WallMart Coursework in R

Code ▾

Installing Packages and Libraries

Hide

```
install.packages("data.table")
library(data.table)
install.packages("ggplot2")
library(ggplot2)
install.packages("psych")
library(psych)
install.packages("corrplot")
library(corrplot)
install.packages("dplyr")
library(dplyr)
install.packages("plyr")
library(plyr)
install.packages("Amelia")
library(Amelia)
library(tidyr)
library(stringr)
library(dummies)
```

Setting the Working Directory & Loading the datasets as data frames

Hide

```
#setwd("D:/RCW/")
train <- read.csv("D:/RCW/Train.csv")
test = read.csv("D:/RCW/Test.csv")
```

Printing the top 6 rows of the train & test data frames by using head()

Hide

```
head(train)
```

	Item_Identifier <fctr>	Item_Weight <dbl>	Item_Fat_Content <fctr>	Item_Visibility <dbl>	Item_Type <fctr>
1	FDA15	9.300	Low Fat	0.01604730	Dairy
2	DRC01	5.920	Regular	0.01927822	Soft Drinks
3	FDN15	17.500	Low Fat	0.01676007	Meat
4	FDX07	19.200	Regular	0.00000000	Fruits and Vegetables
5	NCD19	8.930	Low Fat	0.00000000	Household
6	FDP36	10.395	Regular	0.00000000	Baking Goods

6 rows | 1-7 of 12 columns

Hide

```
head(test)
```

	Item_Identifier <fctr>	Item_Weight <dbl>	Item_Fat_Content <fctr>	Item_Visibility <dbl>	Item_Type <fctr>
1	FDW58	20.750	Low Fat	0.007564836	Snack Foods
2	FDW14	8.300	reg	0.038427677	Dairy
3	NCN55	14.600	Low Fat	0.099574908	Others
4	FDQ58	7.315	Low Fat	0.015388393	Snack Foods
5	FDY38	NA	Regular	0.118599314	Dairy
6	FDH56	9.800	Regular	0.063817206	Fruits and Vegetables

6 rows | 1-7 of 11 columns

Checking the Dimensions of the dataset by using dim() The train dataset has 8523(rows) 12(cols) The test dataset has 5681(rows) 11(cols)

Hide

```
dim(train)
```

```
[1] 8523   12
```

Hide

```
dim(test)
```

```
[1] 5681   11
```

Checking the column names to find the missing columns We find the ‘Item_Outlet_Sales’ is missing from test dataset. This is because we will be predicting the values of ‘Item_Outlet_Sales’

Hide

```
names(train)
```

```
 [1] "Item_Identifier"           "Item_Weight"               "Item_Fat_Content"
 [4] "Item_Visibility"           "Item_Type"                 "Item_MRP"
 [7] "Outlet_Identifier"         "Outlet_Establishment_Year" "Outlet_Size"
[10] "Outlet_Location_Type"      "Outlet_Type"               "Item_Outlet_Sales"
```

Hide

```
names(test)
```

```
 [1] "Item_Identifier"           "Item_Weight"               "Item_Fat_Content"
 [4] "Item_Visibility"           "Item_Type"                 "Item_MRP"
 [7] "Outlet_Identifier"         "Outlet_Establishment_Year" "Outlet_Size"
[10] "Outlet_Location_Type"      "Outlet_Type"
```

Checking if this data has missing values. We are using table to group the values by False and True. We find 1463 NA values in the train dataset.

Hide

```
table(is.na(train))
```

```


 FALSE   TRUE 
100813   1463 
```

Checking the variables wisth the count of NA values. We find only Item_Weight has the 1463 NA values.

Hide

```
colSums(is.na(train))
```

```
          Item_Identifier               Item_Weight          Item_Fat_Content           Item_Vis
ibility 
                        0                      1463                         0                   
      0 
                Item_Type                  Item_MRP         Outlet_Identifier Outlet_Establishme
nt_Year 
                        0                         0                         0                   
      0 
              Outlet_Size      Outlet_Location_Type               Outlet_Type           Item_Outle
t_Sales 
                        0                         0                         0                   
      0 
```

Checking the variables and their types in train dataset

Hide

```
str(train)
```

```
'data.frame':   8523 obs. of  12 variables:
 $ Item_Identifier          : Factor w/ 1559 levels "DRA12","DRA24",..: 157 9 663 1122 1298 759 
697 739 441 991 ...
 $ Item_Weight              : num  9.3 5.92 17.5 19.2 8.93 ...
 $ Item_Fat_Content         : Factor w/ 5 levels "LF","low fat",..: 3 5 3 5 3 5 5 3 5 5 ...
 $ Item_Visibility          : num  0.016 0.0193 0.0168 0 0 ...
 $ Item_Type                : Factor w/ 16 levels "Baking Goods",..: 5 15 11 7 10 1 14 14 6 6 
...
 $ Item_MRP                 : num  249.8 48.3 141.6 182.1 53.9 ...
 $ Outlet_Identifier        : Factor w/ 10 levels "OUT010","OUT013",..: 10 4 10 1 2 4 2 6 8 3 
...
 $ Outlet_Establishment_Year: int  1999 2009 1999 1998 1987 2009 1987 1985 2002 2007 ...
 $ Outlet_Size              : Factor w/ 4 levels "","High","Medium",..: 3 3 3 1 2 3 2 3 1 1 ...
 $ Outlet_Location_Type     : Factor w/ 3 levels "Tier 1","Tier 2",..: 1 3 1 3 3 3 3 3 2 2 ...
 $ Outlet_Type              : Factor w/ 4 levels "Grocery Store",..: 2 3 2 1 2 3 2 4 2 2 ...
 $ Item_Outlet_Sales        : num  3735 443 2097 732 995 ...
```

We will see a summary of the train dataset.

Hide

```
summary(train)
```

```
Item_Identifier  Item_Weight     Item_Fat_Content Item_Visibility                  Item_Type

FDG33  :  10    Min.   : 4.555   LF     : 316     Min.   :0.00000   Fruits and Vegetables:1232

FDW13  :  10    1st Qu.: 8.774   low fat: 112     1st Qu.:0.02699   Snack Foods          :1200

DRE49  :   9    Median :12.600   Low Fat:5089     Median :0.05393   Household            : 910

DRN47  :   9    Mean   :12.858   reg    : 117     Mean   :0.06613   Frozen Foods         : 856

FDD38  :   9    3rd Qu.:16.850   Regular:2889     3rd Qu.:0.09459   Dairy                : 682

FDF52  :   9    Max.   :21.350                    Max.   :0.32839   Canned               : 649

(Other):8467    NA's   :1463                                        (Other)              :2994

   Item_MRP      Outlet_Identifier Outlet_Establishment_Year Outlet_Size   Outlet_Location_Type
Min.   : 31.29   OUT027 : 935      Min.   :1985                    :2410   Tier 1:2388
1st Qu.: 93.83   OUT013 : 932      1st Qu.:1987              High  : 932   Tier 2:2785
Median :143.01   OUT035 : 930      Median :1999              Medium:2793   Tier 3:3350
Mean   :140.99   OUT046 : 930      Mean   :1998              Small :2388
3rd Qu.:185.64   OUT049 : 930      3rd Qu.:2004
Max.   :266.89   OUT045 : 929      Max.   :2009
                 (Other):2937
           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
```

From the above information we can:

Exploring the Numerical Columns:

1.Item_Weight - There are 1463 NA Values

2.Item Visibility - Contains no NA Values, but contains 0 values.

3.Item_MRP - Contains No NA/0 values.Also has an acceptable price range with no outliers.

4.Outlet_Establishment_Year - Contains no NA/0 values. Average mean is 1997, implying mostly old stores.

5.Item_Outlet_Sales - Contains no NA/0 values.

Exploring the Factor Columns:

1.Item_Identifier - Contains 1559 unique values

Hide

```
#install.packages("plyr")
#library(plyr)
#library(dplyr)
train %>%
  summarise(n_distinct(Item_Identifier))
```

n_distinct(Item_Identifier) <int>
1559

1 row

2.Item_Fat_Content - We find the level values Low Fat/low fat/LF are same but typed incorrectly.

Hide

```
train %>%
group_by(Item_Fat_Content) %>% summarise(Count = n()) %>% arrange(desc(Count))
```

Item_Fat_Content <fctr>	Count <int>
Low Fat	5089
Regular	2889
LF	316
reg	117
low fat	112

5 rows

3.Item_Type - Categories of Items with counts

Hide

```
train%>%
group_by(Item_Type) %>%
summarise(Count = n()) %>% arrange(desc(Count))
```

Item_Type <fctr>	Count <int>
Fruits and Vegetables	1232
Snack Foods	1200
Household	910
Frozen Foods	856
Dairy	682

Item_Type <fctr>	Count <int>
Canned	649
Baking Goods	648
Health and Hygiene	520
Soft Drinks	445
Meat	425

1-10 of 16 rows Previous **1** 2 Next

4.Outlet_Identifier - There are Item information from 10 different Outlets

Hide

```
train %>%
group_by(Outlet_Identifier) %>%
summarise(Count = n()) %>% arrange(desc(Count))
```

Outlet_Identifier <fctr>	Count <int>
OUT027	935
OUT013	932
OUT035	930
OUT046	930
OUT049	930
OUT045	929
OUT018	928
OUT017	926
OUT010	555
OUT019	528

1-10 of 10 rows

5.Outlet_Size - Outlet Size data not properly levelled. (2410 counts)

Hide

```
train%>%
group_by(Outlet_Size) %>% summarise(Count = n())
```

Outlet_Size <fctr>	Count <int>

Outlet_Size	**Count**
<fctr>	<int>
	2410
High	932
Medium	2793
Small	2388

4 rows

6.Outlet_Location_Type - Number of Outlet Location type with counts. We find the data is normally distributed.

Hide

```
train%>%
group_by(Outlet_Location_Type) %>%
summarise(Count = n()) %>% arrange(desc(Count))
```

Outlet_Location_Type	**Count**
<fctr>	<int>
Tier 3	3350
Tier 2	2785
Tier 1	2388

3 rows

7.Outlet_Type - We find the Types of Outlet

Hide

```
train%>%
group_by(Outlet_Type)%>%
summarise(Count=n())%>% arrange(desc(Count))
```

Outlet_Type	**Count**
<fctr>	<int>
Supermarket Type1	5577
Grocery Store	1083
Supermarket Type3	935
Supermarket Type2	928

4 rows

Data Manipulation

We are creating a new variable in test dataset Item_Outlet_Sales, to match our number of rows with train dataset.

Hide

```
test$Item_Outlet_Sales <- 1
names(test)
```

```
 [1] "Item_Identifier"           "Item_Weight"               "Item_Fat_Content"
 [4] "Item_Visibility"           "Item_Type"                 "Item_MRP"
 [7] "Outlet_Identifier"         "Outlet_Establishment_Year" "Outlet_Size"
[10] "Outlet_Location_Type"      "Outlet_Type"               "Item_Outlet_Sales"
```

Now, we are combining thee train and test data with rbind function

Hide

```
combi <- rbind(train, test)
dim(combi)
```

```
[1] 14204    12
```

We are imputing the NA values in Item_Weight with the median of the values of the column. To calculate the median of the non-missing values if are passing the argument na.rm=TRUE

Hide

```
combi$Item_Weight[is.na(combi$Item_Weight)] <- median(combi$Item_Weight, na.rm = TRUE)
summary(combi$Item_Weight)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  4.555   9.300  12.600  12.760  16.000  21.350 
```

There are 0 values in in Item_visibility, therefore, we also impute the 0s with median of the column values.

Hide

```
combi$Item_Visibility <- ifelse(combi$Item_Visibility == 0, median(combi$Item_Visibility),combi
$Item_Visibility)
summary(combi$Item_Visibility)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
0.003575 0.033143 0.054023 0.069296 0.094037 0.328391 
```

Renaming the blank level in of Outlet_Size to ‘Other’

Hide

```
levels(combi$Outlet_Size)[1] <- "Other"
table(combi$Outlet_Size)
```

```
 Other   High Medium  Small 
  4016   1553   4655   3980 
```

Renaming the levels of Item_Fat_Content to ‘Low Fat’ & ‘Regular’

Hide

```
#library(plyr)
#combi$Item_Fat_Content <- revalue(combi$Item_Fat_Content,c("LF" = "Low Fat", "reg" = "Regular",
 "low fat" = "Low Fat"))
table(combi$Item_Fat_Content)
```

```
Low Fat Regular 
   9185    5019 
```

Data Visualisation

Hide

```
combi_encoded=as.data.frame(combi)
str(combi_encoded)
```

```
'data.frame':   14204 obs. of  12 variables:
 $ Item_Identifier          : Factor w/ 1559 levels "DRA12","DRA24",..: 157 9 663 1122 1298 759
697 739 441 991 ...
 $ Item_Weight              : num  9.3 5.92 17.5 19.2 8.93 ...
 $ Item_Fat_Content         : Factor w/ 2 levels "Low Fat","Regular": 1 2 1 2 1 2 2 1 2 2 ...
 $ Item_Visibility          : num  0.016 0.0193 0.0168 0.054 0.054 ...
 $ Item_Type                : Factor w/ 16 levels "Baking Goods",..: 5 15 11 7 10 1 14 14 6 6
...
 $ Item_MRP                 : num  249.8 48.3 141.6 182.1 53.9 ...
 $ Outlet_Identifier        : Factor w/ 10 levels "OUT010","OUT013",..: 10 4 10 1 2 4 2 6 8 3
...
 $ Outlet_Establishment_Year: int  1999 2009 1999 1998 1987 2009 1987 1985 2002 2007 ...
 $ Outlet_Size              : Factor w/ 4 levels "Other","High",..: 3 3 3 1 2 3 2 3 1 1 ...
 $ Outlet_Location_Type     : Factor w/ 3 levels "Tier 1","Tier 2",..: 1 3 1 3 3 3 3 3 2 2 ...
 $ Outlet_Type              : Factor w/ 4 levels "Grocery Store",..: 2 3 2 1 2 3 2 4 2 2 ...
 $ Item_Outlet_Sales        : num  3735 443 2097 732 995 ...
```

Dividing data to Train & Test before Label & Hot Encoding

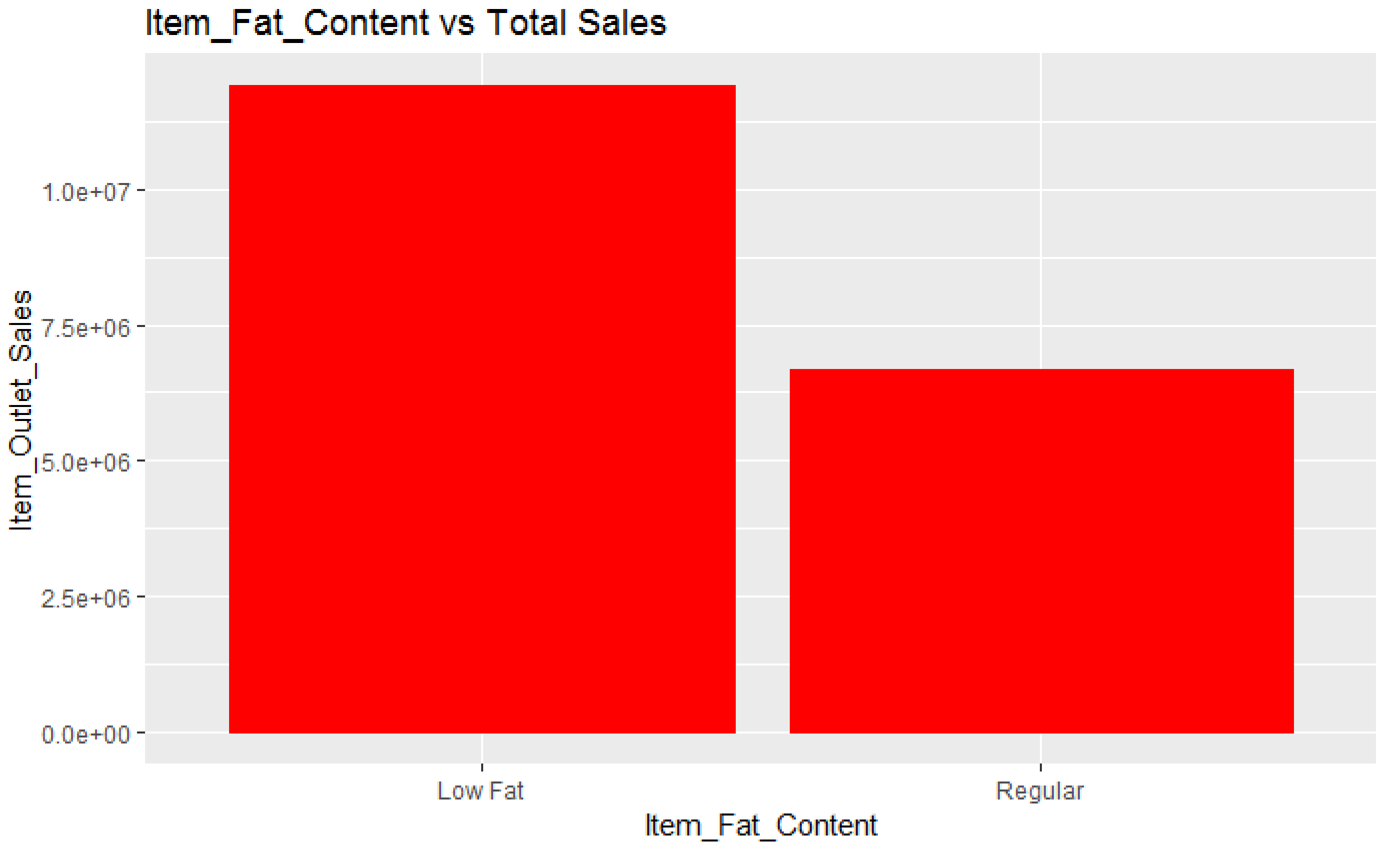
Hide

```
new_train_combi <- combi %>% filter(Item_Outlet_Sales != 1)
new_test_combi <- combi %>% filter(Item_Outlet_Sales == 1)
str(new_train_combi)
str(new_test_combi)
```

We have tried to visualise Item_Outlet_Sales with different Categorical Values

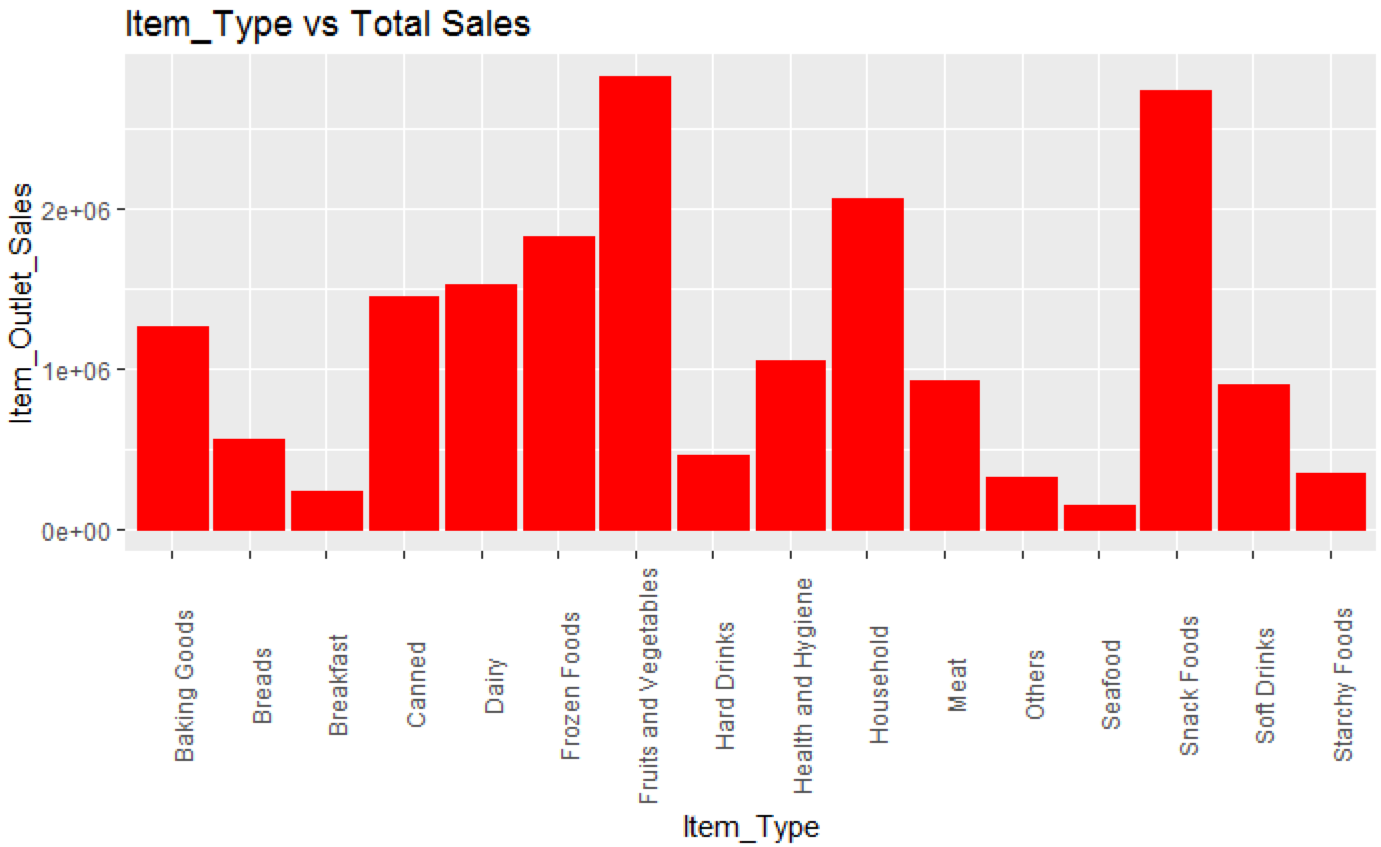
Hide

```
ggplot(new_train_combi, aes(Item_Fat_Content, Item_Outlet_Sales)) + geom_bar(stat = "identity",
 color = "red") + ggtitle("Item_Fat_Content vs Total Sales")
```



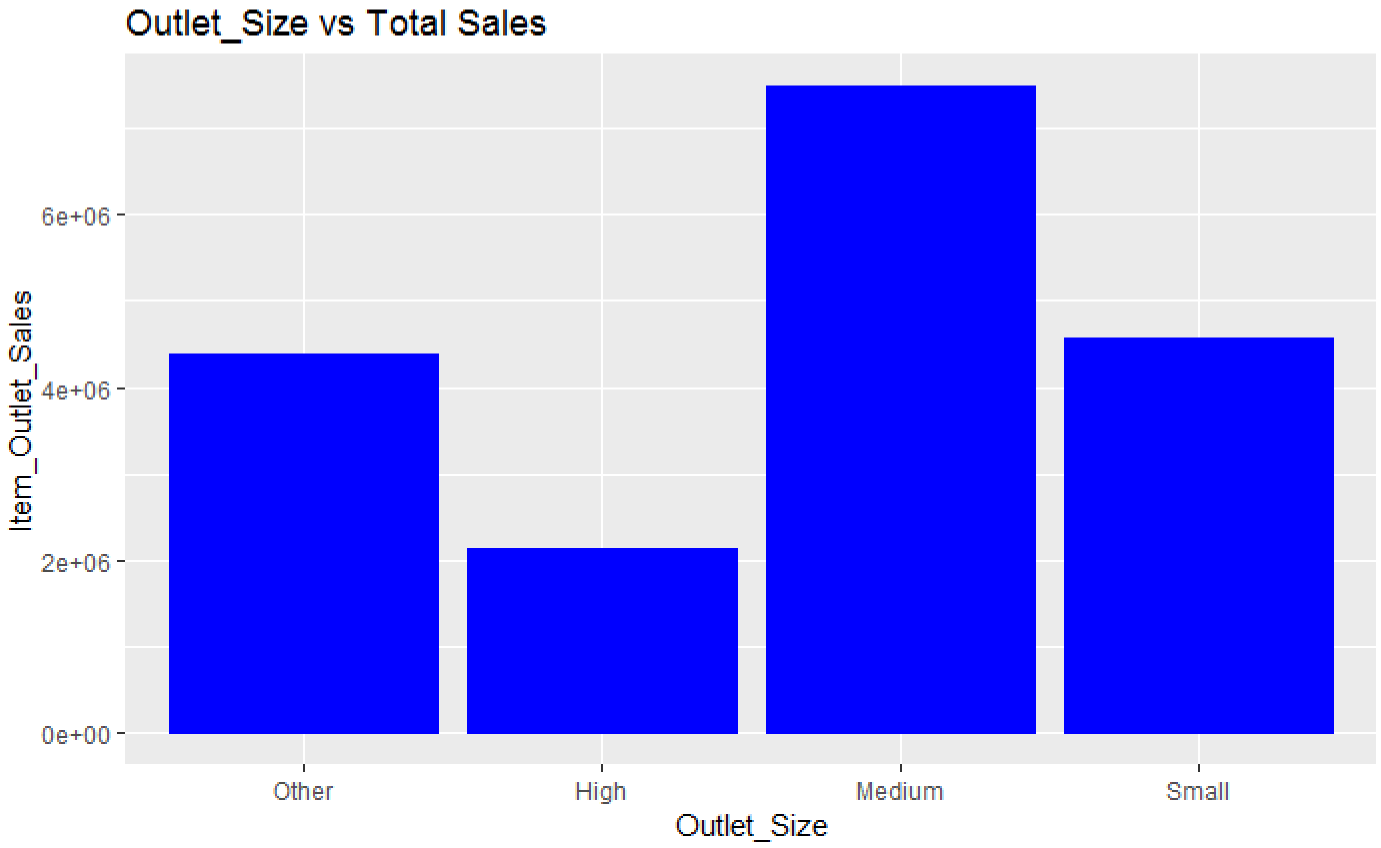
Hide

```
ggplot(new_train_combi, aes(Item_Type, Item_Outlet_Sales)) + geom_bar(stat = "identity", color =
 "red") + theme(axis.text.x = element_text(angle = 90), axis.text.y = element_text(angle = 0)) +
 ggtitle("Item_Type vs Total Sales")
```



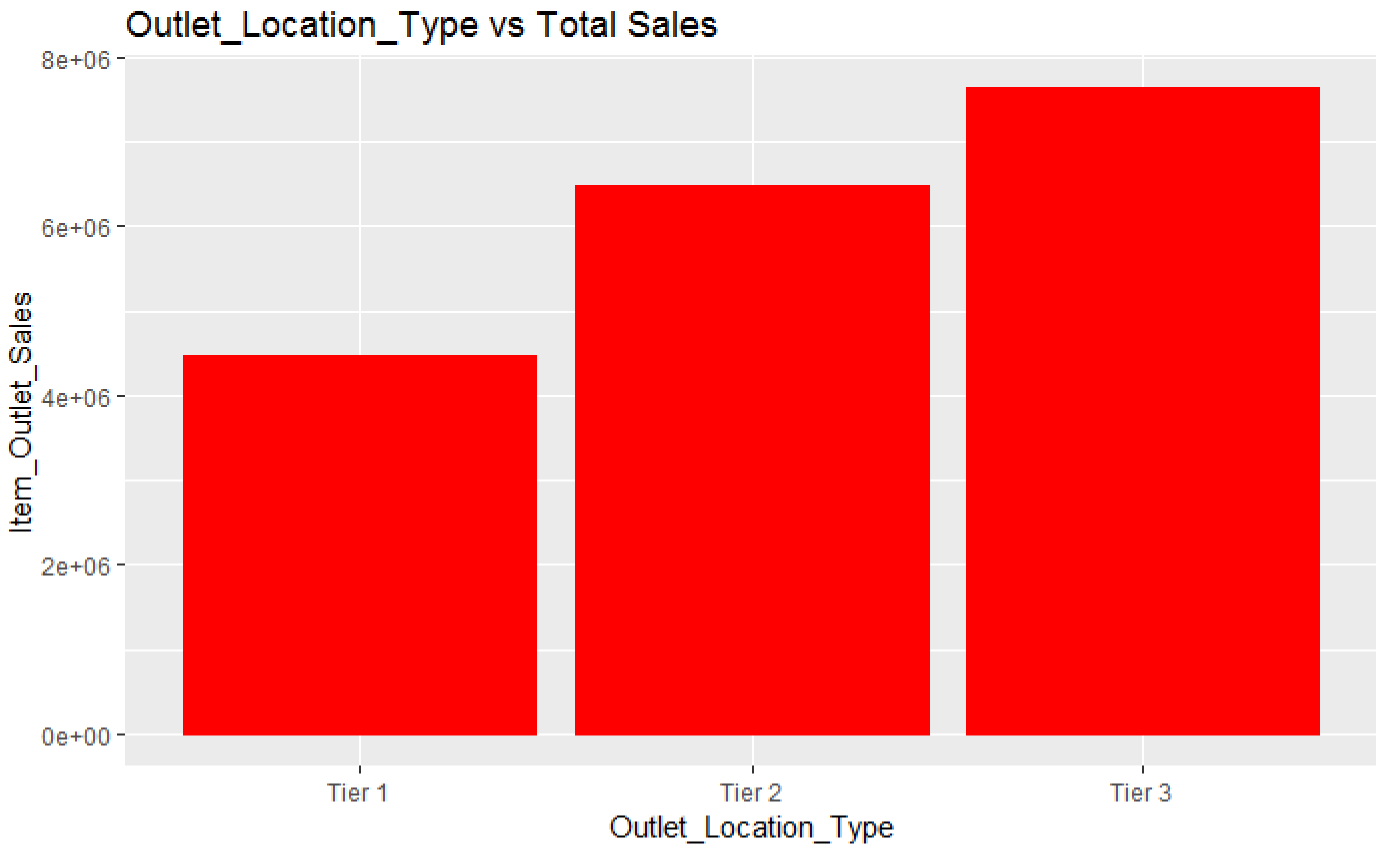
Hide

```
ggplot(new_train_combi, aes(Outlet_Size, Item_Outlet_Sales)) + geom_bar(stat = "identity", color
 = "blue") + ggtitle("Outlet_Size vs Total Sales")
```



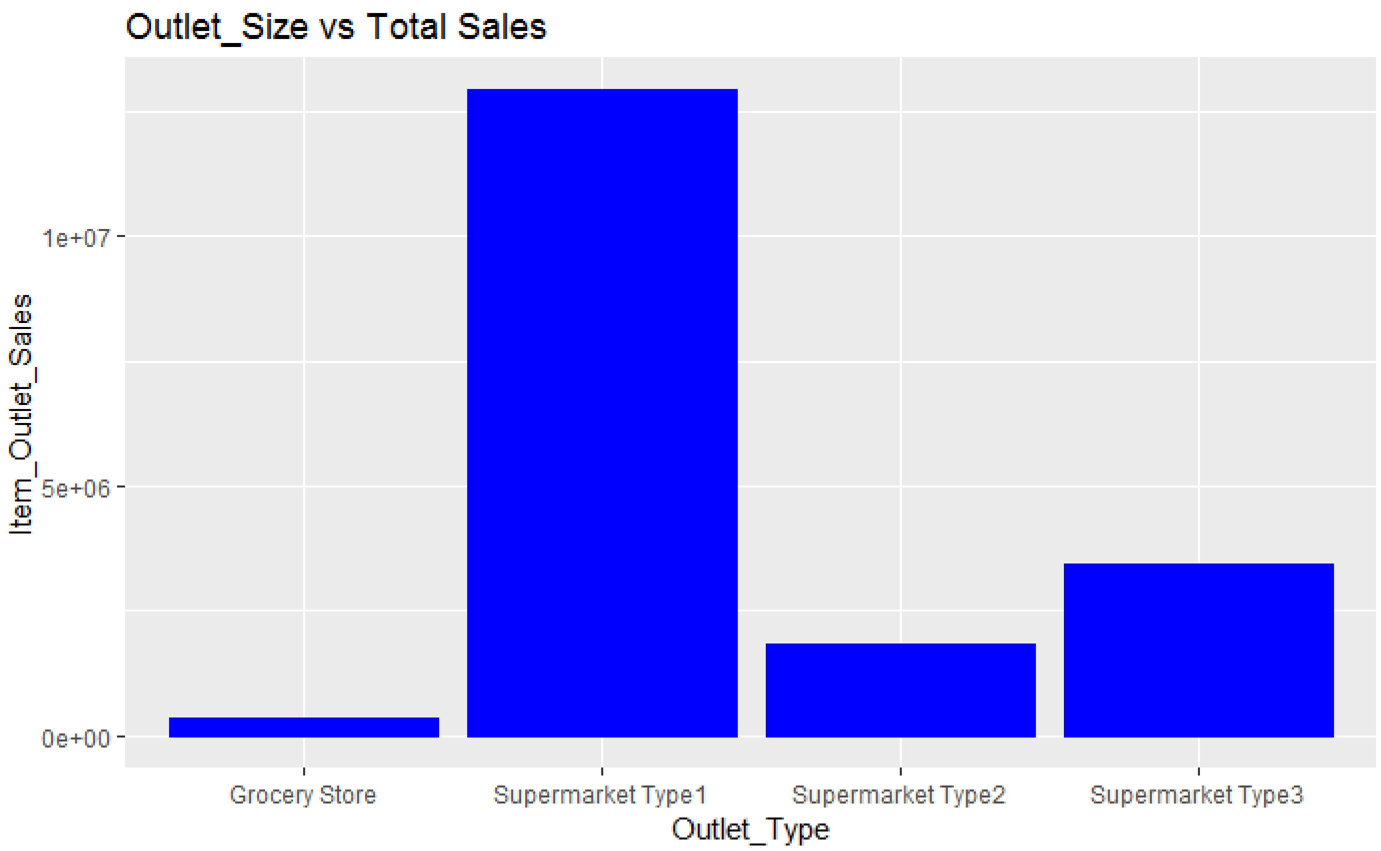
Hide

```
ggplot(new_train_combi, aes(Outlet_Location_Type, Item_Outlet_Sales)) + geom_bar(stat = "identity", color = "red") + ggtitle("Outlet_Location_Type vs Total Sales")
```



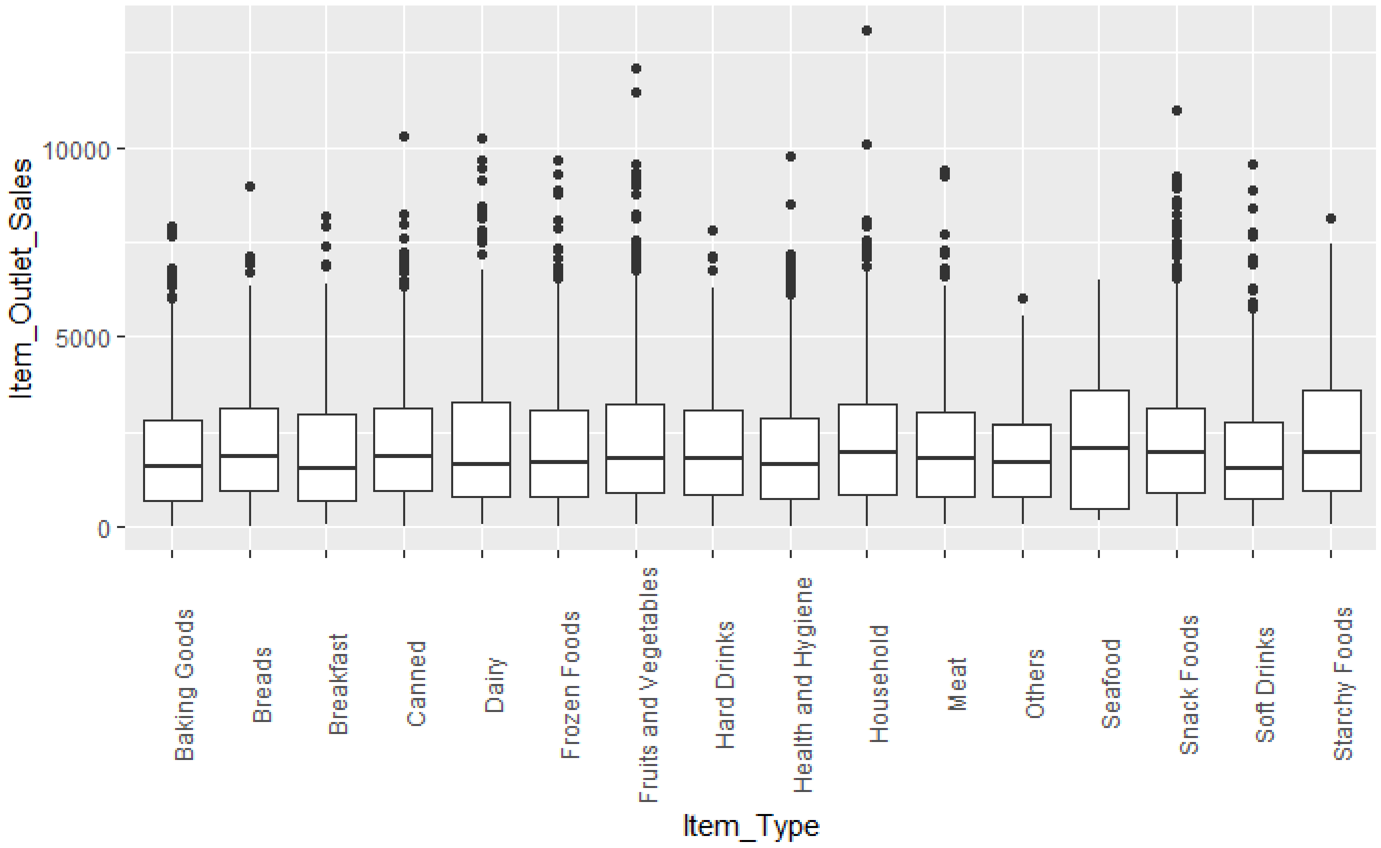
Hide

```
ggplot(new_train_combi, aes(Outlet_Type, Item_Outlet_Sales)) + geom_bar(stat = "identity", color = "blue") + ggtitle("Outlet_Size vs Total Sales")
```


Outlet_Size vs Total Sales

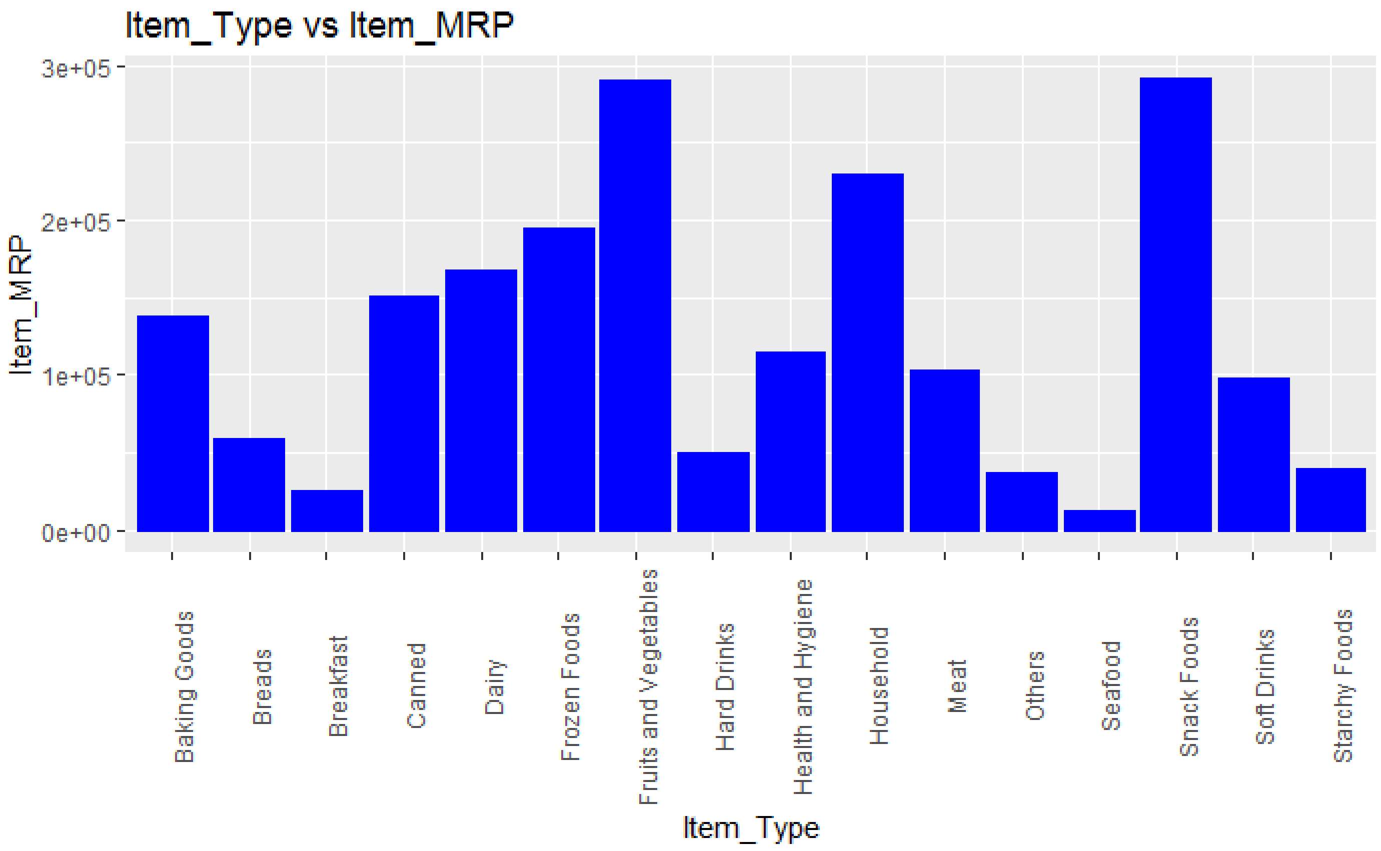
Hide

```
qplot(x=Item_Type,y=Item_Outlet_Sales, data=new_train_combi,geom = "boxplot",) + theme(axis.text.x = element_text(angle = 90), axis.text.y = element_text(angle = 0))
```



Hide

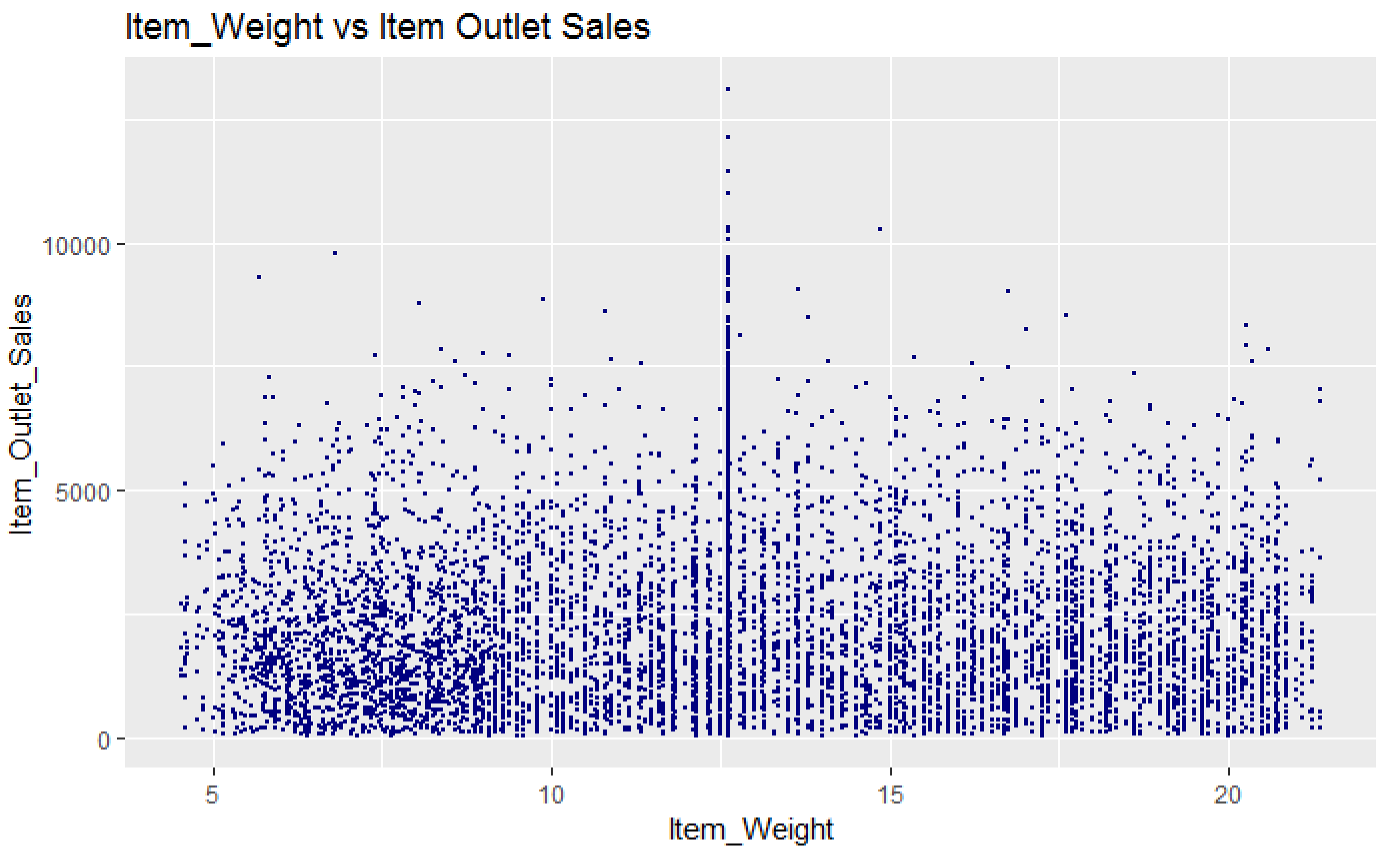
```
ggplot(combi, aes(Item_Type, Item_MRP)) + geom_bar(stat = "identity", color = "blue")+ theme(axi
s.text.x = element_text(angle = 90), axis.text.y = element_text(angle = 0)) + ggtitle("Item_Type
 vs Item_MRP")
```



We have tried to visualise Item_Outlet_Sales with different Continuous Values:

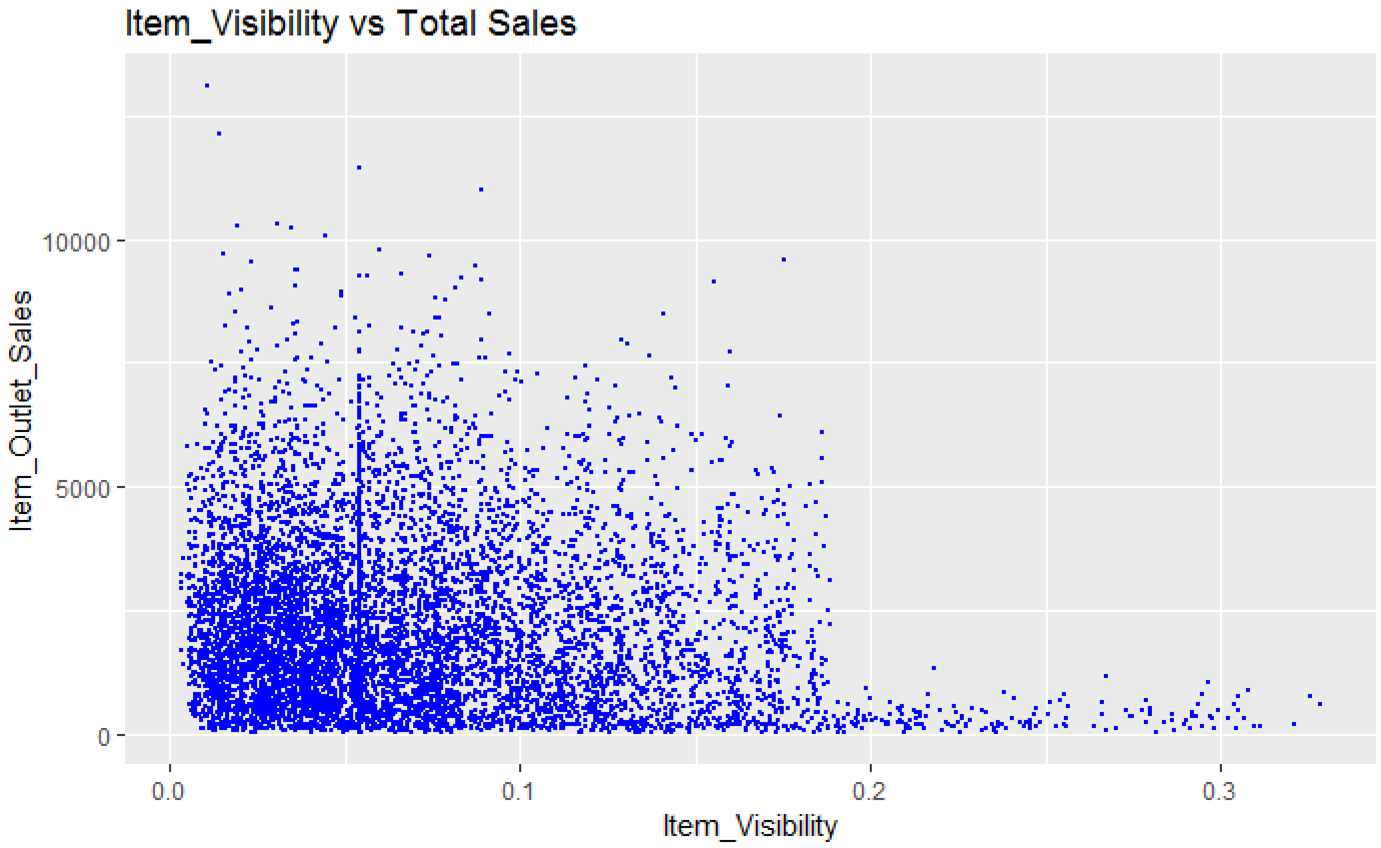
Hide

```
ggplot(new_train_combi, aes(Item_Weight, Item_Outlet_Sales)) + geom_point(size = .5, color="nav
y") + ggtitle("Item_Weight vs Item Outlet Sales")
```



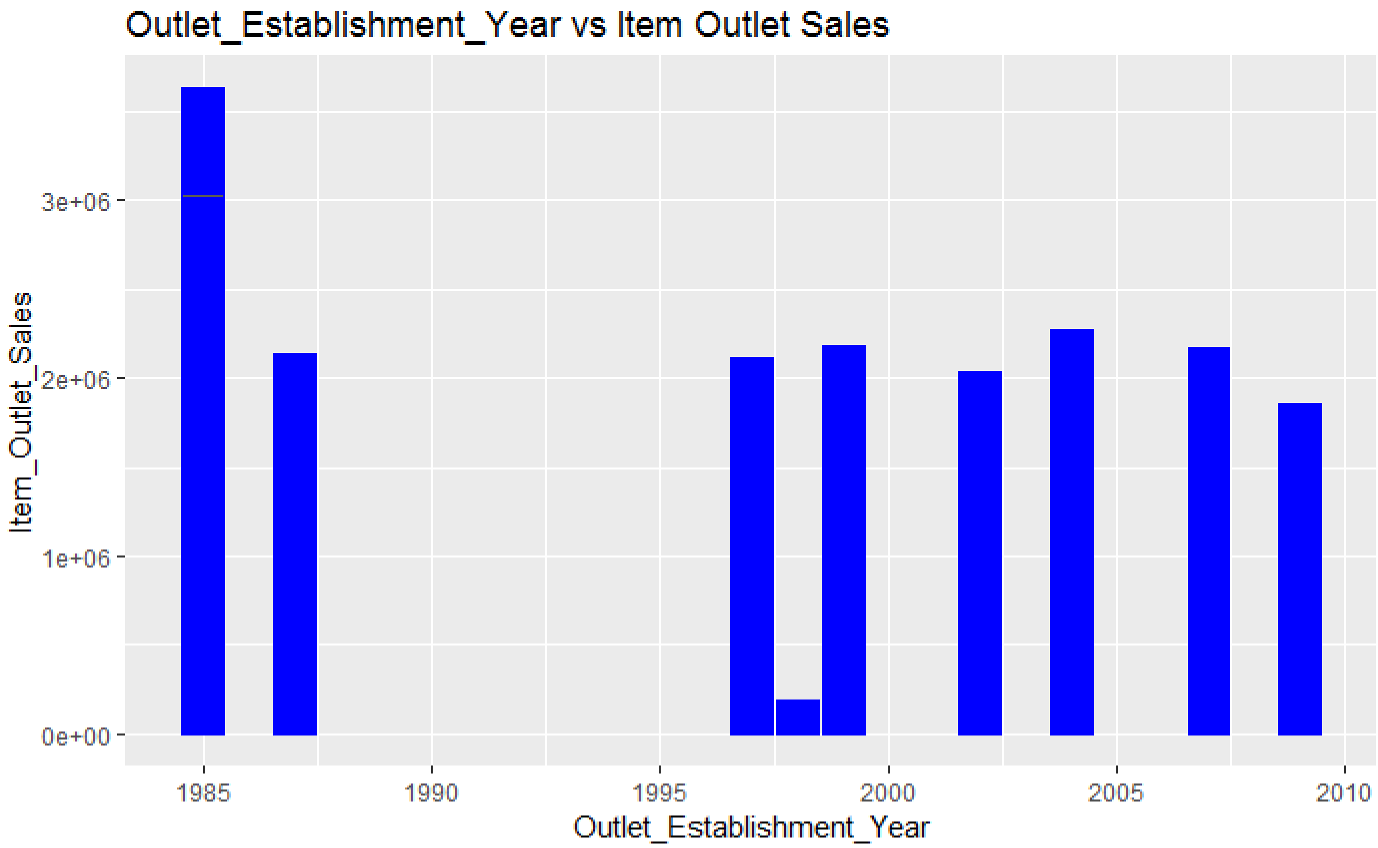
Hide

```
ggplot(new_train_combi, aes(Item_Visibility, Item_Outlet_Sales)) + geom_point(size = .5,  color
 = "blue") + ggtitle("Item_Visibility vs Total Sales")
```



Hide

```
ggplot(new_train_combi, aes(Outlet_Establishment_Year, Item_Outlet_Sales)) + geom_bar(stat = "identity", color="blue")  + ggtitle("Outlet_Establishment_Year vs Item Outlet Sales")
```



Manupulating data with Label Encoding & Hot Encoding.

Creating a dataframe combi_encoded similar to combi.

Hide

```
combi_encoded=as.data.frame(combi)
str(combi_encoded)
```

Label Encoding. We will change the categorical variable Item_Fat_Content to numeric 0 & 1.

Hide

```
combi_encoded$Item_Fat_Content <- ifelse(combi_encoded$Item_Fat_Content == "Regular",1,0)
str(combi_encoded)
```

Hot Encoding. We will use dummy.data.frame() to split the catrgorical variable to a matrix of variables 0 and 1,

Hide

```
library(dummies)
combi_encoded_dummies <- dummy.data.frame(combi_encoded, names = c('Outlet_Size','Outlet_Location_Type','Outlet_Type'),sep = '_')

str(combi_encoded_dummies)
```

Now, We will save the dataframe with all columns with int and num and Drop thecolumns with Categorical variables/Factors.

Item_Identifier, Outlet_Identifier,Item_Type has not been converted to matrix because of the high no of factor levels, which we cannot compute due to limited system resources.

Hence we are removing the 3 variables from the final variable.

Hide

```
combi_encoded_dummies_drop <- select(combi_encoded_dummies, -c(Item_Identifier, Outlet_Identifie
r,Item_Type))
str(combi_encoded_dummies_drop)
```

```
'data.frame':	14204 obs. of  17 variables:
 $ Item_Weight                  : num  9.3 5.92 17.5 19.2 8.93 ...
 $ Item_Fat_Content             : num  0 1 0 1 0 1 1 0 1 1 ...
 $ Item_Visibility              : num  0.016 0.0193 0.0168 0.054 0.054 ...
 $ Item_MRP                     : num  249.8 48.3 141.6 182.1 53.9 ...
 $ Outlet_Establishment_Year    : int  1999 2009 1999 1998 1987 2009 1987 1985 2002 2007 ...
 $ Outlet_Size_Other            : int  0 0 0 1 0 0 0 0 1 1 ...
 $ Outlet_Size_High             : int  0 0 0 0 1 0 1 0 0 0 ...
 $ Outlet_Size_Medium           : int  1 1 1 0 0 1 0 1 0 0 ...
 $ Outlet_Size_Small            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Outlet_Location_Type_Tier 1  : int  1 0 1 0 0 0 0 0 0 0 ...
 $ Outlet_Location_Type_Tier 2  : int  0 0 0 0 0 0 0 0 1 1 ...
 $ Outlet_Location_Type_Tier 3  : int  0 1 0 1 1 1 1 1 0 0 ...
 $ Outlet_Type_Grocery Store    : int  0 0 0 1 0 0 0 0 0 0 ...
 $ Outlet_Type_Supermarket Type1: int  1 0 1 0 1 0 1 0 1 1 ...
 $ Outlet_Type_Supermarket Type2: int  0 1 0 0 0 1 0 0 0 0 ...
 $ Outlet_Type_Supermarket Type3: int  0 0 0 0 0 0 0 1 0 0 ...
 $ Item_Outlet_Sales            : num  3735 443 2097 732 995 ...
 - attr(*, "dummies")=List of 3
  ..$ Outlet_Size         : int  9 10 11 12
  ..$ Outlet_Location_Type: int  13 14 15
  ..$ Outlet_Type         : int  16 17 18 19
```

Hide

```
summary(combi_encoded_dummies_drop)
```

```
  Item_Weight     Item_Fat_Content Item_Visibility      Item_MRP      Outlet_Establishment_Year
 Min.   : 4.555   Min.   :0.0000   Min.   :0.003575   Min.   : 31.29   Min.   :1985             
 1st Qu.: 9.300   1st Qu.:0.0000   1st Qu.:0.033143   1st Qu.: 94.01   1st Qu.:1987             
 Median :12.600   Median :0.0000   Median :0.054023   Median :142.25   Median :1999             
 Mean   :12.760   Mean   :0.3534   Mean   :0.069296   Mean   :141.00   Mean   :1998             
 3rd Qu.:16.000   3rd Qu.:1.0000   3rd Qu.:0.094037   3rd Qu.:185.86   3rd Qu.:2004             
 Max.   :21.350   Max.   :1.0000   Max.   :0.328391   Max.   :266.89   Max.   :2009             
 Outlet_Size_Other Outlet_Size_High Outlet_Size_Medium Outlet_Size_Small Outlet_Location_Type_Ti
er 1
 Min.   :0.0000    Min.   :0.0000   Min.   :0.0000     Min.   :0.0000    Min.   :0.0000         

 1st Qu.:0.0000    1st Qu.:0.0000   1st Qu.:0.0000     1st Qu.:0.0000    1st Qu.:0.0000         

 Median :0.0000    Median :0.0000   Median :0.0000     Median :0.0000    Median :0.0000         

 Mean   :0.2827    Mean   :0.1093   Mean   :0.3277     Mean   :0.2802    Mean   :0.2802         

 3rd Qu.:1.0000    3rd Qu.:0.0000   3rd Qu.:1.0000     3rd Qu.:1.0000    3rd Qu.:1.0000         

 Max.   :1.0000    Max.   :1.0000   Max.   :1.0000     Max.   :1.0000    Max.   :1.0000         

 Outlet_Location_Type_Tier 2 Outlet_Location_Type_Tier 3 Outlet_Type_Grocery Store Outlet_Type_S
upermarket Type1
 Min.   :0.0000             Min.   :0.0000              Min.   :0.0000            Min.   :0.000
0                  
 1st Qu.:0.0000             1st Qu.:0.0000              1st Qu.:0.0000            1st Qu.:0.000
0                  
 Median :0.0000             Median :0.0000              Median :0.0000            Median :1.000
0                  
 Mean   :0.3267             Mean   :0.3931              Mean   :0.1271            Mean   :0.654
3                  
 3rd Qu.:1.0000             3rd Qu.:1.0000              3rd Qu.:0.0000            3rd Qu.:1.000
0                  
 Max.   :1.0000             Max.   :1.0000              Max.   :1.0000            Max.   :1.000
0                  
 Outlet_Type_Supermarket Type2 Outlet_Type_Supermarket Type3 Item_Outlet_Sales
 Min.   :0.0000                Min.   :0.0000                Min.   :    1.0  
 1st Qu.:0.0000                1st Qu.:0.0000                1st Qu.:    1.0  
 Median :0.0000                Median :0.0000                Median :  559.3  
 Mean   :0.1088                Mean   :0.1098                Mean   : 1309.3  
 3rd Qu.:0.0000                3rd Qu.:0.0000                3rd Qu.: 2163.2  
 Max.   :1.0000                Max.   :1.0000                Max.   :13087.0  
```

Dividing data to Train & Test POST Label & Hot Encoding

Hide

```
new_train_combi_encoded_dummies_drop <- combi_encoded_dummies_drop %>% filter(Item_Outlet_Sales
!= 1)
new_test_combi_encoded_dummies_drop <- combi_encoded_dummies_drop %>% filter(Item_Outlet_Sales =
= 1)
str(new_train_combi_encoded_dummies_drop)
```

```
'data.frame':	8523 obs. of  17 variables:
 $ Item_Weight                  : num  9.3 5.92 17.5 19.2 8.93 ...
 $ Item_Fat_Content             : num  0 1 0 1 0 1 1 0 1 1 ...
 $ Item_Visibility              : num  0.016 0.0193 0.0168 0.054 0.054 ...
 $ Item_MRP                     : num  249.8 48.3 141.6 182.1 53.9 ...
 $ Outlet_Establishment_Year    : int  1999 2009 1999 1998 1987 2009 1987 1985 2002 2007 ...
 $ Outlet_Size_Other            : int  0 0 0 1 0 0 0 0 1 1 ...
 $ Outlet_Size_High             : int  0 0 0 0 1 0 1 0 0 0 ...
 $ Outlet_Size_Medium           : int  1 1 1 0 0 1 0 1 0 0 ...
 $ Outlet_Size_Small            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Outlet_Location_Type_Tier 1  : int  1 0 1 0 0 0 0 0 0 0 ...
 $ Outlet_Location_Type_Tier 2  : int  0 0 0 0 0 0 0 0 1 1 ...
 $ Outlet_Location_Type_Tier 3  : int  0 1 0 1 1 1 1 1 0 0 ...
 $ Outlet_Type_Grocery Store    : int  0 0 0 1 0 0 0 0 0 0 ...
 $ Outlet_Type_Supermarket Type1: int  1 0 1 0 1 0 1 0 1 1 ...
 $ Outlet_Type_Supermarket Type2: int  0 1 0 0 0 1 0 0 0 0 ...
 $ Outlet_Type_Supermarket Type3: int  0 0 0 0 0 0 0 1 0 0 ...
 $ Item_Outlet_Sales            : num  3735 443 2097 732 995 ...
 - attr(*, "dummies")=List of 3
  ..$ Outlet_Size         : int  9 10 11 12
  ..$ Outlet_Location_Type: int  13 14 15
  ..$ Outlet_Type         : int  16 17 18 19
```

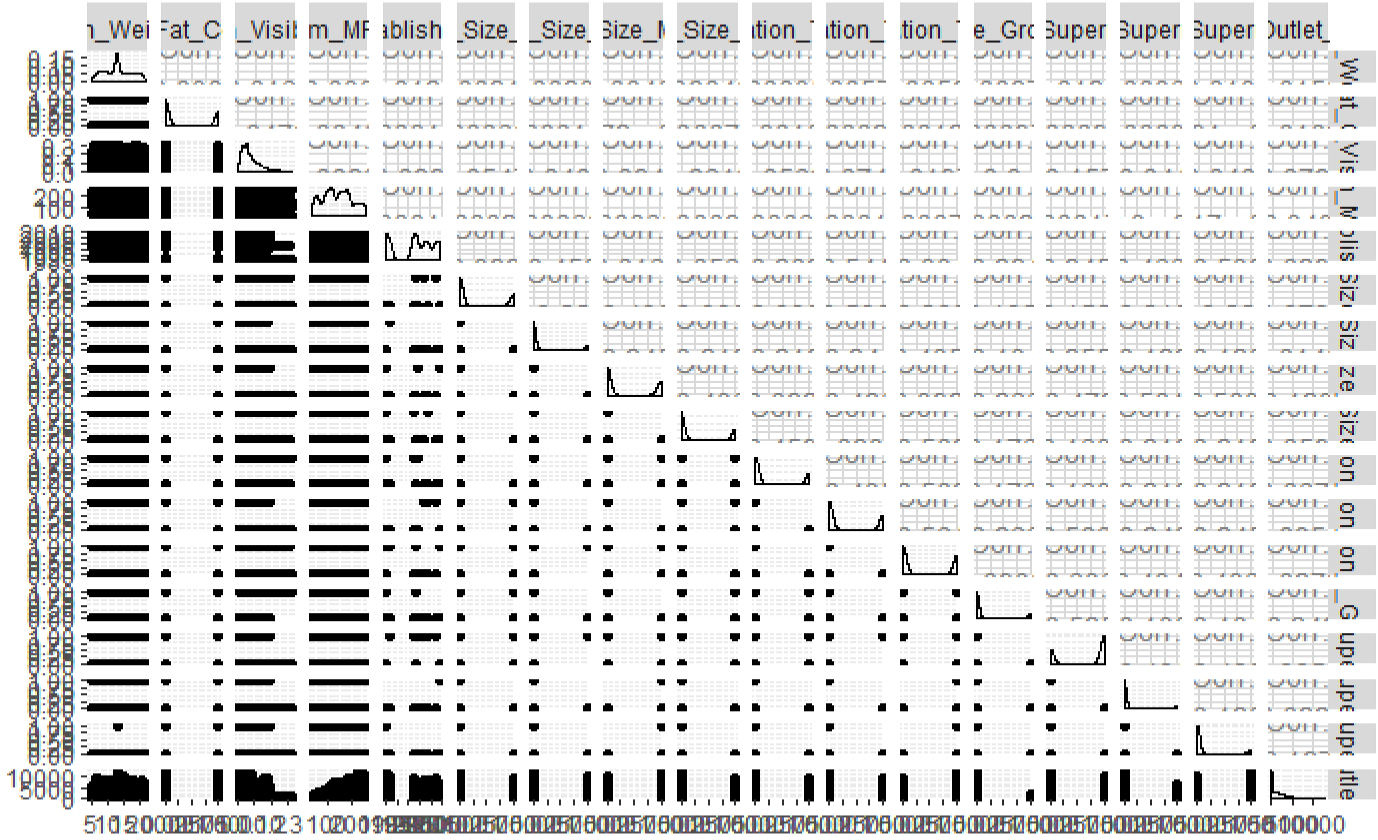
Hide

```
str(new_test_combi_encoded_dummies_drop)
```

```
'data.frame':	5681 obs. of  17 variables:
 $ Item_Weight                  : num  20.75 8.3 14.6 7.32 12.6 ...
 $ Item_Fat_Content             : num  0 1 0 0 1 1 1 0 1 0 ...
 $ Item_Visibility              : num  0.00756 0.03843 0.09957 0.01539 0.1186 ...
 $ Item_MRP                     : num  107.9 87.3 241.8 155 234.2 ...
 $ Outlet_Establishment_Year    : int  1999 2007 1998 2007 1985 1997 2009 1985 2002 2007 ...
 $ Outlet_Size_Other            : int  0 1 1 1 0 0 0 0 1 1 ...
 $ Outlet_Size_High             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Outlet_Size_Medium           : int  1 0 0 0 1 0 1 1 0 0 ...
 $ Outlet_Size_Small            : int  0 0 0 0 0 1 0 0 0 0 ...
 $ Outlet_Location_Type_Tier 1  : int  1 0 0 0 0 1 0 0 0 0 ...
 $ Outlet_Location_Type_Tier 2  : int  0 1 0 1 0 0 0 0 1 1 ...
 $ Outlet_Location_Type_Tier 3  : int  0 0 1 0 1 0 1 1 0 0 ...
 $ Outlet_Type_Grocery Store    : int  0 0 1 0 0 0 0 0 0 0 ...
 $ Outlet_Type_Supermarket Type1: int  1 1 0 1 0 1 0 0 1 1 ...
 $ Outlet_Type_Supermarket Type2: int  0 0 0 0 0 0 1 0 0 0 ...
 $ Outlet_Type_Supermarket Type3: int  0 0 0 0 1 0 0 1 0 0 ...
 $ Item_Outlet_Sales            : num  1 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "dummies")=List of 3
  ..$ Outlet_Size         : int  9 10 11 12
  ..$ Outlet_Location_Type: int  13 14 15
  ..$ Outlet_Type         : int  16 17 18 19
```

Hide

```
library(GGally)
ggpairs(combi_encoded_dummies_drop)
```



Linear (Multiple) Regression

Amount of correlation present in our predictor variables

Hide

```
cor(new_train_combi_encoded_dummies_drop)
```

```
                              Item_Weight Item_Fat_Content Item_Visibility      Item_MRP
Item_Weight                   1.000000000    -0.0210920104    -0.018772588  0.0249505601
Item_Fat_Content             -0.021092010     1.0000000000     0.049793752  0.0060628994
Item_Visibility              -0.018772588     0.0497937522     1.000000000 -0.0045367831
Item_MRP                      0.024950560     0.0060628994    -0.004536783  1.0000000000
Outlet_Establishment_Year     0.007739014     0.0031506634    -0.078272866  0.0050199162
Outlet_Size_Other            -0.005190112    -0.0010847801     0.051641561 -0.0067540030
Outlet_Size_High              0.015976179    -0.0021320419    -0.043643327  0.0024375785
Outlet_Size_Medium           -0.002790703     0.0046714179    -0.083994282 -0.0045100820
Outlet_Size_Small            -0.002980744    -0.0023130394     0.066331288  0.0097927692
Outlet_Location_Type_Tier 1   0.002083178     0.0031548795     0.063767534 -0.0012290862
Outlet_Location_Type_Tier 2  -0.007382332    -0.0032717388    -0.073390251  0.0019513073
Outlet_Location_Type_Tier 3   0.005173738     0.0002410231     0.011843986 -0.0007437168
Outlet_Type_Grocery Store    -0.004778323    -0.0029242966     0.299204770 -0.0042771353
Outlet_Type_Supermarket Type1  0.011747051     0.0005332351    -0.152830781  0.0048854837
Outlet_Type_Supermarket Type2  0.004955601     0.0021294095    -0.033373806  0.0038499211
Outlet_Type_Supermarket Type3 -0.017723313     0.0001825365    -0.053023468 -0.0067136994
Item_Outlet_Sales             0.009692876     0.0187185336    -0.134097091  0.5675744467
                             Outlet_Establishment_Year Outlet_Size_Other Outlet_Size_High Outle
t_Size_Medium
Item_Weight                                 0.007739014      -0.005190112      0.015976179
-0.002790703
Item_Fat_Content                            0.003150663      -0.001084780     -0.002132042
  0.004671418
Item_Visibility                            -0.078272866       0.051641561     -0.043643327
-0.083994282
Item_MRP                                    0.005019916      -0.006754003      0.002437579
-0.004510082
Outlet_Establishment_Year                   1.000000000       0.387635656     -0.453388454
-0.016345705
Outlet_Size_Other                           0.387635656       1.000000000     -0.220008664
-0.438368642
Outlet_Size_High                           -0.453388454      -0.220008664      1.000000000
-0.244633888
Outlet_Size_Medium                         -0.016345705      -0.438368642     -0.244633888
  1.000000000
Outlet_Size_Small                          -0.056566813      -0.391733940     -0.218609151
-0.435580104
Outlet_Location_Type_Tier 1                -0.201690130      -0.391733940     -0.218609151
  0.082072274
Outlet_Location_Type_Tier 2                 0.540819608       0.592969531     -0.244112933
-0.486396549
Outlet_Location_Type_Tier 3                -0.333894725      -0.209236546      0.435418920
  0.391616506
Outlet_Type_Grocery Store                  -0.281195730       0.194602128     -0.133686090
-0.266370372
Outlet_Type_Supermarket Type1               0.245069762       0.152307648      0.254668077
-0.471782330
Outlet_Type_Supermarket Type2               0.466336465      -0.219478216     -0.122480954
  0.500670430
Outlet_Type_Supermarket Type3              -0.538072347      -0.220406028     -0.122998724
  0.502786939
Item_Outlet_Sales                          -0.049134970      -0.131973256      0.024170053
```

```
  0.204701320
                              Outlet_Size_Small Outlet_Location_Type_Tier 1 Outlet_Location_Type
_Tier 2
Item_Weight                        -0.002980744                 0.002083178                -0.00
7382332
Item_Fat_Content                   -0.002313039                 0.003154879                -0.00
3271739
Item_Visibility                     0.066331288                 0.063767534                -0.07
3390251
Item_MRP                            0.009792769                -0.001229086                 0.00
1951307
Outlet_Establishment_Year          -0.056566813                -0.201690130                 0.54
0819608
Outlet_Size_Other                  -0.391733940                -0.391733940                 0.59
2969531
Outlet_Size_High                   -0.218609151                -0.218609151                -0.24
4112933
Outlet_Size_Medium                 -0.435580104                 0.082072274                -0.48
6396549
Outlet_Size_Small                   1.000000000                 0.458963522                 0.08
3381305
Outlet_Location_Type_Tier 1         0.458963522                 1.000000000                -0.43
4652524
Outlet_Location_Type_Tier 2         0.083381305                -0.434652524                 1.00
0000000
Outlet_Location_Type_Tier 3        -0.502066266                -0.502066266                -0.56
0639241
Outlet_Type_Grocery Store           0.176158327                 0.176158327                -0.26
5803129
Outlet_Type_Supermarket Type1       0.163388083                 0.163388083                 0.50
6347158
Outlet_Type_Supermarket Type2      -0.218082078                -0.218082078                -0.24
3524369
Outlet_Type_Supermarket Type3      -0.219003987                -0.219003987                -0.24
4553832
Item_Outlet_Sales                  -0.098402699                -0.111287125                 0.05
8261357
                              Outlet_Location_Type_Tier 3 Outlet_Type_Grocery Store Outlet_Type_
Supermarket Type1
Item_Weight                                  0.0051737383              -0.004778323
     0.0117470508
Item_Fat_Content                             0.0002410231              -0.002924297
     0.0005332351
Item_Visibility                              0.0118439862               0.299204770
    -0.1528307806
Item_MRP                                    -0.0007437168              -0.004277135
     0.0048854837
Outlet_Establishment_Year                   -0.3338947248              -0.281195730
     0.2450697621
Outlet_Size_Other                           -0.2092365463               0.194602128
     0.1523076482
Outlet_Size_High                             0.4354189199              -0.133686090
     0.2546680773
Outlet_Size_Medium                           0.3916165060              -0.266370372
```

```
    -0.4717823295
Outlet_Size_Small                           -0.5020662661               0.176158327
     0.1633880834
Outlet_Location_Type_Tier 1                 -0.5020662661               0.176158327
     0.1633880834
Outlet_Location_Type_Tier 2                 -0.5606392409              -0.265803129
     0.5063471581
Outlet_Location_Type_Tier 3                  1.0000000000               0.093276443
    -0.6364646571
Outlet_Type_Grocery Store                    0.0932764434               1.000000000
    -0.5249424714
Outlet_Type_Supermarket Type1               -0.6364646571              -0.524942471
     1.0000000000
Outlet_Type_Supermarket Type2                0.4343691117              -0.133363769
    -0.4809434894
Outlet_Type_Supermarket Type3                0.4362053421              -0.133927544
    -0.4829766060
Item_Outlet_Sales                            0.0463761913              -0.411727080
     0.1087652555
                              Outlet_Type_Supermarket Type2 Outlet_Type_Supermarket Type3 Item_O
utlet_Sales
Item_Weight                                     0.004955601                 -0.0177233134
0.009692876
Item_Fat_Content                                0.002129410                  0.0001825365
0.018718534
Item_Visibility                                -0.033373806                 -0.0530234675      -
0.134097091
Item_MRP                                        0.003849921                 -0.0067136994
0.567574447
Outlet_Establishment_Year                       0.466336465                 -0.5380723466      -
0.049134970
Outlet_Size_Other                              -0.219478216                 -0.2204060278      -
0.131973256
Outlet_Size_High                               -0.122480954                 -0.1229987236
0.024170053
Outlet_Size_Medium                              0.500670430                  0.5027869391
0.204701320
Outlet_Size_Small                              -0.218082078                 -0.2190039874      -
0.098402699
Outlet_Location_Type_Tier 1                    -0.218082078                 -0.2190039874      -
0.111287125
Outlet_Location_Type_Tier 2                    -0.243524369                 -0.2445538319
0.058261357
Outlet_Location_Type_Tier 3                     0.434369112                  0.4362053421
0.046376191
Outlet_Type_Grocery Store                      -0.133363769                 -0.1339275441      -
0.411727080
Outlet_Type_Supermarket Type1                  -0.480943489                 -0.4829766060
0.108765256
Outlet_Type_Supermarket Type2                   1.000000000                 -0.1227021700      -
0.038058540
Outlet_Type_Supermarket Type3                  -0.122702170                  1.0000000000
0.311192046
```

```
Item_Outlet_Sales                                  -0.038058540                0.3111920462
1.000000000
```

Hide

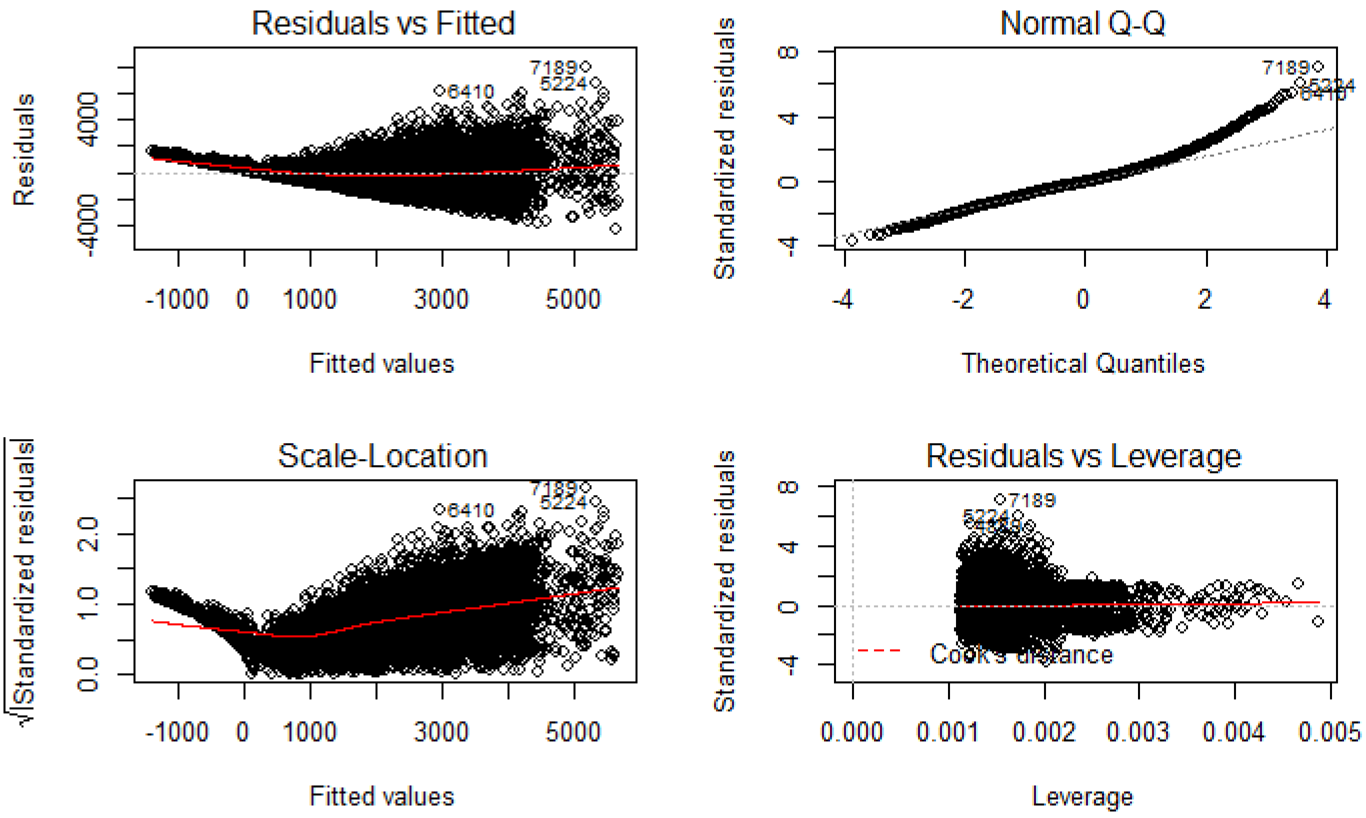
```
summary(linear_model)
```

```


Call:
lm(formula = Item_Outlet_Sales ~ ., data = new_train_combi_encoded_dummies_drop)

Residuals:
    Min      1Q  Median      3Q     Max 
-4313.0  -675.6   -87.6   571.3  7917.6 

Coefficients: (3 not defined because of singularities)
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     -6.666e+04  2.078e+04  -3.207  0.00134 ** 
Item_Weight                     -6.276e-01  2.895e+00  -0.217  0.82837    
Item_Fat_Content                 5.151e+01  2.562e+01   2.010  0.04443 *  
Item_Visibility                 -2.410e+02  2.626e+02  -0.918  0.35880    
Item_MRP                         1.556e+01  1.964e-01  79.232  < 2e-16 ***
Outlet_Establishment_Year        3.433e+01  1.048e+01   3.276  0.00106 ** 
Outlet_Size_Other               -1.440e+02  4.564e+01  -3.156  0.00161 ** 
Outlet_Size_High                 6.953e+02  2.552e+02   2.725  0.00645 ** 
Outlet_Size_Medium               2.911e+01  5.637e+01   0.516  0.60556    
Outlet_Size_Small                       NA         NA      NA       NA    
`Outlet_Location_Type_Tier 1`    3.207e+02  1.547e+02   2.073  0.03824 *  
`Outlet_Location_Type_Tier 2`    2.253e+02  1.005e+02   2.242  0.02496 *  
`Outlet_Location_Type_Tier 3`           NA         NA      NA       NA    
`Outlet_Type_Grocery Store`     -3.633e+03  1.779e+02 -20.415  < 2e-16 ***
`Outlet_Type_Supermarket Type1` -2.154e+03  2.944e+02  -7.319 2.73e-13 ***
`Outlet_Type_Supermarket Type2` -2.551e+03  2.569e+02  -9.930  < 2e-16 ***
`Outlet_Type_Supermarket Type3`         NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1128 on 8509 degrees of freedom
Multiple R-squared:  0.5635,    Adjusted R-squared:  0.5628 
F-statistic: 844.8 on 13 and 8509 DF,  p-value: < 2.2e-16
```

Using Regression plot:

Hide

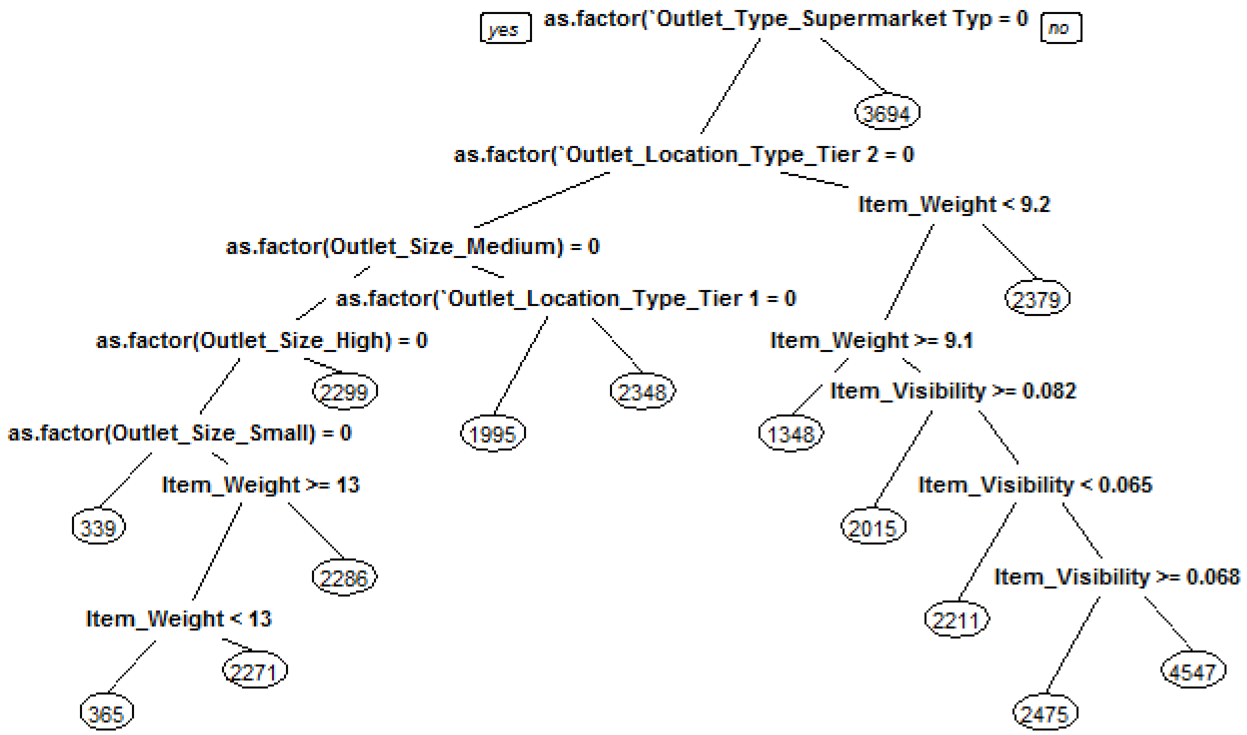
```
par(mfrow=c(2,2))
plot(linear_model)
```



Random Forest

Hide

```
#load randomForest library
library(randomForest)
library(rpart)
library(e1071)
library(rpart.plot)
library(caret)
formula_tree <- as.formula(Item_Outlet_Sales ~ Item_Weight +
Item_Visibility +
as.factor(Outlet_Size_High) +
as.factor(Outlet_Size_Medium) +
as.factor(Outlet_Size_Small) +
as.factor(`Outlet_Location_Type_Tier 1`) +
as.factor(`Outlet_Location_Type_Tier 2`) +
as.factor(`Outlet_Location_Type_Tier 3`) +
as.factor(`Outlet_Type_Supermarket Type3`))
tree1 <- rpart(formula_tree, data = new_train_combi_encoded_dummies_drop, control = rpart.contro
l(cp=0.001))
prp(tree1)
```


yes
as.factor(`Outlet_Type_Supermarket Typ = 0
no
3694
as.factor(`Outlet_Location_Type_Tier 2 = 0
Item_Weight < 9.2
as.factor(Outlet_Size_Medium) = 0
2379
as.factor(`Outlet_Location_Type_Tier 1 = 0
as.factor(Outlet_Size_High) = 0
Item_Weight >= 9.1
2299
1995
2348
1348
Item_Visibility >= 0.082
as.factor(Outlet_Size_Small) = 0
Item_Weight >= 13
339
2015
Item_Visibility < 0.065
2286
Item_Weight < 13
2211
Item_Visibility >= 0.068
2271
365
4547
2475

Hide

```
summary(tree1)
```

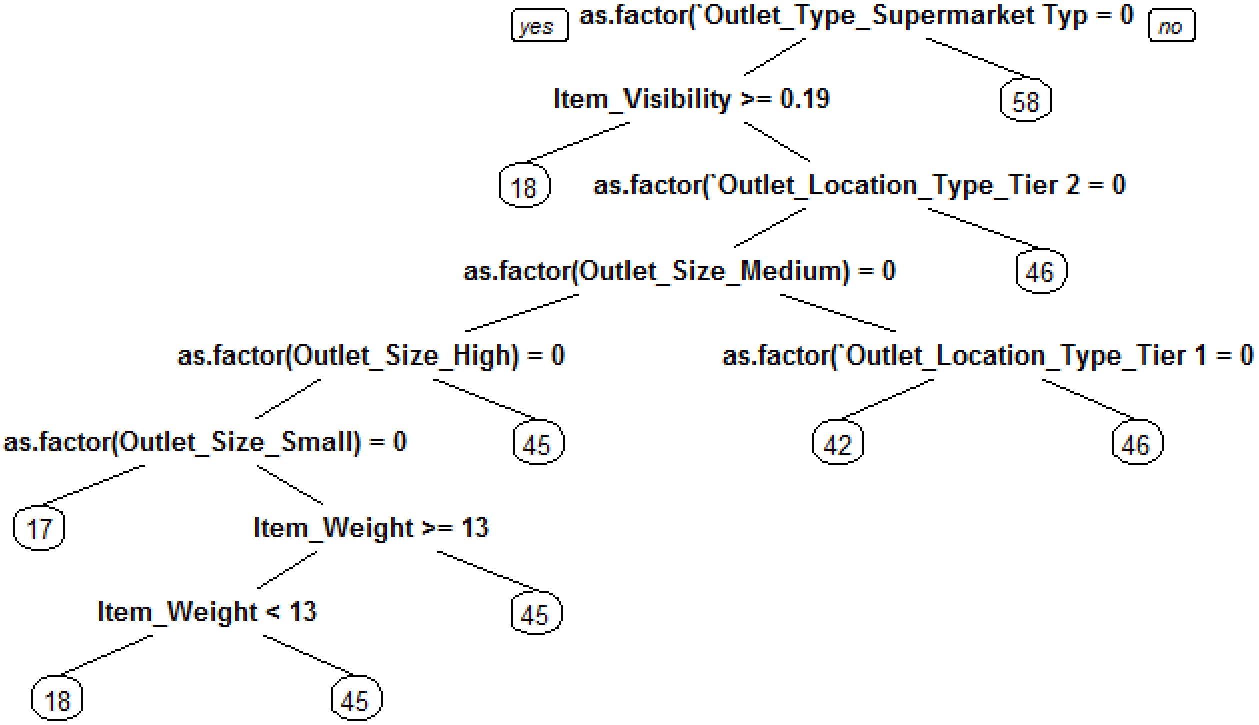
```
Call:
rpart(formula = formula_tree, data = new_train_combi_encoded_dummies_drop,
    control = rpart.control(cp = 0.001))
  n= 8523

           CP nsplit rel error    xerror       xstd
1 0.096840490      0 1.0000000 1.0002935 0.02059661
2 0.021597713      1 0.9031595 0.9037103 0.01738674
3 0.002330376      7 0.7636181 0.7651574 0.01617594
4 0.001007738      8 0.7612877 0.7650304 0.01618859
5 0.001000000     13 0.7562490 0.7712937 0.01629126

Variable importance
as.factor(`Outlet_Type_Supermarket Type3`)                               Item_Weight
                                        33                                        17
              as.factor(Outlet_Size_Small)                as.factor(Outlet_Size_High)
                                        13                                        10
  as.factor(`Outlet_Location_Type_Tier 2`)   as.factor(`Outlet_Location_Type_Tier 1`)
                                         6                                         6
             as.factor(Outlet_Size_Medium)   as.factor(`Outlet_Location_Type_Tier 3`)
                                         6                                         5
                           Item_Visibility
                                         4

Node number 1: 8523 observations,    complexity param=0.09684049
  mean=2181.289, MSE=2911799
  left son=2 (7588 obs) right son=3 (935 obs)
  Primary splits:
      as.factor(`Outlet_Type_Supermarket Type3`) splits as  LR, improve=0.096840490, (0 missing)
      as.factor(Outlet_Size_Medium)              splits as  LR, improve=0.041902630, (0 missing)
      Item_Visibility                            < 0.1876999   to the right, improve=0.02274282
0, (0 missing)
      as.factor(`Outlet_Location_Type_Tier 1`)   splits as  RL, improve=0.012384820, (0 missing)
      as.factor(Outlet_Size_Small)               splits as  RL, improve=0.009683091, (0 missing)

Node number 2: 7588 observations,    complexity param=0.02159771
  mean=1994.887, MSE=2396599
  left son=4 (4803 obs) right son=5 (2785 obs)
  Primary splits:
      as.factor(`Outlet_Location_Type_Tier 2`) splits as  LR, improve=0.026204960, (0 missing)
      Item_Visibility                          < 0.1876999   to the right, improve=0.025033390,
(0 missing)
      as.factor(`Outlet_Location_Type_Tier 3`) splits as  RL, improve=0.013460230, (0 missing)
      Item_Weight                              < 12.625      to the left,  improve=0.008887997,
(0 missing)
      as.factor(Outlet_Size_High)              splits as  LR, improve=0.005403372, (0 missing)
  Surrogate splits:
      as.factor(`Outlet_Location_Type_Tier 3`) splits as  RL, agree=0.685, adj=0.143, (0 split)
      as.factor(`Outlet_Location_Type_Tier 1`) splits as  RL, agree=0.682, adj=0.133, (0 split)

Node number 3: 935 observations
  mean=3694.039, MSE=4522521
```

```
Node number 4: 4803 observations,    complexity param=0.02159771
  mean=1804.057, MSE=2346885 
  left son=8 (2945 obs) right son=9 (1858 obs)
  Primary splits:
      as.factor(Outlet_Size_Medium) splits as  LR, improve=0.036417060, (0 missing)
      Item_Visibility               < 0.1871346   to the right, improve=0.032668820, (0 missing)
      as.factor(Outlet_Size_High)   splits as  LR, improve=0.025130610, (0 missing)
      Item_Weight                   < 12.725      to the left,  improve=0.014730420, (0 missing)
      as.factor(Outlet_Size_Small)  splits as  RL, improve=0.009643201, (0 missing)
  Surrogate splits:
      as.factor(Outlet_Size_Small) splits as  RL, agree=0.690, adj=0.200, (0 split)
      Item_Visibility              < 0.004418756 to the right, agree=0.613, adj=0.001, (0 split)

Node number 5: 2785 observations,    complexity param=0.001007738
  mean=2323.991, MSE=2311222 
  left son=10 (816 obs) right son=11 (1969 obs)
  Primary splits:
      Item_Weight                  < 9.24        to the left,  improve=0.003134485, (0 missing)
      as.factor(Outlet_Size_Small) splits as  LR, improve=0.002861338, (0 missing)
      Item_Visibility              < 0.02369722  to the right, improve=0.002025882, (0 missing)
  Surrogate splits:
      Item_Visibility < 0.006036989 to the left,  agree=0.71, adj=0.009, (0 split)

Node number 8: 2945 observations,    complexity param=0.02159771
  mean=1571.848, MSE=2350193 
  left son=16 (2013 obs) right son=17 (932 obs)
  Primary splits:
      as.factor(Outlet_Size_High)              splits as  LR, improve=1.041629e-01, (0 missing)
      Item_Visibility                          < 0.1478574   to the right, improve=4.266678e-02, 
(0 missing)
      Item_Weight                              < 12.625      to the left,  improve=2.770336e-02, 
(0 missing)
      as.factor(Outlet_Size_Small)             splits as  LR, improve=7.875173e-06, (0 missing)
      as.factor(`Outlet_Location_Type_Tier 1`) splits as  LR, improve=7.875173e-06, (0 missing)
  Surrogate splits:
      as.factor(Outlet_Size_Small)             splits as  RL, agree=0.812, adj=0.405, (0 split)
      as.factor(`Outlet_Location_Type_Tier 1`) splits as  RL, agree=0.812, adj=0.405, (0 split)
      as.factor(`Outlet_Location_Type_Tier 3`) splits as  LR, agree=0.812, adj=0.405, (0 split)
      Item_Visibility                          < 0.008760024 to the right, agree=0.687, adj=0.01
2, (0 split)

Node number 9: 1858 observations,    complexity param=0.002330376
  mean=2172.117, MSE=2120707 
  left son=18 (928 obs) right son=19 (930 obs)
  Primary splits:
      as.factor(`Outlet_Location_Type_Tier 1`) splits as  LR, improve=0.0146775500, (0 missing)
      as.factor(`Outlet_Location_Type_Tier 3`) splits as  RL, improve=0.0146775500, (0 missing)
      Item_Visibility                          < 0.1599065   to the right, improve=0.0021485360, 
(0 missing)
      Item_Weight                              < 21.225      to the left,  improve=0.0007303589, 
(0 missing)
  Surrogate splits:
      Item_Weight     < 8.305       to the left,  agree=0.513, adj=0.026, (0 split)
      Item_Visibility < 0.0418986   to the right, agree=0.512, adj=0.023, (0 split)
```

```
Node number 10: 816 observations,    complexity param=0.001007738
  mean=2191.775, MSE=1960975 
  left son=20 (25 obs) right son=21 (791 obs)
  Primary splits:
      Item_Weight                  < 9.1025      to the right, improve=0.011483350, (0 missing)
      Item_Visibility              < 0.0819964   to the right, improve=0.008333672, (0 missing)
      as.factor(Outlet_Size_Small) splits as  LR, improve=0.003297263, (0 missing)

Node number 11: 1969 observations
  mean=2378.784, MSE=2446127 

Node number 16: 2013 observations,    complexity param=0.02159771
  mean=1235.186, MSE=1992511 
  left son=32 (555 obs) right son=33 (1458 obs)
  Primary splits:
      as.factor(Outlet_Size_Small)             splits as  LR, improve=0.15331690, (0 missing)
      as.factor(`Outlet_Location_Type_Tier 1`) splits as  LR, improve=0.15331690, (0 missing)
      as.factor(`Outlet_Location_Type_Tier 3`) splits as  RL, improve=0.15331690, (0 missing)
      Item_Visibility                          < 0.1053019   to the right, improve=0.04639051, 
(0 missing)
      Item_Weight                              < 12.55       to the right, improve=0.02963642, 
(0 missing)
  Surrogate splits:
      Item_Weight < 21.3        to the right, agree=0.725, adj=0.002, (0 split)

Node number 17: 932 observations
  mean=2298.995, MSE=2349196 

Node number 18: 928 observations
  mean=1995.499, MSE=1891151 

Node number 19: 930 observations
  mean=2348.355, MSE=2287583 

Node number 20: 25 observations
  mean=1347.686, MSE=467144.2 

Node number 21: 791 observations,    complexity param=0.001007738
  mean=2218.453, MSE=1984958 
  left son=42 (245 obs) right son=43 (546 obs)
  Primary splits:
      Item_Visibility              < 0.0819964   to the right, improve=0.009339677, (0 missing)
      Item_Weight                  < 5.4625      to the right, improve=0.004727454, (0 missing)
      as.factor(Outlet_Size_Small) splits as  LR, improve=0.003139342, (0 missing)
  Surrogate splits:
      Item_Weight < 4.795       to the left,  agree=0.702, adj=0.037, (0 split)

Node number 32: 555 observations
  mean=339.3517, MSE=73316.71 

Node number 33: 1458 observations,    complexity param=0.02159771
  mean=1576.193, MSE=2301297 
  left son=66 (999 obs) right son=67 (459 obs)
  Primary splits:
      Item_Weight     < 12.55       to the right, improve=0.1005474, (0 missing)
      Item_Visibility < 0.1475278   to the right, improve=0.0502919, (0 missing)
  Surrogate splits:
      Item_Visibility < 0.01167985  to the right, agree=0.69, adj=0.015, (0 split)

Node number 42: 245 observations
  mean=2015.192, MSE=1690564

Node number 43: 546 observations,    complexity param=0.001007738
  mean=2309.66, MSE=2090200
  left son=86 (444 obs) right son=87 (102 obs)
  Primary splits:
      Item_Visibility              < 0.06471562  to the left,  improve=0.020248240, (0 missing)
      Item_Weight                  < 6.105       to the right, improve=0.003617898, (0 missing)
      as.factor(Outlet_Size_Small) splits as  LR, improve=0.002980943, (0 missing)
  Surrogate splits:
      Item_Weight < 9.05        to the left,  agree=0.815, adj=0.01, (0 split)

Node number 66: 999 observations,    complexity param=0.02159771
  mean=1250.134, MSE=1991107
  left son=132 (535 obs) right son=133 (464 obs)
  Primary splits:
      Item_Weight     < 12.625      to the left,  improve=0.45383950, (0 missing)
      Item_Visibility < 0.1054815   to the right, improve=0.05700259, (0 missing)
  Surrogate splits:
      Item_Visibility < 0.05747654  to the right, agree=0.64, adj=0.224, (0 split)

Node number 67: 459 observations
  mean=2285.85, MSE=2241414

Node number 86: 444 observations
  mean=2211.056, MSE=1716776

Node number 87: 102 observations,    complexity param=0.001007738
  mean=2738.879, MSE=3489142
  left son=174 (89 obs) right son=175 (13 obs)
  Primary splits:
      Item_Visibility              < 0.06799998  to the right, improve=0.13690350, (0 missing)
      Item_Weight                  < 5.7575      to the right, improve=0.02053833, (0 missing)
      as.factor(Outlet_Size_Small) splits as  LR, improve=0.01802642, (0 missing)

Node number 132: 535 observations
  mean=364.8547, MSE=131842.8

Node number 133: 464 observations
  mean=2270.877, MSE=2189312

Node number 174: 89 observations
  mean=2474.734, MSE=2568266

Node number 175: 13 observations
  mean=4547.26, MSE=6045685
```

Hide

```
formula_sqrt_tree <- as.formula(sqrt(Item_Outlet_Sales) ~ Item_Weight +
Item_Visibility +
as.factor(Outlet_Size_High) +
as.factor(Outlet_Size_Medium) +
as.factor(Outlet_Size_Small) +
as.factor(`Outlet_Location_Type_Tier 1`) +
as.factor(`Outlet_Location_Type_Tier 2`) +
as.factor(`Outlet_Location_Type_Tier 3`) +
as.factor(`Outlet_Type_Supermarket Type3`))
tree2 <- rpart(formula_sqrt_tree, data = new_train_combi_encoded_dummies_drop, control = rpart.c
ontrol(cp=0.001))
prp(tree2)
```


yes
as.factor(`Outlet_Type_Supermarket Typ = 0
no
Item_Visibility >= 0.19
58
18
as.factor(`Outlet_Location_Type_Tier 2 = 0
as.factor(Outlet_Size_Medium) = 0
46
as.factor(Outlet_Size_High) = 0
as.factor(`Outlet_Location_Type_Tier 1 = 0
as.factor(Outlet_Size_Small) = 0
45
42
46
17
Item_Weight >= 13
Item_Weight < 13
45
18
45

Hide

```
summary(tree2)
```

```
Call:
rpart(formula = formula_sqrt_tree, data = new_train_combi_encoded_dummies_drop, 
    control = rpart.control(cp = 0.001))
  n= 8523 

           CP nsplit rel error    xerror        xstd
1 0.084734551      0 1.0000000 1.0000587 0.013488517
2 0.033643184      1 0.9152654 0.9155306 0.011935872
3 0.002451204      8 0.6670346 0.6681327 0.009953933
4 0.001000000      9 0.6645834 0.6658112 0.009917513

Variable importance
as.factor(`Outlet_Type_Supermarket Type3`)                                Item_Weight 
                                        20                                         19 
              as.factor(Outlet_Size_Small)                            Item_Visibility 
                                        17                                         11 
               as.factor(Outlet_Size_High)   as.factor(`Outlet_Location_Type_Tier 1`) 
                                        10                                          6 
  as.factor(`Outlet_Location_Type_Tier 2`)              as.factor(Outlet_Size_Medium) 
                                         6                                          6 
  as.factor(`Outlet_Location_Type_Tier 3`) 
                                         6 

Node number 1: 8523 observations,    complexity param=0.08473455
  mean=42.94478, MSE=337.035 
  left son=2 (7588 obs) right son=3 (935 obs)
  Primary splits:
      as.factor(`Outlet_Type_Supermarket Type3`) splits as  LR, improve=0.08473455, (0 missing)
      as.factor(Outlet_Size_Medium)              splits as  LR, improve=0.04830326, (0 missing)
      Item_Visibility                            < 0.1876999   to the right, improve=0.03812571, 
(0 missing)
      as.factor(`Outlet_Location_Type_Tier 1`)   splits as  RL, improve=0.01521102, (0 missing)
      as.factor(Outlet_Size_Small)               splits as  RL, improve=0.01226258, (0 missing)

Node number 2: 7588 observations,    complexity param=0.03364318
  mean=41.06888, MSE=308.234 
  left son=4 (169 obs) right son=5 (7419 obs)
  Primary splits:
      Item_Visibility                          < 0.1876999   to the right, improve=0.040214080, 
(0 missing)
      as.factor(`Outlet_Location_Type_Tier 2`) splits as  LR, improve=0.039407540, (0 missing)
      as.factor(`Outlet_Location_Type_Tier 3`) splits as  RL, improve=0.019028330, (0 missing)
      Item_Weight                              < 12.625      to the left,  improve=0.012605810, 
(0 missing)
      as.factor(Outlet_Size_Medium)            splits as  LR, improve=0.008860972, (0 missing)

Node number 3: 935 observations
  mean=58.16867, MSE=310.4447 

Node number 4: 169 observations
  mean=17.74189, MSE=57.2304 

Node number 5: 7419 observations,    complexity param=0.03364318
```

```
  mean=41.60025, MSE=301.274 
  left son=10 (4636 obs) right son=11 (2783 obs)
  Primary splits:
      as.factor(`Outlet_Location_Type_Tier 2`) splits as  LR, improve=0.032631850, (0 missing)
      as.factor(`Outlet_Location_Type_Tier 3`) splits as  RL, improve=0.017205290, (0 missing)
      Item_Weight                              < 12.625      to the left,  improve=0.009803359,
(0 missing)
      Item_Visibility                          < 0.09709373  to the right, improve=0.008435584,
(0 missing)
      as.factor(Outlet_Size_Medium)            splits as  LR, improve=0.006199232, (0 missing)
  Surrogate splits:
      as.factor(`Outlet_Location_Type_Tier 3`) splits as  RL, agree=0.690, adj=0.173, (0 split)
      as.factor(`Outlet_Location_Type_Tier 1`) splits as  RL, agree=0.685, adj=0.161, (0 split)
      as.factor(Outlet_Size_Medium)            splits as  RL, agree=0.625, adj=0.001, (0 split)
      Item_Weight                              < 4.795       to the right, agree=0.625, adj=0.00
0, (0 split)

Node number 10: 4636 observations,    complexity param=0.03364318
  mean=39.17092, MSE=322.0103 
  left son=20 (2780 obs) right son=21 (1856 obs)
  Primary splits:
      as.factor(Outlet_Size_Medium) splits as  LR, improve=0.04767614, (0 missing)
      as.factor(Outlet_Size_High)   splits as  LR, improve=0.02812387, (0 missing)
      Item_Weight                   < 12.625      to the left,  improve=0.01683184, (0 missing)
      Item_Visibility               < 0.09722637  to the right, improve=0.01237150, (0 missing)
      as.factor(Outlet_Size_Small)  splits as  RL, improve=0.01158969, (0 missing)
  Surrogate splits:
      as.factor(Outlet_Size_Small) splits as  RL, agree=0.697, adj=0.242, (0 split)
      as.factor(Outlet_Size_High)  splits as  RL, agree=0.601, adj=0.004, (0 split)
      Item_Visibility              < 0.004418756 to the right, agree=0.600, adj=0.001, (0 split)

Node number 11: 2783 observations
  mean=45.6471, MSE=240.5229 

Node number 20: 2780 observations,    complexity param=0.03364318
  mean=35.96943, MSE=351.9025 
  left son=40 (1848 obs) right son=41 (932 obs)
  Primary splits:
      as.factor(Outlet_Size_High)              splits as  LR, improve=0.1213224000, (0 missing)
      Item_Weight                              < 12.625      to the left,  improve=0.0302976900,
(0 missing)
      Item_Visibility                          < 0.09764521  to the right, improve=0.0202096200,
(0 missing)
      as.factor(Outlet_Size_Small)             splits as  LR, improve=0.0001383401, (0 missing)
      as.factor(`Outlet_Location_Type_Tier 1`) splits as  LR, improve=0.0001383401, (0 missing)
  Surrogate splits:
      as.factor(Outlet_Size_Small)             splits as  RL, agree=0.829, adj=0.490, (0 split)
      as.factor(`Outlet_Location_Type_Tier 1`) splits as  RL, agree=0.829, adj=0.490, (0 split)
      as.factor(`Outlet_Location_Type_Tier 3`) splits as  LR, agree=0.829, adj=0.490, (0 split)
      Item_Visibility                          < 0.008760024 to the right, agree=0.669, adj=0.01
2, (0 split)
      Item_Weight                              < 4.9         to the right, agree=0.665, adj=0.00
1, (0 split)
```

```
Node number 21: 1856 observations,    complexity param=0.002451204
  mean=43.96625, MSE=238.8889 
  left son=42 (927 obs) right son=43 (929 obs)
  Primary splits:
      as.factor(`Outlet_Location_Type_Tier 1`) splits as  LR, improve=0.015880820, (0 missing)
      as.factor(`Outlet_Location_Type_Tier 3`) splits as  RL, improve=0.015880820, (0 missing)
      Item_Visibility                          < 0.1599065   to the right, improve=0.001985009, 
(0 missing)
      Item_Weight                              < 16.925      to the left,  improve=0.001082007, 
(0 missing)
  Surrogate splits:
      Item_Weight     < 8.305       to the left,  agree=0.513, adj=0.026, (0 split)
      Item_Visibility < 0.0418986   to the right, agree=0.512, adj=0.023, (0 split)

Node number 40: 1848 observations,    complexity param=0.03364318
  mean=31.32921, MSE=334.7085 
  left son=80 (475 obs) right son=81 (1373 obs)
  Primary splits:
      as.factor(Outlet_Size_Small)             splits as  LR, improve=0.20427800, (0 missing)
      as.factor(`Outlet_Location_Type_Tier 1`) splits as  LR, improve=0.20427800, (0 missing)
      as.factor(`Outlet_Location_Type_Tier 3`) splits as  RL, improve=0.20427800, (0 missing)
      Item_Weight                              < 12.55       to the right, improve=0.04010251, 
(0 missing)
      Item_Visibility                          < 0.09722637  to the right, improve=0.02612810, 
(0 missing)
  Surrogate splits:
      Item_Weight < 21.3        to the right, agree=0.744, adj=0.002, (0 split)

Node number 41: 932 observations
  mean=45.17021, MSE=258.6473 

Node number 42: 927 observations
  mean=42.0164, MSE=229.8928 

Node number 43: 929 observations
  mean=45.91191, MSE=240.2863 

Node number 80: 475 observations
  mean=17.27092, MSE=45.40423 

Node number 81: 1373 observations,    complexity param=0.03364318
  mean=36.19279, MSE=342.7676 
  left son=162 (914 obs) right son=163 (459 obs)
  Primary splits:
      Item_Weight     < 12.55       to the right, improve=0.12279120, (0 missing)
      Item_Visibility < 0.1056321   to the right, improve=0.03047616, (0 missing)
  Surrogate splits:
      Item_Visibility < 0.01167985  to the right, agree=0.671, adj=0.015, (0 split)

Node number 162: 914 observations,    complexity param=0.03364318
  mean=31.59534, MSE=336.4522 
  left son=324 (450 obs) right son=325 (464 obs)
  Primary splits:
      Item_Weight     < 12.625      to the left,  improve=0.55950920, (0 missing)
      Item_Visibility < 0.1054815   to the right, improve=0.03801279, (0 missing)
  Surrogate splits:
      Item_Visibility < 0.05747654  to the right, agree=0.606, adj=0.2, (0 split)

Node number 163: 459 observations
  mean=45.34761, MSE=229.4434

Node number 324: 450 observations
  mean=17.66319, MSE=57.44482

Node number 325: 464 observations
  mean=45.10712, MSE=236.225
```

Hide

```
main_predict2 <- predict(tree1, newdata = new_test_combi_encoded_dummies_drop, type = "vector")
sub_file <- data.frame(Item_Identifier = test$Item_Identifier, Outlet_Identifier = test$Outlet_I
dentifier,       Item_Outlet_Sales = main_predict2)
write.csv(sub_file, 'Decision_tree_sales.csv')
```

Hide

```
main_predict3 <- predict(tree2, newdata = new_test_combi_encoded_dummies_drop, type = "vector")
sub_file <- data.frame(Item_Identifier = test$Item_Identifier, Outlet_Identifier = test$Outlet_I
dentifier,       Item_Outlet_Sales = main_predict3)
write.csv(sub_file, 'Decision_tree_sales_sqrt.csv')
```