W271 Group Lab

Bike share demand

Annie DeForge, Hannah Abraham, Mariah Ehmke, Nora Povejsil

Contents

1 The Keeling Curve

1.1 Introduction

The most recent UN Climate Conference known as COP28, resulted in international agreement to limit greenhouse gas emissions with the aim of limiting overall planetary global warming to 1.5 degrees Celsius. The plans rely on reducing greenhouse gas emissions from anthropogenic activities and increasing the uptake of carbon dioxide (CO2) from the atmosphere into natural storage systems, such as forests and soils (United Nations Climate Conference, 2023). Global levels of CO2 have more than doubled around the world compared to pre-industrialization levels (Jiménez-de-la-Cuesta and Mauritsen, 2019). This doubling has occurred despite seasonal growth and contractions in CO2 levels, particularly in areas with forests.

Seasonal variation in CO2 levels reflects the terrestrial cycles related to plant growth and decomposition. Keeling (1960) showed atmospheric local seasonal CO2 levels in Mauna Loa, Hawaii peaked right before a new growing season, steadily decreased as plants absorbed CO2 during the growing season, and reached a low at the end of the growing season. Despite these seasonal adjustments, baseline CO2 levels in Hawaii and throughout the world have grown over time. The growth reflects increased emissions from industrial, agricultural, and transportation systems. The goal of the COP28 is to contain CO2 levels through improved forestry and biodiversity management to absorb more CO2 from the atmosphere.

The objective of this analysis is to measure the trend in atmospheric CO2 levels in Mauna Loa, Hawaii, controlling for seasonal fluctuations. We test the following null hypotheses:

H01: Atmospheric CO2 concentrations at Mauna Loa, Hawaii follow a linear trend from 1957 to 2020. HA1: Atmospheric CO2 concentrations at Mauna Loa, Hawaii follow a non-linear trend from 1957 to 2020.

If the null hypothesis holds, it will provide insight into the rate at which atmospheric CO2 levels would decrease given reductions in CO2 emissions from industrial, agricultural, and transportation technologies. If it the trend is linear, we would expect a proportional decrease in CO2 levels from emission reductions following a reverse linear trend. If it is not linear, then CO2 levels will respond differently to emission caps. We will explore possible growth paths if the null hypothesis is rejected.

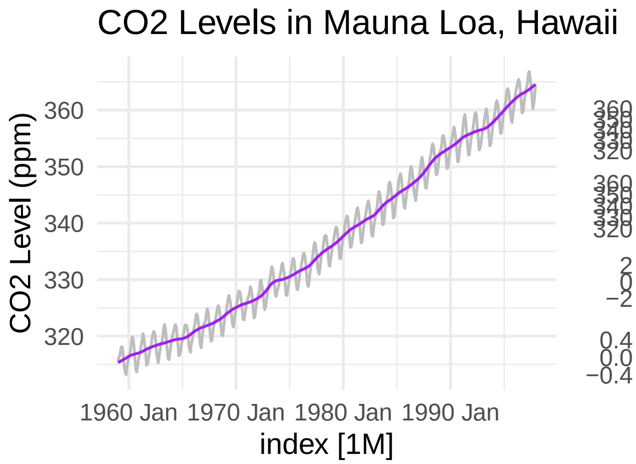
1.2 CO2 data

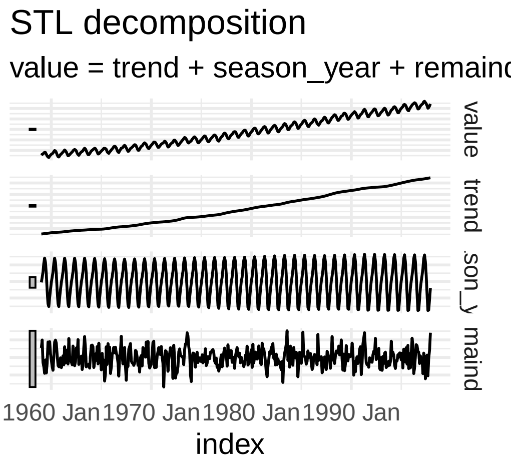
The data were collected at the Mauna Loa, Hawaii Observatory near the Mauna Loa Volcano in Hawaii. The amount of CO_2 is reported as the 'mole fraction' or number of carbon dioxide molecules present in a given number of molecules of air (National Oceanic and Atmospheric Administration, 2023). A CO_2 level of 400 indicates there are 400 parts per million (ppm) CO_2 molecules in every million molecules of dry air. The data collection began in 1957 by Dave Keeling (Keeling, 1960). The data are continuosly collected. We present data in part a that were monthly averages of CO_2 levels from 1957 to 1997. In part b, we extend our analysis using weekly averaged data from 1997 to the present.

1.2.1 Exploratory Data Analysis

The exploratory data analysis includes a linear plot of the raw data and the trend-cycle components, and the additive components of the time series (i.e., trend, seasonality, and the remainder). According to Keeling (YEAR), the data follow a seasonal, cyclical pattern. Each year, CO2 levels are lowest at the end of the growing season following plant growth and CO2 absorption. Then, the CO2 levels peak at the start of the growing season, before plants have begun absorbing excess CO2 from the atmosphere. The raw data follow this pattern, but with an upward trend, illustrated by the purple line in Figure 1.

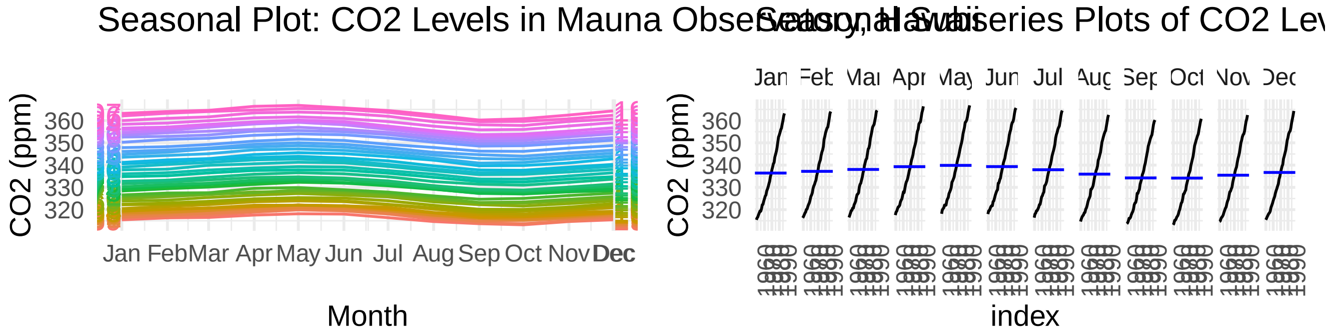
We employ seasonal, trend, and remainder (STL) decomposition to parse out the role of seasonality, trends in the average across time, and remainder components on the time series in the four plots on the right-hand side of Figure 1. The trend is semi-linear and upward sloping. The yearly average is increasing over time. The seasonal variation appears to expand across time as the height and, especially the depth, of the seasonal line plot increases after the mid-1970s. The bottom right-hand line plot of remainder variation does not have a discernible visual pattern.



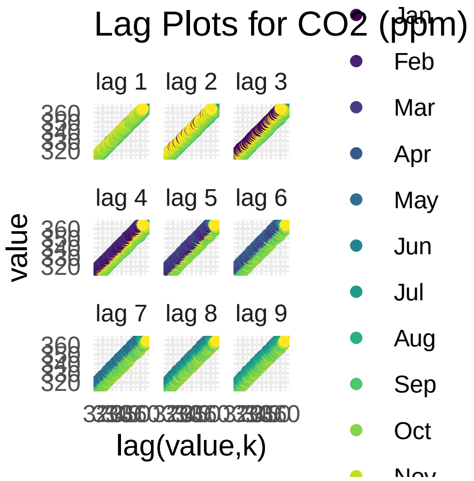


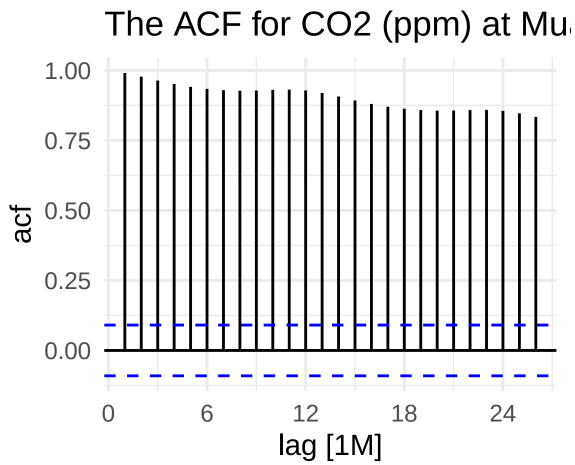
To analyze the trend year-by-year, Figure 2 provides another angle to view the upward trend in CO2 levels at Mauna Loa. The monthly CO2 (ppm) level was close to 315 ppm in 1959 and increased to upward of 365 ppm at the end of 1997. This is approximately a 16 percent increase in the concentration of CO2 at Mauna Loa across the nearly forty years of observations.

A break-down of the overall trend into time-trends in CO2 levels per month is presented in Figure 3. One will note CO2 levels appear to be lowest, on average in September and October and peak in May. This reflects the growing season patterns discussed by Keeling (1960). Although Keeling had fewer data points to consider, his observation transcends time. What is not clear is whether the upward trend in average CO2 levels is constant or increasing at a faster pace than in Keelings day. The trend line in Figure one appears to present as slightly convex or increasing at an increasing rate, especially after 1990. Also, the distance between lines in Figure 2 appears to increase in the 1990s as there is more white space between lines than previous decades, in particular when compared to the 1950s and 1960s.



Thus far, the data visualizations suggest a non-stationary process with strong seasonal trends. We now turn to lag and ACF plots to gauge possible auto-regressive tendencies in the data. We see strong correlation among the lag values in the lag-lag plot on the left-hand side of Figure 4. The first three month lags are nearly perfectly correlated with the time t CO_2 levels. From lags four to seven, there is more dispersion in lag correlates, but the linear, positive correlation remains strong. Finally, lags eight and nine realign with time t. The ACF supports a seasonal auto-regressive process. Lags decrease at a slow, but cyclical rate.





The EDA suggests we

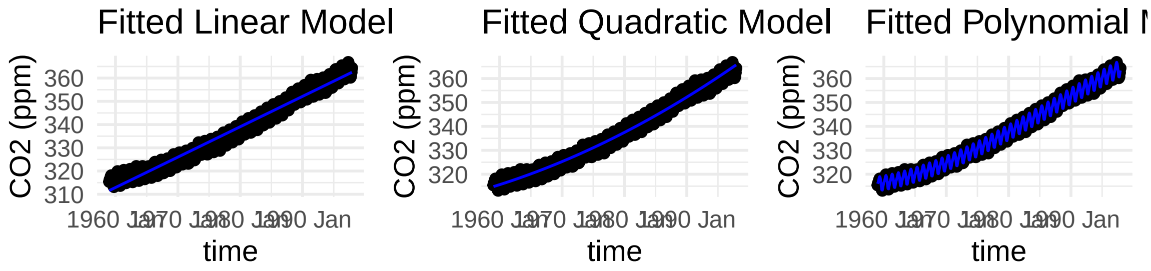
will need to consider a linear trend along with seasonal variation in the models of the carbon dioxide levels over time. Next, we build time-series models of carbon dioxide fluctuations to break down the seasonal variation and time trend to forecast future CO_2 accumulation at Mauna Loa.

##Models

We test a series of models to determine the best model to explain carbon dioxide levels and accumulation path at Mauna Loa. The general formulation of the models follows a linear time trend (equation 1), quadratic transformation (equation 2), and polynomial time series model (equation 3).

1.3 Linear time trend model

Fit a linear time trend model to the `co2` series, and examine the characteristics of the residuals. Compare this to a quadratic 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 to the year 2020.

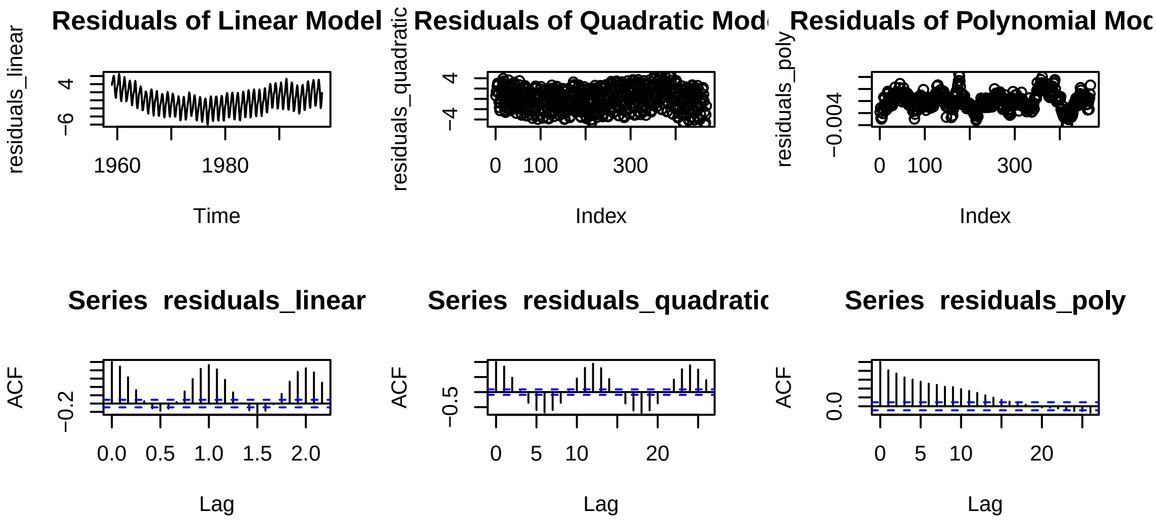

Fitted Linear Model

Fitted Quadratic Model

Fitted Polynomial M

```
##      mod_names    mod_aic              mod_bic
## [1,] "linear"     "2232.96081427566"  "2245.40621916342"
## [2,] "quadratic"  "735.409043710632"  "752.002916894303"
## [3,] "polynomial" "-6116.44609583328" "-6050.0706030986"
```

```
par(mfrow=c(2,3))
plot(residuals_linear, main = "Residuals of Linear Model")
plot(residuals_quadratic, main = "Residuals of Quadratic Model")
plot(residuals_poly, main = "Residuals of Polynomial Model")
acf(residuals_linear)
acf(residuals_quadratic)
acf(residuals_poly)
```


Residuals of Linear Model

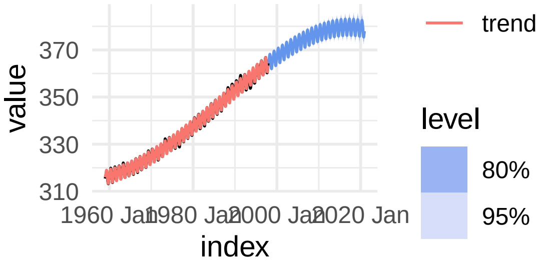
Residuals of Quadratic Mod

Residuals of Polynomial Moc

Series residuals_linear

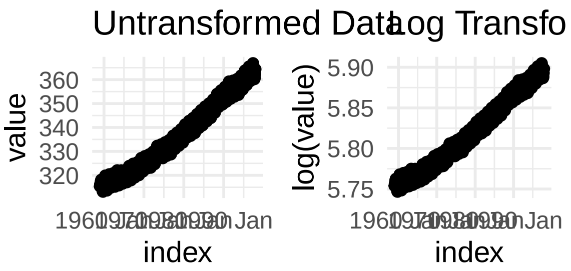
Series residuals_quadratic

Series residuals_poly



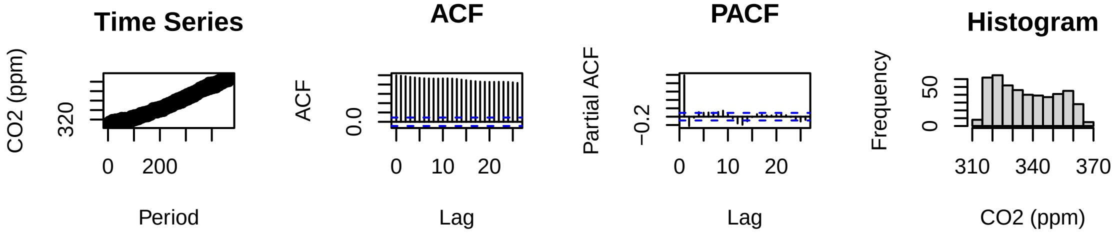
1.3.1 Log-Linear Model consideration

The data has a pretty linear trend with equal variance throughout the data. From the plot, you can see that a log transformation does not help with linearizing the data. Therefore, there is no transformation that is required for this data



1.4 (3 points) Task 3a: ARIMA times series model

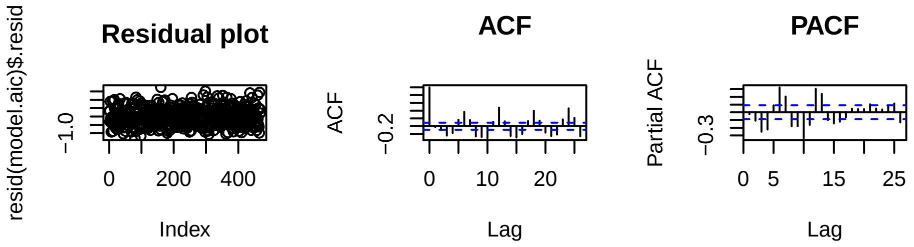
Following all appropriate steps, choose an ARIMA model to fit to the series. Discuss the characteristics of your model and how you selected between alternative ARIMA specifications. Use your model (or models) to generate forecasts to the year 2022.



Core plot write up The time series plot shows an increasing average as time increases. The histogram of Co2 values also shows an non-normal distribution. Both of these point to an increasing trend in the data. The ACF plot shows persistent significant lags, even for very large lag values, this is an attribute of an AR process. Additionally, the PACF plot quickly dies, but maintain an oscillating pattern, which is an attribute of an MA process. All of these factors together point towards an ARIMA process that underlies the Co2 time series that we wish to model.

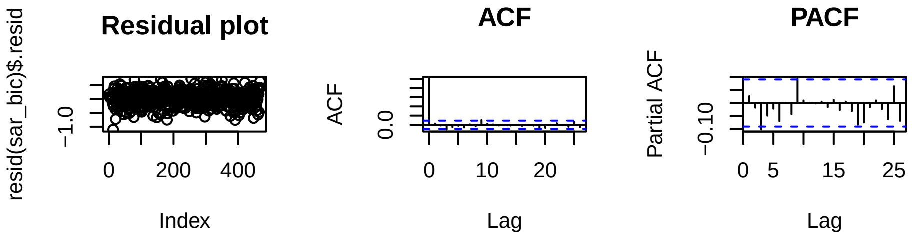
```
## Series: value
## Model: ARIMA(2,1,4) w/ drift
##
## Coefficients:
##          ar1      ar2      ma1     ma2     ma3     ma4  constant
##       1.6886  -0.9587  -1.3228  0.1540  0.1374  0.1909    0.0286
## s.e.  0.0137   0.0134   0.0481  0.0749  0.0902  0.0563    0.0039
##
## sigma^2 estimated as 0.2901:  log likelihood=-373.34
## AIC=762.68   AICc=762.99   BIC=795.85
```

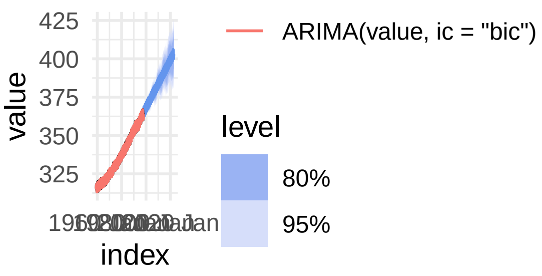
Model Interpretation Model selection across AIC, AICc, and BIC all chose an ARIMA model with 2 AR parameters, 4 MA parameters and linear differencing. However, from looking at the ACF and PACF plots, below there is signs that there are still unaccounted seasonality trends that could be incorporated into the model.



```
## Series: value
## Model: ARIMA(1,1,1)(1,1,2)[12]
##
## Coefficients:
##          ar1      ma1     sar1     sma1     sma2
##       0.2569  -0.5847  -0.5489  -0.2620  -0.5123
## s.e.  0.1406   0.1204   0.5879   0.5701   0.4819
##
## sigma^2 estimated as 0.08576:  log likelihood=-84.39
## AIC=180.78   AICc=180.97   BIC=205.5
```

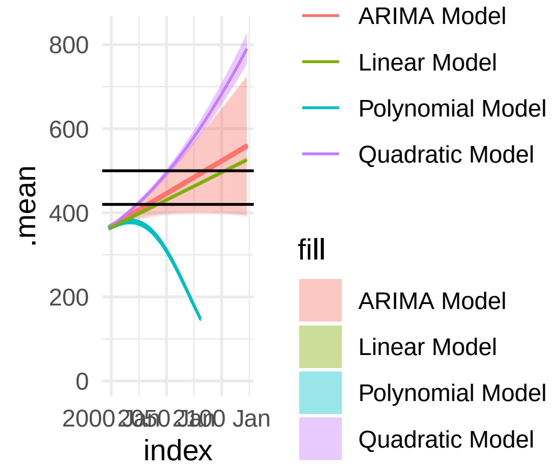
```
## Series: value
## Model: ARIMA(0,1,1)(1,1,2)[12]
##
## Coefficients:
##           ma1     sar1     sma1     sma2
##       -0.3482  -0.4986  -0.3155  -0.4641
## s.e.   0.0499   0.5281   0.5164   0.4366
##
## sigma^2 estimated as 0.08603:  log likelihood=-85.59
## AIC=181.18   AICc=181.32   BIC=201.78
```





1.5 (3 points) Task 4a: Forecast atmospheric CO2 growth

Generate predictions for when atmospheric CO2 is expected to be at 420 ppm and 500 ppm levels for the first and final times (consider prediction intervals as well as point estimates in your answer). Generate a prediction for atmospheric CO2 levels in the year 2100. How confident are you that these will be accurate predictions?



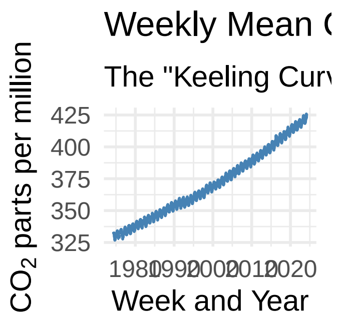
```
##                  420: Predicted first time first time 95% CI
## Linear Model     "2041 Dec"                "( 2037 Dec , 2046 Jan )"
## Quadratic Model "2023 Jun"                "( 2021 Apr , 2025 Oct )"
## ARIMA Model      "2031 May"                "( 2020 May , N/A )"
##                  420: Predicted last last last time 95% CI
## Linear Model     "2042 Aug"               "( 2038 Aug , 2046 Oct )"
## Quadratic Model "2023 Oct"               "( 2021 Aug , 2026 Feb )"
## ARIMA Model      "2035 Oct"               "( 2022 Nov , N/A )"
```

```
##                  500: Predicted first time 95% CI
## Linear Model     "2103 Feb"                "( 2098 Oct , 2107 Aug )"
## Quadratic Model "2051 Nov"                "( 2048 Nov , 2055 Apr )"
## ARIMA Model      "2083 Apr"                "( 2051 Mar , N/A )"
##                  500: Predicted last time 95% CI
## Linear Model     "2103 Nov"               "( 2099 Jun , 2108 Apr )"
## Quadratic Model "2052 Feb"               "( 2049 Jan , 2055 Jul )"
## ARIMA Model      "2087 Sep"               "( 2052 Jan , N/A )"
```

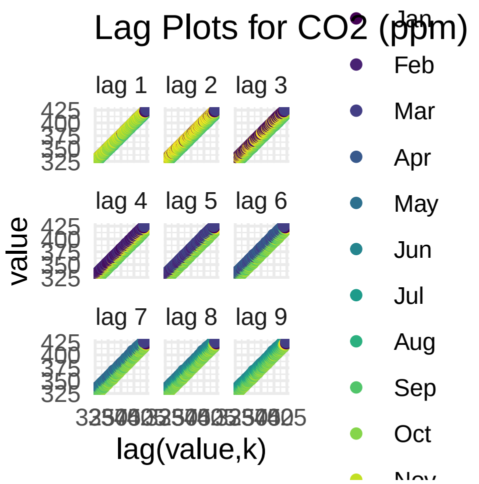
2 Report from the Point of View of the Present

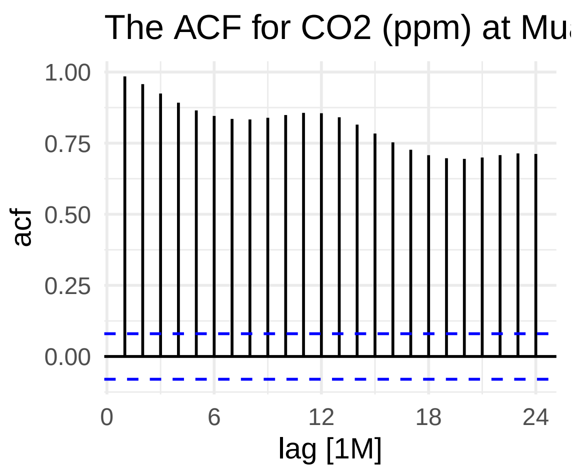
2.1 (1 point) Task 0b: Introduction

2.2 (3 points) Task 1b: Create a modern data pipeline for Mona Loa CO2 data.

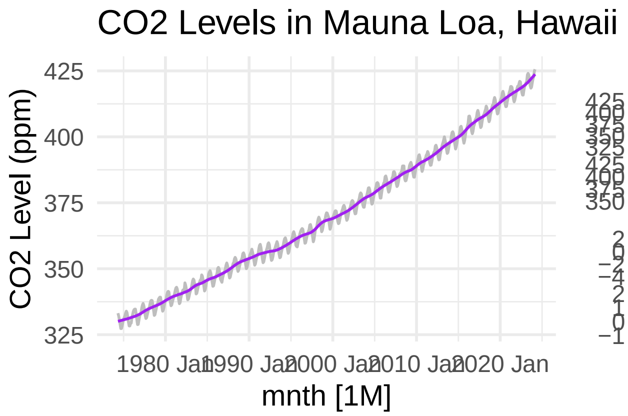


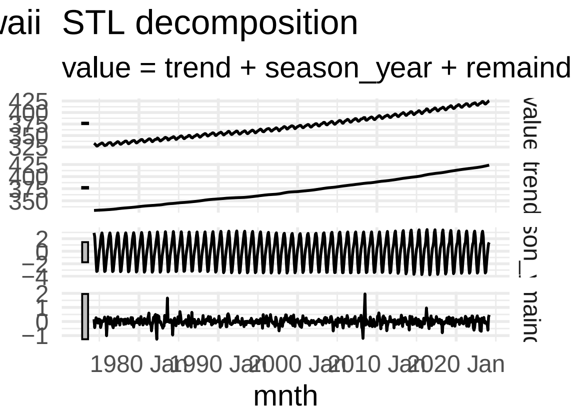
```
## Warning: Removed 10 rows containing missing values (gg_lag).
```

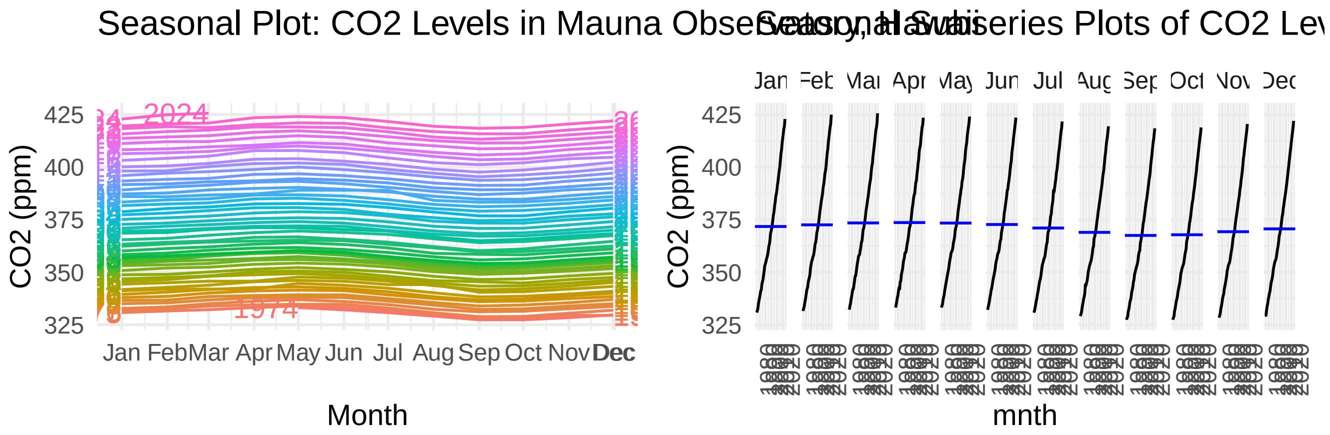




```
## Warning: na.locf will be replaced by na_locf.
##     Functionality stays the same.
##     The new function name better fits modern R code style guidelines.
##     Please adjust your code accordingly.
```

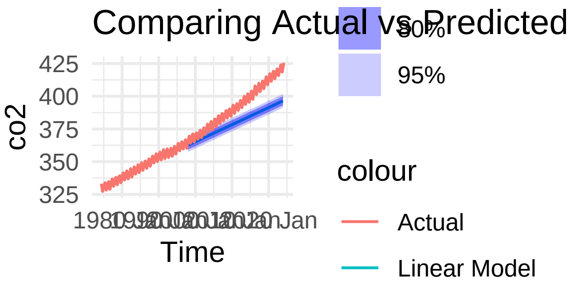




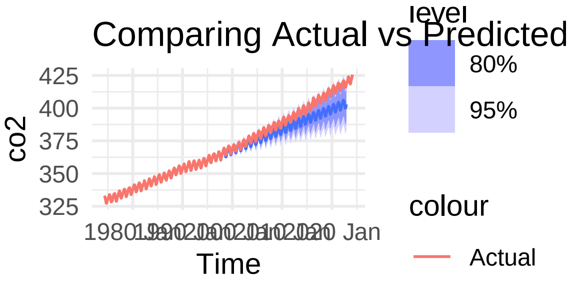


## (1 point) Task 2b: Compare linear model forecasts against realized CO2

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.)



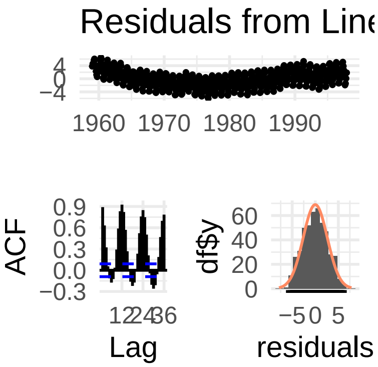
2.3 (1 point) Task 3b: Compare ARIMA models forecasts against realized CO2



2.4 (3 points) Task 4b: Evaluate the performance of 1997 linear and ARIMA models

#Linear Model Eval

```
##                         ME      RMSE       MAE          MPE      MAPE      MASE
## Training set 4.373498e-15  2.612462  2.146882 -0.005352854 0.6392463  1.994736
## Test set     1.269036e+01 14.526501 12.690362  3.166701389 3.1667014 11.791014
##                   ACF1
## Training set 0.8910172
## Test set            NA
```

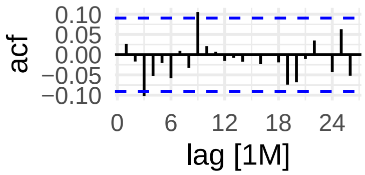


```
##
##  Breusch-Godfrey test for serial correlation of order up to 24
##
## data:  Residuals from Linear regression model
## LM test = 457.32, df = 24, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
##
## data:  residuals_linear
## X-squared = 850.26, df = 10, p-value < 2.2e-16
```

#ARIMA Model Eval

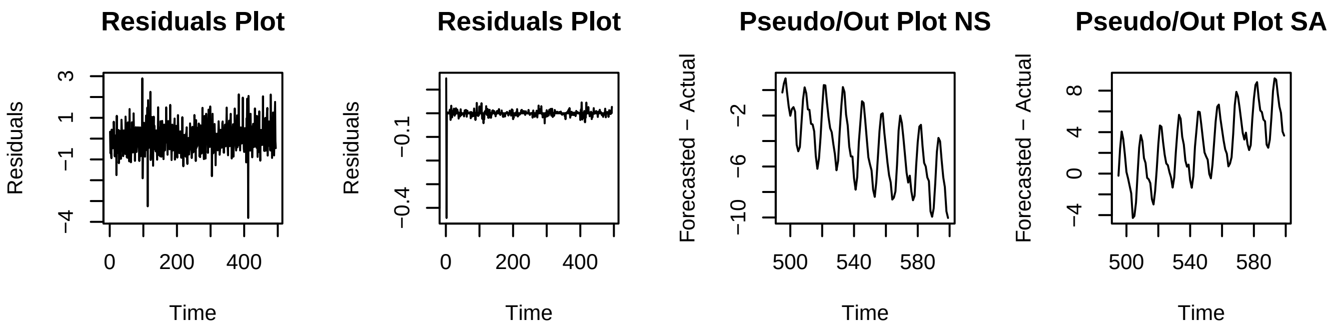
```
##                 ME     RMSE      MAE      MPE     MAPE
## Test set 6.979727 8.573007 6.979727 1.730336 1.730336
```

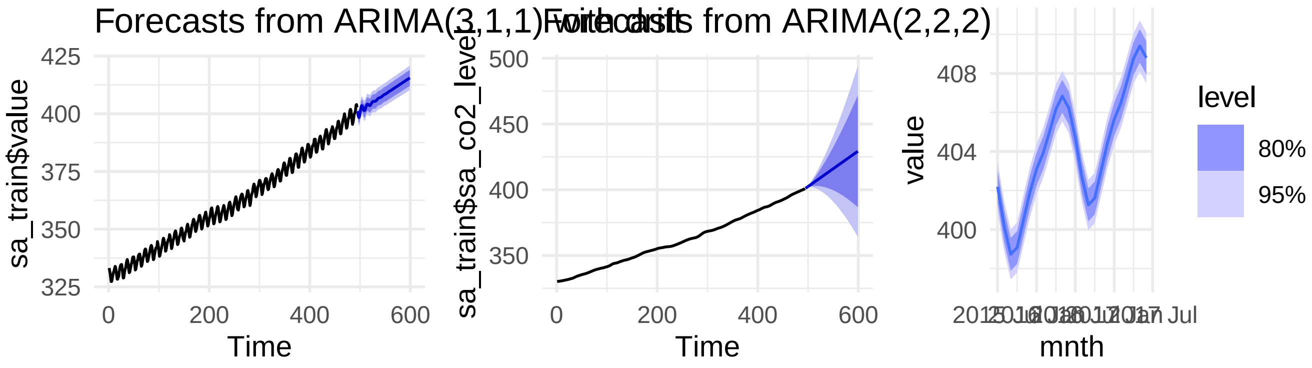


```
##
##  Box-Ljung test
##
## data:  resid.sar
## X-squared = 14.681, df = 10, p-value = 0.1441
```

2.5 (4 points) Task 5b: Train best models on present data

```
## Warning in sa_poly_forecast$.mean - sa_test$value: longer object length is not
## a multiple of shorter object length

## [1] "RMSE for ARIMA model (NSA):"

## [1] 5.13033

## [1] "RMSE for ARIMA model (SA):"

## [1] 4.140413

## [1] "RMSE for Polynomial Time-Trend model (SA):"

## [1] 10.51142
```





2.6 (3 points) Task Part 6b: How bad could it get?

```
##                                   420: Predicted first time
## ARIMA Not Seasonal Adjusted Model "2026 Oct"
## ARIMA Seasonally Adjusted Model   "2021 May"
## Polynomial Model                  "2021 May"
##                                  95% CI
## ARIMA Not Seasonal Adjusted Model "( 2023 Oct , 2030 Mar )"
## ARIMA Seasonally Adjusted Model   "( 2018 Feb , N/A )"
## Polynomial Model                  "( 2022 Jan , 2023 Jan )"
##                                   420: Predicted last time
## ARIMA Not Seasonal Adjusted Model "2027 Apr"
## ARIMA Seasonally Adjusted Model   "2021 Aug"
## Polynomial Model                  "2023 Sep"
##                                  95% CI
## ARIMA Not Seasonal Adjusted Model "( 2024 Mar , 2030 Sep )"
## ARIMA Seasonally Adjusted Model   "( 2018 Mar , N/A )"
## Polynomial Model                  "( 2022 Aug , 2023 Aug )"
```

```
##                  Predicted first time 95% CI
## ARIMA (NSA)      "2073 Jan"           "( 2068 Apr , 2078 Mar )"
## ARIMA (SA)       "2046 Jan"           "( 2024 Aug , N/A )"
## Polynomial Model "2042 May"           "( 2042 Feb , 2043 May )"
##                  Predicted last time 95% CI
## ARIMA (NSA)      "2073 Jul"          "( 2068 Sep , 2078 Sep )"
## ARIMA (SA)       "2046 Mar"          "( 2024 Aug , N/A )"
## Polynomial Model "2043 Oct"          "( 2042 Oct , 2044 Oct )"
```