How can I subset a data set in R?

Authored by

stats writer

June 30, 2024

ARABPSYCHOLOGY.COM

RECOMMENDED CITATION

stats writer (2024). *How can I subset a data set in R?*. PSYCHOLOGICAL SCALES. Retrieved from https://scales.arabpsychology.com/?p=161353

Subsetting a data set in R refers to the process of selecting and extracting specific rows and columns from a larger data set based on certain criteria. This can be achieved by using various functions and operators in R, such as the subset() function, the indexing operator, and logical operators. Subsetting allows for more efficient and targeted analysis of data, as well as creating smaller, more manageable data sets. It is an essential skill for data manipulation and exploration in R.

How can I subset a data set? | R FAQ

The R program (as a text file) for all the code on this page.

Subsetting is a very important component of data management and there are several ways that one can subset data in R. This page aims to give a fairly exhaustive list of the ways in which it is possible to subset a data set in R.

First we will create the data frame that will be used in all the examples. We will call this data frame x.df and it will be composed of 5 variables (V1 - V5) where the values come from a normal distribution with a mean 0 and standard deviation of 1; as well as, one variable (y) containing integers from 1 to 5.

ARABPSYCHOLOGY.COM

```
set.seed(1234)
x <- matrix(rnorm(30, 1), ncol = 5)
y <- c(1, seq(5))

#combining x and y into one matrix
x <- cbind(x, y)

#converting x into a data frame called x.df
x.df <- data.frame(x)
x.df
V1 V2 V3 V4 V5 y
1 -0.2070657 0.425260040 0.22374611 0.1628283 0.30627975 1
2 1.2774292 0.453368144 1.06445882 3.4158352 -0.44820491 1
3 2.0844412 0.435548001 1.95949406 1.1340882 1.57475572 2
4 -1.3456977 0.109962171 0.88971451 0.5093141 -0.02365572 3
5 1.4291247 0.522807300 0.48899049 0.5594521 0.98486170 4
6 1.5060559 0.001613555 0.08880458 1.4595894 0.06405140 5
```

ARABPSYCHOLOGY.COM

In order to verify which names are used for the variables
in the data frame we use the names function.

```
names(x.df)
"V1" "V2" "V3" "V4" "V5" "y"
```

Subsetting rows using the subset function

The subset function with a logical statement will let you subset
the data frame by observations. In the following example the x.sub data frame
contains only the observations for which the values of the variable y is greater than 2.

```
x.sub <- subset(x.df, y > 2)
x.sub

V1 V2 V3 V4 V5 y
4 -1.345698 0.109962171 0.88971451 0.5093141 -0.02365572 3
5 1.429125 0.522807300 0.48899049 0.5594521 0.98486170 4
6 1.506056 0.001613555 0.08880458 1.4595894
```

ARABPSYCHOLOGY.COM

0.06405140 5

Subsetting rows using multiple conditional statements

There is no limit to how many logical statements may be combined to achieve the subsetting
that is desired. The data frame x.sub1 contains only the observations for which the
values of the variable y is greater than 2 and for which the variable V1 is greater than 0.6.

```
x.sub1 <- subset(x.df, y > 2 & V1 > 0.6)
x.sub1
V1 V2 V3 V4 V5 y
5 1.429125 0.522807300 0.48899049 0.5594521 0.9848617
4
6 1.506056 0.001613555 0.08880458 1.4595894 0.0640514
5
```

Subsetting both rows and columns

It is possible to subset both rows and columns using the subset function. The
select argument lets you subset variables (columns). The data frame x.sub2

ARABPSYCHOLOGY.COM

contains only the variables V1 and V4 and then only the observations of these two variables
where the values of variable y are greater than 2 and the values of variable V2 are greater than
0.4.

```
x.sub2 <- subset(x.df, y > 2 & V2 > 0.4, select = c(V1, V4))
x.sub2
```

```
V1 V4
5 1.429125 0.5594521
```

In the data frame x.sub3 contains only the observations in variables V2-V5
for which the values in variable y are greater than 3.

```
x.sub3 <- subset(x.df, y > 3, select = V2:V5)
x.sub3
```

```
V2 V3 V4 V5
5 0.522807300 0.48899049 0.5594521 0.9848617
6 0.001613555 0.08880458 1.4595894 0.0640514
```

ARABPSYCHOLOGY.COM

Subsetting rows using indices

Another method for subsetting data sets is by using the bracket notation which designates
the indices of the data set. The first index is for the rows and the second for the columns.
The x.sub4 data frame contains only the observations for which the values of variable y are
equal to 1. Note that leaving the index for the columns blank indicates that we want x.sub4 to
contain all the variables (columns) of the original data frame.

```
x.sub4 <- x.df
x.sub4

V1 V2 V3 V4 V5 y
1 -0.2070657 0.4252600 0.2237461 0.1628283 0.3062798 1
2 1.2774292 0.4533681 1.0644588 3.4158352 -0.4482049 1
```

Subsetting rows selecting on more than one value

We use the %in% notation when we want to subset on multiple values of y.
The x.sub5 data frame contains only the observations

ARABPSYCHOLOGY.COM

for which the values of variable y are
equal to either 1 or 4.

x.sub5 <- x.df
x.sub5
V1 V2 V3 V4 V5 y
1 -0.2070657 0.4252600 0.2237461 0.1628283 0.3062798 1
2 1.2774292 0.4533681 1.0644588 3.4158352 -0.4482049 1
5 1.4291247 0.5228073 0.4889905 0.5594521 0.9848617 4

Subsetting columns using indices

We can also use the indices to subset the variables (columns) of the data set. The x.sub6
data frame contains only the first two variables of the x.df data frame. Note that leaving the
index for the rows blank indicates that we want x.sub6 to contain all the rows of the original data
frame.

x.sub6 <- x.df
x.sub6
V1 V2
1 -0.2070657 0.425260040
2 1.2774292 0.453368144

ARABPSYCHOLOGY.COM

3 2.0844412 0.435548001

4 -1.3456977 0.109962171

5 1.4291247 0.522807300

6 1.5060559 0.001613555

The x.sub7 data frame contains all the rows but only the 1st, 3rd and 5th variables
(columns) of the x.df data set.

```
x.sub7 <- x.df
x.sub7
```

V1 V3 V5

1 -0.2070657 0.22374611 0.30627975

2 1.2774292 1.06445882 -0.44820491

3 2.0844412 1.95949406 1.57475572

4 -1.3456977 0.88971451 -0.02365572

5 1.4291247 0.48899049 0.98486170

6 1.5060559 0.08880458 0.06405140

Subsetting both rows and columns using indices

The x.sub8 data frame contains the 3rd-6th variables of x.df and only observations number

ARABPSYCHOLOGY.COM

1 and 3.

x.sub8 <- x.df

x.sub8

V3 V4 V5 y

1 0.2237461 0.1628283 0.3062798 1

3 1.9594941 1.1340882 1.5747557 2

ARABPSYCHOLOGY.COM