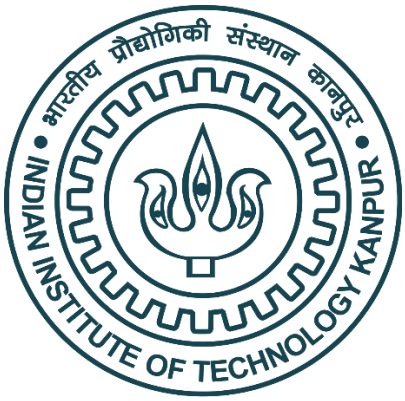
Analyzing and Forecasting using Goldman Sachs Stock Market Data

MTH517A- Course Project



DEPARTMENT OF MATHEMATICS AND STATISTICS

1. Rajdeep Saha (201380)
2. Sagnik Dey (201397)
3. Saumyadip Bhowmick (201408)
4. Shuvam Gupta (201421)
5. Soumik Karmakar (201428)

Course Project of **Time Series (MTH517A)**

SUPERVISED BY DR. AMIT MITRA

STUDENT'S DECLARATION

We, hereby declare that, the project work entitled "Analyzing and Forecasting using Goldman Sachs Stock Market Data" is a record of an original work done by us under the guidance of **Dr. Amit Mitra** and this project work is submitted in the partial fulfillment of the requirements for the paper **MTH517A- Time Series Analysis**.

We affirm that we have identified all our sources and that no part of our dissertation paper uses unacknowledged materials.

ACKNOWLEDGEMENT

We take this opportunity to express our sincere gratitude and deep regards to our project guide *Dr. Amit Mitra* for his exemplary guidance, monitoring and constant encouragement throughout the course of this project. Her blessings, help, time to time guidance will carry us a long way in the journey of life which we are about to embark.

We would also like to thank the institute authorities for giving us the chance to do the project and for providing the environment and necessary facilities required for the completion of our project. We would also like to thank PK Kelkar Library for having provided various reference books related to our project.

We would like to thank our department professors for teaching all the necessary topics with immense care which was needed to make the project fruitful.

Finally, we would like to extend our gratitude and thanks to our parents. Without their constant support and encouragement, it would not have been possible for us to proceed with our effort.

INDEX

1. Introduction:

A **stock market**, **equity market**, or **share market** is the aggregation of buyers and sellers of **stocks** (also called shares), which represent **ownership** claims on businesses; these may include securities listed on a public **stock exchange**, as well as stock that is only traded privately, such as shares of private companies which are sold to **investors** through **equity confounding** platforms. The stock market is very volatile and unpredictable, and a small geographical or socio-economical change can impact the share trends of stocks in stock market, recently we have seen how COVID-19 have impacted the stock market. The randomness and forever fluctuating market make investing a risky business. In order to get ahead of the market, statisticians have always tried to analyze and forecast the stock price that reflects all known and unknown information in the public domain.

The idea of this project is to analyze **Goldman Sachs Stock Market** data. The objective is to analyze Goldman Sachs closing price movement and possibly fit a suitable model to build forecasts. Close Price has been chosen because it is the price at which the market settles down at a given day and it acts as a reference price for the next day.

The data used here is the daily Goldman Sachs stock market data for 3 years, Jan 2018 to Dec 2020 and is collected from the open source of **Yahoo Finance**.

2. Test for Randomness- Turning Point Test:

We shall carry out our analysis with the **Univariate Data** Goldman Sachs stock closing price. We shall denote our data as $\{X_t\}$. Firstly, we shall check whether there is any deterministic component in the data.

To investigate this, we shall go for **"Turning Point Test"**. It is a non-parametric test procedure to check randomness in a time series.

To test,

H_0: The series is purely random against H_1: not H_0

Define,

U_i= 1 if X_i is a turning point

= 0 otherwise

Suppose, P is the total number of turning points, i.e., P= $\sum_{i=2}^{n-1} U_i$

An asymptotic test for H_0 is based on the test statistic,

$$Z = \frac{P - \frac{2}{3}(n-2)}{\sqrt{\frac{16n-29}{90}}}$$

Z asymptotically follows N(0,1) under H_0.

We would reject H_0 if $|Z| > \tau_{\frac{\alpha}{2}}$, where $\tau_{\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$% point of standard normal distribution.

Here, p-value<2.2e-16. So, we reject H_0 and conclude that, the series is not random.

3. Test for Trend- Relative Ordering Test:

Since, the null hypothesis of the Turning Point Test gets rejected, we can conclude that there is deterministic component in our data.

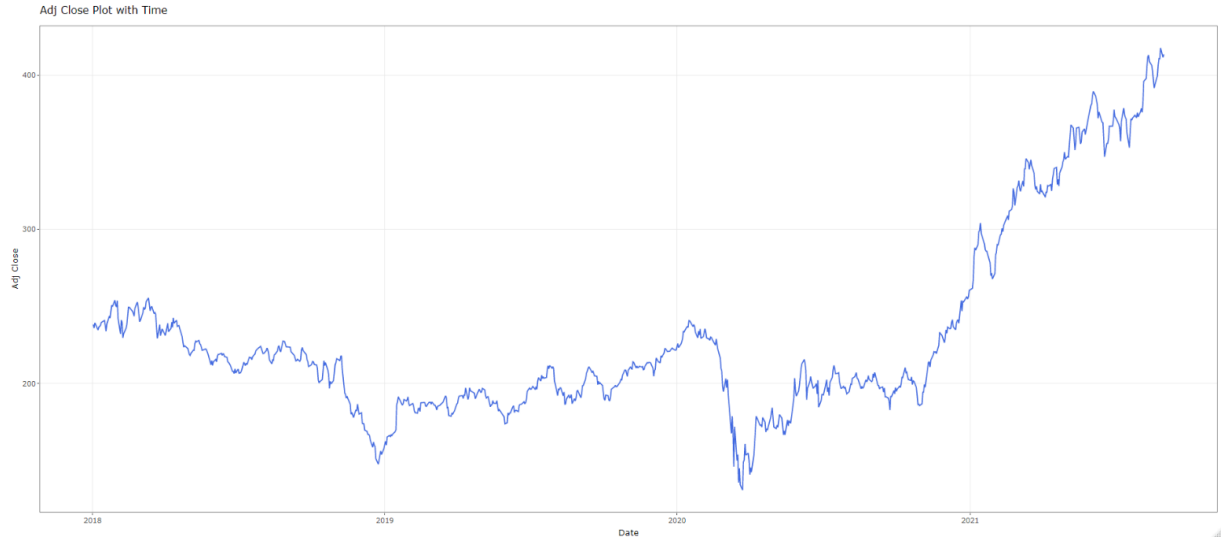


Fig 1: Original Data

From the graph, we can see that, there is an increasing trend in the data. To check the presence of trend, we perform a non- parametric test procedure named **Relative Ordering Test.**

To test,

H_0: There is no trend in the series against H_1: not H_0

Define,

q_{ij}= 1 if $X_i > X_j$

= 0 otherwise

$Q = \sum_{i,j\ i<j} q_{ij}$

Q is related to **Kendall's τ,** through the relationship,

$$\tau = 1 - \frac{4Q}{n(n-1)}$$

Under the null hypothesis of no trend, we can show, $E(\tau)=0$, $Var(\tau)= \frac{2(2n+5)}{9n(n-1)}$

An asymptotic test for H_0 is given by,

$$Z=\frac{\tau - E(\tau)}{\sqrt{Var(\tau)}}$$

We would reject H_0 if $|Z|>\tau_{\frac{\alpha}{2}}$, where $\tau_{\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$% point of standard normal distribution.
Here, the p-value is coming out to be 0(<0.05). So, we reject H_0 and conclude that, trend is present.

4. Trend Elimination- Differencing Method:

In order to remove trend, we applied the method of differencing. We applied a difference operator of lag 1 on our data. In most of the cases, the stock prices are assumed to be dependent on the immediate previous value.

After applying differencing of order 1, we have obtained the series given by,

$$Z_t = \nabla X_t = X_t - X_{t-1}$$

The following figure shows the series after applying differencing.

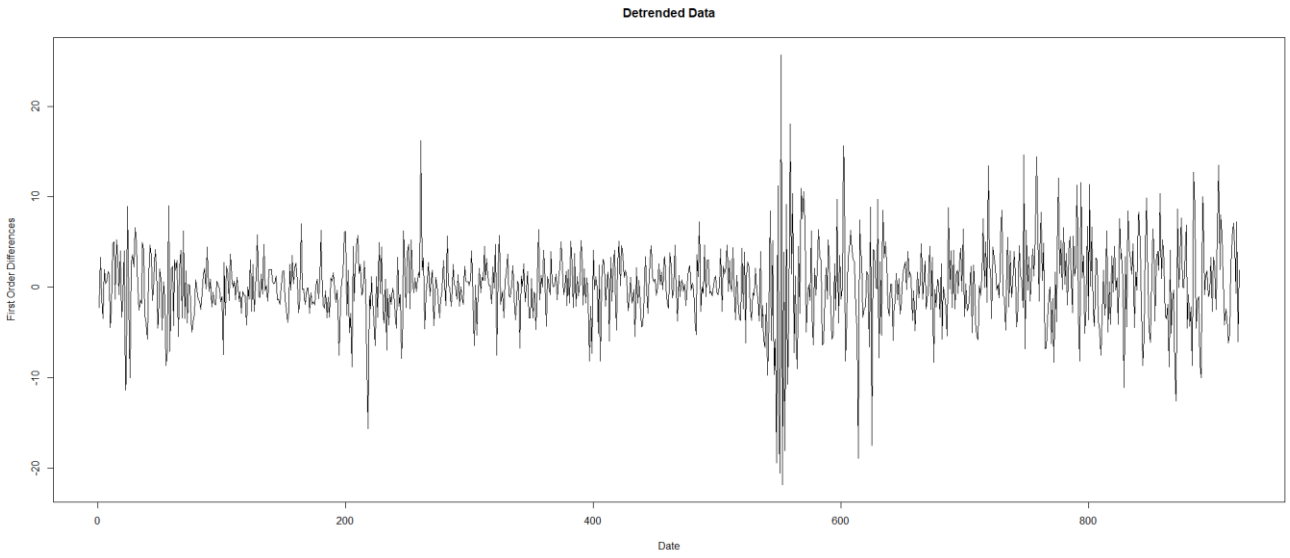


Fig 2: Detrended Data

From the graph we can conclude that, the trend has been removed after applying differencing. We again perform Relative Ordering Test on the data we have got after differencing.

We have obtained Z=2.28<$\tau_{0.025}$(=2.58). Hence, we can accept the null hypothesis of no trend and conclude that the trend is removed from our data.

5. Test for Stationarity- ADF Test:

Augmented Dickey–Fuller Test (ADF) tests the null hypothesis that a unit root is present in a time series. Here, the alternative hypothesis is that the series is stationary.

$$X_t = \phi_0 + \phi_1 X_{t-1} + \epsilon_t$$

To test,

H_0: $\phi_1 = 1$ (series is non- stationary) against, H_1: $\phi_1 \leq 1$ (series is stationary)

It is easy to see that,

$$X_t - X_{t-1} = \phi_0 + (\phi_1 - 1)X_{t-1} + \epsilon_t$$

$$\text{or, } \nabla X_t = \phi_0 + \phi^* X_{t-1} + \epsilon_t$$

Here, we compare the value of test statistic with the value of Dickey Fuller distribution. If the value of test statistic is less than the value of Dickey Fuller distribution, we reject the null hypothesis and conclude that the series is stationary. Here, we have got, p-value=0.01(<0.05), so we reject H_0 and conclude that the series is stationary.

6. ACF and PACF plots for deciding order of AR and MA:

Autocorrelation and partial autocorrelation plots are used to graphically summarize the strength of a relationship between an observation of a time series and observations at prior time steps. Plots of autocorrelation function (ACF) and partial autocorrelation function (PACF) give us different viewpoints of time series.

A partial autocorrelation plot is a summary of the relationship between an observation in a time series with observations at prior time steps with the relationships of intervening observations removed i.e. PACF only describes the direct relationship between an observation and its lag. This would suggest that there would be no correlation for lag values beyond k. While ACF describes the autocorrelation between an observation and another observation at a prior time step that includes direct and indirect dependence information.

We use the ACF and PACF plots to determine the order of AR and MA. For both the plots, we take that value of the lag, after which the values of ACF and PACF are not significantly

different from 0. From the ACF plot, we get we get the order of MA(q) i.e., the value of 'q' and from PACF plot, we get the order of AR(p), i.e., the value of 'p'.

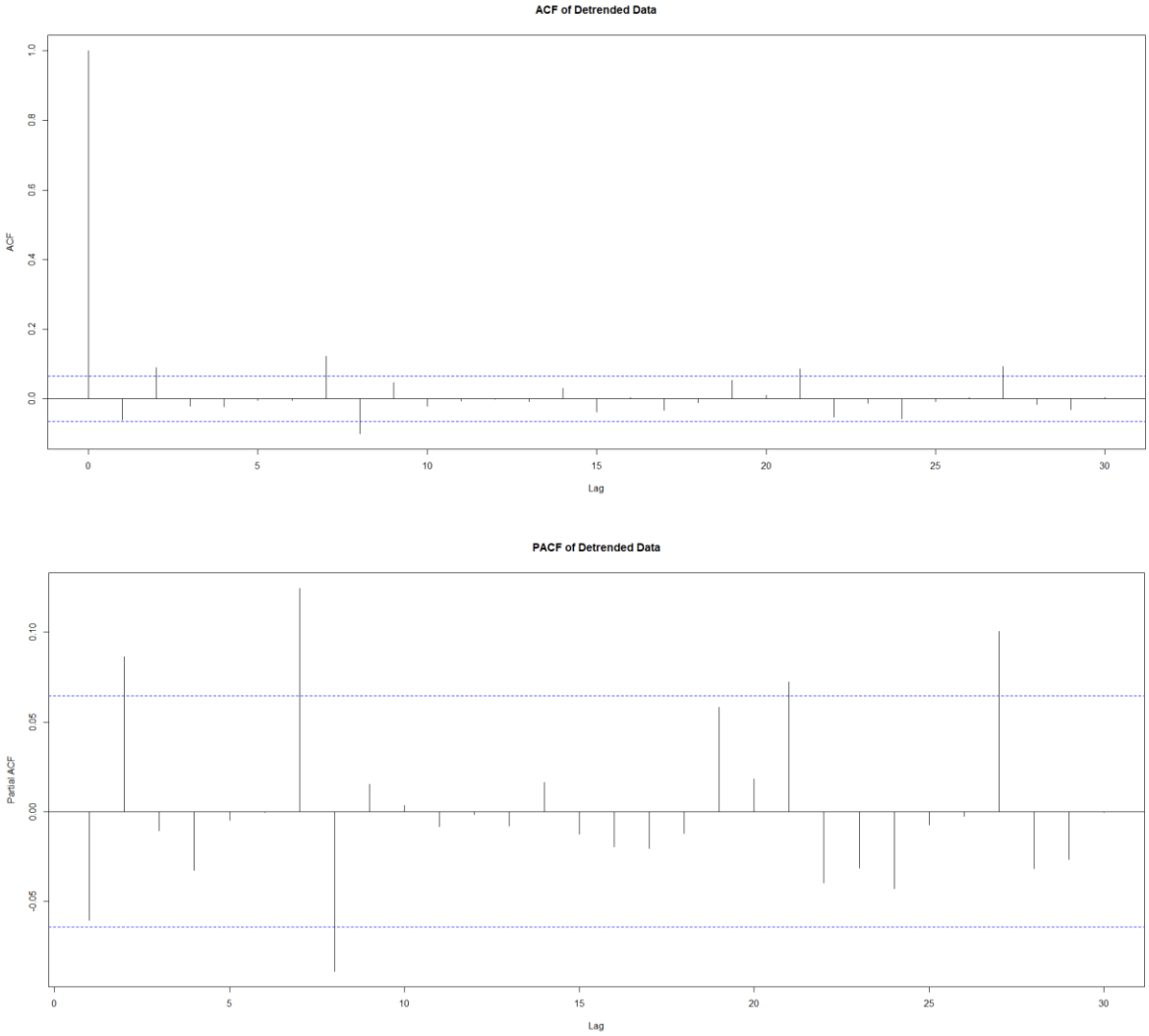


Fig 3&4: ACF and PACF plots of the data

From the graph, we can see that, in both the cases after lag 7, the values of ACF and PACF can be considered indifferent from 0. So, we consider all the 64 models with the 'p' and 'q' both less than or equal to 7. From these 64 models, we will choose that model, for which we get the minimum value of AIC (Akaike Information Criteria).

The results we have obtained shows that, the minimum value of AIC attains for ARMA(5,3) model. So, we choose the value of 'p' and 'q' as 5 and 3 respectively.

7. Model Selection for Original Data:

$\{X_t\}$ is said to follow an **Auto Regressive Integrated Moving Average (ARIMA)** of order (p, d, q) if,

$$Y_t = \nabla^d X_t = (1-B)^d X_t \sim ARMA(p,q)$$

where, 'd' is the smallest integer for which $\nabla^d X_t$ is stationary. From our analysis, d= 1. So, we can use **ARIMA(5,1,3) for fitting and forecasting procedure.**

8. Residual Analysis:

In our data, we have fitted, the ARIMA(5,1,3) model. The model is given by,

$$\widehat{x_t} = 0.0504x_{t-1} + 0.1893x_{t-2} - 0.9567x_{t-3} - 0.1112x_{t-4} + 0.0965x_{t-5} + \\ +\epsilon_t - 0.1099\epsilon_{t-1} - 0.0888\epsilon_{t-2} + 0.9822\epsilon_{t-3}$$

Residuals refer to the information in training or validation set that was not explained by the model fitted. Residual analysis is performed to check whether errors follow **White Noise process**. The auto-correlation function having values within the confidence interval of the corresponding estimates, will indicate that the residuals are uncorrelated.

The residual plot is as shown below,

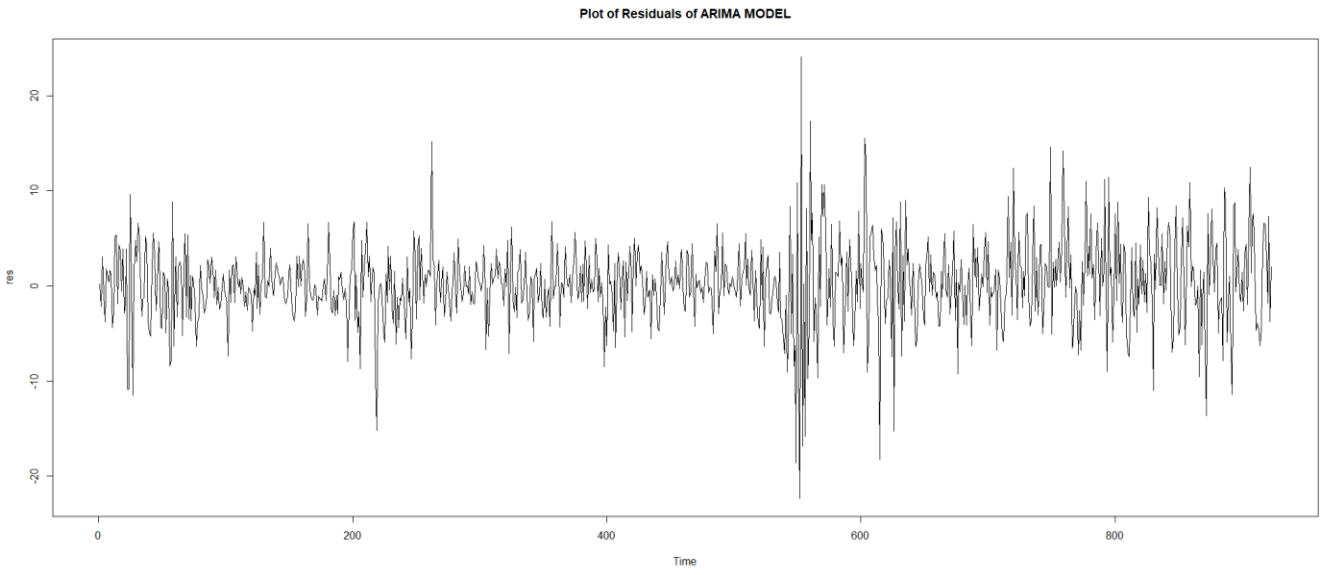


Fig 5: Residual Plot of the Data

We perform **Ljung-Box** test to check whether the errors are independently distributed.

To test, H_0: The data is independently distributed against H_1: not H_0

The test statistic is,

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\widehat{\rho_k}^2}{n-k}$$

The p-value of the test came out to be 0.9714 (≥ 0.05). Thus, we accept at 5% level of significance. Hence, we can conclude that the residuals are independent.

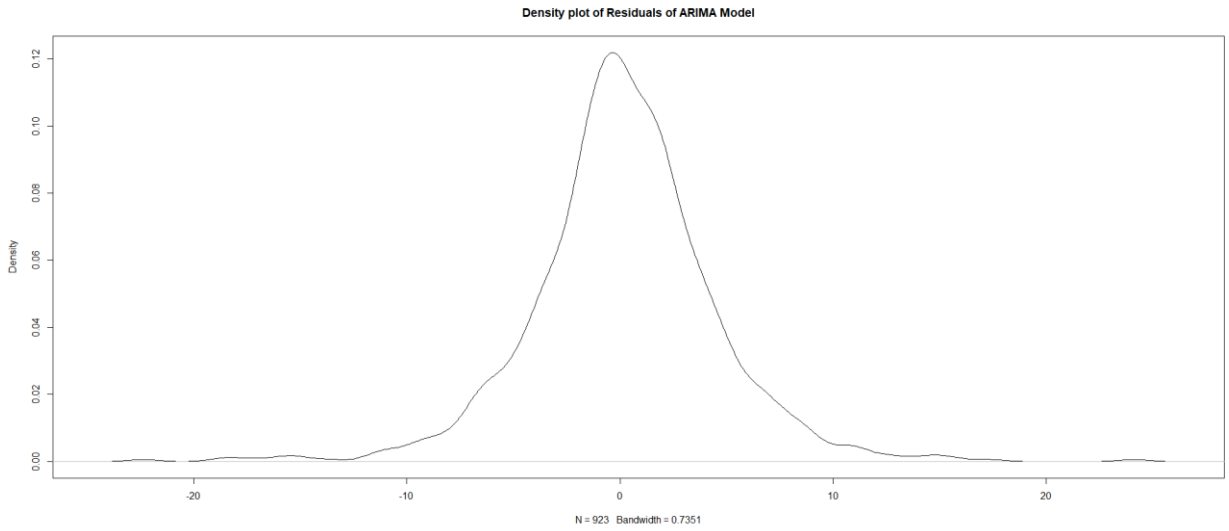


Fig 6: Density Plot of Residuals

From the density plot of the residuals, we observe that the residuals are concentrated at 0 which suggests that the mean of the residuals is close to zero i.e, we have,

$$E(\epsilon_t) = 0 \ \forall t$$

Now, we wish to check whether the residuals are homoscedastic, i.e.,

$$V(\epsilon_t) = constant \ \forall t$$

For this purpose, we wish to check whether ARCH effect is present in the model. The Autoregressive Conditional Heteroscedasticity (ARCH) model is a statistical model for time series data that describes the conditional variance of the current error term as a function of the previous lagged values of the residuals.

The detection of the ARCH effect in a time series is a test of serial independence. lied to the serially uncorrelated fitting error of some model in our case ARIMA model. We have assumed that linear serial dependence inside the original series is removed with an efficient model. Hence, any further serial dependence must be due to some nonlinear mechanism which has not been detected by the model. Here, the nonlinear mechanism we are concerned with is the conditional heteroscedasticity.

Lagrange Multiplier Test:

In this testing process,

H_0: ARCH-Effect is not present H_1: ARCH-Effect is present

This procedure simply involves obtaining the squares of the residual from fitted model and regress them on a constant and p lagged values, where is the ARCH lags. Let us consider the equation,

$$\epsilon_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 + v_t$$

The hypothesis is that, in the absence of ARCH components, we have α_i = 0 for all i=1, 2,…, p, against the alternative that, in the presence of ARCH components, at least one of the estimated α_i, must be significant.

We are seeing that p-value<2.2e-16. So, we reject null hypothesis at 5% level and conclude that ARCH effect is present in the model. First, we plot ACF and PACF of the square of the residuals.

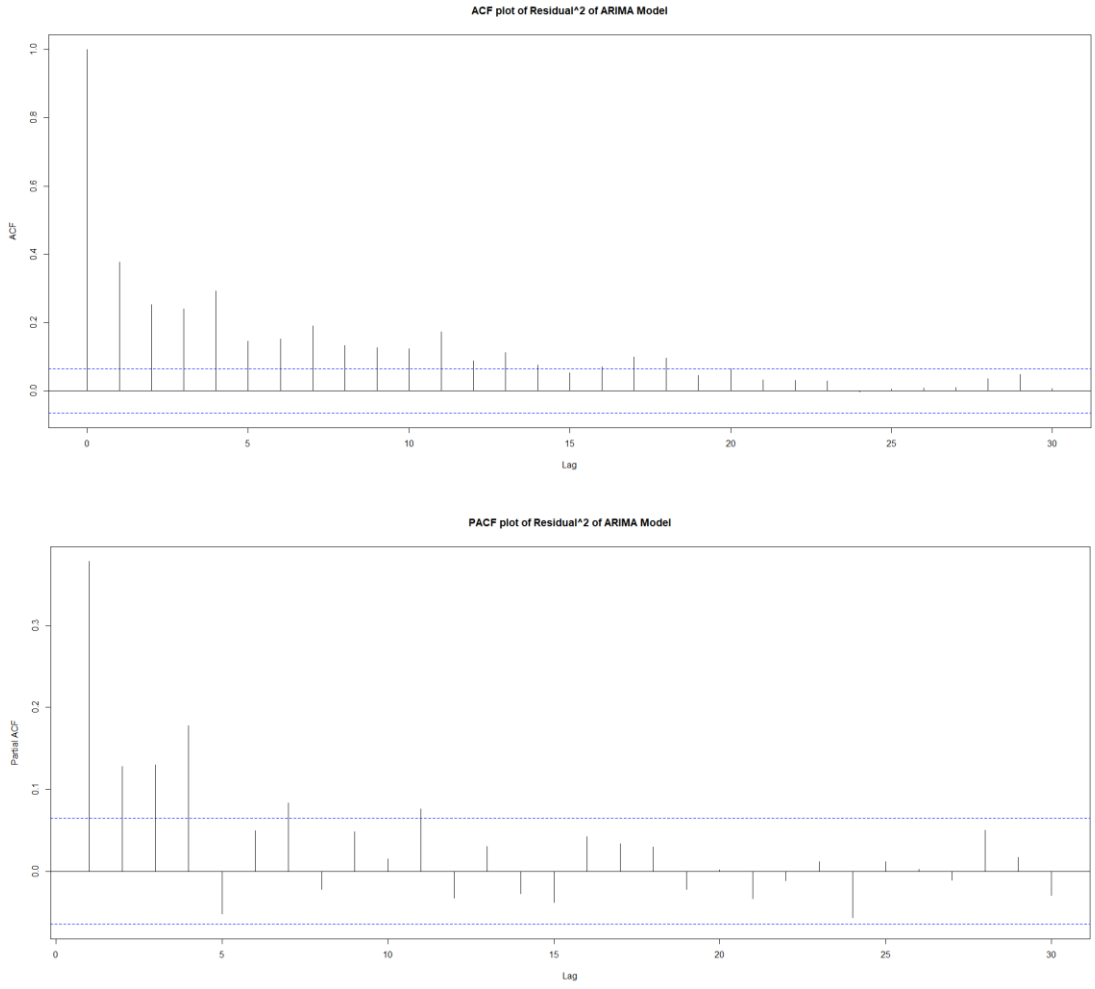


Fig 7&8: ACF and PACF Plots of the Square of the Residuals

From ACF plot, we can see an exponential decay. So, we can say that, the square of the residuals follows Auto Regressive process. From the PACF plot, we can say that, after lag 4, the PACF values are not significantly different from 0. So, we can conclude, that, the square of the residuals follow AR(4) process.

9. Forecasting:

Since, square of residuals follow AR(7) process, we can say, variance of the residuals follow ARCH(7). Using ARCH(7) process, we can estimate the variance of the residual. To obtain the estimated residual, we multiply the estimated variance of the residual with WN(0,1).

Then, the forecasted value of ARIMA-ARCH model is given by,

$$\widehat{x_t'} = \widehat{x_t} + \widehat{\epsilon_t}$$

where, $\widehat{\epsilon_t} = \widehat{\sigma_t} z_t$

$z_t \sim N(0,1)$

The series σ_t^2 was modeled by,

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 , \alpha_0 > 0, \alpha_i \geq 0, i > 0$$

using ARCH(7) process on residuals of the fitted ARIMA model.

Now that an appropriate model is fitted, our next objective is to check how efficiently this model can predict future values, given the knowledge of past observations. For building up the model, we have used data from 1st January 2018 to 31st August 2021 and based on the model we got, we have forecasted value for the next 20 days, i.e., 1st September 2021 to 29th September 2021.

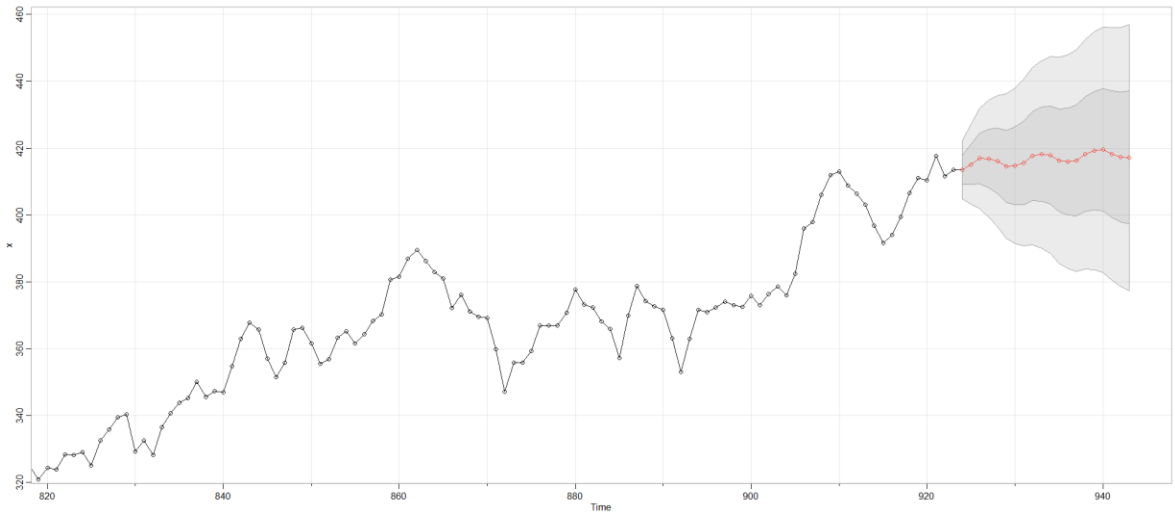


Fig 9: Prediction of Train and Test Data

Fig 10: Observed vs Predicted Plot

We obtain a **Mean Absolute Percentage Error (MAPE)** of 4.596 for this forecasting which is quite low.

10. Conclusion:

Summing up all the statistical analysis, we found that ARIMA(5,1,3) fits the best on our raw data but conditional variances of the residuals are heteroscedastic. To remove this volatility, we use ARCH(7) model and conclude that the ARIMA(5,1,3)-ARCH(7) model fits our data quite accurately. Forecasting for the future 20 days with the final ARIMA(5,1,3)-ARCH(7) model resulted predicted values close to the observed ones ,validating our concern of predicting the Goldman Sachs stock market prices.

11. Appendix:

- R Code:

```
rm(list=ls())


# Required libraries

library(ggplot2)
library(plotly)
library(tseries)
library(FinTS)
library(astsa)
library(rugarch)
library(spgs)

# Loading the data set
path <- 'https://raw.githubusercontent.com/s-karmakar16/Time-Series-
Project/main/GS_yahoo.csv'
raw.data <- read.csv(url(path))
data <- raw.data[raw.data$Date>= "2018-01-01"  & raw.data$Date
<="2021-08-31",]
data$Date <- as.Date.character(data$Date)


# Plotting train data
p <-ggplot(data,aes(x=Date, y=Adj.Close)) +
  geom_line(colour = "royalblue") + labs(x = 'Date', y = 'Adj Close',
title = "Adj Close Plot with Time") +
  theme_bw()
p <- ggplotly(p)
p


# Required Data
x <-  data$Adj.Close
n <- length(x)

#Turning Point Test to check if there is deterministic component
turningpoint.test(x)


#Relative Ordering Test Function
ro.test <- function (y = timeseries){
  n<-length(y)
  q<-0
```

```
  for(i in 1:(n-1))
  {
    for(j in (i+1):n)
    {
      if(y[i]>y[j])
      {
        q<-q+1
      }
    }
  }
  eq<-n*(n-1)/4
  tau<-1-(4*q/(n*(n-1)))
  var_tau<-(2*(2*n+5))/(9*n*(n-1))
  z<-tau/sqrt(var_tau)
  if(z>0)
  {
    p_value<-1-pnorm(z)
  }
  if(z<0)
  {
    p_value<-pnorm(z)
  }
  cat("            Relative Ordering Test for Presence of Trend
\n\n")
  cat("Null Hypothesis: Absence of Trend, and \n")
  cat("Alternative Hypothesis: Presence of Trend. \n\n")
  cat("Test Statistic:",paste(round(z,4)),"\n")
  cat("p_value:", paste(round(p_value,4)),"\n")
  cat("No. of Discordants:",paste(q),"\n")
  cat("Expected No. of Discordants:",paste(eq),"\n")
}



ro.test(x)


#lth order backward difference function
detrend <- function(x,l){
  k <- 1
  y <- array(0)
  for(i in (l+1):length(x)){
    y[k] <-  x[i] - x[i-l]
    k <- k+1
  }
  return(y)
}
```

```
detrended = detrend(x,1)
ro.test(detrended)
plot.ts(detrended , xlab = "Date", ylab = "First Order Differences",
main = "Detrended Data")



# Checking for Stationarity - Augmented Dickey Fuller test
adf.test(detrended)


# acf and pacf plot
acf(detrended, lag.max =30, main  = "ACF of Detrended Data")
pacf(detrended, lag.max =30, main = "PACF of Detrended Data")



# Order Estimation of ARIMA
best.order = c(0,0,0)
best.aic = Inf
for (q in 0:7) for (p in 0:7){
  fit.model = arima(detrended, order = c(p,0,q) ,optim.control =
list(maxit = 1000))
  fit.aic = fit.model$aic
  if(fit.aic < best.aic){
    best.order = c(p,1,q)
    best.aic = fit.aic
  }
}
best.order   # Best order ARIMA is at (5,1,3)
best.aic


# Fitting ARIMA
fit.model = arima(x, order = c(5,1,3) ,optim.control = list(maxit =
1000) )
res = fit.model$residuals
#acf(res, lag.max = 30, main = "ACF plot of Residuals of ARIMA
Model")
#pacf(res, lag.max =  30, main = "PACF plot of Residuals of ARIMA
Model")
plot(density(res), main = "Density plot of Residuals of ARIMA Model")
plot.ts(res, main = "Plot of Residuals of ARIMA MODEL")
```

```
# Checking for autocorelation of residuals - Ljung-Box Test
Box.test(res,lag=20,type='Ljung-Box')


# Estimating conditional volatility
acf(res^2, lag.max = 30, main = "ACF plot of Residual^2 of ARIMA
Model")
pacf(res^2, lag.max =  30, main = "PACF plot of Residual^2 of ARIMA
Model")

# Checking for ARCH effect in model
ArchTest(res,lag=7)


# Forecast using ARIMA
forecast1=sarima.for(x,n.ahead=20,5,1,3)


# Fitting ARCH(7)
model<-ugarchspec(variance.model = list(model ="sGARCH",garchOrder =
c(7, 0)), mean.model =list(armaOrder = c(0,0), include.mean =
FALSE),distribution.model = "norm")
fit1<-ugarchfit(model,res,out.sample = 100)
forecast2<-ugarchforecast(fit1, n.ahead = 20)

b<-forecast2@forecast$sigmaFor
a<-forecast1$pred
set.seed(12)
w<-rnorm(20)
predicted <- a+b*w
observed <- raw.data[raw.data$Date > "2021-08-31",]$Adj.Close[1:20]


#Plot of Observed vs Predicted data
cols = c("Predicted" = "red", "Observed" = "royalblue")
df.new <- data.frame(predicted, observed)
ggplot(df.new, aes(x = 1:20)) + geom_line(aes(y = predicted, col =
"Predicted")) +
  geom_line(aes(y = observed, col = "Observed")) + theme_bw() +
  labs(title = "Predicted vs Oberved Adj Close", x = "Time", y = "Adj
Close", color = "Legend") +
  scale_color_manual(values = cols)



#Measures of Prediction Error Rate
```

```
mape = mean(abs((predicted - observed)/observed))*100
mape

#MAPE is pretty low. Hence we conclude that our Fitted model is quite
good.
```

12. References:

1. https://finance.yahoo.com/quote/GS/history/
2. Davis, Brockwell(1996), Introduction to Time Series and Forecasting
3. https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average
4. https://en.wikipedia.org/wiki/Autoregressive_conditional_heteroskedasticity
5. Arch Models: Properties, Estimation and Testing by Anil K. Bara