Big Mart sales Data Report

Customer Data Analysis

Vignesh Hariharan

Contents

1 Introduction

I chose Big Mart sales prediction data as my data set. The data is from Analyticsvidhya.com. In this part, I will introduce the company briefly and the data set. Then you can understand what I have done in the subsequent part.

Big Mart is a departmental and convenience store retail chain. They stock an expansive range of daily need items including groceries, candies, personal care products, soft drinks, ready-to-eat food, ice-cream, toiletries, tobacco products, magazines, and newspapers etc. Besides basic everyday items, they also offer additional services like phone recharge and wire transfer.

The Big Mart Started small with a single store in April 2007, today Big Mart has become a Ill-known name in the retail industry with more than 40 outlets in world. With our unmatched services, attractive prices and high-quality products, they are touched the lives of young customers.

Big Mart Provide:

1. Grocery Products
2. Dairy Products
3. Organic Foods
4. Bakery Foods
5. Frozen Food
6. Free Home Delivery

Customer satisfaction and value for the stakeholders are our major goals which I try to fulfill with our four pillars: quality of the products, the speed of service, rewarding experience, and environmental responsibility.

We got two excel files of Big Mart Sales. The one is training dataset and the file name is "train". In the training data. In the training data I have sales There are some other columns to indicate the status of stores which are:

- Bigmart dataset has 8524 rows and 12 columns.
- Item_identifier - describes the code of each item and it is Unique product ID.
- Item_light - explains the light of each item.
- Item_Fat_Content – whether each item has low fat or regular fat.
- Item_Visibility – The % of total display area of all products in store allocated to the product.
- Item_type – the category to which the product belongs.
- Item_MRP – Maximum retail price of the product.
- Outlet_identifier - is a Unique store ID.
- Outlet_establishment_year - is the year in which store was established.
- Outlet_Size - The size of the store in terms of ground area covered.

- Outlet_Location_Type - The type of city in which the store is located.
- Outlet_Type - Whether the outlet is just a grocery store or some sort of supermarket.
- Item_Outlet_Sales - Sales of the product in the particulat store. This is the outcome variable to be predicted.

2. Manipulating Dataset

```
> big_mart_imputed <- big_mart_imputed %>%
+   select(Item_Identifier:Item_Outlet_Sales)
> summary(big_mart_imputed)
 Item_Identifier  Item_Weight     Item_Fat_Content Item_Visibility                  Item_Type        Item_MRP       Outlet_Ide
ntifier
 FDG33  :  10    Min.   : 0.00   Low Fat:5517     Min.   :0.003575   Fruits and Vegetables:1232   Min.   : 31.29   OUT027 : 9
35
 FDW13  :  10    1st Qu.: 6.65   Regular:3006     1st Qu.:0.031228   Snack Foods          :1200   1st Qu.: 93.83   OUT013 : 9
32
 DRE49  :   9    Median :11.00                    Median :0.057249   Household            : 910   Median :143.01   OUT035 : 9
30
 DRN47  :   9    Mean   :10.65                    Mean   :0.069941   Frozen Foods         : 856   Mean   :140.99   OUT046 : 9
30
 FDD38  :   9    3rd Qu.:16.00                    3rd Qu.:0.097383   Dairy                : 682   3rd Qu.:185.64   OUT049 : 9
30
 FDF52  :   9    Max.   :21.35                    Max.   :0.328391   Canned               : 649   Max.   :266.89   OUT045 : 9
29
 (Other):8467                                                        (Other)              :2994                    (Other):29
37
 Outlet_Establishment_Year Outlet_Size   Outlet_Location_Type            Outlet_Type    Item_Outlet_Sales
 Min.   :1985              0      :2410  Tier 1:2388          Grocery Store     :1083   Min.   :   33.29
 1st Qu.:1987              High   : 932  Tier 2:2785          Supermarket Type1:5577   1st Qu.:  834.25
 Median :1999              Medium:2793   Tier 3:3350          Supermarket Type2: 928   Median : 1794.33
 Mean   :1998              Small :2388                        Supermarket Type3: 935   Mean   : 2181.29
 3rd Qu.:2004                                                                           3rd Qu.: 3101.30
 Max.   :2009                                                                           Max.   :13086.97
```

In the Item_Fat_Content column there Ire several observations that needed cleaning. All of the content in this column was either **Low Fat** or **Regular**. Holver, some of the observations Ire stored as **LF**, **low fat** or **reg**. The cleaning made sure all observations Ire entered as **Low Fact**or **Regular**.

There Ire also 1463 missing values for the Item_Weight column. These missing values will present problems when trying to create a Machine Learning Model. In this report, kNN imputation was used to impute values for the missing observations. This method imputes a value based on other observations with similar values for the other variables in the dataset.

3. Discovering Way to Impute Values for Outlet_Size

3.1 Outlet Identifier by Outlet Size Table

```
> table(big_mart_imputed$Outlet_Identifier, big_mart_imputed$Outlet_Size)

           0 High Medium Small
  OUT010 555    0      0     0
  OUT013   0  932      0     0
  OUT017 926    0      0     0
  OUT018   0    0    928     0
  OUT019   0    0      0   528
  OUT027   0    0    935     0
  OUT035   0    0      0   930
  OUT045 929    0      0     0
  OUT046   0    0      0   930
  OUT049   0    0    930     0
```

You can see These tables show that there are 10 different Big Mart outlets that are being used in the dataset. Each outlet size is either small, medium or high.

3.2 Outlet Identifier by Outlet_Type Table

```
> table(big_mart_imputed$Outlet_Identifier, big_mart_imputed$Outlet_Type)

         Grocery Store Supermarket Type1 Supermarket Type2 Supermarket Type3
  OUT010           555                 0                 0                 0
  OUT013             0               932                 0                 0
  OUT017             0               926                 0                 0
  OUT018             0                 0               928                 0
  OUT019           528                 0                 0                 0
  OUT027             0                 0                 0               935
  OUT035             0               930                 0                 0
  OUT045             0               929                 0                 0
  OUT046             0               930                 0                 0
  OUT049             0               930                 0                 0
```

You can see the outlet identifier is each outlet type is either Grocery Store, Supermarket Type1, Supermarket Type2 or Supermarket Type3.

3.3 Outlet Type by Outlet Size Table

```
> table(stores.df$Outlet_Type, big_mart_imputed$Outlet_Size)

                       0 High Medium Small
  Grocery Store      555    0      0   528
  Supermarket Type1 1855  932    930  1860
  Supermarket Type2    0    0    928     0
  Supermarket Type3    0    0    935     0
```

The Outlet Type by Outlet Size Table shows that all Grocery Store locations are small. And supermarket Type1 locations is high, medium and small, and supermarket type3 & supermarket Type3 locations is medium.

The Outlet Type by Outlet Size Table shows that all Grocery Store locations are small. Since the OUT010 location is a Grocery Store, all observations that are for this location will have the Outlet_Size variable imputed as Small. Unfortunately, the Outlet Type for both the OUT017 and OUT045 locations are Supermarket Type1. The Outlet Size for Supermarket Type1 locations are either small, medium or high. Since the Outlet Size is only high for one location, in this report, the Outlet Size variable will be set to Small for the OUT017 location and the Outlet Size variable will be set to Medium for the OUT045 location.

3.3 Summary Cleaned Dataset

```
> summary(big_mart_imputed)
 Item_Identifier  Item_Weight     Item_Fat_Content Item_Visibility                   Item_Type        Item_MRP       Outlet_Ide
ntifier
 FDG33  :  10    Min.    : 0.00   Low Fat:5517     Min.    :0.003575   Fruits and Vegetables:1232   Min.    : 31.29   OUT027 : 9
35
 FDW13  :  10    1st Qu.: 6.65    Regular:3006     1st Qu.:0.031228   Snack Foods          :1200   1st Qu.: 93.83   OUT013 : 9
32
 DRE49  :   9    Median :11.00                     Median :0.057249   Household            : 910   Median :143.01   OUT035 : 9
30
 DRN47  :   9    Mean   :10.65                     Mean   :0.069941   Frozen Foods         : 856   Mean   :140.99   OUT046 : 9
30
 FDD38  :   9    3rd Qu.:16.00                     3rd Qu.:0.097383   Dairy                : 682   3rd Qu.:185.64   OUT049 : 9
30
 FDF52  :   9    Max.   :21.35                     Max.    :0.328391   Canned               : 649   Max.    :266.89   OUT045 : 9
29
 (Other):8467                                                          (Other)              :2994                    (Other):29
37
 Outlet_Establishment_Year Outlet_Size   Outlet_Location_Type              Outlet_Type   Item_Outlet_Sales
 Min.    :1985             High  : 932   Tier 1:2388          Grocery Store    :1083   Min.    :   33.29
 1st Qu.:1987             Medium:3722   Tier 2:2785          Supermarket Type1:5577   1st Qu.:  834.25
 Median :1999             Small :3869   Tier 3:3350          Supermarket Type2: 928   Median : 1794.33
 Mean   :1998                                                Supermarket Type3: 935   Mean   : 2181.29
 3rd Qu.:2004                                                                          3rd Qu.: 3101.30
 Max.    :2009                                                                          Max.    :13086.97
```

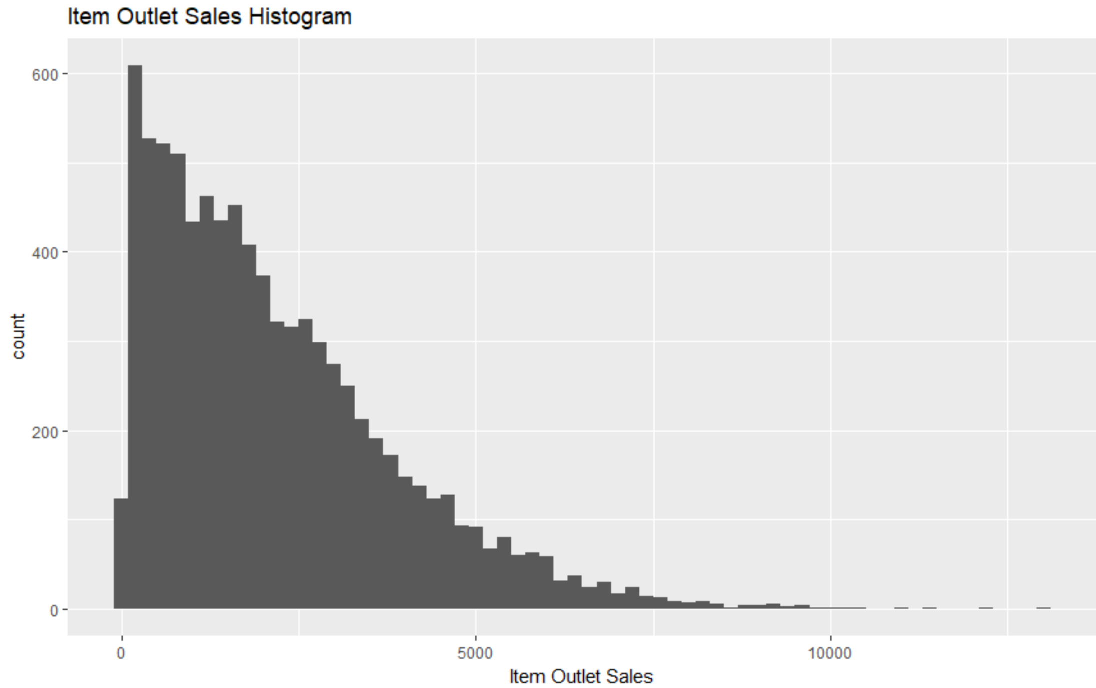
Now you can see All the changes can be seen when comparing the summary of the cleaned dataset with the summary of the original dataset.

4. Descriptive analysis

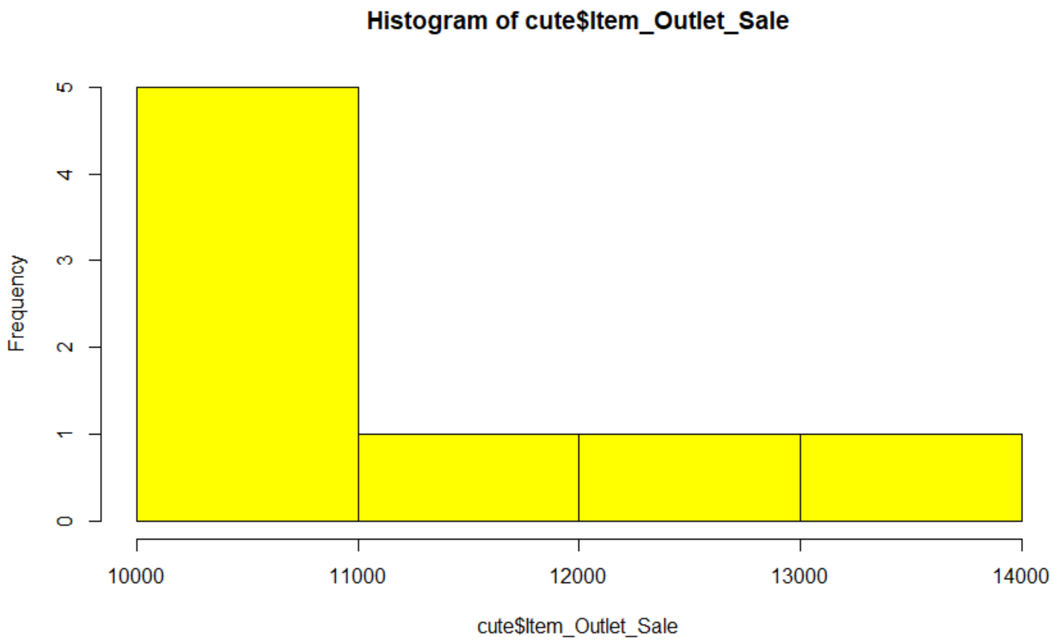
4.1 Item Outlet Sales Histogram

In the training dataset, the average of sales is 2181.29 and the standard deviation is 1706.5. You can see the SD of the sales is a little bit large comparing to the average. And the Max of the sales is 13086 which is very large comparing to the average.

There are many outliers in the training dataset. From the histogram, I can the outliers easily.

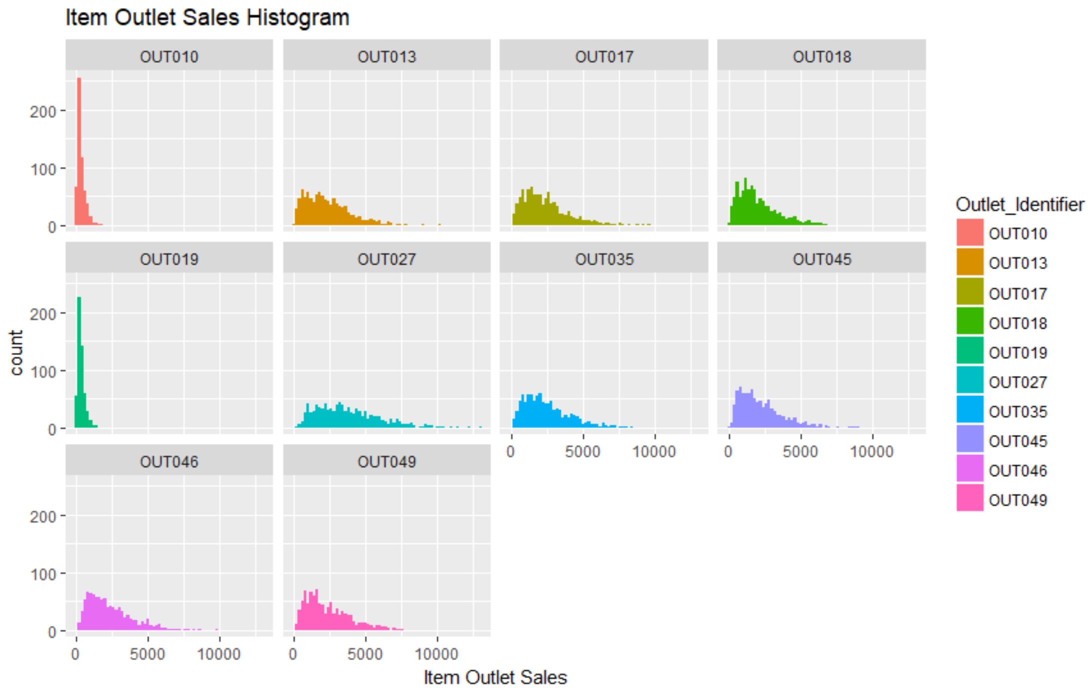


Most of sales are located in the range of (0 to 5000). But some of them are extremely high. The number of observation which the sales is beyond 10000 is 8. The store 909 has maximum sale in the data for Household.



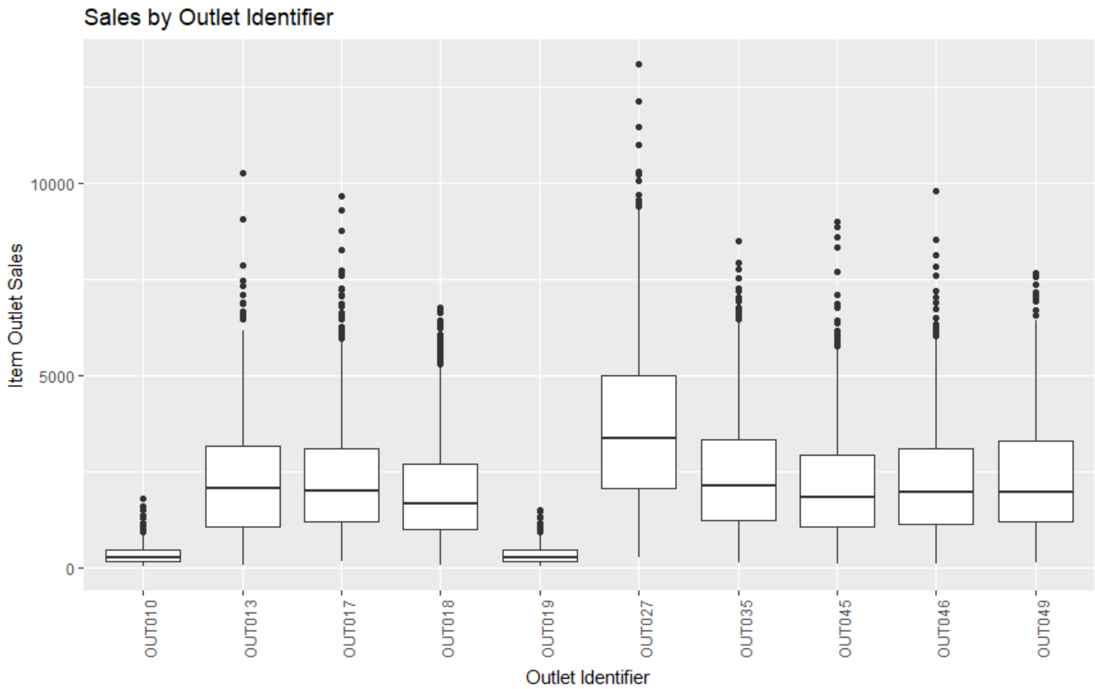
From the histogram of store 909 sales, there are only two extremely high sales which is supermarket type 3 and that Item types is Household and fruits and Vegetables sales high.

4.2 Item Outlet Sales Histogram by Outlet Identifier



The histogram of item outlet sales broken down by Outlet Identifier shows that most of the low item outlet sales Ire in the OUT010 and OUT019 locations. Further examination shows that these two locations Ire the only two locations that Ire Grocery Stores. Therefore, there should be no surprise that they would have the lolst sales.

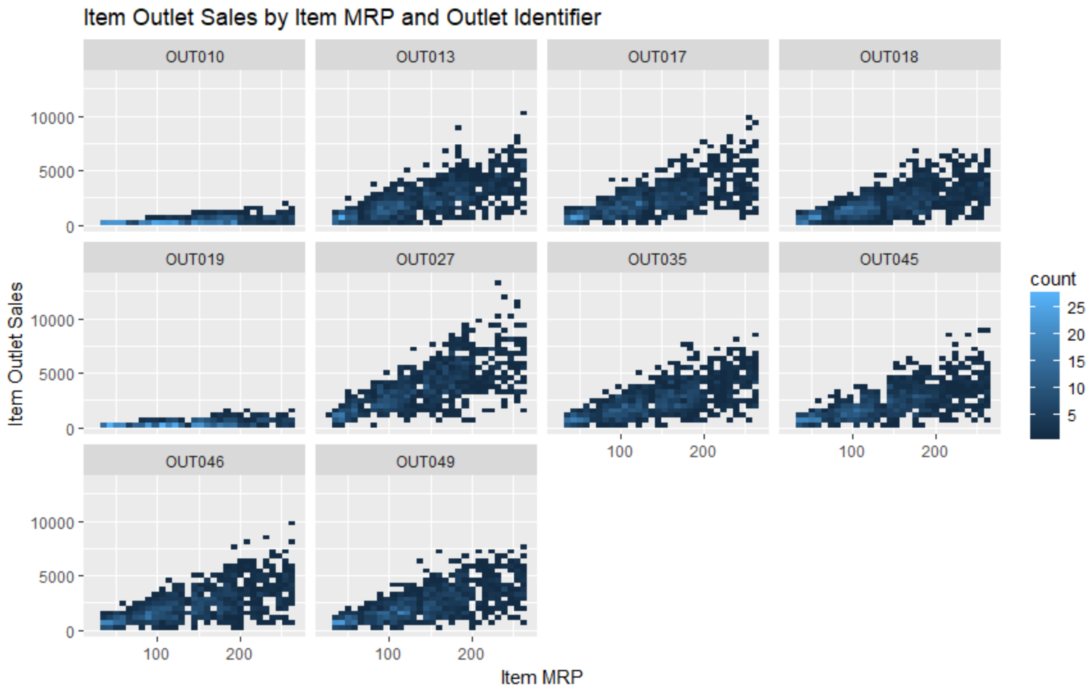
4.3 Sales by Outlet Identifier



The boxplot shows that these two locations had the lolst sales all around. The Outlet that produced the highest sales was the OUT027 location.

Although a person might assume that this outlet was the biggest, its size was only medium. Holver, it was the only outlet that had a Outlet Type of Supermarket Type3. Another item worth noting is that the biggest location was ranked third when looking at median sales by location.

4.4 Item Outlet Sales by Item MRP and Outlet Identifier



You can see this graph, there appears to be a moderate positive correlation betlen Item Outlet Sales and Item MRP.

4.5 Median Sales by Location and Correlation of Item Outlet Sales and Item MRP

```
> #Median Sales by Location
> big_mart_imputed %>%
+   group_by(Outlet_Identifier) %>%
+   summarize(median_sales = median(Item_Outlet_Sales)) %>%
+   arrange(desc(median_sales))
# A tibble: 10 x 2
   Outlet_Identifier median_sales
              <fctr>        <dbl>
 1            OUT027    3364.9532
 2            OUT035    2109.2544
 3            OUT013    2050.6640
 4            OUT017    2005.0567
 5            OUT049    1966.1074
 6            OUT046    1945.8005
 7            OUT045    1834.9448
 8            OUT018    1655.1788
 9            OUT019     265.3213
10            OUT010     250.3408
> cor(big_mart_imputed$Item_MRP, big_mart_imputed$Item_Outlet_Sales)
[1] 0.5675744
>
```

You can see the above picture for assumption is corroborated when running a test for the correlation betIen these two variables. The correlation coefficient of 0.5675744 shows this relationship

5. Conjoint Analysis

In the conjoint analysis, I treat the stores as our products. I have different attributes for the stores. And there are different levels for each attribute. I choose “Item type” And “Outlet Identifier” which are from big mart data set. We did a multiple regression to get the utilities of each level of attributes by using data.

Attribute	Level	Utilities
Item Type	Breads	4.1013
Item Type	Breakfast	5.0694
Item Type	Canned	24.5812
Item Type	Dairy	42.2239
Item Type	Frozen Foods	-27.6101
Item Type	Fruits and Vegetables	29.9455
Item Type	Hard Drinks	-1.2855
Item Type	Health and Hygiene	-9.9106
Item Type	Household	-39.1160
Item Type	Meat	-0.7553
Item Type	Others	-20.8090
Item Type	Seafood	183.3777
Item Type	Snack Foods	-11.6125
Item Type	Soft Drinks	-27.5884
Item Type	Starchy Foods	25.6830
Outlet Identifier	OUT013	1939.7011
Outlet Identifier	OUT017	2013.2696
Outlet Identifier	OUT018	1632.6082
Outlet Identifier	OUT019	9.2710
Outlet Identifier	OUT027	3351.2660
Outlet Identifier	OUT035	2052.7825
Outlet Identifier	OUT045	1838.6680
Outlet Identifier	OUT046	1909.8894
Outlet Identifier	OUT049	2006.6065
Outlet Size	High	NA
Outlet Size	Medium	NA
Outlet Size	Small	NA

In this chart Outlet Identifier OUT027 utilities value is 3321.2660 it is high. This Outlet Identifier was Supermarket Type3 and its size was medium. This outlet performed much better than any other Identifier. The Outlet Identifier second highest utilities value for OUT035 is 2052.7825. This Outlet identifier was Supermarket Type3and its size was medium. As a conclusion, if the Big mart want to open a new store, they could consider “Supermarket Type3”, and the size is Medium.

6. Machine learning models.

Before the model can be built, the columns Item_Identifier and Outlet_Identifier ware removed. These columns had zero variance because they are particular to each item and each outlet.

The next step to build the machine learning model to predict future Item_Outlet_sales was to compare a list of machine learning models. The algorithms in this list included lm, glm, glmnet, treebag, bagEarth, random forest aka ranger and gbm. All of these model types are suitable for regression analysis.

```
> results <- resamples(models)
> summary(results)

Call:
summary.resamples(object = results)

Models: glm, glmnet, lm, ranger, treebag, gbm, bagEarth
Number of resamples: 30

MAE
             Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
glm      775.9943 816.4942 839.0193 839.6854 853.9837 933.2627    0
glmnet   772.7056 813.0107 838.9686 837.0534 850.5198 933.6463    0
lm       775.9943 816.4942 839.0193 839.6854 853.9837 933.2627    0
ranger   714.2322 769.5478 778.4033 783.0977 794.4835 858.3276    0
treebag  741.2490 782.0189 792.2292 796.6196 813.3684 877.2990    0
gbm      707.3616 754.9257 767.6244 768.0749 781.3916 847.8307    0
bagEarth 777.4188 815.8169 841.2400 838.1773 852.7168 934.0947    0

RMSE
              Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
glm      1028.7664 1098.129 1120.050 1131.993 1162.969 1260.622    0
glmnet   1027.9210 1099.337 1118.640 1131.657 1159.880 1262.603    0
lm       1028.7664 1098.129 1120.050 1131.993 1162.969 1260.622    0
ranger   1022.8900 1090.593 1115.802 1117.760 1138.682 1218.262    0
treebag  1011.8209 1075.150 1097.227 1106.062 1128.304 1204.538    0
gbm       998.5647 1053.001 1074.248 1084.177 1110.078 1191.526    0
bagEarth 1030.9190 1097.859 1116.683 1130.479 1159.710 1258.292    0

Rsquared
              Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
glm      0.5297502 0.5458521 0.5567653 0.5601816 0.5649000 0.6150966    0
glmnet   0.5308590 0.5486413 0.5565591 0.5608996 0.5673058 0.6167648    0
lm       0.5297502 0.5458521 0.5567653 0.5601816 0.5649000 0.6150966    0
ranger   0.5385490 0.5600307 0.5701170 0.5733718 0.5839111 0.6185823    0
treebag  0.5392784 0.5653139 0.5733258 0.5802891 0.5978366 0.6303610    0
gbm      0.5630290 0.5813838 0.5936231 0.5967521 0.6095468 0.6480338    0
bagEarth 0.5316263 0.5513855 0.5577866 0.5614351 0.5670542 0.6166787    0
```

You can see the MAE result for all models and the gbm models is low compare to other models, The gbm model MAE is 707.3616 and you can see the above the result MIN, 1^{st} qu, Median, Mean, 3^{rd} qu, Max is all very low for GBM model and the best model for GBM in MAE result.

You can see the RMSE result for this model also the best fit GBM model and you can see the GBM highlighted values the RMSE is 998.5647 compare to other model Gbm is the low RMSV, So the best model for RMSE is GBM.

The final model is RSquared this model also GBM is the Best model because the RSquared value is high for GBM compare to other models. And you can see the Rsquard value for GBM is 0.5630290.

6.1 MODELS

Then I try to see the all models sample errors also because the sample error also important for this model. You can see the all model RMSE results or sample errors.

6.1.1 Generalized Linear Models (glm)

Generalized linear models are fit using the **glm( )** function. The form of the **glm** function is

glm(*formula***, family=***familytype***(link=***linkfunction***), data=)**

```
> models
$glm
Generalized Linear Model

5967 samples
   9 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 5370, 5370, 5370, 5371, 5370, 5371, ...
Resampling results:

  RMSE      Rsquared   MAE
  1131.993  0.5601816  839.6854
```

You can see this model results and this model I try to resampling and cross-valisate (10 fold repeated 3 times) and also you can see the summary of sample size above picture then the final results for RMSE is 1131.993 bit high compare to other models.

6.1.2 Generalized Linear Models NET (glmnet)

Fit a generalized linear model via penalized maximum likelihood. The regularization path is computed for the lasso or elasticnet penalty at a grid of values for the regularization parameter lambda. Can deal with all shapes of data, including very large sparse data matrices. Fits linear, logistic and multinomial, poisson, and Cox regression models.

```
$glmnet
glmnet

5967 samples
   9 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 5370, 5370, 5370, 5371, 5370, 5371, ...
Resampling results across tuning parameters:

  alpha  lambda      RMSE      Rsquared   MAE
  0.10     1.935134  1132.016  0.5601584  839.3104
  0.10    19.351336  1132.545  0.5599452  838.6196
  0.10   193.513361  1164.728  0.5469627  859.5859
  0.55     1.935134  1132.078  0.5601088  839.1439
  0.55    19.351336  1131.657  0.5608996  837.0534
  0.55   193.513361  1214.637  0.5167489  900.0673
  1.00     1.935134  1131.777  0.5603411  838.7579
  1.00    19.351336  1132.192  0.5608161  836.9175
  1.00   193.513361  1265.689  0.4855106  942.3760

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were alpha = 0.55 and lambda = 19.35134.
```

You can see the result above picture and the alpha 0.10 to 1.00 and also lambda value but this model also not fit because the RMSE value is bit High.

6.1.3 Linear regression(lm)

Linear regression is used to predict the value of an outcome variable *Y* based on one or more input predictor variables *X*. The aim is to establish a linear relationship (a mathematical formula) betlen the predictor variable(s) and the response variable, so that, I can use this formula to estimate the value of the response *Y*, when only the predictors (*Xs*) values are known.

$$Y = \beta_1 + \beta_2 X + \epsilon$$

where, β_1 is the intercept and β_2 is the slope. Collectively, they are called *regression* *coefficients*. ϵ is the error term, the part of Y the regression model is unable to explain.

```
$lm
Linear Regression

5967 samples
   9 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 5370, 5370, 5370, 5371, 5370, 5371, ...
Resampling results:

  RMSE      Rsquared   MAE
  1131.993  0.5601816  839.6854

Tuning parameter 'intercept' was held constant at a value of TRUE
```

You can see the linear model results this model also not fit because the RMSE bit high.

6.1.4 Random forest (ranger)

In the **random forest** approach, a large number of **decision** trees are created. Every observation is fed into every **decision** tree. The most common outcome for each observation is used as the final output.

```
$ranger
Random Forest

5967 samples
   9 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 5370, 5370, 5370, 5371, 5370, 5371, ...
Resampling results across tuning parameters:

  mtry  splitrule   RMSE      Rsquared   MAE
   2    variance    1273.262  0.5161956   960.9864
   2    extratrees  1337.972  0.4696238  1021.2260
  14    variance    1117.760  0.5733718   783.0977
  14    extratrees  1118.879  0.5717661   781.5368
  27    variance    1136.161  0.5613899   795.7413
  27    extratrees  1132.067  0.5639571   789.7057

Tuning parameter 'min.node.size' was held constant at a value of 5
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were mtry = 14, splitrule = variance and min.node.size = 5.
```

The random forest resampling results also not fit you can see the RMSE is bit high.

6.1.5 Gradient Boosting Machine(gbm)

The GBM fit misc is an **R** object that is simply passed on to the **gbm** engine. It. can be used for additional data for the specific distribution. Currently it is only used for passing the censoring indicator for the Cox proportional hazards model. w is a vector of lights of the same length as the y.

```
$gbm
Stochastic Gradient Boosting

5967 samples
   9 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 5370, 5370, 5370, 5371, 5370, 5371, ...
Resampling results across tuning parameters:

  interaction.depth  n.trees  RMSE      Rsquared   MAE
  1                   50      1259.058  0.4913015  944.4269
  1                  100      1178.610  0.5351478  872.7497
  1                  150      1154.090  0.5463843  855.8716
  2                   50      1127.136  0.5741083  829.2778
  2                  100      1100.918  0.5850253  802.3857
  2                  150      1096.961  0.5874178  796.2519
  3                   50      1088.117  0.5959459  779.1024
  3                  100      1084.177  0.5967521  768.0749
  3                  150      1086.752  0.5948096  768.0089
```

The GBM modes is the best model for compare to these other models. You can see the RMSE values interaction 1 and 2 bit High but interaction 3 RMSE values is 1084.177 it is low compare to other RMSE values, the RSquared values is 0.5967521 high compare to other RSquared values and also MAE value is low MAE is 768.0749.

Finally, When comparing the RMSE or out of sample error, the best performing model was gbm model. This model had an out of sample error of 1084.177.

Although the gbm model could be used for predictions. Combining these models should produce better results. Hopefully, an ensemble model of these models in the list will use the best parts of each model.

7. ENSEMBLE MODELS

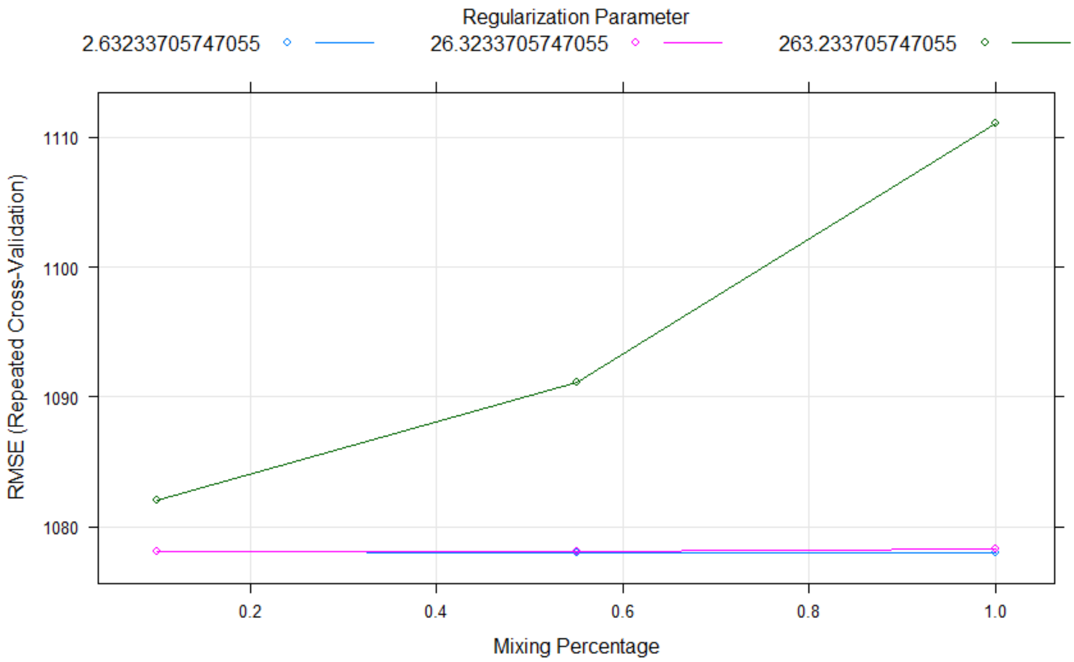
The three different types of ensemble for this report Ire a glmnet ensemble, a random forest ensemble and a bagEarth ensemble. After these ensembles Ire created, they Ire each tested to see which produced the best RMSE.

7.1 GLMNET Ensemble

The GLMNET Ensemble for this tested to see the produced the best RMSE. The test model to getting Predictions. and also calculated RMSE values.

```
> #GLMNET
> stack_glmnet <- caretStack(models, method = "glmnet", trControl = trainControl(method = "repeatedcv", number = 10, repeats
= 3, savePredictions = TRUE))
> stack_glmnet
A glmnet ensemble of 2 base models: glm, glmnet, lm, ranger, treebag, gbm, bagEarth

Ensemble results:
glmnet

17901 samples
    7 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 16111, 16110, 16112, 16110, 16112, 16111, ...
Resampling results across tuning parameters:

  alpha  lambda      RMSE      Rsquared   MAE
  0.10     2.632727  1078.739  0.6001692  760.4628
  0.10    26.327270  1078.874  0.6000871  761.2351
  0.10   263.272696  1082.906  0.5981888  771.1183
  0.55     2.632727  1078.716  0.6002045  760.2970
  0.55    26.327270  1078.844  0.6001880  760.7939
  0.55   263.272696  1091.930  0.5994612  780.6883
  1.00     2.632727  1078.696  0.6002292  760.1716
  1.00    26.327270  1078.989  0.6002252  760.8454
  1.00   263.272696  1111.794  0.5994637  812.4723

RMSE was used to select the optimal model using the smallest value.
The final values used for the model were alpha = 1 and lambda = 2.632727.
> |
```

The glmnet model produced an RMSE of 1083.199.



why I show this plot you can easily understand this model Y is a RMSE (Repeated Cross-Validation) and X is a Maxing percentage that means Alpha values. the lambda value is 2.63233705747055 it is representing for Blue line and the Pink line is

26.3233705747055 and the green line is 263.233725747055.

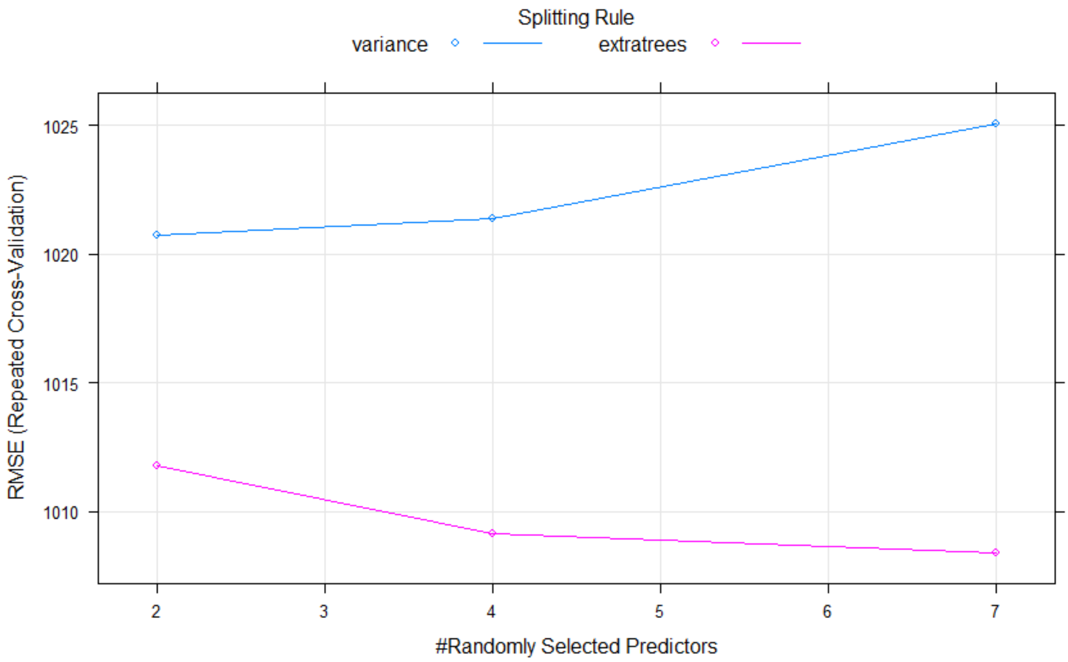
The green line alpha value increases the RMSE (Repeated Cross-Validation) value also increase but the blue and pink line you can see alpha value increase this both line is same but the alpha value is 0.6 plink line is little bit increase 0.2% for RMSE (Repeated Cross-Validation) value.

7.2 Random Forest Ensemble

The Random Forest Ensemble for tested to produce the best RMSE values and this model to getting prediction, and calculated the RMSE value. This Random Forest Ensemble to growing trees, progress percentage then the Estimated remaining times you can see the blow image.

```
> #random
> stack_rf <- caretStack(models, method = "ranger", trControl = trainControl(method = "repeatedcv", number = 10, repeats = 3,
 savePredictions = TRUE))
> stack_rf
A ranger ensemble of 2 base models: glm, glmnet, lm, ranger, treebag, gbm, bagEarth

Ensemble results:
Random Forest

17901 samples
    7 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 16111, 16110, 16110, 16111, 16112, 16109, ...
Resampling results across tuning parameters:

  mtry  splitrule   RMSE      Rsquared   MAE
  2     variance    1020.734  0.6417762  712.5215
  2     extratrees  1011.751  0.6481109  709.6961
  4     variance    1021.385  0.6413952  711.1486
  4     extratrees  1009.108  0.6498883  706.7467
  7     variance    1025.077  0.6388658  713.5233
  7     extratrees  1008.370  0.6503895  705.6616

Tuning parameter 'min.node.size' was held constant at a value of 5
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were mtry = 7, splitrule = extratrees and min.node.size = 5.
```

The random forest ensemble produced an RMSE of 1106.944.



You can see this plot Y is a RMSE (Repeated Cross-Validation) and X is a Randomly Selected Predictors this models I try to splitting the rule the Blue line is representing Variance and pink line is extratrees and you can see this plot random selected value is increase the extratrees is decrease then the variance values is increase.

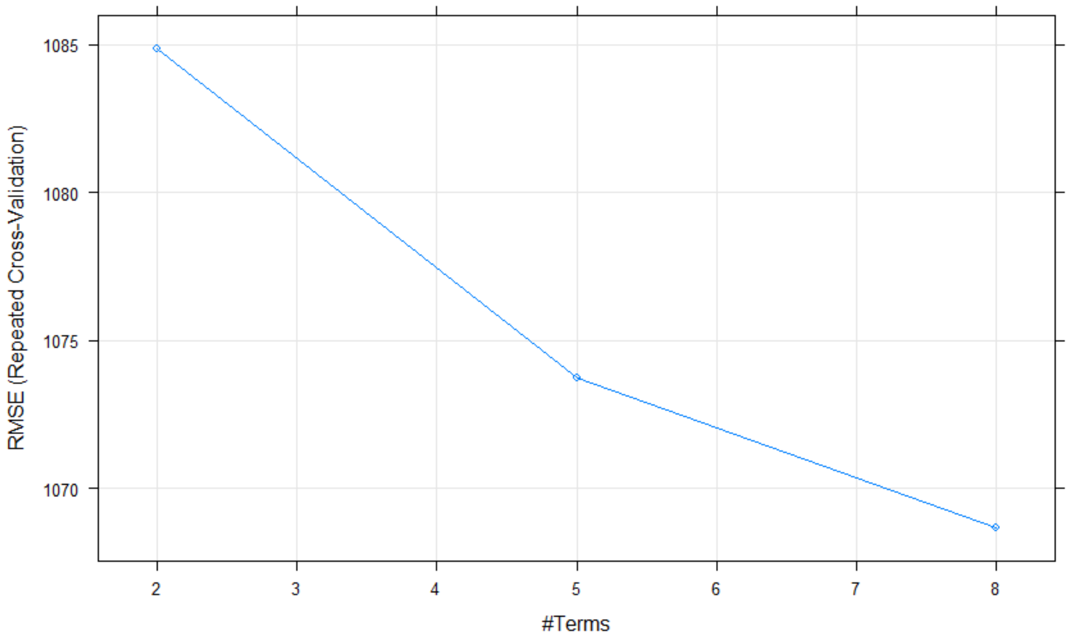
7.3 Bagging Ensemble

The Bagging, aka bootstrap aggregation, is a relatively simple way to increase the poIr of a predictive statistical model by taking multiple random samples(with replacement) from your training data set, and using each of these samples to construct a separate model and separate predictions for your test set.

The bagging Ensemble for tested to produce the best RMSE.

```
> #Bagging Ensemble
>
> stack_bag <- caretStack(models, method = "bagEarth", trControl = trainControl(method = "repeatedcv", number = 10, repeats =
 3, savePredictions = TRUE))
> stack_bag
A bagEarth ensemble of 2 base models: glm, glmnet, lm, ranger, treebag, gbm, bagEarth

Ensemble results:
Bagged MARS

17901 samples
    7 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 16111, 16110, 16110, 16110, 16111, 16112, ...
Resampling results across tuning parameters:

  nprune  RMSE      Rsquared   MAE
  2       1084.380  0.5972818  769.1236
  5       1074.002  0.6038893  758.0163
  8       1069.268  0.6073349  755.4200

Tuning parameter 'degree' was held constant at a value of 1
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were nprune = 8 and degree = 1.
>
```

The bagEarth model produced an RMSE of 1084.196.



You can see this plot Y is a RMSE (Repeated Cross-Validation) and X is a Nprune t is nodes 2 nodes the RMSE (Repeated Cross-Validation) high and the 5 nodes the RMSE (Repeated Cross-Validation) decrease finally 8 nodes the RMSE (Repeated Cross-Validation) decrease.

8. Conclusions

8.1 Prediction of Sales

The stores can organize supply chain management and investment based on the sales prediction. When the number of customer prediction goes up, the sales will increase, which means the demand will also increase. The store needs to equip with enough products for the upcoming increasing demands. When the stores want to do short time promotion, they should prepare more supplies. Deployment of human resources is similar. Based on heat map of average customers, stores can arrange sellers appropriately.

There Ire other conclusions that can be made from this report's analysis. First there is a moderate correlation betIen an Item's MRP at a Big Mart location and that item's sales at that location. Also, the smallest locations produced the loIst sales. HoIver, the largest location did not produce the highest sales. The location that produced the highest sales was the OUT027 location. This location was Supermarket Type3 and its size was medium. This outlet performed much better than any other location. Its median Item_Outlet_Sales Ire 3364.95. The location that was second was the OUT035 location, which had a median Item_Outlet_Sales of 2109.25.

If Big Mart Ire to try to increase sales at all locations, it may consider switching more locations to Supermarket Type3. Other things Big Mart could do to increase sales is to see which Items had the highest sales. They may also consider how product visibility affected outlet sales. HoIver, the model built in this report should be good for helping Big Mart predict future sales at its locations.