(i) Obtaining the S&P500 Data

```
getSymbols("^GSPC", from = "2009-01-01", to = "2020-12-31",warnings =
FALSE, auto.assign = TRUE)
```

Figure 1: Time series plot

Figure 1 shows the plot for the time series of Adjusted Close of S&P 500 companies from the period '2009-01-01' to 2020-12-31'

(ii) Transforming time series to log-returns

```
lret = quantmod::periodReturn(sp500,period="daily", type="log")
plot(lret,main ="", lty = "solid")
title(main="S&P500 Returns",xlab="Time",ylab="Log Returns",)
```

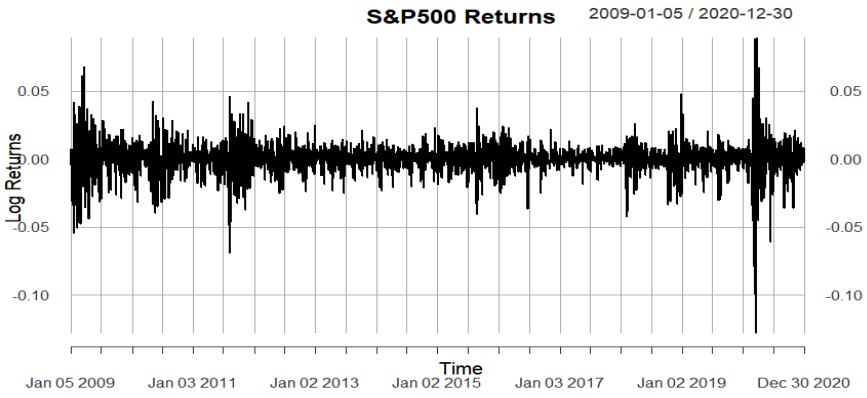

Figure 2: Log Returns for S&P 500

Figure 2 shows the returns for the S&P 500 consider the adjusted close price, the mathematical expression for computing the log returns is as follows:

$$r_t = \ln\left(\frac{S_t}{S_{t-1}}\right) = \ln(S_t) - \ln(S_{t-1})\text{-}$$

'S_t' and 'S_{t-1}' are the asset prices in time 't'

(iii) Examining ACF and PACF functions

```
acf(lret,main = "ACF")
pacf(lret,main = "PACF")
```

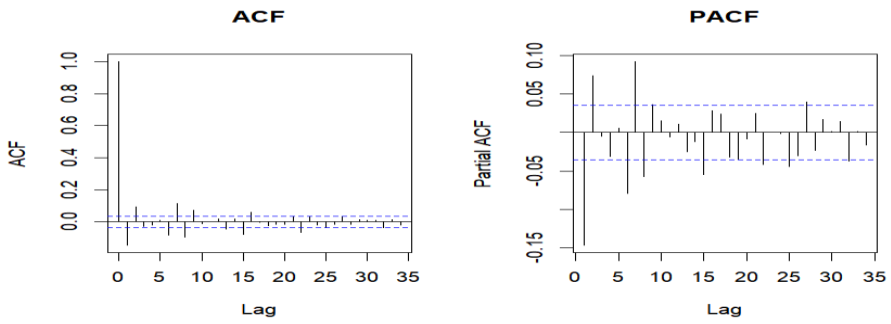

Figure 3: ACF and PACF

Pattern recognition and randomization testing are both aided by the autocorrelation analysis. This area of blue represents the 95% confidence interval. Any value within the blue region is therefore statistically extremely close to zero, but any value beyond the blue region is statistically not zero. In figure 3 we can see a significant number of lines that cross the threshold region for both ACF and the PACF, there +ve correlation at lags 2,7,9 and 16 and -ve autocorrelation at lags 1,6,13,15,22 and it can also be observed that the ACF plot decays slowly to zero. The PACF shows significant autocorrelation before lag 10. Therefore, we can say that the time series has autocorrelation.

(iv) Ljung-Box Test

```
Box.test(lret, type = "Ljung-Box")
```

Data Used	X-squared	df	p-value
Log returns	64.8	1	7.772e-16

Table 1: Ljung-Box Test Results

The Ljung-Box test can be used to check if the time series contain autocorrelation. The test makes uses of the below hypothesis:
H0: The time series data is independently distributed
HA: The time series data is not independently distributed i.e., they display autocorrelation.
From the test results in the table 1, we see that the p-value is quite low i.e., less than 0.05 and we can discard the null hypothesis and say that the time series has autocorrelation.

(v) Testing for Stationarity

```
adf.test(lret)
```

Data Used	Dickey-Fuller	Lag order	p-value
Log returns	-15.327	14	0.01

Table 2: ADF Test Results

According to the Augmented Dickey-Fuller (ADF) test a time series is deemed "stationary" if there is no trend, a constant variance over time, and a constant autocorrelation structure across time. The ADF test makes use of the following hypothesis:
H0: The time series data is non-stationary.
HA: The time series is stationary.
From the test results in table 2 we can infer from the p-value since it is less than 0.05 that we should consider the alternate hypothesis and reject the null hypothesis, therefore the time series is stationary. The ADF test does not test for a trend and hence we cannot say if the series is trend stationary or not.

Testing for Trend Stationarity:

```
tseries::kpss.test(lret,null ="Trend")
```

Data Used	KPSS Trend	Lag parameter	p-value
Log returns	0.017505	9	0.1

Table 3: KPSS Test Results

We use the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test to determine if the time series data is trend stationary or not i.e., if the time series has a changing mean and variance in time. The test makes use of the following hypothesis:
H0: The time series data is trend stationary.
HA: The time series data is not trend stationary.
According to the results obtained in the table 3 since the p-value is larger than 0.05 therefore we have to consider the H0 (null hypothesis) and we can infer the time series is trend stationary.

(vi) Testing for Normality

```
qqnorm(lret) qqline(lret)
```

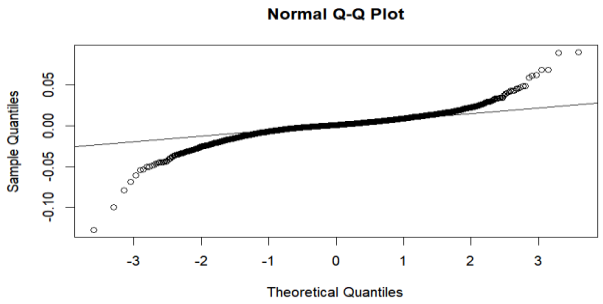

Figure 4: QQ plot for normality testing

In figure 4 we can see that the time series does not follow the normal line on either end of the curve. Hence, we can conclude that the time series is not normal.

```
shapiro.test(as.vector(lret))
```

Data Used	W	p-value
Log returns	0.87502	< 2.2e-16

Table 4: Shapiro-Wilk Test Results

Using Shapiro-Wilk test we can determine whether or not the time series data is regularly distributed. The Shapiro-Wilk test makes use of the following hypothesis:
H0: The data from the time series is normally distributed.
HA: The data from the time series is not normally distributed.
Table 4 shows us that the p-value is very small i.e., it is lower than 0.05 so we reject the null hypothesis(H0) and consider the alternative hypothesis and conclude that the time series is not normally distributed.

(vii) ARIMA Model

```
FitMod <- auto.arima(lret,seasonal = FALSE)
```

We get a model of ARIMA (4,0,4) with mean 0 from this we can concur that the time series is an AR (4) and MA (4) and has a **lag order of 0** since we are using the 'lret' (log returns) to fit a model.

Now we analyse the residuals.

```
checkresiduals(FitMod)
```

we get a p-value = 0.007488 for the Ljung-box test which is less than 0.05 for confirming any pattern in the data. So we explore further with model of different orders and get a model of ARIMA(6,0,2) with a p-value of 0.5599 and can be considered to be a good fit for our model with **lag order of 0**.

(viii) ARIMA Model Coefficient

```
NewFit1 =  arima(lret, order = c(6, 0, 2))
```

ar1	ar2	ar3	ar4	ma1	ma2	ma3	ma4	intercept

-1.3417	-0.6169	0.0081	0.0028	-0.0416	-0.0811	1.2229	0.5372	5e-04

Table 5: ARIMA model Coefficient

Removing the insignificant coefficients:

```
confint(NewFit1, level = 0.98)
```

By running the about function we get to know that the 99% confidence of ar3, ar4 and ar5 contains 0

	1 %	99%
ar3	-0.0689521535	0.085106371
ar4	-0.0754195122	0.080966671
ar5	-0.1153899115	0.032090571

Thus, we can conclude that these coefficients are insignificant.

Fitting the model once again with the new value obtained after the above test.

```
NewFit2 <- arima(lret, order = c(6, 0, 2), fixed =
c(NA,NA,0,0,0,NA,NA,NA,NA))
```

ar1	ar2	ar3	ar4	ar5	ar6	ma1	ma2
-1.3070	-0.5568	0	0	0	-0.0504	1.1918	0.4811

Table 6: refitted model results

From the table 7 we can derive the following equation for our model: -

$$Y_t = c - 1.3070y_{t-1} - 0.5568y_{t-2} - 0.0504y_{t-6} + \varepsilon_t + 1.1918\varepsilon_{t-1} + 0.4811\varepsilon_{t-2} \quad \text{----(2)}$$

(ix) Analysis of Residual (AIRMA (6,0,2))

```
checkresiduals(NewFit2)
```

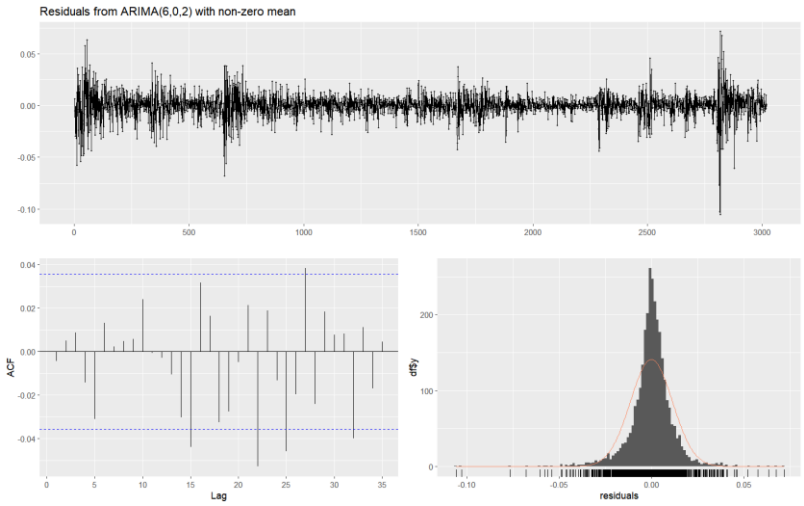

Figure 5: Plots of check residuals

From the figure 4 we can infer that the residuals have a mean 0 since they oscillate around zero and have a uniform variance. The density plot shows a possible normal distribution centred at mean of about zero (0). As for the ACF we can notice there are some points outside the confidence interval, thus indicating auto correlation. The Ljung-Box test give a significantly larger p-value indicating there is no more hidden information in the data. Even though this is not a great fit but is proves to be better than the model generated by auto.arima( ).