Minor Project Report

on

**Big Mart Sales Prediction using Machine Learning**

Submitted in partial fulfillment of the requirements for the award of degree of

**Bachelor of Technology**

in

**Computer Engineering**

By

**Bhavini Raina**

**(19001003030)**

**Chahat Gupta**

**(20001003505)**

Under supervision of

**Dr. Payal Gulati**



**Department of Computer Engineering**

**J. C. BOSE UNIVERSITY OF SCIENCE & TECHNOLOGY, YMCA**

**FARIDABAD-121006**

**May 2020**

**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being carried out in this Minor Project titled **“Big Mart Sales Prediction using Machine Learning”** in fulfillment of the requirement for the degree of Bachelor of Technology in Computer Engineering and submitted to “**J. C. Bose** **University of Science and Technology, YMCA, Faridabad**”*,* is an authentic record of my own work carried out under the supervision of Dr. Payal Gulati

The work contained in this thesis has not been submitted to any other University or Institute for the award of any other degree or diploma by me.

Chahat Gupta

20001003505

**CERTIFICATE**

This is to certify that the work carried out in this project titled **“Big Mart Sales Prediction using Machine Learning”** submitted by **Chahat Gupta** to “**J. C. Bose** **University of Science and Technology, YMCA, Faridabad**” for the award of the degree of Bachelor of Technology in Computer Engineering is a record of bonafide work carried out by her under my supervision. In my opinion, the submitted report has reached the standards of fulfilling the requirements of the regulations to the degree

Dr. Payal Gulati

(Supervisor)

Assistant Professor,

Department of Computer Engg,

J. C. Bose University of Science and Technology, YMCA, Faridabad

Dr. Komal Kumar Bhatia

Chairman,

Department of Computer Engg,

J. C. Bose University of Science and Technology, YMCA, Faridabad

|  |  |  |
| --- | --- | --- |
| **S.NO TITLE PAGE NO.** | | |
| **1** | Introduction | 4 |
| **2** | General Description | 4 |
| **3** | Tool Used | 5 |
| **4** | Architecture | 6 |
| **5** | Data Description | 11 |
| **6** | Observation of data | 13-16 |
| **7** | Process Flow | 17-18 |
| **8** | Project Implementation | 18-21 |
| **9** | Conclusion | 21 |
| **10** | Future Scope | 22 |
| **11** | Reference | 23 |

## INTRODUCTION

]In today’s modern world, huge shopping centers such as big malls and marts are recording data related to sales of items or products with their various dependent or independent factors as an important step to be helpful in the prediction of future demands and inventory management. The dataset built with various dependent and independent variables is a composite form of item attributes, data gathered using a customer, and also data related to inventory management in a data warehouse. The data is thereafter refined to get accurate predictions and gather new as well as interesting results that shed new light on our knowledge concerning the task’s data. This can then further be used for forecasting future sales using employing machine learning algorithms such as random forests and simple or multiple linear regression models.

## GENERAL DESCRIPTION*: -*

### Product Perspective: -

The Store Sales Prediction is an ML-based Web Application that Can predict future product demand by analyzing records. It will give the number that will be the measure of product sales.

### Problem Statement: -

The data scientists at BigMart have collected 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 sales.

So, the idea is to find out the properties of a product, and store which impacts the sales of a product. Let’s think about some of the analyses that can be done

### Proposed Solution: -

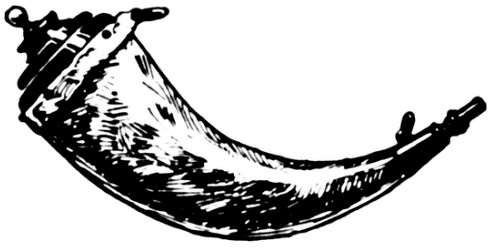
We will use performed EDA to find the important relation between different attributes and will use a machine-learning algorithm to predict the future sales demand. The client will be filled the required feature as input and will get results through the web application. The system will get features and it will be passed into the backend where the features will be validated and pre-processed and then it will be passed to a hyperparameter tuned machine learning model to predict the outcome.

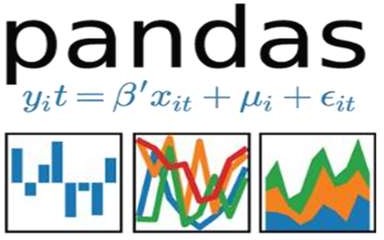
### Data Requirements: -

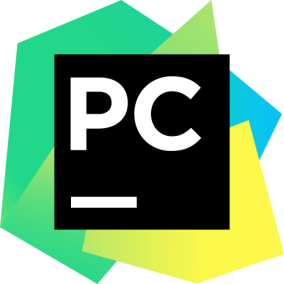
The data is required for the building of the project is already available on the dashboard. The Store Sales Prediction data is recorded many product descriptions along with past sales quantity. For building the ml model we will use the dataset that is given. The data consists of 8523 rows and various information about products like product id, product category, store id, store location, etc.

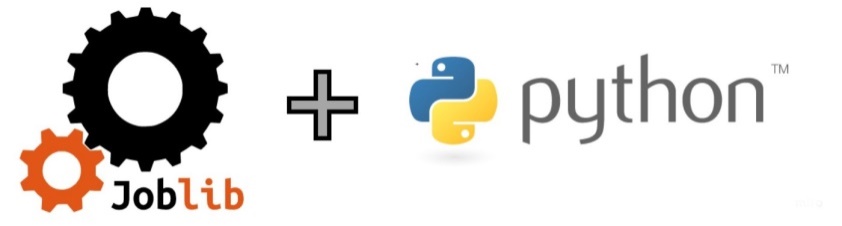
## TOOL USED: -

The programming language is Python is used here, also we will use some other python- based libraries like, for ml, we will use Scikit-Learn library, for data manipulation we will use pandas, for custom APIs creation Flask web frameworks. Visual Studio Code is used as python IDE for all modular coding and custom APIs creation. And storing all code files for publically available we will use GitHub. For the deployment process, we will be using Heroku cloud platforms for hosting our application.





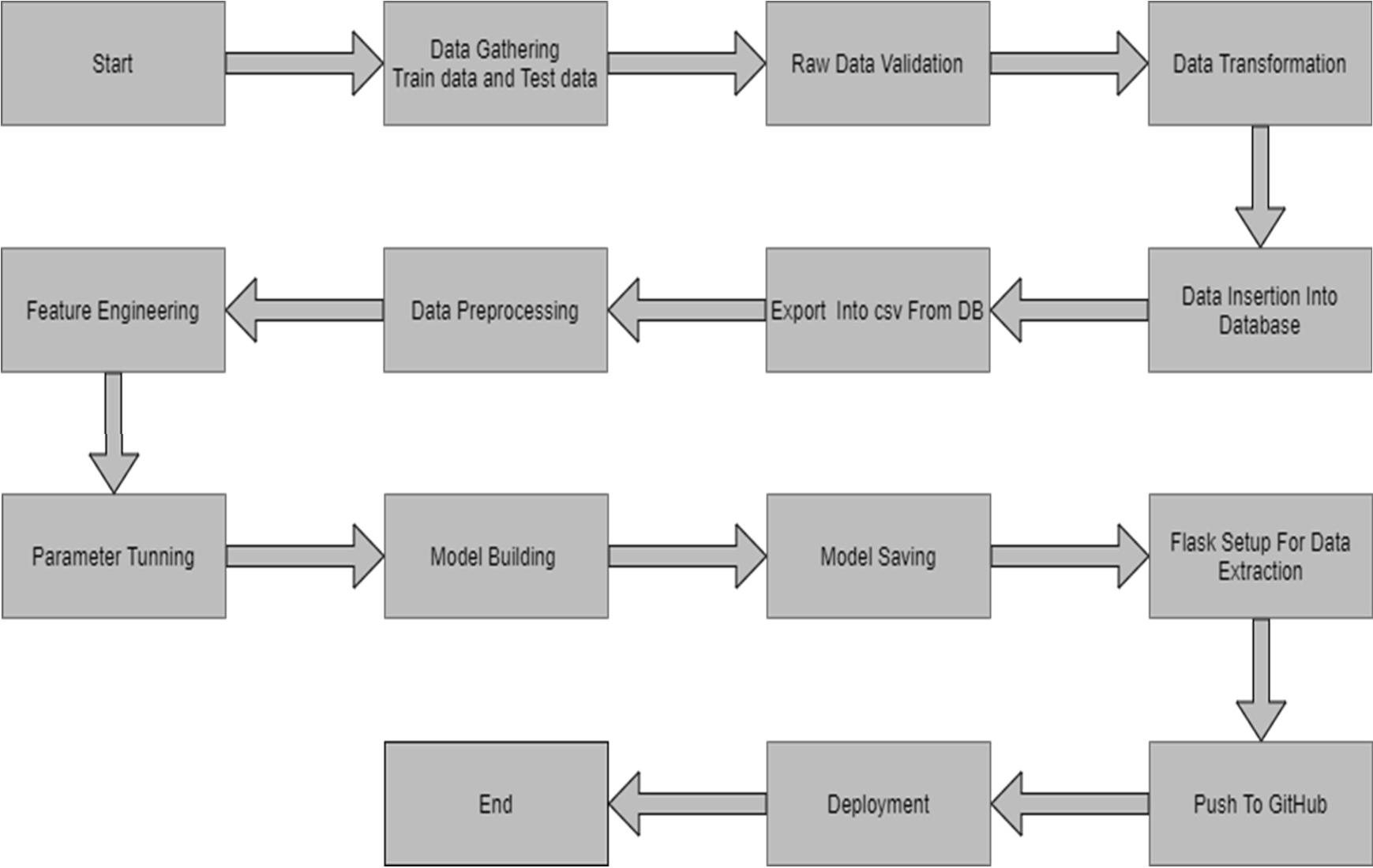




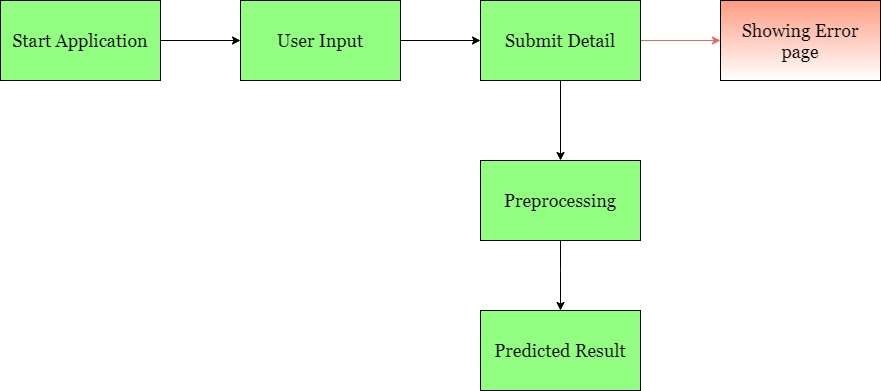
## ARCHITECTURE: -

We will be using the following architecture for this project

### Model Training and Evaluation: -



* + **Deployment architecture: -**



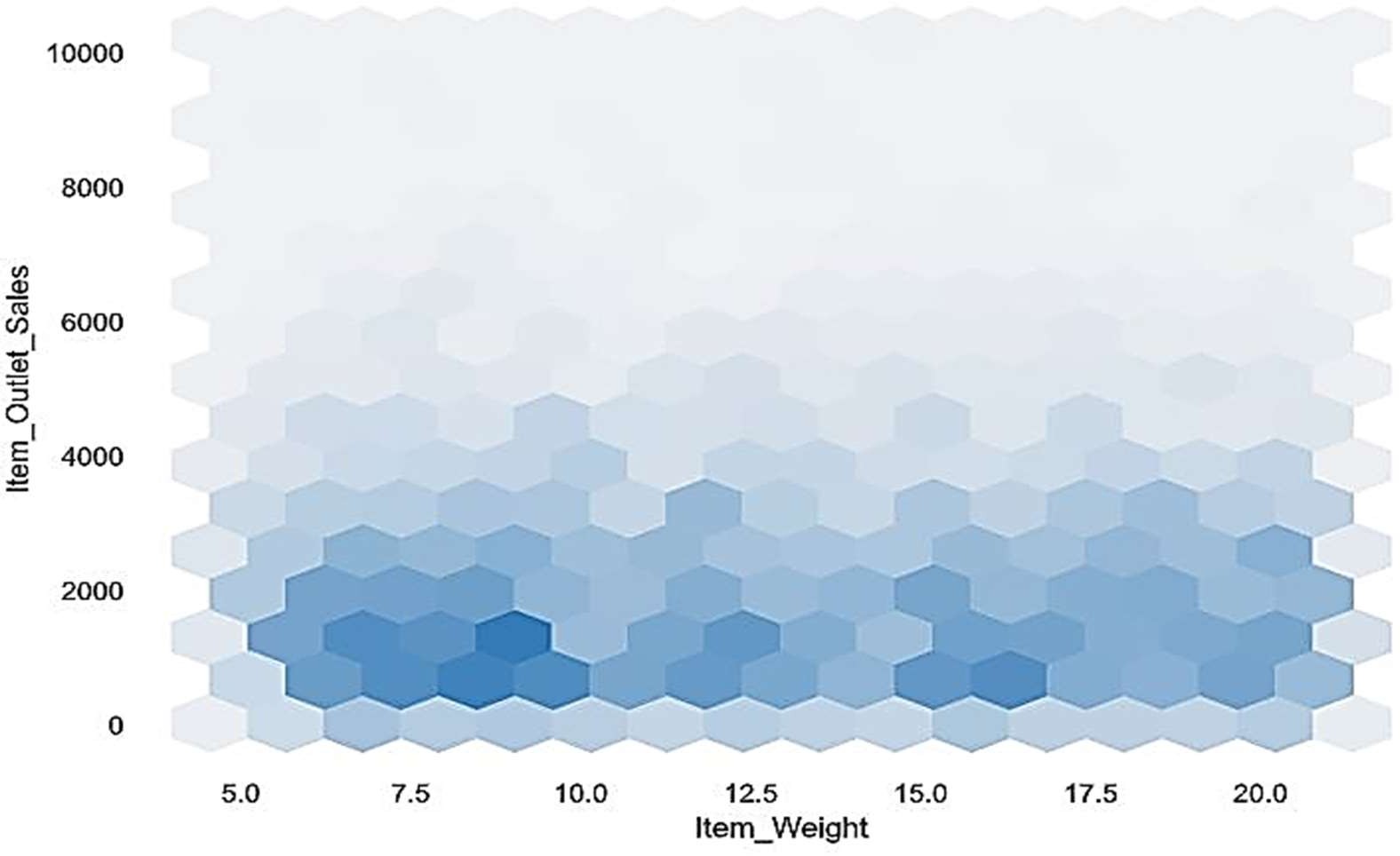
* **DATA DESCRIPTION: -**



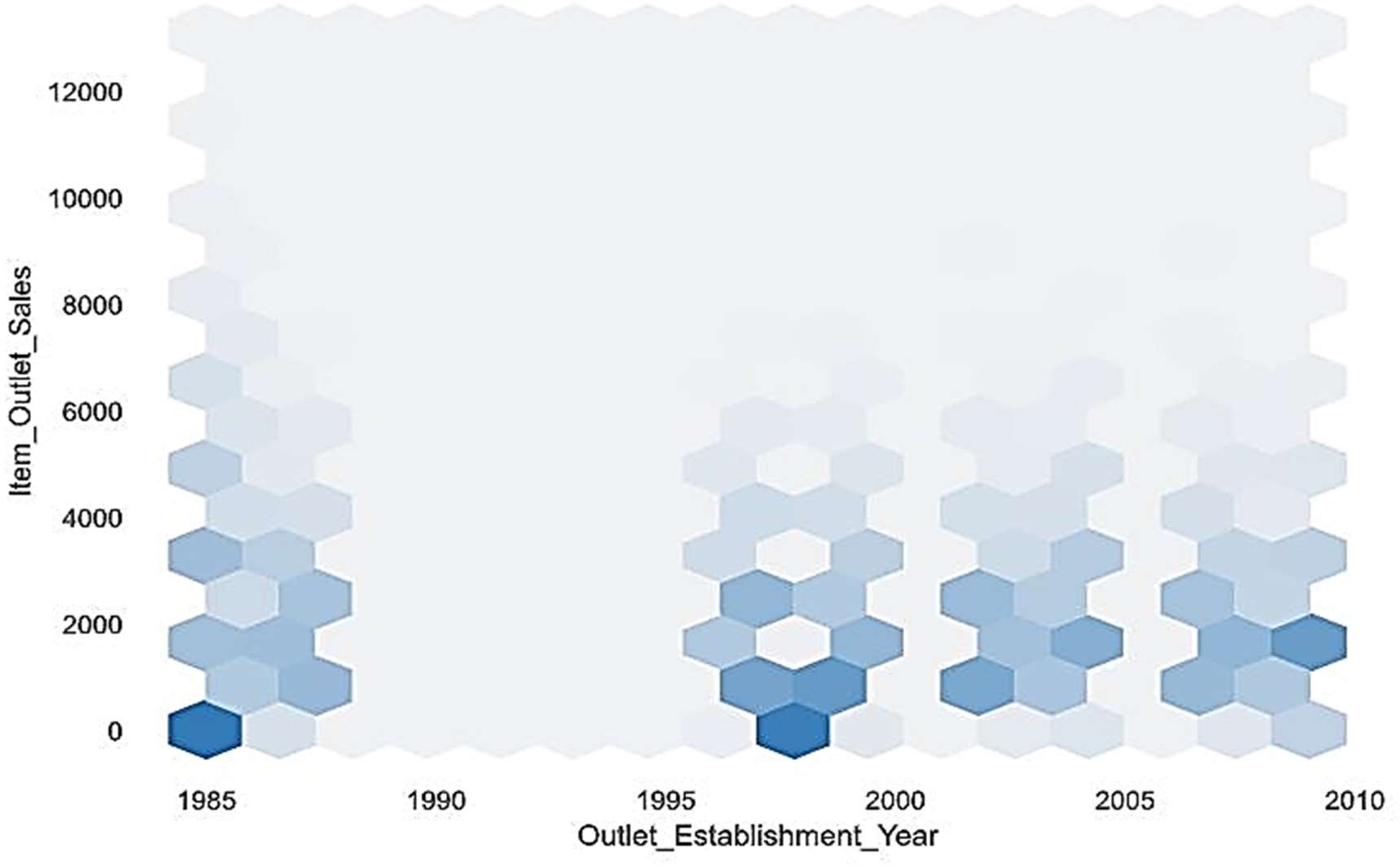
* **OBSERVATION OF DATA: -**

Correlation is used to understand the relation between a target variable and predictors. In this work, Item-Sales is the target variable, and its correlation with other variables is observed.

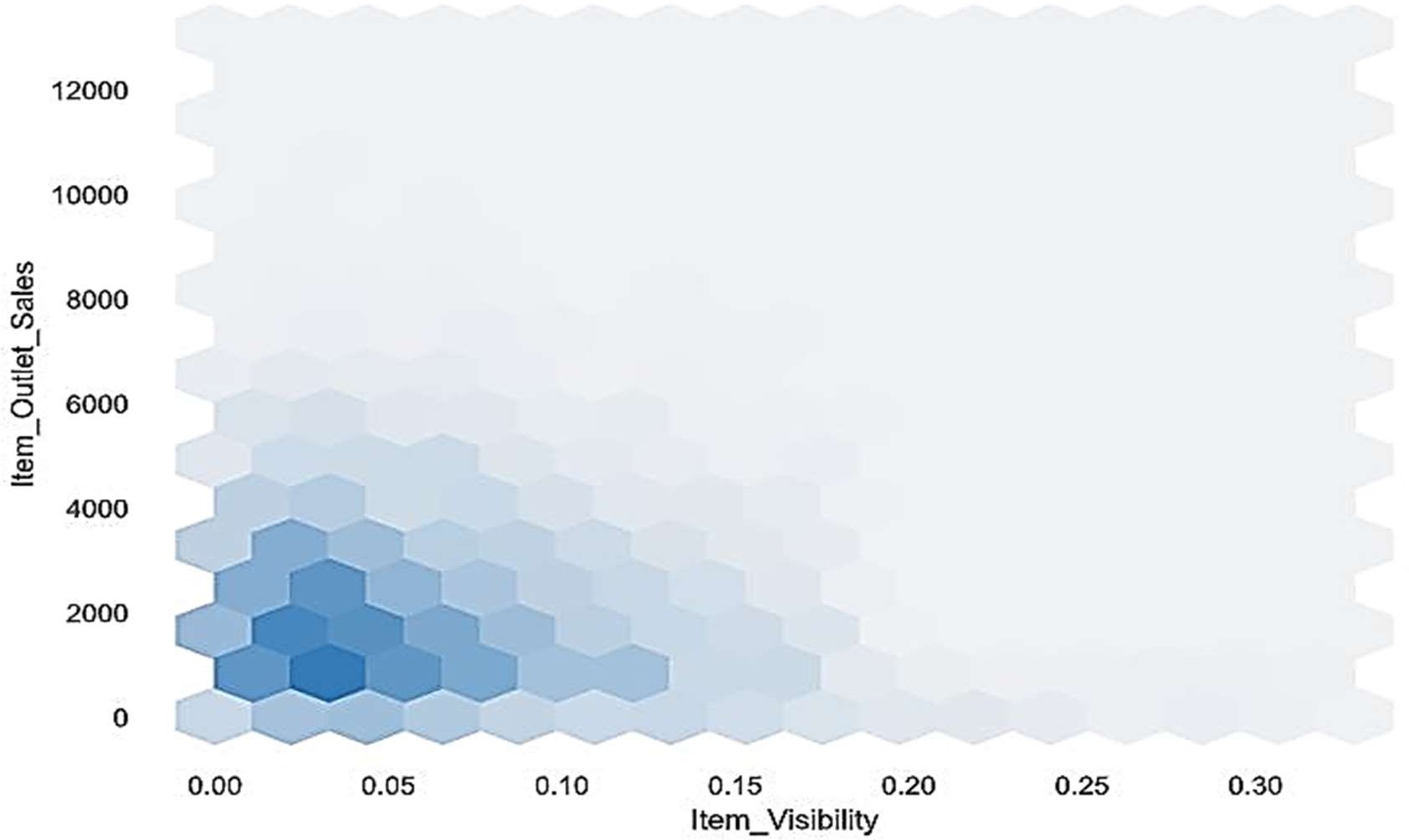
Considering case 1 of Item-Weight, the feature item weight is shown to have a low correlation with the target variable Item-Outlet-Sales in below Fig.



Considering case 2 There is no significant relationship found between the year of store establishment and the sales for the items. Values can also be combined into variables that classify them into periods and give meaningful results.



Considering case 3 The place where an item is placed in a store, referred to as Item visibility affects the sales. However, the plot chart shows that the flow is on the opposite side. One of the reasons might be that daily used products don’t need high visibility.

However, there is an issue that some products have zero visibility, which is quite impossible.

### Correlation: -

* Item visibility is having nearly zero correlation with our dependent variable Item\_Outlet\_Sales and grocery store outlet type. This means that the sales are not affected by the visibility of items which is a contradiction to the general assumption of “more visibility thus, more sales”.
* Item\_MRP (maximum retail price) is positively correlated with sales at an outlet, which indicates that the price quoted by an outlet plays an important factor in sales.
* Variation in MRP quoted by various outlets depends on their sales

## PROCESS FLOW: -

### Data Gathering: -

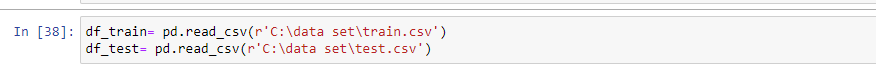
Data source: https://[www.kaggle.com/brijbhushannanda1979/bigmart-sales-data](http://www.kaggle.com/brijbhushannanda1979/bigmart-sales-data) Train and Test data are stored in .csv format.

### Data Exploration: -

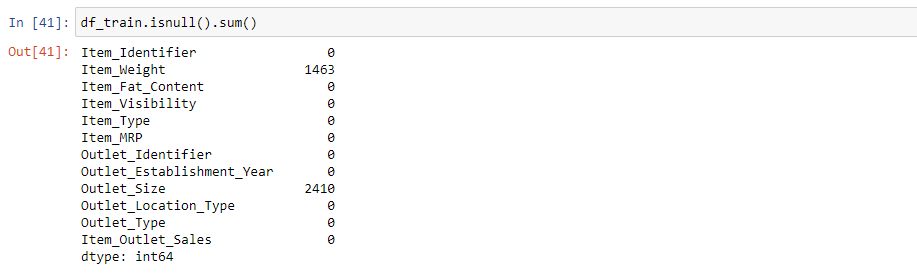
Let’s start by loading the required libraries

**Import Libraries**

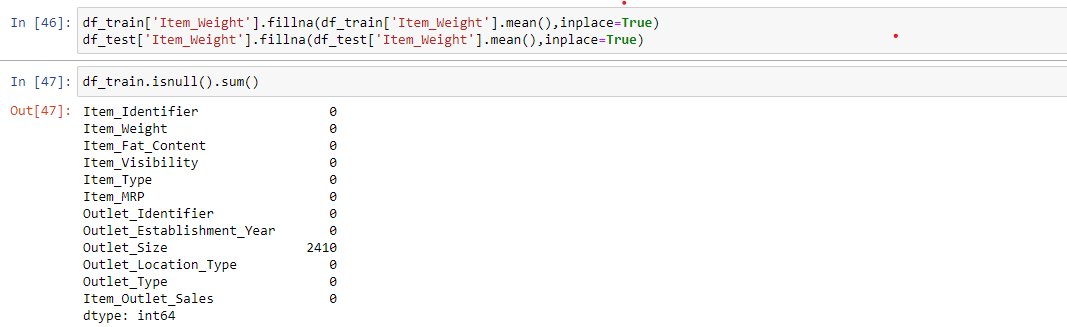
**Import the data set**

****

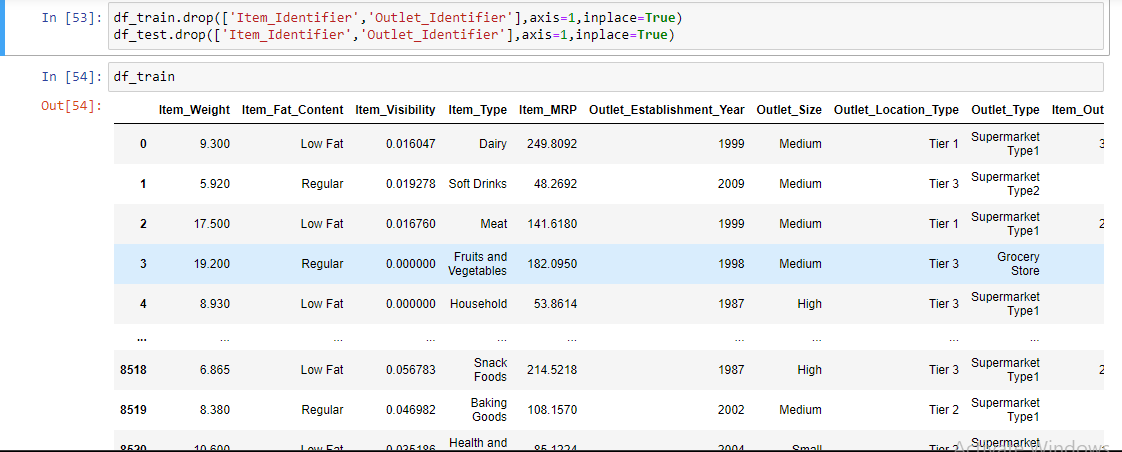
**Finding the null values**

****

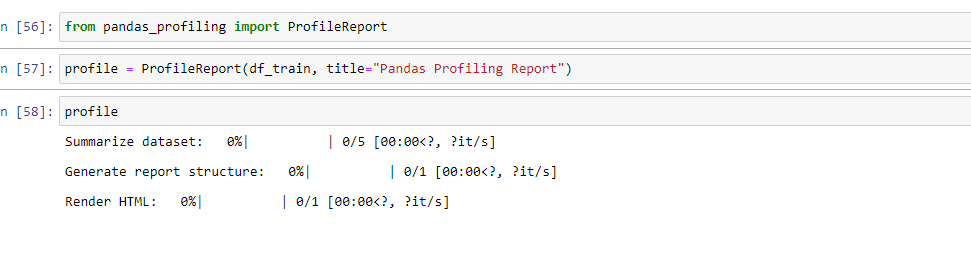
**Fill the null values with the mean values**

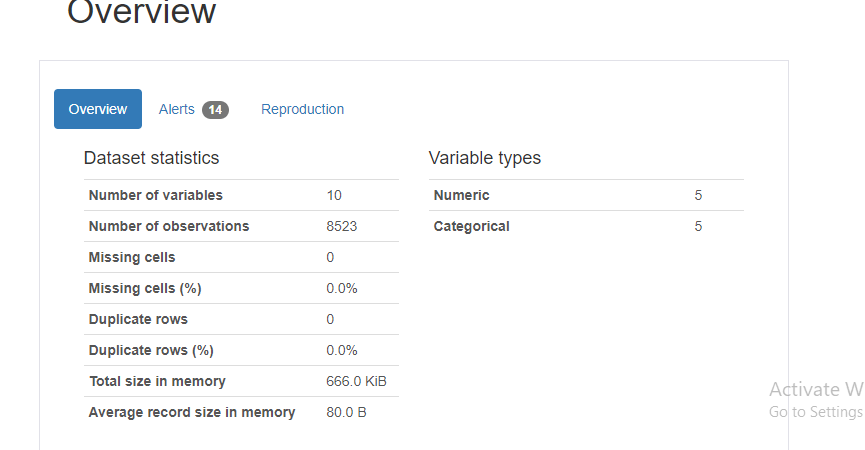
****

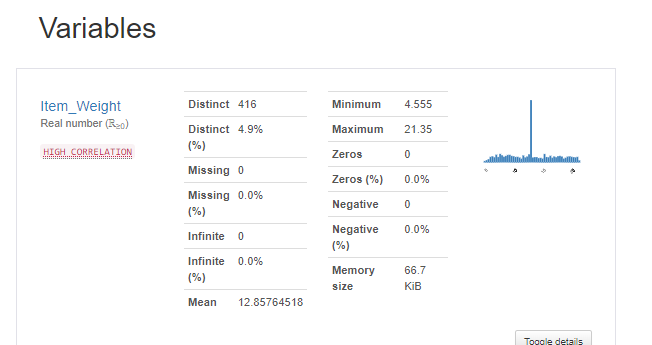
**Drop the unnecessary columns**

****

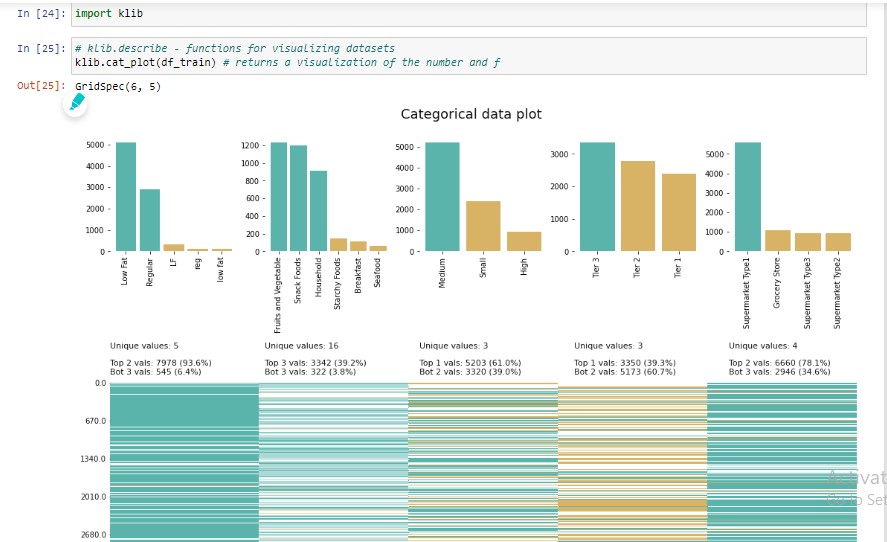
**Applying the pandas profiling**

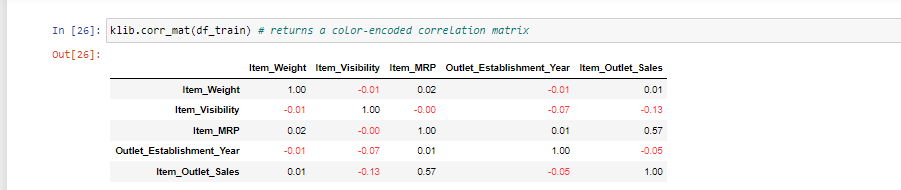
****

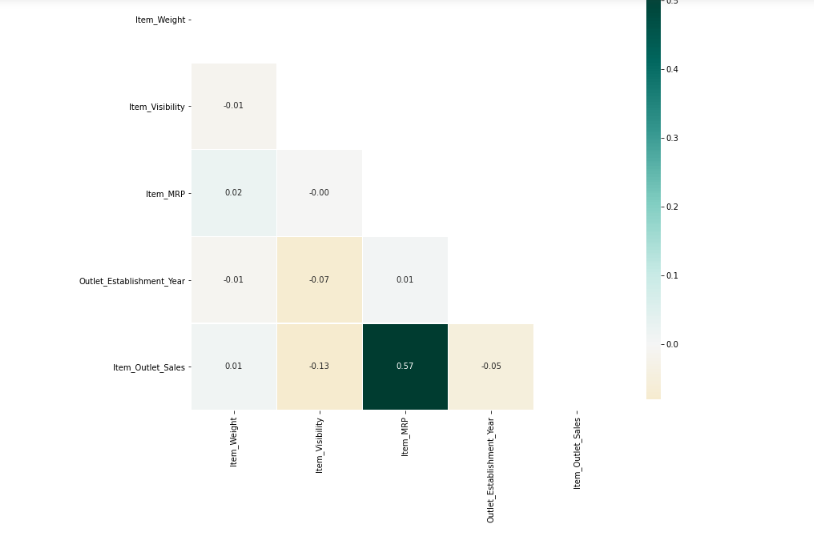
**The report generation of pandas profiling**



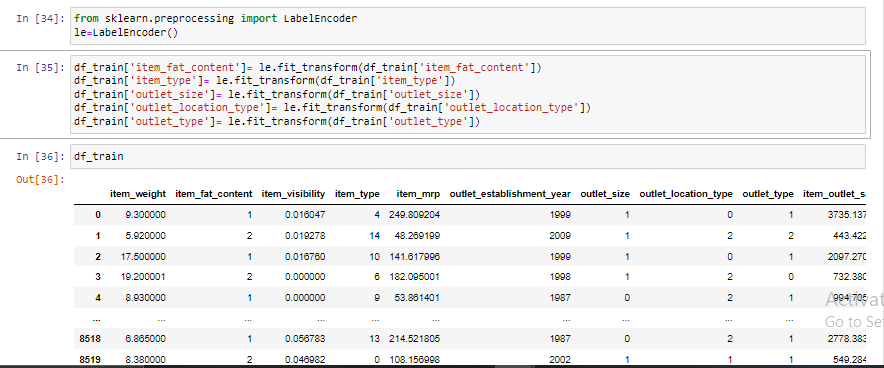
**Importing the klib lib for graph plotting**

****

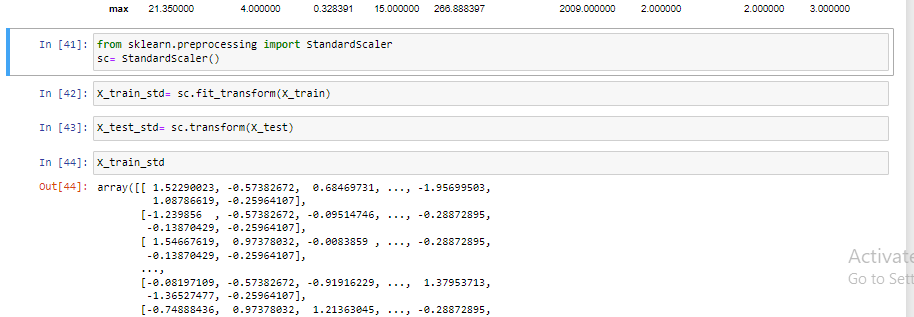
****



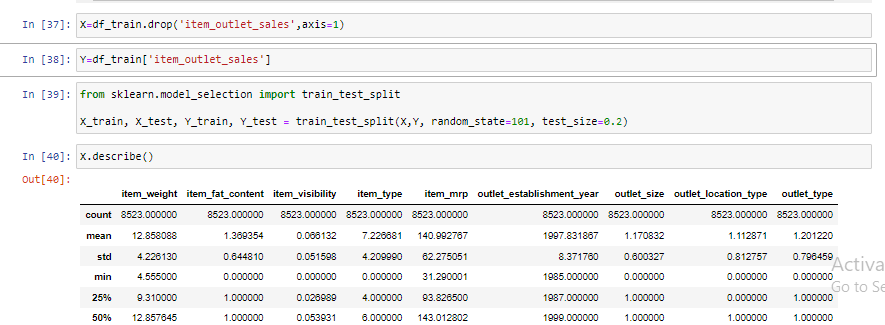
**Importing label encoder model**

****

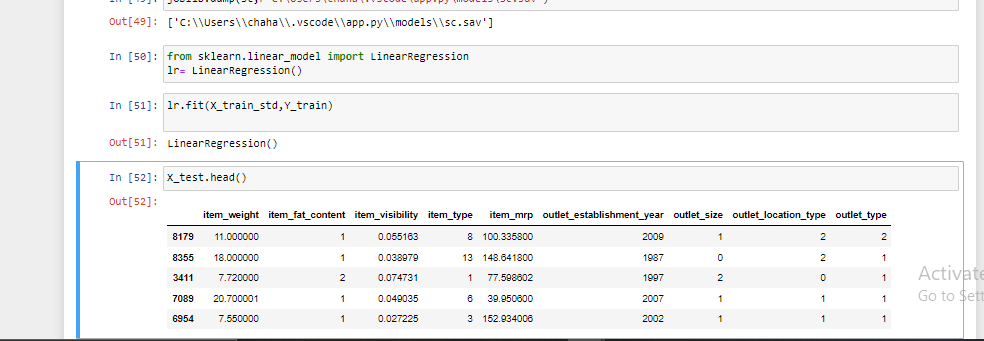
**Importing the standard scaler model**

****

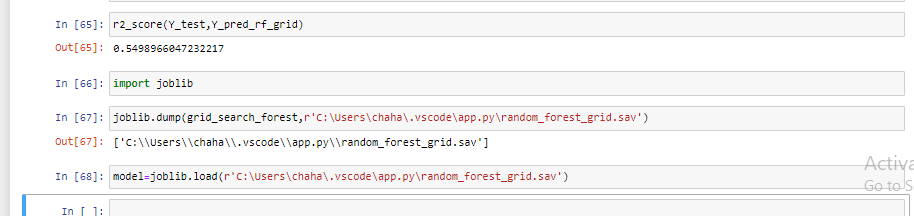
**Split the data into training and testing data**

****

**Appling machine learning model to the data set**

****

**Calculating the r square value of the model**

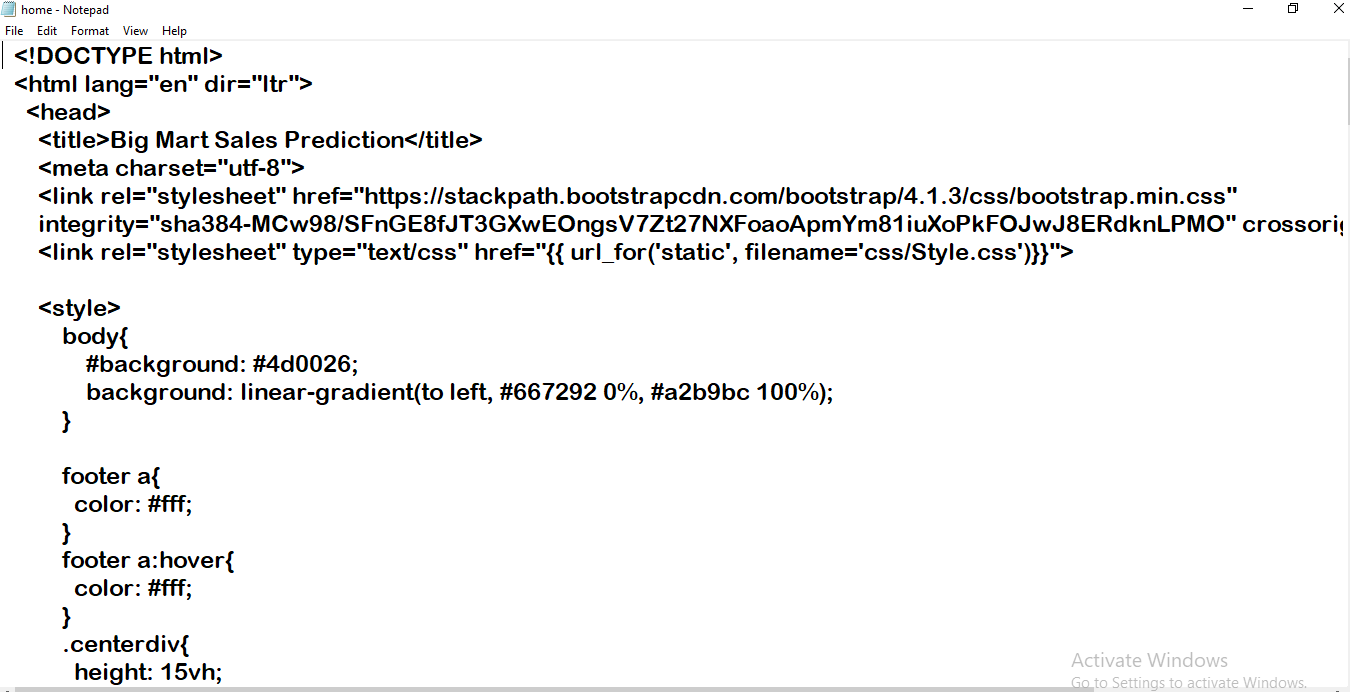
****

**Saving the model to the machine by using joblib**

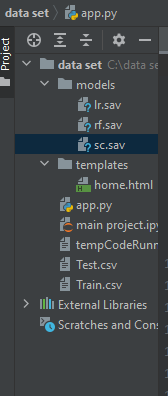
**Saving the model to the computer**

****

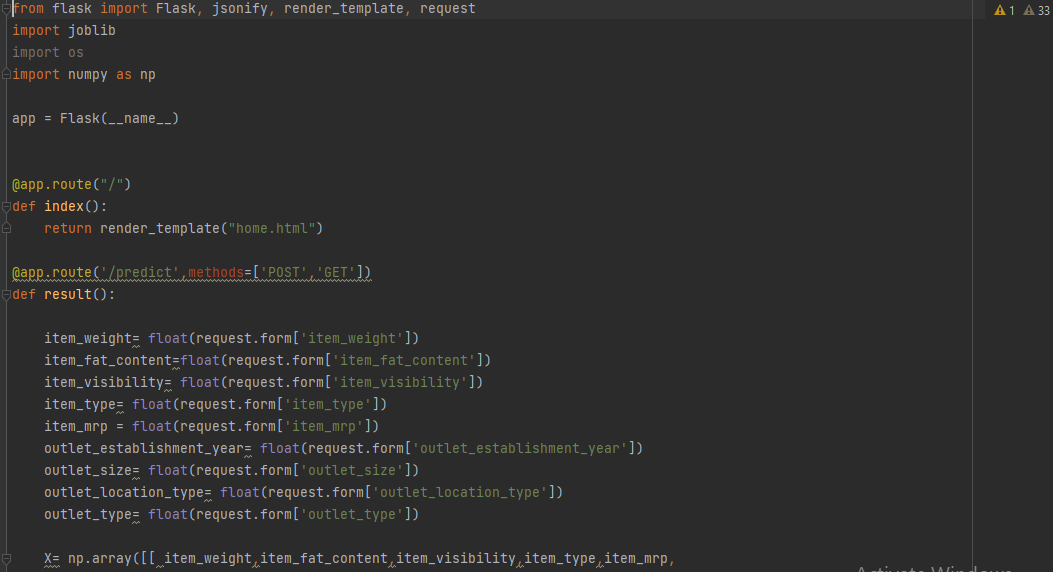
**After that we have to create a html page for our project**

****

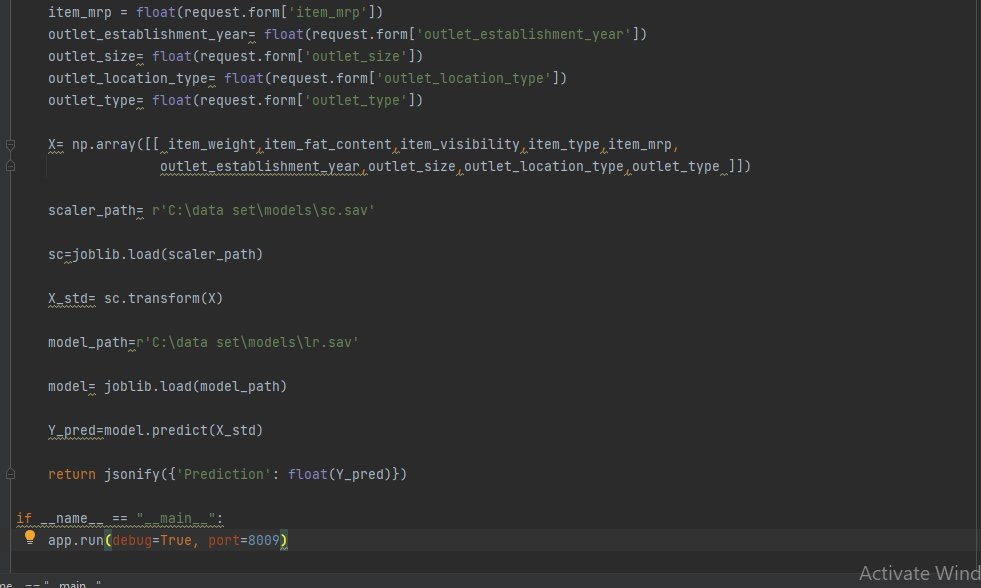
**Add this page , main project templates to our pycharm environment**

****

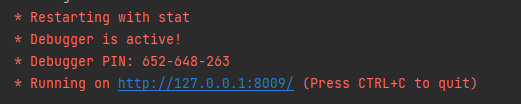
**And deploye , our project to web page 1**

****

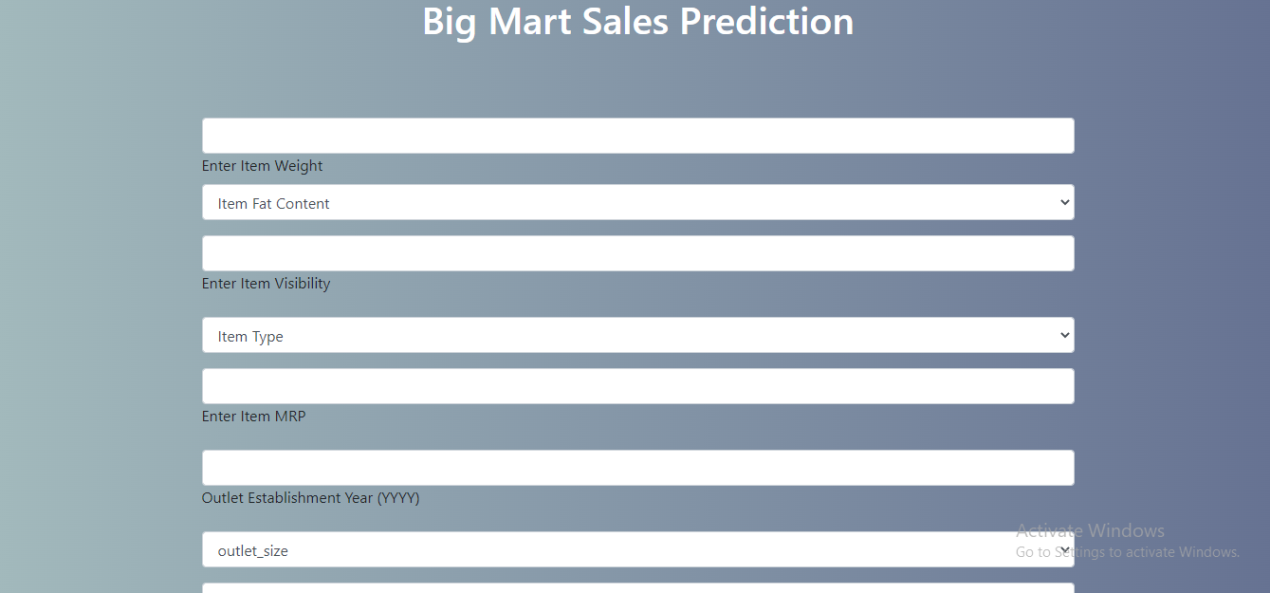
**Page 2**

****

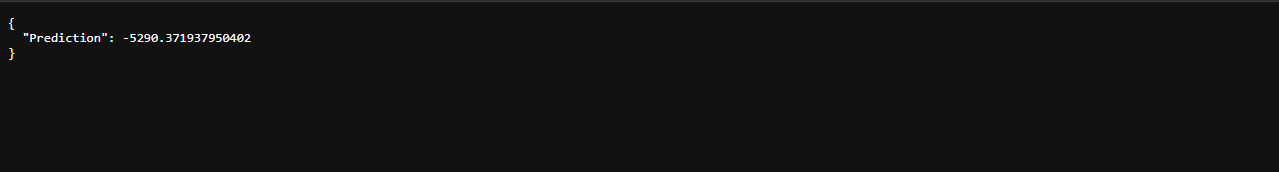
**Link of our project is**

****

**The main output of our project is:- (1)**

****

(2) Prediction page



 **CONCLUSION: -**

In this project, the basics of machine learning and the associated data processing and modeling algorithms have been described, followed by their application for the task of sales prediction in Big Mart shopping centers at different locations. On implementation, the prediction results show the correlation among different attributes considered and how a particular location of medium size recorded the highest sales, suggesting that other shopping locations should follow similar patterns for improved sales.

Also, it can be concluded that more locations should be switched or shifted to Tier-3 in outlet type “Supermarket Type3” to increase the sales of products at Big Mart

By using this method we make any other store model like that to increase the sale of the store

## FUTURE SCOPE: -

The project can be further collaborated in any device supported with an in-built intelligence by the Internet of Things (IoT), to be more feasible for use. Various stakeholders concerned with sales information can also provide more inputs to help in hypothesis generation and more instances can be taken into consideration such that more precise results that are closer to real-world situations are generated. When combined with effective data mining methods and properties, the traditional means could be seen to make a higher and positive effect on the overall development of a simpler algorithms corporation‟s tasks on the whole. One of the main highlights is more expressive regression outputs, which are more understandable bounded with some accuracy. Moreover, the flexibility of the proposed approach can be increased with variants at a very appropriate stage of regression model building.

## REFERENCE: -

* + https://[www.kaggle.com/brijbhushannanda1979/bigmart-sales-data](http://www.kaggle.com/brijbhushannanda1979/bigmart-sales-data)
  + https://ineuron.ai/
  + https://devdocs.io/scikit\_learn/
  + https://xgboost.readthedocs.io/en/stable/
  + https://[www.adamsmith.haus/python/docs/pandas](http://www.adamsmith.haus/python/docs/pandas)
  + https://[www.adamsmith.haus/python/docs/seaborn](http://www.adamsmith.haus/python/docs/seaborn)