Science Meets Business

RUTGERS
Professional Science Master's Program

MBS | Master of Business & Science

OUTLET SALES PREDICTION

-Multiple Linear Regression

Team:

- Shreya Bagchi
- Akanksha Sharma
- Namrita Gupta
- Ridhi Tatineni

What are we AIMING to do?

-We are analysing the dataset of BigMart which has about 1559 products distributed across 10 outlet stores. The aim is to build a predictive model based on the features contributing to the sales of the product in each outlet store.

-From our analysis, we wish to understand what features best contribute to the sales prediction in these outlet stores.

DATA COLLECTION

-This data was collected from Analytics Vidhya.

-This dataset consist of 10 outlets that were established between 1985 to 2009. They are categorised by the size of the Outlets and the individual items.

-The dataset had 8523 rows and was reduced to 8499 rows after data cleaning.

Analytics Vidhya

What **QUESTIONS** can we answer?

What factors play a key role in increasing sales?

Is it the **product**

- Item type ?
- MRP of the products?
- Item fat content?
- Item visibility?
- Item weight

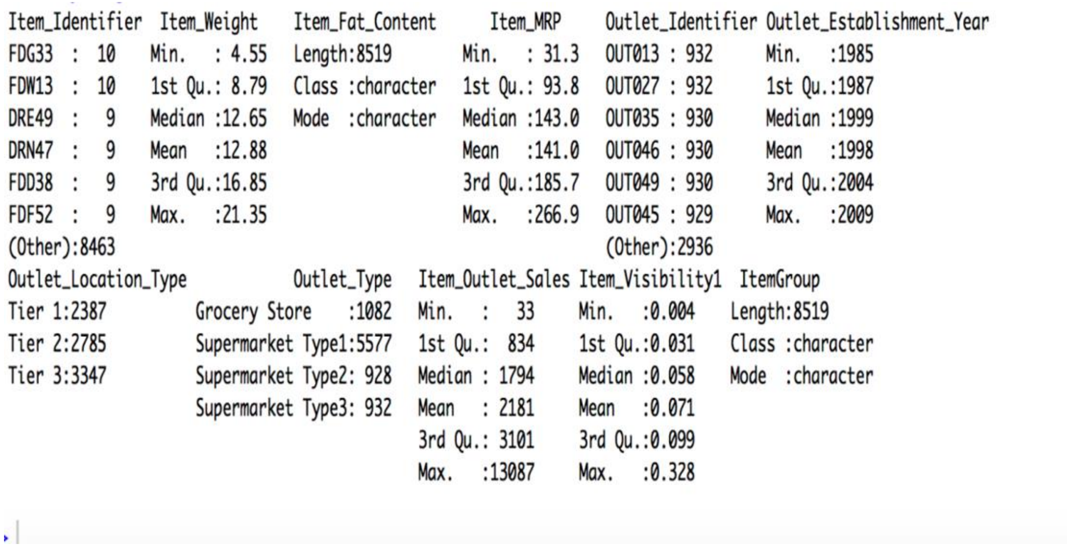
Is it the **outlet**?

- Outlet size?
- Outlet establishment year?
- Outlet type?
- Outlet location type?

DATA CLEANING - Missing value treatment

- **Item_Visibility:** There were some records with 0 visibility. Since, in reality, there cannot be any item which cannot be visible in a store, we have replaced all those values with the mean of the items according to their Item ID (Item_Identifier).
- **Item_Fat_Content:** There were 5 unique values (LF, low fat, reg, Low Fat, Regular) for two types of Fat Content types namely, Low Fat and Regular We have substituted all 'low fat' and 'LF' with 'Low Fat', and the 'reg' with 'Regular'.
- **Item_Type:** There were 16 factors of this variable and we wanted to create a broader class of products (Drinks, Food, Non-consumable). Hence, we have classified them according to the first two letters of the Item_Identifier variable and made a separate variable called Item_Group for them..
- **Item_Weight:** There were some records with blank weights so we have calculated the mean of the weights according to their item identifiers and have replaced the blank values in that manner. 4 records were deleted as they didn't had any reference weights.
- **Outlet_Size:** We have dropped extra column such as Outlet_Size (as there is about ~28% data missing from that column).

OVERALL SUMMARY OF MISSING VALUE TREATED DATA



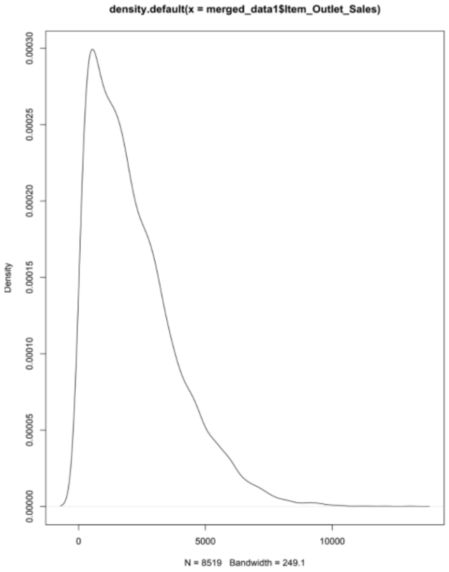
```
Item_Identifier  Item_Weight     Item_Fat_Content       Item_MRP      Outlet_Identifier Outlet_Establishment_Year
FDG33  :  10    Min.   : 4.55   Length:8519        Min.   : 31.3   OUT013 : 932      Min.   :1985
FDW13  :  10    1st Qu.: 8.79   Class :character   1st Qu.: 93.8   OUT027 : 932      1st Qu.:1987
DRE49  :   9    Median :12.65   Mode  :character   Median :143.0   OUT035 : 930      Median :1999
DRN47  :   9    Mean   :12.88                      Mean   :141.0   OUT046 : 930      Mean   :1998
FDD38  :   9    3rd Qu.:16.85                      3rd Qu.:185.7   OUT049 : 930      3rd Qu.:2004
FDF52  :   9    Max.   :21.35                      Max.   :266.9   OUT045 : 929      Max.   :2009
(Other):8463                                                       (Other):2936
Outlet_Location_Type              Outlet_Type    Item_Outlet_Sales Item_Visibility1  ItemGroup
Tier 1:2387          Grocery Store     :1082   Min.   :   33     Min.   :0.004    Length:8519
Tier 2:2785          Supermarket Type1:5577   1st Qu.:  834     1st Qu.:0.031    Class :character
Tier 3:3347          Supermarket Type2: 928   Median : 1794     Median :0.058    Mode  :character
                     Supermarket Type3: 932   Mean   : 2181     Mean   :0.071
                                              3rd Qu.: 3101     3rd Qu.:0.099
                                              Max.   :13087     Max.   :0.328
```

12 variables + 8523 rows

Data cleaning

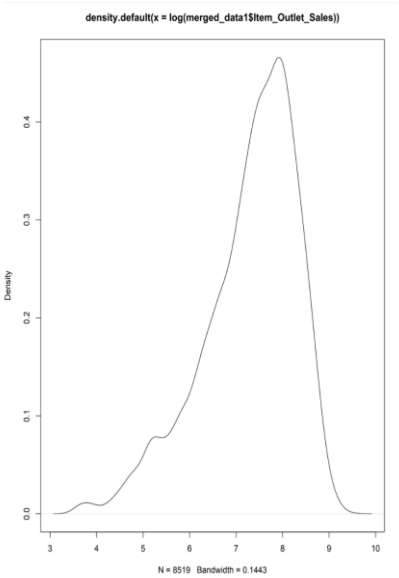
11 variables + 8519 rows

EXPLORATORY DATA ANALYSIS - Response variable



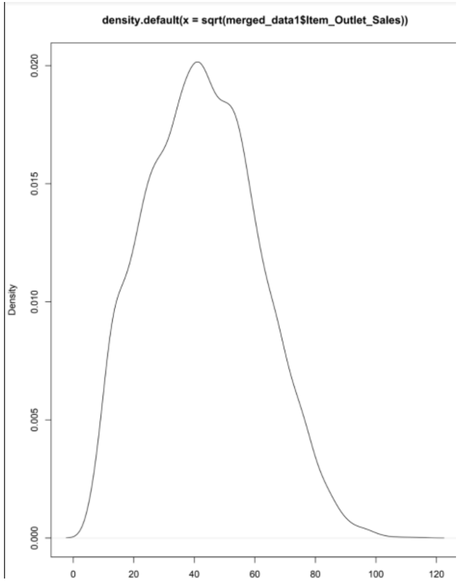
The density distribution of the Item_Outlet_Sales response variable shows right skewness meaning that the Mean is to the right of the Median.

Log Transformation of the variable



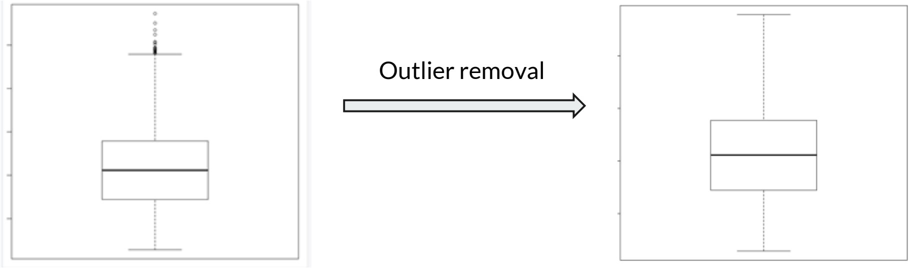
The density distribution of Item_Outlet_Sales after log transformation is slightly better but still shows slight skewness. .

Square root Transformation of the variable

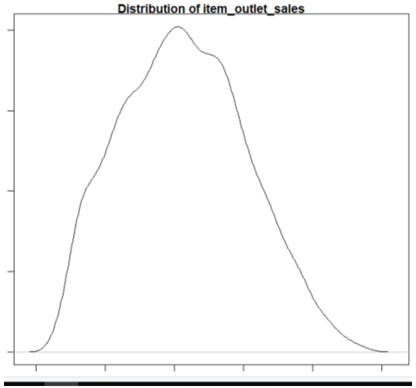


The density distribution has been made much better and more like a bell curve with square root transformation of response variable..

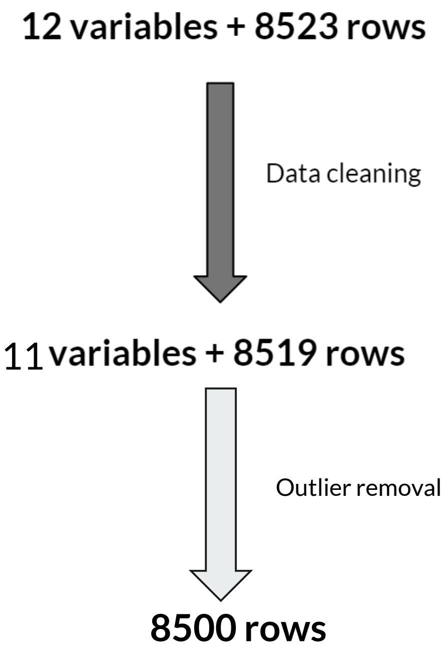
Removal of Outliers



19 data points were observed which were outliers. Those records were removed.



The density distribution of the response variable improved even further after outlier removal.



12 variables + 8523 rows

Data cleaning

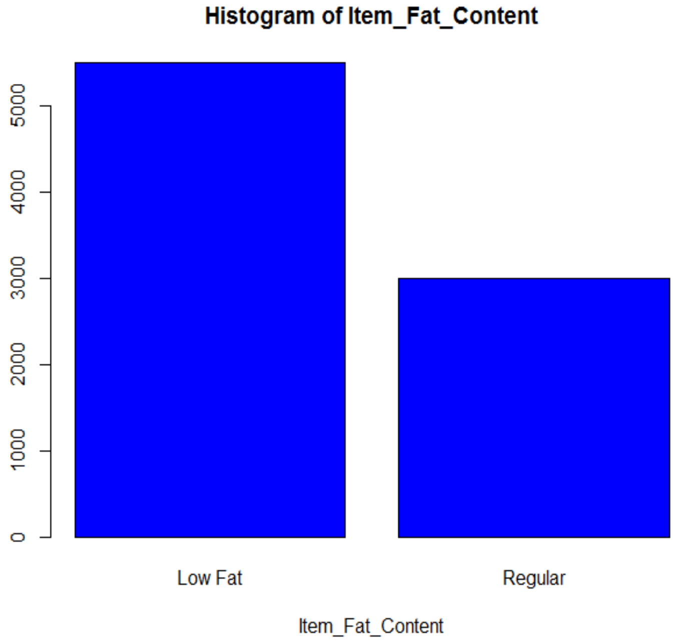
11 variables + 8519 rows

Outlier removal

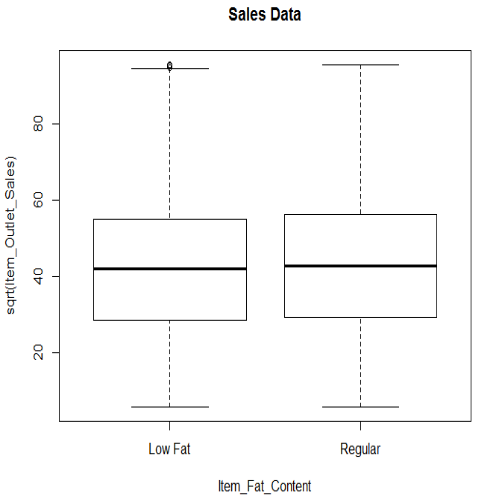
8500 rows

Let’s now explore the data and see how the sales are with respect to each of the variables.

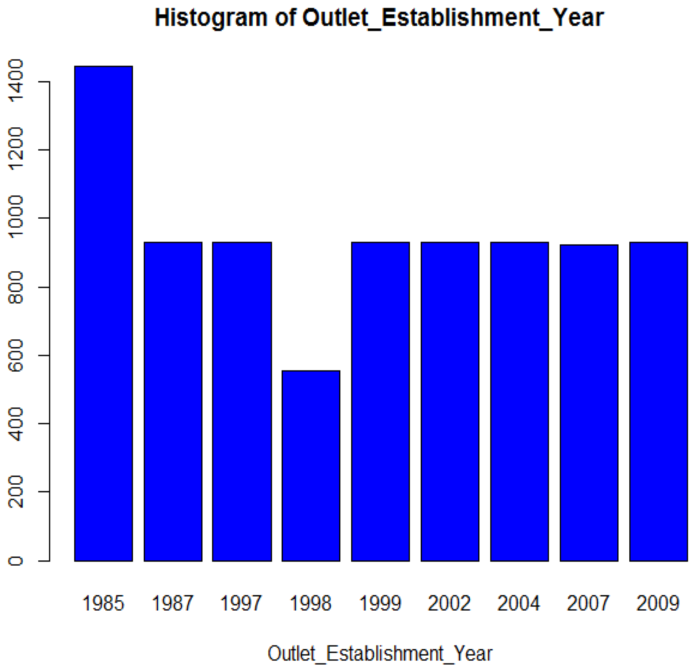
Item’s Fat Content



The Median for both Low Fat and Regular are approximately 40 of the sqrt(Item_Outlet_Sales).

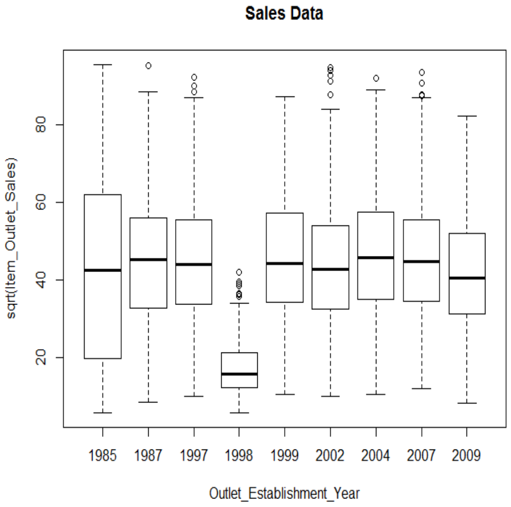


Item’s Establishment Year

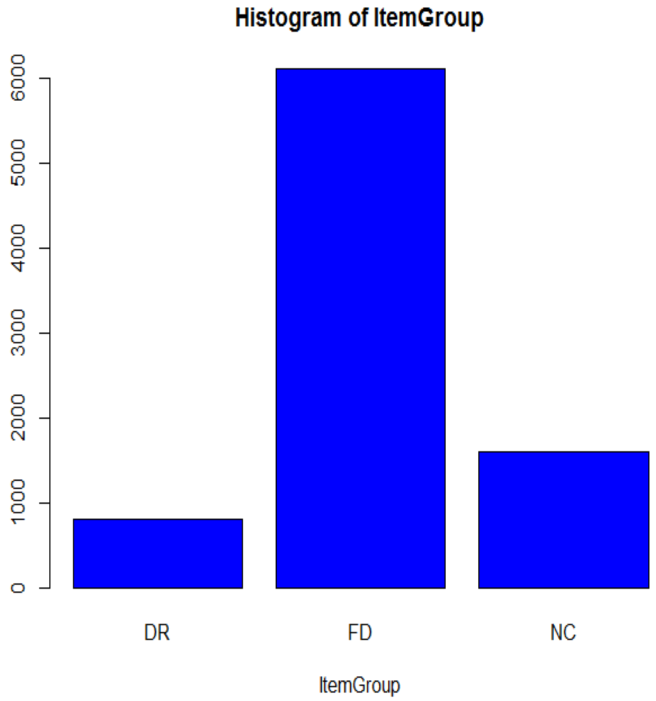


- Year 1998 had the lowest sales among all the years

- Median of rest of the years lie approximately around 40 to 45 of sqrt(Item_Outlet_Sales)

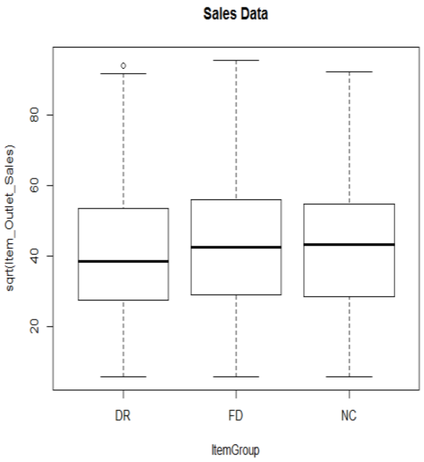


Item Group

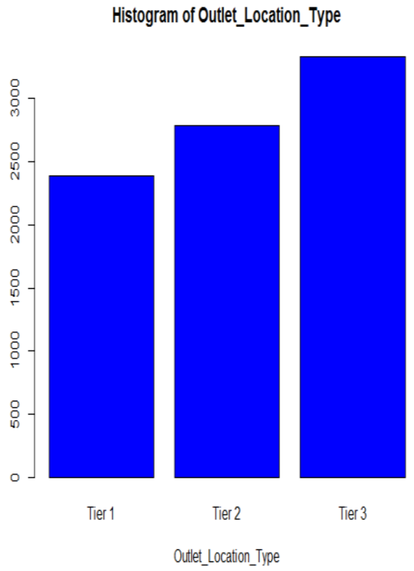


-Histogram of Item Group showed that Food has the highest frequency which is followed by Non-Consumable items and then Drinks.

- Median of Item Group lies approximately around 40 to 45 of sqrt(Item_Outlet_Sales)

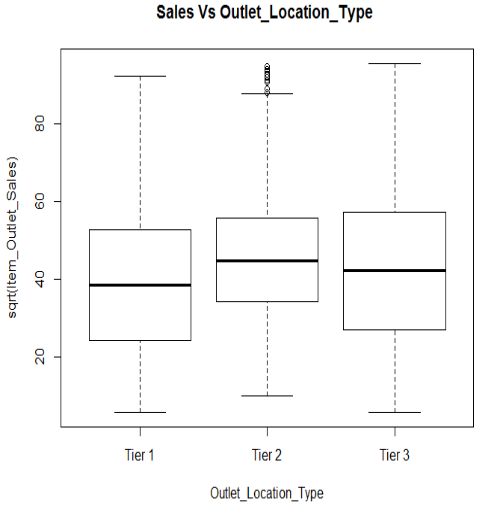


Outlet Location Type - Tiers

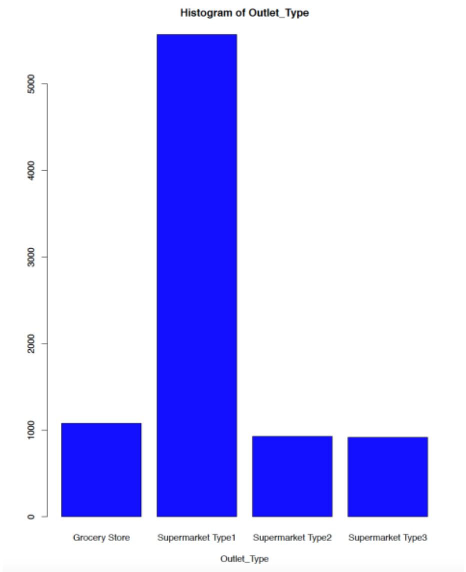


- From histogram, the count of different locations of outlet stores in the data follows a pattern from being lowest in Tier 1 followed by Tier 2 and then Tier 3

- From boxplot, it can be seen as the median value of sales corresponding to Tier1, Tier2 and Tier3 fall between 40 and 45. Moreover, there occurs some outliers pertaining to sales corresponding to Tier 2.



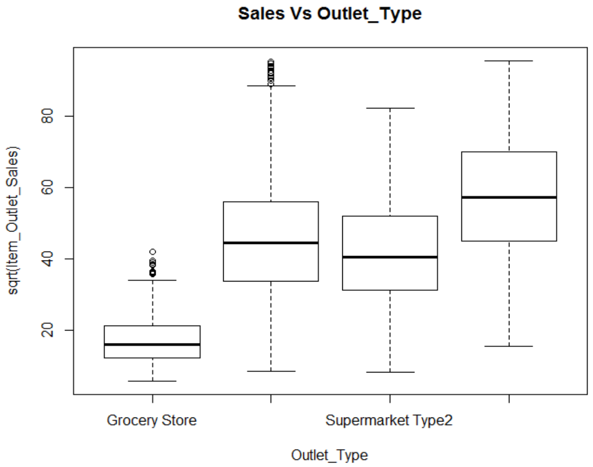
Type of Outlets



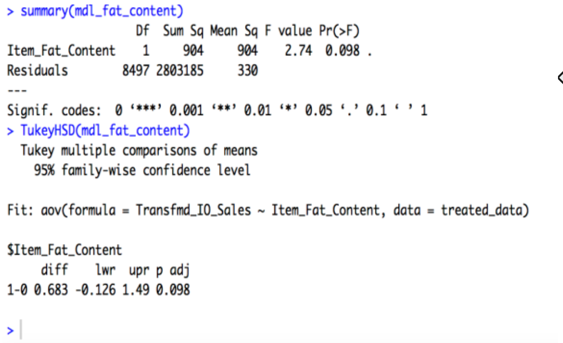
- From histogram, the count of Supermarket Type 1 outlet came out to be significantly higher than that of other three outlet types

- From boxplot, it could be seen that the median sales pertaining to Grocery store is lower than that of other three outlets and lie in range of ~(15-20). Whereas, the median sales of other 3 outlets lie in range of ~(40-60).

- Moreover, outlets corresponding to sales in Grocery store and Supermarket 1 are found.



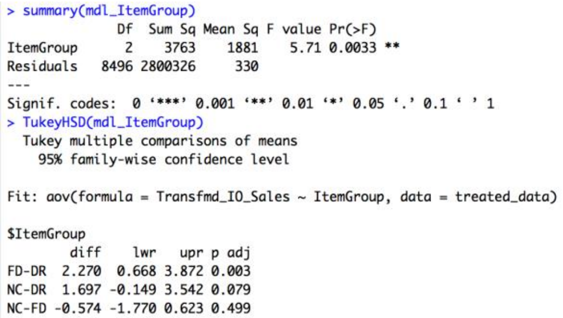
ANOVA - To test if the levels within variables are significantly different



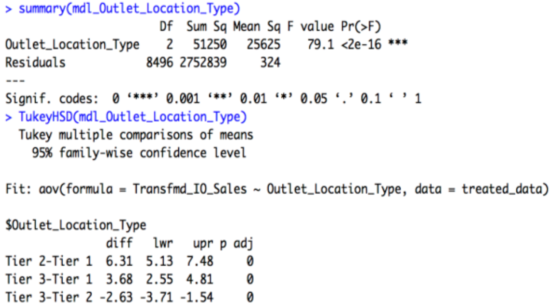
```
> summary(mdl_fat_content)
                  Df  Sum Sq Mean Sq F value Pr(>F)
Item_Fat_Content    1     904     904    2.74  0.098 .
Residuals        8497 2803185     330
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(mdl_fat_content)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Transfmd_IO_Sales ~ Item_Fat_Content, data = treated_data)

$Item_Fat_Content
     diff    lwr  upr p adj
1-0 0.683 -0.126 1.49 0.098

> |
```

At 10% level, the main effects of **Item_Fat_Content** as well as the effects of its different levels are significant towards Item_Outlet_Sales
p = 0.098 (main effect)

At 5% level, the main effect of **ItemGroup** is significant (p-value = 0.0033) as well as for level FD-DR but the effects of its two other levels are insignificant for NC-DR and NC-FD.But, since NC (non consumable) is a broad category, so we decided to keep this at this point and will check in the step of variable selection procedure.



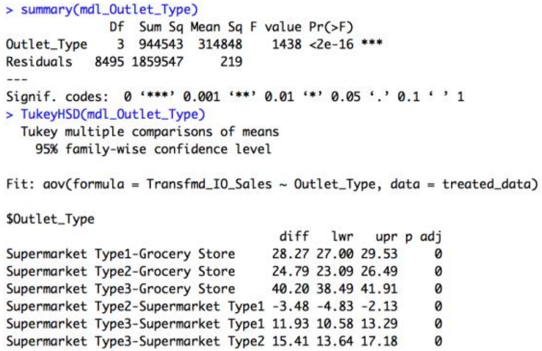
```
> summary(mdl_ItemGroup)
              Df  Sum Sq Mean Sq F value Pr(>F)
ItemGroup      2    3763    1881    5.71 0.0033 **
Residuals   8496 2800326     330
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(mdl_ItemGroup)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Transfmd_IO_Sales ~ ItemGroup, data = treated_data)

$ItemGroup
        diff    lwr   upr p adj
FD-DR  2.270  0.668 3.872 0.003
NC-DR  1.697 -0.149 3.542 0.079
NC-FD -0.574 -1.770 0.623 0.499
```



```
> summary(mdl_Outlet_Location_Type)
                       Df  Sum Sq Mean Sq F value Pr(>F)
Outlet_Location_Type    2   51250   25625    79.1 <2e-16 ***
Residuals            8496 2752839     324
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(mdl_Outlet_Location_Type)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Transfmd_IO_Sales ~ Outlet_Location_Type, data = treated_data)

$Outlet_Location_Type
                 diff   lwr   upr p adj
Tier 2-Tier 1  6.31  5.13  7.48     0
Tier 3-Tier 1  3.68  2.55  4.81     0
Tier 3-Tier 2 -2.63 -3.71 -1.54     0
```

At 5% level, the main effects as well as individual effects of **Outlet_Location_Type** are significant towards its contribution to Item_Outlet_Sales

At 5% level, the main effects of **Outlet_Type** as well as the effects of its different levels are significant towards Item_Outlet_Sales



```
> summary(mdl_Outlet_Type)
              Df  Sum Sq Mean Sq F value Pr(>F)
Outlet_Type    3  944543  314848    1438 <2e-16 ***
Residuals   8495 1859547     219
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(mdl_Outlet_Type)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Transfmd_IO_Sales ~ Outlet_Type, data = treated_data)

$Outlet_Type
                                  diff   lwr   upr p adj
Supermarket Type1-Grocery Store     28.27 27.00 29.53     0
Supermarket Type2-Grocery Store     24.79 23.09 26.49     0
Supermarket Type3-Grocery Store     40.20 38.49 41.91     0
Supermarket Type2-Supermarket Type1 -3.48 -4.83 -2.13     0
Supermarket Type3-Supermarket Type1 11.93 10.58 13.29     0
Supermarket Type3-Supermarket Type2 15.41 13.64 17.18     0
```

REGRESSORS

- Item_Fat_Content: Categorical variable including Low Fat and Regular
- Item_Visibility: Quantitative variable
- Item_Group: Categorical variable including DR, FD & NC and are transposed into columns
- Item_MRP: Quantitative variable
- Outlet_Size: Categorical variable including Small, Medium, High and are transposed into columns
- Outlet_Establishment_Year: Categorical variable from 1985 to 2009. These are also transposed into columns
- Outlet_Type: Categorical variable including Grocery Store and Supermarket Type1, Supermarket Type2 and Supermarket Type3 and are transposed into columns

Dependent/Predictor Variable: Item_Outlet_Sales

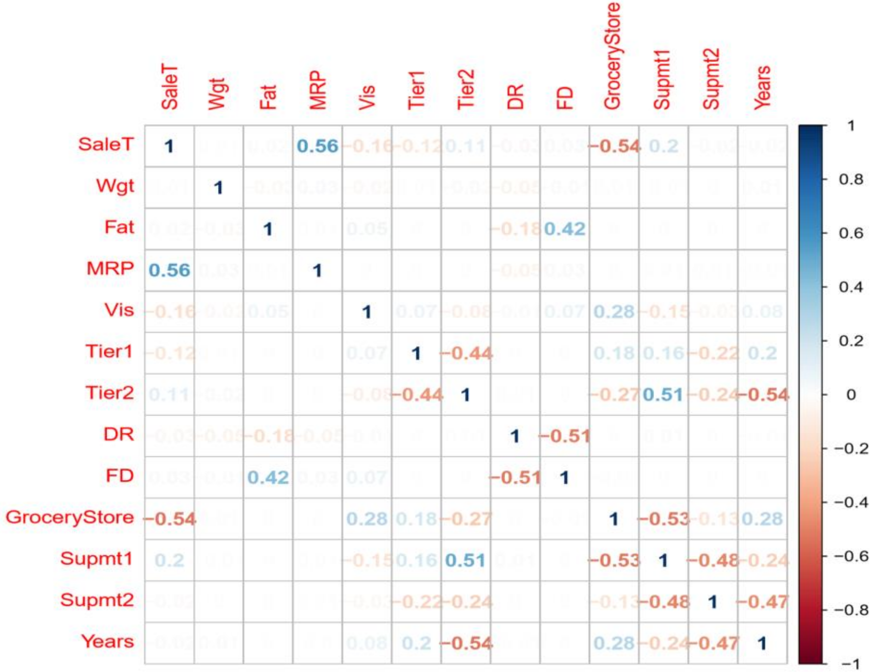
SPLITTING DATA INTO TRAIN AND TEST DATA:

Training data -> 90%

Test data -> 10%

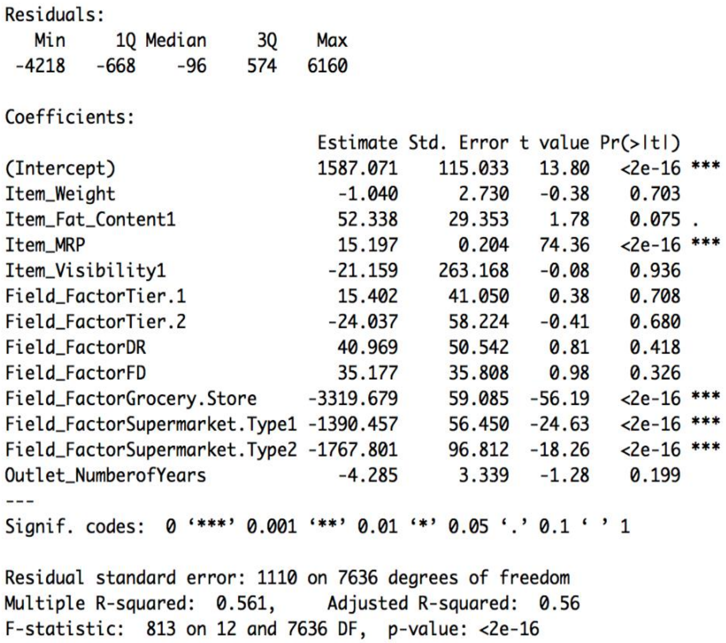
CORRELATION ANALYSIS

How are the regressor and the response variables correlated to each other ?



- ❖ The response variable "Sales" is correlated significantly with the MRP of items and the type of outlets,
- ❖ But,it has not shown any significant correlation with number of years of outlet establishment
- ❖ DR and FD has high correlation among each other.
- ❖ The outlet location type is correlated with outlet type.
- ❖ Out of these regressor variable pairs which has significant correlation , we need to consider only 1 variable out of each pair.

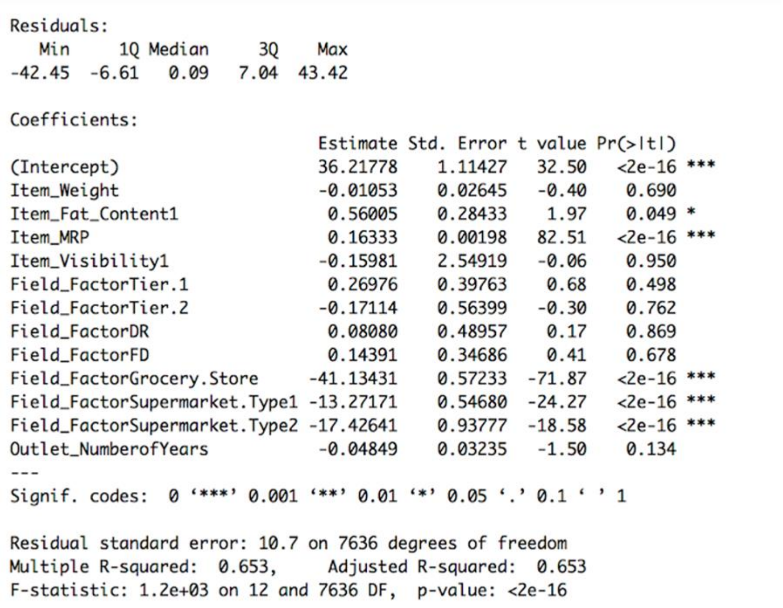
1. MLR Model - Full model with all regressors without transformed Item Outlet Sales



```
Residuals:
   Min     1Q Median     3Q    Max 
 -4218   -668    -96    574   6160 

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    1587.071    115.033   13.80   <2e-16 ***
Item_Weight                      -1.040      2.730   -0.38    0.703    
Item_Fat_Content1                52.338     29.353    1.78    0.075 .  
Item_MRP                         15.197      0.204   74.36   <2e-16 ***
Item_Visibility1                -21.159    263.168   -0.08    0.936    
Field_FactorTier.1               15.402     41.050    0.38    0.708    
Field_FactorTier.2              -24.037     58.224   -0.41    0.680    
Field_FactorDR                   40.969     50.542    0.81    0.418    
Field_FactorFD                   35.177     35.808    0.98    0.326    
Field_FactorGrocery.Store     -3319.679     59.085  -56.19   <2e-16 ***
Field_FactorSupermarket.Type1 -1390.457     56.450  -24.63   <2e-16 ***
Field_FactorSupermarket.Type2 -1767.801     96.812  -18.26   <2e-16 ***
Outlet_NumberofYears             -4.285      3.339   -1.28    0.199    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1110 on 7636 degrees of freedom
Multiple R-squared:  0.561,	Adjusted R-squared:  0.56 
F-statistic:  813 on 12 and 7636 DF,  p-value: <2e-16
```

- Predictor variables used: All 11 variables
- Adjusted R square: 56 %
- But, based on p-values of the individual variables, few of them are insignificant.

2. MLR Model - Full model with all regressors and transformed Item Outlet sales



```
Residuals:
   Min     1Q Median     3Q    Max 
-42.45  -6.61   0.09   7.04  43.42 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   36.21778    1.11427   32.50   <2e-16 ***
Item_Weight                   -0.01053    0.02645   -0.40    0.690    
Item_Fat_Content1              0.56005    0.28433    1.97    0.049 *  
Item_MRP                       0.16333    0.00198   82.51   <2e-16 ***
Item_Visibility1              -0.15981    2.54919   -0.06    0.950    
Field_FactorTier.1             0.26976    0.39763    0.68    0.498    
Field_FactorTier.2            -0.17114    0.56399   -0.30    0.762    
Field_FactorDR                 0.08080    0.48957    0.17    0.869    
Field_FactorFD                 0.14391    0.34686    0.41    0.678    
Field_FactorGrocery.Store    -41.13431    0.57233  -71.87   <2e-16 ***
Field_FactorSupermarket.Type1 -13.27171   0.54680  -24.27   <2e-16 ***
Field_FactorSupermarket.Type2 -17.42641   0.93777  -18.58   <2e-16 ***
Outlet_NumberofYears          -0.04849    0.03235   -1.50    0.134    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.7 on 7636 degrees of freedom
Multiple R-squared:  0.653,	Adjusted R-squared:  0.653 
F-statistic: 1.2e+03 on 12 and 7636 DF,  p-value: <2e-16
```

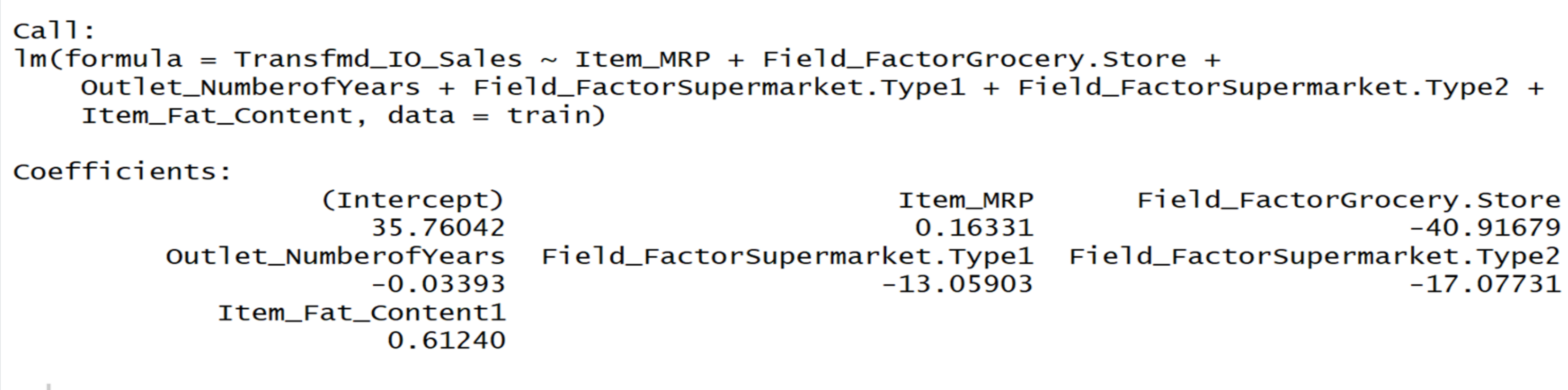
- Predictor variables used: All 11 variables
- But in this case, the response variable used is the transformed Item_Outlet_Sales.
- Adjusted R square increased from 56 to 65.3 %
- But, based on p-values of the individual variables, few of them are insignificant.

SELECTION PROCEDURES

1. Forward, backward, stepwise selection

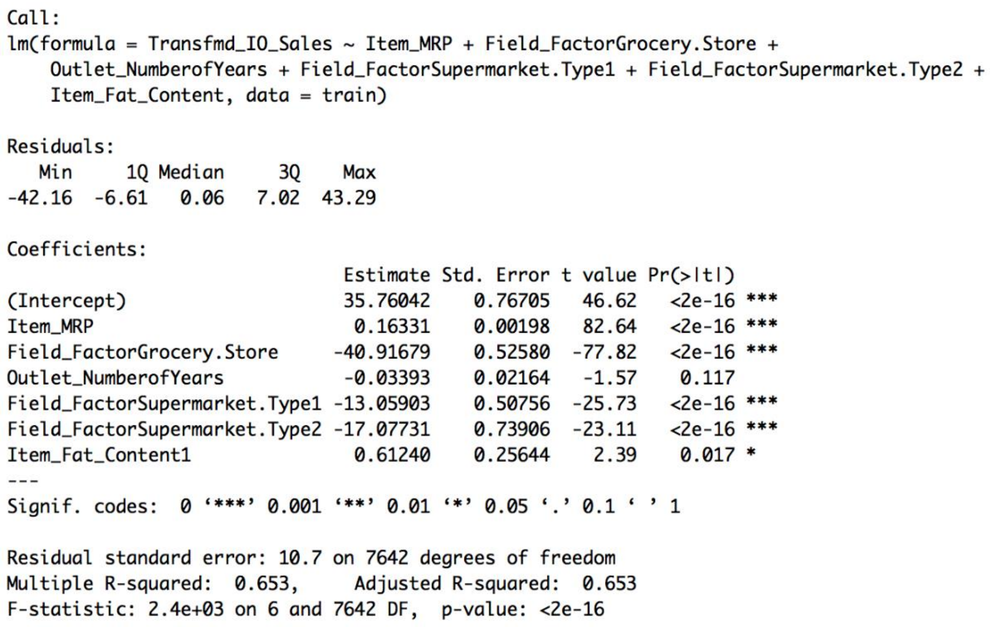
We have performed Forward, Backward and Stepwise Regression procedures and all three have identified the same regressors -

Item_MRP,
Item Fat content,
Outlet Establishment no. of Years
Supermarket Type1
Supermarket Type 2
Grocery Store



```
Call:
lm(formula = Transfmd_IO_Sales ~ Item_MRP + Field_FactorGrocery.Store +
    Outlet_NumberofYears + Field_FactorSupermarket.Type1 + Field_FactorSupermarket.Type2 +
    Item_Fat_Content, data = train)

Coefficients:
                  (Intercept)                       Item_MRP      Field_FactorGrocery.Store
                     35.76042                        0.16331                      -40.91679
         Outlet_NumberofYears  Field_FactorSupermarket.Type1  Field_FactorSupermarket.Type2
                     -0.03393                      -13.05903                      -17.07731
            Item_Fat_Content1
                      0.61240
```

3. MLR Model - With all selected regressors (mdl1 in code)



```
Call:
lm(formula = Transfmd_IO_Sales ~ Item_MRP + Field_FactorGrocery.Store +
    Outlet_NumberofYears + Field_FactorSupermarket.Type1 + Field_FactorSupermarket.Type2 +
    Item_Fat_Content, data = train)

Residuals:
   Min      1Q Median      3Q     Max
-42.16  -6.61   0.06   7.02  43.29

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   35.76042    0.76705   46.62   <2e-16 ***
Item_MRP                       0.16331    0.00198   82.64   <2e-16 ***
Field_FactorGrocery.Store    -40.91679    0.52580  -77.82   <2e-16 ***
Outlet_NumberofYears          -0.03393    0.02164   -1.57    0.117
Field_FactorSupermarket.Type1 -13.05903    0.50756  -25.73   <2e-16 ***
Field_FactorSupermarket.Type2 -17.07731    0.73906  -23.11   <2e-16 ***
Item_Fat_Content1              0.61240    0.25644    2.39    0.017 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.7 on 7642 degrees of freedom
Multiple R-squared:  0.653,     Adjusted R-squared:  0.653
F-statistic: 2.4e+03 on 6 and 7642 DF,  p-value: <2e-16
```

- Predictor variables used: 6 variables
- But in this case, 6 regressors have been selected from the selection procedures.
- Adjusted R square: 65.3 % and p-value is very low for the entire model.
- But, based on p-values of the individual variables, most of them are significant at 5% level, except for Outlet_Establishment_Year with the higher p-value than the level of significance (5%), which makes it insignificant.

SELECTION PROCEDURES

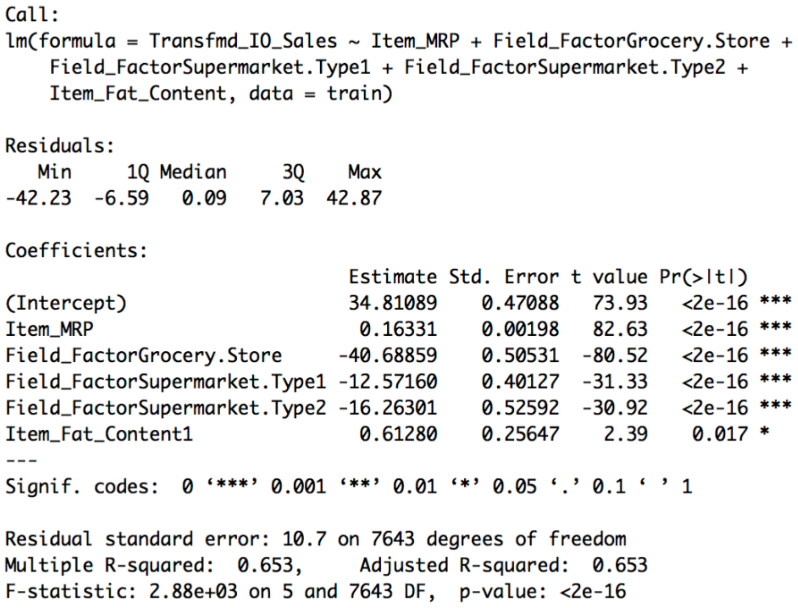
2. **All possible subset models (2 ^ 12).** We got **2 final models**:

- 6 regressors : (which was same as the selection procedure)
  - Item_MRP
  - Field_FactorGrocery.Store
  - Outlet_NumberofYears
  - Field_FactorSupermarket.Type1
  - Field_FactorSupermarket.Type2
  - Item_Fat_Content

- 5 regressors
  - Item_MRP
  - Field_FactorGrocery.Store
  - Field_FactorSupermarket.Type1
  - Field_FactorSupermarket.Type2
  - Item_Fat_Content

On running the MLR we saw Outlet_NumberofYears was coming insignificant (p-value = 0.117) and it didn't showed any strong correlation with response variable also in correlation matrix. Moreover, it was correlated with one of other regressor (Field_FactorGrocery.Store).

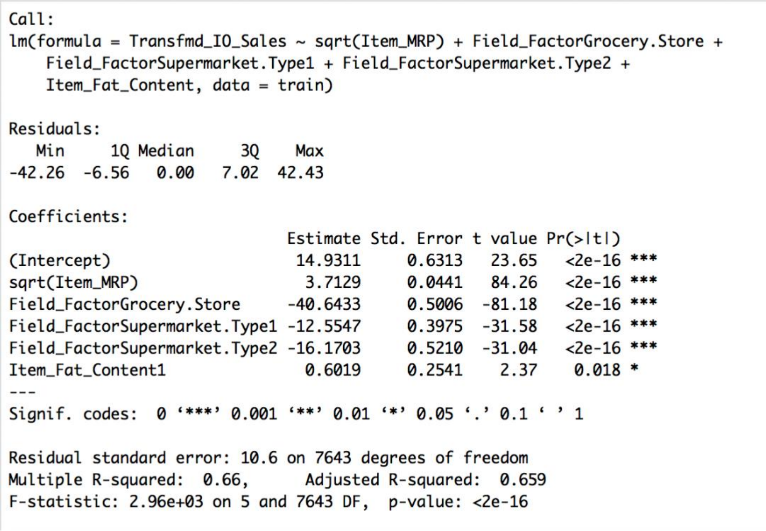
4. MLR Model - Removing year variable (mdl2 in code)



```
Call:
lm(formula = Transfmd_IO_Sales ~ Item_MRP + Field_FactorGrocery.Store +
    Field_FactorSupermarket.Type1 + Field_FactorSupermarket.Type2 +
    Item_Fat_Content, data = train)

Residuals:
   Min      1Q Median      3Q     Max
-42.23  -6.59   0.09    7.03   42.87

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   34.81089    0.47088   73.93   <2e-16 ***
Item_MRP                       0.16331    0.00198   82.63   <2e-16 ***
Field_FactorGrocery.Store    -40.68859    0.50531  -80.52   <2e-16 ***
Field_FactorSupermarket.Type1 -12.57160    0.40127  -31.33   <2e-16 ***
Field_FactorSupermarket.Type2 -16.26301    0.52592  -30.92   <2e-16 ***
Item_Fat_Content1              0.61280    0.25647    2.39    0.017 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.7 on 7643 degrees of freedom
Multiple R-squared:  0.653,     Adjusted R-squared:  0.653
F-statistic: 2.88e+03 on 5 and 7643 DF,  p-value: <2e-16
```

- Predictor variables used: Reduced to 5 variables
- Outlet_NumberofYears was removed.
- Adjusted R square: 65.24 %
- Based on p-values of the individual variables, all of them are significant at 5% level.

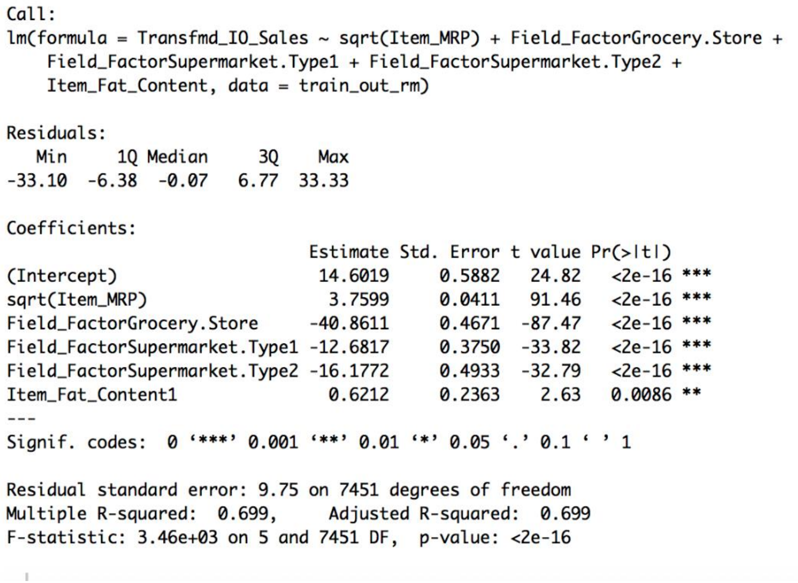
5. MLR Model - Transform MRP (mdl3 in code)



```
Call:
lm(formula = Transfmd_IO_Sales ~ sqrt(Item_MRP) + Field_FactorGrocery.Store +
    Field_FactorSupermarket.Type1 + Field_FactorSupermarket.Type2 +
    Item_Fat_Content, data = train)

Residuals:
   Min     1Q Median     3Q    Max
-42.26  -6.56   0.00   7.02  42.43

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    14.9311     0.6313   23.65   <2e-16 ***
sqrt(Item_MRP)                  3.7129     0.0441   84.26   <2e-16 ***
Field_FactorGrocery.Store     -40.6433     0.5006  -81.18   <2e-16 ***
Field_FactorSupermarket.Type1 -12.5547     0.3975  -31.58   <2e-16 ***
Field_FactorSupermarket.Type2 -16.1703     0.5210  -31.04   <2e-16 ***
Item_Fat_Content1               0.6019     0.2541    2.37    0.018 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.6 on 7643 degrees of freedom
Multiple R-squared:  0.66,      Adjusted R-squared:  0.659
F-statistic: 2.96e+03 on 5 and 7643 DF,  p-value: <2e-16
```

- Predictor variables used: Reduced to 5 variables
- The Item_MRP was transformed to fit the regression line better. It was done through **square root transformation**.
- Adjusted R square: 65.93 %
- Based on p-values of the individual variables, all of them are significant at 5% level.

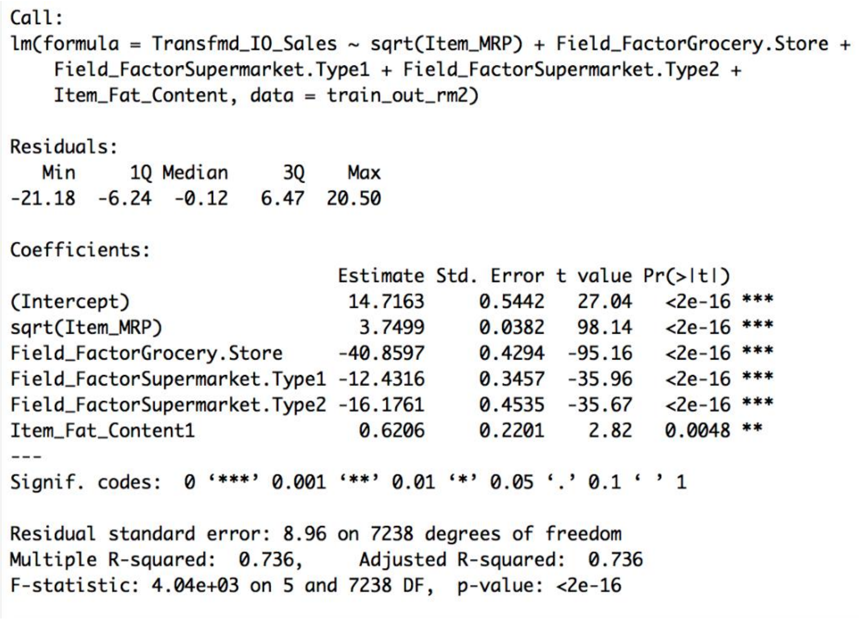
6. MLR Model -Outlier Removal Using Both R-Student & DFFITS (mdl4 in code)



```
Call:
lm(formula = Transfmd_IO_Sales ~ sqrt(Item_MRP) + Field_FactorGrocery.Store +
    Field_FactorSupermarket.Type1 + Field_FactorSupermarket.Type2 +
    Item_Fat_Content, data = train_out_rm)

Residuals:
   Min     1Q Median     3Q    Max 
-33.10  -6.38  -0.07   6.77  33.33 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    14.6019     0.5882   24.82   <2e-16 ***
sqrt(Item_MRP)                  3.7599     0.0411   91.46   <2e-16 ***
Field_FactorGrocery.Store     -40.8611     0.4671  -87.47   <2e-16 ***
Field_FactorSupermarket.Type1 -12.6817     0.3750  -33.82   <2e-16 ***
Field_FactorSupermarket.Type2 -16.1772     0.4933  -32.79   <2e-16 ***
Item_Fat_Content1               0.6212     0.2363    2.63   0.0086 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.75 on 7451 degrees of freedom
Multiple R-squared:  0.699,	Adjusted R-squared:  0.699 
F-statistic: 3.46e+03 on 5 and 7451 DF,  p-value: <2e-16
```

- Predictor variables used: Reduced to 5 variables
- The outliers were removed using the Rstudent and DFFITS criteria.
- Datapoints which were outliers according to both Rstudent and DFFITS criteria were removed.
- Adjusted R square: 69.9 % (increased slightly from the previous model).
- Based on p-values of the individual variables, all of them are significant at 5% level.

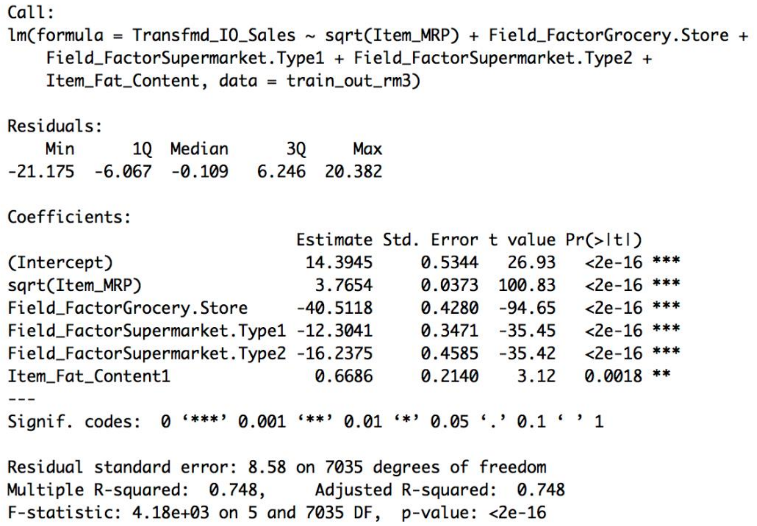
7. MLR Model - Outlier Removal Using R-Student (mdl5 in code)



```
Call:
lm(formula = Transfmd_IO_Sales ~ sqrt(Item_MRP) + Field_FactorGrocery.Store +
    Field_FactorSupermarket.Type1 + Field_FactorSupermarket.Type2 +
    Item_Fat_Content, data = train_out_rm2)

Residuals:
   Min     1Q Median     3Q    Max
-21.18  -6.24  -0.12   6.47  20.50

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    14.7163     0.5442   27.04   <2e-16 ***
sqrt(Item_MRP)                  3.7499     0.0382   98.14   <2e-16 ***
Field_FactorGrocery.Store     -40.8597     0.4294  -95.16   <2e-16 ***
Field_FactorSupermarket.Type1 -12.4316     0.3457  -35.96   <2e-16 ***
Field_FactorSupermarket.Type2 -16.1761     0.4535  -35.67   <2e-16 ***
Item_Fat_Content1               0.6206     0.2201    2.82   0.0048 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.96 on 7238 degrees of freedom
Multiple R-squared:  0.736,	Adjusted R-squared:  0.736
F-statistic: 4.04e+03 on 5 and 7238 DF,  p-value: <2e-16
```

- Predictor variables used: Reduced to 5 variables
- The outliers were removed using only the Rstudent criteria.
- Adjusted R square: 73.6 % (increased from the previous model).
- Based on p-values of the individual variables, all of them are significant at 5% level.

8. MLR Model - Using Either R student OR DFFITS (mdl6 in code)



```
Call:
lm(formula = Transfmd_IO_Sales ~ sqrt(Item_MRP) + Field_FactorGrocery.Store +
    Field_FactorSupermarket.Type1 + Field_FactorSupermarket.Type2 +
    Item_Fat_Content, data = train_out_rm3)

Residuals:
    Min      1Q  Median      3Q     Max
-21.175  -6.067  -0.109   6.246  20.382

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    14.3945     0.5344   26.93   <2e-16 ***
sqrt(Item_MRP)                  3.7654     0.0373  100.83   <2e-16 ***
Field_FactorGrocery.Store     -40.5118     0.4280  -94.65   <2e-16 ***
Field_FactorSupermarket.Type1 -12.3041     0.3471  -35.45   <2e-16 ***
Field_FactorSupermarket.Type2 -16.2375     0.4585  -35.42   <2e-16 ***
Item_Fat_Content1               0.6686     0.2140    3.12   0.0018 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.58 on 7035 degrees of freedom
Multiple R-squared:  0.748,	Adjusted R-squared:  0.748
F-statistic: 4.18e+03 on 5 and 7035 DF,  p-value: <2e-16
```

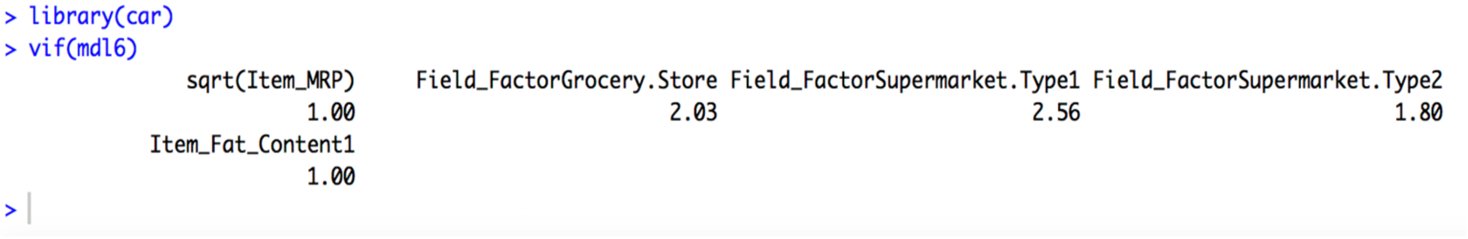
- Predictor variables used: Reduced to 5 variables
- The outliers were removed from the dataset based on either the Rstudent criteria or DFFITS criteria.
- Adjusted R square: 74.8 % (**highest adj-R square** compared to all the models).
- Based on p-values of the individual variables, **all of regressors are significant at 5% level.**

Selecting Final Model

- We tried different approaches to enhance the adj-R square of the model like removing second round of outliers and then re-running the model but the adj-R square was not increasing significantly.
- Based on the adj-R square values, we chose our final model as Model 8 (Using Either R student OR DFFITS (mdl6 in code)) which has the highest adj- R2 value as 74.8% compared to the other models.

VIF (Variance Inflation Factor)

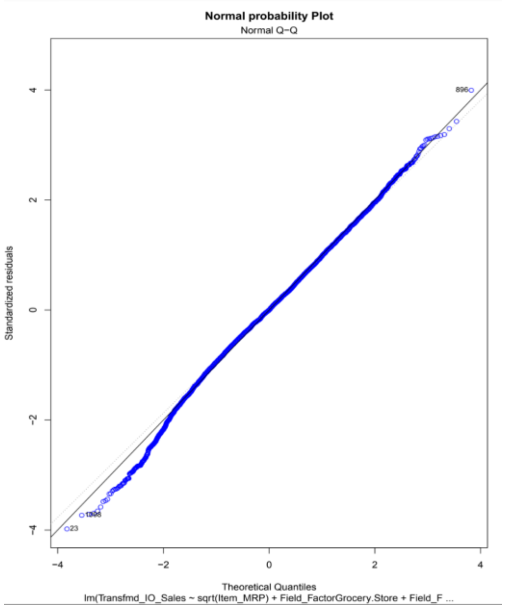
- The model also showed the all the Variance coeff < 5 ==> there is low variance(VIF)
- Conclusion: There is no problem and variance inflation factor is under control for the normal fit



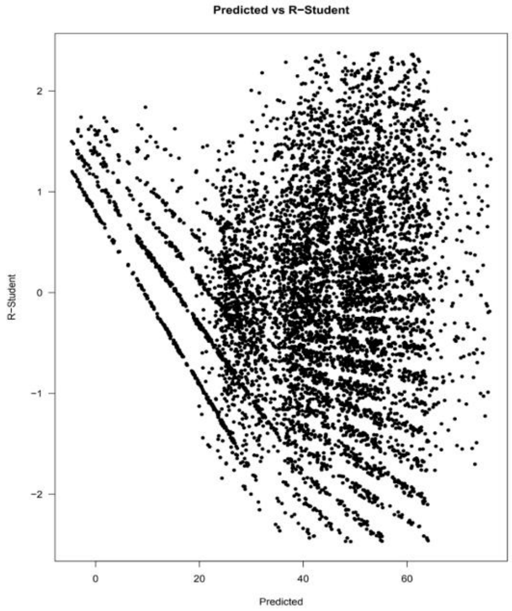
```
> library(car)
> vif(mdl6)
            sqrt(Item_MRP)     Field_FactorGrocery.Store Field_FactorSupermarket.Type1 Field_FactorSupermarket.Type2
                      1.00                          2.03                          2.56                          1.80
         Item_Fat_Content1
                      1.00
> |
```

Residual Analysis

Normal Probability plot

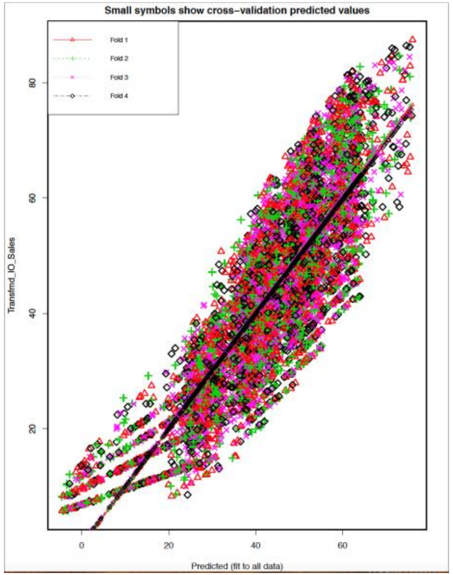


Residual Vs. Predicted Plot



Normal Probability plot showed that the data is almost linear. We can also conclude from the residual vs. predicted plot that the data has nearly constant variance.

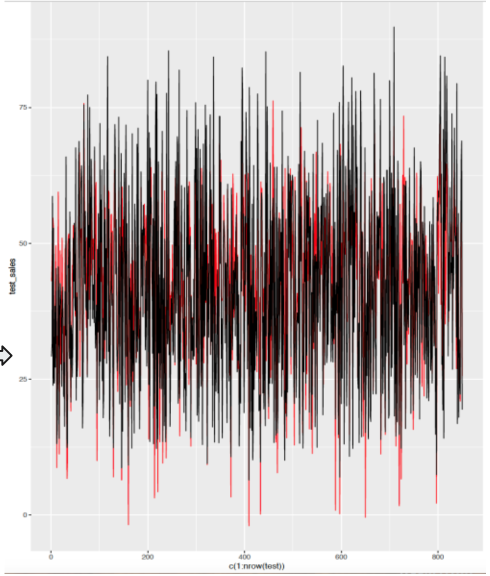
Cross Validation of Final Model



4 fold cross-validation

4 fold cross-validation showing that all 4 folds have almost similar predictions with constant variance.

On comparing the sales prediction with actual sales of test data, there seems to be a good overlap. This should be a good model to propose.



ggplot:
Actual and Predicted sales comparison

Metrics for Model Evaluation

Metrics	Test Data	Train Data
Multiple R-Square	66%	74.8%
Root Mean Square Error(RMSE)	10.4	8.58
Mean Absolute error(MAE)	8.18	7

The corresponding metric values of train and test data are close enough.

CONCLUSION

Final Model:

Transformed Item Outlet Sales = 14.3945 + 3.7654**Square root of Item MRP* - 40.5118**Grocery Store* - 12.3041**Supermarket Type1* - 16.2375**Supermarket Type2* + 0.6686**Item Fat Content*

The final model is the best working model because:

- Only 5 regressors were sufficient to predict the item outlet sales, these regressors were MRP, type of outlet: grocery, supermarket 1 & supermarket 2 and item fat content
- It has all significant regressors at 5% level
- It has the highest adj-R2 of 74.8%
- It has the lowest RMSE = 8.58
- When this model was validated with test data it gave the RMSE of 10.4 which was close enough with the training model

Thank you!

References:

DataSource - https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/

https://www.datacamp.com/community/tutorials/make-histogram-basic-r

https://www.cscu.cornell.edu/news/statnews/stnews68.pdf

https://stats.stackexchange.com/questions/56302/what-are-good-rmse-values

https://s3.amazonaws.com/assets.datacamp.com/production/course_3851/slides/chapter2.pdf

http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software

https://www.rdocumentation.org/packages/SimDesign/versions/1.13/topics/RMSE

https://stackoverflow.com/questions/26237688/rmse-root-mean-square-deviation-calculation-in-r/26237921

https://r4ds.had.co.nz/exploratory-data-analysis.html

http://www.sthda.com/english/articles/38-regression-model-validation/157-cross-validation-essentials-in-r/