



R news and tutorials contributed by (580) R bloggers

- [Home](#)
- [About](#)
- [RSS](#)
- [add your blog!](#)
- [Learn R](#)
- [R jobs](#) ♦ ♦ ♦
- [Contact us](#)

## Welcome!

Follow @rbloggers { 36.5K

Here you will find daily **news and tutorials about R**, contributed by over 573 bloggers.

There are many ways to **follow us -**

[By e-mail:](#)

Your e-mail here
   
 Subscribe
   
 33505 readers
   
BY FEEDBURNER

[On Facebook:](#)

R blog...
   
 40K likes

Like Page

Be the first of your friends to like this

**If you are an R blogger yourself** you are invited to [add your own R content feed to this site](#) (Non-English R bloggers should add themselves- [here](#))

## [Jobs for R-users](#)

- [Data Manager II, CRP](#)
- [Data Scientist for StartupMetrics @ New York](#)
- [Data Mining Analyst Fellow – fixed term 1 year](#)
- [Statistician](#)
- [Data Modeller & Analyst @ Manchester, UK](#)

Search & Hit Enter



## Popular Searches

- [web scraping](#)
- [heatmap](#)
- [twitteR](#)
- [maps](#)
- [time series](#)
- [boxplot](#)
- [animation](#)
- [shiny](#)
- [hadoop](#)
- [how to import image file to R](#)
- [ggplot2](#)
- [trading](#)
- [latex](#)
- [eclipse](#)
- [finance](#)
- [sql](#)
- [quantmod](#)
- [googlevis](#)
- [excel](#)
- [knitr](#)
- [PCA](#)
- [ggplot](#)
- [market research](#)
- [rstudio](#)
- [rattle](#)
- [regression](#)
- [map](#)
- [coplot](#)
- [tutorial](#)
- [rcmdr](#)

## Recent Posts

- [Analyzing World Bank data with WDI, googleVis Motion Charts](#)
- [A few thoughts on the existing code parallelization](#)
- [Fixing “Peer certificate cannot be authenticated”](#)
- [How to add pbapply to R packages](#)
- [anytime 0.0.2: Added functionality](#)
- [Collapsing a bipartite co-occurrence network](#)
- [tidyverse 1.0.0](#)
- [lubridate 1.6.0](#)
- [Network Analysis Part 1 Exercises](#)
- [Why you need version control](#)
- [HIBPwned updated on CRAN](#)
- [Data Science 101, now online](#)
- [Monitoring R](#)



[Applications with RZabbix](#)

- [How I made some Pokémon Business Cards](#)
- [2016-12 'DOM' Version 0.2](#)

## Other sites

- [Jobs for R-users](#)
- [SAS blogs](#)

# Data Manipulation with dplyr

August 20, 2015

By [Teja Kodali](#)

 Like 225  Share  Tweet  Share 30

(This article was first published on [DataScience+](#), and kindly contributed to [R-bloggers](#))

`dplyr` is a package for data manipulation, written and maintained by Hadley Wickham. It provides some great, easy-to-use functions that are very handy when performing exploratory data analysis and manipulation. Here, I will provide a basic overview of some of the most useful functions contained in the package.

For this article, I will be using the `airquality` dataset from the `datasets` package. The `airquality` dataset contains information about air quality measurements in New York from May 1973 – September 1973.

The head of the dataset looks like this:

```
head(airquality)
  Ozone Solar.R Wind Temp Month Day
1    41    190  7.4   67     5    1
2    36    118  8.0   72     5    2
3    12    149 12.6   74     5    3
4    18    313 11.5   62     5    4
5    NA     NA 14.3   56     5    5
6    28     NA 14.9   66     5    6
```

Before we dive into the functions, let's load up the two packages:

```
library(datasets)
library(dplyr)
```

Okay, now let's get to the functions.

## Filter

The filter function will return all the rows that satisfy a following condition. For example below will return all the rows where Temp is larger than 70.

```
filter(airquality, Temp > 70)
  Ozone Solar.R Wind Temp Month Day
1    36    118  8.0   72     5    2
2    12    149 12.6   74     5    3
3     7     NA  6.9   74     5   11
4    11    320 16.6   73     5   22
5    45    252 14.9   81     5   29
6   115    223  5.7   79     5   30
...
```

Another example of filter is to return all the rows where Temp is larger than 80 and Month is after May.

```
filter(airquality, Temp > 80 & Month > 5)
  Ozone Solar.R Wind Temp Month Day
1    NA    286  8.6   78     6    1
2    NA    287  9.7   74     6    2
3    NA    186  9.2   84     6    4
4    NA    220  8.6   85     6    5
5    NA    264 14.3   79     6    6
...
```



## Mutate

Mutate is used to add new variables to the data. For example lets add a new column that displays the temperature in Celsius.

```
mutate(airquality, TempInC = (Temp - 32) * 5 / 9)
  Ozone Solar.R Wind Temp Month Day TempInC
1    41     190  7.4   67    5   1  19.44444
2    36     118  8.0   72    5   2  22.22222
3    12     149 12.6   74    5   3  23.33333
4    18     313 11.5   62    5   4  16.66667
5    NA        NA 14.3   56    5   5  13.33333
...
```

## Summarise

The summarise function is used to summarise multiple values into a single value. It is very powerful when used in conjunction with the other functions in the dplyr package, as demonstrated below. na.rm = TRUE will remove all NA values while calculating the mean, so that it doesn't produce spurious results.

```
summarise(airquality, mean(Temp, na.rm = TRUE))
  mean(Temp)
1    77.88235
```

## Group By

The group\_by function is used to group data by one or more variables. Will group the data together based on the Month, and then the summarise function is used to calculate the mean temperature in each month.

```
summarise(group_by(airquality, Month), mean(Temp, na.rm = TRUE))
  Month mean(Temp)
1     5    65.54839
2     6    79.10000
3     7    83.90323
4     8    83.96774
5     9    76.90000
```

## Sample

The sample function is used to select random rows from a table. The first line of code randomly selects ten rows from the dataset, and the second line of code randomly selects 15 rows (10% of the original 153 rows) from the dataset.

```
sample_n(airquality, size = 10)
sample_frac(airquality, size = 0.1)
```

## Count

The count function tallies observations based on a group. It is slightly similar to the table function in the base package. For example:

```
count(airquality, Month)
  Month n
1     5 31
2     6 30
3     7 31
4     8 31
5     9 30
```

This means that there are 31 rows with Month = 5, 30 rows with Month = 6, and so on.

## Arrange

The arrange function is used to arrange rows by variables. Currently, the airquality dataset is arranged based on Month, and then Day. We can use the arrange function to arrange the rows in the descending order of Month, and then in the ascending order of Day.

```
arrange(airquality, desc(Month), Day)
  Ozone Solar.R Wind Temp Month Day
1    96     167  6.9   91    9   1
2    78     197  5.1   92    9   2
3    73     183  2.8   93    9   3
4    91     189  4.6   93    9   4
5    47      95  7.4   87    9   5
6    32      92 15.5   84    9   6
```

## Pipe

The pipe operator in R, represented by %>% can be used to chain code together. It is very useful when you are performing several operations



on data, and don't want to save the output at each intermediate step.

For example, let's say we want to remove all the data corresponding to Month = 5, group the data by month, and then find the mean of the temperature each month. The conventional way to write the code for this would be:

```
filteredData <- filter(airquality, Month != 5)
groupedData <- group_by(filteredData, Month)
summarise(groupedData, mean(Temp, na.rm = TRUE))
```

With piping, the above code can be rewritten as:

```
airquality %>%
  filter(Month != 5) %>%
  group_by(Month) %>%
  summarise(mean(Temp, na.rm = TRUE))
```

This is a very basic example, and the usefulness may not be very apparent, but as the number of operations/functions performed on the data increase, the pipe operator becomes more and more useful!

That brings us to the end of this article. I hope you enjoyed it and found it useful. If you have questions, feel free to leave a comment or reach out to me on [Twitter](#).

◆ 2 comments on this item ◆ Share on Facebook ◆ Stumble It! ◆ Digg This!

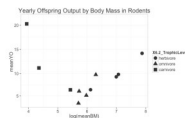
Like 225 Share Tweet Share 30

#### Related



Hands-on dplyr tutorial  
for faster data  
manipulation in R  
In "R bloggers"

Comments: 85  
Introducing dplyr  
In "R bloggers"



Introduction to dplyr:  
data manipulation made  
easy(er) and fun(er)  
In "R bloggers"

225 Like Share Tweet Share 30

To leave a comment for the author, please follow the link and comment on their blog:  
[DataScience+.](#)

R-bloggers.com offers **daily e-mail updates** about R news and **tutorials** on topics such as: [Data science](#), [Big Data](#), [R jobs](#), visualization ([ggplot2](#), [Boxplots](#), [maps](#), [animation](#)), programming ([RStudio](#), [Sweave](#), [LaTeX](#), [SQL](#), [Eclipse](#), [git](#), [hadoop](#), [Web Scraping](#)) statistics ([regression](#), [PCA](#), [time series](#), [trading](#)) and more...

If you got this far, why not **subscribe for updates** from the site?  
Choose your flavor: [e-mail](#), [twitter](#), [RSS](#), or [facebook](#)...

Like 225 Share Tweet Share 30

Comments are closed.

Search & Hit Enter

## Recent popular posts

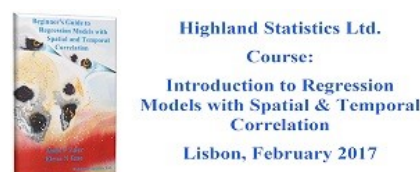
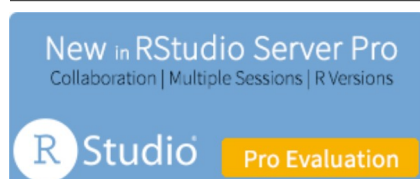
- [Why you need version control](#)
- [How to add pbapply to R packages?](#)

## Most visited articles of the week



1. [How to write the first for loop in R](#)
2. [Installing R packages](#)
3. [Using apply, sapply, lapply in R](#)
4. [R tutorials](#)
5. [Weapons of Math Destruction – A Data Scientist's Guide to Disarmament](#)
6. [How to Make a Histogram with Basic R](#)
7. [In-depth introduction to machine learning in 15 hours of expert videos](#)
8. [How to perform a Logistic Regression in R](#)
9. [Data Science 101, now online](#)

## Sponsors



Quantide: statistical consulting and training

**R Courses for Professionals**  
Download R templates for machine learning

The fastest way to learn data science!

[VIEW SNEAK PEAK](#)

DATA SOCIETY

**[R] Kenntnis-Tage 2016**  
Wissen & Vernetzen by  eoda  
Data Science goes professional

 Kassel  
2.11. – 3.11.

 **STATISTICS**  
VIEWS  
Bringing Statistics Together

 **NYC DATA SCIENCE ACADEMY**

**Become a Data Scientist**  
Develop expertise in R, Python, Hadoop & Spark  
In just 12 Weeks

[Apply for Data Science Bootcamp](#)

**OCTOBER 8-9, 2016**

Data Science Training  
**World-Class Instructors**  
22 Hands-On Workshops

 **ODSG® UK | LONDON** [Sign up](#)

 **#ODSG® WEST**


Data Science Training  
**World-Class Instructors**  
40 Hands-On Workshops

**NOVEMBER 4-6, 2016** [Sign up](#)

**Download a FREE Chapter Today!**

From *Biometrical Science*  
**Extending the Linear Model with R**  
Generalized Linear, Mixed Effects and Nonparametric Regression Models  
SECOND EDITION

**SAVE 25% on All R Books** Promo Code **CZQ33**

 **CRC Press**  
Taylor & Francis Group  
[www.crcpress.com](http://www.crcpress.com)

**STATWORX**

Consulting  
Schulung  
Data Mining



[Mehr erfahren](#)

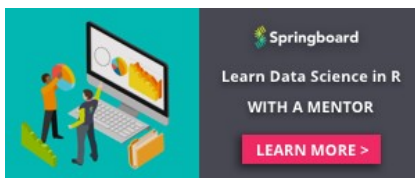
**Try the FASTEST ML for R**



[Click for a Free Trial](#)

 **YOTTAMINE ANALYTICS**



SIGMA  
SXSIGMA  
SX

## Paymetric® White Paper:

How Shifts in ePayment  
Acceptance Practices Will  
Impact Your Business



[Contact us](#) if you wish to help support  
R-bloggers, and place your banner here.

### [Jobs for R users](#)

- [Data Manager II, CRP](#)
- [Data Scientist for StartupMetrics @ New York](#)
- [Data Mining Analyst Fellow – fixed term 1 year](#)
- [Statistician](#)
- [Data Modeller & Analyst @ Manchester, UK](#)
- [Lead Data Scientist for Lumosity @ San Francisco, California, U.S.](#)
- [Environmental Specialist, Data Analyst @ Saint Petersburg, Florida, United States](#)

Search & Hit Enter

[Full list of contributing R-bloggers](#)

**R-bloggers** was founded by [Tal Galili](#), with gratitude to the [R](#) community.

Is powered by [WordPress](#) using a [bavotasan.com](#) design.

Copyright © 2016 **R-bloggers**. All Rights Reserved. [Terms and Conditions](#) for this website