Bigmart Sales

Yatharth Malik

January 31, 2017

```
library(caret)
library(plyr)
library(dplyr)
library(dummies)
library(mlr)
library(rpart)
library(rpart.plot)
library(caret)
library(e1071)
library(Metrics)
library(randomForest)
```

Loading data and exploration

```
train = read.csv("train.csv",na.strings = c(""," ",NA,"NA"))
test = read.csv("test.csv",na.strings = c(""," ",NA,"NA"))

summary(train)
```

```
## Item_Identifier  Item_Weight     Item_Fat_Content Item_Visibility
## FDG33  :  10    Min.   : 4.555   LF     : 316     Min.   :0.00000
## FDW13  :  10    1st Qu.: 8.774   low fat: 112     1st Qu.:0.02699
## DRE49  :   9    Median :12.600   Low Fat:5089     Median :0.05393
## DRN47  :   9    Mean   :12.858   reg    : 117     Mean   :0.06613
## FDD38  :   9    3rd Qu.:16.850   Regular:2889     3rd Qu.:0.09459
## FDF52  :   9    Max.   :21.350                    Max.   :0.32839
## (Other):8467    NA's   :1463
##                 Item_Type       Item_MRP      Outlet_Identifier
## Fruits and Vegetables:1232   Min.   : 31.29   OUT027 : 935
## Snack Foods          :1200   1st Qu.: 93.83   OUT013 : 932
## Household            : 910   Median :143.01   OUT035 : 930
## Frozen Foods         : 856   Mean   :140.99   OUT046 : 930
## Dairy                : 682   3rd Qu.:185.64   OUT049 : 930
## Canned               : 649   Max.   :266.89   OUT045 : 929
## (Other)              :2994                    (Other):2937
## Outlet_Establishment_Year Outlet_Size   Outlet_Location_Type
## Min.   :1985             High  : 932   Tier 1:2388
## 1st Qu.:1987             Medium:2793   Tier 2:2785
## Median :1999             Small :2388   Tier 3:3350
## Mean   :1998             NA's  :2410
## 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
##
```

```
str(train)
```

```
## 'data.frame':    8523 obs. of  12 variables:
##  $ Item_Identifier          : Factor w/ 1559 levels "DRA12","DRA24",..: 157 9 663 1122 1298 7
59 697 739 441 991 ...
##  $ 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/ 3 levels "High","Medium",..: 2 2 2 NA 1 2 1 2 NA NA
 ...
##  $ 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 ...
```

```
summary(test)
```

```
##  Item_Identifier  Item_Weight     Item_Fat_Content Item_Visibility
##  DRF48  :   8    Min.   : 4.555   LF      : 206     Min.   :0.00000
##  FDK57  :   8    1st Qu.: 8.645   low fat:  66     1st Qu.:0.02705
##  FDN52  :   8    Median :12.500   Low Fat:3396     Median :0.05415
##  FDP15  :   8    Mean   :12.696   reg     :  78     Mean   :0.06568
##  FDQ60  :   8    3rd Qu.:16.700   Regular:1935     3rd Qu.:0.09346
##  FDW10  :   8    Max.   :21.350                    Max.   :0.32364
##  (Other):5633    NA's   :976
##                 Item_Type       Item_MRP      Outlet_Identifier
##  Snack Foods          : 789   Min.   : 31.99   OUT027 : 624
##  Fruits and Vegetables: 781   1st Qu.: 94.41   OUT013 : 621
##  Household            : 638   Median :141.42   OUT035 : 620
##  Frozen Foods         : 570   Mean   :141.02   OUT046 : 620
##  Dairy                : 454   3rd Qu.:186.03   OUT049 : 620
##  Baking Goods         : 438   Max.   :266.59   OUT045 : 619
##  (Other)              :2011                    (Other):1957
##  Outlet_Establishment_Year Outlet_Size   Outlet_Location_Type
##  Min.   :1985              High  : 621   Tier 1:1592
##  1st Qu.:1987              Medium:1862   Tier 2:1856
##  Median :1999              Small :1592   Tier 3:2233
##  Mean   :1998              NA's  :1606
##  3rd Qu.:2004
##  Max.   :2009
##
##             Outlet_Type
##  Grocery Store    : 722
##  Supermarket Type1:3717
##  Supermarket Type2: 618
##  Supermarket Type3: 624
##
##
##
```

```
str(test)
```

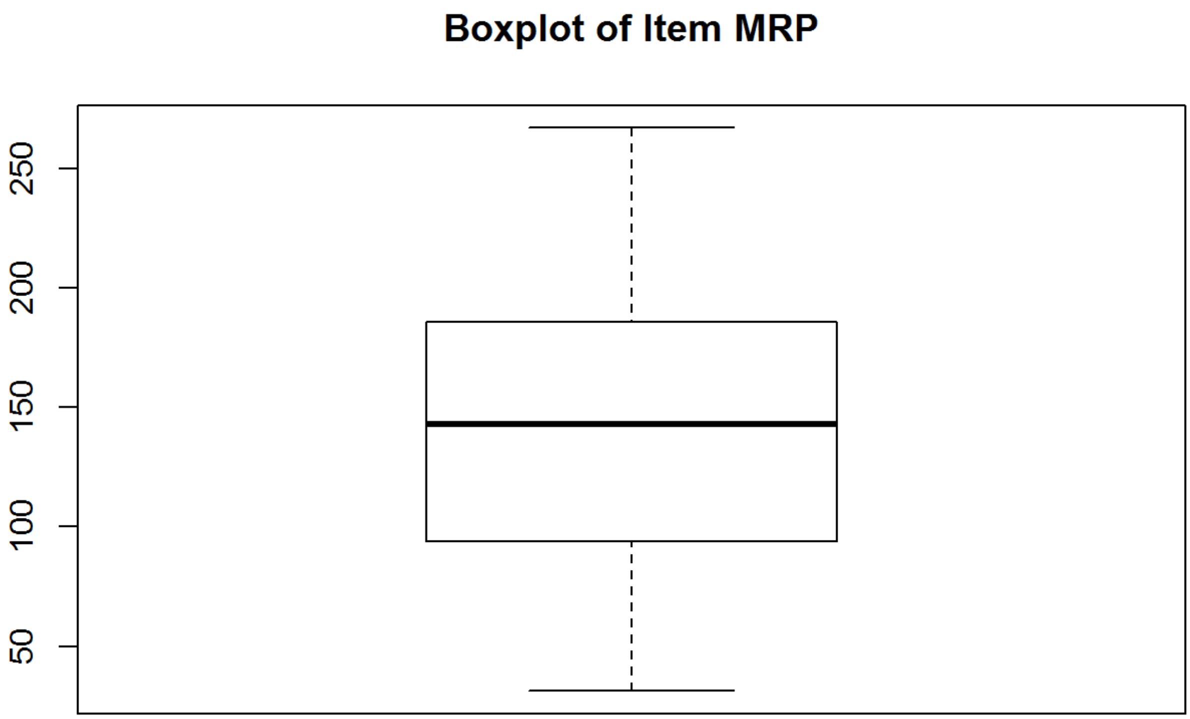
```
## 'data.frame':    5681 obs. of  11 variables:
##  $ Item_Identifier          : Factor w/ 1543 levels "DRA12","DRA24",..: 1104 1068 1407 810 11
85 462 605 267 669 171 ...
##  $ Item_Weight              : num  20.75 8.3 14.6 7.32 NA ...
##  $ Item_Fat_Content         : Factor w/ 5 levels "LF","low fat",..: 3 4 3 3 5 5 5 3 5 3 ...
##  $ Item_Visibility          : num  0.00756 0.03843 0.09957 0.01539 0.1186 ...
##  $ Item_Type                : Factor w/ 16 levels "Baking Goods",..: 14 5 12 14 5 7 1 1 14 1
 ...
##  $ Item_MRP                 : num  107.9 87.3 241.8 155 234.2 ...
##  $ Outlet_Identifier        : Factor w/ 10 levels "OUT010","OUT013",..: 10 3 1 3 6 9 4 6 8 3
 ...
##  $ Outlet_Establishment_Year: int  1999 2007 1998 2007 1985 1997 2009 1985 2002 2007 ...
##  $ Outlet_Size              : Factor w/ 3 levels "High","Medium",..: 2 NA NA NA 2 3 2 2 NA NA
 ...
##  $ Outlet_Location_Type     : Factor w/ 3 levels "Tier 1","Tier 2",..: 1 2 3 2 3 1 3 3 2 2
 ...
##  $ Outlet_Type              : Factor w/ 4 levels "Grocery Store",..: 2 2 1 2 4 2 3 4 2 2 ...
```

Infrences drawn from data exploration :-

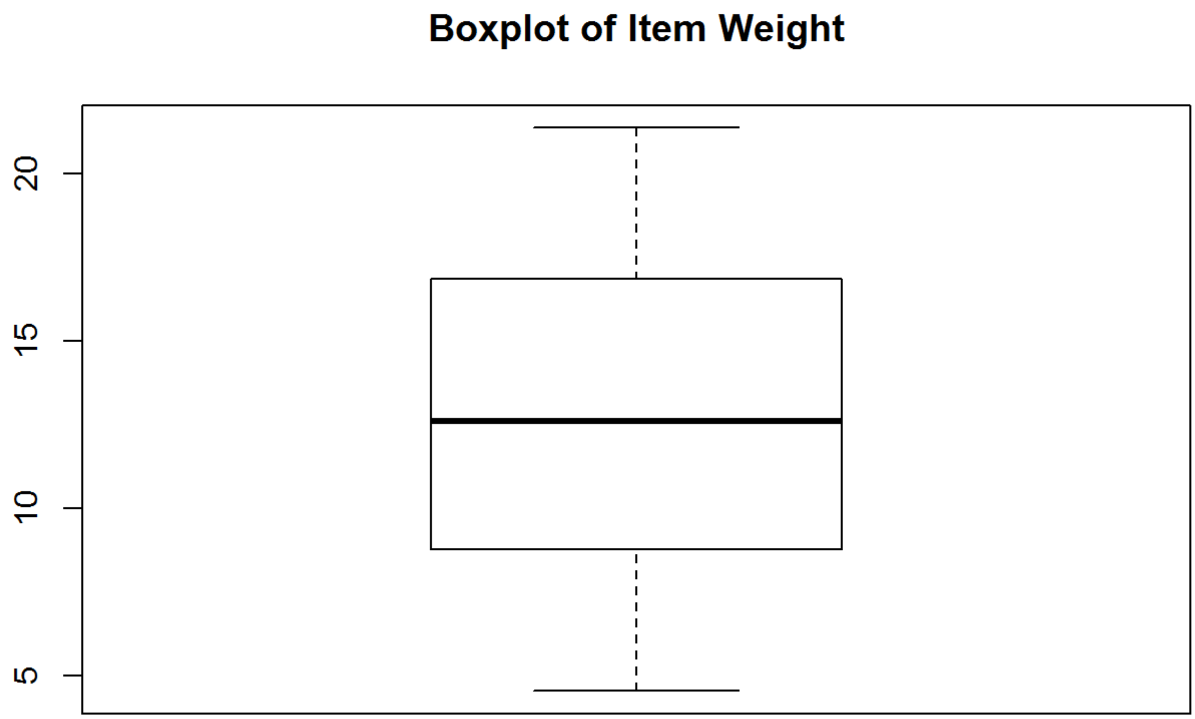
1. Factor mismatch in Item_Fat_Content.
2. Missing values in Item_Weight and Outlet_Size.
3. Minimum value of Item_Visibility is 0,which is not practically possible.Hence,we'll deal them as missing values.

Univariate Analysis

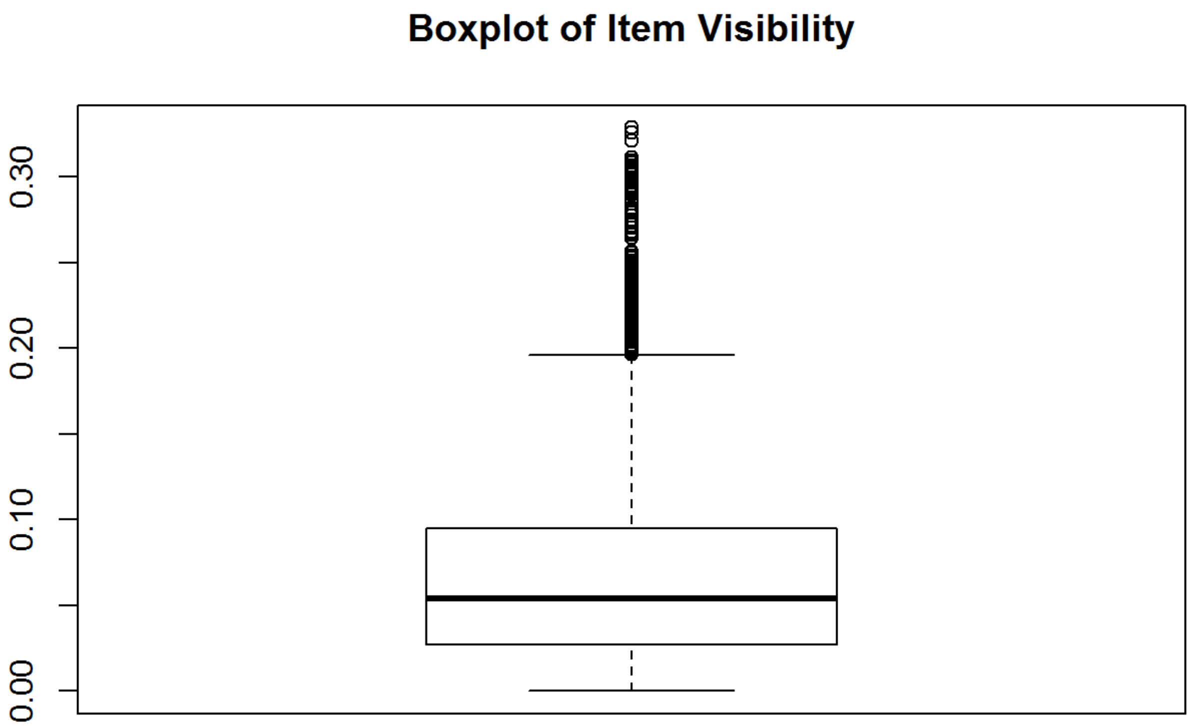
```
boxplot(train$Item_MRP,main = "Boxplot of Item MRP")
```




```
boxplot(train$Item_Weight,main = "Boxplot of Item Weight")
```

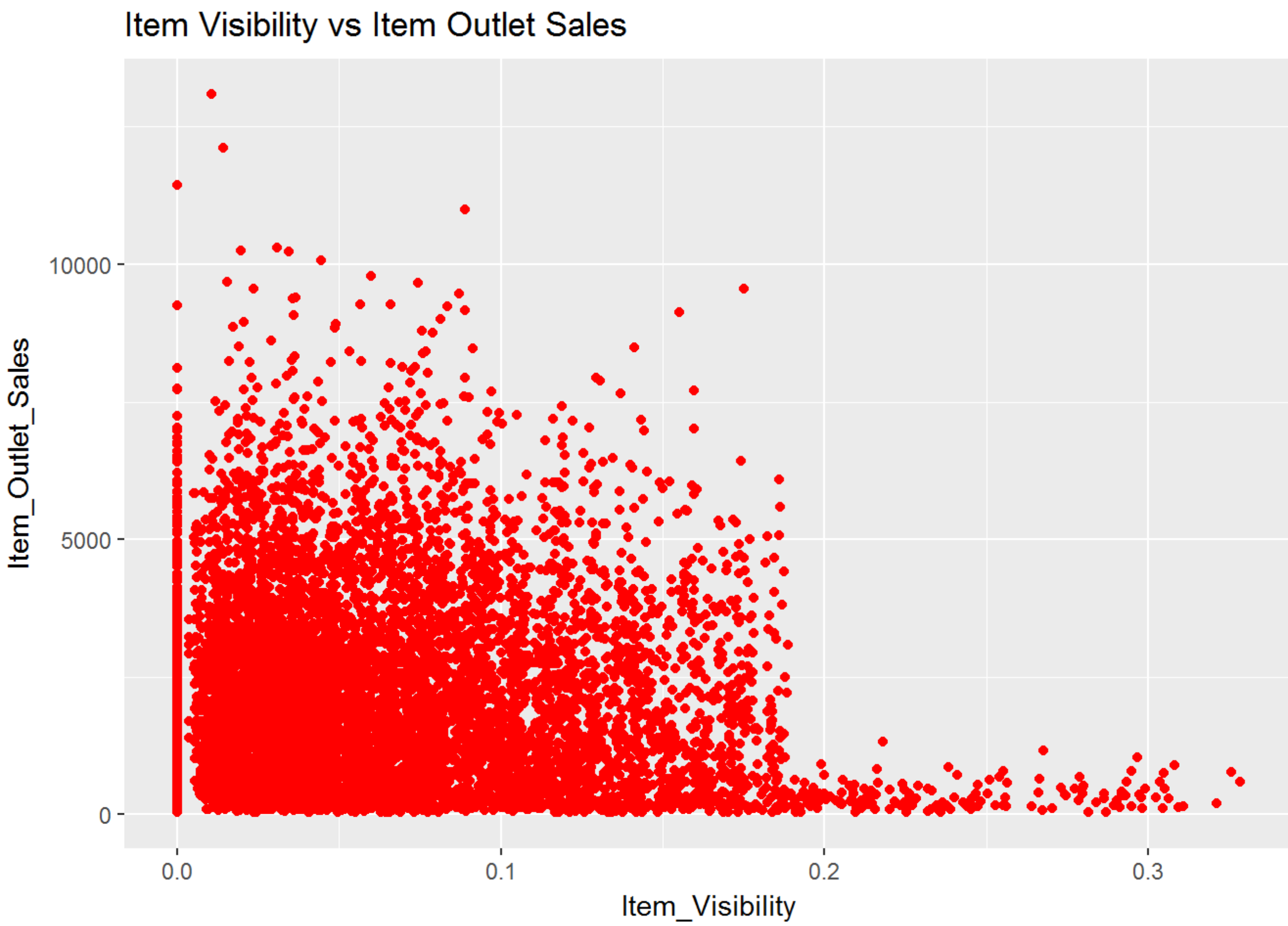



```
boxplot(train$Item_Visibility,main = "Boxplot of Item Visibility")
```

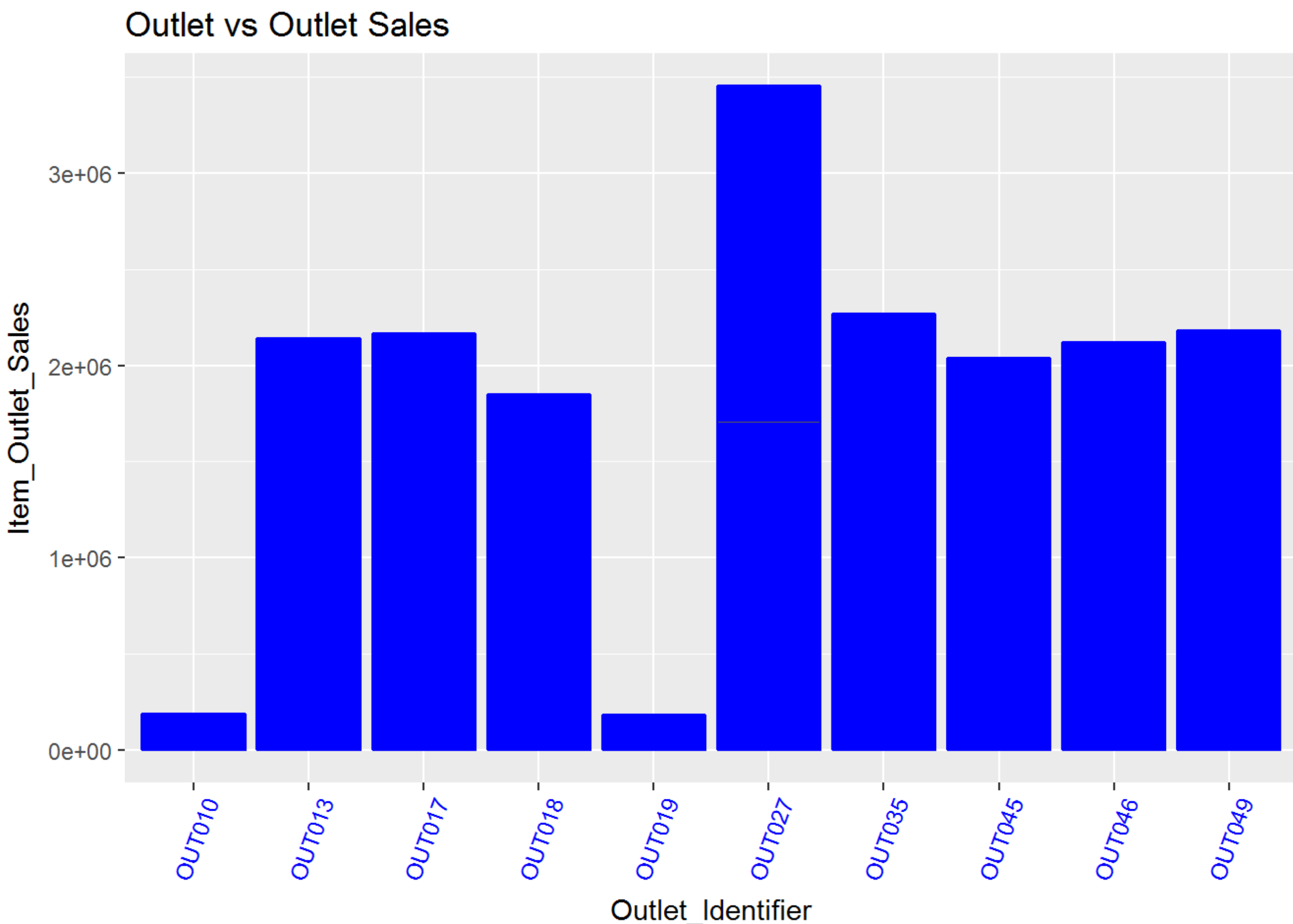


Bivariate Analysis

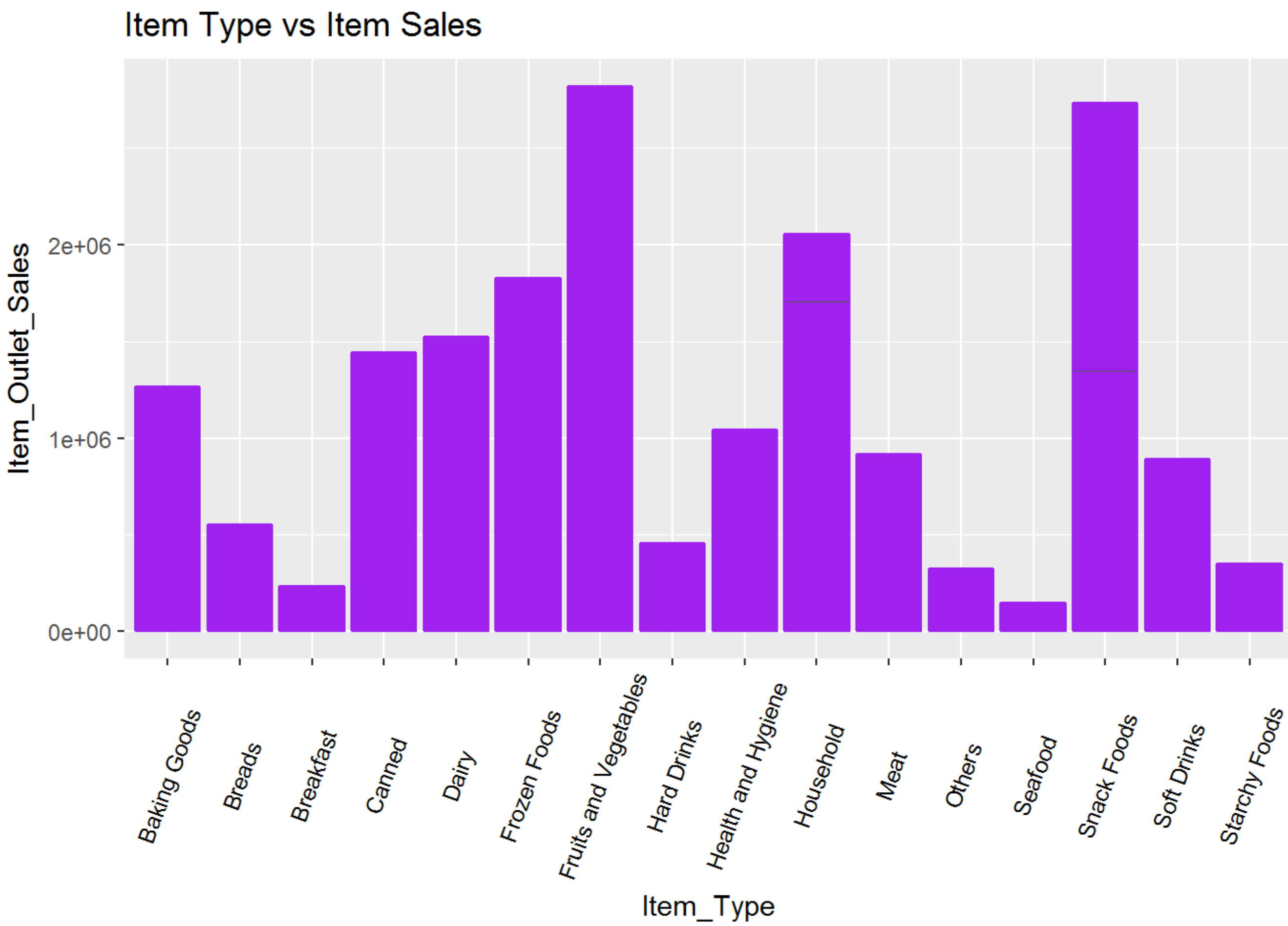
```
ggplot(train,aes(x=Item_Visibility,y=Item_Outlet_Sales)) + geom_point(color = "red") +
ggtitle("Item Visibility vs Item Outlet Sales")
```




```
ggplot(train,aes(x=Outlet_Identifier,y= Item_Outlet_Sales)) + geom_bar(stat="identity",color =
"blue") + ggtitle("Outlet vs Outlet Sales") + theme(axis.text.x = element_text(angle = 70,vjust
= 0.5,color = "blue"))
```




```
ggplot(train,aes(x=Item_Type,y= Item_Outlet_Sales)) + geom_bar(stat="identity",color = "purple")
 + ggtitle("Item Type vs Item Sales") + theme(axis.text.x = element_text(angle = 70,vjust =
0.5,color = "black"))
```



##

Dealing with categorical and continuous variables

We will use median imputation to deal with continuous missing values

```
test$Item_Outlet_Sales = 1
comb = rbind(train,test)

comb$Item_Weight[is.na(comb$Item_Weight)] = median(comb$Item_Weight,na.rm = T)

comb$Item_Visibility = ifelse(comb$Item_Visibility==0,median(comb$Item_Visibility),comb$Item_Vis
ibility)

comb$Outlet_Size = ifelse(is.na(comb$Outlet_Size),"Others",comb$Outlet_Size)
comb$Outlet_Size = as.factor(comb$Outlet_Size)
levels(comb$Outlet_Size)[1] = "High"
levels(comb$Outlet_Size)[2] = "Medium"
levels(comb$Outlet_Size)[3] = "Low"

table(comb$Item_Fat_Content)
```

```
##
##       LF  low fat Low Fat     reg Regular
##      522      178    8485     195    4824
```

```
comb$Item_Fat_Content = revalue(comb$Item_Fat_Content,c("LF" = "Low Fat","reg"="Regular"))
comb$Item_Fat_Content = revalue(comb$Item_Fat_Content,c("low fat"="Low Fat"))

table(comb$Item_Fat_Content)
```

```
##
## Low Fat Regular
##    9185    5019
```

Feature Engineering

```
temp = comb%>%group_by(Outlet_Identifier)%>%tally()
names(temp)[2]  = "Outlet_Count"
comb = full_join(comb,temp,by = "Outlet_Identifier")

temp1 = comb%>%group_by(Item_Identifier)%>%tally()
names(temp1)[2]  =  "Item_Count"
comb = merge(comb,temp1,by = "Item_Identifier")

temp2 = comb%>%select(Outlet_Establishment_Year)%>%mutate(Outlet_Year = 2013 - comb$Outlet_Estab
lishment_Year)
temp2$Outlet_Establishment_Year = NULL
comb = cbind(comb,temp2 )


items = substr(comb$Item_Identifier,1,2)
items = gsub("FD","Food",items)
items = gsub("DR","Drinks",items)
items = gsub("NC","Non Consumable",items)
comb$Item_Type_New = factor(items)

str(comb)
```

```
## 'data.frame':    14204 obs. of  16 variables:
##  $ Item_Identifier          : Factor w/ 1559 levels "DRA12","DRA24",..: 1 1 1 1 1 1 1 1 1 2
 ...
##  $ Item_Weight              : num  11.6 11.6 11.6 11.6 12.6 ...
##  $ Item_Fat_Content         : Factor w/ 2 levels "Low Fat","Regular": 1 1 1 1 1 1 1 1 1 2 ...
##  $ Item_Visibility          : num  0.054 0.041 0.054 0.0409 0.0407 ...
##  $ Item_Type                : Factor w/ 16 levels "Baking Goods",..: 15 15 15 15 15 15 15 15
 15 15 ...
##  $ Item_MRP                 : num  142 141 142 143 140 ...
##  $ Outlet_Identifier        : Factor w/ 10 levels "OUT010","OUT013",..: 7 10 8 9 6 4 1 2 3 9
 ...
##  $ Outlet_Establishment_Year: int  2004 1999 2002 1997 1985 2009 1998 1987 2007 1997 ...
##  $ Outlet_Size              : Factor w/ 4 levels "High","Medium",..: 3 2 4 3 2 2 4 1 4 3 ...
##  $ Outlet_Location_Type     : Factor w/ 3 levels "Tier 1","Tier 2",..: 2 1 2 1 3 3 3 3 2 1
 ...
##  $ Outlet_Type              : Factor w/ 4 levels "Grocery Store",..: 2 2 2 2 4 3 1 2 2 2 ...
##  $ Item_Outlet_Sales        : num  993 1 3829 1 1 ...
##  $ Outlet_Count             : int  1550 1550 1548 1550 1559 1546 925 1553 1543 1550 ...
##  $ Item_Count               : int  9 9 9 9 9 9 9 9 9 10 ...
##  $ Outlet_Year              : num  9 14 11 16 28 4 15 26 6 16 ...
##  $ Item_Type_New            : Factor w/ 3 levels "Drinks","Food",..: 1 1 1 1 1 1 1 1 1 1 ...
```

One Hot Encoding

```
comb = dummy.data.frame(comb,names = c("Outlet_Size","Outlet_Location_Type","Outlet_Type","Item_
Type_New","Item_Fat_Content"),sep='_')
str(comb)
```

```
## 'data.frame':    14204 obs. of  27 variables:
##  $ Item_Identifier              : Factor w/ 1559 levels "DRA12","DRA24",..: 1 1 1 1 1 1 1 1 1
 2 ...
##  $ Item_Weight                  : num  11.6 11.6 11.6 11.6 12.6 ...
##  $ Item_Fat_Content_Low Fat     : int  1 1 1 1 1 1 1 1 1 0 ...
##  $ Item_Fat_Content_Regular     : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ Item_Visibility              : num  0.054 0.041 0.054 0.0409 0.0407 ...
##  $ Item_Type                    : Factor w/ 16 levels "Baking Goods",..: 15 15 15 15 15 15 15
 15 15 15 ...
##  $ Item_MRP                     : num  142 141 142 143 140 ...
##  $ Outlet_Identifier            : Factor w/ 10 levels "OUT010","OUT013",..: 7 10 8 9 6 4 1 2
 3 9 ...
##  $ Outlet_Establishment_Year    : int  2004 1999 2002 1997 1985 2009 1998 1987 2007 1997 ...
##  $ Outlet_Size_High             : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Outlet_Size_Medium           : int  0 1 0 0 1 1 0 0 0 0 ...
##  $ Outlet_Size_Low              : int  1 0 0 1 0 0 0 0 0 1 ...
##  $ Outlet_Size_Others           : int  0 0 1 0 0 0 1 0 1 0 ...
##  $ Outlet_Location_Type_Tier 1  : int  0 1 0 1 0 0 0 0 0 1 ...
##  $ Outlet_Location_Type_Tier 2  : int  1 0 1 0 0 0 0 0 1 0 ...
##  $ Outlet_Location_Type_Tier 3  : int  0 0 0 0 1 1 1 1 0 0 ...
##  $ Outlet_Type_Grocery Store    : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ Outlet_Type_Supermarket Type1: int  1 1 1 1 0 0 0 1 1 1 ...
##  $ Outlet_Type_Supermarket Type2: int  0 0 0 0 0 1 0 0 0 0 ...
##  $ Outlet_Type_Supermarket Type3: int  0 0 0 0 1 0 0 0 0 0 ...
##  $ Item_Outlet_Sales            : num  993 1 3829 1 1 ...
##  $ Outlet_Count                 : int  1550 1550 1548 1550 1559 1546 925 1553 1543 1550 ...
##  $ Item_Count                   : int  9 9 9 9 9 9 9 9 9 10 ...
##  $ Outlet_Year                  : num  9 14 11 16 28 4 15 26 6 16 ...
##  $ Item_Type_New_Drinks         : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Item_Type_New_Food           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Item_Type_New_Non Consumable : int  0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, "dummies")=List of 5
##   ..$ Item_Fat_Content    : int  3 4
##   ..$ Outlet_Size         : int  10 11 12 13
##   ..$ Outlet_Location_Type: int  14 15 16
##   ..$ Outlet_Type         : int  17 18 19 20
##   ..$ Item_Type_New       : int  25 26 27
```

Predictive Modelling

```
comb = select(comb,-c(Item_Identifier,Outlet_Identifier,Item_Type,Outlet_Establishment_Year))

new_train = comb[1:nrow(train),]
new_test = comb[-(1:nrow(train)),]
names(new_train) = make.names(names(new_train))
names(new_test) = make.names(names(new_test))
```

1. 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
## -3436.9 -1096.4   -43.0   791.8  8883.7
##
## Coefficients: (7 not defined because of singularities)
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    3.611e+04  2.202e+04   1.640   0.1011
## Item_Weight                    2.461e+00  3.980e+00   0.618   0.5363
## Item_Fat_Content_Low.Fat      -5.059e+01  3.508e+01  -1.442   0.1493
## Item_Fat_Content_Regular              NA         NA      NA       NA
## Item_Visibility               -1.572e+02  3.501e+02  -0.449   0.6535
## Item_MRP                       9.524e+00  2.631e-01  36.204  < 2e-16 ***
## Outlet_Size_High              -9.254e+02  6.064e+02  -1.526   0.1270
## Outlet_Size_Medium             2.335e+02  1.132e+02   2.063   0.0392 *
## Outlet_Size_Low                1.868e+02  8.841e+01   2.113   0.0346 *
## Outlet_Size_Others                    NA         NA      NA       NA
## Outlet_Location_Type_Tier.1   -1.222e+03  7.101e+02  -1.721   0.0854 .
## Outlet_Location_Type_Tier.2   -1.083e+03  7.101e+02  -1.525   0.1274
## Outlet_Location_Type_Tier.3           NA         NA      NA       NA
## Outlet_Type_Grocery.Store     -1.628e+04  8.919e+03  -1.825   0.0680 .
## Outlet_Type_Supermarket.Type1  2.058e+02  5.888e+02   0.350   0.7267
## Outlet_Type_Supermarket.Type2 -1.314e+03  1.972e+02  -6.661 2.89e-11 ***
## Outlet_Type_Supermarket.Type3         NA         NA      NA       NA
## Outlet_Count                  -2.285e+01  1.417e+01  -1.613   0.1068
## Item_Count                     1.781e+01  2.327e+01   0.765   0.4441
## Outlet_Year                           NA         NA      NA       NA
## Item_Type_New_Drinks          -1.663e+01  4.787e+01  -0.347   0.7283
## Item_Type_New_Food                    NA         NA      NA       NA
## Item_Type_New_Non.Consumable          NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1526 on 8507 degrees of freedom
## Multiple R-squared:  0.2097, Adjusted R-squared:  0.2083
## F-statistic: 150.5 on 15 and 8507 DF,  p-value: < 2.2e-16
```

```
pred_lm = predict(linear_model,type = "response")
rmse(new_train$Item_Outlet_Sales,pred_lm)
```

```
## [1] 1524.375
```

2. Decision Trees

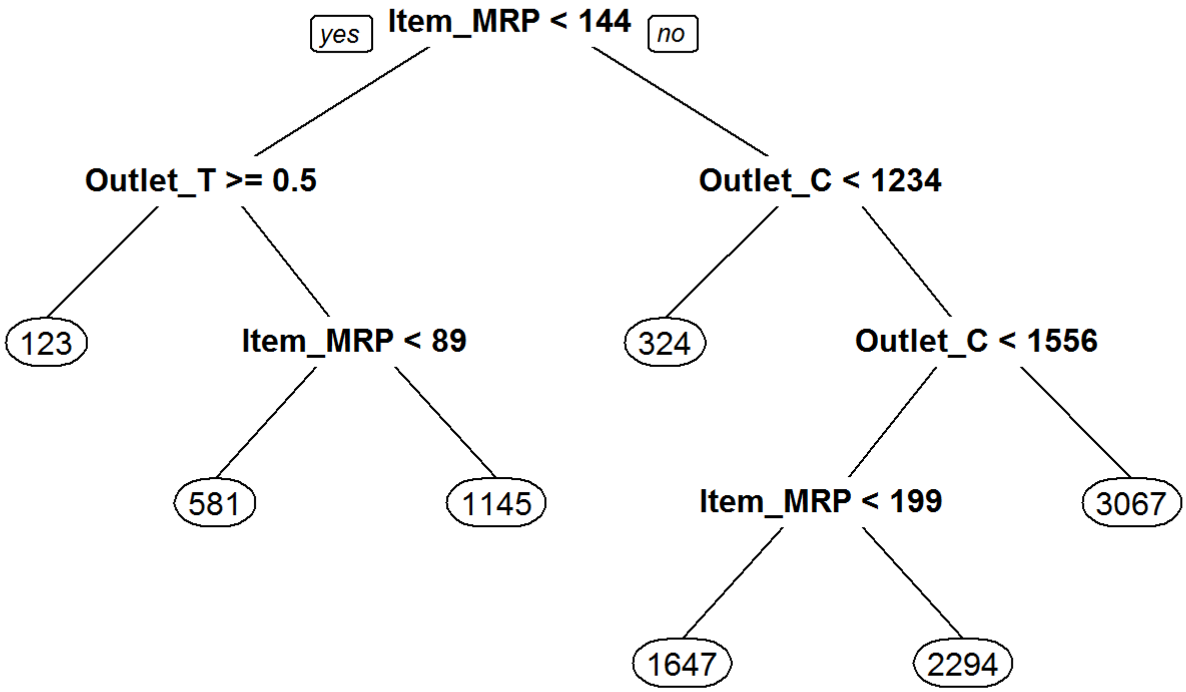
```
tree_model = rpart(Item_Outlet_Sales ~ . ,data = new_test)
summary(tree_model)
```

```
## Call:
## rpart(formula = Item_Outlet_Sales ~ ., data = new_test)
##   n= 5681
##
##           CP nsplit rel error    xerror       xstd
## 1 0.08646909      0 1.0000000 1.0001613 0.02827382
## 2 0.05420419      1 0.9135309 0.9141970 0.02420009
## 3 0.02443891      2 0.8593267 0.8597872 0.02284921
## 4 0.01316430      3 0.8348878 0.8366061 0.02050304
## 5 0.01224056      4 0.8217235 0.8246613 0.02025643
## 6 0.01210477      5 0.8094829 0.8229437 0.02023515
## 7 0.01000000      6 0.7973782 0.8076124 0.01986044
##
## Variable importance
##                      Item_MRP                  Outlet_Count
##                            33                            27
##     Outlet_Type_Grocery.Store Outlet_Type_Supermarket.Type3
##                            20                             7
##                   Outlet_Year               Item_Visibility
##                             7                             3
##                   Item_Weight                    Item_Count
##                             1                             1
##
## Node number 1: 5681 observations,    complexity param=0.08646909
##   mean=1293.524, MSE=2809010
##   left son=2 (2896 obs) right son=3 (2785 obs)
##   Primary splits:
##       Item_MRP                      < 143.797    to the left,  improve=0.08646909, (0 missin
g)
##       Outlet_Count                  < 1234       to the left,  improve=0.06113592, (0 missin
g)
##       Outlet_Type_Grocery.Store     < 0.5        to the right, improve=0.06113592, (0 missin
g)
##       Outlet_Type_Supermarket.Type3 < 0.5        to the left,  improve=0.03814215, (0 missin
g)
##       Outlet_Size_Medium            < 0.5        to the left,  improve=0.01847886, (0 missin
g)
##   Surrogate splits:
##       Item_Weight              < 13.05      to the left,  agree=0.533, adj=0.047, (0 split)
##       Item_Visibility          < 0.05837035 to the left,  agree=0.522, adj=0.025, (0 split)
##       Item_Count               < 8.5        to the right, agree=0.520, adj=0.021, (0 split)
##       Item_Fat_Content_Low.Fat < 0.5        to the right, agree=0.515, adj=0.010, (0 split)
##       Item_Fat_Content_Regular < 0.5        to the left,  agree=0.515, adj=0.010, (0 split)
##
## Node number 2: 2896 observations,    complexity param=0.0131643
##   mean=810.2202, MSE=1049318
##   left son=4 (386 obs) right son=5 (2510 obs)
##   Primary splits:
##       Outlet_Type_Grocery.Store     < 0.5        to the right, improve=0.06913057, (0 missin
g)
##       Outlet_Count                  < 1234       to the left,  improve=0.06913057, (0 missin
g)
##       Item_MRP                      < 76.6512    to the left,  improve=0.05935699, (0 missin
```

```
g)
##       Outlet_Type_Supermarket.Type3 < 0.5        to the left,  improve=0.04433924, (0 missin
g)
##       Outlet_Size_Medium            < 0.5        to the left,  improve=0.02583729, (0 missin
g)
##   Surrogate splits:
##       Outlet_Count    < 1234       to the left,  agree=1.000, adj=1.000, (0 split)
##       Item_Visibility < 0.1756642  to the right, agree=0.885, adj=0.135, (0 split)
##
## Node number 3: 2785 observations,    complexity param=0.05420419
##   mean=1796.091, MSE=4143372
##   left son=6 (349 obs) right son=7 (2436 obs)
##   Primary splits:
##       Outlet_Count                  < 1234       to the left,  improve=0.07496041, (0 missin
g)
##       Outlet_Type_Grocery.Store     < 0.5        to the right, improve=0.07496041, (0 missin
g)
##       Outlet_Type_Supermarket.Type3 < 0.5        to the left,  improve=0.04774003, (0 missin
g)
##       Outlet_Size_Medium            < 0.5        to the left,  improve=0.02061265, (0 missin
g)
##       Item_MRP                      < 220.0456   to the left,  improve=0.01811604, (0 missin
g)
##   Surrogate splits:
##       Outlet_Type_Grocery.Store < 0.5        to the right, agree=1.000, adj=1.000, (0 split)
##       Item_Visibility           < 0.1896654  to the right, agree=0.889, adj=0.112, (0 split)
##
## Node number 4: 386 observations
##   mean=123.4176, MSE=23738.38
##
## Node number 5: 2510 observations,    complexity param=0.01210477
##   mean=915.84, MSE=1123341
##   left son=10 (1020 obs) right son=11 (1490 obs)
##   Primary splits:
##       Item_MRP                      < 88.6185    to the left,  improve=0.06850924, (0 missin
g)
##       Outlet_Type_Supermarket.Type3 < 0.5        to the left,  improve=0.03339728, (0 missin
g)
##       Outlet_Count                  < 1556       to the left,  improve=0.03339728, (0 missin
g)
##       Outlet_Year                   < 27         to the left,  improve=0.03339728, (0 missin
g)
##       Outlet_Type_Supermarket.Type1 < 0.5        to the right, improve=0.01175536, (0 missin
g)
##   Surrogate splits:
##       Item_Visibility < 0.01236591 to the left,  agree=0.599, adj=0.013, (0 split)
##       Item_Count      < 7.5        to the left,  agree=0.596, adj=0.007, (0 split)
##
## Node number 6: 349 observations
##   mean=323.7147, MSE=113127.9
##
## Node number 7: 2436 observations,    complexity param=0.02443891
##   mean=2007.035, MSE=4365689
##   left son=14 (2132 obs) right son=15 (304 obs)
```

```
##   Primary splits:
##       Outlet_Count                  < 1556       to the left,  improve=0.036671610, (0 missin
g)
##       Outlet_Year                   < 27         to the left,  improve=0.036671610, (0 missin
g)
##       Outlet_Type_Supermarket.Type3 < 0.5        to the left,  improve=0.036671610, (0 missin
g)
##       Item_MRP                      < 220.3285   to the left,  improve=0.021380880, (0 missin
g)
##       Outlet_Type_Supermarket.Type1 < 0.5        to the right, improve=0.009034658, (0 missin
g)
##   Surrogate splits:
##       Outlet_Type_Supermarket.Type3 < 0.5        to the left,  agree=1.000, adj=1.00, (0 spli
t)
##       Outlet_Year                   < 27         to the left,  agree=1.000, adj=1.00, (0 spli
t)
##       Outlet_Type_Supermarket.Type1 < 0.5        to the right, agree=0.876, adj=0.01, (0 spli
t)
##
## Node number 10: 1020 observations
##   mean=580.5478, MSE=431194.4
##
## Node number 11: 1490 observations
##   mean=1145.369, MSE=1467517
##
## Node number 14: 2132 observations,    complexity param=0.01224056
##   mean=1855.945, MSE=3569990
##   left son=28 (1443 obs) right son=29 (689 obs)
##   Primary splits:
##       Item_MRP        < 199.0584   to the left,  improve=0.025664040, (0 missing)
##       Item_Weight     < 6.6925     to the left,  improve=0.003830461, (0 missing)
##       Outlet_Count    < 1549       to the left,  improve=0.003305032, (0 missing)
##       Item_Visibility < 0.1859728  to the left,  improve=0.003216576, (0 missing)
##       Outlet_Year     < 7.5        to the left,  improve=0.002334160, (0 missing)
##   Surrogate splits:
##       Item_Weight     < 5.0725     to the right, agree=0.684, adj=0.023, (0 split)
##       Item_Count      < 7.5        to the right, agree=0.682, adj=0.017, (0 split)
##       Item_Visibility < 0.1806009  to the left,  agree=0.680, adj=0.010, (0 split)
##
## Node number 15: 304 observations
##   mean=3066.65, MSE=8663172
##
## Node number 28: 1443 observations
##   mean=1646.788, MSE=2745550
##
## Node number 29: 689 observations
##   mean=2293.991, MSE=5013142
```

```
prp(tree_model)
```




```
pred_tree = predict(tree_model,type= "vector")
rmse(new_train$Item_Outlet_Sales,pred_tree)
```

```
## Warning in actual - predicted: longer object length is not a multiple of
## shorter object length
```

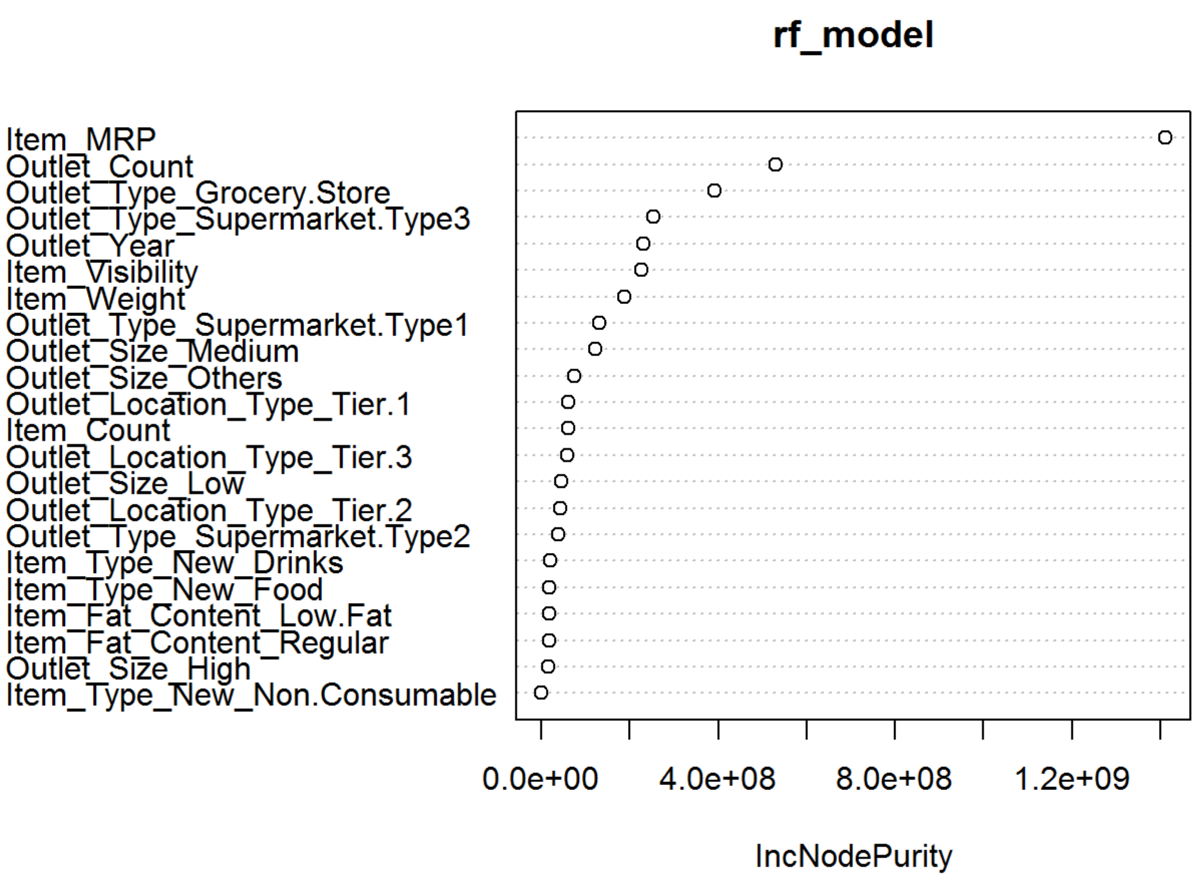
```
## [1] 1899.484
```

3. Random Forest

```
rf_model = randomForest(Item_Outlet_Sales ~ . ,data = new_train,mtry = 2 ,ntree = 1000)
summary(rf_model)
```

```
##                 Length Class  Mode
## call               5   -none- call
## type               1   -none- character
## predicted       8523   -none- numeric
## mse             1000   -none- numeric
## rsq             1000   -none- numeric
## oob.times       8523   -none- numeric
## importance        22   -none- numeric
## importanceSD       0   -none- NULL
## localImportance    0   -none- NULL
## proximity          0   -none- NULL
## ntree              1   -none- numeric
## mtry               1   -none- numeric
## forest            11   -none- list
## coefs              0   -none- NULL
## y               8523   -none- numeric
## test               0   -none- NULL
## inbag              0   -none- NULL
## terms              3   terms  call
```

```
varImpPlot(rf_model)
```




```
pred_rf = predict(rf_model,type="response")
rmse(new_train$Item_Outlet_Sales,pred_rf)
```

```
## [1] 1566.205
```