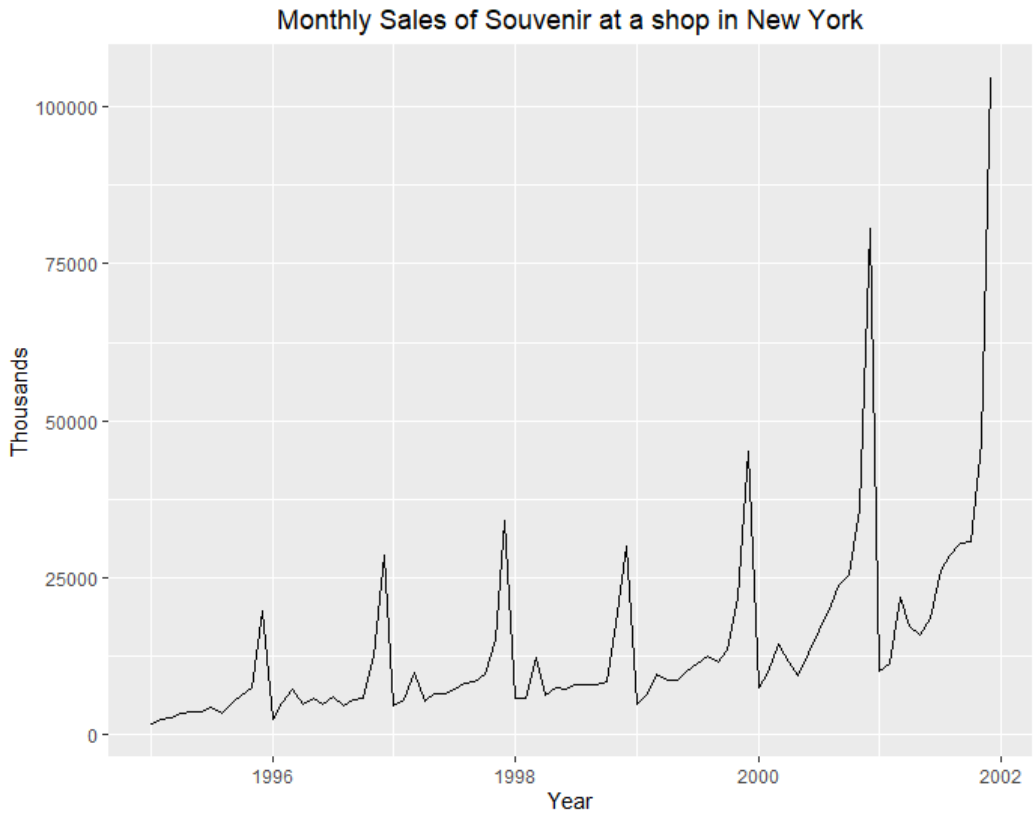
FORECASTING ANALYTICS

Assignment-Report

Name: **Yash Srivastava**
PGID: **12010060**
EMAIL: **Yash_srivastava_ampba2021s@isb.edu**

Q1. A)



The above is the time series plot of Monthly Sale of Souvenirs.

The plot has all three components of Time series:

1. **Level**: The plot has average sales per Month from year 1995 to 2001.

2. **Trend**: As we can see from the plot, the variations are going up with time, hence we can intuitively say that there is an exponential trend.

3. **Seasonality**: There are short term variations which are occurring at regular intervals. Every year towards the end (mostly during the early days of the last month) we can see an increase in average sales and then the sales drop and there is some increase. This cycle is occurring every year from December 1995 to Dec 1996, 1997…….2001.

   Since it is a monthly data, we can say that the seasonality index is 12.

Q1. B

Fit a linear trend model with additive seasonality (**Model A**)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3065.55	2640.262	-1.16108	0.250286
trend	245.3642	34.08279	7.199063	1.24E-09
season2	1119.384	3422.055	0.327109	0.744743
season3	4408.845	3422.564	1.28817	0.202716
season4	1462.567	3423.413	0.427225	0.670771
season5	1446.195	3424.6	0.422296	0.674344
season6	1867.977	3426.126	0.545216	0.587661
season7	2988.563	3427.99	0.871812	0.386845
season8	3227.581	3430.192	0.940933	0.350576
season9	3955.56	3432.731	1.152307	0.253843
season10	4821.657	3435.606	1.403437	0.165728
season11	11524.64	3438.817	3.351338	0.001408
season12	32469.55	3442.362	9.432346	2.19E-13

Fit an Exponential trend model with multiplicative seasonality (**Model B**)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.646362873	0.084119807	90.89848383	4.08E-65
trend	0.021119646	0.001085892	19.44913013	2.42E-27
season2	0.28201487	0.109028054	2.58662666	0.012178334
season3	0.694998267	0.109044275	6.37354199	3.08E-08
season4	0.3738734	0.109071306	3.427788799	0.001115071
season5	0.421709982	0.109109137	3.865029017	0.000279085
season6	0.447046125	0.109157759	4.095413185	0.000130147
season7	0.58337985	0.109217156	5.341467153	1.55E-06
season8	0.546896699	0.10928731	5.004210437	5.37E-06
season9	0.635565046	0.109368202	5.811241611	2.65E-07
season10	0.729490487	0.109459807	6.664459845	9.98E-09
season11	1.200954082	0.109562098	10.9614009	7.38E-16
season12	1.952202221	0.109675046	17.79987601	2.13E-25

Linear tend Model
RMSE: **17451.55**

Exponential Trend Model
RMSE: **7101.444**

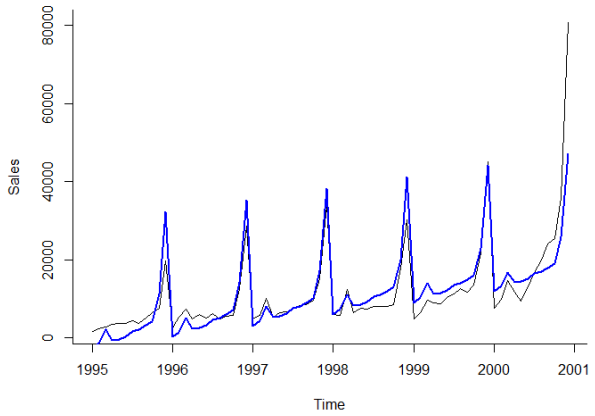
Q1. C

Considering RSME as the metric we can that Exponential Trend model for Multiplicative seasonality is better than linear trend model for additive seasonality as it has low RMSE.

Linear tend Model	Exponential Trend Model
RMSE: **17451.55**	RMSE: **7101.444**

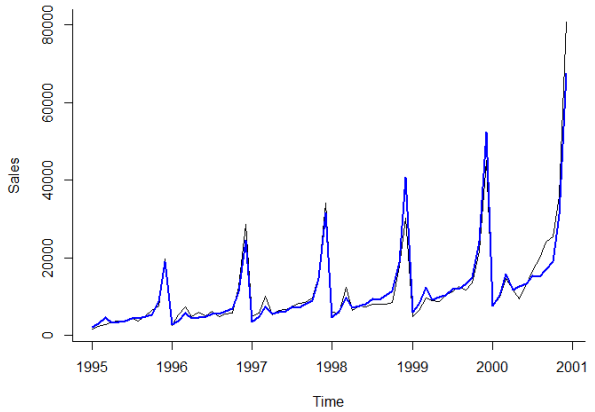
This was evident from the line chart as the Exponential Trend Model was able to fit the data better than Linear tend Model.

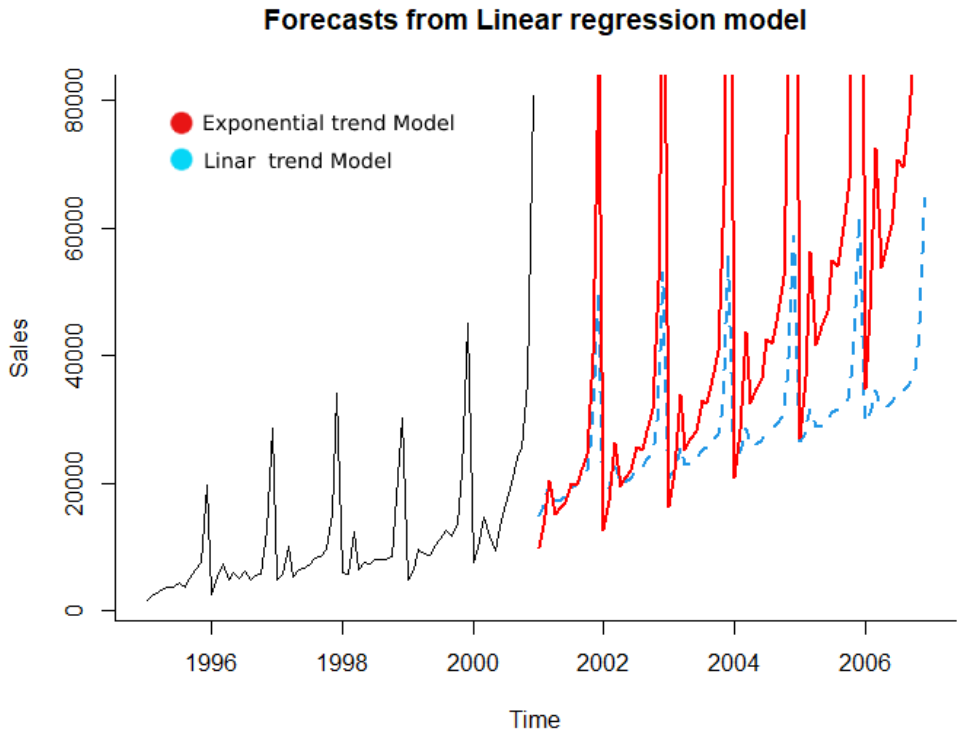
Actual Vs Fitted for Linear Trend Model



Blue line is the fitted line.

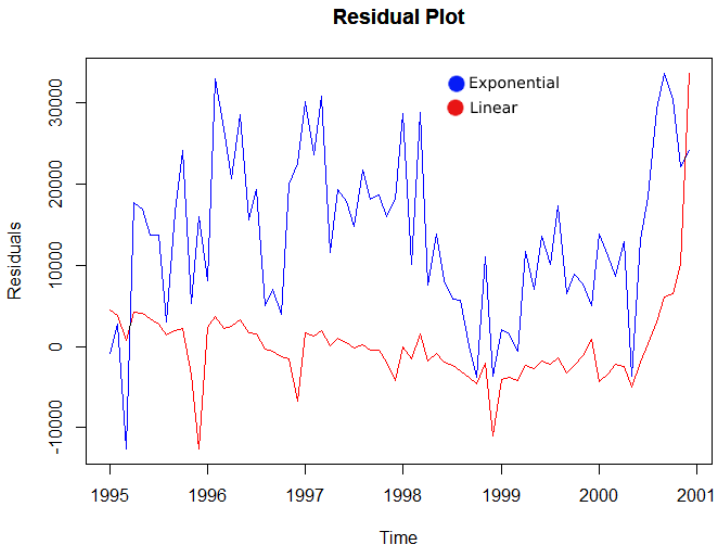
Actual Vs Fitted Exponential Trend Model





From the above plot, we can see that forecast was Linear trend model is slightly going up from the trend which will result in overfitting model, while the forecast for exponential trend model is correctly moving with the trend.

This can be verified using the Residual Plot from the Validation set.



As we can see from the plot the residual from Linear trend is going down, this is because the model the model was capturing some trend which was never there, as a result the residuals went down.

In case of Exponential model, the residuals are moving across the horizontal which indicates a better fit.

Q1 D) **Additive Model:** From the coefficients we can see that Month 12, December has the highest average sales with respect to January.

Season12
Estimate: **32469.55**

This means that for unit increase in time, the average sales for December is increasing by 32469.55 w.r.t January by controlling other factors.

Trend
Estimate: **245.36**

This means that the trend is increasing by the magnitude of 245.36 for a unit increase in time.

Q1 E) Multiplicative Model:
Season10, October
Estimate: **0.729**

This means that for unit increase in time, the average sales for October is increasing by 72.9% w.r.t January by controlling other factors.

Trend
Estimate: **0.021**

This means that the trend is increasing by the magnitude of 2.1% for a unit increase in time.

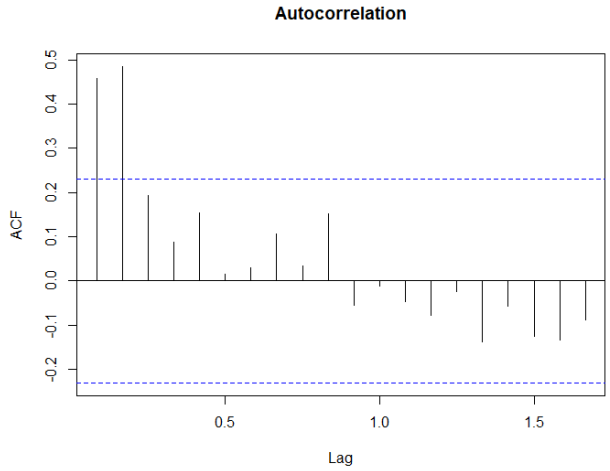
Q1 F) From part C we have seen that multiplicative model fits the data better than additive model. Therefore we use this model to forecast for sales in January in 2002.

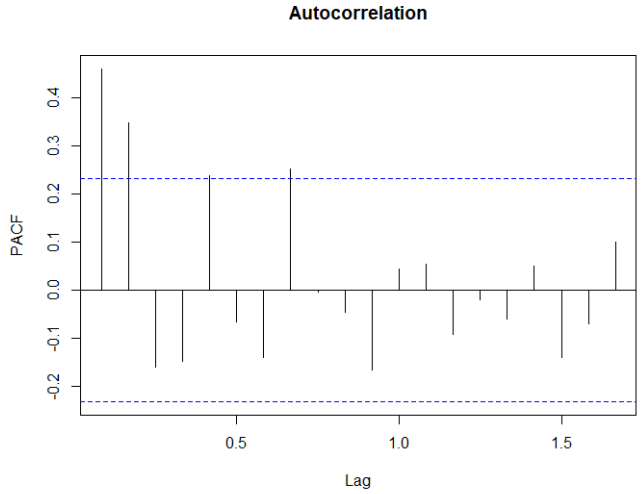
We need to use the entire dataset because the model has already seen the training set and we have validated it also using the validation set. Therefore while using the entire dataset the forecast will have information from recent observations as well as past data.

```
Exp_forcast <- tslm(my_data ~ trend+season, lambda = 0)
forecast(Exp_forcast, h=1, level = 0)
```

= **13484.06**
This means that average sales for January is predicted to be 13484.06.

Q1 G)





From the ACF plot we can see that there is fast decay which signifies that residuals have no trend. The first and second lag of residuals have strong autocorrelation while the remaining are between the significance lines stating there is no correlation between them. So they are random or white noise.

From the PACF plot we can see that first and second lag of residuals are strongly correlated signifying that Yt-1, and Yt-2 have a direct impact on Yt.
So intuitively we can say that AR[p] model will have p = 2.

However, while most of remaining lag of residuals are between the significance lines, we can see 5^{th} and 7^{th} lag are slightly above the line.
This could be due the noise in the data.

Q1 H)

Fitting AR[p] model
error1 =train_et$residuals
train_arima <- Arima(error1, order = c(2,0,0))
summary(train_arima)

Coefficients:

	ar1	ar2	mean
	0.3072	0.3687	-0.0025
s.e.	0.1090	0.1102	0.0489

Using Auto ARIMA we can check our assumptions.
Coefficients:

	ma1	ma2	sma1
	-0.7050	0.4720	0.6439
s.e.	0.1416	0.1617	0.2139

As we can see the auto arima model is estimating coefficients only for 2 lag of residuals thus proving our assumption that AR[p] will have **p = 2.**

Q1 I) Taking the forecast from part F we have 13484.06. Now this is adjusted by estimating the residual from the AR model.

error2 = Exp_forcast$residuals
train_arima2 <- Arima(error2, order = c(2,0,0))
summary(train_arima2)

forecast.auto <- forecast(train_arima2,h=1

Forecasting the residual for January 2002 we get = 0.06501293

Adding this to the previous forecast we get = 13484.06 + 0.06501293

= **13484.125**

Q2 **. A)**
Cross section data is a snapshot of time. It is collected at a single point in time. For eg: The sales of a product for a given Month.

Time-series data are collected over time at regular time intervals. For eg: Monthly sales of a product collected over period of 2 years.

For building a model on a cross sectional data we can do a random split into train and test, while for a time series data we do sequential split into train and test.

B) Seasonality is the short-term variations in the trend due to seasonal factors which occurs at regular intervals, while Cyclicality is the medium term variations which occur at irregular patters.

For eg: Every year during October-November we can a spike in sales, this is mainly due to the festival season and this pattern occurs every year. Hence we can say that there is seasonal cycle occurring after every 12months. In case of cyclicality there is no regular pattern like pandemic or recession which may occur at any period of time without following any seasonal cycle.

C) Centered Moving average is not suitable for forecasting because the we lose information as the first and last few obversions are lost. For time series recent information is very important and if we lose it then model might not fit the data well. Therefore we use trailing moving average as it uses latest information.

D) Stationarity is important in forecasting because the data generated is in statistical equilibrium. This means that the data has constant mean, constant variance and constant covariance. This means that mean, variance and covariance does not change with time. When we say that there is Stationarity it also means that data has no trend and seasonality.

The covariance between Yt and Yt-k will constant if the data has stationarity.
This is the assumption of AR and ARIMA models.

E) The ACF plot helps us to understand if the time series is stationary or not. When the we look at the lag of residuals and we see that there is a smooth decay we say that time series is not stationary. This is because the residuals have trend. For the time series to be stationary the lag of residuals should have a fast decay signifying that there is no trend.

F) Partitioning the data into train test and validation is not suitable in forecasting because we will lose the temporal correlation that is present in the data. This means that the present data is strongly correlated with the past data. So if sequence is broken the forecast will not be accurate.

G) Smoothing and ARIMA do not work with missing values.

H) Additive and Multiplicative decomposition differs in the way trend is computed. For additive decomposition the trend is subtracted while for multiplicative decomposition the trend is divided.

Unadj St = Yt – T <- Unadjusted seasonality is when we subtract Trend - Additive
Unadj St = Yt /T <- Unadjusted seasonality is when we divide Trend – Multiplicative

I) After accounting for trend and seasonality if there is still some correlation left in the residuals then it is a bad news. We estimate the residuals to understand the extra information that they have which might help in making the forecast better. However, after doing everything if we still find some correlation this means that is still some unexplained information that we are not able to analyze, this might affect the forecast.

However we can try to solve it by looking at the PACF plots again and see there might be some residuals which show autocorrelation and build an AR model around it to estimate the coefficients by increasing the value of P.