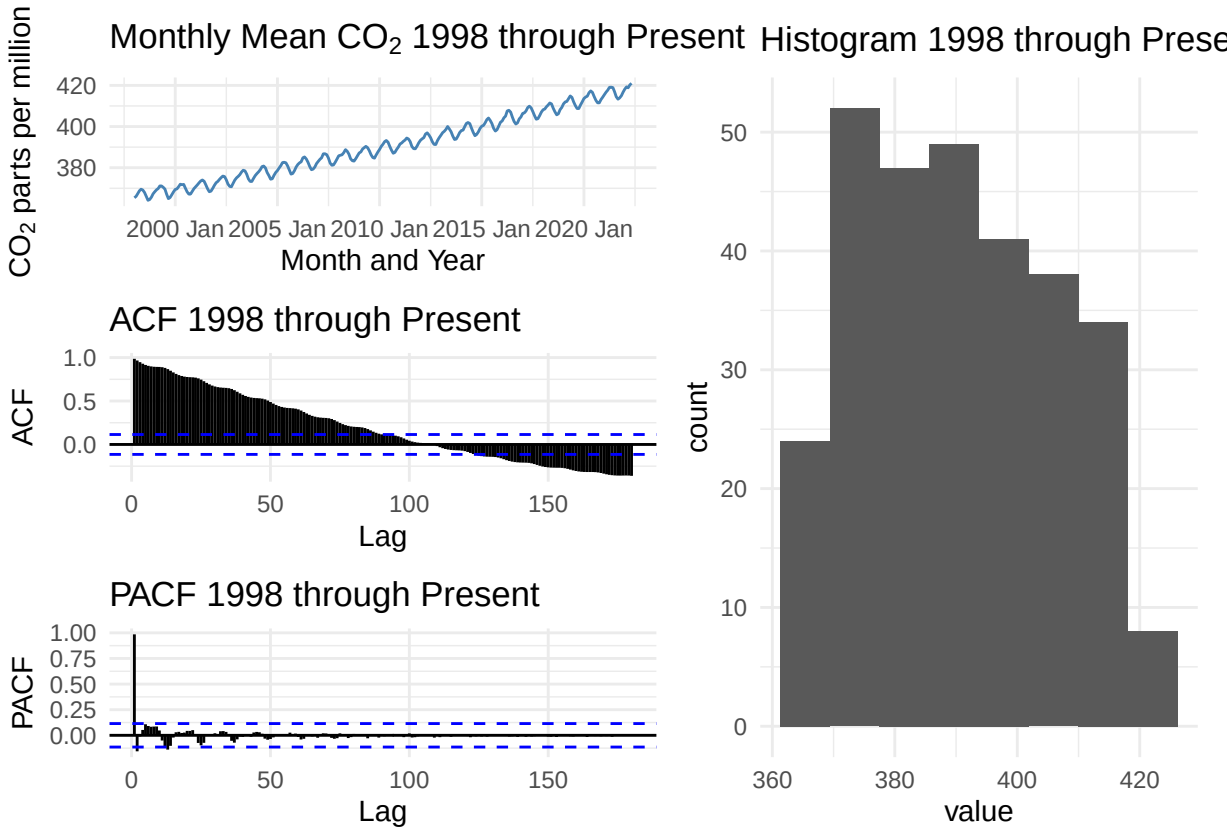
(1 point) Task 0b: Introduction

We now ask the question of whether the data collected from January 1998 to the present has different characteristics which could lead us to additional or different conclusions than those which the data set from 1958 until 1997 yielded.

We assume that the collection process for both datasets is identical.

It is worth noting that the "present" data set contains only about 24 years, or 293 months, compared to the 39 years, or 468 months contained in the "1997" dataset.

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



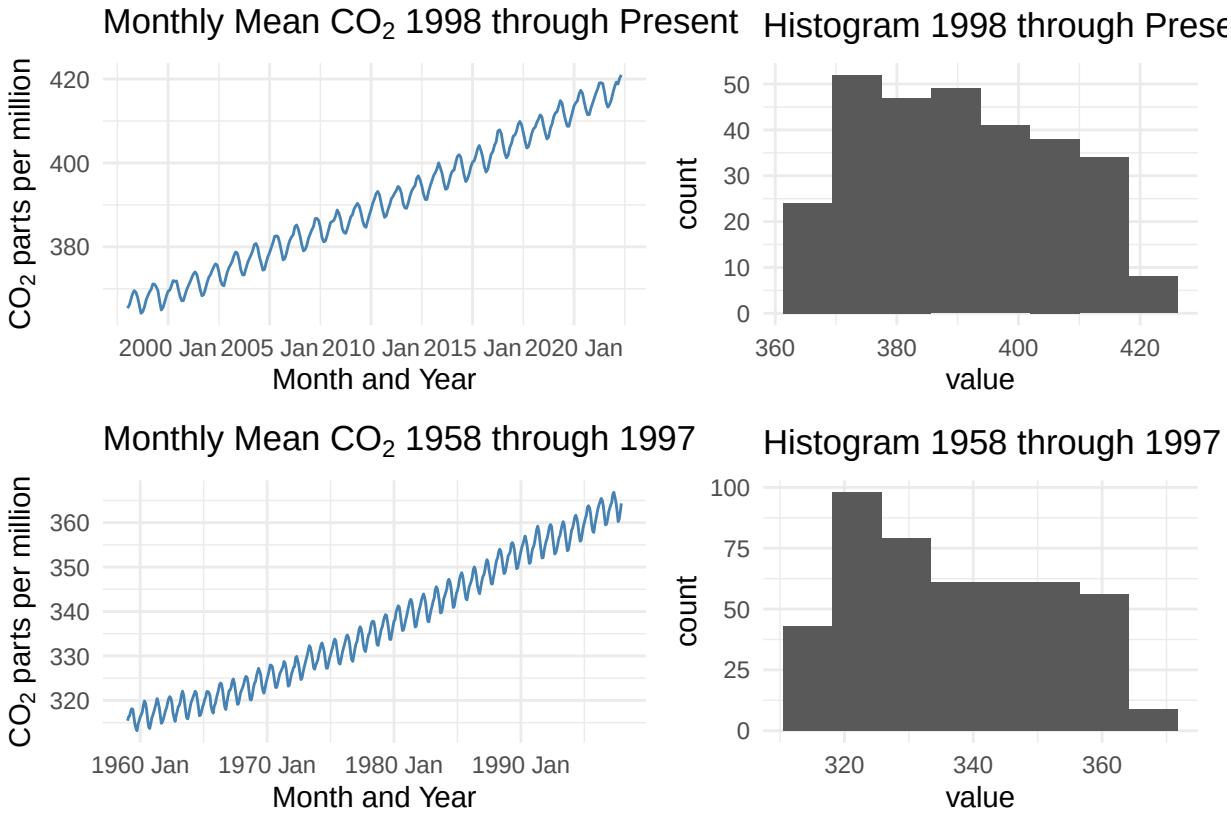
As with the "1997" data set, we see a clear increasing trend with the same seasonal "wave" pattern. Seasonal fluctuation does not appear to increase with time, suggesting, again, that an additive model explains the decomposition in this series. The time series is not stationary, as its mean increases. Variance presents as constant.

The ACF displays a slow, over-time, decline consistent with a trending time series. While the ACF plot of the "1997" plot remains above the significance line until lag 140, this ACF plot crosses the significance line at close to 100. As mentioned above, this "present" data set contains only 24 years, as opposed to 39, which is possibly the cause of this cross in the significance line at close to 100.

The PACF has a large positive spike at lag 1, followed by a smaller negative spike at log 2, then oscillates between positive and negative lags with progressively lower levels of significance.

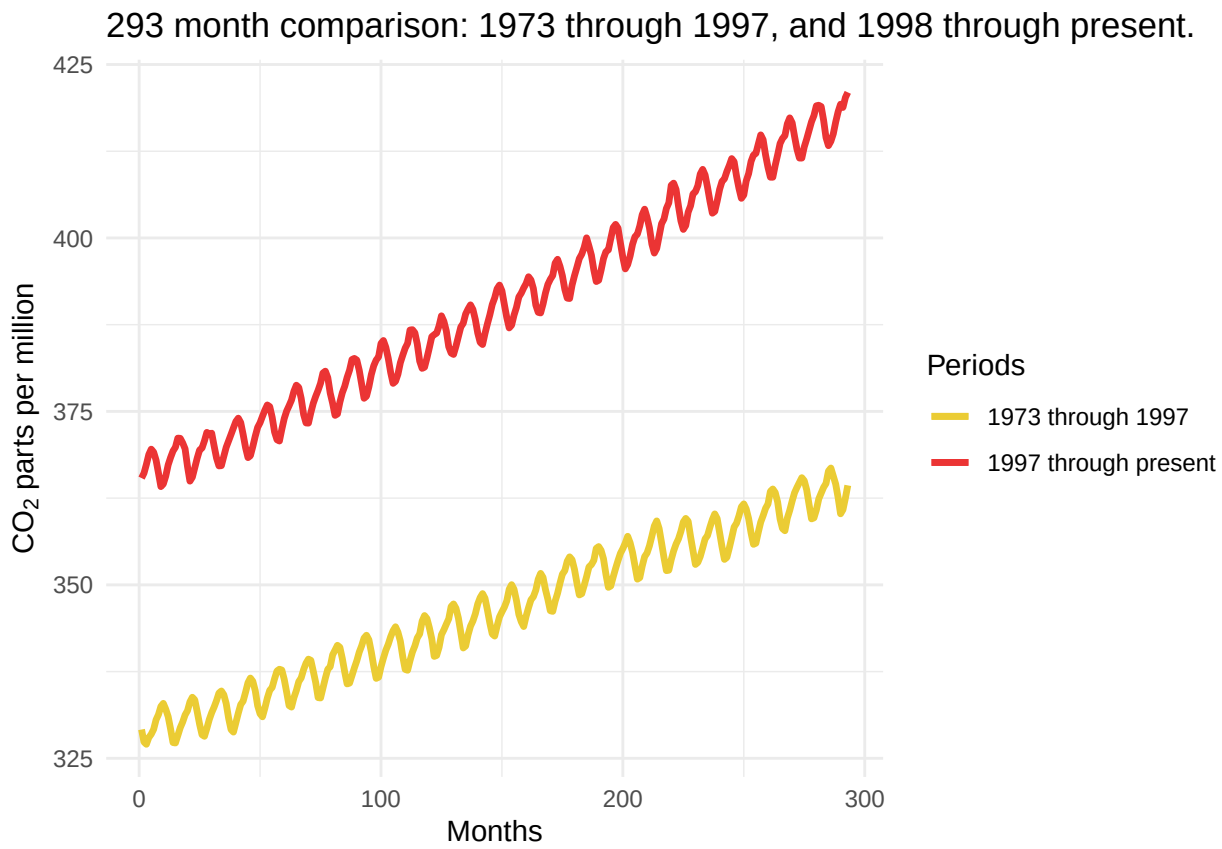
Our histogram of has no values lower than 365, or nigher than 421, and that these values are not normally distributed.

In order to better understand the differences between these two datasets we examine them side by side.



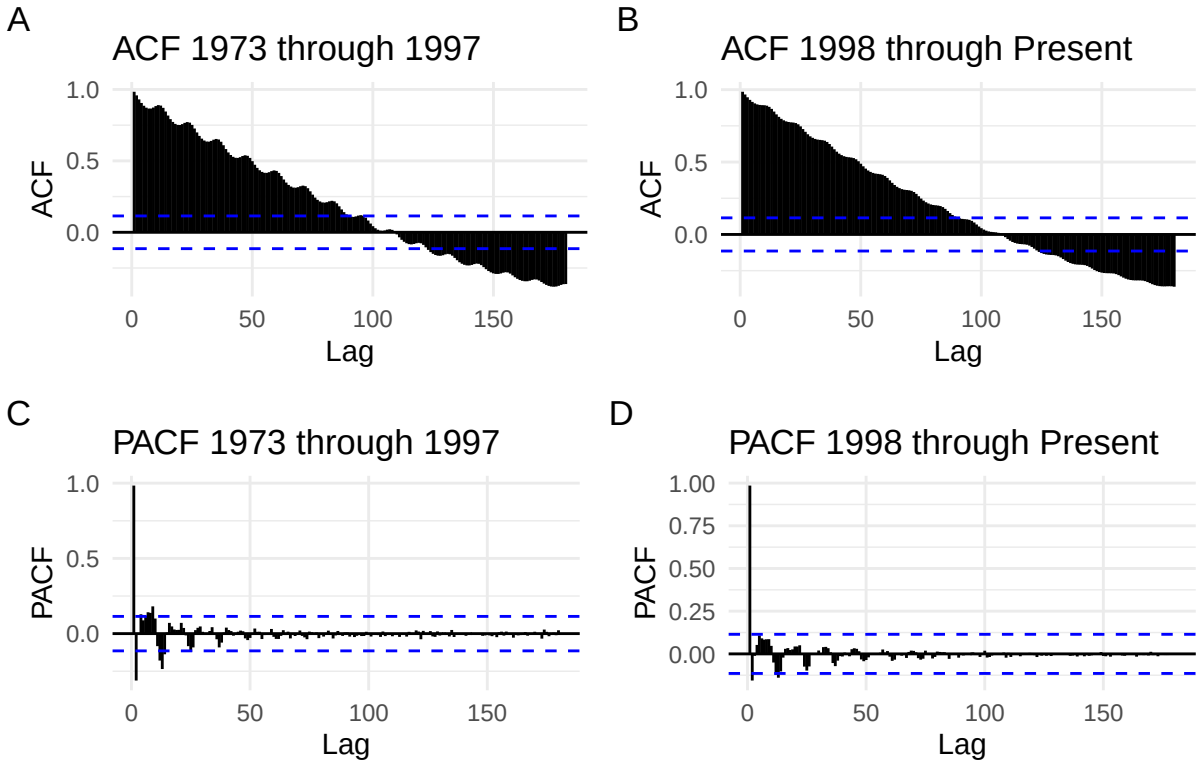
When we start looking graphing the two data sets side by side, the difference of one containing 39 years, and the other 24, begins to hamper our understanding.

We take only the last 293 months of the "1997" data set and graph it on the same space as the "present" data set.



In the 293 months prior to Dec 31, 1997, the atmospheric CO2 rose by 39.82 PPM. By comparison, in the 293 months prior to May 31, 2022 (the most recent month we have a measuring for), the atmospheric CO2 rose by 56.83 PPM. This is a difference of 17.01 PPM during the same time interval. This could indicate an acceleration in the increase in PPM.

We apply the same "parsing" of the original "1997" data set and compare the parsed dataset's ACF and PACF graphs to those of the "present" data set.

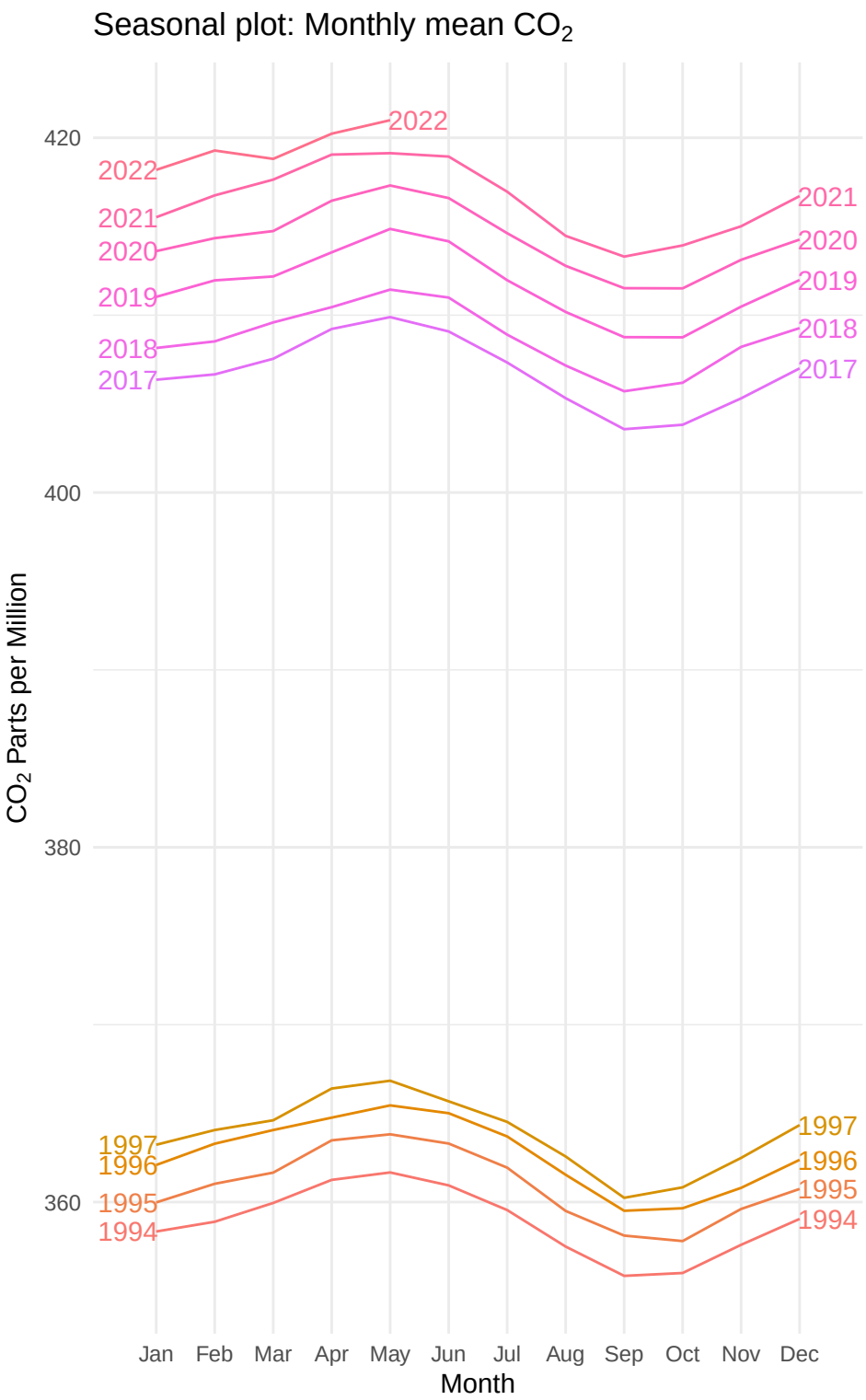


Parsing the "1997" data set to include 293 months only yields a ACF which crosses the significance line at about the same lag as the ACF plot of the "present" data set does. We notice that the degree of oscillation on the "1997" ACF is greater than the degree of osilation on the "present" data set.

Next we examine the seasonality of the "present" comparing it to the "1997" seasonality.

```
## Warning: Removed 228 row(s) containing missing values (geom_path).
```

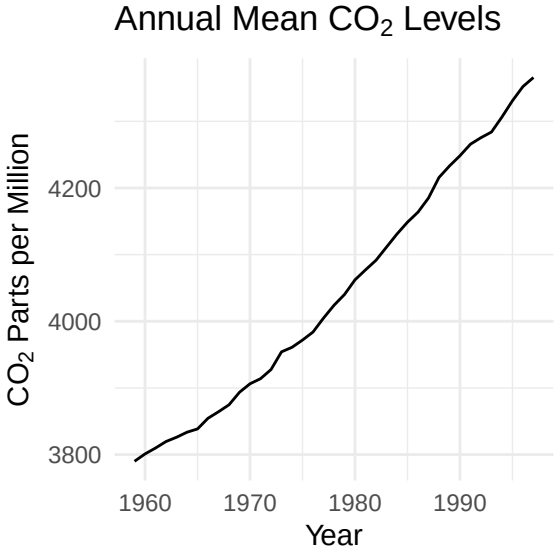
```
## Warning: Removed 38 rows containing missing values (geom_text).
```



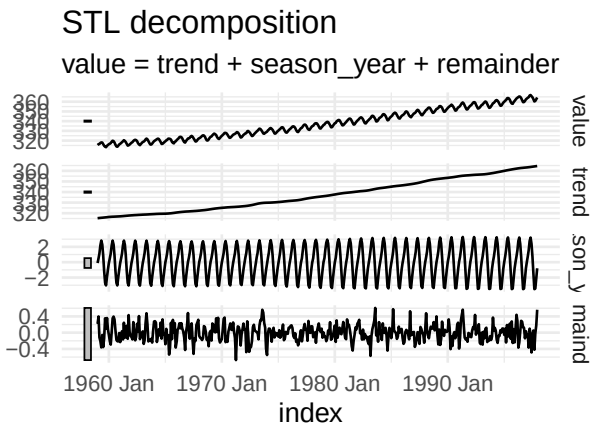
From these plots, we can see that the CO2 levels appear to increase from January through May, then decrease from June through October, hitting a low in October, and then increase again from November through end of the year. The source report does mention that plants and soil absorbing and emitting CO2 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,

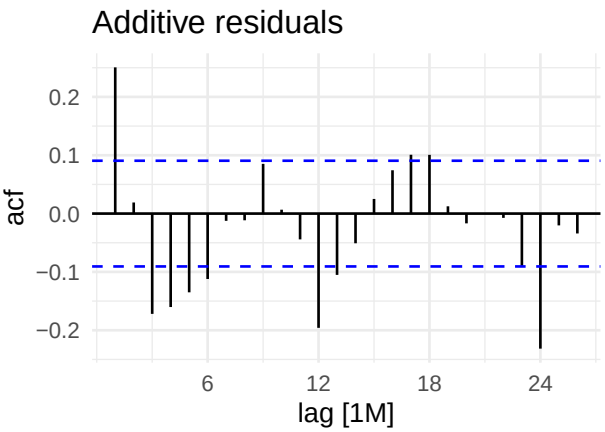
there is variability across the Northern Hemisphere in what counts as "colder" months and when plants thrive - for example, Hawaiian "winter" is temperate and plants can grow year round, so this could explain why the seasonal variability is slight across the year. We can again see the trend of CO2 levels 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.

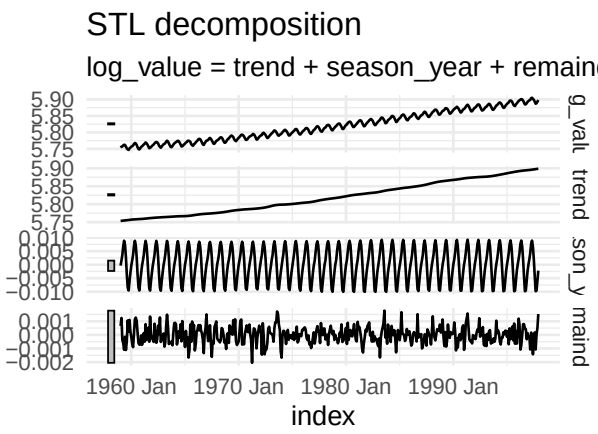
We can confirm that our time series trend is increasing by removing our seasonality components and aggregating our data by year instead of month, as seen below.

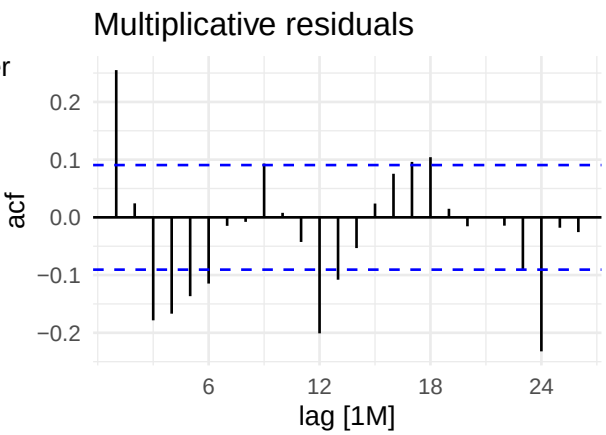


We can also use both additive and multiplicative decomposition to remove both the trend and seasonal movements from our dataset and confirm that the variance is stationary. The results from these decompositions are plotted below.









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