MDA102-Statistical methods using R

Dr Sharon Varghese A

arrange()

A function in dplyr package to sort observations based on a variable

```
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

chicago<-readRDS("chicago.rds")
#head(chicago)
#str(chicago)
chicago_arra<-arrange(chicago,date)
tail(chicago_arra)

     city tmpd dptp       date pm25tmean2 pm10tmean2  o3tmean2 no2tmean2
6935 chic   35 29.6 2005-12-26    8.40000        8.5 14.041667  16.81944
6936 chic   40 33.6 2005-12-27   23.56000       27.0  4.468750  23.50000
6937 chic   37 34.5 2005-12-28   17.75000       27.5  3.260417  19.28563
6938 chic   35 29.4 2005-12-29    7.45000       23.5  6.794837  19.97222
6939 chic   36 31.0 2005-12-30   15.05714       19.2  3.034420  22.80556
6940 chic   35 30.1 2005-12-31   15.00000       23.5  2.531250  13.25000

chicago_arra_desc<-arrange(chicago,desc(date))
head(chicago_arra_desc)

  city tmpd dptp       date pm25tmean2 pm10tmean2  o3tmean2 no2tmean2
1 chic   35 30.1 2005-12-31   15.00000       23.5  2.531250  13.25000
2 chic   36 31.0 2005-12-30   15.05714       19.2  3.034420  22.80556
3 chic   35 29.4 2005-12-29    7.45000       23.5  6.794837  19.97222
4 chic   37 34.5 2005-12-28   17.75000       27.5  3.260417  19.28563
5 chic   40 33.6 2005-12-27   23.56000       27.0  4.468750  23.50000
6 chic   35 29.6 2005-12-26    8.40000        8.5 14.041667  16.81944
```

rename()

To give new names to the variables.New variable name is given before '=' sign and old variable name after '=' sign.

```
chicago_rename<-rename(chicago,dewpoint=dptp,pm25=pm25tmean2)
head(chicago_rename)
```

```
  city tmpd dewpoint       date pm25 pm10tmean2 o3tmean2 no2tmean2
1 chic 31.5   31.500 1987-01-01   NA   34.00000 4.250000  19.98810
2 chic 33.0   29.875 1987-01-02   NA         NA 3.304348  23.19099
3 chic 33.0   27.375 1987-01-03   NA   34.16667 3.333333  23.81548
4 chic 29.0   28.625 1987-01-04   NA   47.00000 4.375000  30.43452
5 chic 32.0   28.875 1987-01-05   NA         NA 4.750000  30.33333
6 chic 40.0   35.125 1987-01-06   NA   48.00000 5.833333  25.77233
```

mutate()

This function is used to create a new variable

```
chicago_mutate<-mutate(chicago,tmpd_prod=tmpd*dptp)#multiple variables can be
created separated by comma
head(chicago_mutate)
```

```
  city tmpd   dptp       date pm25tmean2 pm10tmean2 o3tmean2 no2tmean2
1 chic 31.5 31.500 1987-01-01         NA   34.00000 4.250000  19.98810
2 chic 33.0 29.875 1987-01-02         NA         NA 3.304348  23.19099
3 chic 33.0 27.375 1987-01-03         NA   34.16667 3.333333  23.81548
4 chic 29.0 28.625 1987-01-04         NA   47.00000 4.375000  30.43452
5 chic 32.0 28.875 1987-01-05         NA         NA 4.750000  30.33333
6 chic 40.0 35.125 1987-01-06         NA   48.00000 5.833333  25.77233
  tmpd_prod
1   992.250
2   985.875
3   903.375
4   830.125
5   924.000
6  1405.000
```

transmute()

This function create separate data frame for new variables and exclude all existing variables.

```
chicago_transmute<-transmute(chicago,tmpd_prod=tmpd*dptp,o3_no2=o3tmean2*no2t
mean2)
head(chicago_transmute)
```

```
  tmpd_prod    o3_no2
1   992.250  84.94940
2   985.875  76.63111
3   903.375  79.38492
4   830.125 133.15104
5   924.000 144.08333
6  1405.000 150.33860
```

group_by() and summrize()

These functions are used to group by some variables and summary function to find summary across all those subset. To make subgroups in a data frame with respect to a variable

```
chicago_year<-mutate(chicago,year=as.POSIXlt(date)$year+1900)
head(chicago_year)
```

```
  city tmpd   dptp       date pm25tmean2 pm10tmean2 o3tmean2 no2tmean2 year
1 chic 31.5 31.500 1987-01-01         NA   34.00000 4.250000  19.98810 1987
2 chic 33.0 29.875 1987-01-02         NA         NA 3.304348  23.19099 1987
3 chic 33.0 27.375 1987-01-03         NA   34.16667 3.333333  23.81548 1987
4 chic 29.0 28.625 1987-01-04         NA   47.00000 4.375000  30.43452 1987
5 chic 32.0 28.875 1987-01-05         NA         NA 4.750000  30.33333 1987
6 chic 40.0 35.125 1987-01-06         NA   48.00000 5.833333  25.77233 1987
```

```
chicago_group<-group_by(chicago_year,year)#group by years
summarize(chicago_group,mean(pm25tmean2,na.rm = TRUE),median(o3tmean2,na.rm =
TRUE))#to show year wise summary mean and median of corresponding variables
```

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

```
# A tibble: 19 x 3
    year `mean(pm25tmean2, na.rm = TRUE)` `median(o3tmean2, na.rm = TRUE)`
   <dbl>                            <dbl>                             <dbl>
 1  1987                            NaN                                18.8
 2  1988                            NaN                                20.4
 3  1989                            NaN                                19.3
 4  1990                            NaN                                19.0
 5  1991                            NaN                                18.4
 6  1992                            NaN                                15.2
 7  1993                            NaN                                15.0
 8  1994                            NaN                                16.0
 9  1995                            NaN                                16.8
10  1996                            NaN                                15.8
11  1997                            NaN                                18.2
12  1998                             18.3                              20.2
13  1999                             18.5                              20.5
14  2000                             16.9                              18.1
15  2001                             16.9                              18.8
16  2002                             15.3                              19.9
```

```
17  2003                                15.2                                19.5
18  2004                                14.6                                20.7
19  2005                                16.2                                23.1
```

Managing date and time in R

Date and time are stored in two classes in R. As POSIXct by default and it can be converted into other class POSIXlt. POSIXlt has more metadata about time and date. Underlying information on a R object can be seen using unclass() function.

```
p<-Sys.time()
p

[1] "2020-07-15 11:38:01 IST"

class(p)

[1] "POSIXct" "POSIXt"

c<-as.POSIXlt(p)#to convert into POSIXlt class
class(c)

[1] "POSIXlt" "POSIXt"

names(unclass(c))

 [1] "sec"    "min"    "hour"   "mday"   "mon"    "year"   "wday"   "yday"
 [9] "isdst"  "zone"   "gmtoff"

unclass(c)

$sec
[1] 1.187176

$min
[1] 38

$hour
[1] 11

$mday
[1] 15

$mon
[1] 6

$year
[1] 120

$wday
```

```
[1] 3

$yday
[1] 196

$isdst
[1] 0

$zone
[1] "IST"

$gmtoff
[1] 19800

attr(,"tzone")
[1] ""      "IST"   "+0630"

year=c$year+1900#year will be counted as number of years from 1900
year

[1] 2020
```