**EED-363: Applied Machine Learning Project**

Vinayak Mehrotra- 1710110391- vm897@snu.edu.in

Sachin Mavi- 1710110285- sm790@snu.edu.in

Big Mart Sales Prediction

**Shiv Nadar University**

# INTRODUCTION

This report describes our implementation of the Big Mart Sales Prediction System. Big Mart is a big supermarket chain and this project predicts sales for their products across different stores. Big Mart has collected data of 1559 products in 10 different stores for the year 2013 in order to identify which product and store plays a pivotal role in their profit generation.

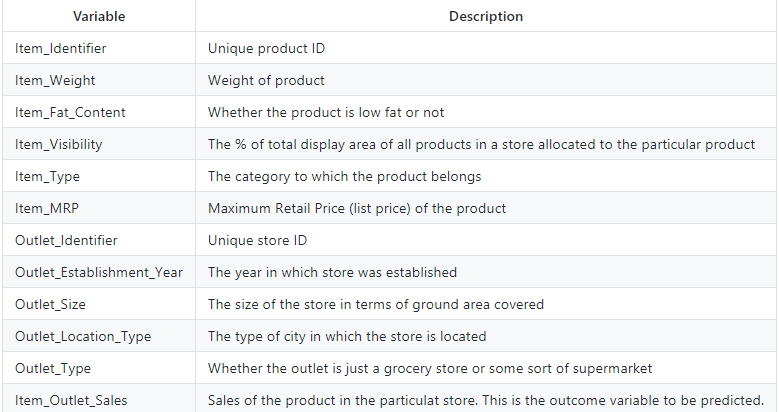
# Description of the Problem

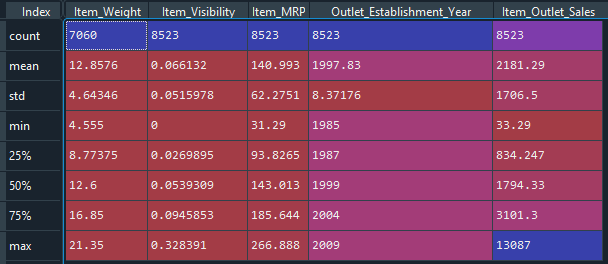
Before selecting the algorithms and the performance metrics, it is important to understand the question. We are dealing here with a regression problem, because the target which is Item Outlet Sales is a continuous variable and not a discrete problem.

1. Supervised Learning: The data is provided with data labels along with the target variable at hand.
2. Plain Batch Learning: This is not a time-series data and new data can be incorporated easily without changing the data much.
3. Performance Measure: Being a regression problem where we fit a line, Root Mean Square Error (RMSE) is an appropriate measure for our problem.

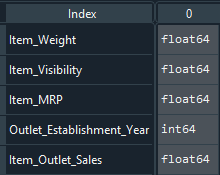
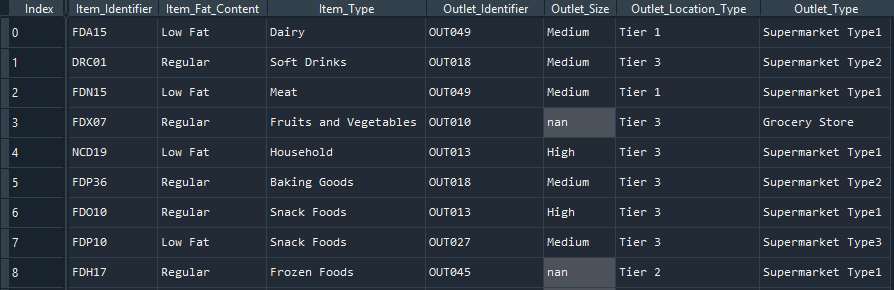
# Description of the Dataset

1. The dataset is already divided into training and test dataset, with the training data containing 8523 examples and 11 features and 1 target variable.

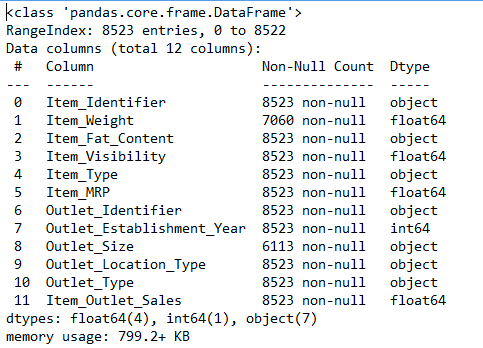




1. The dataset contains 5 numeric features (including one target) and 7 categorical features.

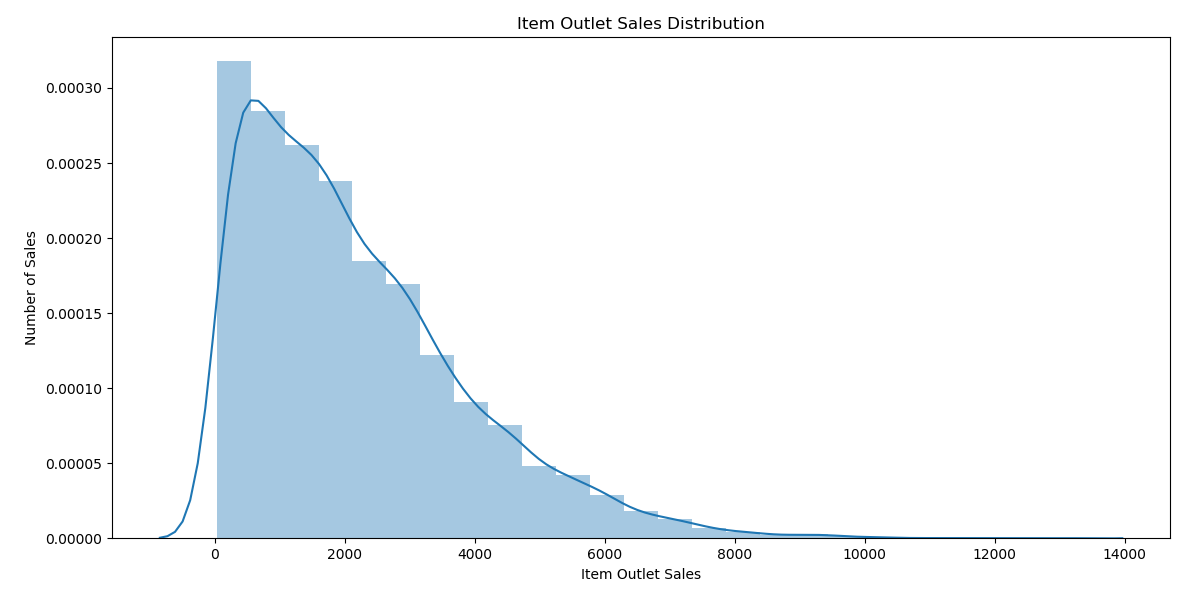
 

1. The dataset is not clean, in the sense that there are some missing values, some redundancies in the dataset, which we will fix and then fit models in it.

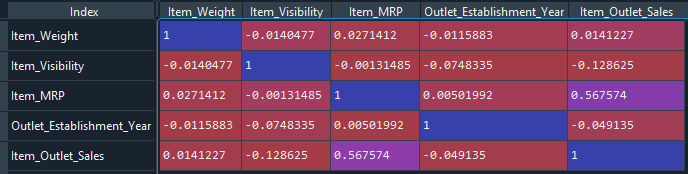


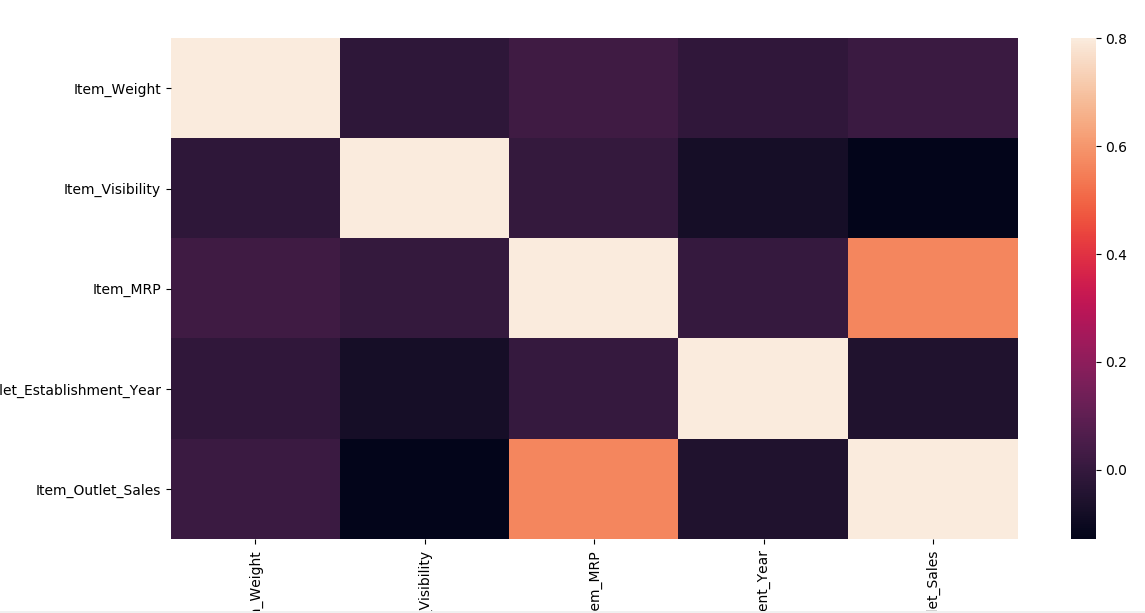
# Complete Analysis of the Dataset

1. Histogram displaying the distribution of the target variable.

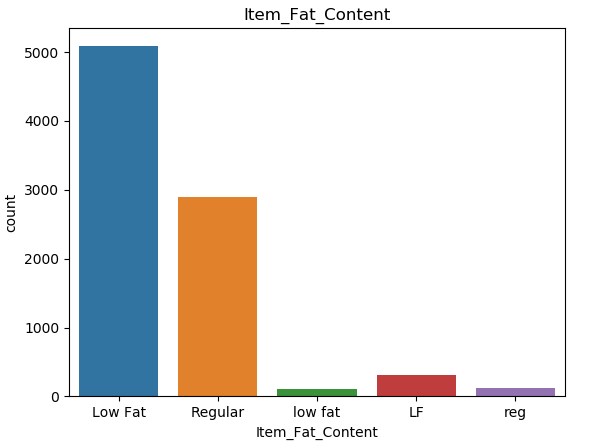
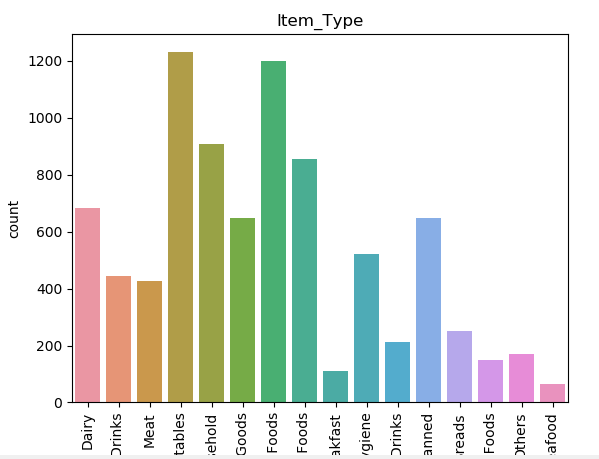


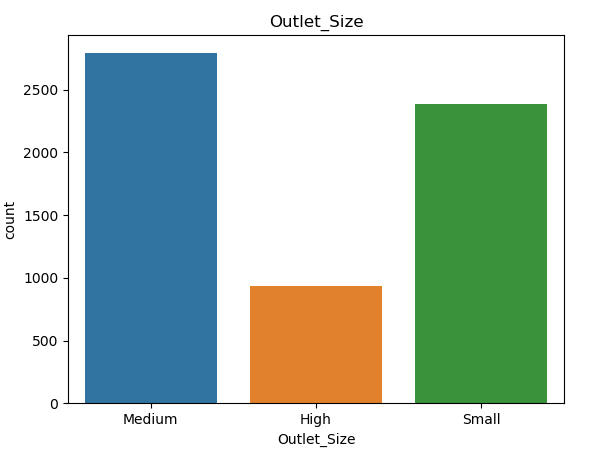
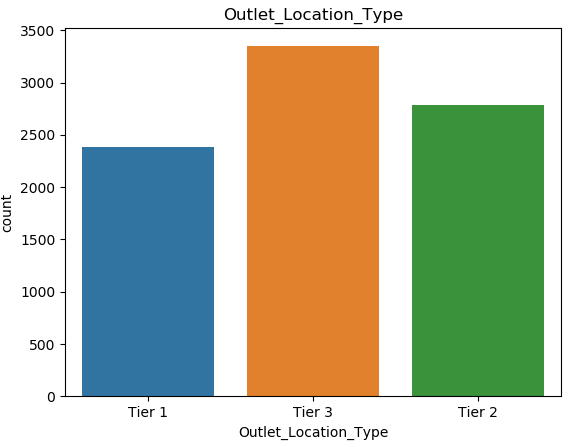
1. Correlation of the numeric features with the target and a heatmap

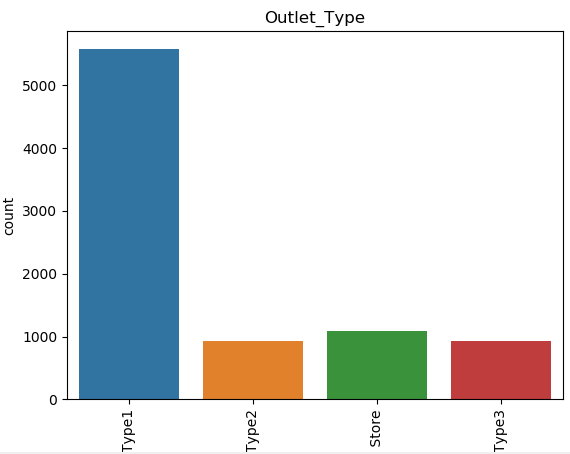




1. Univariate Analysis of categorical features



1. Summary of Univariate analysis

· Upon univariate analysis, we observed that the feature Item\_Fat\_Content had redundancies in its categories, i.e., Low Fat was written in three different ways, which had to be corrected

· We observed that the feature Item\_Identifier indicates three different kinds of items broadly, so it had to be made that way.

· We observed that in the dataset, Item\_Visibility feature has some values as 0, which implies that the store has no such product, so we treated it as a missing value and used the ‘mean’ strategy for the imputer.

· Features Item\_Weight and Outlet\_Size contain some values as NaN.

· There are 1559 unique items in a single store, which had to be kept in mind.

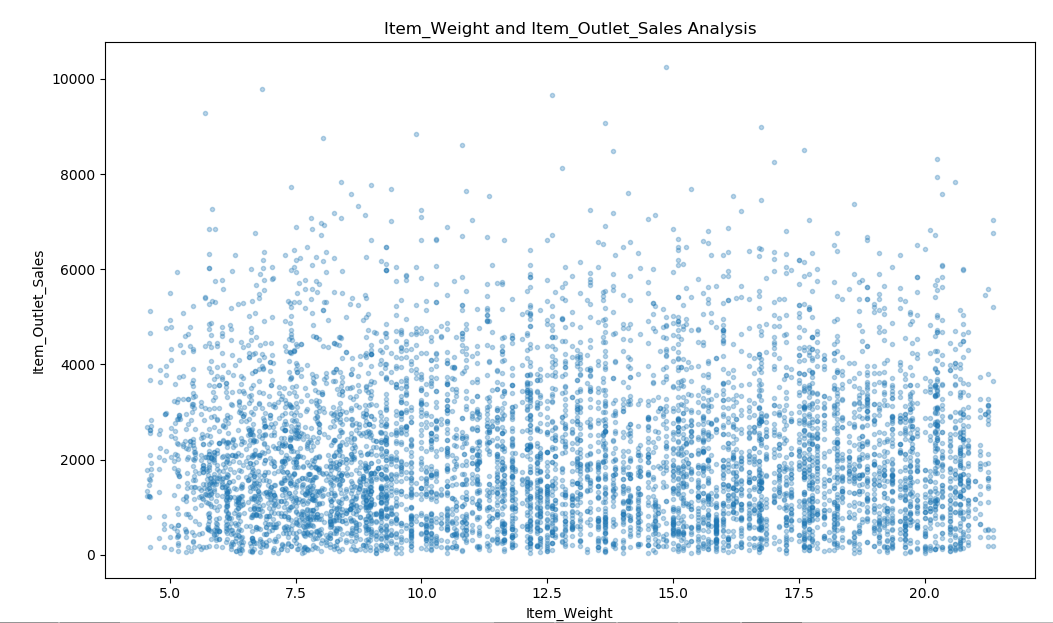
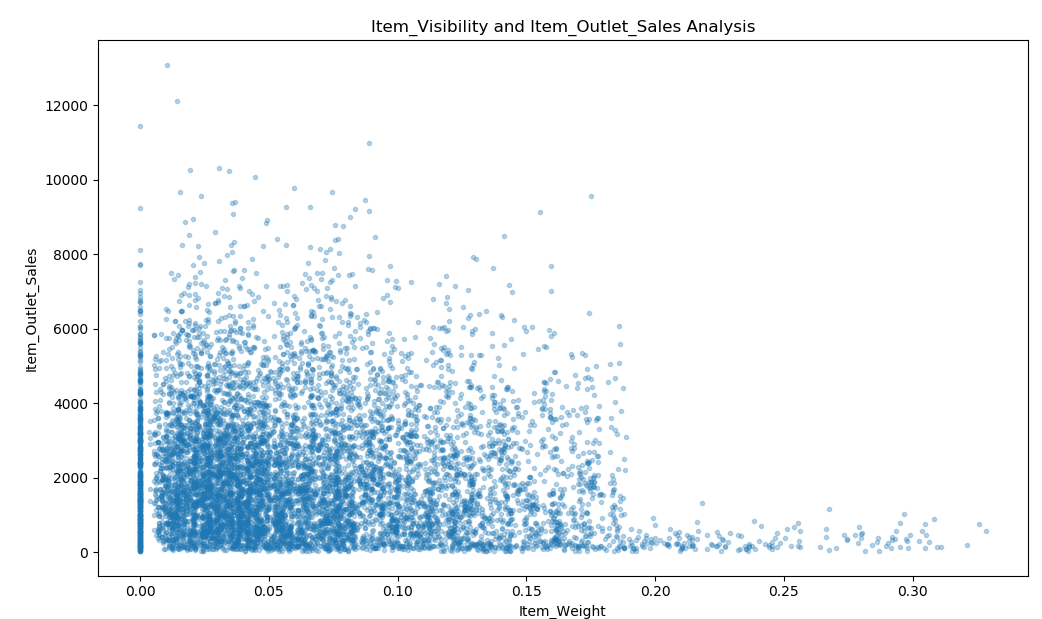
· The Item\_Type feature has different scattered categories, so we can create a new variable which makes this classification more streamlined, so that the encoding is reduced.

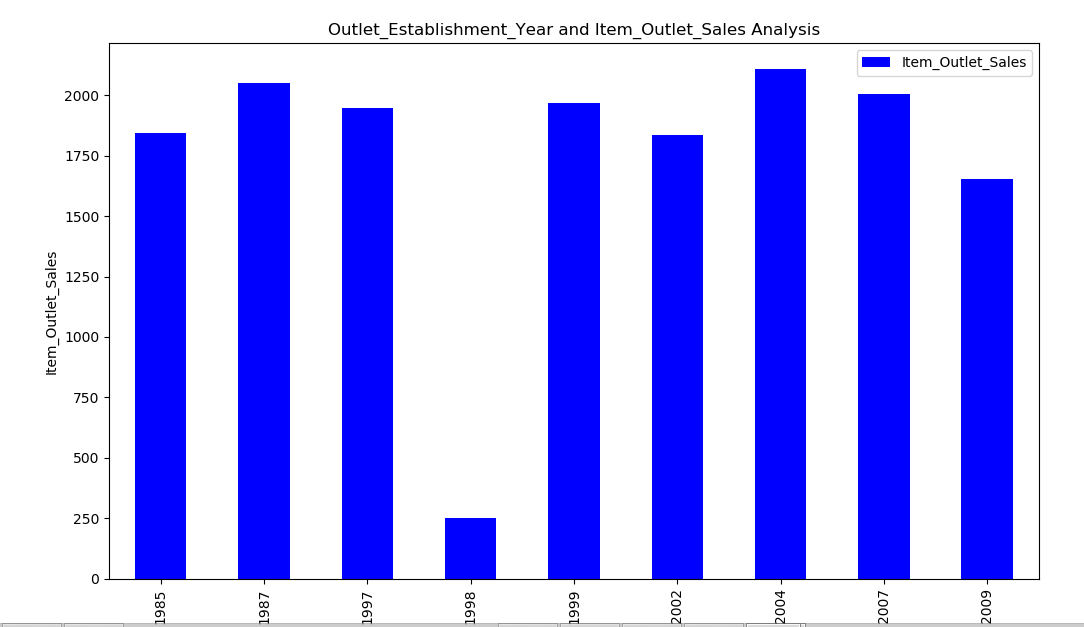
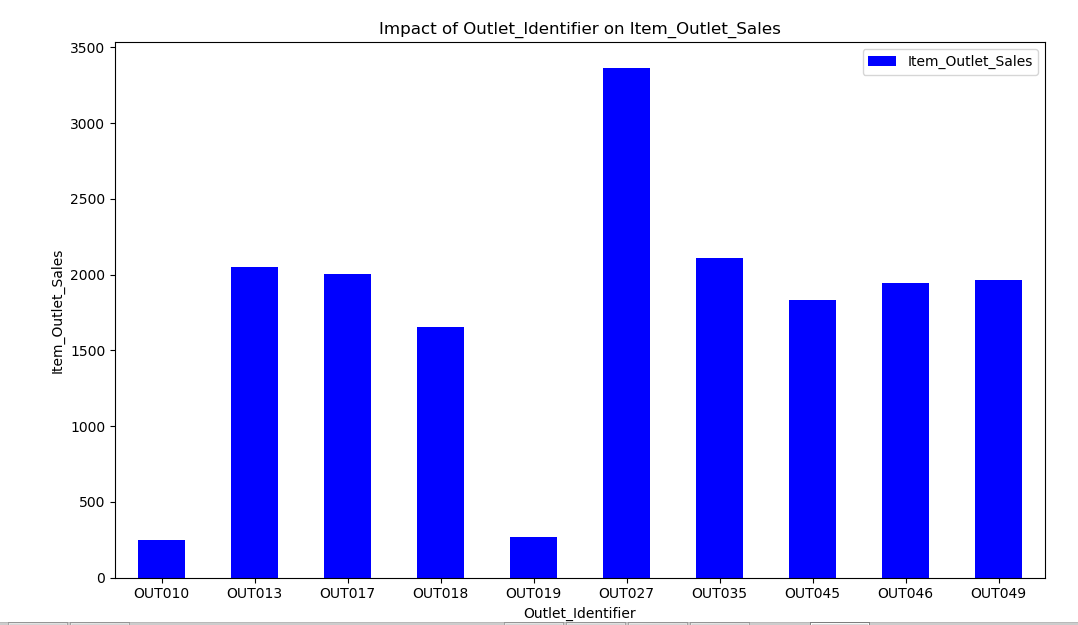
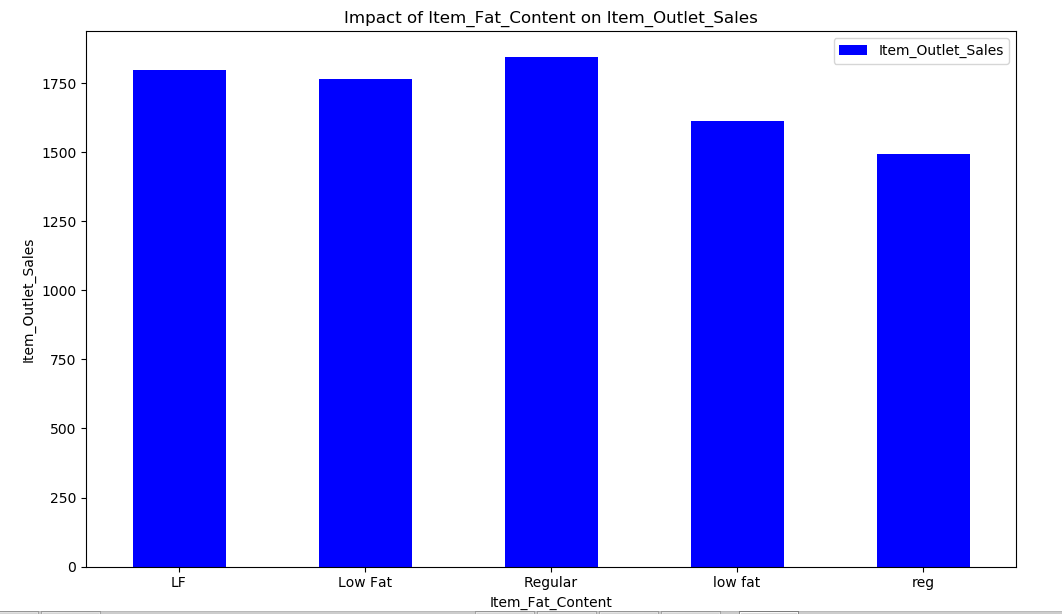
· Feature Outlet\_Establishment\_Year has years which are not understood by the model as they should be. So, we have to convert them in a way that is more intuitive for the models.

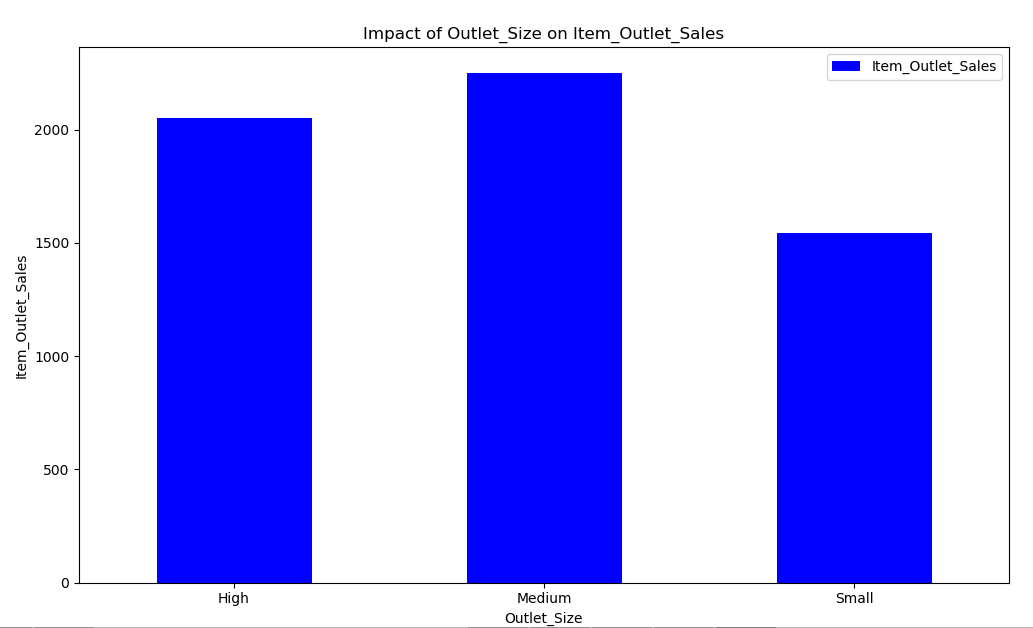
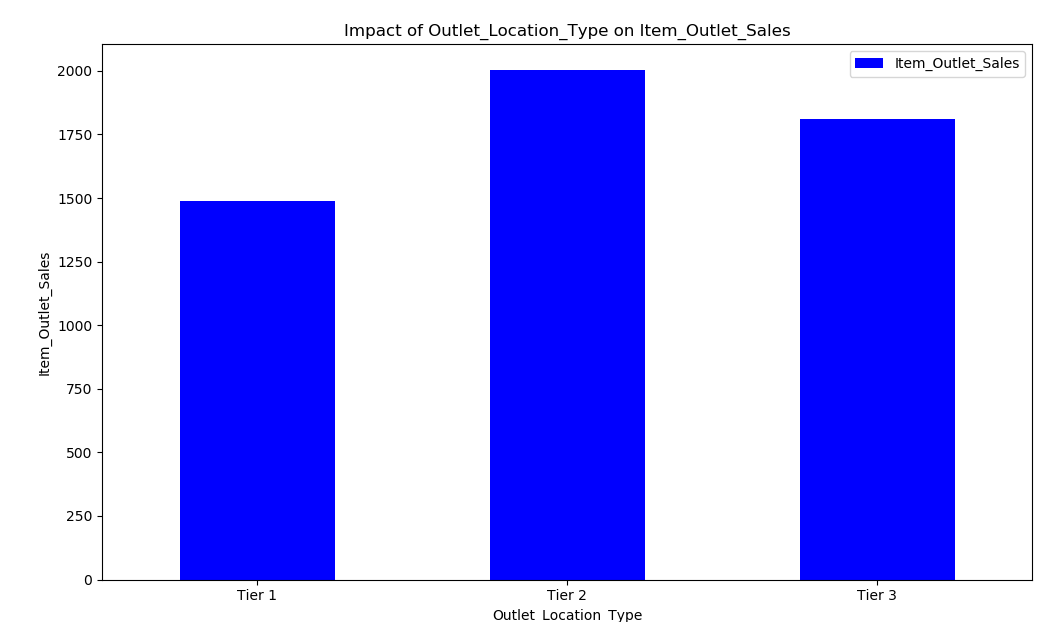
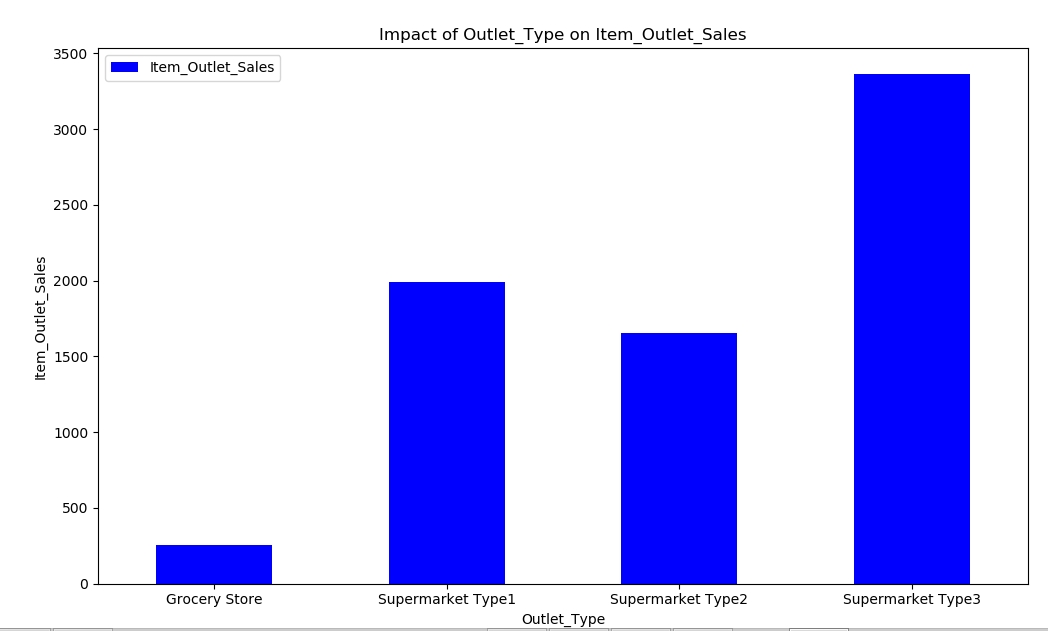
· Using Dummy Encoding for the categorical variables, we have to keep in mind that we take n-1 columns out of the n dummy encoded columns for a categorical feature, to prevent dummy variable trap which brings in Multicollinearity.

· We also observed that certain products were Non-Consumable and we had to replace their values in the Item\_Fat\_Content feature with Non-Edible.

1. Bivariate Analysis of categorical features



1. Summary of Bivariate Analysis

# Item\_Weight and Item\_Outlet\_Sales Analysis: we found out that the correlation was low with the help of a heatmap

# Item\_Visibility and Item\_Outlet\_Sales Analysis: Initial guess based on intuition is that the products that are kept in front will make the sales go high and increase the profit

# Many Products have Visibility = 0

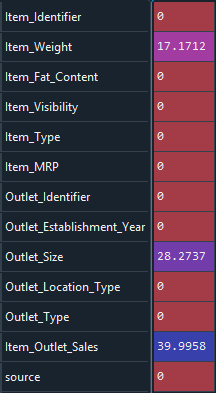
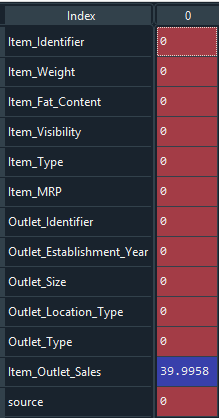
# The data shows a trend that eliminates our hypothesis, which can be due to the fact that important products that control the profit do not need substantial visibility, they are just in demand

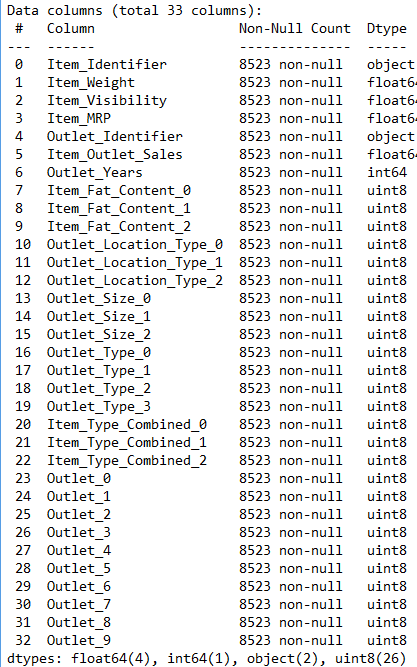
# Year is not related to the target, year 1998 has low sales which may be due to the fact that few stores may have opened in that year

# low fat products seem to have higher sales than regular products

# It is visible that Grocery Stores have less sales, maybe because why will someone go for grocery store and then to a different store when there are big stores having everything available under one roof. Supermarket type 3 has higher sales than Supermarket type 1.

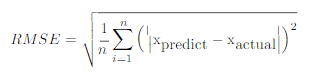
1. Checking the percentage of null values in features

 ------------>>>>>>>> 

1. After imputing the missing values and encoding the categorical features: 

# Model Building

The Performance Measure for Regression problems is RMSE:



This has to be minimized.

1. Multiple Linear Regression: A model which create a linear relationship between the dependent variable and one or more independent variable, mathematically linear regression is deﬁned as:



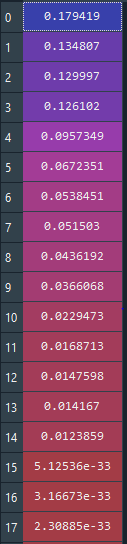
Where y is the target variable and x are the independent variables

1. Ridge Regression: This method fits the training data with a bigger bias, so that the variance is less because of the tradeoff between bias and variance. This ensures that the fitted line has coefficients which are less sensitive to the data points. The cost function for this method is defined as:



Where lambda is the penalty term, higher its value bigger will be the penalty. Controlled in the code by the parameter alpha.

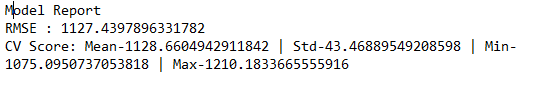
1. PCA: We have used Principal Component Analysis for Feature Extraction to see if we get better performance metrics. But before this, feature scaling had to be performed. We obtained a variance ratio for all principal components for which a Scree Plot can be made:



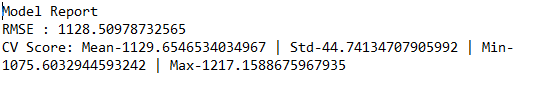
1. Cross-Validation: We performed a 20 fold Cross Validation for gauging the performance of our models

## Results

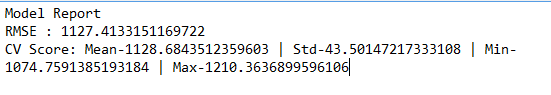
1. Multiple Linear Regression



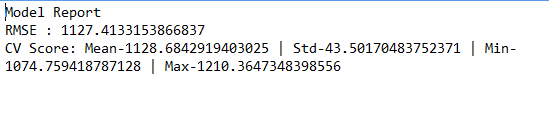
1. Ridge Regression



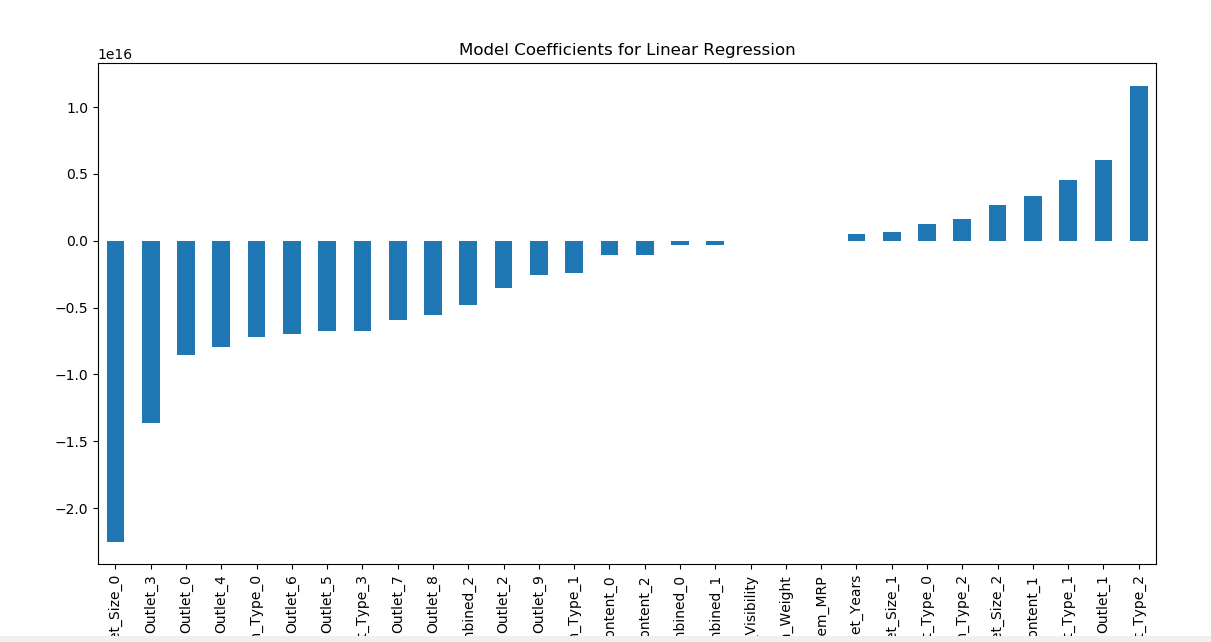
1. Linear Regression after PCA



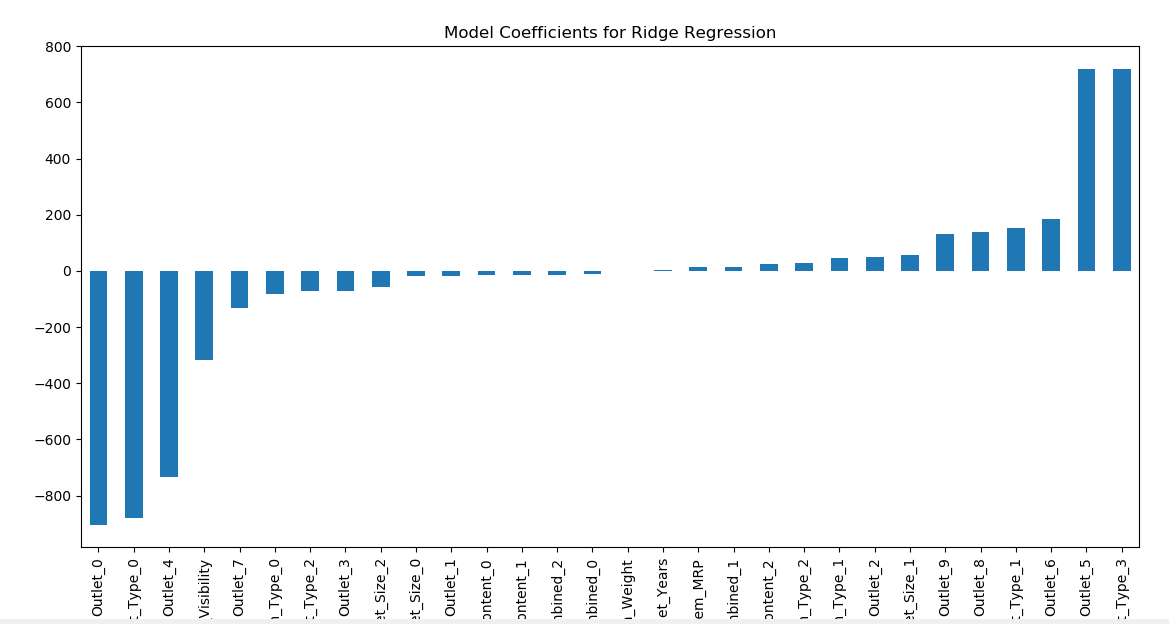
1. Ridge Regression after PCA



1. Model Coefficients for Linear Regression



1. Model Coefficients for Ridge Regression



## 

## 

## Conclusions

1. PCA with lesser components resulted in a worse performance because of the fact that some important components were discarded.
2. Feature Extraction resulted in a tiny bit better result, but not a very good result.
3. Linear Regression after PCA gave the best result with the least RMSE.

## Scope Of Improvement

1. Taking into consideration the value of Adjusted R squared would improve feature selection as it changes when new features are added or removed in backward selection.
2. Hyperparameter Tuning of the parameters of our model using Cross Validation or other methods would result in better accuracy.
3. Using models like Decision Trees and XGBoost and other methods of Ensembling would result in way better accuracy.

We worked on this project whole-heartedly and did what we could given the circumstances. These improvements are not just written for the sake of writing, we will learn these methods and continue with the project. It as a very good learning experience.

## References

1. <https://www.researchgate.net/publication/336530068_A_Comparative_Study_of_Big_Mart_Sales_Prediction>
2. <https://medium.com/diogo-menezes-borges/project-1-bigmart-sale-prediction-fdc04f07dc1e>