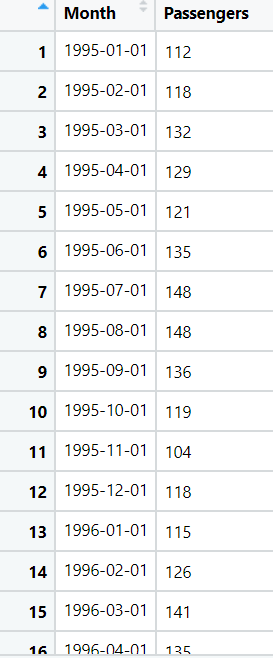
# Topic: Forecasting – Time Series

Forecast the Coca-Cola prices and Airlines Passengers data set. Prepare a document for each model explaining how many dummy variables you have created and RMSE value for each model. Finally which model you will use for Forecasting.

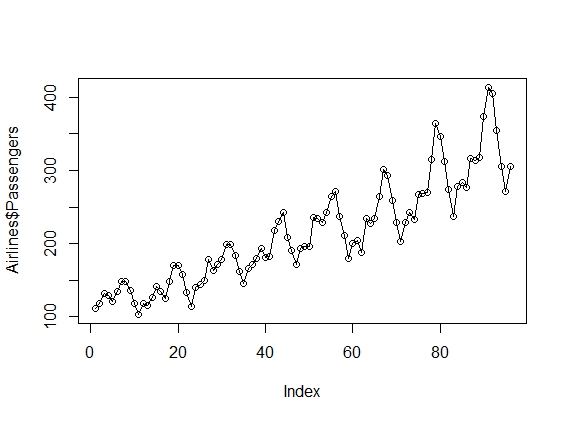
## 1.) Airlines.xlsx



Firstly we need to convert the dataset to csv to remove the time stamp error in python and proceed with data preprocessing steps.

**Data Preprocessing:**

1. The dataset consists of 96 observations with 2 variables of Month-Year and passengers list.
2. Below is the graph that represents passengers list for the given period of time



1. Checking the NA values, as there are no NA values no further imputation is required.
2. As the data is of 12 months, so the frequency is taken as 12 and created 12dummy variables.
3. Now the months names are assigned to the columns names of dummy variables and combined with the airline data.
4. A time column “t” is assigned to the dataset and taking the log of passengers and square of t, so that the whole data is normalized.

**Splitting the data:**

1. Now the data is divided into training and test data with [1:70] and [71:96]
2. Calculating the RMSE value using different models, as we cannot directly tell what is the exact trend followed by the data
3. Only residual values are calculated for all the models to calculate the RMSE values

**Calculating RMSE Using Different Models:**

**Linear Model:**

* RMSE is 48.30 and Adjusted R2 Value is 0.7699

**Exponential Model:**

* RMSE is 43.47 and Adjusted R2 is 0.78
* As predicted values are logged values, we do exponential of expo\_pred$fit to get the actual values.

**Quadratic Model:**

* RMSE is 43.898 and Adjusted R2 Value is 0.769

**Additive Seasonality Model:**

* RMSE is 124.97and Adjusted R2 is 0.083 Hence, it may not be additive seasonality model.

**Additive Seasonality with Linear Model:**

* RMSE is 34.502 and Adjusted R2 is 0.94

**Additive Seasonality with Quadratic Model:**

* RMSE is 30.393 and Adjusted R2 is 0.95

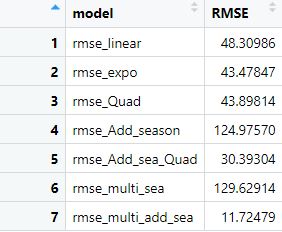
**Multiplicative Seasonality Model:**

* In multiplicative we multiply but we can't multiply directly hence we apply log
* RMSE is 129.6291 & Adjusted R2 is 0.07

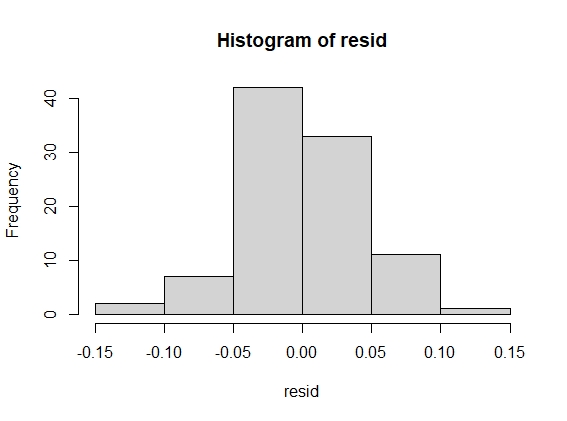
**Multiplicative Seasonality Linear trend:**

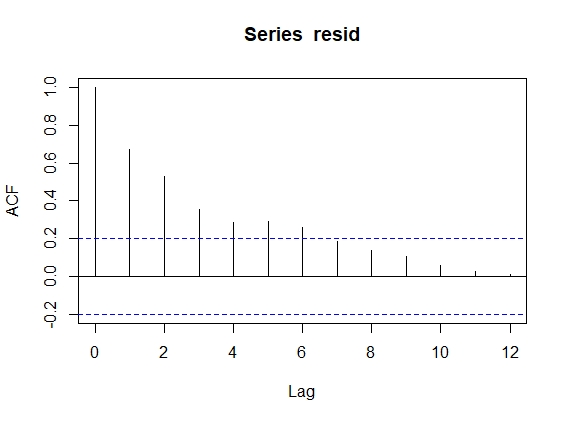
* RMSE is 11.72 and Adjusted R2 is 0.96
* This is the highest R2 & lowest RMSE

**Below are the over all RMSE values for all models**



* Now building the model with the whole dataset of airlines
* Histogram of residual values

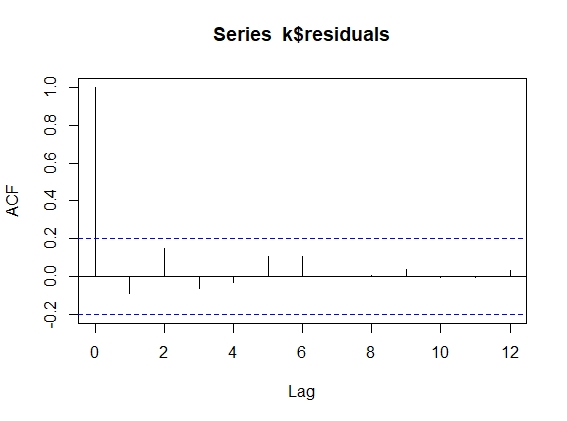




* From the above residual graph, we can say that the lag 1 to 4 are significant, so Arima can be built

**Building the Arima Model:**

* Auto regression is only used to forecast errors
* Performing auto regression with 2nd lag, p=2,d=0,q=0



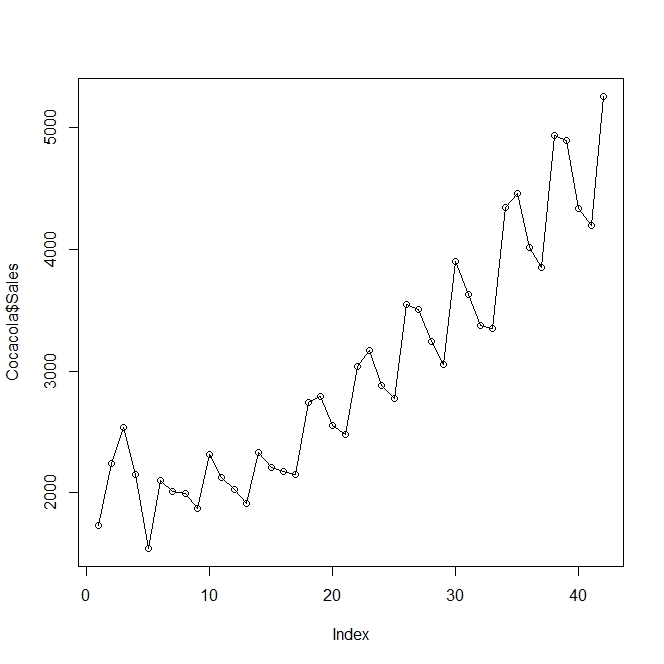
* From the above graph, we can say that significance problem is removed & all are below threshold ACF values.
* Using the function pred\_res$pred then recalling function acf(k$residuals,lag.max = 12) we get the output of the problem statement.

**2.) CocaCola\_Sales\_RawData.xlsx**



**Data Preprocessing:**

1. The dataset consists of 42 observations with 2 variables of Month-Year and Sales for the 4 quarters in a specific period.
2. Below is the graph that represents sales for the given period of time.



1. As the data is of 4 quarters, so the frequency is taken as 4 and created 4 dummy variables
2. Now the quarters names are assigned to the columns names of dummy variables and combined with the data.
3. A time column “t” is assigned to the dataset and taking the log of sales and square of t, so that the whole data is normalized.

**Splitting the data:**

1. Now the data is divided into training and test data with [1:30] and [31:42]
2. Calculating the RMSE value using different models, as we cannot directly tell what is the exact trend followed by the data
3. Only residual values are calculated for all the models to calculate the RMSE values

**Calculating RMSE Using Different Models**

**Linear Model:**

* RMSE is 714.014 and Adjusted R2 Value is 0.69

**Exponential Model:**

* RMSE is 552.28 and Adjusted R2 is 0.69
* RMSE has reduced of the exponential model than linear.
* As predicted values are logged values, we do exponential of expo\_pred$fit to get actual values

**Quadratic Model:**

* RMSE is 646.27 and Adjusted R2 Value is 0.79

**Additive Seasonality Model:**

* RMSE is 1778.00 and Adjusted R2 is 0.05. Hence, it may not be additive seasonality model.

**Additive Seasonality with Linear Model:**

* RMSE is 637.94 and Adjusted R2 is 0.82

**Additive Seasonality with Quadratic Model:**

* RMSE is 586.05 and Adjusted R2 is 0.94

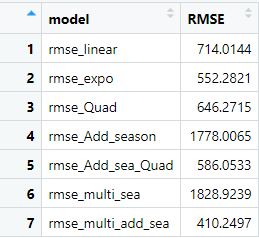
**Multiplicative Seasonality Model:**

* In multiplicative we multiply but we can't multiply directly hence we apply log
* RMSE is 1828.92 and Adjusted R2 is 0.07

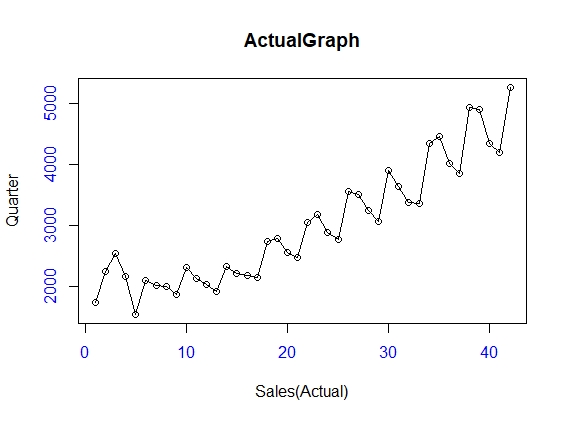
**Multiplicative Seasonality Linear trend:**

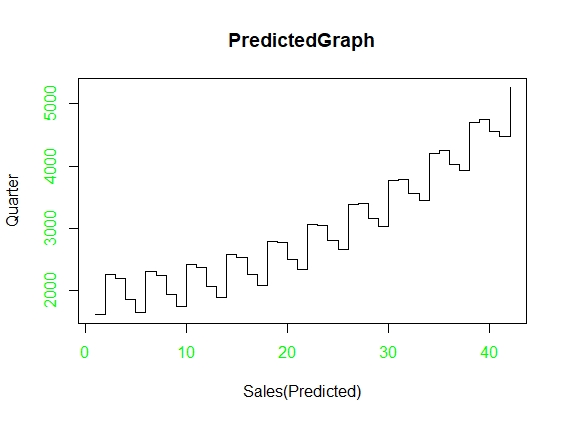
* RMSE is 410.24 and Adjusted R2 is 0.83
* This is the highest R2 & lowest RMSE

**Below are the overall RMSE values for all models**

****

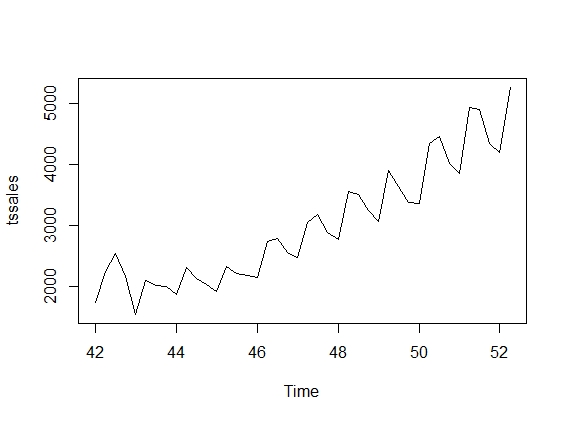
* Here we find that Multiplicative additional Seasonality with Linear trend which has least RMSE value of 410.24
* Now building the model with the whole dataset





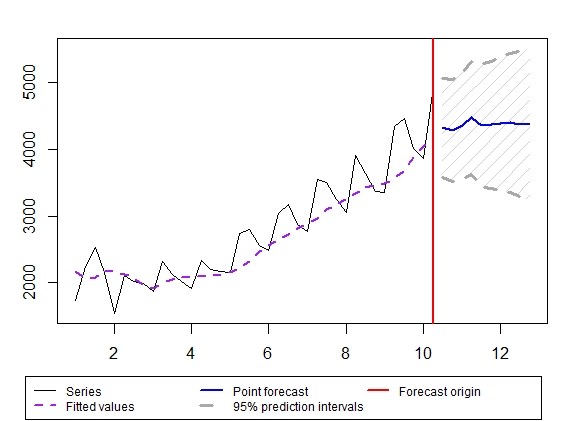
**MAPE and MAE Graphical Representation:**

**Time Series Graph:**

****

* Visualization shows that it has level, trend, seasonality => Additive seasonality

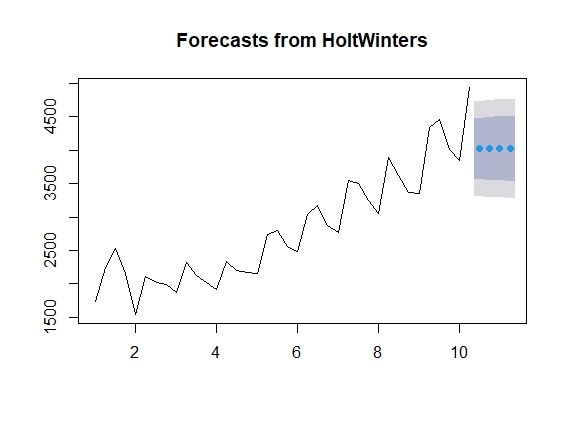
**Moving Average Graph:**

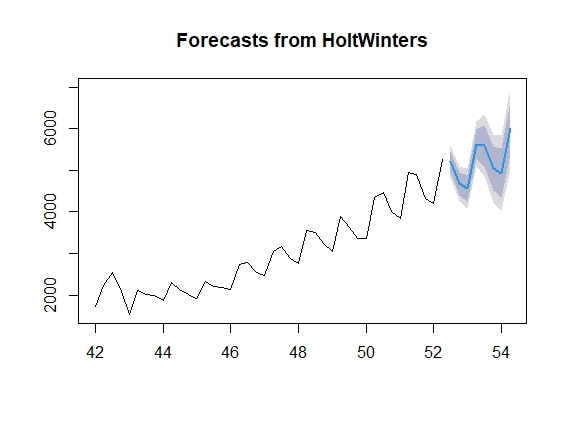
****

* The MAPE value for moving average is 8.908

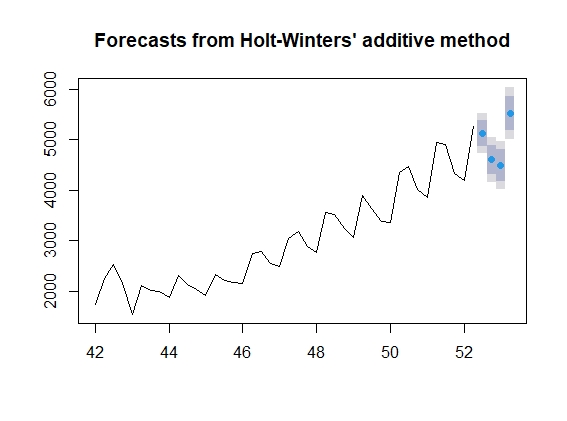
**Using Holtwinters Function:**

* Taking alpha value as 0.2 as a default and assuming time series data has only level parameter





**Additive Graph:**

****

* Based on the MAPE value who choose holts winter exponential technique which assumes the time series
* Data level, trend, seasonality characters with default values of alpha, beta and gamma

# Hints:

1. Business Problem
   1. Objective
   2. Constraints (if any)
2. Data Pre-processing

2.1 Feature Engineering, EDA etc.

1. Model Building
   1. Partition the dataset
   2. Model(s) – Work with all the models (linear, exponential, quadratic etc.)
   3. Model(s) Improvement steps
   4. Model Evaluation
   5. Python and R codes
2. Result Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.

## Note:

1. For each assignment the solution should be submitted in the format
2. Research and Perform all possible steps for improving the model(s) accuracy & reduce the RMSE (also evaluate errors like MAPE, MAE etc.)
3. All the codes (executable programs) are running without errors
4. Documentation of the module should be submitted along with R & Python codes, elaborating on every step mentioned here