**DSCI 5340-001, 003: Predictive Analytics and Business Forecasting**

Project 1 Report

BigMart Outlet Sales Analysis

Nishith Reddy Mannuru: 11467417

**BIGMART ITEM OUTLET SALES PREDICTION MODEL**

I have identified this dataset in Kaggle. [Link](https://www.kaggle.com/brijbhushannanda1979/bigmart-sales-data?select=Train.csv).

Contents

[INTRODUCTION AND BUSINESS OBJECTIVE 3](#_Toc67243312)

[DATA DICTIONARY 3](#_Toc67243313)

[LITERATURE REVIEW: 4](#_Toc67243314)

[HYPOTHESIS GENERATION WITH RESPECT TO PROBLEM STATEMENT 5](#_Toc67243315)

[IMPORTING DATASET INTO SAS 5](#_Toc67243316)

[CHECKING THE MISSING VALUES IN THE DATASET 6](#_Toc67243317)

[CLEANING UP THE DATASET 8](#_Toc67243318)

[RE-CHECKING THE MISSING VALUES IN THE CLEANED DATASET 9](#_Toc67243319)

[EXPLORATORY DATA ANALYSIS (EDA) 10](#_Toc67243320)

[CORRELATION 13](#_Toc67243321)

[REGRESSION MODEL WITH THE QUANTITATIVE VARIABLES 14](#_Toc67243322)

[ModelImprovement 17](#_Toc67243323)

[DUMMY VARIABLE ENCODING FOR THE CATEGORICAL VARIABLES 17](#_Toc67243324)

[RUNNING A REGRESSION MODEL WITH THE QUANTITATIVE VARIABLES ALONG WITH THE CATEGORICAL VARIABLES (DUMMY VARIABLES) 18](#_Toc67243325)

[RUNNING A STEP-WISE REGRESSION 20](#_Toc67243326)

[RUNNING A REGRESSION MODEL WITH ONLY THE VARIABLES GENERATED NY STEPWISE REGRESSION 21](#_Toc67243327)

[RESIDUAL ANALYSIS 23](#_Toc67243328)

[CHECKING IF THE RESIDUAL FOLLOWS A CONSTANT VARIANCE 25](#_Toc67243329)

[TRANFORMING THE TARGET VARIABLE TO MAKE THE RESIDUALS FOLLOW CONSTANT VARIANCE 26](#_Toc67243330)

# INTRODUCTION AND BUSINESS OBJECTIVE

The main aim is to build a predictive model which can predict the item outlet sales of each product at a particular outlet. By performing this model, the department of BigMart Sales can understand the Item\_Outlet\_Sales for each unique product by analyzing different parameters that are affecting. If a new product is to be introduced to the market, for example, Quinoa, then the department wants to understand how the sales are effective due to different factors.

# DATA DICTIONARY

This data set contains a set of **8,523 records** under **12 attributes**.

The attributes being as follows:

|  |  |
| --- | --- |
| **Attribute Name** | **Description** |
| Item\_Identifier | Unique Identification number for each product |
| Item\_Iight | Iight of item or product |
| Item\_Fat\_Content | Checks the Concentration of fat in the product either LOW FAT or HIGH FAT |
| Item\_Visibility | The % of visibility of the product in the store |
| Item\_Type | Product Category |
| Item\_MRP | Maximum Retail Price for a Product |
| Outlet\_Identifier | Store ID |
| Outlet\_Establishment\_Year | The year when store was opened |
| Outlet\_Size | The size of the store (Area Size Category)  High/Medium/Small |
| Outlet\_Location\_Type | In Terms of city Tiers (Size) |
| Outlet\_Type | Grocery store or a type of supermarket |
| Item\_Outlet\_Sales | Sales of the product In the Specific outlet |

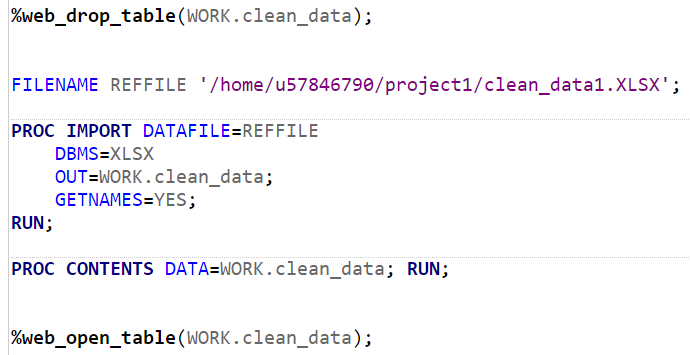
# LITERATURE REVIEW:

* According to Forst (2011), in the research paper both multiple regression and also Box-Jenkins’s analysis to perform sales forecasting for a restaurant. In the model, Forst used 3 predictors, to predict Iekly sales and concluded that multiple regression was the best for forecasting sales. Forst performed different models and considered R square value for determining the model’s variability. Forst conducted seven different types of models where sixth model was considered as best based on all the significance parameters. Forst also calculated the correlation matrix and made use of Dummy variables. The idea was appreciated, and similar approach was folloId in our analysis as Ill.
* In this paper the author H. M. Al-Hamadi was forecasting long term electric poIr load using fuzzy linear regression technique. Forecasting electric load plays a major role in predicting the future electric load and peak load demand. H. M. Al-Hamadi believed using random data and exogenous variables such as Iather-related variables, time factors, economics and also random events can help his predictions to be more accurate. I folloId the same methodologies of using random quantitative and categorical variables along with dummy variables encoding to boost the accuracy of your predictions.
* In this paper the authors talk about the combing content – based and collaborative filters in an online newspaper. Collaborative filtering puts together the options through which humans can make their personal decisions. In this paper the authors put together a new filtering approach which combines the total coverage and the content filters. Ana a similar filtering approach was folloId in our work too. I used filtering approach to filter the categorical variables into groups to maximize the effect of forecasting Item\_Outlet\_Sales.
* In this paper the authors Das, P., Chaudhury, S. predicts the retail sales of footIar using feedforward and recurrent neural networks. They believed fluctuations of any sales over time is one of the major problems faced in the industries. They tried to overcome the problem by taking a neural network approach. The data they worked with is a Iekly retail sale of the particular product. I tried to address the Item\_Outlet\_Sales problem the same way, by gathering the yearly sales information at different outlet locations.
* In this paper the authors Ni, Y. and Fan, F predicted the dynamic sales forecasting model for the fashion retail. They believed that forecasting fashion retail is difficult with number of factors such as season, the region, and the fashion trend. So, they made predictions considering both short term and long term with an adjustment model. I have also folloId the same mechanism while addressing the forecasting issues regarding big market sales prediction. And I have overcome those issues using dynamic mechanism.

# HYPOTHESIS GENERATION WITH RESPECT TO PROBLEM STATEMENT

* The **NULL HYPOTHESIS** in our problem statement is to find if Item\_Outlet\_Sales depend on quantitative variables like Item\_MRP, Item\_Iight, Item\_Visibility. I wanted to check if the prediction of sales increases depending on these numerical variables. For example, If the department plans to introduce a new product like Quinoa, then how the outlet sales differ for the product based on different parameters. So, I will reject null hypothesis if I find that I need other variables like categorical and check the behavior of the outlet sales.
* **ALTERNATE HYPOTHESIS** being not only quantitative variables, but also categorical variables are used to predict the Item\_Outlet\_Sales.

# IMPORTING DATASET INTO SAS



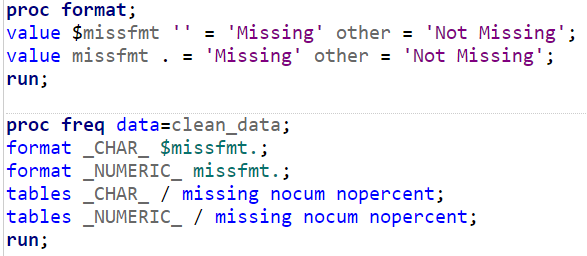


# CHECKING THE MISSING VALUES IN THE DATASET

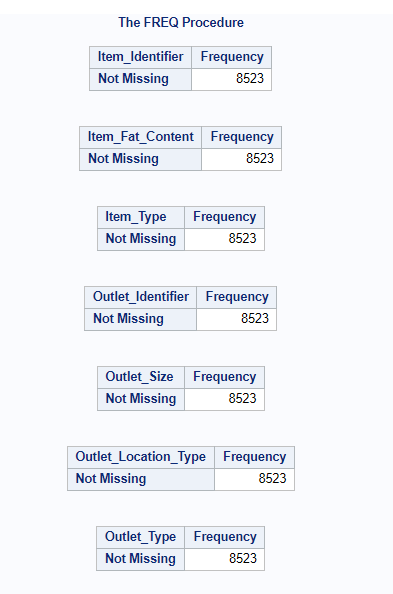
Cleaning is one of the important steps before I run any mode. It is a step of data preparation. The cleaner data I have, the more accurate the model can predict the data. There are many ways to clean the data like making use of mean, median or mode if the missing variable is quantitative. Also, I can predict the missing values sometimes using the existing data points.

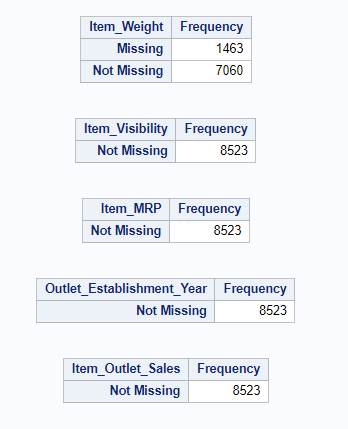
All in all, one must understand the domain of the business case and then accordingly perform the missing value imputation.

SAS CODE – Took reference from the video [Link](https://www.youtube.com/watch?v=JQHAA_AGxrE&t=411s):



OUTPUT





I could see that the item Iight attribute has 1463 missing observations. So, now I should clean up the data to address those missing observations.

As it is Item\_Iight is quantitative type of variable I have used mean method to replace the missing values.

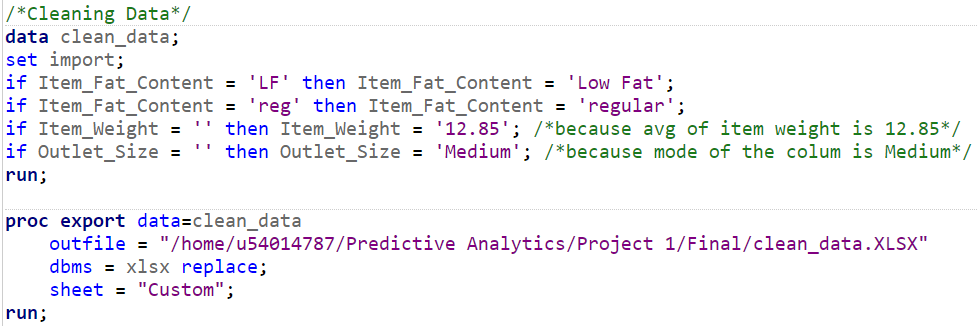
# CLEANING UP THE DATASET

Since the attribute item Iight has 1463 missing values in it, I are replacing all those missing values with “12.85” which is the average item Iight of all the observation excluding the missing values.

Similarly, the column outlet sales have some missing values too. Therefore, I are replacing the missing values with Medium, considering the mode of the column.

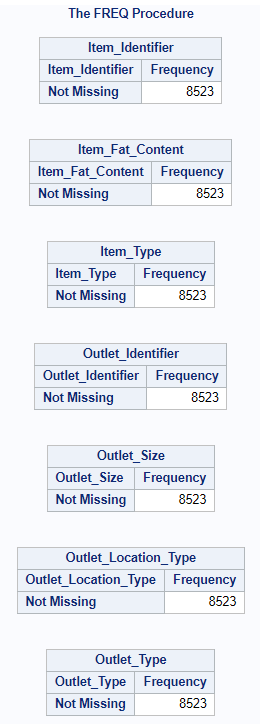
And then I are exporting the cleaned-up dataset into a new excel file, named “clean\_data.XLSX”, and I’ll I are working with this dataset.

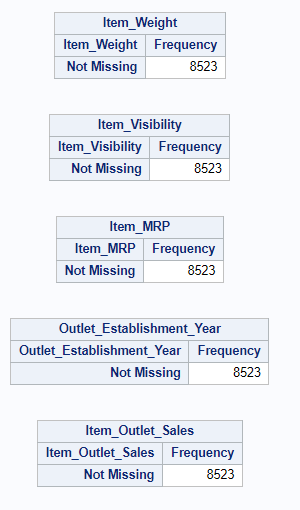
SAS CODE



# RE-CHECKING THE MISSING VALUES IN THE CLEANED DATASET

OUTPUT –



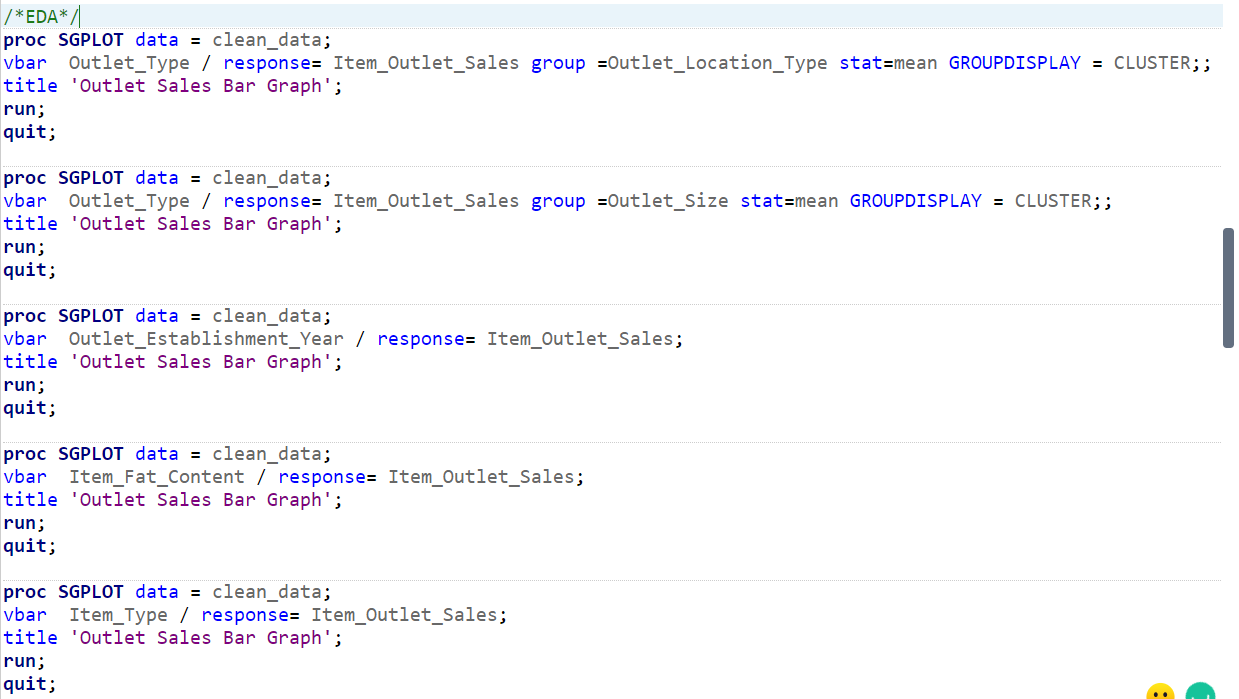


Now, as you can observe there are no missing values in the dataset, and our data set is perfect to be used.

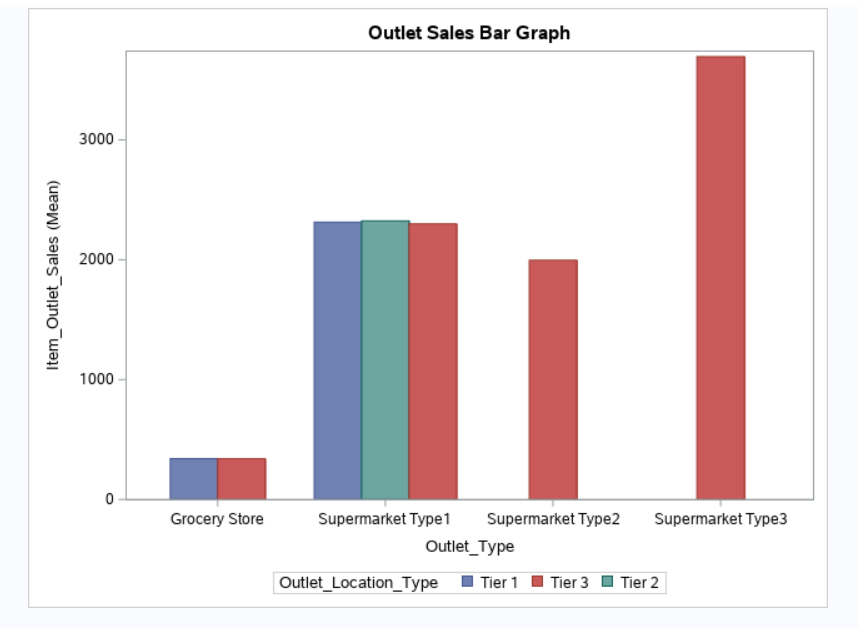
# EXPLORATORY DATA ANALYSIS (EDA)

Now, I are conducting exploratory data analysis to figure out in what ways the data sets are similar or different. I used bar graphs to analyze them better.

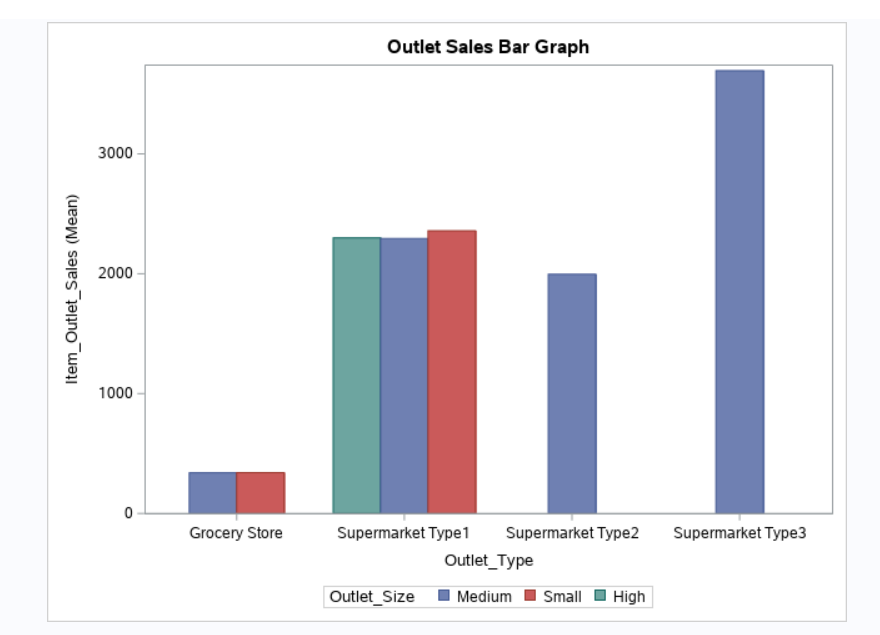
SAS CODE



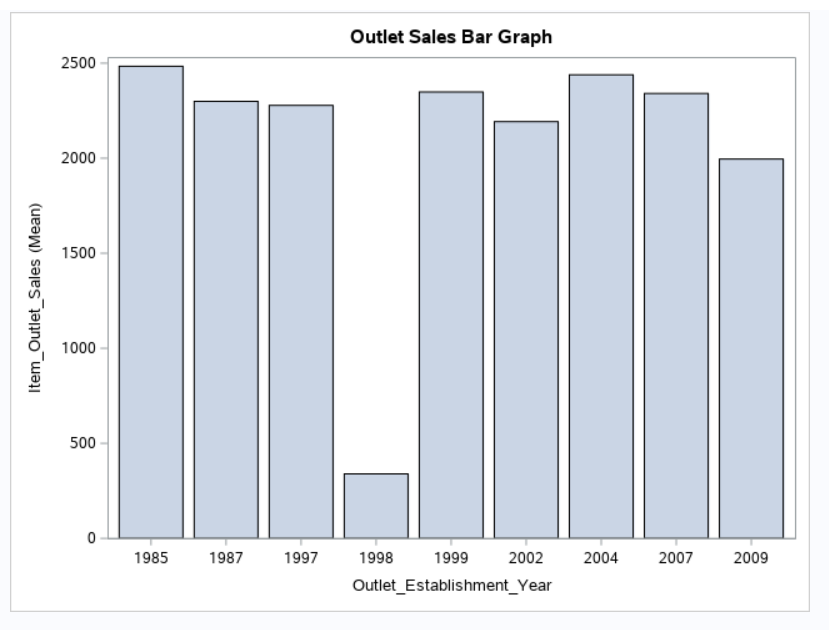
OUTPUT

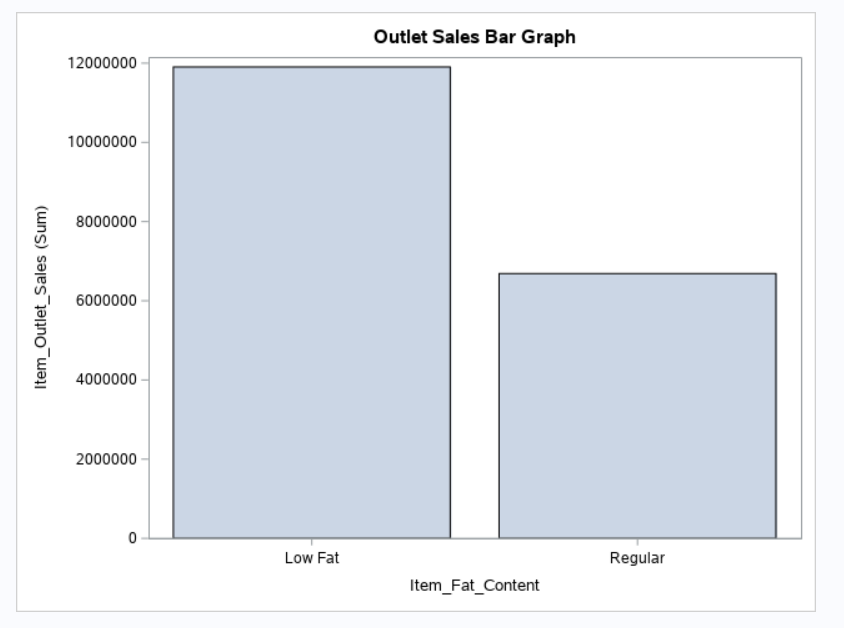


This is a very interesting observation as I can see here that only Supermarket Type1 has all 3 types of locations where as Supermarket type2 and 3 has only Tier 3 location associated with it. This implied that these variables are very important in our sales analysis.

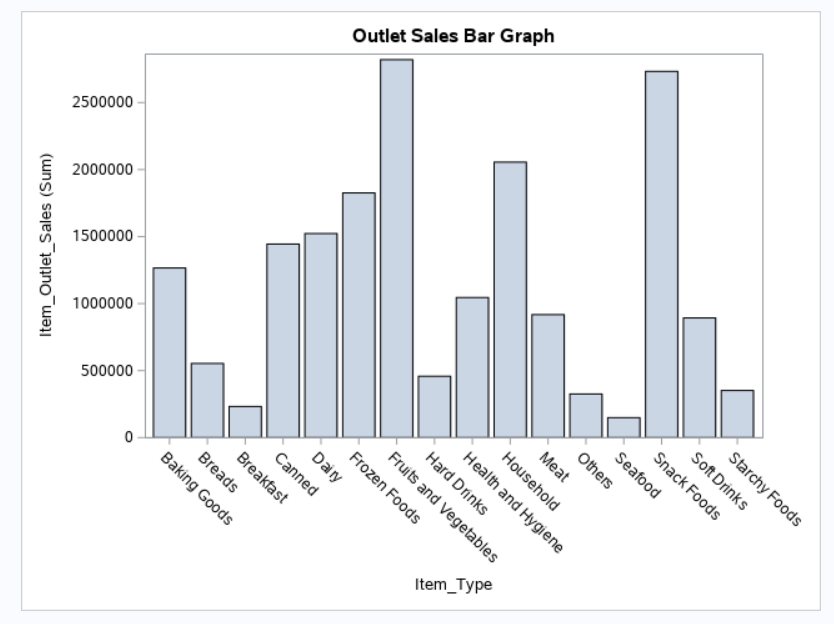


The same goes with the above bar graph as Ill. I can see that only medium size of outlet is associated with Supermarket type 2 and type 3.





Above are few more observations which can help in further in our analysis. The first figure tells us how 1998 year has been the loIst effected sales and also in the second figure it says low fat foods are preferred by customers more hence products are more with low fat.



Based on the category types, fruits and vegetables and snacks foods are more into the market because customers are purchasing more related these products.

Above all are few of our observations before I proceeded with data model building.

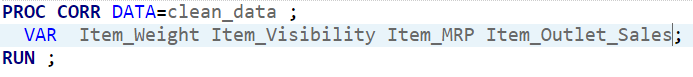
# CORRELATION

I would like to see if there are any variables which are highly correlated with each other.

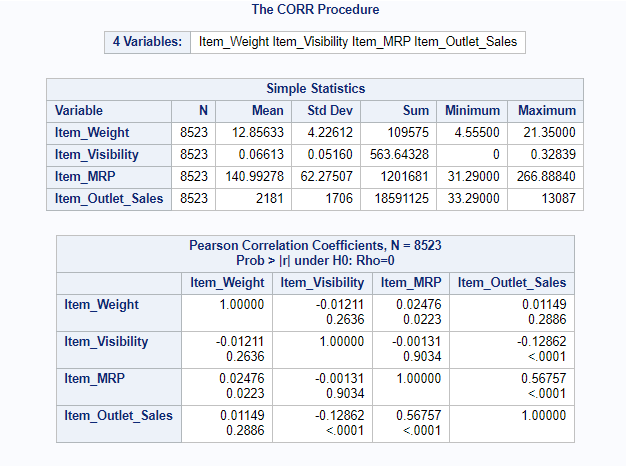
Now, I are going to use the correlation to measure the strength of linear relationship betIen the quantitative variables.

The quantitative variables I have are Item\_Iight, Item\_Visibility, Item\_MRP and Item\_Outlet\_Sales

SAS CODE



OUTPUT



***Inference:*** I can see that the variables are not much highly correlated to each other. The highest correlation is 0.56. Hence, I can use all the variables while building the model. There is no variables which is closer to -1 or 1 which indicates that I are not overfitting the model if I use these variables. For example: if 2 variables are highly correlated and If I use both of them to predict the sales, then this results in overfitting.

# REGRESSION MODEL WITH THE QUANTITATIVE VARIABLES

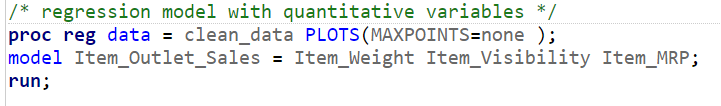
As our target variable is continuous variable, I would like to perform regression model and analyze the dependency of target variable on independent variables.

For our initial analysis, I are taking 3 quantitative variables as input: They are:

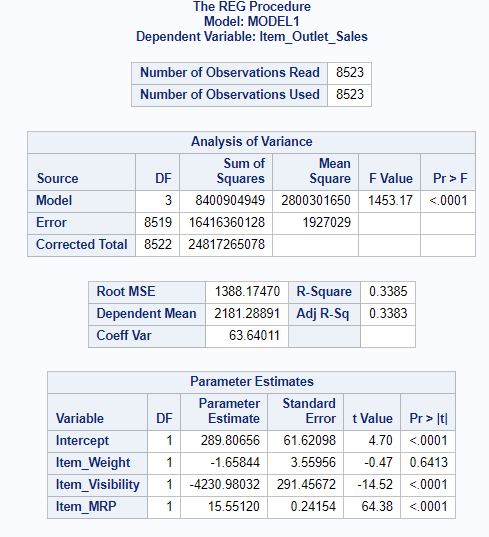
1. Item\_Iight
2. Item\_Visibility
3. Item\_MRP

Putting together a regression model to see if my target variable (Item\_outlet\_sales) can be predicted with input variables.

SAS CODE

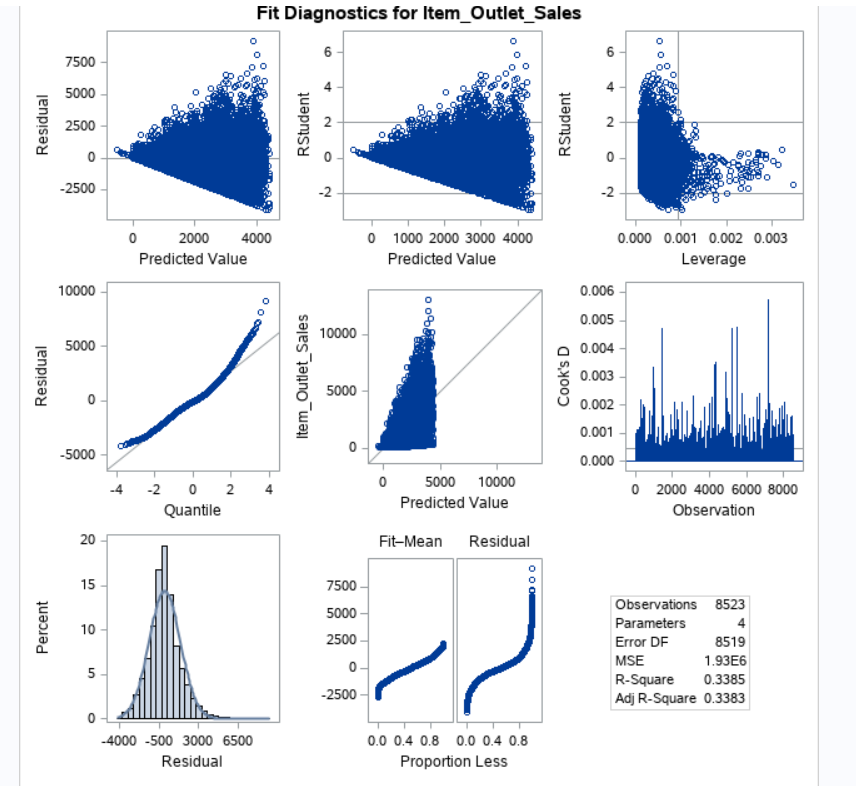


OUTPUT



***Inference:*** I can see that variables Item\_Visibility and Item\_MRP are significant because their P value is less than 0.05 where as variable Item\_Iight is 0.64 Hence this variable is not significant in predicting the target variable in our model

Also, it is to note that the R2 value is only 33% which indicates that 33% of the variability is explained by our model. Hence, I need to add more variables to see if it increases the R2  which can explain more variability.





* When you observe the output, the R square value is 0.3385. Which is 33%.
* Since the R square is so low, I could conclude or state that the item\_outlet\_sales cannot be significantly predicted with only quantitative variables (Item\_Iight, Item\_Visibility, Item\_MRP).

# ModelImprovement

From our previous analysis I have seen that only few variables Ire significant but there are other categorical variables too which I can make use of to make our model predict better.

As I cannot use categorical variables in regression model, I need to transform the categorical variable into numeric by doing hot encoding (Dummy variable encoding)

# DUMMY VARIABLE ENCODING FOR THE CATEGORICAL VARIABLES

* Since the R square value of the model when using only quantitative variables really low, I are going to use categorical variables, to see if it increases our chances of predicting the item\_outlet\_sales.
* I are going to use dummy variable encoding for the categorical variables that I have in the dataset.

The categorical variables in the data set are:

1. Item\_Fat\_Content
2. Outlet\_Location\_Type
3. Outlet\_Size
4. Outlet\_Type.

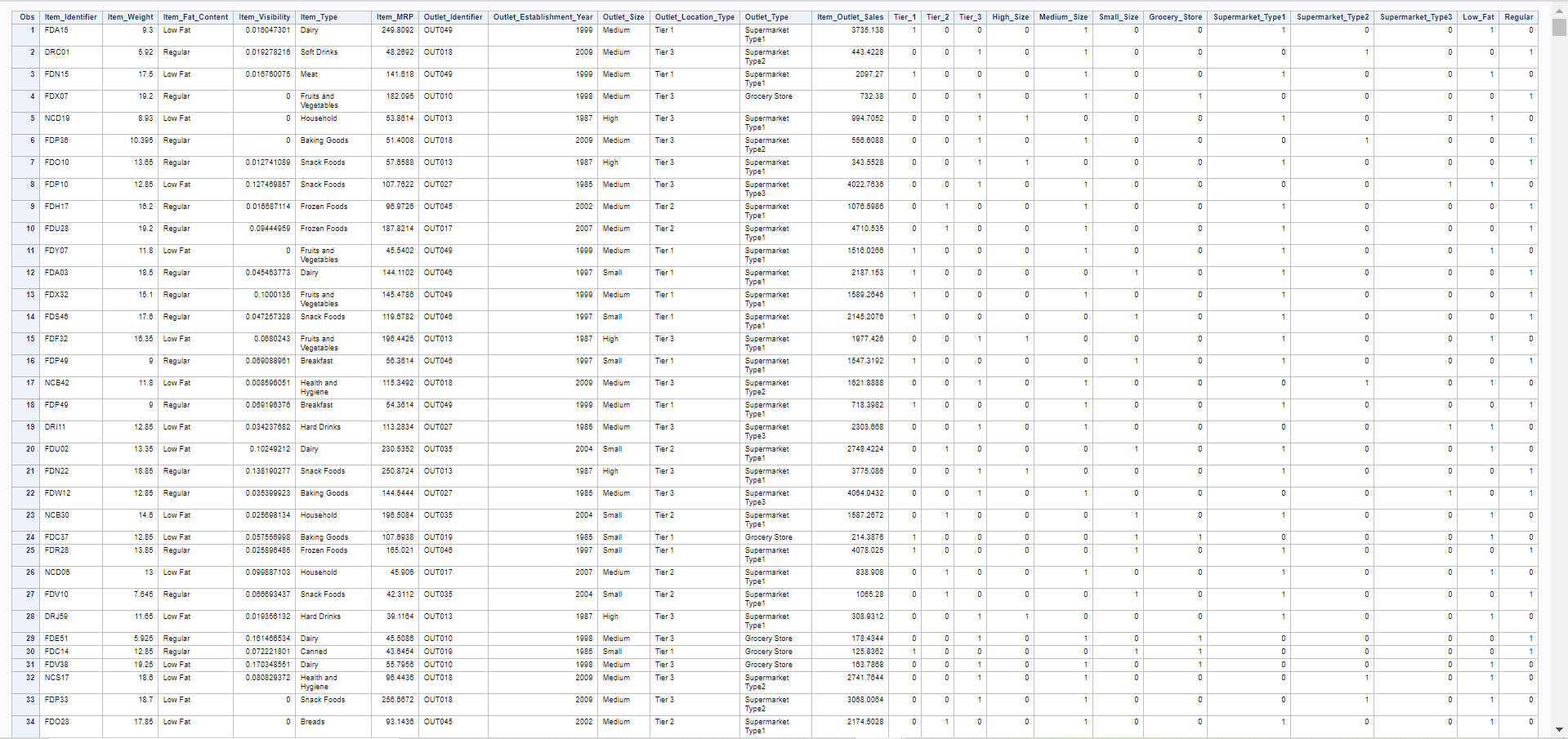
SAS CODE



proc print data = import1;

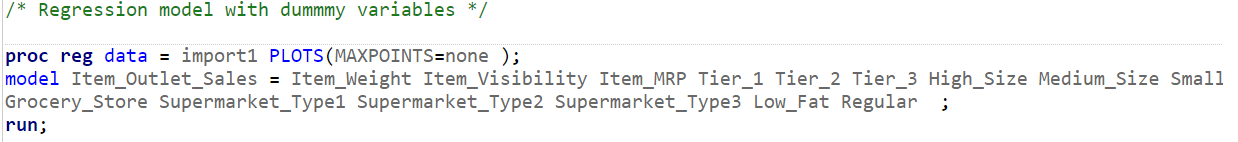
run;

OUTPUT

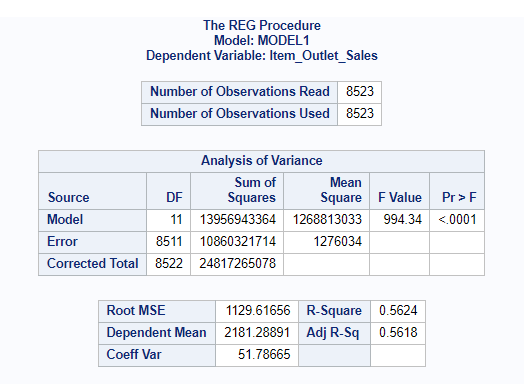


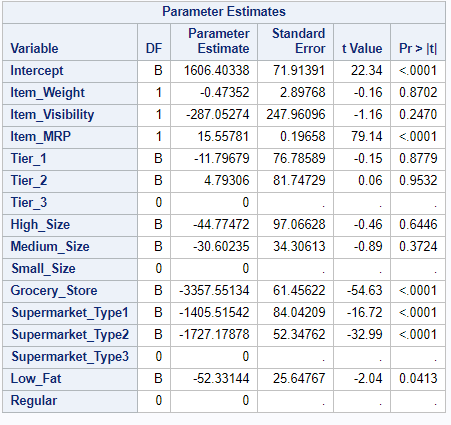
# RUNNING A REGRESSION MODEL WITH THE QUANTITATIVE VARIABLES ALONG WITH THE CATEGORICAL VARIABLES (DUMMY VARIABLES)

SAS CODE



OUTPUT



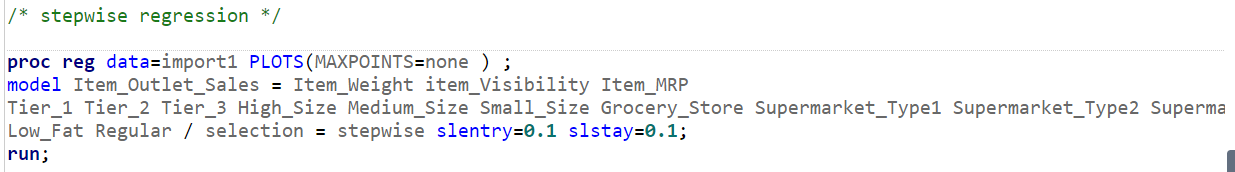


* When you take a look at the R square now, it has been increased to 0.5624. Which is 56%.
* I could conclude or state that the prediction of item\_outlet\_sales has significantly increased when considering quantitative variables along with categorical variables.
* I can also see that many variables are not so significant if I use all the categorical dummy variables. Few variables have p-value greater than 0.05.
* Significant variables are: Item\_MRP, Grocery\_Store, Supermarket\_type1, Supermarket\_type2 and Low-fat.

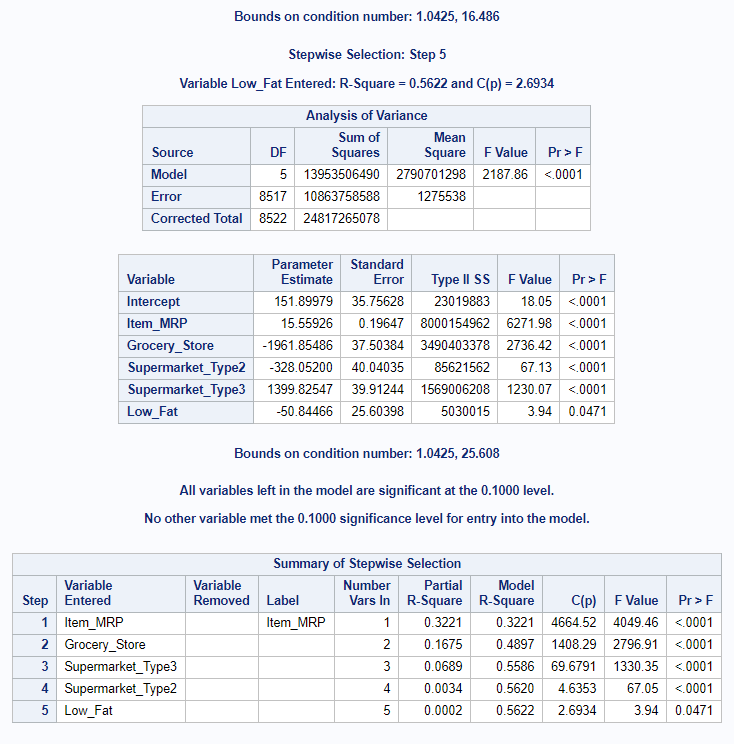
# RUNNING A STEP-WISE REGRESSION

* As I have seen that most of variables Ire not significant, I wanted to explore stepwise regression and use only those variables as input variables which seem to be significant.
* Hence in this step, I will be running a stepwise regression to identity which exact independent variables are being used to predict item\_outlet\_sales significantly.

SAS CODE



OUTPUT

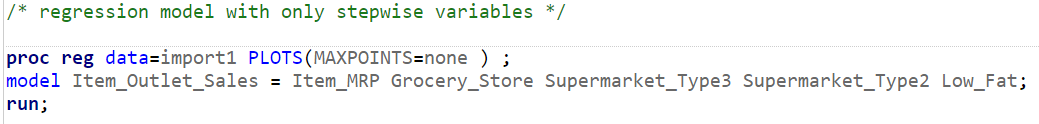


* After stepwise regression, I can observe that the only variables that are a significant measure of item\_outlet\_sales (target variable) are Item\_MRP, Grocery\_Store, Supermarket\_Type3, Supermarket\_Type2 and Low\_Fat.
* Even Though R square remains the same at 0.5622. that is 56%.
* I can only use these independent variables to run the regression model and I will still get the same R square value.

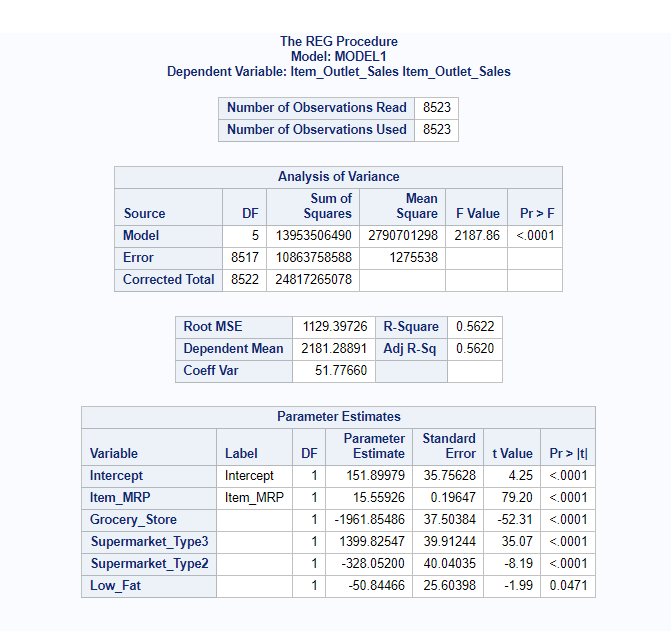
# RUNNING A REGRESSION MODEL WITH ONLY THE VARIABLES GENERATED NY STEPWISE REGRESSION

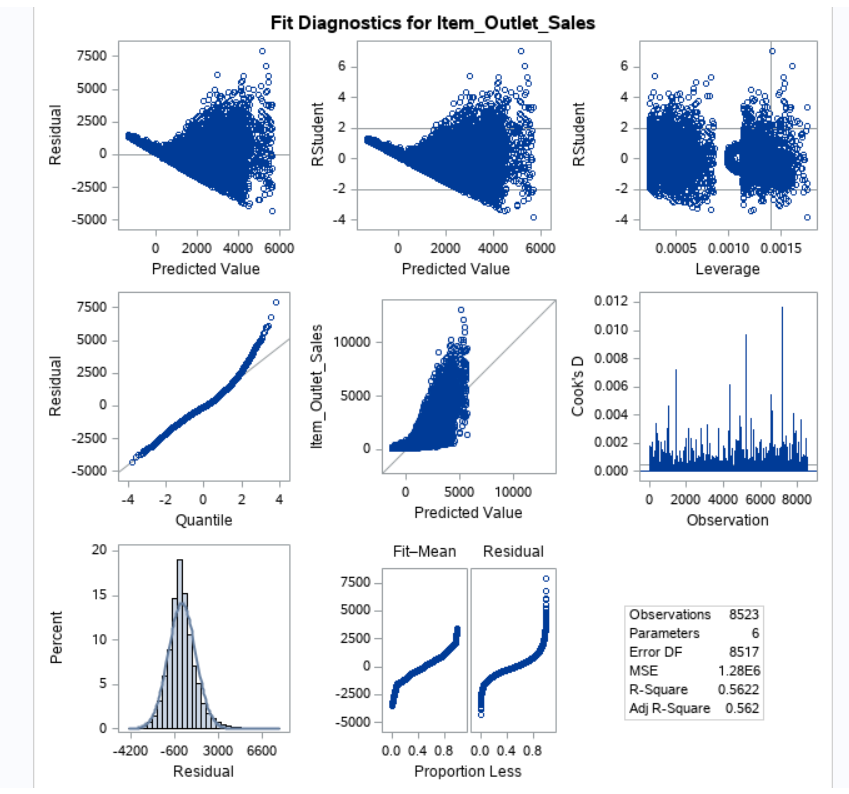
* Running a regression model with only Item\_MRP, Grocery\_Store, Supermarket\_Type3, Supermarket\_Type2 and Low\_Fat to check if I are getting the same R square value.

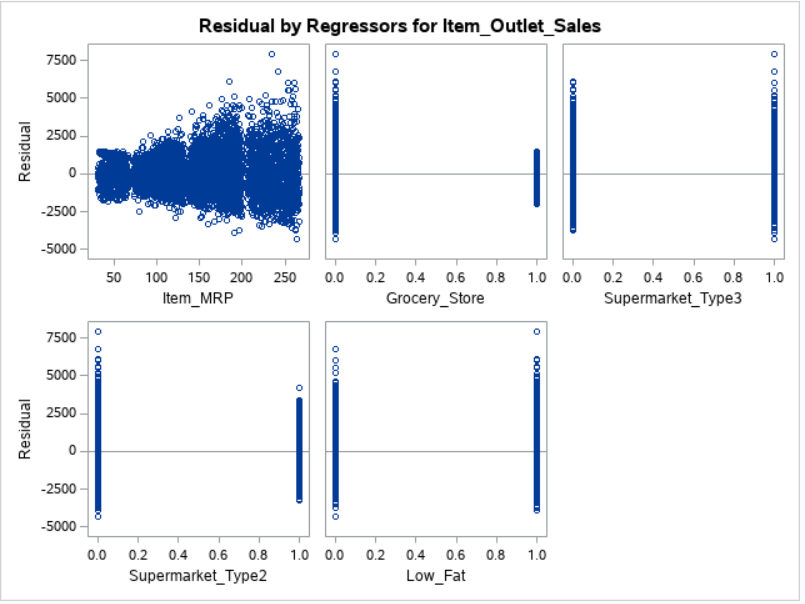
SAS CODE



OUTPUT







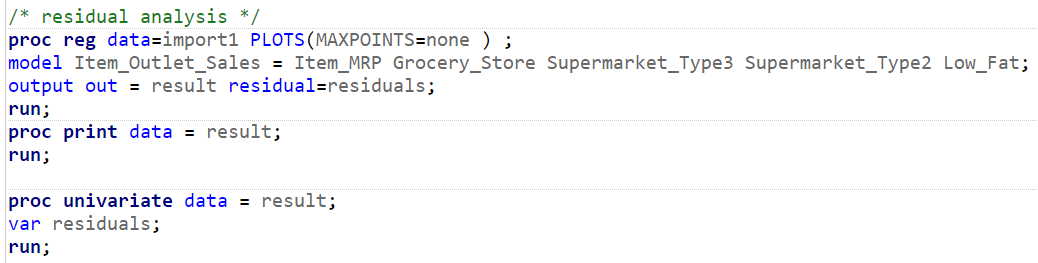
* As you observe the R square is same as the R square value, I got from the stepwise regression model. i.e., 0.5622. Which is 56% and also the variables all are <0.001 hence all variables are significant.

# RESIDUAL ANALYSIS

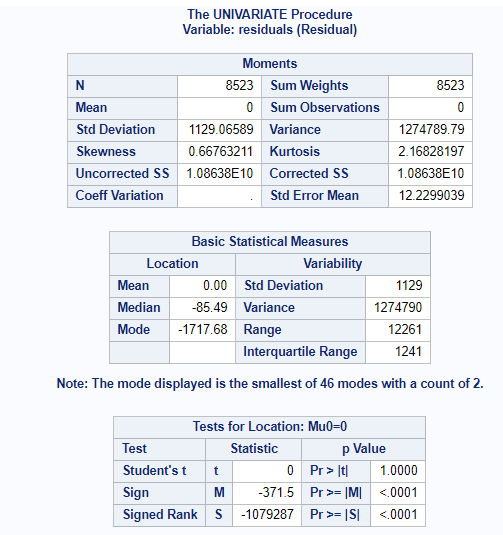
* Residuals should have a mean zero and should follow constant variance.

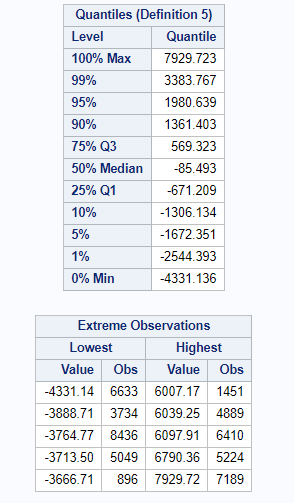
**CHECKING THE MEAN OF RESIDUALS**

SAS CODE



OUTPUT





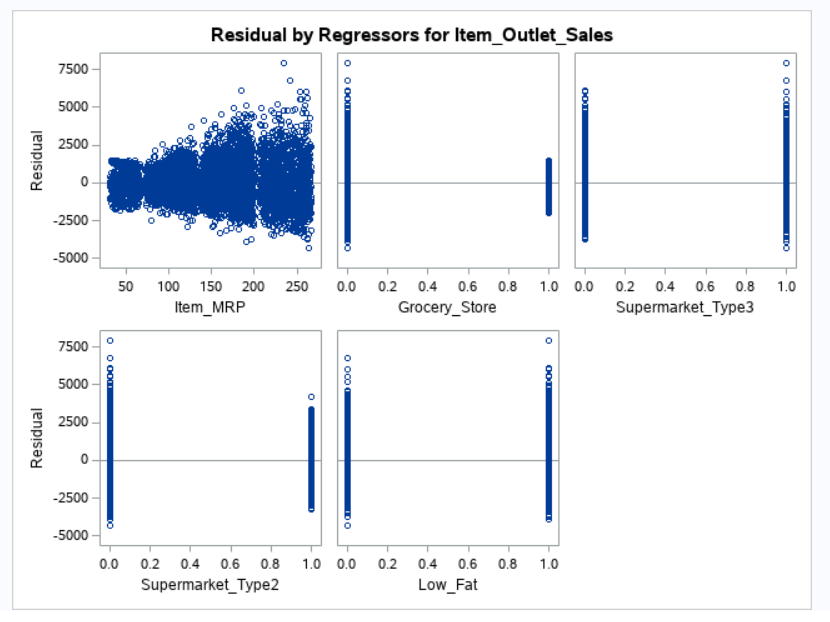
* As you observe, the mean is zero. And hence its perfect.

# CHECKING IF THE RESIDUAL FOLLOWS A CONSTANT VARIANCE

SAS CODE

proc reg data=import1 PLOTS(MAXPOINTS=none);  
model Item\_Outlet\_Sales = Item\_MRP Grocery\_Store Supermarket\_Type3 Supermarket\_Type2 Low\_Fat;  
run;

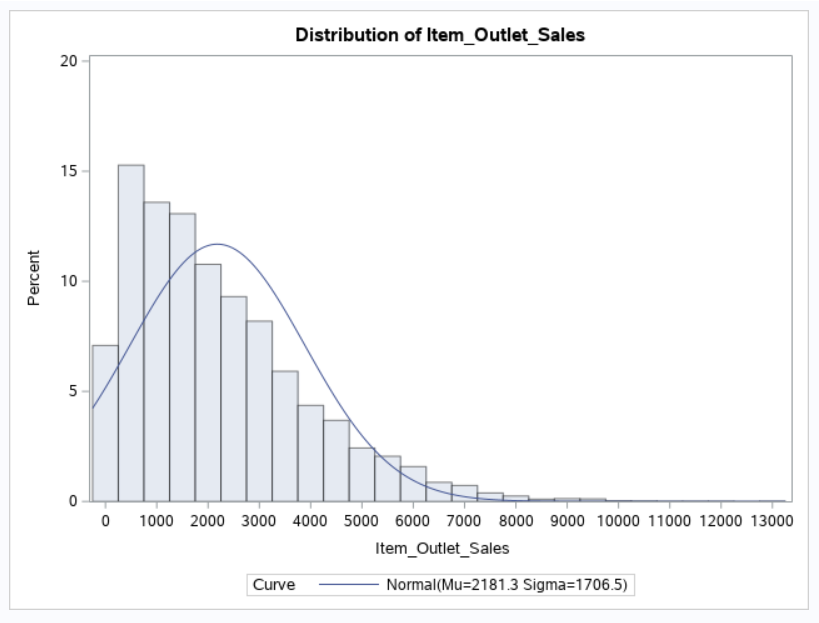
OUTPUT



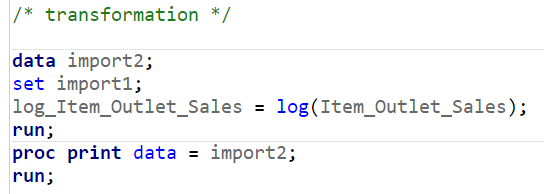
* The residual when plotted versus Item\_MRP and Grocery\_Store, it is not following constant variance which is one of the assumptions of linear regression.

# TRANFORMING THE TARGET VARIABLE TO MAKE THE RESIDUALS FOLLOW CONSTANT VARIANCE

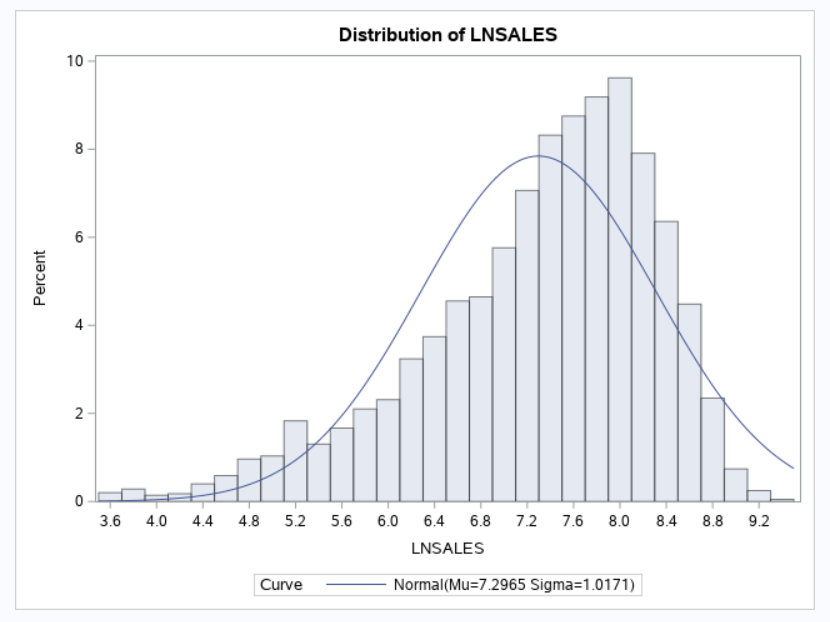
**Before transformation normality check:**



SAS CODE

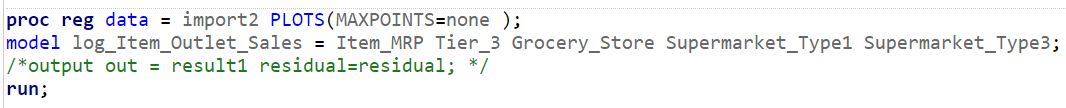


**After Transformation normality check:**

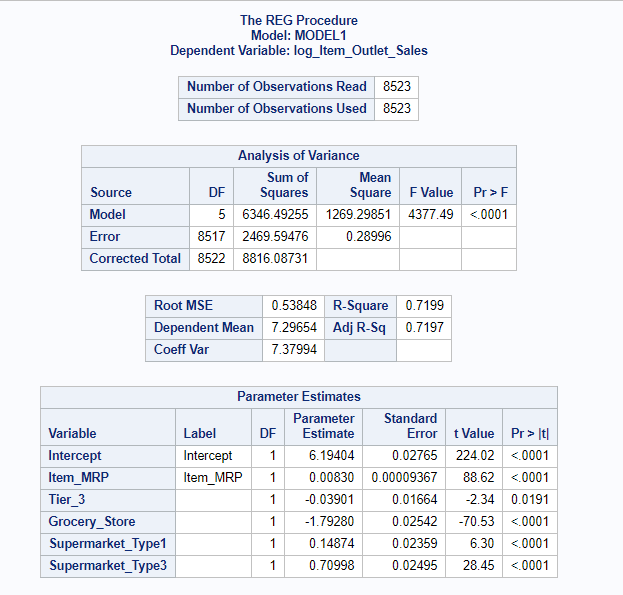


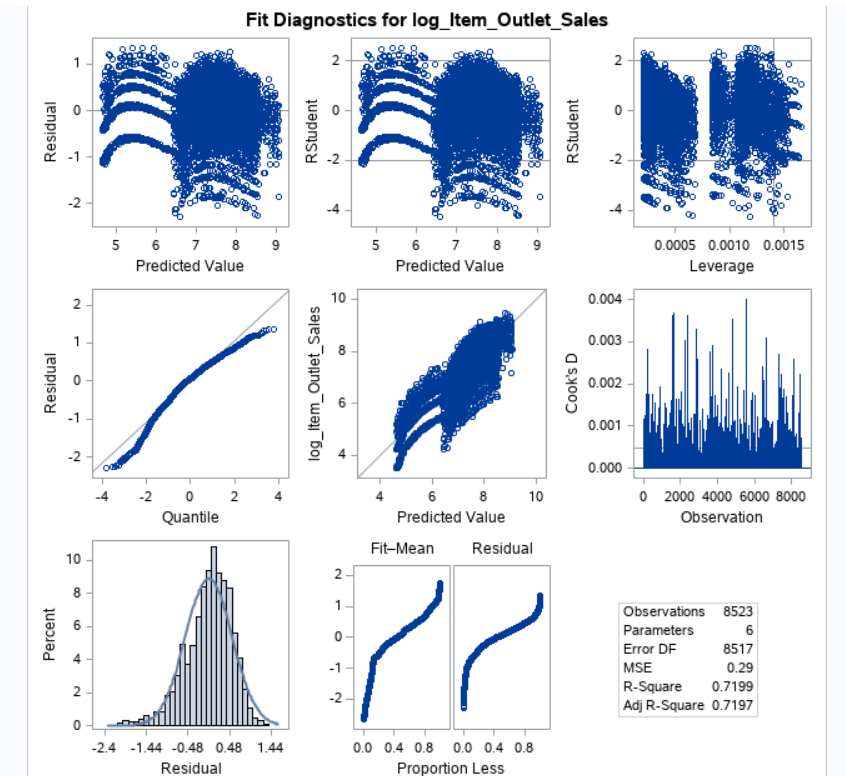
**RUNNING A REGRESSION MODEL WITH THE TRANSFORMED TARGET VARIABLE**

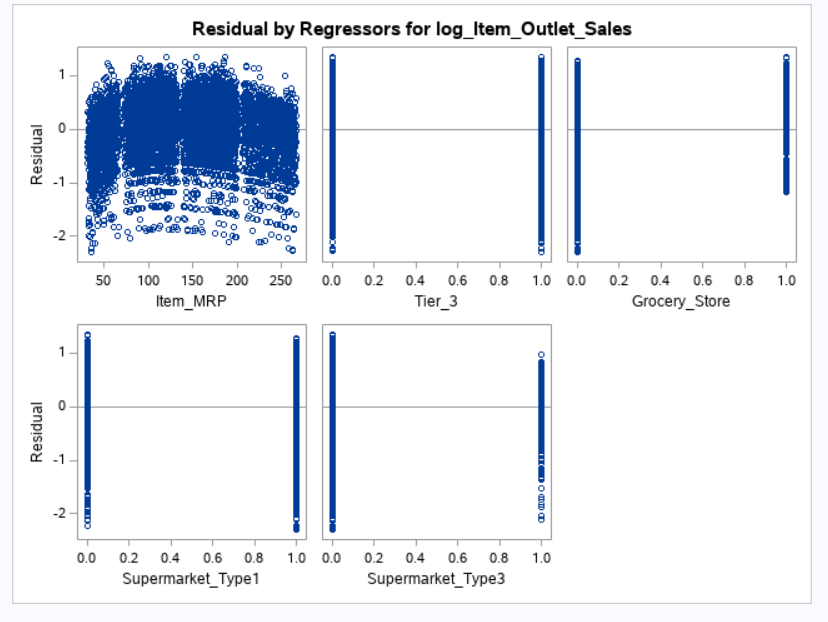
SAS CODE



OUTPUT –







* After log transformation, our R square value has significantly increased to 0.7199. which is 71%.
* And, when you take a look at the residuals right now, it has a constant variance.

**CONCLUSION**

* I reject the null hypothesis because all the P values obtained for the independent variables are less then alpha (0.05). And I also conclude that not only quantitative variables but also categorical variables like Grocery Stores, Tier 3, Supermarket type 1 and type 3 depends on the effect of outlet sales.
* Hence, I suggest adding quantitative variables as Ill as categorical variables as like Item MRP, Grocery Stores, Tier 3, Supermarket type 1 and Supermarket type 3 has achieved more accurate results from the model.

**References:**

Forst, Frank. (2011). *Forecasting Restaurant Sales Using Multiple Regression And Box-Jenkins Analysis.* Journal of Applied Business Research (JABR). 8. 15. 10.19030/jabr.v8i2.6157.

H. M. Al-Hamadi “Long-Term Electric PoIr Load Forecasting Using Fuzzy Linear Regression Technique”, IEEE Mar.2011

Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., Sartin, M.: Combing content-based and collaborative filters in an online newspaper (1999)

Das, P., Chaudhury, S. Prediction of retail sales of footIar using feedforward and recurrent neural networks. *Neural Comput & Applic* **16,**491–502 (2007)

Ni, Y., Fan, F.: A two-stage dynamic sales forecasting model for the fashion retail. Expert Syst. Appl. **38**(3), 1529–1536 (2011)

Kaggle. *BigMart Sales Data*. <https://www.kaggle.com/brijbhushannanda1979/bigmart-sales-data>

Stanford University Libraries. SAS Missing Data. (2015). <https://www.kaggle.com/brijbhushannanda1979/bigmart-sales-data>