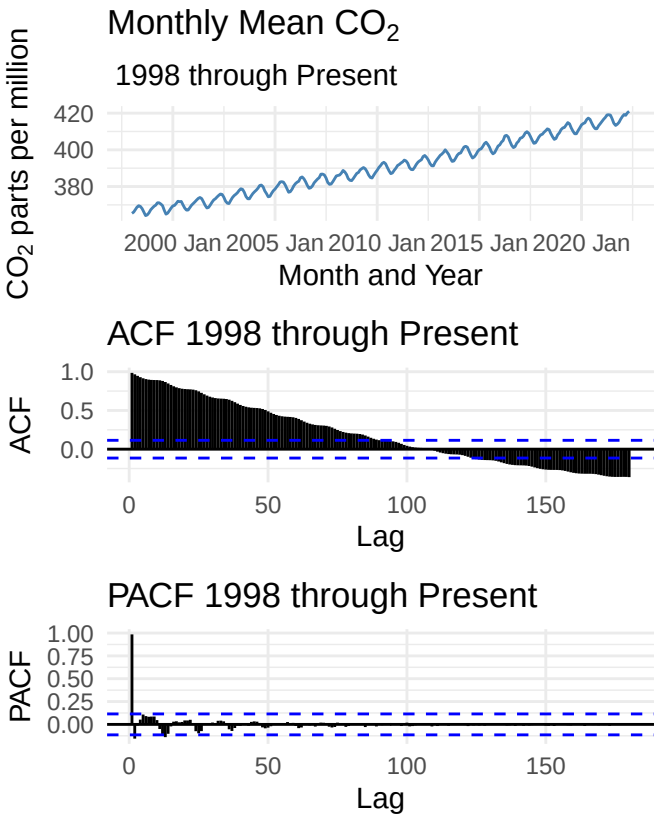
(1 point) Task 0b: Introduction

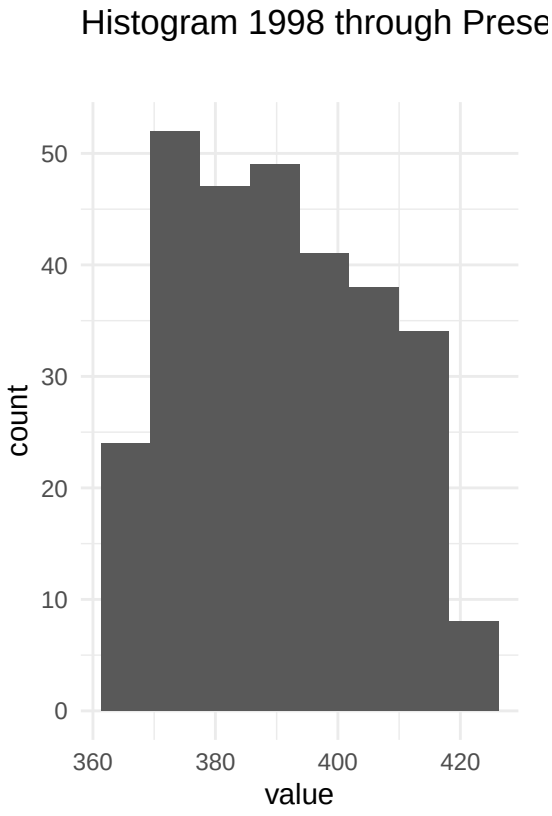
We now ask the question of whether the data collected from January 1998 to the present has different characteristics which could lead us to additional or different conclusions than those which the data set from 1958 until 1997 yielded.

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.





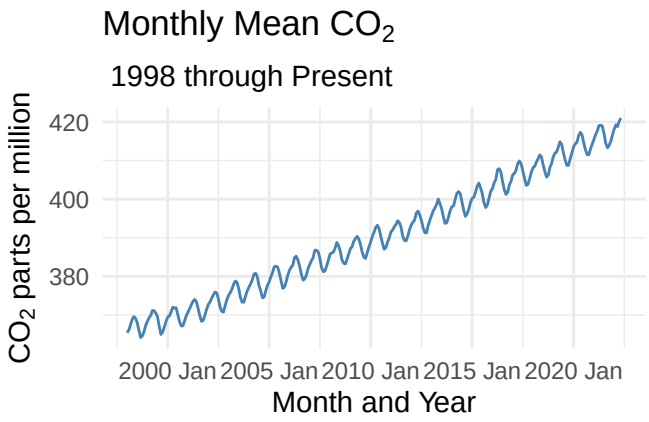
As with the "1997" data set, we see a clear increasing trend with the same seasonal "wave" pattern. Seasonal fluctuation does not appear to increase with time, suggesting, again, that an additive model explains the decomposition in this series. The time series is not stationary, as its mean increases.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

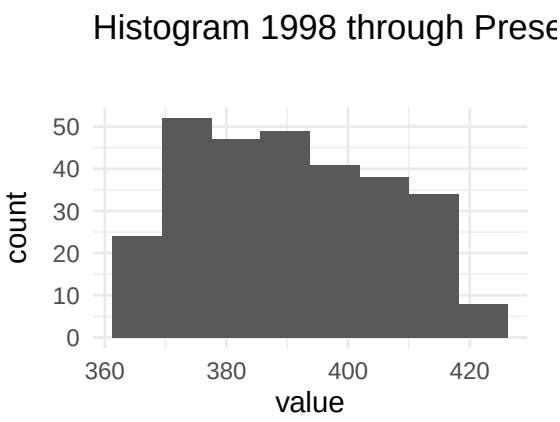
The ACF displays a slow, over-time, decline consistent with a trending time series. While the ACF plot of the "1997" plot remains above the significance line until lag 140, this ACF plot crosses the significance line at close to 100. As mentioned above, this "present" data set contains only 24 years,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

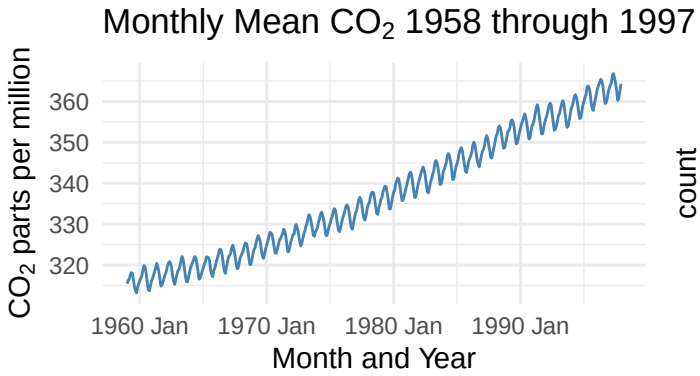
The PACF has a large positive spike at lag 1, followed by a smaller negative spike at log 2, then oscillates between positive and negative lags with progressively lower levels of significance.

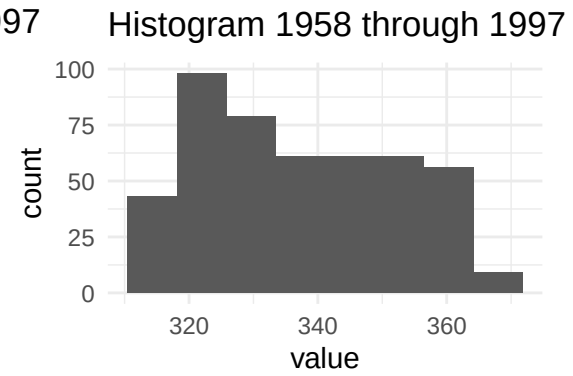
Our histogram of has no values lower than 365, or nigher than 421, and that these values are not normally distributed.

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



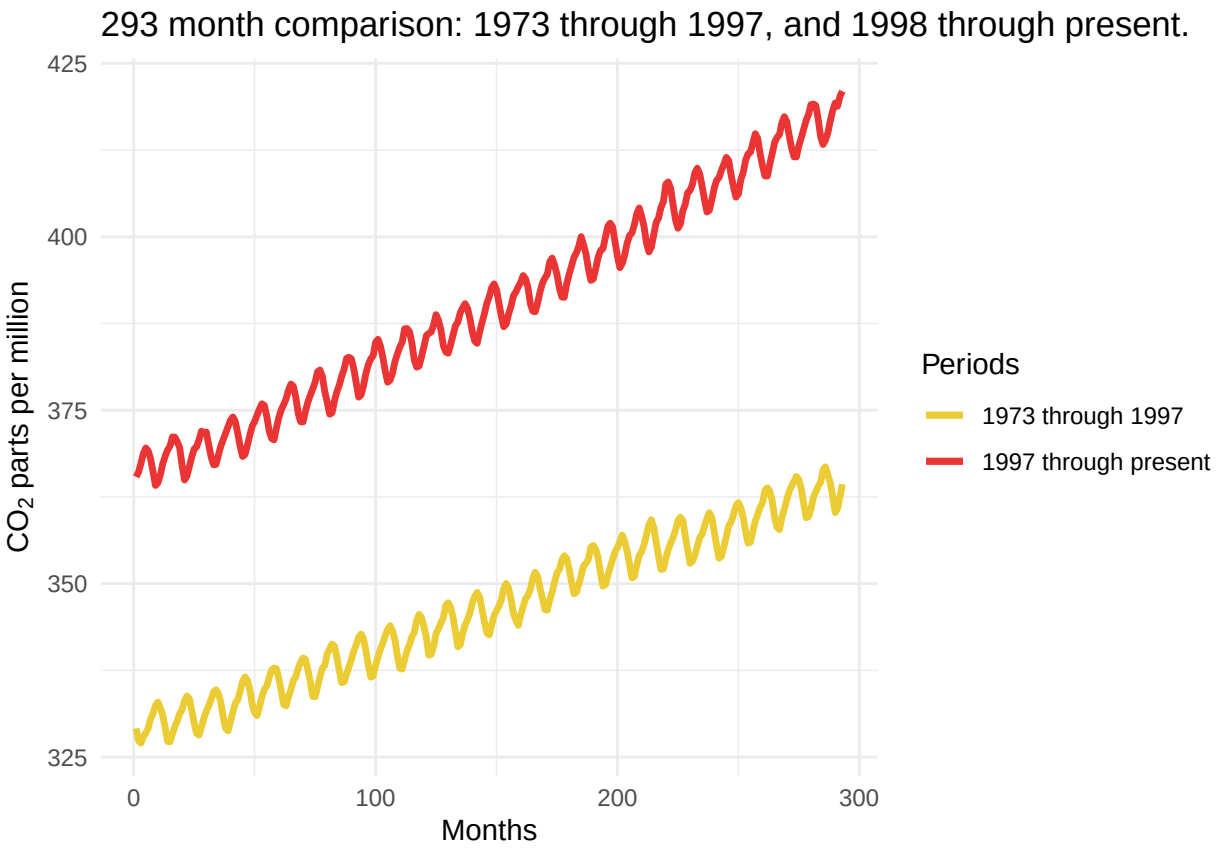






When we start looking graphing the two data sets side by side, the difference of one containing 39 years, and the other 24, begins to hamper our understanding.

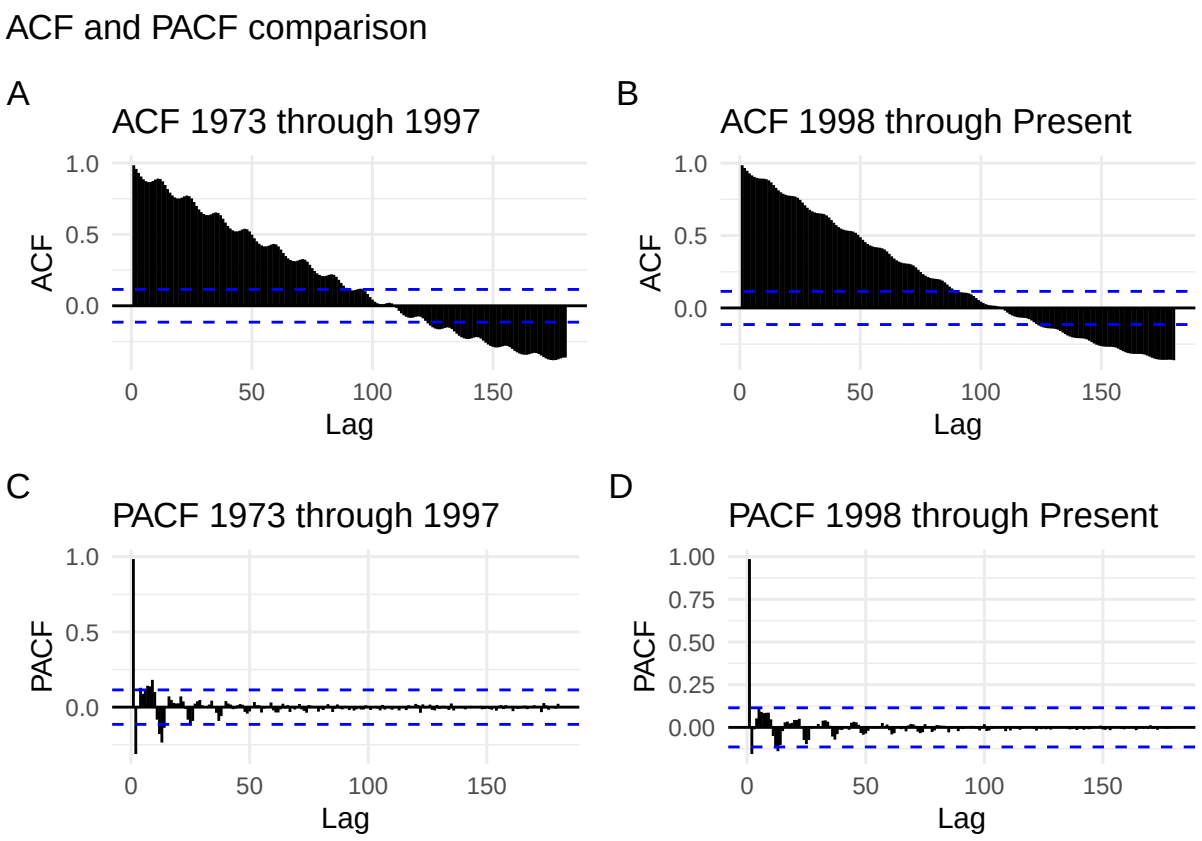
We take only the last 293 months of the "1997" data set and graph it on the same space as the "present" data set.



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

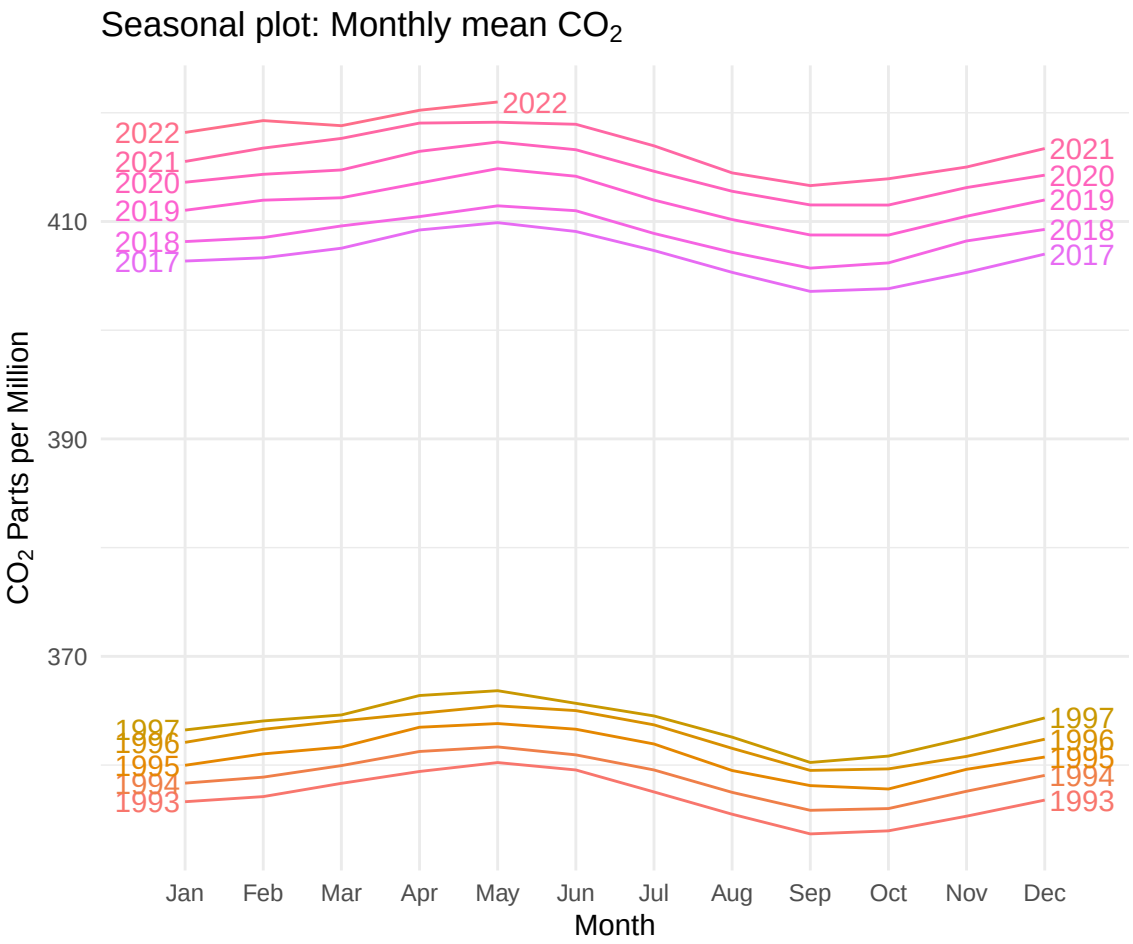
May 31, 2022 (the most recent month we have a measurement for), the atmospheric CO2 rose by 56.83 PPM. This is a difference of 17.01 PPM during the same time interval. This could indicate an acceleration in the increase in PPM.

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.

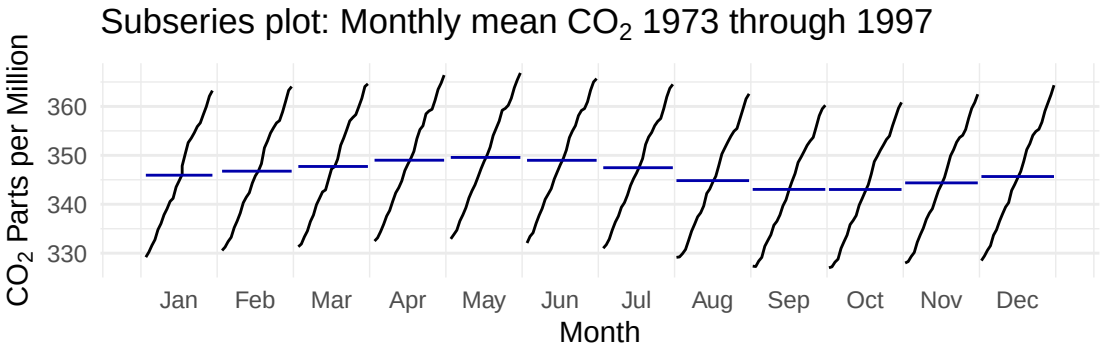
Next we examine the seasonality of the "present" comparing it to the "1997" seasonality.

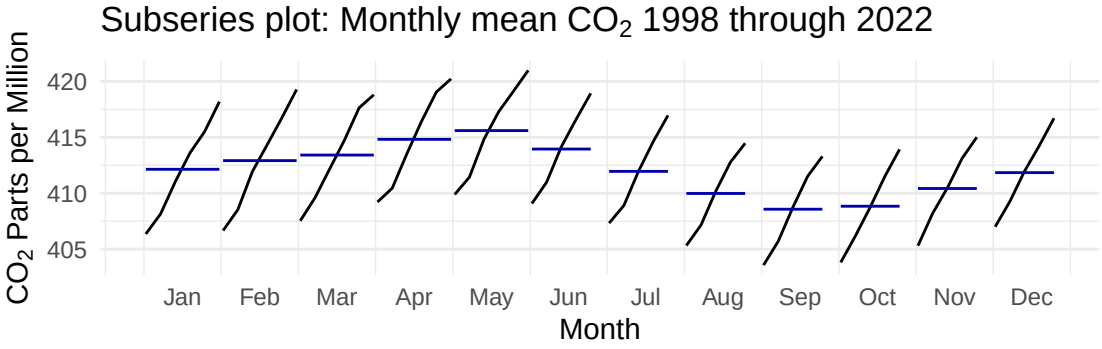


In the plot above we look at the last 5 years in the "1997" data set and the last 5 years of the "present" data set.

What we are trying to ascertain is if the seasonality has become more extreme. It is difficult to ascertain if the seasonality has become more extreme based on the graph above.

We next compare the two subseries graphs looking at each data set.

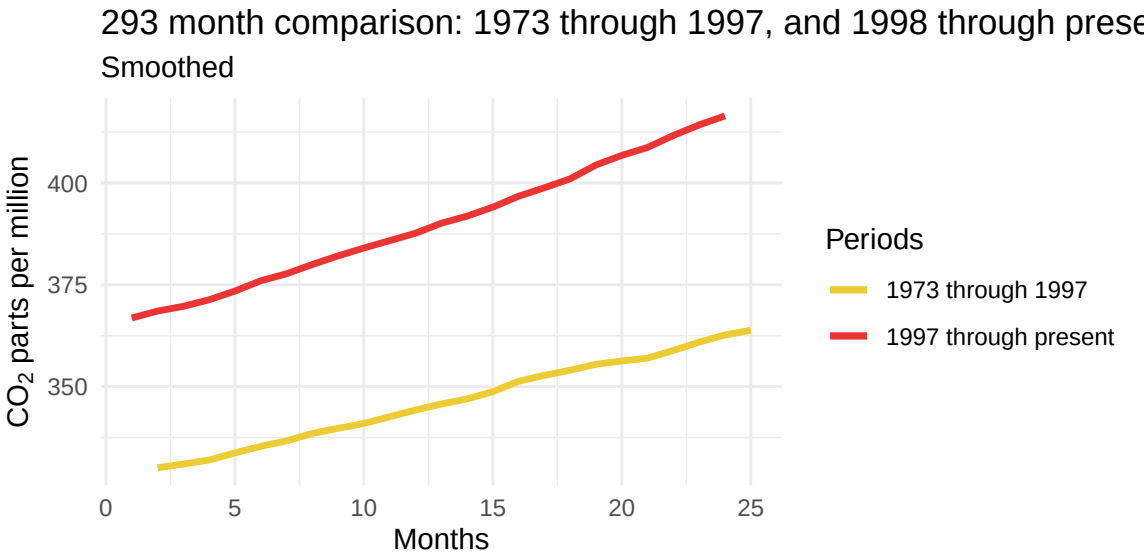




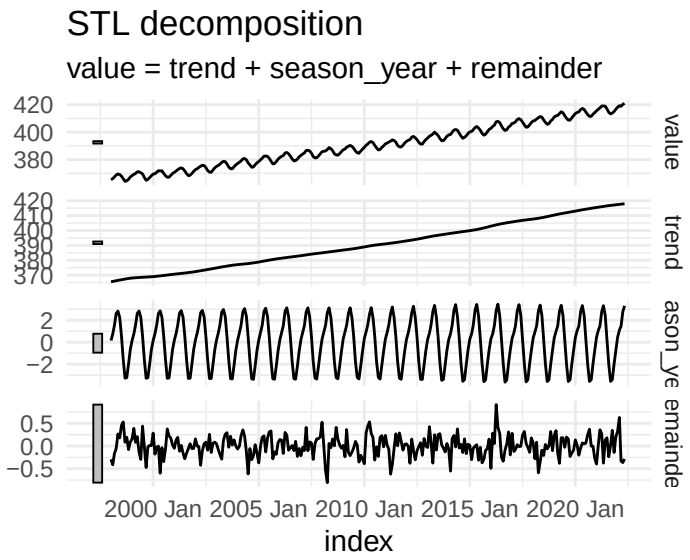
With this new plot it becomes clearer that the yearly oscillations in PPM are more extreme in the present data set than in the one ending in 1997.

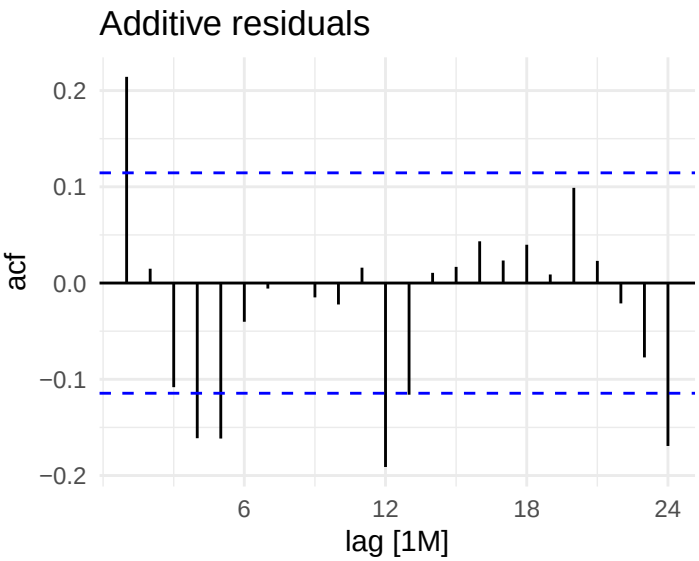
That is to say the seasonality effect shows a noticeable increase in the magnitude of the fluctuations in the second data set.

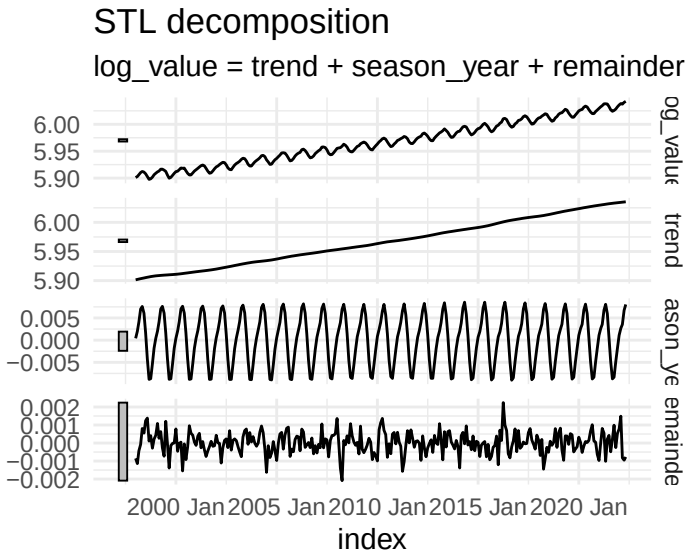
We now remove the seasonal fluctuations in both data sets and graph the increase in atmospheric CO2 PPM.

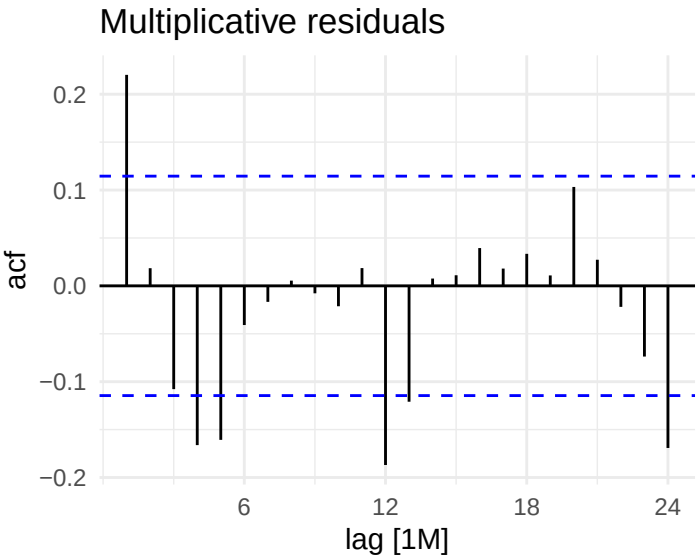


We now perform additive and multiplicative decomposition to remove the trend and seasonal movements from our data set to confirm that the









These graphs confirm that the time series is trending upwards. In the additive decomposition plots (top left), we see that seasonal fluxuations increase over time. This confirms our side-by-side graph above showing the increase in seasonal fluctuations between the two time series.

The multiplicative "season_year" decomposition graph is stable, and not visibly increasing. This means like the "1997" time series, the "present" time series may be multiplicative.

Both residuals plots, again, show our time series as stationary, but not white noise. This means correlation in the data still exists.

```
##
##  Augmented Dickey-Fuller Test
##
## data:  co2_present_ts
## Dickey-Fuller = -8.8282, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary
```

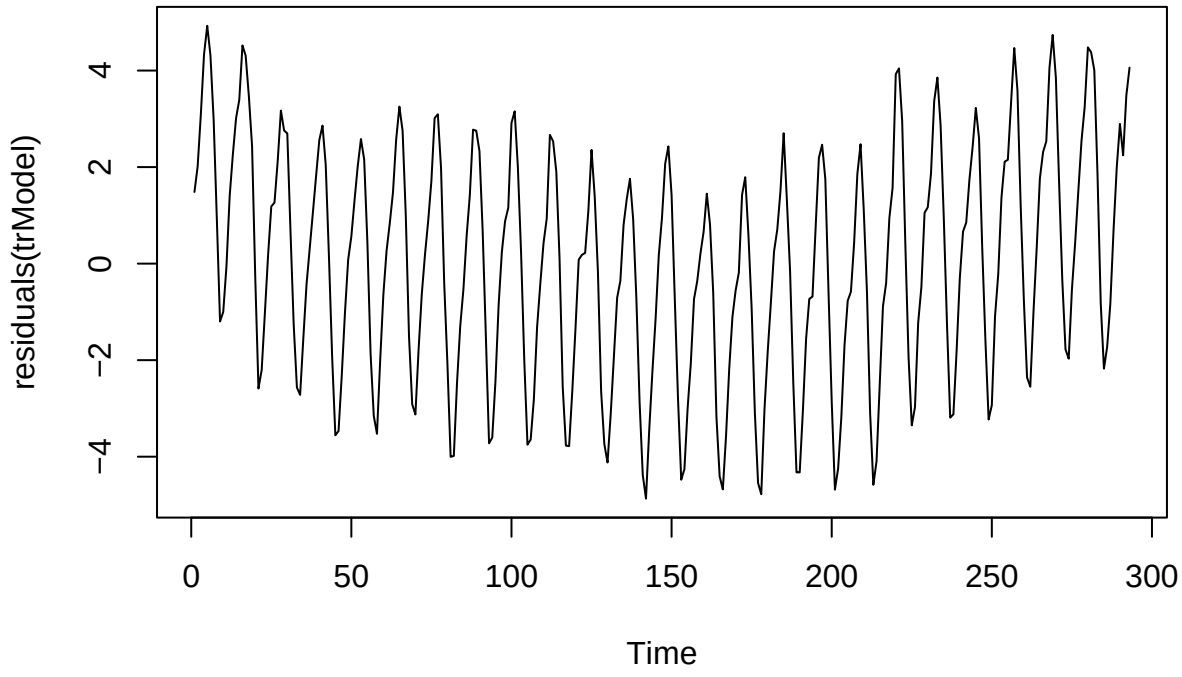
```
##
##  Phillips-Perron Unit Root Test
##
## data:  co2_present_ts
## Dickey-Fuller Z(alpha) = -74.123, Truncation lag parameter = 5, p-value
## = 0.01
## alternative hypothesis: stationary
```

The tests above indicate that we can reject the null hypothesis that the time series is non-stationary.

The Unit Root tests above indicate that this time series is difference stationary.

This contradicts our visual EDA, which shows a changing mean, given the strong upward trend.

We should note that the tests above are testing whether the time series are difference stationary. We apply a linear model and then graph its residuals.



Graphing the residuals appears to show a mean reverting process. That is to say the residuals are stationary and not trending.

Finally, we run an ADF and PP test on the residuals, and we again, see that the null hypothesis of non-trend-stationary can be ignored.

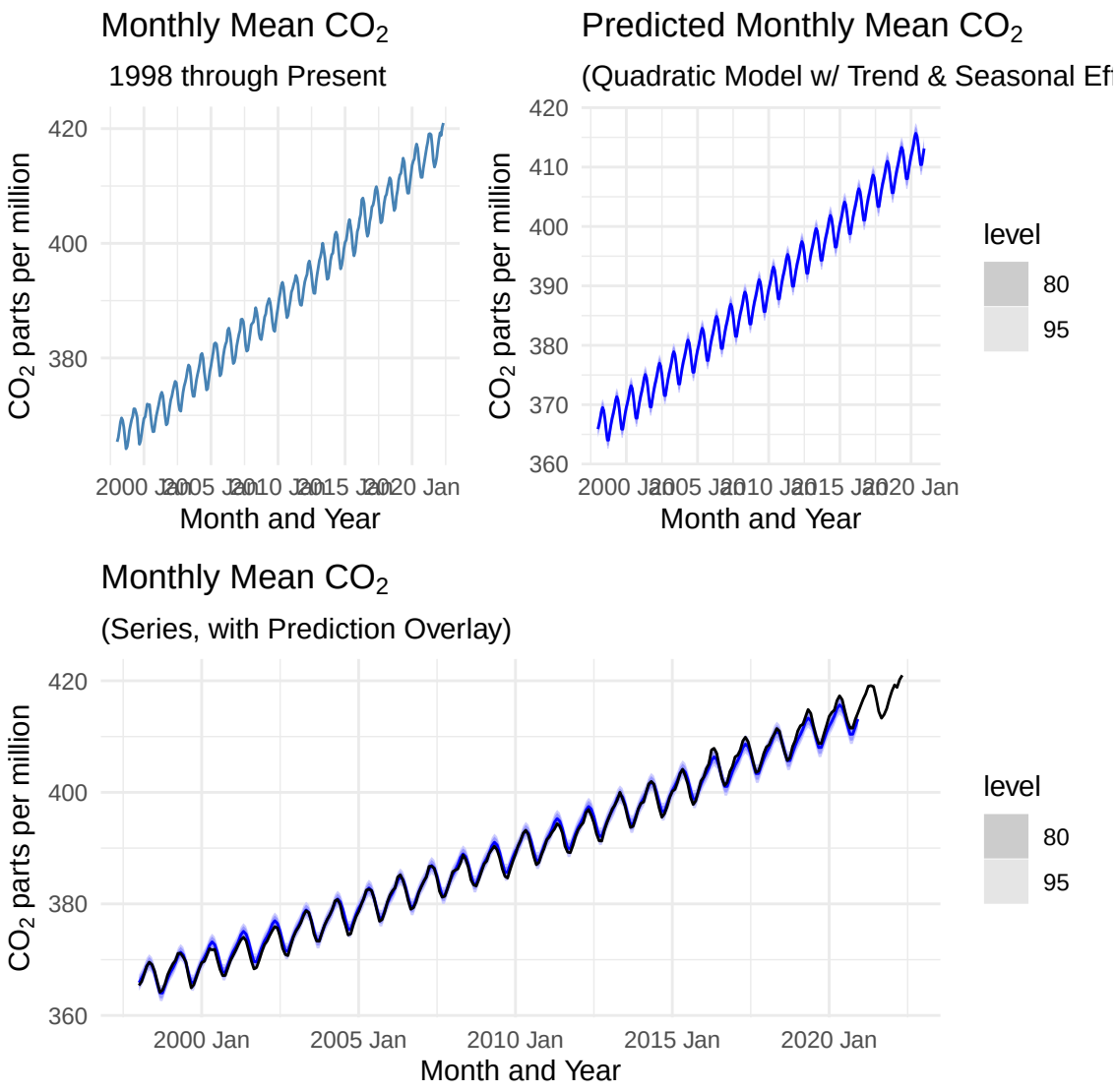
```
## Warning in adf.test(trModel$residuals, alternative = "stationary", k = 5): p-
## value smaller than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  trModel$residuals
## Dickey-Fuller = -8.8282, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary

## Warning in pp.test(trModel$residuals, alternative = "stationary"): p-value
## smaller than printed p-value

##
##  Phillips-Perron Unit Root Test
##
## data:  trModel$residuals
## Dickey-Fuller Z(alpha) = -74.123, Truncation lag parameter = 5, p-value
## = 0.01
## alternative hypothesis: stationary
```

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



In the top-left, we plot the 1998-present time series, next to it, we plot the prediction yielded by the best-fitting we developed in our first section. Beneath the two plots we overlay the original time series with the prediction. Towards the middle of time time series, the fit is nearly perfect. At both the beginning and end of the time series, there are barely visible differences the prediction and the "real" underlying time series.

We predict 23 years from Dec 31, 1997, which means the predict values are through December 31 2021. Our "present" time series contains data "through" May 2022, which means we have an addition 5 months of data, indicated by the solid black line extending upwards out of the prediction.

The model is very accurate.

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