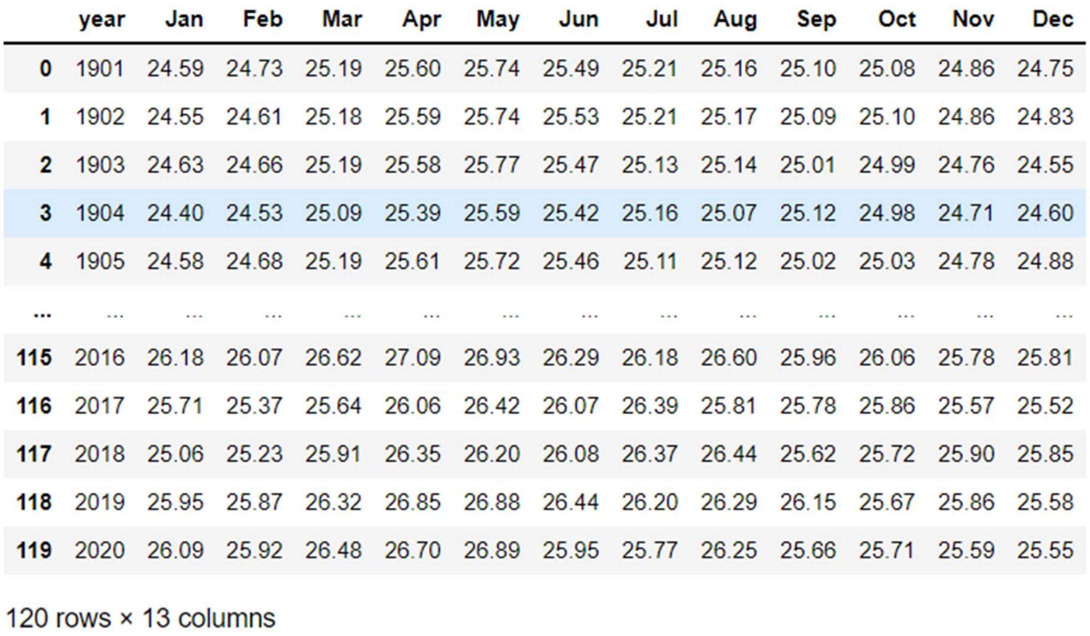
Title: Time series modelling on monthly mean temperature between 1960 and 2020 in Malaysia

1. Main objective of the analysis that also specifies whether your model will be focused on a specific type of Time Series, Survival Analysis, or Deep Learning and the benefits that your analysis brings to the business or stakeholders of this data.

- Climate change threatens the lives and livelihoods of many people through natural disasters, environmental degradation and extreme weather patterns disrupt harvests, deplete fisheries, erode livelihoods and spur infectious diseases [4]. The main objective of the analysis is therefore to find the best out of three SARIMA model that yields prediction closely follows the actual value of monthly mean temperature for climate change detection and monitoring.

2. Brief description of the data set you chose, a summary of its attributes, and an outline of what you are trying to accomplish with this analysis.

- The historical data between 1901 and 2020 on monthly mean temperature of Malaysia is downloadable in CSV file format via Climate Change Knowledge Portal, https://climateknowledgeportal.worldbank.org/download-data.



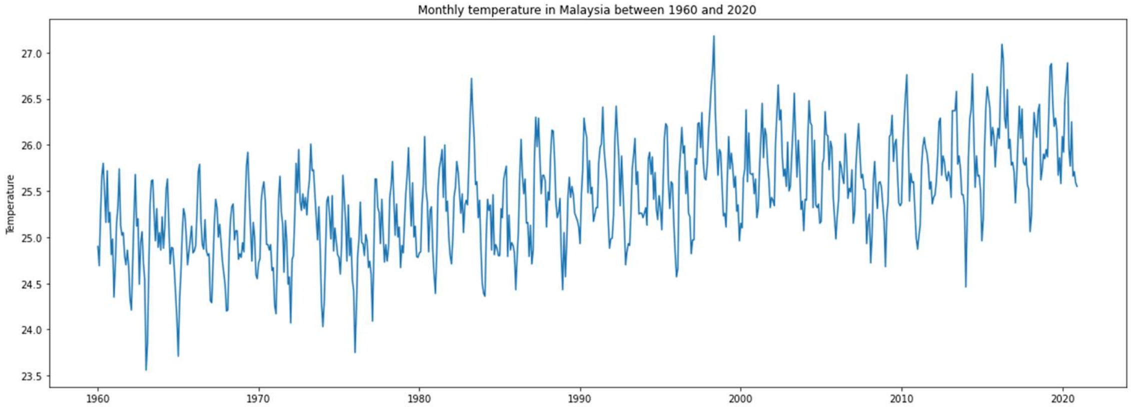
	year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
0	1901	24.59	24.73	25.19	25.60	25.74	25.49	25.21	25.16	25.10	25.08	24.86	24.75
1	1902	24.55	24.61	25.18	25.59	25.74	25.53	25.21	25.17	25.09	25.10	24.86	24.83
2	1903	24.63	24.66	25.19	25.58	25.77	25.47	25.13	25.14	25.01	24.99	24.76	24.55
3	1904	24.40	24.53	25.09	25.39	25.59	25.42	25.16	25.07	25.12	24.98	24.71	24.60
4	1905	24.58	24.68	25.19	25.61	25.72	25.46	25.11	25.12	25.02	25.03	24.78	24.88
...	...	...	...	...	...	...	...	...	...	...	...	...	...
115	2016	26.18	26.07	26.62	27.09	26.93	26.29	26.18	26.60	25.96	26.06	25.78	25.81
116	2017	25.71	25.37	25.64	26.06	26.42	26.07	26.39	25.81	25.78	25.86	25.57	25.52
117	2018	25.06	25.23	25.91	26.35	26.20	26.08	26.37	26.44	25.62	25.72	25.90	25.85
118	2019	25.95	25.87	26.32	26.85	26.88	26.44	26.20	26.29	26.15	25.67	25.86	25.58
119	2020	26.09	25.92	26.48	26.70	26.89	25.95	25.77	26.25	25.66	25.71	25.59	25.55

120 rows × 13 columns

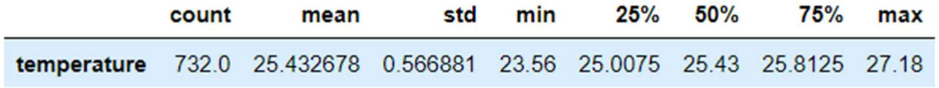
- Data was first reshaped to only two columns: date and temperature for subsequent analysis.
- Logarithm transformation on temperature was performed to achieve constant variances.
- To achieve series stationarity, 12-month seasonal differencing was used based on Augmented Dickey Fuller(ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test for stationary test.
- Final model was recommended according to diagnostic plots and performance metric.

3. Brief summary of data exploration and actions taken for data cleaning or feature engineering.

- Data prior to 1960 was not analysed to avoid introduction of unnecessary noise to time series modelling.
- Data contains neither duplicates nor missing values.
- The variance does not seem to be constant and there seems to be an increasing trend after 1980:

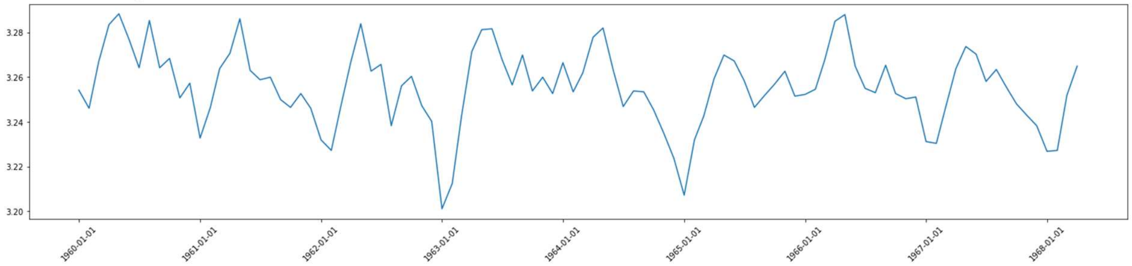


- Summary statistics of the monthly temperature between 1960 and 2020:

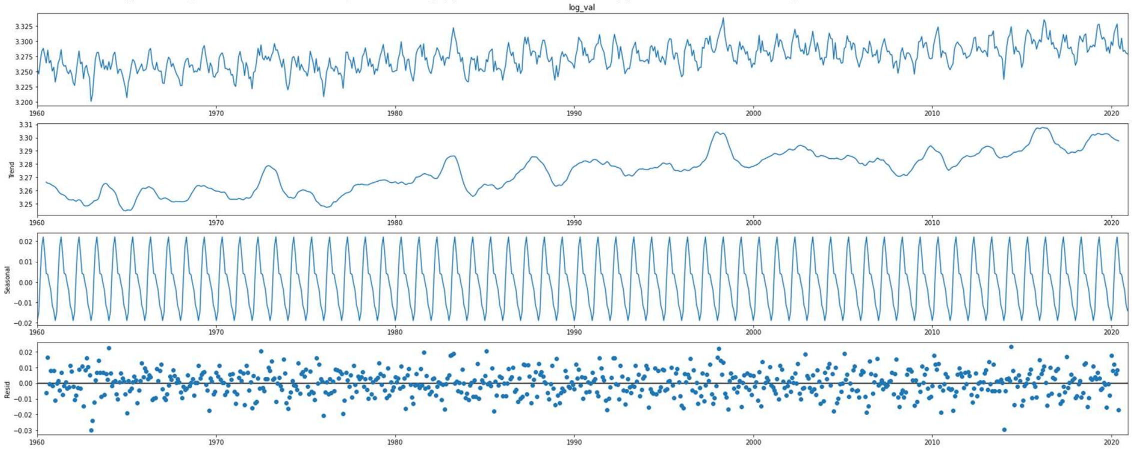


	count	mean	std	min	25%	50%	75%	max
temperature	732.0	25.432678	0.566881	23.56	25.0075	25.43	25.8125	27.18

- After logarithm transformation:



- Decomposing time series by setting period=12 to capture seasonality:



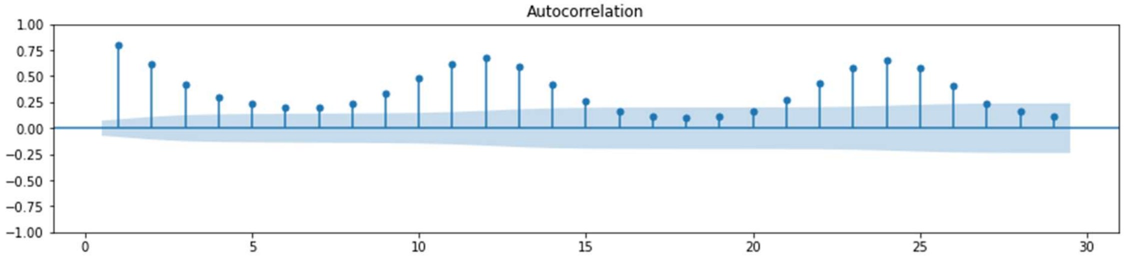
- Test for stationary: Both tests confirm that the given series is stationary.
  i. ADF: The p-value is less than 0.05, thus we reject the Null Hypothesis. Therefore, the series has a unit root and is stationary.

```
ADF Statistic: -7.940554
p-value: 0.000000
Critical Values:
        1%: -3.440
        5%: -2.866
        10%: -2.569
```

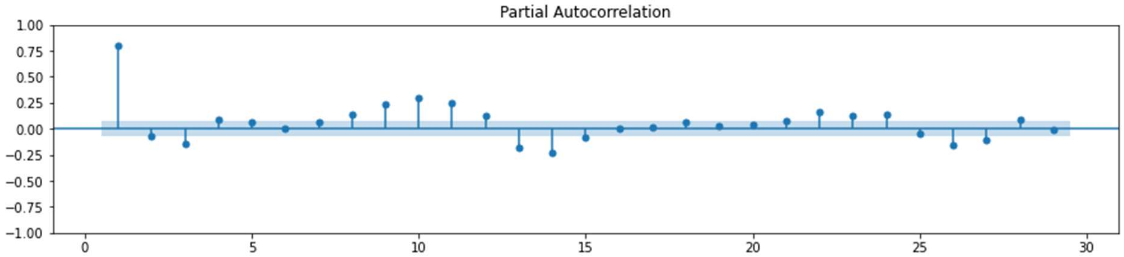
  ii. KPSS: The p-value of the KPSS test is more than 0.05. Thus, we will not reject the null hypothesis that the series is stationary. The KPSS test concludes that the series is stationary.

```
Results of KPSS Test:
Test Statistic              0.024387
p-value                     0.100000
Lags Used                  14.000000
Critical Value (10%)        0.347000
Critical Value (5%)         0.463000
Critical Value (2.5%)       0.574000
Critical Value (1%)         0.739000
dtype: float64
```

- ACF plot:

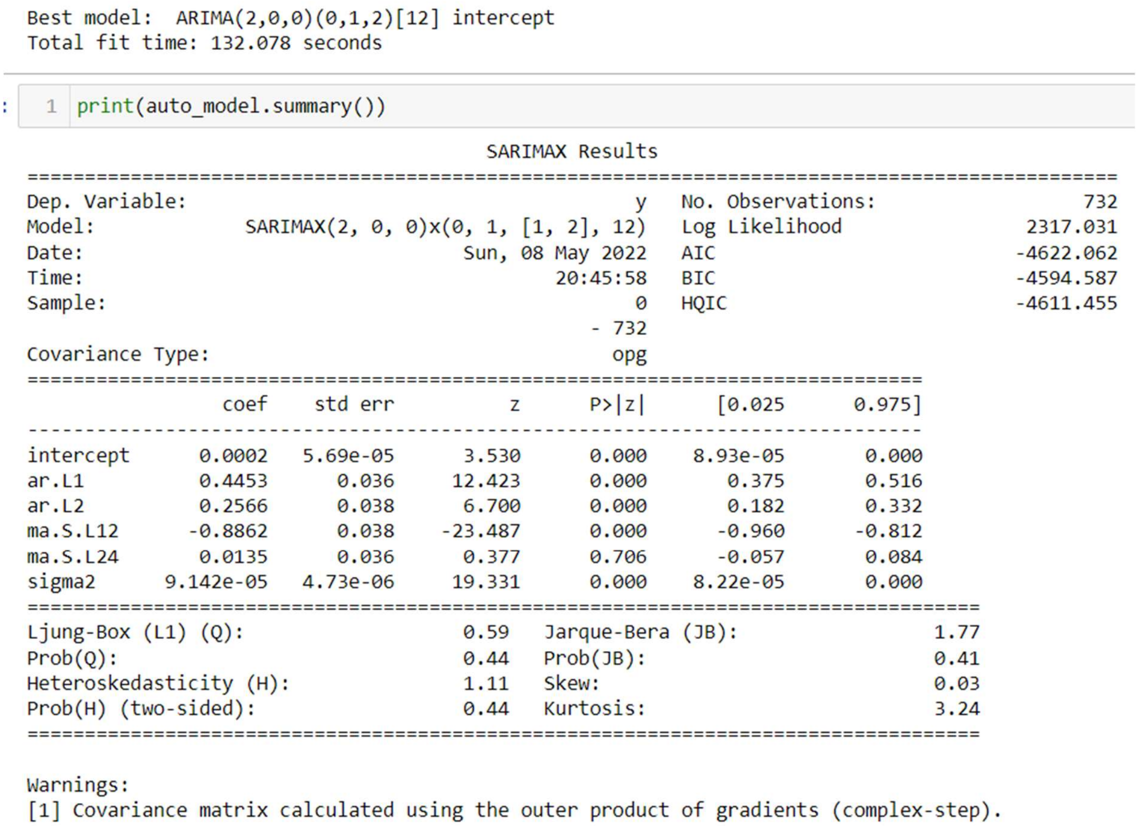


- PACF plot:



4. Summary of training at least three variations of the Time Series, Survival Analysis, or Deep Learning model you selected. For example, you can use different models or different hyperparameters.

   i. Model with optimal p, q, d, P, Q, D and m obtained from auto-ARIMA process with stepwise algorithm by setting D=1 i.e. one seasonal differencing, m=12 i.e. monthly data:



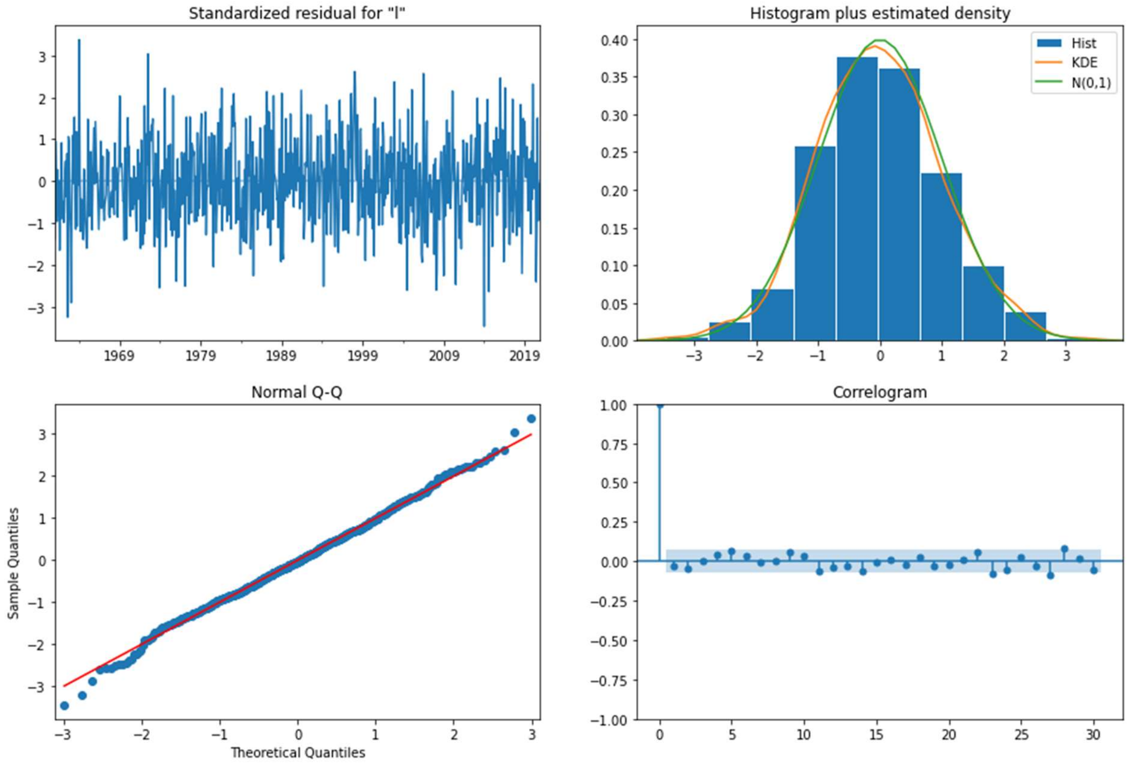
```
Best model:  ARIMA(2,0,0)(0,1,2)[12] intercept
Total fit time: 132.078 seconds
```

```
print(auto_model.summary())
```

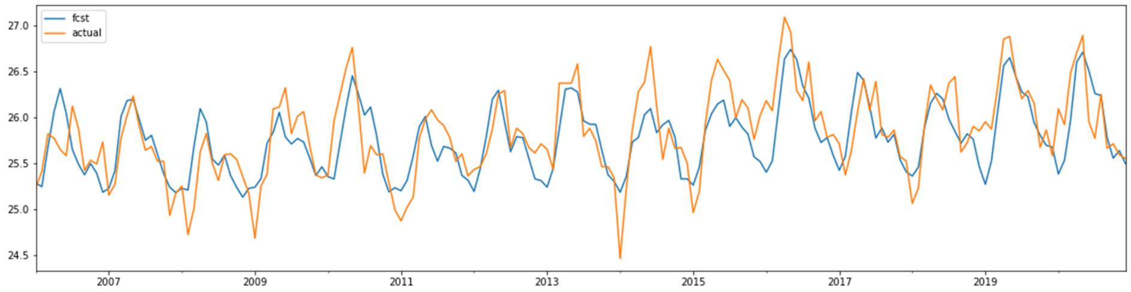
```
                                      SARIMAX Results
================================================================================================
Dep. Variable:                                   y   No. Observations:                      732
Model:             SARIMAX(2, 0, 0)x(0, 1, [1, 2], 12)   Log Likelihood                    2317.031
Date:                             Sun, 08 May 2022   AIC                              -4622.062
Time:                                     20:45:58   BIC                              -4594.587
Sample:                                          0   HQIC                             -4611.455
                                             - 732
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      0.0002   5.69e-05      3.530      0.000    8.93e-05       0.000
ar.L1          0.4453      0.036     12.423      0.000       0.375       0.516
ar.L2          0.2566      0.038      6.700      0.000       0.182       0.332
ma.S.L12      -0.8862      0.038    -23.487      0.000      -0.960      -0.812
ma.S.L24       0.0135      0.036      0.377      0.706      -0.057       0.084
sigma2      9.142e-05   4.73e-06     19.331      0.000    8.22e-05       0.000
===================================================================================
Ljung-Box (L1) (Q):                   0.59   Jarque-Bera (JB):                 1.77
Prob(Q):                              0.44   Prob(JB):                         0.41
Heteroskedasticity (H):               1.11   Skew:                             0.03
Prob(H) (two-sided):                  0.44   Kurtosis:                         3.24
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

The best model with lowest AIC generated from auto-ARIMA process is $(2,0,0)(0,1,2)_{12}$.
According to the model summary, the model meets the condition of independence in the residuals (no correlation) because the p-value of the Ljung-Box test (Prob(Q)) is greater than 0.05, so we cannot reject the null hypothesis of independence. We can say that the residual distribution is homoscedastic (constant variance) because the p-value of the Heteroskedasticity test (Prob(H)) is greater than 0.05.



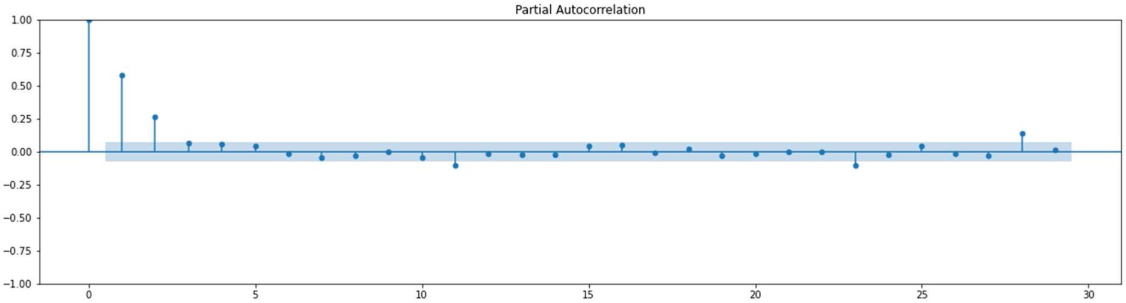
The above four diagnostic plots all look healthy. No significant autocorrelation spikes, normally distributed residuals and homoscedastic values.

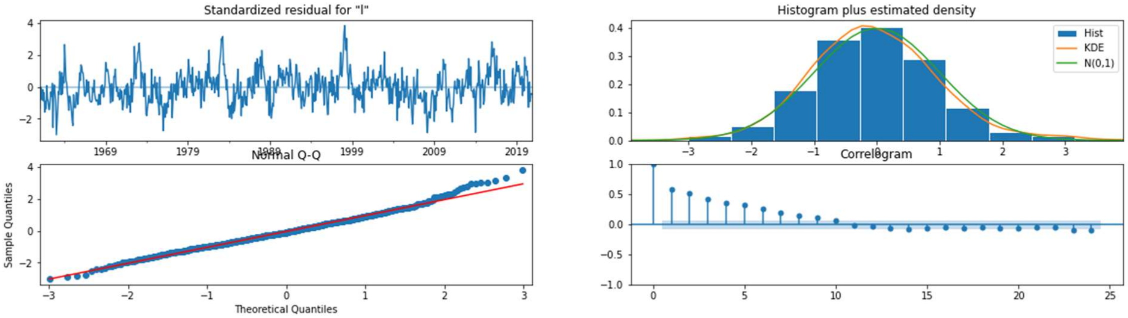


Above plot shows the back transformed forecast from this model after cross validation and the actual values.

The mean absolute percentage error (MAPE) is about 0.94%.

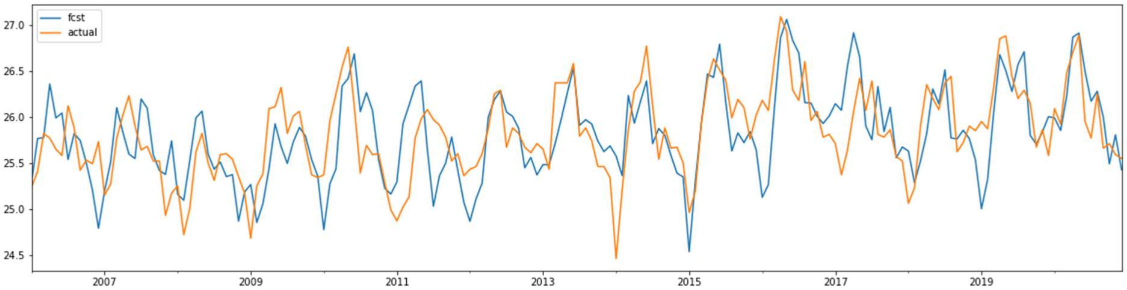
ii. Set only the seasonal order to be P=0, D=1, Q=2, m=12 and the nothing for the non-seasonal order i.e. ARIMA $(0,0,0)(0,1,2)_{12}$:





Normality of residuals and homoscedasticity of variance are healthy but autocorrelation is a problem as depicted by the clear spikes from the autocorrelation plot.

iii. ARIMA $(2,0,0)(0,1,0)_{12}$ by setting seasonal component of P i.e. auto-regressive order and Q i.e. moving average order to 0.



Above plot shows the back transformed forecast from ARIMA $(2,0,0)(0,1,0)_{12}$ model after cross validation and the actual values. The forecast values do not follow as closer the actual values as much than model i.

The mean absolute percentage error (MAPE) is about 1.24% which is slightly worse than model i.

5. A paragraph explaining which of your models you recommend as a final model that best fits your needs in terms of accuracy or explainability.

- ARIMA $(2,0,0)(0,1,2)_{12}$ is recommended due to healthy diagnostic plots and lowest MAPE. The forecast values follow closely the actual values compared to the other two which has either autocorrelation problem or worse MAPE.

6. Summary Key Findings and Insights, which walks your reader through the main findings of your modeling exercise.

   The impact of climate change are global in scope and unprecedented in scale [11]. Hence, time series models are valuable tool to study the variability pattern and prediction of weather data.

   Given our data, the ARIMA(2,0,0)(0,1,2)$_{12}$ model is able to produce prediction as close as to the actual values with very low MAPE based on the historical data between 1960 and 2020 on monthly mean temperature in Malaysia compared to the other two.

7. Suggestions for next steps in analyzing this data, which may include suggesting revisiting this model or adding specific data features to achieve a better model.

- Data is about Malaysia, a tropical country. Therefore, result may not be applicable to other countries especially non-tropical countries.
- Explore other time series algorithms such as deep learning and compare their performance.
- Collect data on other features related to climate change such as precipitation, rainfall and etc.
- Combine previous points and study them together in time series modelling.

References

[1] https://www.coursera.org/learn/time-series-survival-analysis/home/welcome

[2] https://climateknowledgeportal.worldbank.org/download-data

[3] https://www.un.org/humansecurity/wp-content/uploads/2017/10/Human-Security-and-Climate-Change-Policy-Brief-1.pdf

[4] https://www.usgs.gov/science/faqs/climate-and-land-use-change

[5] https://alkaline-ml.com/pmdarima/modules/classes.html

[6] https://www.machinelearningplus.com/time-series/kpss-test-for-stationarity/

[7] https://www.alldatascience.com/time-series/forecasting-time-series-with-auto-arima/

[8] https://towardsdatascience.com/time-series-forecast-error-metrics-you-should-know-cc88b8c67f27

[9] https://medium.com/@ooemma83/how-to-interpret-acf-and-pacf-plots-for-identifying-ar-ma-arma-or-arima-models-498717e815b6

[10] https://towardsdatascience.com/understanding-the-seasonal-order-of-the-sarima-model-ebef613e40fa

[11] https://www.un.org/en/global-issues/climate-change