Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Lab 2

Investigating the Keeling Curve and forecasting CO2 levels in Earth's atmosphere

Contents

0.1 SECTION 2

0.2 Report from the Point of View of the Present

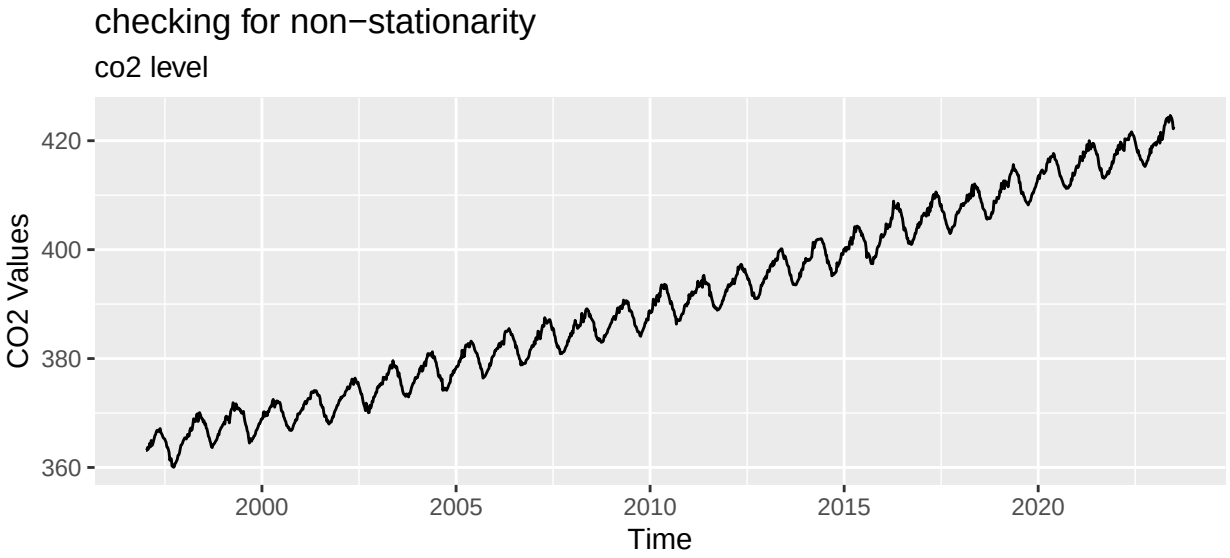
One of the very interesting features of Keeling and colleagues' research is that they were able to evaluate, and re-evaluate the data as new series of measurements were released. This permitted the evaluation of previous models' performance and a much more difficult question: If their models' predictions were "off" was this the result of a failure of the model, or a change in the system?

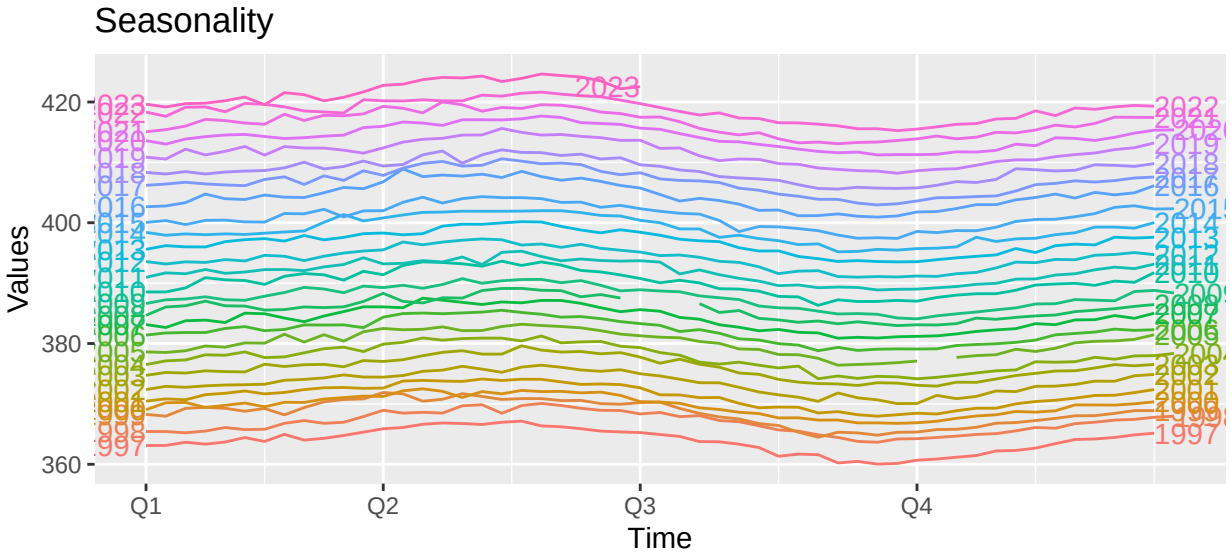
0.3 (1 point) Task 0b: Introduction

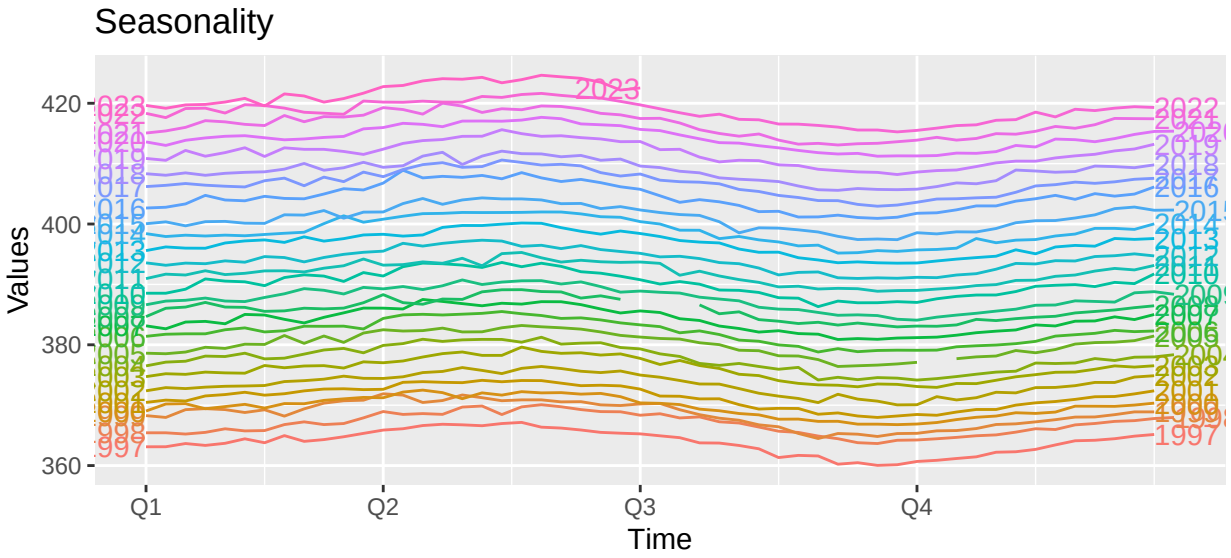
In this introduction, you can assume that your reader will have **just** read your 1997 report. In this introduction, **very** briefly pose the question that you are evaluating, and describe what (if anything) has changed in the data generating process between 1997 and the present.

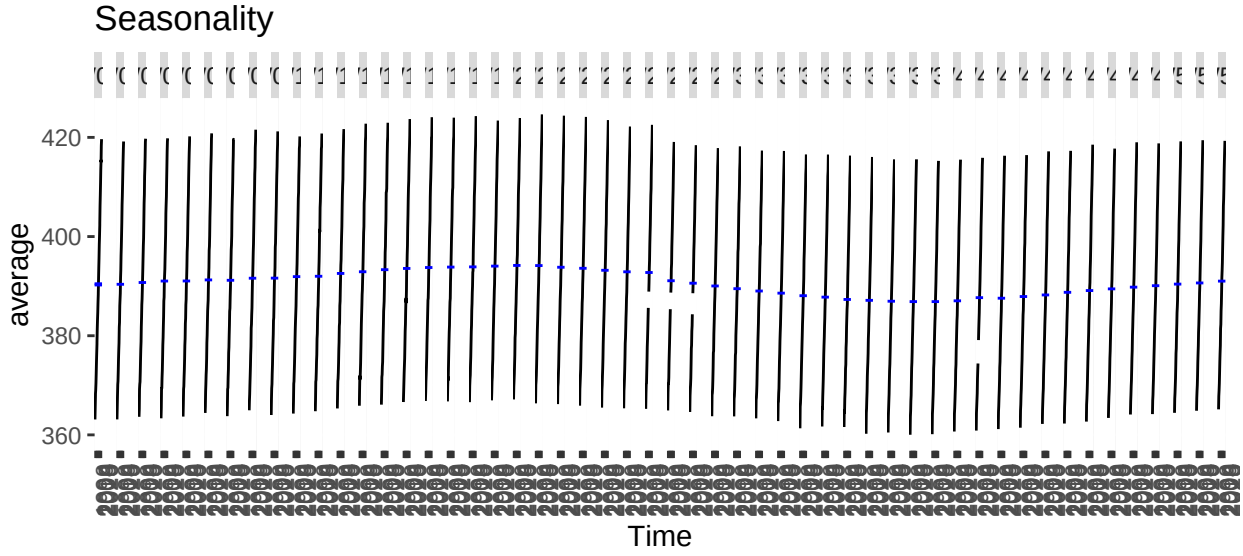
0.3.1 (3 points) Task 1b: Create a modern data pipeline for Mona Loa CO2 data. [Weekly Data] [EDA - ACF, PACF, Seasonality, decomposition]

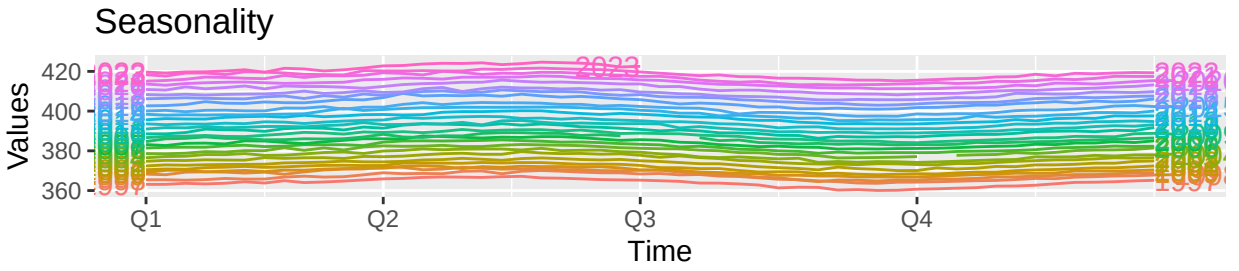
The most current data is provided by the United States' National Oceanic and Atmospheric Administration, on a data page [here]. Gather the most recent weekly data from this page. (A group that is interested in even more data management might choose to work with the hourly data.) Create a data pipeline 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.

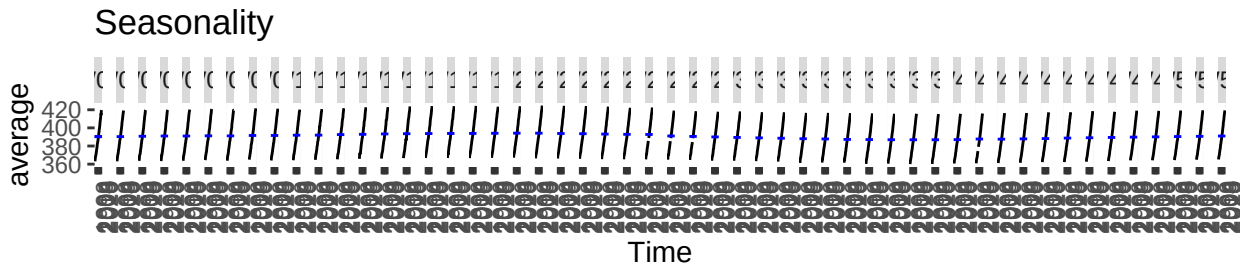

checking for non−stationarity
co2 level
CO2 Values
420
400
380
360
2000
2005
2010
2015
2020
Time


Seasonality
Values
420
400
380
360
Q1
Q2
Q3
Q4
Time
2023
2022
2019
2018
2016
1997

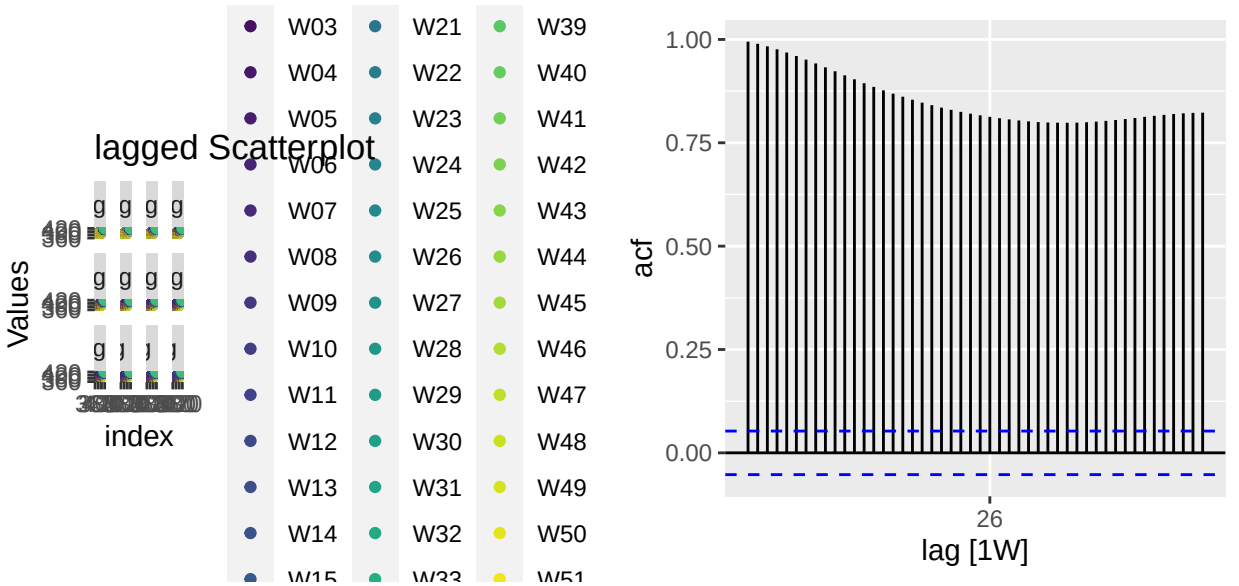

Seasonality
Values
420
400
380
360
2023
2022
1997
Q1
Q2
Q3
Q4
Time


Seasonality
average
420
400
380
360
Time

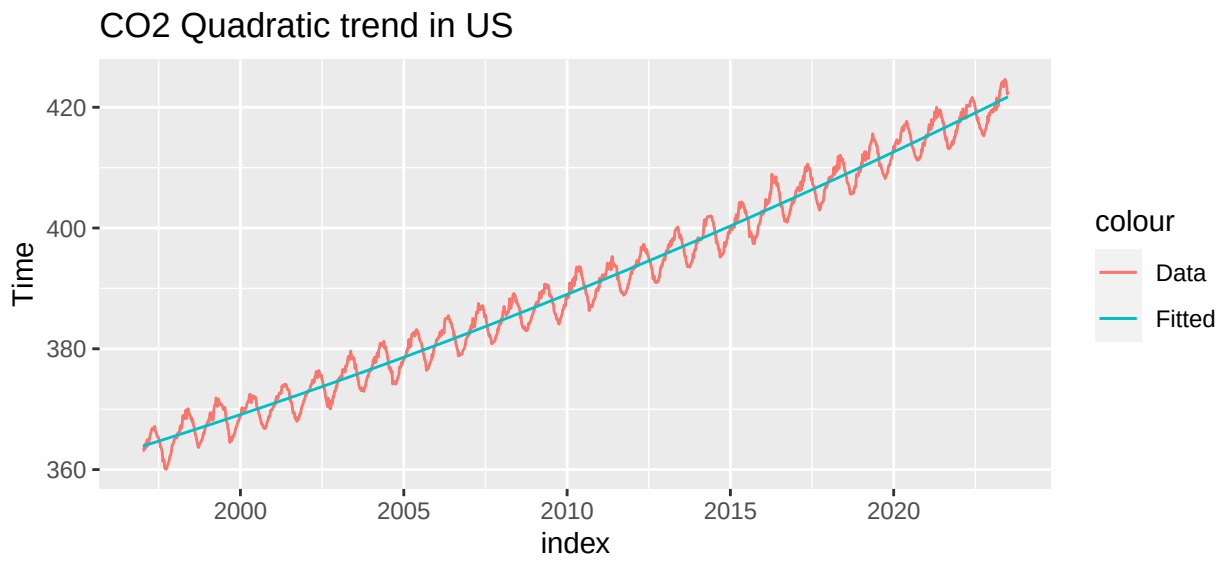





```
## Warning: Removed 4 rows containing missing values (gg_lag).
```



(1 point) Task 2b: Compare linear model forecasts against realized CO2 [Weekly Data] 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.)



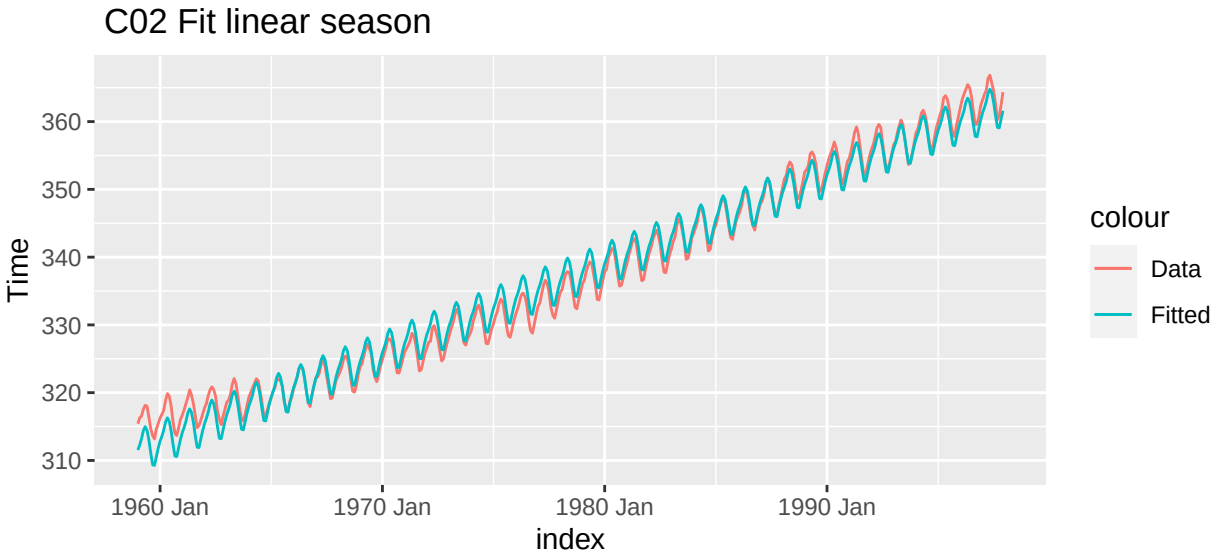
0.3.2 (1 point) Task 3b: Compare ARIMA models forecasts against realized CO2

0.3.3 4B.(3 points) Task 4b: Evaluate the performance of 1997 linear and ARIMA models

**Ques:*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.)

Observations: In the previous section, we conducted exploratory data analysis to visually assess the forecast of both the linear model and ARIMA model for atmospheric C02 levels. These models captured the historic trends and patterns effectively up to a certain point 1997. However as we moved from 1997, the forecasted lines started to deviate, making it challenging to determine which model better fits the data. To quantitatively evaluate the accuracy of the models, we conducted a formal evaluation using the Root Mean Squared Error (RMSE) test. Both the ARIMA and linear models were developed using data from 1974 to 1993. We used these models to forecast the atmospheric CO2 levels until 2023. Since models were built on a monthly dataset, the predictions we obtained were also on a monthly basis for the period from 1993 to 2023.

To ensure a fair comparison of these predictions we aggregated the current dataset to a monthly level. The model performance of ARIMA model , after comparing its predictions with the current dataset, resulted in an RMSE of 51.66 On the other hand, the linear model which considered both the seasonal and trend components, had an RMSE of 14.037. It looks like that data has certain trend and pattern which are better captured by linear regression.




```
## [1] "Root Mean Squared Error (RMSE): 51.6689948143351"
```

```
## [1] "Root Mean Squared Error (RMSE): 14.0391328048566"
```

0.3.4 (4 points) Task 5b: Train best models on present data

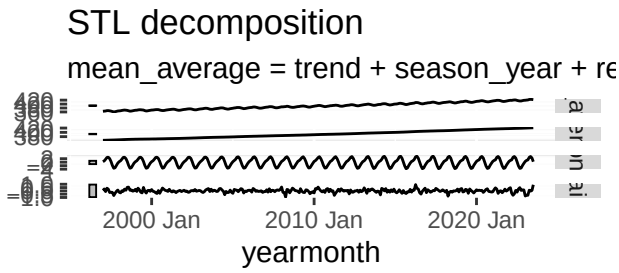
Ques: Seasonally adjust the weekly NOAA data, and split both seasonally-adjusted (SA) and non-seasonally-adjusted (NSA) series into training and test sets, using the last two years of observations as the test sets. For both SA and NSA series, fit ARIMA models using all appropriate steps. Measure and discuss how your models perform in-sample and (psuedo-) out-of-sample, comparing candidate models and explaining your choice. In addition, fit a polynomial time-trend model to the seasonally-adjusted series and compare its performance to that of your ARIMA model.

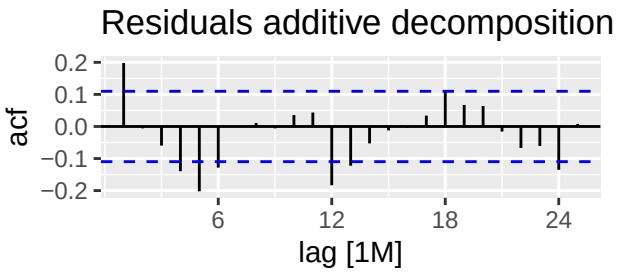
Analysis: Our past evaluation consisted of comparing two old models, ARIMA and linear, on a recent dataset. However if we want to evaluate the long-term trend in the data then we utilized the seasonal adjustment using the sophisticated STL method. STL effectively decomposes the time series into trend, seasonal, and residual components, enabling a more efficient decomposition of time series and analyze each parts.

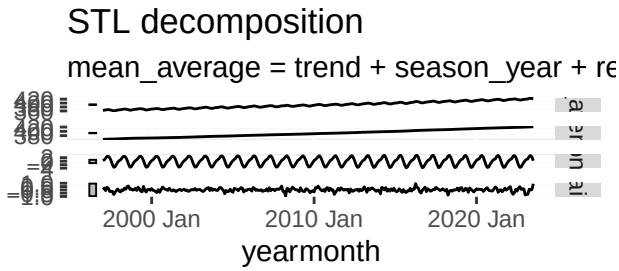
Once we obtained the seasonally adjusted data after decomposition, we proceeded to train the models on data from 1997 to 2021 and evaluated their performance on the 2022-2023 dataset. The results from the STL-based seasonal adjustment showed that the ARIMA model outperformed the linear model. The success of the ARIMA model can be attributed to its effective capture of residual autocorrelation present in the trend of the seasonally adjusted data obtained from STL. The ARIMA model achieved a significantly lower Root Mean Square Error (RMSE) of 0.14029, demonstrating its superior accuracy compared to the linear model, which obtained a higher RMSE of 1.572.

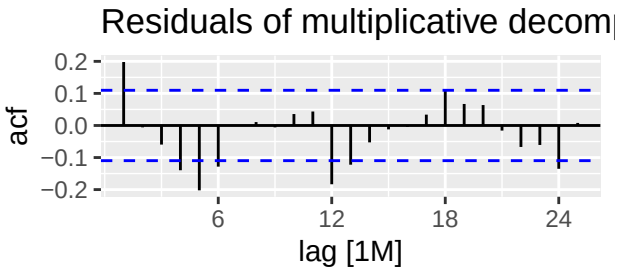
Additionally, we performed a similar analysis on the raw, non-seasonally adjusted weekly dataset, retaining its inherent seasonal pattern. Once again, the ARIMA model outperformed the linear model, as expected. The RMSE was 0.14 which is lower then 1.78 got in linear model evaluation. The ARIMA model's ability to better capture the seasonal pattern is due to its autoregressive component and differencing, which effectively removed the trend component, resulting in lower root

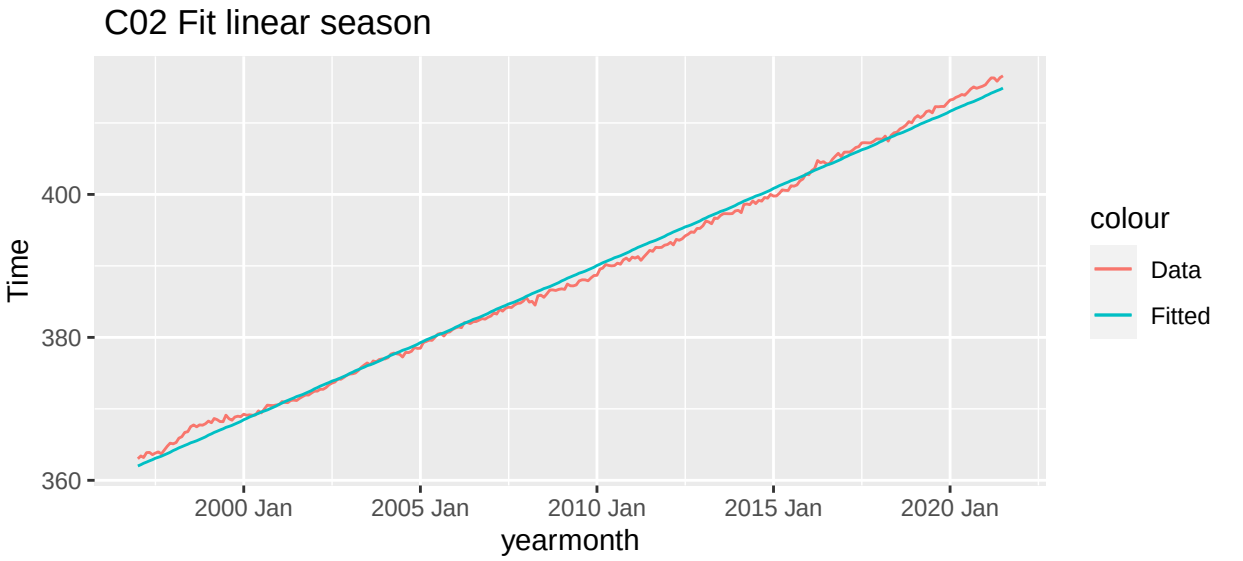
mean square errors. To summarize, our investigation revealed that applying STL-based seasonal adjustment and subsequently utilizing the ARIMA model proved to be the most effective approach for understanding the long-term trend and accurately capturing seasonal patterns in the data.





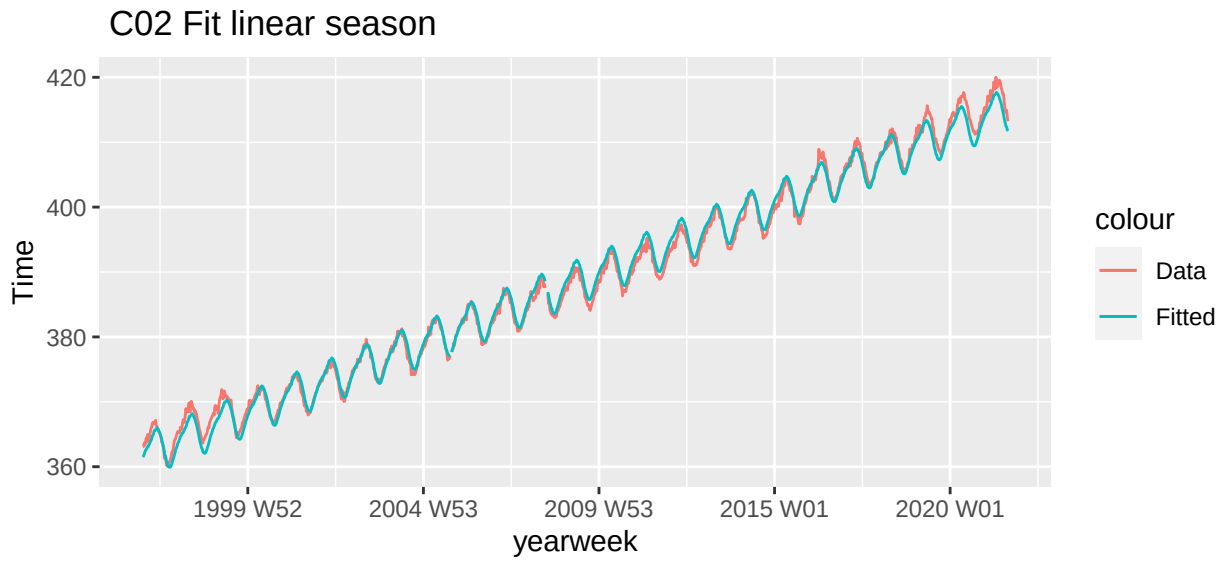







```
## [1] "Root Mean Squared Error (RMSE): 1.57279267689926"
```

```
## [1] "Root Mean Squared Error (RMSE): 0.140291213125284"
```




```
## [1] "Root Mean Squared Error (RMSE): 1.784659353824"
```

0.3.5 (3 points) Task Part 6b: How bad could it get?

With the non-seasonally adjusted data series, 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 2122. How confident are you that these will be accurate predictions?

Generate a prediction for atmospheric CO2 levels in the year 2100. How confident are you that these will be accurate predictions?