Dplyr

Victor Garcia

dplyr is a grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges:

mutate() adds new variables that are functions of existing variables select() picks variables based on their names. filter() picks cases based on their values. summarise() reduces multiple values down to a single summary. arrange() changes the ordering of the rows. These all combine naturally with group_by() which allows you to perform any operation "by group". You can learn more about them in vignette("dplyr"). As well as these single-table verbs, dplyr also provides a variety of two-table verbs, which you can learn about in vignette("two-table").

If you are new to dplyr, the best place to start is the data transformation chapter in R for data science.

Backends In addition to data frames/tibbles, dplyr makes working with other computational backends accessible and efficient. Below is a list of alternative backends:

dtplyr: for large, in-memory datasets. Translates your dplyr code to high performance data.table code.

dbplyr: for data stored in a relational database. Translates your dplyr code to SQL.

sparklyr: for very large datasets stored in Apache Spark.

```
#Installation
# The easiest way to get dplyr is to install the whole tidyverse:
#install.packages("tidyverse")

# Alternatively, install just dplyr:
#install.packages("dplyr")
#Development version
#To get a bug fix or to use a feature from the development version, you can install the development ver

#install.packages("devtools")
#devtools::install_github("tidyverse/dplyr")
#Cheatsheet
```

```
#Usage
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
starwars %>%
  filter(species == "Droid")
```

```
## # A tibble: 6 x 14
##   name  height  mass hair_color skin_color eye_color birth_year sex   gender
##   <chr>  <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
## 1 C-3PO    167    75 <NA>       gold       yellow           112 none  mascu~
## 2 R2-D2     96    32 <NA>       white, bl~ red               33 none  mascu~
## 3 R5-D4     97    32 <NA>       white, red red               NA none  mascu~
## 4 IG-88    200   140 none       metal      red               15 none  mascu~
## 5 R4-P~     96    NA none       silver, r~ red, blue         NA none  femin~
## 6 BB8       NA    NA none       none       black             NA none  mascu~
## # ... with 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

```
#> # A tibble: 6 x 14
#>   name  height  mass hair_color skin_color eye_color birth_year sex   gender
#>   <chr>  <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
#> 1 C-3PO    167    75 <NA>       gold       yellow           112 none  mascu...
#> 2 R2-D2     96    32 <NA>       white, bl... red               33 none  mascu...
#> 3 R5-D4     97    32 <NA>       white, red red               NA none  mascu...
#> 4 IG-88    200   140 none       metal      red               15 none  mascu...
#> 5 R4-P...     96    NA none       silver, r... red, blue         NA none  femin...
#> # ... with 1 more row, and 5 more variables: homeworld <chr>, species <chr>,
#> #   films <list>, vehicles <list>, starships <list>

starwars %>%
  select(name, ends_with("color"))
```

```
## # A tibble: 87 x 4
##    name               hair_color    skin_color  eye_color
##    <chr>              <chr>         <chr>       <chr>
##  1 Luke Skywalker     blond         fair        blue
##  2 C-3PO              <NA>          gold        yellow
##  3 R2-D2              <NA>          white, blue red
##  4 Darth Vader        none          white       yellow
##  5 Leia Organa        brown         light       brown
##  6 Owen Lars          brown, grey   light       blue
##  7 Beru Whitesun lars brown         light       blue
##  8 R5-D4              <NA>          white, red  red
##  9 Biggs Darklighter  black         light       brown
## 10 Obi-Wan Kenobi     auburn, white fair        blue-gray
## # ... with 77 more rows
```

```
#> # A tibble: 87 x 4
#>   name           hair_color skin_color  eye_color
#>   <chr>          <chr>      <chr>       <chr>
#> 1 Luke Skywalker blond      fair        blue
#> 2 C-3PO          <NA>       gold        yellow
#> 3 R2-D2          <NA>       white, blue red
#> 4 Darth Vader    none       white       yellow
#> 5 Leia Organa    brown      light       brown
```

```
#> # ... with 82 more rows

starwars %>%
  mutate(name, bmi = mass / ((height / 100)  ^ 2)) %>%
  select(name:mass, bmi)
```

```
## # A tibble: 87 x 4
##    name               height  mass   bmi
##    <chr>               <int> <dbl> <dbl>
##  1 Luke Skywalker        172    77  26.0
##  2 C-3PO                 167    75  26.9
##  3 R2-D2                  96    32  34.7
##  4 Darth Vader           202   136  33.3
##  5 Leia Organa           150    49  21.8
##  6 Owen Lars             178   120  37.9
##  7 Beru Whitesun lars    165    75  27.5
##  8 R5-D4                  97    32  34.0
##  9 Biggs Darklighter     183    84  25.1
## 10 Obi-Wan Kenobi        182    77  23.2
## # ... with 77 more rows
```

```
#> # A tibble: 87 x 4
#>   name           height  mass   bmi
#>   <chr>           <int> <dbl> <dbl>
#> 1 Luke Skywalker    172    77  26.0
#> 2 C-3PO             167    75  26.9
#> 3 R2-D2              96    32  34.7
#> 4 Darth Vader       202   136  33.3
#> 5 Leia Organa       150    49  21.8
#> # ... with 82 more rows

starwars %>%
  arrange(desc(mass))
```

```
## # A tibble: 87 x 14
##    name  height  mass hair_color skin_color eye_color birth_year sex   gender
##    <chr>  <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
##  1 Jabb~    175  1358 <NA>       green-tan~ orange         600   herm~ mascu~
##  2 Grie~    216   159 none       brown, wh~ green, y~       NA   male  mascu~
##  3 IG-88    200   140 none       metal      red             15   none  mascu~
##  4 Dart~    202   136 none       white      yellow          41.9 male  mascu~
##  5 Tarf~    234   136 brown      brown      blue            NA   male  mascu~
##  6 Owen~    178   120 brown, gr~ light      blue            52   male  mascu~
##  7 Bossk    190   113 none       green      red             53   male  mascu~
##  8 Chew~    228   112 brown      unknown    blue           200   male  mascu~
##  9 Jek ~    180   110 brown      fair       blue            NA   male  mascu~
## 10 Dext~    198   102 none       brown      yellow          NA   male  mascu~
## # ... with 77 more rows, and 5 more variables: homeworld <chr>, species <chr>,
## #   films <list>, vehicles <list>, starships <list>
```

```
#> # A tibble: 87 x 14
#>   name  height  mass hair_color skin_color eye_color birth_year sex   gender
```

```
#>   <chr>  <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
#> 1 Jabb...    175  1358 <NA>       green-tan... orange         600   herm... mascu...
#> 2 Grie...    216   159 none       brown, wh... green, y...     NA   male  mascu...
#> 3 IG-88      200   140 none       metal      red              15   none  mascu...
#> 4 Dart...    202   136 none       white      yellow         41.9 male  mascu...
#> 5 Tarf...    234   136 brown      brown      blue             NA   male  mascu...
#> # ... with 82 more rows, and 5 more variables: homeworld <chr>, species <chr>,
#> #   films <list>, vehicles <list>, starships <list>
```

```
starwars %>%
  group_by(species) %>%
  summarise(
    n = n(),
    mass = mean(mass, na.rm = TRUE)
  ) %>%
  filter(
    n > 1,
    mass > 50
  )
```

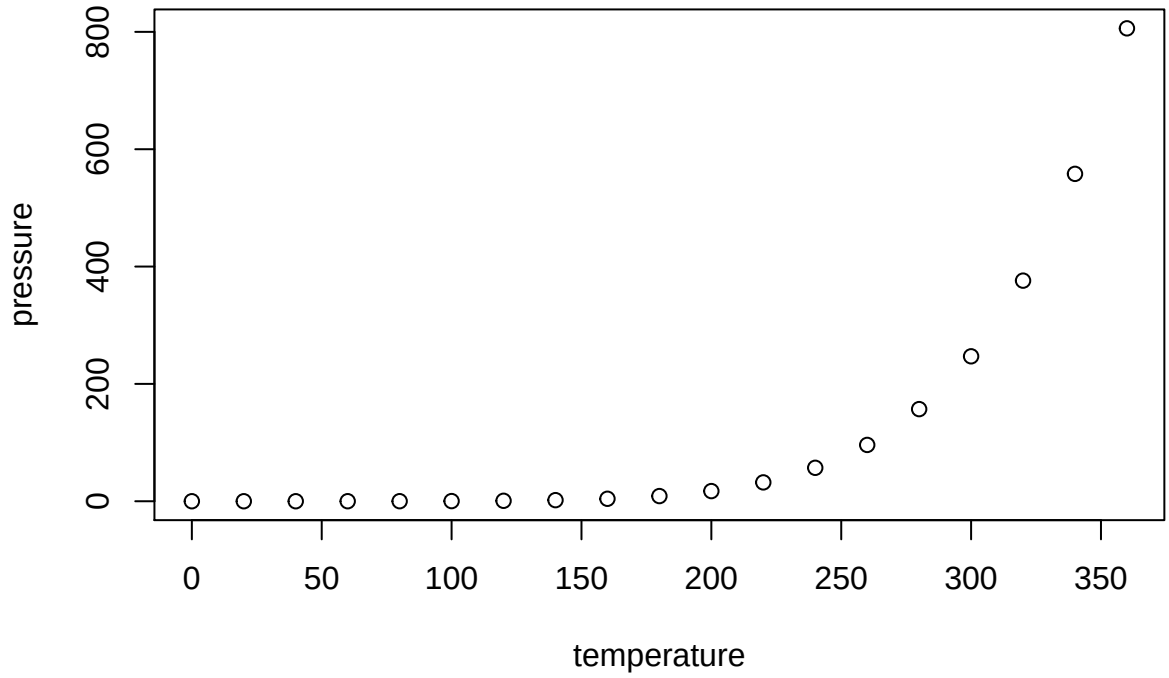
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 8 x 3
##   species      n  mass
##   <chr>    <int> <dbl>
## 1 Droid        6  69.8
## 2 Gungan       3  74
## 3 Human       35  82.8
## 4 Kaminoan     2  88
## 5 Mirialan     2  53.1
## 6 Twi'lek      2  55
## 7 Wookiee      2 124
## 8 Zabrak       2  80
```

```
#> # A tibble: 8 x 3
#>   species      n  mass
#>   <chr>    <int> <dbl>
#> 1 Droid        6  69.8
#> 2 Gungan       3  74
#> 3 Human       35  82.8
#> 4 Kaminoan     2  88
#> 5 Mirialan     2  53.1
#> # ... with 3 more rows
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.