dplyr_filter

October 12, 2020

Introduction to dplyr Package

1 1. The *dplyr* Package

- "dplyr" is short for "data plyer"
- R package for aggregating, summarizing, reshaping, and generally wrangling data
- Extremely popular in the R community
- Authored by Hadley Wickham
- Part of the "tidyverse" set of packages

1.1 1.1 The *dplyr* "Verbs"

- The *dplyr* package is organized around a set of "verbs", which are functions that operator on data
  - `filter()`
  - `summarise()`
  - `select()`
  - `mutate()`
  - `arrange()`

1.2 1.2 The "Pipe" Operator

- Can be used to pipe some object into a function call
- %>%
  - `x %>% f(y)` is the same as `f(x, y)`

2 2. `filter()` Examples with *dplyr*

```
In [3]: library(dplyr)           # load the package

In [4]: arrests_df <- read.csv("data/pvd_arrests_2020-10-03.csv")

In [7]: arrests_df %>%
            filter(gender == "Male")
```

	arrest_date <chr>	year <int>	month <int>	gender <chr>	race <chr>	ethnicity <chr>	yea <int
	2019-08-24T02:23:00.0	2019	8	Male	White	NonHispanic	198
	2019-08-23T23:43:00.0	2019	8	Male	Black	NonHispanic	199
	2019-08-23T23:43:00.0	2019	8	Male	Black	NonHispanic	199
	2019-08-23T23:43:00.0	2019	8	Male	Black	NonHispanic	199
	2019-08-23T21:38:00.0	2019	8	Male	White	Hispanic	199
	2019-08-23T19:50:00.0	2019	8	Male	White	Hispanic	200
	2019-08-23T19:50:00.0	2019	8	Male	White	Hispanic	200
	2019-08-23T18:26:00.0	2019	8	Male	White	Hispanic	199
	2019-08-23T18:26:00.0	2019	8	Male	White	Hispanic	199
	2019-08-23T14:42:00.0	2019	8	Male	White	Hispanic	199
	2019-08-23T14:00:00.0	2019	8	Male	Black	NonHispanic	197
	2019-08-23T12:00:00.0	2019	8	Male	Black	NonHispanic	198
	2019-08-23T12:00:00.0	2019	8	Male	Black	NonHispanic	198
	2019-08-23T11:49:00.0	2019	8	Male	Unknown	Hispanic	196
	2019-08-23T10:59:00.0	2019	8	Male	White	NonHispanic	198
	2019-08-23T00:57:00.0	2019	8	Male	White		199
	2019-08-23T00:00:00.0	2019	8	Male	Black	Hispanic	197
	2019-08-23T00:00:00.0	2019	8	Male	Black	Hispanic	197
	2019-08-23T00:00:00.0	2019	8	Male	Black	NonHispanic	196
	2019-08-23T00:00:00.0	2019	8	Male	Black	NonHispanic	196
	2019-08-23T00:00:00.0	2019	8	Male	White	NonHispanic	198
	2019-08-22T19:37:00.0	2019	8	Male	White	NonHispanic	197
	2019-08-22T16:27:00.0	2019	8	Male	Black	NonHispanic	199
	2019-08-22T15:35:00.0	2019	8	Male	White	NonHispanic	198
	2019-08-22T15:06:00.0	2019	8	Male	Black	NonHispanic	195
	2019-08-22T15:05:00.0	2019	8	Male	Black	NonHispanic	198
	2019-08-22T12:05:00.0	2019	8	Male	White	Hispanic	199
	2019-08-22T11:55:00.0	2019	8	Male	White	Hispanic	199
	2019-08-22T11:20:00.0	2019	8	Male	White	NonHispanic	197
A data.frame: 6804 Œ 18	2019-08-22T11:20:00.0	2019	8	Male	White	NonHispanic	197
	2020-09-28T15:37:00.0	2020	9	Male	Black	NonHispanic	198
	2020-09-28T00:07:00.0	2020	9	Male	White	NonHispanic	197
	2020-09-27T21:25:00.0	2020	9	Male	White	NonHispanic	199
	2020-09-27T21:25:00.0	2020	9	Male	White	NonHispanic	199
	2020-09-27T21:25:00.0	2020	9	Male	White	NonHispanic	199
	2020-09-27T00:00:00.0	2020	9	Male	White	Hispanic	199
	2020-09-26T22:00:00.0	2020	9	Male	Black	NonHispanic	197
	2020-09-26T21:27:00.0	2020	9	Male	White	Hispanic	197
	2020-09-26T18:55:00.0	2020	9	Male	Black	NonHispanic	198
	2020-09-26T18:40:00.0	2020	9	Male	Black	NonHispanic	196
	2020-09-26T18:40:00.0	2020	9	Male	Black	NonHispanic	196
	2020-09-26T16:27:00.0	2020	9	Male	White	NonHispanic	198
	2020-09-26T00:47:00.0	2020	9	Male	White	NonHispanic	199
	2020-09-26T00:47:00.0	2020	9	Male	White	NonHispanic	199
	2020-09-26T00:00:00.0	2020	9	Male	Black	NonHispanic	198
	2020-09-26T00:00:00.0	2020	9	Male	Black	Hispanic	196
	2020-09-25T23:30:00.0	2020	9	Male	Black	NonHispanic	198
	2020-09-25T21:23:00.0	2020	9	Male	Black	Hispanic	199
	2020-09-25T21:23:00.0	2020	9	Male	Black	Hispanic	199
	2020-09-25T15:14:00.0	2020	9	Male	White	NonHispanic	194

2.0.1 2.1.1 Comparing `filter()` with Logical Indexing

In [8]:
```r
# dplyr approach
arrests_df %>%
    filter(gender == "Male")


# "base" R approach
is_male <- arrests_df$gender == "Male"      # create vector of bools

arrests_df[is_male, ]                       # get male
```

arrest_date <chr>	year <int>	month <int>	gender <chr>	race <chr>	ethnicity <chr>	yea <int
2019-08-24T02:23:00.0	2019	8	Male	White	NonHispanic	198
2019-08-23T23:43:00.0	2019	8	Male	Black	NonHispanic	199
2019-08-23T23:43:00.0	2019	8	Male	Black	NonHispanic	199
2019-08-23T23:43:00.0	2019	8	Male	Black	NonHispanic	199
2019-08-23T21:38:00.0	2019	8	Male	White	Hispanic	199
2019-08-23T19:50:00.0	2019	8	Male	White	Hispanic	200
2019-08-23T19:50:00.0	2019	8	Male	White	Hispanic	200
2019-08-23T18:26:00.0	2019	8	Male	White	Hispanic	199
2019-08-23T18:26:00.0	2019	8	Male	White	Hispanic	199
2019-08-23T14:42:00.0	2019	8	Male	White	Hispanic	199
2019-08-23T14:00:00.0	2019	8	Male	Black	NonHispanic	197
2019-08-23T12:00:00.0	2019	8	Male	Black	NonHispanic	198
2019-08-23T12:00:00.0	2019	8	Male	Black	NonHispanic	198
2019-08-23T11:49:00.0	2019	8	Male	Unknown	Hispanic	196
2019-08-23T10:59:00.0	2019	8	Male	White	NonHispanic	198
2019-08-23T00:57:00.0	2019	8	Male	White		199
2019-08-23T00:00:00.0	2019	8	Male	Black	Hispanic	197
2019-08-23T00:00:00.0	2019	8	Male	Black	Hispanic	197
2019-08-23T00:00:00.0	2019	8	Male	Black	NonHispanic	196
2019-08-23T00:00:00.0	2019	8	Male	Black	NonHispanic	196
2019-08-23T00:00:00.0	2019	8	Male	White	NonHispanic	198
2019-08-22T19:37:00.0	2019	8	Male	White	NonHispanic	197
2019-08-22T16:27:00.0	2019	8	Male	Black	NonHispanic	199
2019-08-22T15:35:00.0	2019	8	Male	White	NonHispanic	198
2019-08-22T15:06:00.0	2019	8	Male	Black	NonHispanic	195
2019-08-22T15:05:00.0	2019	8	Male	Black	NonHispanic	198
2019-08-22T12:05:00.0	2019	8	Male	White	Hispanic	199
2019-08-22T11:55:00.0	2019	8	Male	White	Hispanic	199
2019-08-22T11:20:00.0	2019	8	Male	White	NonHispanic	197
A data.frame: 6804 Œ 18 2019-08-22T11:20:00.0	2019	8	Male	White	NonHispanic	197
2020-09-28T15:37:00.0	2020	9	Male	Black	NonHispanic	198
2020-09-28T00:07:00.0	2020	9	Male	White	NonHispanic	197
2020-09-27T21:25:00.0	2020	9	Male	White	NonHispanic	199
2020-09-27T21:25:00.0	2020	9	Male	White	NonHispanic	199
2020-09-27T21:25:00.0	2020	9	Male	White	NonHispanic	199
2020-09-27T00:00:00.0	2020	9	Male	White	Hispanic	199
2020-09-26T22:00:00.0	2020	9	Male	Black	NonHispanic	197
2020-09-26T21:27:00.0	2020	9	Male	White	Hispanic	197
2020-09-26T18:55:00.0	2020	9	Male	Black	NonHispanic	198
2020-09-26T18:40:00.0	2020	9	Male	Black	NonHispanic	196
2020-09-26T18:40:00.0	2020	9	Male	Black	NonHispanic	196
2020-09-26T16:27:00.0	2020	9	Male	White	NonHispanic	198
2020-09-26T00:47:00.0	2020	9	Male	White	NonHispanic	199
2020-09-26T00:47:00.0	2020	9	Male	White	NonHispanic	199
2020-09-26T00:00:00.0	2020	9	Male	Black	NonHispanic	198
2020-09-26T00:00:00.0	2020	9	Male	Black	Hispanic	196
2020-09-25T23:30:00.0	2020	9	Male	Black	NonHispanic	198
2020-09-25T21:23:00.0	2020	9	Male	Black	Hispanic	199
2020-09-25T21:23:00.0	2020	9	Male	Black	Hispanic	199
2020-09-25T15:14:00.0	2020	9	Male	White	NonHispanic	194

A data.frame: 6804 Œ 18

	arrest_date <chr>	year <int>	month <int>	gender <chr>	race <chr>	ethnicity <chr>
1	2019-08-24T02:23:00.0	2019	8	Male	White	NonHispanic
8	2019-08-23T23:43:00.0	2019	8	Male	Black	NonHispanic
9	2019-08-23T23:43:00.0	2019	8	Male	Black	NonHispanic
10	2019-08-23T23:43:00.0	2019	8	Male	Black	NonHispanic
11	2019-08-23T21:38:00.0	2019	8	Male	White	Hispanic
12	2019-08-23T19:50:00.0	2019	8	Male	White	Hispanic
13	2019-08-23T19:50:00.0	2019	8	Male	White	Hispanic
14	2019-08-23T18:26:00.0	2019	8	Male	White	Hispanic
15	2019-08-23T18:26:00.0	2019	8	Male	White	Hispanic
18	2019-08-23T14:42:00.0	2019	8	Male	White	Hispanic
19	2019-08-23T14:00:00.0	2019	8	Male	Black	NonHispanic
20	2019-08-23T12:00:00.0	2019	8	Male	Black	NonHispanic
21	2019-08-23T12:00:00.0	2019	8	Male	Black	NonHispanic
22	2019-08-23T11:49:00.0	2019	8	Male	Unknown	Hispanic
23	2019-08-23T10:59:00.0	2019	8	Male	White	NonHispanic
24	2019-08-23T00:57:00.0	2019	8	Male	White	
25	2019-08-23T00:00:00.0	2019	8	Male	Black	Hispanic
26	2019-08-23T00:00:00.0	2019	8	Male	Black	Hispanic
27	2019-08-23T00:00:00.0	2019	8	Male	Black	NonHispanic
28	2019-08-23T00:00:00.0	2019	8	Male	Black	NonHispanic
29	2019-08-23T00:00:00.0	2019	8	Male	White	NonHispanic
31	2019-08-22T19:37:00.0	2019	8	Male	White	NonHispanic
33	2019-08-22T16:27:00.0	2019	8	Male	Black	NonHispanic
36	2019-08-22T15:35:00.0	2019	8	Male	White	NonHispanic
37	2019-08-22T15:06:00.0	2019	8	Male	Black	NonHispanic
38	2019-08-22T15:05:00.0	2019	8	Male	Black	NonHispanic
40	2019-08-22T12:05:00.0	2019	8	Male	White	Hispanic
41	2019-08-22T11:55:00.0	2019	8	Male	White	Hispanic
42	2019-08-22T11:20:00.0	2019	8	Male	White	NonHispanic
43	2019-08-22T11:20:00.0	2019	8	Male	White	NonHispanic
8717	2020-09-28T15:37:00.0	2020	9	Male	Black	NonHispanic
8718	2020-09-28T00:07:00.0	2020	9	Male	White	NonHispanic
8719	2020-09-27T21:25:00.0	2020	9	Male	White	NonHispanic
8720	2020-09-27T21:25:00.0	2020	9	Male	White	NonHispanic
8721	2020-09-27T21:25:00.0	2020	9	Male	White	NonHispanic
8724	2020-09-27T00:00:00.0	2020	9	Male	White	Hispanic
8725	2020-09-26T22:00:00.0	2020	9	Male	Black	NonHispanic
8728	2020-09-26T21:27:00.0	2020	9	Male	White	Hispanic
8729	2020-09-26T18:55:00.0	2020	9	Male	Black	NonHispanic
8730	2020-09-26T18:40:00.0	2020	9	Male	Black	NonHispanic
8731	2020-09-26T18:40:00.0	2020	9	Male	Black	NonHispanic
8732	2020-09-26T16:27:00.0	2020	9	Male	White	NonHispanic
8733	2020-09-26T00:47:00.0	2020	9	Male	White	NonHispanic
8734	2020-09-26T00:47:00.0	2020	9	Male	White	NonHispanic
8738	2020-09-26T00:00:00.0	2020	9	Male	Black	NonHispanic
8739	2020-09-26T00:00:00.0	2020	9	Male	Black	Hispanic
8741	2020-09-25T23:30:00.0	2020	9	Male	Black	NonHispanic
8742	2020-09-25T21:23:00.0	2020	9	Male	Black	Hispanic
8743	2020-09-25T21:23:00.0	2020	9	Male	Black	Hispanic
8744	2020-09-25T15:14:00.0	2020	9	Male	White	NonHispanic

2.1 2.2 `filter()` Examples (cont.)

```
In [11]: # Here we create a new data.frame from result of filter()

         arrests_males <- arrests_df %>%
             filter(gender == "Male")
```

```
In [12]: head(arrests_males)
```

A data.frame: 6 Œ 18

	arrest_date <chr>	year <int>	month <int>	gender <chr>	race <chr>	ethnicity <chr>	year_of <int>
1	2019-08-24T02:23:00.0	2019	8	Male	White	NonHispanic	1981
2	2019-08-23T23:43:00.0	2019	8	Male	Black	NonHispanic	1991
3	2019-08-23T23:43:00.0	2019	8	Male	Black	NonHispanic	1991
4	2019-08-23T23:43:00.0	2019	8	Male	Black	NonHispanic	1991
5	2019-08-23T21:38:00.0	2019	8	Male	White	Hispanic	1996
6	2019-08-23T19:50:00.0	2019	8	Male	White	Hispanic	2000

2.2 2.2 Using `filter()` with Multiple Conditions

```
In [14]: arrests_teen_male <- arrests_df %>%
             filter(
                 gender == "Male",
                 age < 20
             )

         head(arrests_teen_male)
```

A data.frame: 6 Œ 18

	arrest_date <chr>	year <int>	month <int>	gender <chr>	race <chr>	ethnicity <chr>	year_of_birt <int>
1	2019-08-23T19:50:00.0	2019	8	Male	White	Hispanic	2000
2	2019-08-23T19:50:00.0	2019	8	Male	White	Hispanic	2000
3	2019-08-21T13:09:00.0	2019	8	Male	White	Hispanic	1999
4	2019-08-21T13:09:00.0	2019	8	Male	White	Hispanic	1999
5	2019-08-21T13:09:00.0	2019	8	Male	White	Hispanic	1999
6	2019-08-20T02:00:00.0	2019	8	Male	White	Hispanic	1999

2.2.1 2.2.1 Using `filter()` with Logical OR

- Recall the `||` operator is the logical OR
- The `|` operator performs the same role, but elementwise for columns (or vectors)

```
In [16]: young_old_male <- arrests_df %>%
             filter(
                 gender == "Male",
                 age < 25 | age > 65
             )

         head(young_old_male)
```

A data.frame: 6 Œ 18

	arrest_date <chr>	year <int>	month <int>	gender <chr>	race <chr>	ethnicity <chr>	year_of_birt <int>
1	2019-08-23T21:38:00.0	2019	8	Male	White	Hispanic	1996
2	2019-08-23T19:50:00.0	2019	8	Male	White	Hispanic	2000
3	2019-08-23T19:50:00.0	2019	8	Male	White	Hispanic	2000
4	2019-08-23T18:26:00.0	2019	8	Male	White	Hispanic	1996
5	2019-08-23T18:26:00.0	2019	8	Male	White	Hispanic	1996
6	2019-08-23T14:42:00.0	2019	8	Male	White	Hispanic	1998

2.2.2 2.2.2 Using `filter()` with Logical OR (cont.)

```
In [18]: ptk_young_old_male <- arrests_df %>%
             filter(
                 gender == "Male",
                 age < 25 | age > 65 | from_city == "Pawtucket"
             )

         head(ptk_young_old_male)
```

A data.frame: 6 Œ 18

	arrest_date <chr>	year <int>	month <int>	gender <chr>	race <chr>	ethnicity <chr>	year_of_birt <int>
1	2019-08-23T21:38:00.0	2019	8	Male	White	Hispanic	1996
2	2019-08-23T19:50:00.0	2019	8	Male	White	Hispanic	2000
3	2019-08-23T19:50:00.0	2019	8	Male	White	Hispanic	2000
4	2019-08-23T18:26:00.0	2019	8	Male	White	Hispanic	1996
5	2019-08-23T18:26:00.0	2019	8	Male	White	Hispanic	1996
6	2019-08-23T14:42:00.0	2019	8	Male	White	Hispanic	1998