CO2 Emissions 1997

Team 2

2024-03-18

POV: 1997

Context

Climate science has emerged as a leading field of interest in the 20th century. Efforts to bring awareness to the impact of human intervention on the environment, such as the burning of fossil fuels, have proved fruitful. The Intergovernmental Panel on Climate Change (IPCC) has reinforced these efforts. In 1990 they released the First Climate Assessment Report stating that "human activities are substantially increasing the atmospheric concentration of greenhouse gases" (IPCC, 1990). Greenhouse gases are a group of gases, such as carbon dioxide, methane, and nitrous oxide, that when present in higher concentrations in the atmosphere, raise the surface temperate of the Earth. Carbon dioxide is the most abundant greenhouse gas that is produced from human activity, namely energy production via fossil fuel combustion. Since the industrial revolution, energy consumption from petroleum and natural gas sources has risen dramatically. This report aims to investigate the following questions: How have the levels of atmospheric CO2 changed over time? And, is there an identifiable pattern that will persist into the future? Forecasting atmospheric carbon dioxide concentrations allows scientists to measure the corresponding impact to the global environment and justify the need for human intervention in the opposite direction.

Data and Exploration

Charles Keeling was a research scientist who made it his life's work to survey the atmosphere in hopes of confirming Svante Arrhenius's theory that fossil fuel combustion is increasing the concentration of CO2 in the atmosphere. To this end, Keeling collected atmospheric CO2 concentration measurements at a number of sampling-stations including the Mauna Loa Observatory in Hawaii. These measurements were taken using a CO2 analyzer which detects the amount of infrared absorption present in a air sample and turns it into a mole fraction of CO2, defined as the total CO2 molecules divided by the total non-water vapor molecules in the air, measured in parts per million (ppm). This report uses the data collected at the Mauna Loa Observatory between January 1959 and December 1997. The dataset consists of 468 observations with each observation representing the monthly total atmospheric concentration of CO2 (ppm). The observations for February, March, and April 1964 were unavailable so the values in the dataset were generated via linear interpolation between the observations for January and May 1964.

In order to better understand the characteristics of this time series, we conducted a exploratory analysis prior to modeling. Figure 1 shows the time series plot for CO2 concentration, its autocorrelation plot, and its partial autocorrelation plot. The time series plot shows a clear positive trend as well as the presence of seasonality. The autocorrelation plot provides evidence to support the presence of both trend and seasonality as it decays with increasing lags and shows a spike at about every twelfth lag, indicating a seasonal cycle.

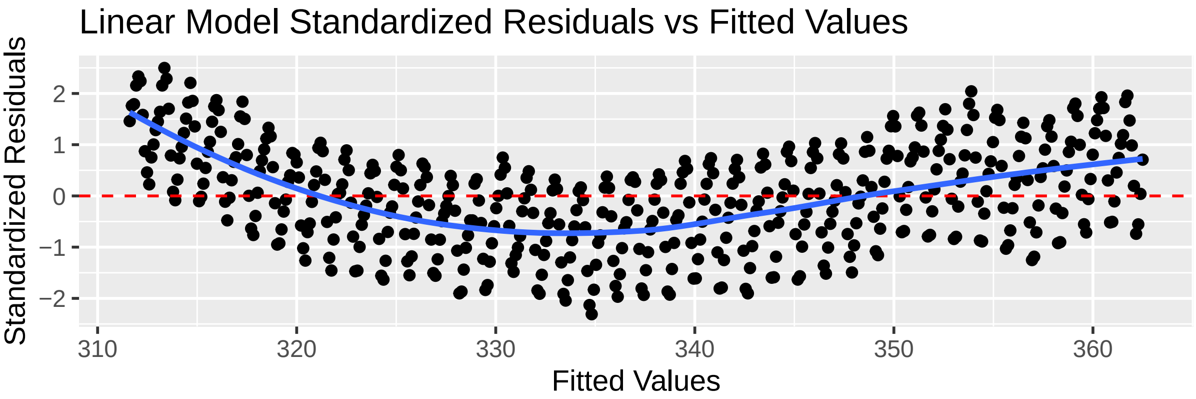
- Discuss seasonal/trend decomposition
- What about growth rates?

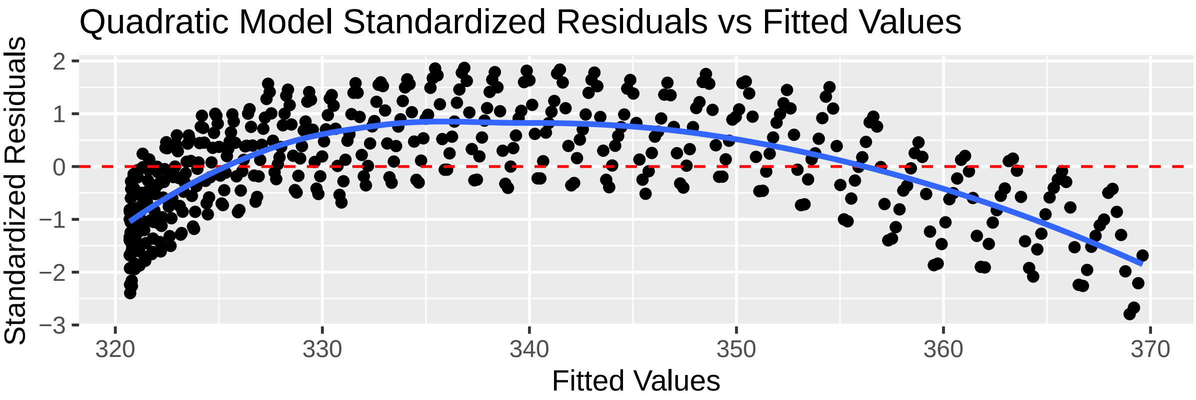
Models and Forecasts

- Why modeling is important to aid understanding?
- Empirical evidence

Linear Model

```
## ‘geom_smooth()‘ using method = ’loess’ and formula = ’y ~ x’
## ‘geom_smooth()‘ using method = ’loess’ and formula = ’y ~ x’
```





```
##
## Call:
## lm(formula = value ~ month_since_start + month, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.768 -1.284 -0.405  1.261  4.337
##
## Coefficients:
##                    Estimate Std. Error  t value Pr(>|t|)
## (Intercept)       311.42208    0.29171 1067.565  < 2e-16 ***
## month_since_start   0.10921    0.00056  195.003  < 2e-16 ***
## month2              0.66336    0.37054    1.790 0.074078 .
## month3              1.40543    0.37054    3.793 0.000169 ***
## month4              2.53597    0.37054    6.844 2.50e-11 ***
## month5              3.01445    0.37054    8.135 3.95e-15 ***
## month6              2.35139    0.37055    6.346 5.36e-10 ***
## month7              0.83039    0.37055    2.241 0.025510 *
## month8             -1.23728    0.37056   -3.339 0.000910 ***
## month9             -3.06161    0.37056   -8.262 1.58e-15 ***
## month10            -3.24441    0.37057   -8.755  < 2e-16 ***
## month11            -2.05490    0.37058   -5.545 4.99e-08 ***
## month12            -0.93744    0.37059   -2.530 0.011755 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.636 on 455 degrees of freedom
## Multiple R-squared:  0.9884, Adjusted R-squared:  0.988
## F-statistic:  3218 on 12 and 455 DF,  p-value: < 2.2e-16
```

- Functional form: y = t + e (co2 concentration = time index + error) & y = tˆ2 + e
- Results
  - High R-squared values
  - Significant p-values
- Interpretation/evaluation?
  - U-shaped (and upside down U-shaped) patterns
  - Suggest linear models are not appropriate for the data
  - Linear model predicting too low middle values
  - Quadratic model predicting too high middle values

ARIMA Model

```
## Warning: Model specification induces a quadratic or higher order polynomial trend.
## This is generally discouraged, consider removing the constant or reducing the number of differences.
## Model specification induces a quadratic or higher order polynomial trend.
## This is generally discouraged, consider removing the constant or reducing the number of differences.
## Model specification induces a quadratic or higher order polynomial trend.
## This is generally discouraged, consider removing the constant or reducing the number of differences.
```

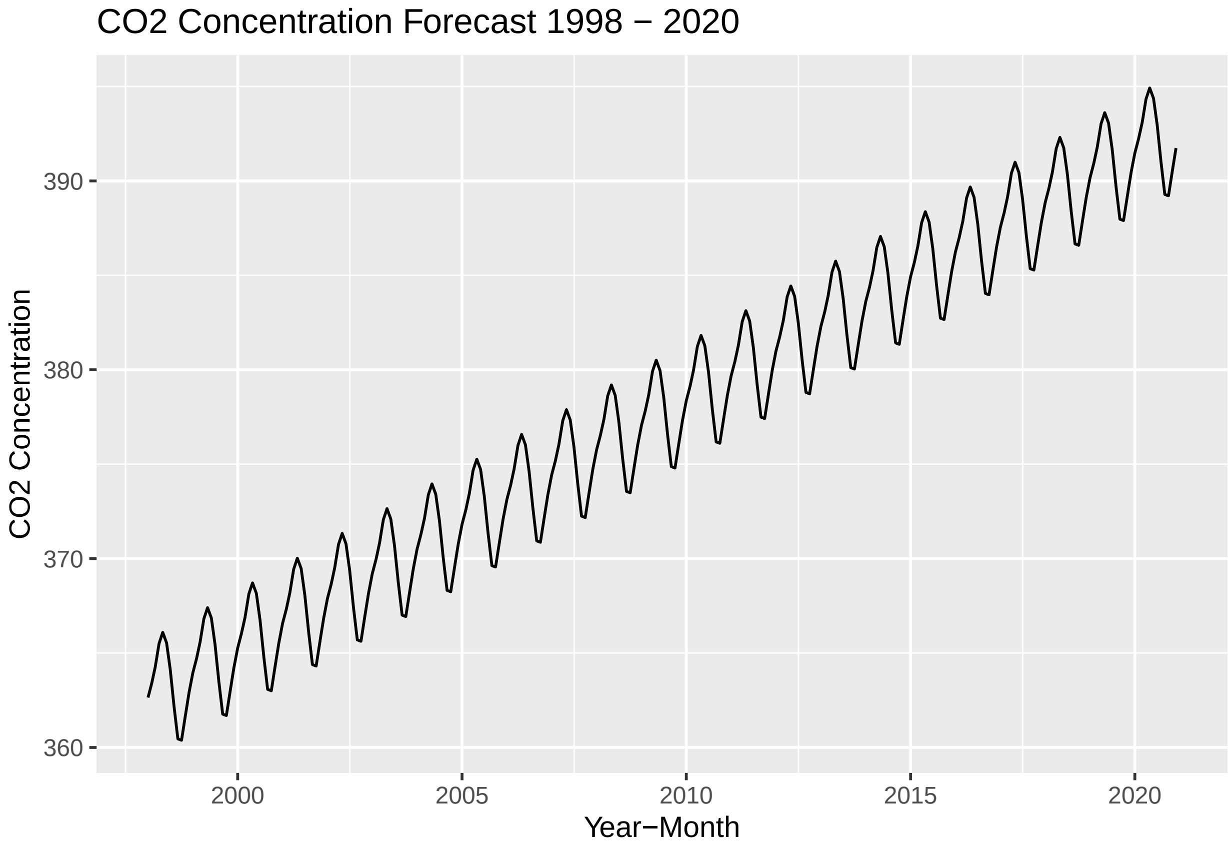
```
## Series: value
## Model: ARIMA(0,1,1)(0,1,1)[12] w/ poly
##
## Coefficients:
##           ma1     sma1  constant
##       -0.3539  -0.8563    0.0021
## s.e.   0.0498   0.0254    0.0015
##
## sigma^2 estimated as 0.08558:  log likelihood=-85.12
## AIC=178.24   AICc=178.33   BIC=194.72
```

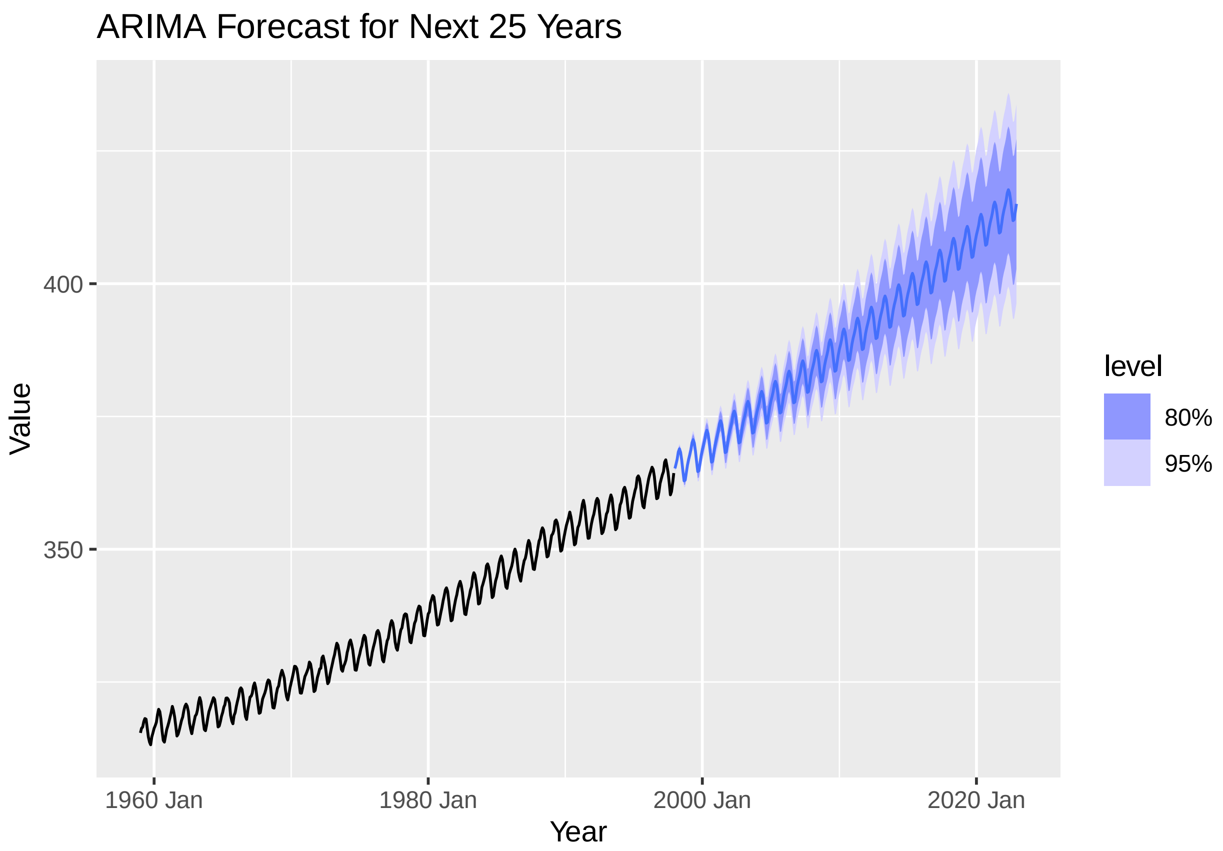
```
## Series: value
## Model: ARIMA(3,1,0)(3,1,0)[12] w/ poly
##
## Coefficients:
##           ar1      ar2      ar3     sar1     sar2     sar3  constant
##       -0.3678  -0.1561  -0.1112  -0.6756  -0.4821  -0.2333    0.0089
## s.e.   0.0469   0.0500   0.0474   0.0477   0.0529   0.0480    0.0146
##
## sigma^2 estimated as 0.09642:  log likelihood=-106.97
## AIC=229.94   AICc=230.27   BIC=262.91
```

```
## Series: value
## Model: ARIMA(0,1,1)(1,1,2)[12] w/ poly
##
## Coefficients:
##           ma1     sar1     sma1     sma2  constant
##       -0.3521  -0.5363  -0.2842  -0.4984    0.0033
## s.e.   0.0501   0.5606   0.5440   0.4621    0.0023
##
## sigma^2 estimated as 0.08579:  log likelihood=-84.63
## AIC=181.26   AICc=181.45   BIC=205.98
```

- Functional form
- Results

Forecasts





- Forecasts
- Predictions for when CO2 is expected to be at 420 ppm and 500 ppm
- Interpretation/evaluation?

#(1 point) Task 0b: Introduction

In this introduction, you can assume that your reader will have just read your 1997 report. In this introduction, very briefly pose the question that you are evaluating, and describe what (if anything) has changed in the data generating process between 1997 and the present.

#(3 points) Task 1b: Create a modern data pipeline for Mona Loa CO2 data.

The most current data is provided by the United States' National Oceanic and Atmospheric Administration, on a data page [here]. Gather the most recent weekly data from this page. (A group that is interested in even more data management might choose to work with the hourly data.) Create a data pipeline that starts by reading from the appropriate URL, and ends by saving an object called co2_present that is a suitable time series object. Conduct the same EDA on this data. Describe how the Keeling Curve evolved from 1997 to the present, noting where the series seems to be following similar trends to the series that you "evaluated in 1997" and where the series seems to be following different trends. This EDA can use the same, or very similar tools and views as you provided in your 1997 report.

```
weekly_co2_url <- "https://gml.noaa.gov/webdata/ccgg/trends/co2/co2_weekly_mlo.csv"

content <- read_lines(weekly_co2_url, skip_empty_rows = TRUE)

header_end_index <- max(grep("^#", content))
```

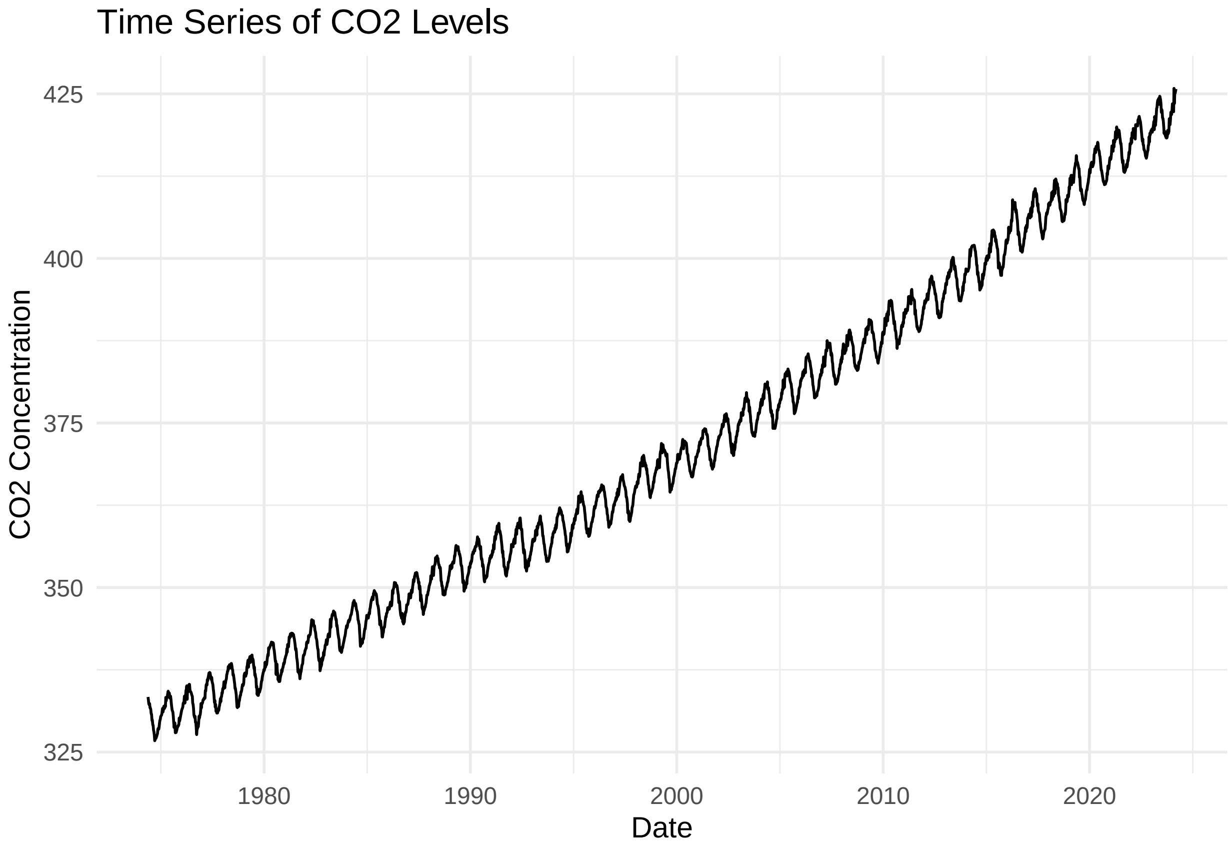
```r
co2_present <- read_csv(weekly_co2_url, skip = header_end_index, show_col_types = FALSE)

co2_present <- co2_present %>%
  mutate(date = make_date(year, month, day))

co2_present <- as_tsibble(co2_present, index = date)

co2_present <- co2_present[co2_present$average != -999.99, ]

head(co2_present)
```

```
## # A tsibble: 6 x 10 [7D]
##    year month   day decimal average ndays ‘1 year ago‘ ‘10 years ago‘
##   <dbl> <dbl> <dbl>   <dbl>   <dbl> <dbl>        <dbl>          <dbl>
## 1  1974     5    19   1974.    333.     5       -1000.         -1000.
## 2  1974     5    26   1974.    333.     6       -1000.         -1000.
## 3  1974     6     2   1974.    332.     5       -1000.         -1000.
## 4  1974     6     9   1974.    332.     7       -1000.         -1000.
## 5  1974     6    16   1974.    332.     7       -1000.         -1000.
## 6  1974     6    23   1974.    332.     5       -1000.         -1000.
## # i 2 more variables: ‘increase since 1800‘ <dbl>, date <date>
```

```r
dim(co2_present)
```

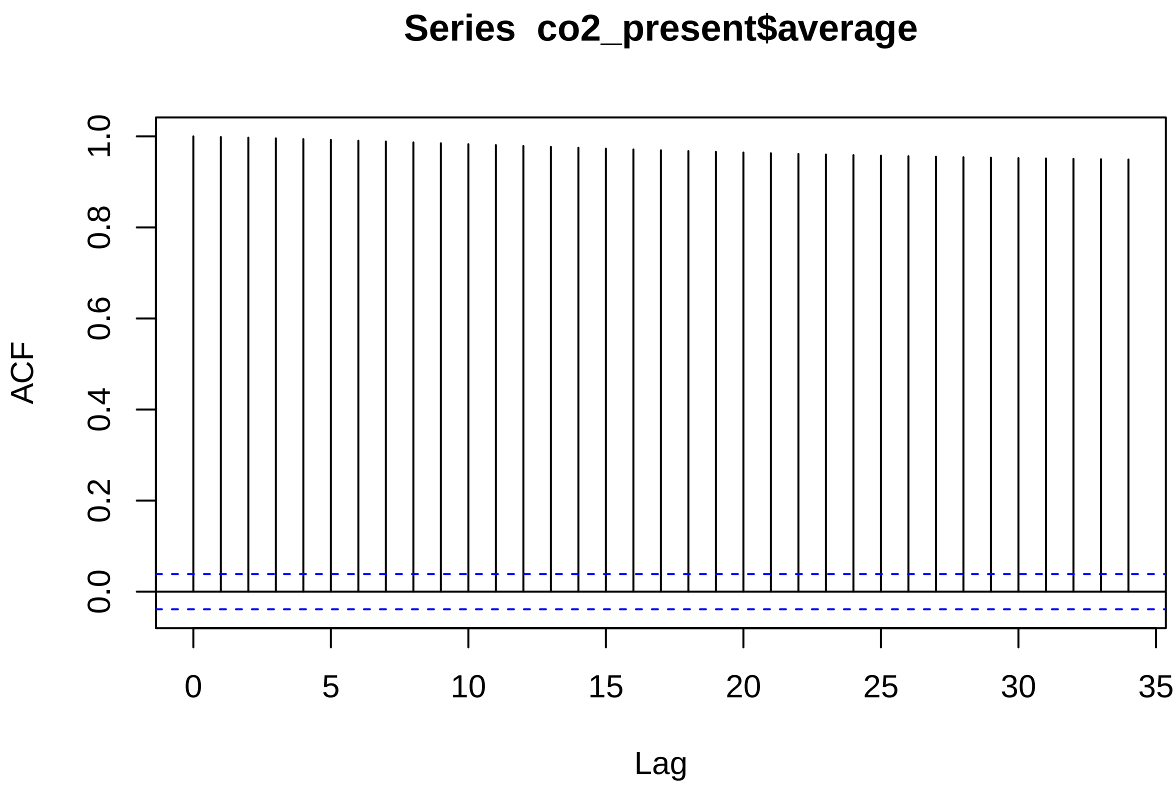
```
## [1] 2582   10
```

```r
ggplot(co2_present, aes(x = date, y = average)) +
  geom_line() +
  theme_minimal() +
  labs(title = "Time Series of CO2 Levels", x = "Date", y = "CO2 Concentration")
```



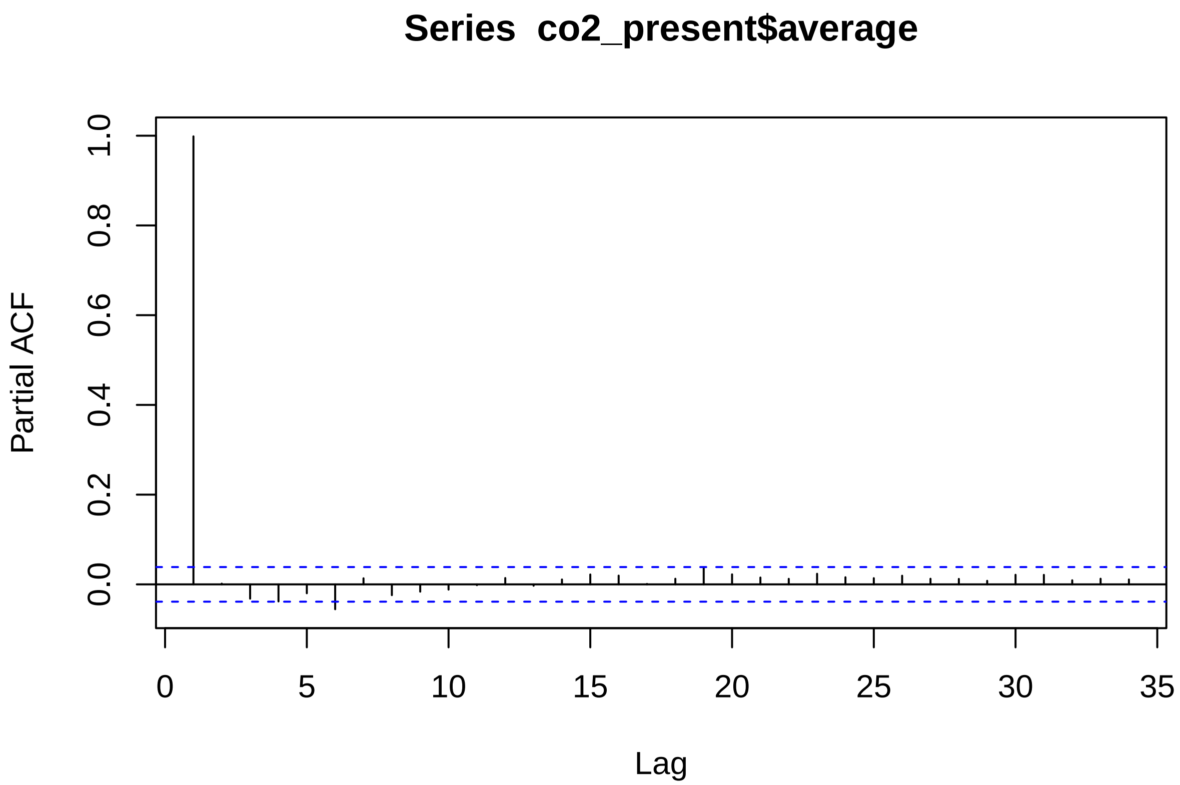
The plot shows a clear upward trend, indicating that CO2 levels have been increasing over the given time period. The pattern also shows seasonal fluctuations within each year, where the CO2 concentration peaks and then drops slightly, before rising again. The overall trend, however, is an increase in CO2 levels.

```r
acf(co2_present$average, na.action = na.pass)
```



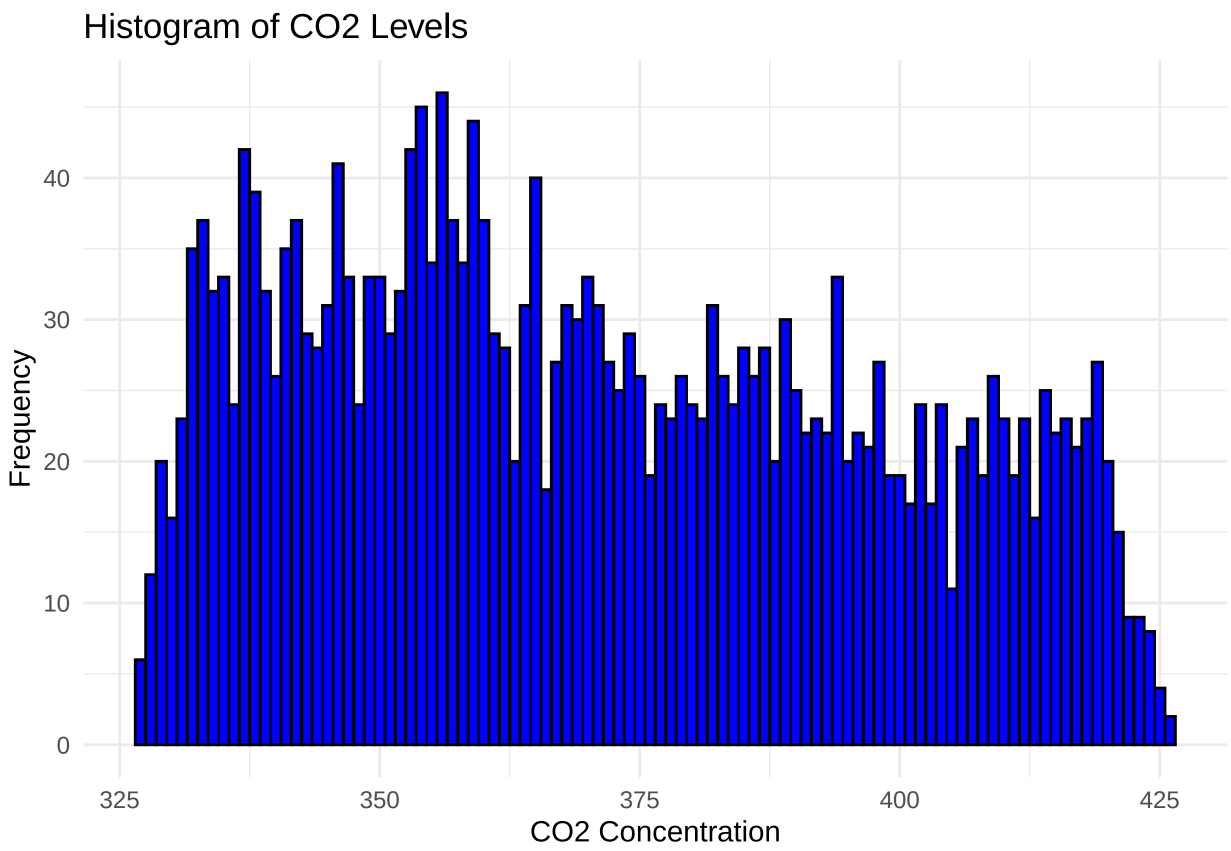
The plot shows a strong positive autocorrelation at all lags up to 35, and all are above the significance level, indicating a very persistent time series with a strong seasonal or cyclic pattern. This might suggest that the time series data of CO2 concentrations have a consistent pattern that repeats over time, with no significant decay in correlation as the lag increases.

```r
pacf(co2_present$average, na.action = na.pass)
```



The plot indicates that there is no almost no significant partial autocorrelation in the data at lags greater than zero. This could suggest that a simple autoregressive model may not be a good fit for the data.

```
ggplot(co2_present, aes(x = average)) +
  geom_histogram(binwidth = 1, fill = 'blue', color = 'black') +
  theme_minimal() +
  labs(title = "Histogram of CO2 Levels", x = "CO2 Concentration", y = "Frequency")
```



The plot shows multuple peaks suggesting more of a multimodal distribution. This could imply that there are multiple common CO2 levels within the data, possibly reflecting different environmental conditions or measurement periods. The data appears to be right-skewed.

#(1 point) Task 2b: Compare linear model forecasts against realized CO2

Descriptively compare realized atmospheric CO2 levels to those predicted by your forecast from a linear time model in 1997 (i.e. "Task 2a"). (You do not need to run any formal tests for this task.)

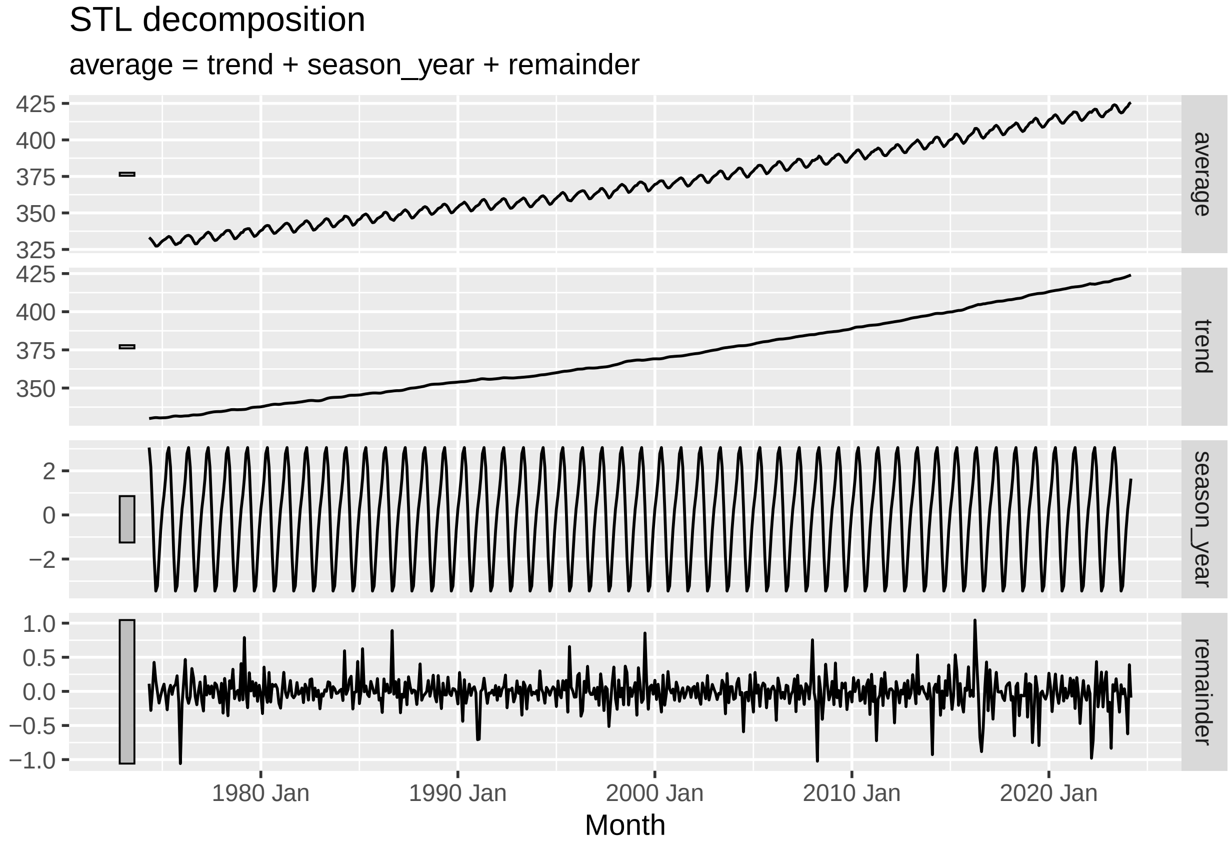
```
co2_present <- read_csv(weekly_co2_url, skip = header_end_index, show_col_types = FALSE)

co2_present <- co2_present[co2_present$average != -999.99, ]

co2_monthly <- co2_present %>%
  group_by(year, month) %>%
  summarise(average = mean(average, na.rm = TRUE),
            .groups = "drop")
```

#(1 point) Task 3b: Compare ARIMA models forecasts against realized CO2 Descriptively compare realized atmospheric CO2 levels to those predicted by your forecast from the ARIMA model that you fitted in 1997 (i.e. "Task 3a"). Describe how the Keeling Curve evolved from 1997 to the present.

#(3 points) Task 4b: Evaluate the performance of 1997 linear and ARIMA models In 1997 you made predictions about the first time that CO2 would cross 420 ppm. How close were your models to the truth? After reflecting on your performance on this threshold-prediction task, continue to use the weekly data to generate a month-average series from 1997 to the present, and compare the overall forecasting performance of your models from Parts 2a and 3b over the entire period. (You should conduct formal tests for this task.)

###Bo Edits starts here

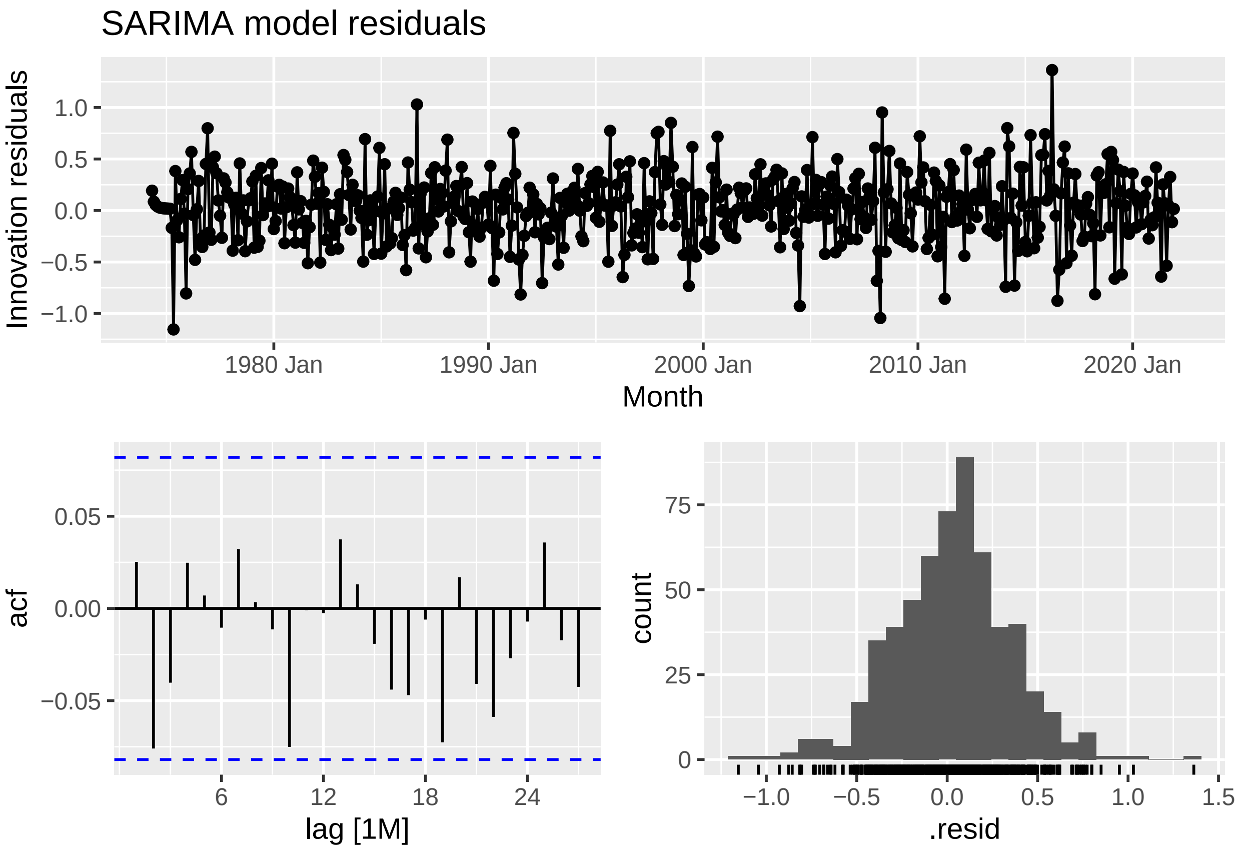
As we have observed in our EDA section and observations from 1997 report, we believe the CO2 levels have both overall upward trend and very possible seasonal trend. We used STL decomposition to decompose the data into 3 components: 1) upward trend using 6 months window, 2) seasonal trend which is observed as yearly trend, and 3) remainder, which the mean is observed close to zero, and fluctuates around reasonably bounded variance. We will proceed with further investigation to check for stationary on the remainder.



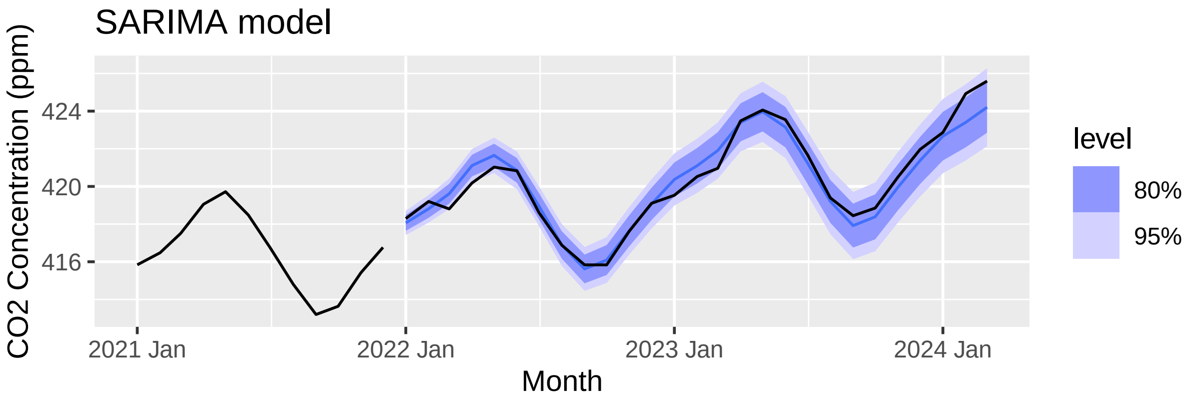
From the residual plots and the Box-Ljung test results in p-value of $0.04 < 0.05$, although the residuals of the decomposition appeared to be stationary, they do not appear to be completely white noise. This means that while the decomposition method eliminates the deterministic components from this specific time series, there are some correlation remains in the data. Combined with the knowledge that we learned from 1997 report, we decided to fit our data using SARIMA and polynomial time models with our new data.

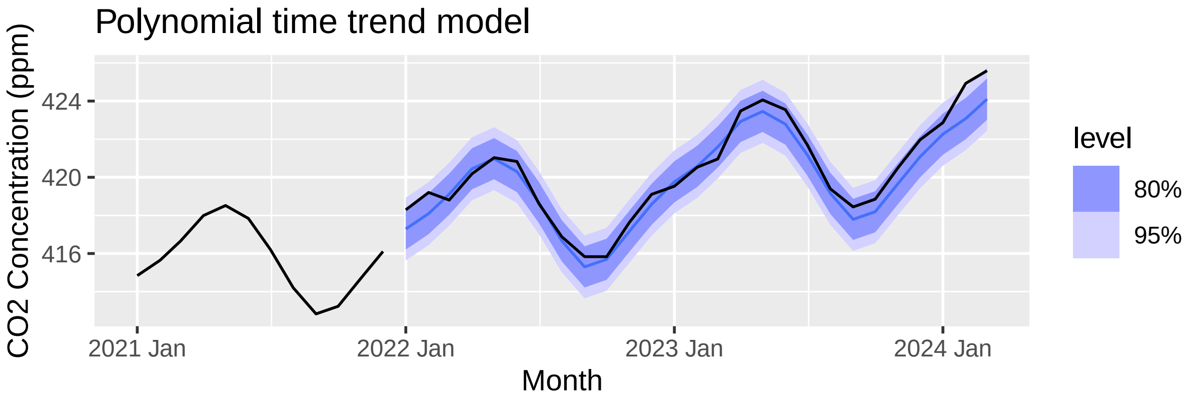
Prior to model building, we have divided up our data to use data points prior to 2022 as our training dataset, and post 2022 as our testing dataset. Then, we start fitting SARIMA and polynomial time models using our training data and test the accuracy using the testing data.

SARIMA Model



The residuals of SARIMA model appear to be close to white noise. There is no significant lags, but with some seasonal pattern that suggests we should run a statistical test on the residuals from the models to see if they are randomly distributed i.e. are white noise, which is what we want for a good model fit, or if they appear to have some serial correlation over time and violate the assumptions for a stationary time series fit. The Box-Ljung test result p-value of $0.53 > 0.05$ confirmed that the residuals of the SARIMA model are stationary and appeared to be white noise. And we performed the similar steps and test on Polynomial model confirmed the same, so we decided to use both models for forecasting and accuracy comparison.

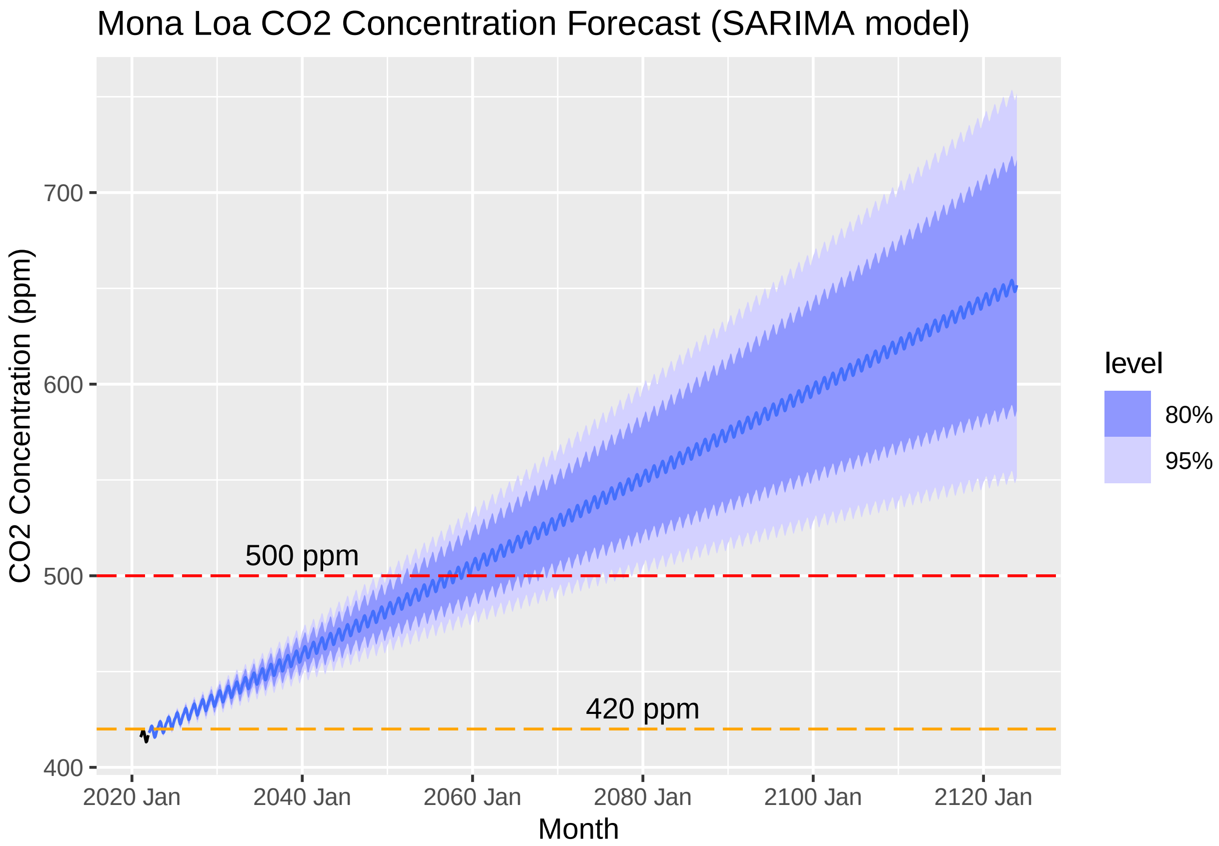




```
## # A tibble: 2 x 10
##   .model      .type      ME  RMSE   MAE    MPE  MAPE  MASE RMSSE  ACF1
##   <chr>       <chr>   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 .model      Test   0.0718 0.611 0.469 0.0167 0.111   NaN   NaN 0.574
## 2 trend_model Test   0.470  0.718 0.584 0.111  0.139   NaN   NaN 0.437
```

From the above accuracy table, it is evident that SARIMA model has outperformed polynomial time-trend model in terms of minimize the forecast error in the training dataset. As a result, we have decided to use SARIMA model to forecast out fo 2122.

Forecast to 2122



We can see that using our SARIMA model the forecasts also have a trend and seasonal movement and fluctuations increase overtime. But these forecasts will get very inaccurate as we move beyond at best 5 year forecasts of the model, as the confidence intervals of the model prediction begin to open up very widely due to the inherent model restriction and the unpredictability of the future. We can be pretty confident to say that we will hit 420 ppm CO2 level in 2023-2025, less confident about 500 ppm CO2 level as the model suggest it can be as early as 2045 or as late as 2060 or beyond.