**TITLE OF PROJECT** : SALES PREDICTION ON WALMART DATA

**NAME : RIYA GARG**

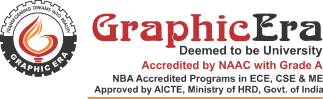
**SECTION : A**

**UNIVERSITY ROLL NO: 2013459**

**CLASS ROLL NO : 47**

**PROBLEM STATEMENT NO: 35**

**SEMESTER: 6**

****

**PROBLEM STATEMENT:**

**Sales Forecasting using Walmart Dataset.**

**DATA DESCRIPTION: The data is taken from Kaggle. According to problem statement:**

You are provided with historical sales data for 45 Walmart stores located in different regions. Each store contains a number of departments, and you are tasked with predicting the department-wide sales for each store.

In addition, Walmart runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of which are the Super Bowl, Labor Day, Thanksgiving, and Christmas. The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks. Part of the challenge presented by this competition is modeling the effects of markdowns on these holiday weeks in the absence of complete/ideal historical data.

stores.csv

This file contains anonymized information about the 45 stores, indicating the type and size of store.

train.csv

This is the historical training data, which covers to 2010-02-05 to 2012-11-01. Within this file you will find the following fields:

* Store - the store number
* Dept - the department number
* Date - the week
* Weekly\_Sales -  sales for the given department in the given store
* IsHoliday - whether the week is a special holiday week

test.csv

This file is identical to train.csv, except we have withheld the weekly sales. You must predict the sales for each triplet of store, department, and date in this file.

features.csv

This file contains additional data related to the store, department, and regional activity for the given dates. It contains the following fields:

* Store - the store number
* Date - the week
* Temperature - average temperature in the region
* Fuel\_Price - cost of fuel in the region
* MarkDown1-5 - anonymized data related to promotional markdowns that Walmart is running. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA.
* CPI - the consumer price index
* Unemployment - the unemployment rate
* IsHoliday - whether the week is a special holiday week

For convenience, the four holidays fall within the following weeks in the dataset (not all holidays are in the data):

Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13  
Labor Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13  
Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13  
Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

**MOTIVATION:** The sales forecast indicates as to how much of a particular product is likely to be sold in a specified future period in a specified market at speci­fied price. Accurate sales forecasting is essential for a business house to enable it to produce the re­quired quantity at the right time. Walmart is one of the largest retailers in the world and it is very important for them to have accurate forecasts for their sales in various departments. Since there can be many factors that can affect the sales for every department, it becomes imperative that we identify the key factors that play a part in driving the sales and use them to develop a model that can help in forecasting the sales with some accuracy. In data science ,sales forecasting is the most common problem that a data scientist perform as a part of day to day activities.

**LANGUAGE USED: PYTHON**

**TOOL USED: Voila, Jupyter notebook**

**Libraries used: pandas, seaborn, matplotlib, NumPy, sklearn, datetime**

**DATA RESOURCE: KAGGLE**

**Project:**

**METHODOLOGY:**

**Step1: Import all necessary libraries as specified above with desired modules/functionalities.**

**Step2: Import the data that we want to work on (three files features.csv, train.csv, stores.csv)**

**Step3: Merge the datasets to make it a single dataset to work on.**

**Step 4: Cleaning the data**

**1.removing the null values**

**2. deleting the repeating column IsHoliday due to merging datasets.**

**3. group different holidays**

**4. deleting negative values of weekly sales.**

**5. Conversion of date into week, month and year.**

**Step 5: Perform exploratory data analysis**

**Major conclusions obtained were that:**

Data is in between 05-02-2010 to 26-10-2012

1. Sales on holidays is more than on non-holidays.
2. Weekly sales are not increased on Christmas which shows that people shop before Christmas and tend to enjoy on holiday rather than shopping.
3. There is positive effect on sales in Super bowl day.
4. There is a significant increase in the average weekly sales on Thanksgiving Day may be due to black Friday sale.
5. The 'type' feature also seems to have a certain relationship with Weekly Sales. Type 'A' has a higher sales median than types 'B' and 'C'.  Type 'C' tends to have lower weekly sales.
6. Some stores have higher sales specifically average store 20 followed by 4th store. Some stores showed a decrease in sales over the years, such as stores 14, 27, 35 and 36.
7. Weekly sales by department are even more irregular, with departments with average sales ranging from 0 to more than 70000. Despite this discrepancy in weekly sales by department, this behavior seems to remain stable over the years. Some departments showed a decrease in sales over the years, such as departments 18, 65 and 73.
8. 92 has highest weekly sale doing department
9. There are no values for CPI between 140 to 180
10. There is no significant pattern in unemployment, fuel price, cpi, temperature on weekly sales
11. On Grouping weekly sales by store size, it seems to indicate a certain trend towards higher sales for larger stores.
12. For all holidays Type A stores has highest sales.
13. it is seen that 2011 has lower sales than 2010 generally. When we look at the mean sales it is seen that 2010 has higher values, but 2012 has no information about November and December which have higher sales.
14. The best Weekly sales are in December and November.
15. It was seen that 51th week and 47th weeks have significantly higher averages of sales.

Step 6: Finding correlations between the data attributes and dropping the not desired attributes before training the model.

Step7: After dropping Temperature, MarkDown4, MarkDown5, CPI, Unemployment again calculate the correlation matrix.

Step 8: Train the model using random forest model

Step 9: calculate the score and apply wmae test on it.

Step 10: Then use feature selection to get better accuracy and less value wmae.

Step 11: Plot the graph for actual vs predicted values.

Step 12: Store the predicted values in submission.csv

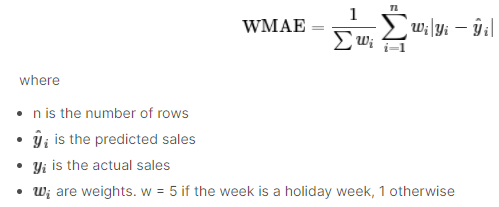
Step 13: Then we can convert this notebook into a dashboard using voila tool and deploy the dashboard on cloud.

Accuracy of the model achieved:

85.3

Best value of wmae(weighted mean absolute error) achieved: 19.74

WMAE TEST(weighted mean absolute error): The metric of the competition is weighted mean absolute error (WMAE). Weight of the error changes when it is holiday.



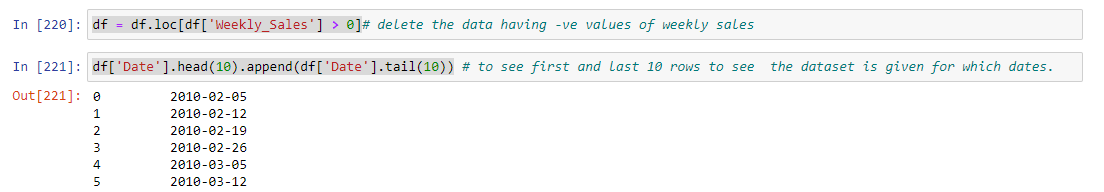
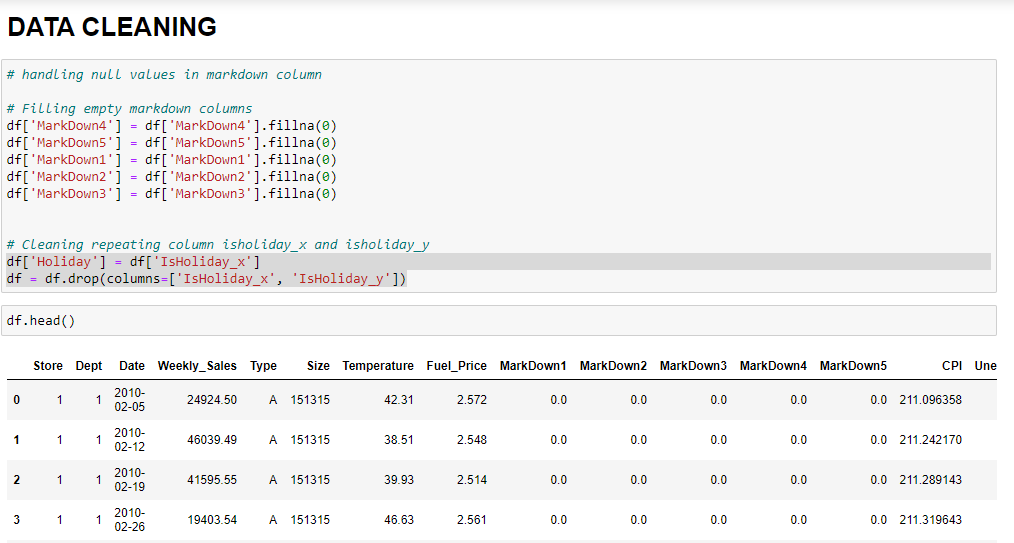
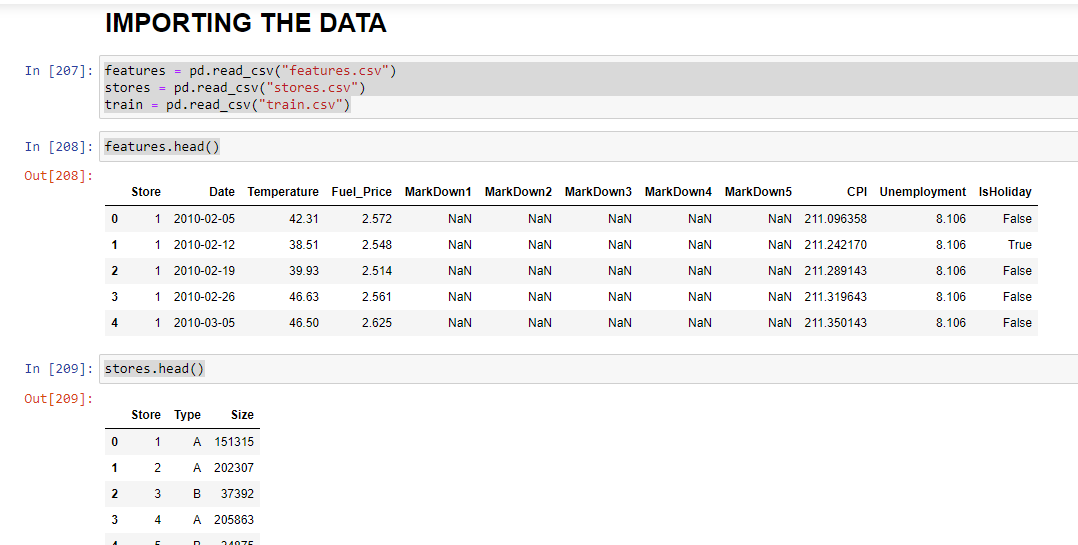
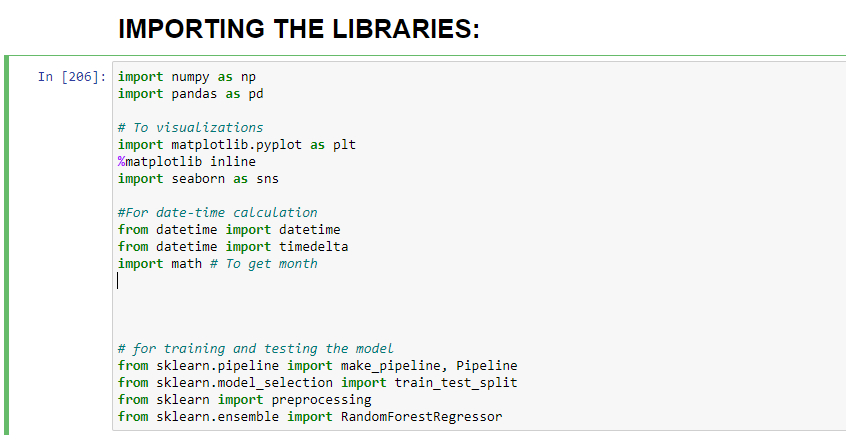
RANDOM FOREST:

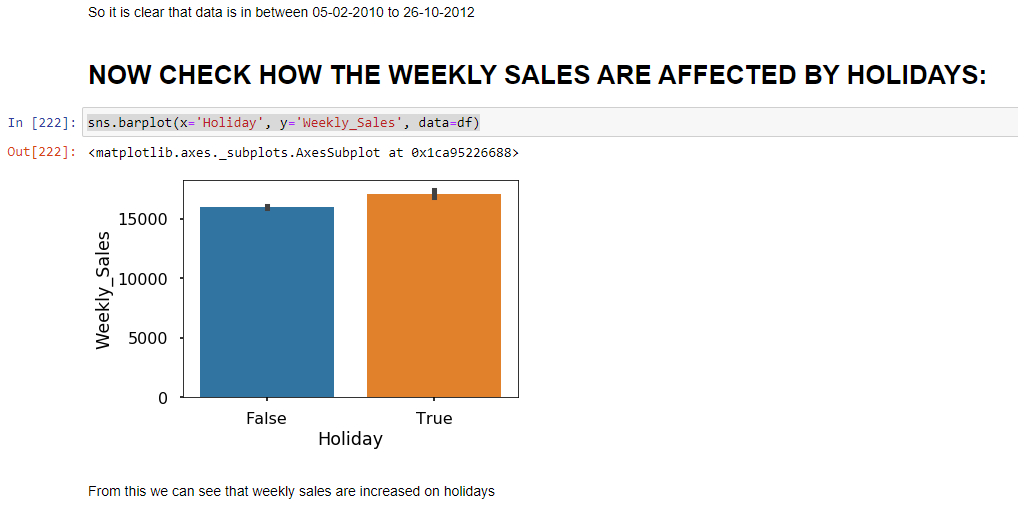
Random forest is a bagging technique and not a boosting technique. The trees in random forests are run in parallel. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

**NEW THINGS LEARNT:**

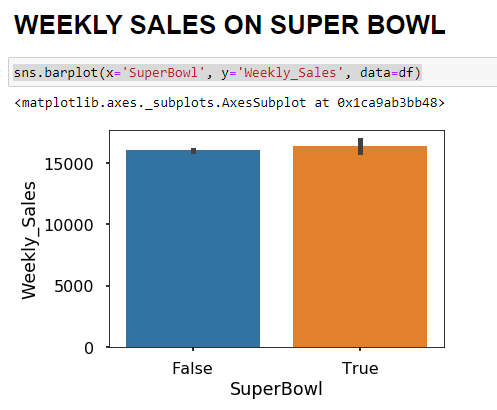
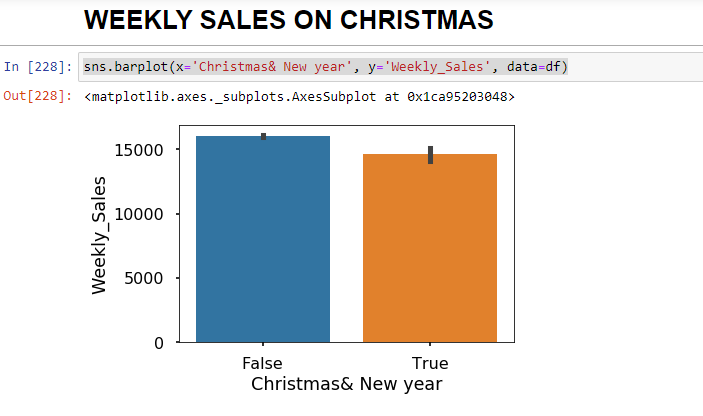
* **Python**
* **Voila**
* **Data visualization through matplotlib, seaborn**
* **Wmae test**
* **Random forest algorithm with feature selection**

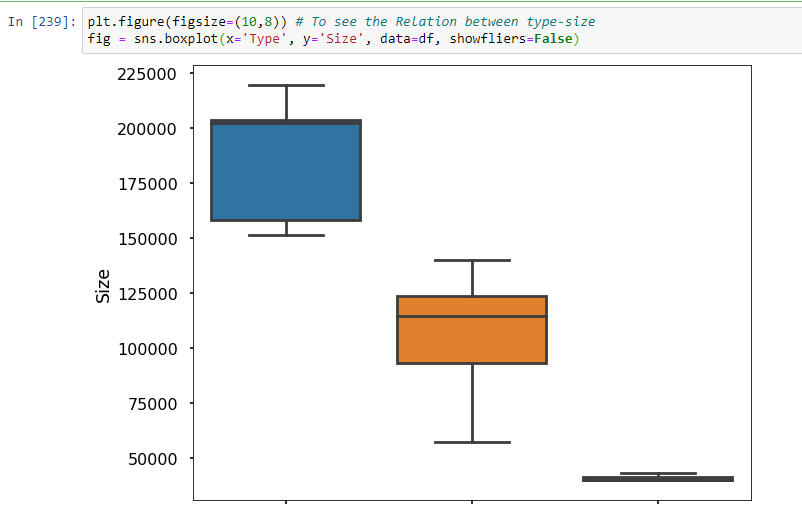
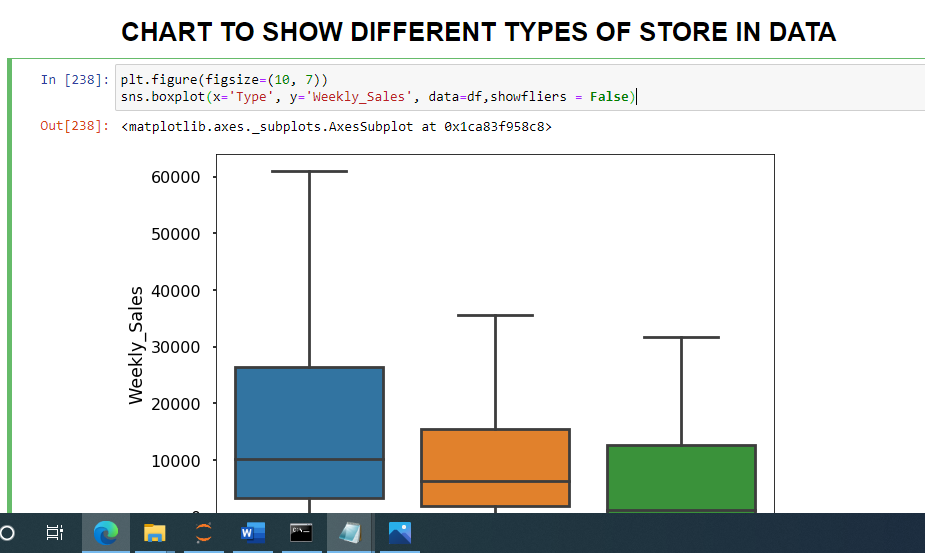
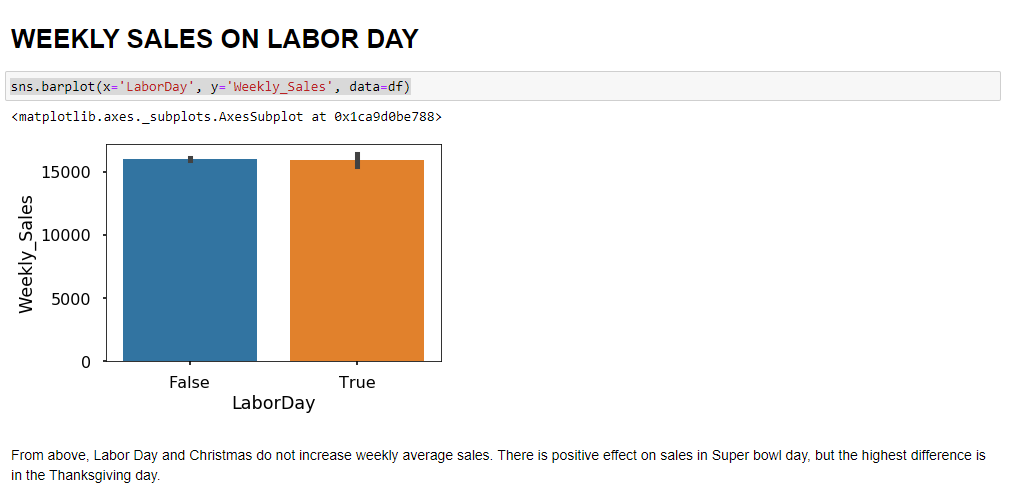
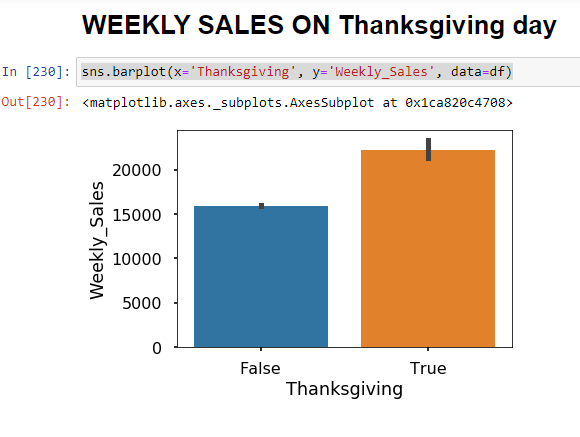
**OUTPUT AND SOURCE CODE:**

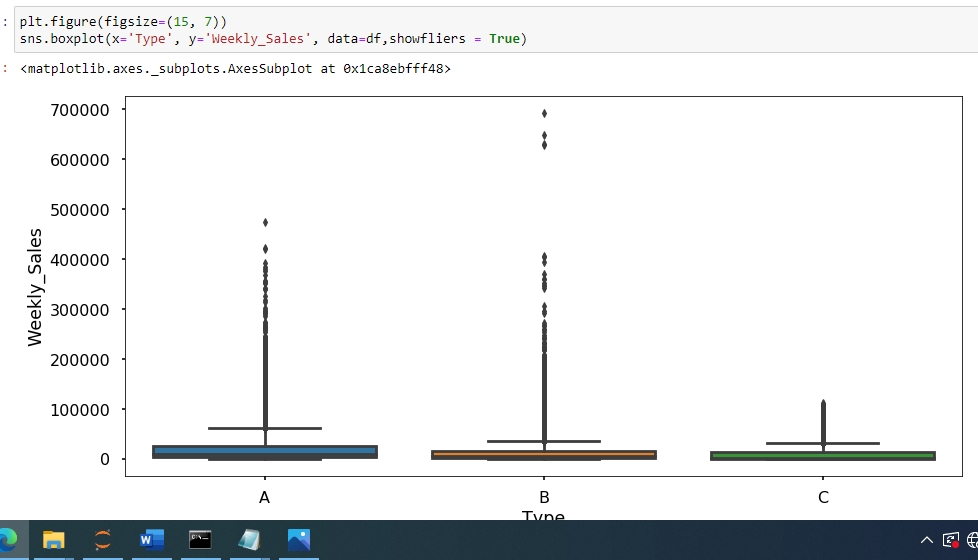


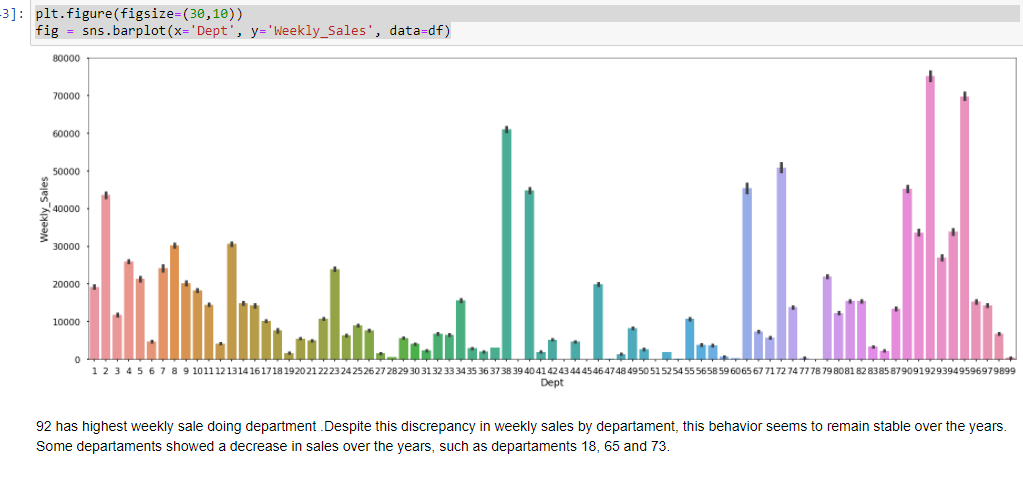
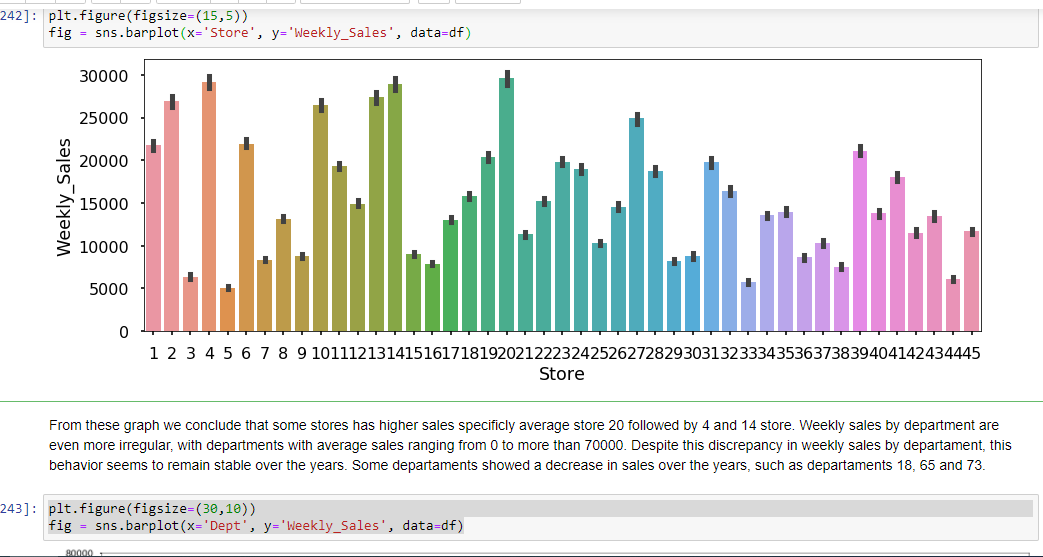


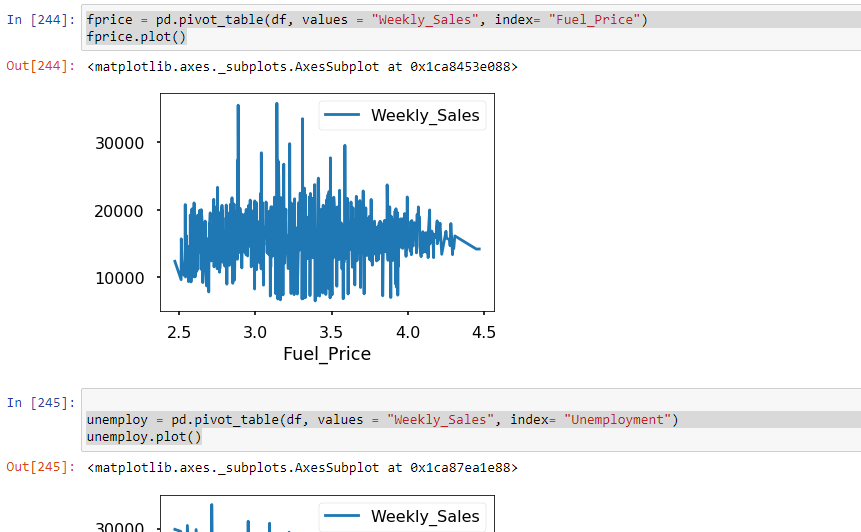


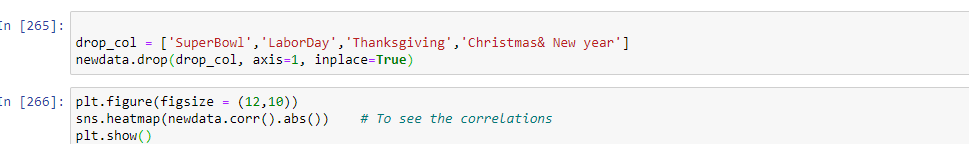
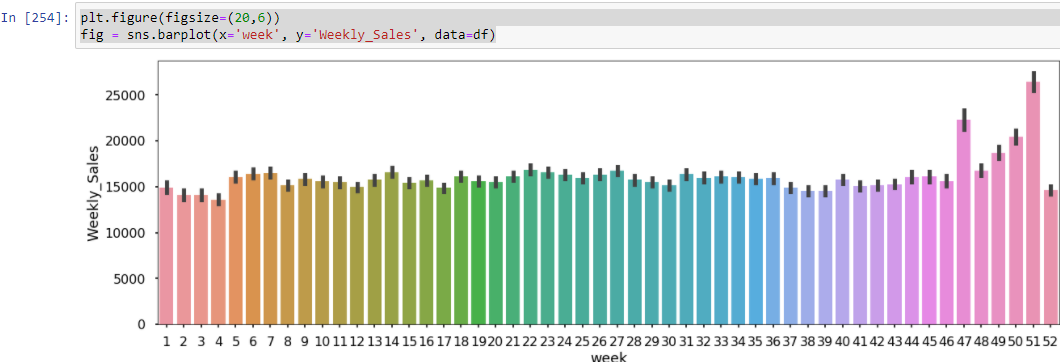
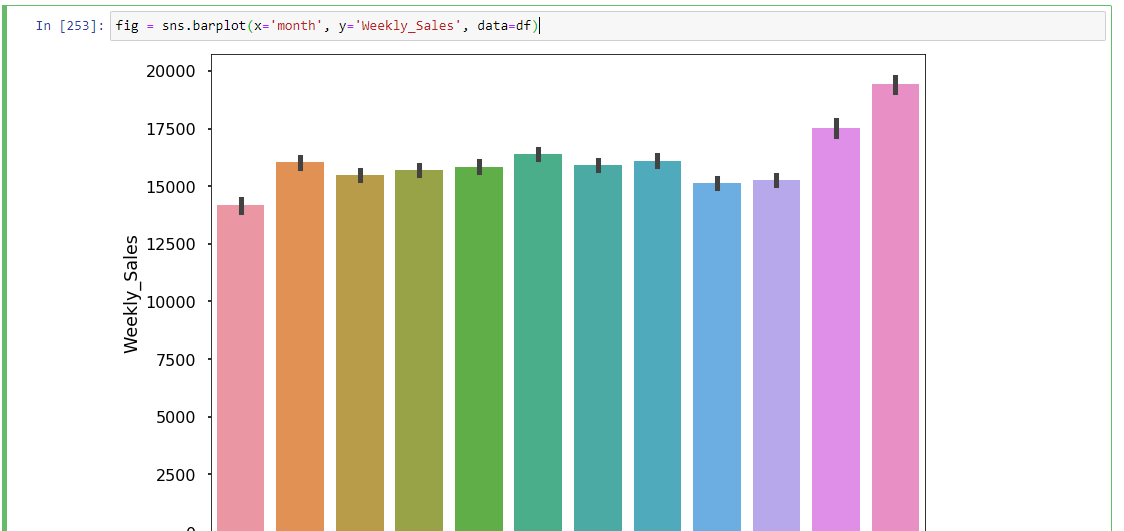
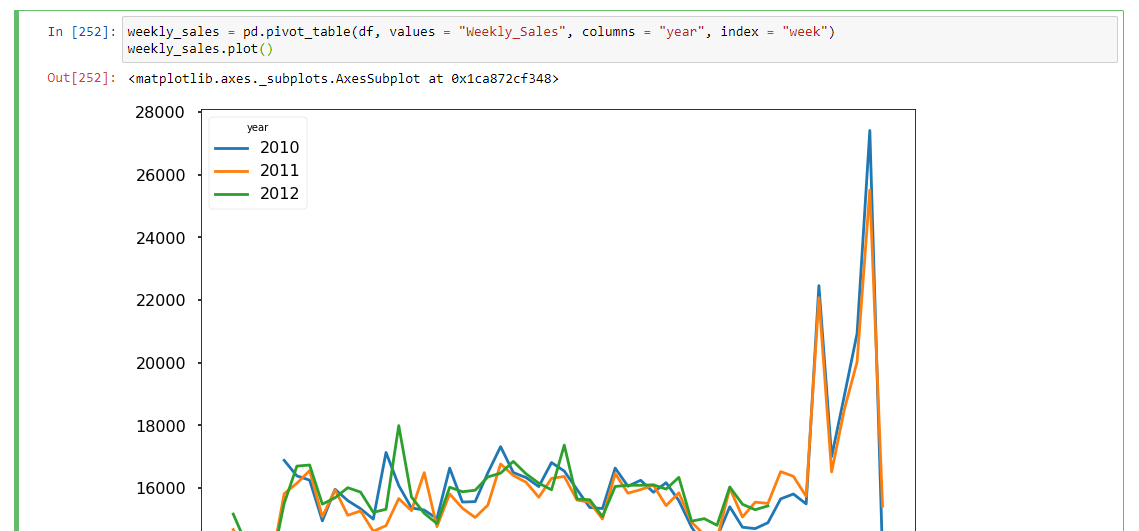
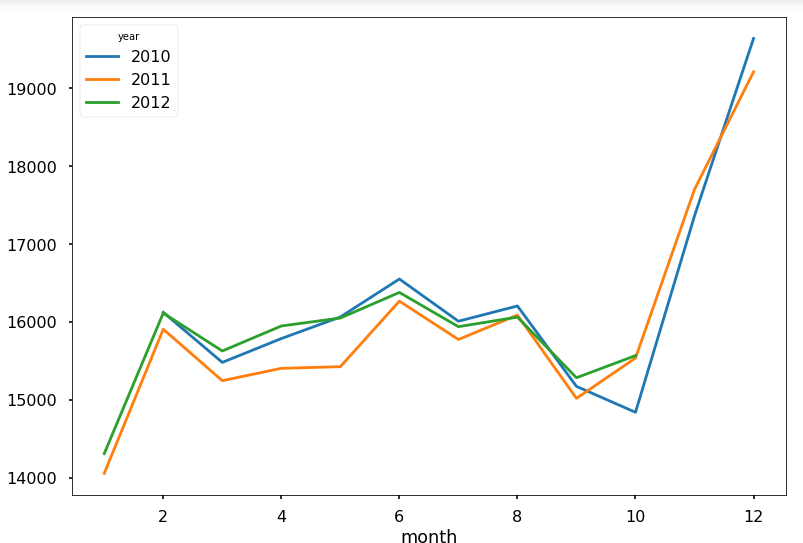
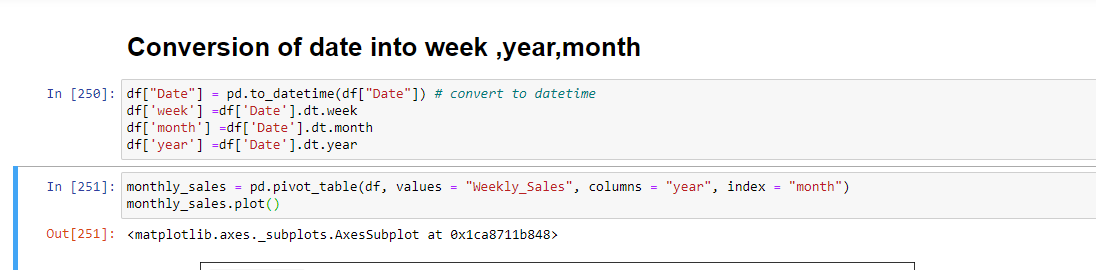
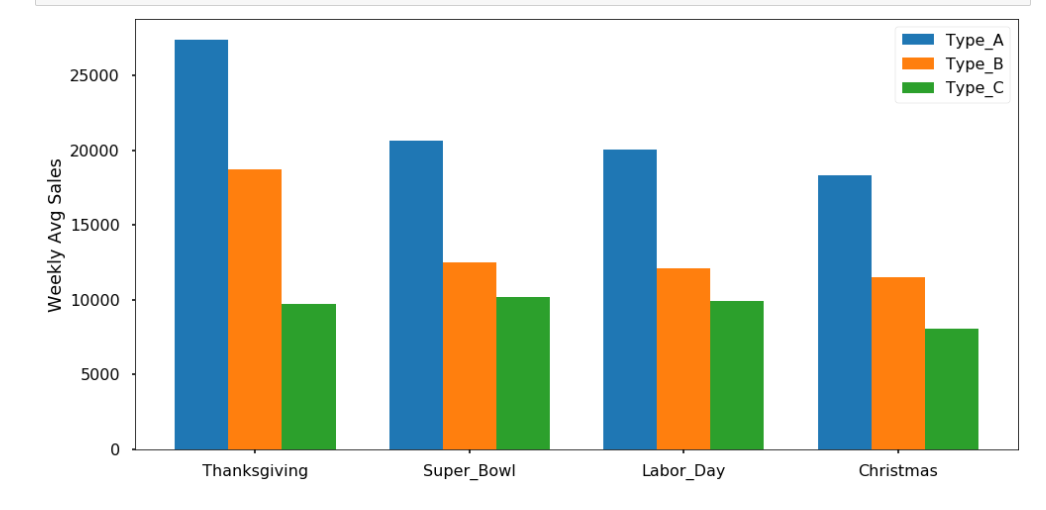
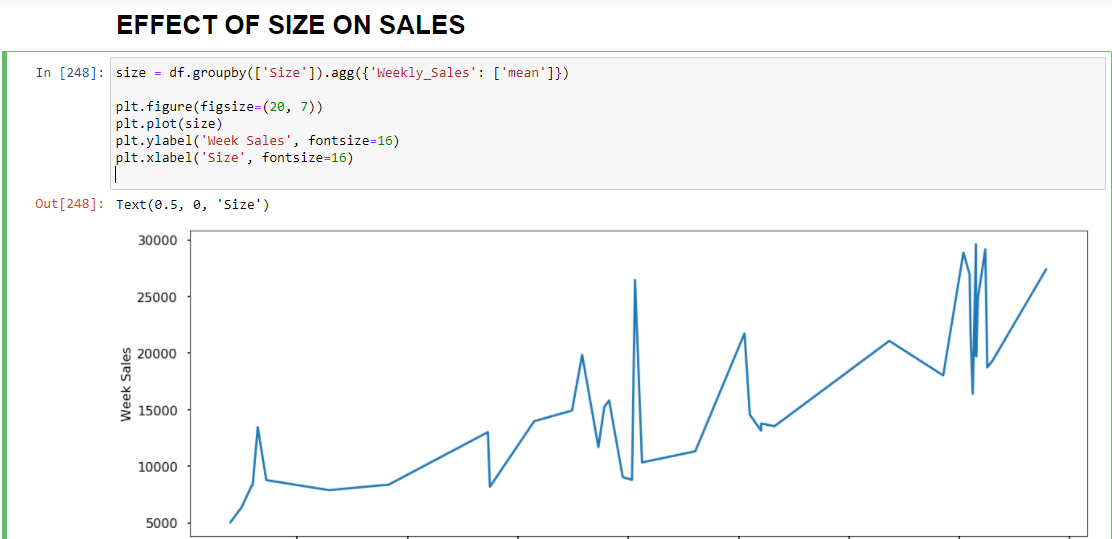
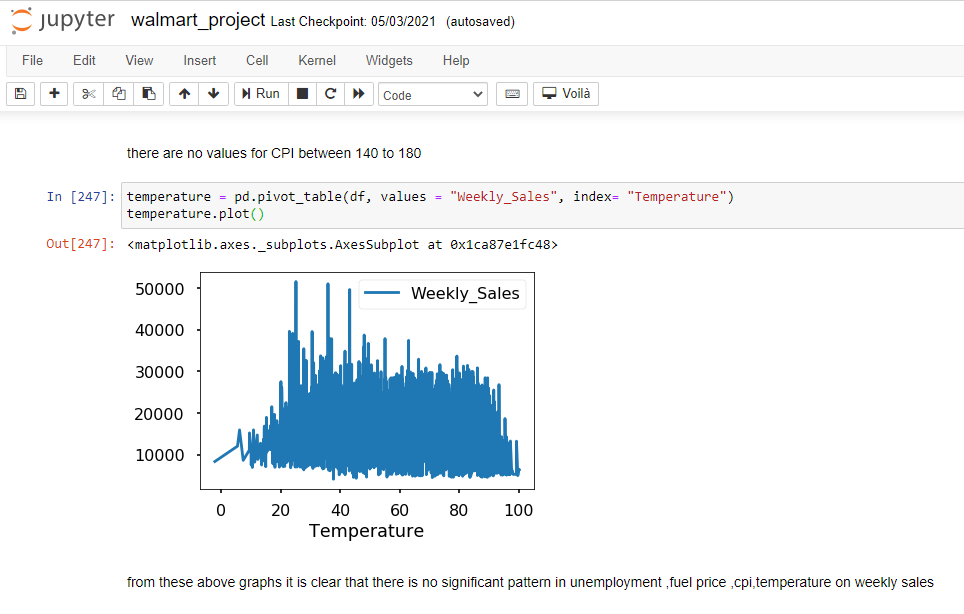
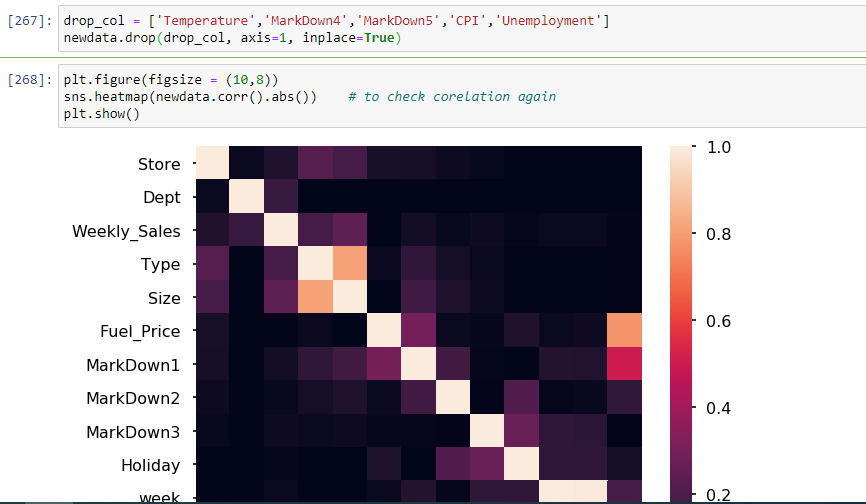
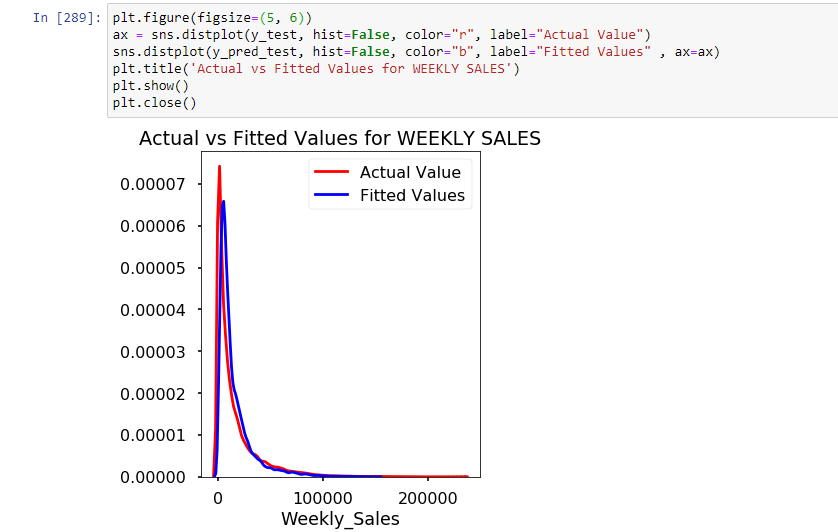
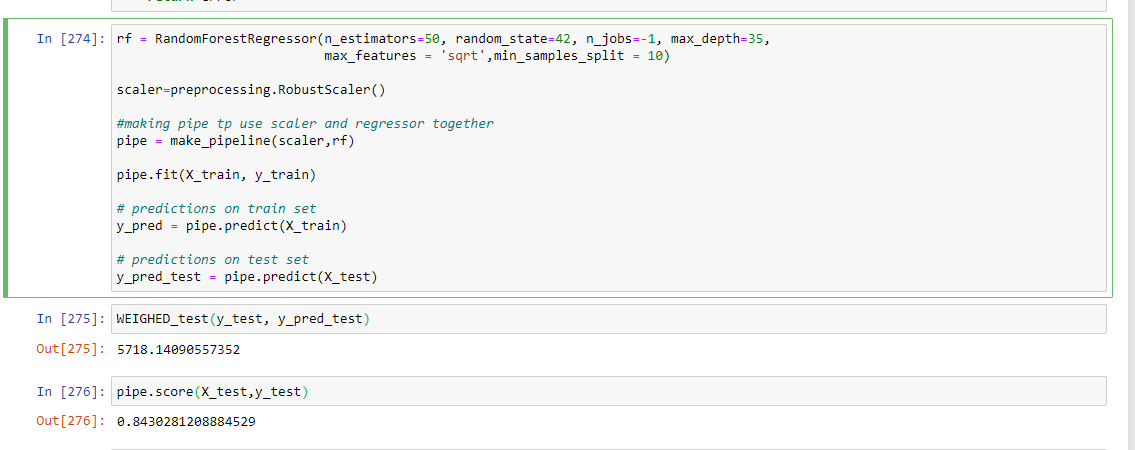
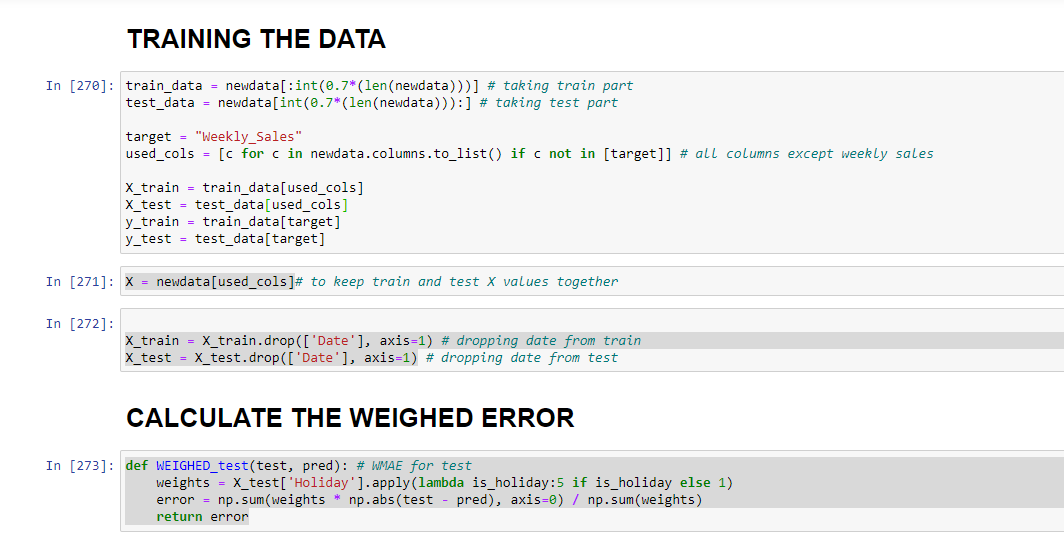










AFTER APPLYING FEATURE IMPORTANCE TO WHOLE DATASET:

