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.

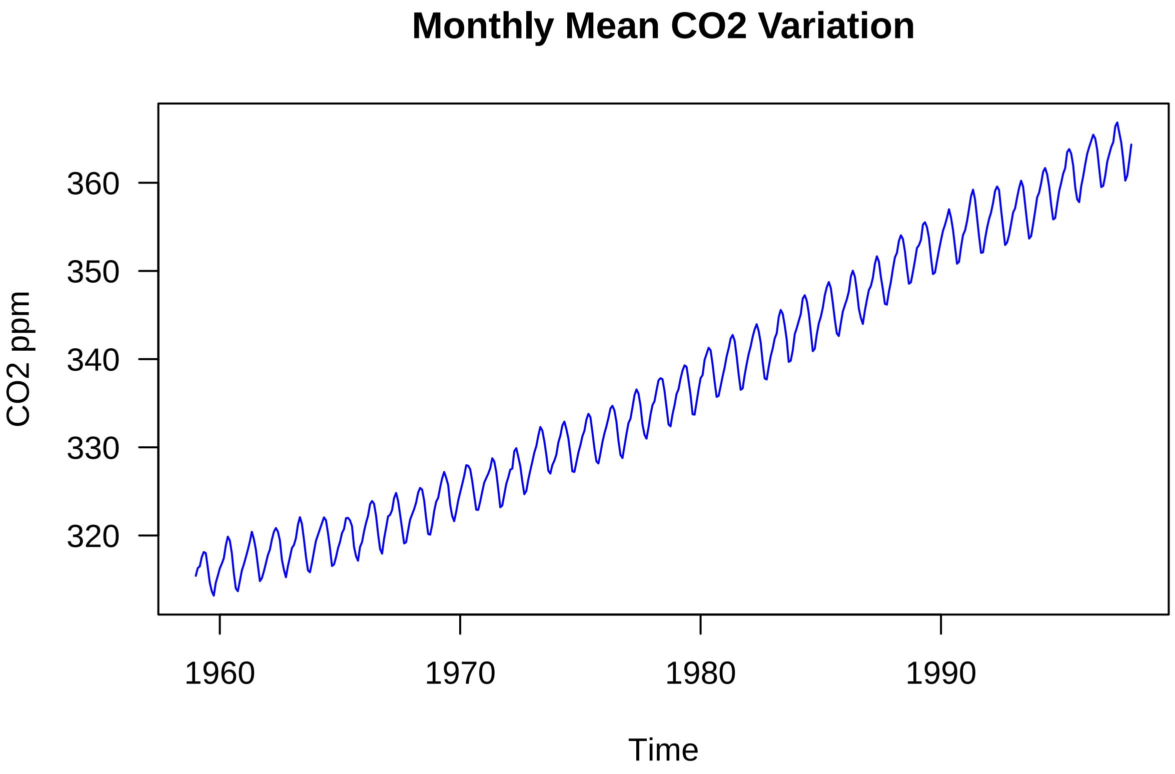
The Keeling Curve

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'.

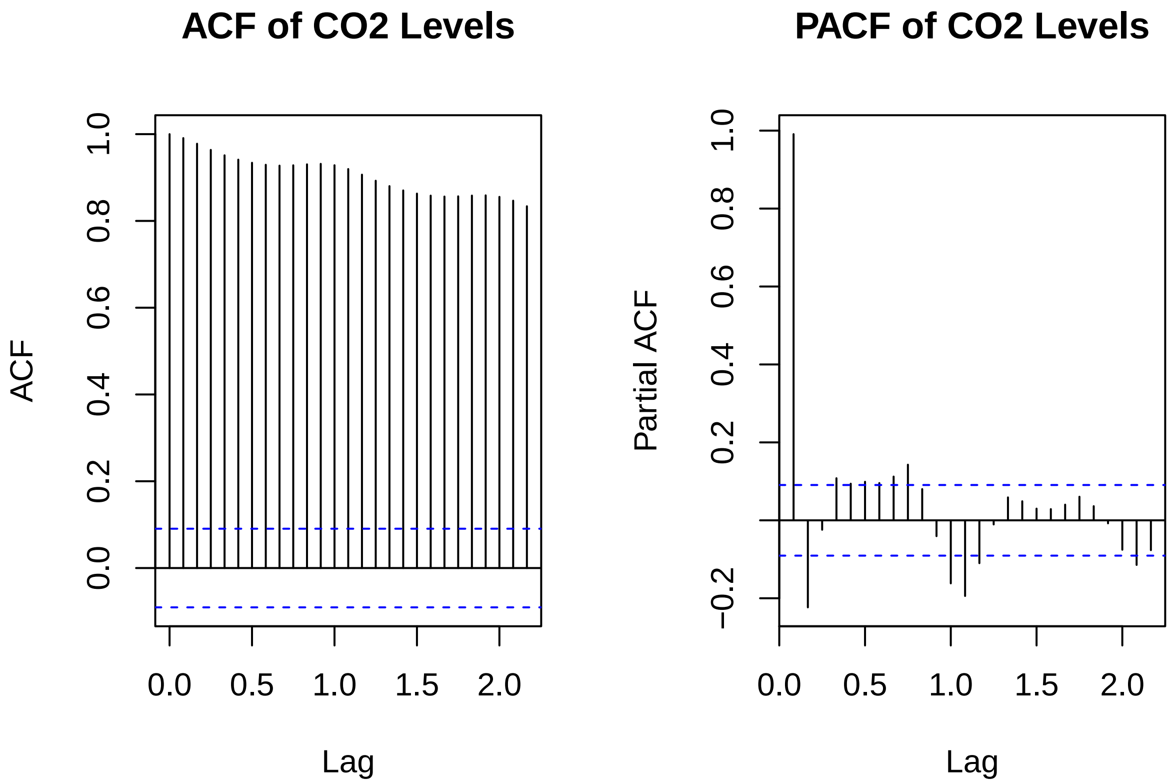
```
# Get co2 as data.frame
co2.ts <- as_tsibble(co2)
co2.ts <- rename(co2.ts, ppm = value, month = index)
plot(co2, ylab = expression("CO2 ppm"), col = "blue", las = 1,
    type = "l")
title(main = "Monthly Mean CO2 Variation")
```



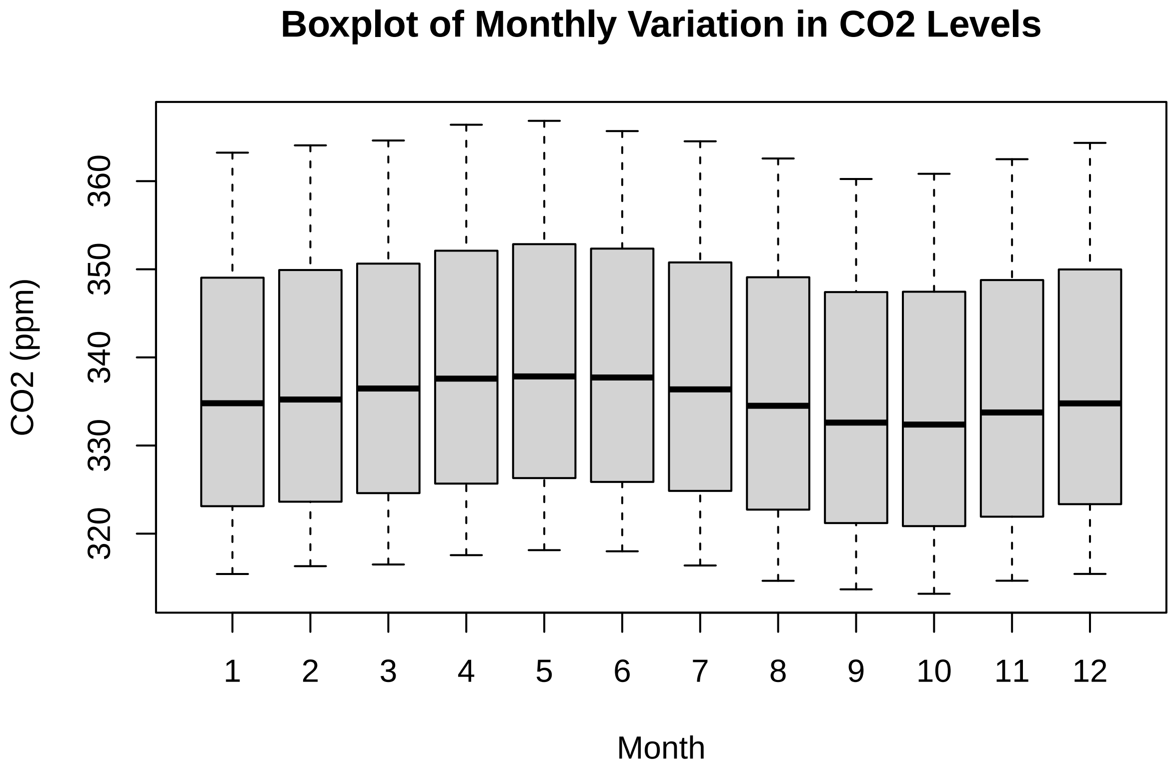
Part 1 (3 points)

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.

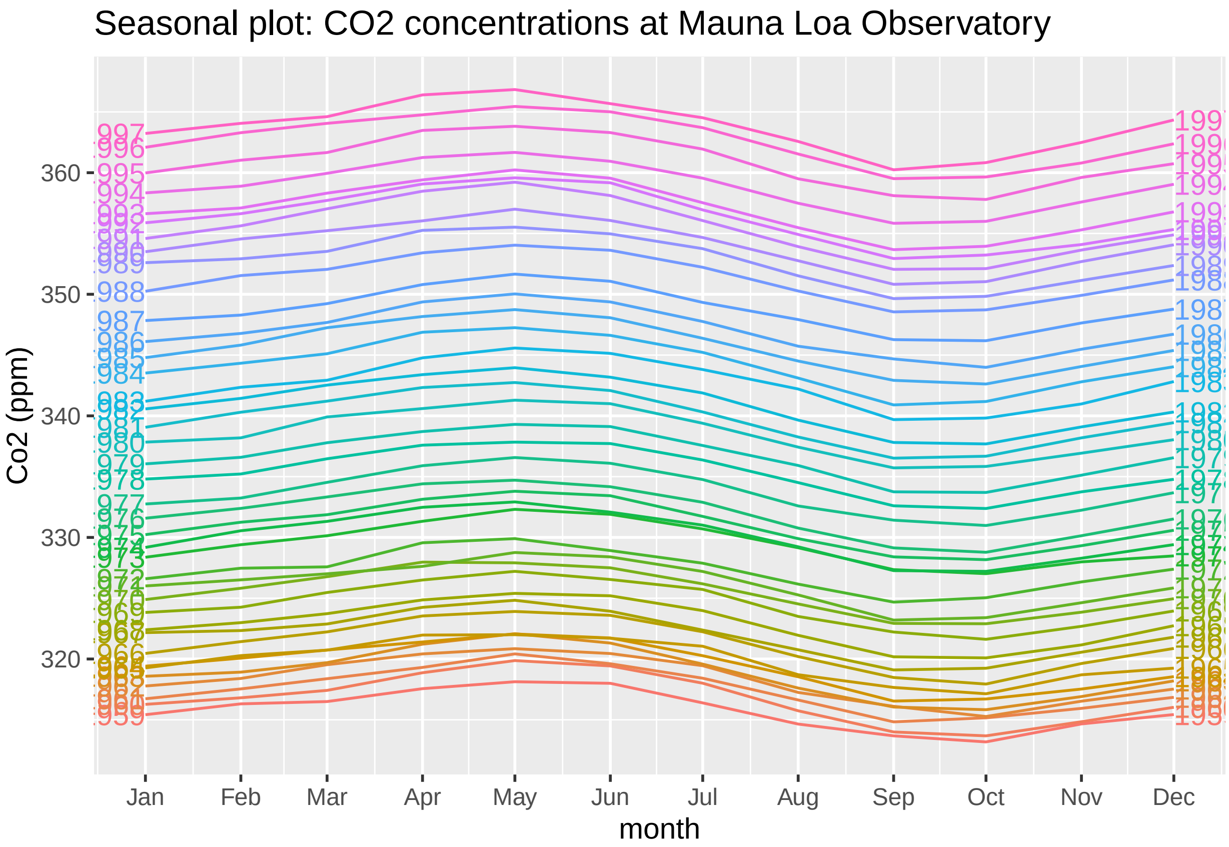
```
# Auto and partial correlation functions
par(mfrow = c(1, 2))
acf(co2, main = "ACF of CO2 Levels")
pacf(co2, main = "PACF of CO2 Levels")
```



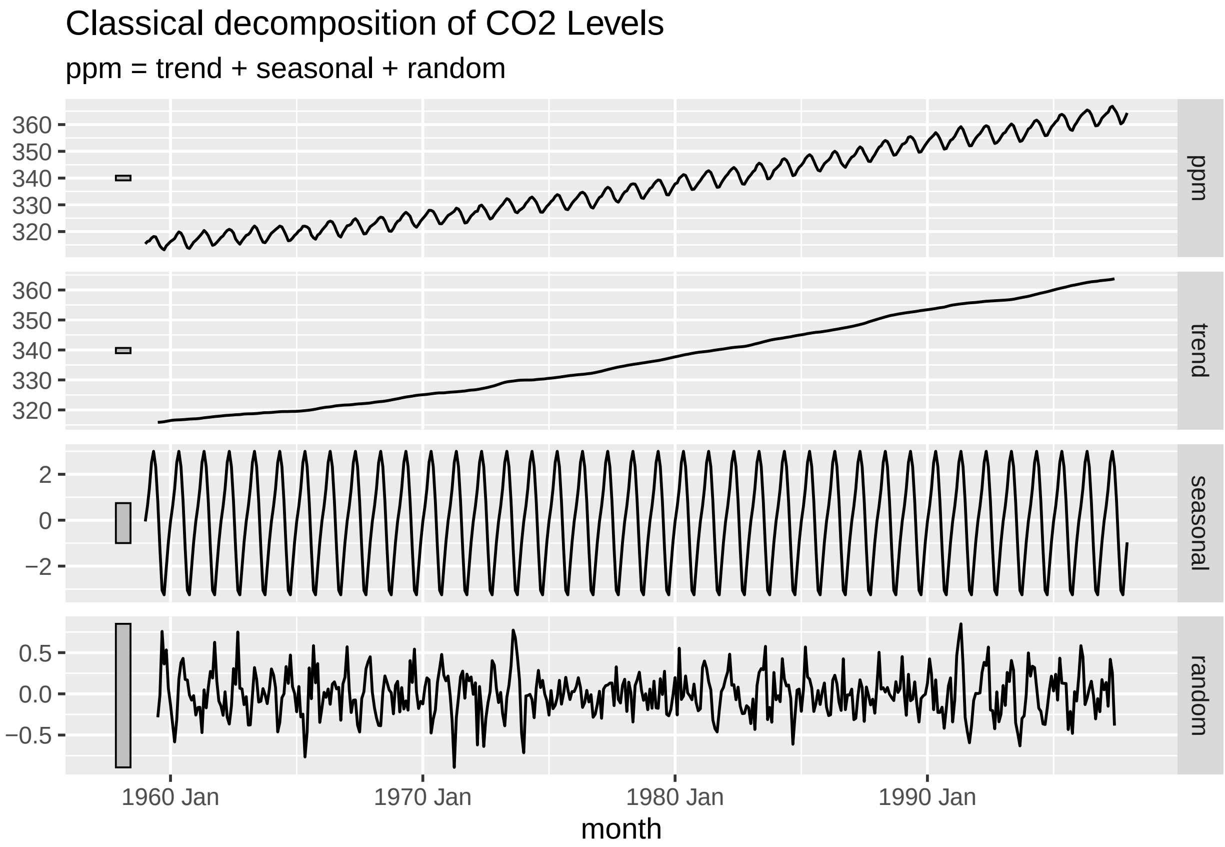
```
boxplot(co2 ~ cycle(co2), xlab = "Month", ylab = "CO2 (ppm)",
    main = "Boxplot of Monthly Variation in CO2 Levels")
```



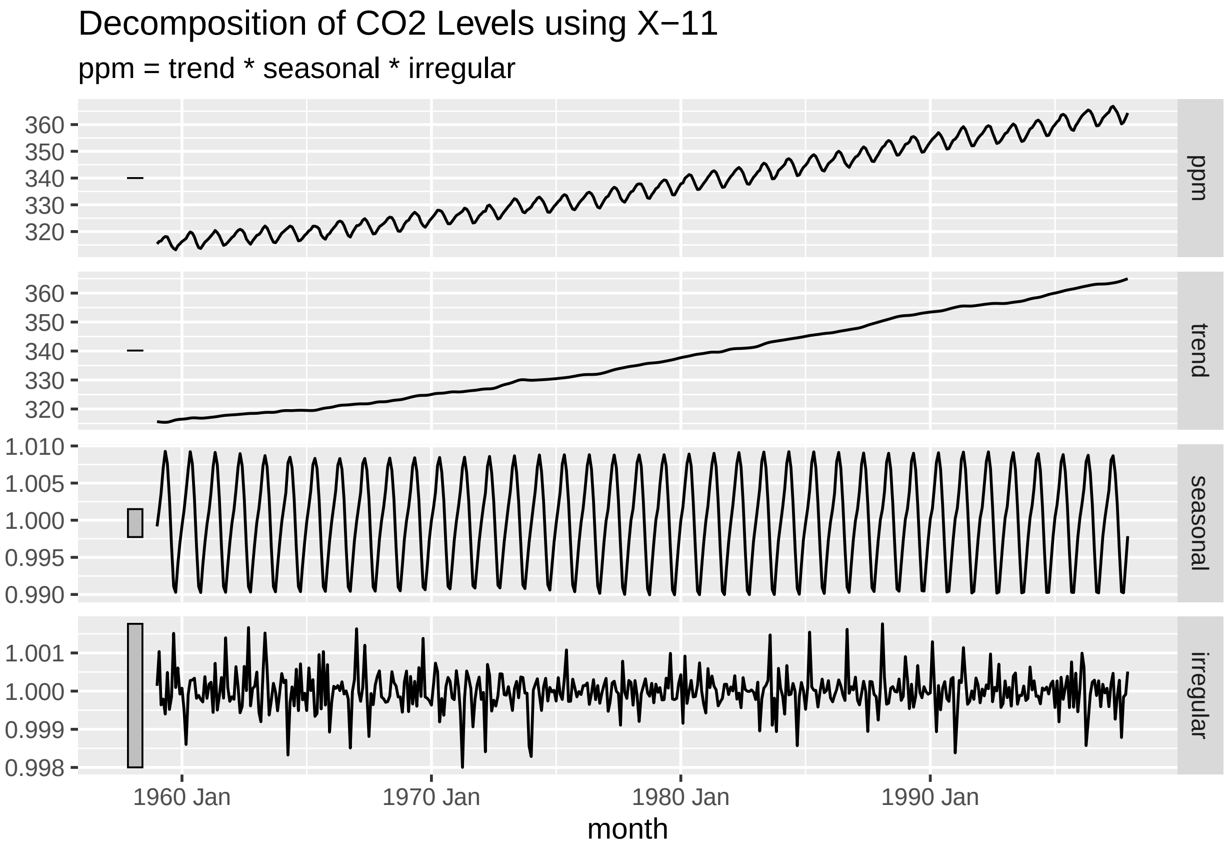
```
# Seasonal Plots
co2.ts %>%
    gg_season(ppm, labels = "both") + labs(y = "Co2 (ppm)", title = "Seasonal plot: CO2 concen
```



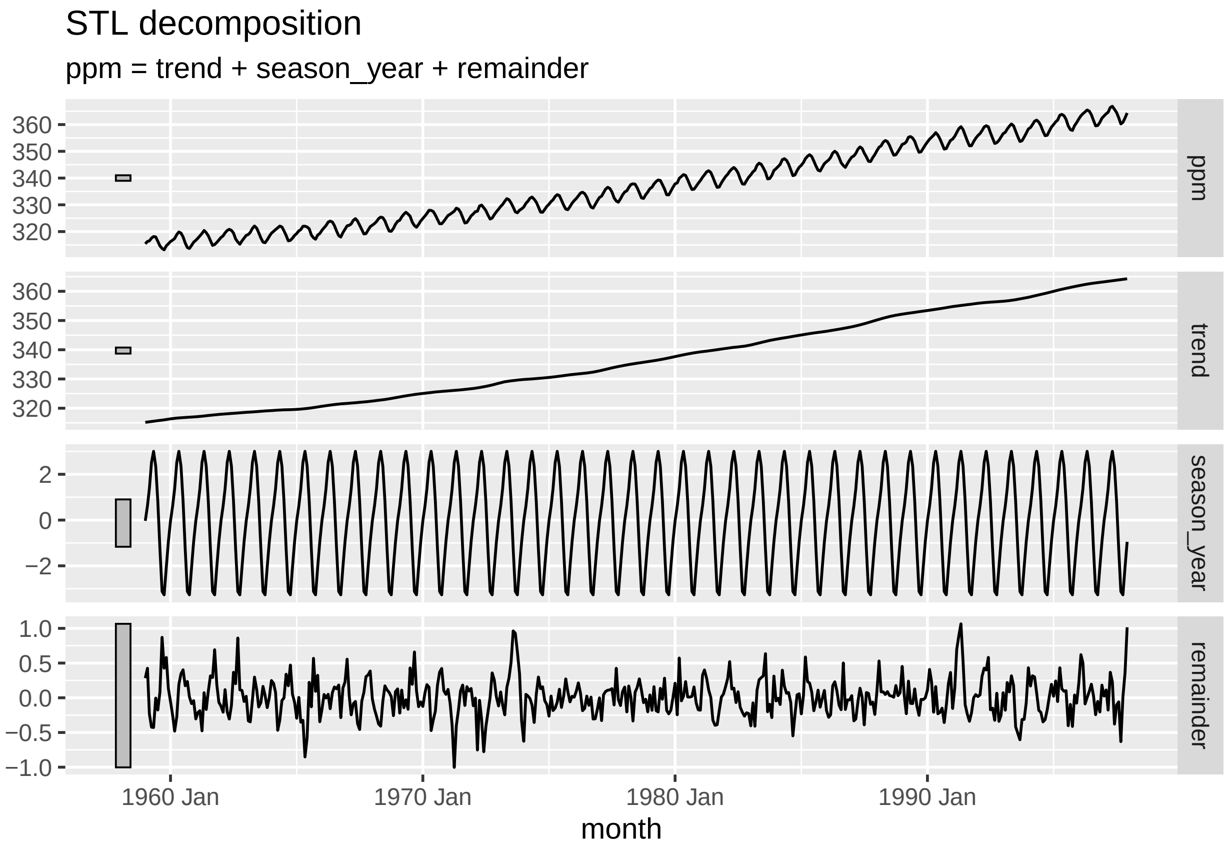
```
co2.ts %>%
    model(classical_decomposition(ppm, type = "additive")) %>%
    components() %>%
    autoplot() + labs(title = "Classical decomposition of CO2 Levels")
```



```
x11_dcmp <- co2.ts %>%
    model(x11 = X_13ARIMA_SEATS(ppm ~ x11())) %>%
    components()
autoplot(x11_dcmp) + labs(title = "Decomposition of CO2 Levels using X-11")
```



```
co2.ts %>%
    model(STL(ppm ~ trend(window = 21) + season(window = "periodic"),
        robust = TRUE)) %>%
    components() %>%
    autoplot()
```



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.fit <- lm(ppm ~ time(month), data = co2.ts)
summary(linear.fit)
```

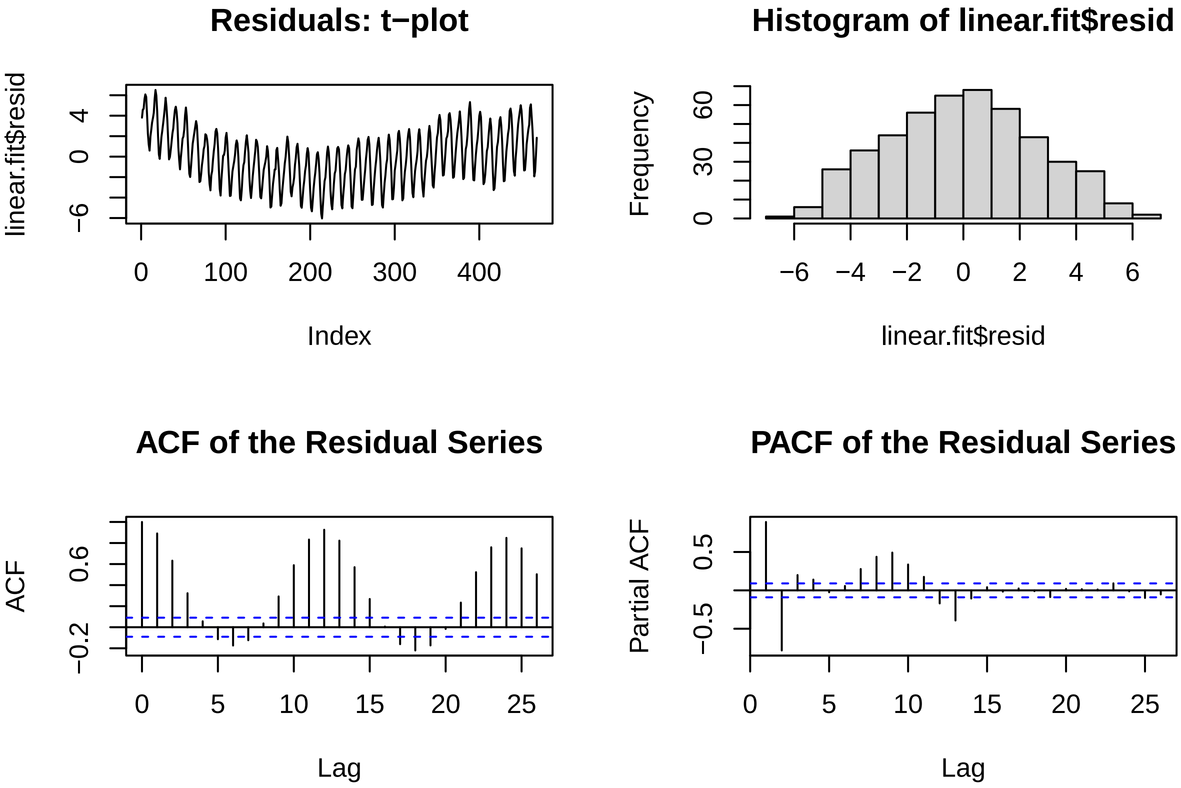
```
##
## Call:
## lm(formula = ppm ~ time(month), data = co2.ts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0399 -1.9476 -0.0017  1.9113  6.5149
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.250e+03  2.127e+01  -105.8   <2e-16 ***
## time(month)  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
```

```
# Residual Diagnostics
summary(linear.fit$resid)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -6.039885 -1.947575 -0.001671  0.000000  1.911271  6.514852
```

```
par(mfrow = c(2, 2))

plot(linear.fit$resid, type = "l", main = "Residuals: t-plot")
hist(linear.fit$resid)
acf(linear.fit$resid, main = "ACF of the Residual Series")
pacf(linear.fit$resid, main = "PACF of the Residual Series")
```



```
Box.test(residuals(linear.fit), lag = 12, type = "Ljung")
```

```
##
##  Box-Ljung test
##
## data:  residuals(linear.fit)
## X-squared = 1598.1, df = 12, p-value < 2.2e-16
```

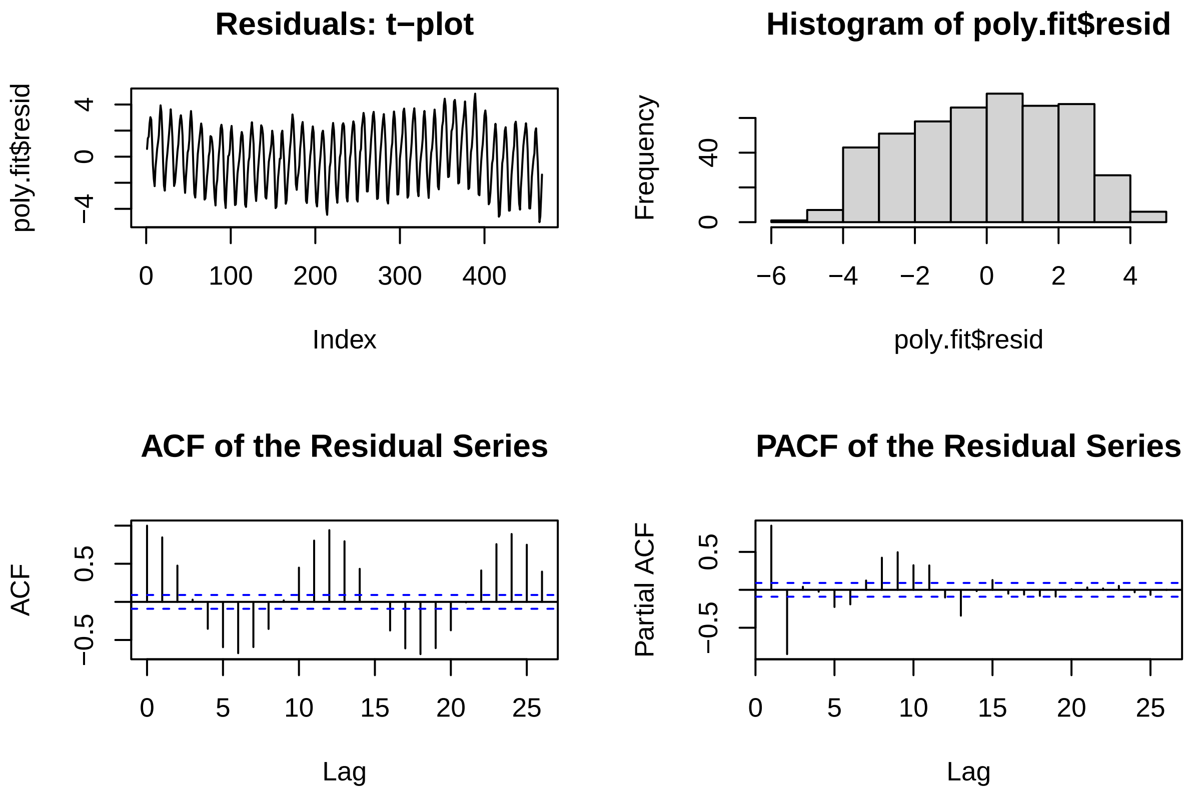
```
poly.fit <- lm(ppm ~ time(month) + I(time(month)^2), data = co2.ts)
summary(poly.fit)
```

```
##
## Call:
## lm(formula = ppm ~ time(month) + I(time(month)^2), data = co2.ts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0195 -1.7120  0.2144  1.7957  4.8345
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.770e+04  3.483e+03   13.70   <2e-16 ***
## time(month)      -4.919e+01  3.521e+00  -13.97   <2e-16 ***
## I(time(month)^2)  1.276e-02  8.898e-04   14.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.182 on 465 degrees of freedom
## Multiple R-squared:  0.9788, Adjusted R-squared:  0.9787
## F-statistic: 1.075e+04 on 2 and 465 DF,  p-value: < 2.2e-16
```

```
# Residual Diagnostics
summary(poly.fit$resid)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -5.0195 -1.7120  0.2144  0.0000  1.7957  4.8345
```

```
par(mfrow = c(2, 2))

plot(poly.fit$resid, type = "l", main = "Residuals: t-plot")
hist(poly.fit$resid)
acf(poly.fit$resid, main = "ACF of the Residual Series")
pacf(poly.fit$resid, main = "PACF of the Residual Series")
```



```
Box.test(residuals(poly.fit), lag = 12, type = "Ljung")
```

```
##
##  Box-Ljung test
##
## data:  residuals(poly.fit)
## X-squared = 1951.9, df = 12, p-value < 2.2e-16
```

Log Transformation of CO2 Levels

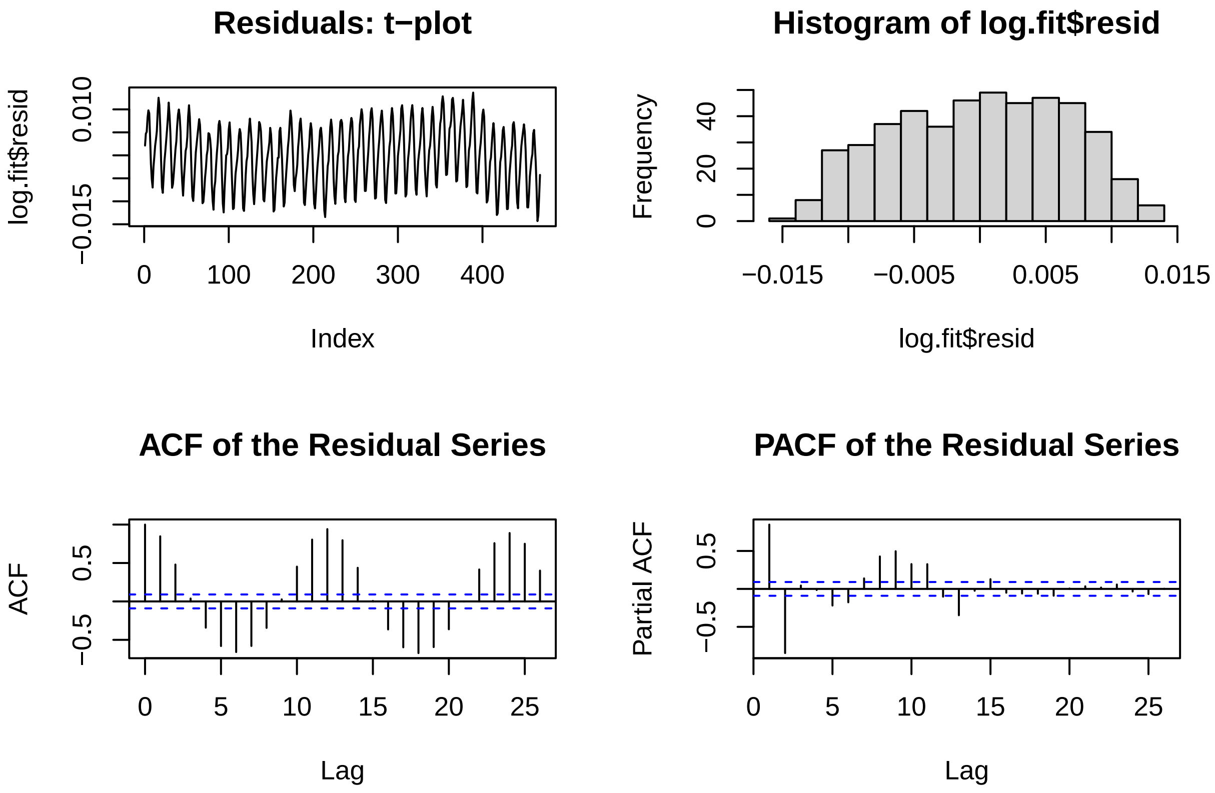
```
log.fit <- lm(log(ppm) ~ time(month) + I(time(month)^2), data = co2.ts)
summary(log.fit)
```

```
##
## Call:
## lm(formula = log(ppm) ~ time(month) + I(time(month)^2), data = co2.ts)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0143052 -0.0050832  0.0005277  0.0052757  0.0136508
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.193e+02  1.036e+01   11.52   <2e-16 ***
## time(month)       -1.186e-01  1.047e-02  -11.32   <2e-16 ***
## I(time(month)^2)   3.094e-05  2.646e-06   11.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00649 on 465 degrees of freedom
## Multiple R-squared:  0.9786, Adjusted R-squared:  0.9785
## F-statistic: 1.061e+04 on 2 and 465 DF,  p-value: < 2.2e-16
```

```
# Residual Diagnostics
summary(log.fit$resid)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -0.0143052 -0.0050832  0.0005277  0.0000000  0.0052757  0.0136508
```

```
par(mfrow = c(2, 2))

plot(log.fit$resid, type = "l", main = "Residuals: t-plot")
hist(log.fit$resid)
acf(log.fit$resid, main = "ACF of the Residual Series")
pacf(log.fit$resid, main = "PACF of the Residual Series")
```



```
Box.test(residuals(log.fit), lag = 12, type = "Ljung")
```

```
##
##  Box-Ljung test
##
## data:  residuals(log.fit)
## X-squared = 1925.9, df = 12, p-value < 2.2e-16
```

Seasonal Time-Trend Model

```
Seas <- cycle(co2)
stt.fit <- lm(ppm ~ 0 + time(month) + I(time(month)^2) + factor(Seas),
    data = co2.ts)
summary(stt.fit)
```

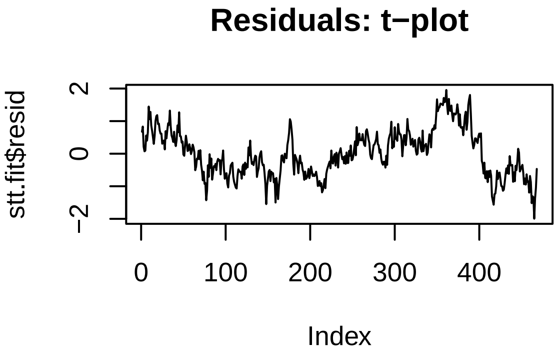
```
## 
## Call:
## lm(formula = ppm ~ 0 + time(month) + I(time(month)^2) + factor(Seas), 
##     data = co2.ts)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.99478 -0.54468 -0.06017  0.47265  1.95480 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## time(month)        -4.920e+01  1.168e+00  -42.12   <2e-16 ***
## I(time(month)^2)    1.277e-02  2.952e-04   43.24   <2e-16 ***
## factor(Seas)1       4.771e+04  1.156e+03   41.29   <2e-16 ***
## factor(Seas)2       4.771e+04  1.156e+03   41.29   <2e-16 ***
## factor(Seas)3       4.771e+04  1.156e+03   41.29   <2e-16 ***
## factor(Seas)4       4.771e+04  1.156e+03   41.29   <2e-16 ***
## factor(Seas)5       4.771e+04  1.156e+03   41.29   <2e-16 ***
## factor(Seas)6       4.771e+04  1.156e+03   41.29   <2e-16 ***
## factor(Seas)7       4.771e+04  1.156e+03   41.29   <2e-16 ***
## factor(Seas)8       4.771e+04  1.156e+03   41.29   <2e-16 ***
## factor(Seas)9       4.771e+04  1.156e+03   41.29   <2e-16 ***
## factor(Seas)10      4.771e+04  1.156e+03   41.29   <2e-16 ***
## factor(Seas)11      4.771e+04  1.156e+03   41.29   <2e-16 ***
## factor(Seas)12      4.771e+04  1.156e+03   41.29   <2e-16 ***
## ---
## 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:      1,  Adjusted R-squared:      1 
## F-statistic: 7.259e+06 on 14 and 454 DF,  p-value: < 2.2e-16
```

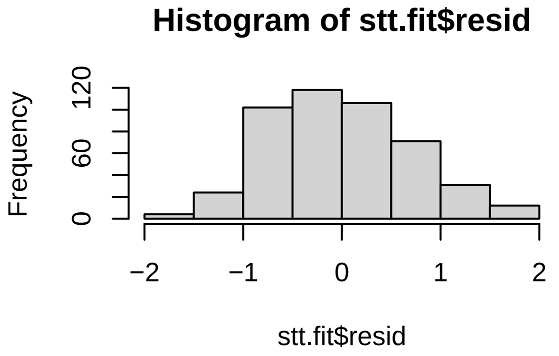
```
# Residual Diagnostics
summary(stt.fit$resid)
```

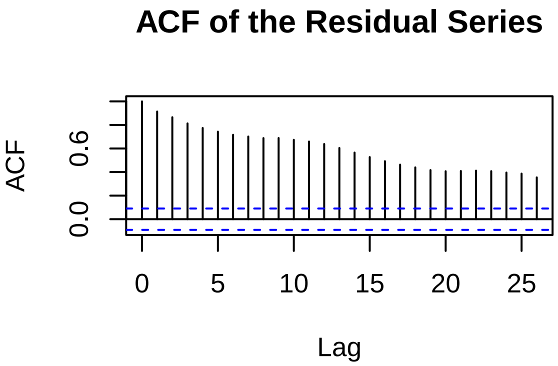
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## -1.99478 -0.54468 -0.06017  0.00000  0.47265  1.95480 
```

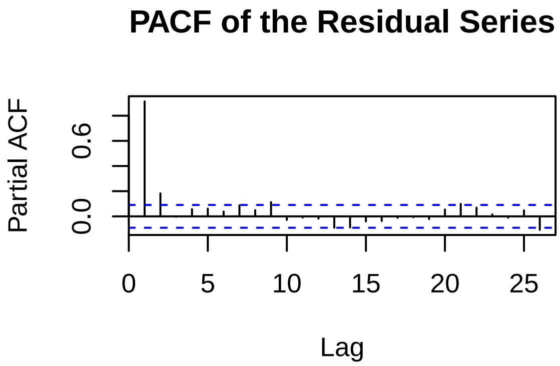
```
par(mfrow = c(2, 2))

plot(stt.fit$resid, type = "l", main = "Residuals: t-plot")
hist(stt.fit$resid)
```

```
acf(stt.fit$resid, main = "ACF of the Residual Series")
pacf(stt.fit$resid, main = "PACF of the Residual Series")
```







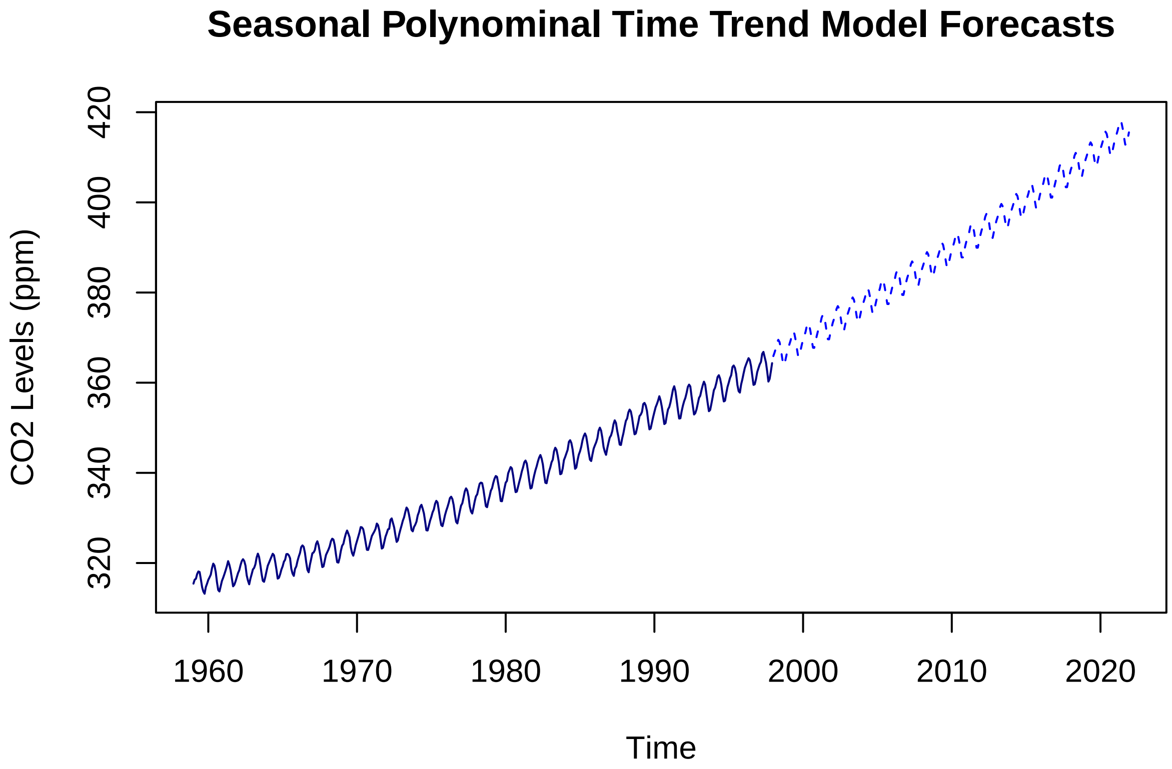


```
Box.test(residuals(stt.fit), lag = 12, type = "Ljung")
```

```
##
##  Box-Ljung test
##
## data:  residuals(stt.fit)
## X-squared = 3163.8, df = 12, p-value < 2.2e-16
```

Seasonal Time-Trend Model Predictions

```
new.t = seq(1998, len = (2021 - 1997) * 12, by = 1/12)
new.Seas <- rep(1:12, (2021 - 1997))
new.dat <- data.frame(month = new.t, Seas = new.Seas)
stt.preds <- ts(predict(stt.fit, new.dat), st = 1998, fr = 12)

ts.plot(co2, stt.preds, lty = 1:2, col = c("navy", "blue"), ylab = "CO2 Levels (ppm)",
    main = "Seasonal Polynominal Time Trend Model Forecasts")
```



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.

```
get.best.arima <- function(x.ts, maxord = c(1, 1, 1, 1, 1, 1)) {
    best.aic <- 1e+08
    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")
        fit.aic <- -2 * fit$loglik + (log(n) + 1) * length(fit$coef)
        if (fit.aic < best.aic) {
            best.aic <- fit.aic
            best.fit <- fit
            best.model <- c(p, d, q, P, D, Q)
        }
    }
    list(best.aic, best.fit, best.model)
}

best.arima.co2 <- get.best.arima(co2.ts$ppm, maxord = c(4, 4,
    4, 4, 4, 4))
```

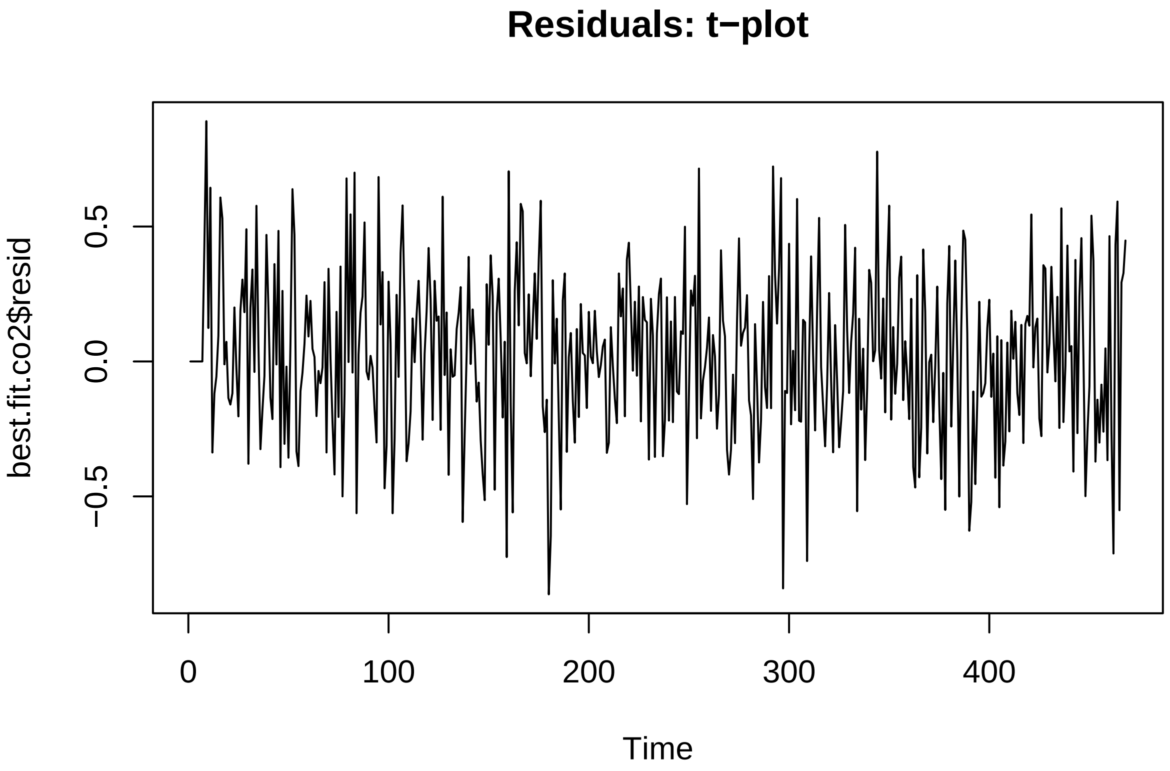
```
best.arima.co2[[1]]
```

```
## [1] 300.0805
```

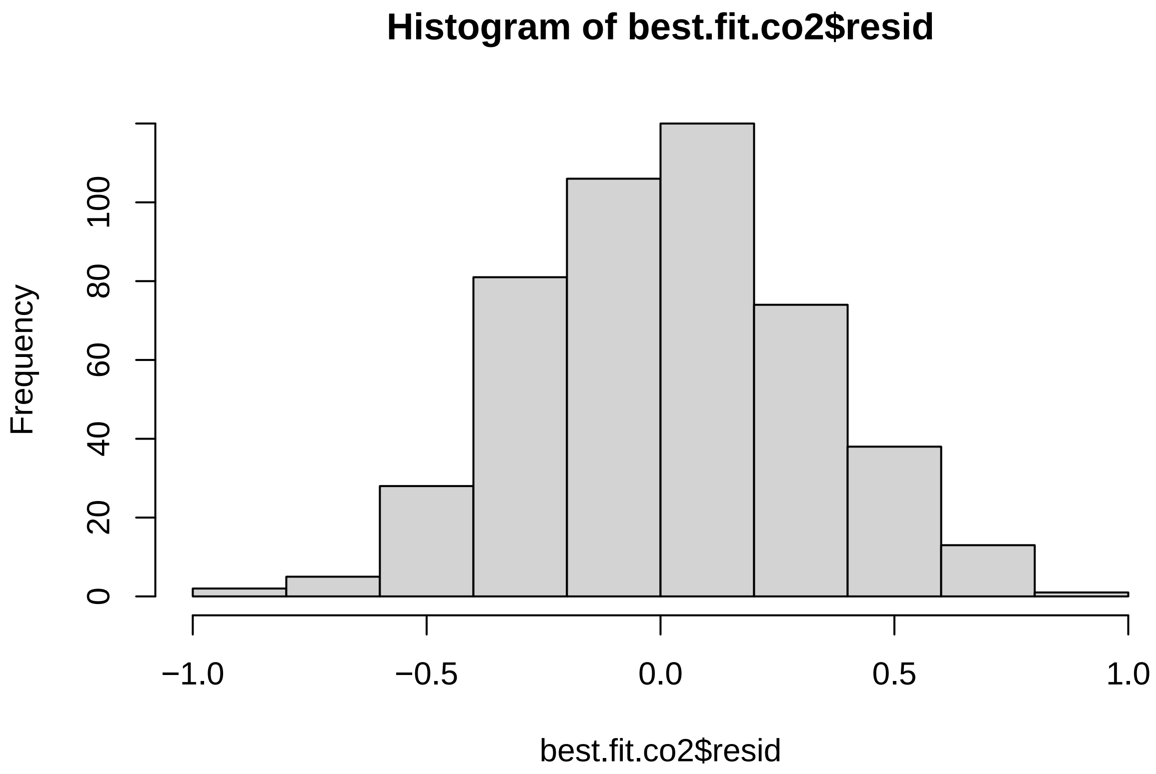
```r
best.fit.co2 <- best.arima.co2[[2]]
best.arima.co2[[3]]
```

```
## [1] 3 0 3 2 2 4
```

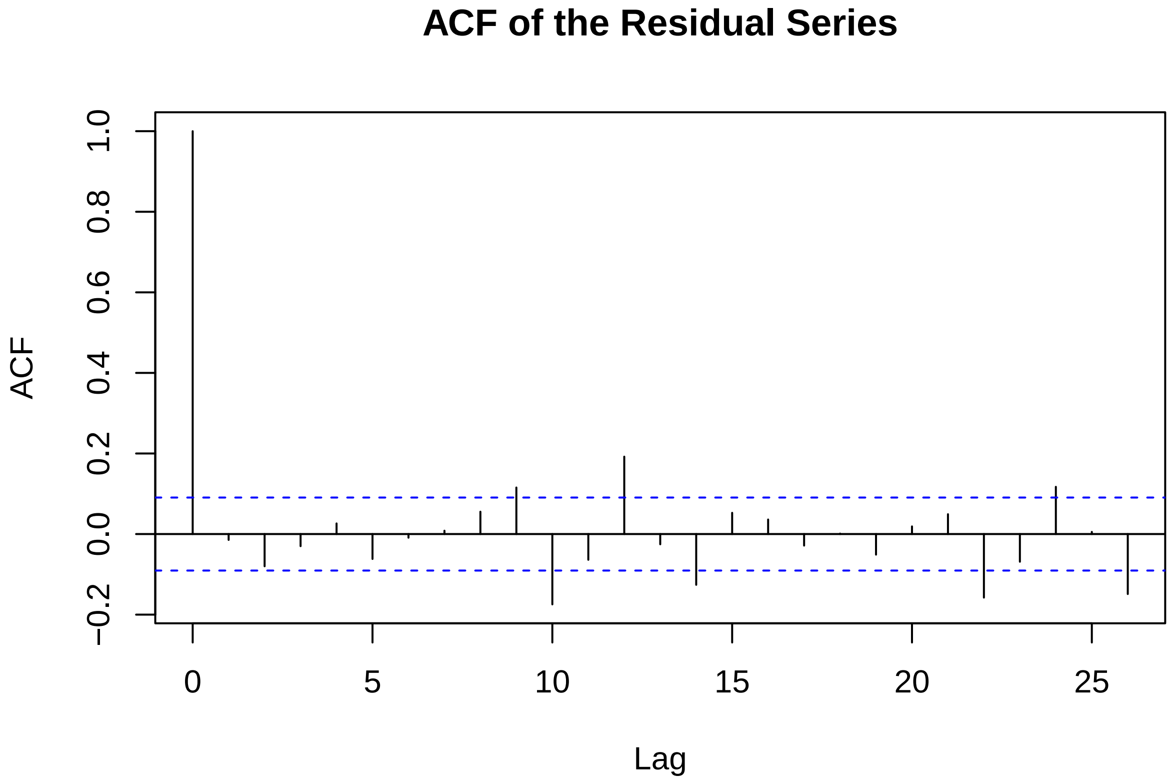
```r
# Residual diagnostics
plot(best.fit.co2$resid, type = "l", main = "Residuals: t-plot")
```



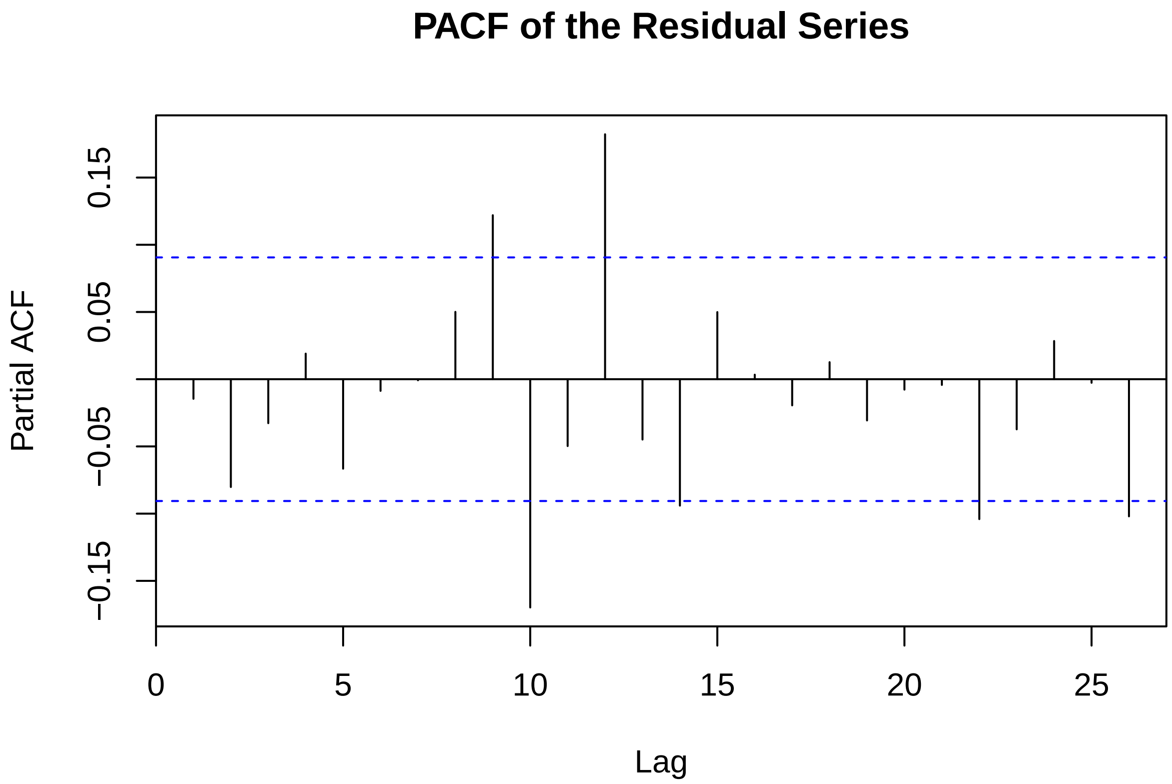
```r
hist(best.fit.co2$resid)
```



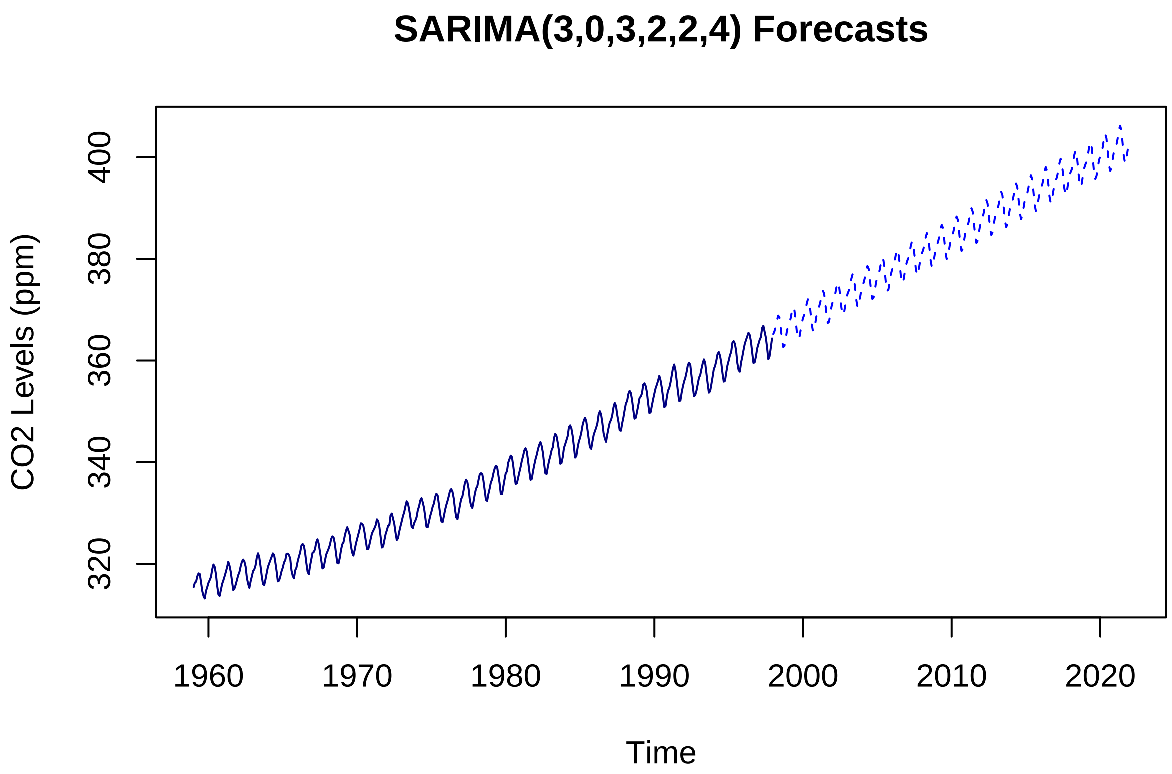
```
acf(best.fit.co2$resid, main = "ACF of the Residual Series")
```



```
pacf(best.fit.co2$resid, main = "PACF of the Residual Series")
```



```
new.t = seq(1998, len = (2021 - 1997) * 12, by = 1/12)
new.dat <- data.frame(month = new.t)
arima.preds <- ts(predict(best.fit.co2, (2021 - 1997) * 12)$pred,
    st = 1998, fr = 12)

ts.plot(co2, arima.preds, lty = 1:2, col = c("navy", "blue"),
    ylab = "CO2 Levels (ppm)", main = "SARIMA(3,0,3,2,2,4) Forecasts")
```



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("yr", "mon", "day", "decimal", "ppm",
    "days", "1yr_ago", "10yrs_ago", "since1800")
summary(co2_weekly)
```

```
##        yr            mon             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
```

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?