TECHNICAL REPORT

32049072

1. Forecast monthly behaviour till December 2021 of 4 datasets

1.1 Data preparation

Given 4 data sets [Table 1.1] are time series with monthly data and varied time-length. Since the data is monthly, it might be useful to use calendar adjusted values to optimize our forecast. So calendar adjusted values are calculated in a separate worksheet (CalAdj). Furthermore, various statistics are computed in the same sheet to get an overview of the data, to check for outliers, to detect missing values, to detect negative values, etc. Outlier detection is done using inter-quantile range method. It is to be noted that several outliers are not replaced or imputed because they showed some periodically recurring pattern which we want to include in our predictions. Variance, standard deviation, skew, and kurtosis are calculated to understand distribution of data points. To detect possible seasonality in the dataset, yearly observations from initial years and recent years were taken and arranged on yearly basis in the sheet "Seas_plot". No data truncation is done for this task.

Abbreviation	Data	Excel File Name	Period
K54D	Monthly Average of Weekly pay	K54Ddata_32049072.xlsx	Jan 2000- Dec2020
EAFV	Retail sales index, household goods, all businesses	EAFVdata_32049072.xlsx	Jan 1988- Dec2020
K226	Index of production – Extraction of crude	K226data_32049072.xlsx	Jan 1997- Dec2020
JQ2J	Wholesale/Retail Trade Turnover	JQ2Jdata_32049072.xlsx	Jan 2000- Dec2020

Table 1.1

1.2 Preliminary Analysis

1.2.1: **K54D** Time series in Fig.1.1 shows moderate upward additive trend with seasonality. The steady growth shunted a little during 2008-10 probably due to 2008 economic crisis. From 2013 onwards it regained its growth trajectory. In 2020, there seems to be another deviation in growth trend, and this is due to the COVID19 pandemic. Seasonality is confirmed by reptitive patterns in the time series after fixed intervals and by observing ACF plot spikes at lag 12,24,36, etc. (Fig.1.2) which implies strong correlation between lagged datasets in every 12th month. Fig.1.3 shows seasonal plots which seem to imitate time-series growth trend. This implies multiplicative seasonality. Residuals of seasonal decompose also show the least Mean Squared Error for multiplicative seasonality. No cyclic pattern observed. Time plot of original data set as well as the calendar adjusted dataset do not seem to stabilize variance after transformation. Calendar adjustments also do not significantly improve forecasts because they tend to increase the variance as shown in the excel file: K54Ddata_32049072.xlsx: Sheet: CalAdj.

1.2.2: Similarly, it is found that **EAFV** Time series in Fig.1.4 shows overall additive growth trend with multiplicative seasonality [Refer Fig 1.5 and Fig 1.6]. The growth took a set-back during 2008-10 due to 2008 economic crisis. It started to regain its growth trajectory from 2014 but due to COVID19, sharp decline in few months after lockdown is seen. Interestingly, the growth-trend resumed drastically after mid-2020. This sharp decline will affect out forecasts. Observing the excel sheet: CalAdj in excel file EAFVdata_32040972.xlsx, these steep values are adjusted by taking the average of adjacent values. Moreover, the descriptive statistics chart in the same sheet shows that the calendar adjusted values, for which these outliers have been adjusted, have decreased in variance. Hence, to minimise forecast errors, calendar and outlier adjusted values will be used.

1.2.3: For **K226** Time series in Fig.1.7, an approximate decline trend is seen. No seasonality can be observed from time-plot analysis suggesting absence of any seasonal patterns. A smooth ACF plot confirms that seasonality is absent [Fig. 1.8]. This time plot [Fig 1.7] exhibits multiple local trends. Initially, some local upward trend is observed following which the graph shows gradual decline until 2012 after which it stabilized for a short period. Thereafter, it exhibited slow growth followed by gradual decline, forming a downward concave pattern. Calendar adjusted values are calculated and analysed. The variance has increased because of these values and hence it is decided to not use calendar adjusted values. Data truncation is not recommended because overall trend will be lost which will lead to inaccurate and erroneous forecasts.

1.2.4: Observing **JQ2J** Time series in Fig.1.9, a very slow linear growth trend can be seen. Various crest and troughs with short periodic recurrence can be seen. The amplitude of these crest and troughs seem to be increasing gradually which suggests that this series has multiplicative seasonality [Further confirmed using analysis in 1.2.1 along with Fig 1.11 and Fig 1.12]. ACF plots [Fig 1.11] seem to show a 6 monthly seasonality but closer inspection reveals that lagged values at intervals of 6 are smaller than lagged values at interval of 12. Hence, seasonal period is 12 months. A local trend from 2008-10 is observed during which the growth declined slightly. A steep drop can be observed towards the end of the graph [Fig 1.9]. This deviation must be adjusted to forecast with higher accuracy. Hence, as in 1.2.2, the relevant values have been adjusted and analysed in the excel sheet: CalAdj in excel file JQ2Jdata_32040972.xlsx. Since, the variance and standard deviation are reduced with calendar-outlier adjusted values, therefore, these values will be used for forecasts.

Descriptive statistics chart does not classify these values as outliers. Hence, visual analysis and judgement is necessary. Fig 1.10 shows a plot of calendar and outlier adjusted values plotted against the original time series.

1.3 Exponential Smoothing (ES) Method:

1.3.1 K54D: From 1.2.1 we know that the trend is additive, and seasonality is multiplicative. Only Holt-Winter's (multiplicative) ES method (HWS) can handle seasonality (multiplicative) and trend in a time series. This is also shown in Pegel's classification scheme in cell B-3. Fig 1.13 & Fig. 1.14 are obtained using file ExponentialSmoothing_K54DExpMethod_32049072.py which uses Holt-Winter's model to forecast for 12 months. The values of alpha (Smoothing Level), beta (Smoothing Slope), and gamma (Smoothing Seasonality) are 0.26697, 0.000, 0.61429 respectively for Model 1. These values are chosen as they have the least Mean Square Error (MSE) of 59.87. Model 2 and 3 are also generated using the same file. Their MSE is higher and therefore these values may overfit or underfit the model, and hence, they are rejected. Another important criterion to measure the forecast accuracy is to observe the residuals' correlogram. Fig 1.15 and Fig 1.16 shows ACF plot for Model 1 HWS and Model 3 HWS respectively. Notice the 1st lag in Model 1 is closer to zero than in Model 3. It denotes that errors in model-1 are more random. Though there is some alternating sine-wave pattern in both plots, it can be ignored as it is within the critical range. Furthermore, Pegel's classification grid shows formula for additive trend and multiplicative seasonality in Cell B-3 which are also called Holt-Winter's (Multiplicative Seasonality) Method. The pattern in residuals indicates that the model is not completely random, and this might not be the best model (as will be shown in the next task).

1.3.2 EAFV: From 1.2.2 we know that the trend is additive, and seasonality is multiplicative. The only difference is the use of calendar-outlier adjusted values. Before calculating MSE or plotting graphs, the values are back transformed to reflect true forecasts. Like 1.3.1, HWS is used and Fig 1.17 & Fig. 1.18 are obtained using file ExponentialSmoothing_EAFVExpMethod_32049072.py. The values of alpha, beta, and gamma are 0.455, 0.00, 0.3955 respectively for Model 1 with MSE 10.79 which is the lowest of all the other possible combinations. All further conclusions are similar as in 1.3.1 and Fig. 1.19, Fig. 1.20 should be referred for analysis.

1.3.3 K226: We know from section 1.2.3 that for this time-plot there is no seasonality and the trend is linear. Holt's Linear method (Cell B-1 in Pegel's classification) is superior to Simple Exponential Smoothing (SES) as it considers the trend component of the time-series. Here only two smoothing constants, alpha and beta, are used. Their values are 0.59 and 0.01 respectively. Again, these parameters are chosen based on the lowest MSE of all possible combinations. ACF plots of residuals in Fig 1.23 and Fig 1.24 also show that the residuals for model 2 appear more random and within critical limits than model 1. Hence, model 2 is more accurate at forecasting. Holt's Linear method is compared against SES using training set and test set in file ExponentialSmoothing_K226Compare_32049072.py. It is observed that SES has MSE of 81.1347, whereas Holt's Linear Smoothing has MSE of 56.00 which is significantly less. Note that the Holt-Winter's method is also computed which shows an MSE of 53. This is not very significant from Holt-Linear model and hence, ignored.

1.3.4 JQ2J: Section 1.2.4 showed how calendar adjustments are made and outliers corrected. The trend was additive and seasonality multiplicative. Hence, like in Section 1.3.2, back transform is required before plotting the graph and calculating residuals. Fig 1.25 and 1.26 shows two plots drawn using HWS method. Optimized values for alpha, beta and gamma are 0.429, 0.0529, and 0.4056, respectively and the MSE obtained is 1.019e+06 which is the least among various combinations of smoothing parameters. From ACF plots in Fig. 1.27 and Fig. 1.28, observe that the 1st lag in model 2 is very close to zero, almost at the critical threshold. This shows that the series is mostly random. Further comparison of non-calendar adjusted values and untreated-outlier values was done and it was found that the minimum MSE is generated only by adjustment of calendar values and outlier averaging.

The Mean Absolute Error (MAE) value is also a good measure of selecting an appropriate model. However, MSE is chosen as a widely accepted statistical measure to determine goodness of fit and, therefore, to standardize the report it is being used throughout. With any exponential smoothing method, there are 3 major limitations which are initialization, optimization and prediction intervals.

2. ARIMA Forecasting

2.1 Data Preparation:

All the data preparation work has been done as described in section 1.1. We have already concluded that the calendar adjusted values won't be used so this task uses sheet "Data" in K54Ddata_3209072.xlsx containing the original time series.

2.2 Preliminary analysis:

Continuing our analysis from section 1.2.1, it is observed that because there is trend and seasonality present, the time-series is not stationary. A PACF plot (shown in Fig 2.1) supports this argument because the lag1 point is large signifying that there is relation between alternate data points. Therefore, the time series is non-stationary.

2.3 ARIMA model:

2.3.1 Before proceeding with ARIMA model, we need to make the series stationary. Seasonal differencing (D=1) is applied to remove seasonality (s=12) (plot shown in Fig 2.2). The ACF and PACF still exhibit pattern as shown in Fig 2.3 and Fig 2.4 and to remove this pattern a first difference is taken (d=1). In Fig. 2.6 (ACF) and Fig. 2.7 (PACF) of seasonal differenced and first differenced data, it can be seen a lot of seasonality is removed, and the trend seems to become more random. There are still some values breaching the critical threshold. This tells us that the model is not pure white noise but very close to it. The spike in ACF value at lag 1 is negative and quickly dies out. And, the PACF at lag 1 is negative and shows exponential decay of the first few lags. This is characteristic of a non-seasonal MA(1) model. It also suggests that the moving average parameter [θ1] > 0. We further observe that the ACF at lag1, the value r_1 is significant – reinforcing the non-seasonal MA(1) model – and r_{12} is significant – suggesting a seasonal MA(1) model. With similar further observation in Fig 2.7, PACF also has significant value at 12th lag which supports the seasonal MA(1). Therefore, our initial estimate of the possible ARIMA(p,d,q)(P,D,Q)s model is: ARIMA(0,1,1)$(0,1,1)_{12}$ (Ref Fig 2.8 and Fig 2.9)

2.3.2 The model's accuracy is tested first with Ljung-Box test (portmanteau test). Because the P-value of the Q-statistics is 0.81 > 0.05, therefore, our model passes the Q-Statistics test. The AIC of our model is 1662.27 and its MSE is 44.1. A few other models with different values of parameters is tried using *ARIMA_K54DMultValue_32049072.py* and an output is created in *Possible_Arima_combinations.xlsx (Also presented in Sheet: ARIMA in Excel: K54Ddata_32049072.xlsx)*. When compared with other models, our AIC is the 4th lowest. The Ljung-Box statistic (portmanteau test) and MSE is generated for the lowest 5 AIC parameters with their respective (p,d,q)(P,D,Q)s. The ARIMA(1,1,1)$(0,1,1)_{12}$ model generates the AIC value of 1650.224 with MSE of 43.89. Its second lowest but requires addition of only 1 parameter. Another ARIMA(1,1,1)(1,1,0) $_{12}$ has lower AIC than our initial ARIMA model but its MSE is higher, hence this is rejected. The ARIMA(1,1,1)(1,1,1) $_{12}$ model with the lowest AIC (1649.49) also has the lowest MSE (43.72). Due to the difference being marginal, ARIMA(1,1,1)$(0,1,1)_{12}$ (Ref Fig. 2.10 and Fig. 2.11) model is the appropriate as it gives and (significantly) lower AIC and (marginally lower) MSE on cost of addition of only 1 parameter.

2.4 ARIMA and exponential smoothing forecasting

Linear exponential smoothing (ES) models are all special cases of ARIMA models. But the non-linear ES models have no equivalent ARIMA equivalents. Likewise, there are ARIMA models that have no ES counterparts. ES and ARIMA are both black-box methods. While some ARIMA models are stationary, all ES models are non-stationary. ARIMA models aim to describe the autocorrelations in the data, while exponential smoothing models are based on a description of the trend and seasonality in the data.

In this instance, our time series has multiplicative seasonality. We used Holt-Winters Multiplicative seasonality to forecast in section 1.3.1 above. Unfortunately, Holt-Winters' Multiplicative method has no equivalent ARIMA model. Though AIC can be generated for both, its value cannot be used as a measure to choose suitable model because AIC is only useful for selecting models between the same class (like ARIMA with ARIMA or ES with ES). However, it will be suitable to compare their respective MSE performance because we are comparing the models on their genuine forecasting ability. The MSE for HWS = 59.86. and the MSE for ARIMA(1,1,1)$(0,1,1)_{s=12}$ is 43.89. Therefore, our chosen ARIMA model performs better than Holt Winter's Multiplicative method for the given time series because the MSE is lower.

3. Regression Prediction

3.1 Data Preparation:

To regress on multiple explanatory variables the time series has to be of standard length. Hence, a common year – 2000- is chosen and taken as base line. Further, in order to find value for coefficients of independent variables the data from 2011 to June2020 is used. This will be used to predict the values for 6months and calculate the MSE. The outliers are not adjusted because they exhibit some pattern. No transformation has been done. Given 4 independent variable and 1 dependent variable, their correlations are observed to understand any underlying relation.

3.2 Preliminary analysis:

Correlation plot and values are observed to estimate possible cases of multicollinearity. A correlation scatter graph plot is shown in Fig 3.1. Variables have moderate to weak correlation. Still some multicollinearity is detected. It does not affect our forecast but we should be careful when eliminating independent variables from our regression equation. An ANOVA table is generated which shows that the R^2(Adjusted) value is only 0.509 which means that about 50% of the data can be explained by our hyperplane. But the F-Statistics is very small confirming that the four variables are accounting for a significant part in the variation in FTSE (Y). Further analysis of the OLS regression result shows that t-statistics are below $P<0.05$ and hence all variables can explain the pattern.

3.3 ARIMA model:

To better fit the model, 11 monthly explanatory variables were introduced in the equation to capture errors due to seasonality. A time variable was also used to lessen the error due to time. Various combinations were tried and the maximum R^2 Adjusted value was 0.680 by removing one variable EAFV and several other monthly variables. An OLS tables is shown in Fig 3.2 showing various outputs. The lower F-statistic and the lower AIC compared to our previous ANOVA table shows that the model is appropriate. Using file Regression_FTSEForecasts_32049072.py, firstly training set is used to observe the MSE with the regression equation and then, in the same file, forecasts for next 12month are made. The figure for training set is shown in fig 3.3 and 3.4. There it is clear that the graph follows forecast trend. There is a lot of scope of improvement in that forecast for example, instead of Holt exponential smoothing, ARIMA could have been used. Data outliers could have been adjusted. Better values for coefficients could be chosen. The residual plots could have showed whether the model has reduced errors in the equations or not.

Appendix A

ExponentialSmoothing_K226TimePlot_32049072.py, ExponentialSmoothing_K54DTimePlot_32049072.py, ExponentialSmoothing_JQ2JTimePlot_32049072.py, ExponentialSmoothing_EAFVTimePlot_32049072.py	Plot the time plot and transformation of the respective datasets
ExponentialSmoothing_K226SeasDecom_32049072.py, ExponentialSmoothing_K54DSeasDecom_32049072.py, ExponentialSmoothing_JQ2JSeasDecom_32049072.py, ExponentialSmoothing_EAFVSeasDecom_32049072.py	Plot ACF, seasonal decompose and seasonal plots.
ExponentialSmoothing_K226ExpMethod_32049072.py, ExponentialSmoothing_K54DExpMethod_32049072.py, ExponentialSmoothing_JQ2JExpMethod_32049072.py, ExponentialSmoothing_EAFVExpMethod_32049072.py	Employs exponential smoothing methods
ExponentialSmoothing_K226CalAdj_32049072.py, ExponentialSmoothing_K54DCalAdj_32049072.py, ExponentialSmoothing_JQ2JCalAdj_32049072.py, ExponentialSmoothing_EAFVCalAdj_32049072.py	Plots calendar adjusted values and their transformations
ExponentialSmoothing_K226Compare_32049072.py	Compares SES and Holt Linear smoothing
ARIMA_K54DMethod_32049072.py	Plot PACF + calculates seasonal difference and first difference and plots it
ARIMA_K54DMultValue_32049072.py	Runs various combinations of ARIMA parameters and outputs valid parameters in an excel file
ARIMA_K54DSARIMA_32049072.py	Calculates ARIMA model for the given parameters
Regression_FTSECorrelation_32049072	Plot correlation scatter graph
Regression_FTSEanova_32049072.py	Calculates ANOVA and finds correlation coefficients
Regression_FTSEIndiTime_32049072.py	Uses indicator variables to solve regression equation
Regression_FTSEForecasts_32049072.py	First uses training set to computer MSE and forecasts and then forecasts for the 12 month period.

APPENDIX B _ Graphs and Diagrams

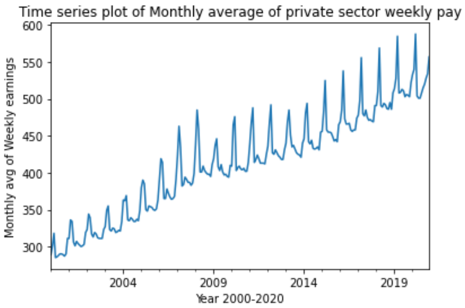


Figure 1.1

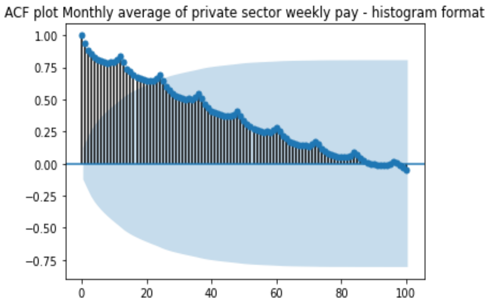


Figure 1.2

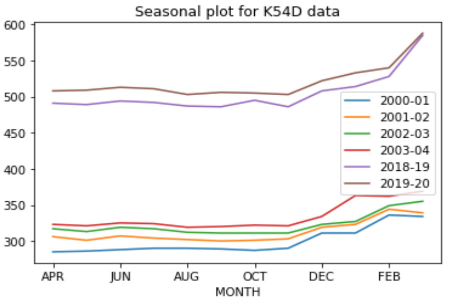


Figure 1.3

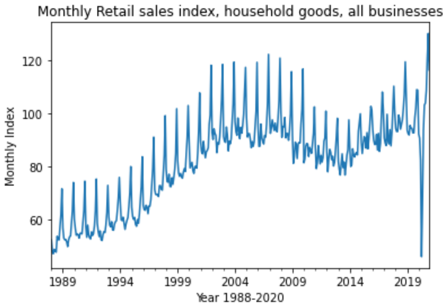


Figure 1.4

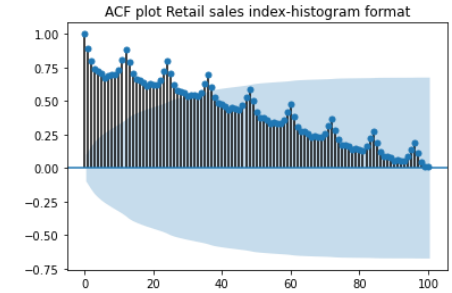


Figure 1.5

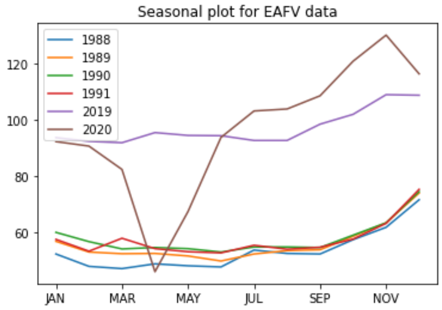


Figure 1.6

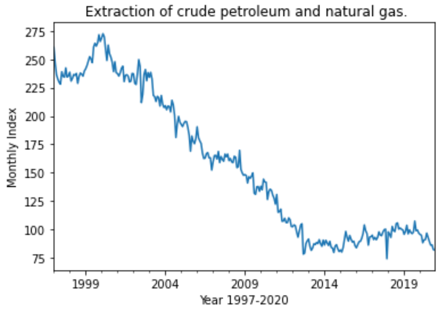


Figure 1.7

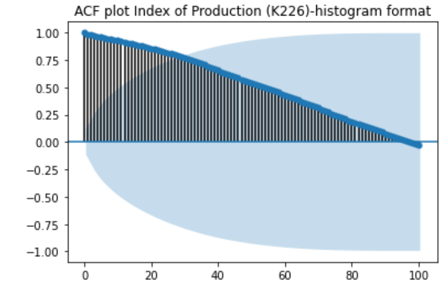


Figure 1.8

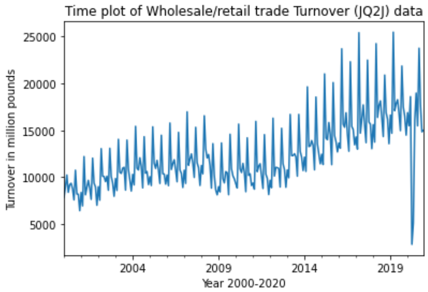


Figure 1.9

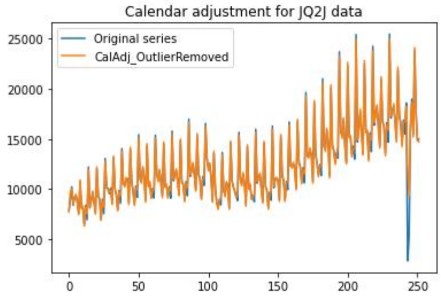


Figure 1.10

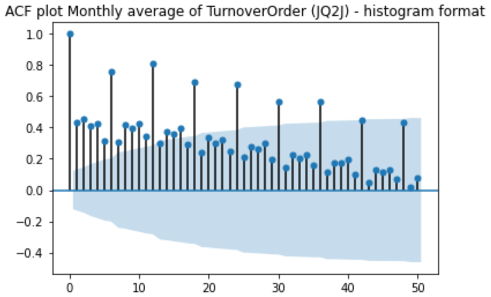


Figure 1.11

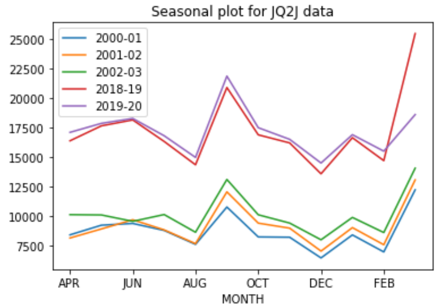


Figure 1.12

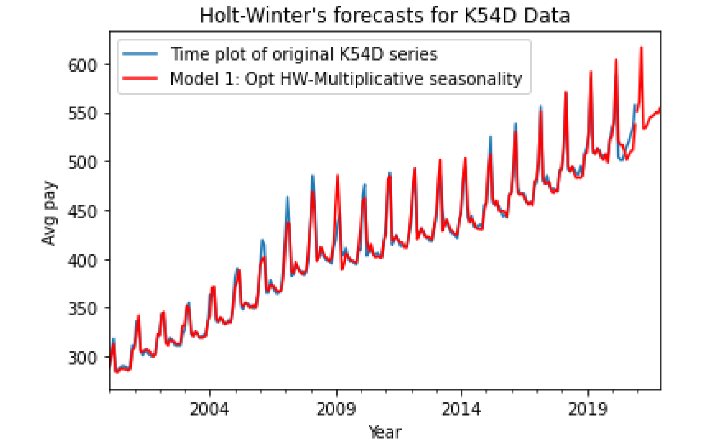


Figure 1.13

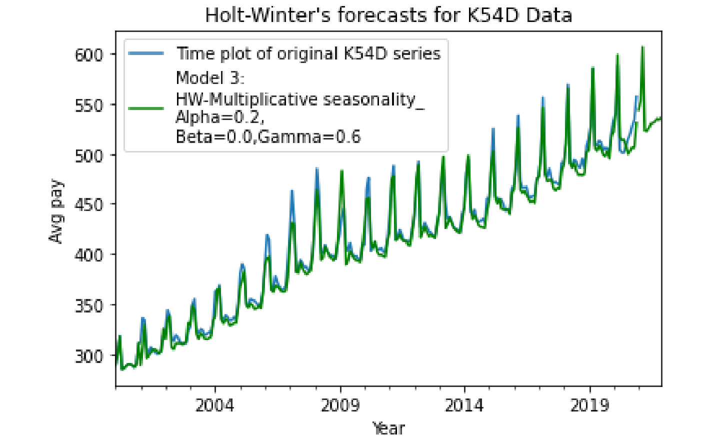


Figure 1.14

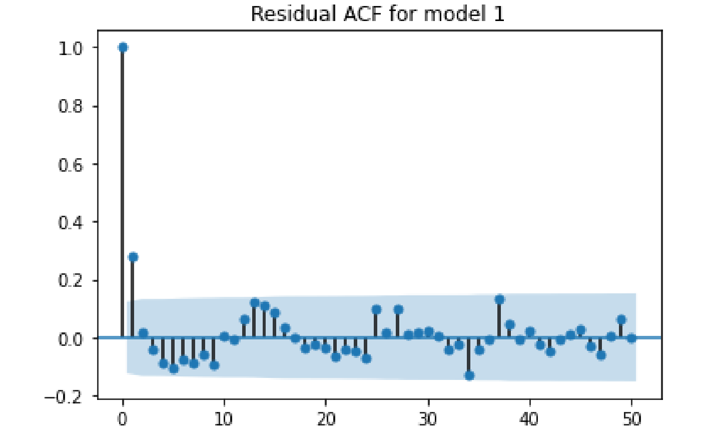


Figure 1.15

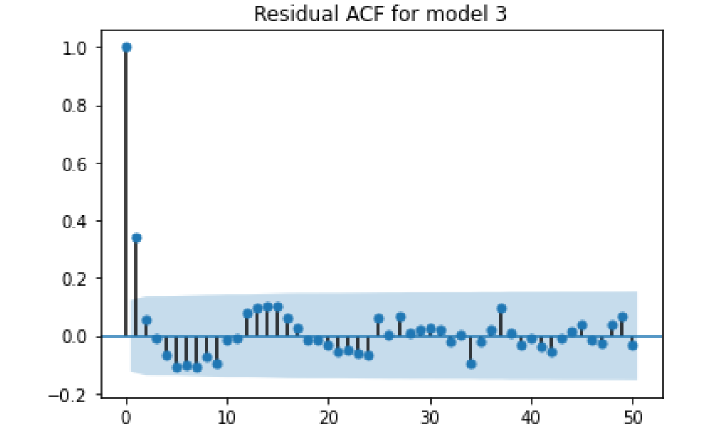


Figure 1.16

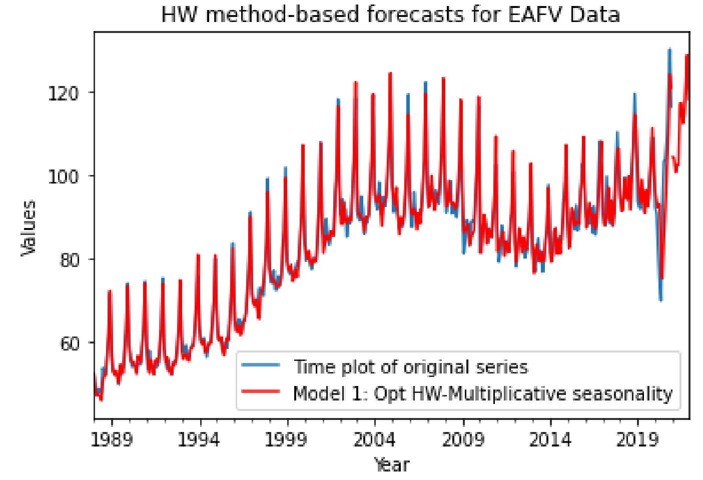


Figure 1.17

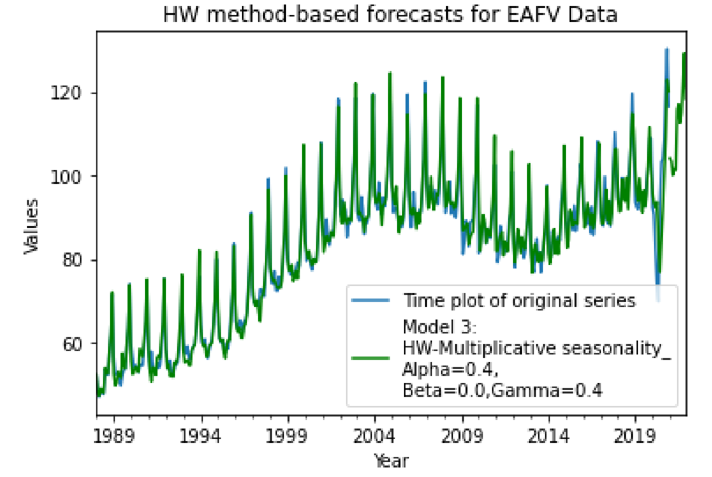


Figure 1.18

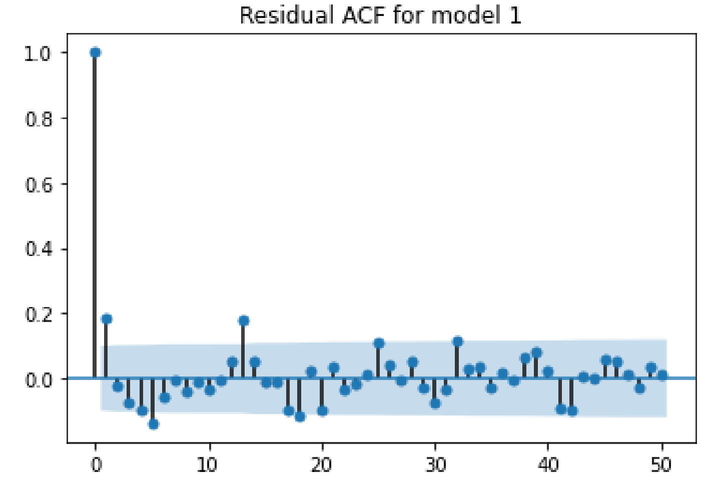


Figure 1.19

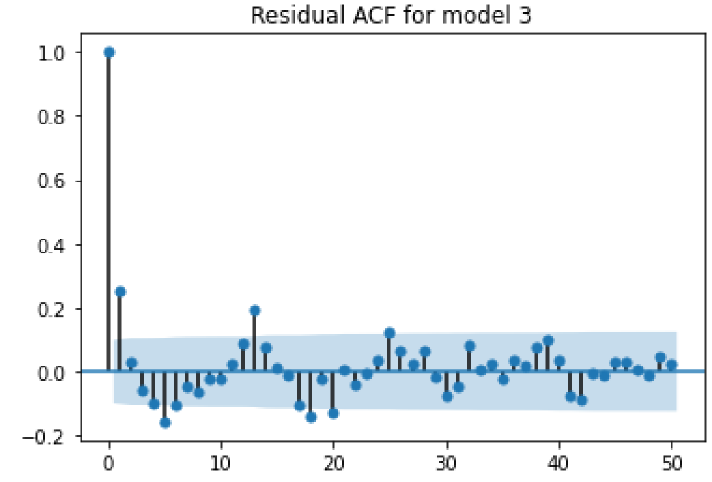


Figure 1.20

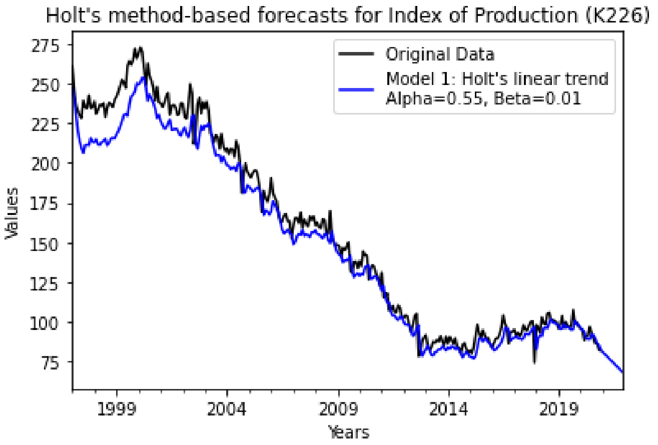


Figure 1.21

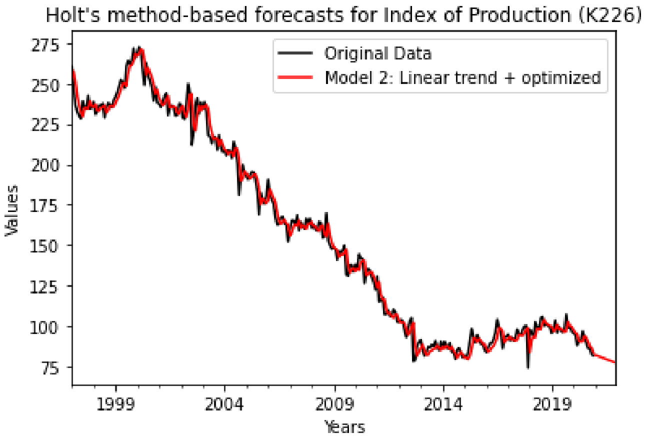


Figure 1.22

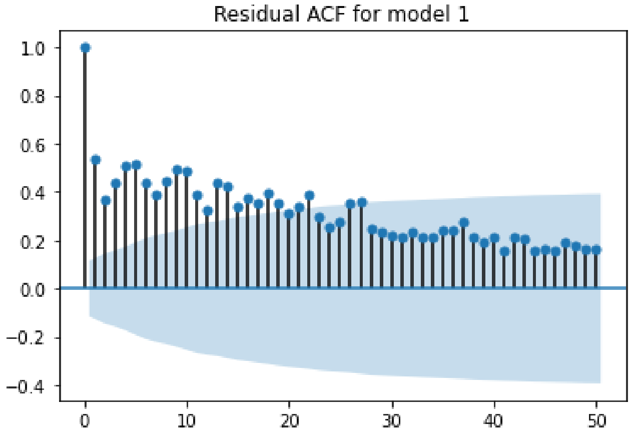


Figure 1.23

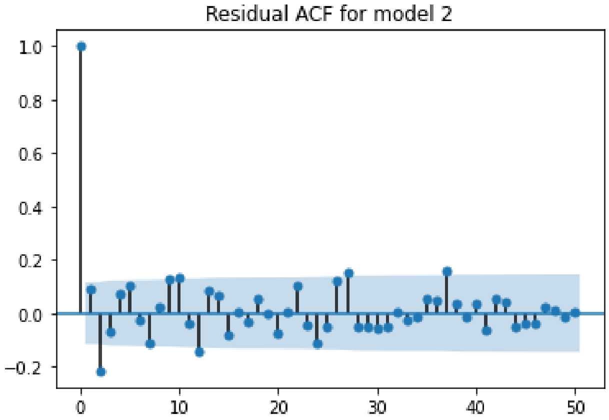


Figure 1.24

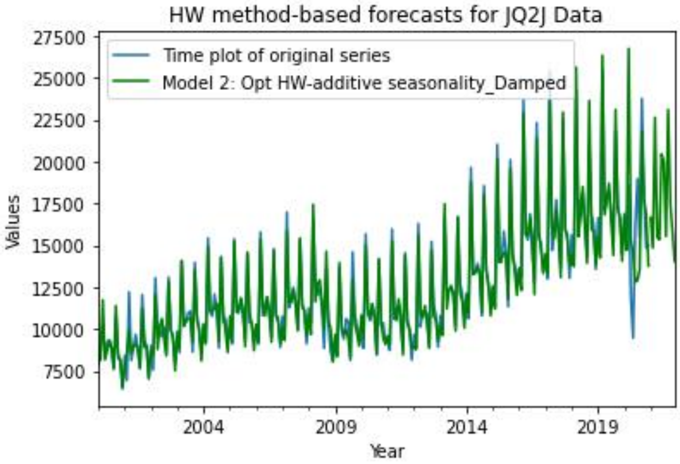


Figure 1.25

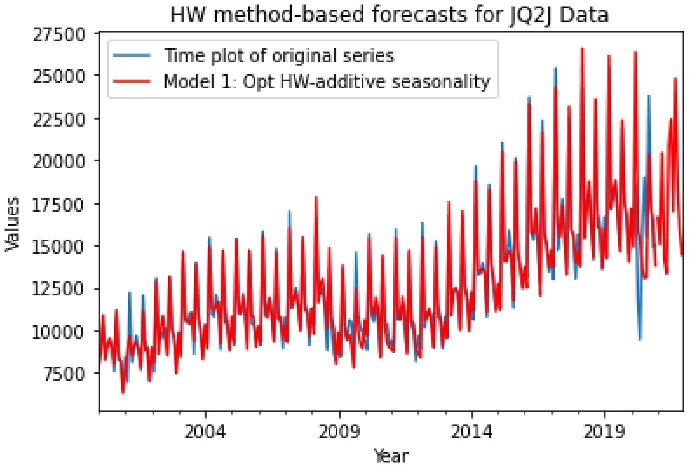


Figure 1.26

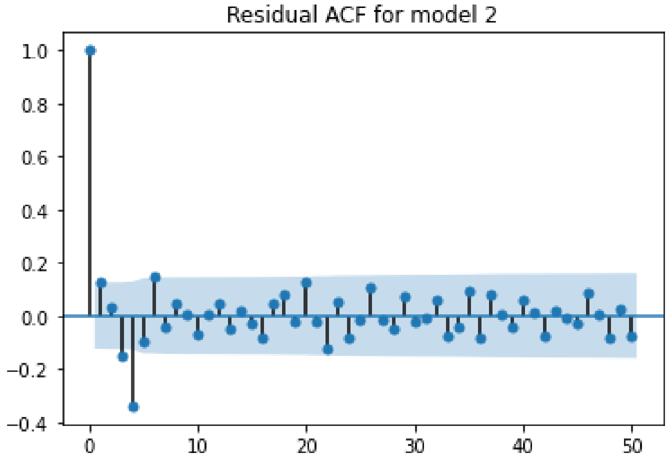


Figure 1.27

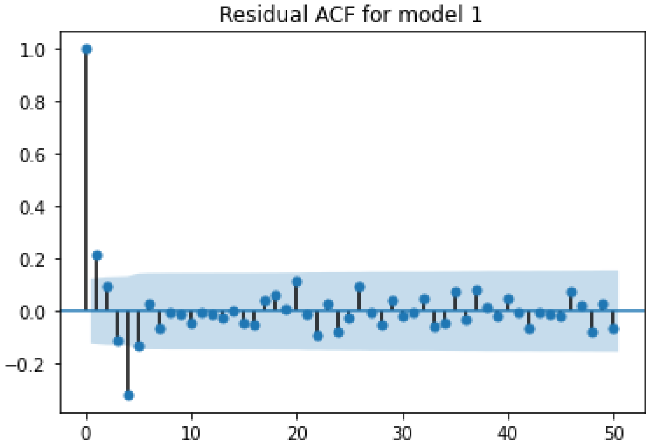


Figure 1.28

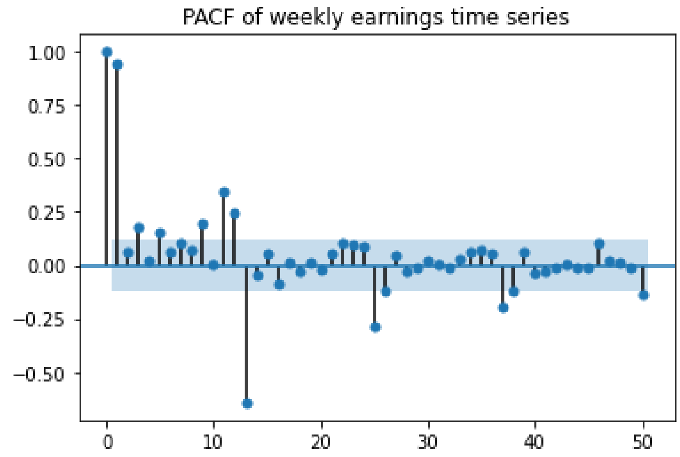


Figure 2.1

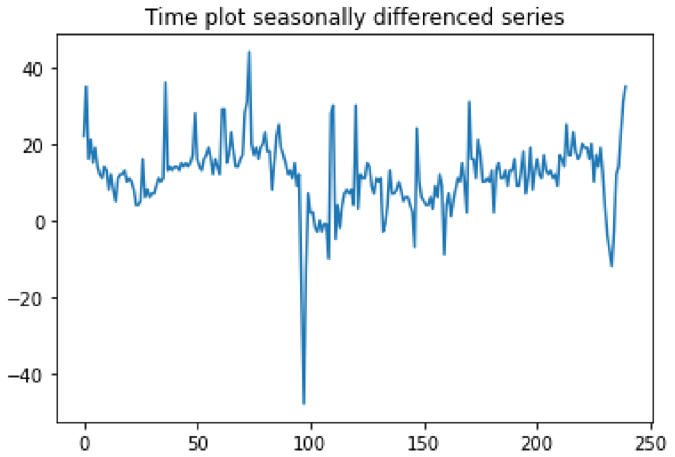


Figure 2.2

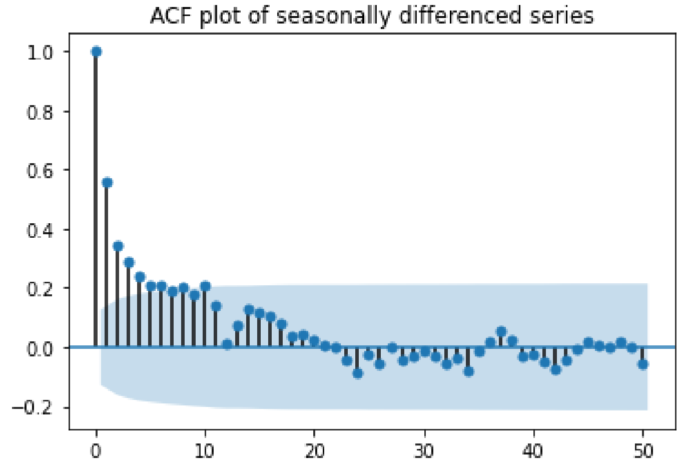


Figure 2.3

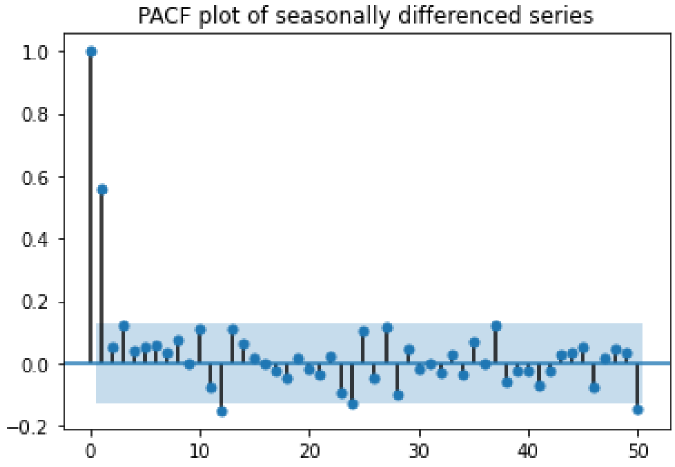


Figure 2.4

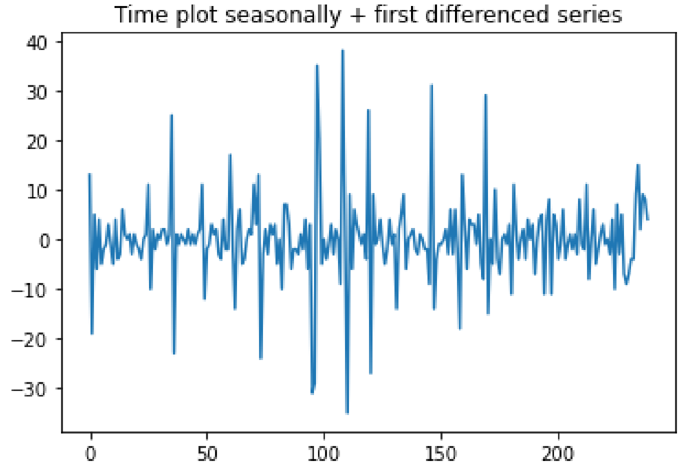


Figure 2.5

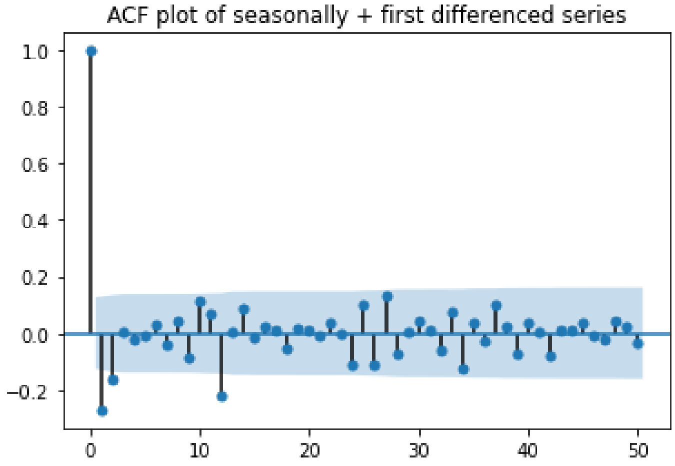


Figure 2.6

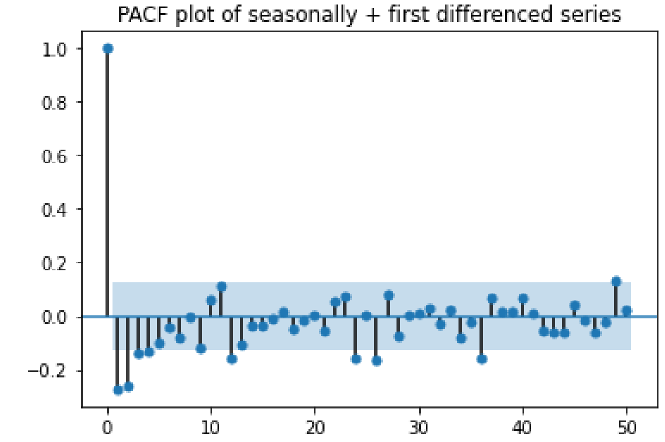


Figure 2.7

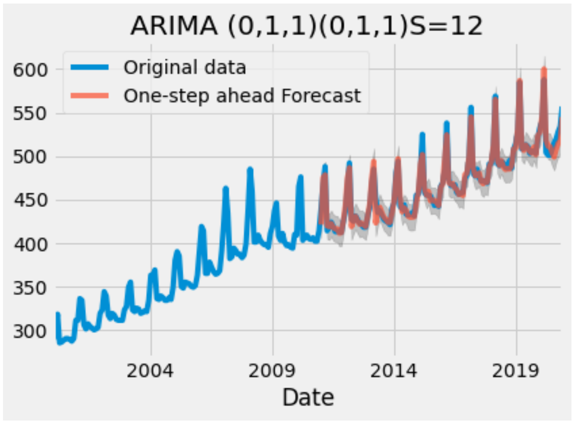


Figure 2.8

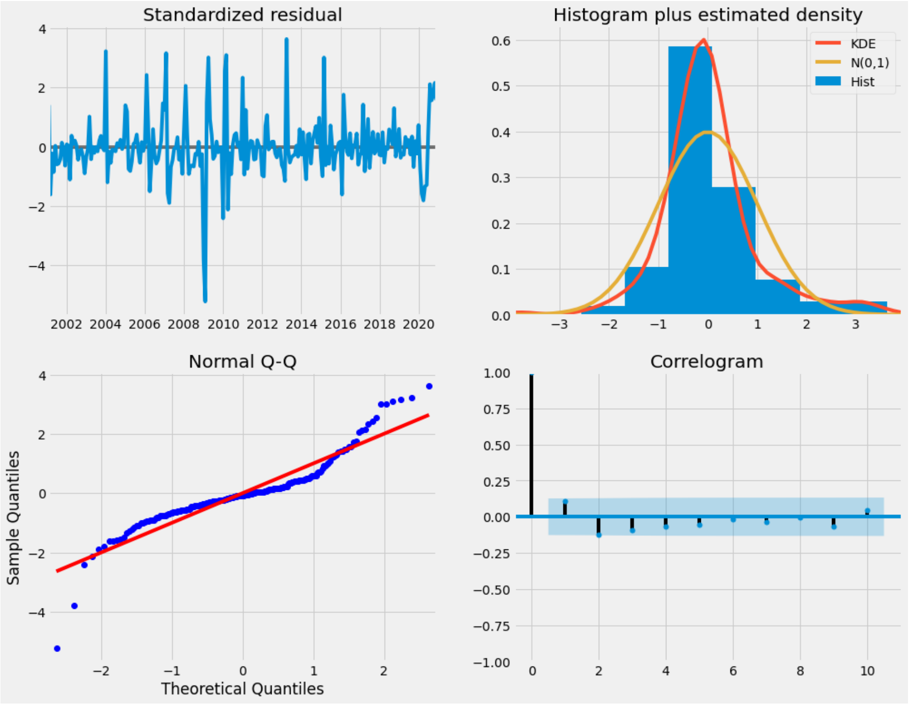


Figure 2.9

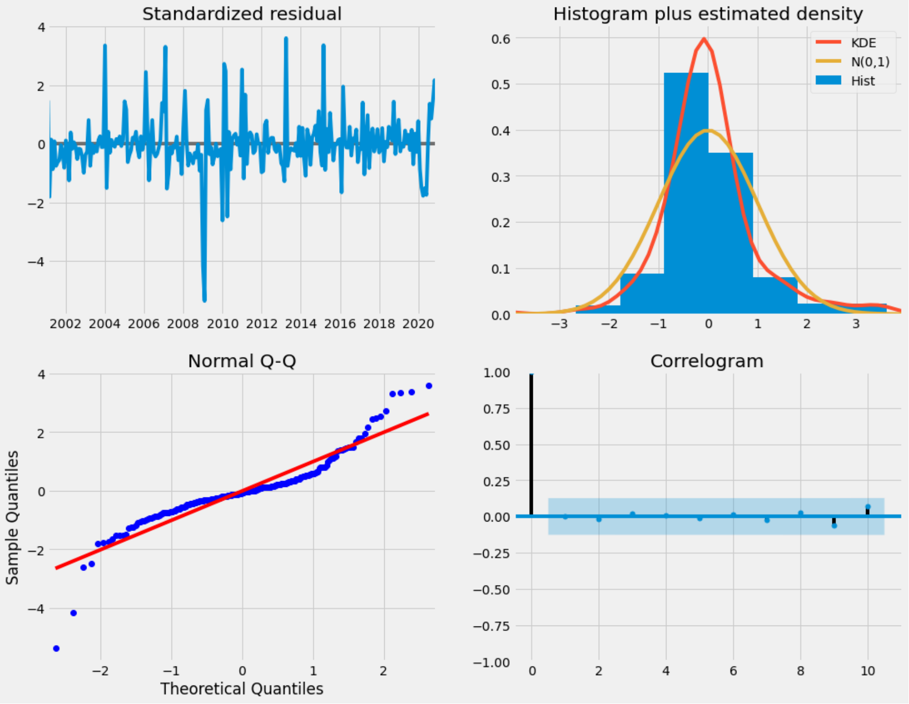


Figure 2.10

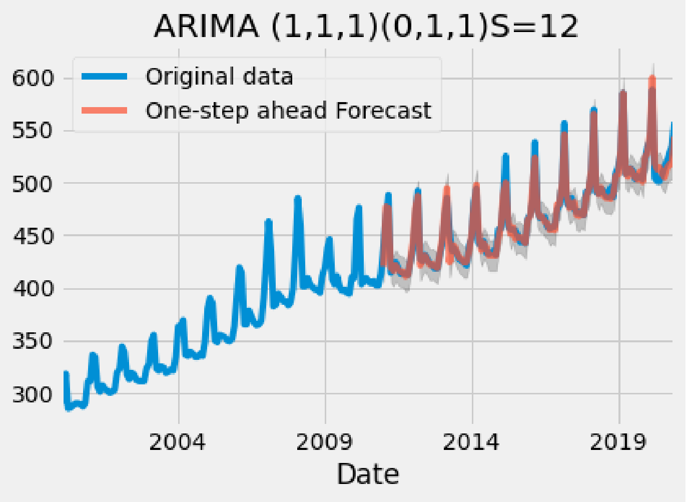


Figure 2.11

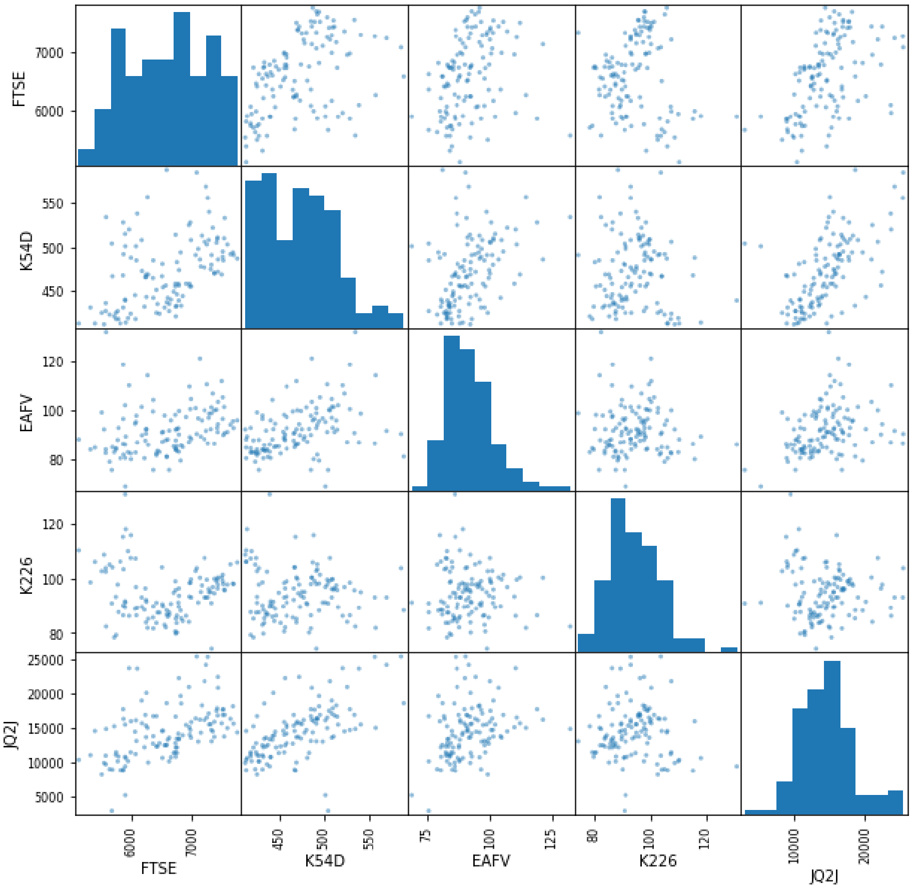
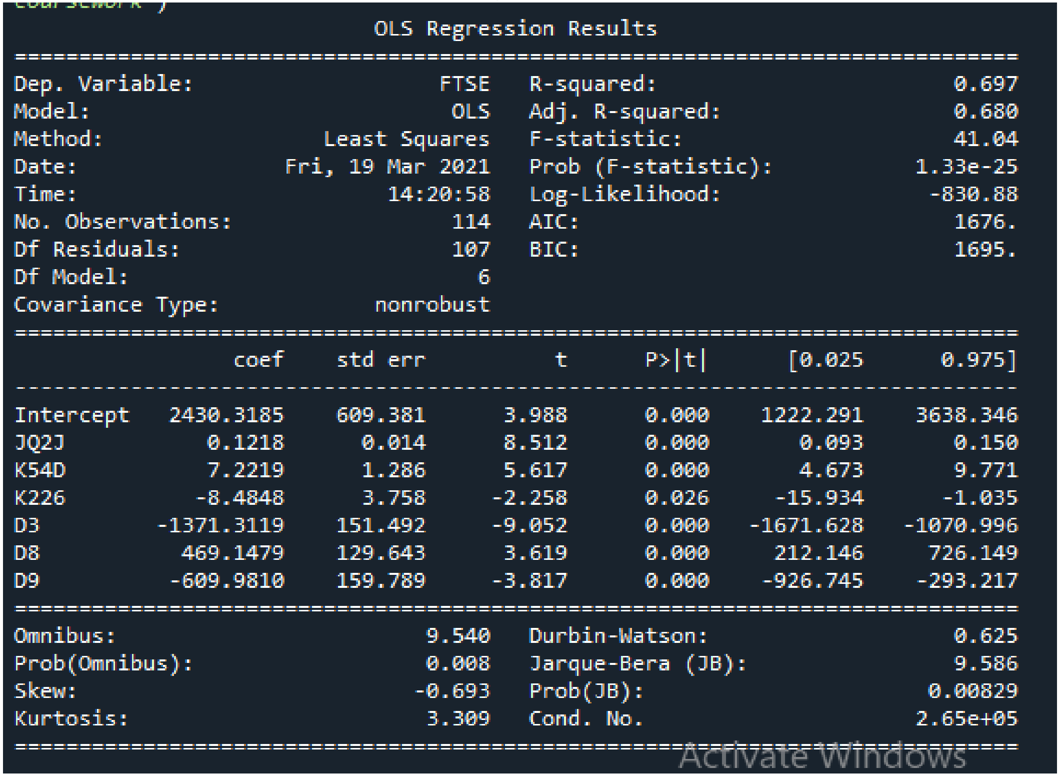


Figure 3.1



```
coursework')
                            OLS Regression Results
==============================================================================
Dep. Variable:                   FTSE   R-squared:                       0.697
Model:                            OLS   Adj. R-squared:                  0.680
Method:                 Least Squares   F-statistic:                     41.04
Date:                Fri, 19 Mar 2021   Prob (F-statistic):           1.33e-25
Time:                        14:20:58   Log-Likelihood:                -830.88
No. Observations:                 114   AIC:                             1676.
Df Residuals:                     107   BIC:                             1695.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   2430.3185    609.381      3.988      0.000    1222.291    3638.346
JQ2J           0.1218      0.014      8.512      0.000       0.093       0.150
K54D           7.2219      1.286      5.617      0.000       4.673       9.771
K226          -8.4848      3.758     -2.258      0.026     -15.934      -1.035
D3         -1371.3119    151.492     -9.052      0.000   -1671.628   -1070.996
D8           469.1479    129.643      3.619      0.000     212.146     726.149
D9          -609.9810    159.789     -3.817      0.000    -926.745    -293.217
==============================================================================
Omnibus:                        9.540   Durbin-Watson:                   0.625
Prob(Omnibus):                  0.008   Jarque-Bera (JB):                9.586
Skew:                          -0.693   Prob(JB):                      0.00829
Kurtosis:                       3.309   Cond. No.                     2.65e+05
==============================================================================
```

Figure 3.2

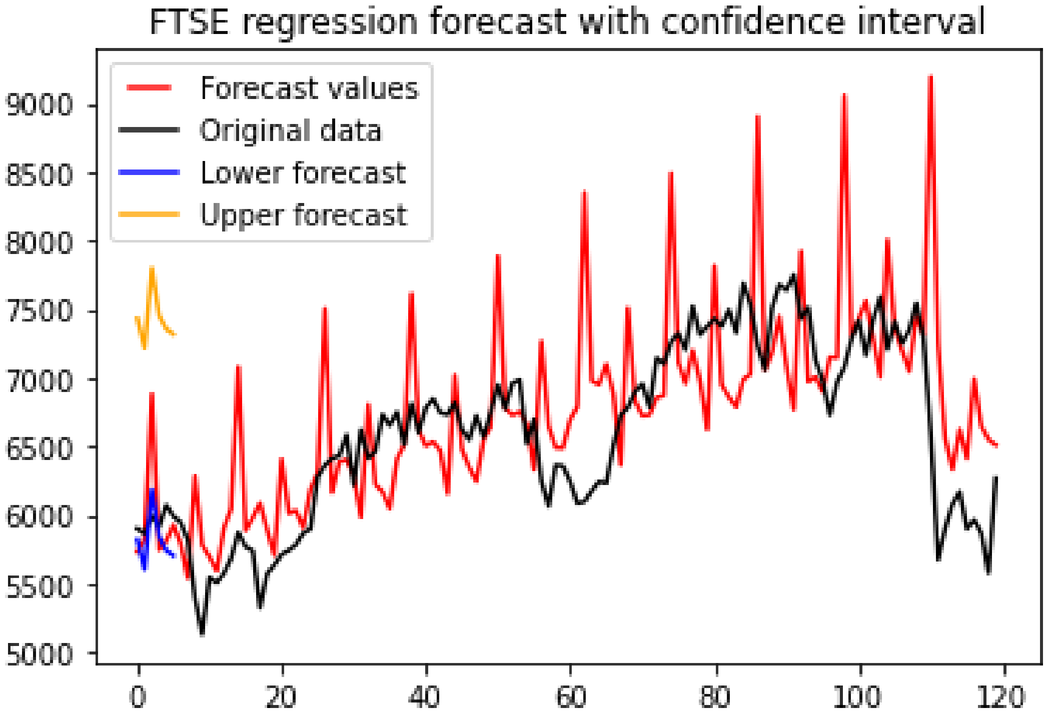


Figure 3.3 (Training)

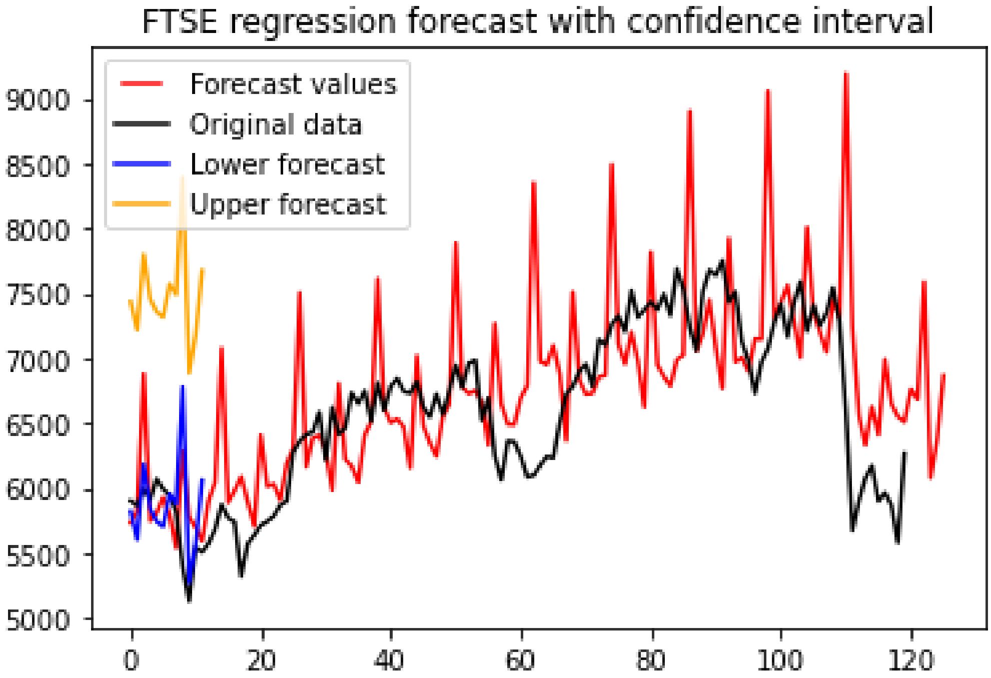


Fig 3.4(Full_forecast)