Proposal for Final Project: Forecasting weather

Dataset Kaggle: https://www.kaggle.com/code/lonnieqin/jena-climate-prediction-with-lstm/input

1. Forecast Variable and Numerical Independent Variables

- **Forecast Variable**: The forecast variable is **Temperature (T (degC))**, which represents the temperature in degrees Celsius over time.
- **Numerical Independent Variables**:
    - **p (mbar)**: Atmospheric pressure.
    - **Tpot (K)**: Potential temperature.
    - **Tdew (degC)**: Dew point temperature.
    - **rh (%)**: Relative humidity.
    - **VPmax (mbar)**: Maximum vapor pressure.
    - **VPact (mbar)**: Actual vapor pressure.
    - **VPdef (mbar)**: Vapor pressure deficit.
    - **sh (g/kg)**: Specific humidity.
    - **H2OC (mmol/mol)**: Water vapor concentration.
    - **rho (g/m^3)**: Air density.
    - **wv (m/s)**: Wind velocity.
    - **max. wv (m/s)**: Maximum wind velocity.
    - **wd (deg)**: Wind direction.

2.

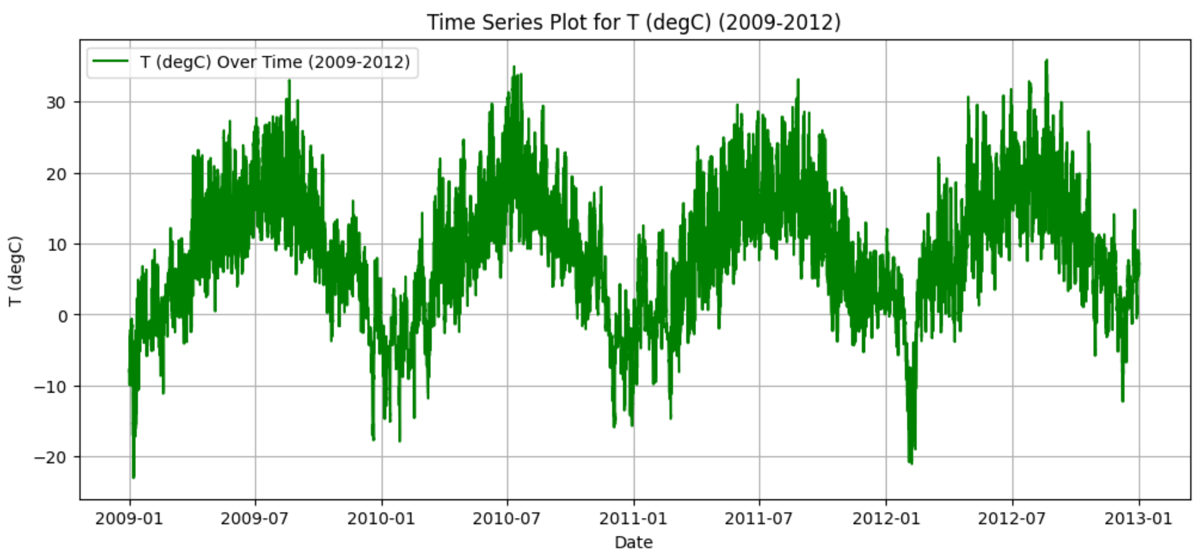


Fig-1 data vs time graph

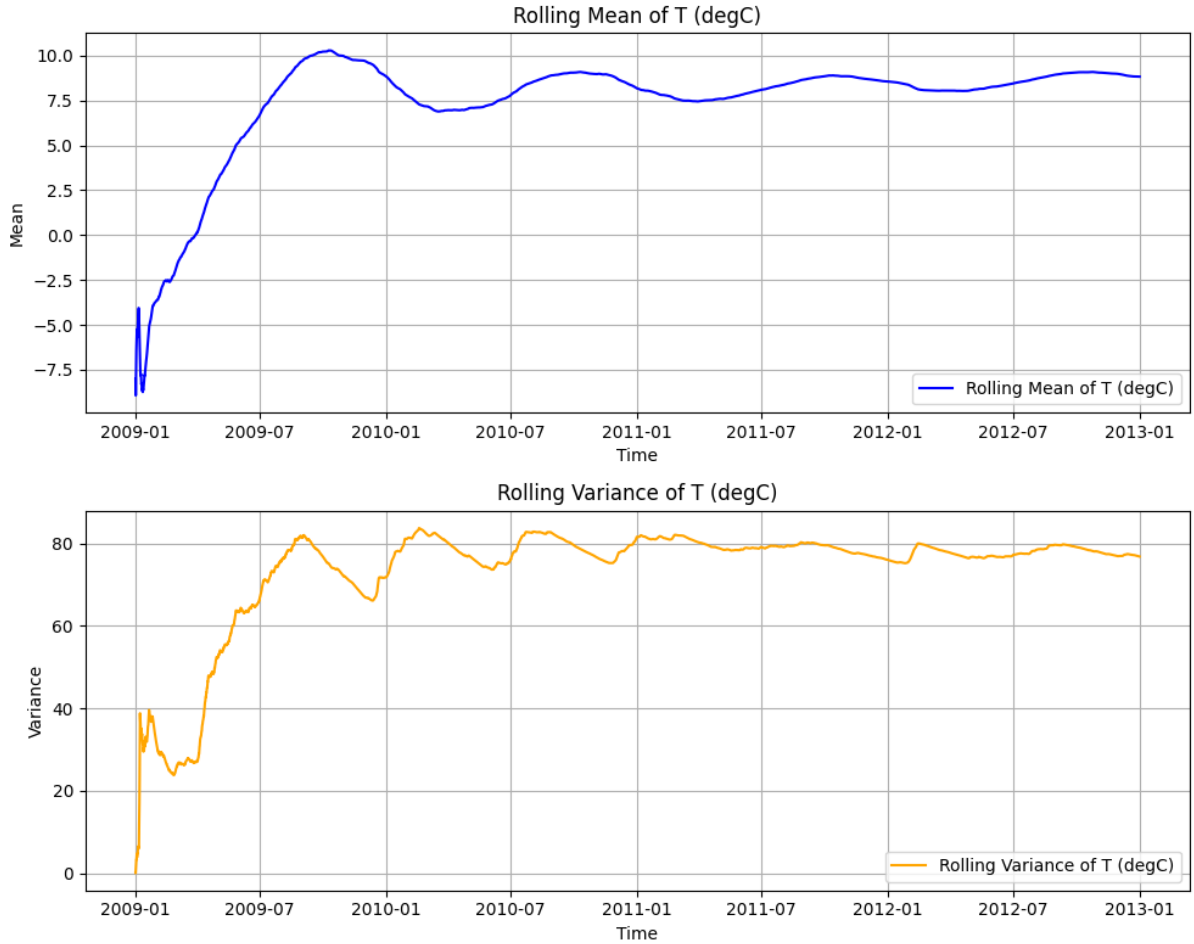


Fig-2 Rolling mean and variance show that data is not stationary

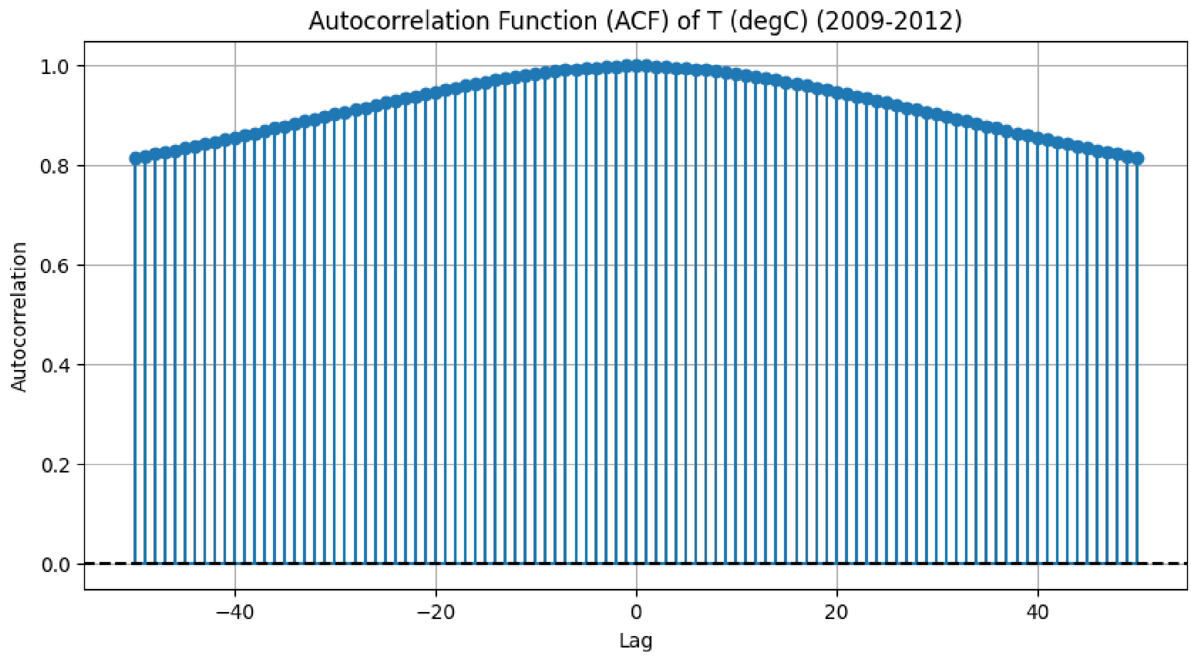


Fig-3 ACF plot shows there is a slow decay of the graph. Proving the data is not stationary

Tests: === Augmented Dickey-Fuller (ADF) Test ===

ADF Statistic: -8.4220

p-value: 0.0000

Critical Values:

1%: -3.4304

5%: -2.8616

10%: -2.5668

✅ The series is likely stationary (reject H0).

=== Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test ===

KPSS Statistic: 1.7739

p-value: 0.0100

Critical Values:

10%: 0.3470

5%: 0.4630

2.5%: 0.5740

1%: 0.7390

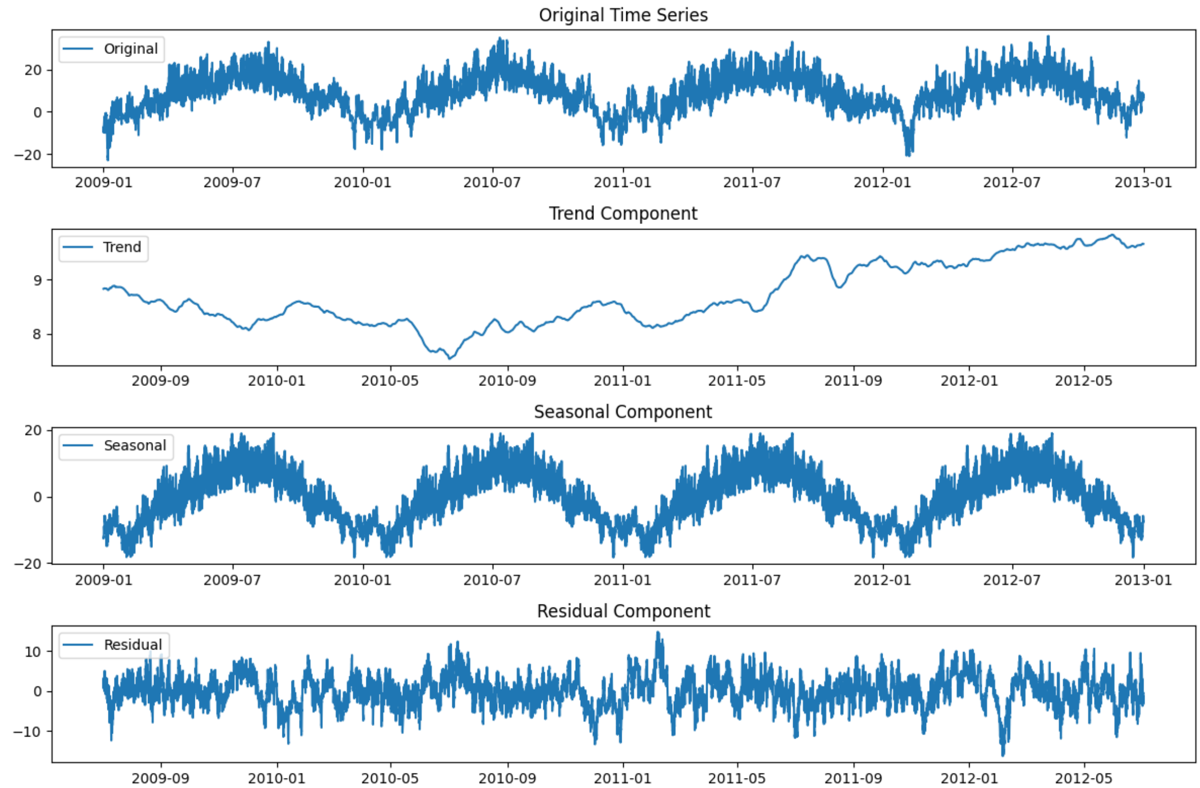
❌ The series is likely non-stationary (reject H0).

Although ADF shows stationary, Kpss and ACF show the data is not stationary this ambiguity due to the seasonality of the data as in Fig-1. The KPSS test, which is more sensitive to non-stationarity, supports this conclusion.

3.

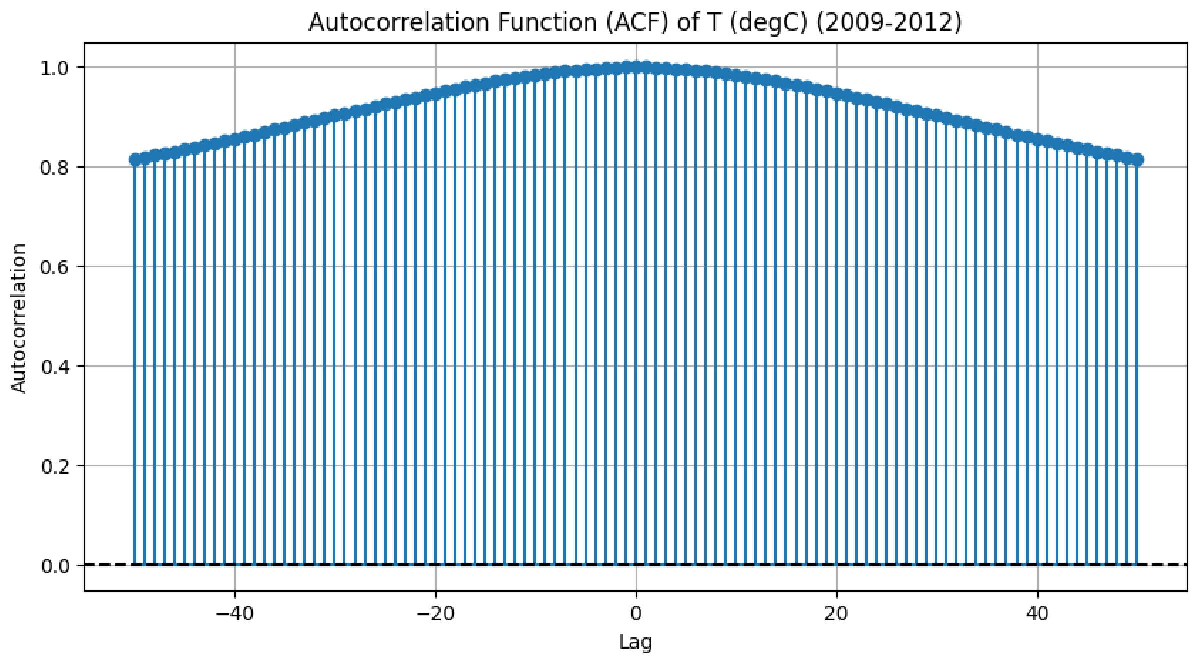
Strength of Trend: 0.0000

Strength of Seasonality: 0.8093



- The **no trend** (strength = 0.0) is not the primary cause of non-stationarity in the series.
- The **strong seasonality** (strength = 0.80635) does significantly contribute to non-stationarity

4.



1. Presence of Autocorrelation

- The ACF plot shows **significant autocorrelation** at multiple lags, indicating that past values of the temperature series are correlated with future values.
- This autocorrelation is a key assumption for using **autoregressive (AR)** models, which rely on the relationship between a value and its lagged values to make predictions.
- The gradual decay of the ACF suggests that the series has a **memory effect**, where past observations influence future ones. This is a strong indication that the series can be modeled using an AR process.

2. Identification of Lag Structure

- The ACF plot helps identify the **lag structure** of the series. For example:
  - Significant spikes at specific lags (e.g., lag 1, lag 2) indicate that these lags are particularly important for modeling the series.
  - The decay pattern in the ACF can help determine the **order of the AR model**.
- In the case of the temperature series, the ACF shows significant autocorrelation at the first few lags, suggesting that an **AR model** with a small order might be appropriate.

3. Detection of Seasonality

- The ACF plot can reveal **seasonal patterns** in the data. For example, if there are significant spikes at regular intervals yearly. this indicates the presence of seasonality.
- In the temperature series, the ACF shows periodic spikes, suggesting that the series has **seasonal components**. This makes the series suitable for **seasonal ARIMA (SARIMA)** models, which can capture both autocorrelation and seasonality.

The ACF of the temperature series is appropriate for modeling using the **Autocorrelation Approach Method** because:

1. It shows **significant autocorrelation**, which is a key assumption for AR models.
2. It helps identify the **lag structure** and determine the order of the AR model.
3. It reveals **seasonal patterns**, making the series suitable for SARIMA models.
4. It confirms **non-stationarity**, indicating the need for differencing or transformation.
5. It can be used for **model validation** by analyzing the residuals.

By leveraging the ACF, we can build a robust AR or SARIMA model to forecast temperature accurately.