TIME SERIES ANALYSIS PROJECT REPORT

TOPIC: Predicting Electricity Consumption using Time Series Analysis

Department of Statistics

Savitribai Phule Pune University

Pune.

Feb-Apr 2022

Name: Mohit Dattatray Jadhav (2118)

CONTENT

1. Introduction……………………………………………

2. Problem Statement…………………………………….

3. Visualizing Time Series………………………………..

4. Make the Time Series Stationary……………………..

5. Modelling……………………………………………….

6. Forecasting……………………………………………..

7. Conclusion……………………………………………..

8. References……………………………………………...

Introduction:

What is Time Series analysis?

Time series forecasting is a technique for the prediction of events through a sequence of time. The technique is used across many fields of study, from geology to behaviour to economics. The techniques predict future events by analysing the trends of the past, on the assumption that future trends will hold similar to historical trends.

There are two main goals of time series analysis:

(a) Identifying the nature of the phenomenon represented by the sequence of observations, and

(b) Forecasting (predicting future values of the time series variable).

Both of these goals require that the pattern of observed time series data is identified and more or less formally described. Once the pattern is established, we can interpret and integrate it with other data (i.e., use it in our theory of the investigated phenomenon, e.g., seasonal commodity prices). Regardless of the depth of our understanding and the validity of our interpretation (theory) of the phenomenon, we can extrapolate the identified pattern to predict future events.

Stages in Time Series Forecasting

Solving a time series problem is a little different as compared to a regular modelling task. A simple/basic journey of solving a time series problem can be demonstrated through the following processes. We will understand about tasks which one needs to perform in every stage.

We will also look at the python implementation of each stage of our problem-solving journey.

Steps are –

1. Visualizing Time Series

In this step, we try to visualize the series. We try to identify all the underlying patterns related to the series like trend and seasonality.

2. Make the Time Series Stationary

A stationary time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time. Most statistical forecasting methods are based on the assumption that the time series can be rendered approximately stationary (i.e., “stationarised”) through the use of mathematical transformations. A stationarised series is relatively easy to predict: you simply predict that its statistical properties will be the same in the future as they have been in the past! Another reason for trying to stationarise a time series is to be able to obtain meaningful sample statistics such as means, variances, and correlations with other variables. Such statistics are useful as descriptors of future behavior only if the series is stationary. For example, if the series is consistently increasing over time, the sample mean and variance will grow with the size of the sample, and they will always underestimate the mean and variance in future periods. And if the mean and variance of a series are not well-defined, then neither are its correlations with other variables

3. Modelling

We need to find optimal parameters for forecasting models one’s we have a stationary series. These parameters come from the ACF and PACF plots. Hence, this stage is more about plotting above two graphs and extracting optimal model parameters based on them.

Once we have our optimal model parameters, we can fit an ARIMA model to learn the pattern of the series. Always remember that time series algorithms work on stationary data only hence making a series stationary is an important aspect.

4. Forecasting

After fitting our model, we will be predicting the future in this stage.

Problem Statement

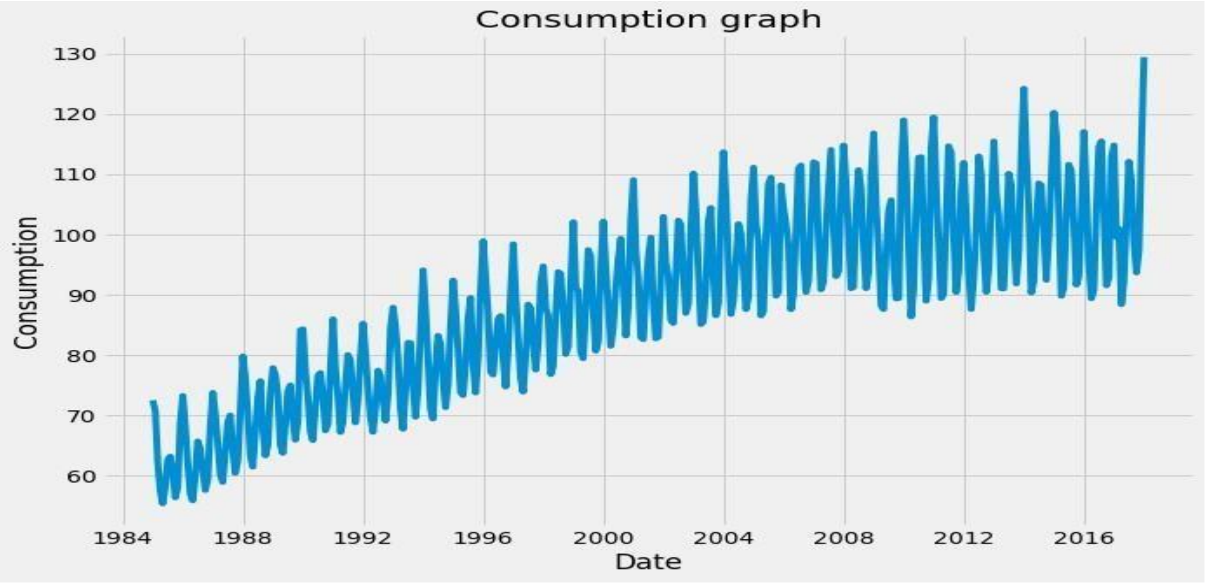
Our Dataset contains only 2 columns, one column is Date and the other column relates to the consumption percentage.

It shows the consumption of electricity from 1985 till 2018. The goal is to predict electricity consumption for the next 6 years i.e. till 2024.

1.Visualizing Time Series -

Our Dataset contains only 2 columns, one column is Date and the other column relates to the consumption percentage.It shows the consumption of electricity from 1985 till 2018.

	Date	Value
0	01-01-1985	72.5052
1	02-01-1985	70.6720
2	03-01-1985	62.4502
3	04-01-1985	57.4714
4	05-01-1985	55.3151



- Remember that for time series forecasting, a series needs to be **stationary**. The series should have a constant mean, variance, and covariance.
- There are few points to note here, the **mean is not constant** in this case as we can clearly see an **upward trend**.
- Hence, we have identified that our series is **not stationary**. We need to have a stationary series to do time series forecasting. In the next stage, we will try to convert this into a stationary series.

Also, a given time series is thought to consist of three systematic components including level, trend, seasonality, and one non-systematic component called noise.

These components are defined as follows:

Level: The average value in the series.

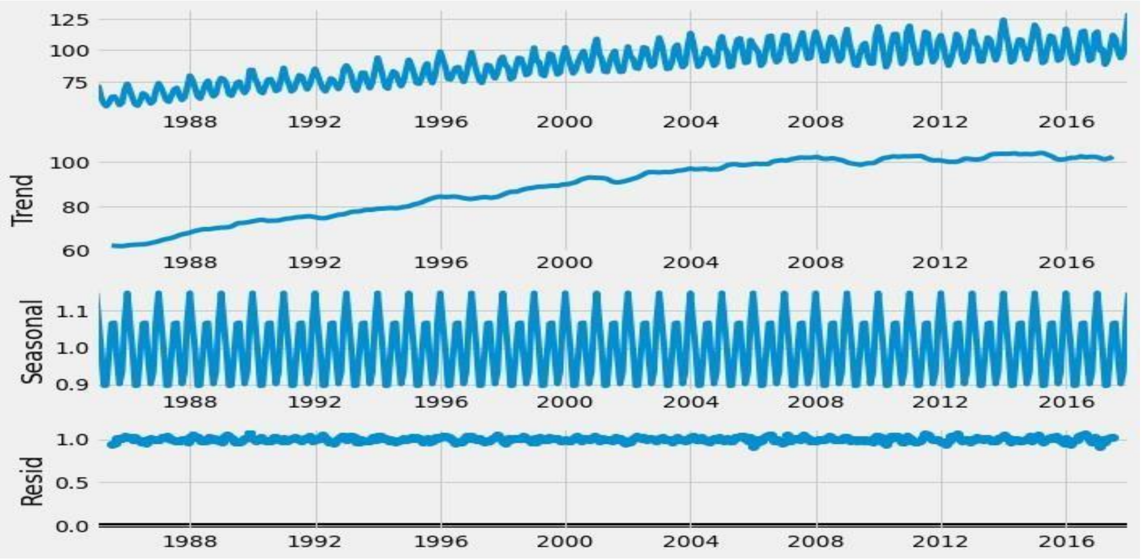
Trend: The increasing or decreasing value in the series.

Seasonality: The repeating short-term cycle in the series.

Noise: The random variation in the series.

In order to perform a time series analysis, we may need to separate seasonality and trend from our series. The resultant series will become stationary through this process.

So let us separate Trend and Seasonality from the time series.



This gives us more insight into our data and real-world actions. Clearly, there is an upward trend and a recurring event where electricity consumption shoots maximum every year.

2. Make the Time Series Stationary -

First, we need to check if a series is stationary or not.

ADF (Augmented Dickey-Fuller) Test

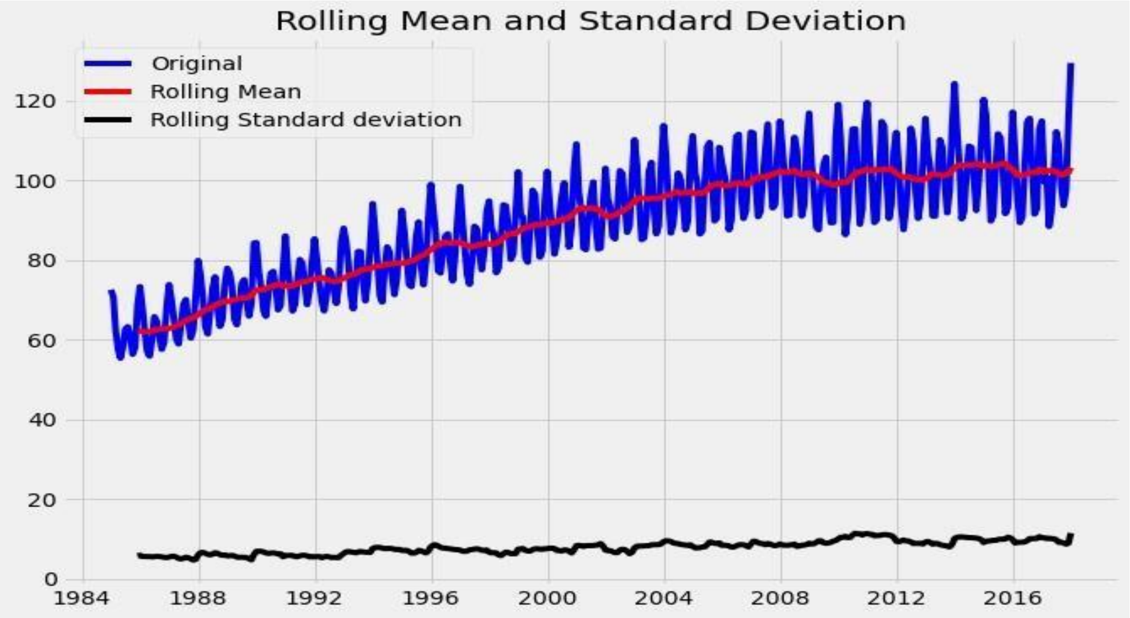
The Dickey-Fuller test is one of the most popular statistical tests. It can be used to determine the presence of unit root in the series, and hence help us understand if the series is stationary or not. The null and alternate hypothesis of this test is:

Null Hypothesis: The series has a unit root (value of a =1)

Alternate Hypothesis: The series has no unit root.

If we fail to reject the null hypothesis, we can say that the series is non-stationary. This means that the series can be linear or difference stationary.

If both mean and standard deviation are flat lines(constant mean and constant variance), the series becomes stationary.



```
    Results of dickey fuller test
Test Statistics                  -2.256990
p-value                           0.186215
No. of lags used                 15.000000
Number of observations used     381.000000
critical value (1%)              -3.447631
critical value (5%)              -2.869156
critical value (10%)             -2.570827
```

Through the above graph, we can see the increasing mean and standard deviation and hence our series is **not stationary**.

We see that the **p-value** is greater than **0.05** so we cannot reject the Null hypothesis. Also, the test statistics is greater than the critical values. so the data is **non-stationary**.

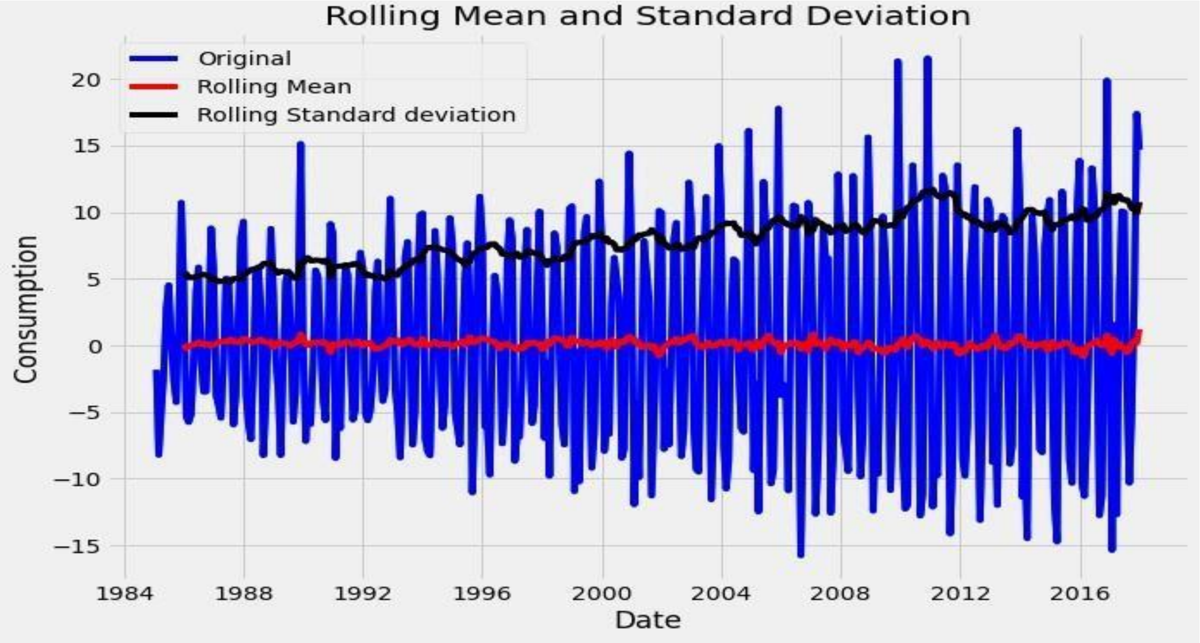
To get a stationary series, we need to eliminate the **trend** and **seasonality** from the series.

Differencing -

One of the most common methods of dealing with both trend and seasonality is differencing. Differencing is a method of transforming a time series dataset. It can be used to remove the series dependence on time, so-called temporal dependence. This includes structures like trends and seasonality. Differencing can help stabilize the mean of the time series by removing changes in the level of a time series, and so eliminating (or reducing) trend and seasonality.

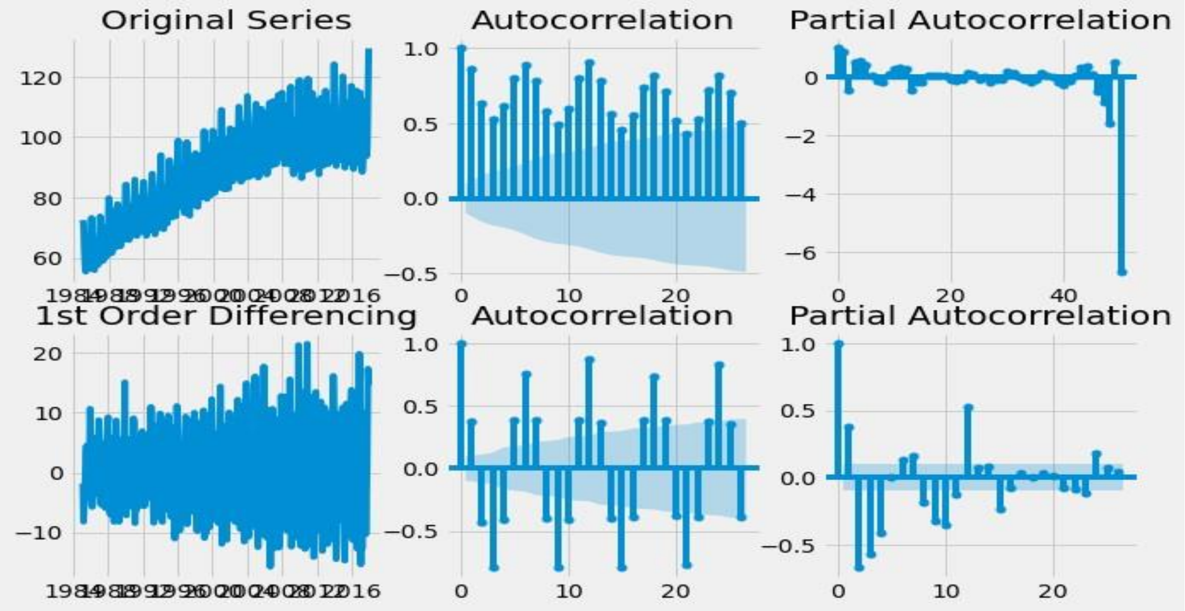
Differencing is performed by subtracting the previous observation from the current observation.

Perform the Dickey-Fuller test (ADFT) once again.



```
         Results of dickey fuller test
Test Statistics                  -7.104891e+00
p-value                           4.077787e-10
No. of lags used                  1.400000e+01
Number of observations used       3.810000e+02
critical value (1%)              -3.447631e+00
critical value (5%)              -2.869156e+00
critical value (10%)             -2.570827e+00
```

From the above graph, we observed that the data attained **stationarity**. As We see that the **p-value** is less than **0.05** so we can reject the Null hypothesis. Also, the test statistics is less than the critical values. so the data is **stationary**.



We have got our **stationary** series and now we can move to fit the best Model for our Data.

3. Modelling

Before we go on to build our forecasting model, we need to determine optimal parameters for our model. Here we use **ARIMA** model.

ARIMA stands for **Auto-Regressive Integrated Moving Averages**. The ARIMA forecasting for a stationary time series is nothing but a linear (like a linear regression) equation. The predictors depend on the parameters **(p,d,q)** of the **ARIMA** model.

An ARIMA model is one where the time series was differenced at least once to make it stationary and you combine the AR and the MA terms. So the equation becomes:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + .. + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + .. + \phi_q \epsilon_{t-q}$$

Number of AR (Auto-Regressive) terms (p): AR terms are just lags of dependent variable. For instance if p is 5, the predictors for y(t) will be y(t- 1)….y(t-5).

Number of MA (Moving Average) terms (q): MA terms are lagged forecast errors in prediction equation. For instance if q is 5, the predictors for x(t) will be e(t-1) ….e(t-5) where e(i) is the difference between the moving average at i^{th} instant and actual value.

Number of Differences (d): These are the number of nonseasonal differences, i.e., in this case we took the first order difference. So here we get d=1.

An importance concern here is how to determine the value of 'p' and 'q'. We use two plots to determine these numbers.

Autocorrelation Function (ACF)

Statistical correlation summarizes the strength of the relationship between two variables. Pearson's correlation coefficient is a number between -1 and 1 that describes a negative or positive correlation respectively. A value of zero indicates no correlation.

We can calculate the correlation for time series observations with previous time steps, called lags. Because the correlation of the time series observations is calculated with values of the same series at previous times, this is called a serial correlation, or an autocorrelation.

A plot of the autocorrelation of a time series by lag is called the AutoCorrelation Function, or the acronym ACF. This plot is sometimes called a correlogram or an autocorrelation plot.

Partial Autocorrelation Function (PACF)

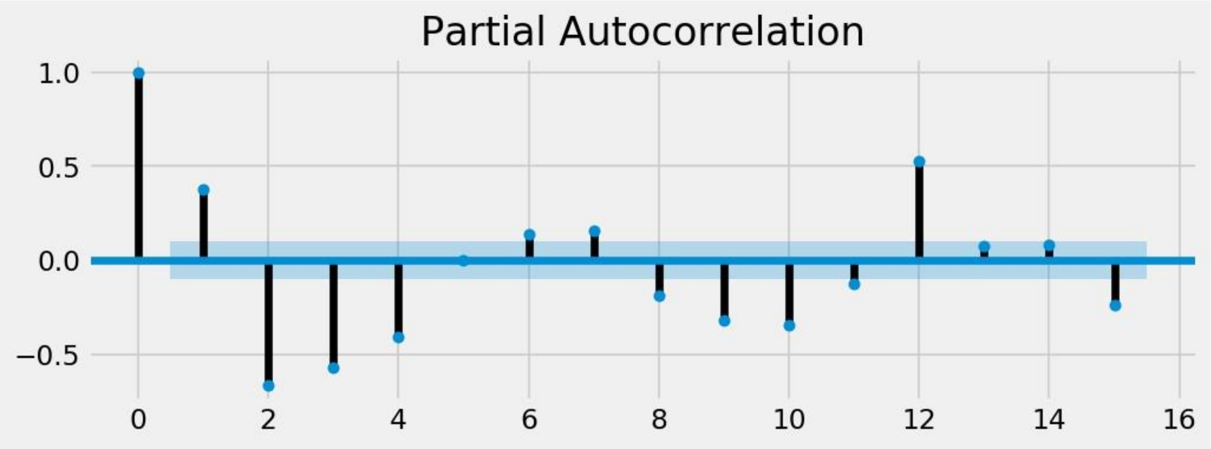
A partial autocorrelation is a summary of the relationship between an observation in a time series with observations at prior time steps with the relationships of intervening observations removed.

The partial autocorrelation at lag k is the correlation that results after removing the effect of any correlations due to the terms at shorter lags.

The autocorrelation for observation and observation at a prior time step is comprised of both the direct correlation and indirect correlations. It is these indirect correlations that the partial autocorrelation function seeks to remove.

we can find out the required number of **AR terms(p)** by inspecting the Partial Autocorrelation (PACF) plot.

Any autocorrelation in a stationarized series can be rectified by adding enough AR terms. So, we initially take the order of AR term to be equal to as many lags that crosses the significance limit in the PACF plot.

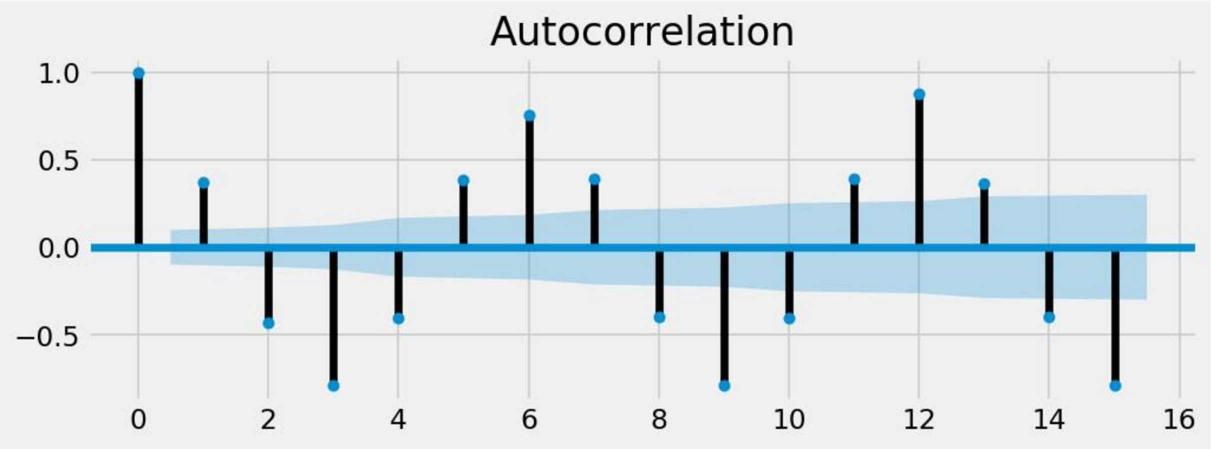


we can observe that the PACF lag 1 is quite significant since is well above the significance line. Lag 2 turns out to be significant as well, slightly managing to cross the significance limit (blue region). But I am going to be conservative and tentatively fix the p as 1.

Just like how we looked at the PACF plot for the number of AR terms, we can look at the ACF plot for the number of MA terms. An MA term is technically, the error of the lagged forecast.

The ACF tells how many MA terms are required to remove any autocorrelation in the stationarized series.

Let’s see the autocorrelation plot of the differenced series.



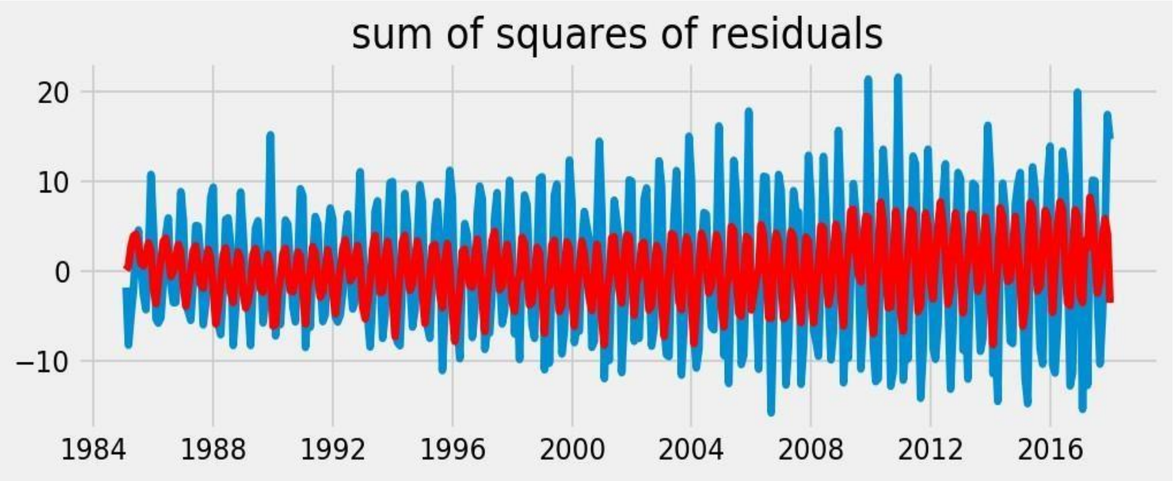
one of lag is well above the significance line. So, let’s tentatively fix q as 1. When in doubt, go with the simpler model that sufficiently explains the Y. Now that we’ve determined the values of p, d and q, we have everything needed to fit the ARIMA model. Let’s use the `ARIMA()` implementation in `statsmodels` package.

1.ARIMA(1,1,1) :

Dep. Variable:	D.Consumption	**No. Observations:**	396
Model:	ARIMA(1, 1, 1)	**Log Likelihood**	-1327.994
Method:	css-mle	**S.D. of innovations**	6.904
Date:	Mon, 16 Aug 2021	**AIC**	2663.989
Time:	00:03:34	**BIC**	2679.914
Sample:	02-01-1985 - 01-01-2018	**HQIC**	2670.298

	coef	std err	z	P |z|	[0.025	0.975]
const	0.1108	0.020	5.607	0.000	0.072	0.149
ar.L1.D.Consumption	0.5414	0.045	11.919	0.000	0.452	0.630

ma.L1.D.Consumption	-0.9767	0.010	-102.359	0.000	-0.995	-0.958



RSS : 18925.738995

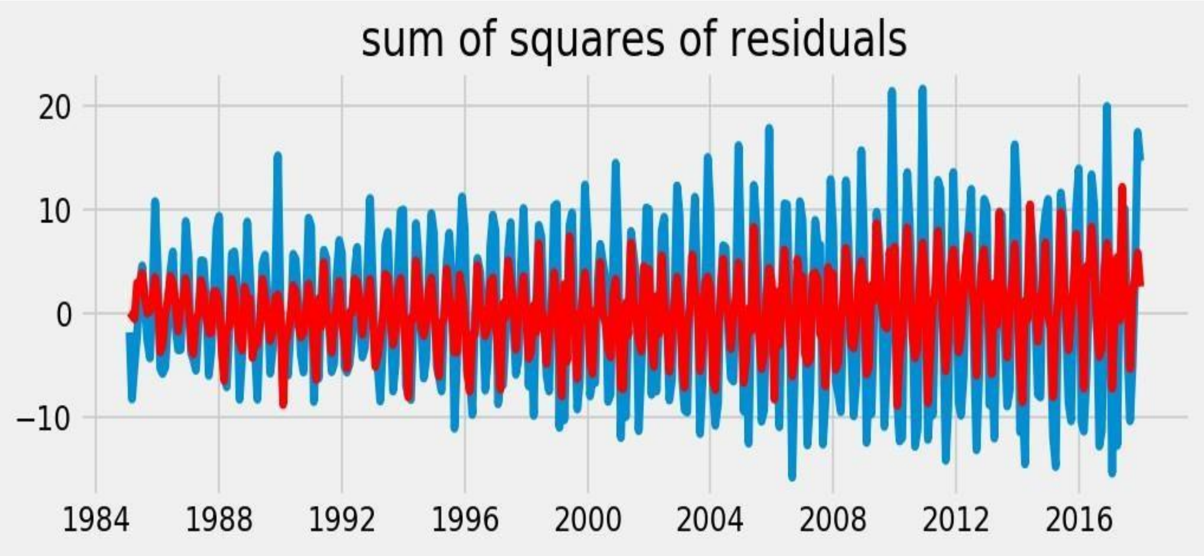
2. ARIMA(1,1,2) :

ARIMA Model Results

Dep. Variable:	D.Consumption	**No. Observations:**	396
Model:	ARIMA(1, 1, 2)	**Log Likelihood**	-1241.559
Method:	css-mle	**S.D. of innovations**	5.547
Date:	Mon, 16 Aug 2021	**AIC**	2493.119
Time:	00:03:11	**BIC**	2513.026
Sample:	02-01-1985 - 01-01-2018	**HQIC**	2501.005

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1108	0.020	5.589	0.000	0.072	0.150
ar.L1.D.Consumption	0.3401	0.055	6.182	0.000	0.232	0.448

ma.L1.D.Consumption	-0.2914	0.034	-8.494	0.000	-0.359	-0.224
ma.L2.D.Consumption	-0.6664	0.032	-20.870	0.000	-0.729	-0.604



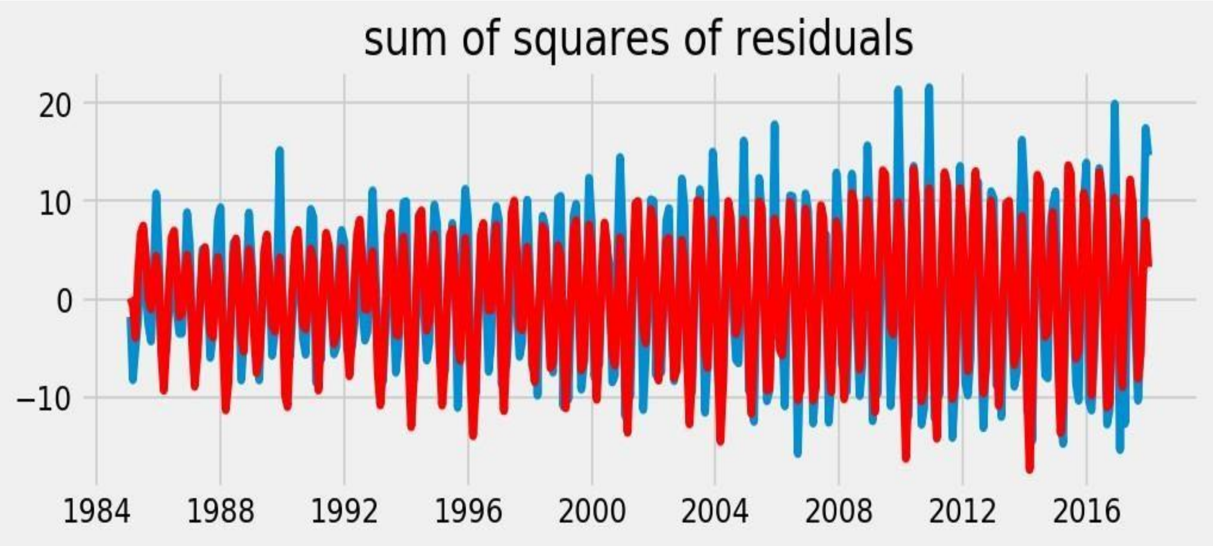
```
RSS : 12227.843382
```

3. ARIMA(3,1,1) :

Dep. Variable:	D.Consumption	**No. Observations:**	396
Model:	ARIMA(3, 1, 1)	**Log Likelihood**	-1103.963
Method:	css-mle	**S.D. of innovations**	3.910
Date:	Mon, 16 Aug 2021	**AIC**	2219.926
Time:	00:02:42	**BIC**	2243.815
Sample:	02-01-1985 - 01-01-2018	**HQIC**	2229.390

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	0.1101	0.024	4.615	0.000	0.063	0.157

ar.L1.D.Consumption	0.7216	0.057	12.685	0.000	0.610	0.833
ar.L2.D.Consumption	-0.6279	0.056	-11.144	0.000	-0.738	-0.517
ar.L3.D.Consumption	-0.2319	0.056	-4.115	0.000	-0.342	-0.121
ma.L1.D.Consumption	-0.8645	0.036	-23.946	0.000	-0.935	-0.794



```
RSS: 6110.656645
```

The model AIC has reduced in ARIMA(3,1,1), which is good. The P Values of the AR1 and MA1 terms have improved and are highly significant (<< 0.05).

Less the **RSS value**, the more **effective** the model is. Here we can see that the ARIMA(1,1,1) and ARIMA(2,1,1) models have quite high RSS compared to ARIMA(3,1,1).

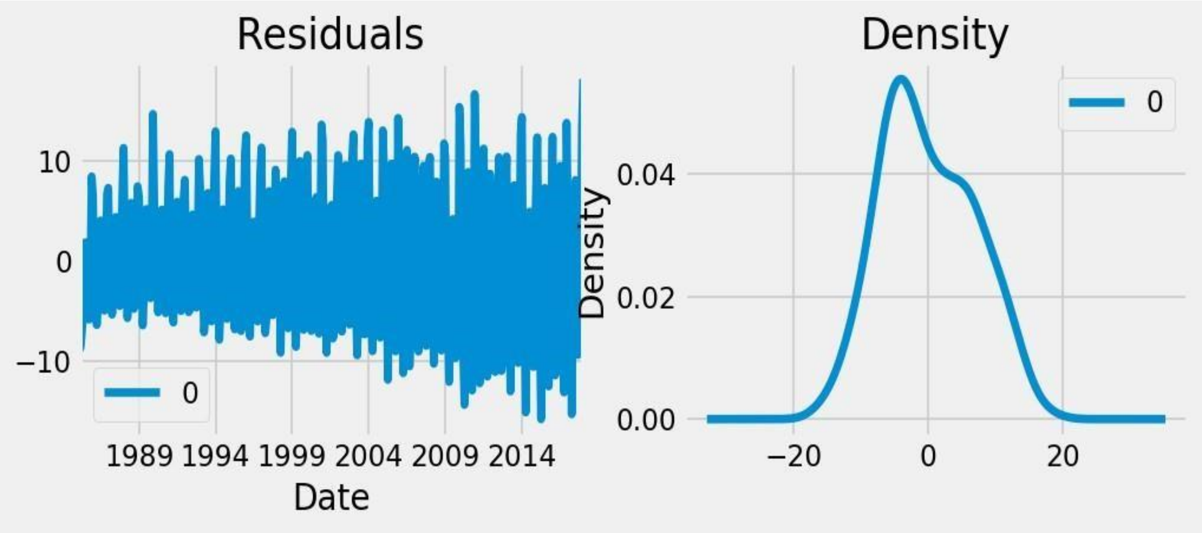
Also if we consider **AIC** criteria, then the model with the lowest AIC is more effective.

Hence **ARIMA(3,1,1)** is best among these three models.

Therefore, the equation of ARIMA(3,1,1) model is

Y(t) = 0.1101 + 0.7216*Y(t-1) – 0.6279*Y(t-2) - 0.2319*Y(t-3) – 0.8645*e(t-1)

Let's plot the residuals to ensure there are no patterns (that is, look for constant mean and variance).

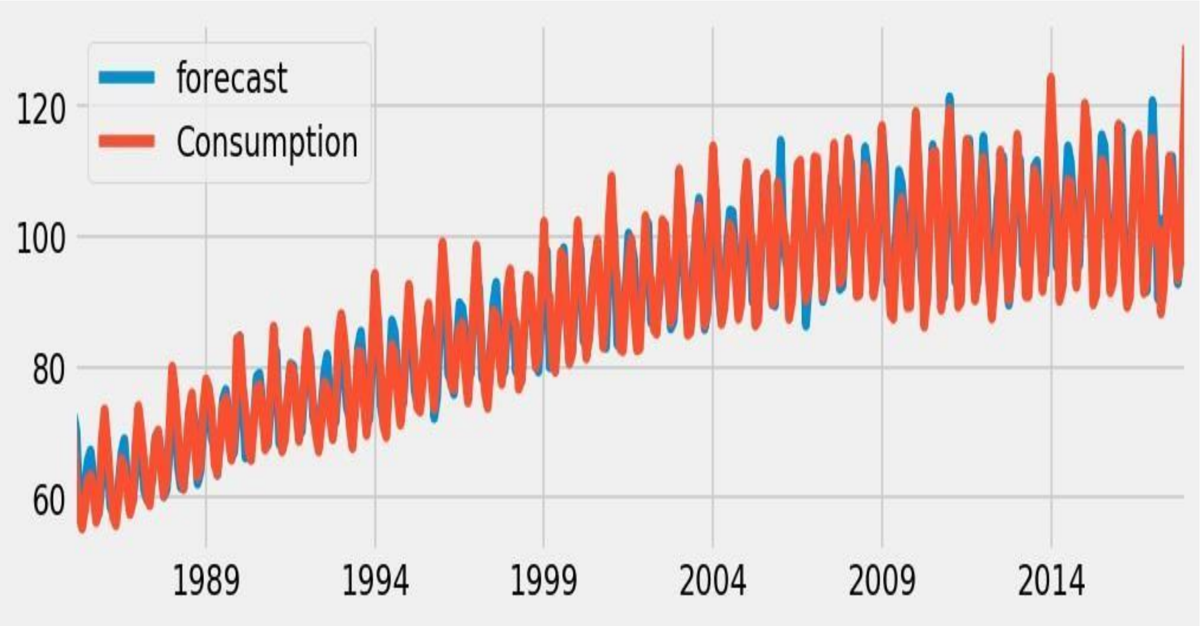


The residual errors seem to fluctuate around a mean of zero and have a uniform variance. The density plot suggest normal distribution with mean zero.

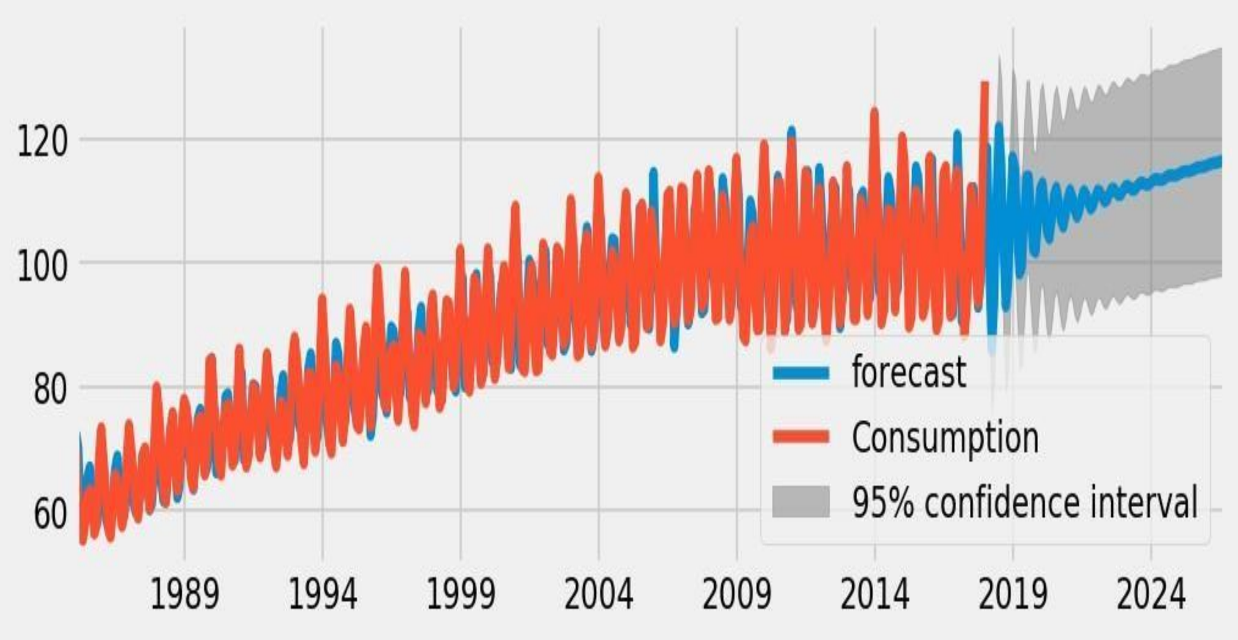
Overall, it seems to be a good fit. Let's forecast.

Forecasting

Actual vs Fitted



Prediction



From the above graph, we calculated the future predictions till 2024 the greyed-out area is the confidence interval that means the predictions will not cross that area.

Conclusion:

Our dataset shows the consumption of electricity from 1985 till 2018 and our goal is to predict electricity consumption for the next 6 years i.e. till 2024.

From the above analysis, **ARIMA(3,1,1)** model provides a decent fit to the consumption of electricity as compared to those provided by ARIMA(1,1,1) and ARIMA(1,1,2) models. On the basis of **RSS** value and **AIC** criteria we get ARIMA(3,1,1) as the best model for the consumption of electricity. Which has the following Mathematical equation,

$$\mathbf{Y(t) = 0.1101 + 0.7216*Y(t-1) - 0.6279*Y(t-2) - 0.2319*Y(t-3) - 0.8645*e(t-1)}$$

By using ARIMA(3,1,1) model we forecasted the electricity consumption for the next 6 years i.e. till 2024.

NOTE: we have used the **Python** software packages for the output presented in this report.

References :

 TIME SERIES ANALYSIS |**Forecasting and Control**|Fifth Edition|GEORGE E. P. BOX, GWILYM M. JENKINS, GREGORY C. REINSEL, GRETA M. LJUNG

- https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average
- https://www.analyticsvidhya.com/blog/2018/09/non-stationary-time-series-python/
- https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/
- https://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/