CAPSTONE PROJECT REPORT

PREDICTING SALE OF DIFFIERENT ITEMS IN BIG MART

GROUP 2

CONTENTS

Topics

Topic: Predictive analysis

Group members:

Shekhar Singh

Neha Mondal

Ritesh Singh

Azmath Ulla Baig

Aprajita Raj

BUSINESS PROBLEM STATEMENT

If in future company want to expand its businesses to other location then which size mart would be more beneficial to open for them, this can be easily comprehensible by analyzing the data of location and size of mart, how it is affecting the sale .

Business Problem Understanding:

- We are using big data of details about Bigmart products(items)
- The sale of the product that are sold in Bigmart depends on various attributes like, it’s size , price , visibility etc .
- The attributes (properties of the items) are correlated with each other, which indicates that change in any one of the properties may effect the sales directly and also indirectly (as other properties are also effected).
- Bigmart items has visibility and also fat content so we can target the consumers who are health conscious.

Business Objective:

- Our objective is to check how the attributes are correlated with each other.
- Do the EDA and check the major factors which are affecting the sales.
- We will also check how each attributes has effect on the sales(target value).
- Build a model to predict the sales of any new product with given attributes.

Business understanding:

The business understanding in the BigMart item sales factor, is the need to optimize sales performance and increase profits for the company. Additionally, the project recognizes the importance of data-driven decision making in achieving this goal. By analyzing historical sales data and building predictive model, this project aims to provide accurate insights into which items are likely to sell well in different locations and under different circumstances. These insights can then be used in form of inventory and marketing decisions, leading to increased sales and revenue for the company.

Approach:

- First we will read , clean and understand the data by doing EDA .
- We will divide the data into test and train set
- And try to use linear regression to develop a model for sales prediction.
- By using data analysis techniques and linear regression approach we aim to extract insights from the item sales from the dataset to determine which factors have the greatest impact on sales.
- We can optimize the company’s inventory and marketing strategies to increase sales and revenue, ultimately leading to a more successful business.
- We can also comprehend about the expansion of the business to other location, how it will affect the sales of company

PROBLEM UNDERSTANDING

- Our target value is the sales and from the data that we have we can see that the sales of the product vary from a maximum of 13000(approx) to a minimum od 33(approx).
- The variation in the sales of the items must depends on the various attributes that are collected in the data.
- The reasons for the terms not having good can be determined by analyzing the trends in the sales due to the effect of other attributes.
- And build a model for future prediction to avoid the items that may have risk of low sales.

Proposed solution to the problem:

- Analyzing the data and fetching the attributes at effect the sales the most.
- Then proposing a model that can predict the sales of any new item from the given attributes.
- This can help the company to deal with items that can have low sales.
- We are trying to solve the problem to check why there is a huge variation inthe sales of items.
- **Inefficient inventory management:** BigMart may not always know which products are likely to sell well in a given location, which can lead to overstocking or understocking of certain products. The project can help addressing this issue by providing accurate predictions of product demand, allowing the company to stock products more efficiently.
- **Ineffective marking strategies:** Without accurate data on which marketing strategies are most effective at driving sales, BigMart may be wasting resources on ineffective promotions. The project can help address this by providing insights into which marketing strategies are most effective for different product types, store locations, and other factors.
- **Lack of understanding of key sales factors:** By analyzing sales data and building predictive model, the project can identify the key factors that influence product sales in BigMart stores. This information can help the company optimize their inventory and marketing strategies, leading to increased profits margins and better customer attraction.

TECHNOLOGIES USED

While the theoretical aspects and knowledge and experience gained from a project are significant, no Data Science project can be finished without the use of technology. It is essential to fulfil the asks of the problem statement. Hence, it is important we address the non-theoretical sides behind the outcomes of this project.

We have primarily come across two types of technologies throughout the creation of our project. One of them is the programming language we have used for various calculations, data cleaning and visualization etc. while the other is the platform which provided us with the facility of making use of the said programming language. They are listed as follows:

- **Python (Programming Language)**

  Python is a high-level, interpreted and general-purpose programming language. It is often known as the "Swiss Army Knife" of programming languages. Commended for its simplicity, readability and versatility, it uses a very simple and easy to understand syntax which makes it a perfect choice for beginners to programming.

- **Jupyter Notebook (Integrated Development Environment)**

  Jupyter Notebook is a web-based programming environment which is open-source. It provides its users the ability to create and edit programs, graphs and comments all in a single document. It is a very popular tool used by data scientists and mentors of data analysis and machine learning. It is used to code in various programming languages, including Python, R, Julia, and many others.

SKILLS USED AND DEVELOPED

- The skills we have used are as follows:

- **Programming:** Basic Python programming has helped us in using libraries, doing calculations by using formulas and looping etc. throughout the timeline of the project.
- **Inferencing:** In the conceptual as well as the dataset-based part, we have made numerous inferences by looking at numbers and graphs.
- **Data Pre-processing:** We have performed outlier detection, dropping and conversion of columns, null removal and imputation on the data provided.
- **Data Visualization:** Various graphs and plots have been created from our knowledge of visualization to dig deeper into the data for information.
- **Statistical Analysis:** We have used statistical concepts for finding out probabilities and testing hypotheses and claims stated in the questions.
- **Domain Knowledge:** For the dataset-based part, the domain knowledge has been used to make significant decisions at times.
- **Exploratory Data Analysis:** To analyse the data for hidden information and to identify patterns and insights from the data, we have made use of the concepts of EDA.
- **Problem-Solving:** While we were mostly told exactly what to do to achieve results, it took us to solve undefined problems at instances.
- **Supervised Learning Regression:** To build models when the target column is numerical.
- **Supervised Learning Classification:** To build models when the target column is categorical.
- **Unsupervised Learning:** To classify all data in few categories together without having any target column.

DATA UNDERSTANDING/ DESCRIPTION:

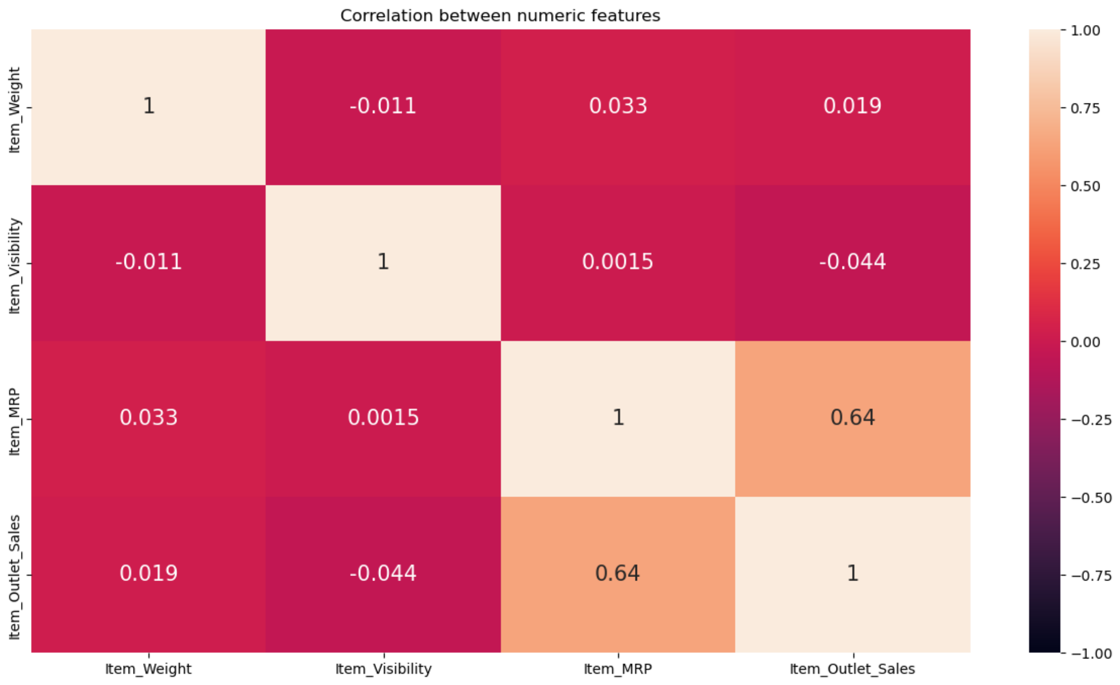
- To understand the data we did data profiling
- Then from there we checked the missing values and outliers
- One of the attribute item-size has a lots of missing values but if we observe the data we can see the item size are missing for items like fruits and vegetables as those items does not have fixed weight (depends on the customer).
- We also plot frequency graph for each attributes
- And to check correlation we plot pair plots or heat map among the attributes.
- And identifying the unique values

• Item_Identifier has a high cardinality: 1559 distinct values	• **High cardinality**
• Item_MRP is highly overall correlated with Item_Outlet_Sales	• **High correlation**
• Outlet_Establishment_Year is highly overall correlated with Outlet_Identifier and 3 other fields	• **High correlation**
• Item_Outlet_Sales is highly overall correlated with Item_MRP	• **High correlation**
• Outlet_Identifier is highly overall correlated with Outlet_Establishment_Year and 3 other fields	• **High correlation**
• Outlet_Size is highly overall correlated with Outlet_Establishment_Year and 3 other fields	• **High correlation**
• Outlet_Location_Type is highly overall correlated with Outlet_Establishment_Year and 3 other fields	• **High correlation**

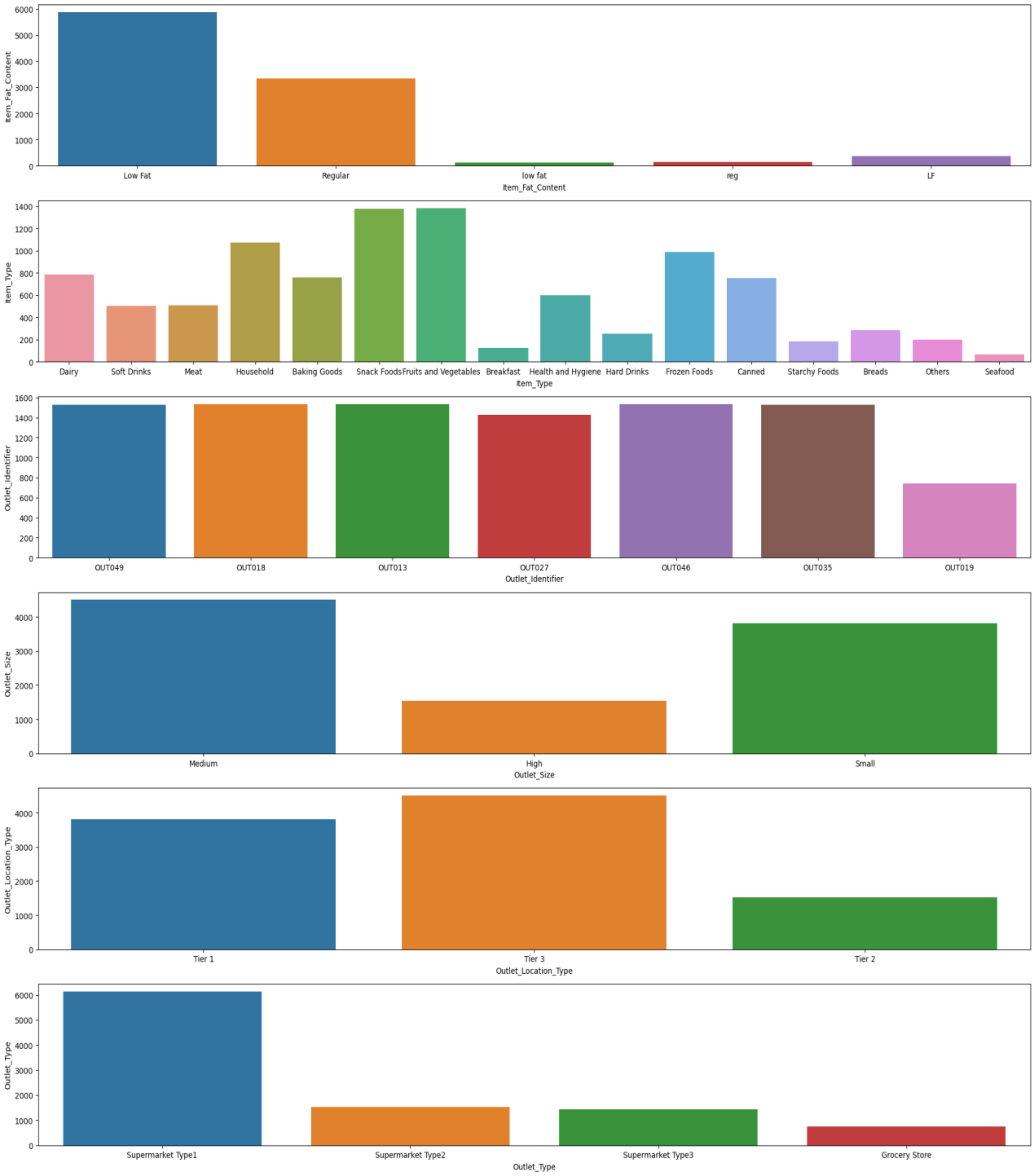
• Outlet_Type is highly overall correlated with Outlet_Establishment_Year and 3 other fields	• **High correlation**
• Item_Identifier is uniformly distributed	• **Uniform**
• Item_Visibility has 633 (6.4%) zeros	

- The attribute item_fat_content the category low fat has very high count in respect to the other category this may be the trend of the population or it is also possible that this trend came due to lack of data (bias error) , to remove this company can take more survey if this trend is seen in other sets as well then we can assume it as the trend of the population.
- Similar situation can be seen in outlet type as well.
- From the pair plot and correlation Metrix we can see that there is no mutual relation between the features. Which is a preferable in a linear regression

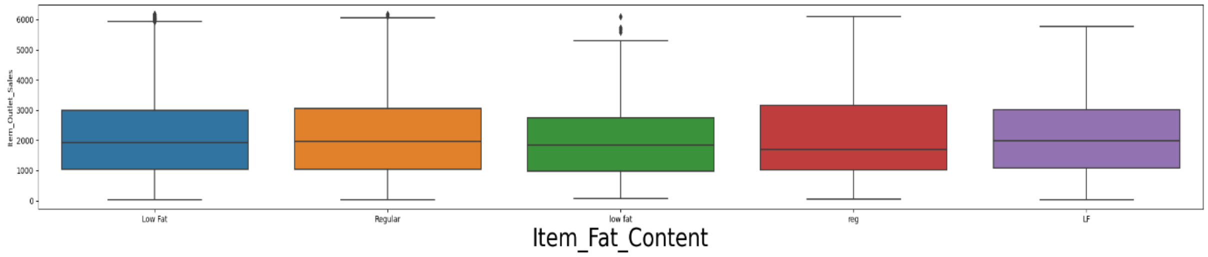
Correlation Metrix:

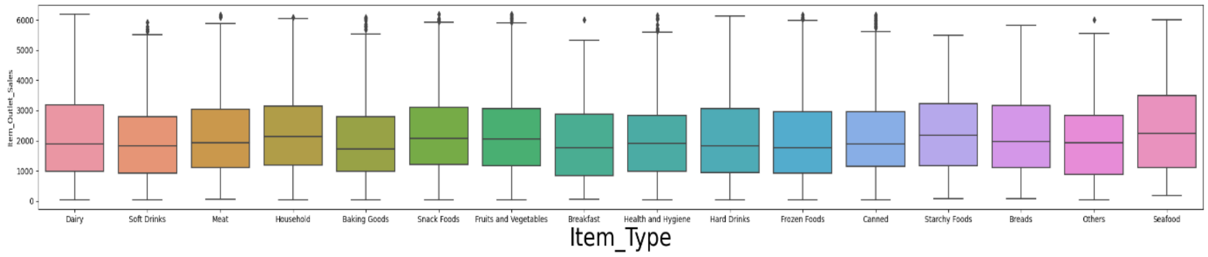


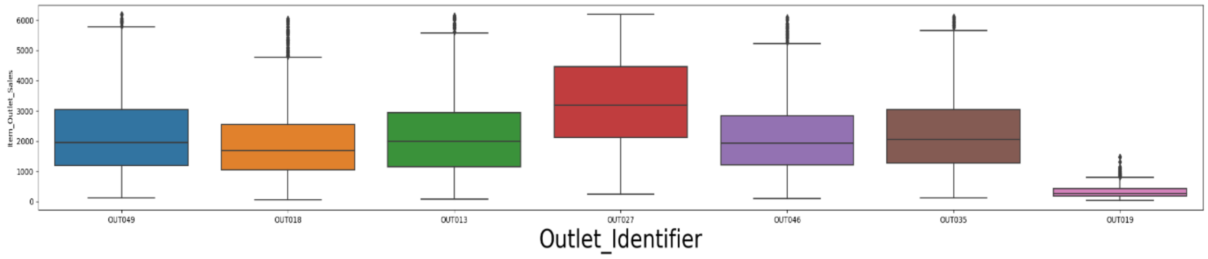
Categorical variables:

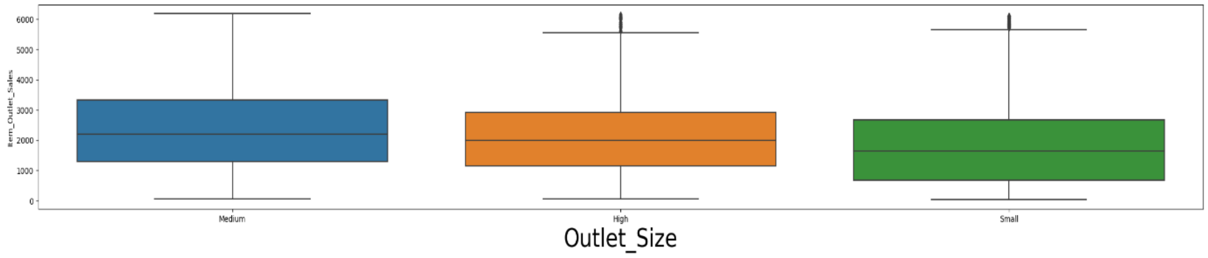


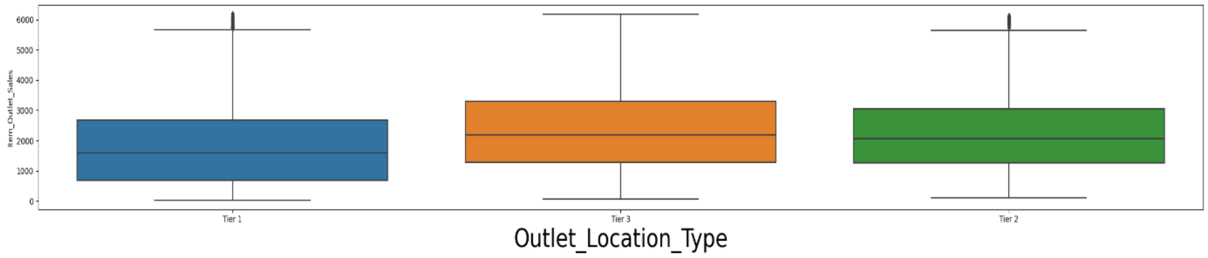
Bivariate analysis :

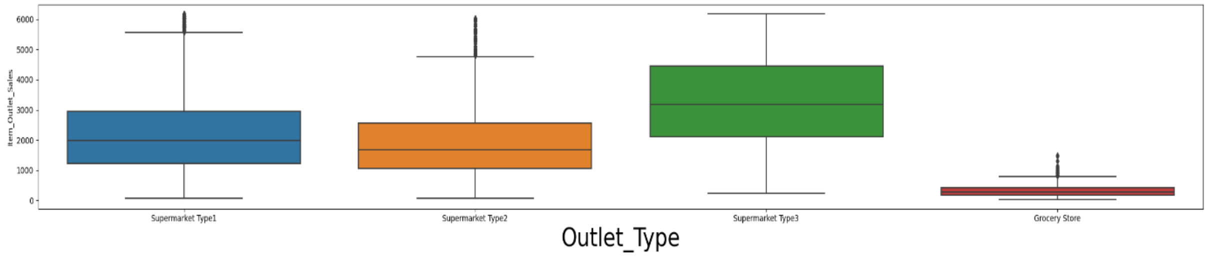








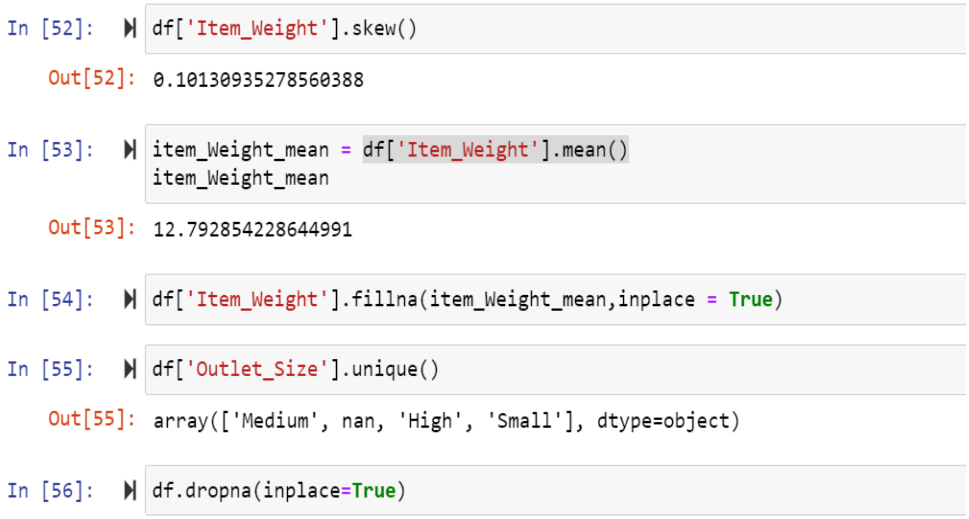




DATA PREPARING /CLEANING:

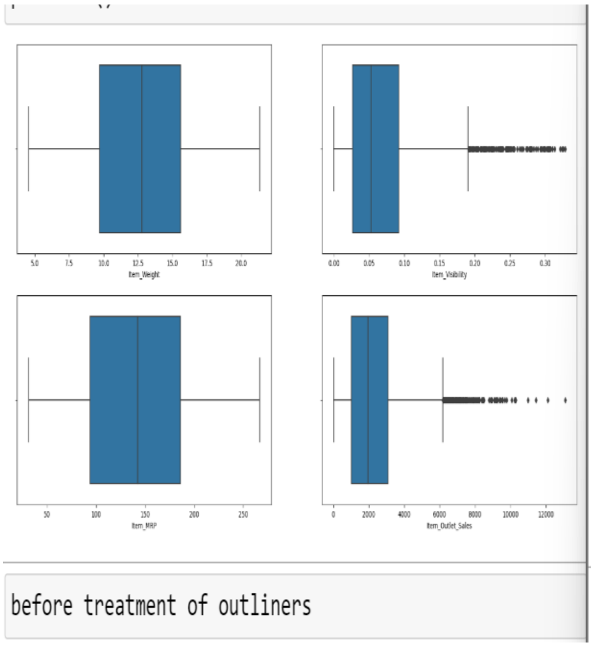
- We might need to change some categorical (like item_type , oultet_type) values to numeric values.
- Filling the missing values
- Dividing the data into test and train set
- Feature selection is also important as we have large number of attributes(12 attributes).
- Feature engineering may be done as well for numerical data's to normalize or standardize the data.
- For numerical data we did the required transformations like log transformation, boxcox
- And for the categorical features we did and dummy encoding'
- we also find out whether outlet location type is significant in affecting Item sales. We found out that outlet type has different average sales which means it does affect our sales.

Missing value treatment:

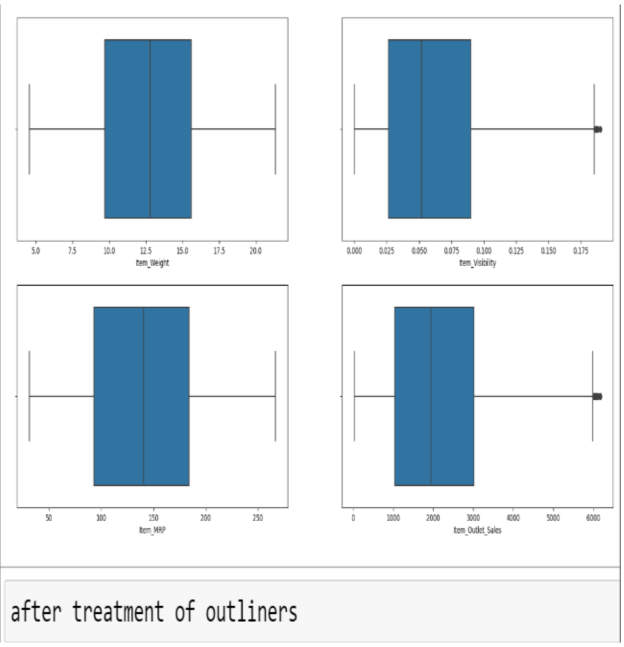


```
In [52]:  df['Item_Weight'].skew()

  Out[52]: 0.10130935278560388

In [53]:  item_Weight_mean = df['Item_Weight'].mean()
           item_Weight_mean

  Out[53]: 12.792854228644991

In [54]:  df['Item_Weight'].fillna(item_Weight_mean,inplace = True)

In [55]:  df['Outlet_Size'].unique()

  Out[55]: array(['Medium', nan, 'High', 'Small'], dtype=object)

In [56]:  df.dropna(inplace=True)
```

Outliner treatment:

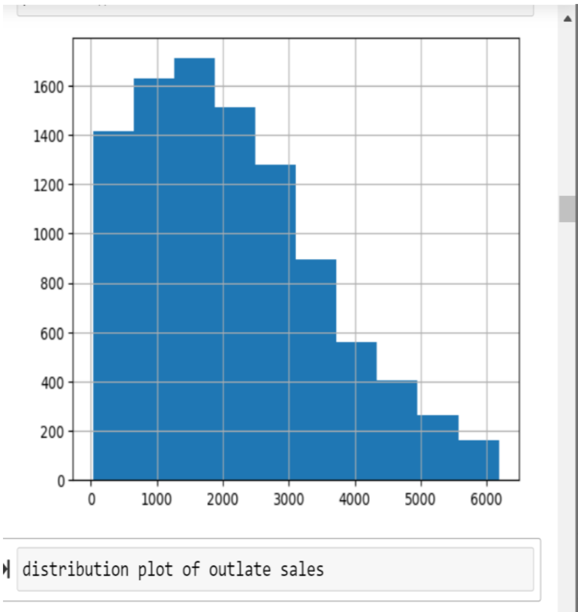


before treatment of outliners

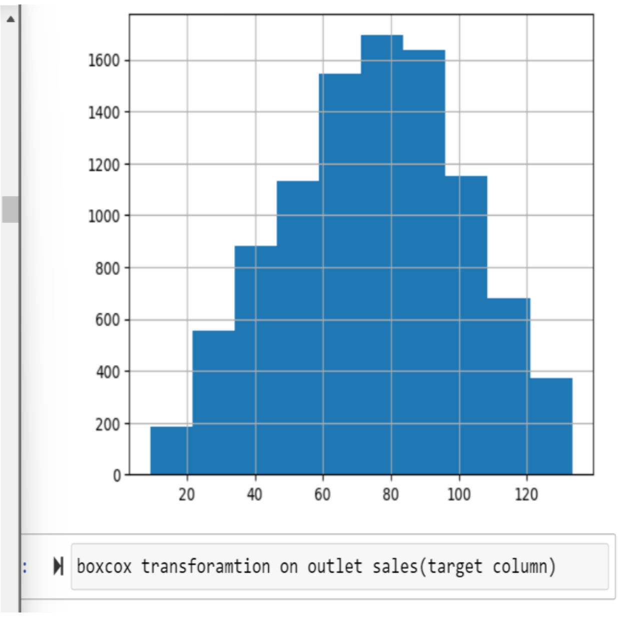


after treatment of outliners

Boxcox transformation on target variable:



distribution plot of outlate sales



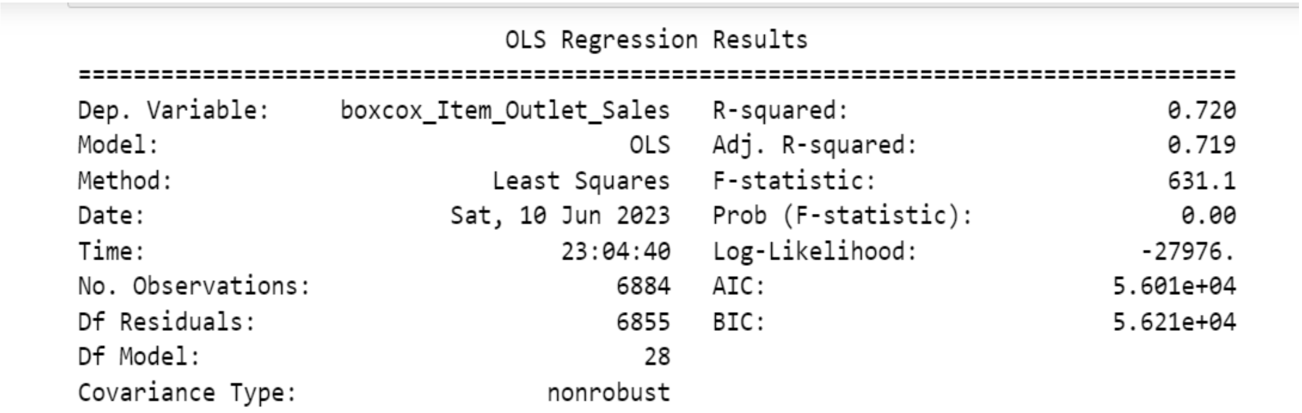
boxcox transforamtion on outlet sales(target column)

PROBLEM SOLVING STEPS /MODEL BUILDING

- As out target value is numerical we will be using linear regression method to build a suitable model.
- Use of Computational techniques, analysis techniques and statistical techniques that can help in designing better store sale performance based on the previous trend performance.
- For our first try model we build a MLR full model using the new data set that has been transformed using transformation of numerical and encoding of categorical data.
- From the r square value, we can see the model is performing nice.
- For numerical columns we checked the VIF values and found all the VIF values less than 10. hence, we can conclude that there is no multicollinearity among the numerical attributes.

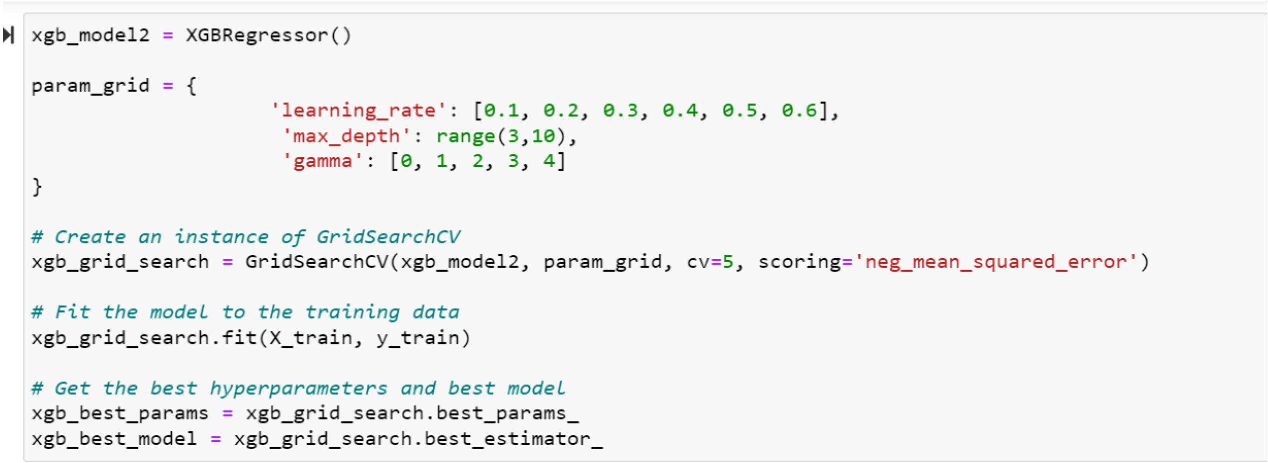
-
  ```
  VIF_Factor        Features
  0    1.001219     Item_Weight
  1    1.000115  Item_Visibility
  2    1.001109        Item_MRP
  ```
- But our model shows there is a strong multicollinearity. So, we will drop unrelated or less significant attributes and then build a model.
- We did build a model on Linear Regression SGD and our R^2 value decreased than the value we got from the MLR full model. It shows signs of under fit.
- And when we used the forward feature selection our R^2 value was above 90 but the RMSE value where around 20
- Then we build models with decision tree, Random Forest and XGboosting , all these models shown overfitting conditions.
- we applied hyperparameter tunning for all three models and after comparing the R^2 , adj. R^2 , RMSE train and RMSE test, we took XGboosting with hyperparameter tunning as the final model.
- Our final model has R^2 0.73 that means our model can explain 73% of the variation in the data.
- RMSE train = 13.6 and RMSE test = 12.9 which implies there is no overfitting conditions

BASE MODEL



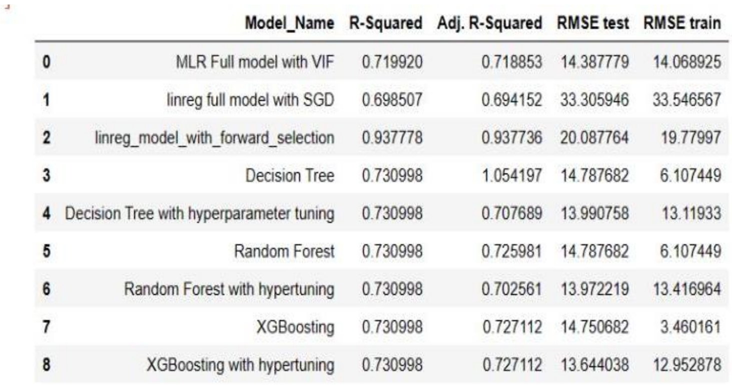
```
                            OLS Regression Results
==============================================================================
Dep. Variable:     boxcox_Item_Outlet_Sales   R-squared:                       0.720
Model:                                  OLS   Adj. R-squared:                  0.719
Method:                       Least Squares   F-statistic:                     631.1
Date:                      Sat, 10 Jun 2023   Prob (F-statistic):               0.00
Time:                              23:04:40   Log-Likelihood:                -27976.
No. Observations:                      6884   AIC:                         5.601e+04
Df Residuals:                          6855   BIC:                         5.621e+04
Df Model:                                28
Covariance Type:                  nonrobust
```

FINAL MODEL



```
xgb_model2 = XGBRegressor()

param_grid = {
                    'learning_rate': [0.1, 0.2, 0.3, 0.4, 0.5, 0.6],
                     'max_depth': range(3,10),
                     'gamma': [0, 1, 2, 3, 4]
}

# Create an instance of GridSearchCV
xgb_grid_search = GridSearchCV(xgb_model2, param_grid, cv=5, scoring='neg_mean_squared_error')

# Fit the model to the training data
xgb_grid_search.fit(X_train, y_train)

# Get the best hyperparameters and best model
xgb_best_params = xgb_grid_search.best_params_
xgb_best_model = xgb_grid_search.best_estimator_
```

MODEL SCORE



	Model_Name	R-Squared	Adj. R-Squared	RMSE test	RMSE train
0	MLR Full model with VIF	0.719920	0.718853	14.387779	14.068925
1	linreg full model with SGD	0.698507	0.694152	33.305946	33.546567
2	linreg_model_with_forward_selection	0.937778	0.937736	20.087764	19.77997
3	Decision Tree	0.730998	1.054197	14.787682	6.107449
4	Decision Tree with hyperparameter tuning	0.730998	0.707689	13.990758	13.11933
5	Random Forest	0.730998	0.725981	14.787682	6.107449
6	Random Forest with hypertuning	0.730998	0.702561	13.972219	13.416964
7	XGBoosting	0.730998	0.727112	14.750682	3.460161
8	XGBoosting with hypertuning	0.730998	0.727112	13.644038	12.952878

According to RMSE test values our best model is XGBoosting with hypertuning , so we will use this as our final model to predict the sales of BIG MART.

TAKEAWAY

- The highest item sale is 13086, lowest item sale is 33 and average sale is 2099 which is comparatively very low than the highest.
- Highest cost of item is 266.88 and average is 141.
- Visibility of item is not more than 40%
- Most item has low fat
- Sales of any item is not hugely affected by the fat content.
- Dairy and seafood has slightly more sales than other items even though no.of dairy item sold is very much higher than seafood.
- Outlet 019 has very less sales compared to other outlet
- Teir 3 has more sales.
- Grocery store has very less sales
- Even though the supermarket type 3 and type 2 have less no.of item sold still its contribution in sale is not less than other type

CONCLUSION AND FUTURE STEPS

Conclusion:

- The sales hugely differ for different locations.
- There are items with low cost, low sales.
- Items like dairy, fruits and vegetables counts the most. Seafood has very less count but has huge sales, hence we can conclude that the overall sales of fresh product is good.
- Grocery product need more attention. We should filter the grocery products based on different location.

Future steps:

- As we concluded sales dependence on the variables are different in driftnet locations hence, we can make necessary modification in our model for different location for better sales
- We can include more type of fresh product in the mart as its sales are seen to be more.
- Taking product feedback from the customers will help to improve in item selection
- Once the sales of the existing store are increased, stores can also we opened in other locations