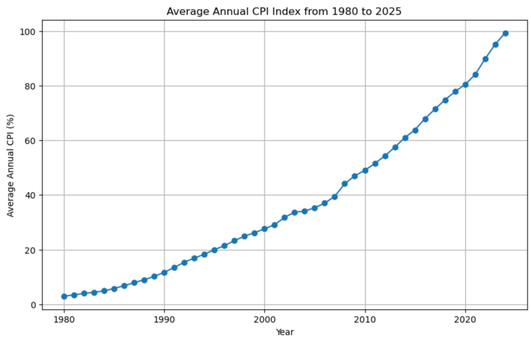
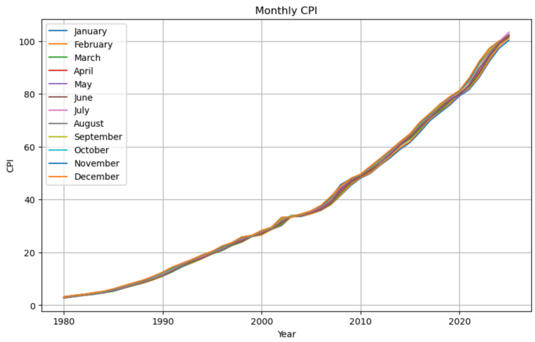
CPI Index Time Series Analysis and Forecasting

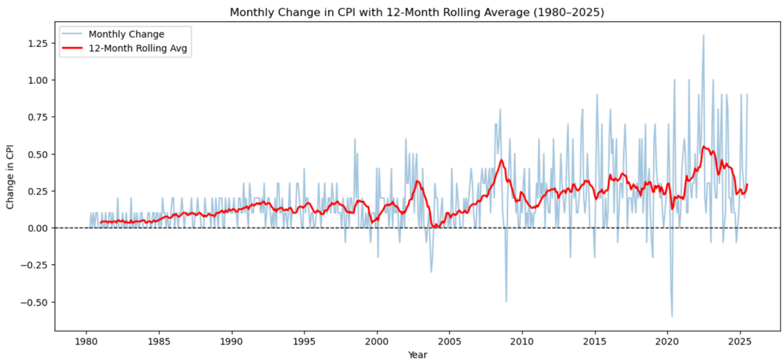
The aim of this project is to analyse South African CPI Index data using time series methods as well as generate forecasts of future values. The models applied include SARIMAX and Prophet and their performance is compared to determine which provides better forecasting accuracy.

The data used in the analysis was obtained from StatsSA which is publicly available on the website in a pdf format. It consists of monthly CPI indices from 1980 to 2025 for each month. There is also an additional column of annual CPI averages.

The data was imported and converted to CSV format. Steps were taken to ensure correct formatting of headings and missing values were removed. The data was also converted to float type and decimal commas were replaced with a decimal point. For subsequent analysis, the data is transformed from wide to long format for time series analysis and forecasting.







The visualisations above present trends in CPI from 1980 to 2025. The top-left graph shows the average annual CPI index which exhibits a steady upward trajectory over the years, highlighting long-term inflationary growth. The top-right graph illustrates the monthly CPI where all months follow a similar increasing pattern, suggesting consistent inflationary pressures across the calendar year. The bottom graph displays monthly changes in CPI along with a 12-month rolling average, providing insights into short-term fluctuations and longer-term trends. Collectively, these figures demonstrate the persistent rise in CPI over time, alongside periods of volatility that may reflect economic shocks or policy interventions.

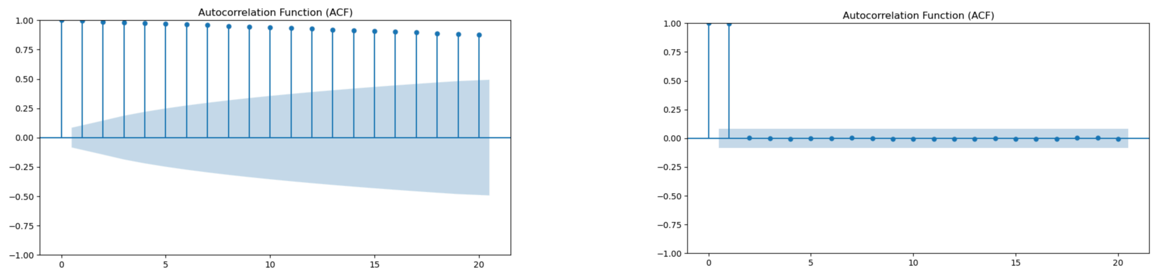
To assess whether the CPI indices are stationary, the Augmented Dickey-Fuller (ADF) test was applied. The ADF test evaluates the null hypothesis that a unit root is present in the time series, meaning the series is non-stationary. If the p-value is greater than the chosen significance level, we fail to reject the null hypothesis indicating non-stationarity. Conversely, a p-value below 0.05 suggests stationarity.

H_0 : Time series is non-stationary.
H_A : Time series is stationary.

Results of the ADF test are as follows:

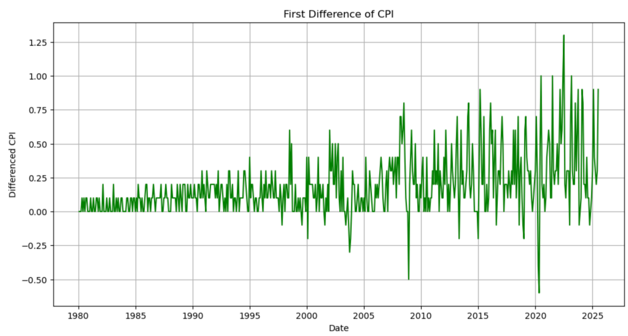
```
ADF Statistic: 4.419182332809767
P-value: 1.0
Critical Values:
        1%: -3.4427957890025533
        5%: -2.867029512430173
        10%: -2.5696937122646926
```

Since the p-value is greater than the chosen significance level of 5%, we fail to reject the null hypothesis. Thus the time series is non-stationary. The following graphs below represent the auto-correlation function (ACF) and partial auto-correlation function (PACF). The ACF graph confirms the ADF test for non-stationarity as it decays slowly and remains significant for many lags.



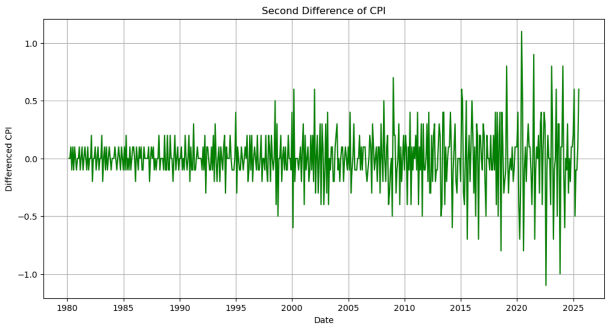
Stationarity is crucial for time series analysis as it facilitates modelling and forecasting of data. Properties of stationarity include constant mean, variance and the lack of significant trends or seasonality. Differencing is a statistical technique in time series analysis that involves subtracting the previous observation from the current one to transform a non-stationary time series into a stationary one.

To obtain stationarity we difference the CPI indices data and run an ADF test to assess for stationarity.



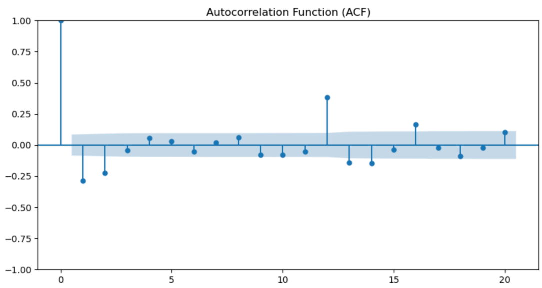

```
ADF Statistic: -2.6789473000161075
P-value: 0.07775269149931198
Critical Values: {'1%': -3.442772146350605, '5%': -2.8670191055991836, '10%': -2.5696881663873414}
```

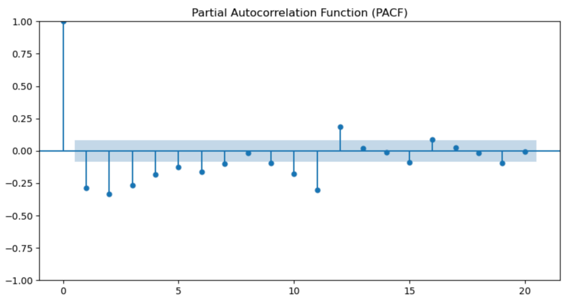
The p-value remains greater than the significance level of 5% when differencing once, indicating non-stationarity, which is also confirmed by the graph above as it is not centered around 0 and there is an obvious increasing trend. Thus we difference a second time to obtain stationarity. The result are as follows:




```
ADF Statistic: -8.670122470963946
P-value: 4.5703530861173713e-14
Critical Values: {'1%': -3.442772146350605, '5%': -2.8670191055991836, '10%': -2.5696881663873414}
```

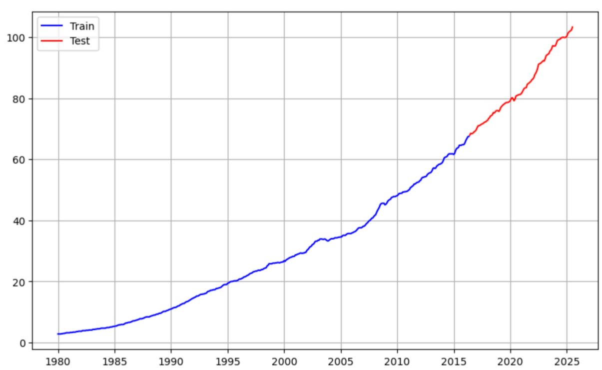
The p-value now is significantly small, indicating the time series is now stationary. This is supported by the above graph as it is centered around 0 and there is no obvious trend. Additionally, this is reinforced by the ACF and PACF graphs below as there is a sudden drop after the first few lags which is a characteristic of stationarity.





Model Selection and Fitting Using SAIMAX

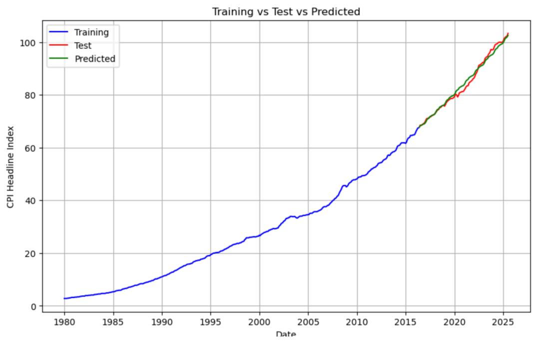
Model selection and fitting were performed using the original dataset in long format, as the forecasting packages applied do not require the series to be stationary. To evaluate model performance, the data was divided into training and testing sets, with an 80/20 split. The training set was used to fit the model while the testing set provided an independent basis for assessing predictive accuracy as illustrated below.



A stepwise search was performed to identify the optimal parameters of the SARIMAX model. Based on the selection criteria, the best-fitting model for the CPI data was found to be SARIMAX(1,1,1)(0,1,1)[12] which accounts for both non-seasonal and seasonal components with a yearly seasonal period of 12. The results of the fitted model are presented below:

```
Best model:  ARIMA(1,1,1)(0,1,1)[12] intercept
Total fit time: 52.819 seconds
                                      SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                 437
Model:             SARIMAX(1, 1, 1)x(0, 1, 1, 12)   Log Likelihood                 230.607
Date:                             Mon, 01 Sep 2025   AIC                           -451.214
Time:                                     13:52:18   BIC                           -430.966
Sample:                                 01-01-1980   HQIC                          -443.214
                                      - 05-01-2016
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      0.0026      0.001      2.267      0.023       0.000       0.005
ar.L1          0.6357      0.087      7.280      0.000       0.465       0.807
ma.L1         -0.3872      0.112     -3.470      0.001      -0.606      -0.168
ma.S.L12      -0.7809      0.028    -28.278      0.000      -0.835      -0.727
sigma2         0.0192      0.001     19.104      0.000       0.017       0.021
===================================================================================
Ljung-Box (L1) (Q):                   0.04   Jarque-Bera (JB):                57.30
Prob(Q):                              0.84   Prob(JB):                         0.00
Heteroskedasticity (H):               5.18   Skew:                            -0.02
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.80
===================================================================================
```

The selected SARIMAX model was then used to generate forecasts of the CPI indices, which were plotted alongside the test data to visually assess the model's fit. In addition to the graphical comparison, evaluation metrics were calculated to quantify the model's predictive accuracy. These metrics provide a measure for determining how well the model captures the underlying patterns in the data.

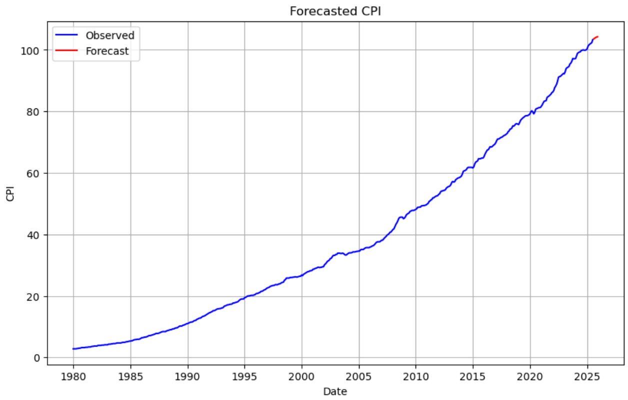



```
R_2 : 0.9868619964116112
Mean Absolute error : 0.9621838269356379
Mean Absolute percentage Error : 0.011144045619313443
```

The performance of the SARIMAX(1,1,1)(0,1,1)[12] model was evaluated by comparing the predicted CPI indices against the test dataset. As shown in the graph, the predicted series closely follows the actual values with minimal deviation in the test period, indicating a strong model fit. The evaluation metrics further support this conclusion. The coefficient of determination ($R^2 = 0.987$) suggests that the model explains nearly 99% of the variance in the data. The Mean Absolute Error (MAE $\approx$ 0.96) indicates that, on average, the predictions deviate from the actual values by less than one index point. Similarly, the Mean Absolute Percentage Error (MAPE $\approx$ 1.1%) confirms that the model's forecasts are highly accurate relative to the scale of the data. These results collectively demonstrate that the selected SARIMAX model provides reliable forecasts of CPI trends and captures both long-term patterns and seasonal dynamics effectively.

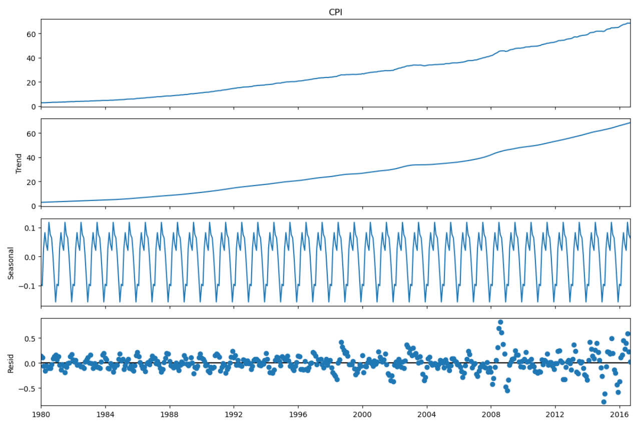
Model Forecasting using SARIMAX

The fitted model is then used to predict future CPI indices from August 2025 until December 2025. The result are as follows:

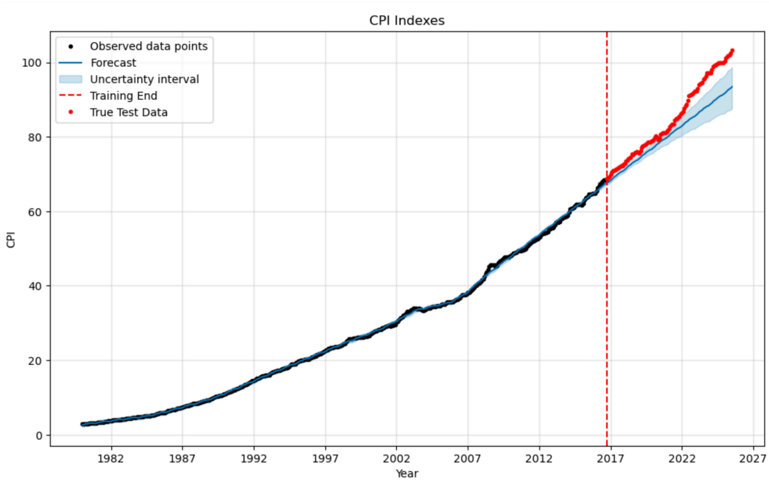



```
Forecasted values:
2025-08-01    103.540885
2025-09-01    103.773767
2025-10-01    104.045307
2025-11-01    104.130902
2025-12-01    104.287623
Freq: MS, Name: predicted_mean, dtype: float64
```

Model Selection and Fitting Using Prophet



The time series data was decomposed into its trend, seasonal and residual components. The analysis reveals an upward trend along with a clear seasonal pattern. The residuals initially fluctuate around zero but gradually exhibit increasing spread over time.



The CPI data was modelled using the Prophet package in Python, and it can be observed that the actual test values deviate from the predicted values.

The results of the evaluation metrics are found below:

```
r_2 : 0.7356888804507352
Mean Absolute error : 4.342034637033902
Mean Absolute percentage Error : 0.04798559412696587
```

The coefficient of determination ($R^2 = 0.74$) suggests that the model explains nearly 74% of the variance in the data. The Mean Absolute Error (MAE ≈ 4.34) indicates that, on average, the predictions deviate from the actual values by about 4 index points. Similarly, the Mean Absolute Percentage Error (MAPE ≈ 4%) shows that the model's forecasts are relatively accurate in relation to the scale of the data. Taken together, these results demonstrate that Prophet provides reliable forecasts of CPI trends and effectively captures both long-term patterns and seasonal dynamics.

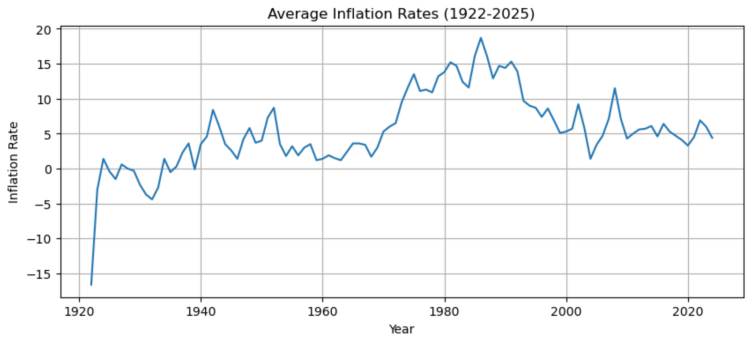
Both the Prophet and SARIMAX(1,1,1)(0,1,1)[12] models provide reliable forecasts of CPI trends, capturing long-term patterns and seasonal dynamics. However, the SARIMAX model outperforms Prophet in terms of predictive accuracy. While Prophet achieves an R^2 of 0.74, an MAE of approximately 4.34, and a MAPE of around 4%, the SARIMAX model attains a much higher R^2 of 0.987, with lower errors (MAE ≈ 0.96, MAPE ≈ 1.1%). This indicates that SARIMAX more closely matches the actual CPI values and produces more precise forecasts, making it the preferred choice for accurate short- and medium-term CPI prediction.

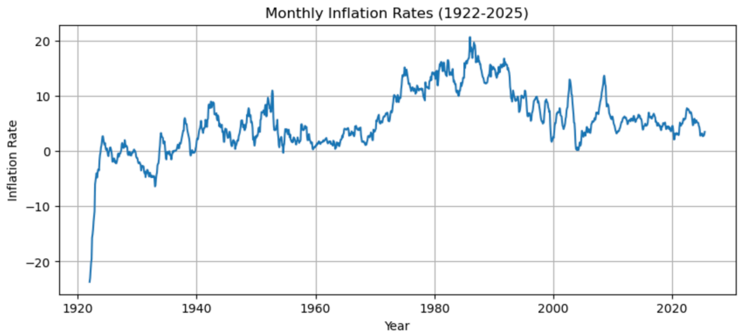
YoY Inflation Rates Analysis

The aim of this project is to analyse South African inflation rate data using time series methods and to generate forecasts of future values. The modelling approach applied includes the use of Prophet.

The data used in the analysis was obtained from StatsSA which is publicly available on the website in pdf format. It consists of monthly inflation rates from 1911 to 2025 for each month, although from 1911 to 1921, the rates are missing. There is also an additional column of annual inflation rate averages.

The data was imported and converted to CSV format. The original data was split into two tables when converting from PDF to CSV format, so the data had to be concatenated and indexing was corrected. Steps were taken to ensure correct formatting of headings and missing values were removed. The data was also converted to float type and decimal commas were replaced with a decimal point. For subsequent analysis, the data is transformed from wide to long format for time series analysis and forecasting.





The graphs show South Africa’s inflation rates from 1922 to 2025. Both the average and monthly series highlight a clear peak during the late 1970s and early 1980s. After this period, inflation steadily declined and has remained more stable since the 2000s, though monthly data reveal continued short-term fluctuations. Overall, the series reflects long-term structural changes with reduced volatility in recent decades

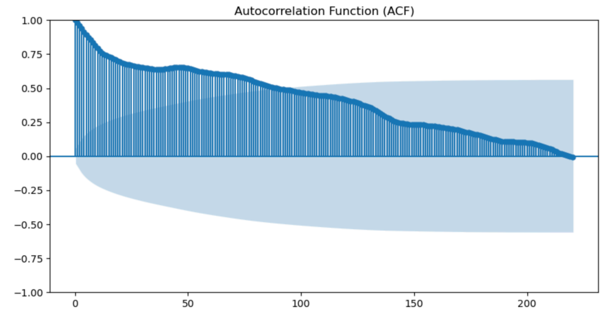
To assess whether the inflation rates are stationary, the ADF test was applied.

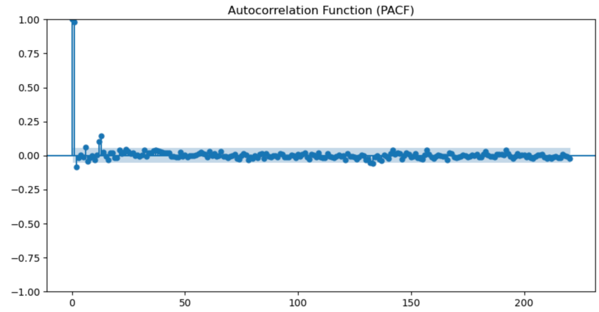
H_0 : Time series is non-stationary.
H_A : Time series is stationary.

Results of the ADF test are as follows:

```
ADF Statistic: -3.0697823609656205
P-value: 0.028871762164032
Critical Values: {'1%': -3.4356863371792095, '5%': -2.8638964938393667, '10%': -2.568024631481501}
```

According to the ADF test, the p-value is less than the significance level of 5%. Thus we reject the null hypothesis and conclude that the inflation rates data is stationary.





The ACF graph suggests that the data is non-stationary, as it takes approximately 220 lags for the autocorrelations to decay to zero, although the PACF shows a sudden drop. This observation appears to contradict the results of the ADF test presented earlier.

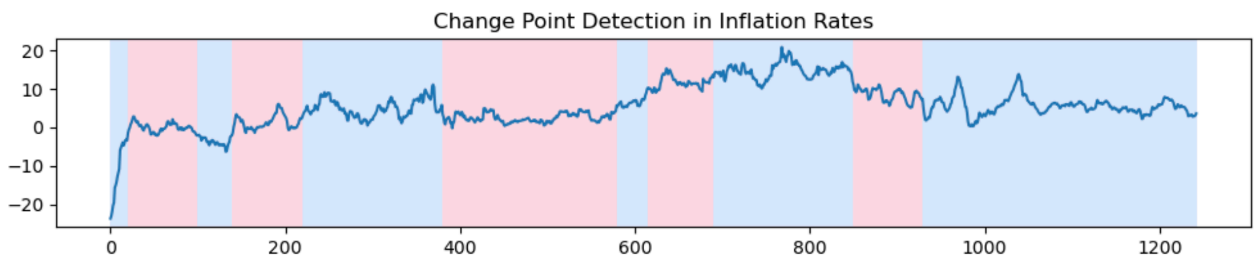
We perform a Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test to further verify the stationarity of the data.

H_0 : The time series is stationary, either level stationary or trend stationary.
H_A : The time series is non-stationary and contains a unit root.

The result of the KPSS test is as follows:

```
KPSS Statistic: 2.3145195099101907, p-value: 0.01
```

The p-value is below the 5% significance level, so we reject the null hypothesis of stationarity. This indicates that the time series is non-stationary, which aligns with the observations from the ACF graph but contradicts the ADF test results. To better understand the structural changes in the series and identify points where the trend or variance shifts, we will now proceed with change point detection.



The change point detection graph displays the inflation rate values over time, with vertical lines indicating detected change points. In this graph, several shifts are visible, showing points where the mean changes significantly. The series can be seen segmented between these points, highlighting periods of relative stability and periods of structural change. This visualisation helps identify multiple structural breaks in the data which can inform more accurate modelling and forecasting.

```
Segment 1: ADF p=0.057,  Not stationary
Segment 1: KPSS p=0.010, Not stationary
Segment 2: ADF p=0.739,  Not stationary
Segment 2: KPSS p=0.100, Stationary
Segment 3: ADF p=0.202,  Not stationary
Segment 3: KPSS p=0.064, Stationary
Segment 4: ADF p=0.088,  Not stationary
Segment 4: KPSS p=0.100, Stationary
Segment 5: ADF p=0.099,  Not stationary
Segment 5: KPSS p=0.100, Stationary
Segment 6: ADF p=0.204,  Not stationary
Segment 6: KPSS p=0.100, Stationary
Segment 7: ADF p=0.991,  Not stationary
Segment 7: KPSS p=0.022, Not stationary
Segment 8: ADF p=0.392,  Not stationary
Segment 8: KPSS p=0.100, Stationary
Segment 9: ADF p=0.019, Stationary
Segment 9: KPSS p=0.100, Stationary
Segment 10: ADF p=0.648,  Not stationary
Segment 10: KPSS p=0.047, Not stationary
Segment 11: ADF p=0.020, Stationary
Segment 11: KPSS p=0.100, Stationary
```

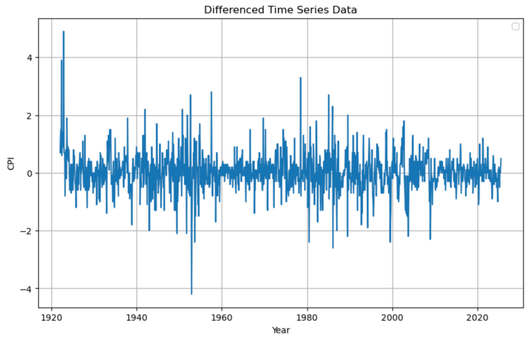
The output above shows the results of stationarity tests applied to each segment identified in the change point detection graph. It can be observed that in some segments, both tests are in agreement, while in other segments, the tests yield conflicting results.

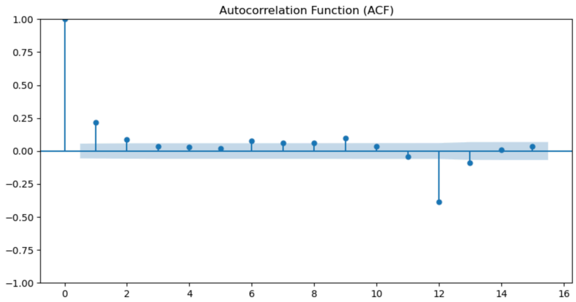
```
        Zivot-Andrews Results
=====================================
Test Statistic                 -4.413
P-value                         0.147
Lags                               14
-------------------------------------

Trend: Constant
Critical Values: -5.28 (1%), -4.81 (5%), -4.57 (10%)
Null Hypothesis: The process contains a unit root with a single structural break.
Alternative Hypothesis: The process is trend and break stationary.
```

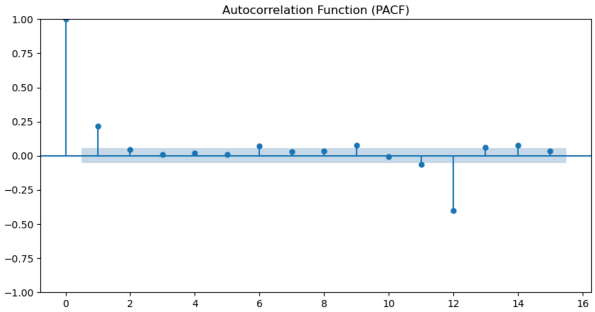
```
      Dickey-Fuller GLS Results
=====================================
Test Statistic                 -0.062
P-value                         0.672
Lags                               14
-------------------------------------

Trend: Constant
Critical Values: -2.58 (1%), -1.96 (5%), -1.64 (10%)
Null Hypothesis: The process contains a unit root.
Alternative Hypothesis: The process is weakly stationary.
```

To further assess stationarity, we applied the Zivot-Andrews test and the ADF-GLS test. For both tests, we fail to reject the null hypothesis. Based on these results, along with the previous analyses, we conclude that the time series is non-stationary. Consequently, we apply differencing to transform the series into a stationary one.

The figure below shows the differenced inflation rate data along with its ACF and PACF plots. The time series is now centered around zero, and both the ACF and PACF decay to zero rapidly, confirming that the series has now become stationary.

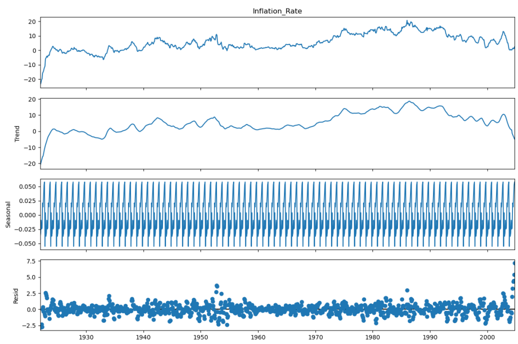


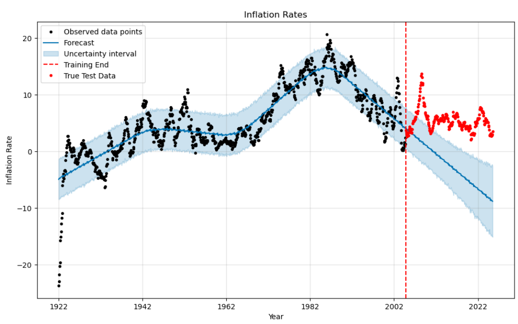




Model Fitting and Forecasting using Prophet

Below, we performed model decomposition on the original data, which reveals no clear trend but a noticeable seasonal pattern. The residuals are generally centered around zero, with a few outliers. This indicates that the model effectively captures the seasonal component, while the remaining variation is mostly random noise.





The above graph presents the inflation rate series, with observed data in black, the training period ending around 2002, and the true test data shown in red. During the training period, the model fits the historical inflation data reasonably well, capturing both the upward movement leading into the 1980s and the subsequent downward trend. The forecast follows the overall direction of the training data, and the uncertainty interval appropriately reflects increasing uncertainty over time.

However, in the testing period, the actual inflation rates deviate noticeably from the model's predicted values, remaining consistently above the forecast. This indicates that while the model performs well in describing and learning from past data, it underestimates more recent inflation dynamics and short-term volatility. Overall, the model is effective in capturing long-term historical patterns but less accurate in predicting out-of-sample behaviour.

While the model used in Prophet captures long-term historical patterns in the inflation rate, its predictive performance on recent data is less accurate, as actual values deviate from the forecast. Further analysis of the stationarity of the series, along with appropriate transformations or alternative modelling approaches, can be undertaken to improve the accuracy and robustness of the results.