Assignment No.3

**Name:** Somesh Gawade

**Roll No.** 8044

LP1 Data Analytics

**Problem statement:** Big-Mart Sales Analysis: For data comprising of transactions records of a sales store. The data has 8523 rows of 12 variables. Predict the sales of a store. Sample Test data set available here

[https://datahack.analyticsvidya.com/contents/practice-problem-big-mart-](https://datahack.analyticsvidya.com/contents/practice-problem-big-mart-sales-iii/) [sales-iii/](https://datahack.analyticsvidya.com/contents/practice-problem-big-mart-sales-iii/)

## Objective:

* To Learn Regression Algorithms.
* To Learn to split dataset into training and test datasets.

**Outcome**: I will be able to develop a predictive model for sales of BigMart.

## Software and Hardware requirements:

* Operating System**:** 64 bit Windows 10
* Programming Language**:** Python(Jupyter Notebook)
* Python Libraries: Pandas ,Numpy, Sklearn, Matpotlib etc.

# Theory Concepts in Brief

**Linear Regression:**

In statistics, linear Regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.

The relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.

If the goal is prediction, forecasting, or error reduction linear regression can be used to fit a predictive model to an observed data set of the values of the response and explanatory variables.

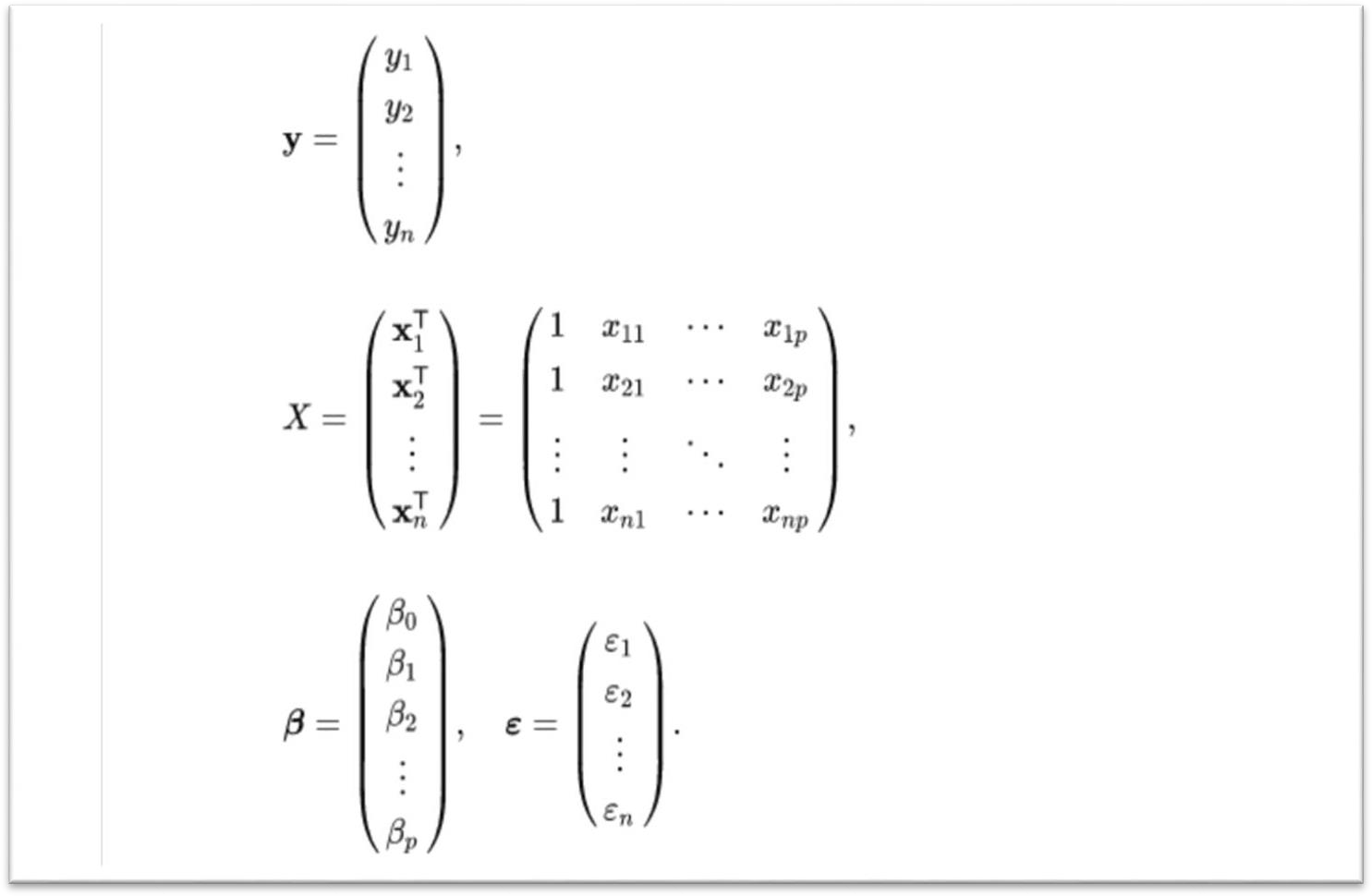
After developing such a model, if additional values of explanatory variables collected without any accompanying response value, the fitted model can be used to make a prediction of the response.

Given a dataset of n statistical units, a linear regression model assumes that the relationship between the dependent variable y and the p-vector of regressions x is linear. This relationship is

modeled through a disturbance term or error variable an unobserved random variable that adds “noise “ to the linear relationship between dependent variable and regressors.

Dataset={yi,xi1,…,xip}ni=1 Model Equation:

yi=xi1+…+pxip+i =xiTI, i=1,2,…,n, Matrix Notation:y=X+,



# Dataset Description:

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.

* **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 ground area covered
* **Outlet\_location\_Type:** The type of city in which 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



Start

Download Dataset

using Kaggle API

Load Dataset in

Pandas DataFrame

Handle Null Values

Handle Redu**a**ndant Values

Label Encoding of

String values

Split in Test and

train

Train the Model

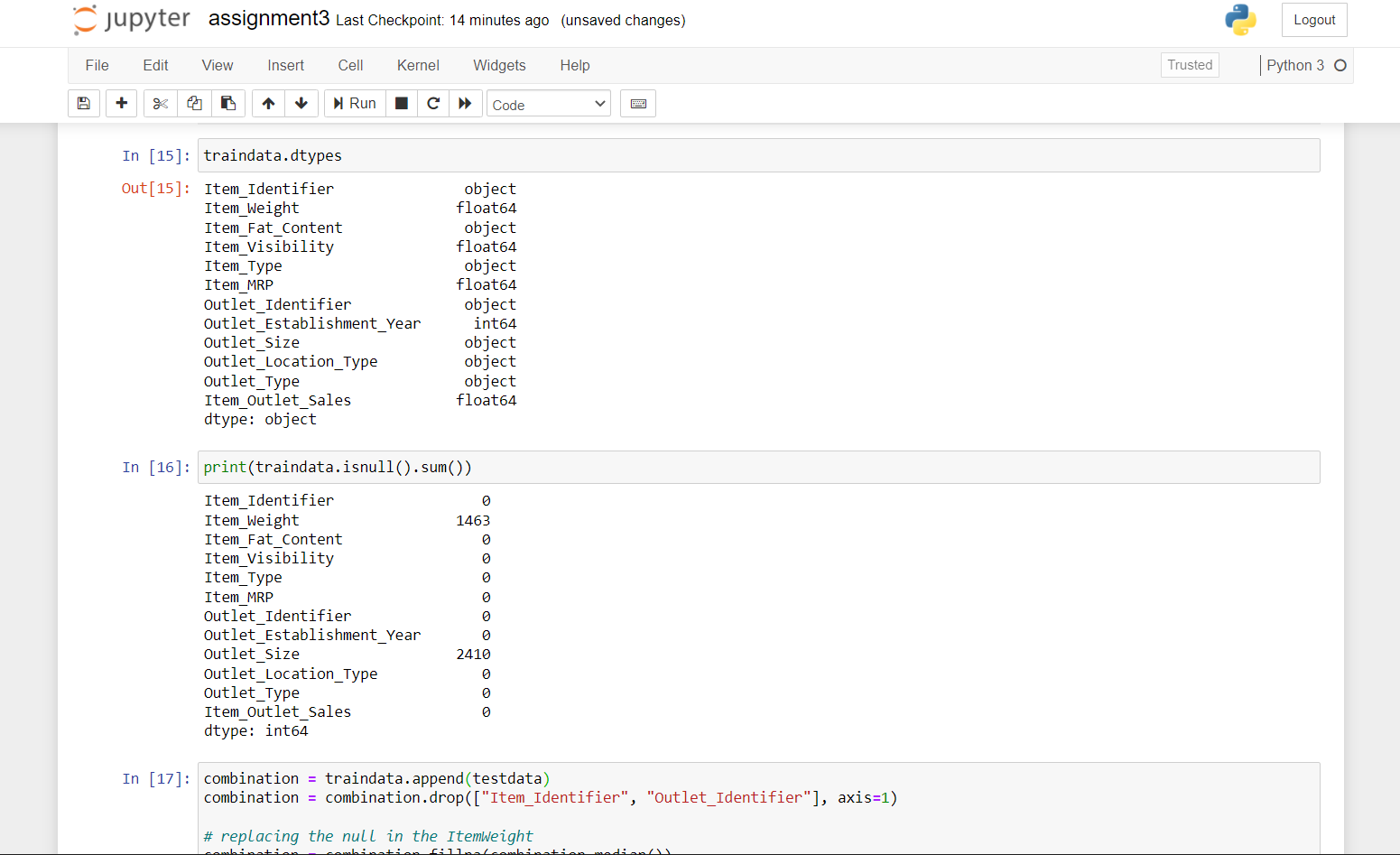
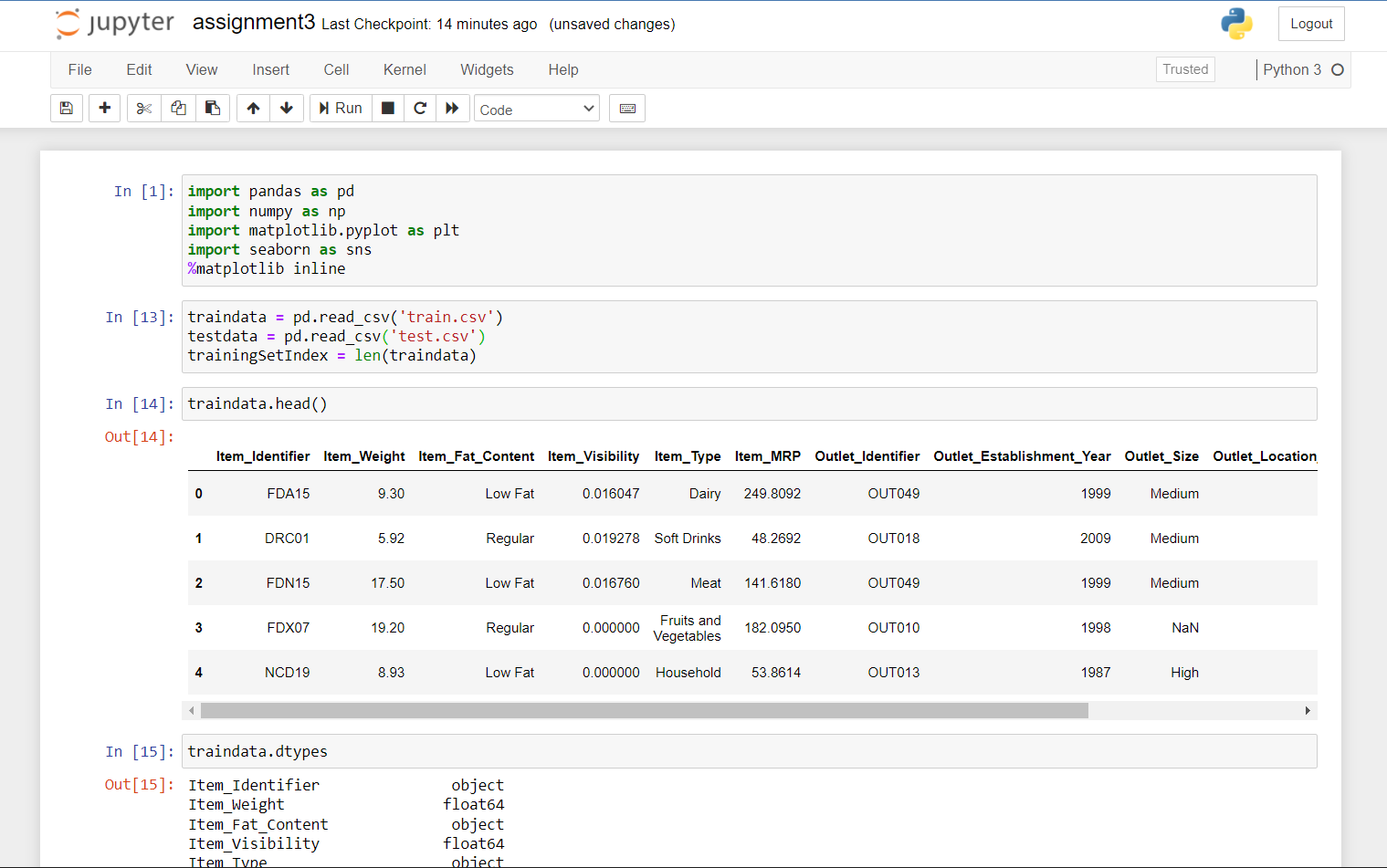
Display

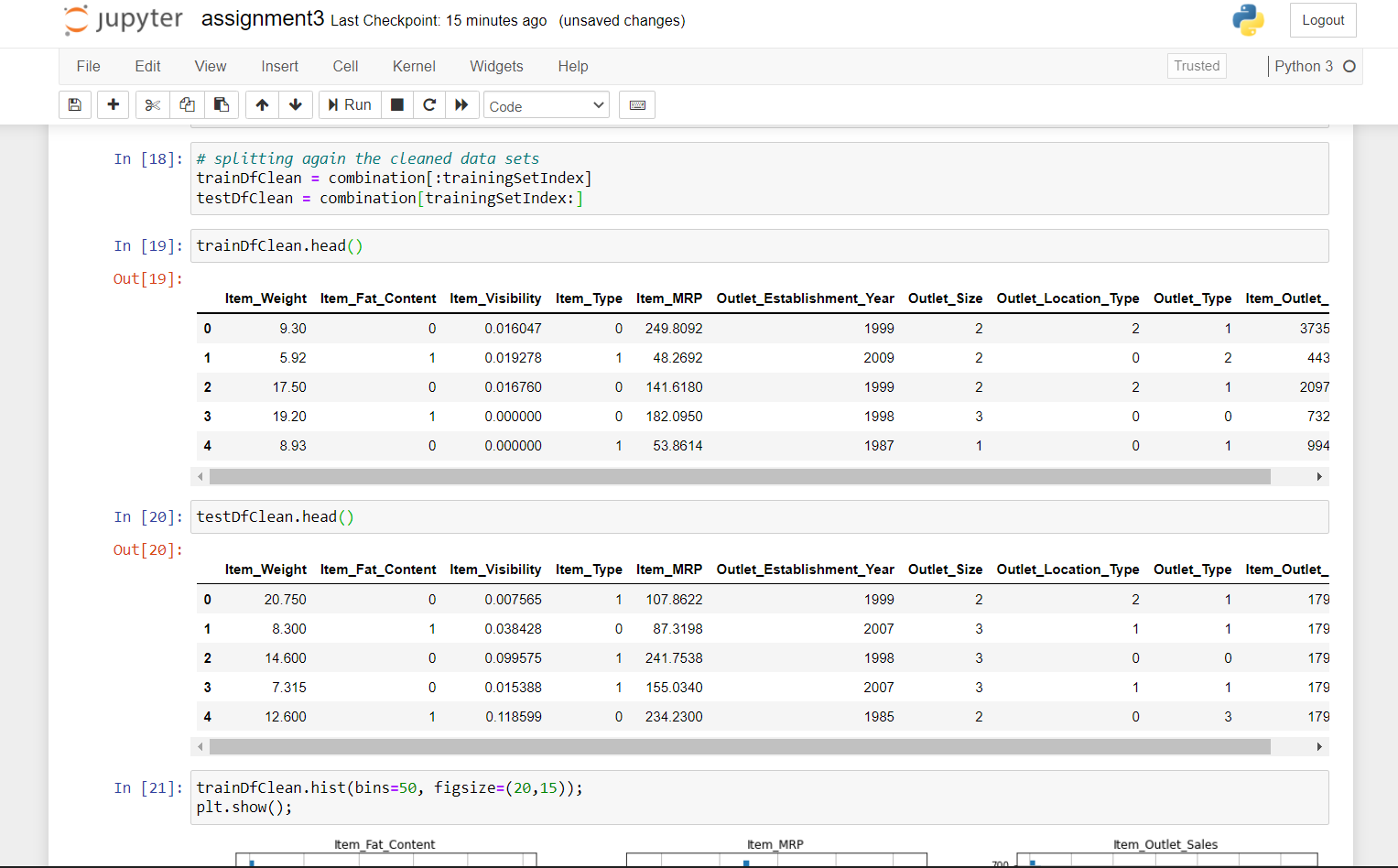
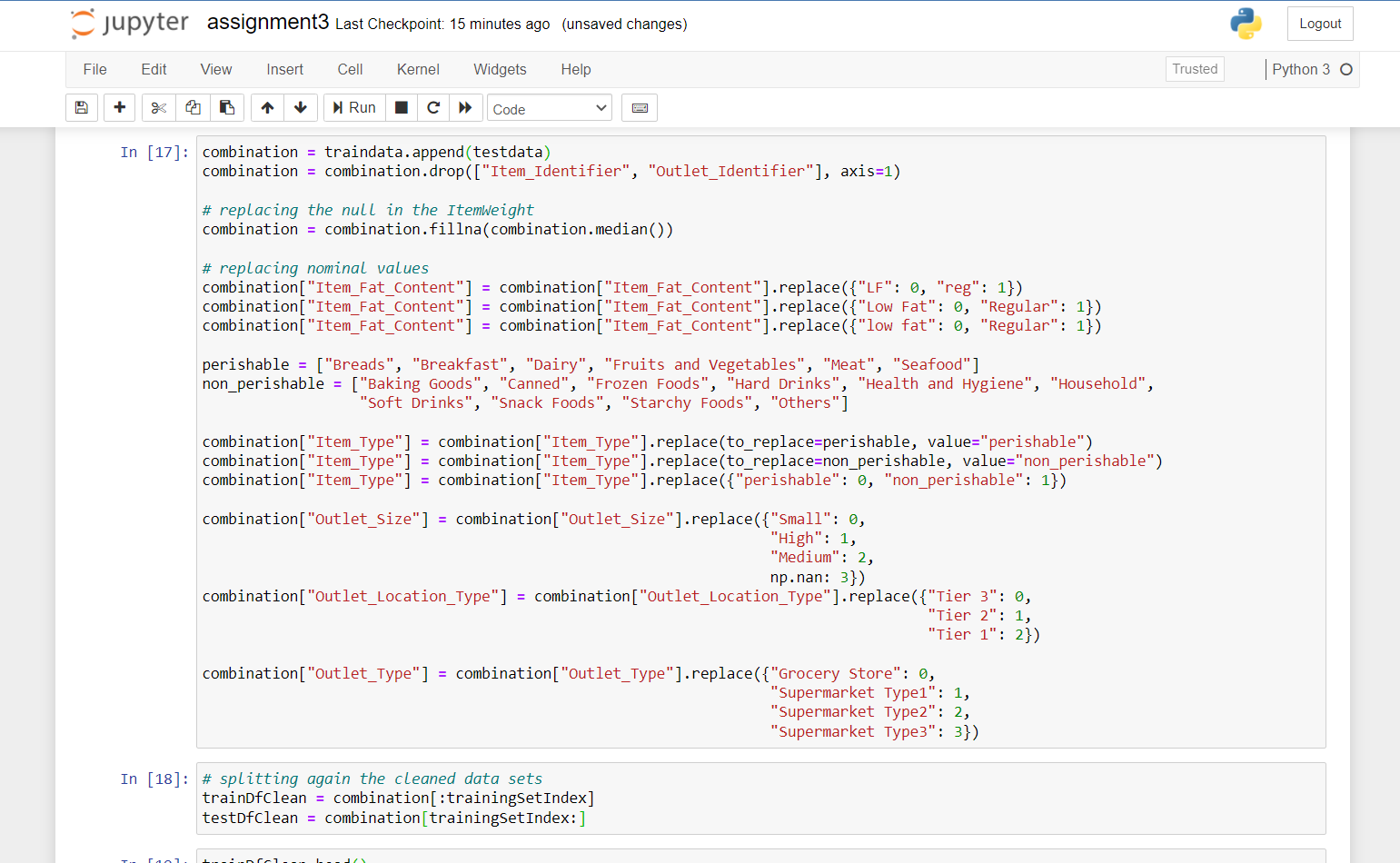
RMSE

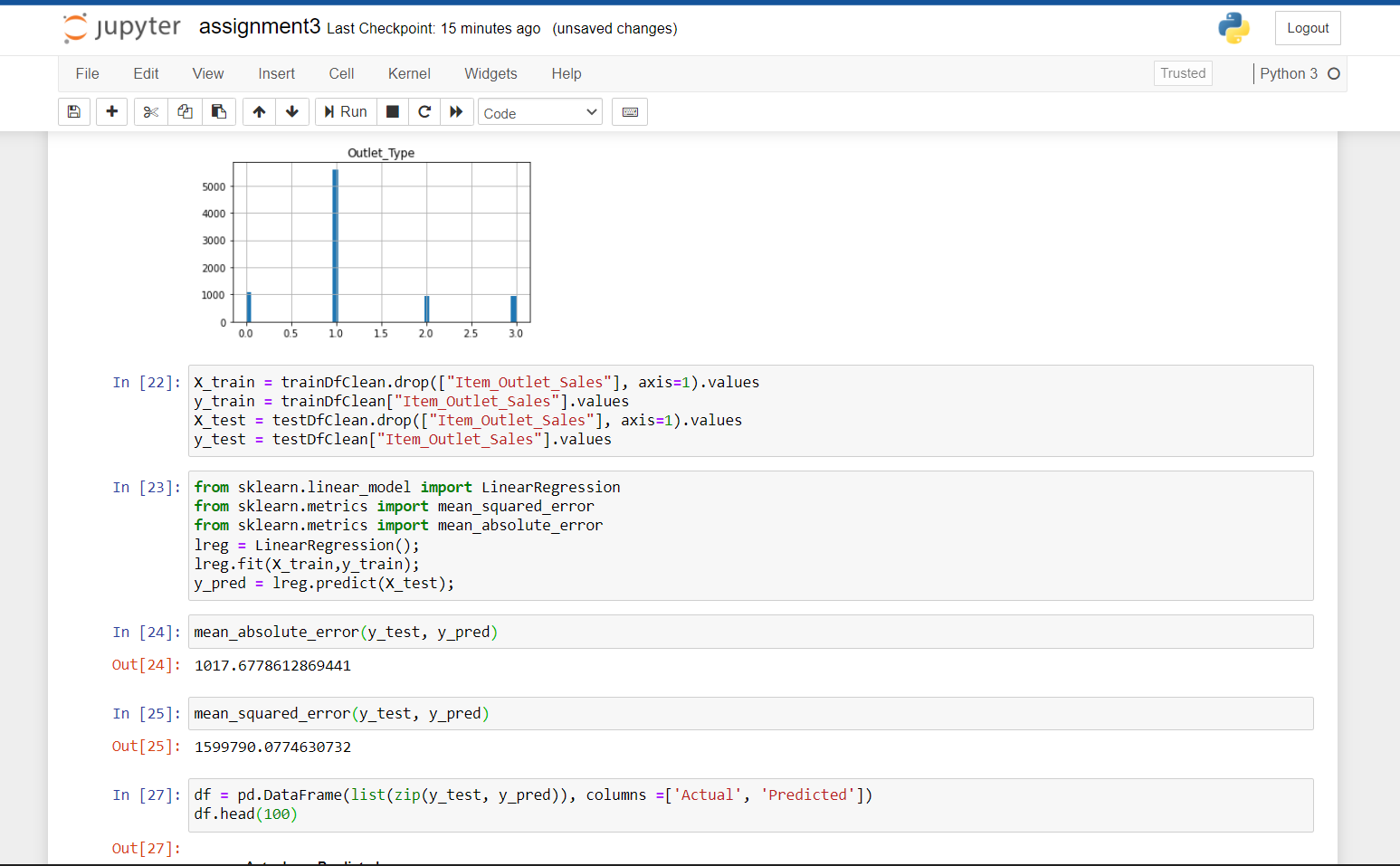
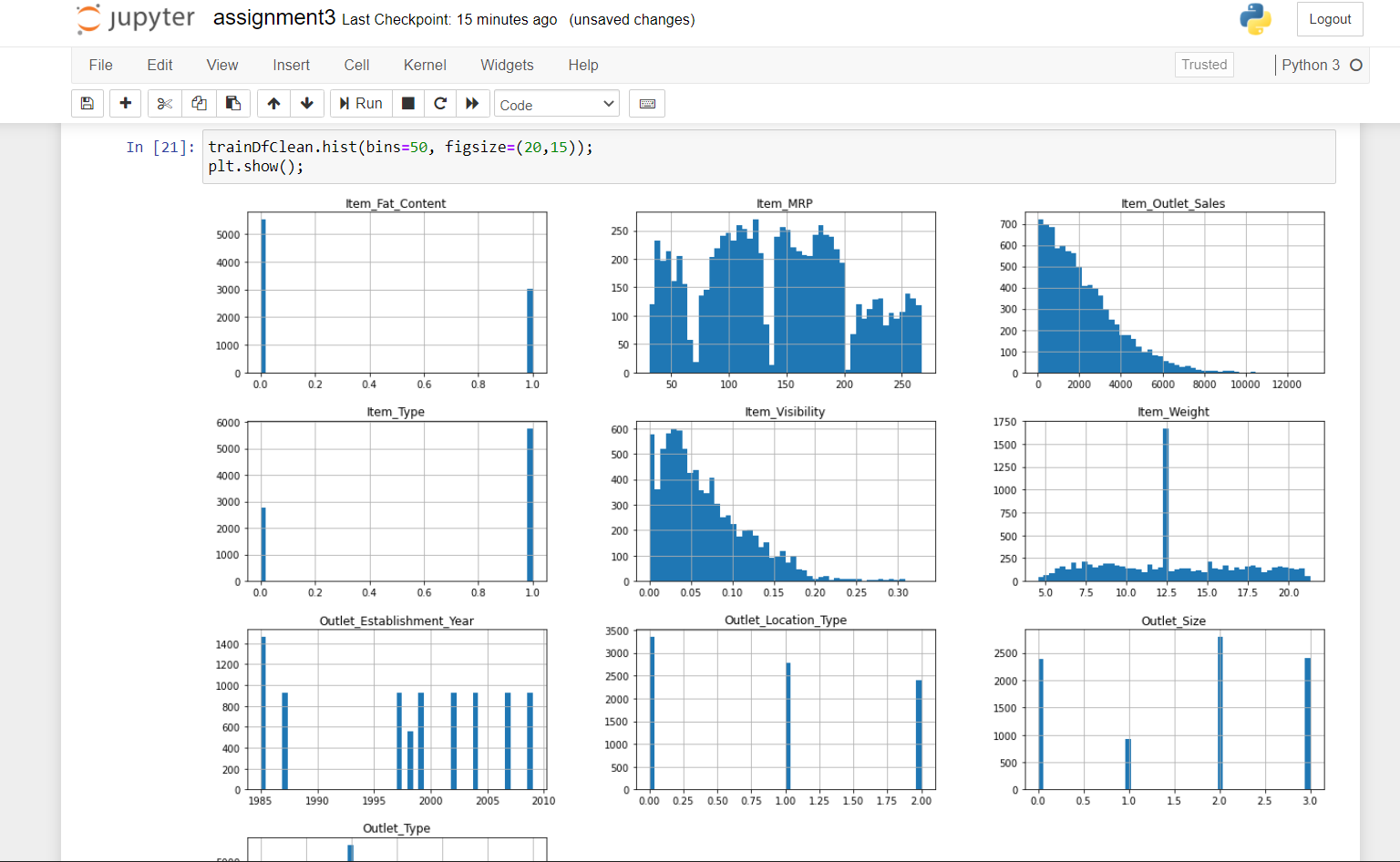
End

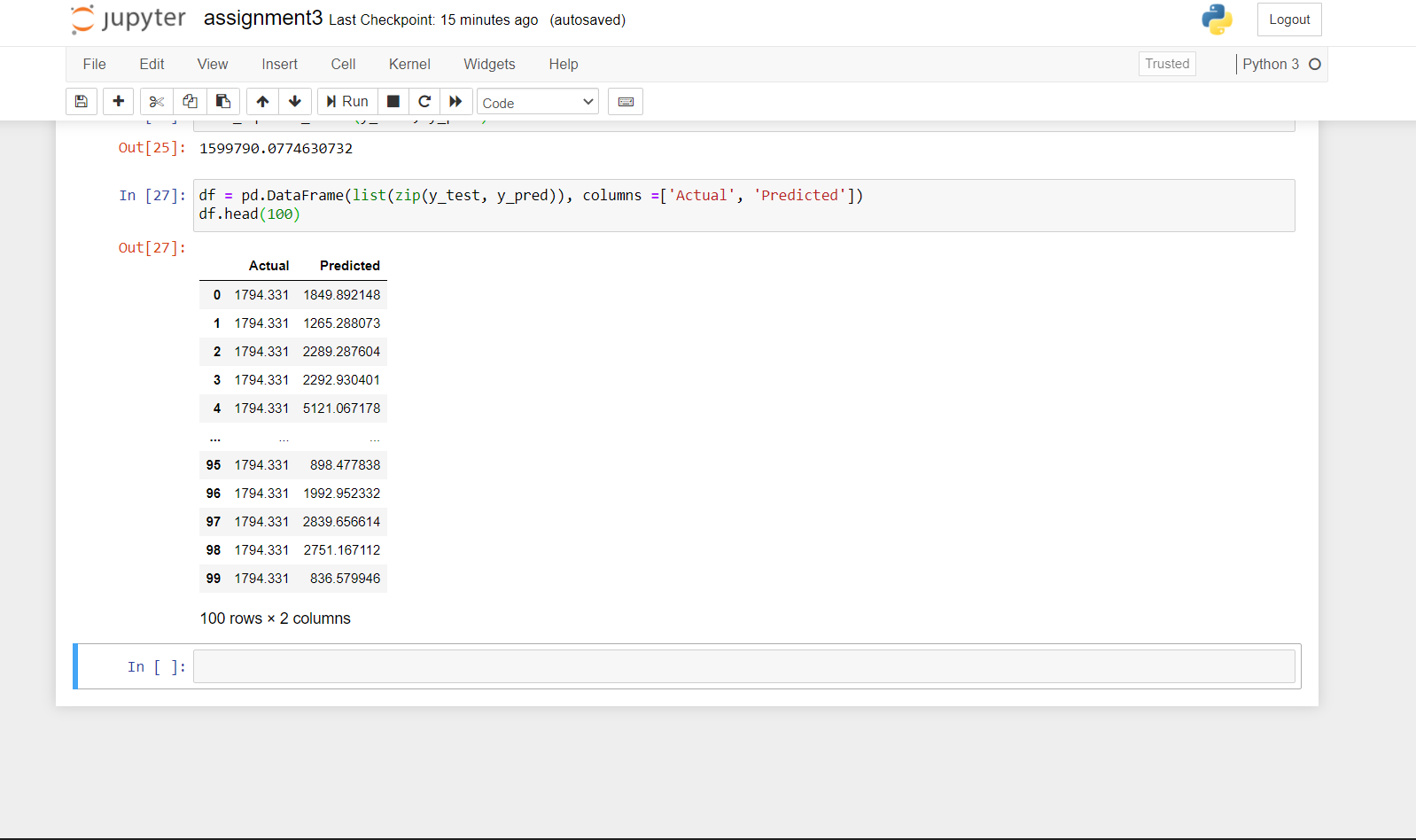
**Conclusion:** I have successfully developed the predictive model for sales of an Item at BigMart.

**Output:**

****

****

****

****