

What is the complete guide to working with the Iris dataset in R?

Authored by
stats writer

May 4, 2024

RECOMMENDED CITATION

stats writer (2024). *What is the complete guide to working with the Iris dataset in R?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=142758>

The complete guide to working with the Iris dataset in R is a comprehensive resource that provides detailed information on how to effectively work with the popular Iris dataset in the R programming language. It covers various aspects of the dataset, including its background, structure, and content. Additionally, the guide offers step-by-step instructions on how to import, manipulate, and analyze the dataset using different techniques and tools available in R. It also includes tips and best practices for data cleaning, visualization, and modeling, making it a valuable reference for both beginners and experienced users. Overall, this guide serves as a one-stop solution for anyone looking to gain a thorough understanding and practical skills in working with the Iris dataset in R.

A Complete Guide to the Iris Dataset in R

The iris dataset is a built-in dataset in R that contains measurements on 4 different attributes (in centimeters) for 50 flowers from 3 different species.

This tutorial explains how to explore and summarize a dataset in R, using the iris dataset as an example.

Load the Iris Dataset

Since the iris dataset is a built-in dataset in R, we can load it by using the following command:

```
data(iris)
```

We can take a look at the first six rows of the dataset by using the head() function:

#view first six rows of iris dataset

head(iris)

**Sepal.Length Sepal.Width Petal.Length Petal.Width
Species**

1 5.1 3.5 1.4 0.2 setosa

2 4.9 3.0 1.4 0.2 setosa

3 4.7 3.2 1.3 0.2 setosa

4 4.6 3.1 1.5 0.2 setosa

5 5.0 3.6 1.4 0.2 setosa

6 5.4 3.9 1.7 0.4 setosa

Summarize the Iris Dataset

We can use the summary() function to quickly summarize each variable in the dataset:

#summarize iris dataset

summary(iris)

Sepal.Length Sepal.Width Petal.Length Petal.Width

Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100

1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300

**Median :5.800 Median :3.000 Median :4.350 Median
:1.300**

Mean :5.843 Mean :3.057 Mean :3.758 Mean :1.199
3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800
Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500
Species
setosa :50
versicolor:50
virginica :50

For each of the numeric variables we can see the following information:

Min: The minimum value.1st Qu: The value of the first quartile (25th percentile).Median: The median value.Mean: The mean value.3rd Qu: The value of the third quartile (75th percentile).Max: The maximum value.

For the only categorical variable in the dataset (Species) we see a frequency count of each value:

setosa: This species occurs 50 times.versicolor: This species occurs 50 times.virginica: This species occurs 50 times.

We can use the dim() function to get the dimensions of

the dataset in terms of number of rows and number of columns:

```
#display rows and columns  
dim(iris)
```

150 5

We can also use the names() function to display the column names of the data frame:

```
#display column names  
names(iris)
```

```
"Sepal.Length" "Sepal.Width" "Petal.Length"  
"Petal.Width" "Species"
```

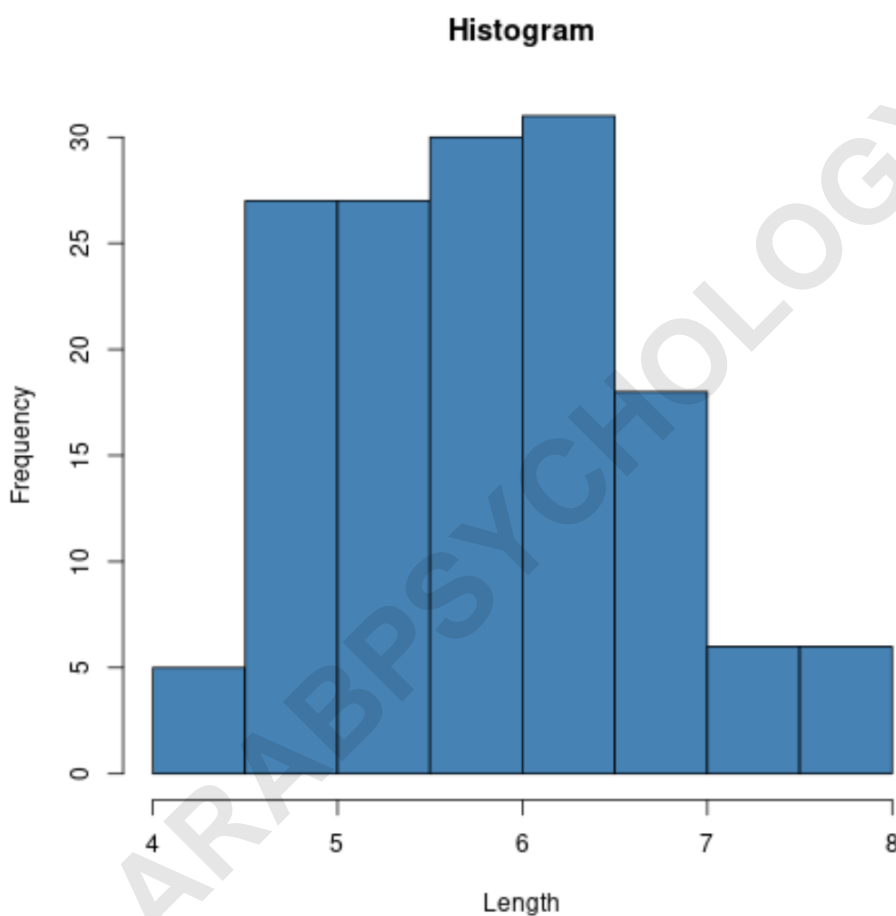
Visualize the Iris Dataset

We can also create some plots to visualize the values in the dataset.

For example, we can use the hist() function to create a histogram of the values for a certain variable:

```
#create histogram of values for sepal length
```

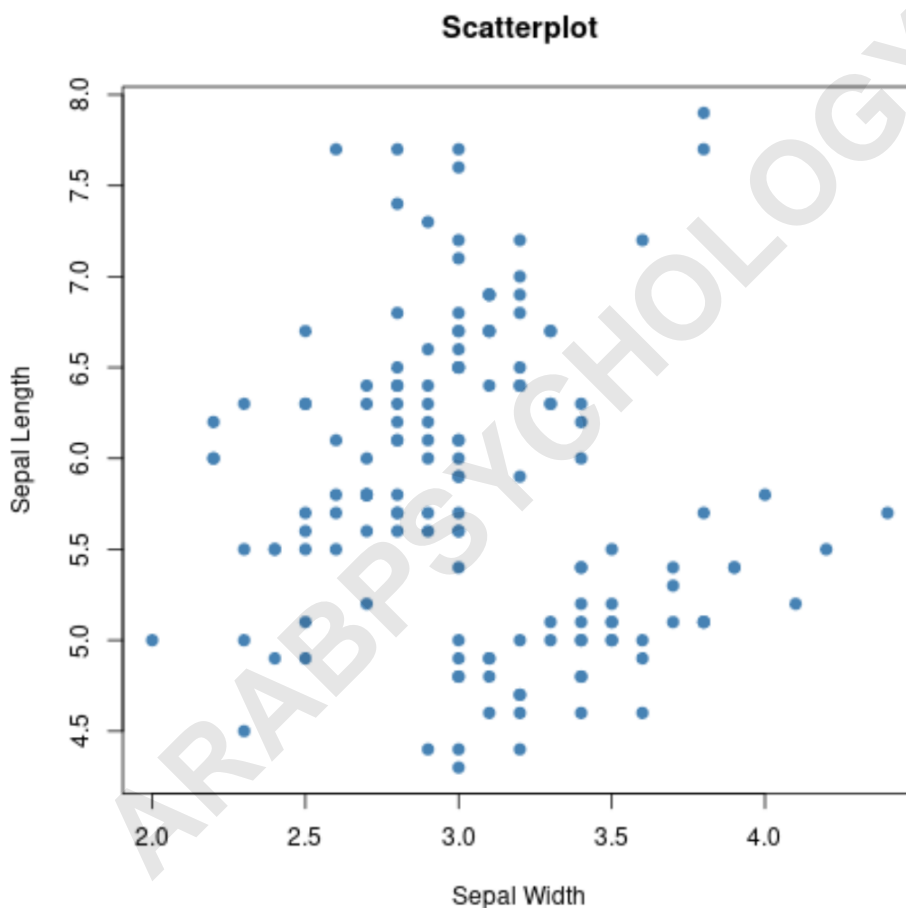
```
hist(iris$Sepal.Length,  
col='steelblue',  
main='Histogram',  
xlab='Length',  
ylab='Frequency')
```



We can also use the plot() function to create a scatterplot of any pairwise combination of variables:

#create scatterplot of sepal width vs. sepal length

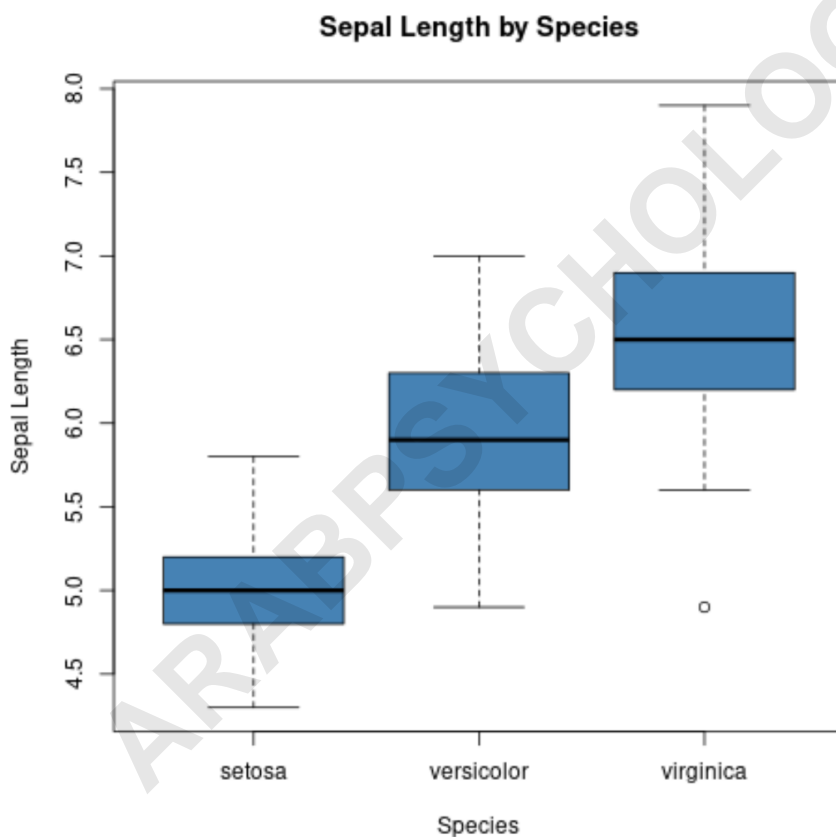
```
plot(iris$Sepal.Width, iris$Sepal.Length,  
col='steelblue',  
main='Scatterplot',  
xlab='Sepal Width',  
ylab='Sepal Length',  
pch=19)
```



We can also use the `boxplot()` function to create a boxplot by group:

#create scatterplot of sepal width vs. sepal length

```
boxplot(Sepal.Length~Species,  
data=iris,  
main='Sepal Length by Species',  
xlab='Species',  
ylab='Sepal Length',  
col='steelblue',  
border='black')
```



The x-axis displays the three species and the y-axis displays the distribution of values for sepal length for each species.

This type of plot allows us to quickly see that the sepal length tends to be largest for the virginica species and smallest for the setosa species.

The following tutorials further explain how to summarize datasets in R:

ARABPSYCHOLOGY.COM