Project1

Brandon Cunningham

2024-06-18

Data Importing and Indexing

```
data_start_ind <- 1
data_end_ind <- 1622
forecast_stary_ind <- 1623
forecast_end_ind <- 1722

path <- paste(getwd(), '/Data Set for Class.xls', sep="")
sheet_name <- 'S03'

# Read the specified sheet from the Excel file
s03 <- read_excel(path, sheet = sheet_name)
```

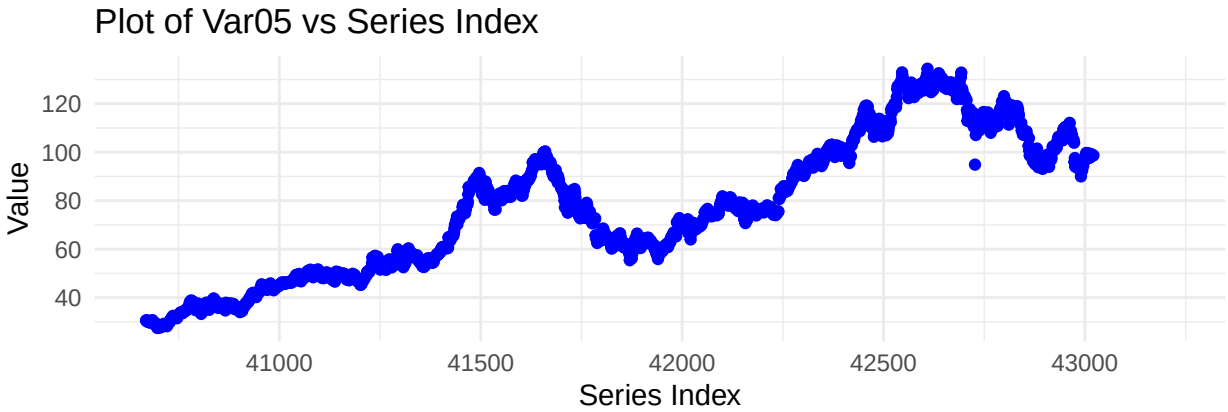
Data Visualization

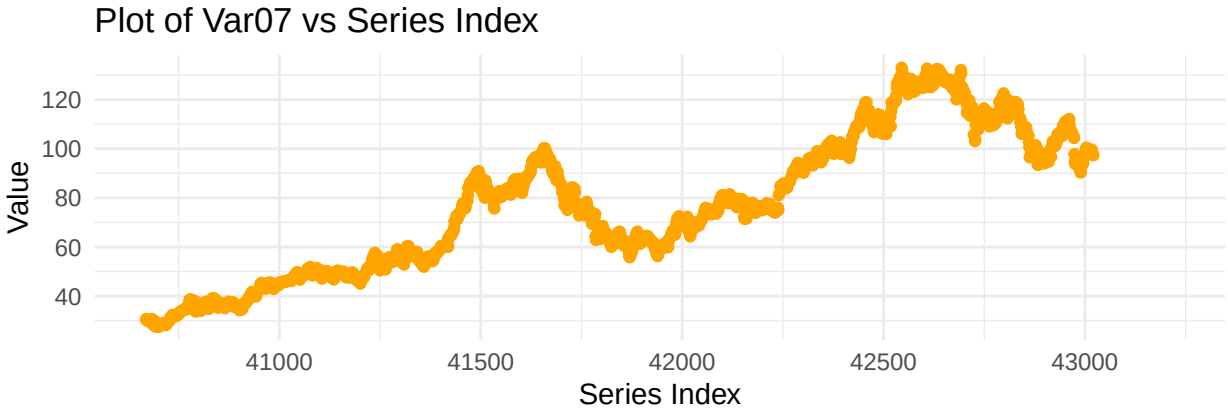
```
var5_plot <- ggplot(s03, aes(x = SeriesInd, y = Var05)) +
  geom_point(color = "blue") +
  labs(title = "Plot of Var05 vs Series Index", x = "Series Index", y = "Value") +
  theme_minimal()
```

```
var7_plot <- ggplot(s03, aes(x = SeriesInd, y = Var07)) +
  geom_point(color = "orange") +
  labs(title = "Plot of Var07 vs Series Index", x = "Series Index", y = "Value") +
  theme_minimal()
```

```
grid.arrange(var5_plot, var7_plot, nrow = 2)
```

```
## Warning: Removed 144 rows containing missing values or values outside the scale range
## (`geom_point()`).
## Removed 144 rows containing missing values or values outside the scale range
## (`geom_point()`).
```





Data Imputation

I'm using linear imputation, so creating a line of best fit between the last two known points and filling in missing values along that line

```
data_range <- which(s03$SeriesInd < 43022)
na_var5 <- which(is.na(s03$Var05[data_range]))
na_var7 <- which(is.na(s03$Var07[data_range]))
```

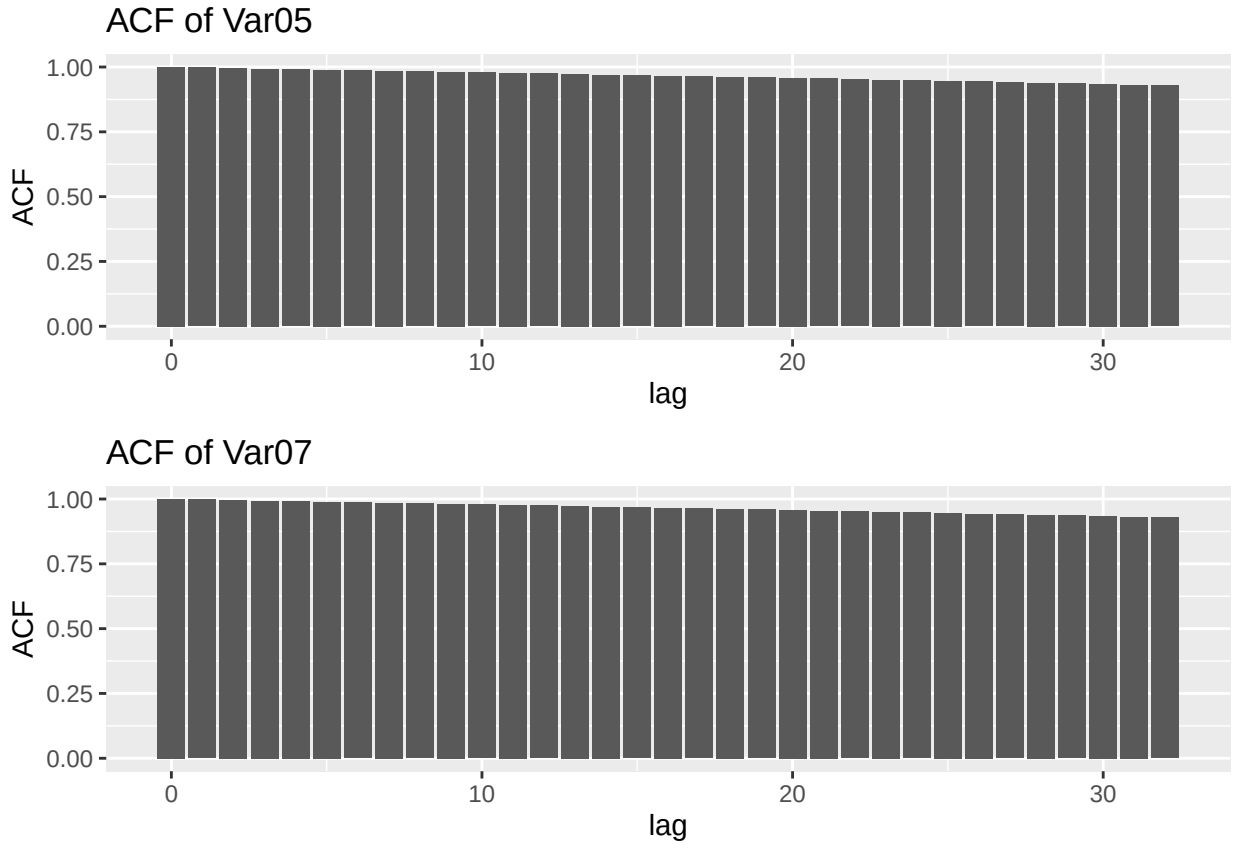
```
imputed_var5 <- approx(x = s03$SeriesInd[data_range], y = s03$Var05[data_range],
                       xout = s03$SeriesInd[data_range])$y
imputed_var7 <- approx(x = s03$SeriesInd[data_range], y = s03$Var07[data_range],
                       xout = s03$SeriesInd[data_range])$y
```

```
s03$Var05[data_range][na_var5] <- imputed_var5[na_var5]
s03$Var07[data_range][na_var7] <- imputed_var7[na_var7]
```

Values to forecast: 43022 - 43221 index numbers: 1623 - 1762

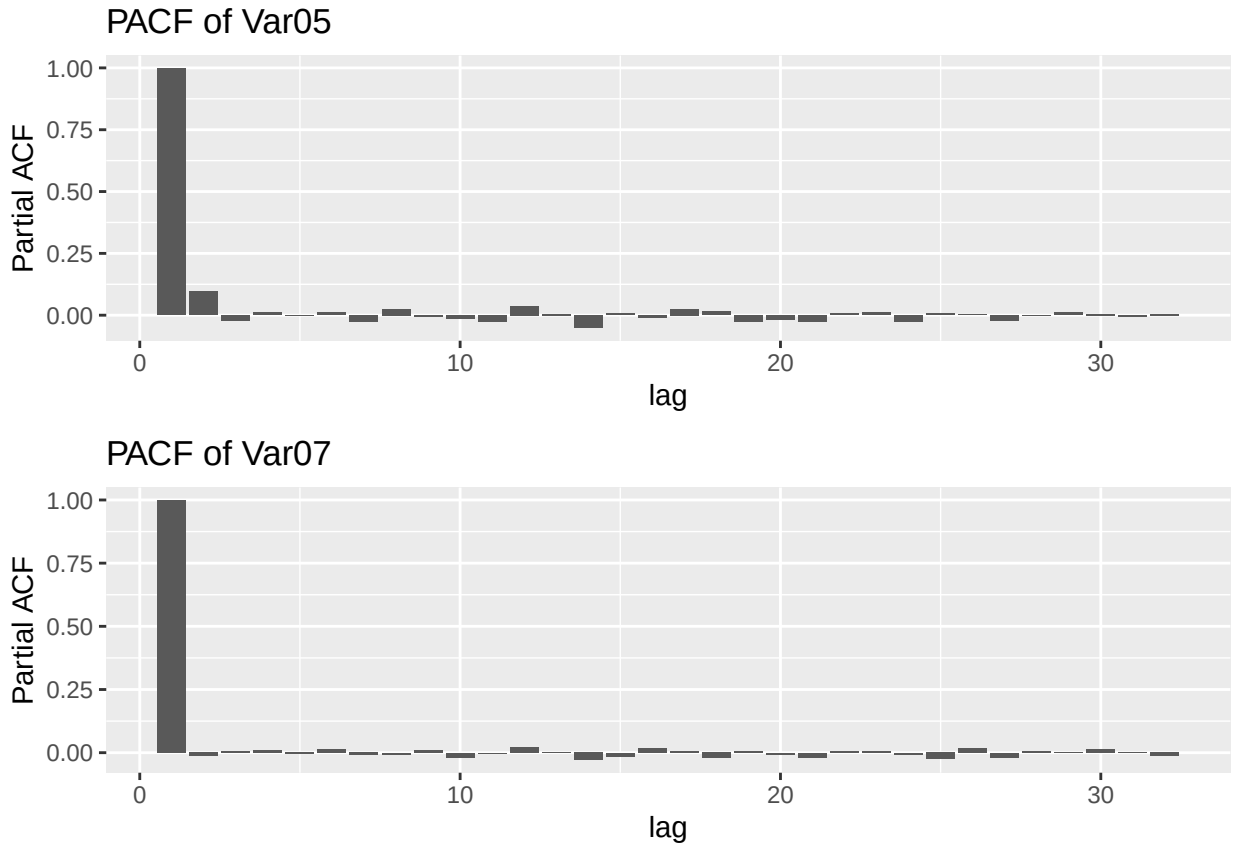
Checking for Stationarity

```
acf_var5 <- acf(s03$Var05[data_range], plot = FALSE)
acf_var7 <- acf(s03$Var07[data_range], plot = FALSE)

acf_var5_df <- data.frame(lag = acf_var5$lag, acf = acf_var5$acf)
acf_var7_df <- data.frame(lag = acf_var7$lag, acf = acf_var7$acf)

acf1 <- ggplot(acf_var5_df, aes(x = lag, y = acf)) +
  geom_bar(stat = "identity") +
  labs(title = "ACF of Var05", y = 'ACF')

acf2 <- ggplot(acf_var7_df, aes(x = lag, y = acf)) +
  geom_bar(stat = "identity") +
  labs(title = "ACF of Var07", y = 'ACF')

grid.arrange(acf1, acf2, nrow=2)
```




```
pacf_var5 <- pacf(s03$Var05[data_range], plot = FALSE)
pacf_var7 <- pacf(s03$Var07[data_range], plot = FALSE)

pacf_var5_df <- data.frame(lag = pacf_var5$lag, pacf = pacf_var5$acf)
pacf_var7_df <- data.frame(lag = pacf_var7$lag, pacf = pacf_var7$acf)
```

```
pacf1 <- ggplot(pacf_var5_df, aes(x = lag, y = pacf)) +
  geom_bar(stat = "identity") +
  labs(title = "PACF of Var05", y = 'Partial ACF')

pacf2 <- ggplot(pacf_var7_df, aes(x = lag, y = pacf)) +
  geom_bar(stat = "identity") +
  labs(title = "PACF of Var07", y = 'Partial ACF')

grid.arrange(pacf1, pacf2, nrow=2)
```



The data is non-stationary. We will preforming differencing to make the data stationary.

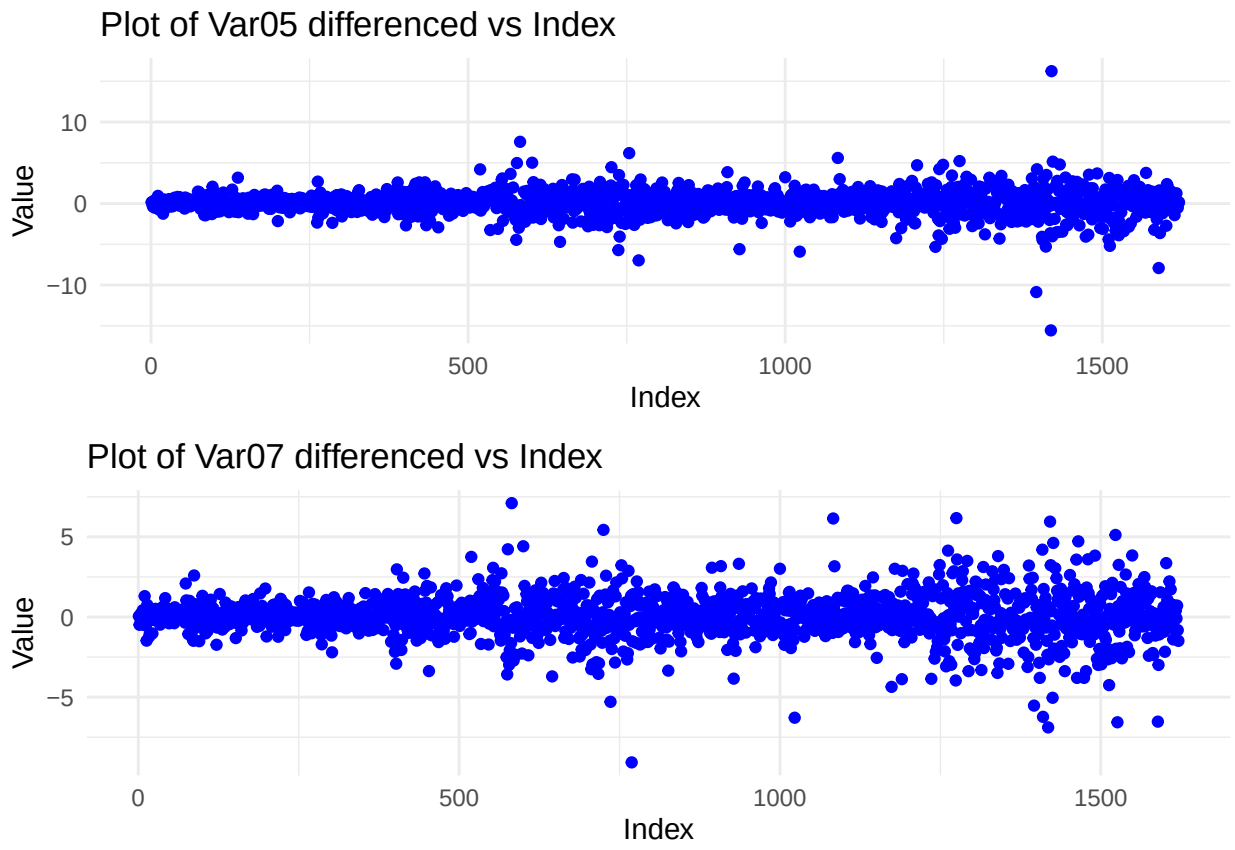
```
var5_diff <- diff(s03$Var05[data_range], differences = 1)
var7_diff <- diff(s03$Var07[data_range], differences = 1)

var5_diff_df <- data.frame(Index = seq_along(var5_diff), Value = var5_diff)
var7_diff_df <- data.frame(Index = seq_along(var7_diff), Value = var7_diff)
```

```
var5_plot <- ggplot(var5_diff_df, aes(x = Index, y = Value)) +
  geom_point(color = "blue") +
  labs(title = "Plot of Var05 differenced vs Index", x = "Index", y = "Value") +
  theme_minimal()
```

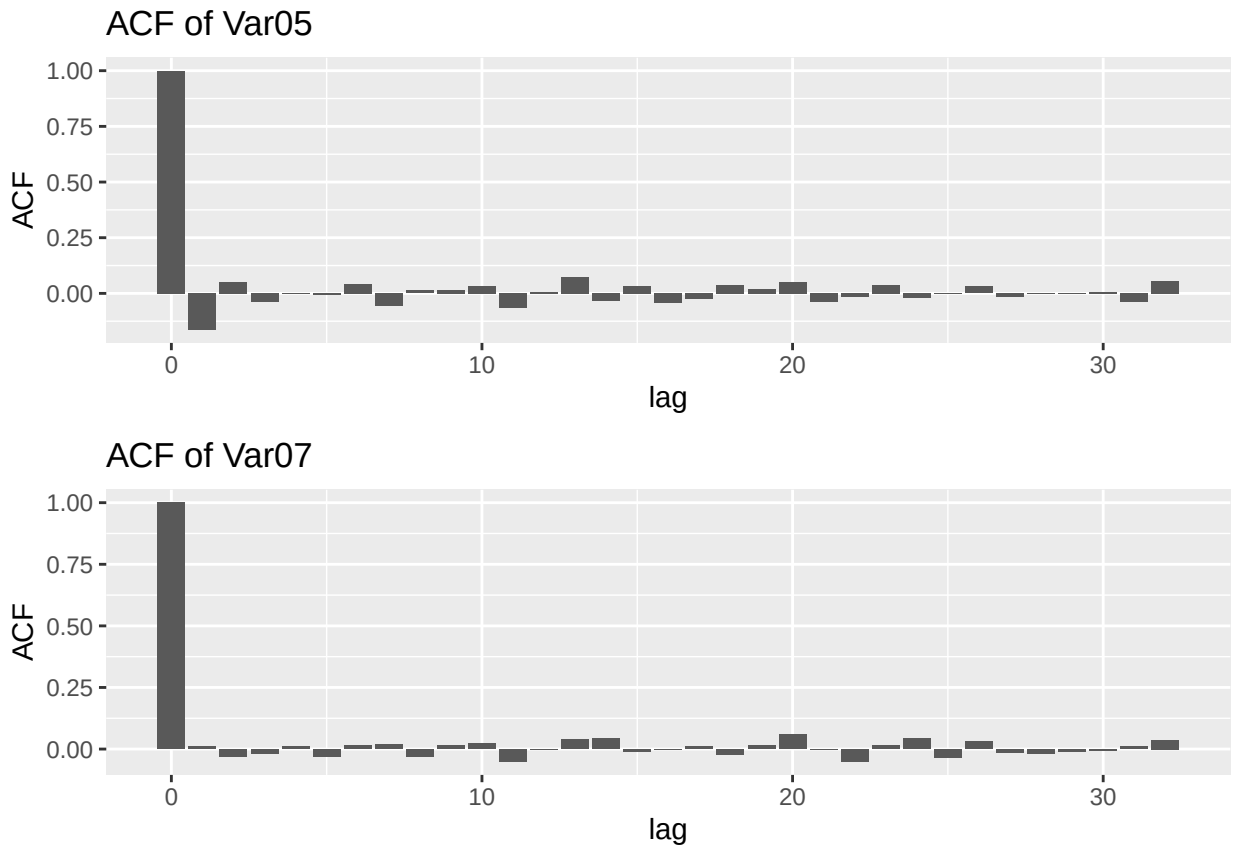
```
var7_plot <- ggplot(var7_diff_df, aes(x = Index, y = Value)) +
  geom_point(color = "blue") +
```

```
  labs(title = "Plot of Var07 differenced vs Index", x = "Index", y = "Value") +
  theme_minimal()
```

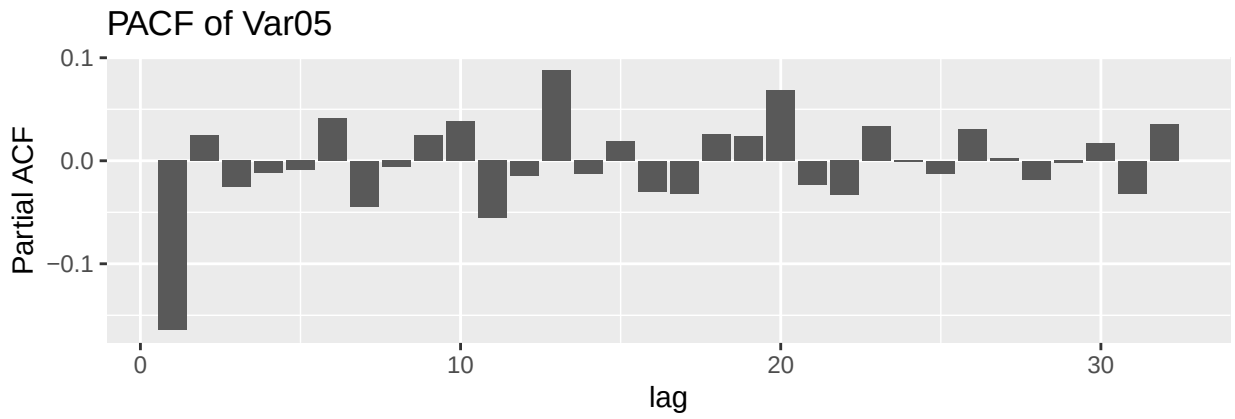
```
grid.arrange(var5_plot, var7_plot, nrow = 2)
```

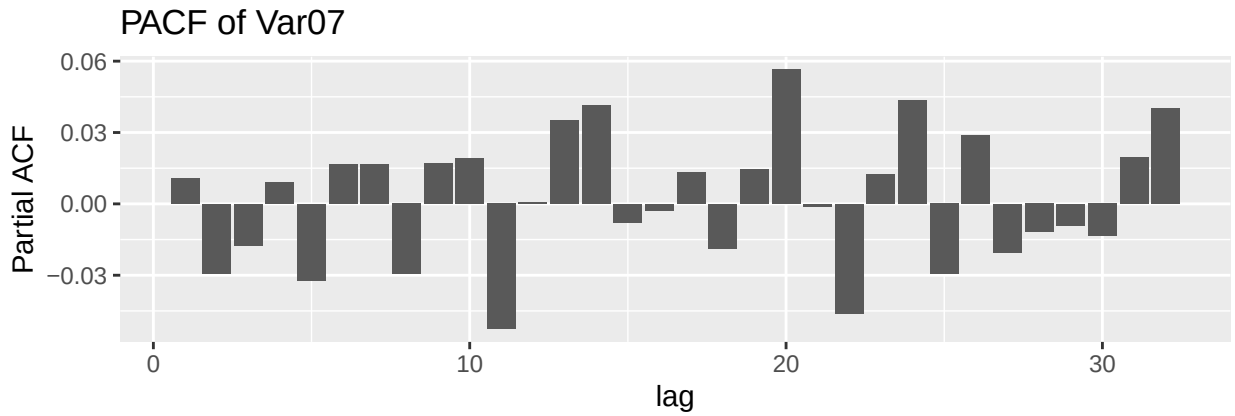



```
acf_var5 <- acf(var5_diff, plot = FALSE)
acf_var7 <- acf(var7_diff, plot = FALSE)

acf_var5_df <- data.frame(lag = acf_var5$lag, acf = acf_var5$acf)
acf_var7_df <- data.frame(lag = acf_var7$lag, acf = acf_var7$acf)

acf1 <- ggplot(acf_var5_df, aes(x = lag, y = acf)) +
  geom_bar(stat = "identity") +
  labs(title = "ACF of Var05", y = 'ACF')

acf2 <- ggplot(acf_var7_df, aes(x = lag, y = acf)) +
  geom_bar(stat = "identity") +
  labs(title = "ACF of Var07", y = 'ACF')

grid.arrange(acf1, acf2, nrow=2)
```




```
pacf_var5 <- pacf(var5_diff, plot = FALSE)
pacf_var7 <- pacf(var7_diff, plot = FALSE)

pacf_var5_df <- data.frame(lag = pacf_var5$lag, pacf = pacf_var5$acf)
pacf_var7_df <- data.frame(lag = pacf_var7$lag, pacf = pacf_var7$acf)

pacf1 <- ggplot(pacf_var5_df, aes(x = lag, y = pacf)) +
  geom_bar(stat = "identity") +
  labs(title = "PACF of Var05", y = 'Partial ACF')

pacf2 <- ggplot(pacf_var7_df, aes(x = lag, y = pacf)) +
  geom_bar(stat = "identity") +
  labs(title = "PACF of Var07", y = 'Partial ACF')

grid.arrange(pacf1, pacf2, nrow=2)
```

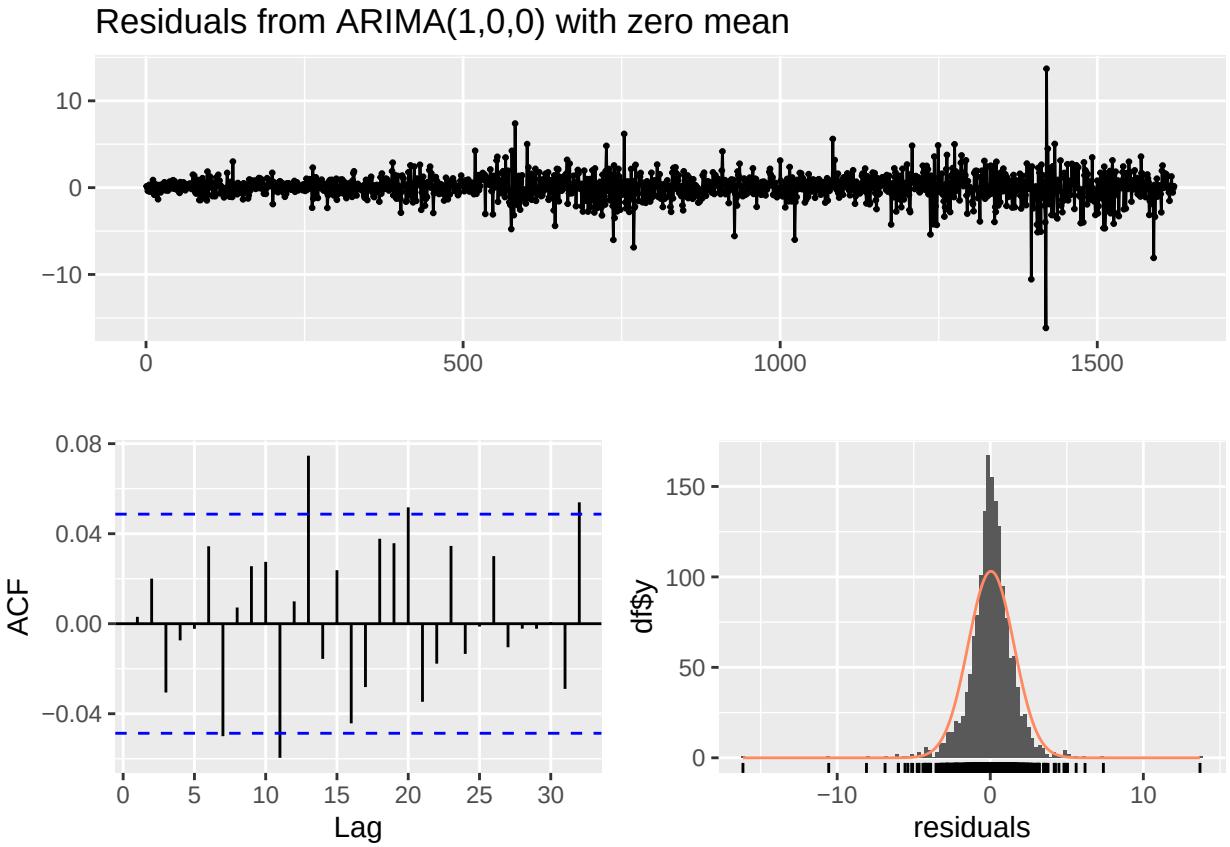





```
fit_var5 <- auto.arima(var5_diff, stationary = TRUE)
summary(fit_var5)
```

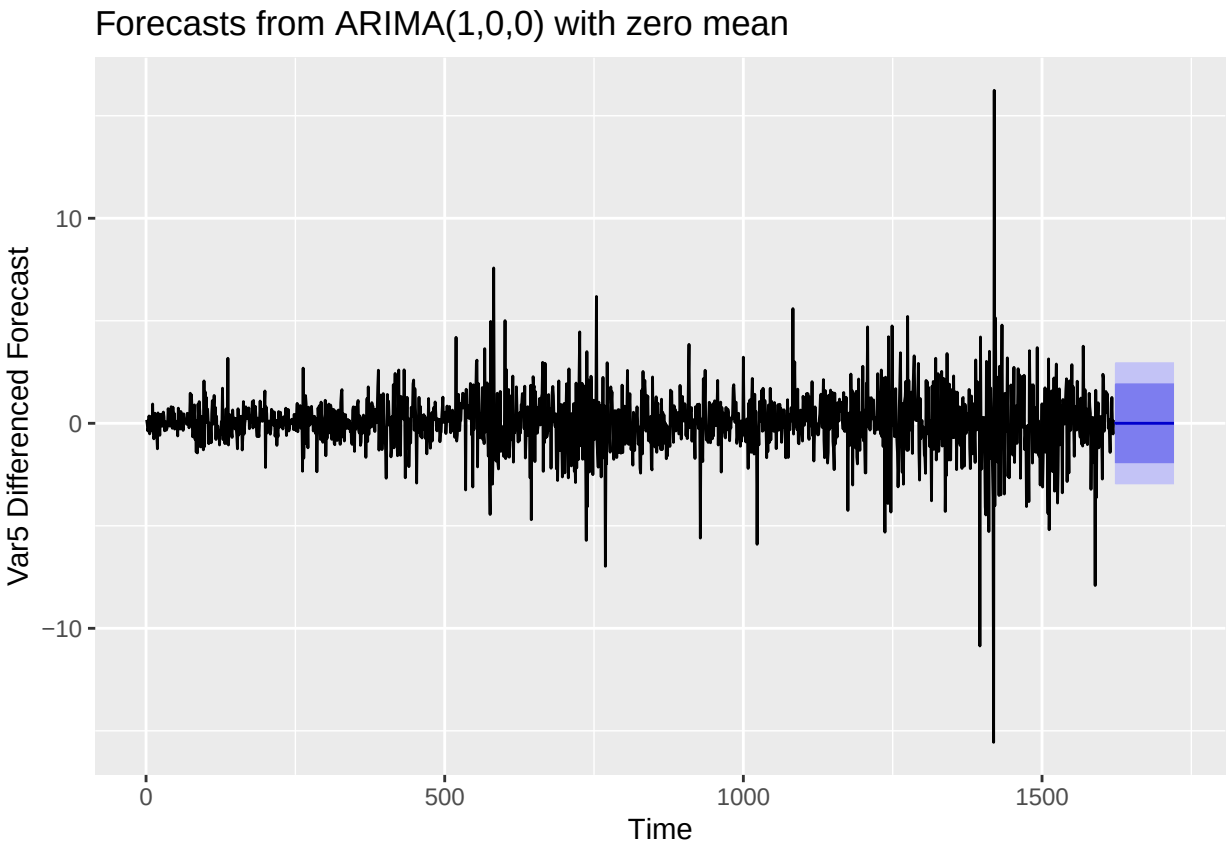
```
## Series: var5_diff
## ARIMA(1,0,0) with zero mean
##
## Coefficients:
##           ar1
##       -0.1632
## s.e.   0.0245
##
## sigma^2 = 2.246:  log likelihood = -2955.41
## AIC=5914.82   AICc=5914.83   BIC=5925.6
##
## Training set error measures:
##                       ME     RMSE      MAE MPE MAPE      MASE        ACF1
## Training set 0.04892235 1.498185 1.004552 NaN  Inf 0.6758614 0.003046622
```

```
checkresiduals(fit_var5)
```

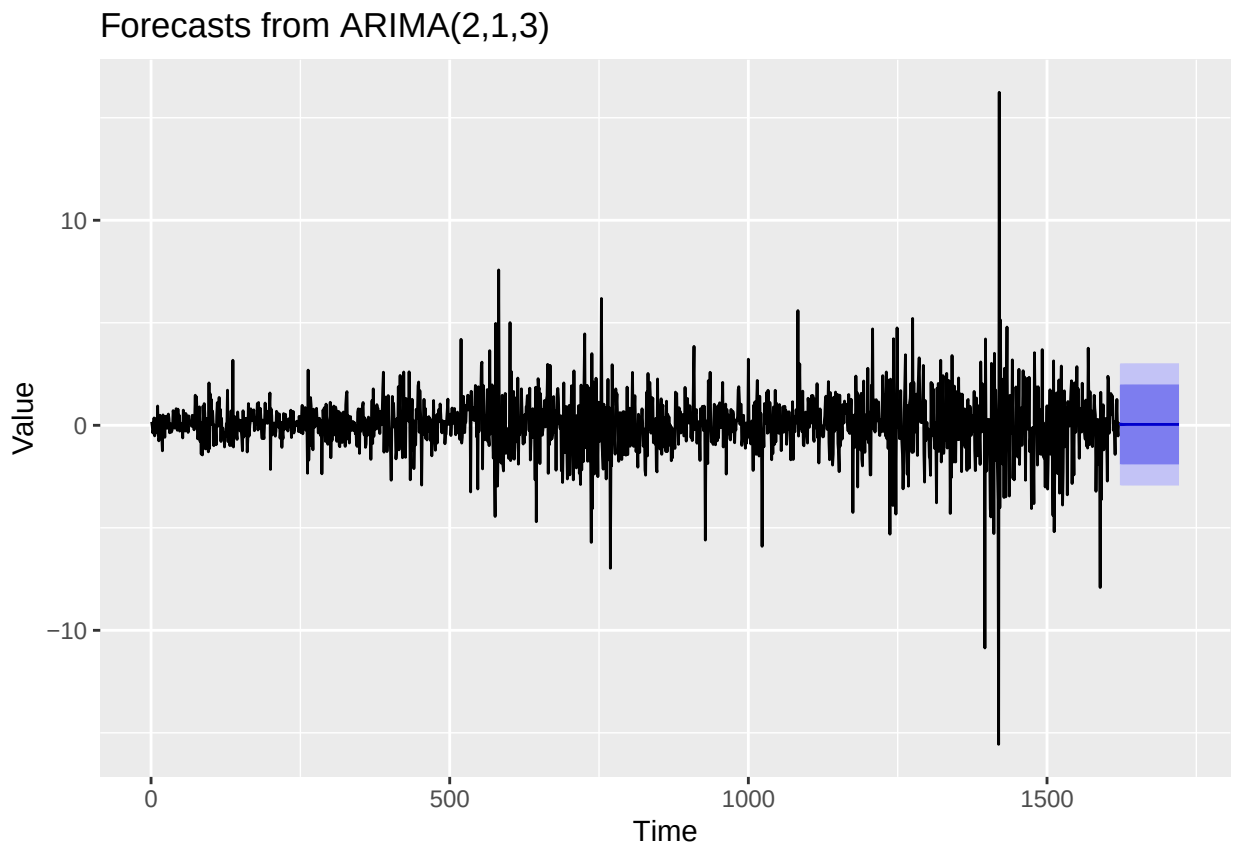



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,0,0) with zero mean
## Q* = 10.671, df = 9, p-value = 0.2989
##
## Model df: 1.   Total lags used: 10
```

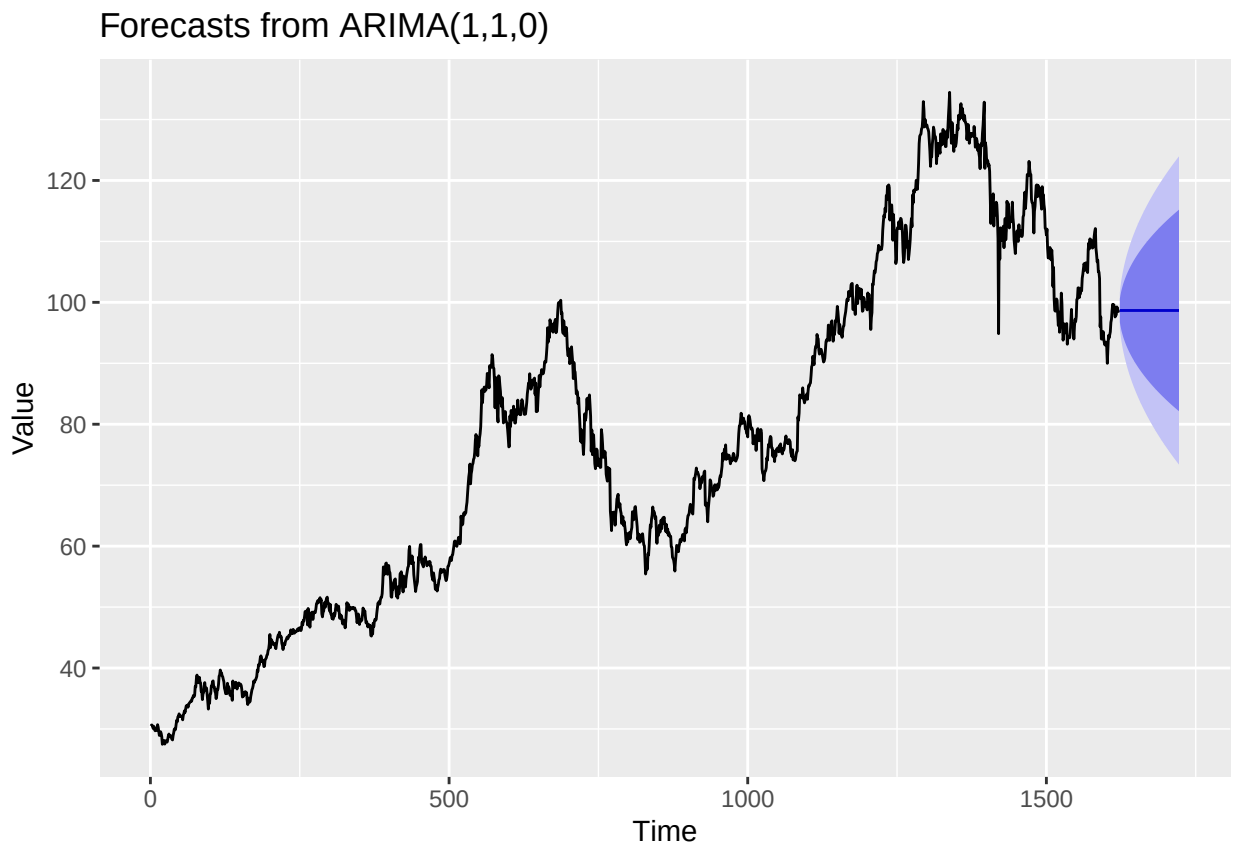
```
fc_var5 <- forecast(fit_var5, h=100)
autoplot(fc_var5) + ylab('Var5 Differenced Forecast')
```



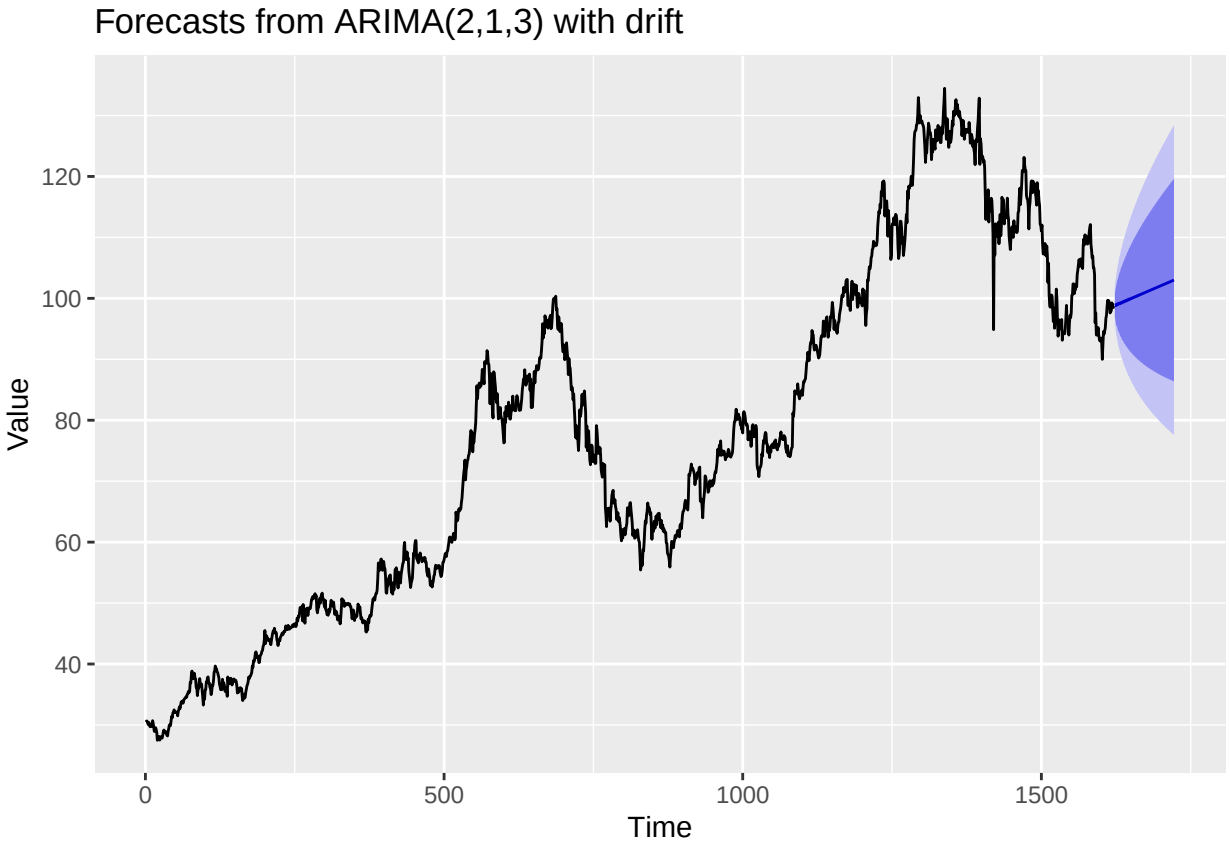

```
fit <- Arima(var5_diff, order=c(2,1,3), include.constant=FALSE)
fc <- forecast(fit, h=100)
autoplot(fc) + ylab('Value')
```




```
fit <- auto.arima(s03$Var05[data_range])
fc <- forecast(fit, h=100)
autoplot(fc) + ylab('Value')
```




```
fit <- Arima(s03$Var05[data_range], order=c(2,1,3), include.drift=TRUE)
fc <- forecast(fit, h=100)
autoplot(fc) + ylab('Value')
```


Forecasts from ARIMA(2,1,3) with drift
Value
120
100
80
60
40
0
500
1000
1500
Time