FORECASTING EARTH SURFACE TEMPERATURE

GROUP 7

DHANVANTH VOONA A20543395
NEERAJ VARDHAN A20545853
SAI CHARAN A20543155

CONTENT

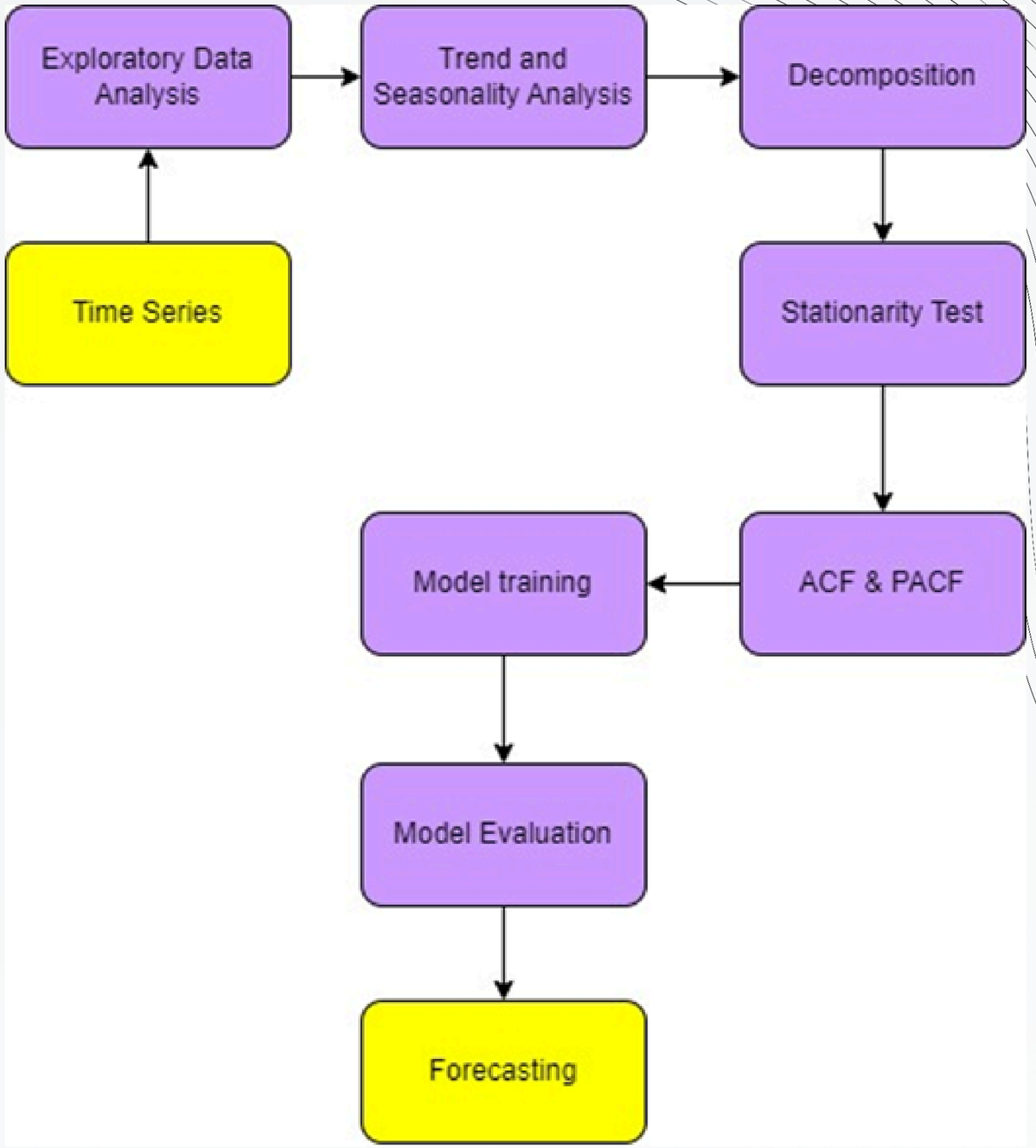
INTRODUCTION

- In the face of rapid climate change, understanding and predicting Earth's surface temperature have never been more critical.
- Our project aims to leverage historical temperature data spanning two centuries to forecast future temperature trends.
- By employing state-of-the-art forecasting models, we seek to provide valuable insights into the impacts of climate change and contribute to mitigation efforts.
- we delve into the past, analyze the present, and forecast the future of Earth's surface temperature.

PROBLEM STATEMENT

- Earth's surface temperature influences numerous aspects of life on our planet, including agriculture, ecosystems, and human health
- Accurate temperature forecasts are essential for planning and mitigating the impacts of climate change, such as extreme weather events, sea level rise, and shifts in agricultural productivity.
- Furthermore, historical temperature trends play a crucial role in analyzing the effectiveness of current climate policies and guiding future actions.
- Therefore, the main objective of our project is to forecast Earth's surface temperature using historical temperature data spanning two centuries.
- By employing state-of-the-art forecasting models, we aim to provide valuable insights into future temperature trends and contribute to the ongoing efforts to combat climate change.

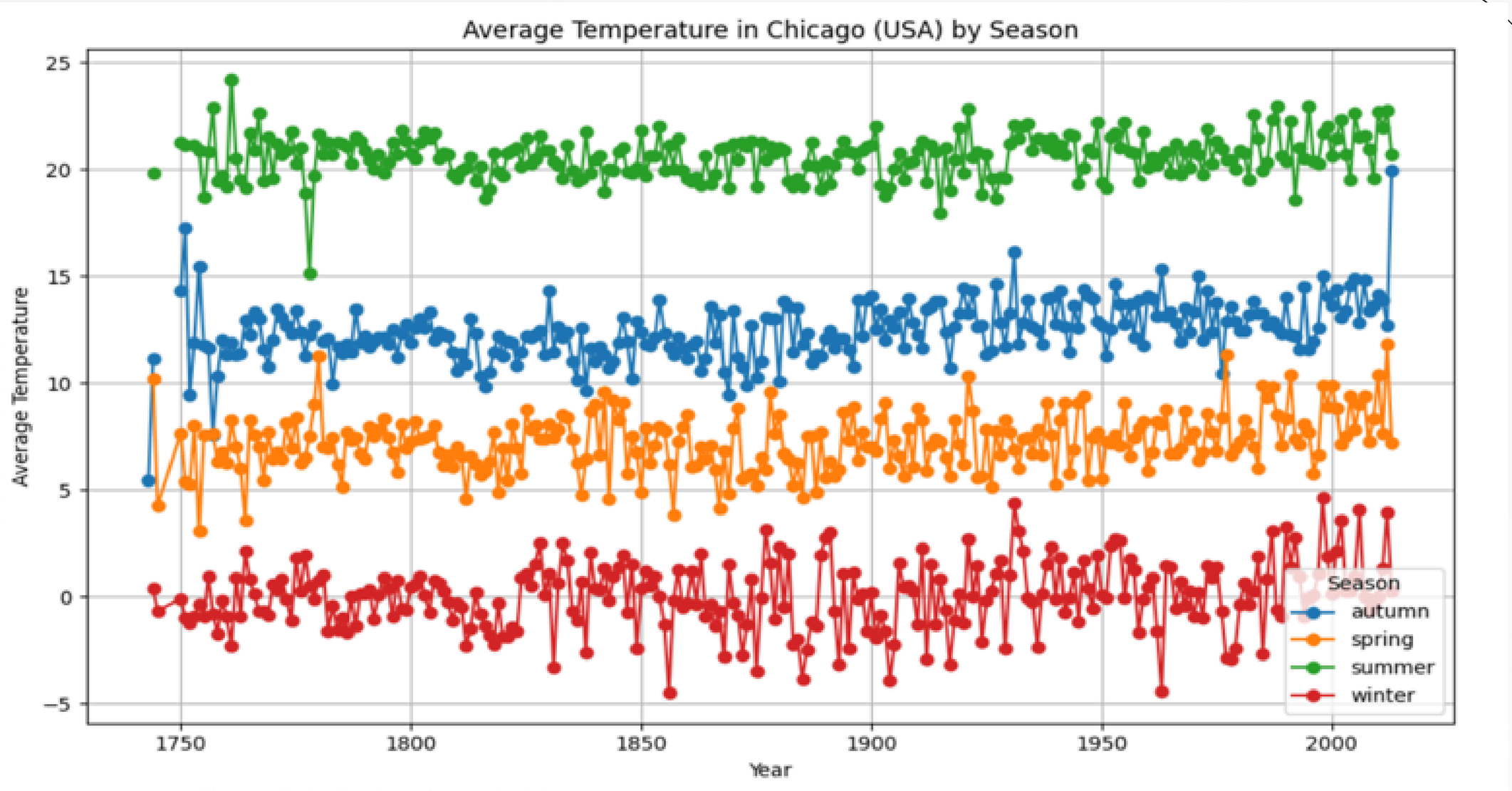
PROPOSED ALGORITHM



DATASET DETAILS

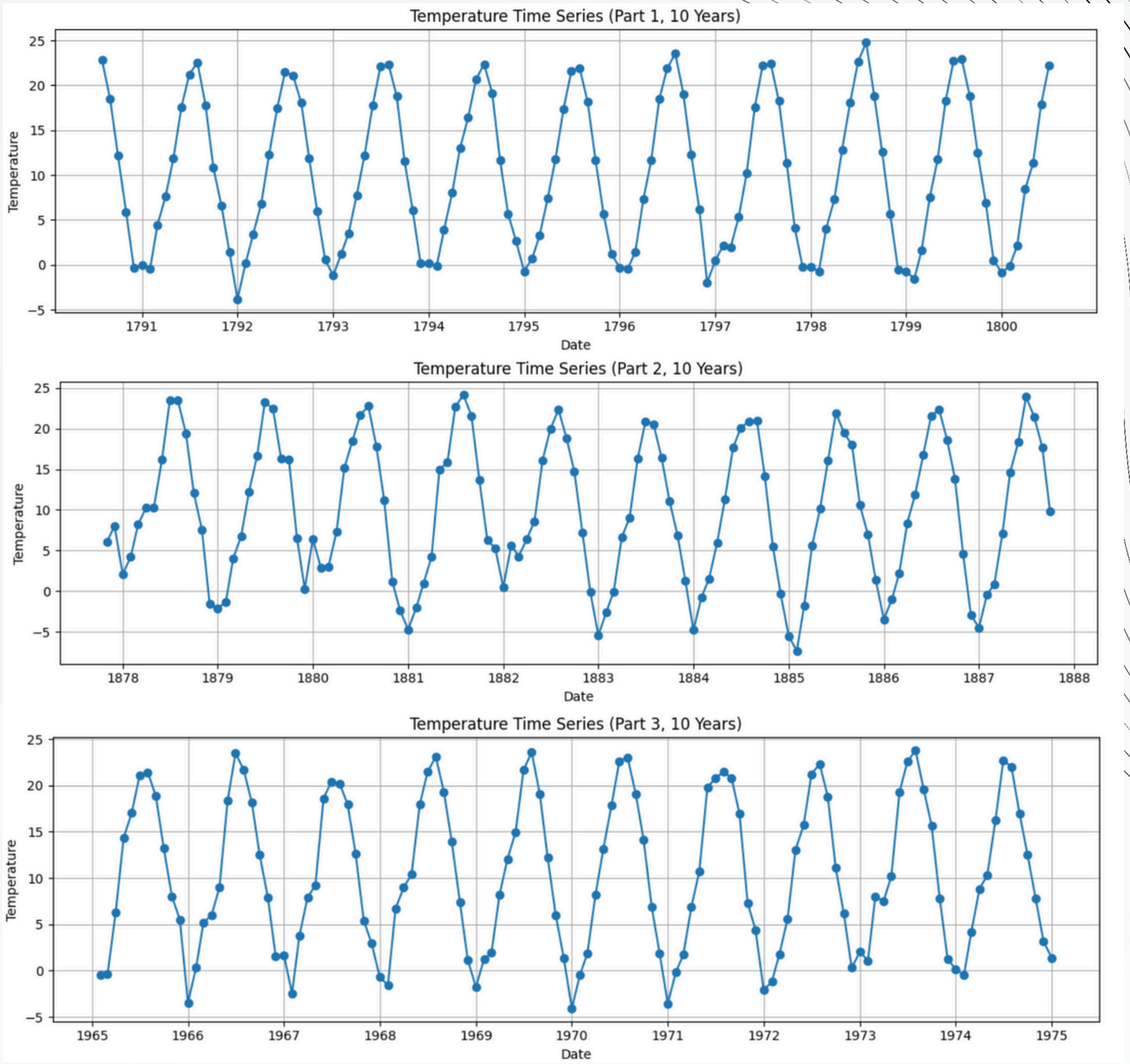
- Date Range: It has observations of average temperature of each month from 1743 to 2013.
- The temperature recordings belong to 159 countries and several cites from each countries.
- Metrics Included:
  - DT(Date): Represents the date of the observation.
  - Average Temperature (Float/ Celsius): The average temperature recorded for the specific date and location.
  - AverageTemperatureUncertainty(Float/Celsius): The uncertainty interval around the average temperature, indicating the potential variability or error in the temperature measurement.
  - City(String): The city in which the temperature was recorded.
  - Country(String): The country in which the city is located.

Analysing Chicago Temperature by Season



SEASONALITY ANALYSIS FOR THE LAST 3 DECADES

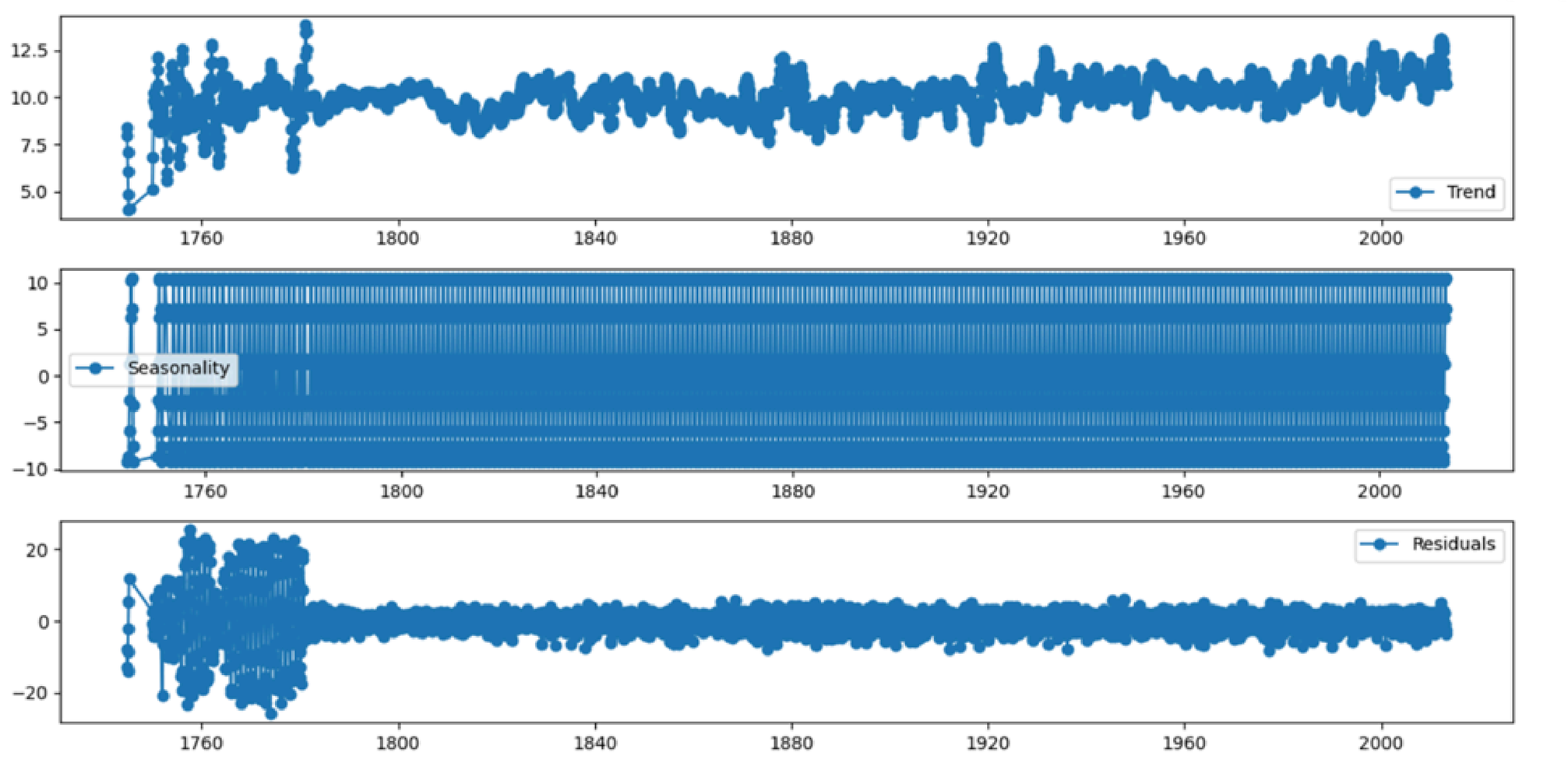
It is very evident that there is presence of seasonality with an approximate time-period of 12 months for each split indicating the time period of seasonality hasn’t changed over the last 3 centuries.



DECOMPOSITION

- Decomposition is breaking down a time series into its constituent components:
- **Trend:**
  - Long-term progression or directionality of the data.
- **Seasonality:**
  - Repeating patterns or fluctuations at regular intervals.
- **Residual:**
  - Unexplained part of the data representing random fluctuations.
- **Accurate Forecasting:**
  - Understanding trend and seasonality is crucial for accurate predictions.
  - Separating components improves forecasting accuracy.
- **Stationarity:**
  - Separating trend and seasonality makes data more stationary.
  - Stationary data is easier to model and forecast accurately.

Trend, Seasonal and Residual Components of the given data



We can observe in trend component is mostly in a plateau or stagnant position without significant upward or downward movement. Whereas in Seasonal component we can clear see pattern or fluctuation that occur at a regular intervals

STATIONARY CHECK

- **Null Hypothesis (Ho)**: that the time series has a unit root and is non-stationary.
- **Alternative hypothesis (Ha)**: that the time series does not have a unit root and is stationary.

Pvalue	Stationary	Non-Stationary
<0.05		
>0.05		

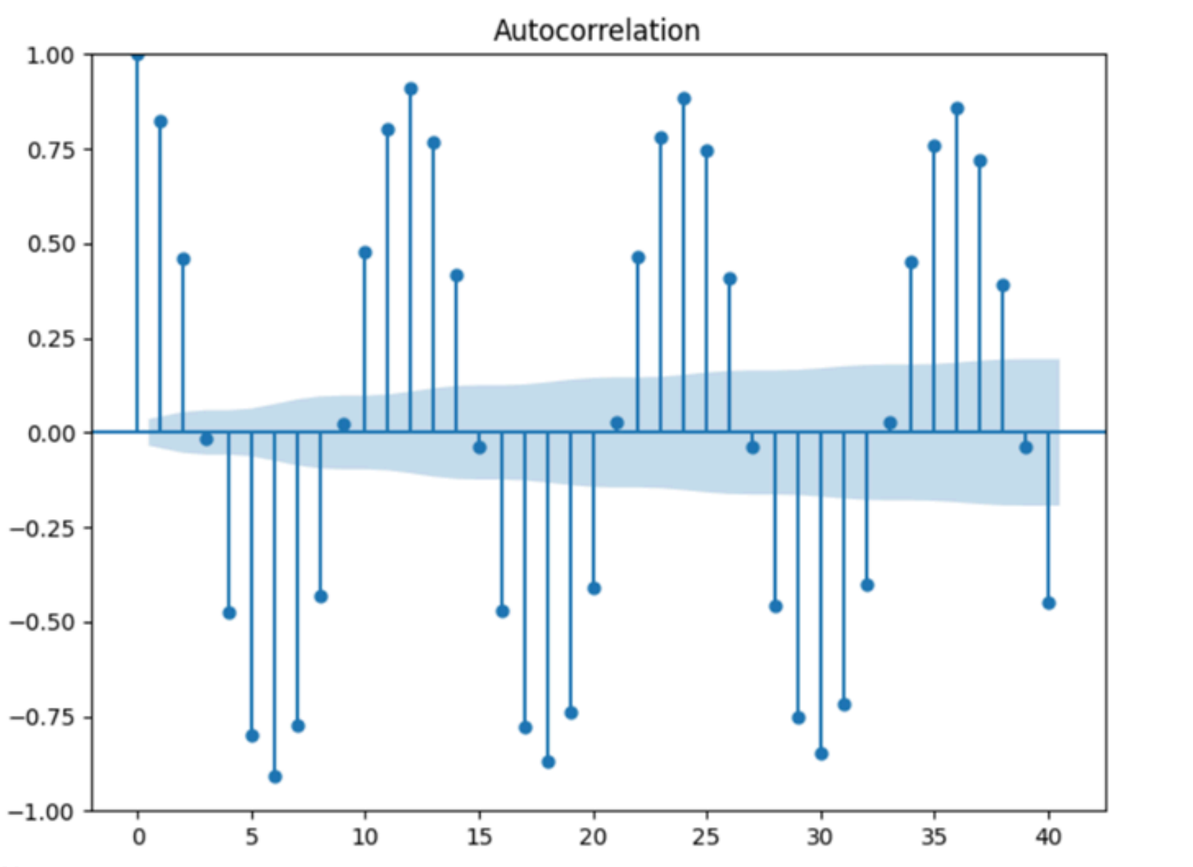
STATIONARY CHECK

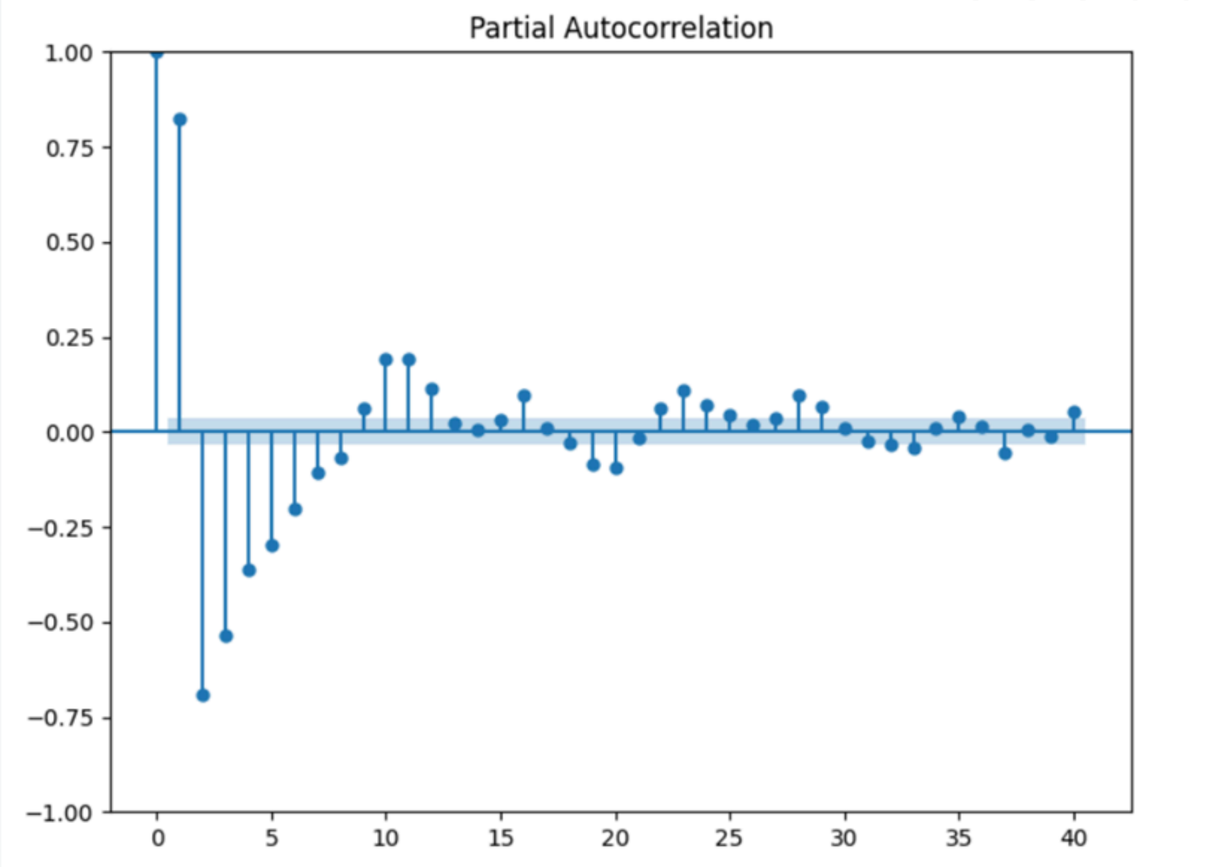
```
adfuller_test(data2['Temp'])
```

```
ADF Test Statistic : -6.136682572298648
p-value : 8.155104183672214e-08
#Lags Used : 28
Number of Observations Used : 3112
strong evidence against the null hypothesis(Ho), reject the null hypothesis. Data has no unit root and is stationary
```

We can see that the p-value <<< 0.05 which rejects the null hypothesis and indicates that time series does not have a unit root and is stationary. Since the time series is stationary we need not deseasonalize the time series.

ACF and PACF





- The seasonal pattern in the ACF suggests that there is a repeating pattern in the data at regular intervals, indicating the presence of seasonality.
- The slow decay of PACF indicates that each observation is related to its seasonal neighbors.

ARIMA

- ARIMA, which stands for Autoregressive Integrated Moving Average, is a robust statistical method utilized for analyzing and forecasting time series data.
- Autoregression (AR):AR captures the dependency between an observation and its lagged observations. Put simply, the current value of the time series depends on its past values.
- Integrated (I):The integration process makes the time series stationary by differencing raw observations. Stationarity ensures consistent statistical properties over time, which is crucial for reliable analysis.
- Moving Average (MA):MA models the relationship between an observation and the residual errors from a moving average model applied to lagged observations. It captures the short-term fluctuations in the data.

$$X_t - \phi X_{t-1} = Z_t + \theta Z_{t-1},$$

where $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$ and $\phi + \theta \neq 0$.

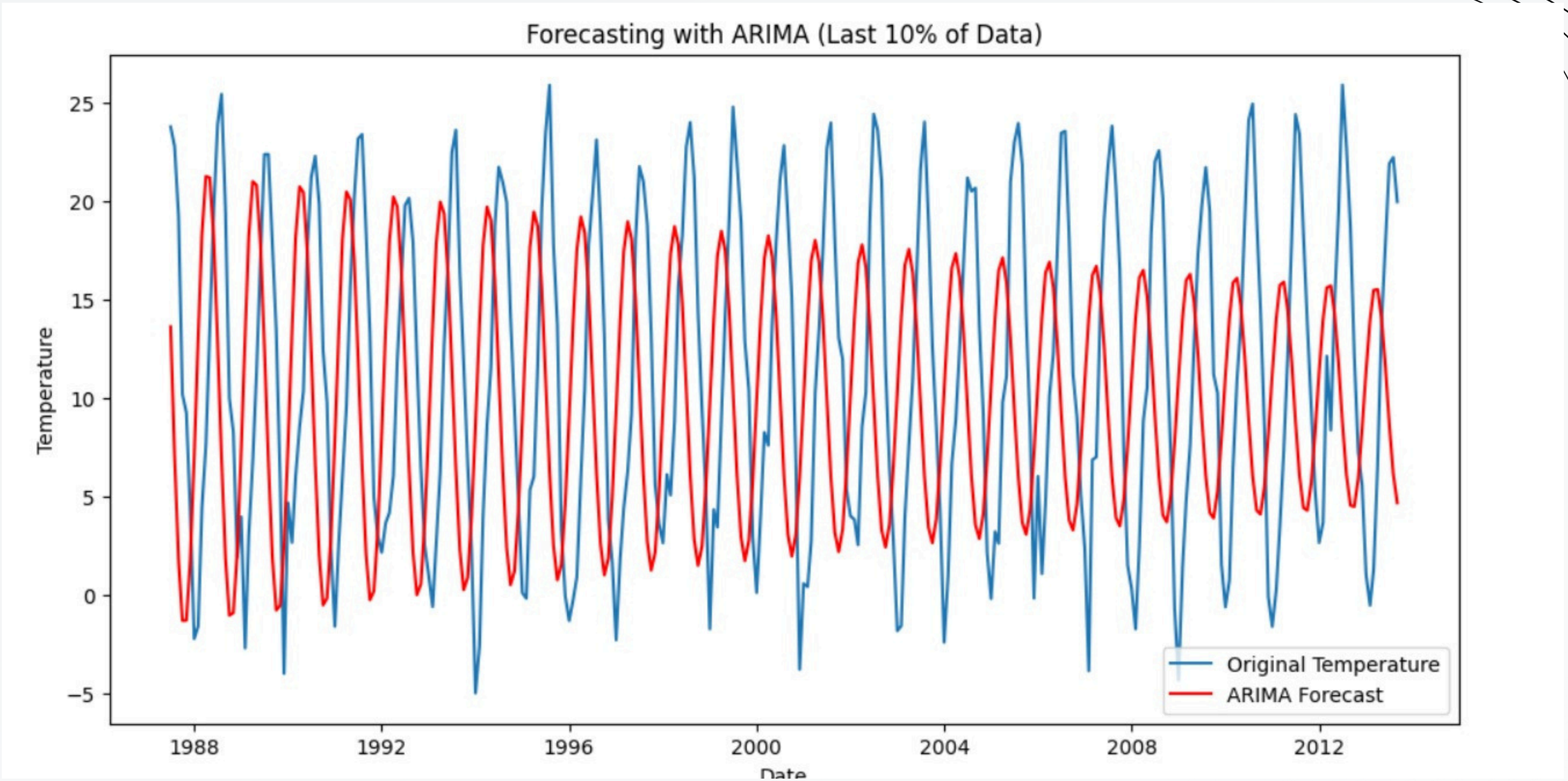
ARIMA

```python
# Grid search for best pdq values
best_aic = float("inf")
best_pdq = None
for p, d, q in product(range(0, 5), range(0, 5), range(0, 4)):
    try:
        model = ARIMA(data2['Temp'], order=(p, d, q))
        model_fit = model.fit()
        aic = model_fit.aic
        if aic < best_aic:
            best_aic = aic
            best_pdq = (p, d, q)
    except:
        continue
```

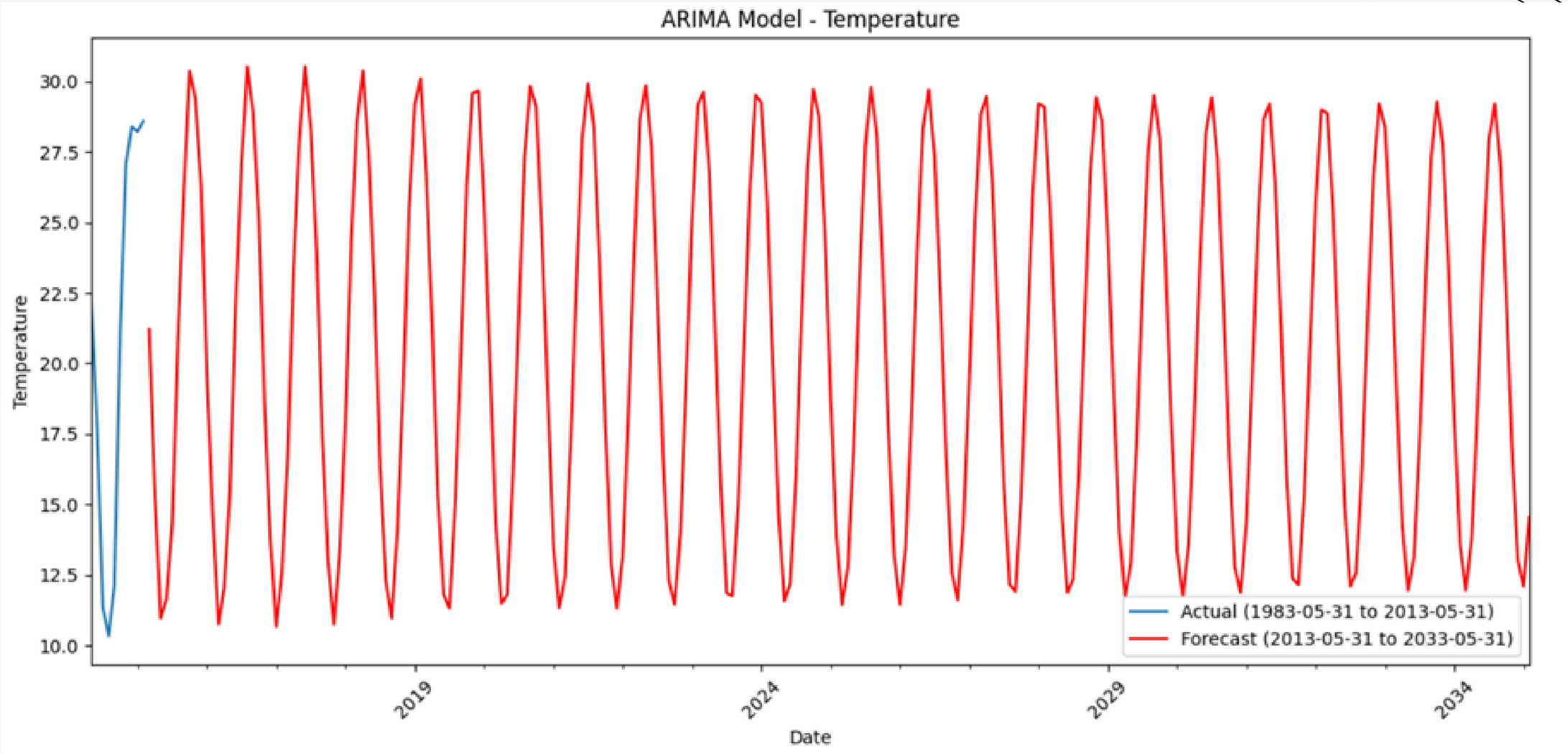
```
Top 10 combinations of parameters with the best AIC scores:
1. AIC: 14503.18796408741, PDQ: (2, 0, 3)
2. AIC: 14549.612198621595, PDQ: (2, 1, 3)
3. AIC: 14668.077459489316, PDQ: (2, 0, 2)
4. AIC: 14946.046733142935, PDQ: (2, 1, 2)
5. AIC: 15082.753426534904, PDQ: (2, 2, 3)
6. AIC: 15094.653432092233, PDQ: (2, 0, 1)
7. AIC: 16665.6416619096, PDQ: (2, 0, 0)
8. AIC: 17271.91124640028, PDQ: (1, 0, 2)
9. AIC: 17446.374980492343, PDQ: (0, 0, 3)
10. AIC: 17598.680010663913, PDQ: (1, 1, 3)
Best PDQ values: (2, 0, 3)
```

- **Best ARIMA Model**: ARIMA(2, 0, 3) - AIC: 14503.18
  - Identified through grid search of parameter combinations (p, d, q).
  - AIC (Akaike Information Criterion) used as the optimization metric.

Forecasting with ARIMA

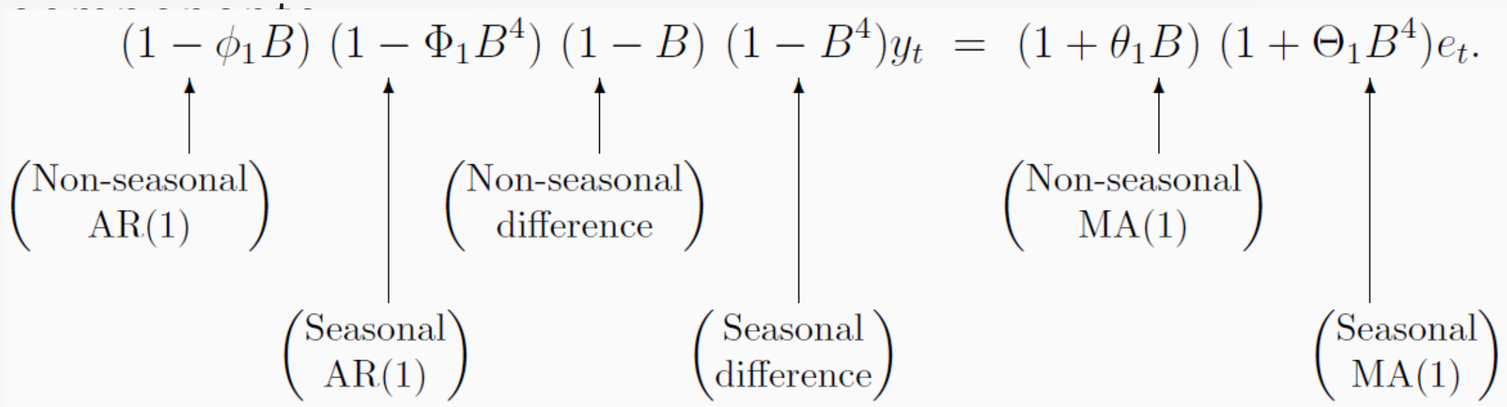


Forecasting with ARIMA



SARIMA

- SARIMA, or Seasonal Autoregressive Integrated Moving Average, is an extension of the ARIMA model that incorporates seasonal components. It is used for analyzing and forecasting time series data with seasonal patterns
- **Autoregression (AR)**: Captures dependence between an observation and its lagged values.
- **Integrated (I)**: Makes the time series stationary by differencing raw observations.
- **Moving Average (MA)**: Models the relationship between an observation and residual errors.
- **Seasonal (S)**: Captures seasonal patterns in the data.
- Enables accurate modelling and forecasting of time series data with both non-seasonal and seasonal components



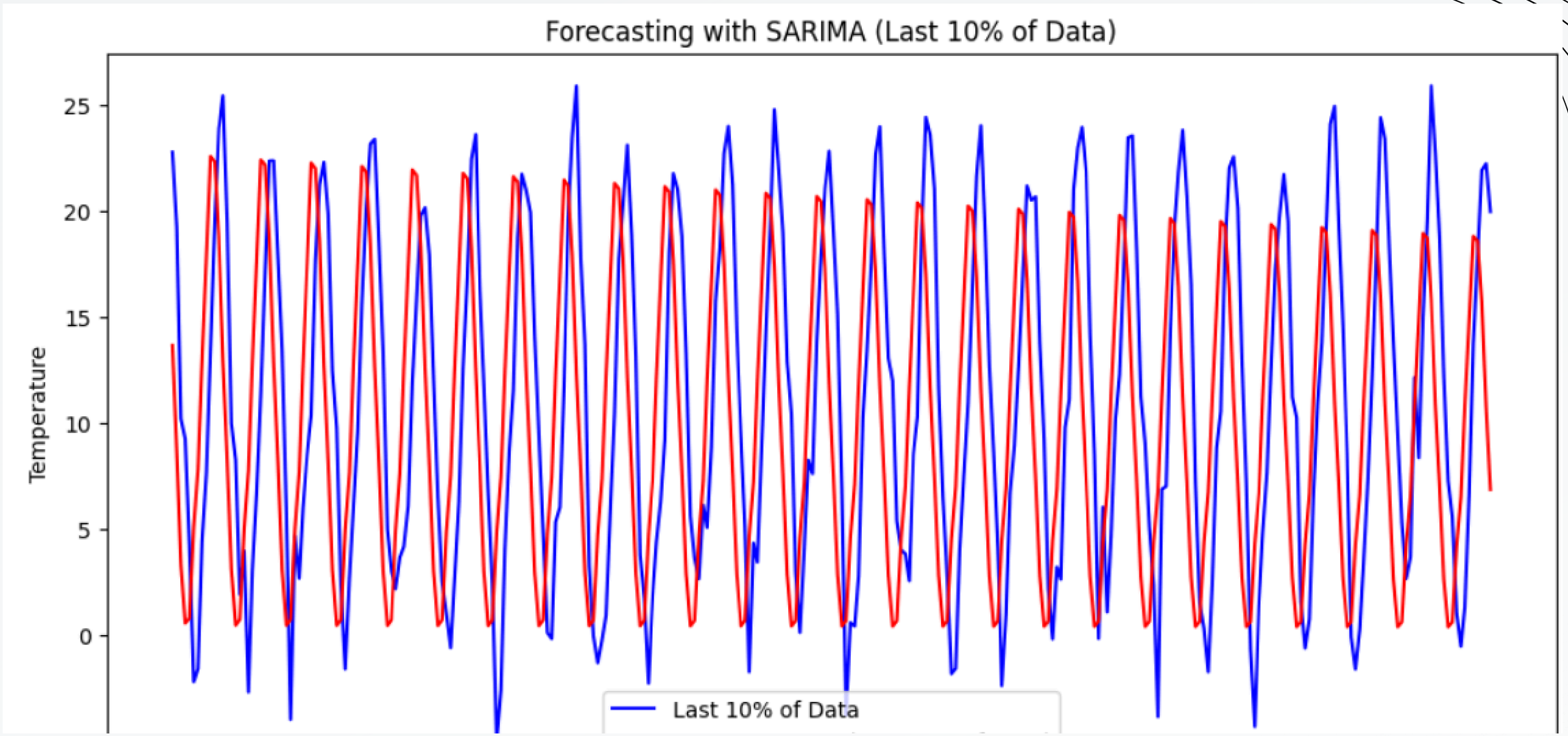
SARIMA

```
for param in pdq:
    for seasonal_param in seasonal_pdq:
        try:
            model = SARIMAX(train_data['Temp'],
                            order=param,
                            seasonal_order=seasonal_param,
                            enforce_stationarity=False,
                            enforce_invertibility=False)
            results = model.fit()
            aic = results.aic
            if aic < best_aic:
                best_aic = aic
                best_order = param
                best_seasonal_order = seasonal_param
#             print(f'SARIMA{param}x{seasonal_param} - AIC: {aic:.2f}')
        except:
            continue

print(f'Best SARIMA model: SARIMA{best_order}x{best_seasonal_order} - AIC: {best_aic:.2f}')
```

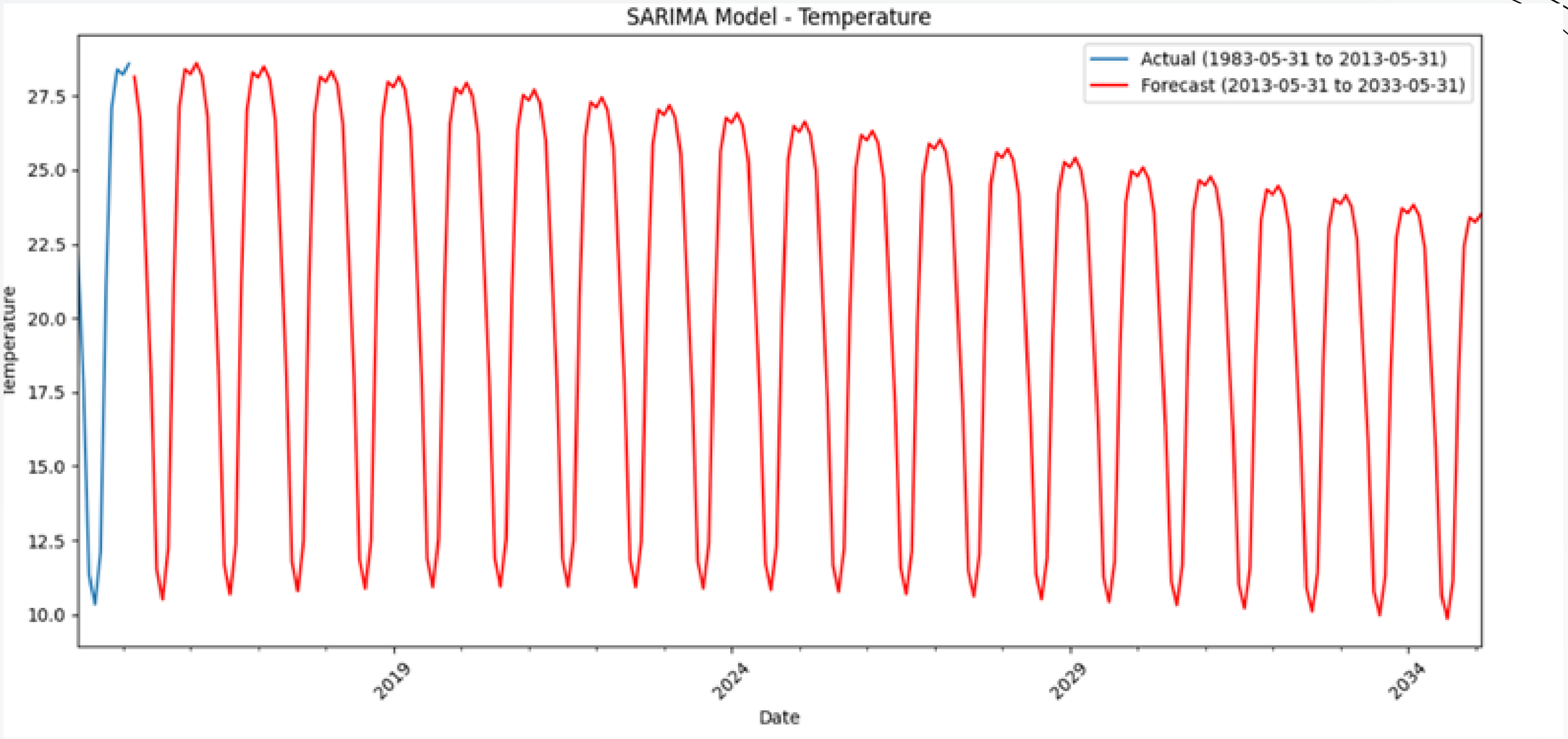
```
Tit   = total number of iterations
Tnf   = total number of function evaluations
Tnint = total number of segments explored during Cauchy searches
Skip  = number of BFGS updates skipped
Nact  = number of active bounds at final generalized Cauchy point
Projg = norm of the final projected gradient
F     = final function value

           * * *

   N    Tit     Tnf  Tnint  Skip  Nact     Projg        F
   9      0      1      0     0     0   0.000D+00  -0.000D+00
 F =  -0.0000000000000000

CONVERGENCE: NORM_OF_PROJECTED_GRADIENT_<=_PGTOL
Best SARIMA model: SARIMA(1, 1, 2)x(0, 0, 0, 12) - AIC: -2.69
```

- **Best SARIMA Model**: SARIMA(1, 1, 2)x(0, 0, 0, 12) – AIC: -2.69
  - Identified through exhaustive search of parameter combinations (p, d, q, P, D, Q, m)
  - AIC (Akaike Information Criterion) used as the optimization metric.

Forecasting with SARIMA

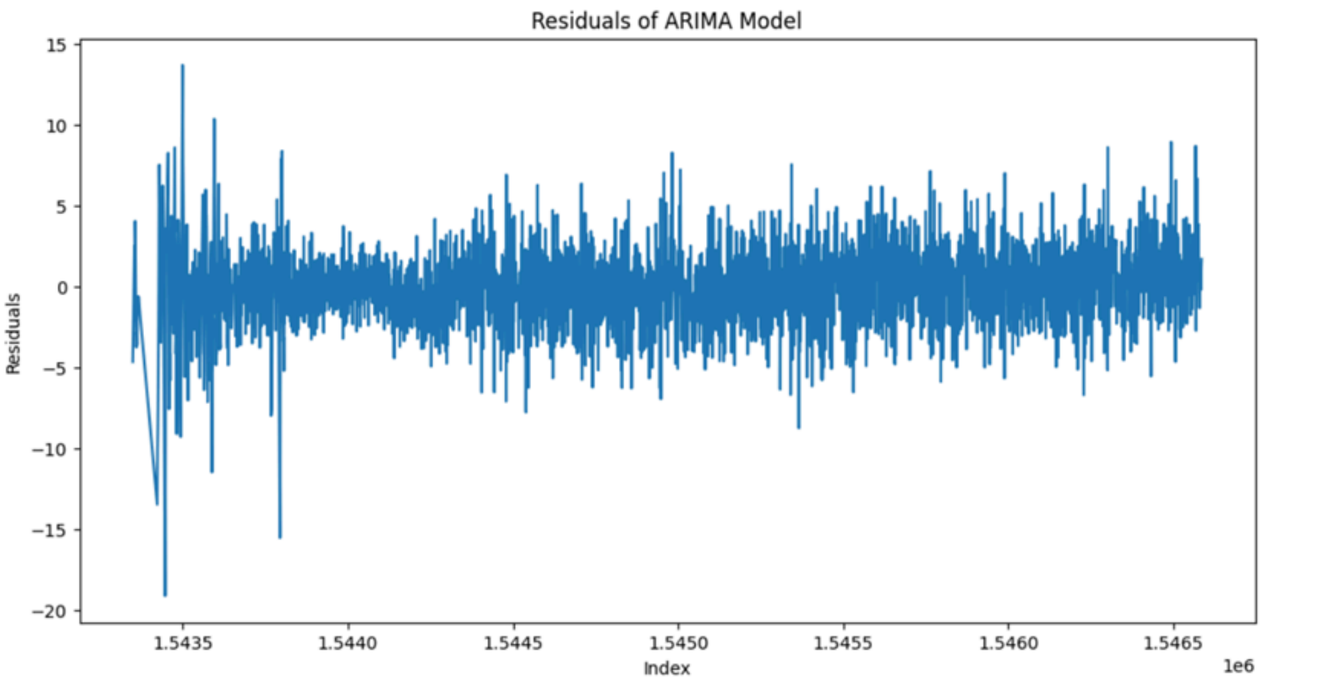


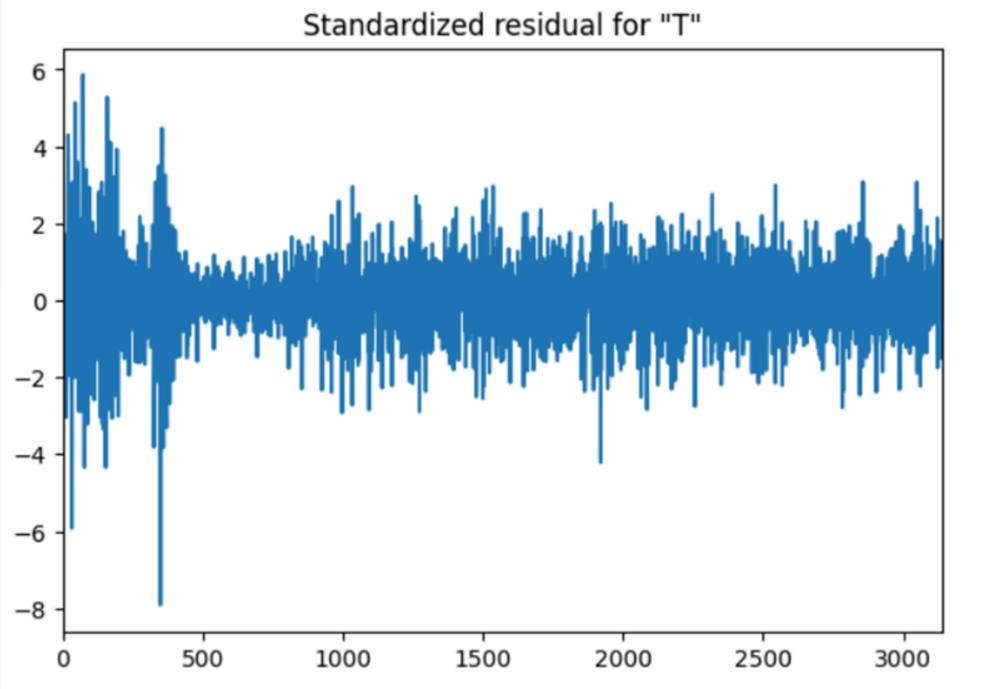
Forecasting with SARIMA



RESULTS AND DISCUSSIONS

Residual Analysis

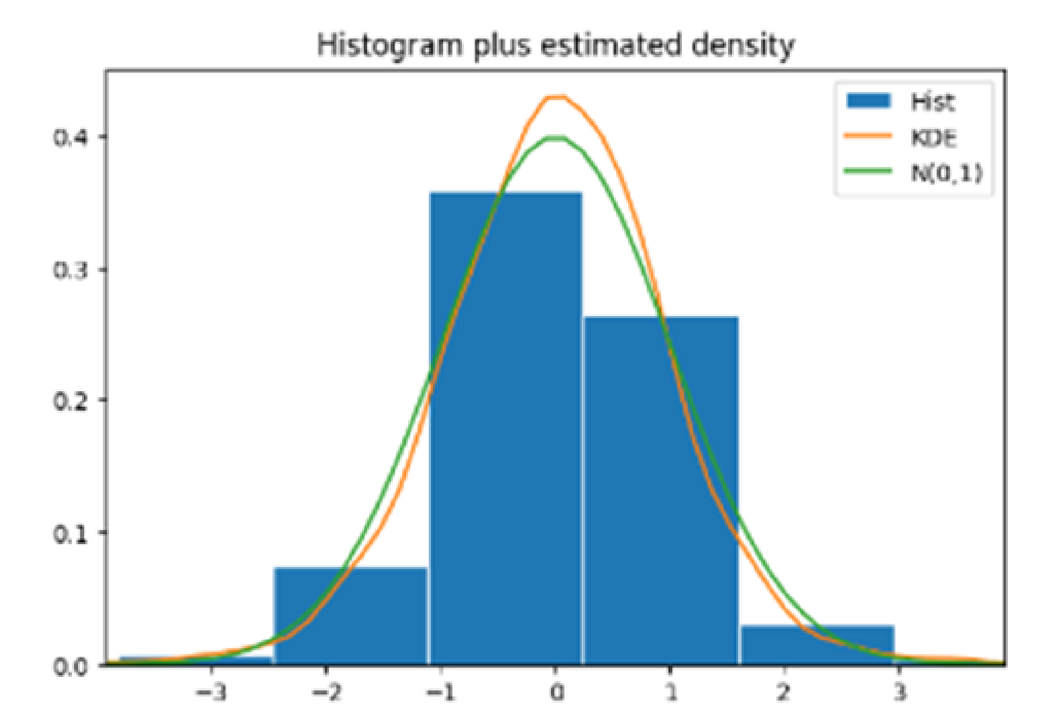


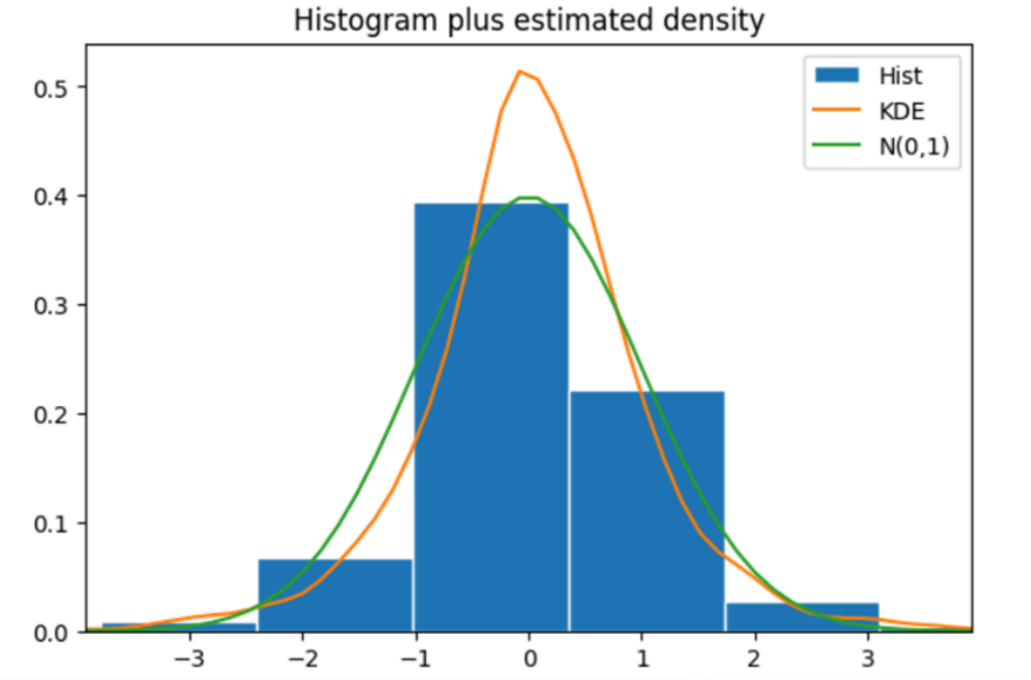


Here the residual plot follows the random process indicating that the model has captured the underlying patterns in the data.

RESULTS AND DISCUSSIONS

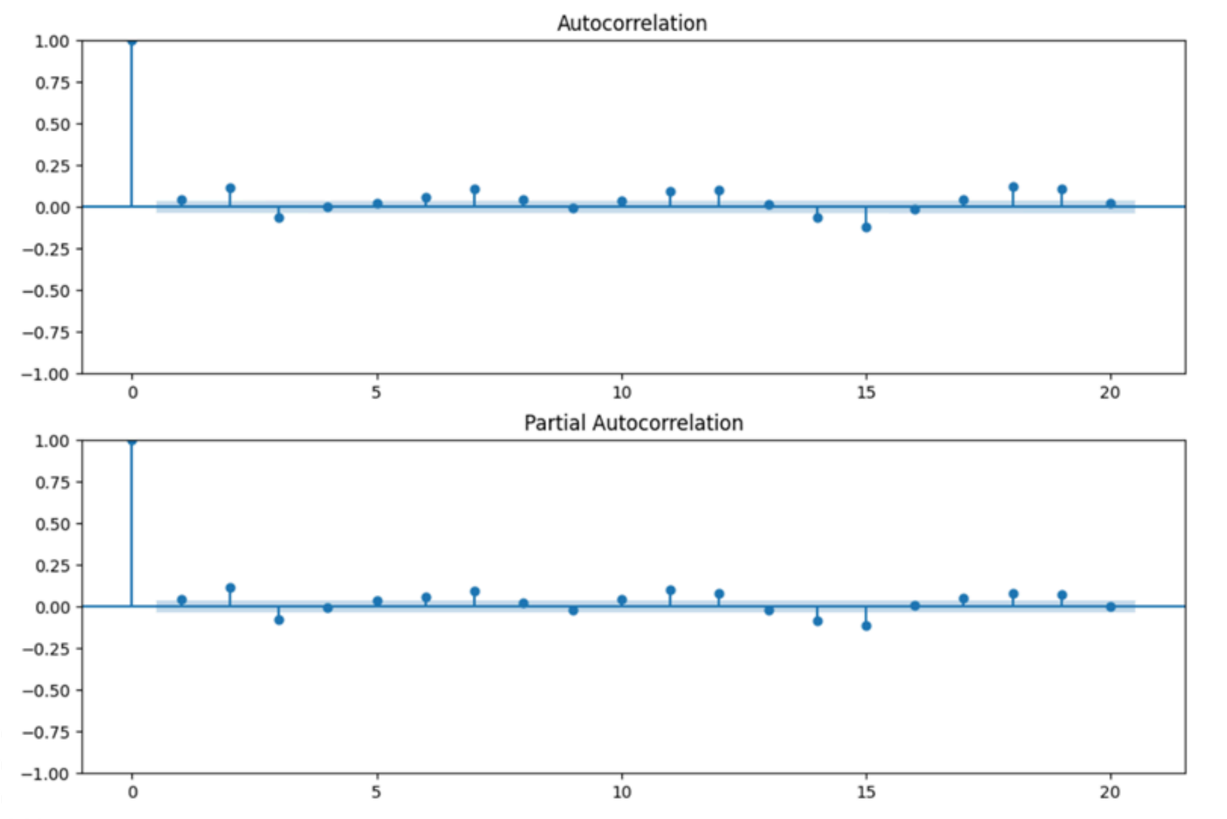
Histogram of Residuals

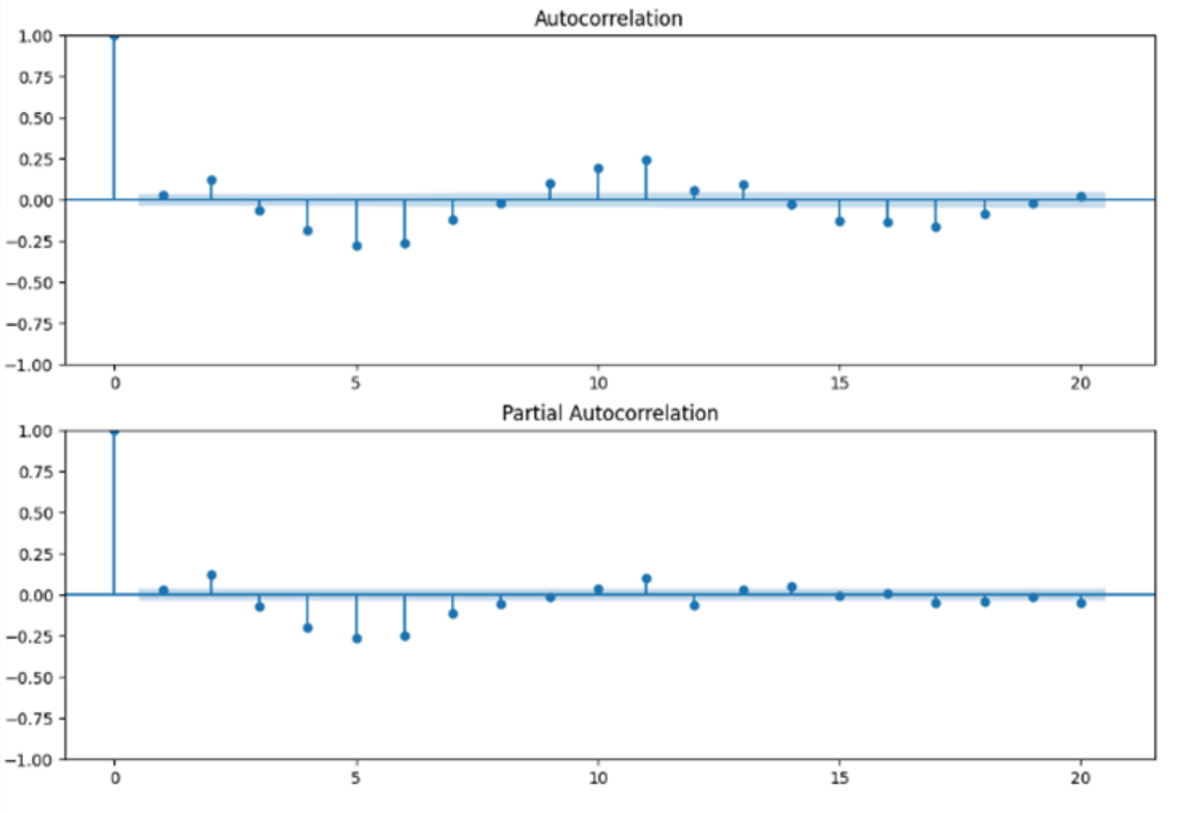




The normal distribution of residuals in the first model suggests overall accuracy, but right skewness indicates overestimation, This could indicate that the model is missing important features or that the model assumptions are violated. SARIMA's less right-skewed residuals imply a better fit than ARIMA.

Autocorrelation and Partial autocorrelation of residuals of ARIMA and SARIMA





We observe a tiny seasonal components in the ACF and PACF of residuals in ARIMA which suggests that there may be residual seasonal patterns in the data that are not captured by ARIMA model but in case of SARIMA there is no presence of seasonality.

PERFORMANCE METRICS

Mean Square Error (MSE)

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Root Mean Square Error (RMSE)

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \hat{x}_i\right)^2}{N}}$$

ARIMA VS SARIMA

	MSE	RMSE
ARIMA	123.55	11.115
SARIMA	48.33761	6.9525

CONCLUSION

- The project involved an extensive exploratory data analysis (EDA) that yielded valuable insights into the dataset.
- Trend and seasonality analysis revealed clear patterns in the data, highlighting the importance of accounting for these factors in the modeling process.
- Both ARIMA and SARIMA models were trained on the dataset, with SARIMA demonstrating superior performance over ARIMA.
- This was evident from the residual analysis, which showed that SARIMA had lower residual errors compared to ARIMA.
- Performance metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) further confirmed that SARIMA outperformed ARIMA in forecasting temperature.
- These results suggest that **SARIMA** is a more suitable model for forecasting temperature.

FUTURE ENHANCEMENTS

- Include external factors such as greenhouse gas emissions, solar activity, volcanic eruptions, and ocean currents in decision-making.
- Implement ensemble modeling techniques, such as combining forecasts from multiple ARIMA and SARIMA models.
- Incorporate long-term trends in Earth's surface temperature, such as global warming, into the forecasting models.

THANKYOU