list of example datasets used in this book at the following repository (https://doi.org/10.6084/m9.figshare.24728073) to download the example CSV file.

Once the user has successfully downloaded and saved the example file (trees.csv) on the computer, we can proceed to conduct the Pearson Correlation test in R.

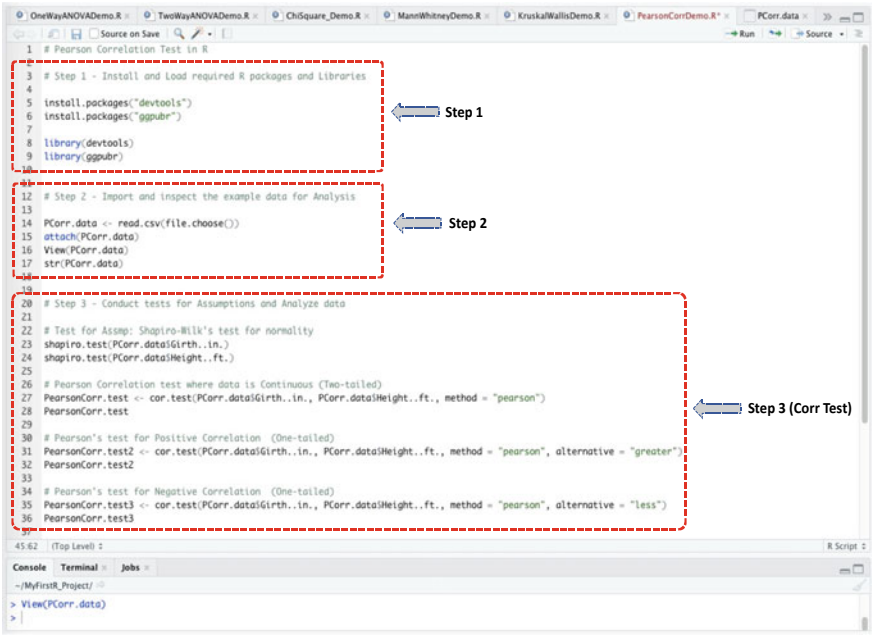

Fig. 12.3 Steps to conducting Pearson correlation test in R

Step 1—Install and Load the Required R Packages and Libraries

Install and **Load** the following *R packages* and *libraries* (see Fig. 12.3, Step 1, Lines 3–9) that will be used to call the different R functions, data manipulations, and graphical visualizations for the Pearson Correlation test.

The syntax and code to install and load the required R packages and libraries are as follows:

```
install.packages("devtools")
install.packages("ggpubr")

library(devtools)
library(ggpubr)
```

Step 2—Import and Inspect the Example Dataset for Pearson Correlation Analysis

As illustrated in Step 2 in Fig. 12.3 (Lines 12–17), import the dataset named "**trees.csv**" that we downloaded earlier and store this as an R object named "**PCorr.data**" in R (the users are welcome to use any name they may choose if they wish to do so).

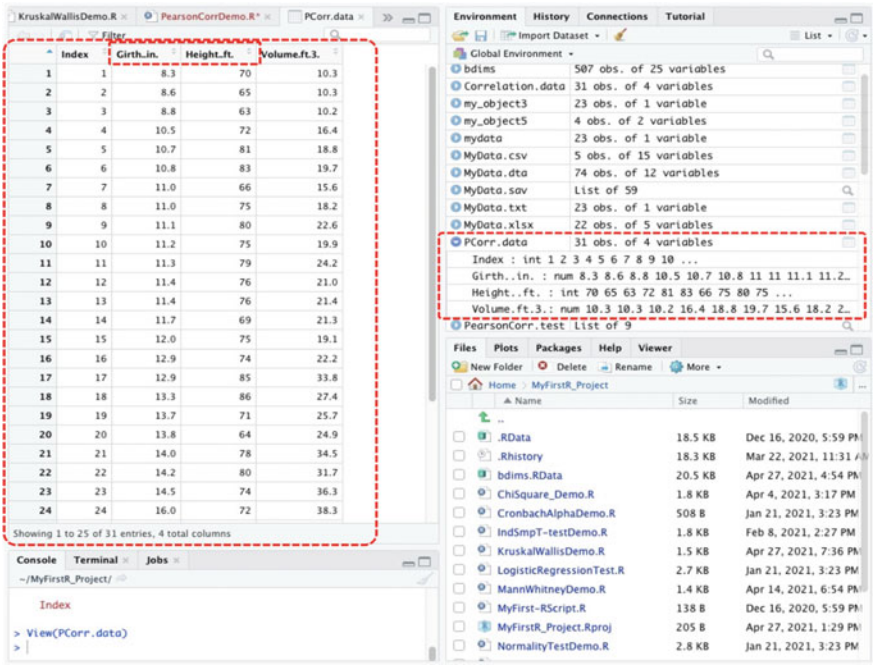

Fig. 12.4 Example of CSV dataset imported and stored as R object in R

Once the user has successfully imported the dataset, you will be able to view the details of the **trees.csv** dataset as shown in Fig. 12.4 with 31 observations and 4 variables in the data sample.

The syntax and code to import and save the data in R is shown below:

```
PCorr.data <- read.csv(file.choose())
attach(PCorr.data)
View(PCorr.data)
str(PCorr.data)
```

Step 3—Conduct Tests for Assumptions and Analyze Data

Now that we have successfully imported the example dataset and stored this in an R object we called "**PCorr.data**", we can proceed to analyze the data.

As defined in Fig. 12.3 (Step 3, Lines 20–36), first we will conduct the tests of assumptions (data normality) (see: Lines 22–24) as discussed earlier in Sect. 12.1 by using the **shapiro.test( )** method, and then perform the **Pearson Correlation** test if all the necessary conditions to conduct the test are met using the **cor.test( )** function in R, respectively (see: Fig. 12.3, Step 3, Lines 26–36).

Also, as defined earlier in the Introduction section (Sect. 12.1);

- **Pearson's correlation** statistics checks whether there exists a *linear relationship* between two independently sampled variables or data.
- The targeted variables must be continuous data type.

To illustrate the above defined tests using the example dataset we stored as "**PCorr.data"** in R (see: highlighted columns in Fig. 12.4):

1. We will test whether there exists a relationship (correlation) between the **Girth..in.** and **Height..ft.** variables of the trees example data? (**two-tailed test**).
2. Also, we will check whether the correlation (if there exist any) is a *positive* or *negative* (direction) correlation? (**one-tailed test**).

The syntax and code to performing the above tests in R is as shown in the codes below (see: Fig. 12.3, Step 3, Lines 20–36):

```
# Test for Assmp: Shapiro-Wilk's test for normality
  shapiro.test(PCorr.data$Girth..in.)
  shapiro.test(PCorr.data$Height..ft.)


# Pearson Correlation test where data is Continuous (Two-tailed)
PearsonCorr.test <- cor.test(PCorr.data$Girth..in.,
PCorr.data$Height..ft., method = "pearson")
PearsonCorr.test


# Pearson's test for Positive Correlation (One-tailed)
PearsonCorr.test2 <- cor.test(PCorr.data$Girth..in.,
PCorr.data$Height..ft., method = "pearson", alternative = "greater")
PearsonCorr.test2


# Pearson's test for Negative Correlation (One-tailed)
PearsonCorr.test3 <- cor.test(PCorr.data$Girth..in.,
PCorr.data$Height..ft., method = "pearson", alternative = "less")
PearsonCorr.test3
```

Useful Tips

- The users should always use the `alternative = "greater"` and `alternative = "less"` options to specify the "positive" and "negative" (direction) correlation tests (one-tailed), respectively.

Once the user has successfully run the codes as defined in the **Step 3** in Fig. 12.3 (Lines 20–36); they will be presented with the results of the "tests for assumptions" and the "Pearson Correlation" tests in the Console as shown in Fig. 12.5a and b, respectively.

In Fig. 12.5a, we conducted the test for assumption (data normality) necessary for the Pearson correlation test or parametric methods. This is done in order to determine if the targeted variables (i.e., **Girth..in.** and **Height..ft.**) are fitting and valid for the test (Pearson correlation, a parametric test) (see Chap. 4).

As highlighted in the figure (Fig. 12.5a); we can see that the *normality test* by using the Shapiro–Wilk's method **shapiro.test( ),** where we assume a value of $p > 0.05$ is normal, shows that the distribution of the two variables (**Girth..in.** and **Height..ft.**) are normal, with **Girth..in.** variable showing a significant value of `p-value=0.08893 (W=0.94117)` and **Height..ft.** showing significant value of `p-value=0.4034 (W=0.96545)`, respectively.

Therefore, with the necessary conditions met, we proceeded to conduct the "Pearson Correlation" as defined in the Step 3 (Fig. 12.3) and the results reported in Fig. 12.5b.

As shown in Fig. 12.5b, the authors performed the Pearson's correlation tests by considering the two variables (**Girth..in.** and **Height..ft.**). We stored the results of the tests in an R objects named "`PearsonCorr.test`" for the **two-tailed** analysis, and "`PearsonCorr.test2`" and "`PearsonCorr.test3`" for the **one-tailed** analysis, respectively.

 # Step 4—Plot and Visualize Correlation Between the Targeted Variables

Another great way to check whether there is a relationship (correlation) between the two specified variables is by plotting them as graph. By so doing, the users will be able to visualize the "linear line" between the variables.

As described in Fig. 12.6 (Step 4, Lines 39–45) and the resultant scatterplot in the same figure (Fig. 12.6); the authors applied the **ggscatter( )** function in R to visualize the relationship between the two variables "**Girth..in.**" and "**Height..ft.**" as contained in the example dataset we stored as "**PCorr.data**" in R.

The syntax and code used to plot the graph is as shown below, and the chart or scatterplot represented in Fig. 12.6.

```
# Step 4 - Visualize Correlation between the two variables

ggscatter(PCorr.data, x = "Girth..in.", y = "Height..ft.",

          add = "reg.line", conf.int = TRUE,

          cor.coef = TRUE, cor.method = "pearson",

          xlab = "Girth (inches)", ylab = "Height (ft)",

          main = "Correlation between Tree Girth and Height")

)
```

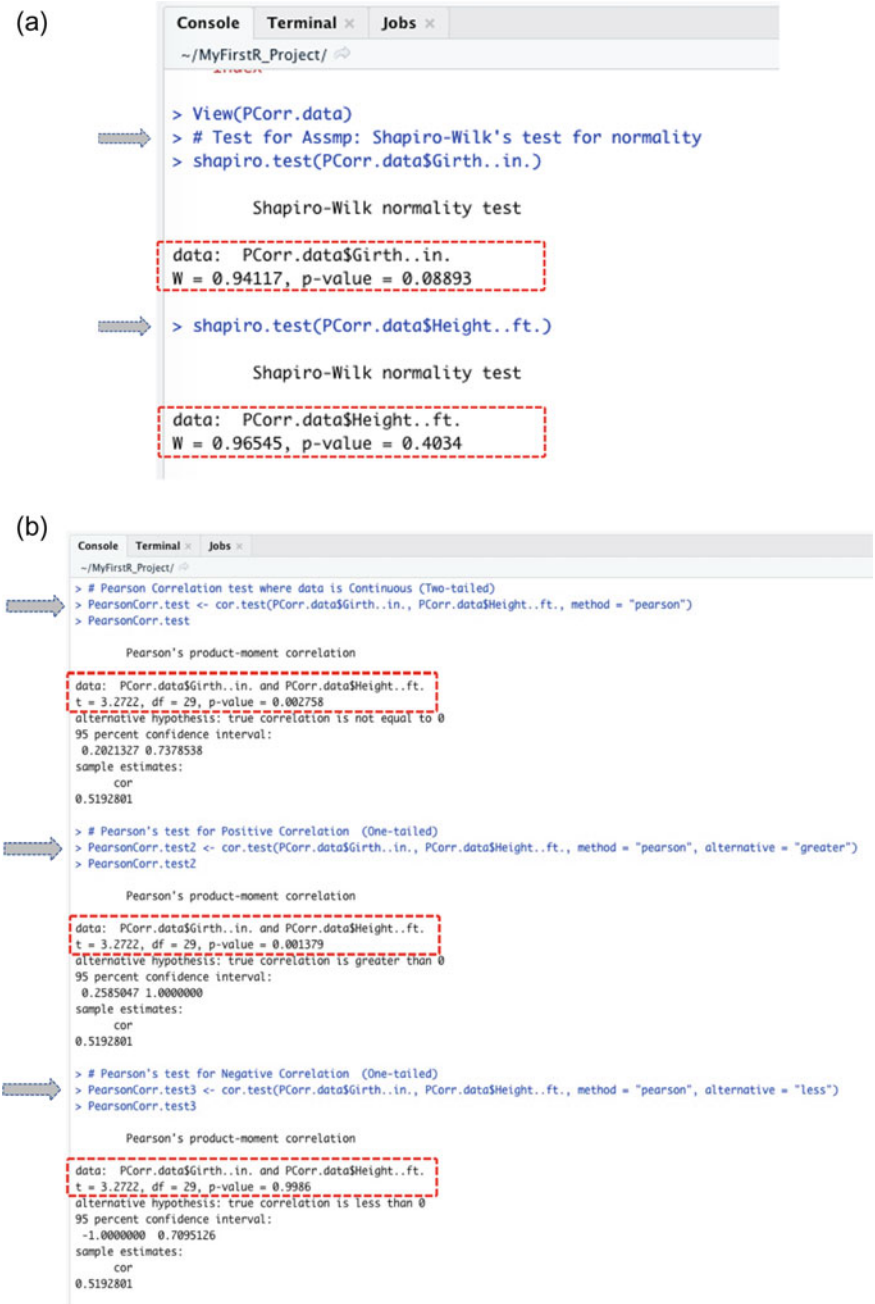

Fig. 12.5 **a** Results of test for data normality displayed in the Console in R. **b** Results of Pearson correlation tests displayed in the Console in R

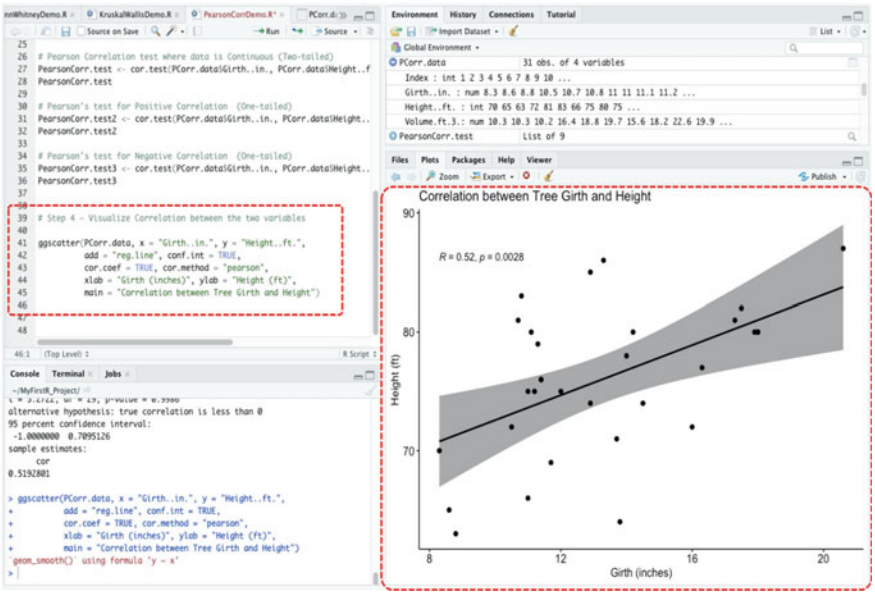

Fig. 12.6 Plot representing correlation (relationship) between two variables in R using the ggscatter( ) function

Step 5—Results Interpretation (Pearson Correlation)

The final step in the Pearson's correlation analysis is to interpret and understand the result of the test.

By default, the hypothesis for conducting the test (Pearson Correlation) by considering the two continuous variables "**Girth..in.**" and "**Height..ft.**" (see: Fig. 12.5b) is as follows;

Two-Tailed Pearson Correlation

- **(H_1)** *IF* the *p*-value of the test is less than or equal to 0.05 ($p \leq 0.05$), *THEN* we can assume that there is a correlation between the two variables (**Girth..in.** and **Height..ft.**). Thus, the population correlation coefficient (ρ) $\neq 0$. Meaning that the population correlation coefficient is not 0, therefore, we can assume that a non-zero correlation exist between the "**Girth..in.**" and "**Height..ft.**" variables.
- **(H_0)** *ELSE IF* the *p*-value is greater than 0.05 ($p > 0.05$) THEN we can say that there is no correlation between the two variables. Therefore, $\rho = 0$. Meaning that the population correlation coefficient is 0, therefore, there is no association (correlation) between the two variables.

One-Tailed Pearson Correlation

- **(H_1)** *IF* the *p*-value of the test is less than or equal to 0.05 ($p \leq 0.05$), *THEN* we can statistically assume that either $\rho > 0$, i.e., the population correlation coefficient is greater than 0, thus, a *positive* correlation may exist.
  OR
  $\rho < 0$, i.e., the population correlation coefficient is less than 0, thus, a *negative* correlation may exist between the two variables (**Girth..in.** and **Height..ft.**).
- **(H_0)** *ELSE IF* the *p*-value is greater than 0.05 ($p > 0.05$) THEN we can conclude that there is no correlation between the two variables. Therefore, $\rho = 0$. Meaning that the population correlation coefficient is 0, thus, there is no association (correlation) between the two variables.

```
> PearsonCorr.test <- cor.test(PCorr.data$Girth..in.,
PCorr.data$Height..ft., method = "pearson")
> PearsonCorr.test

     Pearson's product-moment correlation

data:  PCorr.data$Girth..in. and PCorr.data$Height..ft.
t = 3.2722, df = 29, p-value = 0.002758
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2021327 0.7378538
sample estimates:
      cor
0.5192801)
```

As shown in the above result and gathered in the outcome of the Pearson correlation (two-tailed) test for the example dataset (**PCorr.data**) represented in Fig. 12.5b; the meaning of the results of the **cor.test( )** method we applied by testing the relationship between the **Girth..in.** and **Height..ft**. variables (stored in an R object we called "`PearsonCorr.test`") can be explained as a list containing the following:

- **Statistics**: `t = 3.2722` that denotes the value of the Pearson correlation statistics.
- **Parameter**: `df = 29` which signifies the degrees of freedom for the test statistics.
- ***p*-value**: `p-value = 0.002758` is the *p*-value (significance levels) of the test.
- **Confidence interval**: `Conf.Int(95%, 0.2021327 0.7378538)` represents the confidence interval for the correlation assumed to be appropriate to the specified alternative hypothesis.
- **Sample estimates**: `cor = 0.5192801` is the value of the population correlation coefficient (ρ).

Statistically, the *p*-value of the Pearson correlation test (`PearsonCorr.test`) we conducted is `p = 0.002758` (see Fig. 12.5b). As we can see, the value is

significantly less than the scientifically acceptable significance levels ($p \leq 0.05$). Therefore, we reject the H_0 and accept H_1 by concluding that there is a significant relationship (correlation) between the two sets of variables (**Girth..in.** and **Height..ft.**) in the dataset (**two-tailed test**).

Furthermore, as shown in the next results of the Pearson correlation test presented below and in Fig. 12.5b, done for the "one-tailed" correlation tests, therein;

- We also checked whether the correlation, if any? (in this example case, yes—see result of the correlation described above) is a "positive" or "negative" (direction) correlation, respectively. The results of this particular test (one-tailed) were stored in R objects we called "`PearsonCorr.test2`" and "`PearsonCorr.test3`", respectively.

```
> > PearsonCorr.test2 <- cor.test(PCorr.data$Girth..in.,
PCorr.data$Height..ft., method = "pearson", alternative = "greater")
> PearsonCorr.test2

      Pearson's product-moment correlation

data:  PCorr.data$Girth..in. and PCorr.data$Height..ft.
t = 3.2722, df = 29, p-value = 0.001379
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
 0.2585047 1.0000000
sample estimates:
      cor
0.5192801
```

```
> PearsonCorr.test3 <- cor.test(PCorr.data$Girth..in.,
PCorr.data$Height..ft., method = "pearson", alternative = "less")
> PearsonCorr.test3

      Pearson's product-moment correlation

data:  PCorr.data$Girth..in. and PCorr.data$Height..ft.
t = 3.2722, df = 29, p-value = 0.9986
alternative hypothesis: true correlation is less than 0
95 percent confidence interval:
 -1.0000000  0.7095126
sample estimates:
      cor
0.5192801
```

As reported in the above results of the "one-tailed" tests for *positive correlation* (`PearsonCorr.test2, p=0.001379`), and *negative correlation* (`PearsonCorr.test3, p=0.9986`); we can see based on the *p*-values of the "direction test" as it is called (significant levels, $p \leq 0.05$); that the correlation

we found between the two variables "**Girth..in"** and "**Height..ft.**" (`two-tailed, PearsonCorr.test, p=0.002758`) (see Fig. 12.5b) was a "positive" directed correlation or association (PearsonCorr.test2, p=0.001379).

12.3 Kendall's Tau and Spearman's Rho Correlation Tests in R

Kendall's tau and *Spearman's rho* correlation (non-parametric equivalents or alternatives to the Pearson correlation) measures the strength of dependence or degree of association between two categorical or ordinal variables. In this statistical settings, the methods are used when the dataset the researcher or data analyst wants to investigate or analyze violates the assumptions of the parametric counterpart (Pearson), e.g., non-normally distributed data samples or existence of ordinal data type, etc.

Just like Pearson correlation test, the methods (*Kendall's tau* and *Spearman's rho*) also can be used to calculate the level of change (be it positive or negative) in one variable when another variable changes.

By default, the hypothesis for testing whether there is *correlation* (measure of strength of dependence or degree of association) between the two specified set of (categorical or ordinal) variables is; *IF* the *p*-value of the test is less than or equal to 0.05 ($p \leq 0.05$), *THEN* we can assume that there is a statistically significant strong dependence or association between the two analyzed variables, and that this is not by chance (H_1). *ELSE IF* the *p*-value is greater than 0.05 ($p > 0.05$) THEN we can say that there is no significant dependency or association between the two variables, and any observed dependency or association could only occur by chance (H_0).

Here, the authors will demonstrate how to conduct the *Kendall's tau* and *Spearman's rho* correlation tests in R using the **cor.test( )** function. We will do this following the same steps we have outlined in Fig. 12.1.

To start, **Create a new R Script** and name it "**Tau.Rho.Demo**" or any name the user may preferably choose.

Now, let's proceed to download an example dataset or file that we will use to demonstrate the two tests (*Kendall's tau* and *Spearman's rho*) (*** the users are welcome to use any dataset they may want to use provided the dataset are in the right format and type, and they can follow the example codes provided by the authors accordingly).

As shown in Fig. 12.7, download the example **.dta** dataset named "**lifeexp.dta**" through the following source: https://www.stata-press.com/data/r8/u.html and save the file on the computer or local machine (*** the example file can also be downloaded via the following repository by the authors: https://doi.org/10.6084/m9.figshare.24728073).

Once the user has successfully downloaded and saved the example file on the computer, we can proceed to conduct the *Kendall's tau* and *Spearman's rho* correlation tests using R.

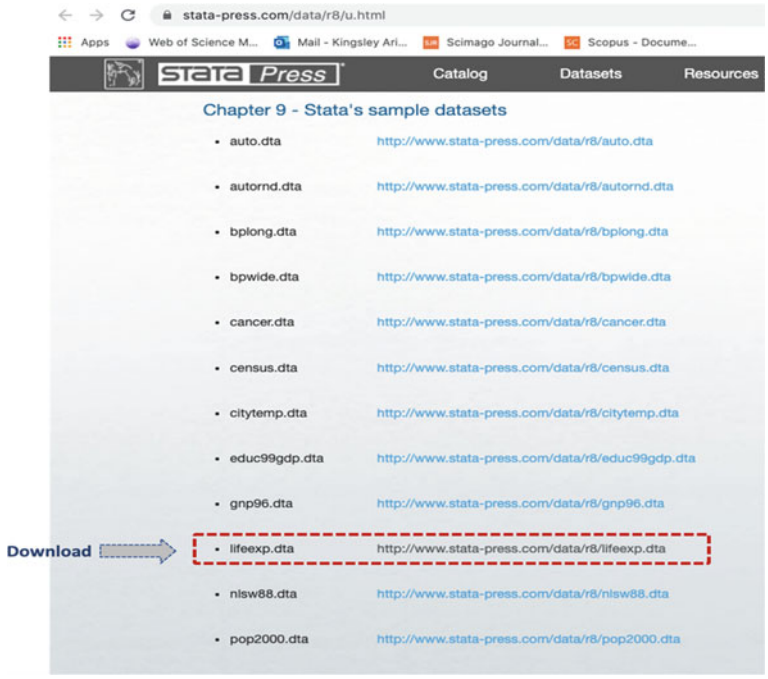

Fig. 12.7 Example of (.dta) sample file download (*Source* https://www.stata-press.com/data/r8/u.html)

Step 1—Install and Load the Required R Packages and Libraries

Install and **Load** the following *R packages* and *libraries* (see Fig. 12.8, Step 1, Lines 3–9) that will be used to call the different R functions, data manipulations, and graphical visualizations for the *Kendall's tau* and *Spearman's rho* Correlation tests.

The syntax and code to install and load the R packages and libraries are as follows: (***Note: if the reader have practiced and implemented the previous example in Sect. 12.2, then you may not need to re-install the following R packages again. New readers that may have directly visited this section will need to install and load the following packages and libraries as described below.)

```
install.packages("devtools")
install.packages("ggpubr")

library(devtools)
library(ggpubr)
```

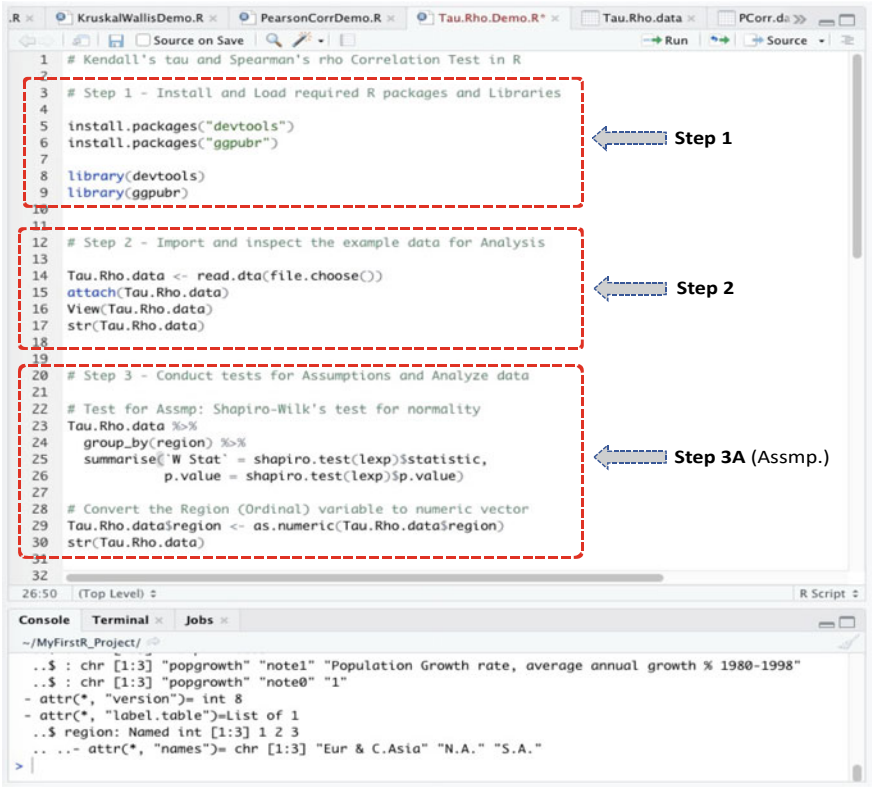

Fig. 12.8 Steps used for conducting *Kendall's tau* and *Spearman's rho* correlation tests in R

Step 2—Import and Inspect the Example Dataset for Correlation Analysis

As defined in Step 2 in Fig. 12.8 (Lines 12–17); import the dataset named "**lifeexp.dta**" that we downloaded earlier, and store this in an R object named "**Tau.Rho.data**" in R (the users are welcome to use any name of choice if they wish to do so).

Once the user has successfully imported the example dataset, they will be able to view the details of the dataset (**lifeexp.dta**) as shown in Fig. 12.9 with 68 observations and 6 variables in the data sample.

The syntax and code for importing and attaching the file in R are as shown below:

```
Tau.Rho.data <- read.dta(file.choose())
attach(Tau.Rho.data)
View(Tau.Rho.data)
str(Tau.Rho.data)
```

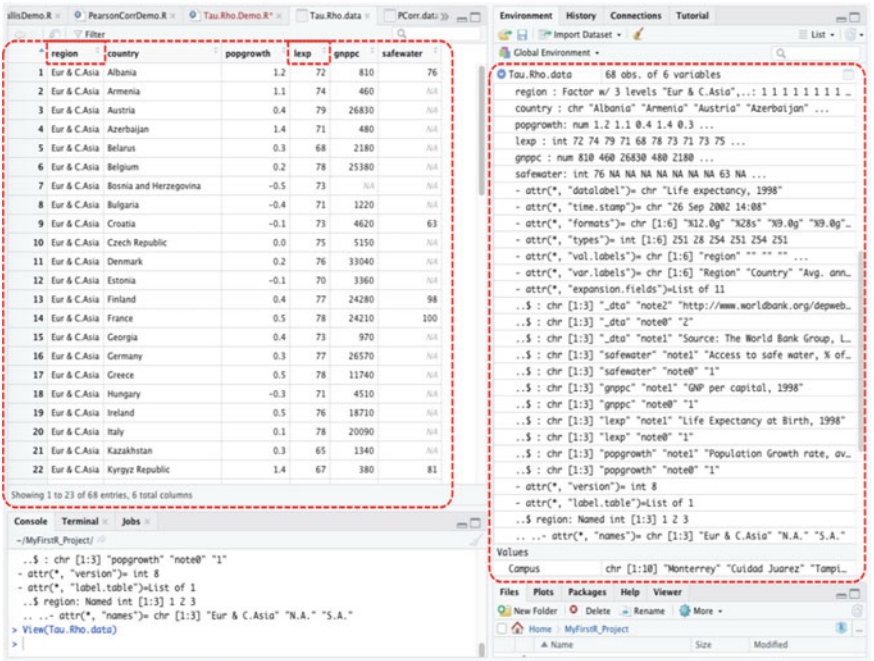

Fig. 12.9 Example of a .dta dataset imported and stored as an R object in R

Step 3—Conduct Tests for Assumptions and Analyze Data

Now that we have imported the example dataset and stored this in an R object we named "**Tau.Rho.data**", we can proceed to analyze the data.

As defined in Step 3A in Fig. 12.8 (Lines 20–30), we will first conduct the test of assumptions (e.g., data normality, and factorization of ordinal data type, etc.), and then perform the *Kendall's tau* and *Spearman's rho* tests (Step 3B, Fig. 12.10, Lines 32–57), if all the necessary conditions are met, by using the **cor.test( )** function in R.

As defined earlier in the Introduction section (Sect. 12.1);

- The **Kendall's tau** and **Spearman's rho** correlation statistics checks whether there exists a *dependency* or *association* between two independently sampled variables.
- The targeted variables should be categorical or ordinal data type.

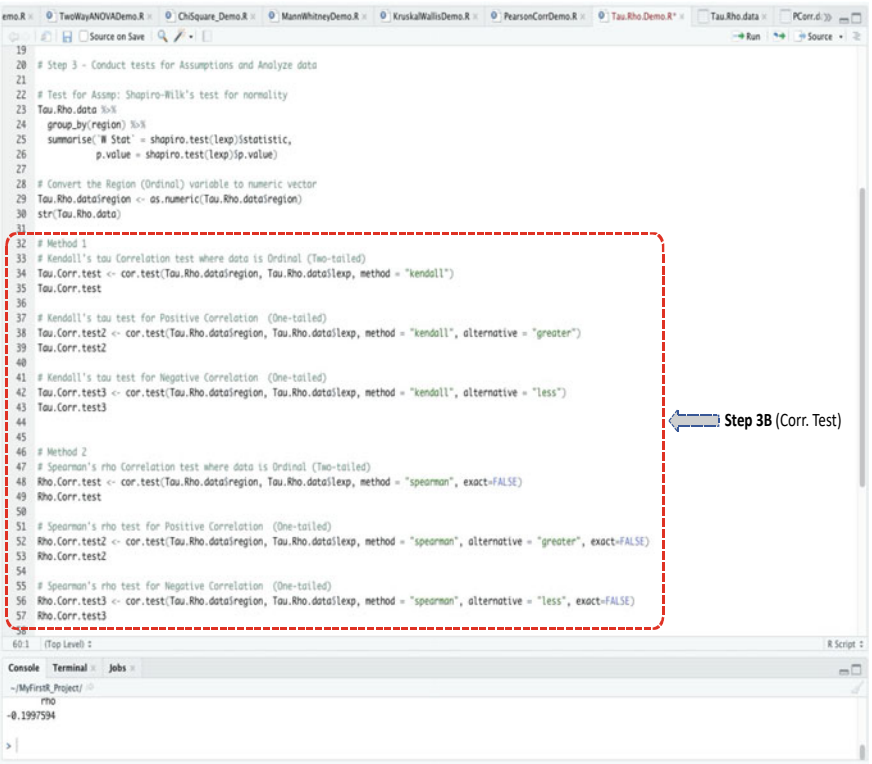

Fig. 12.10 Conducting *Kendall's tau* and *Spearman's rho* correlation tests in R

To illustrate the two tests (*Kendall's tau* and *Spearman's rho*) using the example dataset we stored as "**Tau.Rho.data"** in R (see: highlighted columns in Fig. 12.9):

1. We will test whether there exists a dependency or association (correlation) between the "**region**" and "**lexp**" variables in the example (Tau.Rho.data) life expectancy data (**two-tailed test**).
2. Then, we will also test whether the correlation (if there exist any) is a *positive* or *negative* (direction) correlation (**one-tailed test**).

The syntax to performing the above tests in R is as shown in the codes provided and described below (see: Fig. 12.10, Step 3B, Lines 32–57):

```
# Test for Assmp: Shapiro-Wilk's test for normality
Tau.Rho.data %>%
  group_by(region) %>%
  summarise(`W Stat` = shapiro.test(lexp)$statistic,
            p.value = shapiro.test(lexp)$p.value)

# Convert the Region (Ordinal) variable to numeric vector
Tau.Rho.data$region <- as.numeric(Tau.Rho.data$region)
str(Tau.Rho.data)

# Method 1
# Kendall's tau Correlation test where data is Ordinal (Two-tailed)
Tau.Corr.test <- cor.test(Tau.Rho.data$region, Tau.Rho.data$lexp, method
= "kendall")
Tau.Corr.test

# Kendall's tau test for Positive Correlation  (One-tailed)
Tau.Corr.test2   <-   cor.test(Tau.Rho.data$region,   Tau.Rho.data$lexp,
method = "kendall", alternative = "greater")
Tau.Corr.test2

# Kendall's tau test for Negative Correlation  (One-tailed)
Tau.Corr.test3   <-   cor.test(Tau.Rho.data$region,   Tau.Rho.data$lexp,
method = "kendall", alternative = "less")
Tau.Corr.test3


# Method 2
# Spearman's rho Correlation test where data is Ordinal (Two-tailed)
Rho.Corr.test <- cor.test(Tau.Rho.data$region, Tau.Rho.data$lexp, method
= "spearman", exact=FALSE)
Rho.Corr.test

# Spearman's rho test for Positive Correlation  (One-tailed)
Rho.Corr.test2   <-   cor.test(Tau.Rho.data$region,   Tau.Rho.data$lexp,
method = "spearman", alternative = "greater", exact=FALSE)
Rho.Corr.test2

# Spearman's rho test for Negative Correlation  (One-tailed)
Rho.Corr.test3   <-   cor.test(Tau.Rho.data$region,   Tau.Rho.data$lexp,
method = "spearman", alternative = "less", exact=FALSE)
Rho.Corr.test3
```

Useful Tips and Information

- The users should always use the `alternative = "greater"` and `alternative = "less"` options to specify the "positive" and "negative" (direction) correlation analysis (i.e., for one-tailed test), respectively.
- Another important task the authors conducted which the users may need to do (depending on the readily available dataset) prior to performing the tests (Kendall or Spearman) was to factorize the targeted ordinal data type (e.g., region) into a numeric format (see: Fig. 12.8, Lines 28–30) before applying the **cor.test( )** function or methods.

***Note**: For *Spearman's rho* test (Method 2), we included the R code `exact=FALSE` in the **cor.test( )** function (see: Fig. 12.10, Lines 48, 52, and 56). This was done in order to handle the error "Cannot compute exact *p*-value with ties" when running the method (Method 2—see Fig. 12.10). This is owing to the fact that the *Spearman's rho* method is much more sensitive to error and handling discrepancies in data samples than the *Kendall's tau* method, as we explained and pointed out earlier in Sect. 12.1).

Once the user has successfully run the set of codes and analysis as defined in **Steps 3A and 3B** (Figs. 12.8 and 12.10, Lines 20–57), they will be presented with the results of the "tests for assumptions", followed by the "*Kendall's tau*" (method 1) test, and then "*Spearman's rho*" (method 2) tests in the Console in R as shown in Figs. 12.11a, b, and c, respectively.

Consequentially, in Fig. 12.11a, the authors performed the test for assumption (data normality) for the *Kendall's tau* and *Spearman's rho* correlation analysis in order to determine if the selected or targeted variables "**region**" and "**lexp**" are suitable for conducting the two tests.

As highlighted in the figure (Fig. 12.11a), we can see that the *normality test* using the Shapiro–Wilk's method or function—**shapiro.test( )** (where we assume a value of $p > 0.05$ is normal) shows that the distribution of the two variables was not normally distributed, with *p*-values of the "**region**" variable (with three ranked groups) when analyzed against the "**lexp**" variable showing to be mostly non-normal values ($p \leq 0.05$) whereby the values of `p-value=0.0203` (W=0.938) for "Eur & C.Asia", `p-value=0.0538` (W=0.878) for "N.A", and `p-value=0.308 (W=0.914)` for "S.A", respectively. Therefore, we assume that the dataset or analyzed variables are not normally distributed, and a *distributed-free* method such as the *Kendall's tau* and *Spearman's rho* correlation analysis will be suitable for analyzing the data sample.

Thus, we proceed to conduct the "*Kendall's tau*" and "*Spearman's rho*" correlation analysis as defined in Step 3B (Fig. 12.10, Lines 32–57) and the results are as presented in Figs. 12.11b and c, respectively.

As shown in Figs. 12.11b, c, the authors performed the *Kendall's tau* and *Spearman's rho* tests by considering the two variables **"region"** and "**lexp"** in the example data (stored as Tau.Rho.data in R).

- The results of the *Kendall's tau* tests were stored in an R object we named "`Tau.Corr.test`" for the **two-tailed** analysis, and then

"`Tau.Corr.test2`" and "`Tau.Corr.test3`" for the **one-tailed** analysis, respectively.

- Accordingly, we stored the results of the *Spearman's rho* tests in R objects we called "`Rho.Corr.test`" for the **two-tailed** analysis, and then "`Rho.Corr.test2`" and "`Rho.Corr.test3`" for the **one-tailed** analysis, respectively.

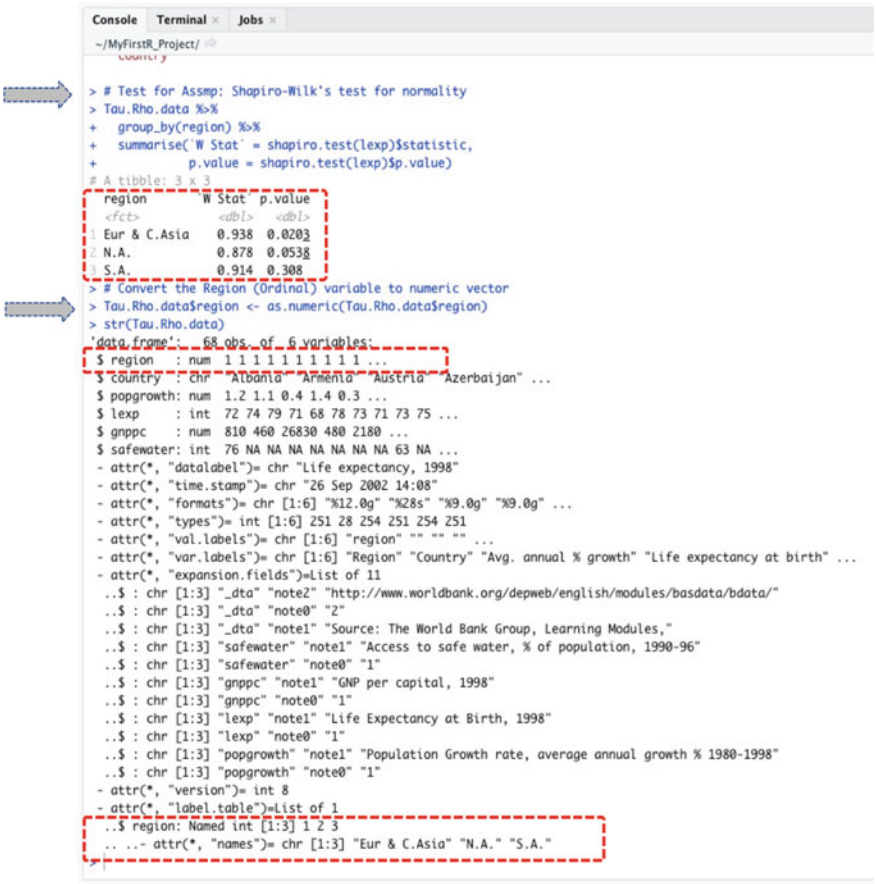

Fig. 12.11 **a** Results of test for data normality and factorization displayed in the Console in R. **b** Results of *Kendall's tau* correlation tests displayed in the Console in R. **c** Results of *Spearman's rho* correlation tests displayed in the Console in R

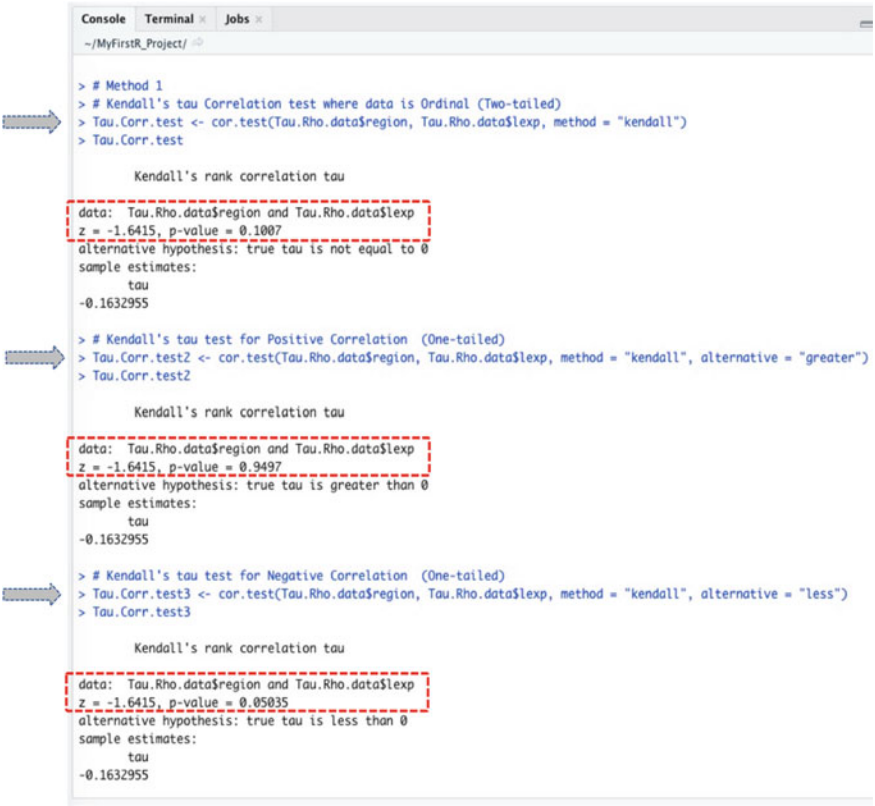

Fig. 12.11 (continued)

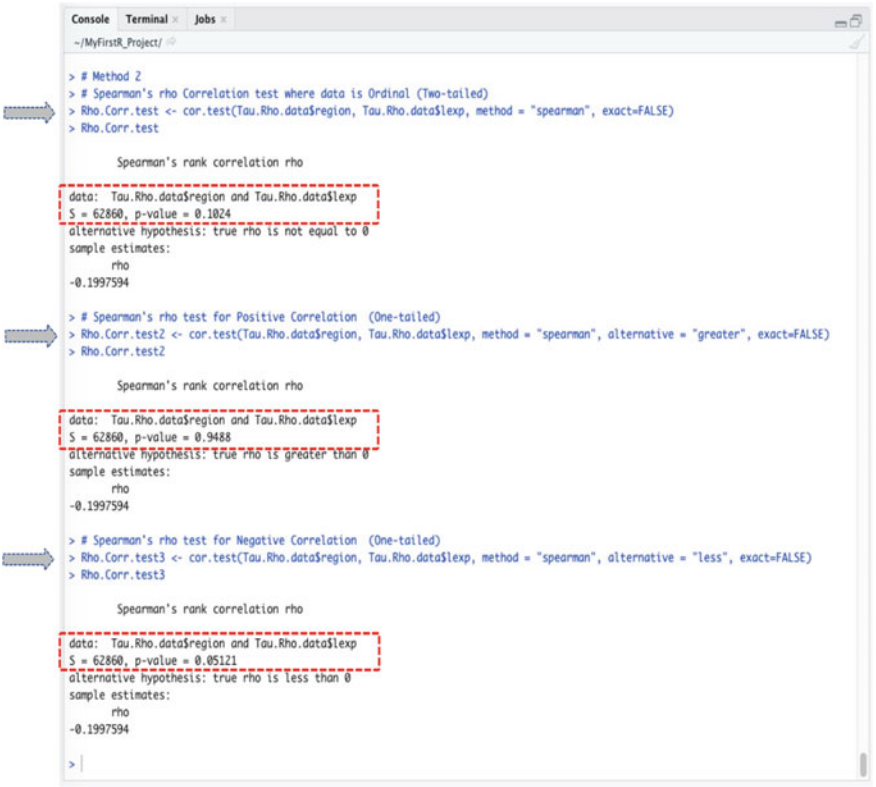

Fig. 12.11 (continued)

Step 4—Plot and Visualize Correlation Between the Variables

As previously illustrated earlier in Sect. 12.2, another way to check whether there is association or relationship (correlation) between two variables is by plotting them as graph. By so doing, the researcher or data analyst are able to visualize the *linear line* (correlation) between the two analyzed variables.

As represented in Figs. 12.12a, b (see Step 4, Lines 60–74) and the resultant scatterplots in the same figures (Fig. 12.12a, b); the authors utilized the **ggscatter( )** function to visualize the association or linearity between the two variables "**region**" and "**lexp**" as contained in the example data we stored as "**Tau.Rho.data**" in R.

The syntax and code we used to plot the graphs for both the Kendall's tau (method 1) and Spearman's rho (method 2) correlation is as shown in the codes below, and the resultant charts are represented in Figs. 12.12a and b, respectively.

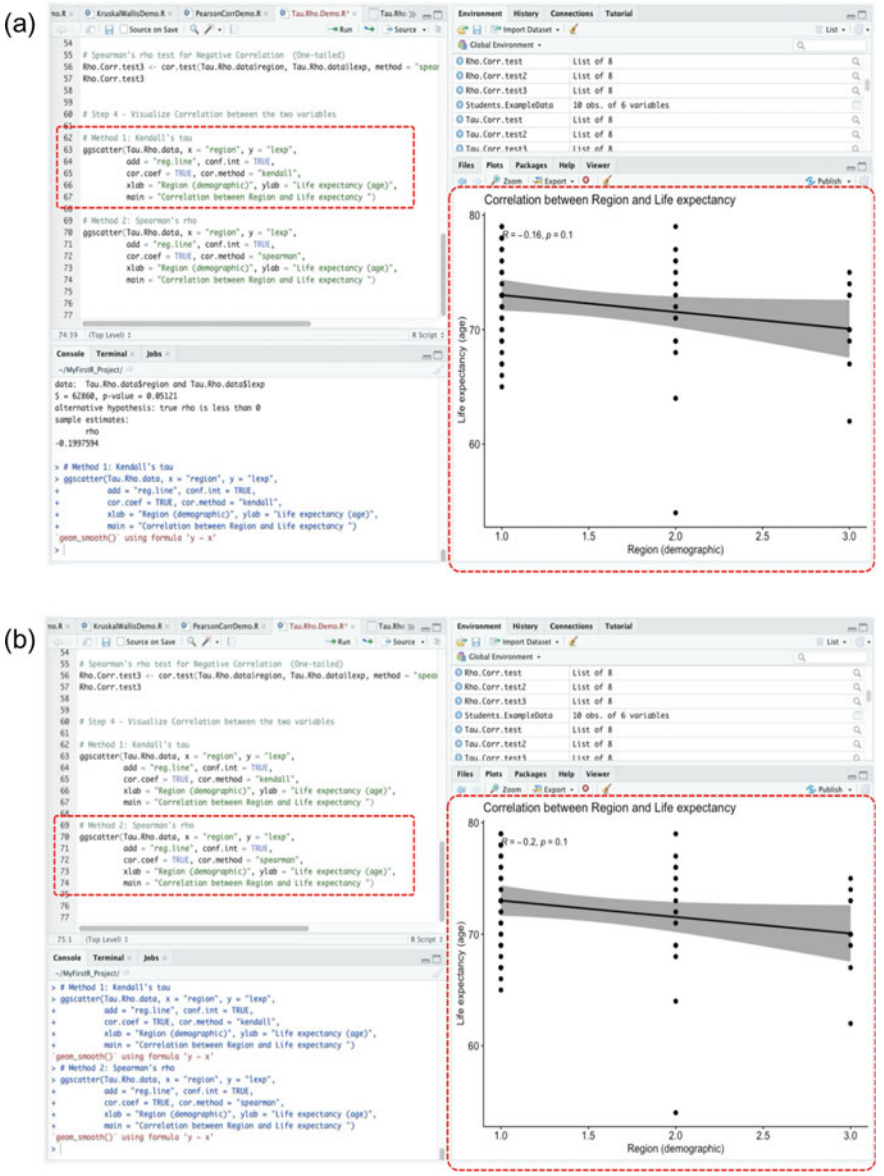

Fig. 12.12 **a** Plot for *Kendall's tau* correlation (test for dependency) between two variables in R using the ggscatter() function. **b** Plot for *Spearman's rho* correlation (test for association) between two variables in R using the ggscatter() function

```
# Step 4 - Visualize Correlation between the two variables

# Method 1: Kendall's tau
ggscatter(Tau.Rho.data, x = "region", y = "lexp",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "kendall",
          xlab = "Region (demographic)", ylab = "Life expectancy (age)",
          main = "Correlation between Region and Life expectancy ")

# Method 2: Spearman's rho
ggscatter(Tau.Rho.data, x = "region", y = "lexp",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "spearman",
          xlab = "Region (demographic)", ylab = "Life expectancy (age)",
          main = "Correlation between Region and Life expectancy ")
```

Step 5—Results Interpretation (*Kendall's Tau* and *Spearman's Rho*)

The final step for the *Kendall's tau* (method 1) and *Spearman's rho* (method 2) correlation analysis is to interpret and understand the results of the tests.

By default, the hypothesis for conducting the tests (*Kendall's tau* and *Spearman's rho*) by considering the analyzed variables "**region**" and "**lexp**" in this particular example (see: Fig. 12.11b, c) is;

Two-Tailed Kendall's Tau and Spearman's Rho Correlation Test

- **(H_1)** *IF* the *p*-value of the tests is less than or equal to 0.05 ($p \leq 0.05$), *THEN* we can assume that there is a dependency or association between the two variables (**region** and **lexp**). Thus, the population correlation coefficient (ρ) $\neq 0$. Meaning that the population correlation coefficient is not 0, and consequently, we can assume that a non-zero correlation exist between the "**region**" and "**lexp**" variables.
- **(H_0)** *ELSE IF* the *p*-value is greater than 0.05 ($p > 0.05$) THEN we can assume that there is no correlation (association) between the two variables. Thus, $\rho = 0$. Meaning that the population correlation coefficient is 0, and therefore, there is no association (correlation) between the two variables.

One-Tailed Kendall's Tau and Spearman's Rho Correlation Test

- **(H_1)** *IF* the *p*-value of the test is less than or equal to 0.05 ($p \leq 0.05$), *THEN* we can statistically assume that the value of $\rho > 0$, i.e., the population correlation coefficient is greater than 0, thus, a positive correlation exist between the two analyzed variables.

  OR

  $\rho < 0$, i.e., the population correlation coefficient is less than 0, thus, a negative correlation exist between the two variables (**region** and **lexp**).

- **(H_0)** *ELSE IF* the *p*-value is greater than 0.05 ($p > 0.05$) *THEN* we can conclude that there is no correlation between the two variables. Thus, $\rho = 0$. Meaning that the population correlation coefficient is 0, and therefore, there is no association (correlation) between the two variables.

```
> Tau.Corr.test <- cor.test(Tau.Rho.data$region, Tau.Rho.data$lexp,
method = "kendall")
> Tau.Corr.test
     Kendall's rank correlation tau
data:  Tau.Rho.data$region and Tau.Rho.data$lexp
z = -1.6415, p-value = 0.1007
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
-0.1632955)
```

```
> Rho.Corr.test <- cor.test(Tau.Rho.data$region, Tau.Rho.data$lexp,
method = "spearman", exact=FALSE)
> Rho.Corr.test
     Spearman's rank correlation rho
data:  Tau.Rho.data$region and Tau.Rho.data$lexp
S = 62860, p-value = 0.1024
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1997594)
```

As shown in the results above which is the outcome of the *Kendall's tau* (method 1) and *Spearman's rho* (method 2) correlation analysis (Two-tailed) for the example dataset (**Tau.Rho.data**) that we have reported in Fig. 12.11b, c; the meaning of the results of the **cor.test( )** method or function that we implemented to test the association or dependency between the **region** and **lexp** variables (stored as R objects "`Tau.Corr.test`" and "`Tau.Corr.test`") can be explained as a list containing the following:

Method 1: Kendall's Tau

- **Statistics**: `z = −1.6415` denotes the value of the Kendall's tau correlation analysis.
- ***p*-value**: `p-value = 0.1007` is the *p*-value (significance level) of the test.
- **Sample estimates**: `tau = −0.1632955` is the value of the population correlation coefficient.

Method 2: Spearman's Rho

- **Statistics**: `s = 62860` signifies the value of the Spearman's rho correlation analysis.
- ***p*-value**: `p-value = 0.1024` is the *p*-value (significance level) of the test.
- **Sample estimates**: `rho = −0.1997594` is the value of the population correlation coefficient.

Statistically, we can see that the *p*-value of both tests, i.e., the *Kendall's tau* (`Tau.Corr.test, z=-1.6415, p=0.1007`, method 1) and *Spearman's rho* (`Rho.Corr.test, s=62860, p = 0.1024`, method 2) correlation analysis (Two-tailed) are conventionally the same (p=0.1) and greater than the stated or scientifically acceptable significance levels ($p \leq 0.05$). Therefore, we reject the H_1 and accept H_0 by supposedly concluding that there is no dependency or association (correlation) between the two sets of analyzed variables (**region** and **lexp**) in the example data (**two-tailed test**).

Also, as shown in the next results reported below, and in Figs. 12.11b, c for the "one-tailed" correlation tests:

- We checked whether the correlation, if any? (in this case, no) may be a *positive* or *negative* (direction) correlation by considering the outcomes or output of the *Kendall's tau* and *Spearman's rho* tests, respectively.

Method 1: Kendall's Tau Test for Positive or Negative Correlation (One-Tailed)

```
> Tau.Corr.test2 <- cor.test(Tau.Rho.data$region, Tau.Rho.data$lexp,
method = "kendall", alternative = "greater")
> Tau.Corr.test2

      Kendall's rank correlation tau

data:  Tau.Rho.data$region and Tau.Rho.data$lexp
z = -1.6415, p-value = 0.9497
alternative hypothesis: true tau is greater than 0
sample estimates:
       tau
-0.1632955
```

```
> > Tau.Corr.test3 <- cor.test(Tau.Rho.data$region,
Tau.Rho.data$lexp, method = "kendall", alternative = "less")
> Tau.Corr.test3

      Kendall's rank correlation tau

data:  Tau.Rho.data$region and Tau.Rho.data$lexp
z = -1.6415, p-value = 0.05035
alternative hypothesis: true tau is less than 0
sample estimates:
       tau
-0.1632955)
```

Method 2: Spearman's Rho Test for Positive or Negative Correlation (One-Tailed)

```
> Rho.Corr.test2 <- cor.test(Tau.Rho.data$region, Tau.Rho.data$lexp,
method = "spearman", alternative = "greater", exact=FALSE)
> Rho.Corr.test2

      Spearman's rank correlation rho

data:  Tau.Rho.data$region and Tau.Rho.data$lexp
S = 62860, p-value = 0.9488
alternative hypothesis: true rho is greater than 0
sample estimates:
       rho
-0.1997594
```

```
> Rho.Corr.test3 <- cor.test(Tau.Rho.data$region, Tau.Rho.data$lexp,
method = "spearman", alternative = "less", exact=FALSE)
> Rho.Corr.test3

      Spearman's rank correlation rho

data:  Tau.Rho.data$region and Tau.Rho.data$lexp
S = 62860, p-value = 0.05121
alternative hypothesis: true rho is less than 0
sample estimates:
       rho
-0.1997594)
```

As gathered in the above results for the "one-tailed" test for *positive* and *negative* correlation (direction test) for the *Kendall's tau* (method 1) and *Spearman's rho* (method 2) tests; we can see that the results of the direction test (one-tailed) based on the *p*-values or estimated significance levels, i.e., $p \leq 0.05$, show that there is a negatively directed correlation between the targeted variables (`Tau.Corr.test3, p = 0.05035`) and (`Rho.Corr.test3, p = 0.05121`), respectively. Indeed, this is also reflected in the outcomes of the **two-tailed** test results (see Fig. 12.11b, c), therein we found that the sample estimates or population correlation coefficient (ρ) is less than 0, (i.e. Kendall tau, $\rho = -0.1632955$) and (Spearman rho, $\rho = -0.1997594$), and thus, it can be said in addition to the fact that there was no correction or association between the two analyzed variables (**region** and **lexp**), that a negatively directed correlation exists between the two variables (**region** and **lexp**).

12.4 Summary

In this chapter, the authors covered and demonstrated to the readers how to conduct the three main types of Correlational Analysis in R. This includes the practical illustration of how to perform the Pearson cor, Kendall's tau, and Spearman's rho correlation tests using R.

We illustrated how to conduct the *Pearson correlation* test, also known as the Pearson product–moment correlation coefficient in Sect. 12.2. While in Sect. 12.3, the chapter covered how to perform the *Kendall's tau* and *Spearman's rho* correlation tests.

Also, the chapter covered in each of the above sections (Sects. 12.2 and 12.3) how to graphically plot or visualize the correlation between two specified variables and/or the results of the correlational analysis. The content of the chapter also discussed in detail how to interpret and understand the results of the three main tests (Pearson, Kendall's tau, and Spearman's rho) in R.

In summary, the main contents covered in this chapter include:

- *Pearson correlation* (also known as Pearson Product–moment correlation coefficient) is a parametric procedure or statistical test of hypothesis used to compare the relationship that exists (linearity) between two sets of continuous (usually normally distributed) variables.
- *Kendall's tau* (also known as Kendall rank correlation coefficient) is described as a non-parametric procedure (distribution-free) or statistical test of hypothesis applied by the researchers to measure the strength of dependence or association between two categorical or ordinal variable types.
- *Spearman's rho* (also known as Spearman rank correlation coefficient) is equally described as non-parametric procedure (distribution-free) or statistical test of hypothesis applied by the researchers to measure the degree of association between two categorical or ordinal variable types.
- Both the *Kendall's tau* and *Spearman's rho* correlation tests are considered as the non-parametric versions or alternative to the *Pearson's correlation* test.

When choosing whether to conduct a Pearson, Kendall tau, or Spearman's rho correlation tests? The researcher or data analyst should:

- Perform the "*Pearson correlation*" if the targeted variables come from an independently sampled population, are normally distributed, in continuous data format, and shows or presents to be linearly related when plotted.
- Perform the "*Kendall's tau* or *Spearman's rho*" tests if the targeted variables come from an independently sampled population, are distribution-free (i.e., non-normally distributed), and in categorical or ordinal data format. Although it is noteworthy to mention that the two tests (i.e., Kendall's and Spearman's) can also be applied for discrete or interval datasets, as long as the dataset being analyzed has violated the test of assumptions such as data normality or homoscedasticity.
- In any case (be it Pearson, Kendall's tau, or Spearman's rho); the researchers or data analyst can perform a "one-tailed" correlational analysis to determine

the direction test (positive or negative) of the linear relationship or association/dependency (if there exist any) between the analyzed variables.

References

Akoglu, H. (2018). User's guide to correlation coefficients. In *Turkish journal of emergency medicine* (Vol. 18, Issue 3, pp. 91–93). Emergency Medicine Association of Turkey. https://doi.org/10.1016/j.tjem.2018.08.001

Brossart, D. F., Laird, V. C., & Armstrong, T. W. (2018). Interpreting Kendall's Tau and Tau-U for single-case experimental designs. *Cogent Psychology, 5*(1), 1518687. https://doi.org/10.1080/23311908.2018.1518687

Couso, I., Strauss, O., & Saulnier, H. (2018). Kendall's rank correlation on quantized data: An interval-valued approach. *Fuzzy Sets and Systems, 343*, 50–64. https://doi.org/10.1016/j.fss.2017.09.003

Hauke, J., & Kossowski, T. (2011). Comparison of values of Pearson's and Spearman's correlation coefficient on the same sets of data. *Quaestiones Geographicae*, *30*(2), 87–93. https://repozytorium.amu.edu.pl/handle/10593/15580

Privitera, G. J. (2023). *Statistics for the behavioral sciences* (4th ed.). SAGE Publications, Inc. https://us.sagepub.com/en-us/nam/statistics-for-the-behavioral-sciences/book265576#contents

Puth, M. T., Neuhäuser, M., & Ruxton, G. D. (2014). Effective use of Pearson's product-moment correlation coefficient. In *Animal behaviour* (Vol. 93, pp. 183–189). Academic Press. https://doi.org/10.1016/j.anbehav.2014.05.003

Schober, P., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia and Analgesia, 126*(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Wang, B., Wang, R., & Wang, Y. (2019). Compatible matrices of Spearman's rank correlation. *Statistics and Probability Letters, 151*, 67–72. https://doi.org/10.1016/j.spl.2019.03.015

Zar, J. H. (2014). Spearman rank correlation: Overview. In *Wiley StatsRef: Statistics reference online*. Wiley. https://doi.org/10.1002/9781118445112.stat05964