Time Series Analysis on Daily Air Quality

Abstract

The aim of the analysis is to identify the key factors that contribute to changes in air quality and to develop a predictive model that can forecast air quality levels in the future. We used various regression techniques to evaluate the performance of each technique's prediction. The performance of these models is evaluated using metrics such as mean squared error and Mean Absolute Percentage Error. Finally, the study concludes with a discussion of the results, including the key factors that contribute to air quality changes and the accuracy of the predictive model. The findings of this study have implications for policy-makers, urban planners, and public health officials, who can use the insights gained to develop strategies to improve air quality in cities.

Introduction:

The dataset is an air quality time series dataset that was taken from Kaggle. The original dataset contains hourly averaged data for 9358 instances, while our dataset is a daily averaged version of the same dataset, with 392 instances. The dataset contains responses from a chemical multi-sensor device measuring the concentration of five metal oxides. The temperature is considered as the dependent variable.

During the data preprocessing stage, the dataset was found to contain a few outliers, which were replaced by the average daily values. Additionally, the last 28 observations were considered as test data.

To test for the presence of white noise in the dataset, a significance level of 0.05 was used. This means that any patterns or correlations observed in the data that have a probability of occurring by chance of less than 5% are considered significant. The analysis showed that the dataset was not a white noise series.

Goal

Daily temperature changes can be predicted by considering several factors that influence temperature, including the concentration of metal oxides in the air and humidity levels. These predictions can be of great help to people, allowing them to plan ahead and prepare for potential temperature fluctuations, which can impact their health and result in increased costs.

Factors such as metal oxide concentrations in the air and humidity levels can have a significant effect on temperature, making it important to consider them when predicting daily temperature changes. By doing so, individuals can better plan their daily activities, such as choosing the appropriate clothing to wear or deciding when to go outside. Additionally, businesses and organizations can use this information to prepare for temperature changes and minimize any associated costs.

Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more predictor variables

We first used all the variables and ran a regression model. Then we removed insignificant variables and kept only the 5 significant variables.

We obtained the following model summary:

```
Call:
lm(formula = temperature ~ carbon_monoxide + benzene + nitric_oxide +
    nitrogen_dioxide + relative_humidity, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-19.6739  -3.6898   0.2677   3.8098  14.0096

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       31.750311   1.478028  21.482  < 2e-16 ***
carbon_monoxide    0.170702   0.065160   2.620  0.00915 **
benzene            0.892882   0.070934  12.588  < 2e-16 ***
nitric_oxide      -0.016861   0.003225  -5.228 2.81e-07 ***
nitrogen_dioxide  -0.086136   0.011914  -7.230 2.61e-12 ***
relative_humidity -0.210694   0.022138  -9.517  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.222 on 386 degrees of freedom
Multiple R-squared:  0.5719,    Adjusted R-squared:  0.5664
F-statistic: 103.1 on 5 and 386 DF,  p-value: < 2.2e-16
```

We can see that all the predictor variables are highly significant in describing the dependent variable. The Adj R2 of this model is 56% and the Standard error is 5.22%.

We can infer from the model summary as follows:

A one-unit increase in carbon_monoxide is associated with an expected increase of 0.170702 in the dependent variable, holding all other variables constant.

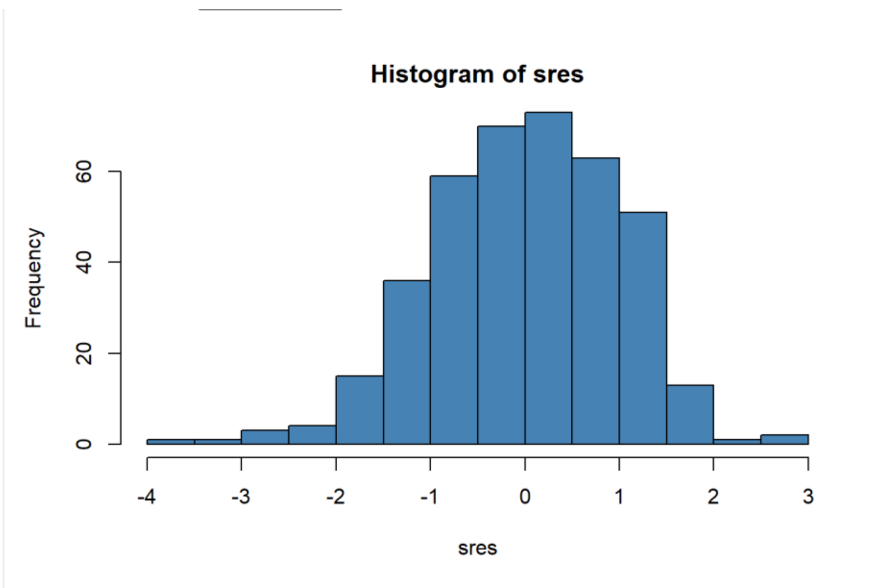
A one-unit increase in Benzene is associated with an expected increase of 0.892882 in the dependent variable, holding all other variables constant.

A one-unit increase in Nitric_oxide is associated with an expected decrease of -0.016861 in the dependent variable, holding all other variables constant.

A one-unit increase in nitrogen_dioxide is associated with an expected decrease of -0.086136 in the dependent variable, holding all other variables constant.

A one-unit increase in relative_humidity is associated with an expected decrease of -0.21069 in the dependent variable, holding all other variables constant.

Normality:



From the above histogram we cannot confidently indicate that the distribution is not a normal distribution or not.

So we performed Shapiro Wilk Normality test. Which gave us a p value of 0.00739 which is less than 0.005 Hence rejecting the null hypothesis. So we can confirm that the series is not a normal distribution.

H_0: Series are normally distributed

H_a: Series are not normally distributed

```
        Shapiro-Wilk normality test

data:  sres
W = 0.98969, p-value = 0.007399
```

Autocorrelation:

```
        Box-Pierce test

data:  sres
X-squared = 1766.8, df = 20, p-value < 2.2e-16
```

H_0:$\rho 1=\rho 2=\cdots=\rho 20=0$(all autocorrelations are zero)

H_a:at least one $\rho k\neq 0$(at least one autocorrelation is not Zero)

From the Box test, with p value =~0 which is less than 0.05. We reject the null hypothesis and conclude that the residuals are autocorrelated.

Non Constant Variance:

A tibble: 1 × 5

statistic <dbl>	p.value <dbl>	parameter <dbl>	method <chr>	alternative <chr>
71.63433	9.830995e-08	20	White's Test	greater

1 row

H_0:There is no Heteroscedasticty (constant variance)
H_a: There is Heteroscedasticty (non-constant variance)
P-value=9.830995e-08
Since the p-value is less than 0.05, we reject the null hypothesis. The residuals are heteroscedastic.(non-constant variance)

Multicollinearity:

```
  carbon_monoxide           benzene      nitric_oxide   nitrogen_dioxide
relative_humidity
         1.031811          1.270924          3.057695           2.189840
1.360262
```

All the VIFs are less than 10. So there might be no multicollinearity.

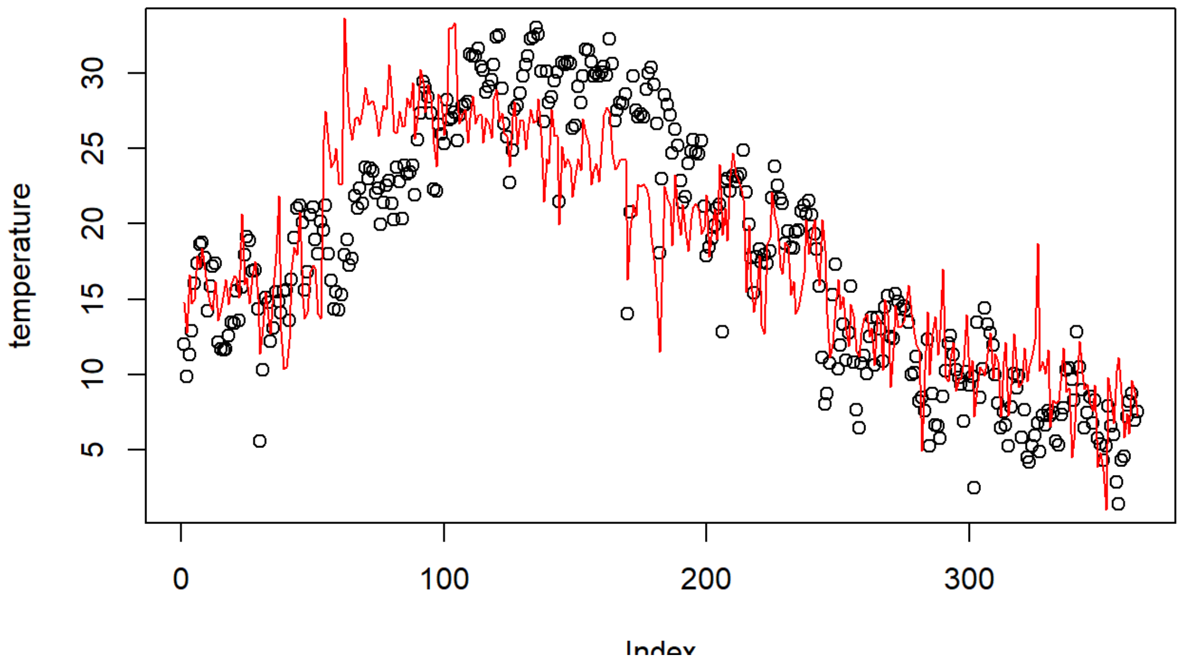
Deterministic Time Series Models:

Seasonal Model:

For the Seasonal model, we considered two seasonal variables, Summer and Winter as there are k=3 seasons we considered (k-1)=2 seasons.

```
Call:
lm(formula = temperature ~ time + (summer) + (winter) + carbon_monoxide +
    benzene + nitric_oxide + nitrogen_dioxide + relative_humidity,
    data = train_seasonal)

Residuals:
     Min       1Q   Median       3Q      Max
-15.7066  -2.7075   0.1396   3.3066  11.7587

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       22.740745   1.508250  15.078  < 2e-16 ***
time              -0.039094   0.006325  -6.180 1.75e-09 ***
summer            10.745493   1.656941   6.485 2.97e-10 ***
winter            12.180181   0.953254  12.777  < 2e-16 ***
carbon_monoxide    0.213045   0.057462   3.708 0.000243 ***
benzene            0.619325   0.071300   8.686  < 2e-16 ***
nitric_oxide      -0.009679   0.003412  -2.837 0.004817 **
nitrogen_dioxide  -0.053463   0.011527  -4.638 4.95e-06 ***
relative_humidity -0.118487   0.020198  -5.866 1.02e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.3 on 355 degrees of freedom
Multiple R-squared:  0.7276,    Adjusted R-squared:  0.7215
F-statistic: 118.6 on 8 and 355 DF,  p-value: < 2.2e-16

[1] 2105.711

        Shapiro-Wilk normality test

data:  mlr_seasonal$residuals
W = 0.99166, p-value = 0.03836


        Box-Pierce test

data:  mlr_seasonal$residuals
X-squared = 1395.7, df = 20, p-value < 2.2e-16
```

From the above model summary, We notice that the seasonal variables are significant as their respective p values are less than 0.05. The adj R2 is 72%. This is higher than the MLR model.



The above graph is the Model fit for seasonal model. As the data is only for an year we are not able to see the seasonal effect of summer and winter in this model.

```
        Box-Pierce test

data:  mlr_seasonal$residuals
X-squared = 1395.7, df = 20, p-value < 2.2e-16
```

H0: $\rho 1=\rho 2=\cdots=\rho 20=0$(all autocorrelations are zero)

Ha: at least one $\rho k\neq 0$(at least one autocorrelation is different than zero) for k=1,...,20

As the p value from Box pierce test is less than 0.05. We can say that the residuals are autocorrelated and not white noise.

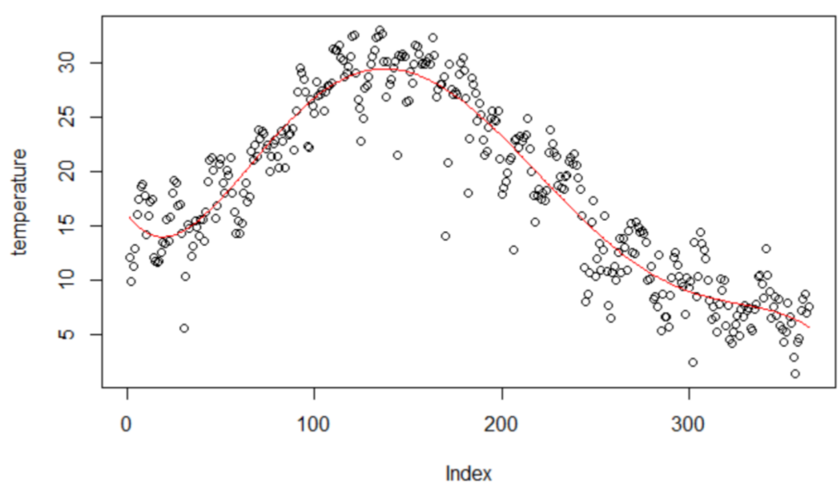
Polynomial Model:

For the Polynomial model, we iterated the value of K upto 10. But the Adj R2 was saturated and had no impact after k=5. Also the K values greater than 5 were not significant. So we considered k =5 for our model which also had the lowest AIC value.

```
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       18.1907     0.1471 123.688   <2e-16 ***
poly(time, k)1   -82.8762     2.8059 -29.536   <2e-16 ***
poly(time, k)2  -102.9847     2.8059 -36.703   <2e-16 ***
poly(time, k)3    48.2476     2.8059  17.195   <2e-16 ***
poly(time, k)4    29.5466     2.8059  10.530   <2e-16 ***
poly(time, k)5   -24.5886     2.8059  -8.763   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.806 on 358 degrees of freedom
Multiple R-squared:  0.883,	Adjusted R-squared:  0.8814 
F-statistic: 540.6 on 5 and 358 DF,  p-value: < 2.2e-16

[1] 1792.036
```

All the polynomial coefficients were statistically significant.

The model fit is as shown below:



Autocorrelation test:

```
        Box-Pierce test

data:  poly_m$residuals
X-squared = 222.29, df = 20, p-value < 2.2e-16
```

H0: $\rho 1=\rho 2=\cdots=\rho 20=0$ (all autocorrelations are zero)

Ha: at least one $\rho k \neq 0$ (at least one autocorrelation is different than zero) for k=1,...,20

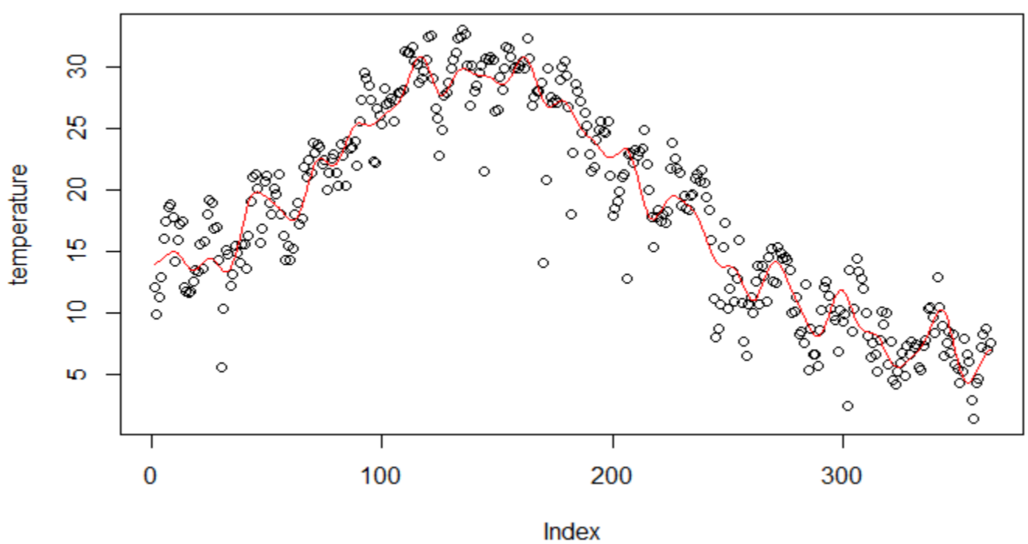
As the P-value < 0.05, we reject the H0 . And conclude that the residuals are autocorrelated and not white noise.

Harmonic Model:

```
Call:
lm(formula = temperature ~ time + sin1 + cos1 + sin2 + cos2 +
    sin4 + cos4 + sin6 + cos6 + sin9 + cos9 + sin10 + cos10 +
    sin11 + cos11)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.1572  -1.4682   0.2033   1.6530   6.1269 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 21.588907   0.522223  41.340  < 2e-16 ***
time        -0.018620   0.002767  -6.728 7.07e-11 ***
sin1         4.950245   0.371593  13.322  < 2e-16 ***
cos1        -7.983597   0.187829 -42.505  < 2e-16 ***
sin2        -1.530334   0.246922  -6.198 1.62e-09 ***
cos2        -0.080975   0.187829  -0.431  0.66666    
sin4        -0.507853   0.195251  -2.601  0.00969 ** 
cos4         0.071178   0.187829   0.379  0.70495    
sin6        -0.523414   0.188269  -2.780  0.00573 ** 
cos6        -0.088830   0.187829  -0.473  0.63656    
sin9         0.810962   0.190512   4.257 2.67e-05 ***
cos9        -0.574989   0.187829  -3.061  0.00238 ** 
sin10        0.396082   0.188861   2.097  0.03670 *  
cos10        0.876552   0.187829   4.667 4.37e-06 ***
sin11        0.860970   0.190044   4.530 8.10e-06 ***
cos11       -0.052624   0.187829  -0.280  0.77951    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.534 on 348 degrees of freedom
Multiple R-squared:  0.9073,	Adjusted R-squared:  0.9033 
F-statistic: 227.1 on 15 and 348 DF,  p-value: < 2.2e-16
```

For the Harmonic Model, we identified the peaks from the periodogram and took the sine cosine pairs accordingly. After removing the insignificant sine cosine pairs we arrived at the above model which had an Adj R2 of 90% and Standard Error of 2.5%.

The Model fit is as shown below:



Autocorrelation test:

```
        Box-Pierce test

data:  tri_m1$residuals
X-squared = 137.63, df = 20, p-value < 2.2e-16
```

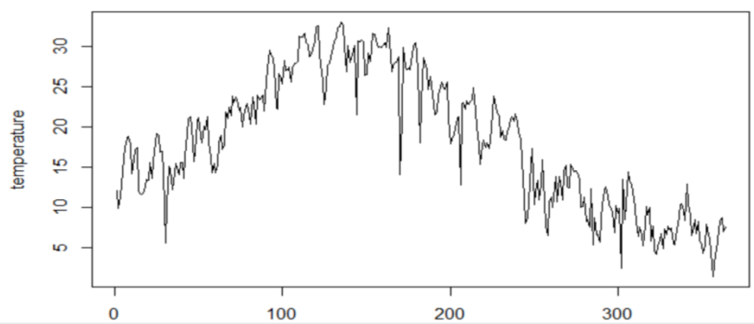
$$H0: \rho1=\rho2=\cdots=\rho20=0 \text{(all autocorrelations are zero)}$$

Ha: at least one $\rho k \neq 0$(at least one autocorrelation is different than zero) for k=1,...,20

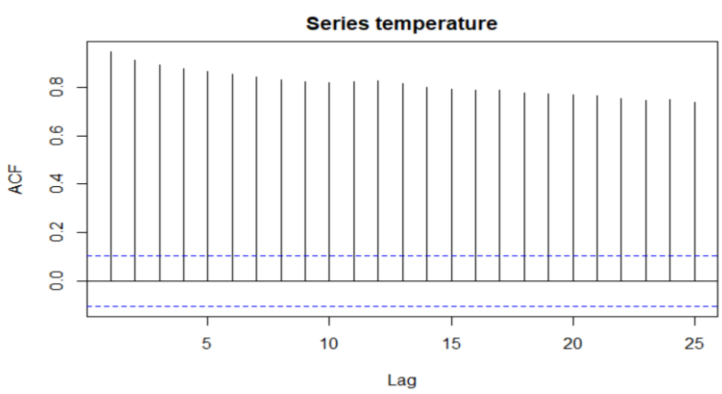
As the P-value < 0.05, we reject the H0 . And conclude that the residuals are autocorrelated and not white noise.

Stochastic Time Series Model:

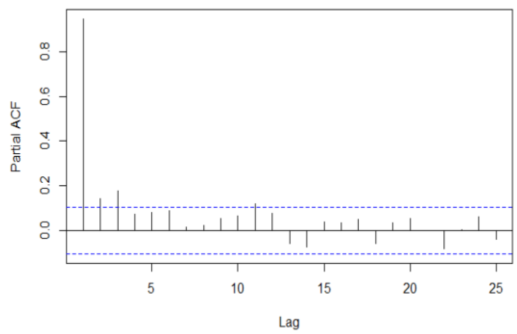
Time Series Plot:



ACF Plot:



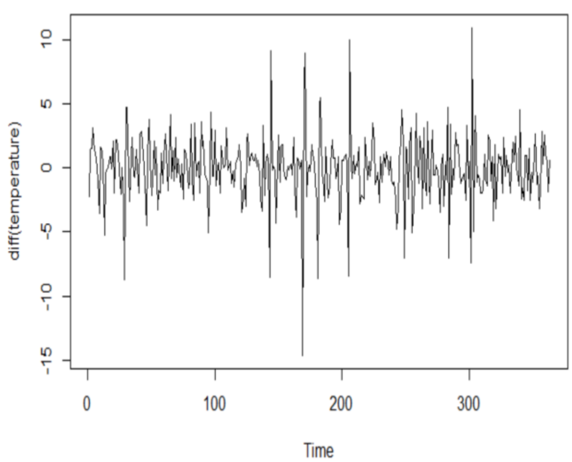
Pacf Plot:



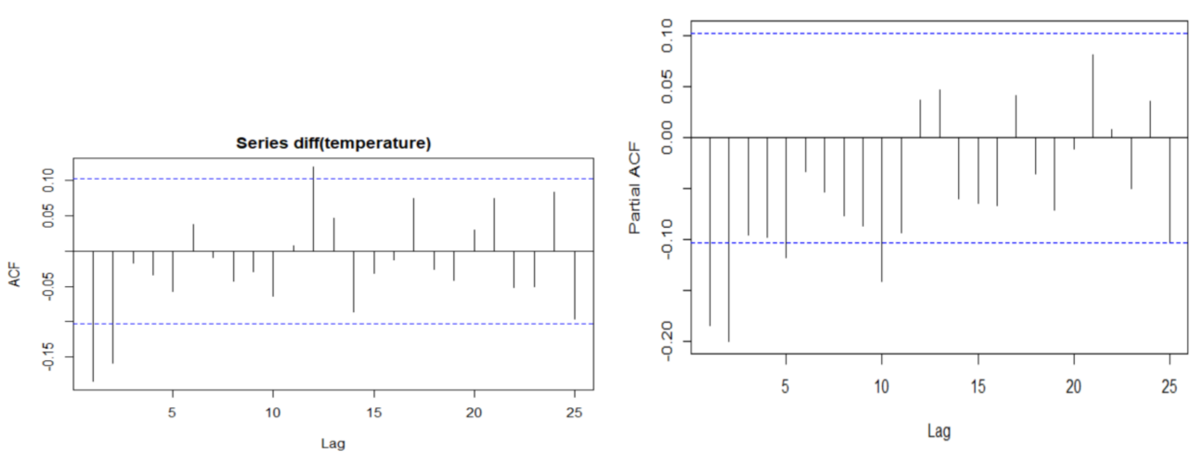
We can see that the Acf exhibits a slow decay which suggests that the series is not stationary.

The time series plot also does not have constant mean and variance. So in order to stabilize the series we used the first differential.

Time Series plot:



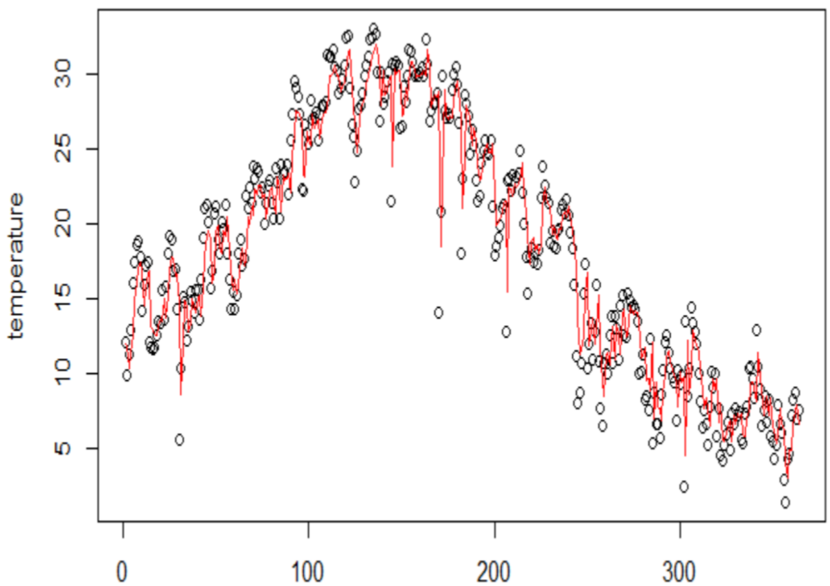
ACF and PACF:



From the above time series plot, we observe that there is constant mean and variance. The ACF plot is also chopped off and there is no slow decay.

As we have higher order AR and MA processes, we considered an ARIMA model with AR(2) and MA(1) which gave us the lowest AIC value.

```
Call:
arima(x = temperature, order = c(2, 1, 1))

Coefficients:
         ar1      ar2      ma1
      0.5714  -0.0274  -0.8669
s.e.  0.0644   0.0572   0.0382

sigma^2 estimated as 5.842:  log likelihood = -835.68,  aic = 1677.37

        Box-Pierce test

data:  arima_fit$residual^2
X-squared = 20.522, df = 20, p-value = 0.4257
```



Autocorrelation Test:

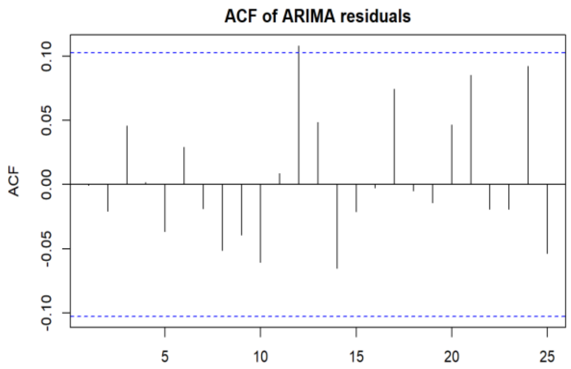
```
        Box-Pierce test

data:  arima_fit$residual^2
X-squared = 20.522, df = 20, p-value = 0.4257
```

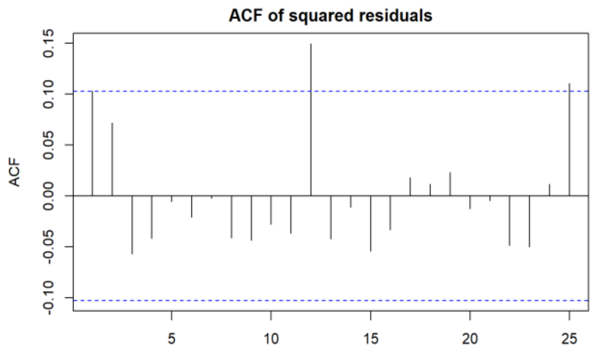
H0: $\rho 1=\rho 2=\cdots=\rho 20=0$ (all autocorrelations are zero)

Ha: at least one $\rho k \neq 0$ (at least one autocorrelation is different than zero) for $k=1,\ldots,20$

As we can see that the p value is greater than 0.05. We accept H0. Hence the Residuals are autocorrelated and are white noise.

We plotted the ACF plots for residuals and squared residuals of ARIMA.





As the residuals are already in white noise. They dont exhibit any ARCH/GARCH Model.

Re-estimating the models:

Reestimating the MLR Model with ARIMA(1,1,1)

```
Call:
arima(x = mlrfit$residuals, order = c(1, 1, 1))

Coefficients:
         ar1      ma1
      0.5544  -0.9184
s.e.  0.0543   0.0247

sigma^2 estimated as 9.961:  log likelihood = -1004.59,  aic = 2013.18
```

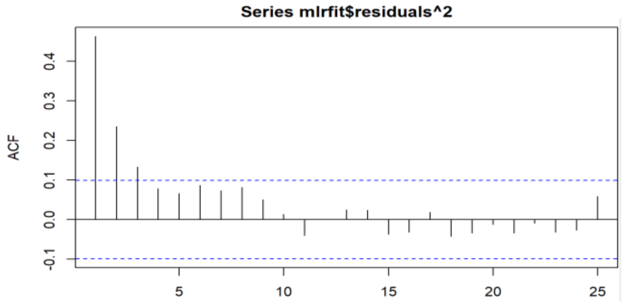
```
	Box-Ljung test

data:  rearima_fit$residuals
X-squared = 6.5173, df = 10, p-value = 0.7701
```

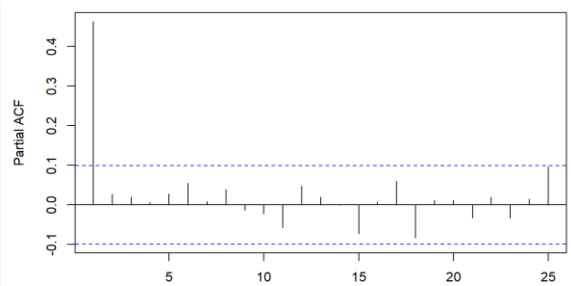
H0:$\rho 1=\rho 2=\cdots=\rho 20=0$(all autocorrelations are zero)

Ha: at least one $\rho k\neq 0$(at least one autocorrelation is different than zero) for k=1,...,20

As we can see that the p value is greater than 0.05. We accept H0. Hence the Residuals are autocorrelated and are white noise.

ACF and PACF:





Reestimating the Seasonal MLR Model with ARIMA(1,1,1):

```
Call:
arima(x = mlr_seasonal$residuals, order = c(1, 1, 1))

Coefficients:
         ar1      ma1
      0.5912  -0.9107
s.e.  0.0684   0.0386

sigma^2 estimated as 6.816:  log likelihood = -863.74,  aic = 1731.49
[1] 1.073862
```

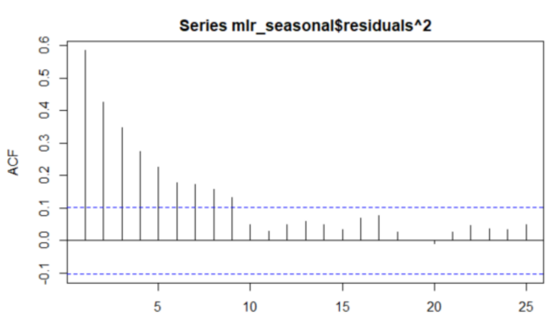
```
        Box-Ljung test

data:  rearima_fit1$residuals
X-squared = 16.833, df = 10, p-value = 0.07814
```

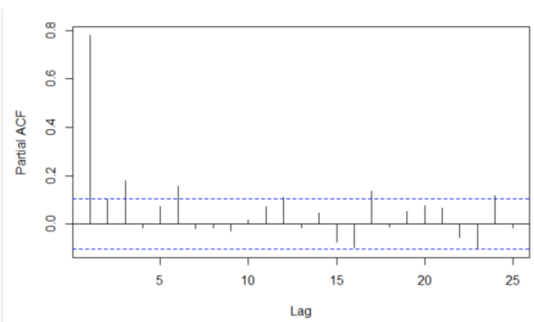
$H0: \rho 1=\rho 2=\cdots=\rho 20=0$(all autocorrelations are zero)

Ha: at least one $\rho k \neq 0$(at least one autocorrelation is different than zero) for k=1,...,20

As we can see that the p value is greater than 0.05. We accept H0. Hence the Residuals are autocorrelated and are white noise.

ACF and PACF:





Reestimating the Polynomial Model with ARIMA(2,0,0):

```
Call:
arima(x = poly_m$residuals, order = c(2, 0, 0))

Coefficients:
         ar1      ar2  intercept
      0.5924  -0.0394    -0.0056
s.e.  0.0523   0.0526     0.2671

sigma^2 estimated as 5.22:  log likelihood = -817.43,  aic = 1640.87
```

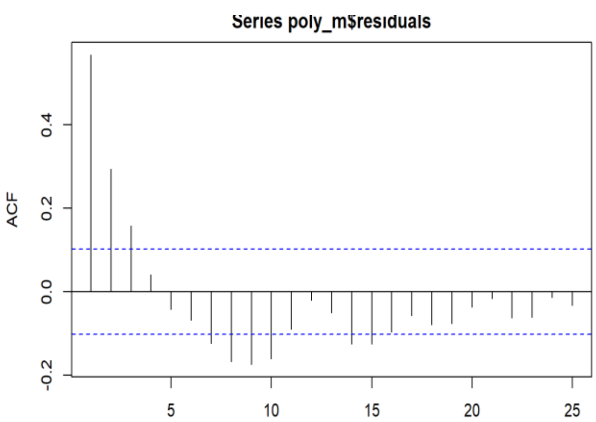
```
        Box-Ljung test

data:  rearima_fit2$residuals
X-squared = 9.7311, df = 10, p-value = 0.4644
```

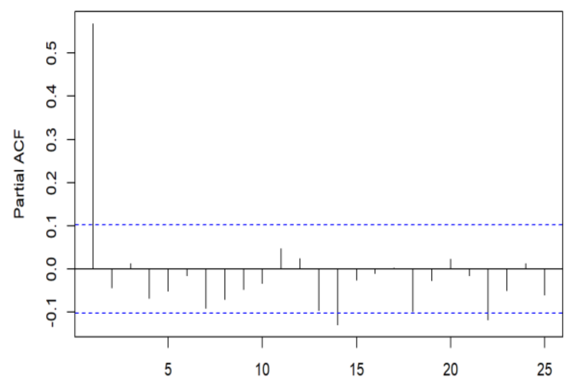
$H0: \rho 1 = \rho 2 = \cdots = \rho 20 = 0$ (all autocorrelations are zero)

Ha: at least one $\rho k \neq 0$ (at least one autocorrelation is different than zero) for k=1,...,20

As we can see that the p value is greater than 0.05. We accept H0. Hence the Residuals are autocorrelated and are white noise

ACF and PACF plots:





Reestimating the Polynomial Model with ARIMA(2,0,0):

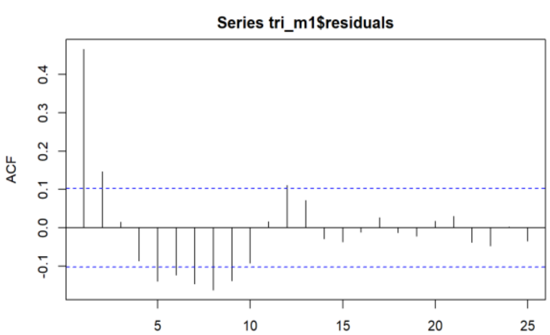
```
Call:
arima(x = tri_m1$residuals, order = c(2, 0, 0))

Coefficients:
         ar1      ar2  intercept
      0.5068  -0.0889    -0.0010
s.e.  0.0521   0.0523     0.1962

sigma^2 estimated as 4.764:  log likelihood = -800.75,  aic = 1607.5

        Box-Ljung test

data:  rearima_fit3$residuals
X-squared = 12.106, df = 10, p-value = 0.278
```

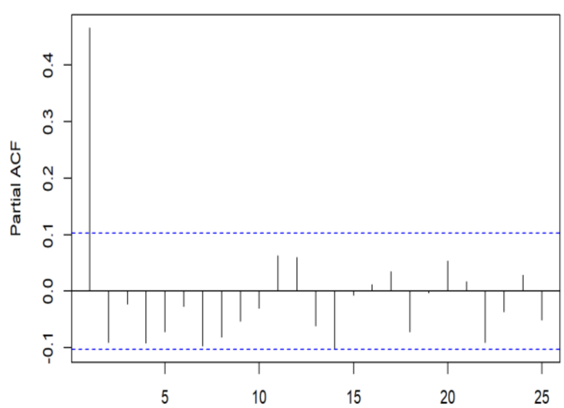
$H0: \rho 1=\rho 2=\cdots=\rho 20=0$(all autocorrelations are zero)

Ha: at least one $\rho k \neq 0$(at least one autocorrelation is different than zero) for k=1,...,20

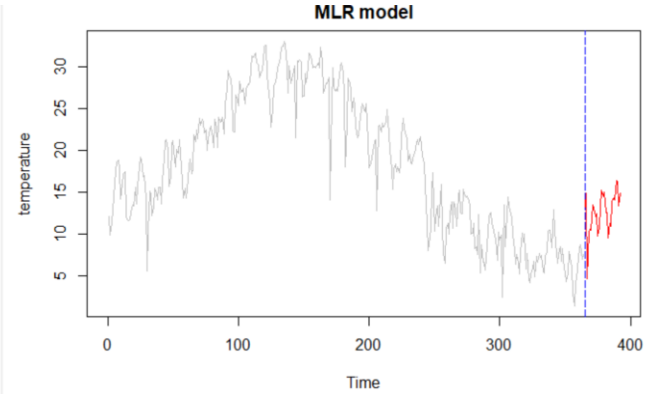
As we can see that the p value is greater than 0.05. We accept H0. Hence the Residuals are autocorrelated and are white noise

ACF and PACF plots:

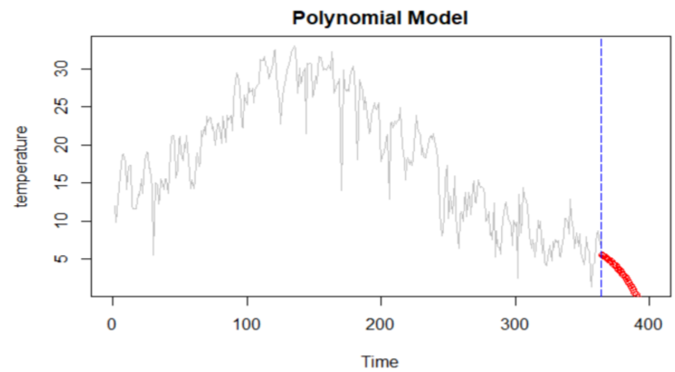




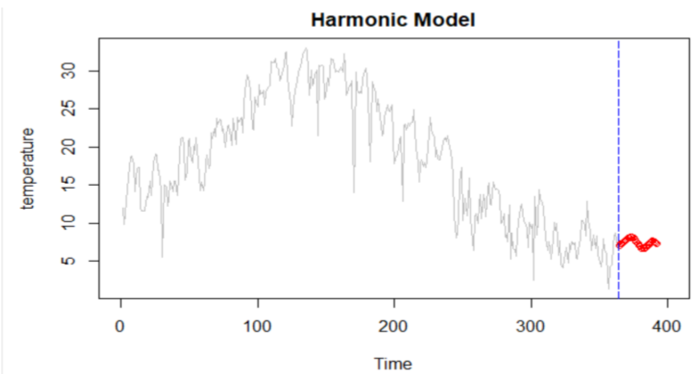
Predictive Comparison:



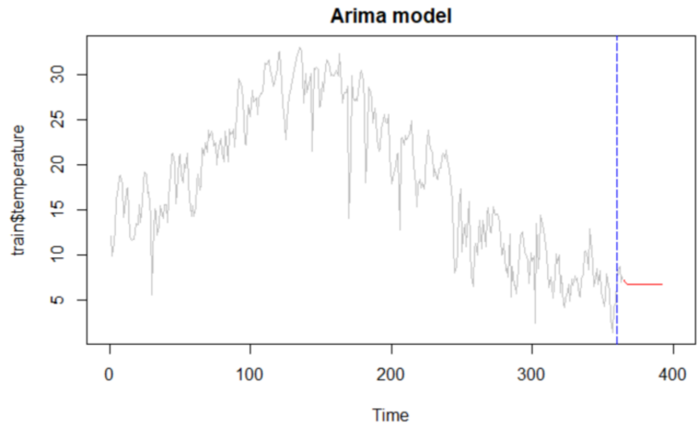
The MLR model prediction in the graph is representing that there is an upward trend from the actual data,



The prediction on the polynomial model is that there is a drop from actual data to the predicted data on temperature.



The Harmonic model is usually used to predict the seasonality in the data. The prediction on the harmonic model states that there is a part of seasonality from the actual data.



The Arima Model predictions here may be directionally correct but not accurate.

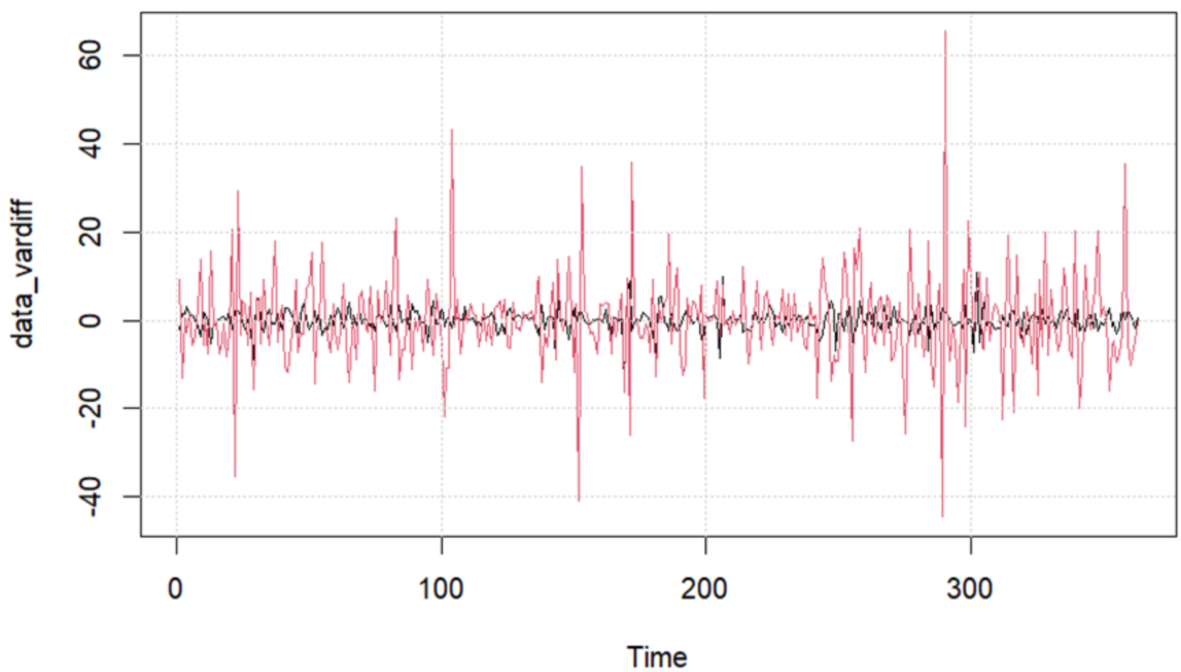
Models	MAPE
MLR	0.2686651
Polynomial Model	0.8151284
Harmonic Model	0.5876667
ARIMA Model	0.6208912

The Mean Absolute Percentage Error on the model we performed. We can state from the comparison that the MLR model is having the lowest MAPE rate at 26% and the Polynomial model is estimated high at 81%.

VARMA Model

VARMA is useful when there are multiple time series that influence each other when the relationship between variables is not straightforward. It is helpful in identifying causal relationships among the variables, as well as it forecasts the future values and estimates the impact of change in one variable on the other.

The dependent variables are Temperature and Relative Humidity. And other variables are reactive with other oxides like (Carbon monoxide, Benzene, Nitric oxide, etc). This is the Time series plot on the Temperature and relative humidity with respect to time.

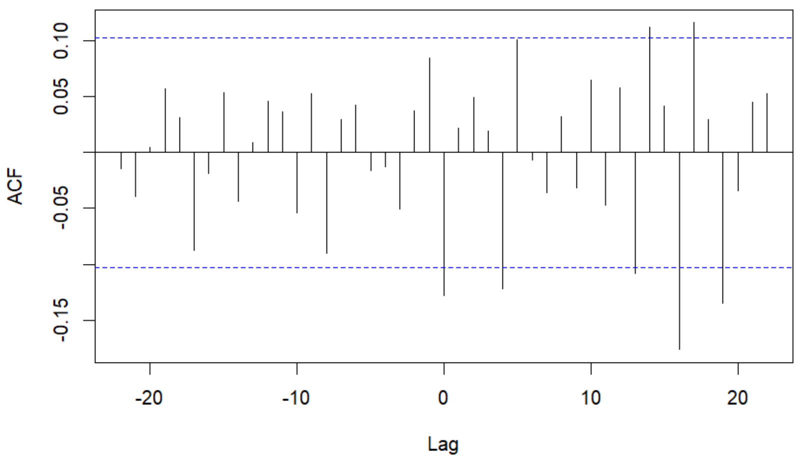


Varma Model is a valuable tool for understanding the complex interdependencies among time series and for making predictions about the future behavior.

In VARMA model with both the temperature and relative humidity are the cross coefficients represent the causal relationship between temperature and relative humidity as well as their interaction with other variables in the model

The coefficients for this model took the number of parameters as 30 and we see the strong coefficient at 14th parameter from VARorder.

The Cross Correlation between the Temperature and the Relative Humidity in Cross Correlation plot defines the below:




```
Coefficient(s):
              Estimate  Std. Error  t value Pr(>|t|)
diff.temp.  -0.013159    0.022107   -0.595   0.5517
diff.rh.    -0.045878    0.189674   -0.242   0.8089
diff.temp.   0.350689    0.341282    1.028   0.3042
diff.rh.     0.004466    0.013144    0.340   0.7341
diff.temp.   0.094624    0.192728    0.491   0.6234
diff.rh.     0.022558    0.023821    0.947   0.3437
diff.temp.   0.049306    0.060535    0.815   0.4154
diff.rh.     0.005736    0.015044    0.381   0.7030
diff.temp.  -0.019556    0.066094   -0.296   0.7673
diff.rh.    -0.021244    0.015557   -1.366   0.1721
diff.temp.  -0.007386    0.063643   -0.116   0.9076
diff.rh.     0.034676    0.016927    2.049   0.0405 *
diff.temp.  -0.608710    1.097986   -0.554   0.5793
diff.rh.    -0.350870    0.053714   -6.532 6.48e-11 ***
```

```
Residuals cov-matrix:
           [,1]      [,2]
[1,]  5.862703 -2.289615
[2,] -2.289615 97.014563
----
aic=  6.488482
bic=  6.788877
```

The AIC value on the VARMA model with order AR(7) is 6.488482.

When VARMA with order (1,1) which is AR(1) MA(1) for Multivariate gives the coefficients with better significance with the coefficients.

```
Coefficient(s):
                           Estimate  Std. Error  t value Pr(>|t|)
diff.temperature.          0.543201    0.067427    8.056 8.88e-16 ***
diff.relative_humidity.    0.009773    0.015373    0.636   0.5250
diff.temperature.          0.606122    0.336024    1.804   0.0713 .
diff.relative_humidity.    0.552622    0.054063   10.222  < 2e-16 ***
                          -0.854713    0.046033  -18.568  < 2e-16 ***
                          -0.004218    0.010208   -0.413   0.6794
                          -0.299217    0.190644   -1.570   0.1165
                          -0.946473    0.023283  -40.651  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
---

Estimates in matrix form:
AR coefficient matrix
AR( 1 )-matrix
      [,1]    [,2]
[1,] 0.543 0.00977
[2,] 0.606 0.55262
MA coefficient matrix
MA( 1 )-matrix
      [,1]    [,2]
[1,] 0.855 0.00422
[2,] 0.299 0.94647

Residuals cov-matrix:
          [,1]      [,2]
[1,]  5.852501 -2.678774
[2,] -2.678774 92.630145
----
aic=  6.326236
bic=  6.412063
```

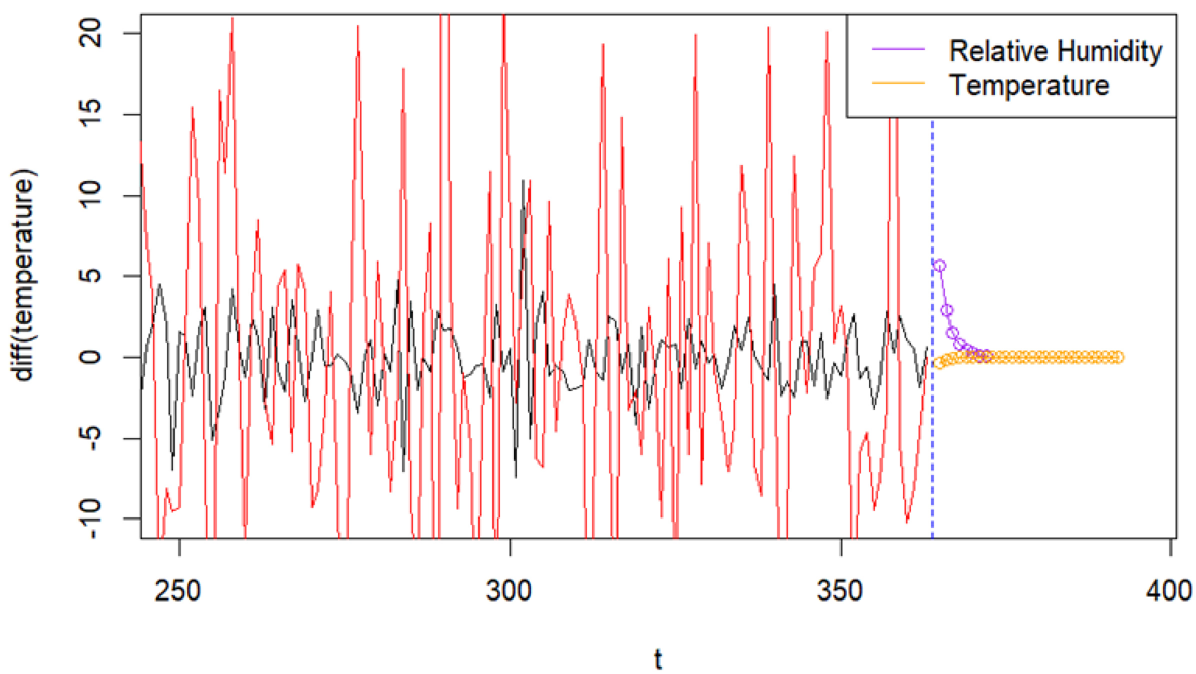
In comparison from the from VARMA(7,0) and VARMA(1,1) gives the efficient from the when comparing with AIC values So the lowest AIC from VARMA(7,0) and VARMA(1,1)

VARMA (7,0)	VARMA (1,1)
AIC = 6.4884	AIC = 6.3262

BIC = 6.7888	BIC = 6.4120

So, the best model from the comparison of AIC evaluated 6.3262 which is VARMA (1,1).

The prediction with best model which is VARMA(1,1) on the Relative Humidity and Temperature is indicating the black line indicates temperature and red line indicates the Relative Humidity on actual data, and indicating purple line is Relative Humidity and Orange line is Temperature represents prediction on VARMA model.

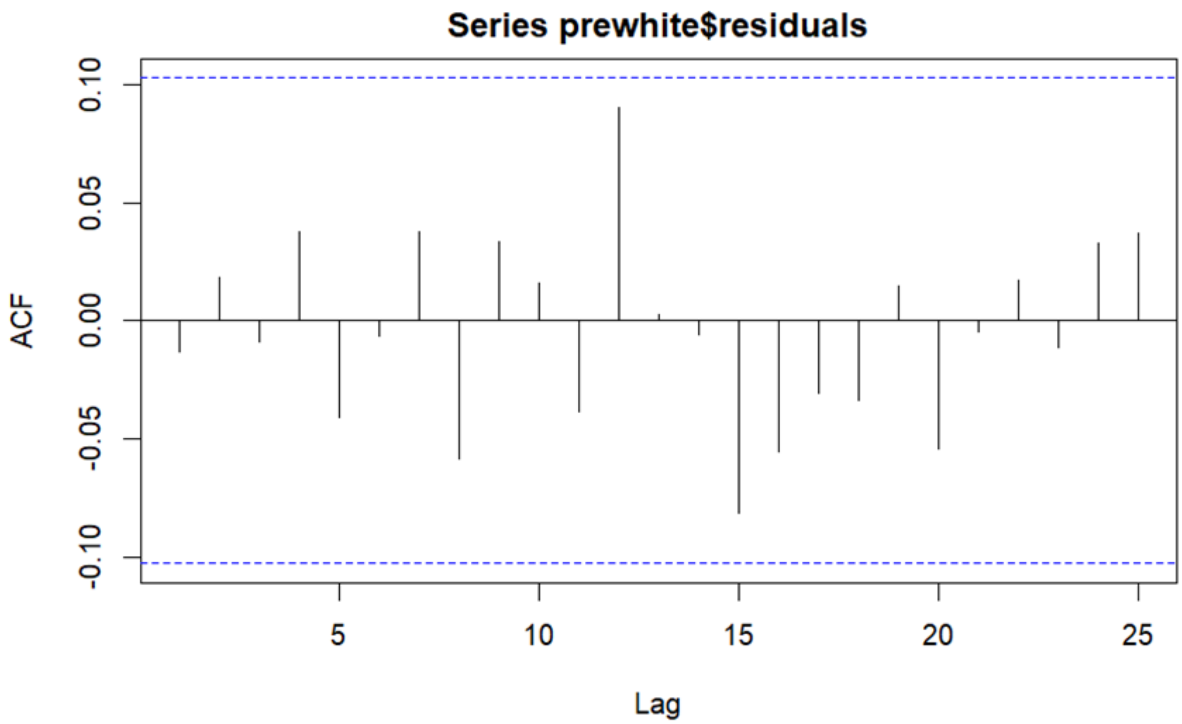


Transfer Function(TF) model:

Transfer function model provides a flexible and powerful framework for modeling multivariate time series.

The pre-whiten with the input variable is chosen as Relative Humidity and the and the output variable as Temperature.

The pre-whiten of the Relative-Humidity with AR (1) and MA (1) and the ACF shows as below



The coefficient on the pre-whitening on Relative Humidity is evaluated by the AIC as 2683.46.

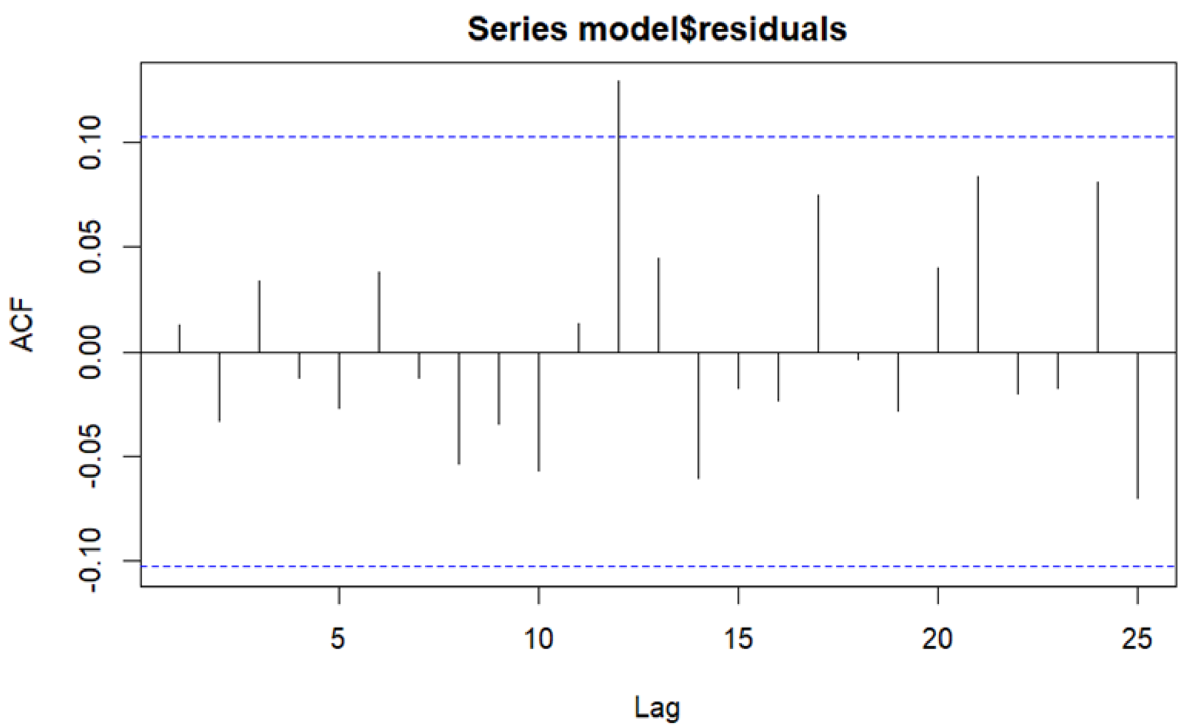
```
Call:
arima(x = diff(relative_humidity), order = c(1, 0, 1))

Coefficients:
         ar1      ma1  intercept
      0.5503  -0.9362    -0.0028
s.e.  0.0539   0.0223     0.0748

sigma^2 estimated as 93.25:  log likelihood = -1338.73,  aic = 2683.46
```

we fit a Transfer Function (TF) model using the arimax() function from the "forecast" package, where the "Temperature" variable is the dependent variable, the "Relative Humidity" variable is the independent variable, and we specify an AR(1) and MA(1) transfer function with a coefficient.

```
Call:
arimax(x = Yn, order = c(1, 0, 1), include.mean = TRUE, xtransf = data.frame(Xn),
    transfer = list(c(1, 0)))

Coefficients:
         ar1     ma1  intercept   Xn-AR1   Xn-MA0
      0.5782  -0.880    -0.0068  -0.0004  -0.0281
s.e.  0.0610   0.032     0.0365   0.3581   0.0130

sigma^2 estimated as 5.767:  log likelihood = -833.32,  aic = 1676.64
```

The AIC with Transfer Function model estimated AIC as 1676.64.

We check the residuals on the model whether the model is white noise or not. We used the Ljung-Box test for estimating the residuals in the model.

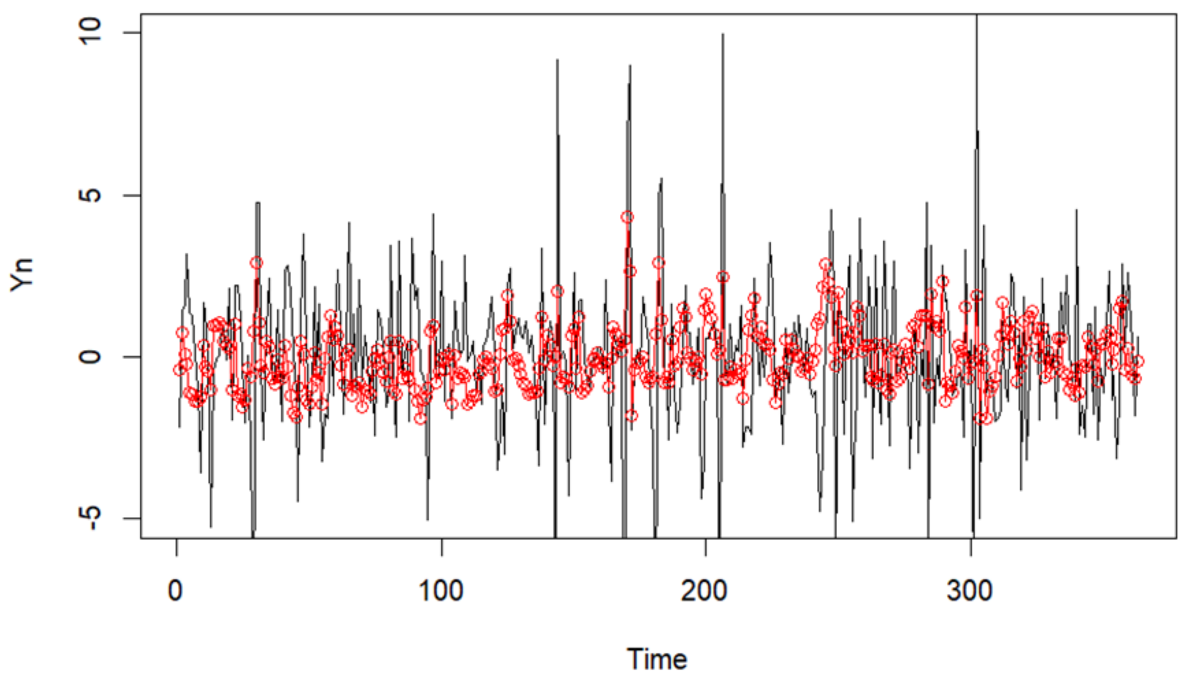



```
	Box-Ljung test

data:  model$residuals
X-squared = 0.059768, df = 1, p-value = 0.8069
```

As we can observe in the Ljung-Box test on finding the residuals on TF model the p-value as 0.8069. which is greater than 0.05 by the significance level. Which implies it has white noise as we fail to reject the null hypothesis.

To perform a corrected model on the transfer model it isn't required as the transfer model we built is having White Noise.

This graph indicates whether the model is fitted on the actual data.



We can observe the fitted model is underfitted to the actual data.

The Mean Absolute Percentage Error on VARMA and TF model as below

VARMA Model	TF Model
MAPE: 29% [1] 0.292254	MAPE: 30% [1] 0.3049856

We can conclude with the evidence that the VARMA model is the best Model in predicting the Multivariate time series.

Future Scope and Conclusion:

- After estimating regression, deterministic and stochastic time series models, ARIMA model gives the white noise for the series.

- By comparing the MAPE values of all models, MLR value gives the lowest Mean Absolute percentage error value.

- Multivariate has the MAPE of 29% in combination of Temperature and Relative Humidity with the VARMA model.

- Predict the temperature changes by including the other factors that affect the temperature such as weather patterns, geography, human activities etc.

- Would like to explore different time periods of the same data sets and compare how the results are changing.

- If data is not imbalanced we would get the good fit of the model.