**Retail Analysis with Walmart Data**

This project is to predict sales and demand accurately.

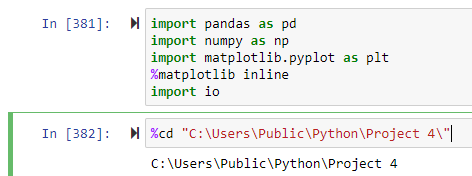
Dataset Description

This is the historical data that covers sales from 2010-02-05 to 2012-11-01, in the file Walmart\_Store\_sales.

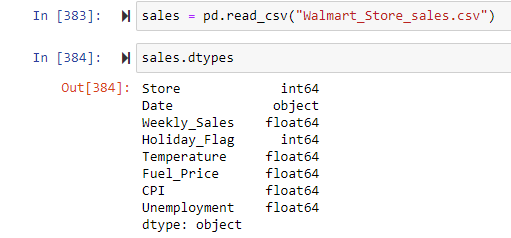
sales is the dataframe we are using. We have the following fields:

* Store - the store number
* Date - the week of sales
* Weekly\_Sales - sales for the given store
* Holiday\_Flag - whether the week is a special holiday week 1 – Holiday week 0 – Non-holiday week
* Temperature - Temperature on the day of sale
* Fuel\_Price - Cost of fuel in the region
* CPI – Prevailing consumer price index
* Unemployment - Prevailing unemployment rate

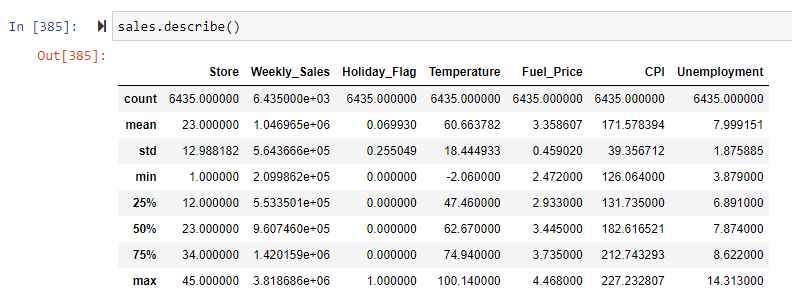
1., We import all the necessary files and we set the directory.



2. Obtaining the data into pandas dataframe and checking the datatypes of the variables.



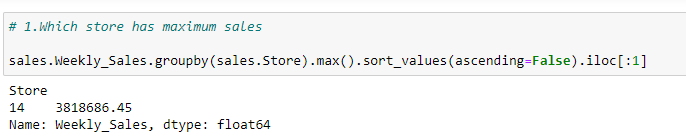
3. Description of the data



Tasks

1. Which store has maximum sales

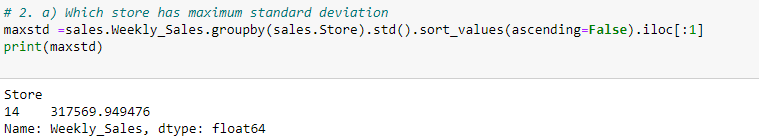
Grouping the Weekly sales based on the Store and finding the maximum sales . Sorted the data and took the first row from the output.



Answer : Store 14

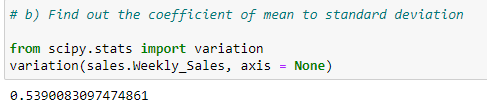
1. a) Which store has maximum standard deviation?

Grouping the Weekly sales based on the Store and finding the std deviation of the weekly sales . Sorted the data and took the first row from the output as maxstd.



Answer : Store 14

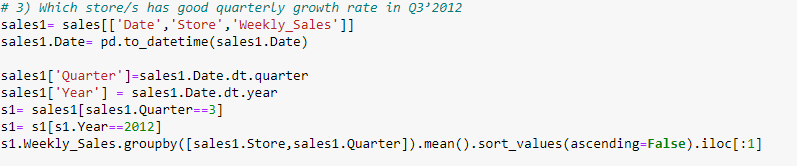
b) Find out the coefficient of mean to standard deviation



Answer : 0.539

3)Which store/s has good quarterly growth rate in Q3’2012

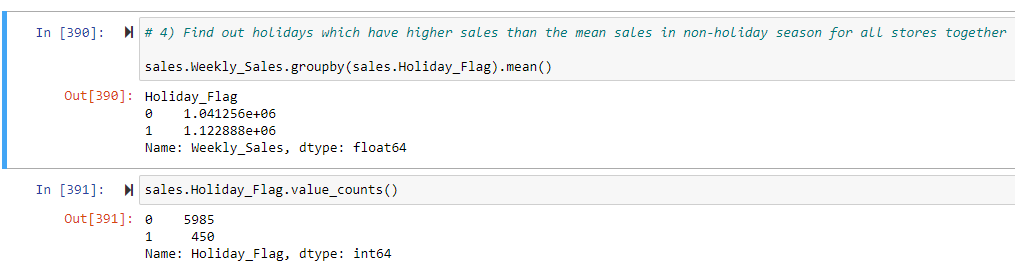
I took the Date,Store and Weekly Sales into sales1. Then converted the Date into datetime type.I added 2 new columns Quarter and Year to find the quarter & the year from the date. I filtered out the data first based on the quarter and then the year. I then grouped the weekly sales based on the store and quarter and found the mean. Sorted the data and took the first row from the output.



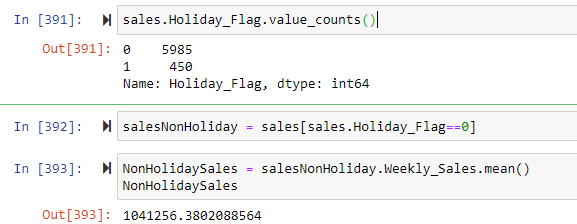


Answer : Store 4

4)Find out holidays which have higher sales than the mean sales in non-holiday season for all stores together.

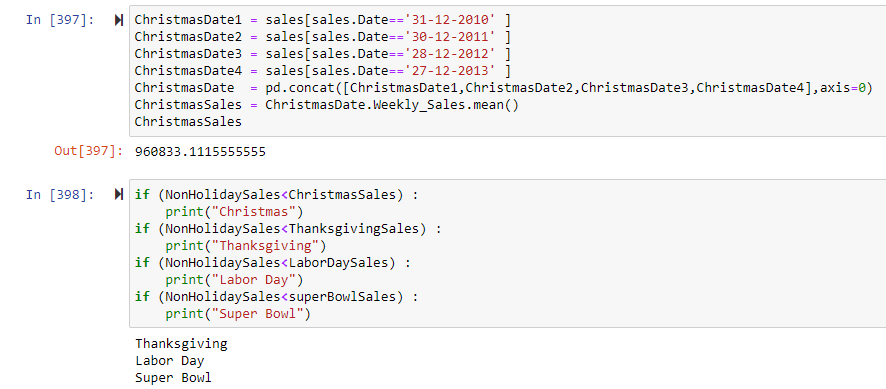


Here, we first found the mean sales for all stores together,



Found the mean of each of the holiday seasons.

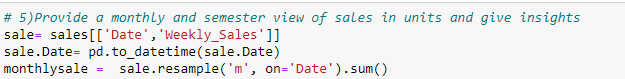




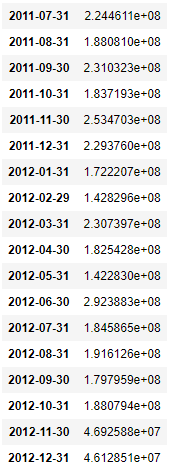
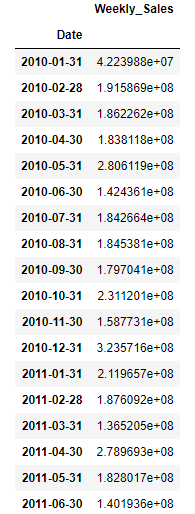
Then we compared the mean of each of the holiday sales with non holiday sales we had found earlier.

Answer: Thanksgiving , Labor day and Super bowl had higher sales than the mean sales in non-holiday season for all stores together.

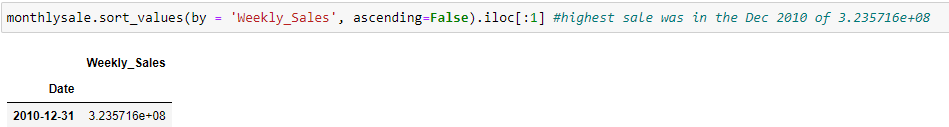
5) Provide a monthly and semester view of sales in units and give insights



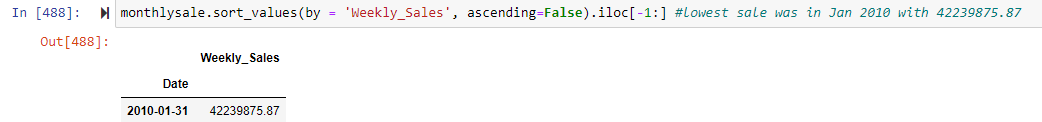
We took the Date and Weekly sales from sales,corrected the format for date into datetime and resampled the data on month.



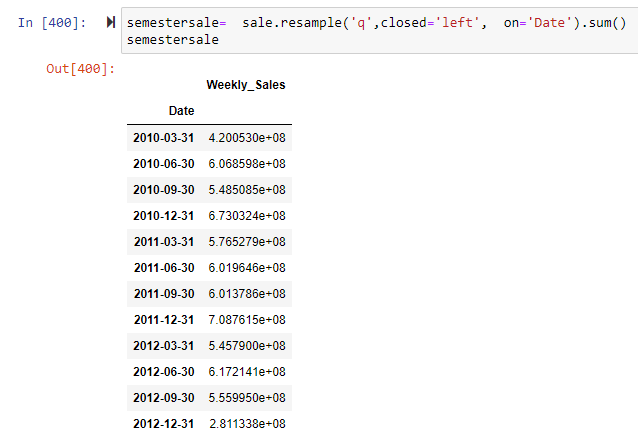
Highest sale was in the month of Dec 2010 with 3.235716e+08 sales



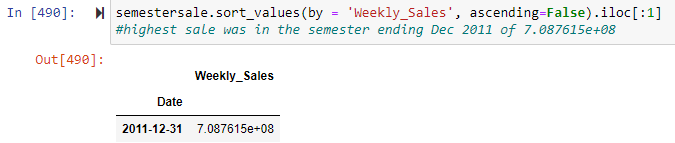
Lowest sale was in the month of Jan 2010 with 42239875.87 sales



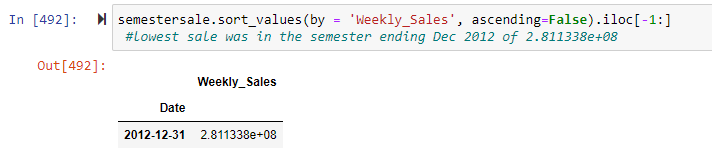
Similarly found the sum of the Weekly sales per semester



Highest sale was in the semester ending Dec 2011 with 7.087615e+08 sales

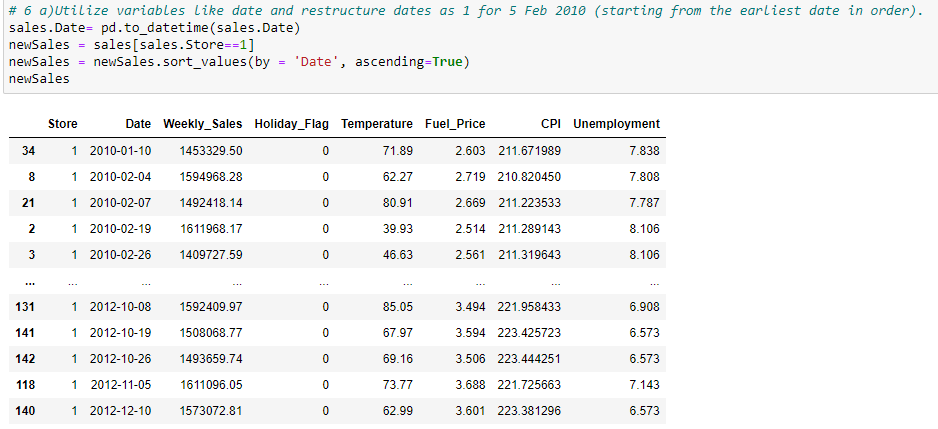


Lowest sale was in the semester ending Dec 2012 with 2.811338e+08

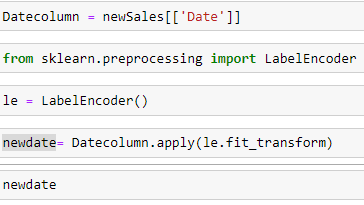


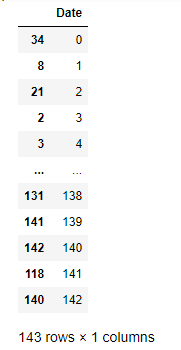
6) For Store 1 – Build prediction models to forecast demand

a)Setting it to only store 1



b)Restructuring the date

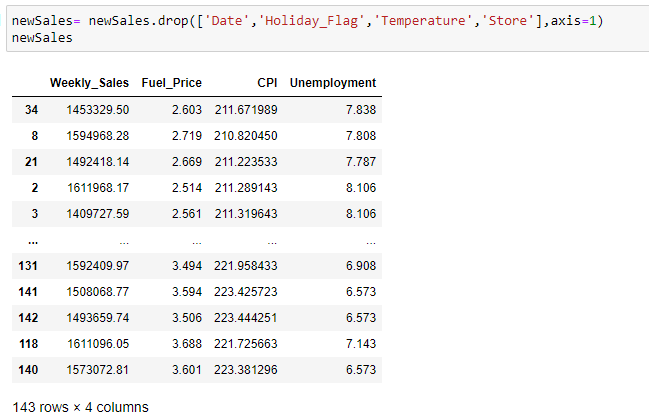




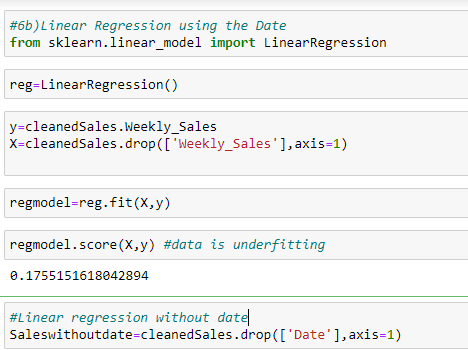
We used Label encoder to assign a dummy variable for the date ,then we dropped the date from the dataframe and then joined the new date .

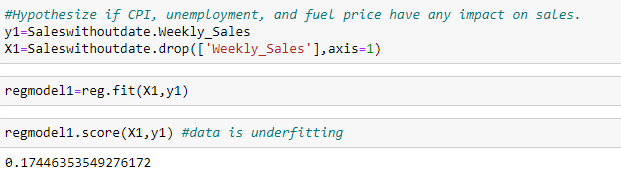
c)Hypothesize if CPI, unemployment, and fuel price have any impact on sales.





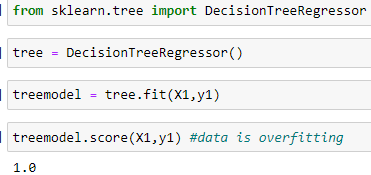
Linear Regression





As the score is less than 0.60, it is underfitting

-->Decision Tree



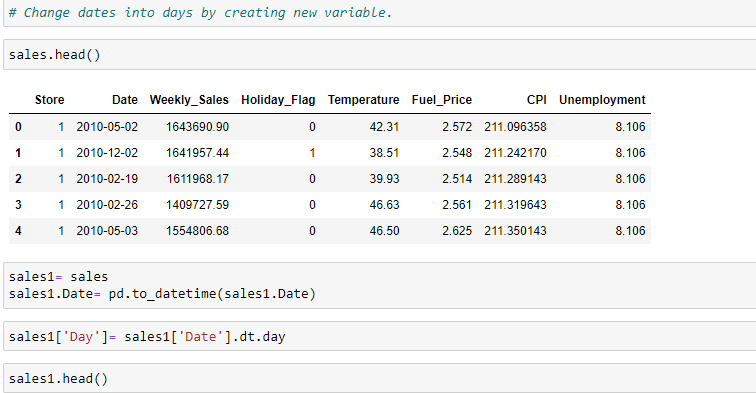
As the score is more than 0.0.95 , it is overfitting.

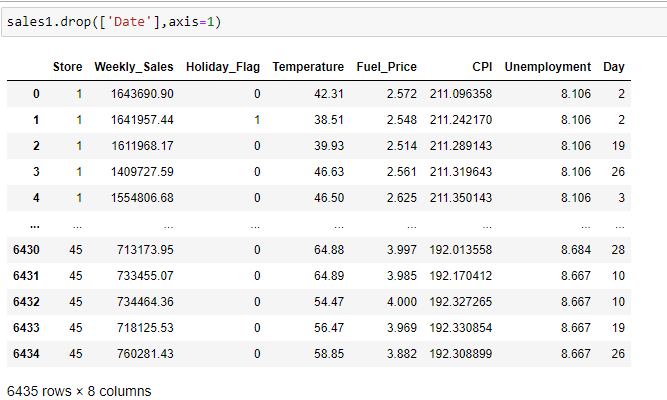
KNeighbors Regression model



Using KNN we got score of 0.3577. It is underfitting.

7)Change date into day





Values after dropping Date.