Four Page dplyr

Andy Grogan-Kaylor

2018-10-16

Contents

1 Background

`dplyr` is a very powerful R library for managing and processing data.

While `dplyr` is very powerful, learning to use `dplyr` can be very confusing. This two page guide aims to present some of the most common `dplyr` functions and commands in the form of a brief cheatsheet.

```
library(dplyr)
```

2 Sample Data

year	x	y	z
2003	NA	Group B	108.6
2012	51.23	Group A	115
2002	34.42	Group A	87.77

year	x	y	z
2008	51.98	Group B	107.5
2010	49.38	Group A	112.7

3 Piping

Pipes %>% connect pieces of a command e.g. *data* to *data wrangling* to a *graph command*.

4 Select A Subset of Variables: `select()`

```
mynewdata <- mydata %>% select(x,y) # select only x and y
```

x	y
NA	Group B
51.23	Group A
34.42	Group A
51.98	Group B
49.38	Group A

5 Filter A Subset of Rows: `filter()`

```
mynewdata <- mydata %>%
  filter(year > 2010) # filter on year
```

year	x	y	z
2012	51.23	Group A	115

6 Create New Variables: `mutate()`

```
mynewdata <- mydata %>%
  mutate(myscale = x + z) # create a new variable e.g. a scale
```

year	x	y	z	myscale
2003	NA	Group B	108.6	NA
2012	51.23	Group A	115	166.2
2002	34.42	Group A	87.77	122.2

year	x	y	z	myscale
2008	51.98	Group B	107.5	159.5
2010	49.38	Group A	112.7	162

7 Recode Variables: `mutate()`

```
mynewdata <- mydata %>%
  mutate(zcategorical = cut(z, # use mutate to recode
                            breaks=c(-Inf, 100, Inf),
                 labels = c("low", "high")))
```

year	x	y	z	zcategorical
2003	NA	Group B	108.6	high
2012	51.23	Group A	115	high
2002	34.42	Group A	87.77	low
2008	51.98	Group B	107.5	high
2010	49.38	Group A	112.7	high

8 Rename Variables: `rename()`

```
newdata <- mydata %>%
  rename(age = x, # rename
         mental_health = z)
```

year	age	y	mental_health
2003	NA	Group B	108.6
2012	51.23	Group A	115
2002	34.42	Group A	87.77
2008	51.98	Group B	107.5
2010	49.38	Group A	112.7

9 Drop Missing Values: `filter()`

```
newdata <- mydata %>%
  filter(!is.na(x)) # filter by x is not missing
```

year	x	y	z
2012	51.23	Group A	115
2002	34.42	Group A	87.77
2008	51.98	Group B	107.5
2010	49.38	Group A	112.7

10 Connecting To Other Packages Like `ggplot`

Notice how, in the code below, I never actually create the new data set `mynewdata`. I simply pipe `mydata` into a `dplyr` command, and pipe the result directly to `ggplot2`.

```
library(ggplot2)

mydata %>% # my data
  mutate(myscale = x + z) %>% # dplyr command
  ggplot(aes(x = year, # the rest is ggplot
             y = myscale)) +
  geom_point() +
  geom_smooth(se = FALSE) + # smoother without confidence interval
  labs(title = "My Scale By Year") +
  scale_x_continuous(breaks = scales::pretty_breaks(n = 10)) +
  theme(axis.text.x = element_text(angle = 90))
```