Sales Forecasting

In Big-Mart

By

R.Jhansi

S180620

Cse-2A

**Aim:**

The aim of the project is to build a predictive model and find out the sales of a product at particular store using machine learning.

**Purpose:**

In today’s world big malls and marts record sales data of individual items for predicting future demand and inventory management. This data Stores a large number of attributes of the item as well as individual customer data together in a data warehouse. This data is mined for detecting frequent patterns as well as anomaliies. This data can be used for forecasting future sales with the help of xgboost regression model.

Using this model, Big mart will try to understand the properties of products and stores which plays a key role in increasing sales.

**Introduction:**

Global malls and stores chains and the increase in the number of electronic payment customers, the competition among the rival organizations is becoming more serious day by day.

Each organization is trying to attract more customers using personalized and short- time offers which makes the prediction of future volume of sales of every item an important asset in the planning and inventory management of every organization, transport service, etc.

Due to the cheap availability of computing and storage, it has become possible to use sophisticated machine learning algorithms for this purpose. In this paper, we are providing forecast for the sales data of big mart in a number of big mart stores across various location types which is based on the historical data of sales volume.

**Algorithm used:**

Xgboost regression is used to fit the model and predict output of the model

**Input:**

Item identifier, Item weight, Item fat content, Item visibility, Item type, Item MRP, Outlet identifier, Outlet establishment year, Outlet size, Outlet location type, Outlet type

**Output:**

Item outlet sales

**Libraries used:**

Numpy Library

Pandas Library

Matplot Library

Seaborn Library

Scikit-learn Library

Xgboost

**Building the model:**

1.Collect the data

2.Import required packages

3.Exploratory data analysis

4.Data preprocessing

5.Splitting data to train and test data

6.Regression model

7.Evaluation

8.Make prediction using the model

**About the dataset:**

Item\_Identifier: To find the item category

Item\_weight: Weight of the item

Item\_Fat\_Content: The amout of fat the item contains

Item\_Visibility: The location of product in a store will impact sales. Ones which are right at entrance will catch the eye of customer first rather than the ones in back.

Item\_Type: Tells what type of item it is

Item\_MRP: Price of the item

Outlet\_Identifier: Similar to product id, Used for store identification

Outlet\_Establishment\_Year: On which year the store is established

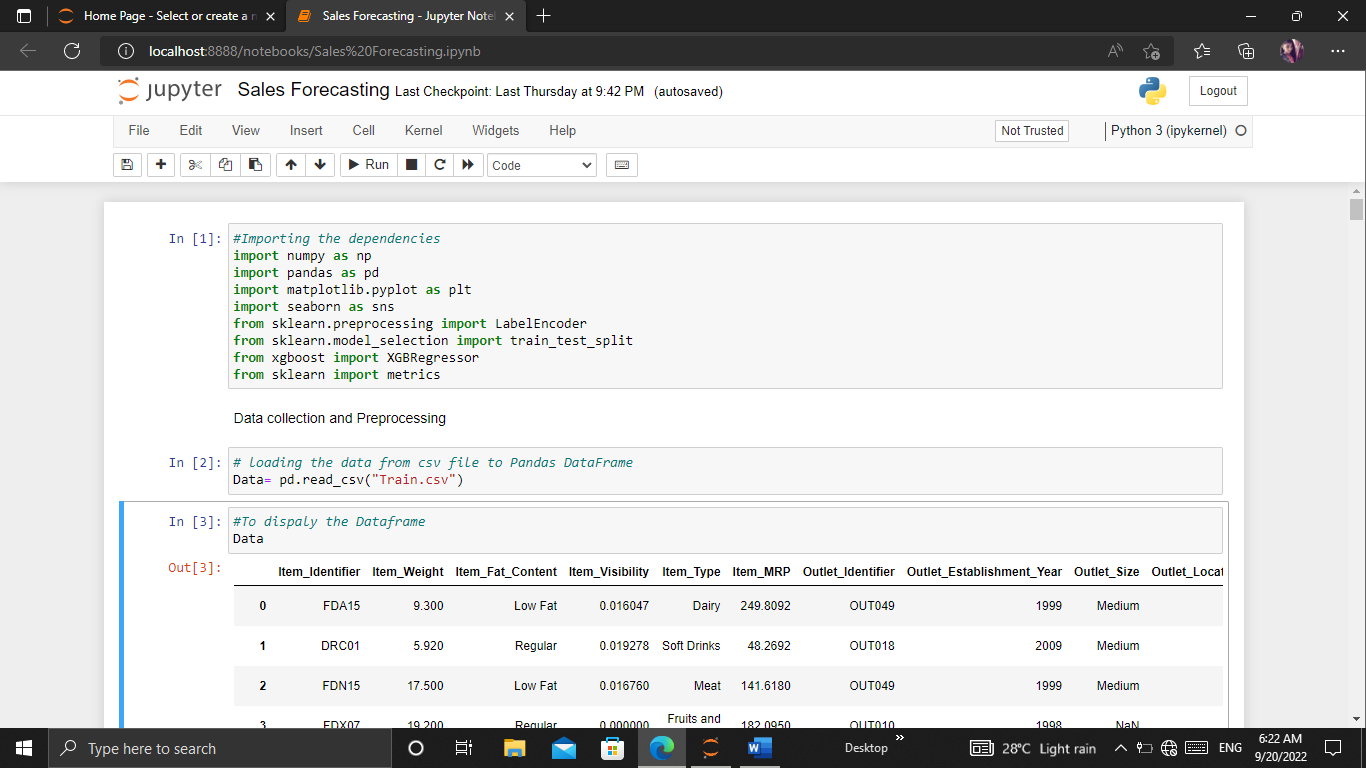
Outlet\_Size: Size of the store

Outlet\_Location\_Type: Where the outlet located whether in populated area,less popular area etc

Outlet\_Type: Type of the outlet

Item\_Outlet\_Sales: What is the sales of particular product. This is the outcome variable to be predicted.

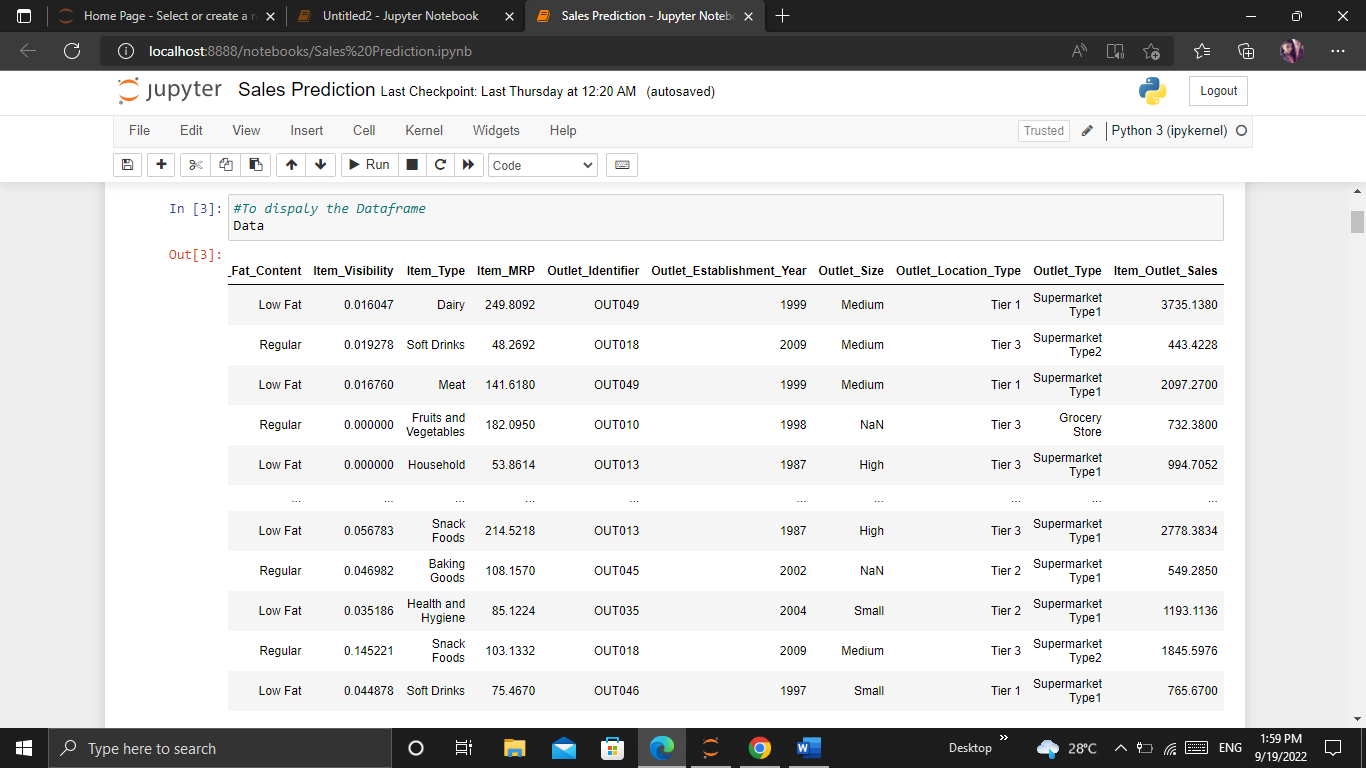
**Importing Packages:**



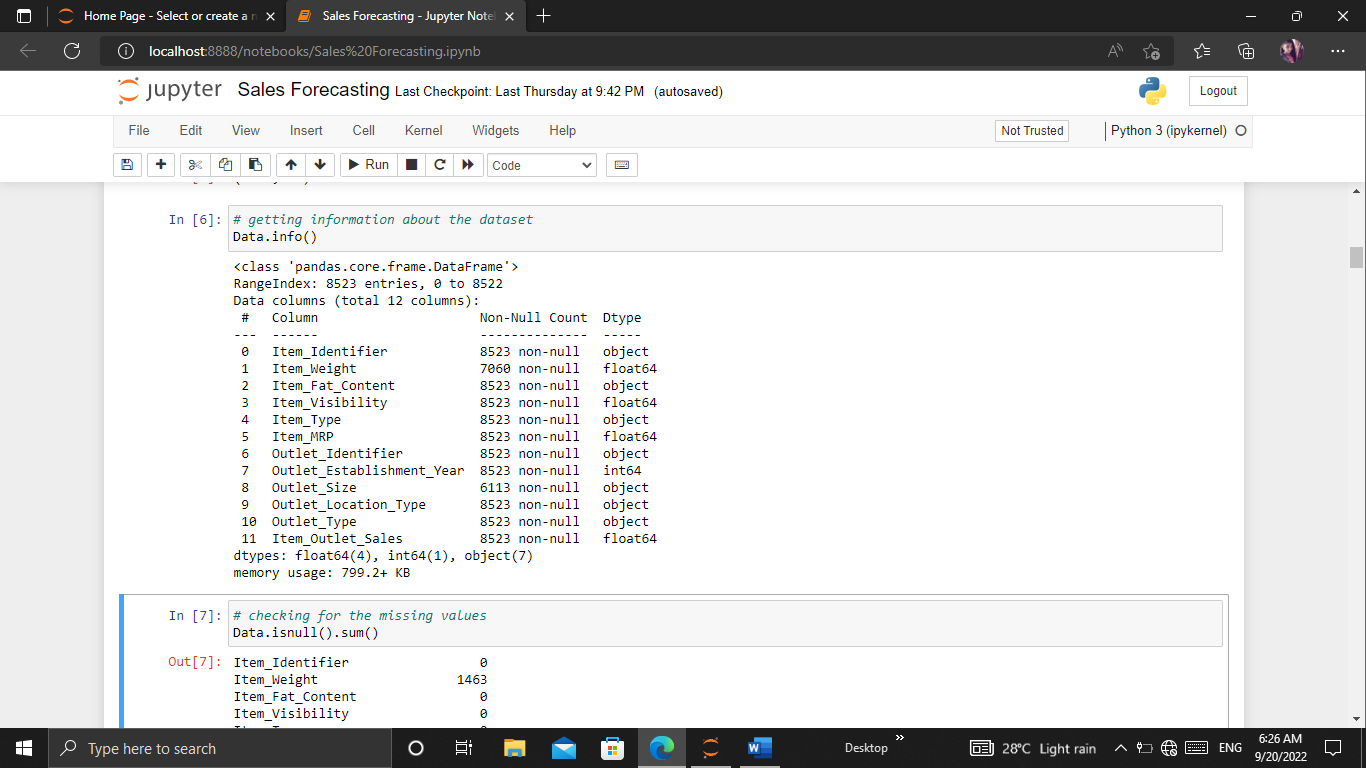
**Data Analysis:**

I have taken a data set which has 8523 rows and 12 columns.

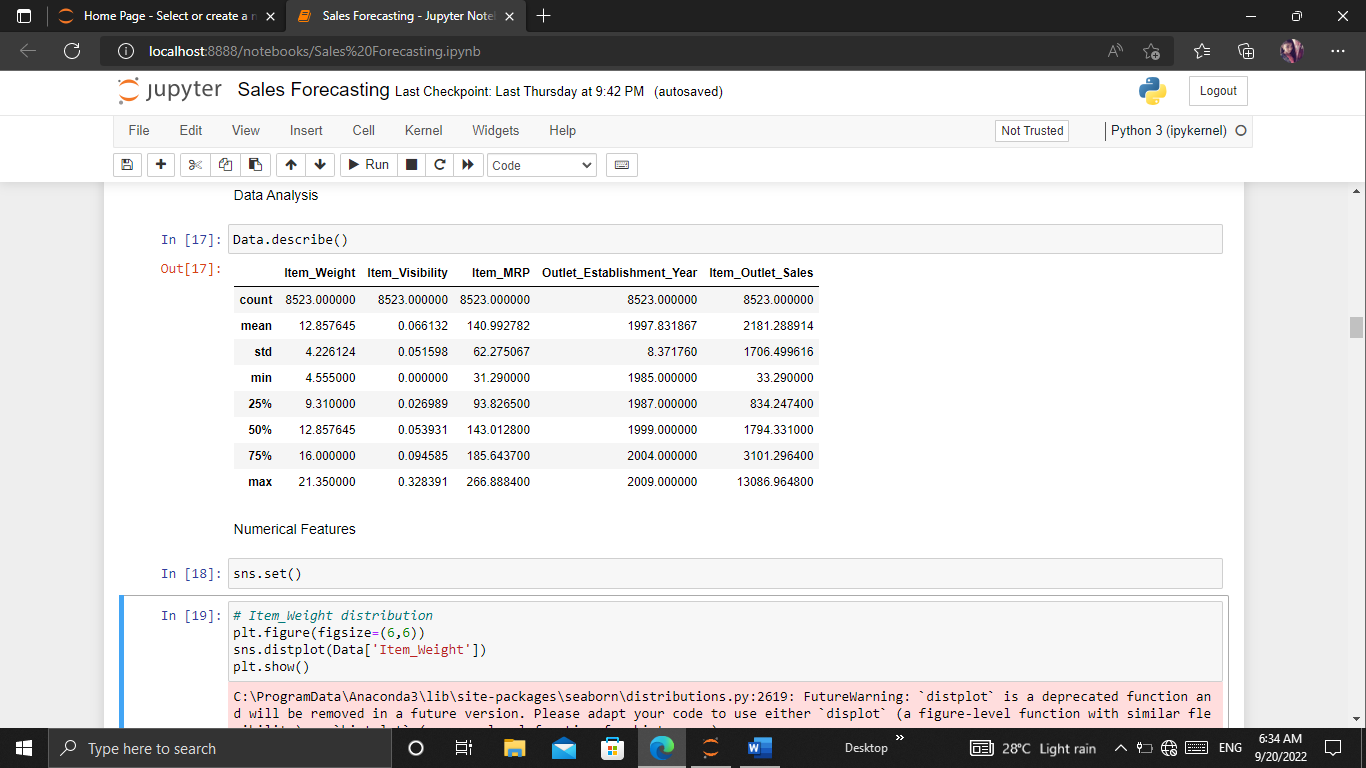
The goal for the Data analysis is to get some insight and if any irregularities are found we will correct that in the next section, **Data Pre-Processing.**



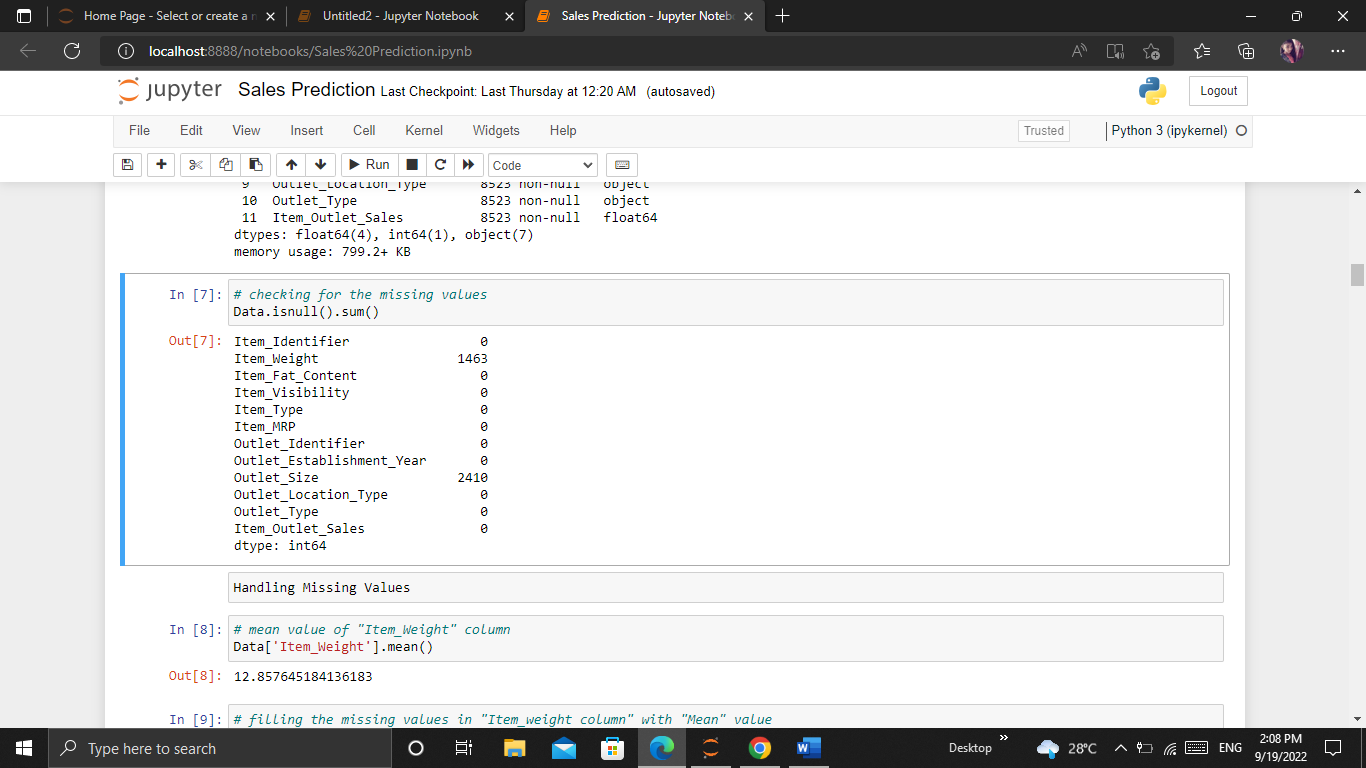
Getting information about the dataset:



Discriptive statistics of dataset:



Checking for Missing values:



We can observe that we are having 1463 missing values in the Item\_Weight column & we are having about 2410 missing values in the Outlet\_Size column.

**Handing Missing values:**

1.Data['Item\_Weight'].mean)

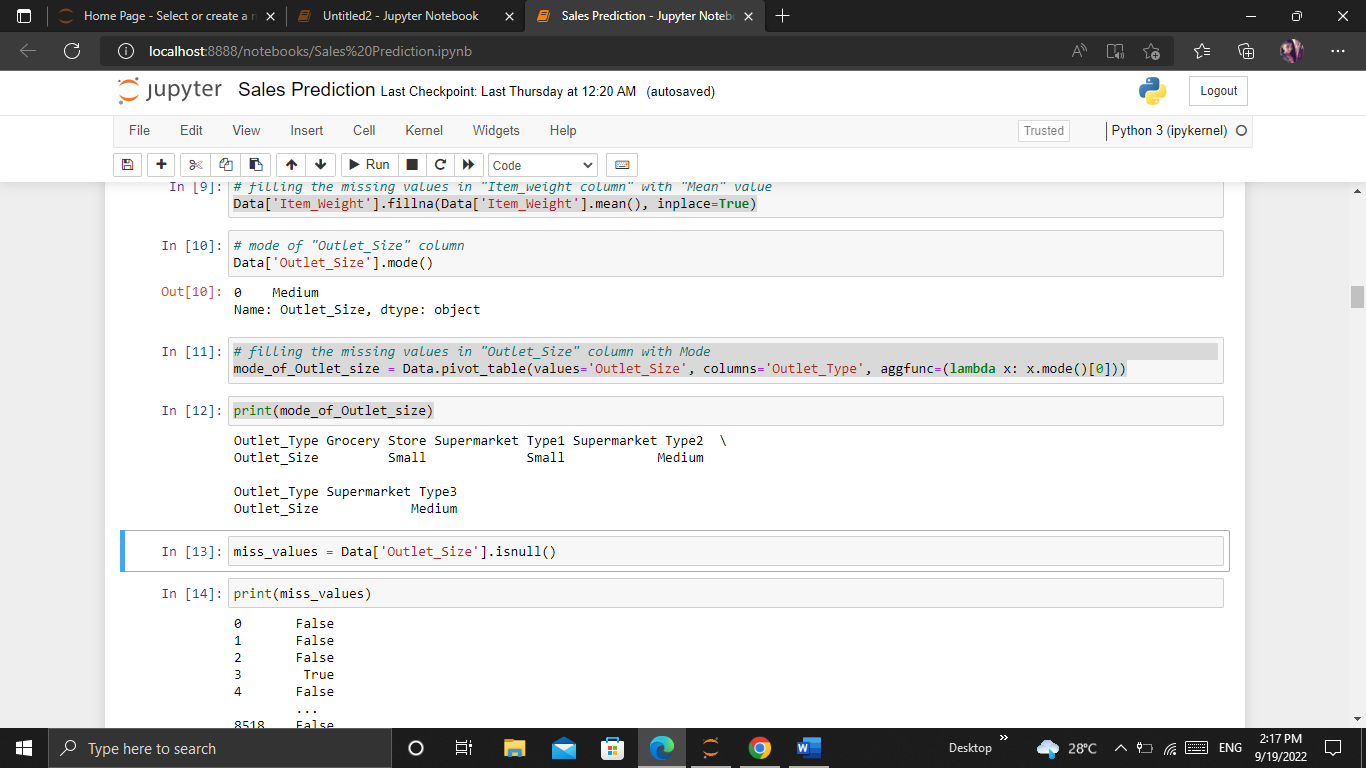
o/p:12.857645184136183

2.Data['Item\_Weight'].fillna(Data['Item\_Weight'].mean(), inpl ace=True)

# filling the missing values in "Outlet\_Size" column with Mode

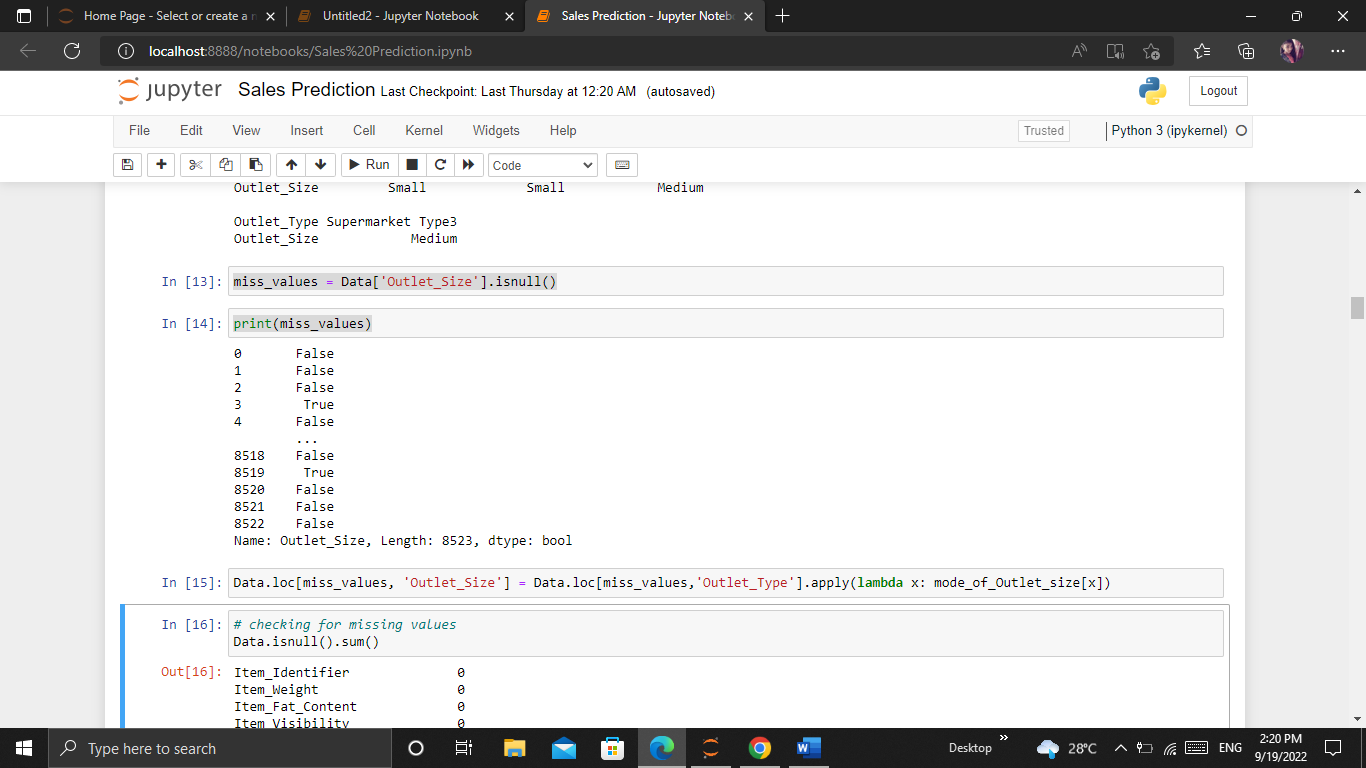
3.mode\_of\_Outlet\_size=Data.pivot\_table(values='Outlet\_Size',columns='Outlet\_Type',aggfunc=(lambda x: x.mode()[0]))

4. print(mode\_of\_Outlet\_size)



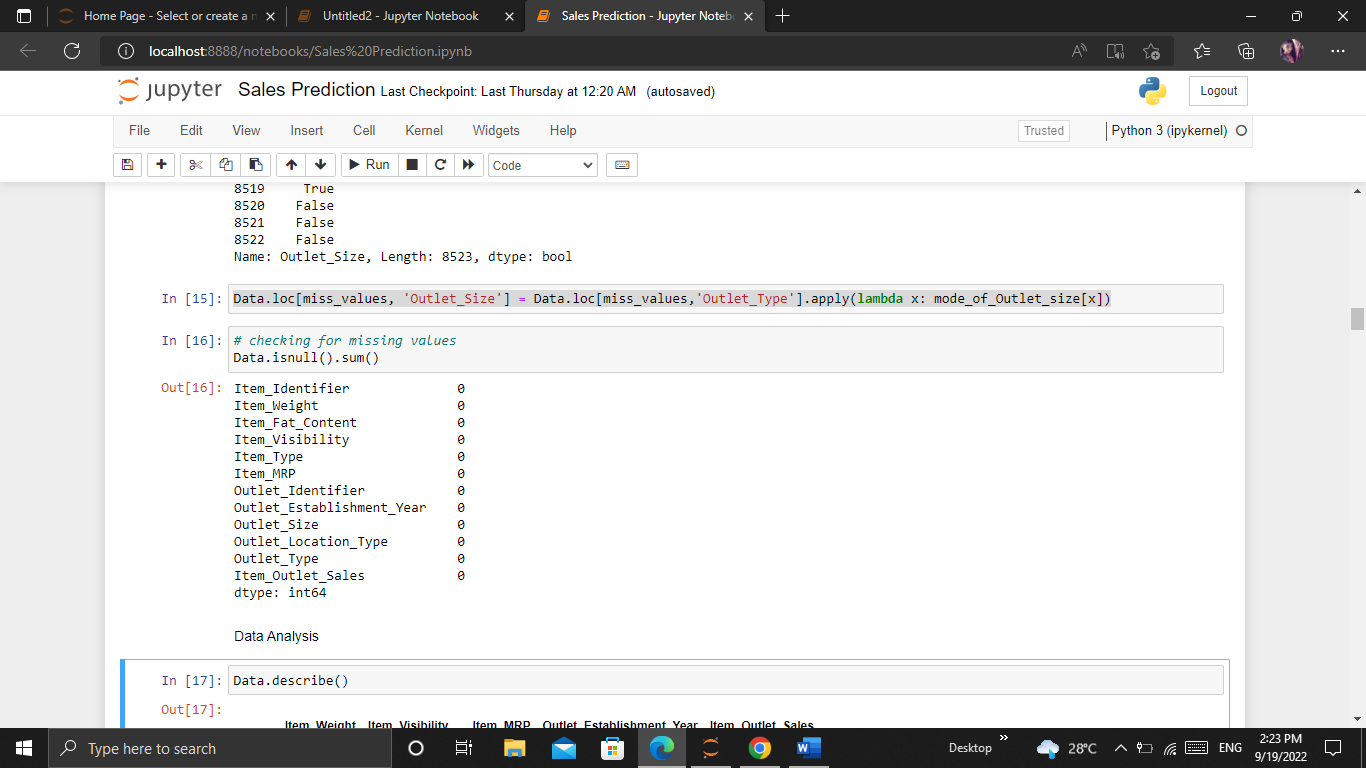
5. miss\_values = Data['Outlet\_Size'].isnull()

print(miss\_values)



6.Data.loc[miss\_values,'Outlet\_Size']=Data.loc[miss\_values,'Outlet\_Type'].apply(lambda x:mode\_of\_Outlet\_size[x])

7. Data.isnull().sum()



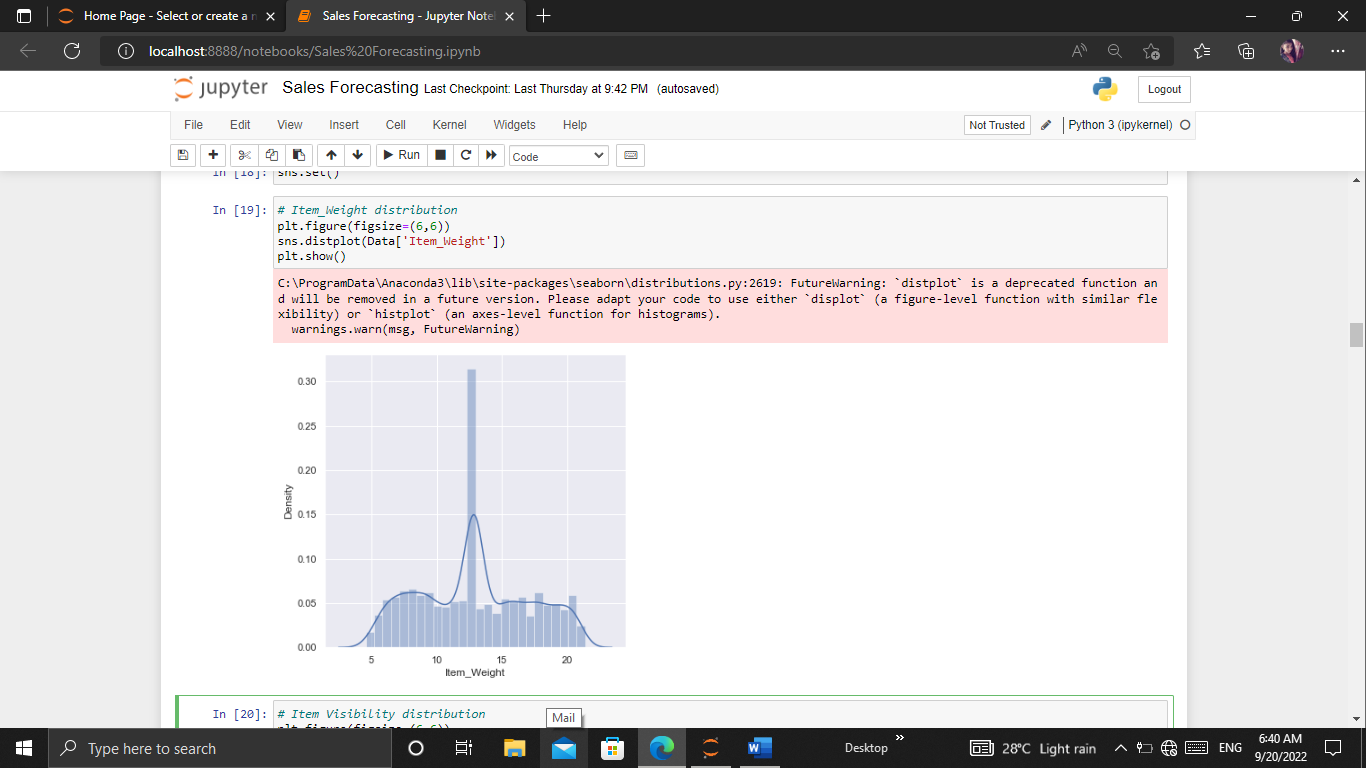
Now the data is cleaned

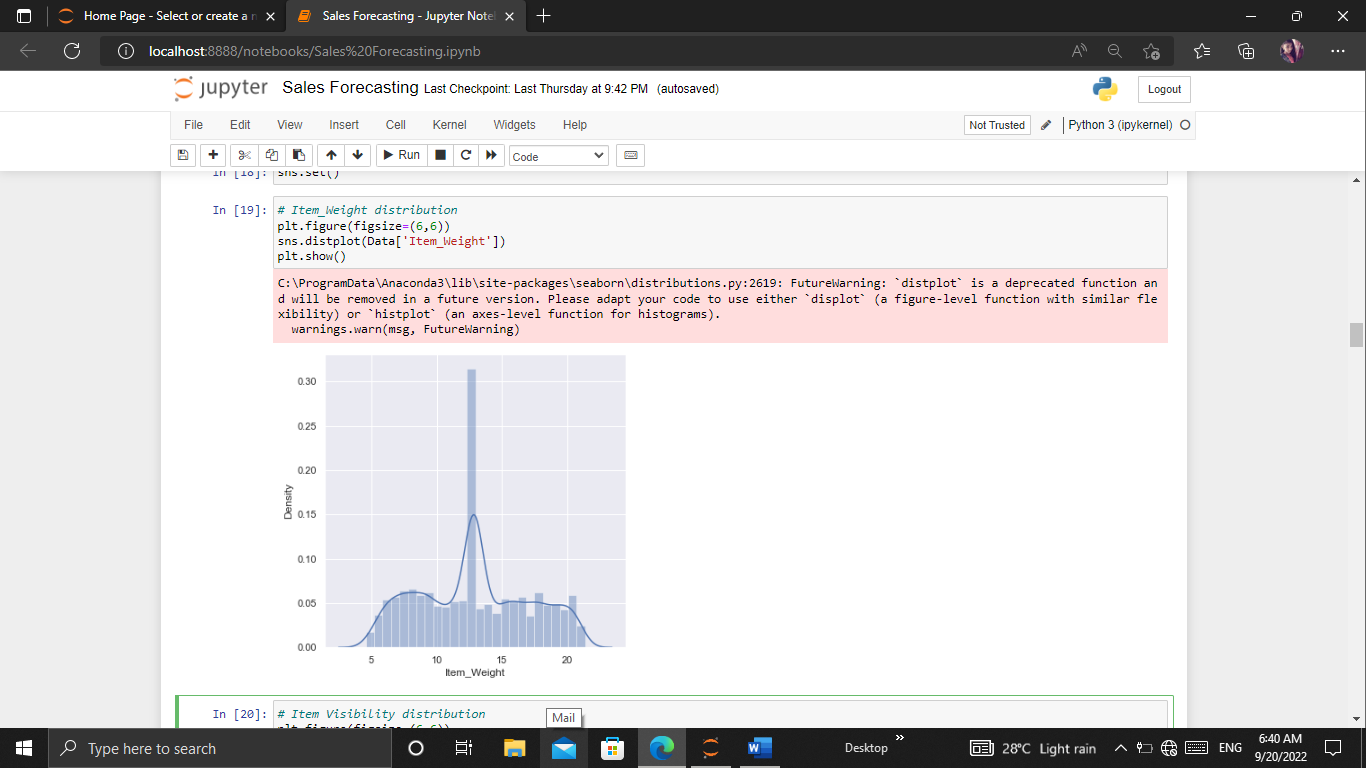
**Data visualization:**

Data visualization is the graphical representation of information and data.

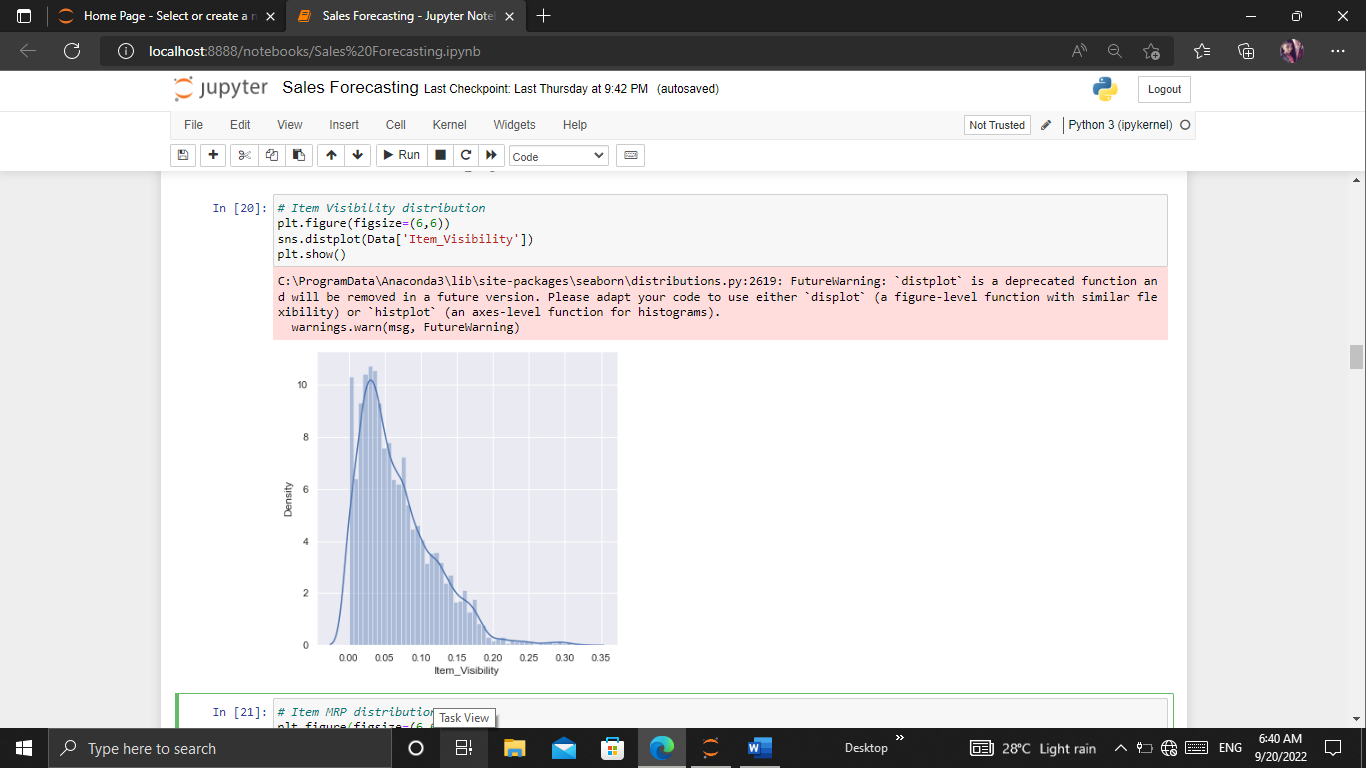
It enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns.

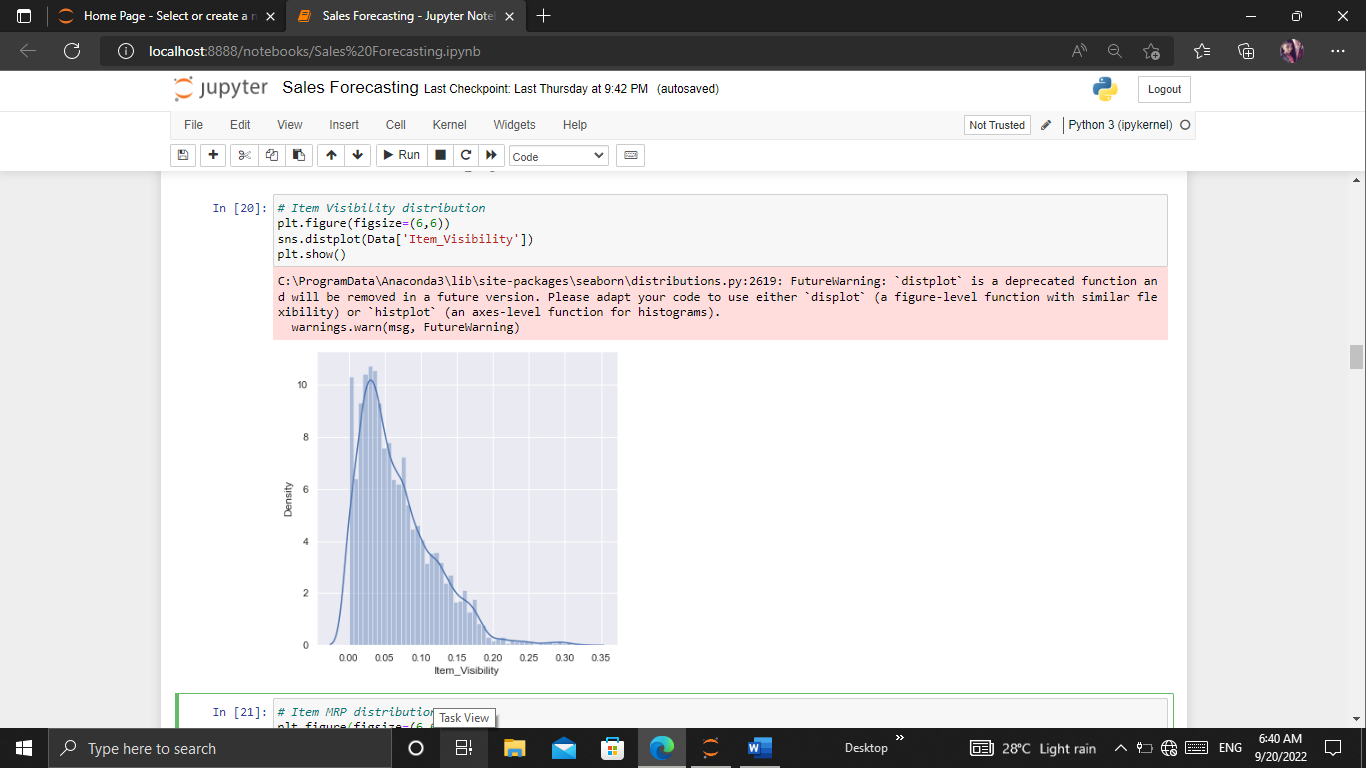
Here visualized the numerical data using distribution plots to comparing range and distribution for groups of numerical data and visualized the categorical data in Count plot to represent the occurrence of the observation present in categorical variable.



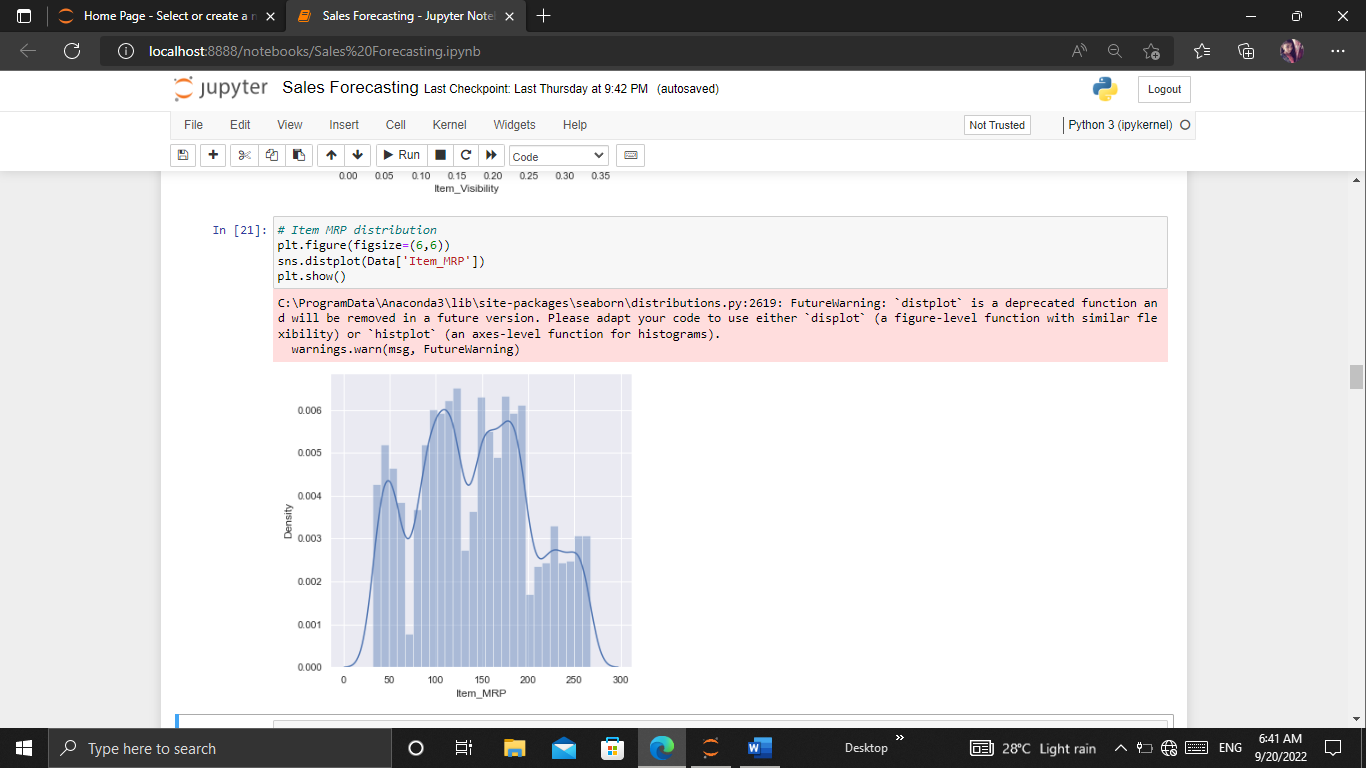


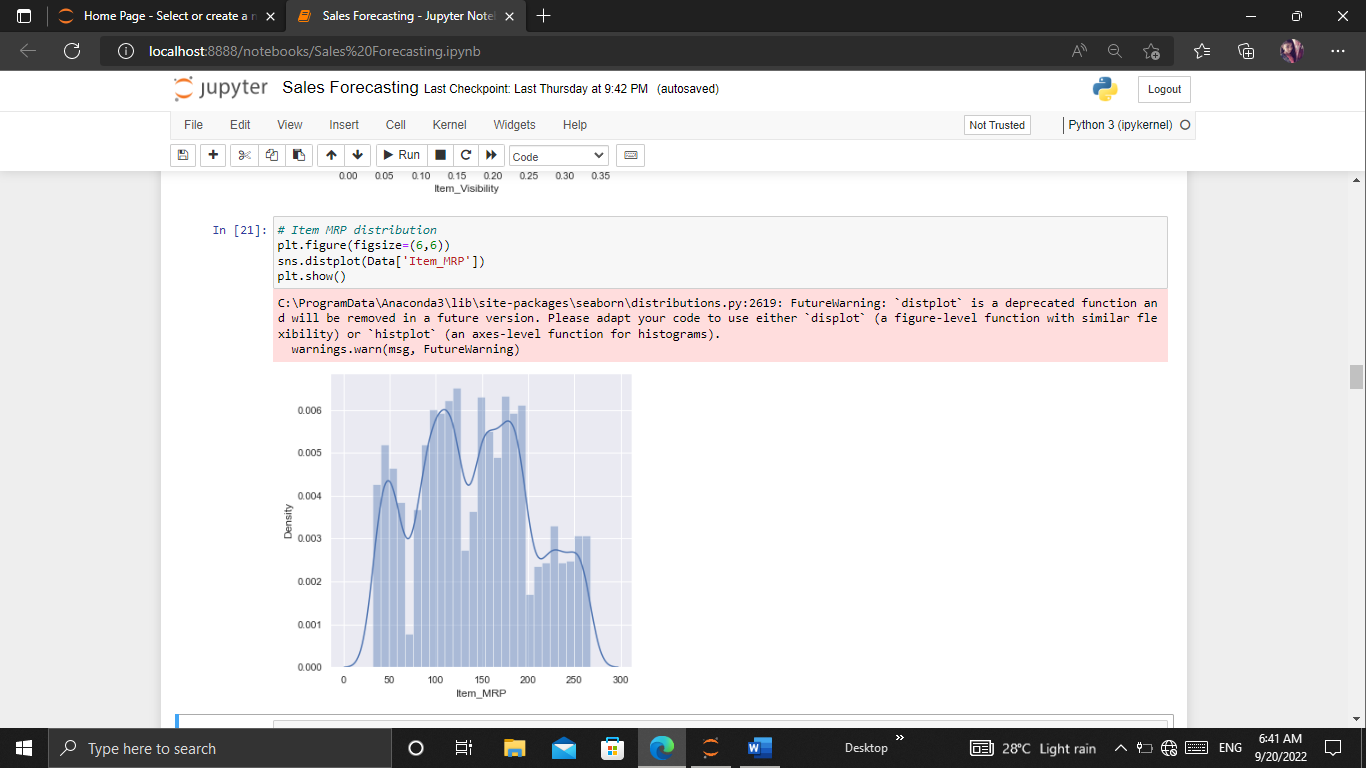
Hence from the above graph we can observe that we have the item weight from 5 Kg to 20 Kg & we have maximum values around 12 Kg where the mean is 12.85 Kg. Therefore in this 8523 products the average weight is about 12.8 Kg



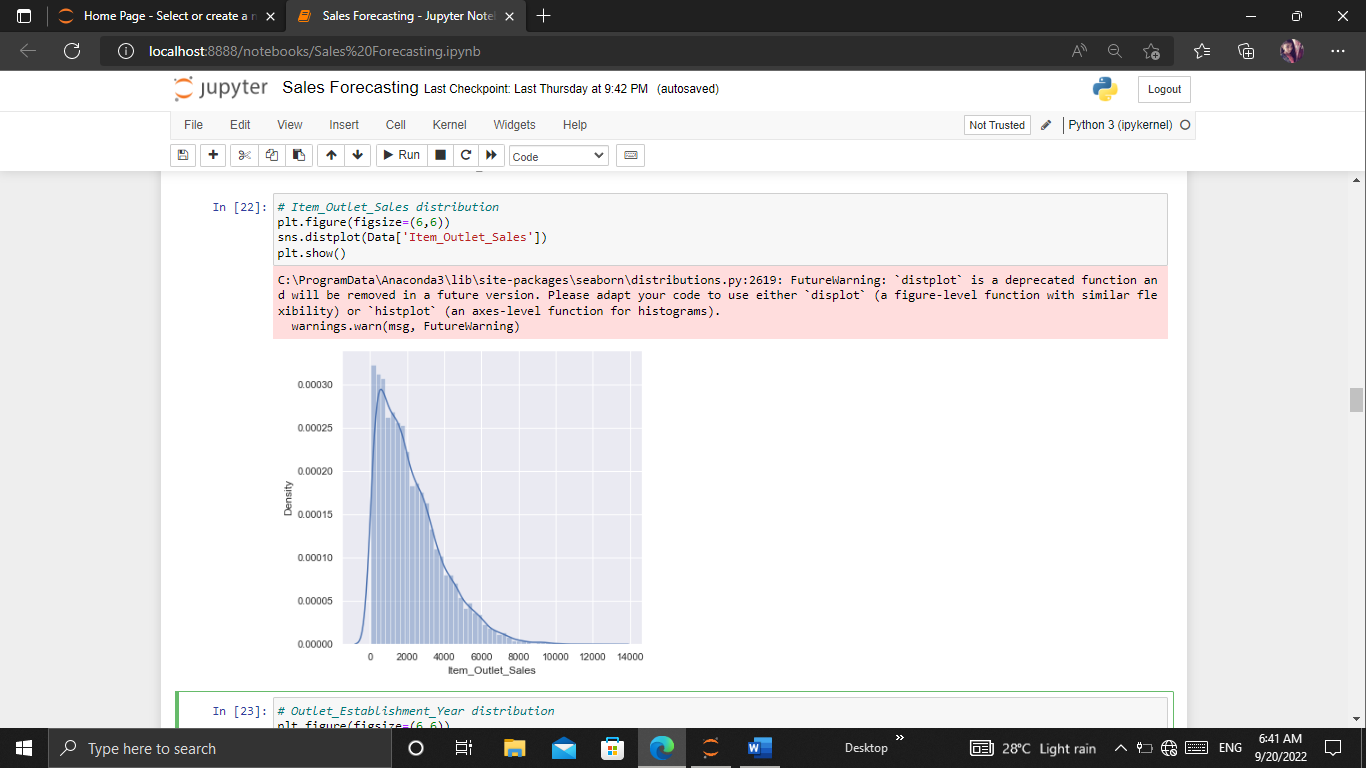


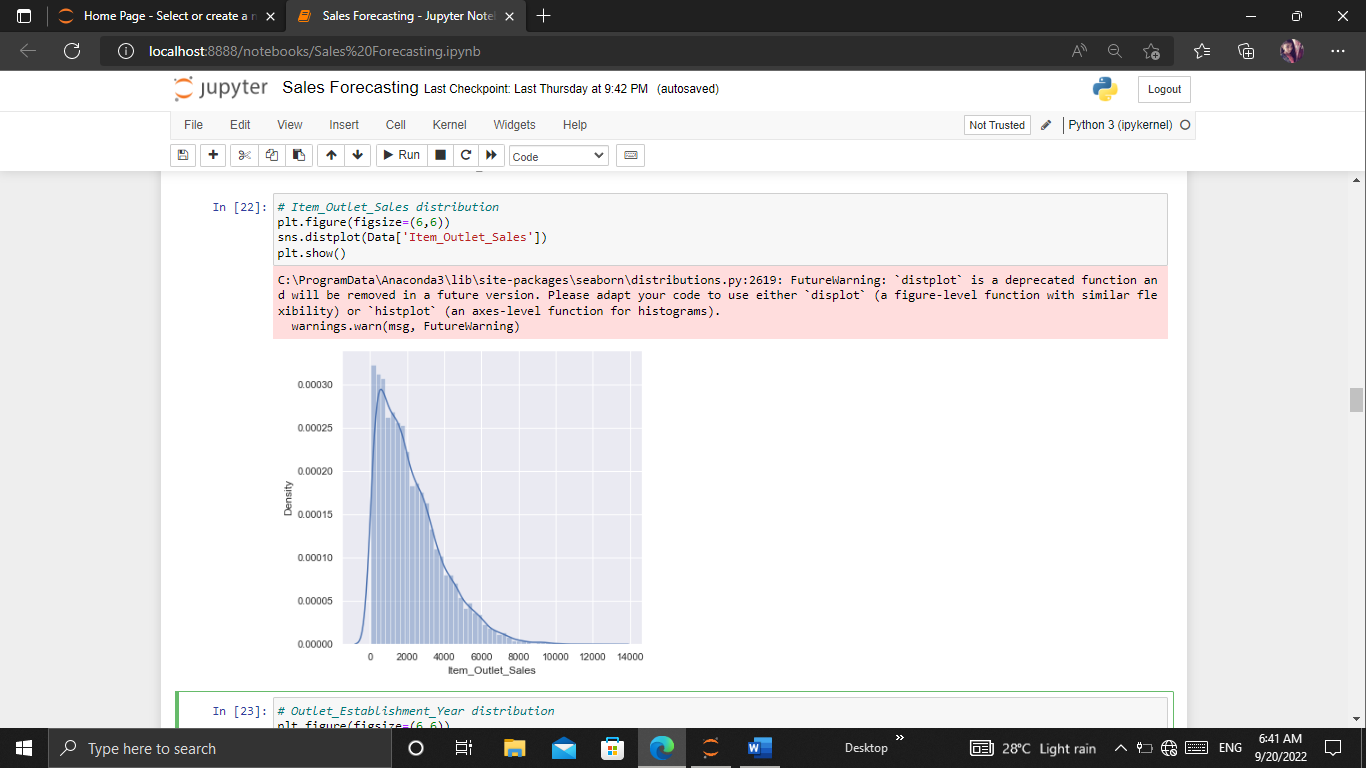
Hence from the above graph we can observe that Item\_Visibility feature is positively skewed



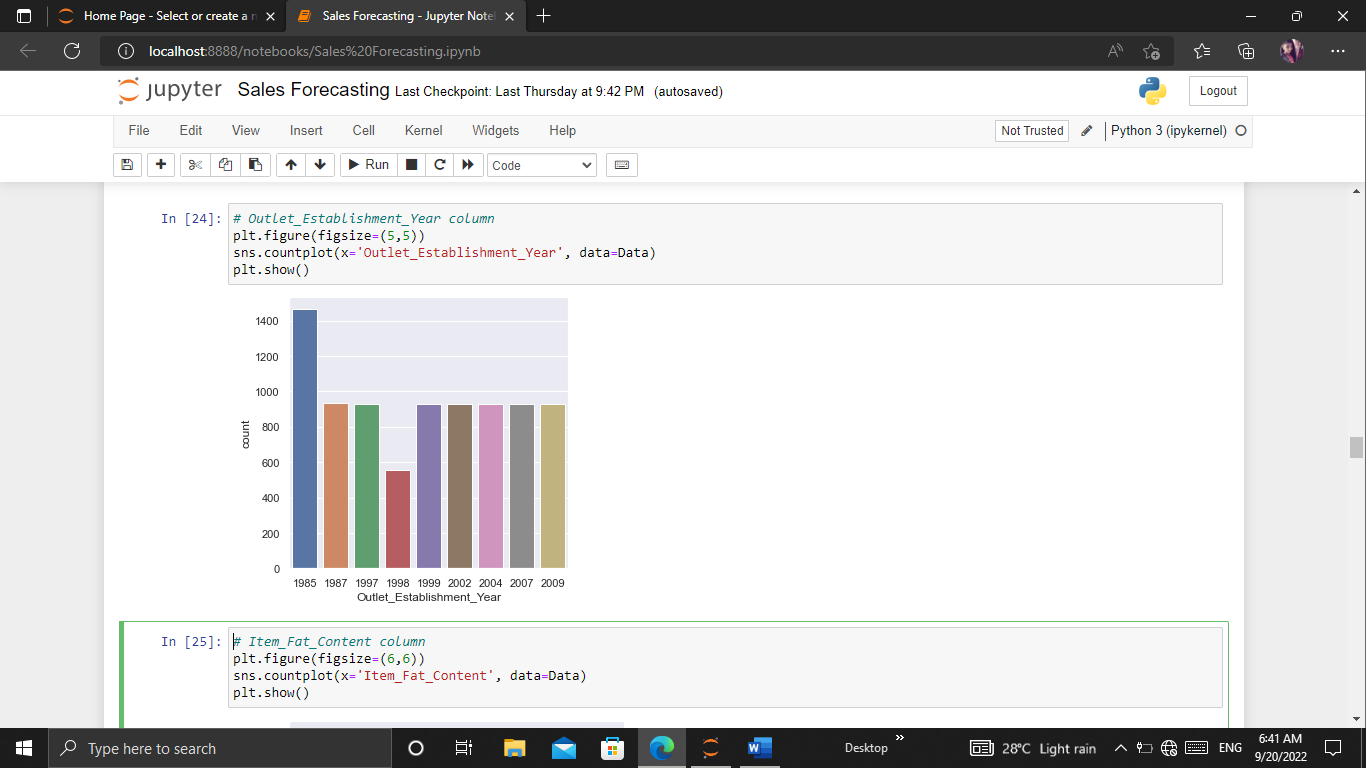


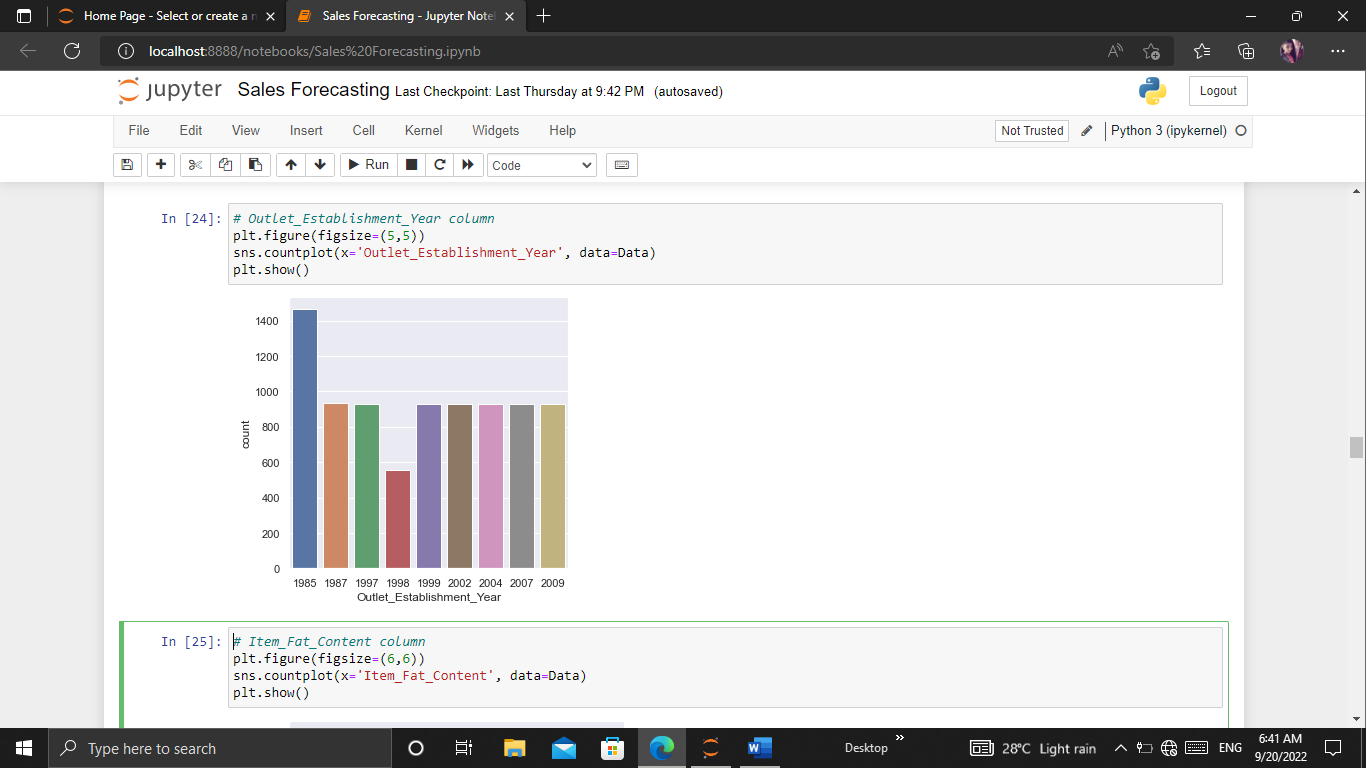
From the above graph, we can observe that we have good amount of products for 50 MRP, 100 MRP ,200 MRP & then we have less products.Hence we have more products in the range of 100 MRP - 180 MRP.



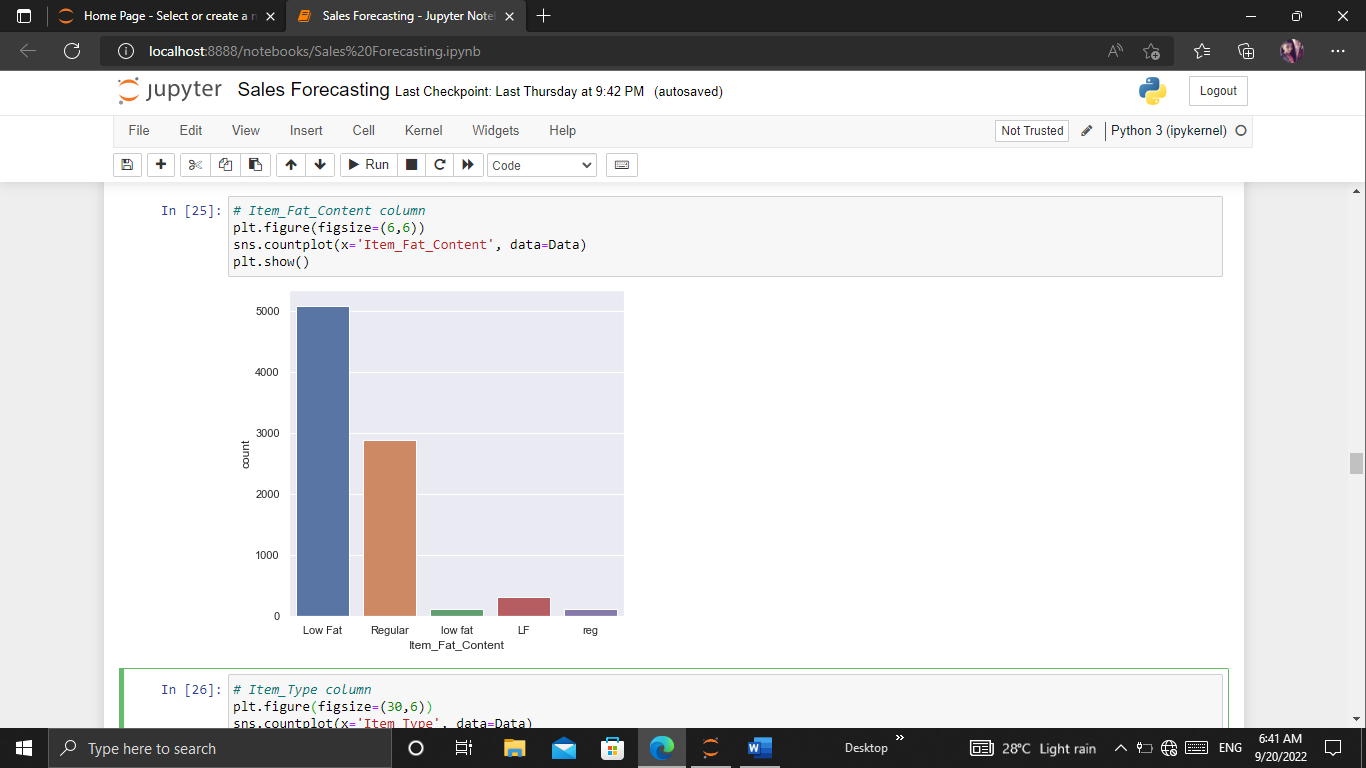


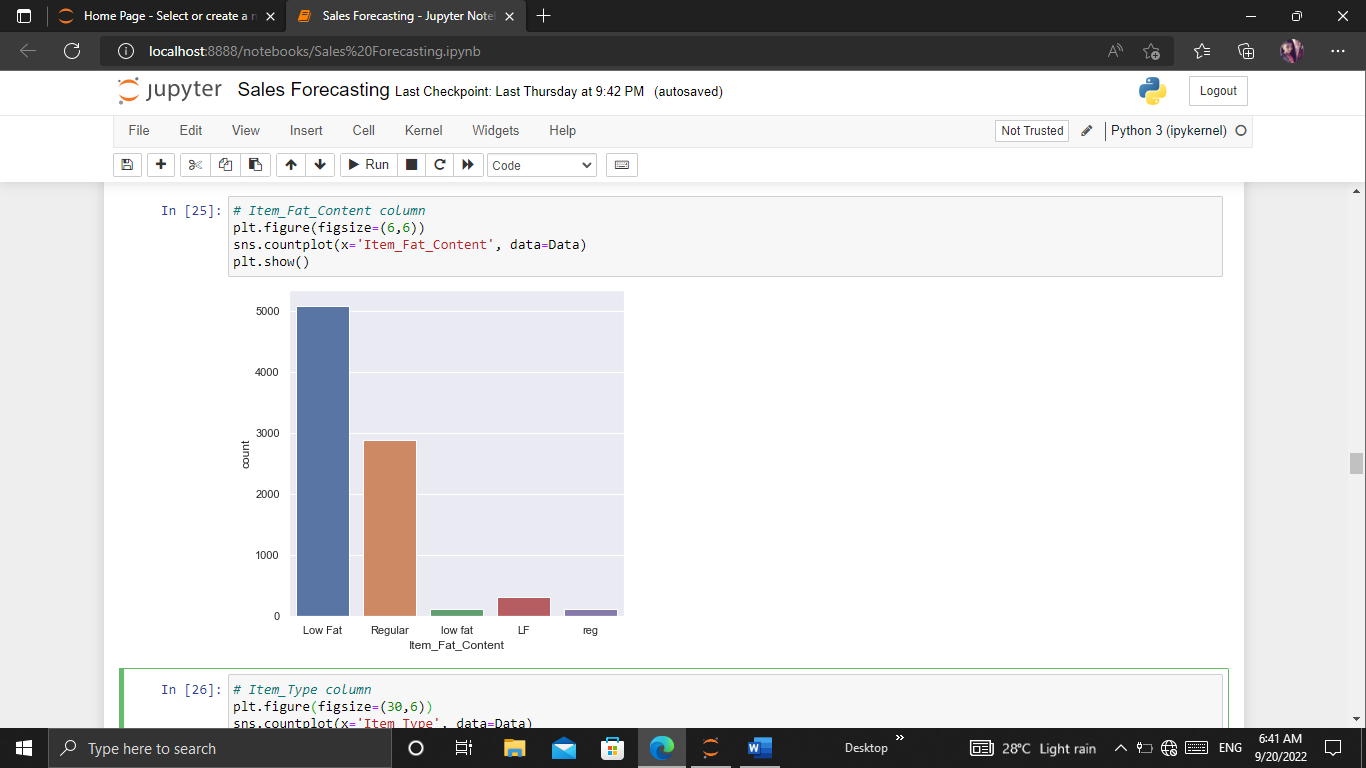
Hence from the above graph we can observe that Item\_Outlet\_Sales feature is positively skewed.





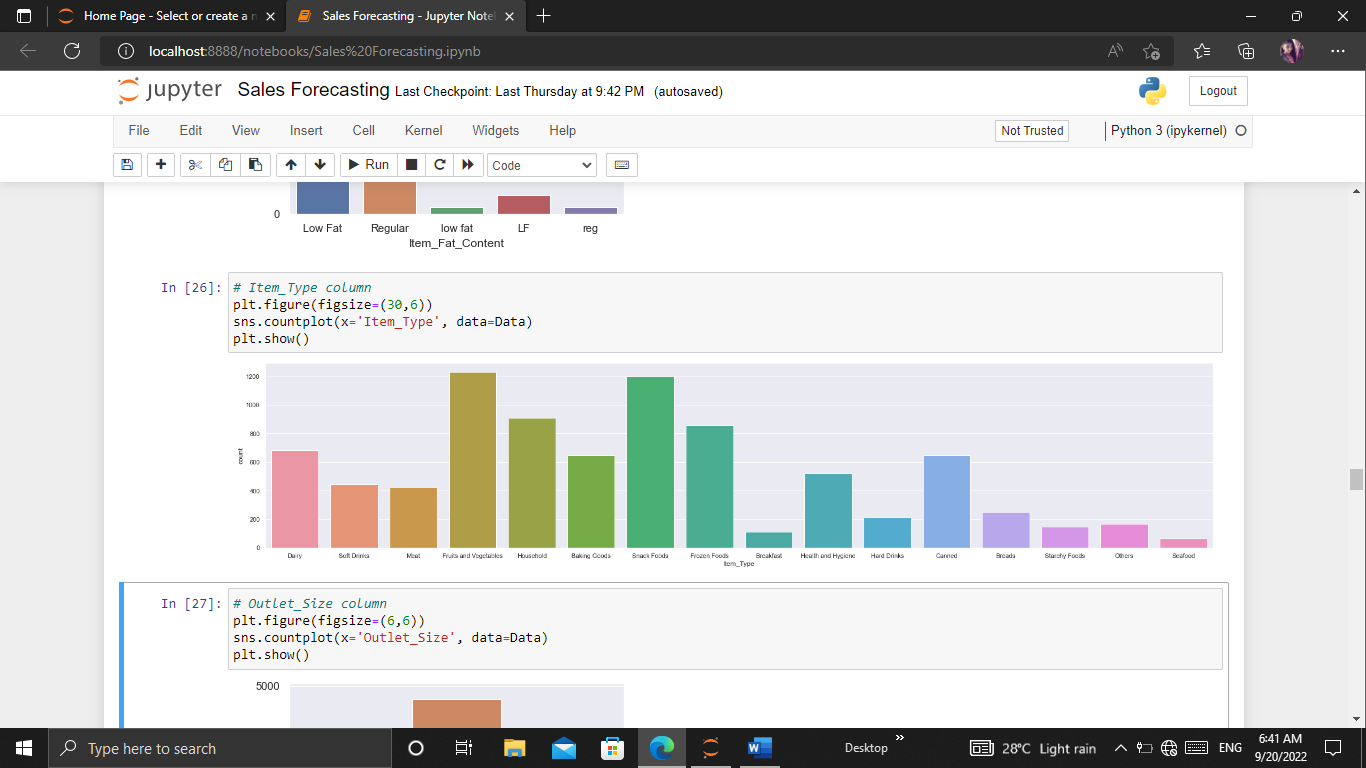
Hence from the above graph we can observe that we have the outlet establishment from the year 1985, 1987 and all the way to 2009.Therefore these are the years on which different outlets or different stores have been established.We can also observe that a lots of stores are established in the year 1985 & less in the year 1998 & all the others years are almost same.

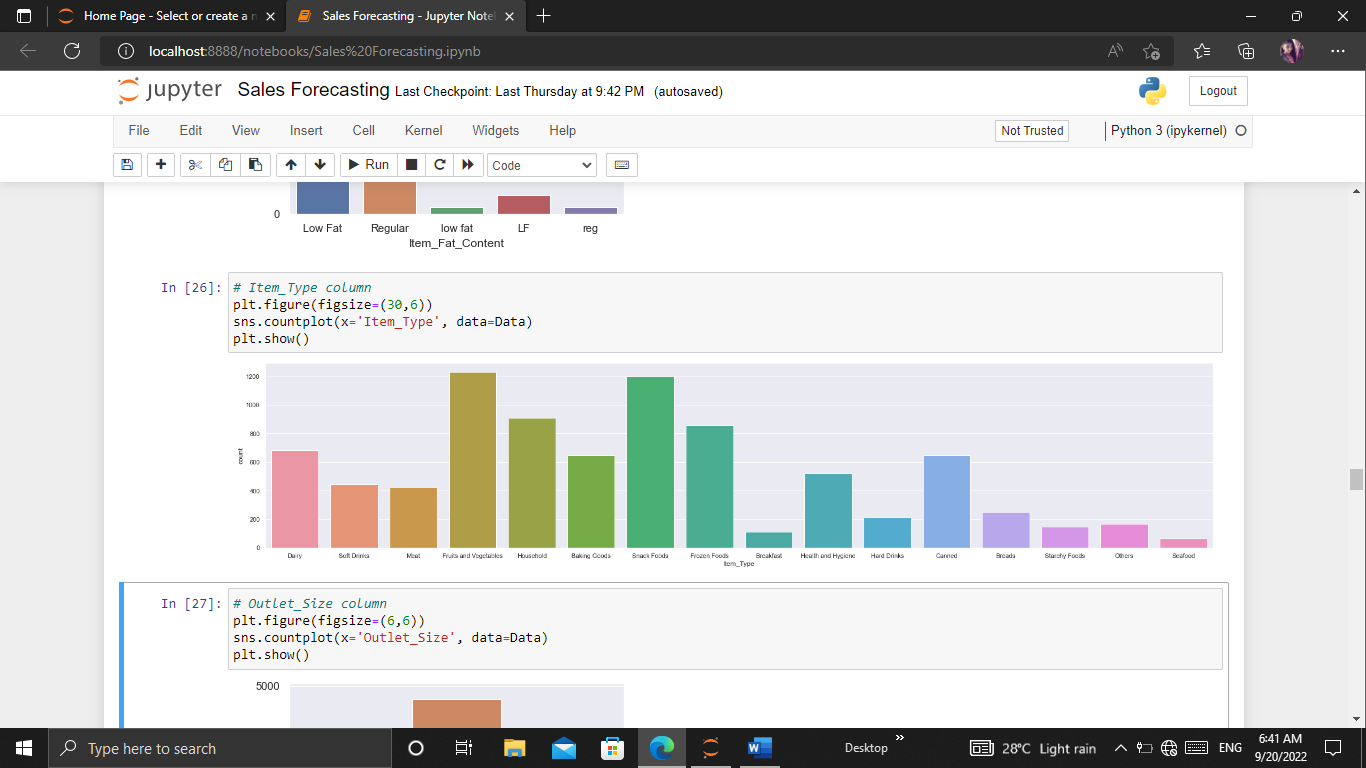




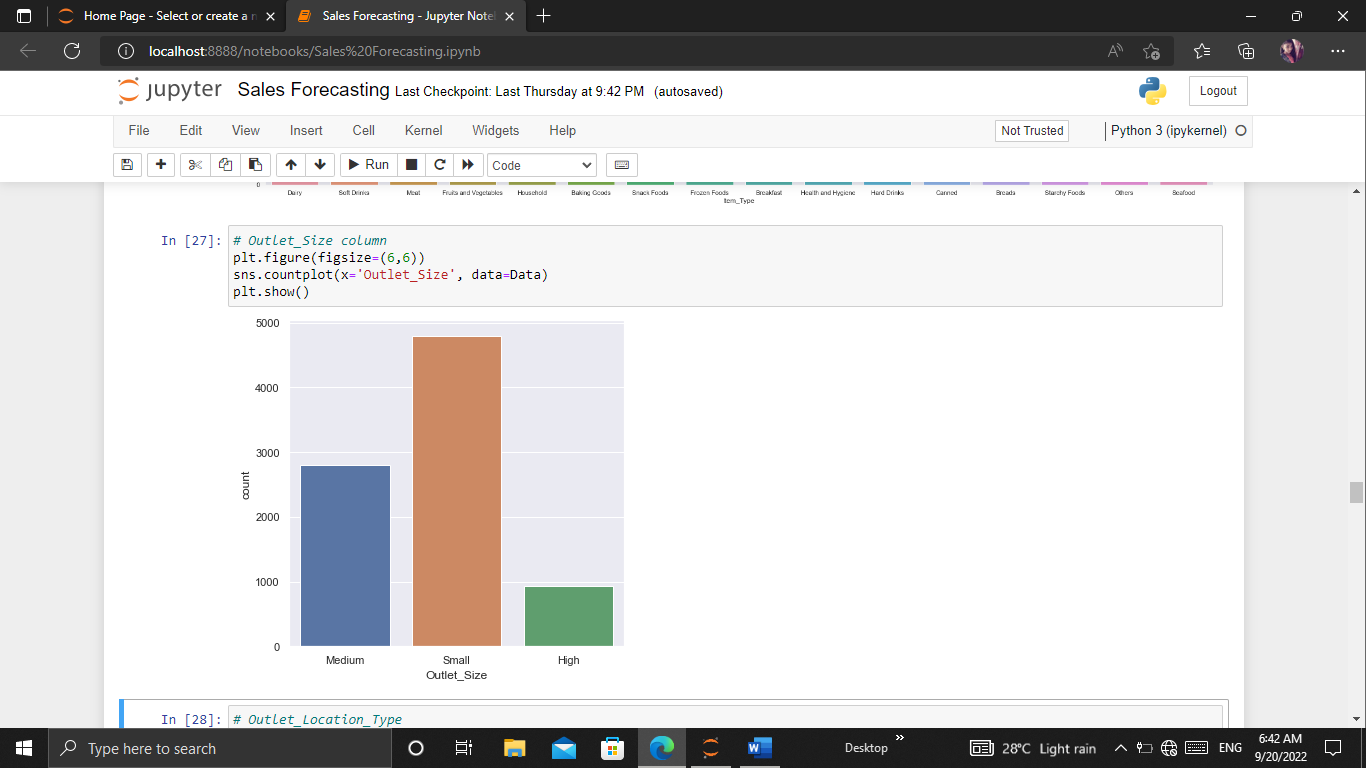
From the above graph we can observe that the data in the Item\_Fat\_Content column has to be cleaned since we have columns such as Low fat,low fat & Lf which is same & must be put into a single particular label.Similarly we have Regular & reg where we need to put this into a single entity.

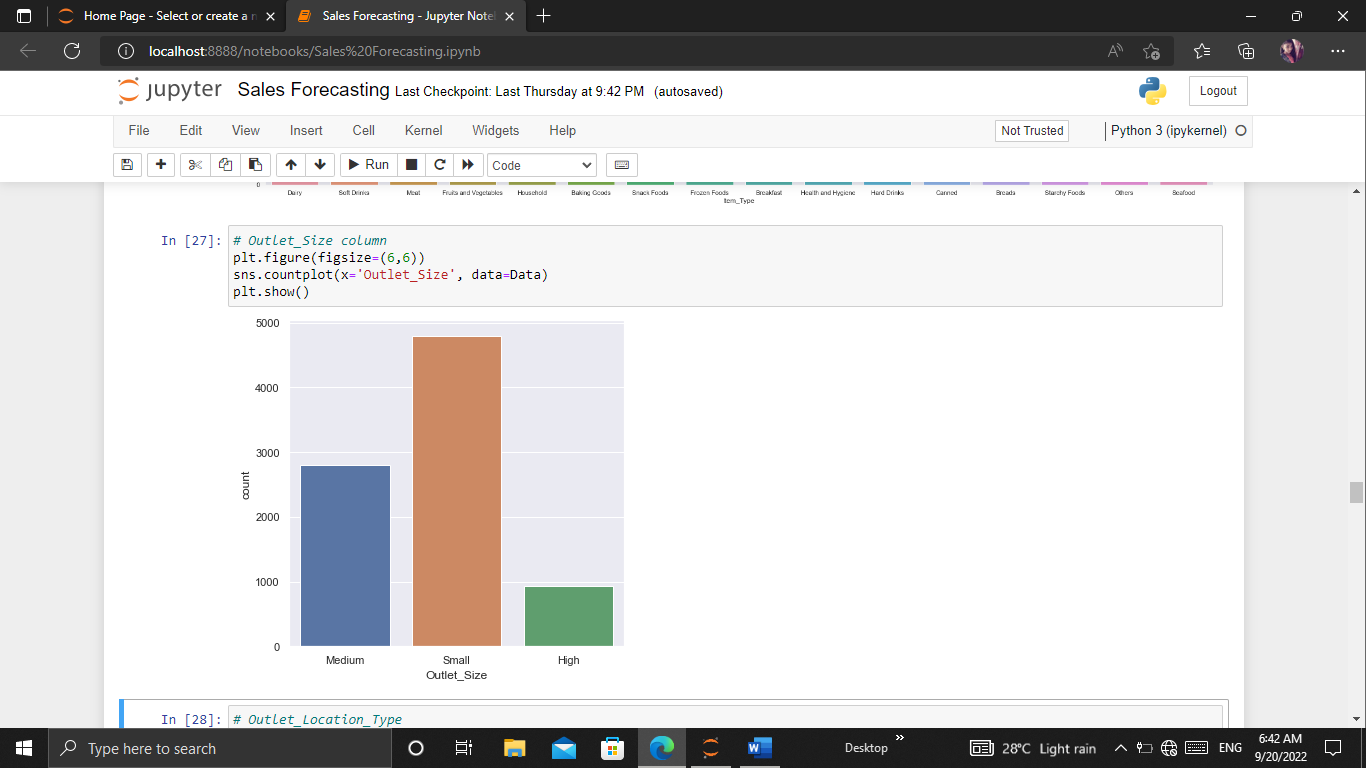
Hence, we need to pre process this data so we will be dealing with this in a later point of time after the visualization of the data.





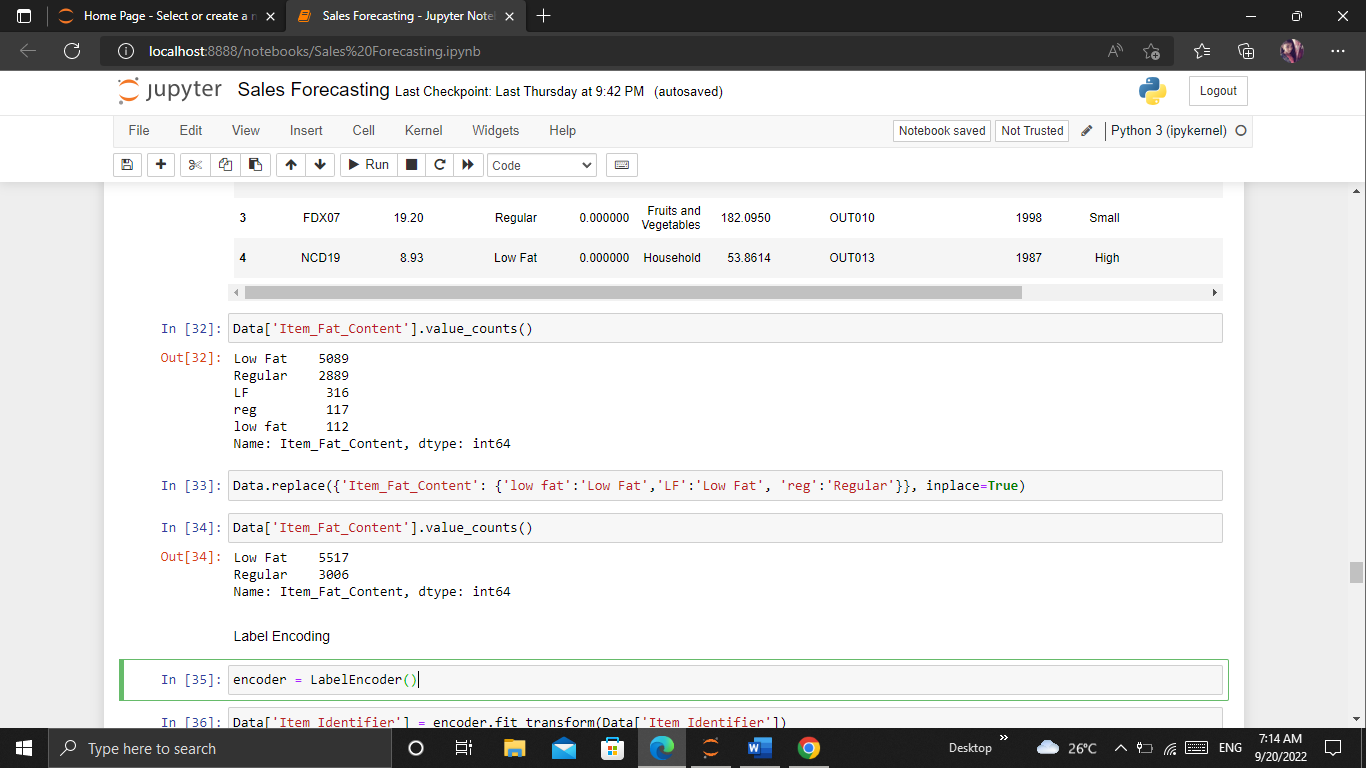
From the above graph we can observe the different items or food types we have such as dairy, soft drinks, meat, fruits & vegetables, household etc. Hence totally we have about 16 Item\_Type values in this case where we have more values in the fruits & vegetables column and snack foods column.





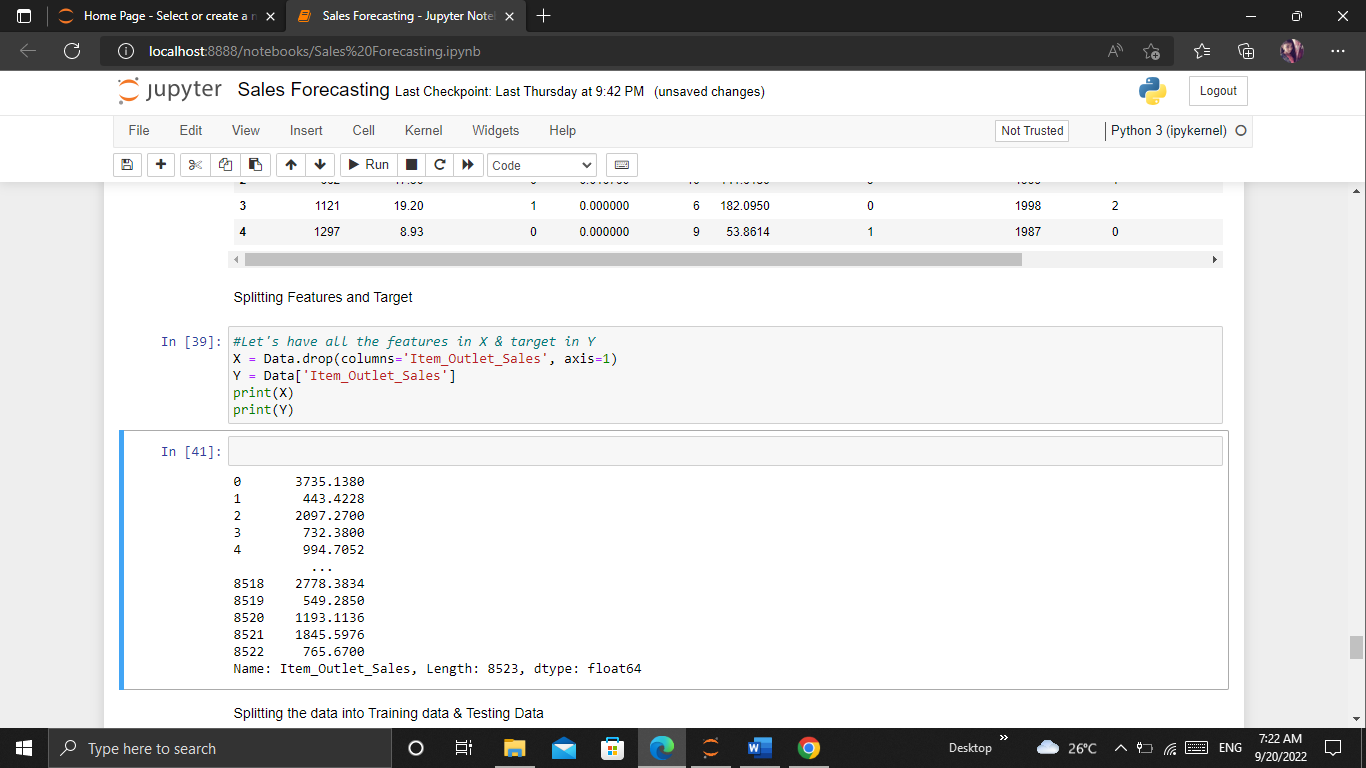
From the above graph, we can observe that we have three outlet\_Size in this case which is medium, small & high

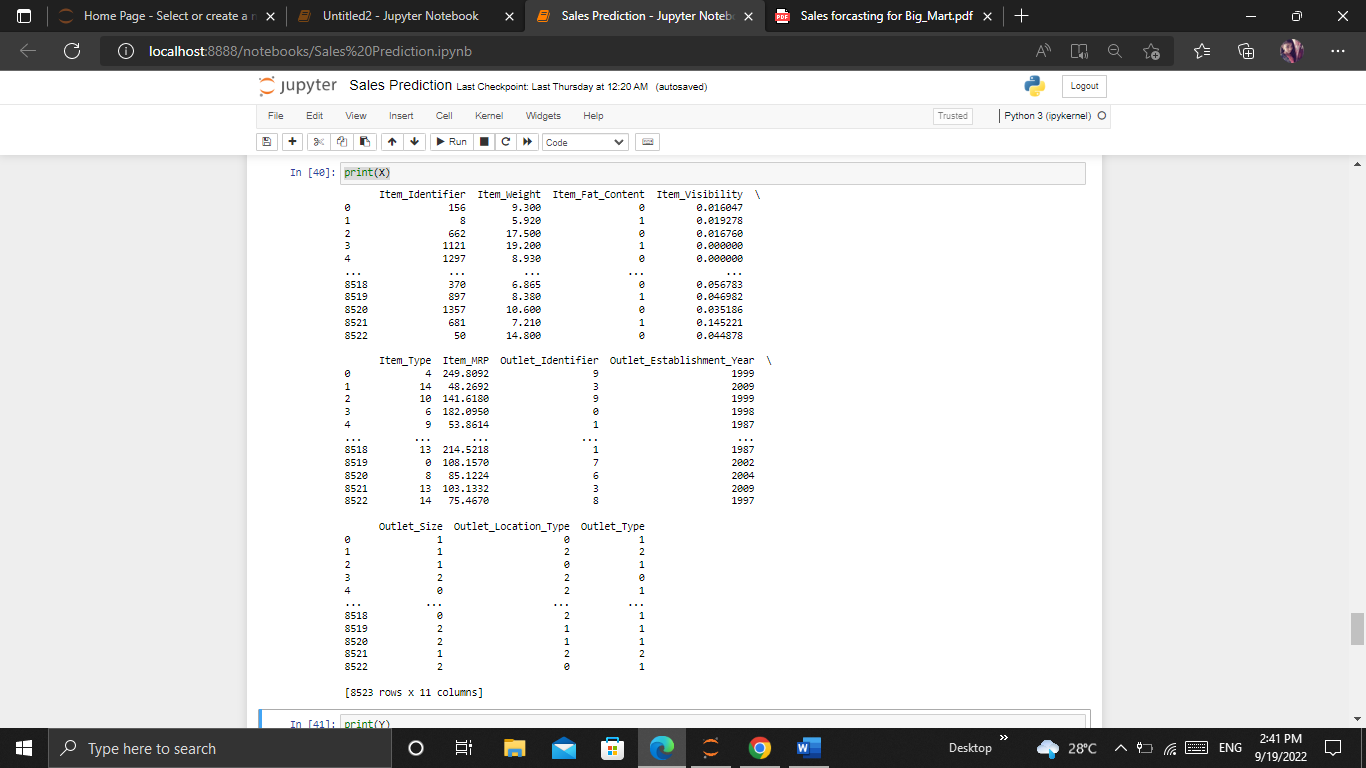
**Preprocessing the data**

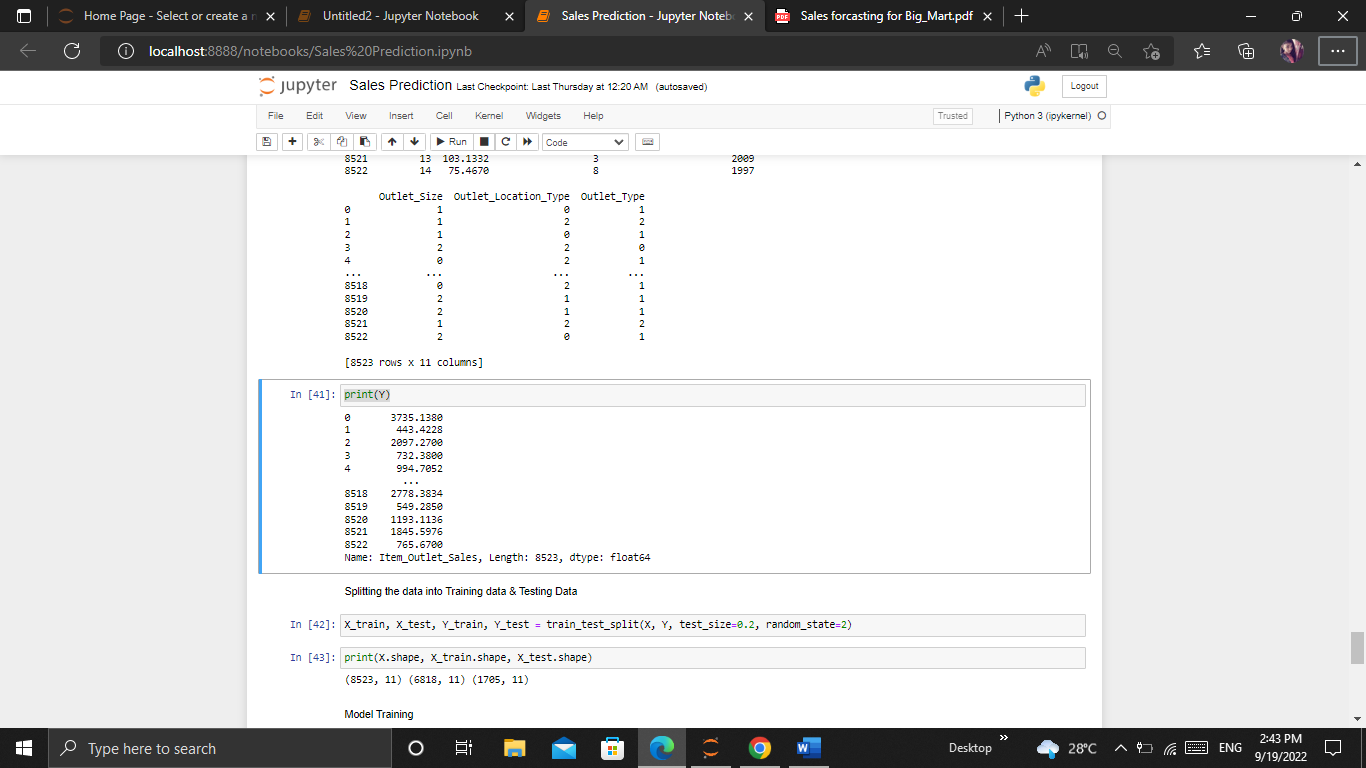


**Splitting Features and Target:**

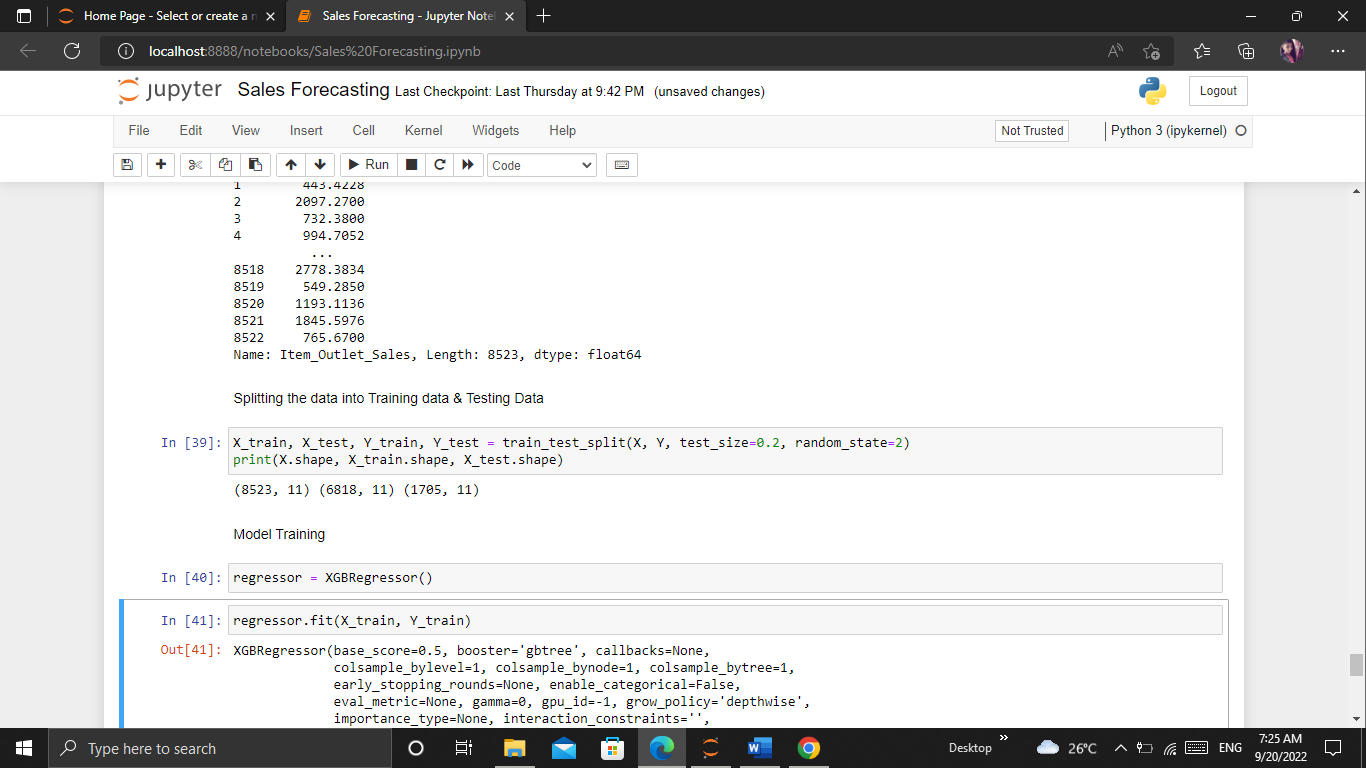
We know that the data in the "Item\_Outlet\_Sales" column is the target& remaining are the features.



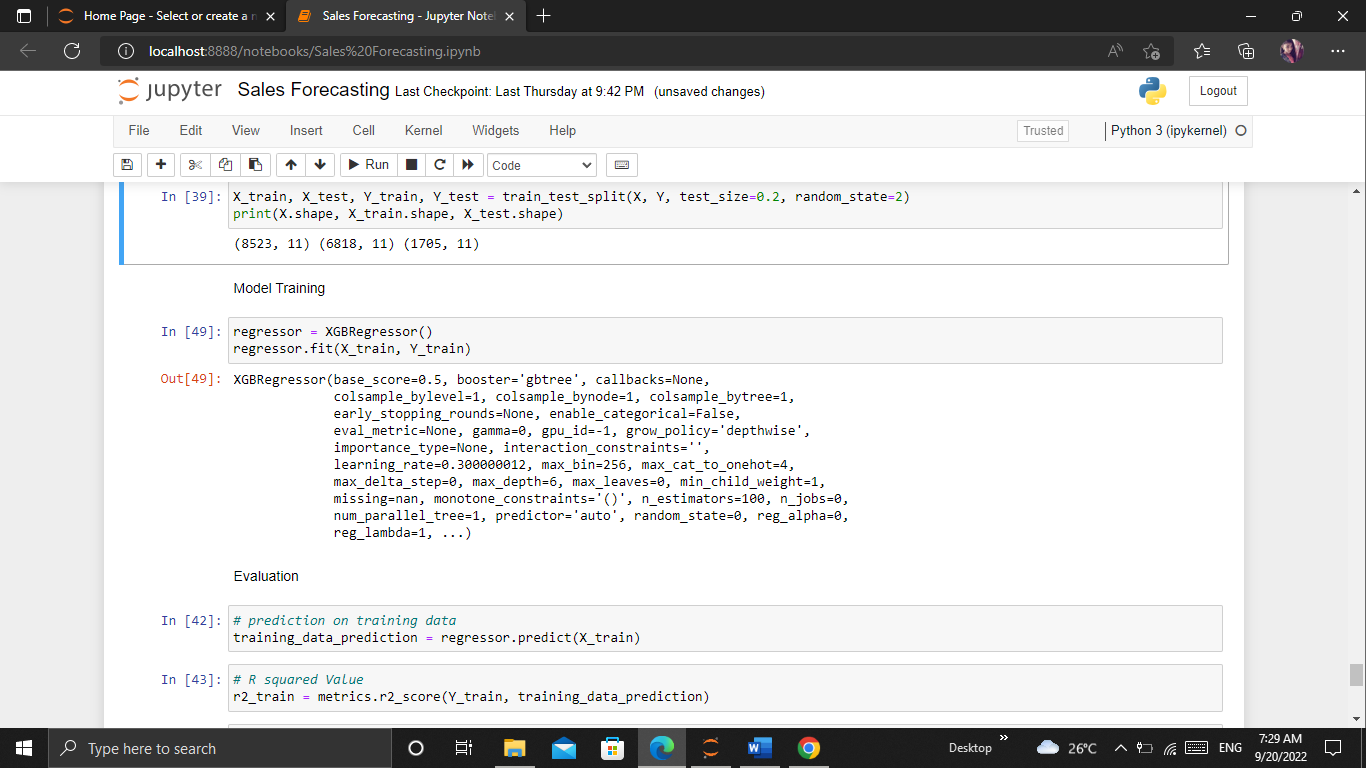




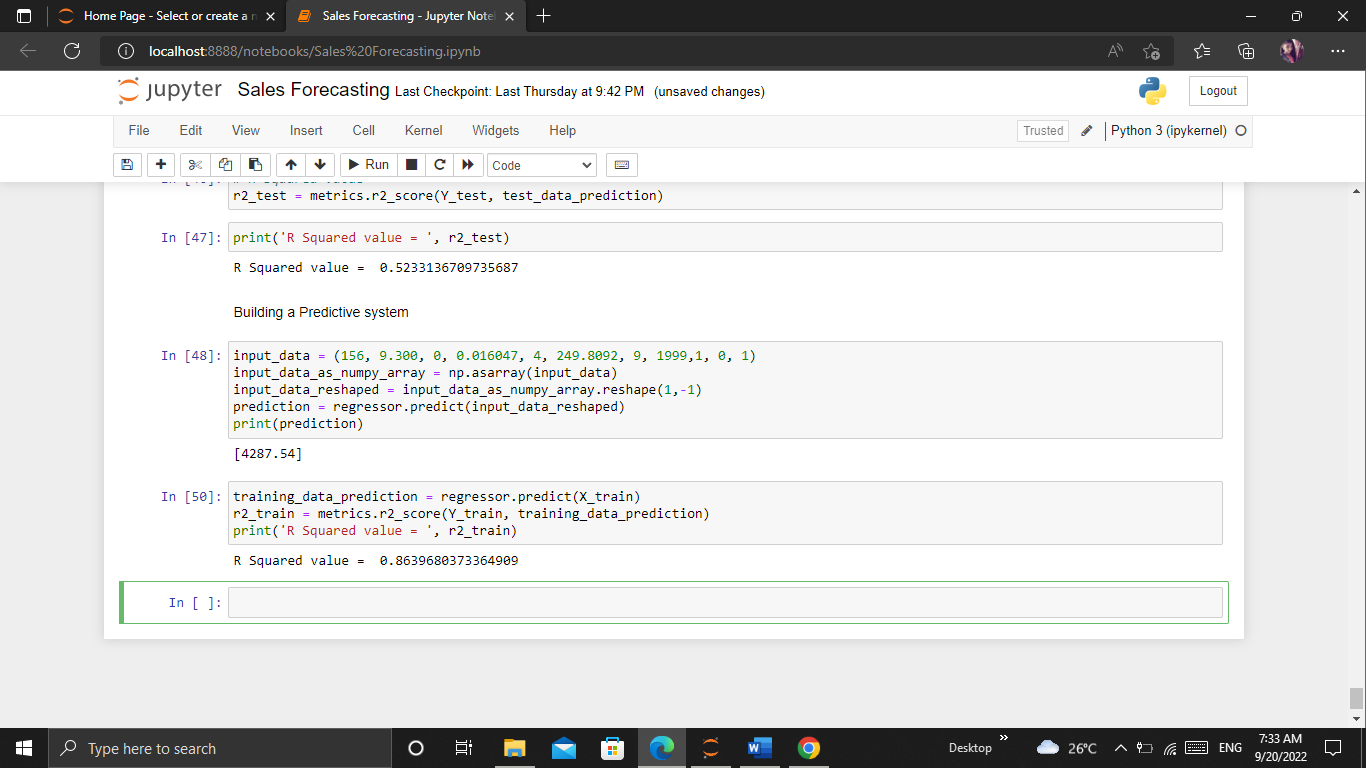
**Splitting the data into Training data & Testing Data:**



Model Training:



**Evaluation:**



**Conclusion:**

Most of the Marts plan to attract the customers to the store and make profit to the maximum extent by them. Once the customers enter the stores if they are attracted then definitely they shop more by the special offers and obtain the desired items which are available in the favorable cost and satisfy them. If the products as per the needs of the customers then it can make maximum profit the retailers can also make the changes in the operations, objectives of the store that cause loss and efficient methods can be applied to gain more profit by observing the history of data in existing stores and has a clear idea of sales can be known.

Sales forecasting plays a vital role in the business sector in every field. With the help of the sales forecasts, will help to estimate the income.