ASSIGNMENT SUBMISSION FORM

This will be the first page of your assignment

Course Name : **Forecasting Analytics**
Assignment Title : **Individual Assignment**
Submitted by : **Unnati Khinvasara**

Student Name	PG ID
Unnati Khinvasara	12120097

ISB Honour Code

- I will represent myself in a truthful manner.
- I will not fabricate or plagiarise any information with regard to the curriculum.
- I will not seek, receive or obtain an unfair advantage over other students.
- I will not be a party to any violation of the ISB Honour Code.
- I will personally uphold and abide, in theory and practice, the values, purpose and rules of the ISB Honour Code.
- I will report all violations of the ISB Honour Code by members of the ISB community.
- I will respect the rights and property of all in the ISB community.
- I will abide by all the rules and regulations that are prescribed by ISB.

Note: Lack of awareness of the ISB Honour Code is never an excuse for a violation. Please go through the Honour Code in the student handbook, understand it completely. Please also pay attention to the following points:

- Please do not share your assignment with your fellow students under any circumstances if the Honour Code scheme prohibits it. The HCC considers both parties to be guilty of an Honour Code violation in such circumstances.
- If the assignment allows you to refer to external sources, please make sure that you cite all your sources. Any material that is taken verbatim from an external source (website, news article etc.) must be in quotations. A much better practice is to paraphrase the source material (it still must be cited).

(Please start writing your assignment below)

Forecasting Analytics - Assignment

Unnati Khinvasara

2022-09-25

Q1 - Background

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.

Importing required Libraries *(Code & Output hidden for better presentation)*

Reading and exploring the dataset

```
df <- read_excel("F:\\ISB\\3 - Term 3 - (3-7 Aug) - M\\Forecasting Analytics\\
Individual Assignment (25th Sept)\\SouvenirSales.xlsx")
str(df)
```

```
## tibble [84 × 2] (S3: tbl_df/tbl/data.frame)
##  $ Date : POSIXct[1:84], format: "1995-01-01" "1995-02-01" ...
##  $ Sales: num [1:84] 1665 2398 2841 3547 3753 ...
```

Converting dataset from tibble to dataframe

```
df <- as.data.frame(df)
summary(df)
```

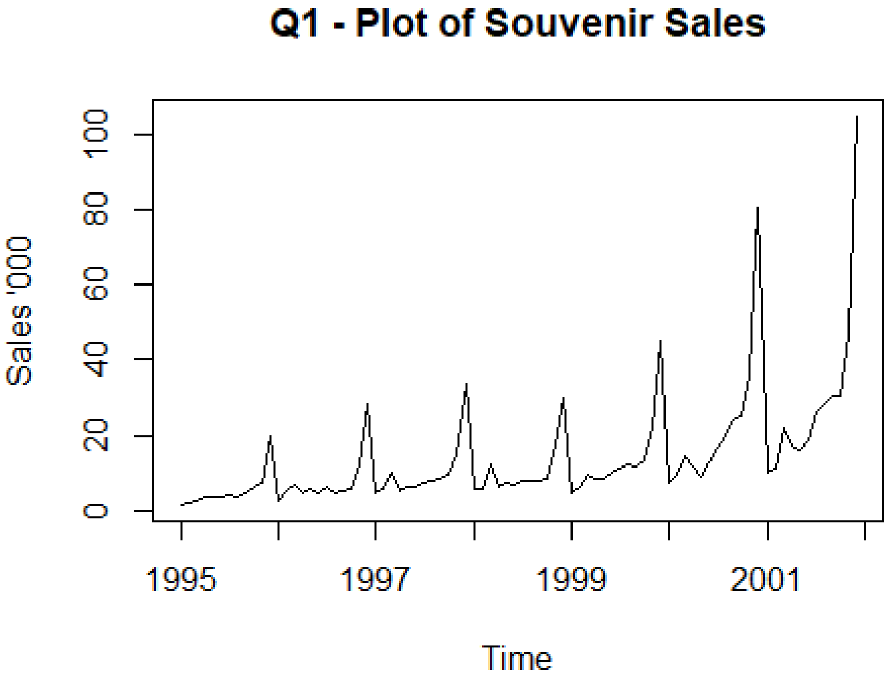
```
##       Date                             Sales       
##  Min.   :1995-01-01 00:00:00.000   Min.   :  1665  
##  1st Qu.:1996-09-23 12:00:00.000   1st Qu.:  5884  
##  Median :1998-06-16 00:00:00.000   Median :  8772  
##  Mean   :1998-06-16 14:17:08.570   Mean   : 14316  
##  3rd Qu.:2000-03-08 18:00:00.000   3rd Qu.: 16889  
##  Max.   :2001-12-01 00:00:00.000   Max.   :104661  
```

```
head(df)
```

```
##         Date   Sales
## 1 1995-01-01 1664.81
## 2 1995-02-01 2397.53
## 3 1995-03-01 2840.71
## 4 1995-04-01 3547.29
## 5 1995-05-01 3752.96
## 6 1995-06-01 3714.74
```

Q1(a) - Plot the time series of the original data. Which time series components appear from the plot.

```
q1.ts <- ts(df$Sales, start= c(1995,1), frequency = 12)
q1.ts

##            Jan       Feb       Mar       Apr       May       Jun       Jul
## 1995   1664.81   2397.53   2840.71   3547.29   3752.96   3714.74   4349.61
## 1996   2499.81   5198.24   7225.14   4806.03   5900.88   4951.34   6179.12
## 1997   4717.02   5702.63   9957.58   5304.78   6492.43   6630.80   7349.62
## 1998   5921.10   5814.58  12421.25   6369.77   7609.12   7224.75   8121.22
## 1999   4826.64   6470.23   9638.77   8821.17   8722.37  10209.48  11276.55
## 2000   7615.03   9849.69  14558.40  11587.33   9332.56  13082.09  16732.78
## 2001  10243.24  11266.88  21826.84  17357.33  15997.79  18601.53  26155.15
##            Aug       Sep       Oct       Nov       Dec
## 1995   3566.34   5021.82   6423.48   7600.60  19756.21
## 1996   4752.15   5496.43   5835.10  12600.08  28541.72
## 1997   8176.62   8573.17   9690.50  15151.84  34061.01
## 1998   7979.25   8093.06   8476.70  17914.66  30114.41
## 1999  12552.22  11637.39  13606.89  21822.11  45060.69
## 2000  19888.61  23933.38  25391.35  36024.80  80721.71
## 2001  28586.52  30505.41  30821.33  46634.38 104660.67

plot(q1.ts/1000, main = "Q1 - Plot of Souvenir Sales", xlab = "Time", ylab="Sa
les '000" )
```

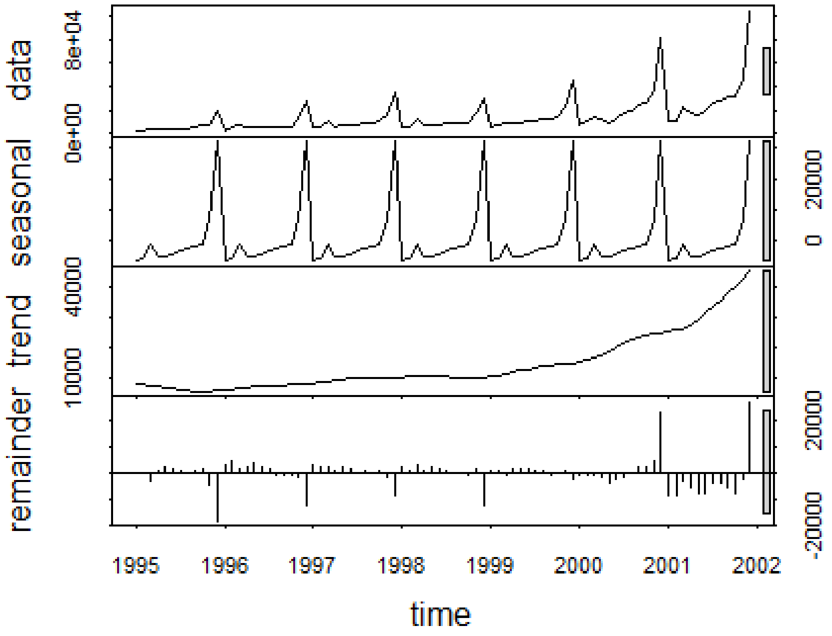


The following is a gist of the time series components which appear from the above charts -

- We can firstly spot the **seasonality** in the dataset. There seem to be 2 seasonal peaks in each year. One smaller peak in March and another significant peak in December. This could be attributed to the fact of holiday & vacation season of Easter and Christmas.
- Secondly, we can see that the **trend** is increasing with years and seems to be *exponential* in nature since it is increasing with a wider factor each year.
- Currently, we are not able to observe any **cyclicality** from the plot above.
- Lastly, we can also observe a slight dip in the sales for the year of 1999.

We can check for our observations above by plotting a decomposition of the time series. It reconfirms our observations with respect to **seasonality** and **trend**

```
plot(stl(q1.ts, "per"))
```



Q1(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.

Partitioning the data

```
train <- window(q1.ts/1000,end=c(2000,12), frequency=12)
#autoplot(train) + ylab("Sales '000") + ggtitle("Training Dataset over time")

val <- window(q1.ts/1000,start=c(2001,1), frequency=12)
```

Building Models

Model A - Linear Trend Model with Additive Seasonality

```
model_a <- tslm(train ~ trend+season)
model_a_pred <- forecast(model_a, h = 12, level = 0 )

#Regression Coefficients
summary(model_a)

##
## Call:
## tslm(formula = train ~ trend + season)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.592  -2.359  -0.411   1.940  33.651
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.06555    2.64026  -1.161  0.25029
## trend        0.24536    0.03408   7.199 1.24e-09 ***
## season2      1.11938    3.42206   0.327  0.74474
## season3      4.40884    3.42256   1.288  0.20272
## season4      1.46257    3.42341   0.427  0.67077
## season5      1.44619    3.42460   0.422  0.67434
## season6      1.86798    3.42613   0.545  0.58766
## season7      2.98856    3.42799   0.872  0.38684
## season8      3.22758    3.43019   0.941  0.35058
## season9      3.95556    3.43273   1.152  0.25384
## season10     4.82166    3.43561   1.403  0.16573
## season11    11.52464    3.43882   3.351  0.00141 **
## season12    32.46955    3.44236   9.432 2.19e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.927 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

#Validation Set Errors
accuracy(val, model_a_pred$mean)

##                 ME     RMSE      MAE       MPE     MAPE      ACF1 Theil's U
## Test set -8.251513 17.45155 10.05528 -23.82776 35.30901 0.3206228  2.663002

accuracy(model_a_pred,val)
```

```
##                          ME      RMSE       MAE       MPE     MAPE     MASE
## Training set 2.470535e-16  5.365199  3.205089  6.967778 36.75088 0.855877
## Test set     8.251513e+00 17.451547 10.055276 10.533974 26.66568 2.685130
##                   ACF1 Theil's U
## Training set 0.4048039        NA
## Test set     0.3206228 0.9075924
```

Model B - Exponential Trend Model with Multiplicative Seasonality

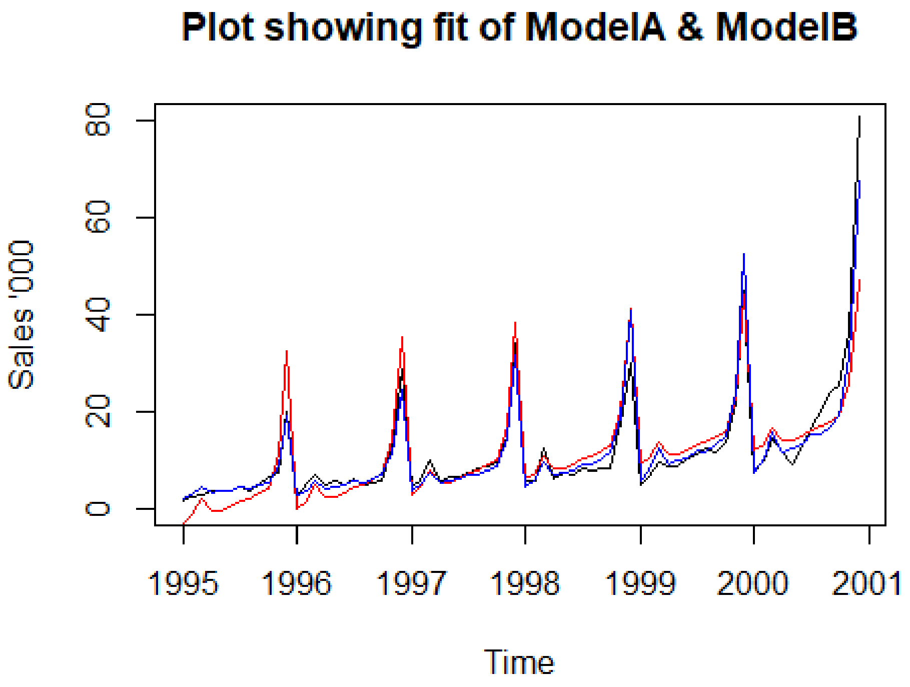
```
model_b <- tslm(train ~ trend+season, lambda=0)
model_b_pred <- forecast(model_b, h = 12, level = 0 )
#model_b_pred

#Regression Coefficients
model_b

##
## Call:
## tslm(formula = train ~ trend + season, lambda = 0)
##
## Coefficients:
## (Intercept)        trend      season2      season3      season4      season
5
##     0.73861      0.02112      0.28201      0.69500      0.37387      0.4217
1
##     season6      season7      season8      season9     season10     season1
1
##     0.44705      0.58338      0.54690      0.63557      0.72949      1.2009
5
##    season12
##     1.95220

#Validation Set Errors
accuracy(val, model_b_pred$mean)

##                 ME     RMSE      MAE       MPE     MAPE      ACF1 Theil's U
## Test set -4.824494 7.101444 5.191669 -16.53324 19.25336 0.4245018 0.6427617

accuracy(model_b_pred,val)

##                    ME     RMSE      MAE       MPE     MAPE     MASE      AC
F1
## Training set 0.197519 2.865154 1.671185 -1.472819 13.94047 0.446268 0.43813
70
## Test set     4.824494 7.101444 5.191669 12.359434 15.51910 1.386367 0.42450
18
##              Theil's U
## Training set        NA
## Test set     0.4610253
```

Q1(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)

Considering RMSE (solved in above part b), we observe that RMSE of Model A (Additive Model) is higher than that of Model B (Exponential Model) and hence Model B is preferred. The same conclusion of exponential trend was observed from the line chart plotted above as the increase in sales for each year was growing with a wider gap continually.

Plotting both models showing forecast values

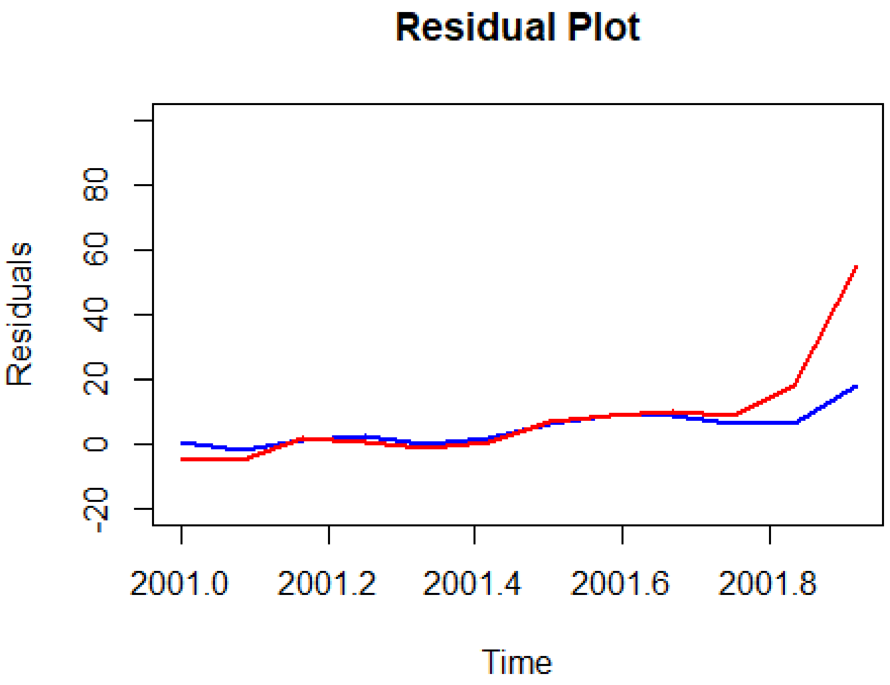
```
#Plotting both Models
plot(train, xlab = "Time", ylab = "Sales '000", ylim= c(0,80), main = "Plot sh
owing fit of ModelA & ModelB")
lines(model_a_pred$fitted, col="red")
lines(model_b_pred$fitted, col="blue")
```



From the above, we can see that the blue line of model B (exponential model) captures the increasing trend of the data better than model A (additive model). This corroborates our theory above.

Residual Plot of Validation Set

```
#Residual Plot for entire timeseries
#plot(train-model_b$fitted.values,main= "Residual Plot", ylab= "Residuals", co
l="blue",lwd=2, ylim=c(-20,50))
#lines(val-model_b_pred$mean, col="blue",lwd=2)
#lines(train-model_a$residuals,col="red",lwd=2)
#lines(val-model_a_pred$mean, col="red",lwd=2)

plot(val-model_b_pred$mean,main= "Residual Plot", ylab= "Residuals", col="blue
",lwd=2, ylim=c(-20,100))
lines(val-model_a_pred$mean, col="red",lwd=2)
```



We can notice that the residual plot for the Model B is closer to zero which that of Model A seems to be higher. Thus, Model B is the preferred trend.

Q(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?

```
model_a_pred$mean

##           Jan      Feb      Mar      Apr      May      Jun      Jul      Au
g
## 2001 14.84603 16.21078 19.74560 17.04469 17.27368 17.94083 19.30678 19.7911
6
##           Sep      Oct      Nov      Dec
## 2001 20.76450 21.87597 28.82431 50.01459
```

The additive model follows the trend seen in the data with a small peak in March and a higher peak in December.

Further, we can see that December has the highest average sales during the year. This is in line with our previous observation of seasonality during the end months.

```
model_a$coefficients

## (Intercept)       trend     season2     season3     season4     season5
##  -3.0655544   0.2453642   1.1193842   4.4088450   1.4625675   1.4461950
##     season6     season7     season8     season9    season10    season11
##   1.8679775   2.9885633   3.2275808   3.9555600   4.8216574  11.5246383
##    season12
##  32.4695508
```

We also observe that the estimated trend coefficient is 0.2453. It means that there is increase of $24,000 in sales per month.

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

```
model_b_pred$mean

##            Jan       Feb       Mar       Apr       May       Jun       Jul
## 2001  9.780022 13.243095 20.441749 15.143541 16.224628 16.996137 19.894424
##            Aug       Sep       Oct       Nov       Dec
## 2001 19.591112 21.864492 24.530299 40.144775 86.908868
```

The multiplicative model also follows the previous observation of dataset with smaller peak in March and higher peak in December. It is to be noted that the sale values are closer to the original dataset with wider gap in end of year sale values.

```
model_b$coefficients

## (Intercept)       trend     season2     season3     season4     season5
##  0.73860759  0.02111965  0.28201487  0.69499827  0.37387340  0.42170998
##     season6     season7     season8     season9    season10    season11
##  0.44704613  0.58337985  0.54689670  0.63556505  0.72949049  1.20095408
##    season12
##  1.95220222
```

Co-efficient of October (season 10) is 0.7294 which means that the sales in the month of October is higher than January (reference month) sales by 72.94%.

Further, the estimated trend co-efficient is 0.0211, which is the Beta 1 of the model. Effectively it means that sales increases by 2.11% every month in the current model.

Q(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.

We use the best model (Model B - Exponential Trend Model) to forecast the sales in January. It is to be noted that we will use the entire original dataset to forecast values for Jan 2002.

```
model_b_jan02 <- tslm(q1.ts ~ trend + season, lambda=0)
jan_02 <- forecast(model_b_jan02, h = 1)
```

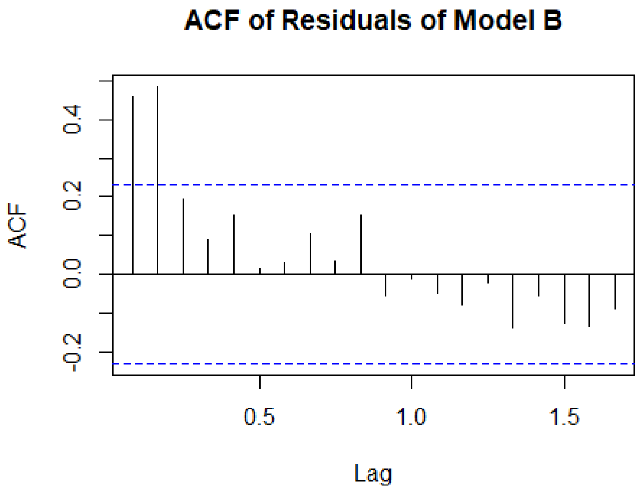
The forecast for the month of Jan 2002 using the model B on original dataset (1995 - 2001) is as follows -

```
jan_02$mean[1]

## [1] 13484.06
```

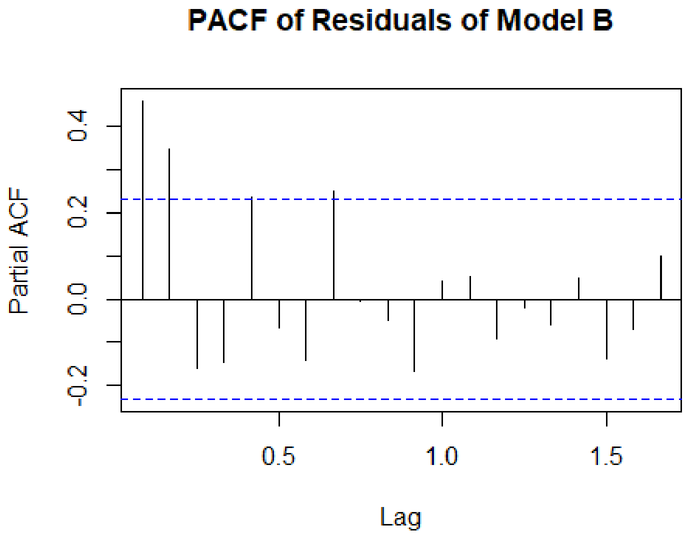
Q(g) Plot the ACF and PACF plot until lag 20 of the residuals obtained from training set of the best model chosen. Comment on these plots and think what AR(p) model could be a good choice?

The ACF & PACF plot of the exponential model (best chosen model = Model B) is as follows-

```
acf(model_b$residuals,na.action=na.pass,lag.max = 20, main="ACF of Residuals o
f Model B")
```



```
pacf(model_b_pred$residuals,na.action=na.pass,lag.max = 20, main="PACF of Resi
duals of Model B")
```



Observing the above ACF plot, we note that there are 3 observations above the significant line, post which the observatins are around the zero mark. Further, even for PACF plot, we have 3 observation above the line (with the 4th observation barely on the significane line).

Hence we believe that AR(3) model would be appropriate.

Q(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?

```
ar_model <- Arima(model_b$residuals, order = c(3,0,0))
summary(ar_model)

## Series: model_b$residuals
## ARIMA(3,0,0) with non-zero mean
##
## Coefficients:
##          ar1     ar2      ar3     mean
##       0.3469  0.3996  -0.1208  -0.0013
## s.e.  0.1155  0.1138   0.1236   0.0426
##
## sigma^2 = 0.02053:  log likelihood = 39.5
## AIC=-69   AICc=-68.09   BIC=-57.62
##
## Training set error measures:
##                       ME      RMSE       MAE      MPE     MAPE      MASE
## Training set 0.003079309 0.1392322 0.1138666 2205.758 2506.715 0.5945732
##                     ACF1
## Training set -0.02678944
```

We can see that the RMSE is 0.13 and the mean is very close to 0. Hence our intuition of AR(3) model was right.

Q(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.

Using Model B and the AR(3) Model, we have forecasted sales for Jan 2002 -

```
#Arima Forecast
ar_jan <- forecast(ar_model, h = 1, level = 0)
ar_jan$mean*1000

##          Jan
## 2001 85.0434
```

The final sales value for Jan 2002 is as follows -

```
final_jan_sales <- jan_02$mean[1] + ar_jan$mean[1]*1000
final_jan_sales

## [1] 13569.11
```

Q2 - SHORT ANSWER TYPE QUESTIONS

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

- Cross Sectional Data is data which is collected at one point of time relating to multiple objects, for eg – data of stocks at a given day. Whereas Time Series Data relates to multiple observations of a single object at different time intervals, for eg – annual data of TCS stock
- Cross Sectional Data is static in terms of time where focus is on the multiple variables, whereas Time Series Data is based on data gained over time interval and main focus is same variable over time period.
- Time-series help us forecast data for upcoming intervals, whereas cross-sectional helps us understand the pattern of the data at a given point of time.
- Use case of both types of data is different. Time Series could help forecast a future condition, whereas cross-sectional is better to understand current position and near future of given objects.
- Time series is normally depicted using a forward line graph, whereas a cross sectional data could be depicted using bar graphs / pie chart / columnar graph, etc.

(b) Explain the difference between seasonality and cyclicality.

Seasonality is the characteristic of a time series data wherein there are regular & predictable fluctuations that recur every year/quarter/month/etc. These occur due to rhythmic forces which are periodical in nature. These fluctuations may be due to seasonal conditions, festivals, etc.

Cyclicality is a characteristic of a time series data wherein fluctuation is neither regular nor fixed and is usually occurring between a larger period of time. Example can be business cycles such as growth / depressions phase, etc which last for multiple years and length of cycle is not known.

Key differences between seasonality & Cyclicality are as follows –

- Regularity – Seasonal variations are regular in nature whereas period of regularity of cyclical variations is now known
- Period – Periods in seasonal variation are shorter than a year, however, cyclical variations are usually longer than a year or 2
- Cause of change – Seasonal variations can be attributed to causes such as festivals, weather, etc. Whereas, cyclical changes are caused to macro-economic factors, business cycles, etc.
- Accuracy in Measurement – Seasonality can be spotted within a year and is relatively easier to measure than cyclicality.

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

Centered moving average is not considered suitable for forecasting because the first & last few observations are lost while calculation of the window area. And hence the trend line ends before the forecast period thus not taking into account last couple of data point trends till the horizon edge. Further, it also heavily depends on past data and fails to change as per upcoming trends which is required for forecasting purposes. Also, it should be noted that moving averages suppresses seasonality.

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

Stationarity means that the time series data does not have a trend or seasonality. It refers to the assumption that the statistical properties of the process creating the times series forecast does not change and is stationary in nature. Thus, we can jot down the time series function as a linear algebraic function with a constant slope & intercept. This does not mean that the values are same over time, it simply means that the method of calculating the value would remain the same.

This assumption helps greatly in time series forecasting, since we have a base foundation for laying the future forecast values. It is said that Stationarity series have constant statistical properties over time and hence forecasting the same is easier.

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

Stationarity refers to having no trend in the time series data. It means that there is a constant variance & auto-correlation pattern in the data. Thus, when the Autocorrelation Function shows exponential decay and closes rapidly towards the zero line, it helps us identify that the time series is stationary. Whereas for a non-stationary time series, ACF drops slowly with a gradual tail.

(f) Why partitioning time series data into training, validation, and test set is not recommended? Describe briefly two considerations for choosing the width of validation period?

Partitioning time series data into training, validation & test set is not recommended because in a time-series data subsequent observation depends on the previous observation by nature of the data & time. And hence if the data is randomly partitioned into training, validation & test set, it loses this relationship with the previous observation and there is no correlation present in the data set.

In general parlance, 70% of the dataset is used for training and the balance 30% is sequentially divided between validation & testing. Considerations for choosing the width of the validation period –

- Number of hyperparameters – If there are more hyperparameters, validation period should be higher and vice versa.
- Recent Data – The most recent time frame should be the validation set since the model would learn based on previous patterns and sequentially apply them on the latest time frame.

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

False. Both ARIMA and smoothing method of forecasting cannot handle time series data with missing values. We have to either drop or impute the values using methods such as interpolation, rolling stats, LOCF, NECF methods.

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

True. Additive & Multiplicative decomposition differ in the way trends are computed. While in additive models, it differs by an absolute value and we need to subtract the series to get the trend. In multiplicative the trend differs by a factor (multiplicative) and we need to divide the same to arrive at the trend.

(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.

If there is correlation left amongst the residuals of the time series data after accounting for trend and seasonality, it is generally a bad sign for the analyst because it means that there is scope of improvement in the forecasting method & parameters used and some modification has to be carried out to the current method to explain the remaining correlation in the time series data.