Data Cleaning

Carlos Tejeda Ocampo A011170840

9 de julio de 2020

# R Data Cleaning

## Ticket Sales Dataset

Data Cleaning of messy Ticket Sales Dataset

#sales <- read.csv("sales.csv", stringsAsFactors = FALSE)  
  
url\_sales <- 'http://s3.amazonaws.com/assets.datacamp.com/production/course\_1294/datasets/sales.csv'  
sales <- read.csv(url\_sales, stringsAsFactors=F)

### Examine Data

dim(sales)

## [1] 5000 46

#head(sales)  
names(sales)

## [1] "X" "event\_id" "primary\_act\_id"   
## [4] "secondary\_act\_id" "purch\_party\_lkup\_id" "event\_name"   
## [7] "primary\_act\_name" "secondary\_act\_name" "major\_cat\_name"   
## [10] "minor\_cat\_name" "la\_event\_type\_cat" "event\_disp\_name"   
## [13] "ticket\_text" "tickets\_purchased\_qty" "trans\_face\_val\_amt"   
## [16] "delivery\_type\_cd" "event\_date\_time" "event\_dt"   
## [19] "presale\_dt" "onsale\_dt" "sales\_ord\_create\_dttm"   
## [22] "sales\_ord\_tran\_dt" "print\_dt" "timezn\_nm"   
## [25] "venue\_city" "venue\_state" "venue\_postal\_cd\_sgmt\_1"  
## [28] "sales\_platform\_cd" "print\_flg" "la\_valid\_tkt\_event\_flg"  
## [31] "fin\_mkt\_nm" "web\_session\_cookie\_val" "gndr\_cd"   
## [34] "age\_yr" "income\_amt" "edu\_val"   
## [37] "edu\_1st\_indv\_val" "edu\_2nd\_indv\_val" "adults\_in\_hh\_num"   
## [40] "married\_ind" "child\_present\_ind" "home\_owner\_ind"   
## [43] "occpn\_val" "occpn\_1st\_val" "occpn\_2nd\_val"   
## [46] "dist\_to\_ven"

### Sumarize Data

#str(sales)  
#summary(sales)  
glimpse(sales)

## Observations: 5,000  
## Variables: 46  
## $ X <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 1...  
## $ event\_id <chr> "abcaf1adb99a935fc661", "6c56d7f08c95f2aa453...  
## $ primary\_act\_id <chr> "43f0436b905bfa7c2eec", "1a3e9aecd0617706a79...  
## $ secondary\_act\_id <chr> "b85143bf51323b72e53c", "f53529c5679ea6ca5a4...  
## $ purch\_party\_lkup\_id <chr> "7dfa56dd7d5956b17587", "4f9e6fc637eaf7b736c...  
## $ event\_name <chr> "Xfinity Center Mansfield Premier Parking: F...  
## $ primary\_act\_name <chr> "XFINITY Center Mansfield Premier Parking", ...  
## $ secondary\_act\_name <chr> "NULL", "Dave Matthews Band", "NULL", "NULL"...  
## $ major\_cat\_name <chr> "MISC", "MISC", "MISC", "MISC", "MISC", "MIS...  
## $ minor\_cat\_name <chr> "PARKING", "CAMPING", "PARKING", "PARKING", ...  
## $ la\_event\_type\_cat <chr> "PARKING", "INVALID", "PARKING", "PARKING", ...  
## $ event\_disp\_name <chr> "Xfinity Center Mansfield Premier Parking: F...  
## $ ticket\_text <chr> " THIS TICKET IS VALID FOR PARKING ...  
## $ tickets\_purchased\_qty <int> 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 2, 4, 1, 1,...  
## $ trans\_face\_val\_amt <dbl> 45, 75, 5, 20, 20, 10, 30, 28, 20, 25, 20, 9...  
## $ delivery\_type\_cd <chr> "eTicket", "TicketFast", "TicketFast", "Mail...  
## $ event\_date\_time <chr> "2015-09-12 23:30:00", "2009-09-05 01:00:00"...  
## $ event\_dt <chr> "2015-09-12", "2009-09-04", "2006-04-21", "2...  
## $ presale\_dt <chr> "NULL", "NULL", "NULL", "NULL", "2005-03-02"...  
## $ onsale\_dt <chr> "2015-05-15", "2009-03-13", "2006-02-25", "2...  
## $ sales\_ord\_create\_dttm <chr> "2015-09-11 18:17:45", "2009-07-06 00:00:00"...  
## $ sales\_ord\_tran\_dt <chr> "2015-09-11", "2009-07-05", "2006-04-05", "2...  
## $ print\_dt <chr> "2015-09-12", "2009-09-01", "2006-04-05", "2...  
## $ timezn\_nm <chr> "EST", "PST", "MST", "CST", "PST", "PST", "E...  
## $ venue\_city <chr> "MANSFIELD", "QUINCY", "PHOENIX", "DALLAS", ...  
## $ venue\_state <chr> "MASSACHUSETTS", "WASHINGTON", "ARIZONA", "T...  
## $ venue\_postal\_cd\_sgmt\_1 <chr> "02048", "98848", "85003", "75210", "98092",...  
## $ sales\_platform\_cd <chr> "www.concerts.livenation.com", "NULL", "NULL...  
## $ print\_flg <chr> "T ", "T ", "T ", "T ", "T ", "T ", "T ", "T...  
## $ la\_valid\_tkt\_event\_flg <chr> "N ", "N ", "N ", "N ", "N ", "N ", "N ", "N...  
## $ fin\_mkt\_nm <chr> "Boston", "Seattle", "Arizona", "Dallas", "S...  
## $ web\_session\_cookie\_val <chr> "7dfa56dd7d5956b17587", "4f9e6fc637eaf7b736c...  
## $ gndr\_cd <chr> NA, NA, NA, NA, NA, NA, "M", NA, NA, NA, "M"...  
## $ age\_yr <chr> NA, NA, NA, NA, NA, NA, "28", NA, NA, NA, "8...  
## $ income\_amt <chr> NA, NA, NA, NA, NA, NA, "112500", NA, NA, NA...  
## $ edu\_val <chr> NA, NA, NA, NA, NA, NA, "High School", NA, N...  
## $ edu\_1st\_indv\_val <chr> NA, NA, NA, NA, NA, NA, "High School", NA, N...  
## $ edu\_2nd\_indv\_val <chr> NA, NA, NA, NA, NA, NA, "NULL", NA, NA, NA, ...  
## $ adults\_in\_hh\_num <chr> NA, NA, NA, NA, NA, NA, "4", NA, NA, NA, "2"...  
## $ married\_ind <chr> NA, NA, NA, NA, NA, NA, "0", NA, NA, NA, "1"...  
## $ child\_present\_ind <chr> NA, NA, NA, NA, NA, NA, "1", NA, NA, NA, "NU...  
## $ home\_owner\_ind <chr> NA, NA, NA, NA, NA, NA, "0", NA, NA, NA, "1"...  
## $ occpn\_val <chr> NA, NA, NA, NA, NA, NA, "NULL", NA, NA, NA, ...  
## $ occpn\_1st\_val <chr> NA, NA, NA, NA, NA, NA, "Craftsman Blue Coll...  
## $ occpn\_2nd\_val <chr> NA, NA, NA, NA, NA, NA, "NULL", NA, NA, NA, ...  
## $ dist\_to\_ven <int> NA, 59, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...

### Remove unnecesary columns

The first column is redundant, the second, third and fourth columns contains no relevant information and the last sixteen columns have too many missing values, so we remove them.

sales2 <- sales[,c(6:30)]  
glimpse(sales2)

## Observations: 5,000  
## Variables: 25  
## $ event\_name <chr> "Xfinity Center Mansfield Premier Parking: F...  
## $ primary\_act\_name <chr> "XFINITY Center Mansfield Premier Parking", ...  
## $ secondary\_act\_name <chr> "NULL", "Dave Matthews Band", "NULL", "NULL"...  
## $ major\_cat\_name <chr> "MISC", "MISC", "MISC", "MISC", "MISC", "MIS...  
## $ minor\_cat\_name <chr> "PARKING", "CAMPING", "PARKING", "PARKING", ...  
## $ la\_event\_type\_cat <chr> "PARKING", "INVALID", "PARKING", "PARKING", ...  
## $ event\_disp\_name <chr> "Xfinity Center Mansfield Premier Parking: F...  
## $ ticket\_text <chr> " THIS TICKET IS VALID FOR PARKING ...  
## $ tickets\_purchased\_qty <int> 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 2, 4, 1, 1,...  
## $ trans\_face\_val\_amt <dbl> 45, 75, 5, 20, 20, 10, 30, 28, 20, 25, 20, 9...  
## $ delivery\_type\_cd <chr> "eTicket", "TicketFast", "TicketFast", "Mail...  
## $ event\_date\_time <chr> "2015-09-12 23:30:00", "2009-09-05 01:00:00"...  
## $ event\_dt <chr> "2015-09-12", "2009-09-04", "2006-04-21", "2...  
## $ presale\_dt <chr> "NULL", "NULL", "NULL", "NULL", "2005-03-02"...  
## $ onsale\_dt <chr> "2015-05-15", "2009-03-13", "2006-02-25", "2...  
## $ sales\_ord\_create\_dttm <chr> "2015-09-11 18:17:45", "2009-07-06 00:00:00"...  
## $ sales\_ord\_tran\_dt <chr> "2015-09-11", "2009-07-05", "2006-04-05", "2...  
## $ print\_dt <chr> "2015-09-12", "2009-09-01", "2006-04-05", "2...  
## $ timezn\_nm <chr> "EST", "PST", "MST", "CST", "PST", "PST", "E...  
## $ venue\_city <chr> "MANSFIELD", "QUINCY", "PHOENIX", "DALLAS", ...  
## $ venue\_state <chr> "MASSACHUSETTS", "WASHINGTON", "ARIZONA", "T...  
## $ venue\_postal\_cd\_sgmt\_1 <chr> "02048", "98848", "85003", "75210", "98092",...  
## $ sales\_platform\_cd <chr> "www.concerts.livenation.com", "NULL", "NULL...  
## $ print\_flg <chr> "T ", "T ", "T ", "T ", "T ", "T ", "T ", "T...  
## $ la\_valid\_tkt\_event\_flg <chr> "N ", "N ", "N ", "N ", "N ", "N ", "N ", "N...

### Separating information from columns

event\_date\_time and sales\_ord\_create\_dttm include both date and hour in the same value. We need to separate these values to facilitate the analysis.

sales3 <- separate(sales2, event\_date\_time, c('event\_dt', 'event\_time'), sep = " ")  
col <- str\_detect(names(sales3),'event')  
glimpse(sales3[,col])

## Observations: 5,000  
## Variables: 6  
## $ event\_name <chr> "Xfinity Center Mansfield Premier Parking: F...  
## $ la\_event\_type\_cat <chr> "PARKING", "INVALID", "PARKING", "PARKING", ...  
## $ event\_disp\_name <chr> "Xfinity Center Mansfield Premier Parking: F...  
## $ event\_dt <chr> "2015-09-12", "2009-09-05", "2006-04-22", "2...  
## $ event\_time <chr> "23:30:00", "01:00:00", "01:30:00", "00:00:0...  
## $ la\_valid\_tkt\_event\_flg <chr> "N ", "N ", "N ", "N ", "N ", "N ", "N ", "N...

head(sales3$sales\_ord\_create\_dttm)

## [1] "2015-09-11 18:17:45" "2009-07-06 00:00:00" "2006-04-05 00:00:00"  
## [4] "2011-07-01 17:38:50" "2005-06-18 00:00:00" "2012-07-21 17:20:18"

# Split sales\_ord\_create\_dttm: sales5  
sales4 <- separate(sales3, sales\_ord\_create\_dttm, c('ord\_create\_dt', 'ord\_create\_time'), sep = " ")

## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 4 rows [2516,  
## 3863, 4082, 4183].

## check new columns  
col <- str\_detect(names(sales4),'ord\_create')  
glimpse(sales4[,col])

## Observations: 5,000  
## Variables: 2  
## $ ord\_create\_dt <chr> "2015-09-11", "2009-07-06", "2006-04-05", "2011-07-...  
## $ ord\_create\_time <chr> "18:17:45", "00:00:00", "00:00:00", "17:38:50", "00...

### Identifying and formatting date and time

date\_cols <- str\_detect(colnames(sales4), "dt")  
glimpse(sales4[,date\_cols])

## Observations: 5,000  
## Variables: 6  
## $ event\_dt <chr> "2015-09-12", "2009-09-05", "2006-04-22", "2011-0...  
## $ presale\_dt <chr> "NULL", "NULL", "NULL", "NULL", "2005-03-02", "NU...  
## $ onsale\_dt <chr> "2015-05-15", "2009-03-13", "2006-02-25", "2011-0...  
## $ ord\_create\_dt <chr> "2015-09-11", "2009-07-06", "2006-04-05", "2011-0...  
## $ sales\_ord\_tran\_dt <chr> "2015-09-11", "2009-07-05", "2006-04-05", "2011-0...  
## $ print\_dt <chr> "2015-09-12", "2009-09-01", "2006-04-05", "2011-0...

sales4[, date\_cols] <- lapply(sales4[, date\_cols], ymd)

## Warning: 2892 failed to parse.

## Warning: 101 failed to parse.

## Warning: 4 failed to parse.

## Warning: 424 failed to parse.

glimpse(sales4[,date\_cols])

## Observations: 5,000  
## Variables: 6  
## $ event\_dt <date> 2015-09-12, 2009-09-05, 2006-04-22, 2011-09-03, ...  
## $ presale\_dt <date> NA, NA, NA, NA, 2005-03-02, NA, NA, NA, NA, NA, ...  
## $ onsale\_dt <date> 2015-05-15, 2009-03-13, 2006-02-25, 2011-04-22, ...  
## $ ord\_create\_dt <date> 2015-09-11, 2009-07-06, 2006-04-05, 2011-07-01, ...  
## $ sales\_ord\_tran\_dt <date> 2015-09-11, 2009-07-05, 2006-04-05, 2011-07-01, ...  
## $ print\_dt <date> 2015-09-12, 2009-09-01, 2006-04-05, 2011-07-06, ...

time\_cols <- str\_detect(colnames(sales4), "\_time")  
glimpse(sales4[,time\_cols])

## Observations: 5,000  
## Variables: 2  
## $ event\_time <chr> "23:30:00", "01:00:00", "01:30:00", "00:00:00", "01...  
## $ ord\_create\_time <chr> "18:17:45", "00:00:00", "00:00:00", "17:38:50", "00...

sales4[, time\_cols] <- lapply(sales4[, time\_cols], hms)

## Warning in .parse\_hms(..., order = "HMS", quiet = quiet): Some strings failed to  
## parse, or all strings are NAs

glimpse(sales4[,time\_cols])

## Observations: 5,000  
## Variables: 2  
## $ event\_time <Period> 23H 30M 0S, 1H 0M 0S, 1H 30M 0S, 0S, 1H 0M 0S, 2...  
## $ ord\_create\_time <Period> 18H 17M 45S, 0S, 0S, 17H 38M 50S, 0S, 17H 20M 18...

### Combine venue\_city and venue\_state columns

sales5 <- unite(sales4, venue\_city\_state, venue\_city, venue\_state, sep = ", ")  
glimpse(sales5)

## Observations: 5,000  
## Variables: 25  
## $ event\_name <chr> "Xfinity Center Mansfield Premier Parking: F...  
## $ primary\_act\_name <chr> "XFINITY Center Mansfield Premier Parking", ...  
## $ secondary\_act\_name <chr> "NULL", "Dave Matthews Band", "NULL", "NULL"...  
## $ major\_cat\_name <chr> "MISC", "MISC", "MISC", "MISC", "MISC", "MIS...  
## $ minor\_cat\_name <chr> "PARKING", "CAMPING", "PARKING", "PARKING", ...  
## $ la\_event\_type\_cat <chr> "PARKING", "INVALID", "PARKING", "PARKING", ...  
## $ event\_disp\_name <chr> "Xfinity Center Mansfield Premier Parking: F...  
## $ ticket\_text <chr> " THIS TICKET IS VALID FOR PARKING ...  
## $ tickets\_purchased\_qty <int> 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 2, 4, 1, 1,...  
## $ trans\_face\_val\_amt <dbl> 45, 75, 5, 20, 20, 10, 30, 28, 20, 25, 20, 9...  
## $ delivery\_type\_cd <chr> "eTicket", "TicketFast", "TicketFast", "Mail...  
## $ event\_dt <date> 2015-09-12, 2009-09-05, 2006-04-22, 2011-09...  
## $ event\_time <Period> 23H 30M 0S, 1H 0M 0S, 1H 30M 0S, 0S, 1H 0...  
## $ presale\_dt <date> NA, NA, NA, NA, 2005-03-02, NA, NA, NA, NA,...  
## $ onsale\_dt <date> 2015-05-15, 2009-03-13, 2006-02-25, 2011-04...  
## $ ord\_create\_dt <date> 2015-09-11, 2009-07-06, 2006-04-05, 2011-07...  
## $ ord\_create\_time <Period> 18H 17M 45S, 0S, 0S, 17H 38M 50S, 0S, 17H...  
## $ sales\_ord\_tran\_dt <date> 2015-09-11, 2009-07-05, 2006-04-05, 2011-07...  
## $ print\_dt <date> 2015-09-12, 2009-09-01, 2006-04-05, 2011-07...  
## $ timezn\_nm <chr> "EST", "PST", "MST", "CST", "PST", "PST", "E...  
## $ venue\_city\_state <chr> "MANSFIELD, MASSACHUSETTS", "QUINCY, WASHING...  
## $ venue\_postal\_cd\_sgmt\_1 <chr> "02048", "98848", "85003", "75210", "98092",...  
## $ sales\_platform\_cd <chr> "www.concerts.livenation.com", "NULL", "NULL...  
## $ print\_flg <chr> "T ", "T ", "T ", "T ", "T ", "T ", "T ", "T...  
## $ la\_valid\_tkt\_event\_flg <chr> "N ", "N ", "N ", "N ", "N ", "N ", "N ", "N...

## MBTA Ridership Data

Data Cleaning of messy MBTA Ridership Data

#url\_mbta <- 'http://s3.amazonaws.com/assets.datacamp.com/production/course\_1294/datasets/mbta.xlsx'  
#download.file(url\_mbta, 'mbta.xlsx')  
mbta <- read\_excel('mbta.xlsx', skip=1)

## New names:  
## \* `` -> ...1

glimpse(mbta)

## Observations: 11  
## Variables: 60  
## $ ...1 <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11  
## $ mode <chr> "All Modes by Qtr", "Boat", "Bus", "Commuter Rail", "Heav...  
## $ `2007-01` <chr> "NA", "4", "335.819", "142.2", "435.294", "227.231", "0.0...  
## $ `2007-02` <chr> "NA", "3.6", "338.675", "138.5", "448.271", "240.262", "-...  
## $ `2007-03` <dbl> 1187.653, 40.000, 339.867, 137.700, 458.583, 241.444, 0.1...  
## $ `2007-04` <chr> "NA", "4.3", "352.162", "139.5", "472.201", "255.557", "-...  
## $ `2007-05` <chr> "NA", "4.9", "354.367", "139", "474.579", "248.262", "0.0...  
## $ `2007-06` <dbl> 1245.959, 5.800, 350.543, 143.000, 477.032, 246.108, 0.09...  
## $ `2007-07` <chr> "NA", "6.521", "357.519", "142.391", "471.735", "243.286"...  
## $ `2007-08` <chr> "NA", "6.572", "355.479", "142.364", "461.605", "234.907"...  
## $ `2007-09` <dbl> 1256.571, 5.469, 372.598, 143.051, 499.566, 265.748, -0.0...  
## $ `2007-10` <chr> "NA", "5.145", "368.847", "146.542", "457.741", "241.434"...  
## $ `2007-11` <chr> "NA", "3.763", "330.826", "145.089", "488.348", "250.497"...  
## $ `2007-12` <dbl> 1216.890, 2.985, 312.920, 141.585, 448.268, 233.379, -0.0...  
## $ `2008-01` <chr> "NA", "3.175", "340.324", "142.145", "472.624", "241.223"...  
## $ `2008-02` <chr> "NA", "3.111", "352.905", "142.607", "492.1", "249.306", ...  
## $ `2008-03` <dbl> 1253.522, 3.512, 361.155, 137.453, 494.046, 253.132, 0.05...  
## $ `2008-04` <chr> "NA", "4.164", "368.189", "140.389", "513.204", "271.07",...  
## $ `2008-05` <chr> "NA", "4.015", "363.903", "142.585", "507.952", "258.351"...  
## $ `2008-06` <dbl> 1314.821, 5.189, 362.962, 142.057, 518.349, 266.961, 0.06...  
## $ `2008-07` <chr> "NA", "6.016", "370.921", "145.731", "512.309", "270.158"...  
## $ `2008-08` <chr> "NA", "5.8", "361.057", "144.565", "476.99", "239.344", "...  
## $ `2008-09` <dbl> 1307.041, 4.587, 389.537, 141.907, 517.324, 258.171, 0.02...  
## $ `2008-10` <chr> "NA", "4.285", "357.974", "151.957", "523.644", "250.063"...  
## $ `2008-11` <chr> "NA", "3.488", "345.423", "152.952", "487.115", "232.068"...  
## $ `2008-12` <dbl> 1232.655, 3.007, 325.767, 140.810, 446.743, 205.420, -0.0...  
## $ `2009-01` <chr> "NA", "3.014", "338.532", "141.448", "461.004", "215.66",...  
## $ `2009-02` <chr> "NA", "3.196", "360.412", "143.529", "482.407", "228.737"...  
## $ `2009-03` <dbl> 1209.792, 3.330, 353.686, 142.893, 467.224, 222.844, -0.0...  
## $ `2009-04` <chr> "NA", "4.049", "359.38", "142.34", "493.152", "238.232", ...  
## $ `2009-05` <chr> "NA", "4.119", "354.75", "144.225", "475.634", "224.962",...  
## $ `2009-06` <dbl> 1233.085, 4.900, 347.865, 142.006, 473.099, 226.259, -0.0...  
## $ `2009-07` <chr> "NA", "6.444", "339.477", "137.691", "470.828", "230.308"...  
## $ `2009-08` <chr> "NA", "5.903", "332.661", "139.158", "466.676", "231.783"...  
## $ `2009-09` <dbl> 1230.461, 4.696, 374.260, 139.087, 500.403, 250.922, -0.0...  
## $ `2009-10` <chr> "NA", "4.212", "385.868", "137.104", "513.406", "230.739"...  
## $ `2009-11` <chr> "NA", "3.576", "366.98", "129.343", "480.278", "214.711",...  
## $ `2009-12` <dbl> 1207.845, 3.113, 332.394, 126.066, 440.925, 194.446, -0.0...  
## $ `2010-01` <chr> "NA", "3.207", "362.226", "130.91", "464.069", "204.396",...  
## $ `2010-02` <chr> "NA", "3.195", "361.138", "131.918", "480.121", "213.136"...  
## $ `2010-03` <dbl> 1208.857, 3.481, 373.443, 131.252, 483.397, 211.693, 0.01...  
## $ `2010-04` <chr> "NA", "4.452", "378.611", "131.722", "502.374", "227.246"...  
## $ `2010-05` <chr> "NA", "4.415", "380.171", "128.8", "487.4", "217.805", "0...  
## $ `2010-06` <dbl> 1244.409, 5.411, 363.275, 129.144, 490.263, 215.922, 0.00...  
## $ `2010-07` <chr> "NA", "6.513", "353.04", "122.935", "488.587", "218.729",...  
## $ `2010-08` <chr> "NA", "6.269", "343.688", "129.732", "473.731", "210.53",...  
## $ `2010-09` <dbl> 1225.527, 4.699, 381.622, 132.892, 521.099, 236.368, 0.00...  
## $ `2010-10` <chr> "NA", "4.402", "384.987", "131.033", "532.403", "236.366"...  
## $ `2010-11` <chr> "NA", "3.731", "367.955", "130.889", "502.887", "221.881"...  
## $ `2010-12` <dbl> 1216.262, 3.156, 326.338, 121.422, 450.433, 196.211, -0.0...  
## $ `2011-01` <chr> "NA", "3.14", "334.958", "128.396", "468.418", "198.45", ...  
## $ `2011-02` <chr> "NA", "3.284", "346.234", "125.463", "504.068", "219.886"...  
## $ `2011-03` <dbl> 1223.452, 3.674, 380.399, 134.374, 516.730, 227.935, 0.05...  
## $ `2011-04` <chr> "NA", "4.251", "380.446", "134.169", "528.631", "242.28",...  
## $ `2011-05` <chr> "NA", "4.431", "385.289", "136.14", "528.122", "225.776",...  
## $ `2011-06` <dbl> 1302.414, 5.474, 376.317, 135.581, 529.528, 221.865, 0.05...  
## $ `2011-07` <chr> "NA", "6.581", "361.585", "132.41", "532.888", "231.01", ...  
## $ `2011-08` <chr> "NA", "6.733", "353.793", "130.616", "508.145", "220.164"...  
## $ `2011-09` <dbl> 1290.549, 5.003, 388.271, 136.901, 550.137, 244.949, 0.04...  
## $ `2011-10` <chr> "NA", "4.484", "398.456", "128.72", "554.932", "237.768",...

### Remove unnecesary rows and columns

# Remove rows unuseful data  
keep <- !(mbta$mode %in% c('All Modes by Qtr', 'Pct Chg / Yr', 'TOTAL'))  
  
mbta2 <- mbta[keep,]  
mbta3 <- mbta2[,-1]  
glimpse(mbta3)

## Observations: 8  
## Variables: 59  
## $ mode <chr> "Boat", "Bus", "Commuter Rail", "Heavy Rail", "Light Rail...  
## $ `2007-01` <chr> "4", "335.819", "142.2", "435.294", "227.231", "4.772", "...  
## $ `2007-02` <chr> "3.6", "338.675", "138.5", "448.271", "240.262", "4.417",...  
## $ `2007-03` <dbl> 40.000, 339.867, 137.700, 458.583, 241.444, 4.574, 5.500,...  
## $ `2007-04` <chr> "4.3", "352.162", "139.5", "472.201", "255.557", "4.542",...  
## $ `2007-05` <chr> "4.9", "354.367", "139", "474.579", "248.262", "4.768", "...  
## $ `2007-06` <dbl> 5.800, 350.543, 143.000, 477.032, 246.108, 4.722, 5.600, ...  
## $ `2007-07` <chr> "6.521", "357.519", "142.391", "471.735", "243.286", "3.9...  
## $ `2007-08` <chr> "6.572", "355.479", "142.364", "461.605", "234.907", "3.9...  
## $ `2007-09` <dbl> 5.469, 372.598, 143.051, 499.566, 265.748, 4.329, 5.609, ...  
## $ `2007-10` <chr> "5.145", "368.847", "146.542", "457.741", "241.434", "4.3...  
## $ `2007-11` <chr> "3.763", "330.826", "145.089", "488.348", "250.497", "4.0...  
## $ `2007-12` <dbl> 2.985, 312.920, 141.585, 448.268, 233.379, 3.708, 5.062, ...  
## $ `2008-01` <chr> "3.175", "340.324", "142.145", "472.624", "241.223", "4.1...  
## $ `2008-02` <chr> "3.111", "352.905", "142.607", "492.1", "249.306", "4.079...  
## $ `2008-03` <dbl> 3.512, 361.155, 137.453, 494.046, 253.132, 4.175, 6.164, ...  
## $ `2008-04` <chr> "4.164", "368.189", "140.389", "513.204", "271.07", "4.14...  
## $ `2008-05` <chr> "4.015", "363.903", "142.585", "507.952", "258.351", "4.0...  
## $ `2008-06` <dbl> 5.189, 362.962, 142.057, 518.349, 266.961, 4.400, 6.294, ...  
## $ `2008-07` <chr> "6.016", "370.921", "145.731", "512.309", "270.158", "4.5...  
## $ `2008-08` <chr> "5.8", "361.057", "144.565", "476.99", "239.344", "4.286"...  
## $ `2008-09` <dbl> 4.587, 389.537, 141.907, 517.324, 258.171, 4.878, 6.503, ...  
## $ `2008-10` <chr> "4.285", "357.974", "151.957", "523.644", "250.063", "4.5...  
## $ `2008-11` <chr> "3.488", "345.423", "152.952", "487.115", "232.068", "4.3...  
## $ `2008-12` <dbl> 3.007, 325.767, 140.810, 446.743, 205.420, 3.420, 5.957, ...  
## $ `2009-01` <chr> "3.014", "338.532", "141.448", "461.004", "215.66", "3.58...  
## $ `2009-02` <chr> "3.196", "360.412", "143.529", "482.407", "228.737", "3.6...  
## $ `2009-03` <dbl> 3.330, 353.686, 142.893, 467.224, 222.844, 2.586, 6.675, ...  
## $ `2009-04` <chr> "4.049", "359.38", "142.34", "493.152", "238.232", "2.824...  
## $ `2009-05` <chr> "4.119", "354.75", "144.225", "475.634", "224.962", "2.73...  
## $ `2009-06` <dbl> 4.900, 347.865, 142.006, 473.099, 226.259, 2.763, 6.790, ...  
## $ `2009-07` <chr> "6.444", "339.477", "137.691", "470.828", "230.308", "2.6...  
## $ `2009-08` <chr> "5.903", "332.661", "139.158", "466.676", "231.783", "2.5...  
## $ `2009-09` <dbl> 4.696, 374.260, 139.087, 500.403, 250.922, 2.716, 6.690, ...  
## $ `2009-10` <chr> "4.212", "385.868", "137.104", "513.406", "230.739", "2.5...  
## $ `2009-11` <chr> "3.576", "366.98", "129.343", "480.278", "214.711", "2.72...  
## $ `2009-12` <dbl> 3.113, 332.394, 126.066, 440.925, 194.446, 2.365, 6.454, ...  
## $ `2010-01` <chr> "3.207", "362.226", "130.91", "464.069", "204.396", "2.68...  
## $ `2010-02` <chr> "3.195", "361.138", "131.918", "480.121", "213.136", "2.4...  
## $ `2010-03` <dbl> 3.481, 373.443, 131.252, 483.397, 211.693, 2.550, 7.066, ...  
## $ `2010-04` <chr> "4.452", "378.611", "131.722", "502.374", "227.246", "2.6...  
## $ `2010-05` <chr> "4.415", "380.171", "128.8", "487.4", "217.805", "2.686",...  
## $ `2010-06` <dbl> 5.411, 363.275, 129.144, 490.263, 215.922, 2.646, 7.460, ...  
## $ `2010-07` <chr> "6.513", "353.04", "122.935", "488.587", "218.729", "2.54...  
## $ `2010-08` <chr> "6.269", "343.688", "129.732", "473.731", "210.53", "2.43...  
## $ `2010-09` <dbl> 4.699, 381.622, 132.892, 521.099, 236.368, 2.652, 7.349, ...  
## $ `2010-10` <chr> "4.402", "384.987", "131.033", "532.403", "236.366", "2.6...  
## $ `2010-11` <chr> "3.731", "367.955", "130.889", "502.887", "221.881", "2.5...  
## $ `2010-12` <dbl> 3.156, 326.338, 121.422, 450.433, 196.211, 2.379, 7.199, ...  
## $ `2011-01` <chr> "3.14", "334.958", "128.396", "468.418", "198.45", "2.213...  
## $ `2011-02` <chr> "3.284", "346.234", "125.463", "504.068", "219.886", "2.5...  
## $ `2011-03` <dbl> 3.674, 380.399, 134.374, 516.730, 227.935, 2.559, 8.387, ...  
## $ `2011-04` <chr> "4.251", "380.446", "134.169", "528.631", "242.28", "2.76...  
## $ `2011-05` <chr> "4.431", "385.289", "136.14", "528.122", "225.776", "2.77...  
## $ `2011-06` <dbl> 5.474, 376.317, 135.581, 529.528, 221.865, 2.815, 8.377, ...  
## $ `2011-07` <chr> "6.581", "361.585", "132.41", "532.888", "231.01", "2.671...  
## $ `2011-08` <chr> "6.733", "353.793", "130.616", "508.145", "220.164", "2.6...  
## $ `2011-09` <dbl> 5.003, 388.271, 136.901, 550.137, 244.949, 2.843, 8.318, ...  
## $ `2011-10` <chr> "4.484", "398.456", "128.72", "554.932", "237.768", "2.96...

### Restructuring the data

The table is stored backwards, we should restructure it to improve readability

mbta4 <- gather(mbta3, month, thou\_riders, -mode)  
#change character to numeric data  
mbta4$thou\_riders <- as.numeric(mbta4$thou\_riders)  
mbta5 <- spread(mbta4, mode, thou\_riders)  
glimpse(mbta5)

## Observations: 58  
## Variables: 9  
## $ month <chr> "2007-01", "2007-02", "2007-03", "2007-04", "20...  
## $ Boat <dbl> 4.000, 3.600, 40.000, 4.300, 4.900, 5.800, 6.52...  
## $ Bus <dbl> 335.819, 338.675, 339.867, 352.162, 354.367, 35...  
## $ `Commuter Rail` <dbl> 142.200, 138.500, 137.700, 139.500, 139.000, 14...  
## $ `Heavy Rail` <dbl> 435.294, 448.271, 458.583, 472.201, 474.579, 47...  
## $ `Light Rail` <dbl> 227.231, 240.262, 241.444, 255.557, 248.262, 24...  
## $ `Private Bus` <dbl> 4.772, 4.417, 4.574, 4.542, 4.768, 4.722, 3.936...  
## $ RIDE <dbl> 4.900, 5.000, 5.500, 5.400, 5.400, 5.600, 5.253...  
## $ `Trackless Trolley` <dbl> 12.757, 12.913, 13.057, 13.444, 13.479, 13.323,...

head(mbta5)

## # A tibble: 6 x 9  
## month Boat Bus `Commuter Rail` `Heavy Rail` `Light Rail` `Private Bus`  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 2007~ 4 336. 142. 435. 227. 4.77  
## 2 2007~ 3.6 339. 138. 448. 240. 4.42  
## 3 2007~ 40 340. 138. 459. 241. 4.57  
## 4 2007~ 4.3 352. 140. 472. 256. 4.54  
## 5 2007~ 4.9 354. 139 475. 248. 4.77  
## 6 2007~ 5.8 351. 143 477. 246. 4.72  
## # ... with 2 more variables: RIDE <dbl>, `Trackless Trolley` <dbl>

mbta6 <- separate(mbta5, month, c('year', 'month'))  
  
head(mbta6)

## # A tibble: 6 x 10  
## year month Boat Bus `Commuter Rail` `Heavy Rail` `Light Rail`  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 2007 01 4 336. 142. 435. 227.  
## 2 2007 02 3.6 339. 138. 448. 240.  
## 3 2007 03 40 340. 138. 459. 241.  
## 4 2007 04 4.3 352. 140. 472. 256.  
## 5 2007 05 4.9 354. 139 475. 248.  
## 6 2007 06 5.8 351. 143 477. 246.  
## # ... with 3 more variables: `Private Bus` <dbl>, RIDE <dbl>, `Trackless  
## # Trolley` <dbl>

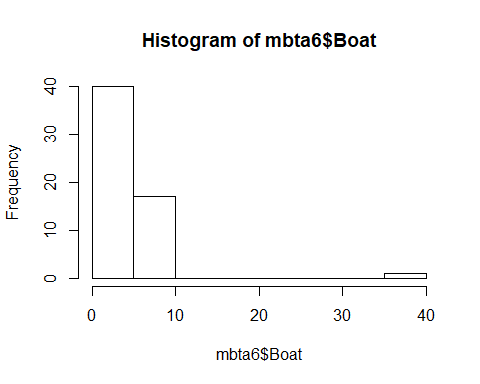
### Data entry error

Thre seems to be an entry error in boat

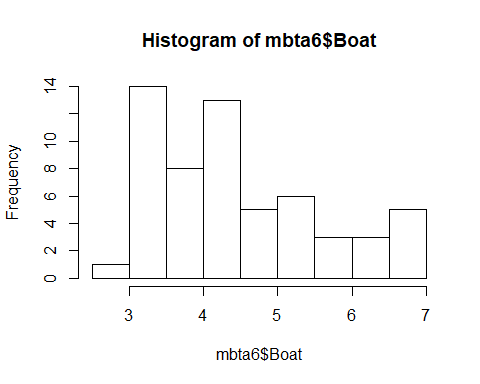
summary(mbta6)

## year month Boat Bus   
## Length:58 Length:58 Min. : 2.985 Min. :312.9   
## Class :character Class :character 1st Qu.: 3.494 1st Qu.:345.6   
## Mode :character Mode :character Median : 4.293 Median :359.9   
## Mean : 5.068 Mean :358.6   
## 3rd Qu.: 5.356 3rd Qu.:372.2   
## Max. :40.000 Max. :398.5   
## Commuter Rail Heavy Rail Light Rail Private Bus   
## Min. :121.4 Min. :435.3 Min. :194.4 Min. :2.213   
## 1st Qu.:131.4 1st Qu.:471.1 1st Qu.:220.6 1st Qu.:2.641   
## Median :138.8 Median :487.3 Median :231.9 Median :2.820   
## Mean :137.4 Mean :489.3 Mean :233.0 Mean :3.352   
## 3rd Qu.:142.4 3rd Qu.:511.3 3rd Qu.:244.5 3rd Qu.:4.167   
## Max. :153.0 Max. :554.9 Max. :271.1 Max. :4.878   
## RIDE Trackless Trolley  
## Min. :4.900 Min. : 5.777   
## 1st Qu.:5.965 1st Qu.:11.679   
## Median :6.615 Median :12.598   
## Mean :6.604 Mean :12.125   
## 3rd Qu.:7.149 3rd Qu.:13.320   
## Max. :8.598 Max. :15.109

hist(mbta6$Boat)

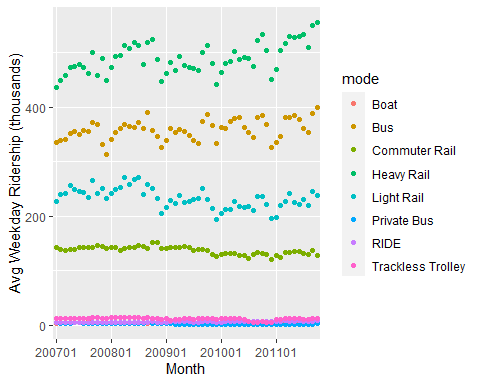
 Fixing the error

i <- which(mbta6$Boat == 40)  
  
  
mbta6$Boat[i] <- 4  
  
  
hist(mbta6$Boat)



### Plot the data

mbta\_all <- mbta6 %>%  
 unite(year\_mon, year, month, sep = "") %>%  
 gather(mode, thou\_riders, -year\_mon)  
  
ggplot(mbta\_all, aes(x = year\_mon, y = thou\_riders, col = mode)) + geom\_point() +   
 scale\_x\_discrete(name = "Month", breaks = c(200701, 200801, 200901, 201001, 201101)) +   
 scale\_y\_continuous(name = "Avg Weekday Ridership (thousands)")



## World Food Facts Sales Dataset

Data Cleaning of messy World Food Facts

url\_food <- 'http://s3.amazonaws.com/assets.datacamp.com/production/course\_1294/datasets/food.csv'  
food <- fread(url\_food)  
  
food <- data.frame(food)  
glimpse(food)

## Observations: 1,500  
## Variables: 160  
## $ V1 <int> 1, 2, 3, 4, 5, 6, 7, 8, ...  
## $ code <int> 100030, 100050, 100079, ...  
## $ url <chr> "http://world-en.openfoo...  
## $ creator <chr> "sebleouf", "foodorigins...  
## $ created\_t <int> 1424747544, 1450316429, ...  
## $ created\_datetime <chr> "2015-02-24T03:12:24Z", ...  
## $ last\_modified\_t <int> 1438445887, 1450817956, ...  
## $ last\_modified\_datetime <chr> "2015-08-01T16:18:07Z", ...  
## $ product\_name <chr> "Confiture de fraise fra...  
## $ generic\_name <chr> "", "", "PÃ¢tes de fruit...  
## $ quantity <chr> "265 g", "375g", "1 kg",...  
## $ packaging <chr> "Bocal,Verre", "Plastic,...  
## $ packaging\_tags <chr> "bocal,verre", "plastic,...  
## $ brands <chr> "Casino DÃ©lices", "Guyl...  
## $ brands\_tags <chr> "casino-delices", "guyli...  
## $ categories <chr> "Aliments et boissons Ã ...  
## $ categories\_tags <chr> "en:plant-based-foods-an...  
## $ categories\_en <chr> "Plant-based foods and b...  
## $ origins <chr> "", "", "", "", "Argenti...  
## $ origins\_tags <chr> "", "", "", "", "argenti...  
## $ manufacturing\_places <chr> "France", "Belgium", "",...  
## $ manufacturing\_places\_tags <chr> "france", "belgium", "",...  
## $ labels <chr> "", "", "", "Vegetariano...  
## $ labels\_tags <chr> "", "", "", "en:vegetari...  
## $ labels\_en <chr> "", "", "", "Vegetarian,...  
## $ emb\_codes <chr> "EMB 78015", "", "", "",...  
## $ emb\_codes\_tags <chr> "emb-78015", "", "", "",...  
## $ first\_packaging\_code\_geo <chr> "48.983333,2.066667", ""...  
## $ cities <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ cities\_tags <chr> "andresy-yvelines-france...  
## $ purchase\_places <chr> "Lyon,France", "NSW,Aust...  
## $ stores <chr> "Casino", "", "", "El Co...  
## $ countries <chr> "France", "Australia", "...  
## $ countries\_tags <chr> "en:france", "en:austral...  
## $ countries\_en <chr> "France", "Australia", "...  
## $ ingredients\_text <chr> "Sucre de canne, fraises...  
## $ allergens <chr> "", "", "", "", "", "", ...  
## $ allergens\_en <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ traces <chr> "Lait,Fruits Ã  coque", ...  
## $ traces\_tags <chr> "en:milk,en:nuts", "", "...  
## $ traces\_en <chr> "Milk,Nuts", "", "", "",...  
## $ serving\_size <chr> "15 g", "", "", "", "", ...  
## $ no\_nutriments <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ additives\_n <int> 1, NA, 2, 5, 0, NA, NA, ...  
## $ additives <chr> "[ sucre-de-canne -> fr:...  
## $ additives\_tags <chr> "en:e440", "", "en:e440,...  
## $ additives\_en <chr> "E440 - Pectins", "", "E...  
## $ ingredients\_from\_palm\_oil\_n <int> 0, NA, 0, 0, 0, NA, NA, ...  
## $ ingredients\_from\_palm\_oil <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ ingredients\_from\_palm\_oil\_tags <chr> "", "", "", "", "", "", ...  
## $ ingredients\_that\_may\_be\_from\_palm\_oil\_n <int> 0, NA, 0, 1, 0, NA, NA, ...  
## $ ingredients\_that\_may\_be\_from\_palm\_oil <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ ingredients\_that\_may\_be\_from\_palm\_oil\_tags <chr> "", "", "", "e471-mono-e...  
## $ nutrition\_grade\_uk <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ nutrition\_grade\_fr <chr> "d", "", "", "d", "d", "...  
## $ pnns\_groups\_1 <chr> "Sugary snacks", "Sugary...  
## $ pnns\_groups\_2 <chr> "Sweets", "Chocolate pro...  
## $ states <chr> "en:to-be-checked, en:co...  
## $ states\_tags <chr> "en:to-be-checked,en:com...  
## $ states\_en <chr> "To be checked,Complete,...  
## $ main\_category <chr> "en:plant-based-foods-an...  
## $ main\_category\_en <chr> "Plant-based foods and b...  
## $ image\_url <chr> "http://en.openfoodfacts...  
## $ image\_small\_url <chr> "http://en.openfoodfacts...  
## $ energy\_100g <dbl> 918, NA, NA, 766, 2359, ...  
## $ energy\_from\_fat\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ fat\_100g <dbl> 0.00, NA, NA, 16.70, 45....  
## $ saturated\_fat\_100g <dbl> 0.000, NA, NA, 9.900, 5....  
## $ butyric\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ caproic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ caprylic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ capric\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ lauric\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ myristic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ palmitic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ stearic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ arachidic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ behenic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ lignoceric\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ cerotic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ montanic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ melissic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ monounsaturated\_fat\_100g <dbl> NA, NA, NA, 2.9, 9.5, NA...  
## $ polyunsaturated\_fat\_100g <dbl> NA, NA, NA, 3.9, 32.8, N...  
## $ omega\_3\_fat\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ alpha\_linolenic\_acid\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ eicosapentaenoic\_acid\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ docosahexaenoic\_acid\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ omega\_6\_fat\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ linoleic\_acid\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ arachidonic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ gamma\_linolenic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ dihomo\_gamma\_linolenic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ omega\_9\_fat\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ oleic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ elaidic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ gondoic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ mead\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ erucic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ nervonic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ trans\_fat\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ cholesterol\_100g <dbl> NA, NA, NA, 0.00020, NA,...  
## $ carbohydrates\_100g <dbl> 54.00, NA, NA, 5.70, 17....  
## $ sugars\_100g <dbl> 54.00, NA, NA, 4.20, 2.7...  
## $ sucrose\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ glucose\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ fructose\_100g <int> NA, NA, NA, NA, NA, NA, ...  
## $ lactose\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ maltose\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ maltodextrins\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ starch\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ polyols\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ fiber\_100g <dbl> NA, NA, NA, 0.2, 9.0, NA...  
## $ proteins\_100g <dbl> 0.00, NA, NA, 2.90, 18.2...  
## $ casein\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ serum\_proteins\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ nucleotides\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ salt\_100g <dbl> 0.0000000, NA, NA, 0.050...  
## $ sodium\_100g <dbl> 0.00000000, NA, NA, 0.02...  
## $ alcohol\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ vitamin\_a\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ beta\_carotene\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ vitamin\_d\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ vitamin\_e\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ vitamin\_k\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ vitamin\_c\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ vitamin\_b1\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ vitamin\_b2\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ vitamin\_pp\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ vitamin\_b6\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ vitamin\_b9\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ vitamin\_b12\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ biotin\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ pantothenic\_acid\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ silica\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ bicarbonate\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ potassium\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ chloride\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ calcium\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ phosphorus\_100g <dbl> NA, NA, NA, NA, 1.155, N...  
## $ iron\_100g <dbl> NA, NA, NA, NA, 0.00380,...  
## $ magnesium\_100g <dbl> NA, NA, NA, NA, 0.129, N...  
## $ zinc\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ copper\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ manganese\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ fluoride\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ selenium\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ chromium\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ molybdenum\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ iodine\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ caffeine\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ taurine\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ ph\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ fruits\_vegetables\_nuts\_100g <dbl> 54, NA, NA, NA, NA, NA, ...  
## $ collagen\_meat\_protein\_ratio\_100g <int> NA, NA, NA, NA, NA, NA, ...  
## $ cocoa\_100g <int> NA, NA, NA, NA, NA, NA, ...  
## $ chlorophyl\_100g <lgl> NA, NA, NA, NA, NA, NA, ...  
## $ carbon\_footprint\_100g <dbl> NA, NA, NA, NA, NA, NA, ...  
## $ nutrition\_score\_fr\_100g <int> 11, NA, NA, 11, 17, NA, ...  
## $ nutrition\_score\_uk\_100g <int> 11, NA, NA, 11, 17, NA, ...

### Remove duplicate and useless info

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)  
  
food2 <- food[,-duplicates]  
  
useless <- c(1, 2, 3, 32:41)  
food3 <- food2[,-useless]  
glimpse(food3)

## Observations: 1,500  
## Variables: 123  
## $ created\_t <int> 1424747544, 1450316429, 14286...  
## $ last\_modified\_t <int> 1438445887, 1450817956, 14287...  
## $ last\_modified\_datetime <chr> "2015-08-01T16:18:07Z", "2015...  
## $ product\_name <chr> "Confiture de fraise fraise d...  
## $ generic\_name <chr> "", "", "PÃ¢tes de fruits", "...  
## $ packaging <chr> "Bocal,Verre", "Plastic,Box",...  
## $ brands <chr> "Casino DÃ©lices", "Guylian",...  
## $ categories <chr> "Aliments et boissons Ã  base...  
## $ origins <chr> "", "", "", "", "Argentina", ...  
## $ manufacturing\_places <chr> "France", "Belgium", "", "Ita...  
## $ labels <chr> "", "", "", "Vegetariano,Vega...  
## $ emb\_codes <chr> "EMB 78015", "", "", "", "ES ...  
## $ emb\_codes\_tags <chr> "emb-78015", "", "", "", "es-...  
## $ cities <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ cities\_tags <chr> "andresy-yvelines-france", ""...  
## $ purchase\_places <chr> "Lyon,France", "NSW,Australia...  
## $ countries <chr> "France", "Australia", "Franc...  
## $ countries\_en <chr> "France", "Australia", "Franc...  
## $ allergens <chr> "", "", "", "", "", "", "", "...  
## $ traces <chr> "Lait,Fruits Ã  coque", "", "...  
## $ traces\_en <chr> "Milk,Nuts", "", "", "", "Nut...  
## $ serving\_size <chr> "15 g", "", "", "", "", "", "...  
## $ no\_nutriments <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ additives <chr> "[ sucre-de-canne -> fr:sucre...  
## $ additives\_en <chr> "E440 - Pectins", "", "E440 -...  
## $ ingredients\_from\_palm\_oil <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ ingredients\_from\_palm\_oil\_tags <chr> "", "", "", "", "", "", "", "...  
## $ ingredients\_that\_may\_be\_from\_palm\_oil <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ image\_small\_url <chr> "http://en.openfoodfacts.org/...  
## $ energy\_from\_fat\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ fat\_100g <dbl> 0.00, NA, NA, 16.70, 45.50, N...  
## $ saturated\_fat\_100g <dbl> 0.000, NA, NA, 9.900, 5.200, ...  
## $ butyric\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ caproic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ caprylic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ capric\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ lauric\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ myristic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ palmitic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ stearic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ arachidic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ behenic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ lignoceric\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ cerotic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ montanic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ melissic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ monounsaturated\_fat\_100g <dbl> NA, NA, NA, 2.9, 9.5, NA, NA,...  
## $ polyunsaturated\_fat\_100g <dbl> NA, NA, NA, 3.9, 32.8, NA, NA...  
## $ omega\_3\_fat\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ alpha\_linolenic\_acid\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ eicosapentaenoic\_acid\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ docosahexaenoic\_acid\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ omega\_6\_fat\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ linoleic\_acid\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ arachidonic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ gamma\_linolenic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ dihomo\_gamma\_linolenic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ omega\_9\_fat\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ oleic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ elaidic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ gondoic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ mead\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ erucic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ nervonic\_acid\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ trans\_fat\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ cholesterol\_100g <dbl> NA, NA, NA, 0.00020, NA, NA, ...  
## $ carbohydrates\_100g <dbl> 54.00, NA, NA, 5.70, 17.30, N...  
## $ sugars\_100g <dbl> 54.00, NA, NA, 4.20, 2.70, NA...  
## $ sucrose\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ glucose\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ fructose\_100g <int> NA, NA, NA, NA, NA, NA, NA, N...  
## $ lactose\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ maltose\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ maltodextrins\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ starch\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ polyols\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ fiber\_100g <dbl> NA, NA, NA, 0.2, 9.0, NA, NA,...  
## $ proteins\_100g <dbl> 0.00, NA, NA, 2.90, 18.20, NA...  
## $ casein\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ serum\_proteins\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ nucleotides\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ salt\_100g <dbl> 0.0000000, NA, NA, 0.0508000,...  
## $ sodium\_100g <dbl> 0.00000000, NA, NA, 0.0200000...  
## $ alcohol\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ vitamin\_a\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ beta\_carotene\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ vitamin\_d\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ vitamin\_e\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ vitamin\_k\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ vitamin\_c\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ vitamin\_b1\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ vitamin\_b2\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ vitamin\_pp\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ vitamin\_b6\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ vitamin\_b9\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ vitamin\_b12\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ biotin\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ pantothenic\_acid\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ silica\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ bicarbonate\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ potassium\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ chloride\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ calcium\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ phosphorus\_100g <dbl> NA, NA, NA, NA, 1.155, NA, NA...  
## $ iron\_100g <dbl> NA, NA, NA, NA, 0.00380, NA, ...  
## $ magnesium\_100g <dbl> NA, NA, NA, NA, 0.129, NA, NA...  
## $ zinc\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ copper\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ manganese\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ fluoride\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ selenium\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ chromium\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ molybdenum\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ iodine\_100g <dbl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ caffeine\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ taurine\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ ph\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ fruits\_vegetables\_nuts\_100g <dbl> 54, NA, NA, NA, NA, NA, NA, N...  
## $ collagen\_meat\_protein\_ratio\_100g <int> NA, NA, NA, NA, NA, NA, NA, N...  
## $ cocoa\_100g <int> NA, NA, NA, NA, NA, NA, NA, N...  
## $ chlorophyl\_100g <lgl> NA, NA, NA, NA, NA, NA, NA, N...  
## $ nutrition\_score\_fr\_100g <int> 11, NA, NA, 11, 17, NA, NA, 1...  
## $ nutrition\_score\_uk\_100g <int> 11, NA, NA, 11, 17, NA, NA, 1...

### Correct missing data to 0

i <- sapply(food3, is.numeric) # Identify all numeric variables  
food3[i] <- lapply(food3[i], as.character)   
food3[is.na(food3)] <- 0 # Replace NA with 0  
food3[i] <- lapply(food3[i], as.factor)   
glimpse(food3)

## Observations: 1,500  
## Variables: 123  
## $ created\_t <fct> 1424747544, 1450316429, 14286...  
## $ last\_modified\_t <fct> 1438445887, 1450817956, 14287...  
## $ last\_modified\_datetime <chr> "2015-08-01T16:18:07Z", "2015...  
## $ product\_name <chr> "Confiture de fraise fraise d...  
## $ generic\_name <chr> "", "", "PÃ¢tes de fruits", "...  
## $ packaging <chr> "Bocal,Verre", "Plastic,Box",...  
## $ brands <chr> "Casino DÃ©lices", "Guylian",...  
## $ categories <chr> "Aliments et boissons Ã  base...  
## $ origins <chr> "", "", "", "", "Argentina", ...  
## $ manufacturing\_places <chr> "France", "Belgium", "", "Ita...  
## $ labels <chr> "", "", "", "Vegetariano,Vega...  
## $ emb\_codes <chr> "EMB 78015", "", "", "", "ES ...  
## $ emb\_codes\_tags <chr> "emb-78015", "", "", "", "es-...  
## $ cities <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ cities\_tags <chr> "andresy-yvelines-france", ""...  
## $ purchase\_places <chr> "Lyon,France", "NSW,Australia...  
## $ countries <chr> "France", "Australia", "Franc...  
## $ countries\_en <chr> "France", "Australia", "Franc...  
## $ allergens <chr> "", "", "", "", "", "", "", "...  
## $ traces <chr> "Lait,Fruits Ã  coque", "", "...  
## $ traces\_en <chr> "Milk,Nuts", "", "", "", "Nut...  
## $ serving\_size <chr> "15 g", "", "", "", "", "", "...  
## $ no\_nutriments <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ additives <chr> "[ sucre-de-canne -> fr:sucre...  
## $ additives\_en <chr> "E440 - Pectins", "", "E440 -...  
## $ ingredients\_from\_palm\_oil <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ ingredients\_from\_palm\_oil\_tags <chr> "", "", "", "", "", "", "", "...  
## $ ingredients\_that\_may\_be\_from\_palm\_oil <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ image\_small\_url <chr> "http://en.openfoodfacts.org/...  
## $ energy\_from\_fat\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ fat\_100g <fct> 0, 0, 0, 16.7, 45.5, 0, 0, 25...  
## $ saturated\_fat\_100g <fct> 0, 0, 0, 9.9, 5.2, 0, 0, 17, ...  
## $ butyric\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ caproic\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ caprylic\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ capric\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ lauric\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ myristic\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ palmitic\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ stearic\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ arachidic\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ behenic\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ lignoceric\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ cerotic\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ montanic\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ melissic\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ monounsaturated\_fat\_100g <fct> 0, 0, 0, 2.9, 9.5, 0, 0, 0, 0...  
## $ polyunsaturated\_fat\_100g <fct> 0, 0, 0, 3.9, 32.8, 0, 0, 0, ...  
## $ omega\_3\_fat\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ alpha\_linolenic\_acid\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ eicosapentaenoic\_acid\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ docosahexaenoic\_acid\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ omega\_6\_fat\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ linoleic\_acid\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ arachidonic\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ gamma\_linolenic\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ dihomo\_gamma\_linolenic\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ omega\_9\_fat\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ oleic\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ elaidic\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ gondoic\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ mead\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ erucic\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ nervonic\_acid\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ trans\_fat\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ cholesterol\_100g <fct> 0, 0, 0, 2e-04, 0, 0, 0, 0, 0...  
## $ carbohydrates\_100g <fct> 54, 0, 0, 5.7, 17.3, 0, 0, 0....  
## $ sugars\_100g <fct> 54, 0, 0, 4.2, 2.7, 0, 0, 0.5...  
## $ sucrose\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ glucose\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ fructose\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ lactose\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ maltose\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ maltodextrins\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ starch\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ polyols\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ fiber\_100g <fct> 0, 0, 0, 0.2, 9, 0, 0, 0, 0, ...  
## $ proteins\_100g <fct> 0, 0, 0, 2.9, 18.2, 0, 0, 15,...  
## $ casein\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ serum\_proteins\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ nucleotides\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ salt\_100g <fct> 0, 0, 0, 0.0508, 3.9878, 0, 0...  
## $ sodium\_100g <fct> 0, 0, 0, 0.02, 1.57, 0, 0, 0....  
## $ alcohol\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ vitamin\_a\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ beta\_carotene\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ vitamin\_d\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ vitamin\_e\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ vitamin\_k\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ vitamin\_c\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ vitamin\_b1\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ vitamin\_b2\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ vitamin\_pp\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ vitamin\_b6\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ vitamin\_b9\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ vitamin\_b12\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ biotin\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ pantothenic\_acid\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ silica\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ bicarbonate\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ potassium\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ chloride\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ calcium\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ phosphorus\_100g <fct> 0, 0, 0, 0, 1.155, 0, 0, 0, 0...  
## $ iron\_100g <fct> 0, 0, 0, 0, 0.0038, 0, 0, 0, ...  
## $ magnesium\_100g <fct> 0, 0, 0, 0, 0.129, 0, 0, 0, 0...  
## $ zinc\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ copper\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ manganese\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ fluoride\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ selenium\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ chromium\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ molybdenum\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ iodine\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ caffeine\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ taurine\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ ph\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ fruits\_vegetables\_nuts\_100g <fct> 54, 0, 0, 0, 0, 0, 0, 0, 0, 0...  
## $ collagen\_meat\_protein\_ratio\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ cocoa\_100g <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ chlorophyl\_100g <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ nutrition\_score\_fr\_100g <fct> 11, 0, 0, 11, 17, 0, 0, 10, 0...  
## $ nutrition\_score\_uk\_100g <fct> 11, 0, 0, 11, 17, 0, 0, 15, 0...