dataframe basic maniputation

Peilin Chen

11/7/2017

dataframe is a collection of different class of data

Create a dataframe

```
quarters<-rep(paste('q',seq=(1:4),sep=''),3)
year<-rep(seq(2000,2002),each=4)
mydataframe<-cbind.data.frame(year,quarters)
mydataframe
```

```
##    year quarters
## 1  2000       q1
## 2  2000       q2
## 3  2000       q3
## 4  2000       q4
## 5  2001       q1
## 6  2001       q2
## 7  2001       q3
## 8  2001       q4
## 9  2002       q1
## 10 2002       q2
## 11 2002       q3
## 12 2002       q4
```

access each column by calling the column name

```
mydataframe$year
```

```
##  [1] 2000 2000 2000 2000 2001 2001 2001 2001 2002 2002 2002 2002
```

```
mydataframe$quarters
```

```
##  [1] q1 q2 q3 q4 q1 q2 q3 q4 q1 q2 q3 q4
## Levels: q1 q2 q3 q4
```

access each column by indexing location

```
mydataframe[,1]
```

```
##  [1] 2000 2000 2000 2000 2001 2001 2001 2001 2002 2002 2002 2002
```

```
mydataframe[,2]
```

```
##  [1] q1 q2 q3 q4 q1 q2 q3 q4 q1 q2 q3 q4
## Levels: q1 q2 q3 q4
```

access rows by indexing location

```
mydataframe[1:5,] # first 5 rows
```

```
##   year quarters
## 1 2000       q1
## 2 2000       q2
## 3 2000       q3
## 4 2000       q4
## 5 2001       q1
```

select rows based on column values, using slicing, or subset() function

```
mydataframe[quarters=='q1',]
```

```
##   year quarters
## 1 2000       q1
## 5 2001       q1
## 9 2002       q1
```

```
subset(mydataframe,quarters=='q1')
```

```
##   year quarters
## 1 2000       q1
## 5 2001       q1
## 9 2002       q1
```

using logic operator |(or) &(and)

```
mydataframe[quarters=='q1'|quarters=='q2',]
```

```
##    year quarters
## 1  2000       q1
## 2  2000       q2
## 5  2001       q1
## 6  2001       q2
## 9  2002       q1
## 10 2002       q2
```

```
mydataframe[quarters=='q1'&year==2000,]
```

```
##   year quarters
## 1 2000       q1
```

how to change row / column names:

```
row.names(mydataframe)
```

```
##  [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12"
```

```
row.names(mydataframe)<-letters[1:12]
row.names(mydataframe)
```

```
##  [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j" "k" "l"
```

```
colnames(mydataframe)
```

```
## [1] "year"     "quarters"
```

```
colnames(mydataframe)[1]<-'YEAR' #change the first column's name
mydataframe
```

```
##   YEAR quarters
## a 2000       q1
## b 2000       q2
## c 2000       q3
## d 2000       q4
## e 2001       q1
## f 2001       q2
## g 2001       q3
## h 2001       q4
## i 2002       q1
## j 2002       q2
## k 2002       q3
## l 2002       q4
```

create a new column in a dataframe:

```
set.seed(1)
mydataframe$newcolumn1<-rnorm(12,mean = 0,sd =1)
mydataframe$newcolumn2<-runif(12,min = 0,max = 1)
mydataframe
```

```
##   YEAR quarters newcolumn1 newcolumn2
## a 2000       q1 -0.6264538 0.26722067
## b 2000       q2  0.1836433 0.38611409
## c 2000       q3 -0.8356286 0.01339033
## d 2000       q4  1.5952808 0.38238796
## e 2001       q1  0.3295078 0.86969085
## f 2001       q2 -0.8204684 0.34034900
## g 2001       q3  0.4874291 0.48208012
## h 2001       q4  0.7383247 0.59956583
## i 2002       q1  0.5757814 0.49354131
## j 2002       q2 -0.3053884 0.18621760
## k 2002       q3  1.5117812 0.82737332
## l 2002       q4  0.3898432 0.66846674
```

if our newcolumn3 is newcolumn1+newcolumn2

```
mydataframe$newcolumn3<-mydataframe$newcolumn1+mydataframe$newcolumn2
mydataframe
```

```
##   YEAR quarters newcolumn1 newcolumn2 newcolumn3
## a 2000       q1 -0.6264538 0.26722067 -0.3592331
## b 2000       q2  0.1836433 0.38611409  0.5697574
## c 2000       q3 -0.8356286 0.01339033 -0.8222383
## d 2000       q4  1.5952808 0.38238796  1.9776688
## e 2001       q1  0.3295078 0.86969085  1.1991986
## f 2001       q2 -0.8204684 0.34034900 -0.4801194
## g 2001       q3  0.4874291 0.48208012  0.9695092
## h 2001       q4  0.7383247 0.59956583  1.3378905
## i 2002       q1  0.5757814 0.49354131  1.0693227
## j 2002       q2 -0.3053884 0.18621760 -0.1191708
## k 2002       q3  1.5117812 0.82737332  2.3391545
## l 2002       q4  0.3898432 0.66846674  1.0583100
```

if our newcolumn4 is an index column : when newcolumn1>0, it is 1, otherwise 0

```
mydataframe$newcolumn5<- ifelse(mydataframe$newcolumn1>0,1,0)
mydataframe
```

```
##   YEAR quarters newcolumn1 newcolumn2 newcolumn3 newcolumn5
## a 2000       q1 -0.6264538 0.26722067 -0.3592331          0
## b 2000       q2  0.1836433 0.38611409  0.5697574          1
## c 2000       q3 -0.8356286 0.01339033 -0.8222383          0
## d 2000       q4  1.5952808 0.38238796  1.9776688          1
## e 2001       q1  0.3295078 0.86969085  1.1991986          1
## f 2001       q2 -0.8204684 0.34034900 -0.4801194          0
## g 2001       q3  0.4874291 0.48208012  0.9695092          1
## h 2001       q4  0.7383247 0.59956583  1.3378905          1
## i 2002       q1  0.5757814 0.49354131  1.0693227          1
## j 2002       q2 -0.3053884 0.18621760 -0.1191708          0
## k 2002       q3  1.5117812 0.82737332  2.3391545          1
## l 2002       q4  0.3898432 0.66846674  1.0583100          1
```

Here we introduce how to two data manipulation packages in R

dplyr tidyr #### install packages and library the packages

```
install.packages('dplyr')
install.packages('tidyr')
library.packages('dplyr')
library.packages('tidyr')
```

what if we have a list of packages needs to install and library? Someone posted in Github https://gist.github.com/stevenworthington/3178163

```
ipak <- function(pkg){
    new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
    if (length(new.pkg))
```

```
        install.packages(new.pkg, dependencies = TRUE)
    sapply(pkg, require, character.only = TRUE)
}

# usage
packages_toinstall <- c("ggplot2", "plyr", "reshape2", "RColorBrewer", "scales", "grid",'dplyr','tidyr'
ipak(packages_toinstall)
```

library multiple packages

```
packages_toload<-c("ggplot2", "plyr", "reshape2", "RColorBrewer", "scales", "grid")
lapply(packages_toload, require, character.only = TRUE)
```

```
packages_toload<-c('dplyr','tidyr')
lapply(packages_toload,require,character.only=TRUE)
```

```
## [[1]]
## [1] TRUE
##
## [[2]]
## [1] TRUE
```

piping

using the output of the privious step as the input of the current step

```
input %>% function1 ()  %>% function2 ()

#equal to : function2(function1(input))
```

```
iris
```

```
##     Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
## 1            5.1         3.5          1.4         0.2     setosa
## 2            4.9         3.0          1.4         0.2     setosa
## 3            4.7         3.2          1.3         0.2     setosa
## 4            4.6         3.1          1.5         0.2     setosa
## 5            5.0         3.6          1.4         0.2     setosa
## 6            5.4         3.9          1.7         0.4     setosa
## 7            4.6         3.4          1.4         0.3     setosa
## 8            5.0         3.4          1.5         0.2     setosa
## 9            4.4         2.9          1.4         0.2     setosa
## 10           4.9         3.1          1.5         0.1     setosa
## 11           5.4         3.7          1.5         0.2     setosa
## 12           4.8         3.4          1.6         0.2     setosa
## 13           4.8         3.0          1.4         0.1     setosa
## 14           4.3         3.0          1.1         0.1     setosa
## 15           5.8         4.0          1.2         0.2     setosa
## 16           5.7         4.4          1.5         0.4     setosa
## 17           5.4         3.9          1.3         0.4     setosa
## 18           5.1         3.5          1.4         0.3     setosa
## 19           5.7         3.8          1.7         0.3     setosa
## 20           5.1         3.8          1.5         0.3     setosa
## 21           5.4         3.4          1.7         0.2     setosa
## 22           5.1         3.7          1.5         0.4     setosa
## 23           4.6         3.6          1.0         0.2     setosa
## 24           5.1         3.3          1.7         0.5     setosa
## 25           4.8         3.4          1.9         0.2     setosa
## 26           5.0         3.0          1.6         0.2     setosa
## 27           5.0         3.4          1.6         0.4     setosa
## 28           5.2         3.5          1.5         0.2     setosa
## 29           5.2         3.4          1.4         0.2     setosa
## 30           4.7         3.2          1.6         0.2     setosa
## 31           4.8         3.1          1.6         0.2     setosa
## 32           5.4         3.4          1.5         0.4     setosa
## 33           5.2         4.1          1.5         0.1     setosa
## 34           5.5         4.2          1.4         0.2     setosa
## 35           4.9         3.1          1.5         0.2     setosa
## 36           5.0         3.2          1.2         0.2     setosa
## 37           5.5         3.5          1.3         0.2     setosa
## 38           4.9         3.6          1.4         0.1     setosa
## 39           4.4         3.0          1.3         0.2     setosa
## 40           5.1         3.4          1.5         0.2     setosa
## 41           5.0         3.5          1.3         0.3     setosa
## 42           4.5         2.3          1.3         0.3     setosa
## 43           4.4         3.2          1.3         0.2     setosa
## 44           5.0         3.5          1.6         0.6     setosa
## 45           5.1         3.8          1.9         0.4     setosa
## 46           4.8         3.0          1.4         0.3     setosa
## 47           5.1         3.8          1.6         0.2     setosa
## 48           4.6         3.2          1.4         0.2     setosa
## 49           5.3         3.7          1.5         0.2     setosa
## 50           5.0         3.3          1.4         0.2     setosa
## 51           7.0         3.2          4.7         1.4 versicolor
## 52           6.4         3.2          4.5         1.5 versicolor
## 53           6.9         3.1          4.9         1.5 versicolor
## 54           5.5         2.3          4.0         1.3 versicolor
## 55           6.5         2.8          4.6         1.5 versicolor
## 56           5.7         2.8          4.5         1.3 versicolor
## 57           6.3         3.3          4.7         1.6 versicolor
## 58           4.9         2.4          3.3         1.0 versicolor
## 59           6.6         2.9          4.6         1.3 versicolor
## 60           5.2         2.7          3.9         1.4 versicolor
## 61           5.0         2.0          3.5         1.0 versicolor
## 62           5.9         3.0          4.2         1.5 versicolor
## 63           6.0         2.2          4.0         1.0 versicolor
## 64           6.1         2.9          4.7         1.4 versicolor
## 65           5.6         2.9          3.6         1.3 versicolor
## 66           6.7         3.1          4.4         1.4 versicolor
## 67           5.6         3.0          4.5         1.5 versicolor
## 68           5.8         2.7          4.1         1.0 versicolor
## 69           6.2         2.2          4.5         1.5 versicolor
## 70           5.6         2.5          3.9         1.1 versicolor
## 71           5.9         3.2          4.8         1.8 versicolor
## 72           6.1         2.8          4.0         1.3 versicolor
## 73           6.3         2.5          4.9         1.5 versicolor
## 74           6.1         2.8          4.7         1.2 versicolor
## 75           6.4         2.9          4.3         1.3 versicolor
## 76           6.6         3.0          4.4         1.4 versicolor
## 77           6.8         2.8          4.8         1.4 versicolor
## 78           6.7         3.0          5.0         1.7 versicolor
## 79           6.0         2.9          4.5         1.5 versicolor
## 80           5.7         2.6          3.5         1.0 versicolor
## 81           5.5         2.4          3.8         1.1 versicolor
## 82           5.5         2.4          3.7         1.0 versicolor
## 83           5.8         2.7          3.9         1.2 versicolor
## 84           6.0         2.7          5.1         1.6 versicolor
## 85           5.4         3.0          4.5         1.5 versicolor
## 86           6.0         3.4          4.5         1.6 versicolor
## 87           6.7         3.1          4.7         1.5 versicolor
## 88           6.3         2.3          4.4         1.3 versicolor
## 89           5.6         3.0          4.1         1.3 versicolor
## 90           5.5         2.5          4.0         1.3 versicolor
## 91           5.5         2.6          4.4         1.2 versicolor
## 92           6.1         3.0          4.6         1.4 versicolor
## 93           5.8         2.6          4.0         1.2 versicolor
## 94           5.0         2.3          3.3         1.0 versicolor
## 95           5.6         2.7          4.2         1.3 versicolor
## 96           5.7         3.0          4.2         1.2 versicolor
## 97           5.7         2.9          4.2         1.3 versicolor
## 98           6.2         2.9          4.3         1.3 versicolor
## 99           5.1         2.5          3.0         1.1 versicolor
## 100          5.7         2.8          4.1         1.3 versicolor
## 101          6.3         3.3          6.0         2.5  virginica
## 102          5.8         2.7          5.1         1.9  virginica
## 103          7.1         3.0          5.9         2.1  virginica
## 104          6.3         2.9          5.6         1.8  virginica
## 105          6.5         3.0          5.8         2.2  virginica
## 106          7.6         3.0          6.6         2.1  virginica
## 107          4.9         2.5          4.5         1.7  virginica
## 108          7.3         2.9          6.3         1.8  virginica
## 109          6.7         2.5          5.8         1.8  virginica
## 110          7.2         3.6          6.1         2.5  virginica
## 111          6.5         3.2          5.1         2.0  virginica
## 112          6.4         2.7          5.3         1.9  virginica
## 113          6.8         3.0          5.5         2.1  virginica
## 114          5.7         2.5          5.0         2.0  virginica
## 115          5.8         2.8          5.1         2.4  virginica
## 116          6.4         3.2          5.3         2.3  virginica
## 117          6.5         3.0          5.5         1.8  virginica
## 118          7.7         3.8          6.7         2.2  virginica
## 119          7.7         2.6          6.9         2.3  virginica
## 120          6.0         2.2          5.0         1.5  virginica
## 121          6.9         3.2          5.7         2.3  virginica
## 122          5.6         2.8          4.9         2.0  virginica
## 123          7.7         2.8          6.7         2.0  virginica
## 124          6.3         2.7          4.9         1.8  virginica
## 125          6.7         3.3          5.7         2.1  virginica
## 126          7.2         3.2          6.0         1.8  virginica
## 127          6.2         2.8          4.8         1.8  virginica
## 128          6.1         3.0          4.9         1.8  virginica
## 129          6.4         2.8          5.6         2.1  virginica
## 130          7.2         3.0          5.8         1.6  virginica
## 131          7.4         2.8          6.1         1.9  virginica
## 132          7.9         3.8          6.4         2.0  virginica
## 133          6.4         2.8          5.6         2.2  virginica
## 134          6.3         2.8          5.1         1.5  virginica
## 135          6.1         2.6          5.6         1.4  virginica
## 136          7.7         3.0          6.1         2.3  virginica
## 137          6.3         3.4          5.6         2.4  virginica
## 138          6.4         3.1          5.5         1.8  virginica
## 139          6.0         3.0          4.8         1.8  virginica
## 140          6.9         3.1          5.4         2.1  virginica
## 141          6.7         3.1          5.6         2.4  virginica
## 142          6.9         3.1          5.1         2.3  virginica
## 143          5.8         2.7          5.1         1.9  virginica
## 144          6.8         3.2          5.9         2.3  virginica
## 145          6.7         3.3          5.7         2.5  virginica
## 146          6.7         3.0          5.2         2.3  virginica
## 147          6.3         2.5          5.0         1.9  virginica
## 148          6.5         3.0          5.2         2.0  virginica
## 149          6.2         3.4          5.4         2.3  virginica
## 150          5.9         3.0          5.1         1.8  virginica
```

```
iris %>% group_by(Species)
```

```
## # A tibble: 150 x 5
## # Groups:   Species [3]
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##           <dbl>       <dbl>        <dbl>       <dbl>  <fctr>
##  1          5.1         3.5          1.4         0.2  setosa
##  2          4.9         3.0          1.4         0.2  setosa
##  3          4.7         3.2          1.3         0.2  setosa
##  4          4.6         3.1          1.5         0.2  setosa
##  5          5.0         3.6          1.4         0.2  setosa
##  6          5.4         3.9          1.7         0.4  setosa
##  7          4.6         3.4          1.4         0.3  setosa
##  8          5.0         3.4          1.5         0.2  setosa
##  9          4.4         2.9          1.4         0.2  setosa
## 10          4.9         3.1          1.5         0.1  setosa
## # ... with 140 more rows
```

using slice

```
iris%>%slice(1:5)
```

```
## # A tibble: 5 x 5
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##          <dbl>       <dbl>        <dbl>       <dbl>  <fctr>
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
```

using summarize

```
iris%>%group_by(Species)%>%summarize(n_obs=n(),min_length=min(Sepal.Length),max_length=max(Sepal.Length
```

```
## # A tibble: 3 x 6
##      Species n_obs min_length max_length min_petal_width max_petal_width
##       <fctr> <int>      <dbl>      <dbl>           <dbl>           <dbl>
## 1     setosa    50        4.3        5.8             0.1             0.6
## 2 versicolor    50        4.9        7.0             1.0             1.8
## 3  virginica    50        4.9        7.9             1.4             2.5
```

using select

```
iris%>%group_by(Species)%>%summarize(n_obs=n(),min_length=min(Sepal.Length),max_length=max(Sepal.Length
```

```
## # A tibble: 3 x 2
##      Species n_obs
##       <fctr> <int>
## 1     setosa    50
## 2 versicolor    50
## 3  virginica    50
```

There are many other useful methods you can call from the packages for example:

```
rename,
mutate
filter,
left_join,
right_join,
inner_join,
full_join ......
```