Week 3

Aditya

6/15/2020

Subsetting and sorting

Subsetting

Creating Dataset

```
set.seed(13435)
X <- data.frame("var1" = sample(1:5),"var2" = sample(6:10), "var3" = sample(11:15))
X <- X[sample(1:5),]; X$var2[c(1,3)] = NA
X
```

```
##   var1 var2 var3
## 5    2   NA   11
## 4    4   10   12
## 1    3   NA   14
## 2    1    7   15
## 3    5    6   13
```

Subset an specific column by doing

```
X[,1]
```

```
## [1] 2 4 3 1 5
```

or, by doing with column name

```
X[,'var1']
```

```
## [1] 2 4 3 1 5
```

or, i can select specific rows, and columns by

```
X[1:3,'var2']
```

```
## [1] NA 10 NA
```

or with columns index

```
X[c(1,3,5),c(3,1)]
```

```
##   var3 var1
## 5   11    2
## 1   14    3
## 3   13    5
```

Logicals

Selecting all rows which var1 value is less than or equal to 3 and var3 value is greater than 11

```
X[(X$var1<=3 & X$var3>11),]
```

```
##   var1 var2 var3
## 1    3   NA   14
## 2    1    7   15
```

similarly,

```
X[(X$var1 <=3 | X$var3>15),]
```

```
##   var1 var2 var3
## 5    2   NA   11
## 1    3   NA   14
## 2    1    7   15
```

Dealing with missing values

which command gets rid of the missing values

```
X[which(X$var2 > 8),]
```

```
##   var1 var2 var3
## 4    4   10   12
```

Sorting

```
sort(X$var1)
```

```
## [1] 1 2 3 4 5
```

by default it's in increasing order. we can sort it in decreasing order to, by passing decreasing parameter

```
sort(X$var1, decreasing = T)
```

```
## [1] 5 4 3 2 1
```

while sorting with NA values, they are put first by default. If selected na.last = TRUE, then they will be put last

```
sort(X$var2,na.last = T)
```

```
## [1]  6  7 10 NA NA
```

Ordering

We can actually order a data frame by one or more particular columns

```
X[order(X$var1),]
```

```
##   var1 var2 var3
## 2    1    7   15
## 5    2   NA   11
## 1    3   NA   14
## 4    4   10   12
## 3    5    6   13
```

sorting with multiple variables. In these case, sort first with value 1, if there is a tie then, sort by value 2

```
X[order(X$var2,X$var3,na.last = F),]
```

```
##   var1 var2 var3
## 5    2   NA   11
## 1    3   NA   14
## 3    5    6   13
## 2    1    7   15
## 4    4   10   12
```

Ordering with plyr

We can do the same thing with the plyr library

```
library(plyr)
arrange(X,var1)
```

```
##   var1 var2 var3
## 1    1    7   15
## 2    2   NA   11
## 3    3   NA   14
## 4    4   10   12
## 5    5    6   13
```

or to arrange in decreasing order

```
arrange(X,desc(var1))
```

```
##   var1 var2 var3
## 1    5    6   13
## 2    4   10   12
## 3    3   NA   14
## 4    2   NA   11
## 5    1    7   15
```

Adding rows and columns in the data frame

```
X$var4 <- rnorm(5)
X
```

```
##   var1 var2 var3       var4
## 5    2   NA   11 -0.4150458
## 4    4   10   12  2.5437602
## 1    3   NA   14  1.5545298
## 2    1    7   15 -0.6192328
## 3    5    6   13 -0.9261035
```

also can add with the cbind command

```
Y <- cbind(X,hakunaMatata=rnorm(5))
Y
```

```
##   var1 var2 var3       var4 hakunaMatata
## 5    2   NA   11 -0.4150458  -0.66549949
## 4    4   10   12  2.5437602  -0.02166735
## 1    3   NA   14  1.5545298  -0.17411953
## 2    1    7   15 -0.6192328   0.23900438
## 3    5    6   13 -0.9261035  -1.83245959
```

can add a new row with the rbind command

```
Y <- rbind(Y, rnorm(5))
Y
```

```
##          var1      var2      var3       var4 hakunaMatata
## 5  2.00000000        NA 11.000000 -0.4150458  -0.66549949
## 4  4.00000000 10.000000 12.000000  2.5437602  -0.02166735
## 1  3.00000000        NA 14.000000  1.5545298  -0.17411953
## 2  1.00000000  7.000000 15.000000 -0.6192328   0.23900438
## 3  5.00000000  6.000000 13.000000 -0.9261035  -1.83245959
## 6 -0.03718739 -0.440517 -1.448264 -0.5182457   0.75852718
```

Summarizing the data

Downloading the dataset

```
if(!file.exists('restaurants.csv'))
{
  fileURL <-'https://data.baltimorecity.gov/api/views/k5ry-ef3g/rows.csv?accessType=DOWNLOAD'
  download.file(fileURL, destfile = 'restaurants.csv',method = 'curl')
}
restData <- read.csv('restaurants.csv')
```

Checking the data with head. By default n = 6, but we can set n to view the data.

```
head(restData, n =3)
```

```
##    name zipCode neighborhood councilDistrict policeDistrict
## 1   410   21206    Frankford               2   NORTHEASTERN
## 2  1919   21231  Fells Point               1   SOUTHEASTERN
## 3 SAUTE   21224       Canton               1   SOUTHEASTERN
##                        Location.1 X2010.Census.Neighborhoods
## 1 4509 BELAIR ROAD\nBaltimore, MD                         NA
## 2    1919 FLEET ST\nBaltimore, MD                         NA
## 3   2844 HUDSON ST\nBaltimore, MD                         NA
##   X2010.Census.Wards.Precincts Zip.Codes
## 1                           NA        NA
## 2                           NA        NA
## 3                           NA        NA
```

To view the column names, there are two methods.

```
names(restData)
```

```
## [1] "name"                          "zipCode"
## [3] "neighborhood"                  "councilDistrict"
## [5] "policeDistrict"                "Location.1"
## [7] "X2010.Census.Neighborhoods"    "X2010.Census.Wards.Precincts"
## [9] "Zip.Codes"
```

```
# or,
colnames(restData)
```

```
## [1] "name"                          "zipCode"
## [3] "neighborhood"                  "councilDistrict"
## [5] "policeDistrict"                "Location.1"
## [7] "X2010.Census.Neighborhoods"    "X2010.Census.Wards.Precincts"
## [9] "Zip.Codes"
```

Both of them will return a list.

Also we can view data, from tail of the dataset

```
tail(restData, n = 5)
```

```
##                          name zipCode    neighborhood councilDistrict
## 1323 ZEN WEST ROADSIDE CANTINA   21212        Rosebank               4
## 1324                  ZIASCOS   21231 Washington Hill               1
## 1325             ZINK'S CAFÂ\220   21213   Belair-Edison              13
## 1326             ZISSIMOS BAR   21211         Hampden               7
## 1327                    ZORBAS   21224       Greektown               2
##      policeDistrict                       Location.1 X2010.Census.Neighborhoods
## 1323       NORTHERN      5916 YORK RD\nBaltimore, MD                         NA
## 1324   SOUTHEASTERN     1313 PRATT ST\nBaltimore, MD                         NA
## 1325   NORTHEASTERN 3300 LAWNVIEW AVE\nBaltimore, MD                         NA
## 1326       NORTHERN      1023 36TH ST\nBaltimore, MD                         NA
```

```
## 1327   SOUTHEASTERN  4710 EASTERN Ave\nBaltimore, MD                          NA
##      X2010.Census.Wards.Precincts Zip.Codes
## 1323                           NA        NA
## 1324                           NA        NA
## 1325                           NA        NA
## 1326                           NA        NA
## 1327                           NA        NA
```

Making Summary

```
summary(restData)
```

```
##      name              zipCode       neighborhood       councilDistrict
##  Length:1327        Min.   :-21226   Length:1327        Min.   : 1.000
##  Class :character   1st Qu.: 21202   Class :character   1st Qu.: 2.000
##  Mode  :character   Median : 21218   Mode  :character   Median : 9.000
##                     Mean   : 21185                      Mean   : 7.191
##                     3rd Qu.: 21226                      3rd Qu.:11.000
##                     Max.   : 21287                      Max.   :14.000
##  policeDistrict      Location.1        X2010.Census.Neighborhoods
##  Length:1327        Length:1327        Mode:logical
##  Class :character   Class :character   NA's:1327
##  Mode  :character   Mode  :character
##
##
##
##  X2010.Census.Wards.Precincts Zip.Codes
##  Mode:logical                 Mode:logical
##  NA's:1327                    NA's:1327
##
##
##
##
```

We can use str command to see more information about the data frame. The depth, type, and demo.

```
str(restData)
```

```
## 'data.frame':    1327 obs. of  9 variables:
##  $ name                       : chr  "410" "1919" "SAUTE" "#1 CHINESE KITCHEN" ...
##  $ zipCode                    : int  21206 21231 21224 21211 21223 21218 21205 21211 21205 21231 ..
##  $ neighborhood               : chr  "Frankford" "Fells Point" "Canton" "Hampden" ...
##  $ councilDistrict            : int  2 1 1 14 9 14 13 7 13 1 ...
##  $ policeDistrict             : chr  "NORTHEASTERN" "SOUTHEASTERN" "SOUTHEASTERN" "NORTHERN" ...
##  $ Location.1                 : chr  "4509 BELAIR ROAD\nBaltimore, MD" "1919 FLEET ST\nBaltimore, M
##  $ X2010.Census.Neighborhoods : logi  NA NA NA NA NA NA ...
##  $ X2010.Census.Wards.Precincts: logi  NA NA NA NA NA NA ...
##  $ Zip.Codes                  : logi  NA NA NA NA NA NA ...
```

Quantiles

Quantiles on quantitative variables

```
quantile(restData$councilDistrict, na.rm = T)
```

```
##   0%  25%  50%  75% 100%
##    1    2    9   11   14
```

Quantile values with different probabilities

```
quantile(restData$councilDistrict, prob = c(0.5,0.75, 0.9))
```

```
## 50% 75% 90%
##   9  11  12
```

Tables

We can also make table

```
table(restData$zipCode, useNA = 'ifany')
```

```
##
## -21226  21201  21202  21205  21206  21207  21208  21209  21210  21211  21212
##      1    136    201     27     30      4      1      8     23     41     28
##  21213  21214  21215  21216  21217  21218  21220  21222  21223  21224  21225
##     31     17     54     10     32     69      1      7     56    199     19
##  21226  21227  21229  21230  21231  21234  21237  21239  21251  21287
##     18      4     13    156    127      7      1      3      2      1
```

Here, we can see how many zipcode are there with what value. For example 21201 has 136 entries and so on. The parameter useNA will NA value in the end if there is any NA value, because by default it doesn't count the missing values

We can also make 2D tables.

```
table(restData$councilDistrict, restData$zipCode)
```

```
##
##      -21226 21201 21202 21205 21206 21207 21208 21209 21210 21211 21212 21213
##   1       0     0    37     0     0     0     0     0     0     0     0     2
##   2       0     0     0     3    27     0     0     0     0     0     0     0
##   3       0     0     0     0     0     0     0     0     0     0     0     2
##   4       0     0     0     0     0     0     0     0     0     0    27     0
##   5       0     0     0     0     0     3     0     6     0     0     0     0
##   6       0     0     0     0     0     0     0     1    19     0     0     0
##   7       0     0     0     0     0     0     0     1     0    27     0     0
##   8       0     0     0     0     0     1     0     0     0     0     0     0
##   9       0     1     0     0     0     0     0     0     0     0     0     0
##   10      1     0     1     0     0     0     0     0     0     0     0     0
##   11      0   115   139     0     0     0     1     0     0     0     1     0
```

```
##   12      0    20    24     4     0     0     0     0     0     0     0    13
##   13      0     0     0    20     3     0     0     0     0     0     0    13
##   14      0     0     0     0     0     0     0     0     4    14     0     1
##
##      21214 21215 21216 21217 21218 21220 21222 21223 21224 21225 21226 21227
##   1      0     0     0     0     0     0     7     0   140     1     0     0
##   2      0     0     0     0     0     0     0     0    54     0     0     0
##   3     17     0     0     0     3     0     0     0     0     0     0     1
##   4      0     0     0     0     0     0     0     0     0     0     0     0
##   5      0    31     0     0     0     0     0     0     0     0     0     0
##   6      0    15     1     0     0     0     0     0     0     0     0     0
##   7      0     6     7    15     6     0     0     0     0     0     0     0
##   8      0     0     0     0     0     0     0     2     0     0     0     2
##   9      0     0     2     8     0     0     0    53     0     0     0     0
##   10     0     0     0     0     0     1     0     0     0    18    18     0
##   11     0     0     0     9     0     0     0     1     0     0     0     0
##   12     0     0     0     0    26     0     0     0     0     0     0     0
##   13     0     1     0     0     0     0     0     0     5     0     0     1
##   14     0     1     0     0    34     0     0     0     0     0     0     0
##
##      21229 21230 21231 21234 21237 21239 21251 21287
##   1      0     1   124     0     0     0     0     0
##   2      0     0     0     0     1     0     0     0
##   3      0     0     0     7     0     0     2     0
##   4      0     0     0     0     0     3     0     0
##   5      0     0     0     0     0     0     0     0
##   6      0     0     0     0     0     0     0     0
##   7      0     0     0     0     0     0     0     0
##   8     13     0     0     0     0     0     0     0
##   9      0    11     0     0     0     0     0     0
##   10     0   133     0     0     0     0     0     0
##   11     0    11     0     0     0     0     0     0
##   12     0     0     2     0     0     0     0     0
##   13     0     0     1     0     0     0     0     1
##   14     0     0     0     0     0     0     0     0
```

We can actually view relationship between different data using this 2D table data.

Missing values

```
sum(is.na(restData$councilDistrict))
```

```
## [1] 0
```

or we can use any() function

```
any(is.na(restData$Location.1))
```

```
## [1] FALSE
```

the any() function checks if there is any FALSE in the given **logical array**

```
all(restData$zipCode > 0)
```

```
## [1] FALSE
```

There was a negative zipcode. Remember?

Row and column sums

```
colSums(is.na(restData))
```

```
##                          name                       zipCode
##                             0                             0
##                  neighborhood               councilDistrict
##                             0                             0
##                policeDistrict                   Location.1
##                             0                             0
##   X2010.Census.Neighborhoods X2010.Census.Wards.Precincts
##                          1327                          1327
##                     Zip.Codes
##                          1327
```

Check if there are any missing values in the entire dataset

```
sum(is.na(restData))
```

```
## [1] 3981
```

Values with specific characteristics

```
table(restData$zipCode %in% c('21212') )
```

```
##
## FALSE  TRUE
##  1299    28
```

```
table(restData$zipCode %in% c('21212','21213'))
```

```
##
## FALSE  TRUE
##  1268    59
```

We can also get subset of the data using same procedure

```
restData[restData$zipCode %in% c('21212','21213'),  ]
```

```
##                                      name zipCode                neighborhood
## 29                       BAY ATLANTIC CLUB   21212                    Downtown
## 39                             BERMUDA BAR   21213                Broadway East
## 92                               ATWATER'S   21212   Chinquapin Park-Belvedere
## 111            BALTIMORE ESTONIAN SOCIETY   21213          South Clifton Park
## 187                              CAFE ZEN   21212                    Rosebank
## 220                  CERIELLO FINE FOODS   21212   Chinquapin Park-Belvedere
## 266    CLIFTON PARK GOLF COURSE SNACK BAR   21213                 Darley Park
## 276                CLUB HOUSE BAR & GRILL   21213 Orangeville Industrial Area
## 289                CLUBHOUSE BAR & GRILL   21213 Orangeville Industrial Area
## 291                             COCKY LOU'S   21213                Broadway East
## 362       DREAM TAVERN, CARRIBEAN  U.S.A.   21213                Broadway East
## 373                          DUNKIN DONUTS   21212                    Homeland
## 383        EASTSIDE  SPORTS  SOCIAL  CLUB   21213                Broadway East
## 417                       FIELDS OLD TRAIL   21212                   Mid-Govans
## 475                             GRAND CRU   21212   Chinquapin Park-Belvedere
## 545                           RANDY'S BAR   21213                Broadway East
## 604     MURPHY'S NEIGHBORHOOD BAR & GRILL   21212                   Mid-Govans
## 616                                NEOPOL   21212   Chinquapin Park-Belvedere
## 620             NEW CLUB THUNDERBIRD INC.   21213                  Middle East
## 626                   NEW MAYFIELD, INC.   21213                Belair-Edison
## 678                           IKAN SEAFOOD   21212   Chinquapin Park-Belvedere
## 711                           KAY-CEE CLUB   21212                    Homeland
## 763                                LA'RAE   21213                      Oliver
## 777                 LEMONGRASS BALTIMORE   21213                 Little Italy
## 779                  LEN'S SANDWICH SHOP   21213                Broadway East
## 845                            MCDONALD'S   21213          South Clifton Park
## 852                            MCDONALD'S   21212              Radnor-Winston
## 873                 NEW REX LIQUORS,INC.   21212                 Wilson Park
## 895                             OK TAVERN   21213                Biddle Street
## 919                           PANERA BREAD   21212                  Lake Walker
## 940                    PEIWEI ASIAN DINER   21212                   Cedarcroft
## 949                  PERGUSA ENTERPRISES   21212                    Rosebank
## 957              PHANTOM'S BAR AND GRILL   21213                Belair-Edison
## 976          POPEYES FAMOUS FRIED CHICKEN   21212               Winston-Govans
## 994                          ROBBIE'S NEST   21213                Broadway East
## 1017                          RUTLAND BAR   21213                Broadway East
## 1018                       RYAN'S DAUGHTER   21212   Chinquapin Park-Belvedere
## 1022         saigon remembered restaurant   21212                   Mid-Govans
## 1053                SHIRLEY'S  HONEY HOLE   21213                Broadway East
## 1120                     STEEPLE CHASE II   21213                Biddle Street
## 1122                                SUBWAY   21213                      Oliver
## 1153                              TAM-TAM   21212                   Mid-Govans
## 1155                                TASTE   21212                   Mid-Govans
## 1159                         TAYLORS EAST   21213                        Berea
## 1186                THE EDGE BAR & LOUNGE   21213                Broadway East
## 1187 THE EDGE BAR & LOUNGE - KITCHEN AREA   21213                Broadway East
## 1198               THE HOLLOW BAR & GRILL   21212                    Rosebank
## 1209            THE NEW BUCKETT'S LOUNGE   21213                Broadway East
## 1232                         THREE  ACE'S   21213                Belair-Edison
## 1246                TORAIN'S  HIDE-A-WAY   21213                Broadway East
## 1259                   TSUNAMI BALTIMORE   21213                 Little Italy
## 1287                         VITO'S PIZZA   21212                   Cedarcroft
## 1298 WENDY'S OLD FASHIONED HAMBURGERS #96   21212                    Homeland
```

```
## 1304                 WHITTEN'S  (4502-04)   21213          Claremont-Freedom
## 1312                           wozi lounge   21212                   Guilford
## 1319          YETI RESTAURANT & CARRYOUT   21212                   Rosebank
## 1320                    YORK CLUB TAVERN   21212                   Homeland
## 1323           ZEN WEST ROADSIDE CANTINA   21212                   Rosebank
## 1325                       ZINK'S CAFÂ\220   21213             Belair-Edison
##      councilDistrict policeDistrict                        Location.1
## 29                11        CENTRAL      206 REDWOOD ST\nBaltimore, MD
## 39                12        EASTERN      1801 NORTH AVE\nBaltimore, MD
## 92                 4       NORTHERN   529 BELVEDERE AVE\nBaltimore, MD
## 111               12        EASTERN      1932 BELAIR RD\nBaltimore, MD
## 187                4       NORTHERN   438 BELVEDERE AVE\nBaltimore, MD
## 220                4       NORTHERN   529 BELVEDERE AVE\nBaltimore, MD
## 266               14   NORTHEASTERN       2701 ST LO DR\nBaltimore, MD
## 276               13        EASTERN     4217 ERDMAN AVE\nBaltimore, MD
## 289               13        EASTERN     4217 ERDMAN AVE\nBaltimore, MD
## 291               12        EASTERN      2101 NORTH AVE\nBaltimore, MD
## 362               13        EASTERN  2300 LAFAYETTE AVE\nBaltimore, MD
## 373                4       NORTHERN        5422 YORK RD\nBaltimore, MD
## 383               13        EASTERN 1203 COLLINGTON AVE\nBaltimore, MD
## 417                4       NORTHERN        5723 YORK RD\nBaltimore, MD
## 475                4       NORTHERN   527 BELVEDERE AVE\nBaltimore, MD
## 545               12        EASTERN      2135 NORTH AVE\nBaltimore, MD
## 604                4       NORTHERN        5847 YORK RD\nBaltimore, MD
## 616                4       NORTHERN   529 BELVEDERE AVE\nBaltimore, MD
## 620               13        EASTERN       2201 CHASE ST\nBaltimore, MD
## 626               13   NORTHEASTERN      3349 BELAIR RD\nBaltimore, MD
## 678                4       NORTHERN   529 BELVEDERE AVE\nBaltimore, MD
## 711                4       NORTHERN    201 HOMELAND AVE\nBaltimore, MD
## 763               12        EASTERN     1000 HOFFMAN ST\nBaltimore, MD
## 777                1   SOUTHEASTERN    1300 BANK STREET\nBaltimore, MD
## 779               12        EASTERN  1500 WASHINGTON ST\nBaltimore, MD
## 845               12        EASTERN       2001 BROADWAY\nBaltimore, MD
## 852                4       NORTHERN        5100 YORK RD\nBaltimore, MD
## 873                4       NORTHERN        4637 YORK RD\nBaltimore, MD
## 895               13        EASTERN      2301 BIDDLE ST\nBaltimore, MD
## 919                4       NORTHERN    6307 1 2 YORK RD\nBaltimore, MD
## 940                4       NORTHERN        6302 YORK RD\nBaltimore, MD
## 949                4       NORTHERN        5928 YORK RD\nBaltimore, MD
## 957                3   NORTHEASTERN      3539 BELAIR RD\nBaltimore, MD
## 976                4       NORTHERN        5002 YORK RD\nBaltimore, MD
## 994               12        EASTERN      2250 NORTH AVE\nBaltimore, MD
## 1017              12        EASTERN    1508 RUTLAND AVE\nBaltimore, MD
## 1018               4       NORTHERN   600 BELVEDERE AVE\nBaltimore, MD
## 1022               4       NORTHERN        5857 york rd\nBaltimore, MD
## 1053              13        EASTERN      2300 OLIVER ST\nBaltimore, MD
## 1120              13        EASTERN       2401 CHASE ST\nBaltimore, MD
## 1122              12        EASTERN      1400 NORTH AVE\nBaltimore, MD
## 1153               4       NORTHERN        5722 YORK RD\nBaltimore, MD
## 1155               4       NORTHERN   510 BELVEDERE AVE\nBaltimore, MD
## 1159              13        EASTERN     1201 POTOMAC ST\nBaltimore, MD
## 1186              12        EASTERN     2015 FEDERAL ST\nBaltimore, MD
## 1187              12        EASTERN     2015 FEDERAL ST\nBaltimore, MD
## 1198               4       NORTHERN        5921 YORK RD\nBaltimore, MD
```

```
## 1209              13        EASTERN     1432 CHESTER ST\nBaltimore, MD
## 1232               3   NORTHEASTERN      3534 belair RD\nBaltimore, MD
## 1246              12        EASTERN   1701 ELLSWORTH ST\nBaltimore, MD
## 1259               1   SOUTHEASTERN        1300 BANK ST\nBaltimore, MD
## 1287               4       NORTHERN        6304 YORK RD\nBaltimore, MD
## 1298               4       NORTHERN        5615 YORK RD\nBaltimore, MD
## 1304              13   NORTHEASTERN     4502 ERDMAN AVE\nBaltimore, MD
## 1312               4       NORTHERN        4515 YORK RD\nBaltimore, MD
## 1319               4       NORTHERN        5926 YORK RD\nBaltimore, MD
## 1320               4       NORTHERN        5407 YORK RD\nBaltimore, MD
## 1323               4       NORTHERN        5916 YORK RD\nBaltimore, MD
## 1325              13   NORTHEASTERN   3300 LAWNVIEW AVE\nBaltimore, MD
##      X2010.Census.Neighborhoods X2010.Census.Wards.Precincts Zip.Codes
## 29                           NA                           NA        NA
## 39                           NA                           NA        NA
## 92                           NA                           NA        NA
## 111                          NA                           NA        NA
## 187                          NA                           NA        NA
## 220                          NA                           NA        NA
## 266                          NA                           NA        NA
## 276                          NA                           NA        NA
## 289                          NA                           NA        NA
## 291                          NA                           NA        NA
## 362                          NA                           NA        NA
## 373                          NA                           NA        NA
## 383                          NA                           NA        NA
## 417                          NA                           NA        NA
## 475                          NA                           NA        NA
## 545                          NA                           NA        NA
## 604                          NA                           NA        NA
## 616                          NA                           NA        NA
## 620                          NA                           NA        NA
## 626                          NA                           NA        NA
## 678                          NA                           NA        NA
## 711                          NA                           NA        NA
## 763                          NA                           NA        NA
## 777                          NA                           NA        NA
## 779                          NA                           NA        NA
## 845                          NA                           NA        NA
## 852                          NA                           NA        NA
## 873                          NA                           NA        NA
## 895                          NA                           NA        NA
## 919                          NA                           NA        NA
## 940                          NA                           NA        NA
## 949                          NA                           NA        NA
## 957                          NA                           NA        NA
## 976                          NA                           NA        NA
## 994                          NA                           NA        NA
## 1017                         NA                           NA        NA
## 1018                         NA                           NA        NA
## 1022                         NA                           NA        NA
## 1053                         NA                           NA        NA
## 1120                         NA                           NA        NA
## 1122                         NA                           NA        NA
```

```
## 1153                        NA                          NA        NA
## 1155                        NA                          NA        NA
## 1159                        NA                          NA        NA
## 1186                        NA                          NA        NA
## 1187                        NA                          NA        NA
## 1198                        NA                          NA        NA
## 1209                        NA                          NA        NA
## 1232                        NA                          NA        NA
## 1246                        NA                          NA        NA
## 1259                        NA                          NA        NA
## 1287                        NA                          NA        NA
## 1298                        NA                          NA        NA
## 1304                        NA                          NA        NA
## 1312                        NA                          NA        NA
## 1319                        NA                          NA        NA
## 1320                        NA                          NA        NA
## 1323                        NA                          NA        NA
## 1325                        NA                          NA        NA
```

Cross tabs

Can view Data from datasets with the summary

```
data("UCBAdmissions")
DF <- as.data.frame(UCBAdmissions)
summary(DF)
```

```
##       Admit       Gender   Dept       Freq
##  Admitted:12   Male  :12   A:4   Min.   :  8.0
##  Rejected:12   Female:12   B:4   1st Qu.: 80.0
##                            C:4   Median :170.0
##                            D:4   Mean   :188.6
##                            E:4   3rd Qu.:302.5
##                            F:4   Max.   :512.0
```

Cross tabs. Breaking down Freq by gender and admit column data

```
xt<- xtabs(Freq ~ Gender + Admit, data = DF)
xt
```

```
##         Admit
## Gender   Admitted Rejected
##   Male       1198     1493
##   Female      557     1278
```

We can even crosstab for larger amount of data and variables. But it is hard to see. warpkreaks(standard dataset) break with all other columns

```
warpbreaks$replicate <- rep(1:9, len = 54)
xt = xtabs(breaks ~., data = warpbreaks)
xt
```

```
## , , replicate = 1
##
##     tension
## wool  L  M  H
##    A 26 18 36
##    B 27 42 20
##
## , , replicate = 2
##
##     tension
## wool  L  M  H
##    A 30 21 21
##    B 14 26 21
##
## , , replicate = 3
##
##     tension
## wool  L  M  H
##    A 54 29 24
##    B 29 19 24
##
## , , replicate = 4
##
##     tension
## wool  L  M  H
##    A 25 17 18
##    B 19 16 17
##
## , , replicate = 5
##
##     tension
## wool  L  M  H
##    A 70 12 10
##    B 29 39 13
##
## , , replicate = 6
##
##     tension
## wool  L  M  H
##    A 52 18 43
##    B 31 28 15
##
## , , replicate = 7
##
##     tension
## wool  L  M  H
##    A 51 35 28
##    B 41 21 15
##
## , , replicate = 8
##
##     tension
## wool  L  M  H
##    A 26 30 15
```

```
##    B 20 39 16
##
## , , replicate = 9
##
##     tension
## wool  L  M  H
##    A 67 36 26
##    B 44 29 28
```

We can actually Flat the output. So that, we can see the data in a more compact format

```
ftable(xt)
```

```
##              replicate  1  2  3  4  5  6  7  8  9
## wool tension
## A    L                 26 30 54 25 70 52 51 26 67
##      M                 18 21 29 17 12 18 35 30 36
##      H                 36 21 24 18 10 43 28 15 26
## B    L                 27 14 29 19 29 31 41 20 44
##      M                 42 26 19 16 39 28 21 39 29
##      H                 20 21 24 17 13 15 15 16 28
```

Dataset Size

Size of the dataset in memory

```
fakeData = rnorm(1e5)
object.size(fakeData)
```

```
## 800048 bytes
```

Even we can print the size in different scale

```
print(object.size(fakeData),units = 'MB')
```

```
## 0.8 Mb
```

Creating variables

creating variables for data analysis. Like if there is any missing values, if there is any value greater than some value, predicting values etc.

Creating sequence

using by command, a sequence can be created containing a constant distance

```
s1 <- seq(1,10,by =2); s1
```

```
## [1] 1 3 5 7 9
```

similarly we can give seq() length, and min max values. It'll then create exactly that length sequence

```
s2 <- seq(1,10,length = 3); s2
```

```
## [1]  1.0  5.5 10.0
```

or, we can create a sequence og length equal to another list

```
x <- rnorm(5)
seq(along=x)
```

```
## [1] 1 2 3 4 5
```

Subsetting variables

Subsetting the restaurants that are near me

```
restData$nearMe = restData$neighborhood %in% c('Roland Park','Homeland')
table(restData$nearMe)
```

```
##
## FALSE  TRUE
##  1314    13
```

Subsetting restaurants with wrong zipcode. Binary variable

```
restData$zipWrong <- ifelse(restData$zipCode<0,T,F)
table(restData$zipWrong, restData$zipCode <0)
```

```
##
##         FALSE TRUE
##   FALSE  1326    0
##   TRUE      0    1
```

```
table(restData$zipWrong, restData$zipCode >0)
```

```
##
##         FALSE TRUE
##   FALSE     0 1326
##   TRUE      1    0
```

Creating categorical variables

We can create a group of zipcodes.

```
restData$zipGroups <- cut(restData$zipCode, breaks = quantile(restData$zipCode))
table(restData$zipGroups)
```

```
##
## (-2.123e+04,2.12e+04]  (2.12e+04,2.122e+04] (2.122e+04,2.123e+04]
##                   337                   375                   282
## (2.123e+04,2.129e+04]
##                   332
```

Viewing the zipcodes

```
table(restData$zipGroups,restData$zipCode)
```

```
##
##                         -21226 21201 21202 21205 21206 21207 21208 21209 21210
##   (-2.123e+04,2.12e+04]      0   136   201     0     0     0     0     0     0
##   (2.12e+04,2.122e+04]       0     0     0    27    30     4     1     8    23
##   (2.122e+04,2.123e+04]      0     0     0     0     0     0     0     0     0
##   (2.123e+04,2.129e+04]      0     0     0     0     0     0     0     0     0
##
##                         21211 21212 21213 21214 21215 21216 21217 21218 21220
##   (-2.123e+04,2.12e+04]     0     0     0     0     0     0     0     0     0
##   (2.12e+04,2.122e+04]     41    28    31    17    54    10    32    69     0
##   (2.122e+04,2.123e+04]     0     0     0     0     0     0     0     0     1
##   (2.123e+04,2.129e+04]     0     0     0     0     0     0     0     0     0
##
##                         21222 21223 21224 21225 21226 21227 21229 21230 21231
##   (-2.123e+04,2.12e+04]     0     0     0     0     0     0     0     0     0
##   (2.12e+04,2.122e+04]      0     0     0     0     0     0     0     0     0
##   (2.122e+04,2.123e+04]     7    56   199    19     0     0     0     0     0
##   (2.123e+04,2.129e+04]     0     0     0     0    18     4    13   156   127
##
##                         21234 21237 21239 21251 21287
##   (-2.123e+04,2.12e+04]     0     0     0     0     0
##   (2.12e+04,2.122e+04]      0     0     0     0     0
##   (2.122e+04,2.123e+04]     0     0     0     0     0
##   (2.123e+04,2.129e+04]     7     1     3     2     1
```

We can even cut the groups into more segments using the cut2() function of the Himsc library

```
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:plyr':
##
##     is.discrete, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
restData$zipGroups = cut2(restData$zipCode, g =5)
table(restData$zipGroups)
```

```
##
## [-21226,21205) [ 21205,21214) [ 21214,21225) [ 21225,21231) [ 21231,21287]
##            338            193            445            210            141
```

Creating factor variables

```
restData$zcf <- factor(restData$zipCode)
restData$zcf[1:10]
```

```
##  [1] 21206 21231 21224 21211 21223 21218 21205 21211 21205 21231
## 32 Levels: -21226 21201 21202 21205 21206 21207 21208 21209 21210 ... 21287
```

Levels

```
yesno <- sample(c('yes','no'), size = 10, replace = TRUE)
yesnof <- factor(yesno, levels = c('yes','no'))
relevel(yesnof, ref = 'yes')
```

```
##  [1] no  yes yes yes no  yes yes no  yes yes
## Levels: yes no
```

```
as.numeric(yesnof)
```

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

Creating new DataFrame with new variable

Creating a copy of restData but adding another varibale called zipGroups. mutate function is under plyr library

```
restData2 <- mutate(restData,zipGroups = cut2(zipCode,g=4))
table(restData2$zipGroups)
```

```
##
## [-21226,21205) [ 21205,21220) [ 21220,21227) [ 21227,21287]
##            338            375            300            314
```

Setting significance

```r
x <- rnorm(10)
x
```

```
##  [1]  0.5428679  0.5144002 -0.5557001  1.1938073  0.4106490  1.2332349
##  [7]  0.6730770  0.8580998  0.6304570  0.2445746
```

```r
signif(x,digits = 2)
```

```
##  [1]  0.54  0.51 -0.56  1.20  0.41  1.20  0.67  0.86  0.63  0.24
```