Big mart

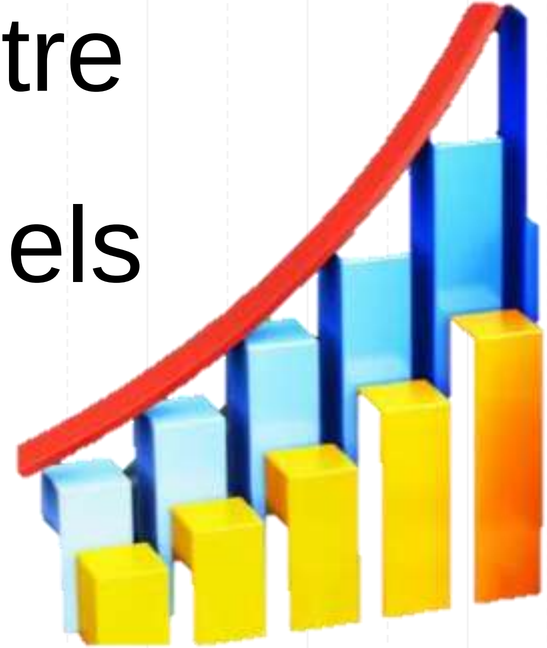
Shop smarter. Live better.

Big Mart Data Analysis

Roll no :- 31031821012

Name :- Ronak Kalantre

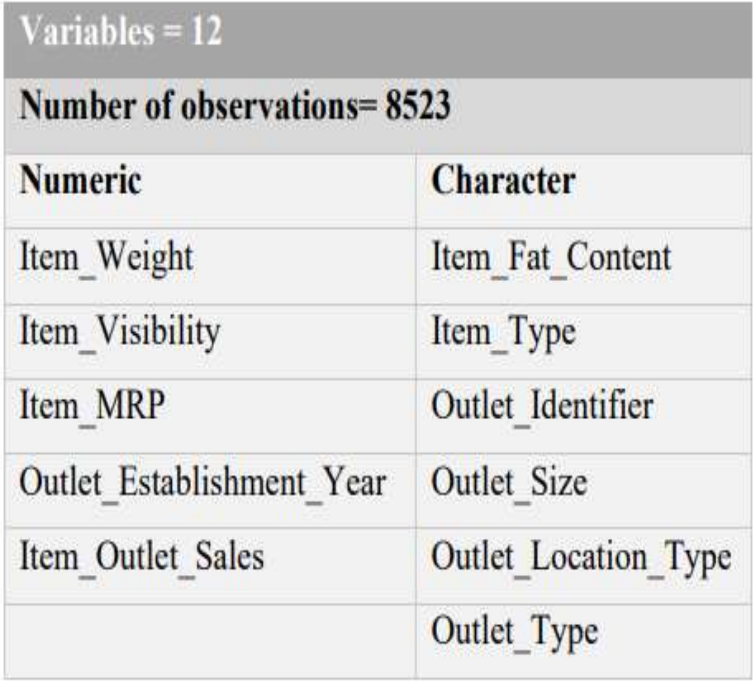
Project :- Linear Models



INTRODUCTION

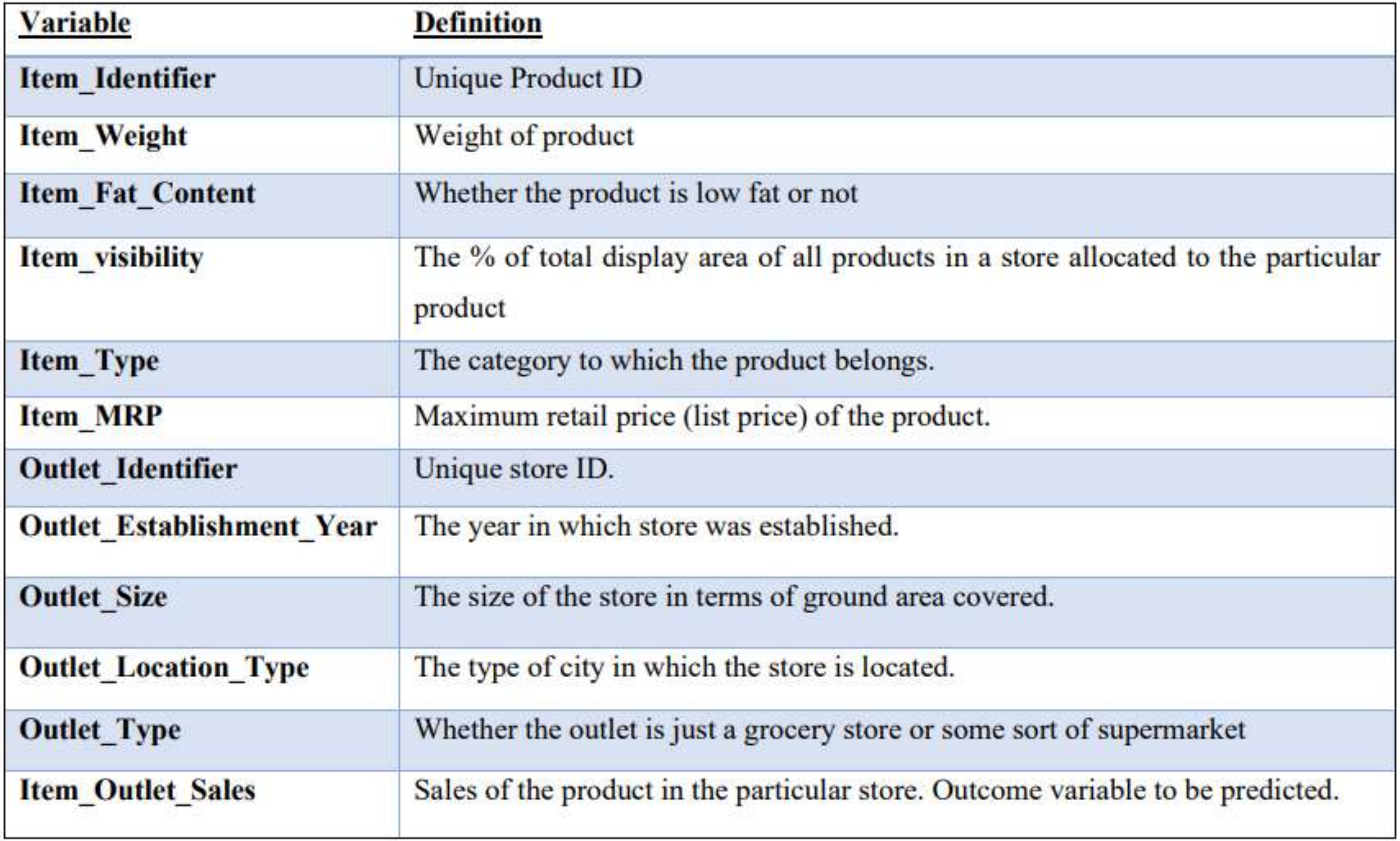
- In today's modern world, huge shopping centers such as big malls and marts are recording data related to sales of items or products with their various dependent or independent factors as an important step to be helpful in prediction of future demands and inventory management. Big Mart is an immense network of shops virtually all over the world. Trends in Big Mart are very relevant and data scientists evaluate those trends per Product and store in order to create potential centers.
- The dataset built with various dependent and independent variables is a composite form of item attributes, data gathered by means of customer, and also data related to inventory management in a data warehouse. The data is thereafter refined in order to get accurate predictions and gather new as well as interesting results that shed a new light on our knowledge with respect to the task's data. This can then further be used for forecasting future sales by means of employing machine learning algorithms such as the random forests and simple or multiple linear regression model. We are addressing the problem of big mart sales prediction or forecasting of an item on customer's future demand in different big mart stores across various locations and products based on the previous record. As good sales are the life of every organization so the forecasting of sales plays an important role in any shopping complex and a better prediction is always helpful, to develop as well as to enhance the strategies of business about the marketplace which is also helpful to improve the knowledge of marketplace.

ABOUT THE DATA SET



Variables = 12	
Number of observations= 8523	
Numeric	**Character**
Item_Weight	Item_Fat_Content
Item_Visibility	Item_Type
Item_MRP	Outlet_Identifier
Outlet_Establishment_Year	Outlet_Size
Item_Outlet_Sales	Outlet_Location_Type
	Outlet_Type

- I have collected the data form kaggle.
- Big Mart is a store that have collected its 2018 data for 1559 products across 10 stores in different cities.
- The data set consists of 8523 products across different cities and locations.
- There are missing values in the data.
- There are about 1463 missing values in Item_weight and 2410 in Outlet_size.



Variable	Definition
Item_Identifier	Unique Product ID
Item_Weight	Weight of product
Item_Fat_Content	Whether the product is low fat or not
Item_visibility	The % of total display area of all products in a store allocated to the particular product
Item_Type	The category to which the product belongs.
Item_MRP	Maximum retail price (list price) of the product.
Outlet_Identifier	Unique store ID.
Outlet_Establishment_Year	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 particular store. Outcome variable to be predicted.

OBJECTIVES

Through this project I intend to achieve the following objectives:

1) To analyse the data and evaluate meaningful insights from the given data.

2) To check whether Item_Outlet_Sales differs significantly for different levels of Item_Fat_Content, Outlet_size, Outlet_Location_Type, Outlet_Type and Outlet_Identifier.

METHODOLOGY

To conduct the project, I have collected the data on Kaggle and for the statistical study, the following tools will be used.

- Data Visualization
- Model Fitting
  a) Multiple Regression Model
  b) Random Forest Method
- Testing Of Hypothesis
  a) Analysis of Variance using ANOVA Test
  b) Chi Square Test

- ***Data Visualization***

Visualization of data is a clear way to make sense of data rather than the numerical summaries. A wide range of charts are used to demonstrate comparisons, simple- histogram, barplot, boxplot and multivariate-scatter, boxplot, bar plot, along with correlation plots are used. Plots allow us to analyze the relationship between two data.

- ***Model Fitting***

a) Multiple Linear Regression

The technique of modelling a linear relationship between a response (dependent) variable and one or more explanatory (independent) variables is called linear regression.
The equation of a multiple linear regression is given by:

$$\mathbf{Y = \beta 0 + \beta 1X1 + \beta 2X2+ \ldots + \beta kXk + \varepsilon}$$

where,

Y : the dependent variable
Xi’s : independent variables (i = 1 to k)
&
ε is the error term

Assumption :-

There are four assumption associated with a linear regression model

- **Linearity**
- **Homoscedasticity**
- **Normality of Residuals**
- **Independence**

b) Random Forest Method

A random forest is a machine learning technique that is used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.

A random forest eradicates the limitations of a decision tree algorithm. It reduces the overfitting of datasets and increases precision. It generates predictions without requiring many configurations.

- **Testing of Hypothesis**

a) Analysis of Variance (ANOVA) Test

ANOVA is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. It checks the impact of one or more factors by comparing the means of different samples. This test is basically based on separation of variance in the data into components where each of the components is a measure of variation due to some specific independent cause. It consists of comparison of estimate of variance due to assignable causes with estimate of variance due to chance causes.

- One Way ANOVA

The scheme of classification due to one factor is called one-way classification of ANOVA.

- Two Way ANOVA

The scheme of classification due to two factors is called two-way classification of ANOVA.

b) Chi Square Test

The Chi-Square test of independence is used to determine if there is a significant relationship between two nominal (categorical) variables. The frequency of each category for one nominal variable is compared across the categories of the second nominal variable. The data can be displayed in a contingency table where each row represents a category for one variable and each column represents a category for the other variable. The chi-squared test of independence compares our sample data in the contingency table to the distribution of values we’d expect if the null hypothesis is correct.

ANALYSIS

OBJECTIVE 1 :- To analysis the data and evaluate meaningful insights from the given data

- **Data Preprocessing**

Data preprocessing involves transforming raw data to well-formed data sets. It involves both data validation and data imputation

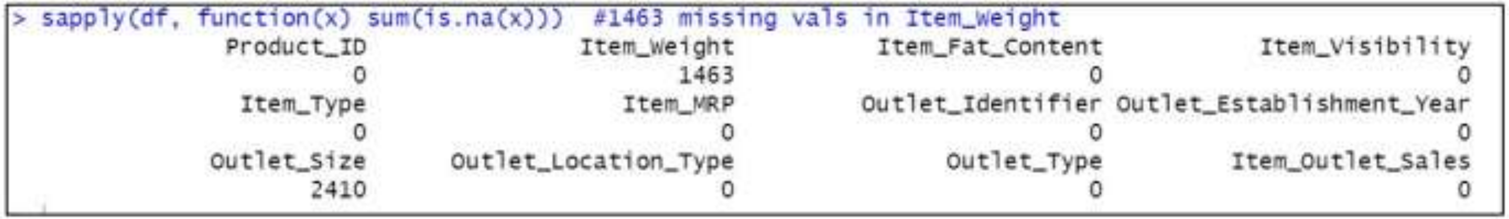
- **Imputation of Missing Values**

The goal of data imputation is to correct errors and input missing values. Missing data in the data set can reduce the power / fit of a model or can lead to a biased model because we have not analyzed the behavior and relationship with other variables correctly. It can lead to wrong prediction or classification.

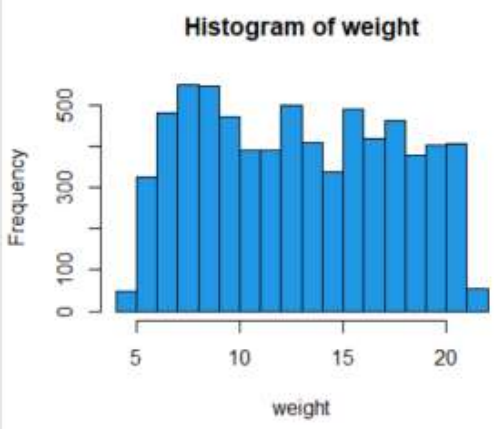
Mean / Mode / Median imputation is one of the most frequently used methods. It consists of replacing the missing data by

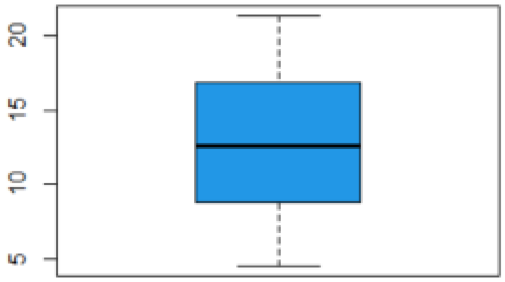
- mean, if the data is more clustered
- median, if we have skewness/ outliers in the data
- mode, if the data is categorical

After analyzing, there are 1463 missing values in Item_Weight (numerical variable) and 2410 missing values in Outlet_Size (categorical variable).

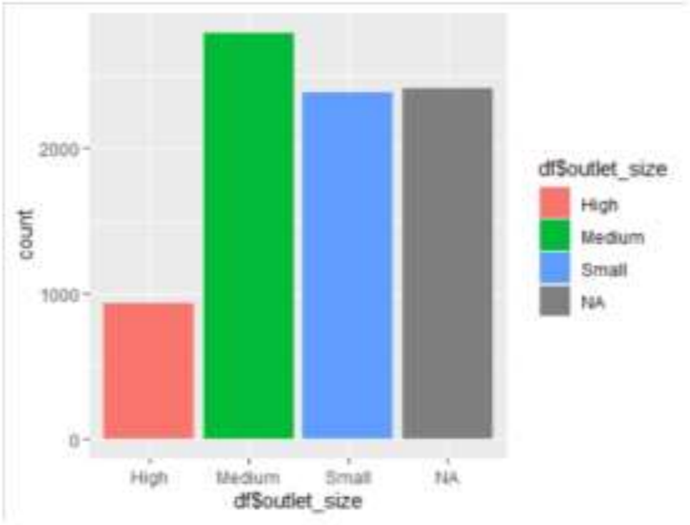


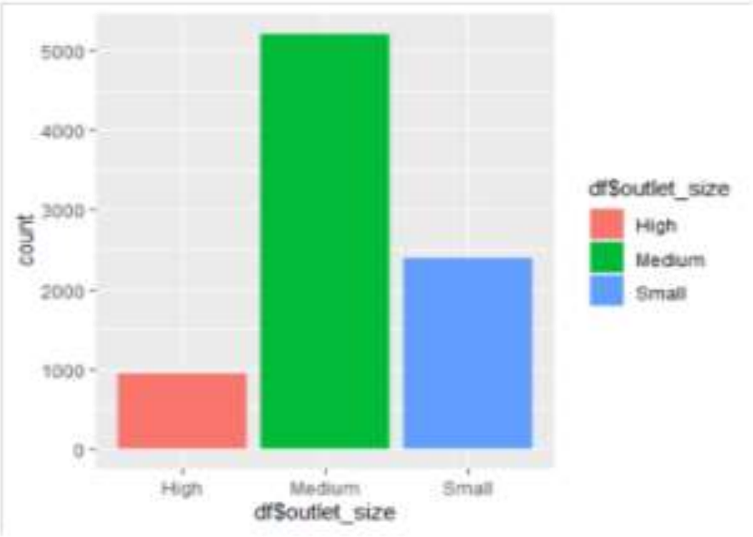
```
> sapply(df, function(x) sum(is.na(x)))  #1463 missing vals in Item_Weight
               Product_ID               Item_Weight          Item_Fat_Content           Item_Visibility
                        0                      1463                         0                         0
                Item_Type                  Item_MRP         Outlet_Identifier Outlet_Establishment_Year
                        0                         0                         0                         0
              Outlet_Size      Outlet_Location_Type               Outlet_Type         Item_Outlet_Sales
                     2410                         0                         0                         0
```





- For variable Item_Weight, there are no outliers therefore imputing mean of the observations for the missing values.





- For outlet size, since it was a categorical variable, missing values were replaced by mode of the observations. Clearly, the mode is the category Medium hence missing values were imputed by Medium.

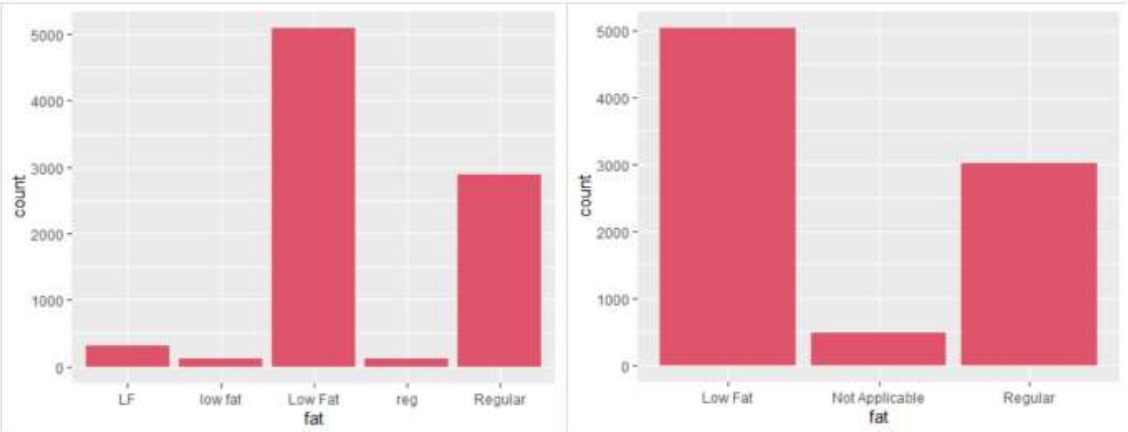
- **Establishment year transformation:**

  The variable estatblishment_year was transformed to year_established by subtracting from 2021 to get more meaning from the data, since year_established would imply the age of the outlet which might impact Outlet_Sales (Response Variable)

- **Data Cleaning- Item_Type and Item_Fat_Content:**

  Item_Fat_Content is uncleaned since it has ‘LF’, ‘low fat’ and ‘Low Fat’ observations for the same values and ‘reg’ and ‘Regular’ represent the same values as well.

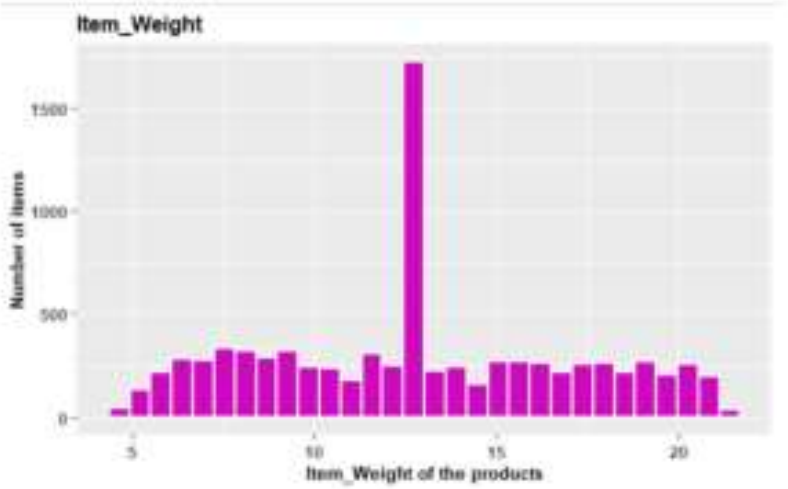
  Also, introducing a new category as ‘Not Applicable’ in Item_Fat_Content for Item_type: ‘health and hygiene’, ‘household’ and ‘others’.

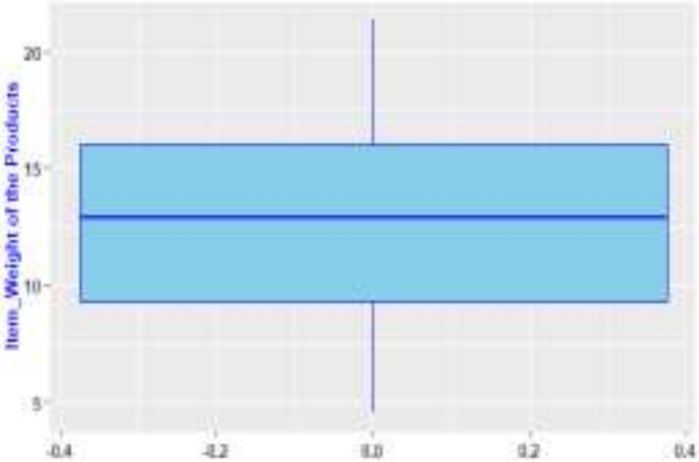


UNIVARIATE ANALYSIS:

1.) Item_Weight

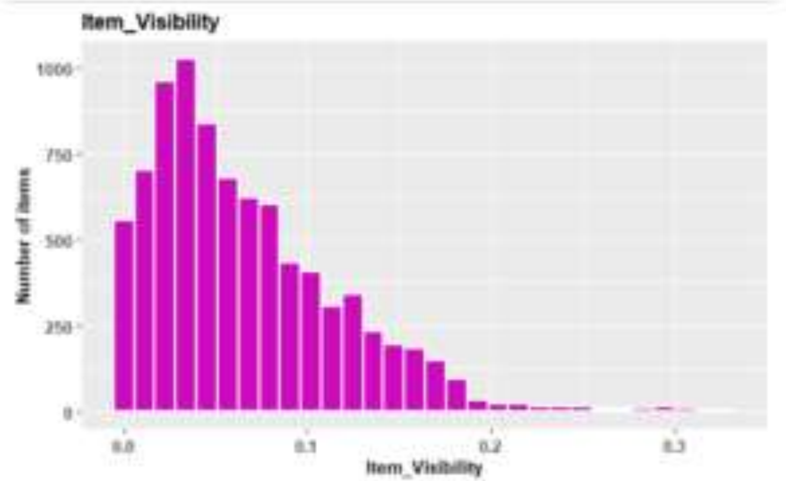
- The weight of the items lies between the range of 4 - 22. No. of items having weight below 5 and above 22 are comparatively lesser in number. The average weight of the items is 12.
- Item weight is symmetric. No outliers are present.

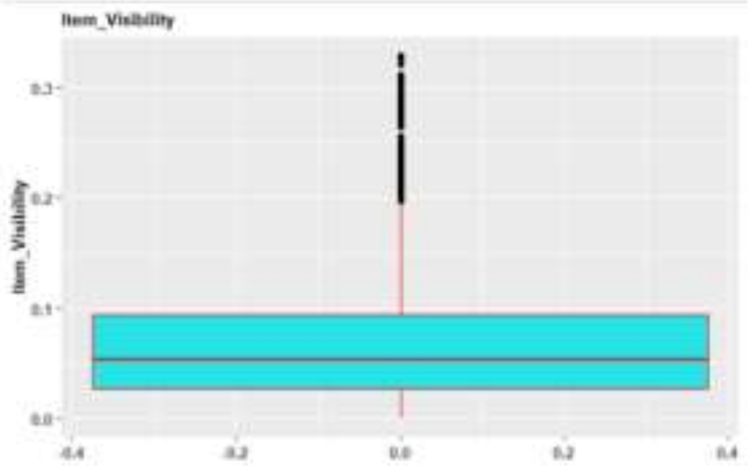




2) Item_Visibility

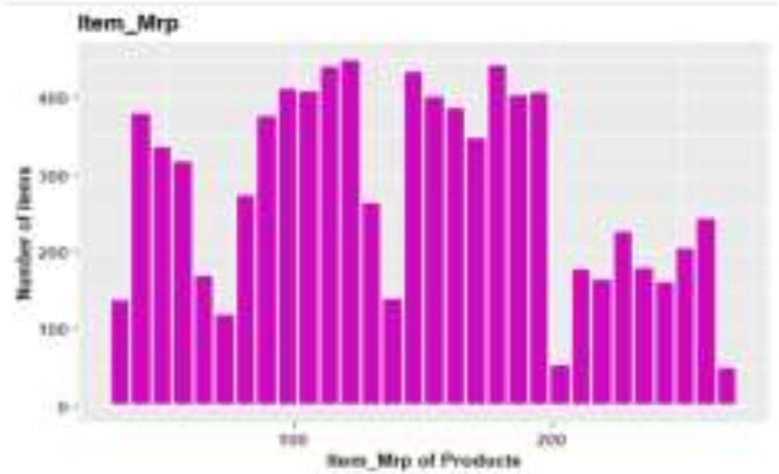
- Maximum number of products have product item_visibility between 0-0.1.
- The graph is right skewed. Outliers are present.

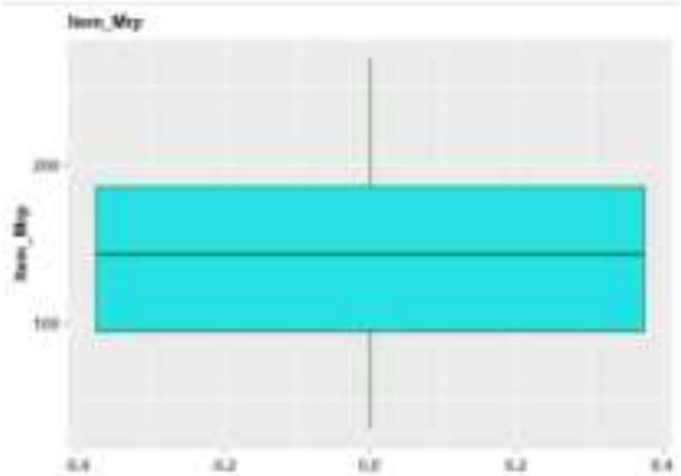




3.) Item_Mrp

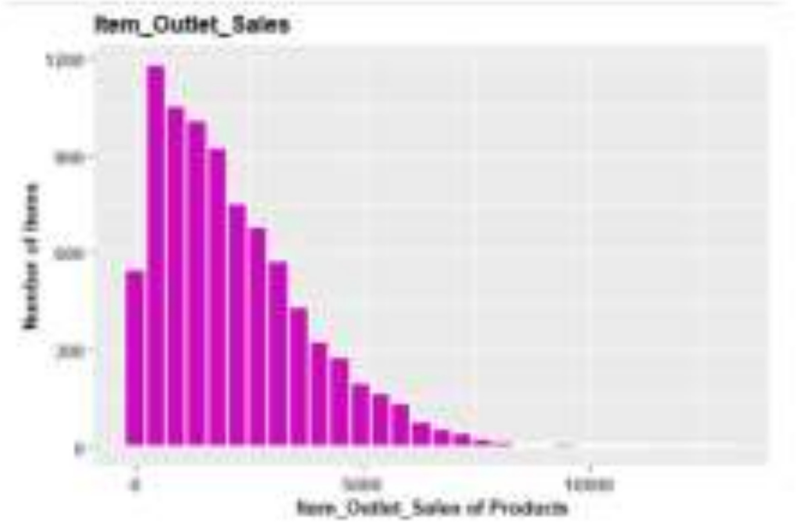
- The maximum number of products have Item_Mrp less than 200.
- The graph is symmetric and has no Outliers.

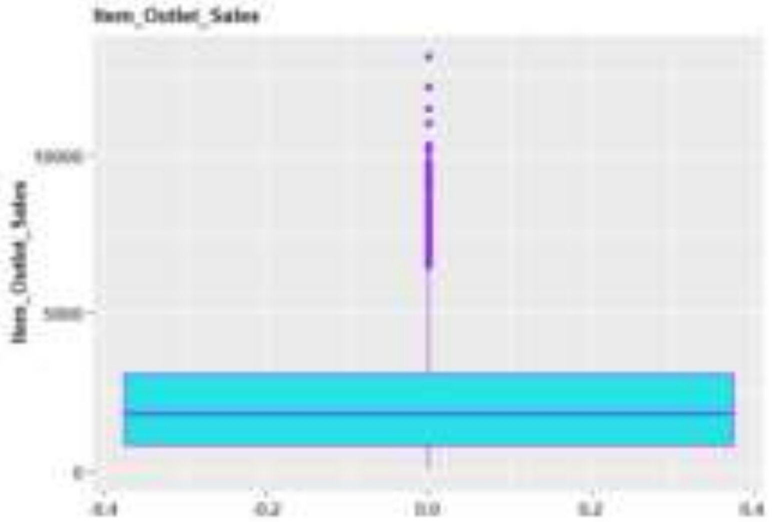




4.) Item_Outlet_Sales

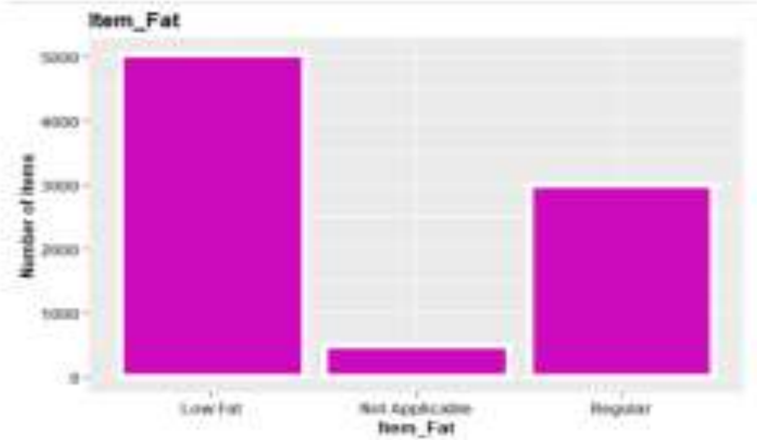
- The graph might suggest that maximum number of products in store are sold in the price range of 0 to 5000.
- The graph is right skewed, outliers are present.





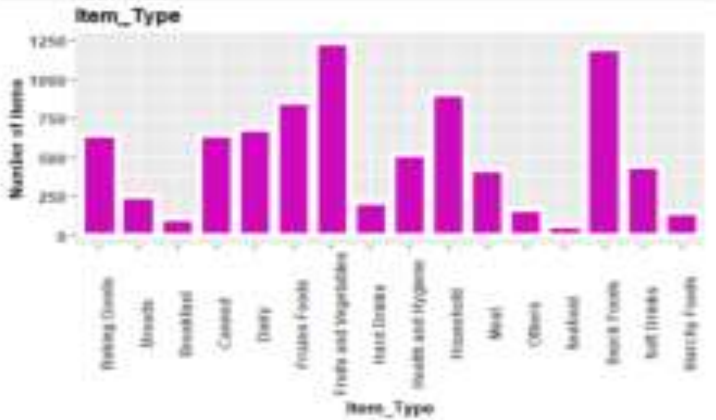
5.) Item_Fat

- Items having low fat content are highest in number (approximately 59%), followed by regular fat content (approximately 35%)
- The items with no fat content are least in number (approximately 6%)



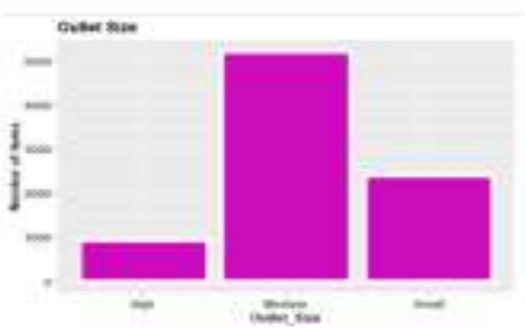
6.) Item_Type

- More than 1200 items (approximately 14%) sold are of type fruits & vegetables and snack foods.
- Number of items for type breakfast and seafood are very less sold



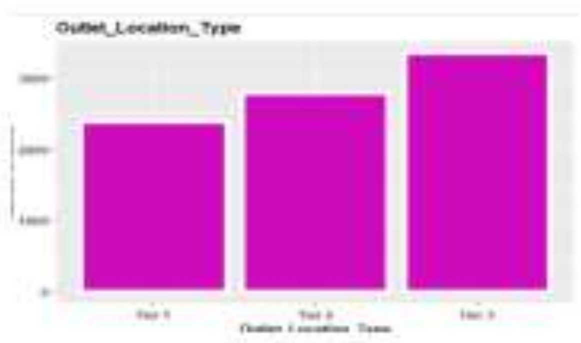
7.) Outlet_Size

- Maximum number of items were sold from medium size outlets (approximately 59%), followed by small size outlets (approximately 29%)
- Least number of items were sold from outlets of high size (approximately 12%)
- Medium size outlets are more popular



8.) Outlet_Location_Type

- Items in outlets located in Tier3 cities were sold more (approximately 39%), followed by outlets located in Tier2 cities (approximately 33%)
- Least number of items were sold from outlets located in Tier1 cities (approximately 28%).

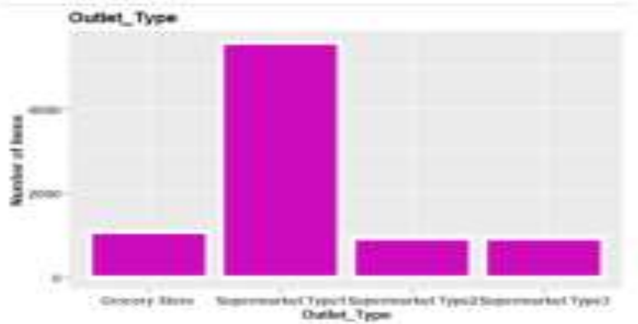


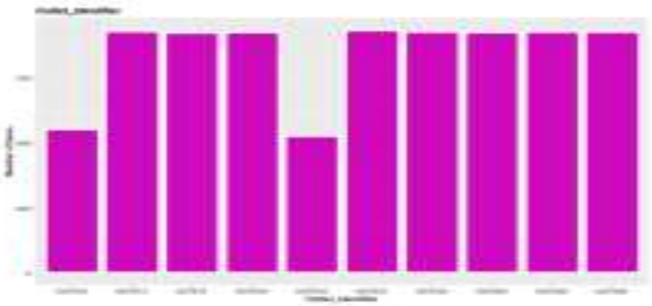
9.) Outlet_Type

- Maximum number of items were sold in Supermarket Type1(approximately 66%) which implies most of the customers prefer to buy the items from the Supermarket Type 1 stores.
- Less number of items were sold in Supermarket Type2, Supermarket Type3 and Grocery Stores

10) Outlet_Identifier

- Outlets OUT013, OUT017, OUT018, OUT027, OUT035, OUT045, OUT046, and OUT049 are equally popular (have more number of items sold) while outlets OUT010 and OUT019 are less popular (have less number of items sold)

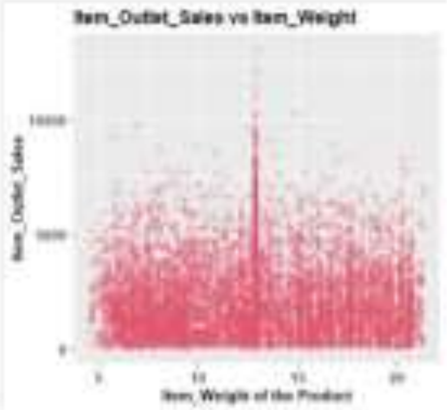




Bivariate Data Visualization

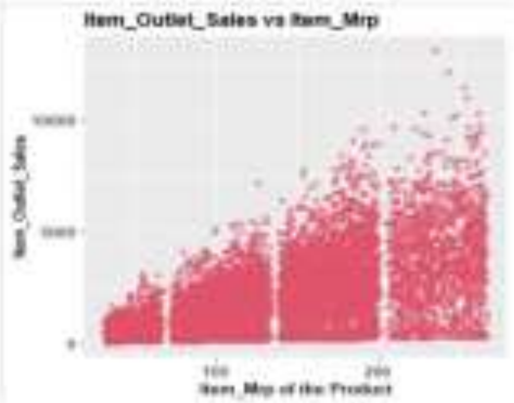
1.) Item_Outlet_Sales and Item_Weight

- Item_Outlet_Sales is spread well across the entire range of the Item_Weight without any pattern.



2.) Item_Outlet_Sales and Item_Mrp

- We can clearly see 4 different distributions for Item_MRP. . Maximum number of products in store were sold which were had more MRP.

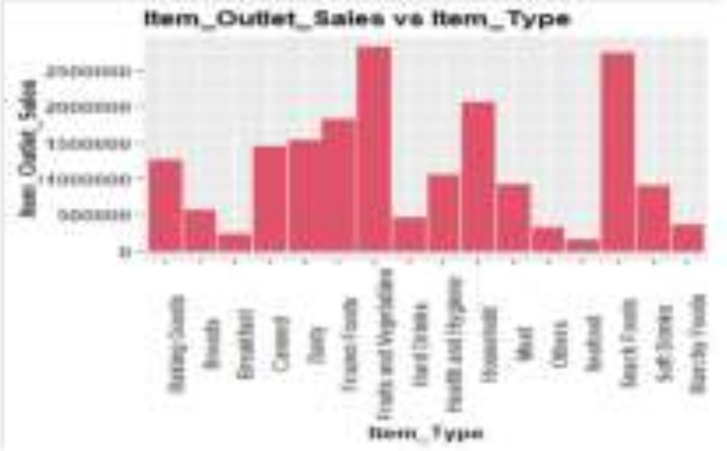


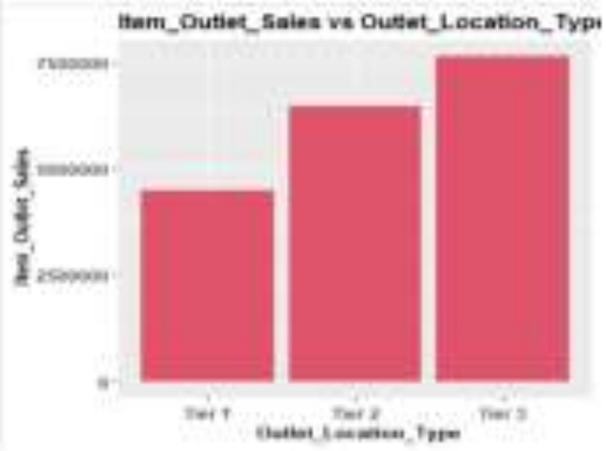
3.) Item_Outlet_Sales and Item_Type

- The maximum sales have been generated by mainly Baking Goods, Sea food, Starchy food and other food items.
- Baking goods and snack foods have approximately more than 2500000 sales.

4.) Item_Outlet_Sales and Outlet_Location_Type

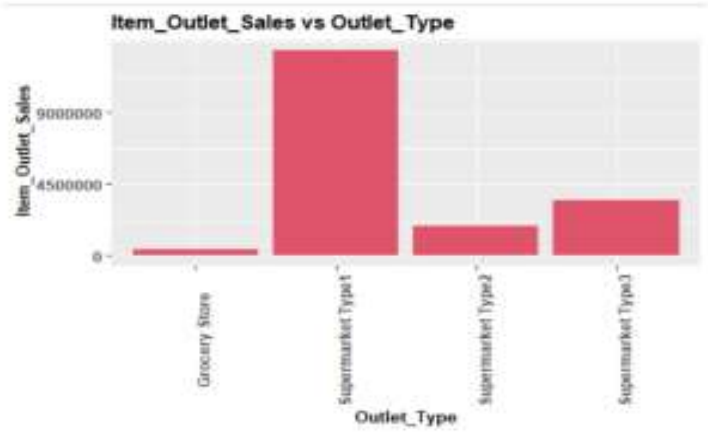
- Most of the products in store were sold in Tier 3 city. Tier 1 remained the lowest.





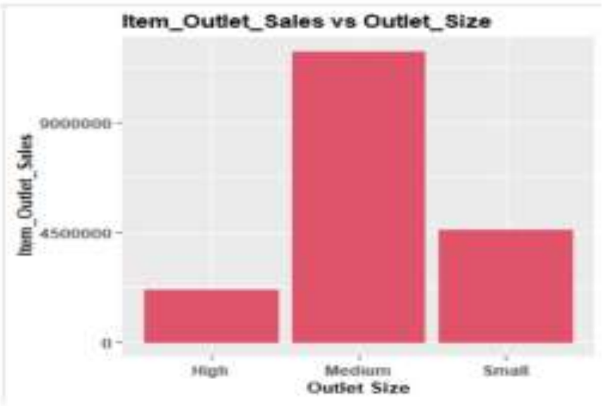
5.) Outlet type and Item outlet sales

- The above graph tells us that the Supermarket Type 1 has maximum number of sales while grocery store has the lowest sale.

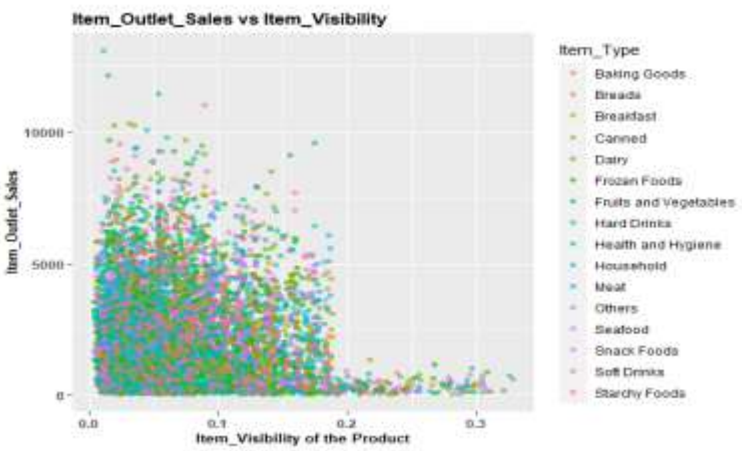


6.) Outlet size and Item outlet sale

- Maximum products are in medium sized outlet store

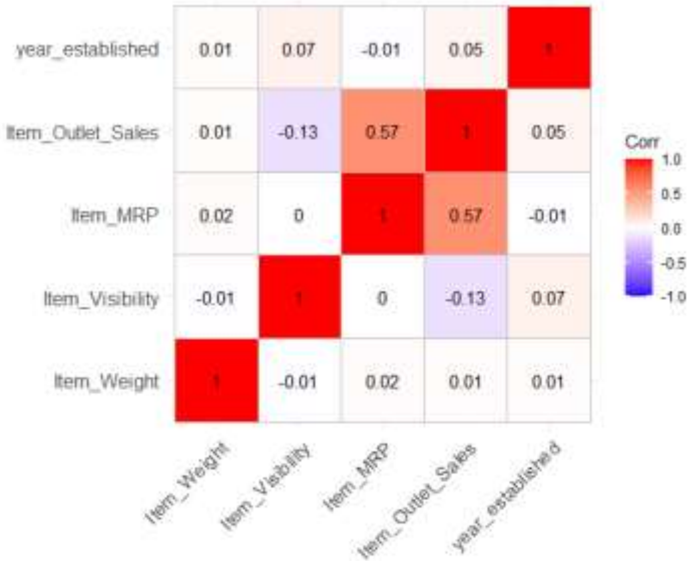


7.) Item outlet and sales Item visibility



Item_Visibility vs Item_Outlet_Sales indicates that the more visible a product is the less high its Item_Outlet_Sales will be. This might be due to the fact that a great number of daily use products, which do not need high Item_Visibility, control the top of the Item_Outlet_Sales chart. Furthermore, there is a concerning number of products with Item_Visibility 0.

- **Correlation Plot**



	Item_Weight	Item_Visibility	Item_MRP	Item_Outlet_Sales	year_established
year_established	0.01	0.07	-0.01	0.05	1
Item_Outlet_Sales	0.01	-0.13	0.57	1	0.05
Item_MRP	0.02	0	1	0.57	-0.01
Item_Visibility	-0.01	1	0	-0.13	0.07
Item_Weight	1	-0.01	0.02	0.01	0.01

- Item_Mrp has high degree of positive correlation with Item_Outlet_Sales.

- Item_Visibility has negative correlation with Item_Outlet_Sales.

OBJECTIVE 2: To check whether Item_Outlet_Sales differs significantly for different levels of Item_Fat_Content ,Outlet_size ,Outlet_Location_Type, Outlet_Type and Outlet_Identifier

Hypothesis Testing - One Way ANOVA

Model:

$$yij=\mu+\alpha i+\mathcal{E}ij;\ i=1,2,3,\ldots.,k\ ,\ j=1,2,\ \ldots,\ r$$

where:

yij: Sales of ith factor replicated jth time
μ: General mean
αi: effect of ith factor
$\mathcal{E}$ij: random error and follows N(0, σ^2)

a) Item_Fat_Content (Levels: Low Fat, Regular, Not Applicable)

Assumptions :-

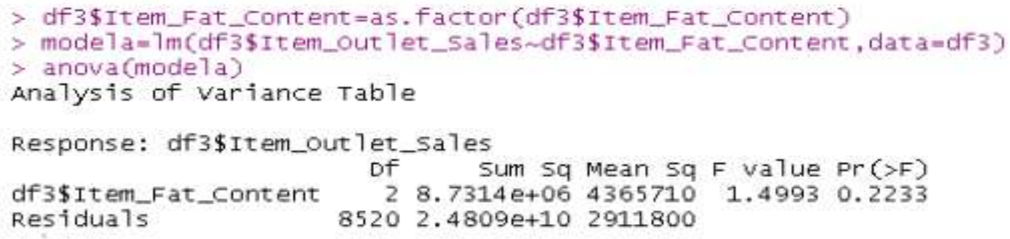
1. Normality of residuals: Using Normal Q-Q plot, since majority of the points lie along the line, eij's are normally distributed.
2. Homoscedasticity: Errors have constant variance
3. Independence of Residuals: eij's are independent since no pattern is observed in residual plot.

Hypothesis :-

H0: There is no significant difference between the average sales of different levels of Item_Fat_Content, $\alpha i=0\ \forall i$

H1: $\alpha i \neq 0$ for atleast one I

Analysis :-



```
> df3$Item_Fat_Content=as.factor(df3$Item_Fat_Content)
> modela=lm(df3$Item_Outlet_Sales~df3$Item_Fat_Content,data=df3)
> anova(modela)
Analysis of Variance Table

Response: df3$Item_Outlet_Sales
                       Df     Sum Sq Mean Sq F value Pr(>F)
df3$Item_Fat_Content    2 8.7314e+06 4365710  1.4993 0.2233
Residuals            8520 2.4809e+10 2911800
```

Inference:

Testing H0: αi=0 ∀i: p value = 0.2233 > 0.05, hence we do not reject H0 at 5% level of significance. This means αi is not significant, all the levels of Item_Fat_Content have the same average sales.

b) Outlet_Size (Levels: Small, Medium, High)

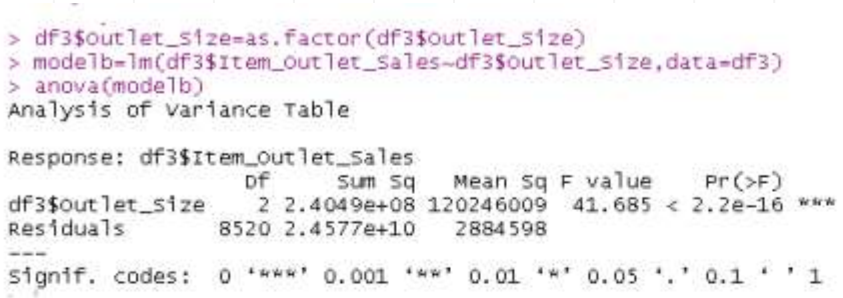
Assumptions: All the 3 assumptions were checked and are approximately satisfied.

Hypothesis:

H0: There is no significant difference between the average sales of different levels of Outlet_Size,
αi=0 ∀i

H1: αi ≠0 for atleast one i

Analysis:



```
> df3$outlet_size=as.factor(df3$outlet_size)
> modelb=lm(df3$Item_Outlet_Sales~df3$outlet_size,data=df3)
> anova(modelb)
Analysis of Variance Table

Response: df3$Item_Outlet_Sales
                  Df     Sum Sq   Mean Sq F value    Pr(>F)    
df3$outlet_size    2 2.4049e+08 120246009  41.685 < 2.2e-16 ***
Residuals       8520 2.4577e+10   2884598                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

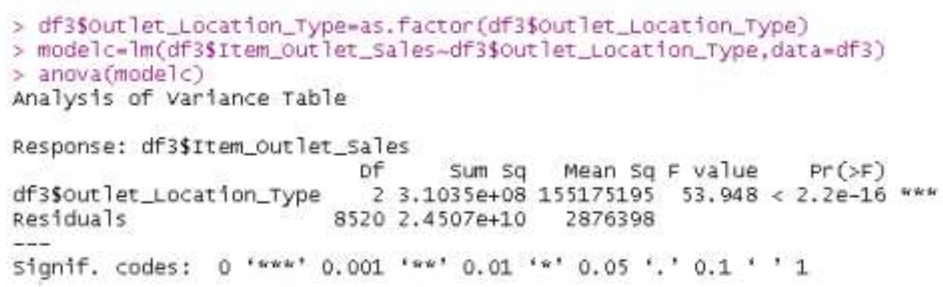
Inference:
Testing H0: αi=0 ∀i: p value = 2.2e-16 < 0.05, hence we reject H0 at 5% level of significance.
This means αi is significant, all the levels of Outlet Size do not have the same average sales.

c) Outlet_Location_Type (Levels: Tier1, Tier2, Tier3)

Assumptions: All the 3 assumptions were checked and are approximately satisfied.

Hypothesis:
H0: There is no significant difference between the average sales of different levels of Outlet_Location_Type, αi=0 ∀i
H1: αi ≠0 for atleast one i

Analysis :-



```
> df3$Outlet_Location_Type=as.factor(df3$Outlet_Location_Type)
> modelc=lm(df3$Item_Outlet_Sales~df3$Outlet_Location_Type,data=df3)
> anova(modelc)
Analysis of Variance Table

Response: df3$Item_Outlet_Sales
                           Df     Sum Sq   Mean Sq F value    Pr(>F)    
df3$Outlet_Location_Type    2 3.1035e+08 155175195  53.948 < 2.2e-16 ***
Residuals                8520 2.4507e+10   2876398                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Inference:

Testing H0: αi=0 ∀i: p value = 2.2e-16 < 0.05, hence we reject H0 at 5% level of significance. This means αi is significant, all the levels of Outlet_Location_Type do not have the same average sales.

d) Outlet_Type (Levels: Grocery Store, Supermarket Type1, Supermarket Type2, Supermarket Type3)

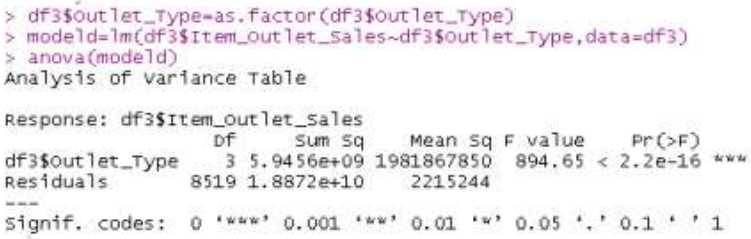
Assumptions: All the 3 assumptions were checked and are approximately satisfied.

Hypothesis:

H0: There is no significant difference between the average sales of different levels of Outlet Type,
αi=0 ∀i

H1: αi ≠0 for atleast one i

Analysis :-



```
> df3$Outlet_Type=as.factor(df3$Outlet_Type)
> modeld=lm(df3$Item_Outlet_Sales~df3$Outlet_Type,data=df3)
> anova(modeld)
Analysis of Variance Table

Response: df3$Item_Outlet_Sales
                  Df     Sum Sq    Mean Sq F value    Pr(>F)
df3$Outlet_Type    3 5.9456e+09 1981867850  894.65 < 2.2e-16 ***
Residuals       8519 1.8872e+10    2215244
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Inference:

Testing H0: αi=0 ∀i: p value = 2.2e-16 < 0.05, hence we reject H0 at 5% level of significance. This means αi is significant, all the levels of Outlet_Type do not have the same average sales.

e) Outlet_Identifier (Levels: OUT010, OUT013, OUT017, OUT018, OUT019, OUT027, OUT035, OUT045, OUT046, OUT049)

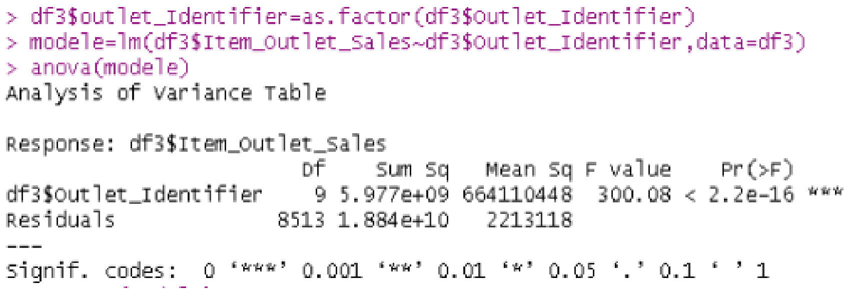
Assumptions: All the 3 assumptions were checked and are approximately satisfied.

Hypothesis:

H0: There is no significant difference between the average sales of different levels of Outlet_Identifier, αi=0 ∀i

H1: αi ≠0 for atleast one I

Analysis:



```
> df3$outlet_Identifier=as.factor(df3$Outlet_Identifier)
> modele=lm(df3$Item_Outlet_Sales~df3$Outlet_Identifier,data=df3)
> anova(modele)
Analysis of Variance Table

Response: df3$Item_Outlet_Sales
                        Df    Sum Sq   Mean Sq F value    Pr(>F)    
df3$Outlet_Identifier    9 5.977e+09 664110448  300.08 < 2.2e-16 ***
Residuals             8513 1.884e+10   2213118                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

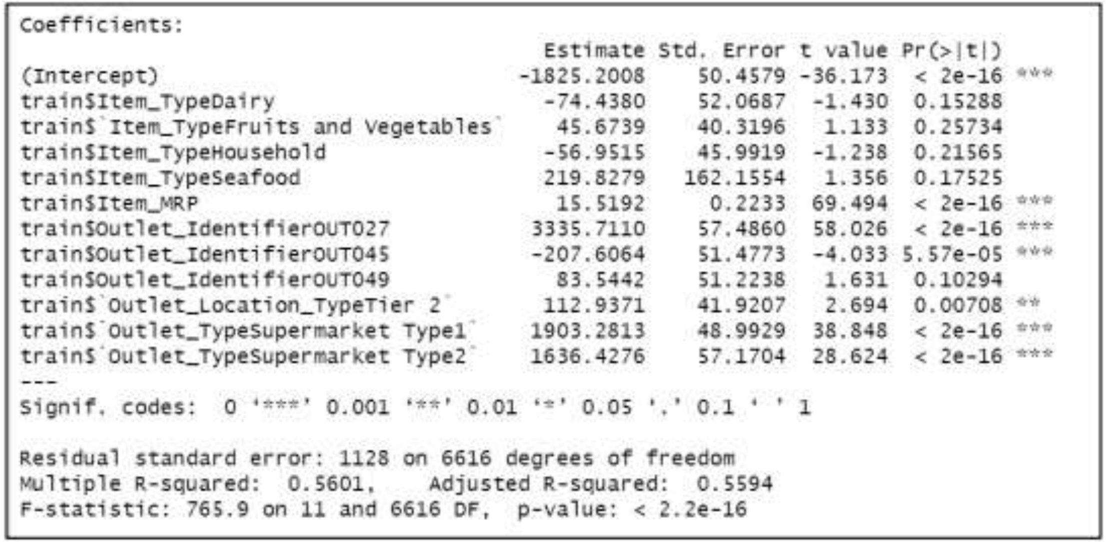
Inference:

Testing H0: αi=0 $\forall i$: p value = 2.2e-16 < 0.05, hence we reject H0 at 5% level of significance. This means αi is significant, all the levels of Outlet_Identifier do not have the same average sales

MODEL FITTING:

a) Multiple Linear Regression:

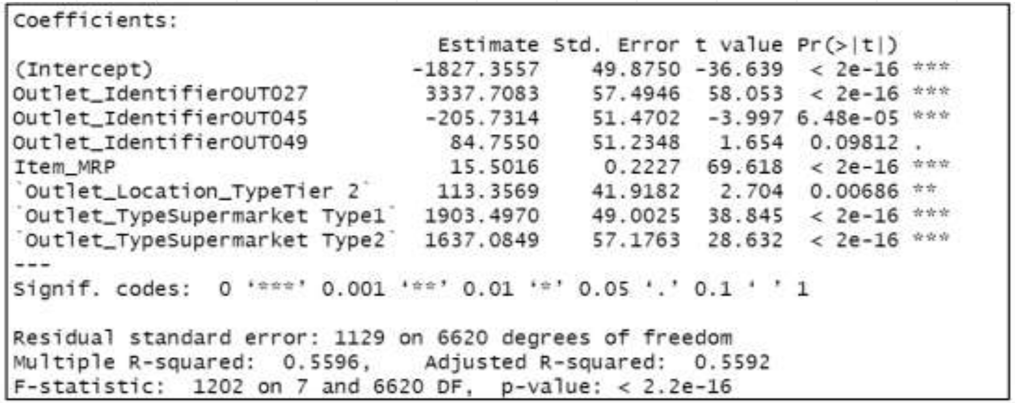
Fitting the model based on best subset selection:



```
Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)                            -1825.2008    50.4579 -36.173  < 2e-16 ***
train$Item_TypeDairy                     -74.4380    52.0687  -1.430  0.15288    
train$`Item_TypeFruits and Vegetables`    45.6739    40.3196   1.133  0.25734    
train$Item_TypeHousehold                 -56.9515    45.9919  -1.238  0.21565    
train$Item_TypeSeafood                   219.8279   162.1554   1.356  0.17525    
train$Item_MRP                            15.5192     0.2233  69.494  < 2e-16 ***
train$Outlet_IdentifierOUT027           3335.7110    57.4860  58.026  < 2e-16 ***
train$Outlet_IdentifierOUT045           -207.6064    51.4773  -4.033 5.57e-05 ***
train$Outlet_IdentifierOUT049             83.5442    51.2238   1.631  0.10294    
train$`Outlet_Location_TypeTier 2`       112.9371    41.9207   2.694  0.00708 ** 
train$`Outlet_TypeSupermarket Type1`    1903.2813    48.9929  38.848  < 2e-16 ***
train$`Outlet_TypeSupermarket Type2`    1636.4276    57.1704  28.624  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1128 on 6616 degrees of freedom
Multiple R-squared:  0.5601,	Adjusted R-squared:  0.5594 
F-statistic: 765.9 on 11 and 6616 DF,  p-value: < 2.2e-16
```

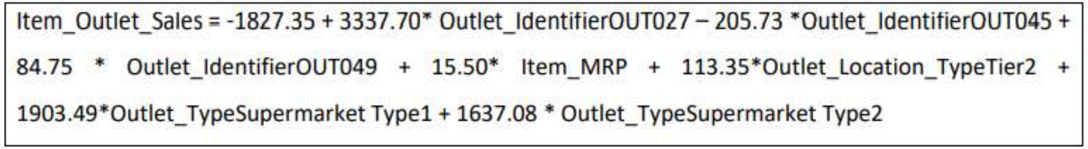
Reducing the model by selecting significant variables, our model becomes:



```
Coefficients:
                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  -1827.3557    49.8750 -36.639  < 2e-16 ***
Outlet_IdentifierOUT027       3337.7083    57.4946  58.053  < 2e-16 ***
Outlet_IdentifierOUT045       -205.7314    51.4702  -3.997 6.48e-05 ***
Outlet_IdentifierOUT049         84.7550    51.2348   1.654  0.09812 .  
Item_MRP                        15.5016     0.2227  69.618  < 2e-16 ***
`Outlet_Location_TypeTier 2`   113.3569    41.9182   2.704  0.00686 ** 
`Outlet_TypeSupermarket Type1` 1903.4970    49.0025  38.845  < 2e-16 ***
`Outlet_TypeSupermarket Type2` 1637.0849    57.1763  28.632  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1129 on 6620 degrees of freedom
Multiple R-squared:  0.5596,	Adjusted R-squared:  0.5592 
F-statistic:  1202 on 7 and 6620 DF,  p-value: < 2.2e-16
```

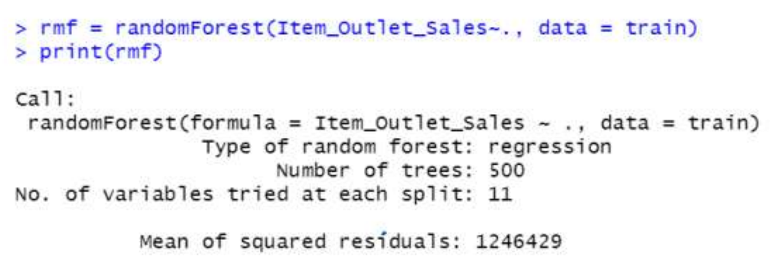
The Equation Becomes



Item_Outlet_Sales = -1827.35 + 3337.70* Outlet_IdentifierOUT027 – 205.73 *Outlet_IdentifierOUT045 + 84.75 * Outlet_IdentifierOUT049 + 15.50* Item_MRP + 113.35*Outlet_Location_TypeTier2 + 1903.49*Outlet_TypeSupermarket Type1 + 1637.08 * Outlet_TypeSupermarket Type2

The R2 for the Model is 0.5596 and Adj R2 is 0.5592 implying the proportionate variation caused to the response variable i.e Item_Outlet_Sales.

b) Random forest in Regression



```
> rmf = randomForest(Item_Outlet_Sales~., data = train)
> print(rmf)

Call:
 randomForest(formula = Item_Outlet_Sales ~ ., data = train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 11

          Mean of squared residuals: 1246429
```

- **R2 = 0.60**
- **RMSE : 1105**
- **MAPE : 0.53**

FINAL MODEL SELECTION:

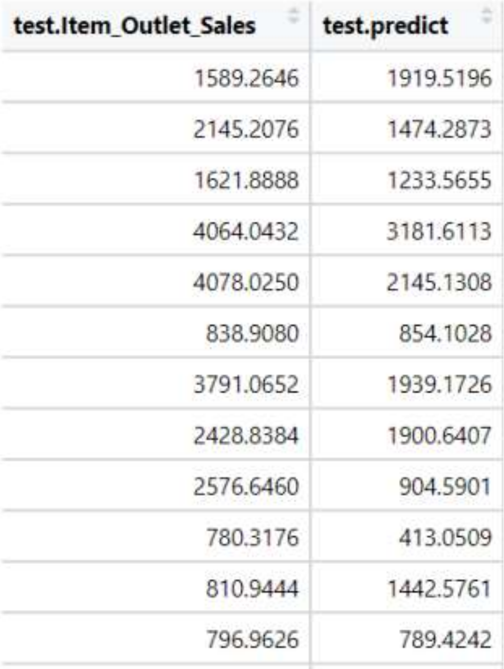
- Multiple linear regression after box cox transformation and Random Forest:

Evaluation	Multiple Linear Regression	Random Forest
R2	0.71	0.60
RMSE	1096	1105

Since R^2 is greater and RMSE is smaller for multiple Linear Regression, thus Multiple linear regression is selected as our final model.

PREDICTION:

Prediction implies forecasting values of the response variable on the basis of explanatory variables. We used the predict() function to make predictions from that model on test data. The test data has all of the columns from the training data, but they can be in a different order with different values. Here, rather than re-predicting on the training set, we predict on the test set.



test.Item_Outlet_Sales	test.predict
1589.2646	1919.5196
2145.2076	1474.2873
1621.8888	1233.5655
4064.0432	3181.6113
4078.0250	2145.1308
838.9080	854.1028
3791.0652	1939.1726
2428.8384	1900.6407
2576.6460	904.5901
780.3176	413.0509
810.9444	1442.5761
796.9626	789.4242

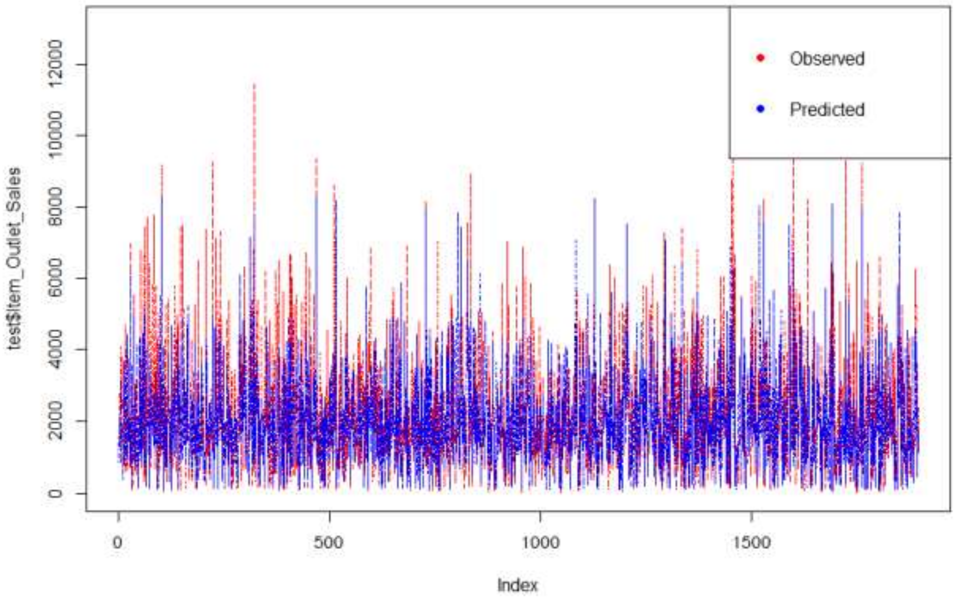
Evaluation Metrics:

Evaluation Metrics	Values
RMSE	1129.58
R^2	0.7157
$Adj\ R^2$	0.7154
MAPE	48%

INFERENCE:

- $\boldsymbol{R}^2$ $for\ both\ test\ and\ train\ data\ is\ almost\ equal.\ Hence\ our\ model\ is\ efficient.$

- Plot for the observed and predicted values is given below where the red lines are the observed values and blue lines are expected observations. We can see the blue lines are almost covering all the red lines giving us a insight that the model has a good accuracy.



CONCLUSIONS:

- **Item weight** and **Item_Establishment_Year** do not give any meaningful insight hence may not have any relationship with Item_Outlet_Sales.

- The effect of factor **Item_fat_content** on **Item_Outlet_Sales** is not found to be significant at 5% Level Of Significance. Hence, **the average Item_Outlet_Sales is same for low, regular fat and Not applicable products**. Therefore, it may not be an important factor for forecasting Item_Outlet_Sales.

- Big mart has maximum number of products with visibility between 0-0.1. Increase in **Item_visibility** can decrease the item outlet sales because it is having negative correlation.

- There is a significance difference in the Item Outlet sales of different **Item_type**. So this can be important for predicting Item Outlet sales. Consumable items contribute to more sales followed by no consumable. In consumable items, the maximum sales have been generated by mainly Baking Goods, Sea food, Starchy food and other food items.

- The effect of factor **Outlet_Identifier** on Item_Outlet_Sales is found to be significant at 5% Level of Significance. Hence, this means all the levels of Outlet_Identifier do not have the same average sales.

- The effect of factor **Outlet_Size** on Item_Outlet_Sales is found to be significant at 5% Level Of Significance. Hence, **the average Sales is not same for different outlet size**. From the graph, Medium size stores have maximum sales while high has minimum sales. Thus, Customers might prefer Outlets of medium size more as compared to high. **Hence, Big Mart should prefer Medium size stores over high**.

- The effect of **Outlet_Type** on Item_Outlet_Sales is found to be significant at 5% Level of Significance. Hence, the average Sales is not same for different Outlet Type. **Supermarket Type 1 has more Sales**.

- FINAL MODEL selected was **Multiple Linear Regression**:

The findings reveal that the factors that significantly affect the Item_Outlet_Sales are :

➢ Outlet Identifier OUT27
➢ Outlet Identifier OUT45
➢ Outlet Identifier OUT49
➢ Outlet Location Type Tier2
➢ Outlet Type Supermarket Type 1
➢ Outlet Type Supermarket Type 2
➢ Item_MRP

- Rmse for the model is 1129. 58 and MAPE is 48%. Implying a good model accuracy, since our response variable is continous.

- The final model satisifes all the assumptions i.e. linearity, normality, homoscedasticity and independence of residuals.

REFERENCES :-

- Data: https://www.kaggle.com/brijbhushannanda1979/bigmart-sales-data
- https://www.rpubs.com/shubh2565/bigmart-dataset-analysis
- https://www.kaggle.com/aishwarya2490/bigmart-sales-prediction-model
- https://d1wqtxts1xzle7.cloudfront.net/64640730/IRJET-V7I6676-with-cover-page-v2.pdf?Expires=1636397544&Signature=OSS2bFNJkuZ9IVLm6ZNxRZwOpOz74XeJzOv6M5g

Thank

You