# The Next Panademic Could be Hiding in the Permafrost- Act Fast on Climate Change

Sung Hee(Sunny),Hong

04 May, 2020

## Abstract or Executive Summary

The real key motivation of this project is to be educated about an issue that does not only impact my community directly but impacts the future of humanity. As today (May 3, 2019) marks as a day in the pandemic of corvid-19, it has come of my conscious to work on a project related to it. Although permafost is unlikely the cause of coronavirus, it is one of the known ways of how such deadly virus can appear and spread like fire. In conclusion, I have decided to work on a small scale project to simply predict the climate change in Delhi, India. I believe learning the change of such trend and to forecast climate will help determine future climate expectations and prepare the extend of impact.

The questions I addressed were:

- How much impact does this measurement of the average change in the climate will impact humanity and to what degree?

In order to execute this project it required my knowledge of time series such as the understanding of ACF and ACF datas, differencing, transformation, modeling, and residual diagnostic testing. Key results I found was.

# Main Body

## *Introduction*

This project will focus on forecasting the future mean temperature in Delhi, India to help further the investigation of the shift of average. The Kaggle dataset I obtain provides columns of Date from January 1st 2013 to April 24th 2017 in the city of Delhi, India. The 4 parameters this data provides are meantemp (mean temperature), humidity, "wind_speed", and "meanpressure". It is important to note that this is a daily collect dataset, which will allow us to extract seasonality.

The dataset has been collected by Weather Underground API. Tools such as histogram, time series plot, ACF, PACF, and transformation will be used for model identification and model estimation.

The final model I ended up with is SARIMA (0,0,2) X $(1,0,1)_{12}$. Briefly state the conclusion:

I used a software language (R) to proceed this project from the start to end.

# Sections

## *Splitting Train Test*

Before doing data exploratory analysis, I decided to split the data into a training dataset and testing data set. The training data set contains total observation of 1,462. The testing data set contains 114 rows. In addition I compressed the daily mean temperature data to monthly mean temperature data. By compressing I ended up with 49 observation of train data and 4 rows of test data. Further analysis will help identify important features on a time series plot, interpret ACF and PACF, identify a stationary time series, and identify how to take differences (transformation).

## *Seasonality*

There is a regularly repeating pattern of highs and los related to months. There is upward trend from the months Januaray to June. From the end of June to December there is a decrease in trend.

### *Trend*

We can make few observations by looking at the time series plot of the training data and (figure 1). In figure 1, we can observe that there is a visible trend. On average by season, the measurements tend to either increase of decrease.Meaning, on average the measurements tend to rop or jump more. We can observe unequal variance as we can distinguish more drastic drops and drastic jumps by season.

Fig1:Train Time Series

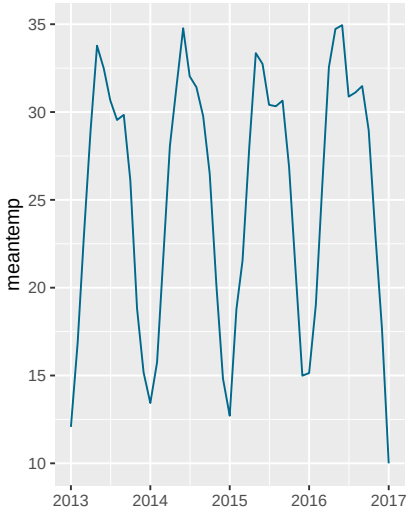
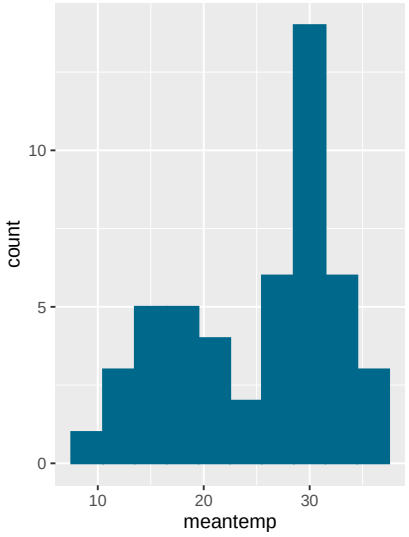



Fig2: Train Histogram



### *Stationarity*

In order to use ACF and PACF to identify possible structure of time series data, the data must be stationary. The mean should be same for all time. The variace of mean temperature should be same for all time. The covariance between mean temperature at t and lag t-1 is the same for all t.

The histogram shown in figure 2 does not outline a bell-curve, further confirming the nonstationarity of the original data.(Figure 3, Left) The autocorrelation function does not cut off but decays very slowly and remains above the confidence interval range (above the blue dotted lines). This is a strong indication of a non-stationary series.

**Figure 3: Original Train Data:**

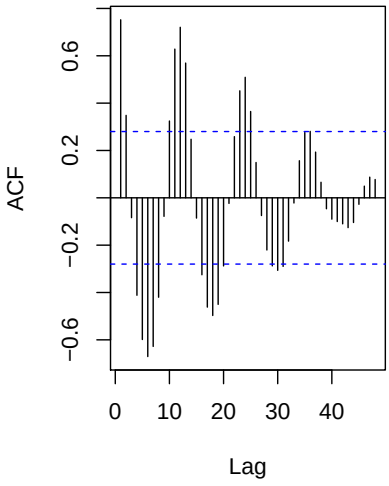
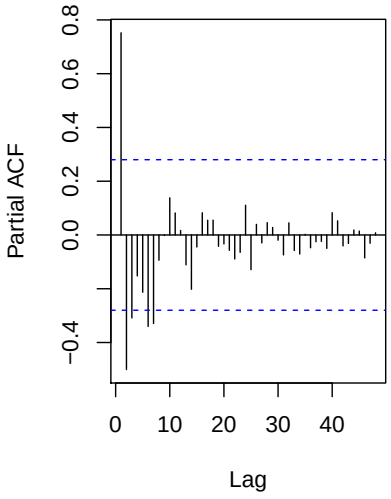



**Figure 4:Original Train Data:**



## *Box-Cox Transformation*

Box-Cox transformation is used to stabilize the variance and attempt to make data approximately normal. Figure 4 shows that instead of guessing *lambda*, it chooses the maximum log-likelihood that will produce the best transformation to the dependent variable (total number of cars parked) to minimize the variance. In this case, $\lambda = 1.676768$.

Analysis of the transformation...

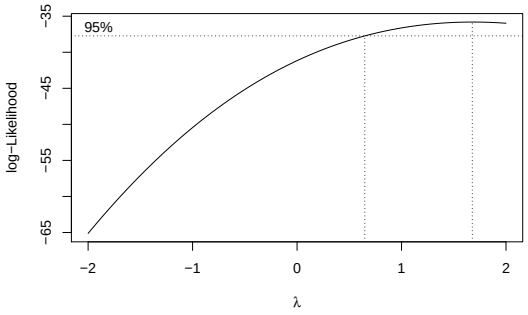



Figure 1:Original data

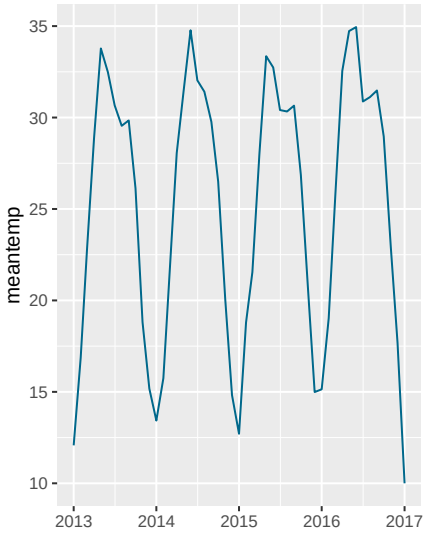



Figure 5:Box−Cox tranformed data

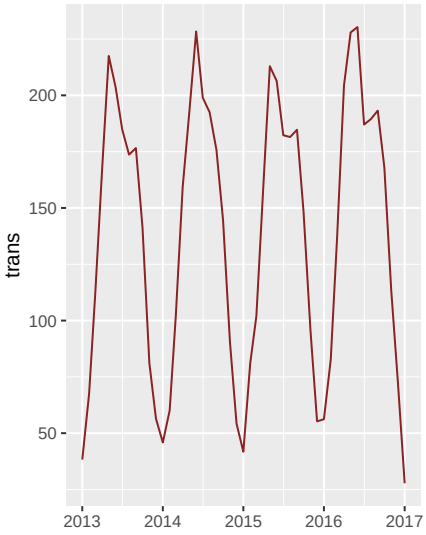


Fig2: Train Histogram

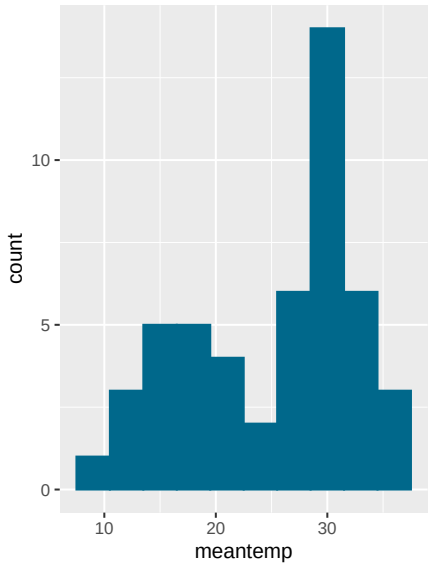



Fig6: Transformed Histogram

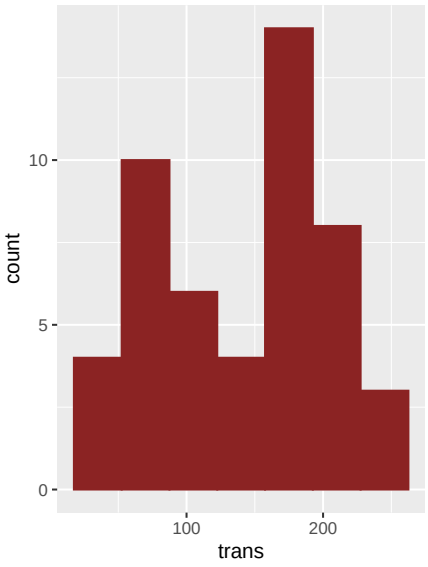



**Figure 6: Trans Train Data:**

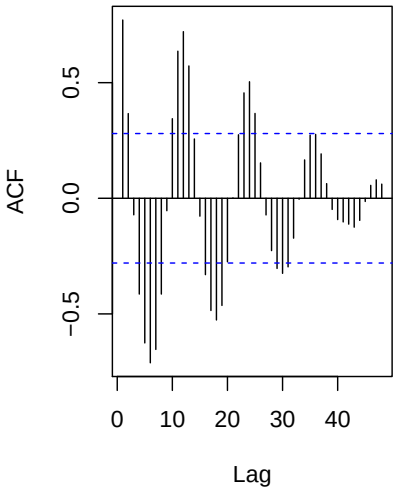
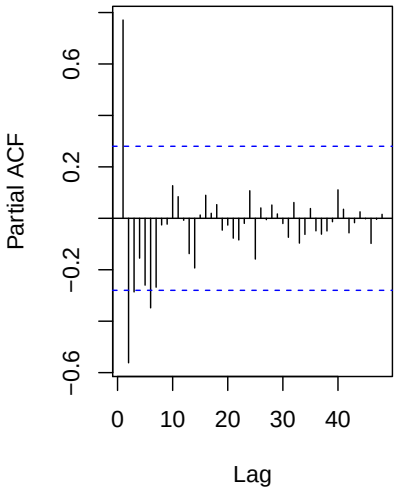



**Figure 7:Trans Train Data:**



## *Seasonality and Differencing*

Figure 7: Diff_12 Time Series

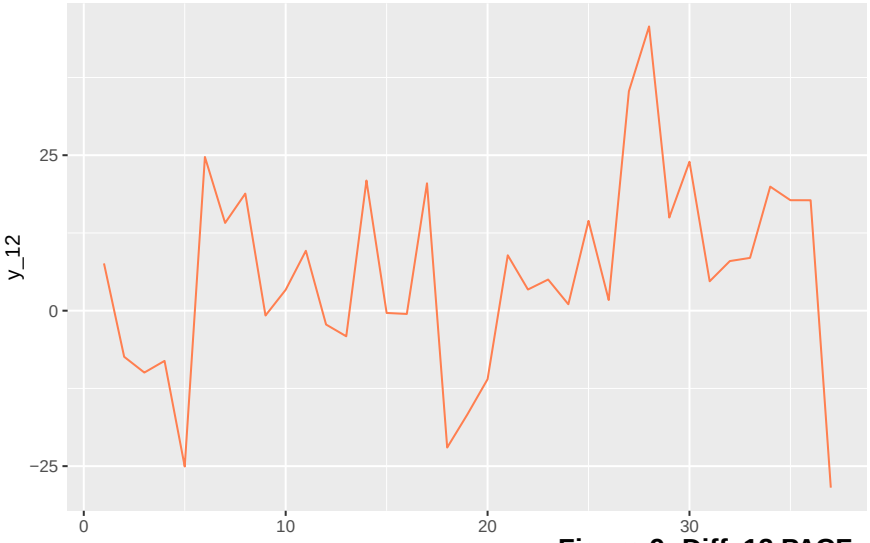



**Figure 8: Diff_12 ACF :**

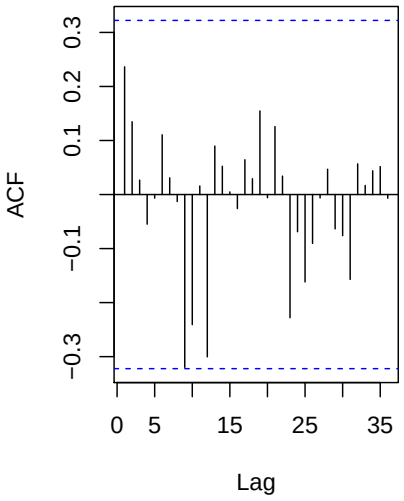
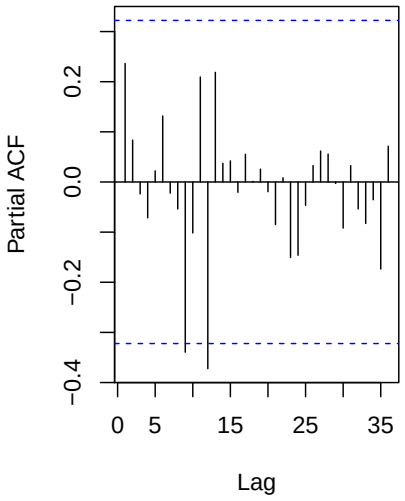



**Figure 9: Diff_12 PACF :**




```
## [1] 254.3784
```

## *Modeling*

*Modeling the non-seasonal part (p,d,q):*

In this case focus on the within season lags h=1,. . . ,11 We applied one differencing to remove the trend: d=0 * ACF seems to be tailing off. A good choice for the MA could be q=0

- PACF cuts off at lag 1,2 A good choice for the AR part could be p=1 or 2

**Figure 6: Transformed Train Data**

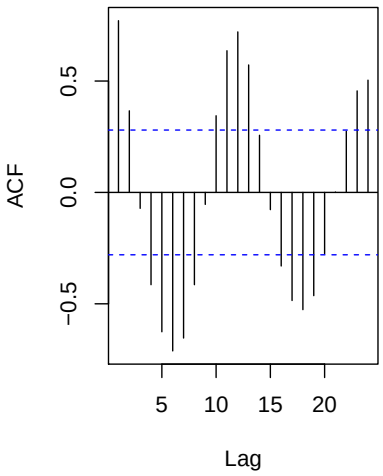
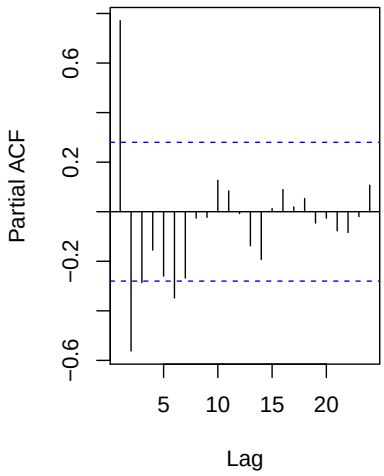



**Figure 7:Transformed Train Data**



*Modeling the seasonal part (P,D,Q):*

For this part, focus on the seasonal lags, h=1,2,3, etc. Seasonaal differencing so D=0,s=12:

- The ACF shows strong peak at h=9: A good choice for the MA part could be Q=9
- PACF shows tow strong peaks at h= 11 A good choice for the AR part could be P=1,11

**Figure 8: Diff_12 ACF :**

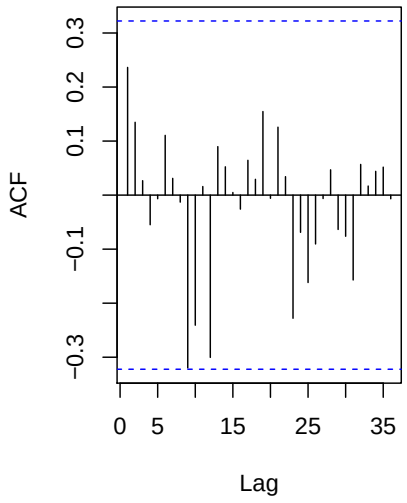
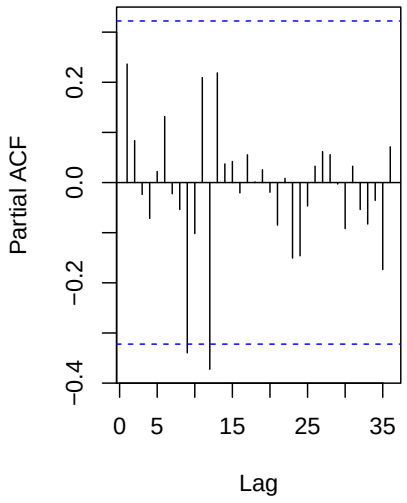



**Figure 9: Diff_12 PACF :**




## AIC Checking & Modeling

```
##
## Call:
## arima(x = train_ts_trans, order = c(0, 0, 2), seasonal = list(order = c(1, 0,
##     1), period = 12), method = "ML")
##
## Coefficients:
##          ma1     ma2    sar1     sma1  intercept
##       0.6337  0.4289  0.9999  -0.9514   140.2446
## s.e.  0.1408  0.1612  0.0011   0.1944    19.6619
##
## sigma^2 estimated as 124.9:  log likelihood = -208.66,  aic = 427.32
```

## *Diagnostic Checking / Analysis of Residual*

Plot of residual should resemble WN (no trend, no seasonality, no change of variance, etc)
* step 1: plot residual: should resemble WN * step 2: Plot a histogram of residual: should resemble Gaussian * Step 3: Examine Normal QQ plot- should be close to straight line * Step 4: Tan Shapir Wilk test of normality
* Step 5: Check Sample ACF/PACF of residual: should resemble WN
* Step 6: Use Yule-Walker Estimation : should fit into AR(0)
* Box Pierce Test, Ljung-Box Test, h= $\sqrt{n}$= $\sqrt{188}$=

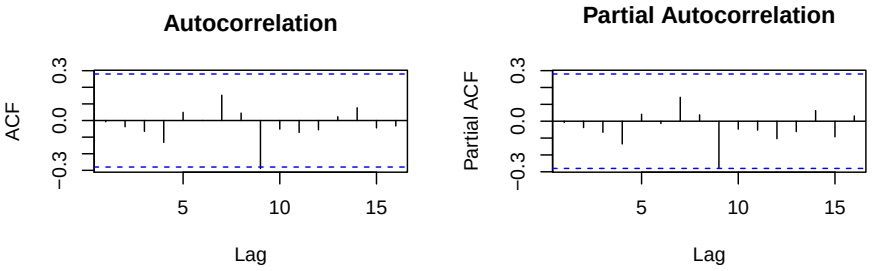
___Final Model: SARIMA (0,0,2) X (1,0,1)_12 ___

```
## Warning in stats::arima(x = x, order = order, seasonal = seasonal, xreg =
## xreg, : possible convergence problem: optim gave code = 1
```

```
##
## Call:
## arima(x = train_ts_trans, order = c(0, 0, 2), seasonal = list(order = c(1, 0,
##     1), period = 12), method = "CSS")
##
## Coefficients:
##          ma1     ma2    sar1     sma1  intercept
##       0.2770  0.3133  0.9971  -0.9737   1403.195
## s.e.  0.1567  0.1680  0.0051   0.1476   2224.688
##
## sigma^2 estimated as 143.5:  part log likelihood = -191.2
```

```
## [1] -0.1033715
```

```
## [1] 110.6042
```


```
##
##  Box-Pierce test
##
## data:  res
## X-squared = 7.1136, df = 12.7, p-value = 0.8837
```

```
##
##  Box-Ljung test
##
## data:  res
## X-squared = 8.8568, df = 12.71, p-value = 0.765
```

```
##
##  Box-Ljung test
##
## data:  res^2
## X-squared = 9.7411, df = 13.71, p-value = 0.7629
```

All the test passes. Because

## *Residual Analysis*

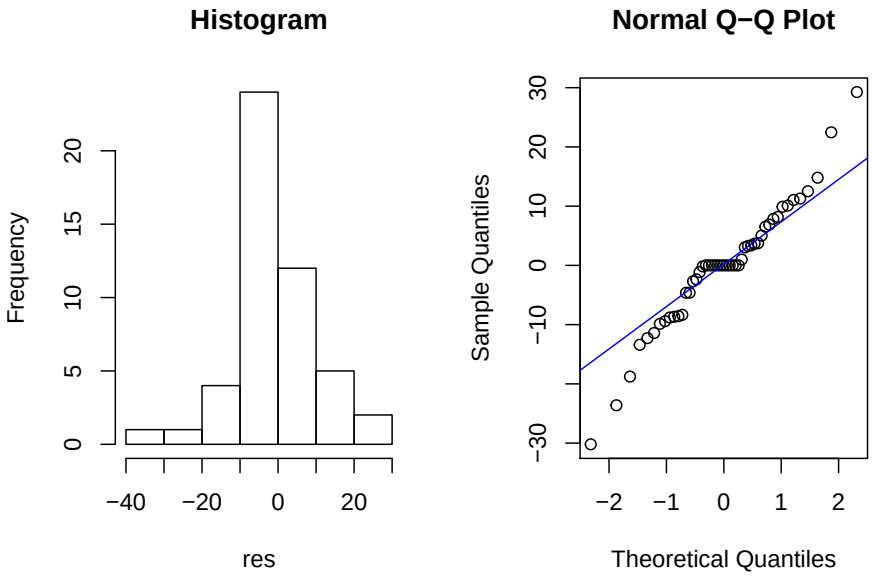


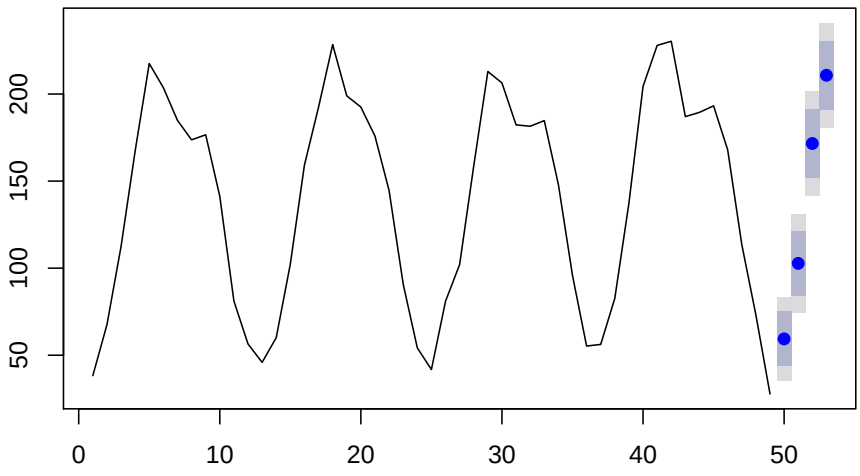

Figure 1: Figure 17

## *Data Forecasting*

**Final Model:**

```
##
## Call:
## arima(x = train_ts_trans, order = c(0, 0, 2), seasonal = list(order = c(1, 0,
##     1), period = 12), method = "ML")
##
## Coefficients:
##          ma1     ma2    sar1     sma1  intercept
##       0.6337  0.4289  0.9999  -0.9514   140.2446
## s.e.  0.1408  0.1612  0.0011   0.1944    19.6619
##
## sigma^2 estimated as 124.9:  log likelihood = -208.66,  aic = 427.32
```

**Forecasting:**



## *Conclusion*

The objective of this project is to predict and forecast future monthly climate changes in Delhi, India. In order to stabelize the variance I used an Box-Cox transformation method. The time series, ACF, and PACF plots of the transformed data showed no specific trend but an seasonality. Then I went on to proceed to deduct the seasonality. I removed the seasonality by 12 month period. The modeling process takes analysis of both exploration of the transformed time series, ACF, PACF plots as well as the seaonality deducted time series, ACF, and PACF plots. In conclusion to analyzing AIC of different models, the best model I concluded with is SARIMA (0,0,2) X (1,0,1)_12 . The forecasted model was successful. The values I forecasted are relatively close to original data. After running through the residual diagnostic test, because all p values where over the significant level of 0.05, I failed to reject the normality. In conclusion the prediciton was within the confidence interval proving the model worked.

*mathematical formula of this model:*

```
## Time Series:
## Start = 50
## End = 53
## Frequency = 1
##    testdat$meantemp forecast_value lower_CI.80% lower_CI.95% upper_CI.80%
## 50         15.71087       15.63849     13.06030     11.54596     17.95572
## 51         18.34998       21.63376     19.23256     17.87579     23.86651
## 52         23.75376       29.33566     27.29051     26.16485     31.28851
## 53         30.75366       33.14963     31.27553     30.25208     34.95459
##    upper_CI.95%
## 50     19.10339
## 51     24.99276
## 52     32.28950
## 53     35.88494
```