EXPLORATORY DATA ANALYSIS AND APPLYING MACHINE LEARNING CLASSIFIERS TO PREDICT BIG MART SALES

TEAM MEMBERS

Jaya Sree Myla-AP19110010367
Manasa Nunna-AP19110010382
Jayanth Balla-AP19110010394
Tanmayee Chandanam-AP19110010402

PROJECT REPORT

A report submitted in partial fulfillment of the degree B-tech in Computer Science Engineering

Mentor: ANABIK PAL

DEPARTMENT OF COMPUTER SCIENCE

SRM UNIVERSITY, AP, AMARAVATI

APRIL-2022

DECLARATION

We solemnly declare that the project report is based on our own work carried out as a part of the **“Undergraduate Research Opportunities”** under the guidance of **“Dr. ANABIK PAL”** . We also imply that different techniques are performed on the dataset using different approaches and conclusions are drawn as an outcome of research work.

- The work contained in the report is obtained from various sources online and they are cited for reference at the end of the report.
- The project has been done by us under the general guidance of our mentor.
- The work has not been submitted to any other Institution for any other degree/diploma/certificate in any other University of India or abroad.
- We have followed the guidelines provided by the university in writing the report.

ABSTRACT

Nowadays, shopping malls and arcades maintain *track* of their sales data for each individual item in order to *forecast future* client demand and adjust inventory management. In a data warehouse, these data stores fundamentally include a vast quantity of consumer data and individual item information. Anomalies and common patterns are discovered through mining the data warehouse's data storage. For businesses like Big Mart, the resulting data can be utilized to anticipate future sales volume using various machine learning approaches.

The case of Big Mart, a one-stop-shopping center, the data has been explored in this paper in order to predict the sales of various types of python libraries like dtale, klib and pandas profiling and to comprehend the effects of various elements on the items sales. Using several components of a dataset gathered for Big Mart, and results with high degrees of accuracy are created, and these observations can be used to make decisions to boost sales, thanks to the methods used to build a predictive model using machine learning techniques.

Machine Learning is a class of methods that allows software to improve its accuracy in predicting events without having to be explicitly coded. Machine learning's primary assumption is to create models and algorithms that can take in data and apply statistical analysis to predict an output while updating outputs as new data becomes available. These models can be used in a variety of situations and trained to match management's expectations so that precise procedures can be taken to meet the organization's goals.

In this research, we present a predictive model for predicting the sales of a company like Big Mart using linear regression, random forest, to modify the business model to predict outcomes, the sales estimate is based on Big Mart sales for various outlets. Using various machine learning approaches, the resulting data can be utilized to anticipate possible sales volumes for shops like Big Mart.

The proposed system's estimation should take into account the item_weight and outlet_size. Various machine-learning methods, such as “linear regression” and “random forest” algorithms are used in a variety of networks. Finally, *hyperparameter tweaking* is utilized to assist in selecting relevant hyperparameters as an *optimization technique* in the data science process that allows the algorithm to shine and give the best results.

ACKNOWLEDGEMENT

Conducting this project study has been one of the most enlightening and interesting internships. This is an honest effort towards putting forward whatever we have gained as a valuable experience that will surely help us move up the learning curve towards a path we have chosen.

It is our pleasure to be indebted to our friends, family who directly or indirectly contributed to the development of this work and who influenced our thinking, behavior and acts during the course of study.

We are thankful to **DR. ANABIK PAL** for providing the valuable time and guidance in elaborating views of studying the project details and getting the right vision for its implementation.

TABLE OF CONTENTS

CHAPTER-1

INTRODUCTION

1.1 Aim & Motivation of the project:

- The goal of this project is to perform the data science processes on the Big Mart dataset and apply the machine learning model to predict which item the owner should focus on to boost their sales.
- To come up with the optimal model that is more efficient and produces accurate results quickly using Machine Learning,
- To implement random forest and Linear Regression models on the "Big Mart Sales" dataset.
- To determine what elements can help them enhance sales and what adjustments could be made to the product or store's design characteristics.
- Finding different performance metrics for each model after training and deciding which method to use for the given dataset in future.

Motivation:

Data science experts, like any other machine learning endeavor, require data to work with. Researchers determine what data they need to collect based on their objectives. The data is then prepped, preprocessed, and translated into a format that can be used to develop machine learning models. Another important aspect of the job is determining the best ways for training machines, fine-tuning the models, and picking the best performers. After selecting a model that provides the most accurate predictions, it can be put into production.
The following is an example of the type of work data scientists do to construct ML-powered systems that can predict client attrition:

- Understanding a problem and final goal
- Data collection
- Data preparation and preprocessing
- Modeling and testing

1.2 Problem Statement:

Using a machine learning technique to forecast BigMart transactions allows data scientists to investigate different trends by shop and product to find the most effective solutions. Many organizations are heavily reliant on their data and require market forecasts. Every shopping center or store tries to help the individual and current proprietor attract more customers depending on the day, so that the business volume for everything may be assessed for organization stock administration, logistics and transportation administration, and so on.

For assessing deal volume in various BigMarts across different domains, several Machine Learning techniques such as linear regression, random forest are used to handle the issue of deals anticipation of goods based on client's future requests.

To analyze, explore, and apply various EDA techniques(preprocessing) on the dataset and find the relationship between the columns in the dataset. Then the various machine learning models are applied to the dataset after preprocessing. In later steps, hyper parameter tuning is done so as to optimize the model and the model is evaluated using several performance metrics. To derive the accuracy of various models and predict which model is best to use on the dataset to get most accurate results.

1.3 Proposed Model:

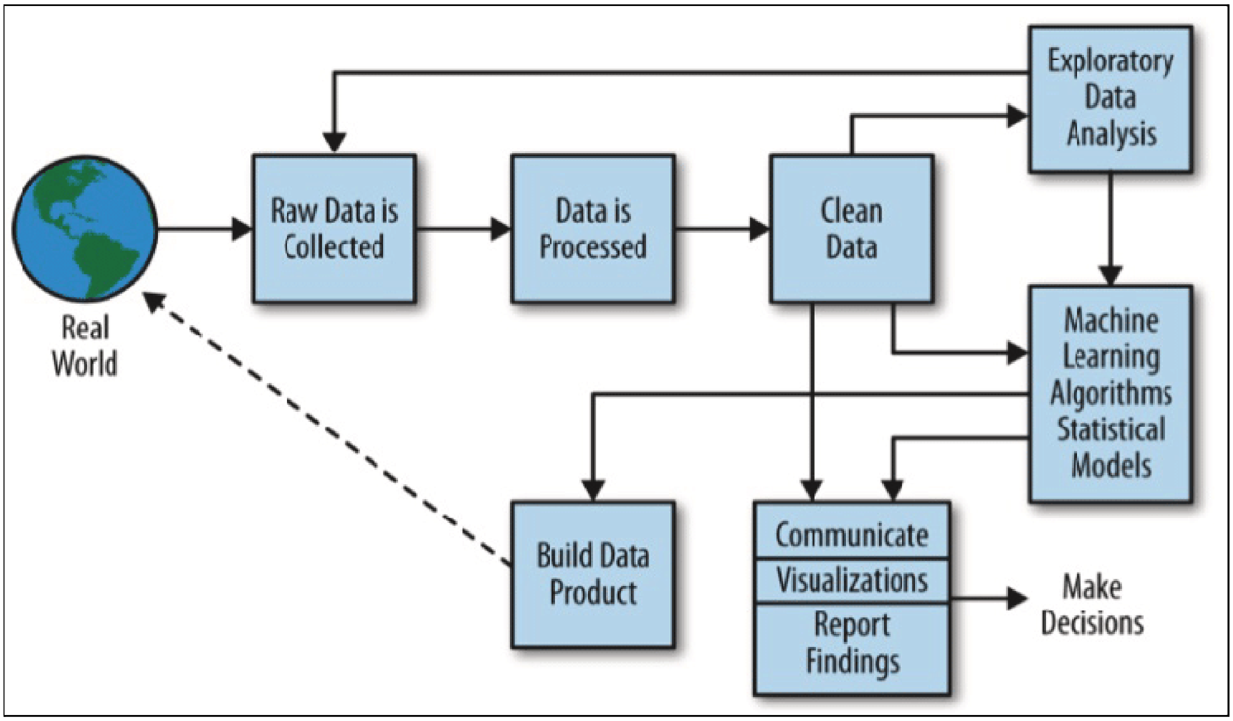


Fig 1. Data Science Process

1.4 Different initiatives taken to solve the problem:

From the above mentioned data processes, we are going to perform the following methods in our project

- Data Collection
- Data preparation
- Model Training
- Model Optimization
- Model Evaluation

CHAPTER-2

METHODOLOGY

The steps followed in this work, right from the dataset preparation to obtaining results are represented in Fig 2.

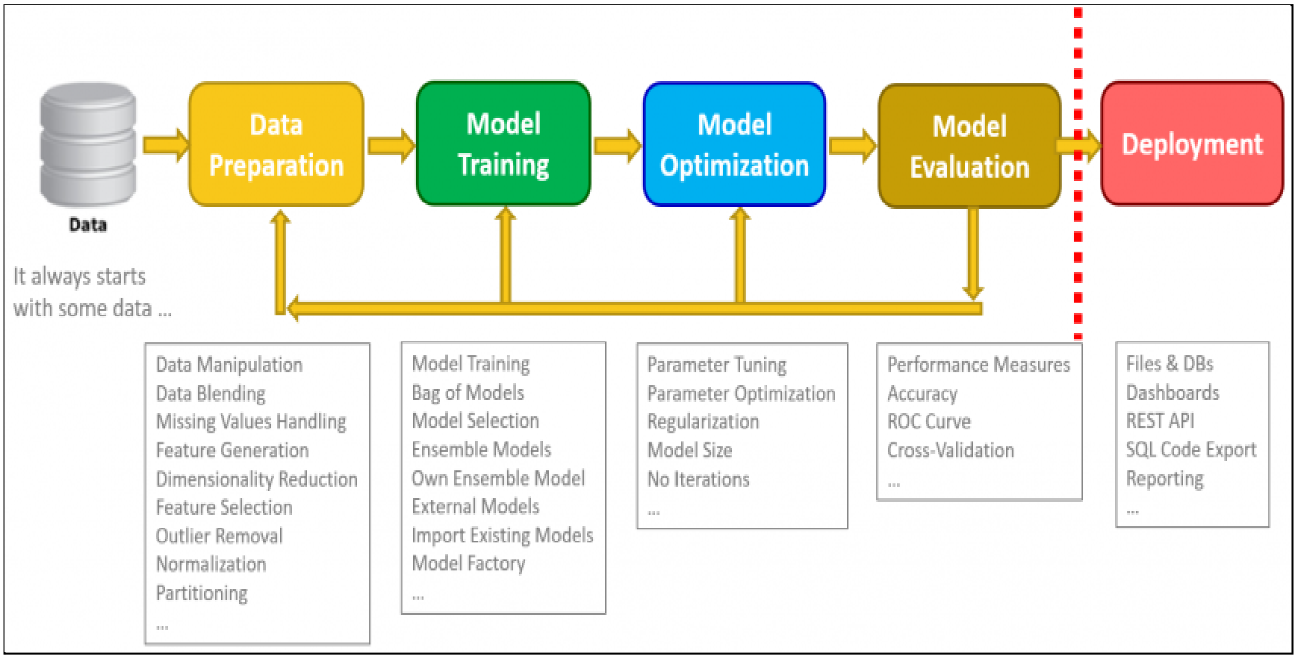


Fig 2. Data Processing Steps

2.1 Dataset Description:

The dataset contains the *train (8523)* and *test (5681)* rows-, and the train data set has both input and output variable(s). For the test data set, we need to forecast sales.

Table 1:

Variable	Description
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. This is the outcome variable to be predicted

The process starts with data collection, preparing the data and then training the data using our model, optimizing and evaluating the model with performance metrics. If the performance of our model is not accurate then again optimize the model or train the model or prepare data based on requirement to improve its perfomance. The main aim is to predict the "*item_outlet sales"*.

There are two types in our dataset train and test. But we will work majorly on the train dataset rather than test data. Because the actual test dataset does not contain target labels. The accuracy of the model can't be verified without the target variables to compare it with in the test dataset. In general, after testing data on the model, a column is created in the dataset to put the predicted values into that dataset. The accuracy of the model can't be predicted by simply adding data into the .csv file.

So the train data is divided into train and test and then the predicted result is compared with test data. First the preprocessing techniques are performed on the train and test dataset, so as to prepare the data for model implementation. Then the training dataset is divided into train and test and train the model using training data and testing the model using the test dataset that is derived from the training dataset. As the derived test dataset contains target variables it is easier to find the accuracy of the model.

2.2. Data Science Process:

To perform the data science processes step by step.

1. **Data Collection :** Data is collected and loaded from kaggle and all the coding part is done in google colaboratory by importing necessary python libraries.

2. **Data Preparation :**

   A. **Data Manipulation** - It is the process of adjusting data to make it organized and easier to read. Data is put in csv format/json format. The big mart dataset is already organized. So the data manipulation is not performed.

   B. **Data Blending -** Data blending is a method for combining data from multiple sources.We have only train data for training and test data for testing. There is no multiple data to merge and training. This step is therefore unnecessary.

   C. **Missing Values Handling -** Missing data is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset. The dataset contains some missing values.The missing values should be handled by either dropping them or filling them.

   Missing values in the train dataset are depicted in table 2 (same variables are having missing values in the test data with different count) :

Table 2

Variable	Number of Missing Values
Item_Weight	1463
Outlet_Size	2410

According to context, it's not good to drop those values. The shape of train dataset is only 8523*12. If 2410 rows are dropped the dataset will become relatively small. So missing values are handled without dropping them.

Handling Missing Values:

- To find if the missing values are either numerical or categorical.
- Item_weight is numerical values containing null values.
- Outlet_size is object Dtype so it is categorical

Imputation is a technique for replacing missing data with a substitute value while retaining the majority of the dataset's data/information.

- ☐ Mean imputation is performed on numerical values
- ☐ Mode imputation is performed on categorical values

- Item_weight is numerical so mean imputation is done
- Outlet_size is categorical so mode imputation is done

D. **Feature Generation -** Feature Generation involves creating new features which can improve model accuracy. Selecting important features or to generate more features(columns).

E. **Dimensionality Reduction -** Dimensionality reduction refers to techniques that reduce the number of input variables in a dataset.Dimensionality reduction transforms features into a lower dimension.

F. **Feature Selection -** Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data.
More input features often make a predictive modeling task more challenging to model, so the unwanted features/columns are dropped. Item_identifier and

outlet_Identifier are not important because outlet_sales are not dependent on the id of the item. So, those columns are dropped.

G. **Outlier_Removal -** Some datasets contain extreme values that are outside the range of what is expected and unlike the other data. When any machine learning model is being applied on any dataset, it is better to remove these outliers so as to increase the efficiency of the model and increase the model skill.
Here comes the **Exploratory Data Analysis**
EDA is performed using 3 methods in the dataset.
   - Dtale library
   - Pandas profiling
   - Klib library

   After the Exploratory analysis of data the statistical analysis of the data, correlation heatmaps, graphs describing their distribution, etc, are found out.

H. **Normalization -** It is the method of organizing data to appear similarly across all records and fields. Normalized data set is ranging in a very vast way. Normalization is done so that mean is 0 and standard deviation is 1. To keep the values close to model and to converge better standardization is required. Data needs to be normalized for linear regression, deep learning algorithms, CNN, K-means, logistic regression whereas data is not required to be normalized for random forest and decision tree. Here a random forest model is applied, so the data is normalized.

I. **Label Encoding -** Label Encoding is an encoding technique for handling categorical variables. In this technique, each label is assigned to a unique integer based on alphabetical ordering. Here label encoding is performed on categorical columns.

J. **Partitioning** - Partitioning is used when the model for the data is being chosen from a broad set of models. The main idea behind data partitioning is to retain a subset of accessible data out of the analysis process and use it afterwards for model testing.

2. Model Training :

Models are built using pre-processed data and multivariate analysis is performed using learning algorithms. To train a machine learning model is to feed an ML algorithm with data to help identify and learn good values for all attributes involved. There are various types of machine learning models, with supervised and unsupervised learning being the most common.The Random Forest, linear regression, XG boot regression are used for model training.

3. Model Optimization :

Hyper parameter tuning is performed on the model to find which parameters give better results and they are used for model evaluation.

4. Model Evaluation :

Every model is evaluated using metrics like accuracy, error, root mean square error, r squared value etc. All these are represented in the graphs below in the results(chapter 3).

2.3 Machine learning Algorithms:

Linear Regression:

A statistical process for estimating the relationship between a *dependent variable* and *independent variable* (predictor). It is a method of modeling a target value based on independent predictors. This method is mostly used for forecasting and finding out the cause and effect relationship between variables.

Regression Equation:

$$\boxed{Y_i = \beta_0 + \beta_1 X_i}$$

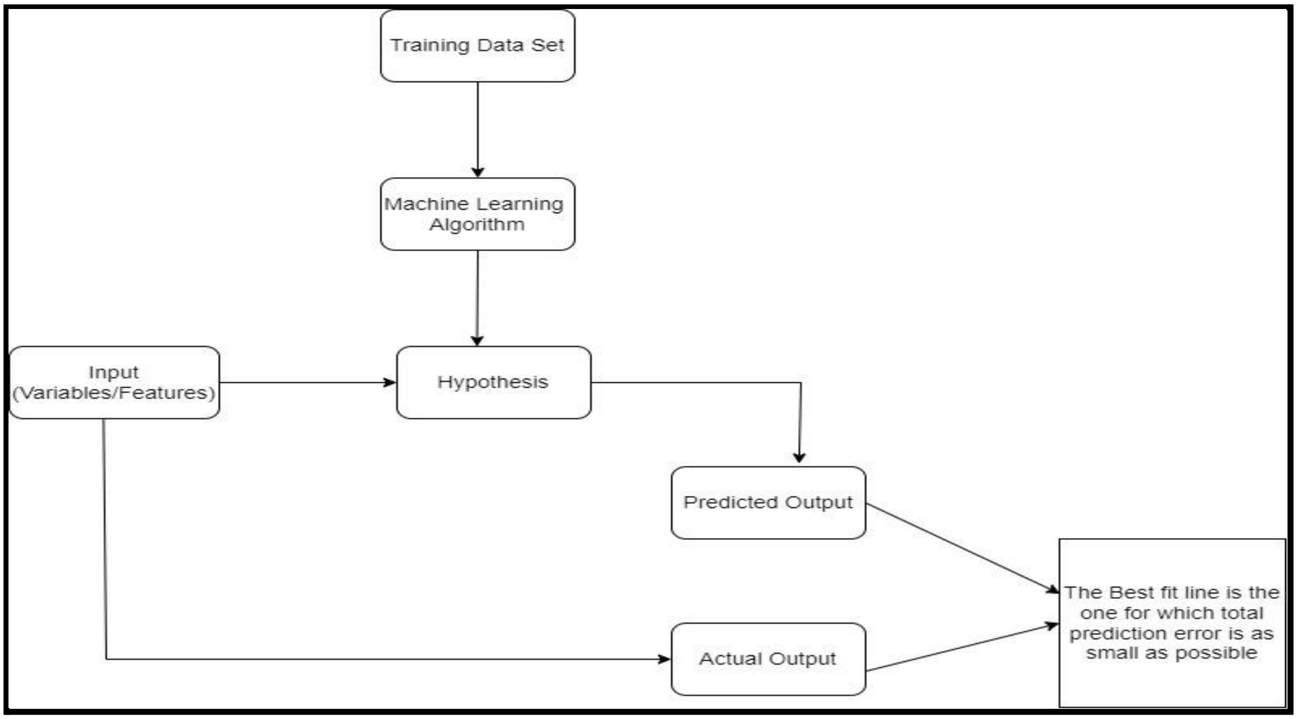

Training Data Set
Machine Learning Algorithm
Input (Variables/Features)
Hypothesis
Predicted Output
Actual Output
The Best fit line is the one for which total prediction error is as small as possible

Fig 3. Algorithm for linear regression

Random Forest:

It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

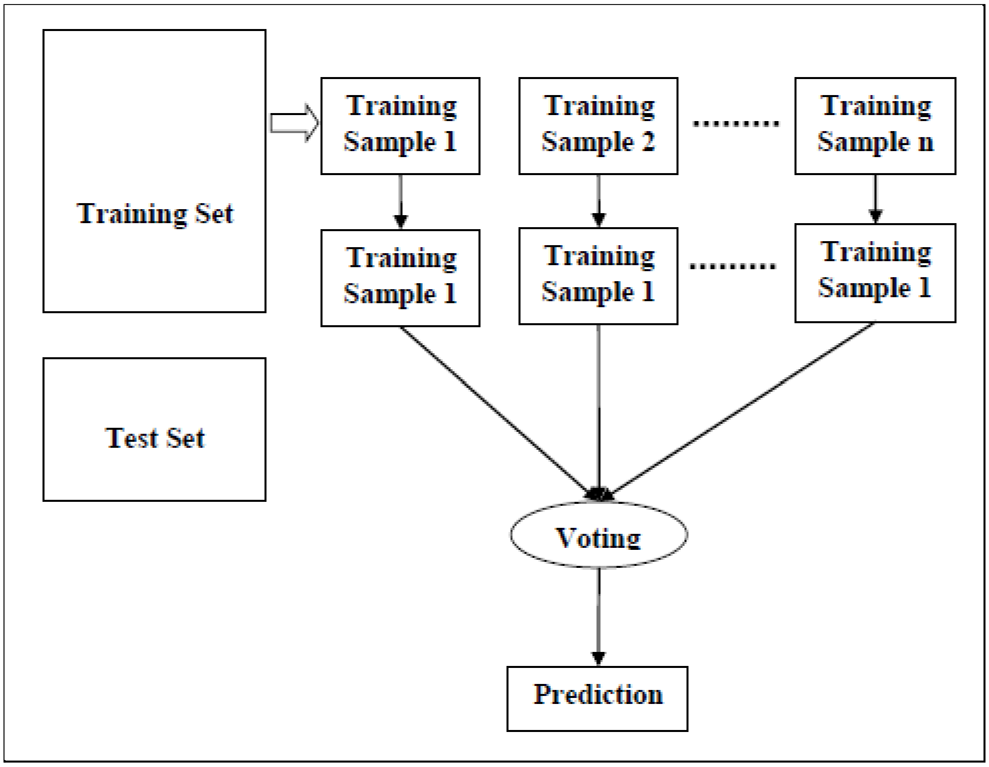


Fig 4. Algorithm for Random Forest

The dataset is split into 80% for training data and 20% testing data. Then the preprocessing methods are applied on the dataset to prepare data to apply the machine learning models. Machine learning models Linear Regression and Random Forest are implemented as training data as input and they are further hyper parameter tuning is performed on these models to optimize those models and to give the best results.

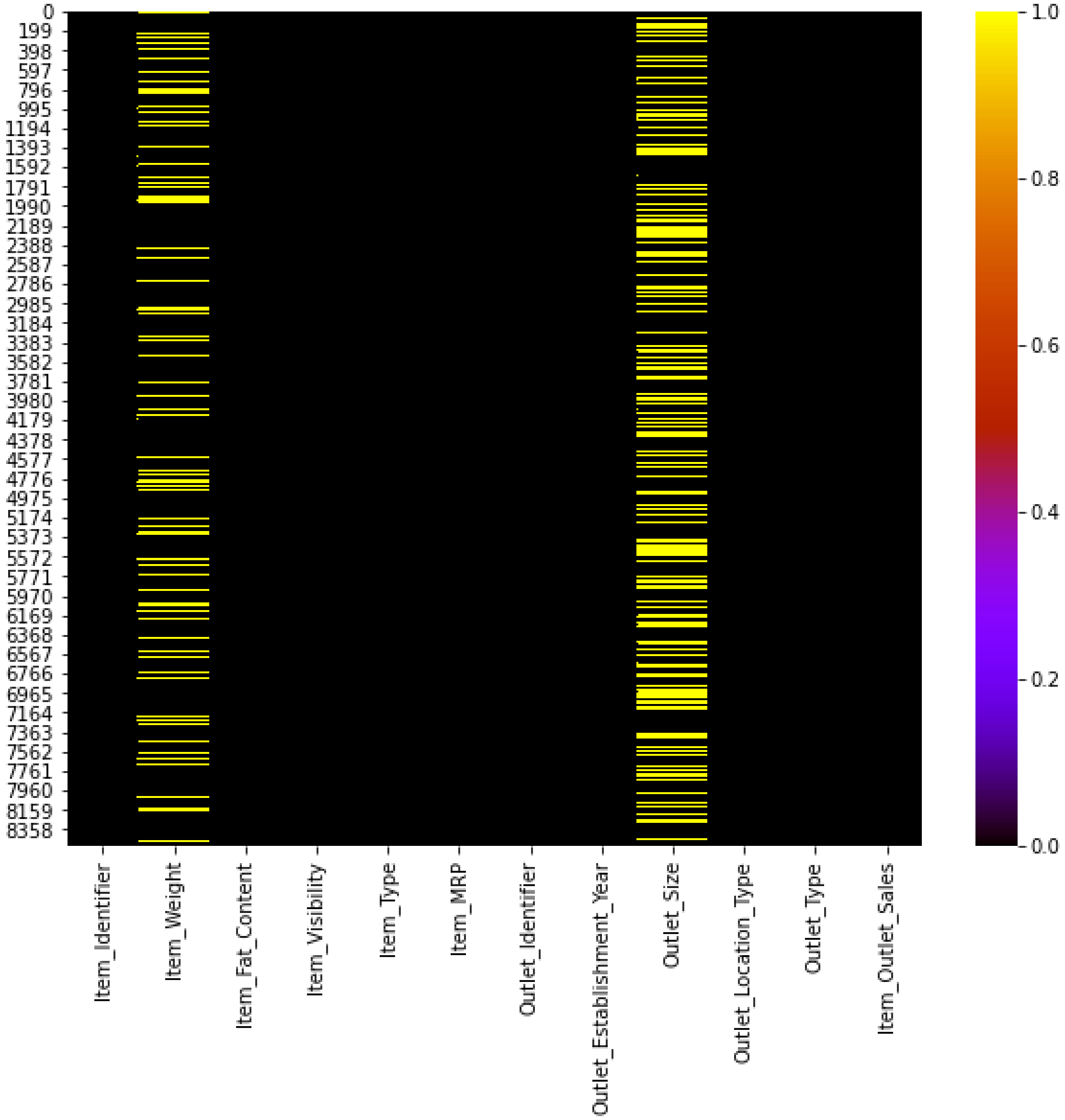
CHAPTER-3

RESULTS & OUTPUTS

3.1 Results - Graphs, Performance

A. Analyzing Null values using Heat Map

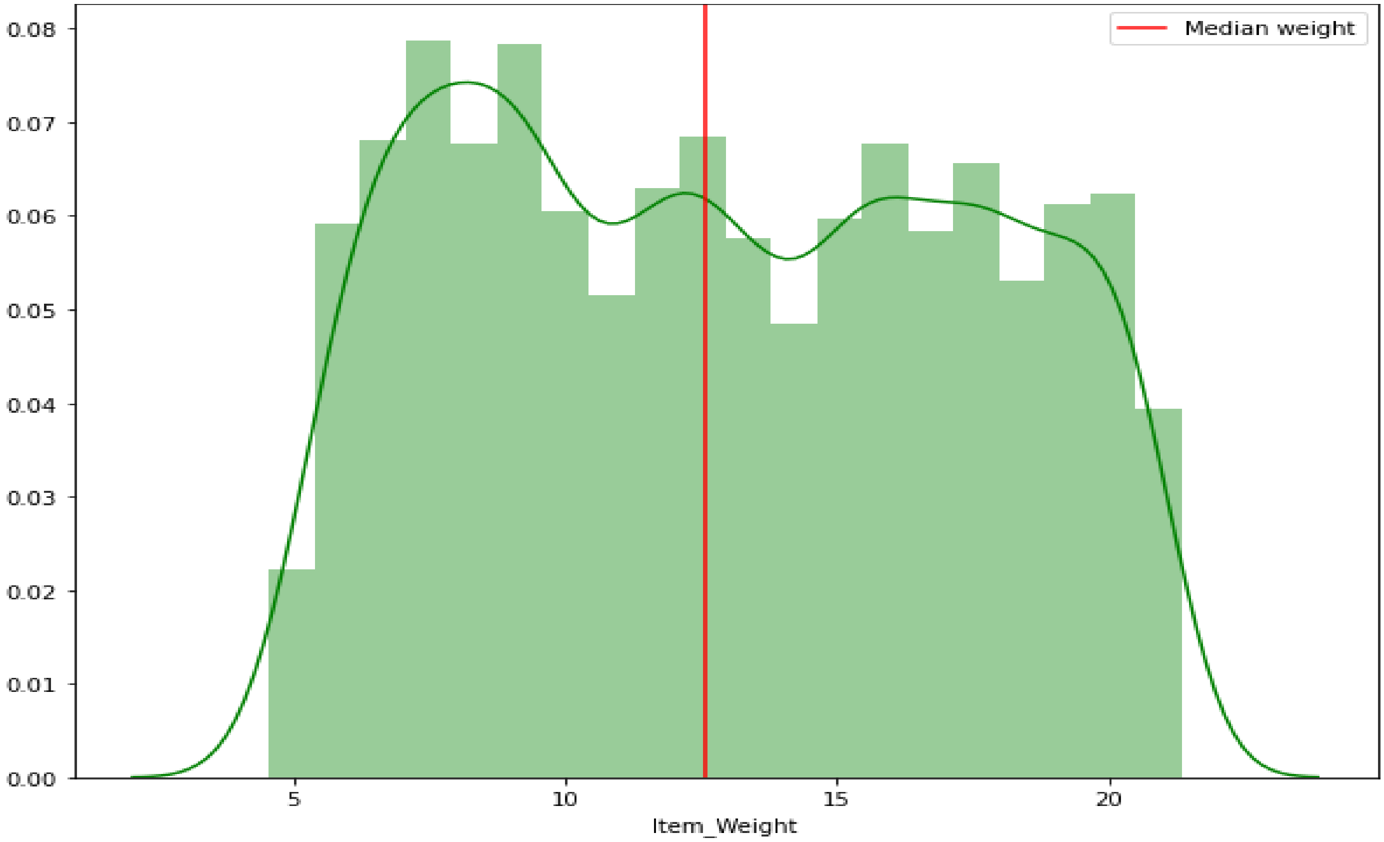
Null values are present in Item_Weight and Outlet_Size. Let us visualize it with a heatmap.



The yellow marks show the presence of null values at particular rows. There are quite a few null values in these two columns that will need to be sorted out.

B. Item_Weigth Distribution

The seaborn distplot below depicts the variation in the data distribution. It represents the overall distribution of continuous data variables. So, it represents the weight distribution of the items and the median weight is 12.60 from the graph.



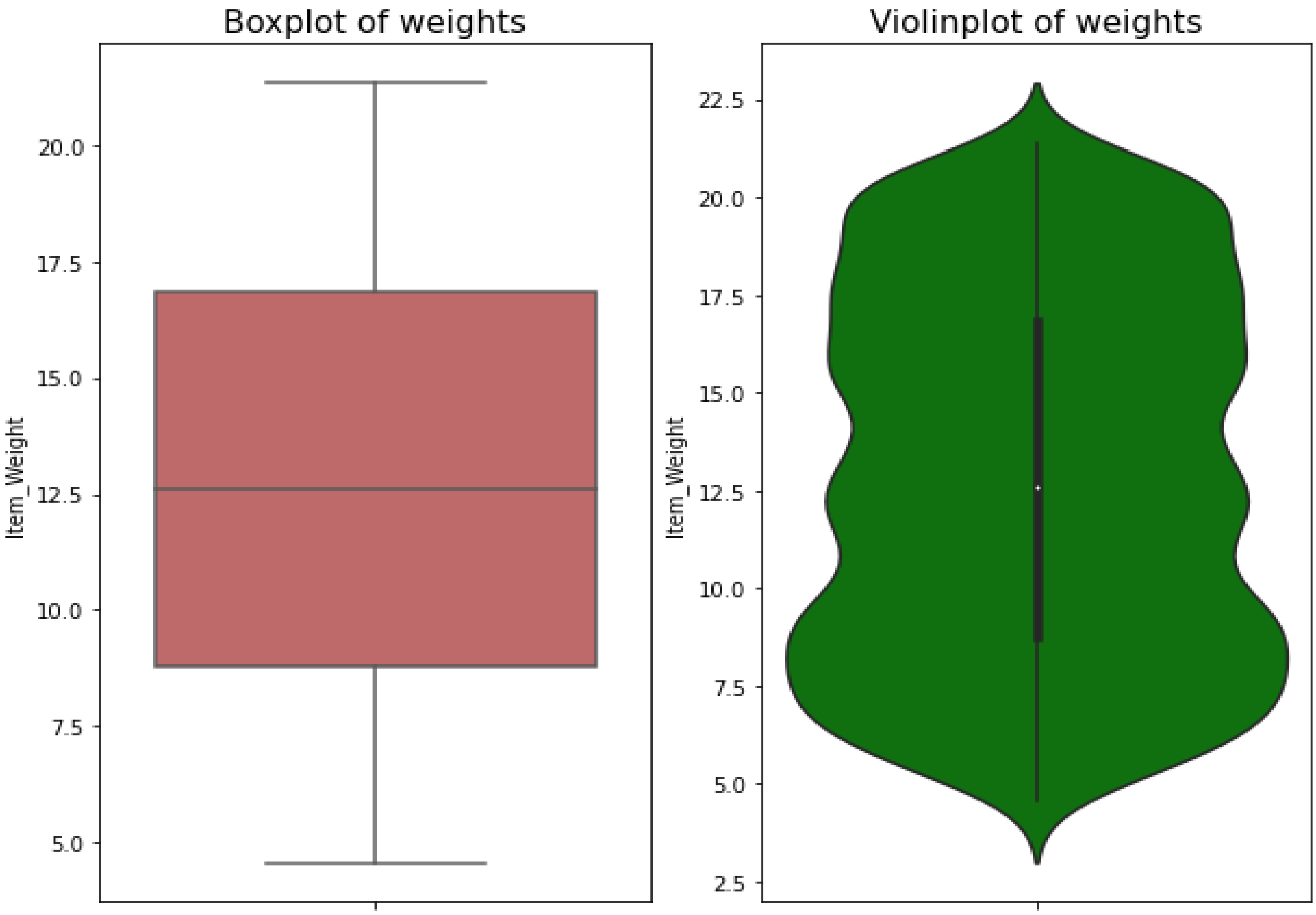
WEIGHT DISTRIBUTION OF ITEMS
Median Weight : 12.60

Statistical Analysis: Name: Item_Weight, dtype: float64

count	7060.000000
mean	12.857645
std	4.643456

min	4.555000
25%	8.773750
50%	12.600000
75%	16.850000
max	21.350000

Below are the violin and boxplot for item_weight. The curve plateaus over a large range of weights. Hence, it is simply not possible to assume a weight for the null values. So, they are left alone as it is or drop them if it is later deemed to not be too important in our analysis.

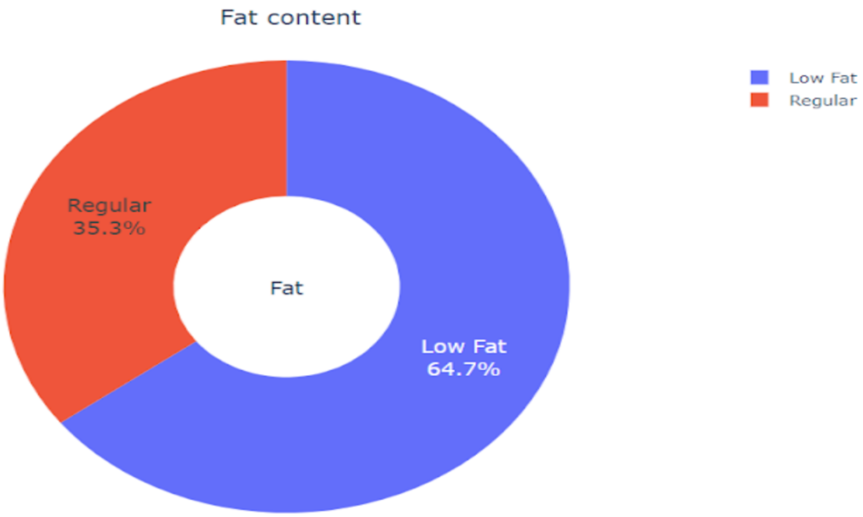


Boxplot of weights

Violinplot of weights

C. Item_Fat_Content

There are only two unique values in this column which are either low fat or regular. The percentage of each value is depicted in the pie chart below.



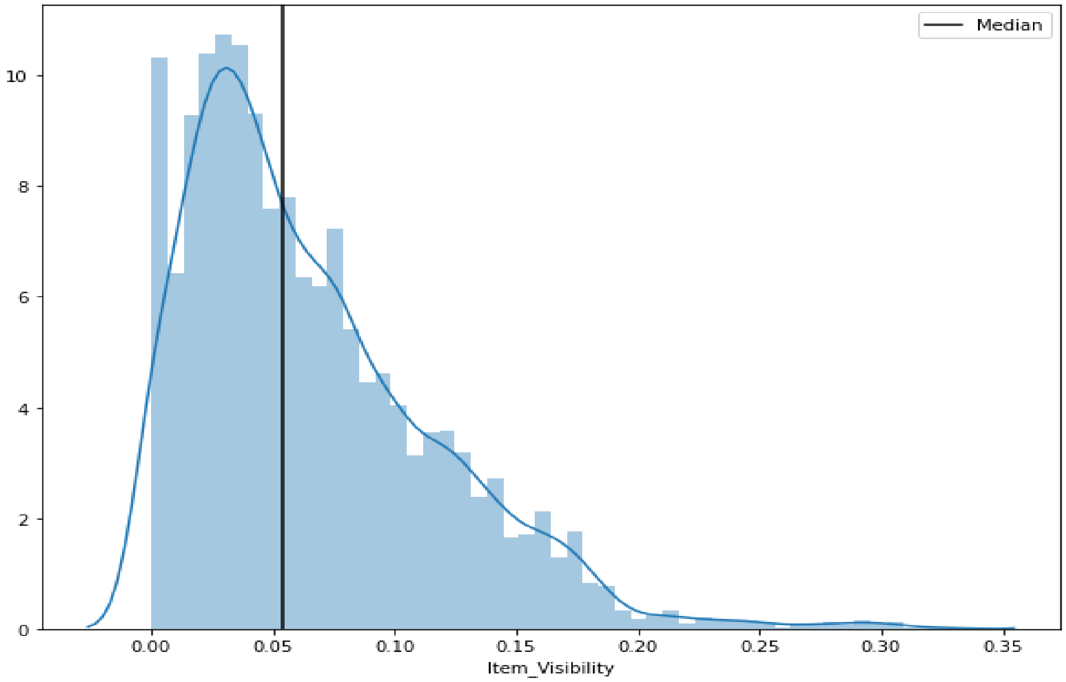
From the pie chart, 64.7 % items are low fat. This states that the majority of customers are health conscious and prefer food with lower fat rating.

D. Item_Visibility

Statistical Analysis: Item_Visibility, dtype: float64

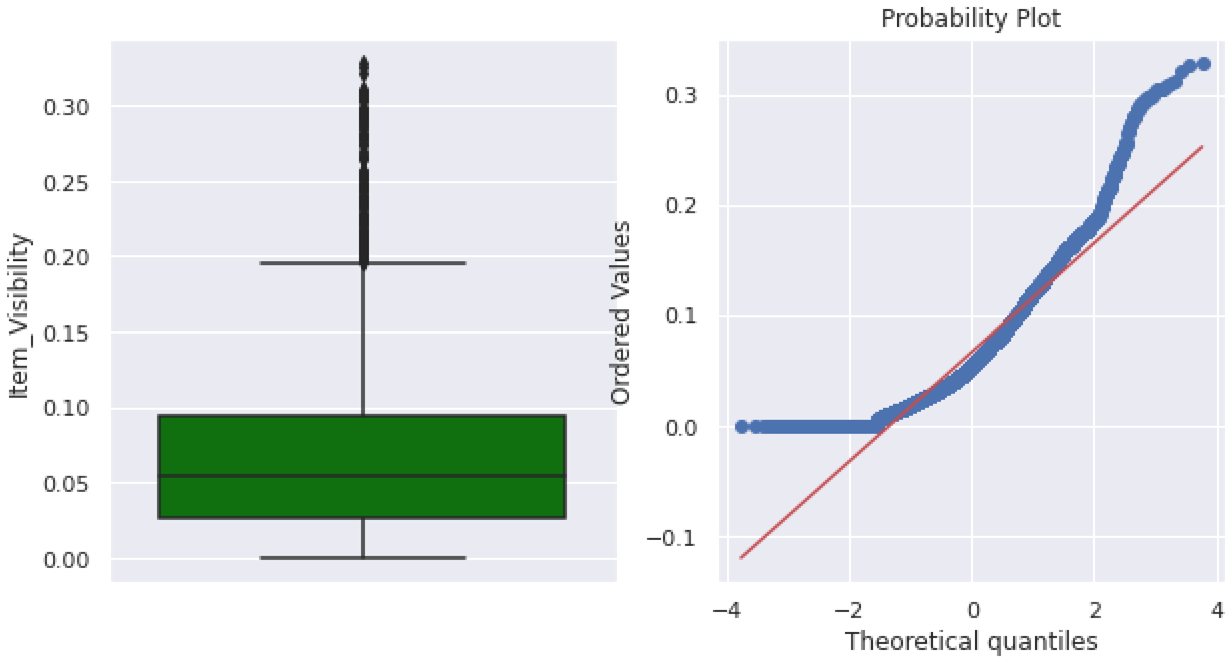
count	8523.000000
mean	0.066132
std	0.051598
min	0.000000
25%	0.026989
50%	0.053931
75%	0.094535
max	0.328391

The distplot of Item_Visibility Distribution is depicted in the graph below:

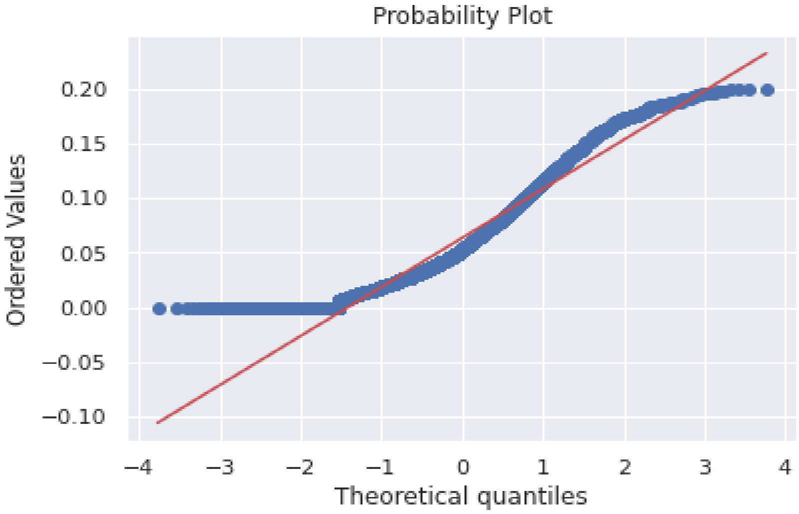


From the above graph, the median is 0.05. The curve is *right skewed*, so the *median* gives a better indicator than mean value.

To check any outliers in Item_Visibility from box plot and probability plot :



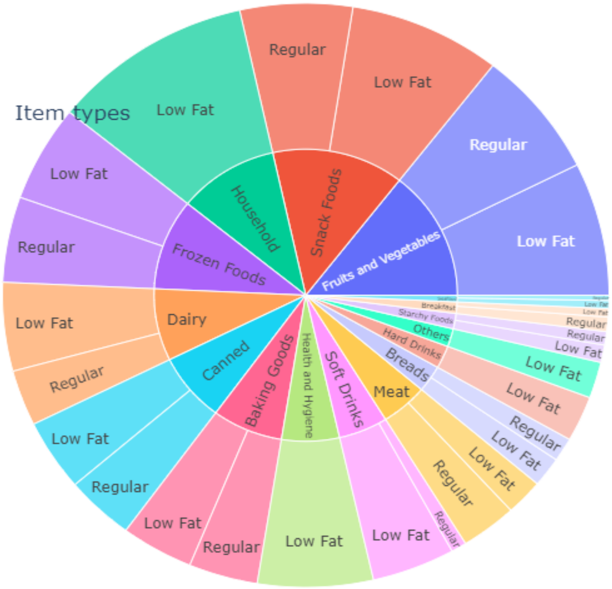
- From the above box plot, the range is 0 to 0.2 and there are some outliers present.
- Also, from the probability plot the values are deviating(blue line) from 0.2.
- The Presence of outliers doesn't bode well with machine learning models. So, the outliers need to be removed.
- From 8300 entries, there are approximately 134 outliers.

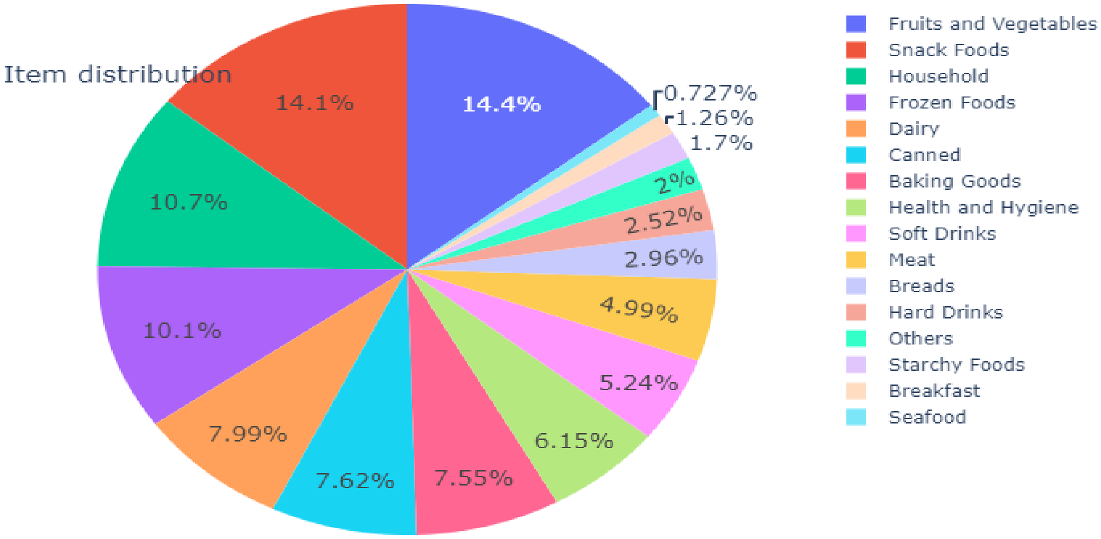


- The probability plot above represents after the outliers are removed from the item_visibilty.

E. Item_Type

Below are the pie plots depicting item_type and their distributions

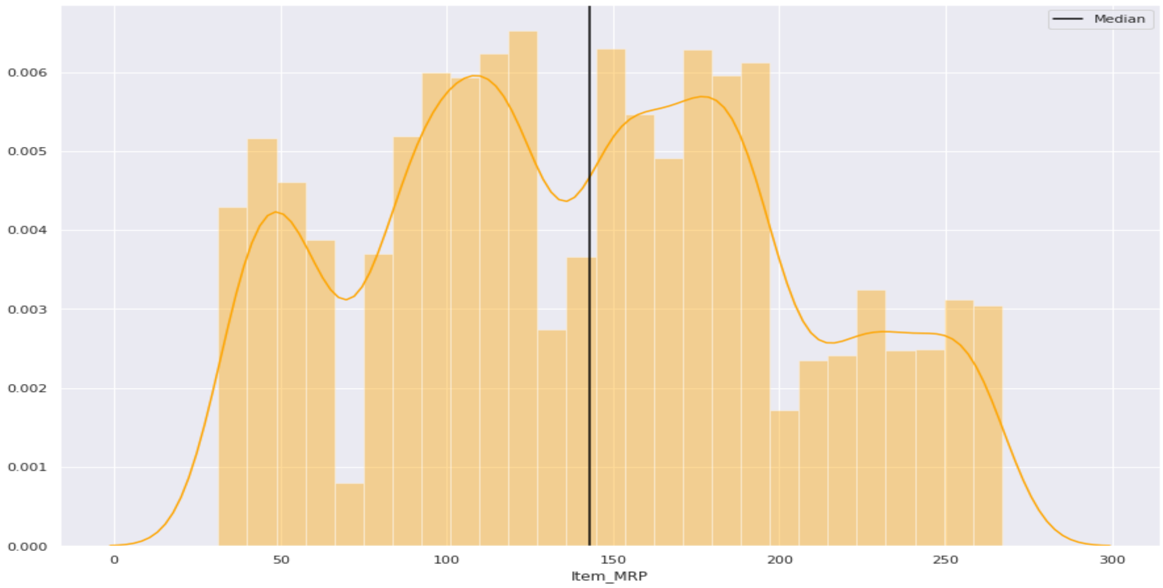




From the above plot, fruits and vegetables are the highest sold item followed by Snack foods.

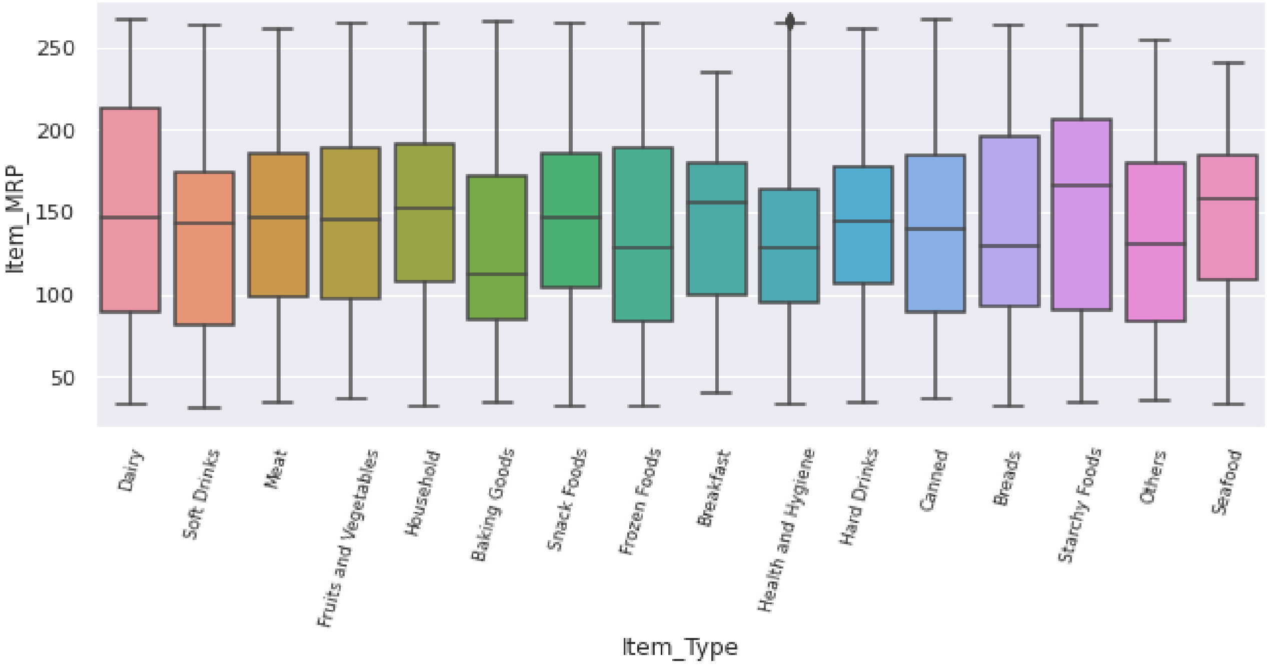
F. Item_MRP

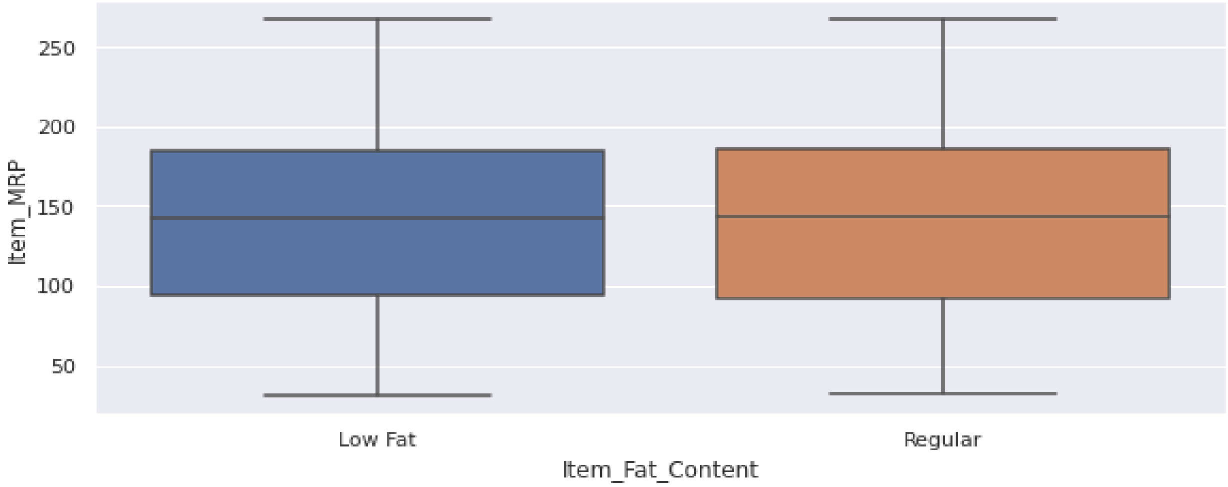
Below graph depicts the Item_MRP distribution in distplot.



From the above distplot, the median is 142.92.

Below is the box plot of item_MRPs with items.



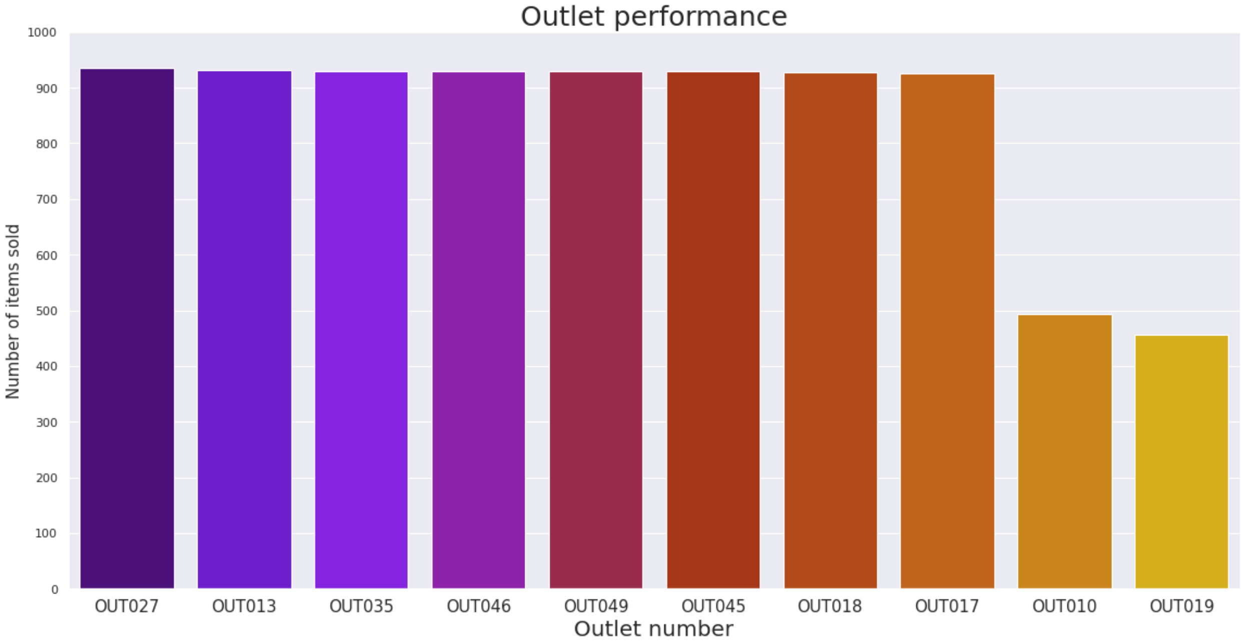


From the above plot, Dairy products and Starchy foods have a higher median price than the rest.

Both low and regular food have almost identical median prices.

G. Outlet_Identifier

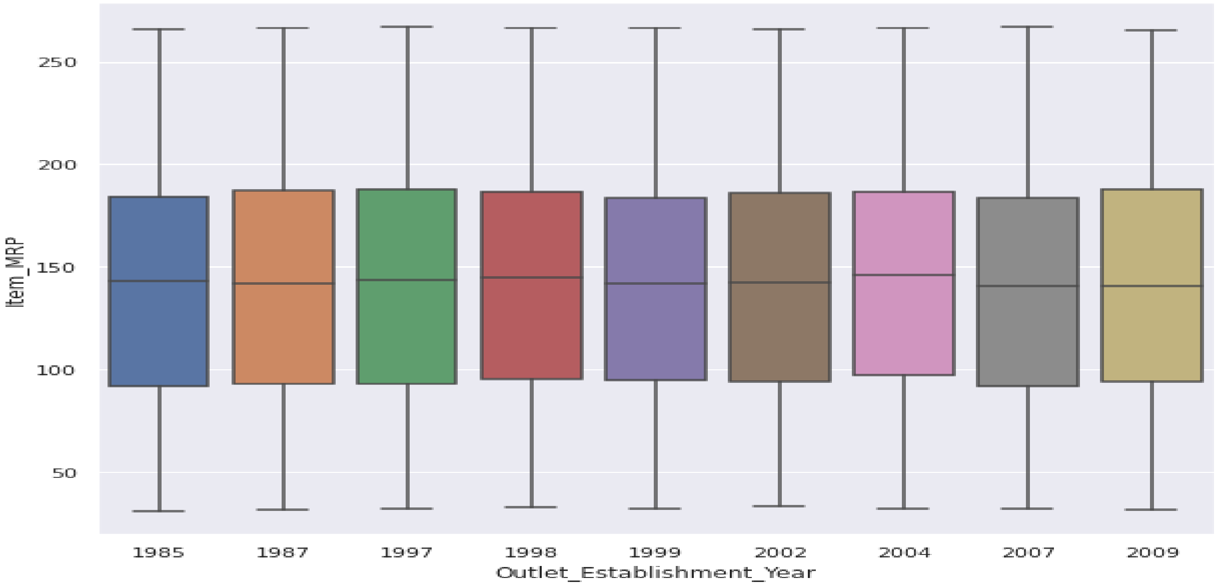
Below graph is the catplot of outlet number vs number of items sold



Most of the outlets have performed similarly with approximately 950 items sold. Whereas, outlet 10 and 19 are however lagging behind in sales.

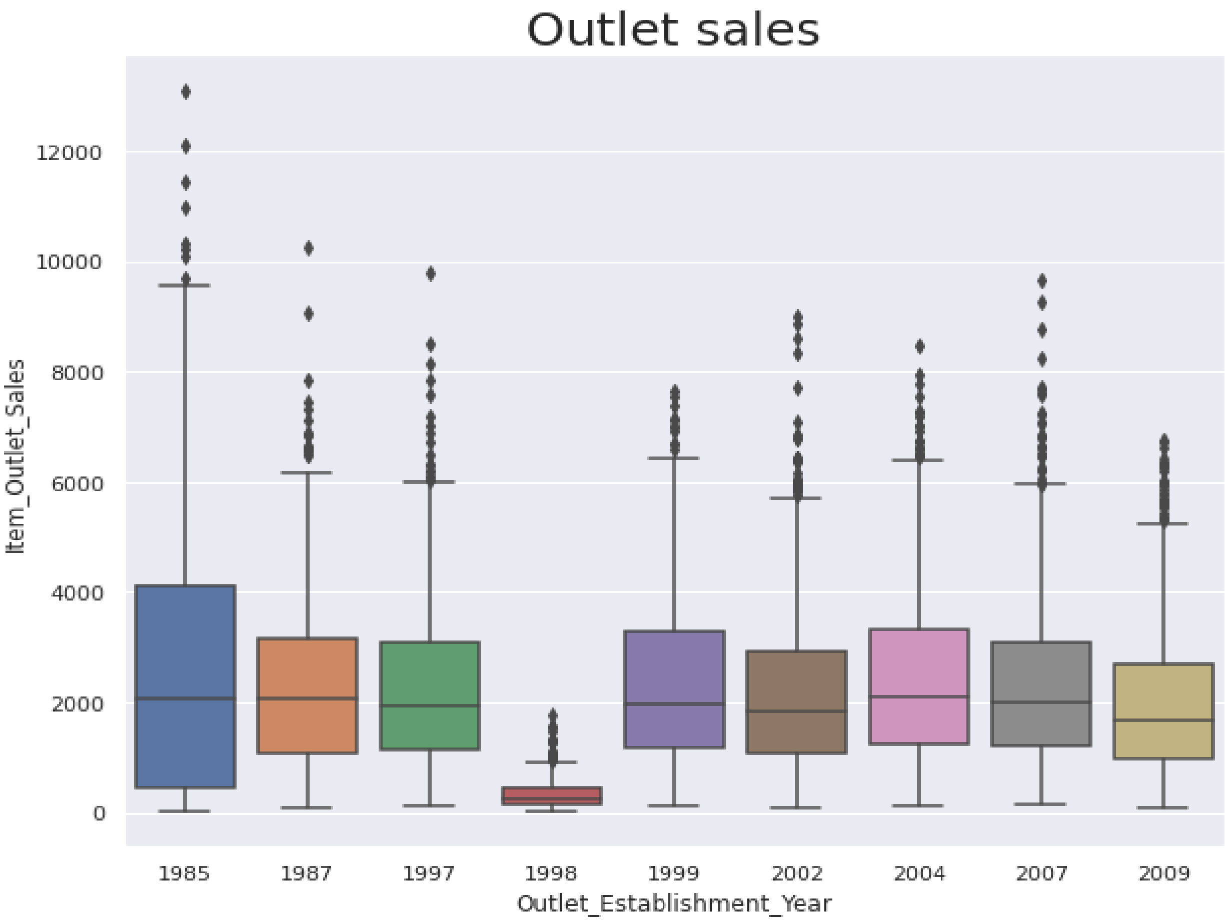
H. Outlet_Establishment_Year

Below plot is the box plot between outlet_Establishment_Year and Item_MRP.



From the above plot, no matter how old the shops are, the median prices of items sold is nearly the same. Hence, customers have no bias to buy more expensive products from older or newer markets.

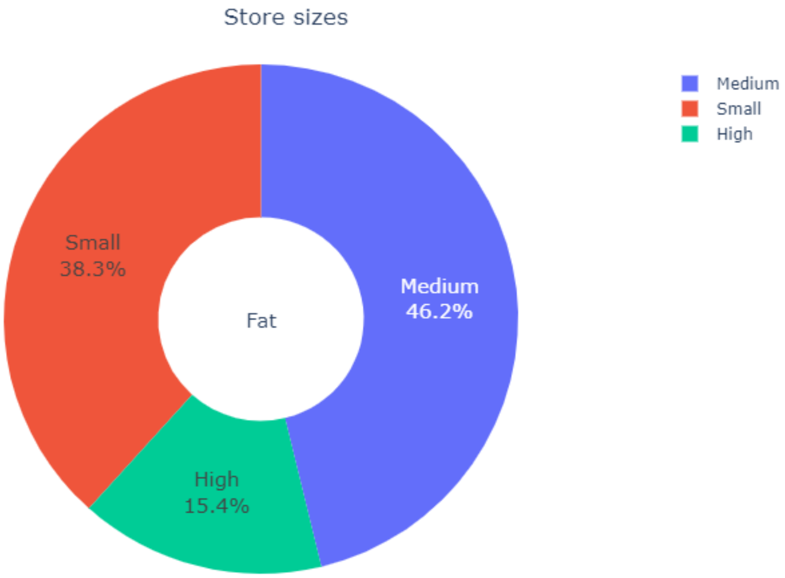
To check if the establishment year has anything to do with number of outlet sales:



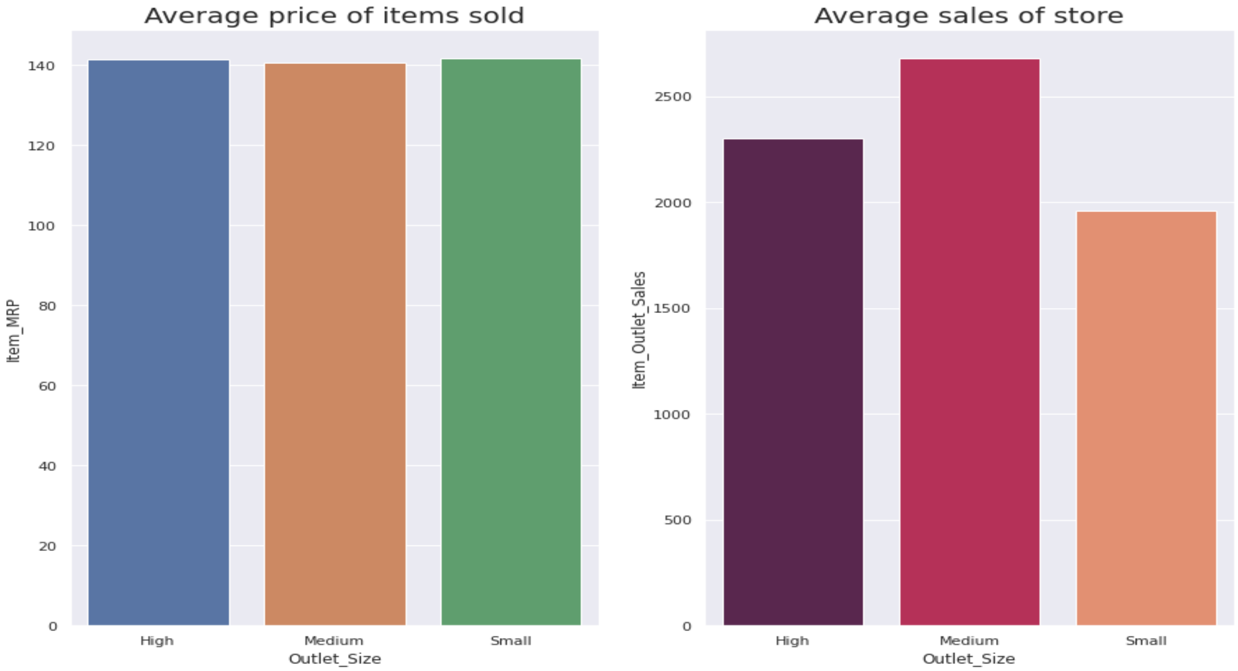
From above, it can be seen, the sales reported by the older stores are higher than the relatively newer stores (except for the 1998 established store).

I. Outlet_Size

This feature deals with how big is the size of the outlet store. This field also contains empty values.



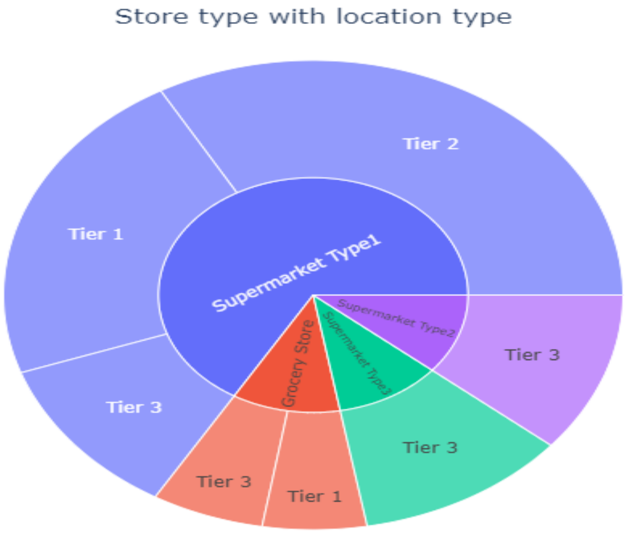
From the pie chart, maximum stores fall into the medium category followed by small. Only 15% of stores are high sized stores.



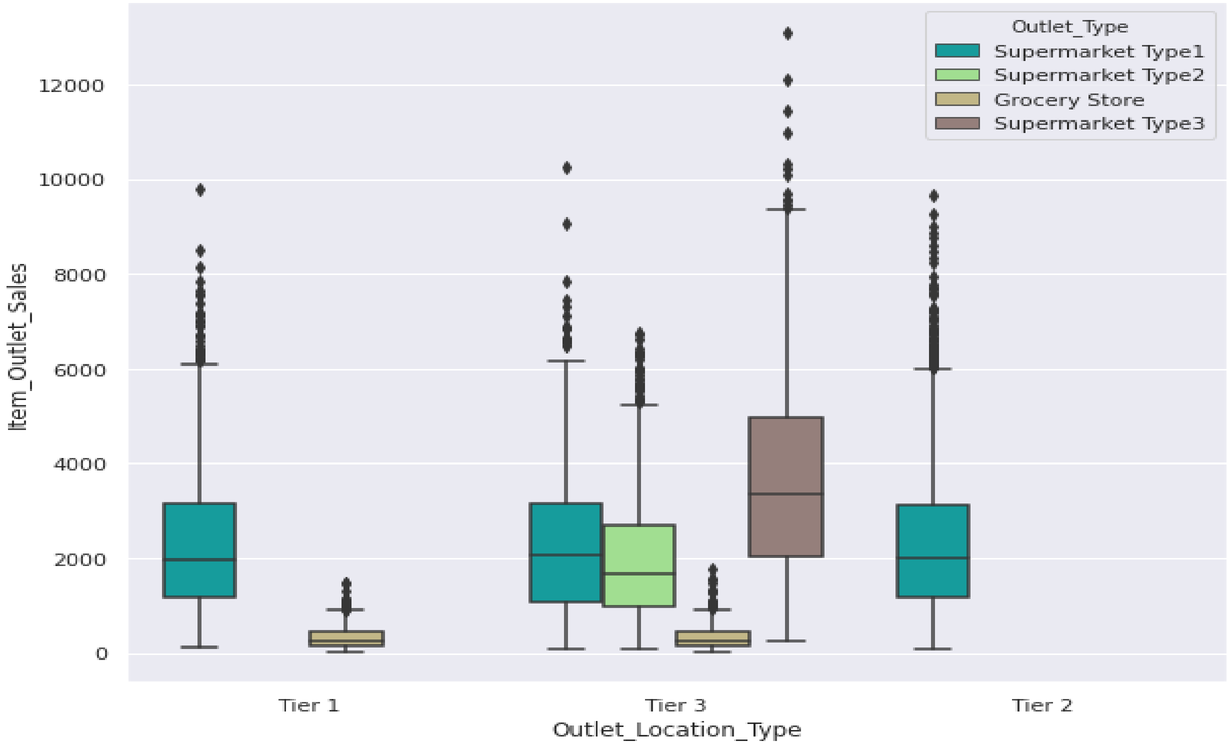
From the above graph, the average price of items sold in each outlet store size is nearly the same which is Rs 140. However, The medium stores seem to sell better followed by high sized and then small sized stores.

J. Outlet_Location_Type and Outlet_Type

As seen from below pie chart, majority of the stores are of type 1 supermarket distributed over various location tiers. Supermarket type 2 and 3 are confined to only tier 3 locations. Very small section of the stores are actually grocery stores.



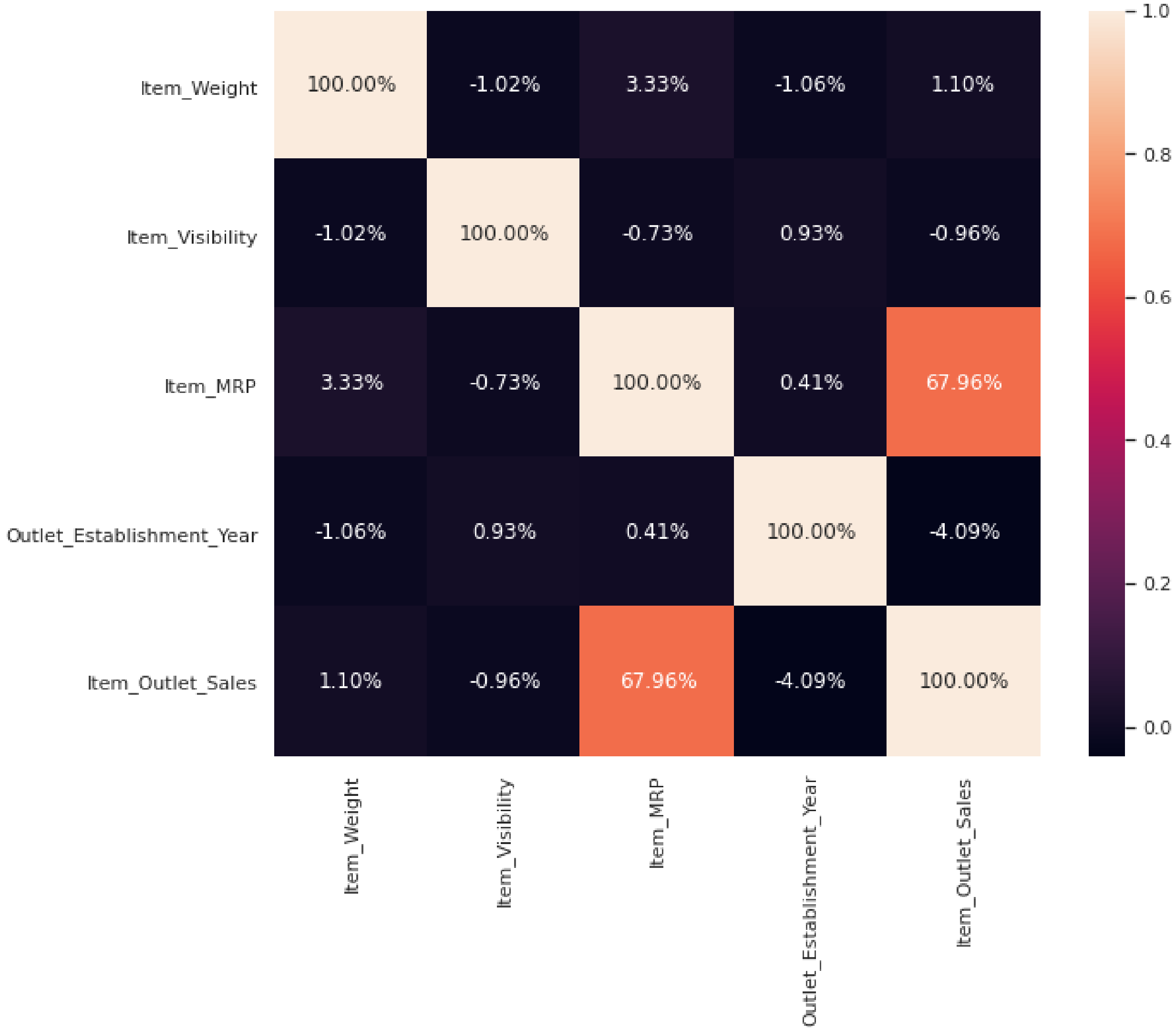
The below represented box plot between Item_Outlet_Sales and Outlet_Location_Type



From above, tier 3 locations seem to be selling better than both tier 2 and tier 1. It is also to be noted that tier 3 has more stores in it. Hence, the sales are better too

Correlation heatmap:

Now that each of the features are inspected individually, we will try to check the correlations of each term with the other.



From the above, the correlation of Item_Weight is extremely low. Hence, simply drop this column and get done with the issues of null values. Similarly, remove the order_size as there is no way to deal with the null values here as well. Also, get rid of the item_identifier and outlet_indetifier since it is of no consequence to us.

Machine learning Performance Metrics :

- **Root Mean Square Error :** It is the square root of the mean of the squared differences between actual outcomes and predicted outcomes.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{\left(\widehat{y_i} - y_i\right)^2}{n}}$$

- **Mean Absolute Error :** The mean or average of absolute values of the errors. And the absolute error is the magnitude of difference between the prediction of an observation and the true value of that observation

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

- **Mean Squared Error :** It is the sum, over all the data points, of the square of the difference between the predicted and actual target variables, divided by the number of data points.

$$MSE = \frac{1}{n} \Sigma \left(y - \widehat{y} \right)^2$$

Performance of machine Learning Models on data:

Performance Metrics	**Linear Regression**	**Random Forest**
RMSE	0.5041875773270632	0.3473443624488158
MAE	880.9999044084501	783.2320127534548
MSE	1162.4412631603454	1110.6987486230885

Inference:

The RMSE, MAE and MSE scores for Random Forest model are comparatively less so, Random Forest model can be used for the further sales prediction of this data.

CHAPTER-4

CONCLUSION & FUTURE WORK

The fundamentals of machine learning, as well as the accompanying data processing and modeling techniques, are discussed in this study, followed by their application to the problem of sales forecast in several shopping complexes. When the number of parameters employed is raised, accuracy, which is important in prediction-based systems, can be considerably improved. A look at how the sub-models work can also lead to a boost in system productivity.

The evaluation of the model is a critical component in developing a successful machine learning model. As a result, it's critical to build a model and acquire metrics recommendations from it. It will take time and effort until we attain good accuracy based on the results of metric improvements. The findings of one model are described by evaluation metrics.

A key element of the evaluation metrics is the capacity to distinguish between model results. For this examination, we employed the Root Mean Squared Error (RMSE) statistic. For regression situations, the RMSE is the most often used evaluation approach. Because of the square root's power, this statistic has a lot of fluctuation in percentages.The metrics squared feature tends to produce more stable results, avoiding the cancellation of positive or negative error values. In the further research process, K-NN, XGBoost Regressor are going to be used to train the model and find the RMSE scores for each method and find which method is even better than Random Forest.

CHAPTER-5

REFERENCES

[1] https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/

[2] https://towardsdatascience.com/random-forest-in-python-24d0893d51c0

[3] https://realpython.com/linear-regression-in-python/

[4] T. Alexander and D. Christopher, quot;An Ensemble Based Predictive Modeling in Forecasting Sales of Big Martquot;, International Journal of Scientific Research, vol. 5, no. 5, pp. 1-4, 2016. [Accessed 10 October 2019]

[5] M. Wistuba, N. Schilling and L. Schmidt-Thieme, quot;Hyperparameter Optimization Machines,quot; 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, 2016, pp. 41- 50.

[6] K. Punam, R. Pamula and P. K. Jain, quot;A Two-Level Statistical Model for Big Mart Sales Prediction,quot; 2018 International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, Uttar Pradesh, India, 2018, pp. 617-620

[7] C. M. Wu, P. Patil and S. Gunaseelan: Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data (2018).

[8] Das, P., Chaudhury, S.: Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data (2007)

[9] V. Shrivastava and P. Arya, quot;A study of various clustering algorithms on retail sales dataquot;, International Journal of Computing, Communications and Networking, vol. 1, no. 2, pp. 1-7, 2012. [Accessed 20 april 2022].

[10] A. Krishna, A. V, A. Aich and C. Hegde, quot;Salesforecasting of Retail Stores using Machine Learning Techniques,quot; 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India, 2018, pp. 160-166