Name: Vedant Devidas Patil

IST 652 – Scripting for Data Analysis.

**MINI PROJECT 1 – WALMART DATASET**

Data and Source:

The Walmart Dataset which has the Walmart stores sales data can be found on the Kaggle dictionary of datasets by going to the below link: <https://www.kaggle.com/datasets/rutuspatel/walmart-dataset-retail>

The dataset consists of data of sales from 2010-02-05 to 2012-11-01, in the file Walmart Store sales. Within this file you will find the following fields as mentioned on the Kaggle:

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

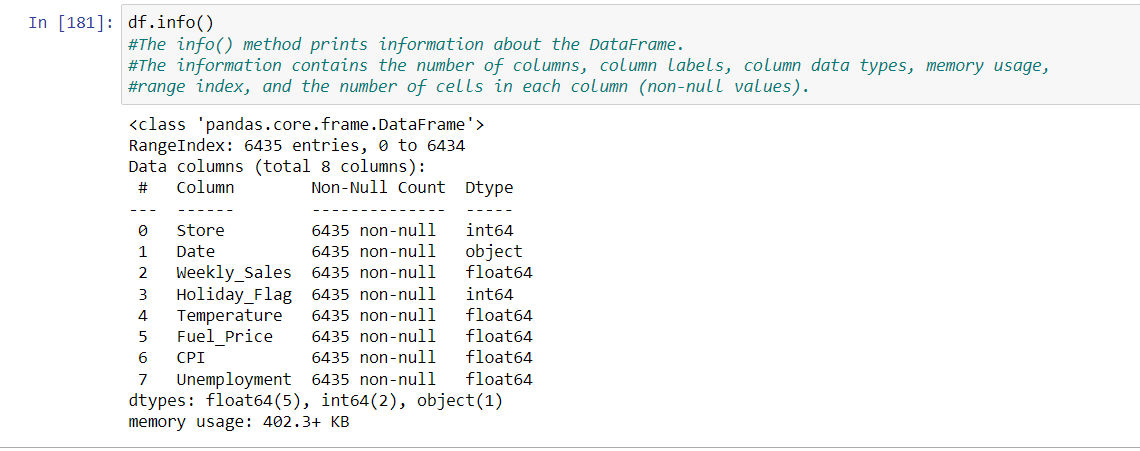
Fuel\_Price - Cost of fuel in the region

CPI – Prevailing consumer price index

Unemployment – The prevailing unemployment rate

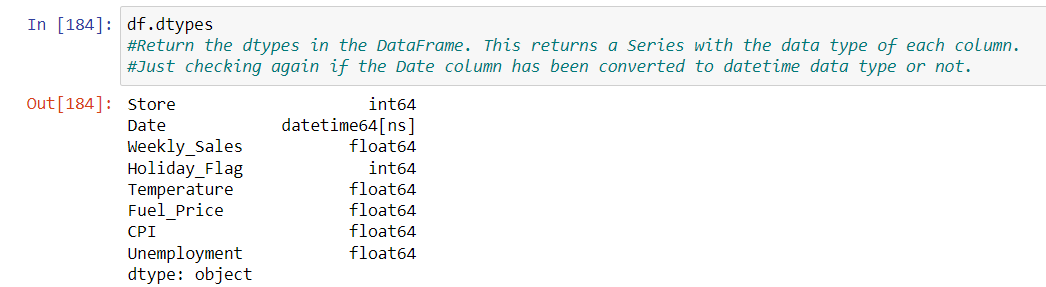
Description of Data Exploration and Data Cleaning:

In order to proceed with the analysis, it is very important to identify and understand the data. The structure and size of data, role that each column plays in the dataset, the datatype of the data.



The above screenshot shows the info() method that gives an information about all the Columns and datatypes, from which we can infer which column datatypes should be changed according to our requirements. Also, it shows us about how many non-null values are present in each column.

Dtypes: Another way of data exploration that shows the datatype of each column as below.



The describe() method returns description of the data in the DataFrame. If the DataFrame contains numerical data, the description contains these information for each column: count - The number of not-empty values.

mean - The average (mean) value.

std - The standard deviation.

min-The minimum value in that column

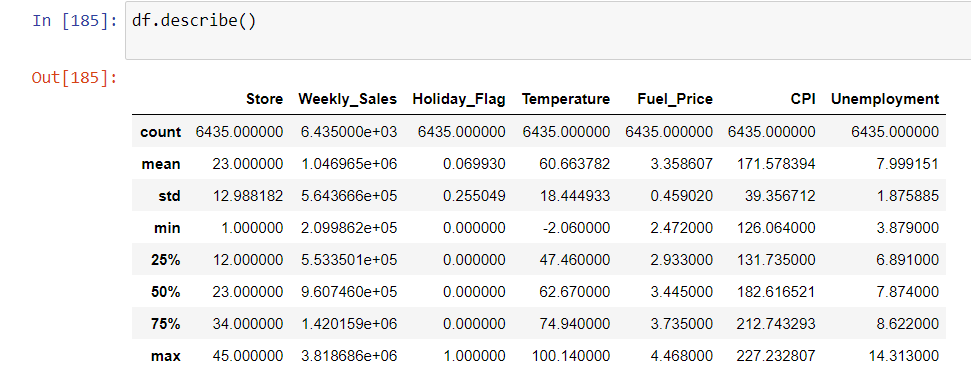
25%: what is 25th percent of the data in that column.

50%: what is 50th percent of the data in that column.

75%: what is 75th percent of the data in that column.

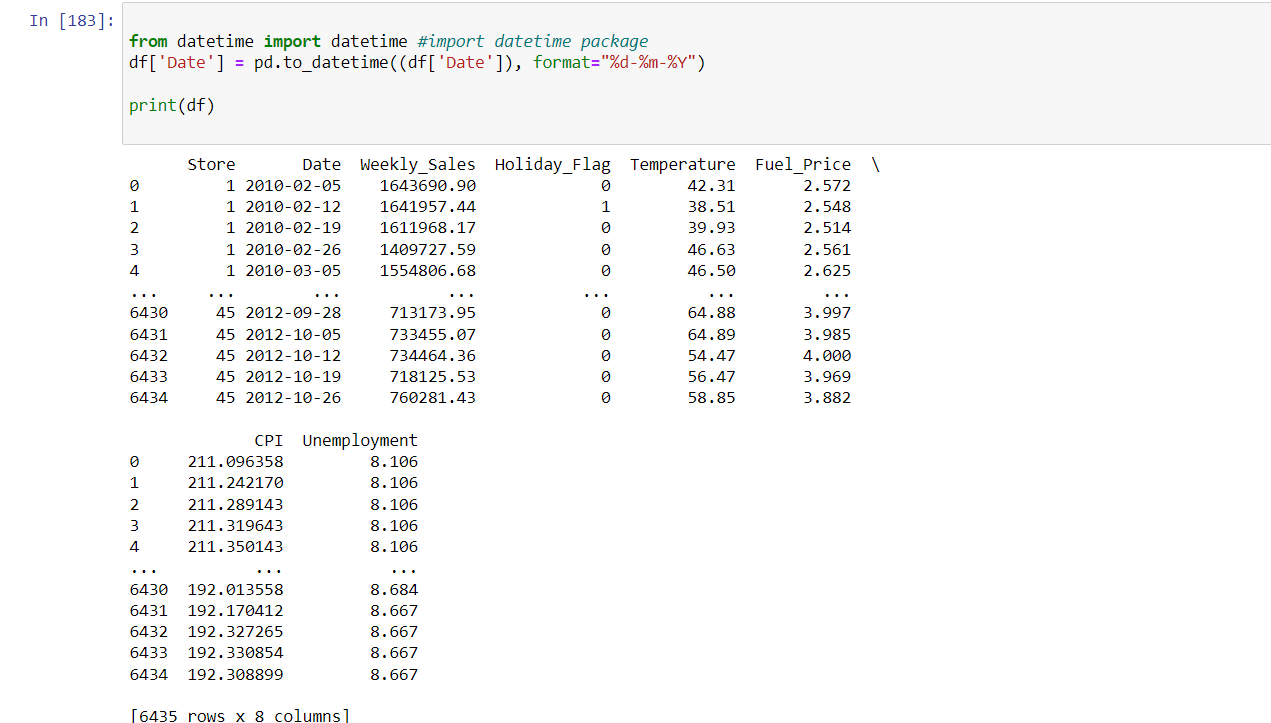
max: The maximum value of that column.

This helps us to find required outliers.

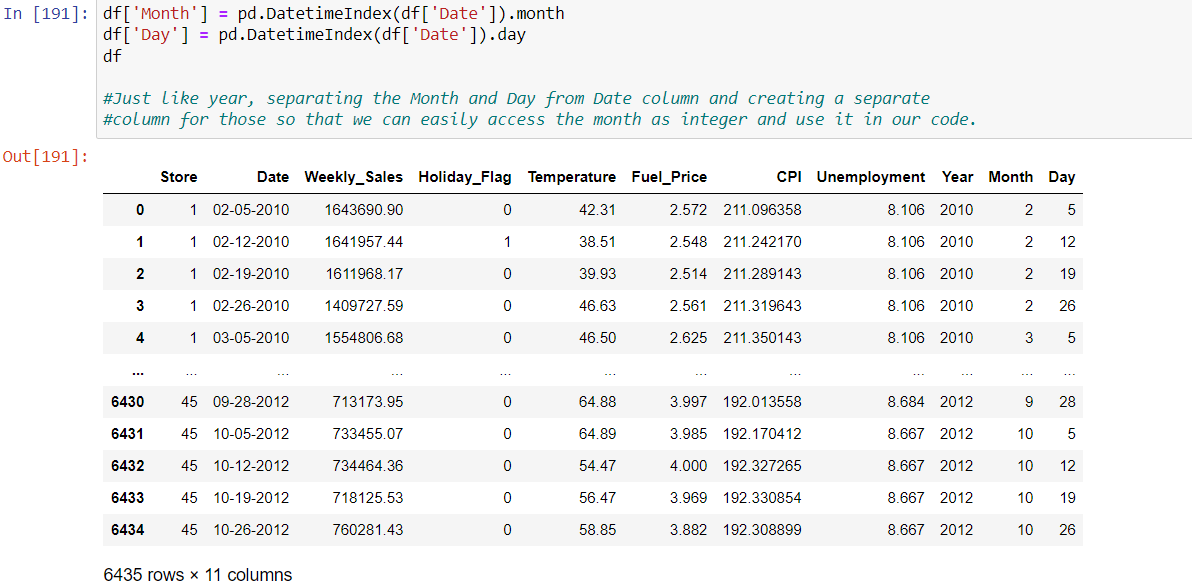


Data Cleaning and Formatting

Now, since the column “Date” is of object/string type, hence we need to convert it to required datetime64(ns) format so that we perform required operations on Date column through pandas. Changing the date column datatype from object to Datetime using below code. Using to\_datetime function of the datetime package, we are converting its data type and specifying the original format that already exist in our imported dataset which is dd-mm-yyyy. Important step in data cleaning and formatting.



Separating the month, date and year column from Date Column and creating separate column for each individual information using below code.



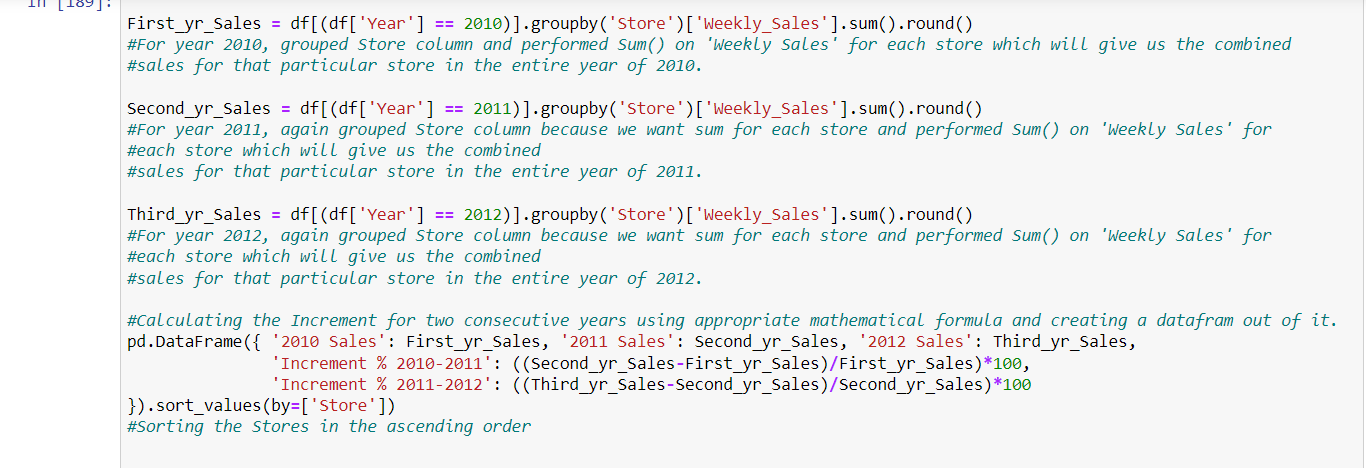
Now lets begin with the analysis wherein I have solved two comparison questions

Q1. Find out the annual increment/decrement percentage of sales of each store every consecutive year for a period of 3 years from 2010 to 2012 and determine which financial year the stores performed worst/best.

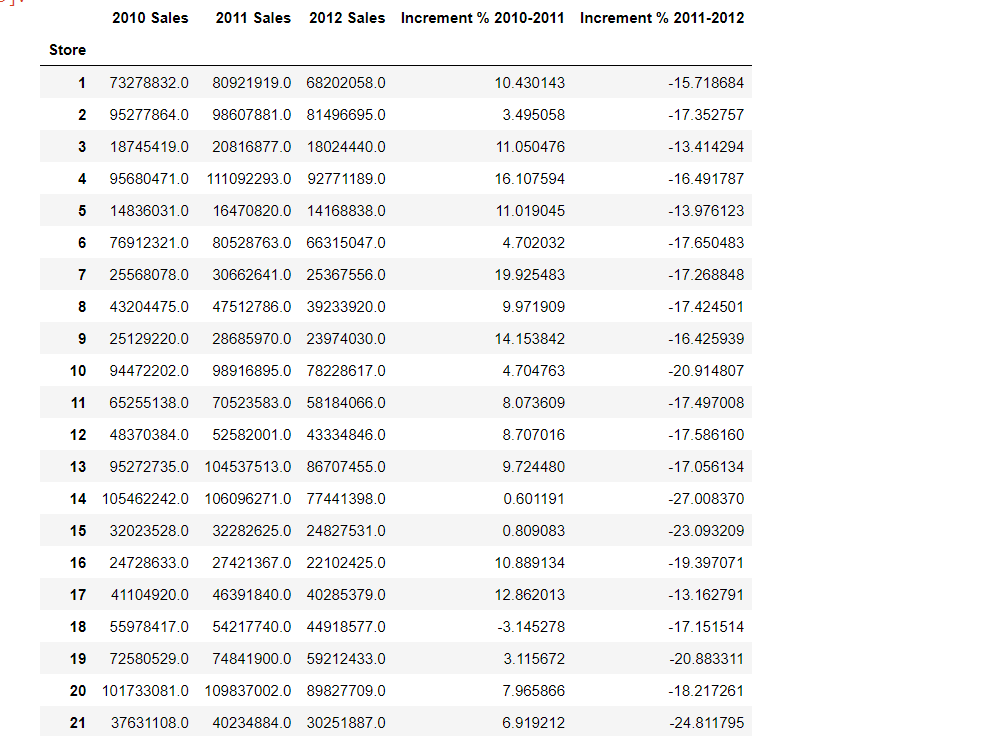
Here, we are finding out for each store that what was the increment or decrement for each store in two consecutive years i.e from 2010 to 2011 and other from 2011 to 2012. This will help us find which year was best for the stores overall, and how much stores actually progressed over time. If not progressed, what was their decrement percentage?

Here the unit of analysis is total sales for each year and the comparison factor is all the stores from 1 to 45. We can compute the comparison between two stores can comparing their increment/decrement over two years and decide which store is better over other.

Program & Description:



Output:



From the ouput, we can infer that almost all of the stores performed positively from year 2010-2011, meaning, their growth is positive and their sales increased over time. However, from 2011-2012, the sales declines for almost all of the stores which shows that 2012 was the worst year for the stores and their Sales was highest in 2011, showing that it was the best year for all the stores.

Q2. Find out top 5 stores and which particular month for that store has the highest sales in year 2010 and 2011.

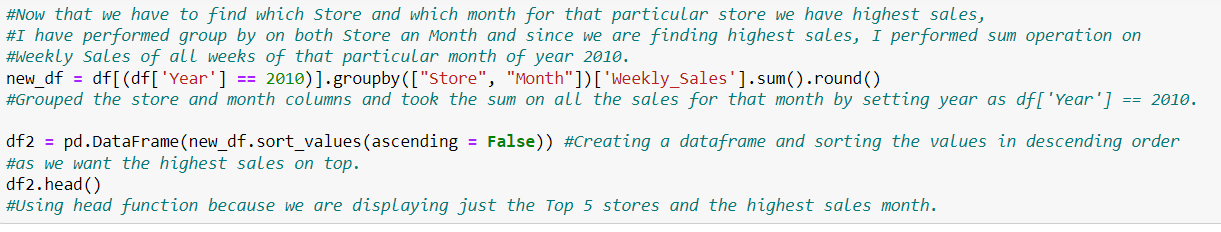
Here, we are finding out which store performed the best in 2010 and 2011 along with which month was the most successful month for that particular store. In this way, we are finding the top 5 stores that performed the best for a particular month according to their Sales amount.

Here, the unit of analysis is Store and Month as we are finding out which store and what particular month that store gave their highest output and Comparison factor should be Sum of Monthly Sales as we are displaying the store with highest sales at the top.

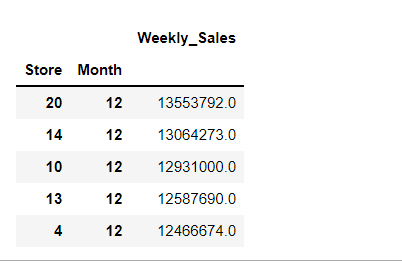
In order to proceed with this, first, we had to find out the sum of all the sales in the entire month, for all months, for all the stores. Once we find that, we can find out which month had the highest sales and display that month for that store along with its sales for that month.

Program & Explaination:

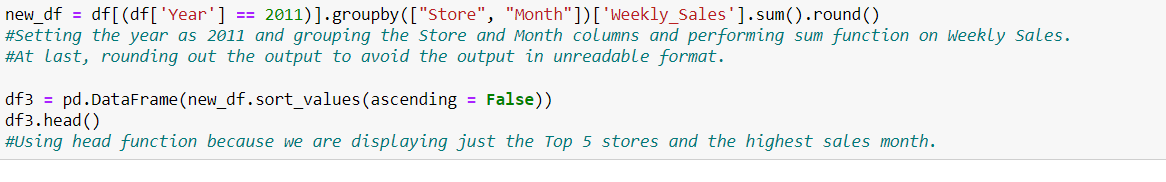
For year 2010



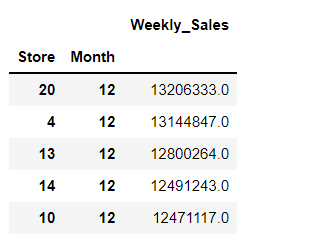
Output:



For year 2011:



Output:



From the above screenshots, we can infer that, for all top 5 stores, the 12th Month, December, had the highest sales with Store 20 bagging the topmost position. Also, we can see that for both year, there is some similarity in the store numbers, meaning that store 20,4,13,14,10 are only appearing in top 5 list for both years. Hence, these are some best performing stores with Store 20 being at Top.