1. Exponential Smoothing

First, Excel took the data of four variables (K54D, EAFV, K226 and JQ2J) from January 2000 to December 2019, next, reduced the value of JQ2J by 100 times so that it is on the same scale as the other three variables.

The Time Plot, ACF and PACF curve are shown in ***Figure 1-1, Figure 1-2, Figure 1-3*** respectively. K226 shows a Multiplicative trend for all 20 years; however, for the latest 7 years, it shows an additive trend. Hold on to the regulation that the later the time is, the more significant it is, I hold the view that it shows an additive trend. Through these three figures using Pegels(1969)'s method, the following feature of data can be inferred, which is the basis for exponential smoothing model selection.

Variables	Trend-Cycle	Seasonal	Numerical-Trend	Suitable Model
K54D	Additive	Additive	Increase Significantly	Holt-Winters' additive method
EAFV	None	Additive	Stable	Holt-Winters' additive method
K226	Additive	None	Decrease	Holt's linear trend method
JQ2J(*.01)	Additive	Multiplicative	Increase	Holt-Winters' multiplicative method

Because K54D, EAFV and K226 all have additive or Multiplicative seasonality, Holt-Winters' method is suitable for them; meanwhile, Holt's linear trend method is suitable for K226, not only because K226 has no seasonal trend, also the prediction result of Holt's linear trend method is a straight line with non-zero slope, which is also consistent with the terminal character of K226 curve. K226 curve looks like a ReLU function that folds along the mirror, according to the characteristic of ReLU, I consider K226 will also show an approximate linear trend in 2020.

The results of exponential smoothing are shown in ***Figure 1-4***. I have predicted the change of these four variables in 2020. The predicted numerical trend is basically the same as the original data trend, and their MSE is 52.36 at the highest and 14.49 at the lowest. Therefore, the fitting is effectiveness.

For further testing the effectiveness of exponential smoothing models, I separate these four variables (K54D, EAFV, K226 and JQ2J) to training dataset(2000 JAN—2018 DEC) and testing dataset (2019 JAN—2019 DEC). By applying the data from the training set into the established four models, I compare the predicted results in 2019 with the testing dataset and calculate the MSE. Through ***Figure 1-5***, the prediction results are basically consistent with the testing dataset; therefore, the model fitting is effective. In addition, the MSE of these four variables (Testing MSE for K226 = 32.84 Testing MSE for K54D = 13.75, Testing MSE for JQ2J(*.01) = 37.84, Testing MSE for EAFV

= 27.49) is smaller than the overall MSE and within the acceptable range, which further enhances the confidence of the model validity.

The forecasting value of these four variables and overall MSE are shown in the following table.

t	K54D	EAFV	K226	JQ2J
2020 JAN	525.62	102.1462	107.96	16721.46
2020 FEB	541.19	100.3935	108.03	14596.97
2020 MAR	598.48	99.90794	108.10	26959.71
2020 APR	521.83	105.6927	108.16	17515.42
2020 MAY	521.71	103.7251	108.23	18475.75
2020 JUN	525.92	101.0906	108.30	19390.44
2020 JUL	523.09	102.2277	108.37	17575.24
2020 AUG	515.78	100.5272	108.44	15195.3
2020 SEP	517.20	105.3662	108.51	23826.6
2020 OCT	518.05	111.2199	108.58	18433.15
2020 NOV	514.33	120.3574	108.64	17513.01
2020 DEC	533.86	113.1208	108.71	14753.16
MSE	52.36	14.49	45.26	51.76

In the field of machine learning, many workers directly use training dataset to predict results. However, due to the particularity of time series (mentioned above), the closer time is, the more significant it is. Therefore, it is necessary to use the overall data to predict the value in 2020.

2. ARIMA

Before applying the data to ARIMA model, I create a for loop to find the optimal p,d,q by decreasing AIC. The output of my code (***Table 2-1***) shows that SARIMAX (1, 1, 1) x (0, 1, 1, 12) produces a minimum AIC value of 1551.61. For this reason, I think it's the best choice of all the models I've considered.

Coef column informs the significance of each characteristic weight. Here, the P-value for each weight is less than 0.05, so it's reasonable to keep all the weights in our model.

In ***Figure 2-1*****:**

- In the top right figure, the red KDE line and the n (0,1) line (where n (0,1)) are the standard symbols of the normal distribution, with an average value of 0 and a standard deviation of 1. This is a good indication of the normal distribution of residues.

- the QQ graph in the lower left corner shows that the ordered distribution of the residual (blue dot) similarly follows the linear trend of sampling with the standard normal distribution of n (0,1). Again, this is a strong indication of the normal distribution of residues.

- the residuals over time (top left) do not show any significant seasonality and appear to be white noise. This is confirmed by autocorrelation (i.e. correlation graph) in the lower right corner, which indicates that the time series residual has a low correlation with its own lag value. These observations lead to the conclusion that the model selection is satisfactory, which can help to understand time series data and predict future value.

Using One-step ahead Forecast from 2005, the code took dynamic=False to Ensure that the forecast for each point will use all previous historical observations. the curve shown in the ***Figure 2-2*** shows an effective fit, and the MSE is 57.67. The predicted value from the dynamic prediction results in an MSE of 477.60. This is higher than One-step ahead, which is expected, because the model relies on less historical data in the time series (***Figure 2-3***).

Both of One-step ahead and Dynamic Forecasting confirm that ARIMA time series model is satisfactory. The forecast values for the next 12 months are 527.09 543.34, 600.54, 523.93, 523.69, 528.09, 525.53, 518.54, 520.27, 521.54, 518.09, 537.30.(as shown in ***Figure 2-4***)

ARMA model considers the trend, seasonality and random volatility of time series in the process of modeling. By analyzing the sequence value of the same time point and the change trend of the sequence value in each season, the seasonal and non-seasonal components of the time series can be extracted respectively to predict the future value. However, the prediction steps of ARMA model are complex, which is suitable for the prediction of the short-term trend of events. For example, if the one-step ahead starts from 2019 Jan, the MSE will decrease to 8.7.

Exponential smoothing uses the weighted average of the historical value to predict the future value of the series. This model considers that the short-term data has a great impact on the prediction, while the long-term data has a smaller impact on the prediction. Therefore, exponential smoothing is not suitable for the prediction of time series with large fluctuations over time in theory.

3. Multi Linear Regression

The four economic indicators (K54D, EAFV, K226, JQ2J) can be used to forecast FTSE. According to the results of multiple linear regression, the value of FTSE will drop sharply from 7000 to about 4000 in 2020.

First, four independent variables and one dependent variable (FTSE) are analyzed by descriptive statistic (***Table 3-1***), and the dispersion degree of data is further observed by boxplot (***Figure 3-1***). For the observability of boxplot, two variables (IQ2J and FTSE) are reduced by 100 times. It can be found from the boxplot that the data distribution of EAFV and JQ2J is relatively scattered, and there may be influential points.

Through corr matrix (***Table 3-2***), the person correlation coefficient of independent variables (K54D, EAFV, K226, JQ2J) to FTSE are 0.63, -0.00, -0.56 and 0.53 respectively. Therefore, there are nearly no linear relationship between EAFV and FTSE, then drop EAFV.

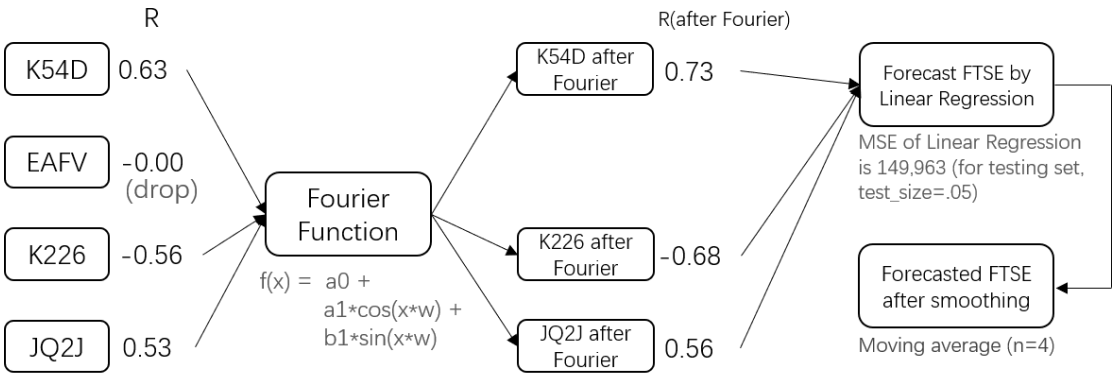
Figure 3-2 shows the linear regression lines of different factors to FTSE (confidence = 95%) which confirms the existence of influential points. As the research object of this report is time series, it is not a good way to delete influential points directly. In this report, Fourier function [f(x) = a0 + a1*cos(x*w) + b1*sin(x*w)] is used to eliminate the existence of influential points. The essence of this method is to weaken the outlier trend of extreme values and pull these influential points back to the collective (cftool of Matlab is used in this step).

After that, train_test_split separates the dataset to train set and test set by training_size=0.95, because there are 12 values to be predicted, using 12 values to test the model not only fits the number to be predicted, but also makes the model get effective training.

Although MSE of Linear Regression is 149,963 (for testing set, test_size=.05) which is a huge number, probably because the value of EAFV is huge. The fitting curve shows an outstanding fit (***Figure 3-3***).

The prediction results as shown in the ***Figure 3-4*** are obtained by substituting the predicted values of four dependent variables into the model. In order to make these results fit the original data better, this report smooth them using Moving average (n=4). The predicted curve after smoothing are shown in ***Figure 3-5.*** Predicted values are shown in ***Table 3-3***.

The overall steps of linear regression can be summarized as the following flow chart.



APPENDIX A: Description of Codes

1. Exponential Smoothing

Obtaining sample features by Time-plot, PACF and ACF then the appropriate exponential smoothing method is selected for these four variables. Evaluate the effectiveness of the model through the overall MSE and MSE for testing dataset, after that, use overall data set for prediction.

2. ARIMA

By minimizing AIC, the code create a for loop traversal to find the optimal combination of parameters, and then confirm that the model fitting effect is satisfactory through ARIMA Output, one-step ahead and dynamic forecasting. Next make prediction and visualization of prediction results.

3. Multi Linear Regression

Descriptive statistics points the existence of influential points; therefore, Fourier function is used to converge data moreover improve linear regression coefficient. After evaluating the model using test set, applying the results of exponential smoothing to this model, finally, smoothing the prediction results of linear regression model [Moving average (n=4)].

APPENDIX B: Analysis and Forecast Graphs

SECTION 1

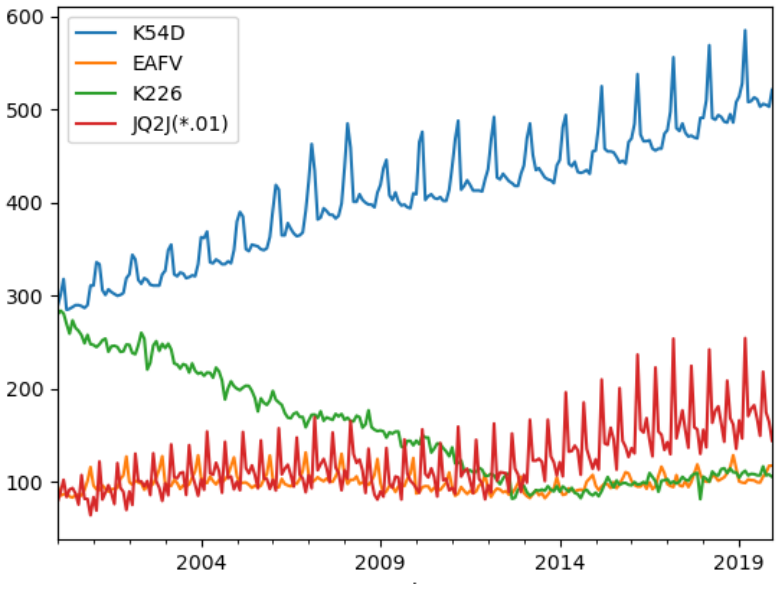


Figure 1-1 *Time Plot*

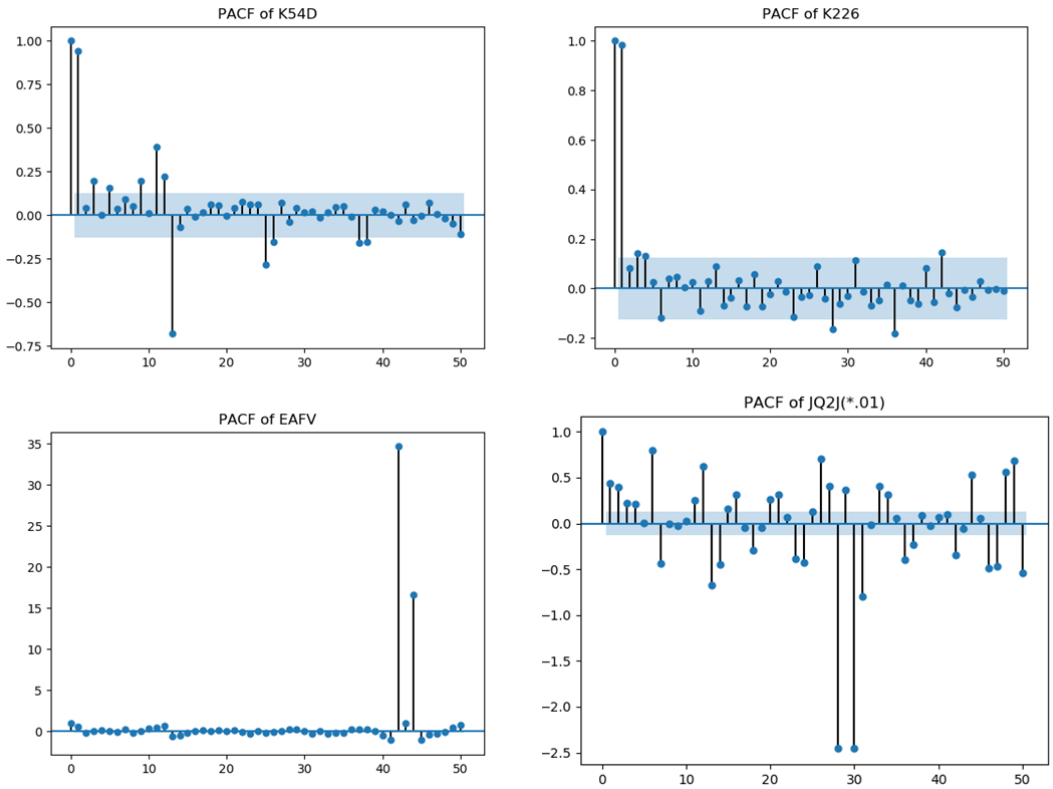


Figure 1-2 *PACF Plot*

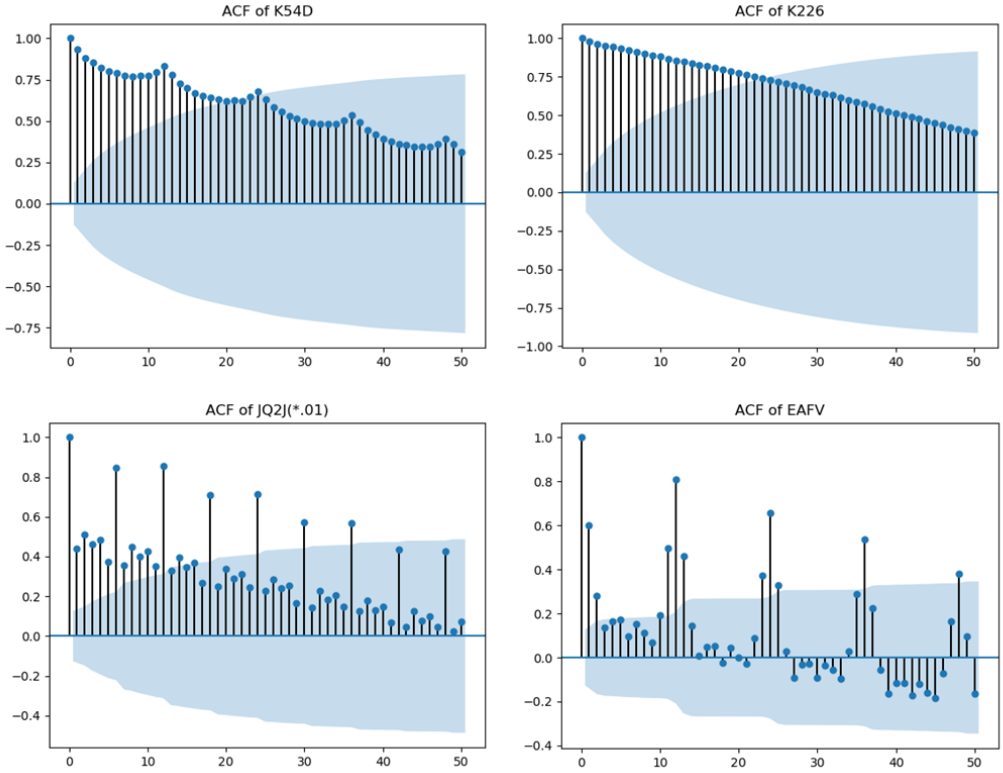


Figure 1-3 *ACF Plot*

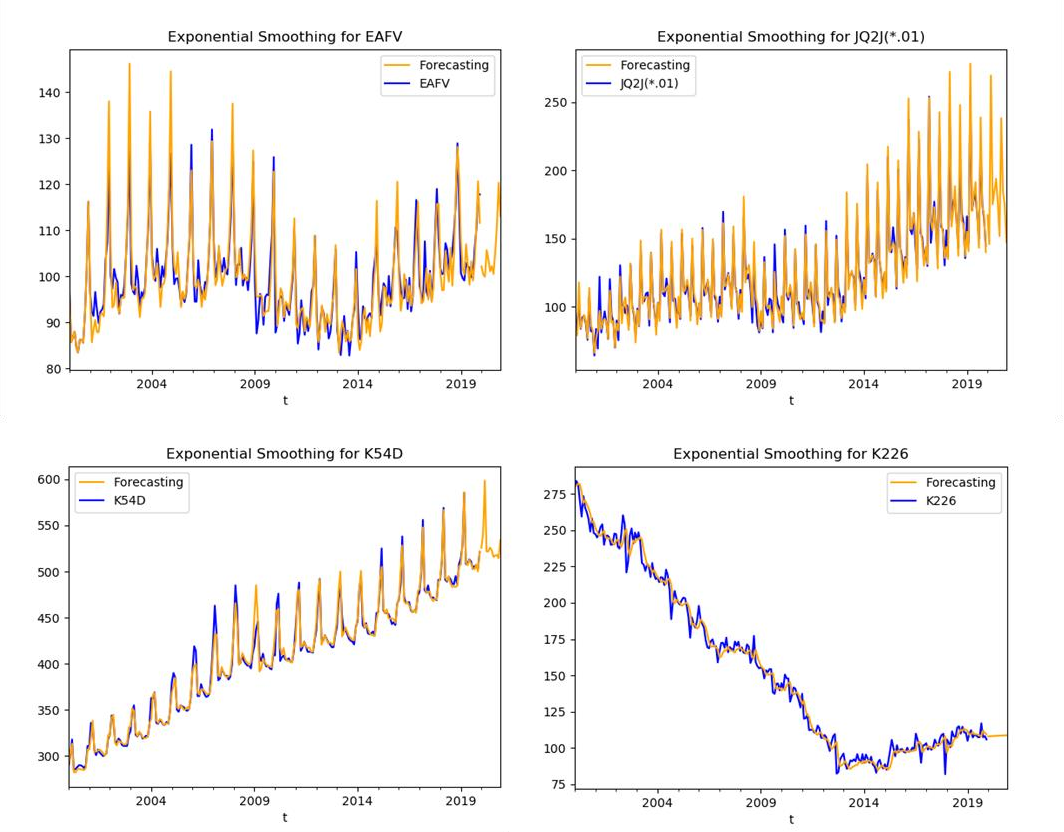


Figure 1-4 *Exponential Smoothing Plot*

Figure 1-5 *Testing Result*

SECTION 2

Table 2-1 *SARIMAX Results*

SARIMAX Results

Dep. Variable:	K54D	No. Observations:	240
Model:	SARIMAX(1, 1, 1)x(0, 1, 1, 12)	Log Likelihood	-771.807
Date:	Fri, 13 Mar 2020	AIC	1551.614
Time:	00:12:21	BIC	1565.314
Sample:	01-01-2000 - 12-01-2019	HQIC	1557.142
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.2464	0.074	3.350	0.001	0.102	0.390
ma.L1	-0.8166	0.051	-16.094	0.000	-0.916	-0.717
ma.S.L12	-0.3395	0.032	-10.621	0.000	-0.402	-0.277
sigma2	52.0550	2.736	19.027	0.000	46.693	57.417

Ljung-Box (Q):	30.91	Jarque-Bera (JB):	674.91
Prob(Q):	0.85	Prob(JB):	0.00
Heteroskedasticity (H):	0.60	Skew:	-0.20
Prob(H) (two-sided):	0.03	Kurtosis:	11.44

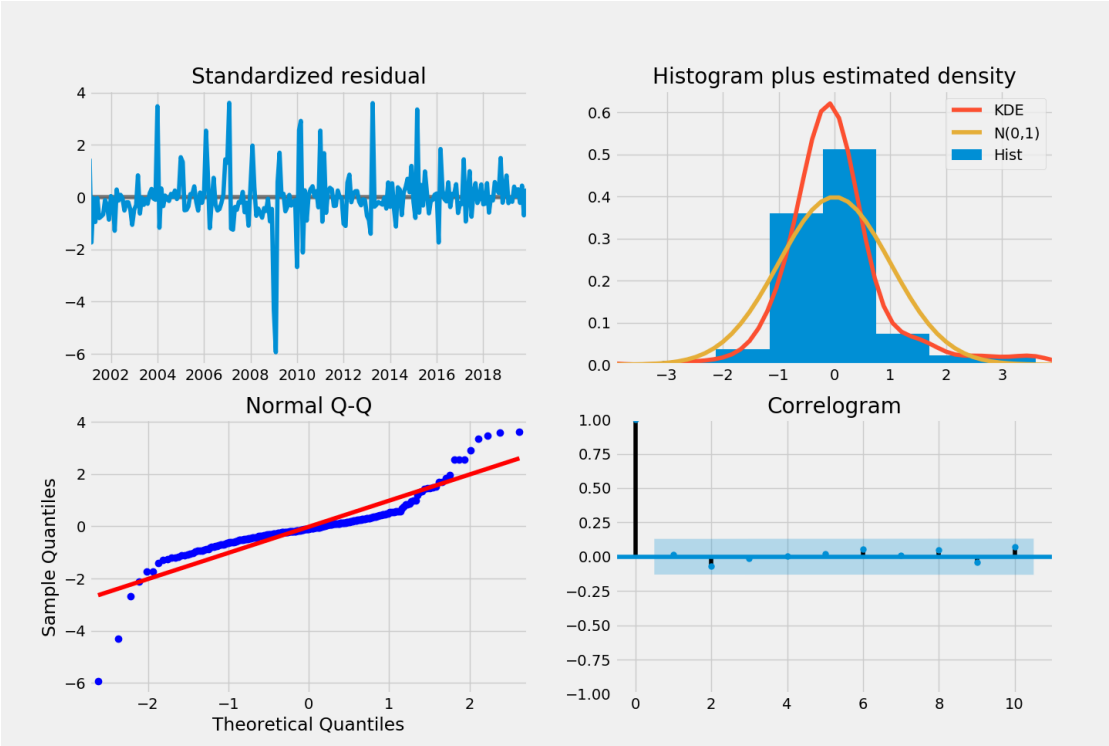


Figure 2-1 *SARIMAX Plot*

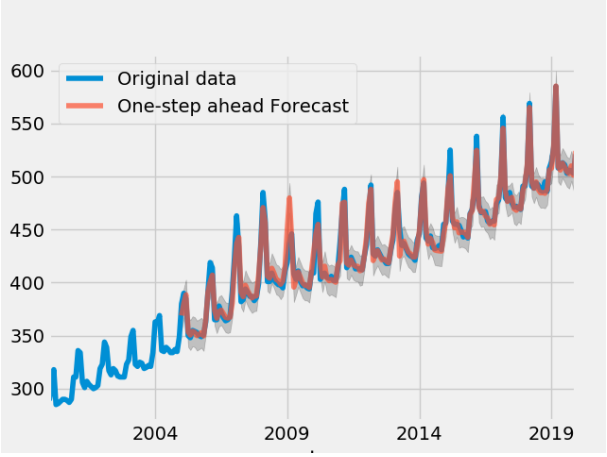


Figure 2-2 *One-step ahead Forecast*

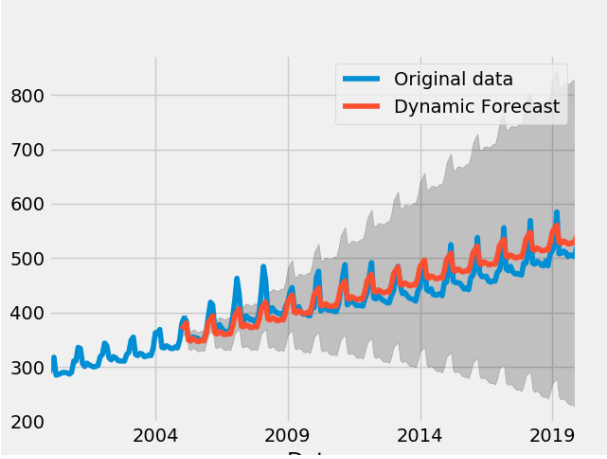


Figure 2-3 *Dynamic Forecast*

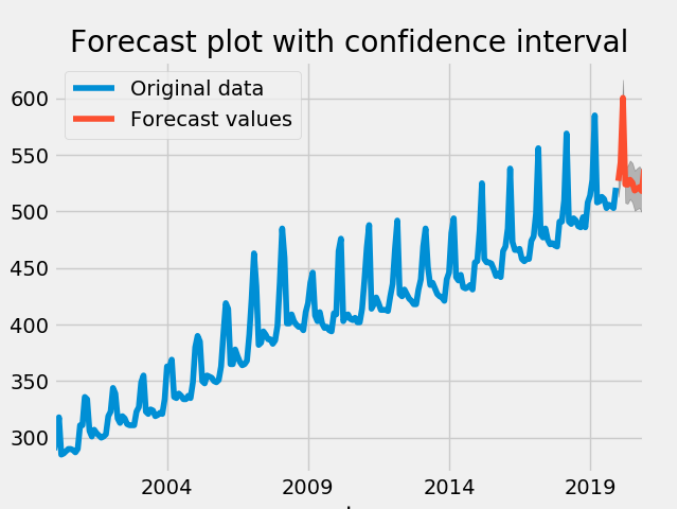


Figure 2-4 *Forecasted Value Plot*

SECTION 3

Table 3-1 *Descriptive Statistic*

head:	K54D	EAFV	K226	JQ2J	FTSE
t					
2000-01-01	289	96.5	280.0	7873.2	6930.200195
2000-02-01	301	85.7	284.0	8950.8	6268.500000
2000-03-01	318	86.8	281.1	10250.7	6232.600098
2000-04-01	285	88.0	269.8	8400.3	6540.200195
2000-05-01	286	84.5	259.4	9216.6	6327.399902

Shape: (240, 5)

	K54D	EAFV	K226	JQ2J	FTSE
count	240.000000	240.000000	240.000000	240.000000	240.000000
mean	408.395833	99.374167	154.974167	12283.336250	5881.445419
std	66.674925	9.572061	57.972269	3524.293448	1014.509944
min	285.000000	82.800000	81.800000	6442.100000	3567.399902
25%	350.750000	92.875000	103.575000	9807.725000	5188.474976
50%	413.000000	98.050000	144.250000	11216.350000	5962.699951
75%	458.250000	103.500000	201.950000	14362.625000	6629.675171
max	585.000000	131.900000	284.000000	25459.500000	7748.799805

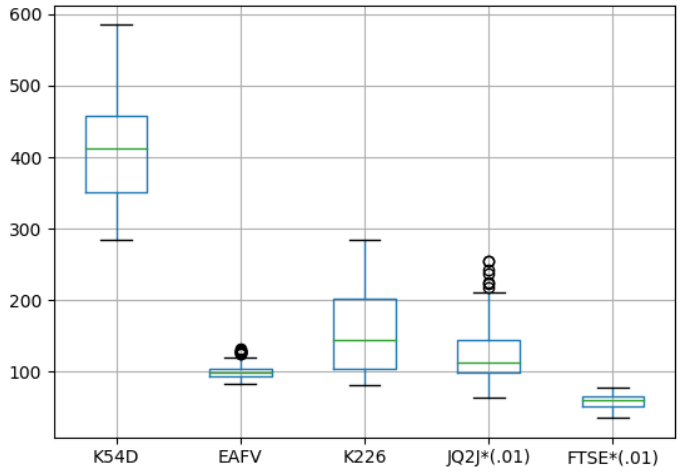


Figure 3-1 *Boxplot*

Table 3-2 *Corr Matrix*

	K54D	EAFV	K226	JQ2J	FTSE
K54D	1.000000	0.015817	-0.881190	0.685027	0.630622
EAFV	0.015817	1.000000	0.066406	-0.002333	-0.004220
JQ2J	0.685027	-0.002333	-0.523213	1.000000	0.529084
FTSE	0.630622	-0.004220	-0.560529	0.529084	1.000000

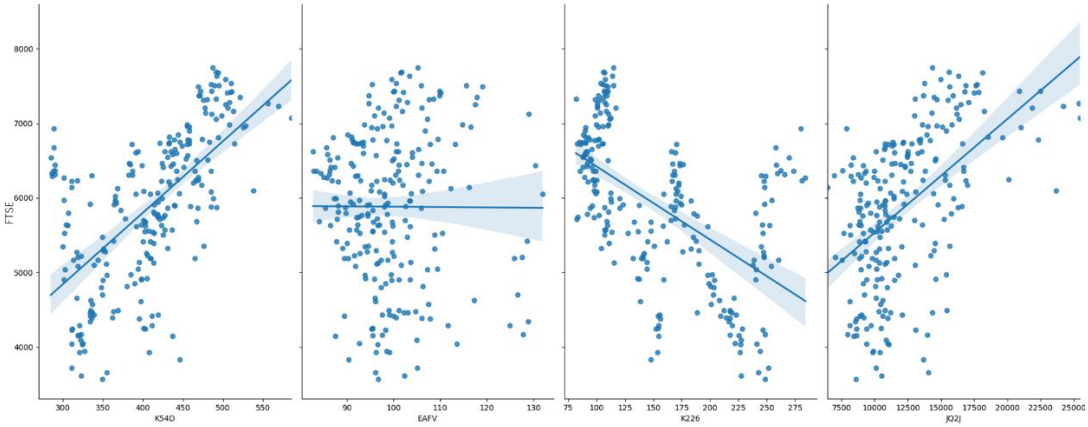


Figure 3-2 *Predict linear regression lines*

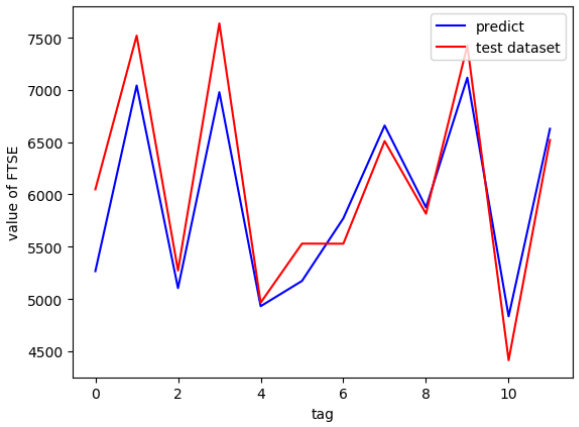


Figure 3-3 *Testing Result*

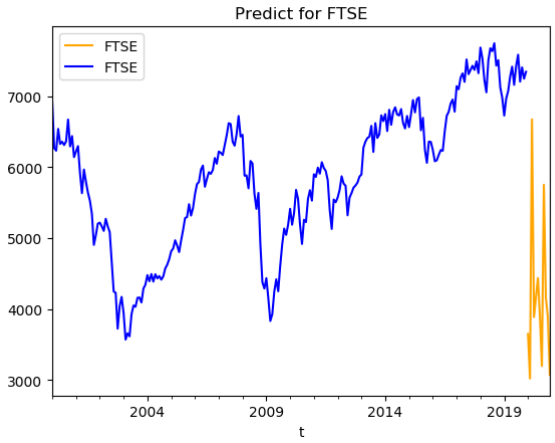


Figure 3-4 *Predicting Curve*

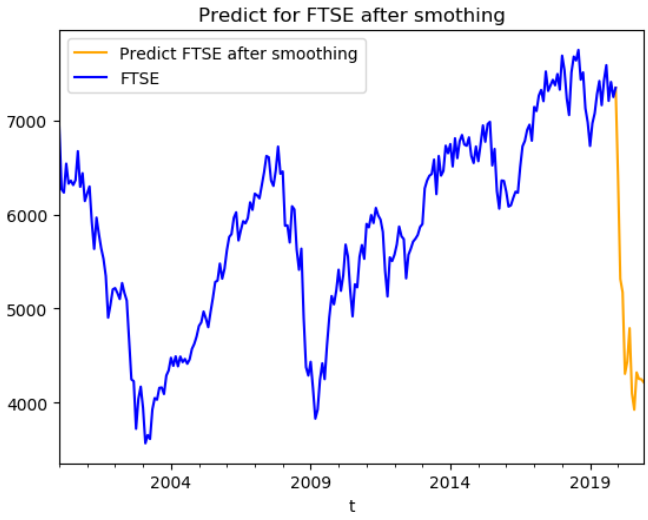


Figure 3-5 *Predicting Curve after smoothing*

Table 3-3 *Predicted value*

t	Predicted FTSE	Predict FTSE after smoothing
2020 JAN	3647.27	6412.59
2020 FEB	3017.75	5314.98
2020 MAR	6675.49	5171.75
2020 APR	3884.19	4306.17
2020 MAY	4167.30	4436.18
2020 JUN	4436.56	4790.88
2020 JUL	3900.10	4097.04
2020 AUG	3195.02	3924.75
2020 SEP	5751.55	4320.81
2020 OCT	4158.87	4251.39
2020 NOV	3891.56	4249.25
2020 DEC	3071.09	4218.27