Fall 10-1-2018

Using social media data in demand forecasting: the case of Walmart

Aybike Akici
St. Mary's University, reflib@stmarytx.edu

USING SOCIAL MEDIA DATA IN DEMAND FORECASTING:

THE CASE OF WALMART

APPROVED:

Rafael Moras, Ph.D., P.E., Supervising Professor

Paul Uhlig, Ph.D., Committee Member

Benjamin Jurewicz, Ph.D., Committee Member

APPROVED:

Winston Erevelles, Ph.D.
Dean of the Graduate School

Date:

USING SOCIAL MEDIA DATA IN DEMAND FORECASTING:

THE CASE OF WALMART

A

THESIS

Presented to the faculty of the School of Science, Engineering and Technology

of

St. Mary's University

in partial fulfillment

of the requirements

for the degree of

MASTER OF SCIENCE

in

Industrial Engineering

by

Aybike Akıcı, M.S.

San Antonio, Texas

October 2018

To my love, my friend, my husband Fatih

ABSTRACT

USING SOCIAL MEDIA DATA IN DEMAND FORECASTING: THE CASE OF WALMART

Aybike Akıcı

St. Mary's University, 2018

Supervising Professor: Rafael Moras, Ph.D., P.E.

We describe a fully empirical study on demand forecasting, that is applicable to any real-world data. This is a hands-on case study on the power of social media in demand forecasting. We implement a Box-Jenkins methodology with exogenous variables, namely ARIMAX, to forecast Walmart's future sales. The social media components that we utilize are the number of likes and comments on the official Facebook page of Walmart. The details of the empirical investigation for fitting the best ARIMAX model are presented, and the results are discussed. With this thesis, we demonstrate that social media information should be considered in forecasting, as it is very valuable for any company when performing demand planning, and inventory management.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

It has been a very long journey, and I would like to thank many people who helped me finish this study and will always be remarkable for me.

Firstly, I would like to present my sincere thanks to my advisor, Dr. Rafael Moras for his encouragement, support, and understanding. I will always appreciate his valuable contributions and guidance to this thesis. I am also exceedingly grateful to Professors Paul Uhlig and Benjamin Jurewicz for serving in my thesis committee and sharing their valuable insights. I am also very thankful to Dr. David Reilly for his help and support.

I cannot end without thanking my family, on whose constant encouragement and love I have relied. In this context, I would like to express my sincere thanks to my parents, Sultan and Meriç Özdemirel, and my siblings, Ayça and Tuğşad.

Last, but certainly not the least, I am most grateful to my dear husband Fatih Akıcı for his infinite support, guidance, and love. This study is dedicated to him. I am also very thankful to my beautiful children; my two-and-a-half-year-old daughter, Gül Şirin, and my son, Korkut Alp, who joined our happy little family six months ago.

CHAPTER 1

INTRODUCTION

Demand forecasting is an important industrial engineering/operations management topic because forecasting demand accurately and effectively tends to lead to higher customer satisfaction while simultaneously optimizing warehousing and inventories by keeping the right amount of the right product on shelves. In today’s highly interconnected world, it is a clearly visible fact that the relationship between retailers and consumers has been transformed by social media. Brands can use social media data that enable them to optimize their buying and selling processes in ways that both the brands and customers could benefit from. By 2011, more than 83 percent of the Inc. 500 companies used at least one of the social media platforms (Hameed, 2011). Besides, consumers affect the purchasing decisions of their networks by expressing their preferences in social media sites such as Facebook and Twitter. According to a study, 74 percent of consumers reported that they use social media to make decisions on purchasing (Barbera, 2016). This finding typifies the impact of social media on retailing decisions.

Traditional forecasting methods generally utilize solely historical sales data to predict future demand. A problem with this plain approach is that it does not appreciate the importance of “shocks” arising from factors such as sudden changes in customer satisfaction and sentiments, marketing campaigns, celebrity endorsements, scandals, and the like, on future sales. These factors can significantly shift the future sales trajectory, which the traditional model would be unable to capture, because simply it does not include such elements.

In order to include the effects of the aforementioned external shocks, social media data may be taken into consideration, since this source of information can represent a composition of these sudden changes. People communicate their ideas and feelings about various brands by tweeting in Twitter, commenting and liking on Facebook, following and liking on Instagram and so on. Hence, one needs to take into account the social media component besides the historical sales data to be able to perform forecasting more accurately.

Contrary to the widespread Autoregressive Integrated Moving Average (ARIMA) models, where only the past records of the dependent variable and residual terms are used, we implemented the Box-Jenkins methodology with the usual auto regressive and moving average components, as well as exogenous variables, namely ARIMAX to forecast Walmart's sales. Exogenous explanatory variables are independent variables other than the past records of the dependent variable and residuals, which, in our case, were social media characteristics. The social media components were the number of likes and comments on the official Facebook page of Walmart. The idea that we explored was, in a time-series fashion, whether Walmart sales increase or decrease in conjunction with the number of Facebook likes and comments. This way, Facebook activity became the X in the ARIMAX, and we sought its significance when the full ARIMA components were present.

In the next chapter, we discuss literature relevant to the problem at hand. In Chapter 3, we present the methodology that we implemented. In Chapter 4, we explain the details of the data and prepare them for the ARIMAX modeling. Once the data are fully clean, we perform an empirical investigation in Chapter 5. We discuss the results of this investigation in Chapter 6. All supplementary materials are presented in Appendices.

CHAPTER 2

LITERATURE REVIEW

The idea of utilizing social media data in explaining demand and sales is very new and has been attracting significant attention in recent years. Therefore, the empirical literature incorporating social media variables into a demand forecasting framework is still thin. The ground-breaking study in the related literature has been Asur and Huberman (2010). They regard social media as a form of collective wisdom and prove that it can predict the future (movie box-office revenues, in their specific case) better than the conventional gold standards in the industry. Most of the work in the operations management area studying social network effects is based on theoretical models. Candogan et al. (2012) consider the effect of the social network in explaining consumption behavior of individuals. Zhang et al. (2015) study the managing of services in the presence of social interactions. Papanastasiou and Savva (2014) present pricing strategies for new products when customers decide to purchase later to learn more information about the products. Behesti-Kashi et al. (2015) present a very comprehensive literature of the usage of social media in forecasting, with a special focus in sales forecasting of fashion industry. Chen et al. (2011) study the evolution of the relationship between social media and sales, and find that the relationship is different between the early and mature stages of internet usage. Stephen and Galak (2012) compare the effectiveness of two types of “earned media”, namely the traditional (e.g., publicity and press mentions) and social (e.g., blog and online community posts) in affecting sales. They find that the impact of the social earned media is larger than the traditional one, moreover, social earned media drives traditional earned media activity. In a very recent paper, John et al. (2017) take a controversial position and question whether “liking” a brand on Facebook causes a person to view it more favorably, and their answer is negative. In a recent paper that shares the same motivation with us, Kumar et al. (2016) examine the effect of firm-generated content (FGC) in social media on three key customer metrics: spending, cross-buying, and customer profitability. They use an extensive novel data set and find that after accounting for the effects of television advertising and e-mail marketing, FGC has a positive and significant effect on customers behavior. In their pioneering empirical work, Cui et al. (2017) investigate whether using publicly available social media data can improve the accuracy of daily sales forecasts. They implement various models to forecast sales and find that using social media information yields a statistically significant improvement in the out-of-sample forecast accuracy, with relative improvements ranging from 13 percent to 23 percent over different forecast horizons. Even more recently, Boone et al. (2018) claim that another type of user-generated content (customer search data, specifically one obtained from Google Trends) can be used to reduce out-of-sample forecast accuracy. They support Cui et al. (2017) by showing that adding customer search data to time series models improves their accuracy.

The study presented in this thesis, in which we study the empiricial value of social media information in forecasting the future demand, resembles Cui et al. (2017)'s work. It differs from theirs in terms of the methods we use and the dramatically different characteristics of the data sets we considered.

CHAPTER 3

METHODOLOGY

We followed the Box-Jenkins methodology to fit an analytical model to the time series. With various versions, such as Autoregressive Moving Average (ARMA), Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), Vector ARIMA (VARIMA), Fractional ARIMA (FARIMA), ARIMA with exogenous regressors (ARIMAX), Setc., the Box-Jenkins methodology is a well-established way of understanding and forecasting time series. We employed an ARIMAX model in our thesis. The basis of the ARIMAX model is the autoregressive moving average (ARMA) models, which were first developed by Box and Jenkins in 1970 (Box et al., 2008). Descriptions of the ARMA, ARIMA, and ARIMAX models follow.

The ARMA model consists of two building blocks, namely the AR and MA components, as its name suggests. For a series Y_t, the AR component refers to its relationship with its past values. This formulates the level of its current observations in terms of the level its lagged observations. The justification of the model stems from the fact that some time series mark the evolution of a phenomenon that evolves according to its history. For instance, a smoker may be guessed to have smoked yesterday and is expected to smoke tomorrow as well. Similarly, a factory with a high level of throughput is very likely to have produced large amounts yesterday and is expected to produce similar amounts tomorrow too. This serial dependence concept is formulated by the auto regressive (AR) model (Hyndman and Athanasopoulos, 2012). A description of the model follows.

The notation AR(p) refers to the autoregressive model of order p. The AR(p) model is represented as

$$Y_t = c + \sum_{i=1}^{p} \varphi_i \, Y_{t-i} + \epsilon_t$$

where φ_1,…, φ_p are the parameters of the model, c is a constant, and ϵ_t is the residual.

The second building block refers to the fact that the observations of a random variable at time t are not only affected by the shock at time t, but also the shocks that occur before time t. Thus, if we observed a negative shock to the production of an industry through new tariffs on imported raw materials, new regulations on labor market, or an unanticipated entry of a big competitor to the market, then we would expect that this negative effect to affect the production in the future. This concept can be represented by a moving average (MA) model (Hyndman and Athanasopoulos, 2012), which is described next.

The notation MA(q) refers to the moving average model of order q:

$$Y_t = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \, \epsilon_{t-i}$$

where θ_1,…, θ_q are the parameters of the model, μ is the expectation of Y_t (often assumed to equal 0), and the ϵ_t,…, ϵ_{t-1} are again white noise error terms.

The general autoregressive moving average process of orders p and q or ARMA(p,q) combines the AR and MA models into a unique representation. The ARMA process of orders p and q is defined as

$$Y_t = \; c + \epsilon_t + \sum_{i=1}^{p} \varphi_i \, Y_{t-i} + \sum_{i=1}^{q} \theta_i \, \epsilon_{t-i}$$

The ARMA formulation draws a more complete picture than the individual AR and MA models, because in actuality we rarely observe a random variable depending exclusively on either its past values or its past shocks.

ARMA models are vulnerable to factors such as trends and seasonality. An example is two completely unrelated time series that happen to either increase over time or exhibit seasonality which could falsely be identified to be related. A stationary time series is one whose properties do not depend on the time at which the series is observed (Diebold, 2007). It is important to adjust the series for seasonality or trend behavior because the seasonality and trend will affect the values of the time series at different times. Box et al. (2008) also claimed that nonstationary data could be turned to stationary data by differencing the series. Differencing consists in calculating the differences between consecutive observations. If we combine differencing with autoregression and a moving average model, we obtain an autoregressive integrated moving average (ARIMA) model. According to Hyndman and Athanasopoulos (2012), the model can be written as

$$Y'_t = \ c + \epsilon_t + \sum_{i=1}^{p} \varphi_i \, Y'_{t-i} + \sum_{i=1}^{q} \theta_i \, \epsilon_{t-i}$$

where Y'_t is the differenced series. The formulation is called an ARIMA(*p*,*d*,*q*) model, where

p = order of the autoregressive part

d = degree of first differencing involved

q = order of the moving average part

The significance of ARIMAX models is that they afford the ability to account for the impact of external variables on a time series. ARIMAX models consist of ARIMA, that is, the history of a series explained by its AR and MA components; and the X, which represents the external variables that we believe to have an effect on our time series. In short, an ARIMAX model simply adds in the covariate on the right-hand side (Hyndman, 2010):

$$Y_t = \beta X_t + c + \epsilon_t + \sum_{i=1}^{p} \varphi_i \, Y_{t-i} + \sum_{i=1}^{q} \theta_i \, \epsilon_{t-i}$$

where X_t is a causal/exogenous variable at time t, and β is its coefficient.

In the next chapter, we discuss the details of data preparation for the ARIMAX modelling.

CHAPTER 4

DATA PREPARATION

4.1. Gathering the data

The data set utilized in this thesis is the weekly sales series of Walmart from February 2010 to October 2012 for 45 stores located across the US. The data are publicly available at https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting. Also available is the store size variable. An initial exploration of the aggregate (45 stores combined) sales data reveals the behavior described in Figure 1.

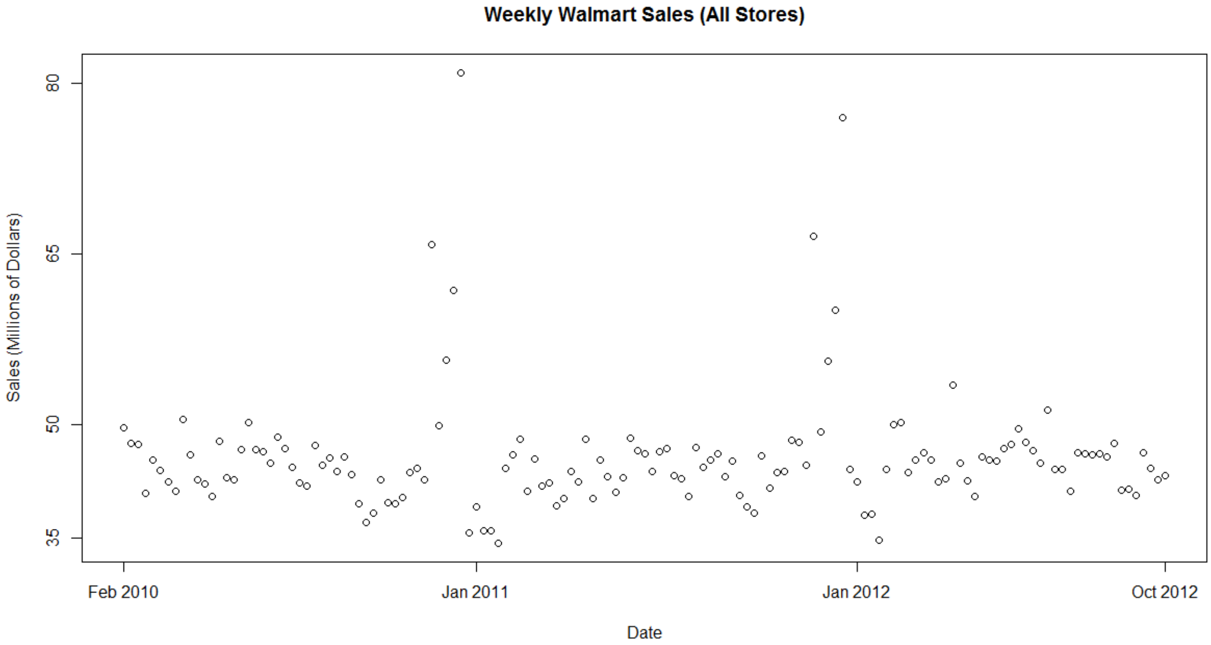


Figure 1. Aggregate weekly Walmart sales data over time

Every Walmart store has 99 departments. We firstly aggregated all department sales data for every store to obtain one sales data point per store per week. The store size variable ranged between 34,875 and 219,622 square feet, with a mean of 130,287 square feet. We believed that the store area might play a role in the behavior of sales series, i.e. stores of different sizes might react differently to factors such as holidays, social media stimulus, etc. Therefore, we incorporated this information into the analysis by grouping the stores into four buckets: Large, Upper Medium, Lower Medium, and Small and estimating four models for each group. This grouping scheme is illustrated in Table 1. A histogram that reflects the distribution of the stores in terms of their sizes is furnished in Figure 2.

Bins*	**Frequency**	**Group**
<50	10	Small
50-75	2	Lower Medium
75-100	2	Lower Medium
100-125	7	Lower Medium
125-150	4	Upper Medium
150-175	5	Upper Medium
175-200	2	Upper Medium
>200	13	Large
*in thousand sq. ft		

Table 1. The groups of the stores

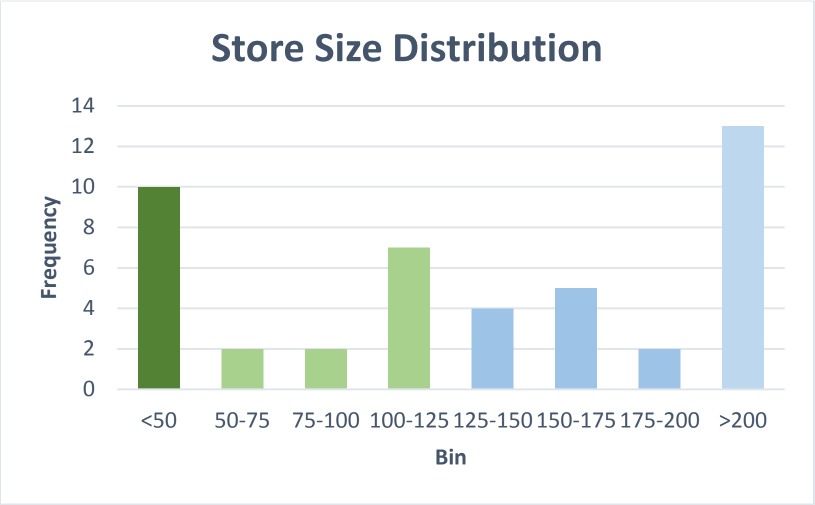


Figure 2. Distribution of the stores

We separated whole sales data into four groups, according to store type. The sales graphs for each store type are displayed in Figures 3 to 6, respectively.

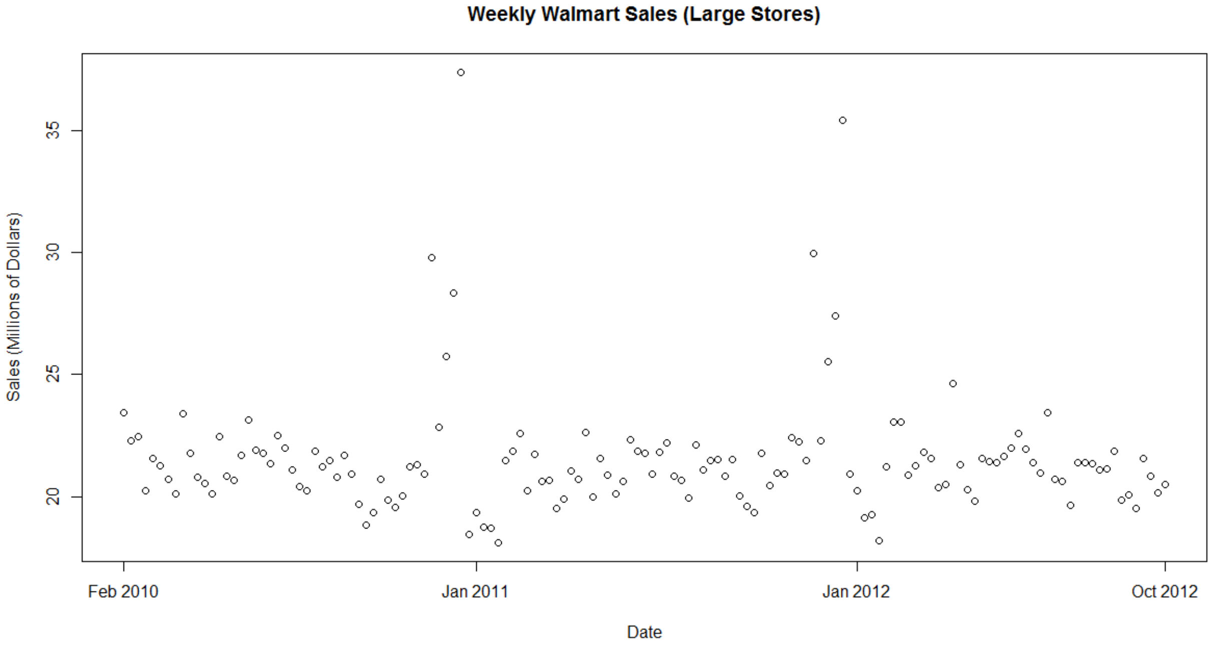


Figure 3. Weekly Walmart sales data for the large stores

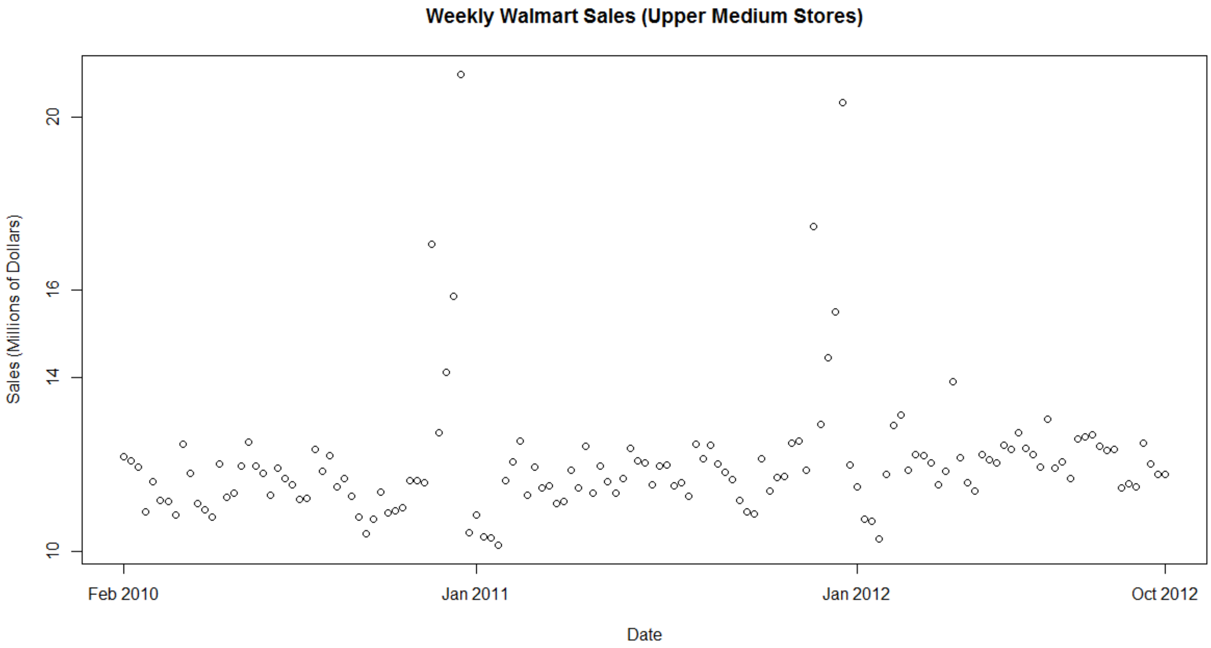


Figure 4. Weekly Walmart sales data for the upper medium stores

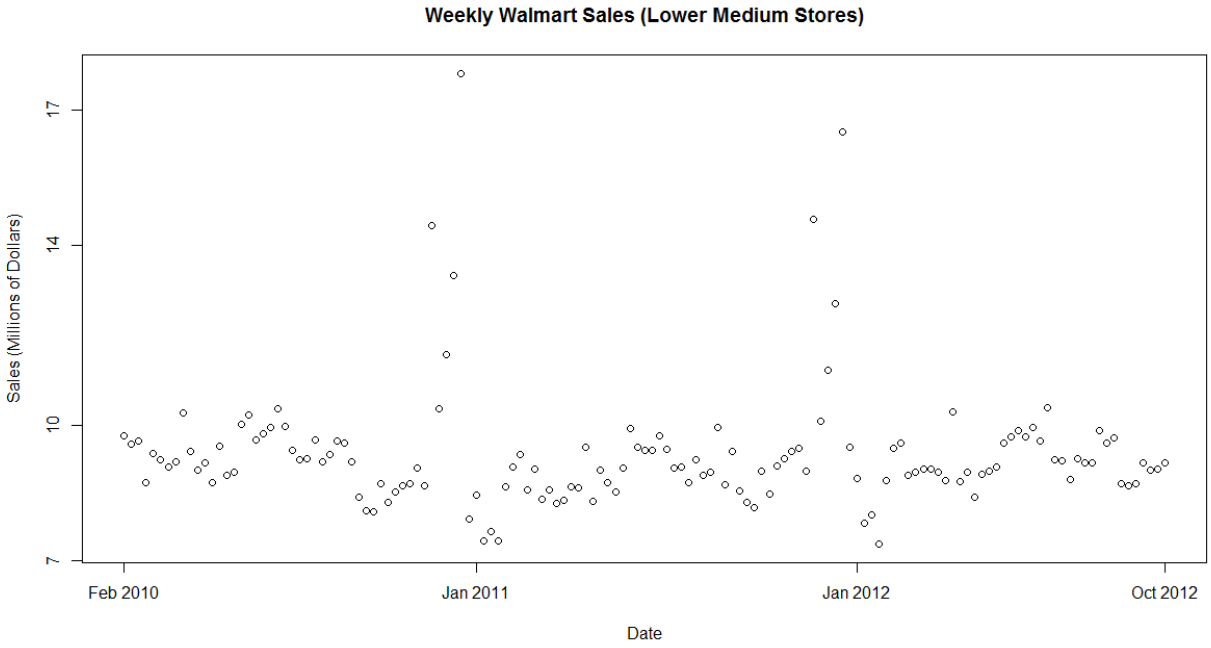


Figure 5. Weekly Walmart sales data for the lower medium stores

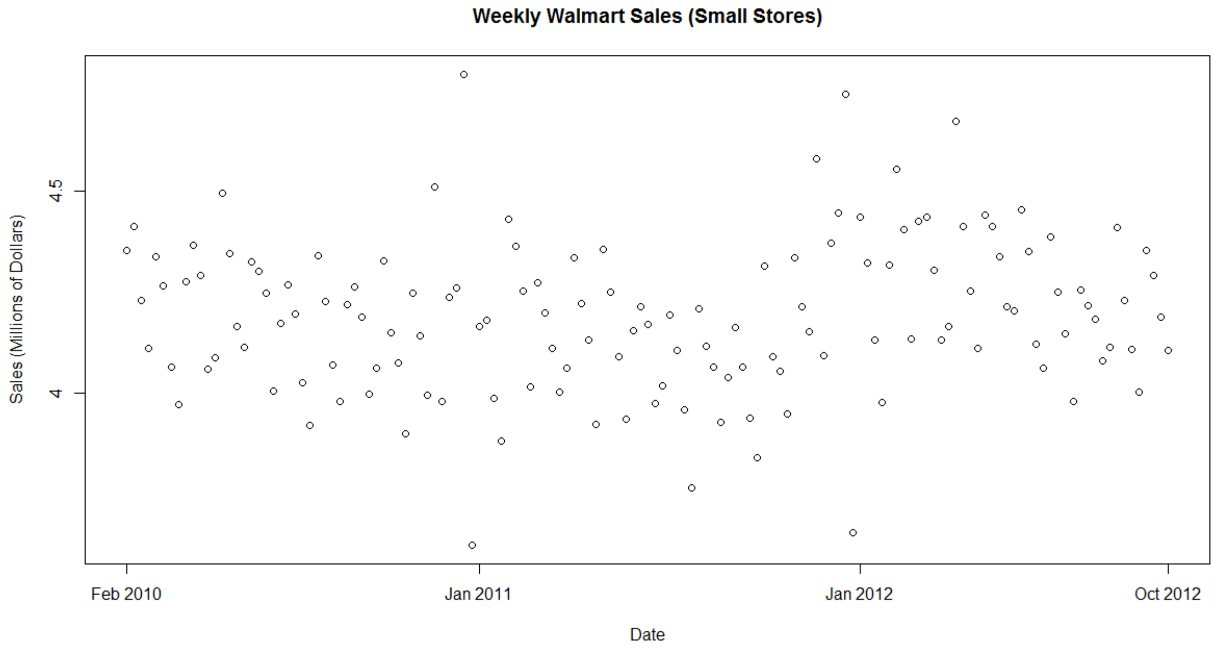
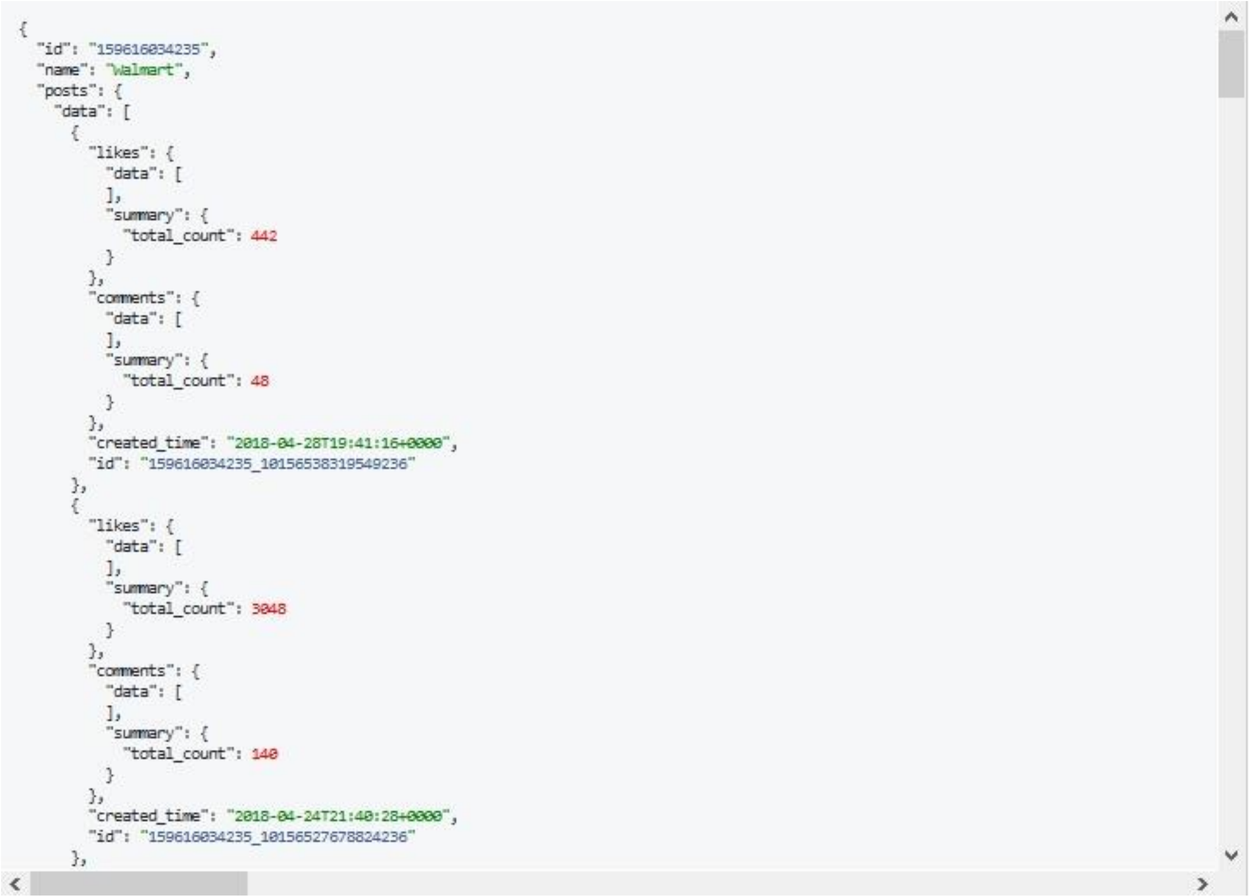


Figure 6. Weekly Walmart sales data for the small stores

Facebook data was the social media component considered in this thesis. The company's official Facebook account is https://www.facebook.com/WalmartcomUS/. The explanatory variables consist of the time series of number of likes, and comments for posts. The advantage of using Facebook is that it offers a public application programming interface (API) to access the complete data set of activities of Walmart on Facebook. This Facebook API (https://developers.facebook.com/) gives an access token to anybody who wants to query it, and the data shown on each page of API includes the date, number of likes, and number of comments of every single post for the specified time period. In Figure 7, we provide an example of Facebook API page. We attained the Facebook activity data through Facebook API using the Python programming language, following the techniques clearly discussed by Russell (2014). A sample of final data file is shown in Appendix 1.



```
{
  "id": "159616034235",
  "name": "Walmart",
  "posts": {
    "data": [
      {
        "likes": {
          "data": [
          ],
          "summary": {
            "total_count": 442
          }
        },
        "comments": {
          "data": [
          ],
          "summary": {
            "total_count": 48
          }
        },
        "created_time": "2018-04-28T19:41:16+0000",
        "id": "159616034235_10156538319549236"
      },
      {
        "likes": {
          "data": [
          ],
          "summary": {
            "total_count": 3048
          }
        },
        "comments": {
          "data": [
          ],
          "summary": {
            "total_count": 140
          }
        },
        "created_time": "2018-04-24T21:40:28+0000",
        "id": "159616034235_10156527678824236"
      },
```

Figure 7. A sample Facebook API page

We noticed that at the beginning of the time span we studied, the number of likes and number of comments were low, and that they increased over time. The company's activity was rare, too. For instance, Walmart published an average of one post every two days at the beginning of the study period, whereas later there could be up to five daily posts. Hence, we detected a scale effect that reflects a behavior of a higher volume of Facebook activity at the end. We removed the scale effect by taking the average number of likes and comments per post into consideration. In Figures 8 and 9 we show the behavior of the average weekly number of likes and comments from February 2010 to October 2012, respectively. The next step was to address issues such as trends, unit roots, and seasonality.

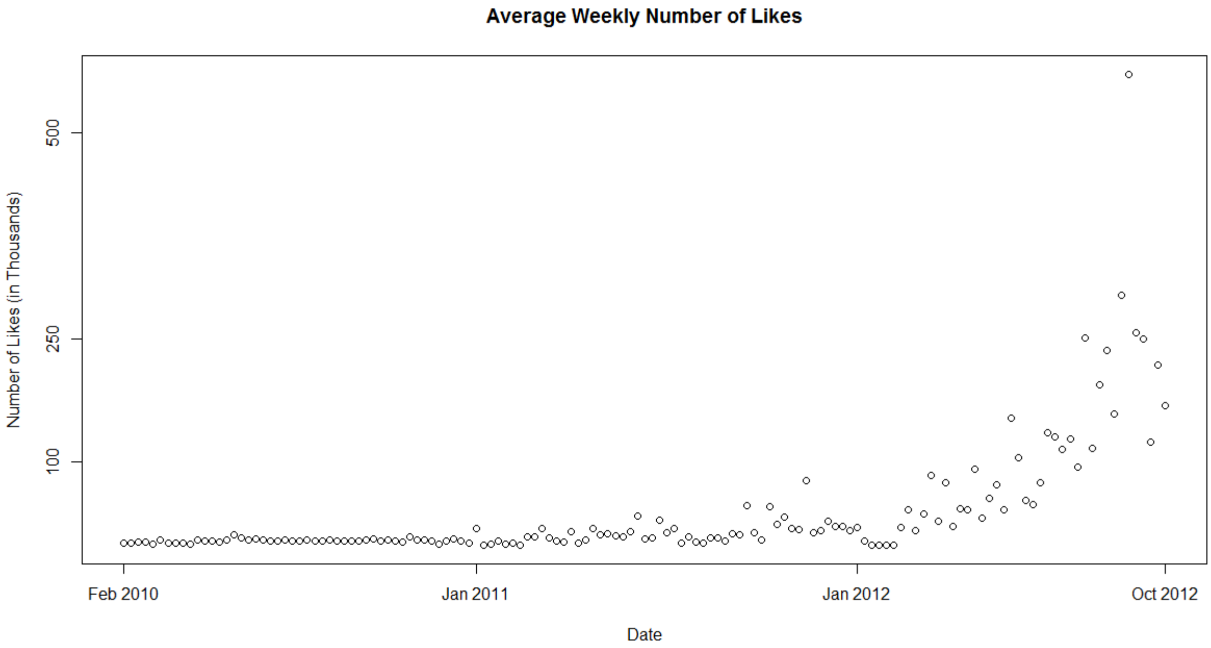


Figure 8. Average weekly number of likes

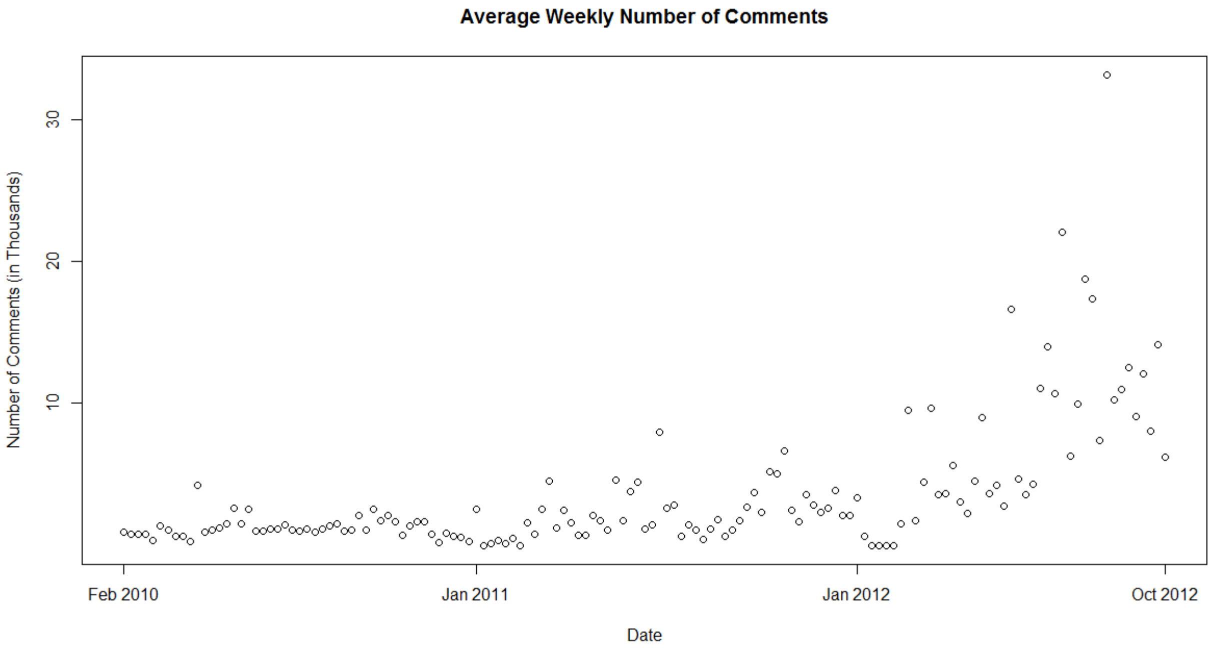


Figure 9. Average weekly number of comments

4.2. Data analysis

The analysis of seasonality, trends, and cycles are crucial aspects in time series analysis. These components capture the historical patterns in the time series. One series does not have to have all three components necessarily, but if they exist, they should be removed before analyzing the series. Seasonal components are the fluctuations in the data related to calendar cycles. Trend refers to an overall pattern of the series. Cycles are decreasing and increasing patterns that are not seasonal (Diebold, 2007). The process of removing these components from the series is referred to as decomposition. Once we know the patterns seasonality, trends, and cycles, we should check if the series is stationary or not.

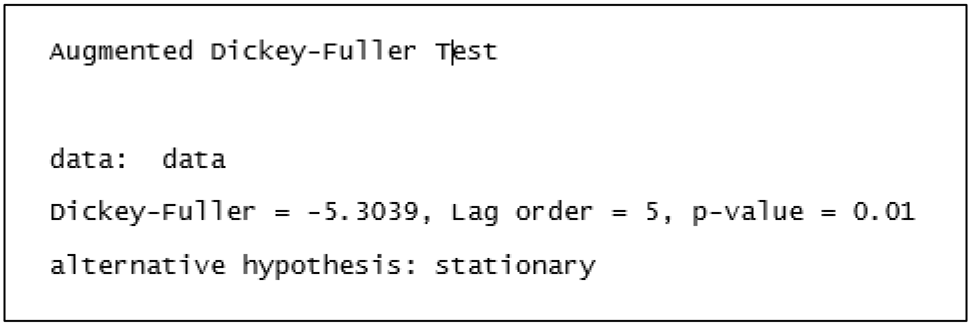
Fitting an ARIMAX model requires the series to be stationary (Hyndman, 2010). A series to be classified as stationary should meet the condition that its mean, variance, and autocovariance are time invariant. As the ARIMAX model uses previous lags of series to model its behavior besides an exogenous variable, modeling a stable series which has consistent properties provides less uncertainty (Hyndman and Athanasopoulos, 2012). We used the Augmented Dickey-Fuller Test (ADF) to check for stationarity. The null hypothesis is that the series is non-stationary; or, in other words, integrated, mathematically, the null hypothesis is that there is a unit-root in the series. The alternative hypothesis is that the series is stationary. Following the ADF procedure, we tested whether the change in Y can be explained by lagged values and a linear trend (Hyndman and Athanasopoulos, 2012). If the contribution of the lagged value to the change in Y is zero, then the lagged value will have no effect on the change in Y, which will imply that the series is not going to be mean-reverting from today to tomorrow. Therefore, the series is going to be deemed non-stationary.

We used the **adf.test** function on R to test for stationarity of all the series. In the summary tables that we provide, it suffices to check whether the resulting *p*-values of the test are less or greater than the significance level of 0.05. The former means stationarity, and the latter, non-stationarity.

Non-stationary variables are not allowed to enter the statistical model as is. They need to be made stationary through a transformation of differencing or log-differencing, and the transformed series should be tested again by the ADF test to make sure that they become stationary.

We first tested the stationarity of the aggregate weekly sales series. As can be seen from Figure 10, the ADF test resulted in a conclusion of stationarity. However, an analysis of the graph (Figure 1) reveals the presence of outliers that do not obey a stationary behavior. Therefore, the test merely tolerates the existence of these outliers, which distorts stationarity.

We tended to remove these outliers, but before that, we checked whether there was seasonality in the data, which would give us the possibility of eliminating them through seasonal differencing. We examined the Autocorrelation Function (ACF) plot of the series (Figure 11). The ACF plot is a graphical representation of the autocorrelation coefficients of a time series, which plots the correlations between its current and lagged values. This depicts a measure of the linear relationship of a series with its past records (Hyndman and Athanasopoulos, 2012).



```
Augmented Dickey-Fuller Test


data:  data
Dickey-Fuller = -5.3039, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

Figure 10. ADF test result for weekly Walmart sales for all stores

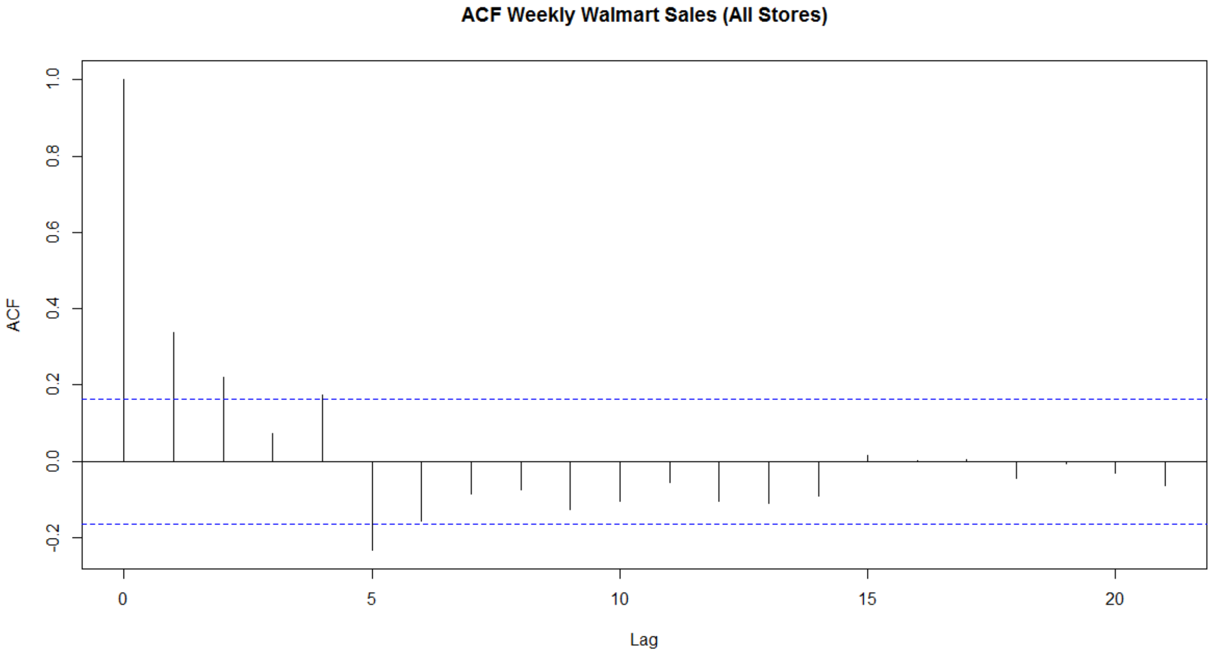


Figure 11. ACF plot for weekly Walmart sales for all stores

The sales graph appeared to spike around the same time every year, which led us to suspect annual seasonality. Even though we see co-movement between the Christmas season sales, this is not enough for annual seasonality. An annual seasonality is said to exist if the same months or seasons co-move during the entire year, between two separate years. For example, we would say that the series had seasonality if the sales on February 2010 were correlated with that on February 2011, March 2010 sales were correlated with March 2011 sales, and so on. Thus, seasonality would result in an autocorrelation behavior with lag 12. The ACF plot did not reveal that kind of a pattern, which suggested that there was no seasonality. This meant that the same months of different years do not correlate, and the annual spikes in the sales series only pertain to the Christmas season. Therefore, the Christmas season merely constituted an outlier behavior. In conclusion, since the high Christmas season sales were not a part of seasonality and were so sparse that the ADF test did not notice them, we opted to remove those data points from the data series.

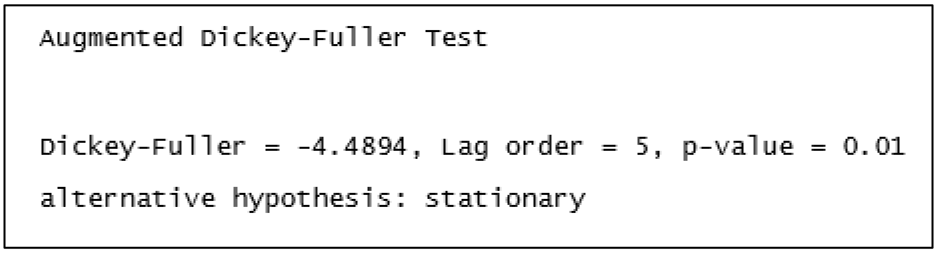
As we mentioned before, the application of an ARIMAX model requires a series to be stationary. In the next section we describe whether the sales series for all store types and weekly average number of likes and comments were stationary.

4.3. Checking the data for stationarity

We performed separate ADF tests for sales series for every store type, and average number of likes and comments series to check stationarity.

When we performed ADF test for sales data for large stores, we obtained the following results (Figure 12). Since the *p*-value (0.01) was less than 0.05, the null hypothesis was rejected in favor of stationarity.

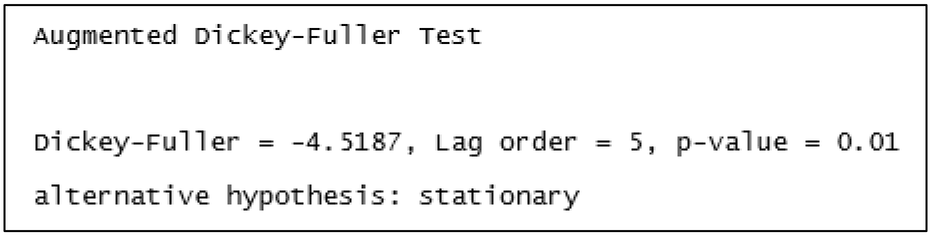
The result of ADF test for sales data of upper medium stores is shown in Figure 13. Again, since the *p*-value (0.01) was less than 0.05, we reject the null hypothesis. Similarly, we concluded that the time series for lower medium stores was stationary, and that for small stores it was not stationary.



```
Augmented Dickey-Fuller Test


Dickey-Fuller = -4.4894, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

Figure 12. ADF test result for the large stores

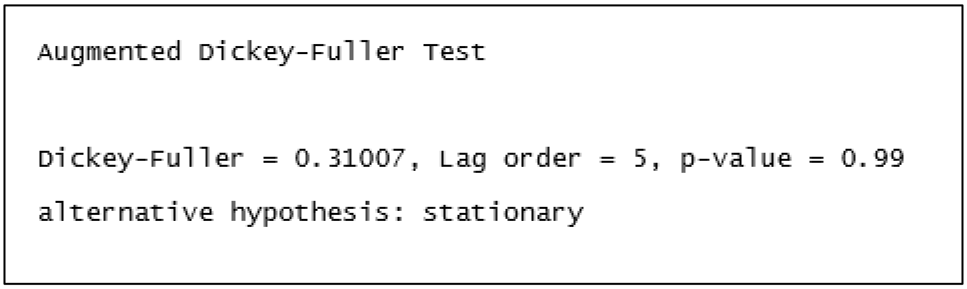


```
Augmented Dickey-Fuller Test


Dickey-Fuller = -4.5187, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

Figure 13. ADF test result for the upper medium stores

When we implemented ADF test for the average number of likes data, we obtained the results depicted in Figure 14. Since the *p*-value (0.99) was greater than 0.05, we failed to reject the null hypothesis, and concluded that the average number of likes was non-stationary.



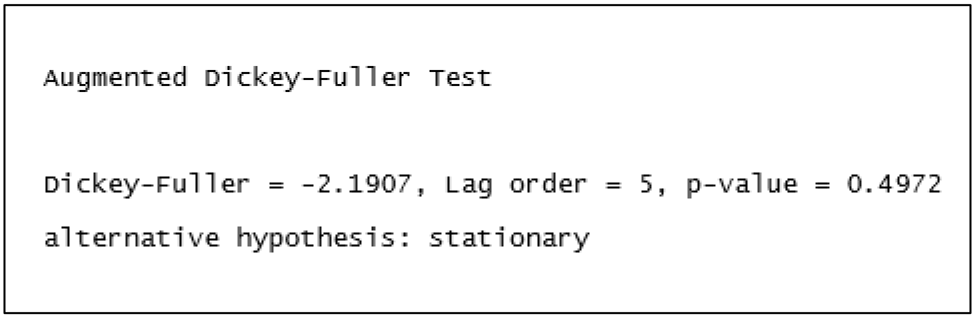
```
Augmented Dickey-Fuller Test


Dickey-Fuller = 0.31007, Lag order = 5, p-value = 0.99
alternative hypothesis: stationary
```

Figure 14. ADF test result for the average number of likes

In Figure 15 we show the result of the ADF test for the average number of likes. We failed to reject the null hypothesis for the series of average number of comments, as the *p*-value (0.49) is greater than 0.05. On the other hand, sales series for small stores was non-stationary, for that reason we had to stationarize the data before starting the actual analysis.

As the average number of likes and comments series were non-stationary, we must perform a decomposition process to make them suitable for ARIMAX analysis.



```
Augmented Dickey-Fuller Test


Dickey-Fuller = -2.1907, Lag order = 5, p-value = 0.4972
alternative hypothesis: stationary
```

Figure 15. ADF test result for the average number of comments

We next discuss the details of decomposition process of series of small store sale, and average number of likes and comments.

We first subtracted the original time series from its lagged series to extract trends or cycles from the data. The original series (Y_t) was subtracted from its lagged series (Y_{t-n}). The formulas are as follows (Diebold, 2007):

Not differencing (d=0) $Y_t^d = Y_t$

First differencing (d=1) $Y_t^d = Y_t - Y_{t-1}$

We attempted to remove the trend through first order differencing for the small store sales, and for the average number of likes and comments series. Plots of the differenced series are furnished in Figures 16, 17, and 18. As noticed, the trend component of the series was extracted and the differenced data (residual) did not show any trend after first-order differencing. The series was not found to be stationary on variance as evidenced by the changing levels of variation. Further analysis was thus necessary.

The following equation represents the log transform process (Diebold, 2007):

Log of sales $Y_t^l = \log(Y_t)$

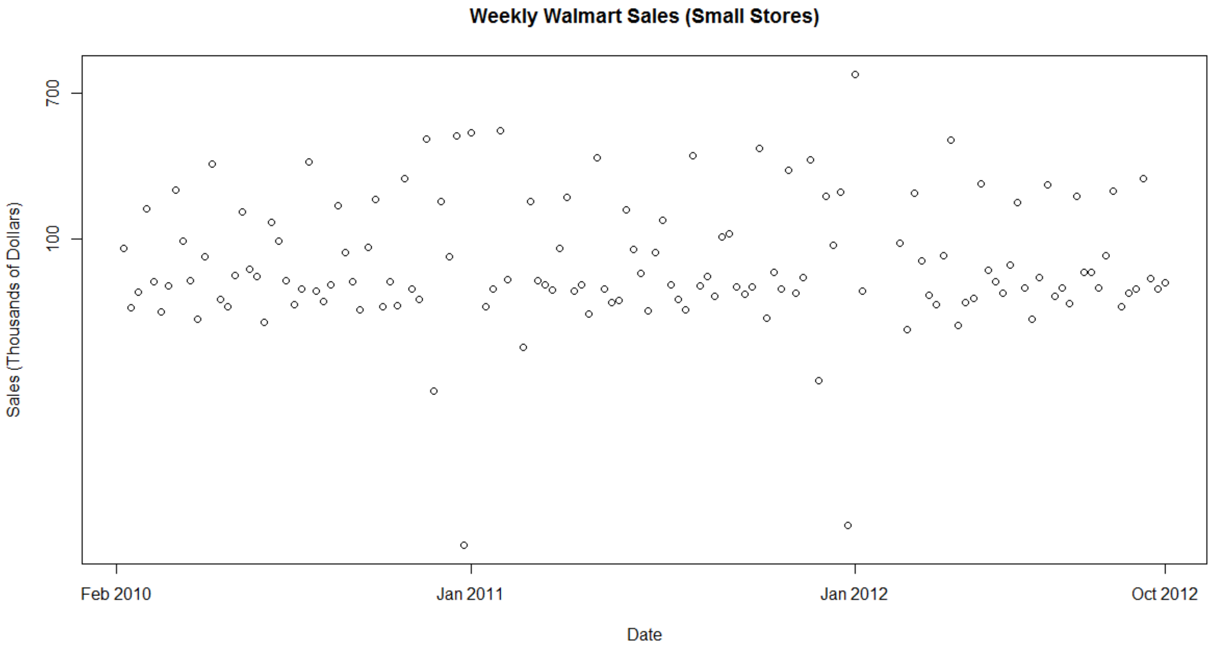


Figure 16. Weekly Walmart sales data for the small stores after differencing

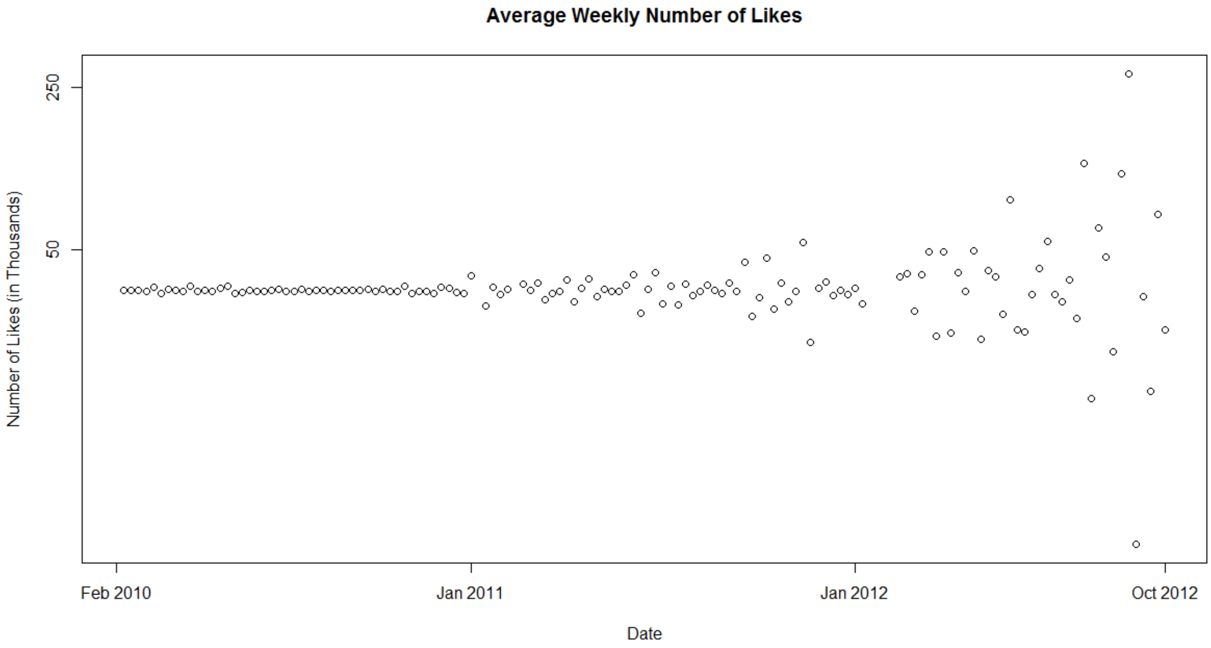


Figure 17. Average weekly number of likes after differencing

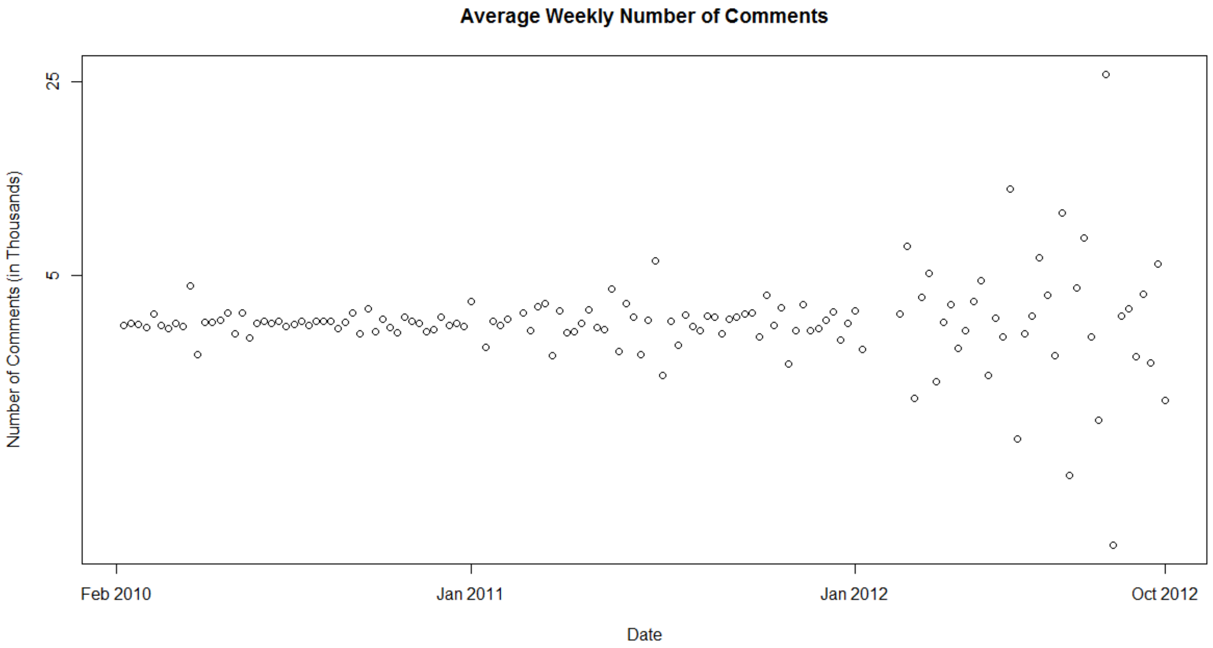


Figure 18. Average weekly number of comments after differencing

In Figures 19 to 21 we show the output plots for small store sales, and average number of likes and comments, respectively. The new series seemed stationary on variance. Differencing and log transform operations solved only one part of non-stationary problem separately; the former made the series stable on the mean whereas the latter transformed the series to a stationary one on variance. To obtain a fully stationary series, both operations must be applied together.

In order to reconfirm that the series were stationary on mean and variance, we looked at the differenced plot for log - transformed series. The mathematical representation of the difference log transform process is, according to Diebold (2007),

1st differencing (d=1) of log of series $Y_t' = \log\ (Y_t) - \log(Y_{t-1})$

In Figures 22, 23, and 24 we show the plots for the -aforementioned mathematical equation. The series seemed stationary on mean and variance. We still needed to determine whether the series were indeed stationary by performing an ADF test for on each series.

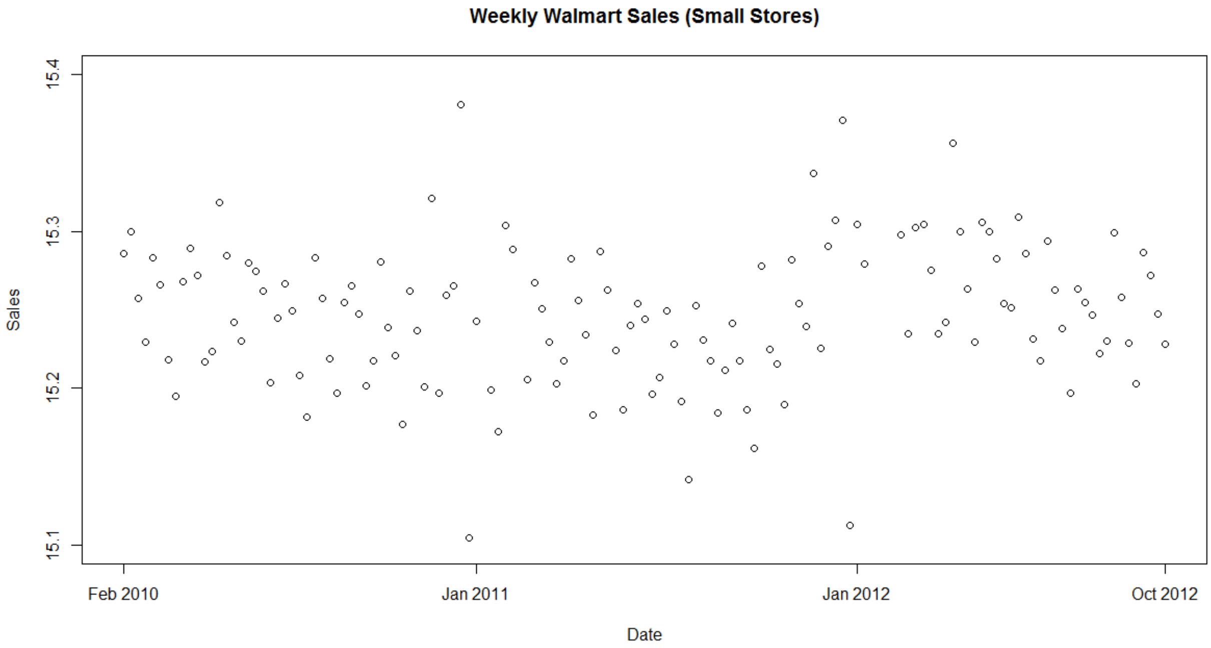


Figure 19. Weekly Walmart sales data for the small stores after log transform

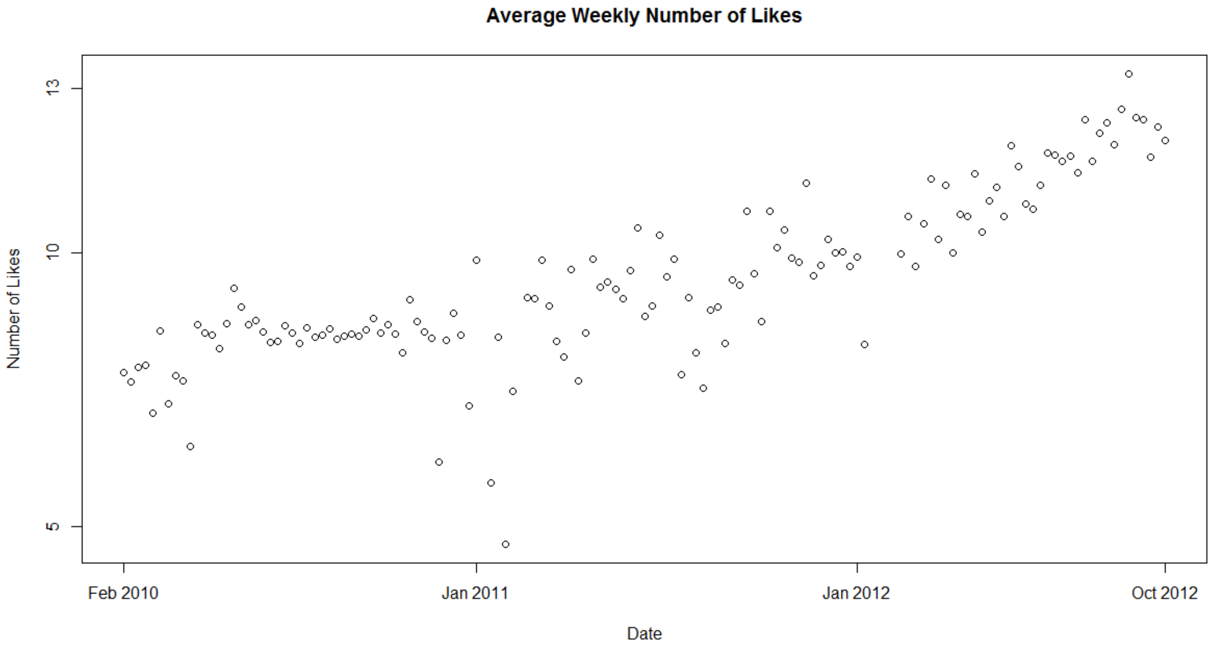


Figure 20. Average weekly number of likes after log transform

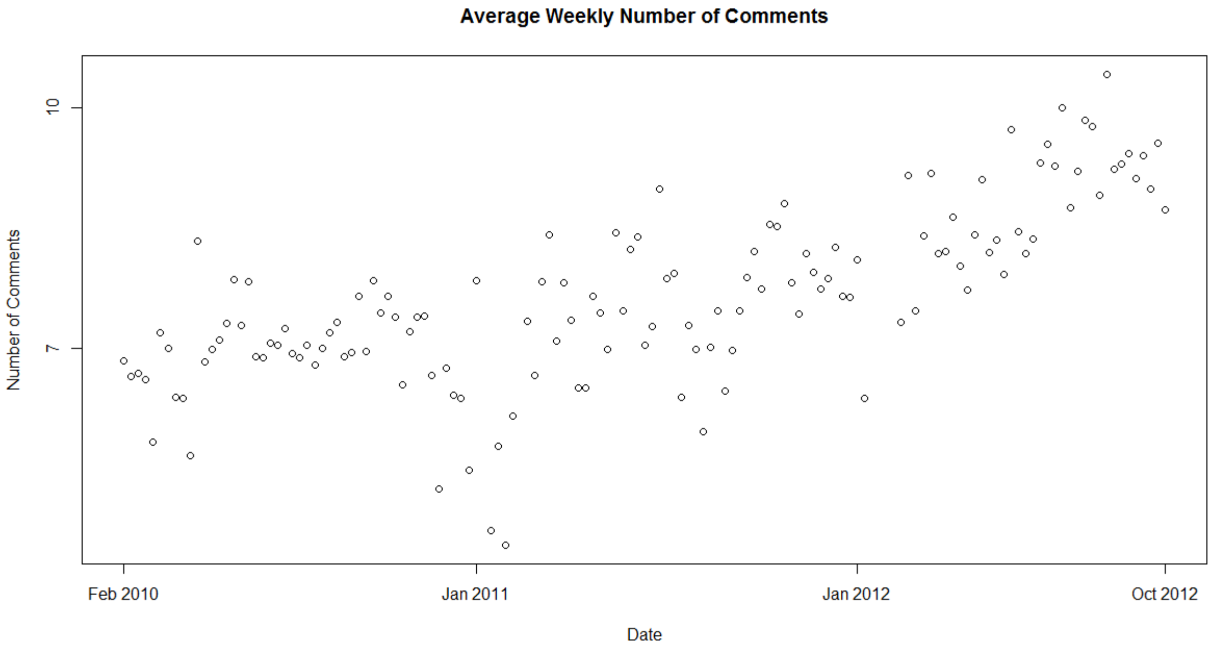


Figure 21. Average weekly number of comments after log transform

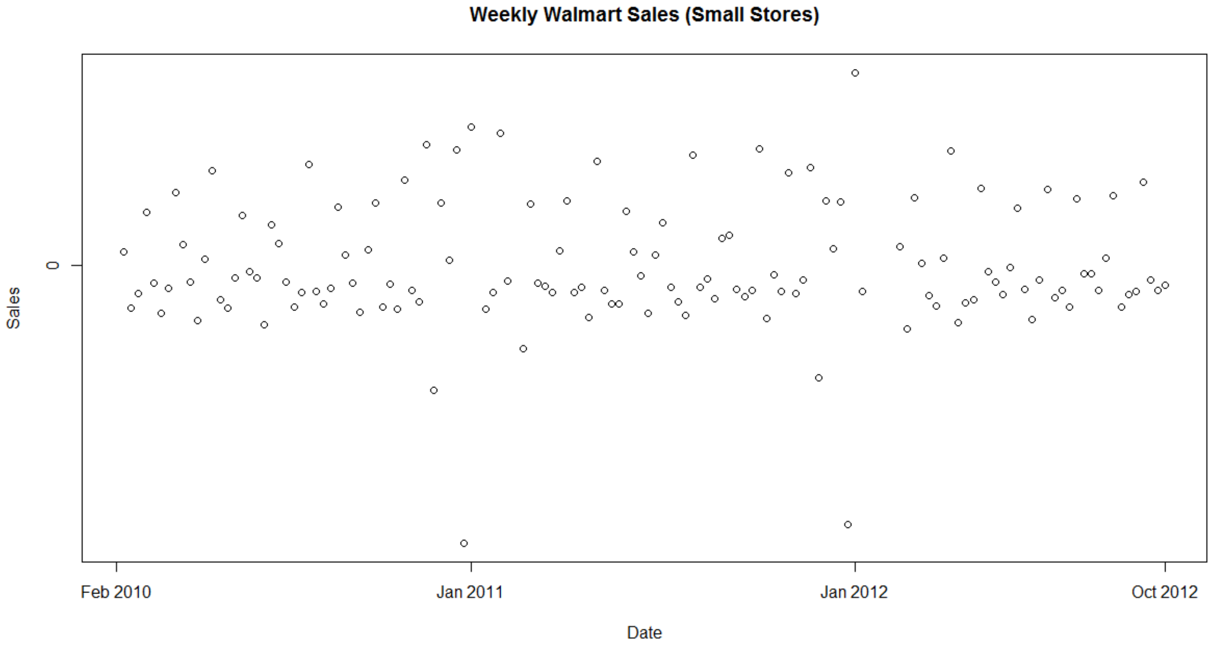


Figure 22. Weekly Walmart sales data for the small stores after differencing and log transform

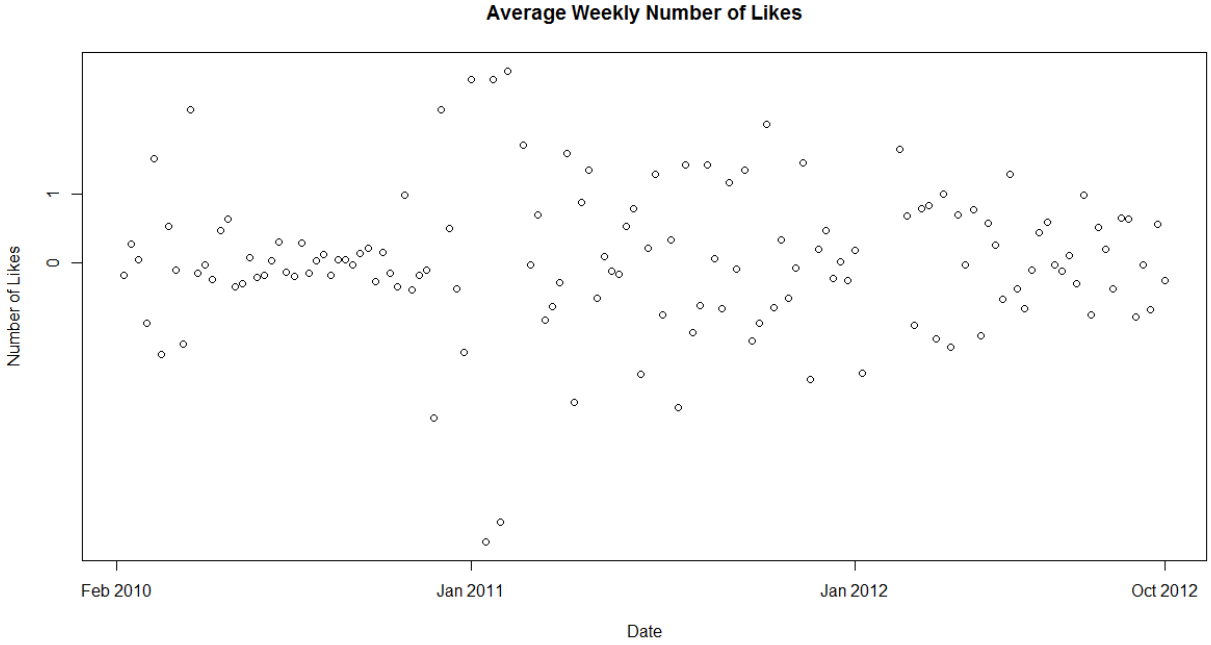


Figure 23. Average weekly number of likes after differencing and log transform

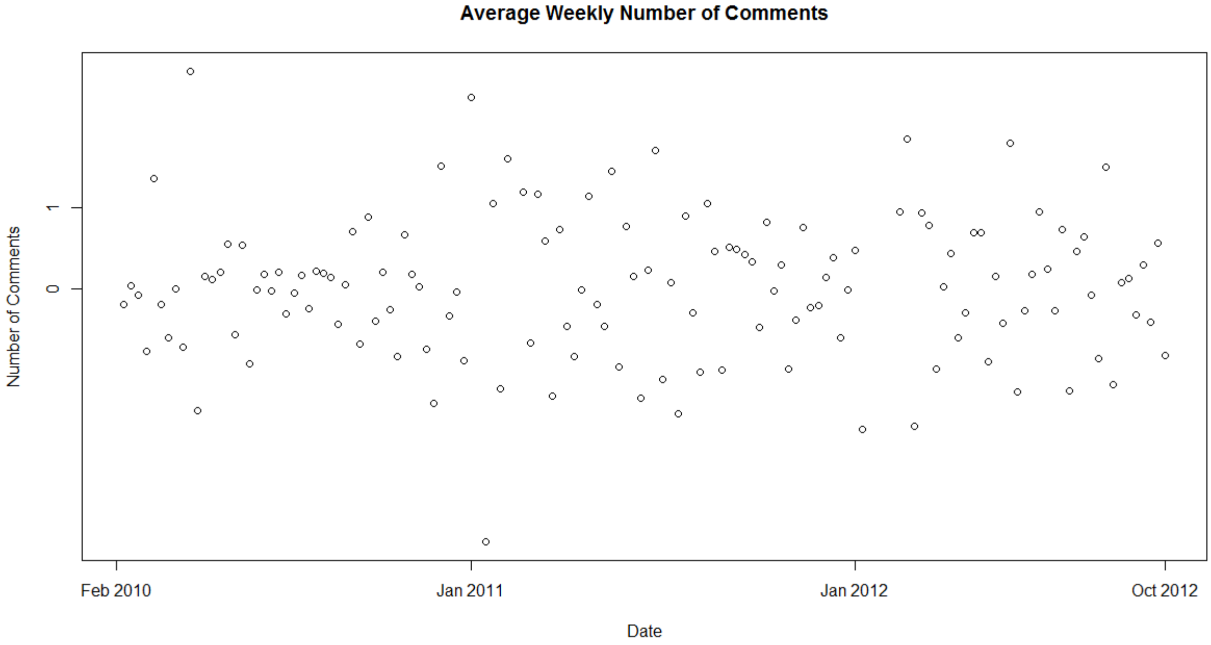


Figure 24. Average weekly number of comments after differencing and log transform

An ADF test was performed for the differenced and log transformed of (1) the small store sales series, (2) the average number of likes, and (3) the average number of comments. In all cases, the tests yielded the conclusion that the series were stationary. The data were deemed suitable for ARIMAX analysis, as they satisfied the stationary conditions. The R code for the data preparation phase is shown in Appendix 2. Also, the R code for plotting all the figures is displayed in Appendix 3.

In Chapter 5, we discuss the details of our ARIMAX investigation.

CHAPTER 5

EMPIRICAL INVESTIGATION

5.1. Development of the ARIMAX model

In this section, we describe the use of the **auto.arima()** function on R to find the most appropriate parameters of an ARIMAX model. The function allows users to automatically produce a set of optimal (*p*, *d*, *q*). It achieves this by searching through multiple alternatives and it uses a variation of the Hyndman and Khandakar algorithm (Hyndman and Khandakar, 2008) that combines unit root tests, minimization of the Akaike Information Criterion (AICc) and Bayesian Information Criterion (BIC). In essence, this algorithm optimizes the ARIMA model fit by looping over its different specifications, presented in a schematic form in Figure 25.

The exogenous variables were the average weekly number of likes and the average weekly number of comments. As discussed in the data preparation section, we divided the raw sales data into four in terms of store sizes. Hence, we performed the ARIMAX analysis for each store type data.

As discussed in the previous chapter, we used the sales data without any differencing or log transform operation for the large, upper medium, and lower medium store types, as they were stationary. We considered differenced and log transformed versions of the series of sales for small stores, the average number of likes and comments, since only these versions achieved stationarity. Differencing subtracts the lagged values from the original series, as a result, the number of data points is one less than the original one after differencing. As we performed differencing for exogenous variables but not for sales, the number of data points in exogenous variables series was

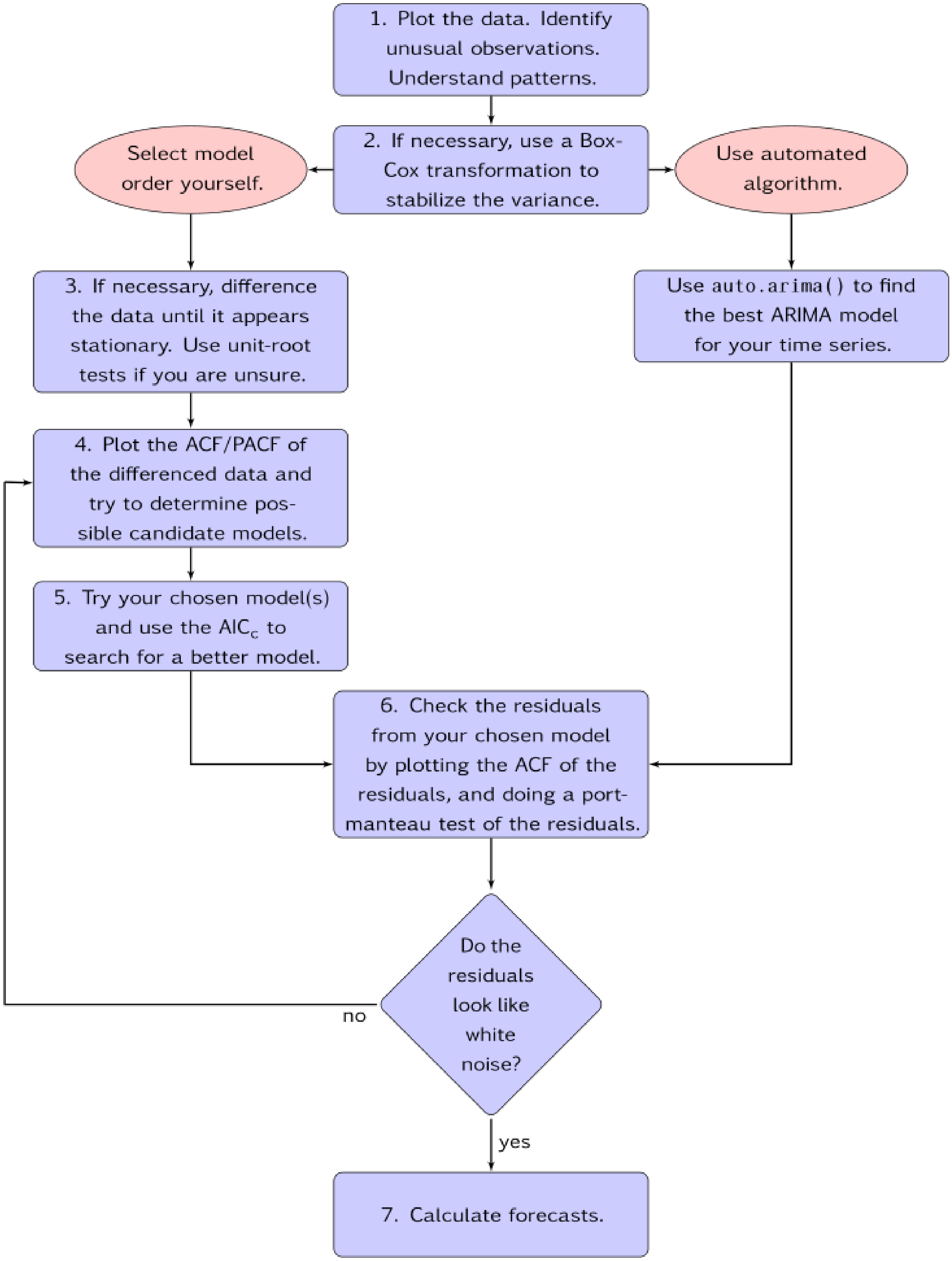

1. Plot the data. Identify unusual observations. Understand patterns.
2. If necessary, use a Box-Cox transformation to stabilize the variance.
Select model order yourself.
Use automated algorithm.
3. If necessary, difference the data until it appears stationary. Use unit-root tests if you are unsure.
Use auto.arima() to find the best ARIMA model for your time series.
4. Plot the ACF/PACF of the differenced data and try to determine possible candidate models.
5. Try your chosen model(s) and use the AICc to search for a better model.
6. Check the residuals from your chosen model by plotting the ACF of the residuals, and doing a portmanteau test of the residuals.
Do the residuals look like white noise?
no
yes
7. Calculate forecasts.

Figure 25. General process for forecasting using an ARIMA model

one less than the number of data points of the sales series for the large, upper medium and lower medium stores. For that reason, in order to eliminate this inequality, we just removed the first data point of sales series that were already stationary.

In conclusion, we needed to set up two different ARIMAX models: one would be for the sales series of large, upper medium, and lower medium stores which were already stationary, and another one would be for the sales series of small stores which was not originally stationary.

The ARIMAX model that we set up and tested for small stores is represented with the following equation:

$$Y_t' = \beta X_t' + c + \epsilon_t + \sum_{i=1}^{p} \varphi_i \, Y'_{t-i} + \sum_{i=1}^{q} \theta_i \, \epsilon_{t-i}$$

where

$Y_t' = logY_t - logY_{t-1}$,

$X_t' = [logL_t - logL_{t-1}, logC_t - logC_{t-1}]$,

$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$

L_t stands for average number of likes, and

C_t stands for average number of comments.

The following equation represents the ARIMAX model that was set up and tested for the sales series of large, upper medium, and lower medium stores:

$$Y_t = \beta X_t' + c + \epsilon_t + \sum_{i=1}^{p} \varphi_i \, Y_{t-i} + \sum_{i=1}^{q} \theta_i \, \epsilon_{t-i}$$

where X_t' and β follow the aforementioned definitions.

5.1.1. ARIMAX model for the large stores

In Table 2 we show the coefficient for an ARIMAX model for the large stores with exogenous variables of average number of likes and comments. The results of a coefficient test for a *z* test is also shown in the table. An ARIMA (1,0,0) configuration was recommended with an auto regressive term and exogenous variables parameters. In the light of this information, the model can be represented by the following formula:

$$Y_t = -183836.2(logL_t - logL_{t-1}) + 108821.6(logC_t - logC_{t-1}) + 21050508.1 + \epsilon_t + 0.2349Y_{t-1}$$

```
Regression with ARIMA(1,0,0) errors


Coefficients:
         ar1   intercept  average_likes  average_comments
      0.2349  21050508.1      -183836.2          108821.6
s.e.  0.0862    117905.5       107238.4          127752.8


                  Pr(>|z|)
ar1               0.006432 **
intercept        < 2.2e-16 ***
average_likes     0.086478 .
average_comments  0.394317
```

Table 2. ARIMAX outputs for the large stores

According to the results, the intercept was significant at 0.1% level of significance, while the AR(1) component (which is the lagged value) was significant at 1% level of significance. The *p*-values of coefficients for the average number of likes and comments are 0.086 and 0.394, respectively, which means that the average number of comments were not significant at any level, and likes were significant only at 10% significance level.

5.1.2. ARIMAX model for the upper medium stores

In Table 3 we show the coefficient for an ARIMAX model for the upper medium stores with exogenous variables of average number of likes and comments with the results of a coefficient test for a *z* test. An ARIMA (1,0,2) configuration was recommended with an auto regressive term, two moving average terms and exogenous variables parameters. The model can be represented by the following formula:

$$Y_t = -250071.27(logL_t - logL_{t-1}) + 54700.57(logC_t - logC_{t-1}) + 12054984.5 + \epsilon_t$$
$$- 0.7712Y_{t-1} + 1.2955\epsilon_{t-1} + 0.6854\epsilon_{t-2}$$

The intercept, AR(1), MA(1), and MA(2) coefficients were significant even at 0.1% level of significance. The average number of likes was found to have a significant effect on explaining sales since it had a *p*-value of 0.009. The average number of comments turned out to be insignificant.

```
Regression with ARIMA(1,0,2) errors


Coefficients:
          ar1     ma1     ma2   intercept  average_likes  average_comments
      -0.7712  1.2955  0.6854  12054984.5     -250071.27          54700.57
s.e.   0.0745  0.0727  0.0724    183147.3       96785.38         118373.68


               Pr(>|z|)
ar1           < 2.2e-16 ***
ma1           < 2.2e-16 ***
ma2           < 2.2e-16 ***
intercept     < 2.2e-16 ***
avg_likes      0.009773 **
avg_comments   0.644009
```

Table 3. ARIMAX outputs for the upper medium stores

5.1.3. ARIMAX model for the lower medium stores

In Table 4 we show the coefficient for an ARIMAX model for the lower medium stores with exogenous variables of average number of likes and comments. The results of a coefficient test are also shown in the table. An ARIMA (2,0,2) configuration was recommended with two auto regressive term, two moving average terms and exogenous variables parameters. In light of this information, the model can be represented by the following formula:

$$Y_t = -270085.24(logL_t - logL_{t-1}) + 99127.9(logC_t - logC_{t-1}) + 9408152.5 + \epsilon_t$$
$$- 0.9138Y_{t-1} - 0.1264Y_{t-2} + 1.3953\epsilon_{t-1} + 0.7455\epsilon_{t-2}$$

```
Regression with ARIMA(2,0,2) errors


Coefficients:
          ar1      ar2     ma1     ma2  intercept  average_likes
      -0.9138  -0.1264  1.3953  0.7455  9408152.5     -270085.24
s.e.   0.1223   0.1168  0.0818  0.0764   155530.4       92748.12

      average_comments
               99127.9
s.e.          111578.2


                   Pr(>|z|)
ar1               7.852e-14 ***
ar2                0.278919
ma1               < 2.2e-16 ***
ma2               < 2.2e-16 ***
intercept         < 2.2e-16 ***
average_likes      0.003591 **
average_comments   0.374317
```

Table 4. ARIMAX outputs for the lower medium stores

Again, the intercept, AR(1), MA(1) and MA(2) coefficients all had *p*-values less than 0.001, Since the average number of likes had a *p*-value 0.003, it had a significant effect in explaining the model at 1% level of significance. The *p*-values for AR(2) and average comments were greater than 0.10, hence they have no power on explaining the model.

5.1.4. ARIMAX model for the small stores

In Table 5 we show the coefficients for an ARIMAX model for the small stores with exogenous variables of average number of likes and comments with the results of a coefficient test for a *z* test. An ARIMA (3,0,1) configuration was recommended with three auto regressive term, one moving average term and exogenous variables parameters. The model can be represented by the following formula:

$$logY_t - logY_{t-1} = -0.0010(logL_t - logL_{t-1}) - 6e - 04(logC_t - logC_{t-1}) + \epsilon_t - 0.5632Y_{t-1} - 0.5637Y_{t-2} - 0.4761Y_{t-3} - 0.5088\epsilon_{t-1}$$

According to the results, the autoregressive components up to level 3, and the MA(1) coefficients have *p*-values less than 0.001. It was surprising that the model did not estimate an intercept parameter. Also, the *p*-values for the average number of likes and the average number of comments are greater than 0.05, which means that neither exogenous variable has a significant effect in explaining the model.

After specifying the best ARIMAX models whose R code is shown in Appendix 2, we conducted a post-modeling diagnosis by checking whether each of the models were statistically adequate. Thus, we examined the ACF plot of their residuals, and tested whether the residuals were white noise by applying the Ljung-Box test. A description of these tests follows.

```
Regression with ARIMA(3,0,1) errors


Coefficients:
          ar1      ar2      ar3      ma1  average_likes  average_comments
      -0.5632  -0.5637  -0.4761  -0.5088        -0.0010            -6e-04
s.e.   0.1125   0.0970   0.0929   0.1306         0.0065             7e-03


                   Pr(>|z|)
ar1               5.499e-07 ***
ar2               6.277e-09 ***
ar3               2.980e-07 ***
ma1               9.757e-05 ***
average_likes        0.8734
average_comments     0.9344
```

Table 5. ARIMAX outputs for the small stores

5.2. ACF plots for the residuals of the ARIMAX models and Ljung-Box tests

In this section, we describe the process for ascertaining that no linear relationship exists between the lagged values of the residuals of each of the estimated models. If that is the case, we can conclude that the residuals are random with no information left for extraction. Equivalently, we may infer that the model has been successful in explaining all the variability in the dependent variable by utilizing the variability of the independent variables. A series that shows no autocorrelation is called “white noise” (Hyndman and Athanasopoulos, 2012). We expected white noise residuals for all ARIMAX models that we developed to claim them as good fits. For a white noise series, each autocorrelation is expected to be close to zero. To be more specific, we expect 95 percent of the spikes in the ACF to fall inside the confidence interval (Hyndman and Athanasopoulos, 2012). The Ljung-Box test statistics is computed to examine the null hypothesis of independence in a given time series (Ljung and Box, 1978). This test is sometimes known as “portmanteau” test.

In Figure 26 we show the ACF plot of the residuals of the ARIMAX model for the large stores. The corresponding result of Ljung-Box test is included in Figure 27. Since the *p*-value was 0.2935, we failed to reject the null hypothesis of randomness.

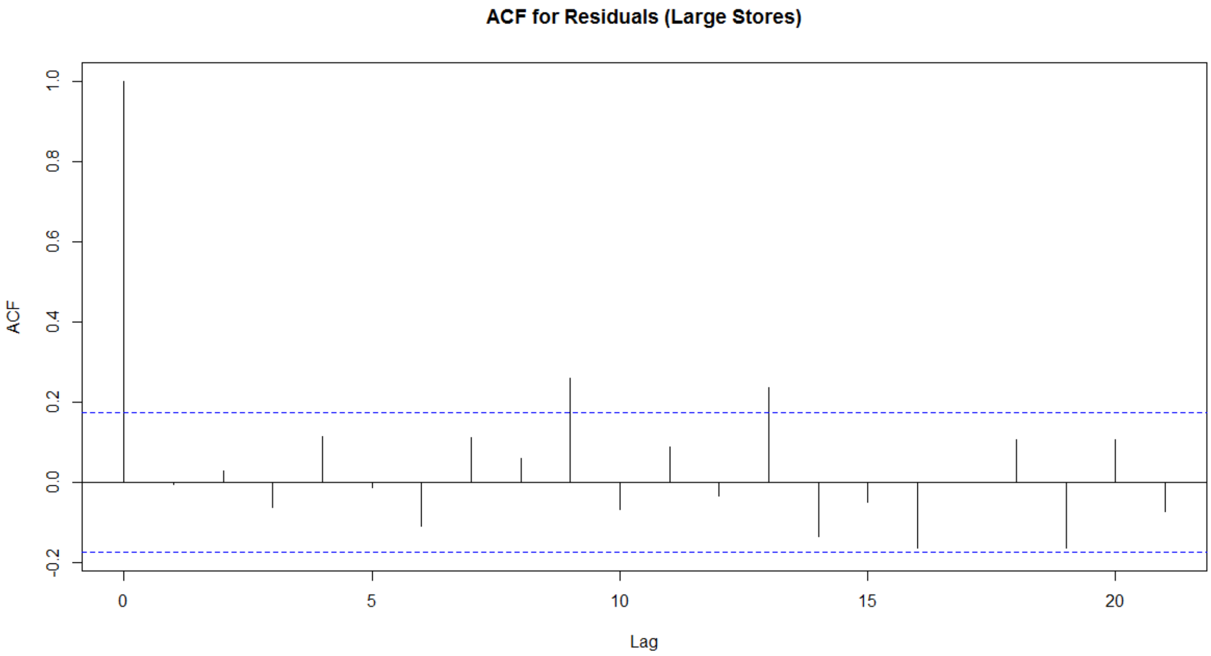
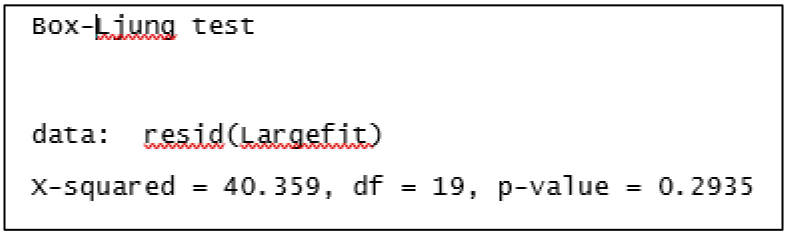


Figure 26. ACF plot of the residuals of the ARIMAX model for the large stores



```
Box-Ljung test


data:  resid(Largefit)
X-squared = 40.359, df = 19, p-value = 0.2935
```

Figure 27. Ljung-Box test result for the large stores

The ACF residual plots of the ARIMAX models for the upper medium, lower medium, and small stores are displayed in Figure 28 to 30, respectively. An analysis of the corresponding *p*-values reveals that we failed to reject the null hypothesis. Thus, we had enough statistical evidence to conclude that all the residuals were random.

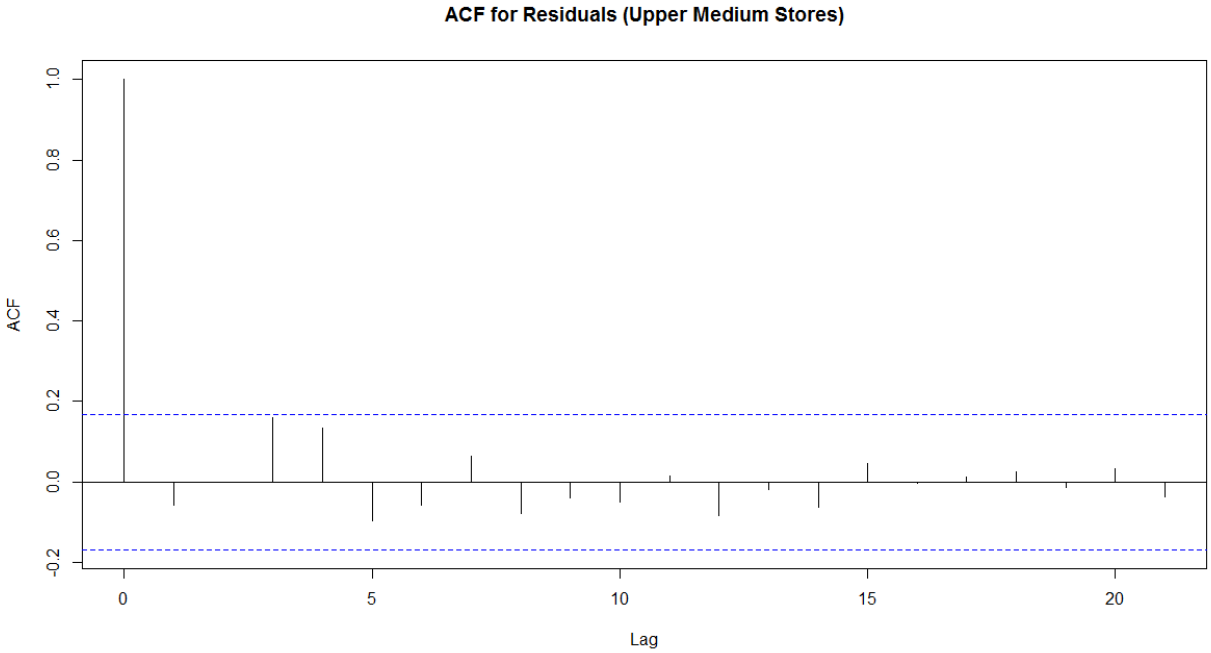


Figure 28. ACF plot of the residuals of the ARIMAX model for the upper medium stores

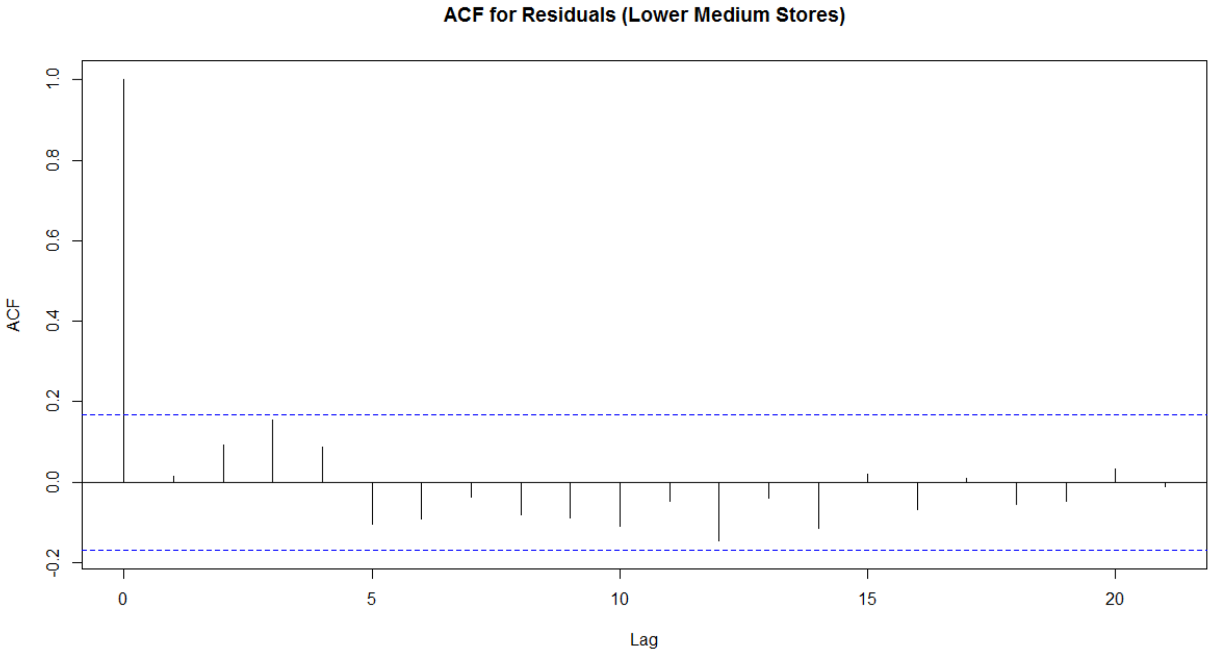


Figure 29. ACF plot of the residuals of the ARIMAX model for the lower medium stores

Because the residuals of the ARIMAX models for all store types were independent and random, we concluded that all ARIMAX models provided an adequate fit to the data.

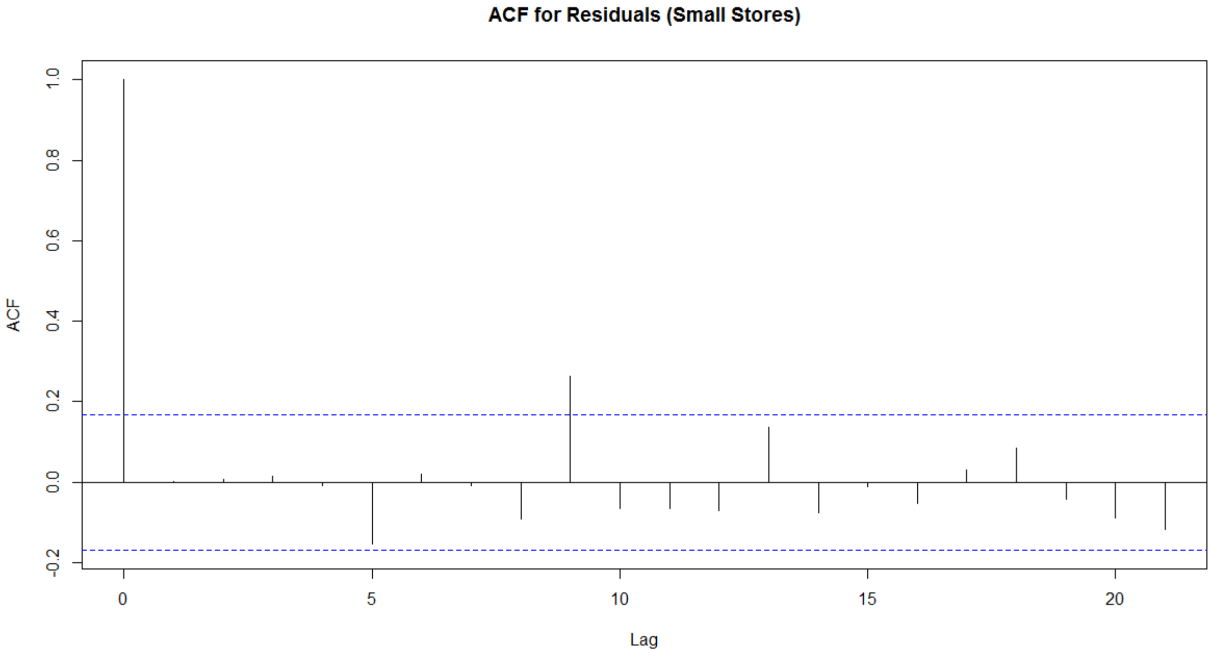


Figure 30. ACF plot of the residuals of the ARIMAX model for the small stores

In the next chapter, we present the interpretation of the results obtained from the application of the aforementioned ARIMAX models.

CHAPTER 6

RESULTS AND CONCLUSION

In this section, we discuss the results that we obtained from our ARIMAX investigation and present the conclusions and business implications of the study. We furnish recommendations for future work as well.

The decision to separate the Walmart stores into four different sizes and perform a size-level study appears to have been justified. The small Walmart stores had completely different patterns than the others; in particular, they had a nonstationary sales trajectory. This meant that the small stores featured a strong growth route. If we lumped all stores together, we would not have been able to capture this difference.

The heterogeneity of store types was also visible from the different ARIMA characteristics they were found to have. The large stores had only AR(1) components; upper medium and lower medium stores had AR(1), MA(1) and MA(2); finally, the small stores had AR components up to level 3, and an MA(1) component. These results appeared to make much sense because the large stores tend to be established, only fluctuate around a basis (i.e., intercept term), depend only on the past step (hence the finding of AR(1)), and are not frequently subject to past shocks carried inside MA components. In contrast, small Walmart stores are usually in the process of growing and may be heavily dependent on their past records (hence the finding of AR(1), AR(2), and AR(3)). Past shocks are also good indicators of today’s sales for them, as the MA(1) finding suggests. We then had the upper medium and lower medium stores, which indeed behave somewhere in between large and small stores. Their sales were explained by AR and MA components. It would be impossible to make this heterogeneity visible if we did not perform a detailed analysis based on store sizes. Much of this wealth of interpretation would be lost in a lumped data set.

We observed very interesting patterns when it came to the effect of social media on sales. The average number of comments did not have a significant impact on sales in any store size level. This was a highly anticipated result, given the blend of positive and negative comments that appeared to cancel each other out. We didn't know if people were saying good or bad things.

We observed a non-linear picture of significance on the effect of the average number of likes on sales based on store sizes. The number of likes had no impact on small stores, a strong significance on lower medium and upper medium stores, and a weakened but significant effect on large stores. We can interpret this as follows. The sales values of small stores have a very stable behavior over time, as seen in Figure 6. It can be inferred that their sales are not affected by any external factor such as social media. Therefore, the number of likes and comments have no power on explaining the future sales of small Walmart stores.

Lower and upper medium Walmart stores usually have started establishing their customer base and are on their way to becoming large stores. These stores are at the point where sales performance may either improve or deteriorate based on customer satisfaction and loyalty. Therefore, the finding of number of likes to be a significant factor on their sales made sense.

Large Walmart stores usually have the best-established customer base. It might be reasonable to think that some of their customers shop from them no matter what, but some are more reactive to how the company runs its stores. We believe this can explain the relatively weaker but still significant impact of social media likes on large store sales.

Our findings should motivate retailers to keep a social media presence and trace their customer’s reactions on their social media page. If analyzed correctly, social media data become a reliable indicator on their customer attitudes toward them. Moreover, the results of this research would seem to encourage retailers to focus on their medium-sized stores the most when forming their policies. Designing the marketing strategy based on how social media receives it can be a smart and inexpensive method of policy design, which can lead to optimized sales and revenue.

We observed a consistent negative sign in front of both AR coefficients and the number of likes. Our interpretation of this somewhat surprising result is that sales follow a cyclic path, where high sales are usually followed by low sales, and vice versa. The negative coefficients of numbers of likes might be the result of a lag effect. Since a log-differenced series of likes subtracts today’s log series from yesterday’s, the negative coefficient might mean that an increase in today’s sales is observed when yesterday’s likes are higher than today’s. When we experimented with lagged values of log-differenced number of likes, we noticed that the coefficient became positive. We do not present an elaboration of this result in this thesis, but it can be explored in future work.

Several extensions to the work presented here are suggested. Firstly, sentiment analysis of comments can bring value to understanding the behavior of the sales time series. Future work can be directed at analyzing the effects of comments and consider not only the quantity but the essence of such comments. In that regard, creating two separate variables as the number of positive comments and negative comments, or a sentiment value for each comment may improve the model. Secondly, a more detailed ARIMAX model with a search through lags of the independent variables and their nonlinear transformations can improve the proposed model. Thirdly, a more granular study in which more than four store size groups can uncover additional hidden patterns. Lastly, the Box-Jenkins methodology employed to model sales using social media should be extended to organizations similar to Walmart and any company who has a social media presence.

BIBLIOGRAPHY

Allon, G., & Zhang, D. J. (2015). Managing Service Systems in the Presence of Social Networks. *Available at SSRN 2673137*.

Asur, S., & Huberman, B.A. (2010). Predicting the Future with Social Media. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 492–499.

Barbera, S. (2016). *How Retailers Use Social Media to Predict Consumer Demand* [Blog Post]. Retrieved from https://www.cgsinc.com/blog/how-retailers-use-social-media-predict-consumer-demand

Beheshti-Kashi, S., Karimi, H.R., Thoben, K.D., Lütjen, M., & Teucke, M. (2015). A Survey on Retail Sales Forecasting and Prediction in Fashion Markets. *Systems Science & Control Engineering*, 3(1), 154-161.

Boone, T., Ganeshan, R., Hicks, R. L., & Sanders, N.R. (2018). Can Google Trends Improve Your Sales Forecast? *Production and Operations Management*. Retrieved from https://doi.org/10.1111/poms.12839

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008). *Time Series Analysis: Forecasting and Control* (4th ed.). Hoboken, NJ: John Wiley & Sons Inc.

Candogan, O., Bimpikis, K., & Ozdaglar, A. (2012). Optimal Pricing in Networks with Externalities. *Operations Research*, 60(4), 883-905.

Chen, Y., Fay, S., & Wang Q. (2011). The Role of Marketing in Social Media: How Online Consumer Reviews Evolve. *Journal of Interactive Marketing*, 25(2), 85-94.

Cui, R., Gallino, S., Moreno, A., & Zhang, D. (2017). The Operational Value of Social Media Information. *Production and Operations Management*. Retrieved from https://doi.org/10.1111/poms.12707

Diebold, F. X. (2007). *Elements of Forecasting* (4th ed.). Mason, OH: Thomson Higher Education

Hameed, B. (2011). *Social Media Usage Exploding amongst Fortune 500 Companies* [Blog Post]. Retrieved from http://www.adweek.com/digital/social-media-usage-exploding-amongst-fortune-500-companies/

Hyndman, R.J. (2010). The ARIMAX model muddle [Blog Post]. Retrieved from https://robjhyndman.com/hyndsight/arimax/

Hyndman. R. J., & Athanasopoulos, G. (2012). *Forecasting: Principles and Practice*. Retrieved from https://www.otexts.org/fpp

Hyndman, R.J., & Khandakar. Y. (2008). Automatic Time Series Forecasting: The Forecast Package for R. *Journal of Statistical Software*, 27(3).

John, L. K., Emrich, O., Gupta, S., & Norton, M. I. (2017). Does "Liking" Lead to Loving? The Impact of Joining a Brand's Social Network on Marketing Outcomes. *Journal of Marketing Research*, 54(1), 144-155.

Kumar, A., Bezawada, R., Rishika, R., Janakiraman, R., & Kannan, P.K. (2016). From Social to Sale: The Effects of Firm-Generated Content in Social Media on Customer Behavior. *Journal of Marketing*, 80(1), 7-25.

Ljung, G. M., & Box, G. E. P. (1978). *On a Measure of a Lack of Fit in Time Series Models*. Biometrika 65(2), 297–303.

Papanastasiou, Y., & Savva, N. (2014). Dynamic Pricing in the Presence of Social Learning and Strategic Consumers, *Working Paper*.

Russell, M. A. (2013). *Mining the Social Web* (2nd ed). Sebastopol, CA: O'Reilly Media Inc.

Stephen, A. T., & Galak, J. (2012). The Effects of Traditional and Social Earned Media on Sales: A Study of a Microlending Marketplace. *Journal of Marketing Research*, 49(5), 624-639.

APPENDICES

APPENDIX 1 – A data file example

Date	store_size	sales	comments_avg	likes_avg
2/5/2010	Large	23444736.14	943	2492
2/12/2010	Large	22273846.64	781	2075
2/19/2010	Large	22474784.83	809	2696
2/26/2010	Large	20234521.78	752	2807
3/5/2010	Large	21568615.3	346	1169
3/12/2010	Large	21271741.54	1338	5330
3/19/2010	Large	20726570.6	1097	1393
3/26/2010	Large	20130605.86	597	2354
4/2/2010	Large	23411624.39	593	2105
4/9/2010	Large	21774496.54	290	642
4/16/2010	Large	20801375.54	4188	5886
4/23/2010	Large	20543443.09	938	5045
4/30/2010	Large	20112594.01	1092	4859
5/7/2010	Large	22451986.27	1226	3819
5/14/2010	Large	20839531.78	1508	6084
5/21/2010	Large	20656799.62	2592	11513
5/28/2010	Large	21670794.45	1475	8127
6/4/2010	Large	23128781.3	2526	5945
6/11/2010	Large	21887438.65	999	6364
6/18/2010	Large	21773694.7	983	5162

APPENDIX 2 – R code for the data preparation and the fitting of the best ARIMAX models

```
# Importing the libraries
library(forecast)
library(lmtest)
library(tseries)

# Declaring the working directory
setwd("C:\\Users\\aybike\\Desktop\\Thesis\\Data\\FinalData\\")

# Declaring the data
large <- read.csv(file="large.csv", header=TRUE, sep=",")
data <- large[, 'sales']

# Removing the Christmas season from the data
data <- large[large[,'sales']<25000000,]
data <- data[data[,'likes_avg']!=0,]

# ADF tests to check stationarity
adf.test(data[,'sales'])
adf.test(data[,'likes_avg'])
adf.test(data[,'comments_avg'])

# Differencing the data to make data stationary on mean (remove trend)
diff(data[,'sales'])
diff(data[,'likes_avg'])
diff(data[,'comments_avg'])

# Log transforming the data to make data stationary on variance
log(data[,'sales'])
log(data[,'likes_avg'])
log(data[,'comments_avg'])

# Differencing and log transforming the data to make data stationary on both mean and variance
diff(log(data[,'sales']))
diff(log(data[,'likes_avg']))
diff(log(data[,'comments_avg']))

# ADF tests to check stationarity after decomposition processes
adf.test(diff(log(data[,'likes_avg'])))
adf.test(diff(log(data[,'comments_avg'])))

# Identification of best fit ARIMAX model (Sales & Log Differencing Likes and Comments)
average_likes = diff(log(data[,'likes_avg']))
average_comments = diff(log(data[,'comments_avg']))
my_y = data[,'sales'][c(2:length(data[,'sales']))]
Largefit <- auto.arima(my_y, xreg=cbind(average_likes,average_comments))
summary(Largefit)
coeftest(Largefit)

# Drawing the ACF plot for residuals of ARIMAX model to ensure no more information is left for extraction
acf(ts(Largefit$residuals),main='ACF Residual')
```

```
# Ljung Box test to check whether the residuals are random and independent
Box.test(resid(Largefit),type="Ljung",lag=20,fitdf=1)
```

APPENDIX 3 – R code for plotting

```
library(forecast)

library(lmtest)
library(tseries)

# AGGREGATE WEEKLY SALES
setwd("C:\\Users\\aybike\\Desktop\\Thesis\\Data\\")
data <- read.csv(file="AggregateWeeklySales.csv", header=TRUE, sep=",")
myDates <- as.Date(as.character(data[,'Date']), format="%m/%d/%Y")

v1 <- as.Date(c("2010-02-05","2011-01-07","2012-01-06", "2012-10-26"))
w1 <- c("40000000","50000000", "65000000", "80000000")

v2 <- c("Feb 2010","Jan 2011","Jan 2012","Oct 2012")
w2 <- c("35","50", "65", "80")

plot(myDates, data[,'sales'], xaxt = "n", yaxt = "n", xlab='Date', ylab='Sales (Millions of Dollars)', main="Weekly
Walmart Sales (All Stores)")
axis(side = 1, at = v1, labels = v2, tck=-.02)
axis(side = 2, at = w1, labels = w2, tck=-.02)

# INITIAL SALES PLOTS (INCLUDING THE CHRISTMAS SEASON)

# LARGE STORES
setwd("C:\\Users\\aybike\\Desktop\\Thesis\\Data\\FinalData\\")
data <- read.csv(file="large.csv", header=TRUE, sep=",")
myDates <- as.Date(as.character(data[,'Date']), format="%m/%d/%Y")

v1 <- as.Date(c("2010-02-05","2011-01-07","2012-01-06", "2012-10-26"))
w1 <- c("20000000","25000000", "30000000", "35000000")

v2 <- c("Feb 2010","Jan 2011","Jan 2012","Oct 2012")
w2 <- c("20","25", "30", "35")

plot(myDates, data[,'sales'], xaxt = "n", yaxt = "n", xlab='Date', ylab='Sales (Millions of Dollars)', main="Weekly
Walmart Sales (Large Stores)")
axis(side = 1, at = v1, labels = v2, tck=-.02)
axis(side = 2, at = w1, labels = w2, tck=-.02)

# UPPER MEDIUM STORES
setwd("C:\\Users\\aybike\\Desktop\\Thesis\\Data\\FinalData\\")
data <- read.csv(file="uppermedium.csv", header=TRUE, sep=",")
myDates <- as.Date(as.character(data[,'Date']), format="%m/%d/%Y")

v1 <- as.Date(c("2010-02-05","2011-01-07","2012-01-06", "2012-10-26"))
w1 <- c("10000000","14000000", "16000000", "20000000")

v2 <- c("Feb 2010","Jan 2011","Jan 2012","Oct 2012")
w2 <- c("10","14", "16", "20")

plot(myDates, data[,'sales'], xaxt = "n", yaxt = "n", xlab='Date', ylab='Sales (Millions of Dollars)', main="Weekly
Walmart Sales (Upper Medium Stores)")
axis(side = 1, at = v1, labels = v2, tck=-.02)
axis(side = 2, at = w1, labels = w2, tck=-.02)

# LOWER MEDIUM STORES
setwd("C:\\Users\\aybike\\Desktop\\Thesis\\Data\\FinalData\\")
data <- read.csv(file="lowermedium.csv", header=TRUE, sep=",")
myDates <- as.Date(as.character(data[,'Date']), format="%m/%d/%Y")

v1 <- as.Date(c("2010-02-05","2011-01-07","2012-01-06", "2012-10-26"))
w1 <- c("7000000","10000000", "14000000", "17000000")

v2 <- c("Feb 2010","Jan 2011","Jan 2012","Oct 2012")
w2 <- c("7","10", "14", "17")
plot(myDates, data[,'sales'], xaxt = "n", yaxt = "n", xlab='Date', ylab='Sales (Millions of Dollars)', main="Weekly
Walmart Sales (Lower Medium Stores)")

axis(side = 1, at = v1, labels = v2, tck=-.02)
axis(side = 2, at = w1, labels = w2, tck=-.02)

# SMALL STORES
setwd("C:\\Users\\aybike\\Desktop\\Thesis\\Data\\FinalData\\")
data <- read.csv(file="small.csv", header=TRUE, sep=",")
myDates <- as.Date(as.character(data[,'Date']), format="%m/%d/%Y")

v1 <- as.Date(c("2010-02-05","2011-01-07","2012-01-06", "2012-10-26"))
w1 <- c("4000000","4500000")

v2 <- c("Feb 2010","Jan 2011","Jan 2012","Oct 2012")
w2 <- c("4","4.5")

plot(myDates, data[,'sales'], xaxt = "n", yaxt = "n", xlab='Date', ylab='Sales (Millions of Dollars)', main="Weekly
Walmart Sales (Small Stores)")
axis(side = 1, at = v1, labels = v2, tck=-.02)
axis(side = 2, at = w1, labels = w2, tck=-.02)

# AVERAGE NUMBER OF LIKES
setwd("C:\\Users\\aybike\\Desktop\\Thesis\\Data\\FinalData\\")
data <- read.csv(file="small.csv", header=TRUE, sep=",")
myDates <- as.Date(as.character(data[,'Date']), format="%m/%d/%Y")

v1 <- as.Date(c("2010-02-05","2011-01-07","2012-01-06", "2012-10-26"))
w1 <- c("100000","250000", "500000")

v2 <- c("Feb 2010","Jan 2011","Jan 2012","Oct 2012")
w2 <- c("100","250","500")

plot(myDates, data[,'likes_avg'], xaxt = "n", yaxt = "n", xlab='Date', ylab='Number of Likes (in Thousands)',
main="Average Weekly Number of Likes")

axis(side = 1, at = v1, labels = v2, tck=-.02)
axis(side = 2, at = w1, labels = w2, tck=-.02)

# AVERAGE NUMBER OF COMMENTS
setwd("C:\\Users\\aybike\\Desktop\\Thesis\\Data\\FinalData\\")
data <- read.csv(file="small.csv", header=TRUE, sep=",")
myDates <- as.Date(as.character(data[,'Date']), format="%m/%d/%Y"
v1 <- as.Date(c("2010-02-05","2011-01-07","2012-01-06", "2012-10-26"))
w1 <- c("10000","20000", "30000")

v2 <- c("Feb 2010","Jan 2011","Jan 2012","Oct 2012")
w2 <- c("10","20","30")

plot(myDates, data[,'comments_avg'], xaxt = "n", yaxt = "n", xlab='Date', ylab='Number of Comments (in
Thousands)', main="Average Weekly Number of Comments")
axis(side = 1, at = v1, labels = v2, tck=-.02)
axis(side = 2, at = w1, labels = w2, tck=-.02)

# PLOTS FOR THE SERIES AFTER DECOMPOSITION PROCESSES (CHRISTMAS SEASON REMOVED)

# SMALL STORE SALES - AFTER DIFFERENCING
setwd("C:\\Users\\aybike\\Desktop\\Thesis\\Data\\FinalData\\")
small <- read.csv(file="small.csv", header=TRUE, sep=",")
data <- small[small[,'sales']<25000000,]
data <- data[data[,'likes_avg']!=0,]

myDates <- as.Date(as.character(data[,'Date']), format="%m/%d/%Y")
myDates = myDates[c(2:length(data[,'sales']))]

v1 <- as.Date(c("2010-02-05","2011-01-07","2012-01-06", "2012-10-26"))
w1 <- c("100000","700000")

v2 <- c("Feb 2010","Jan 2011","Jan 2012","Oct 2012")
w2 <- c("100","700")

plot(myDates, diff(data[,'sales']), xaxt = "n", yaxt = "n", xlab='Date', ylab='Sales (Thousands of Dollars)',
main="Weekly Walmart Sales (Small Stores)")
axis(side = 1, at = v1, labels = v2, tck=-.02)
axis(side = 2, at = w1, labels = w2, tck=-.02)

# AVERAGE NUMBER OF LIKES - AFTER DIFFERENCING
setwd("C:\\Users\\aybike\\Desktop\\Thesis\\Data\\FinalData\\")
small <- read.csv(file="small.csv", header=TRUE, sep=",")
data <- small[small[,'sales']<25000000,]
data <- data[data[,'likes_avg']!=0,]

myDates <- as.Date(as.character(data[,'Date']), format="%m/%d/%Y")
myDates = myDates[c(2:length(data[,'likes_avg']))]

v1 <- as.Date(c("2010-02-05","2011-01-07","2012-01-06", "2012-10-26"))
w1 <- c("50000","250000")

v2 <- c("Feb 2010","Jan 2011","Jan 2012","Oct 2012")
w2 <- c("50","250")

plot(myDates, diff(data[,'likes_avg']), xaxt = "n", yaxt = "n", xlab='Date', ylab='Number of Likes (in Thousands)',
main="Average Weekly Number of Likes")
axis(side = 1, at = v1, labels = v2, tck=-.02)
axis(side = 2, at = w1, labels = w2, tck=-.02)

# AVERAGE NUMBER OF COMMENTS - AFTER DIFFERENCING
setwd("C:\\Users\\aybike\\Desktop\\Thesis\\Data\\FinalData\\")
```

```
small <- read.csv(file="small.csv", header=TRUE, sep=",")
data <- small[small[,'sales']<25000000,]
data <- data[data[,'likes_avg']!=0,]

myDates <- as.Date(as.character(data[,'Date']), format="%m/%d/%Y")
myDates = myDates[c(2:length(data[,'comments_avg']))]

v1 <- as.Date(c("2010-02-05","2011-01-07","2012-01-06", "2012-10-26"))
w1 <- c("5000","25000")

v2 <- c("Feb 2010","Jan 2011","Jan 2012","Oct 2012")
w2 <- c("5","25")

plot(myDates, diff(data[,'comments_avg']), xaxt = "n", yaxt = "n", xlab='Date', ylab='Number of Comments (in Thousands)', main="Average Weekly Number of Comments")
axis(side = 1, at = v1, labels = v2, tck=-.02)
axis(side = 2, at = w1, labels = w2, tck=-.02)

# SMALL STORE SALES - AFTER LOG TRANSFORM
setwd("C:\\Users\\aybike\\Desktop\\Thesis\\Data\\FinalData\\")
small <- read.csv(file="small.csv", header=TRUE, sep=",")
data <- small[small[,'sales']<25000000,]
data <- data[data[,'likes_avg']!=0,]

myDates <- as.Date(as.character(data[,'Date']), format="%m/%d/%Y")

v1 <- as.Date(c("2010-02-05","2011-01-07","2012-01-06", "2012-10-26"))
w1 <- c("15.1", "15.2", "15.3", "15.4")

v2 <- c("Feb 2010","Jan 2011","Jan 2012","Oct 2012")
w2 <- c("15.1", "15.2", "15.3", "15.4")

plot(myDates, log(data[,'sales']), xaxt = "n", yaxt = "n", ylim=c(15.1,15.4), xlab='Date', ylab='Sales', main="Weekly Walmart Sales (Small Stores)")
axis(side = 1, at = v1, labels = v2, tck=-.02)
axis(side = 2, at = w1, labels = w2, tck=-.02)

# AVERAGE NUMBER OF LIKES - AFTER LOG TRANSFORM
setwd("C:\\Users\\aybike\\Desktop\\Thesis\\Data\\FinalData\\")
small <- read.csv(file="small.csv", header=TRUE, sep=",")
data <- small[small[,'sales']<25000000,]
data <- data[data[,'likes_avg']!=0,]

myDates <- as.Date(as.character(data[,'Date']), format="%m/%d/%Y")

v1 <- as.Date(c("2010-02-05","2011-01-07","2012-01-06", "2012-10-26"))
w1 <- c("5","10","13")

v2 <- c("Feb 2010","Jan 2011","Jan 2012","Oct 2012")
w2 <- c("5","10","13")

plot(myDates, log(data[,'likes_avg']), xaxt = "n", yaxt = "n", xlab='Date', ylab='Number of Likes', main="Average Weekly Number of Likes")
axis(side = 1, at = v1, labels = v2, tck=-.02)
axis(side = 2, at = w1, labels = w2, tck=-.02)
```

```
# AVERAGE NUMBER OF COMMENTS - AFTER LOG TRANSFORM
setwd("C:\\Users\\aybike\\Desktop\\Thesis\\Data\\FinalData\\")
small <- read.csv(file="small.csv", header=TRUE, sep=",")
data <- small[small[,'sales']<25000000,]
data <- data[data[,'likes_avg']!=0,]

myDates <- as.Date(as.character(data[,'Date']), format="%m/%d/%Y")

v1 <- as.Date(c("2010-02-05","2011-01-07","2012-01-06", "2012-10-26"))
w1 <- c("4","7","10")

v2 <- c("Feb 2010","Jan 2011","Jan 2012","Oct 2012")
w2 <- c("4","7","10")

plot(myDates, log(data[,'comments_avg']), xaxt = "n", yaxt = "n", xlab='Date', ylab='Number of Comments',
main="Average Weekly Number of Comments")
axis(side = 1, at = v1, labels = v2, tck=-.02)
axis(side = 2, at = w1, labels = w2, tck=-.02)

# SMALL STORE SALES - AFTER DIFFERENCING AND LOG TRANSFORM
setwd("C:\\Users\\aybike\\Desktop\\Thesis\\Data\\FinalData\\")
small <- read.csv(file="small.csv", header=TRUE, sep=",")
data <- small[small[,'sales']<25000000,]
data <- data[data[,'likes_avg']!=0,]

myDates <- as.Date(as.character(data[,'Date']), format="%m/%d/%Y")
myDates = myDates[c(2:length(data[,'sales']))]

v1 <- as.Date(c("2010-02-05","2011-01-07","2012-01-06", "2012-10-26"))
w1 <- c("0","0.5")

v2 <- c("Feb 2010","Jan 2011","Jan 2012","Oct 2012")
w2 <- c("0","0.5")

plot(myDates, diff(log(data[,'sales'])), xaxt = "n", yaxt = "n", xlab='Date', ylab='Sales', main="Weekly Walmart
Sales (Small Stores)")
axis(side = 1, at = v1, labels = v2, tck=-.02)
axis(side = 2, at = w1, labels = w2, tck=-.02)

# AVERAGE NUMBER OF LIKES - AFTER DIFFERENCING AND LOG TRANSFORM
setwd("C:\\Users\\aybike\\Desktop\\Thesis\\Data\\FinalData\\")
small <- read.csv(file="small.csv", header=TRUE, sep=",")
data <- small[small[,'sales']<25000000,]
data <- data[data[,'likes_avg']!=0,]

myDates <- as.Date(as.character(data[,'Date']), format="%m/%d/%Y")
myDates = myDates[c(2:length(data[,'likes_avg']))]

v1 <- as.Date(c("2010-02-05","2011-01-07","2012-01-06", "2012-10-26"))
w1 <- c("0","1")

v2 <- c("Feb 2010","Jan 2011","Jan 2012","Oct 2012")
w2 <- c("0","1")
plot(myDates, diff(log(data[,'likes_avg'])), xaxt = "n", yaxt = "n", xlab='Date', ylab='Number of Likes',
main="Average Weekly Number of Likes")
axis(side = 1, at = v1, labels = v2, tck=-.02)
axis(side = 2, at = w1, labels = w2, tck=-.02)

# AVERAGE NUMBER OF COMMENTS - AFTER DIFFERENCING AND LOG TRANSFORM
setwd("C:\\Users\\aybike\\Desktop\\Thesis\\Data\\FinalData\\")
small <- read.csv(file="small.csv", header=TRUE, sep=",")
data <- small[small[,'sales']<25000000,]
data <- data[data[,'likes_avg']!=0,]

myDates <- as.Date(as.character(data[,'Date']), format="%m/%d/%Y")
myDates = myDates[c(2:length(data[,'comments_avg']))]

v1 <- as.Date(c("2010-02-05","2011-01-07","2012-01-06", "2012-10-26"))
w1 <- c("0","1")

v2 <- c("Feb 2010","Jan 2011","Jan 2012","Oct 2012")
w2 <- c("0","1")

plot(myDates, diff(log(data[,'comments_avg'])), xaxt = "n", yaxt = "n", xlab='Date', ylab='Number of Comments',
main="Average Weekly Number of Comments")
axis(side = 1, at = v1, labels = v2, tck=-.02)
axis(side = 2, at = w1, labels = w2, tck=-.02)

# ACF PLOTS FOR RESIDUALS AFTER ARIMAX

# LARGE STORES
setwd("C:\\Users\\aybike\\Desktop\\Thesis\\Data\\FinalData\\")
large <- read.csv(file="large.csv", header=TRUE, sep=",")
data <- large[, 'sales']
data <- large[large[,'sales']<25000000,]
data <- data[data[,'likes_avg']!=0,]

average_likes = diff(log(data[,'likes_avg']))
average_comments = diff(log(data[,'comments_avg']))
my_y = data[,'sales'][c(2:length(data[,'sales']))]
Largefit <- auto.arima(my_y, xreg=cbind(average_likes,average_comments))

acf(ts(Largefit$residuals),main='ACF for Residuals (Large Stores)')

# UPPER MEDIUM STORES
setwd("C:\\Users\\aybike\\Desktop\\Thesis\\Data\\FinalData\\")
uppermedium <- read.csv(file="uppermedium.csv", header=TRUE, sep=",")
data <- uppermedium[uppermedium[,'sales']<25000000,]
data <- data[data[,'likes_avg']!=0,]

average_likes = diff(log(data[,'likes_avg']))
average_comments = diff(log(data[,'comments_avg']))
my_y2 = data[,'sales'][c(2:length(data[,'sales']))]
Upperfit <- auto.arima(my_y2, xreg=cbind(average_likes,average_comments))

acf(ts(Upperfit$residuals),main='ACF for Residuals (Upper Medium Stores)')

# LOWER MEDIUM STORES
setwd("C:\\Users\\aybike\\Desktop\\Thesis\\Data\\FinalData\\")
```

```
lowermedium <- read.csv(file="lowermedium.csv", header=TRUE, sep=",")
data <- lowermedium[lowermedium[,'sales']<25000000,]
data <- data[data[,'likes_avg']!=0,]

average_likes = diff(log(data[,'likes_avg']))
average_comments = diff(log(data[,'comments_avg']))
my_y3 = data[,'sales'][c(2:length(data[,'sales']))]
Lowerfit <- auto.arima(my_y3, xreg=cbind(average_likes,average_comments))

acf(ts(Lowerfit$residuals),main='ACF for Residuals (Lower Medium Stores)')

# SMALL STORES
setwd("C:\\Users\\aybike\\Desktop\\Thesis\\Data\\FinalData\\")
small <- read.csv(file="small.csv", header=TRUE, sep=",")
data <- small[small[,'sales']<25000000,]
data <- data[data[,'likes_avg']!=0,]

average_likes = diff(log(data[,'likes_avg']))
average_comments = diff(log(data[,'comments_avg']))
my_y4 = diff(log(data[,'sales']))
Smallfit <- auto.arima(my_y4, xreg=cbind(average_likes,average_comments))

acf(ts(Smallfit$residuals),main='ACF for Residuals (Small Stores)')
```

VITA

CENSUS: Aybike Akici was born on March 22, 1988, in Cine, Turkey. Her parents are Sultan and Meric Ozdemirel. She is married to Fatih Akici with two children, Gül Şirin and Korkut Alp.

TRAINING: Aybike Akici got her B.S. degree in Logistics Management from Izmir University of Economics, Izmir, Turkey in 2011. She received her M.S degree in Intelligent Engineering Systems from the same university in 2014.

EXPERIENCE: From 2011 to 2014, she was employed at Izmir University of Economics, Izmir, Turkey, as a teaching and research assistant. She served in the same position at St. Mary’s University, San Antonio, Texas from 2015 to 2017.

ADDRESS: 1202 Hidden Ridge Apt 2009
Irving, TX 75038