Abstract:

Walmart is one of the biggest supermarket chains in the United States. Annually, Walmart earns billions of dollars in sales. However, huge unrealized potential profits can be missed due to external factors such as economic recessions and climate. Natural events and high unemployment rate can cause certain branches to close due to weak profits. Thus, we would like to examine the relationship between the external factors and profit though basic data analysis techniques and conduct predictive statistical analysis to forecast future profits.

Table of Contents

1. Introduction

One challenge in modeling retail data is the need to make decisions based on limited history. Since festivals such as Christmas only occur once a year, it is difficult to see how strategic decisions will affect sales. In this regard, data analytics is important to help businesses understand the issues they face and explore data in a meaningful way.

Furthermore, predicting future sales for retail is one of the most important aspects of strategic planning. Organizations use predictive analytics to turn data into future insights, which is critical for making better-informed business decisions. With predictive analytics, companies can anticipate market trends, optimize sales resources, design effective personalized offers, and more.

In our project, we will be using the historical weekly sales data for 45 Walmart stores in different regions, paired with other variables such as store type and size, consumer price index (CPI), unemployment rate, temperature of the region, fuel costs, and whether the week is a special holiday week.

Based on this dataset, we seek to answer the following popular questions around the weekly Walmart sales:

1. Does the size of the stores affect weekly sales?
2. Does the type of the stores affect the weekly sales?
3. Does the temperature in the region the weekly sales?
4. Does fuel price in the Walmart affect the weekly sales?
5. Does CPI in the region affect the weekly sales?
6. Is there a single most important type and size of the store that most important in affecting the weekly sales in the store?
7. Does the weekly sales increase when the week is a special holiday week?
8. How is unemployment rate related to the sales in the stores?

This report will cover the data description and analysis using R language. For each of our research objectives, we performed statistical analysis and drew conclusions in the most appropriate approach, together with explanations and elaborations. Additionally, our project seeks to develop a statistical model based on the historical dataset, to predict the weekly sales for each store.

2. Data Description

The dataset for the historical weekly sales for the 45 Walmart stores is obtained from Kaggle.com, titled as “Walmart Recruiting – Store Sales Forecasting”. The data collected ranges from 2010 to 2012, and it consists of four csv data frames, titled “stores.csv”, “train.csv”, “test.csv”, “features.csv”.

This dataset is open for competition in the public in 2014. In total, 688 teams and 688 competitors participated in this competition and there were 12,240 entries received for submission. Currently, this dataset is used for study and research purposes.

Before proceeding to data analysis, we first performed a preliminary data cleaning to ensure that:

- Irrelevant columns in features are eliminated, eg. “Markdown1”, “Markdown2”,…, “Markdown5”
- Boolean data type is converted into numeric type for *IsHoliday* variable (TRUE=1 and FALSE=0) across train and features
- Weekly sales of all departments for each store per week were combined in features
- Date in train and features represented as a string data type is split into Year, Month and Day
- Included a column “Week” retrieved from Date in train and features
- Removed data in features from Weeks 144 – 169 not reflected in train

After all the preparation, **6435** observations (weekly sales) with **14** variables are retained for analysis.

1. Stores.csv
   - Store - Store number
   - Type – Type of stores (A, B, or C)
   - Size – Size of stores
2. Train.csv
   - Weekly_Sales – Sales for the given store in that week (Main Variable of interest)
   - Week – Particular week in the timeframe
   - Date – Date when *Weekly_Sales* was recorded for the different stores
   - Year – Year when *Weekly_Sales* were recorded for the different stores
   - Month – Month of the year when *Weekly_Sales* were recorded for the different stores
   - Day – Day of the month when *Weekly_Sales* were recorded for the different stores

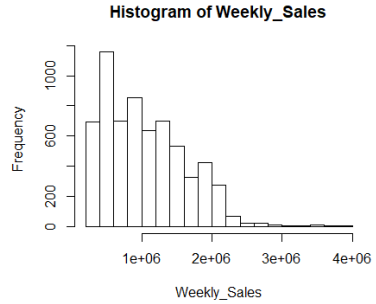
3. Features.csv
    - Temperature – Average temperature in the region during that week
    - Fuel_Price – Cost of fuel in the region during that week
    - CPI – Consumer Price Index during that week
    - Unemployment – Unemployment rate in the region during that week
    - IsHoliday – Whether that week is a special holiday week (1= Yes, 0 = No).

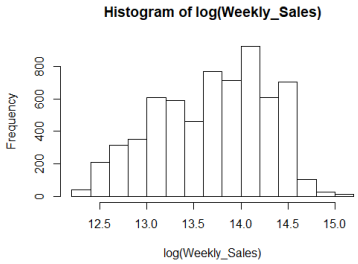
3. Description and Cleaning of Dataset

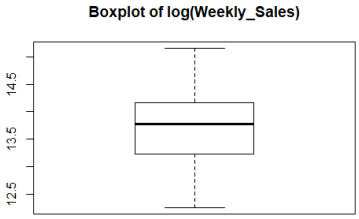
In this section, we shall investigate the data in more detail. Each variable is investigated individually to look for possible outliers, and/or to perform a transformation to avoid highly skewed data.

3.1 Summary statistics for the main variable of interest, Weekly_Sales

The following plots show the overall distribution of the variable *Weekly_Sales*







Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
12.25	13.22	13.78	13.86	14.17	15.16

It appears that the main variable *Weekly_Sales* is highly skewed, hence we apply a log-transformation (base *e*) to the variable.

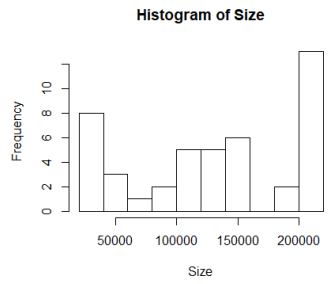
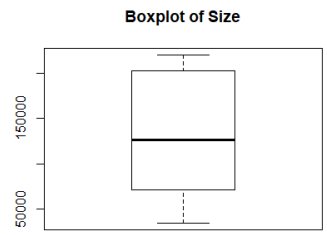
The histogram and boxplot of the log-transformed variable, *log*(*Weekly_Sales)* are shown below with summary statistics. The log-transformed data appears to be more normally distributed as it is more symmetric. Furthermore, there are no outliers in the boxplot of the log-transformed data that is required to be removed.

We shall proceed to the next session with this transformed dataset.

3.2 Summary statistics for other variables

The histogram, the boxplot, and the transformation applied are tabulated in the following subsections. Some additional observations about the data and its outliers are recorded.

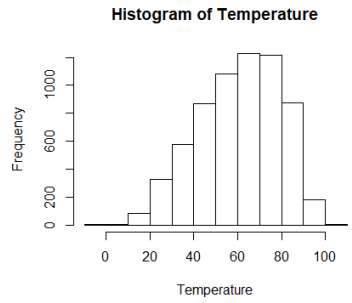
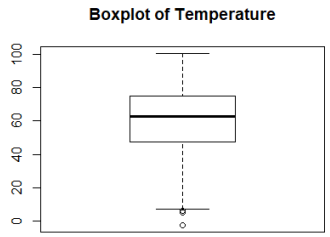
3.2.1 Size of store, Store

		• No outlying values of *Size* is removed • There seems to be more stores that have size larger than 200,000 square feet

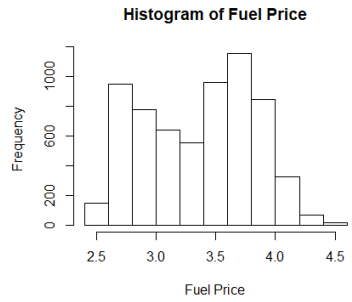
3.2.1 Type of stores, Type

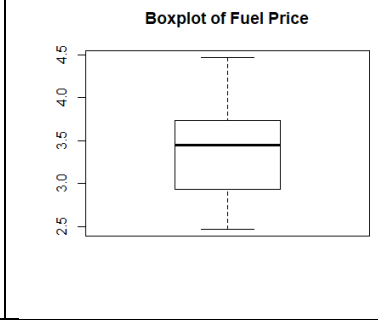
	• Most stores are of type A, and there are a little number of stores of type C.

3.2.3 Average weekly temperature, Temperature

		• There exist outliers with temperatures of -2.06, 5.54, 6.23 • However, we will not remove the outliers as it could affect our data analysis of *Weekly_Sales* and affect the accuracy of the data since some regions can have extreme temperatures.

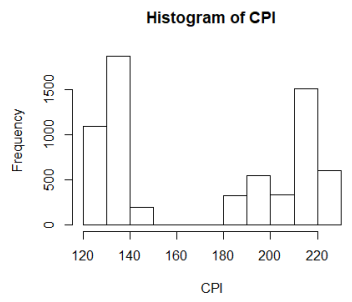
3.2.4 Average weekly fuel price, Fuel_Price

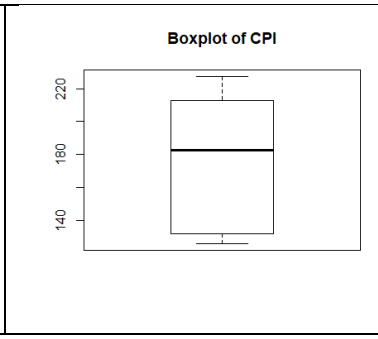




- No outlying fuel price is removed.
- Most of the fuel prices range within 3.5 to 4.0.

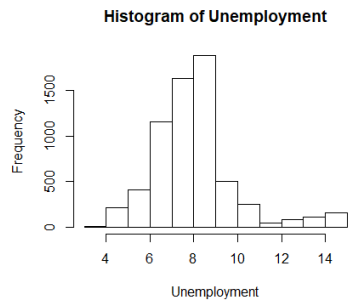
3.2.5 Average weekly CPI, CPI

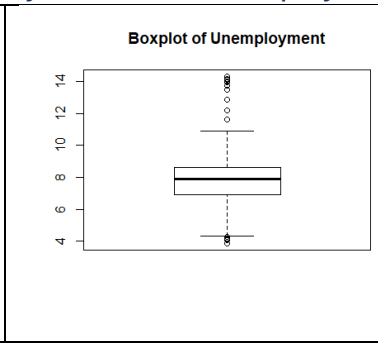




- No outlying CPI is removed.

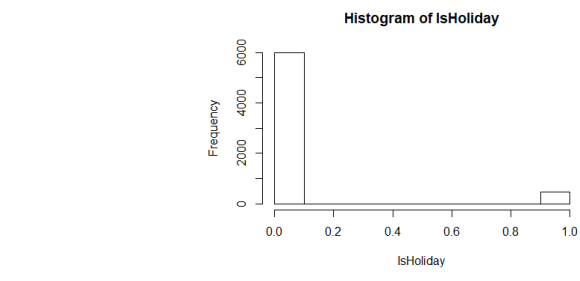
3.2.6 Average weekly unemployment rate, Unemployment





- There are many outliers in Unemployment.
- However, just like for Temperature, removing the outliers would negatively affect the data analysis of *Weekly_Sales*.

3.2.7 Holiday week, IsHoliday



- There are significantly more times that the week is not a holiday.

3.3 Final Dataset for Analysis

Based on the above analysis, the dataset still has 6435 observations, and the log transformation is applied to the main variable *Weekly_Sales*.

4 Statistical Analysis

4.1 Correlations between log(Weekly_Sales) and other continuous variables

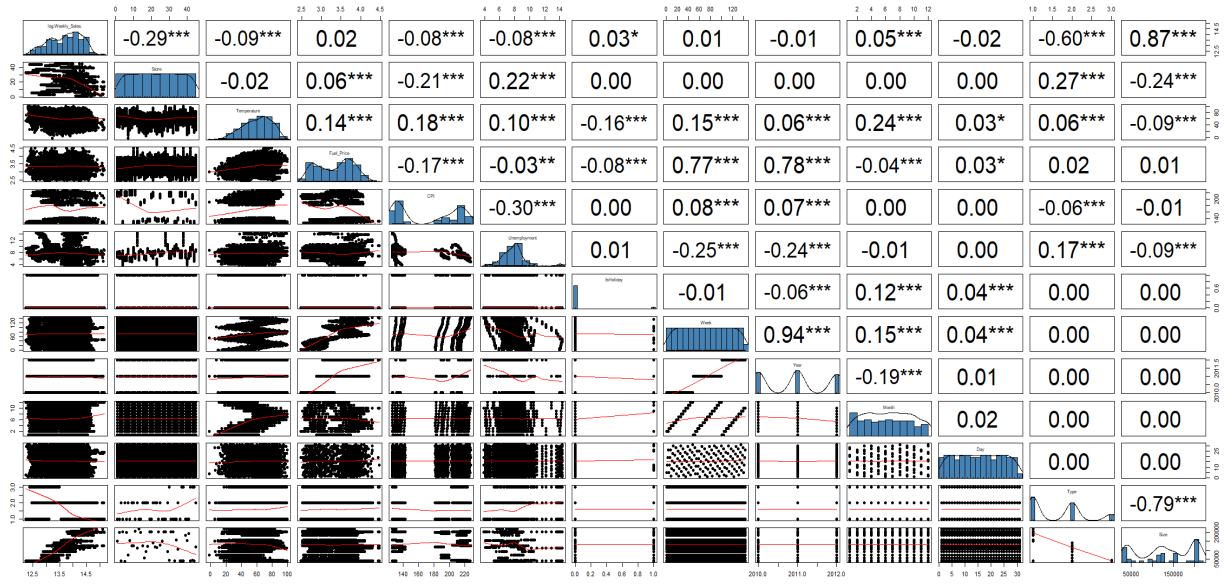


Figure 4.1: Correlation Plot between all variables in the final dataset

Scatter plots and correlation coefficients are useful in studying the possible linear relationships between store's *Weekly_Sales* and other variables.

Based on the correlation plot obtained in Figure 4.1, we make some basic observations between the main variable Weekly Sales and the other variables in our final dataset.

- Weekly sales and Size of store are highly positively correlated (r = 0.87)
- Weekly sales and the Store(number) are moderately negatively correlated (r = -0.29)
- Weekly sales and Type of store are highly negatively correlated (r=-0.60)
- Weekly sales and Unemployment are negatively correlated (r = -0.08)
- Weekly sales and CPI are negatively correlated (r=-0.08)
- Weekly sales and Temperature are negatively correlated (r=-0.09)
- Weekly sales and Month are positively correlated (r=0.05)

Additionally, from the correlation plot, we can answer the two questions:

1. “Does the weekly sales increase when the week is a special holiday week?”

   Since *log(Weekly_sales)* and *IsHoliday* is positively correlated (r=0.03), we conclude that weekly sales increases when the week is a special holiday week.

2. “How is unemployment rate related to the sales in the stores?”

   Since *log(Weekly_sales)* and *Unemployment* is negatively correlated (r=-0.08), we conclude that as unemployment rate increases, weekly sales in the stores decreases.

Some interesting observations amongst the other variables of interest are:

- Fuel Price and Week are highly positively correlated (r = 0.77)
- Fuel Price and CPI are negatively correlated (r=-0.17)
- Temperature and Fuel Price are positively correlated (r = 0.14)
- CPI and Temperature are positively correlated (r = 0.18)

4.2 Statistical Tests

We classify the remaining variables in two sections: internal and external variables. Our analysis will focus on these two factors individually. The internal variables include size of store and type of store. These factors are decided upon and controlled by Walmart internally. External variables include Temperature, CPI, Fuel Price and other factors of time such as day or month. These variables are not directly controlled by Walmart. Pairwise analysis of log(Weekly_Sales) and each variable is conducted below.

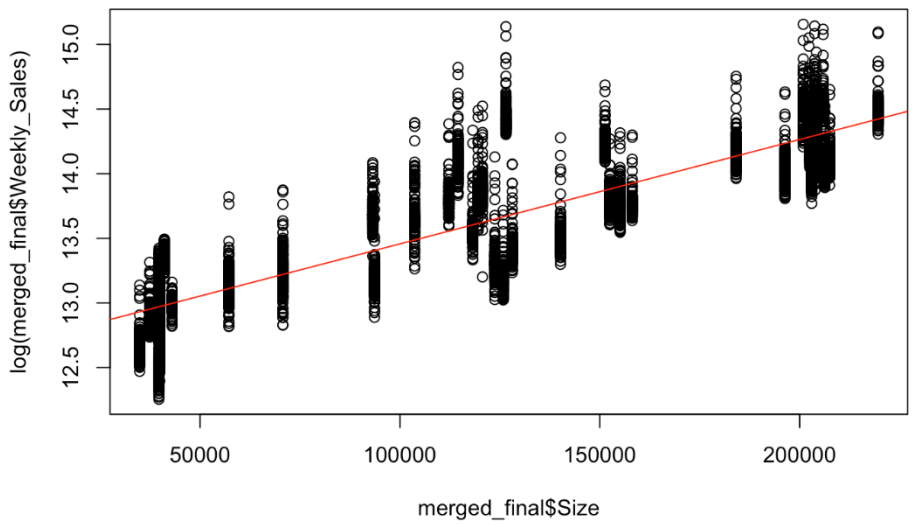
4.2.1 Relation between *log(Weekly_Sales)* and Size of Store

In this section, we try to answer the question “Is the weekly sales of a store dependent on the size of store?”

It is observed from Figure 4.1 that *log(Weekly_Sales)* and Size of store is highly positively correlated with a correlation coefficient of r = 0.87, which implies that *log(Weekly_Sales)* increase with an increase in size of store. This shows that larger stores are able to generate higher weekly sales compared to smaller stores. In order to study this relationship further, a simple linear regression is performed between *log(Weekly_Sales)* and the size of the store.

The regression model gives a p-value of 2e-16, which shows that the relationship between *log(Weekly_Sales)* and Size is statistically significant. Further, the R-squared value of 0.7507 indicates that Size explains about 75% of the variation in *log(Weekly_Sales)*, thus reinforcing the strong correlation coefficient between the two variables as seen in the correlation plot in Figure 4.1

Therefore, the Size of the store statistically affects the weekly sales of Walmart and explains about 75.07% of the variation in weekly sales.

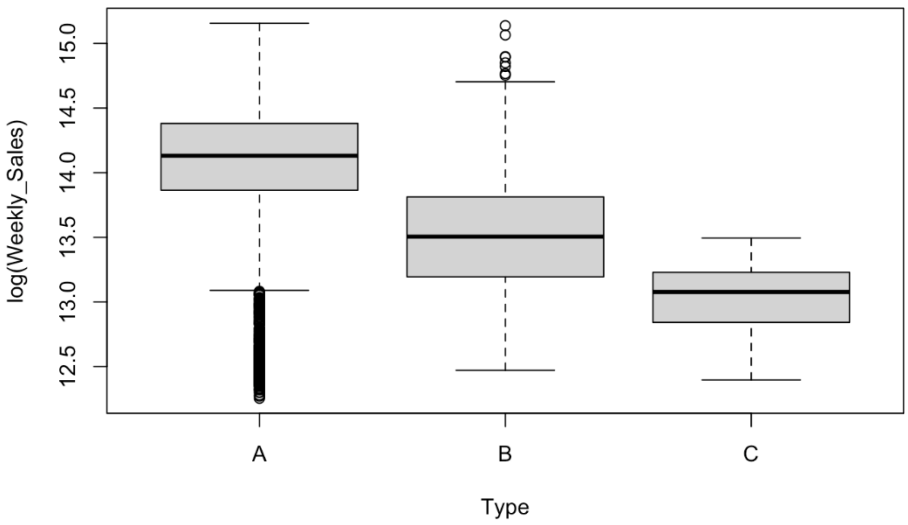



```
Call:
lm(formula = log(Weekly_Sales) ~ Size, data = merged_final)

Residuals:
     Min       1Q   Median       3Q      Max
-0.71549 -0.21936 -0.03457  0.20070  1.46560

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.265e+01  8.401e-03  1505.8   <2e-16 ***
Size        8.076e-06  5.803e-08   139.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2938 on 6433 degrees of freedom
Multiple R-squared:  0.7507,    Adjusted R-squared:  0.7506
F-statistic: 1.937e+04 on 1 and 6433 DF,  p-value: < 2.2e-16
```

4.2.2 Relation between *log(Weekly_Sales)* and Type of Store

In this section, we want to understand the relationship between Weekly Sales and the Type of Store. We want to answer the question “Is the weekly sales of a store dependent on the type of store?”. The dataset categorizes the store into 3 types: A, B and C.

An ANOVA test will be conducted to determine whether *log(Weekly_Sales)* is dependent on the type of Store. The following plot gives us a boxplot which depicts the distribution of *log(Weekly_Sales)* for the three different store types:



From visual judgement, it is apparent that the mean Weekly Sales for each type of store are vastly different, with Type A showing the most weekly sales, followed by Type B and Type C. To statistically analyze this inequality of means, an ANOVA Test is appropriate.

We test the following hypotheses:

$H_0: \mu_A = \mu_B = \mu_C$ against $H_1: \mu_i \neq \mu_j$ for i,j = A, B, C

```
              Df Sum Sq Mean Sq F value Pr(>F)
Type           2  806.5   403.2    1826 <2e-16 ***
Residuals   6432 1420.3     0.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA test produces a p-value of 2e-16, and so we reject the null hypothesis at the 5% significance level (as well as 1% significance level). Thus, the mean weekly sales are not equal for all store Types.

A pairwise t-test is used to analyze the pairwise relationship between all three types of stores as shown below:

```
        Pairwise comparisons using t tests with pooled SD

data:  logweeklysales and factortype

  A      B
B <2e-16 -
C <2e-16 <2e-16

P value adjustment method: none
```

From the pairwise t-test, it is evident that the mean weekly sales for each type of store are vastly different from each other, which confirms our observations from the boxplot. Therefore, we can conclude that the weekly sales of a store is dependent on the type of store (A, B and C).

4.2.3 Relation between *log(*Weekly_*Sales) and* Temperature

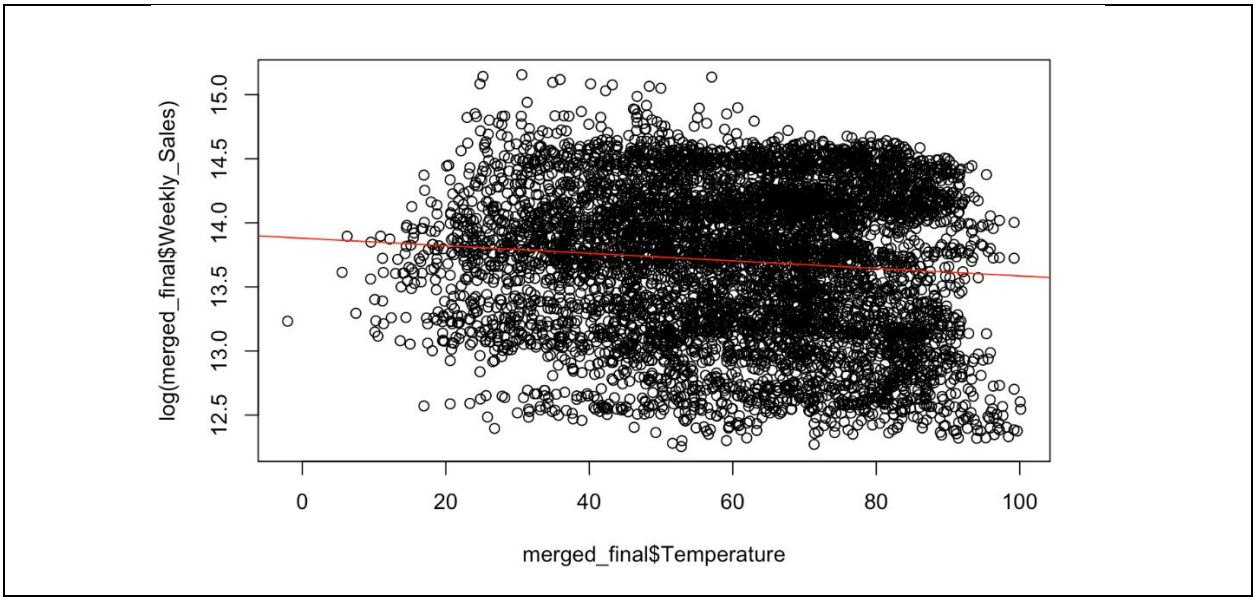
In this section, we determine whether the weekly sales of a store depend on the temperature of the region.

Log(Weekly_Sales) and Temperature is negatively correlated with a coefficient of –0.09, which implies that with an increase in temperature, there is a decrease in *log(Weekly_Sales)* and vice versa. This makes logical sense as if it is too hot outside, people may choose to not go out as compared to when temperatures are lower and more pleasant. However, if it is too cold (i.e., temperatures are very low), people would not go out either and this would also impact the weekly sales Walmart stores.

We perform a linear regression between *log(Weekly_Sales)* and Temperature to further analyze the relationship between the two variables.

The regression model displays a slightly downward trend in *log(Weekly_Sales)* as temperatures rise. The p-value of temperature in the regression model is given as 1.04e-13, which indicates that the variable is statistically significant at 0.05 level of significance. However, the R-sq value at 0.856% indicates that temperature explains about 0.856% of the variation in *log(Weekly_Sales)*.

Thus, while temperature statistically affects *log(Weekly_Sales)*, it only explains about 0.86% of the variation in *log(Weekly_Sales)*. It is not practically significant.



```
Call:
lm(formula = log(Weekly_Sales) ~ Temperature, data = merged_final)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4703 -0.4729  0.0624  0.4831  1.4245

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.880921   0.025106 552.894  < 2e-16 ***
Temperature -0.002951   0.000396  -7.453 1.04e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5858 on 6433 degrees of freedom
Multiple R-squared:  0.00856,   Adjusted R-squared:  0.008406
F-statistic: 55.54 on 1 and 6433 DF,  p-value: 1.036e-13
```

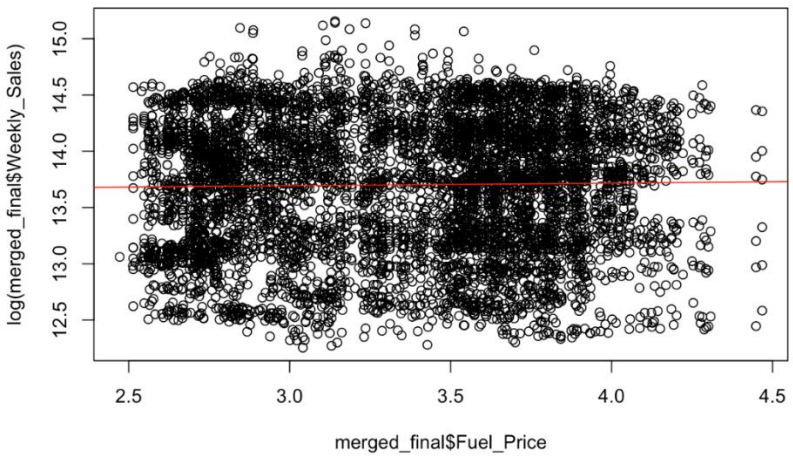
4.2.4 Relation between log(Weekly_Sales) and Fuel Price

In this section, we determine whether the weekly sales of a store depend on the fuel price of the region.

Weekly Sales and Fuel price is slightly positively correlated with a coefficient of 0.02, which implies that if there is an increase in Fuel Price, there would be an increase in weekly sales. However, the correlation is not very significant.

A simple linear regression model implemented produces a p-value of 0.132, which is higher than an alpha of 0.10. Thus, Fuel Price is not a statistically significant variable. Its R-sq value too is too low (0.035%) which indicates that it explains very little variation in *log(Weekly_Sales)*.

Hence, there is not much of a relationship between Fuel Price and *log(Weekly_Sales)*. It is not practically significant.



```
Call:
lm(formula = log(Weekly_Sales) ~ Fuel_Price, data = merged_final)

Residuals:
     Min       1Q   Median       3Q      Max
-1.43947 -0.47916  0.07093  0.46342  1.45874

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.62111    0.05416 251.507   <2e-16 ***
Fuel_Price   0.02406    0.01598   1.506    0.132
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5882 on 6433 degrees of freedom
Multiple R-squared:  0.0003523, Adjusted R-squared:  0.0001969
F-statistic: 2.267 on 1 and 6433 DF,  p-value: 0.1322
```

4.2.5 Relation between log(Weekly_Sales) and CPI

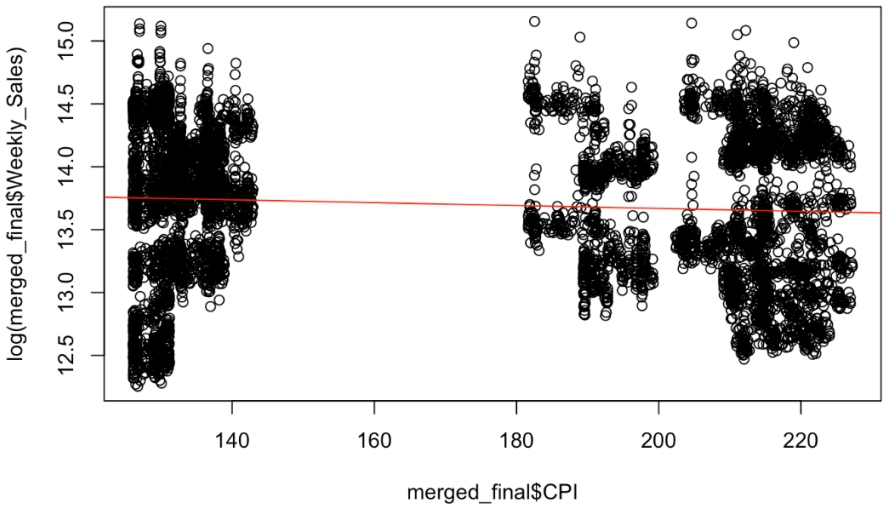
In this section, we determine whether the weekly sales of a store are dependent on the Consumer Price Index (CPI).

Log(Weekly_Sales) and CPI are slightly negatively correlated with a correlation coefficient of –0.08. This is reasonable, since if the consumer price index increases, i.e., the cost of goods increases, the weekly sales would decrease. To study this relationship further, we implement a simple linear regression between *log(Weekly_Sales)* and CPI.

The scatter plot displays a slight downward trend, in correspondence with the negative correlation coefficient between the two variables. The regression model displays a very low p-value of 4.86e-10, which indicates that CPI is statistically significant. However, the R-squared value of 0.6003% indicates otherwise, and tells us that only about 0.6% of the variation in *log(Weekly_Sales)* is explained by the CPI.

Thus, although CPI is statistically significant, it shows a low R-squared value which indicates that it only explains about 0.6003% of the variation in *log(Weekly_Sales)*. It is practically insignificant in affecting the weekly sales.

This indicates that in reality, people would prefer to shop from Walmart despite a higher CPI and thus revealing either a sign of brand loyalty for Walmart or a competitive price offering by Walmart, both characteristics of which Walmart is known for.

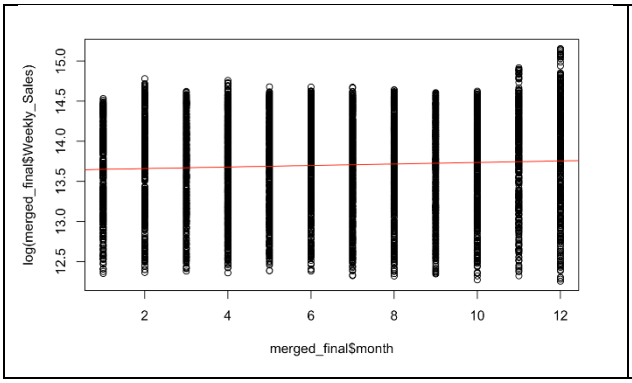


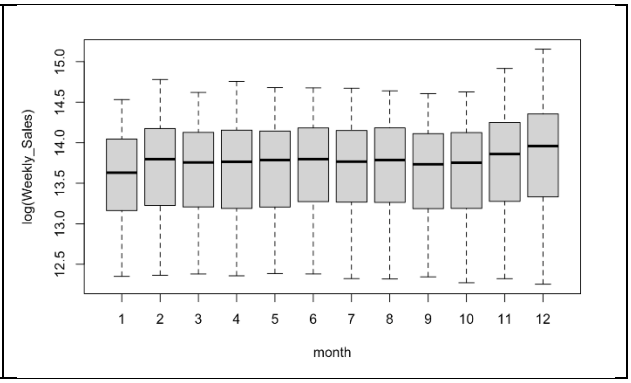

```
Call:
lm(formula = log(Weekly_Sales) ~ CPI, data = merged_final)

Residuals:
     Min       1Q   Median       3Q      Max
-1.49905 -0.46832  0.04435  0.48306  1.47809

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.9006262  0.0327087 424.983  < 2e-16 ***
CPI         -0.0011582  0.0001858  -6.233 4.86e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5866 on 6433 degrees of freedom
Multiple R-squared:  0.006003,  Adjusted R-squared:  0.005849
F-statistic: 38.85 on 1 and 6433 DF,  p-value: 4.861e-10
```

4.2.6 Relation between log(Weekly_Sales) and Month

In this section, we determine whether weekly sales of a store are dependent on the month of the year.

Log(Weekly_Sales) and month are slightly positively correlated with a correlation coefficient of 0.05. This is reasonable, since people usually shop more during the end of the year, but general shopping is still carried out throughout the year. To study this relationship further, we compare a linear regression model between *log(Weekly_Sales)* and Month, along with an ANOVA model between *log(Weekly_Sales)* and the month as a factor or categorical variable.





The linear regression model shows a slightly upward trend, possibly owing to the marginal increase in shopping during the holidays and Christmas season as November and December approaches. In addition, there is also a noticeable increase in the weekly sales in February from January, possibly owing to the marginal increase in shopping during the Super Bowl Holiday which falls every February. The model produces a p-value of 2.07e -05, indicating that month is a statistically significant variable. However, it only explains about 0.2813% of the variation in *log(Weekly_Sales)*, which is relatively lower than most other external variables. As a result, we can conclude that month does not play a significant role in affecting the weekly sales.

```
Call:
lm(formula = log(Weekly_Sales) ~ month, data = merged_final)

Residuals:
     Min       1Q   Median       3Q      Max
-1.50062 -0.47403  0.07082  0.46541  1.40000

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.639779   0.016319  835.81  < 2e-16 ***
month        0.009636   0.002262    4.26 2.07e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5875 on 6433 degrees of freedom
Multiple R-squared:  0.002813,  Adjusted R-squared:  0.002658
F-statistic: 18.15 on 1 and 6433 DF,  p-value: 2.071e-05
```

Now, it is also of interest to check if the mean *log(Weekly_Sales)* is the same for each month. We test the following hypotheses:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_{12} \quad \text{against} \quad H_1: \mu_i \neq \mu_j \text{ for } i,j = 1, 2, \ldots, 12$$

Where μ_j is the mean *log(Weekly_Sales)* for month j.

The summary of the ANOVA test is given below:

```
                Df Sum Sq Mean Sq F value   Pr(>F)
factor(month)   11     21  1.9065   5.552 6.52e-09 ***
Residuals     6423   2206  0.3434
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The p-value of the ANOVA test at 6.52e - 09 is lower than a 1% significance level, and we reject the null hypothesis. Therefore, the mean *log(Weekly_Sales)* is not equal for all months of the year.

In order to understand the difference in mean *log(Weekly_Sales)* for each month, we implement a pairwise t-test as shown below:

```
        Pairwise comparisons using t tests with pooled SD

data:  logweeklysales and factormonth

   1       2       3       4       5       6       7       8       9       10      11
2  0.00393 -       -       -       -       -       -       -       -       -       -
3  0.03496 0.35695 -       -       -       -       -       -       -       -       -
4  0.01834 0.49088 0.79963 -       -       -       -       -       -       -       -
5  0.01266 0.66231 0.63418 0.81365 -       -       -       -       -       -       -
6  0.00082 0.64054 0.15658 0.23448 0.36162 -       -       -       -       -       -
7  0.00914 0.68251 0.58942 0.77082 0.96477 0.36648 -       -       -       -       -
8  0.00386 0.96488 0.37065 0.51061 0.68817 0.60202 0.70987 -       -       -       -
9  0.11917 0.12372 0.52803 0.37004 0.27392 0.04061 0.23712 0.12697 -       -       -
10 0.07401 0.19946 0.71183 0.52870 0.40216 0.07410 0.35975 0.20601 0.79361 -       -
11 3.1e-05 0.09306 0.01153 0.01924 0.03846 0.19704 0.03639 0.08097 0.00209 0.00439 -
12 8.8e-11 3.8e-05 4.0e-07 8.9e-07 5.7e-06 0.00018 3.3e-06 2.3e-05 1.6e-08 6.2e-08 0.03526
```

We observe that the mean *log(Weekly_Sales)* for January (month 1) is significantly different from all months except September (month 9) which shows a higher p-value of 0.119. The mean *log(Weekly_Sales)* is similar for months between February to October (month 2 to month 10), indicating a normal shopping period. Meanwhile, the mean *log(Weekly_Sales)* for November (month 11) and December (month 12) is significantly different from the rest.

Thus, this confirms our guess that customers shop more during the holiday season which is thanksgiving in September and Christmas in November, December, and January.

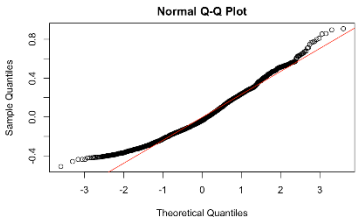
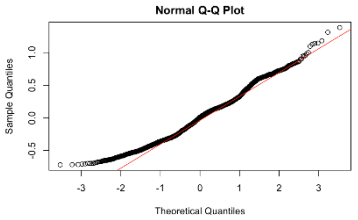
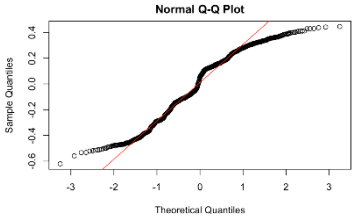
4.2.7 The single most important Type and Size of store that affects log(Weekly_Sales)?

We have seen from section 4.2.1 that the size of the store affects *log(Weekly_Sales)* the most. Each type of store (A, B and C) has a different store size. We now use simple linear regression models to understand which type of store has the most impact on *log(Weekly_Sales).*

$$log(Weekly\_Sales) = \beta_0 + \beta_1 * X + \varepsilon$$

Where X could be the size of Type A store, Type B store or Type C store. The summary of the analysis is listed in the table below.

By comparing the R-squared value and the residual plot, the size of Type A store affects the *log(Weekly_Sales)* the most. Thus, we reinstate the result that larger stores lead to higher *log(Weekly_Sales)*.

Variable (X)	Fitted Model, Y = *log(Weekly_Sales)*	p-value of X	R-squared value	QQ-plot of residuals
Size of Type A Store	$\hat{Y}$ = 12.31 + 0.00000972X	< 2e-16	0.8107	
Size of Type B Store	$\hat{Y}$ = 12.56 + 0.00000941X	< 2e-16	0.4175	
Size of Type C Store	$\hat{Y}$ = 11.93 + 0.00002735X	0.000206	0.01597	

4.3 Multiple Linear Regression

In this section, we build a multiple linear model for *log(Weekly_Sales)* on both internal and external factors. We use a backward elimination method to select the most appropriate variables as shown below.

We observe that the final model uses Size of Store, month, CPI and Type of store.

We conclude that Size, Month, CPI and Type are the significant measures that could be used to model *log(Weekly_Sales)*, whilst Store, Week, Year, Temperature, Fuel_Price, Unemployment, IsHoliday are not.

The fitted model is given as:

$\widehat{Log(Weekly\_Sales)}$ = 12.41 + 0.000009575 * Size + 0.009685 * Month – 0.0008150* CPI + 0.2072*TypeB + 0.3105* TypeC

```
> lmmodel =
lm(log(Weekly_Sales)~Temperature+Size+month+Fuel_Price+CPI+Type, data=
merged_final)
> step(lmmodel, direction = 'backward')
```

```
Start:  AIC=-16401.6
log(Weekly_Sales) ~ Temperature + Size + month + Fuel_Price +
    CPI + Type

              Df Sum of Sq     RSS      AIC
- Fuel_Price   1      0.00  501.81 -16403.5
- Temperature  1      0.07  501.88 -16402.7
<none>                     501.80 -16401.6
- CPI          1      5.91  507.71 -16328.3
- month        1      6.22  508.02 -16324.4
- Type         2     36.04  537.85 -15959.2
- Size         1    875.37 1377.18  -9906.9

Step:  AIC=-16403.54
log(Weekly_Sales) ~ Temperature + Size + month + CPI + Type

              Df Sum of Sq     RSS      AIC
- Temperature  1      0.08  501.89 -16404.5
<none>                     501.81 -16403.5
- CPI          1      6.09  507.90 -16327.9
- month        1      6.30  508.11 -16325.2
- Type         2     36.08  537.89 -15960.8
- Size         1    876.94 1378.75  -9901.6

Step:  AIC=-16404.47
log(Weekly_Sales) ~ Size + month + CPI + Type

        Df Sum of Sq     RSS      AIC
<none>                501.89 -16404.5
- month  1      6.33  508.22 -16325.8
- CPI    1      6.53  508.42 -16323.3
- Type   2     36.40  538.29 -15958.0
- Size   1    884.16 1386.05  -9869.6

Call:
lm(formula = log(Weekly_Sales) ~ Size + month + CPI + Type, data = merged_final)

Coefficients:
(Intercept)         Size        month          CPI        TypeB        TypeC
  1.241e+01    9.575e-06    9.685e-03   -8.150e-04    2.072e-01    3.105e-01
```

4.4 Time Series

In this section, a thorough time series analysis of Walmart's weekly sales data is performed to forecast the Weekly Sales for each types of stores (A, B and C). A time series is a sequence of observations over time, and time series analysis can be useful to see how earnings change over time. Additionally, time series data are often examined in hopes of discovering a historical pattern that can be exploited in the preparation of a forecast.

Since the weekly sales of stores are captured weekly over two years, we can use the two year's weekly sales historical data to help forecast the revenue generated in the next following years.

To start of analyzing the time series of the weekly sales, we answer the three essential questions:

a) Is the data stationary?
b) Is there seasonality in the data?
c) Is the target variable (Weekly Sales) autocorrelated?

4.4.1 Data Preparation

Sales data is recorded for 45 stores of three types (A, B and C), over 143 weeks from 5th February 2010 to 26th October 2012. The mean of the weekly sales data for of the three different store types A, B and C across the 45 stores is aggregated and is divided into three different datasets to ease the process of analysis.

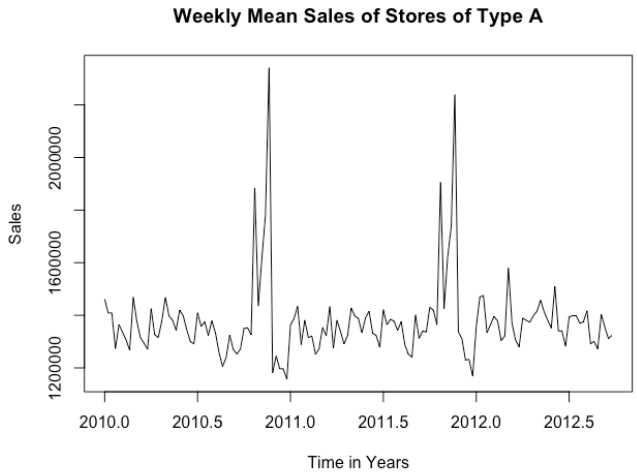
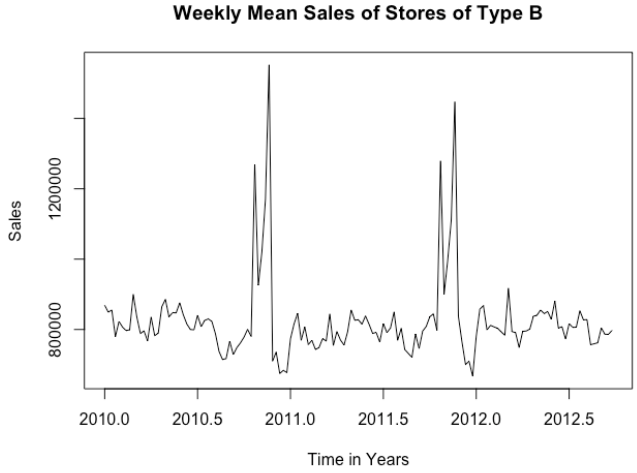
Description of each store type is presented below:

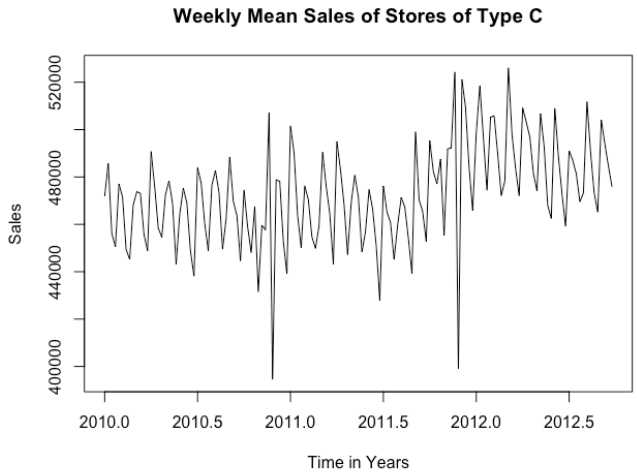
- Store Type A - Walmart Discount Stores
- Store Type B - Walmart Supercenters
- Store Type C - Walmart Neighborhood Markets

4.4.2 Comparison of Time Plots of Store Types A, B and C

A comparison of the time series plots of the sales of the three store types is drawn to analyse how the weekly sales change with time.

It is important to take note that "trend" refers to the upward or downward movement that characterizes a time series over a period of time while "seasonal" refers to a periodic pattern in a time series that complete themselves within a calendar year and are then repeated on a yearly basis.

Time Plot of Store Type A 1. The data seems to be mostly stationary, with only four outliers. 2. There appears to be a rather large spike in the sales near the end of 2010 and 2011, giving an indication of a seasonal trend. 3. The reasons for the spike could be the earthquake in Haiti at the end of 2010, and the storm Sandy in the Caribbean, causing people to panic-buy necessities at the discount stores.	
Time Plot of Store Type B 1. The data seems to be mostly stationary, with only four outliers. 2. There appears to be a rather large spike in the sales near the end of 2010 and 2011, giving an indication of a seasonal trend. 3. The reasons for the spike could be the earthquake in Haiti at the end of 2010, and the storm Sandy in the Caribbean, causing people to panic-buy necessities at the supercenters.	

Time Plot of Store Type C	

1. The data does not appear to be stationary and has about two outliers.
2. The high frequency could be explained by the constant shopping in neighborhood stores by residents and overall growth in trend could be explained by increased population over the years.
3. There appears to be a decline in sales around the end of 2010 and 2011, which could be explained by increased sales in discount stores and supercenters at the same time due to the natural calamities.

4.4.3 Test for Stationarity of Sales Data

In order to build a time series model for forecasting, we need to ensure that the data is stationary, and this can be achieved through taking differences of the data. To check whether a time series is stationary, there are three ways to identify stationarity:

1. Look at plots: Review a time series plot of the weekly sales to visually check if there are any obvious trends or seasonality

2. Summary statistics: Review the summary statistics for the weekly sales time series. There are three basic criteria for a series to be classified as stationary series:

- The mean of the series should not be a function of time rather should be a constant.
- The variance of the series should not a be a function of time.
- The covariance of the i^{th} term and the $(i + m)^{th}$ term should not be a function of time.

3. Statistical tests: Use statistical tests to check if the expectations of stationarity are met or have been violated.

Using statistical test, we check the stationarity of the time series data using the Augmented Dickey-Fuller test. This tests the null hypothesis that a unit root is present in the time series sample, and alternative hypothesis that the unit root is absent, resulting in stationary data. The significance level of the test is taken to be 5%.

Test for Stationarity in Sales for Store Type A:

```
        Augmented Dickey-Fuller Test

data:  aSales
Dickey-Fuller = -5.3166, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

We observe that the p-value = 0.01 is smaller than alpha = 0.05. Hence, we reject the null hypothesis at the 5% significance level. Thus, the data is taken to be stationary.

Test for Stationarity in Sales for Store Type B:

```
        Augmented Dickey-Fuller Test

data:  bSales
Dickey-Fuller = -5.2846, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

We observe that the p-value = 0.01 is smaller than alpha = 0.05. Hence, we reject the null hypothesis at the 5% significance level. Thus, the data is taken to be stationary.

Test for Stationarity in Sales for Store Type C:

```
        Augmented Dickey-Fuller Test

data:  cSales
Dickey-Fuller = -2.1119, Lag order = 5, p-value = 0.5298
alternative hypothesis: stationary
```

We observe that the p-value = 0.5298 is greater than alpha = 0.05. Hence, we do reject the null hypothesis at the 5% significance level, as there is not enough evidence to support the alternative hypothesis. Thus, the data is not stationary.

4.4.3.1. Transforming Non-Stationary Data into Stationary Data

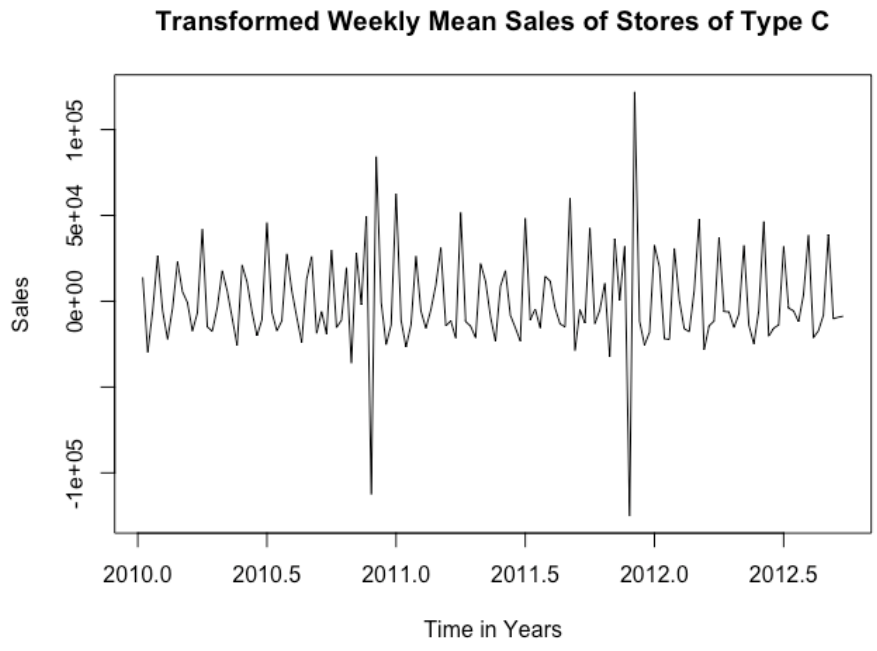
It is observed that the Sales data for store type C is not stationary. Therefore, the data is transformed using the differencing operator to make it stationary so that time series analysis may be performed on it. We now check if the differenced data is stationary

Test for Stationary in Sales for Store Type C after Differencing (d=1):

```
        Augmented Dickey-Fuller Test

data:  c
Dickey-Fuller = -7.7901, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

We observe that the p-value = 0.01 is much lesser than alpha = 0.05. Hence, we reject the null hypothesis at the 5% significance level. Thus, the data is taken to be stationary.

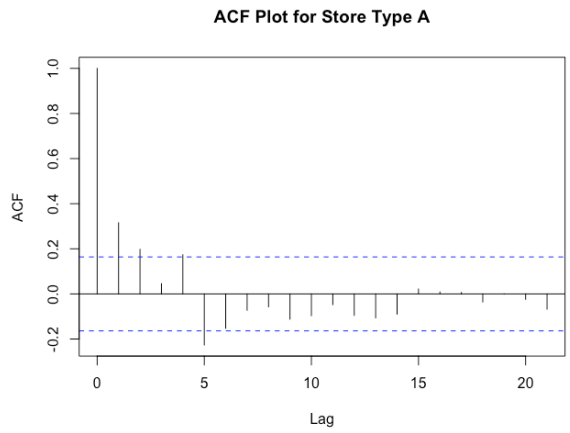
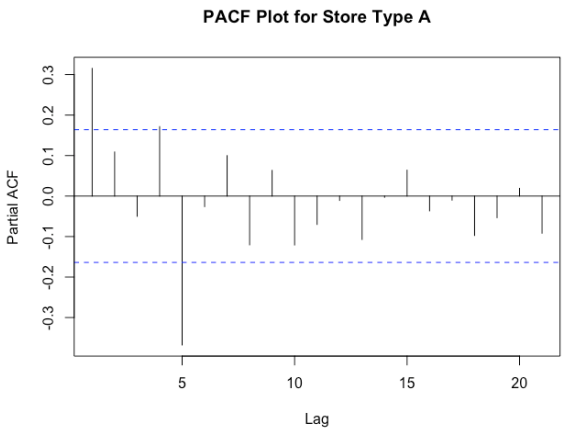
The time series plot of the data can be visualized as follows:

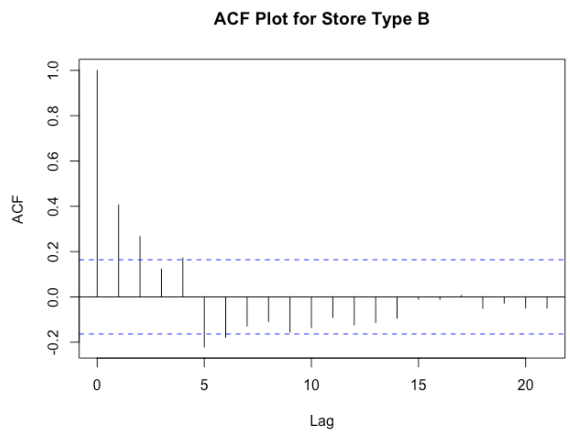


4.4.4. Comparison of ACF and PACF of the Sales Data

The Auto Correlation Function (ACF) is the coefficient of correlation between two values in a time series. The ACF is a way to measure the linear relationship between an observation at time *t* and the observations at previous times. On the other hand, the Partial Auto Correlation (PACF) is a summary of the relationship between an observation in a time series with observations at prior time steps with the relationships of intervening observations removed.

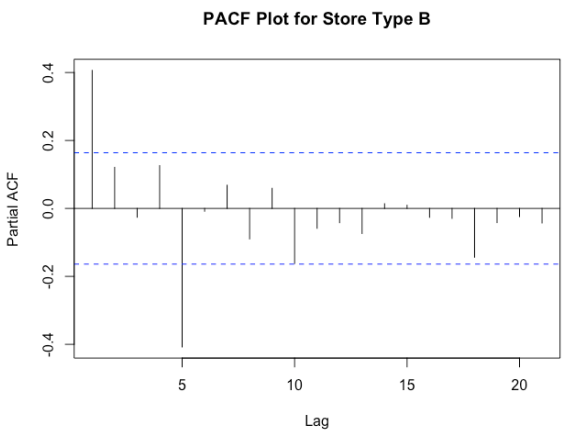
A comparison of the ACFs and PACFs of the three sales data is performed to estimate a suitable model for the three time series data.

ACF Plots	PACF Plots
	
<u>Observation:</u> There are significant correlations at the first, second, third and fifth lag, followed by correlations that are not significant. <u>Indication:</u> The ACF is significant up to lag 2, which suggests and MA(5) model.	<u>Observation:</u> There are significant correlations at the first, third and fifth lag, followed by correlations that are not significant. <u>Indication:</u> The PACF is significant up to lag 2, which suggests and AR(5) model.



Observation:
There are significant correlations at the first, second, third and fifth lag, followed by correlations that are not significant.

Indication:
The ACF is significant up to lag 2, which suggests and MA(5) model.

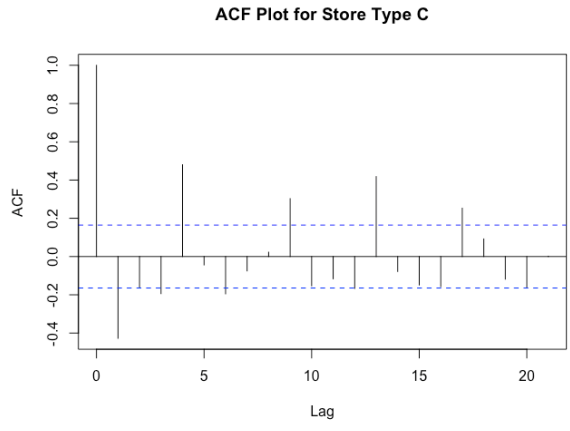


Observation:
There are significant correlations at the first and fifth lag, followed by correlations that are not significant.

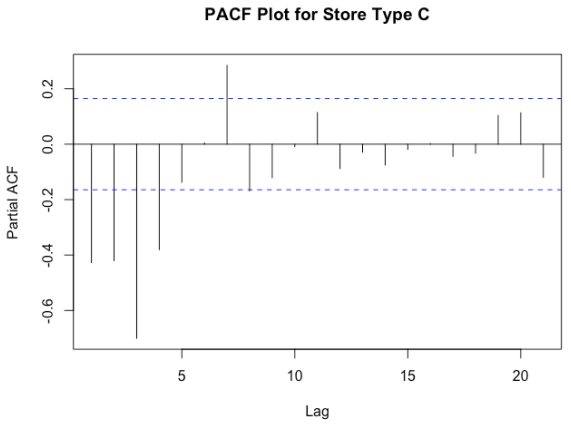
Indication:
The PACF is significant up to lag 2, which suggests and AR(5) model.

The observations and indications are consistent with the previous data analysis of the three models.

4.4.4.1. ACF and PACF of Differenced Store Type C Sales Data



Observation:
The ACF appears to spike at lags 4, 9, 13 and 16 after which it cuts off.



Observation:
The PACF is significant at lags 3 and 7 after which it dies down.

Indication: The ACF is significant up to lag 16, which suggests SMA(4) model with non-seasonal MA(1).	Indication: The PACF is significant up to lag 7 which suggests and SAR(2) model with non-seasonal AR(2).

4.4.5. Fitting Times Series Model to the Sales Data

The best model for the three sales data will be fitted using the arima() and auto.arima() function in R. We fit the best model to the data from among MA(q), AR(p), ARMA(p,q) and ARIMA(p,d,q) processes.

It is important to take note of the following notations:

- *t* refers to the period of time (in weeks)
- X_t refers to the discrete-time random variable at any time t
- Z_t is a white noise process: A sequence of uncorrelated random variables, $\{Z_t\}_{t=1}^{n}$with mean 0 and finite variance σ_Z^2

The Moving Average (MA) Process:

The process $\{X_t\}$ defined by

$$X_t = \sum_{j=0}^{q} \theta_j Z_{t-j}, \theta_0 = 1$$

where $Z_t \sim WN(0, \sigma^2)$, is called a moving average process of order q and is denoted by MA(q). We always assume that $\theta_i = 0$ for $i \leq -1$.

The Seasonal Moving Average (SMA) Process:

The process $\{X_t\}$ defined by

$$y_t = Z_t + \widetilde{\theta_1} Z_{t-s} + \cdots + \widetilde{\theta_Q} Z_{t-Qs}$$

Is called the seasonal moving average model of order Q. s is called the seasonal period.

The Autoregressive (AR) Process:

Assume that $\{Z_t\}$ is white noise. Then the process $\{X_t\}$ defined by

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + Z_t$$

is called an autoregressive process of order p, denoted by AR(p).

The Seasonal Autoregressive (SAR) Process:

The process $\{X_t\}$ defined by

$$y_t = \widetilde{\phi_1} y_{t-s} + \cdots + \widetilde{\phi_P} y_{t-Ps} + Z_t$$

Is called the seasonal autoregressive model of order P. s is called the seasonal period.

The ARMA(p,q) Process:

The process $\{X_t\}$ is said to be an autoregressive moving average or ARMA(p,q) process if $\{X_t\}$ is stationary and for every t,

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

where $Z_t \sim WN(0, \sigma^2)$.

Akaike's Information Criteria (AIC):

The AIC is an estimator of prediction error and thereby relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

Of the potential models, the model with the least AIC value is chosen at the model of best fit for the sales data.

Ljung-Box Test:

The Box–Pierce (and Ljung–Box) test examines the Null of independently distributed residuals. It's derived from the idea that the residuals of a "correctly specified" model are independently distributed. If the residuals are not, then they come from a miss–specified model.

The Ljung-Box test is performed on the best fitted model to ensure that it is adequate.

H_0: The data are independently distributed.

H_a: The data are not independently distributed; they exhibit serial correlation.

Fitted Model for Store Type A:

1. Fitting MA(5) model

```
arima(x = aSales, order = c(0, 0, 5))

Coefficients:
         ma1     ma2      ma3     ma4      ma5   intercept
      0.3895  0.2724  -0.0032  0.4063  -0.1703  1377528.25
s.e.  0.0884  0.0828   0.0991  0.0949   0.1057    20542.67

sigma^2 estimated as 1.701e+10:  log likelihood = -1888.23,  aic = 3790.46
```

2. Fitting AR(5) model

```
Call:
arima(x = aSales, order = c(5, 0, 0))

Coefficients:
         ar1     ar2      ar3     ar4      ar5   intercept
      0.3557  0.0651  -0.0600  0.2742  -0.3597  1377228.21
s.e.  0.0772  0.0789   0.0789  0.0789   0.0765    15444.87

sigma^2 estimated as 1.777e+10:  log likelihood = -1890.84,  aic = 3795.69
```

3. Fitting a model using the auto.arima() function in R:

```
Series: aSales[1:length(aSales)]
ARIMA(2,0,2) with non-zero mean

Coefficients:
          ar1      ar2     ma1     ma2        mean
      -1.0003  -0.4002  1.4093  0.9078  1377329.89
s.e.   0.1343   0.1301  0.0919  0.0933    15540.53

sigma^2 estimated as 1.879e+10:  log likelihood=-1892.81
AIC=3797.62   AICc=3798.24   BIC=3815.4
```

Observation:

The AIC value of MA(5) at 3790.46 is the least. Hence, MA(5) is chosen at the model of best fit for the sales data of store type A.

Ljung-Box Test:

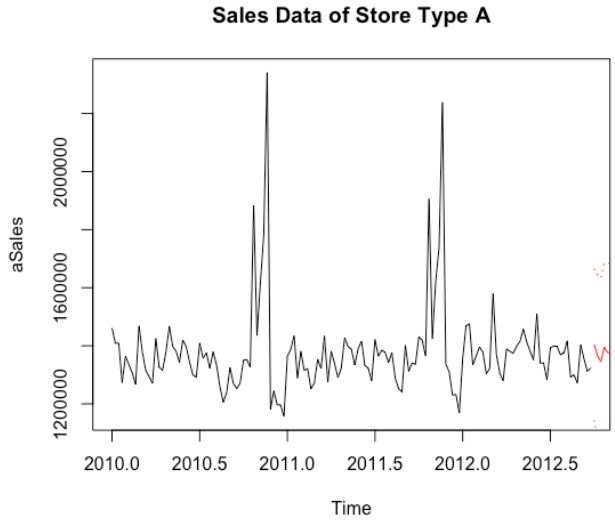
```
        Box-Pierce test

data:  afit1$residuals
X-squared = 0.040907, df = 1, p-value = 0.8397
```

Since the p-value = 0.8397 is much greater than the 5% significance level, we do not reject the null hypothesis, concluding that our data is independently distributed.

Predicting future 8 week sales values for Store Type A:

```
$pred
Time Series:
Start = c(2012, 40)
End = c(2012, 47)
Frequency = 52
[1] 1404849 1364984 1345227 1394854 1376081 1377528 1377528 1377528
```

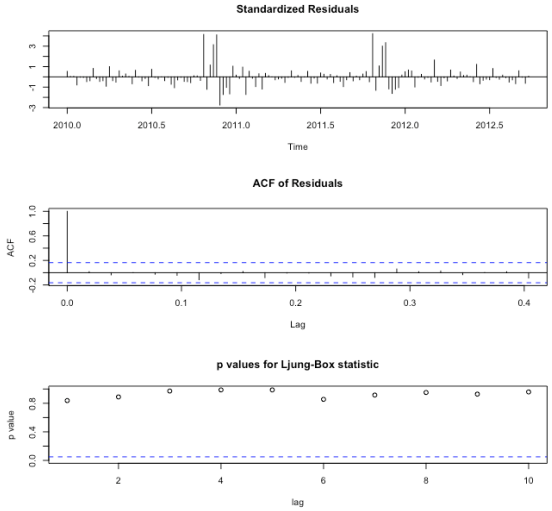
The plot of the predicted weekly sales for the next 8 weeks with an approximated 95 % confidence interval is displayed below.



Proposed Model:

$$\widehat{X_t} = Z_t + 0.3895Z_{t-1} + 0.2724Z_{t-2} - 0.0032Z_{t-3} + 0.4063Z_{t-4} - 0.1703Z_{t-5}$$

Diagnostic Checking of the Fitted MA(5) Model:



Observation:

As the ACF plot of the residuals shows that the all the ACF values are within the confidence intervals, and the p-values are sufficiently high, the MA(5) model is adequate for the data.

Fitted Model for Store Type B:

1. Fitting MA(5) model

```
Call:
arima(x = bSales, order = c(0, 0, 5))

Coefficients:
         ma1     ma2     ma3     ma4      ma5  intercept
      0.4467  0.3246  0.1522  0.4470  -0.0973  823339.16
s.e.  0.0873  0.0831  0.0958  0.0947   0.0950   18093.18

sigma^2 estimated as 9.217e+09:  log likelihood = -1844.2,  aic = 3702.41
```

2. Fitting AR(5) model

```
Call:
arima(x = bSales, order = c(5, 0, 0))

Coefficients:
         ar1     ar2      ar3     ar4      ar5  intercept
      0.4123  0.0842  -0.0245  0.2675  -0.3964  823397.63
s.e.  0.0757  0.0796   0.0798  0.0794   0.0748   12384.33

sigma^2 estimated as 9.382e+09:  log likelihood = -1845.26,  aic = 3704.52
```

3. Fitting a model using the auto.arima() function in R:

```
Series: bSales[1:length(bSales)]
ARIMA(2,0,2) with non-zero mean

Coefficients:
          ar1      ar2     ma1     ma2      mean
      -0.8452  -0.2019  1.3284  0.7602  823094.4
s.e.   0.1262   0.1486  0.0856  0.1096   12788.7

sigma^2 estimated as 1.069e+10:  log likelihood=-1852.02
AIC=3716.05   AICc=3716.66   BIC=3733.82
```

Observation:

The AIC value of MA(5) at 3702.41 is the least. Hence, MA(5) is chosen at the model of best fit for the sales data of store type B.

Ljung-Box Test:

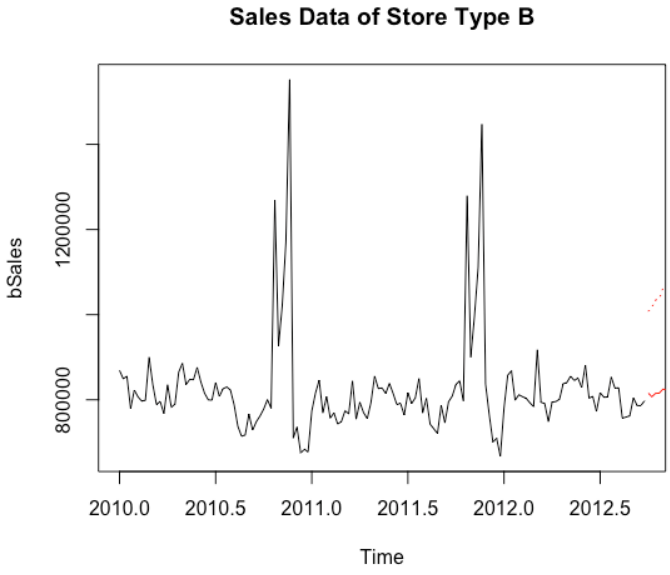
```
        Box-Pierce test

data:  bfit1$residuals
X-squared = 0.012338, df = 1, p-value = 0.9116
```

Since the p-value = 0.9116 is much greater than the 5% significance level, we do not reject the null hypothesis, concluding that our data may be independently distributed.

Predicting future 8 week sales values for Store Type B:

```
$pred
Time Series:
Start = c(2012, 40)
End = c(2012, 47)
Frequency = 52
[1] 815611.0 807365.3 815013.6 815734.1 825302.6 823339.2 823339.2 823339.2
```

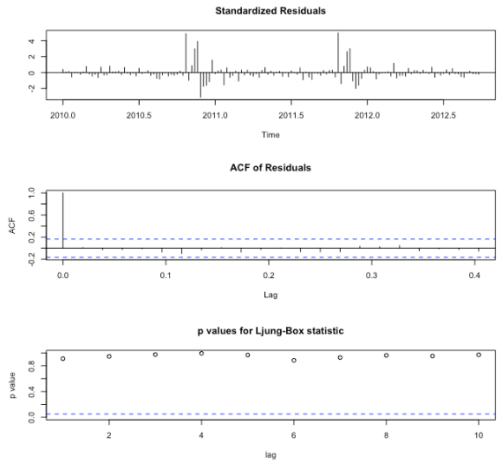
The plot of the predicted weekly sales for the next 8 weeks with an approximated 95 % confidence interval is displayed below.



Proposed Model:

$$\widehat{X_t} = Z_t + 0.4467Z_{t-1} + 0.3246Z_{t-2} + 0.1522Z_{t-3} + 0.4470Z_{t-4} - 0.0973Z_{t-5}$$

Diagnostic Checking of the Fitted MA(5) Model:



Observation:

As the ACF plot of the residuals shows that the all the ACF values are within the confidence intervals, and the p-values are sufficiently high, the MA(5) model is adequate for the data.

Fitted Model for Store Type C

1. Fitting SMA(4) model with MA(1):

```
Call:
arima(x = cSales, order = c(0, 0, 1), seasonal = list(order = c(0, 1, 4), period = 2))

Coefficients:
          ma1     sma1    sma2     sma3    sma4
      -0.2812  -0.9159  0.6406  -0.5496  0.1664
s.e.   0.1006   0.1085  0.1043   0.0892  0.0945

sigma^2 estimated as 270902090:  log likelihood = -1570.33,  aic = 3152.66
```

2. Fitting SAR(2) model with AR(2):

```
Call:
arima(x = cSales, order = c(2, 0, 0), seasonal = list(order = c(2, 1, 0), period = 2))

Coefficients:
          ar1      ar2     sar1    sar2
      -0.3232  -0.3369  -0.6748  0.0994
s.e.   0.0855   0.1933   0.2200  0.1701

sigma^2 estimated as 261885364:  log likelihood = -1568.01,  aic = 3146.02
```

3. Fitting SARIMA(2,0,1,2,1,4,2):

```
Call:
arima(x = cSales, order = c(2, 0, 1), seasonal = list(order = c(2, 1, 4), period = 2))

Coefficients:
          ar1      ar2     ma1    sar1    sar2     sma1    sma2     sma3    sma4
      -0.5510  -0.3583  0.2743  0.0238  0.5098  -0.7584  0.1228  -0.3748  0.1261
s.e.   0.5512   0.1567  0.5706  0.2356  0.1430   0.3143  0.2645   0.1491  0.1488

sigma^2 estimated as 241581449:  log likelihood = -1562.82,  aic = 3145.64
```

Observation:

The AIC value of **SARIMA(2,0,1,2,1,4,2)** with period = 2, at 3145.64 is the least. Hence, this is chosen at the model of best fit for the sales data of store type C.

Proposed Model:

$$(\mathbf{1 - 0.5510B - 0.3583B^2})(\mathbf{1 + 0.0238B + 0.5098B^2})\boldsymbol{X_t}$$
$$= (\mathbf{1 + 0.2743B})(\mathbf{1 - 0.7584B + 0.1228B^2 - 0.3748B^3 + 0.1261B^4})\boldsymbol{Z_t}$$

Ljung-Box Test:

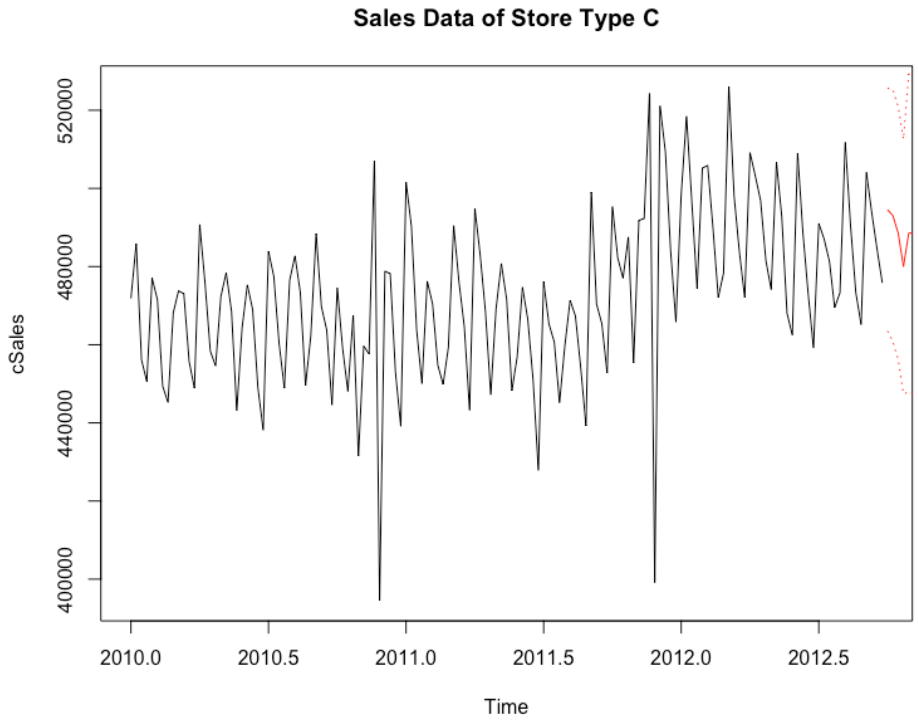
```
        Box-Pierce test

data:  cfit3$residuals
X-squared = 0.015261, df = 1, p-value = 0.9017
```

Since the p-value = 0.9017 is much greater than the 5% significance level, we do not reject the null hypothesis, concluding that our data may be independently distributed.

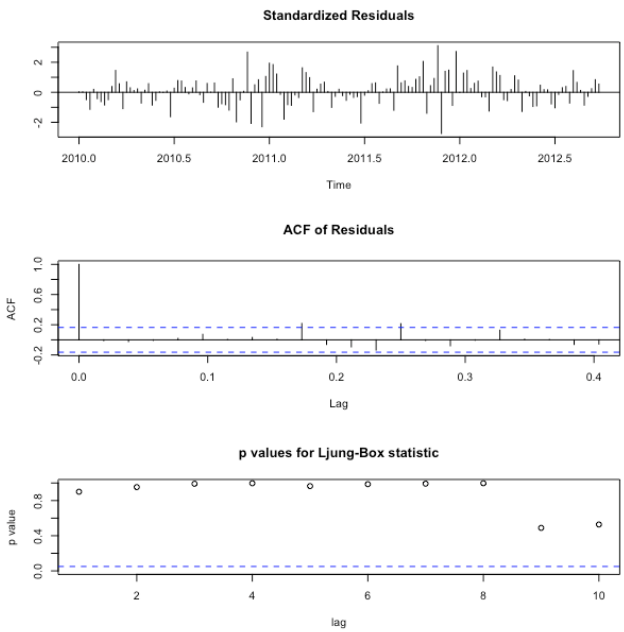
Predicting future 8 week sales values for Store Type C:

```
$pred
Time Series:
Start = c(2012, 40)
End = c(2012, 47)
Frequency = 52
[1] 494574.0 493022.8 488603.1 479989.0 488566.8 488525.0 488175.2 482116.6
```

The plot of the predicted weekly sales for the next 8 weeks with an approximated 95 % confidence interval is displayed below.



Diagnostic Checking of the proposed SARIMA(2,0,1,2,1,4,2) Model:



Observation:

As the ACF plot of the residuals shows that the all the ACF values are within the confidence intervals, and the p-values are sufficiently high, the SARIMA(2,0,1,2,1,4,2) model is adequate for the data.

5. Conclusion and Discussion

Walmart is one of the largest supermarket chains in the United States, which earns hundreds of billions annually from the sales of goods such as food, lifestyle goods and general merchandise. To maximize the profits, the firm has to strategically set up branches based on external factors such as temperature, unemployment rate of the area and fuel prices. In this report, we attempt to answer some of the questions regarding profit based on 2010-2012 data.

We conclude that:

- The weekly sales of a store are dependent on the size of the store
- The weekly sales of a store are dependent on the type of store (A,B,C)
- The temperature of the region does not affect the weekly sales of a store
- The fuel price of the region does not affect the weekly sales of a store
- The weekly sales of a store are not dependent on the Consumer Price Index (CPI)
- Among the variables, the Type of store and the Size of the store are the most important factors affecting the weekly sales. In particular, the sizes of Type A store affect weekly sales the most.

Additionally, we see that the size of the store, month, CPI and type of store can be used to model the weekly profit via multiple linear regression. The measures on store number, week, year, temperature, fuel price, unemployment rate and holiday do not seem to have strong correlation with weekly profit.

Although the findings of this report are interesting, it is important to note that this report is based only on data from 2010 to 2012. This suggests that the results we derived in our report may be irrelevant for Walmart to make business decisions to remain relevant and competitive in 2021, especially since the COVID-19 pandemic has caused a significant shift in consumer preferences to shop online instead. This also means that the datasets are becoming more complex compared to 2010-2012 data due to the need for data analysis on both physical shopping and online shopping. It has been reported that Walmart sales have increased by 24% since the first U.S. lockdowns began in early March 2020 (Wahba, 2020). This trend is critical to understanding changing consumer behavior and more appropriate and insightful actions can be derived to support the sales in the pandemic.

All in all, deeper and wider analysis of the current Walmart Sales data, with advance analytical techniques would be needed to make a stronger statement about the relationships around weekly sales.

6. Appendix

6.1 Data Cleaning & Preparation

```
#Importing datasets
train<-read.csv("C:/Users/User/Desktop/MH3511/train.csv",header=T)
str(train)
head(train)

features<-read.csv("C:/Users/User/Desktop/MH3511/features.csv",header=T)
str(features)
head(features)

stores<-read.csv("C:/Users/User/Desktop/MH3511/stores.csv",header=T)
str(stores)
head(stores)

#Convert True/False values to Binary Numbers in features and train
features$IsHoliday[features$IsHoliday==T]<-1
features$IsHoliday[features$IsHoliday==F]<-0

train$IsHoliday[train$IsHoliday==T]<-1
train$IsHoliday[train$IsHoliday==F]<-0

#Remove Markdown columns in features
features<-features[,c(-5:-9)]
head(features)

#Omit NA values in features
features<-na.omit(features)

#Combine weekly sales of all depts for each store per week in train
Weekly_Sales<-
aggregate(train$Weekly_Sales,by=list(train$Store,train$Date),FUN=sum)
names(Weekly_Sales)=c("Store","Date","Weekly_Sales")


#Merge train and features dataset together
library("dplyr")
merged<-inner_join(features,Weekly_Sales,by=c("Date","Store"))
#Comment: 6435 Obversations and 8 Variables

#Add a week column for train and features
merged <- merged %>% group_by(Store) %>% mutate(Week = row_number(Store))
#Comment: 6435 Observations and 9 Variables

#Change chr type into date type in merged
merged$Date<-as.Date(merged$Date)
```

```
#Separate Date
library("lubridate")
merged$year<-year(ymd(merged$Date))
merged$month<-month(ymd(merged$Date))
merged$day<-day(ymd(merged$Date))
#Comment: 6435 Observations and 12 Variables

#Merge merged and stores together
merged_final<-inner_join(merged,stores,by=c("Store"))
#Comment: 6435 Obversations and 14 Variables
```

6.2 Summary Statistics

```
#Weekly_Sales
hist(merged$Weekly_Sales, xlab="Weekly_Sales",main="Histogram of
Weekly_Sales")
hist(log(merged$Weekly_Sales),xlab="log(Weekly_Sales)",main="Histogram of
log(Weekly_Sales)")
boxplot(log(merged$Weekly_Sales),main="Boxplot of log(Weekly_Sales)")

#Size of store
hist(stores$Size,xlab="Size",main="Histogram of Size")
boxplot(stores$Size,main="Boxplot of Size")

#Temperature
hist(merged$Temperature,xlab="Temperature",main="Histogram of Temperature")
boxplot(merged$Temperature,main="Boxplot of Temperature")

#Fuel_Price
hist(merged$Fuel_Price,xlab="Fuel Price",main="Histogram of Fuel Price")
boxplot(merged$Fuel_Price,main="Boxplot of Fuel Price")

#CPI
hist(merged$CPI,xlab="CPI",main="Histogram of CPI")
boxplot(merged$CPI,main="Boxplot of CPI")

#Unemployment
hist(merged$Unemployment,xlab="Unemployment",main="Histogram of
Unemployment")
boxplot(merged$Unemployment,main="Boxplot of Unemployment")

#IsHoliday
hist(merged$IsHoliday,xlab="IsHoliday",main="Histogram of IsHoliday")

#Type of store
table(stores$Type)
```

```
barplot(table(stores$Type),xlab="Type of Store",ylab="Frequency",main="Barplot of Type of Stores")
boxplot(number$Freq~number$Var1)
```

6.3 Statistical Analysis

```
#Correlation chart
aadt<-
data.frame("log(Weekly_Sales)"=log(merged$Weekly_Sales),Store=merged$Store,Temperature=merged$Temperature,Fuel_Price=merged$Fuel_Price,CPI=merged$CPI,Unemployment=merged$Unemployment,IsHoliday=merged$IsHoliday)
pairs.panels(aadt,
             method="pearson",
             hist.col="steelblue",
             pch=19,
             density=T,
             ellipses=F,stars=T,scale=F)

#Pairwise Regression Models

lmsize = lm(log(Weekly_Sales) ~ Size, data = merged_final)
summary(lmsize)
plot(merged_final$Size, log(merged_final$Weekly_Sales))
abline(lmsize, col = 'red')

res = residuals(lmsize)
qqnorm(res)
qqline(res,col = 'red')

lmtemp = lm(log(Weekly_Sales) ~ Temperature, data = merged_final)
summary(lmtemp)
plot(merged_final$Temperature, log(merged_final$Weekly_Sales))
abline(lmtemp, col = 'red')

lmfuel = lm(log(Weekly_Sales) ~ Fuel_Price, data = merged_final)
summary(lmfuel)
plot(merged_final$Fuel_Price, log(merged_final$Weekly_Sales))
abline(lmfuel, col = 'red')

lmcpi = lm(log(Weekly_Sales) ~ CPI, data = merged_final)
summary(lmcpi)
plot(merged_final$CPI, log(merged_final$Weekly_Sales))
abline(lmcpi, col = 'red')

lmmonth = lm(log(Weekly_Sales) ~ month, data = merged_final)
summary(lmmonth)
plot(merged_final$month, log(merged_final$Weekly_Sales))
abline(lmmonth, col = 'red')
```

```
boxplot(log(Weekly_Sales)~month, data = merged_final)

#Check if the mean weekly sales is the same across all months of the year

#H0: mu1 = mu2 = ... = mu12 for all months of the year, where mui = mean
log(Weekly_Sales) for month i
#H1: not all mui are the same
aov(log(Weekly_Sales)~factor(month), data = merged_final)
summary(aov(log(Weekly_Sales)~factor(month), data = merged_final))

#Since p value is very small, we reject H0.
#Thus the mean log weekly sales in not the same for all months

logweeklysales = log(merged_final$Weekly_Sales)
factormonth = factor(merged_final$month)

pairwise.t.test(logweeklysales, factormonth, p.adjust.method = "none")
#Check if mean log(Weekly_Sales) is same across types of stores

aovtype = aov(log(Weekly_Sales)~Type, data = merged_final)
summary(aovtype)

boxplot(log(Weekly_Sales)~Type, data = merged_final)

#Since p value is extremely low, we reject H0.
#muA, muB and muC are not equal.

TypeA = subset(merged_final, Type == 'A')
A = log(TypeA$Weekly_Sales)

TypeB = subset(merged_final, Type == 'B')
B = log(TypeB$Weekly_Sales)

TypeC = subset(merged_final, Type == 'C')
C = log(TypeC$Weekly_Sales)

factortype = factor(merged_final$Type)

pairwise.t.test(logweeklysales, factortype, p.adjust.method = "none")
#they are all significantly different from each other, as is observed visually from the
boxplot

#Which type and size affects log(Weekly_Sales) the most?

sizeA = TypeA$Size
```

```
lmA = lm(log(Weekly_Sales)~sizeA, data = TypeA)
summary(lmA)


sizeB = TypeB$Size
lmB = lm(log(Weekly_Sales)~sizeB, data = TypeB)
summary(lmB)
resb = residuals(lmB)
#qqnorm(resb)
#qqline(resb, col = 'red')

sizeC = TypeC$Size
lmC = lm(log(Weekly_Sales)~sizeC, data = TypeC)
summary(lmC)
resc = residuals(lmC)
qqnorm(resc)
qqline(resc, col='red')
```

6.3.1 Multiple Linear Regression Model

```
lmmodel = lm(log(Weekly_Sales)~Temperature+Size+month+Fuel_Price+CPI+Type,
data= merged_final)
step(lmmodel, direction = 'backward')
```

6.3.2 Time Series Model

```
 # Time Series Analysis for Walmart Weekly Sales Data

# Splitting the data into three for the Store Types A, B, C

# Extracting the Week, Type of Store and Weekly Sales from the merged_final
dataset
sales <- merged_final[, c(9,13,8)]
str(sales)

combined = aggregate(sales$Weekly_Sales, by = list(sales$Type, sales$Week), FUN
= mean)
names(combined) = c("Type", "Weeks", "Sales")
# extracting the above data for Store Type A
A <- combined[combined$Type == "A",]
head(A)

# extracting the above data for Store Type B
B <- combined[combined$Type == "B",]
head(B)

# Extracting the above data for Store Type = C
C <- combined[combined$Type == "C",]
head(C)
```

```
#-------------------------------------------------------------
# Time Series Analysis for Walmart Weekly Sales Data

# Splitting the data into three for the Store Types A, B, C

# Extracting the Week, Type of Store and Weekly Sales from the merged_final
dataset
sales <- merged_final[, c(9,13,8)]
str(sales)

combined = aggregate(sales$Weekly_Sales, by = list(sales$Type, sales$Week), FUN
= mean)
names(combined) = c("Type", "Weeks", "Sales")
# extracting the above data for Store Type A
A <- combined[combined$Type == "A",]
head(A)

# extracting the above data for Store Type B
B <- combined[combined$Type == "B",]
head(B)

# Extracting the above data for Store Type = C
C <- combined[combined$Type == "C",]
head(C)

#-------------------------------------------------------------
# Time Series Analysis of the weekly sales for type A stores

# Creating the time series data
aSales <- ts(A$Sales, start=2010, frequency=52)

# Visualising the time series plot
ts.plot(aSales, xlab="Time in Years", ylab="Sales", main="Weekly Mean Sales of
Stores of Type A")
plot(1:143, aSales)
lines(1:143, aSales, type ="l")

# Stationarity
adf.test(aSales)

# ACF plot
acf(aSales[1:length(aSales)], main="ACF Plot for Store Type A")

# PACF plot
pacf(aSales[1:length(aSales)], main="PACF Plot for Store Type A")
```

```
# Fitting appropriate time series model to the data
afit1 <- arima(aSales, order =c(0,0,5))
afit2 <- arima(aSales, order =c(5,0,0))
afit3 <- auto.arima(aSales[1:length(aSales)])

# Ljung-Box Test
Box.test(afit1$residuals)

# Predicting future sales values for Store Type A
par(mfrow=c(1,1))
apred<-predict(afit1,n.ahead=8)

# Confidence Interval
ts.plot(aSales, main= "Sales Data of Store Type A")
lines(apred$pred,col="red")
lines(apred$pred+2*apred$se,col="red",lty=3)
lines(apred$pred-2*apred$se,col="red",lty=3)

# Diagnostic Checking
tsdiag(afit1)

#-----------------------------------------------------------
# Time Series Analysis of the weekly sales for type B stores

# Creating the time series data
bSales <- ts(B$Sales, start=2010, frequency=52)

# Visualising the time series plot
ts.plot(bSales, xlab="Time in Years", ylab="Sales", main="Weekly Mean Sales of
Stores of Type B")
plot(1:143, bSales)
lines(1:143, bSales, type ="l")

# Stationarity
adf.test(bSales)

# ACF plot
acf(bSales[1:length(bSales)], main="ACF Plot for Store Type B")

# PACF plot
pacf(bSales[1:length(bSales)], main="PACF Plot for Store Type B")

# Fitting appropriate time series model to the data
bfit1 <- arima(bSales, order =c(0,0,5))
bfit2 <- arima(bSales, order =c(5,0,0))
bfit3 <- auto.arima(bSales[1:length(bSales)])
```

```
# Ljung-Box Test
Box.test(bfit1$residuals)

# Predicting future sales values for Store Type B
par(mfrow=c(1,1))
bpred<-predict(bfit1,n.ahead=8)

# Confidence Interval
ts.plot(bSales, main= "Sales Data of Store Type B")
lines(bpred$pred,col="red")
lines(bpred$pred+2*bpred$se,col="red",lty=3)
lines(bpred$pred-2*bpred$se,col="red",lty=3)

# Diagnostic Checking
tsdiag(bfit1)


#-----------------------------------------------------------
# Time Series Analysis of the weekly sales for type C stores

# Creating the time series data
cSales <- ts(C$Sales, start=2010, frequency=52)

# Visualising the time series plot
ts.plot(cSales, xlab="Time in Years", ylab="Sales", main="Weekly Mean Sales of
Stores of Type C")
plot(1:143, cSales)
lines(1:143, cSales, type ="l")

# Stationarity
adf.test(cSales)

# Transforming the Sales data for Store Type C due to Non-Stationarity
c <- diff(cSales, lag=1)
adf.test(c)
ts.plot(c, xlab="Time in Years", ylab="Sales", main="Transformed Weekly Mean Sales
of Stores of Type C")
plot(1:139, c)
lines(1:139, c, type ="l")

# ACF plot
acf(cSales[1:length(cSales)], main="ACF Plot for Store Type C")
acf(c[1:length(c)], main="ACF Plot for Store Type C")

# PACF plot
```

```
pacf(cSales[1:length(cSales)], main="PACF Plot for Store Type C")
pacf(c[1:length(c)], main="PACF Plot for Store Type C")


# Fitting appropriate time series model to the data

cfit1 <- arima(cSales, order =c(0, 0, 1),
               seasonal =list(order = c(0,1,4), period=2))

cfit2 <- arima(cSales, order =c(2, 0, 0),
                     seasonal =list(order = c(2,1,0), period=2))

cfit3 <- arima(cSales, order =c(2, 0, 1),
               seasonal =list(order = c(2,1,4), period=2))

# Ljung-Box Test
Box.test(cfit3$residuals)

# Predicting future sales values for Store Type B
cpred<-predict(cfit3,n.ahead=8)
cpred

# Confidence Interval
ts.plot(cSales, main= "Sales Data of Store Type C")
lines(cpred$pred,col="red")
lines(cpred$pred+2*cpred$se,col="red",lty=3)
lines(cpred$pred-2*cpred$se,col="red",lty=3)

# Diagnostic Checking
tsdiag(cfit3)
```

7. References

Aima, A. (2019, March 19). *WALMART Sales Data Analysis & Sales Prediction using Multiple Linear Regression in R programming Language*. Retrieved from https://medium.datadriveninvestor.com/walmart-sales-data-analysis-sales-prediction-using-multiple-linear-regression-in-r-programming-adb14afd56fb

Pillai, S. (2018, July 14). *Walmart Store Sales Forecast*. Retrieved from RPubs by RStudio: https://rpubs.com/spillai/walmart_store_sales_forecast

Sreejith Nair, T. J. (2020, Februrary 20). *Project Report - Walmart Sales Prediction*. Retrieved from RPubs by RStudio: https://rstudio-pubs-static.s3.amazonaws.com/580728_b2bb958fa0d341c587001c3935702b46.html

Wahba, P. (2020, May 20). *Fortune*. Retrieved from Walmart's online sales surge during the pandemic, bolstering its place as a strong No.2 to Amazon: https://fortune.com/2020/05/19/walmart-online-sales-amazon-ecommerce/