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