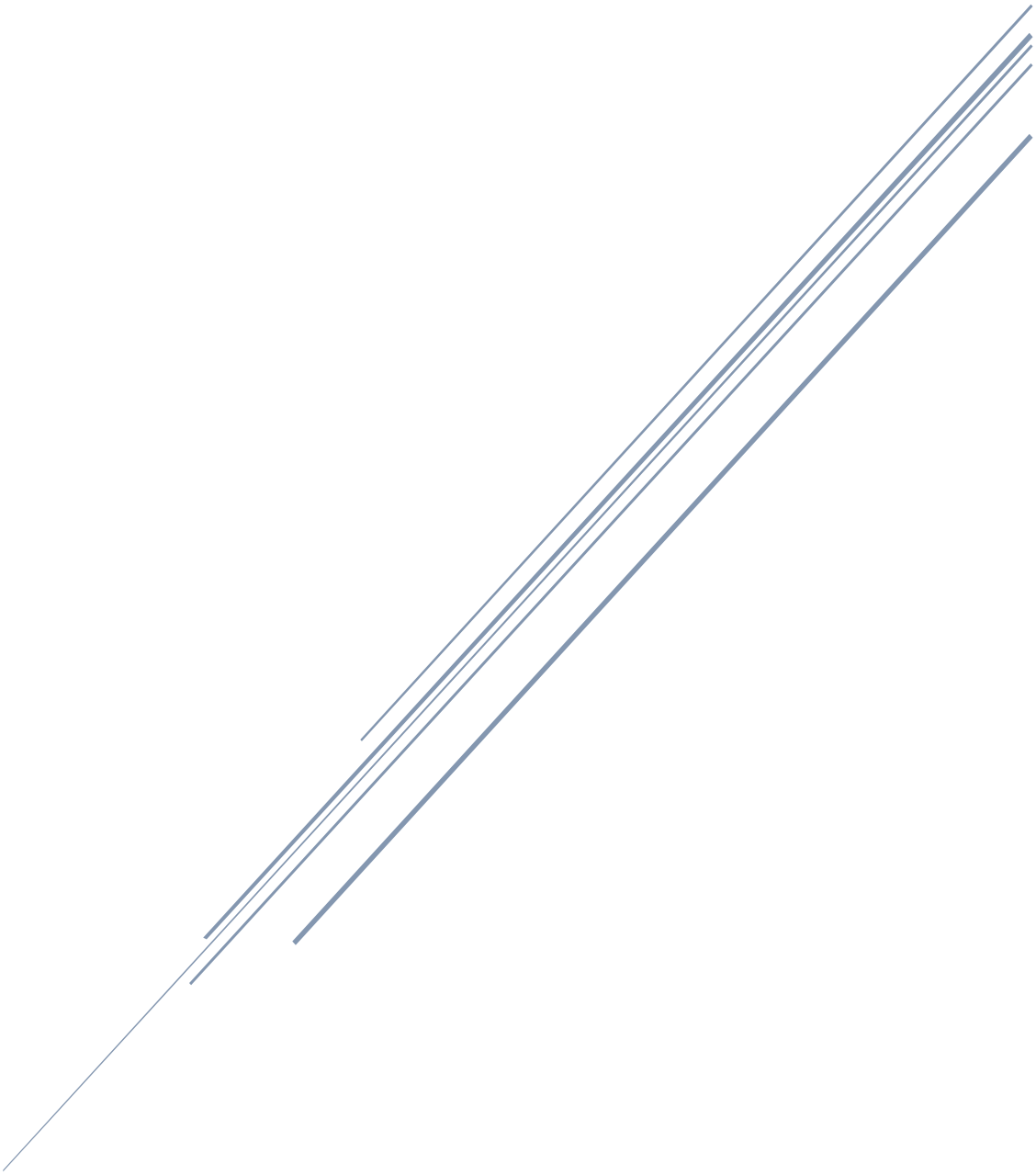
FORECASTING ANALYTICS

Assignment



Submitted By:
Ajay Nathani - 11920092

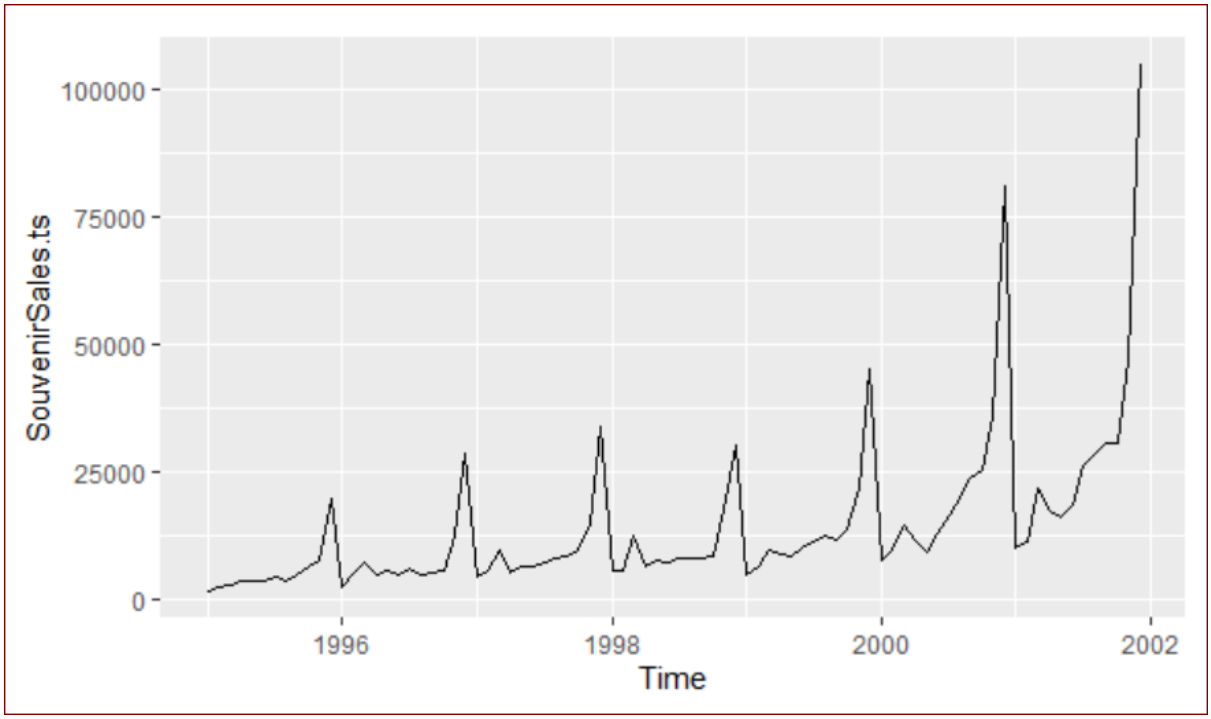
1. Consider the data set SouvenirSales.xls (1995 Jan -2001 Dec) that gives the monthly sales of souvenir at a shop in New York. Back in 2001, an analyst was appointed to forecast sales for the next 12 months (Year 2002). The analyst portioned the data by keeping the last 12 months of data (year 2001)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

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

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

SouvenirSales.ts

autoplot(SouvenirSales.ts)



From the timeseries plot, one can observe that there is level, trend and seasonality present in data.

(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.

#train test split

train <- window(SouvenirSales.ts, end = c(2000,12), frequency = 12)

test <- window(SouvenirSales.ts,start = c(2001,1), frequency = 12)

Model A

#linear trend model with additive seasonality (Model A)

train.modelA <- tslm(train ~ 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 model with multiplicative seasonality (Model B)

train.modelB <- tslm(train ~ 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 *
```

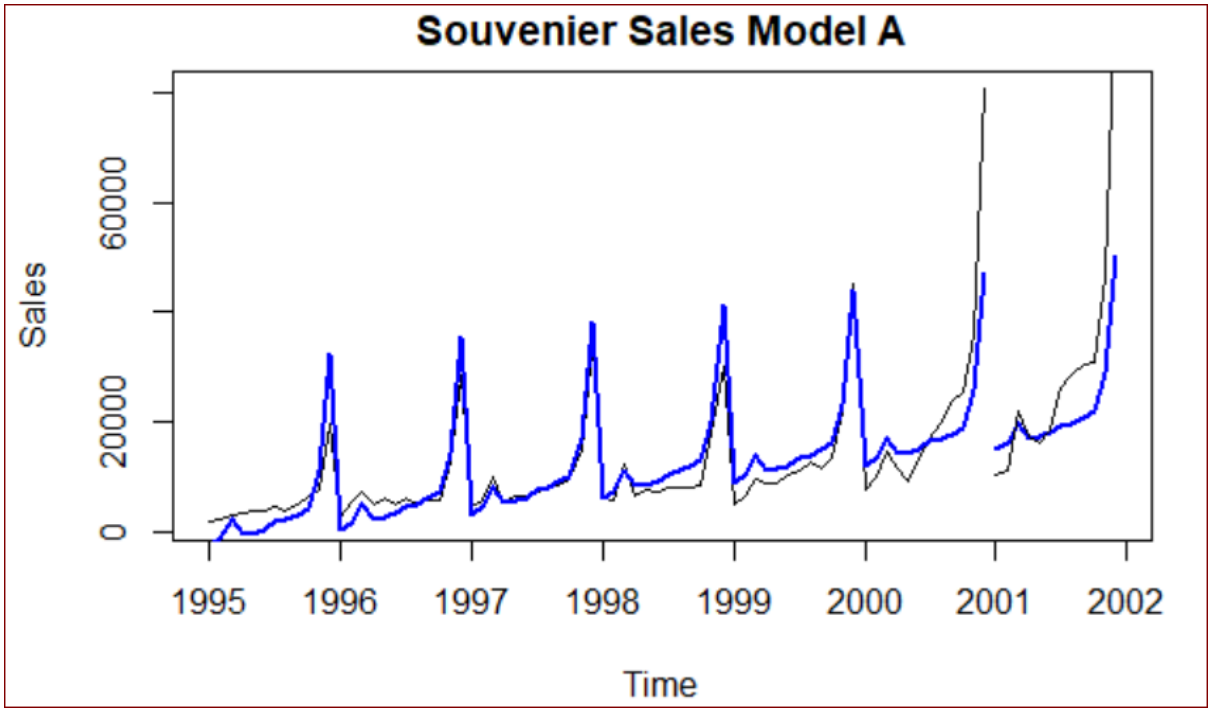
```
season3      0.694998   0.109044   6.374 3.08e-08 ***
season4      0.373873   0.109071   3.428 0.001115 **
season5      0.421710   0.109109   3.865 0.000279 ***
season6      0.447046   0.109158   4.095 0.000130 ***
season7      0.583380   0.109217   5.341 1.55e-06 ***
season8      0.546897   0.109287   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
```

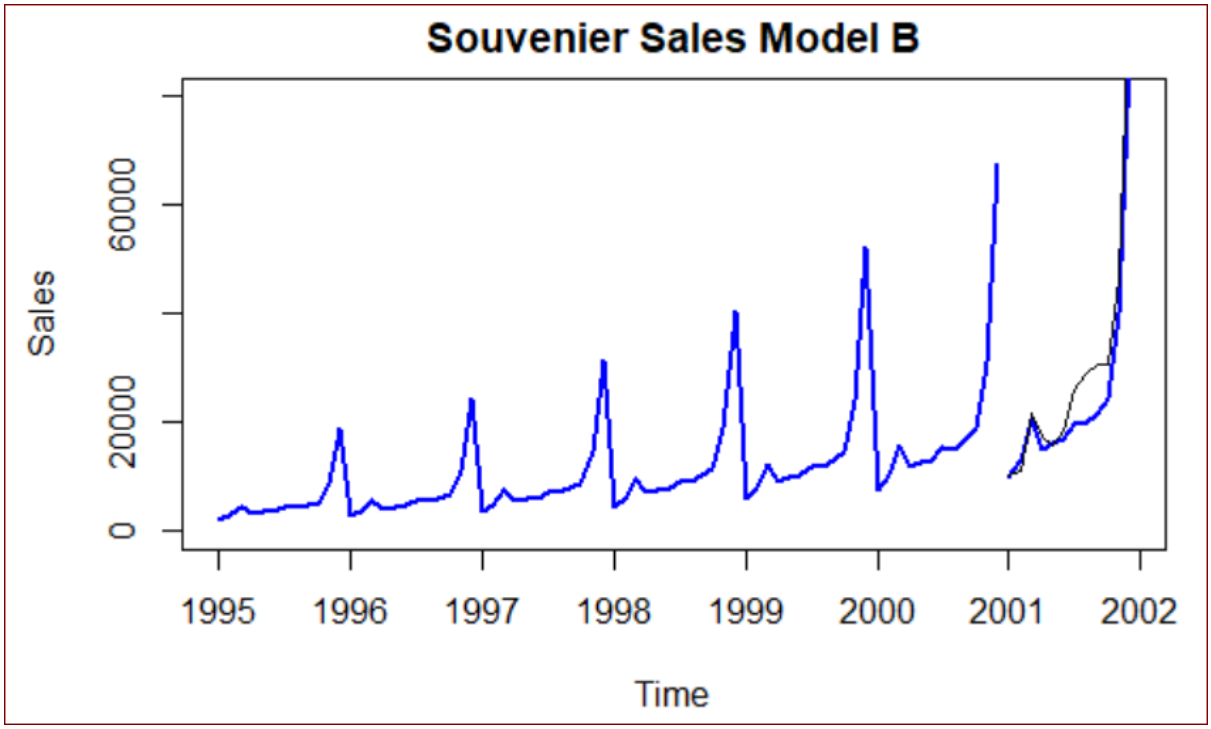
(c) Which model is the best model considering RMSE as the metric? Could you have understood this from the line chart? Explain. Produce the plot showing the forecasts from both models along with actual data. In a separate plot, present the residuals from both models (consider only the validation set residuals).

Model B(exponential trend and multiplicative seasonality) is better considering the RMSE as metric. This was also understood from linechart as we can observe the magnitude of seasonality was increasing with trend, which is better captured in model B.

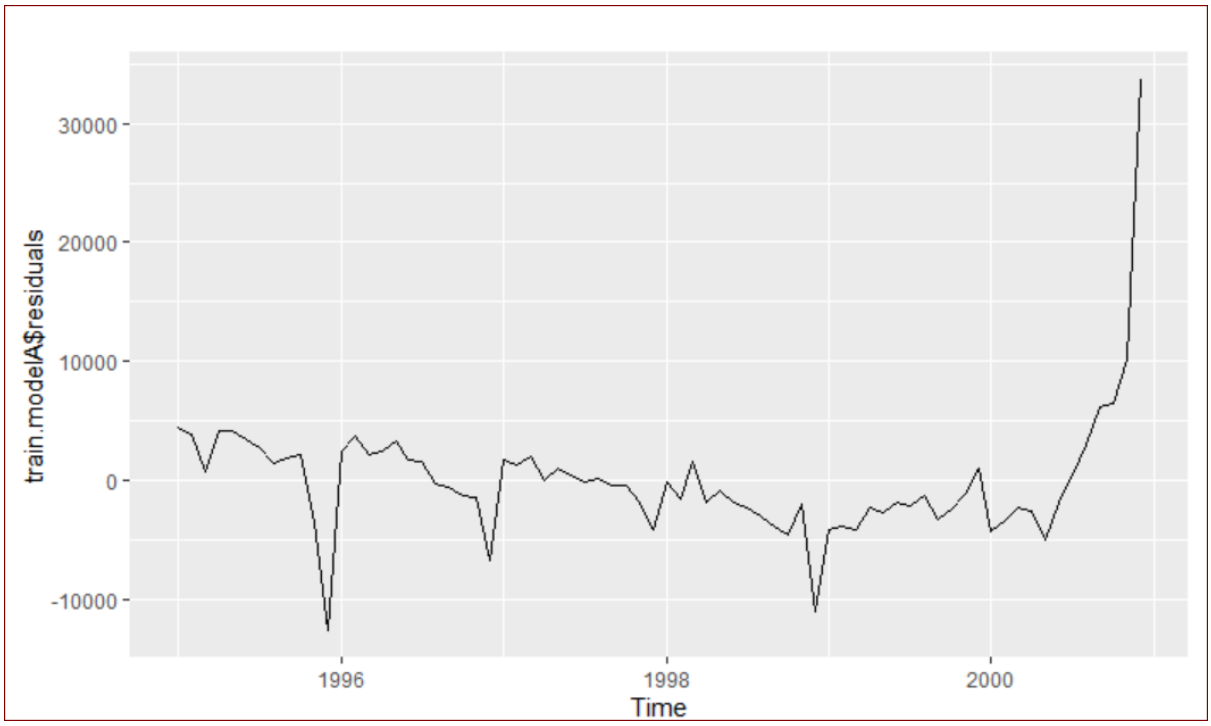
Model A, plot showing forecast with actual data



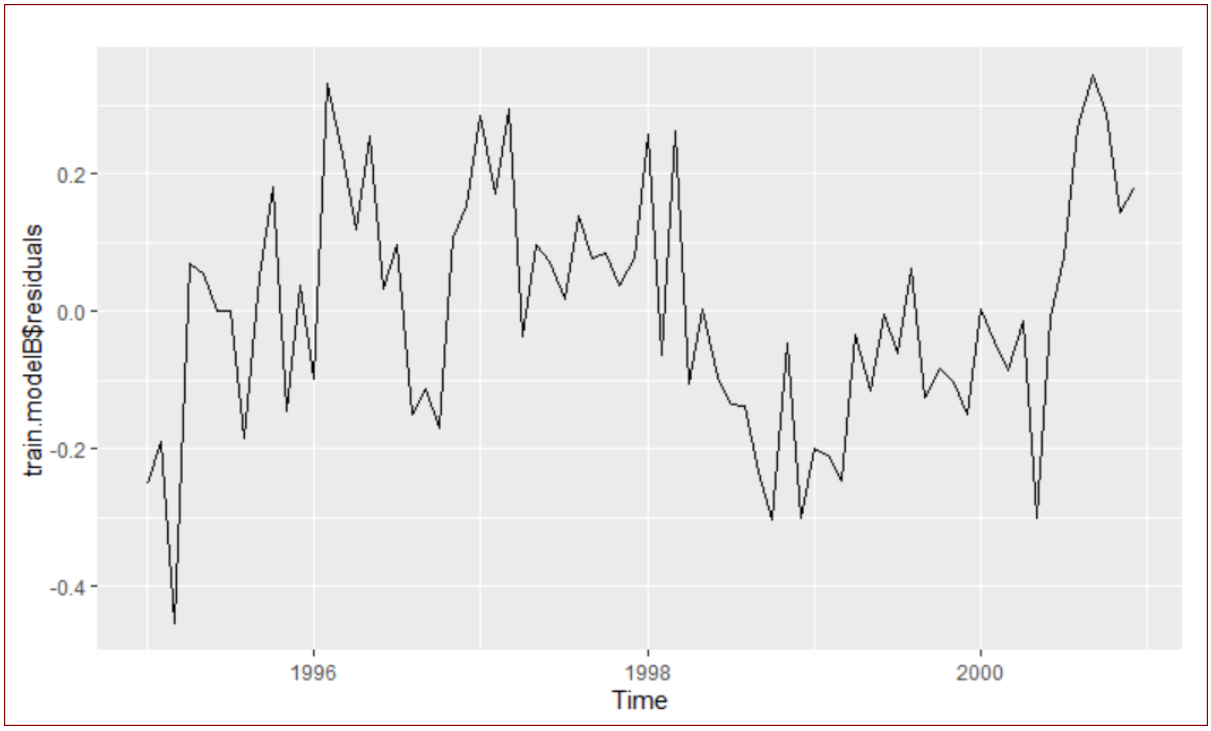
Model B, , plot showing forecast with actual data



Residuals – Model A



Residuals – Model B



(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?

December has the highest average sales during the year. The coefficient of trend (245.36) measures the annual change in Sales keeping all other variables constant, i.e removing seasonality.

summary(train.modelA)

```
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
```

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

The coefficient of October (0.729490) means sales in October are .7294*100 = 72.94% higher than sales in January, keeping other variables constant.

Trend Coefficient (0.021120) means annual change in Sales is 2.11% keeping other variables constant.

```
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
```

(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.

modelB.full <- tslm(SouvenirSales.ts ~ trend + season,lambda = 0)

modelB.predict <- forecast(modelB.full,h=1)

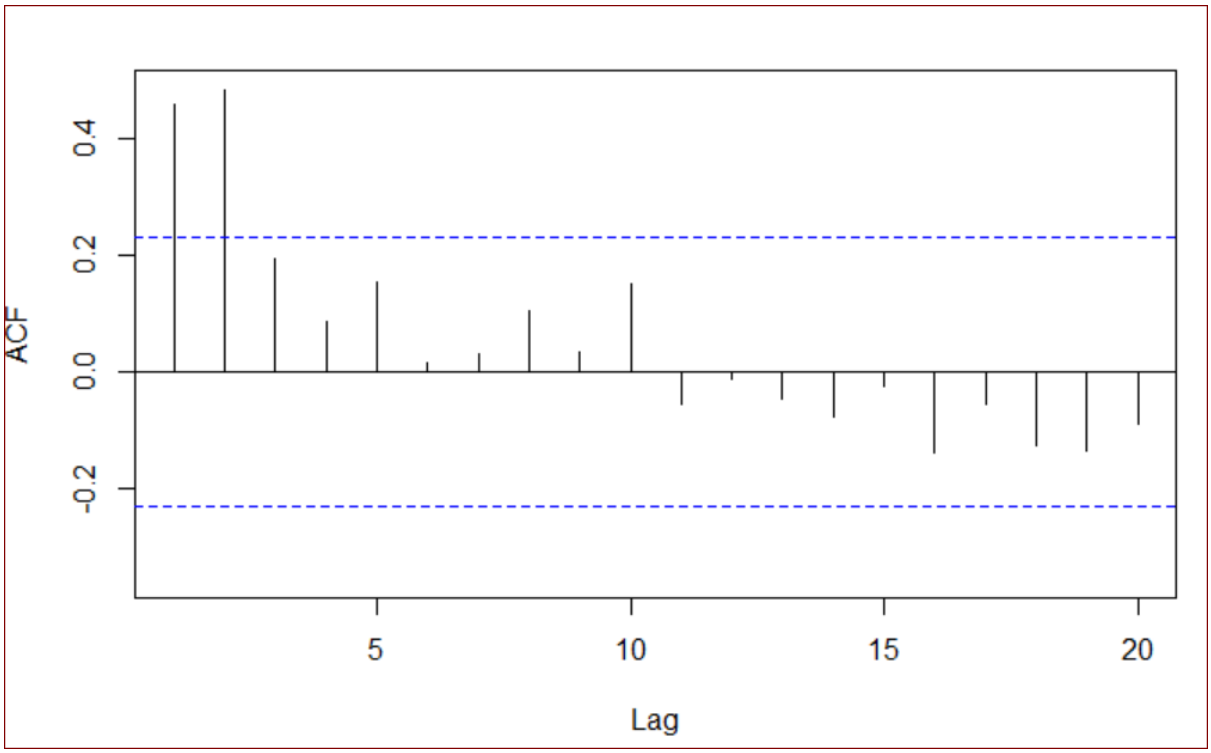
modelB.predict

	Point Forecast <dbl>	Lo 80 <dbl>	Hi 80 <dbl>	Lo 95 <dbl>	Hi 95 <dbl>
Jan 2002	13484.06	10373.35	17527.6	9000.202	20201.76

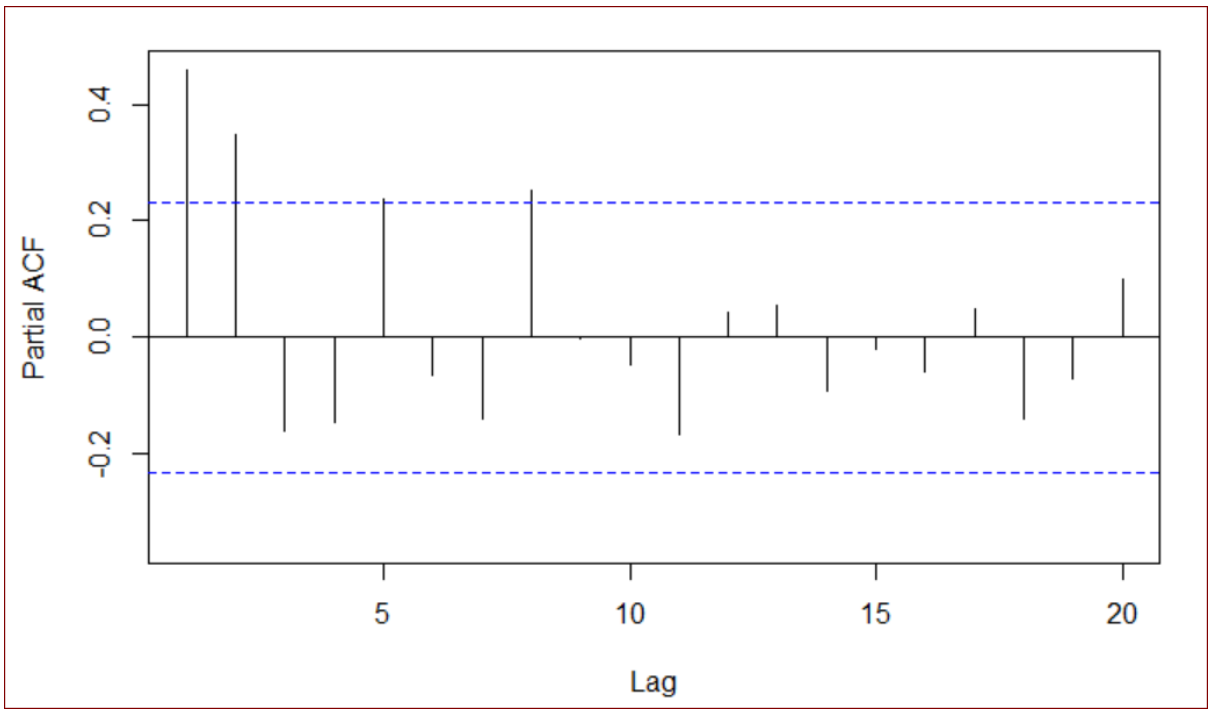
Forecast for Jan 2002 is 13484.06

(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?

ACF plot



PACF plot



From the plot, one can observe the 1st and 2nd value as most significant in both plots, will use pacf to determine p i.e. p =2 and q = 2. Hence, AR(2) seems to be a good choice.

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

```
Call:
arima(x = residuals.modelB, order = c(2, 0, 0))

Coefficients:
         ar1     ar2  intercept
      0.3072  0.3687    -0.0025
s.e.  0.1090  0.1102     0.0489

sigma^2 estimated as 0.01965:  log likelihood = 39.03,  aic = -72.05
```

The intution that it fits AR(2) model is correct. This is evident from standard error , we see that both coeffiecints are more two standard deviation away from zero. This mean that parameters passed the t-test.

(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.

Regression Model

Jan 2002 13484.06

Arima model

Jan 2002 0.06501293

Final value = 13484.06 + .06501293 = 13,484.1250129

2. Short answer type questions: 9 x 2

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

Cross Sectional Data	Time Series Data
Data is being looked at a given point of time.	Series of observation are looked over time
There is IID assumption, independently and identically distributed.	Every observation has a time correlation with previous observation.
We do prediction for cross sectional data.	We forecast the future for time series data.

b) Explain the difference between seasonality and cyclicality.

Seasonality	Cyclicality
Short term variation due to seasonal factors.	Medium term variation repeating at irregular intervals.
Magnitude of variation is not very high.	Magnitude of variation is high.
Pattern will repeat at regular time interval.	Patter happens at irregular interval.
Two types of Seasonality- Additive and Multiplicative	Generally treated along with trend.

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

Centered moving average can be used to remove seasonality. It is not considered suitable for forecasting because it starts early in forecast but also end early. This is not ideal for forecasting as in timeseries, the latest observations are usually most important.

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

A timeseries is considered stationary if its mean, variance and covariance are constant with time. Constant mean implies no trend, constant variance and covariance implies seasonality effect is minimal.

It is important for time series forecasting methods like ARIMA, because it assumes that there is no trend and seasonality in data and we only have to check the impact of residual to correctly predict the future.

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

If a time series is stationary, ACF plot falls very sharply i.e. it will drop to 0 very quickly. And, if it does not fall very sharply, that means timeseries is not stationary.

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 and test set is not recommended because in timeseries the latest observations are most important. And if we exclude the test set while training our model, then we are possibly missing out on most important piece of information. This is usually not the problem with cross-sectional data.

Consideration while choosing the width of validation period:

1. Atleast 1 full seasonal cycle is captured in validation set.
2. Validation period should be atleast equal to forecasting horizon. A shorter period will fail to mimic actual scenario.

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

False, Smoothing and ARIMA does not work with missing values.Both these models depend on the latest previous value to forecast future value, hence if the timeseries has missing value, these values are to be imputed first for smoothing and ARIMA to work.

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

False, additive and multiplicative decomposition does not differ in way the trend is computed. Trend is increase or decrease in timeseries over long period of time and it follows the same process to calculate whether in additive or multiplicative.

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 analyst observes that correlation is still left among residuals after removing trend and seasonality, this is a good news. It means that there is autocorrelation present and it can be used to make our predictions even better. It can be either positive or negative and will help in improving the forecast and also to evaluate the predictability of series. If the residuals are not autocorrelated, then it is simply noise and we cannot further improve our results.