Retail Analysis with Walmart Data

How holiday markdowns impact Walmart store sales and what factors can help to predict store sales

Analysing impact of holiday markdowns on Walmart store sales

Every year, the retail chain Walmart organises markdown sales around major holiday events in the US viz Super Bowl, Labour Day, Thanksgiving and Christmas. The given dataset uses weekly sales data for three years (2010-2012) from 45 Walmart stores to analyse the impact of holidays and other factors like Consumer Price Inflation (CPI), Unemployment (in the region where the store is located), Fuel Price etc on the sales. These factors are used to build a model which predicts sales for a given store.

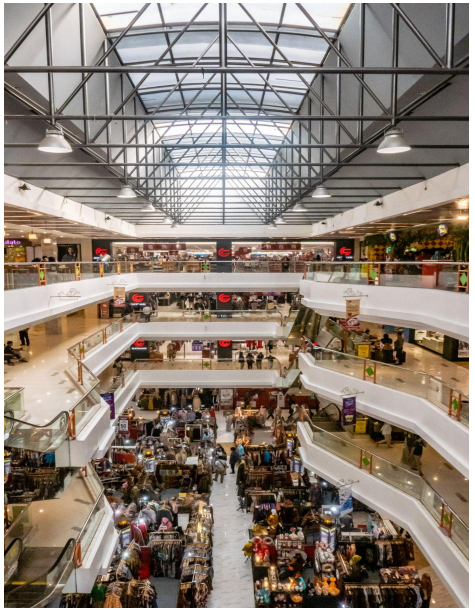


Photo by Sangga Rima Roman Selia on Unsplash

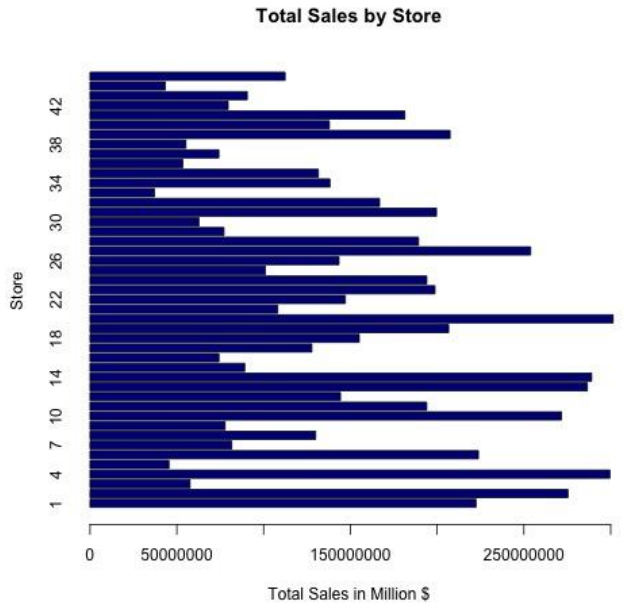
Business Questions

With a little pre-processing of the data and some exploratory analysis, the following business questions are answered.

1. **Which store has maximum sales?**

Store 20, with a little over USD 300m in sales.

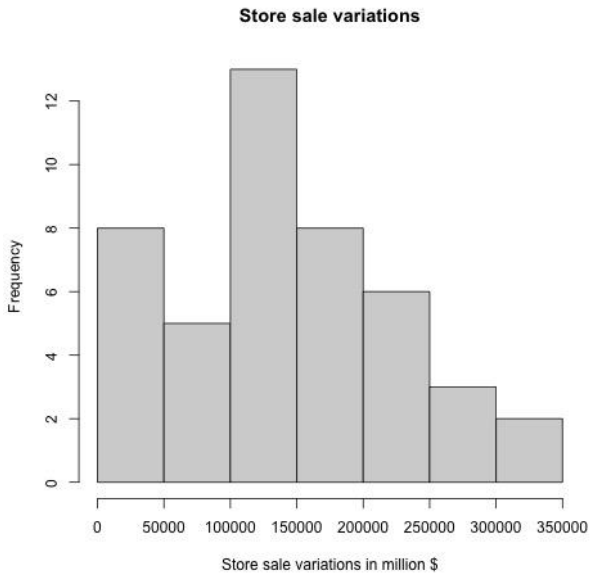
```
     Store             x
Min.   : 1   Min.   : 37160222
1st Qu.:12   1st Qu.: 79565752
Median :23   Median :138249763
Mean   :23   Mean   :149715977
3rd Qu.:34   3rd Qu.:199613906
Max.   :45   Max.   :301397792
```



2. **Which store has maximum standard deviation? What is the coefficient of mean sales to standard deviation?**

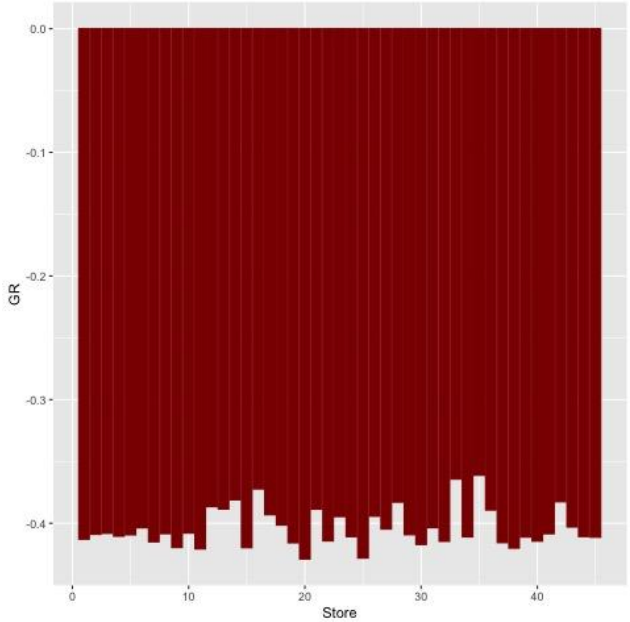
Store 14 - with slightly more than USD 317m in sales, showing maximum variation than sales of other stores. Please refer to attached R script for coefficient.

```
[1] 317569.9
```



3. Which store(s) had good quarterly growth rate in Q3’2012?

All store sales showed a negative growth rate in Q3’2012 versus Q2. This was likely caused by item returns, though the dataset does not have any variables to prove that.



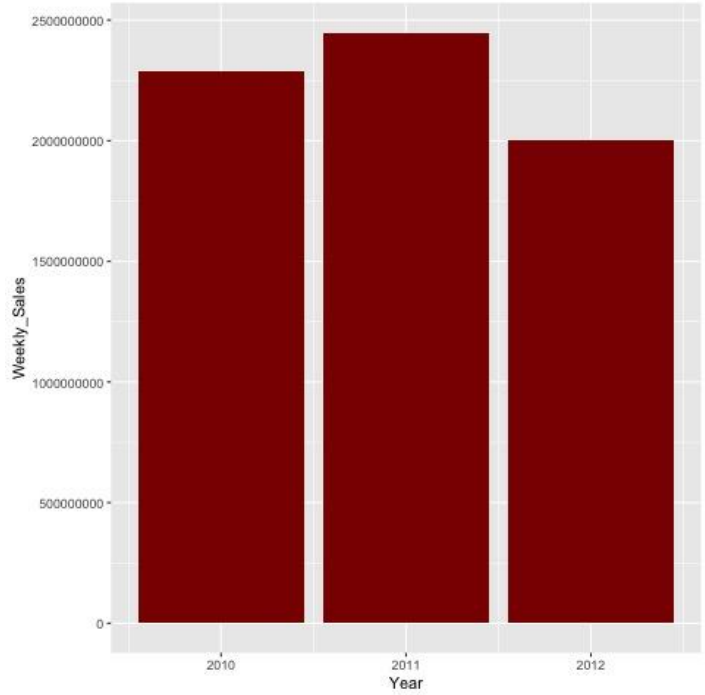
4. **Some holidays have negative impact on sales. Which holidays have higher sales than mean sales in non-holiday season for all stores together?**

The following holiday periods have higher mean sales than non-holiday sales combined:

1. Christmas 2010
2. Labour Day 2010
3. Thansksgiving 2011
4. Labour Day 2011
5. Labour Day 2012

Insights

Overall, December 2010 and June 2012 were the best months for sales. As such, the marketing strategy adopted by the company during these periods may be replicated in other periods to generate better sales in other periods as well.



Sales Forecast

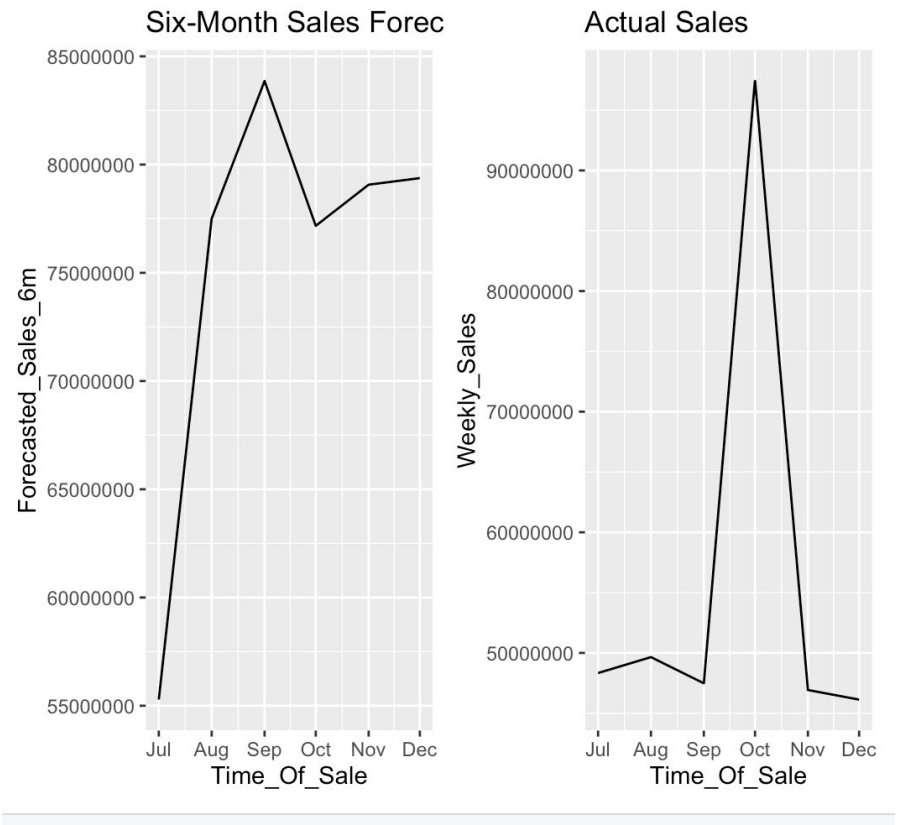
Two models were used to predict sales for all the stores, since subsetting for just one store would have led to an insufficient sample size. In the first model, linear regression was used to generate a sales forecast. In the initial test, only Holiday Flag (whether the sales period is a holiday or not), CPI and Unemployment emerged as significant factors influencing sales.

```
Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept)  1726523.4    79763.5  21.646 < 0.0000000000000002 ***
Holiday_Flag   74891.7    27639.3   2.710              0.00675 **
Temperature     -724.2      400.5  -1.808              0.07060 .
Fuel_Price    -10167.9    15762.8  -0.645              0.51891
CPI            -1598.9      195.1  -8.194 0.00000000000000302 ***
Unemployment  -41552.3     3972.7 -10.460 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 557400 on 6429 degrees of freedom
Multiple R-squared:  0.02544,   Adjusted R-squared:  0.02469
F-statistic: 33.57 on 5 and 6429 DF,  p-value: < 0.00000000000000022
```

On the basis of this, the model was further refined to drop Temperature and Fuel Price as independent variables, with the null hypothesis that Holiday Flag, CPI and Unemployment as factors influencing Weekly Sales. As such, even this model produced a p-value much below 0.05 and R-squared number that can explain less than 5% of the variations in sales, leading us to reject the hypothesis.

```
Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept)  1664939.2    51224.4  32.503 < 0.0000000000000002 ***
Holiday_Flag   84509.5    27250.4   3.101              0.00194 **
CPI            -1652.8      185.2  -8.923 < 0.0000000000000002 ***
Unemployment  -42542.2     3886.5 -10.946 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 557500 on 6431 degrees of freedom
Multiple R-squared:  0.02479,   Adjusted R-squared:  0.02434
F-statistic:  54.5 on 3 and 6431 DF,  p-value: < 0.00000000000000022
```

The second approach involved time series forecasting using the ARIMA model where moving averages are considered. The forecasts were for a six-month period only. In this model as well, the forecast was not very accurate.



The sale forecast using ARIMA did not track the actual sales well, given a high overall value of the mean absolute percentage error (MAPE) of 0.51. For some stores specifically, the error in terms of deviation was quite high. The reason for this being time series forecasting requires at least four years of data, whereas the given dataset contains only three years of sales data.

Conclusion

It is reasonable to conclude that the dataset does not contain sufficient information in terms of variables to make an accurate prediction in terms of weekly store sales. The linear regression model does not work since all but one of the variables are business specific in nature or the set of 45 stores may have other factors like department-wise markdowns, store location etc influencing sales. On the other hand, the time series fails because at least four years of data is needed, as mentioned earlier.