summary(train)

Item\_Identifier Item\_Weight Item\_Fat\_Content Item\_Visibility Item\_Type

Length:8523 Min. : 4.555 Length:8523 Min. :0.00000 Length:8523

Class :character 1st Qu.: 8.774 Class :character 1st Qu.:0.02699 Class :character

Mode :character Median :12.600 Mode :character Median :0.05393 Mode :character

Mean :12.858 Mean :0.06613

3rd Qu.:16.850 3rd Qu.:0.09459

Max. :21.350 Max. :0.32839

NA's :1463

Item\_MRP Outlet\_Identifier Outlet\_Establishment\_Year Outlet\_Size

Min. : 31.29 Length:8523 Min. :1985 Length:8523

1st Qu.: 93.83 Class :character 1st Qu.:1987 Class :character

Median :143.01 Mode :character Median :1999 Mode :character

Mean :140.99 Mean :1998

3rd Qu.:185.64 3rd Qu.:2004

Max. :266.89 Max. :2009

Outlet\_Location\_Type Outlet\_Type Item\_Outlet\_Sales

Length:8523 Length:8523 Min. : 33.29

Class :character Class :character 1st Qu.: 834.25

Mode :character Mode :character Median : 1794.33

Mean : 2181.29

3rd Qu.: 3101.30

Max. :13086.97

Too many NAs in Item weight. Almost 20%. Ideally this should be dropped from modelling

sapply(train,function(x){sum(ifelse(str\_trim(x)=="",1,0))})

Item\_Identifier Item\_Weight Item\_Fat\_Content

0 NA 0

Item\_Visibility Item\_Type Item\_MRP

0 0 0

Outlet\_Identifier Outlet\_Establishment\_Year Outlet\_Size

0 0 2410

Outlet\_Location\_Type Outlet\_Type Item\_Outlet\_Sales

0 0 0

2410 blanks in outlet size

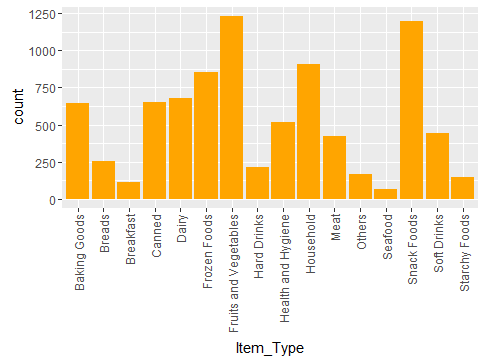
> length(unique(train$Item\_Identifier))

[1] 1559

There are total 1559 products and 8523 observations

|  |
| --- |
| unique(train$Item\_Type)  [1] "Dairy" "Soft Drinks" "Meat" "Fruits and Vegetables"  [5] "Household" "Baking Goods" "Snack Foods" "Frozen Foods"  [9] "Breakfast" "Health and Hygiene" "Hard Drinks" "Canned"  [13] "Breads" "Starchy Foods" "Others" "Seafood" |
|  |
| |  | | --- | | > | |

16 different item types



table(train$Outlet\_Identifier)

OUT010 OUT013 OUT017 OUT018 OUT019 OUT027 OUT035 OUT045 OUT046 OUT049

555 932 926 928 528 935 930 929 930 930

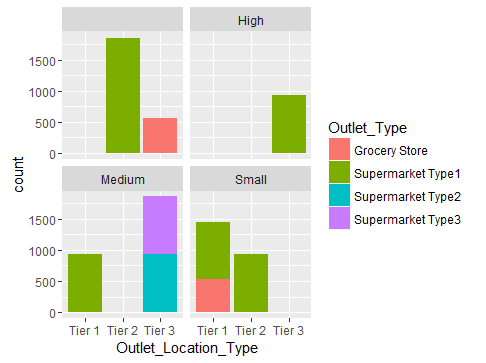
> unique(train$Outlet\_Location\_Type)

[1] "Tier 1" "Tier 3" "Tier 2"

> table(train$Outlet\_Location\_Type)

Tier 1 Tier 2 Tier 3

2388 2785 3350



We will need to impute the missing outlet size before making any conclusions.

table(train$Outlet\_Identifier,train$Outlet\_Size) ##Size not available for outlets 10,17 and 45

High Medium Small

OUT010 555 0 0 0

OUT013 0 932 0 0

OUT017 926 0 0 0

OUT018 0 0 928 0

OUT019 0 0 0 528

OUT027 0 0 935 0

OUT035 0 0 0 930

OUT045 929 0 0 0

OUT046 0 0 0 930

OUT049 0 0 930 0

#Size not available for outlets 10,17 and 45

|  |
| --- |
| > sales <- train %>%  + group\_by(Outlet\_Identifier) %>%  + summarize(total\_outlet\_sales=sum(Item\_Outlet\_Sales))  > sales  Source: local data frame [10 x 2]  Outlet\_Identifier total\_outlet\_sales  (fctr) (dbl)  1 OUT010 188340.2  2 OUT013 2142663.6  3 OUT017 2167465.3  4 OUT018 1851822.8  5 OUT019 179694.1  6 OUT027 3453926.1  7 OUT035 2268122.9  8 OUT045 2036725.5  9 OUT046 2118395.2  10 OUT049 2183969.8 |
|  |
| |  | | --- | | > | |

Cannot identify size based upon total sales

High Medium Small

Tier 1 0 0 930 1458

Tier 2 1855 0 0 930

Tier 3 555 932 1863 0

Of the missing outlet sizes, 1855 are from tier 2, 555 are from tier 3

