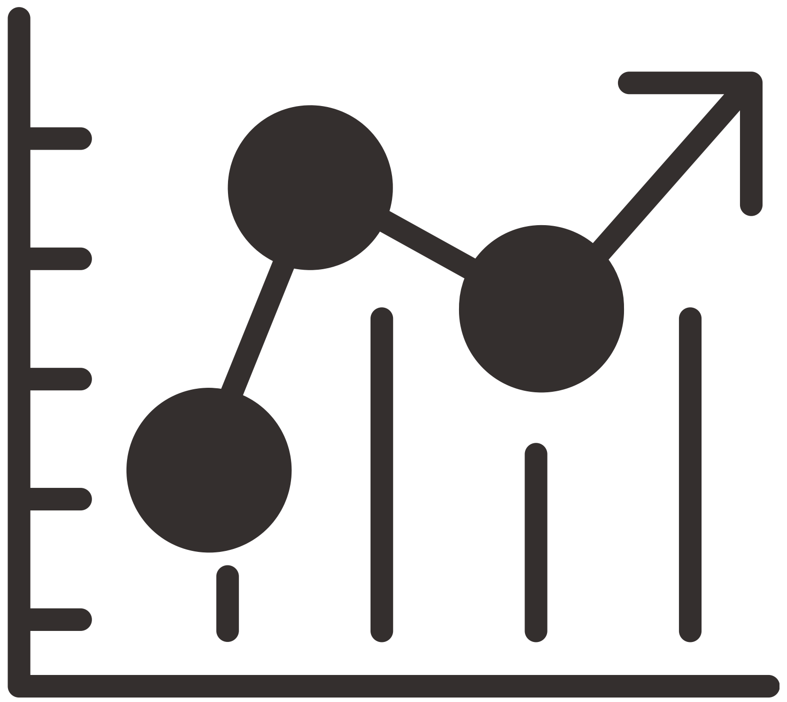
Time Series Forecasting Final Project

USA Credit Card Deliquency Rate Prediction

Group 10 - Arunabh Choudhury, Priyal Desai, Purvi Panchal, Sahana Kumar, Yash Gupta

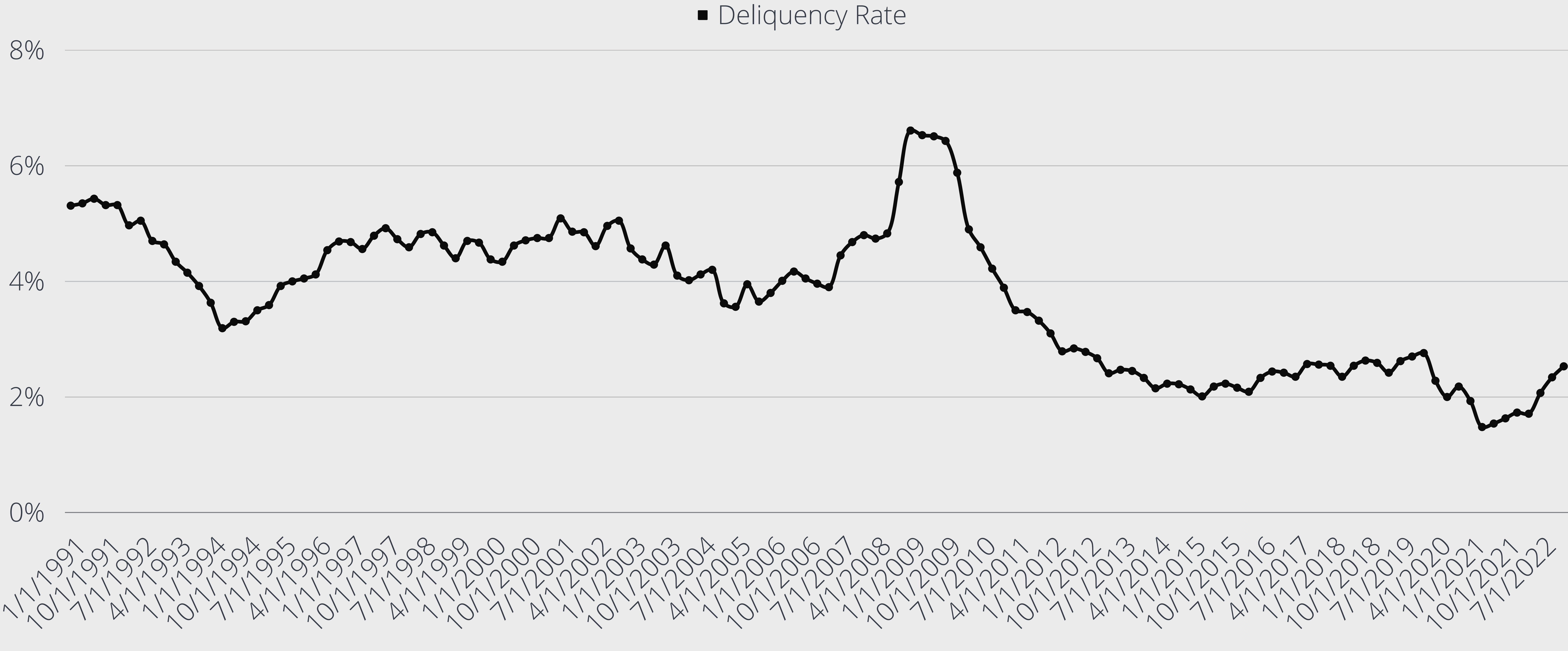


PROBLEM STATEMENT

The Credit Card Delinquency Rate USA Forecasting project aims to address the pressing issue of predicting and managing credit card delinquency rates in the United States. As credit card usage continues to rise, so does the concern for delinquent payments, which have significant implications for both consumers and financial institutions. The lack of accurate forecasting models hampers the ability of credit card companies and banks to effectively mitigate risks, leading to potential financial losses and negative consequences for individuals' credit scores. Therefore, there is a critical need to develop a reliable and robust forecasting framework that can accurately predict credit card delinquency rates, enabling stakeholders to make informed decisions and implement proactive measures to minimize defaults and financial distress.

The goal of this project is to leverage historical credit card data, and time series forecasting techniques to build an accurate forecasting model for credit card delinquency rates in the USA.

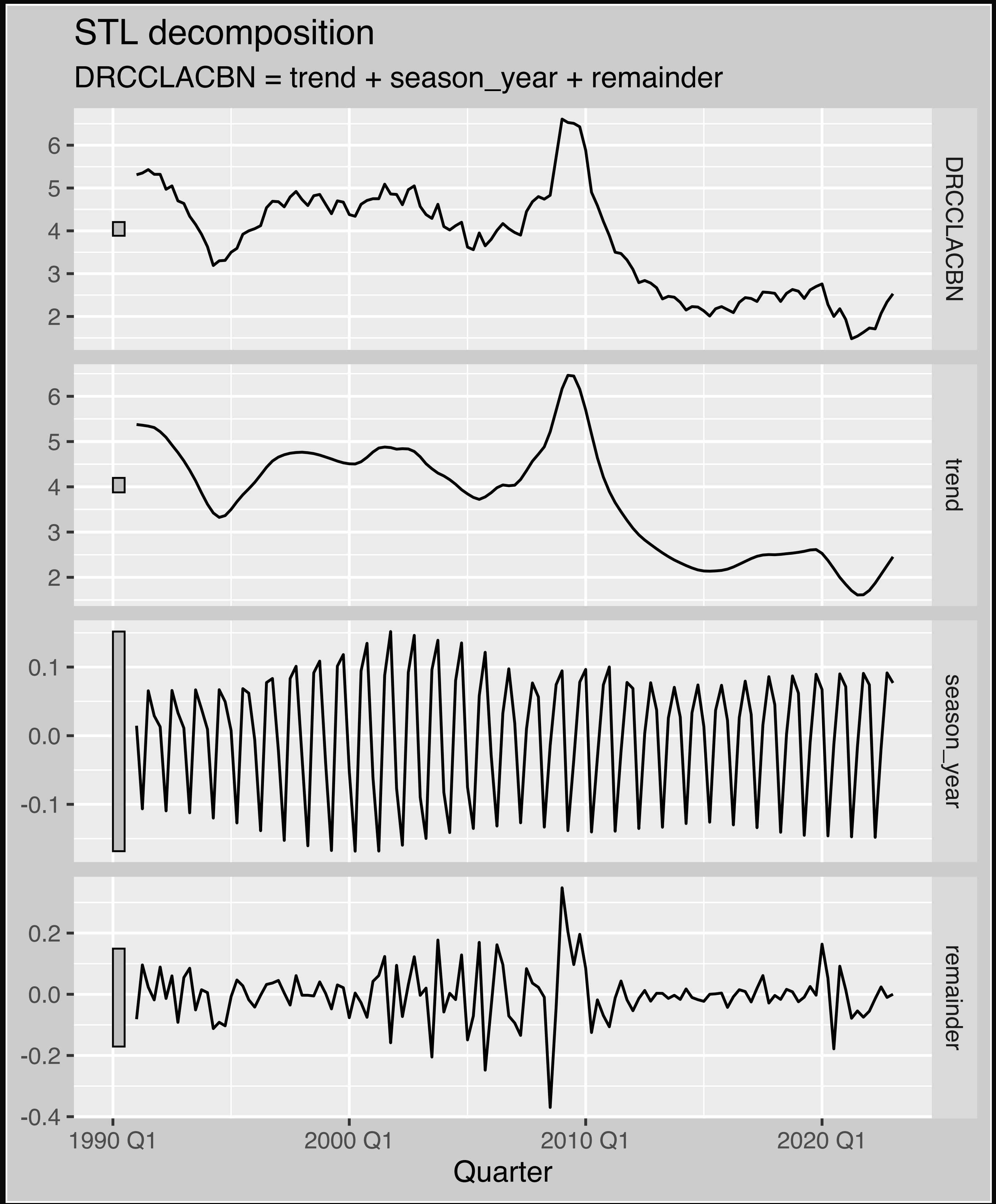
The Quarterly Data on the US Credit Card Deliquency Rate has been taken from the FRED Website. Below is the trend of Default rate from 1991 to 2022



WORKFLOW

STL Decomposition	Helps us understand the Trend, Seasonality, Irregularity in the Time Series Data for future Modeling.
Benchmark Models	Commonly Used Models like Mean, Naive, Seasonal Naive, Drift, Linear Model, STLF for Time Series Forecasting
Stationarity Check	Stationarity check is necessary in time series analysis to ensure that the statistical properties of the data, such as mean and variance, remain constant over time.
Smoothing Models	Smoothing is needed in time series forecasting to reduce noise, identify trends, remove seasonal effects, detect outliers, and enhance data visualization.
ACF PACF	ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots are used to determine the order of autoregressive (AR) and moving average (MA) terms
ARIMA Models	Choosing the Best ARIMA Model for Forecasting based on Accuracy Measures
Final Forecast	Final Forecasting Based using the Best ETS Model with Suitable Accuracy Measures

STL DECOMPOSITION



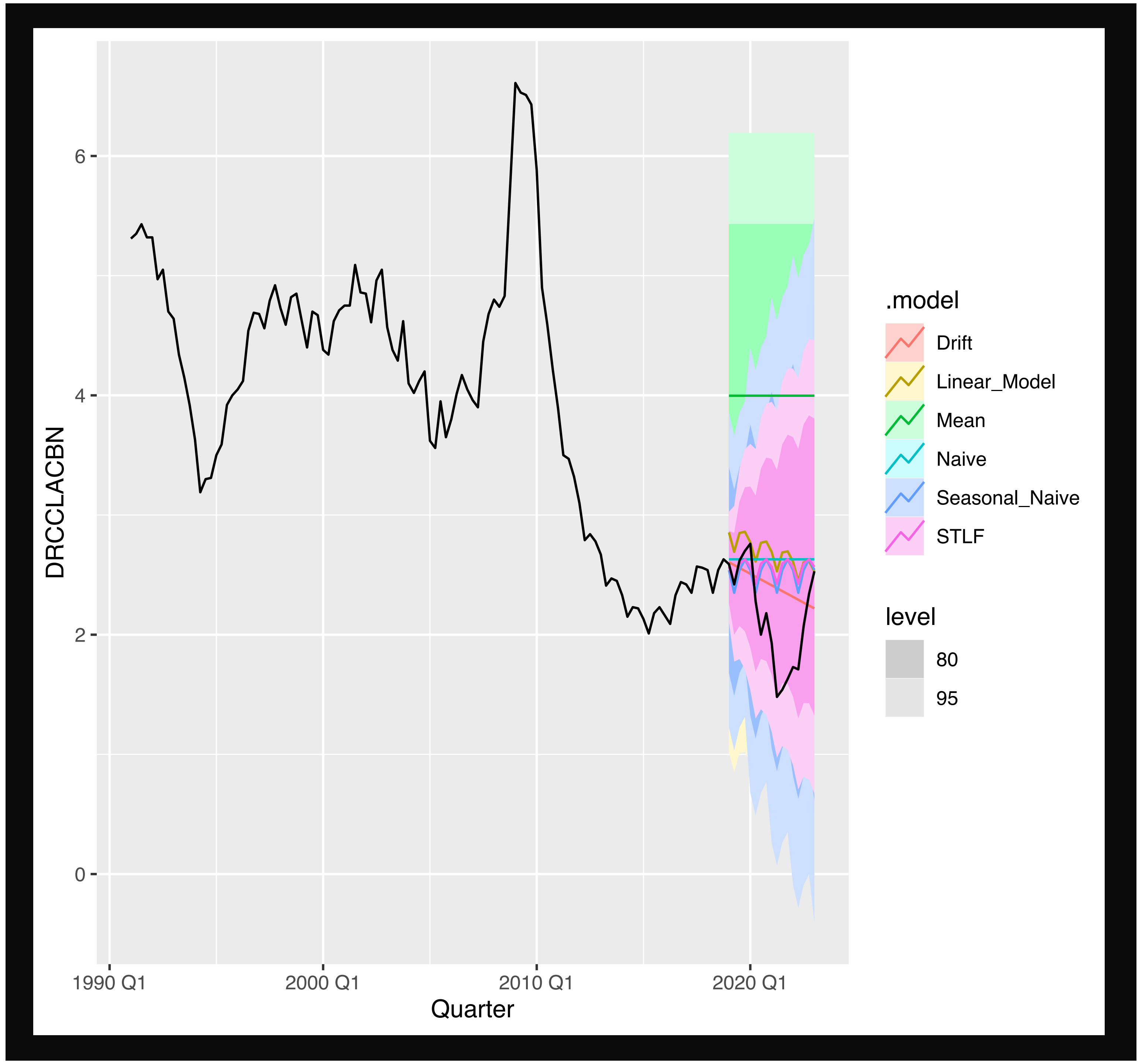
- **STL (Seasonal and Trend decomposition using Loess) is a method for decomposing time series data into seasonal, trend, and remainder components.**
- It helps in analyzing and understanding the underlying patterns in the data.
- STL decomposition separates the seasonality, trend, and irregular variations, allowing for a deeper understanding of each component.
- It is a flexible and effective approach for time series analysis.
- STL decomposition is valuable for tasks like forecasting, anomaly detection, and further modeling of the time series data.
- It provides insights into the seasonal patterns, long-term trends, and residual variations in the data.

Trend Component: Trend- The trend component of the graph shows overall the delinquency rate has decreased over the years. There was a major rise in the years 2008-2010 and has been on a steady decrease ever since.

Seasonal Component: The seasonality component is mostly steady with an increase until 2002 and a decrease thereafter. The seasonality remains steady post the rise

Residual Component: The remainder component of a time series is the random variation in the data. We can see an increase in variation of the remainder component between the years 2001 to 2010 with its peak around 2008-2010

BENCHMARK MODELS



There are a few models which are widely used for time series forecasting, like **Random Walk with Drift, Naive, Mean Models** et al.

We implemented these models to forecast values from 2019 and measured the **RMSE** which was the lowest for Random Walk with Drift Model.

However, when we did the **LJung Box Test,** the test used to assess whether a time series exhibits significant autocorrelation beyond what would be expected under the assumption of **white noise**. If the test result shows a lack of significant autocorrelations, it suggests that the data resembles white noise, where the observations are uncorrelated and independent of each other.

We found that the residuals are not white noise suggesting that there is information left uncaptured.
Hence, these models may not be the best to proceed with for final forecasting models.

STATIONARITY CHECK

Stationarity implies that the statistical properties of the series, such as **mean, variance, and autocovariance, do not change over time.**

The **KPSS (Kwiatkowski-Phillips-Schmidt-Shin)** test is a statistical test used to assess the stationarity of a time series. It is employed to determine whether the series exhibits a unit root, indicating non-stationarity, or is stationary.

The test examines the null hypothesis that the series is stationary against the alternative hypothesis of a unit root or non-stationarity. The need for the KPSS test arises from the importance of stationarity in time series analysis.

If the KPSS Pvalue is less than 0.05, the series is not - stationary. Hence we need to convert the original data to a seasonally differenced data to check for stationary.

KPSS Result on Original Data

	kpss_stat	kpss_pvalue
	<dbl>	<dbl>
1	1.72	0.01

SEASONAL DIFFERENCING

Seasonal differencing is often necessary in time series analysis to remove the seasonal component and achieve stationarity.

Seasonality refers to regular patterns that occur at fixed intervals, such as daily, weekly, or yearly cycles. These patterns can introduce non-stationarity, where the mean, variance, or autocorrelation structure of the time series changes over time.

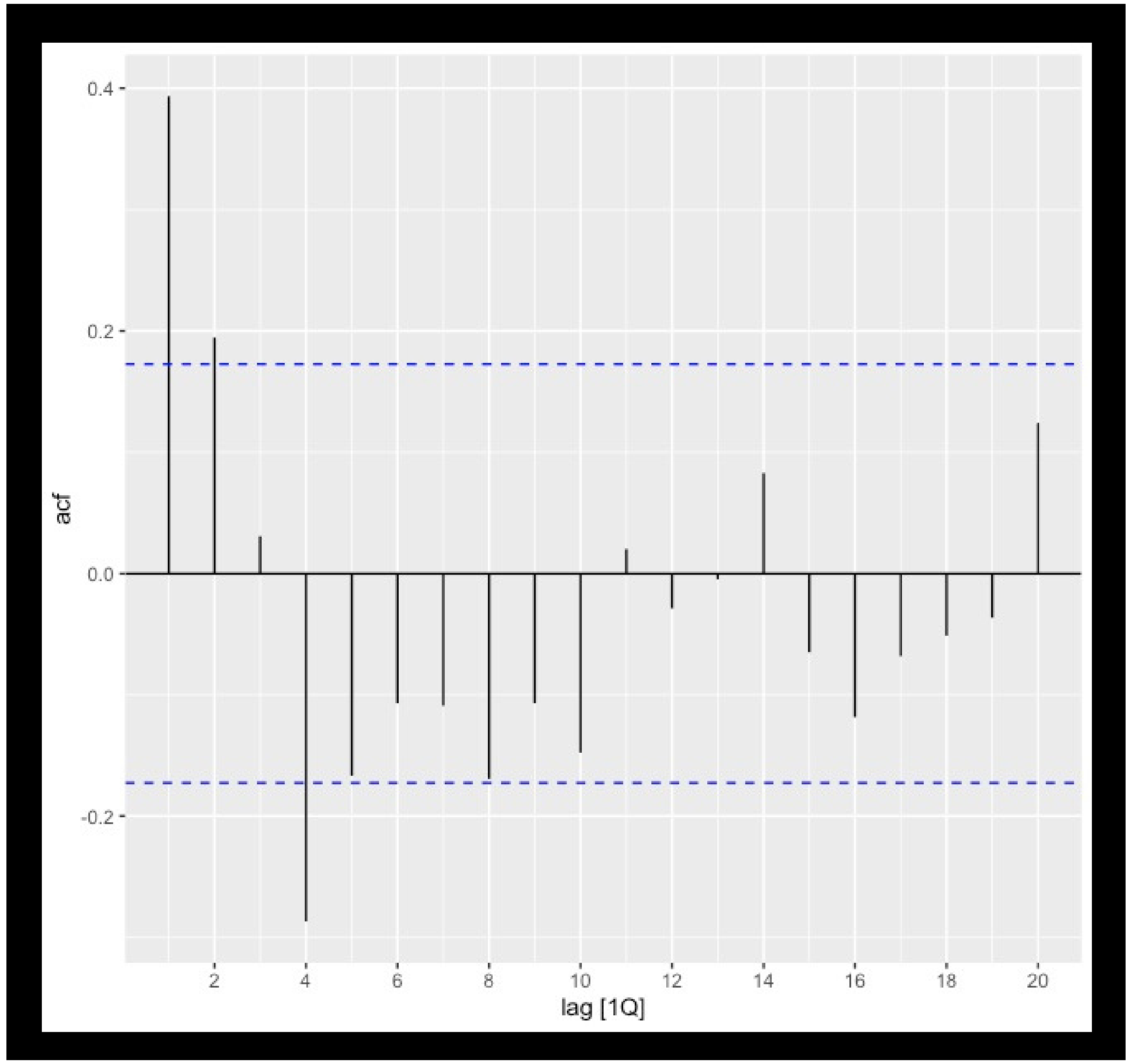
By applying seasonal differencing, the seasonal patterns are removed by subtracting the observation at the current time step from the observation at the same time step in the previous season.

This differencing process helps in eliminating the effects of seasonality and making the series stationary, which is a fundamental assumption for many time series models and statistical tests

KPSS Result on Seasonally Differenced Data

	kpss_stat	kpss_pvalue
	<dbl>	<dbl>
1	0.0389	0.1

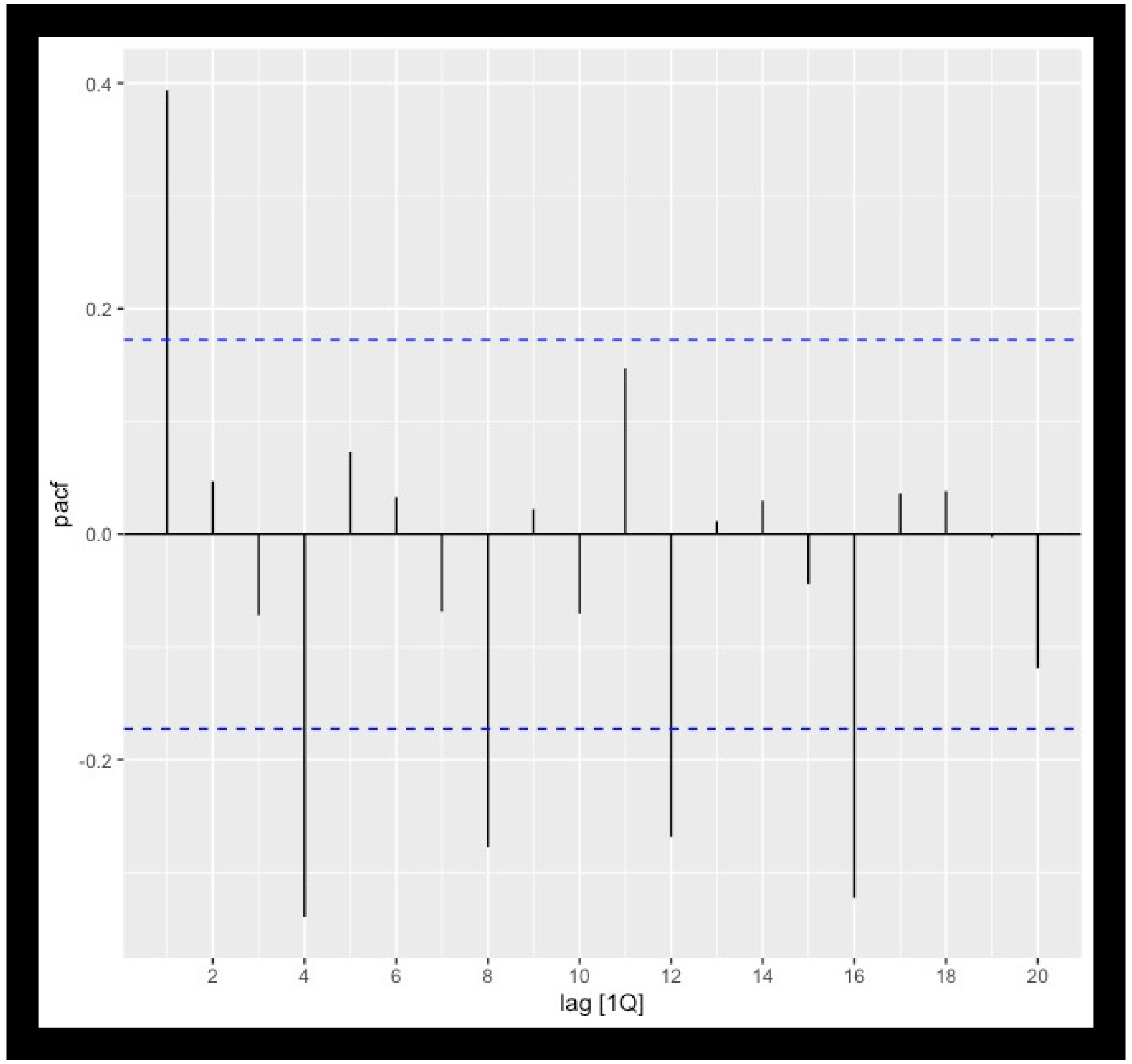
ACF



ACF (Autocorrelation Function) plots are used to visualize and analyze the autocorrelation structure of a time series. The ACF plot shows the correlation coefficients between the observations at different lags, indicating the relationship between each observation and its past values.

The need for ACF plots arises from the importance of understanding the autocorrelation patterns in a time series. Autocorrelation refers to the correlation between observations at different time points, and it can provide insights into the underlying patterns, trends, and seasonality present in the data

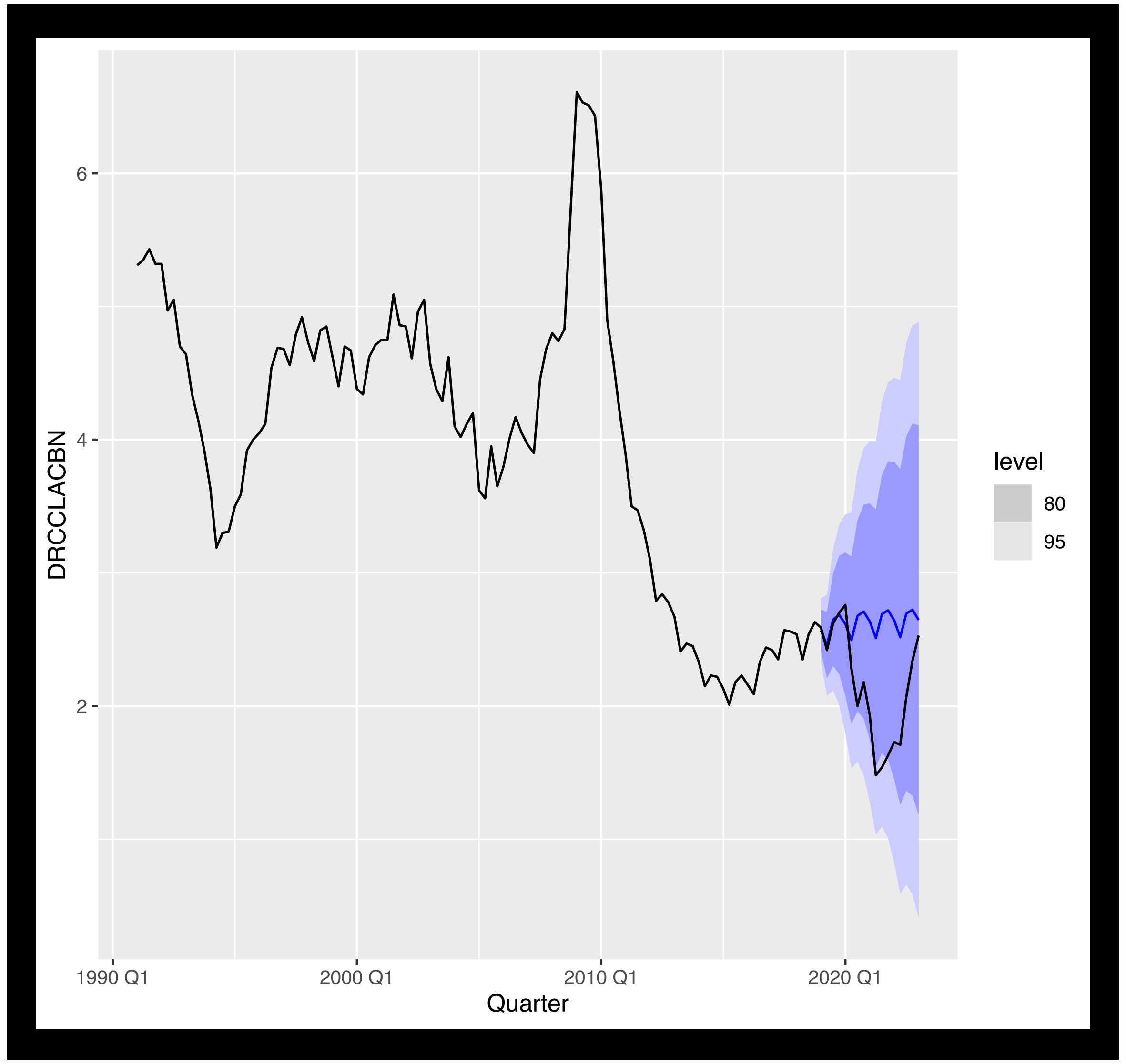
PACF



PACF (Partial Autocorrelation Function) plots are used to analyze the partial correlation between observations at different lags in a time series, after removing the correlation explained by intervening lags. PACF plots are particularly useful in identifying the appropriate lag order for autoregressive models.

The need for PACF plots arises from the complexity of the autocorrelation structure in time series data. While the ACF plot provides information about the overall autocorrelation at different lags, the PACF plot helps to isolate and understand the direct relationship between an observation and its specific lag.

SMOOTHING MODELS



Forecast Using ETS Smoothing

There are various smoothing techniques used like HOLT, Damped, SES, et al.

Smoothing techniques are commonly applied in time series analysis to reduce noise, remove short-term fluctuations, and highlight underlying trends or patterns in the data. Here are the essential details in bullet points:

Noise Reduction: Smoothing helps to eliminate or reduce the impact of random fluctuations or noise in the time series, making it easier to identify and interpret the underlying patterns and trends.

Trend Identification: Smoothing techniques can help reveal the long-term trend in the data by reducing the influence of short-term fluctuations. This is especially useful when analyzing data with high-frequency or irregular variability.

Outlier Detection: Smoothing methods can help identify and handle outliers by attenuating their impact on the overall trend. By reducing the impact of extreme values, smoothing can provide a more accurate representation of the underlying data patterns.

Data Compression: Smoothing can be used to compress the time series data by reducing the number of data points while preserving the essential characteristics of the series. This can be beneficial for data storage, visualization, or computational efficiency.

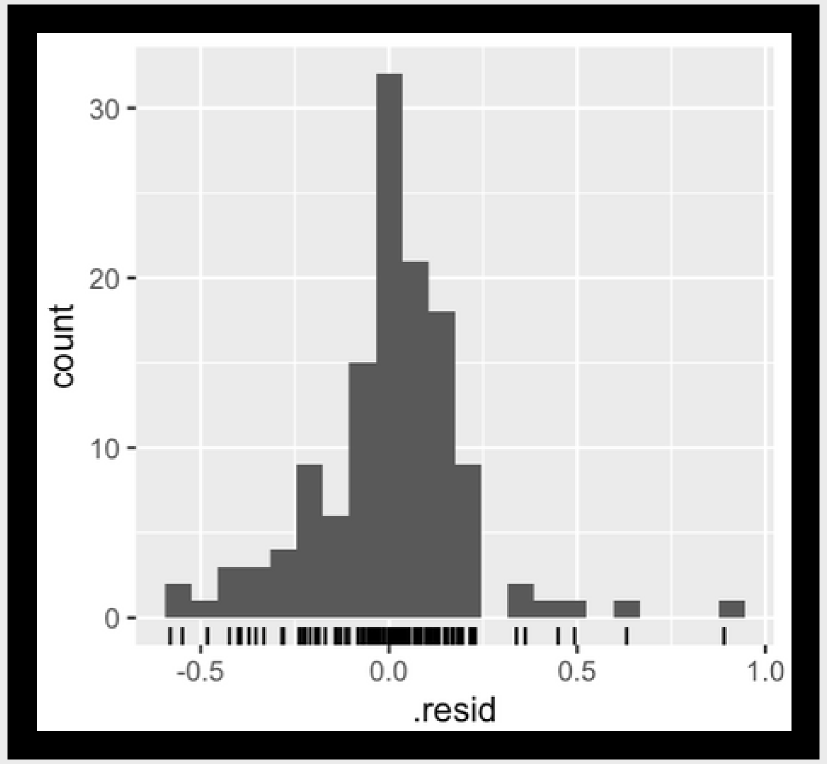
Forecasting: Smoothing techniques, such as moving averages or exponential smoothing, are often employed as a basis for forecasting future values. By removing noise and short-term fluctuations, smoothing can provide a more stable and reliable basis for making predictions.

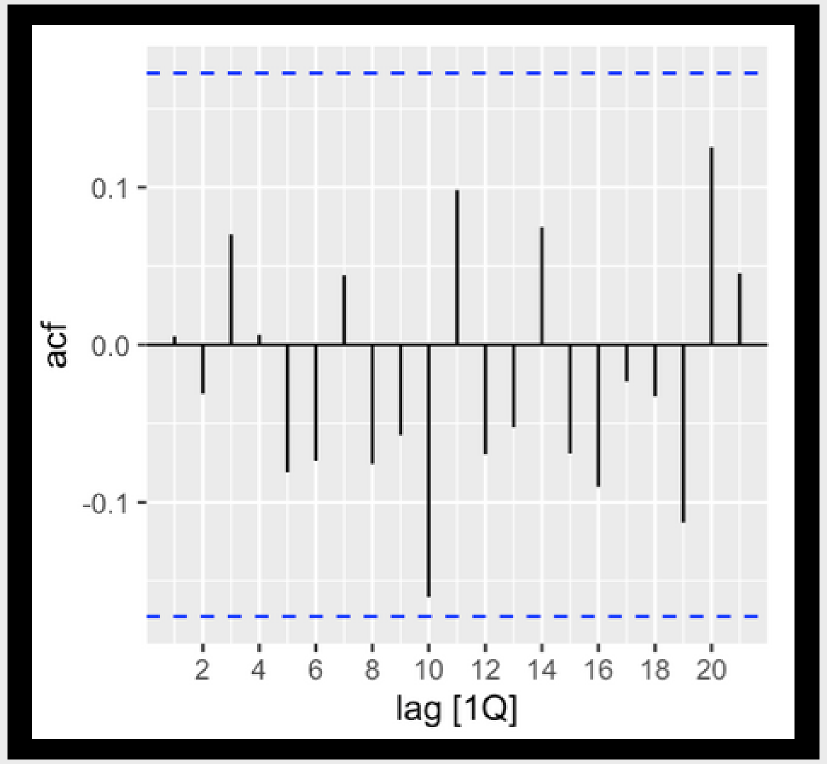
ARIMA MODELS

- ARIMA models combine AR (autoregressive), MA (moving average), and I (differencing) components to capture patterns and dependencies in time series data.
- The **AR component represents the relationship between the current observation and its past values, capturing autocorrelation.**
- The **MA component models the relationship between the current observation and past error terms, accounting for remaining autocorrelation.**
- The **I component applies differencing to remove trend and seasonality, ensuring stationarity of the time series.**
- ARIMA model parameters are denoted by "pdq" notation: "p" is the AR order, "d" is the differencing order, and "q" is the MA order.
- Lower values of "p", "d", and "q" indicate simpler models, while higher values capture more complex dependencies.
- The "PDQ" notation is used for seasonal ARIMA models, where "P", "D", and "Q" represent the seasonal AR, differencing, and MA orders, respectively. The general form of an ARIMA is **ARIMA(p,d,q)(P,D,Q)**

For this time series, we tried ARIMA Models with combination of pdq and PDQ values and assessed the performance using RMSE and AICc Also, performed a LJung Box Test to see verify is Residuals are white noise and Homoscedastic. Hence ARIMA(1,1,1)(0,1,1) is the Best Model.

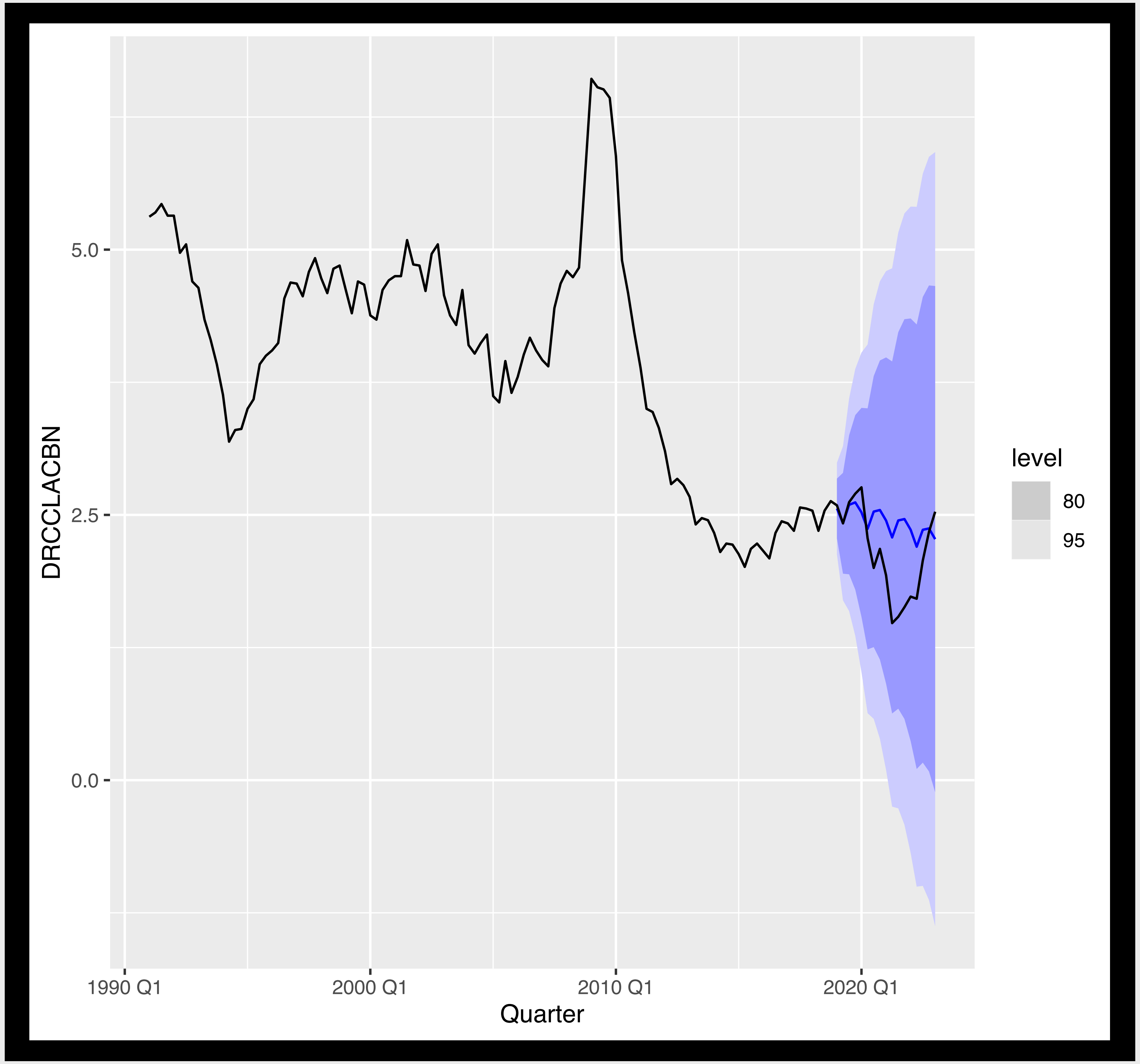
.model	.type	RMSE	MAE	MAPE	MASE	RMSSE
<chr>	*<chr>*	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*	*<dbl>*
arima112211	Test	0.467	0.359	20.1	0.742	0.695
arima111011	Test	0.467	0.360	20.2	0.744	0.696
arima_initial	Test	0.470	0.362	20.3	0.748	0.700
arima212111	Test	0.489	0.378	21.2	0.780	0.729
arima111111	Test	0.496	0.381	21.5	0.788	0.739
arima112111	Test	0.496	0.381	21.5	0.788	0.739
arima211111	Test	0.496	0.381	21.5	0.788	0.739
auto	Test	0.725	0.582	32.4	1.20	1.08





Final Forecasted Plot

We have chosen the final forecasts from 2019 using ARIMA(1,1,1)(0,1,1)



Thank You