Statistical Techniques and Time Series

Task 1: IMPORTING THE DATA SET

The S&P 500 dataset for the time period 2009/01/01 to 2020/12/31 was collected from Yahoo Finance. The dataset has 3020 rows and 6 columns with the date as the index. The function names(dataset_name) is used to run the column names.
names (GSPC) "GSPC.Open" "GSPC.High" "GSPC.Low" "GSPC.Close" "GSPC.Volume" "GSPC.Adjusted"

	GSPC.Open	GSPC.High	GSPC.Low	GSPC.Close	GSPC.Volume	GSPC.Adjusted
2009-01-02	902.99	934.73	899.35	931.80	4048270000	931.80
2009-01-05	929.17	936.63	919.53	927.45	5413910000	927.45
2009-01-06	931.17	943.85	927.28	934.70	5392620000	934.70
2009-01-07	927.45	927.45	902.37	906.65	4704940000	906.65
2009-01-08	905.73	910.00	896.81	909.73	4991550000	909.73
2009-01-09	909.91	911.93	888.31	890.35	4716500000	890.35
2009-01-12	890.40	890.40	864.32	870.26	4725050000	870.26
2009-01-13	869.79	877.02	862.02	871.79	5567460000	871.79
2009-01-14	867.28	867.28	836.93	842.62	5407880000	842.62
2009-01-15	841.99	851.59	817.04	843.74	7807350000	843.74
2009-01-16	844.45	858.13	830.66	850.12	6786040000	850.12

Showing 1 to 12 of 3,020 entries, 6 total columns

FIGURE 1: HEAD OF DATA THAT WE HAVE LOADED AND LINE GRAPH OF THE ADJUSTED COL WITH RESPECTIVE TO TIME

The line graph shows that the Adjusted stock price was increasing constantly from the year 2009 to 2020

Task 2: PERFORMING LOG RETURNS

On the dataset, the log-returns method is used. Log returns are advantageous since they merely eliminate unstable components from the data set, standardize the dataset, and increase its stability.

$$r_t = \log(y_t)\text{-}\log(y_{t-1})$$

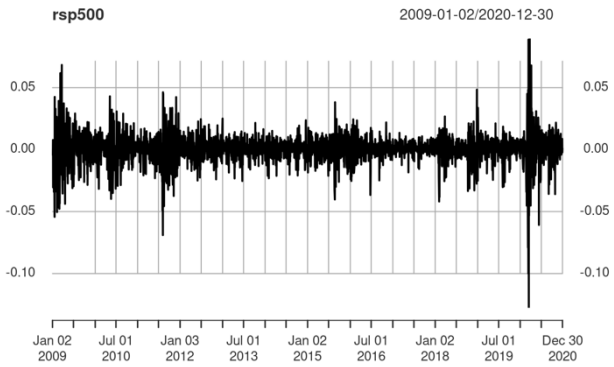

FIGURE 2:LOG RETURNS

The log returns show that the mean is near zero and data volatility increases after July-01-2019

Task 3: PERFORMING ACF & PACP

ACF, short for autocorrelation function, is a function used to plot a graph and determine the correlation and lag values. As its name implies, the partial autocorrelation function (PACF) yields partial correlation values as well as the following lag values.

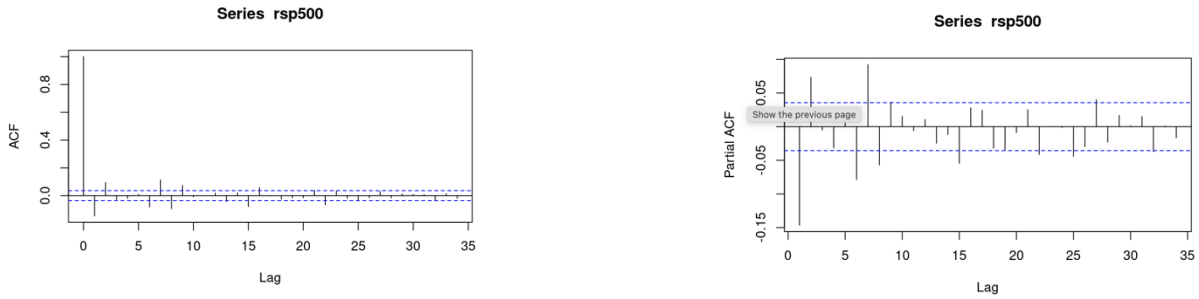

FIGURE 3: ACF & PACF PLOT

In ACF plot Lag starts with zero which means time series data has a correlation and the blue area here indicates data is statistically zero.

Task 4: LJUNG-BOX TEST

The ljung-box test is used to determine whether the variables are correlated and to determine the validity of the hypothesis. If the p-value for the hypothesis is less than 0.05, we reject the null hypothesis and accept the alternative hypothesis.

SYNTAX: Box.test(rsp500, lag = 20, type = "Ljung-Box")

RESULTS:
Box-Ljung test
data: rsp500
X-squared = 237.69, df = 20, p-value < 2.2e-16

As the p value is 2.2e-16 which is lesser than 0.05 we can say that the variables are correlated and accept the alternative hypothesis.

Task 5: STATIONARITY TEST

R's augmented Dickey-Fuller test, A time series is deemed "stationary" if there is no trend, a constant variance over time, and a constant autocorrelation structure across time and it is resulted as

SYNTAX: adf.test(Price) kpss.test(rsp500)

RESULTS:

Augmented Dickey-Fuller Test

data: Price

Dickey-Fuller = -3.8165, Lag order = 14, p-value =0.01815, alternative hypothesis: stationary

KPSS Test for Level Stationarity
data: rsp500
KPSS Level = 0.016899, Truncation lag parameter = 9, p-value = 0.1

Output indicates the Augmented Dickey-Fuller Test is successful and the data is stationary

Task 6: NORMALITY TEST

The Shapiro-Wilk test is a statistical test that determines whether the data distribution as a whole deviates from a comparable normal distribution. If the test is non-significant (p>. 05), it means that the sample's distribution is not significantly different from a normal distribution.

Shapiro-Wilk normality test
data: as.vector(rsp500)
W = 0.87502, p-value < 2.2e-16

The p-value indicates that the normal distribution is true for the given data.

Task 7: FITTING ARIMA MODEL

To better comprehend data or make predictions about the future, AutoRegressive Integrated Moving Average (ARIMA) models are fitted to historical time series data. The three components that make up an ARIMA model are the Auto Regressive (AR) section, Moving Average (MA), which indicates that the error is a linear mixture of prior errors, and Integrated (I), which describes the differencing used to convert non-stationary time-series into stationary. The name ARIMA(p,d,q) stands for ARIMA with an autoregressive model.

- p- lag value
- d- differential value
- q – error value

$$\hat{Y}_t = \mu + \varphi_1 Y_{t-1} + \ldots + \varphi_p Y_{t-p} - \theta_1 e_{t-1} - \ldots - \theta_q e_{t-q}$$

The data is fitted into ARIMA Model and the lag value is determined

ARIMA (2,1,4) with drift:27514.18

Best model: ARIMA (2,1,4) with drift

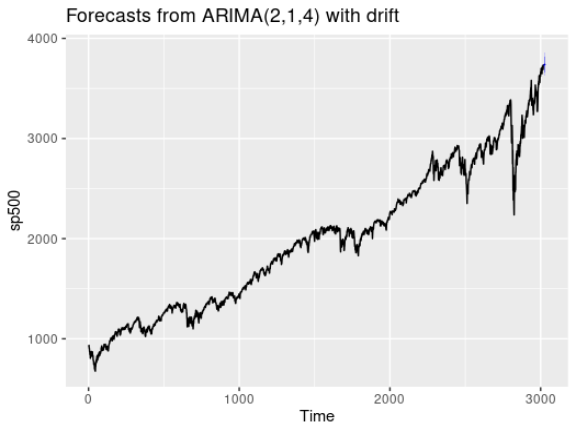

FIGURE 4: FORECAST OF THE FIT

Task 8: COEFFICIENTS FOR THE CHOSEN ARIMA MODEL

The coefficients are executed by using summary() followed by the model in the parameter.

SYNTAX: Summary(arimafit)

RESULTS:

Coefficients:	Ar1	Ar2	Ma1	Ma2	Ma3	Ma4	Drift
S.E	-1.7374 0.0256	-0.8815 0.0223	1.6171 0.0310	0.7773 0.0395	0.0695 0.0382	0.0378 0.0248	0.9411 0.4049

In equation: -1.73y_{t-1} −0.88y_{t-2}+1.61e_{t-1} +0.77e_{t-2}+0.06e_{t-3} + 0.03e_{t-4} + 0.94

TRAINING SET ERROR MEASURE:

Error measures:	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-0.01344049	22.98904	14.21841	-0.01580041	0.7545495	0.998118	0.001213312

Task 9: RESIDUALS

data: Residuals from ARIMA(2,1,4) with drift

Q* = 35.784, df = 4, p-value = 3.206e-07
Model df: 6. Total lags used: 10

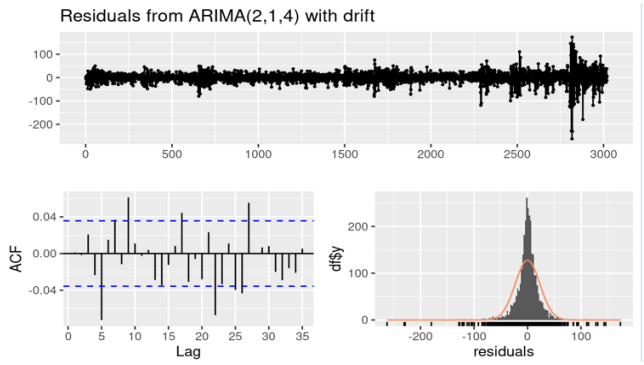

FIGURE 5: RESIDUALS

The residuals mean $E[e_t]$ =-0.01344049

The residuals variance $Var[e_t]$ = 528.6707

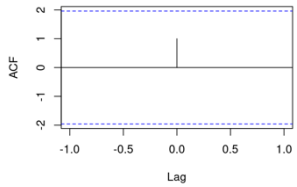

FIGURE 6:THE AUTOCOVARIANCE OF THE RESIDUALS = ZERO

As all the data points are near to regression line and mean is zero, we can say the data is statistically significant