Forecasting State Revenue

Exploring Texas Tax Trends

About the data

The Quarterly Summary of State and Local Government Tax Revenue provides quarterly estimates of state and local government tax revenue at a national level, as well as detailed tax revenue data for individual states. The information contained in this survey is the most current information available on a nationwide basis for government tax collections.

ECONOMIC RESEARCH

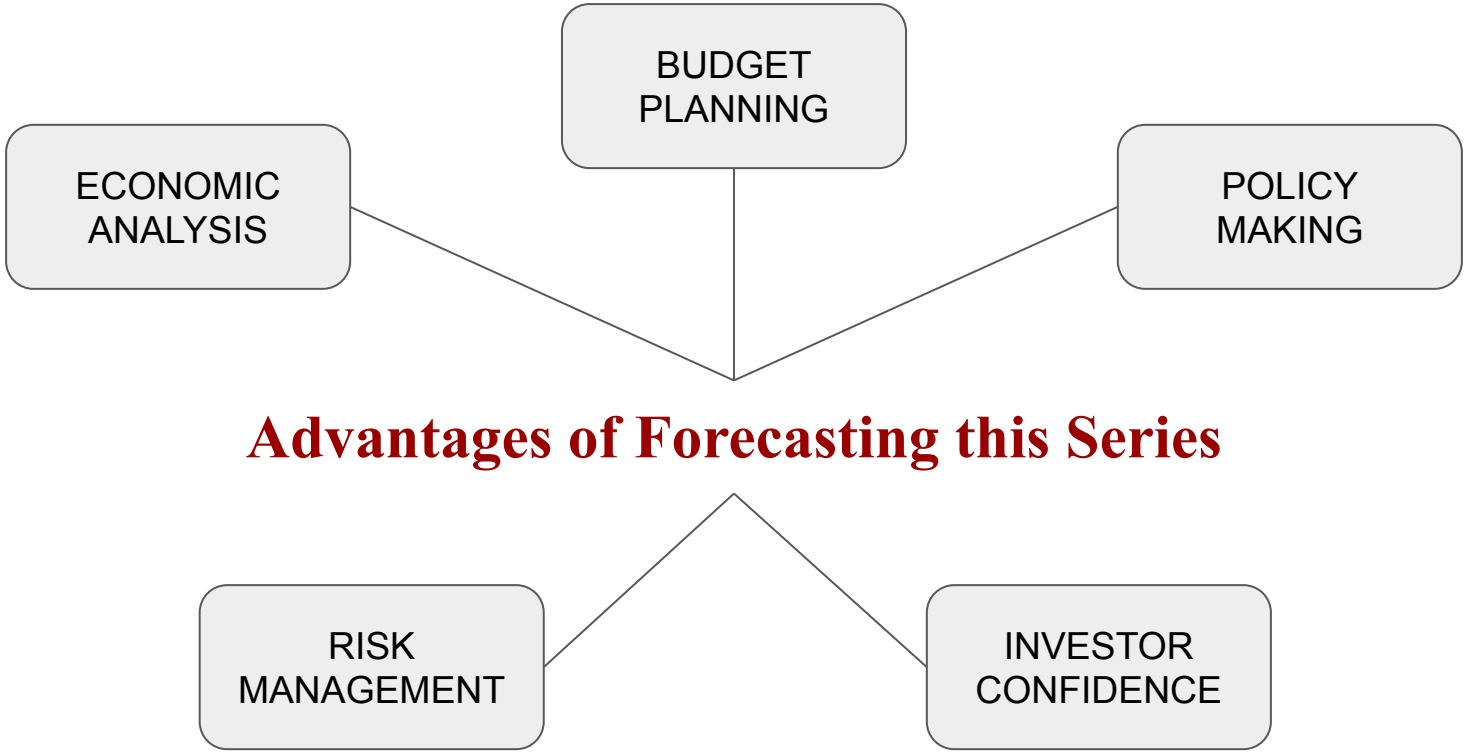
FRED®

ECONOMIC DATA | ST. LOUIS FED

FACTS

- Non Seasonally Adjusted Data
- No. of Rows: 120 — Quarterly Data 1994 - 2023
- No. of Columns: 2 — Date & Total Taxes

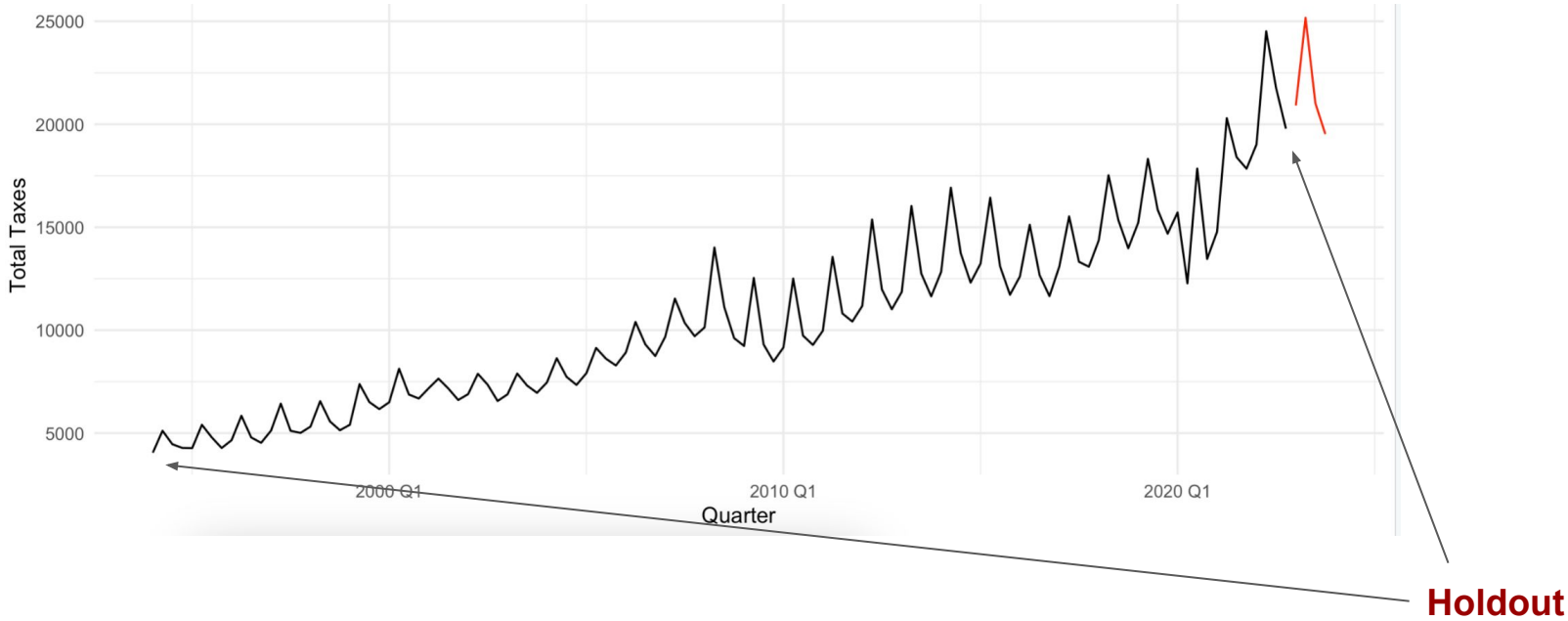
(https://fred.stlouisfed.org/series/QTAXTOTALQTAXCAT3TXNO)



Our aim is to achieve valuable insights into the financial health of governments and the broader economy, guiding decision-making at both the public and private sector levels

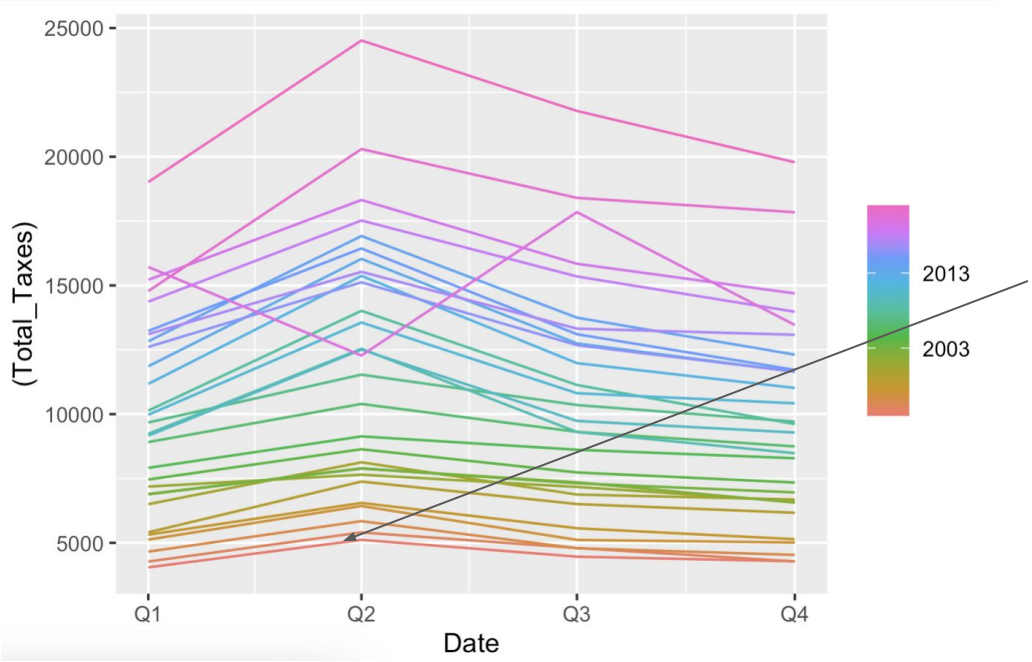
Holdout Period

Our dataset spans from Q1 1994 to Q4 2023. To ensure the robustness of our analysis, we've implemented a holdout period excluding the last four observations, which comprise all four quarters of 2023.
Thus, **our holdout period includes data only up to Q4 2022**.



Transformation

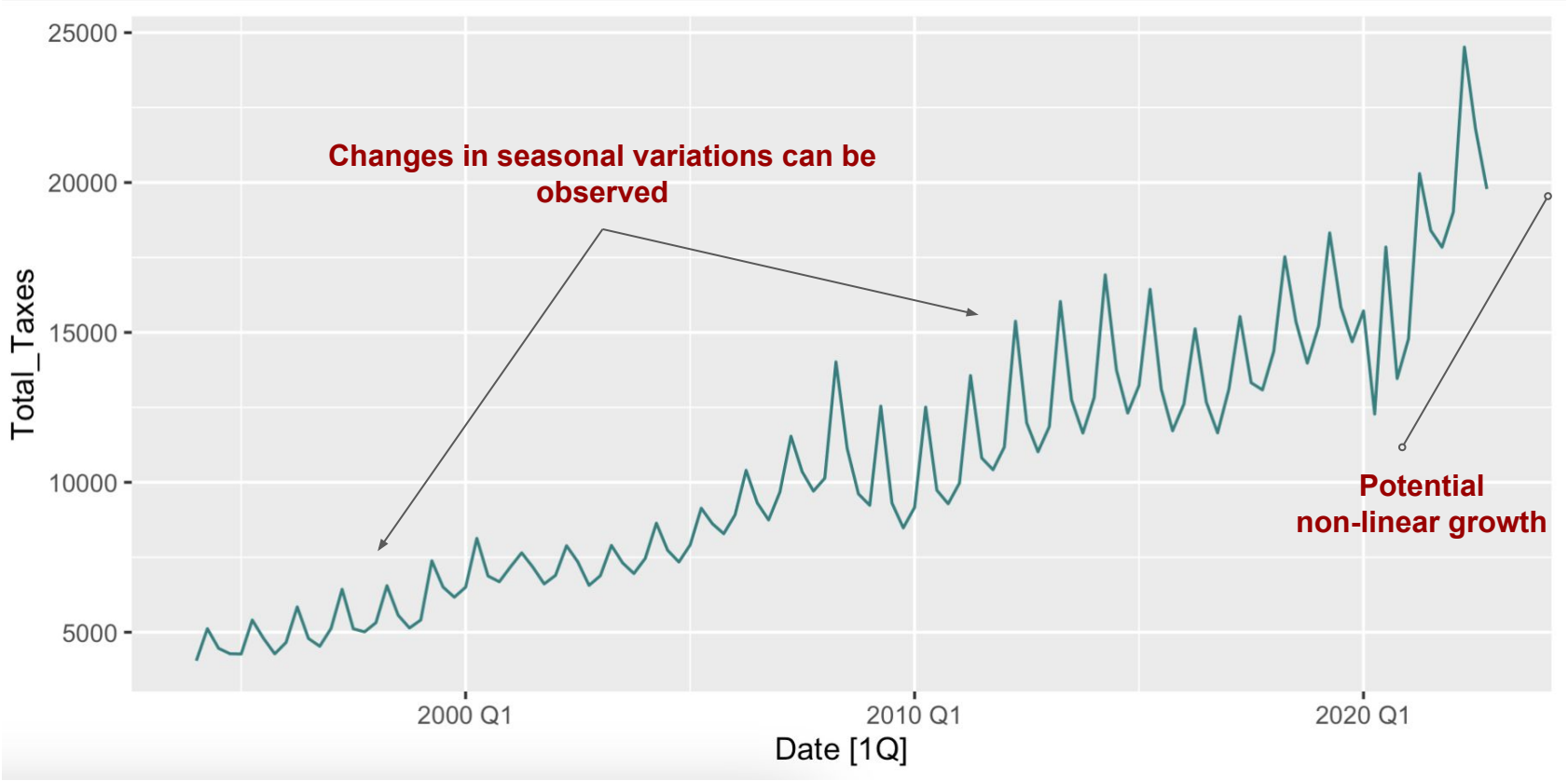
gg_season plot of our holdout series



It's evident from the data that the peak in total tax collection consistently occurs at the beginning of Q2 and tapers off later, aligning with expectations since many taxes are due to be submitted to the government by April 15.

This pattern remains consistent across all years except for 2020, suggesting a **seasonality pattern**.

Autoplot for our series

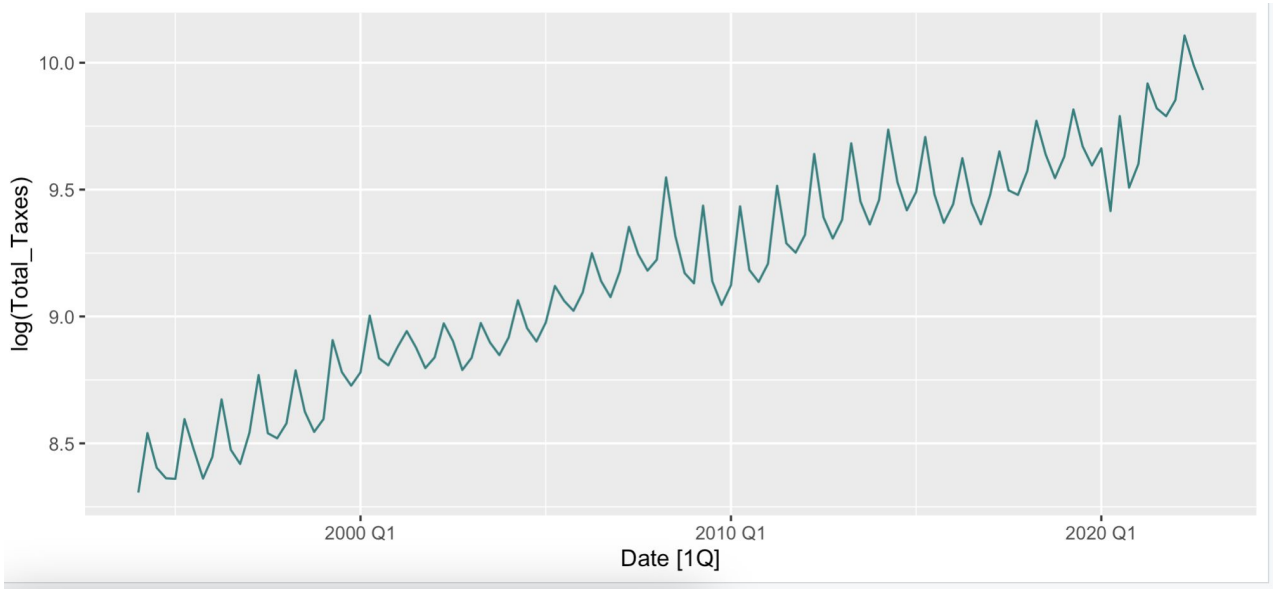


Transformation

Observing our series we believe a **Box-Cox transformation is necessary**.

We also checked value of **λ** using the guerrero method in R which gave us a value of **-0.213** which is **≈ 0** and vary from 1, hence we chose to continue with **log transformation**.

Transformed series



Model 1 : SARIMA

Does our series need seasonal and non-seasonal differencing?

Let's check using **nsdiffs()** and **ndiffs()** function

```
TaxHold%>%features(transform,unitroot_nsdiffs)
A tibble: 1 x 1
nsdiffs
  <int>
      1
#is there a non-seasonal unit root
TaxHold%>%features(transform,unitroot_ndiffs)
A tibble: 1 x 1
ndiffs
  <int>
      1
```

The procedures indicate that our data might possess both a seasonal and non-seasonal **unit root**. Let's do a unit root test on this to check if our series is trend stationary

KPSS test

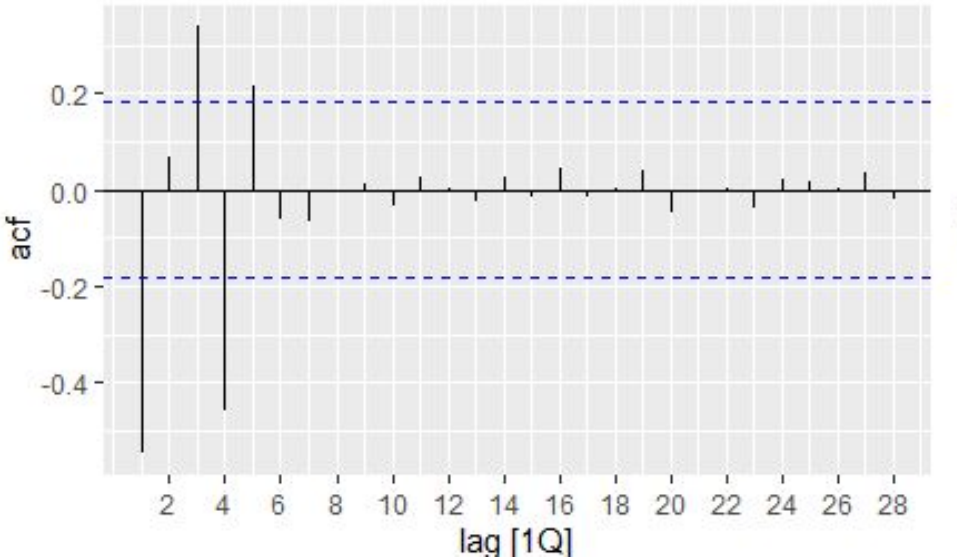
The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test is a type of **unit root test** that assesses the **null hypothesis that a univariate time series is trend stationary.**

```
print(seasonal_diff)
A tibble: 1 × 2
kpss_stat kpss_pvalue
    <dbl>       <dbl>
     2.35        0.01
print(non_seasonal_diff)
A tibble: 1 × 2
kpss_stat kpss_pvalue
    <dbl>       <dbl>
     2.35        0.01
```

Here, for both of our KPSS tests (seasonal and non-seasonal), we obtain a p-value of 0.01, which is less than 0.05 (our significance level). Therefore, we will reject the null hypothesis that the series is trend stationary.

We will continue with **d=1** and **D=1**

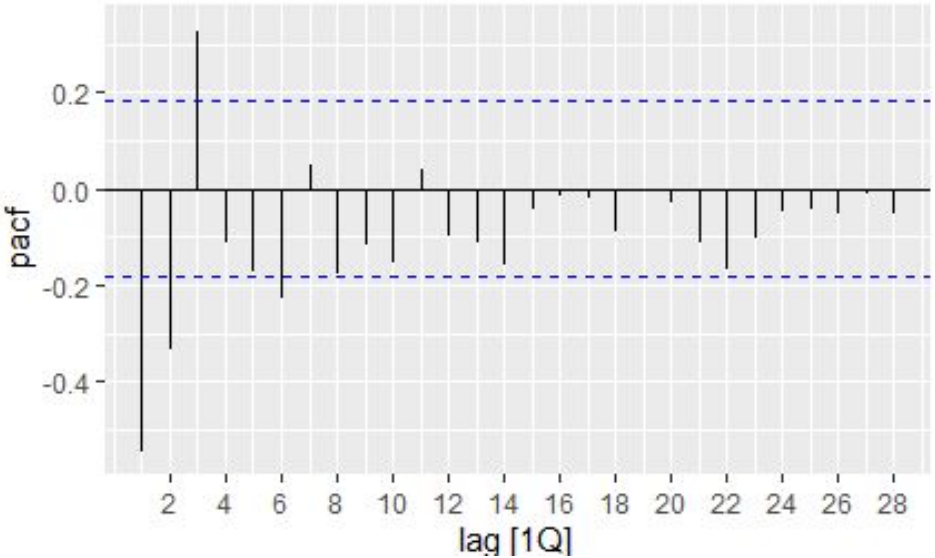
ACF and PACF charts



In our ACF chart, we can see a significant non-seasonal spike at the start of our plot, this might suggest lack of significant autocorrelation in the data beyond the first lag. We can consider a **AR(1)** model, but we will also check our pacf chart.

We can consider **one seasonal spike at lag 4**, which suggests that there might be a seasonal autoregressive term present in the data. This pattern is consistent with a seasonal autoregressive model of order **(SAR(1))**

ACF and PACF charts



In our PACF chart. We can see four non-seasonal spikes and we might want to consider setting a higher order AR model to start with.

There are no significant seasonal lags in the partial autocorrelation function (PACF), it suggests that there might not be a seasonal autoregressive or moving average component present in the data.

SARIMA models

We tried **several models** with our initial guesses for p,q & P,Q and here are three of those models which we think did a great job fitting the data

```
sarima1 = TaxHold%>%model(ARIMA(log(Total_Taxes)~0+pdq(1,1,2)+
                                PDQ(0,1,1)))%>%report()

sarima2 = TaxHold%>%model(ARIMA(log(Total_Taxes)~0+pdq(2,1,2)+
                                PDQ(0,1,1)))%>%report()

sarima3 = TaxHold%>%model(ARIMA(log(Total_Taxes)~0+pdq(3,1,2)+
                                PDQ(0,1,1)))%>%report()
```

Model	AIC	BIC
sarima1	-253.24	-239.69
sarima2	-255.06	-238.81
sarima3	-255.12	-236.16

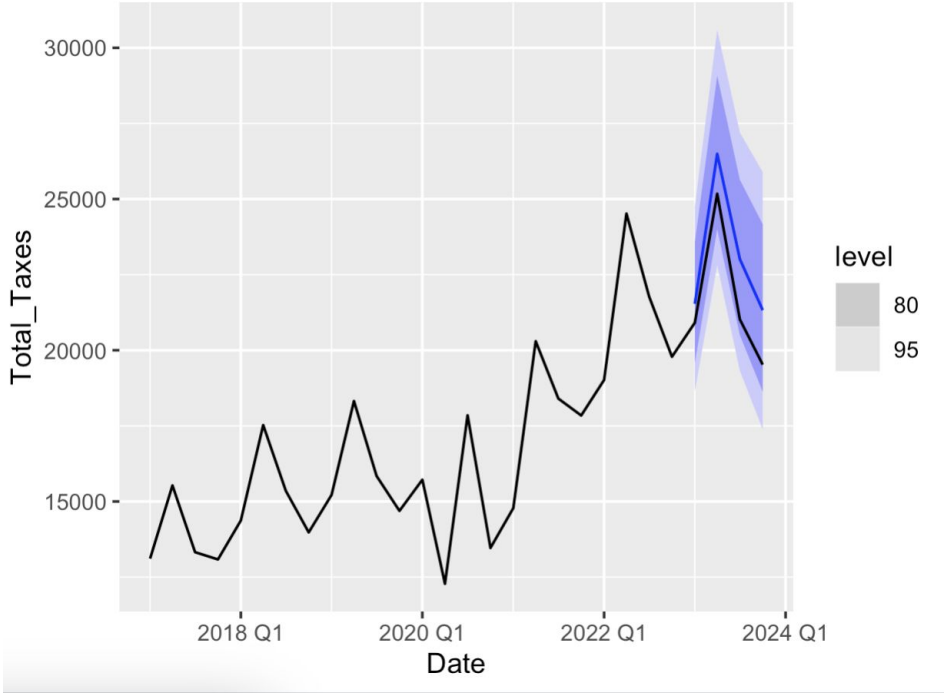
Selected Model: **sarima1(1,1,2)x(0,1,1)**

with the lowest BIC value to prefer a simpler model

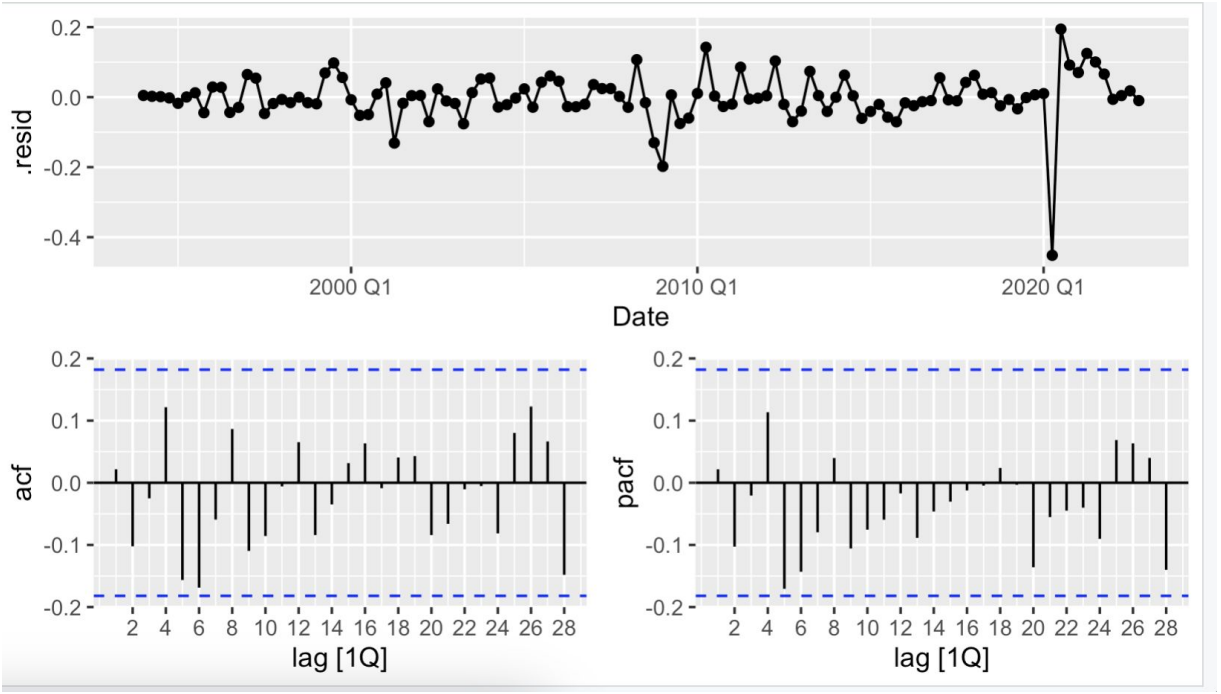
SARIMA (1,1,2) x (0,1,1)

Here, we present a forecast for the last four quarters in our holdout data, spanning from **Q1 2023 to Q4 2023**.
Our model demonstrates proficiency in both forecasting and accurately capturing the patterns in our historical data.

Given its performance, we intend to proceed with this model to check our residuals charts.



Correlogram of residuals



Here we can see that **all the values of ρ(k) are within 95% bands**, it tells us that the model adequately captures the temporal dependencies in the data (how past observations influence future observations within the same series.) and the forecasts derived from the model are likely to be reliable.

Ljung-Box test

Null Hypothesis (H0): There is no residual autocorrelation in the time series at different lags.
Alternative Hypothesis (Ha): There is residual autocorrelation present in the time series at different lags.

On performing a Ljung-Box test on our selected model, we got a p value of **0.723**

When significance level (α) is set at **0.05**, then **0.723 > 0.05**, **we fail to reject the null hypothesis.**

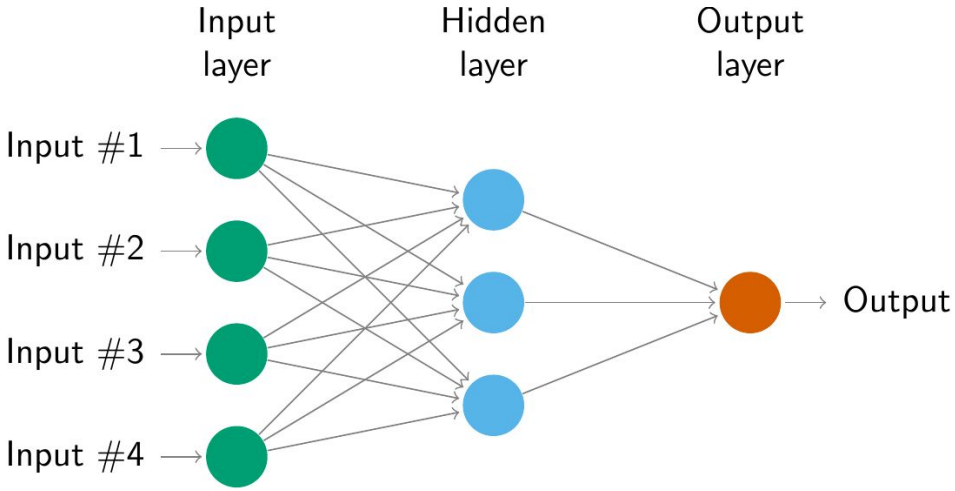
This shows the effectiveness of our model in capturing the underlying structure of the time series data.

Model 2 : Neural Networks

For our second model, we will be looking at a **Feed forward neural network** using R's (**NNETAR**) tool.

Unlike **traditional linear autoregressive** models, **neural networks** can **capture complex nonlinear relationships** in the data, allowing them to model more intricate patterns and dependencies

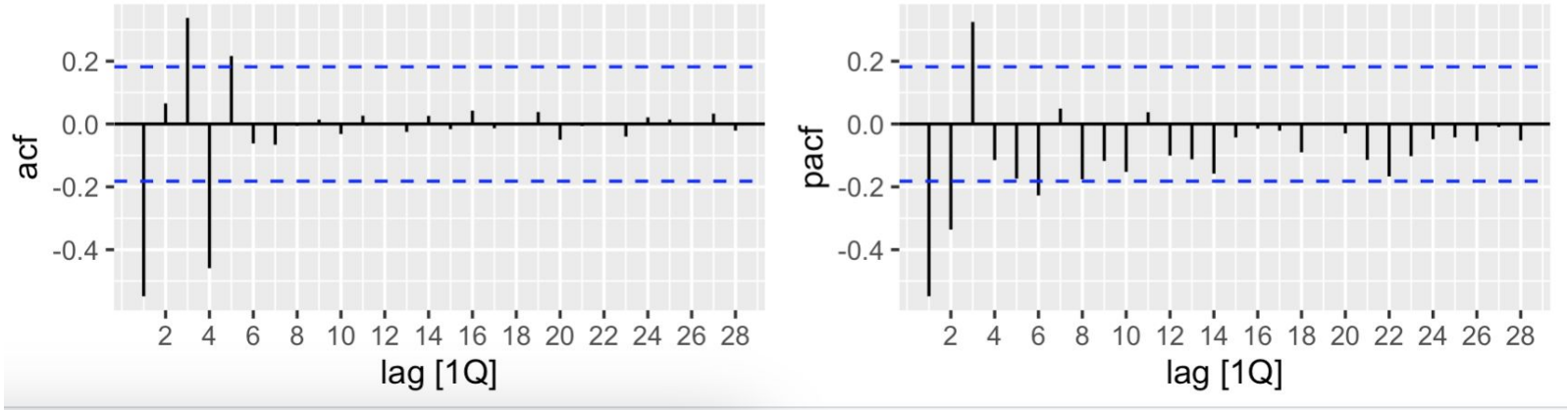
A neural network autoregression (**NNETAR**) model works by incorporating **lagged values** of the time series data (**autoregressive terms**) as input features to predict future values.



Steps for NNETAR model

Check for missing values → Our data has **none**!

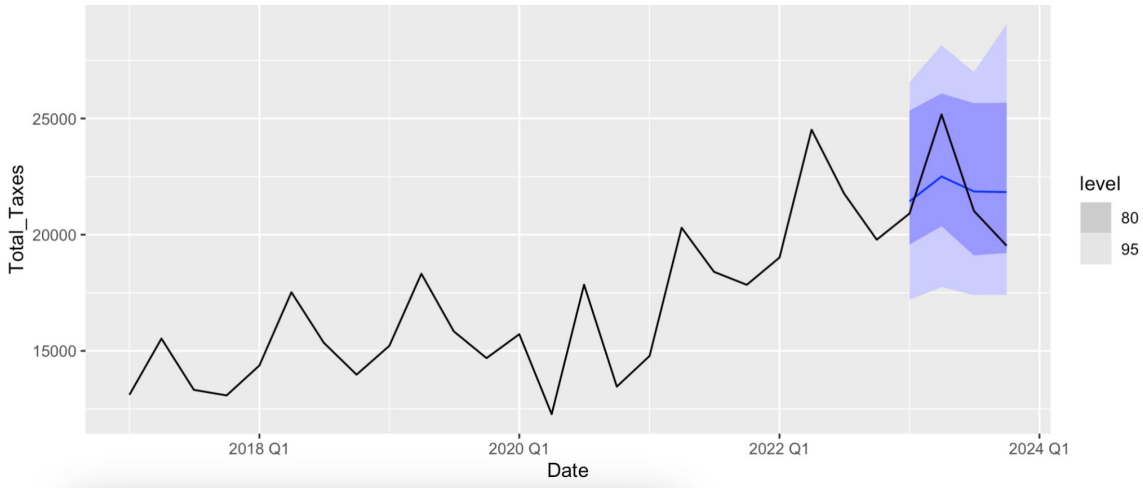
Check our acf and pacf charts (Infer p and P)



1st - Preliminary model

```
prelimNET=TaxHold%>%model(NNETAR(log(Total_Taxes)~
          AR(P=0,p=2),n_networks=200))
```

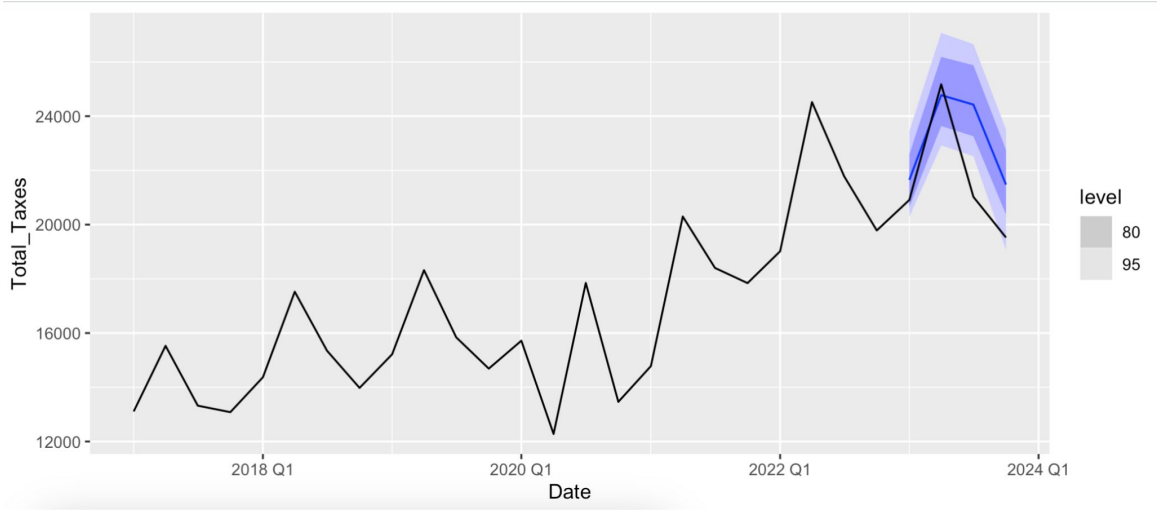
For our preliminary model we have selected **p=2** and **P=0**. Also selected networks as 200. Here we have not used any fourier terms.



2nd - Model with Fourier terms

```
SMALLfourierNET=TaxHold%>%model(NNETAR(log(Total_Taxes)~AR(P=0,p=3)+
                fourier(period=4,K=2),n_networks=200))
```

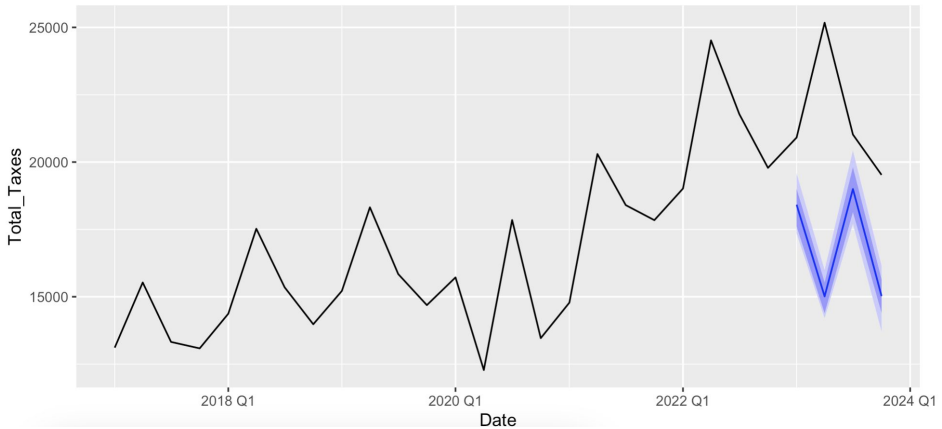
Here we have used **quarterly fourier terms** as we can suspect quarterly seasonality might be at play. The Fourier terms will capture the seasonality in the data with this periodicity. Here we have included **2 fourier terms**.

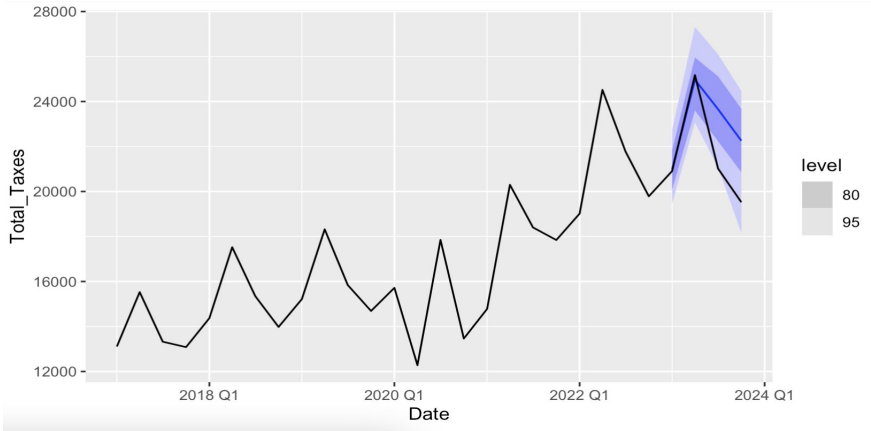


3rd - Model with additional Fourier terms

```
BigfourierNET=TaxHold%>%model(NNETAR(log(Total_Taxes)~AR(P=0,p=2)+
                fourier(period=4,K=2)+fourier(period=12,K=2),n_networks=200))
```

```
fourier(period=8,K=2),
```





Selecting a Model

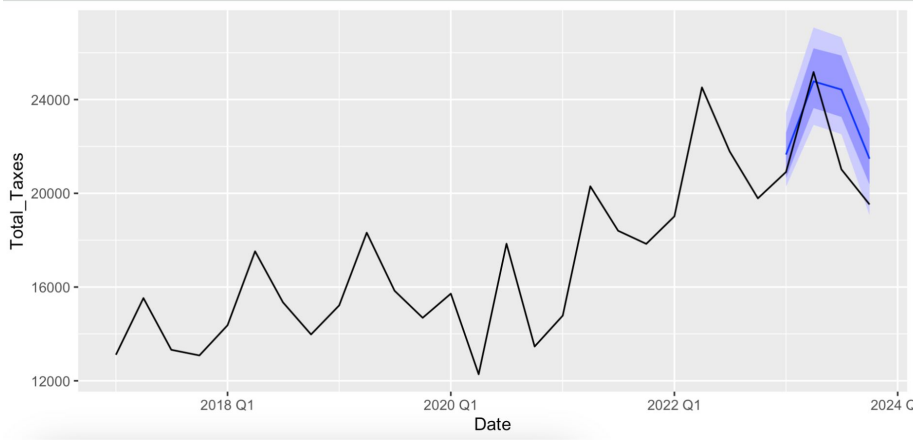
Here let's check **SMALLfourierNET** and **BigfourierNET** models σ2 and RMSE values.

Final Model for NNET

SMALLfourierNET

$\sigma^2 = 0.0032$

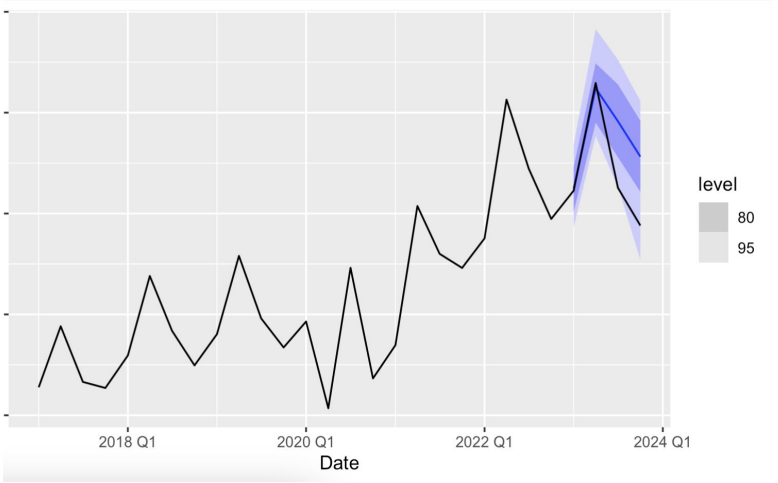
RMSE = 2546



BigfourierNET

$\sigma^2 = 0.0021$

RMSE = 2266



Comparing our SARIMA and NNETAR models w.r.t RMSE

SARIMA (1,1,2) x (0,1,1) → RMSE = 1529

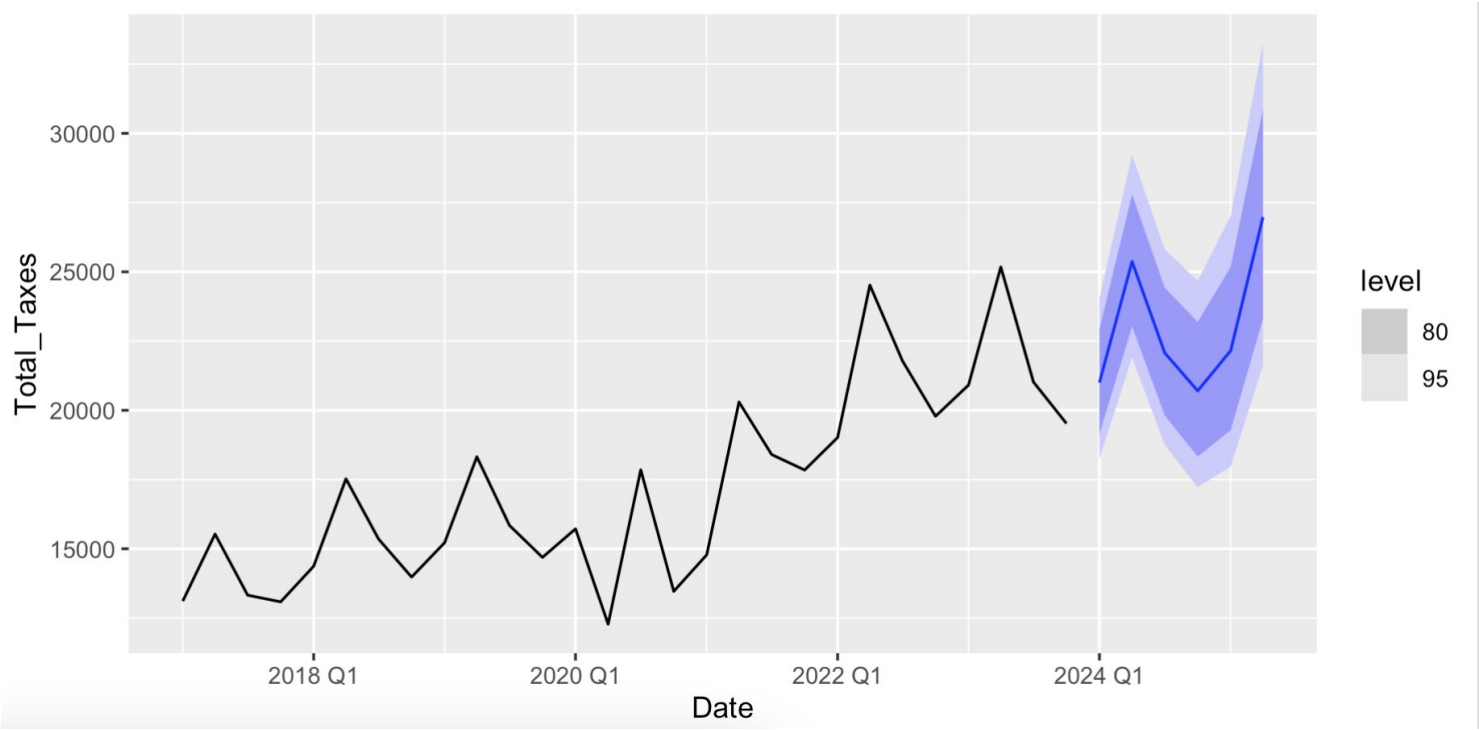
BigfourierNET → RMSE = 1683

Lower RMSE indicates that the model's predictions are closer to the actual values in the dataset.

Hence we will select **SARIMA (1,1,2) x (0,1,1)** for our future predictions

6-step ahead forecast

Here we have forecasted 6-step ahead using the entire sample. The forecast is provided for next six quarters from **Q1 2024** to **Q2 2025**



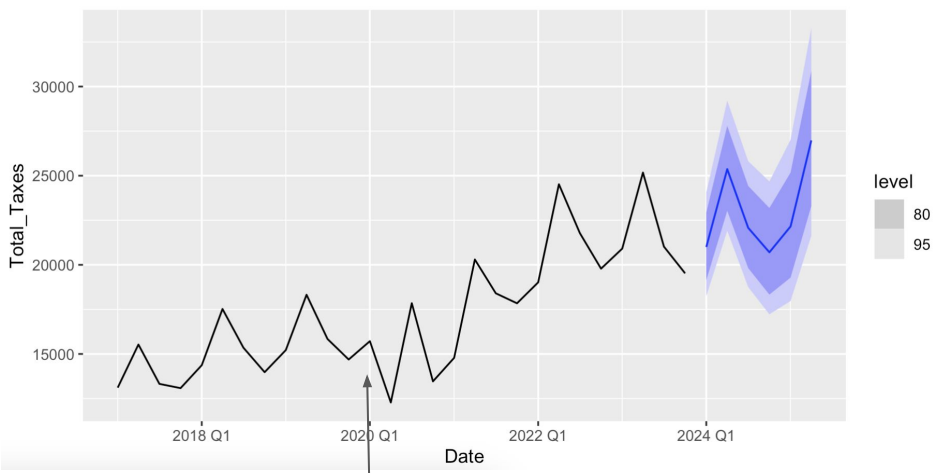
Assessing the Implications of the Forecast on Stakeholders and Decision-Making

Revenue Trends:

This forecast provides insights into the expected trajectory of tax revenues in Texas over the next six quarters.

Impact on Financial Planning:

The forecasted tax revenues can significantly influence budgetary planning for the state government. Understanding how revenues are projected to change can help policymakers allocate resources effectively and prioritize spending on key initiatives.



Unforeseen events:

It's important to recognize the forecast's limitations. Unforeseen events like the **COVID-19 pandemic** in 2020 demonstrate that forecasting models cannot predict every eventuality. It's essential to remain prepared for unexpected disruptions and to maintain flexibility in our planning strategies.