Predictive Analytics Mini Project

Spring 2022

Project Report – Walmart Sales Forecasting

<u>Group 3:</u>

Shreyashi Mukhopadhyay

Shalmali Joshi

Namit Srivastava

Nhi Nguyen

Introduction:

Walmart (https://www.walmart.com/) 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. This project aims to describe the exploratory study of historical time series sales data for 45 Walmart stores spanning across 99 departments located in different regions of the country and tuning the best algorithms to obtain the best possible results for weekly sales forecasting. The overall data showed a 52-week seasonal frequency.

Holidays can create a huge impact on sales. So, if there is a good prediction on Sales then Walmart can calculate how much product to order during Holiday time. It will help in predicting which products need to be purchased during the holiday time as customers are planning to buy the products and they need to be available immediately. Through this prediction they can figure out which product to be made available at what time. This problem can also solve the issue of Marketing Campaigns as forecasting is often used to adjust ads and marketing campaigns can influence the number of sales. Walmart runs several markdown events throughout the year and these markdown events precede to the prominent holidays and it becomes even more important to schedule Markdown events preceding or at the time of popular Holiday events more efficiently.

Project description:

Predicting future sales for a company is one of the most important aspects of strategic planning.

This project aims to analyze how internal and external factors would affect the future Weekly Sales of Walmart, one of the biggest companies in the US.

This project also encompasses a complete analysis of data through data visualization, data exploration, pattern discovery and time series analysis in R as the programming language.

Data Availability:

The project data and relevant information are available at https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting.

The data collected ranges from 2010 to 2012, where 45 Walmart stores across the country were included in this analysis. Each store contains several departments, and this data aims to are tasked with predicting the department-wide sales for each store. It is important to note that we also have external data available like CPI, Unemployment Rate and Fuel Prices in the region of each store which, hopefully, helps us to make a more detailed analysis.

The following 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

Data description:

We have utilized three data files for our analysis from Kaggle.com, namely Train.csv, Stores.csv and Features.csv which contain the following information:

Train.csv

- Store: The store number. Range from 1–45.
- Type: Three types of stores 'A', 'B' or 'C'.
- Size: Sets the size of a Store would be calculated by the no. of products available in the particular store ranging from 34,000 to 210,000.

Stores.csv

-Date: The date of the week where this observation was taken.

-Weekly_Sales: The sales recorded during that Week.

-Store: The store which observation in recorded 1–45.

-Dept: One of 1–99 that shows the department.

-IsHoliday: Boolean value representing a holiday week or not.

Features.csv

-Temperature: Temperature of the region during that week.

-Fuel_Price: Fuel Price in that region during that week.

-MarkDown 1:5: Represents the Type of markdown and what quantity was available during that week.

-CPI: Consumer Price Index during that week.

-Unemployment: The unemployment rate during that week in the region of the store.

Data Manipulation:

→ We first read the data files Train.csv, Stores.csv and Features.csv in R and convert them into individual dataframes in R.

→ We then merge the 3 files into a single file **dfTrainMerged** and remove NA values.

→ We convert the Boolean field IsHoliday to a numeric field 0 or 1.

→ We then split the date field into four data columns of Day, Month, Year and Week.

→ The resultant final data frame **dfTrainMerged1** has the following data structure.

```{r}
str(dfTrainMerged1)
```

```
'data.frame':   421570 obs. of  19 variables:
 $ Store        : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Year         : chr  "2012" "2012" "2012" "2012" ...
 $ Month        : chr  "01" "01" "01" "01" ...
 $ Day          : chr  "13" "13" "13" "13" ...
 $ IsHoliday    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Dept         : int  1 25 55 92 21 17 19 42 4 44 ...
 $ Weekly_Sales : num  16894 6384 10418 158155 7091 ...
 $ Type         : chr  "A" "A" "A" "A" ...
 $ Size         : int  151315 151315 151315 151315 151315 151315 151315 151315 151315 151315 ...
 $ Temperature  : num  48.5 48.5 48.5 48.5 48.5 ...
 $ Fuel_Price   : num  3.26 3.26 3.26 3.26 3.26 ...
 $ MarkDown1    : num  5183 5183 5183 5183 5183 ...
 $ MarkDown2    : num  8026 8026 8026 8026 8026 ...
 $ MarkDown3    : num  42.2 42.2 42.2 42.2 42.2 ...
 $ MarkDown4    : num  453 453 453 453 453 ...
 $ MarkDown5    : num  3719 3719 3719 3719 3719 ...
 $ CPI          : num  220 220 220 220 220 ...
 $ Unemployment : num  7.35 7.35 7.35 7.35 7.35 ...
 $ week         : chr  "02" "02" "02" "02" ...
```

Data Visualization:

Distribution of Continuous variables
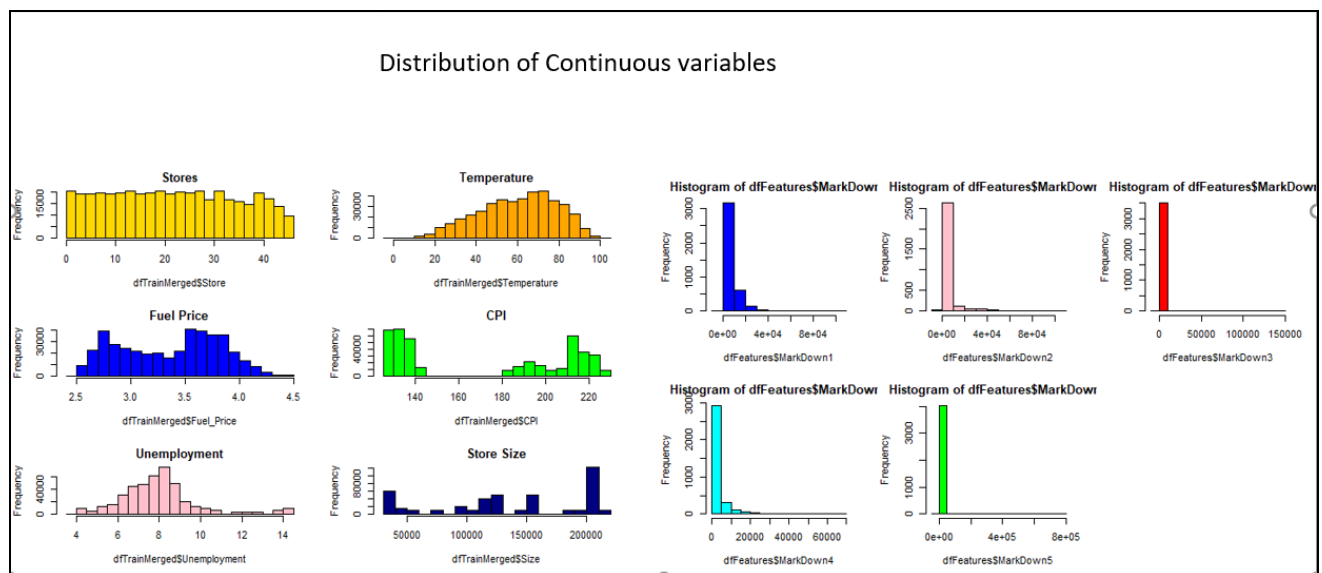
→ The plots of the continuous variables in the data show us that except Temperature, all the other variables are either positively skewed or have uneven distribution or have extreme values.

→ We thus would need to apply log transformation to the above plots to view a more even spread of the data.

Distribution of Weekly Sales
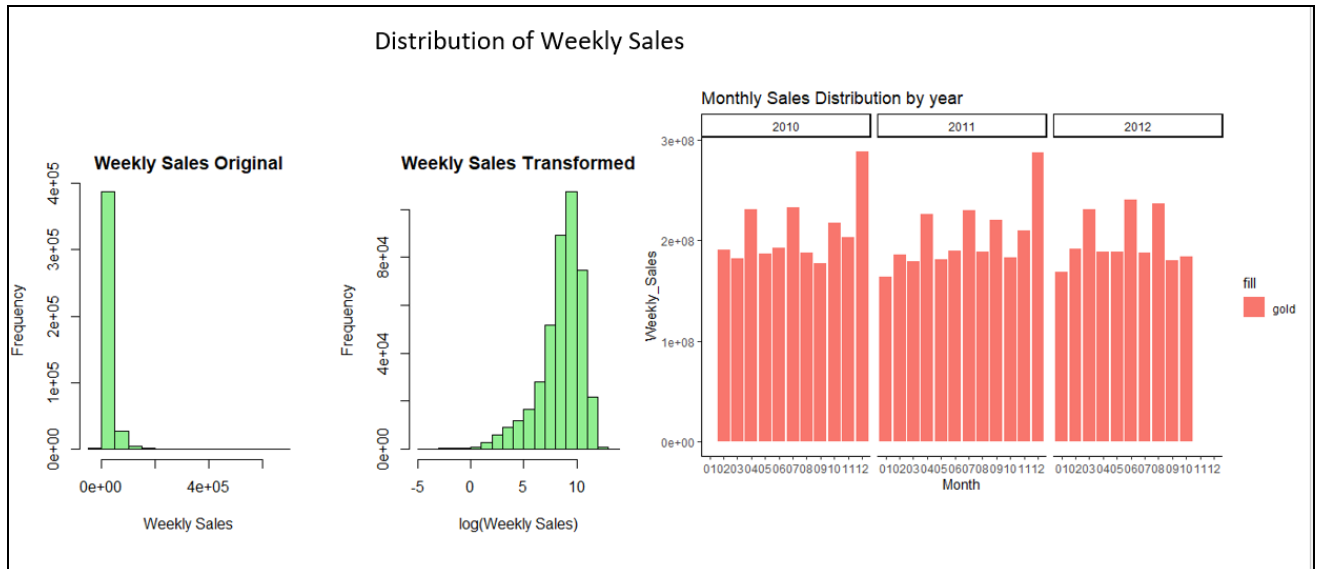
→ After applying log transformation, we now see a more evenly distributed spread of the data.

→ Also, upon visualizing the total weekly sales per year, we can observe clear seasonality in the data wherein the data peaks starting October, November, and December for both years 2010 and 2011 owing to the festive shopping during the Thanksgiving and Christmas Holidays. (Note that we have not been provided data for 11/12 and 12/12 since we need to forecast the sales for that time of the year).

Distribution of Number of stores by Store Type
Distribution of Store size by Store Type

Distribution of Total weekly sales by by Store Type
Distribution of Total weekly sales by year

→ The distribution of the number of stores by store type shows that the store belonging to format A is highest in number followed by stores of format B and format C.

→ The distribution of the store size by store type shows that the store of format A has the highest store size with the highest number of items sold in that store followed by store type B and C.

→ The plot of distribution of total weekly sales per week for 52 weeks shows that the weekly sales starts peaking from weeks 47 through 52 showing a clear seasonality in the sales data.

→ The plot of distribution of department wise mean weekly sales shows that department 92 and 95 have high mean weekly sales.

→ The plot of the distribution of weekly sales by store shows that store 20 has the highest weekly sales.

→ The plot of the distribution of weekly sales by day shows that the sales are higher on weekdays compared to weekends and dip significantly on the last day of the month.
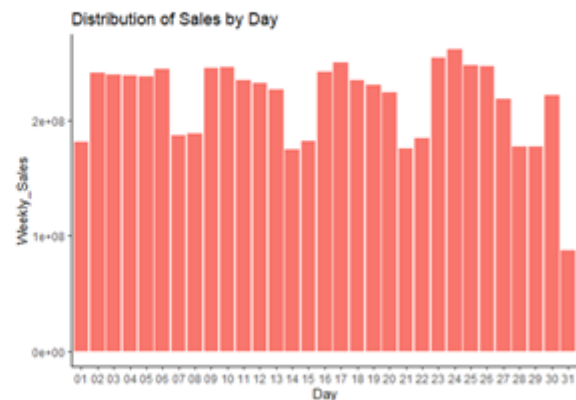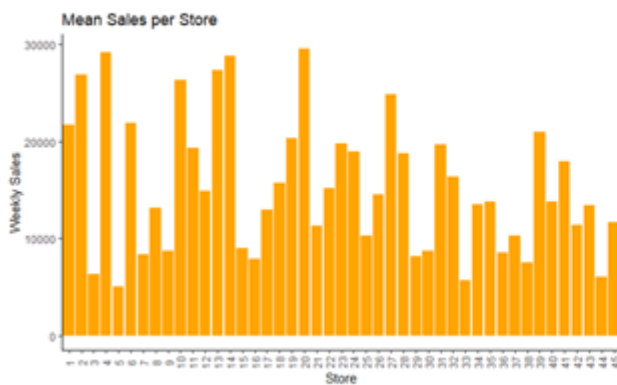
Distribution of Total Sales per week for 52 weeks
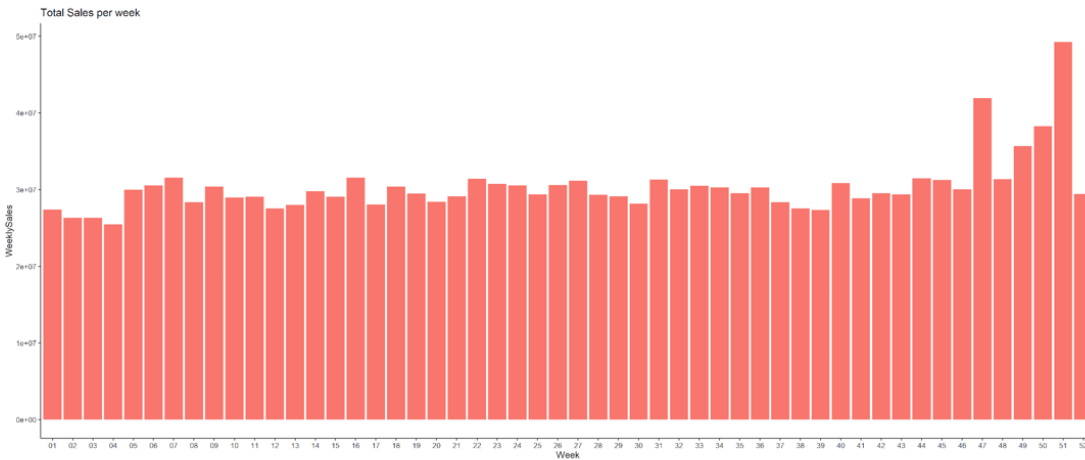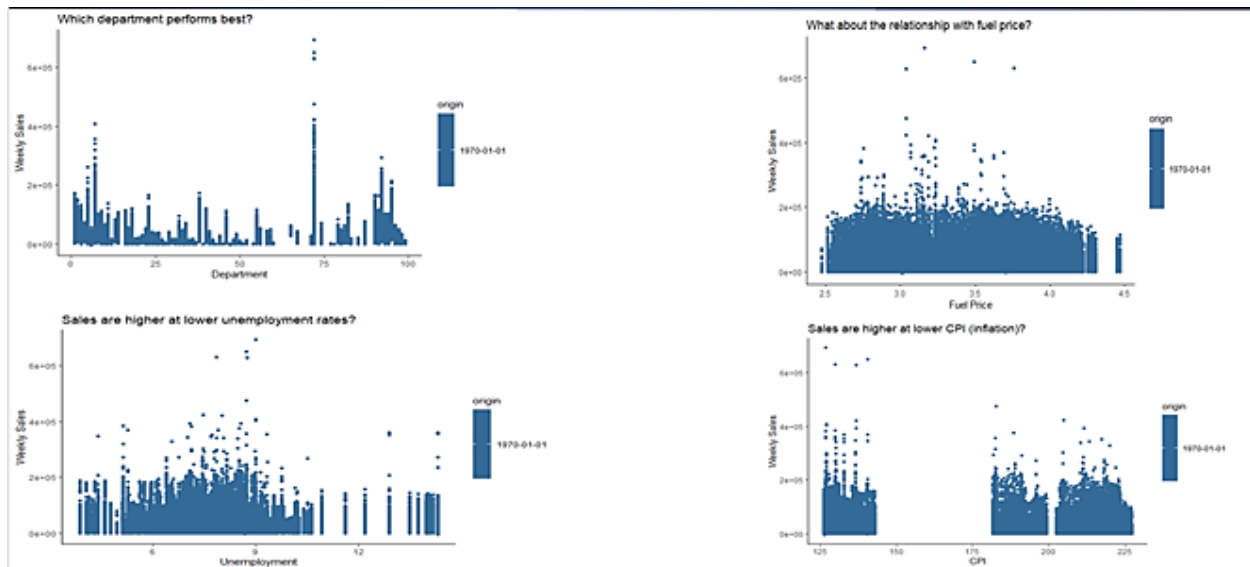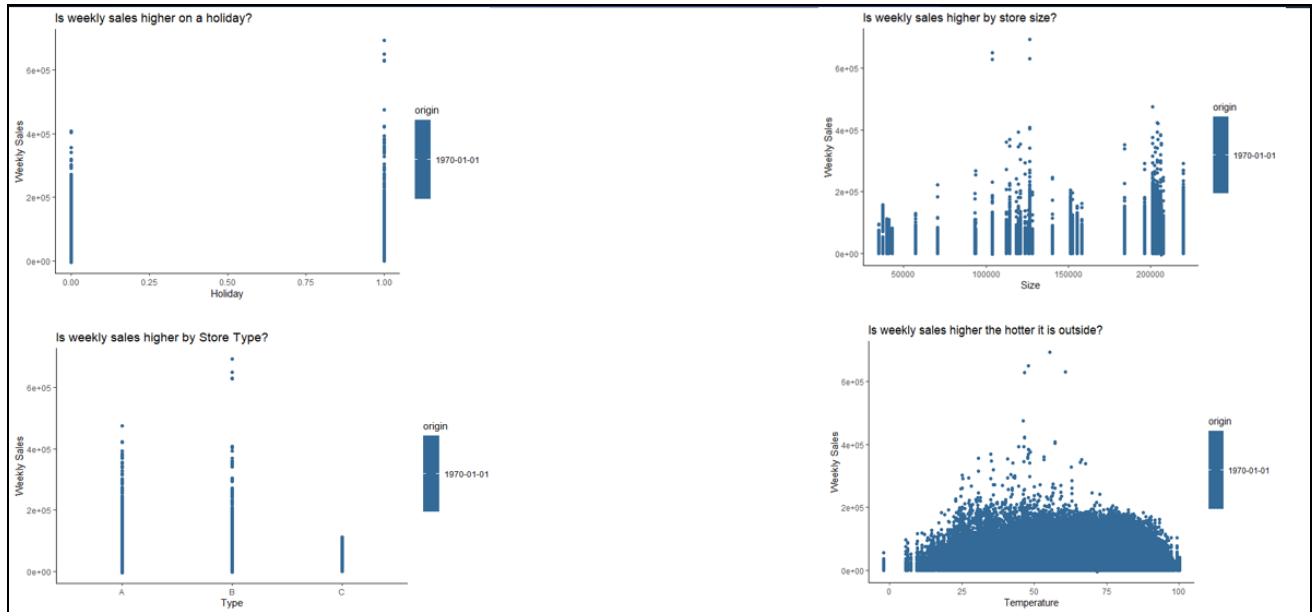

Distribution of Department wise Mean Sales
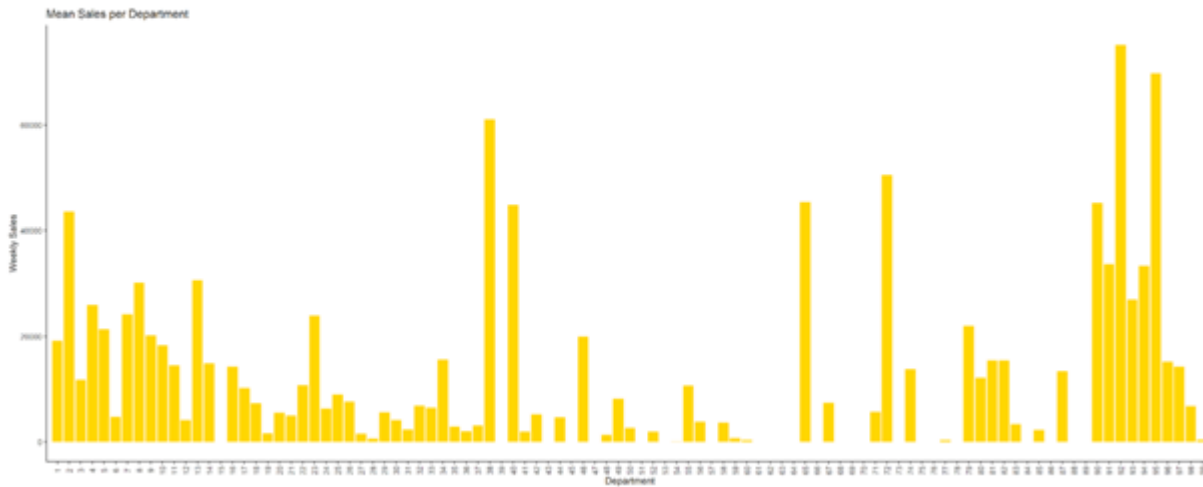
Distribution of Mean weekly sales per store and
Distribution of sales by days of the month

Data Exploration

→ Exploring the data at a more granular level shows us that:

1. Weekly Sales are higher on a holiday than a non-holiday.

2. Weekly sales is higher by store size.

3. Weekly sales is highest for store type A compared to B and C except a few outliers.

4. Weekly sales are highest when the temperature is moderate around 50 -60 degrees.

5. Department 75 performs best in terms of weekly sales.

6. Weekly sales are higher when the fuel prices are lower.

7. Weekly sales also reported higher levels when the unemployment was lower.

8. Weekly Sales were higher for lower CPI values.

HOLIDAY AND MARKDOWN ANALYSIS
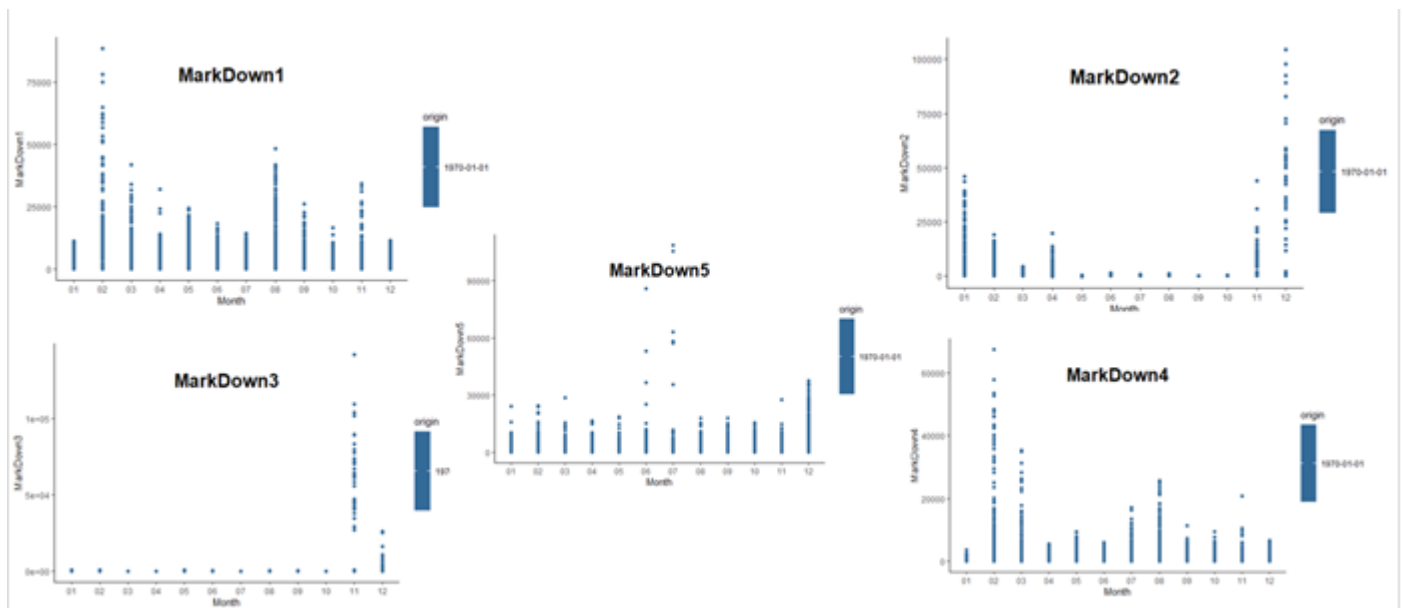
9. Markdown1 and Markdown4 are offered in February during the Super Bowl Holiday, and it shows the sale of around 75,000 – 80,000 units.

10. Markdown 2 is offered in December during the Christmas Holiday, and it shows the sale of around 100,000 units.

11. Markdown3 is offered during the Thanksgiving Holiday, and it shows the maximum sale of more than 1 million units.

12. Markdown 5 is offered during the 4th of July Holiday, and it shows a sale of around 30,000 to 90,000 units.

CORRELATION PLOTS AND HEATMAP:
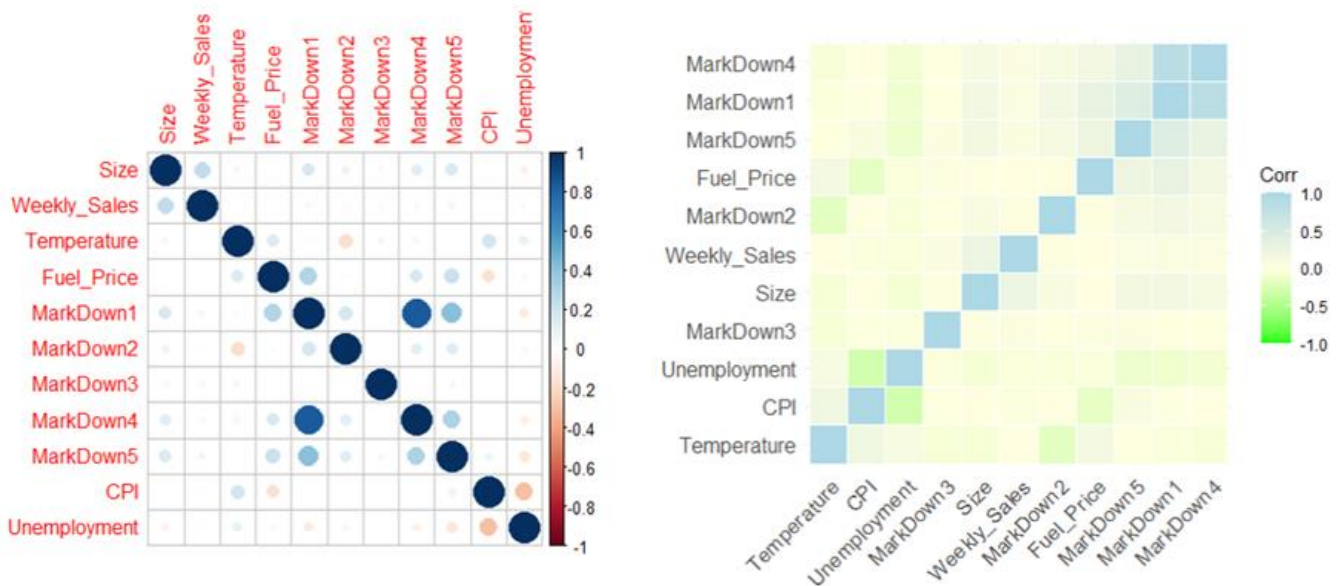
13. The correlation plot and the correlation heatmaps show that Markdown1 and Markdown 4 are highly positively correlated since both the Markdowns are offered during the Super Bowl Holiday in February.

14. Markdown5 and Markdown1 are positively correlated. Markdown5 is offered during the 4th of July which follows the Markdown 1 offered during February and March for the Super Bowl Holiday, which shows that Markdown items unsold during the Markdown1 are further Marked down during the 4th of July holiday to enhance sales.

15. CPI and unemployment are highly negatively correlated since the CPI normally remains low when the unemployment is lower.