Week 3

Tomás

19 de noviembre de 2019

# Subsetting

library(plyr)  
set.seed(123)  
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[,1] # Primera columna del dataframe

## [1] 3 4 1 5 2

x[,"var1"] # Columna var1 del dataframe

## [1] 3 4 1 5 2

x[1:2,"var2"]#Columna var 2 y filas 1:2

## [1] NA 10

x[(x$var1 <= 3 & x$var3 > 11),] # Busco filas que tengan var 1 mayor/igual que 3 y var 3 menor que 11

## var1 var2 var3  
## 1 3 NA 12  
## 5 1 NA 15  
## 2 2 6 13

x[(x$var1 <= 3 | x$var3 > 15),] # Busco filas que var 1 sea mayor o igual a 3 o que var3 sea mayor que 15

## var1 var2 var3  
## 1 3 NA 12  
## 5 1 NA 15  
## 2 2 6 13

x[which(x$var2 >8),] #La funcion which entrega los indices donde se comple esa condicion sin contar los NA

## var1 var2 var3  
## 4 4 10 14

x[x$var2 >8,] # Sin la funcion which considera los NA

## var1 var2 var3  
## NA NA NA NA  
## 4 4 10 14  
## NA.1 NA NA NA

sort(x$var1) # Ordena la columna de menor a mayor

## [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 7 10 NA NA

x[order(x$var1),] #Me ordena el dataframe en funcion de la variable 1

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

x[order(x$var1,x$var3),] # Si existe empate en la variable 1, se orden por la variable 3

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

arrange(x,var1) # Mismo caso que order, pero mas compracto

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

arrange(x,desc(var1)) # Decreciente

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

x$var4 <- rnorm(5) # Añadir columna nueva  
y <- cbind(x, "var5" = rnorm(5)) # Lo mismo pero debo crear otro dataframe, rbind agrega filas

# Summarazing data

data <- read.csv("Restaurants.csv")  
head(data,3) # Visualizar las 3 primeras lineas

## 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  
## 1 4509 BELAIR ROAD\nBaltimore, MD  
## 2 1919 FLEET ST\nBaltimore, MD  
## 3 2844 HUDSON ST\nBaltimore, MD

tail(data,3) # Visualizar las 3 ultimas lineas

## name zipCode neighborhood councilDistrict policeDistrict  
## 1325 ZINK'S CAFÂ\220 21213 Belair-Edison 13 NORTHEASTERN  
## 1326 ZISSIMOS BAR 21211 Hampden 7 NORTHERN  
## 1327 ZORBAS 21224 Greektown 2 SOUTHEASTERN  
## Location.1  
## 1325 3300 LAWNVIEW AVE\nBaltimore, MD  
## 1326 1023 36TH ST\nBaltimore, MD  
## 1327 4710 EASTERN Ave\nBaltimore, MD

summary(data) # Resumen de la data

## name zipCode neighborhood  
## MCDONALD'S : 8 Min. :-21226 Downtown :128   
## POPEYES FAMOUS FRIED CHICKEN: 7 1st Qu.: 21202 Fells Point : 91   
## SUBWAY : 6 Median : 21218 Inner Harbor: 89   
## KENTUCKY FRIED CHICKEN : 5 Mean : 21185 Canton : 81   
## BURGER KING : 4 3rd Qu.: 21226 Federal Hill: 42   
## DUNKIN DONUTS : 4 Max. : 21287 Mount Vernon: 33   
## (Other) :1293 (Other) :863   
## councilDistrict policeDistrict  
## Min. : 1.000 SOUTHEASTERN:385   
## 1st Qu.: 2.000 CENTRAL :288   
## Median : 9.000 SOUTHERN :213   
## Mean : 7.191 NORTHERN :157   
## 3rd Qu.:11.000 NORTHEASTERN: 72   
## Max. :14.000 EASTERN : 67   
## (Other) :145   
## Location.1   
## 1101 RUSSELL ST\nBaltimore, MD: 9   
## 201 PRATT ST\nBaltimore, MD : 8   
## 2400 BOSTON ST\nBaltimore, MD : 8   
## 300 LIGHT ST\nBaltimore, MD : 5   
## 300 CHARLES ST\nBaltimore, MD : 4   
## 301 LIGHT ST\nBaltimore, MD : 4   
## (Other) :1289

str(data) #Entrega la clase de las variables y algunas caracteristicas del df

## 'data.frame': 1327 obs. of 6 variables:  
## $ name : Factor w/ 1277 levels "#1 CHINESE KITCHEN",..: 9 3 992 1 2 4 5 6 7 8 ...  
## $ zipCode : int 21206 21231 21224 21211 21223 21218 21205 21211 21205 21231 ...  
## $ neighborhood : Factor w/ 173 levels "Abell","Arlington",..: 53 52 18 66 104 33 98 133 98 157 ...  
## $ councilDistrict: int 2 1 1 14 9 14 13 7 13 1 ...  
## $ policeDistrict : Factor w/ 9 levels "CENTRAL","EASTERN",..: 3 6 6 4 8 3 6 4 6 6 ...  
## $ Location.1 : Factor w/ 1210 levels "1 BIDDLE ST\nBaltimore, MD",..: 835 334 554 755 492 537 505 530 507 569 ...

quantile(data$councilDistrict, na.rm = TRUE) # Me entrega los quantiles

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

table(data$zipCode, useNA = "ifany") # Entrega una tabla con las frecuencias de las observaciones (useNA = "ifany", entrega una columna extra si encuentra observaciones NA)

##   
## -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

table(data$councilDistrict,data$zipCode) # Entrega una tabla de dos dimensiones con dos variables

##   
## -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

sum(is.na(data$councilDistrict)) # Verifica si existen NA y los suma, en este caso no existe (Si hay NA is.na entrega un TRUE, TRUE se interpreta como 1)

## [1] 0

any(is.na(data$councilDistrict)) # Pregunta si existe algun dato NA, any verifica si existe algun TRUE

## [1] FALSE

all(data$zipCode > 0) #Verifica que todas las observaciones de esa variables son mayores que 0, si alguna es menor que 0 entrega FALSE

## [1] FALSE

colSums(is.na(data)) # Suma por cada columna la cantidad de NA (se puede hacer por filas tambien "rowSums)

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

table(data$zipCode %in% c("21212","21213")) #Permite encontrar cuantas observaciones son parte de esos dos strings

##   
## FALSE TRUE   
## 1268 59

data[data$zipCode %in% c("21212","21213"),] #Con esto puedo crear un subset que contiene solo las filas que pertenencen al character

## 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 CAFÂ\220 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  
## 39 1801 NORTH AVE\nBaltimore, MD  
## 92 529 BELVEDERE AVE\nBaltimore, MD  
## 111 1932 BELAIR RD\nBaltimore, MD  
## 187 438 BELVEDERE AVE\nBaltimore, MD  
## 220 529 BELVEDERE AVE\nBaltimore, MD  
## 266 2701 ST LO DR\nBaltimore, MD  
## 276 4217 ERDMAN AVE\nBaltimore, MD  
## 289 4217 ERDMAN AVE\nBaltimore, MD  
## 291 2101 NORTH AVE\nBaltimore, MD  
## 362 2300 LAFAYETTE AVE\nBaltimore, MD  
## 373 5422 YORK RD\nBaltimore, MD  
## 383 1203 COLLINGTON AVE\nBaltimore, MD  
## 417 5723 YORK RD\nBaltimore, MD  
## 475 527 BELVEDERE AVE\nBaltimore, MD  
## 545 2135 NORTH AVE\nBaltimore, MD  
## 604 5847 YORK RD\nBaltimore, MD  
## 616 529 BELVEDERE AVE\nBaltimore, MD  
## 620 2201 CHASE ST\nBaltimore, MD  
## 626 3349 BELAIR RD\nBaltimore, MD  
## 678 529 BELVEDERE AVE\nBaltimore, MD  
## 711 201 HOMELAND AVE\nBaltimore, MD  
## 763 1000 HOFFMAN ST\nBaltimore, MD  
## 777 1300 BANK STREET\nBaltimore, MD  
## 779 1500 WASHINGTON ST\nBaltimore, MD  
## 845 2001 BROADWAY\nBaltimore, MD  
## 852 5100 YORK RD\nBaltimore, MD  
## 873 4637 YORK RD\nBaltimore, MD  
## 895 2301 BIDDLE ST\nBaltimore, MD  
## 919 6307 1 2 YORK RD\nBaltimore, MD  
## 940 6302 YORK RD\nBaltimore, MD  
## 949 5928 YORK RD\nBaltimore, MD  
## 957 3539 BELAIR RD\nBaltimore, MD  
## 976 5002 YORK RD\nBaltimore, MD  
## 994 2250 NORTH AVE\nBaltimore, MD  
## 1017 1508 RUTLAND AVE\nBaltimore, MD  
## 1018 600 BELVEDERE AVE\nBaltimore, MD  
## 1022 5857 york rd\nBaltimore, MD  
## 1053 2300 OLIVER ST\nBaltimore, MD  
## 1120 2401 CHASE ST\nBaltimore, MD  
## 1122 1400 NORTH AVE\nBaltimore, MD  
## 1153 5722 YORK RD\nBaltimore, MD  
## 1155 510 BELVEDERE AVE\nBaltimore, MD  
## 1159 1201 POTOMAC ST\nBaltimore, MD  
## 1186 2015 FEDERAL ST\nBaltimore, MD  
## 1187 2015 FEDERAL ST\nBaltimore, MD  
## 1198 5921 YORK RD\nBaltimore, MD  
## 1209 1432 CHESTER ST\nBaltimore, MD  
## 1232 3534 belair RD\nBaltimore, MD  
## 1246 1701 ELLSWORTH ST\nBaltimore, MD  
## 1259 1300 BANK ST\nBaltimore, MD  
## 1287 6304 YORK RD\nBaltimore, MD  
## 1298 5615 YORK RD\nBaltimore, MD  
## 1304 4502 ERDMAN AVE\nBaltimore, MD  
## 1312 4515 YORK RD\nBaltimore, MD  
## 1319 5926 YORK RD\nBaltimore, MD  
## 1320 5407 YORK RD\nBaltimore, MD  
## 1323 5916 YORK RD\nBaltimore, MD  
## 1325 3300 LAWNVIEW AVE\nBaltimore, MD

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

xtabs(Freq ~ Gender + Admit , data = df) # Crea una tabla cruzada con el centro "Freq" y de variables Gender y Admit

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

data("warpbreaks")  
df1 <- as.data.frame(warpbreaks)  
warpbreaks$replicate <- rep(1:9, len = 54)  
xt <- xtabs(breaks ~ . , data = warpbreaks) # Le entrego todas las variables me hace todas las tablas posibles  
ftable(xt) #Encuentro un resumen de lo anterior

## 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

data <- read.csv("Restaurants.csv")  
s1 <- seq(1,10,by=2) ; s1 #Crea una secuencia de numeros del 1 al 10 cada dos

## [1] 1 3 5 7 9

s2 <- seq(1,10, lenght = 3); s2 #Crea una secuencia de exactamente 3 numeros

## Warning: In seq.default(1, 10, lenght = 3) :  
## extra argument 'lenght' will be disregarded

## [1] 1 2 3 4 5 6 7 8 9 10

x <- c(1,3,8,25,100); seq(along = x) #Crea una secuencia empezando por 1 del mismo largo que el vector

## [1] 1 2 3 4 5

data$nearme <- data$neighborhood %in% c("Roland Park", "Homeland") #Con esto he creadro una variable que indica cuales restaurantes esta cerca mio segun los vecindarios que he espicificado  
table(data$nearme)

##   
## FALSE TRUE   
## 1314 13

data$zipwrong <- ifelse(data$zipCode < 0,TRUE,FALSE) # Crea una nueva variable si el zip code es menor que 0  
table(data$zipwrong,data$zipCode<0)

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

data$zipgroups <- cut(data$zipCode,breaks = quantile(data$zipCode)) #Divide en los cuartiles los zipcode  
table(data$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(data$zipgroups,data$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

library(Hmisc)

## Warning: package 'Hmisc' was built under R version 3.6.1

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.6.1

##   
## 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

data$zipgroups <- cut2(data$zipCode,g=4) # Divide los zip code en 4 grupos  
table(data$zipgroups)

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

data$zcf <- factor(data$zipCode) ; data$zcf[1:10] # Convierte una variable que es integer en un factor, asi se pueden realizar regresiones lineales por ejemplo donde no necesariamente "mas zipcode" implica algo mejor

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

class(data$zcf)

## [1] "factor"

library(plyr)  
data2 <- mutate(data, zipGroups=cut2(zipCode,g=4))

# Reshaping Data

library(reshape2)  
library(datasets)  
data("mtcars")  
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

mtcars$carname <- rownames(mtcars)  
carMelt <- melt(mtcars, id=c("carname","gear","cyl"), measure.cars = c("mpg","hp"))  
cylData <- dcast(carMelt, cyl ~ variable) ; cylData

## Aggregation function missing: defaulting to length

## cyl mpg disp hp drat wt qsec vs am carb  
## 1 4 11 11 11 11 11 11 11 11 11  
## 2 6 7 7 7 7 7 7 7 7 7  
## 3 8 14 14 14 14 14 14 14 14 14

#ddply() # Se puede splitear un dataframe, mandarle una funcion y despues lo deja como un dataframe.

# Dplyr

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.6.1

##   
## 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("chicago.rds")  
head(select(chicago, city:dptp)) # Select me permite utilizar los nombres de las variables y no sus indices

## 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))) # Puedo seleccionar todas menos la que le indico -()

## 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

chic.f <- filter(chicago, chicago$pm25tmean2 > 30) # Filtrar con una condicion  
chic.f <- filter(chicago, chicago$pm25tmean2 > 30 & chicago$tmpd > 80) # Mas de una condicion  
chicago <- arrange(chicago,date) # Puedo organizar por alguna variable  
chicago <- rename(chicago, pm25 = pm25tmean2, dewpoint = dptp) #Renombra una variable  
chicago <- mutate(chicago, pm25detrend = pm25-mean(pm25, na.rm = TRUE)) #Permite crear otra variable con informacion de alguna existente  
chicago <- mutate(chicago, tempcat = factor(1\*(tmpd >80), labels = c("cold","hot")))  
hot\_cold <- group\_by(chicago,tempcat)

## Warning: Factor `tempcat` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

summarize(hot\_cold, pm25 = mean(pm25, na.rm = TRUE), o3 = max(o3tmean2), n02 = mean(no2tmean2))

## # A tibble: 3 x 4  
## tempcat pm25 o3 n02  
## <fct> <dbl> <dbl> <dbl>  
## 1 cold 16.0 66.6 25.2  
## 2 hot 26.5 63.0 25.2  
## 3 <NA> 47.7 9.42 37.4