**News Sentiment Analysis on Stock Prices**

**Group 2**

Meghana Kasula

Pruthvi Nagaraju

Shaloma Ghosh

Vinit Gupta

Weining Zheng

Contents

[Project Objective 2](#_Toc480868048)

[Data Set Details 3](#_Toc480868049)

[Data Preprocessing – Stock prices 3](#_Toc480868050)

[Sentiment Analysis (News Data) 4](#_Toc480868051)

[Modeling - Forecasting Without Sentiments 9](#_Toc480868052)

[Modelling - Forecasting with Sentiments 19](#_Toc480868053)

[Comparison between Models in Part 1 and Part 2 27](#_Toc480868054)

[Volume prediction 27](#_Toc480868056)

[Business Applications 31](#_Toc480868057)

[Conclusion 32](#_Toc480868058)

[References 32](#_Toc480868059)

[Appendix 33](#_Toc480868060)

Executive Summary

* Stock Prices are often used as an important indication of overall strength and health of a company. In general if the stock prices continues to climb, the company is said to be growing.
* We analyze the stock prices and observe that if everyone is happy, especially the customers, there is a likelihood that the stock prices will increase or at least not decrease.
* This is where the integration of fundamental and technical analysis seems inevitable. Fundamental analysis takes the underlying causes in price shift into consideration and technical causes taking previous stock prices into consideration.

# Project Objective

* Public mood on news/twitter about a company is representative of overall sentiment about the respective company.
* The overall sentiment about a company can have a direct or indirect impact on its stock prices. As a result, analyzing sentiments along with the traditional stock prices forecasting is a very important aspect.
* One of the biggest source of extracting public reaction to a company or its product or any stock price related variables are Twitter, Google trends, Reddit, etc.
* Time series analysis is the most common and fundamental method used to perform this task.
* Our project aims to combine the conventional time series analysis with news from Reddit to predict weekly changes in stock price.

# Data Set Details

* Our Data set was taken from the website:

<https://www.kaggle.com/aaron7sun/stocknews>

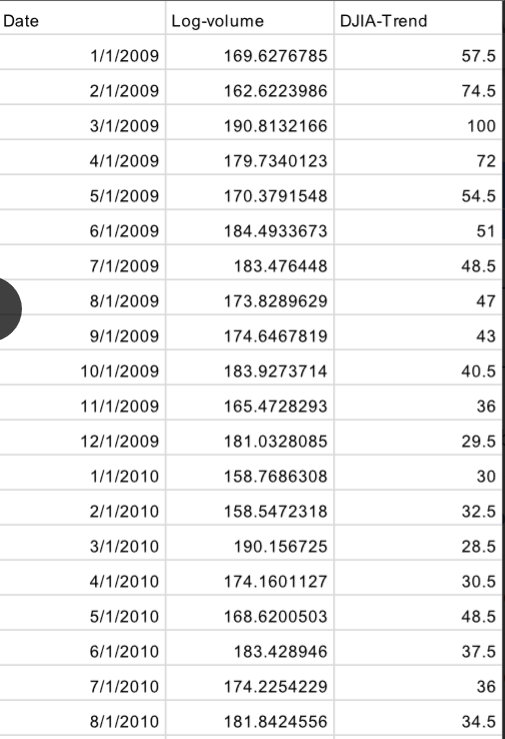
* There are two channels of data provided in this dataset:

1. News data: It is the historical news headlines crawled from [Reddit World News Channel](https://www.reddit.com/r/worldnews?hl) . They are ranked by Reddit users' votes, and only the top 25 headlines are considered for a single date. (Range: 2008-06-08 to 2016-07-01)
2. Stock data: Dow Jones Industrial Average (DJIA) is used to "prove the concept". (Range: 2008-08-08 to 2016-07-01)

* The Stock data has 6 variable namely, open price, close price, high price, low price, volume and adjusted close price.
* The News Data has the top 25 news for that day as ranked by Reddit users.

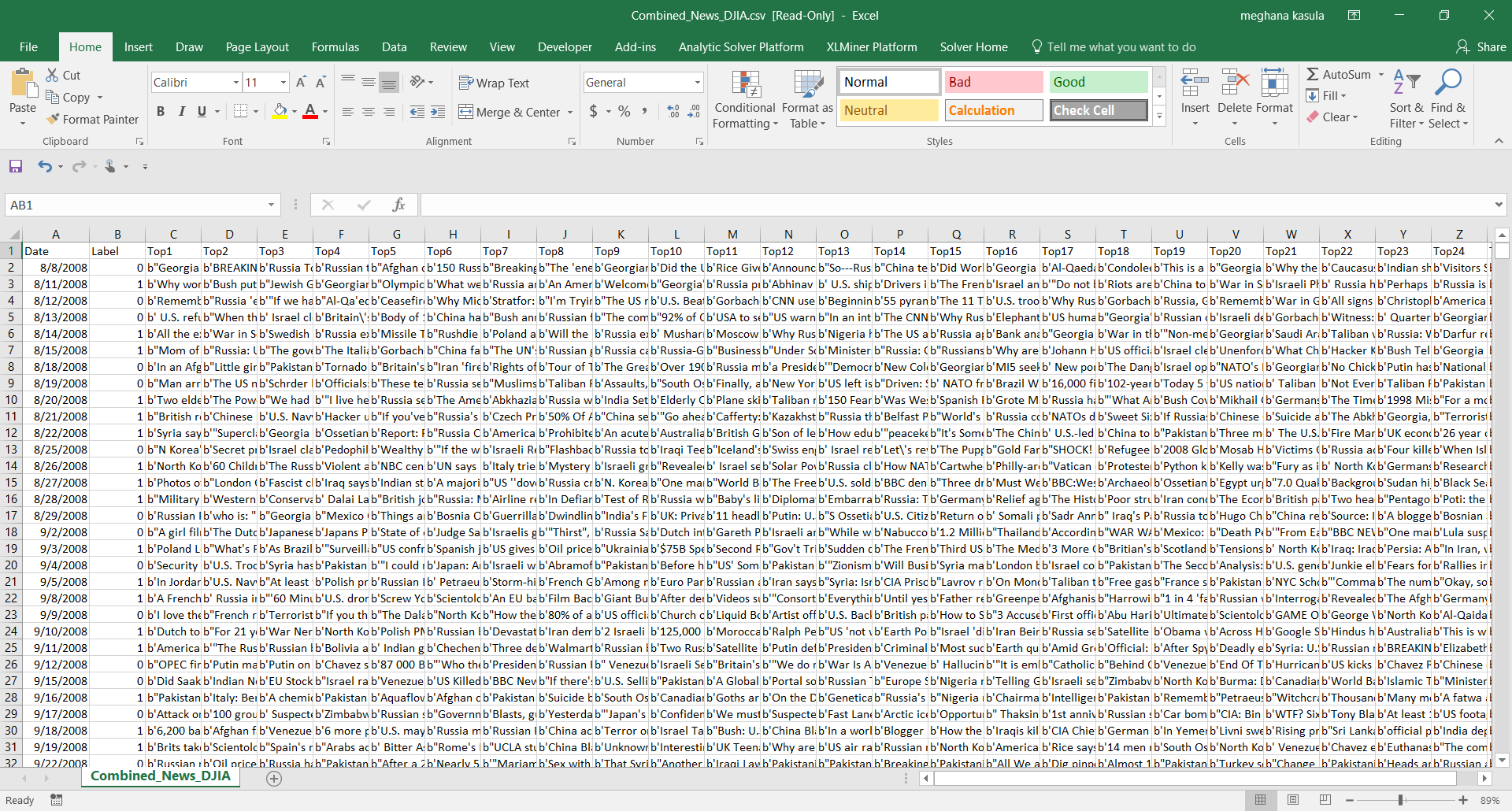
# Data Preprocessing – Stock prices

In data consisting of stock volume traded and the monthly DJIA trend, since the numerical value of the volume data was very high, we did log transformation of it to easy analysis.

Below is the screenshot of the data after taking log transformation

# Sentiment Analysis (News Data)

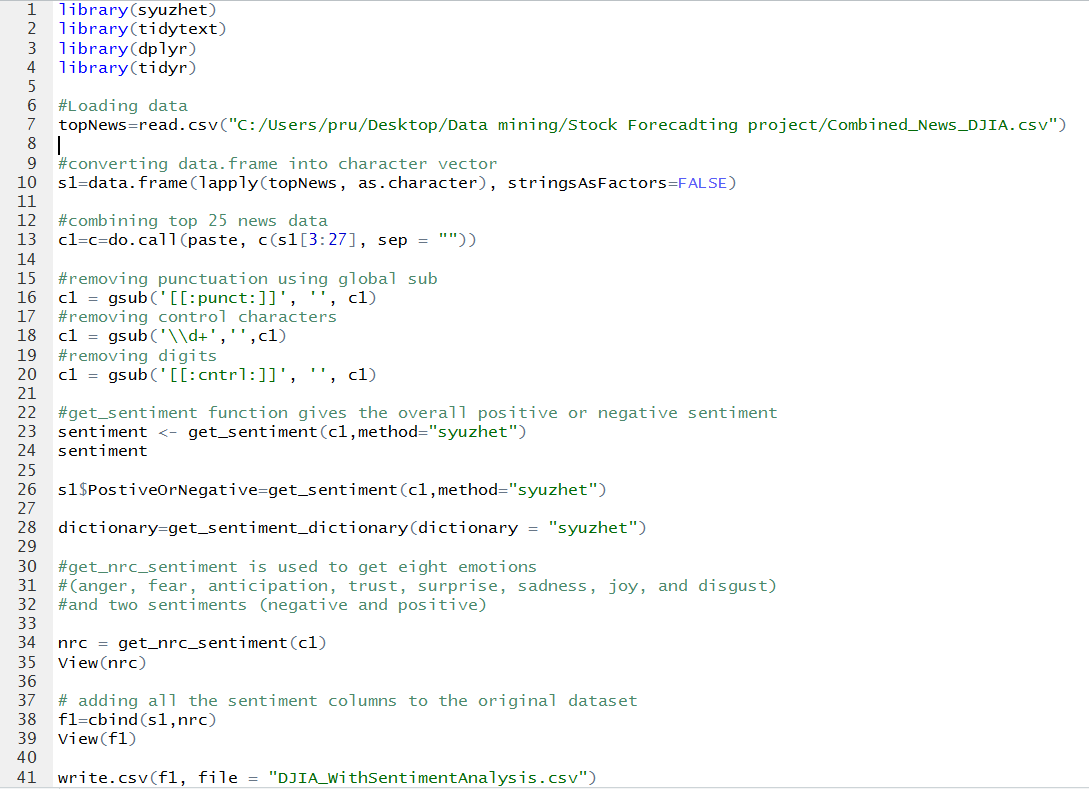
* The News Data has a total of 27 columns with Data and label and 25 top news included.
* Following is the screenshot of the raw data.



* Following is the dataset after preprocessing (performing sentiment analysis for each day)

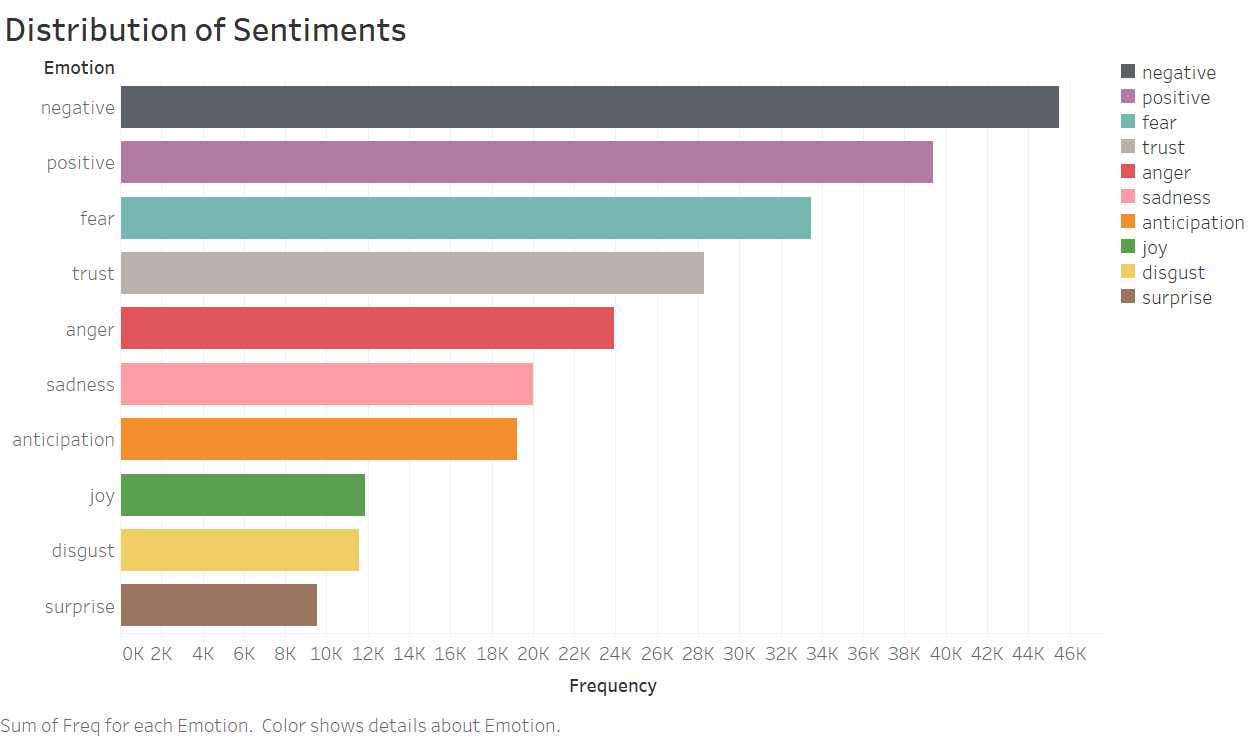


* The steps involved in processing it are as follows:
* Combined all 25 news Data in one column.
* Removed punctuation, control characters like \b,\t etc and digits.
* Used a package called “Syuzhet” on R to do an sentiment analysis of the combined news
* Used two Functions like ‘get\_sentiment’ (meant for positive and negative sentiment) and ‘get\_nrc\_sentiment’ (meant for emotions like Trust, joy, anger, sadness etc.) to get sentiments of combines news each day.
* We have now 11 sentiments namely - Positive or Negative (sign indicated both), Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, Trust, Negative and Positive.
* We will be using all sentiments to explore the news data on stocks
* We will be only using Positive or Negative Dataset for forecasting since it can store information of all kinds of news.

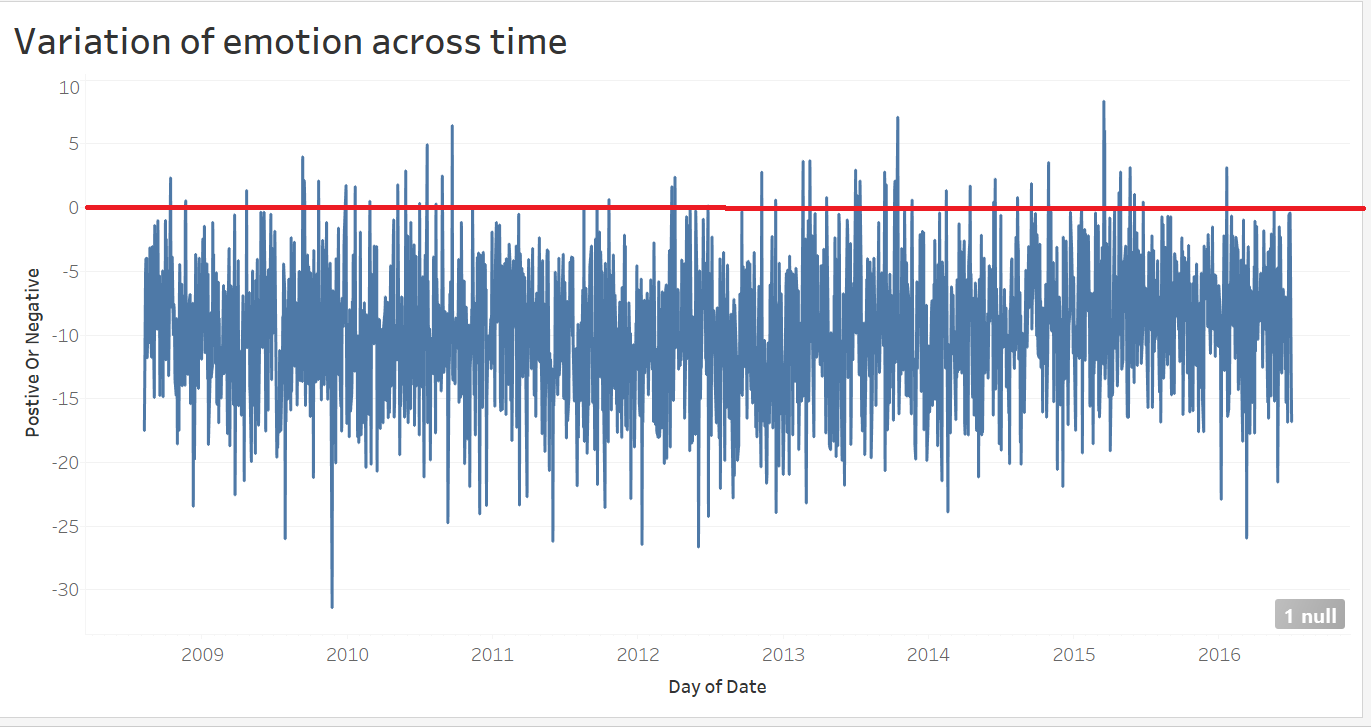
Following is the code we used to perform sentiment analysis

Data exploration

* Frequency of all the emotions was calculated and it was observed that overall, the frequency of negative emotion was the highest.

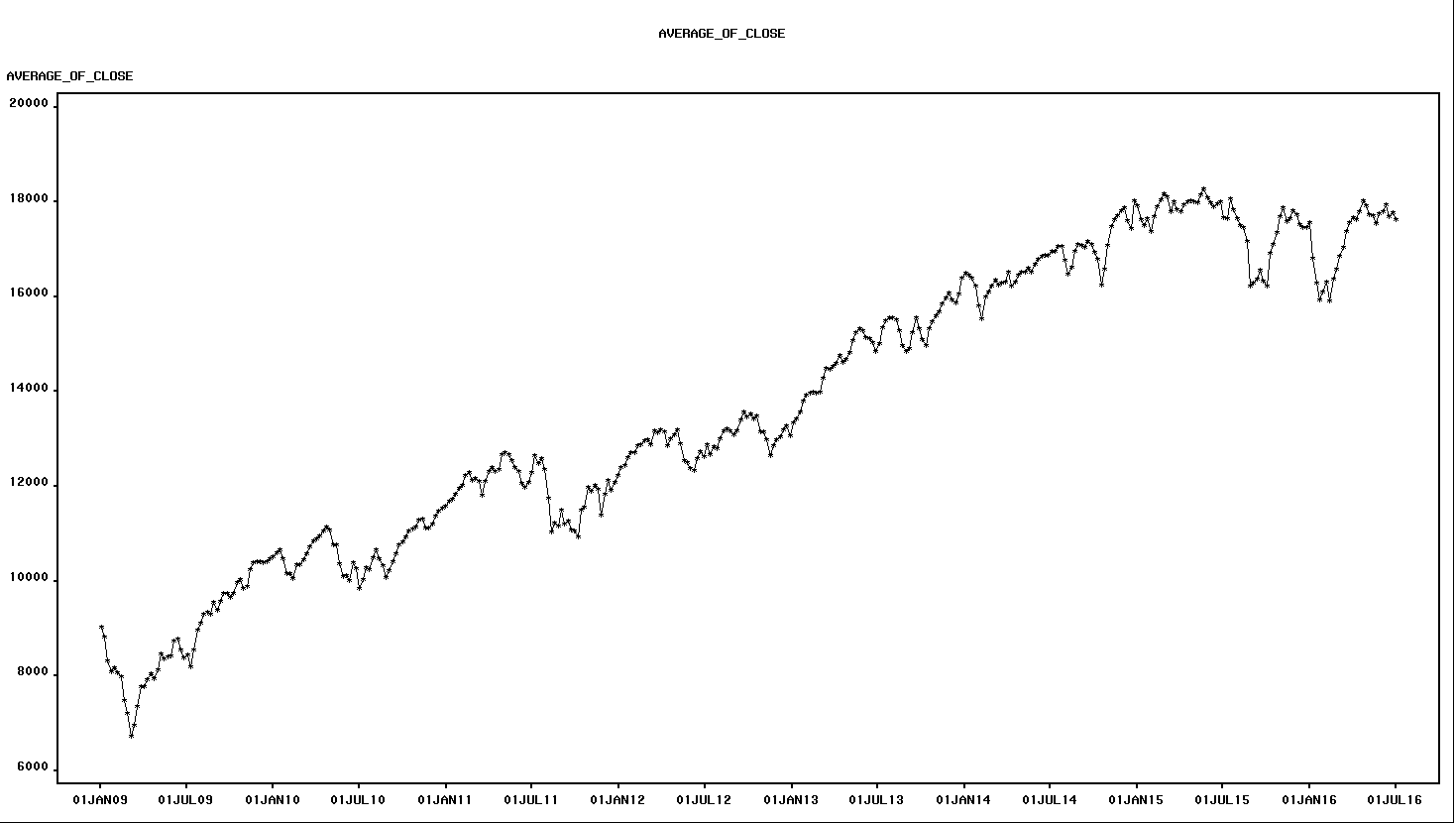


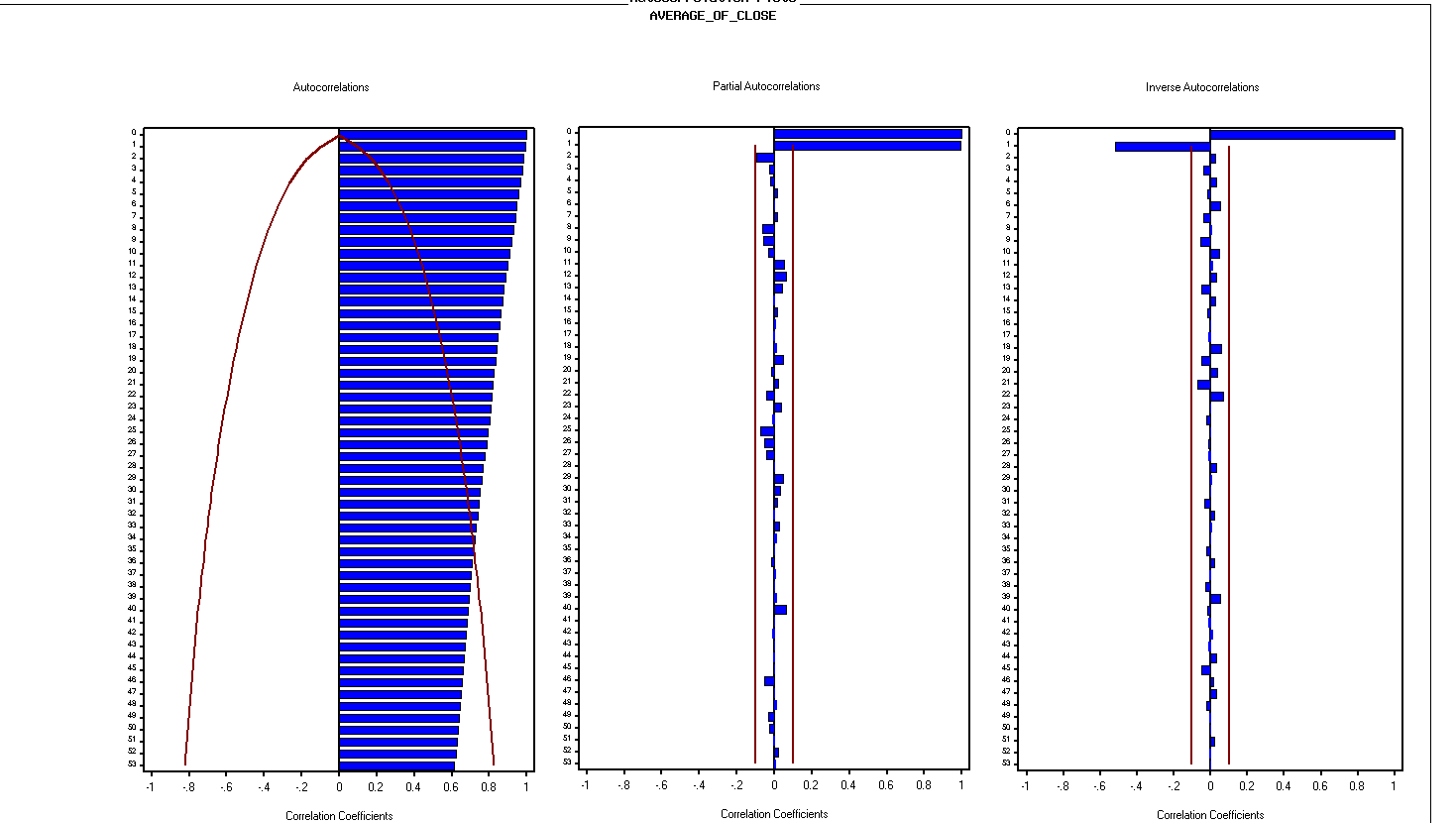
* The graph below shows the correlation between volume of the stocks and the DJIA trend. Peaks in the DJIA trend coincides with the peaks in the stock volume at most places.
* Most of the sentiments are Negative, this shows that the news data could be biased.



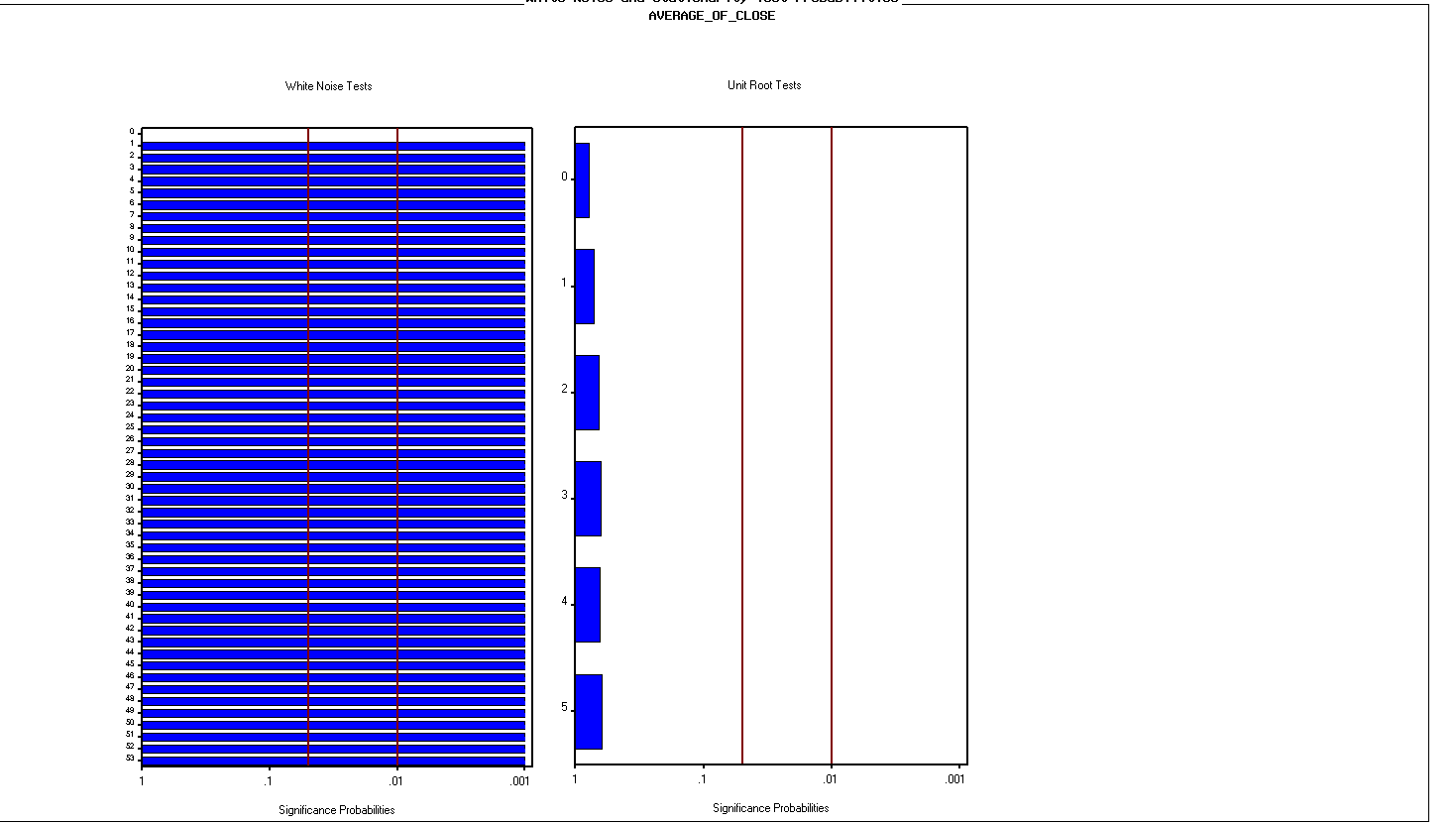
# Modeling - Forecasting Without Sentiments

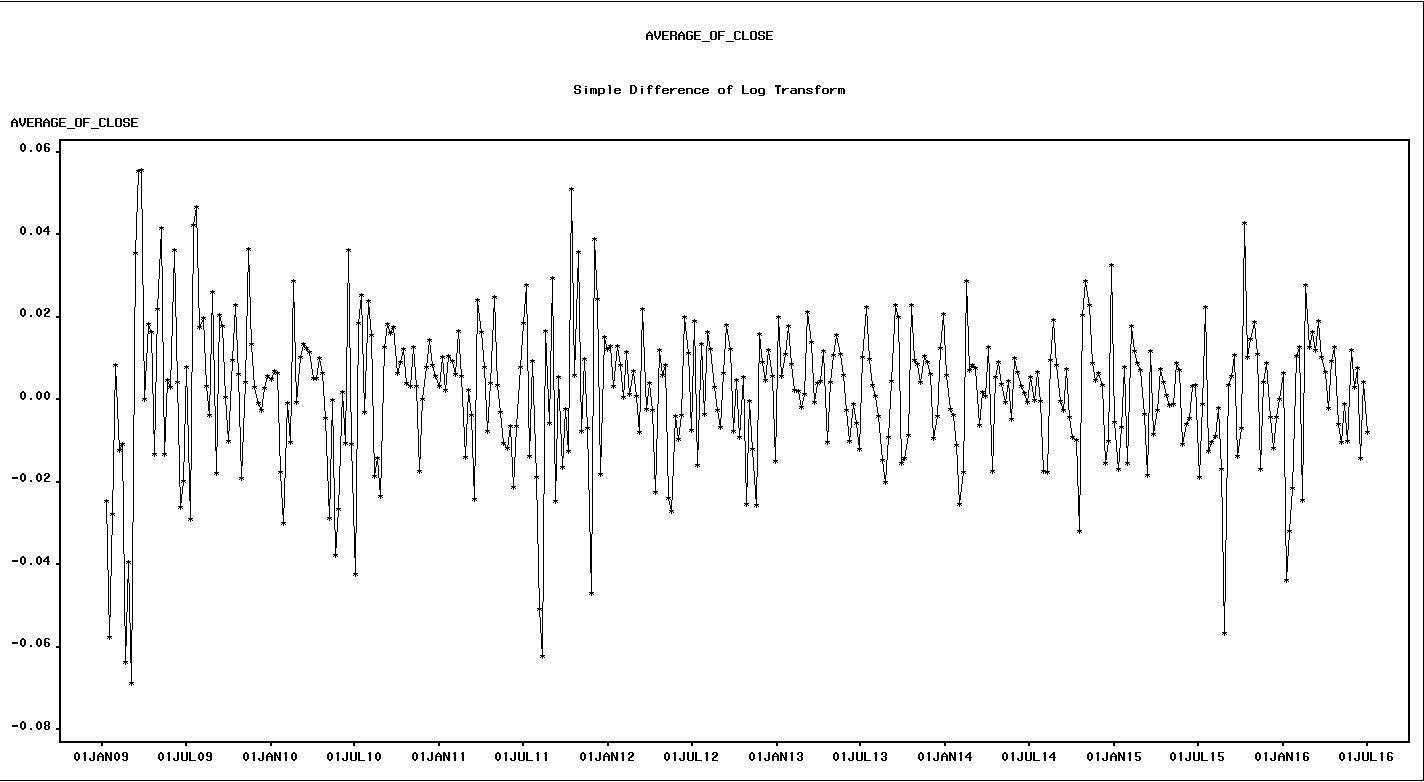
* Objective The objective of the project is to forecast the DJIA stock prices.



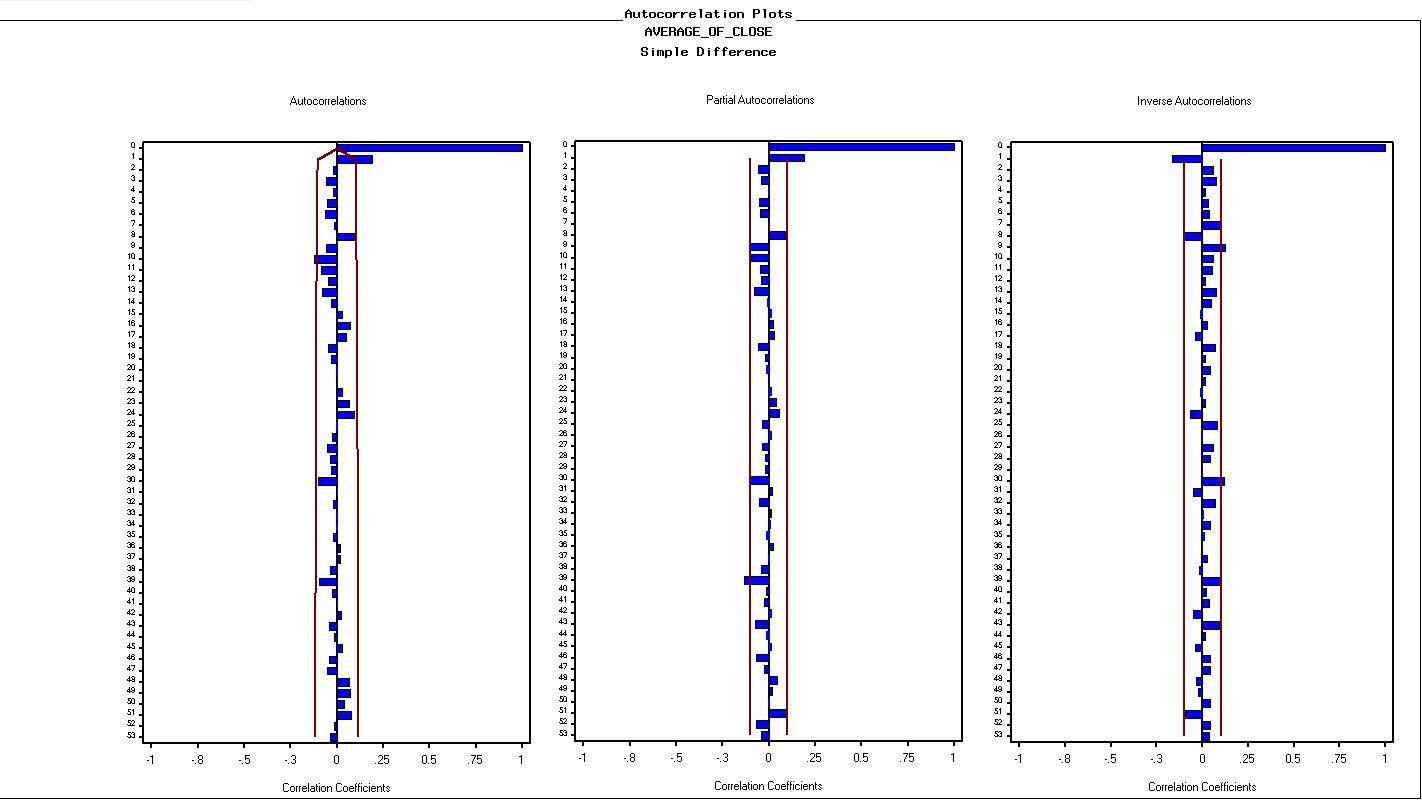
The trend indicates that the stock value has increased over time with certain dips. This is due to general improvement the market and the diminishing value of money over time.

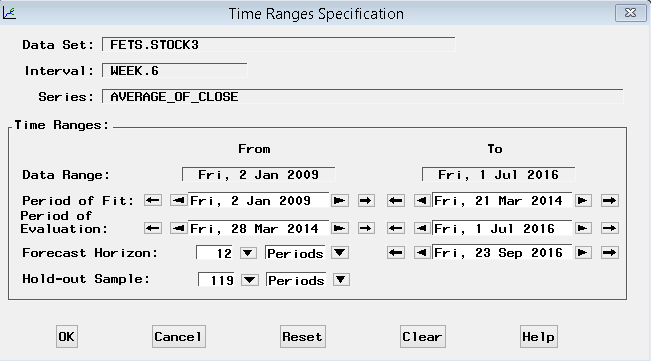
The autocorrelation plot indicates that most of the values are beyond the 95% confidence.

There is significant white noise.



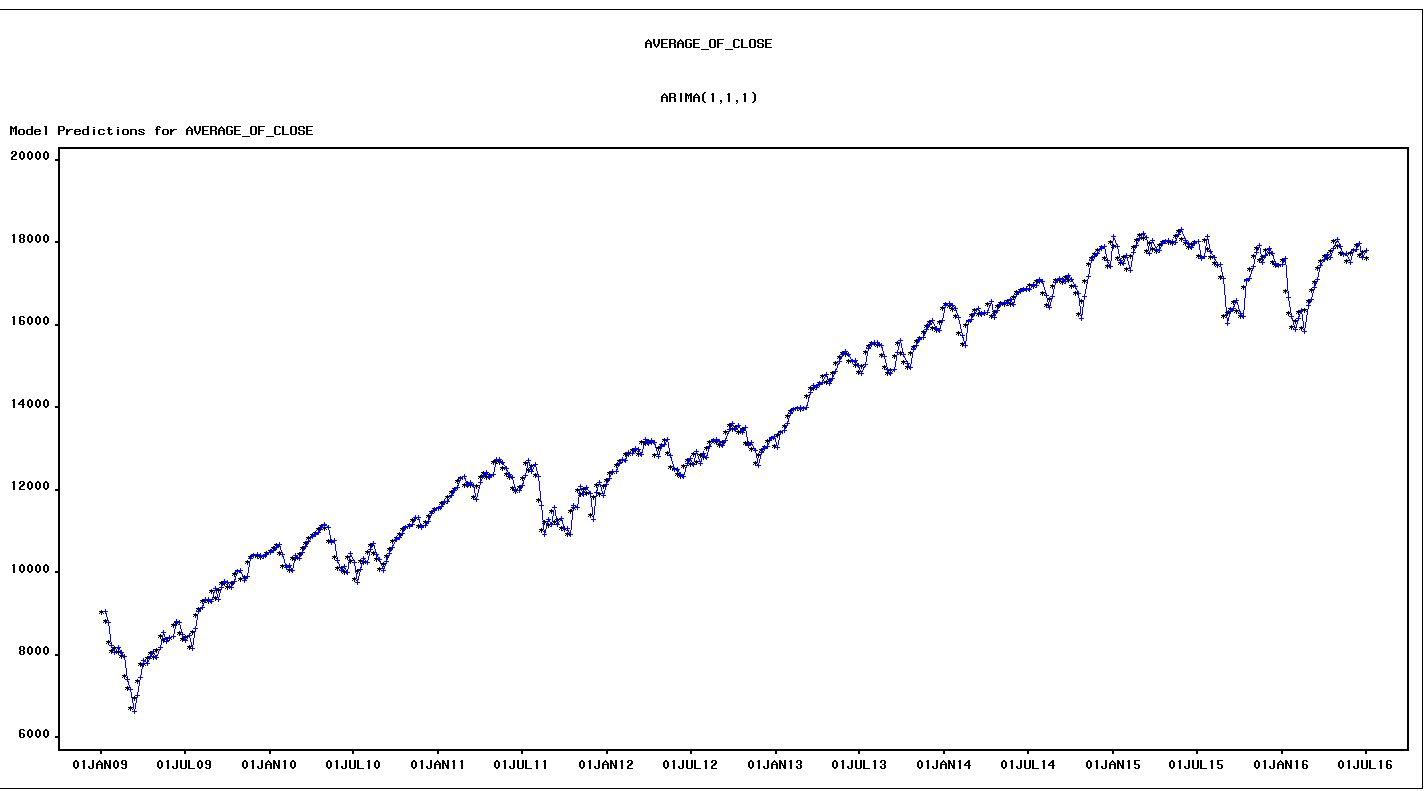
First differences have been applied of log transfers.



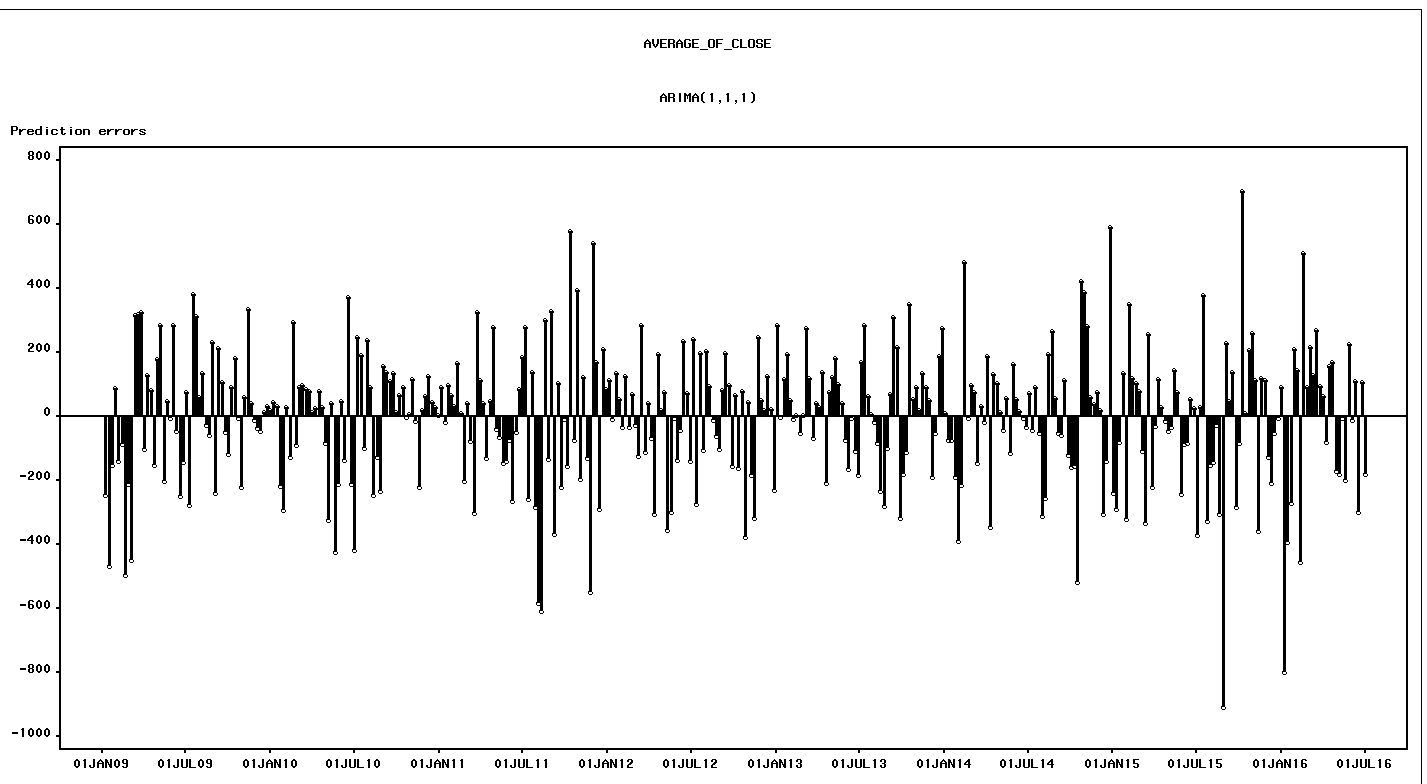
The correlation and auto correlation plots look much better with spikes at p=1 and q=1

We divided the data into training and holdout data set with 70% training and 30 holdout .

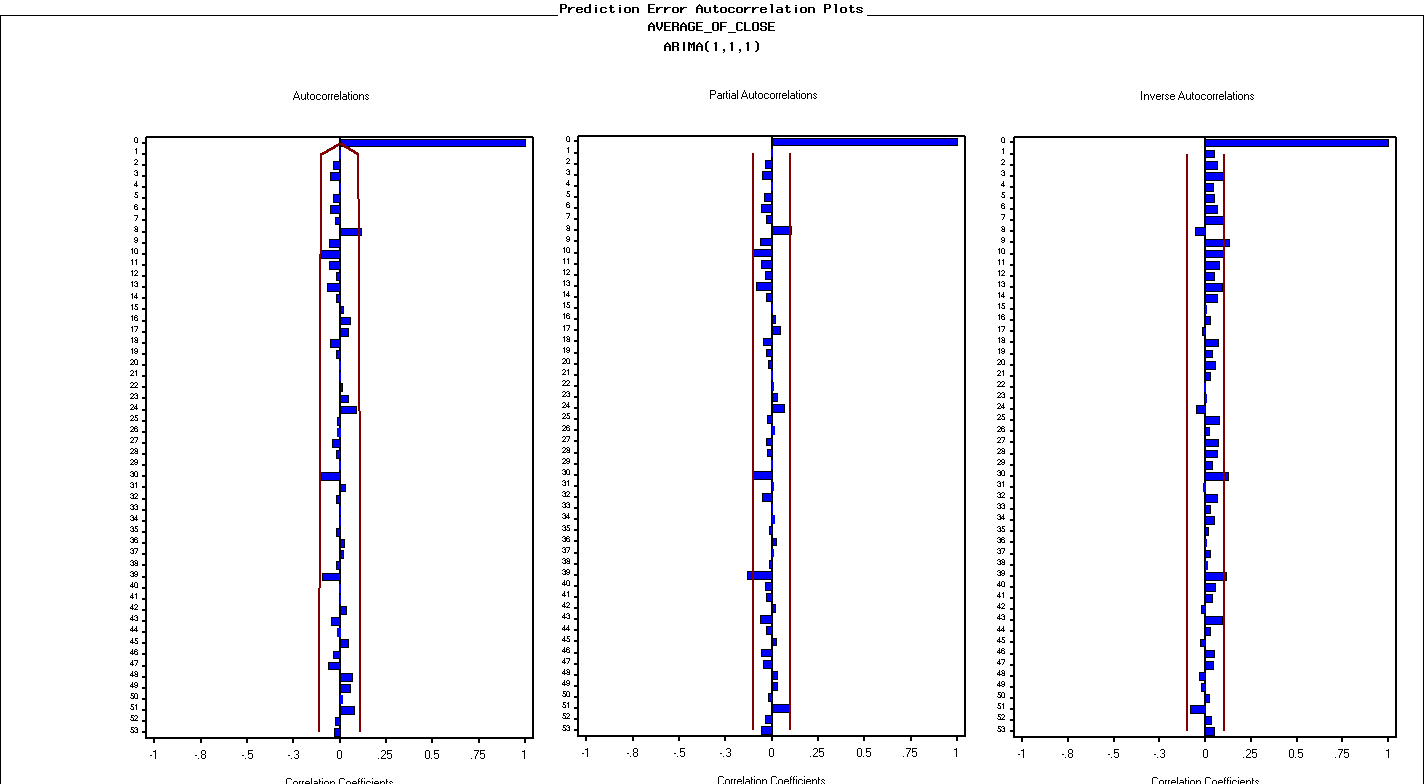


There were 3 models which we developed out of which ARIMA (1, 1, 1) and Damped Trend Exponential smoothing showed least RMSE. 

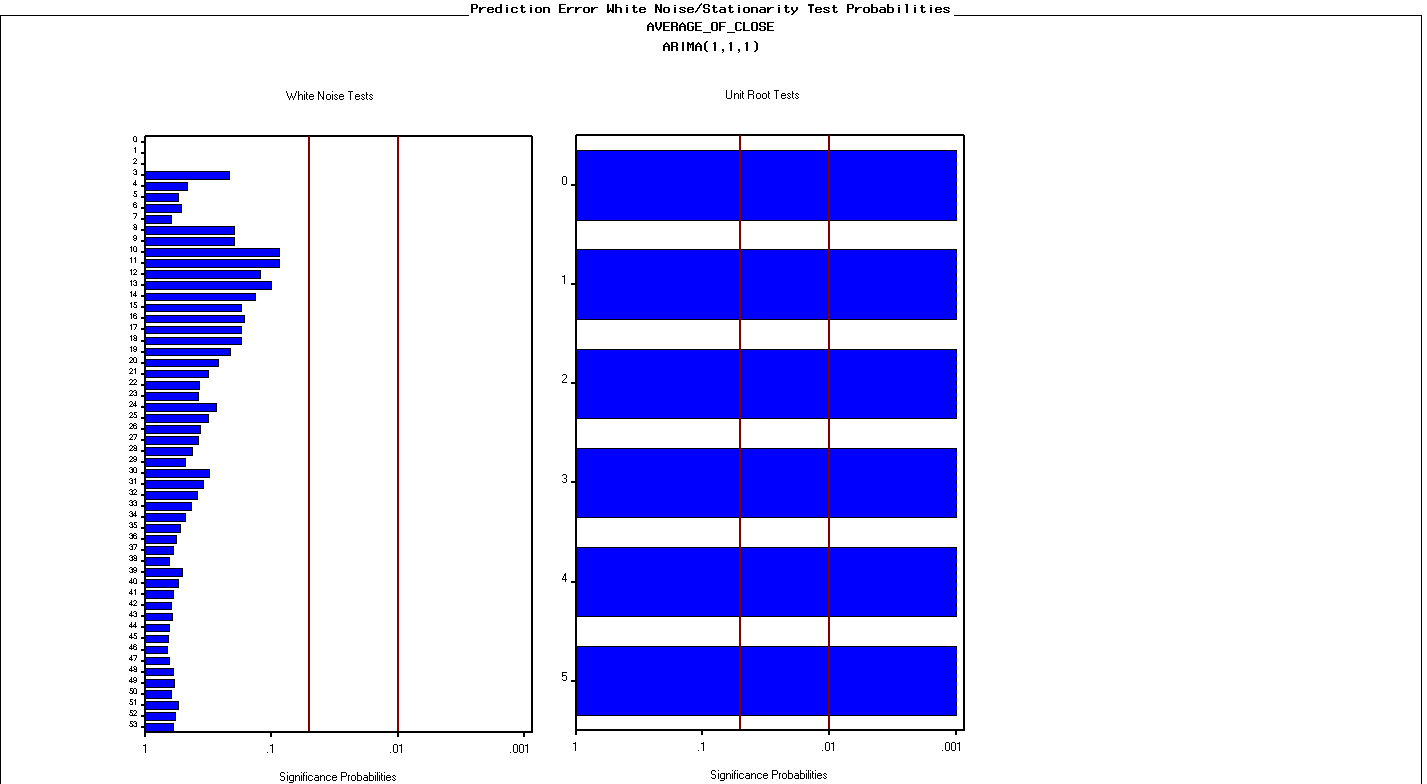
This is the trend of the ARIMA(1, 1, 1) model.

**Model 1**

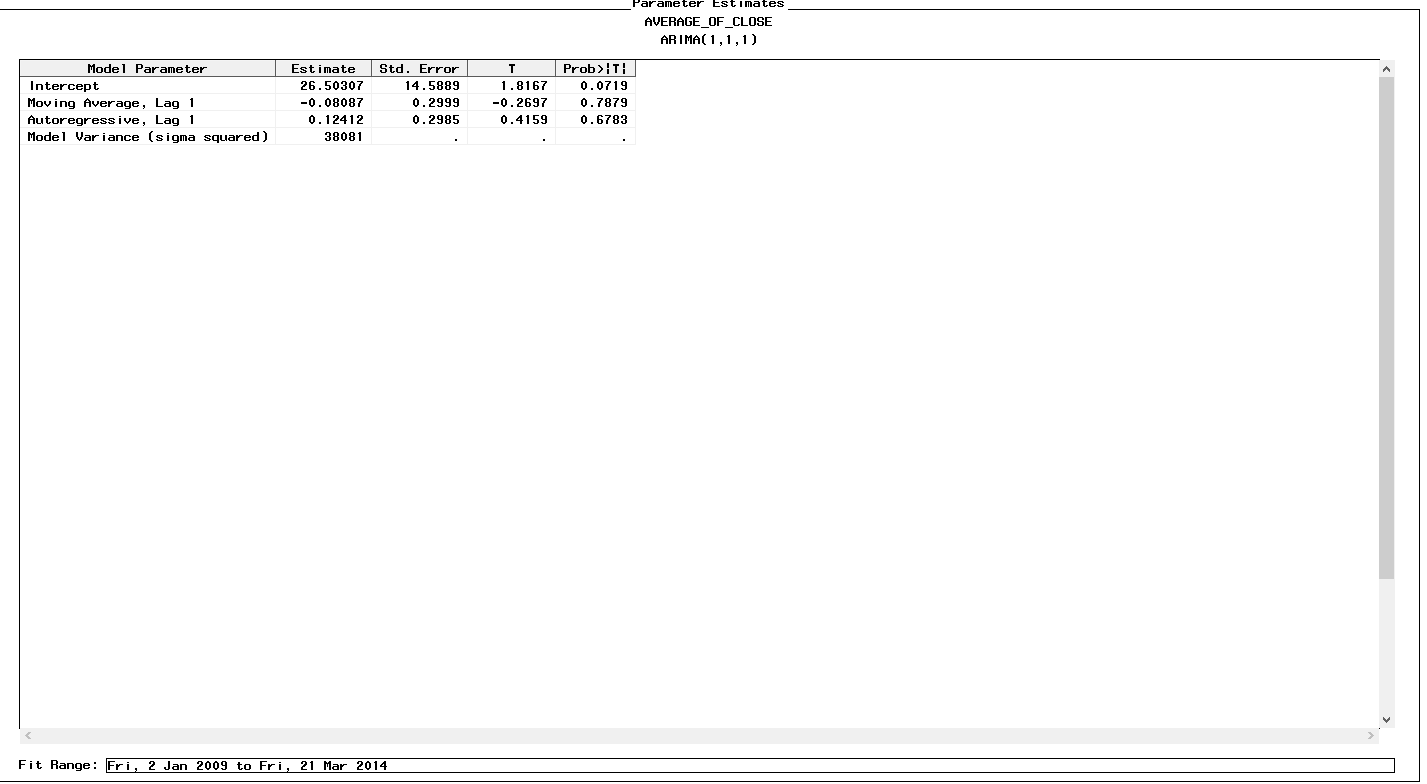
Prediction errors: The prediction errors look symmetrical along the x axis apart from a few lags towards the end which may be due to an event.



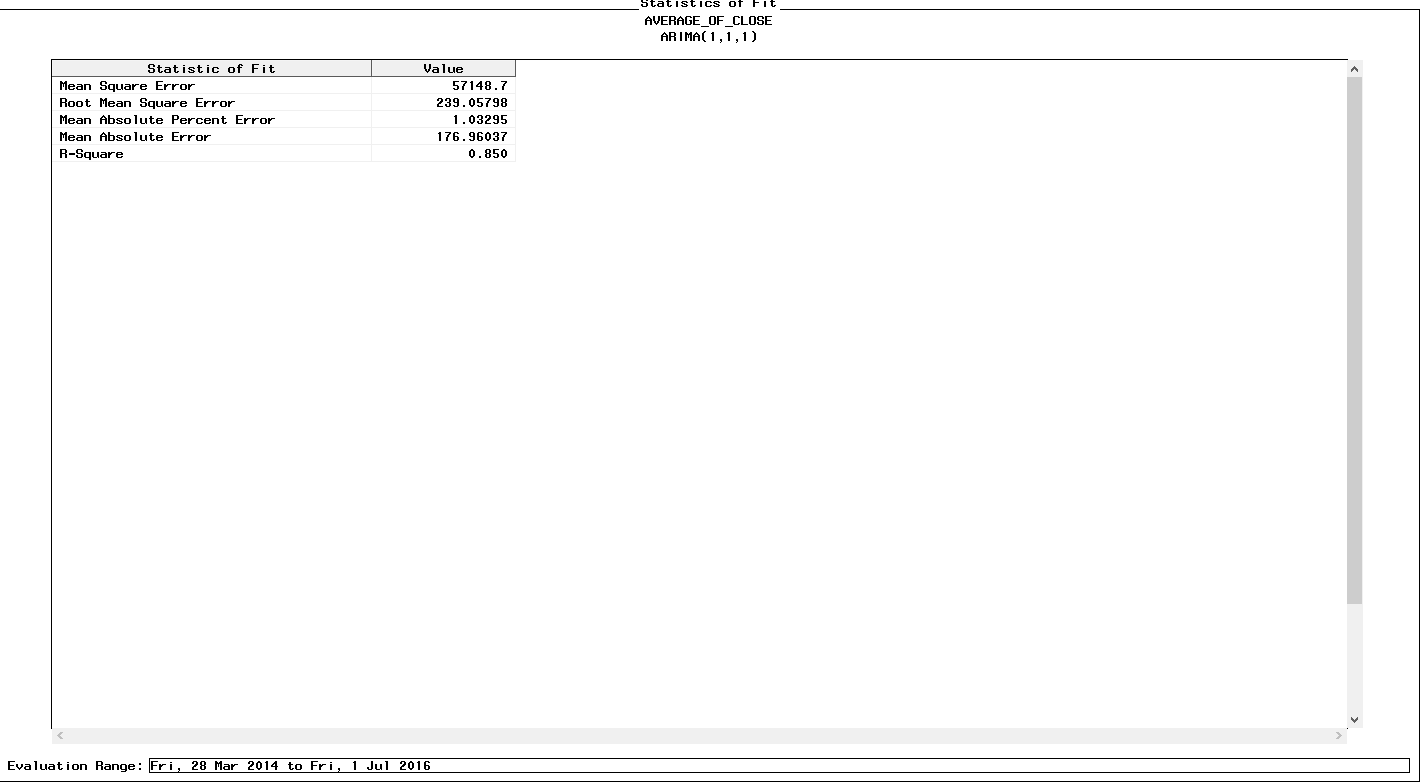
Prediction autocorrelation plots: The autocorrelation and correlation plots look reasonable with all values within the 95% confidence interval



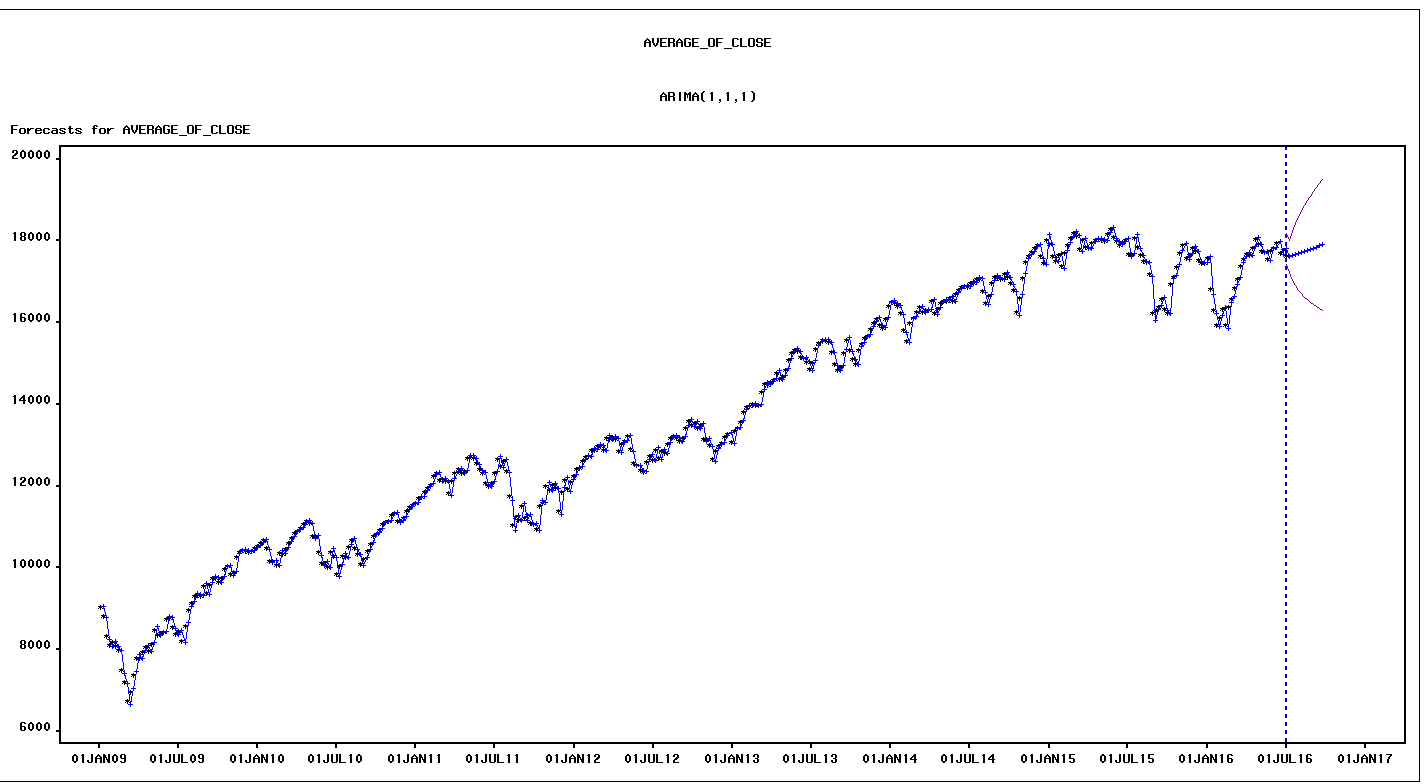
Prediction error white noise/ Unit root test: The white noise test is insignificant. And unit test indicates that the series is stationary.



Parameter Estimates: The variable intercepts have significance in the model forecast



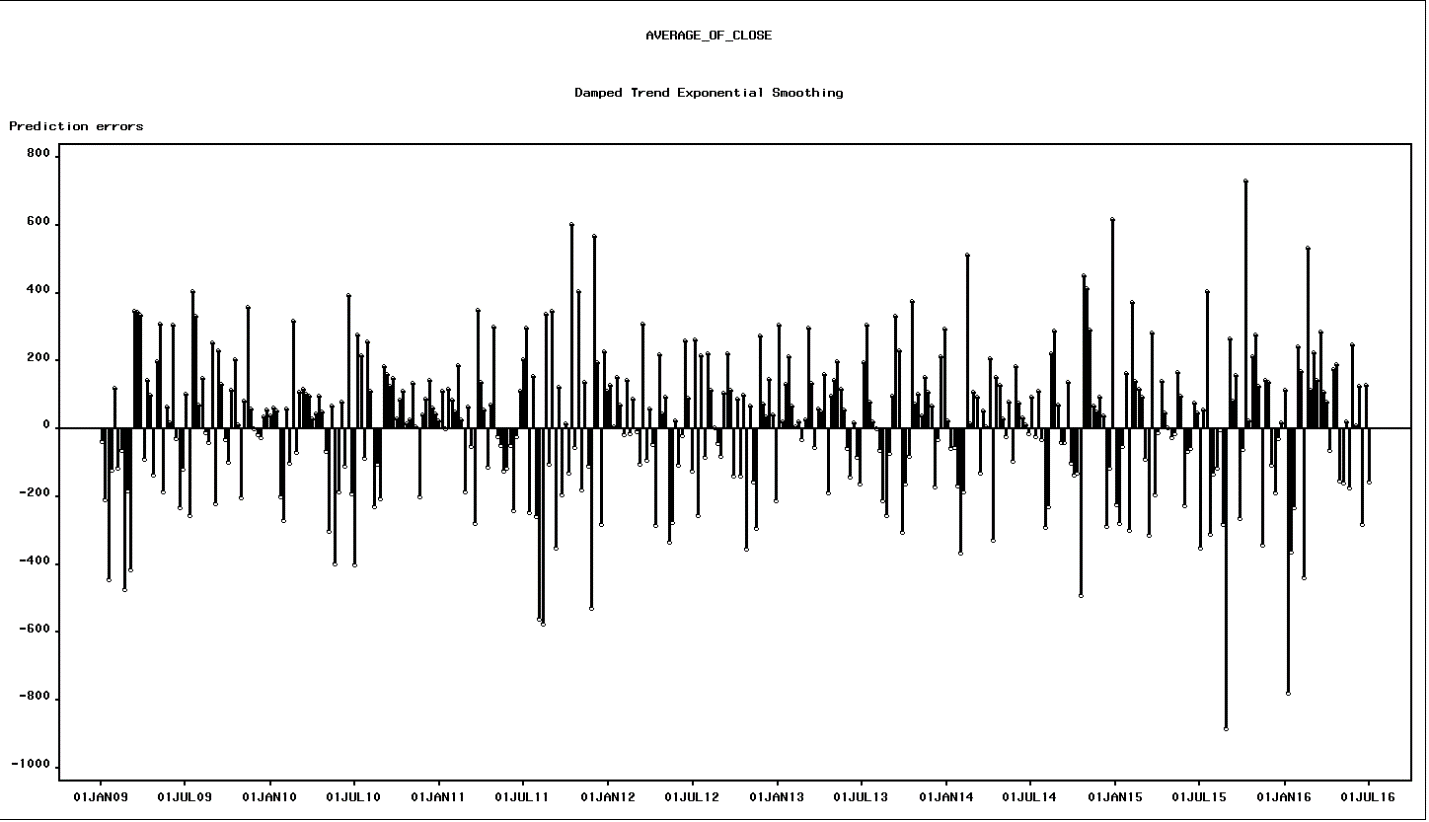
Statistics of fit: MAPE is 1.03%, which is good.



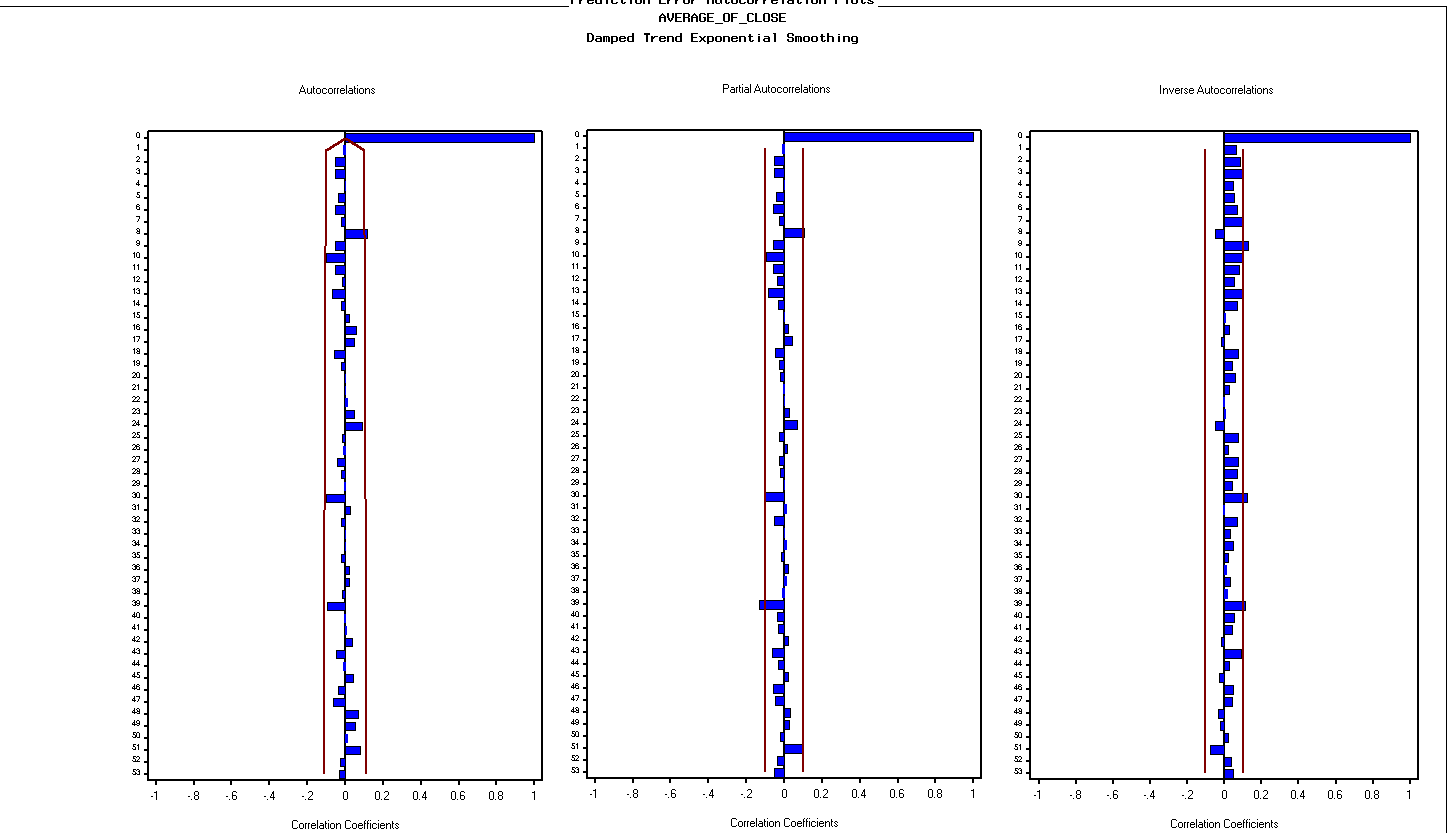
Forecast plot: The forecasted plot captures the trend well.

**Model 2 : Damped Trend Exponential Smoothing**

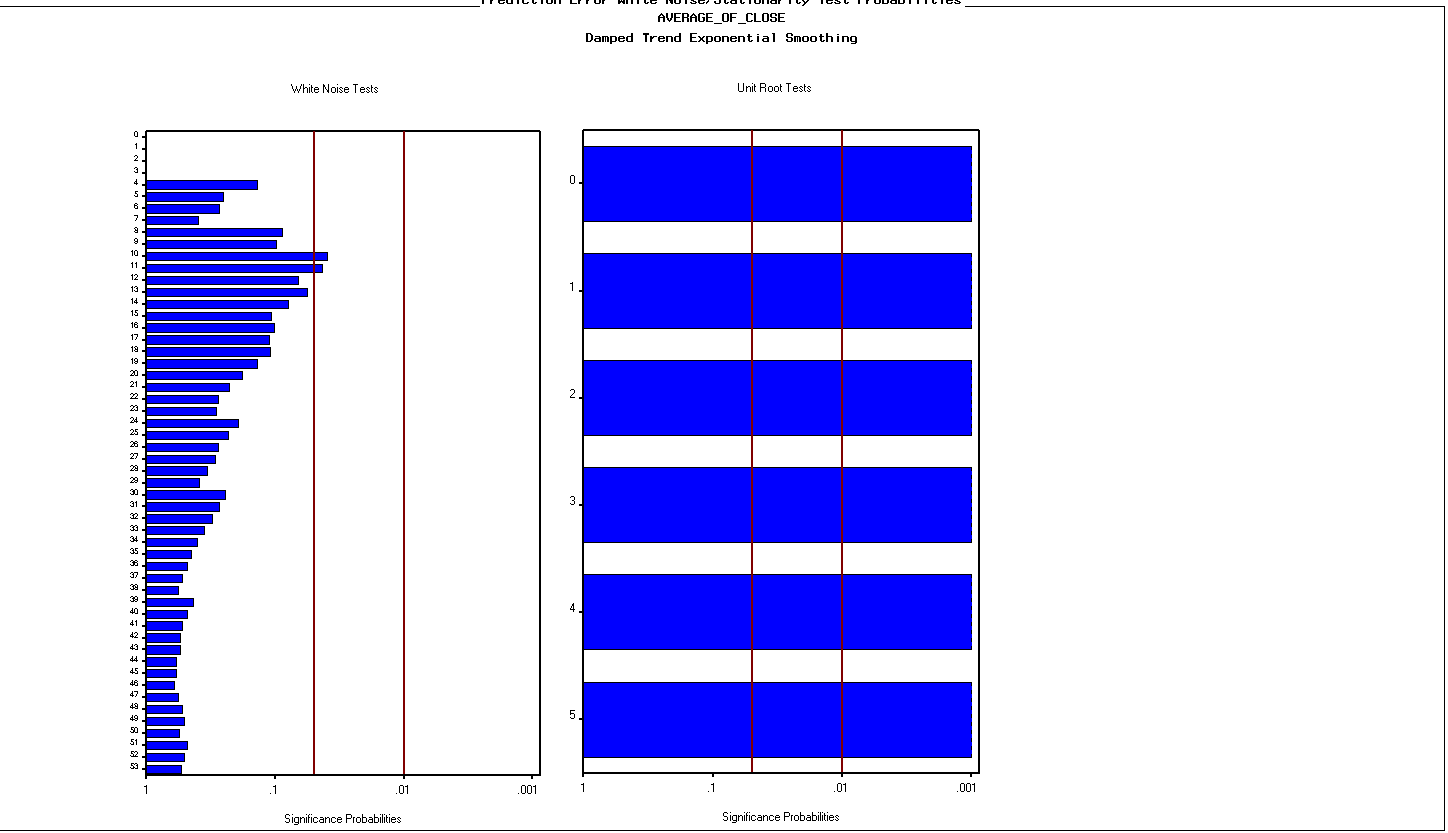
We then observed the parameters of the next best model in terms of RMSE to compare both models and finally chose one model over the other.



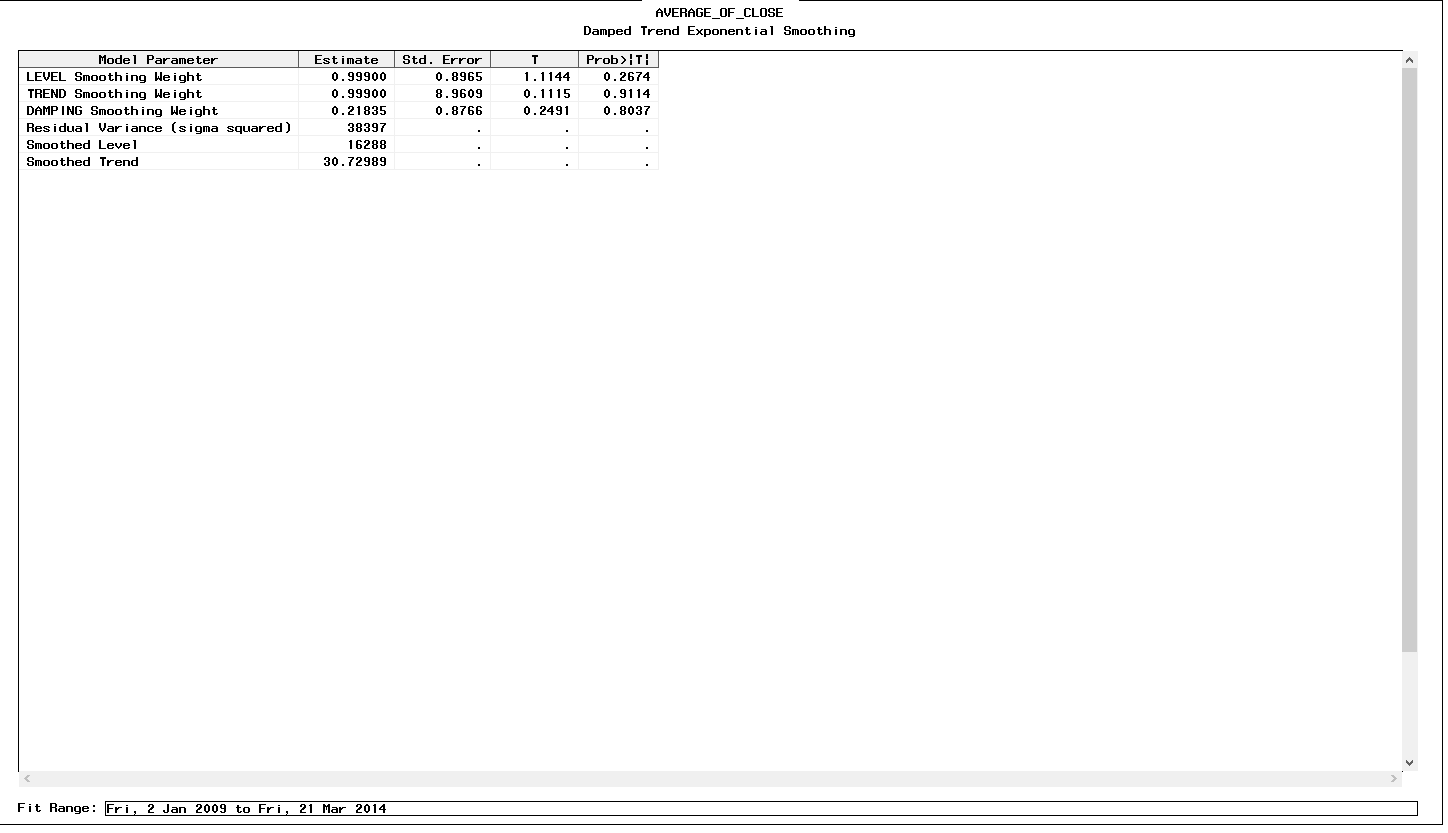
Prediction errors: The prediction errors look symmetrical along the x axis apart from a few lags towards the end which may be due to an event like the previous model.



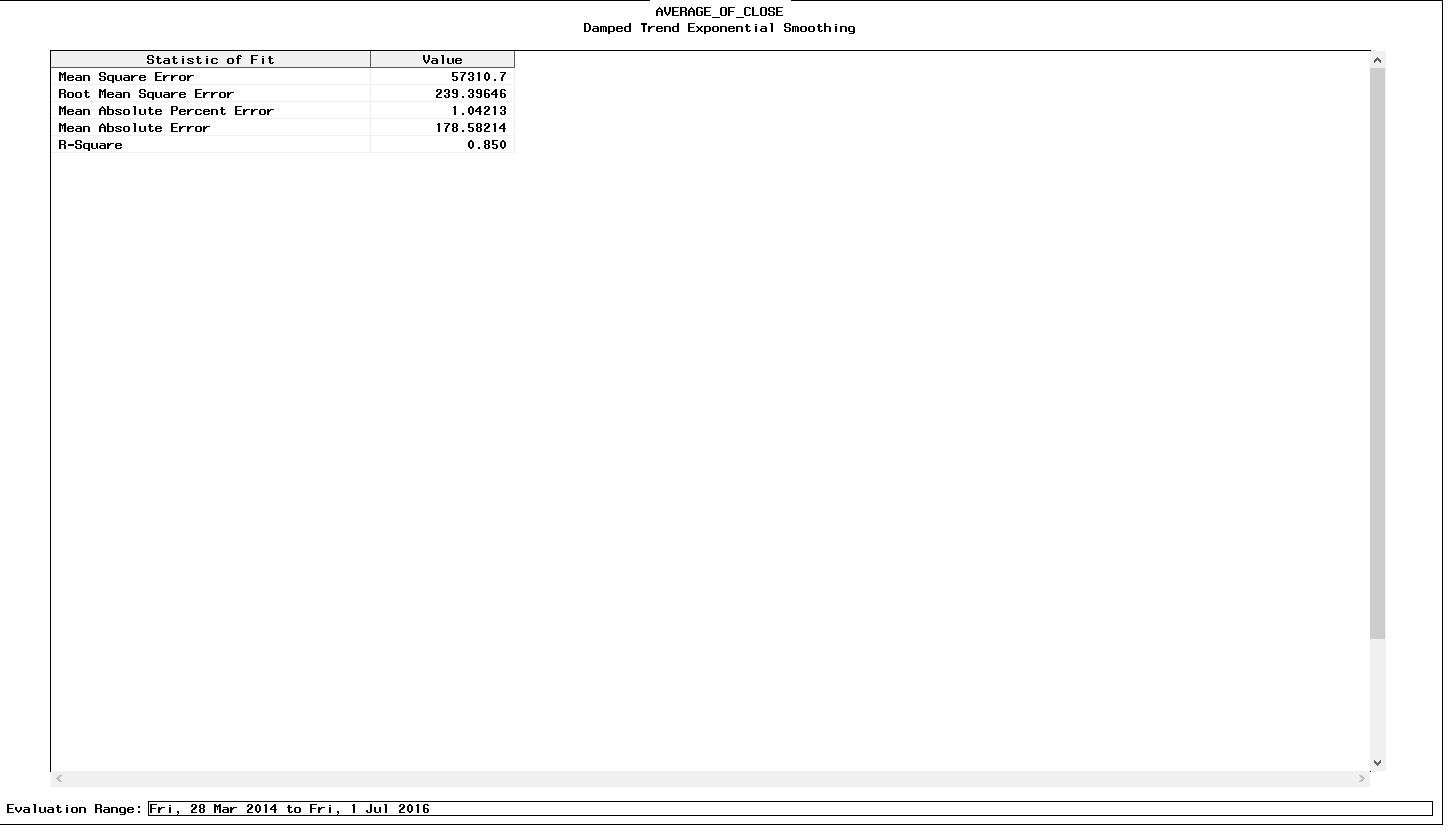
Prediction autocorrelation plots: The autocorrelation and correlation plots look reasonable with most values within the 95% confidence interval



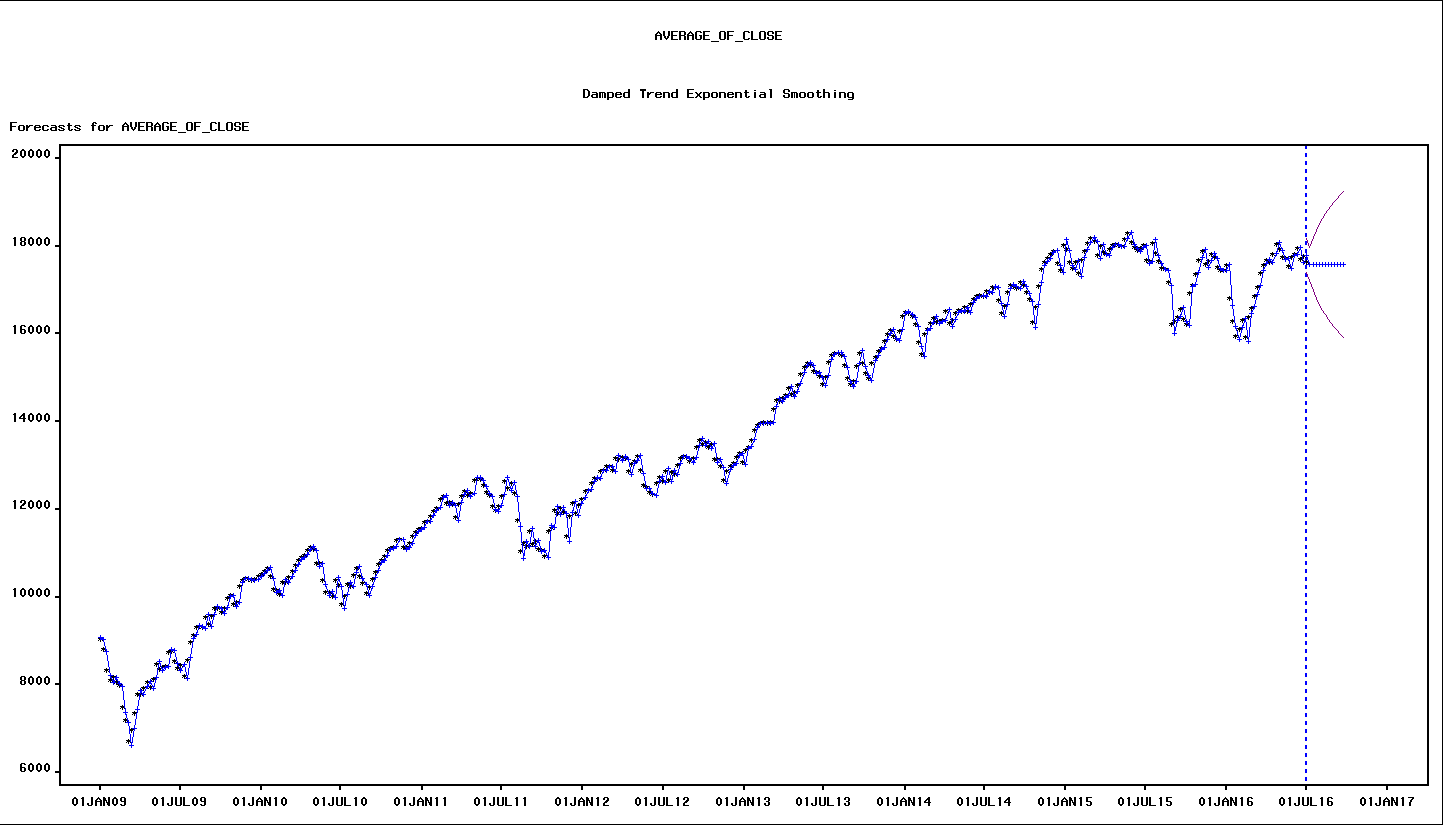
Prediction error white noise/ Unit root test: There is some significant white noise, the unit test indicates that the series is stationary.



Parameter Estimates: The variables have less significance in the model forecast, however we don’t reject the model based on this.



The statistics of fit looks good.

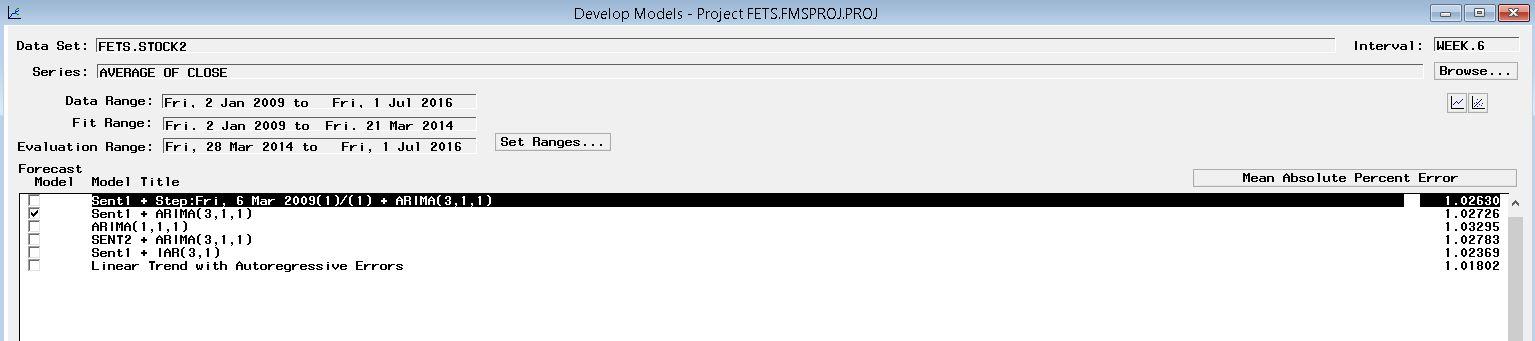


Forecast plot: The forecast shows no trend and is simply the mean. Thus we reject the model

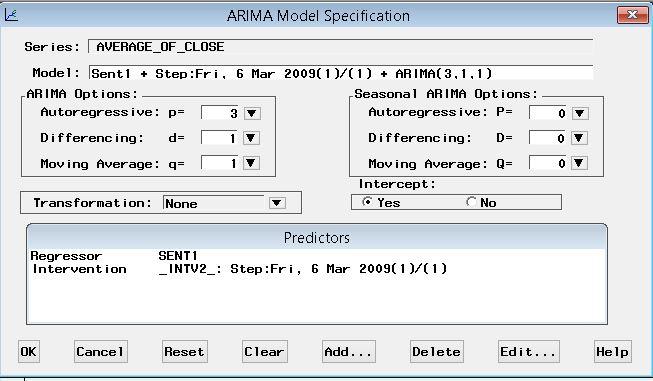
**Conclusion**

Out of the two models we select the ARIMA model as it shows better forecasting plots and has better RMSE.

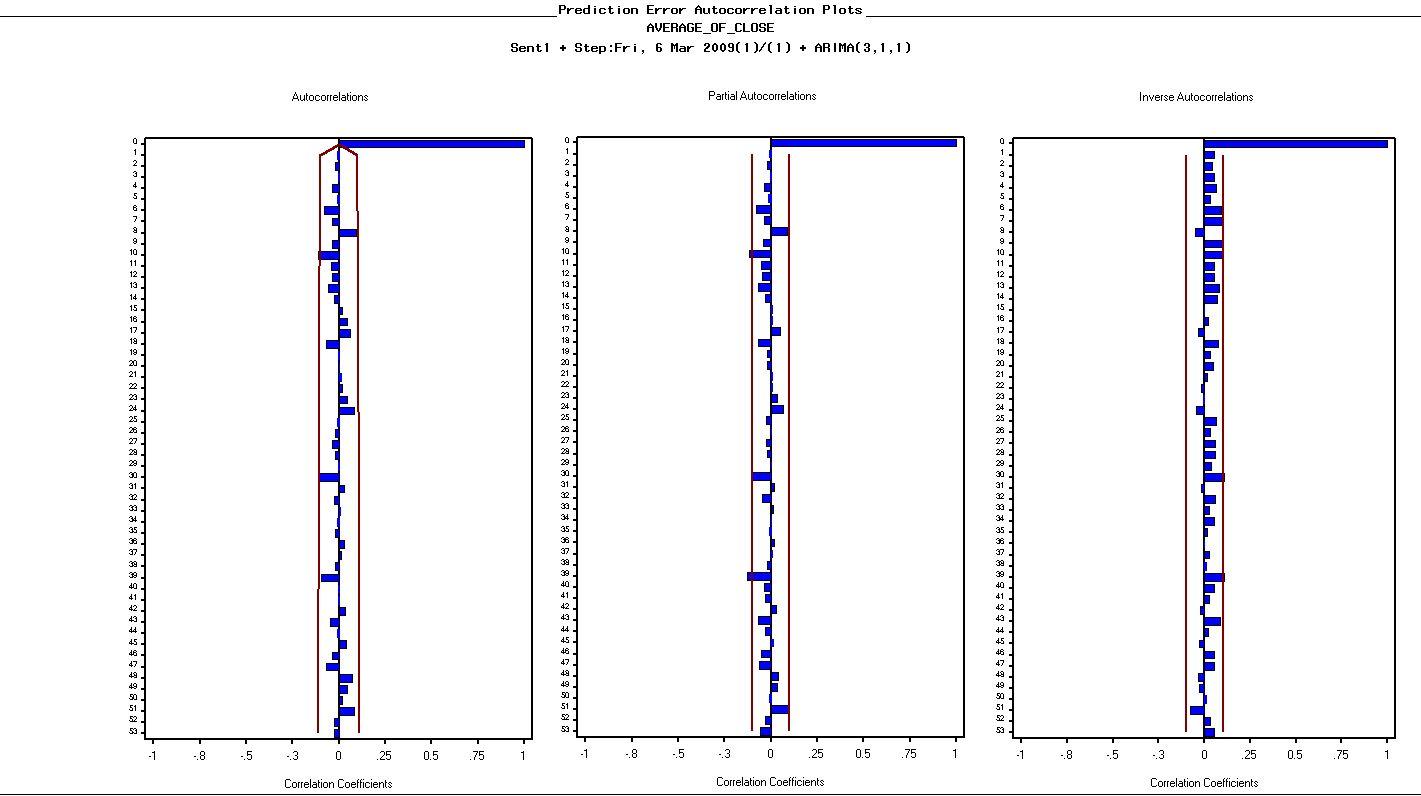
# Modelling - Forecasting with Sentiments

* Objective: To predict the average closing of DJIA stock data, now incorporating sentiments from reddit.

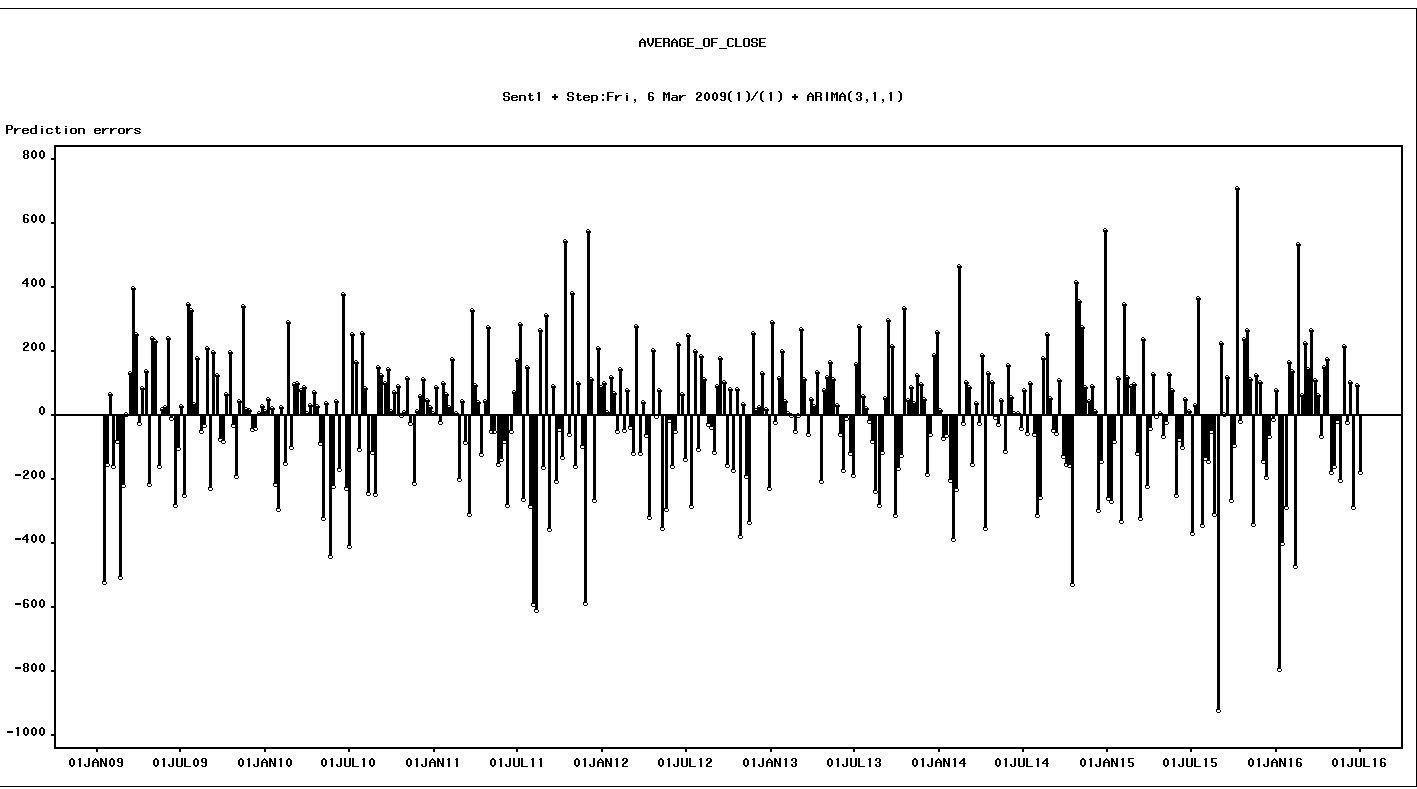
Primary 2 models stand out model with sentiment as regressor and Intervention at 6th March 2009. There was a drastic dip in the stock prices due to the recession and a model without intervention.

**Model 1**

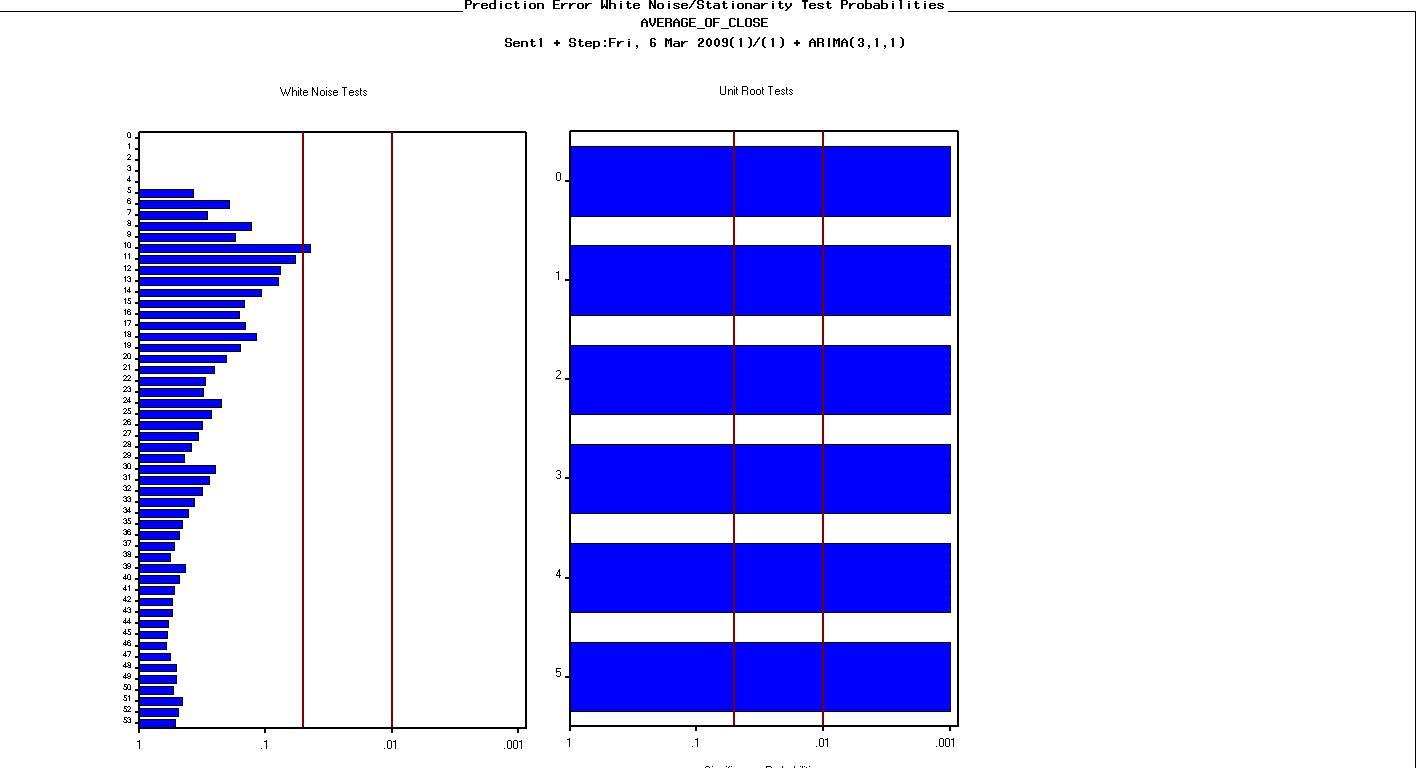
We have incorporated step intervention at 6th March 2009 and Sentiment 1 as regressor



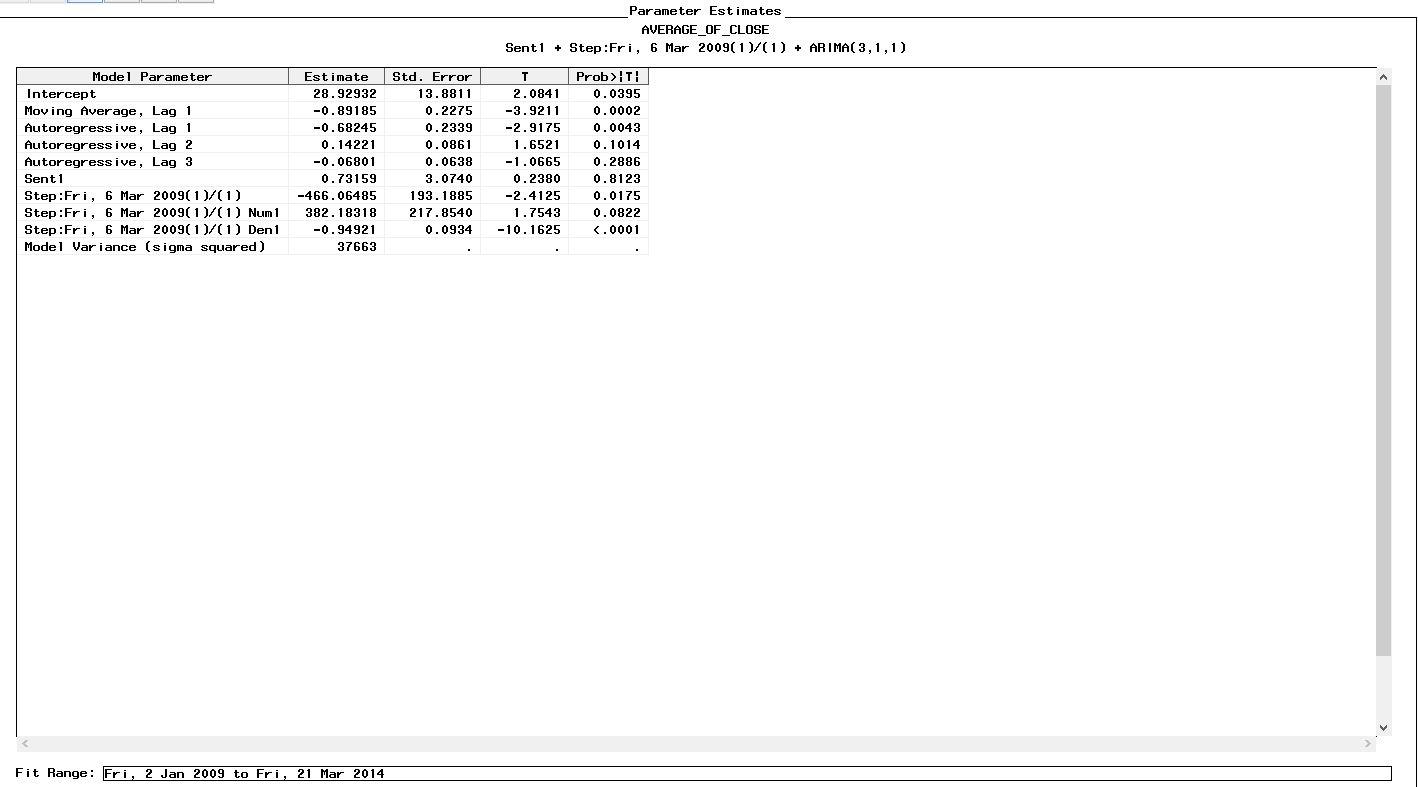
Prediction autocorrelation plots: The autocorrelation and correlation plots look reasonable with most values within the 95% confidence interval



Prediction errors: The prediction errors look symmetrical along the x axis apart from a few lags towards the end which may be due to an event like the previous models



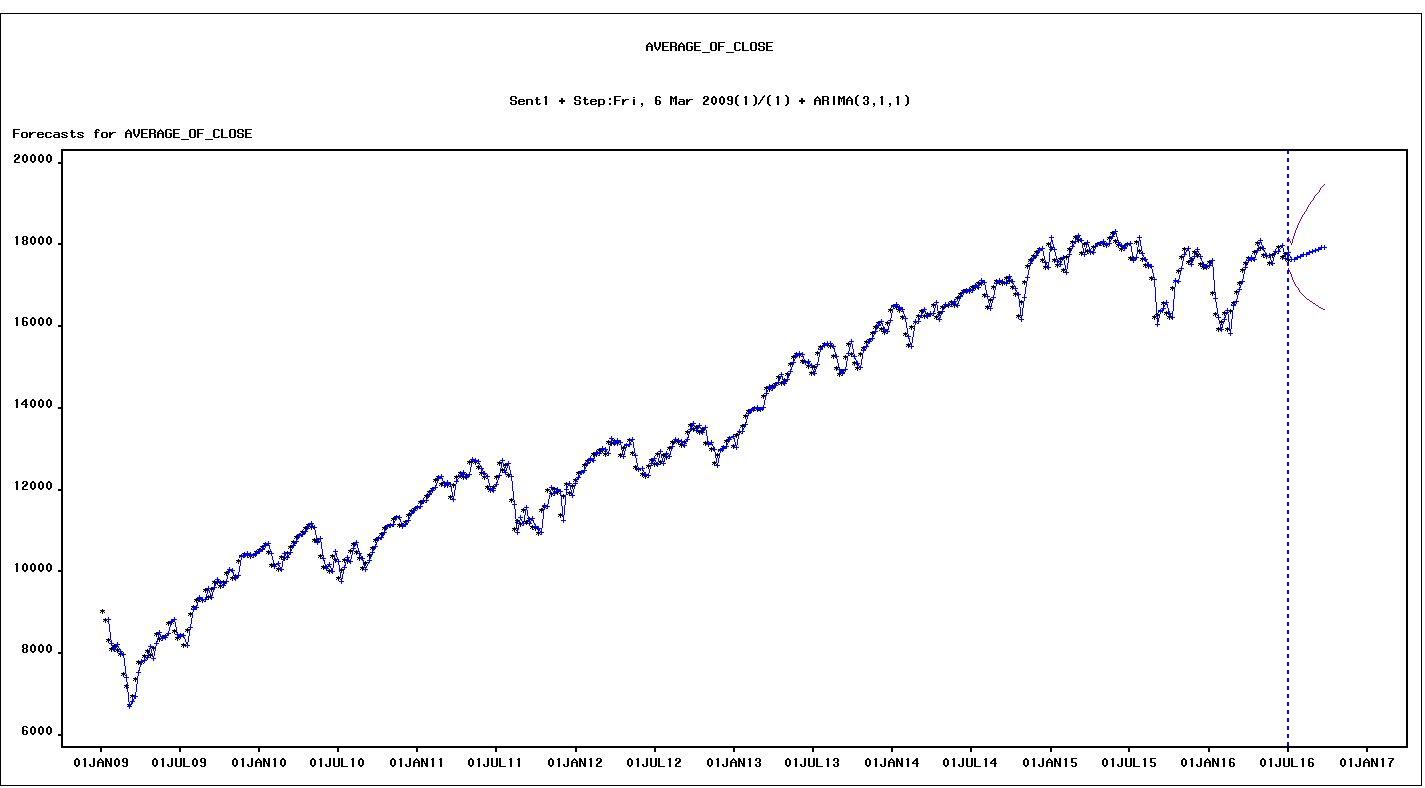
Prediction error white noise/ Unit root test: The white noise test is insignificant. And unit test indicates that the series is stationary.



Parameter estimates: The sentiment has an impact of .73 on the model.



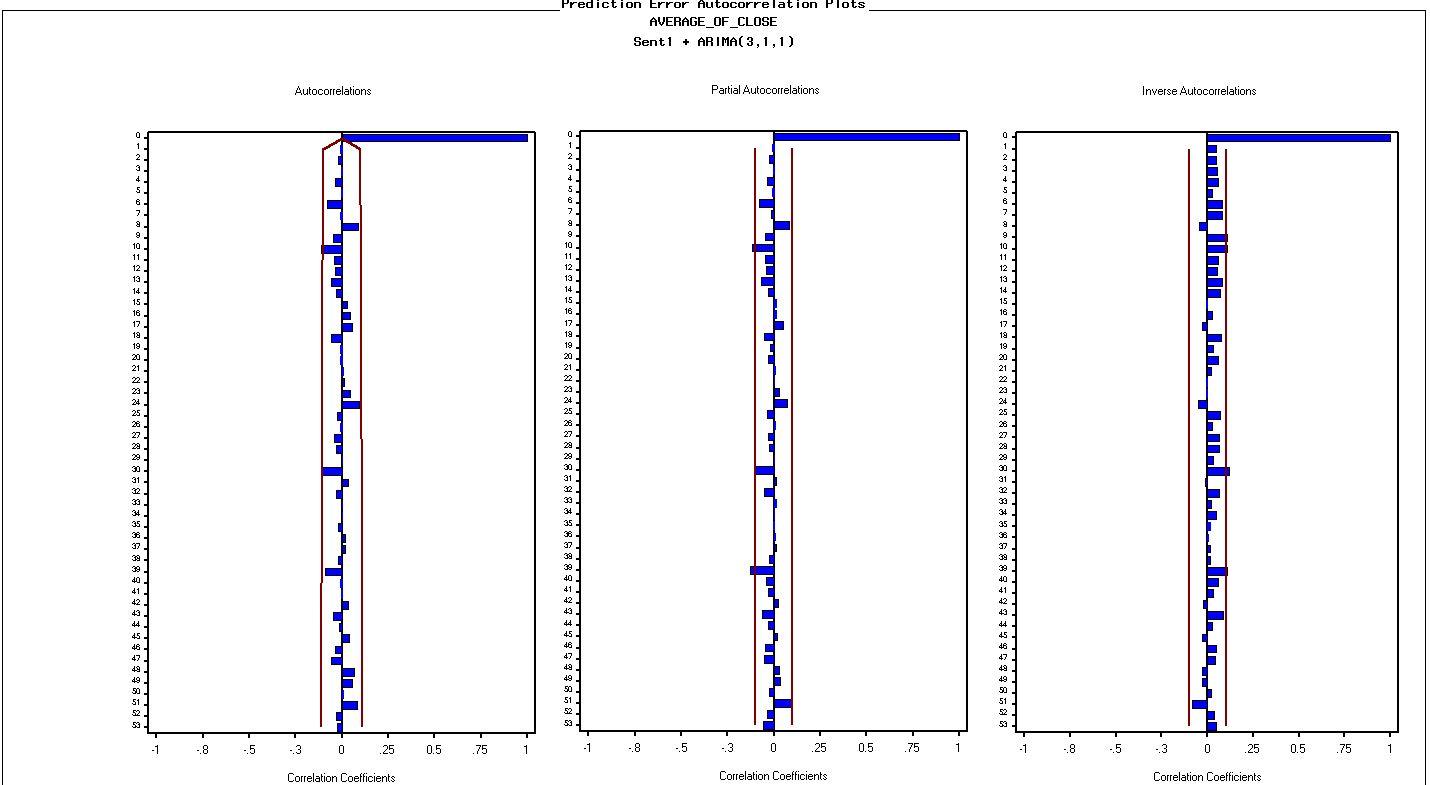
The Statistics of fit looks good



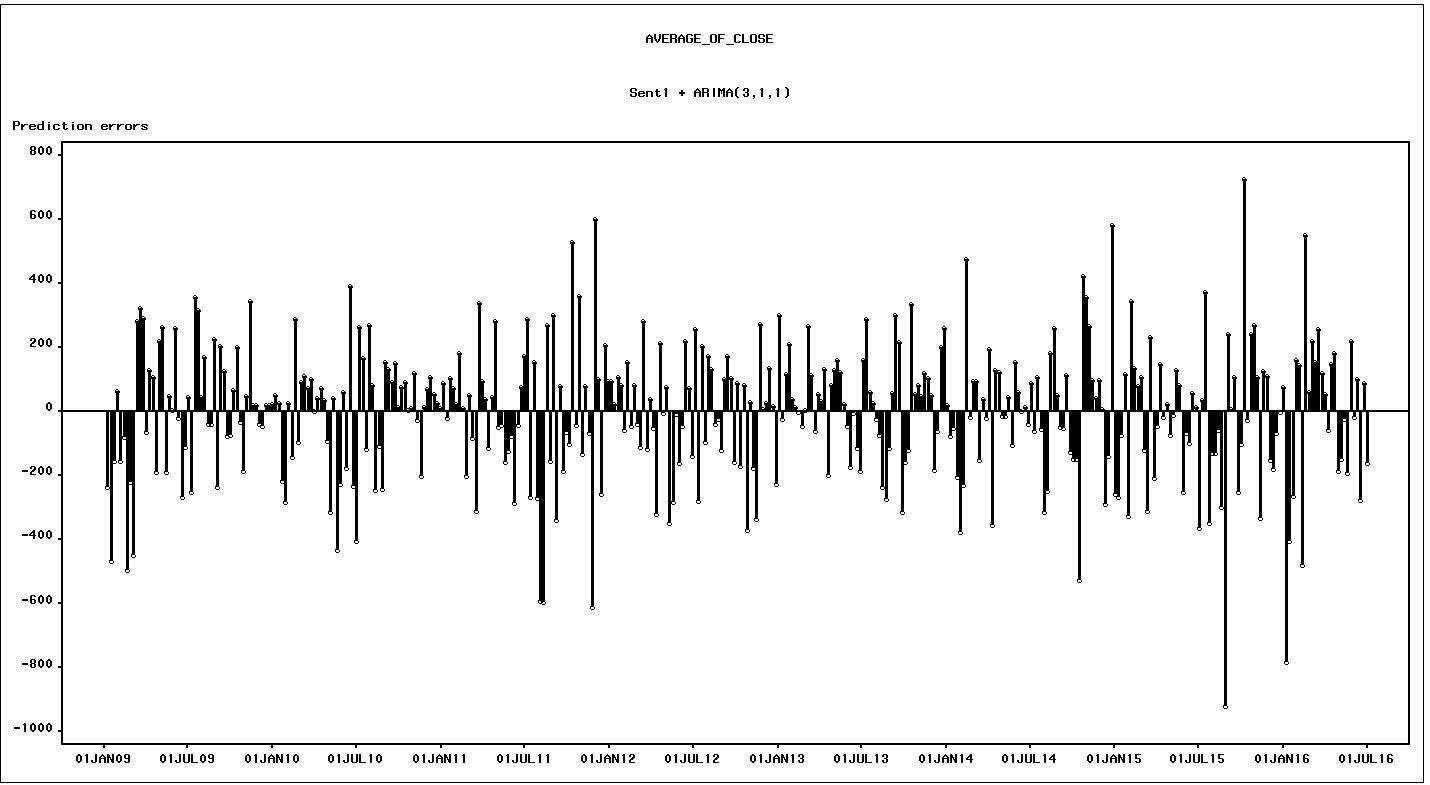
Forecast plot: The forecasted plot captures the trend well.

**Model** **2**

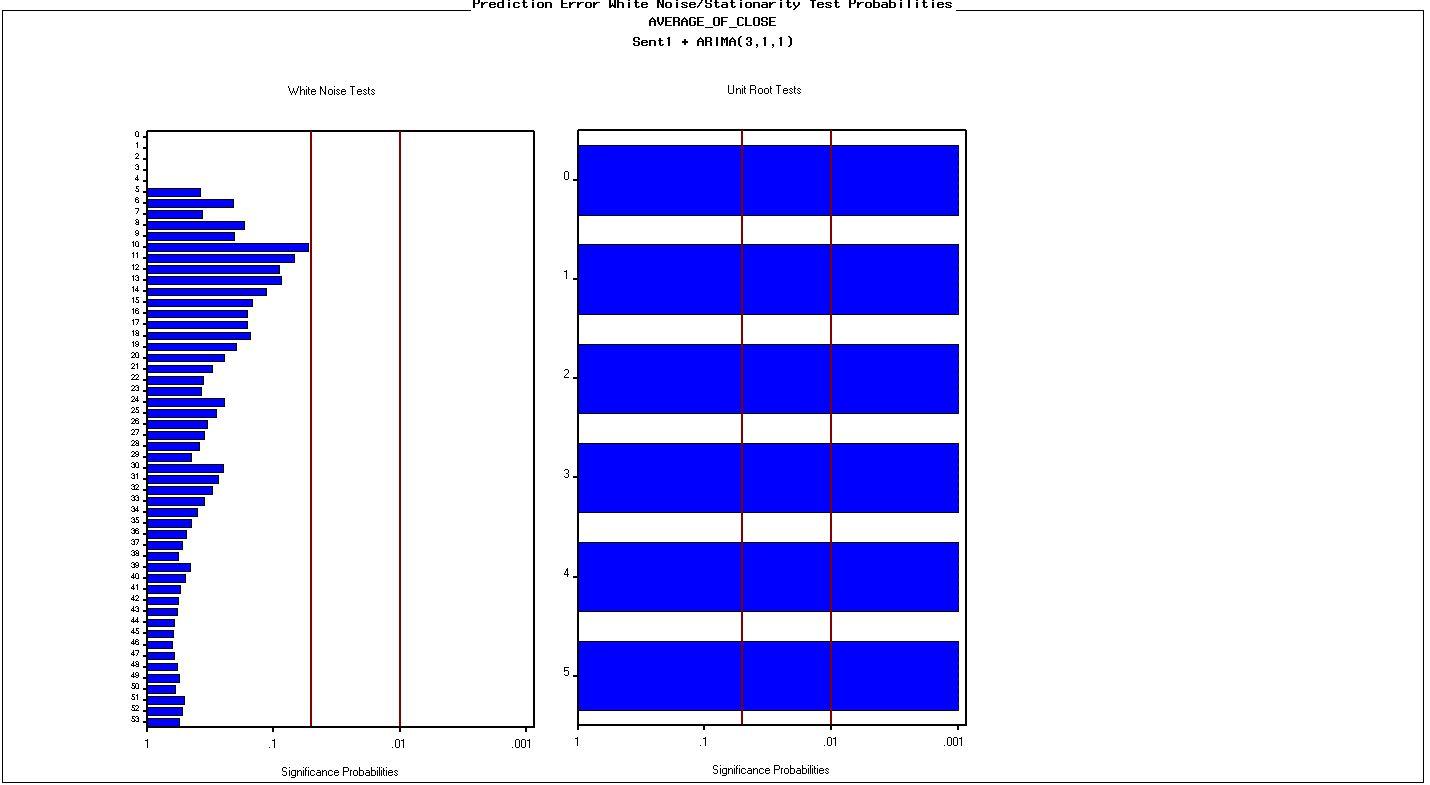
Analysing the model with sentiments as regressors however without interventions



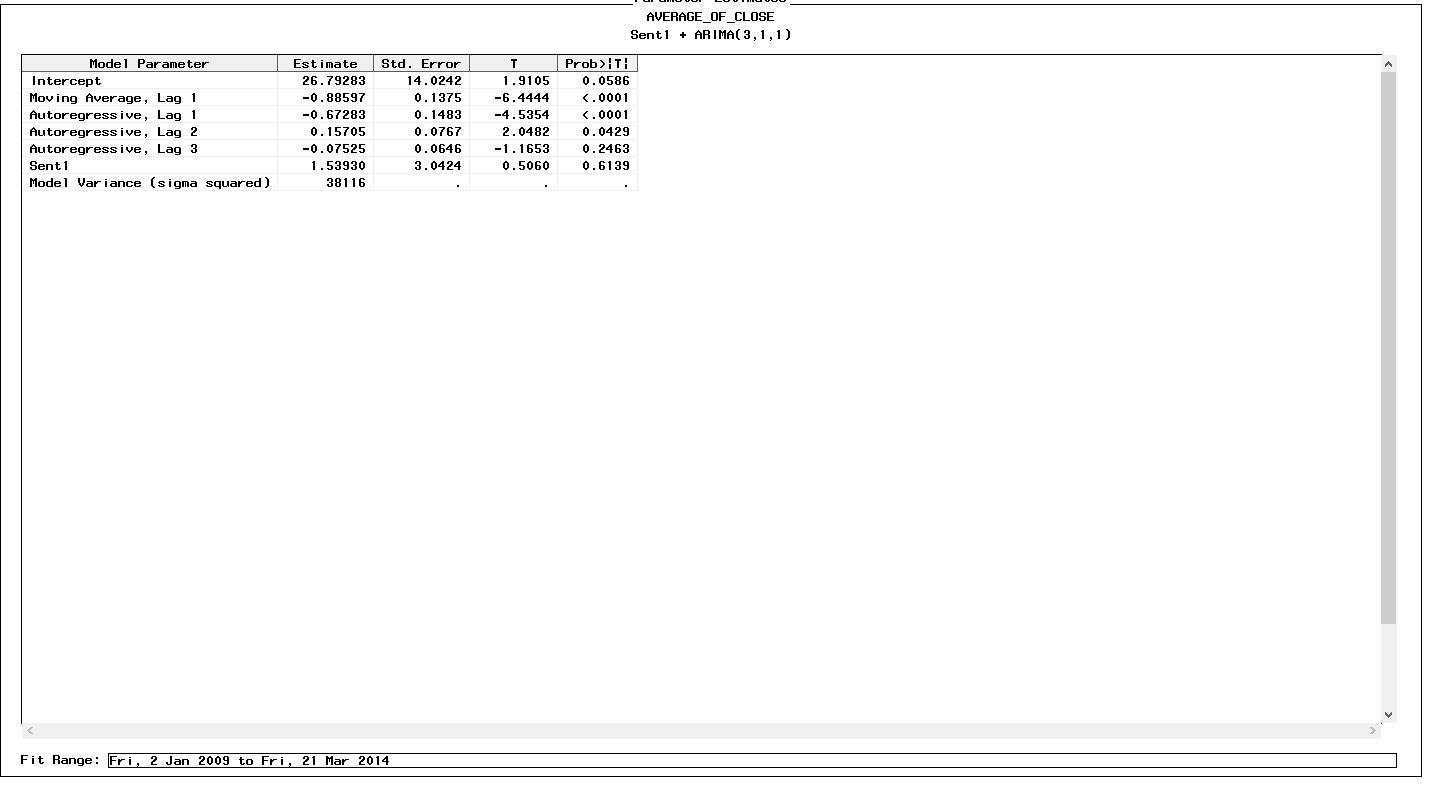
Prediction autocorrelation plots: The autocorrelation and correlation plots look reasonable with most values within the 95% confidence interval



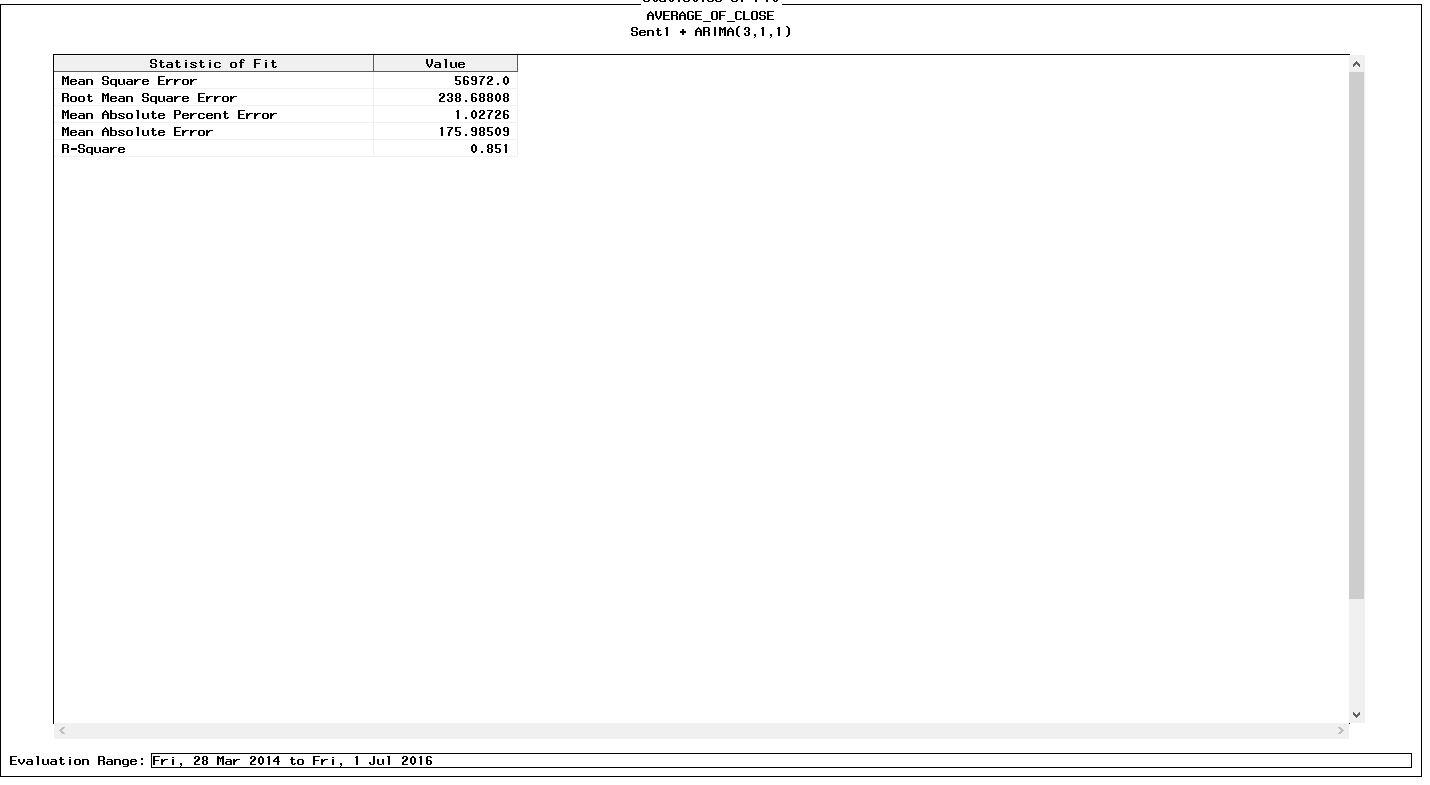
Prediction errors: The prediction errors look symmetrical along the x axis apart from a few lags towards the end which may be due to an event like the previous models



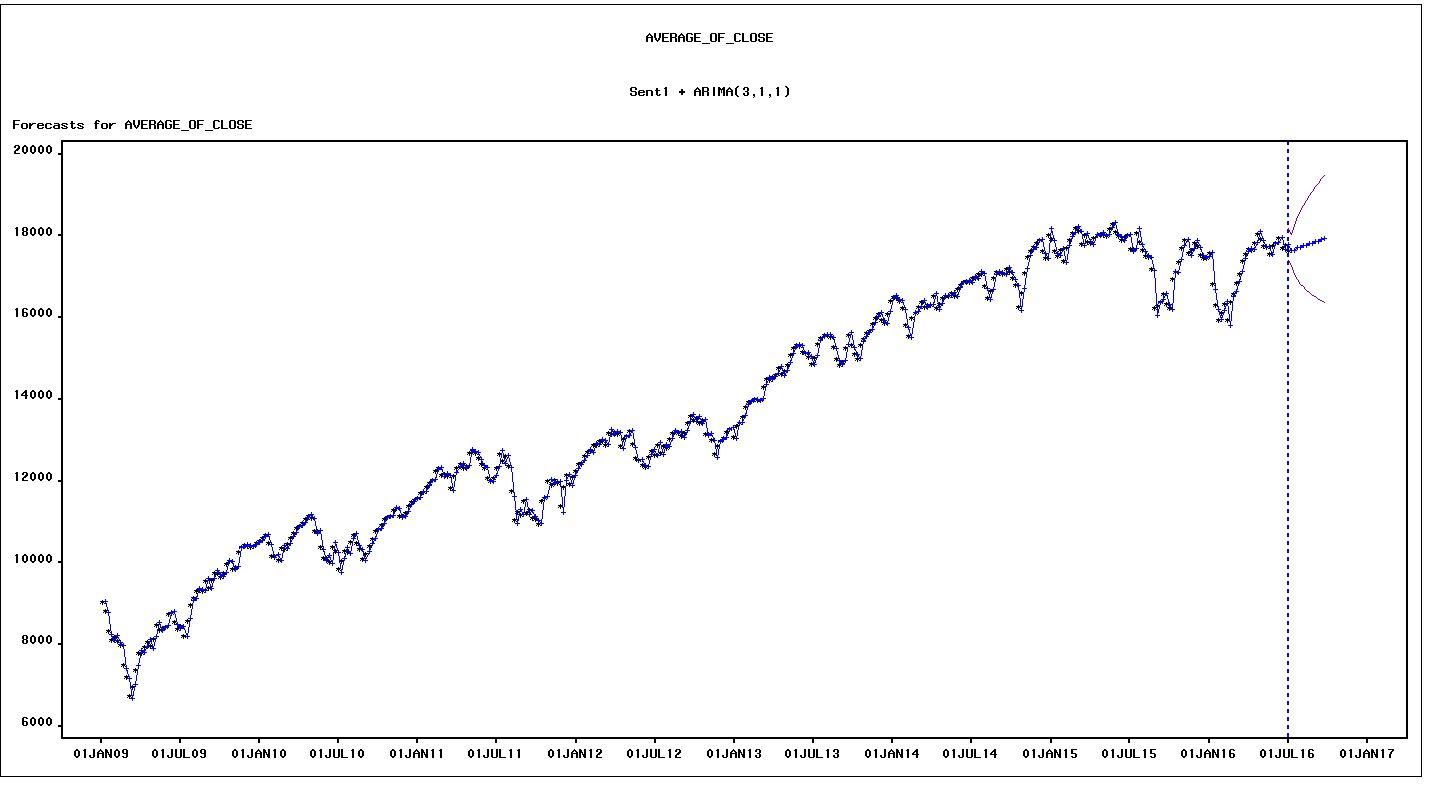
Prediction error white noise/ Unit root test: The white noise test is insignificant. And unit test indicates that the series is stationary.



Parameter estimates: The sentiment has an impact of 1.53 on the model.



The statistics of fit looks good



Forecast plot: The forecast plot captures the trend well

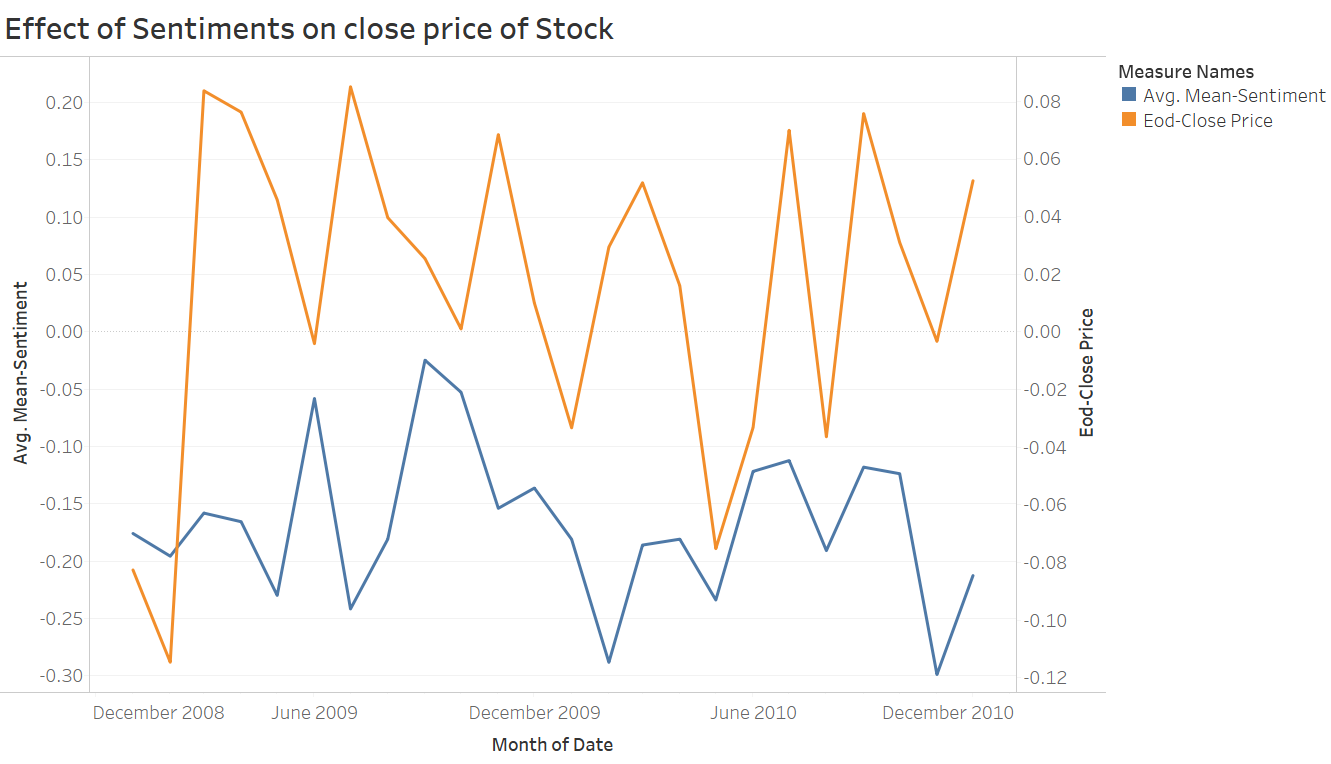
Conclusion: Based on the RMSE values we conclude that the model with intervention performs better.

# Comparison between Models in Part 1 and Part 2

# 

**Conclusion: We observe that the model with sentiments is better in terms of both RMSE and MAPE**

# Volume prediction

As seen from the graph below, the positive and negative sentiments have high correlation with the closing stock price. As the mean sentiment starts to decrease, the close price also tends to decrease and vice versa.

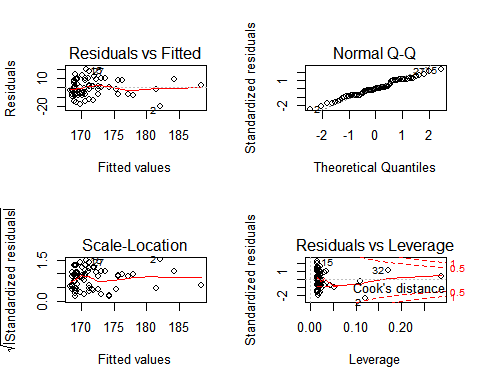
Our objective is to try and predict the volume of stock that will be traded based on the Google Trends obtained during the relevant time period.

Load the data set containing the log-volume of stocks traded and Google Trends during that time. Divide the data into train and test data set. Build a simple linear regression model.

DJIA\_Monthly\_VolTrend <- **read.csv**("~DJIA\_Monthly\_VolTrend.csv")  
djia.train = DJIA\_Monthly\_VolTrend[1:70,]  
djia.test = DJIA\_Monthly\_VolTrend[71:90,]  
  
reg1 = **lm**(Log.volume~DJIA.Trend, data=djia.train)  
**summary**(reg1)

##   
## Call:  
## lm(formula = Log.volume ~ DJIA.Trend, data = djia.train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.4927 -6.7835 -0.1204 8.0891 19.4154   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 163.69456 2.36267 69.284 < 2e-16 \*\*\*  
## DJIA.Trend 0.24726 0.06643 3.722 0.000403 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.652 on 68 degrees of freedom  
## Multiple R-squared: 0.1692, Adjusted R-squared: 0.157   
## F-statistic: 13.85 on 1 and 68 DF, p-value: 0.000403

Observe the error plots of the model to see if they are normally distributed or not

opar = **par**()   
**par**(mfrow=**c**(2,2))   
**plot**(reg1) 

Calculate the cooks distance to find the observations skewing the model.

z=**cooks.distance**(reg1)  
**round**(z,4)

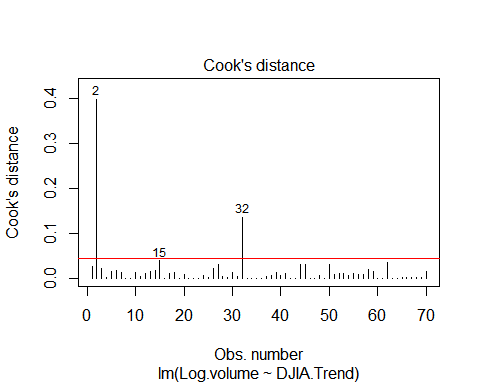
cutoff = 4/**nrow**(DJIA\_Monthly\_VolTrend)  
**length**(z[z>cutoff])

## [1] 2

**plot**(reg1,which = 4, cook.levels = cutoff)

**abline(**h=cutoff,col=”red”)

The plot shows the observation which are above the cooks distance cutoff. Build another model after removing the observations flagged above.



reg2 = **lm**(Log.volume~DJIA.Trend, data=djia.train[-**c**(2,15,32),])  
**summary**(reg2)

##   
## Call:  
## lm(formula = Log.volume ~ DJIA.Trend, data = djia.train[-c(2,   
## 15, 32), ])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.1914 -6.3723 -0.2104 7.4772 17.8346   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 162.52572 2.43230 66.820 < 2e-16 \*\*\*  
## DJIA.Trend 0.28105 0.07259 3.871 0.000254 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.073 on 65 degrees of freedom  
## Multiple R-squared: 0.1874, Adjusted R-squared: 0.1749   
## F-statistic: 14.99 on 1 and 65 DF, p-value: 0.0002536

It can be observed that the R2 clearly improved in the second model.

Let's have a look at AIC of both models.

**AIC**(reg1,reg2)

## df AIC  
## reg1 3 504.7169  
## reg2 3 473.9715

Model 2 definitely performs better than model 1. Let's make prediction on test data set based on our model 2. Add the predicted values as a column in our existing dataset

pred = **predict**(reg2,newdata = djia.test)  
djia.test$prediction = pred

Thus, we can conclude that, up to a certain extent we can predict the volume of stock to be traded during a given month based on the google trends.

# Business Applications

The main purpose of our project is to use text analytics results in improvement of prediction accuracy for stock prices. Both Trading Institutions and Individual Investors can be benefit from it.

1. **Trading Institutions**

Many traders take decisions in the financial market solely based on what other people think and what they recommend. By understanding the impact of the public sentiment, Trading Institutions can prepare the trend changes of the stock price in advance. Moreover, by set the opening price to an investor satisfied level, more and more investors will be attracted to invest. In this way, they can alter their investment strategies and portfolio to make the maximum profit.

1. **Individual Investors**

The use of text analytics provides new revenue streams for traditional publishing outlets and new sources of insight and efficiency for individual investors. The relationship between text and stock price allows information to move quicker and deeper. Common sentiments, such as happy and anger, of certain news can allow the individual investors to understand what the price trend will be in the future and allow them to change their own investment portfolios in order to gain the maximum profit.

# Conclusion

We have investigated the causative relation between public mood as measured from a large scale collection of news from reddit.com and the DJIA values. Our results show that firstly public mood can indeed be captured from the large-scale news feeds by means of simple natural language processing techniques. We haven’t been able to obtain high percentage result as expected, however we have obtained MAPE of about 1% using SAS Time Series Forecasting System. It is worth mentioning that our analysis doesn’t take into account many factors. Firstly, our dataset doesn’t really map the real public sentiment, it only considers the reddit using, english speaking people. It’s possible to obtain a higher correlation if the actual mood is studied. It may be hypothesized that people’s mood indeed affect their investment decisions, hence the correlation. But in that case, there’s no direct correlation between the people who invest in stocks and who use twitter more frequently, though there certainly is an indirect correlation - investment decisions of people may be affected by the moods of people around them, ie. the general public sentiment. All these remain as areas of future research.

# References

[1][*https://pdfs.semanticscholar.org/4ecc/55e1c3ff1cee41f21e5b0a3b22c58d04c9d6.pdf*](https://pdfs.semanticscholar.org/4ecc/55e1c3ff1cee41f21e5b0a3b22c58d04c9d6.pdf)

[2][*http://www.kdnuggets.com/2016/01/sentiment-analysis-predictive-analytics-trading-mistake.html*](http://www.kdnuggets.com/2016/01/sentiment-analysis-predictive-analytics-trading-mistake.html)

[3] [*https://www.kaggle.com/aaron7sun/stocknews*](https://www.kaggle.com/aaron7sun/stocknews)

[4][*https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html*](https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html)

[5][*https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf*](https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf)

[6][*https://trends.google.com/trends/explore?q=djia*](https://trends.google.com/trends/explore?q=djia)

[7][*https://sites.google.com/site/miningtwitter/questions/sentiment/sentiment*](https://sites.google.com/site/miningtwitter/questions/sentiment/sentiment)

[8] [*http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know*](http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know)

[9] [*https://www.econ.berkeley.edu/sites/default/files/Selene%20Yue%20Xu.pdf*](https://www.econ.berkeley.edu/sites/default/files/Selene%20Yue%20Xu.pdf)

# Appendix

**Stock Data**

- Before Data Processing



- After Data Processing

****

**News Data**

- Before Data Processing



- After Data Processing

