Datacamp_Importing & Cleaning Data in R: Case Studies___World Food Facts

dizhen

2019/4/5

Importing the data

```
# Load data.table
library(data.table)

# Import food.csv as a data frame: food
food<-fread("data/food.csv",data.table = FALSE)
```

Examining the data

```
# View summary of food
summary(food)

# View head of food
head(food)

# View structure of food
str(food)
```

Inspecting variables

```
# Load dplyr
library(dplyr)

# View a glimpse of food
glimpse(food)

# View column names of food
names(food)
```

Removing duplicate info

```
# Define vector of duplicate cols (don't change)
duplicates <- c(4, 6, 11, 13, 15, 17, 18, 20, 22,
                24, 25, 28, 32, 34, 36, 38, 40,
                44, 46, 48, 51, 54, 65, 158)

# Remove duplicates from food: food2
food2 <- food[,-duplicates]
```

Removing useless info

```r
# Define useless vector (don't change)
useless <- c(1, 2, 3, 32:41)

# Remove useless columns from food2: food3
food3 <- food2[,-useless]
```

Finding columns

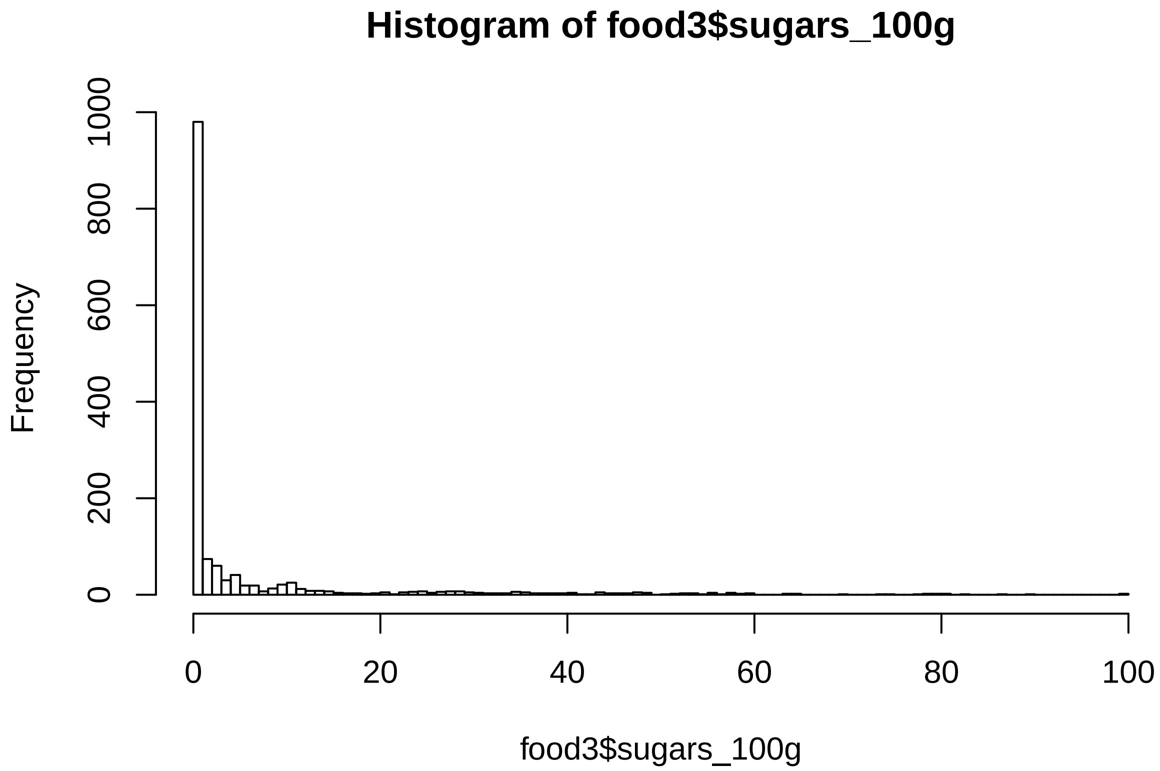
```r
# Create vector of column indices: nutrition
library("stringr")
nutrition <- str_detect(names(food3),"100g")

# View a summary of nutrition columns
summary(food3[,nutrition])
```

```
##  energy_from_fat_100g    fat_100g      saturated_fat_100g
##  Min.   :   0.00      Min.   :  0.00   Min.   : 0.000
##  1st Qu.:  35.98      1st Qu.:  0.90   1st Qu.: 0.200
##  Median : 237.00      Median :  6.00   Median : 1.700
##  Mean   : 668.41      Mean   : 13.39   Mean   : 4.874
##  3rd Qu.: 974.00      3rd Qu.: 20.00   3rd Qu.: 6.500
##  Max.   :2900.00      Max.   :100.00   Max.   :57.000
##  NA's   :1486         NA's   :708      NA's   :797
##  butyric_acid_100g caproic_acid_100g caprylic_acid_100g capric_acid_100g
##  Mode:logical      Mode:logical      Mode:logical       Mode:logical
##  NA's:1500         NA's:1500         NA's:1500          NA's:1500
##
##
##
##
##
##  lauric_acid_100g myristic_acid_100g palmitic_acid_100g stearic_acid_100g
##  Mode:logical     Mode:logical       Mode:logical       Mode:logical
##  NA's:1500        NA's:1500          NA's:1500          NA's:1500
##
##
##
##
##
##  arachidic_acid_100g behenic_acid_100g lignoceric_acid_100g
##  Mode:logical        Mode:logical      Mode:logical
##  NA's:1500           NA's:1500         NA's:1500
##
##
##
##
##
##  cerotic_acid_100g montanic_acid_100g melissic_acid_100g
##  Mode:logical      Mode:logical       Mode:logical
##  NA's:1500         NA's:1500          NA's:1500
##
##
##
```

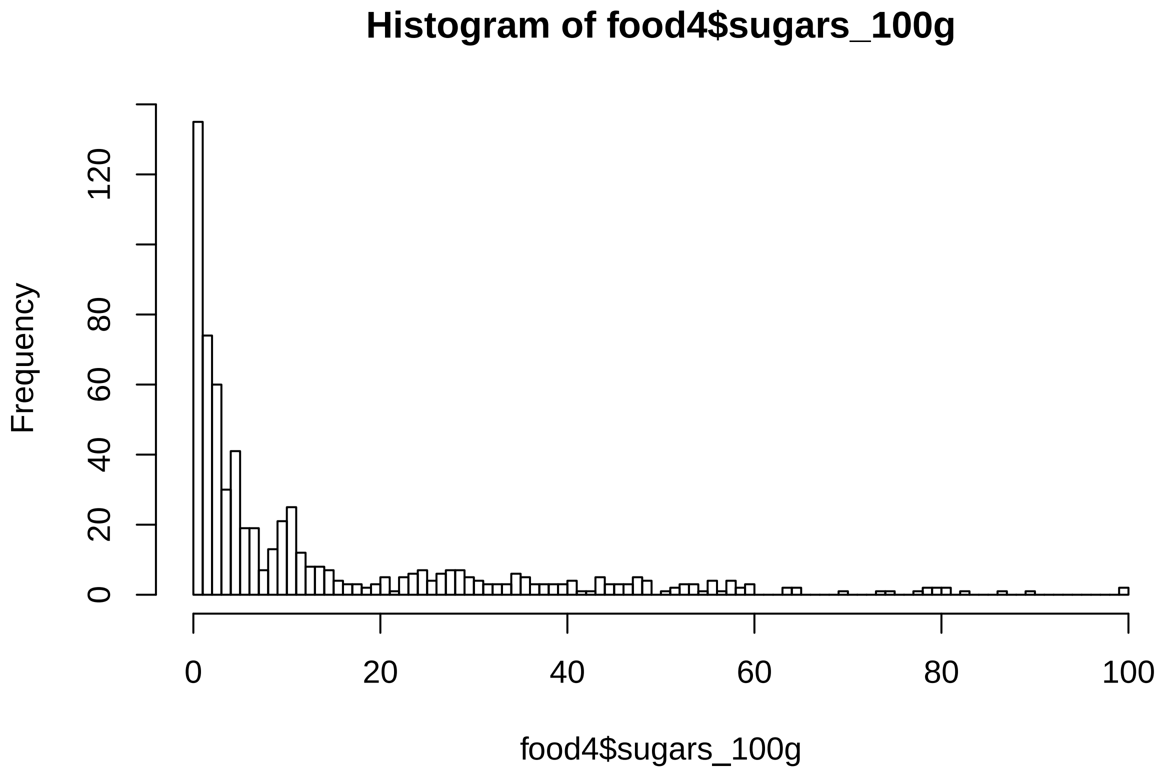
```
##
##
##  monounsaturated_fat_100g polyunsaturated_fat_100g omega_3_fat_100g
##  Min.   : 0.00            Min.   : 0.400           Min.   : 0.033
##  1st Qu.: 3.87            1st Qu.: 1.653           1st Qu.: 1.300
##  Median : 9.50            Median : 3.900           Median : 3.000
##  Mean   :19.77            Mean   : 9.986           Mean   : 3.726
##  3rd Qu.:29.00            3rd Qu.:12.700           3rd Qu.: 3.200
##  Max.   :75.00            Max.   :46.200           Max.   :12.400
##  NA's   :1465             NA's   :1464             NA's   :1491
##  alpha_linolenic_acid_100g eicosapentaenoic_acid_100g
##  Min.   :0.0800            Min.   :0.721
##  1st Qu.:0.0905            1st Qu.:0.721
##  Median :0.1010            Median :0.721
##  Mean   :0.1737            Mean   :0.721
##  3rd Qu.:0.2205            3rd Qu.:0.721
##  Max.   :0.3400            Max.   :0.721
##  NA's   :1497              NA's   :1499
##  docosahexaenoic_acid_100g omega_6_fat_100g linoleic_acid_100g
##  Min.   :1.09              Min.   :0.25     Min.   :0.5000
##  1st Qu.:1.09              1st Qu.:0.25     1st Qu.:0.5165
##  Median :1.09              Median :0.25     Median :0.5330
##  Mean   :1.09              Mean   :0.25     Mean   :0.5330
##  3rd Qu.:1.09              3rd Qu.:0.25     3rd Qu.:0.5495
##  Max.   :1.09              Max.   :0.25     Max.   :0.5660
##  NA's   :1499              NA's   :1499     NA's   :1498
##  arachidonic_acid_100g gamma_linolenic_acid_100g
##  Mode:logical          Mode:logical
##  NA's:1500             NA's:1500
##
##
##
##
##
##  dihomo_gamma_linolenic_acid_100g omega_9_fat_100g oleic_acid_100g
##  Mode:logical                     Mode:logical     Mode:logical
##  NA's:1500                        NA's:1500        NA's:1500
##
##
##
##
##
##  elaidic_acid_100g gondoic_acid_100g mead_acid_100g erucic_acid_100g
##  Mode:logical      Mode:logical      Mode:logical   Mode:logical
##  NA's:1500         NA's:1500         NA's:1500      NA's:1500
##
##
##
##
##
##  nervonic_acid_100g trans_fat_100g   cholesterol_100g carbohydrates_100g
##  Mode:logical       Min.   :0.0000   Min.   :0.0000   Min.   :  0.000
##  NA's:1500          1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:  3.792
##                     Median :0.0000   Median :0.0000   Median : 13.500
##                      Mean   :0.0105   Mean   :0.0265   Mean   : 27.958  
##                      3rd Qu.:0.0000   3rd Qu.:0.0026   3rd Qu.: 55.000  
##                      Max.   :0.1000   Max.   :0.4300   Max.   :100.000  
##                      NA's   :1481     NA's   :1477     NA's   :708      
##   sugars_100g     sucrose_100g   glucose_100g   fructose_100g
##  Min.   :  0.00   Mode:logical   Mode:logical   Min.   :100   
##  1st Qu.:  1.00   NA's:1500      NA's:1500      1st Qu.:100   
##  Median :  4.05                                 Median :100   
##  Mean   : 12.66                                 Mean   :100   
##  3rd Qu.: 14.70                                 3rd Qu.:100   
##  Max.   :100.00                                 Max.   :100   
##  NA's   :788                                    NA's   :1499  
##   lactose_100g   maltose_100g   maltodextrins_100g  starch_100g   
##  Min.   :0.000   Mode:logical   Mode:logical       Min.   : 0.00  
##  1st Qu.:0.250   NA's:1500      NA's:1500          1st Qu.: 9.45  
##  Median :0.500                                     Median :39.50  
##  Mean   :2.933                                     Mean   :30.73  
##  3rd Qu.:4.400                                     3rd Qu.:42.85  
##  Max.   :8.300                                     Max.   :71.00  
##  NA's   :1497                                      NA's   :1493  
##   polyols_100g     fiber_100g     proteins_100g     casein_100g  
##  Min.   : 8.60   Min.   : 0.000   Min.   : 0.000   Min.   :1.1   
##  1st Qu.:59.10   1st Qu.: 0.500   1st Qu.: 1.500   1st Qu.:1.1   
##  Median :67.00   Median : 1.750   Median : 6.000   Median :1.1   
##  Mean   :56.06   Mean   : 2.823   Mean   : 7.563   Mean   :1.1   
##  3rd Qu.:69.80   3rd Qu.: 3.500   3rd Qu.:10.675   3rd Qu.:1.1   
##  Max.   :70.00   Max.   :46.700   Max.   :61.000   Max.   :1.1   
##  NA's   :1491    NA's   :994      NA's   :710      NA's   :1499  
##  serum_proteins_100g nucleotides_100g   salt_100g         sodium_100g    
##  Mode:logical        Mode:logical     Min.   :  0.0000   Min.   : 0.0000  
##  NA's:1500           NA's:1500        1st Qu.:  0.0438   1st Qu.: 0.0172  
##                                       Median :  0.4498   Median : 0.1771  
##                                       Mean   :  1.1205   Mean   : 0.4409  
##                                       3rd Qu.:  1.1938   3rd Qu.: 0.4700  
##                                       Max.   :102.0000   Max.   :40.0000  
##                                       NA's   :780        NA's   :780      
##   alcohol_100g   vitamin_a_100g   beta_carotene_100g vitamin_d_100g
##  Min.   : 0.00   Min.   :0.0000   Mode:logical       Min.   :0e+00  
##  1st Qu.: 0.00   1st Qu.:0.0000   NA's:1500          1st Qu.:0e+00  
##  Median : 5.50   Median :0.0001                      Median :0e+00  
##  Mean   :10.07   Mean   :0.0003                      Mean   :0e+00  
##  3rd Qu.:13.00   3rd Qu.:0.0006                      3rd Qu.:0e+00  
##  Max.   :50.00   Max.   :0.0013                      Max.   :1e-04  
##  NA's   :1433    NA's   :1477                        NA's   :1485   
##  vitamin_e_100g   vitamin_k_100g vitamin_c_100g  vitamin_b1_100g 
##  Min.   :0.0005   Min.   :0      Min.   :0.000   Min.   :0.0001  
##  1st Qu.:0.0021   1st Qu.:0      1st Qu.:0.002   1st Qu.:0.0003  
##  Median :0.0044   Median :0      Median :0.019   Median :0.0004  
##  Mean   :0.0069   Mean   :0      Mean   :0.025   Mean   :0.0006  
##  3rd Qu.:0.0097   3rd Qu.:0      3rd Qu.:0.030   3rd Qu.:0.0010  
##  Max.   :0.0320   Max.   :0      Max.   :0.217   Max.   :0.0013  
##  NA's   :1478     NA's   :1498   NA's   :1459    NA's   :1478    
##  vitamin_b2_100g  vitamin_pp_100g  vitamin_b6_100g  vitamin_b9_100g
##  Min.   :0.0002   Min.   :0.0006   Min.   :0.0001   Min.   :0e+00  
##  1st Qu.:0.0003   1st Qu.:0.0033   1st Qu.:0.0002   1st Qu.:0e+00  
##  Median :0.0009   Median :0.0069   Median :0.0008   Median :1e-04  
##  Mean   :0.0011   Mean   :0.0086   Mean   :0.0112   Mean   :1e-04  
##  3rd Qu.:0.0013   3rd Qu.:0.0140   3rd Qu.:0.0012   3rd Qu.:2e-04  
##  Max.   :0.0066   Max.   :0.0160   Max.   :0.2000   Max.   :2e-04  
##  NA's   :1483     NA's   :1484     NA's   :1481     NA's   :1483   
##  vitamin_b12_100g  biotin_100g   pantothenic_acid_100g  silica_100g   
##  Min.   :0        Min.   :0      Min.   :0.0000        Min.   :8e-04  
##  1st Qu.:0        1st Qu.:0      1st Qu.:0.0007        1st Qu.:8e-04  
##  Median :0        Median :0      Median :0.0020        Median :8e-04  
##  Mean   :0        Mean   :0      Mean   :0.0027        Mean   :8e-04  
##  3rd Qu.:0        3rd Qu.:0      3rd Qu.:0.0051        3rd Qu.:8e-04  
##  Max.   :0        Max.   :0      Max.   :0.0060        Max.   :8e-04  
##  NA's   :1489     NA's   :1498   NA's   :1486          NA's   :1499   
##  bicarbonate_100g potassium_100g   chloride_100g     calcium_100g    
##  Min.   :0.0006   Min.   :0.0000   Min.   :0.0003   Min.   :0.0000  
##  1st Qu.:0.0678   1st Qu.:0.0650   1st Qu.:0.0006   1st Qu.:0.0450  
##  Median :0.1350   Median :0.1940   Median :0.0009   Median :0.1200  
##  Mean   :0.1692   Mean   :0.3288   Mean   :0.0144   Mean   :0.2040  
##  3rd Qu.:0.2535   3rd Qu.:0.3670   3rd Qu.:0.0214   3rd Qu.:0.1985  
##  Max.   :0.3720   Max.   :1.4300   Max.   :0.0420   Max.   :1.0000  
##  NA's   :1497     NA's   :1487     NA's   :1497     NA's   :1449    
##  phosphorus_100g    iron_100g      magnesium_100g     zinc_100g     
##  Min.   :0.0430   Min.   :0.0000   Min.   :0.0000   Min.   :0.0005  
##  1st Qu.:0.1938   1st Qu.:0.0012   1st Qu.:0.0670   1st Qu.:0.0009  
##  Median :0.3185   Median :0.0042   Median :0.1040   Median :0.0017  
##  Mean   :0.3777   Mean   :0.0045   Mean   :0.1066   Mean   :0.0016  
##  3rd Qu.:0.4340   3rd Qu.:0.0077   3rd Qu.:0.1300   3rd Qu.:0.0022  
##  Max.   :1.1550   Max.   :0.0137   Max.   :0.3330   Max.   :0.0026  
##  NA's   :1488     NA's   :1463     NA's   :1479     NA's   :1493    
##   copper_100g    manganese_100g fluoride_100g  selenium_100g 
##  Min.   :0e+00   Min.   :0      Min.   :0      Min.   :0      
##  1st Qu.:1e-04   1st Qu.:0      1st Qu.:0      1st Qu.:0      
##  Median :1e-04   Median :0      Median :0      Median :0      
##  Mean   :1e-04   Mean   :0      Mean   :0      Mean   :0      
##  3rd Qu.:1e-04   3rd Qu.:0      3rd Qu.:0      3rd Qu.:0      
##  Max.   :1e-04   Max.   :0      Max.   :0      Max.   :0      
##  NA's   :1498    NA's   :1499   NA's   :1498   NA's   :1499   
##  chromium_100g  molybdenum_100g  iodine_100g   caffeine_100g 
##  Mode:logical   Mode:logical    Min.   :0      Mode:logical  
##  NA's:1500      NA's:1500       1st Qu.:0      NA's:1500     
##                                 Median :0                    
##                                 Mean   :0                    
##                                 3rd Qu.:0                    
##                                 Max.   :0                    
##                                 NA's   :1499                 
##  taurine_100g   ph_100g        fruits_vegetables_nuts_100g
##  Mode:logical   Mode:logical   Min.   : 2.00              
##  NA's:1500      NA's:1500      1st Qu.:11.25              
##                                Median :42.00              
##                                Mean   :36.88              
##                                3rd Qu.:52.25              
##                                Max.   :80.00              
##                                NA's   :1470               
##  collagen_meat_protein_ratio_100g   cocoa_100g   chlorophyl_100g
##  Min.   :12.00                    Min.   :30     Mode:logical
##  1st Qu.:13.50                    1st Qu.:47     NA's:1500
##  Median :15.00                    Median :60
##  Mean   :15.67                    Mean   :57
##  3rd Qu.:17.50                    3rd Qu.:70
##  Max.   :20.00                    Max.   :81
##  NA's   :1497                     NA's   :1491
##  nutrition_score_fr_100g nutrition_score_uk_100g
##  Min.   :-12.000         Min.   :-12.000
##  1st Qu.:  1.000         1st Qu.:  0.000
##  Median :  7.000         Median :  6.000
##  Mean   :  7.941         Mean   :  7.631
##  3rd Qu.: 15.000         3rd Qu.: 16.000
##  Max.   : 28.000         Max.   : 28.000
##  NA's   :825             NA's   :825
```

Replacing missing values

```r
# Find indices of sugar NA values: missing
missing <- is.na(food3$sugars_100g)

# Replace NA values with 0
food3$sugars_100g[missing] <- 0

# Create first histogram
hist(food3$sugars_100g, breaks = 100)
```



```
# Create food4
food4 <- food3[food3$sugars_100g > 0, ]

# Create second histogram
hist(food4$sugars_100g, breaks = 100)
```



Dealing with messy data

```
# Find entries containing "plasti": plastic
plastic <- str_detect(food3$packaging,"plasti")

# Print the sum of plastic
sum(plastic)
```

```
## [1] 232
```