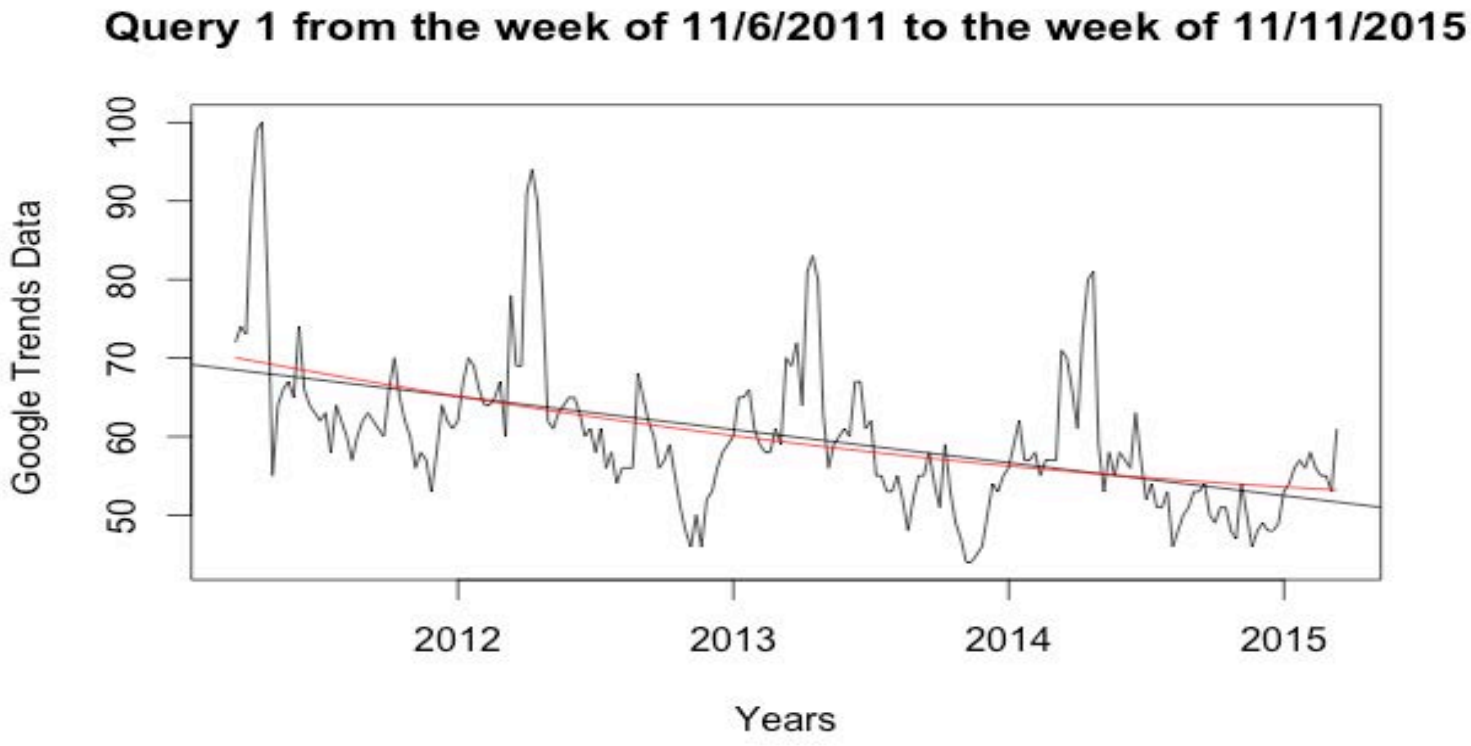
Midterm Two

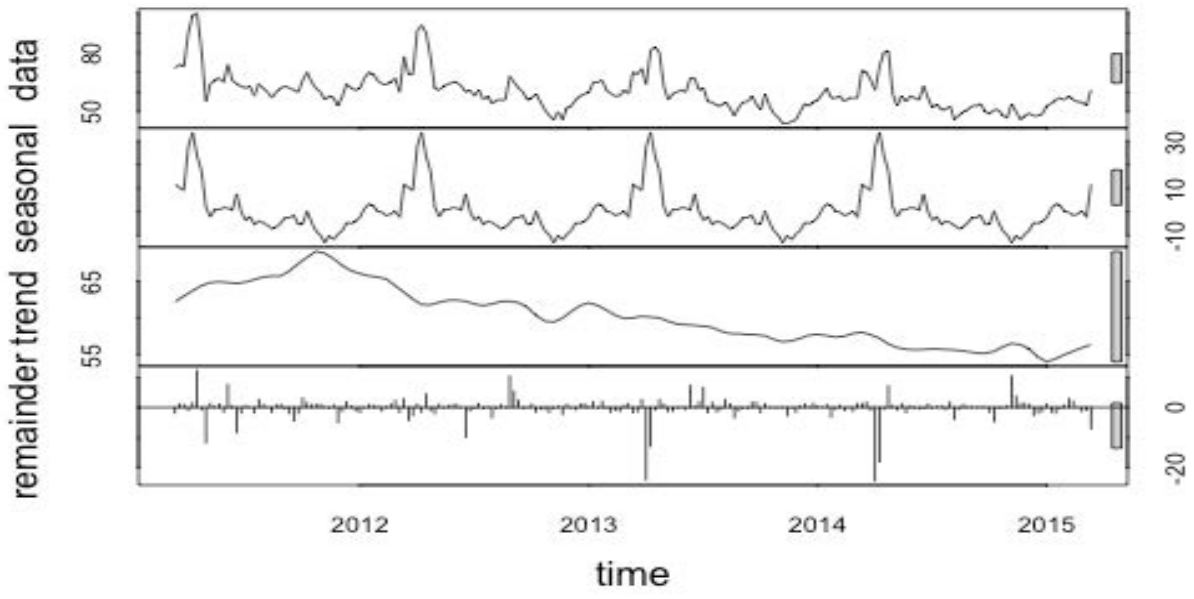
Ryan Chui

Nov 16, 2016

** Exploratory Analysis **



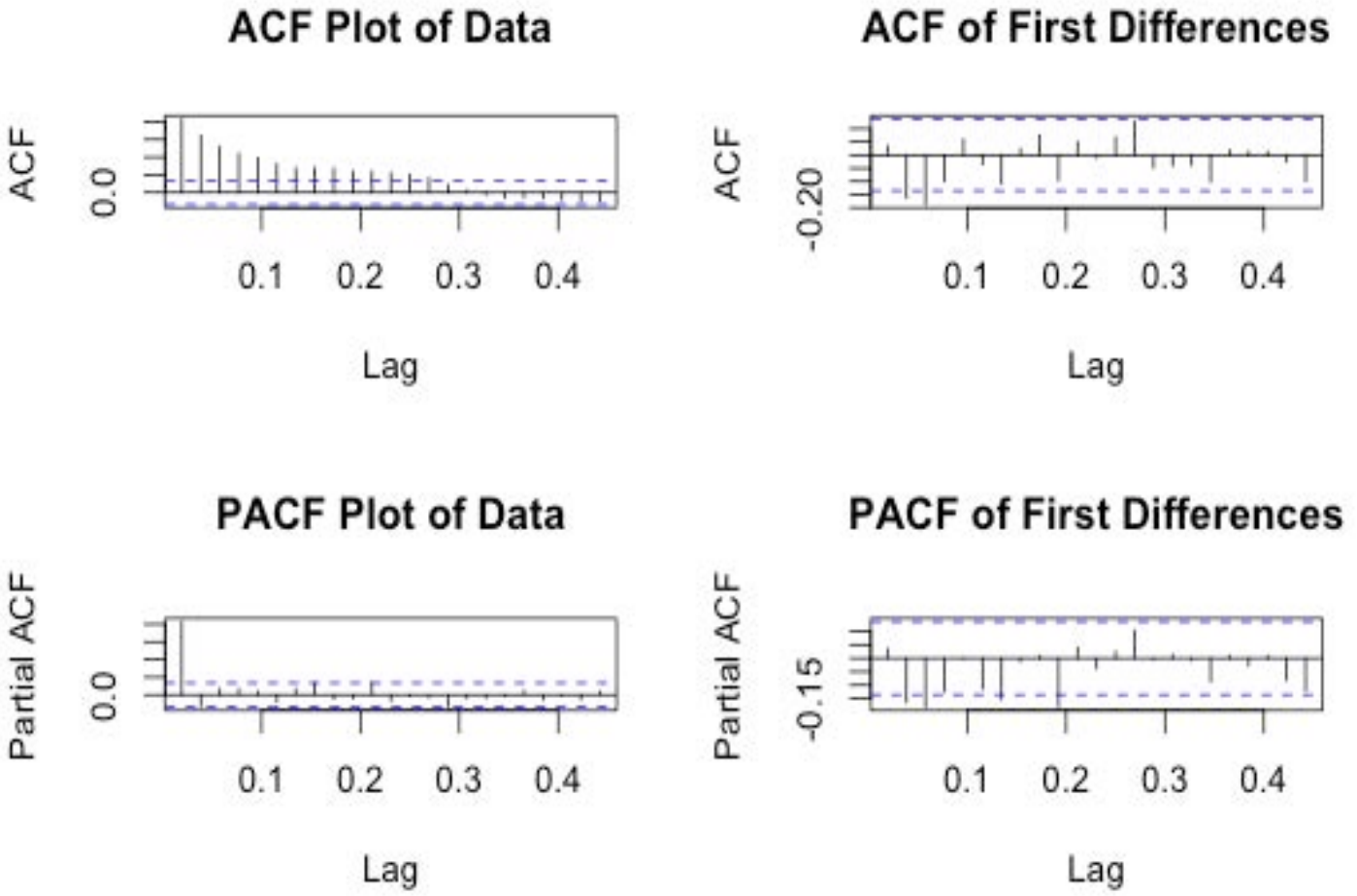
The plot in above is the Google Trend Data from the week of November 06, 2011 to the week of November 11, 2015. The fit line of linear and quadratical line looks identical.



The four graphs are the original data, seasonal component, trend component and the remainder component. The periodic seasonal pattern is extracted out from the original data and the trend moves around between 54 and 66. In the seasonal component, we can see a regular up and down of seasonal pattern of data. The change in trend is less than the variation doing to the weekly variation. In general, the remainder can explain most of the variance in the data and tell if there is an apparent pattern to the white noise. If we take a series with seasonality, the relative variance of the seasonal component is much more relevant. Note that the window controls the wiliness of the trend component.

Regression Analysis

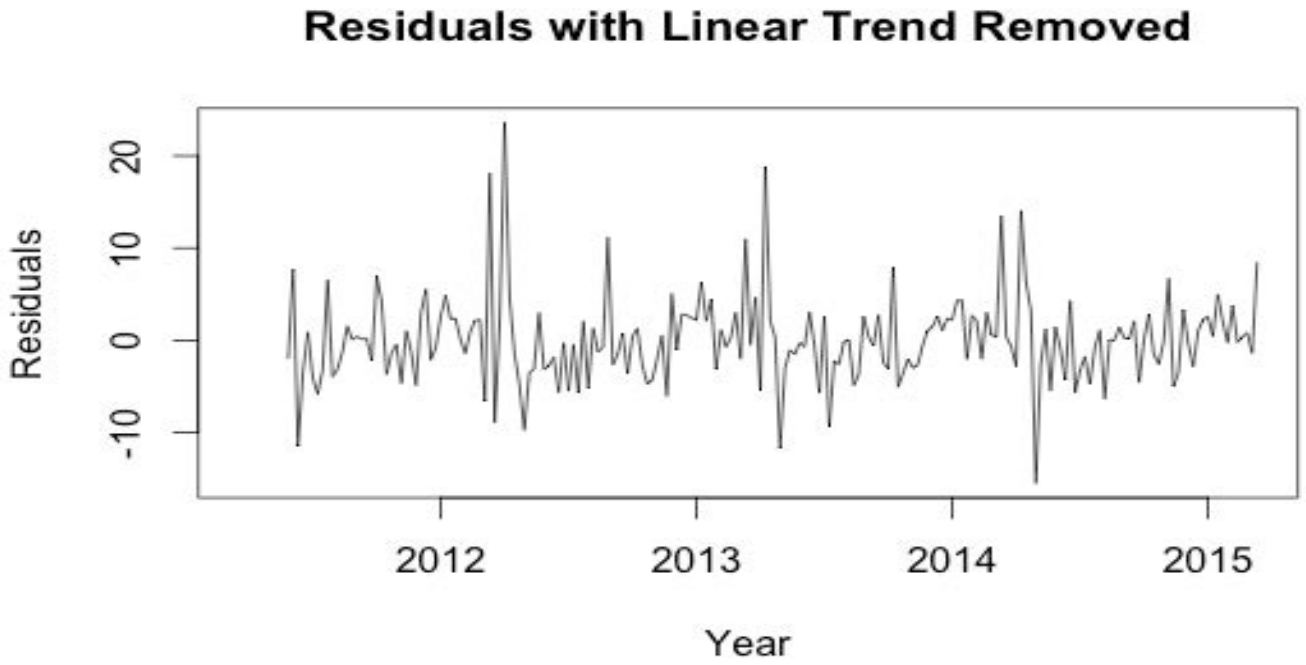
Autocorrelation is used commonly to determine if the time series is stationary or not. A stationary time series will have the autocorrelation fall to zero fairly quickly, but for a non-stationary series it drops gradually. A Partial Autocorrelation is the correlation of the time series with a lag of itself, with the linear dependence of all the lags between them removed.

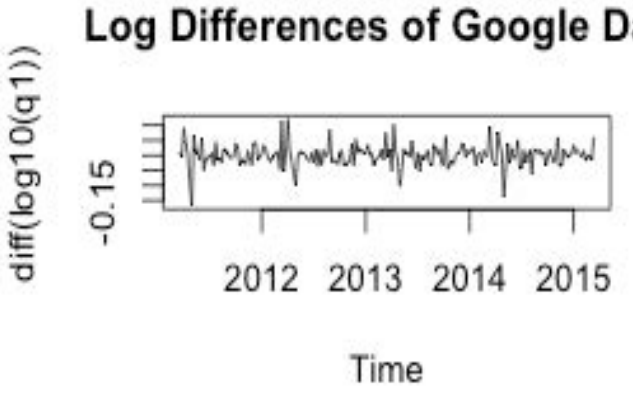


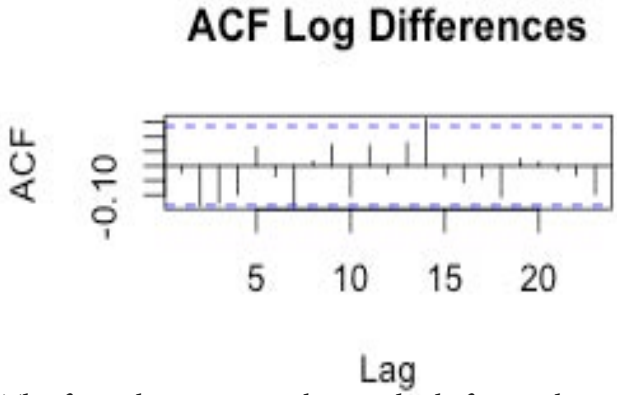
From the above plots, The blue line is confidence intervals where anything outside the boundaries is a significant relationship. If the little bar lies within the blue line, it would mean it is not significantly different from zero or not significantly different from one being uncorrelated, so we can see it's correlated overtime. If we look at the ACF plot of the data, we see a slow decaying function; this is an indication of non-stationary series as the lag remains above the significance range, suggesting that the data follow a long-memory process. If we look at the difference from time periods of ACF plot, it is truly random and has equally likely chance of moving up or down, indicating this is a random walk. This model of the first differences of ACF with significant correlations at lags 2 and 3, but non-significant autocorrelation for higher lags. To make the series stationary, we want to remove the upward trend through 1st order differencing of the series. Therefore, we note a difference between the autocorrelation and the autocorrelation of the difference from one period to another. Our goal is to verify if there is a random walk pattern, and we need to make the series stationary on variance to produce reliable forecasts through ARIMA models.

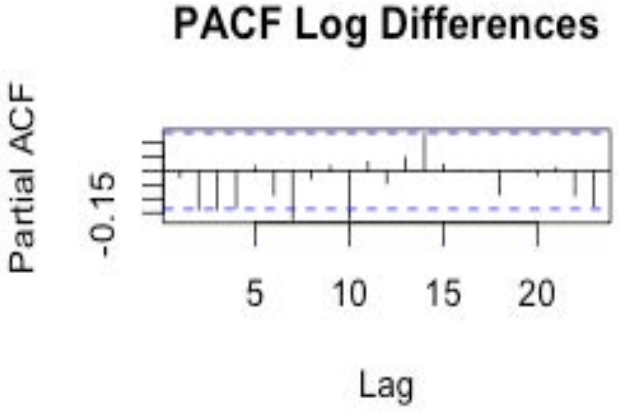
If we look at the PACF of this model, we have a very strong first lag, but the correlation and some of the lags are not too significant. If we look at the difference of the PACF variable, we can see that the lags are significant because they are outside of the confidence interval, indicated by the blue line. When we compare the partial autocorrelation to the autocorrelation of this model, we can see that the autocorrelation shows more an ongoing effect than the PACF. At lag 1, the autocorrelation is the same as the PACF, while the first 4 lags of ACF is very similar the first 4 lags of PACF.

Let's look at the residual plots and consider the first differences of logs and second differences.

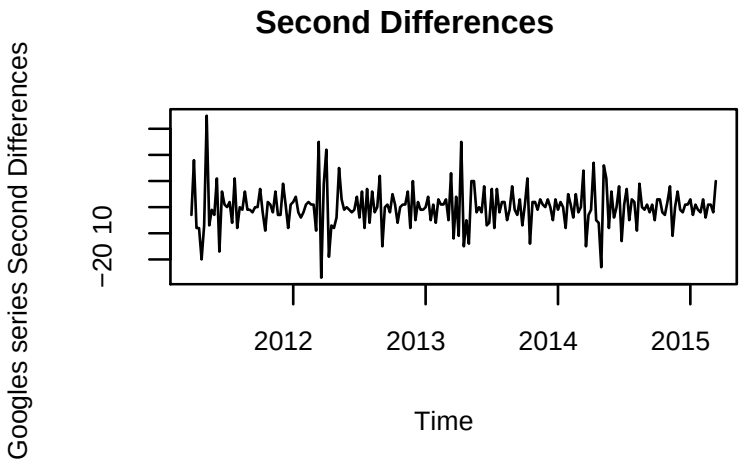


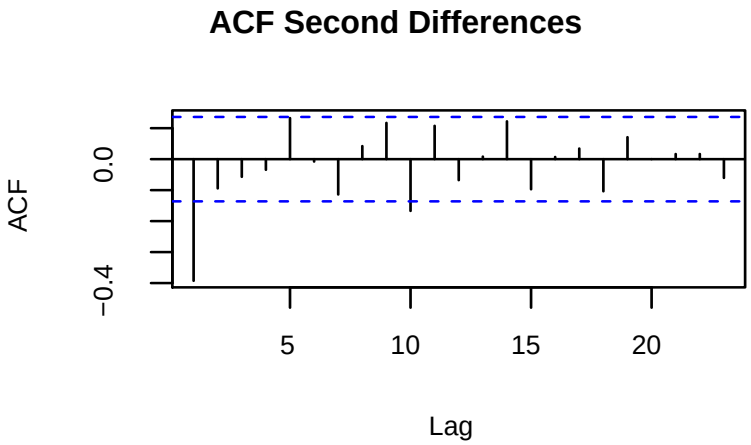


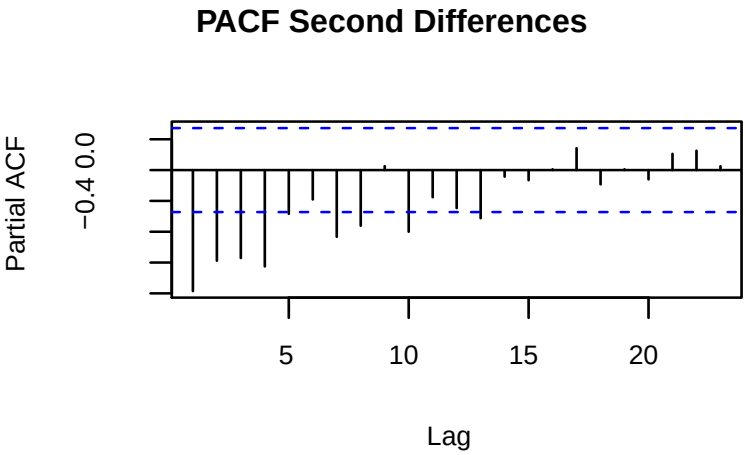




The first plot contains the residuals from a linear fit to the original data. After removing the linear trend, the run sequence plot indicates that the data have a constant location and variance, and the pattern of the residuals shows that the data depart from the model in a fairly systematic way. An ACF and PACF plot of first differenced natural log of series is shown above. Since there are not enough spikes in the plots outside the insignificant zone (dotted horizontal lines), we can conclude that the residuals are random. This implies that there is not enough information available in residuals to be extracted by AR and MA models.



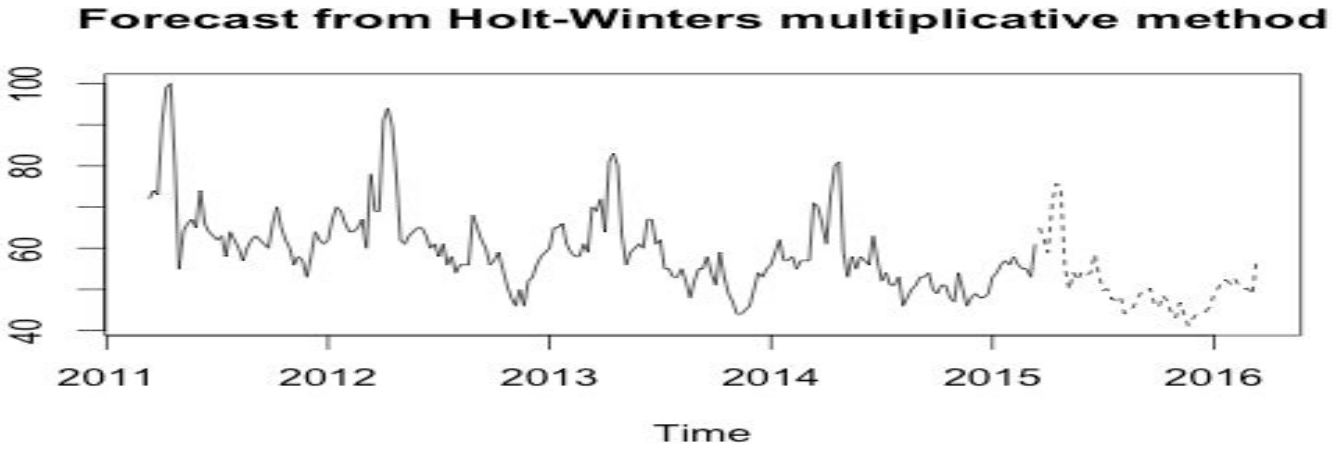


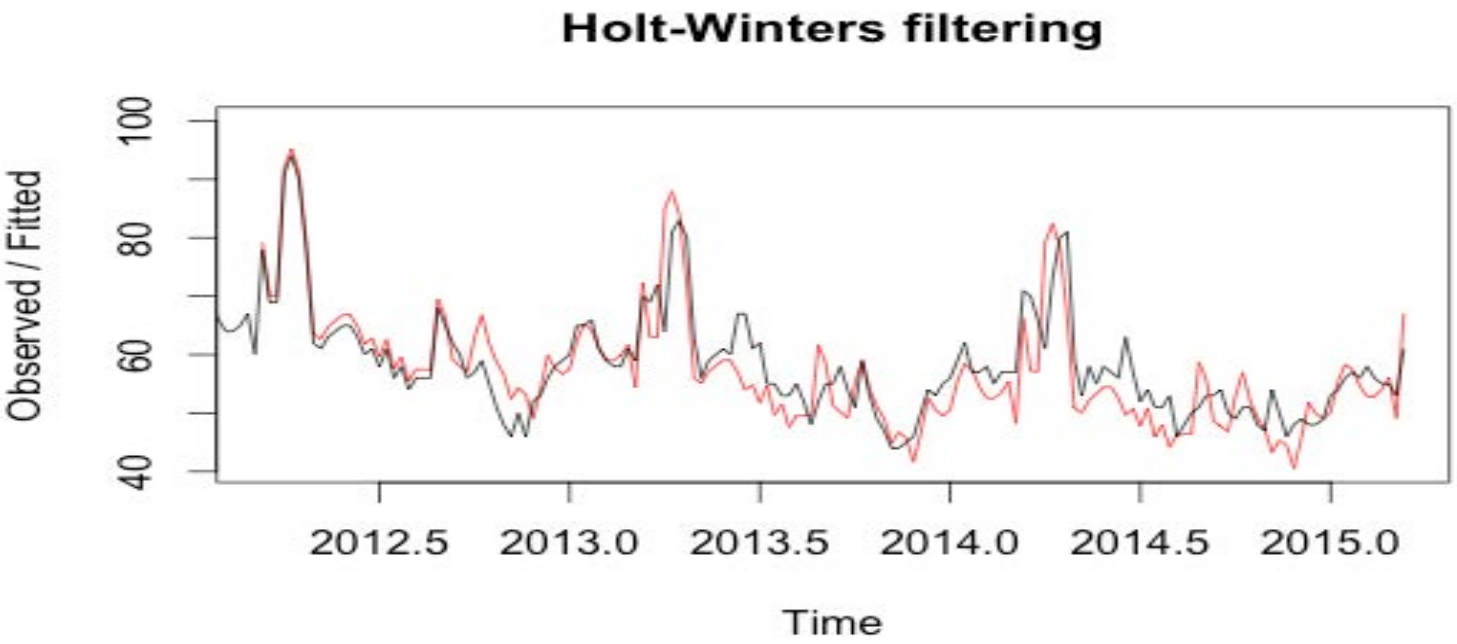


The first plot contains the residuals from a linear fit to the original data. After removing the linear trend, the run sequence plot indicates that the data have a constant location and variance, and the pattern of the residuals shows that the data depart from the model in a fairly systematic way. An ACF and PACF plot of first differenced natural log of series is shown above. Since there are not enough spikes in the plots outside the insignificant zone (dotted horizontal lines), we can conclude that the residuals are random. This implies that there is not
enough information available in residuals to be extracted by AR and MA models.

Notice that differencing once removed linear trends and differencing twice removes quadratic trends, and so on with higher order of polynomials. Let's consider differencing twice and analysis them. In the above plot, the time series of the second differences does appear to be stationary in mean and variance, as the level and variances of the series stays roughly constant over time. We need to difference the time series of the diameter twice in order to achieve a stationary series. As the lags below the significance range in ACF second differences, the correlation is at least significant and should be theoretically zero. In addition, the variation in volatility of differencing twice is larger for the series with logs and first difference. Thus, differencing twice can be interpreted as a means for stabilizing the variance, but first differences of logs could still show a considerable variation in the volatility

Now, let's look at the underlying patterns in temporal data to use in more sophisticated analysis like Holt-Winters seasonal method or ARIMA. We can do predictions using the predict function. The predict function takes q1 as an argument using the Holt-Winter model and we're predicting forward by 52 periods.





The plot above is a prediction which essentially takes the latest wave cycle, keeps repeating it along with the projected downward trend as the time series goes forward. After we plot the data, the black line is the actual data and the red line is the Holt-Winters overlay or fitting of the data. We can see that the Holt Winters model fit fairly close to the actual data partly because this has a more regular seasonal component and a larger seasonal component. We can also tweak parameters if we adjust alpha, beta and gamma to yield different fit.

Forecast data points using the best fit ARIMA model

The next step is to predict tractor sales for next 52 weeks years i.e. for 210 to 262 through the above model. The following R code does this job for us.

```
arima1 = auto.arima(q1, trace = TRUE, test = "kpss", ic = "aic")
summary(arima1)
# Series: q1
# ARIMA(1,1,1)(0,0,1)[52]

# Coefficients:

#         ar1         sma1
#      -0.1049  0.5826
# s.e.  0.0722   0.0974

# sigma^2 estimated as 0.004792:  log likelihood=251.5
# AIC=1264.43  AICc=1264.54    BIC=1274.44

# Training set error measures:
#                       ME      RMSE       MAE       MPE     MAPE      MASE
# Training set -0.008804156 0.06855808 0.05038499 -0.2430769 1.230855 0.6940898
#                     ACF1
# Training set -0.03520171
# confint(arima1)
# 2.5 %     97.5 %
# ar1    -0.2464474 0.03657126
# sma1 0.3916366 0.77347973
```

The system implemented the following models recall in minimum AIC criterion. Based upon the AIC criterion, we see ARIMA(1,1,0)(0,0,1)[52] is the best model. We see that there is a autoregressive parameter and a difference component, but there is no moving average. We can also see that there is no seasonal autoregressive component and no seasonal difference component, but there is one moving average seasonality. In the summary, we can see the values of AR1, seasonal moving average and its standard error. Notice that if we divide the coefficient with the standard error, we will get the t-value. If the T-value is greater than 3, the coefficient is significant or the coefficient is significant far away from general. We can also see the T-values and their interpretation in ARIMA model in quest of a better model. Let's construct a 95% confidence interval, we can see that there is 95% probability that AR1 ranges between -0.25 and 0.036. Since zero is within this interval, we can conclude that the AR1 component is insignificant.

We assume that the residuals are white noise (uncorrelated, mean zero, constant variance). If they aren't, then there is information left in the residuals that should be used in computing forecasts. So a standard residual diagnostic is to check the ACF of the residuals of a forecasting method. We expect these to look like white noise. One way to check is to do some residual diagnostic and look at the plots.
Box.test(arima1$residuals, lag = 52, type = "Ljung-Box")

Box-Ljung test

data: arima1$residuals

X-squared = 66.613, df = 52, p-value = 0.08362
Box.test(arima1$residuals^2, lag = 52, type = "Ljung-Box")

Box-Ljung test

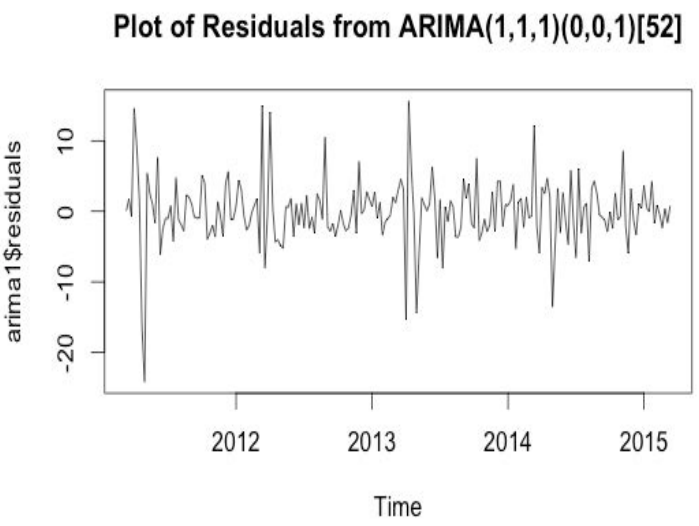
data: arima1$residuals^2

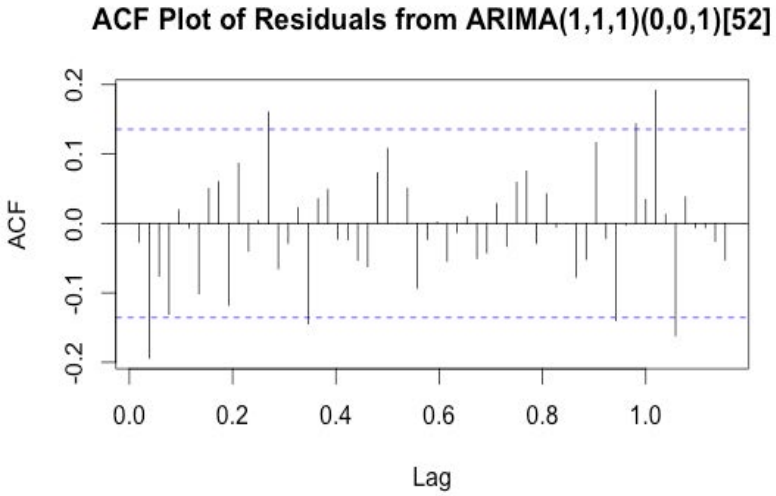
X-squared = 70.438, df = 52, p-value = 0.04517
jarque.bera.test(arima1$residuals)

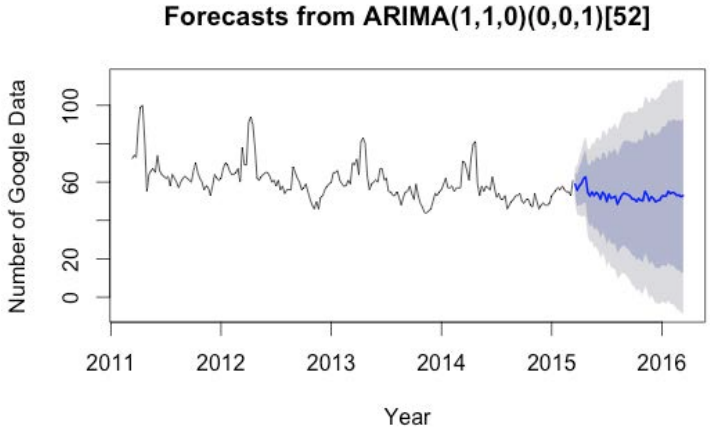
Jarque Bera Test

data: arima1$residuals

X-squared = 228.41, df = 2, p-value < 2.2e-16







As you can see in above, I take the residual from the ARIMA and plot it as a time series. We can visually see that it is somehow stationary in zero but it is not a good estimate or say anything from the figure in general. In the residual plot, we can clearly see a few spikes Thus, I test the Ljung-Box test for the autocorrelation with 52 lags. For 52 lags, it shows autocorrelation at the p-value is slightly greater than 0.05. This suggests there is no autocorrelation and the ACF plot shows some spikes crushing in the boundary values. There are also problems of normality with residuals but anyways, we'll forecast for the next 52 period.

arima2 = auto.arima(q1, trace = T, test = "kpss", ic = "bic") # Developing a SARIMA model and Analysis of model # summary(arima2)
# Series: q1
# ARIMA(0,1,0)(0,0,1)[52]

Coefficients:

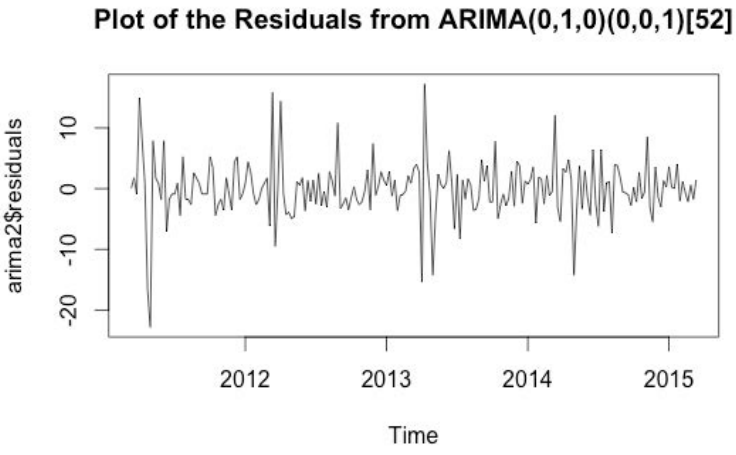
sma1

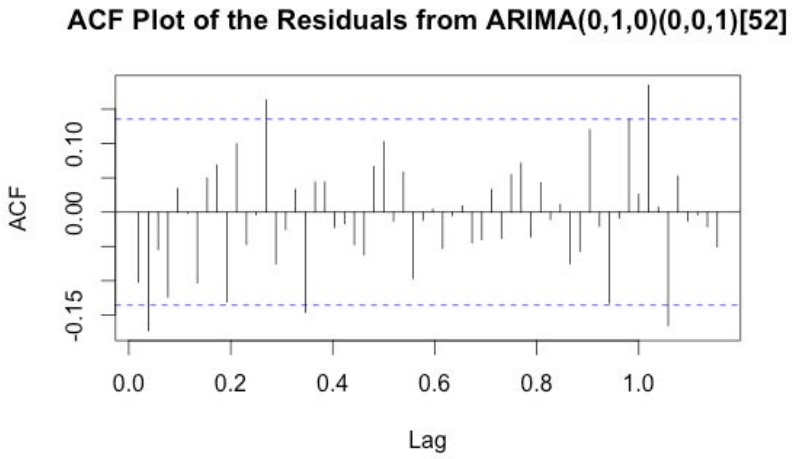
0.5431
s.e. 0.0883

sigma^2 estimated as 23.11: log likelihood=-630.26
AIC=1264.53 AICc=1264.59 BIC=1271.2

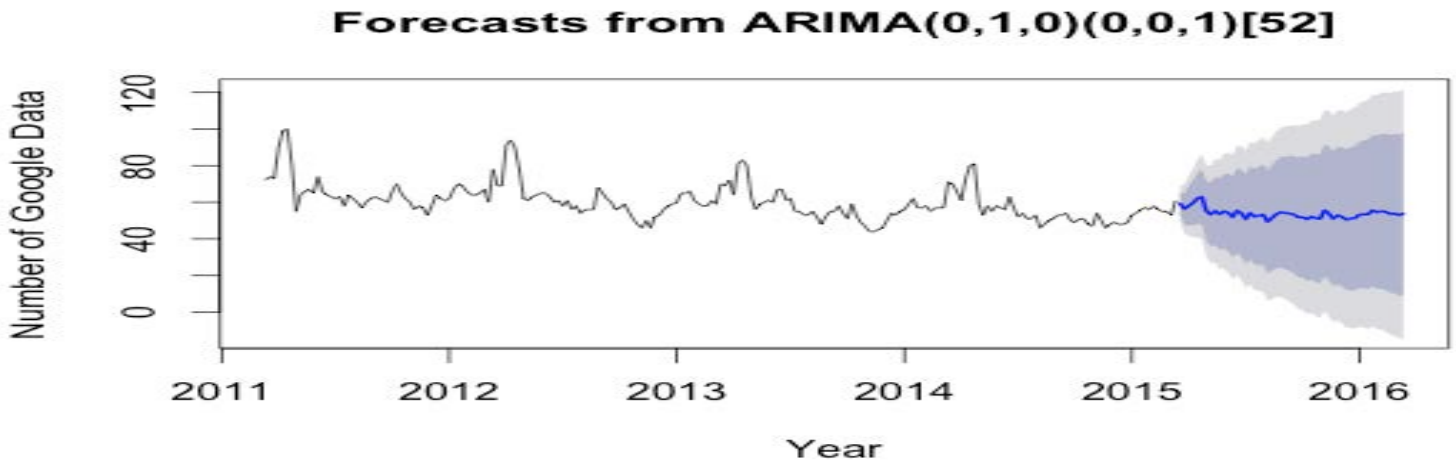
Training set error measures:
ME RMSE MAE MPE MAPE MASE ACF1
Training set -0.05690334 4.78442 3.278019 -0.3792422 5.370565 0.7524108 -0.1021737
confint(arima2)
2.5 % 97.5 %
sma1 0.3701048 0.7161614
```
```

The variable arima2 is implemented the following models recall in minimum BIC criterion. Based upon the BIC criterion, we see that ARIMA(0,1,0)(0,0,1)[52] is the best model. Notice that there is a no autoregressive parameter and moving average but there is a difference. We can also see that there is no seasonal autoregressive component and no seasonal difference component, but there is one moving average seasonality. Using the summary function, we can see the values of seasonal moving average and its standard error. Notice that if we divide the coefficient with the standard error, we will get the T-value. If the T-value is greater than 3, the coefficient is significant or the coefficient is significant far away from general. As we computed 95% confidence interval, we can see that there is 95% probability that sma1 ranges between 0.370 and .716. Since the interval does not contain zero, we can conclude that there is evidence that the arima2 component is significant. Now, let's do some diagnostic residuals for arima2.





As you can see in the Ljung-Box test, this suggests there is no autocorrelation and it is weaker than the previous model. From the previous model, the p-value of Ljung-Box test is 0.08362 and the value for this model is 0.07541. The ACF plot of Residulas shows some spikes crushing in the boundary values and the residuals are also not normal. If we square the residuals and take lag at 52, the Ljung-Box test shows that the p-value is equal to 0.01597 which is less than 0.05 which suggests us to reject the null hypothesis. But anyways, let's forecast for the next 52 periods.



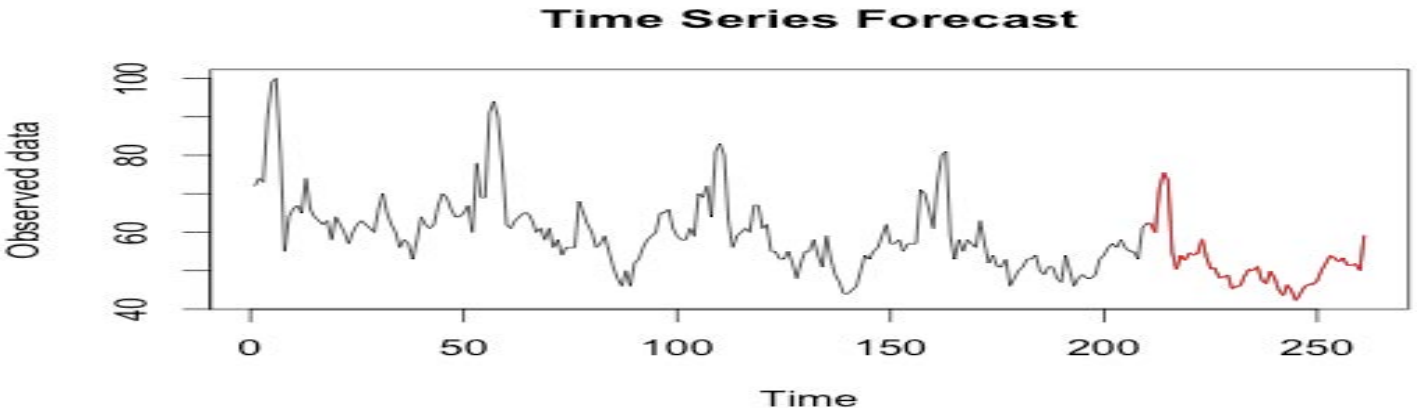
Let's look at the forecast accuracy value for the AIC and BIC model and compare them

```
accuracy(arima.forecast)
#                    ME    RMSE     MAE       MPE    MAPE     MASE       ACF1
#Training set -0.06516037 4.722994 3.233724 -0.4133793  5.29619 0.8838561 -0.02713508
#Test set      4.52770006 9.667339 6.829436  6.1159634 10.86853       NA         NA
accuracy(arima2.forecast, query1[209-52:209,1])
#Training set -0.05690334 4.784420 3.278019 -0.3792422  5.370565 0.8959632 -0.1021737
#Test set      3.93599975 9.382938 6.546722  5.0654387 10.448978       NA         NA
```

Both plots of the AIC and BIC looks very similar, but BIC seems to fit better. In summary, we note that the result of BIC criterion's work better than the AIC criterion because the Mean Absolute Percentage Error (MAPE) has a lower error in percentage, 10.44% (BIC) as opposed to 10.86% (AIC). In addition, BIC has a lower volatility (as measured by sigma^2), along with lower error statistics and MPE(5.076% as apposed to 6.115%). Our mean error in statitics is also lower meaning that any forecast with the second are more likely to be accurate than the first.

Prediction Accuracy Analysis

```
m2 = arima(ts.log, order = c(0,1,2), seasonal = list(order = c(0,1,1), period = 52))
m4 = arima(ts.log, order = c(0,1,4), seasonal = list(order = c(0,1,1), period = 52))
AIC(m2) # -397.1538
AIC(m4) # -400.4293 <- Pick m4 under AIC criterion
BIC(m2) # -384.9544 <- Pick m2 under BIC criterion
BIC(m4) # -382.1302
MSE2 <- computeCVmse(c(0,1,2), c(0, 1, 1)) # 11.67564 26.56129 <- Set K = 1;  Pick m2
MSE4 <- computeCVmse(c(0,1,4), c(0, 1, 1)) # 11.95002 23.04304 <-  Set K = 2; Pick m4; LESS BIASED
predictions <- exp(predict(m4, n.ahead = 52)$pred) # <- under BIC Criterion
predictions first 5 values are listed below
 62.34071 61.35692 58.17968 70.55995 75.28463
```



In general, our goal is to minimize AIC an BIC, but they are actually not a good estimators for the MSE. AIC and BIC are only interesting compared to other models and it does not matter whether they are positive or negative; MSE provides us the quality of an estimator in terms of variation and unbiasedness. For example, I have computed different value for m2 such that arima(ts.log, order = c(0,1,2), seasonal = list(order = c(1,0, 0), period = 52)) and the result of AIC(m2) = -509.9312 which is smaller compared to the result of AIC(m2) = -397.1538. However, note that this is not a good estimate since the prediction accuracy estimate is not close to the prediction estimate of the Holt-Winter model. In this model, we can see that BIC penalizes models with more parameters than AIC. Therefore, it leads to choosing more parsimonious models with fewer parameters. In other words, we can set K = 1 to utilize a larger training set than the test set. Note that I have set K = 2 in this model, it is a better fit because it serves as an extra evidence and validation other than K = 1 and provides us less bias towards overestimating the true expected error. In addition, setting K = 2 give me an additional sample to estimate that yields to a more accurate confidence interval on the estimate. However, if I make an argument of k = 3 or k with higher order, it involves more computation and leads to higher variance and longer running time. Therefore, when K gets bigger, k-fold CV can simulate arbitrary hard samples of the training set.

********** Reference **********

"8.7 ARIMA Modelling in R." 8.7 ARIMA Modelling in R | OTexts. N.p., n.d. Web. 16 Nov. 2016.

"Time Series Analysis." Time Series Analysis. N.p., n.d. Web. 16 Nov. 2016. Roopam.

"Regression with ARIMA Errors to Test E˙ective Marketing? - Case Study Example." YOU CANalytics. N.p., 14 Sept. 2016. Web. 16 Nov. 2016.

Shumway, Robert H., and David S. Sto˙er. Time Series Analysis and Its Applications: With R Examples. New York: Springer, 2006. Print.

"Forecasting Time Series with R." Dataiku RSS. N.p., n.d. Web. 16 Nov. 2016.

Wu, Jason. "Lecture Slides." Bcourses.berkeley.edu. Snoren, n.d. Web.

"Model Selection." SpringerReference (n.d.): n. pag. Web.