Final Report

Topic: Global Warming

INFO6105 41642 Data Sci Engineering Methods SEC 01 Summer 1 2021

Professor: Dr. Handan Liu

Team number: 11

Submitted by:
Maharshi Jinandra

Submission Date:
June 26, 2021

Table of Contents

INTRODUCTION

Rising global average temperature is associated with widespread changes in weather patterns. Scientific studies indicate that extreme weather events such as heat waves and large storms are likely to become more frequent or more intense with human-induced climate change.

Long-term changes in climate can directly or indirectly affect many aspects of society in potentially disruptive ways. For example, warmer average temperatures could increase air conditioning costs and affect the spread of diseases like Lyme disease, but could also improve conditions for growing some crops. More extreme variations in weather are also a threat to society. More frequent and intense extreme heat events can increase illnesses and deaths, especially among vulnerable populations, and damage some crops. While increased precipitation can replenish water supplies and support agriculture, intense storms can damage property, cause loss of life and population displacement, and temporarily disrupt essential services such as transportation, telecommunications, energy, and water supplies.

Weather forecasting is the application of science and technology to predict the conditions of the atmosphere for a given location and time. People have attempted to predict the weather informally for millennia and formally since the 19th century. Weather forecasts are made by collecting quantitative data about the current state of the atmosphere, land, and ocean and using meteorology to project how the atmosphere will change at a given place.

METHODOLOGY

As stated above, our goal is to come up with a successful model to forecast temperature for the defined place and time stamp.

For prediction of weather at regional and national levels, various precipitation forecasting methods are available. Regression analysis, Auto Regression Integrated Moving Average (ARIMA), genetic algorithm, Adaptive Splines Threshold Autoregressive (ASTAR), Support Vector Machines (SVMs), K-nearest neighbor (K-NN) are among the best methods available for weather forecasting. Regression analysis determines the strength of relation between a dependent variable and a series of independent predictor variables by fitting a regression model. The regression analysis is called multiple regression analysis, if it caters to more than two predictor variables. However, this regression analysis is not recommended for most of the practical problems as it tends to oversimplify the real-world situations. ARIMA model forecasts weather variables which are kind of time series data by linearly combining their historic values. ARIMA model as a tool deals with all the aspects related to univariate time series model identification and its parameter estimation and forecasting. ARIMA model has the chances of over-fitting and misidentification if not used carefully. Genetic algorithm approach applies concept of natural evolution in problem solving. It combines a random wide selection of solutions within a given population. Then it evaluates them to get best solution by repeatedly simulating the process of evolution. This entire process gives the desired number of generations. The succeeding generation is an improvement upon the preceding one and in the end, best solution to the problem is obtained. Genetic algorithm can find good quality solution in a short computation time, but it cannot guarantee for an optimal solution to the problem in hand.

The ARIMA model is a modification of the ARMA model. We would be taking a statical approach to forecasting the temperature by using the ARMA model.

There are many flavors to this ARMA model let us understand the basic one first.

AR – Autoregressive Model:

AR(p) can be described as:

$$Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + Z_t, \qquad (1)$$

Where $Z_t \sim (0, \sigma^2)$, c is an unknown constant term, and equal to 1, are the parameters of the AR model.

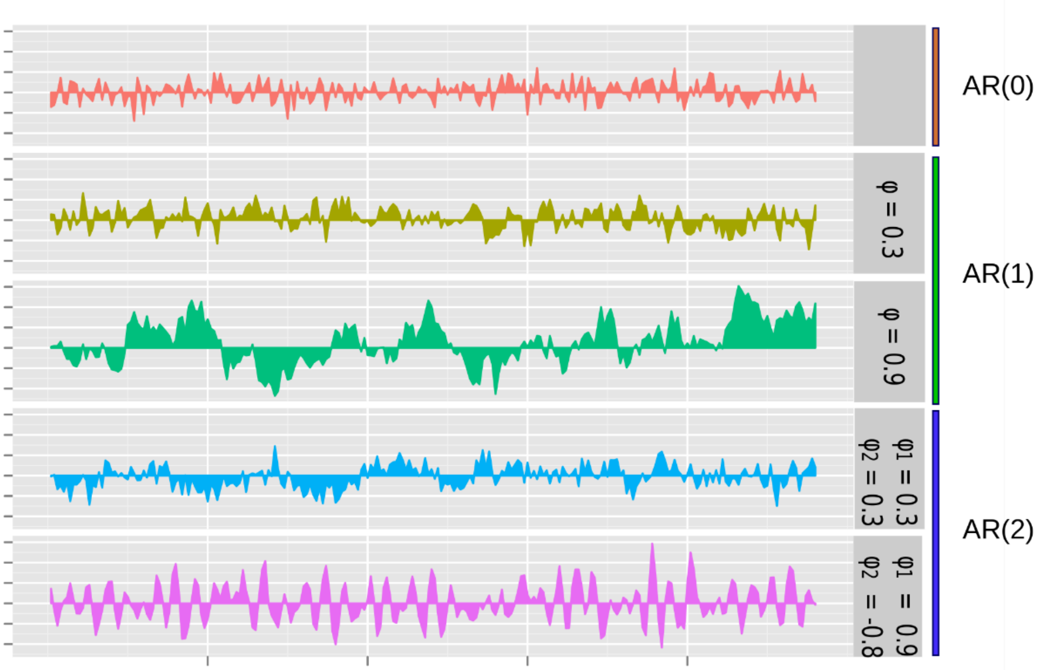
An autoregressive model can thus be viewed as the output of an all-pole infinite impulse response filter whose input is white noise.

Some parameter constraints are necessary for the model to remain wide-sense stationary. For example, processes in the AR(1) model with $|\varphi_1| \geq 1$ are not stationary.

The simplest AR process is AR(0), which has no dependence between the terms. Only the error/innovation/noise term contributes to the output of the process, so in the figure, AR(0) corresponds to white noise.

For an AR(1) process with a positive φ, only the previous term in the process and the noise term contribute to the output. If φ is close to 0, then the process still looks like white noise, but as φ approaches 1, the output gets a larger contribution from the previous term relative to the noise. This results in a "smoothing" or integration of the output, similar to a low pass filter.

For an AR(2) process, the previous two terms and the noise term contribute to the output. If both φ1 and φ2 are positive, the output will resemble a low pass filter, with the high frequency part of the noise decreased. If φ1 is positive while φ2 is negative, then the process favors changes in sign between terms of the process. The output oscillates. This can be likened to edge detection or detection of change in direction.



MA – Moving Average:

MA(q) refers to the moving average model of order q:

$$Y_t = c + Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \qquad (2)$$

Where $Z_t \sim (0, \sigma^2)$, c is an unknown constant term, and θ_1, ..., θ_q are the parameters of the model, and the $\varepsilon_t, \varepsilon_{t-1}$, are white noise error terms.

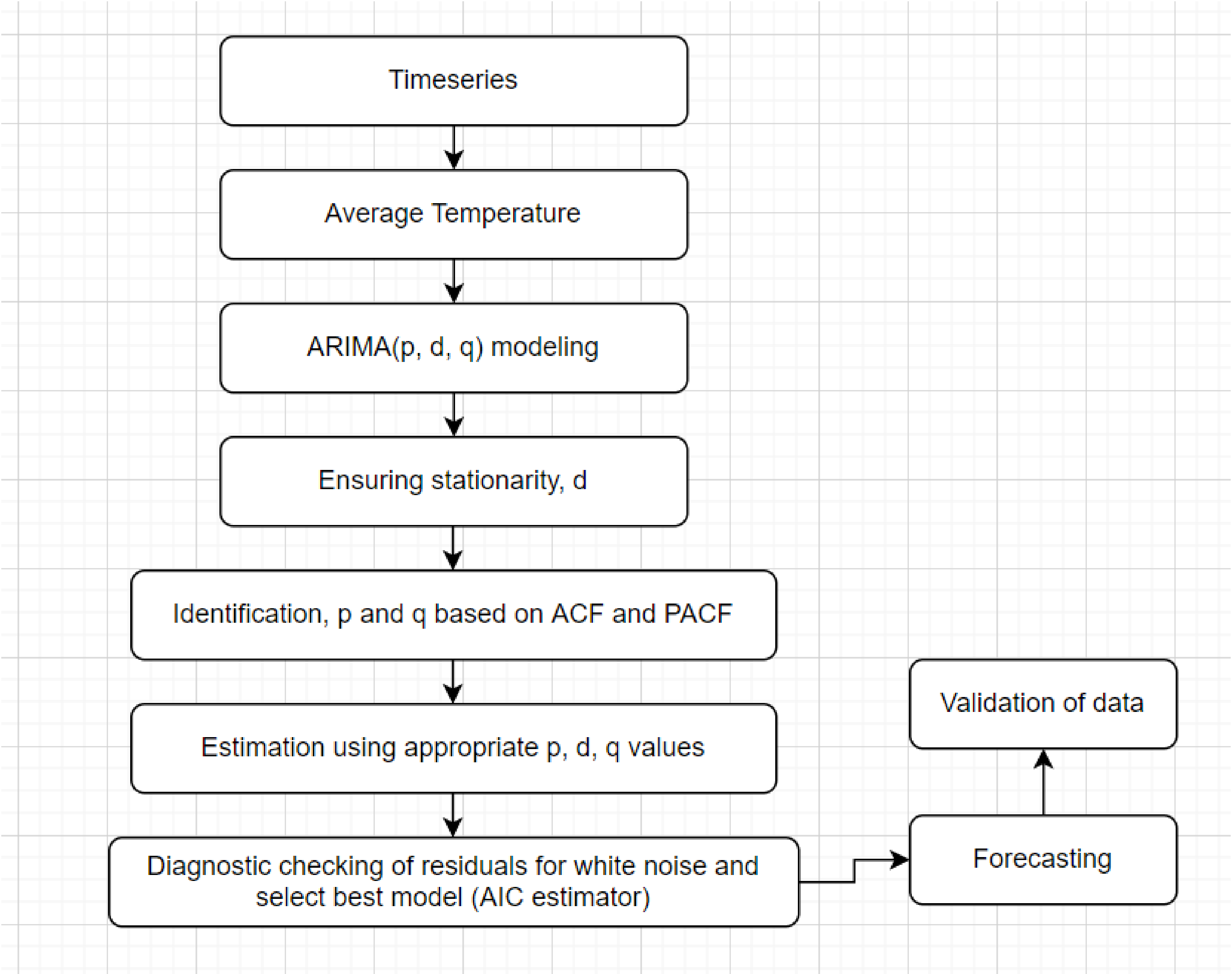
In time series analysis, to better understand the data and for future forecasting, auto-regressive (p) integrated (d) moving average (q) (ARIMA) model is used. The basic idea of using ARIMA model is to remove trend of the series by differencing so that a stationary series is obtained by transforming a non-stationary series (P, D, Q)m, where m is the number of periods in each season, and the uppercase P, D, Q refer to the autoregressive (AR), differencing (I), and moving average (MR) terms respectively, for the seasonal part of the ARIMA model. ARIMA methodology has its own limitations of relying on past values; however, it works best for long and stable time series. It does not explain the structure of the underlying data mechanism but simply approximates the historical patterns.

If we combine the differencing with ARMA model, we get the autoregressive integrated moving average model, i.e., the ARIMA(p, d, q), where d is the order of differencing. So, an ARIMA model corresponds to an ARMA after differencing Yt, d times. This means that Yt satisfies the difference equation

$$\begin{aligned}&\left(1 - \varphi_1 B - \cdots \varphi_p B^p\right)(1 - B)^d Y_t \\ &\quad = c + \left(1 + \theta_1 B + \cdots \theta_q B^q\right) Z_t,\end{aligned}$$

$$\varphi(B)(1 - B)^d Y_t = c + \theta(B) Z_t, \quad Z_t \sim \left(0, \sigma^2\right)$$

The below image summarizes the methodology


Timeseries
Average Temperature
ARIMA(p, d, q) modeling
Ensuring stationarity, d
Identification, p and q based on ACF and PACF
Estimation using appropriate p, d, q values
Diagnostic checking of residuals for white noise and select best model (AIC estimator)
Forecasting
Validation of data

Autocorrelation function (ACF)

The autocorrelation function is a measure of the correlation between observations of a time series that are separated by k time units (y_t and y_{t-k}).

Partial autocorrelation function (PACF):

The partial autocorrelation function is a measure of the correlation between observations of a time series that are separated by k time units (y_t and y_{t-k}), after adjusting for the presence of all the other terms of shorter lag (y_{t-1}, y_{t-2}, ..., y_{t-k-1}).

To identify the appropriate SARIMA model following steps are followed.

Ensuring stationarity (d):

ADF — Augmented Dickey Fuller Test

The Augmented Dickey Fuller Test (ADF) is unit root test for stationarity. Unit roots can cause unpredictable results in your time series analysis.
It uses an autoregressive model and optimizes an information criterion across multiple different lag values.
Null Hypothesis: The series has a unit root. **(not stationary)**
Alternate Hypothesis: The series has no root. **(is stationary)**

The first and foremost step is to determine the order of differencing (d) to stationarise the series. The order of differencing (d) is selected such that it minimizes the standard deviation. This is done by fitting different ARIMA models having various orders of differencing, but a constant coefficient is selected. An already differenced series which is now a stationery series might still have some auto-correlated errors which can be removed by adding AR terms (p >= 1) and MA terms (q >= 1) in the forecasting equation. To compensate for any mild 'under-differencing', AR terms are added to the model, while to compensate any mild 'over-differencing', MA terms are added instead.

Estimations using appropriate p, d, q values:

ARIMA models are estimated for values of p, d, q which are given most appropriate residuals.

Selecting best SARIMA model:

The most appropriate model is determined once all the model parameters are calculated (p, d and q, P, D and Q). The residuals of estimated SARIMA models are checked for white noise and most well-behaved residual model is selected. The selected SARIMA models are tested under AIC criteria for its ability to evaluate the relative quality of statical model for a given dataset.

Akaike Information Criterion (AIC) is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the true likelihood function of the model, so that a lower AIC means a model is considered to be closer to the truth. In other words, AIC estimates the relative amount of information lost by a given model, i.e., the less information a model loses, the higher the quality of that model.

$$\mathrm{AIC} \;=\; 2k - 2\ln(\hat{L})$$

Rolling Forecast Origin:

It is important to evaluate forecast accuracy using genuine forecasts. Consequently, the size of the residuals is not a reliable indication of how large true forecast errors are likely to be. The accuracy of forecasts can only be determined by considering how well a model performs on new data that were not used when fitting the model.

The following points should be noted.

- A model which fits the training data well will not necessarily forecast well.
- A perfect fit can always be obtained by using a model with enough parameters.
- Over-fitting a model to data is just as bad as failing to identify a systematic pattern in the data.

Rolling-forecast type model is required from the problem definition. This is where one-step forecasts are needed given all available data.

The walk-forward validation will work as follows:

- The first part of the dataset will be held back to train the model.
- The remaining part of the dataset will be iterated and test the model.
- For each step in the test dataset:
    - A model will be trained.
    - A one-step prediction made and the prediction stored for later evaluation.
    - The actual observation from the test dataset will be added to the training dataset for the next iteration.
- The predictions made during the iteration of the test dataset will be evaluated and an RMSE score reported

DATASET

[https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data/download]

GlobalLandTemperaturesByCity.csv **(508.15 MB)**

```
DatetimeIndex: 8599212 entries, 1743-11-01 to 2013-09-01
Data columns (total 6 columns):
 #   Column                         Dtype
---  ------                         -----
 0   AverageTemperature             float64
 1   AverageTemperatureUncertainty  float64
 2   City                           object
 3   Country                        object
 4   Latitude                       object
 5   Longitude                      object
dtypes: float64(2), object(4)
memory usage: 459.2+ MB
```

GlobalLandTemperaturesByCountry.csv **(21.63 MB)**

```
DatetimeIndex: 577462 entries, 1743-11-01 to 2013-09-01
Data columns (total 3 columns):
 #   Column                         Non-Null Count   Dtype
---  ------                         --------------   -----
 0   AverageTemperature             544811 non-null  float64
 1   AverageTemperatureUncertainty  545550 non-null  float64
 2   Country                        577462 non-null  object
dtypes: float64(2), object(1)
memory usage: 17.6+ MB
```

GlobalTemperatures.csv **(201.05 KB)**

```
DatetimeIndex: 3192 entries, 1750-01-01 to 2015-12-01
Data columns (total 8 columns):
 #   Column                                     Non-Null Count  Dtype
---  ------                                     --------------  -----
 0   LandAverageTemperature                     3180 non-null   float64
 1   LandAverageTemperatureUncertainty          3180 non-null   float64
 2   LandMaxTemperature                         1992 non-null   float64
 3   LandMaxTemperatureUncertainty              1992 non-null   float64
 4   LandMinTemperature                         1992 non-null   float64
 5   LandMinTemperatureUncertainty              1992 non-null   float64
 6   LandAndOceanAverageTemperature             1992 non-null   float64
 7   LandAndOceanAverageTemperatureUncertainty  1992 non-null   float64
dtypes: float64(8)
memory usage: 224.4 KB
```

[https://www.kaggle.com/andradaolteanu/country-mapping-iso-continent-region?select=continents2.csv]

continents2.csv **(22.05 KB)**

```
RangeIndex: 249 entries, 0 to 248
Data columns (total 11 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Country                   249 non-null    object 
 1   alpha-2                   248 non-null    object 
 2   alpha-3                   249 non-null    object 
 3   country-code              249 non-null    int64  
 4   iso_3166-2                249 non-null    object 
 5   region                    248 non-null    object 
 6   sub-region                248 non-null    object 
 7   intermediate-region       107 non-null    object 
 8   region-code               248 non-null    float64
 9   sub-region-code           248 non-null    float64
 10  intermediate-region-code  107 non-null    float64
dtypes: float64(3), int64(1), object(7)
memory usage: 21.5+ KB
```

[https://raw.githubusercontent.com/melanieshi0120/COVID-19_global_time_series_panel_data/master/data/countries_latitude_longitude.csv]

countries_latitude_longitude.csv **(8.00 KB)**

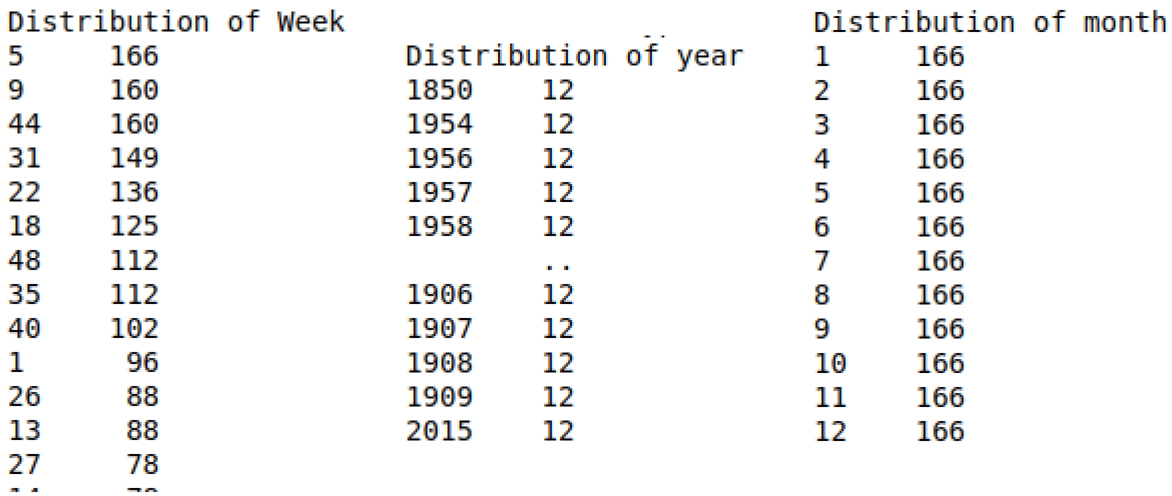
```
RangeIndex: 158 entries, 0 to 157
Data columns (total 3 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   name       158 non-null    object 
 1   latitude   158 non-null    float64
 2   longitude  158 non-null    float64
dtypes: float64(2), object(1)
memory usage: 3.8+ KB
```

RESULT AND ANALYSIS

Data Analysis

Data analysis is the process of exploring the data and performing cleansing, transforming and modeling data with a goal of discovering useful information.

Let's start with the 'GlobalTemperatures.csv' file, this file has a date field lets get to know the distribution of the time period. Looking at the distribution we will get know where we have flexibility to further explore the data.



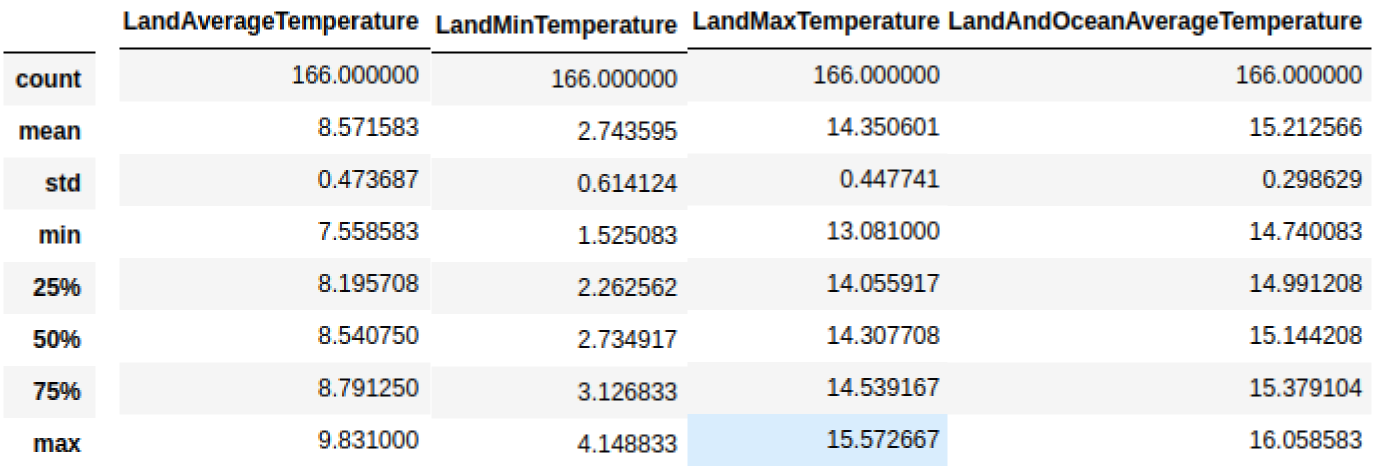
```
Distribution of Week
5      166
9      160
44     160
31     149
22     136
18     125
48     112
35     112
40     102
1       96
26      88
13      88
27      78
```

```
Distribution of year
1850     12
1954     12
1956     12
1957     12
1958     12
         ..
1906     12
1907     12
1908     12
1909     12
2015     12
```

```
Distribution of month
1      166
2      166
3      166
4      166
5      166
6      166
7      166
8      166
9      166
10     166
11     166
12     166
```

As seen the images above the weekly distribution is not evenly spread out, whereas the yearly and monthly are evenly distributed.

Next let's describe the day and see what we can infer from it.



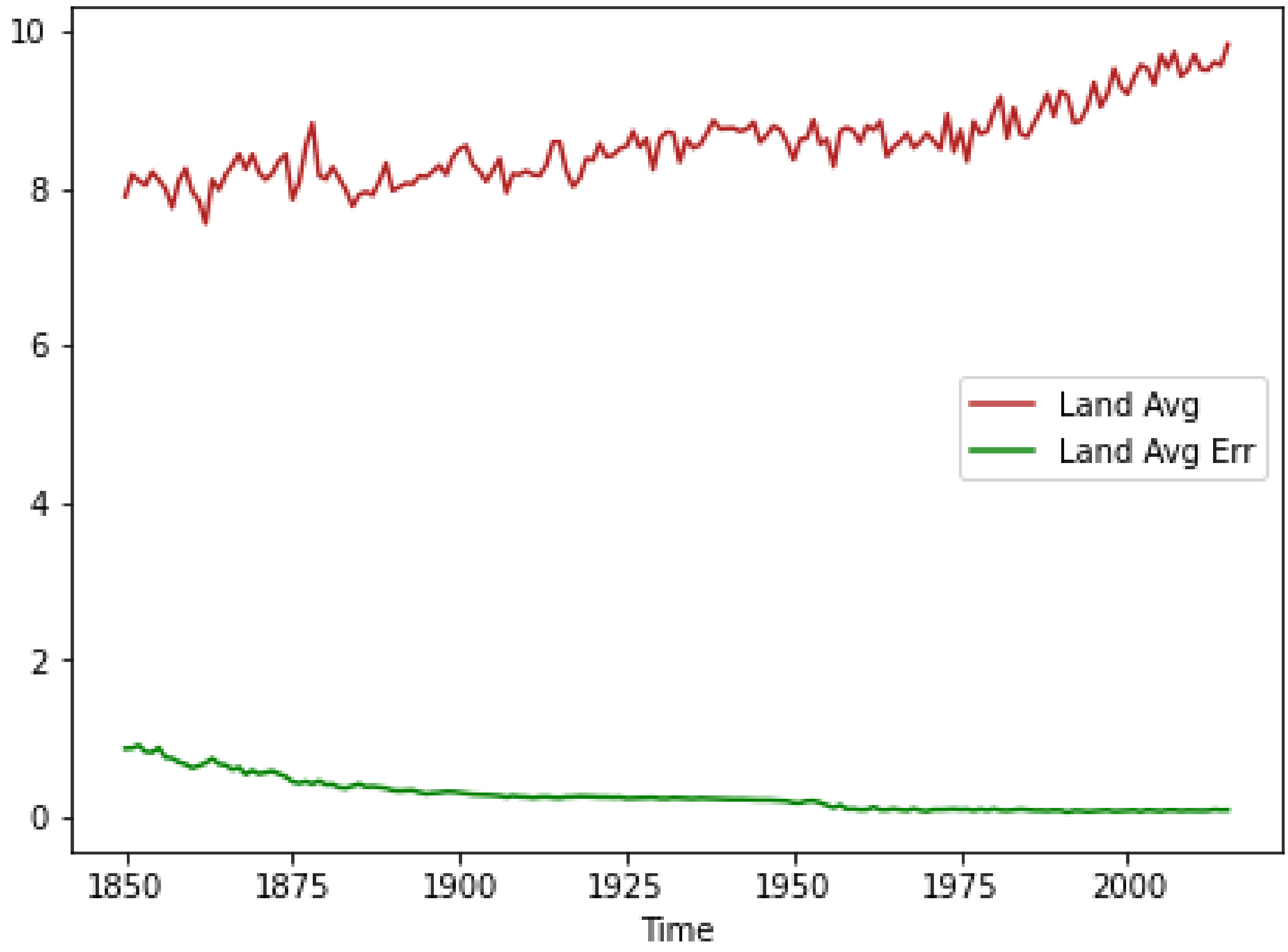
	LandAverageTemperature	LandMinTemperature	LandMaxTemperature	LandAndOceanAverageTemperature
count	166.000000	166.000000	166.000000	166.000000
mean	8.571583	2.743595	14.350601	15.212566
std	0.473687	0.614124	0.447741	0.298629
min	7.558583	1.525083	13.081000	14.740083
25%	8.195708	2.262562	14.055917	14.991208
50%	8.540750	2.734917	14.307708	15.144208
75%	8.791250	3.126833	14.539167	15.379104
max	9.831000	4.148833	15.572667	16.058583

Inferences:

- *LandAverageTemperature:* We can see that, the median and the min, max values does not differ by a considerable amount, due to which the deviation is also a bit less (means, I am saying on the basis of the normal deviation values, maybe this small scale also holds a high importance, in practical scenarios).
- *LandMaxTemperature:* Similar is the case of LandMaxTemperature.

- *LandMinTemperature:* This is an important column, and we can see that, the deviation is the highest among all the other columns, hence we can say that with the years, the min temperature has also increased, min is around 1.5, and has a max value of 4.14.
- *LandAndOceanAverageTemperature:* This column, has the lowest deviation, and hence, we can say that with advent in years, basically there has not been a significant increase in this quantity.

Next let's see the plot of the timeseries

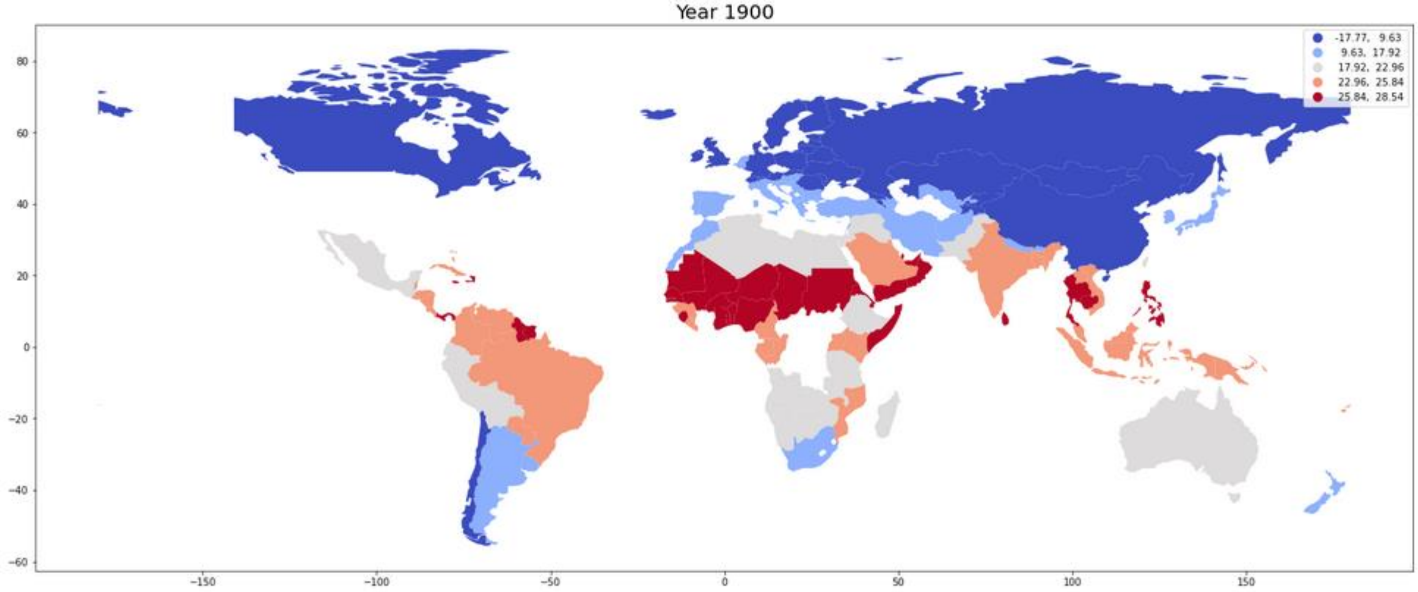


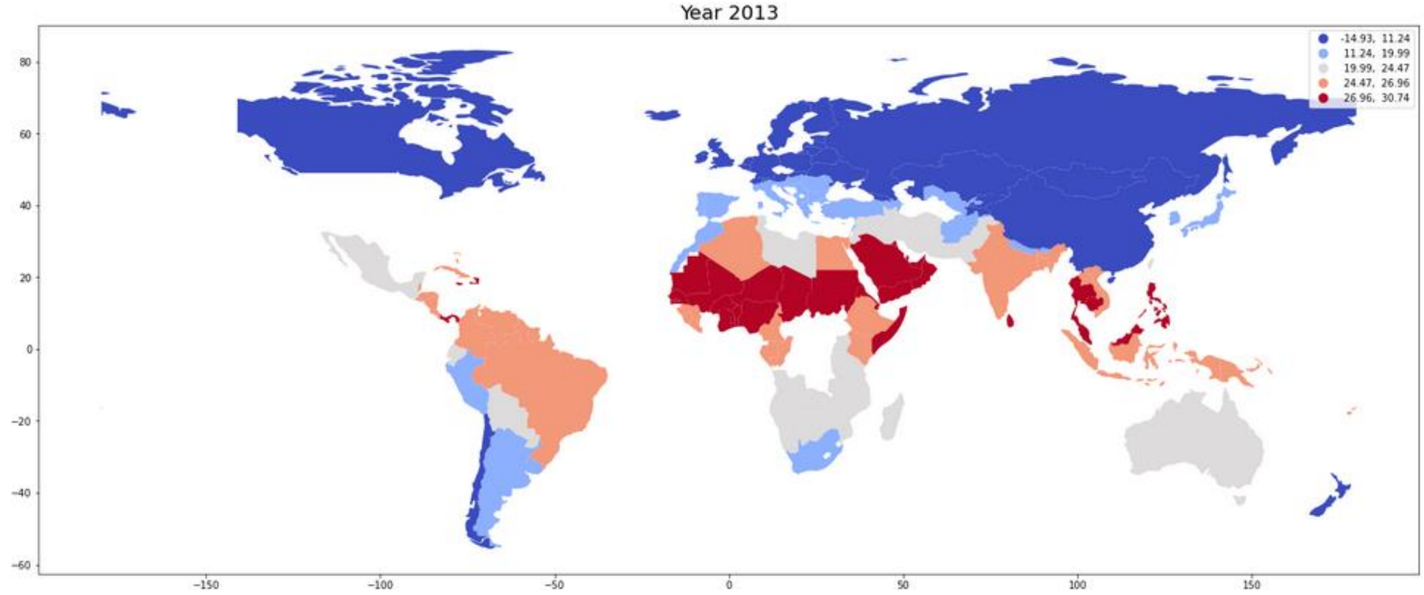
We can see that, with the increase in the years, the average land temperature is increasing, and the spread of the data is becoming saturated. The meaning of saturation is, that each year, the variability in the temperature across all the months is decreasing.

What could be the reason for the increase in the temperature:

- Increase proportion of Carbon Dioxide. What are the reasons for increase in Carbon Dioxide, the industries are primary reasons, then from here, we can link it to industrial revolution, machine automations, etc. And, the Ozone layer depleted, which can be accounted to the CFC

We clearly see the upwards rising trend in the average temperature of the globe, the world map plot will clearly sow us which areas are being affected the most



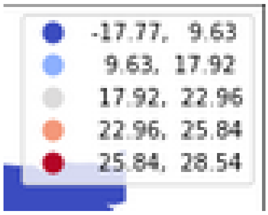


The first image is of year 1900 and the next is after 113 years, look for the following places to see drastic changes, although in those 113 years the range of the average temperature has increased.

- Southeast Africa
- New Zealand
- North of south America
- Middle east
- Southeast Asia

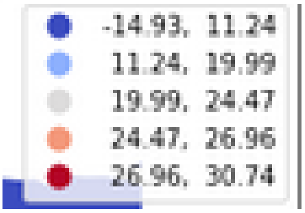
An inspection of the legend from both the legends shows us that the range of average temperature has increased.

1900



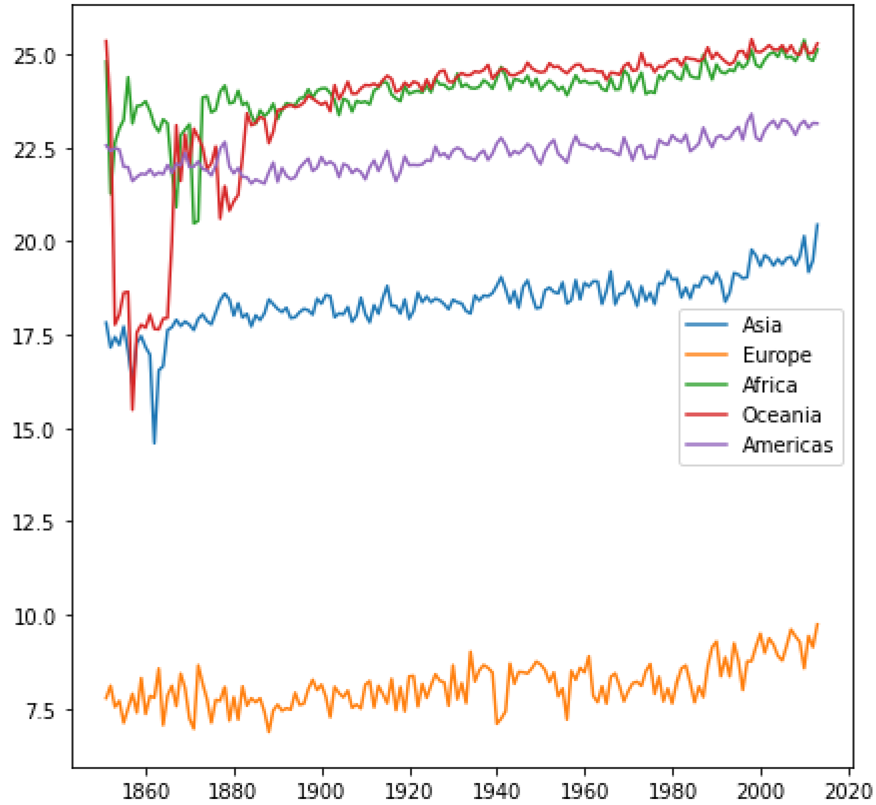
● -17.77, 9.63
● 9.63, 17.92
● 17.92, 22.96
● 22.96, 25.84
● 25.84, 28.54

2013



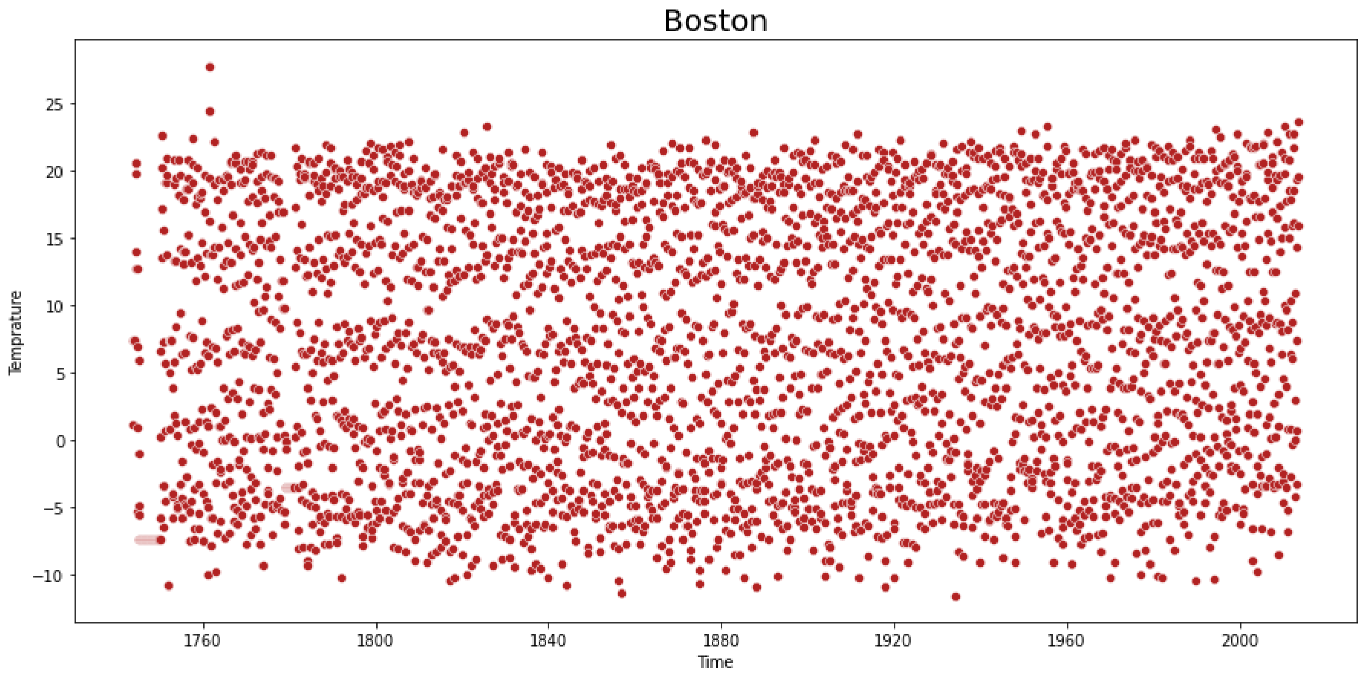
● -14.93, 11.24
● 11.24, 19.99
● 19.99, 24.47
● 24.47, 26.96
● 26.96, 30.74

Average temperature continent vise,

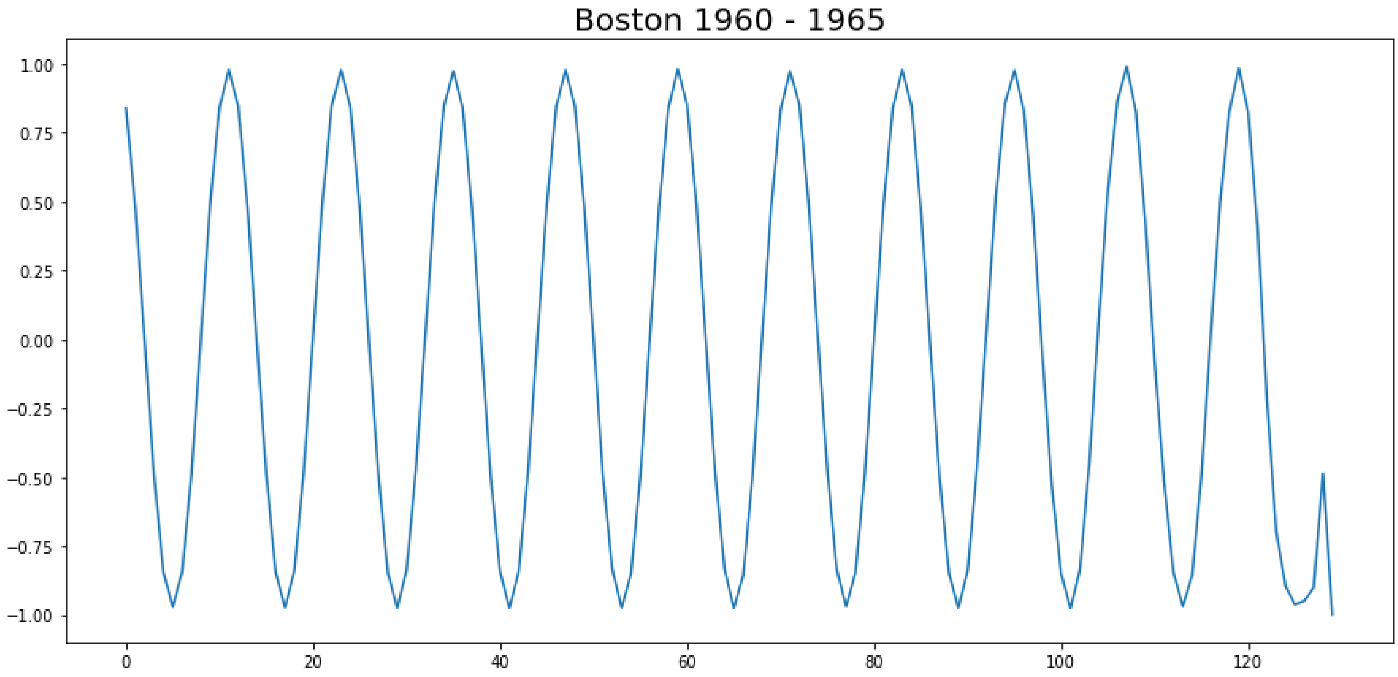


Earth is clearly warming, let take a closer look by zooming into a particular location.

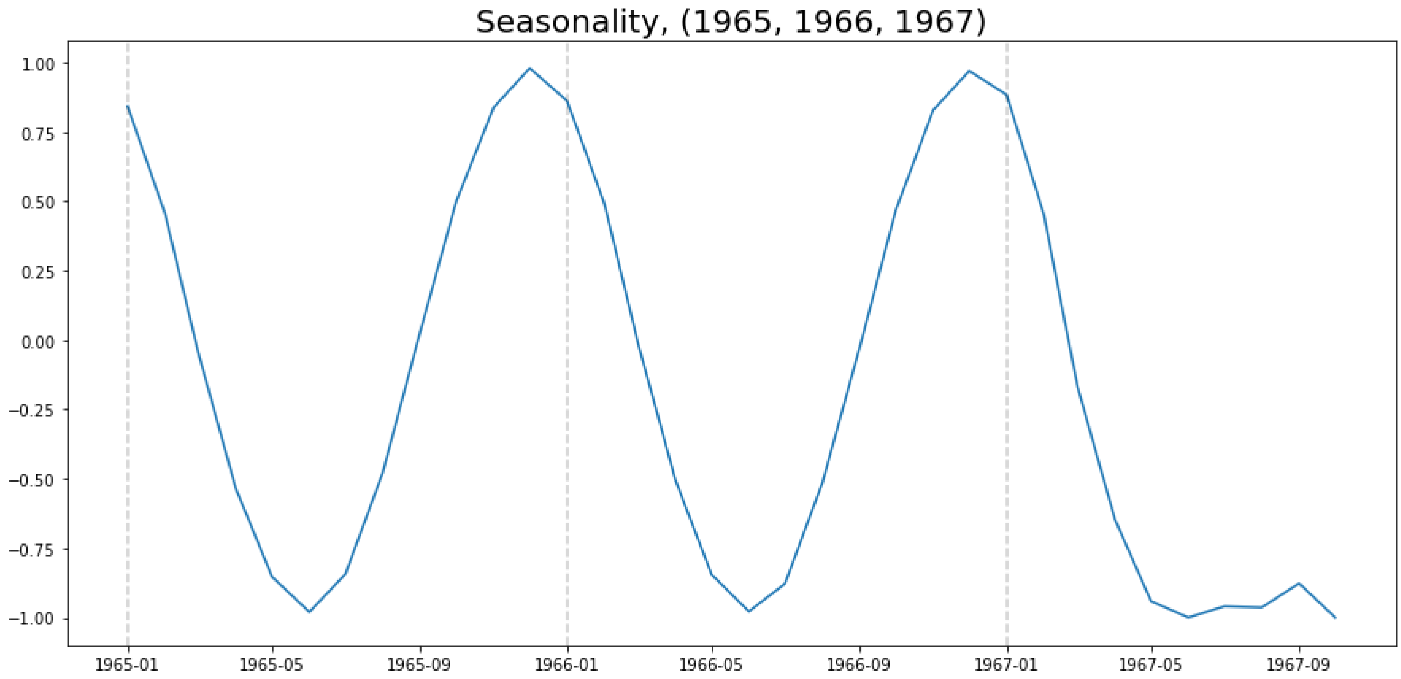
In the next part we will inspect Boston, which is a great example as Boston gets all the four seasons (spring, summer, fall, winter)



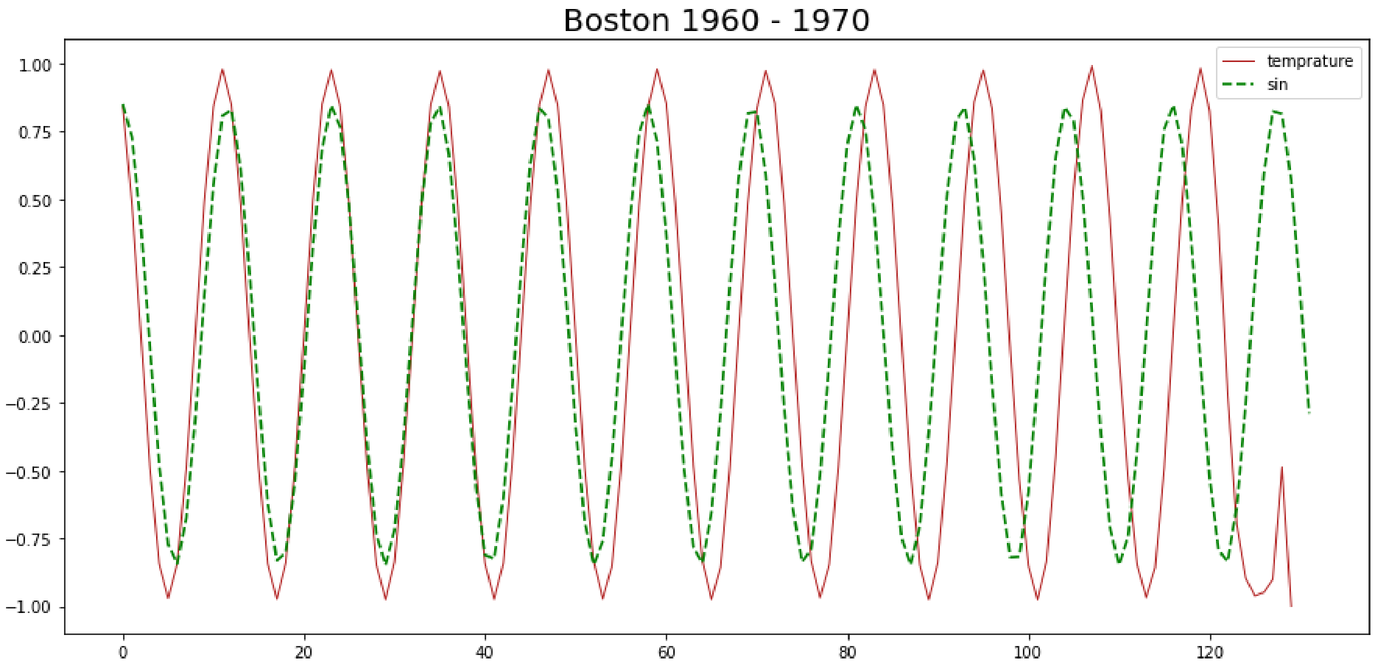
The scatter plot is a little tough to interpret, but by the year 2000 you see that the minimum temperature is moving away from what is was previously, and the temperature is moving upwards.



We can clearly see the pattern in the above image of all the weathers occurring in the span of 5 years, lets try and see if this pattern is consistent and if there is seasonality in the data. The data in the above plot is autocorrelated with lag 1.

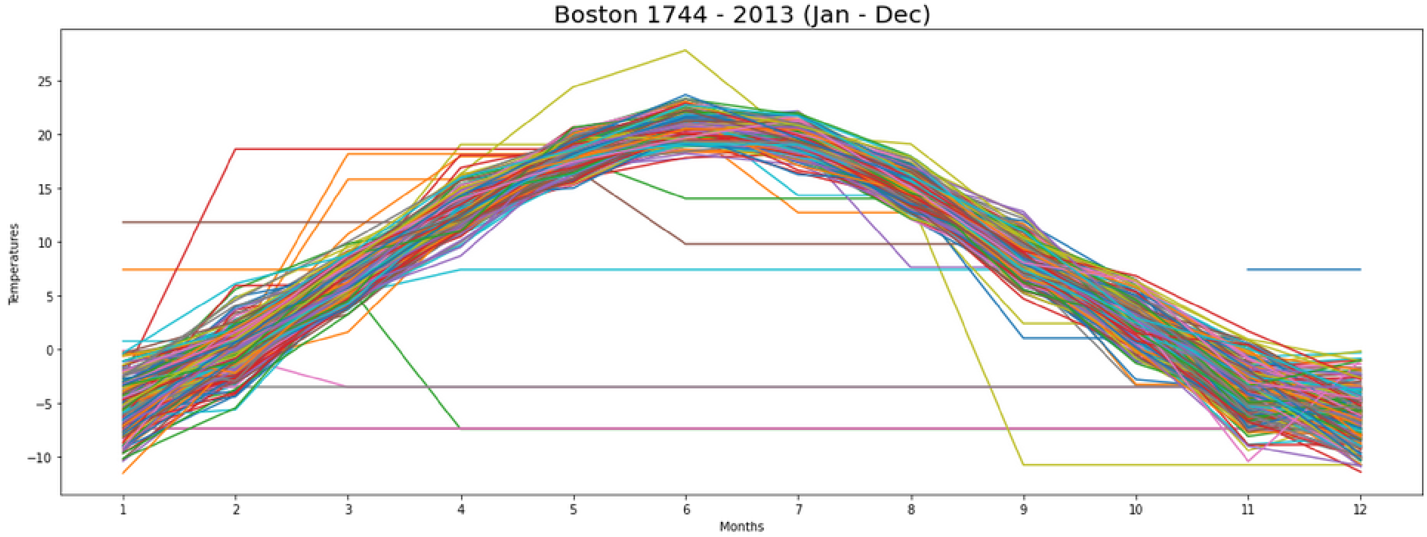


The image above shows that the data is seasonal, we need to zoom out a see if this seasonality holds.

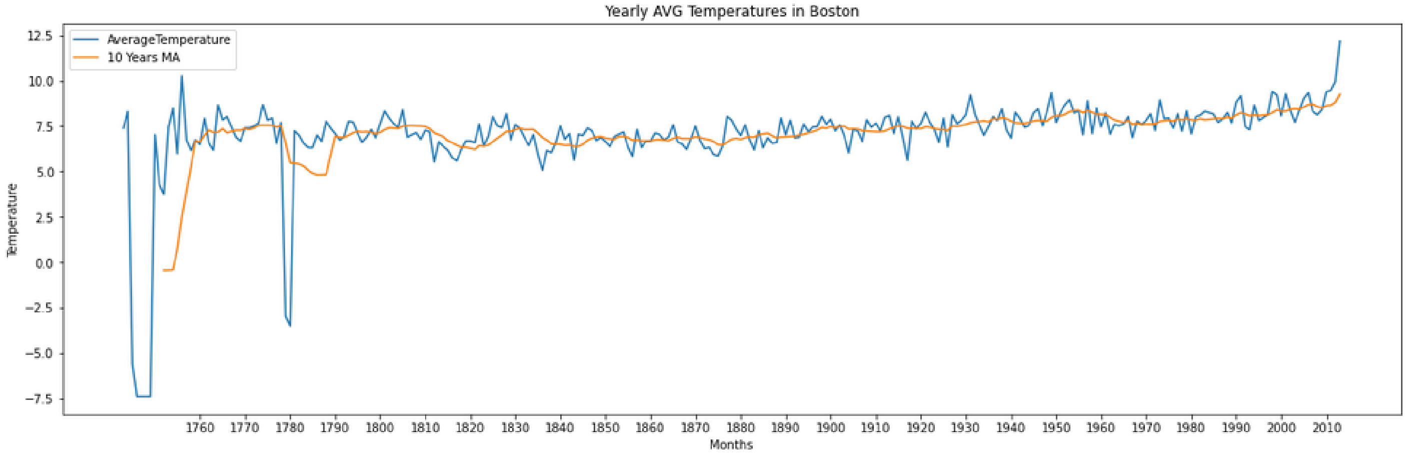


We fit a sin wave above the data to check weather the seasonality holds, and as we see in the image there is a shift in the data, which indicates that the data is seasonal only for short span than the seasonality shifts.

Let’s drill down on the Boston weather a bit more and see how the data is spread out over the months.



The data is more spread out around the ends and closely knit in the middle over the years.

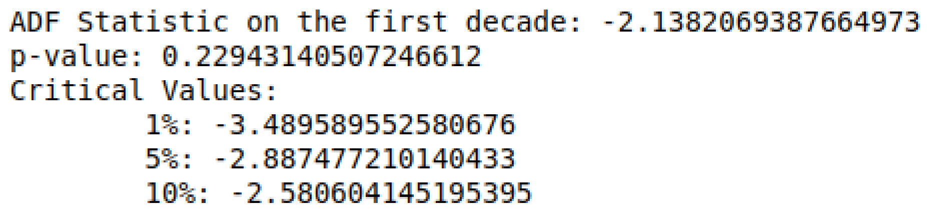


Again, we see clearly that there is an upwards trend.

Modelling and Prediction

By now we know that we will be using ARIMA model as the clearly the timeseries is not stationary.

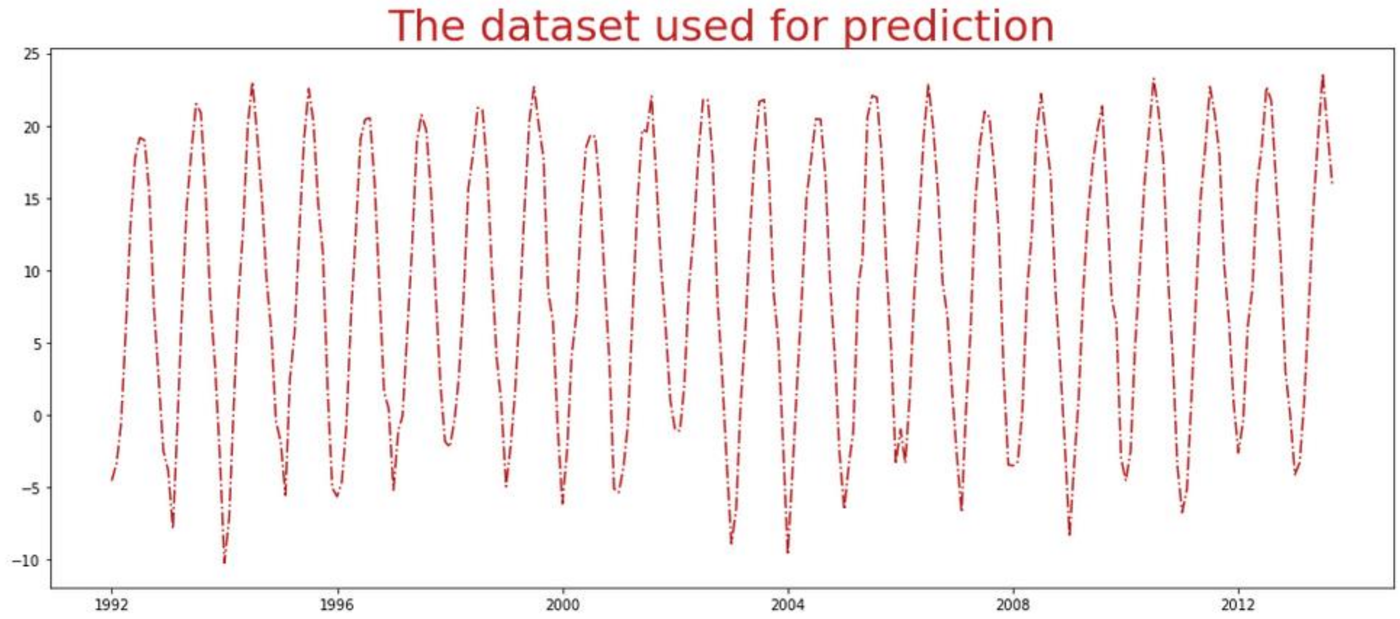
Let’s see the ADF test on the dataset:



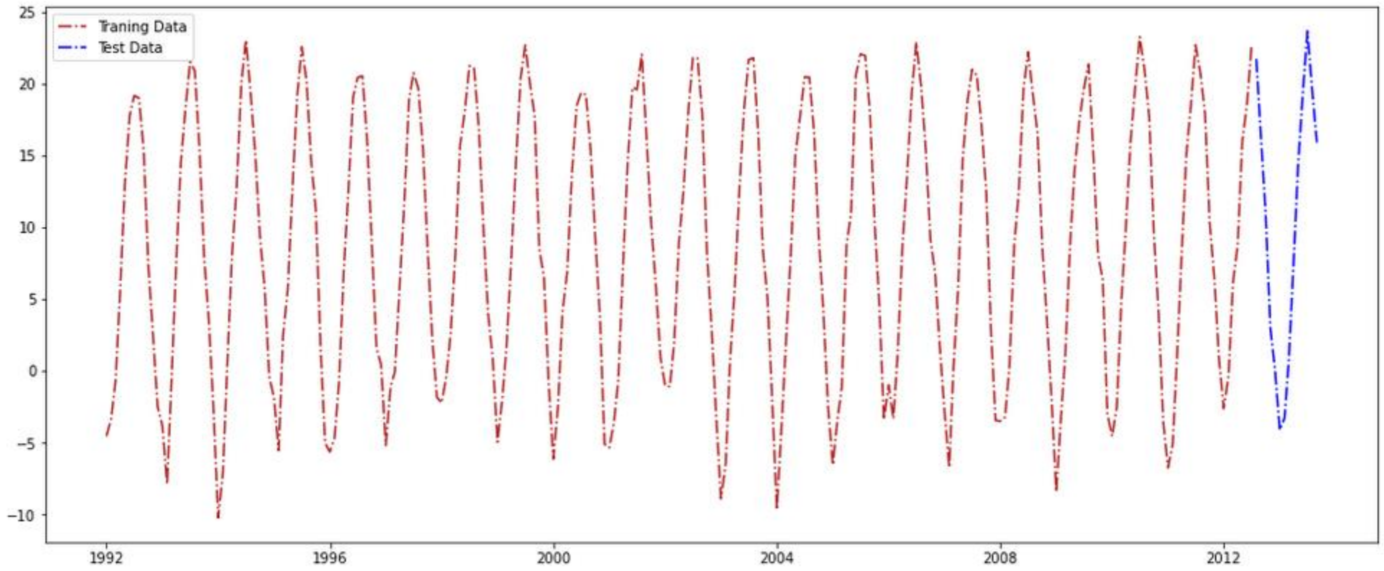
```
ADF Statistic on the first decade: -2.1382069387664973
p-value: 0.22943140507246612
Critical Values:
        1%: -3.489589552580676
        5%: -2.887477210140433
        10%: -2.580604145195395
```

The p-value is not less than 0.05 thus we cannot reject the null hypothesis, thus the dataset is not stationary.

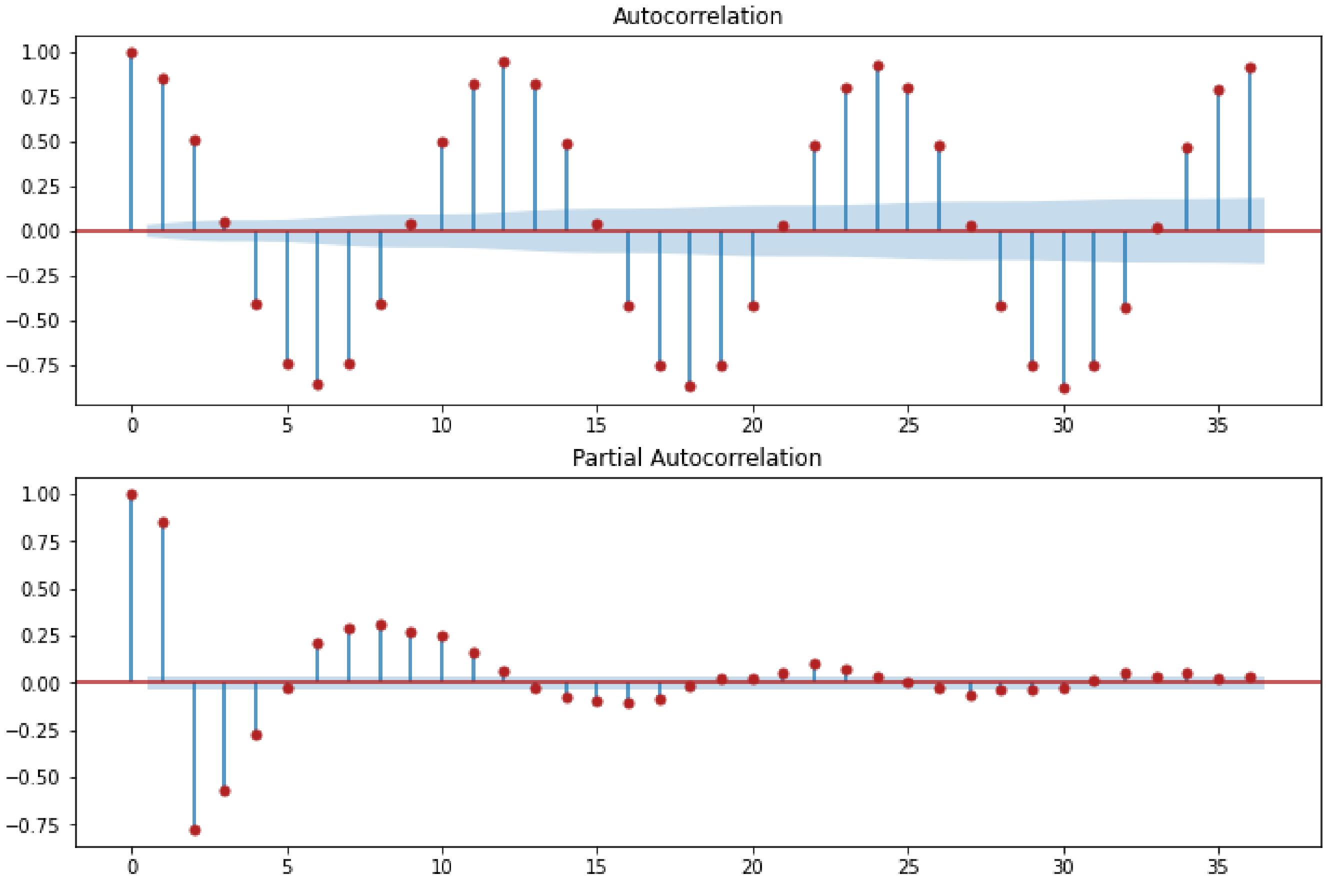
The following is the dataset we would be using for forecasing:



Line in red is training data and blue line is test data.

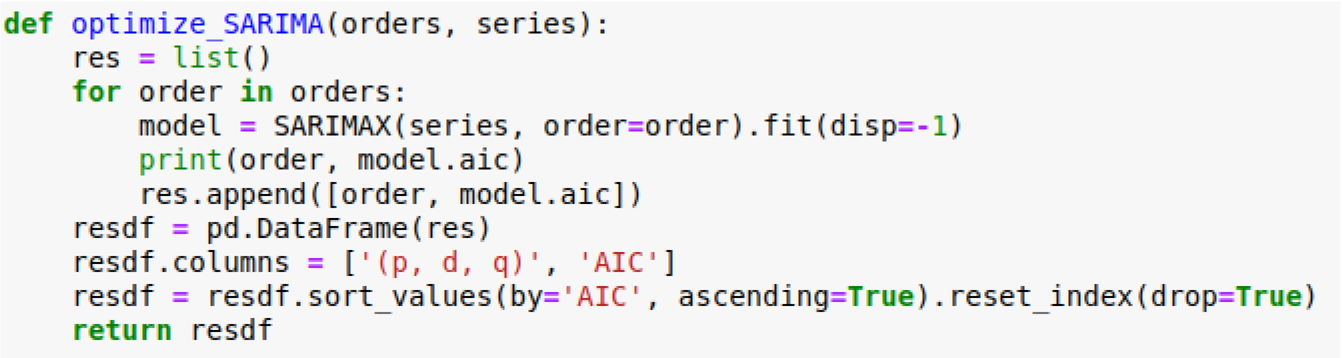


Lets perform the ACF & PACF on the data set to get reasonable values for AR and MA.



In the ACF we see strong correlation till the end so we would have to test MA for multiple values.

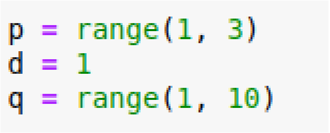
PACF has a strong correlation until 3 so we can test for range from 1 – 3 for AR.



```python
def optimize_SARIMA(orders, series):
    res = list()
    for order in orders:
        model = SARIMAX(series, order=order).fit(disp=-1)
        print(order, model.aic)
        res.append([order, model.aic])
    resdf = pd.DataFrame(res)
    resdf.columns = ['(p, d, q)', 'AIC']
    resdf = resdf.sort_values(by='AIC', ascending=True).reset_index(drop=True)
    return resdf
```

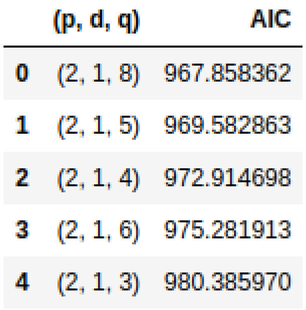
This is the function we would be using for fitting and comparing multiple ARIMA models.

These are the ranges for q, d, q that we would be using according to the ACF & PACF tests.



```python
p = range(1, 3)
d = 1
q = range(1, 10)
```

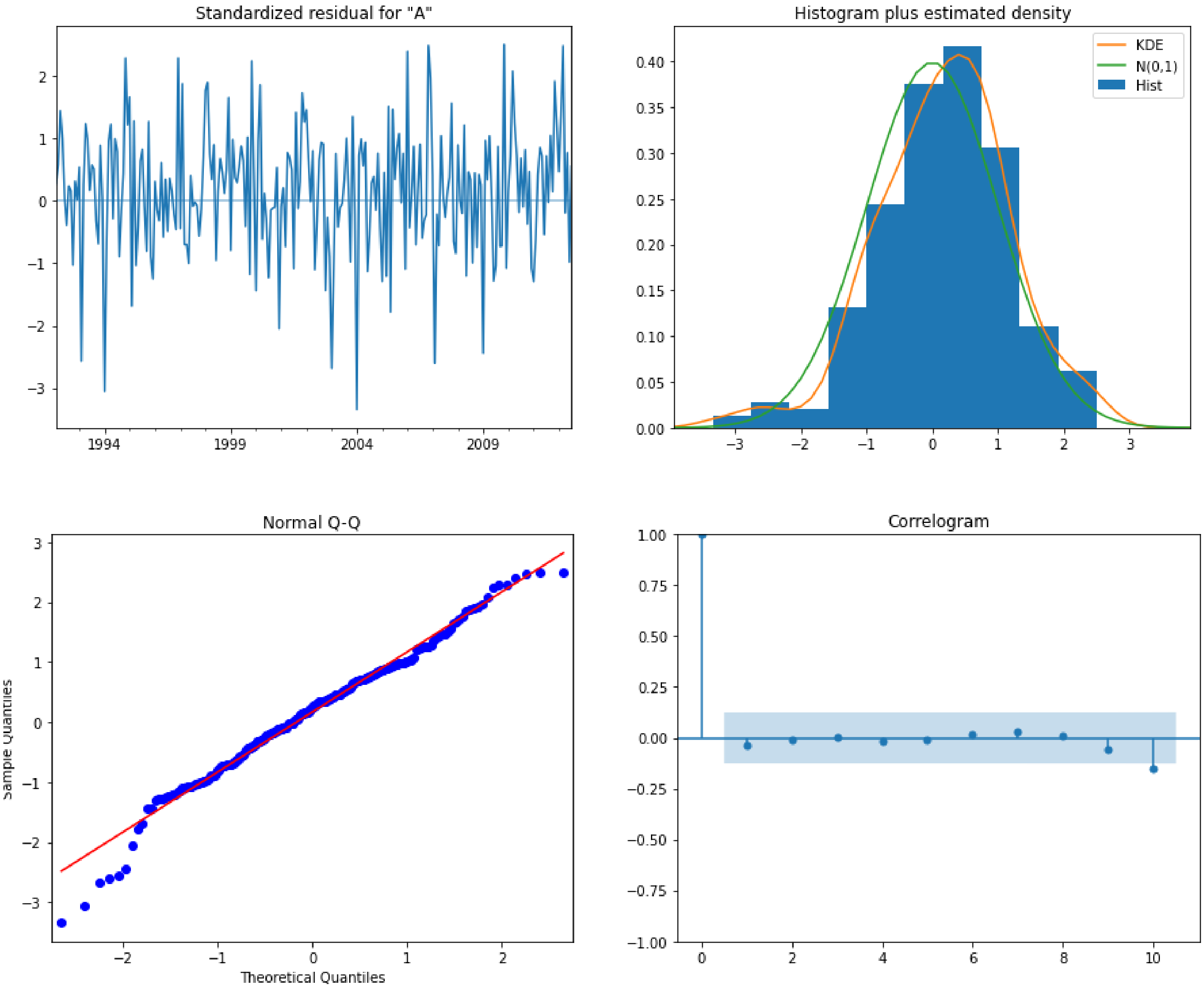
The best 5 fitting models according to the AIC are:



	(p, d, q)	AIC
0	(2, 1, 8)	967.858362
1	(2, 1, 5)	969.582863
2	(2, 1, 4)	972.914698
3	(2, 1, 6)	975.281913
4	(2, 1, 3)	980.385970

The best fitting model is ARIMA(2, 1, 8)

The following are the diagnostics for the model fit



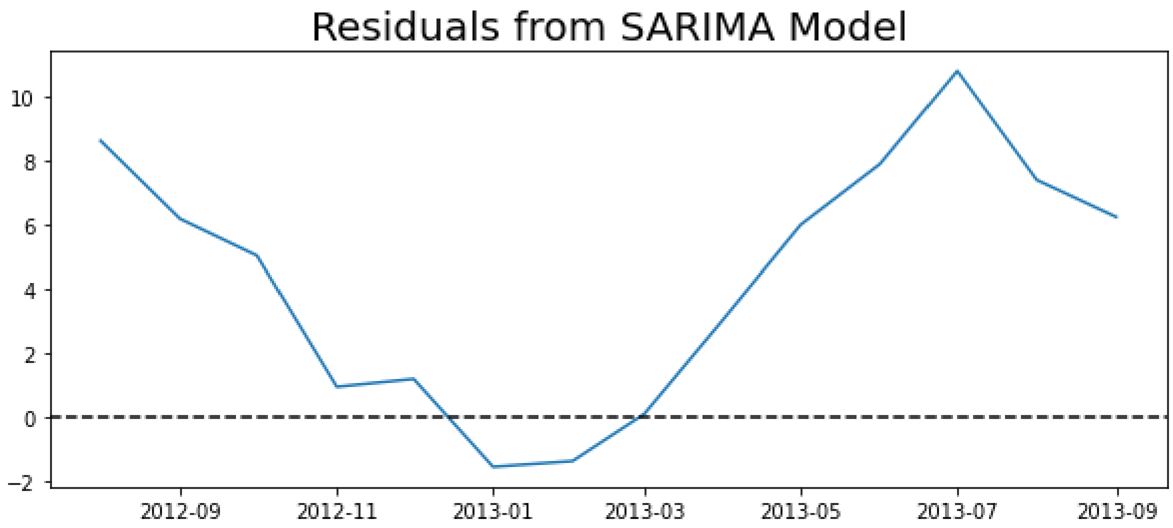
Standardized residual for "A"

Histogram plus estimated density

Normal Q-Q

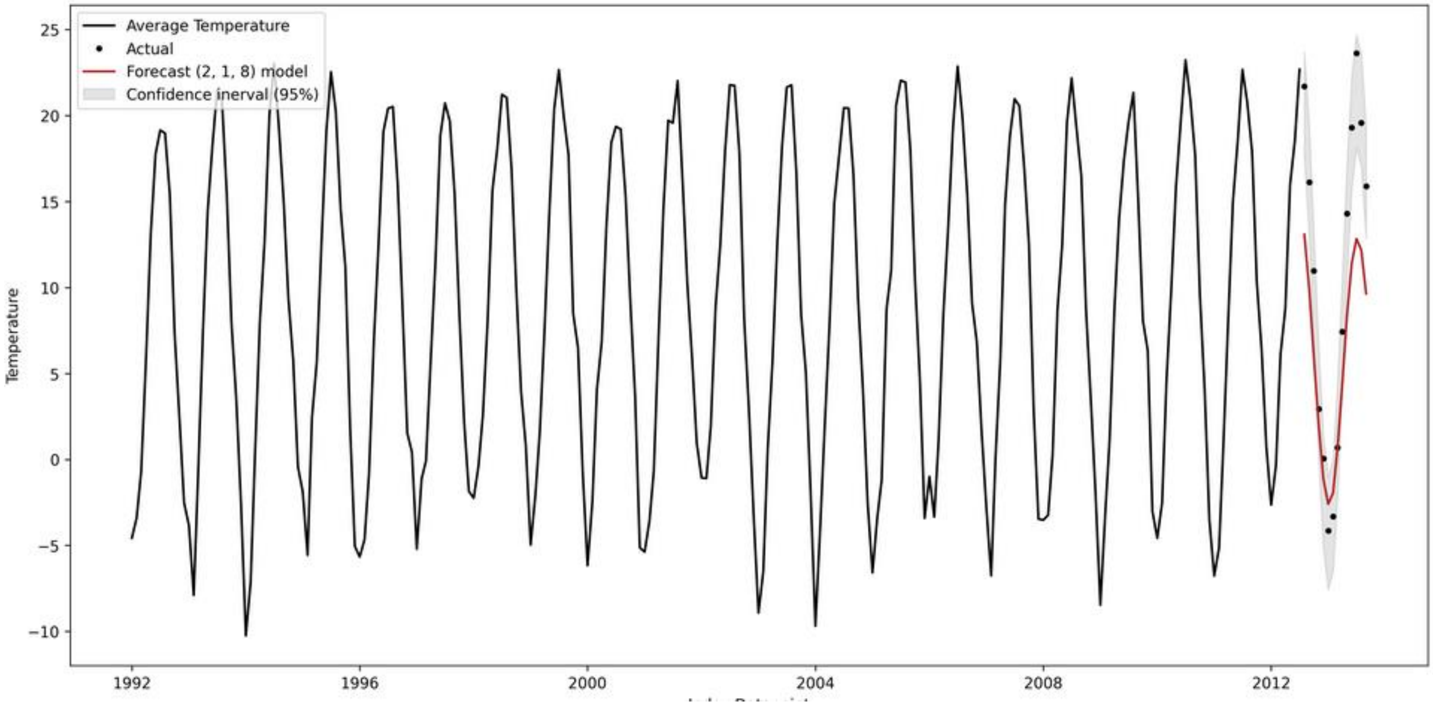
Correlogram

The residuals are as follows

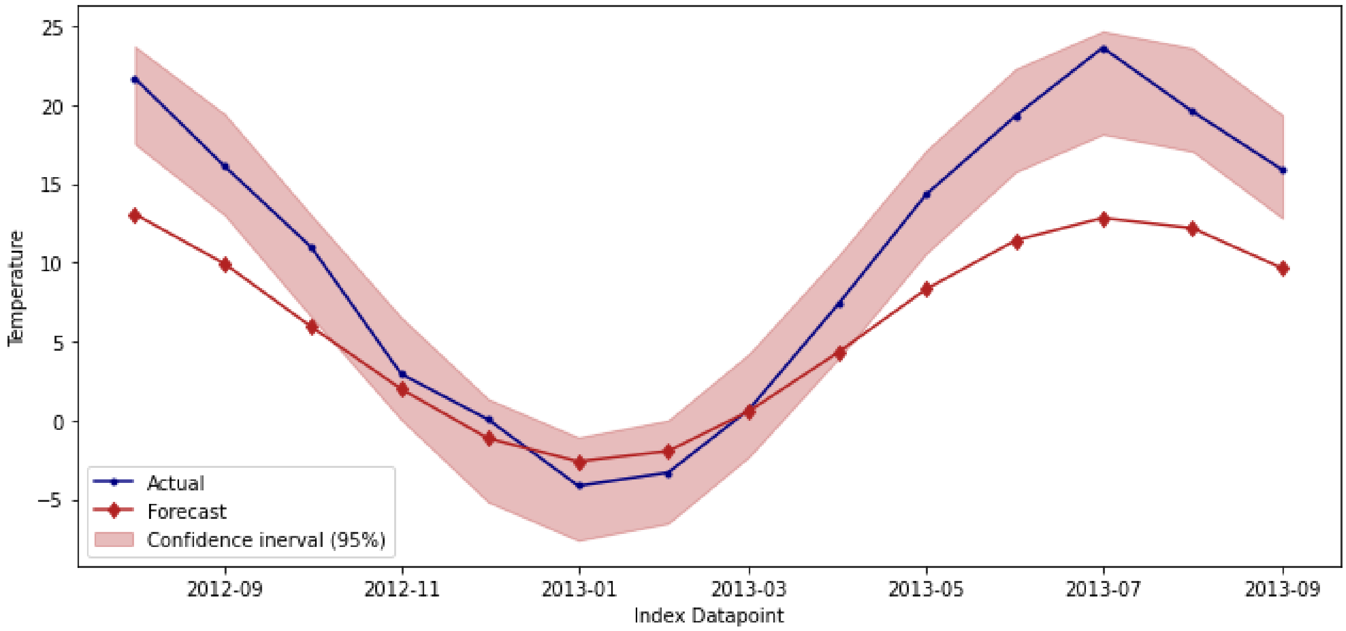


As we can see residuals are above mean, not the ideal fit, but this is the best fitting model we could get for forecasting for a period of 12 months.

This is how the model performs for prediction



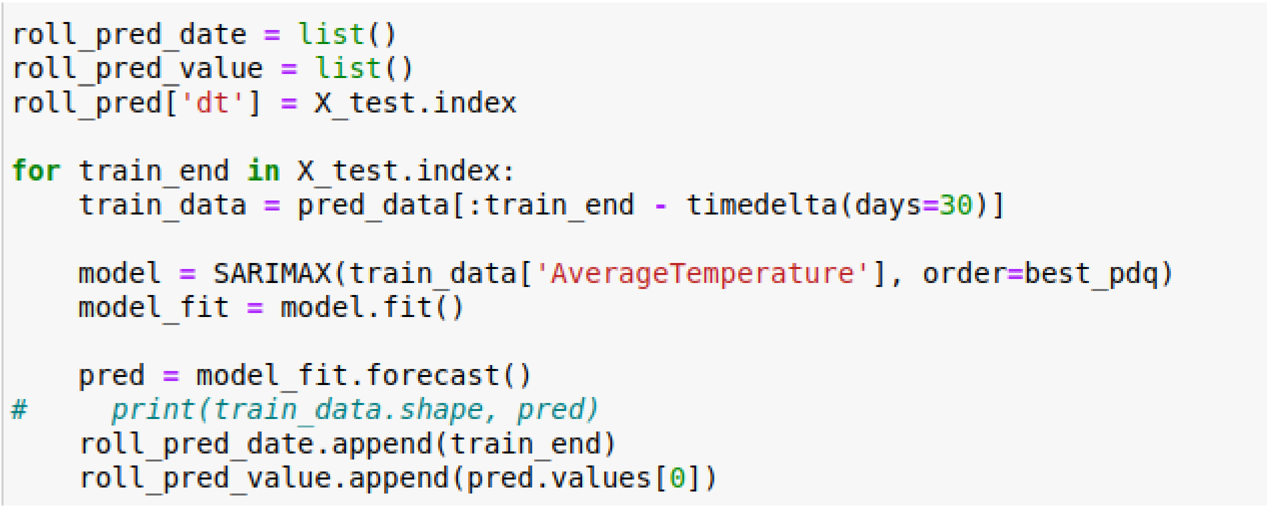
A close look, with 95% confidence interval



I think we could do better if we cut down on the prediction timeline, a reasonable argument is that we rarely need the weather to be predicated 12 months early, due to the nature of the weather i.e. how volatile it is the.

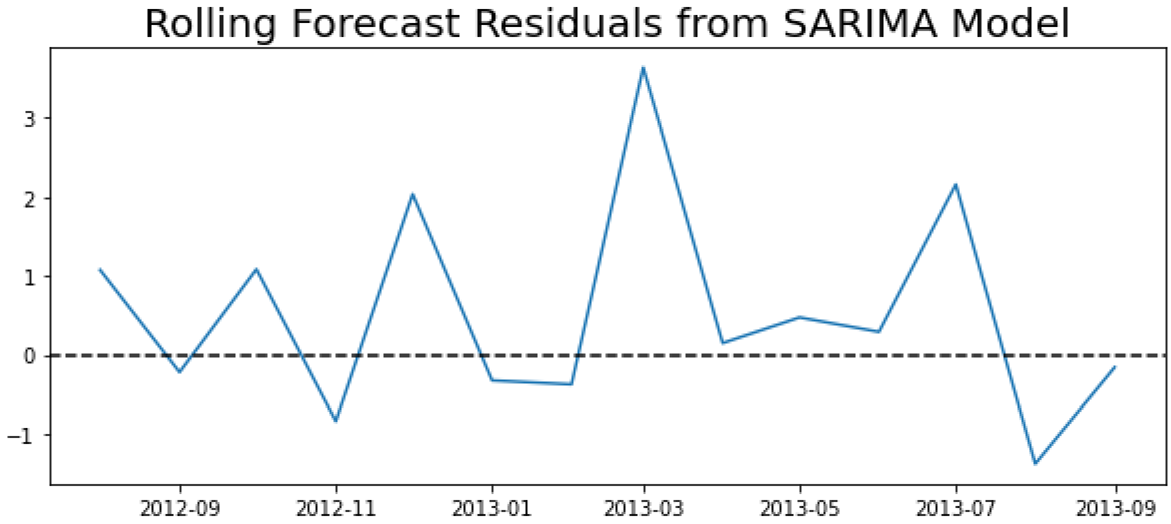
It would be better if we could perform predictions a month at a time, this way we could get better predictions.

We will achieve this by using the Rolling Forecast Origin method.

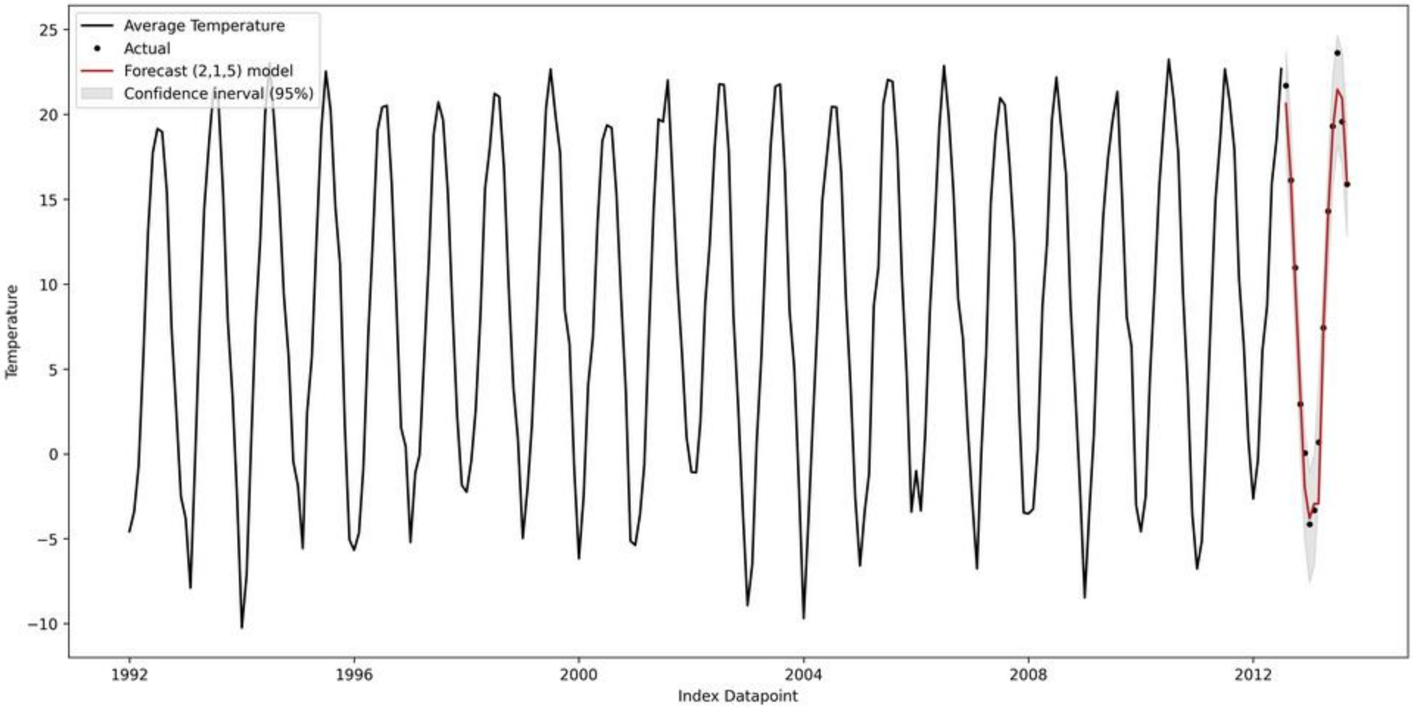


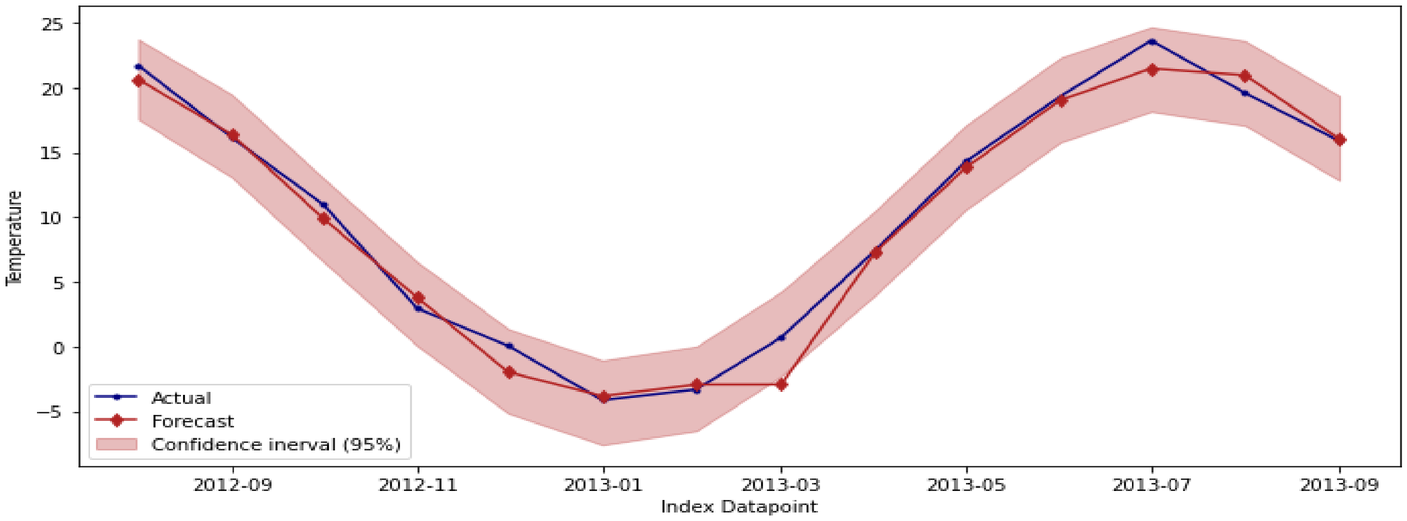
```
roll_pred_date = list()
roll_pred_value = list()
roll_pred['dt'] = X_test.index

for train_end in X_test.index:
    train_data = pred_data[:train_end - timedelta(days=30)]

    model = SARIMAX(train_data['AverageTemperature'], order=best_pdq)
    model_fit = model.fit()

    pred = model_fit.forecast()
#     print(train_data.shape, pred)
    roll_pred_date.append(train_end)
    roll_pred_value.append(pred.values[0])
```

Let’s fit the above model and look a the residuals



Much better results than before.





CONCLUSION:

The average temperature timeseries were studied and best fitted ARIMA model is found after the removal of non-stationarity, and forecasting was done using the same model.

The best fit model (p, d, q) we found was (2, 1, 8), this model was found with the help of ACF, PACF & AIC tests.

We also found out that we can get better predictions if we choose a more realistic timeline (for prediction) given the volatile nature of the weather.

It is concluded from the study that the results achieved from ARIMA modelling for temperature forecasting will assist scientists in gaining the understand that the world is changing due to human intervention and can gauge the severity of this by forecasting the weather.

REFERENCES:

- https://otexts.com/fpp2/accuracy.html
- https://otexts.com/fpp2/prediction-intervals.html
- https://otexts.com/fpp2/AR.html
- https://otexts.com/fpp2/MA.html
- https://en.wikipedia.org/wiki/Autoregressive%E2%80%93moving-average_model
- https://towardsdatascience.com/introduction-to-aic-akaike-information-criterion-9c9ba1c96ced
- https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html
- https://link.springer.com/article/10.1007/s12040-020-01408
- http://www-stat.wharton.upenn.edu/~stine/insr260_2009/lectures/arma_forc.pdf
- https://machinelearningmastery.com/time-series-forecast-study-python-monthly-sales-french-champagne/
- https://www.ibm.com/blogs/internet-of-things/what-is-seasonal-and-subseasonal-forecasting/