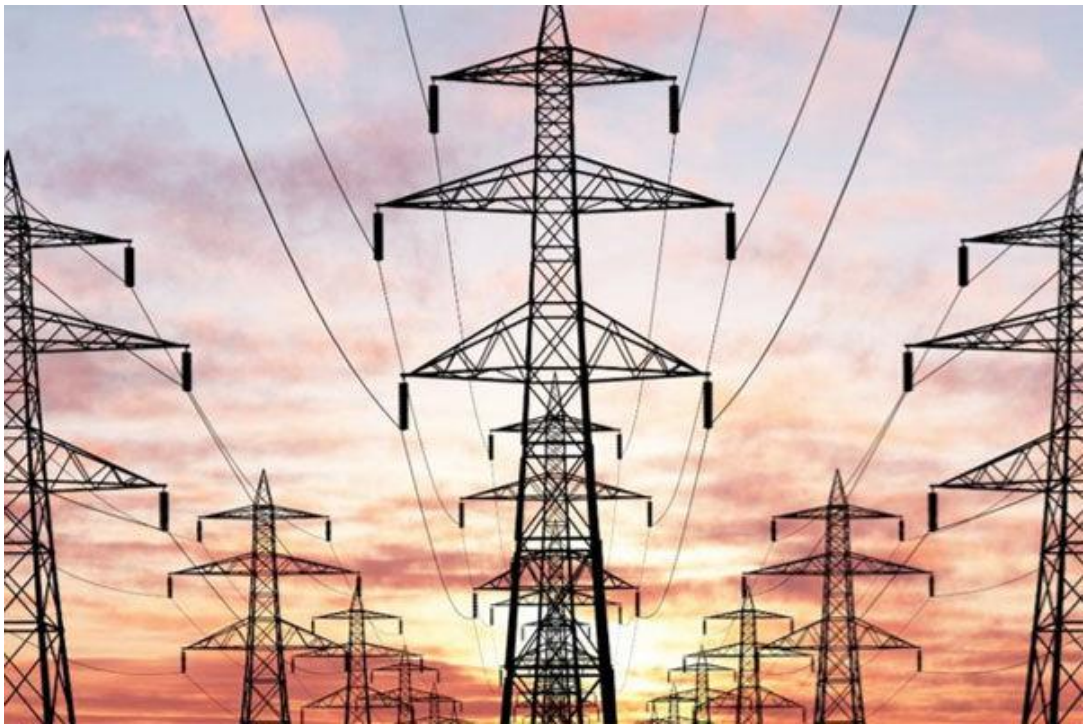
Forecasting and Time Series Methods

Final Project

# "Forecasting Energy Consumption in the US"



# Executive Summary

## Objective:

Global energy consumption grew significantly in last decade, spurred by the sustained economic growth and rising demand in the United States which is one of the world's largest energy consumers. Total energy consumption in the United States reached a record high of 2.3 Gtoe (Gigatonne of Oil Equivalent) in 2018, up 3.5% from 2017.

If we put no cap on this alarming environmental problem we need to be prepared for the repercussions, but if can use our expertise in predicting the trends for next 5 years then we can start act on it to control this harmful climate impacts.

For this very reason we are interested in forecasting the future trend for energy consumption in the United States. We are concentrating our analysis on the following 3 categories,

- Total Primary Energy Consumption
- Total Renewable Energy Consumption
- Total Fossil Fuels Consumption

## Approach:

This energy consumption data set is seasonal time series. For each of the categories,

- We split the dataset into 1973 to 2018 as training data and 2019 as test data
- Checked the stationarity using Augmented Dickey-Fuller Test
- Based on ACF and PACF plots, selected the ARIMA model
- Performed the Residuals test, Ljung-Box tests on the fitted model
- Forecasted for the 2019 test data and calculated the MSE & MAE for the fitted values

## Final Models:

Based on the results of the diagnostic tests, we conclude with the following models as the optimal models to forecast the energy consumption,

| Category | ARIMA | Seasonal Component |
|---|---|---|
| Total Primary Energy Consumption | (1,0,2) | (2,1,1) |
| Total Renewable Energy Consumption | (2,1,2) | (1,1,2) |
| Total Fossil Fuels Consumption | (1,0,2) | (2,1,1) |

We developed a shiny web app which displays the forecast plot, ACF & PACF plots separately for each of the 3 categories of energy consumption. Also, the user can change the order values and check the forecast manually.

# Introduction

Energy is one of the most important of all the worlds' resources. The demand for energy across the world is constantly rising. Primary energy consumption in the United States reached a record high of 101.3 quadrillion British thermal units (Btu) in 2018, up 4% from 2017 and 0.3% above the previous record set in 2007. The increase in 2018 was the largest increase in energy consumption, in both absolute and percentage terms, since 2010. Consumption of fossil fuels—petroleum, natural gas, and coal—grew by 4% in 2018 and accounted for 80% of U.S. total energy consumption. Natural gas consumption reached a record high, rising by 10% from 2017. This increase in natural gas, along with relatively smaller increases in the consumption of petroleum fuels, renewable energy, and nuclear electric power, more than offset a 4% decline in coal consumption.

| | |
|---|---|
| Data Source | **U.S Energy Information and Administration**<br>https://www.eia.gov/totalenergy/data/browser/index.php?tbl=T01.03#/?f=A&start=1949&end=2019&charted=12 |

| | | |
|---|---|---|
| Variables | 7 | |
| Observations | 1834 | |
| Time stamp (Monthly) | 1973 - 2019 | |
| Statistics on time stamped Variable (Quadrillion Btu) | **Min** | 5.43 |
| | **Max** | 9.66 |
| | **Mean** | 7.40 |
| | **Standard Deviation** | 0.95 |

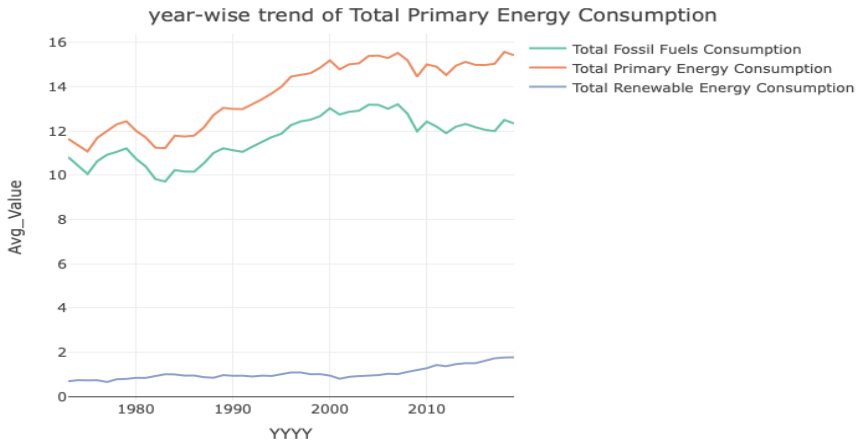
Table: Brief Summary of the Dataset

# Exploratory Data Analysis

A deep dive into basic EDA could give better understanding on the data. Let us look at the structure of the data.

- Time stamped from 1973 to 2019
- For each year, its sum of values is included as month 13. This is should be removed for our analysis
- Value is treated as a character type, which needs to be converted into Numeric, and is the most important attribute for forecasting.
- We can subset some of these attributes as not all contribute for the forecast.

**Understanding Year wise average energy consumption**




Inference:

- Significant growth in consumption of Fossil Fuels
- A slight decreasing trend in Total Energy Consumption
- Trend for Renewable Energy consumption is growing but not significant when compared with the total energy trend.

These basic insights can help us to start our analysis to forecast the future trend. We will now investigate and prepare our data by following steps.

**Sampling the Data**

After exploratory data analysis, we filter our dataset to have only the required columns. We are carrying out time series modelling for each category of energy consumption separately. After the data wrangling, our sample data will look like below,

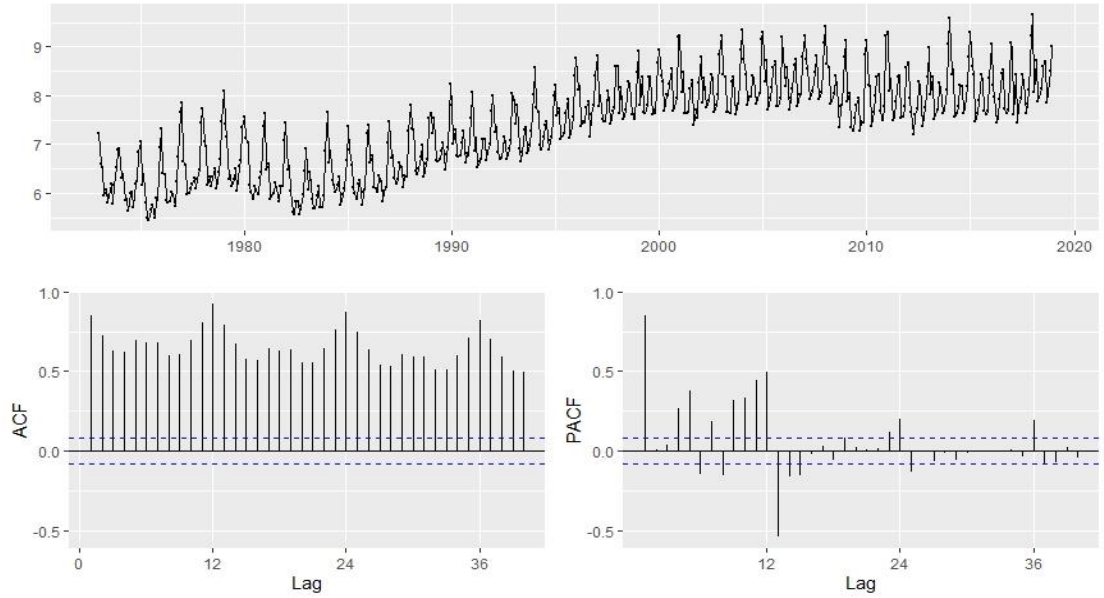
```
##      Year Month Consumption
## 1    1973     1    1.166191
## 2    1973     2    1.050588
## 3    1973     3    1.046856
## 4    1973     4    0.998979
## 5    1973     5    1.014502
## 6    1973     6    1.050077
## 7    1973     7    1.110438
## 8    1973     8    1.135679
## 9    1973     9    1.057732
## 10   1973    10    1.083177
```

Before starting modelling, we split our data into training and testing. Training data will have data from January 1973 to December 2018 and Testing data will have data from January 2019 to December 2019.

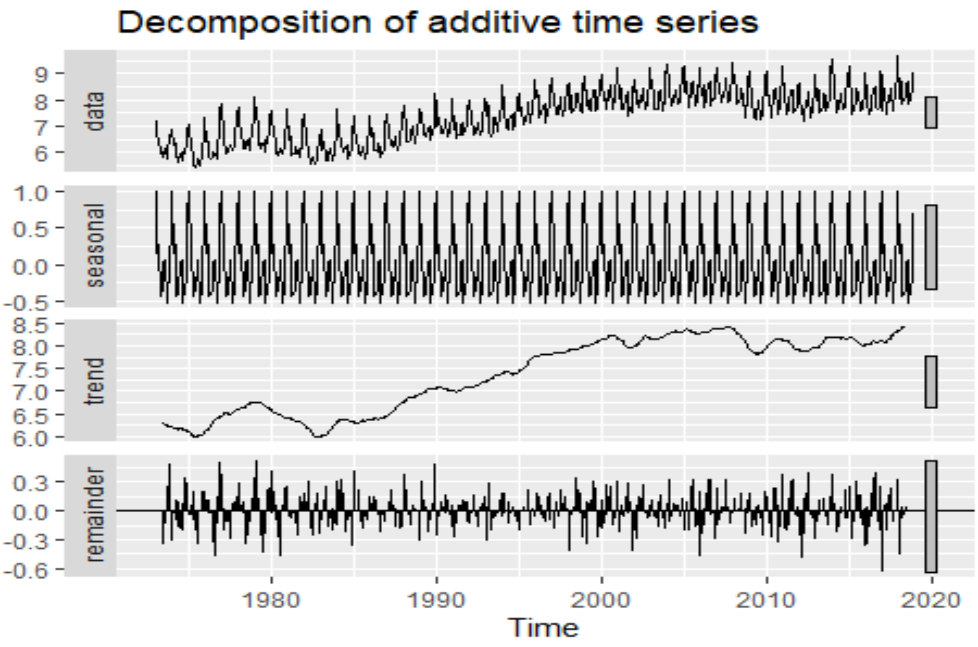
# Model Development

**Category: Total Primary Energy Consumption**

We will look at Decomposition plot and ACF, PACF plots to understand about stationarity of data and investigate about the seasonality.




From the above time series plot, the data does not seem to be stationary as the mean and variance does not seem to be constant across the time period. Also, there is an upward trend in consumption till 2000 and after that there is slow downward trend in the consumption. From the ACF & PACF plots, we do not see any decaying or decreasing pattern.



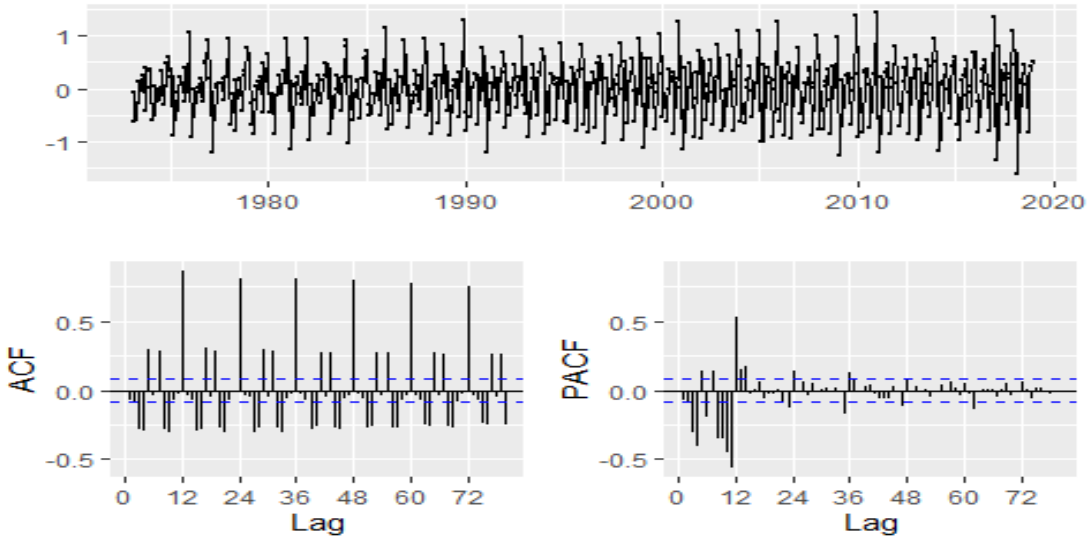
From the above plots, we can see that there is seasonality in the data and increasing and decreasing trend.

Even though, we can clearly interpret that the data is non-stationary from the time series plot, we have not proved it statistically. So, let us perform Augmented Dickey-Fuller Test to statistically prove whether the data is stationary or non-stationary.

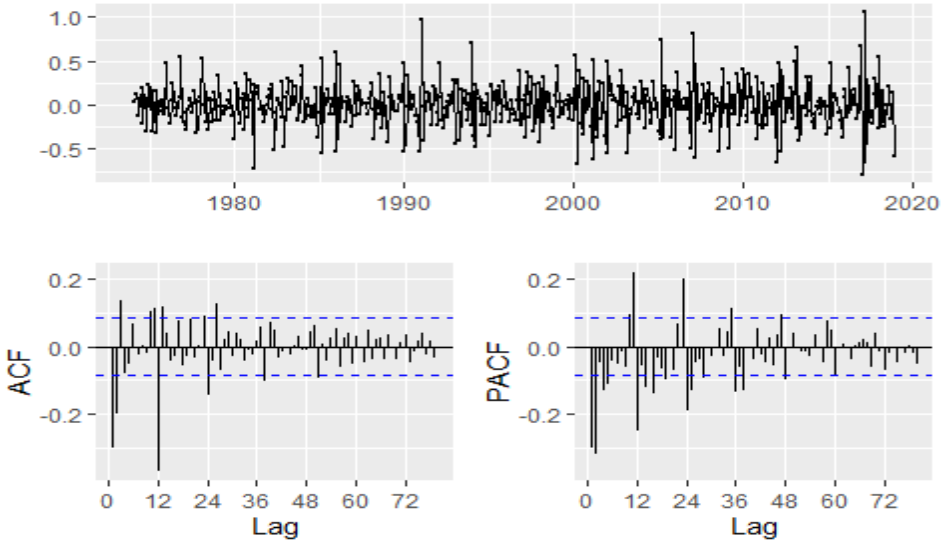
```
aug <- ec_ts %>% adf.test(k=10)
aug

##
##  Augmented Dickey-Fuller Test
##
## data:  .
## Dickey-Fuller = -1.9815, Lag order = 10, p-value = 0.5861
## alternative hypothesis: stationary
```

From the results of the ADF test for different lag values, p-value is 0.5861 for lag k=10 i.e., greater than 0.05. Hence, we cannot reject the null hypothesis which mean data is non-stationary.

We can see that there is trend in the timeseries data, let us do first order differencing to remove the trend. After performing first order differencing let us observe ACF and PACF plots.




We can observe that trend is removed. From the ACF plot, we can see a spike at lag 12 and there is correlation with the 12th lag element each time, indicating seasonality. Also, Auto Correlation is slowly decreasing at lags 12,24,36,48,60 & 72. So, we can take seasonal first order differencing to make it stationary. After taking seasonal differencing with lag 12, our ACF and PACF plots looks like below



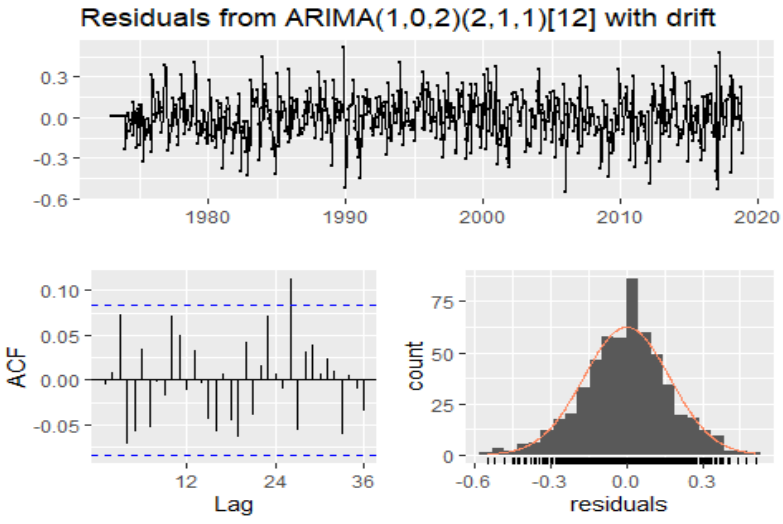

From the ACF plots, we can see that there is significant spikes at lag 1&2 is the non-seasonal MA(2) component and spike at 12 is the Seasonal MA(3) component and from the PACF plot, we can see that there are spikes at lags 12,24,36 & 48 is the seasonal AR(4) component and spikes at lags 1,2&9 are non-seasonal AR(3) components.

So, when we perform seasonal difference and normal difference for removing trend, the series is stationary. Our model estimate is ARIMA(3,1,2)(4,1,3)(12).

We fit the above model and get the model from auto arima function. We will evaluate both the models and pick the best one based on AIC values.

Auto arima returned us the best model as ARIMA(1,0,2)(2,1,2)[12] and the AIC value for this model is **-341.41** which is better than our estimate model which has AIC of **-331.86**. Therefore, we will go ahead with model returned by auto arima.

Next step is to perform residual analysis and Ljung-box test to see if the fitted model is adequate.

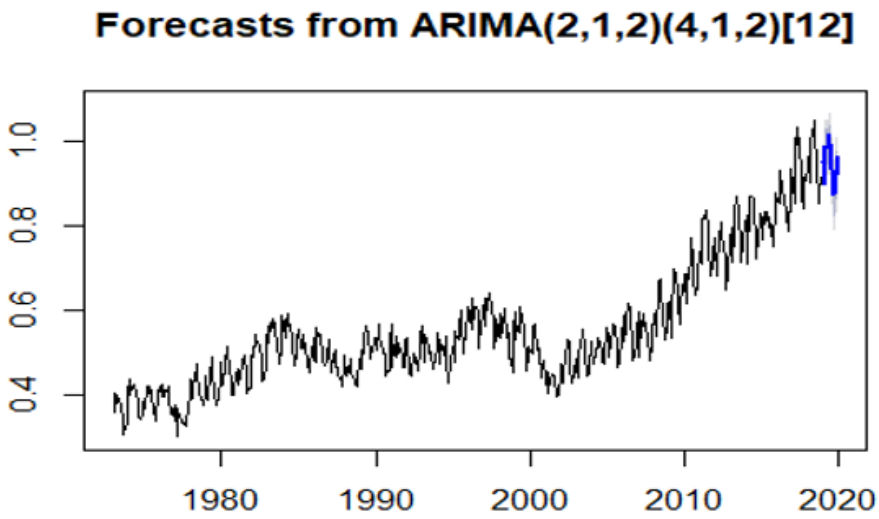


```
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,0,2)(2,1,1)[12] with drift
## Q* = 26.764, df = 17, p-value = 0.06162
##
## Model df: 7.   Total lags used: 24
```

From the above plot, we can see that residuals are normally distributed, and our test also returns p value greater than 0.05. Hence, the model is a good fit.

**Forecasting**

After having the final model, we use this model to forecast the energy consumption on test data set and evaluate the metrics.




MSE for the above model on test data set is **4.808**

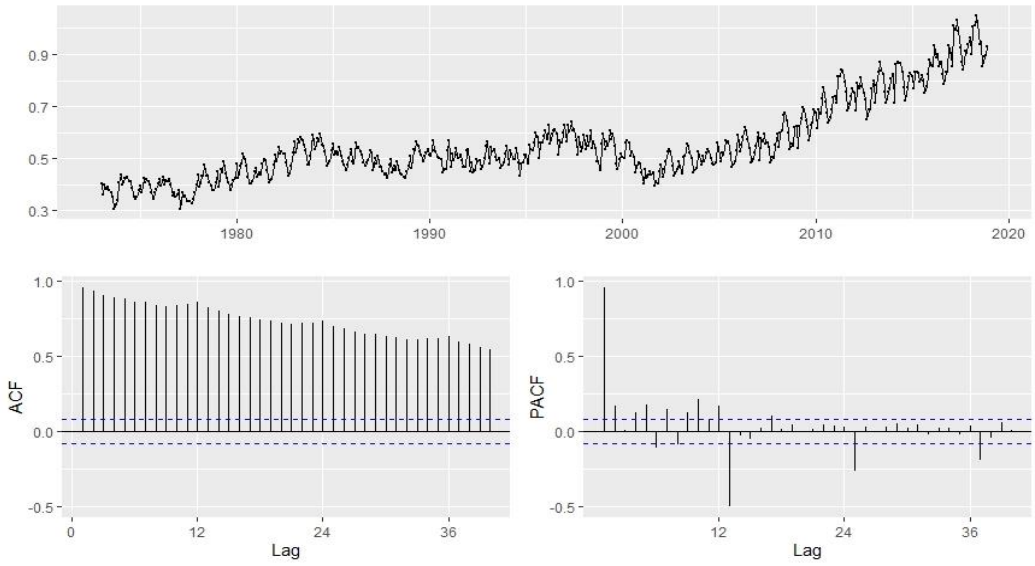
MAE for the above model on test data set is **2.174**

---

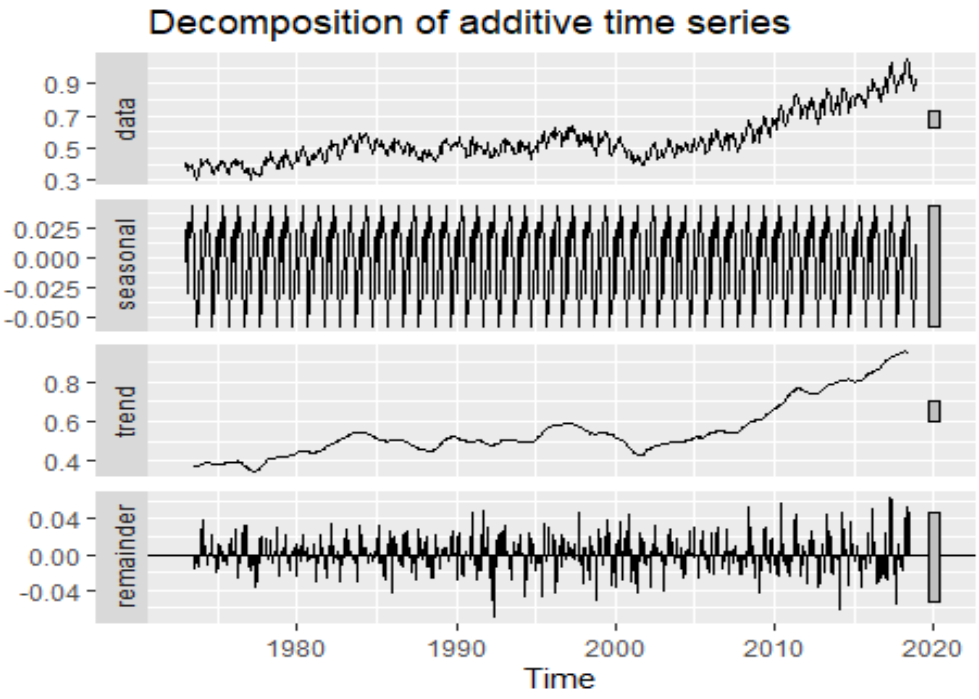
**Category: Total Renewable energy consumption**

We will look at Decomposition plot and ACF, PACF plots to understand about stationarity of data and investigate about the seasonality.

We can see that there is a drastic upward trend in the renewable energy consumption starting from the year 2000. From the ACF, we do not see any identifiable pattern. From PACF plot, we can see that it decayed exponentially.

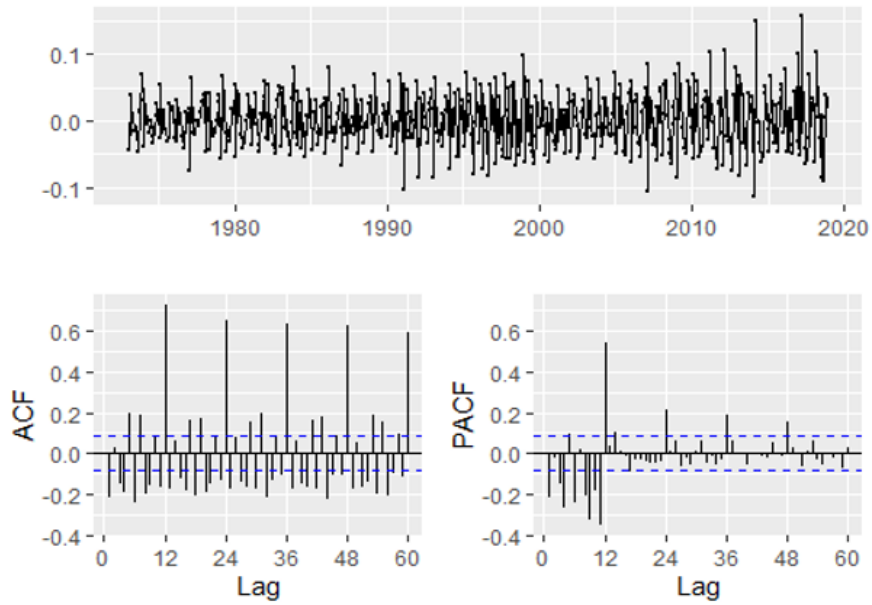



We can see that the data is not stationary for the time series data for total renewable energy consumption and that is confirmed by ADF test which gives the p-value is greater than 0.05.

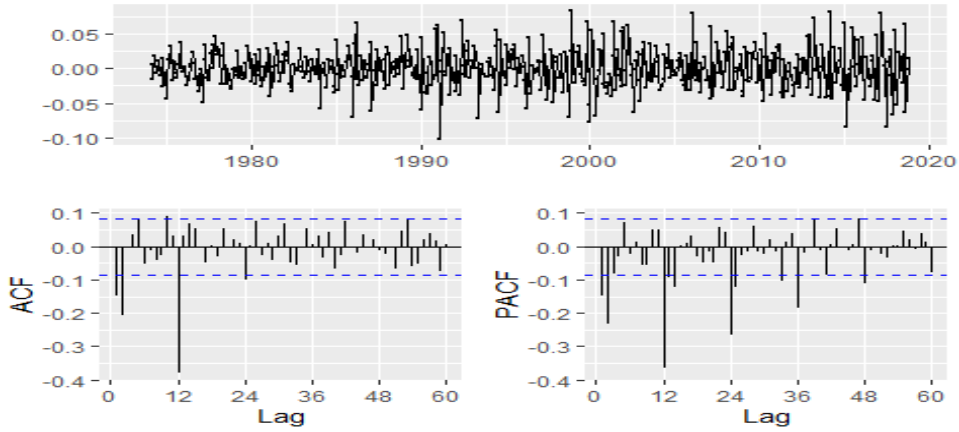



We can see that there is trend in the timeseries data, let us do first order differencing to remove the trend.

Now, we can see that the trend is removed from the time series data for total renewable energy consumption.




We can see that there is Spike at lag 12, there is correlation with the 12th lag element each time, indicating seasonality. Also, Auto Correlation is slowly decreasing at lags 12,24,36,48 & 60. So, we can take seasonal first order differencing to make it stationary. After performing the above steps, ACF and PACF plots look like

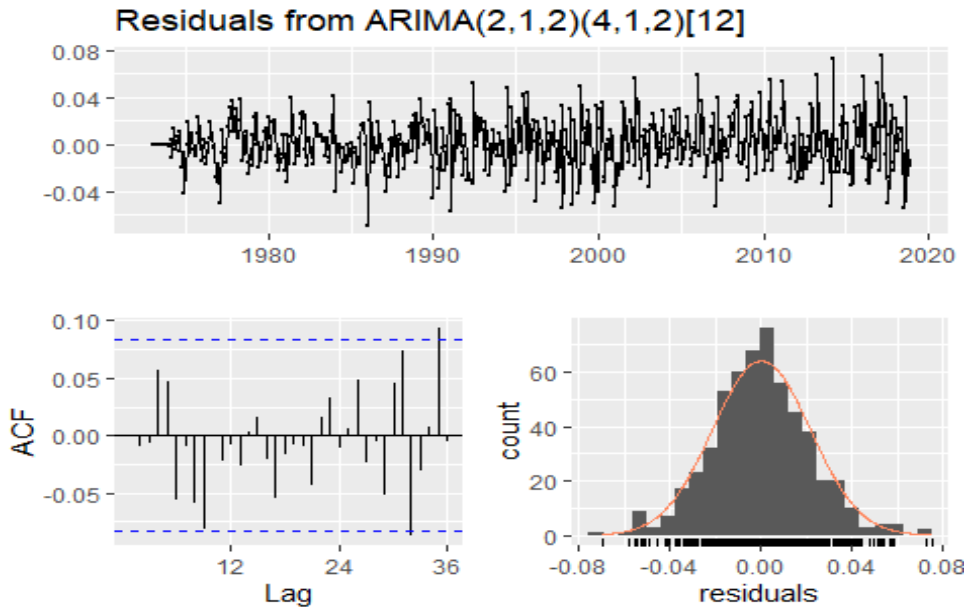



From the ACF plots, we can see that the spike at lag 1&2 is the non-seasonal MA(2) component and spike at 12&24 is the Seasonal MA(2) component and from the PACF plot, we can see that there are spikes at lags 12,24,36 & 48 is the seasonal AR(4) component and spikes at lags 1 & 2 are non-seasonal AR(2) components.

Our model estimate is ARIMA(2,1,2)(4,1,2)(12).

Let us fit auto.arima to the time series data and compare the results for both the models. Auto arima returns ARIMA(2,1,2)(1,1,2)[12] and the AIC value for this model is **-2563.41** which is better than our estimate model which has AIC of **-2557.86**. Therefore, we will go ahead with model returned by auto arima.

Next step is to perform residual analysis and Ljung-box test to see if the fitted model is adequate.

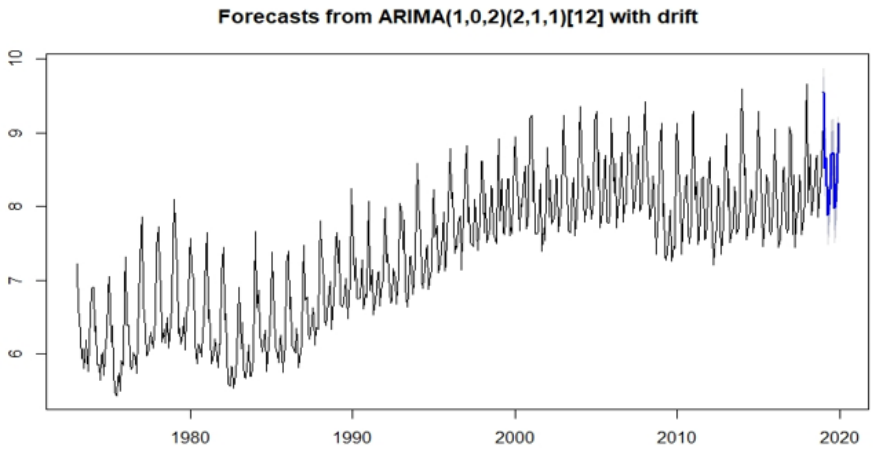



```
 ##  Ljung-Box test
##
## data:  Residuals from ARIMA(2,1,2)(4,1,2)[12]
## Q* = 15.5, df = 14, p-value = 0.3449
##
## Model df: 10.   Total lags used: 24
```

From the above plot, we can see that residuals are normally distributed, and our test also returns p value greater than 0.05. Hence, the model is a good fit.

**Forecasting**

After having the final model, we use this model to forecast the energy consumption on test data set and evaluate the metrics.

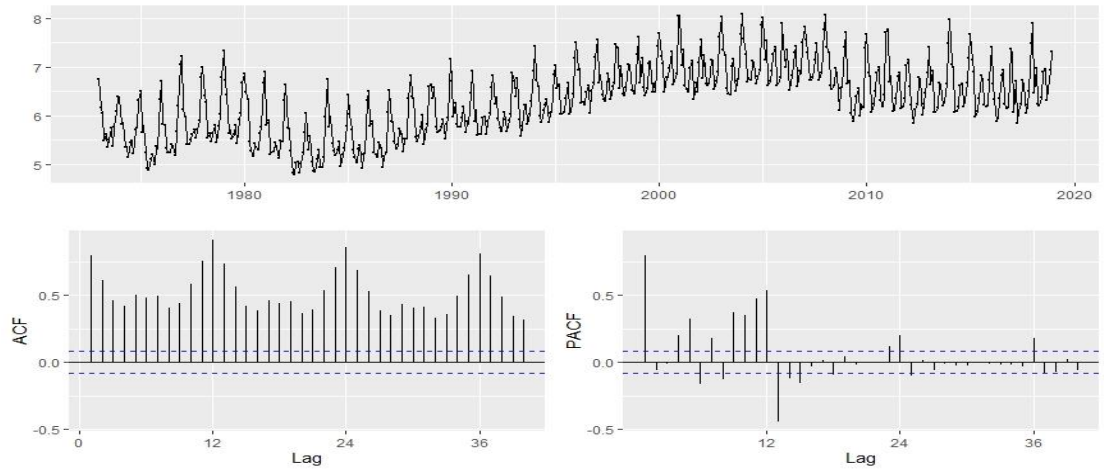



MSE for the above model on test data set is **0.338**

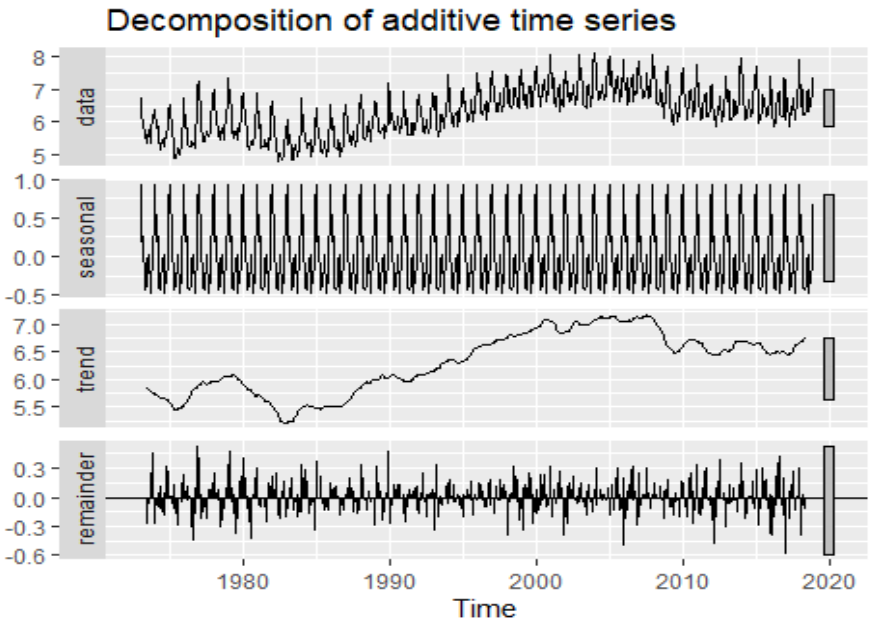
MAE for the above model on test data set is **0.581**

**Category: Total Fossil Fuels Consumption**

We will look at Decomposition plot and ACF, PACF plots to understand about stationarity of data and investigate about the seasonality.

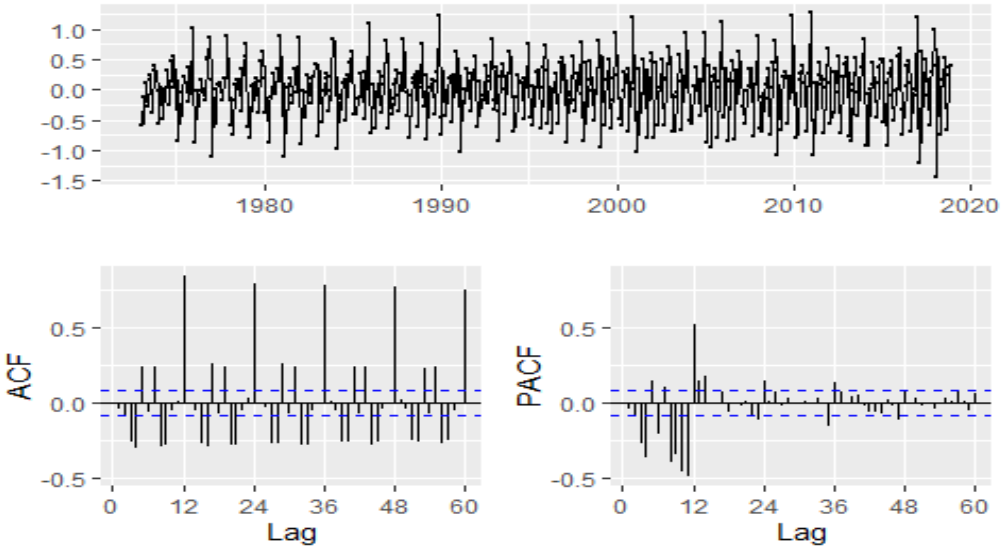



We can see that there is an upward trend in the fossil fuel consumption at around the year 1982 and there is a downward trend starting from the year 2000. From the ACF and PACF plots, we do not see any identifiable pattern.



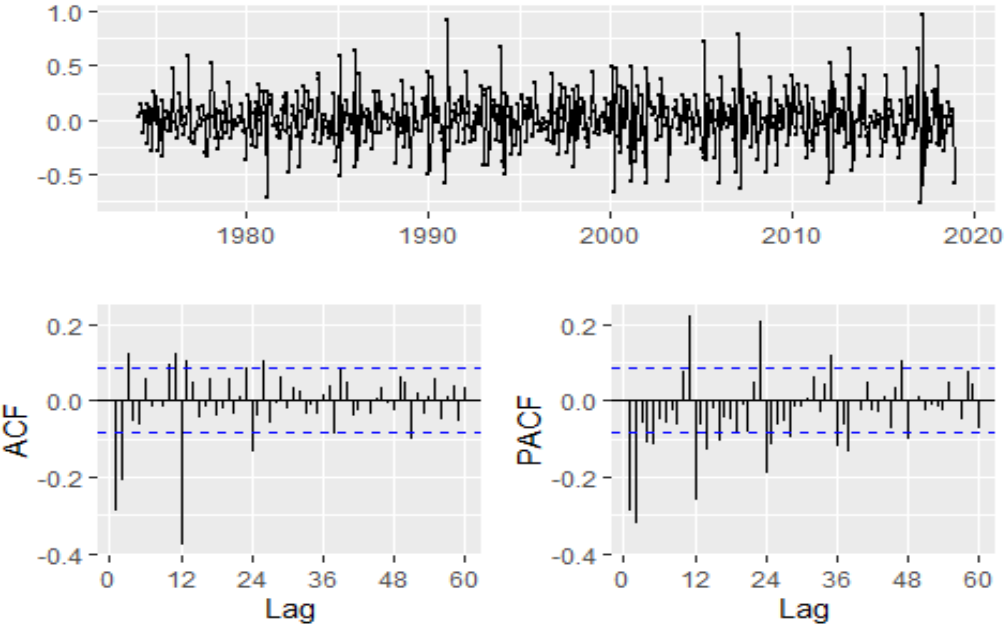

We can see that the data is not stationary for the time series data for total fossil fuels consumption and it is also confirmed by the p-value of ADF test which is greater than 0.05.

We can see that there is trend in the timeseries data, let us do first order differencing to remove the trend.




Now, we can see that the trend is removed from the time series data for total fossil fuel consumption.

We can see that there is Spike at lag 12, there is correlation with the 12th lag element each time, indicating seasonality. Also, Auto Correlation is slowly decreasing at lags 12,24,36,48 & 60. So, we can take seasonal first order differencing to make it stationary. After performing above steps, plot look like below.



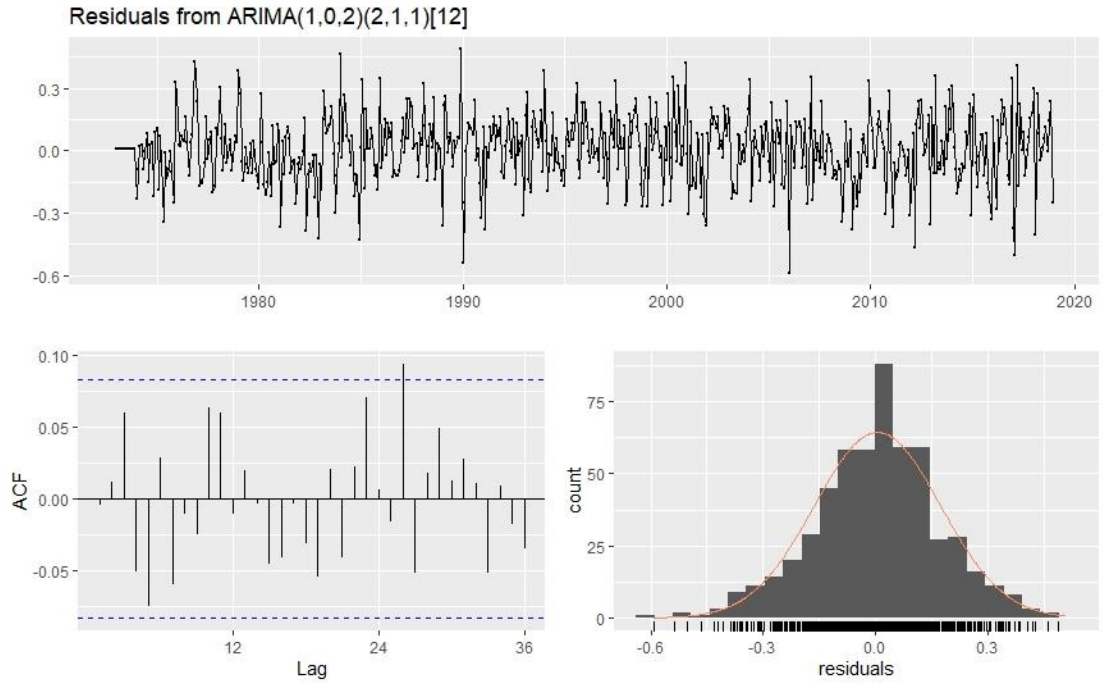

From the ACF plots, we can see that the spike at lag 1&2 is the non-seasonal MA(2) component and spike at 12&24 is the Seasonal MA(2) component and from

the PACF plot, we can see that there are significant spikes at lags 12&24 is the seasonal AR(2) component and spikes at lags 1 & 2 are non-seasonal AR(2) components.

Our estimated model is ARIMA(2,1,2)(2,1,2)(12).

Let us fit auto.arima to the time series data and compare the results for both the models. Auto arima returns ARIMA(1,0,2)(2,1,1)[12] and the AIC value for this model is **-353.41** which is better than our estimate model which has AIC of **-346.86**. Therefore, we will go ahead with model returned by auto arima.

Next step is to perform residual analysis and Ljung-box test to see if the fitted model is adequate.

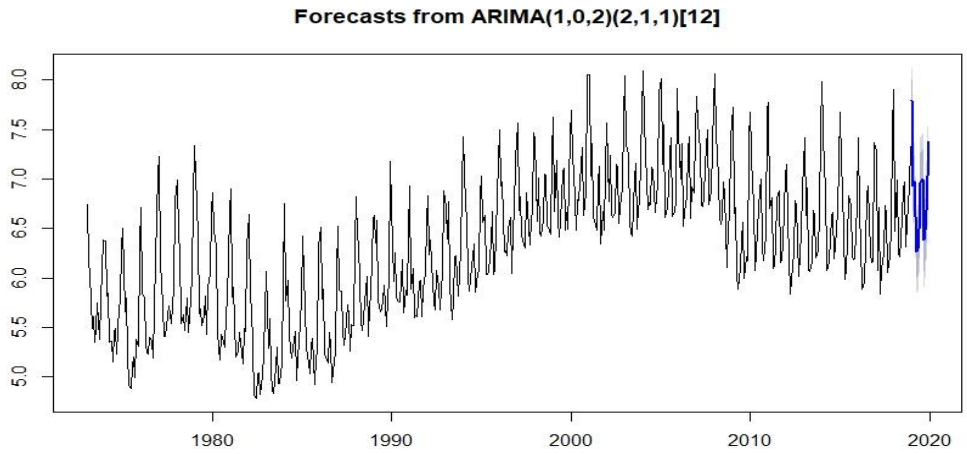



```
##  Ljung-Box test
##
## data:  Residuals from ARIMA(2,1,2)(2,1,2)[12]
## Q* = 23.144, df = 16, p-value = 0.11
##
## Model df: 8.   Total lags used: 24
```

From the results of the test, we can see that p-value is greater than 0.05. Hence, the model is a good fit.

**Forecasting**

After having the final model, we use this model to forecast the energy consumption on test data set and evaluate the metrics.




MSE for the above model on test data set is **0.956**

MAE for the above model on test data set is **0.947**

---

# Conclusion

According to the U.S. Energy Information Administration's (EIA) International Energy Outlook 2019 (IEO2019), global electric power generation from renewable sources will increase more than 20% throughout the projection period (2018–2050), providing almost half of the world's electricity generation in 2050.

In that same period, global coal-fired generation will decrease 13%, representing only 22% of the generation mix in 2050. EIA projects that worldwide electricity generation will grow by 1.8% per year through 2050.

Based on our forecasting it is evident and strongly bolster the above statement. From out forecast plots we can observer that the consumption of renewable energy is increasing and the consumption of energy from fossil fuels is gradually decreasing.

---

# Shiny Web App

- With this energy consumption dataset, we developed a shiny web app that shows the forecast and plots from our final model
- Also, the user can change the ARIMA values
- The category of energy consumption can be chosen from the category dropdown