# Predicting AUD/THB Currency Trends

SpringBoard Data Science Intensive Capstone Project Report

## 1 INTRODUCTION

There are primarily two categories of people who are interested in the movements of currency exchange prices:

1. The Foreign Exchange (FOREX, FX) professionals – these are the people who are dedicated to trading daily in the FX markets, and include brokers, banks and licensed agencies.

2. The currency exchange layman, like most of us – whose interest in currency exchange rates is occasional, such as when we are travelling abroad, making overseas investments and purchase, and perhaps more regularly, the remittance of money for family support.

The first category of FX professionals has esoteric knowledge of the complex FX market, where unlike commodity or stock market, hard cash is actually traded with no room for a later liquidity strategy. They use sophisticated technical tools for market analysis and forecasting, continually improving on their 'secret sauce' algorithms to compete and succeed.

The second category of FX layman tend to be dumbfounded by sudden movements of the FX currency pair they are interested in. They do not have the 'inside' knowledge and rely on the news media and financial reports available to the public to make judgements on when to buy/sell. For these people, the indicators of FX trend can be mythical and confusing: Is it the gold price to watch? Perhaps crude oil? Or the stock market? Or the Reserve Bank's announcement of interest rate change? Or performance of the retail sector? Or trade balance? The list goes on…

The objective of this study is to provide simple and reliable FX indicators for the layman.

In particular, I have chosen the AUDTHB currency pair with the target audience of the Thai community in Australia in mind. As a current practice, they go to a few privately licensed Thai agencies that operate in Bangkok and Sydney/Melbourne, where the AUD/THB exchange rates are several pips better than most banks and public money traders.

Large sums of money are exchanged daily, eg:

- Overseas students receiving money from parents in Thailand
- Workers sending money back to Thailand for family support
- Business people investing in Thailand, and vice versa.

Hence, the ability to predict AUD/THB trends would help this community demystify the possible influences on the AUDTHB price, so they can make better judgements on the timing of their currency exchange to make better savings or minimise loss.

# 2 BACKGROUND ON FX (FOREX) MARKET

The FX market is elusive and seemingly complicated to the layman. It is in fact, a complex market. More so than stock, commodity and other financial market.

More than anything else, it is a market that is open around the clock six days a week, enabling traders to act on news and events as they happen. It is a market where half-billion-dollar trades can be executed in a matter of seconds, and may not even move prices noticeably. Try buying or selling a half billion of anything in another market and see how prices react.

Average daily currency trading volumes exceed $2 trillion per day. That's a mind-boggling number - $2,000,000,000,000. To give some perspective on that size, it is about 10 to 15 times the size of daily trading volume on ALL the world's stock markets combined!

Estimates are that upwards of 90 percent of daily trading volume is derived from speculation (meaning, commercial or investment-based FX trades account for less than 10 percent of daily global volume). The depth and breadth of the speculative market means that the liquidity of the overall FX market is unparalleled among global financial markets.

The bulk of spot currency trading, about 75 percent by volume, takes place in the so-called "major currencies," which represent the world's largest and most developed economies. These 7 major currency pairs all involve the USD on one side of the deal:

- EUR/USD
- USD/JPY
- GBP/USD
- USD/CHF
- USD/CAD
- AUD/USD
- NZD/USD

Additionally, activity in the FX market frequently functions on a regional "currency bloc" basis, where the bulk of trading takes place between the USD bloc, JPY bloc, and EUR bloc, representing the three largest global economic regions.

Because of the size of the Japanese market and the importance of Japanese data to the market, much of the action during the Asia-Pacific session is focused on the Japanese yen currency pairs, such as USDJPY and the JPY crosses, like EUR/JPY and AUD/JPY.

The FX market does not exist in a vacuum. There is a fair amount of noise and misinformation about the interrelationship among other markets (gold, oil, stocks, bonds., etc.) and currencies or individual currency pairs. Important fundamental and psychological relationships exist and generally the influences are macro- and micro-economic factors shown in the table below.

**Macro-Economic Factors**

- Major currency pairs, eg: AUD/USD
- Asia Pacific bloc, eg: THB/JPY, AUD/JPY
- Gold
- Oil
- Stock (Equity)
- Government Bonds
- Political / Social events *

**Micro-Economic Factors**

- Interest rates
- Trade balance,
- GDP
- Consumption (retail)
- CPI
- Monetary policies *
- Central bank action *

# 3 DATA SETS

The data sources I need for this project are primarily historical AUDTHB prices, and the base or cross currency pairs likely to impact the AUDTHB rate. The 6 currency pairs needed are shown in the matrix below:

| | AUD | THB | JPY | USD |
|---|---|---|---|---|
| AUD | x | 1 | 2 | 3 |
| THB | x | x | 4 | 5 |
| JPY | x | x | x | 6 |
| USD | x | x | x | x |

Whilst there is no shortage of FX data on the internet, most sites require paid subscription. In looking for good FX data sources that provides me the data qualities of Relevency, Recency, Range, Reliability, etc. I narrowed this down to CSV downloads that is available from: http://www.investing.com/currencies/aud-thb-historical-data

Macro and micro-economics data are also readily available from many sites, including Government Open Data sources. However, the data quality are not as good as the FX data, particularly the range and recency of these data do not match the FX data that I have already downloaded.

Eventually, I landed with:

- Gold Prices in AUD:
  https://www.quandl.com/collections/markets/gold

- Stock Prices (Nasdaq):
  https://finance.yahoo.com/quote/%5EIXIC/history?period1=1136034000&period2=1474639200&interval=1d&filter=history&frequency=1d

- Crude Oil price in USD:
  https://www.quandl.com/collections/markets/crude-oil

All the data sources are daily data points, with a >10 years range, from 02Jan06 to 9Sep16

## 3.1 Tools Used

The tools that I used for the data analytics are:

- Jupyter Notebook
- Python 3.5 and common Ananconda packages.
- Statsmodel
- Pyflux
- PyBrain

There are other tools that I would have liked to test (eg. keras), but was constrained by my Windows10 OS, and issues with Python 2.x and Python 3.x compatibility.

## 3.2 Data Wrangling

The data sets I used are daily time-stamped data. They have gaps in public holidays, no-trading weekends etc. My data wrangling effort in this project consists mainly of:

- Merging different dataframes from FX, stock and commodity data
- Dropping null values
- Slicing data to appropriate range, eg. for interrupted time series analysis in ARIMAX
- Parsing datasets to transform time-stamps to time-series format,
- Creating dataframe indices
- All sorts of data transformation that are required to achieve modeling fit, eg. log, first-order differencing.

## 3.3 Exploratory Data Analysis

The primary data set of interest in this project is the AUDTHB. Hence most of the EDA work is focused on this data set which will be used for the more detailed analytics and modeling later on.

Examples of the EDA that was carried out include:

- Summary stats
- Box plot – visualize variance
- Histogram plot
- Scatter plot

Essentially, I am dealing with a time series data set, and the analytics to be carried out will be focused on time series techniques. This include the need to ‘stationarise’ the time series, i.e. remove the components of trend, seasonal and cyclic patterns, so as to render the residual data as closely aligned with the Central Limit Theorem as possible. Only then can the ‘stationarised’ data be meaningfully analysed and modelled using linear statistical methods.

The need to stationarise can also be visually surmised from the plot of raw AUDTHB time series data shown in Figure 1 below.

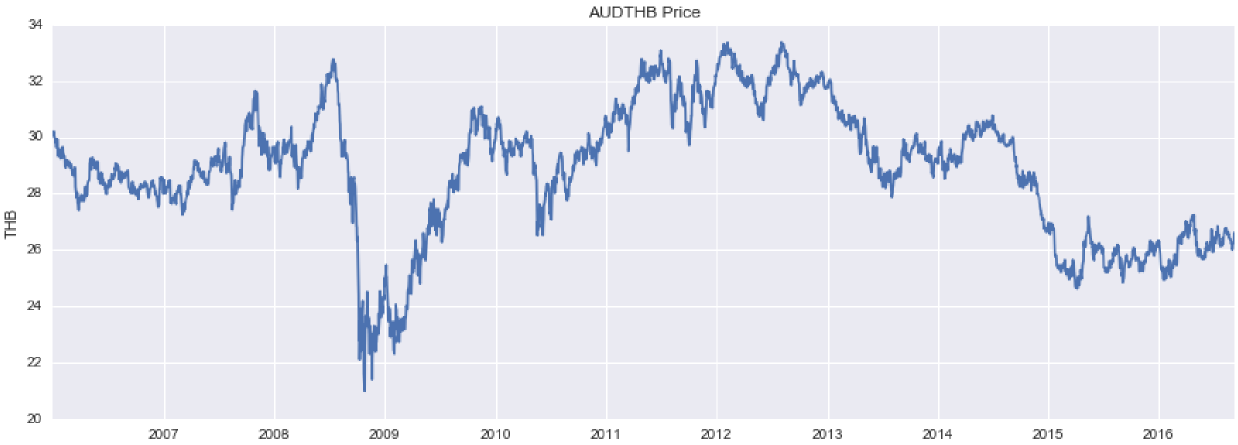


Figure 1 – AUDTHB Price from 1Jan06 to 9Sep16

Using visualization as part of Exploratory Data Analysis, we can also detect a characteristic that can be found in financial market time series data, known as Interrupted Time Series (ITS). This is when the time series can abruptly dip or rise, in response to 'traumatizing' events.

In the case of the 10 years history of the AUDTHB, we can visually identify from Figure1 two such events that happened in mid-2008 and mid-2014. My google research was able to relate the timing of these 2 events with:

1. Mid-2008 is when the GFC started to hit the AUDTHB. One would think that both AUD and THB are equally impacted by this global event, so the AUDTHB price should perform on even keel. However, during the GFC, the Australian economy was robust because of strong trade with China, particularly fuelling the Chinese demand for iron and coal. Hence, when the Thai economy went through the woes during the GFC, the Australian economy 'escaped' the GFC enjoying good financial growth, employment and a relatively strong AUD.

2. Mid-2014 is when the AUD was slaughted by a combination of several macro- and micro-economic events, all happening at once, mainly:

   - Commodity prices - dropped significantly, which impacted Australia as the largest exporter of iron, coal and copper.
   - US economy recovery - the US Federal Reserve stimulus started to see results, and with the return of US financial market confidence, the greenback rallied forward.
   - Low interest rate in Australia – at 2.5% OCR (Official Cash Rate), which is what the Australian Reserve Bank charges commercial banks for overnight loans, cash rate was so low and retail loans were also very low, and AUD became cheap during that time.

The occurrence of these kind of events is a challenging problem for time series analysis, even with rigorous stationarising, as this can then overfit a model. There are specialisations of ITS analytics to handle time series interrupted events, such as available from SAS packages, but this is beyond the scope of this project.

# 4 APPROACH

My approach to the data analysis will follow the common industry FX analysts' view of looking at influences of FX prices based on:

- Fundamental Analysis
- Technical Analysis

Fundamental analysis uses data that is available from local and global, macro and micro-economics trends and events, to explain the behavior of FX currency prices. For this, I will be using the data that I have for the currency pairs that are likely to influence AUDTHB, as well as gold, crude oil and stock prices. This class of analysis are more easily understood by the FX layman.

Technical analysis is more dependent on algorithms and machine learning techniques to try and more accurately model FX price movements. The analysis may use other source data (other currency pairs, oil and stock price, etc.) or derived statistics (eg. standard errors, mean values, etc.) as features to build the models. Technical analysis of FX can get quite complex, even as the 'secret sauce' of professional FX traders, well beyond the understanding of the layman.

The following sections will discuss the work that I carried out for both Fundamental Analysis and Technical Analysis of AUDTHB currency pair.

# 5 FUNDAMENTAL ANALYSIS

Two pieces of work was carried out for the fundamental analysis of AUDTHB.

## 5.1 CORRELATION WITH OTHER KEY CURRENCY PAIRS

The other key currency pairs of interest are:

1. AUDJPY – JPY is the main currency traded in the AsiaPac trading bloc
2. AUDUSD – USD is the main currency in the US trading bloc, but with strong global influence
3. THBJPY – is a cross pair in the AsiaPac bloc, where JPY is main currency traded.
4. USDTHB – USD has a strong global influence
5. USDJPY – Both USD and JPY are the main currencies traded in their respective blocs.

The chart below was plotted to provide the visualization of the 6 currency pairs. No anomaly can be observed, although scale has masked the trends of AUDUSD, THBJPY, and to a lesser extent AUDTHB and USDTHB.

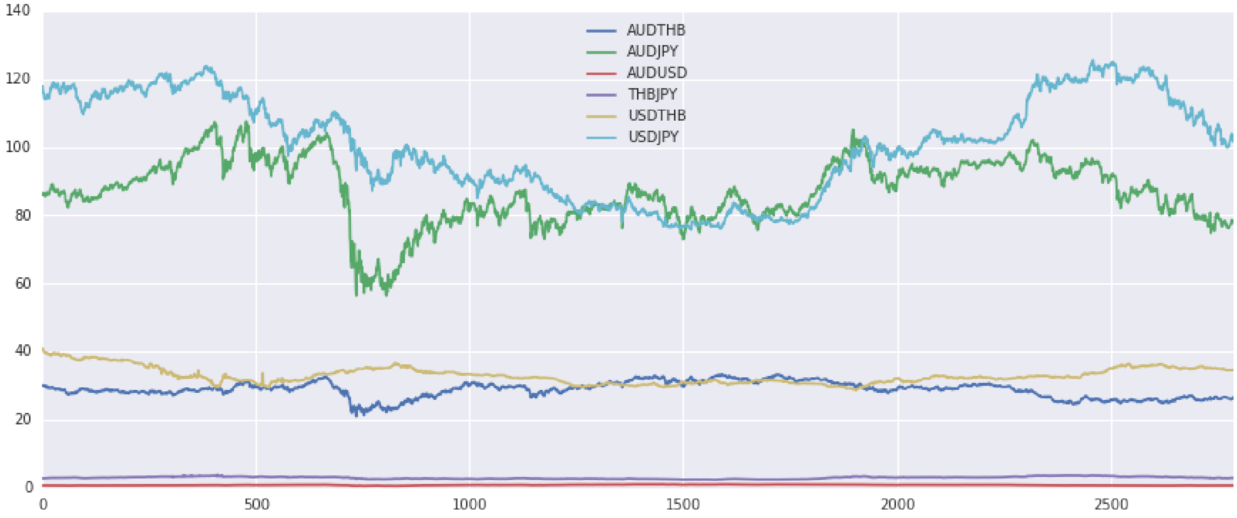


Figure 2 – Plot of 6 currency pairs

A multivariate regression was then carried out to determine R-square values. The result is shown in the table and heatmap chart below:

```
 -------------- CURRENCY PAIRS CORRELATION MATRIX ---------------
          AUDTHB    AUDJPY    AUDUSD    THBJPY    USDTHB    USDJPY
AUDTHB  1.000000  0.324982  0.904801 -0.393204 -0.572754 -0.552115
AUDJPY  0.324982  1.000000  0.234446  0.736821 -0.234069  0.510160
AUDUSD  0.904801  0.234446  1.000000 -0.408834 -0.832943 -0.709598
THBJPY -0.393204  0.736821 -0.408834  1.000000  0.173804  0.886822
USDTHB -0.572754 -0.234069 -0.832943  0.173804  1.000000  0.583960
USDJPY -0.552115  0.510160 -0.709598  0.886822  0.583960  1.000000
```

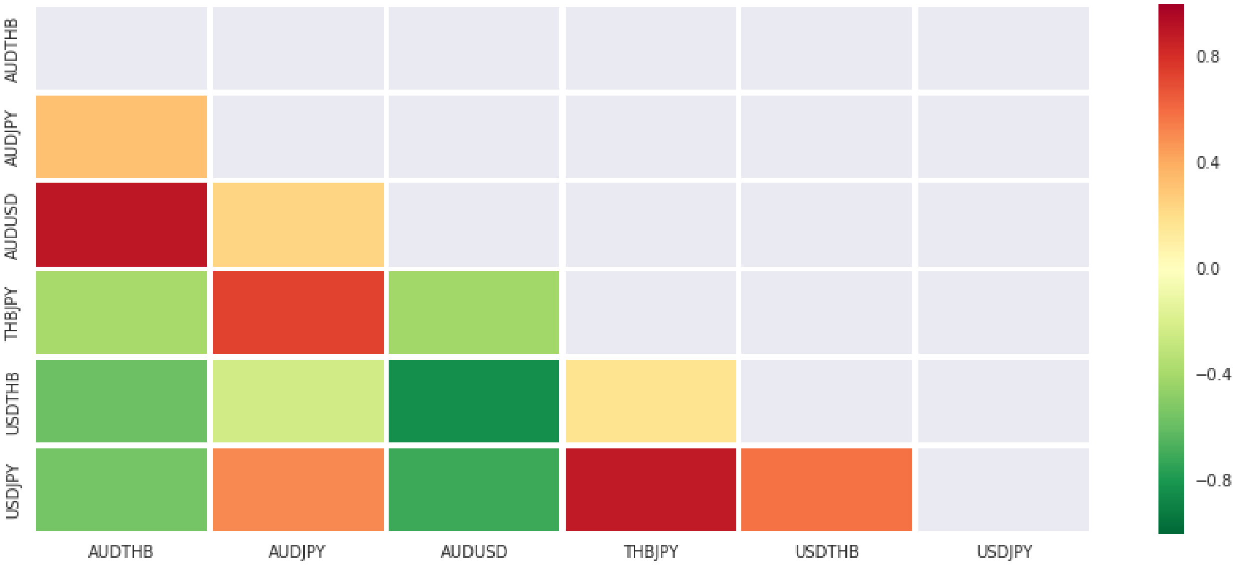


Figure 3 – Currency Pairs Correlation Heatmap

### 5.1.1 Observations

1. AUDTHB and AUDUSD shows the strongest correlation (R-square=0.90)
2. THBJPY and USDJPY also has a strong correlation (R-square=0.88)
3. USDTHB and AUDUSD shows a strong negative correlation (R-square=-0.83)

### 5.1.2 Interpretation

- AUDTHB is strongly influenced by AUDUSD and moderately by AUDJPY. This means AUD although in AsiaPac bloc is more correlated with US trends and events, ie. when AUDUSD goes up, AUDTHB also goes up. AUD is the stronger currency compared to THB.
- THBJPY is strongly influenced by USDJPY. This means although THB and JPY are in the same AsiaPac bloc, THB rallies more strongly with USD than with JPY.
- USDTHB and AUDUSD are negatively correlated. This when the price for one goes up, the other goes down.

## 5.2 Correlation With Macro Economic Factors

The macro-economic factors considered are:

1. Crude Oil prices – based on OPEC crude oil in USD
2. Gold Prices – based on AUD
3. Stock Prices – based on Nasdaq day close prices, in USD.

The chart below was plotted to provide the visualization of the commodity and stock trends. Stock prices have been bullish for past 8 years, whereas gold price increase have been less significant.

The scale has masked the trends of AUDTHB and crude oil prices, so the data was rescaled and the adjusted data re-plotted.

The new plot shows higher variances for oil prices with significant peaks and dips during the past 10 years.

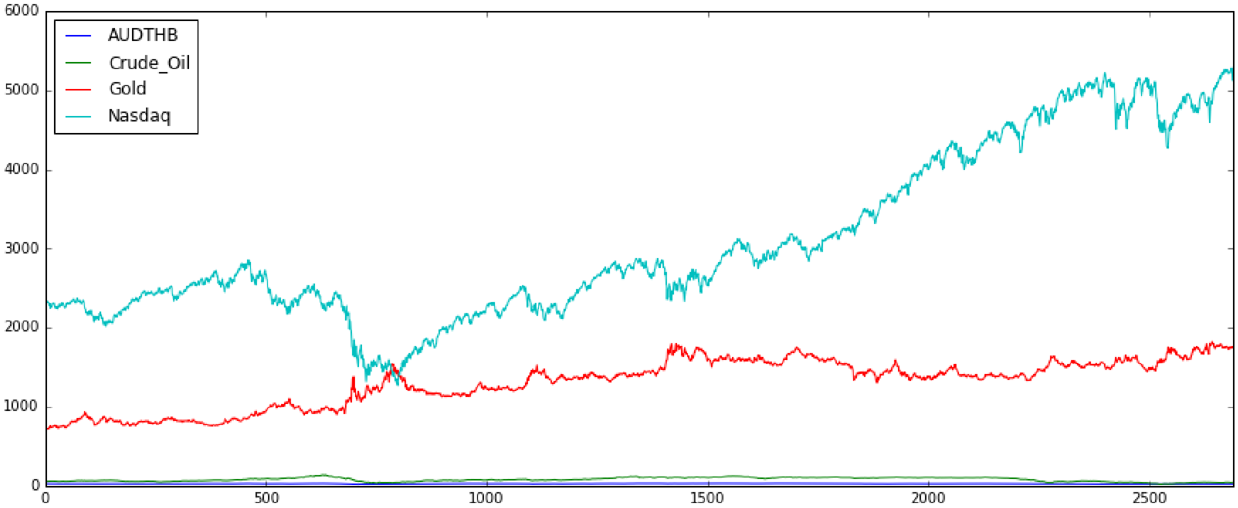


Figure 4 – AUDTHB vs Crude Oil, Gold and Nasdaq Stock Prices

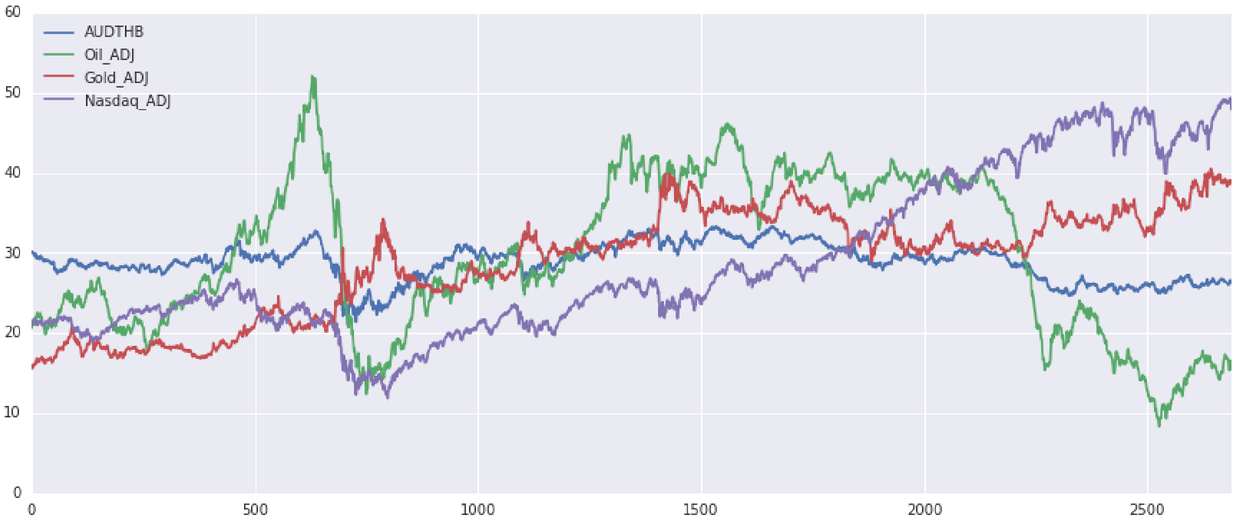


Figure 5 – AUDTHB vs Crude Oil, Gold and Nasdaq Stock Prices (rescaled data)

A multivariate regression was then carried out to determine R-square values. The result is shown in the table and heatmap chart below:

```
-- AUDTHB AND MACRO FACTORS CORRELATION MATRIX ---
              AUDTHB  Crude_Oil      Gold    Nasdaq
AUDTHB      1.000000   0.832489  0.021426 -0.208851
Crude_Oil   0.832489   1.000000  0.137366 -0.151656
Gold        0.021426   0.137366  1.000000  0.553728
```

```
Nasdaq    -0.208851  -0.151656  0.553728  1.000000
```

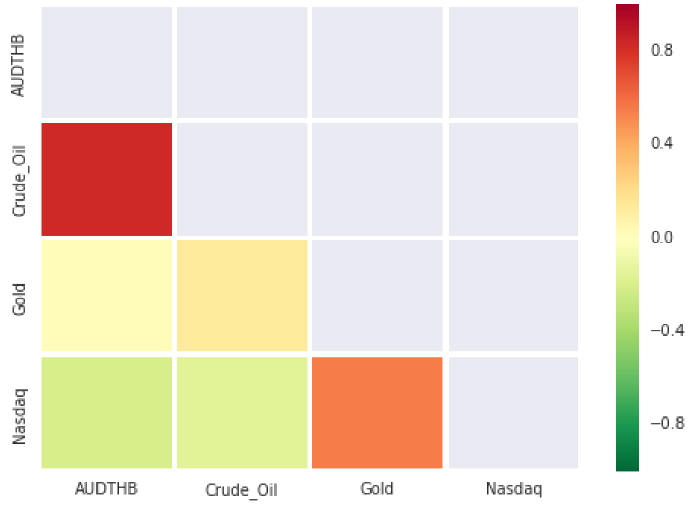


Figure 6 – AUDTHB and Commodity/Stock Prices Correlation Heatmap

### 5.2.1 Observations

1. AUDTHB and Crude Oil shows a strong correlation (R-square=0.83)
2. AUDTHB is poorly correlated with Gold (R-square=0.02) and stock prices (R-square=0.21)

### 5.2.2 Interpretation

- AUDTHB is strongly influenced by crude oil prices. This may be due to the fact that oil prices around the world is tagged to USD. Australia is an oil importer, and as shown in previous analysis AUDTHB is strongly influenced by AUDUSD. So when oil prices goes up, AUDTHB will also go up.
- Unlike popular belief, gold and stock prices are poor indicators of FX rates.

## 6 TECHNICAL ANALYSIS – LINEAR METHODS

Several linear methods were used during technical analysis to determine a well fitted model for the AUDTHB currency pair:

- Mean Variable Model
- Linear Trend Model
- Random Walk Model
- Moving Average / Smoothing Model
- Exponential Smoothing Model
- Improved Linear Trend Model with AUDTHB Price Regressor

- Improved Linear Trend Model with Crude Oil Price Regressor
- Improved Linear Trend Model with AUDTHB and Crude Oil Price Regressors

These models are generally used for trending of historical time series, rather than for forecasting (which will be covered in the advanced linear models in the next section).

Linear models work best with data that are well aligned with the Central Limit Theorem. If not, the raw data needs to be transformed.

Like all FX time-series data, the AUDTHB variance is quite high (see Figure 7 below). A histogram plot of raw AUDTHB data (Figure 8)  also validates the less than normal distribution.

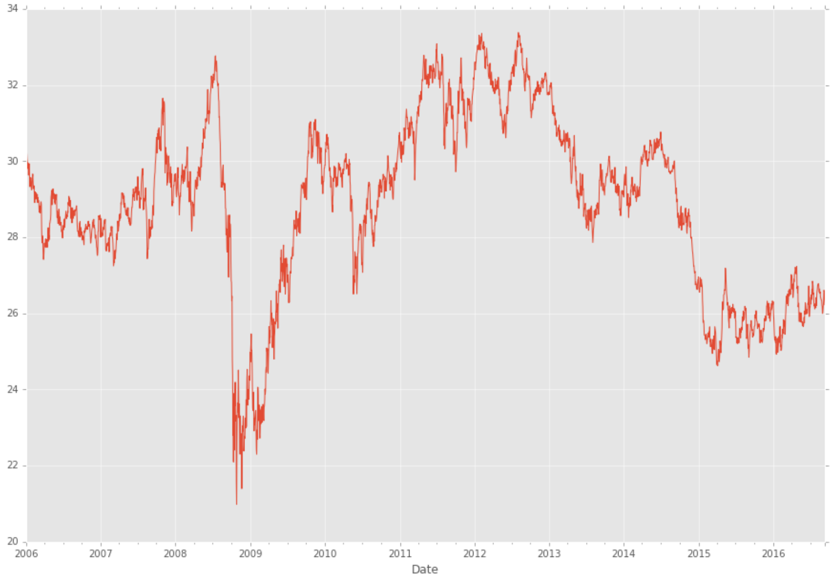


Figure 7 – Time Series of AUDTHB Raw Data Has High Variance

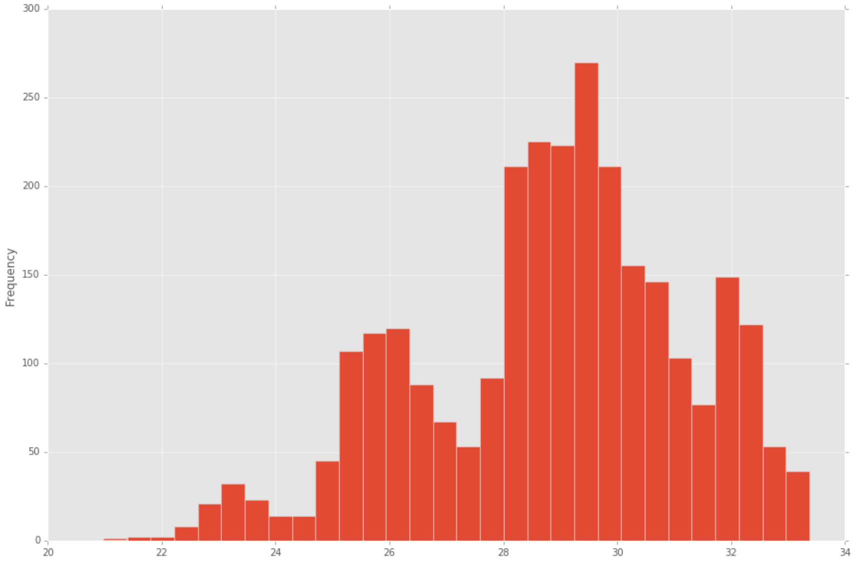


Figure 8 – Histogram of AUDTHB Raw Data

Since log transformations are typically used to help stabilize the variance of a time series data to make them more suitable for linear modeling methods, this was applied to the AUDTHB data. Figure 9 below shows a slight improvement in the histogram, and the AUDTHB log data can be further improved during the later steps on linear modeling.

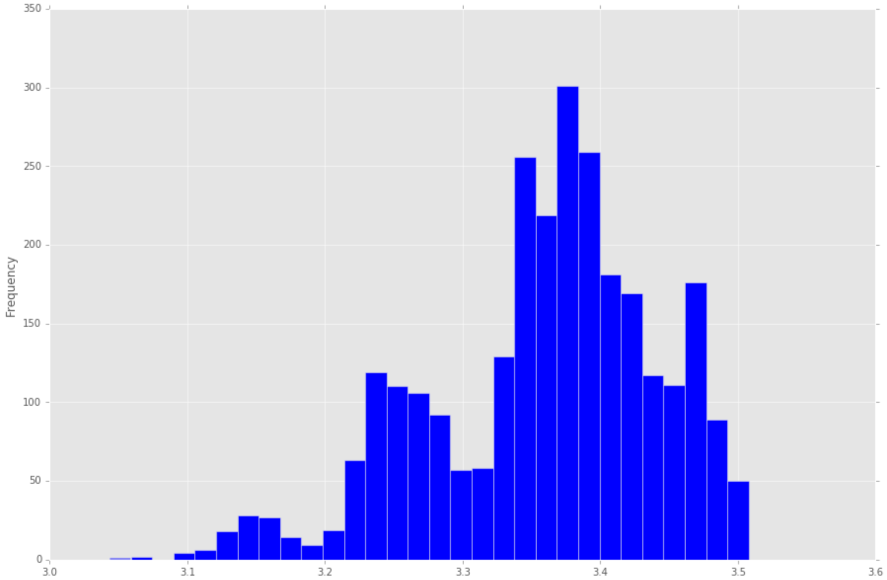


Figure 9 – Histogram of AUDTHB Log Data

## 6.1 Model1 – Mean Variable

This model is a simple calculation of the AUDTHB population mean.

```
data2.plot(kind="line", y = ["AUDTHB", "AUDTHB_mean"])
```

The objective is to provide a baseline model for comparison with other later models. In particular is the calculation of the Root Mean Square Error (RMSE) term, which was captured in a table of comparison to help determine which model has the best fit.

RMSE can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable. Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and is the most important criterion for fit if the main purpose of the model is prediction.

The RMSE is calculated according to this function call:

```
# use Root Mean Squared Error (RMSE) to calculate the error values of this model.
def RMSE(predicted, actual):
   mse = (predicted - actual)**2
   rmse = np.sqrt(mse.sum()/mse.count())
   return rmse
```

Figure 10 below shows the plot of the mean variable model.

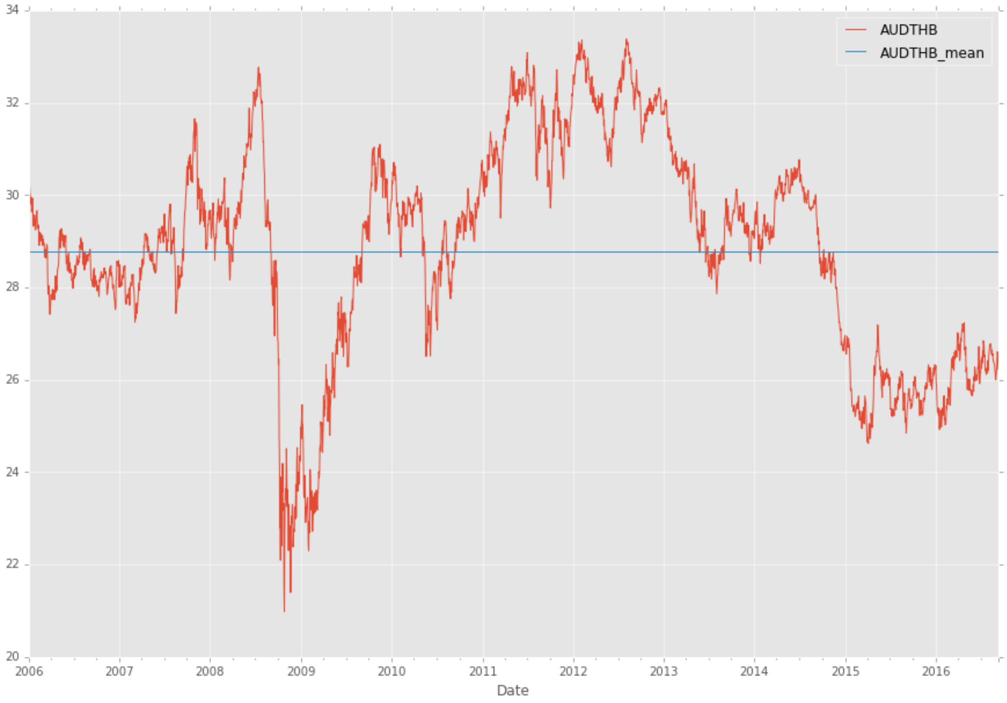


Figure 10 – Chart of Model 1

### 6.1.1 Observations

1. Mean AUDTHB Price for the >10 years period (2006 – 2016) is 28.76
2. RMSE = 2.33

### 6.1.2 Interpretation

- AUDTHB has been quite consistent around the mean price of 28 during the past 10 years.

## 6.2 MODEL2 – LINEAR TREND

This model is a straight forward linear regression, y = mx + c model.

```
# plotting a linear trend (regression) model between AUDTHB_log and Date.
model_linear = smf.ols('AUDTHB_log ~ timeIndex', data = data2).fit()
```

Figure 11 below shows the plot of the model.

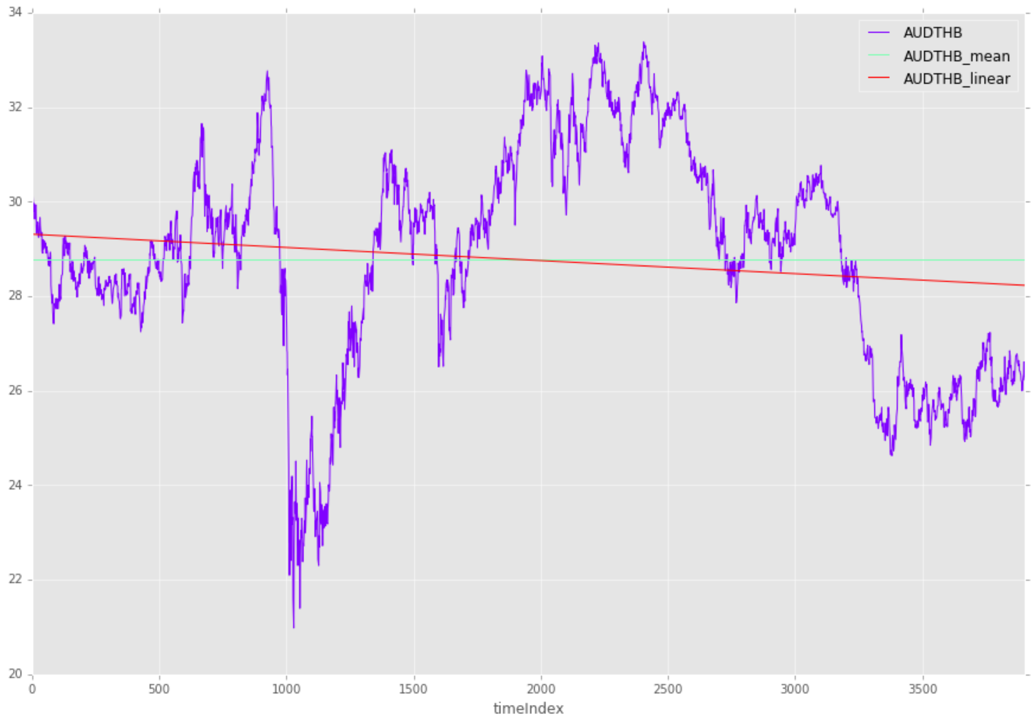


Figure 11 – Chart of Models 1 - 2

A further test was done to plot the residuals of the linear model. The result is shown in Figure 12.

```
#plotting the resid to test
model_linear.resid.plot(kind = "bar")
```

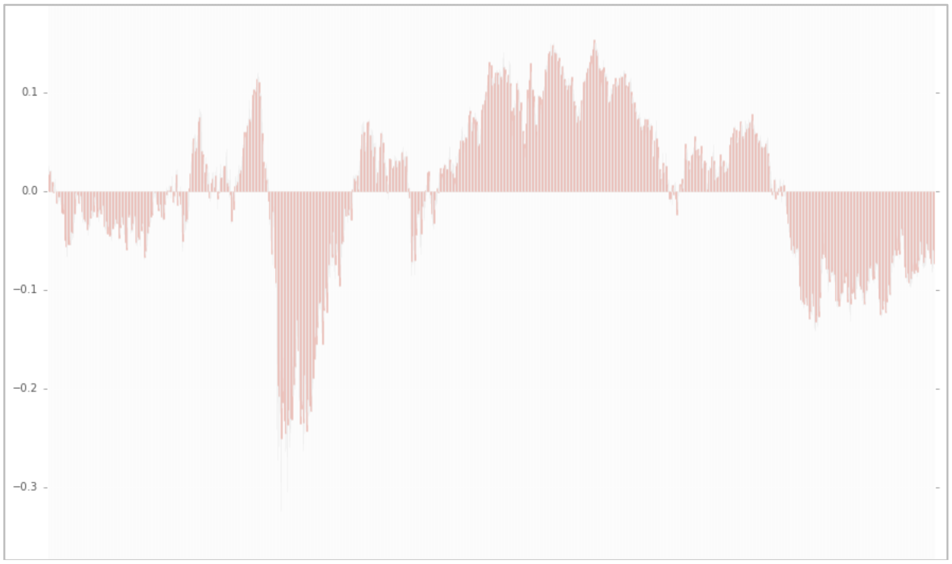


Figure 12 – Plot of Linear Model Residuals.

### 6.2.1 Observations

1. RMSE = 2.32
2. Analysis of residuals shows that the model has a tendency to make an error of the same sign for many periods in a row.

### 6.2.2 Interpretation

- RMSE of 2.32 is not a good model fit.
- If the model has succeeded in extracting all the "signal" from the data, there should be no pattern at all in the residual error plot; the error in the next period should not be correlated with any previous errors. Hence, the linear trend model has obviously failed the autocorrelation test in this case.

## 6.3 Model3 – Random Walk

The best strategy for a time series that shows irregular growth may not be to try to directly predict the level of the series at each period (*Yt*). Rather, it may be better to try to predict the change that occurs from one time period to the next (*Yt - Yt-1*).

Thus, it may be better to look at the first-order differencing of the series, to see if a predictable pattern can be found there, as in a Random Walk model.

For purposes of one-period-ahead forecasting, it is just as good to predict the next change as to predict the next level of the series, since the predicted change can be added to the current level to yield a predicted level.

The simplest case of such a model is one that always predicts that the next change will be zero, as if the series is equally likely to go up or down in the next period regardless of what it has done in the past.

In this model, I created a 5-days feature for the random walk period. The assumption is that 5 days is a good FX trading lag time for the market to fully respond to changes. Moreover, the trading blocs work on 5 trading days a week.

```
# create the random walk shift = 5 days feature
  data2["AUDTHB_logShift1"] = data2.AUDTHB_log.shift(5)
  data2.head()
  data2["AUDTHB_logDiff1"] = data2.AUDTHB_log - data2.AUDTHB_logShift1
  data2.AUDTHB_logDiff1.plot()
```

The Figure below shows the plot of the 5-days first-order difference log plot.

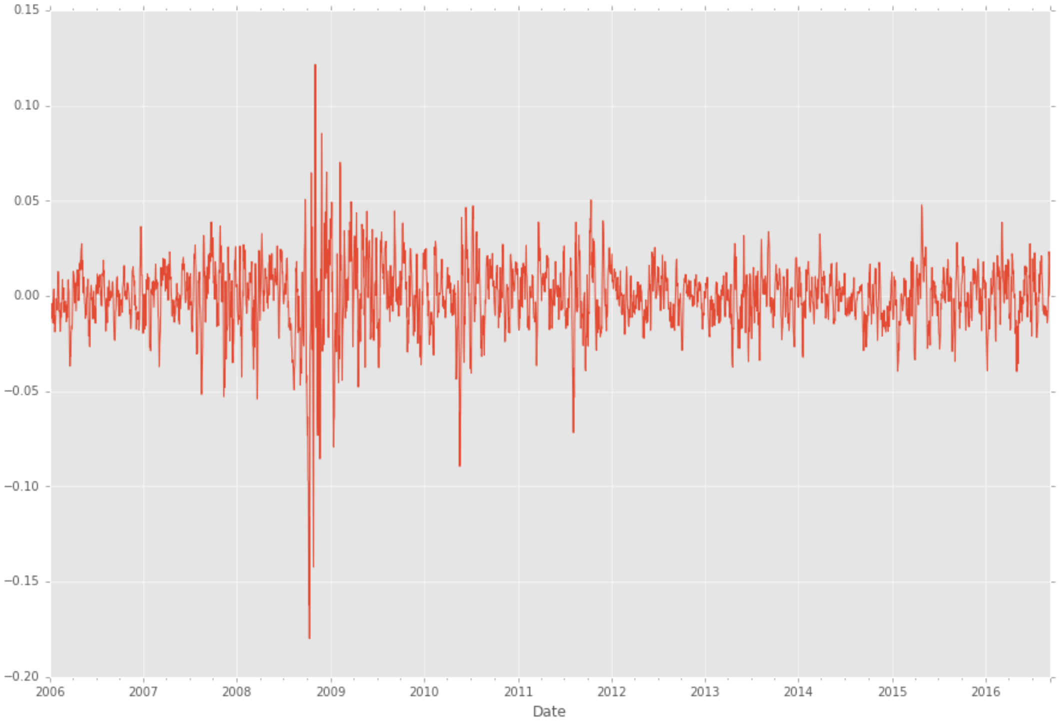


Figure 13 – Chart of 5-days Differencing of AUDTHB Log Data

The log data was converted back to original values and the model fit is shown in Figure 14 below:

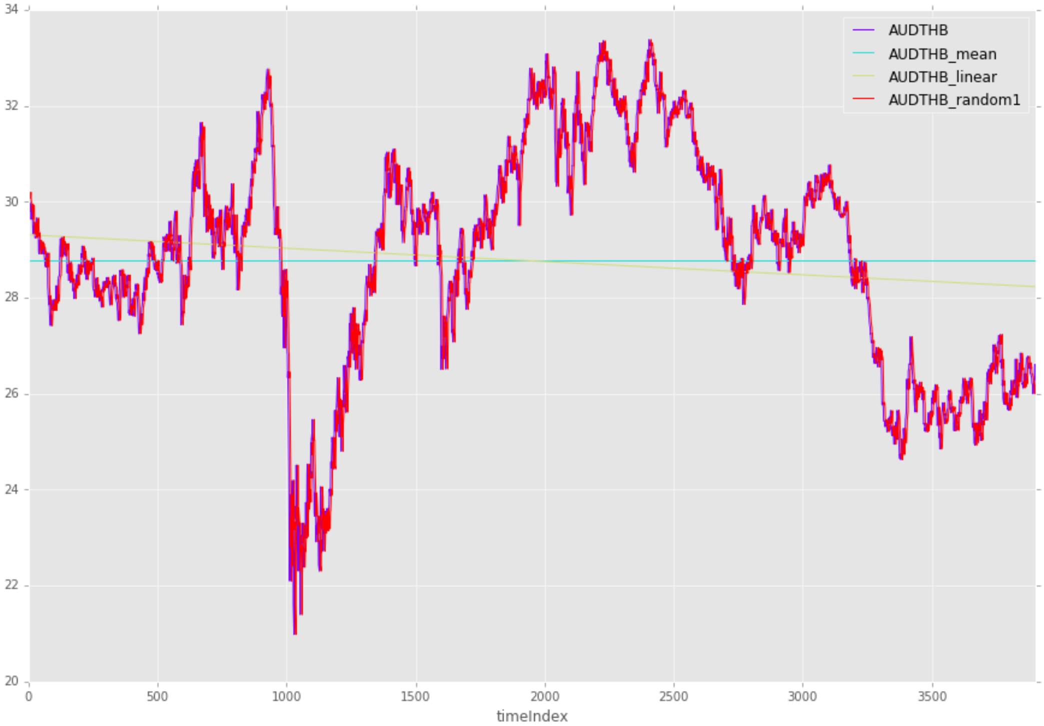


Figure 14 – Chart of Models 1 – 3

### 6.3.1 Observations

1. RMSE = 0.48
2. The 5-days first-order differencing log data appears to have stationarised the AUDTHB time-series.

### 6.3.2 Interpretation

- RMSE of 0.48 is vast improvement in the model fit
- Compared to the previous 2 basic linear models, the random walk model uses a sequencing technique to predict the next value based on a window of previous 5 days pattern. The lag value does not seem to have much impact, as the model fits closely with the pattern of original data values.

## 6.4 Model4 – Moving Average / Smoothing Model

Beyond basic mean model, linear model and random walk model, non-seasonal patterns and trends can be extrapolated using a moving-average or smoothing model.

The assumption behind averaging and smoothing models is that the time series is locally stationary with a slowly varying mean. It works by taking a moving average to estimate the current value of the mean, and then use that as the forecast for the near future. This can be considered as a compromise between the mean model and the random-walk-without-drift-model. The same strategy can be used to estimate and extrapolate a local trend.

A moving average is often called a "smoothed" version of the original series because short-term averaging has the effect of smoothing out the bumps in the original series. By adjusting the degree of smoothing (the window of the moving average), some kind of optimal balance between the performance of the mean and random walk models can be achieved.

The Moving Average model expression is:

$$ \hat{y_t} = \frac{y_{t-1} + y_{t-2} + y_{t-3} + ... + y_{t-m}}{m} \\$$

Since the raw data is a daily time series of > 10 years, a monthly (30-days) moving average is reasonably assumed.

```
data2['AUDTHB_log30days'] = pd.Series.rolling(data2.AUDTHB_log, window = 30,
center=False).mean()
```

The Figure below shows the model fit as compared to all the other previous models.

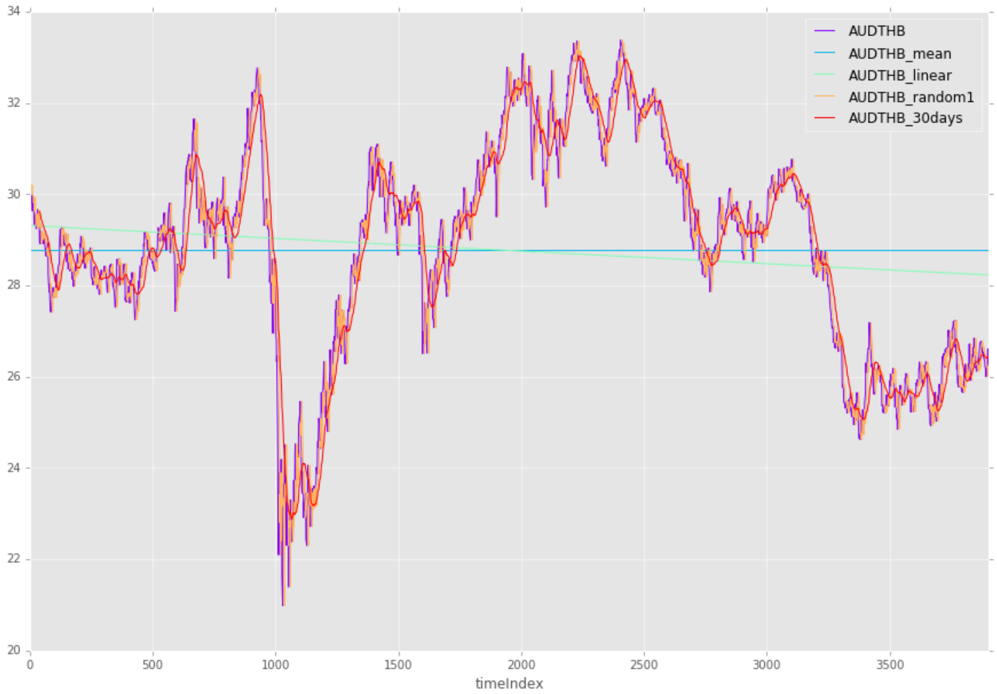


Figure 15 – Chart of Models 1 – 4

### 6.4.1 Observations

1. RMSE = 0.64
2. The moving average model fit is very similar to the random walk model.

### 6.4.2 Interpretation

- RMSE of 0.64 is vast improvement in the model fit
- As with random walk, the moving average model uses a sequencing technique. Even though a 30-days window is used here, and the fitted line is smoother, the lag value does not seem to have much impact and trends are very close to the original AUDTHB time series.

## 6.5 Model5– Exponential Smoothing

Instead of equally weighting each of the observation, the Exponential Smoothing model give more weightage to the recent observations and less to the older ones. This is done by the using a smoothing variable like alpha:

$$ \hat{y_t} = \alpha y_{t-1} + (1-\alpha)\hat{y_{t-1}} \\$$

To be consistent with the previous moving average model, a 30-days half-life is used for the calculation of the alpha value.

*# get alpha value, assume halflife = 30 days, to be consistent with previous model*

*halflife = 30*
*alpha = 1 - np.exp(np.log(0.5)/halflife)*
*alpha*

To model is derived from the expression below and plotted in Figure 16.

*model_ES30_forecast = alpha * y_exp + (1 - alpha) * y_for*

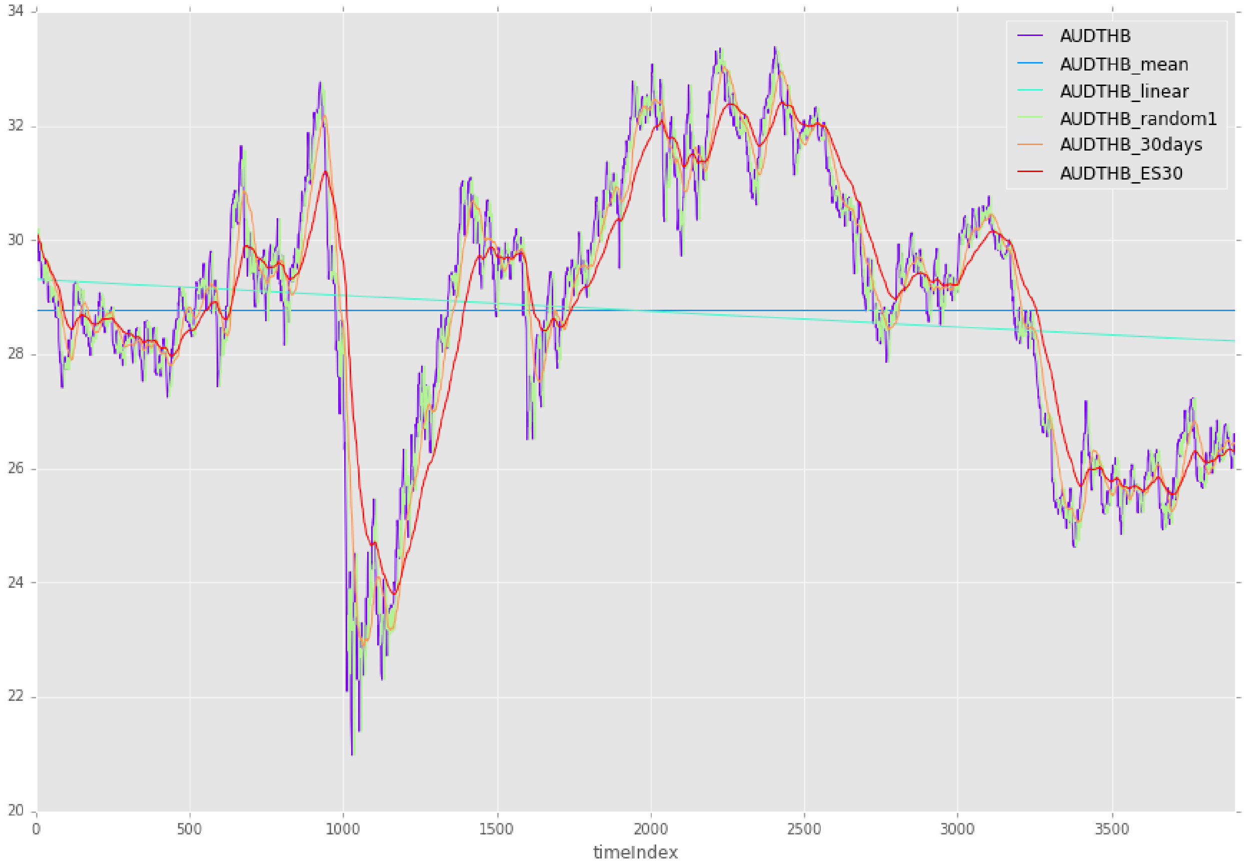


Figure 16 – Chart of Models 1 – 5

### 6.5.1 Observations

1. RMSE = 0.94
2. The Exponential Smoothing model fit is smoother and follows the same trend as AUDTHB time series.

### 6.5.2 Interpretation

- RMSE of 0.94 indicates a poorer fit than the moving average and random walk model.
- Like the previous models using window sequencing to predict the next value, the lag value does not seem to have much impact, as the model follows the original AUDTHB time series trend quite closely.

## 6.6 Model6 – Improve on Linear Model with AUDUSD as additional regressor

Linear modelling can be improved by adding more independent variables, or regressors, to the model.

Given that AUDTHB is shown to be highly correlated with AUDUSD during the fundamental analysis done previously, this will be a good candidate feature to be added to the linear model.

```
## plot linear regression between AUDTHB and timeIndex, with AUDUSD as a regressor
model_linear_USD = smf.ols('AUDTHB_log ~ timeIndex + np.log(AUDUSD)', data = data4).fit()
```

The model was tested again for the multiplicative effects of the independent variables.

```
## What if I multiply the independent variables ?
model_linear_USD2 = smf.ols('AUDTHB_log ~ timeIndex * np.log(AUDUSD)', data = data4).fit()
```

Figure 17 below shows a plot of the model fit with the original AUDTHB data, in comparison with all the other previous models.

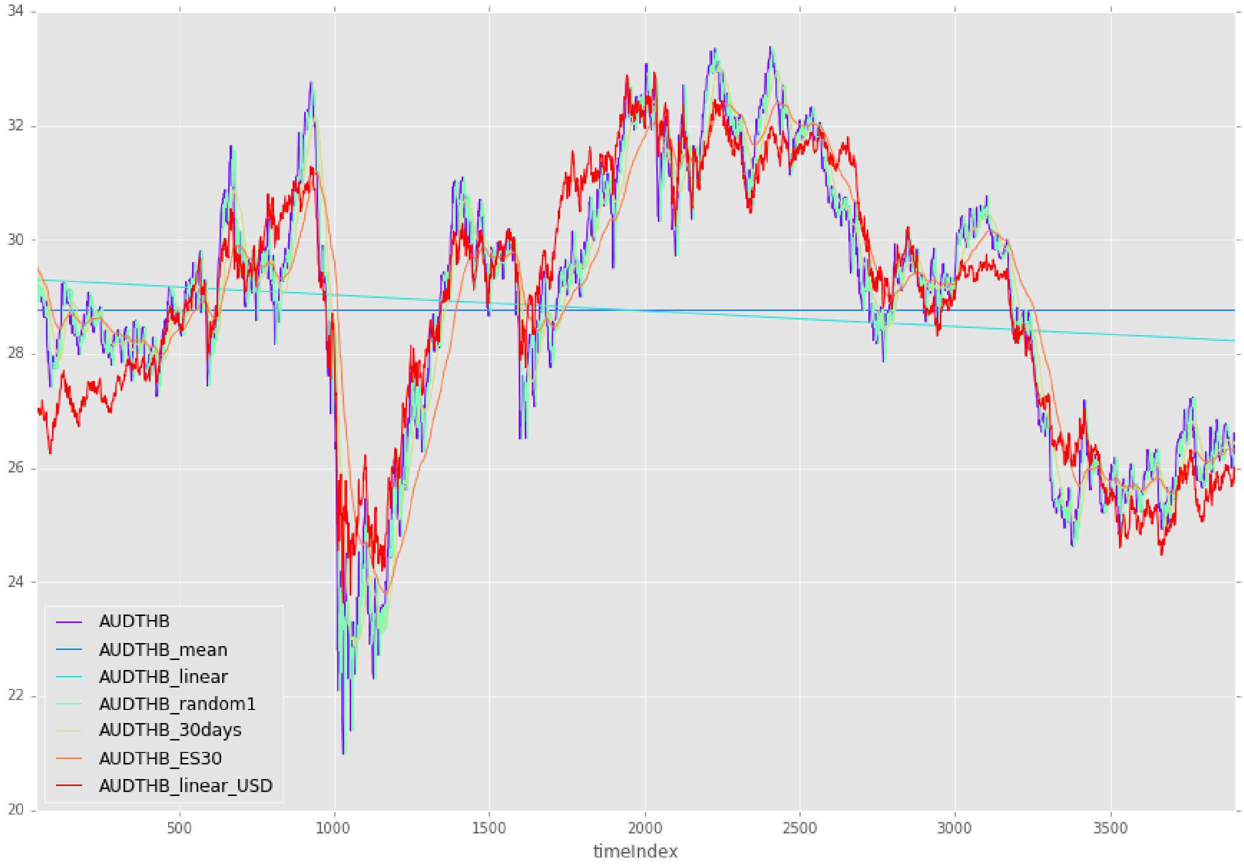


Figure 17 – Chart of Models 1 – 6

### 6.6.1 Observations

1. RMSE = 0.79
2. R-square for additive model = 0.88
3. R-square for additive model = 0.88
4. Fitted model follows the same trend as AUDTHB time series.

### 6.6.2 Interpretation

- RMSE of 0.79 indicates a vast improvement in the linear fit by bringing in a strongly correlated feature such as AUDUSD.
- The improved linear model fit is as good as the models using sequencing, probably due to the strong AUDTHB and AUDUSD correlation.
- This strong correlation is probably validated by similar R-square values for either additive or multiplicative tests of the independent variables in the model.

## 6.7 MODEL7 – IMPROVE ON LINEAR MODEL WITH CRUDE OIL AS ADDITIONAL REGRESSOR

From previous fundamental analysis work, AUDTHB was also shown to be tracking extremely well with OPEC crude oil prices.

This model will be similar to Model 6, but using Crude Oil as the additional independent variable for the regression model.

```
# test for additive effect
model_linear_Oil = smf.ols('AUDTHB_log ~ timeIndex + np.log(Oil_Price)', data = data5).fit()

# test for multiplicative effect
model_linear_Oil2 = smf.ols('AUDTHB_log ~ timeIndex * np.log(Oil_Price)', data = data5).fit()
```

Figure 18 below shows a plot of the model fit with the original AUDTHB data, in comparison with all the other previous models.

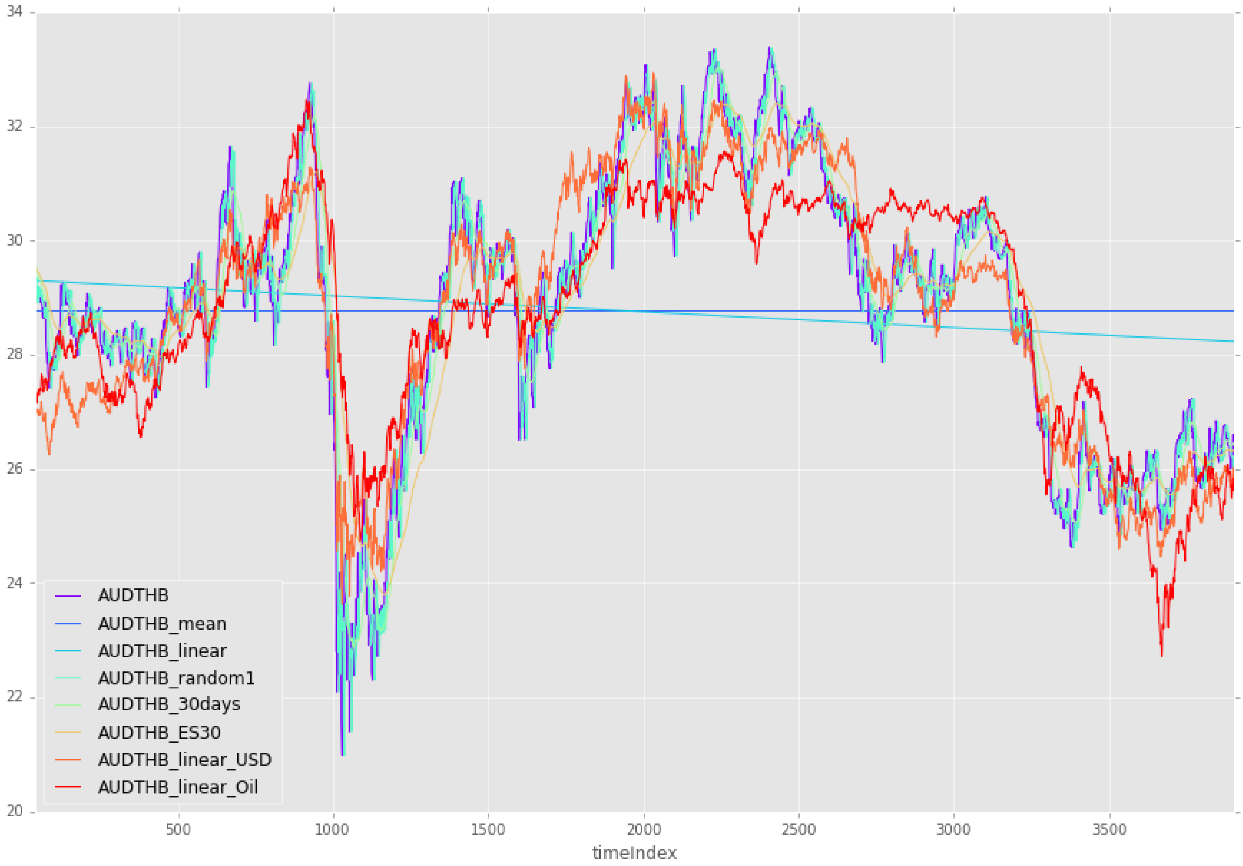


Figure 18 – Chart of Models 1 – 7

### 6.7.1 Observations

1. RMSE = 1.27
2. R-square for additive model = 0.70
3. R-square for additive model = 0.72
4. Fitted model follows the same trend as AUDTHB time series.

### 6.7.2 Interpretation

- RMSE of 1.27 indicates that whilst there is an improvement over the basic linear fit, using crude oil as a feature does not perform as well as using AUDTHB. This is validate by the stronger correlation between AUDTHB/AUDUSD (R-square=0.90 ), as compared with AUDTHB/Crude Oil (R-square=0.83 )
- The improved linear model fit is as good as the models using sequencing, again probably due to the strong AUDTHB and Crude Oil correlation.
- Very similar R-square values for either additive or multiplicative effect of Crude Oil and AUDTHB validates their strong correlation.

## 6.8 Model8 – Improve on Linear Model with Both AUDUSD and Crude Oil as additional regressors

Will the linear modelling be even further improved by including both AUDUSD and Crude Oil as additional features?

This model is set up to test the hypothesis.

```
# Test additive effect of AUDTHB,  AUDUSD and Oil Price
model_linear_USD_Oil = smf.ols('AUDTHB_log ~ timeIndex + np.log(AUDUSD) + np.log(Oil_Price)',
data = data5).fit()

# Test multiplicative effect of AUDTHB,  AUDUSD and Oil Price
model_linear_USD_Oil2 = smf.ols('AUDTHB_log ~ timeIndex * np.log(AUDUSD) * np.log(Oil_Price)',
data = data5).fit()
```

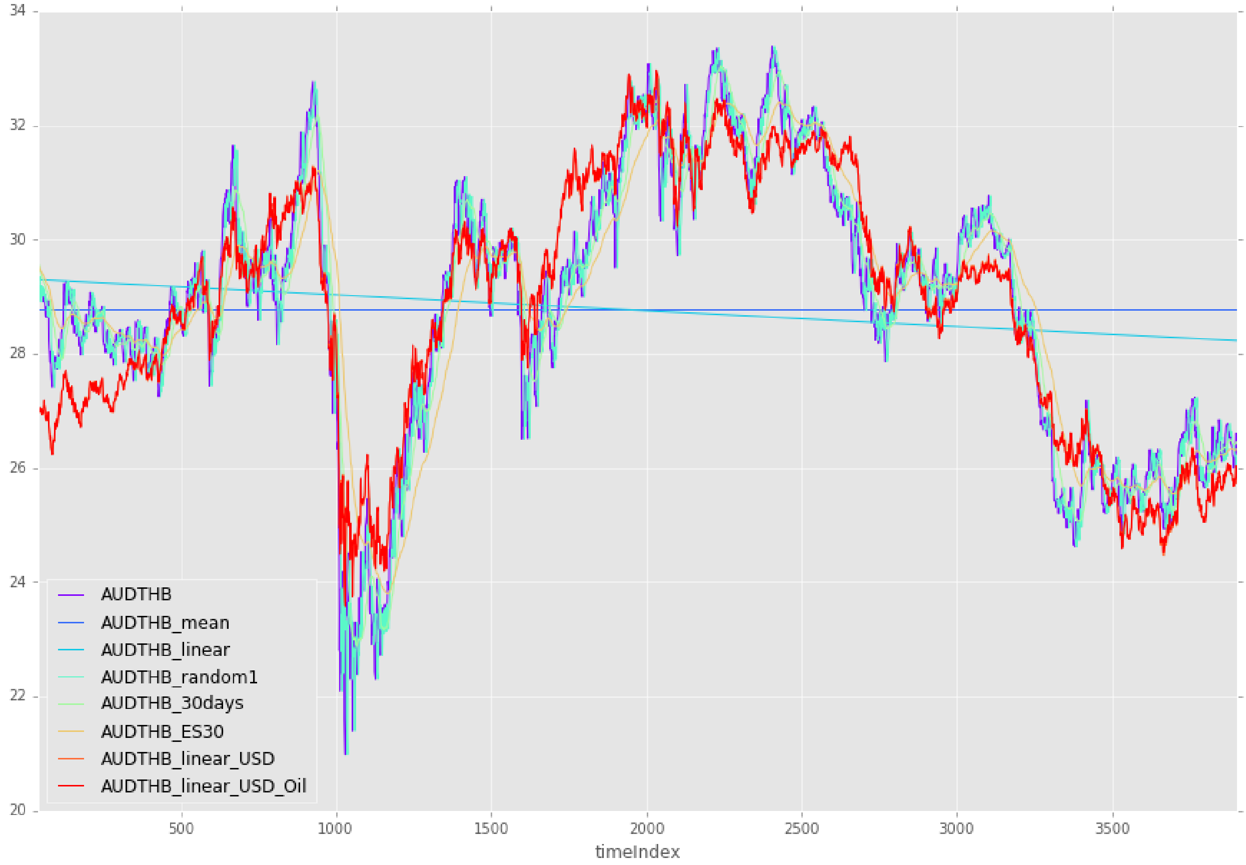


Figure 19 – Chart of Models 1 – 8

### 6.8.1 Observations

1. RMSE = 0.80
2. R-square for additive model = 0.88
3. R-square for additive model = 0.90
4. Fitted model follows the same trend as AUDTHB time series.

### 6.8.2 Interpretation

- RMSE of 1.27 indicates that whilst there is an improvement over the basic linear fit, using crude oil as a feature does not perform as well as using AUDTHB. This is validate by the stronger correlation between AUDTHB/AUDUSD (R-square=0.90 ), as compared with AUDTHB/Crude Oil (R-square=0.83 )
- The improved linear model fit is as good as the models using sequencing, again probably due to the strong AUDTHB and Crude Oil correlation.
- Very similar R-square values for either additive or multiplicative effect of Crude Oil and AUDTHB validates their strong correlation.

# 7 TECHNICAL ANALYSIS – ADVANCED LINEAR METHODS

The final stage of my analysis of AUDTHB price data for the Capstone project includes advanced linear methods that are commonly applied to time series data for forecasting.

The analytics will cover:

- Stationarising method using data transformation
- Stationarising method by decomposition
- ARIMA modeling
- ARIMAX modelling

## 7.1 TIME SERIES STATIONARITY

Most of the time series models work on the assumption that the time series is stationary. Intuitively, we can see that if a time series has a particular behaviour over time, there is a very high probability that it will follow the same in the future. Also, the theories related to stationary series are more mature and easier to implement as compared to non-stationary series

A stationary time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time. Most statistical forecasting methods are based on the assumption that the time series can be rendered approximately stationary (i.e., "stationarized") through the use of mathematical transformations.

A stationarized series is relatively easy to predict: you simply predict that its statistical properties will be the same in the future as they have been in the past.

There are three basic criteria for a series to be classified as stationary series:

- The mean of the series should not be a function of time rather should be a constant.
- The variance of the series should not a be a function of time. This property is known as homoscedasticity.
- The covariance of the *i th* term and the *(i + m) th* term should not be a function of time.

There are 3 major reasons behind non-stationarity:

1. **Trend** - a trend exists when there is a long-term increase or decrease in the data. It does not have to be linear. Sometimes a trend can change direction, when it might go from an increasing trend to a decreasing trend.

2. **Seasonal** - a seasonal pattern exists when a series is influenced by seasonal factors e.g., the quarter of the year, the month, or day of the week. Seasonality is always of a fixed and known period.

3. **Cyclic**: A cyclic pattern exists when data exhibit rises and falls that are not of fixed period. The duration of these fluctuations is usually of at least 2 years.

The mathematical expression is $y_t = S_t + T_t + E_t$ , where:

*y* is the data at period t,
*S* is the seasonal component at period t,
*T* is the trend-cycle component at period t, and
*E* is the remainder (or irregular or error) component at period t.

## 7.2 Stationarity Methods

The underlying principle of Time Series modeling using advanced linear methods is to model or estimate the trend and seasonality in the series, then remove them from the series to get a stationary series. Then statistical forecasting techniques can be implemented on this series.

The final step would be to convert the forecasted values into the original scale by applying trend and seasonality constraints back.

Although log transformation and aggregation (taking average for a time period) are preliminary data wrangling steps that could help stationarise the time series, the main methods to eliminate trend and seasonality are:

1. Differencing – taking the difference of the observation at a particular instant with that at another particular time lag, eg. the previous instant.

2. Decomposition – modeling both trend and seasonality and removing them from the model.

The tests for stationarity can be accomplished by using the following methods.

### 7.2.1 Plotting Rolling Statistics

We can plot the differenced statistic (moving average, moving variance, etc) and see if it varies with time.

Moving average or variance means that at any instant ‘t’, we take the average/variance of the last sequence window, e.g. last 5 days, 30 days (but this is more of a visual technique).

### 7.2.2 Dickey-Fuller Test

This is one of the commonly used statistical tests for checking stationarity.

Here the null hypothesis is that the time series is non-stationary. The test results comprise of a “Test Statistic” and some Critical Values for confidence levels. If the ‘Test Statistic’ is less than the ‘Critical Value’, we can reject the null hypothesis and say that the series is stationary.

## 7.3 Stationarise AUDTHB Time Series Using Rolling Mean and Rolling STD

This was carried out using statsmodels; with the code shown below, for a window of 5 days:

```
#from statsmodels.tsa.stattools import adfuller
from pandas import Series
def test_stationarity(timeseries):

    #Determing rolling statistics
    rolmean = pd.rolling_mean(timeseries, window=5)
    rolstd = pd.rolling_std(timeseries, window=5)

    #Plot rolling statistics:
    #fig = plt.figure(figsize=(12, 8))
    orig = plt.plot(timeseries, color='blue',label='Original')
    mean = plt.plot(rolmean, color='red', label='Rolling Mean')
    std = plt.plot(rolstd, color='black', label = 'Rolling Std')
    plt.legend(loc='best')
    plt.title('Rolling Mean & Standard Deviation')
    plt.show()
```

This output is shown in the Figure below:

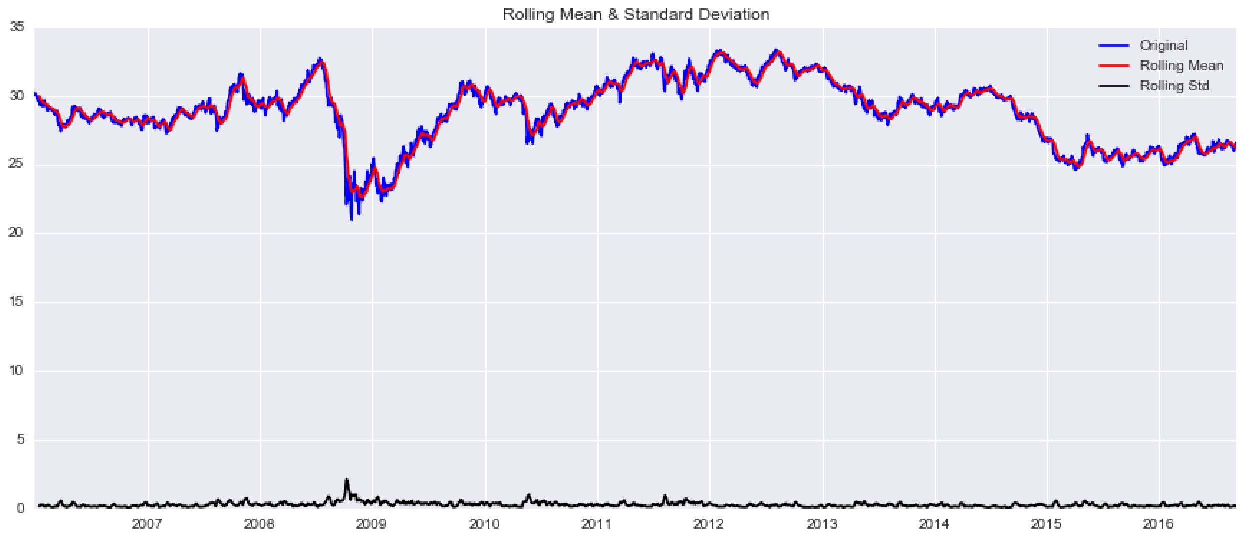


Figure 20 – Plot of Rolling Mean and Rolling STD

The statsmodel output also produces the following Dickey-Fuller statistics:

```
Results of Dickey-Fuller Test:
Test Statistic                  -2.152211
p-value                          0.224045
#Lags Used                       6.000000
Number of Observations Used   2783.000000
Critical Value (1%)             -3.432702
Critical Value (5%)             -2.862579
Critical Value (10%)            -2.567323
dtype: float64
```

### 7.3.1 Observations

1. Test Statistic = -2.15
2. Critical Value (10%) = -2.56

### 7.3.2 Interpretation

- Test Stats of -2.15 is greater than the 10% Critical Value.
- We accept HO: the time series is non-stationary, ie. further work can be done to transform the data to make it stationary for modelling.
- This is validated by visualizing the chart – the rolling mean is tracking the original AUDTHB very closely.

## 7.4 Stationarise Using First Difference

This was carried out using statsmodels; with the code shown below, for a window of 5 days:

```
#create the first order difference data
data2['first_difference'] = data2.AUDTHB - data2.AUDTHB.shift(1)
#print (data2.first_difference.head())
test_stationarity(data2.first_difference.dropna(inplace=False))
```

This output is shown in the Figure below:

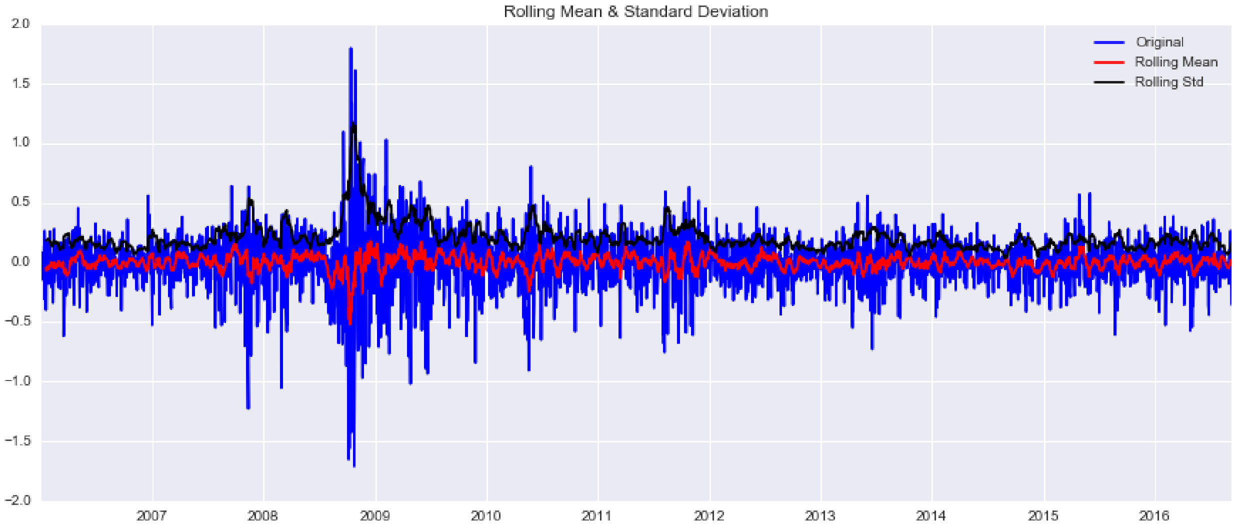


Figure 21 – Plot of First Order Differencing Rolling Mean and Rolling STD

```
Results of Dickey-Fuller Test:
Test Statistic                  -23.218867
p-value                           0.000000
#Lags Used                        5.000000
Number of Observations Used    2783.000000
Critical Value (1%)              -3.432702
Critical Value (5%)              -2.862579
Critical Value (10%)             -2.567323
dtype: float64
```

### 7.4.1 Observations

1. Test Statistic = -23.22
2. Critical Value (1%) = -3.43

### 7.4.2 Interpretation

- Test Stats of -23.22 is lesser than the 1% Critical Value.
- We can reject HO: the time series is stationary
- This is validated by visualizing the chart – the variance of differenced rolling mean is quite constant of the > 10 years time series period.

## 7.5 Time Series Decomposition

Another interesting technique to stationarise a time series is decomposition. This is a technique that attempts to break down a time series into trend, seasonal, and residual factors.

Statsmodels comes with a decompose function out of the box that I used.

```
# decompose the ts data
decomp = sm.tsa.seasonal_decompose(data2.AUDTHB)
fig = plt.figure()
fig = decomp.plot()
fig.set_size_inches(15,10)
```

The output is shown in the Figure below.

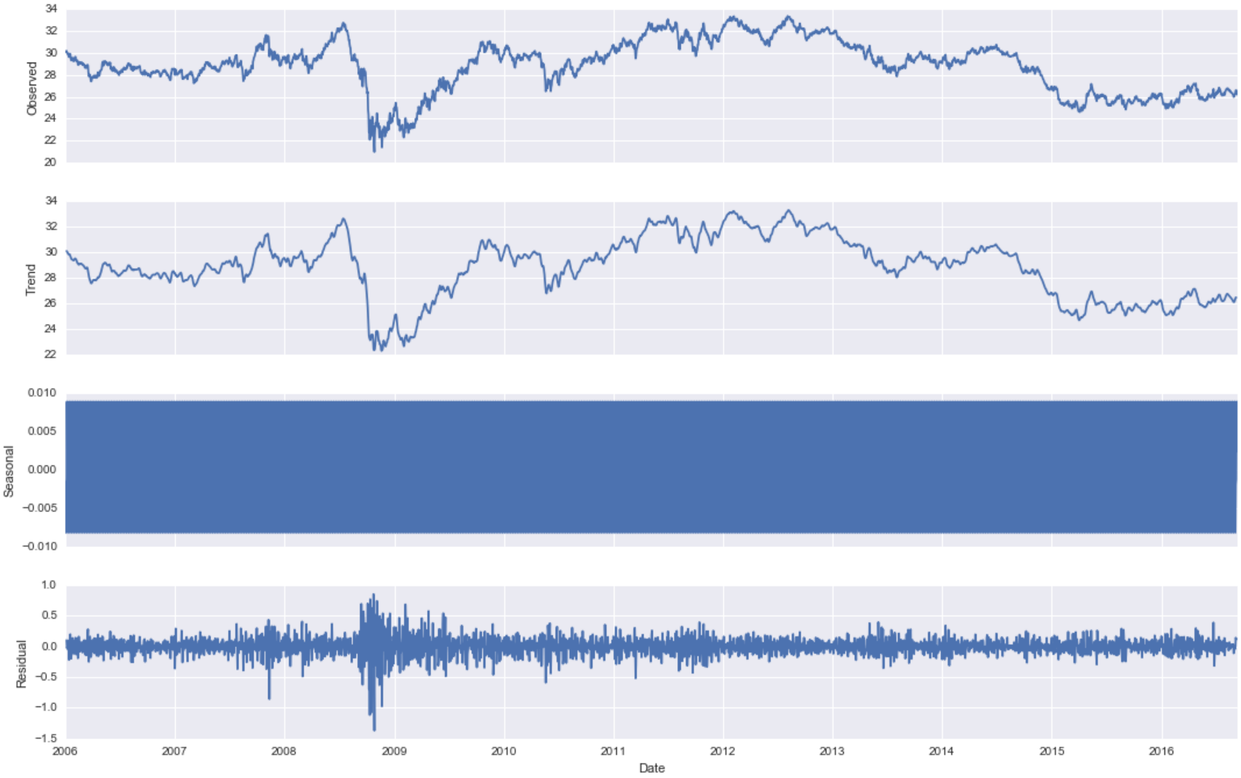


Figure 22 – Decomposition of AUDTHB Time Series

### 7.5.1 Observations

1. Cannot see any real cycle in the data
2. Absolutely no seasonality component.
3. Residual component is flattened out somewhat

### 7.5.2 Interpretation

- FX time series do not exhibit seasonality or cyclic trends, i.e. makes it hard to create useful simple predictive models.
- More can be done to further stationarise the data

## 7.6 ACF AND PACF

Before diving into the ARIMA and ARIMAX modelling, the tuning parameters of these models have to be determined by looking at the autocorrelation and partial autocorrelation factors.

- **Autocorrelation Function (ACF)**: It is a measure of the correlation between the time series with a lagged version of itself. For instance at lag 5, ACF would compare series at time instant 't1'…'t2' with series at instant 't1-5'…'t2-5' (t1-5 and t2 being end points).

- **Partial Autocorrelation Function (PACF)**: This measures the correlation between the time series with a lagged version of itself, but after eliminating the variations already explained by the intervening comparisons. e.g. at lag 5, it will check the correlation but remove the effects already explained by lags 1 to 4.

Since it is impractical to test every possible lag value manually, there is a class of functions that can systematically do this for us. Statsmodel provides these, as shown in the codes that I used below:

```
# ACF and PACF values will be derived from step1 first difference data
# correlation is computed at each lag step that is NOT already explained by previous, lower-order lag steps.
# cos the diff is calculated per day lag, iloc here is 1,
# and it didn't make any difference (tested) if I changed iloc to > 1; the plots shows this as well
fig = plt.figure(figsize=(15,8))
ax1 = fig.add_subplot(211)
fig = sm.graphics.tsa.plot_acf(data2.first_difference.iloc[1:], lags=20, ax=ax1)
ax2 = fig.add_subplot(212)
fig = sm.graphics.tsa.plot_pacf(data2.first_difference.iloc[1:], lags=20, ax=ax2)
```

The output is shown in Figure 23.

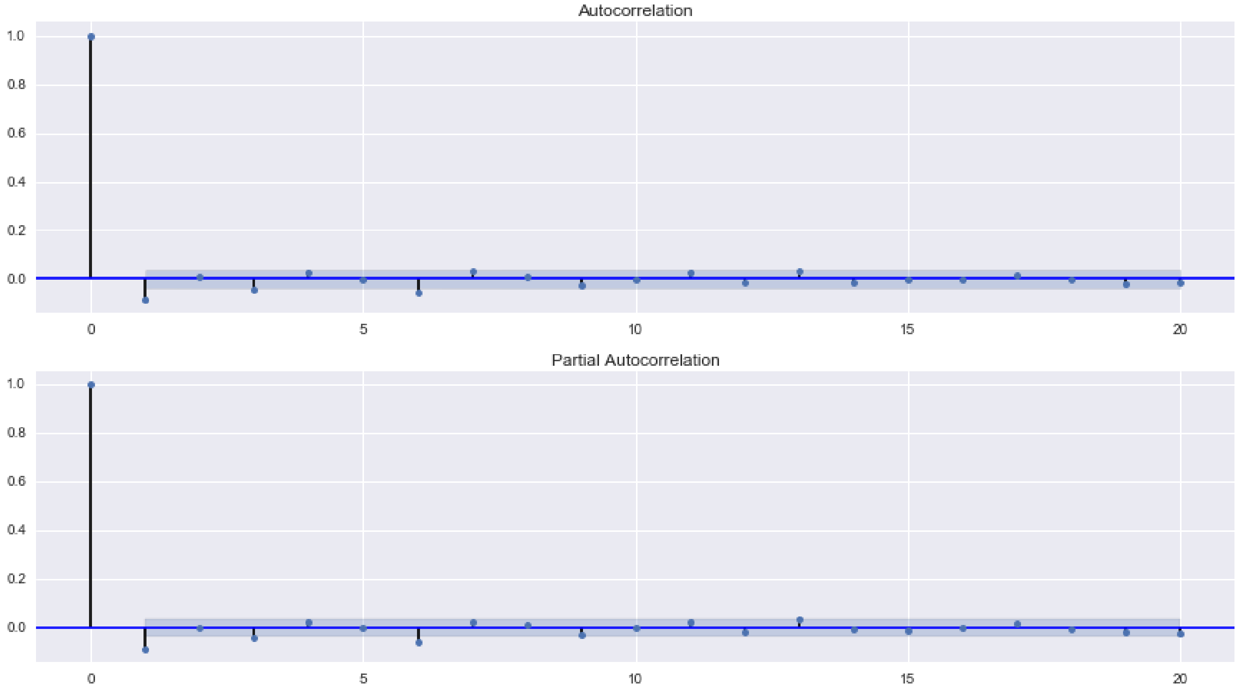


Figure 23 – ACF and PACF Chart

To interpret the charts, the following table can be used to determine the proper terms for AR(p) and MA(q) parameters, bearing in mind that using too many terms can overfit the ARIMA and ARMIAX models.

| Model | ACF | PACF |
|---|---|---|
| White Noise | All zeros | All zeros |
| AR(*p*) | Exponential Decay | *p* significant lags before dropping to zero |
| MA(*q*) | *q* significant lags before dropping to zero | Exponential Decay |
| ARMA(*p,q*) | Decay after *qth* lag | Decay after *pth* lag |

### 7.6.1 Observations

1. ACF shows 1 significant lag before dropping to 0
2. PACF also shows 1 significant lag before dropping to 0

### 7.6.2 Interpretation

- p = 1 and q = 1
- The time series is already differenced by first order, so d = 1
- Hence the optimal (p, d, q) terms is (1, 0, 1)
- The results is not surprising, as in FX data, we do not expect correlation between the change in FX price value from one day to the next.

## 7.7 ARIMA MODEL

If we combine differencing with autoregression and a moving average model, we obtain a non-seasonal ARIMA model.

ARIMA is an acronym for AutoRegressive Integrated Moving Average model ("integration" in this context is the reverse of differencing). The full model can be written as:

- **Number of AR (Auto-Regressive) terms (p):** AR terms are just lags of dependent variable. The term autoregression indicates that it is a regression of the variable against itself. We forecast the variable of interest using a linear combination of past values of the variable. For instance if p is 5, the predictors for *y(t)* will be *y(t-1)….y(t-5).*

- **Number of MA (Moving Average) terms (q)**: MA terms are lagged forecast errors in prediction equation. Rather than use past values of the forecast variable in a regression, a MA model uses past forecast errors in a regression-like model. Each forecast value can be thought of as a weighted moving average of the past few forecast error. For instance if q is 5, the predictors for *y(t)* will be *e(t-1)….e(t-5)* where *e(i)* is the difference between the moving average at *ith* instant and actual value.

  Moving average models should not be confused with moving average smoothing (as in the model 4 and 5 tested previously). A moving average model is used for forecasting future values while moving average smoothing is used for estimating the trend-cycle of past values.

  In MA model, noise / shock quickly vanishes with time, whereas the AR model has a much lasting effect of the shock.

- **Number of Differences (d)**: These are the number of nonseasonal differences, i.e. in this case we took the first order difference. So either we can pass that variable and put d=0 or pass the original variable and put d=1. Both will generate same results.

Going back to Models 3, 4 and 5 from the previous section, I have already observed that the lagged values (5-days, 30-days) do not seem to have much impact. However, I still tried fitting some ARIMA models to find out what they will look like.

The chart below shows the output of a ARIMA (1,0,1) model that I ran. The code used Pyflux's ARIMA function:

```
model = pf.ARIMA(data=data2,ar=1,ma=1,integ=0,target='AUDTHB')
```

```
x = model.fit("MLE")
x.summary()
```

The output shows a reasonably good fit over the 10+ years data range.

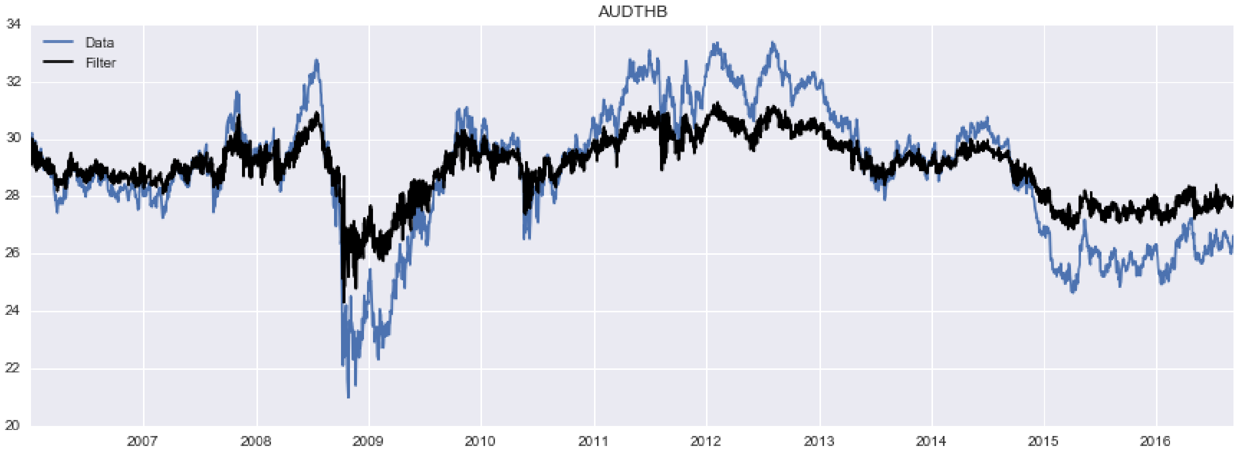


Figure 24 – ARIMA Model of AUDTHB Time Series (variance)

The fitted model was then used to make a forecast, and the output shown in Figure 25 below.

```
model.plot_predict(h=20,past_values=20,figsize=(15,5))
```

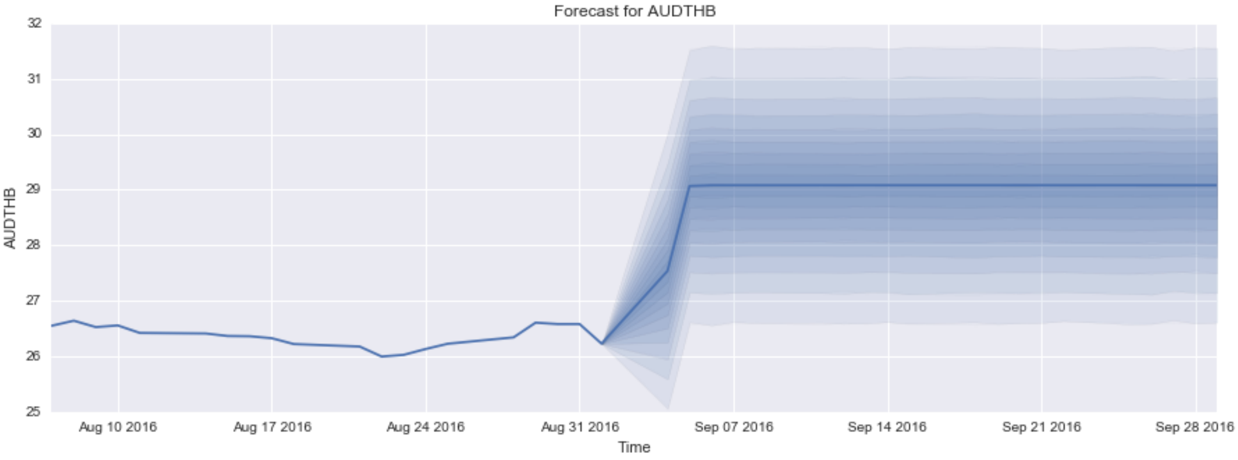


Figure 25 – AUDTHB Forecast with ARIMA Model

### 7.7.1 Observations

1. The forecast is not effective after 1-2 days, and flattens out with the same confidence intervals.
2. Very similar forecast results were obtained after many iterations of different p and q values, even when d=1 is used to difference the raw data and return better stationarity for the modelling.

### 7.7.2 Interpretation

- ARIMA model does not work well for FX time series forecasting.

## 7.8 ARIMAX MODEL

The ARIMAX model promises to be an improvement over the ARIMA model. It allows the input of up to 12 feature events that might have impacted the time series, a sort of Interrupted Time Series (ITS) modelling.

Pyflux also provided an ARIMAX function, which I then ran, using the following codes:

```
# Run ARIMA with 2 events – GFC in 2008 and start of AUDUSD dip in 2014
data2.loc[(data2.index>='2014-03-01'), 'event1'] = 1;
data2.loc[(data2.index<'2014-03-01'), 'event1'] = 0;
data2.loc[(data2.index>='2008-10-01'), 'GFC'] = 1;
data2.loc[(data2.index<'2008-06-01'), 'GFC'] = 0;
# build model
model = pf.ARIMAX(data=data2,formula='AUDTHB~1+event1+GFC',ar=1,ma=1)
x = model.fit()
x.summary()
```

The 2 events that picked are related to the impact of GFC in 2008, and the impact of dropping commodity prices in Australia in 2014. The fitted ARIMAX model is shown below.

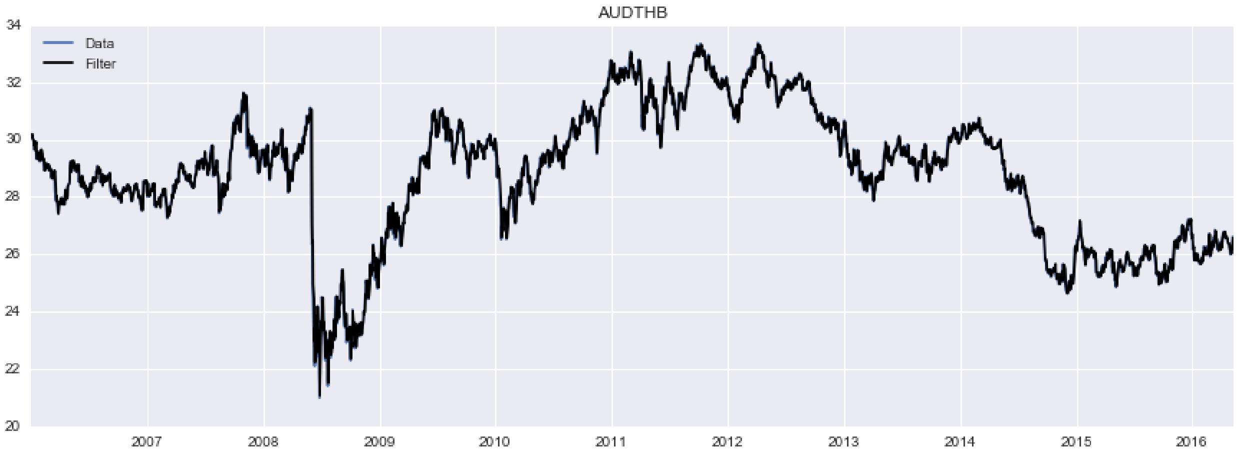


Figure 26 – ARIMAX Model of AUDTHB Time Series

The ARIMAX forecasting function was then used, and the following plot was returned.

*model.plot_predict(h=10,oos_data=data2.iloc[-12:],past_values=50,figsize=(15,5))*

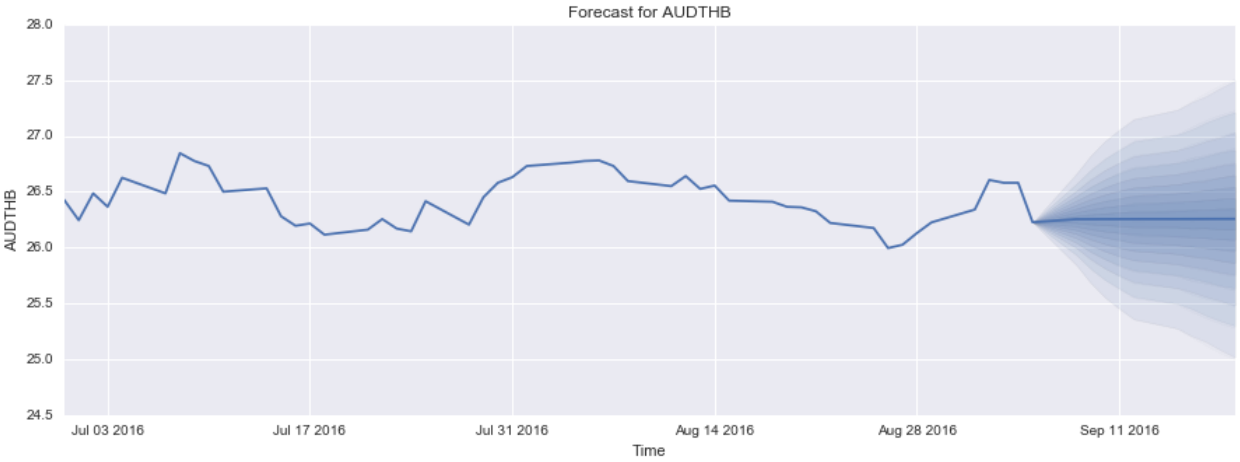


Figure 27 – AUDTHB Forecast with ARIMAX Model

### 7.8.1 Observations

1. The forecast is not effective after 1-2 days, and flattens out with increasing range in the confidence interval.
2. Very similar forecast results were obtained after many iterations of different p and q values.

### 7.8.2 Interpretation

- The ARIMAX model is also not useful for FX time series forecasting.

# 8 Summary of Results

The conclusions that can be drawn from this project are:

1. AUDTHB is strongly correlated with AUDUSD
   - In spite of both Australia and Thailand trading in the AsiaPac bloc, where JPY is the main currency traded
   - Could be due to the strong historical, economics, politics and cultural ties between Australia and USA

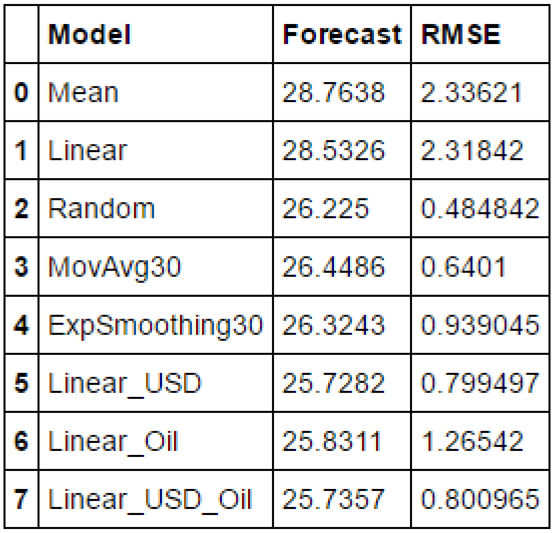
- Also means that when USD goes up, AUDTHB dips because AUD is negatively correlated with USD, eg. when US investments are diverted to Australia when financial confidence are low in the US, and vice versa. Hence, USDTHB and AUDUSD are negatively correlated.

2. JPY has secondary influence on AUDTHB
   - THBJPY is strongly correlated with USDJPY, indicating a strong influence of both JPY and USD on THB.
   - However, given the stronger ties between Australia and US, the THBJPY and AUDJPY correlation is only moderate.

3. AUDTHB strongly trend with Crude Oil Prices
   - This is due to the fact that crude oil prices around the world are tagged to USD.
   - US is a major oil importer; so when crude oil prices goes up, USD goes down.
   - When US goes down, investments are usually diverted to Australia, buffering up AUD.
   - Hence when crude oil prices goes up, this is favorable for AUDTHB

4. Gold and Stock prices no impact
   - AUDTHB is poorly correlated with gold and stock prices.
   - Debunks the myth that gold and stock prices may impact currency exchange rates.

5. Linear methods are improved by sequencing techniques
   - Sequencing techniques such as random walk, moving average/smoothing and exponential smoothing (Models 3, 4, 5) showed clear improvements in RMSE values. See comparison table below.
   - However, the lag value does not seem to have much impact, as the models fitted closely with the pattern of original data values, confirming that these models are suitable for forecasting a time series.



| | Model | Forecast | RMSE |
|---|---|---|---|
| 0 | Mean | 28.7638 | 2.33621 |
| 1 | Linear | 28.5326 | 2.31842 |
| 2 | Random | 26.225 | 0.484842 |
| 3 | MovAvg30 | 26.4486 | 0.6401 |
| 4 | ExpSmoothing30 | 26.3243 | 0.939045 |
| 5 | Linear_USD | 25.7282 | 0.799497 |
| 6 | Linear_Oil | 25.8311 | 1.26542 |
| 7 | Linear_USD_Oil | 25.7357 | 0.800965 |

6. Linear methods are improved by adding regressor features

- Adding either AUDTHB or crude oil price as additional regressor features have improved on the basic linear model of AUDTHB (Models 6,7,8)
- The stronger correlation comes from AUDTHB, as adding AUDTHB to crude oil feature alone improved the RMSE to the same value as correlation with AUDTHB alone.
- The RMSE values obtained are as good as the sequencing methods (Models 3,5,5), probably because of the strong correlation of AUDTHB and crude oil prices with AUDTHB.

7. No multiplicative improvement over additive
   - The results of models 6,7,8 showed there is no improvement in the model (R-square values) using a multiplicative model instead of additive.
   - The additive model is most appropriate if the magnitude of the seasonal fluctuations or the variation around the trend-cycle does not vary with the level of the time series.
   - When the variation in the seasonal pattern, or the variation around the trend-cycle, appears to be proportional to the level of the time series, then a multiplicative model is more appropriate.
   - Since a FX time series does not have seasonal or cyclic components, one would expect an additive model to be similar to a multiplicative model. Hence the similarities of these R-square values within each of models 6, 7 and 8.

8. There is no seasonal component of FX time series
   - The lack of seasonality in FX time series is validated in the decomposition tests conducted.
   - Stationarity tests also validated the non-stationarity characteristics of FX time series data.

9. ARIMA modeling is not suitable for FX forecasting
   - The poor results obtained from the ARIMA model indicates that FX forecasting is complex, and may require the use of more features to explain the FX pattern.

10. ARIMAX modeling is also not suitable for FX forecasting.
    - After adding 2 feature events to account for the time series interruptions in 2008 and 2014, the ARIMAX also appear to be a poor FX time series model that could be used for forecasting.
    - Apart from known macro and micro-economic factors, much of FX price fluctuations has been attributed to human psychology as well. This makes FX forecasting extremely difficult, perhaps lending it to more advanced and sophisticated machine learning techniques, such as ANN, SVM, etc.

# 9 RECOMMENDATIONS

Given that the objective of this study is to provide simple and reliable FX indicators for the layman, the following are my recommendations for the target audience of Thai community in Australia:

1. We predict AUDTHB trend by observing AUDUSD prices

   *AUDUSD is a highly reliable indicator. When USD goes up, AUDTHB price will trend down.*

2. We can reliable predict AUDTHB trend by observing Crude Oil prices

*Crude Oil price is a fairly reliable indicator. When oil prices goes up, AUDTHB price will trend up.*

3. Unlike popular belief, gold and stock prices are poor indicators of FX rates.

   *Do not waste time watch movements in gold and stock prices.*

4. The AUDTHB price is quite insulated from the AsiaPac currency bloc

   *Other THB cross currency pairs with AsiaPac countries may be strongly impacted by JPY, but AUDTHB price is more influenced by Australia-US relationship.*

# 10 FUTURE WORK

This project has given me the opportunity to embark on the learning curve of the rich field of time series analytics. It has provided me an appreciation of the added complexity of human emotional component in FX time series forecasting, that financial technical experts are still trying to get on top of. It is said that if forecasting FX market were that easy, everyone would be doing it!

So in terms of future work… there is potentially a lot!

However, in progressing baby steps, I would like to next:

1. Test clustering methods such as ANN and SVM, which appears to be also widely used for financial time series analytics.
2. Install Ubuntu on my PC so as to be able to install cool Python packages like keras (not supported on Windows) in order to do the sexy clustering analytics.
3. Explore Interrupted Time Series (ITS) methods, as abrupt time series interruption is a common feature of historical FX data.
4. Be able to install an ITS package (like PC SAS) in order to do ITS analytics.
5. Finally, be able to forecast AUDTHB with some accuracy, and become super rich!

# 11 ACKNOWLEDGEMENTS

I would like to thank my course mentor, Roy Shubhabrata, for his advice regarding this study, stepping me through the learning curves, and suggesting great Python packages.

# 12 REFERENCES

1. Pyflux

2. Forex economic indicators list

3. Stasmodel manual

4. PyBrain.org

5. Statistical forecasting: Notes on regression and time series analysis

6. De Facto Classification of Exchange Rate Regimes and Monetary Framework

7. On the determinants of the THB/USD exchange rate

8. 3 Factors That Drive The U.S. Dollar

9. Australian dollar explainer: why is it falling

# 13 Appendix – Python Codes

I have culled the best iterations of my test and learn code, tidied them up to line up with what is covered in this report, and uploaded them to this location on GitHub (Python Codes).