Assignment: FA Assignment

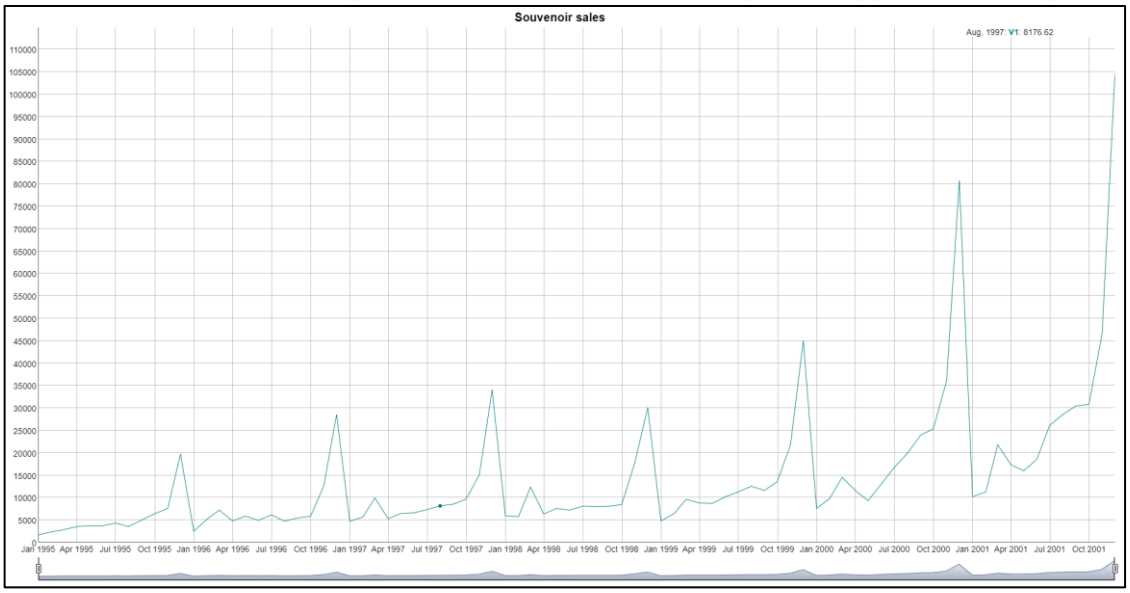
Batch – AMPBA-2021S

ID – 12010066

Name – Debjit Ray

1. Consider the data set SouvenirSales.xls (1995 Jan -2001 Dec) that gives the monthly sales of souvenir at a shop in New York. Back in 2001, an analyst was appointed to forecast sales for the next 12 months (Year 2002). The analyst portioned the data by keeping the last 12 months of data (year 2001) as validation set, and the remaining data as training set. Answer the following questions. Use R to solve the questions.

   a. Plot the time series of the original data. Which time series components appear from the plot.



   From the graph, we can notice the following:

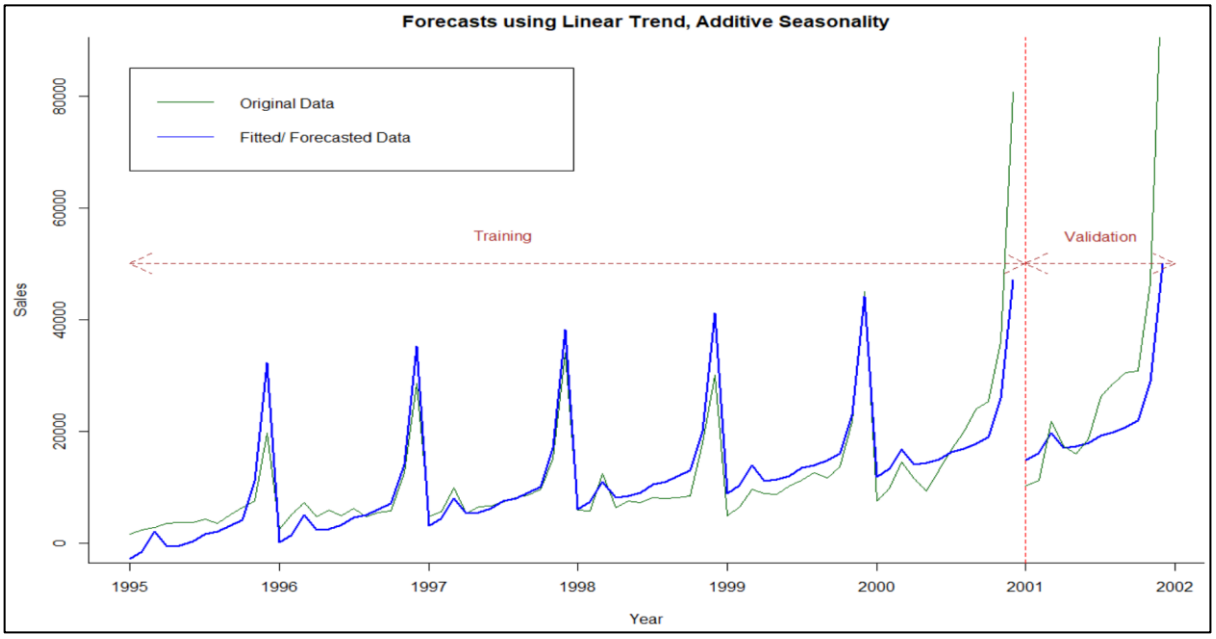
   - There is an **upper trend** in Sales across the years.
   - Every year there is a **peak in sales in the month of March and December**.
   - The **seasonality** for the month of December **seems to be Multiplicative** in nature.

   b. Fit a linear trend model with additive seasonality (Model A) and exponential trend model with multiplicative seasonality (Model B). Consider January as the reference group for each model. Produce the regression coefficients and the validation set errors. Remember to fit only the training period.

   Model A – Linear Trend with Additive Seasonality model

```r
# Define the model with linear trend and additive seasonality
modelA <- tslm(trainSet ~ trend +  season)
# Check the model summary for the regression coefficients
summary(modelA)
# Predict for the validation time period using this model
modelA_Forecast <- forecast(modelA, h=length(testSet), level = 0)
```

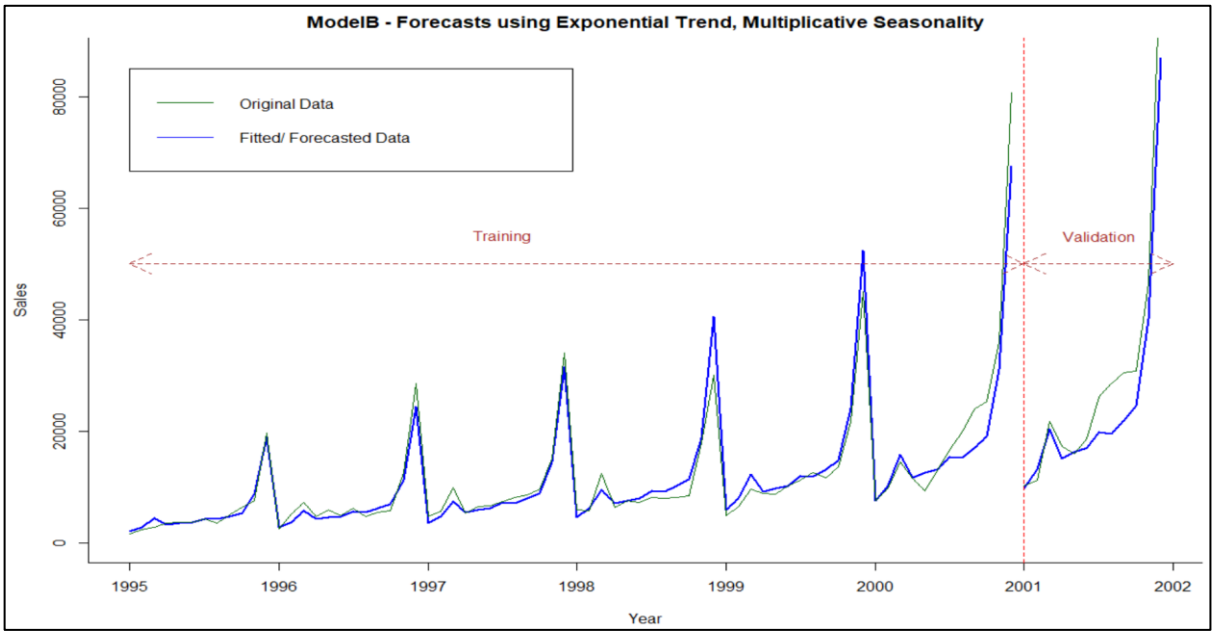
```
Call:
tslm(formula = trainSet ~ trend + season)

Residuals:
   Min     1Q Median     3Q    Max
-12592  -2359   -411   1940  33651

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3065.55    2640.26  -1.161  0.25029
trend         245.36      34.08   7.199 1.24e-09 ***
season2      1119.38    3422.06   0.327  0.74474
season3      4408.84    3422.56   1.288  0.20272
season4      1462.57    3423.41   0.427  0.67077
season5      1446.19    3424.60   0.422  0.67434
season6      1867.98    3426.13   0.545  0.58766
season7      2988.56    3427.99   0.872  0.38684
season8      3227.58    3430.19   0.941  0.35058
season9      3955.56    3432.73   1.152  0.25384
season10     4821.66    3435.61   1.403  0.16573
season11    11524.64    3438.82   3.351  0.00141 **
season12    32469.55    3442.36   9.432 2.19e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5927 on 59 degrees of freedom
Multiple R-squared:  0.7903,    Adjusted R-squared:  0.7476
F-statistic: 18.53 on 12 and 59 DF,  p-value: 9.435e-16
```



Model B – Exponential Trend with Multiplicative Seasonality model

```r
# Define the model with linear trend and additive seasonality
modelB <- tslm(trainSet ~ trend + season, lambda = 0)
# Check the model summary for the regression coefficients
summary(modelB)
# Predict for the validation time period using this model
modelB_Forecast <- forecast(modelB, h=length(testSet), level = 0)
```

```
Call:
tslm(formula = trainSet ~ trend + season, lambda = 0)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4529 -0.1163  0.0001  0.1005  0.3438

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.646363   0.084120  90.898  < 2e-16 ***
trend       0.021120   0.001086  19.449  < 2e-16 ***
season2     0.282015   0.109028   2.587 0.012178 *
season3     0.694998   0.109044   6.374 3.08e-08 ***
season4     0.373873   0.109071   3.428 0.001115 **
season5     0.421710   0.109109   3.865 0.000279 ***
season6     0.447046   0.109158   4.095 0.000130 ***
season7     0.583380   0.109217   5.341 1.55e-06 ***
season8     0.546897   0.109287   5.004 5.37e-06 ***
season9     0.635565   0.109368   5.811 2.65e-07 ***
season10    0.729490   0.109460   6.664 9.98e-09 ***
season11    1.200954   0.109562  10.961 7.38e-16 ***
season12    1.952202   0.109675  17.800  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1888 on 59 degrees of freedom
Multiple R-squared:  0.9424,    Adjusted R-squared:  0.9306
F-statistic:  80.4 on 12 and 59 DF,  p-value: < 2.2e-16
```



As evident from the graphs, the Model B (Exponential Trend with Multiplicative Seasonality) is fitting the original data better than Model A (Linear Trend with Additive Seasonality).

c. Which model is the best model considering RMSE as the metric? Could you have understood this from the line chart? Explain. Produce the plot showing the forecasts from both models along with actual data. In a separate plot, present the residuals from both models (consider only the validation set residuals).

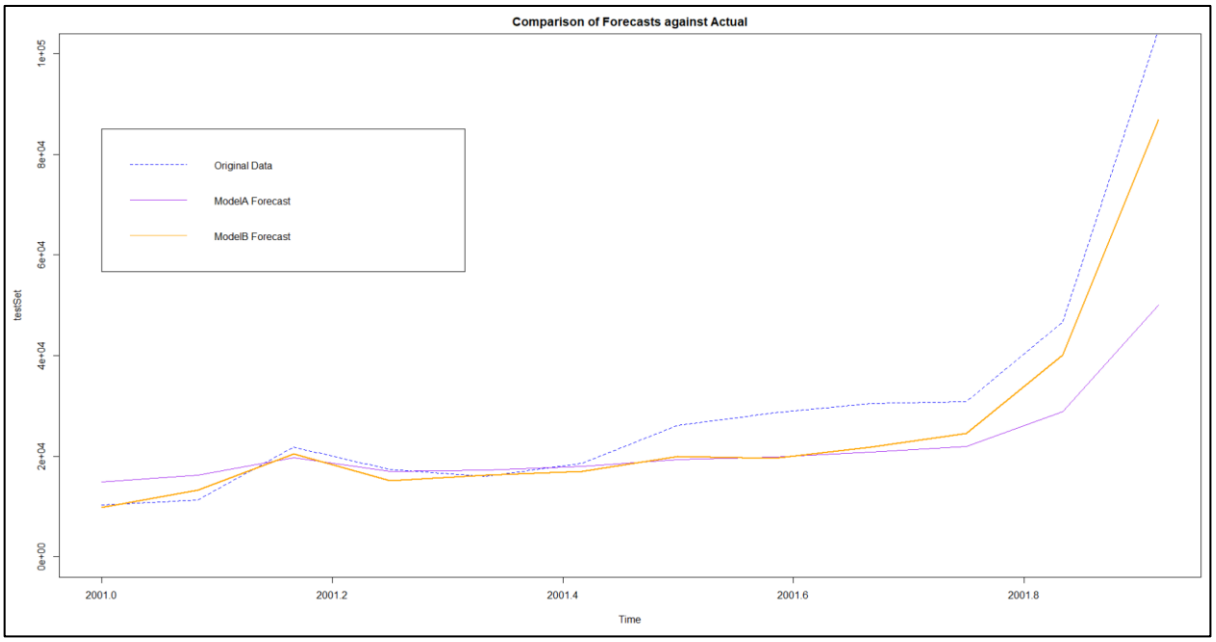
```
198 ```{r Compare the models}
199 print ("Model A's Metrics")
200 accuracy(modelA_Forecast, testSet)
201 print ("Model B's Metrics")
202 accuracy(modelB_Forecast, testSet)
203 ```
```

```
[1] "Model A's Metrics"
                      ME      RMSE       MAE       MPE     MAPE     MASE      ACF1 Theil's U
Training set -5.684342e-14  5365.199  3205.089  6.967778 36.75088 0.855877 0.4048039        NA
Test set      8.251513e+03 17451.547 10055.276 10.533974 26.66568 2.685130 0.3206228 0.9075924
[1] "Model B's Metrics"
                      ME          RMSE          MAE          MPE      MAPE         MASE      ACF1 Theil's U
Training set 4.935341e-17 1.709374e-01 1.382015e-01 -0.03661472  1.534989     0.444076 0.4593573        NA
Test set     3.021146e+04 3.885273e+04 3.021146e+04 99.95338392 99.953384 97076.962013 0.3182420  2.631072
```

Based on RMSE, the Exponential Trend with Multiplicative Seasonality model is performing much better than the Linear Trend with Additive Seasonality model.

Even though our hunch regarding the Seasonality being multiplicative seems to be true, our hunch related to Trend seems to be incorrect. From the original line chart of the Sales data this was difficult to guess, as the trend of sales data for months other than December seems to be increasing gradually.

Plotting the residuals of the training period for both the models the above conclusion becomes even more evident.



As evident from the above graph, Model B (Orange Line) performs much better and is closer to the actual data (Blue Dotted Line).

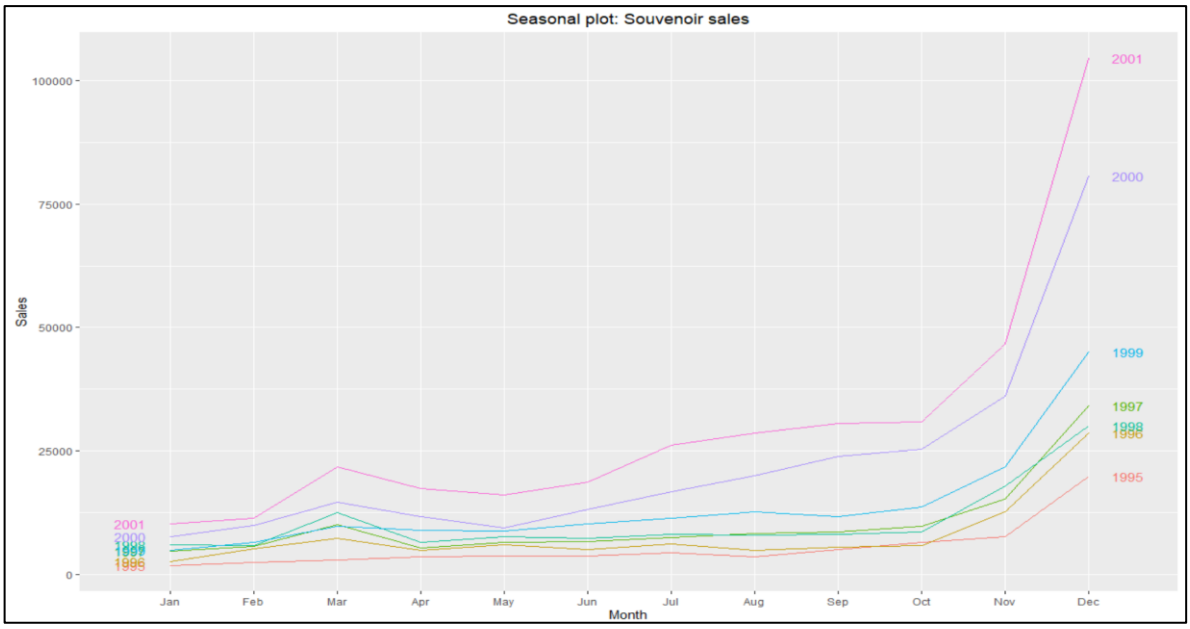
d. Examine the additive model. Which month has the highest average sales during the year. What does the estimated trend coefficient in the model A mean?

The additive model based on the training data is as below:

$$Y_t = -3065.55 + 245.36 * t + 1119.38 * season2 + 4408.84 * season3 + 1462.57 * season4 + 1446.19 * season5 + 1867.98 * season6 + 2988.56 * season7 + 3227.58 * season8 + 3955.56 * season9 + 4821.66 * season10 + 11524.64 * season11 + 32469.55 * season12$$

To determine the month with the highest average sales during a year, we can directly review the coefficients in the regression equation for each month. In our model, we see that January has been taken as reference and for each of the other months; the coefficients have been calculated. These coefficients signify the relative increase in sales in the corresponding month compared to January.

In our model, the highest coefficient is for season12 i.e. December. **Hence, we can deduce December being the month with the highest average sales during the year.** Our deduction can be cross verified against the seasonal graph plotted below, which shows that for every year December has the highest sales compared to the remaining months of the year.



The estimated trend coefficient in the Model A signifies the increase in sales over unit increase in time (here, 1 month) on average. In our model, the coefficient is 245.36 and this signifies; that considering all other factors remaining constant, there is an average increase of 245.36 units for every month over its' preceding month.

e. Examine the multiplicative model. What does the coefficient of October mean? What does the estimated trend coefficient in the model B mean?

The multiplicative model based will take the below form:

$$Y_t = \alpha_1 e^{\beta t} \varepsilon * \alpha_2 e^{\beta_2 season_2} * \alpha_3 e^{\beta_3 season_3} \ldots * \alpha_{12} e^{\beta_{12} season_{12}}$$

Based on the training data it takes the shape of the below mathematical equation:

$$log(Y_t) = 7.646363 + 0.021120 * \text{t} + 0.282015 * \text{season2} + 0.694998 * \text{season3} + 0.373873 * \text{season4} + 0.421710 * \text{season5} + 0.447046 * \text{season6} + 0.583380 * \text{season7} + 0.546897 * \text{season8} + 0.635565 * \text{season9} + 0.729490 * \text{season10} + 1.200954 * \text{season11} + 1.952202 * \text{season12}$$

The coefficient of October i.e. season10 is **0.7297490**. This signifies that compared to the reference month i.e. season1 (January), **the sales in October is greater by ~7.29%.**

Here, the **coefficient of Trend is 0.21120**. This signifies that considering all other factors remaining constant; **on average, the sales is ~2.11% more for every month over the preceding month**.

f. Use the best model type from part (c) to forecast the sales in January 2002. Think carefully which data to use for model fitting in this case.

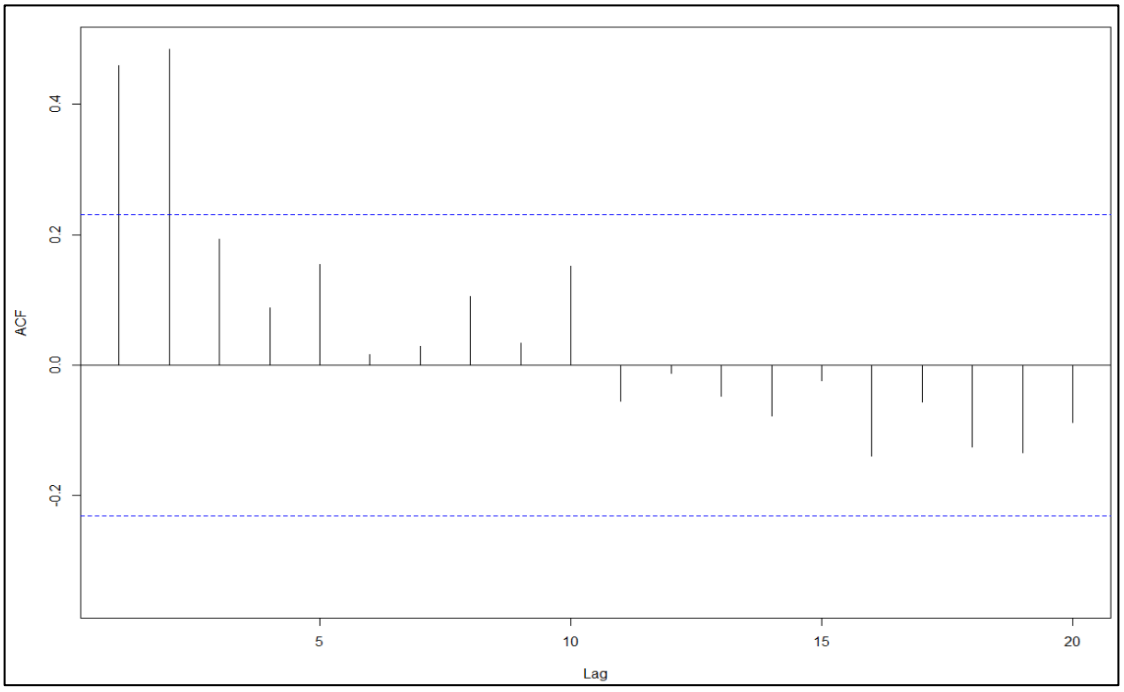
Based on RMSE, as concluded earlier the Exponential Trend with Multiplicative Seasonality is the better fit. Now, before we can use the model, we need to realign our model by training it on the entire dataset (Jan 1995 to Dec 2001).

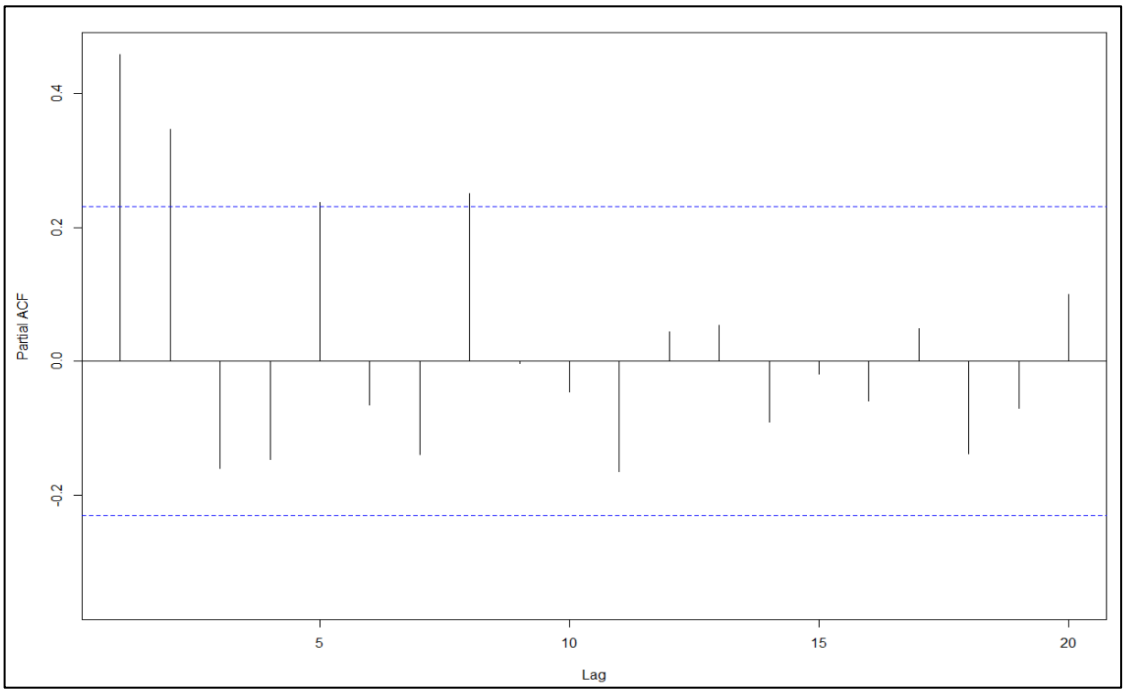
```
264 ```{r ModelB Retrain}
265 # Define the model with linear trend and additive seasonality
266 modelB_Retrained <- tslm(salesData ~ trend + season, lambda = 0)
267 # Check the model summary for the regression coefficients
268 summary(modelB_Retrained)
269 # Predict for the validation time period using this model
270 modelB_Retrained_Forecast <- forecast(modelB_Retrained, h=1, level = 0)
271 modelB_Retrained_Forecast$mean
272 ```
```

```
Call:
tslm(formula = salesData ~ trend + season, lambda = 0)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41644 -0.12619  0.00608  0.11389  0.38567

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.6058604  0.0768740  98.939  < 2e-16 ***
trend       0.0223930  0.0008448  26.508  < 2e-16 ***
season2     0.2510437  0.0993278   2.527 0.013718 *
season3     0.6952066  0.0993386   6.998 1.18e-09 ***
season4     0.3829341  0.0993565   3.854 0.000252 ***
season5     0.4079944  0.0993817   4.105 0.000106 ***
season6     0.4469625  0.0994140   4.496 2.63e-05 ***
season7     0.6082156  0.0994534   6.116 4.69e-08 ***
season8     0.5853524  0.0995001   5.883 1.21e-07 ***
season9     0.6663446  0.0995538   6.693 4.27e-09 ***
season10    0.7440336  0.0996148   7.469 1.61e-10 ***
season11    1.2030164  0.0996828  12.068  < 2e-16 ***
season12    1.9581366  0.0997579  19.629  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1858 on 71 degrees of freedom
Multiple R-squared:  0.9527,    Adjusted R-squared:  0.9447
F-statistic: 119.1 on 12 and 71 DF,  p-value: < 2.2e-16

          Jan
2002 13484.06
```

Thus, based on Model B, the forecast for January 2002 is 13,484.06.

g. Plot the ACF and PACF plot until lag 20 of the residuals obtained from training set of the best model chosen. Comment on these plots and think what AR(p) model could be a good choice?





From the ACF plot, we see the lag at 1 and 2 are significant. **This signifies MA(2) model.**

Similarly, PACF plot is also showing the partial correlation at lags 1 and 2 are significant. **This signifies AR(2) model.**

Assuming the Regression has already removed the Seasonality and Trend leaving the residuals being a stationary time series[1], the model, we can predict an ARMA(2,2) or

[1] From the ACF and PACF plot, we see the spikes to decay rapidly, hence, this is a fair assumption to make.

ARIMA(2,0,2) as the preferable model to extract the relevant information from the residuals.

h. Fit an AR(p) model as you think appropriate from part (g) to the training set residuals and produce the regression coefficients. Was your intuition at part (g) correct?

```
307 ```{r Fit ARMA model}
308 # Fit a ARIMA model based on our intuition from the above question.
309 residualModel <- Arima(modelB$residuals, order = c(2,0,2)) ## (ARIMA(2,0,2) = ARMA(2,2))
310 residualModel_Forecast <- forecast(residualModel,h=12)
311 summary(residualModel)
312
313 ```
```

```
Series: modelB$residuals
ARIMA(2,0,2) with non-zero mean

Coefficients:
         ar1      ar2     ma1     ma2    mean
      0.0895  -0.0966  0.2934  0.7722  0.0013
s.e.  0.1486   0.1778  0.0873  0.1431  0.0315

sigma^2 estimated as 0.01864:  log likelihood=42.96
AIC=-73.92   AICc=-72.63   BIC=-60.26

Training set error measures:
                      ME      RMSE       MAE      MPE     MAPE      MASE         ACF1
Training set 0.0003670075 0.1317171 0.1060178 4027.564 4574.364 0.5535896 0.003910029
```

On fitting the ARMA(2,2) or ARIMA(2,0,2) model as estimated in part (g) we get the above coefficients. The training set metrics like RMSE, MASE look quite promising.

i. Now, using the best regression model and AR(p) model, forecast the sales in January 2002. Think carefully which data to use for model fitting in this case.

We have already retrained Model B on the entire dataset to forecast the sales in January 2002 based on Trend and Seasonality. We can better our forecast by now using the AR model created above.

```
319 ```{r Train AR model on entire data and forecast}
320
321 residualModel_Retrained <- Arima(modelB_Retrained$residuals, order = c(2,0,2)) ## (ARIMA(2,0,2) = ARMA(2,2))
322 residualModel_Retrained_Forecast <- forecast(residualModel_Retrained,h=1)
323 summary(residualModel_Retrained)
324 residualModel_Retrained_Forecast$mean
325 ```
```

```
Series: modelB_Retrained$residuals
ARIMA(2,0,2) with non-zero mean

Coefficients:
         ar1     ar2     ma1     ma2    mean
      0.2002  0.1503  0.1953  0.3472  0.0006
s.e.  0.2629  0.2284  0.2664  0.2919  0.0350

sigma^2 estimated as 0.01995:  log likelihood=47.47
AIC=-82.95   AICc=-81.86   BIC=-68.36

Training set error measures:
                      ME      RMSE       MAE      MPE    MAPE      MASE         ACF1
Training set 0.0009884091 0.1369625 0.1108288 79.17213 159.616 0.6343777 -0.01409401
             Jan
2002 0.05264667
```

```
326
327 ####  Thus, the forecast for January 2002 will be ModelB Forecast + Residual Model Forecast i.e. 13,484.06 + 0.05264667 ~
    13484.11.
```

As seen above, on retraining the AR model and using it for forecasting the Jan 2002 residuals it gives us a forecast of **0.05264667.**

This when added to the Regression model (Model B) forecast provides us the final forecast for Jan 2002 sales = 13,484.06 + 0.05264667 ~ 13,484.11

Corresponding R files:


Debjit_Ray_12010066_
FA_Assignment.Rmd


Debjit_Ray_12010066_
FA_Assignment.html

2. Short answer type questions:

a. Explain the key difference between cross sectional and time series data.

Cross sectional data consists of observations that are collected for a single point of time. By single point of time, it means a period during which all the observations collected are relevant and comparable within that period. For e.g. the data collected for an opinion poll during an election can be collected over days/ 1-2 weeks is a Cross-sectional data.
Time series data consists of observations that are collected over spaced time intervals. For e.g., if during the election, opinions are collected for every fortnight and then used to understand the trend of results varying over fortnights.

So, in the example of opinion poll above, the data collected for an instance of opinion poll is a Cross sectional data. However, if we consider all the data collected over multiple such opinion polls, it will be considered as Time series data.

Key differences between Cross sectional and Time series data are as below:

Cross Sectional Data	**Time Series Data**
Collected over a specific period of time.	Collected over spaced time intervals
For testing models, the input data set is split into Train, Test and Validation datasets.	For testing models, the input data set is split only into Train and Validation datasets.
While splitting the input data set into Train, Test and Validation datasets, the observations are selected randomly for each of the split datasets.	While splitting the data set into Train and Validation datasets, the observations are split by referring to a specific observation/ point of time. All continuous observations prior to this specific point are considered for Train and all after the specific point are considered for Validation dataset.
For making predictions, the model selected on the basis of Test accuracy is used.	For making predictions, the model after selection on the basis of Training accuracy is retrained on the entire set of available data (Training + Validation) to help the model to utilize the latest available data.
Statistical assumption for cross sectional data while modelling is that the data is Independent and Identically Distributed (IID).	Usually, the data has high correlation with data for previous period(s) and hence, the data cannot be assumed to independent.

b. Explain the difference between seasonality and cyclicality.

Seasonality and **Cyclicality** are both components of a Time Series.

Seasonality is the regular repetitive fluctuation of data over short periods of time. For example, sales in a Pizza restaurant will see spikes during Lunch, Evening Snacks and Dinner every day. So, in this case, there will be 3 seasonalities detected if hourly sales data is plotted, each with a regular interval of roughly 24 hours.
Cyclicality is the irregular fluctuation of data over long periods of time. For example, sales in the Pizza restaurant might get affected due to strikes/ thunderstorms/ other unpredictable events.
Key differences between Seasonality and Cyclicality are as below:

Factor	**Seasonality**	**Cyclicality**
Predictability	Highly predictable.	Unpredictable
Time Duration	The fluctuations are equally spaced over short period of time.	The fluctuations are unequally spaced over long duration of time.
Modelling	Can be considered and attributed in the Time series model being created.	Cannot be considered/ attributed in the time series model.

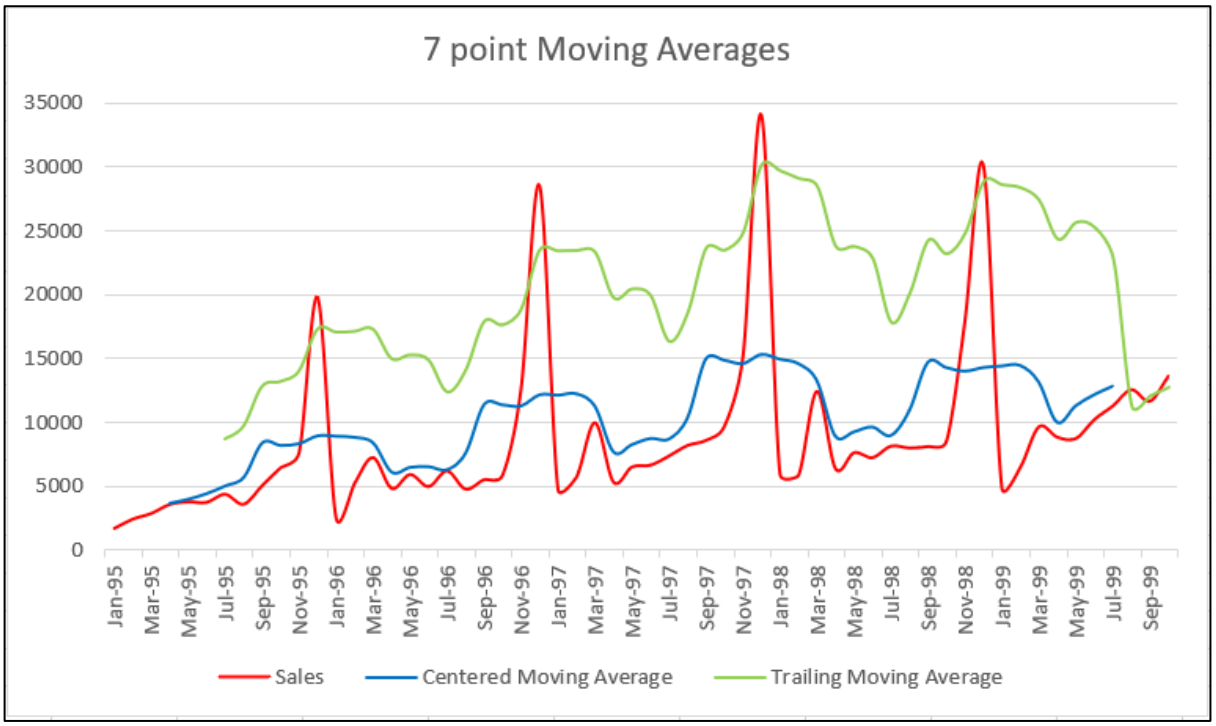
c. Explain why centered moving average is not-considered suitable for forecasting.

Moving Average is used to remove the impact of seasonality and noise from a time series data to help reveal the trend of the time series.

Centered Moving Average is calculated by taking the average of 'n' consecutive observations and mapping it to the $(n/2)^{th}$ time point. For example in the chart below, when we plotted the 7 point centered moving average (Blue line in chart below), the average for the first 7 months (Jan'95 to Jul'95) is plotted against Apr'95. Similarly, the last observation of the moving average line graph is mapped to Jul'99 by considering the observations from Apr'99 to Oct'99.

In contrast, for the trailing moving average (Green line in chart below) the earliest calculated data point is Jul'95 (by taking average from Jan'95 to Jul'95) and the last data point is Oct'99 (by taking average from Apr'99 to Oct'99).

Now for forecasting using the centered moving average, the last data point available is Jul'99 whereas for trailing moving average is Oct'99. In case, we need to forecast the data for Jan'00 using centered moving average, we are trying to forecast a data point which is 6 time periods (months) ahead. Whereas, if we try to forecast using trailing moving average, it is 3 time periods (months) ahead. Hence, the chances of a better forecast would be by using a trailing moving average rather than centered moving average. Thus, centered moving averages are not used to forecast.



d. Explain stationarity and why is it important for some time series forecasting methods?

Strictly/ Strongly stationary time series means that the joint probability distribution of the series does not change at any point when selected across the time axis, t.
For example, let us imagine a time-series Y = f(t); and 3 points at random, t1, t2 and t3 along the time axis. If the joint probabilities for all values of lag, k remain constant we call it a Strictly/ Strongly stationary i.e. $P(Y_{t1,} Y_{(t1-k)}) = P(Y_{t2,} Y_{(t2-k)}) = P(Y_{t3,} Y_{(t3-k)})$.

Weakly stationary time series is one which satisfies the below conditions:

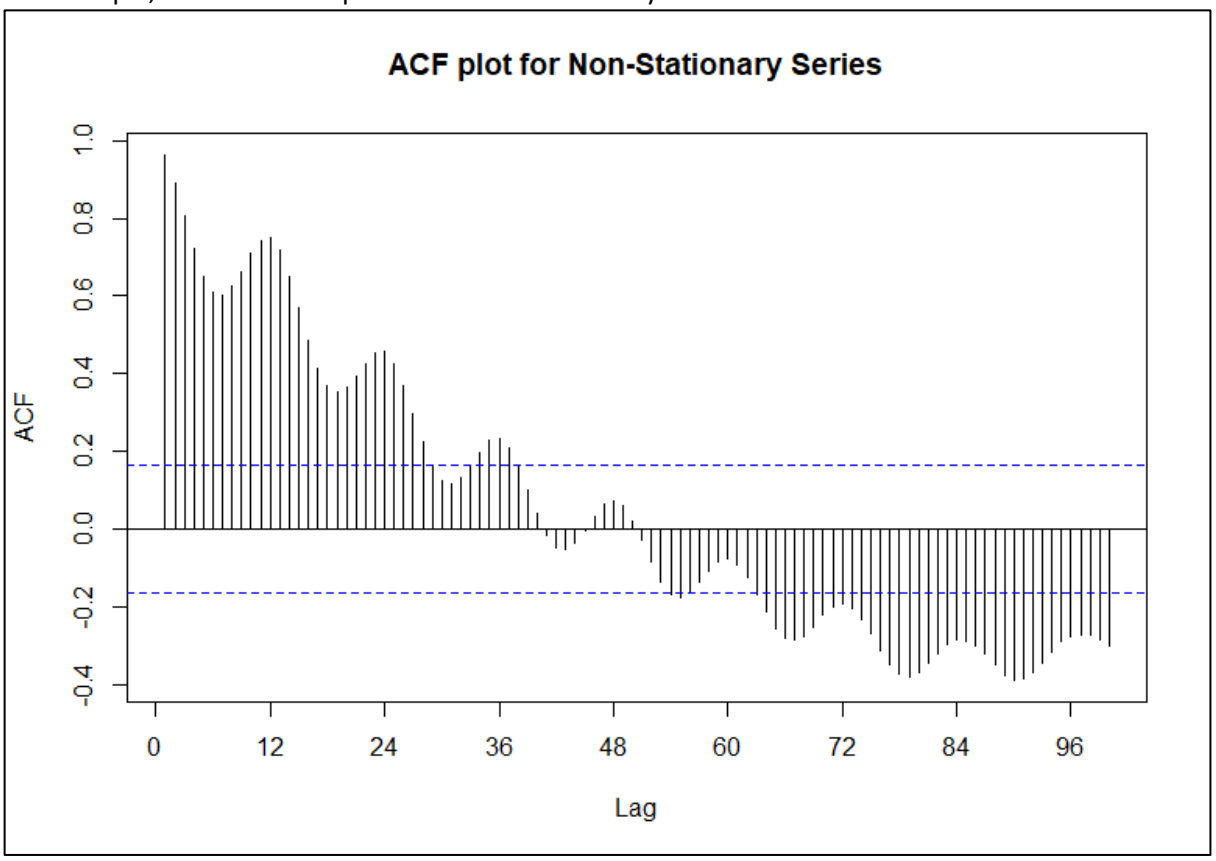
- The mean $E(Y_t)$ is constant throughout all values of 't'. In other words, the mean is not a function of time rather is a constant value.
- The variance $Var(Y_t)$ is constant throughout all values of 't'. In other words, the series is Homoskedastic.
- The covariance and correlation between Y_t and $Y_{(t-k)}$ is same for all values of t provided the lag, k is same. i.e. $Cov(Y_{t1}, Y_{t1-k}) = Cov(Y_{t2}, Y_{t2-k}) = Cov(Y_{tn}, Y_{tn-k})$. In other words, the Covariance is dependant on the lag, k and not on time, t.

Stationarity in time series helps to establish that the statistical properties of the system remains constant over time. This is important to help forecast future values as if the statistical properties change, there will be no mathematical model possible to calculate future values. Hence, for any time series after removing the Trend and Seasonality, we try to make the residuals stationary first by using Differencing. With a stationary series, mathematically, we can help arrive at a model based on auto-correlations to forecast the future values of residuals. With a stationary series, considering the mean, variance and covariance are all constant, it helps us forecast the future values. Hence, stationarity leads to statistical equilibrium, which is the basic assumption/ requirement for a lot of models like AR, ARMA, ARIMA, etc.

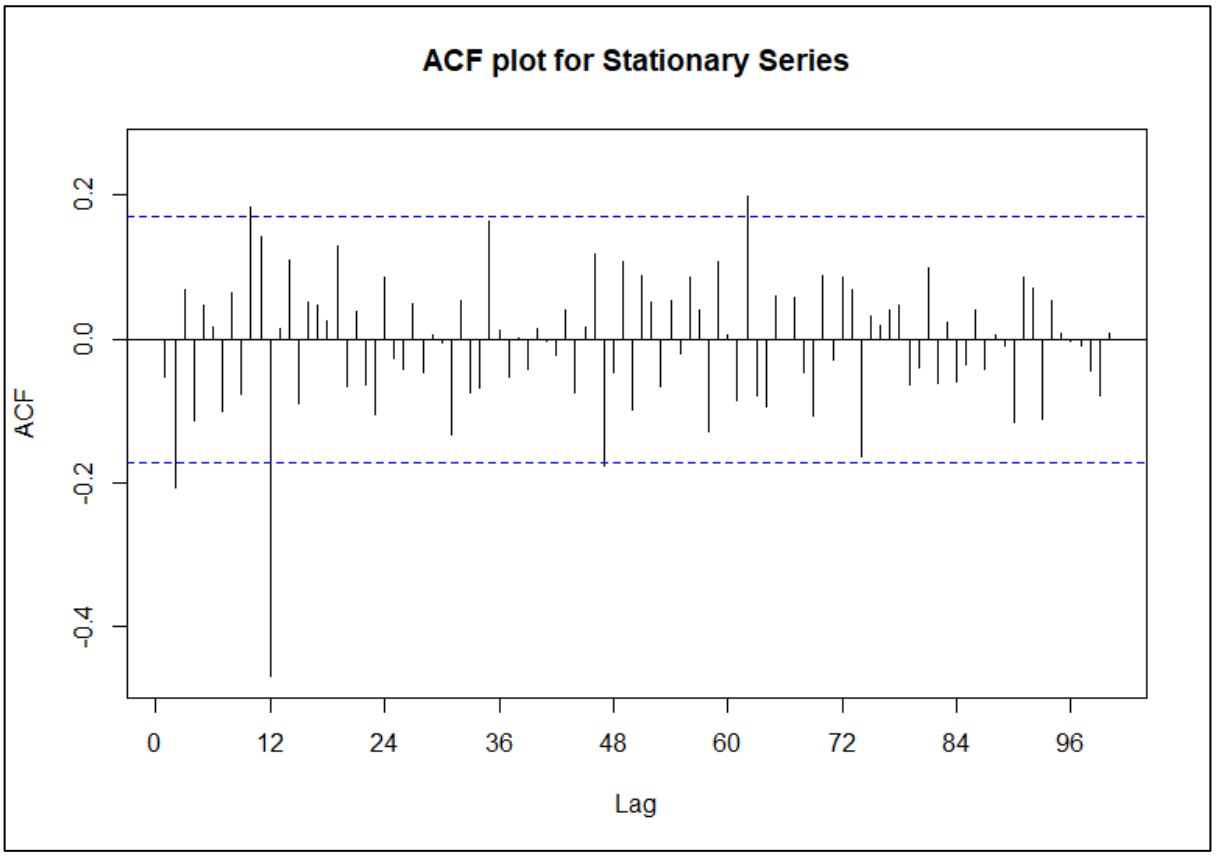
e. How does an ACF plot help to identify whether a time series is stationary or not?

ACF plot can help in identifying if the time series is stationary or not. If the time series is non-stationary, the plot decays slowly and gradually. Whereas, if the time series is stationary, the plot decays rapidly.

For example, the below ACF plot is for a non-stationary series.



In contrast, below is the ACF plot for the same data series after applying Differencing the above non stationary series.



f. Why partitioning time series data into training, validation, and test set is not recommended? Briefly describe two considerations for choosing the width of validation period.

In contrast to cross-sectional data, where we partition our data in Train, Test and Validation datasets by randomly selecting records for each of these subsets; for Time series data, we split it only into Training and Validation datasets. For time-series data, while splitting we ensure that we do not randomly select records, rather we select records for a continuous time period before a specific time point for Training and beyond this specific point for Validation.
The reason for following this practice are as follows:

- For time series analysis/ forecasting, each observation is somehow correlated/ dependant on the previous observation. To have the model extract most about the underlying series, we need continuous period of data for the model to learn from.
- Again for a time series, more the data it trains, more the chances of better forecast accuracy. So, any available data that the time-series model does not see is in a way leads to the model not being able to extract all the underlying pattern from the data.
- The practice of having only Training and Validation dataset allows the model to extract maximum information while ensuring the Validation dataset is reserved for estimating the accuracy of the model.

Two considerations for choosing the width of the validation period are:

- **Forecast Horizon** – The time-period ahead to be forecasted is important in deciding the validation period. The validation period should be at least equal to the forecast horizon. This is because, if our model is expected to predict for a period of up to 1 year in the future, it makes sense to validate our model for at least this period.
- **Length of the Series** – The length of the available time series is an important factor in deciding the width of the validation period. Depending upon the width of the available time series, we need to ensure the training period is not too short, while again, retaining the validation period to be at least that of the forecast horizon.

g. Both smoothing and ARIMA method of forecasting can handle time series data with missing value. True/False. Explain.

False. Neither smoothing nor ARIMA can handle time series data with missing values. This is because both are dependent upon adjacent data.
For e.g. smoothing techniques such as Moving average are dependent upon 'w' consecutive data points to calculate the trend data point. If any of these 'w' data points is missing, the calculated value will be incorrect and thus, lead to incorrect trend calculation.
Similarly, ARIMA depends upon auto-correlation i.e. correlation of data at time 't' with data at different lags. In case of any missing data the correlation calculated will be erroneous and thus, lead to incorrect forecasts.

h. Additive and multiplicative decomposition differ in the way the trend is computed. True /False. Explain.

False.
During decomposition of a time series, the first component calculated/ derived is the Trend. To calculate the trend, we use smoothing techniques like Moving average; which does not have any relation with the time-series being additive or multiplicative. Once, we have calculated the Trend, to estimate the Seasonality and Noise we require to consider whether the time series is Additive or Multiplicative. Accordingly, we subtract the Trend from the actual time series to get Seasonality + Noise in Additive Model. Or, divide original observations by Trend to get Seasonality * Noise in Additive Model.

i. After accounting for trend and seasonality in a time series data, the analyst observes that there is still correlation left amongst the residuals of the time series. Is that a good or a bad news for the analyst? Explain.

After removing the trend and seasonality, if the analyst finds correlation amongst the residuals of the time series, it signifies that more insight can be derived and our Forecast model can be further improved. The analyst needs to use the auto-correlation to build an AR model and this can help to even forecast the residuals. Thus, the overall model can be further improved as the analyst will be able to forecast all three components of the time-series – Trend, Seasonality and Residual; increasing the accuracy% of the overall model. **Hence, finding correlation amongst the residuals is a good news for the analyst.**