Walmart Sales Prediction

UNIVERSITY OF SOUTH FLORIDA

ISM 6137

Statistical Data Mining

Project by Group 1

Shreyas Dalvi
Poonam Krishna
Rahul Panthappallil Jose
Shruti Pareek
Yashjeet Singh

ACKNOWLEDGEMENT

It is a great pleasure to have the opportunity to extend our heartiest felt gratitude to everybody who helped us throughout the course of this project.

It is distinct pleasure to express our deep sense of gratitude and indebtedness to our learned professor, **Dr. Anol Bhattacharjee** for their invaluable guidance and encouragement. With their continuous inspiration only, it becomes possible to complete this Project.

Shreyas Dalvi
Poonam Krishna
Rahul Panthappallil Jose
Shruti Pareek
Yashjeet Singh

TABLE OF CONTENTS

Contents

1. Executive Summary

Predicting the future sales for a company is the key aspect of any company in retail industry. This helps company to strategize their decisions and increase sales further. This project helps Walmart Inc. a big retailer in US to predict their Sales using their store details and other socioeconomic factors. Our goal from this project is to provide key factors boosting company's sales which retailer can control and make more profits in future.

We had data from 2010 to 2012, where 45 Walmart stores with corresponding department details for analysis. The data includes internal retailer factors like Store Type, Size etc. and external factors like Holidays, Unemployment index, consumer price index, Fuel Prices, Temperature, sales in various offer times given by Walmart. Walmart is having four holidays Super Bowl, Labor Day, Thanksgiving and Christmas.

We saw the relationship between the various external and internal factors influencing sales and noticed few factors which influence the weekly sales way higher those are Holidays, Store and Department details, offers given by Walmart. While some other factors like CPI, Unemployment rate and Temperature slightly influences the sales but can't be controlled by company. Collectively Walmart can inspect and control the main influencing factors like Store Size, Promotional offers while they can understand how external factors like holiday seasons, Unemployment rate, CPI (Consumer Price Index), Temperature and strategize their decisions like when to offer a Promotional offers or discounts which eventually maximize their profits.

We built around four models which can give us idea about various important factors boosting sales.
First linear model shows:

1) Promotional offers given with Holidays boosts sales
2) Walmart Superstores make more sales than other types of stores
3) Sales in Holiday seasons is more
4) Sales majorly depends on Stores and respective departments.

Second linear model shows:
Christmas and Thanksgiving have much more impact on sales when promotional offers are given when compared with other holidays.

Third Linear model:
Third linear model analyzed how weekly sales is impacted for various other factors like unemployment, CPI , temperature etc.

2. Problem Significance

Walmart Inc. is an American multinational retail corporation. It is one of the largest retail corporation in the world, which was founded by Sam Walton in year 1962. And, it has revenue over $485.87 billion dollars recorded in 2016. Walmart runs three types of stores based on number items namely hypermarkets, discount department stores, and grocery stores.

Since it is a huge competitor in the retail sector it is intriguing for us to find out what are the factors that drives the sales. Recently, we came across Walmart sales data on Kaggle. This data was for a competition posted by Walmart for recruiting. The challenge was to predict weekly sales of store located in 45 difference regions. Based on this data containing details of weekly sales, store size, department code, consumer price index, unemployment of that region and promotional markdown we would predict the sales of stores by department. Here we have 99 different departments such as Dry Grocery, Sporting Goods, Frozen Goods etc.

Our goal in this project is to predict the sales for Walmart stores by analyzing the store and other socioeconomic factors.

3. Data Source/Preparation

We had access to four different data sets from Kaggle.com about the company. These data sets contained information about the stores, departments, temperature, unemployment etc. Below are the Various CSV Datasets Metadata details:

Below are the Various CSV Datasets Metadata details:

Stores:

Name	**Description**	**Sample Data**
Store	Store number ranging from 1 to 45	1
Type	We have three types of Walmart Store. A - Walmart Hypermarkets B-DiscountStores C- Neighbourhood Markets	A
Size of Store	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	11000

Train:

Name	**Description**	**Sample Data**
Date	The date of the week where this observation was taken. Friday of that week. We had data from 2010 to 2012.	"2010-02-05"
Store	Store number ranging from 1 to 45	1
Dept	One of 1-99 that shows the department	12
IsHoliday	Boolean value representing a holiday week or not.	TRUE

Features:

Name	Description	Sample Data
Store	Store number ranging from 1 to 45	1
Date	The date of the week where this observation was taken. Friday of that week. We had data from 2010 to 2012.	"2010-02-05"
Temperature	Temperature of the region during that week	42.3
Fuel_Price	Fuel Price in that region during that week.	2.57
MarkDown1 - Markdown5	Represents the Type of markdown and what quantity was available during that week.	134.2
CPI	Consumer Price Index during that week.	211
Unemployment	The unemployment rate during that week in the region of the store	8.11
IsHoliday	Boolean value representing a holiday week or not.	TRUE

List of Holidays:

1) Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12
2) Labor Day: 10-Sep-10, 9-Sep-11, 7-Sep-12
3) Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12
4) Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12

As a part of data preparation, we merged all three data frames by the date provided using SQL in R.

1) Train
2) Features
3) Store

Also, We have generated variable “holiday” as descriptive variable giving detail listing of Holiday. This will be used to check if impact of sales is greater for specific holiday or not.

4. Hypothesis

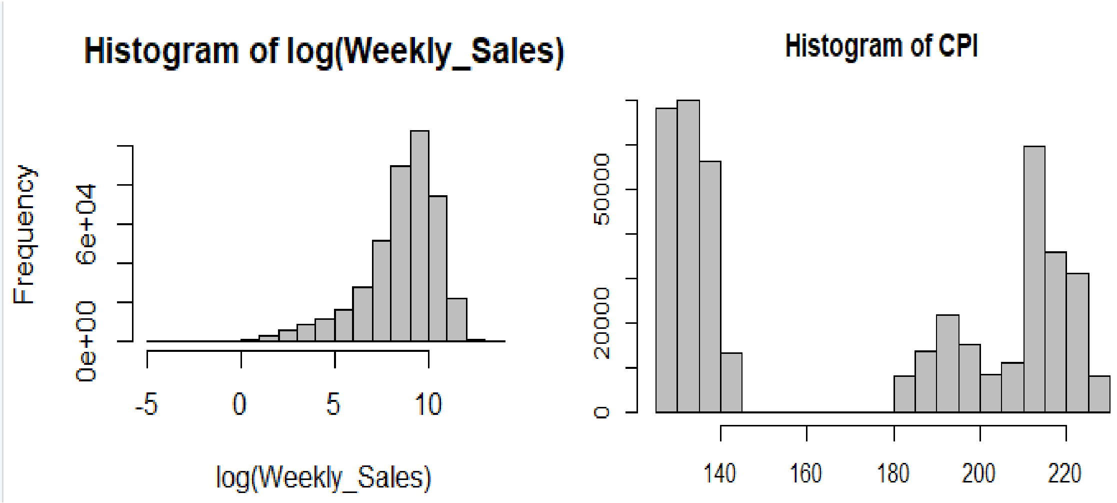
Hypothesis 1: Looking for combined impact for Holiday season and Promotional Offers
H01 = With increase in Markdown (Collective Promotional offers) in holiday seasons there is significant increase in the Weekly sales amount for respective Store and Department. (i.e. Interaction of IsHoliday and Markup)

Hypothesis 2: Looking for significance for various Holidays
H02 = For different holidays along with Markdowns there is significant change in the Weekly sales for respective Store and Department. (i.e. Interaction of Holiday Description and Markup)

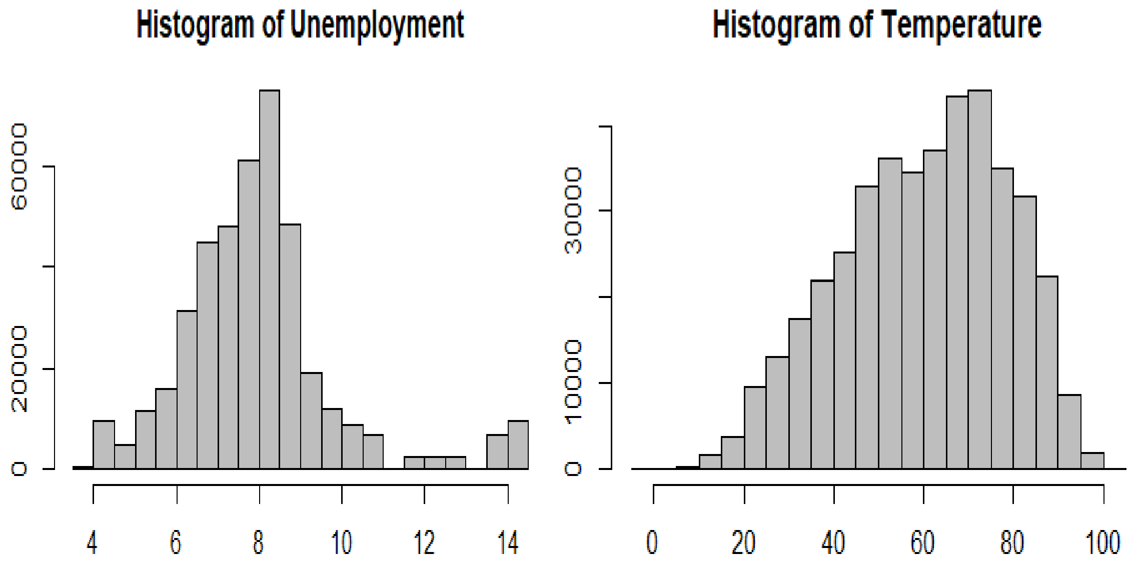
Hypothesis 3: Analyzing the external factors like unemployment, CPI , Temperature.
H03= Weekly sales would be lower significantly for increase in Unemployment rate and CPI index.

5. Descriptive Analysis

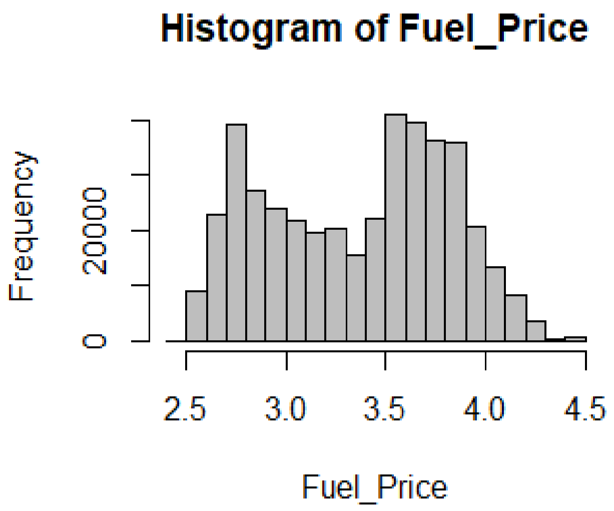
After merging we analyzed the Weekly Sales for three years. The total cumulative sales for all the stores are not continuously increasing through the three years. We plotted the distribution of various variables to be included in the linear regression to check the linearity. We have log transformed the Weekly Sales data as it was highly skewed.

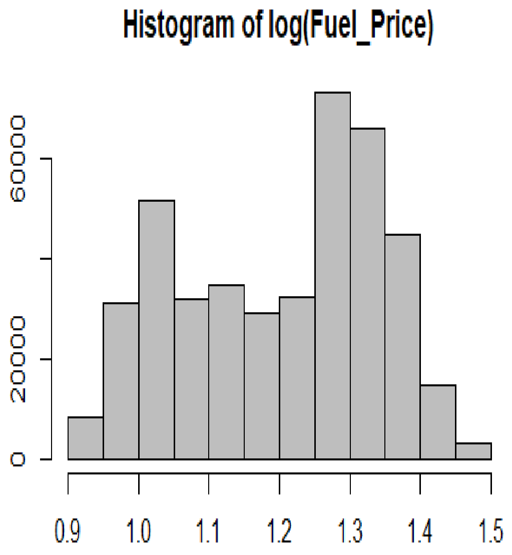


Below are the histograms for the Unemployment and Temperature, data seems to be in normal distribution.

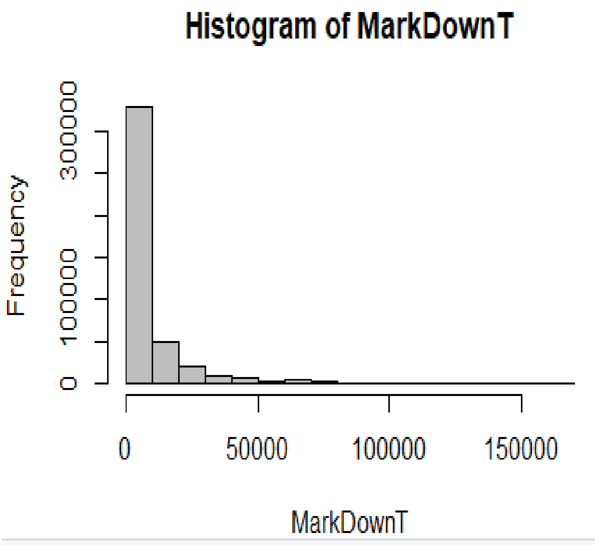


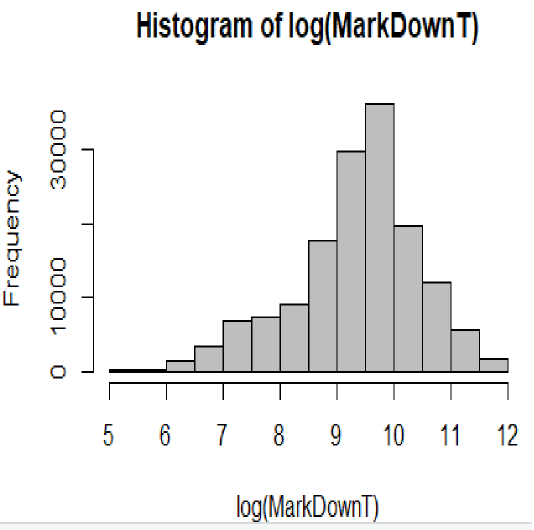
Looking at Fuel price histogram the data seems to be bimodal, so we have done log transform to use the variable in model. Employment and Temperature data seems to be normal. Fuel price seems to be bimodal, have log transformed it to make more normal.



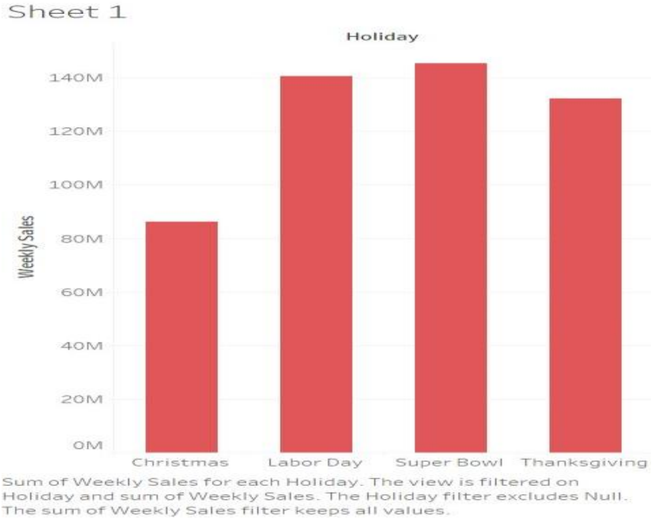


We have aggregated all the markdown variables to one variable MarkdownT. MarkdownT data seems to be highly skewed with some outliers, we have taken for log for MarkDownT. The histogram for log(MarkdownT) seems normal now.

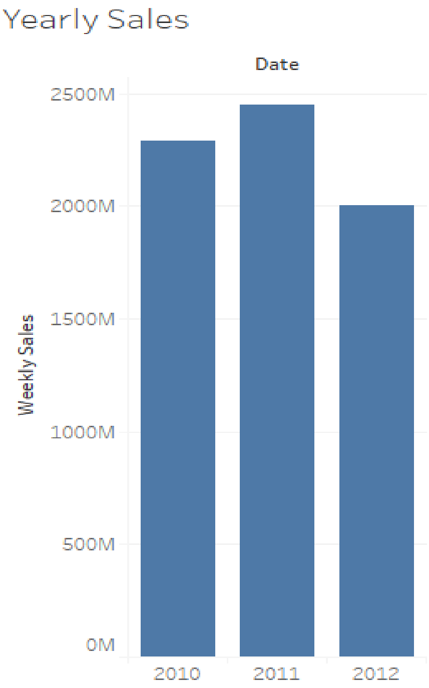




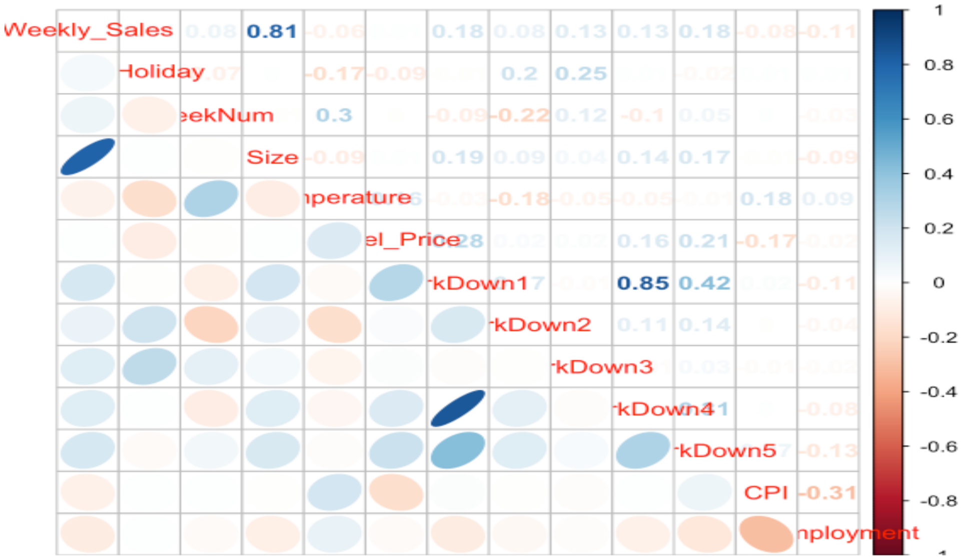
Weekly Sales vs Holiday



Weekly_Sales Vs Year



Correlation matrix for the independent variables:



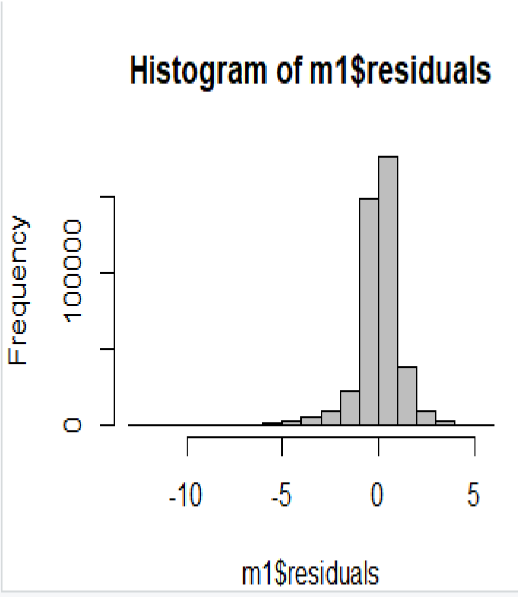
We don't have much correlation within the variables, and it rules out the multicollinearity issue in the model. Weekly sales is highly correlated with the Size of store. Also, MarkDown1 is correlated with MarkDown3, this will not affect the model as we are combining all the MarkDowns together to create MarkDownT variable.

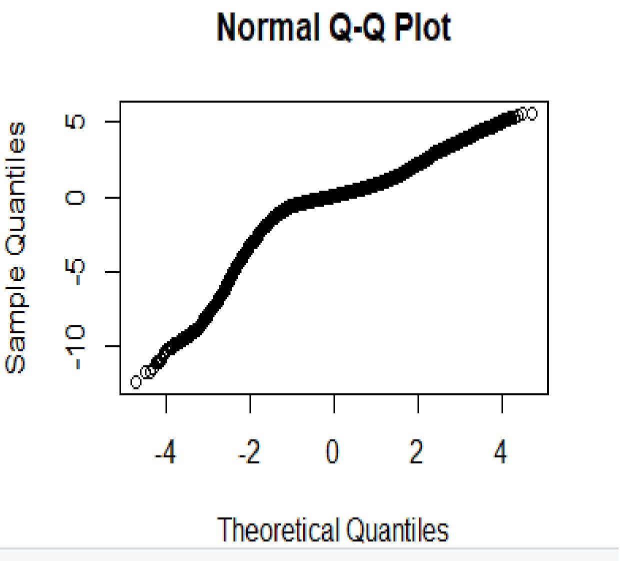
6. Regression Models:

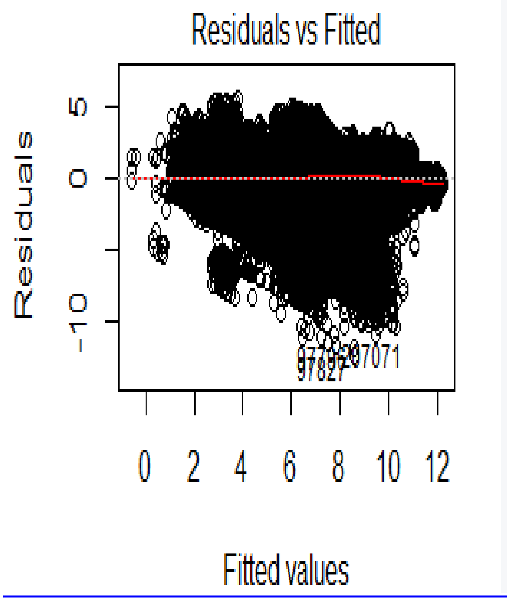
Model 1:
Basic model considering the impact of Holidays, unemployment, Markdown offers provided by Walmart and type of store, store, dept details.

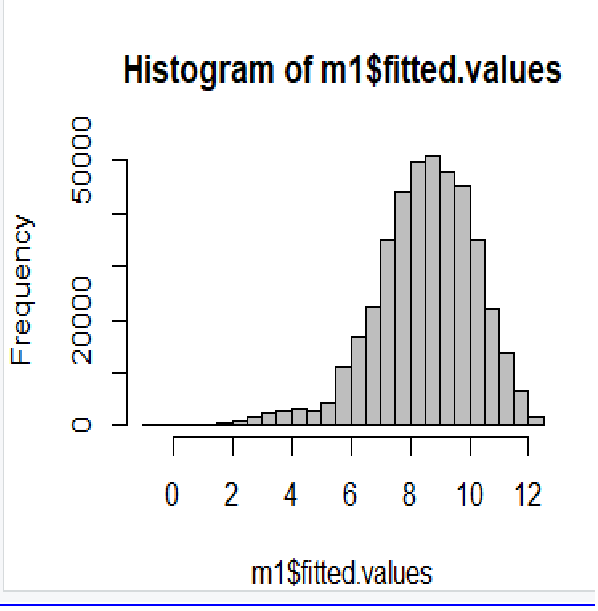
Variable Name	Estimate	Error Term
(Intercept)	10.3934058	0.0192183
Markdown	-0.0034345	0.0004321
Type B	-0.6586907	0.0171617
Type C	-2.8206726	0.0186562
Department 2	0.8584731	0.0213072
Department 3	-0.8371947	0.0213072
Department 4	0.3978068	0.0213072
…	…	…
Store 2	0.2824771	0.0168909
Store 3	-0.8393447	0.0177095
Store 4	0.356457	0.0168765
…	…	…
Holiday	0.0340828	0.0093572
Markdown*Holiday	0.0086489	0.0014656

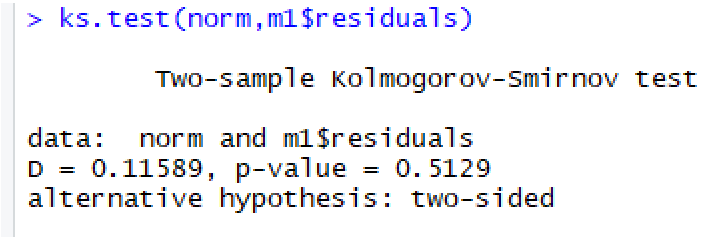
From this model we can see that, Weekly sales largely depends on Type of Store, Holiday week, and interaction between holiday and Markdown. Model explains 66% variability for the Weekly_Sales on the above stated variables.











```
> ks.test(norm,m1$residuals)

        Two-sample Kolmogorov-Smirnov test

data:  norm and m1$residuals
D = 0.11589, p-value = 0.5129
alternative hypothesis: two-sided
```

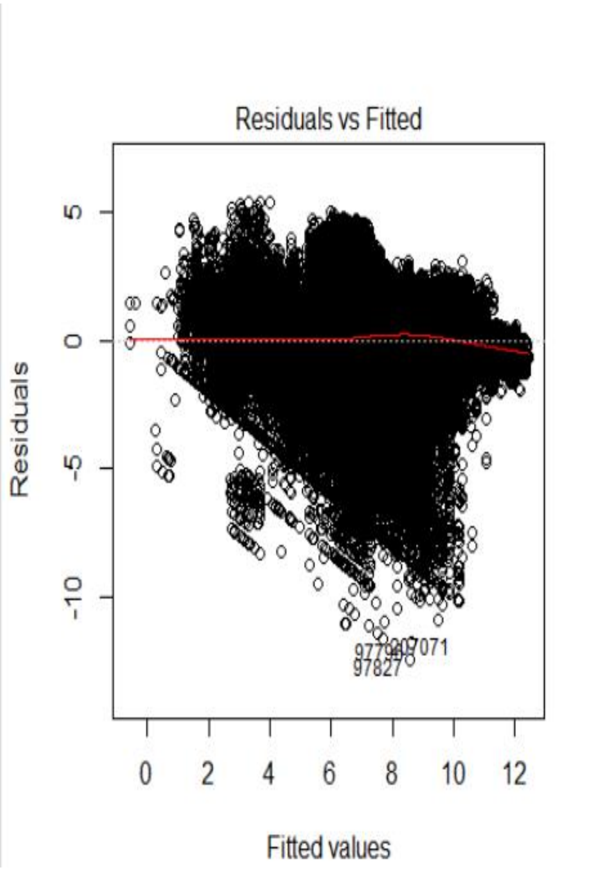
From the above graphs we could see that residuals are normally distributed. We have also performed KS test for residuals normality (P value is 0.5129). Residual vs Fitted plot shows that model is homoskedastic in nature.

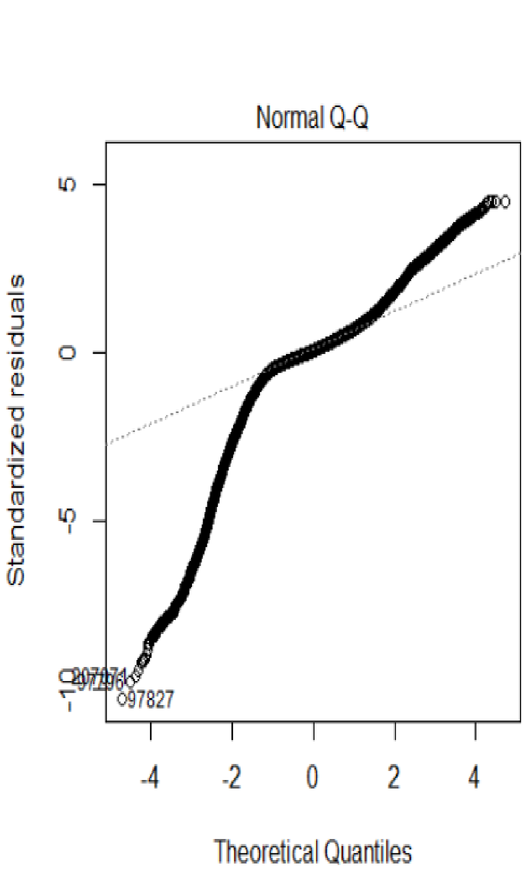
Model2:

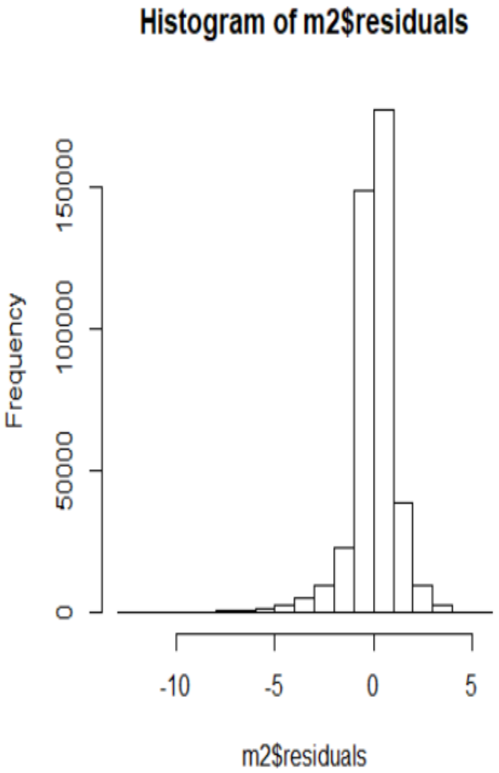
Variable Name	Estimate	Error Term	Significance
(Intercept)	10.0667756	0.037382	***
Markdown	0.0003476	0.0005694	
Department 2	0.8584731	0.0212952	***
Department 3	-0.8371947	0.0212952	***
Department 4	0.3978068	0.0212952	***
…	…	…	
Store 2	0.2815218	0.0168815	***
Store 3	-1.478656	0.0175554	***
Store 4	0.4239211	0.0181198	***
…	…	…	
Christmas	-0.1310742	0.0222782	
Labor Day	-0.0456749	0.0158823	**

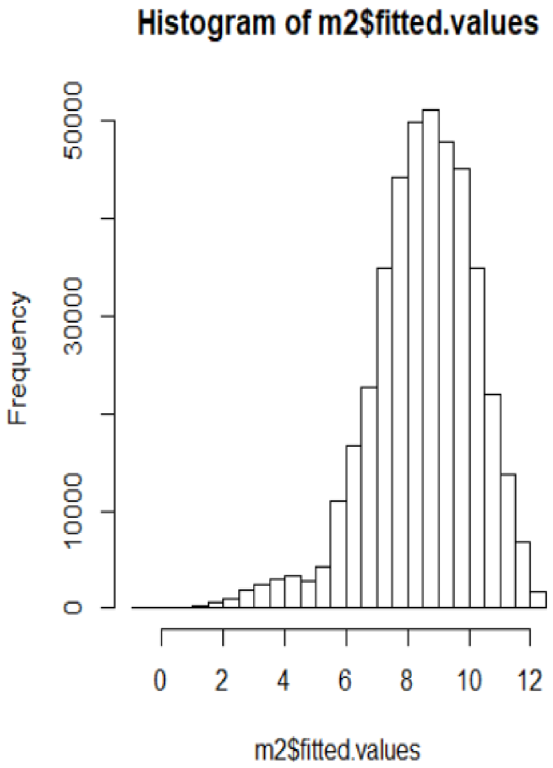
Super Bowl	0.0724526	0.0158929	***
Thanksgiving	0.2781413	0.0221549	***
Markdown*Christmas	0.0150538	0.0030044	***
Markdown*Labor Day	0.0010347	0.0029276	
Markdown*Super Bowl	0.0032744	0.0027019	
Markdown*Thanksgiving	0.0050197	0.0029585	

Model2 explains about 67% of variability using the above variables. We could see that Thanksgiving holiday has highest impact on the Weekly Sales followed by SuperBowl , Christmas and Labour Day considering rest all other effects. Each Store and Department is having its significance on the Weekly sales.









Residual plot is having approximate mean as 0 but its having high kurtosis. Fitted values are slightly left skewed as seen in fitted values histogram this also affects the QQ plot fit line.

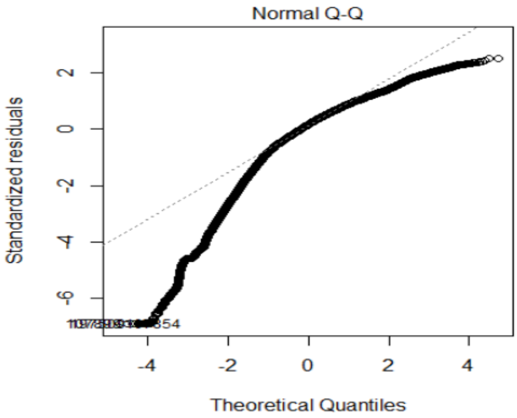
<u>Model 3:</u>

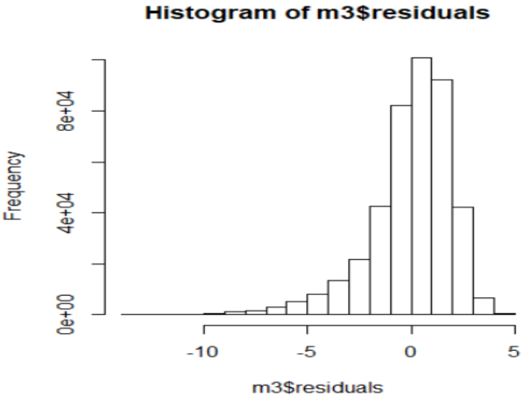
Variable Name	Estimate	Error Term	Significance
Intercept	-4.4230	0.0681	***
Markdown	-0.0026	0.0007	***
Size	1.1290	0.0052	***
Unemployment	-0.0259	0.0018	***
CPI	-0.0002	0.0001	*
Holiday	0.0177	0.0156	

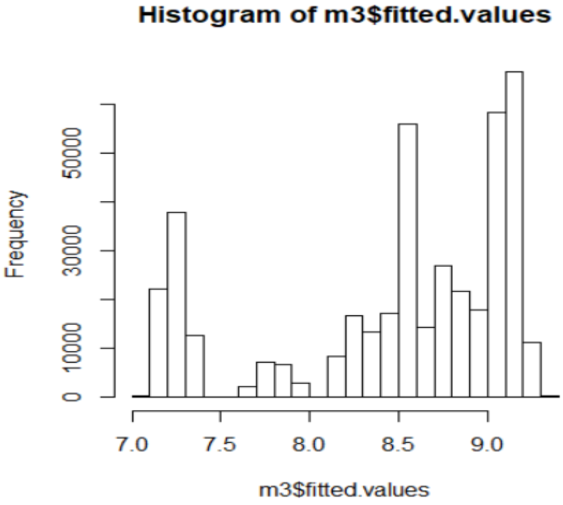
Temperature	-0.0004	0.0002	*
Markdown*Holiday	0.0059	0.0024	*

Model 3 is used to identify the relationship between the external factors and Weekly sales. Unemployment, CPI is having negative impact on the Weekly_Sales. Analysing these factors in the Holiday and Markdown Offer times would help the customer to predict or maximize their sales.

This model explains weak variability of 10%, but it gives important information regarding the external factors affecting the sales for walmart.







This model gives us some significance related to the other socioeconomic factors affecting the sales, but the strength of model is weak. Fitted values don’t follow normal nature rather they have multi-modal distribution. Residuals are left skewed. We won’t be able to accurately predict the Weekly sales by using this model.

Model Comparison:

Model	**Variability Explained**	**AIC**	**BIC**
Model1	66%	1356244	1357657
Model2	66%	1355774	1357354
Model3	10%	1777185	1777283

Interpretation:
Comparing Model1, Model2 and Model3, Model2 is the simple and explains good variability using independent variables, its also having low AIC and BIC.

Conclusion:
Model 2 would be the good choice among these three, as it explains the individual importance of holidays and its having less AIC and BIC values which are one of the key parameters of model stability and strength. These model still have some issues with homoscedasticity and may better work with the Fixed effect model , Time series modelling.
Quality Check
For the quality check we used Residual plots, various statistical tests and plots to verify the strength of the model.

6. Recommendations

Recommendation from the above model interpretations are as below :

1) To increase the Markdown offers during all the Holidays.
2) Increase supply to larger stores in Holiday seasons.
3) Expand the No of Supermarkets.

Appendix:

```
Call:
lm(formula = train_final$lg_wk_sales ~ lg_markdownT + as.factor(Type) +
    as.factor(Dept) + as.factor(Store) + as.factor(IsHoliday) +
    as.factor(IsHoliday) * lg_markdownT)

Residuals:
     Min      1Q  Median      3Q     Max
-12.4768  -0.3477  0.0664  0.5625  5.6481

Coefficients: (2 not defined because of singularities)
                             Estimate Std. Error  t value Pr(>|t|)
(Intercept)                 10.3934058  0.0192183  540.806  < 2e-16 ***
lg_markdownT                 -0.0034345  0.0004321   -7.949 1.89e-15 ***
as.factor(Type)B            -0.6586907  0.0171617  -38.381  < 2e-16 ***
as.factor(Type)C            -2.8206726  0.0186562 -151.192  < 2e-16 ***
as.factor(Dept)2            0.8584731  0.0213072   40.290  < 2e-16 ***
as.factor(Dept)3            -0.8371947  0.0213072  -39.292  < 2e-16 ***
as.factor(Dept)99           -7.0584064  0.0439095 -160.749  < 2e-16 ***
as.factor(Store)2           0.2824771  0.0168909   16.724  < 2e-16 ***
as.factor(Store)3           -0.8393447  0.0177095  -47.395  < 2e-16 ***
as.factor(Store)4           0.3564570  0.0168765   21.122  < 2e-16 ***
……
as.factor(IsHoliday)TRUE                0.0340828  0.0093572          3.642  0.00027 ***
lg_markdownT:as.factor(IsHoliday)TRUE  0.0086489  0.0014656     5.901 3.61e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.209 on 421442 degrees of freedom
Multiple R-squared:  0.6695,  Adjusted R-squared:  0.6694
F-statistic:  6721 on 127 and 421442 DF,  p-value: < 2.2e-16

> AIC(m1)
[1] 1356244
```

Model2:

```
Call:
lm(formula = train_final$lg_wk_sales ~ lg_markdownT +
    as.factor(Dept) + as.factor(Store) + as.factor(Holiday) +
    lg_markdownT * as.factor(Holiday))

Residuals:
     Min      1Q  Median      3Q     Max
-12.4606  -0.3470  0.0673  0.5624  5.4270
```

```
Coefficients:
                                      Estimate Std. Error  t value Pr(>|t|)
(Intercept)                          10.0667756  0.0373820  269.295  < 2e-16 ***
lg_markdownT                          0.0003476  0.0005694    0.611 0.541527
as.factor(Dept)2                      0.8584731  0.0212952   40.313  < 2e-16 ***
……
as.factor(Dept)99                    -7.0556844  0.0438856 -160.775  < 2e-16 ***
as.factor(Store)2                     0.2815218  0.0168815   16.676  < 2e-16 ***
as.factor(Store)3                    -1.4786560  0.0175554  -84.228  < 2e-16 ***
as.factor(Store)4                     0.4239211  0.0181198   23.395  < 2e-16 ***
……
as.factor(Store)45                   -0.7014568  0.0176596  -39.721  < 2e-16 ***
as.factor(Holiday)Christmas          -0.1310742  0.0222782   -5.884 4.02e-09 ***
as.factor(Holiday)Labor Day          -0.0456749  0.0158823   -2.876 0.004030 **
as.factor(Holiday)Super Bowl          0.0724526  0.0158929    4.559 5.15e-06 ***
as.factor(Holiday)Thanksgiving        0.2781413  0.0221549   12.554  < 2e-16 ***
lg_markdownT:as.factor(Holiday)Christmas  0.0150538  0.0030044    5.011 5.43e-07
***
lg_markdownT:as.factor(Holiday)Labor Day  0.0010347  0.0029276    0.353 0.723754

lg_markdownT:as.factor(Holiday)Super Bowl      0.0032744  0.0027019     1.212
0.225544
lg_markdownT:as.factor(Holiday)Thanksgiving  0.0050197  0.0029585    1.697 0.089761 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.208 on 421435 degrees of freedom
Multiple R-squared:  0.6698,  Adjusted R-squared:  0.6697
F-statistic:  6381 on 134 and 421435 DF,  p-value: < 2.2e-16

> AIC(m2)
[1] 1355774
```

Model 3:

```
Call:
lm(formula = train_final$lg_wk_sales ~ lg_markdownT + log(Size) +
    Unemployment + CPI + as.factor(IsHoliday) + as.factor(IsHoliday) *
    lg_markdownT + Temperature)

Residuals:
     Min       1Q   Median      3Q      Max
-13.7405  -0.8792   0.3377  1.3521   4.9926

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -4.423e+00  6.811e-02 -64.948  < 2e-16 ***
lg_markdownT            -2.619e-03  7.270e-04  -3.603 0.000315 ***
```

```
log(Size)                          1.129e+00  5.245e-03 215.252  < 2e-16 ***
Unemployment                      -2.593e-02  1.793e-03 -14.464  < 2e-16 ***
CPI                               -2.130e-04  8.442e-05  -2.524 0.011614 *
as.factor(IsHoliday)TRUE           1.772e-02  1.556e-02   1.139 0.254615
Temperature                       -3.862e-04  1.743e-04  -2.216 0.026675 *
lg_markdownT:as.factor(IsHoliday)TRUE  5.887e-03  2.414e-03   2.438 0.014765 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.991 on 421562 degrees of freedom
Multiple R-squared:  0.1024,  Adjusted R-squared:  0.1023
F-statistic:  6867 on 7 and 421562 DF,  p-value: < 2.2e-16

> AIC(m5)
[1] 1777185
```