## Project Summary

In this project, I will use R language to analyze the sales data for Big Mart Outlets. The data set has taken from [Big Mart Sales Prediction](https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/#ProblemStatement). The project goes through 5 phases, Data Exploration, Data Visualization, Data Cleaning, Data manipulation, and Predictive modeling using Machine Learning

### Data Exploration

# Loading Datasets  
train <- read.csv("C:\\data\\train\_v9rqX0R.csv")  
test <- read.csv("C:\\data\\test\_AbJTz2l.csv")

# Check the sizes  
dim(train)

## [1] 8523 12

dim(test)

## [1] 5681 11

# Check the column names and types  
str(train)

## 'data.frame': 8523 obs. of 12 variables:  
## $ Item\_Identifier : chr "FDA15" "DRC01" "FDN15" "FDX07" ...  
## $ Item\_Weight : num 9.3 5.92 17.5 19.2 8.93 ...  
## $ Item\_Fat\_Content : chr "Low Fat" "Regular" "Low Fat" "Regular" ...  
## $ Item\_Visibility : num 0.016 0.0193 0.0168 0 0 ...  
## $ Item\_Type : chr "Dairy" "Soft Drinks" "Meat" "Fruits and Vegetables" ...  
## $ Item\_MRP : num 249.8 48.3 141.6 182.1 53.9 ...  
## $ Outlet\_Identifier : chr "OUT049" "OUT018" "OUT049" "OUT010" ...  
## $ Outlet\_Establishment\_Year: int 1999 2009 1999 1998 1987 2009 1987 1985 2002 2007 ...  
## $ Outlet\_Size : chr "Medium" "Medium" "Medium" "" ...  
## $ Outlet\_Location\_Type : chr "Tier 1" "Tier 3" "Tier 1" "Tier 3" ...  
## $ Outlet\_Type : chr "Supermarket Type1" "Supermarket Type2" "Supermarket Type1" "Grocery Store" ...  
## $ Item\_Outlet\_Sales : num 3735 443 2097 732 995 ...

# Convert the categorical columns from character to factor   
train$Item\_Fat\_Content <- as.factor(train$Item\_Fat\_Content)   
train$Outlet\_Size <- as.factor(train$Outlet\_Size)   
train$Outlet\_Location\_Type <- as.factor(train$Outlet\_Location\_Type)   
train$Outlet\_Type <- as.factor(train$Outlet\_Type)   
  
levels(train$Item\_Fat\_Content)

## [1] "LF" "low fat" "Low Fat" "reg" "Regular"

levels(train$Outlet\_Size)

## [1] "" "High" "Medium" "Small"

levels(train$Outlet\_Location\_Type)

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

levels(train$Outlet\_Type)

## [1] "Grocery Store" "Supermarket Type1" "Supermarket Type2"  
## [4] "Supermarket Type3"

# Check for missing values  
table(is.na(train))

##   
## FALSE TRUE   
## 100813 1463

# Check the variables which has those missing values  
colSums(is.na(train))

## Item\_Identifier Item\_Weight Item\_Fat\_Content   
## 0 1463 0   
## Item\_Visibility Item\_Type Item\_MRP   
## 0 0 0   
## Outlet\_Identifier Outlet\_Establishment\_Year Outlet\_Size   
## 0 0 0   
## Outlet\_Location\_Type Outlet\_Type Item\_Outlet\_Sales   
## 0 0 0

summary(train)

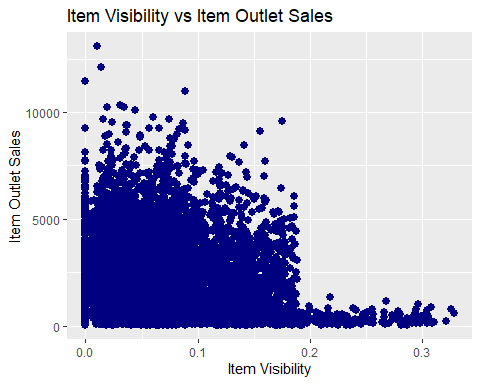
## Item\_Identifier Item\_Weight Item\_Fat\_Content Item\_Visibility   
## Length:8523 Min. : 4.555 LF : 316 Min. :0.00000   
## Class :character 1st Qu.: 8.774 low fat: 112 1st Qu.:0.02699   
## Mode :character Median :12.600 Low Fat:5089 Median :0.05393   
## Mean :12.858 reg : 117 Mean :0.06613   
## 3rd Qu.:16.850 Regular:2889 3rd Qu.:0.09459   
## Max. :21.350 Max. :0.32839   
## NA's :1463   
## Item\_Type Item\_MRP Outlet\_Identifier   
## Length:8523 Min. : 31.29 Length:8523   
## Class :character 1st Qu.: 93.83 Class :character   
## Mode :character Median :143.01 Mode :character   
## Mean :140.99   
## 3rd Qu.:185.64   
## Max. :266.89   
##   
## Outlet\_Establishment\_Year Outlet\_Size Outlet\_Location\_Type  
## Min. :1985 :2410 Tier 1:2388   
## 1st Qu.:1987 High : 932 Tier 2:2785   
## Median :1999 Medium:2793 Tier 3:3350   
## Mean :1998 Small :2388   
## 3rd Qu.:2004   
## Max. :2009   
##   
## Outlet\_Type Item\_Outlet\_Sales   
## Grocery Store :1083 Min. : 33.29   
## Supermarket Type1:5577 1st Qu.: 834.25   
## Supermarket Type2: 928 Median : 1794.33   
## Supermarket Type3: 935 Mean : 2181.29   
## 3rd Qu.: 3101.30   
## Max. :13086.97   
##

Here are some quick inferences drawn from variables in train data set:

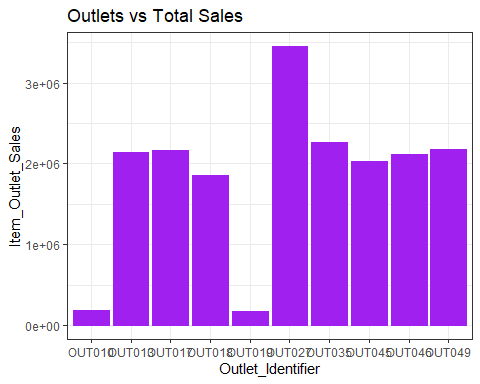
1. Item\_Fat\_Content has mis-matched factor levels.
2. Minimum value of item\_visibility is 0. Practically, this is not possible. If an item occupies shelf space in a grocery store, it ought to have some visibility. We’ll treat all 0’s as missing values.
3. Item\_Weight has 1463 missing values (already explained above).
4. Outlet\_Size has a unmatched factor levels.

### Data Visualization

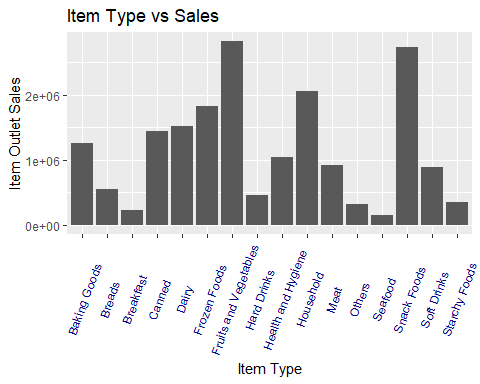
library(ggplot2)  
ggplot(train, aes(x= Item\_Visibility, y = Item\_Outlet\_Sales)) + geom\_point(size = 2.5, color="navy") + xlab("Item Visibility") + ylab("Item Outlet Sales") + ggtitle("Item Visibility vs Item Outlet Sales")

 We can see that majority of sales has been obtained from products having visibility less than 0.2. This suggests that item\_visibility < 2 must be an important factor in determining sales. Let’s plot few more interesting graphs and explore such hidden stories.

ggplot(train, aes(Outlet\_Identifier, Item\_Outlet\_Sales)) + geom\_bar(stat = "identity", color = "purple") +theme(axis.text.x = element\_text(angle = 70, vjust = 0.5, color = "black")) + ggtitle("Outlets vs Total Sales") + theme\_bw()

 Here, we infer that OUT027 has contributed to majority of sales followed by OUT35. OUT10 and OUT19 have probably the least footfall, thereby contributing to the least outlet sales.

ggplot(train, aes(Item\_Type, Item\_Outlet\_Sales)) + geom\_bar( stat = "identity") +theme(axis.text.x = element\_text(angle = 70, vjust = 0.5, color = "navy")) + xlab("Item Type") + ylab("Item Outlet Sales")+ggtitle("Item Type vs Sales")

 From this graph, we can infer that Fruits and Vegetables contribute to the highest amount of outlet sales followed by snack foods and household products.

### Data Cleaning

#Combine the 2 datasets (Train & Test)  
dim(train)

## [1] 8523 12

dim(test)

## [1] 5681 11

test$Item\_Outlet\_Sales <- 1  
combi <- rbind(train, test)

Impute missing value by median. I’m using median because it is known to be highly robust to outliers. Moreover, for this problem, our evaluation metric is RMSE which is also highly affected by outliers. Hence, median is better in this case.

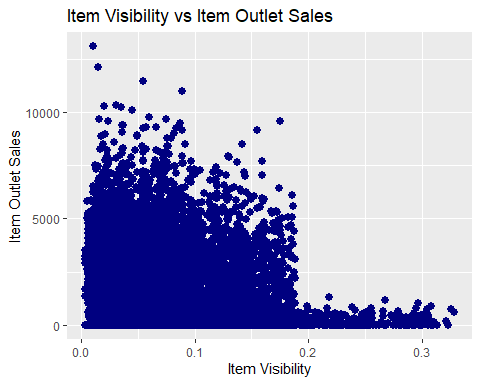
combi$Item\_Weight[is.na(combi$Item\_Weight)] <- median(combi$Item\_Weight, na.rm = TRUE)  
table(is.na(combi$Item\_Weight))

##   
## FALSE   
## 14204

Let’s take up Item\_Visibility. In the graph above, we saw item visibility has zero value also, which is practically not feasible. Hence, we’ll consider it as a missing value and once again make the imputation using median

combi$Item\_Visibility <- ifelse(combi$Item\_Visibility == 0,  
 median(combi$Item\_Visibility), combi$Item\_Visibility)

library(ggplot2)  
ggplot(combi, aes(x= Item\_Visibility, y = Item\_Outlet\_Sales)) + geom\_point(size = 2.5, color="navy") + xlab("Item Visibility") + ylab("Item Outlet Sales") + ggtitle("Item Visibility vs Item Outlet Sales")

 Handling the mismatched levels in variables

levels(combi$Outlet\_Size)[1] <- "Other"  
library(plyr)  
combi$Item\_Fat\_Content <- revalue(combi$Item\_Fat\_Content,  
c("LF" = "Low Fat", "reg" = "Regular"))  
combi$Item\_Fat\_Content <- revalue(combi$Item\_Fat\_Content, c("low fat" = "Low Fat"))  
table(combi$Item\_Fat\_Content)

##   
## Low Fat Regular   
## 9185 5019

levels(combi$Item\_Fat\_Content)

## [1] "Low Fat" "Regular"

levels(combi$Outlet\_Size)

## [1] "Other" "High" "Medium" "Small"

### Data Manipulation

# Count of Outlet Identifiers  
library(dplyr)

##   
## Attaching package: 'dplyr'

## 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

a <- combi%>%  
 group\_by(Outlet\_Identifier)%>%  
 tally()  
head(a)

## # A tibble: 6 x 2  
## Outlet\_Identifier n  
## <chr> <int>  
## 1 OUT010 925  
## 2 OUT013 1553  
## 3 OUT017 1543  
## 4 OUT018 1546  
## 5 OUT019 880  
## 6 OUT027 1559

names(a)[2] <- "Outlet\_Count"  
combi <- full\_join(a, combi, by = "Outlet\_Identifier")

head(combi)

## # A tibble: 6 x 13  
## Outlet\_Identifi~ Outlet\_Count Item\_Identifier Item\_Weight Item\_Fat\_Content  
## <chr> <int> <chr> <dbl> <fct>   
## 1 OUT010 925 FDX07 19.2 Regular   
## 2 OUT010 925 FDE51 5.92 Regular   
## 3 OUT010 925 FDV38 19.2 Low Fat   
## 4 OUT010 925 FDM39 6.42 Low Fat   
## 5 OUT010 925 FDC46 17.7 Low Fat   
## 6 OUT010 925 FDW20 20.8 Low Fat   
## # ... with 8 more variables: Item\_Visibility <dbl>, Item\_Type <chr>,  
## # Item\_MRP <dbl>, Outlet\_Establishment\_Year <int>, Outlet\_Size <fct>,  
## # Outlet\_Location\_Type <fct>, Outlet\_Type <fct>, Item\_Outlet\_Sales <dbl>

# Count of Item Identifiers  
b <- combi%>%  
 group\_by(Item\_Identifier)%>%  
 tally()  
  
names(b)[2] <- "Item\_Count"

head(b)

## # A tibble: 6 x 2  
## Item\_Identifier Item\_Count  
## <chr> <int>  
## 1 DRA12 9  
## 2 DRA24 10  
## 3 DRA59 10  
## 4 DRB01 8  
## 5 DRB13 9  
## 6 DRB24 8

combi <- merge(b, combi, by = "Item\_Identifier")  
head(combi)

## Item\_Identifier Item\_Count Outlet\_Identifier Outlet\_Count Item\_Weight  
## 1 DRA12 9 OUT049 1550 11.6  
## 2 DRA12 9 OUT010 925 11.6  
## 3 DRA12 9 OUT017 1543 11.6  
## 4 DRA12 9 OUT027 1559 12.6  
## 5 DRA12 9 OUT035 1550 11.6  
## 6 DRA12 9 OUT046 1550 11.6  
## Item\_Fat\_Content Item\_Visibility Item\_Type Item\_MRP  
## 1 Low Fat 0.04100956 Soft Drinks 141.0154  
## 2 Low Fat 0.06853504 Soft Drinks 143.0154  
## 3 Low Fat 0.04117751 Soft Drinks 140.3154  
## 4 Low Fat 0.04074762 Soft Drinks 140.0154  
## 5 Low Fat 0.05402054 Soft Drinks 141.9154  
## 6 Low Fat 0.04094590 Soft Drinks 142.9154  
## Outlet\_Establishment\_Year Outlet\_Size Outlet\_Location\_Type Outlet\_Type  
## 1 1999 Medium Tier 1 Supermarket Type1  
## 2 1998 Other Tier 3 Grocery Store  
## 3 2007 Other Tier 2 Supermarket Type1  
## 4 1985 Medium Tier 3 Supermarket Type3  
## 5 2004 Small Tier 2 Supermarket Type1  
## 6 1997 Small Tier 1 Supermarket Type1  
## Item\_Outlet\_Sales  
## 1 1.0000  
## 2 283.6308  
## 3 2552.6772  
## 4 1.0000  
## 5 992.7078  
## 6 1.0000

c <- combi%>%  
 select(Outlet\_Establishment\_Year)%>%   
 mutate(Outlet\_Year = 2013 - combi$Outlet\_Establishment\_Year)  
  
c = c %>% distinct(Outlet\_Establishment\_Year,Outlet\_Year)  
head(c)

## Outlet\_Establishment\_Year Outlet\_Year  
## 1 1999 14  
## 2 1998 15  
## 3 2007 6  
## 4 1985 28  
## 5 2004 9  
## 6 1997 16

combi <- merge(c, combi, by = "Outlet\_Establishment\_Year")  
head(combi)

## Outlet\_Establishment\_Year Outlet\_Year Item\_Identifier Item\_Count  
## 1 1985 28 FDW37 8  
## 2 1985 28 FDA55 9  
## 3 1985 28 FDR33 10  
## 4 1985 28 FDM12 10  
## 5 1985 28 DRF48 9  
## 6 1985 28 DRD12 8  
## Outlet\_Identifier Outlet\_Count Item\_Weight Item\_Fat\_Content Item\_Visibility  
## 1 OUT027 1559 12.6 Low Fat 0.12344967  
## 2 OUT027 1559 12.6 Regular 0.05671306  
## 3 OUT019 880 12.6 Low Fat 0.04690397  
## 4 OUT027 1559 12.6 Regular 0.06957842  
## 5 OUT027 1559 12.6 Low Fat 0.05155062  
## 6 OUT027 1559 12.6 Low Fat 0.05402054  
## Item\_Type Item\_MRP Outlet\_Size Outlet\_Location\_Type  
## 1 Canned 89.7488 Medium Tier 3  
## 2 Fruits and Vegetables 223.8088 Medium Tier 3  
## 3 Snack Foods 110.6570 Small Tier 1  
## 4 Baking Goods 190.2214 Medium Tier 3  
## 5 Soft Drinks 186.3898 Medium Tier 3  
## 6 Soft Drinks 89.4146 Medium Tier 3  
## Outlet\_Type Item\_Outlet\_Sales  
## 1 Supermarket Type3 1901.525  
## 2 Supermarket Type3 5592.720  
## 3 Grocery Store 659.142  
## 4 Supermarket Type3 1.000  
## 5 Supermarket Type3 1.000  
## 6 Supermarket Type3 2645.223

q <- substr(combi$Item\_Identifier,1,2)  
q <- gsub("FD","Food",q)  
q <- gsub("DR","Drinks",q)  
q <- gsub("NC","Non-Consumable",q)  
table(q)

## q  
## Drinks Food Non-Consumable   
## 1317 10201 2686

combi$Item\_Type\_New <- q

# Label Encoding  
combi$Item\_Fat\_Content <- ifelse(combi$Item\_Fat\_Content == "Regular",1,0)

# Convert the type of the new variable to factor  
combi$Item\_Type\_New <- as.factor(combi$Item\_Type\_New)  
str(combi)

## 'data.frame': 14204 obs. of 16 variables:  
## $ Outlet\_Establishment\_Year: int 1985 1985 1985 1985 1985 1985 1985 1985 1985 1985 ...  
## $ Outlet\_Year : num 28 28 28 28 28 28 28 28 28 28 ...  
## $ Item\_Identifier : chr "FDW37" "FDA55" "FDR33" "FDM12" ...  
## $ Item\_Count : int 8 9 10 10 9 8 8 10 9 9 ...  
## $ Outlet\_Identifier : chr "OUT027" "OUT027" "OUT019" "OUT027" ...  
## $ Outlet\_Count : int 1559 1559 880 1559 1559 1559 1559 1559 1559 1559 ...  
## $ Item\_Weight : num 12.6 12.6 12.6 12.6 12.6 12.6 12.6 12.6 12.6 12.6 ...  
## $ Item\_Fat\_Content : num 0 1 0 1 0 0 0 0 0 1 ...  
## $ Item\_Visibility : num 0.1234 0.0567 0.0469 0.0696 0.0516 ...  
## $ Item\_Type : chr "Canned" "Fruits and Vegetables" "Snack Foods" "Baking Goods" ...  
## $ Item\_MRP : num 89.7 223.8 110.7 190.2 186.4 ...  
## $ Outlet\_Size : Factor w/ 4 levels "Other","High",..: 3 3 4 3 3 3 3 3 3 3 ...  
## $ Outlet\_Location\_Type : Factor w/ 3 levels "Tier 1","Tier 2",..: 3 3 1 3 3 3 3 3 3 3 ...  
## $ Outlet\_Type : Factor w/ 4 levels "Grocery Store",..: 4 4 1 4 4 4 4 4 4 4 ...  
## $ Item\_Outlet\_Sales : num 1902 5593 659 1 1 ...  
## $ Item\_Type\_New : Factor w/ 3 levels "Drinks","Food",..: 2 2 2 2 1 1 2 2 2 2 ...

# One Hot Encoding  
library(dummies)

## dummies-1.5.6 provided by Decision Patterns

combi <- dummy.data.frame(combi, names = c('Outlet\_Size','Outlet\_Location\_Type','Outlet\_Type', 'Item\_Type\_New'), sep='\_')

## Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts = FALSE):  
## non-list contrasts argument ignored  
  
## Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts = FALSE):  
## non-list contrasts argument ignored  
  
## Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts = FALSE):  
## non-list contrasts argument ignored  
  
## Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts = FALSE):  
## non-list contrasts argument ignored

str(combi)

## 'data.frame': 14204 obs. of 26 variables:  
## $ Outlet\_Establishment\_Year : int 1985 1985 1985 1985 1985 1985 1985 1985 1985 1985 ...  
## $ Outlet\_Year : num 28 28 28 28 28 28 28 28 28 28 ...  
## $ Item\_Identifier : chr "FDW37" "FDA55" "FDR33" "FDM12" ...  
## $ Item\_Count : int 8 9 10 10 9 8 8 10 9 9 ...  
## $ Outlet\_Identifier : chr "OUT027" "OUT027" "OUT019" "OUT027" ...  
## $ Outlet\_Count : int 1559 1559 880 1559 1559 1559 1559 1559 1559 1559 ...  
## $ Item\_Weight : num 12.6 12.6 12.6 12.6 12.6 12.6 12.6 12.6 12.6 12.6 ...  
## $ Item\_Fat\_Content : num 0 1 0 1 0 0 0 0 0 1 ...  
## $ Item\_Visibility : num 0.1234 0.0567 0.0469 0.0696 0.0516 ...  
## $ Item\_Type : chr "Canned" "Fruits and Vegetables" "Snack Foods" "Baking Goods" ...  
## $ Item\_MRP : num 89.7 223.8 110.7 190.2 186.4 ...  
## $ Outlet\_Size\_Other : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Outlet\_Size\_High : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Outlet\_Size\_Medium : int 1 1 0 1 1 1 1 1 1 1 ...  
## $ Outlet\_Size\_Small : int 0 0 1 0 0 0 0 0 0 0 ...  
## $ Outlet\_Location\_Type\_Tier 1 : int 0 0 1 0 0 0 0 0 0 0 ...  
## $ Outlet\_Location\_Type\_Tier 2 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Outlet\_Location\_Type\_Tier 3 : int 1 1 0 1 1 1 1 1 1 1 ...  
## $ Outlet\_Type\_Grocery Store : int 0 0 1 0 0 0 0 0 0 0 ...  
## $ Outlet\_Type\_Supermarket Type1: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Outlet\_Type\_Supermarket Type2: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Outlet\_Type\_Supermarket Type3: int 1 1 0 1 1 1 1 1 1 1 ...  
## $ Item\_Outlet\_Sales : num 1902 5593 659 1 1 ...  
## $ Item\_Type\_New\_Drinks : int 0 0 0 0 1 1 0 0 0 0 ...  
## $ Item\_Type\_New\_Food : int 1 1 1 1 0 0 1 1 1 1 ...  
## $ Item\_Type\_New\_Non-Consumable : int 0 0 0 0 0 0 0 0 0 0 ...  
## - attr(\*, "dummies")=List of 4  
## ..$ Outlet\_Size : int [1:4] 12 13 14 15  
## ..$ Outlet\_Location\_Type: int [1:3] 16 17 18  
## ..$ Outlet\_Type : int [1:4] 19 20 21 22  
## ..$ Item\_Type\_New : int [1:3] 24 25 26

### Predictive Modeling using Machine Learning

# drop the columns which have either been converted using other variables or are identifier variables  
combi <- select(combi, -c(Item\_Identifier, Outlet\_Identifier, Item\_Fat\_Content, Outlet\_Establishment\_Year,Item\_Type))  
str(combi)

## 'data.frame': 14204 obs. of 21 variables:  
## $ Outlet\_Year : num 28 28 28 28 28 28 28 28 28 28 ...  
## $ Item\_Count : int 8 9 10 10 9 8 8 10 9 9 ...  
## $ Outlet\_Count : int 1559 1559 880 1559 1559 1559 1559 1559 1559 1559 ...  
## $ Item\_Weight : num 12.6 12.6 12.6 12.6 12.6 12.6 12.6 12.6 12.6 12.6 ...  
## $ Item\_Visibility : num 0.1234 0.0567 0.0469 0.0696 0.0516 ...  
## $ Item\_MRP : num 89.7 223.8 110.7 190.2 186.4 ...  
## $ Outlet\_Size\_Other : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Outlet\_Size\_High : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Outlet\_Size\_Medium : int 1 1 0 1 1 1 1 1 1 1 ...  
## $ Outlet\_Size\_Small : int 0 0 1 0 0 0 0 0 0 0 ...  
## $ Outlet\_Location\_Type\_Tier 1 : int 0 0 1 0 0 0 0 0 0 0 ...  
## $ Outlet\_Location\_Type\_Tier 2 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Outlet\_Location\_Type\_Tier 3 : int 1 1 0 1 1 1 1 1 1 1 ...  
## $ Outlet\_Type\_Grocery Store : int 0 0 1 0 0 0 0 0 0 0 ...  
## $ Outlet\_Type\_Supermarket Type1: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Outlet\_Type\_Supermarket Type2: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Outlet\_Type\_Supermarket Type3: int 1 1 0 1 1 1 1 1 1 1 ...  
## $ Item\_Outlet\_Sales : num 1902 5593 659 1 1 ...  
## $ Item\_Type\_New\_Drinks : int 0 0 0 0 1 1 0 0 0 0 ...  
## $ Item\_Type\_New\_Food : int 1 1 1 1 0 0 1 1 1 1 ...  
## $ Item\_Type\_New\_Non-Consumable : int 0 0 0 0 0 0 0 0 0 0 ...

# devide the dataset into train and test  
new\_train <- combi[1:nrow(train),]  
new\_test <- combi[-(1:nrow(train)),]

# Build our Linear Regression model  
linear\_model <- lm(Item\_Outlet\_Sales ~ ., data = new\_train)  
summary(linear\_model)

##   
## Call:  
## lm(formula = Item\_Outlet\_Sales ~ ., data = new\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3378.2 -1082.9 -37.3 742.2 10050.3   
##   
## Coefficients: (8 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2145.9881 827.4669 -2.593 0.009518 \*\*   
## Outlet\_Year 0.4943 14.6869 0.034 0.973151   
## Item\_Count 28.3996 24.1163 1.178 0.238986   
## Outlet\_Count 1.7458 0.2229 7.831 5.41e-15 \*\*\*  
## Item\_Weight 2.2582 4.2823 0.527 0.597971   
## Item\_Visibility -227.4356 340.4925 -0.668 0.504177   
## Item\_MRP 9.0449 0.2711 33.359 < 2e-16 \*\*\*  
## Outlet\_Size\_Other -862.1740 348.8665 -2.471 0.013480 \*   
## Outlet\_Size\_High -785.9564 106.8290 -7.357 2.05e-13 \*\*\*  
## Outlet\_Size\_Medium 39.9653 69.0525 0.579 0.562762   
## Outlet\_Size\_Small NA NA NA NA   
## `Outlet\_Location\_Type\_Tier 1` -786.0833 214.7842 -3.660 0.000254 \*\*\*  
## `Outlet\_Location\_Type\_Tier 2` NA NA NA NA   
## `Outlet\_Location\_Type\_Tier 3` NA NA NA NA   
## `Outlet\_Type\_Grocery Store` NA NA NA NA   
## `Outlet\_Type\_Supermarket Type1` NA NA NA NA   
## `Outlet\_Type\_Supermarket Type2` NA NA NA NA   
## `Outlet\_Type\_Supermarket Type3` NA NA NA NA   
## Item\_Type\_New\_Drinks -26.2149 67.2210 -0.390 0.696560   
## Item\_Type\_New\_Food 56.5611 43.7396 1.293 0.196000   
## `Item\_Type\_New\_Non-Consumable` NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1550 on 8510 degrees of freedom  
## Multiple R-squared: 0.233, Adjusted R-squared: 0.2319   
## F-statistic: 215.5 on 12 and 8510 DF, p-value: < 2.2e-16

Adjusted R² measures the goodness of fit of a regression model. Higher the R², better is the model. Our R² = 0.2319. It means we really did something drastically wrong.

In our case, I could find our new variables aren’t helping much i.e. Item count, Outlet Count and Item\_Type\_New. Neither of these variables are significant. Significant variables are denoted by ‘\*’ sign.

As we know, correlated predictor variables brings down the model accuracy. Let’s find out the amount of correlation present in our predictor variables. This can be simply calculated using:

cor(new\_train)

## Warning in cor(new\_train): the standard deviation is zero

## Outlet\_Year Item\_Count Outlet\_Count  
## Outlet\_Year 1.000000e+00 0.018579412 -0.0764674894  
## Item\_Count 1.857941e-02 1.000000000 -0.2409393610  
## Outlet\_Count -7.646749e-02 -0.240939361 1.0000000000  
## Item\_Weight -1.142695e-02 -0.014645842 0.0079478331  
## Item\_Visibility 3.633614e-02 0.099253984 -0.3431642520  
## Item\_MRP 7.871695e-05 0.002860730 0.0003212289  
## Outlet\_Size\_Other -4.805563e-01 0.109355994 -0.4562232941  
## Outlet\_Size\_High 4.033892e-01 -0.059700160 0.2451236809  
## Outlet\_Size\_Medium 6.263751e-02 -0.095618832 0.3978369126  
## Outlet\_Size\_Small -1.383795e-02 0.062459897 -0.2560607514  
## Outlet\_Location\_Type\_Tier 1 -3.790734e-01 0.012168590 -0.0465361119  
## Outlet\_Location\_Type\_Tier 2 -3.694618e-01 -0.029470703 0.1257332242  
## Outlet\_Location\_Type\_Tier 3 5.536320e-01 0.001789128 -0.0130085334  
## Outlet\_Type\_Grocery Store 6.837141e-02 0.240665631 -0.9991606831  
## Outlet\_Type\_Supermarket Type1 -4.929801e-01 -0.152272216 0.6323974722  
## Outlet\_Type\_Supermarket Type2 NA NA NA  
## Outlet\_Type\_Supermarket Type3 5.510294e-01 -0.061821874 0.2563903066  
## Item\_Outlet\_Sales 8.489250e-02 -0.064512029 0.3174451568  
## Item\_Type\_New\_Drinks -1.374789e-02 -0.018903693 0.0075252718  
## Item\_Type\_New\_Food 2.581601e-03 0.019908905 -0.0006507127  
## Item\_Type\_New\_Non-Consumable 7.327272e-03 -0.008785802 -0.0048897683  
## Item\_Weight Item\_Visibility Item\_MRP  
## Outlet\_Year -0.011426953 0.036336140 7.871695e-05  
## Item\_Count -0.014645842 0.099253984 2.860730e-03  
## Outlet\_Count 0.007947833 -0.343164252 3.212289e-04  
## Item\_Weight 1.000000000 -0.015188496 3.158494e-02  
## Item\_Visibility -0.015188496 1.000000000 -4.023071e-03  
## Item\_MRP 0.031584938 -0.004023071 1.000000e+00  
## Outlet\_Size\_Other -0.005495133 0.150960369 -1.912157e-04  
## Outlet\_Size\_High 0.007884721 -0.082522485 1.042092e-03  
## Outlet\_Size\_Medium -0.004102252 -0.137696357 1.173405e-03  
## Outlet\_Size\_Small 0.002181739 0.092393204 -1.983833e-03  
## Outlet\_Location\_Type\_Tier 1 0.009441659 0.014712664 -7.860293e-04  
## Outlet\_Location\_Type\_Tier 2 -0.008736642 -0.041676402 -1.548301e-03  
## Outlet\_Location\_Type\_Tier 3 -0.005299221 0.005023847 1.518190e-03  
## Outlet\_Type\_Grocery Store -0.007849184 0.342548444 -2.541984e-04  
## Outlet\_Type\_Supermarket Type1 0.018223058 -0.221046994 8.694544e-05  
## Outlet\_Type\_Supermarket Type2 NA NA NA  
## Outlet\_Type\_Supermarket Type3 -0.014744574 -0.082541353 1.587179e-04  
## Item\_Outlet\_Sales 0.014507268 -0.113986271 3.179097e-01  
## Item\_Type\_New\_Drinks -0.039792683 -0.008989388 -3.984833e-02  
## Item\_Type\_New\_Food -0.014407529 0.068395337 2.043851e-02  
## Item\_Type\_New\_Non-Consumable 0.046434498 -0.072117913 6.301057e-03  
## Outlet\_Size\_Other Outlet\_Size\_High  
## Outlet\_Year -0.4805563499 0.4033892392  
## Item\_Count 0.1093559938 -0.0597001603  
## Outlet\_Count -0.4562232941 0.2451236809  
## Item\_Weight -0.0054951331 0.0078847210  
## Item\_Visibility 0.1509603686 -0.0825224848  
## Item\_MRP -0.0001912157 0.0010420916  
## Outlet\_Size\_Other 1.0000000000 -0.2120336514  
## Outlet\_Size\_High -0.2120336514 1.0000000000  
## Outlet\_Size\_Medium -0.3403977562 -0.3577013458  
## Outlet\_Size\_Small -0.2836762656 -0.2980965066  
## Outlet\_Location\_Type\_Tier 1 -0.4204415428 -0.4418140335  
## Outlet\_Location\_Type\_Tier 2 0.5592858905 -0.1185874295  
## Outlet\_Location\_Type\_Tier 3 0.1554053236 0.4975876569  
## Outlet\_Type\_Grocery Store 0.4778375488 -0.2446740677  
## Outlet\_Type\_Supermarket Type1 -0.2313172485 0.3811663632  
## Outlet\_Type\_Supermarket Type2 NA NA  
## Outlet\_Type\_Supermarket Type3 -0.2125343493 -0.2233382018  
## Item\_Outlet\_Sales -0.1767836330 0.0264245918  
## Item\_Type\_New\_Drinks 0.0191501378 -0.0017474057  
## Item\_Type\_New\_Food -0.0032845865 0.0004561897  
## Item\_Type\_New\_Non-Consumable -0.0105656247 0.0007836792  
## Outlet\_Size\_Medium Outlet\_Size\_Small  
## Outlet\_Year 0.0626375075 -0.013837950  
## Item\_Count -0.0956188316 0.062459897  
## Outlet\_Count 0.3978369126 -0.256060751  
## Item\_Weight -0.0041022523 0.002181739  
## Item\_Visibility -0.1376963571 0.092393204  
## Item\_MRP 0.0011734054 -0.001983833  
## Outlet\_Size\_Other -0.3403977562 -0.283676266  
## Outlet\_Size\_High -0.3577013458 -0.298096507  
## Outlet\_Size\_Medium 1.0000000000 -0.478562631  
## Outlet\_Size\_Small -0.4785626307 1.000000000  
## Outlet\_Location\_Type\_Tier 1 0.0479683577 0.674710362  
## Outlet\_Location\_Type\_Tier 2 -0.1903796622 -0.158656133  
## Outlet\_Location\_Type\_Tier 3 0.0421735792 -0.599083403  
## Outlet\_Type\_Grocery Store -0.3927985163 0.232409476  
## Outlet\_Type\_Supermarket Type1 -0.1655021840 0.042064022  
## Outlet\_Type\_Supermarket Type2 NA NA  
## Outlet\_Type\_Supermarket Type3 0.6243705943 -0.298800434  
## Item\_Outlet\_Sales 0.2282643448 -0.119614419  
## Item\_Type\_New\_Drinks -0.0047055981 -0.009343336  
## Item\_Type\_New\_Food 0.0001061572 0.002216111  
## Item\_Type\_New\_Non-Consumable 0.0034043318 0.004447561  
## Outlet\_Location\_Type\_Tier 1  
## Outlet\_Year -0.3790734262  
## Item\_Count 0.0121685897  
## Outlet\_Count -0.0465361119  
## Item\_Weight 0.0094416593  
## Item\_Visibility 0.0147126637  
## Item\_MRP -0.0007860293  
## Outlet\_Size\_Other -0.4204415428  
## Outlet\_Size\_High -0.4418140335  
## Outlet\_Size\_Medium 0.0479683577  
## Outlet\_Size\_Small 0.6747103622  
## Outlet\_Location\_Type\_Tier 1 1.0000000000  
## Outlet\_Location\_Type\_Tier 2 -0.2351470227  
## Outlet\_Location\_Type\_Tier 3 -0.8879119637  
## Outlet\_Type\_Grocery Store 0.0213640017  
## Outlet\_Type\_Supermarket Type1 0.3324141807  
## Outlet\_Type\_Supermarket Type2 NA  
## Outlet\_Type\_Supermarket Type3 -0.4428573368  
## Item\_Outlet\_Sales -0.0817912407  
## Item\_Type\_New\_Drinks -0.0111873246  
## Item\_Type\_New\_Food 0.0024391251  
## Item\_Type\_New\_Non-Consumable 0.0055724627  
## Outlet\_Location\_Type\_Tier 2  
## Outlet\_Year -3.694618e-01  
## Item\_Count -2.947070e-02  
## Outlet\_Count 1.257332e-01  
## Item\_Weight -8.736642e-03  
## Item\_Visibility -4.167640e-02  
## Item\_MRP -1.548301e-03  
## Outlet\_Size\_Other 5.592859e-01  
## Outlet\_Size\_High -1.185874e-01  
## Outlet\_Size\_Medium -1.903797e-01  
## Outlet\_Size\_Small -1.586561e-01  
## Outlet\_Location\_Type\_Tier 1 -2.351470e-01  
## Outlet\_Location\_Type\_Tier 2 1.000000e+00  
## Outlet\_Location\_Type\_Tier 3 -2.383247e-01  
## Outlet\_Type\_Grocery Store -1.302231e-01  
## Outlet\_Type\_Supermarket Type1 2.028685e-01  
## Outlet\_Type\_Supermarket Type2 NA  
## Outlet\_Type\_Supermarket Type3 -1.188675e-01  
## Item\_Outlet\_Sales 8.274422e-05  
## Item\_Type\_New\_Drinks 2.921426e-02  
## Item\_Type\_New\_Food -4.372725e-04  
## Item\_Type\_New\_Non-Consumable -2.139119e-02  
## Outlet\_Location\_Type\_Tier 3  
## Outlet\_Year 0.553631965  
## Item\_Count 0.001789128  
## Outlet\_Count -0.013008533  
## Item\_Weight -0.005299221  
## Item\_Visibility 0.005023847  
## Item\_MRP 0.001518190  
## Outlet\_Size\_Other 0.155405324  
## Outlet\_Size\_High 0.497587657  
## Outlet\_Size\_Medium 0.042173579  
## Outlet\_Size\_Small -0.599083403  
## Outlet\_Location\_Type\_Tier 1 -0.887911964  
## Outlet\_Location\_Type\_Tier 2 -0.238324701  
## Outlet\_Location\_Type\_Tier 3 1.000000000  
## Outlet\_Type\_Grocery Store 0.040285564  
## Outlet\_Type\_Supermarket Type1 -0.428163854  
## Outlet\_Type\_Supermarket Type2 NA  
## Outlet\_Type\_Supermarket Type3 0.498762664  
## Item\_Outlet\_Sales 0.081686923  
## Item\_Type\_New\_Drinks -0.002648243  
## Item\_Type\_New\_Food -0.002230228  
## Item\_Type\_New\_Non-Consumable 0.004556092  
## Outlet\_Type\_Grocery Store  
## Outlet\_Year 0.0683714126  
## Item\_Count 0.2406656314  
## Outlet\_Count -0.9991606831  
## Item\_Weight -0.0078491836  
## Item\_Visibility 0.3425484440  
## Item\_MRP -0.0002541984  
## Outlet\_Size\_Other 0.4778375488  
## Outlet\_Size\_High -0.2446740677  
## Outlet\_Size\_Medium -0.3927985163  
## Outlet\_Size\_Small 0.2324094758  
## Outlet\_Location\_Type\_Tier 1 0.0213640017  
## Outlet\_Location\_Type\_Tier 2 -0.1302230671  
## Outlet\_Location\_Type\_Tier 3 0.0402855636  
## Outlet\_Type\_Grocery Store 1.0000000000  
## Outlet\_Type\_Supermarket Type1 -0.6419088654  
## Outlet\_Type\_Supermarket Type2 NA  
## Outlet\_Type\_Supermarket Type3 -0.2452518430  
## Item\_Outlet\_Sales -0.3155994416  
## Item\_Type\_New\_Drinks -0.0073476481  
## Item\_Type\_New\_Food 0.0004438185  
## Item\_Type\_New\_Non-Consumable 0.0049951790  
## Outlet\_Type\_Supermarket Type1  
## Outlet\_Year -4.929801e-01  
## Item\_Count -1.522722e-01  
## Outlet\_Count 6.323975e-01  
## Item\_Weight 1.822306e-02  
## Item\_Visibility -2.210470e-01  
## Item\_MRP 8.694544e-05  
## Outlet\_Size\_Other -2.313172e-01  
## Outlet\_Size\_High 3.811664e-01  
## Outlet\_Size\_Medium -1.655022e-01  
## Outlet\_Size\_Small 4.206402e-02  
## Outlet\_Location\_Type\_Tier 1 3.324142e-01  
## Outlet\_Location\_Type\_Tier 2 2.028685e-01  
## Outlet\_Location\_Type\_Tier 3 -4.281639e-01  
## Outlet\_Type\_Grocery Store -6.419089e-01  
## Outlet\_Type\_Supermarket Type1 1.000000e+00  
## Outlet\_Type\_Supermarket Type2 NA  
## Outlet\_Type\_Supermarket Type3 -5.859337e-01  
## Item\_Outlet\_Sales 6.600228e-02  
## Item\_Type\_New\_Drinks 7.986969e-03  
## Item\_Type\_New\_Food -3.280370e-05  
## Item\_Type\_New\_Non-Consumable -5.948206e-03  
## Outlet\_Type\_Supermarket Type2  
## Outlet\_Year NA  
## Item\_Count NA  
## Outlet\_Count NA  
## Item\_Weight NA  
## Item\_Visibility NA  
## Item\_MRP NA  
## Outlet\_Size\_Other NA  
## Outlet\_Size\_High NA  
## Outlet\_Size\_Medium NA  
## Outlet\_Size\_Small NA  
## Outlet\_Location\_Type\_Tier 1 NA  
## Outlet\_Location\_Type\_Tier 2 NA  
## Outlet\_Location\_Type\_Tier 3 NA  
## Outlet\_Type\_Grocery Store NA  
## Outlet\_Type\_Supermarket Type1 NA  
## Outlet\_Type\_Supermarket Type2 1  
## Outlet\_Type\_Supermarket Type3 NA  
## Item\_Outlet\_Sales NA  
## Item\_Type\_New\_Drinks NA  
## Item\_Type\_New\_Food NA  
## Item\_Type\_New\_Non-Consumable NA  
## Outlet\_Type\_Supermarket Type3 Item\_Outlet\_Sales  
## Outlet\_Year 0.5510294093 8.489250e-02  
## Item\_Count -0.0618218737 -6.451203e-02  
## Outlet\_Count 0.2563903066 3.174452e-01  
## Item\_Weight -0.0147445739 1.450727e-02  
## Item\_Visibility -0.0825413526 -1.139863e-01  
## Item\_MRP 0.0001587179 3.179097e-01  
## Outlet\_Size\_Other -0.2125343493 -1.767836e-01  
## Outlet\_Size\_High -0.2233382018 2.642459e-02  
## Outlet\_Size\_Medium 0.6243705943 2.282643e-01  
## Outlet\_Size\_Small -0.2988004341 -1.196144e-01  
## Outlet\_Location\_Type\_Tier 1 -0.4428573368 -8.179124e-02  
## Outlet\_Location\_Type\_Tier 2 -0.1188674628 8.274422e-05  
## Outlet\_Location\_Type\_Tier 3 0.4987626644 8.168692e-02  
## Outlet\_Type\_Grocery Store -0.2452518430 -3.155994e-01  
## Outlet\_Type\_Supermarket Type1 -0.5859336589 6.600228e-02  
## Outlet\_Type\_Supermarket Type2 NA NA  
## Outlet\_Type\_Supermarket Type3 1.0000000000 2.500875e-01  
## Item\_Outlet\_Sales 0.2500874995 1.000000e+00  
## Item\_Type\_New\_Drinks -0.0023328818 -2.355702e-02  
## Item\_Type\_New\_Food -0.0004275673 2.261246e-02  
## Item\_Type\_New\_Non-Consumable 0.0022413908 -8.415274e-03  
## Item\_Type\_New\_Drinks Item\_Type\_New\_Food  
## Outlet\_Year -0.013747887 0.0025816013  
## Item\_Count -0.018903693 0.0199089054  
## Outlet\_Count 0.007525272 -0.0006507127  
## Item\_Weight -0.039792683 -0.0144075292  
## Item\_Visibility -0.008989388 0.0683953368  
## Item\_MRP -0.039848327 0.0204385129  
## Outlet\_Size\_Other 0.019150138 -0.0032845865  
## Outlet\_Size\_High -0.001747406 0.0004561897  
## Outlet\_Size\_Medium -0.004705598 0.0001061572  
## Outlet\_Size\_Small -0.009343336 0.0022161109  
## Outlet\_Location\_Type\_Tier 1 -0.011187325 0.0024391251  
## Outlet\_Location\_Type\_Tier 2 0.029214265 -0.0004372725  
## Outlet\_Location\_Type\_Tier 3 -0.002648243 -0.0022302278  
## Outlet\_Type\_Grocery Store -0.007347648 0.0004438185  
## Outlet\_Type\_Supermarket Type1 0.007986969 -0.0000328037  
## Outlet\_Type\_Supermarket Type2 NA NA  
## Outlet\_Type\_Supermarket Type3 -0.002332882 -0.0004275673  
## Item\_Outlet\_Sales -0.023557016 0.0226124609  
## Item\_Type\_New\_Drinks 1.000000000 -0.5155493472  
## Item\_Type\_New\_Food -0.515549347 1.0000000000  
## Item\_Type\_New\_Non-Consumable -0.155080348 -0.7665419068  
## Item\_Type\_New\_Non-Consumable  
## Outlet\_Year 0.0073272722  
## Item\_Count -0.0087858016  
## Outlet\_Count -0.0048897683  
## Item\_Weight 0.0464344978  
## Item\_Visibility -0.0721179127  
## Item\_MRP 0.0063010573  
## Outlet\_Size\_Other -0.0105656247  
## Outlet\_Size\_High 0.0007836792  
## Outlet\_Size\_Medium 0.0034043318  
## Outlet\_Size\_Small 0.0044475613  
## Outlet\_Location\_Type\_Tier 1 0.0055724627  
## Outlet\_Location\_Type\_Tier 2 -0.0213911887  
## Outlet\_Location\_Type\_Tier 3 0.0045560924  
## Outlet\_Type\_Grocery Store 0.0049951790  
## Outlet\_Type\_Supermarket Type1 -0.0059482059  
## Outlet\_Type\_Supermarket Type2 NA  
## Outlet\_Type\_Supermarket Type3 0.0022413908  
## Item\_Outlet\_Sales -0.0084152744  
## Item\_Type\_New\_Drinks -0.1550803482  
## Item\_Type\_New\_Food -0.7665419068  
## Item\_Type\_New\_Non-Consumable 1.0000000000

cor(new\_train$Outlet\_Count, new\_train$`Outlet\_Type\_Grocery Store`)

## [1] -0.9991607

Let’s try to create a more robust regression mode

#load data  
train <- read.csv("C:\\data\\train\_v9rqX0R.csv")  
test <- read.csv("C:\\data\\test\_AbJTz2l.csv")

#create a new variable in test file  
test$Item\_Outlet\_Sales <- 1  
  
#combine train and test data  
combi <- rbind(train, test)

#impute missing value in Item\_Weight  
combi$Item\_Weight[is.na(combi$Item\_Weight)] <- median(combi$Item\_Weight, na.rm = TRUE)

#impute 0 in item\_visibility  
combi$Item\_Visibility <- ifelse(combi$Item\_Visibility == 0, median(combi$Item\_Visibility), combi$Item\_Visibility)

combi$Item\_Fat\_Content <- as.factor(combi$Item\_Fat\_Content)   
combi$Outlet\_Size <- as.factor(combi$Outlet\_Size)   
combi$Outlet\_Location\_Type <- as.factor(combi$Outlet\_Location\_Type)   
combi$Outlet\_Type <- as.factor(combi$Outlet\_Type)   
  
#rename level in Outlet\_Size  
levels(combi$Outlet\_Size)[1] <- "Other"  
  
#rename levels of Item\_Fat\_Content  
library(plyr)  
combi$Item\_Fat\_Content <- revalue(combi$Item\_Fat\_Content,c("LF" = "Low Fat", "reg" = "Regular"))  
combi$Item\_Fat\_Content <- revalue(combi$Item\_Fat\_Content, c("low fat" = "Low Fat"))

levels(combi$Item\_Fat\_Content)

## [1] "Low Fat" "Regular"

levels(combi$Outlet\_Size)

## [1] "Other" "High" "Medium" "Small"

#create a new column 2013 - Year  
combi$Year <- 2013 - combi$Outlet\_Establishment\_Year

#drop variables not required in modeling  
library(dplyr)  
combi <- select(combi, -c(Item\_Identifier, Outlet\_Identifier, Outlet\_Establishment\_Year))

#divide data set  
new\_train <- combi[1:nrow(train),]  
new\_test <- combi[-(1:nrow(train)),]

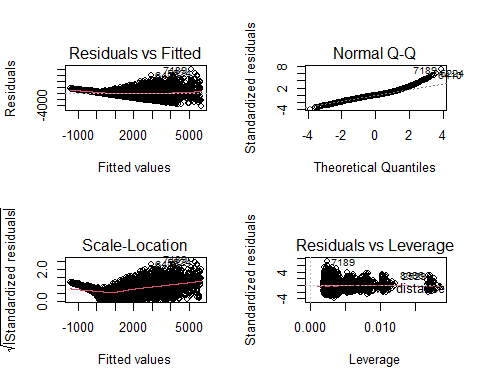
#linear regression  
linear\_model <- lm(Item\_Outlet\_Sales ~ ., data = new\_train)  
summary(linear\_model)

##   
## Call:  
## lm(formula = Item\_Outlet\_Sales ~ ., data = new\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4336.2 -680.8 -89.8 568.3 7946.1   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -977.43100 304.34139 -3.212 0.001325 \*\*   
## Item\_Weight -0.51139 2.91490 -0.175 0.860738   
## Item\_Fat\_ContentRegular 40.50509 28.23186 1.435 0.151401   
## Item\_Visibility -255.77401 263.78199 -0.970 0.332253   
## Item\_TypeBreads 5.25222 84.04714 0.062 0.950173   
## Item\_TypeBreakfast 6.26997 116.59546 0.054 0.957115   
## Item\_TypeCanned 25.28447 62.76864 0.403 0.687091   
## Item\_TypeDairy -41.05177 62.22667 -0.660 0.509456   
## Item\_TypeFrozen Foods -28.00838 58.86986 -0.476 0.634252   
## Item\_TypeFruits and Vegetables 29.84893 54.95872 0.543 0.587065   
## Item\_TypeHard Drinks -0.07955 90.18745 -0.001 0.999296   
## Item\_TypeHealth and Hygiene -10.09434 68.02370 -0.148 0.882035   
## Item\_TypeHousehold -39.12266 59.92886 -0.653 0.513891   
## Item\_TypeMeat -0.35463 70.66270 -0.005 0.995996   
## Item\_TypeOthers -21.27266 98.62482 -0.216 0.829232   
## Item\_TypeSeafood 184.55373 148.00848 1.247 0.212464   
## Item\_TypeSnack Foods -11.48660 55.25066 -0.208 0.835312   
## Item\_TypeSoft Drinks -27.46544 70.16973 -0.391 0.695501   
## Item\_TypeStarchy Foods 21.86315 103.04028 0.212 0.831971   
## Item\_MRP 15.56477 0.19769 78.733 < 2e-16 \*\*\*  
## Outlet\_SizeHigh 849.13529 256.29910 3.313 0.000927 \*\*\*  
## Outlet\_SizeMedium 170.93952 71.08120 2.405 0.016200 \*   
## Outlet\_SizeSmall 144.05363 45.67140 3.154 0.001615 \*\*   
## Outlet\_Location\_TypeTier 2 -100.36636 90.23496 -1.112 0.266050   
## Outlet\_Location\_TypeTier 3 -326.31974 154.89294 -2.107 0.035169 \*   
## Outlet\_TypeSupermarket Type1 1473.06659 140.73922 10.467 < 2e-16 \*\*\*  
## Outlet\_TypeSupermarket Type2 1076.77568 136.08116 7.913 2.83e-15 \*\*\*  
## Outlet\_TypeSupermarket Type3 3639.29114 178.13923 20.429 < 2e-16 \*\*\*  
## Year -34.82155 10.49271 -3.319 0.000908 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1129 on 8494 degrees of freedom  
## Multiple R-squared: 0.5637, Adjusted R-squared: 0.5623   
## F-statistic: 392 on 28 and 8494 DF, p-value: < 2.2e-16

Now we have got R² = 0.5623 which means it improved than before.

Let’s check out regression plot to find out more ways to improve this model.

par(mfrow=c(2,2))  
plot(linear\_model)

 The shape of the Residuals graph shows that the model is suffering from heteroskedasticity (unequal variance in error terms), which can be solved by taking the log of response variable:

linear\_model <- lm(log(Item\_Outlet\_Sales) ~ ., data = new\_train)  
summary(linear\_model)

##   
## Call:  
## lm(formula = log(Item\_Outlet\_Sales) ~ ., data = new\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.30630 -0.29386 0.06823 0.37926 1.35991   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.763e+00 1.450e-01 32.857 < 2e-16 \*\*\*  
## Item\_Weight -4.195e-04 1.388e-03 -0.302 0.76254   
## Item\_Fat\_ContentRegular 1.359e-02 1.345e-02 1.011 0.31208   
## Item\_Visibility -3.162e-02 1.256e-01 -0.252 0.80129   
## Item\_TypeBreads 2.819e-02 4.003e-02 0.704 0.48137   
## Item\_TypeBreakfast -6.944e-02 5.554e-02 -1.250 0.21120   
## Item\_TypeCanned 2.532e-02 2.990e-02 0.847 0.39712   
## Item\_TypeDairy -6.904e-02 2.964e-02 -2.329 0.01987 \*   
## Item\_TypeFrozen Foods -5.422e-02 2.804e-02 -1.934 0.05318 .   
## Item\_TypeFruits and Vegetables -4.600e-03 2.618e-02 -0.176 0.86051   
## Item\_TypeHard Drinks -2.201e-02 4.296e-02 -0.512 0.60849   
## Item\_TypeHealth and Hygiene 1.126e-02 3.240e-02 0.347 0.72826   
## Item\_TypeHousehold -2.632e-02 2.855e-02 -0.922 0.35654   
## Item\_TypeMeat 2.239e-02 3.366e-02 0.665 0.50586   
## Item\_TypeOthers 2.354e-03 4.698e-02 0.050 0.96004   
## Item\_TypeSeafood 5.572e-03 7.050e-02 0.079 0.93700   
## Item\_TypeSnack Foods -1.425e-03 2.632e-02 -0.054 0.95683   
## Item\_TypeSoft Drinks -2.241e-02 3.342e-02 -0.671 0.50247   
## Item\_TypeStarchy Foods -4.744e-02 4.908e-02 -0.967 0.33376   
## Item\_MRP 8.316e-03 9.416e-05 88.313 < 2e-16 \*\*\*  
## Outlet\_SizeHigh 3.547e-01 1.221e-01 2.905 0.00368 \*\*   
## Outlet\_SizeMedium 8.223e-02 3.386e-02 2.429 0.01517 \*   
## Outlet\_SizeSmall 6.748e-02 2.175e-02 3.102 0.00193 \*\*   
## Outlet\_Location\_TypeTier 2 -4.215e-02 4.298e-02 -0.981 0.32673   
## Outlet\_Location\_TypeTier 3 -1.566e-01 7.378e-02 -2.122 0.03387 \*   
## Outlet\_TypeSupermarket Type1 1.744e+00 6.704e-02 26.013 < 2e-16 \*\*\*  
## Outlet\_TypeSupermarket Type2 1.540e+00 6.482e-02 23.753 < 2e-16 \*\*\*  
## Outlet\_TypeSupermarket Type3 2.607e+00 8.485e-02 30.728 < 2e-16 \*\*\*  
## Year -1.495e-02 4.998e-03 -2.992 0.00278 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5378 on 8494 degrees of freedom  
## Multiple R-squared: 0.7214, Adjusted R-squared: 0.7205   
## F-statistic: 785.4 on 28 and 8494 DF, p-value: < 2.2e-16

We have got an improved model with R² = 0.72

# Check our RMSE  
library(Metrics)  
rmse(new\_train$Item\_Outlet\_Sales, exp(linear\_model$fitted.values))

## [1] 1140.004

# Building a decision tree  
  
# Loading required libraries  
library(rpart)  
library(e1071)  
library(rpart.plot)  
library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following objects are masked from 'package:Metrics':  
##   
## precision, recall

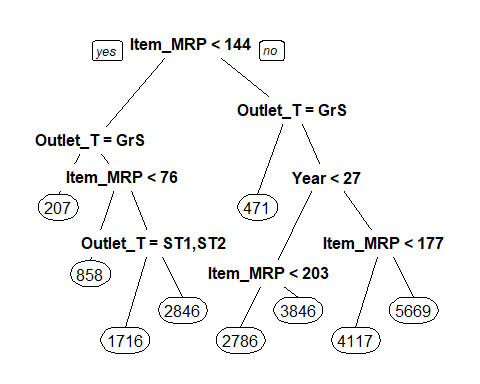
# Setting the tree control parameters  
fitControl <- trainControl(method = "cv", number = 5)  
cartGrid <- expand.grid(.cp=(1:50)\*0.01)  
  
tree\_model <- train(Item\_Outlet\_Sales ~ ., data = new\_train, method = "rpart", trControl = fitControl, tuneGrid = cartGrid)

## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo, :  
## There were missing values in resampled performance measures.

print(tree\_model)

## CART   
##   
## 8523 samples  
## 9 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold)   
## Summary of sample sizes: 6820, 6818, 6817, 6819, 6818   
## Resampling results across tuning parameters:  
##   
## cp RMSE Rsquared MAE   
## 0.01 1124.619 0.5664181 809.8138  
## 0.02 1175.964 0.5256572 862.8547  
## 0.03 1190.628 0.5135925 873.8514  
## 0.04 1253.982 0.4605989 948.1363  
## 0.05 1253.982 0.4605989 948.1363  
## 0.06 1308.878 0.4119165 1008.7161  
## 0.07 1316.301 0.4055564 1019.1364  
## 0.08 1316.301 0.4055564 1019.1364  
## 0.09 1462.634 0.2651056 1114.8159  
## 0.10 1491.496 0.2360265 1131.3207  
## 0.11 1491.496 0.2360265 1131.3207  
## 0.12 1491.496 0.2360265 1131.3207  
## 0.13 1491.496 0.2360265 1131.3207  
## 0.14 1491.496 0.2360265 1131.3207  
## 0.15 1491.496 0.2360265 1131.3207  
## 0.16 1491.496 0.2360265 1131.3207  
## 0.17 1491.496 0.2360265 1131.3207  
## 0.18 1491.496 0.2360265 1131.3207  
## 0.19 1491.496 0.2360265 1131.3207  
## 0.20 1491.496 0.2360265 1131.3207  
## 0.21 1491.496 0.2360265 1131.3207  
## 0.22 1491.496 0.2360265 1131.3207  
## 0.23 1491.496 0.2360265 1131.3207  
## 0.24 1706.059 NaN 1346.5860  
## 0.25 1706.059 NaN 1346.5860  
## 0.26 1706.059 NaN 1346.5860  
## 0.27 1706.059 NaN 1346.5860  
## 0.28 1706.059 NaN 1346.5860  
## 0.29 1706.059 NaN 1346.5860  
## 0.30 1706.059 NaN 1346.5860  
## 0.31 1706.059 NaN 1346.5860  
## 0.32 1706.059 NaN 1346.5860  
## 0.33 1706.059 NaN 1346.5860  
## 0.34 1706.059 NaN 1346.5860  
## 0.35 1706.059 NaN 1346.5860  
## 0.36 1706.059 NaN 1346.5860  
## 0.37 1706.059 NaN 1346.5860  
## 0.38 1706.059 NaN 1346.5860  
## 0.39 1706.059 NaN 1346.5860  
## 0.40 1706.059 NaN 1346.5860  
## 0.41 1706.059 NaN 1346.5860  
## 0.42 1706.059 NaN 1346.5860  
## 0.43 1706.059 NaN 1346.5860  
## 0.44 1706.059 NaN 1346.5860  
## 0.45 1706.059 NaN 1346.5860  
## 0.46 1706.059 NaN 1346.5860  
## 0.47 1706.059 NaN 1346.5860  
## 0.48 1706.059 NaN 1346.5860  
## 0.49 1706.059 NaN 1346.5860  
## 0.50 1706.059 NaN 1346.5860  
##   
## RMSE was used to select the optimal model using the smallest value.  
## The final value used for the model was cp = 0.01.

main\_tree <- rpart(Item\_Outlet\_Sales ~ ., data = new\_train, control = rpart.control(cp=0.01))  
prp(main\_tree)



pre\_score <- predict(main\_tree, type = "vector")  
rmse(new\_train$Item\_Outlet\_Sales, pre\_score)

## [1] 1102.774

# Predict the test data using the decision tree  
main\_predict <- predict(main\_tree, newdata = new\_test, type = "vector")  
sub\_file <- data.frame(Item\_Identifier = test$Item\_Identifier, Outlet\_Identifier = test$Outlet\_Identifier, Item\_Outlet\_Sales = main\_predict)  
write.csv(sub\_file, 'Decision\_tree\_sales.csv')