Wallmart Project

- Upamanyu Mondal

Business Scenario

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 some times, due to the inappropriate machine learning algorithm.

Objectives

To create a ML algorithm will predict demand accurately and ingest factors like economic conditions including CPI, Unemployment Index, etc.

Data Availibity

- The historical data which covers sales from 2010-02-05 to 2012-11-01. The file has 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

Additional Field added

- The date is divided into days, months and year.

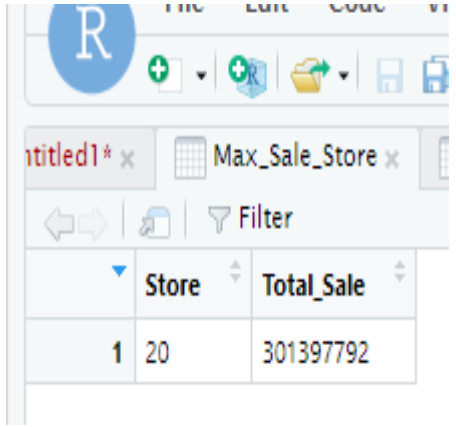
```
walmart_df$Date <- as.Date(walmart_df$Date, format="%d-%m-%Y")
walmart_df$Month=month(walmart_df$Date)
walmart_df$Year=year(walmart_df$Date)
walmart df$Day=day(walmart df$Date)
```

Data Exploration

The weekly sales is given, along with factors such as Holidays, Temperature, Fuel Price, Consumer Price Index and Unemployment on which the sales are expected to be dependent on

- Which store has maximum sales

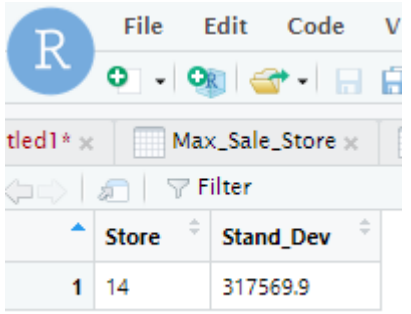
```
Max_Sale_Store =walmart_df %>% group_by(Store) %>% summarise(Total_Sale = sum(Weekly_Sales)) %>%
filter(Total_Sale == max(Total_Sale))
Max_Sale_Store
```



Store 20 was found to have the higher Total sales

Which store has maximum standard deviation i.e., the sales vary a lot.

```
Max_Stand_Dev <-walmart_df %>% group_by(Store) %>% summarise(Stand_Dev = sd(Weekly_Sales)) %>%
  filter(Stand_Dev == max(Stand_Dev))
Max_Stand_Dev
```



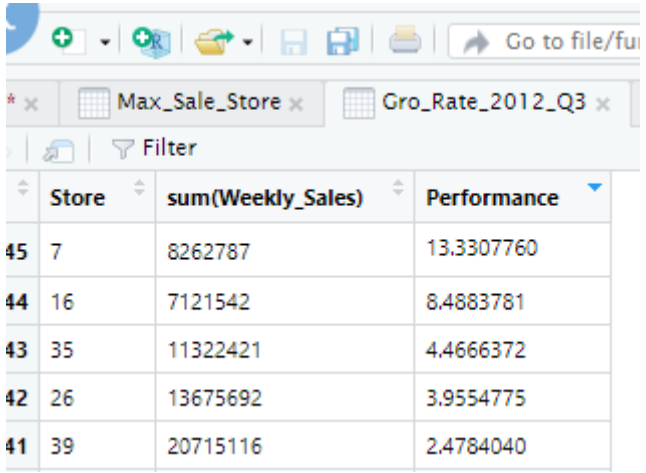
Store 14 was found to have the highest std deviation in sales

Which store/s has good quarterly growth rate in Q3'2012
As we want to calculate growth rte of Q3 2012, we need to use below formula#
Gro_Rate= (Weekly_Sale of 2012Q3 - Weekly_Sale of 2012Q2)/ Weekly_Sale 2012Q2

- The data is divided into quarters

```
str(walmart_df)
YQ = as.yearqtr(walmart_df$Date,format="%Y-%m-%d
YQ
str(walmart_df)
walmart_df$Year_Quart <- YQ
View(walmart_df)
```

```
W_Sale_2012_Q2 <-walmart_df %>% group_by(Store) %>% filter(Date >= as.Date("2012-04-01") & Date <= as.Date("2012-06-
  summarise(sum(Weekly_Sales))
W_Sale_2012_Q2

#2. Will get 2012 Q3 data

W_Sale_2012_Q3 <-walmart_df %>% group_by(Store) %>% filter(Date >= as.Date("2012-07-01") & Date <= as.Date("2012-09-
  summarise(sum(Weekly_Sales))

W_Sale_2012_Q3



# Grwoth Rate
#Gro_Rate= (Weekly_Sale of 2012Q3 - Weekly_Sale of 2012Q2)/ Weekly_Sale 2012Q2

Gro_Rate_2012_Q3 =mutate(W_Sale_2012_Q3,Performance = ((W_Sale_2012_Q3$`sum(Weekly_Sales)`
                                                       -W_Sale_2012_Q2$`sum(Weekly_Sales)`)
                                                      /W_Sale_2012_Q2$`sum(Weekly_Sales)`)
                                                      *100)
arrange(Gro_Rate_2012_Q3,desc(Performance))
```



	Store	sum(Weekly_Sales)	Performance
45	7	8262787	13.3307760
44	16	7121542	8.4883781
43	35	11322421	4.4666372
42	26	13675692	3.9554775
41	39	20715116	2.4784040

Store no 7 was found to have the highest growth followed by 16 and 35

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

```
Mean_N_Holiday_Sale <-walmart_df %>% filter(Holiday_Flag == '0') %>%
  summarise(Total_Non_Holiday_Sales =mean(Weekly_Sales))
Mean_N_Holiday_Sale

Holiday_Sales <- walmart_df %>% group_by(Date) %>% filter(Holiday_Flag == '1') %>%
  summarise(Total_Holiday_Sales = sum(Weekly_Sales)) %>%
  mutate(Higher_Holiday_Sales_Than_Non_Holiday_Sales =Total_Holiday_Sales > Mean_N_Holiday_Sale)
Holiday_Sales
```

None of the holidays were found to have a negative impact on sales…Super Bowl, Labour Day and Thanksgiving was found to have the top sales

Date	Total_Holiday_Sales	Higher_Holiday_Sales_Than_Non_Holiday_Sales
2010-02-12	48336678	TRUE
2010-09-10	45634398	TRUE
2010-11-26	65821003	TRUE
2010-12-31	40432519	TRUE
2011-02-11	47336193	TRUE
2011-09-09	46763228	TRUE
2011-11-25	66593605	TRUE
2011-12-30	46042461	TRUE
2012-02-10	50009408	TRUE
2012-09-07	48330059	TRUE

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

```
Monthly_View <- walmart_df %>% mutate(Month = month(Date)) %>%
  group_by(Month) %>% summarise(Weekly_Sales = sum(Weekly_Sales))

Monthly_View


Monthly_Yearly_View <- walmart_df %>% mutate(Month = month(Date), Year = year(Date)) %>%
  group_by(Month, Year) %>% summarise(Weekly_Sales = sum(Weekly_Sales)) %>% arrange(Year)


Monthly_Yearly_View


#Now will find Semester View of Sale

Semester_View <-walmart_df %>% mutate(Semester = semester(Date,2010)) %>% group_by(Semester)%>%
  summarise(Weekly_Sales_Semester = sum(Weekly_Sales))
Semester_View
```

	Month	Weekly_Sales
1	1	332598438
2	2	568727890
3	3	592785901
4	4	646859785
5	5	557125572
6	6	622629887
7	7	650000977
8	8	613090209
9	9	578761179
10	10	584784788
11	11	413015725
12	12	576838635

Filter

	Semester	Weekly_Sales_Semester
1	2010.1	982622260
2	2010.2	1306263860
3	2011.1	1127339797
4	2011.2	1320860210
5	2012.1	1210765416
6	2012.2	789367443

The first semester mostly have lower sales due to mainly lower sales in January

Statistical model 1

- #H0 : Null hypothesis : There is no any impact of Temperature, CPI, Fuel_Price, unemployment on Sales of store 1 #H1 : Hypothesis : There is impact of Temerature, CPI, Fuel_Price, unemployment on weekly_Sales of store 1

```
Store_1 <-filter(walmart_df, Store == 1)
head(Store_1)
Store_1=Store_1[-2]
Store_1=Store_1[-8]
Store_1=Store_1[-3]
cor(Store_1)
model_Store_1 <- lm(Weekly_Sales ~ Temperature + Fuel_Price + CPI + Unemployment, data = Store_1)
model_Store_1
summary(model_Store_1)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2727200.0  1759518.7  -1.550  0.12344
Temperature     -2426.5      917.8  -2.644  0.00915 **
Fuel_Price     -31637.1    47551.8  -0.665  0.50696
CPI             17872.1     6807.0   2.626  0.00963 **
Unemployment    90632.0    58925.1   1.538  0.12632
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 147700 on 138 degrees of freedom
Multiple R-squared:  0.1291,    Adjusted R-squared:  0.1039
F-statistic: 5.114 on 4 and 138 DF,  p-value: 0.0007142
```

P-Value = 0.00915 < alpha(0.05) -> Reject H0, as temperature has effect on sales
P-Value = 0.50696 >alpha(0.05) -> Don't reject H0, Fuel price has no effect on sales
P-Value = 0.00963 <alpha(o.o5) -> Reject H0, CPI has effect on slaes
P-Value = 0.12632 >alpha(0.05) -> Don't Reject H0, Unemployment has no effect on sales

Statistical Model 2

#H0 : Null hypothesis : There is no any impact of Temperature, CPI, Fuel Price , Holiday, Year and Date unemployment on Sales of store 1 #H1 : Hypothesis : There is impact of Temperature, CPI, Fuel Price, unemployment on weekly_Sales of store 1

```
Store_1a <-filter(walmart_df, Store == 1)
head(Store_1a)
Store_1a=Store_1a[,-9]
model_Store_1a = lm(Weekly_Sales ~ Temperature + Fuel_Price + CPI + Unemployment +Year +Day  + Holiday_Flag, data =
summary(model_Store_1a)
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.647e+08  1.127e+08   3.235  0.00153 **
Temperature  -3.050e+03  9.163e+02  -3.329  0.00112 **
Fuel_Price    7.625e+04  5.370e+04   1.420  0.15789
CPI           3.712e+04  8.751e+03   4.242 4.10e-05 ***
Unemployment  1.157e+04  5.870e+04   0.197  0.84406
Year         -1.846e+05  5.666e+04  -3.258  0.00142 **
Day          -5.299e+03  1.302e+03  -4.071 7.92e-05 ***
Holiday_Flag  8.782e+04  4.581e+04   1.917  0.05732 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 135400 on 135 degrees of freedom
Multiple R-squared:  0.2836,    Adjusted R-squared:  0.2465
F-statistic: 7.636 on 7 and 135 DF,  p-value: 9.408e-08
```

P-Values for Year and date were lesser than 0.05 which shows that sales are dependent on particular in addition to Temperature, CPI

This model also has a better R score in comparison to the previous one.

Conclusion

- Hence we can conclude that the weekly_sales are dependent on Temperatures from the model. This observation also matches with the “monthly and semester view of sales” where we saw that sales goes down mainly during winter or Jan.
- CPI is also strongly related as sales can go down with rising inflation
- Sales are dependent on days and from the hypothesis of model 2 we saw that holidays can affect sales.
- In Both the models Unemployment did not affect sales mainly because Walmart has a huge section of FMCG and also Medical units and are necessary items