# Big Data Mart Sales Problem.





Bigmart is a big supermarket chain, with stores all around the country. The management of the shop had set out a challenge to all Data Scientist to help them create a model that can predict the sales per product for each store. The shop has collected sales data of products across 10 stores in different cities over a given period of time.

### Problem Statement: -

##### *The data scientists at BigMart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined. The aim is to build a predictive model and find out the sales of each product at a particular store.*

**Using this model, BigMart will try to understand the properties of products and stores which play a key role in increasing the sales of their products.**

**The dataset includes two files:**

**- bigdatamart\_Train.csv: Use this file for the model building purpose.**

**- bigdatamart\_Test.csv: Use this file for getting predictions from the trained model.**

# ****COLUMNS DISCRIPTION:****

**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**

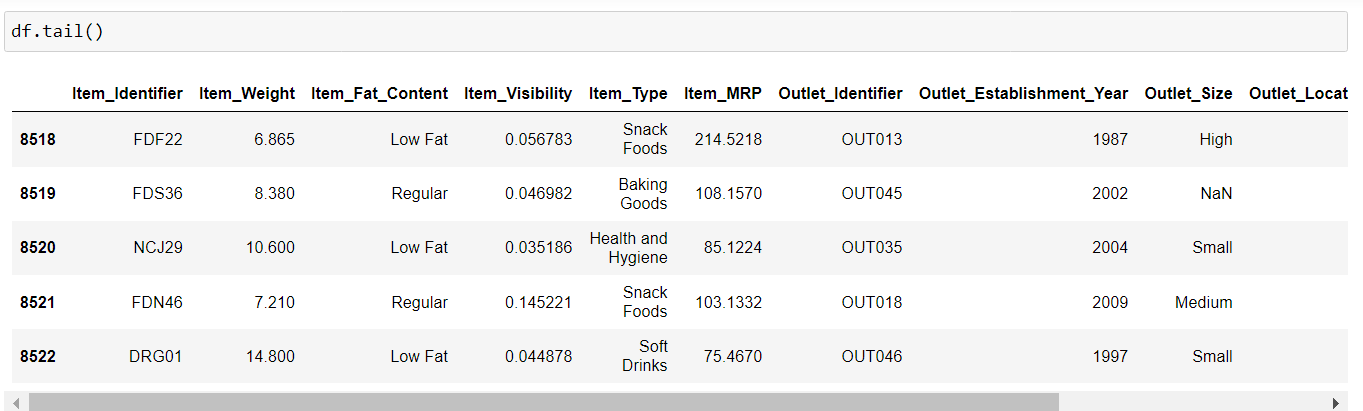
**Outlet\_Type : Whether the outlet is just a grocery store or some sort of supermarket**

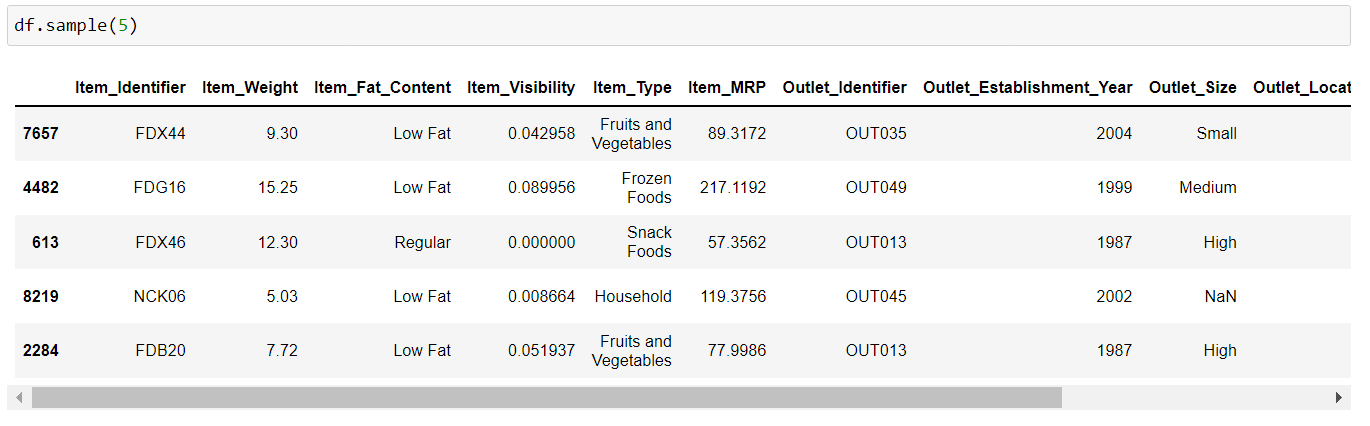
**Item\_Outlet\_Sales : Sales of the product in the particular store.**

# Let’s start exploring the dataset by loading some important Libraries and Train Dataset :-

# 

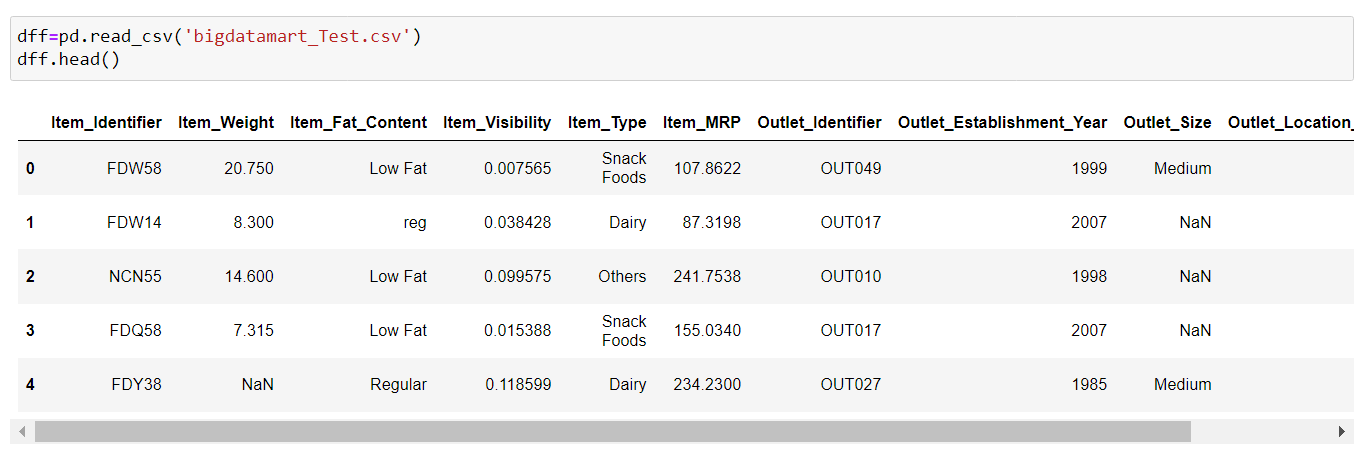
# Here I have imported my libraries and Train dataset





Now I have shuffled the dataset to see whether nan values or abnormal values are present.

Now As we know that we have test dataset also available so let’s load it:-



**We can Clearly see that the data in test dataset we don’t have a standard data, it has null data as well and it don’t contain output or target data**

# Checking Null Value In The Data Frame :-

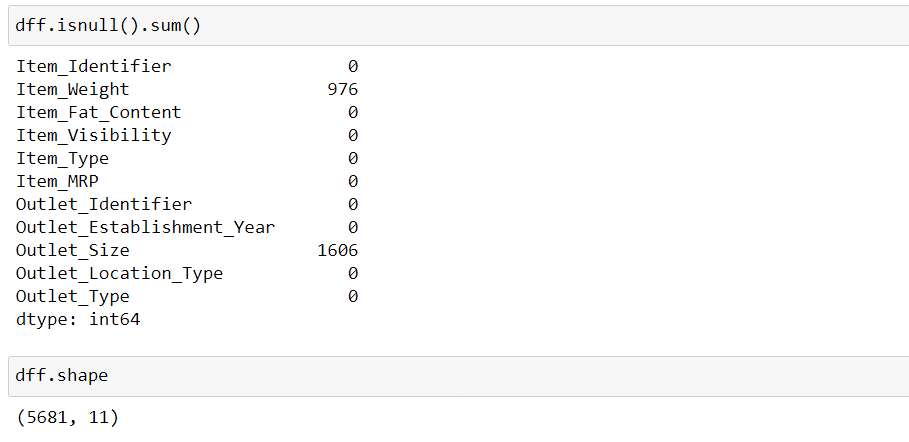
# df -> for Training data

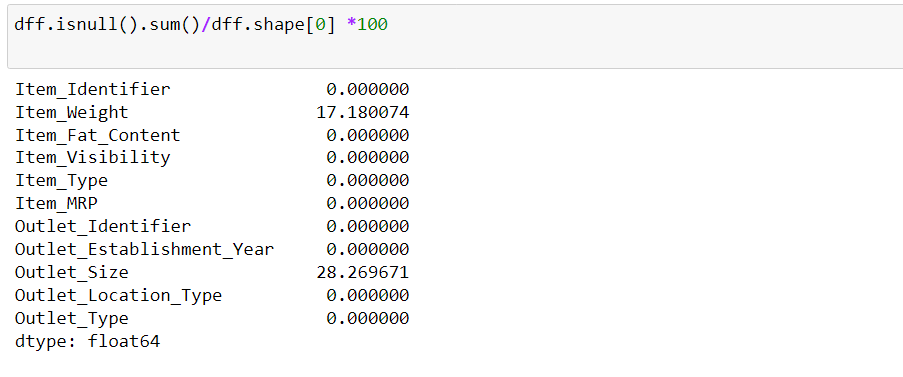
# dff -> for Testing Data

# 

# 

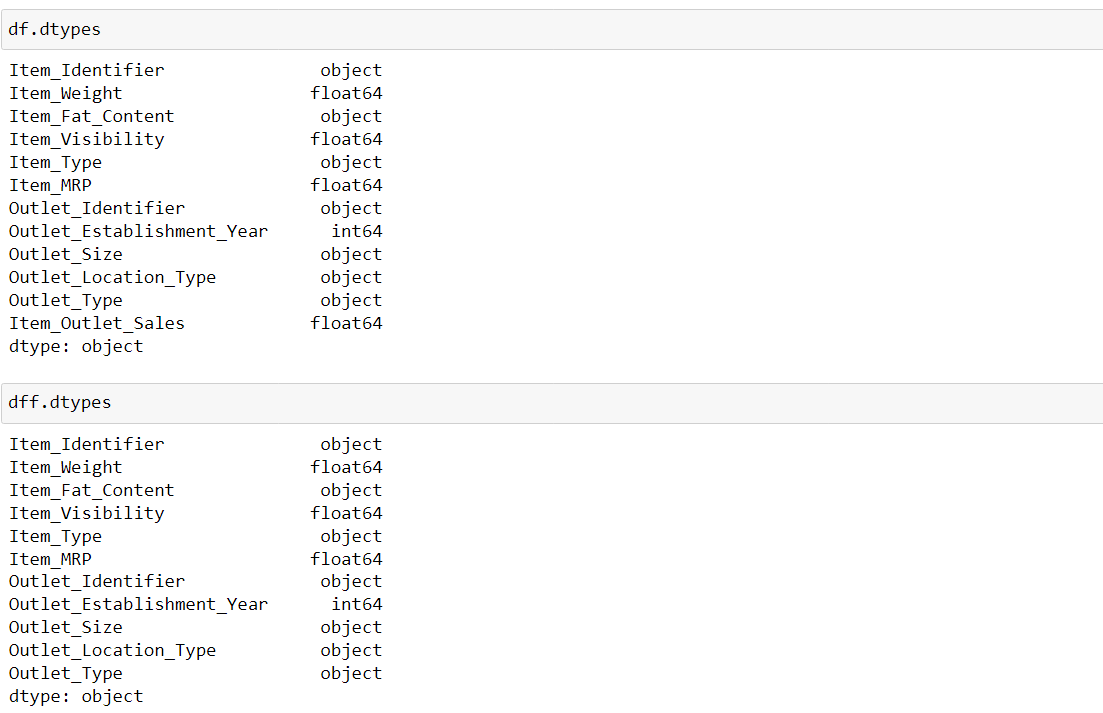
**df(training) :- Item\_Weight=17% ,Outlet\_Size=28% have Null Values.**



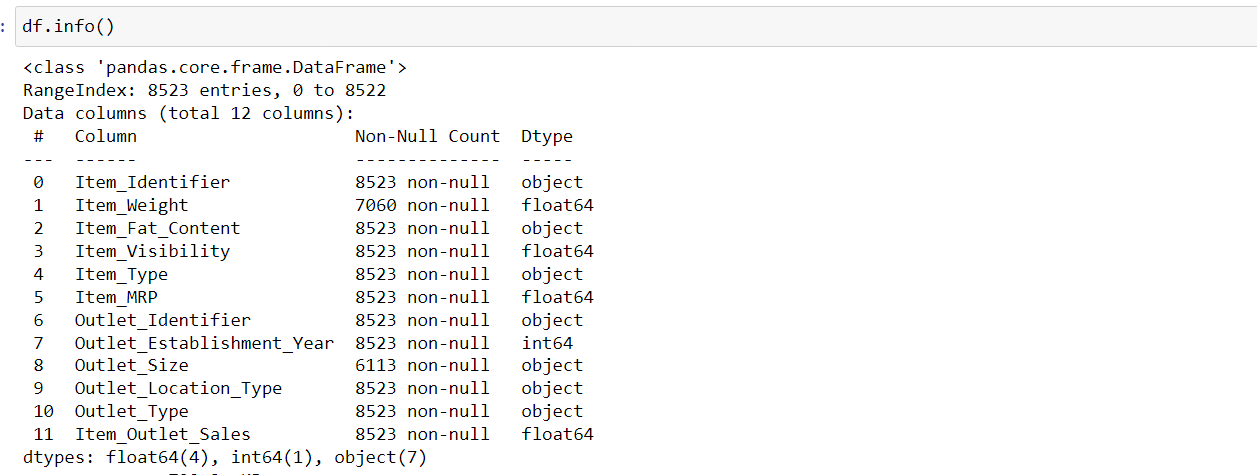


**dff(testing) :- Item\_Weight=17% ,Outlet\_Size=28% have Null Values.**

Checking the data types of both training and testing columns in the dataset :-



* **We have 7 object columns in both training(df) and testing(dff) namely: - Item\_Identifier , Item\_Fat\_Content , Item\_Type , Outlet\_Identifier , Outlet\_Size , Outlet\_Location\_Type , Outlet\_Type.**
* **We have 4 float64 and int64 columns in both training(df) and testing(dff) namely :- Item\_Weight , Item\_Visibility , Item\_MRP , Outlet\_Establishment\_Year ..**

****

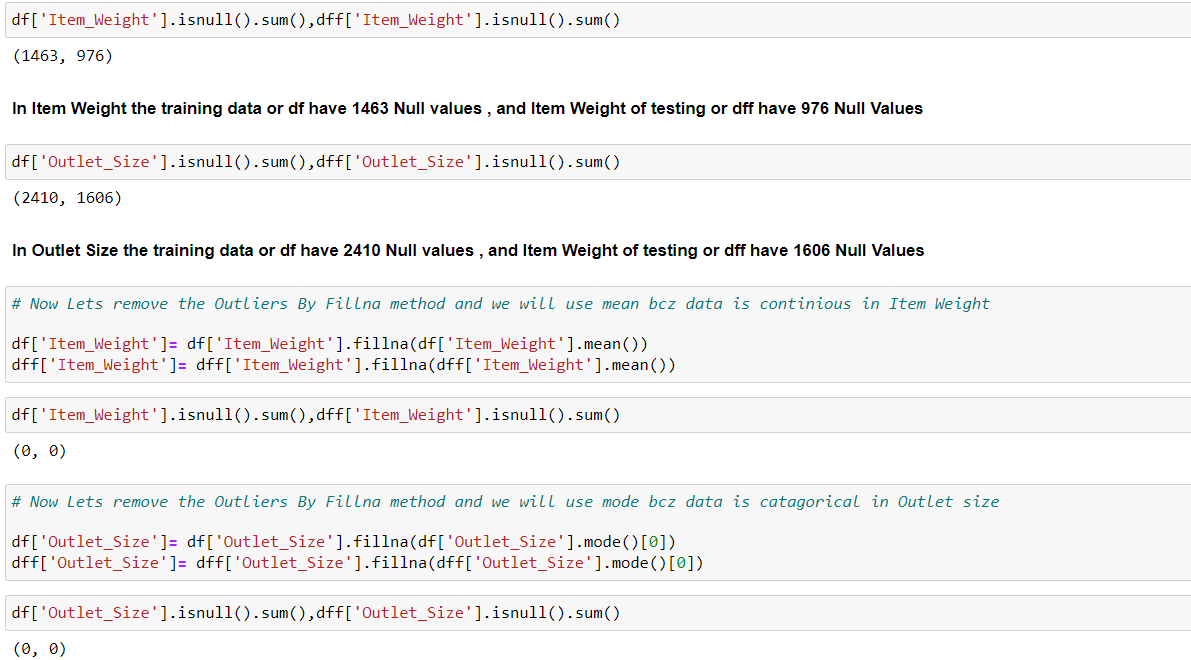
We here get the Information about the dataset and here we get to know about Columns, Dtypes, Non-Null counts of the columns and Memory usage of the dataset.

* In training and testing dataset column Item\_Fat\_Content has a unique values which seems to be repeating so we will group them :-



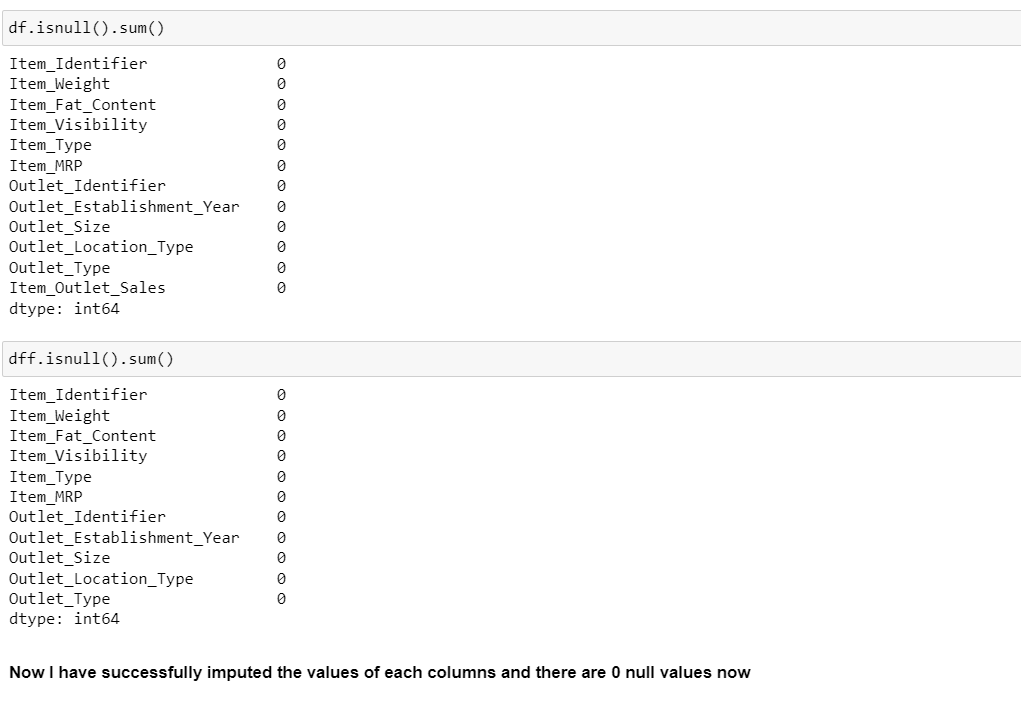
So here we can see that we have grouped them in Low fat and Regular category .

# Now Let’s fill the NaN Values.



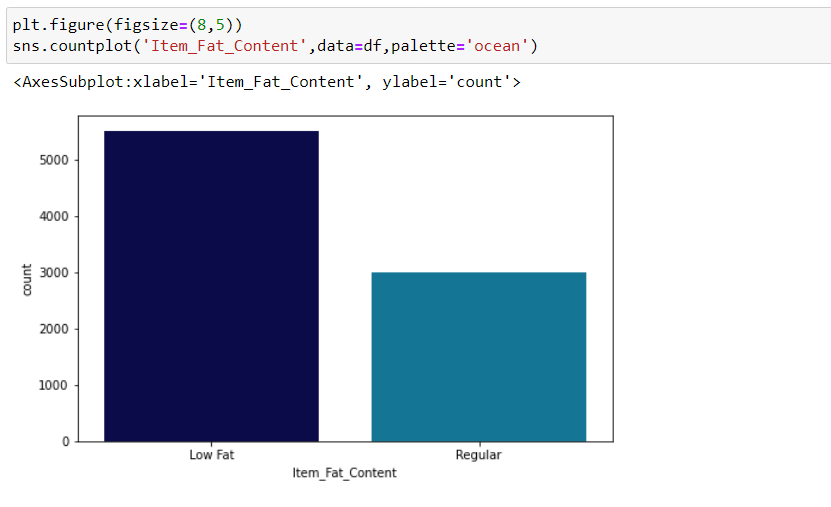
So here I have filled all the Nan values from training and testing dataset accordingly

Now Let’s check again whether any nan value is left or not?

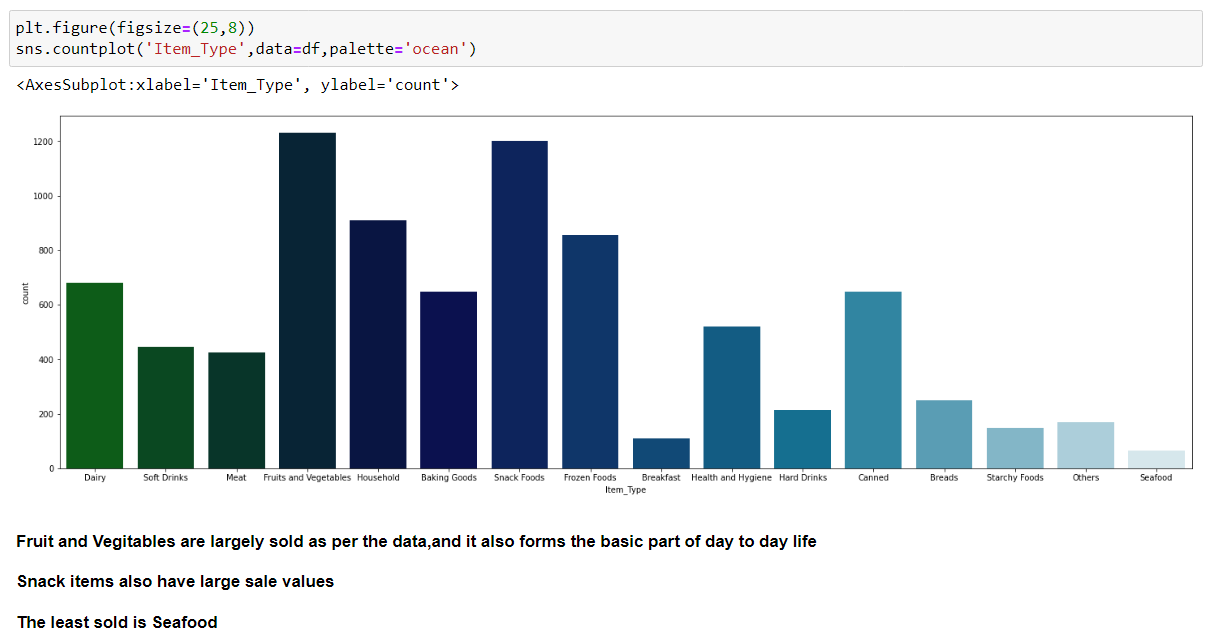


Here we can see that Zero Nan Values is present in the dataset of both training and testing .

## A. Univariate Analysis



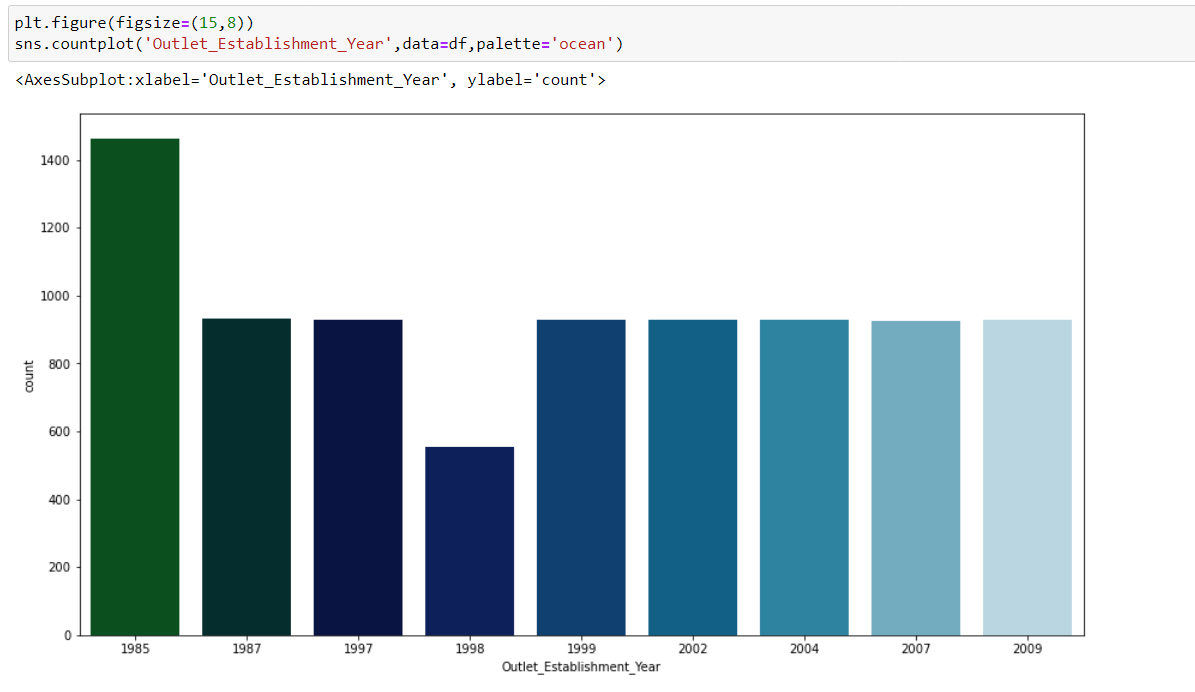
We can see that Low fat content are consumed more by the people . As we know that is good for health too.



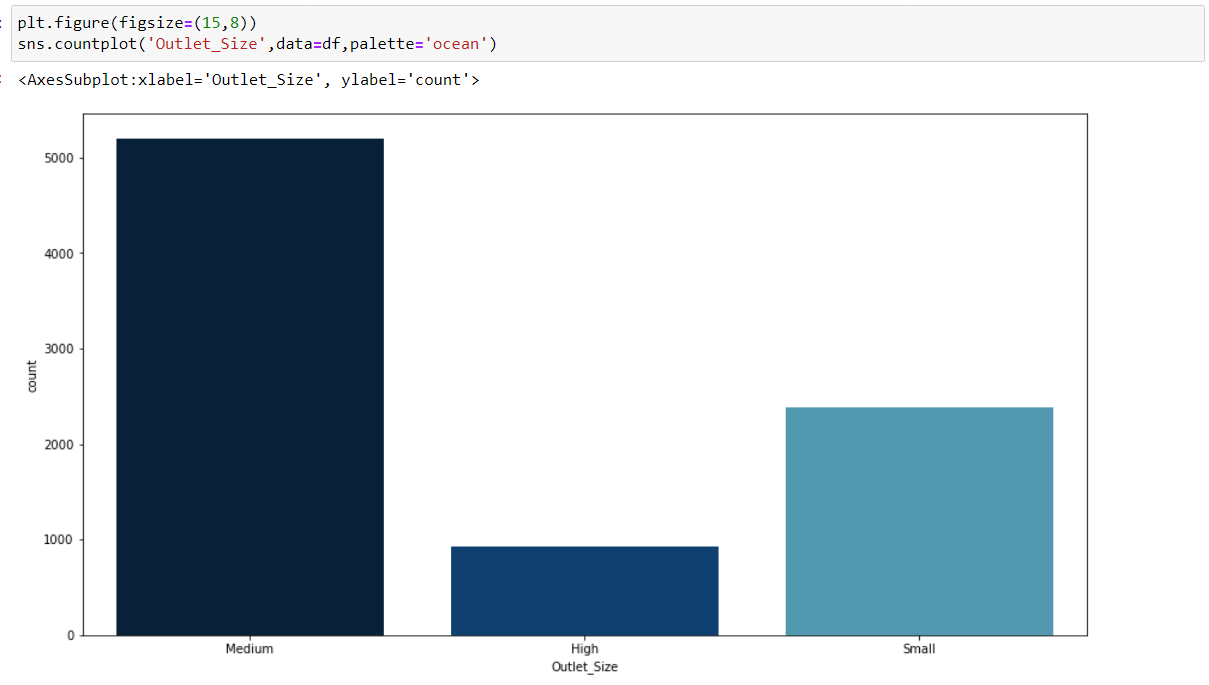
Here we can see that Fruits and Vegetable are sold largely and we know that it is required in day-to-day life.

Snack item also have large sales value.

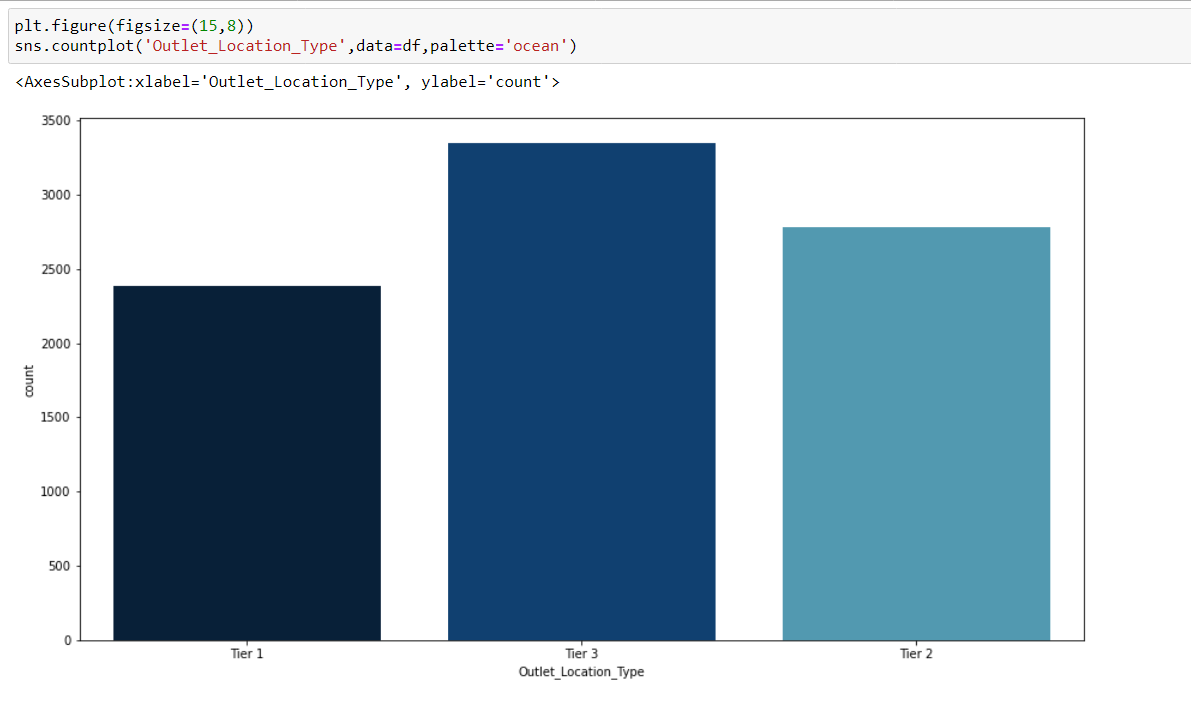
The least sold item is Seafood.



**WE can see maximum companies are established between 1985 and are running.**



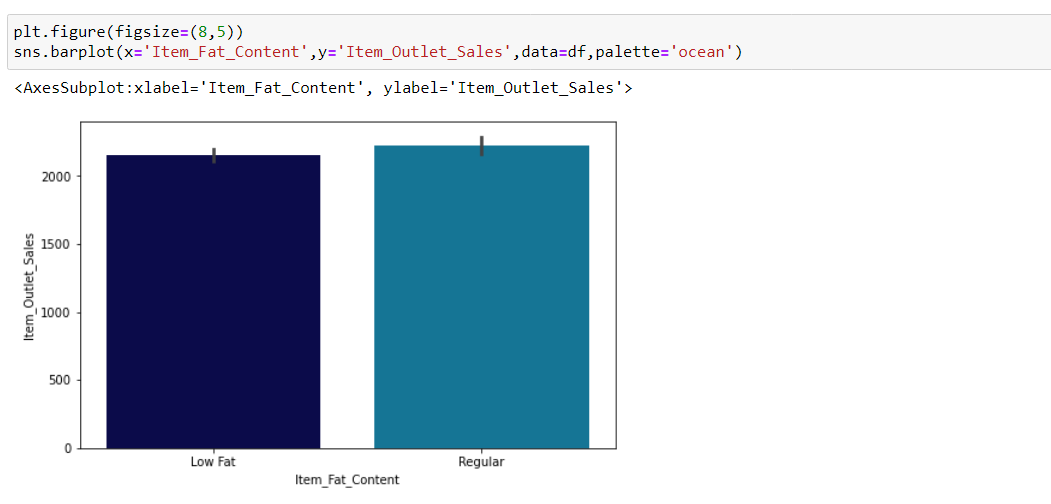
**Medium size Outlet are more in Number as Compared to High and small**



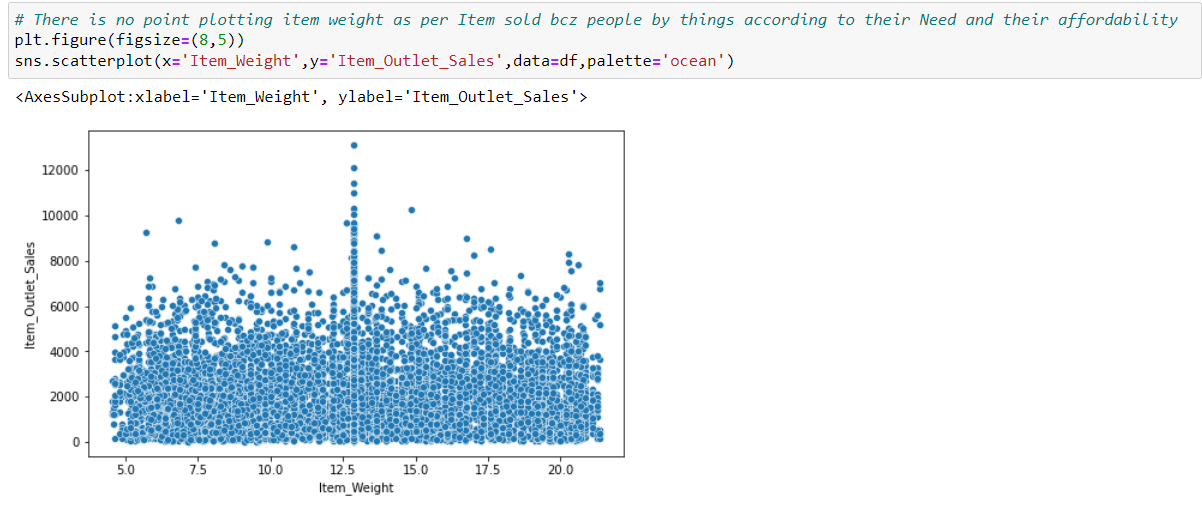
**Outlets are located most in Tier3 cities than followed by Tier2 and at last Tier1.**

**This may be because Tier 3 cities have least rent**

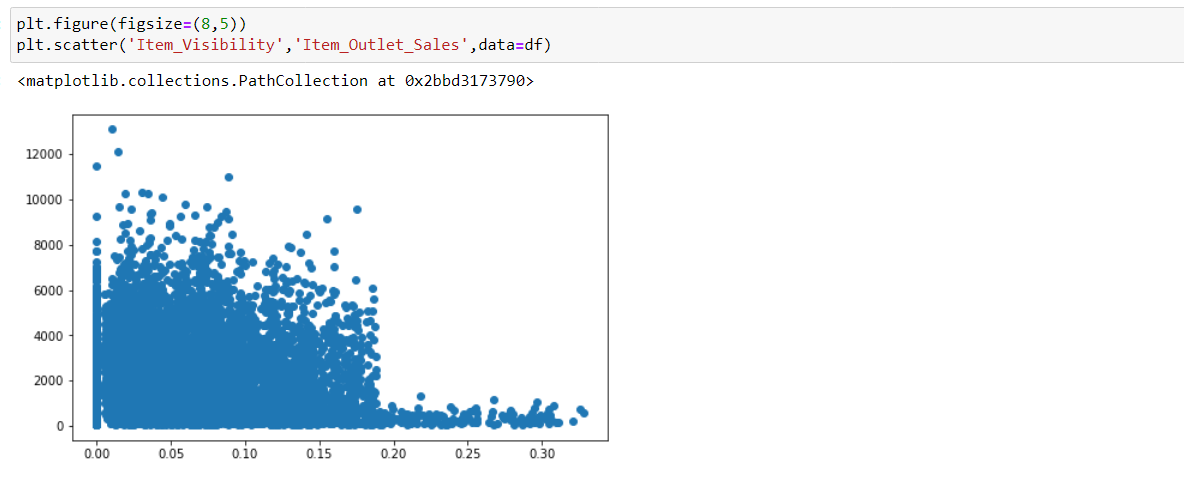
## B. Bivariate Analysis



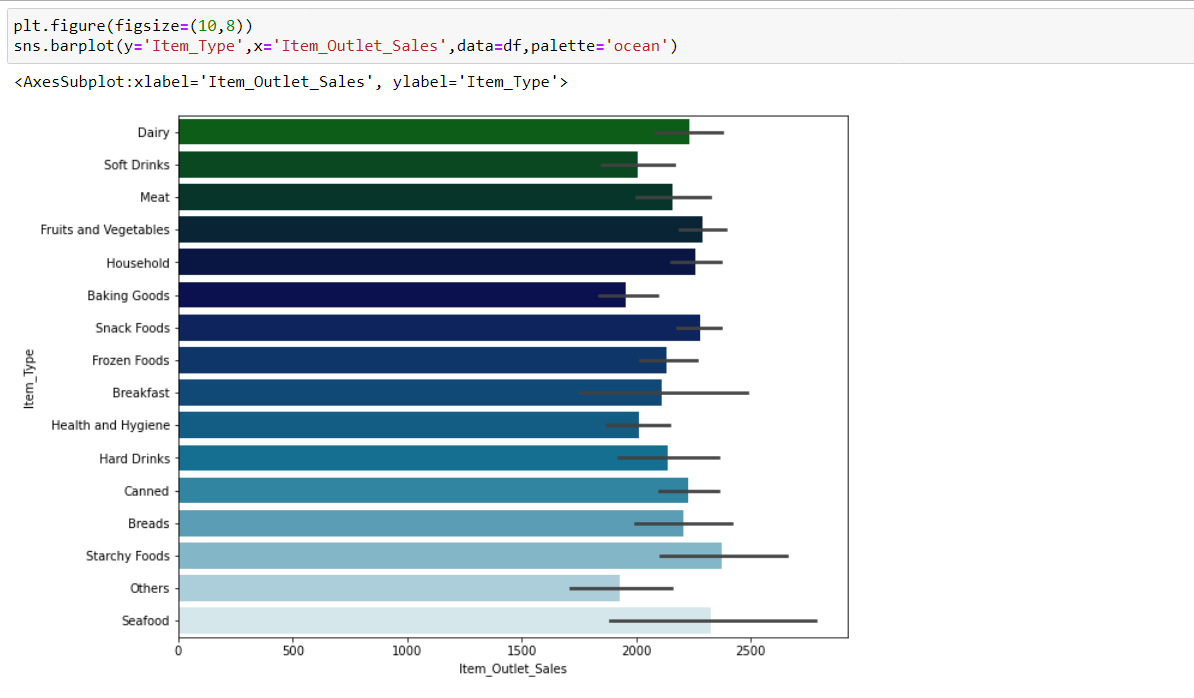
**We can see that Both the Items (Low Fat and Regular) are Equally sold in Item outlet sale.**



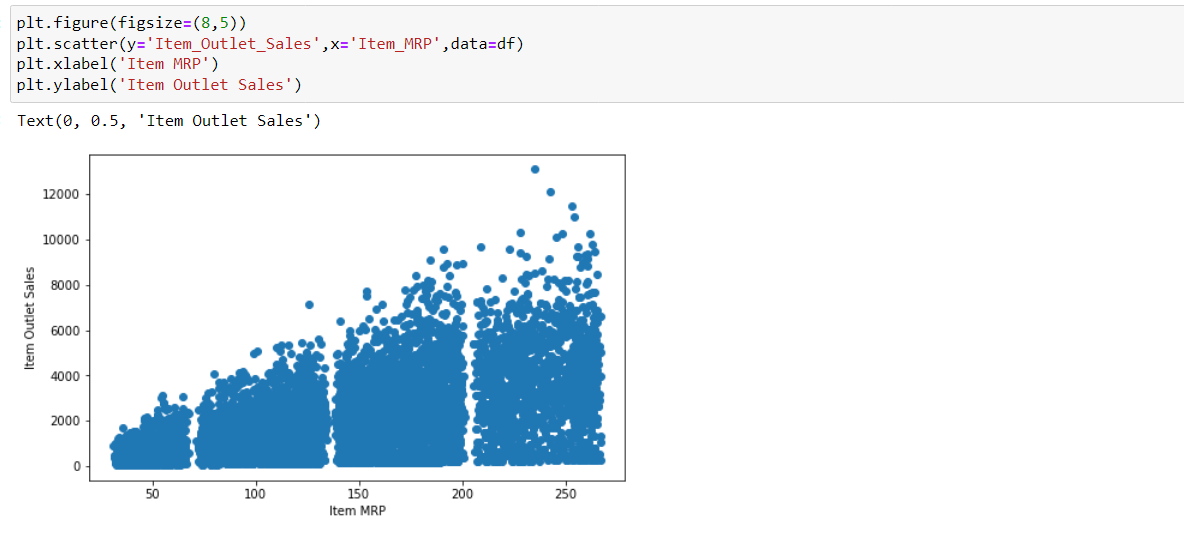
**At Weight 12.5 Maximum items are sold in the Outlet.**



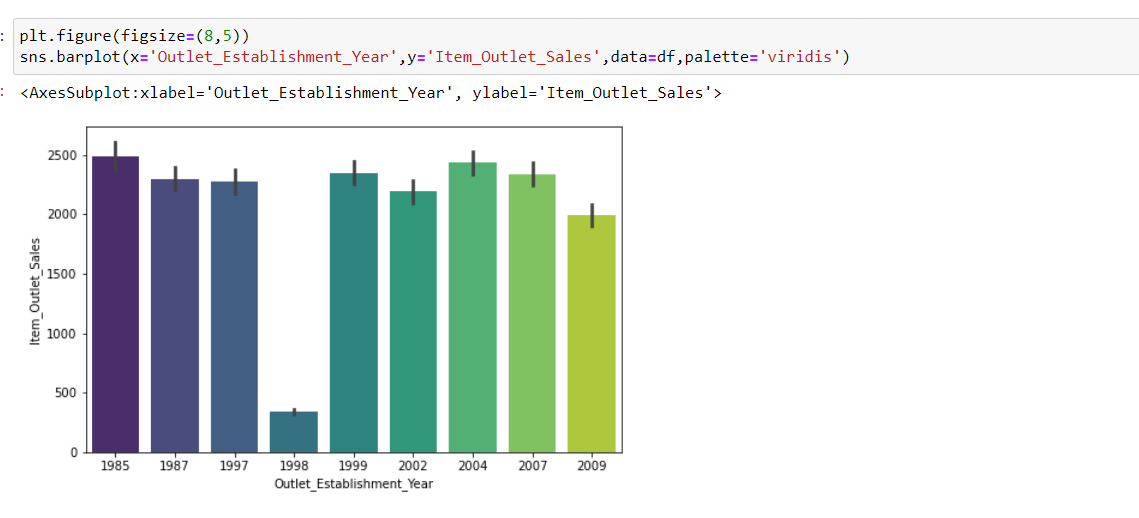
**This plot is Not making any sense as we know that the more the Item is exposed it has good chances of being sold because of its visibility. And here we are watching that at 0 visibility also it has a good sale and at 0.30 visibility the item have least sales ...so this data irrelevant to Outlet sale**



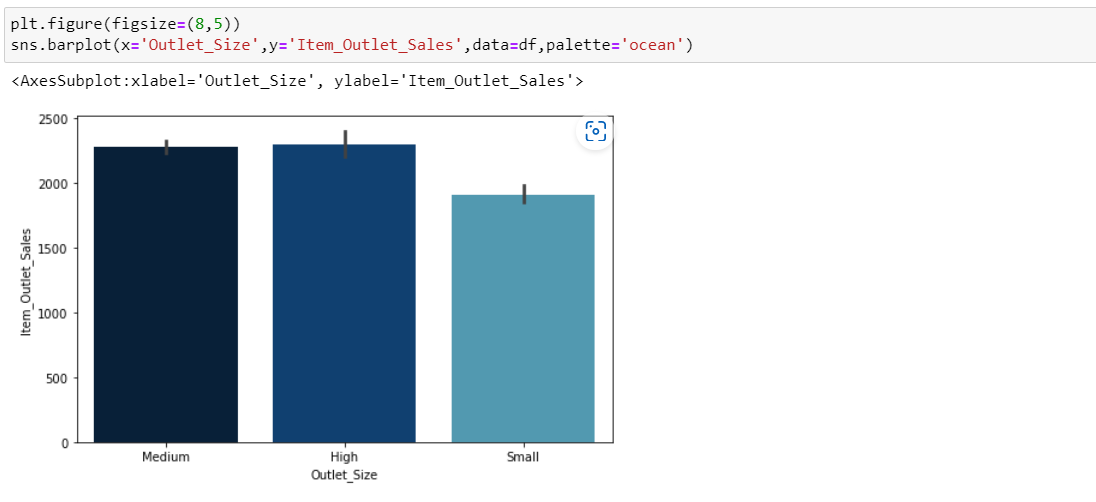
**The products available were Fruits-Veggies and Snack Foods but the sales of Seafood and Starchy Foods seems higher and hence the sales can be improved with having stock of products that are most bought by customers.**



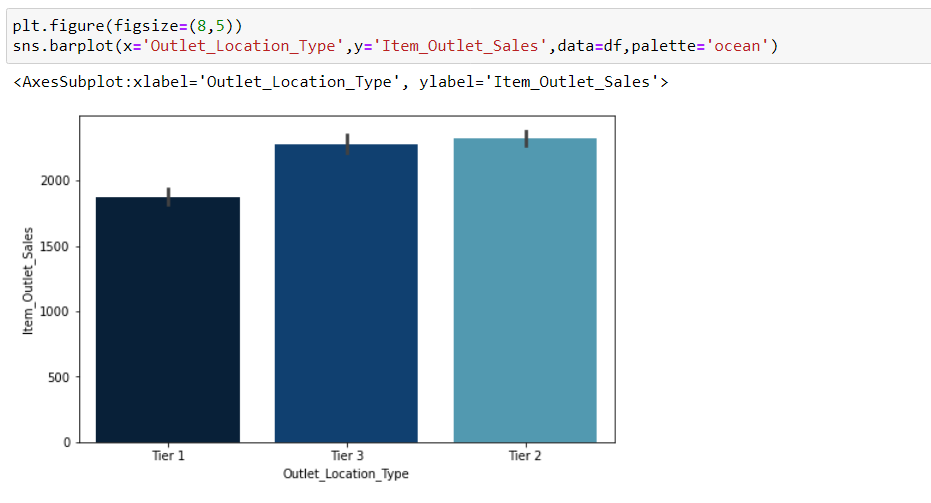
**We can see that the maximum sale for the product is which are having MRP between 200 to 250 .**



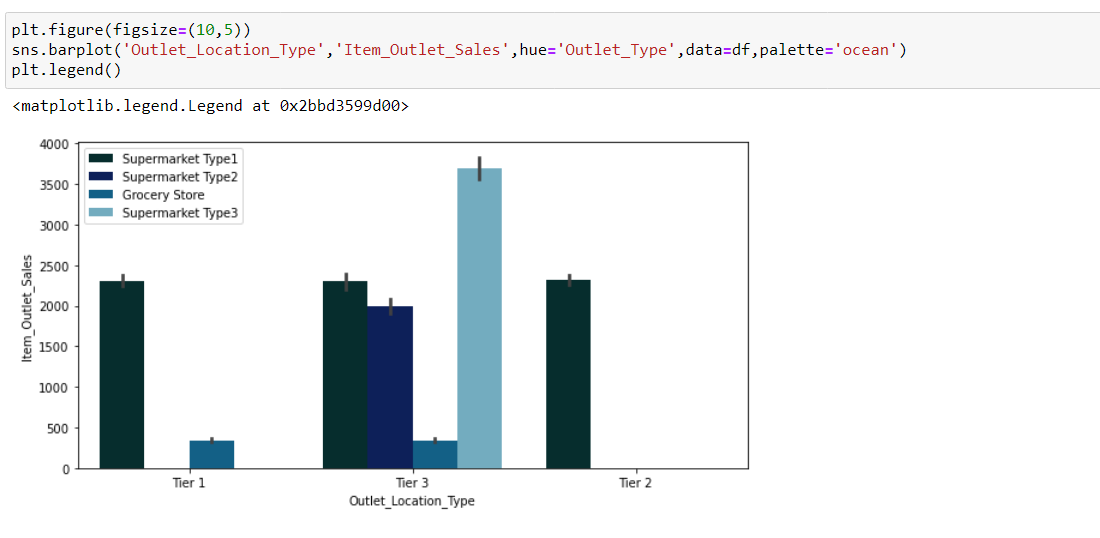
**It is quite evident that Outlets established in 1985 is having good Sales margin because they are in the market and they know how to operate. Established years wouldn't improve the Sales unless the products are sold according to customer's interest because we can see that year 1998 is having least sales as compared to before and after years. Here we not getting out any relevant information from the Establishment year.**



The Maximum sale is given by High and Medium Outlet and the least sale is given by Small Outlets.

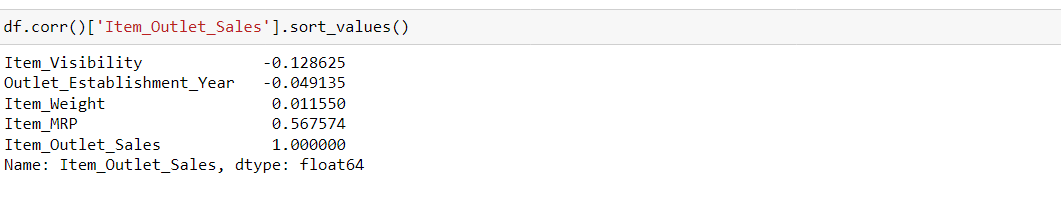


**We can see that the sale in Tier 2 and Tier 3 is almost equal and High but Tier 1 is bit less which can even be improved easily.**

****

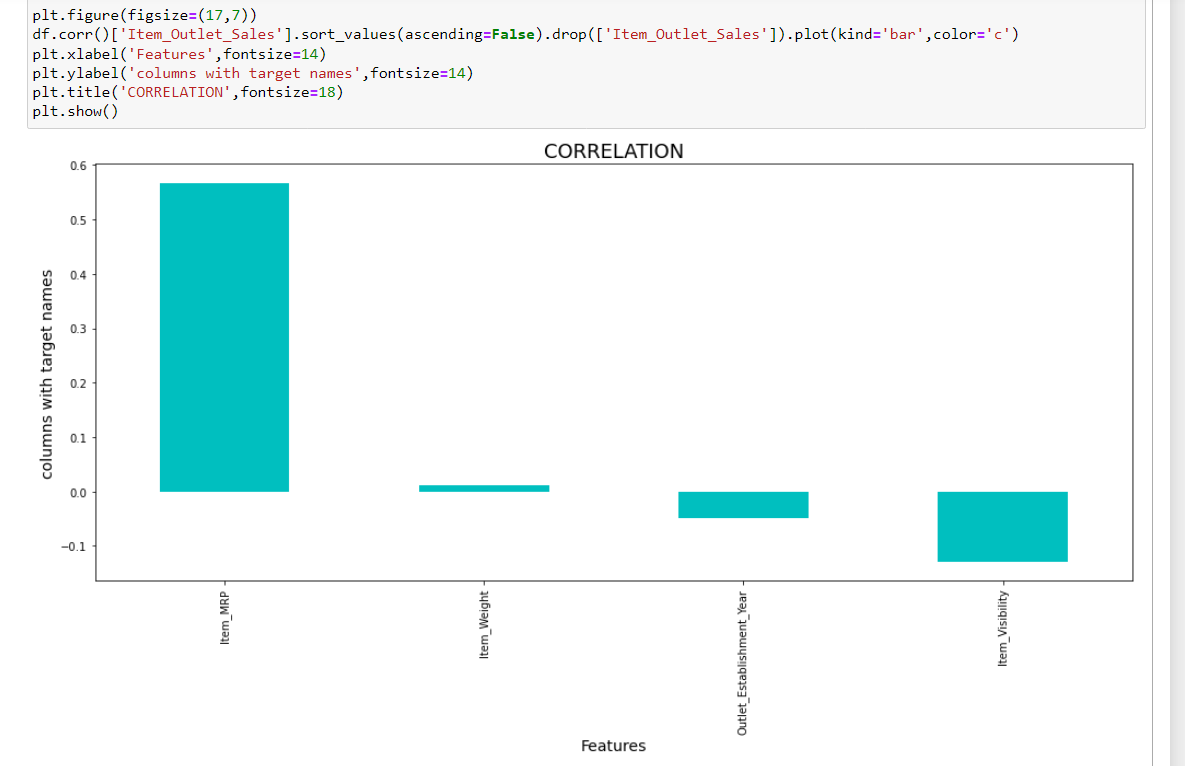
**The Tier-3 location type has all types of Outlet type and has high sales margin.**

## CORRELATION



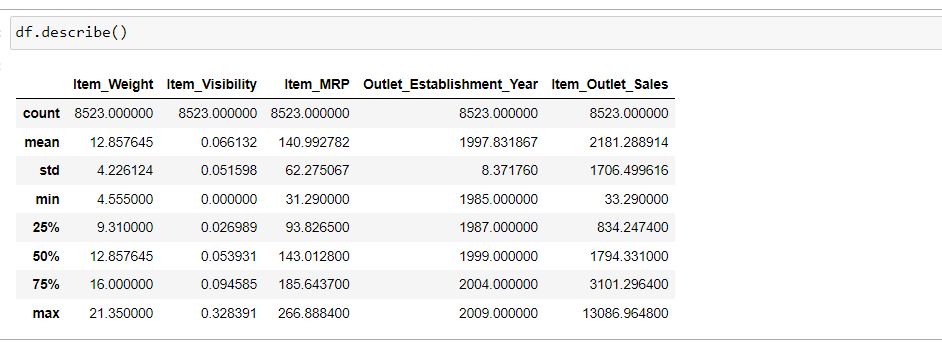
**Item Visibility have very less correlation to taget variable and this is true also we can see it on above too**

**Outlet\_Establishment\_Year also don’t have much impact on target variable ,as it does not matter when it is established ,what matters is it should according to customer interest for successful life ahead..**



**Here we can clearly see two of the columns having negative relationship.**

## DESCRIPTIVE STATISTICS

****

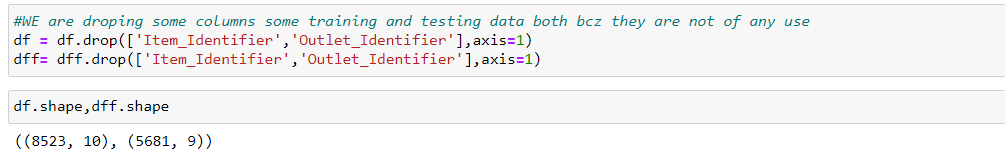
Here we get idea about the dataset and here we come to know about count of each columns, mean of each columns, standard deviation of each columns, [25percentile, 50percentile, 75percentile] of each columns and min & max of each columns.

**Total number rows and column in dataset: -**

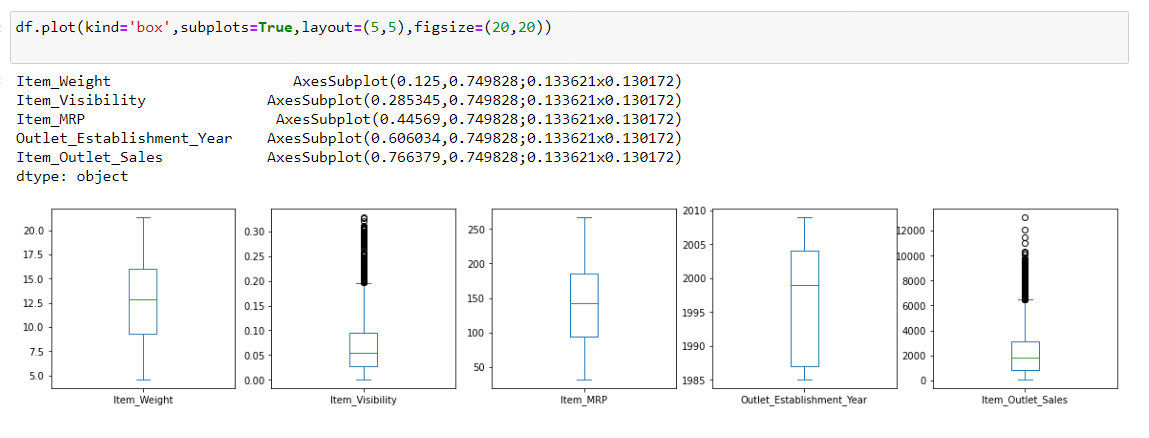
**Rows: - 8523**

**columns: - 12**

**I am deleting two columns as there are of no use for target variable: -**

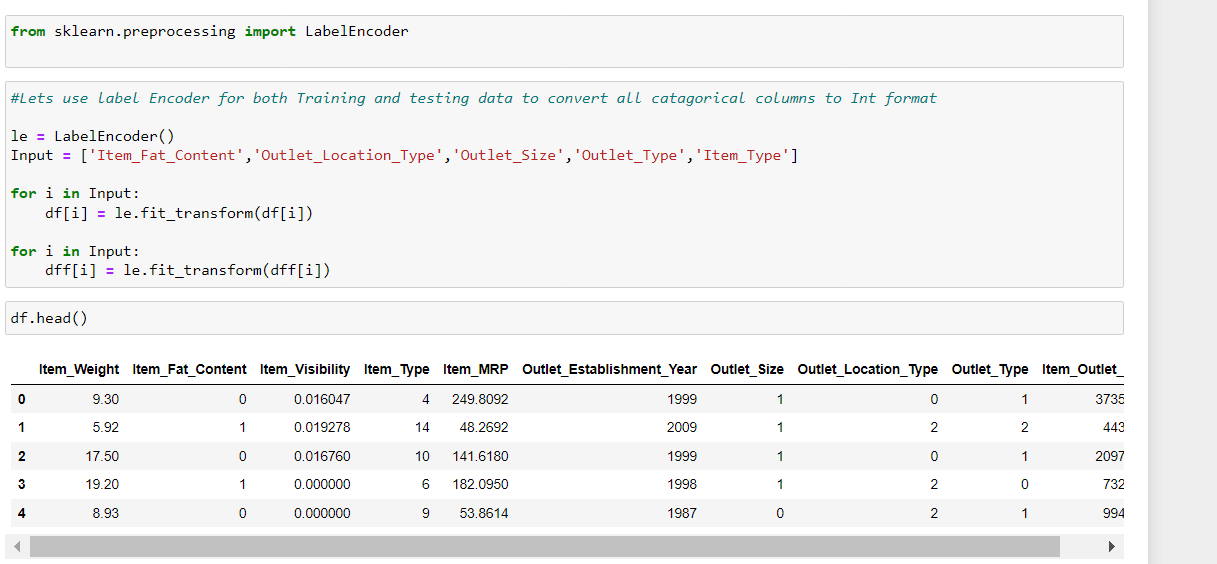
****

I have deleted it from both training and testing dataset

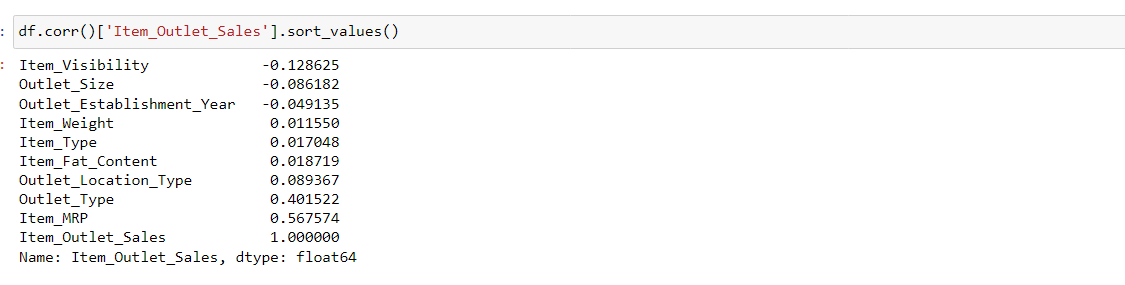


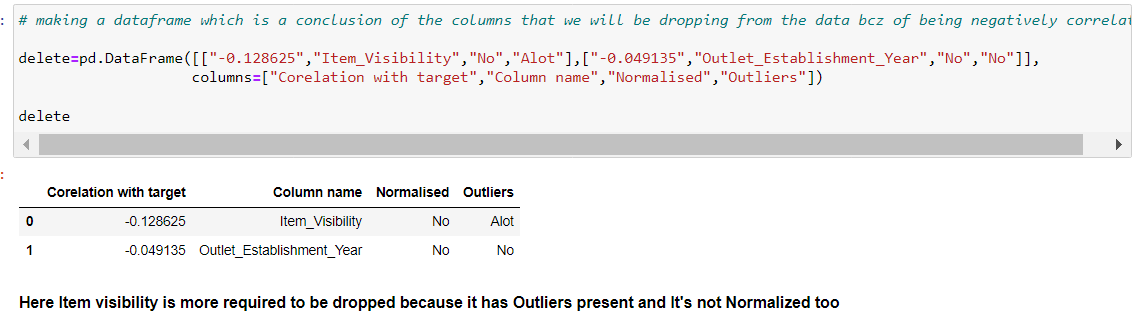
We can see that Outliers are present in the Numerical Columns.

Now Using Encoding technique to convert all categorical columns to int format: -

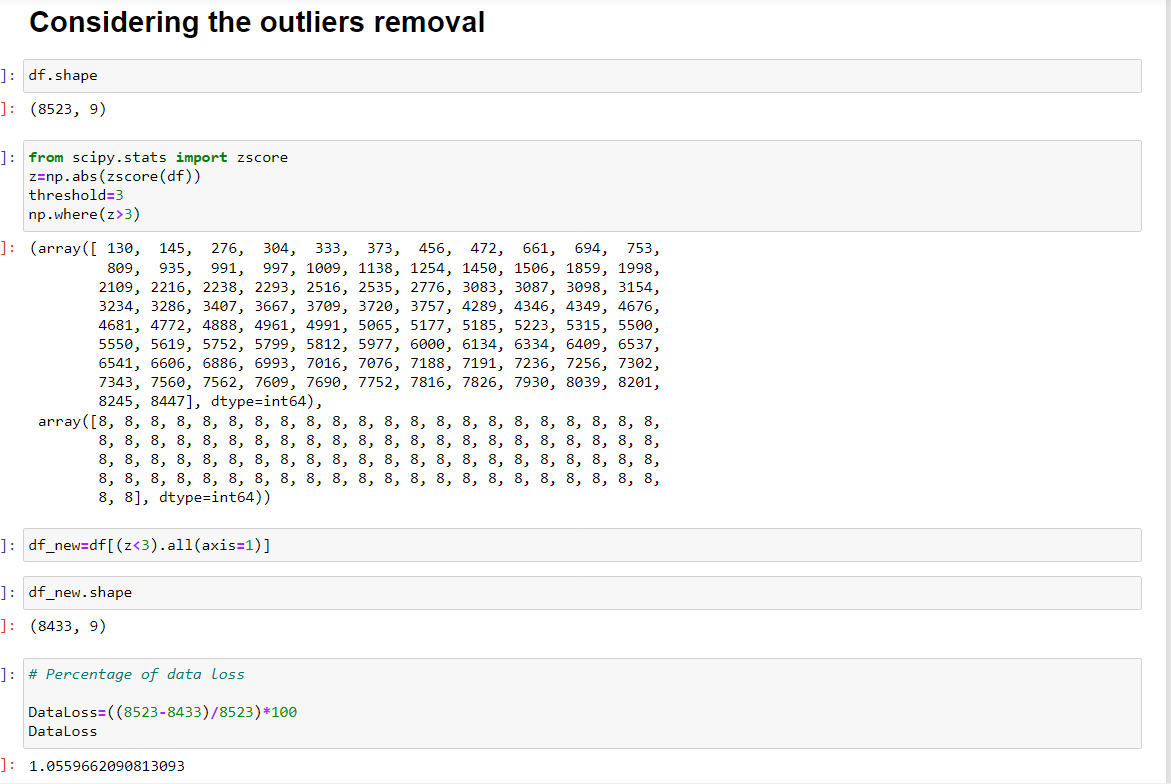


Here I have converted categorical columns to int Format

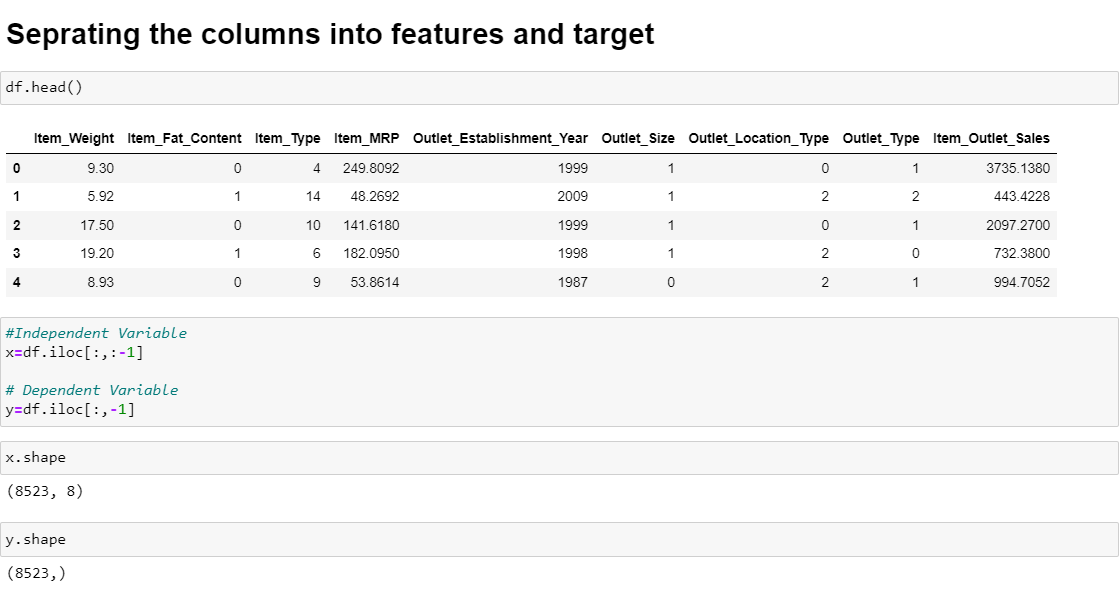




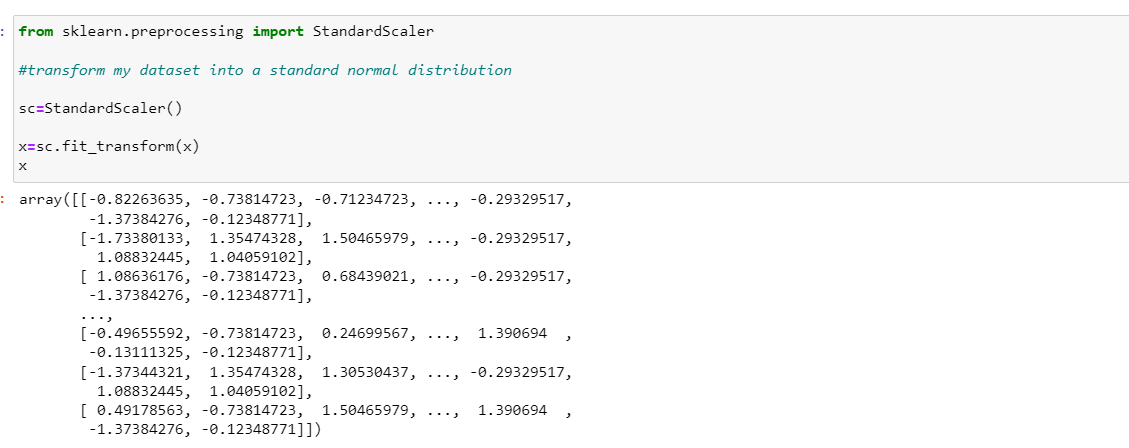
According to the correlation the columns which were acting very negatively to target variable have been dropped.



I have removed the Outlier using Zscore Method and the data that I loosed while removing outlier is only 1% ,that why I have not gone for IQR method.

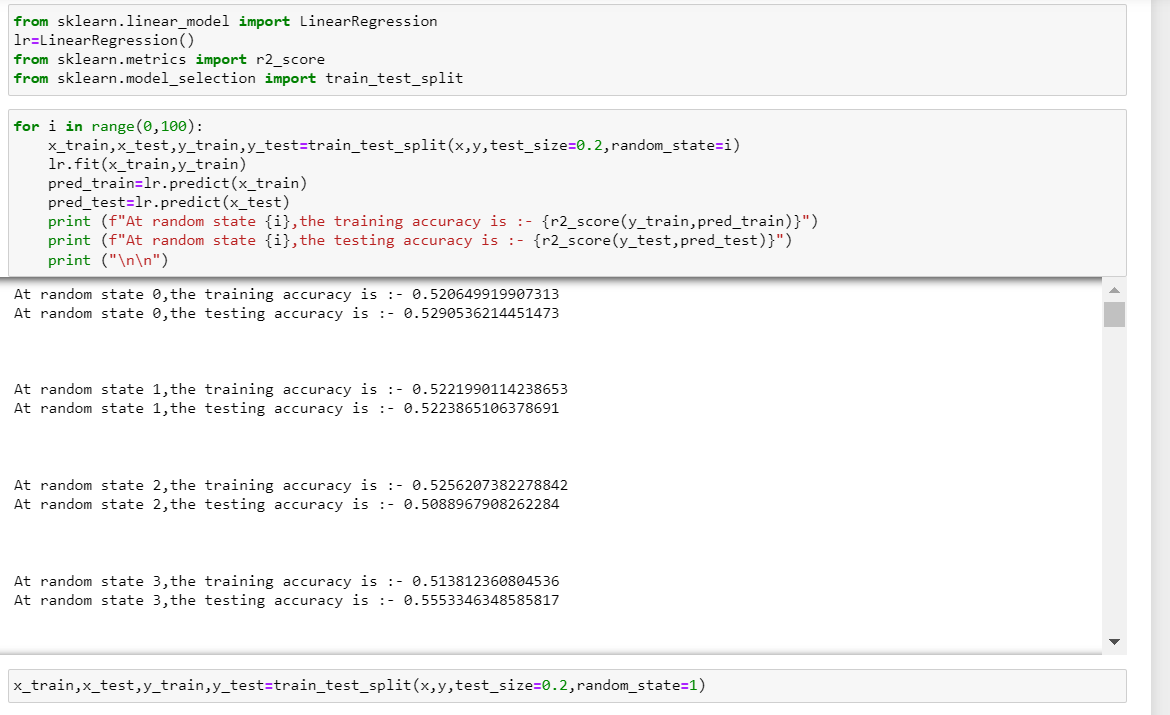


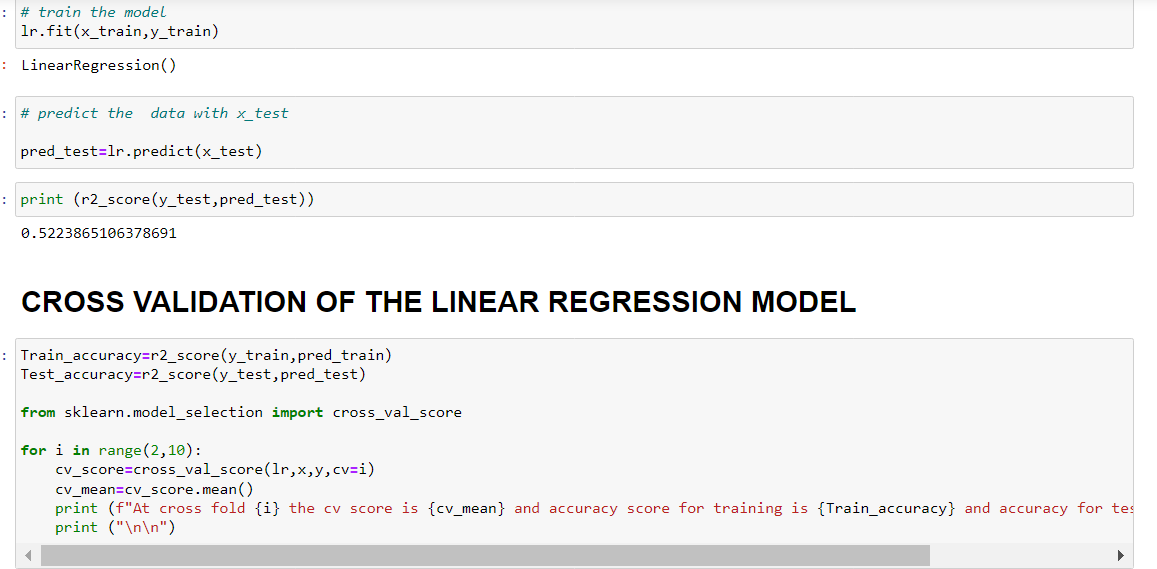
I have Separated my x(independent variable) and y(dependent variable) from the dataset.

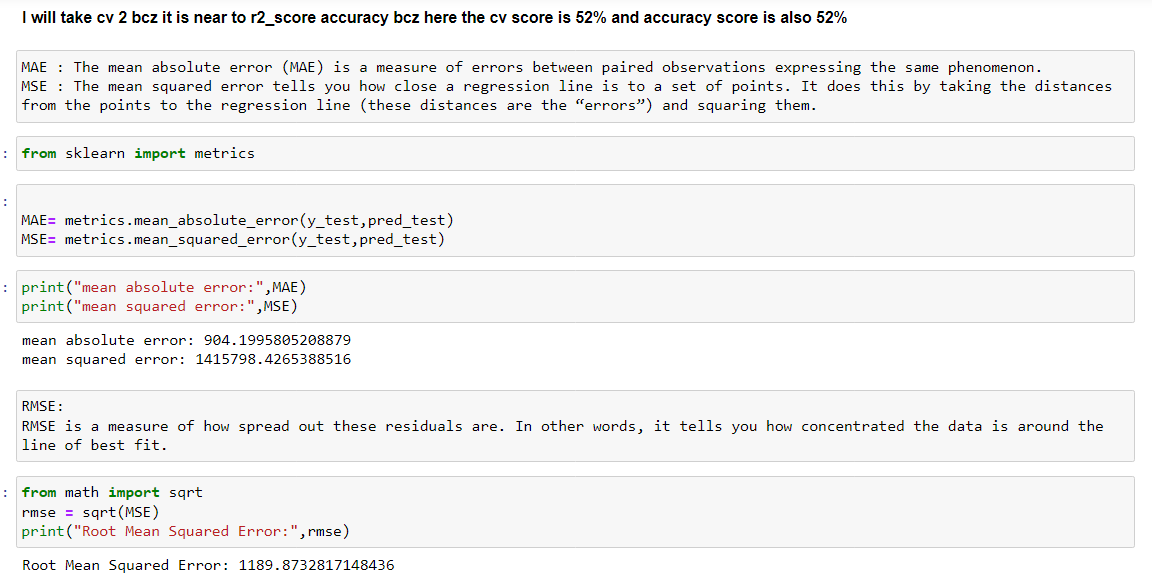


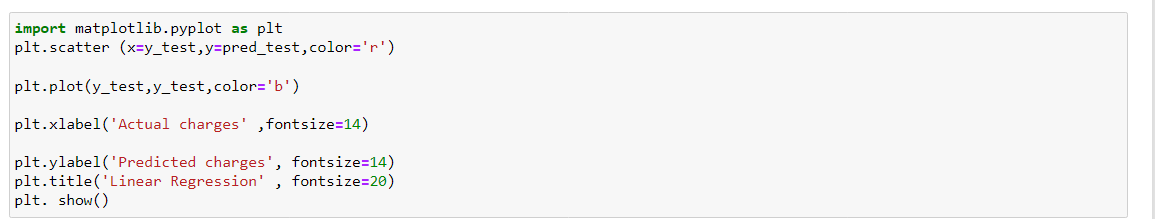
I have scaled my input variable by standard scaler, so that they all come under same format.

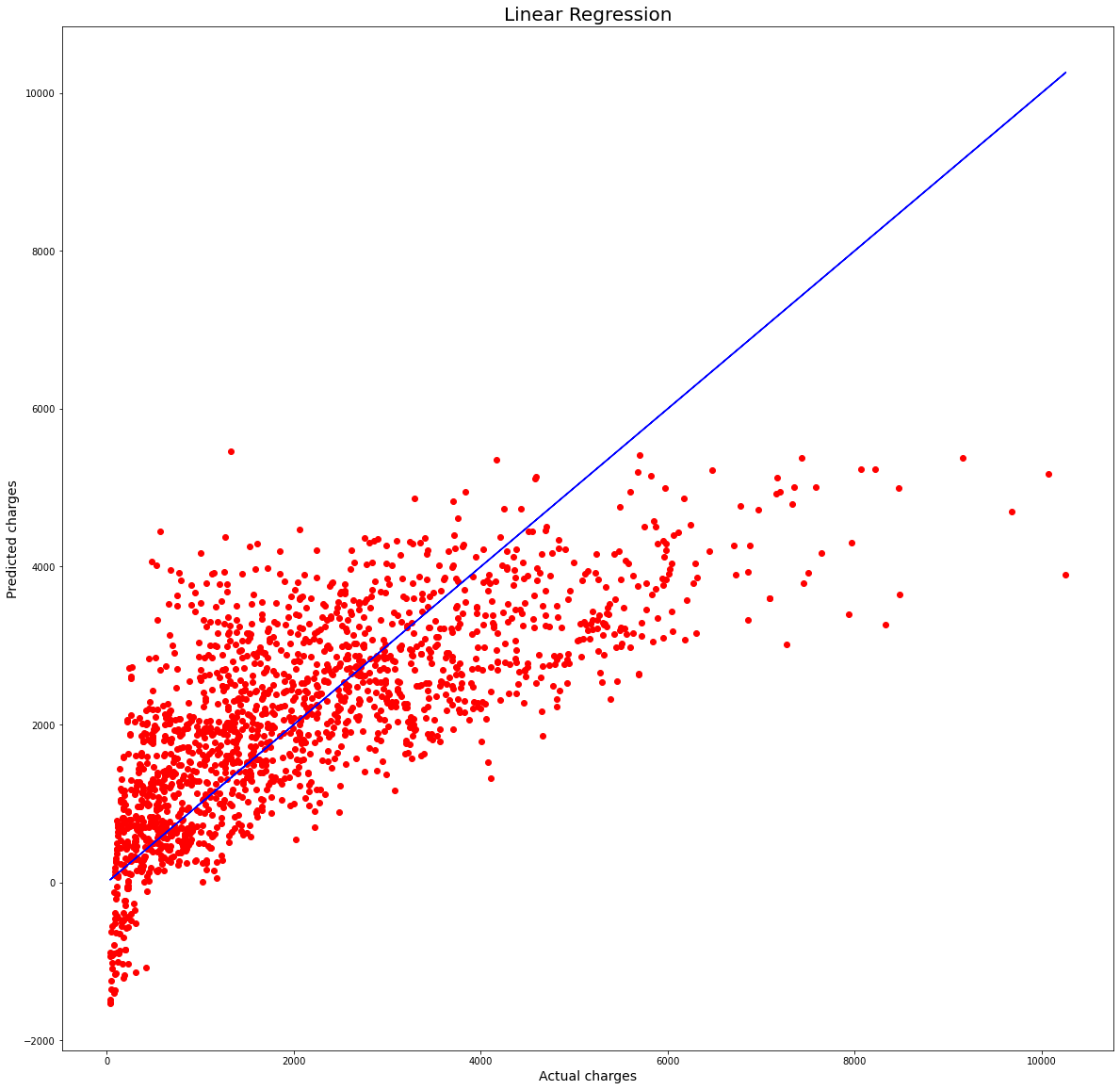
Model training and Testing: -







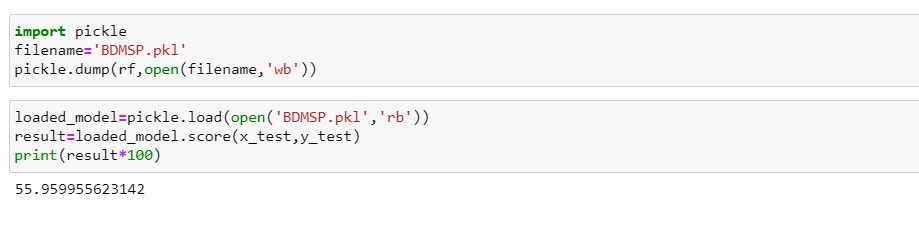




Same Like this I have done it on 4 regression model and the best was Random forest regressor so far. So, I will take it for hyper parameter tuning.



Random forest was giving 56% which was best till now out of all my models.



I have saved the model and predicted it to see the accuracy and it was performing exactly as my last model on Hyper parameter tuning.

### Summery: -

### What I did:

1. Replaced the Nans and zero values, identified outliers, feature selection and normalization - for both train and test data.
2. Visualised the data, studied the correlation amongst the data and chose the required features.
3. Built the models: I created a single model function to which I passed various different models such as Linear Regression, Decision Trees and Random Forests.
4. Calculated the Root Mean Squared Error (RMSE), predicted the sales, cross validated the scores.
5. Classified the train data and imported the results for respective machine learning models to separate csv files (which have been attached above). The Random Forest Regressor algorithm proved to be a clear winner with the lowest RMSE value.

I believe that there is further scope of improvement in this project will surely keep it in mind while doing other projects.