# Topic: Forecasting – Time Series

A picture containing shape, arrow

Description automatically generated

**Instructions**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: Prajay B. Urkude**

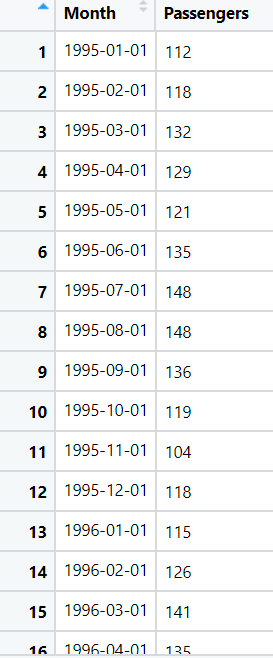
**Batch ID: 16092021**

**Topic: Forecasting – Time Series**

**Problem Statement: -**

1. The dataset consists of monthly totals of international airline passengers from 1995 to 2002. Our main aim is to predict the number of passengers for the next five years using time series forecasting. Prepare a document for each model explaining how many dummy variables you have created and also include the RMSE value for each model.

## File: - Airlines.xlsx



A picture containing shape, arrow

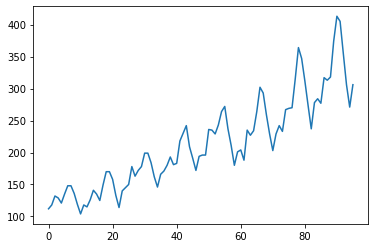
Description automatically generated Ans :- Business Objectives:-

to predict the number of passengers for the next five years using time series forecasting.

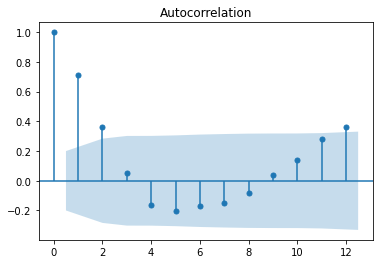
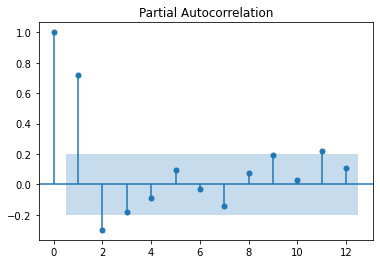
|  |  |  |  |
| --- | --- | --- | --- |
| Name of features | Description | Type | Relevance |
| Month | Date | Quantitative, Nominal | Relevant |
| Passengers | No. of passengers with date | Quantitative, Ratio | Relevant |

**Steps To build the Forecasting model.**

**1) Model Based Approach:**

* Import the libraries such as pandas, numpy, matplotlib, statsmodels . From statsmodel library import formula package and import api function.
* Import the data and do the data manipulation.
* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.
* Add t column which has the value in the serial no. to the data. Add another column which has the column square value of t. Add another one column which takes the log value of t.
* Visualize the timeplot between time on x axis and passengers on the Y axis.

From the time plot we can see that there is level present, trend is upward and the seasonality may be additive or multiplicative which we can find from the regression model.

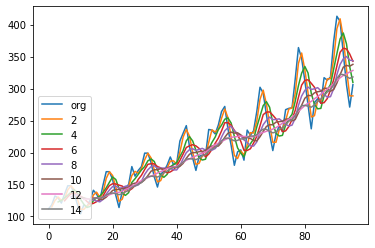
* Partition of the data to be done as train and test. The data to be partition in such a way that the test data should be equal to greater than equal to seasonality.
* To prepare different types of regression models such as linear, exponential, quadratic, additive seasonality, additive seasonality with the quadratic trend, multiplicative seasonality, multiplicative seasonality with linear trend by predicting the model on test data and calculate the root mean squared error values i.e. RMSE value for each model.
* Lesser the RMSE value better is the model. In the given data **additive seasonality with quadratic trend gives the lesser RMSE value**, so we will use this model for forecasting on the new data.
* In the given data the seasonality is additive.
* Import the new data for forecasting the passengers for next 5 years.
* Create a model by using the Additive seasonality with quadratic trend regression model on the previous full data and evaluate the data on the new data. It gives the predicted values for the next 5 years.
* The Total forecasting value is the predicted value + error between predicted value and the actual value.
* So to calculate the error we use Autoregression model.
* Calculate the full residual from the full data and plot the ACF and PACF plot for the 12 lags values.
* ACF is an (complete) auto-correlation function gives values of auto-correlation of any time series with its lagged values.
* PACF is a partial auto-correlation function. It finds correlations of present with lags of the residuals of the time series.

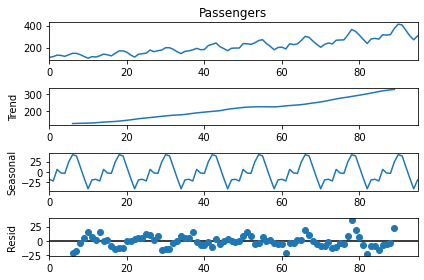
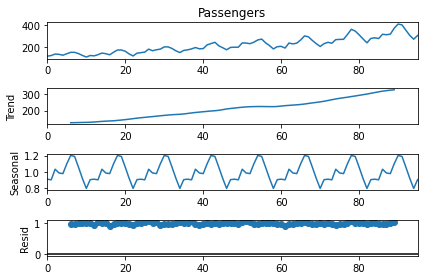
From both the plots we can see that 1st lag value has the highest autocorrelation function.

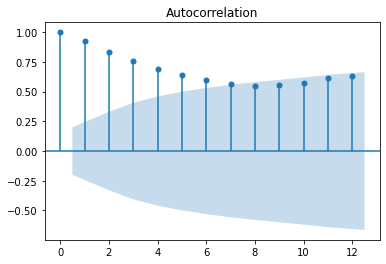
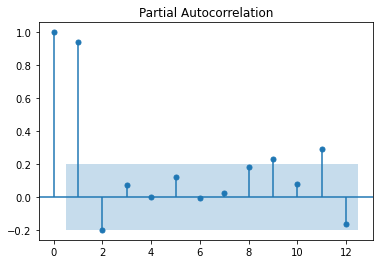
* Build the autoregression model by taking the lag = 1 and evaluate the model on the new data to find out the residuals of the new predicted data.
* Calculate the final prediction by adding the prediction values and the predicted residuals.
* Finally, we get the next 5 years forecasted values monthwise.

**2) Data Driven Approach:**

* Import the statsmodels libraries and the functions which are useful for the data driven technique.
* Creating a function to calculate the mean absolute percentile error value (MAPE).
* Calculate the MAPE value by calculating the moving average for the 12 month and visualize with the timeplot.

From the graph we can see that at the moving average of the 12 values, the data has no seasonality.

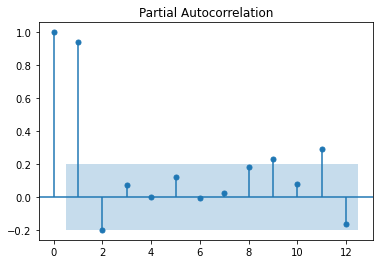
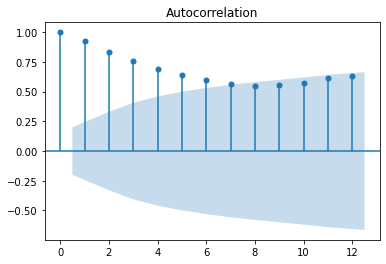
* Time series decomposition is the process of separating data into its core components. Time series decomposition plot using Moving Average for additive and multiplicative model respectively.
* ACF and PACF plot on original data sets



1st lag value has the highest correlation function value than rest of the 12 value.

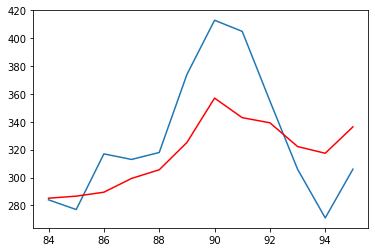
* Build the data driven model by using the different methods such as simple Exponential Method, Holt method, Holts winter exponential smoothing with additive seasonality and additive trend, Holts winter exponential smoothing with multiplicative seasonality and additive trend and evaluate the model on the test datasets and calculate the MAPE values.
* Lesser the MAPE value better is the model.
* For the given training data Holts winter exponential smoothing with multiplicative seasonality and additive trend model has the least MAPE value so we finalize this model for forecasting.
* Prepare the model by using the Holts winter exponential smoothing with multiplicative seasonality and additive trend method on the complete original data and then evaluate the model on the new data for which we have to forecast the passengers’ values and find out the predicted values.

**3) ARIMA Method: -**

* Import the statsmodels libraries and the ARIMA functions which are useful for the ARIMS technique.
* Plot the ACF and the PACF plot and find out the lag value which has the highest autocorrelation Function.
* ****

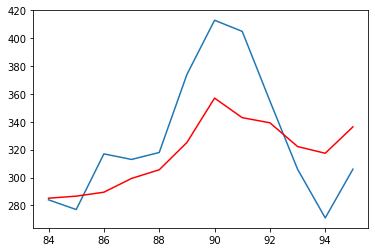
From the above plot we can see that 1st lag value has the highest correlation function value than rest of the 12 value.

* Fit the ARIMA function by taking AR = 1, I,1 and MA =12 and forecast for the next 12 months and calculate the RMSE values.
* Plot the forecasted values with the actual outcomes.



The red line indicates the forecasted values for the next 12 months.

* Auto-ARIMA - Automatically discover the optimal order for an ARIMA model by using the pmdarima library. By using this function, we get the best parameters for ARIMA model is AR = 2, I =1, MA =1
* Again, Fit the ARIMA model by taking the best parameters and forecast for the next 12 months and find the RMSE values. Plot the graph which has the forecasted values with the red line and original value of test datasets with blue line.



* Forecast for the next 5 years i.e. for next 60 month by using this final model.

**Problem Statement: -**

1. The dataset consists of quarterly sales data of Coca-Cola from 1986 to 1996. Predict sales for the next two years by using time series forecasting and prepare a document for each model explaining how many dummy variables you have created and also include the RMSE value for each model.

**File:- CocaCola\_Sales\_RawData.xlsx**



A picture containing shape, arrow

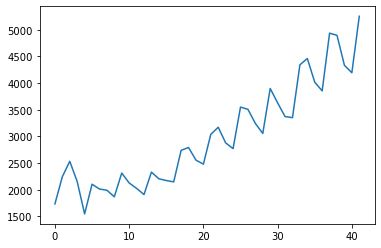
Description automatically generatedAns :- Business Objectives:-

To predict the sales for the next two years by using time series forecasting

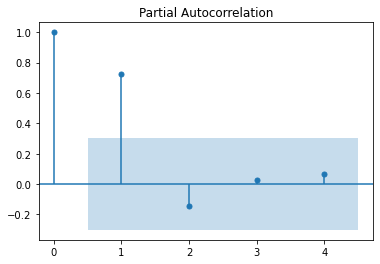
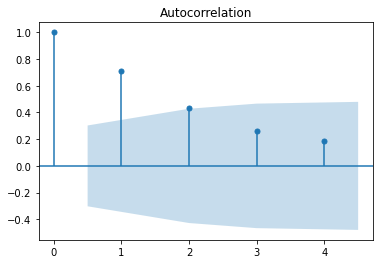
|  |  |  |  |
| --- | --- | --- | --- |
| **Name of features** | **Description** | **Type** | **Relevance** |
| Quarter | Quarter of the year | Quantitative, Nominal | Relevant |
| Sales | Quarterly sales of the coca cola | Quantitative, Ratio | Relevant |

**Steps To build the Forecasting model.**

**1) Model Based Approach:**

* Import the libraries such as pandas, numpy, matplotlib, statsmodels . From statsmodel library import formula package and import api function.
* Import the data and do the data manipulation.
* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.
* Add t column which has the value in the serial no. to the data. Add another column which has the column square value of t. Add another one column which takes the log value of t.
* Visualize the timeplot between time on x axis and sales on the Y axis.

From the time plot we can see that there is level present, trend is upward and the seasonality may be additive or multiplicative which we can find from the regression model.

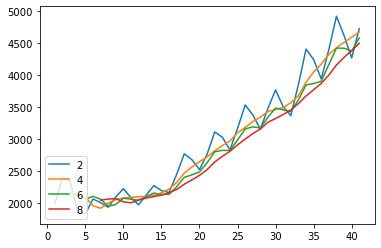
* Partition of the data to be done as train and test. The data to be partition in such a way that the test data should be equal to greater than equal to seasonality.
* To prepare different types of regression models such as linear, exponential, quadratic, additive seasonality, additive seasonality with the quadratic trend, multiplicative seasonality, multiplicative seasonality with linear trend by predicting the model on test data and calculate the root mean squared error values i.e. RMSE value for each model.
* Lesser the RMSE value better is the model. In the given data **Multiplicative seasonality with quadratic trend gives the lesser RMSE value**, so we will use this model for forecasting on the new data.
* In the given data the seasonality is multiplicative.
* Import the new data for forecasting the passengers for next 5 years.
* Create a model by using the Additive seasonality with quadratic trend regression model on the previous full data and evaluate the data on the new data. It gives the predicted values for the next 5 years.
* The Total forecasting value is the predicted value + error between predicted value and the actual value.
* So to calculate the error we use Autoregression model.
* Calculate the full residual from the full data and plot the ACF and PACF plot for the 12 lags values.
* ACF is an (complete) auto-correlation function gives values of auto-correlation of any time series with its lagged values.
* PACF is a partial auto-correlation function. It finds correlations of present with lags of the residuals of the time series.

From both the plots we can see that 1st lag value has the highest autocorrelation function.

* Build the autoregression model by taking the lag = 1 and evaluate the model on the new data to find out the residuals of the new predicted data.
* Calculate the final prediction by adding the prediction values and the predicted residuals.
* Finally, we get the next 2 years forecasted values month wise.

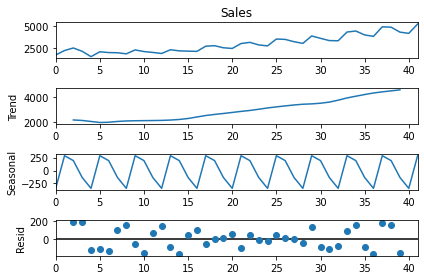
**2) Data Driven Approach:**

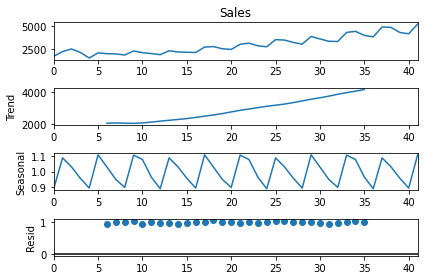
* Import the statsmodels libraries and the functions which are useful for the data driven technique.
* Creating a function to calculate the mean absolute percentile error value (MAPE).
* Calculate the MAPE value by calculating the moving average for the 4 quarters and visualize with the timeplot.



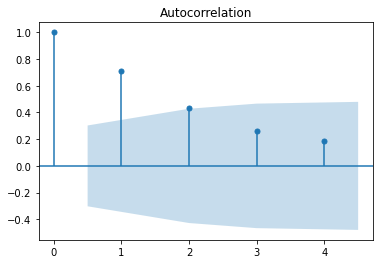
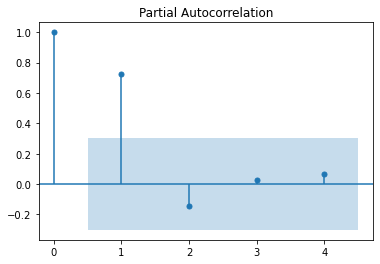
From the graph we can see that at the moving average of the 4 and 8 values, the data has no seasonality.

* Time series decomposition is the process of separating data into its core components. Time series decomposition plot using Moving Average for additive and multiplicative model respectively.





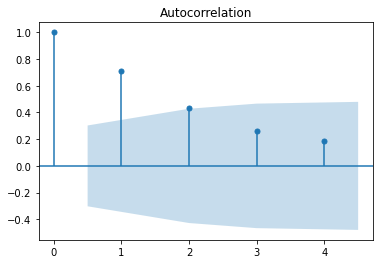
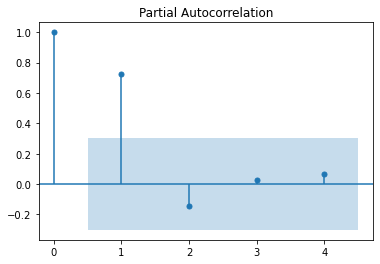
* ACF and PACF plot on original data sets



1st lag value has the highest correlation function value than rest of the 4 value.

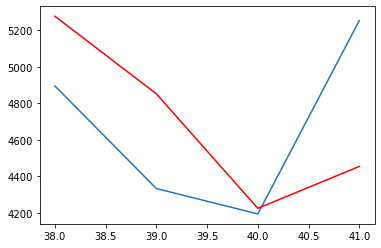
* Build the data driven model by using the different methods such as simple Exponential Method, Holt method, Holts winter exponential smoothing with additive seasonality and additive trend, Holts winter exponential smoothing with multiplicative seasonality and additive trend and evaluate the model on the test datasets and calculate the MAPE values.
* Lesser the MAPE value better is the model.
* For the given training Holts winter exponential smoothing with additive seasonality and additive trend has the least MAPE value so we finalize this model for forecasting.
* Prepare the model by using the Holts winter exponential smoothing with multiplicative seasonality and additive trend method on the complete original data and then evaluate the model on the new data for which we have to forecast the sales’ values and find out the predicted values.

**3) ARIMA Method: -**

* Import the statsmodels libraries and the ARIMA functions which are useful for the ARIMS technique.
* Plot the ACF and the PACF plot and find out the lag value which has the highest autocorrelation Function.
* 

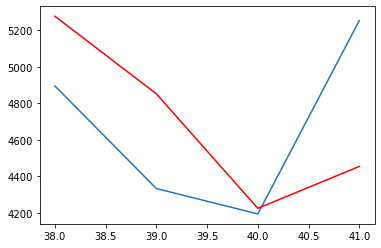
From the above plot we can see that 1st lag value has the highest correlation function value than rest of the 4 value.

* Fit the ARIMA function by taking AR = 1, I=1 and MA =4 and forecast for the next 4 quarters and calculate the RMSE values.
* Plot the forecasted values with the actual outcomes.

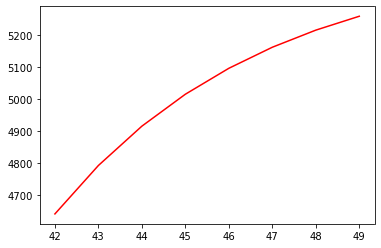


The red line indicates the forecasted values for the next 4 months.

* Auto-ARIMA - Automatically discover the optimal order for an ARIMA model by using the pmdarima library. By using this function, we get the best parameters for ARIMA model is AR = 0, I =1, MA =2
* Again, Fit the ARIMA model by taking the best parameters and forecast for the next 12 months and find the RMSE values. Plot the graph which has the forecasted values with the red line and original value of test datasets with blue line.



* Forecast for the next 2 years i.e. for next 8 quarters by using this final model.



**Problem Statement: -** A picture containing shape, arrow

Description automatically generated

A plastics manufacturing plant has recorded their monthly sales data from 1949 to 1953. Perform forecasting on the data and bring out insights from it and forecast the sale for the next year.

Plastic Sales.csv

A picture containing table

Description automatically generated

**Business Objectives: -**

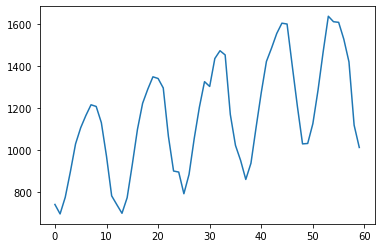
To predict the sales for the next years based on the previous data by using time series forecasting

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of features** | **Description** | **Type** | **Relevance** |
| Month | Month of the year | Quantitative, Nominal | Relevant |
| Sales | Monthly sales of the data | Quantitative, Ratio | Relevant |

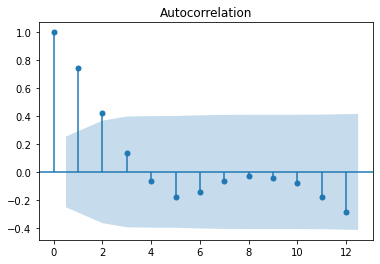
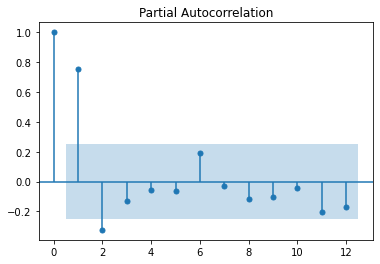
**Steps To build the Forecasting model.**

**1) Model Based Approach/ Regression Model :**

* Import the libraries such as pandas, numpy, matplotlib, statsmodels . From statsmodel library import formula package and import api function.
* Import the data and do the data manipulation.
* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.
* Add t column which has the value in the serial no. to the data. Add another column which has the column square value of t. Add another one column which takes the log value of t.
* Visualize the timeplot between time on x axis and sales on the Y axis.



From the time plot we can see that there is level present, trend is upward and the seasonality may be additive or multiplicative which we can find from the regression model.

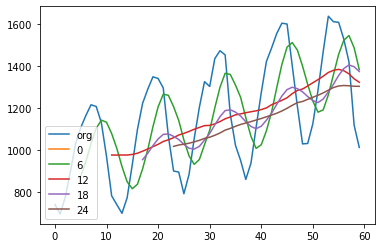
* Partition of the data to be done as train and test. The data to be partition in such a way that the test data should be equal to greater than equal to seasonality.
* To prepare different types of regression models such as linear, exponential, quadratic, additive seasonality, additive seasonality with the quadratic trend, multiplicative seasonality, multiplicative seasonality with linear trend by predicting the model on test data and calculate the root mean squared error values i.e. RMSE value for each model.
* Lesser the RMSE value better is the model. In the given data **Additive Seasonality Quadratic Trend gives the lesser RMSE value**, so we will use this model for forecasting on the new data.
* In the given data the seasonality is additive
* Import the new data for forecasting the passengers for next 5 years.
* Create a model by using the Additive seasonality with quadratic trend regression model on the previous full data and evaluate the data on the new data. It gives the predicted values for the next 5 years.
* The Total forecasting value is the predicted value + error between predicted value and the actual value.
* So to calculate the error we use Autoregression model.
* Calculate the full residual from the full data and plot the ACF and PACF plot for the 12 lags values.
* ACF is an (complete) auto-correlation function gives values of auto-correlation of any time series with its lagged values.
* PACF is a partial auto-correlation function. It finds correlations of present with lags of the residuals of the time series.

From both the plots we can see that 1st lag value has the highest autocorrelation function.

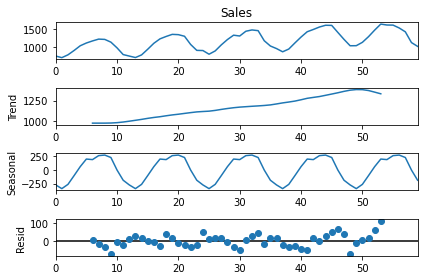
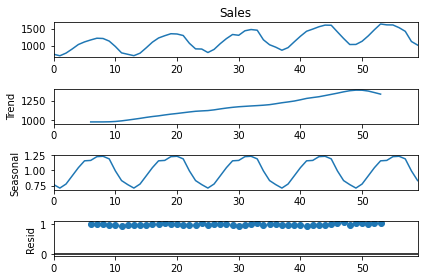
* Build the autoregression model by taking the lag = 1 and evaluate the model on the new data to find out the residuals of the new predicted data.
* Calculate the final prediction by adding the prediction values and the predicted residuals.
* Finally, we get the next 1 years forecasted values month wise.

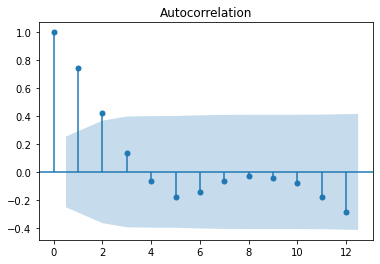
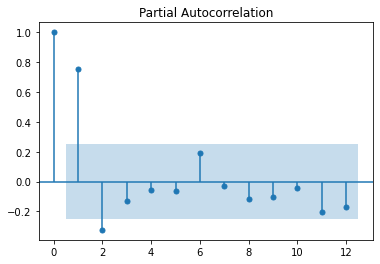
**2) Data Driven Approach:**

* Import the statsmodels libraries and the functions which are useful for the data driven technique.
* Creating a function to calculate the mean absolute percentile error value (MAPE).
* Calculate the MAPE value by calculating the moving average for the 12 months and visualize with the timeplot.



From the graph we can see that at the moving average of the 12 and 24 values, the data has no seasonality.

* Time series decomposition is the process of separating data into its core components. Time series decomposition plot using Moving Average for additive and multiplicative model respectively.
* ACF and PACF plot on original data sets

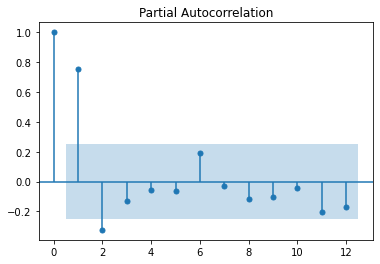
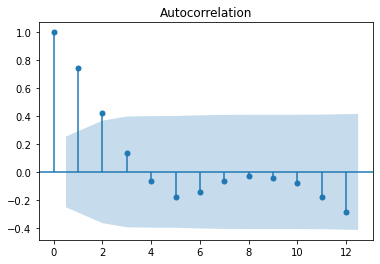
****

1st lag value has the highest correlation function value than rest of the 12 value.

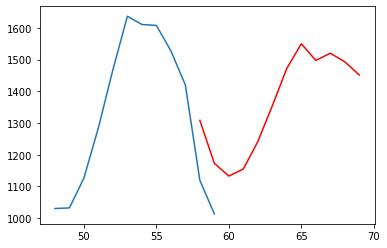
* Build the data driven model by using the different methods such as simple Exponential Method, Holt method, Holts winter exponential smoothing with additive seasonality and additive trend, Holts winter exponential smoothing with multiplicative seasonality and additive trend and evaluate the model on the test datasets and calculate the MAPE values.
* Lesser the MAPE value better is the model.
* For the given Holts winter exponential smoothing with additive seasonality and additive trendhas the least MAPE value so we finalize this model for forecasting.
* Prepare the model by using the Holts winter exponential smoothing with multiplicative seasonality and additive trend method on the complete original data and then evaluate the model on the new data for which we have to forecast the sales’ values and find out the predicted values.

**3) ARIMA Method: -**

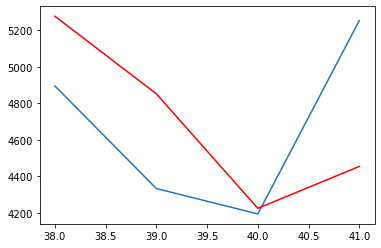
* Import the statsmodels libraries and the ARIMA functions which are useful for the ARIMS technique.
* Plot the ACF and the PACF plot and find out the lag value which has the highest autocorrelation Function.



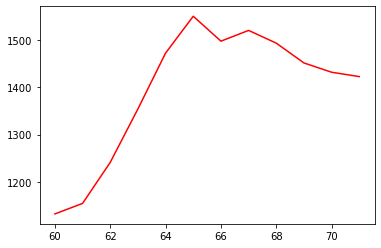
From the above plot we can see that 1st lag value has the highest correlation function value than rest of the 4 value.

* Fit the ARIMA function by taking AR = 1, I=1 and MA =12 and forecast for the next 12 months and calculate the RMSE values.
* Plot the forecasted values with the actual outcomes.

The red line indicates the forecasted values for the next 4 months.

* Auto-ARIMA - Automatically discover the optimal order for an ARIMA model by using the pmdarima library. By using this function, we get the best parameters for ARIMA model is AR = 12, I =1, MA =1
* Again, Fit the ARIMA model by taking the best parameters and forecast for the next 12 months and find the RMSE values. Plot the graph which has the forecasted values with the red line and original value of test datasets with blue line.

* Forecast for the next 2 years i.e.for next 8 quarters by using this final model.



A picture containing shape, arrow

Description automatically generated**Problem Statement: -**

Solar power consumption has been recorded by city councils at regular intervals. The reason behind doing so is to understand how businesses are using solar power so that they can cut down on nonrenewable sources of energy and shift towards renewable energy. Based on the data, build a forecasting model and provide insights on it.

Solarpower.csv

A picture containing table

Description automatically generated

**Business Objectives: -**

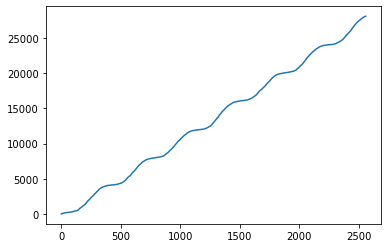
To predict the cumulative power of the for the next years based on the previous data by using time series forecasting

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of features** | **Description** | **Type** | **Relevance** |
| date | date | Quantitative, Nominal | Relevant |
| Cum\_power | Cumulative power consumption | Quantitative, Ratio | Relevant |

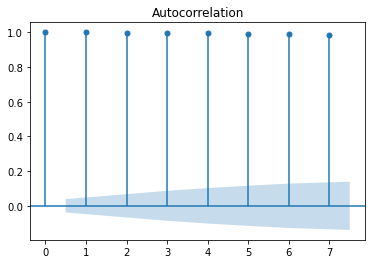
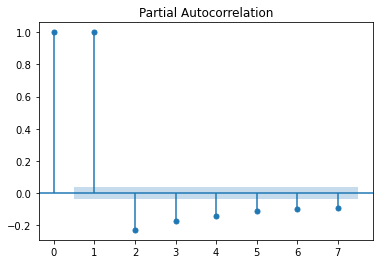
**Steps To build the Forecasting model.**

**1) Model Based Approach/ Regression Model :**

* Import the libraries such as pandas, numpy, matplotlib, statsmodels . From statsmodel library import formula package and import api function.
* Import the data and do the data manipulation.
* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.
* Add t column which has the value in the serial no. to the data. Add another column which has the column square value of t. Add another one column which takes the log value of t.
* Visualize the timeplot between time on x axis and sales on the Y axis.



From the time plot we can see that there is level present, trend is upward and the seasonality may be additive or multiplicative which we can find from the regression model.

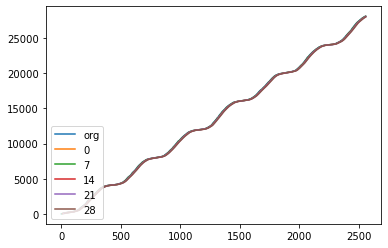
* Partition of the data to be done as train and test. The data to be partition in such a way that the test data should be equal to greater than equal to seasonality.
* To prepare different types of regression models such as linear, exponential, quadratic, additive seasonality, additive seasonality with the quadratic trend, multiplicative seasonality, multiplicative seasonality with linear trend by predicting the model on test data and calculate the root mean squared error values i.e. RMSE value for each model.
* Lesser the RMSE value better is the model. In the given data **linear model gives the lesser RMSE value**, so we will use this model for forecasting on the new data.
* In the given data the seasonality is additive
* Import the new data for forecasting the cumulative power for next 1 years.
* Create a model by using the Additive seasonality with quadratic trend regression model on the previous full data and evaluate the data on the new data. It gives the predicted values for the next 1 years.
* The Total forecasting value is the predicted value + error between predicted value and the actual value.
* So to calculate the error we use Autoregression model.
* Calculate the full residual from the full data and plot the ACF and PACF plot for the 7 lags values.
* ACF is an (complete) auto-correlation function gives values of auto-correlation of any time series with its lagged values.
*  PACF is a partial auto-correlation function. It finds correlations of present with lags of the residuals of the time series.

From both the plots we can see that 1st lag value has the highest autocorrelation function.

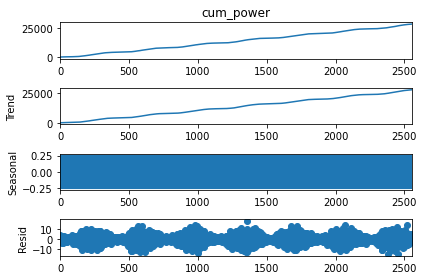
* Build the autoregression model by taking the lag = 1 and evaluate the model on the new data to find out the residuals of the new predicted data.
* Calculate the final prediction by adding the prediction values and the predicted residuals.
* Finally, we get the next 1 years forecasted values day wise.

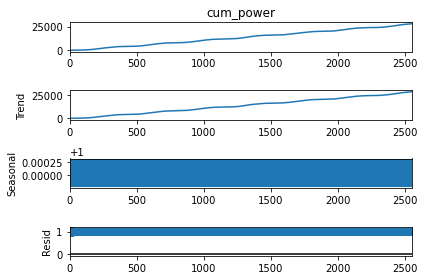
**2) Data Driven Approach:**

* Import the statsmodels libraries and the functions which are useful for the data driven technique.
* Creating a function to calculate the mean absolute percentile error value (MAPE).
* Calculate the MAPE value by calculating the moving average for the 7 days and visualize with the timeplot.

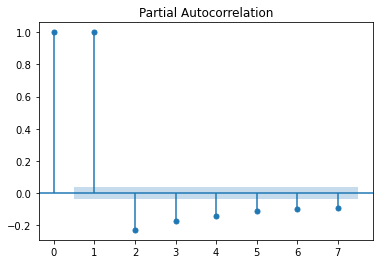
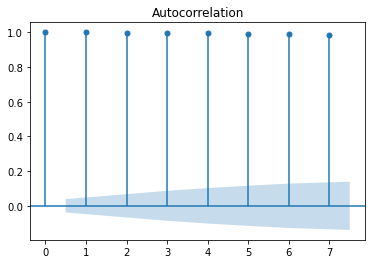


From the graph we can see that at the moving average of the 7 and 14 values, the data has no seasonality.

* Time series decomposition is the process of separating data into its core components. Time series decomposition plot using Moving Average for additive and multiplicative model respectively.
* 



* ACF and PACF plot on original data sets

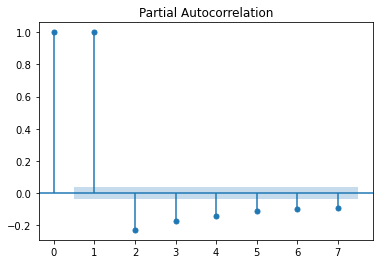
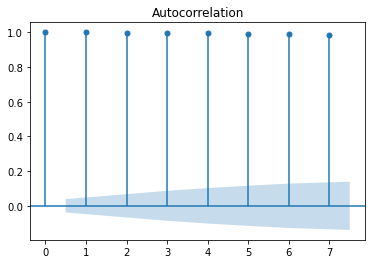
****

1st lag value has the highest correlation function value than rest of the 7 value.

* Build the data driven model by using the different methods such as simple Exponential Method, Holt method, Holts winter exponential smoothing with additive seasonality and additive trend, Holts winter exponential smoothing with multiplicative seasonality and additive trend and evaluate the model on the test datasets and calculate the MAPE values.
* Lesser the MAPE value better is the model.
* For the given Holts winter exponential smoothing with multiplicative seasonality and additive trend least MAPE value so we finalize this model for forecasting.
* Prepare the model by using the Holts winter exponential smoothing with multiplicative seasonality and additive trend method on the complete original data and then evaluate the model on the new data for which we have to forecast the sales’ values and find out the predicted values.

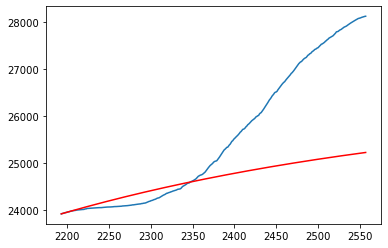
**3) ARIMA Method: -**

* Import the statsmodels libraries and the ARIMA functions which are useful for the ARIMS technique.
* Plot the ACF and the PACF plot and find out the lag value which has the highest autocorrelation Function.

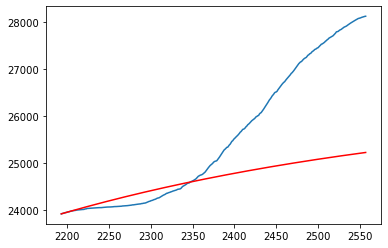


From the above plot we can see that 1st lag value has the highest correlation function value than rest of the 4 value.

* Fit the ARIMA function by taking AR = 1, I=1 and MA =7 and forecast for the next 365 days and calculate the RMSE values.
* Plot the forecasted values with the actual outcomes.



The red line indicates the forecasted values for the next 4 months.

* Auto-ARIMA - Automatically discover the optimal order for an ARIMA model by using the pmdarima library. By using this function, we get the best parameters for ARIMA model is AR = 2, I =1, MA =1
* Again, Fit the ARIMA model by taking the best parameters and forecast for the next 365 days and find the RMSE values. Plot the graph which has the forecasted values with the red line and original value of test datasets with blue line.

* Forecast for the next 2 years i.e.for next 8 quarters by using this final model.

