ANLY482-Anlytics Practicum
Interim Report
Group 06- KYY Market

KYY Market

Forex Currency Pair Predictive Modelling

Chong Zi Hao Kevin
Lim Yan Ling
Tee Yu Xuan

Table of Contents

1. Introduction

1.1 The Finance Market shift to Quant Fund

Proprietary trading has long relied on computers to help automate and execute trades. Data scientists, or more commonly known as Quants by Wall Street, have developed huge statistical models for the purpose of this automation. These models though complex, are somewhat static and as the market changes, a commonality in finance markets, they do not work as well as they do in the past.

As technology advances, we are entering an era of Artificial Intelligence and Machine Learning. Systems have capabilities to analyze large amounts of data at enormous speed and improve themselves through the process. This evolutionary computation and deep learning is seen to be able to automatically recognize changes in the market and adapt in ways the previous statistical models fail to do so.

1.2 About pH7 Global

pH7 is a private investment and consultancy firm that serves clients who are keen to appreciate their wealth and grow their capital. It has its humble beginnings in 2013 in Singapore and has been working hard to build relationship with clients to understand their business and personal concerns. With its strong information analysis capabilities and experience, pH7 provides business opportunities and solutions that are customized to clients' needs. In addition, pH7 leverages on cutting-edge technology in their work, excelling in professionalism and productivity.

By partnering with market platforms that boasts state-of-the-art technology and competitive market access, pH7 aims to capitalize on every investment and business opportunity present in the markets. It aspires to be a firm of excellence and distinction which boasts of its professionalism in dealings and partnerships with clients.

1.3 Opportunity Identification

Noting the rise of the AI and algorithms in the financial services sector, pH7 Global has poised itself to take advantage of this technological shift by expanding its services to providing trading expertise with the help with machine learning itself. Starting with the setup of a data farm for finance instruments, pH7 moved towards including a Data Manager and Data Engineer in their team.

With the infrastructure and personnel for data science in place, pH7 Global now looks towards understanding the data collected and using them for machine learning to aid their business goals.

2. Project Motivation

The team's motivation for doing this project is primarily an interest in undertaking a challenging project in an interesting area of research which has been a hot topic among the finance industry, Machine Learning in Financial Services. The opportunity to learn and put into practice a new area of machine learning not covered in our academics was appealing. Machine Learning is expected to take a huge role in trading, causing a notable shift in the trading markets. Utilizing past data, the opportunity pH7 Global has given us allows us to tap on their expertise in trading of financial instruments and the existing market data they have collected. This gives us a whole new experience of applying analytics in the financial markets.

3. Project Objectives

Utilizing the minute tick data from our sponsor, we would like to discover useful and practical insights which will allow traders to make more informed decisions in their trading. We would be coming up with a predictive modelling for currency pair.

The team and our sponsor pH7 Global have identified 2 areas of focus for this project.

1. Preliminary Data Analysis and Information Research
2. Predictive Algorithm Modeling and Strategy Testing

At the end of the project, the team aims to design a unique predictive model from the data insights discovered during the analysis.

4. Data and Tools used

4.1 Data source

Our project sponsor gave us access to the company's Amazon Web Services RDS, where they had collected historical data on the prices of multiple currency pairs. The dataset includes multiple timeframes of the same period of time series data for a 2 years' time period; 1st July 2015 to 30th June 2017.

The data fields include:

- Timestamp (timestamp of the data)
- High (High point of the currency pair for the minute)
- Low (Low point of the currency pair for the minute)
- Open (open price of the currency pair for the minute)
- Close (closing price of the currency pair for the minute)

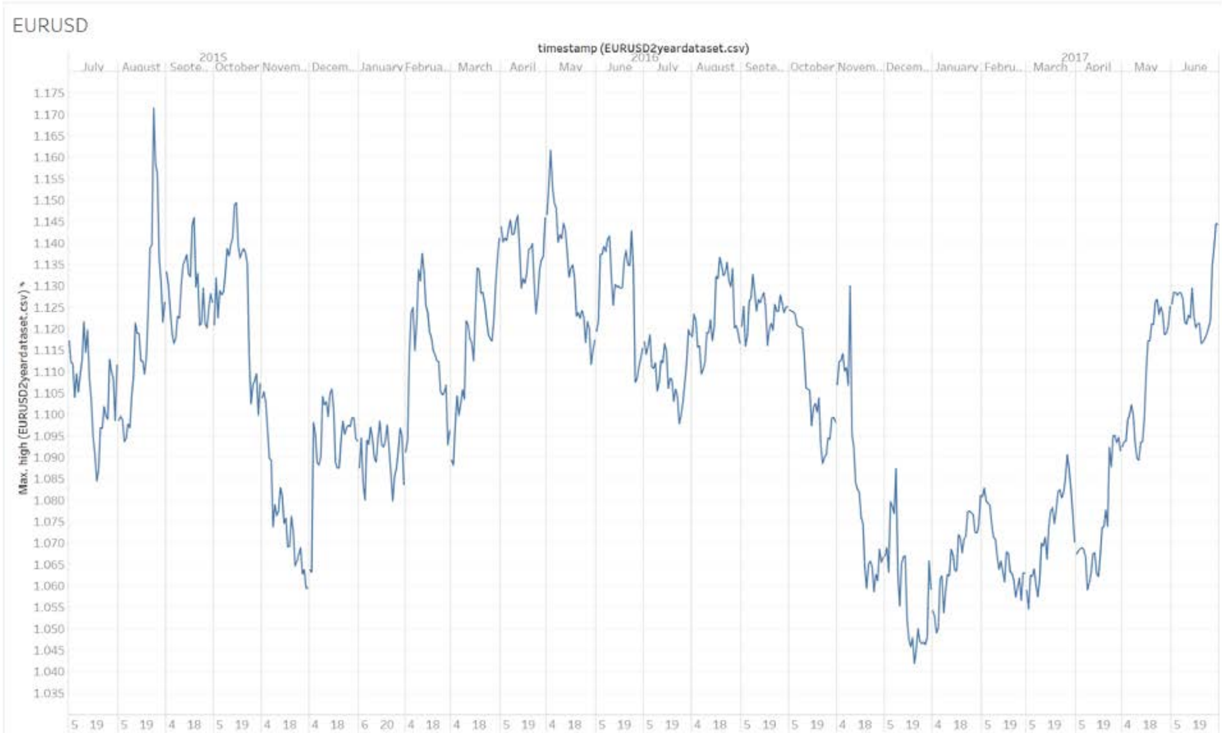
Timestamp	Open	High	Low	Close
2017-10-25 00:01:00	1.31303	1.31308	1.31303	1.31308
2017-10-25 00:02:00	1.31308	1.31313	1.31306	1.31307

To access our client's database, we used Rstudio codes to directly access the AWS servers and retrieve the data as needed for our analysis. This gave us the flexibility of choosing time periods we want to work with for our analysis.

The resulting data retrieval for 2 years worth of minute tick data for 1 currency pairs comes close to 750,000 rows.

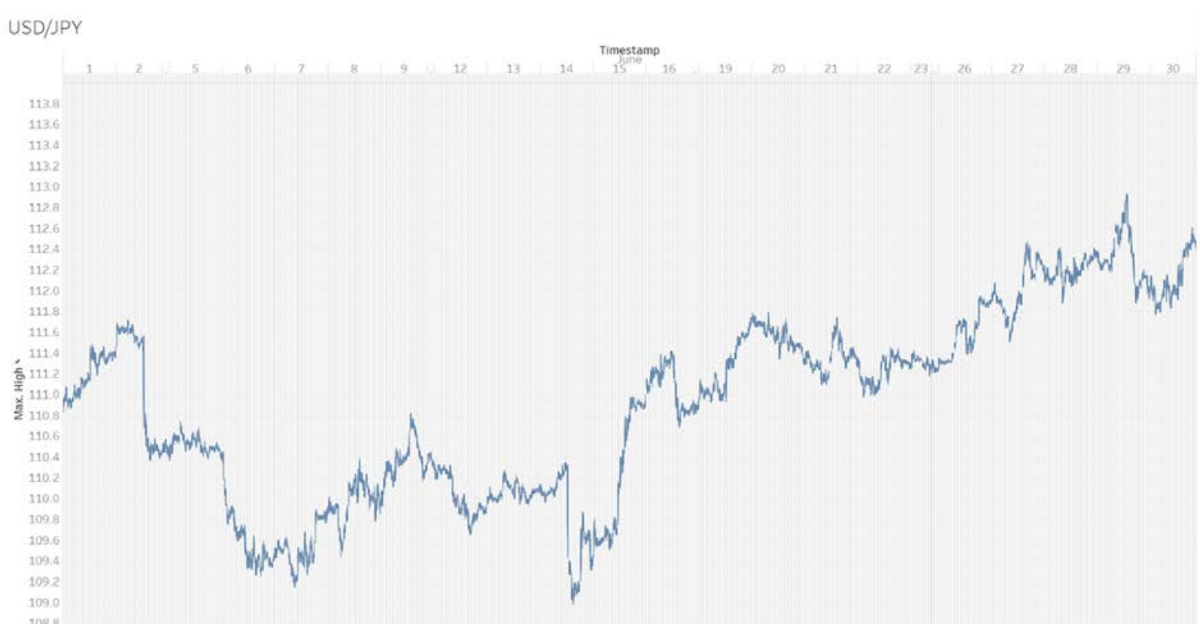
1	rownum	timestamp	open	high	low	close
749855	749854	30/6/2017 20:41	112.481	112.485	112.478	112.478
749856	749855	30/6/2017 20:42	112.478	112.48	112.472	112.472
749857	749856	30/6/2017 20:43	112.472	112.472	112.448	112.448
749858	749857	30/6/2017 20:44	112.449	112.454	112.448	112.449
749859	749858	30/6/2017 20:45	112.441	112.445	112.439	112.443

Below are initial visualizations of two sets of the data by Tableau, with the use of the original data without any transformation:



Initial observations of the data revealed incomplete dataset as the market is closed during weekends. Future analysis of the data will take this information into consideration.

Attempting to visualize minute tick data is restricted to a maximum of 1-month time periods due to the volume of data. The result of this visualization is as shown below:



4.2 Data Cleaning and Preparation

For our data cleaning and preparation, we used the following software to both visualize and ETL the data into other forms:

1. JNP Pro
2. Tableau
3. SQL Server Data Tools 2015 (MSSQL)
4. Microsoft Excel

Through the visualization seen earlier in the report, we realized that there is a need to perform data transformation to visualize all the data. Therefore, the data was prepared with MSSQL instead to produce a ‘day aggregated’ data-set for analysis on the day-time period basis.

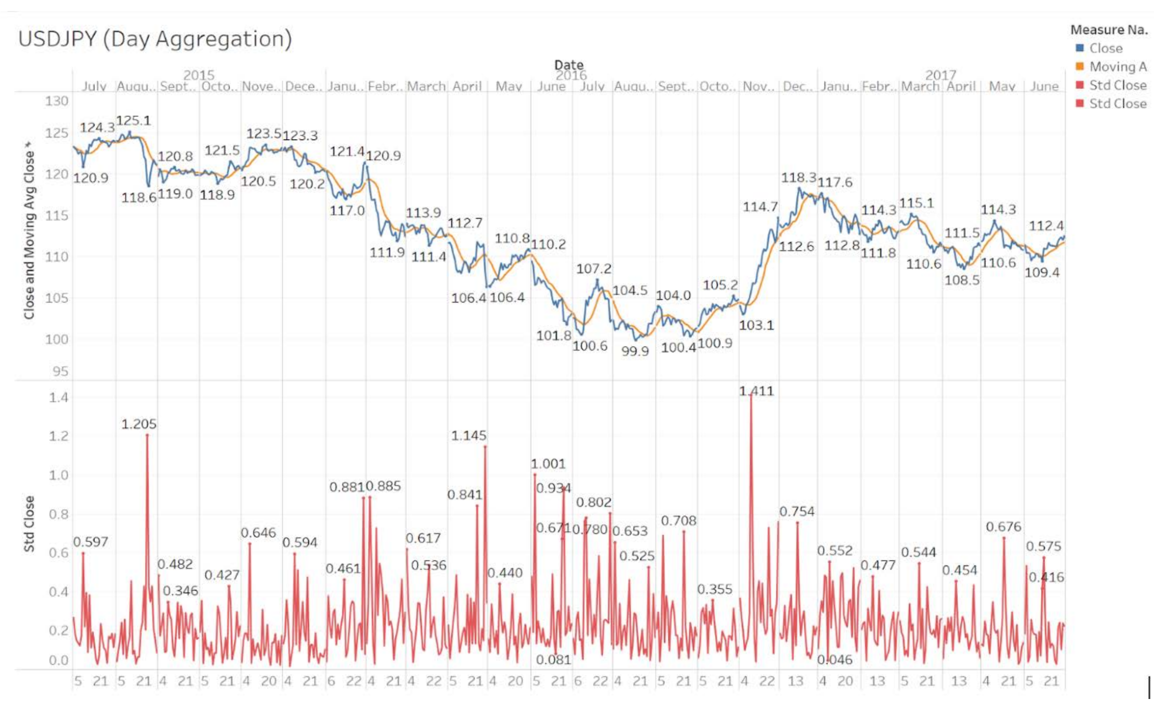
Our initial methodology was to use the clustering method to identify clusters which could be treated as baskets for investment. As the currency values of USDJPY and the rest are vastly different, there was a need to transform into percentage change and standard deviation for clustering.

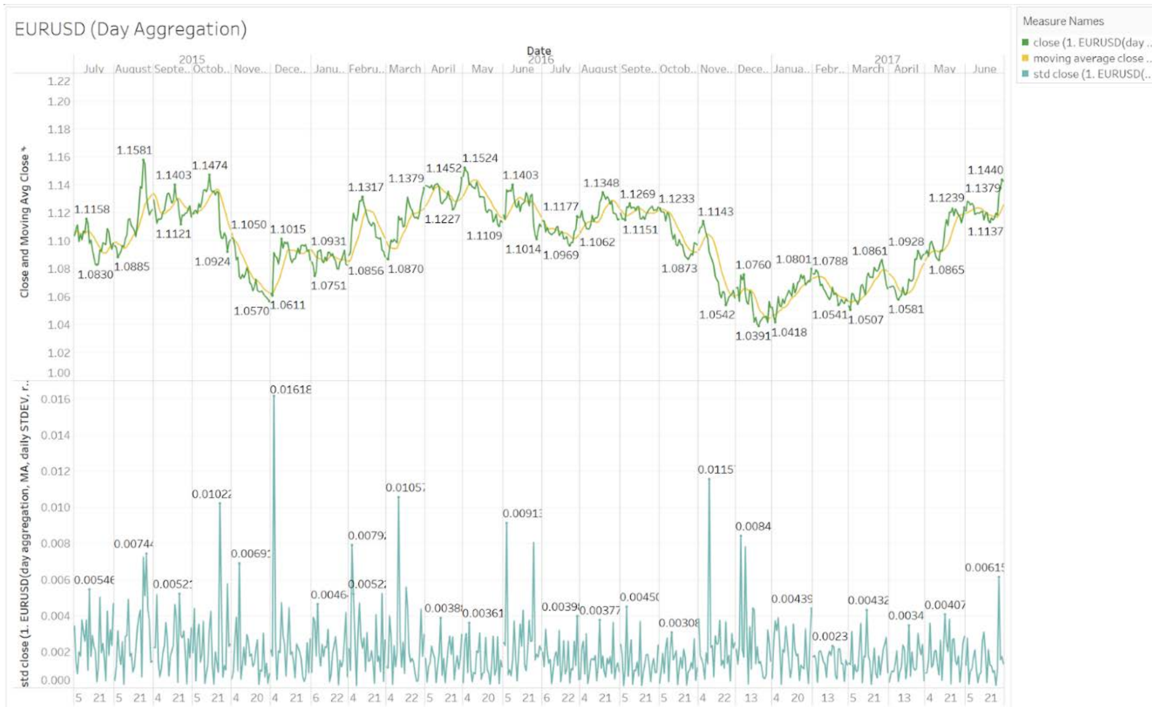
However, our client does not have 15-20 or more currency pairs in their database. Hence, we would be focusing on forecasting with these 5 currency pairs. Our team used the ARIMA forecasting method and thus the data transformation method would not be required as the ARIMA model uses its own unique method to transform data.

The image below shows the result of our first data transformation.

1	rownum	high	low	open	close	year	month	day	std close	moving average close	Date
621	620	111.432	111.165	111.3	111.244	2017	6	23	0.046088	110.9308	23/6/2017
622	621	111.334	111.169	111.227	111.241	2017	6	25	0.026631	111.048	25/6/2017
623	622	111.946	111.214	111.239	111.872	2017	6	26	0.215207	111.2903	26/6/2017
624	623	112.467	111.466	111.873	112.2	2017	6	27	0.241656	111.4105	27/6/2017
625	624	112.417	111.832	112.201	112.357	2017	6	28	0.101104	111.5563	28/6/2017
626	625	112.927	111.816	112.353	111.996	2017	6	29	0.23855	111.6582	29/6/2017
627	626	112.607	111.736	111.999	112.399	2017	6	30	0.221595	111.7282	30/6/2017

We performed data transformation to allows us to visualize the data differently and derive new insights on the data:





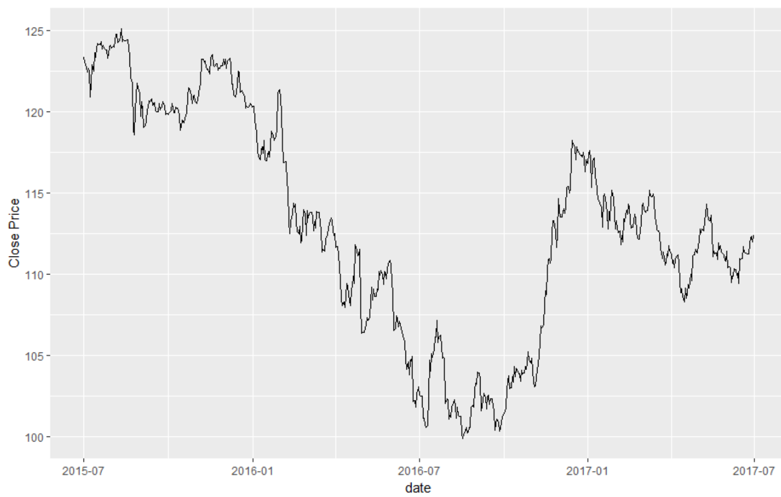
As seen in the visualization of the data of the same currency pair and time period, we can see the trends and price movements for the entire time period of 2 years for USDJPY data.

This provided additional data discoveries which we observe significant shifts in the price movements and their variations throughout the time period. This allows us to visually compare across multiple currency pairs to spot any prominent similarities and trends between them.
Although nothing of significance was identified through the visualization charts as shown below, we could identify periods of time which could increase the granularity of the data points to allow deeper analysis for our forecasting.

5. Exploratory Data Analysis

5.1 Fundamentals

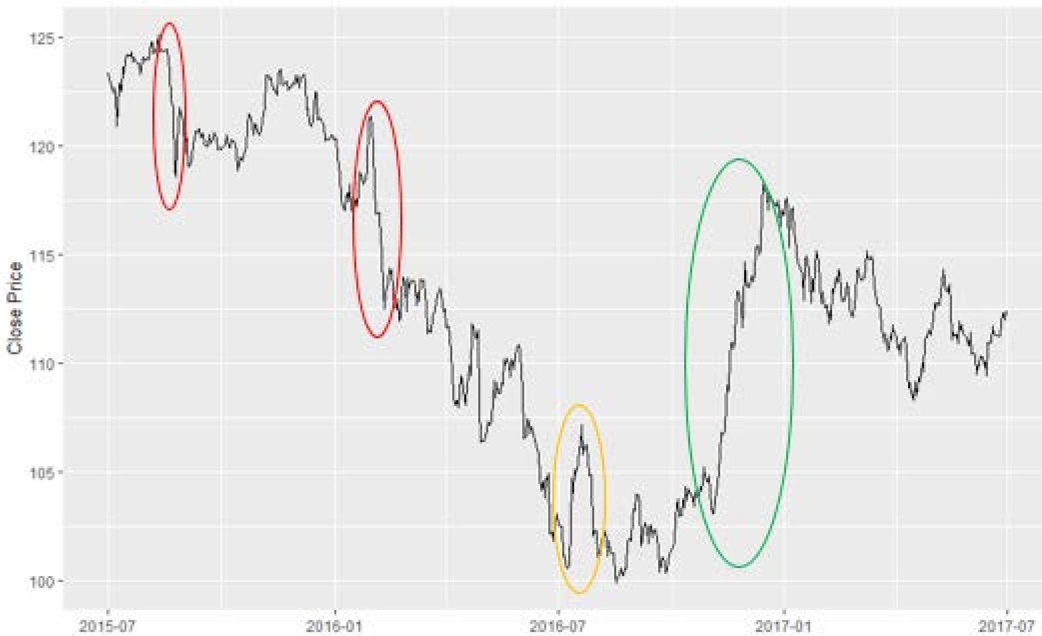
The Japanese yen is one of the "safe haven" currency, together with the swiss franc, it is a go to for investors whenever there is major uncertainty in the market. The Japanese yen has a 13.6% weight of the US Dollar index, 2nd behind the EUR (ICE Futures U.S., N.D.). It is also the world's 3rd biggest economy behind the united states and china, with 5.9% of the world's GDP.



(Chart 1: USDJPY closing chart July 2015 to June 2017)

In the initial data exploration, it is observed that the peak of the USDJPY currency pair has an overall 2-year high point of around 125 and a 2 year low of around 100. With a range of 25 approximately (25%), it implies that it is a very volatile pair. From the graph, the currency pair has very fast movements, with huge price changes in a short-time periods generally.

In the graph below, we have identified 4 very high movement periods. In August 2015 and February 2016, there were huge drops in the USD currency value relative to the JPY. In July 2016 there was a huge rise followed by a drop in the USDJPY value. Lastly, in November till December, there was a huge rise in the USDJPY value which is attributed to trump's election campaign win.



(Chart 2: USDJPY closing chart July 2015 to June 2017 with highlighted periods)

Legend: Monthly Moving Average, Quarterly Moving Average
(Chart 3: USDJPY closing chart with monthly and quarterly moving averages from tradingview.com)

The chart above consists of 2 other lines of moving averages which include the monthly moving average (20 days) and the quarterly moving average of 70 days.

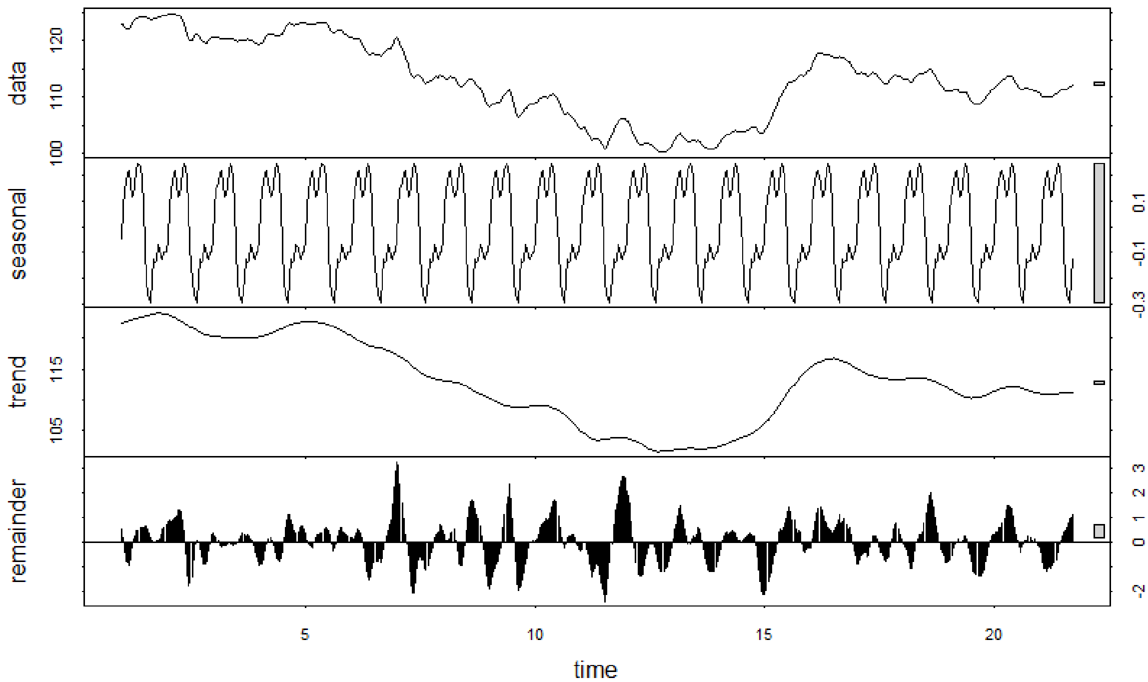
The quarterly moving average 2-year trend is as follows:

Trend	Time Period
Up	July – August 2015
Down	September 2015 – October 2016
Up	November 2016 – February 2017
Down	March 2017 – July 2016

Monthly moving average 2-year trend:

Trend	Time Period
Down	August - September 2015
Up	November 2015 – December 2015
Down	December 2015 – July 2016
Up	September 2016 – December 2016
Down	January 2017 – April 2017

5.2 R Library Forecast: Graph of Plots



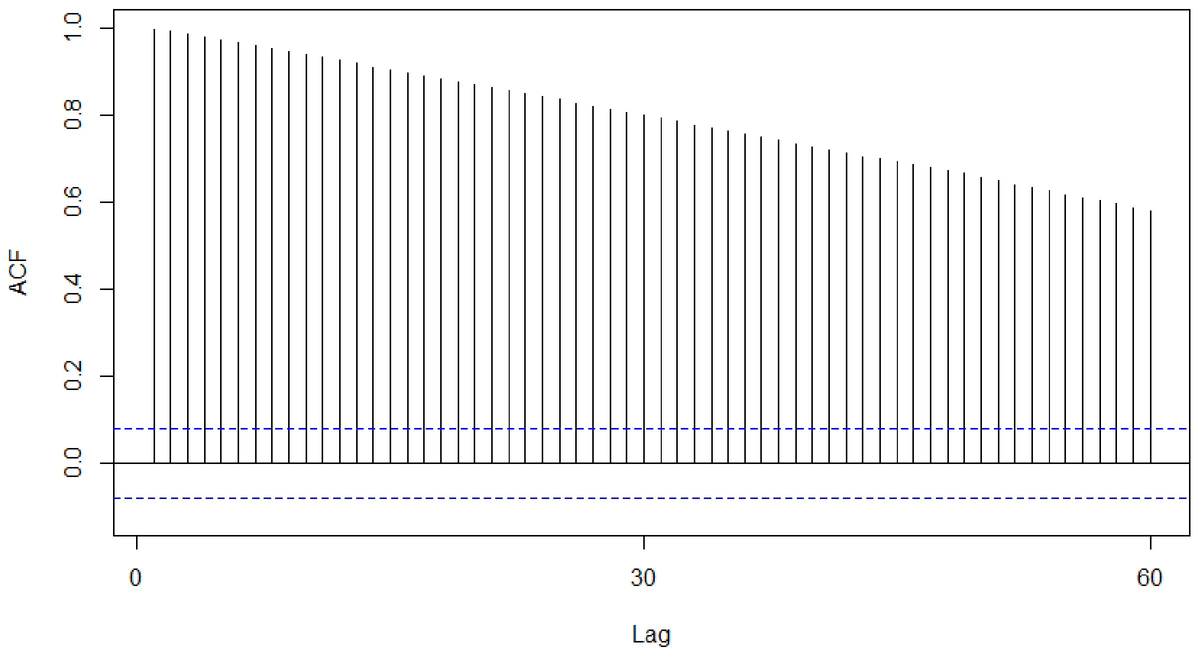
This graph of plots is generated by the decomposition function of the R library forecast with the STL function. The decomposition function does the following steps, by decomposing the existing plot into 3 different types of data, the seasonal component, the trend component and the cycle component.

From the seasonal component, it is evident that there are 21 cycles of seasonal components, in these2 years, which equates to approximately one season a month.

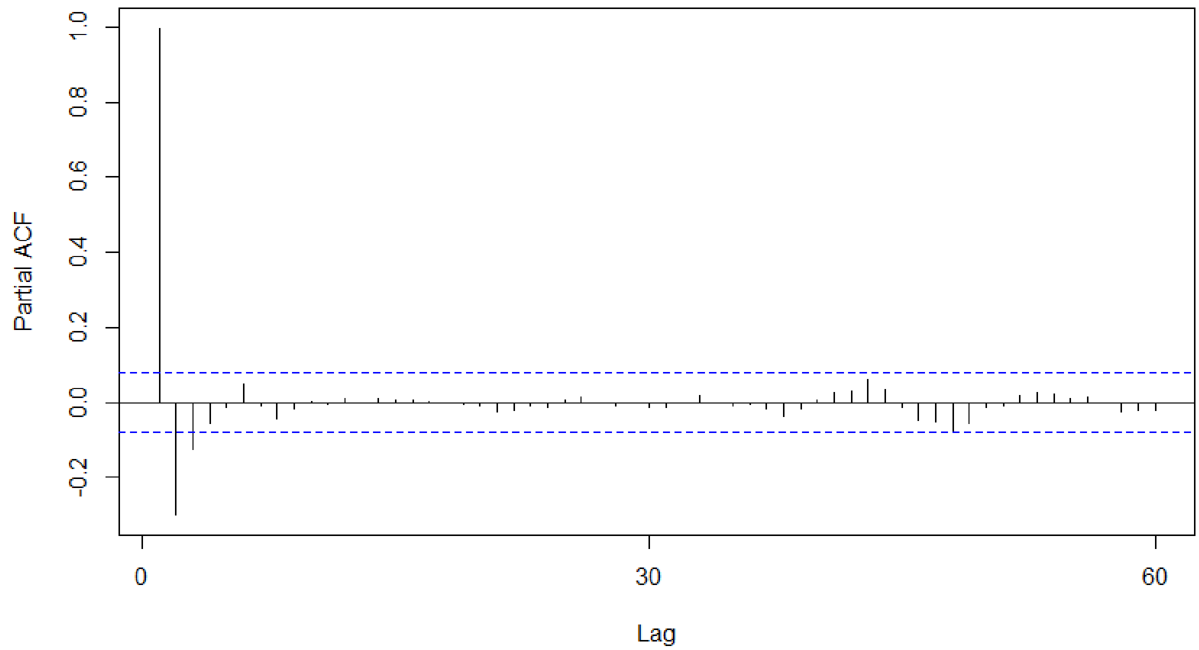
From the trend aspect, there is no certain indicator of increment or decrement in this time period of 2 years. Overall, it reduced slightly in the 2 years after a decrement followed by an increment. A more detailed breakdown is as shown in the graph above.

From the remainder, it can be seen that the remainder is relatively near to the average, with only a few points of outliers and the outliers are gradual spikes and drops.

5.3 Autocorrelation Function



The autocorrelation function (ACF) does not give a good gauge of the lag as all the lag values are all above. Hence, there is a need to use the partial ACF.



The partial ACF shows a lag order of 1 and 2 to be considered. However, with the high amount of spike in order 1, it is evident that there is a need to differentiate the data by the first order.

5.4 Augmented Dickey-Fuller Test using R

The following table is our R output:

Data	Dickey-Fuller	Lag order	p-value	Alternative Hypothesis:
count_d1	-7.5253	8	0.01	stationary

Using the augmented Dickey-Fuller Test on the differenced first order data, the non-stationarity is rejected. Hence the hypothesis is stationary as a requirement for using the ARIMA model.

Using the auto.arima function, the value of ARIMA(3,1,2) is reflected. This implies that P, D and Q are 3,1 and 2 respectively. This model has an AR(3), I(1) and MA(2).

6. Moving Forward

Moving forward, we will commence on to the next phase of our project which is to train and test the model used as explained in the methodology. Currently our model results in only a good predictor on the daily aggregated data with EMA 90 (quarterly) predictor.

We would aim to get an optimal forecast for minute tick data by using the train, test and validate approach. Additionally, Monte Carlo Package of R could also be used to do testing to another level using the vignette package.

Upon further research, should our current ARIMA model not be useful in achieving our objective of a good forecast for minute tick model, we would be considering exploring the 2 following methods namely:

1. Kalman Filter Model (Continuous Dynamic Bayesian Network)
2. Bayesian Network Model (general)

These 2 methods will cover the weaknesses of the current ARIMA model of the assumption of constant variance which usually does not apply in financial time series data and usually have more complex structures than ARIMA. (Cristina et. al. ,2016) Also according to (Zhang ,2010) it is difficult to predict the future as it is based on only a finite window, known as extrapolation.

7. References

B. (n.d.). Dynamic Bayesian Networks. Retrieved November 3, 2010, from https://bi.snu.ac.kr/Courses/g-ai10f/Ch9_DBN.pdf.

Gray, A. (2017, March 9). The world's 10 biggest economies in 2017. Retrieved from https://www.weforum.org/agenda/2017/03/worlds-biggest-economies-in-2017/

PETRICĂ, A., STANCU, S., & TINDECHE, A. (2016). Limitation of ARIMA models in financial and monetary economics. Retrieved from http://store.ectap.ro/articole/1222.pdf

US, I. F. (2011). U.S. Dollar Index. Retrieved from https://www.theice.com/publicdocs/ICE_USDX_Brochure.pdf.