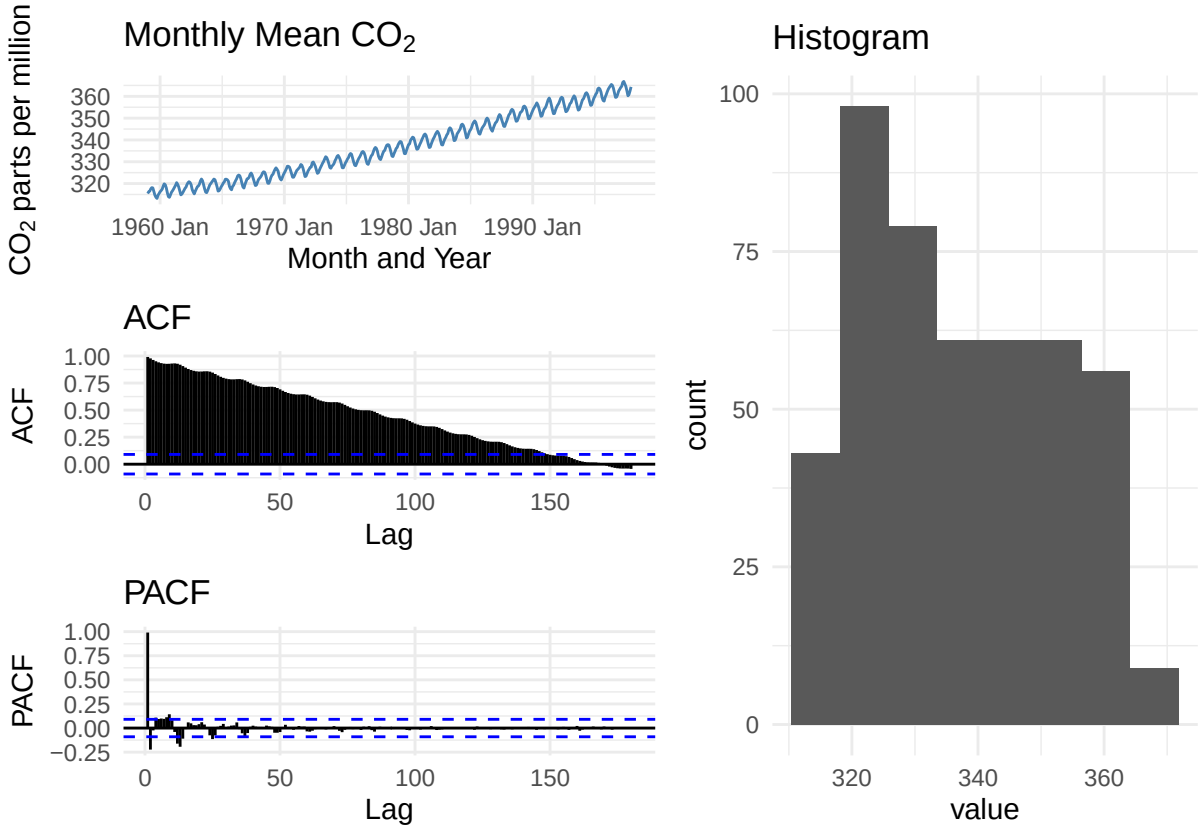
The data that we will analyze in this report is monthly data on the CO2 levels made at the Mauna Loa observatory from Jan 1959 to Dec 1997. According to the data documentation, the air at Mauna Loa is thought to be representative of much of the Northern Hemisphere and potentially the globe as well, as the observatory is at an altitude of 3400 meters and surrounded by bare lava, which allows for measurement of "background" air that is resistent to day-to-day fluctuations in CO2 levels.

The data is in units of "mole fraction", which according to the data source is "defined as the number of carbon dioxide molecules in a given number of molecules of air, after removal of water vapor. For example, 413 parts per million of CO2 (abbreviated as ppm) means that in every million molecules of (dry) air there are on average 413 CO2 molecules." The data that makes up our dataset was measured daily from Jan 1959 to Dec 1997, but in our dataset appears as an averaged mean per month in ppm units.

Let's take a look at the data. In raw form, it appears as a matrix of doubles that represent the ppm measurements per month and year combination, as appears below. Let's create some initial EDA plots that will allow us to better understand the data. Let's start by analyzing time series, histogram, auto-correlation function (ACF), and partial auto-correlation function (PACF) plots.



As we can see above, the data seems to follow a clear and increasing trend, with a distinct seasonal pattern that appears as "waves" that we would like to further analyze. The magnitude of the fluctuations do not appear to vary with the time series level, so in terms of decomposition, an additive model would likely fit this series best. The time series does not appear stationary from this plot, as the mean does not appear constant and in fact appears to increase over time, but the variance appears to be constant.

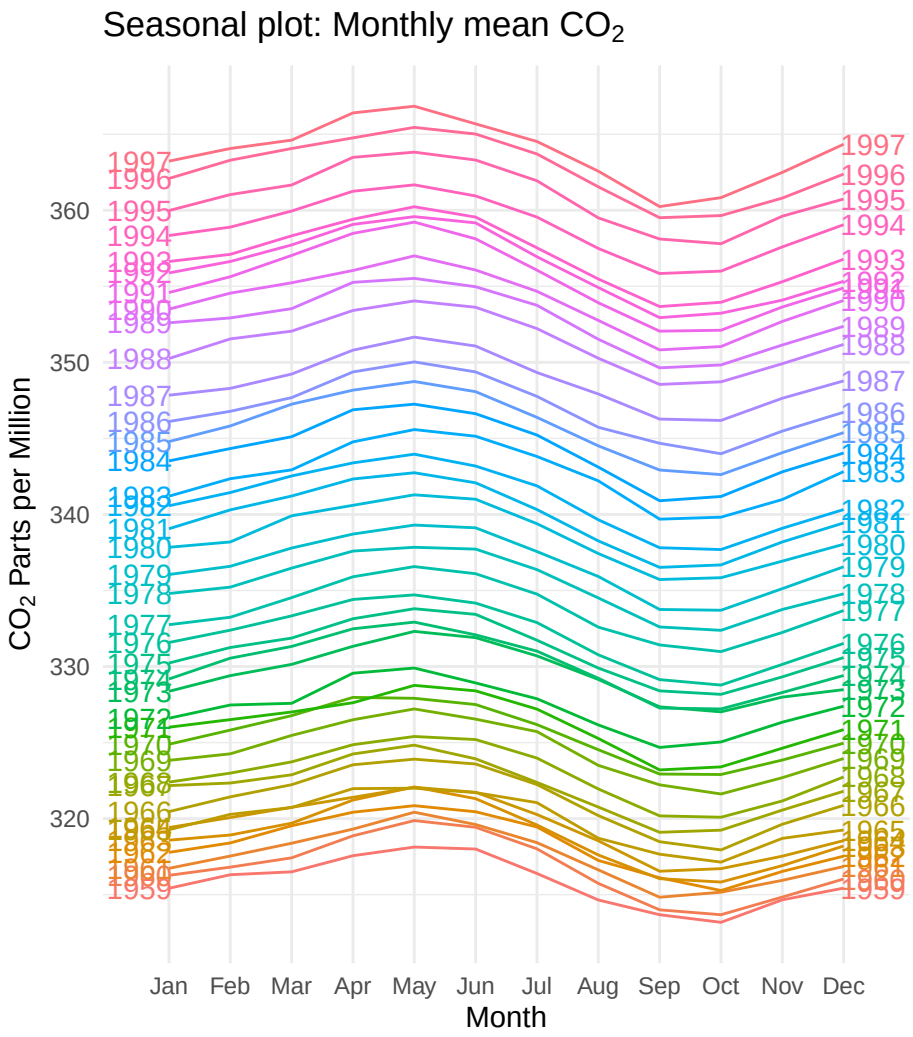
The ACF starts out close to 1 and declines slowly over time, losing significance but staying mostly positive and above the significance line until around lag 140. This slow decline in ACF is what we should see in a time series with a pronounced trend effect, and tracks with what we noticed in the previous time series plot. There appear to be "waves" in the ACF plot similar to the "waves" we also noticed in the time series plot, indicating that the there is a seasonal or cyclic component to our data. Thinking ahead to our modeling, the

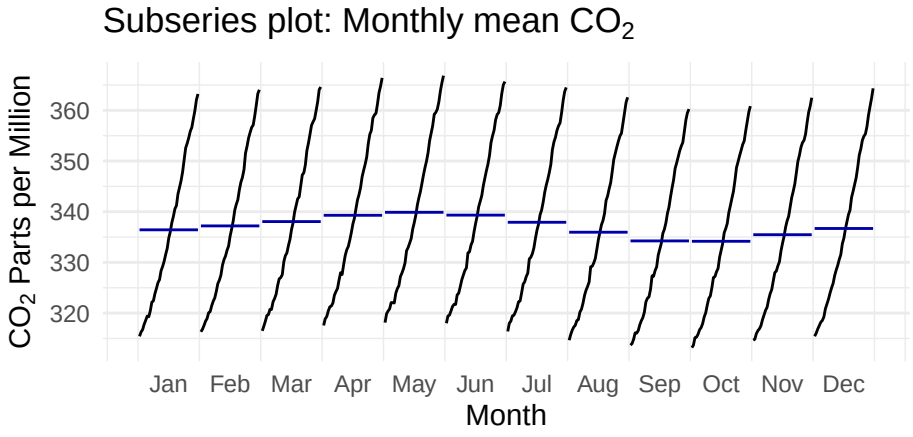
ACF seems to indicate an autoregressive component in our data generating process, as it declines slowly over time.

The PACF starts out with a single significant positive spike at lag 1, followed by a (relatively smaller) significant negative spike at lag 2, with oscillating clusters of positive and negative lags with much lower levels of significance as the lag number increases. Thinking ahead to our modeling, the PACF seems to indicate an moving average component in our data generating process, as it oscillates between positive and negative over time.

The histogram shows that the CO2 values seem to range between 300 and 380 ppm and do not appear normally distributed, with most values in between these two ranges.

Let's take a closer look at the seasonality in the data. We'll start by creating a season plot and a subseries plot of the data.

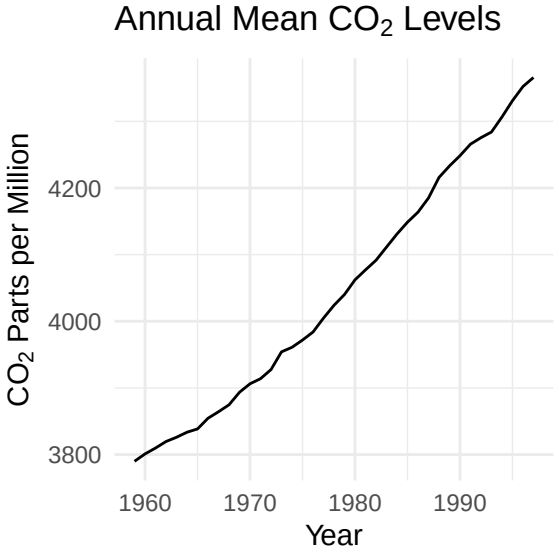




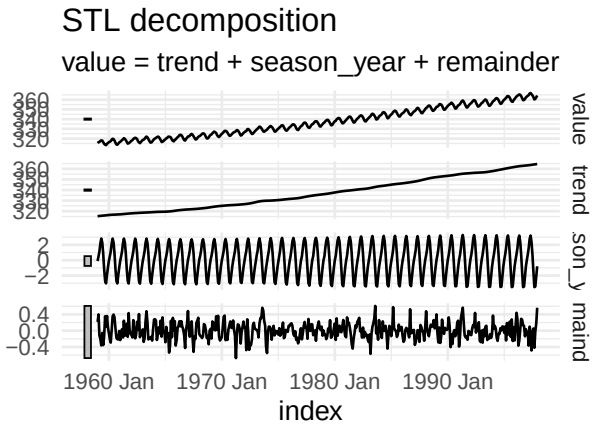
From these plots, we can see that the CO2 levels appear to increase from January through May, then decrease from June through October, hitting a low in October, and then increase again from November through end of the year. The source report does mention that plants and soil absorbing and emitting CO2 could influence these measurements. A possible explanation could be that, in the colder months, we can expect higher CO2 levels as plants die off, and in the warmer months, we can expect lower CO2 levels as plants thrive. Of course,

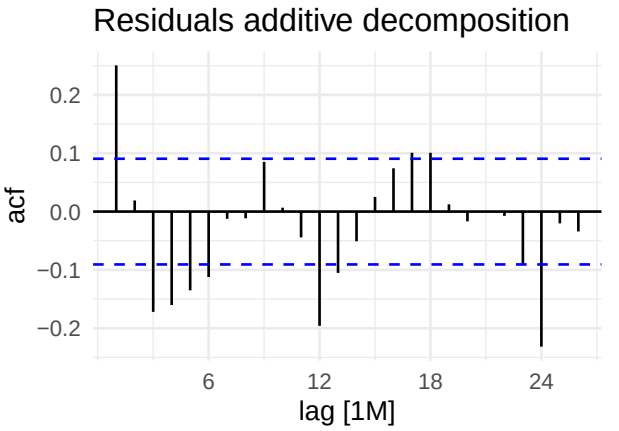
there is variability across the Northern Hemisphere in what counts as "colder" months and when plants thrive - for example, Hawaiian "winter" is temperate and plants can grow year round, so this could explain why the seasonal variability is slight across the year. We can again see the trend of CO2 levels increasing year by year, but the seasonality effect seems constant year after year without a noticeable increase in the magnitude of the fluctuations across years, again supporting an additive model.

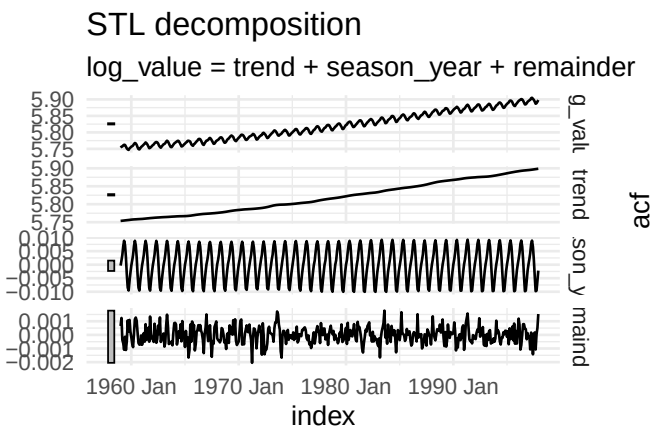
We can confirm that our time series trend is increasing by removing our seasonality components and aggregating our data by year instead of month, as seen below.

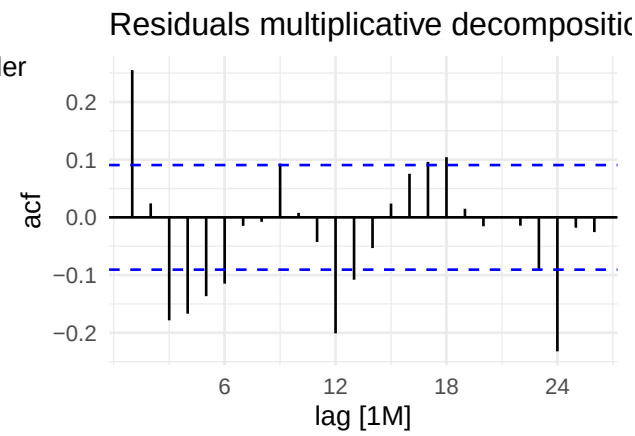


We can also use both additive and multiplicative decomposition to remove both the trend and seasonal movements from our dataset and confirm that the variance is stationary. The results from these decompositions are plotted below.









From these plots, we can confirm again that the time series is trending upwards, as seen in the "trend" sections of the decomposition plots. However, upon closer look, the fluctuations within the seasonality of the time series seem to grow slightly larger over time, which we can see in the "season_year" section of the top left plot. In the multiplicative plot on the bottom left, the "season_year" plot appears more stable, supporting the idea that this might actually be a multiplicative time series.

Looking at the residual plots, the residuals on both appear stationary, meaning that the decomposition methods we are using was able to eliminate deterministic components from the time series. However, they do not appear to be white noise, meaning that there is still correlation in the data.

Let's complete our EDA by running statistical tests to determine whether our model is stationary or non-stationary. We will run both the Augmented Dickey-Fuller (ADF) test and the Phillips Perron (PP) test to do this, under the following hypotheses:

H0: Time series is non-stationary

H1: Time series is stationary

```
##
##  Augmented Dickey-Fuller Test
##
## data:  co2
## Dickey-Fuller = -6.842, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary
```

```
##
##  Phillips-Perron Unit Root Test
##
```

```
## data:  co2
## Dickey-Fuller Z(alpha) = -92.68, Truncation lag parameter = 5, p-value
## = 0.01
## alternative hypothesis: stationary
```

Based on the ADF and PP tests, we can reject the null hypothesis that the time series is non-stationary. This is surprising as from our visual analysis of the time series plots, the time series does not appear to be stationary as the mean trends upwards. Because we know that both the ADF and PP tests have low power, we will move forward with the assumption that this time series is non-stationary based on our visual EDA.

(3 points) Task 2a: 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.

> Would be appropriate - look at week 7, because we think this is additive we should not take the logarithm
>
> https://github.com/mids-271/summer_22_central/blob/master/Live_session_and_solutions/LS_7_Solutions/LS-7-Solutions.pdf - pg 12 & 13