TIME SERIES ANALYSIS

Popularity of Air Conditioners and Heaters

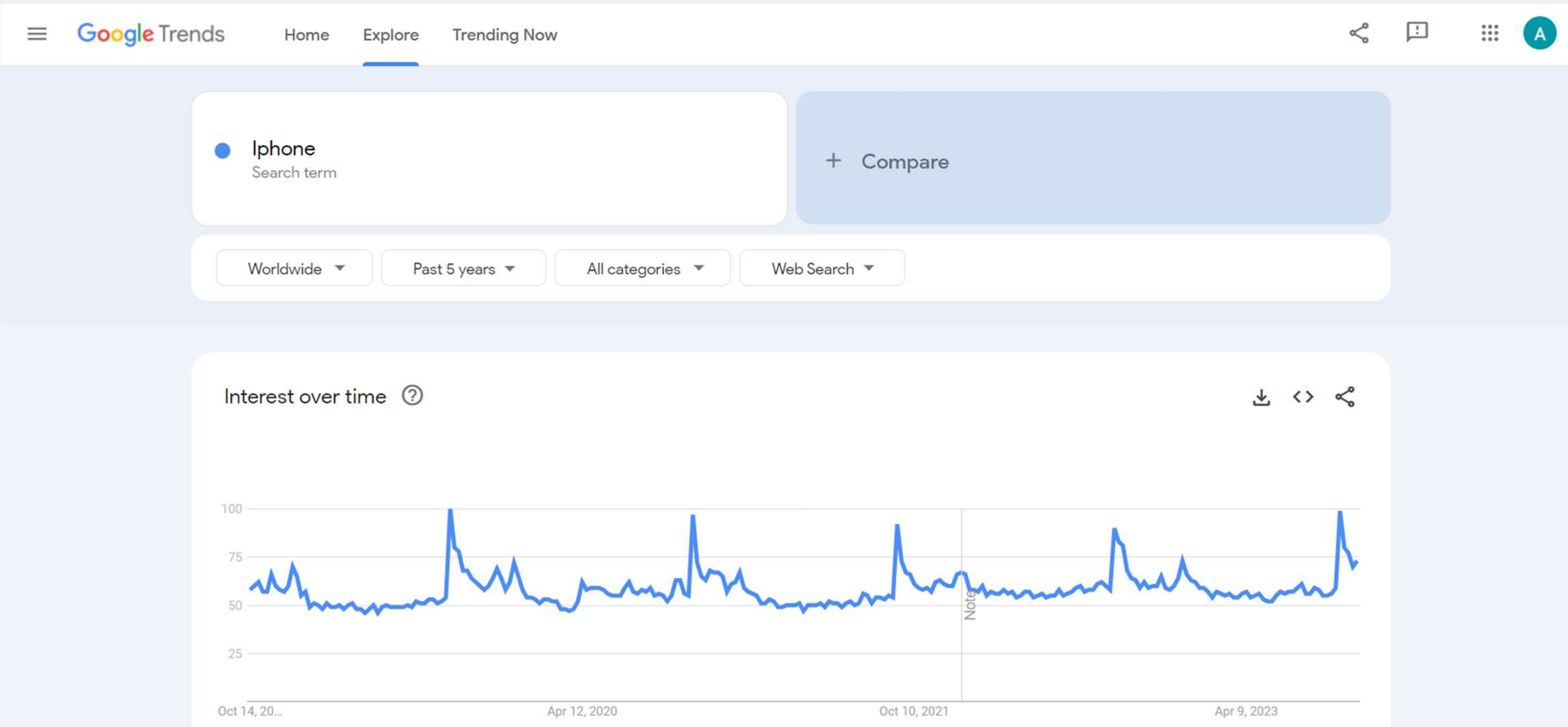
Dataset

Our dataset is made up of Google Trends data on the monthly world-wide popularity of air conditioners and heaters from January 2005 to October 2023. By popularity, we mean the number of web searches for these two items that occurred over the given time period.

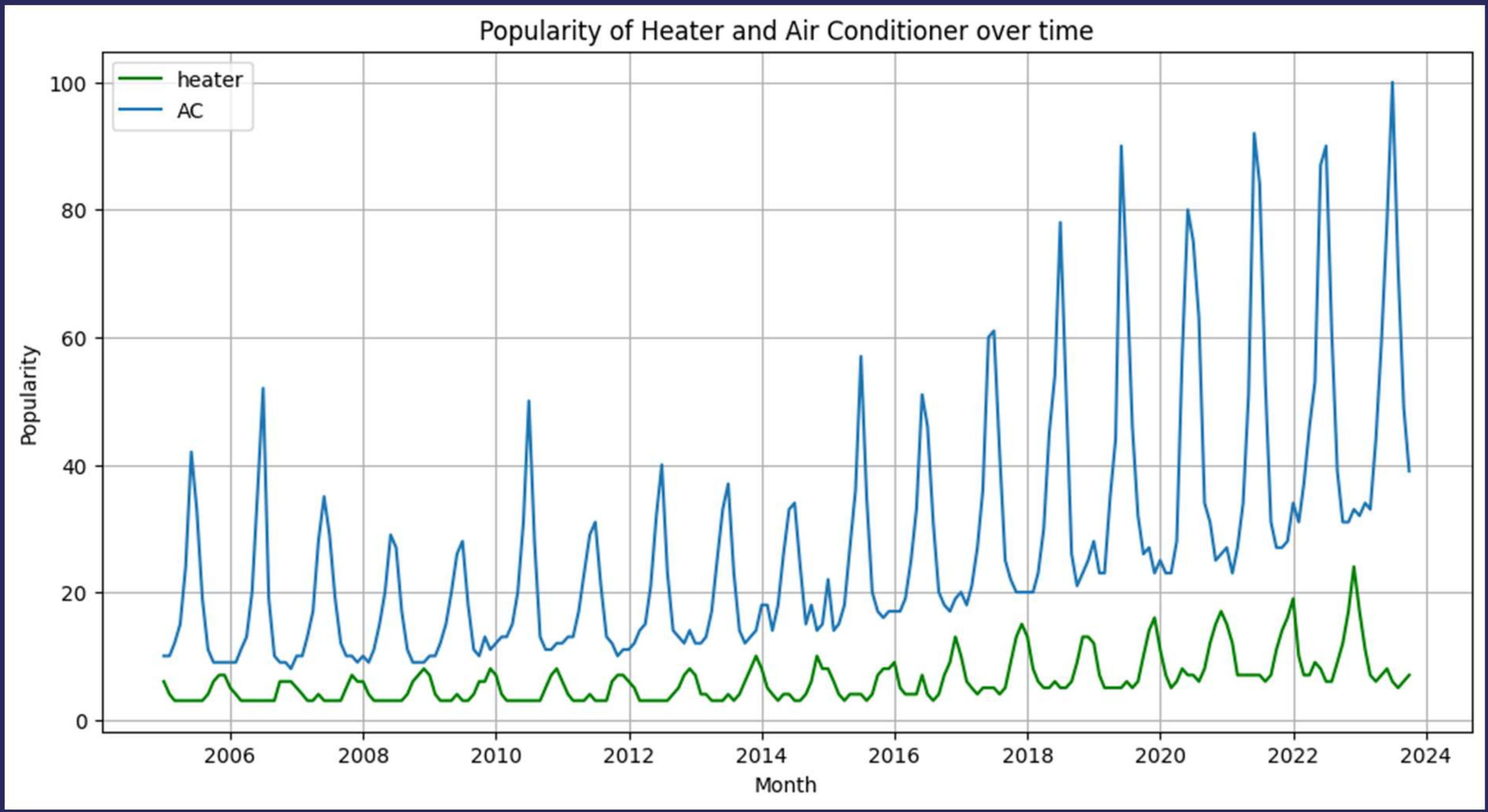
Google Trends

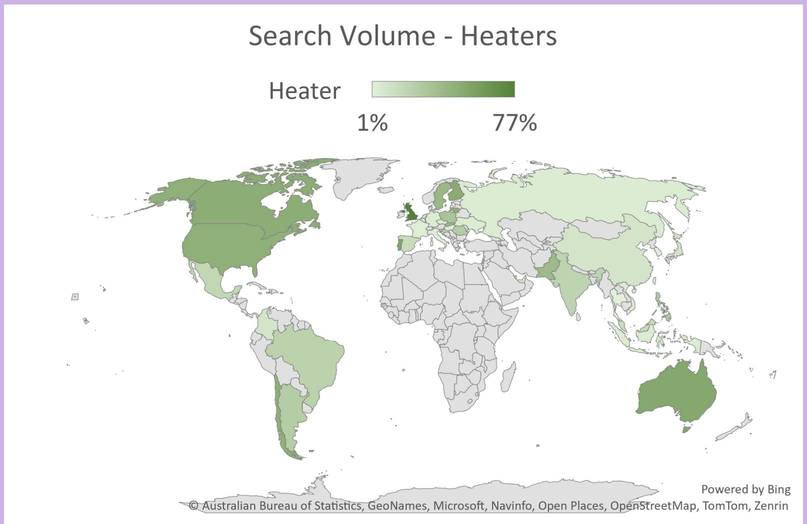
Google Trends is a useful search trends feature that displays how frequently a given search term is entered into Google's search engine in comparison to the site's total search volume over a specified time period. Numbers on the graph don't represent absolute search volume numbers, because the data is normalized and presented on a scale from 0-100

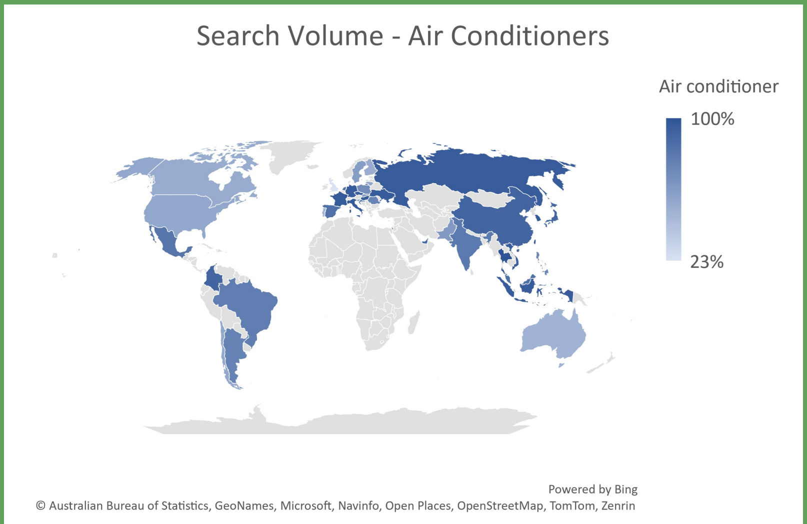
Google Trends



Visualization





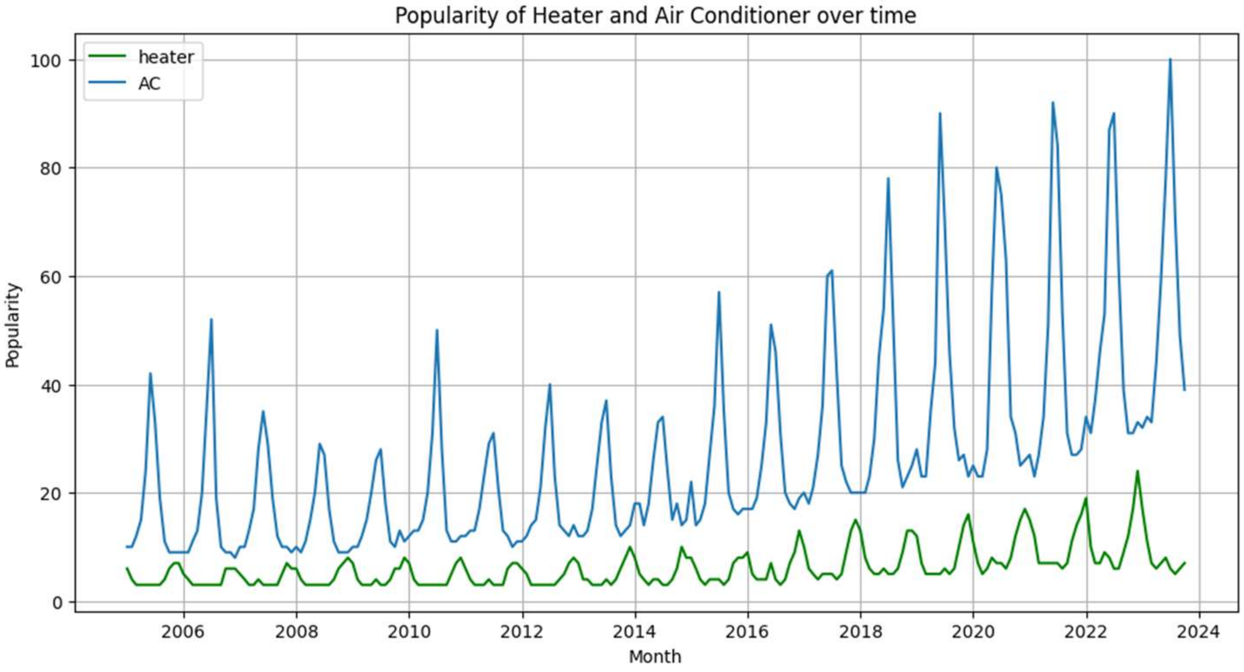


Popularity based on Country

Countries near the equator search more frequently about ACs while the farthest from equator search more about heaters. US, Canada, Australia generally have very cold winters and thus maybe heaters are more frequently searched than Acs

Motivation

A time series consists of various components like trends, seasonality, irregular fluctuations. Trends can result in a varying mean over time, whereas seasonality can result in a changing variance over time, both which define a time series as being non-stationary. Stationary datasets are those that have a stable mean and variance, and are in turn much easier to model.



Key Points:

Seasonal Influences: *Seasonal fluctuations significantly impact the popularity of "Heater" and "AC." These patterns may be reflective of changing climate conditions and consumer behaviors.*

Trend Identification: *We seek to identify any long-term trends that have emerged over the years, shedding light on the evolving consumer preferences for these appliances.*

Estimating Trend and Seasonality

1. Parametric Methods
2. Non-Parametric Methods

Parametric

Multiple Linear Regression

Model:

$X_t = a_0 + a_1 t + b_1 \sin(\pi t/6) + c_1 \cos(\pi t/6)$

- Here, $m_t = a_0 + a_1 t$ corresponds to the trend of popularity over the years
- And, $s_t = b_1 \sin(\pi t/6) + c_1 \cos(\pi t/6)$ corresponds to the seasonal effects of the popularity.

Estimation:

By the differentiating the residual sum of squares we can estimate the coefficients as

$$\boldsymbol{\beta} = (B^T B)^{-1} B^T X$$

Where, $\boldsymbol{\beta} = [a_0, a_1, b_1, c_1]^T$

$B_i = [\ 1, i, \cos(\pi i/6)\ , \sin(\pi i/6)\ ]$ for $i = 1,2,\ldots,n$

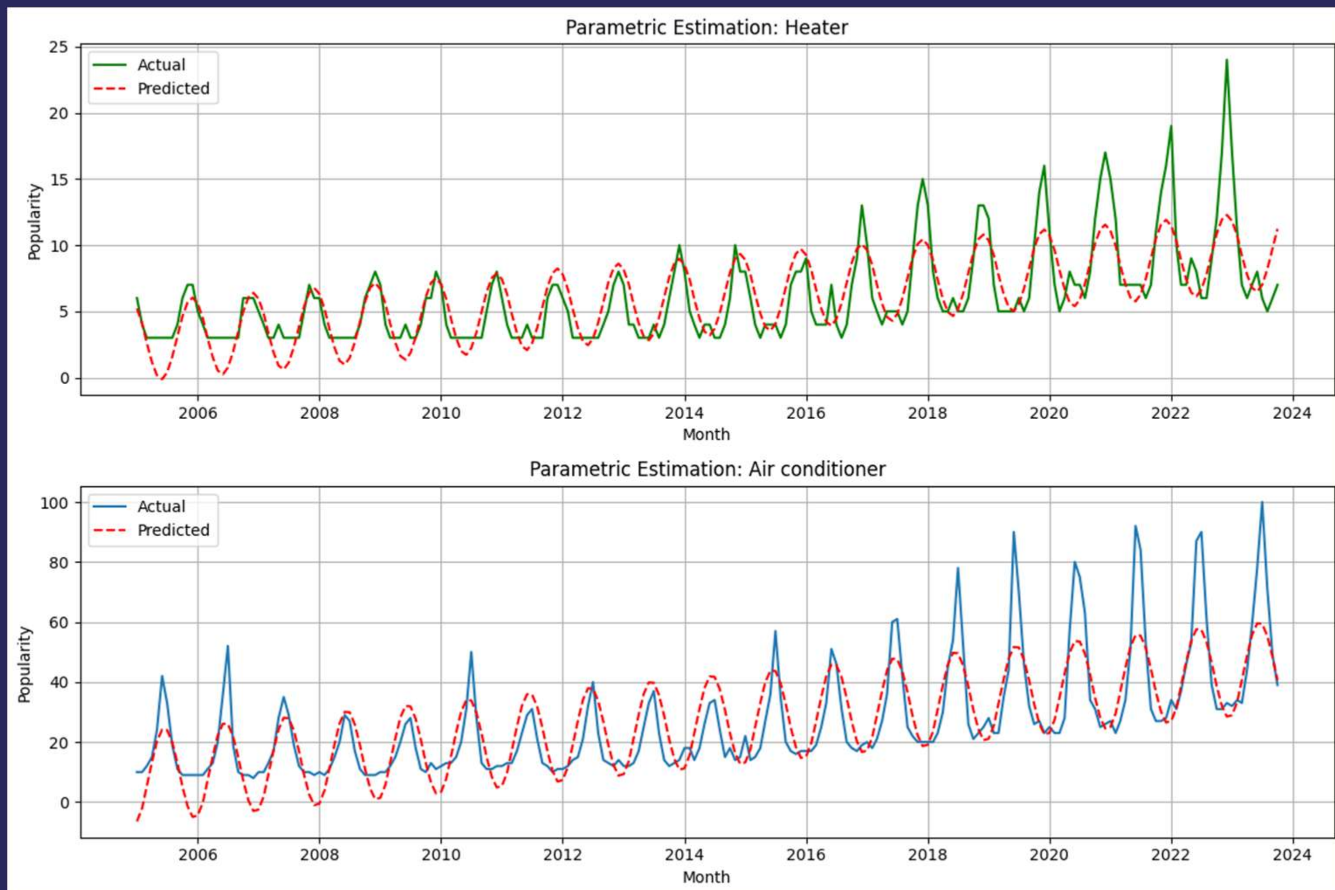
```
# Preparing the design matrix
m = np.array([np.ones(len(df)), (df.index.values+1)]).T
s = np.array([np.sin(np.pi*(df.index.values+1)/q), np.cos(np.pi*(df.index.values+1)/q)]).T
m_s = np.hstack((m, s))

coefficients_heater = np.linalg.inv(m_s.T @ m_s) @ m_s.T @ df['Heater']
coefficients_AC = np.linalg.inv(m_s.T @ m_s) @ m_s.T @ df['Air conditioner']

# Predicting the values
y_hat_heater = m_s @ coefficients_heater
y_hat_AC = m_s @ coefficients_AC
```

Parametric

Multiple Linear Regression

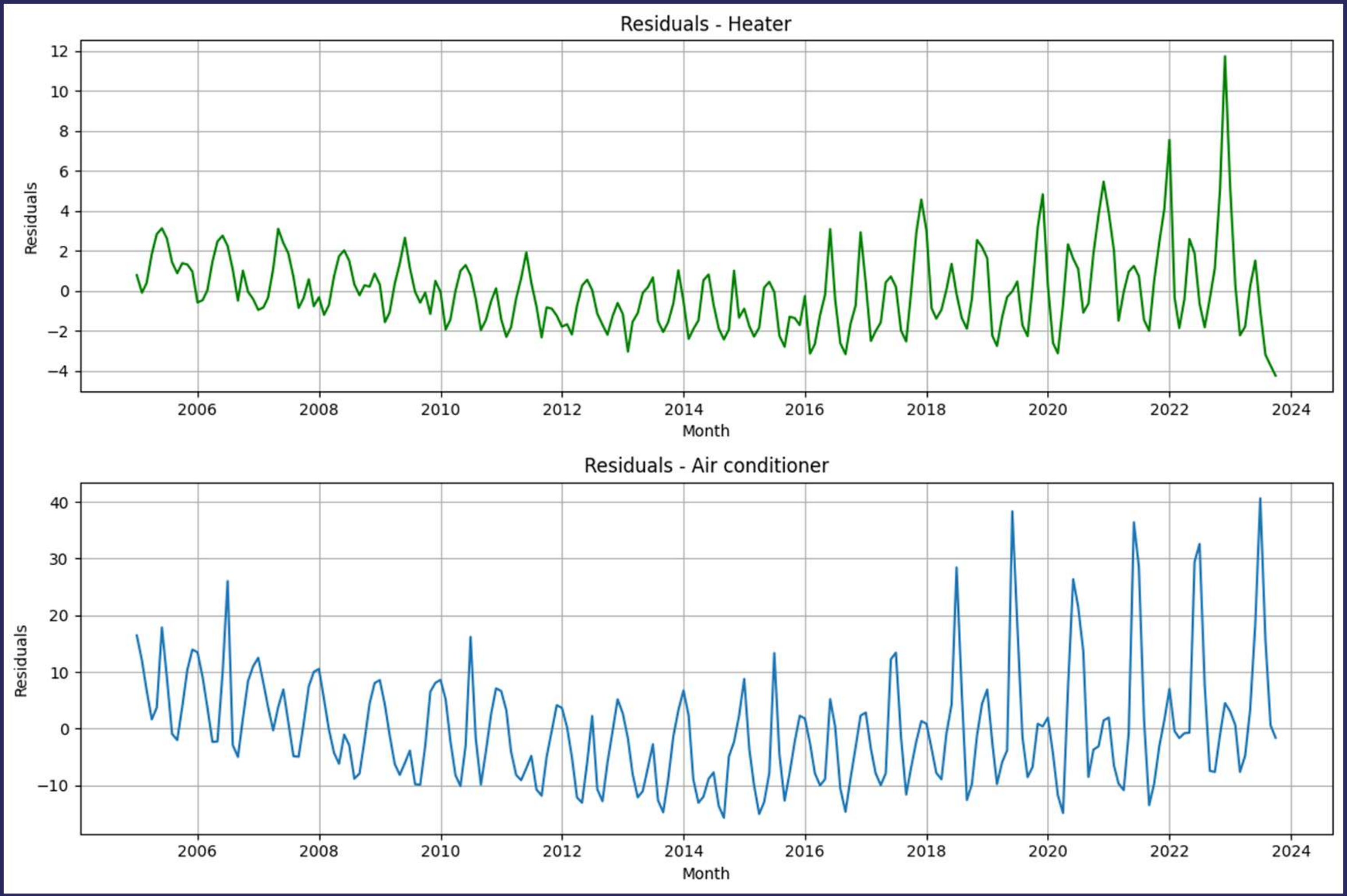


Coefficients:
β = [2.68 0.03 -0.16 2.99]
RMSE: 2.01

Coefficients:
β = [8.12 0.16 -3.4 -15.04]
RMSE: 10.08

Parametric

Multiple Linear Regression



Non-Parametric Estimation

Moving Average Filter: Estimate Trend

To estimate the trend of the data, moving average filter is applied. The trend component m_t is estimated as:

$$\mathbf{m_t = (0.5 * x_{t-q} + x_{t-q-1} + \ldots\ldots + x_{t+q-1} + 0.5 * x_{t+q})/d}$$

where , $d = 2q$ and $q < t \leq n - q$.

Since we are analyzing monthly data across the years, **d =12** and hence **q = 6**.

Moving Average Filter: Estimate Seasonality

We estimate the seasonal component using the trend obtained from moving average filter. We calculate the average of each month across the years

$$\mathbf{w_k = average(\, x_{k+jd} - m_{k+jd}\,)}$$

where $k=1,2,\ldots,d$ and $q < k + jd \leq n - q$

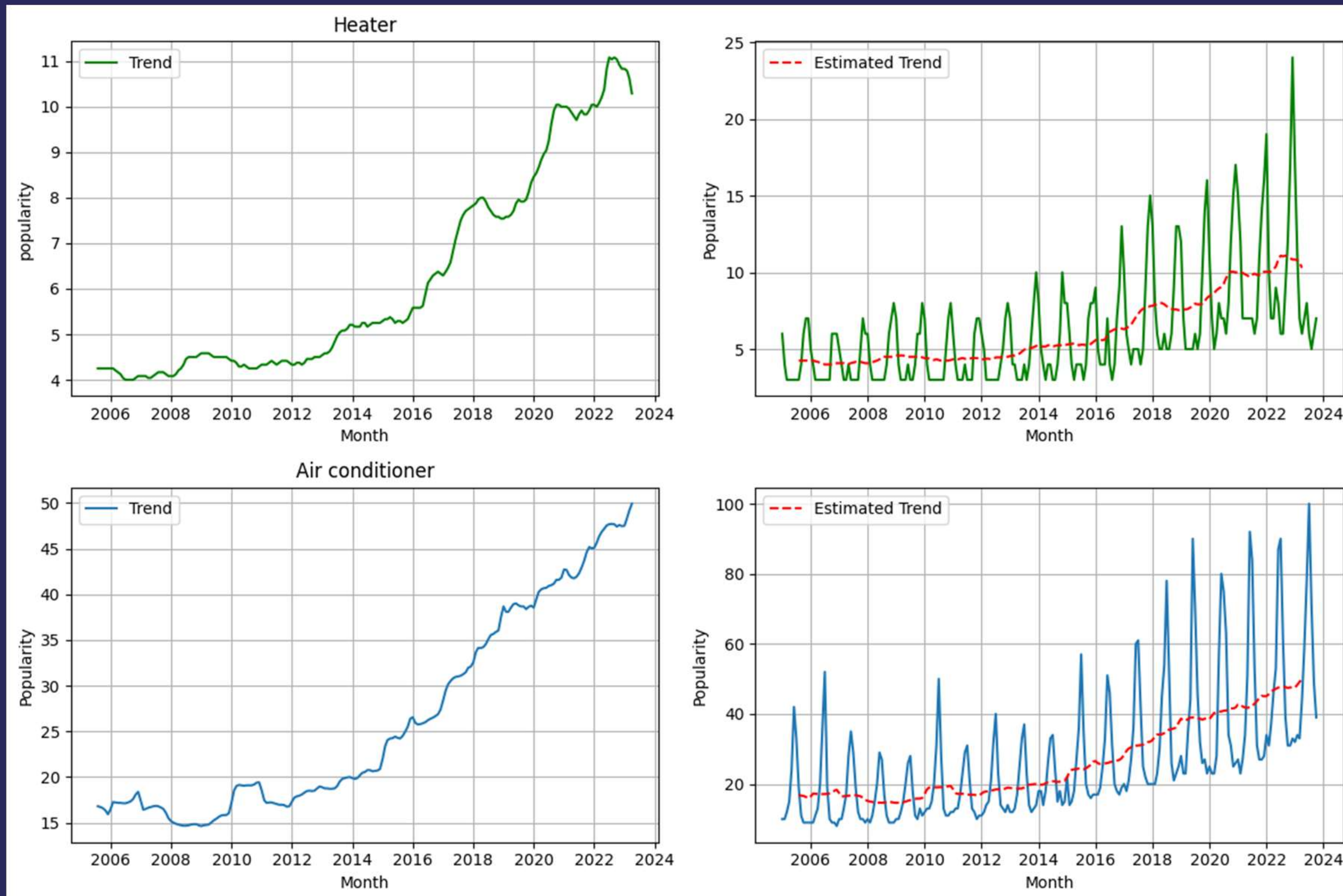
Now, estimating the seasonal component as

$$\mathbf{s_k = w_k - (\Sigma_{i=1}^{k} w_i)/d}$$

where $k=1,2,\ldots,d$

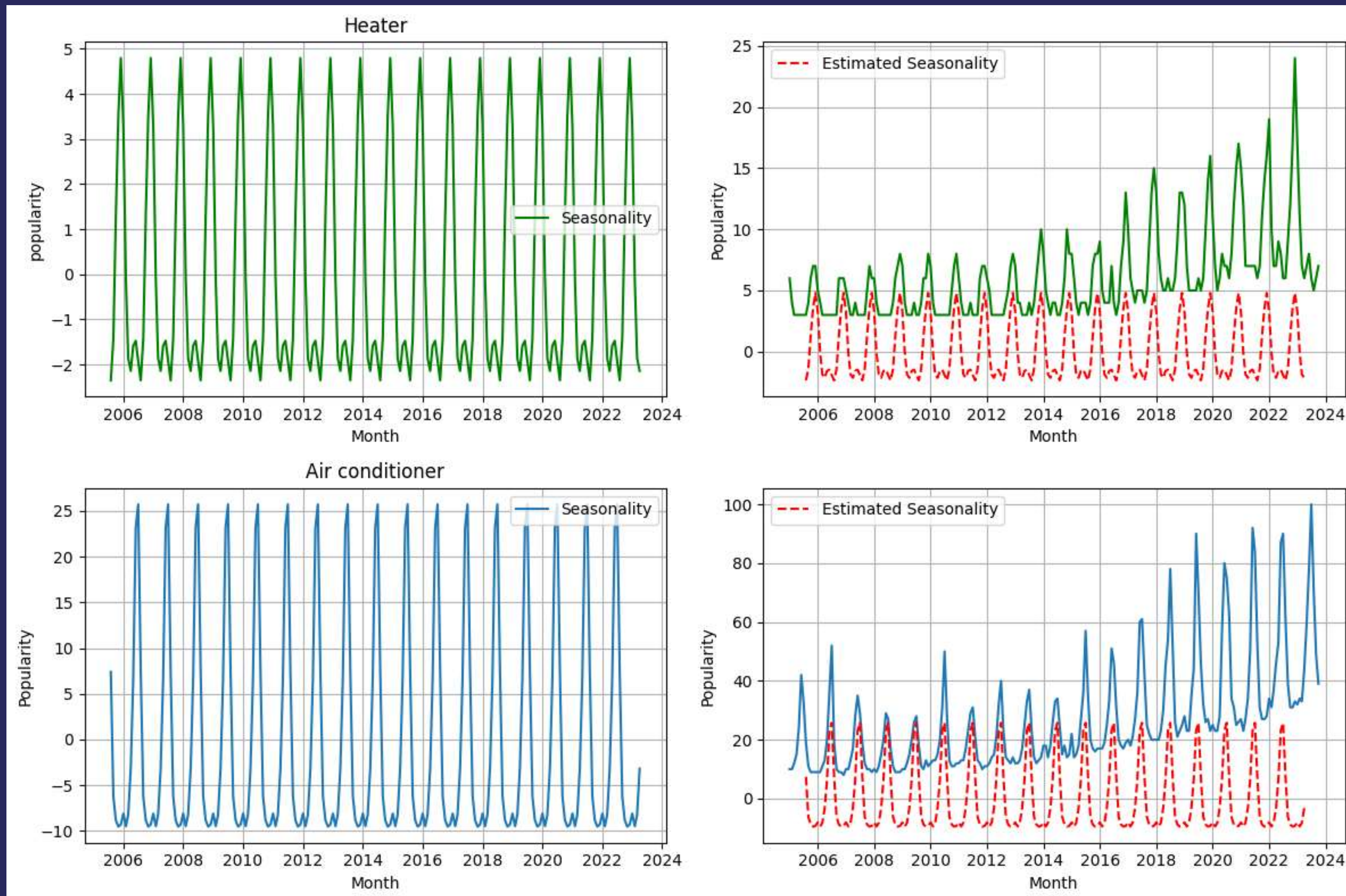
Moving Average Filter

Trend Estimation

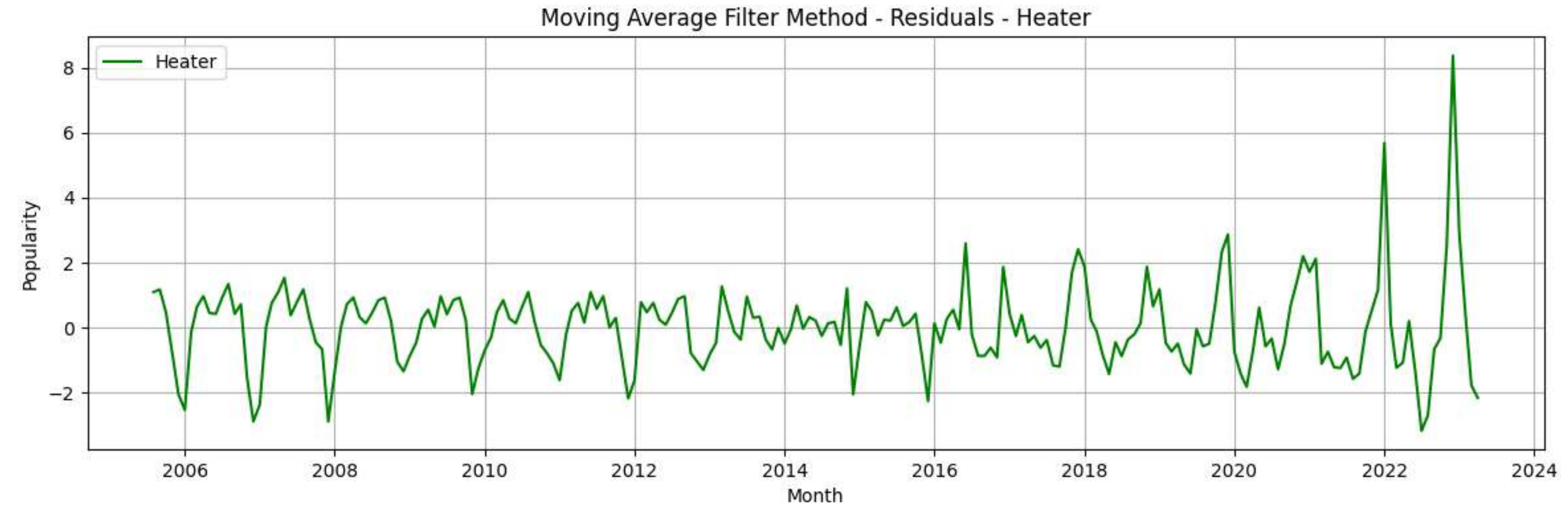


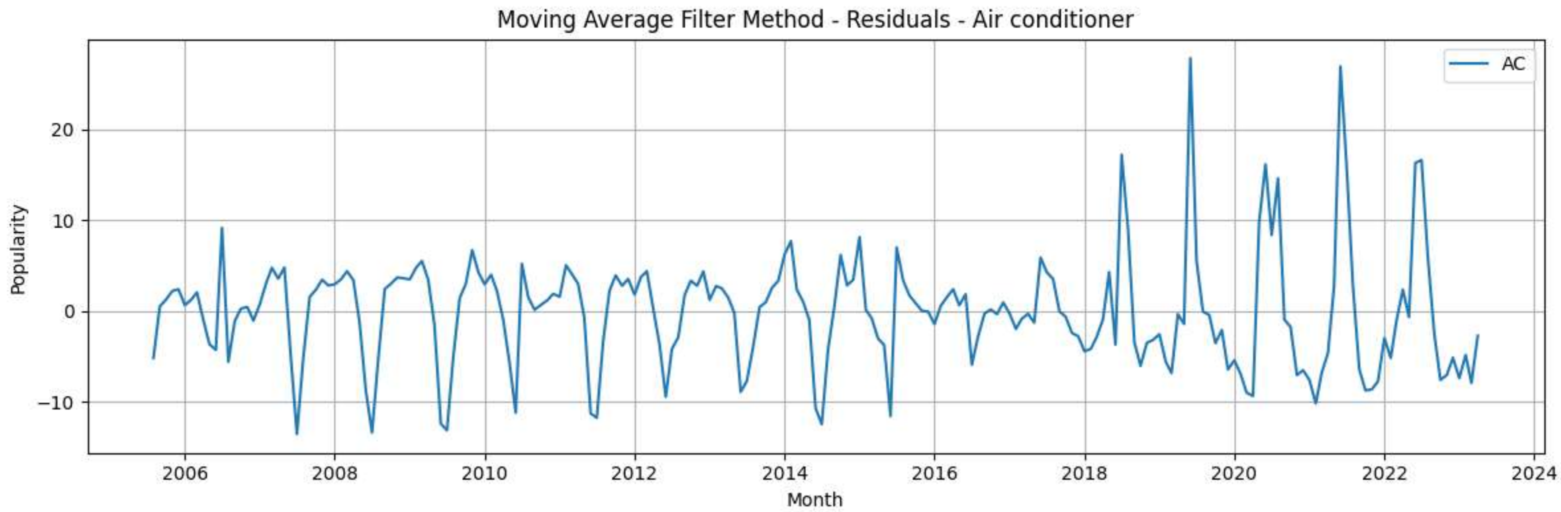
Moving Average Filter

Seasonality Estimation



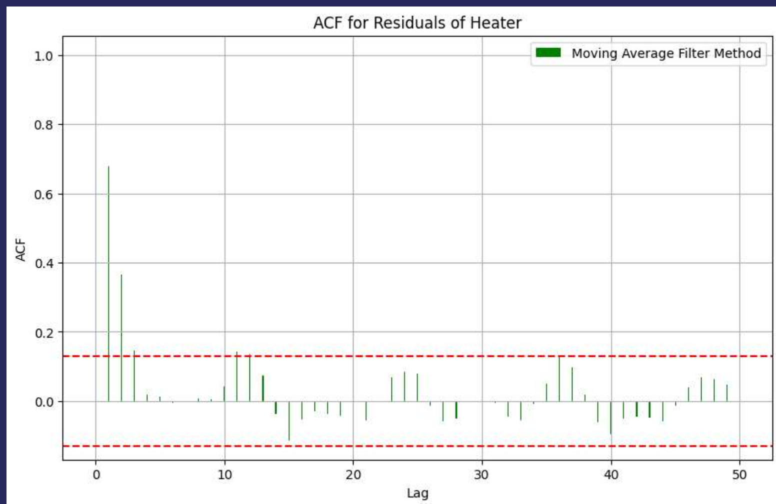
Moving Average Filter Residuals



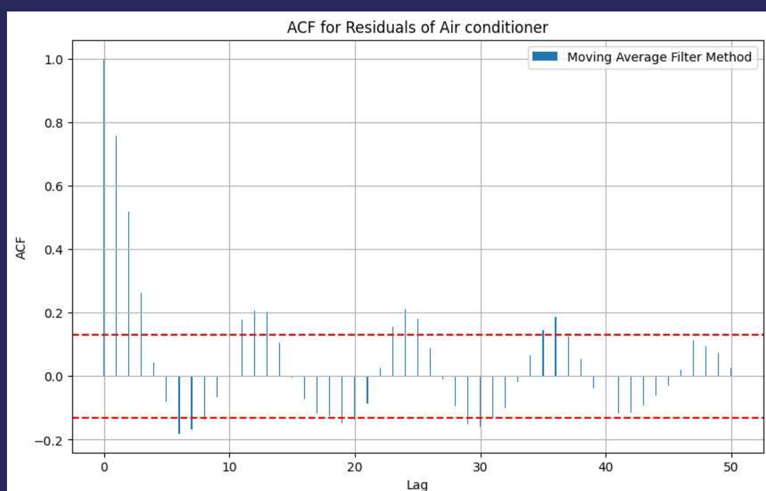


Moving Average Filter

Autocorrelation Function



Interpretation:
ACF at h = 1, 2 are the only significant lags while the rest are zero , suggesting that the underlying process is stationary and is most likely to be a moving average of order 2.



Interpretation:
ACF is gradually decreasing suggesting our process could be Auto regressive process however its seasonally repeating after around 5 lags, it suggests there might be some seasonality.

Non-Parametric Estimation

Method of Differencing to remove trend and Seasonality

Instead of estimating trend, seasonality and subtracting them from original, we can use the differencing operator to directly estimate the residual of a time series.

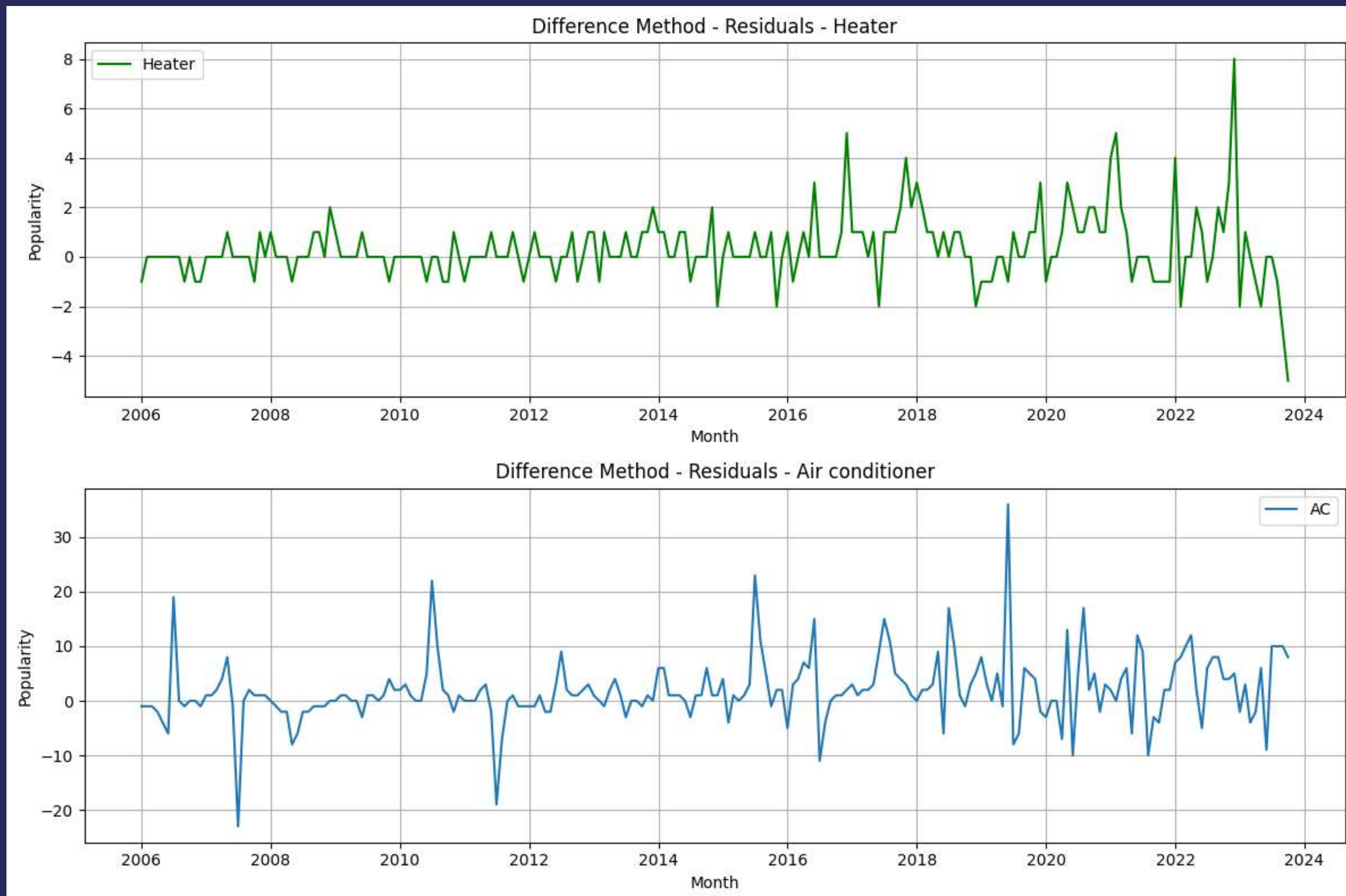
Residual can be estimated as:

$$\nabla_d X_t = (m_t - m_{t-d}) + (Y_t - Y_{t-d}) = a_1 d + (Y_t - Y_{t-d})$$

Where,

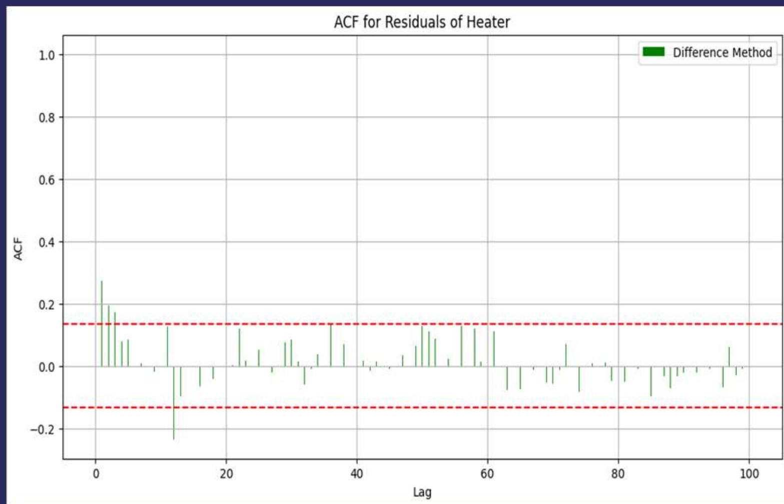
$X_t = m_t + s_t + Y_t$, where $m_t = a_0 + a_1 t$ (trend component), s_t is the seasonal component and Y_t is the residual term.

Method of Differencing Residuals

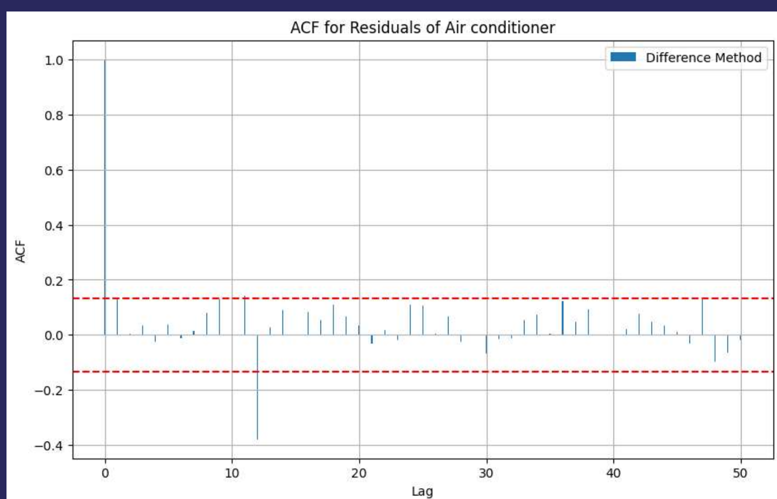


Method of Differencing

Autocorrelation Function

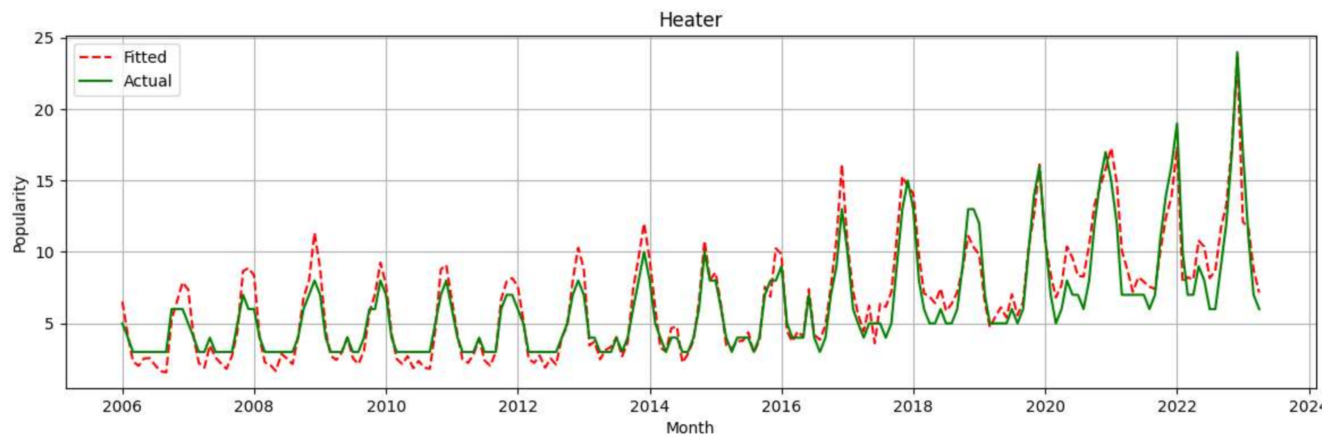


Interpretation:
Process is most likely stationary. The underlying process can be a moving average process of order 2 or 3 as lag at approximately 3 points seems to be significantly not equal to 0.

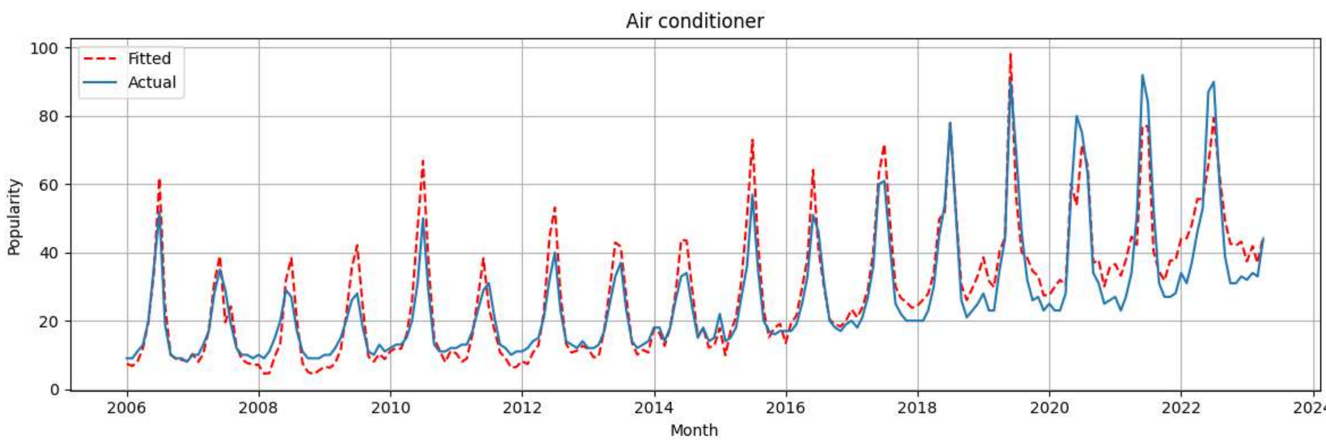


Interpretation:
Process is most likely stationary. There is a lag significantly not equal to 0, the process is likely to be Moving Average of order 1

Fitting

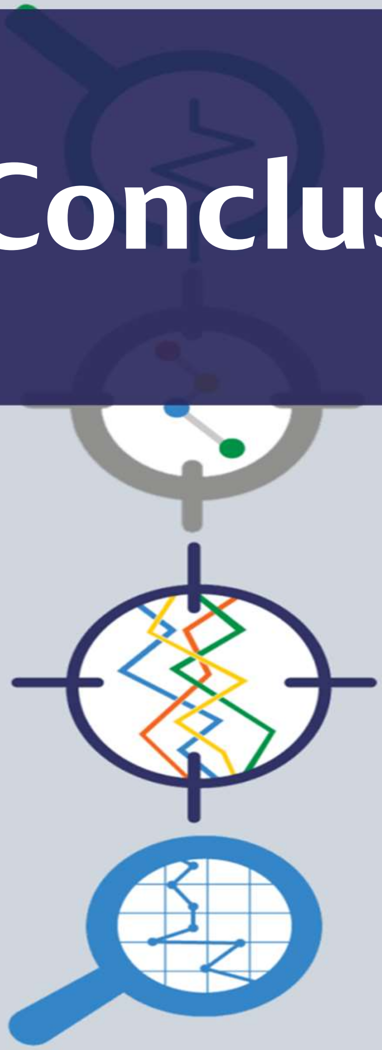


RMSE: 1.29
Adjusted R2: 0.87



RMSE: 6.55
Adjusted R2: 0.85

Conclusion



We estimate the time series for the heater and air conditioner dataset using parametric and non-parametric methods. The parametric approach is unable to model real-world data effectively. It fails to capture the increasing seasonal fluctuations for both datasets. Using the non-parametric approach to estimate the time series for the heater and air conditioner dataset, we obtained more than 85% explanation in variance. Differencing seems to be the superior method of estimation in this case as the auto-correlogram shows that the noise estimated by differencing has almost no cyclic fluctuations in the non-significant values. The underlying noise is stationary, we can estimate the trend and seasonal fluctuations in any time series and effectively evaluate that series.

Thank you

Done by: Group-4

1. Anita Dash
2. Karthik Kurugodu
3. Tanay Yadav
4. N.Dilip Kumar Reddy