dplyr Package

Mr. Sachin B.

Introduction

- `dplyr` package specially designed to help you to with data frames.
- `dplyr` package developed by Hadley Wickham of RStudio
- It is optimized version of `plyr` package.
- It does not provide conceptually any "new" functionality but greatly simplifies existing functionality in R.
- `dplyr` is fast because many key operations are coded in c++
- Everything you learn using dataframe will apply equally to other formats such as datatables, dataframes and multidimentional arrays.

Operations in R

1. `select` : returns a subset of the column of dataframe.
2. `filter` : extract subset of rows based on logical conditions.
3. `arrange` : reorder rows of dataframe.
4. `rename` : rename variables in a dataframe.
5. `mutate` : add new column or transform existing variable.
6. `summarise`/`summarize` : generate summary statistics of variable.

dplyr properties

- The first argument is a dataframe.
- The subsequent describes what to do with it.
- you can refer to column in a dataframe directly without using the `$` operator.
- The result is a new data frame

Download data

```
if(!file.exists("./data")){dir.create("./data")}

fileUrl <- "https://github.com/b-sachin/R-Programming/blob/main/chicago.rds?raw=true"


download.file(fileUrl,destfile = "./data/chicago.rds",method = "curl",extra = '-L')

chicago <- readRDS("./data/chicago.rds")
```

Understand data

```
#dim
dim(chicago)
```

```
## [1] 6940    8
```

```
#structure
str(chicago)
```

```
## 'data.frame':    6940 obs. of  8 variables:
##  $ city      : chr  "chic" "chic" "chic" "chic" ...
##  $ tmpd      : num  31.5 33 33 29 32 40 34.5 29 26.5 32.5 ...
##  $ dptp      : num  31.5 29.9 27.4 28.6 28.9 ...
##  $ date      : Date, format: "1987-01-01" "1987-01-02" ...
##  $ pm25tmean2: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ pm10tmean2: num  34 NA 34.2 47 NA ...
##  $ o3tmean2  : num  4.25 3.3 3.33 4.38 4.75 ...
##  $ no2tmean2 : num  20 23.2 23.8 30.4 30.3 ...
```

1. select

```
# Load the `dplyr` package
library(dplyr)
```

select using column name

```
## Warning: package 'dplyr' was built under R version 4.0.4
```

```
##
## 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
```

```
#names of columns
names(chicago)
```

```
## [1] "city"       "tmpd"       "dptp"       "date"       "pm25tmean2"
## [6] "pm10tmean2" "o3tmean2"   "no2tmean2"
```

```r
# Select 'city' to 'dptp' columns from chicago dataset
head(select(chicago, city:dptp)) # First Several rows
```

```
##   city tmpd   dptp
## 1 chic 31.5 31.500
## 2 chic 33.0 29.875
## 3 chic 33.0 27.375
## 4 chic 29.0 28.625
## 5 chic 32.0 28.875
## 6 chic 40.0 35.125
```

```r
# Select 'city' to 'dptp' and 'o3tmean2' columns from chicago dataset
tail(select(chicago, city:date,o3tmean2)) # Last Several rows
```

```
##      city tmpd dptp       date  o3tmean2
## 6935 chic   35 29.6 2005-12-26 14.041667
## 6936 chic   40 33.6 2005-12-27  4.468750
## 6937 chic   37 34.5 2005-12-28  3.260417
## 6938 chic   35 29.4 2005-12-29  6.794837
## 6939 chic   36 31.0 2005-12-30  3.034420
## 6940 chic   35 30.1 2005-12-31  2.531250
```

```r
# Select all columns except 'city' to 'dptp' from chicago dataset
head(select(chicago, -c(tmpd,date))) # First Several rows
```

```
##   city   dptp pm25tmean2 pm10tmean2 o3tmean2 no2tmean2
## 1 chic 31.500         NA   34.00000 4.250000  19.98810
## 2 chic 29.875         NA         NA 3.304348  23.19099
## 3 chic 27.375         NA   34.16667 3.333333  23.81548
## 4 chic 28.625         NA   47.00000 4.375000  30.43452
## 5 chic 28.875         NA         NA 4.750000  30.33333
## 6 chic 35.125         NA   48.00000 5.833333  25.77233
```

```r
#names of columns
paste(1:length(colnames(chicago)),names(chicago),sep = ".")
```

select using column number

```
## [1] "1.city"       "2.tmpd"       "3.dptp"       "4.date"       "5.pm25tmean2"
## [6] "6.pm10tmean2" "7.o3tmean2"   "8.no2tmean2"
```

```r
# Select 'city' to 'dptp' columns from chicago dataset
head(select(chicago, 1:3)) # First Several rows
```

```
##   city tmpd   dptp
## 1 chic 31.5 31.500
## 2 chic 33.0 29.875
## 3 chic 33.0 27.375
```

```
## 4 chic 29.0 28.625
## 5 chic 32.0 28.875
## 6 chic 40.0 35.125
```

```
# Select 'city' to 'dptp' and 'o3tmean2' columns from chicago dataset
tail(select(chicago, 1:4,7)) # Last Several rows
```

```
##      city tmpd dptp       date  o3tmean2
## 6935 chic   35 29.6 2005-12-26 14.041667
## 6936 chic   40 33.6 2005-12-27  4.468750
## 6937 chic   37 34.5 2005-12-28  3.260417
## 6938 chic   35 29.4 2005-12-29  6.794837
## 6939 chic   36 31.0 2005-12-30  3.034420
## 6940 chic   35 30.1 2005-12-31  2.531250
```

```
# Select all columns except 'city' to 'dptp' from chicago dataset
head(select(chicago, -c(2,4))) # First Several rows
```

```
##   city   dptp pm25tmean2 pm10tmean2 o3tmean2 no2tmean2
## 1 chic 31.500         NA   34.00000 4.250000  19.98810
## 2 chic 29.875         NA         NA 3.304348  23.19099
## 3 chic 27.375         NA   34.16667 3.333333  23.81548
## 4 chic 28.625         NA   47.00000 4.375000  30.43452
## 5 chic 28.875         NA         NA 4.750000  30.33333
## 6 chic 35.125         NA   48.00000 5.833333  25.77233
```

```
#names of columns
paste(1:length(colnames(chicago)),names(chicago),sep = ".")
```

select using column name and number

```
## [1] "1.city"       "2.tmpd"       "3.dptp"       "4.date"       "5.pm25tmean2"
## [6] "6.pm10tmean2" "7.o3tmean2"   "8.no2tmean2"
```

```
# Select 'city' to 'dptp' and 'o3tmean2' columns from chicago dataset
tail(select(chicago, 1:4,o3tmean2)) # Last Several rows
```

```
##      city tmpd dptp       date  o3tmean2
## 6935 chic   35 29.6 2005-12-26 14.041667
## 6936 chic   40 33.6 2005-12-27  4.468750
## 6937 chic   37 34.5 2005-12-28  3.260417
## 6938 chic   35 29.4 2005-12-29  6.794837
## 6939 chic   36 31.0 2005-12-30  3.034420
## 6940 chic   35 30.1 2005-12-31  2.531250
```

```
# dplyr -> Select 'city' to 'dptp' columns from chicago dataset
head(select(chicago, city:dptp)) # First Several rows
```

Select equivalent process in base R

```
##   city tmpd   dptp
## 1 chic 31.5 31.500
## 2 chic 33.0 29.875
## 3 chic 33.0 27.375
## 4 chic 29.0 28.625
## 5 chic 32.0 28.875
## 6 chic 40.0 35.125
```

```
# base R ->  Select 'city' to 'dptp' columns from chicago dataset
i <- match("city", names(chicago))
j <- match("dptp", names(chicago))

head(chicago[, (i:j)])
```

```
##   city tmpd   dptp
## 1 chic 31.5 31.500
## 2 chic 33.0 29.875
## 3 chic 33.0 27.375
## 4 chic 29.0 28.625
## 5 chic 32.0 28.875
## 6 chic 40.0 35.125
```

2. filter

```
chic.f <- filter(chicago, pm25tmean2 > 30)
head(select(chic.f, 1:3, pm25tmean2), 10)
```

```
##    city tmpd dptp pm25tmean2
## 1  chic   23 21.9      38.10
## 2  chic   28 25.8      33.95
## 3  chic   55 51.3      39.40
## 4  chic   59 53.7      35.40
## 5  chic   57 52.0      33.30
## 6  chic   57 56.0      32.10
## 7  chic   75 65.8      56.50
## 8  chic   61 59.0      33.80
## 9  chic   73 60.3      30.30
## 10 chic   78 67.1      41.40
```

```
chic.f <- filter(chicago, pm25tmean2 > 30 & tmpd > 80)
head(select(chic.f, 1:3, pm25tmean2), 10)
```

```
##    city tmpd dptp pm25tmean2
## 1  chic   81 71.2    39.6000
## 2  chic   81 70.4    31.5000
## 3  chic   82 72.2    32.3000
## 4  chic   84 72.9    43.7000
## 5  chic   85 72.6    38.8375
```

```
## 6  chic   84 72.6    38.2000
## 7  chic   82 67.4    33.0000
## 8  chic   82 63.5    42.5000
## 9  chic   81 70.4    33.1000
## 10 chic   82 66.2    38.8500
```

3. arrange

Reordering rows of a data frame (while preserving corresponding order of other columns) is normally a pain to do in R.

```
chicago <- arrange(chicago, date)
head(select(chicago, date, pm25tmean2), 3)
```

Ascending Order

```
##         date pm25tmean2
## 1 1987-01-01         NA
## 2 1987-01-02         NA
## 3 1987-01-03         NA
```

```
tail(select(chicago, date, pm25tmean2), 3)
```

```
##            date pm25tmean2
## 6938 2005-12-29    7.45000
## 6939 2005-12-30   15.05714
## 6940 2005-12-31   15.00000
```

```
chicago <- arrange(chicago, desc(date))
head(select(chicago, date, pm25tmean2), 3)
```

Decending Order

```
##         date pm25tmean2
## 1 2005-12-31   15.00000
## 2 2005-12-30   15.05714
## 3 2005-12-29    7.45000
```

```
tail(select(chicago, date, pm25tmean2), 3)
```

```
##            date pm25tmean2
## 6938 1987-01-03         NA
## 6939 1987-01-02         NA
## 6940 1987-01-01         NA
```

4. rename

Renaming a variable in a data frame in R is surprising hard to do!

```
head(chicago[, 1:5], 3)
```

```
##   city tmpd dptp       date pm25tmean2
## 1 chic   35 30.1 2005-12-31   15.00000
## 2 chic   36 31.0 2005-12-30   15.05714
## 3 chic   35 29.4 2005-12-29    7.45000
```

```
chicago <- rename(chicago, dewpoint = dptp, pm25 = pm25tmean2)
head(chicago[, 1:5], 3)
```

```
##   city tmpd dewpoint       date     pm25
## 1 chic   35     30.1 2005-12-31 15.00000
## 2 chic   36     31.0 2005-12-30 15.05714
## 3 chic   35     29.4 2005-12-29  7.45000
```

5. mutate

```
chicago <- mutate(chicago,pm25detrend=pm25-mean(pm25, na.rm=TRUE))
head(select(chicago, pm25, pm25detrend))
```

```
##       pm25 pm25detrend
## 1 15.00000   -1.230958
## 2 15.05714   -1.173815
## 3  7.45000   -8.780958
## 4 17.75000    1.519042
## 5 23.56000    7.329042
## 6  8.40000   -7.830958
```

6. group_by Summarize

Generating summary statistics

```
chicago <- mutate(chicago, tempcat = factor(1 * (tmpd > 80), labels = c("cold", "hot")))
hotcold <- group_by(chicago, tempcat)
summarize(hotcold, pm25 = mean(pm25, na.rm = TRUE),
          o3 = max(o3tmean2),
          no2 = median(no2tmean2))
```

```
## # A tibble: 3 x 4
##   tempcat  pm25    o3   no2
## * <fct>   <dbl> <dbl> <dbl>
## 1 cold     16.0 66.6   24.5
## 2 hot      26.5 63.0   24.9
## 3 <NA>     47.7  9.42  37.4
```

Generating summary statistics

```
chicago <- mutate(chicago,
                  year = as.POSIXlt(date)$year + 1900)
years <- group_by(chicago, year)
summarize(years, pm25 = mean(pm25, na.rm = TRUE),
          o3 = max(o3tmean2, na.rm = TRUE),
          no2 = median(no2tmean2, na.rm = TRUE))
```

```
## # A tibble: 19 x 4
##     year  pm25    o3   no2
##  * <dbl> <dbl> <dbl> <dbl>
##  1  1987 NaN    63.0  23.5
##  2  1988 NaN    61.7  24.5
##  3  1989 NaN    59.7  26.1
##  4  1990 NaN    52.2  22.6
##  5  1991 NaN    63.1  21.4
##  6  1992 NaN    50.8  24.8
##  7  1993 NaN    44.3  25.8
##  8  1994 NaN    52.2  28.5
##  9  1995 NaN    66.6  27.3
## 10  1996 NaN    58.4  26.4
## 11  1997 NaN    56.5  25.5
## 12  1998  18.3  50.7  24.6
## 13  1999  18.5  57.5  24.7
## 14  2000  16.9  55.8  23.5
## 15  2001  16.9  51.8  25.1
## 16  2002  15.3  54.9  22.7
## 17  2003  15.2  56.2  24.6
## 18  2004  14.6  44.5  23.4
## 19  2005  16.2  58.8  22.6
```

```
chicago$year <- NULL  ## Can't use mutate to create an existing variable
```

%>%

```
chicago %>% mutate(month = as.POSIXlt(date)$mon + 1)
    %>% group_by(month)
    %>% summarize(pm25 = mean(pm25, na.rm = TRUE),
          o3 = max(o3tmean2, na.rm = TRUE),
          no2 = median(no2tmean2, na.rm = TRUE))
```

```
## # A tibble: 12 x 4
##    month  pm25    o3   no2
##  * <dbl> <dbl> <dbl> <dbl>
##  1     1  17.8  28.2  25.4
##  2     2  20.4  37.4  26.8
##  3     3  17.4  39.0  26.8
##  4     4  13.9  47.9  25.0
##  5     5  14.1  52.8  24.2
##  6     6  15.9  66.6  25.0
```

```
##  7     7  16.6  59.5  22.4
##  8     8  16.9  54.0  23.0
##  9     9  15.9  57.5  24.5
## 10    10  14.2  47.1  24.2
## 11    11  15.2  29.5  23.6
## 12    12  17.5  27.7  24.5
```