# Technical Report

## Time Series Analysis of Beer Sales

Douglas Fedorczyk

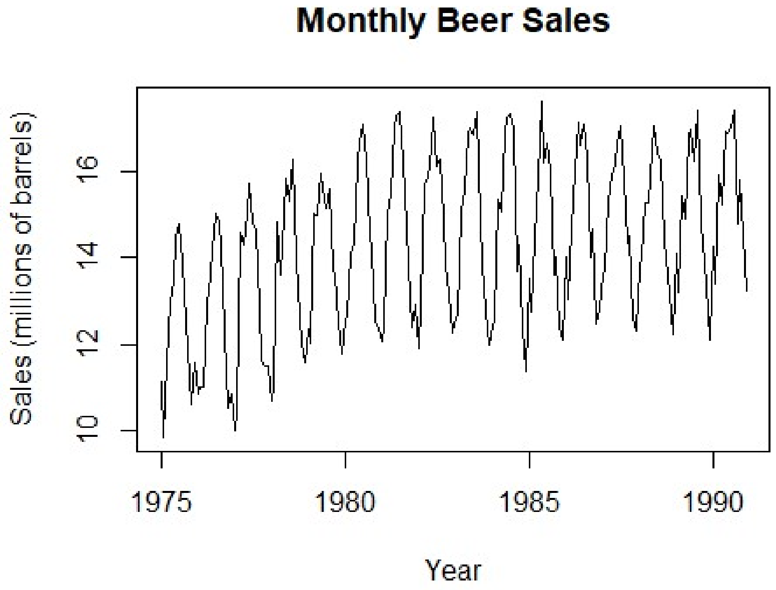
5/16/2021

## Introduction

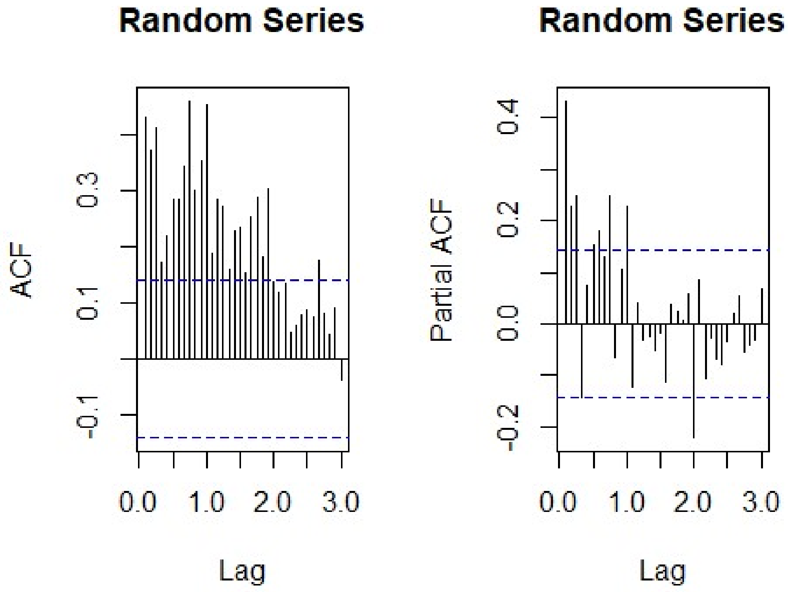
Beer sales data from 1975 to 1990 is provided as part of the TSA package available in R. It provides monthly beer sales in the form of millions of barrels sold. It is a time series data set that can be easily broken down into trend, seasonality, and random components as part of a model fitting process. This report goes through such a process to determine what form the best fitting model will have for the beer sales dataset. It begins by looking at simpler models such as an Autoregressive Moving Average (ARMA) before moving into slightly more complicated models such as Seasonal Autoregressive Integrated Moving Average (SARIMA) models. Additionally, it will also look at comparing two different methods in forecasting beer sales over the course of two years. The first includes using the resulting best fitting model and second includes using the Holt-Winters' (HW) smoothing method.

## Method & Results

Below is a time plot of the beersales dataset. It shows a distinct seasonal effect for beer sales with an increasing trend prior to 1980, before leveling off. It is suggestive of a seasonal type of model, which is indicative of a SARIMA type model for fitting the data. At the same time, the variation in the data remains stable, which is important when looking at the HW method as part of a forecasting model.



In deciding which method to use for the analysis, i.e., an ARMA or ARIMA model, an indicator of which model to use is based on the stationarity of the data. To get the time series to a stationary state the trend and seasonality were removed. This was done first by fitting a linear model to determine the overall trend. The resulting residuals represent the data with the linear trend removed. From there another linear model was fit and the resulting residuals, which represented the removal of the seasonality, provided the stationary state of the data necessary for determining an appropriate ARMA or ARIMA model. Beginning with an ARMA model, the autocorrelation function (ACF) and partial autocorrelation function (PACF) were plotted to help inform the orders of p and q for the AR and MA parts of the model, respectively.

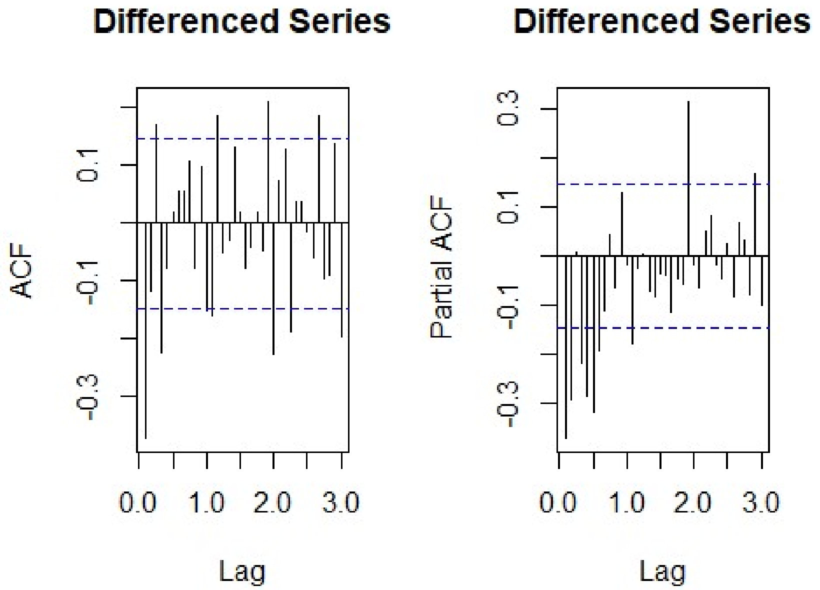


The ACF and PACF plots indicated a couple different p and q orders. In the model development different orders were used to search for the best fitting model using the Akaike information criterion (AIC) score. The AIC score considers how well the model fits the data while also considering the complexity of the model. The lower the AIC score, the better the model. Important points from the above figure include the trending downward of the ACF and the cutoff after the fourth lag in the PACF which was close to significance to warrant initial testing of an AR(4) model. In the ARMA model development the following models were also considered: AR(2), AR(3), ARMA(1,1), ARMA(3,1), and ARMA(4,2) for comparison purposes. Of these models, the best fitting was an ARMA(4,2) which returned an AIC score of 320.

After determining the best model based on an AIC score, the next step in the process was to check diagnostics of the model. This included plotting the ACF and PACF of the residuals, checking the p values for the Ljung-Box statistic, as well as reviewing the QQ plot to ensure that the normality assumption has not been violated. The best fitting ARMA model returned a residual ACF and PACF that were close to white noise. Reviewing the QQ plot, it did not appear that the normality assumption was violated. However, the p values for the Ljung-Box statistic suggested there was likely a better fitting model with some of the p values at lag 7 and above nearing 0. Plots for diagnostics can be found in the appendix.

With the distinct seasonal patterns that exist in the time series it only seemed reasonable to test an ARIMA model, especially given that the best fitting ARMA model showed weak correlation at later lags. Seasonality can be accounted for in the ARIMA model by differencing the series and allowing for more versatility with data that is non-stationary. First, the time series was differenced to control for the trend that exists in the data, then the series was differenced a second time to remove the seasonality, and finally the ACF and PACF plots were examined for the differenced series to find appropriate AR and MA orders to insert into a

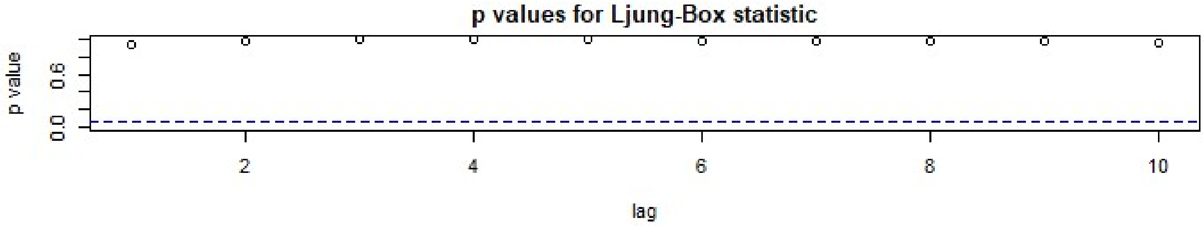
SARIMA model. To find the p and q orders the ACF and PACF plots below provided insight into the respective values.



The ACF and PACF plots for the differenced series suggested models with the following AR components, $ARIMA(1,1,0)X(1,1,0)_{12}$ and MA components, $ARIMA(0,1,1)X(1,1,1)_{12}$. This was in part because of the significant lag that was observed at Lag 1.0, suggestive of a yearly seasonal component. This was expected given the cyclical nature of the dataset. In addition to the previously mentioned models, a model was fitted with the AR and MA components that performed best in the ARMA model, namely ARMA(4,2), to see if those findings carried over to the ARIMA model. Using the AIC scores from the different models the lowest score was taken as the selection criteria of which model to use. The AIC scores with each respective model are shown below.

| AIC Score | Model |
|---|---|
| 394.631935587556 | ARIMA(1,1,0) X (1,1,0)$_{12}$ |
| 335.205670498012 | ARIMA(0,1,1) X (0,1,1)$_{12}$ |
| 326.330603602146 | ARIMA(1,1,1) X (1,1,1)$_{12}$ |
| 329.234451674845 | ARIMA(0,1,1) X (1,1,1)$_{12}$ |
| 312.94357690026 | ARIMA(4,1,2) X (1,1,1)$_{12}$ |

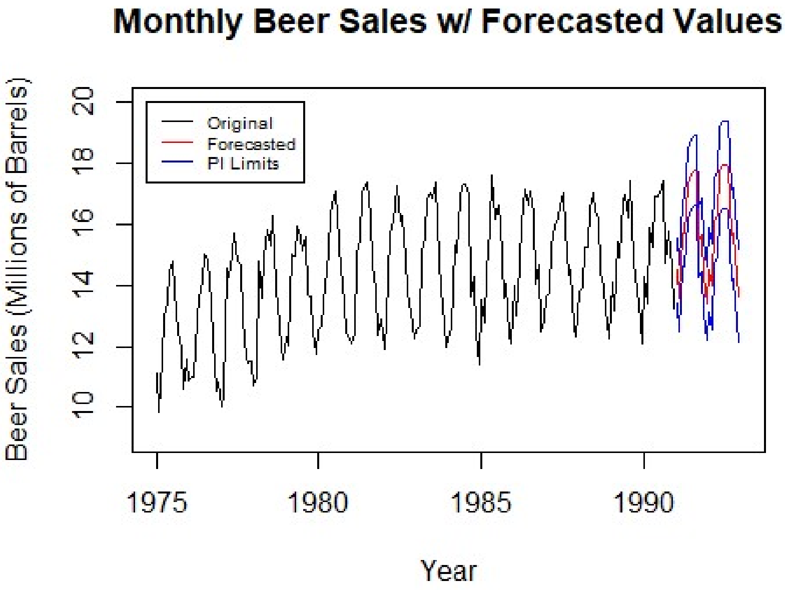
From the models the $ARIMA(4,1,2)X(1,1,1)_{12}$ received the lowest AIC score with a value of 312.9, lower than the score received from the ARMA(4,2) model that was fitted first. While this ARIMA model scored better diagnostics still needed to be performed to ensure the overall fit. The full residual tests are shown in the appendix, the key plot highlighted below is the p values from the Ljung-Box statistic.



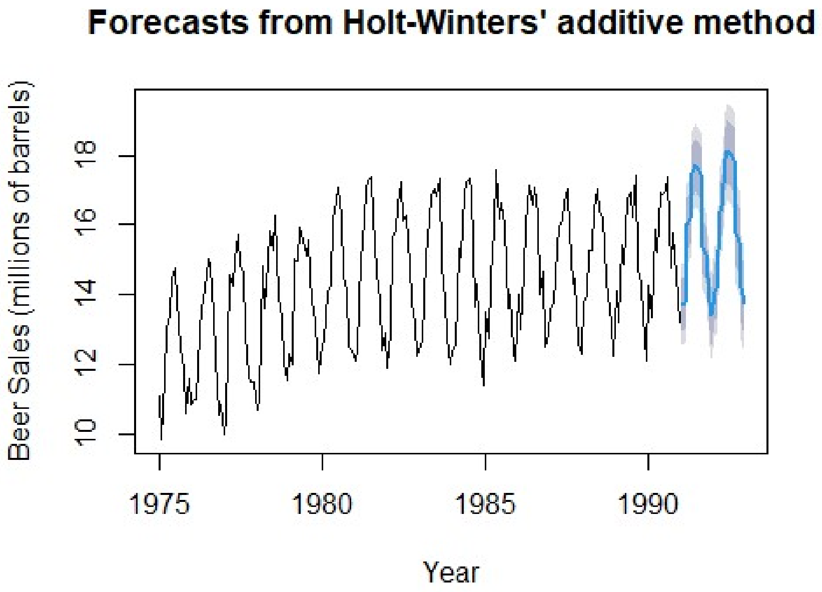
The p values here indicated strong evidence of a model that fit the data well with nearly all the lags having a p value of greater than 0.8.

## Forecasting

In forecasting two years of data, two different methods were considered; the SARIMA model from the previous section and the Holt-Winters seasonal smoothing method. In the figure below are the forecasted values in red using the $SARIMA(4,1,2)X(1,1,1)_{12}$ model. Also included in the figure are the 95% prediction interval limits, which are captured in blue. As can be seen in the figure, the model captures the seasonality of the data quite well where the peaks tend to happen around summer and the troughs tend to happen towards the end/beginning of the year similar to the historical trends.



In the figure below is the forecasted beer sales using the HW method for two years. An additive method was used since the seasonality of the data seemed to be relatively stable and the variation did not increase over time. It displays a similar trend as the SARIMA model with the peaks happening in the summer months while the troughs taking place in the winter. The plot also shows the 80% and 95% confidence bands for the data, which are also similar to the SARIMA model.



## Conclusion/Discussion

In reviewing the beer sales data a suitable SARIMA model was determined. In the process of finding a suitable model, a method of deconstructing the data was employed to get to a level of stationarity. Simpler AR, MA, and ARMA models were considered prior to moving onto more complex SARIMA models. While an ARMA(4,2) model appeared to be suitable in most aspects, it was determined that based on the Ljung-Box statitic that a more suitable model was necessary. However, it did help in finding a suitable SARIMA model by using similar p and q orders for the AR and MA components respectively.

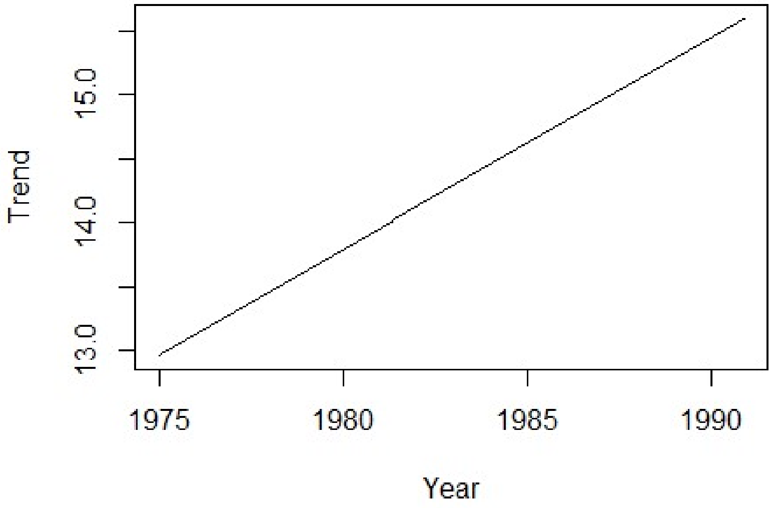
When forecasting beer sales, both the selected SARIMA model as well as the Holt-Winters seasonal smoothing method were considered. Both model and method provided similar results suggesting that both are reasonable methods in forecasting beer sales. The SARIMA model provides some additional insight into the trend and seasonal aspects of the data as these two items have to be removed prior to determining the p and q orders. The HW method on the other hand is much simpler to implement and just as good as the SARIMA model in fitting the data. If understanding certain aspects of the time series is more important, then the method of deconstructing the data and looking for an appropriate model is a suitable method. However, if simplicity and understanding the data are less important, then the HW method may be a more appropriate approach.

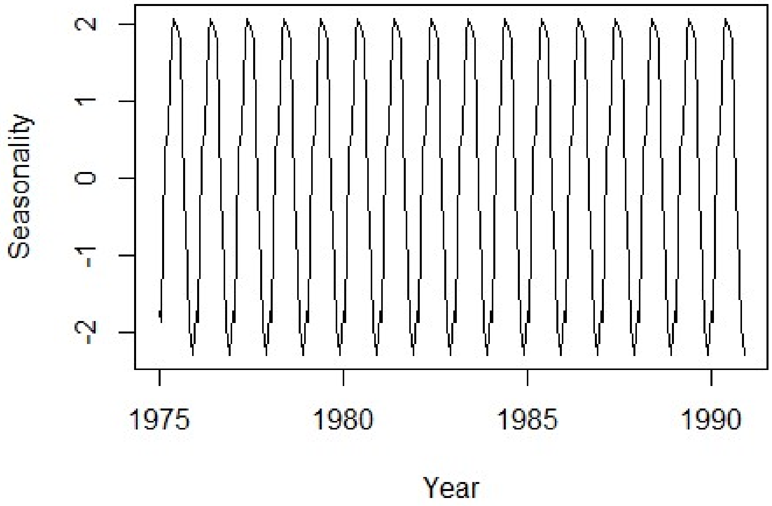
## References

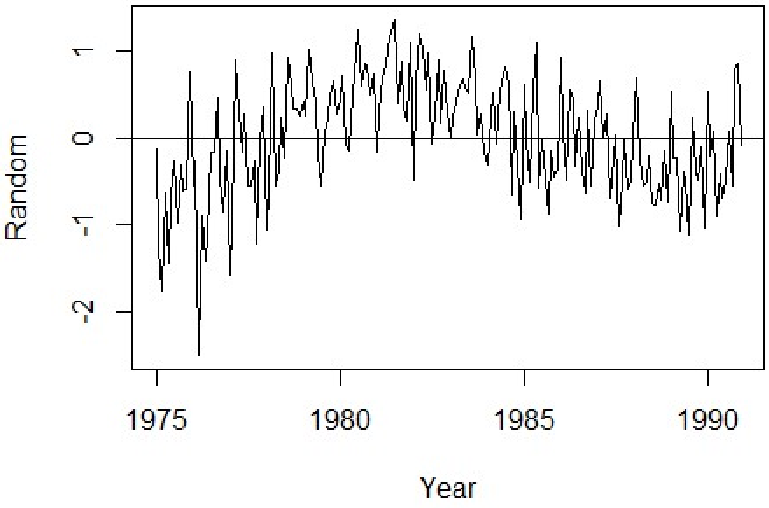
Chisholm, B. and Fedorczyk, D. “Time Series Analysis of Beer Sales.” ST566: Time Series Analytics, Winter 2021.

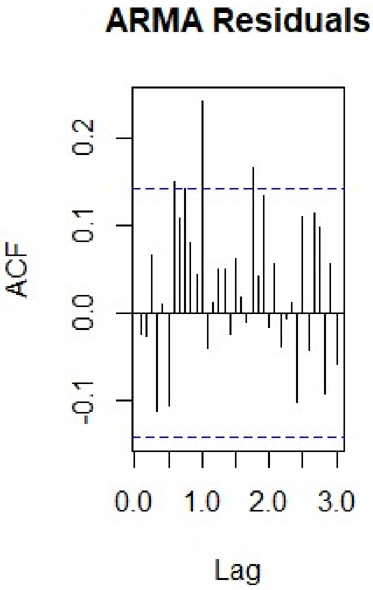
## Appendix

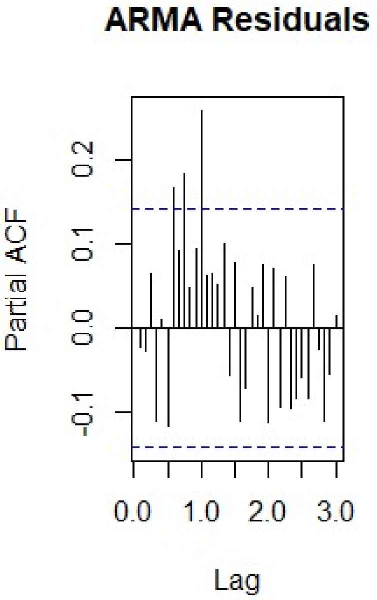
### Additional Plots

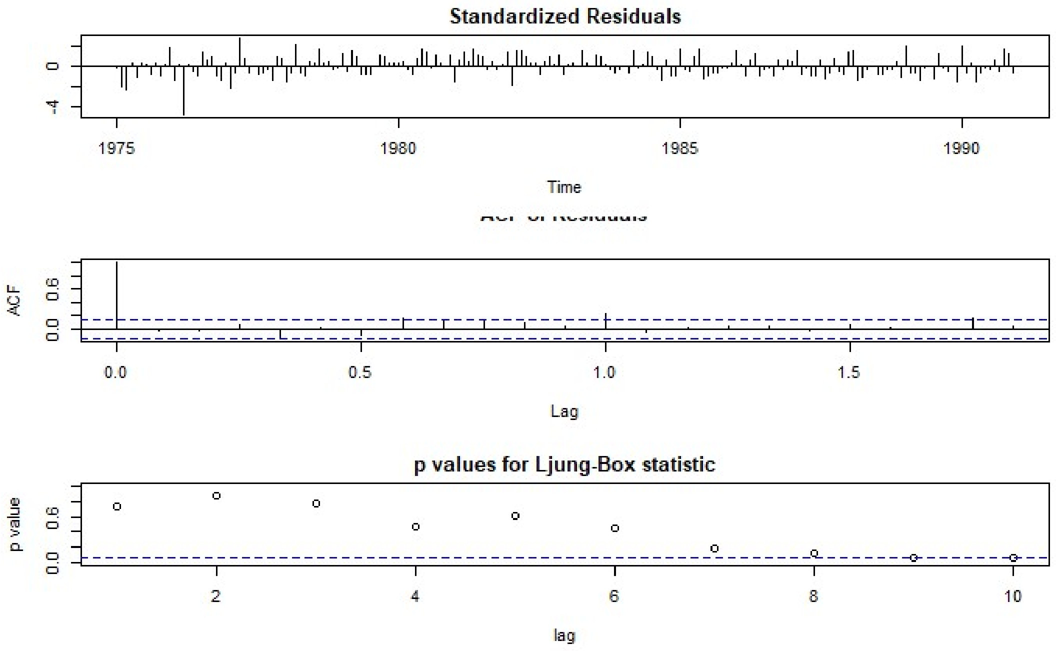


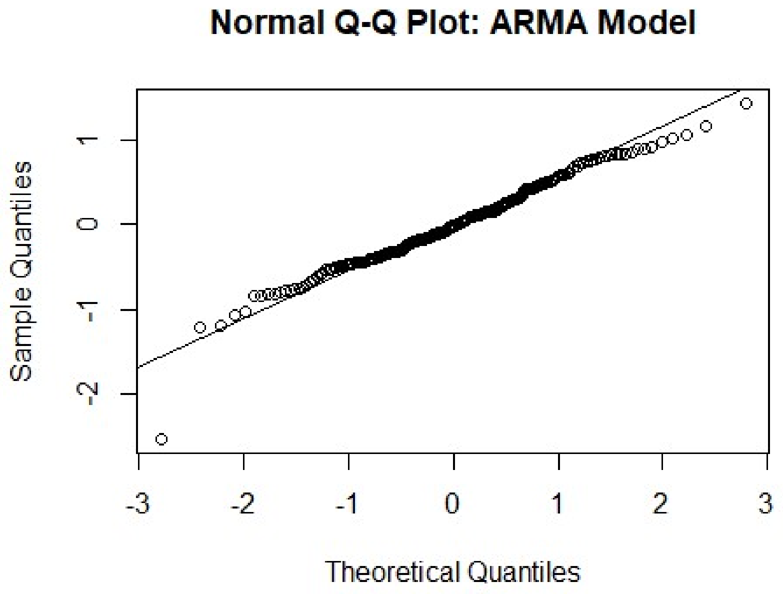


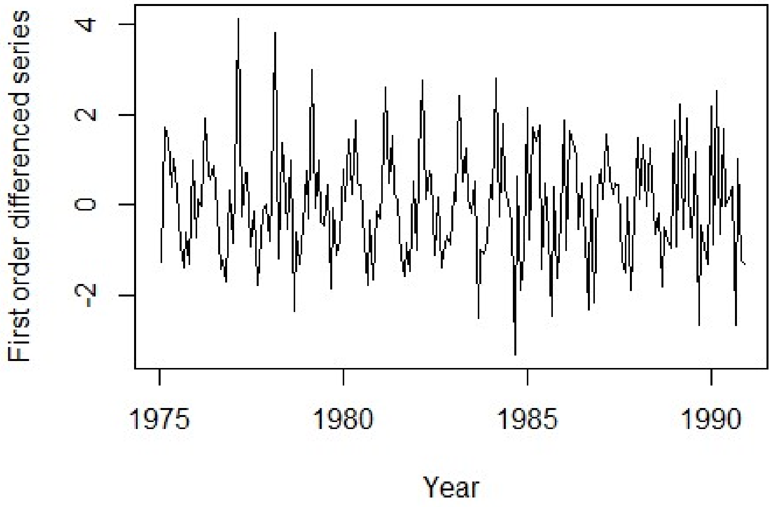

Random
1
0
-1
-2
1975
1980
1985
1990
Year

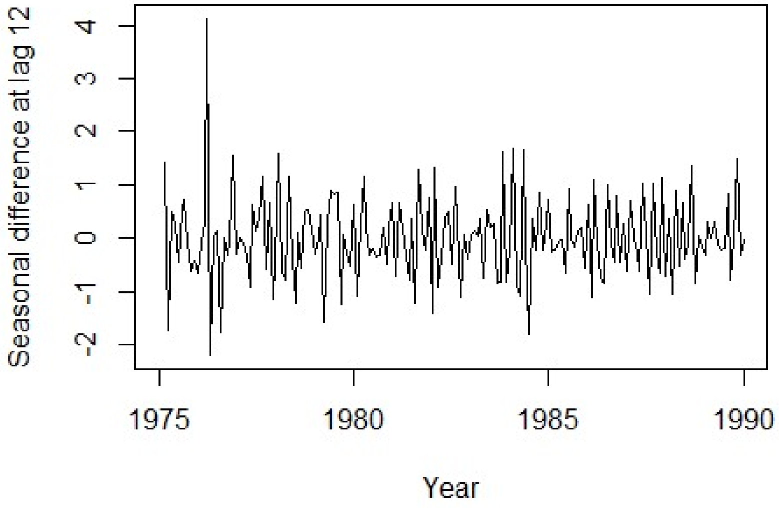

ARMA Residuals
ACF
0.2
0.1
0.0
-0.1
0.0
1.0
2.0
3.0
Lag

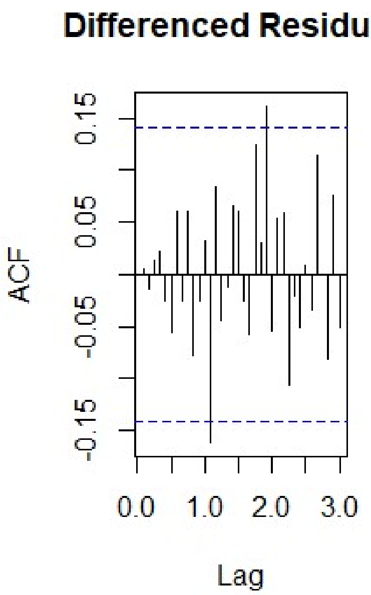

ARMA Residuals
Partial ACF
0.2
0.1
0.0
-0.1
0.0
1.0
2.0
3.0
Lag

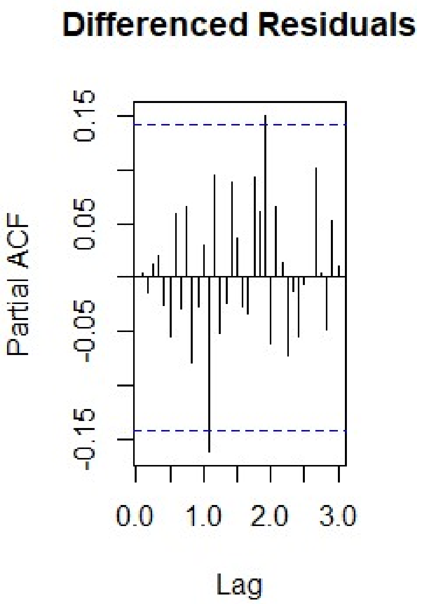

Standardized Residuals
0
-4
1975
1980
1985
1990
Time
ACF of Residuals
ACF
0.0
0.6
0.0
0.5
1.0
1.5
Lag
p values for Ljung-Box statistic
p value
0.0
0.6
2
4
6
8
10
lag

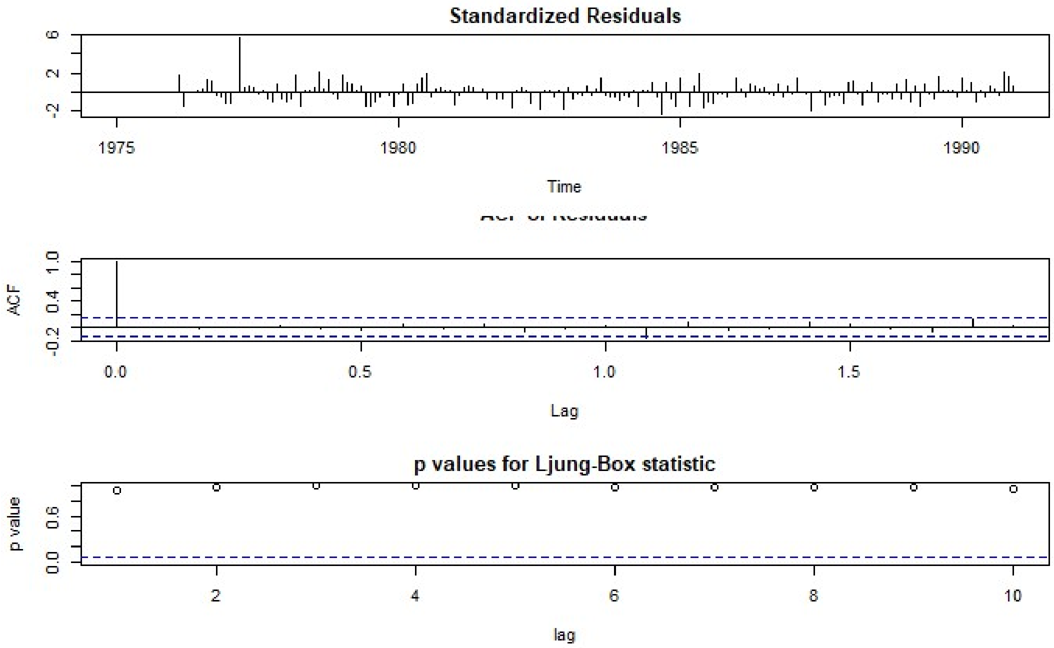

Normal Q-Q Plot: ARMA Model
Sample Quantiles
1
0
-1
-2
-3
-2
-1
0
1
2
3
Theoretical Quantiles

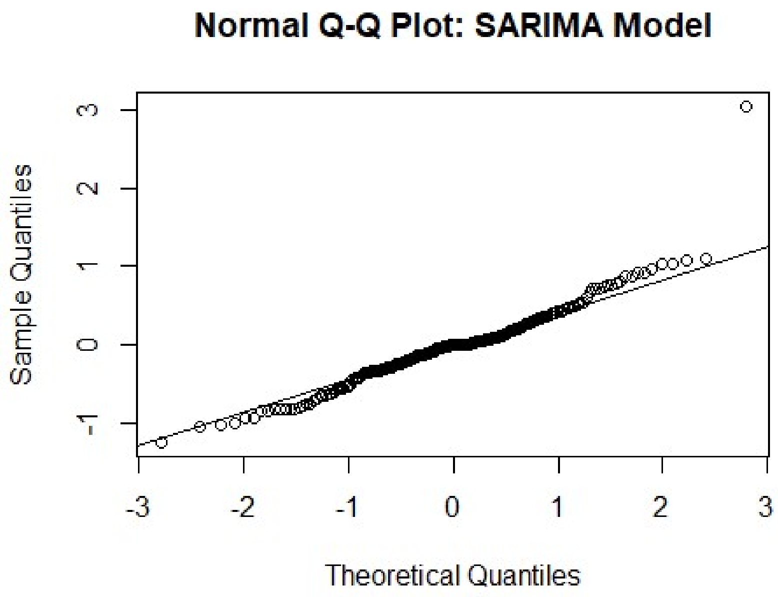

First order differenced series
4
2
0
-2
1975
1980
1985
1990
Year


Seasonal difference at lag 12
4
3
2
1
0
-1
-2
1975
1980
1985
1990
Year


Differenced Residuals
ACF
0.15
0.05
-0.05
-0.15
0.0
1.0
2.0
3.0
Lag


Differenced Residuals
Partial ACF
0.15
0.05
-0.05
-0.15
0.0
1.0
2.0
3.0
Lag


Standardized Residuals
6
2
-2
1975
1980
1985
1990
Time
ACF of Residuals
ACF
1.0
0.6
0.2
-0.2
0.0
0.5
1.0
1.5
Lag
p values for Ljung-Box statistic
p value
0.6
0.0
2
4
6
8
10
lag



### R Code

```
# Load Time Series libraries
library(TSA)
library(forecast)
library(tidyverse)

# Load beersales data set
data(beersales)

# Plot data
plot(beersales, xlab = "Year", ylab = "Sales (millions of barrels)",
     main = "Monthly Beer Sales")

# Nominal ARMA Model Fitting

## Trend and Seasonality Estimation

# fit a linear model
beer.time <- time(beersales)
fit.trend <- lm(beersales ~ beer.time)
beer.trend <- ts(fit.trend$fitted.values, start = c(1975,1), deltat = 1/12)

# linear trend removed
fit.res <- fit.trend$residuals

# format it into time series
fit.res <- ts(fit.res, start = c(1975,1), deltat = 1/12)

# Seasonality removal
```

```r
beer.month <- factor(cycle(fit.res))
fit.season <- lm(fit.res ~ beer.month)
beer.season <- ts(fit.season$fitted, start = c(1975,1), deltat = 1/12)

# With both linear trend and seasonality removed
beer.rand <- ts(fit.res - beer.season, start = c(1975,1), deltat = 1/12)

# Plot the decomposition
plot(beer.trend, xlab = "Year", ylab = "Trend")
plot(beer.season, xlab = "Year", ylab = "Seasonality")
plot(beer.rand, xlab = "Year", ylab = "Random")
abline(h=0)

## Fit Nominal ARMA Model and Residual Series

# check acf and pacf of the residual series
par(mfrow = c(1,2))
acf(beer.rand, main = "Random Series", lag.max=36)
acf(beer.rand, type = "partial", main = "Random Series", ylab = "Partial ACF"
, lag.max=36)

# Fit different models for the residual series: AR(2), AR(3), AR(4), ARMA(1,1
), ARMA(3,1)
(fit.ar3 <- arima(beer.rand, order = c(p = 3, d = 0, q = 0), method = "ML", i
nclude.mean = F))

(fit.ar4 <- arima(beer.rand, order = c(p = 4, d = 0, q = 0), method = "ML", i
nclude.mean = F))

(fit.ar5 <- arima(beer.rand, order = c(p = 5, d = 0, q = 0), method = "ML", i
nclude.mean = F))

(fit.arma11 <- arima(beer.rand, order = c(p = 1, d = 0, q = 1), method = "ML"
, include.mean = F))

(fit.arma41 <- arima(beer.rand, order = c(p = 4, d = 0, q = 1), method = "ML"
, include.mean = F))

(fit.arma42 <- arima(beer.rand, order = c(p = 4, d = 0, q = 2), method = "ML"
, include.mean = F))

## Fitting residuals
par(mfrow = c(1, 2))
res <- fit.arma42$residuals
acf(res, lag.max = 36, main = "ARMA Residuals")
pacf(res, lag.max = 36, main = "ARMA Residuals")

## Diagnostics
par(mfrow = c(1,1), fin = c(5,10))
```

```
tsdiag(fit.arma42)

qqnorm(res, main = "Normal Q-Q Plot: ARMA Model")
qqline(res)

# Differencing

## Difference the Data

# Generate difference of series and plot results
diff1 <- c(NA, diff(beersales))
diff1 <- ts(diff1, start = c(1975,1), deltat = 1/12)
plot(diff1, xlab = "Year", ylab = "First order differenced series")

# Generate seasonal difference and plot results
diff12 <- c(NA, diff(diff1, lag = 12))
diff12 <- ts(diff12, start = c(1975,1), deltat = 1/12)
plot(diff12, xlab = "Year", ylab = "Seasonal difference at lag 12")

## Examine ACF and PACF of Differenced Series

# Generate ACF and PACF
par(mfrow = c(1,2))
acf(diff12, lag.max = 36, na.action = na.pass, main = "Differenced series")
acf(diff12, lag.max = 36, na.action = na.pass, type = 'partial', main = "Diff
erenced series", ylab = " Partial ACF")

## Fit a SARIMA Model

# Fit SARIMA models
n <- length(beersales)

# Period = 12
(fit.sar <- arima(beersales, order = c(1, 1, 0), seasonal = list(order = c(1,
1, 0), period = 12)))

(fit.sma <- arima(beersales, order = c(0, 1, 1), seasonal = list(order = c(0,
1, 1), period = 12)))

(fit.sarima1 <- arima(beersales, order = c(1, 1, 1), seasonal = list(order =
c(1, 1, 1), period = 12)))

(fit.sarima2 <- arima(beersales, order = c(0, 1, 1), seasonal = list(order =
c(1, 1, 1), period = 12)))

(fit.sarima3 <- arima(beersales, order = c(4, 1, 2), seasonal = list(order =
c(1, 1, 1), period = 12)))
```

```r
aic <- rbind(fit.sar$aic, fit.sma$aic, fit.sarima1$aic, fit.sarima2$aic, fit.
sarima3$aic)
model <- rbind("ARIMA(1,1,0) X (1,1,0)", "ARIMA(0,1,1) X (0,1,1)", "ARIMA(1,1
,1) X (1,1,1)", "ARIMA(0,1,1) X (1,1,1)", "ARIMA(4,1,2) X (1,1,1)")
table <- cbind(aic, model)
colnames(table) <- c("aic score", "model")
table

## Model Diagnostics

## Fitting residuals
par(mfrow = c(1,2))
res.diff <- fit.sarima3$residuals
acf(res.diff, lag.max = 36, main = "Differenced Residuals")
pacf(res.diff, lag.max = 36, main = "Differenced Residuals")

par(mfrow = c(1,1))
tsdiag(fit.sarima3)

qqnorm(res.diff, main = "Normal Q-Q Plot: SARIMA Model")
qqline(res.diff)

## Forecasting

# Forecast 1-year Aheard
pred <- predict(fit.sarima3, n.ahead = 24)
plot(beersales, xlim = c(1975, 1993), ylim = c(9.0, 20.0), main = "Monthly Be
er Sales w/ Forecasted Values for 1991-92", xlab = "Year", ylab = "Beer Sales
(Millions of Barrels)")

# Plot Forecasted Values
lines(pred$pred, col = "red")

# Plot 95% Forecasting Limits
lines(pred$pred-2*pred$se,col='blue')
lines(pred$pred+2*pred$se,col='blue')

# Include Legend
legend(x=1974.7, y=20, c("Original", "Forecasted", "PI Limits"), cex = 0.6, l
ty = 1, lwd = 0.8, col = c("black", "red", "blue"))

# Additive Model
fore.beer1 <- hw(beersales, seasonal = "additive", h = 24)
plot(fore.beer1, xlab = "Year", ylab = "Beer Sales (millions of barrels)")
```