

Data transformation: dplyr package

Leonard Maaya, Martina Vandebroek

2023-07-20

Introduction

Data transformation is an important step before starting to visualize and analyze your data.
Transformation may involve:

- creating new variables and summaries
- renaming some variables
- reordering observations
- format variables
- select variables
- perform analyses by group
- filter/ subset observations
- etc ...

The *dplyr* package is the workhorse for data transformation in R.

Installing dplyr

- as a standalone package:

```
install.packages("dplyr")
```

- which is then called in R by:

```
library(dplyr)
```

- as part of the *tidyverse* package

```
install.packages('tidyverse')
```

- Note: *tidyverse* is a collection of packages for data science:

- dplyr
- ggplot2
- tidyr
- ...

- here, loading *tidyverse* automatically loads *dplyr* plus the rest of the packages

```
library(tidyverse)
```

We will load *dplyr* as part of *tidyverse* package and show some of its uses on the iris data

```
#install.packages('tidyverse') # uncomment to install tidyverse  
  
library(tidyverse)  
  
iris = iris
```

A subset of the iris dataset is shown in the table below:

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
numeric	numeric	numeric	numeric	factor
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

In the next slides, we will show some key dplyr functions and their examples on the iris data

Dplyr functions

Pipes

arrange()

Sorts rows according to one or more columns

- by default, *arrange()* sorts in an ascending order

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
numeric	numeric	numeric	numeric	factor
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

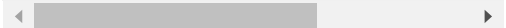
```
# Notes
## pipes: %>%
## select()

irissubset = iris %>%
  select(c(Sepal.Length, Sepal.Width))
```

Sepal.Length	Sepal.Width
numeric	numeric
5.1	3.5
4.9	3.0
4.7	3.2
4.6	3.1
5.0	3.6
5.4	3.9

```
# Notes:
## pipes: %>%
## arrange()

arranged_iris = irissubset %>% ar
```



Sepal.Length	Sepal.Width
--------------	-------------

Resources

- <https://r4ds.had.co.nz/transform.html>
- <https://cran.r-project.org/web/packages/dplyr/dplyr.pdf>