# PROJECT: RETAIL ANALYSIS WITH WALMART DATA

**Business scenario:** One of the leading retail stores in the US, Walmart, would like to predict the sales and demand accurately. There are certain events and holidays which impact sales on each day. There are sales data available for 45 stores of Walmart. The business is facing a challenge due to unforeseen demands and runs out of stock sometimes, due to the inappropriate machine learning algorithm. An ideal ML algorithm will predict demand accurately and ingest factors like economic conditions including CPI, Unemployment Index, etc.

Walmart runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of all, which are the Super Bowl, Labour 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 modelling the effects of markdowns on these holiday weeks in the absence of complete/ideal historical data. Historical sales data for 45 Walmart stores located in different regions are available.

Holiday Events are:
Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13
Labour 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

**Objectives:** Perform following analysis using data available:

- Which store has maximum sales.
- Which store has maximum standard deviation i.e., the sales vary a lot. Also, find out the coefficient of variance i.e. ratio of standard deviation to mean.
- Which store/s has good quarterly growth rate in Q3'2012
- Some holidays have a negative impact on sales. Find out holidays which have higher sales than the mean sales in non-holiday season for all stores together
- Provide a monthly and semester view of sales in units and give insights
- For Store 1 – Build prediction models to forecast demand. Hypothesize if CPI, unemployment, and fuel price have any impact on sales.

**Data available:** We have historical data available which covers sales from 2010-02-05 to 2012-11-01, in the file Walmart Store sales. This file has 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

**R Code:**

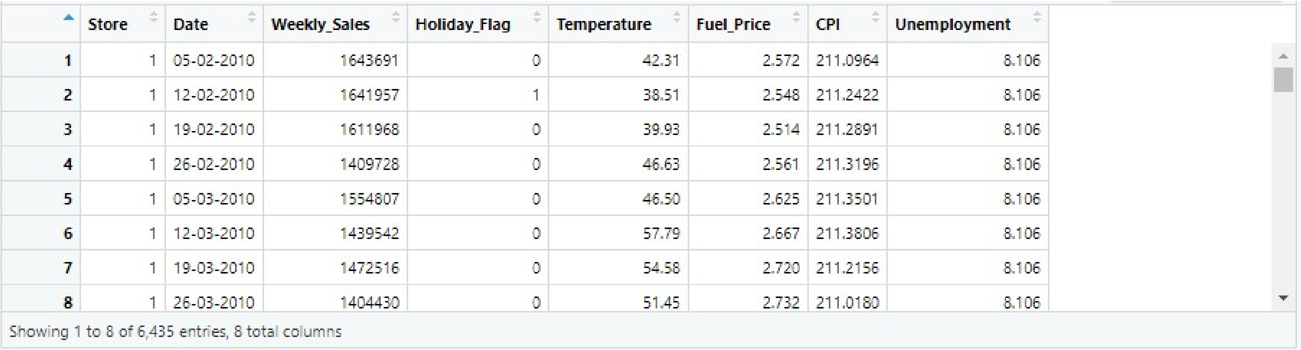
```
##### Project towards completion "Data Science using R" by Manish Gupta - Walmart Sale Data

rm(list=ls())
setwd("D:/SimpliLearn-DataScience/2) Post Graduate Program in Data Science/3. PG DS - Data Science with R/MG Project")
getwd()
walmart = read.csv("Walmart_Store_sales.csv")
View(walmart)
summary(walmart)
str(walmart)

## Data Preparation - Converting Store and Holiday_Flag to factor and Date to Date format
```

```
walmart$Store <- as.factor(walmart$Store)
walmart$Date =as.Date(walmart$Date,format="%d-%m-%Y")
walmart$Holiday_Flag <- as.factor(walmart$Holiday_Flag)

str(walmart)

#### Q1: Which store has maximum sales?
store_sales = aggregate(Weekly_Sales~Store,data=walmart, sum) # Aggregate sales data storewise and get total sale

# Method-I
which.max(store_sales$Weekly_Sales)        # Get index position of maximum value of Weekly_Sales
store_sales[which.max(store_sales$Weekly_Sales),1]   # Get Store name corresponding to maximum value of Weekly_Sales

# Method - II
library(dplyr)
arrange(store_sales, desc(Weekly_Sales))
# Answer-1: Store 20 has highest sale. (sale value = 301397792)

#### Q2: Which store has maximum standard deviation i.e., the sales vary a lot. Also, find out the coefficient of mean to standard deviation?
# Typing error in second part of question. We will find coefficient of variation for each store which is the ratio of standard deviation to mean.

store_sales$sales_mean <- aggregate(Weekly_Sales~Store,data=walmart, mean)$Weekly_Sales # Aggregate sales data storewise and get mean value and assign values to new variable sales_mean in store_sales
store_sales$sales_sd <- aggregate(Weekly_Sales~Store,data=walmart, sd)$Weekly_Sales     # Agreegate sales data storewise and get standard deviation and assign values to new variable sales_sd in store_sales
store_sales$cov = store_sales$sales_sd/ store_sales$sales_mean

str(store_sales)

arrange(store_sales, desc(sales_sd))
## Store 14 has highest standard deviation = 317569.95

arrange(store_sales, desc(cov))
## Store 35 has highest coefficient of variation = 0.22968111

#### Q3: Which store/s has good quarterly growth rate in Q3'2012?
walmart_q <- walmart
Q2_start <- as.Date("01-04-2012","%d-%m-%Y")
Q2_end <- as.Date("30-06-2012","%d-%m-%Y")
Q3_start <- as.Date("01-07-2012","%d-%m-%Y")
Q3_end <- as.Date("30-09-2012","%d-%m-%Y")

# Converting dates to quarter
walmart_q$Quarter = ifelse(Q2_start<=walmart_q$Date & walmart_q$Date <= Q2_end,"Q2-2012", ifelse(Q3_start<=walmart_q$Date & walmart_q$Date < Q3_end,"Q3-2012","Other"))

View(walmart_q)

library(tidyr)
walmart_g <- walmart_q %>%          ## The source dataset
  group_by(Store, Quarter) %>%     ## Grouping variables
  summarise(Weekly_Sales = sum(Weekly_Sales)) %>%  ## aggregation of the Weekly_Sales column
  ungroup() %>%                  ## spread doesn't seem to like groups
  spread(Quarter, Weekly_Sales)    ## spread makes the data wide
```

```r
walmart_g = data.frame(walmart_g)
walmart_g$growth_perct = round((walmart_g$Q3.2012-walmart_g$Q2.2012)/walmart_g$Q2.2012*100,2)
arrange(walmart_g, desc(walmart_g$growth_perct))
## Store 7 had highest growth rate of 13.33%

#### Q4: Some holidays have a negative impact on sales. Find out holidays which have higher sales than the mean
sales in non-holiday season for all stores together?
SuperBowl <- as.Date(c("2010-02-12","2011-02-11","2012-02-10","2013-02-08"))
LabourDay <- as.Date(c("2010-09-10", "2011-09-09", "2012-09-07", "2013-09-06"))
Thanksgiving <- as.Date(c("2010-11-26", "2011-11-25", "2012-11-23", "2013-11-29"))
Christmas <- as.Date(c("2010-12-31", "2011-12-30", "2012-12-28", "2013-12-27"))

walmart_h <- select(walmart,Date,Weekly_Sales)
walmart_h$hflag <- ifelse(walmart_h$Date %in% SuperBowl, "SB", ifelse(walmart_h$Date %in% LabourDay, "LD",
ifelse(walmart_h$Date %in% Thanksgiving, "TG", ifelse(walmart_h$Date %in% Christmas, "CH","None"))))
aggregate(Weekly_Sales~hflag,data=walmart_h, mean) # Aggregate sales data holiday-wise and get mean value.
## Mean sales in non-holiday season for all stores together is 1041256.4 and except Christmas all holidays have higher
sales than average sale in non-holiday sale.

##### Q5: Provide a monthly and semester view of sales in units and give insights
walmart_s <- walmart
walmart_s$Date =as.Date(walmart_s$Date,format=c("%d-%m-%Y"))
View(walmart_s)
walmart_s_month_year = transform(walmart_s,Year_Sale =as.numeric(format(Date,"%Y"))
                              ,Month_Sale =as.numeric(format(Date,"%m")))
View(walmart_s_month_year)

Summarized_View = aggregate(Weekly_Sales~Month_Sale+Year_Sale,walmart_s_month_year,sum)
View(Summarized_View)

Insight_data = arrange(Summarized_View,desc(Weekly_Sales))
View(Insight_data)
## Insights - Walmart booked highest sales in Dec 2010 and Dec 2011 and lowest sales in Jan 2011 and Jan 2012 post
that it was in June 2012. THe company need to adopt marketing strategy similar
## So December is month of highest sale and is followed by lowest sale in month of January. Walmart can plan its
inventory accordingly.

###### Q6: For Store 1 – Build  prediction models to forecast demand
library(dplyr)
walmart_store1 <- select(filter(walmart, Store==1),-1) ## Filtering data for Store 1 for building linear model
View(walmart_store1)
str(walmart_store1)

## Linear Model
walmart_lm = lm(Weekly_Sales ~ Holiday_Flag + Temperature + Fuel_Price+ CPI + Unemployment , walmart_store1)
summary(walmart_lm)

## Drop most insignificant variable Fuel_Price (p value = 60.80%)
walmart_lm1 = lm(Weekly_Sales ~ Holiday_Flag + Temperature + CPI + Unemployment , walmart_store1)
summary(walmart_lm1)

## Drop most insignificant variable Unemployment (p value = 20.54%)
walmart_lm2 = lm(Weekly_Sales ~ Holiday_Flag + Temperature + CPI , walmart_store1)
summary(walmart_lm2)

## Drop most insignificant variable Holiday_Flag1 (p value = 5.15%)
walmart_lm3 = lm(Weekly_Sales ~ Temperature + CPI , walmart_store1)
summary(walmart_lm3)
```

**R execution Output Screenshots and interpretation:**

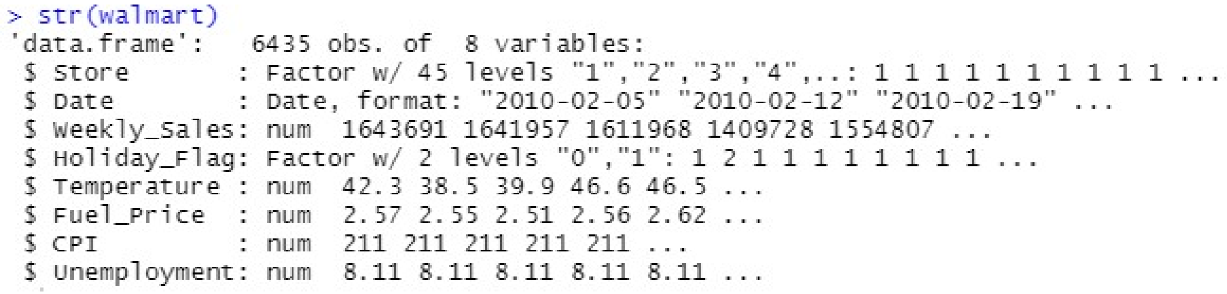
1. Screenshot of data imported:



| | Store | Date | Weekly_Sales | Holiday_Flag | Temperature | Fuel_Price | CPI | Unemployment |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 05-02-2010 | 1643691 | 0 | 42.31 | 2.572 | 211.0964 | 8.106 |
| 2 | 1 | 12-02-2010 | 1641957 | 1 | 38.51 | 2.548 | 211.2422 | 8.106 |
| 3 | 1 | 19-02-2010 | 1611968 | 0 | 39.93 | 2.514 | 211.2891 | 8.106 |
| 4 | 1 | 26-02-2010 | 1409728 | 0 | 46.63 | 2.561 | 211.3196 | 8.106 |
| 5 | 1 | 05-03-2010 | 1554807 | 0 | 46.50 | 2.625 | 211.3501 | 8.106 |
| 6 | 1 | 12-03-2010 | 1439542 | 0 | 57.79 | 2.667 | 211.3806 | 8.106 |
| 7 | 1 | 19-03-2010 | 1472516 | 0 | 54.58 | 2.720 | 211.2156 | 8.106 |
| 8 | 1 | 26-03-2010 | 1404430 | 0 | 51.45 | 2.732 | 211.0180 | 8.106 |

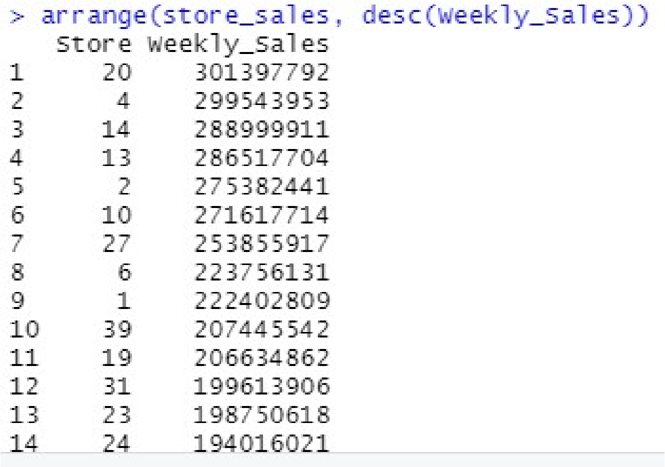
Showing 1 to 8 of 6,435 entries, 8 total columns

2. Data Preparation – Before starting data analysis we will convert Store and Holiday_Flag to factor and Date to Date format. Here is structure of converted data:



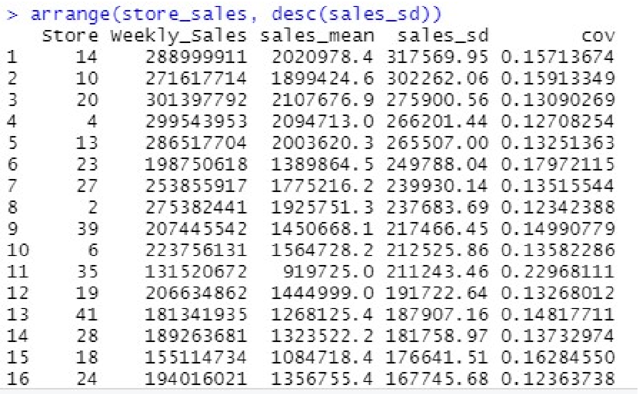
```
> str(walmart)
'data.frame':	6435 obs. of  8 variables:
 $ Store       : Factor w/ 45 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Date        : Date, format: "2010-02-05" "2010-02-12" "2010-02-19" ...
 $ Weekly_Sales: num  1643691 1641957 1611968 1409728 1554807 ...
 $ Holiday_Flag: Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 1 ...
 $ Temperature : num  42.3 38.5 39.9 46.6 46.5 ...
 $ Fuel_Price  : num  2.57 2.55 2.51 2.56 2.62 ...
 $ CPI         : num  211 211 211 211 211 ...
 $ Unemployment: num  8.11 8.11 8.11 8.11 8.11 ...
```

3. First question can be solved using aggregate command which aggregates Weekly Sales data Store-wise and give us total sale for each Store:

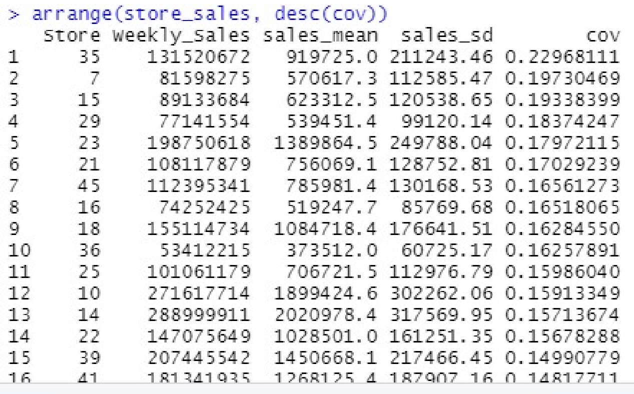


```
> arrange(store_sales, desc(Weekly_Sales))
   Store Weekly_Sales
1     20    301397792
2      4    299543953
3     14    288999911
4     13    286517704
5      2    275382441
6     10    271617714
7     27    253855917
8      6    223756131
9      1    222402809
10    39    207445542
11    19    206634862
12    31    199613906
13    23    198750618
14    24    194016021
```

4. Screenshot of output generated for second question:

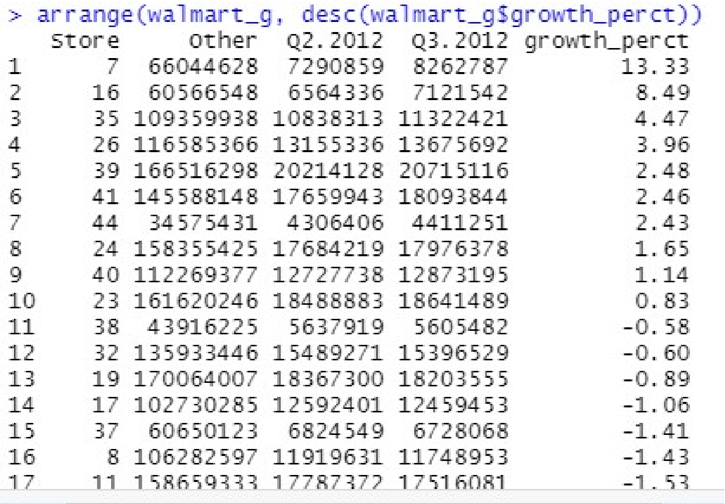


```
> arrange(store_sales, desc(sales_sd))
   Store Weekly_Sales sales_mean  sales_sd        cov
1     14    288999911  2020978.4 317569.95 0.15713674
2     10    271617714  1899424.6 302262.06 0.15913349
3     20    301397792  2107676.9 275900.56 0.13090269
4      4    299543953  2094713.0 266201.44 0.12708254
5     13    286517704  2003620.3 265507.00 0.13251363
6     23    198750618  1389864.5 249788.04 0.17972115
7     27    253855917  1775216.2 239930.14 0.13515544
8      2    275382441  1925751.3 237683.69 0.12342388
9     39    207445542  1450668.1 217466.45 0.14990779
10     6    223756131  1564728.2 212525.86 0.13582286
11    35    131520672   919725.0 211243.46 0.22968111
12    19    206634862  1444999.0 191722.64 0.13268012
13    41    181341935  1268125.4 187907.16 0.14817711
14    28    189263681  1323522.2 181758.97 0.13732974
15    18    155114734  1084718.4 176641.51 0.16284550
16    24    194016021  1356755.4 167745.68 0.12363738
```



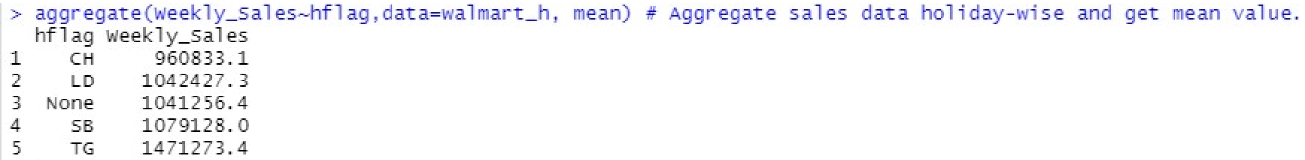
```
> arrange(store_sales, desc(cov))
   Store Weekly_Sales sales_mean  sales_sd        cov
1     35    131520672   919725.0 211243.46 0.22968111
2      7     81598275   570617.3 112585.47 0.19730469
3     15     89133684   623312.5 120538.65 0.19338399
4     29     77141554   539451.4  99120.14 0.18374247
5     23    198750618  1389864.5 249788.04 0.17972115
6     21    108117879   756069.1 128752.81 0.17029239
7     45    112395341   785981.4 130168.53 0.16561273
8     16     74252425   519247.7  85769.68 0.16518065
9     18    155114734  1084718.4 176641.51 0.16284550
10    36     53412215   373512.0  60725.17 0.16257891
11    25    101061179   706721.5 112976.79 0.15986040
12    10    271617714  1899424.6 302262.06 0.15913349
13    14    288999911  2020978.4 317569.95 0.15713674
14    22    147075649  1028501.0 161251.35 0.15678288
15    39    207445542  1450668.1 217466.45 0.14990779
16    41    181341935  1268125.4 187907.16 0.14817711
```

5. For solving third question, first introduce a new column for quarter. We will 3 type of values in this column Q2-2012, Q3-2012 and other. Then we group data to get sale figure for Q-2012 and Q3-2012 for each store. Here is screenshot of final output:



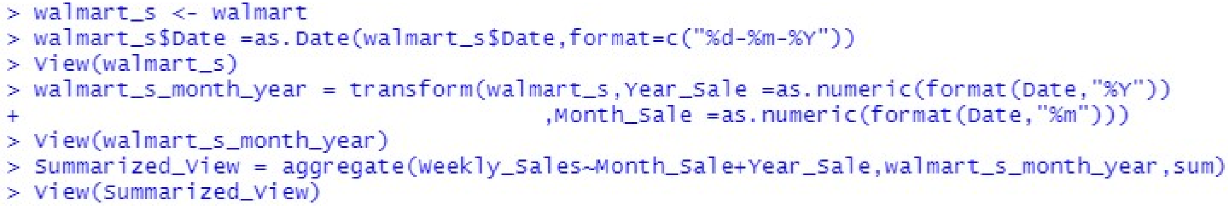
```
> arrange(walmart_g, desc(walmart_g$growth_perct))
   Store     Other   Q2.2012  Q3.2012 growth_perct
1      7  66044628  7290859  8262787        13.33
2     16  60566548  6564336  7121542         8.49
3     35 109359938 10838313 11322421         4.47
4     26 116585366 13155336 13675692         3.96
5     39 166516298 20214128 20715116         2.48
6     41 145588148 17659943 18093844         2.46
7     44  34575431  4306406  4411251         2.43
8     24 158355425 17684219 17976378         1.65
9     40 112269377 12727738 12873195         1.14
10    23 161620246 18488883 18641489         0.83
11    38  43916225  5637919  5605482        -0.58
12    32 135933446 15489271 15396529        -0.60
13    19 170064007 18367300 18203555        -0.89
14    17 102730285 12592401 12459453        -1.06
15    37  60650123  6824549  6728068        -1.41
16     8 106282597 11919631 11748953        -1.43
17    11 158659333 17787372 17516081        -1.53
```

6. In order to solve fourth question, we again introduce a new column for holiday type which contact value coded for respective holiday if applicable, otherwise none. Then we aggregate Weekly sale data holiday wise to conclude. Here is the screenshot of output:

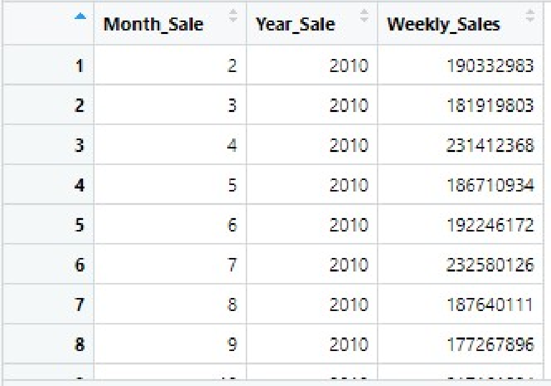


```
> aggregate(Weekly_Sales~hflag,data=walmart_h, mean) # Aggregate sales data holiday-wise and get mean value.
  hflag Weekly_Sales
1    CH     960833.1
2    LD    1042427.3
3  None    1041256.4
4    SB    1079128.0
5    TG    1471273.4
```

7. We introduce 2 new columns for month and year for each week for solution of fifth problem. Then we aggregate sale data month-wise to derive conclusion. Screenshot of code and output generated is as below:

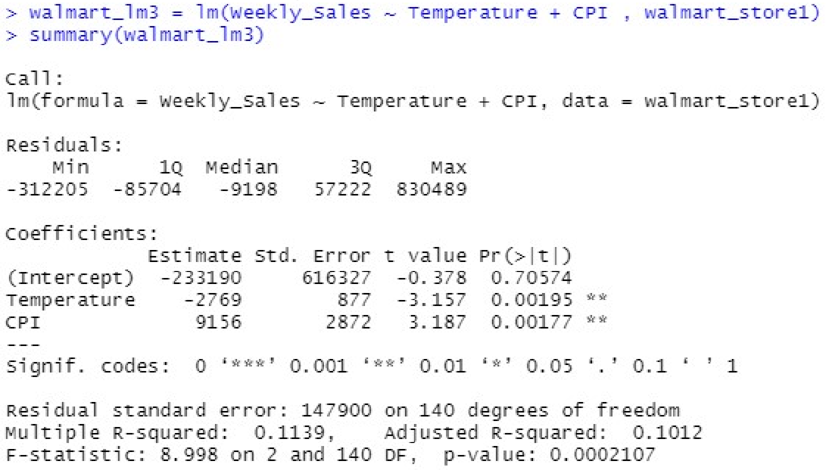


```
> walmart_s <- walmart
> walmart_s$Date =as.Date(walmart_s$Date,format=c("%d-%m-%Y"))
> View(walmart_s)
> walmart_s_month_year = transform(walmart_s,Year_Sale =as.numeric(format(Date,"%Y"))
+                                   ,Month_Sale =as.numeric(format(Date,"%m")))
> View(walmart_s_month_year)
> Summarized_View = aggregate(Weekly_Sales~Month_Sale+Year_Sale,walmart_s_month_year,sum)
> View(Summarized_View)
```



| | Month_Sale | Year_Sale | Weekly_Sales |
|---|---|---|---|
| 1 | 2 | 2010 | 190332983 |
| 2 | 3 | 2010 | 181919803 |
| 3 | 4 | 2010 | 231412368 |
| 4 | 5 | 2010 | 186710934 |
| 5 | 6 | 2010 | 192246172 |
| 6 | 7 | 2010 | 232580126 |
| 7 | 8 | 2010 | 187640111 |
| 8 | 9 | 2010 | 177267896 |

8. For solving sixth question, first we filter data for only Store 1. Then we start building linear model with Weekly sale data as dependent variable and all other as independent variable. Then we start eliminating independent variable which are not significant i.e. whose p-value is more than 0.05 and we get final output as below:



```
> walmart_lm3 = lm(Weekly_Sales ~ Temperature + CPI , walmart_store1)
> summary(walmart_lm3)

Call:
lm(formula = Weekly_Sales ~ Temperature + CPI, data = walmart_store1)

Residuals:
    Min      1Q  Median      3Q     Max 
-312205  -85704   -9198   57222  830489 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  -233190     616327  -0.378  0.70574   
Temperature    -2769        877  -3.157  0.00195 **
CPI             9156       2872   3.187  0.00177 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 147900 on 140 degrees of freedom
Multiple R-squared:  0.1139,	Adjusted R-squared:  0.1012 
F-statistic: 8.998 on 2 and 140 DF,  p-value: 0.0002107
```

## Results:

- **Q1:** Which store has maximum sales?
  ***Answer:*** *Store 20 has highest sale. (sale value = 301397792)*
- **Q2:** Which store has maximum standard deviation i.e., the sales vary a lot. Also, find out the coefficient of variance i.e. ratio of standard deviation to mean.
  ***Answer:*** *Store 14 has highest standard deviation = 317569.95 and Store 35 has highest coefficient of variation = 0.2297*
- **Q3:** Which store/s has good quarterly growth rate in Q3'2012?
  ***Answer:*** *Store 7 had highest growth rate of 13.33%*
- **Q4:** Some holidays have a negative impact on sales. Find out holidays which have higher sales than the mean sales in non-holiday season for all stores together.
  ***Answer:*** *Average sales in non-holiday season for all stores together is 1041256.4 and except Christmas all holidays have higher sales than average sale in non-holiday sale.*
- **Q5:** Provide a monthly and semester view of sales in units and give insights
  ***Answer:*** *Walmart booked highest sales in Dec 2010 and Dec 2011 and lowest sales in Jan 2011 and Jan 2012. So December is month of highest sale and is followed by lowest sale in month of January. Walmart can plan its inventory accordingly.*
- **Q6:** For Store 1 – Build prediction models to forecast demand. Hypothesize if CPI, unemployment, and fuel price have any impact on sales.
  ***Answer:*** *Our linear model is built with Weekly sale data as dependent variable and Temperature and CPI as independent variable.*

**Declaration: This project report has been prepared and is being submitted for assessment towards completion of module "PG DS - Data Science with R".**

**Date: 01-05-2021** **Name: Manish Gupta**

**Place: New Delhi (India)**