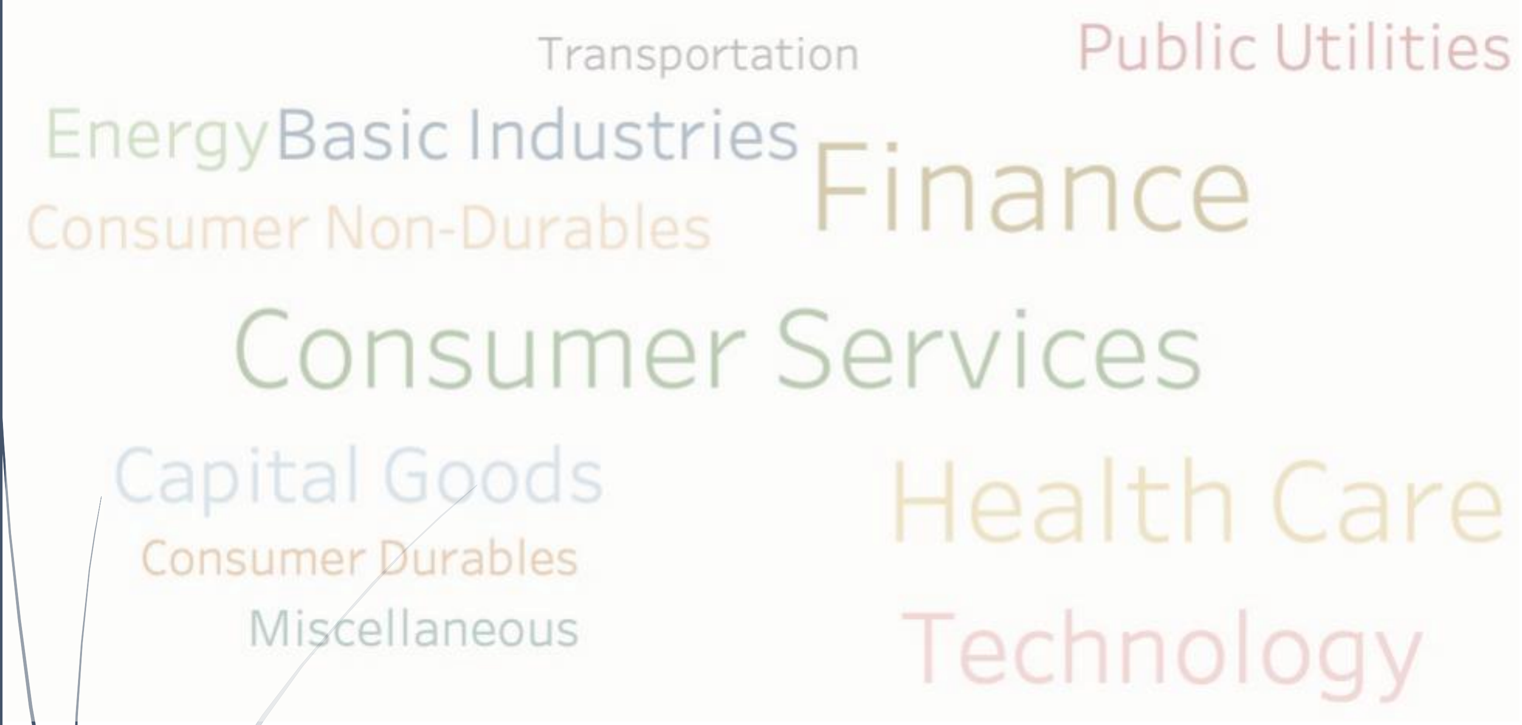
7/29/2019

# Stock Market Forecasting

Group 11



Jesse Arkebauer
Manav Bhalla
Ashish Mehra
Mandeep Singh Narang
Abhinav Saluja

# Table of Contents

Executive Summary ........ 3

1. Introduction ........ 4

2. Data Description ........ 4

3. Preprocessing Data ........ 4

3.1 Auto Regressive Integrated Moving Average (Exogenous) (ARIMA(X)) ........ 4

3.2 Unobservable Components Model (UCM) ........ 5

3.3 Exponential Smoothing Model (ESM) ........ 5

4. Exploratory Data Analysis ........ 5

5. Empirical Analysis ........ 9

5.1 ARIMA(X) ........ 12

5.2 Unobservable Components Model (UCM) ........ 13

5.2.1 Model Derivation and Assessment ........ 13

5.2.2 Model Prediction of Other Stocks & Sector Forecasting ........ 16

5.3 Exponential Smoothing Model (ESM) ........ 17

5.3.1 Model Derivation and Assessment ........ 17

5.3.2 Model Prediction of Other Stocks & Sector Forecasting ........ 18

6. Conclusions ........ 19

7. Citations ........ 20

8. Appendix ........ 21

8.1.1 3 Stages to Building and ARIMA(X) Model – Identification, Estimation and Diagnostics, Forecasting ........ 21

8.1.1.1 Identification ........ 21

8.1.1.2 Tests for Stationarity ........ 21

8.1.1.3 Identification of the Differenced/Log-transformed Series: ........ 22

8.1.1.4 Tests of Statistical Independence ........ 24

8.1.1.5 Estimation and Diagnostics ........ 24

8.1.1.6 Forecasting ........ 27

8.2 ESM ........ 28

# Executive Summary

Time series models can be applied to a myriad of forecasting challenges. In this project, we explored the use of such models for forecasting stocks and using the forecasts to predict the performance of stock sectors. Specifically, we used the ARIMA(X), UCM and ESM models to predict the performance of fifteen stocks in the Oil and Gas, Public Utilities and Transportation sectors, and using the predicted stocks performance we predicted an overall forecast for the three sectors.

We found that the performance of the models varied from highly accurate (0% prediction error) to highly inaccurate (~40% error), although the overall prediction of the trend (up or down) for the sectors was roughly 66% accurate. We also observed that the volatility of the sectors impacted the accuracy of the predictions. In addition to this, we found the inclusion of variables such as price of oil improved the accuracy of our models.

Based on model performance, we conclude the ESM model performed the best, although we also conclude that using these models to decide on investment decisions would be a highly risky endeavor. Human emotion remains a key driver/influencer of the stock market performance and it is difficult to account for this using computational models. Nevertheless, these models can serve as a valuable tool in an investor's tool belt to provide an additional source of data when making investment decisions.

# 1. Introduction

The group project for Predictive Analytics required students to select the task of either forecasting stock market data, or car sale prices. For our project, we decided to forecast the stock market, being motivated by the following key factors:

- Real world applicability
- Quality of Data
- Expansion Potential
- Dynamic Nature of Data

# 2. Data Description

The data available comprises two categories - stocks and ETFs. With these categories, the attributes are consistent and are made up of:

- Date
- Open
- High
- Low
- Close
- Volume
- OpenInt

There are 7195 stocks and 1344 ETFs contained within the raw data files.

# 3. Preprocessing Data

Given the various models we plan to use for forecasting, the pre-processing required for the data was largely driven by the model selected.

## 3.1 Auto Regressive Integrated Moving Average (Exogenous) (ARIMA(X))

The ARIMA model is made up of three main parts: The Auto Regressive (p), the Integrated (d), and the Moving Average (q). These three components form our *ARIMA (p,d,q)* model, which is a linear combination of past values and past errors to determine the future value of a variable. ARIMAX allows for the addition of exogenous variables.

1. The Auto Regressive portion of the model will analyze the historical closing price of stocks to predict the current and future closing price of stocks.
2. The Moving Average portion of the model analyzes past errors to predict future errors. This is a key part of the model, since the AR element analyzes what is explained in our model, the MA element analyzes what is unexplained.
3. The Integrated portion of the model provides us with a way of making our data stationary so that we can draw meaningful insights from it.

4. The three stages of building an ARIMA model include identification, estimation and diagnostics, and forecasting.

### 3.2 Unobservable Components Model (UCM)

The UCM model allows for the inclusion of additional variables to be considered when forecasting the dependent variable. In our case, we included the price of gold, oil and natural gas as part of the model. However, there were certain days that had missing data, which created errors when running the UCM model. Therefore, we used the time series procedure to impute/fill the missing values with the value of the preceding day.

The following process was followed for the pre-processing of data.

1. Import variables gold, oil and natural gas
2. Perform a left join, using date as the primary key
3. Impute missing data using the time series procedure
4. Import stock data, and join with the other variables to create one table to be used for modelling
5. As part of the model, we also restricted the data to 2005-2017. Modelling an unnecessarily large dataset was not expected to yield better results, and the range specified captured the preceding years of the import GFC.

### 3.3 Exponential Smoothing Model (ESM)

The idea of exponential smoothing is that often, time series evolve in such a way that the level, trend, or seasonality changes slowly over time. This renders older data less relevant for forecasting than more recent data. For our ESM model, differencing was the only preprocessing required for the stock market data.

The following process was followed for the pre-processing of data.

1. Import different files containing Stocks data
2. Impute missing data using the time series procedure
3. Aggregated all the data at Weekday level
4. As part of the model, we also restricted the data to 2005-2017. Modelling an unnecessarily large dataset was not expected to yield better results.

## 4. Exploratory Data Analysis

As a preliminary part of our analysis in order to analyze and better understand our data, we perform a decomposition of series into its various components as shown below:

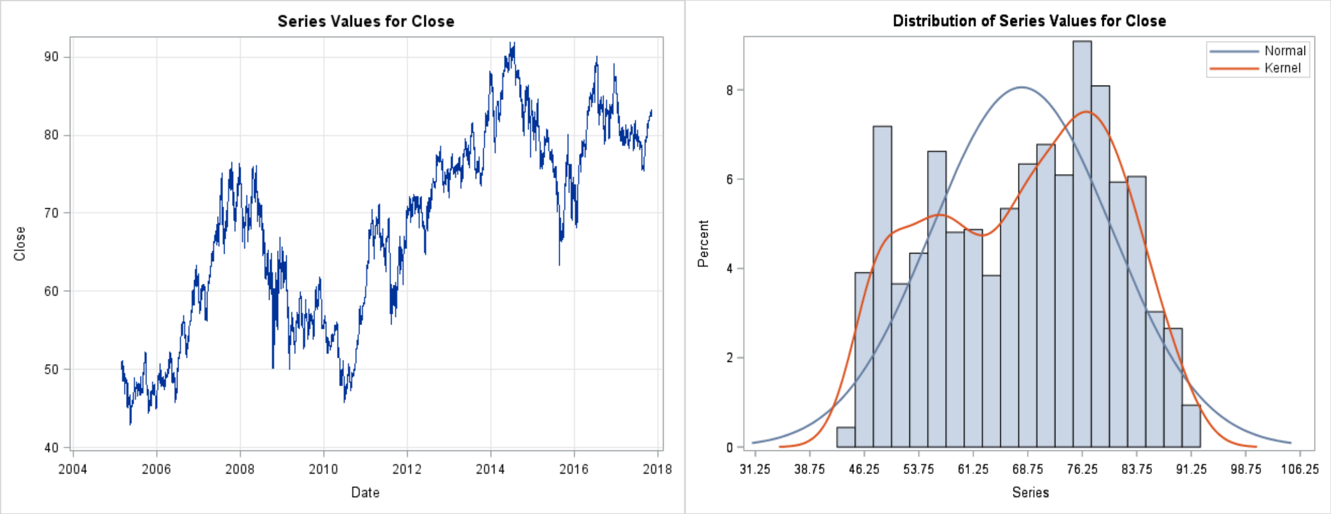


*Figure 1: Distribution of Close*

The above charts show the trend of the series along with its distribution. As we observe the trend, we can see a sharp decrease in the stock prices post 2008 which is due to the financial crisis pertaining to that year, post which the series shows an increasing trend in from 2010 onwards. The distribution of series is positively skewed due to the reason of sharp increase in stock prices past 2010.

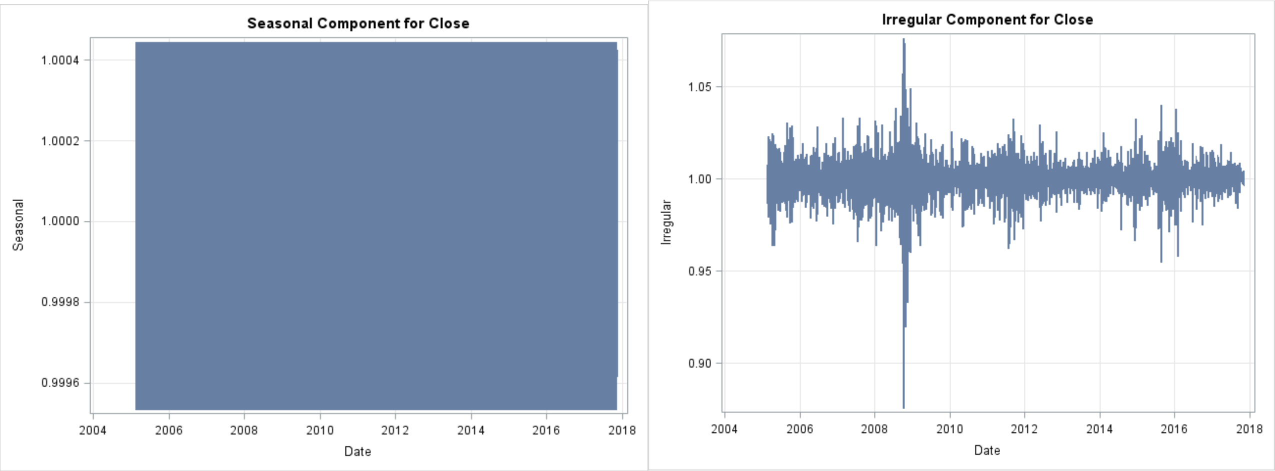


*Figure 2: Season and Irregular Component*

Looking at the above charts we can infer that there is not a seasonal component present within our series which confers with our economic theory as stock prices mostly follow a random walk and do not contain a seasonal component. On the other hand the irregular component seems to be oscillating around the value of 1 with surges indicating shocks within the stock prices in some years including the biggest one in 2008.

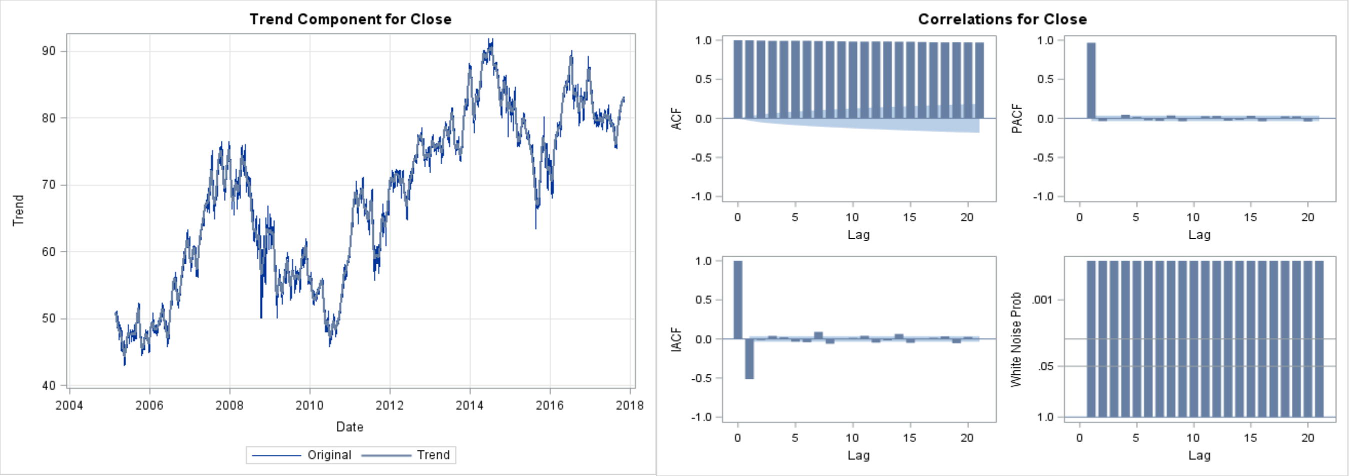
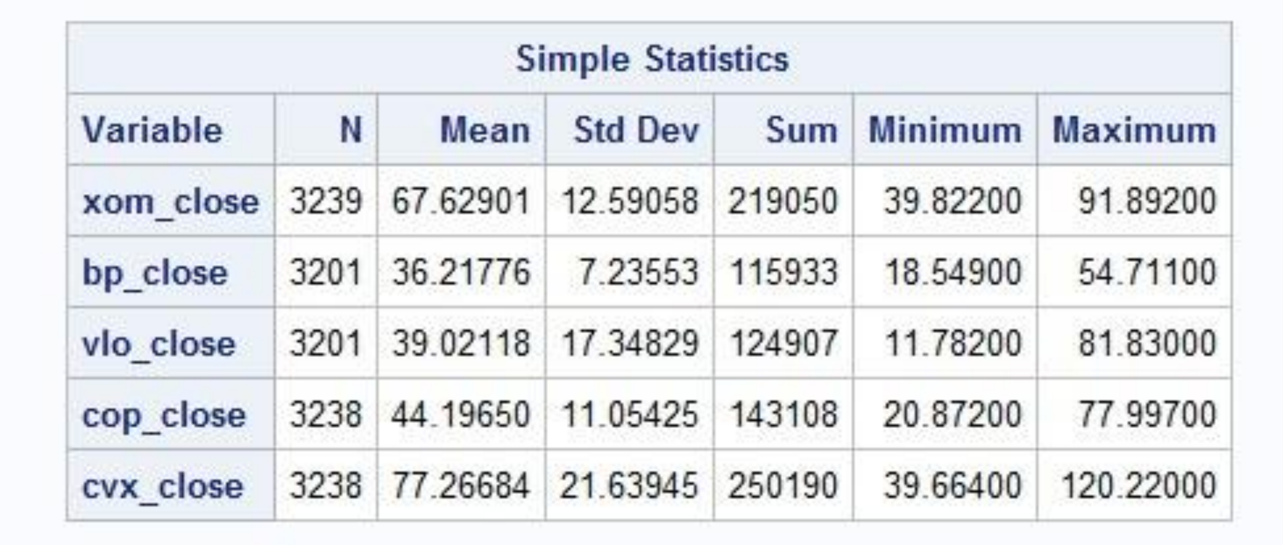


*Figure 3: Trend and Correlations for Close*

From the trend above we see a rising trend from 2006 to 2008 post which there is a steep decline due to the 2008 market crisis, after which the oil prices seem to have an increasing trend with another sharp decline in 2015-16 due to declining oil commodity prices.

From our correlation charts and the white noise probability chart we can clearly infer that our current series is not white noise and there is high autocorrelation within our data and further, our data is not stationary. We will further deep dive into this below.

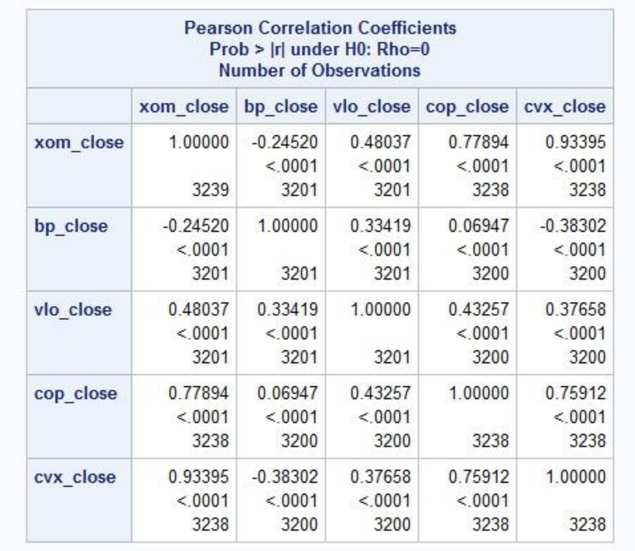


Simple Statistics

| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum |
|---|---|---|---|---|---|---|
| xom_close | 3239 | 67.62901 | 12.59058 | 219050 | 39.82200 | 91.89200 |
| bp_close | 3201 | 36.21776 | 7.23553 | 115933 | 18.54900 | 54.71100 |
| vlo_close | 3201 | 39.02118 | 17.34829 | 124907 | 11.78200 | 81.83000 |
| cop_close | 3238 | 44.19650 | 11.05425 | 143108 | 20.87200 | 77.99700 |
| cvx_close | 3238 | 77.26684 | 21.63945 | 250190 | 39.66400 | 120.22000 |

*Figure 4: High level statistics for Oil Stocks*

From simple statistics, it is clear that Chevron(CVX) and Exxon Mobil(XOM) are the top two oil stocks in terms of closing stock price. On the other hand, the lowest average closing stock price is that of British Multinational Oil and Gas Co (BP) and Valero Energy Corporation (VLO).



Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

| | xom_close | bp_close | vlo_close | cop_close | cvx_close |
|---|---|---|---|---|---|
| xom_close | 1.00000<br><br>3239 | -0.24520<br><.0001<br>3201 | 0.48037<br><.0001<br>3201 | 0.77894<br><.0001<br>3238 | 0.93395<br><.0001<br>3238 |
| bp_close | -0.24520<br><.0001<br>3201 | 1.00000<br><br>3201 | 0.33419<br><.0001<br>3201 | 0.06947<br><.0001<br>3200 | -0.38302<br><.0001<br>3200 |
| vlo_close | 0.48037<br><.0001<br>3201 | 0.33419<br><.0001<br>3201 | 1.00000<br><br>3201 | 0.43257<br><.0001<br>3200 | 0.37658<br><.0001<br>3200 |
| cop_close | 0.77894<br><.0001<br>3238 | 0.06947<br><.0001<br>3200 | 0.43257<br><.0001<br>3200 | 1.00000<br><br>3238 | 0.75912<br><.0001<br>3238 |
| cvx_close | 0.93395<br><.0001<br>3238 | -0.38302<br><.0001<br>3200 | 0.37658<br><.0001<br>3200 | 0.75912<br><.0001<br>3238 | 1.00000<br><br>3238 |

*Figure 5: Correlation Coefficients*

It is interesting to note the correlation coefficients between these stocks, as presented in the table above. We see that Exxon and Chevron have a high degree of correlation. Given their brand reputation, one can imagine that ExxonMobil and Chevron are two high-quality stocks that are well worth owning, at the right price. The reason for high correlation between them (as noted in Figure 5) may be because both Exxon and Chevron offer very safe returns, which makes them excellent for risk averse investors.

Furthermore, the reason for low correlation between COP and BP may be because ConocoPhillips is an independent oil and gas exploration and production company whereas BP is an integrated oil major. When the price of oil collapsed, the independent drillers faced major problems as compared to integrated oil majors.

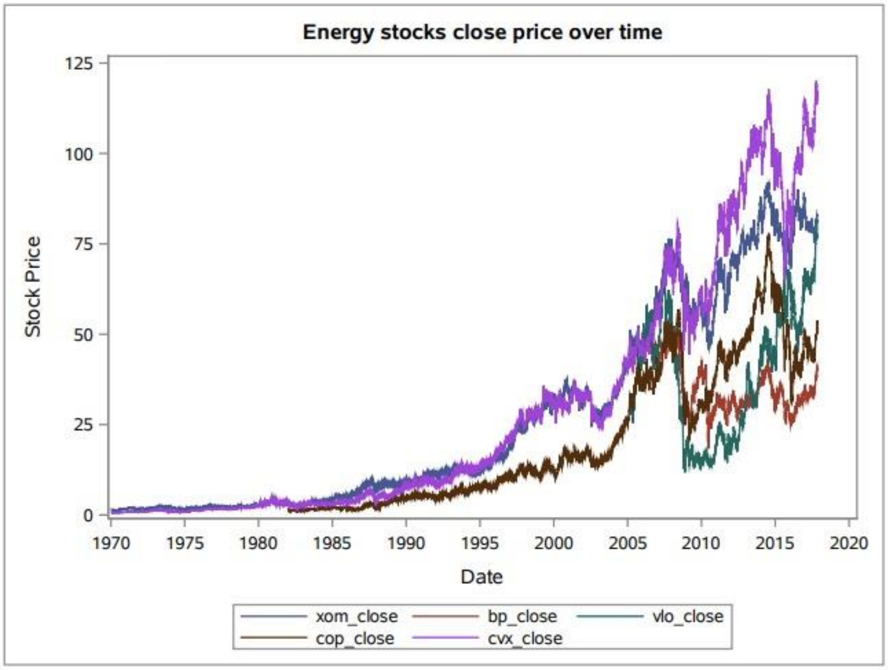


*Figure 6: Oil and Gas Stocks*

Charting the Oil and Gas stocks' close price, we note that they roughly follow the same overall trend up or down. This serves as the basis for our hypothesis that if we can predict the direction these stocks are heading, then we can ascertain the direction for the overall sector.

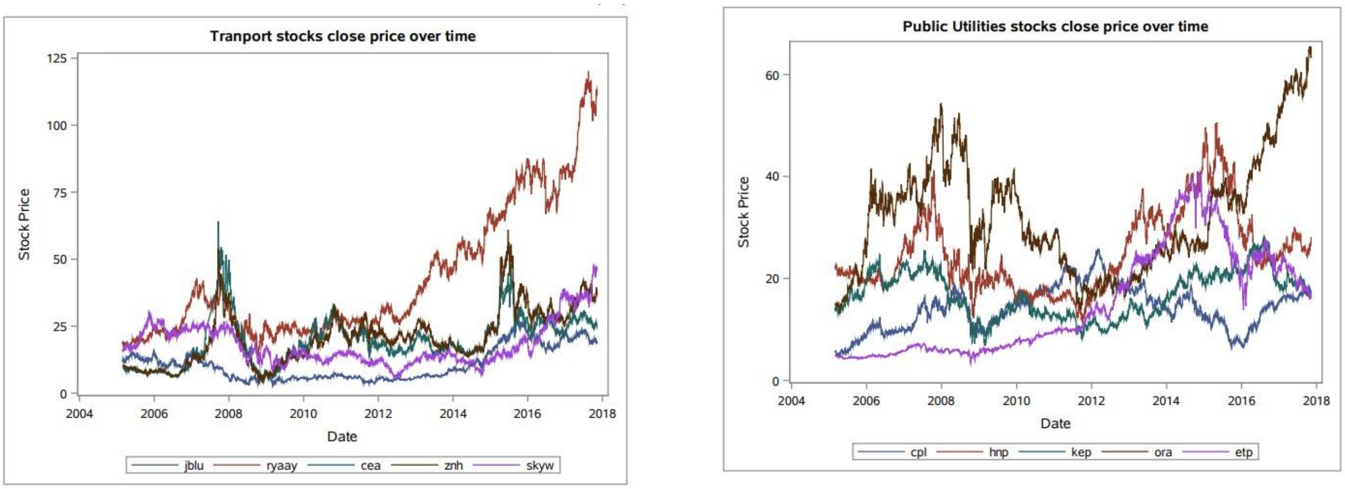


*Figure 7: Transport and Public Sector Stocks*

In contrast to the Oil and Gas sector, we see that the transport sector has much more disparate trends in the stocks. We see that the Closing Stock Price of China Eastern Airlines (CEA) is highly volatile. The closing stock price of China Southern Airlines (ZNH) starts from the bottom and reaches its peak around 2015.

The only stock that remains constant over this period is JetBlue Airways (JBLU). This means the stock volatility is low and there is low amount of risk involved for this stock. On average, The Closing Stock Price remains low for all the companies except for Ryanair Holdings (RYAAY) which is consistently increasing after year 2012.

As with the transport sector, we observe that there is no obvious pattern that can be discerned from the close price of the public utilities stocks. It will be interesting to observe how these sectors are predicted given the overall volatility.

## 5. Empirical Analysis

Given the large data size, the analysis carried out was in the context of an investor looking at three sectors, namely Oil and Gas, Public Utilities and Transportation. In reviewing the time-series plots from above, we form the following hypotheses:

1. Energy stocks tend for follow a very similar trend, therefore should yield the highest accuracy in predictability.
2. The stocks chosen to analyze in the transportation sector and public utilities sector do not appear to follow a trend, as compared to energy stocks, and therefore we anticipate a much higher degree of variance in the accuracy of their predictions.

In the following table, Table 1: Stock and Sector Analysis Summary, we display the performance of our analysis and forecast for each sector and stock. The primary KPI we focused our analysis on is whether an investor would buy, sell or hold a stock, based on the 11/10/17 close price, against our forecasted 11/30/17 close price. Using various model estimation parameters (as noted in 5.1-5.3), the best-fit models are listed below.

| Sector | Stock | 11/10 Value | 11/30 Actual | Model | AIC | BIC | Predicted | Prediction Error | Predicted Direction | Guidance Correct? |
|---|---|---|---|---|---|---|---|---|---|---|
| Oil | XOM | $82.94 | $83.29 | ARIMA | 8746 | 8764 | $83.14 | -0.18% | Buy | Correct |
| | | | | ARIMAX | 8066 | 8090 | $83.14 | -0.18% | Buy | Correct |
| | | | | UCM | 8069 | 8118 | $83.12 | -0.21% | Buy | Correct |
| | | | | ESM | -342 | -323 | $83.08 | -0.25% | Buy | Correct |
| | COP | $52.99 | $50.88 | ARIMA | 7690 | 7702 | $53.13 | 4.42% | Buy | Not Correct |
| | | | | ARIMAX | 6787 | 6811 | $53.15 | 4.47% | Buy | Not Correct |
| | | | | UCM | 6770 | 6818 | $53.17 | 4.50% | Buy | Not Correct |
| | | | | ESM | -1427 | -1409 | $53.10 | 4.36% | Buy | Not Correct |
| | CVX | $117.18 | $118.99 | ARIMA | 9962 | 9974 | $117.61 | -1.16% | Buy | Correct |
| | | | | ARIMAX | 9103 | 9127 | $117.60 | -1.17% | Buy | Correct |
| | | | | UCM | 9094 | 9142 | $117.39 | -1.34% | Buy | Correct |
| | | | | ESM | 870 | 888 | $117.64 | -1.13% | Buy | Correct |
| | BP | $40.30 | $40.07 | ARIMA | 5905 | 5923 | $40.31 | 0.60% | Buy | Not Correct |
| | | | | ARIMAX | 5248 | 5272 | $40.32 | 0.63% | Buy | Not Correct |
| | | | | UCM | 5226 | 5274 | $40.45 | 0.96% | Buy | Not Correct |
| | | | | ESM | -3163 | -3145 | $40.07 | 0.00% | Sell | Correct |
| | VLO | $81.37 | $85.62 | ARIMA | 8340 | 8352 | $81.69 | -4.59% | Buy | Correct |
| | | | | ARIMAX | 8174 | 8198 | $81.69 | -4.59% | Buy | Correct |
| | | | | UCM | 8176 | 8224 | $81.47 | -4.84% | Buy | Correct |
| | | | | ESM | -736 | -718 | $81.91 | -4.33% | Buy | Correct |
| Public Utilities | CPL | $16.75 | $11.96 | ARIMA | 1550 | 1562 | $16.83 | 40.72% | Buy | Not Correct |
| | | | | ARIMAX | 1337 | 1361 | $16.84 | 40.78% | Buy | Not Correct |
| | | | | UCM | 1358 | 1406 | $16.80 | 40.43% | Buy | Not Correct |
| | | | | ESM | -7529 | -7511 | $16.80 | 40.48% | Buy | Not Correct |
| | ET | $17.47 | $16.20 | ARIMA | 2215 | 2227 | $16.63 | 2.63% | Sell | Correct |
| | | | | ARIMAX | 2083 | 2107 | $16.63 | 2.65% | Sell | Correct |
| | | | | UCM | 2102 | 2150 | $16.61 | 2.56% | Sell | Correct |
| | | | | ESM | -6811 | -6793 | $16.40 | 1.23% | Sell | Correct |
| | HNP | $28.10 | $25.65 | ARIMA | 6242 | 6254 | $28.14 | 9.69% | Buy | Not Correct |
| | | | | ARIMAX | 6185 | 6209 | $28.11 | 9.57% | Buy | Not Correct |
| | | | | UCM | 6176 | 6224 | $27.89 | 8.73% | Sell | Correct |
| | | | | ESM | -2811 | -2793 | $28.06 | 9.39% | Sell | Correct |
| | KEP | $17.31 | $17.45 | ARIMA | 2261 | 2273 | $17.33 | -0.69% | Buy | Correct |
| | | | | ARIMAX | 2190 | 2215 | $17.33 | -0.70% | Buy | Correct |
| | | | | UCM | 2151 | 2200 | $17.40 | -0.29% | Buy | Correct |
| | | | | ESM | -6815 | -6797 | $17.28 | -0.97% | Sell | Not Correct |
| | ORA | $63.25 | $65.55 | ARIMA | 7289 | 7302 | $63.55 | -3.05% | Buy | Correct |
| | | | | ARIMAX | 7138 | 7163 | $63.57 | -3.02% | Buy | Correct |
| | | | | UCM | 7163 | 7212 | $63.30 | -3.44% | Buy | Correct |
| | | | | ESM | -1783 | -1765 | $63.74 | -2.76% | Buy | Correct |
| Transportation | JBLU | $18.99 | $21.47 | ARIMA | 1233 | 1245 | $19.04 | -11.33% | Buy | Correct |
| | | | | ARIMAX | 1211 | 1235 | $19.03 | -11.35% | Buy | Correct |
| | | | | UCM | 1219 | 1268 | $18.98 | -11.59% | Sell | Not Correct |
| | | | | ESM | -7839 | -7821 | $19.04 | -11.31% | Buy | Correct |
| | RYAAY | $111.25 | $121.94 | ARIMA | 8208 | 8220 | $111.85 | -8.28% | Buy | Correct |
| | | | | ARIMAX | 8210 | 8235 | $111.83 | -8.29% | Buy | Correct |
| | | | | UCM | 8204 | 8253 | $111.29 | -8.73% | Buy | Correct |
| | | | | ESM | -914 | -895 | $112.14 | -8.04% | Buy | Correct |
| | SKYW | $45.90 | $52.05 | ARIMA | 4367 | 4379 | $46.04 | -11.55% | Buy | Correct |
| | | | | ARIMAX | 4366 | 4390 | $46.04 | -11.54% | Buy | Correct |
| | | | | UCM | 4387 | 4435 | $45.82 | -11.97% | Sell | Not Correct |
| | | | | ESM | -914 | -895 | $46.29 | -11.07% | Buy | Correct |
| | CEA | $25.17 | $29.51 | ARIMA | 8173 | 8185 | $25.28 | -14.35% | Buy | Correct |
| | | | | ARIMAX | 8143 | 8167 | $25.27 | -14.35% | Buy | Correct |
| | | | | UCM | 7800 | 7848 | $25.25 | -14.42% | Buy | Correct |
| | | | | ESM | -615 | -597 | $25.18 | -14.68% | Buy | Correct |
| | ZNH | $37.95 | $44.97 | ARIMA | 7663 | 7675 | $38.14 | -15.19% | Buy | Correct |
| | | | | ARIMAX | 7645 | 7669 | $38.14 | -15.20% | Buy | Correct |
| | | | | UCM | 7610 | 7658 | $38.00 | -15.51% | Buy | Correct |
| | | | | ESM | -1341 | -1323 | $38.09 | -15.30% | Buy | Correct |

*Table 1: Stock and Sector Analysis Summary*

For the UCM models, the overall guidance on the stocks is correct 10 out of 15 times (67%). ARIMA(X) was accurate 22 out of 30 times (73%), and ESM was accurate 12 out of 15 times (80%). Overall performance in the oil and transportation sectors both yielded an error rate of 35% across all models, whilst overall performance in the Public Utilities sector yielded an error rate of only 10% across all sectors. Against the hypothesis formed previously, that the oil sector would lend itself to higher accuracy in forecasting due to its trend, we were incorrect – even though the Public Utilities sector appeared to display more volatility and randomness in its behaviors, our close price forecasting was more accurate for this sector, as compared to the other sectors, namely oil.

In reviewing the prediction error for specific stocks, apparent outliers include 0% (BP [ESM Model]) and ~40% (CPL [All models]). Given our domain knowledge regarding historical accuracy and the complexity of stock market forecasting, we can attribute a perfect prediction to the ESM BP value to sheer luck, whereas the ~40% prediction error for all models on CPL raises the question – what potential economic shock caused this stock to shift so drastically against what each of our models predicted? Research into the topic indicates anticipated delisting of CPL on the stock market due to the looming acquisition of China's State Grid Corp led to the plummeting stock price of CPFL in Nov. 2017 (CPFL Energia).

The methodology for the model derivation and application was based on an approach which iteratively tuned and refined the models to develop the best fit. In the case of UCM, this resulted in model 4 (as referenced in Table 4) and developed based on the Oil and Gas sector. Referencing Table 3, we can see that the best fit model for predicting XOM using ARIMAX is model 4. The ESM process yielded the best results as referenced in Table 1, with a BIC of 323 (the next closest BIC being 8090 [ARIMAX]). Given the similarities of the sectors (Oil, Transport and Utilities), and the reliance of energy, we applied the best model of each of the algorithms (ARIMA, UCM and ESM) to assess the prediction performance of the models.

With regards to the sectors analyzed we note the performance as follows:

| Sector | ETF | Trend/Guidance Provided | Actual 11/10/2017 | Actual 11/30/2017 | Profit/Loss |
|---|---|---|---|---|---|
| Oil & Gas | IEO | Upward -> Buy | $62.31 | $61.20 | -1.78% |
| Utilities | IDU | Upward -> Buy | $138.92 | $142.02 | 2.23% |
| Transportation | XTN | Upward -> Buy | $58.05 | $63.36 | 9.15% |

*Table 2: Sector Performance*

We can see that while a potential investor may have fared positively if they had spread their investments across these three sectors evenly, the prediction is incorrect for Oil, and would have resulted in a loss, had the investor decided to invest solely on this sector. Overall the stock market is impacted by a multitude of components, many of which can be very tough to capture and model, of which human emotion is an extremely important part. It has been well established that human emotion leads to illogical bubbles and crashes, and algorithms have a difficult time accounting for this. So, while these models can prove reliable on forecasting data such as traffic, capacity etc., an investor would be well advised to take an extreme degree of caution when hoping to make investment decisions based on these models.

This section has summarized the overall performance of our models as they pertain to the stocks and associated sectors. The following sections contain a deep dive into the Oil & Gas sector, specifically the stock for Exxon (XOM) which was used to iteratively refine and develop the prediction models utilized. We note that while this deep dive is focused on one sector/stock, this process can be replicated across many sectors and stocks to conduct analysis and predictions across additional industries and sectors.

### 5.1 ARIMA(X)

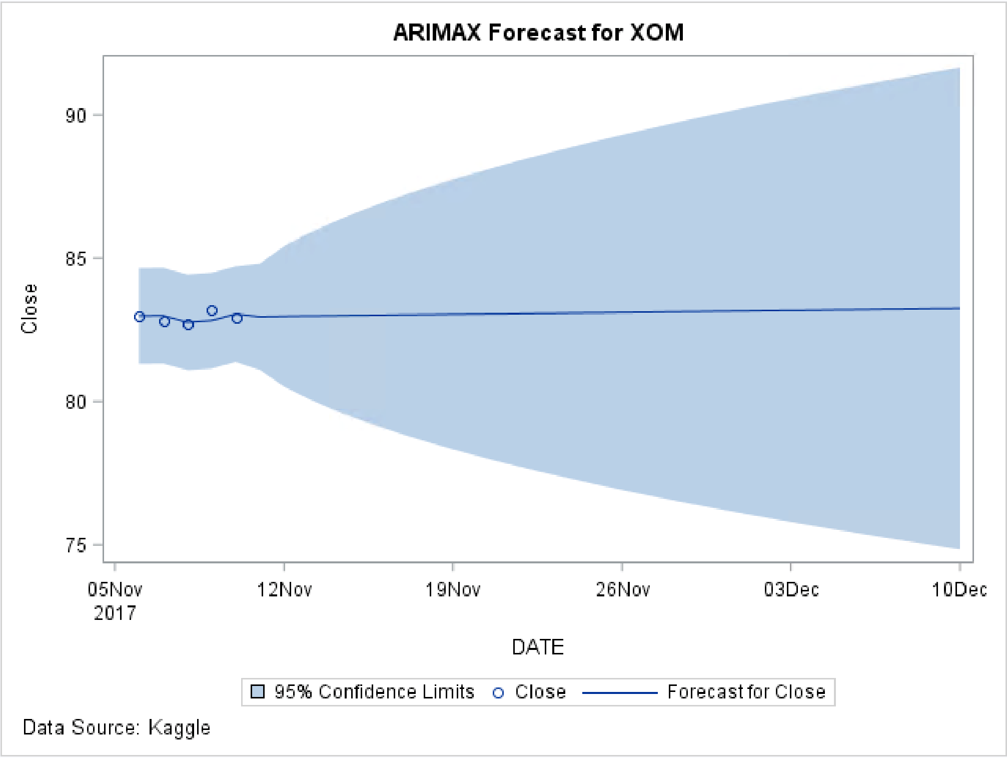
Focused efforts towards ExxonMobil data yielded best-in-class results with a non-log transformed ARIMAX (0,1,2) model. With a prediction error of only -0.18%, and a best fit BIC value of 8,090, we have proven that the ARIMAX model produces one of the most accurate forecast out of the four models tested (ESM, UCM, ARIMA, ARIMAX). Given these statistics, we maintain a strong buy rating on XOM with a 20-day forecast towards growth while heading into the holiday season.

A quick study of our model diagnostics are captured in Table 3 - ARIMA(X) Model Comparison. As previously indicated, the ARIMAX(0,1,2) model has the best fit, and therefore is most accurate in overall forecasting- however the log-transformed ARIMA(0,1,2) model achieved the closest prediction for the 11/30/2017 XOM stock price ($83.32 predicted vs $83.29 actual).

| Model Type | Model | Company | Inclusions | AIC | BIC / SBC | Predicted Nov 30th Price | Actual Nov 30th Price |
|---|---|---|---|---|---|---|---|
| **ARIMA** | **1** | XOM | Log (0,1,2) | -17840.60 | -17822.40 | $83.32 | $83.29 |
| | **2** | XOM | Non-log (0,1,2) | 8746 | 8764.43 | $83.15 | $83.29 |
| | **3** | XOM | Non-log (0,1,0) | 8806 | 8812.27 | $83.14 | $83.29 |
| **ARIMAX** | **4** | XOM | Non-log (0,1,2) | 8066 | 8090 | $83.14 | $83.29 |

*Table 3 - ARIMA(X) Model Comparison*

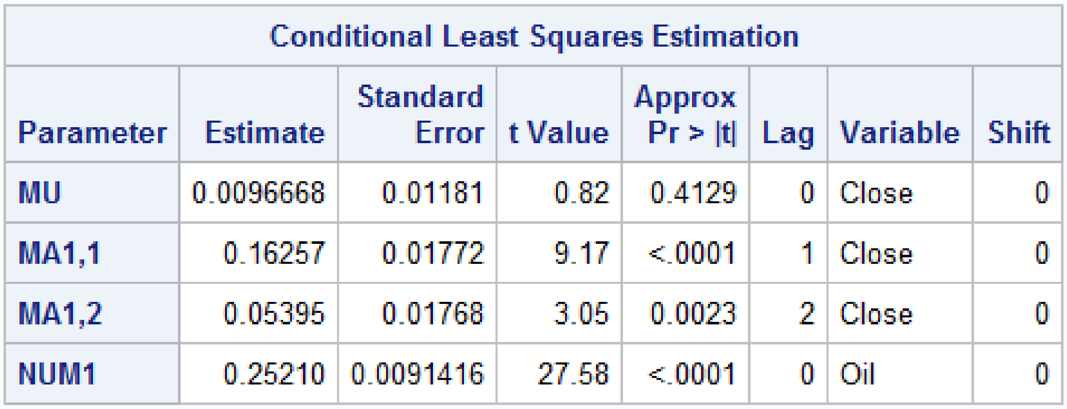
The last observed value on 11/10/17 was $82.94 for XOM; analyzing historical performance and market trends, we forecasted a 11/30/17 value of $83.14. The market responded in favor of our analysis, and our investors yielded portfolio growth .18% higher than we predicted.



Data Source: Kaggle

*Figure 8: ARIMAX Forecast*

As seen in Figure 9: ARIMAX Model, the model used to derive our results included a moving average component for our close variable, and an exogenous variable – oil. **Error! Reference source not found.**



Conditional Least Squares Estimation

| Parameter | Estimate | Standard Error | t Value | Approx Pr > \|t\| | Lag | Variable | Shift |
|---|---|---|---|---|---|---|---|
| MU | 0.0096668 | 0.01181 | 0.82 | 0.4129 | 0 | Close | 0 |
| MA1,1 | 0.16257 | 0.01772 | 9.17 | <.0001 | 1 | Close | 0 |
| MA1,2 | 0.05395 | 0.01768 | 3.05 | 0.0023 | 2 | Close | 0 |
| NUM1 | 0.25210 | 0.0091416 | 27.58 | <.0001 | 0 | Oil | 0 |

*Figure 9: ARIMAX Model*

All variables are statistically significant at alpha = .05, apart from our mean value, which is to be expected. More thorough analysis can be referenced in the appendix.

## 5.2 Unobservable Components Model (UCM)

### 5.2.1 Model Derivation and Assessment

The UCM model can be written as:

$$y_t = \mu_t + \gamma_t + \psi_t + \sum_{j=1}^{m} \beta_j x_{jt} + \varepsilon_t$$

The model allows for various components including trend, seasonal, cycle, irregular and regressors. The various components in the UCM equation include:

- $\mu_t$ is the trend component
- $\gamma_t$ is the season component
- $\sum_{j=1}^{m} \beta_j x_{jt}$ represents the regressors
- $\psi_t$ is the cyclical component
- $\varepsilon_t$ is the irregular component

It should be noted that within the UCM model, SAS has the ability to include a Random Walk (RW) trend model (uses random walk with drift as a baseline model - reported in the summary as RW R$^2$). With this model in mind, and the ability to include other factors that impact the forecast of the time series, we used the following explanatory variables and assessed their suitability:

- Daily closing price of Oil (Spot Crude Oil Price: West Texas Intermediate )
- Gold Fixing Price 10:30 A.M. (London time)
- Henry Hub Natural Gas Spot Price

In addition to this, given UCM flexibility, we also experimented with the inclusion and exclusion of the various components as well as the variables above to determine which model yields the best result. The model with the best performance is analyzed in detail, but we also summarize some of the models tried, as well as other key KPIs observed. This summary is presented below:

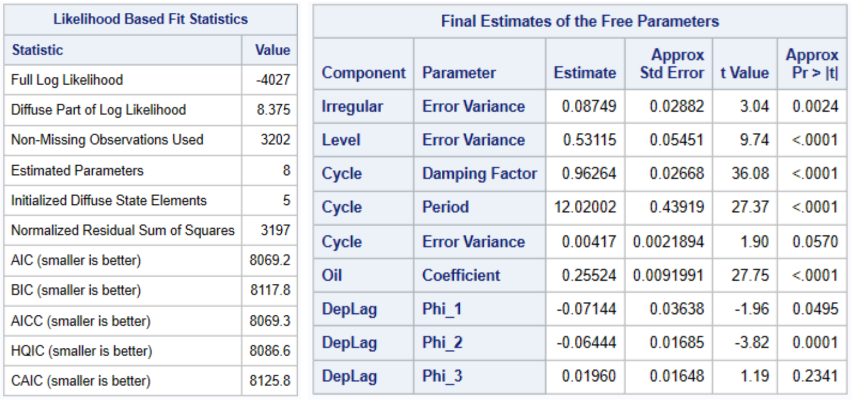
| Model | Inclusions | AIC | BIC | Adjusted R$^2$ | MAPE | RW R$^2$ | Predicted Nov 30th Price | Actual Nov 30th Price |
|---|---|---|---|---|---|---|---|---|
| 1 | Irregular<br>Level<br>Slope<br>Season<br>Cycle<br>Lag 1-3 | 8802.60 | 8863.30 | 0.99 | 1.03 | 0.00 | 83.06 | 83.29 |
| 2 | Irregular<br>Level<br>Cycle<br>Lag 1-3 | 8747.20 | 8795.70 | 0.99 | 1.02 | 0.02 | 82.94 | |
| 3 | Irregular<br>Level<br>Slope<br>Season<br>Cycle<br>Lag 1-3<br>Oil<br>Gas<br>Gold | 8139.10 | 8199.80 | 1.00 | 0.93 | 0.18 | 83.24 | |
| 4 | Irregular<br>Level<br>Cycle<br>Lag 1-3<br>Oil | 8069.20 | 8117.80 | 1.00 | 0.91 | 0.21 | 83.12 | |

*Table 4: UCM Model Performance Summary*

The factors we used during the experimentation stage to determine which models to try were primarily composed of evaluating the significance of the various components as well as the metrics listed in Table 4.

Based on the evaluation metrics, model 4 produced the best metrics. Specifically, the AIC, BIC, Adjusted $R^2$, MAPE, and RW $R^2$ all exhibit the best values in Model 4. However, it is interesting to note, that despite slightly poorer metrics, model 3 produces the closest predicted for the XOM stock close price for Nov 30th. This is an arbitrarily selected date we used for the prediction. Despite this, one observes that the Mean Absolute Percentage Error (MAPE) for model 4 is slightly lower than model 3 (0.93 vs 0.91), so overall model 4 produces a lower percentage error.

The output produced for model 4 is given below:



Likelihood Based Fit Statistics

| Statistic | Value |
|---|---|
| Full Log Likelihood | -4027 |
| Diffuse Part of Log Likelihood | 8.375 |
| Non-Missing Observations Used | 3202 |
| Estimated Parameters | 8 |
| Initialized Diffuse State Elements | 5 |
| Normalized Residual Sum of Squares | 3197 |
| AIC (smaller is better) | 8069.2 |
| BIC (smaller is better) | 8117.8 |
| AICC (smaller is better) | 8069.3 |
| HQIC (smaller is better) | 8086.6 |
| CAIC (smaller is better) | 8125.8 |

Final Estimates of the Free Parameters

| Component | Parameter | Estimate | Approx Std Error | t Value | Approx Pr > \|t\| |
|---|---|---|---|---|---|
| Irregular | Error Variance | 0.08749 | 0.02882 | 3.04 | 0.0024 |
| Level | Error Variance | 0.53115 | 0.05451 | 9.74 | <.0001 |
| Cycle | Damping Factor | 0.96264 | 0.02668 | 36.08 | <.0001 |
| Cycle | Period | 12.02002 | 0.43919 | 27.37 | <.0001 |
| Cycle | Error Variance | 0.00417 | 0.0021894 | 1.90 | 0.0570 |
| Oil | Coefficient | 0.25524 | 0.0091991 | 27.75 | <.0001 |
| DepLag | Phi_1 | -0.07144 | 0.03638 | -1.96 | 0.0495 |
| DepLag | Phi_2 | -0.06444 | 0.01685 | -3.82 | 0.0001 |
| DepLag | Phi_3 | 0.01960 | 0.01648 | 1.19 | 0.2341 |

*Figure 10: UCM - Model 4 Likelihood and Final Estimates*

As reported earlier, the AIC, and BIC as reported in Figure 10, was used as a key evaluation criterion to identify Model 4 as the best performing model, out of the other four UCM models. It should be also noted, that regardless of the significance factor reported for the irregular component, it was included in the model to account for the overall random error in the models.

Furthermore, we used the final estimates presented in Figure 10 and their corresponding t-values to infer whether the corresponding component is non-stochastic (the null hypothesis) or is stochastic (the alternative hypothesis). The components reported as significant to highly significant (level, cycle, oil and lag: 1,2) lead us to assume that these components are best modeled as being stochastic.

The cyclical behavior of the stock price reported by UCM, namely its period, the damping factor, and the variance of the disturbance terms in its stochastic equations report that the period of the cycle is 12 days. However, it can be seen that while the p-values associated with the cycle period and the error variance of cycle are small (<0.0001), meaning they may be suitably modeled as being stochastic rather than nonstochatic (fixed), the cycle error variance appears to be deterministic because its error variance is insignificant @ 0.05 level.

Using this model, we were able to forecast the price of Exxon as depicted in Figure 11. As reported earlier, this model performs quite well and has a MAPE value of 0.91.

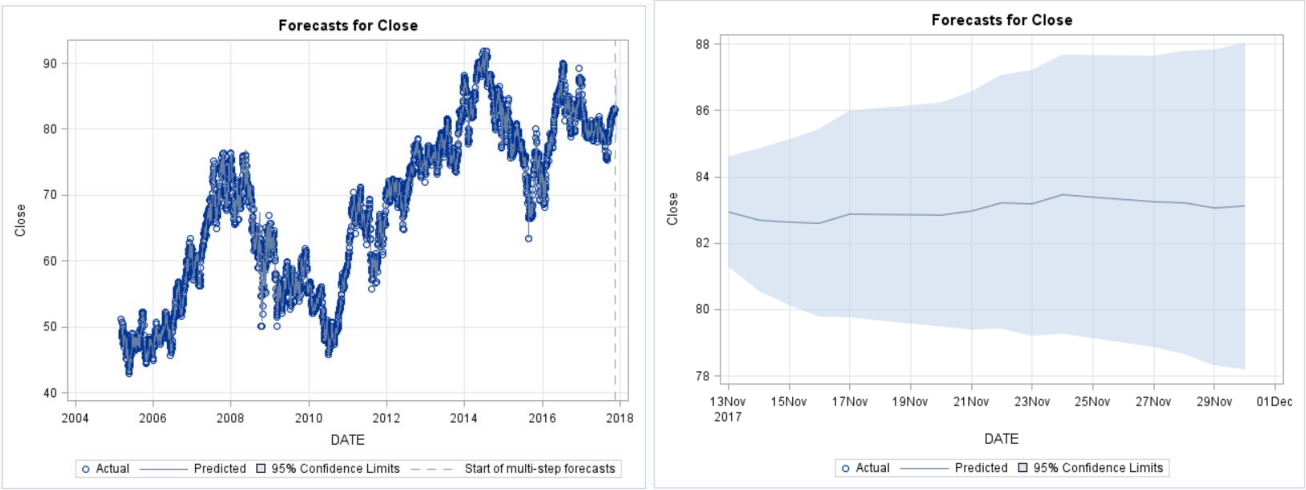


*Figure 11: Forecast for Exxon Close Price*

Finally, with regards to the inclusion/exclusion of other variables we made assessments from a layperson's perspective given a potential investor using such a model may not have a detailed understanding of how the oil industry works. We hypothesized that other commodities such as Oil, Natural Gas, and Gold may influence/impact the price of energy stocks. From our analysis, we see that while Oil certainly appears to impact the price of Exxon stock price, the same cannot be said about natural gas or gold. This appears to be an acceptable conclusion on these commodities as they are not directly related to Oil or Exxon, other than the fact that they are commodities themselves.

### 5.2.2 Model Prediction of Other Stocks & Sector Forecasting

Using model 4, we carried out forecasting of four other oil company stocks, namely, British Petroleum (BP), ConocoPhillips (COP), Chevron (CVX), and Valero Energy (VLO). As before, we have used Nov 30th as an arbitrary date to forecast, and based on the results, we assessed whether an investor would buy, hold or sell an oil/energy-based ETF. For this we selected IEO, "ishares Dow Jones US Oil and Gas Exploration and Production Index Fund."

| Stock | Last Day Observed (11/10/2017) | Forecast 11/30/2017 | Actual 11/30/2017 | Predicted Direction Based on Forecast | Guidance Correct? |
|---|---|---|---|---|---|
| BP | $ 40.30 | $ 40.45 | $ 40.07 | Buy | No |
| CVX | $ 117.18 | $ 117.39 | $ 118.99 | Buy | Yes |
| VLO | $ 81.37 | $ 81.47 | $ 85.62 | Buy | Yes |
| COP | $ 52.99 | $ 53.17 | $ 50.88 | Buy | No |
| XOM | $ 82.94 | $ 83.12 | $ 83.29 | Buy | Yes |

*Table 5: Stock Forecasts*

Based on the summary above, we concluded an investor should have bought the ETF IEO on Nov 10th 2017 under the guidance of the predictions, noted as Buy across all 5 stocks, the net result of the investment is given below:

| ETF | Actual 11/10/2017 | Actual 11/30/2017 |
|---|---|---|
| IEO | $ 62.31 | $ 61.2 |

*Table 6: ETF Performance*

We can clearly see the investor would have lost ~1.8% on their investment. This highlights one of the key findings from this project, namely that the stock market is impacted by many different factors, of which human emotion is one critical contributor. While we included oil prices to improve our model performance, and for future analysis could also account for additional factors, the difficulty in forecasting stock market prices remains dependent on the human factor. The great bubbles and crashes of the stock market have been examined and studied in many books, PHD dissertations as well as research papers, yet we continue to see crashes and bubbles. So, although the UCM model can be a powerful tool, especially as it accounts for unobserved components, it may be better suited and reliable for forecasting data such as traffic, capacity etc. Investors hoping to make stock purchase decisions based on UCM can rely on it as a tool to garner an additional statistical perspective, rather than going with an uneducated "gut" decision, however a high degree of caution would be advised. As highlighted in Table 5, out of 5 stocks, the model predicted with a 60% accuracy.

## 5.3 Exponential Smoothing Model (ESM)

### 5.3.1 Model Derivation and Assessment

The idea of exponential smoothing is to estimate a model locally by using all the observations and weighting the most recent ones more heavily than those in the past. Models can have a local mean, local trend, and local seasonal pattern. Interestingly, many of these models are equivalent to certain subsets of the ARIMA models. This equivalency is exploited in this section. An example of a local-level model is the random walk model The ESM model can be written as

$$Y_t = Y_{t-1} + e_t$$

We have used the SAS ESM procedure for ESM analysis on all the stocks. We are considering a seasonality factor of 7 days and forecasting 30 days of closing price ahead of the last observed day.
The results for XOM close price are referenced in Figure 12:

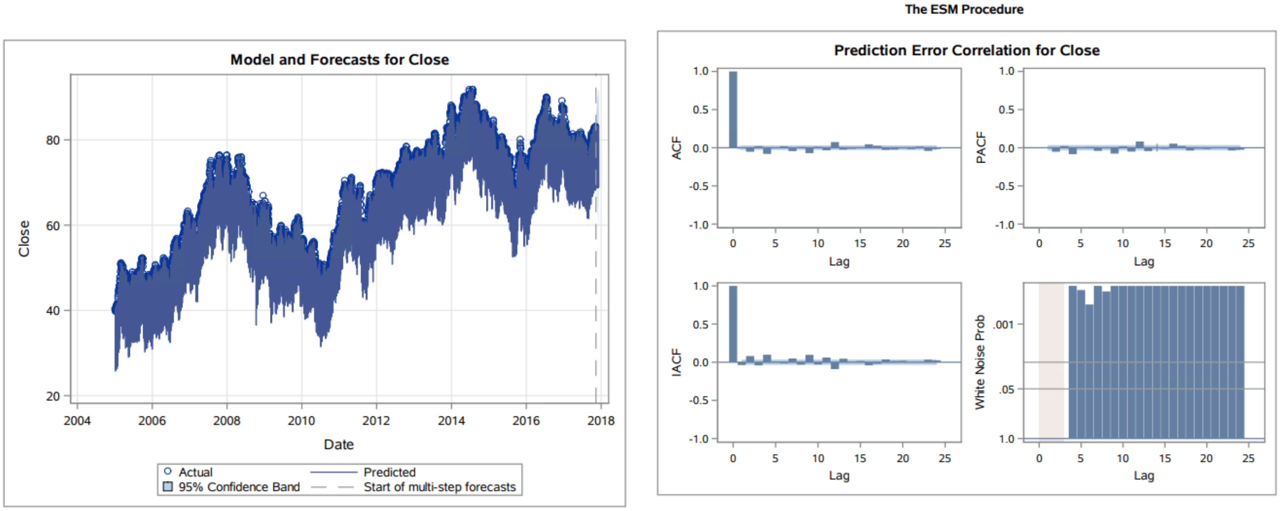
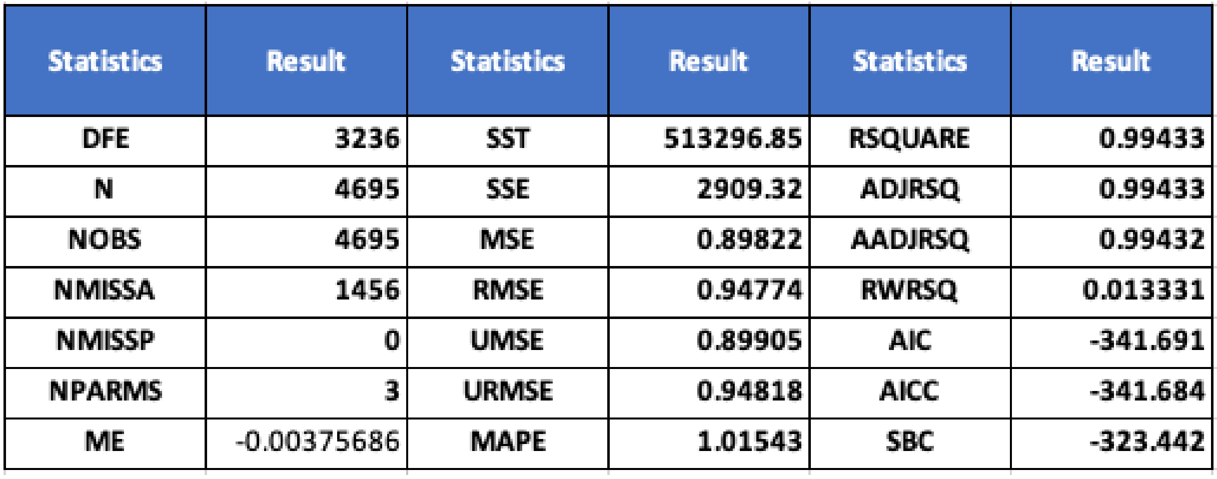


*Figure 12: Lag Analysis*



| Statistics | Result | Statistics | Result | Statistics | Result |
|---|---|---|---|---|---|
| DFE | 3236 | SST | 513296.85 | RSQUARE | 0.99433 |
| N | 4695 | SSE | 2909.32 | ADJRSQ | 0.99433 |
| NOBS | 4695 | MSE | 0.89822 | AADJRSQ | 0.99432 |
| NMISSA | 1456 | RMSE | 0.94774 | RWRSQ | 0.013331 |
| NMISSP | 0 | UMSE | 0.89905 | AIC | -341.691 |
| NPARMS | 3 | URMSE | 0.94818 | AICC | -341.684 |
| ME | -0.00375686 | MAPE | 1.01543 | SBC | -323.442 |

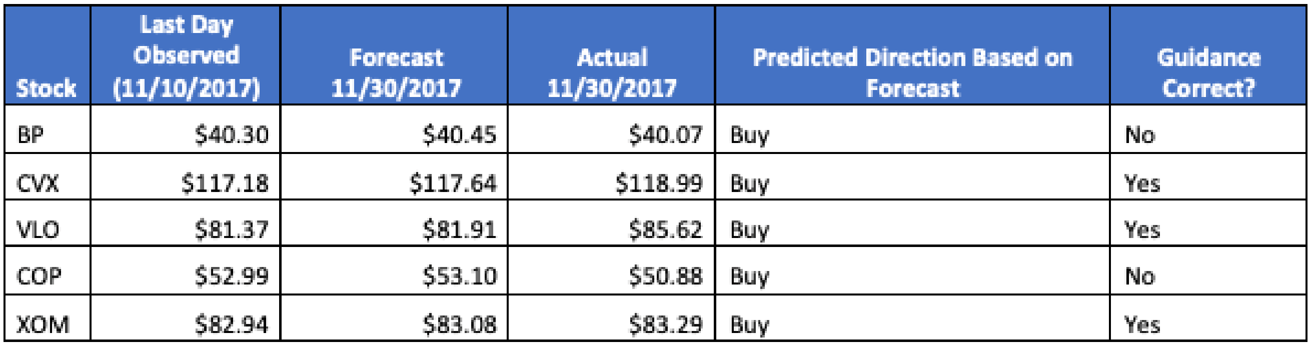
*Figure 13: Key Performance KPIs*

In Figure 12 we can see that the correlation among lags is very weak. This means the likelihood of day wise price dependence on the present day is very low.

Figure 12: Lag Analysis also provides the information that the errors are not white noise. This means the model is not able to explain all factors of time series. Fit statistics can be seen in Figure 13: Key Performance KPIs.

### 5.3.2 Model Prediction of Other Stocks & Sector Forecasting

Using the same methodology as UCM, we predicted the performance for the other stocks within the Oil and Gas sector:



| Stock | Last Day Observed (11/10/2017) | Forecast 11/30/2017 | Actual 11/30/2017 | Predicted Direction Based on Forecast | Guidance Correct? |
|---|---|---|---|---|---|
| BP | $40.30 | $40.45 | $40.07 | Buy | No |
| CVX | $117.18 | $117.64 | $118.99 | Buy | Yes |
| VLO | $81.37 | $81.91 | $85.62 | Buy | Yes |
| COP | $52.99 | $53.10 | $50.88 | Buy | No |
| XOM | $82.94 | $83.08 | $83.29 | Buy | Yes |

*Table 7 - ESM Model Predictions*

We note that when applied to the Oil and Gas Sector, the performance of the ESM model is very similar to the UCM model and yields a 60% accuracy rate on the overall guidance.

## 6. Conclusions

In the preceding sections, we presented a detailed assessment of the ARIMA, UCM and ESM models we used to forecast stock prices and overall sector performance for Oil and Gas, Public Utilities and Transportation. While the details of the model performances are presented in their respective sections, we conclude overall that the ESM model resulted in the best performance.

We also observed that at a stock level, there was large variation in the forecast accuracy, and note the intuitive nature of this result. Considering the various factors that impact the stock market, including human emotion, we also observed the inability of these forecasting models to capture and account for all such factors. However, given the improvement in accuracy by the inclusion of other factors such as Oil prices in the case of the Oil and Gas sector, we also hypothesize that as next steps to this project, students could improve on the model accuracy with the inclusion of other factors. It is interesting to note that startups such as "Social Market Analytics" carry out NLP and sentiment analysis on social media to try and account for the human emotion aspect and use this to predict stock market direction. Such analysis is beyond the scope of this project.

# 7. Citations

1. "The UCM Procedure." *The UCM Procedure*, SAS, support.sas.com/documentation/onlinedoc/ets/132/ucm.pdf.

2. "SAS/ETS Examples." *Analysis of Unobserved Component Models Using PROC UCM*, 22 June 2016, support.sas.com/rnd/app/ets/examples/melanoma/index.htm.

3. Fomby, Tom. "The Unobservable Components Model." *The Unobservable Components Model*, faculty.smu.edu/tfomby/eco5375/data/Unobservable%20Components%20Models/The%20Unobservable%20Components%20Model.pdf.

4. Adebiyi, Ayodele A., Adewumi, Aderemi O. *"Stock Price Prediction Using the ARIMA Model"* 2014, http://ijssst.info/Vol-15/No-4/data/4923a105.pdf.

5. Alwadi, Sadam & Almasarweh, Mohammad & Alsarairah, Ahmed. 29 October 2018 *"Predicting Closed Price Time Series Data Using ARIMA Model."* Modern Applied Science. 12. 181. 10.5539/mas.v12n11p181.

6. Nau, Robert. "*Statistical Forecasting: Notes on Regression and Time Series Analysis,"* https://people.duke.edu/~rnau/411home.htm.

7. CPFL Energia. *"CPFL Energia Q3 Earnings Results Slides."* https://seekingalpha.com/article/4127534-cpfl-energia-s-2017-q3-results-earnings-call-slides

# 8. Appendix

This section contains some of the analysis that was conducted on the models. While this analysis was important, we have placed it in the appendix to allow the overall document to read better.

### 8.1.1 3 Stages to Building and ARIMA(X) Model – Identification, Estimation and Diagnostics, Forecasting

#### 8.1.1.1 Identification

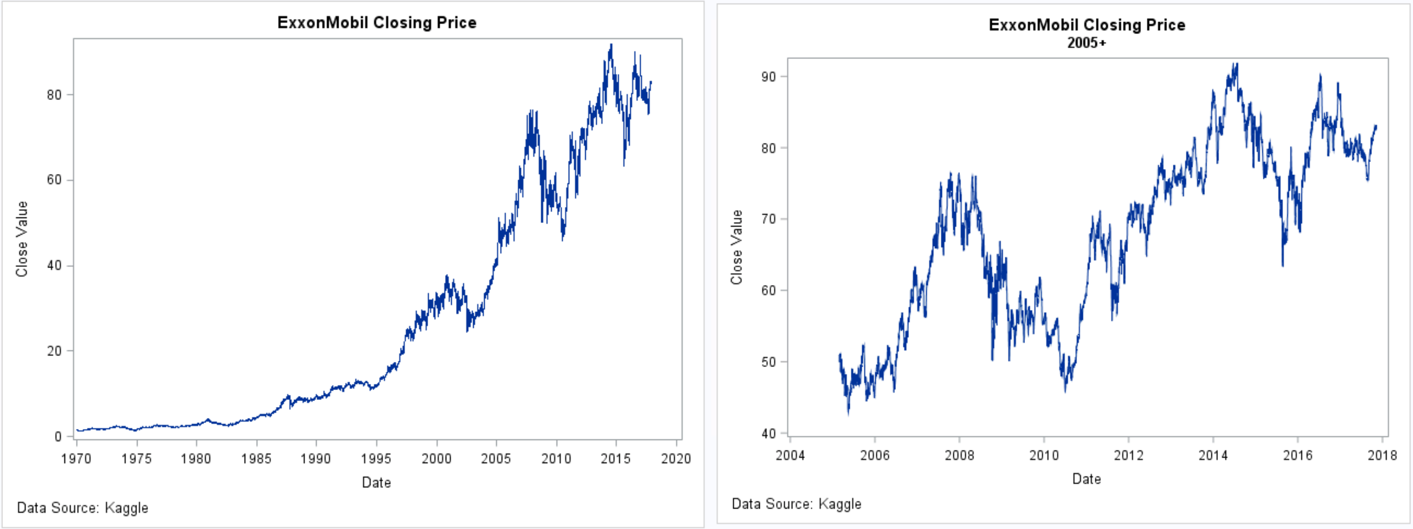


*Figure 14: Overall Exxon Close Price Trend*

Looking at these two plots, we decide to subset our data from 2005 onward, since that is when the most abundant, and somewhat recent fluctuations start in our data. Our data resembles a random walk, and so we will need to make the data stationary prior to performing any sort of meaningful analysis.

#### 8.1.1.2 Tests for Stationarity

A stationary time series is one where statistical properties like the mean and variance are constant over time. Before moving on to our model identification and specification stage or running any forecast we need to make our series stationary. For the purposes of testing stationarity within out Exxon stock series we have used the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test.

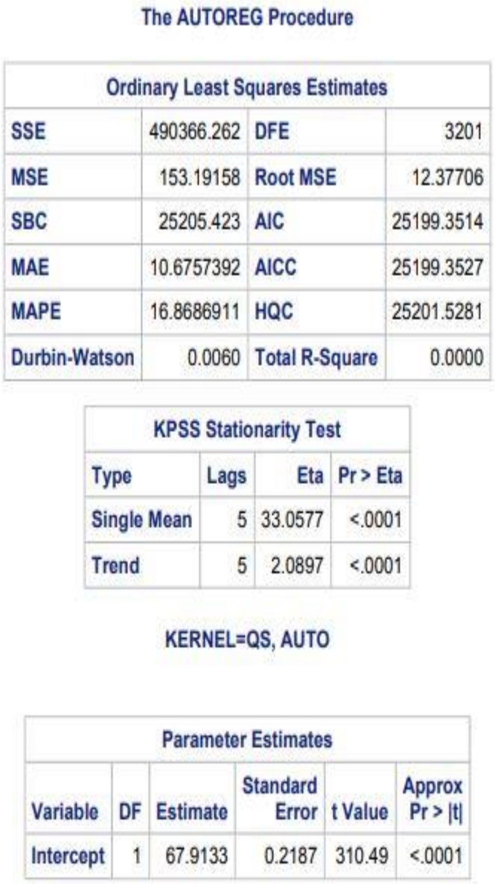
The Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test figures out if a time series is stationary around a mean or linear trend, or is non-stationary due to a unit root. The null hypothesis for the test is that the data is trend-stationary.

- The alternate hypothesis for the test is that the data is *not* stationary.

It breaks up a series into three parts: a deterministic trend (βt), a random walk ($r_t$), and a stationary error ($\varepsilon_t$), with the regression equation:

$$x_t = r_t + B_t + \varepsilon_1$$

The KPSS test generates the following output for the case of the QS kernel and automatic bandwidth selection:



The AUTOREG Procedure

| Ordinary Least Squares Estimates | | | |
|---|---|---|---|
| SSE | 490366.262 | DFE | 3201 |
| MSE | 153.19158 | Root MSE | 12.37706 |
| SBC | 25205.423 | AIC | 25199.3514 |
| MAE | 10.6757392 | AICC | 25199.3527 |
| MAPE | 16.8686911 | HQC | 25201.5281 |
| Durbin-Watson | 0.0060 | Total R-Square | 0.0000 |

KPSS Stationarity Test

| Type | Lags | Eta | Pr > Eta |
|---|---|---|---|
| Single Mean | 5 | 33.0577 | <.0001 |
| Trend | 5 | 2.0897 | <.0001 |

KERNEL=QS, AUTO

Parameter Estimates

| Variable | DF | Estimate | Standard Error | t Value | Approx Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 67.9133 | 0.2187 | 310.49 | <.0001 |

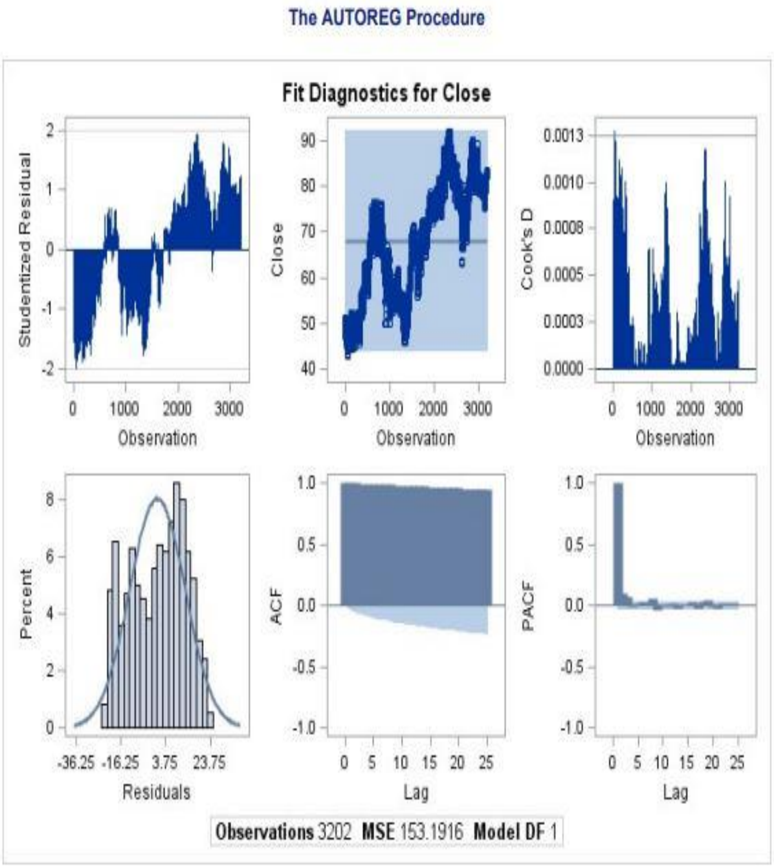


*Figure 15: Fit Diagnostics for Close*

On reading the output, we can start with the most general model (constant and trend) and move to the most specific (no constant and no trend). In the above output, the results for the trend, constant model are summarized. The eta statistic is the test that β1=0 or our series is trend stationary and we can reject the null of stationarity at 1% levels of significance, concluding that currently our series is not stationary or mean reverting.

### 8.1.1.3 Identification of the Differenced/Log-transformed Series:

To make our series stationary we perform differencing of the series. The correct amount of differencing is the lowest order of differencing that yields a time series which fluctuates around a well-defined mean value and whose autocorrelation function (ACF) plot decays fairly rapidly to zero, either from above or below. From the above plots we can clearly see that in current state our series if not stationary as our ACF shows plot decay is quiet slow along with the irregular oscillation within the standardized residuals chart.

We performed differencing to our Exxon series on both log and non-log models and below are the correlations plots :

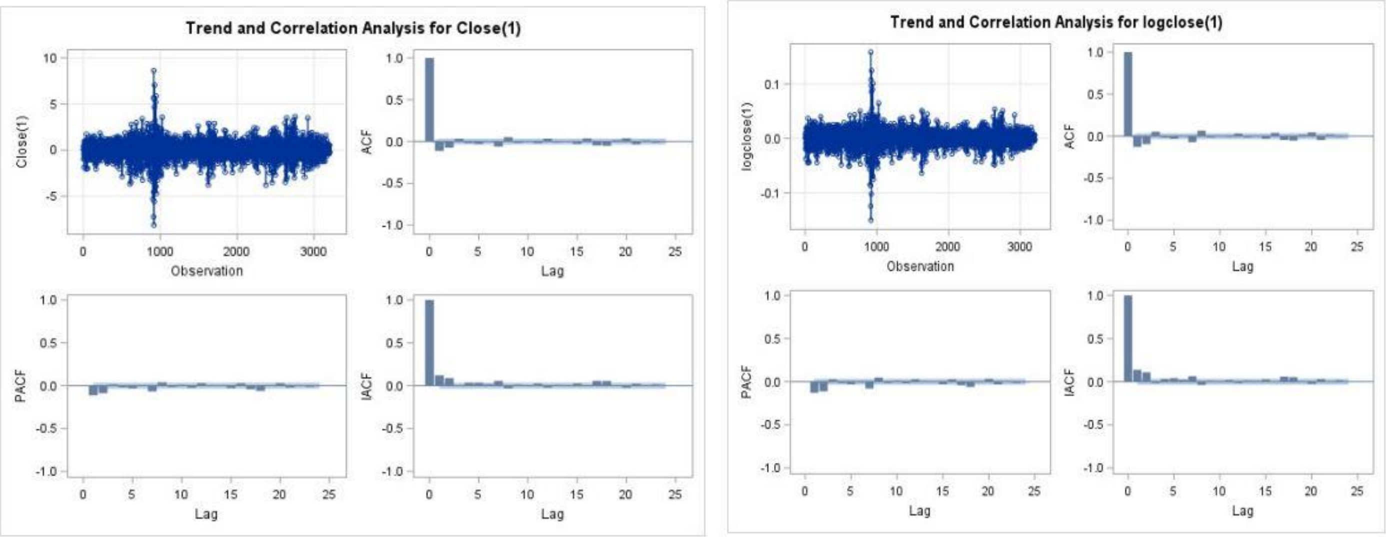
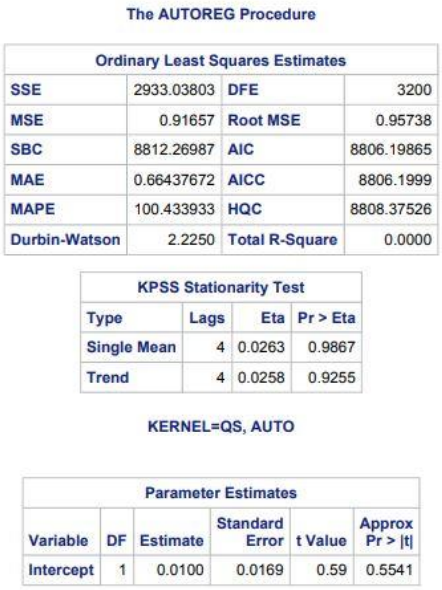


*Figure 16: Trend and Correlation Analysis*

After differencing our data, we can see that our closing price now appears to be centered around a mean of zero with some standard shocks as explored within our preliminary data analysis. The other thing to note would be that our autocorrelations also decrease drastically as seen within our ACF and PACF plots. Our autocorrelations seemed to have subsided to negative or even nearly zero in some lags. Based on the above plots we can safely assume that our series for both log and non-log models are stationary, We have verified this assumption by running the KPSS tests again on the differenced series and the results are stated below:



The AUTOREG Procedure

| Ordinary Least Squares Estimates | | | |
|---|---|---|---|
| SSE | 2933.03803 | DFE | 3200 |
| MSE | 0.91657 | Root MSE | 0.95738 |
| SBC | 8812.26987 | AIC | 8806.19865 |
| MAE | 0.66437672 | AICC | 8806.1999 |
| MAPE | 100.433933 | HQC | 8808.37526 |
| Durbin-Watson | 2.2250 | Total R-Square | 0.0000 |

KPSS Stationarity Test

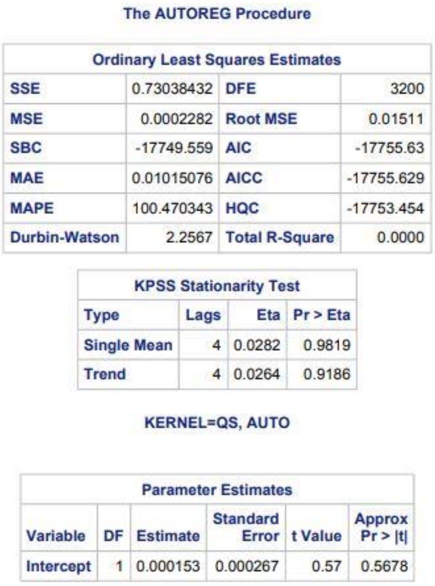
| Type | Lags | Eta | Pr > Eta |
|---|---|---|---|
| Single Mean | 4 | 0.0263 | 0.9867 |
| Trend | 4 | 0.0258 | 0.9255 |

KERNEL=QS, AUTO

Parameter Estimates

| Variable | DF | Estimate | Standard Error | t Value | Approx Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 0.0100 | 0.0169 | 0.59 | 0.5541 |

KPSS – Non-log Model



The AUTOREG Procedure

| Ordinary Least Squares Estimates | | | |
|---|---|---|---|
| SSE | 0.73038432 | DFE | 3200 |
| MSE | 0.0002282 | Root MSE | 0.01511 |
| SBC | -17749.559 | AIC | -17755.63 |
| MAE | 0.01015076 | AICC | -17755.629 |
| MAPE | 100.470343 | HQC | -17753.454 |
| Durbin-Watson | 2.2567 | Total R-Square | 0.0000 |

KPSS Stationarity Test

| Type | Lags | Eta | Pr > Eta |
|---|---|---|---|
| Single Mean | 4 | 0.0282 | 0.9819 |
| Trend | 4 | 0.0264 | 0.9186 |

KERNEL=QS, AUTO

Parameter Estimates

| Variable | DF | Estimate | Standard Error | t Value | Approx Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 0.000153 | 0.000267 | 0.57 | 0.5678 |

KPSS – Log Model

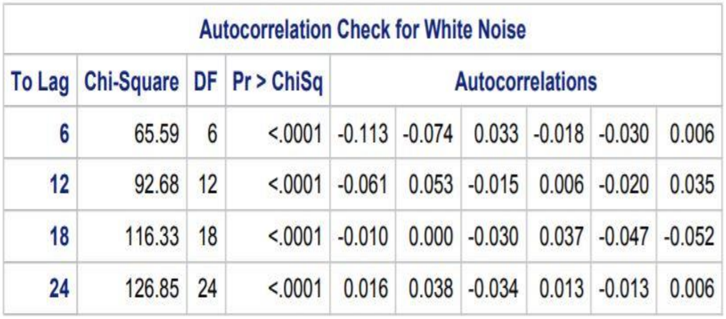
From the above results and p values we fail to reject our null at 1% and 5% levels of significance (H0 = Series is stationary) concluding that our series is stationary after first level of differencing.

Now that our series is stationary we can move forward to conduct tests for statistical independence by observing a white noise plot along with ACF and PACF in order to further refine our model to include any lags for autocorrelations that may help our model to better forecast.

### 8.1.1.4 Tests of Statistical Independence

Independence tests are prominently used in model analysis and diagnostics because models are usually based on the assumption of independently distributed errors. If a given time series is independent, then no determining model is necessary for this completely random process; otherwise, there must exist some relationship in the series to be addressed. For our analysis we would be using the Ljung Box – White Noise Test

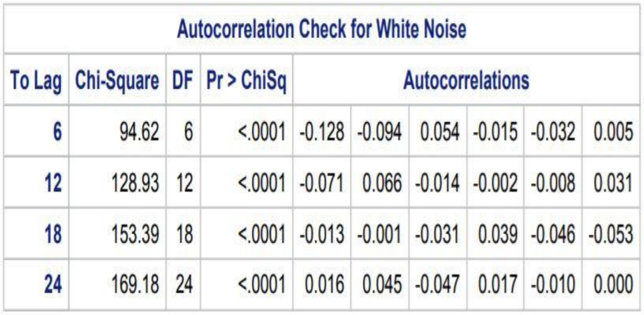
The null hypothesis is that the series is white noise, and the alternative hypothesis is that one or more autocorrelations up to lag m are not zero H0: The series is white noise, H1: The series is not white noise.



Autocorrelation Check for White Noise

| To Lag | Chi-Square | DF | Pr > ChiSq | Autocorrelations | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 6 | 65.59 | 6 | <.0001 | -0.113 | -0.074 | 0.033 | -0.018 | -0.030 | 0.006 |
| 12 | 92.68 | 12 | <.0001 | -0.061 | 0.053 | -0.015 | 0.006 | -0.020 | 0.035 |
| 18 | 116.33 | 18 | <.0001 | -0.010 | 0.000 | -0.030 | 0.037 | -0.047 | -0.052 |
| 24 | 126.85 | 24 | <.0001 | 0.016 | 0.038 | -0.034 | 0.013 | -0.013 | 0.006 |

Non-Log Model



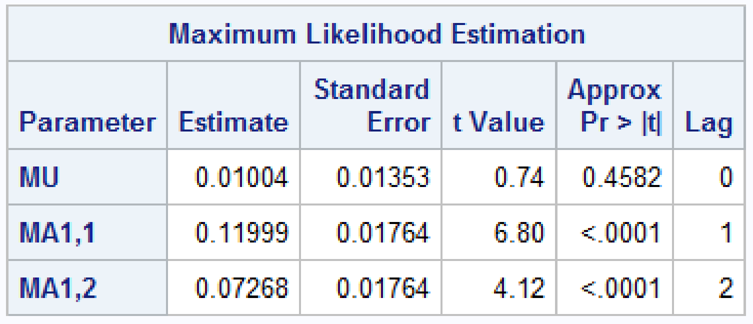
Autocorrelation Check for White Noise

| To Lag | Chi-Square | DF | Pr > ChiSq | Autocorrelations | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 6 | 94.62 | 6 | <.0001 | -0.128 | -0.094 | 0.054 | -0.015 | -0.032 | 0.005 |
| 12 | 128.93 | 12 | <.0001 | -0.071 | 0.066 | -0.014 | -0.002 | -0.008 | 0.031 |
| 18 | 153.39 | 18 | <.0001 | -0.013 | -0.001 | -0.031 | 0.039 | -0.046 | -0.053 |
| 24 | 169.18 | 24 | <.0001 | 0.016 | 0.045 | -0.047 | 0.017 | -0.010 | 0.000 |

Log-Model

The white noise test is an approximate statistical test of the hypothesis that none of the autocorrelations of the series up to a specified lag (30) are different from zero. With all p-values <.0001, we reject the null hypothesis in favor of the alternative. The autocorrelations for the first 30 lags are different from zero. We can see that the autocorrelations although are near to zero or negative but are significant thus indicating that there must be some relationship within the lags of the series that needs to be addressed within our model.

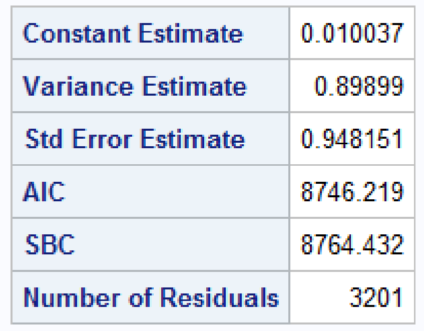
### 8.1.1.5 Estimation and Diagnostics



Maximum Likelihood Estimation

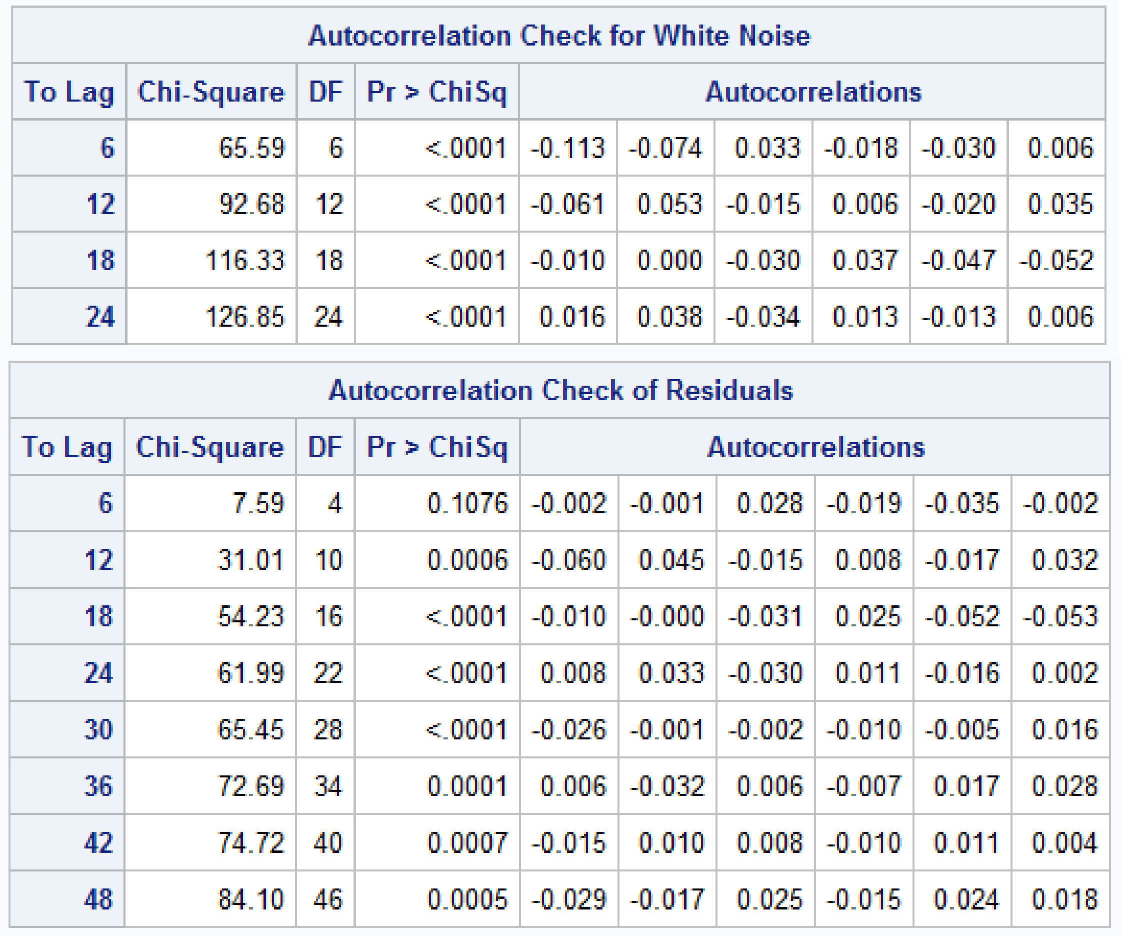
| Parameter | Estimate | Standard Error | t Value | Approx Pr > \|t\| | Lag |
|---|---|---|---|---|---|
| MU | 0.01004 | 0.01353 | 0.74 | 0.4582 | 0 |
| MA1,1 | 0.11999 | 0.01764 | 6.80 | <.0001 | 1 |
| MA1,2 | 0.07268 | 0.01764 | 4.12 | <.0001 | 2 |

The table above is our Maximum Likelihood Estimation with an ARIMA(0,1,2) estimation on our non-log model. Based on a t-value of .74 and p-value of .45, we determine that the mean value is not statistically significant from zero and does not need to be included in the model. All of the lagged variables are significant at alpha = .05, which is to be expected, given the results of our PACF.



| | |
|---|---|
| Constant Estimate | 0.010037 |
| Variance Estimate | 0.89899 |
| Std Error Estimate | 0.948151 |
| AIC | 8746.219 |
| SBC | 8764.432 |
| Number of Residuals | 3201 |

Looking at our Goodness-of-Fit Statistics, for our ARIMA(0,1,2) model, we see the AIC of 8746 and SBC of 8764, whicn were the lowest values attainable with the ARIMA model.



Autocorrelation Check for White Noise

| To Lag | Chi-Square | DF | Pr > ChiSq | Autocorrelations | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 6 | 65.59 | 6 | <.0001 | -0.113 | -0.074 | 0.033 | -0.018 | -0.030 | 0.006 |
| 12 | 92.68 | 12 | <.0001 | -0.061 | 0.053 | -0.015 | 0.006 | -0.020 | 0.035 |
| 18 | 116.33 | 18 | <.0001 | -0.010 | 0.000 | -0.030 | 0.037 | -0.047 | -0.052 |
| 24 | 126.85 | 24 | <.0001 | 0.016 | 0.038 | -0.034 | 0.013 | -0.013 | 0.006 |

Autocorrelation Check of Residuals

| To Lag | Chi-Square | DF | Pr > ChiSq | Autocorrelations | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 6 | 7.59 | 4 | 0.1076 | -0.002 | -0.001 | 0.028 | -0.019 | -0.035 | -0.002 |
| 12 | 31.01 | 10 | 0.0006 | -0.060 | 0.045 | -0.015 | 0.008 | -0.017 | 0.032 |
| 18 | 54.23 | 16 | <.0001 | -0.010 | -0.000 | -0.031 | 0.025 | -0.052 | -0.053 |
| 24 | 61.99 | 22 | <.0001 | 0.008 | 0.033 | -0.030 | 0.011 | -0.016 | 0.002 |
| 30 | 65.45 | 28 | <.0001 | -0.026 | -0.001 | -0.002 | -0.010 | -0.005 | 0.016 |
| 36 | 72.69 | 34 | 0.0001 | 0.006 | -0.032 | 0.006 | -0.007 | 0.017 | 0.028 |
| 42 | 74.72 | 40 | 0.0007 | -0.015 | 0.010 | 0.008 | -0.010 | 0.011 | 0.004 |
| 48 | 84.10 | 46 | 0.0005 | -0.029 | -0.017 | 0.025 | -0.015 | 0.024 | 0.018 |

*Figure 17: Autocorrelation check of Residuals*

The autocorrelation check of residuals table is very helpful in indicating whether or not our residuals are white-noise, and if our ARIMA(0,1,2) model is adequate in explaining our data. As interpreted previously, we can see that the test statistics fail reject the no-autocorrelation hypotheses at a high level of significance (p < .05).

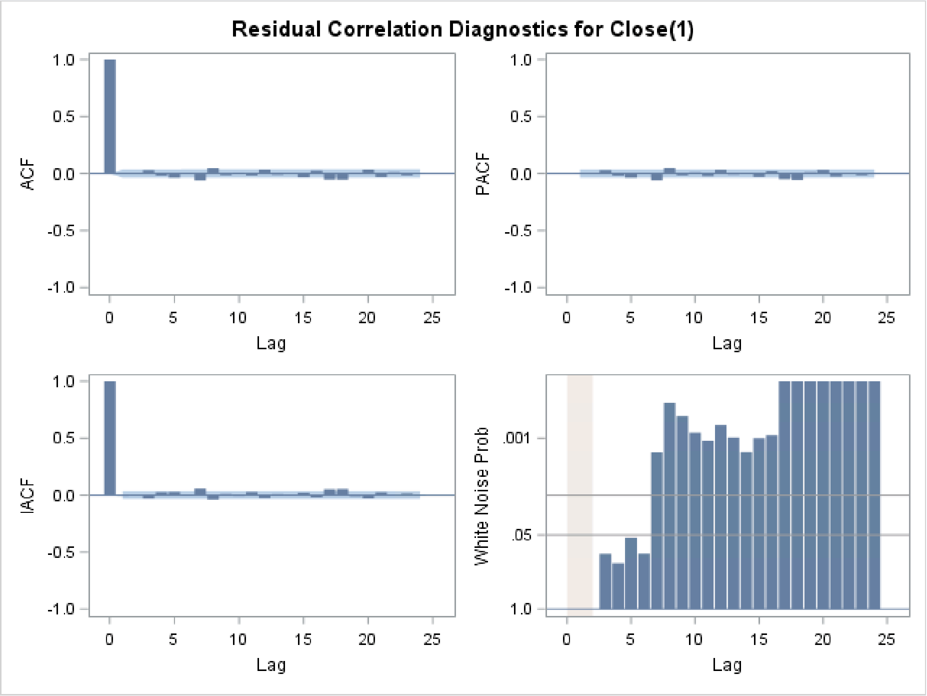


*Figure 18: Assessment of Lags*

Looking at our White Noise Probability chart, we are able to verify our assumptions that the residuals are not white noise, however this was the best fitting ARIMA model we were able to develop within the scope of our resources.

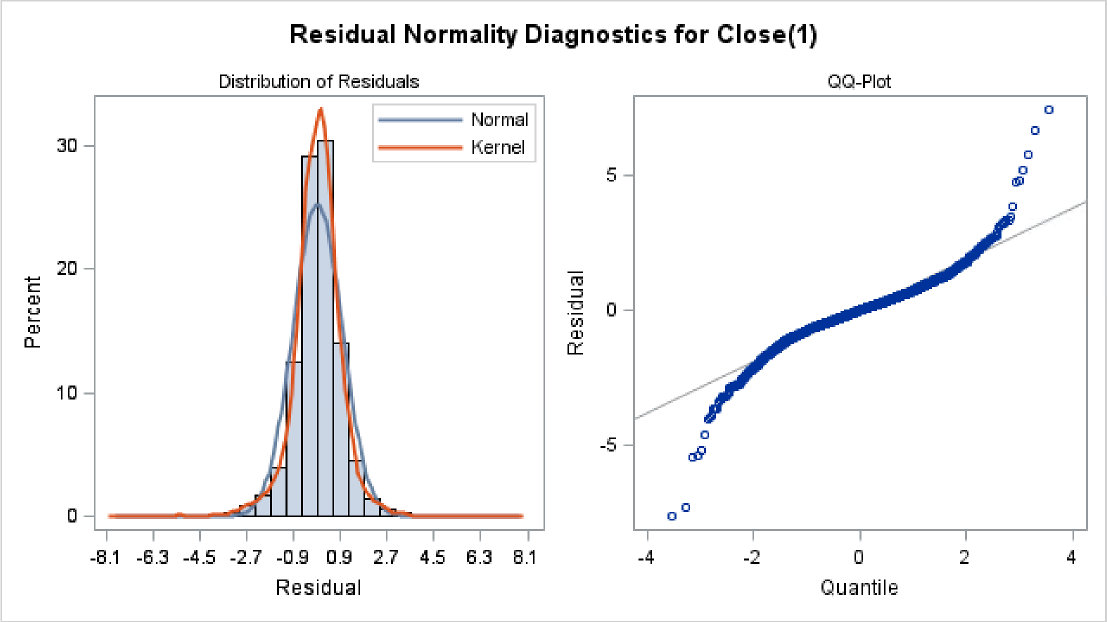


*Figure 19: Residuals Analysis*

Our residuals follow a distribution that has a slightly lower variance than a normal distribution.

### 8.1.1.6 Forecasting

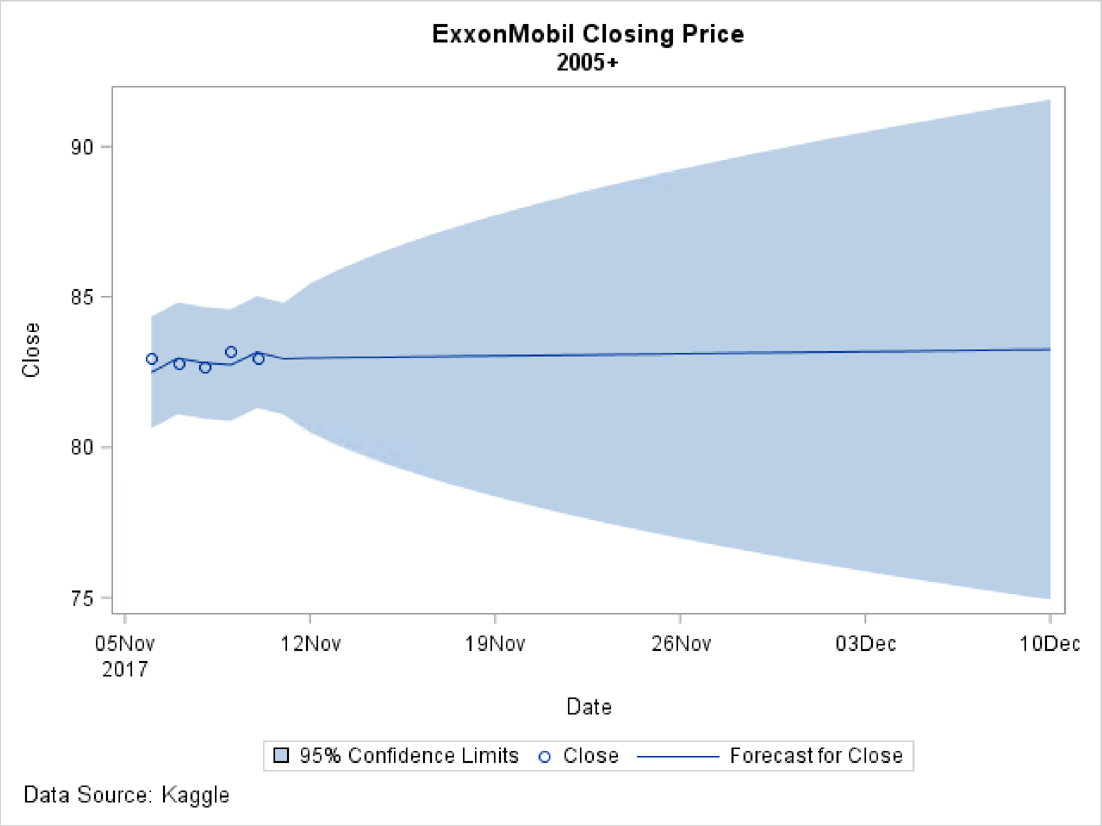


*Figure 20: Close Price Forecast*

The predicted 11/30/2017 value for our ARIMA(0,1,2) model is $83.149 with an actual value of $83.29. Since our last observed value on 11/10/2017 was valued at $82.94, our recommendation would have been to buy XOM stocks on 11/10/2017, yielding a .42% gain as the outcome. With a $1000.00 investment, our net gain (USD) would have been $4.20.

### 8.2 ESM

An ARIMA model that is like the random walk model, but more flexible is Yt − Yt−1 = et −θet−1 with |θ|<1. This is the integrated moving average model of order 1, ARIMA(0,1,1) or IMA(1,1). In fact, when differencing, it is a good idea to start the modeling by using a moving average term at the lag implied by whatever orders of differencing are used. In that way, a value of θ close to 1 can be an indicator of over-differencing. Assuming a positive value 0 < θ < 1, a series of back substitutions yields an interesting representation for Y as follows. The IMA(1,1) model holds at all times, so that Yt−j − Yt−j−1 = et−j −θet−j−1 for any j. Thus, each et−j is et−j = Yt−j − Yt−j−1 + θet−j−1. Substituting for et−1 in et = (Yt − Yt−1) + θet−1 result in the following:

$$e_t = (Y_t - Y_{t-1}) + \theta e_{t-1} = (Y_t - Y_{t-1}) + \theta\left((Y_{t-1} - Y_{t-2}) + \theta e_{t-2}\right) = (Y_t - Y_{t-1}) + \theta(Y_{t-1} - Y_{t-2}) + \theta^2 e_{t-2}$$

Substituting for et−2 gives this:

$$e_t = (Y_t - Y_{t-1}) + \theta(Y_{t-1} - Y_{t-2}) + \theta^2 e_{t-2} = Y_t - (1-\theta)Y_{t-1} - \theta(1-\theta)Y_{t-2} - \theta^2 Y_{t-3} + \theta^3 e_{t-3}$$

An alternate way of writing the equation is:

$$\hat{Y}_{t+1} = (1-\theta)\sum_{j=0}^{\infty}\theta^j Y_{t-j}$$