Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Lab 2

Brittany Dougall, Steve Hall, Prabhu Narsina, and Edward Salinas

Instructions (Please Read Carefully):

- Submit by the due date. **Late submissions will not be accepted**
- No page limit, but be reasonable
- Do not modify fontsize, margin or line-spacing settings
- One student from each group should submit the lab to their student github repo by the deadline
- Submit two files:
  1. A pdf file that details your answers. Include all R code used to produce the answers
  2. The R markdown (Rmd) file used to produce the pdf file

  The assignment will not be graded unless **both** files are submitted
- Name your files to include all group members names. For example, if the students' names are Stan Cartman and Kenny Kyle, name your files as follows:
  - `StanCartman_KennyKyle_Lab2.Rmd`
  - `StanCartman_KennyKyle_Lab2.pdf`
- Although it sounds obvious, please write your name on page 1 of your pdf and Rmd files
- All answers should include a detailed narrative; make sure that your audience can easily follow the logic of your analysis. All steps used in modelling must be clearly shown and explained; do not simply 'output dump' the results of code without explanation
- If you use libraries and functions for statistical modeling that we have not covered in this course, you must provide an explanation of why such libraries and functions are used and reference the library documentation
- For mathematical formulae, type them in your R markdown file. Do not e.g. write them on a piece of paper, snap a photo, and use the image file
- Incorrectly following submission instructions results in deduction of grades
- Students are expected to act with regard to UC Berkeley Academic Integrity.

The Keeling Curve

In the 1950s, the geochemist Charles David Keeling observed a seasonal pattern in the amount of carbon dioxide present in air samples collected over the course of several years. He attributed this pattern to varying rates of photosynthesis throughout the year, caused by differences in land area and vegetation cover between the Earth's northern and southern hemispheres.

In 1958 Keeling began continuous monitoring of atmospheric carbon dioxide concentrations from the Mauna Loa Observatory in Hawaii. He soon observed a trend increase carbon dioxide levels in addition to the seasonal cycle, attributable to growth in global rates of fossil fuel combustion. Measurement of this trend at Mauna Loa has continued to the present.

The `co2` data set in R's `datasets` package (automatically loaded with base R) is a monthly time series of atmospheric carbon dioxide concentrations measured in ppm (parts per million) at the Mauna Loa Observatory from 1959 to 1997. The curve graphed by this data is known as the 'Keeling Curve'.

Part 1 (3 points)

Conduct a comprehensive Exploratory Data Analysis on the `co2` series. This should include (without being limited to) a thorough investigation of the trend, seasonal and irregular elements.

```
opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE,
    warning = FALSE, message = FALSE)

str(co2)
```

```
##  Time-Series [1:468] from 1959 to 1998: 315 316 316 318 318 ...
```

```
summary(co2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   313.2   323.5   335.2   337.1   350.3   366.8
```

```
co2.decompose = decompose(co2)
co2.diff = diff(co2, differences = 1)
co2.seasdiff = diff(co2, lag = 12)
co2.bothdiff = diff(co2.diff, lag = 12)

co2.deseasoned = co2 - co2.decompose$seasonal
co2.detrended = co2 - co2.decompose$trend

par(mfrow = c(3, 1))

plot(co2, ylab = expression("CO2 ppm"), col = "blue", las = 1)
title(main = "Figure1: Monthly Mean CO2 Variation")

boxplot(co2 ~ cycle(co2), main = "Boxplot of CO2 (ppm) by month")

plot(co2.deseasoned, main = expression("Figure2: Presence of CO2 in air  after removing season"
    xlab = "year", ylab = expression("CO2 ppm"))
```

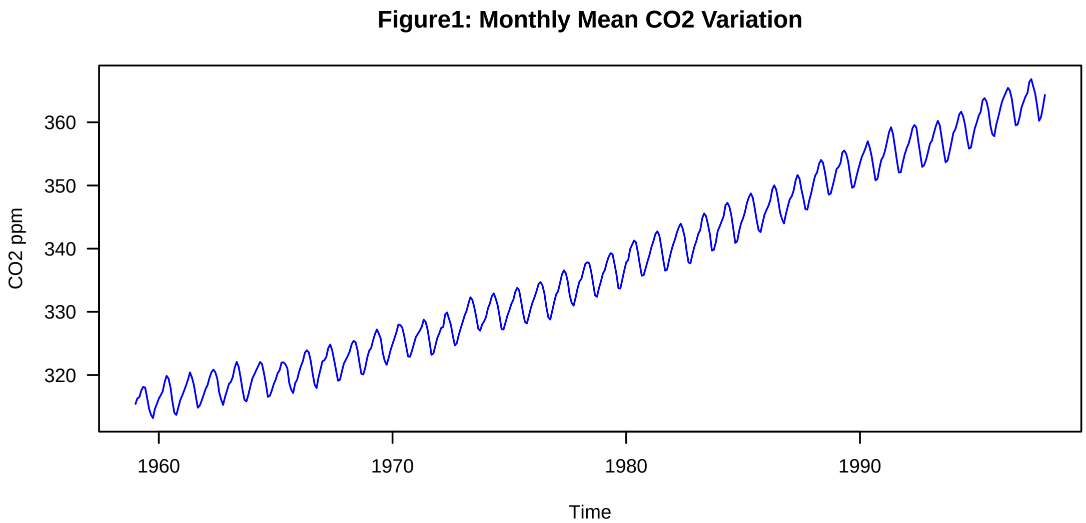
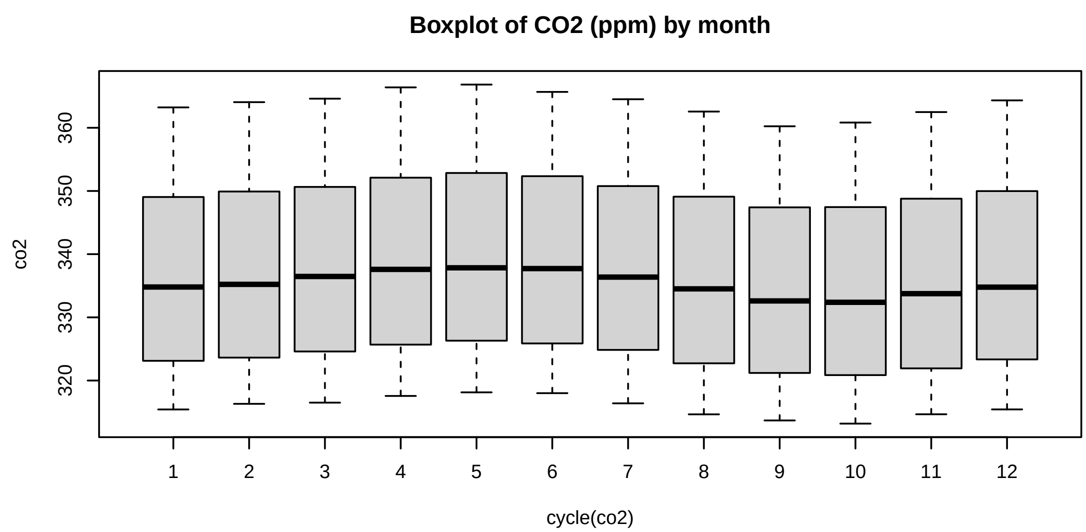


Figure1: Monthly Mean CO2 Variation
CO2 ppm
320
330
340
350
360
1960
1970
1980
1990
Time


Boxplot of CO2 (ppm) by month
co2
320
330
340
350
360
1
2
3
4
5
6
7
8
9
10
11
12
cycle(co2)

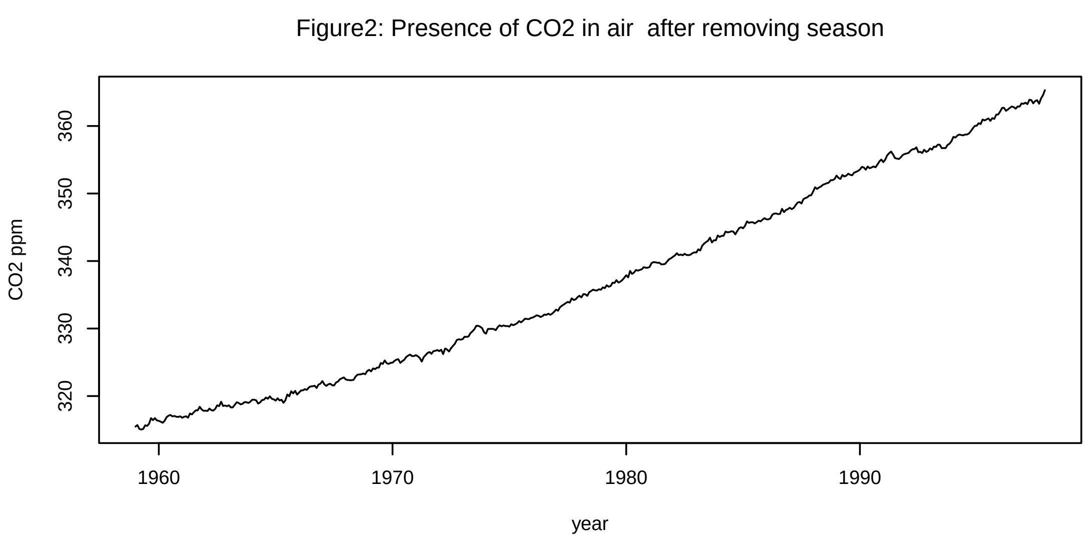


Figure2: Presence of CO2 in air after removing season
CO2 ppm
320
330
340
350
360
1960
1970
1980
1990
year

```r
plot(co2.detrended, main = expression("Figure3: Presence of CO2 in air after removing trend"),
    xlab = "year", ylab = expression("CO2 ppm"), col = "red",
    las = 1)

abline(h = 0)

plot(co2.diff, main = expression("Figure4: Presence of CO2 in air after differencing"),
    xlab = "year", ylab = expression("CO2 ppm"), col = "red",
    las = 1)

abline(h = 0)

plot(co2.seasdiff, main = expression("Figure5: Presence of CO2 in air after seasonal differenci
    xlab = "year", ylab = expression("CO2 ppm"), col = "red",
    las = 1)
abline(h = 0)
```

Figure3: Presence of CO2 in air after removing trend

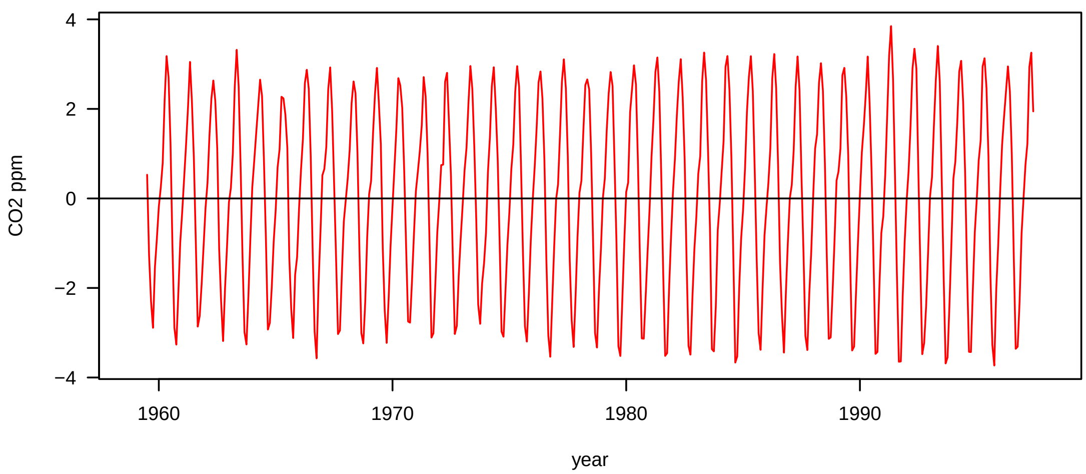


Figure4: Presence of CO2 in air after differencing

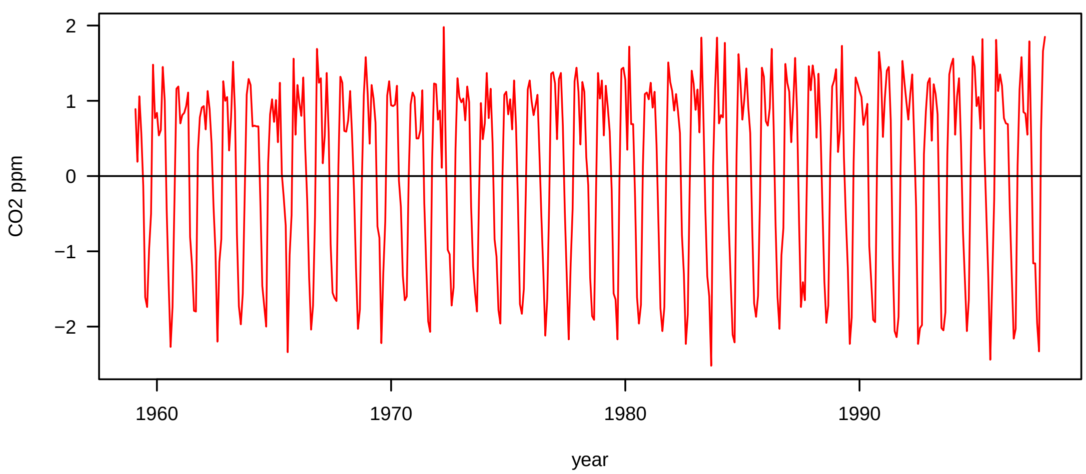
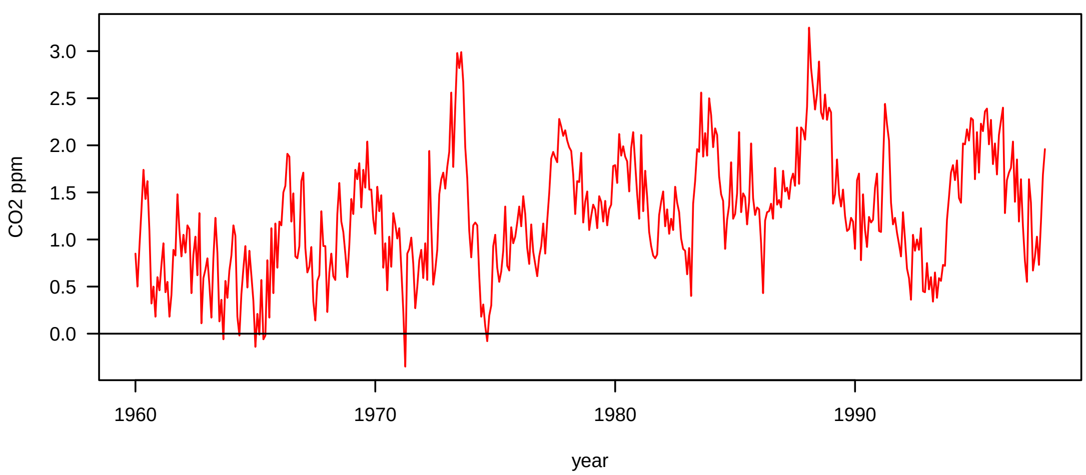


Figure5: Presence of CO2 in air after seasonal differencing



```
plot(co2.bothdiff, main = expression("Figure6: Presence of CO2 in air non-seasonal and seasonal
    xlab = "year", ylab = expression("CO2 ppm"), col = "red",
    las = 1)
abline(h = 0)
```

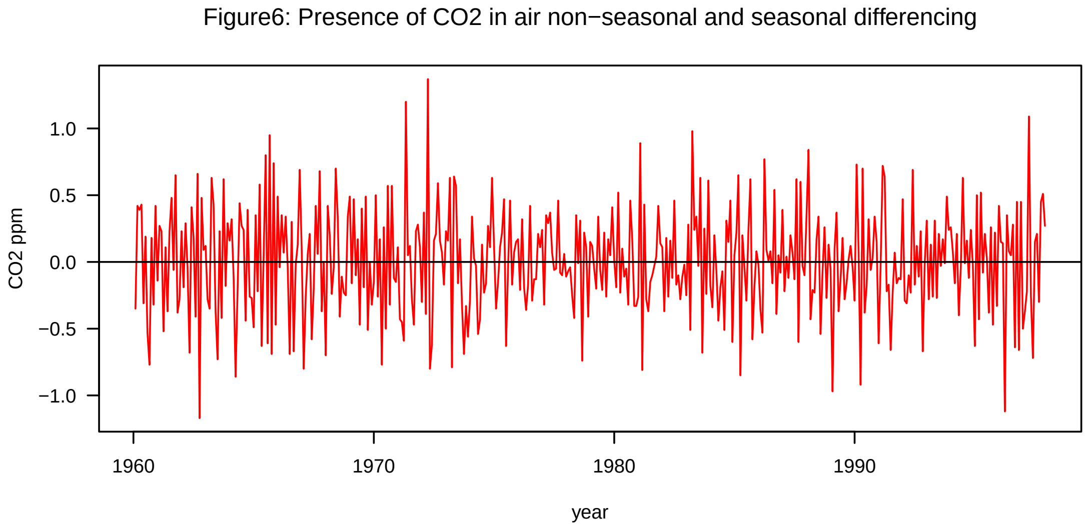
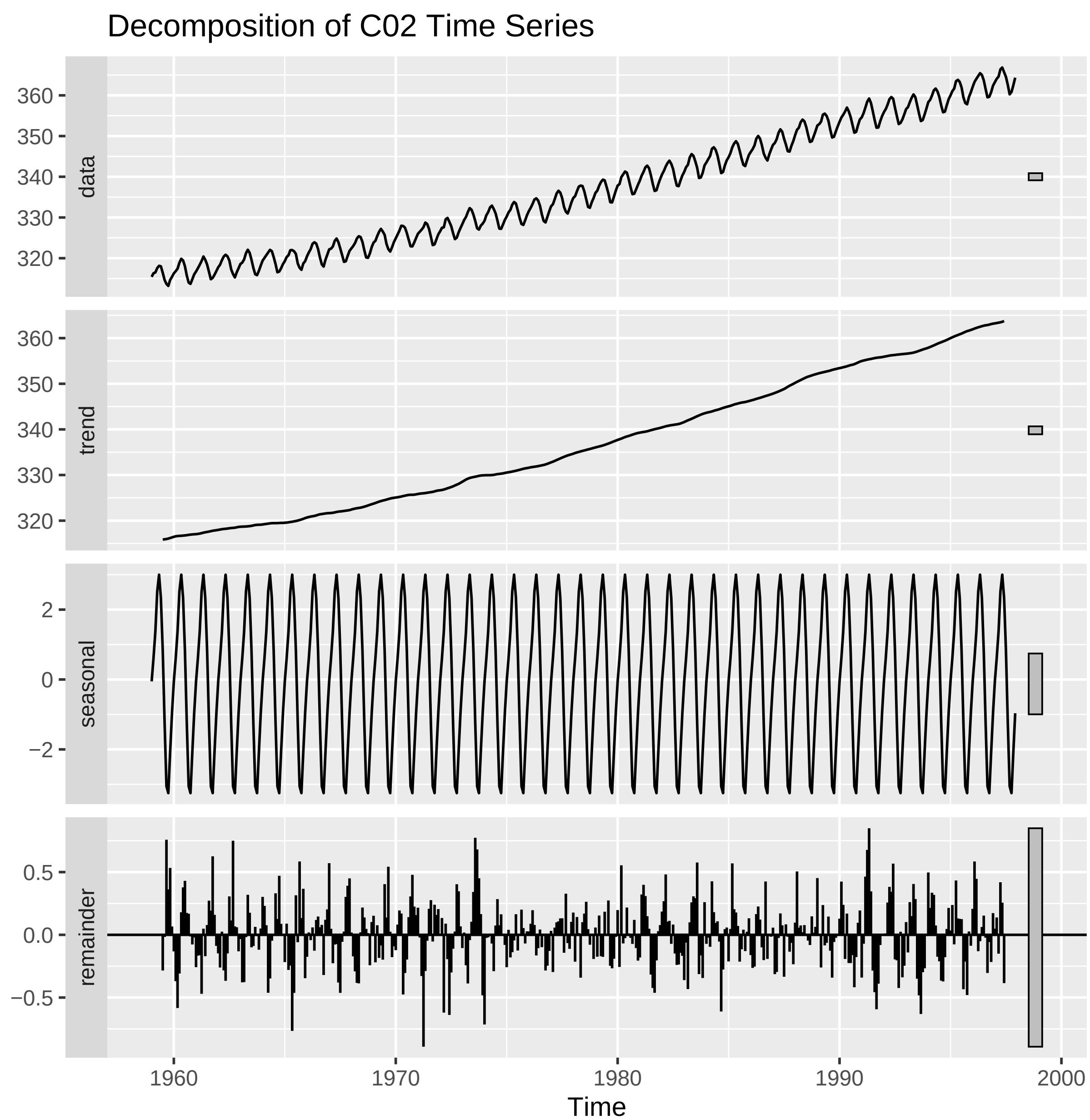


Figure6: Presence of CO2 in air non−seasonal and seasonal differencing
CO2 ppm
1.0
0.5
0.0
−0.5
−1.0
1960
1970
1980
1990
year

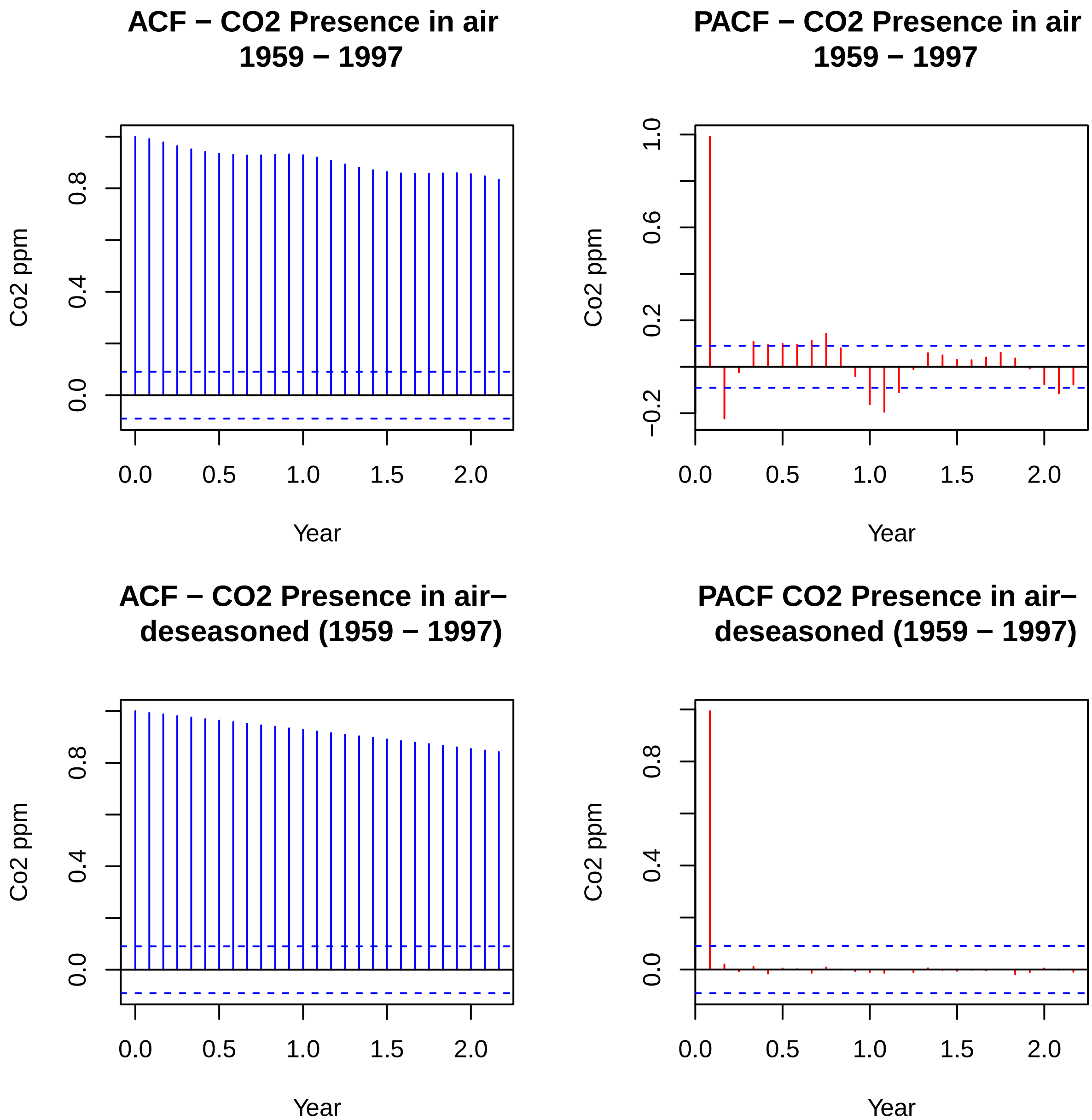
Data provided has CO2 presence in the air (parts per million) in monthly time series format from 1959 to 1998.
From Figure1: The time series plot of the mean of co2 presence in the air indicates a clear trend and seasonal effect. We also observe that the variance is constant over time, which suggests no need for transformation.
From Figure2: We see a clear upward trend in the mean of the presence of Co2 in the air
From Figure3: Co2 presence in the air after removing the trend component from the time series indicates the persistent yearly seasonal effect.
From Figure4: Trend is abstracted after taking the 2-period difference of the time series. It suggests we use ARIMA with integration/difference of 2
From Figure5: Seasonality absent after applying difference of 12 lags for the season. We still see trends present.
From Figure6: Seasonality and trend are absent after difference at two lags and 12 lags for the season. It is much closer to white noise series with non-constant variance. It suggests a possible need of Seasonal adjustment for the ARIMA model

```
autoplot(co2.decompose, main = "Decomposition of CO2 Time Series")
```

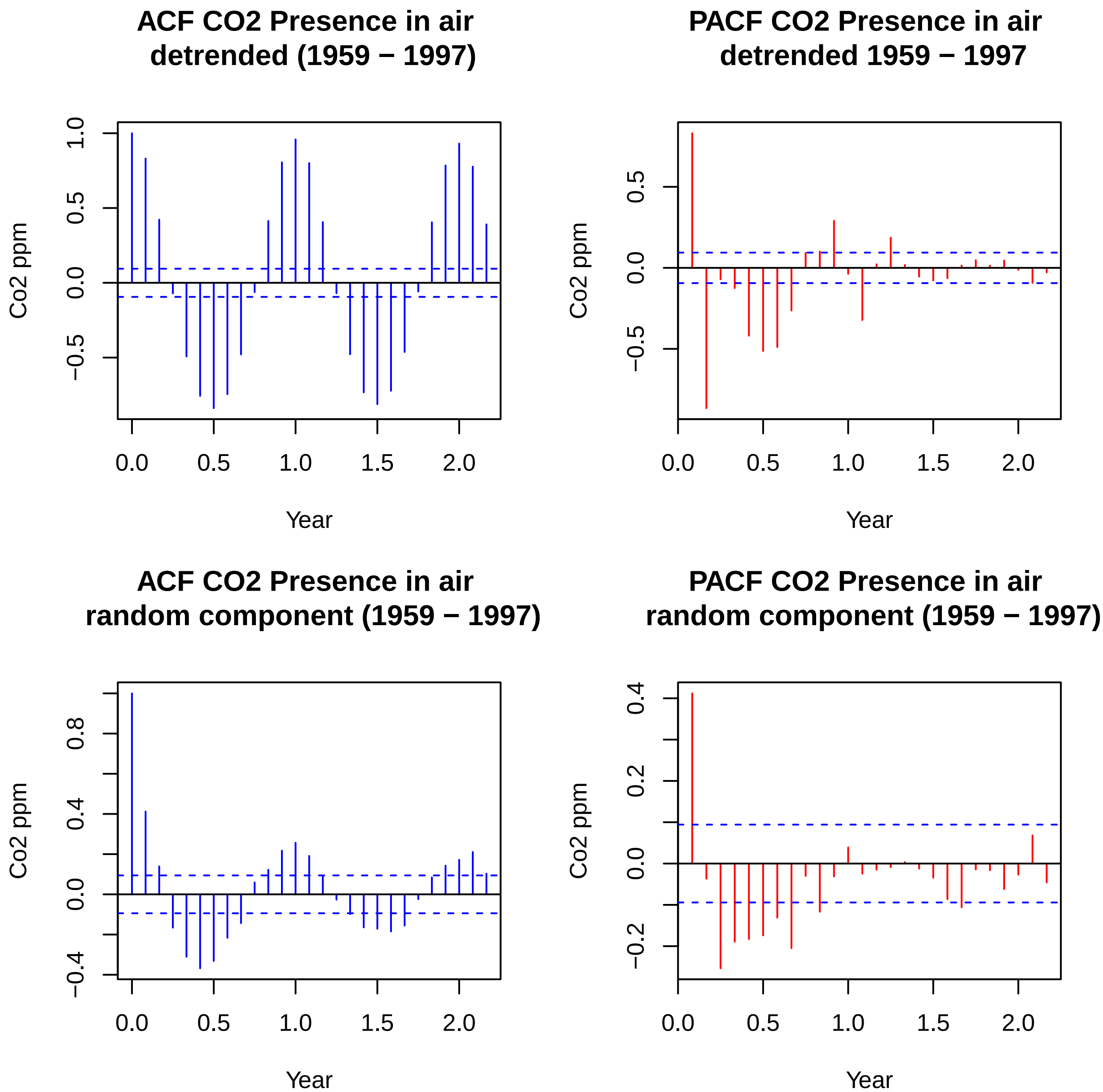



```
plot.acf.alldata = acf(co2, plot = FALSE)
plot.pacf.alldata = pacf(co2, plot = FALSE)

plot.acf.deseasoned = acf(co2.deseasoned, plot = FALSE)
plot.pacf.deseasoned = pacf(co2.deseasoned, plot = FALSE)

plot.acf.detrended = acf(window(co2.detrended, start = c(1960),
    end = c(1996)), plot = FALSE)
plot.pacf.detrended = pacf(window(co2.detrended, start = c(1960),
    end = c(1996)), plot = FALSE)

plot.acf.residual = acf(window(co2.decompose$random, start = c(1960),
```

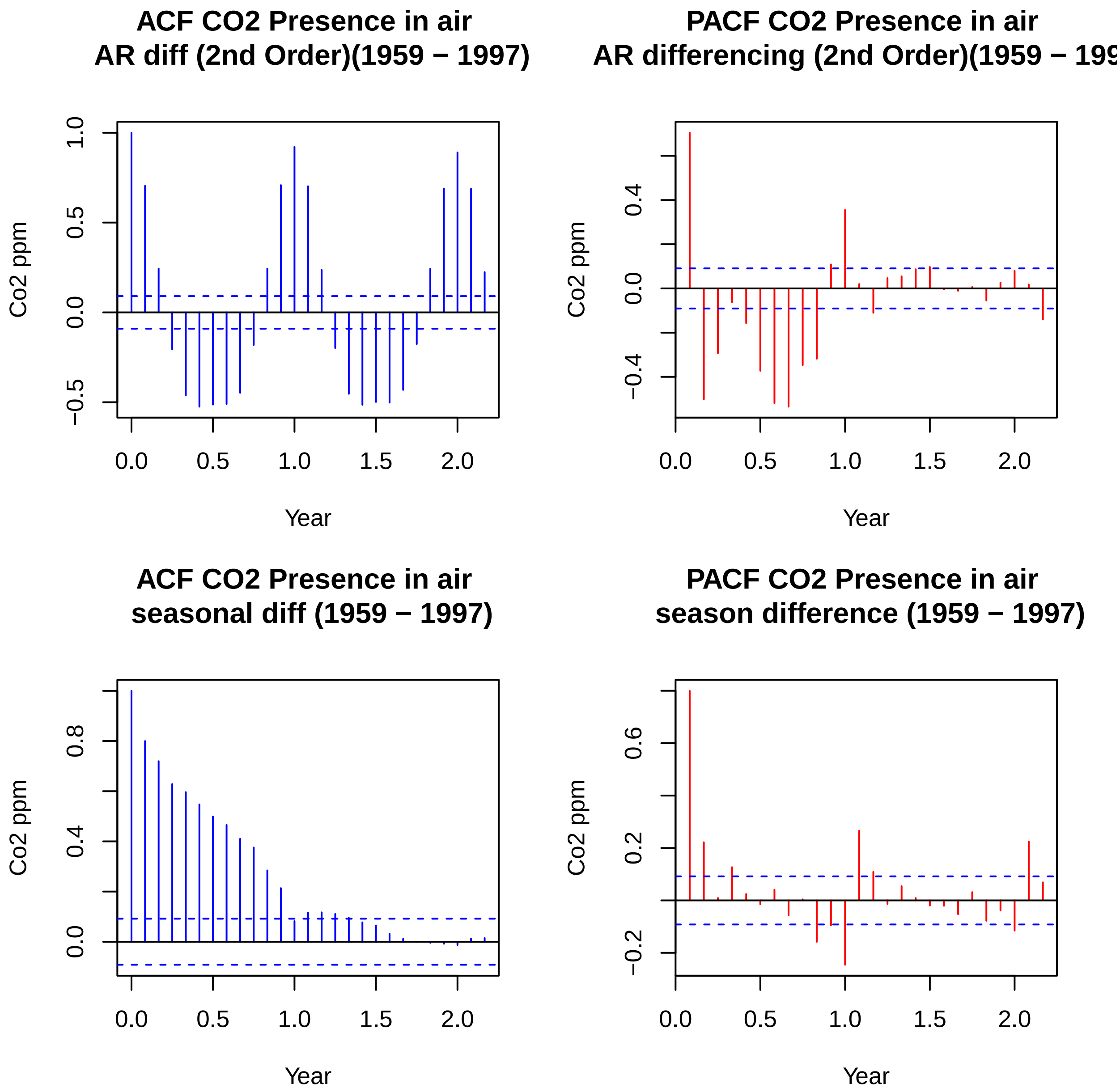
```
    end = c(1996)), plot = FALSE)
plot.pacf.residual = pacf(window(co2.decompose$random, start = c(1960),
    end = c(1996)), plot = FALSE)

plot.acf.diff = acf(co2.diff, plot = FALSE)
plot.pacf.diff = pacf(co2.diff, plot = FALSE)

plot.acf.seasondiff = acf(co2.seasdiff, plot = FALSE)
plot.pacf.seasondiff = pacf(co2.seasdiff, plot = FALSE)

plot.acf.bothdiff = acf(co2.bothdiff, plot = FALSE)
plot.pacf.bothdiff = pacf(co2.bothdiff, plot = FALSE)

par(mfrow = c(2, 2))
plot(plot.acf.alldata, main = "ACF - CO2 Presence in air \n 1959 - 1997",
    xlab = "Year", ylab = "Co2 ppm", col = "blue", cex.main = 0.5)
plot(plot.pacf.alldata, main = "PACF - CO2 Presence in air \n 1959 - 1997",
    xlab = "Year", ylab = "Co2 ppm", col = "red", cex.main = 0.5)


plot(plot.acf.deseasoned, main = "ACF - CO2 Presence in air- \n deseasoned (1959 - 1997)",
    xlab = "Year", ylab = "Co2 ppm", col = "blue")
plot(plot.pacf.deseasoned, main = "PACF CO2 Presence in air- \n deseasoned (1959 - 1997)",
    xlab = "Year", ylab = "Co2 ppm", col = "red", cex.main = 0.5)
```



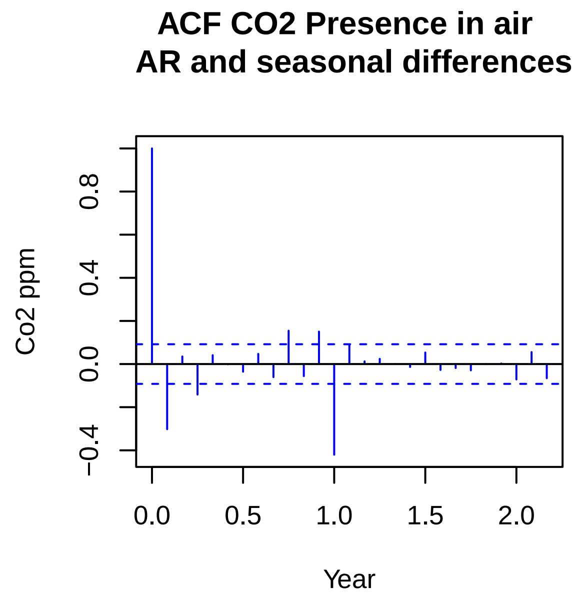

```r
plot(plot.acf.detrended, main = "ACF CO2 Presence in air \n detrended (1959 - 1997)",
    xlab = "Year", ylab = "Co2 ppm", col = "blue")
plot(plot.pacf.detrended, main = "PACF CO2 Presence in air \n detrended 1959 - 1997",
    xlab = "Year", ylab = "Co2 ppm", col = "red", cex.main = 0.5)

plot(plot.acf.residual, main = "ACF CO2 Presence in air \n random component (1959 - 1997)",
    xlab = "Year", ylab = "Co2 ppm", col = "blue")
plot(plot.pacf.residual, main = "PACF CO2 Presence in air \n random component (1959 - 1997)",
    xlab = "Year", ylab = "Co2 ppm", col = "red", cex.main = 0.5)
```

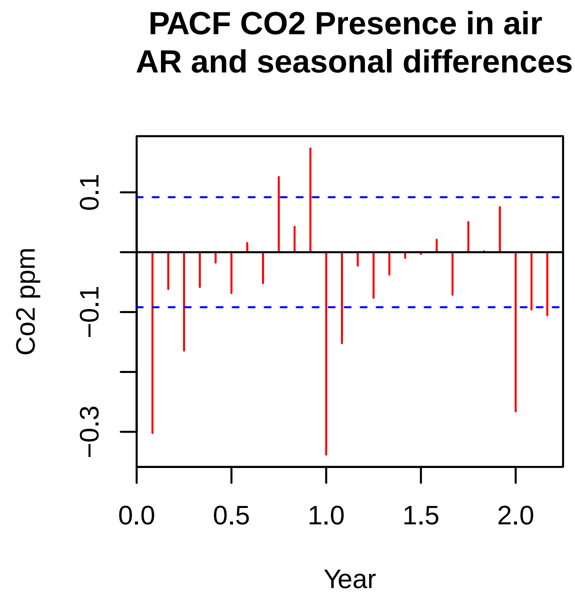



```
plot(plot.acf.diff, main = "ACF CO2 Presence in air \n AR diff (2nd Order)(1959 - 1997)",
    xlab = "Year", ylab = "Co2 ppm", col = "blue")
plot(plot.pacf.diff, main = "PACF CO2 Presence in air \n AR differencing (2nd Order)(1959 - 199
    xlab = "Year", ylab = "Co2 ppm", col = "red", cex.main = 0.5)

plot(plot.acf.seasondiff, main = "ACF CO2 Presence in air \n seasonal diff (1959 - 1997)",
    xlab = "Year", ylab = "Co2 ppm", col = "blue")
plot(plot.pacf.seasondiff, main = "PACF CO2 Presence in air \n season difference (1959 - 1997)"
    xlab = "Year", ylab = "Co2 ppm", col = "red", cex.main = 0.5)
```



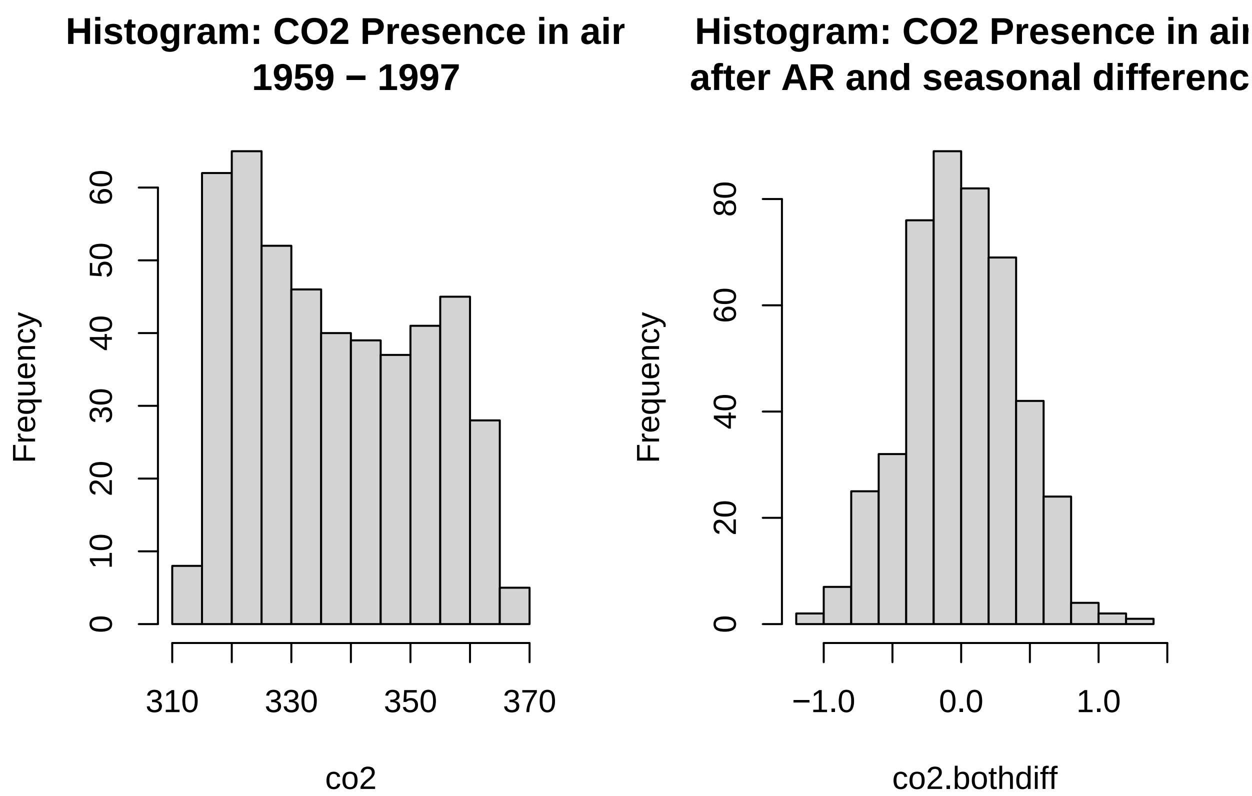

```r
plot(plot.acf.bothdiff, main = "ACF CO2 Presence in air \n AR and seasonal differences",
    xlab = "Year", ylab = "Co2 ppm", col = "blue")
plot(plot.pacf.bothdiff, main = "PACF CO2 Presence in air \n AR and seasonal differences",
    xlab = "Year", ylab = "Co2 ppm", col = "red", cex.main = 0.5)
```





Decomposition graph confirms the findings from EDA, trend and seasonality are present in the time series.
Above ACF and PACF graph shows for different adjustments of time series: 1) original series 2) de-seasoned 3)de-trended 4) random component of time series 5) Two period differenced for trend 5) Two period difference and seasonal differenced time series. Few observations from above graphs
* PACF graph shows autocorrelation dying off at second log after de-seasoned. This suggests to use only 1st order Auto regressive model. This also suggests removing seasonality is important
* ACF graph shows clear seasonal effect after removing trend
* ACF graph after performing auto regressive (AR) and seasonal differences looks closer to white noise ACF graph. This confirms the need for seasonal and Integrated treatment for our model

```r
par(mfrow = c(1, 2))
hist(co2, main = "Histogram: CO2 Presence in air \n 1959 - 1997")
hist(co2.bothdiff, main = "Histogram: CO2 Presence in air\n after AR and seasonal difference")
```



Histogram after applying seasonal and regressive difference looks close to Gaussian distribution.

Part 2 (3 points)

Fit a linear time trend model to the `co2` series, and examine the characteristics of the residuals. Compare this to a higher-order polynomial time trend model. Discuss whether a logarithmic transformation of the data would be appropriate. Fit a polynomial time trend model that incorporates seasonal dummy variables, and use this model to generate forecasts up to the present.

Linear Time Trend Model

```r
# First fit a linear time trend model
par(mfrow = c(3, 1))
co2.ts.lm.linear = lm(co2 ~ time(co2))
summary(co2.ts.lm.linear)
```

```
##
## Call:
```

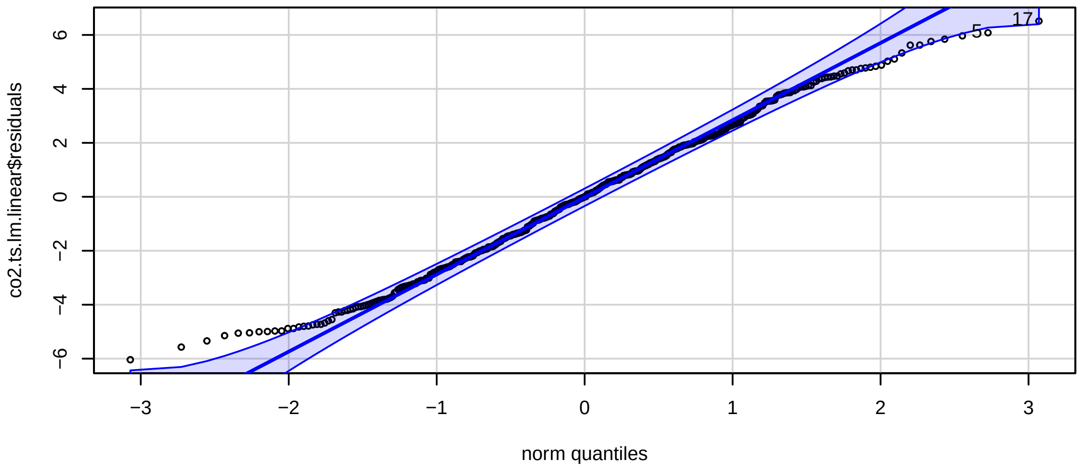
```
## lm(formula = co2 ~ time(co2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0399 -1.9476 -0.0017  1.9113  6.5149
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.250e+03  2.127e+01  -105.8   <2e-16 ***
## time(co2)    1.308e+00  1.075e-02   121.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.618 on 466 degrees of freedom
## Multiple R-squared:  0.9695, Adjusted R-squared:  0.9694
## F-statistic: 1.479e+04 on 1 and 466 DF,  p-value: < 2.2e-16
```

```
qqPlot(co2.ts.lm.linear$residuals, main = expression("Linear Model co2 ~ time(co2) "))
```

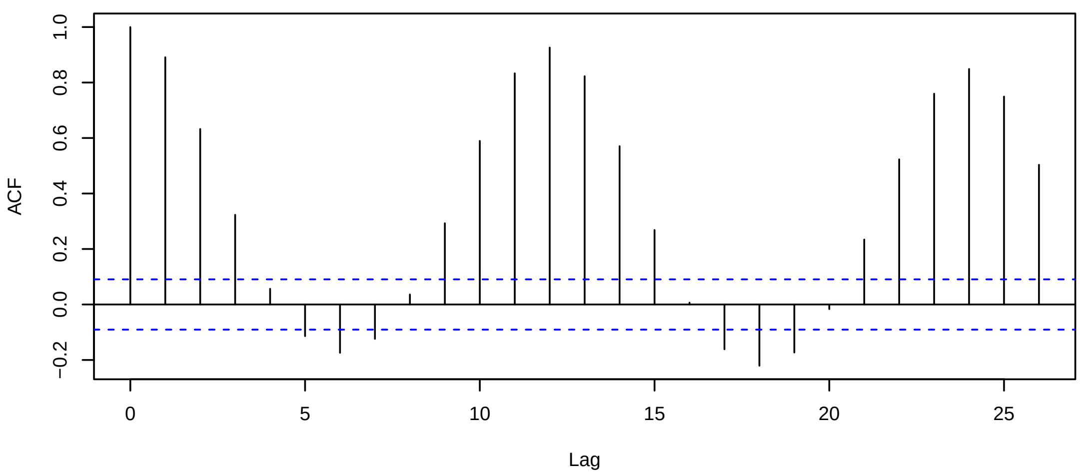
```
## [1] 17  5
```

```
plt.acf = acf(co2.ts.lm.linear$residuals, plot = FALSE)
plt.pacf = pacf(co2.ts.lm.linear$residuals, plot = FALSE)
plot(plt.acf, main = expression("ACF - Linear Model co2 ~ time(co2) "))
plot(plt.pacf, main = expression("PACF - Linear Model co2 ~ time(co2) "))
```

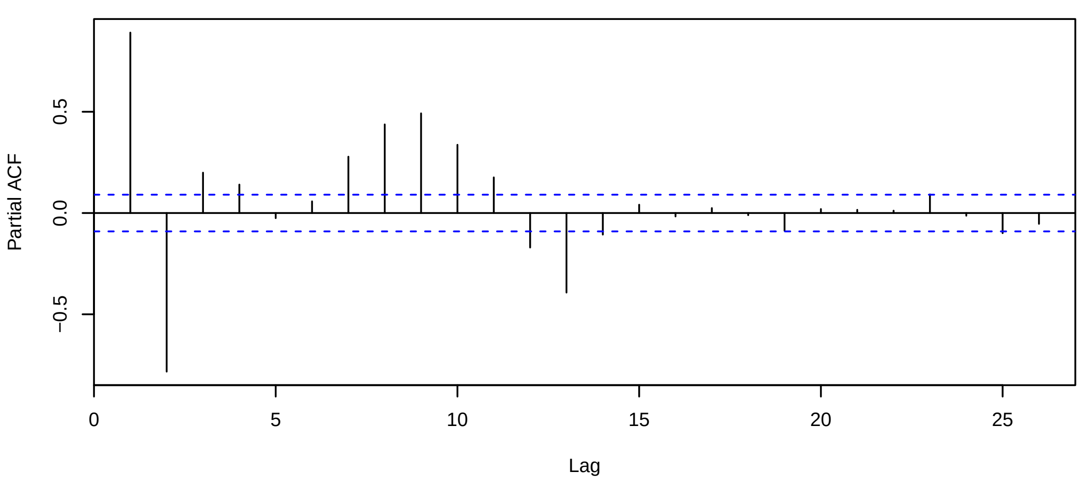
Linear Model co2 ~ time(co2)



ACF − Linear Model co2 ~ time(co2)



PACF − Linear Model co2 ~ time(co2)



```
Box.test(co2.ts.lm.linear$residuals, type = "Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  co2.ts.lm.linear$residuals
## X-squared = 373.94, df = 1, p-value < 2.2e-16
```

Because the variance around the trend line appeared constant, we chose not to take the log of the values of our time series observations.

After fitting a linear model of time, we performed several checks to assess model fit. As seen above, the plot of the residuals against the normal distribution shows skewing in the tails, suggesting that the linear model residuals are not normally distributed.

The ACF and PACF plots do not resemble those of white noise, suggesting poor model fit, and show evidence of autocorrelation in the residuals. This latter finding is supported by the results of the Ljung-Box test, which has a small p-value (< 0.05) - meaning that we fail to reject the null hypothesis that the residuals are correlated.

Seasonal Time-Trend Model

```
# Add seasonal dummy to data.frame
co2.df = data.frame(ppm = c(co2), time = c(time(co2)))
co2.df$season = as.factor(cycle(co2))

par(mfrow = c(3, 1))
co2.ts.lm.stt = lm(ppm ~ time + I(time(co2)^2) + season, data = co2.df)
summary(co2.ts.lm.stt)
```

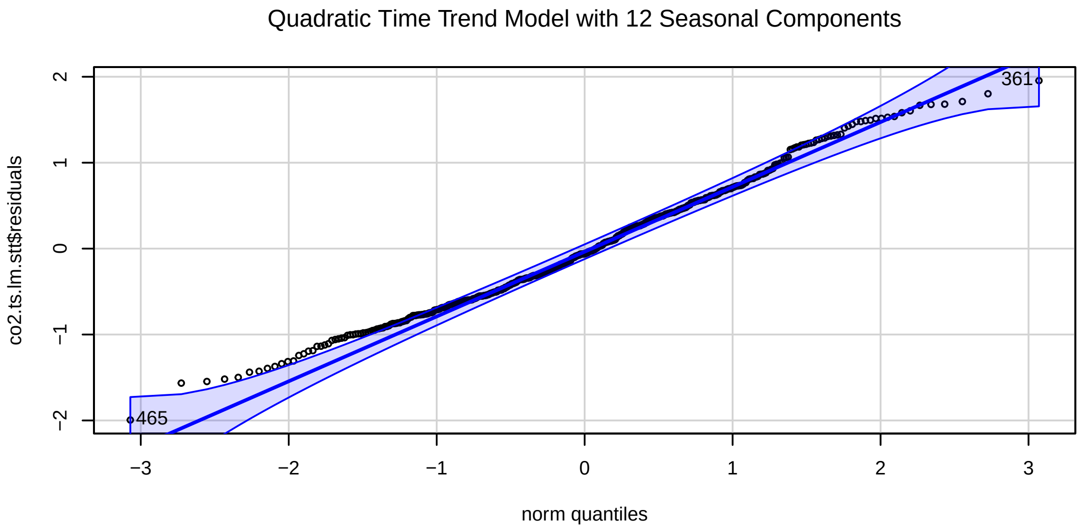
```
##
## Call:
## lm(formula = ppm ~ time + I(time(co2)^2) + season, data = co2.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.99478 -0.54468 -0.06017  0.47265  1.95480
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.771e+04  1.156e+03  41.289  < 2e-16 ***
## time           -4.920e+01  1.168e+00 -42.120  < 2e-16 ***
## I(time(co2)^2)  1.277e-02  2.952e-04  43.242  < 2e-16 ***
## season2         6.642e-01  1.640e-01   4.051 5.99e-05 ***
## season3         1.407e+00  1.640e-01   8.582  < 2e-16 ***
## season4         2.538e+00  1.640e-01  15.480  < 2e-16 ***
```

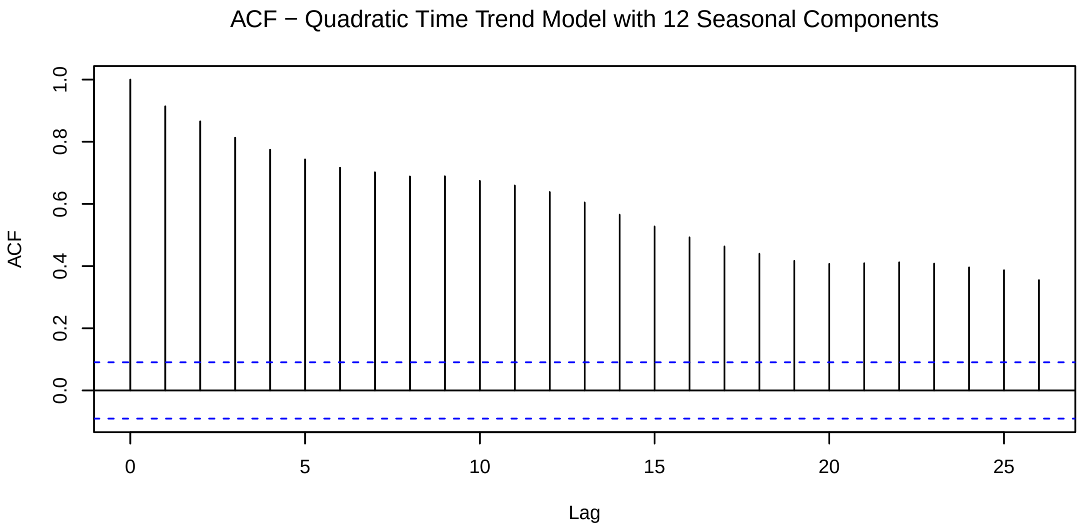
```
## season5         3.017e+00  1.640e-01  18.400  < 2e-16 ***
## season6         2.354e+00  1.640e-01  14.357  < 2e-16 ***
## season7         8.331e-01  1.640e-01   5.081 5.50e-07 ***
## season8        -1.235e+00  1.640e-01  -7.531 2.75e-13 ***
## season9        -3.059e+00  1.640e-01 -18.659  < 2e-16 ***
## season10       -3.243e+00  1.640e-01 -19.777  < 2e-16 ***
## season11       -2.054e+00  1.640e-01 -12.526  < 2e-16 ***
## season12       -9.374e-01  1.640e-01  -5.717 1.97e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.724 on 454 degrees of freedom
## Multiple R-squared:  0.9977, Adjusted R-squared:  0.9977
## F-statistic: 1.531e+04 on 13 and 454 DF,  p-value: < 2.2e-16
```

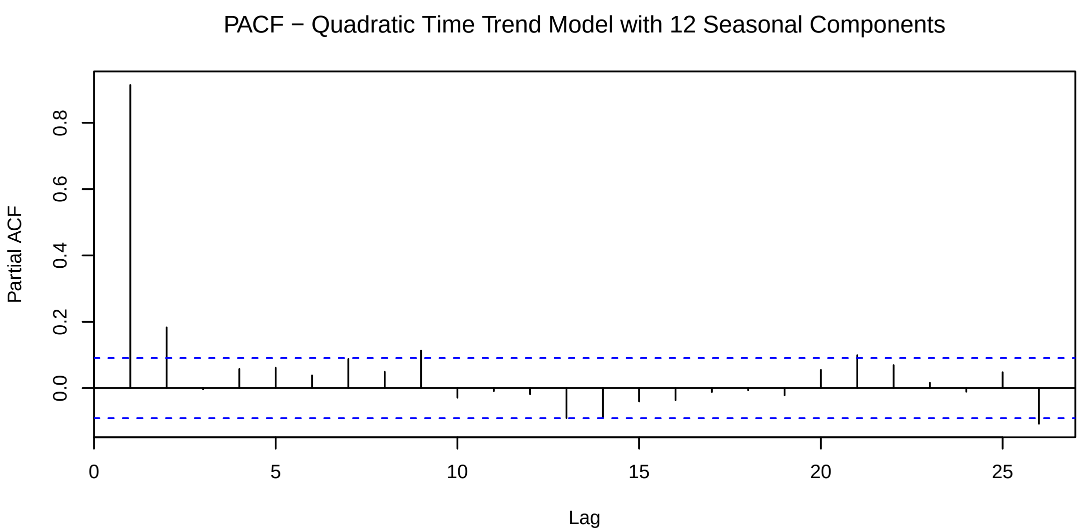
```
qqPlot(co2.ts.lm.stt$residuals, main = expression("Quadratic Time Trend Model with 12 Seasonal
```

```
## [1] 465 361
```

```
plt.acf = acf(co2.ts.lm.stt$residuals, plot = FALSE)
plt.pacf = pacf(co2.ts.lm.stt$residuals, plot = FALSE)
plot(plt.acf, main = expression("ACF - Quadratic Time Trend Model with 12 Seasonal Components")
plot(plt.pacf, main = expression("PACF - Quadratic Time Trend Model with 12 Seasonal Components
```


Quadratic Time Trend Model with 12 Seasonal Components
co2.ts.lm.stt$residuals
norm quantiles
−3
−2
−1
0
1
2
3
465
361


ACF − Quadratic Time Trend Model with 12 Seasonal Components
ACF
Lag
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25


PACF − Quadratic Time Trend Model with 12 Seasonal Components
Partial ACF
Lag
0.0
0.2
0.4
0.6
0.8
0
5
10
15
20
25

```r
Box.test(co2.ts.lm.stt$residuals, type = "Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  co2.ts.lm.stt$residuals
## X-squared = 393.48, df = 1, p-value < 2.2e-16
```

Based upon residual plots, the quadratic model with time and seasonal dummy variables appears to be a better fit. The residual tails are less skewed away from the qqline in the plot against the normal distribution. However, the ACF plot of the residuals, like those of the time linear model, show a trend not captured by our model - the majority of autocorrelations are significant and there is a gradual decay in values over the lags. The PACF shows fewer significant autocorrelations. Again, we find that the model fails the Ljung-Box test, indicating correlation in the residuals.

Despite these inadequacies, the model predictions in the short term do not appear unreasonable, as seen in our forecast plot.

Part 3 (4 points)

Following all appropriate steps, choose an ARIMA model to fit to this `co2` series. Discuss the characteristics of your model and how you selected between alternative ARIMA specifications. Use your model to generate forecasts to the present.

SARIMA Model Selection

```r
# Find the number of seasonal and non-seasonal differences
# needed for stationarity 1 non-seasonal difference and 0
# seasonal differences are required
unitroot_ndiffs(co2)
```

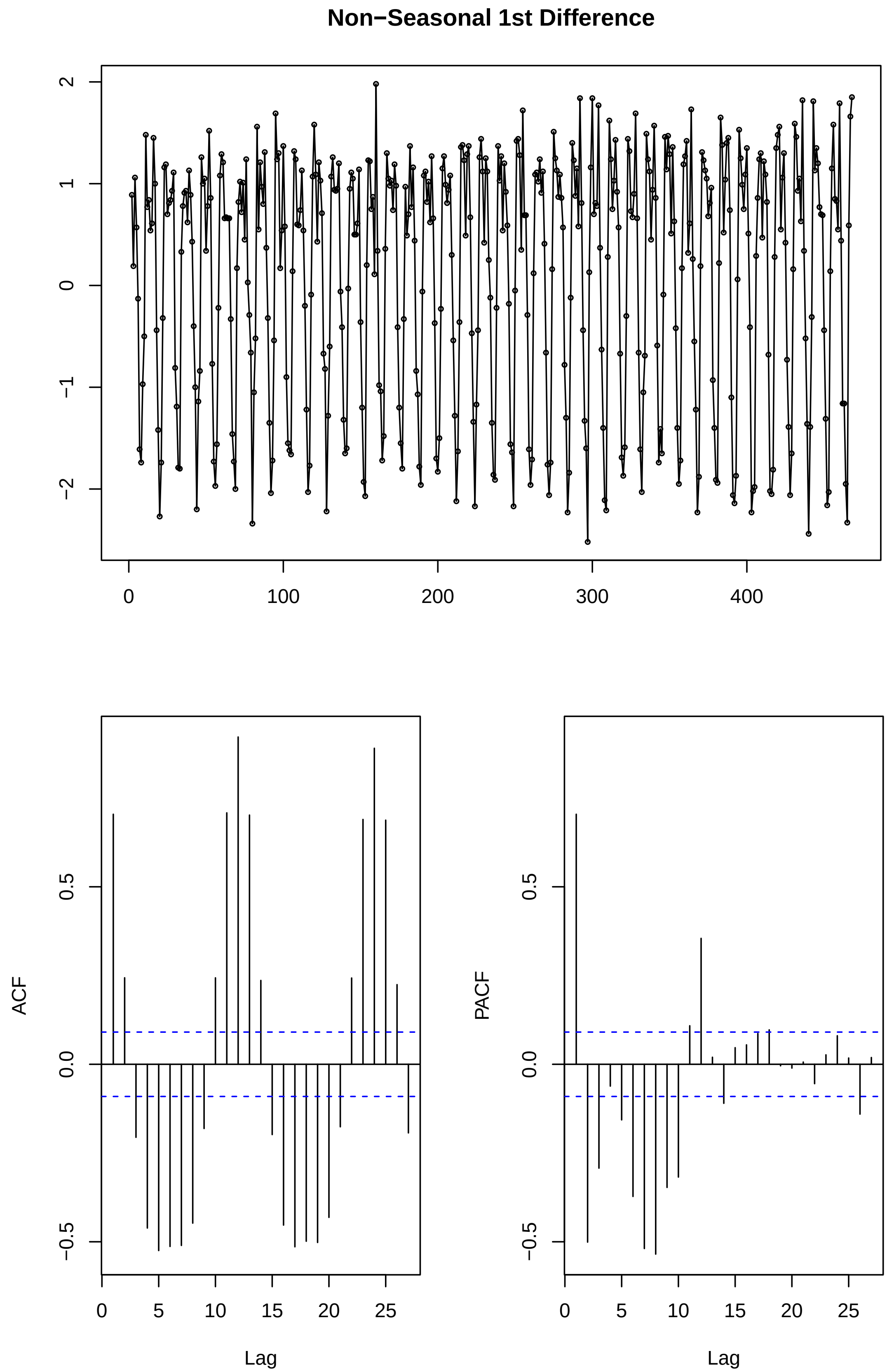
```
## ndiffs
##      1
```

```r
unitroot_nsdiffs(co2)
```

```
## nsdiffs
##       0
```

```r
# Plot the residuals, ACF, and PACF of the
# first-differenced series The PACF chart has fewer
# repeated significant spikes at seasonal lags than the ACF
# does so we'll use it for the seasonal part of the model
# in our initial estimate The PACF only a seasonal spike at
# a lag of 12 - (1,0,0) Since we used the PACF for the
```

```r
# seasonal part, we'll estimate the non-seasonal with the
# ACF The first 2 autocorrelations in the ACF are
# significant, so we'll estimate an MA(2)
tsdisplay(difference(co2), main = "Non-Seasonal 1st Difference")
```


Non−Seasonal 1st Difference
ACF
PACF
Lag
Lag

```r
# Create an Arima model based upon our observations
co2.sarima = arima(co2, order = c(0, 1, 2), seas = list(order = c(1,
    0, 0), frequency(co2)), method = "CSS")
```

The above model can be expressed as auto-regressive equation of
$x_t = x_{t-1} + ($ 0.9803824 $) * x_{t-12} + w_t + ($ -0.3501415 $) * w_{t-1} + ($ -0.0577398 $) * w_{t-2}$
where x_{t-12} represents 12th lag of time series and x_{t-1} is the results of first difference of time series i.e. $x_t^1 = x_t - x_{t-1}$
w_t is white noise from current time step, w_{t-1} is white noise from the previous time step and w_{t-2} is the white noise from 2 steps before. This is the result of moving average component of our model.

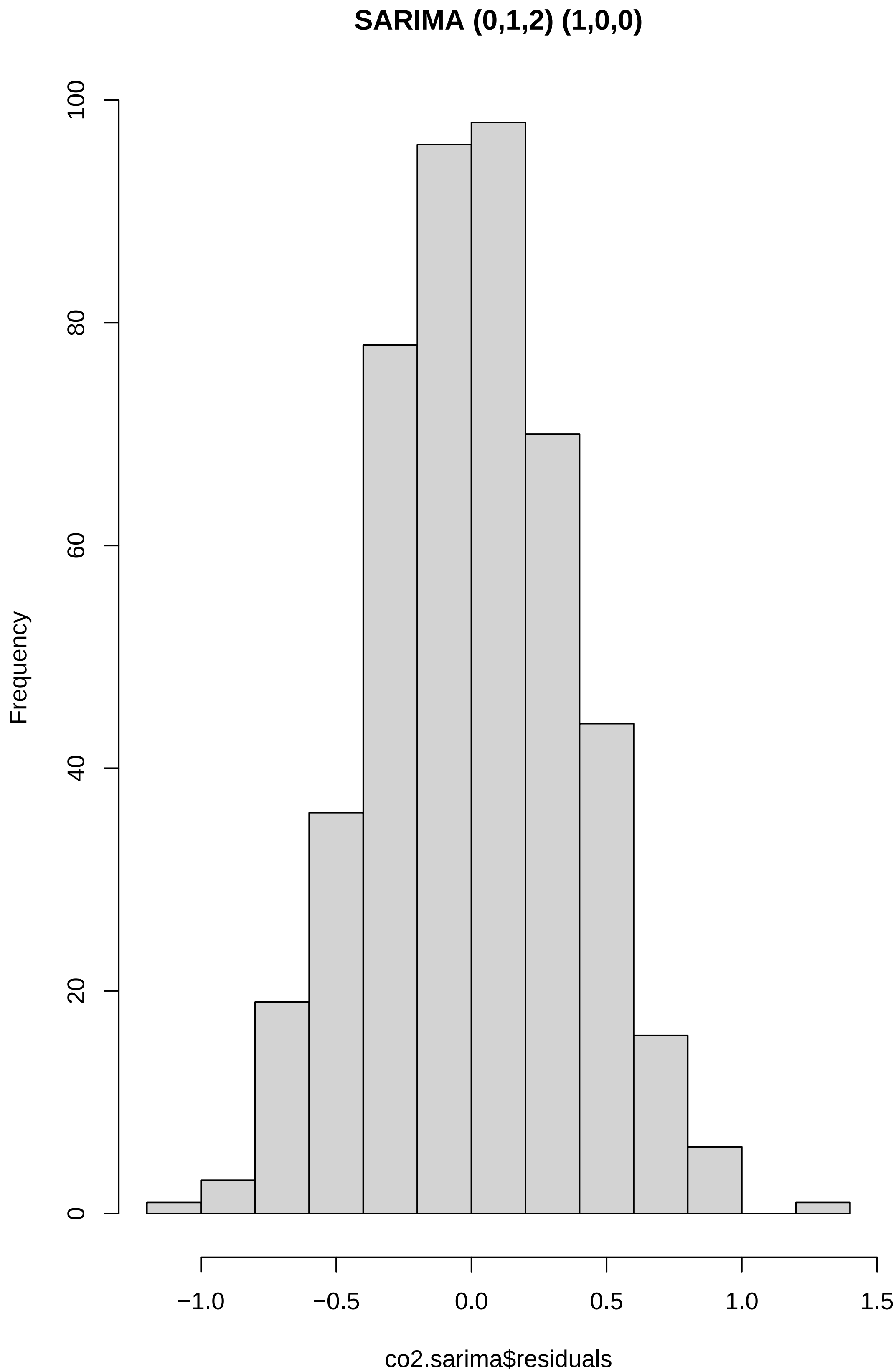
```r
# Find the AIC of the Arima model, check the residuals, and
# perform Ljung-Box
co2.sarima.aicc <- -2 * co2.sarima$loglik + log(length(co2) +
    1) * (length(co2.sarima$coef))
co2.sarima.aicc
```

```
## [1] 413.4629
```

```r
# Look at the estimated coefficients
summary(co2.sarima)
```

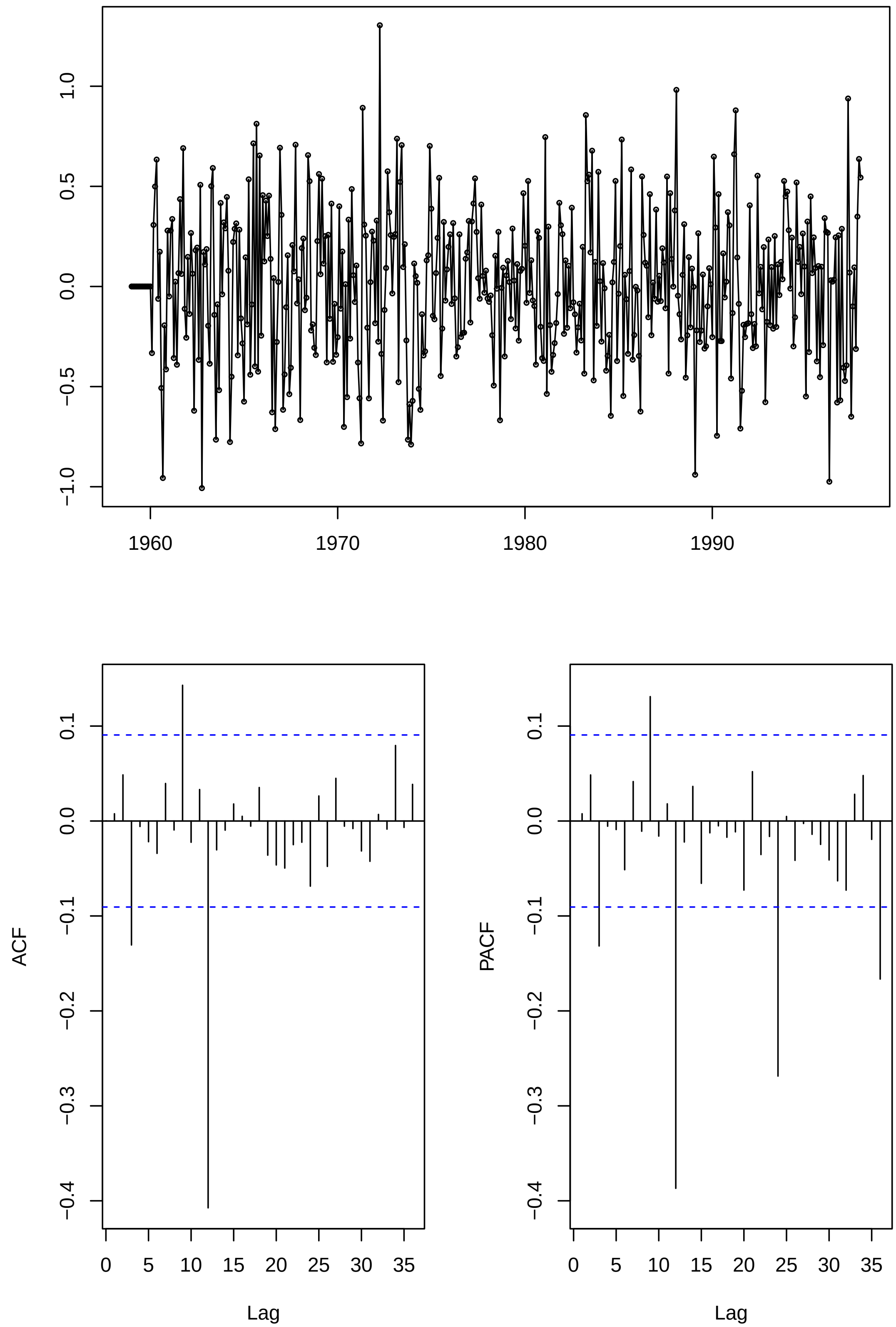
```
##
## Call:
## arima(x = co2, order = c(0, 1, 2), seasonal = list(order = c(1, 0, 0), frequency(co2)),
##     method = "CSS")
##
## Coefficients:
##           ma1      ma2    sar1
##       -0.3501  -0.0577  0.9804
## s.e.   0.0462   0.0444  0.0108
##
## sigma^2 estimated as 0.1364:  part log likelihood = -197.51
##
## Training set error measures:
##                      ME      RMSE       MAE         MPE       MAPE      MASE
## Training set 0.00639654 0.3641826 0.2888305 0.001826364 0.08591893 0.2683615
##                     ACF1
## Training set 0.007648558
```

```r
# The histogram plot looks approximately normal
hist(co2.sarima$residuals, main = "SARIMA (0,1,2) (1,0,0)")
```


SARIMA (0,1,2) (1,0,0)
Frequency
0
20
40
60
80
100
−1.0
−0.5
0.0
0.5
1.0
1.5
co2.sarima$residuals

```r
# A time series plot of the residuals appears to have a
# constant mean The ACF and PACF plots still have a few
# significant autocorrelations
tsdisplay(co2.sarima$residuals, main = "SARIMA (0,1,2) (1,0,0)")
```

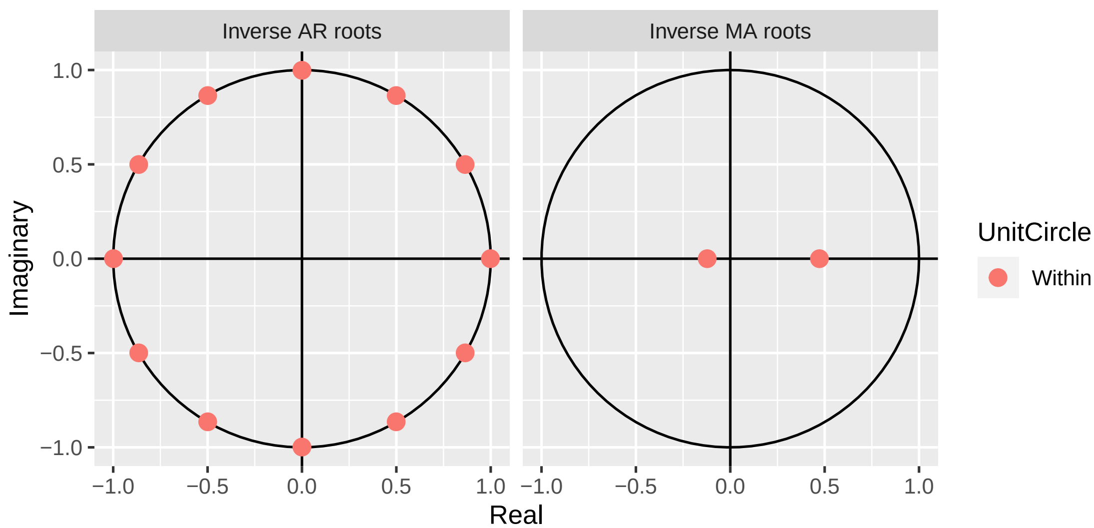
SARIMA (0,1,2) (1,0,0)



```r
# However, the model passes the Ljung-Box test
Box.test(co2.sarima$residuals, type = "Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  co2.sarima$residuals
## X-squared = 0.027554, df = 1, p-value = 0.8682
```

```r
# Check the inverse unit roots for stationarity The inverse
# unit roots are near non-stationarity
autoplot(co2.sarima)
```


Inverse AR roots
Inverse MA roots
1.0
0.5
0.0
−0.5
−1.0
Imaginary
−1.0
−0.5
0.0
0.5
1.0
Real
UnitCircle
Within

To create

our initial model, we first ran unit root tests to check the number of seasonal and non-seasonal differences required for stationarity. These tests returned 1 non-seasonal difference and 0 seasonal differences required, so we used these values as our d and D to estimate our initial Arima model. To obtain p, q, P, and Q, we took a first non-seasonal difference and plotted the ACF, PACF, and differenced values as a time series. The time series plot of the differenced values appeared relatively stationary. The ACF and PACF still showed evidence of autocorrelation. Since the PACF had fewer repeating seasonal lags, we used this plot to estimate the seasonal part of the Arima model. The PACF plot showed a significant autocorrelation at only the first seasonal lag, at 12, so we estimated $(1, 0, 0)$ for the seasonal part of the model. For the non-seasonal part of the Arima model, the ACF showed significant autocorrelation at lags 1 and 2, so we estimated an MA model of order 2, or $(0, 1, 2)$ for the non-seasonal component (with a difference of 1 since we took 1 non-seasonal difference).

The ACF and PACF plots of the residuals of this estimated model ($(0, 1, 2)(1, 0, 0)_{12}$) shows several significant autocorrelations (notably at 1 year in the ACF and PACF and at 2 years in the PACF), although the majority of values fall within the confidence interval for white noise values.

The Ljung-Box test shows a p-value > 0.05, meaning that we reject the null hypothesis that the residuals are auto-correlated.

Since the ACF and PACF plots still showed several strong autocorrelations and the plot of the inverse unit roots showed values near unity, we proceeded to iterate over model parameters to see if we could improve the AIC score and create a model with residuals that better approximated white noise.

Model Selection Algorithm

```
get.best.arima <- function(x.ts, maxord = c(1, 1, 1, 1, 1, 1)) {
    best.aic <- 1e+08
    df.results = data.frame()
    n <- length(x.ts)
    for (p in 0:maxord[1]) for (d in 0:maxord[2]) for (q in 0:maxord[3]) for (P in 0:maxord[4]
        fit <- arima(x.ts, order = c(p, d, q), seas = list(order = c(P,
            D, Q), frequency(x.ts)), method = "CSS")
        # consistent AIC
        fit.aicc <- -2 * fit$loglik + (log(n) + 1) * length(fit$coef)
        # regular AIC
        fit.aic <- -2 * fit$loglik + 2 * (length(fit$coef) +
            1)
        # BIC
        fit.bic <- -2 * fit$loglik + log(n) * (length(fit$coef) +
            1)
        df <- data.frame(model = paste(p, d, q, P, D, Q), AICc = fit.aicc,
            AIC = fit.aic, BIC = fit.bic)
        df.results <- rbind(df.results, df)
    }
    # list(best.aic, best.fit, best.model)
    df.results
```

```
}

arima.search <- get.best.arima(co2, maxord = c(2, 2, 2, 2, 2,
    2))
```

To find a parsimonious seasonal Arima model that better fit the time series, we looped over values in the range of 0 to 2 for the parameters p, q, P, and Q. We also chose the range of 0 to 2 for the number of seasonal and non-seasonal differences, since differencing beyond order 2 is rarely required.

For the best fit model, we chose to use the model with the lowest AICc, as seen in our table below (using AICc since it penalizes the model fit with increasing parameters and corrects for the bias in predictor selection introduced by AIC). As seen below, the best fitting model is $(0, 1, 1)(1, 1, 2)$.

Table 1: Top 10 Models.

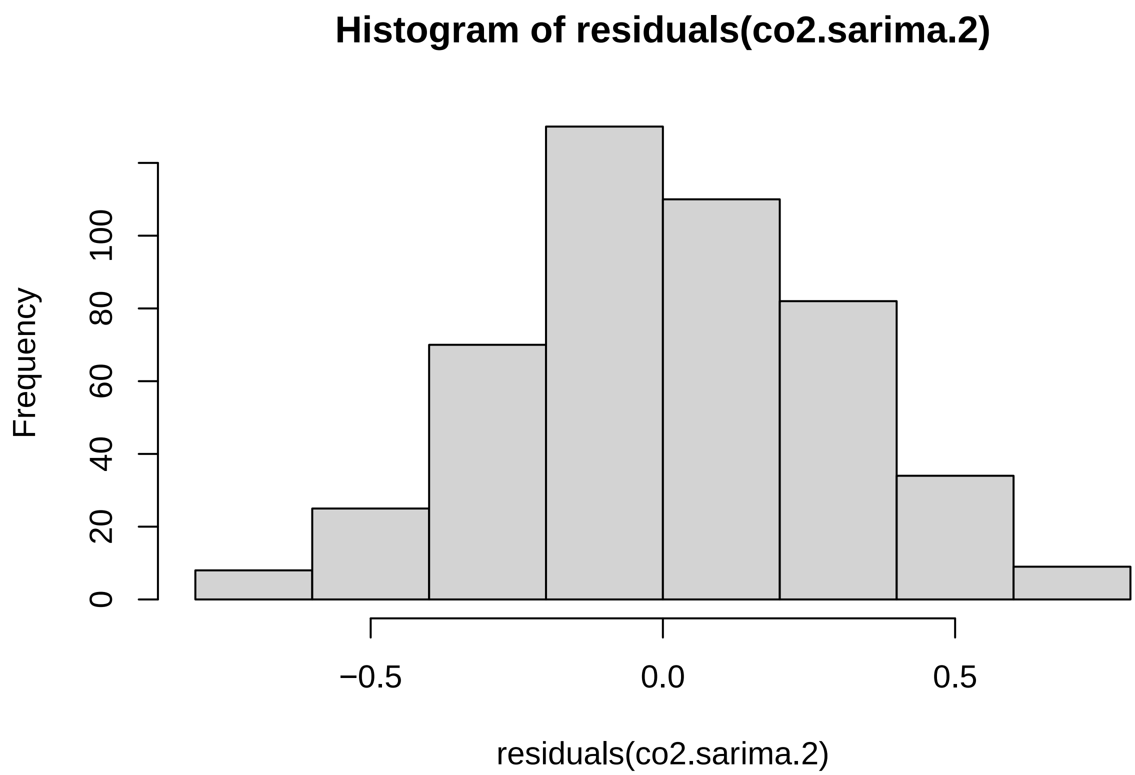
model	AICc	AIC	BIC
0 1 1 1 1 2	193.5103	174.9164	195.6587
0 1 1 2 0 2	196.8432	173.1008	197.9916
0 1 1 1 1 1	197.0927	183.6473	200.2412
1 1 1 1 1 2	197.6924	173.9500	198.8408
0 1 2 1 1 2	199.2472	175.5049	200.3957
1 0 1 1 1 2	199.8742	176.1318	201.0227
1 1 0 1 1 2	201.1810	182.5871	203.3294
1 1 1 1 1 1	201.1950	182.6012	203.3435
1 1 1 2 0 2	201.2614	172.3706	201.4099
0 1 1 2 1 2	201.3117	177.5693	202.4602

```
# Estimate an Arima model with the parameters of the model
# with the lowest AICc found from our parameter search
pdqPDQ <- as.list(unlist(strsplit(best10.arima[1, 1], "[[:space:]]")))
p <- strtoi(pdqPDQ[[1]])
d <- strtoi(pdqPDQ[[2]])
q <- strtoi(pdqPDQ[[3]])
P <- strtoi(pdqPDQ[[4]])
D <- strtoi(pdqPDQ[[5]])
Q <- strtoi(pdqPDQ[[6]])

# Estimate the model
co2.sarima.2 <- arima(co2, order = c(p, d, q), seasonal = list(order = c(P,
    D, Q)), method = "CSS")
```

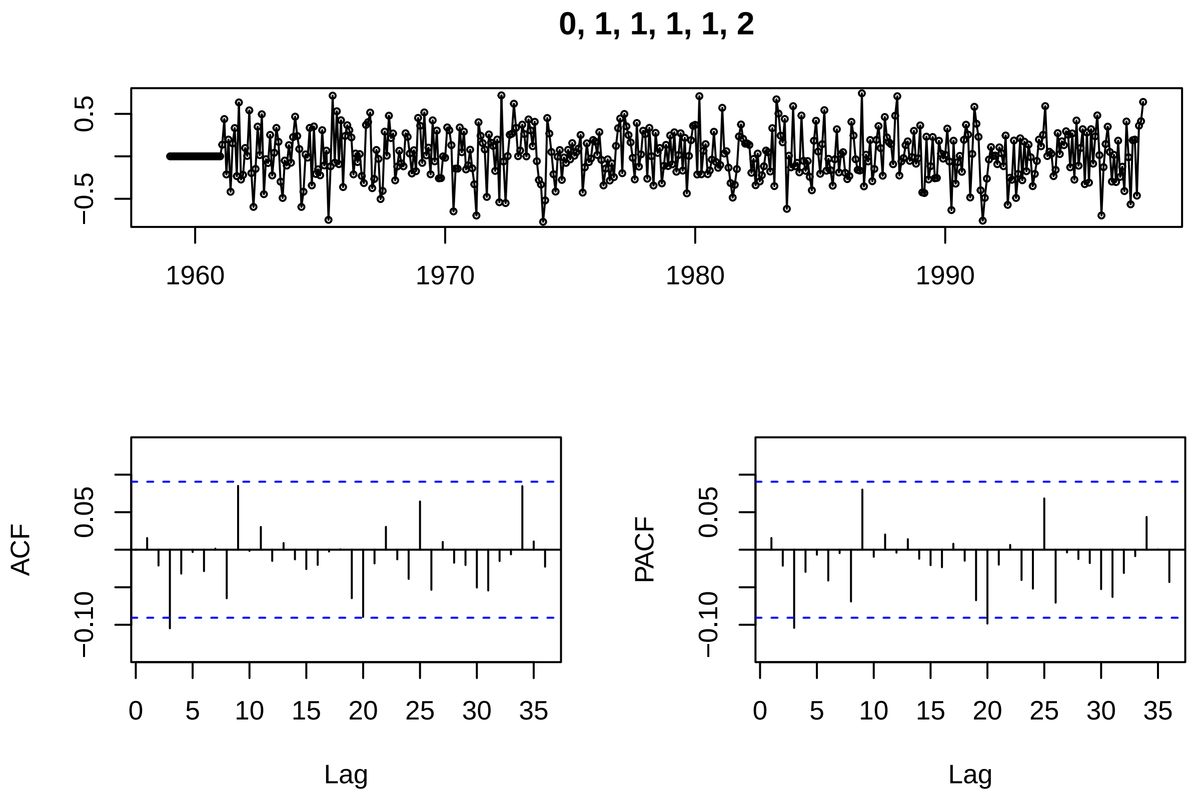
Our best sarima model can be expressed as below
$x_t = x_{t-1} +$ (-0.380699) $* x_{t-12} + w_t +$ (-0.3600677) $* w_{t-1} +$ (-0.4157137) $* w_{t-12} +$ (-0.3558124) $* w_{t-13}$
where x_{t-12} represents 12th lag of time series and x_{t-1} is the results of first difference of time series

i.e. $x_t^1 = x_t - x_{t-1}$
w_t is white noise from current time step, w_{t-1} is white noise from the previous time step, which is the result of AR moving average. w_{t-12} is the white noise from 12 steps before (seasonal) current time step and w_{t-13} is the white noise from 13 steps before current time step. This is the result of seasonal moving average component of our model.

```
# Inspect the residual plots and find the estimated AICc
sarima2.aicc <- -2 * co2.sarima.2$loglik + (log(length(co2)) +
    1) * length(co2.sarima.2$coef)
hist(residuals(co2.sarima.2))
```




```
tsdisplay(co2.sarima.2$residuals, main = {
    toString(pdqPDQ)
})
```



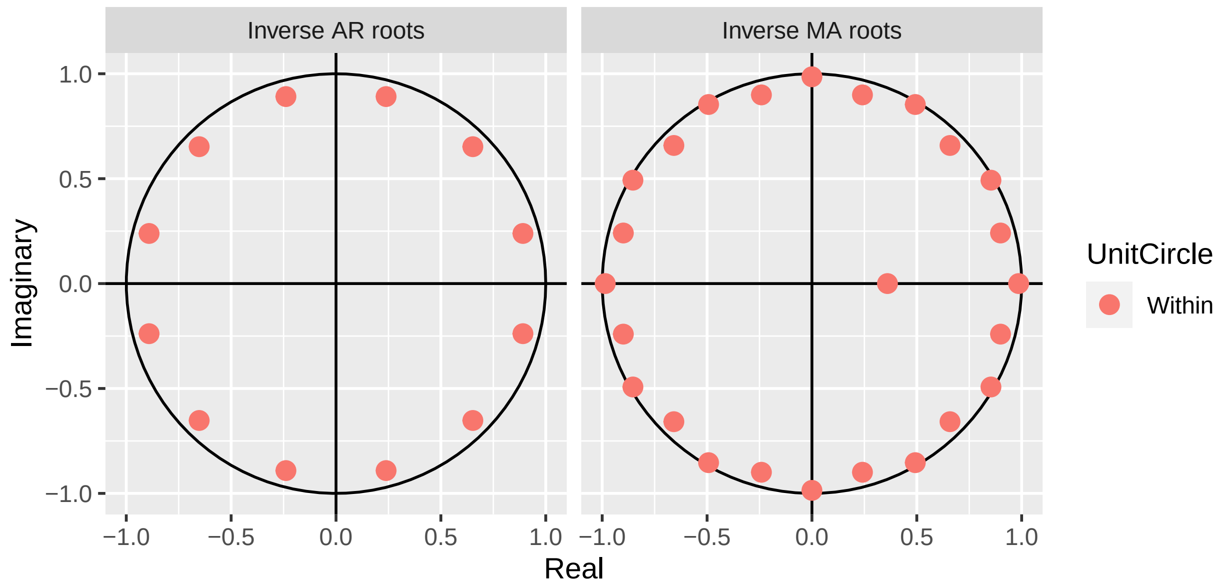

```r
sarima2.aicc
```

```
## [1] 193.5103
```

```r
Box.test(co2.sarima.2$residuals, type = "Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  co2.sarima.2$residuals
## X-squared = 0.11422, df = 1, p-value = 0.7354
```

```r
autoplot(co2.sarima.2)
```



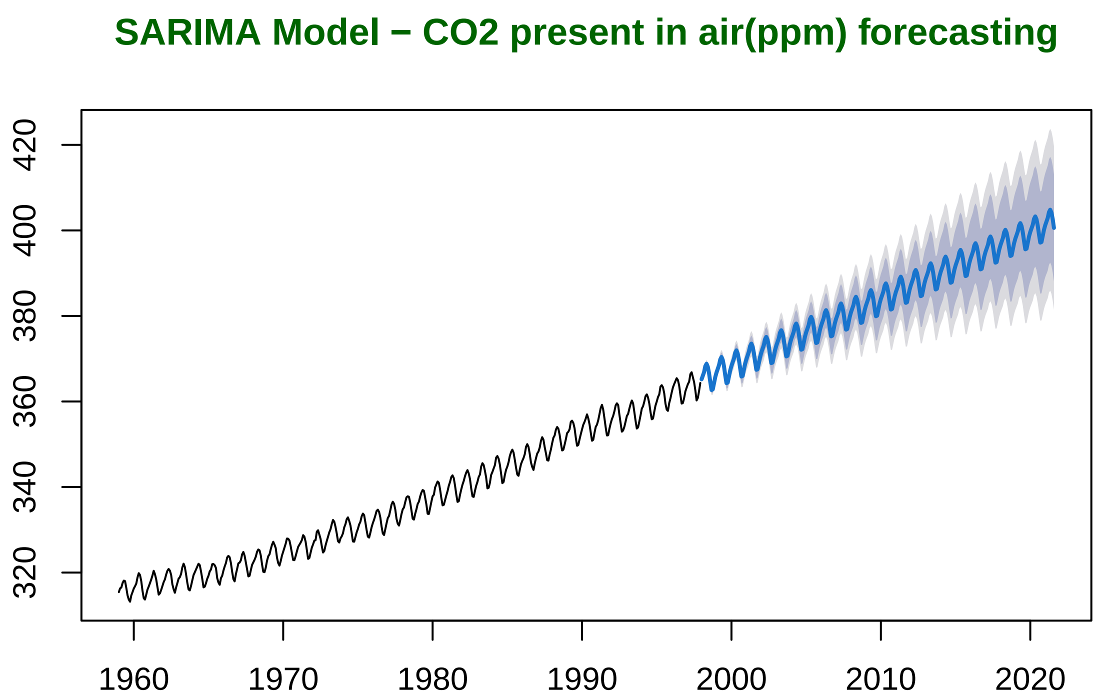
The AICc value is smaller than that of our initial model estimate, and the majority of ACF and PACF values fall within the 95% confidence interval bounds for white noise. In addition, the Ljung-Box test indicates that the data are independently distributed since we fail to reject the null hypothesis.

The histogram of the residuals shows them to be approximately normally distributed and the plot of the residuals as a time series resembles white noise.

Since this model has a lower AICc than our initial estimate, the residuals resemble white noise, and we have not found significant evidence of residual autocorrelation, we proceed with using this model in our forecast. As seen in the plots of the inverse unit roots, the absolute value of the inverse unit roots are less than unity, meaning that the residuals are stationary.

Best Model Forecasts

```r
co2.forecast <- forecast(co2.sarima.2, 284)
co2_forecast_ts <- co2.forecast[4]$mean
plot(co2.forecast, main = "SARIMA Model - CO2 present in air(ppm) forecasting",
    col.main = "darkgreen")
```



Part 4 (5 points)

The file `co2_weekly_mlo.txt` contains weekly observations of atmospheric carbon dioxide concentrations measured at the Mauna Loa Observatory from 1974 to 2020, published by the National Oceanic and Atmospheric Administration (NOAA). Convert these data into a suitable time series object, conduct a thorough EDA on the data, addressing the problem of missing observations and comparing the Keeling Curve's development to your predictions from Parts 2 and 3. Use the weekly data to generate a month-average series from 1997 to the present and use this to generate accuracy metrics for the forecasts generated by your models from Parts 2 and 3.

```
co2_weekly <- read.table("co2_weekly_mlo.txt", header = FALSE)
colnames(co2_weekly) <- c("year", "month", "day", "decimal",
    "ppm", "days", "1yr_ago", "10yrs_ago", "since1800")
summary(co2_weekly)
```

```
##       year          month             day           decimal
##  Min.   :1974   Min.   : 1.00   Min.   : 1.00   Min.   :1974
##  1st Qu.:1986   1st Qu.: 4.00   1st Qu.: 8.00   1st Qu.:1986
##  Median :1997   Median : 7.00   Median :16.00   Median :1998
##  Mean   :1997   Mean   : 6.52   Mean   :15.72   Mean   :1998
##  3rd Qu.:2009   3rd Qu.:10.00   3rd Qu.:23.00   3rd Qu.:2010
##  Max.   :2021   Max.   :12.00   Max.   :31.00   Max.   :2021
##       ppm               days           1yr_ago          10yrs_ago
```

```
##  Min.   :-1000.0   Min.   :0.000   Min.   :-1000.0   Min.   : -999.99
##  1st Qu.:  347.1   1st Qu.:5.000   1st Qu.:  345.6   1st Qu.:  331.48
##  Median :  365.2   Median :6.000   Median :  363.5   Median :  350.18
##  Mean   :  358.3   Mean   :5.871   Mean   :  328.4   Mean   :   59.61
##  3rd Qu.:  388.4   3rd Qu.:7.000   3rd Qu.:  386.2   3rd Qu.:  368.45
##  Max.   :  420.0   Max.   :7.000   Max.   :  417.8   Max.   :  395.23
##    since1800
##  Min.   : -999.99
##  1st Qu.:   66.95
##  Median :   84.55
##  Mean   :   80.38
##  3rd Qu.:  108.07
##  Max.   :  136.87
```

```
describe(co2_weekly)
```

```
## co2_weekly
##
##  9  Variables      2458  Observations
## ---------------------------------------------------------------------------------
## year
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2458        0       48        1     1997    15.71     1976     1979
##      .25      .50      .75      .90      .95
##     1986     1997     2009     2016     2019
##
## lowest : 1974 1975 1976 1977 1978, highest: 2017 2018 2019 2020 2021
## ---------------------------------------------------------------------------------
## month
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2458        0       12    0.993     6.52    3.965        1        2
##      .25      .50      .75      .90      .95
##        4        7       10       11       12
##
## lowest :  1  2  3  4  5, highest:  8  9 10 11 12
##
## Value          1     2     3     4     5     6     7     8     9    10    11
## Frequency    208   190   208   201   211   205   208   208   202   207   202
## Proportion 0.085 0.077 0.085 0.082 0.086 0.083 0.085 0.085 0.082 0.084 0.082
##
## Value         12
## Frequency    208
## Proportion 0.085
## ---------------------------------------------------------------------------------
## day
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2458        0       31    0.999    15.72    10.16        2        4
```

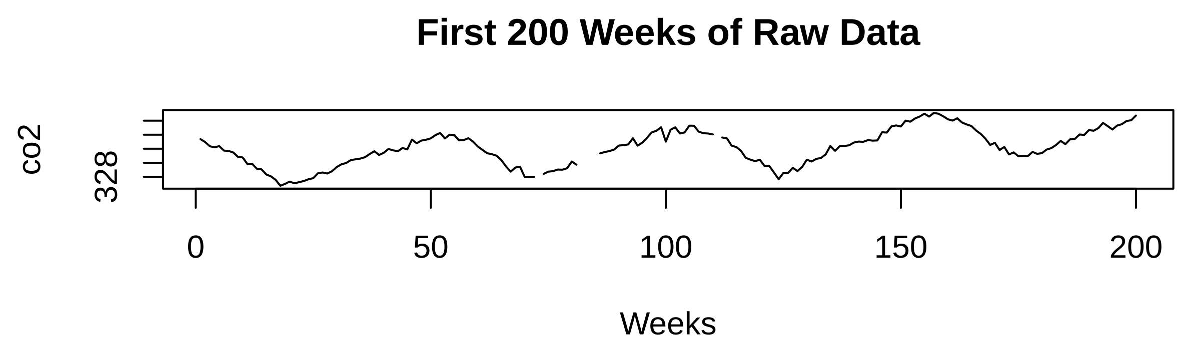
```
##      .25      .50      .75      .90      .95
##        8       16       23       28       29
##
## lowest :  1  2  3  4  5, highest: 27 28 29 30 31
## --------------------------------------------------------------------------------
## decimal
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2458        0     2458        1     1998    15.71     1977     1979
##      .25      .50      .75      .90      .95
##     1986     1998     2010     2017     2019
##
## lowest : 1974.380 1974.399 1974.418 1974.437 1974.456
## highest: 2021.390 2021.410 2021.429 2021.448 2021.467
## --------------------------------------------------------------------------------
## ppm
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2458        0     2148        1    358.3    47.87    332.4    336.1
##      .25      .50      .75      .90      .95
##    347.1    365.2    388.4    404.6    410.6
##
## lowest : -999.99  326.72  326.99  327.07  327.23
## highest:  419.28  419.47  419.53  419.55  420.01
##
## Value      -1000   320   340   360   380   400   420
## Frequency     18    45   638   662   527   435   133
## Proportion 0.007 0.018 0.260 0.269 0.214 0.177 0.054
##
## For the frequency table, variable is rounded to the nearest 20
## --------------------------------------------------------------------------------
## days
##        n  missing distinct     Info     Mean      Gmd
##     2458        0        8    0.896    5.871    1.378
##
## lowest : 0 1 2 3 4, highest: 3 4 5 6 7
##
## Value          0     1     2     3     4     5     6     7
## Frequency     18    14    36   101   176   402   648  1063
## Proportion 0.007 0.006 0.015 0.041 0.072 0.164 0.264 0.432
## --------------------------------------------------------------------------------
## 1yr_ago
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2458        0     2097        1    328.4    101.7    330.5    334.4
##      .25      .50      .75      .90      .95
##    345.6    363.5    386.2    402.0    408.2
##
## lowest : -999.99  326.73  326.84  326.98  327.21
## highest:  417.09  417.10  417.21  417.46  417.83
##
```

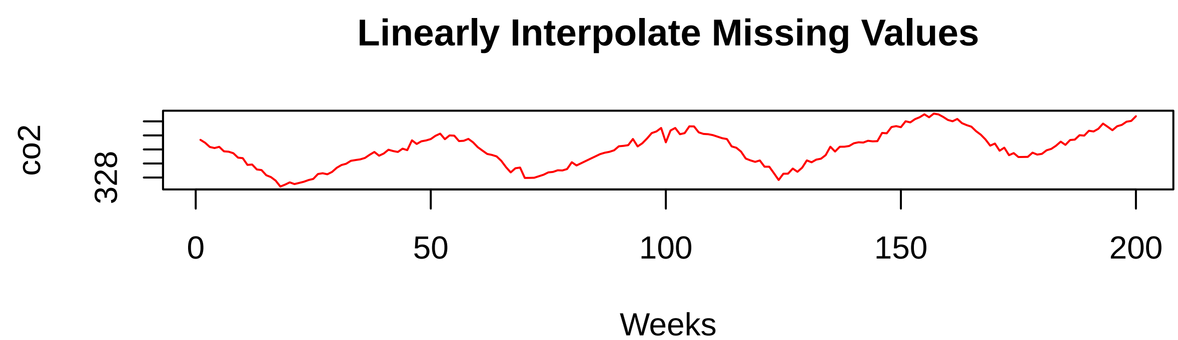
```
## Value      -1000   320   340   360   380   400   420
## Frequency     70    45   638   665   523   436    81
## Proportion 0.028 0.018 0.260 0.271 0.213 0.177 0.033
##
## For the frequency table, variable is rounded to the nearest 20
## --------------------------------------------------------------------------------
## 10yrs_ago
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2458        0     1644    0.989    59.61    479.1  -1000.0  -1000.0
##      .25      .50      .75      .90      .95
##    331.5    350.2    368.5    382.4    387.0
##
## lowest : -999.99  326.66  327.04  327.10  327.26
## highest:  394.08  394.15  394.43  395.13  395.23
##
## Value      -1000   330   340   350   360   370   380   390   400
## Frequency    541   196   328   343   339   286   248   175     2
## Proportion 0.220 0.080 0.133 0.140 0.138 0.116 0.101 0.071 0.001
##
## For the frequency table, variable is rounded to the nearest 10
## --------------------------------------------------------------------------------
## since1800
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     2458        0     2086        1    80.38    43.66    52.11    55.81
##      .25      .50      .75      .90      .95
##    66.95    84.55   108.07   125.10   130.75
##
## lowest : -999.99   49.60   49.65   49.72   49.95
## highest:  136.49  136.61  136.64  136.74  136.87
##
## Value      -1000    50    60    70    80    90   100   110   120   130   140
## Frequency     18   194   326   325   371   270   260   245   200   216    33
## Proportion 0.007 0.079 0.133 0.132 0.151 0.110 0.106 0.100 0.081 0.088 0.013
##
## For the frequency table, variable is rounded to the nearest 10
## --------------------------------------------------------------------------------
```

NOAA data provided in the file has 2458 weekly observations from 1974 to 2021 with 10 variables. Variable `ppm` tracks weekly co2 presence. We will be using `ppm` values for our analysis. It appears that NOAA uses -999 to represent missing values. For `ppm`, there are 18 observations missing. and we have 18 observations that have `ppm` value as a null, we will fill them in before developing time series model.

Impute Missing Values Linearly

```r
co2_weekly <- co2_weekly %>%
    mutate(ppm = ifelse(test = (ppm <= 0), NA, no = ppm))
co2_weekly2 <- data.frame(lapply(co2_weekly, function(X) approxfun(seq_along(X),
    X)(seq_along(X))))
par(mfrow = c(2, 1))
plot(co2_weekly$ppm[1:200], type = "l", xlab = "Weeks", ylab = "co2",
    main = "First 200 Weeks of Raw Data")
plot(co2_weekly2$ppm[1:200], type = "l", col = "red", xlab = "Weeks",
    ylab = "co2", main = "Linearly Interpolate Missing Values")
```

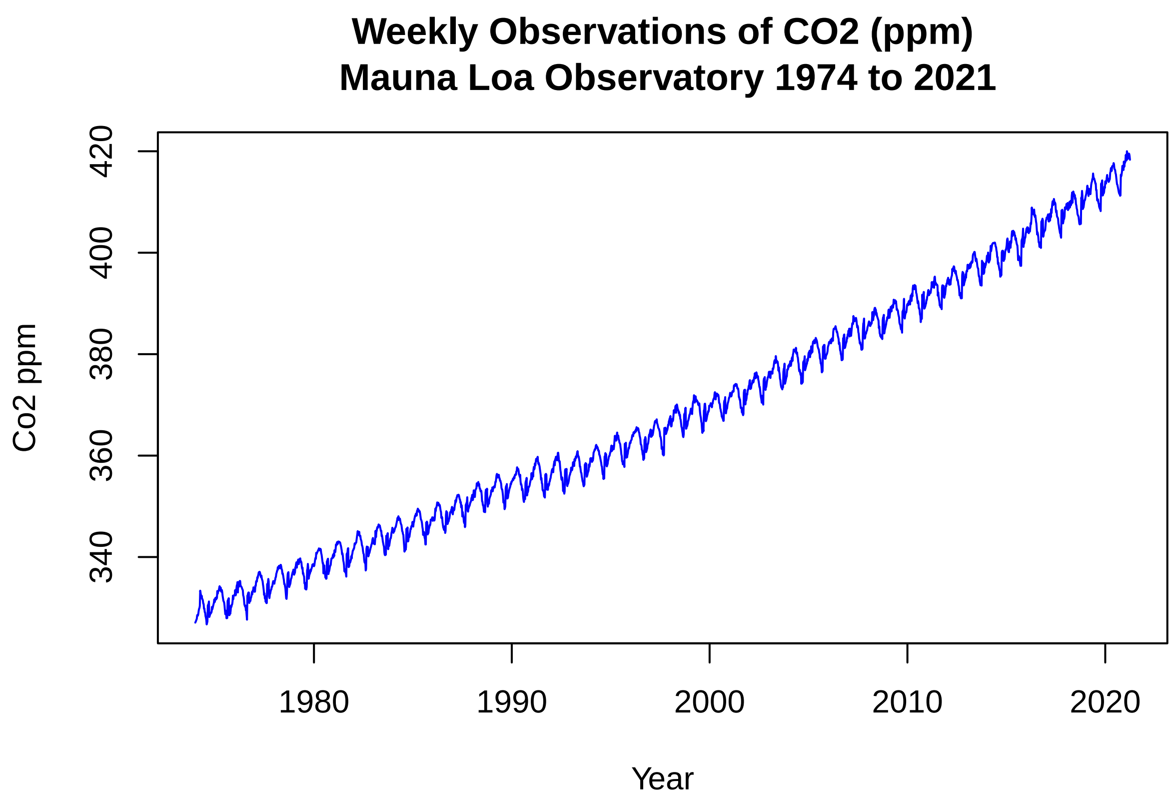




After careful observation of the data, most of the missing points are spread out across the data set (i.e. we do not need to impute 18 weeks in a row). As a result, we suggest it is reasonable to simply interpolate the missing values linearly. The plot above shows the first 200 weeks of the original data series with missing data and a new time series with missing values imputed.

```r
# Get monthly averages for replacement after imputing
# missing values
co2_monthly <- co2_weekly2 %>%
    group_by(year, month) %>%
    summarise(ppm_month_avg = mean(ppm))

# join to add monthly averages
```

```r
co2_merged <- merge(co2_weekly2, co2_monthly, by = c("year",
    "month"))

# Create weekly time series
co2_noaa_weekly_ts <- ts(co2_merged$ppm, start = c(1974), frequency = 52)

# Plot weekly time series
plot(co2_noaa_weekly_ts, main = "Weekly Observations of CO2 (ppm)\n Mauna Loa Observatory 1974
    xlab = "Year", ylab = "Co2 ppm", col = "blue")
```

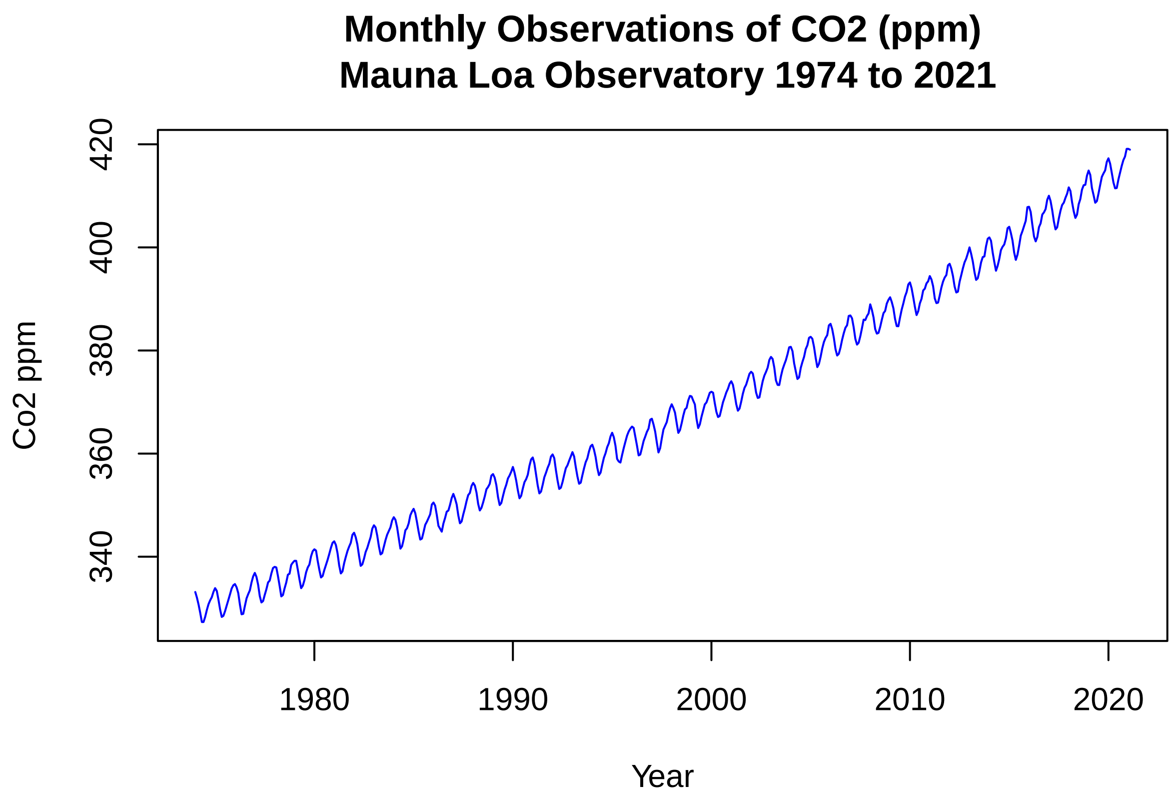



```r
# Calculate monthly averages as our forecast is only on
# monthly basis
co2_noaa_monthly_df <- co2_merged %>%
    group_by(year, month) %>%
    summarise(ppm_month_avg = mean(ppm))
summary(co2_noaa_monthly_df)
```

```
##       year          month        ppm_month_avg
##  Min.   :1974   Min.   : 1.000   Min.   :327.3
##  1st Qu.:1986   1st Qu.: 4.000   1st Qu.:347.2
##  Median :1997   Median : 6.000   Median :365.1
##  Mean   :1997   Mean   : 6.496   Mean   :368.2
```

```
##   3rd Qu.:2009   3rd Qu.: 9.000   3rd Qu.:388.1
##   Max.   :2021   Max.   :12.000   Max.   :419.1
```

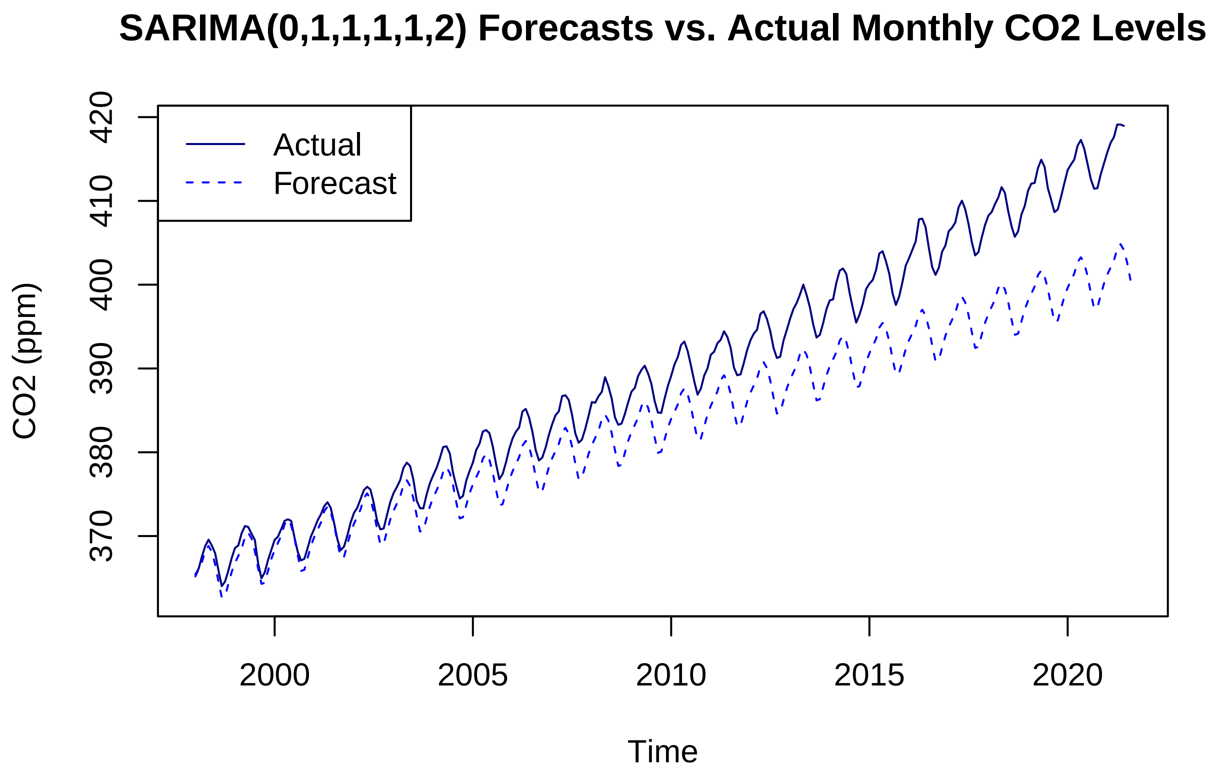
```r
# Create monthly ts object (all observations)
co2_noaa_monthly_ts <- ts(co2_noaa_monthly_df$ppm_month_avg,
    start = c(1974), frequency = 12)

# Plot monthly time series
plot(co2_noaa_monthly_ts, main = "Monthly Observations of CO2 (ppm)\n Mauna Loa Observatory 19
    xlab = "Year", ylab = "Co2 ppm", col = "blue")
```



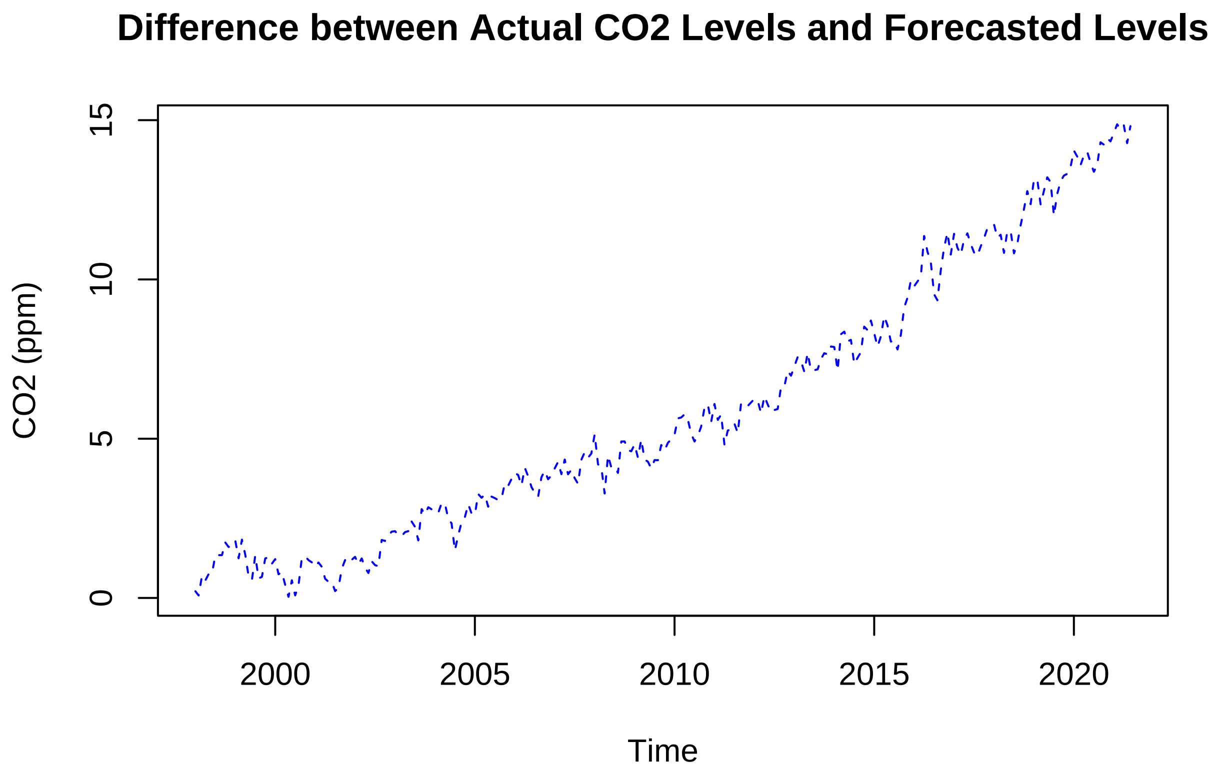
The monthly time series plotted above looks like a smoothed version of the weekly time series.

```r
# transforming time series data to dataframe, so that we
# can join
co2_actuals_filtered <- co2_noaa_monthly_df %>%
    filter(year > 1997)

co2_actuals_ts <- ts(co2_actuals_filtered$ppm_month_avg, start = c(1998),
    frequency = 12)

ts.plot(co2_actuals_ts, co2_forecast_ts, lty = 1:2, col = c("navy",
    "blue"), ylab = "CO2 (ppm)", main = "SARIMA(0,1,1,1,1,2) Forecasts vs. Actual Monthly CO2 I
```

```
legend("topleft", legend = c("Actual", "Forecast"), col = c("navy",
    "blue"), lty = 1:2)
```




```
actuals_fore_diff <- co2_actuals_ts - co2_forecast_ts

ts.plot(actuals_fore_diff, lty = 2, col = c("blue"), ylab = "CO2 (ppm)",
    main = "Difference between Actual CO2 Levels and Forecasted Levels")
```



The difference between the actual measured CO2 levels from 1998 to present and our forecasts is stark. It is clear from the plot above that we underestimated the growth of the series over the subsequent 20+ years. Given that our best model's residuals were stationary and resembled to white noise, we would conclude that the forecast error was not necessarily due to a model misspecification, but rather a change in the underlying CO2 generating process. We hypothesize this could be due to the rapid growth of China's economy and other emerging market economies through the 2000s and 2010s[1]. This could be the subject of a deeper, causal understanding of what is driving the ever-increasing concentrations of atmospheric CO2.

Part 5 (5 points)

Split the NOAA series into training and test sets, using the final two years of observations as the test set. Fit an ARIMA model to the series following all appropriate steps, including comparison of how candidate models perform both in-sample and (psuedo-) out-of-sample. Generate predictions for when atmospheric CO2 is expected to reach 450 parts per million, considering the prediction intervals as well as the point estimate. Generate a prediction for atmospheric CO2 levels in the year 2100. How confident are you that these will be accurate predictions?

[1] https://climateactiontracker.org/countries/china/