Coursera_Cleaning Data_ Week3

Subsetting and sorting

```
#Warm up
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
## 1    2   NA   15
## 4    1   10   11
## 2    3   NA   12
## 3    5    6   14
## 5    4    9   13
```

```
X[,1]
```

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

```
X[,"var1"]
```

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

```
X[["var1"]]
```

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

```
X[1:2, "var2"]
```

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

```
#subset with logical
X[X$var1 <= 3 & X$var3 >11,]
```

```
##   var1 var2 var3
## 1    2   NA   15
## 2    3   NA   12
```

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

```
##   var1 var2 var3
## 1    2   NA   15
## 4    1   10   11
## 2    3   NA   12
```

```
#Dealing with missing values: USe which

X[which(X$var2>8), ] #no  NA
```

```
##   var1 var2 var3
## 4    1   10   11
## 5    4    9   13
```

```
X[X$var2>8, ]# with NA
```

```
##      var1 var2 var3
## NA     NA   NA   NA
## 4       1   10   11
```

```
## NA.1   NA   NA   NA
## 5       4    9   13
```

```
#Sort

sort(X$var1)
```

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

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

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

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

```
## [1]  6  9 10 NA NA
```

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

```
##   var1 var2 var3
## 4    1   10   11
## 1    2   NA   15
## 2    3   NA   12
## 5    4    9   13
## 3    5    6   14
```

```
#ordering with plyr

library(plyr)

arrange(X,var1)
```

```
##   var1 var2 var3
## 1    1   10   11
## 2    2   NA   15
## 3    3   NA   12
## 4    4    9   13
## 5    5    6   14
```

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

```
##   var1 var2 var3
## 1    5    6   14
## 2    4    9   13
## 3    3   NA   12
## 4    2   NA   15
## 5    1   10   11
```

```
#adding rows and cols

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

```
##   var1 var2 var3       var4
## 1    2   NA   15  0.1875960
## 4    1   10   11  1.7869764
## 2    3   NA   12  0.4966936
## 3    5    6   14  0.0631830
## 5    4    9   13 -0.5361329
```

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

```
##   var1 var2 var3       var4    rnorm(5)
## 1    2   NA   15  0.1875960  0.62578490
## 4    1   10   11  1.7869764 -2.45083750
## 2    3   NA   12  0.4966936  0.08909424
## 3    5    6   14  0.0631830  0.47838570
## 5    4    9   13 -0.5361329  1.00053336
```

Summarizing data

```
rest<- read.csv("/users/andrewhu/desktop/Coursera/Restaurants.csv")
#head(data)
#tail(data)
#str(data)
#quantile(data$var, na.rm=TRUE)
#quantile(data$var, probs=c(0.5,0.7,0.9))



#make a table
table(rest$zipCode, useNA="ifany") #not missing the NAs.
```

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

```
#two dimensional table
table(rest$councilDistrict, rest$zipCode)
```

```
##
##      -21226 21201 21202 21205 21206 21207 21208 21209 21210 21211 21212
##   1       0     0    37     0     0     0     0     0     0     0     0
##   2       0     0     0     3    27     0     0     0     0     0     0
##   3       0     0     0     0     0     0     0     0     0     0     0
##   4       0     0     0     0     0     0     0     0     0     0    27
##   5       0     0     0     0     0     3     0     6     0     0     0
##   6       0     0     0     0     0     0     0     1    19     0     0
##   7       0     0     0     0     0     0     0     1     0    27     0
##   8       0     0     0     0     0     1     0     0     0     0     0
##   9       0     1     0     0     0     0     0     0     0     0     0
##   10      1     0     1     0     0     0     0     0     0     0     0
##   11      0   115   139     0     0     0     1     0     0     0     1
##   12      0    20    24     4     0     0     0     0     0     0     0
##   13      0     0     0    20     3     0     0     0     0     0     0
##   14      0     0     0     0     0     0     0     0     4    14     0
##
##      21213 21214 21215 21216 21217 21218 21220 21222 21223 21224 21225
##   1      2     0     0     0     0     0     0     7     0   140     1
```

```
##   2      0     0     0     0     0     0     0     0     0    54     0
##   3      2    17     0     0     0     3     0     0     0     0     0
##   4      0     0     0     0     0     0     0     0     0     0     0
##   5      0     0    31     0     0     0     0     0     0     0     0
##   6      0     0    15     1     0     0     0     0     0     0     0
##   7      0     0     6     7    15     6     0     0     0     0     0
##   8      0     0     0     0     0     0     0     0     2     0     0
##   9      0     0     0     2     8     0     0     0    53     0     0
##   10     0     0     0     0     0     0     1     0     0     0    18
##   11     0     0     0     0     9     0     0     0     1     0     0
##   12    13     0     0     0     0    26     0     0     0     0     0
##   13    13     0     1     0     0     0     0     0     0     5     0
##   14     1     0     1     0     0    34     0     0     0     0     0
##
##      21226 21227 21229 21230 21231 21234 21237 21239 21251 21287
##   1      0     0     0     1   124     0     0     0     0     0
##   2      0     0     0     0     0     0     1     0     0     0
##   3      0     1     0     0     0     7     0     0     2     0
##   4      0     0     0     0     0     0     0     3     0     0
##   5      0     0     0     0     0     0     0     0     0     0
##   6      0     0     0     0     0     0     0     0     0     0
##   7      0     0     0     0     0     0     0     0     0     0
##   8      0     2    13     0     0     0     0     0     0     0
##   9      0     0     0    11     0     0     0     0     0     0
##   10    18     0     0   133     0     0     0     0     0     0
##   11     0     0     0    11     0     0     0     0     0     0
##   12     0     0     0     0     2     0     0     0     0     0
##   13     0     1     0     0     1     0     0     0     0     1
##   14     0     0     0     0     0     0     0     0     0     0
```

```
#check for missing values

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

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

```
any(is.na(rest$councilDistrict))
```

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

```
colSums(is.na(rest))
```

```
##            name         zipCode    neighborhood councilDistrict
##               0               0               0               0
##  policeDistrict      Location.1
##               0               0
```

```
all(colSums(is.na(rest))==0)
```

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

```
#finding values with specific characteristics

table(rest$zipCode %in% c("21212","21213"))
```

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

```r
#use this logical var to subset

rest[rest$zipCode %in% c("21212","21213"),]
```

```
##                                     name zipCode
## 29                      BAY ATLANTIC CLUB   21212
## 39                            BERMUDA BAR   21213
## 92                              ATWATER'S   21212
## 111           BALTIMORE ESTONIAN SOCIETY   21213
## 187                             CAFE ZEN   21212
## 220                  CERIELLO FINE FOODS   21212
## 266    CLIFTON PARK GOLF COURSE SNACK BAR   21213
## 276                CLUB HOUSE BAR & GRILL   21213
## 289                 CLUBHOUSE BAR & GRILL   21213
## 291                           COCKY LOU'S   21213
## 362       DREAM TAVERN, CARRIBEAN  U.S.A.   21213
## 373                          DUNKIN DONUTS   21212
## 383        EASTSIDE  SPORTS  SOCIAL  CLUB   21213
## 417                       FIELDS OLD TRAIL   21212
## 475                             GRAND CRU   21212
## 545                           RANDY'S BAR   21213
## 604     MURPHY'S NEIGHBORHOOD BAR & GRILL   21212
## 616                                NEOPOL   21212
## 620             NEW CLUB THUNDERBIRD INC.   21213
## 626                    NEW MAYFIELD, INC.   21213
## 678                           IKAN SEAFOOD   21212
## 711                           KAY-CEE CLUB   21212
## 763                                LA'RAE   21213
## 777                  LEMONGRASS BALTIMORE   21213
## 779                   LEN'S SANDWICH SHOP   21213
## 845                             MCDONALD'S   21213
## 852                             MCDONALD'S   21212
## 873                  NEW REX LIQUORS,INC.   21212
## 895                              OK TAVERN   21213
## 919                           PANERA BREAD   21212
## 940                    PEIWEI ASIAN DINER   21212
## 949                   PERGUSA ENTERPRISES   21212
## 957               PHANTOM'S BAR AND GRILL   21213
## 976          POPEYES FAMOUS FRIED CHICKEN   21212
## 994                           ROBBIE'S NEST   21213
## 1017                           RUTLAND BAR   21213
## 1018                       RYAN'S DAUGHTER   21212
## 1022         saigon remembered restaurant   21212
## 1053                SHIRLEY'S  HONEY HOLE   21213
## 1120                     STEEPLE CHASE II   21213
## 1122                                SUBWAY   21213
## 1153                              TAM-TAM   21212
## 1155                                TASTE   21212
## 1159                         TAYLORS EAST   21213
## 1186                THE EDGE BAR & LOUNGE   21213
## 1187 THE EDGE BAR & LOUNGE - KITCHEN AREA   21213
## 1198               THE HOLLOW BAR & GRILL   21212
## 1209             THE NEW BUCKETT'S LOUNGE   21213
## 1232                          THREE  ACE'S   21213
```

```
## 1246                TORAIN'S  HIDE-A-WAY   21213
## 1259                   TSUNAMI BALTIMORE   21213
## 1287                        VITO'S PIZZA   21212
## 1298 WENDY'S OLD FASHIONED HAMBURGERS #96   21212
## 1304                WHITTEN'S  (4502-04)   21213
## 1312                          wozi lounge   21212
## 1319          YETI RESTAURANT & CARRYOUT   21212
## 1320                    YORK CLUB TAVERN   21212
## 1323           ZEN WEST ROADSIDE CANTINA   21212
## 1325                    ZINK'S CAF\u0090   21213
##                     neighborhood councilDistrict policeDistrict
## 29                      Downtown             11       CENTRAL
## 39                 Broadway East             12       EASTERN
## 92     Chinquapin Park-Belvedere              4       NORTHERN
## 111           South Clifton Park             12       EASTERN
## 187                     Rosebank              4       NORTHERN
## 220    Chinquapin Park-Belvedere              4       NORTHERN
## 266                  Darley Park             14  NORTHEASTERN
## 276  Orangeville Industrial Area             13       EASTERN
## 289  Orangeville Industrial Area             13       EASTERN
## 291                Broadway East             12       EASTERN
## 362                Broadway East             13       EASTERN
## 373                     Homeland              4       NORTHERN
## 383                Broadway East             13       EASTERN
## 417                   Mid-Govans              4       NORTHERN
## 475    Chinquapin Park-Belvedere              4       NORTHERN
## 545                Broadway East             12       EASTERN
## 604                   Mid-Govans              4       NORTHERN
## 616    Chinquapin Park-Belvedere              4       NORTHERN
## 620                  Middle East             13       EASTERN
## 626                Belair-Edison             13  NORTHEASTERN
## 678    Chinquapin Park-Belvedere              4       NORTHERN
## 711                     Homeland              4       NORTHERN
## 763                       Oliver             12       EASTERN
## 777                 Little Italy              1  SOUTHEASTERN
## 779                Broadway East             12       EASTERN
## 845           South Clifton Park             12       EASTERN
## 852               Radnor-Winston              4       NORTHERN
## 873                  Wilson Park              4       NORTHERN
## 895                Biddle Street             13       EASTERN
## 919                  Lake Walker              4       NORTHERN
## 940                   Cedarcroft              4       NORTHERN
## 949                     Rosebank              4       NORTHERN
## 957                Belair-Edison              3  NORTHEASTERN
## 976               Winston-Govans              4       NORTHERN
## 994                Broadway East             12       EASTERN
## 1017               Broadway East             12       EASTERN
## 1018   Chinquapin Park-Belvedere              4       NORTHERN
## 1022                  Mid-Govans              4       NORTHERN
## 1053               Broadway East             13       EASTERN
## 1120               Biddle Street             13       EASTERN
## 1122                      Oliver             12       EASTERN
## 1153                  Mid-Govans              4       NORTHERN
## 1155                  Mid-Govans              4       NORTHERN
```

```
## 1159                      Berea             13        EASTERN
## 1186              Broadway East             12        EASTERN
## 1187              Broadway East             12        EASTERN
## 1198                   Rosebank              4       NORTHERN
## 1209              Broadway East             13        EASTERN
## 1232              Belair-Edison              3   NORTHEASTERN
## 1246              Broadway East             12        EASTERN
## 1259               Little Italy              1   SOUTHEASTERN
## 1287                 Cedarcroft              4       NORTHERN
## 1298                   Homeland              4       NORTHERN
## 1304          Claremont-Freedom             13   NORTHEASTERN
## 1312                   Guilford              4       NORTHERN
## 1319                   Rosebank              4       NORTHERN
## 1320                   Homeland              4       NORTHERN
## 1323                   Rosebank              4       NORTHERN
## 1325              Belair-Edison             13   NORTHEASTERN
##                                Location.1
## 29        206 REDWOOD ST\nBaltimore, MD\n
## 39        1801 NORTH AVE\nBaltimore, MD\n
## 92     529 BELVEDERE AVE\nBaltimore, MD\n
## 111       1932 BELAIR RD\nBaltimore, MD\n
## 187    438 BELVEDERE AVE\nBaltimore, MD\n
## 220    529 BELVEDERE AVE\nBaltimore, MD\n
## 266        2701 ST LO DR\nBaltimore, MD\n
## 276      4217 ERDMAN AVE\nBaltimore, MD\n
## 289      4217 ERDMAN AVE\nBaltimore, MD\n
## 291       2101 NORTH AVE\nBaltimore, MD\n
## 362   2300 LAFAYETTE AVE\nBaltimore, MD\n
## 373         5422 YORK RD\nBaltimore, MD\n
## 383  1203 COLLINGTON AVE\nBaltimore, MD\n
## 417         5723 YORK RD\nBaltimore, MD\n
## 475    527 BELVEDERE AVE\nBaltimore, MD\n
## 545       2135 NORTH AVE\nBaltimore, MD\n
## 604         5847 YORK RD\nBaltimore, MD\n
## 616    529 BELVEDERE AVE\nBaltimore, MD\n
## 620        2201 CHASE ST\nBaltimore, MD\n
## 626       3349 BELAIR RD\nBaltimore, MD\n
## 678    529 BELVEDERE AVE\nBaltimore, MD\n
## 711     201 HOMELAND AVE\nBaltimore, MD\n
## 763      1000 HOFFMAN ST\nBaltimore, MD\n
## 777     1300 BANK STREET\nBaltimore, MD\n
## 779   1500 WASHINGTON ST\nBaltimore, MD\n
## 845        2001 BROADWAY\nBaltimore, MD\n
## 852         5100 YORK RD\nBaltimore, MD\n
## 873         4637 YORK RD\nBaltimore, MD\n
## 895       2301 BIDDLE ST\nBaltimore, MD\n
## 919     6307 1 2 YORK RD\nBaltimore, MD\n
## 940         6302 YORK RD\nBaltimore, MD\n
## 949         5928 YORK RD\nBaltimore, MD\n
## 957       3539 BELAIR RD\nBaltimore, MD\n
## 976         5002 YORK RD\nBaltimore, MD\n
## 994       2250 NORTH AVE\nBaltimore, MD\n
## 1017    1508 RUTLAND AVE\nBaltimore, MD\n
## 1018   600 BELVEDERE AVE\nBaltimore, MD\n
```

```
## 1022        5857 york rd\nBaltimore, MD\n
## 1053      2300 OLIVER ST\nBaltimore, MD\n
## 1120       2401 CHASE ST\nBaltimore, MD\n
## 1122      1400 NORTH AVE\nBaltimore, MD\n
## 1153        5722 YORK RD\nBaltimore, MD\n
## 1155   510 BELVEDERE AVE\nBaltimore, MD\n
## 1159     1201 POTOMAC ST\nBaltimore, MD\n
## 1186     2015 FEDERAL ST\nBaltimore, MD\n
## 1187     2015 FEDERAL ST\nBaltimore, MD\n
## 1198        5921 YORK RD\nBaltimore, MD\n
## 1209     1432 CHESTER ST\nBaltimore, MD\n
## 1232      3534 belair RD\nBaltimore, MD\n
## 1246   1701 ELLSWORTH ST\nBaltimore, MD\n
## 1259        1300 BANK ST\nBaltimore, MD\n
## 1287        6304 YORK RD\nBaltimore, MD\n
## 1298        5615 YORK RD\nBaltimore, MD\n
## 1304     4502 ERDMAN AVE\nBaltimore, MD\n
## 1312        4515 YORK RD\nBaltimore, MD\n
## 1319        5926 YORK RD\nBaltimore, MD\n
## 1320        5407 YORK RD\nBaltimore, MD\n
## 1323        5916 YORK RD\nBaltimore, MD\n
## 1325   3300 LAWNVIEW AVE\nBaltimore, MD\n
```

```
#Cross tabs
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
```

```
x1<- xtabs(Freq~Gender+Admit, data=DF)
x1
```

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

```
#Flat tables
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
```

```
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
```

Creating new variables

```
#Create sequences: need an index for data set
s1<- seq(1,10,by=2) ; s1 #specify the interval
```

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

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

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

```
x<- c(1,3,8,25,100); seq(along=x) #create index for the 5 values in x
```

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

```
#subsetting variables
rest$nearme = rest$neighborhood %in% c("Roland Park", "Homeland")

table(rest$nearme)
```

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

```
#Create binary variables
rest$zipWrong = ifelse(rest$zipCode<0, TRUE, FALSE)
table(rest$zipWrong, rest$zipCode<0)
```

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

```
#Create categorical variables
rest$zipGroups = cut(rest$zipCode, breaks= quantile(rest$zipCode))

table(rest$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
```

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

```
##
##                         -21226 21201 21202 21205 21206 21207 21208 21209
##   (-2.123e+04,2.12e+04]      0   136   201     0     0     0     0     0
```

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

```
#Easier cutting
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
```

```
rest$zipGroups = cut2(rest$zipCode, g=4) #break them up according to the quantiles

table(rest$zipGroups)
```

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

```
#Create factor variables
rest$zcf <- factor(rest$zipCode)
rest$zcf[1:10]
```

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

```r
class(rest$zcf)
```

```
## [1] "factor"
```

```r
#Levels of factor variables
yesno<- sample(c("yes", "no"), size=10, replace=TRUE)
yesnofac<- factor(yesno, levels= c("yes","no")) #default for levels is no
relevel(yesnofac, ref="yes")
```

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

```r
#Mutate function: transform dataset
library(Hmisc); library(plyr)
rest2 <- mutate(rest, zipGroups= cut2(zipCode,g=4))
table(rest2$zipGroups)
```

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

```r
#common transforms

#abs(x) absolute value
#sqrt() square root
#ceiling 3.5 -->4
#floor 3.5 --> 3
#round(3.475,digits=2) is 3.48
#signif(3.475, digits=2) is 3.5
```

Data Reshaping: each var per col, each obs per row

```r
library(reshape2)
library(datasets)
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```r
#melt the dataset

mtcars$carname <- rownames(mtcars)
carMelt <- melt(mtcars, id=c("carname","gear","cyl"), measure.vars=c("mpg","hp"))

head(carMelt, n=3)
```

```
##         carname gear cyl variable value
## 1     Mazda RX4    4   6      mpg  21.0
## 2 Mazda RX4 Wag    4   6      mpg  21.0
## 3    Datsun 710    4   4      mpg  22.8
```

```r
tail(carMelt, n=3)
```

```
##          carname gear cyl variable value
```

```
## 62  Ferrari Dino    5   6       hp   175
## 63 Maserati Bora    5   8       hp   335
## 64    Volvo 142E    4   4       hp   109
```

```
#Casting data frames

cylData <- dcast(carMelt, cyl~variable)# for cyl 4, there are 11 measurements for mpg..
```

```
## Aggregation function missing: defaulting to length
```

```
cylData
```

```
##   cyl mpg hp
## 1   4  11 11
## 2   6   7  7
## 3   8  14 14
```

```
cylData <- dcast(carMelt, cyl~variable,mean)
cylData
```

```
##   cyl      mpg        hp
## 1   4 26.66364  82.63636
## 2   6 19.74286 122.28571
## 3   8 15.10000 209.21429
```

```
#Averaging values
head(InsectSprays)
```

```
##   count spray
## 1    10     A
## 2     7     A
## 3    20     A
## 4    14     A
## 5    14     A
## 6    12     A
```

```
#apply to count along the index spray, with sum
#sums up the count for each index spray
tapply(InsectSprays$count, InsectSprays$spray,sum)
```

```
##   A   B   C   D   E   F
## 174 184  25  59  42 200
```

```
#Another way using plyr package
ddply(InsectSprays, .(spray),summarise, sum=sum(count))
```

```
##   spray sum
## 1     A 174
## 2     B 184
## 3     C  25
## 4     D  59
## 5     E  42
## 6     F 200
```

Intro to dplyr

```
#Introduction to dplyr

library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.1
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:Hmisc':
##
##     src, summarize
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
chicago<- readRDS("/users/andrewhu/desktop/Coursera/chicago.rds")
dim(chicago)
```

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

```
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 ...
```

```
names(chicago)
```

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

```
#look at subsets of columns
head(select(chicago, city:dptp))
```

```
##   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
```

```
head(select(chicago, -(city:dptp)))
```

```
##         date pm25tmean2 pm10tmean2 o3tmean2 no2tmean2
## 1 1987-01-01         NA   34.00000 4.250000  19.98810
## 2 1987-01-02         NA         NA 3.304348  23.19099
```

```
## 3 1987-01-03         NA   34.16667 3.333333  23.81548
## 4 1987-01-04         NA   47.00000 4.375000  30.43452
## 5 1987-01-05         NA         NA 4.750000  30.33333
## 6 1987-01-06         NA   48.00000 5.833333  25.77233
```

```
#filter
#subset using multiple conditions
chic.f <- filter(chicago, pm25tmean2>30 & tmpd>80)
head(chic.f)
```

```
##   city tmpd dptp       date pm25tmean2 pm10tmean2 o3tmean2 no2tmean2
## 1 chic   81 71.2 1998-08-23    39.6000       59.0 45.86364  14.32639
## 2 chic   81 70.4 1998-09-06    31.5000       50.5 50.66250  20.31250
## 3 chic   82 72.2 2001-07-20    32.3000       58.5 33.00380  33.67500
## 4 chic   84 72.9 2001-08-01    43.7000       81.5 45.17736  27.44239
## 5 chic   85 72.6 2001-08-08    38.8375       70.0 37.98047  27.62743
## 6 chic   84 72.6 2001-08-09    38.2000       66.0 36.73245  26.46742
```

```
#arrange: re order
chicago<- arrange(chicago,date)
head(chicago)
```

```
##   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
```

```
tail(chicago)
```

```
##      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<- arrange(chicago,desc(date))
head(chicago)
```

```
##   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 (new name = old name)
chicago<- rename(chicago,pm25 = pm25tmean2, dewpoint=dptp)
head(chicago)
```

```
##   city tmpd dewpoint       date     pm25 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
```

```
#Mutate:transform and create new var
chicago<- mutate(chicago, pm25detrend= pm25- mean(pm25,na.rm = TRUE))
head(chicago)
```

```
##   city tmpd dewpoint       date     pm25 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
##   pm25detrend
## 1   -1.230958
## 2   -1.173815
## 3   -8.780958
## 4    1.519042
## 5    7.329042
## 6   -7.830958
```

```
#group by: split a data frame according to categorical variables
chicago<- mutate(chicago, tempcat= factor(1* (tmpd>80), labels=c("cold","hot")))
hotcold<- group_by(chicago,tempcat)
hotcold
```

```
## # A tibble: 6,940 x 10
## # Groups:   tempcat [3]
##    city   tmpd dewpoint date        pm25 pm10tmean2 o3tmean2 no2tmean2
##    <chr> <dbl>    <dbl> <date>     <dbl>      <dbl>    <dbl>     <dbl>
##  1 chic     35     30.1 2005-12-31 15          23.5     2.53      13.2
##  2 chic     36     31   2005-12-30 15.1        19.2     3.03      22.8
##  3 chic     35     29.4 2005-12-29  7.45       23.5     6.79      20.0
##  4 chic     37     34.5 2005-12-28 17.8        27.5     3.26      19.3
##  5 chic     40     33.6 2005-12-27 23.6        27       4.47      23.5
##  6 chic     35     29.6 2005-12-26  8.4         8.5    14.0       16.8
##  7 chic     35     32.1 2005-12-25  6.7         8      14.4       13.8
##  8 chic     37     35.2 2005-12-24 30.8        25.2     1.77      32.0
##  9 chic     41     32.6 2005-12-23 32.9        34.5     6.91      29.1
## 10 chic     22     23.3 2005-12-22 36.6        42.5     5.39      33.7
## # ... with 6,930 more rows, and 2 more variables: pm25detrend <dbl>,
## #   tempcat <fct>
```

```
#summarize
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
```

```r
#summarize based on year
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), no2= median(no2tmean2))
```

```
## # 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
```

```r
#Pipeline operator
chicago %>% mutate(month = as.POSIXlt(date)$mon +1) %>% group_by(month) %>% summarize(pm25= mean(pm25, r
```

```
## # 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
```

```r
#---------Merging data
review<- read.csv("/users/andrewhu/desktop/reviews.csv")
solu<- read.csv("/users/andrewhu/desktop/solutions.csv")


head(review)
```

```
##   id solution_id reviewer_id      start       stop time_left accept
```

```
## 1  1           3          27 1304095698 1304095758      1754      1
## 2  2           4          22 1304095188 1304095206      2306      1
## 3  3           5          28 1304095276 1304095320      2192      1
## 4  4           1          26 1304095267 1304095423      2089      1
## 5  5          10          29 1304095456 1304095469      2043      1
## 6  6           2          29 1304095471 1304095513      1999      1
```

```
head(solu)
```

```
##   id problem_id subject_id      start       stop time_left answer
## 1  1        156         29 1304095119 1304095169      2343      B
## 2  2        269         25 1304095119 1304095183      2329      C
## 3  3         34         22 1304095127 1304095146      2366      C
## 4  4         19         23 1304095127 1304095150      2362      D
## 5  5        605         26 1304095127 1304095167      2345      A
## 6  6        384         27 1304095131 1304095270      2242      C
```

```
names(review)
```

```
## [1] "id"          "solution_id" "reviewer_id" "start"       "stop"
## [6] "time_left"   "accept"
```

```
names(solu)
```

```
## [1] "id"         "problem_id" "subject_id" "start"      "stop"
## [6] "time_left"  "answer"
```

```
#merge by solution_id and id

mergedata= merge(review,solu,by.x="solution_id", by.y="id",all=TRUE)
head(mergedata)
```

```
##   solution_id id reviewer_id    start.x     stop.x time_left.x accept
## 1           1  4          26 1304095267 1304095423        2089      1
## 2           2  6          29 1304095471 1304095513        1999      1
## 3           3  1          27 1304095698 1304095758        1754      1
## 4           4  2          22 1304095188 1304095206        2306      1
## 5           5  3          28 1304095276 1304095320        2192      1
## 6           6 16          22 1304095303 1304095471        2041      1
##   problem_id subject_id    start.y     stop.y time_left.y answer
## 1        156         29 1304095119 1304095169        2343      B
## 2        269         25 1304095119 1304095183        2329      C
## 3         34         22 1304095127 1304095146        2366      C
## 4         19         23 1304095127 1304095150        2362      D
## 5        605         26 1304095127 1304095167        2345      A
## 6        384         27 1304095131 1304095270        2242      C
```

```
#use plyr to merge

#e.g. arrange(join(df1,df2),id)


#Multiple dfs

df1 = data.frame(id=sample(1:10),x=rnorm(10))
df2 = data.frame(id=sample(1:10),y=rnorm(10))
df3 = data.frame(id=sample(1:10),z=rnorm(10))
dfList <- list(df1,df2,df3)
```

```
join_all(dfList)
```

```
## Joining by: id
## Joining by: id
```

```
##    id          x          y          z
## 1   9 -1.0105164  0.7686927 -1.6992814
## 2   7  0.6095613 -0.9986890 -0.4737391
## 3   2  0.5041528 -0.8324297  0.7678650
## 4   5  1.3798872 -0.7852223 -1.9271914
## 5   1  0.4906615  0.2855734  0.9866399
## 6   6  1.4912935 -1.8563536  0.4195678
## 7   3 -0.1842727  0.7842468  0.5716195
## 8   8  0.5127249 -0.6440821 -0.4520897
## 9  10 -0.9409096 -0.4483046 -0.7917472
## 10  4 -0.3808710 -0.4178425  1.0365973
```