Regression CP-2: #1

07/06/2021

Andrew Ouyang

Table of Contents

Problem #1- Real-life Data Analysis for Variable Selections

1.1 Data Background Description

The data in this report will be taken from monthly sale data for a souvenir shop at a beach resort town in Queensland. We will be using the data from 1987 to 1992 to predict the sales data per month for 1993. We will be using an exponential weighted moving average (EWMA) approach for forecasting the 1993 sales data and will compare the actual 1993 sales data to our prediction.

Logically, we can expect that there will be a cyclical trend in sales, as more tourists would visit the souvenir shop during the summer and spring months and less tourists would visit during the winter months. Although we can't really draw any predictions for sales over the years given the limited information, we can reasonably expect that sales trends will, for the most part, increase over the years as population and inflation rises.

1.2 Data Plot

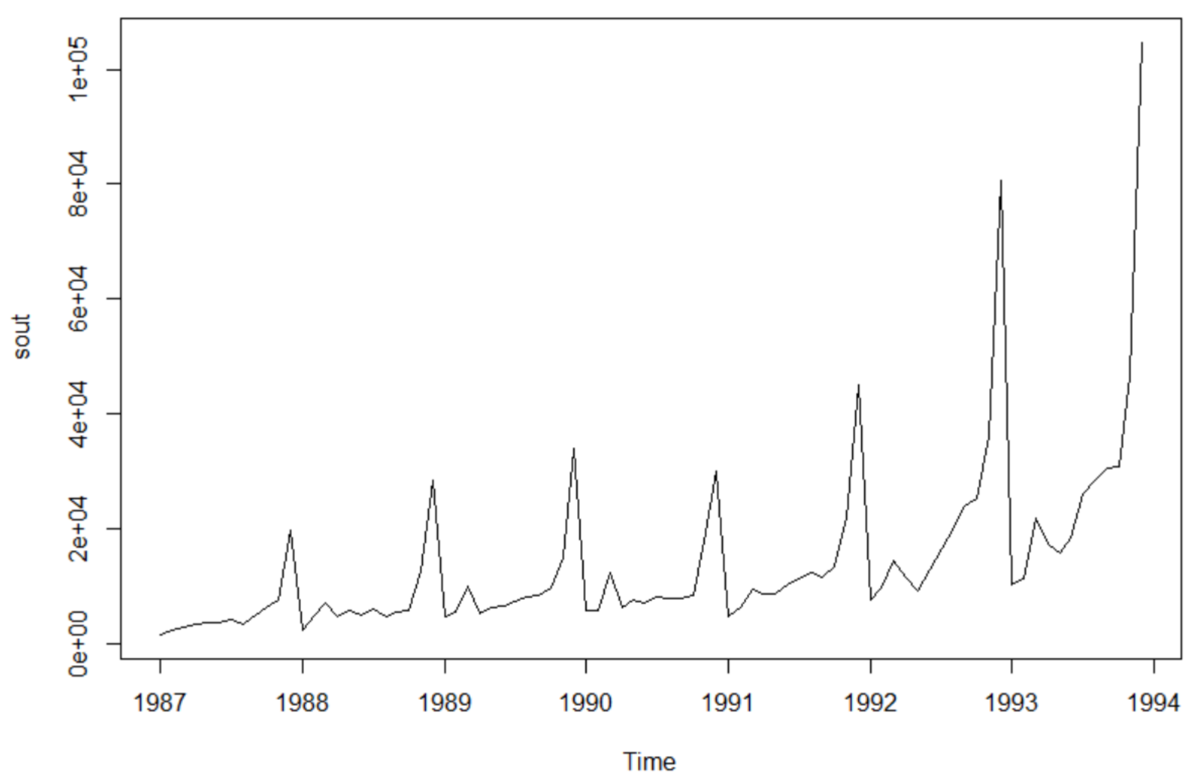

Figure 1: Original sales over time data plot

Looking at the plotted sales data, we can see that there is indeed a cyclical pattern of high sales followed by low sales for each year. However, it doesn't seem to be constant as the trend suggests that there is an exponential trend in sales over the years as opposed to a linear trend. Because the sales increase exponentially, we need to perform a log transformation to further analyze the data. The results of a log transformation of the sales plot is shown in figure 2.

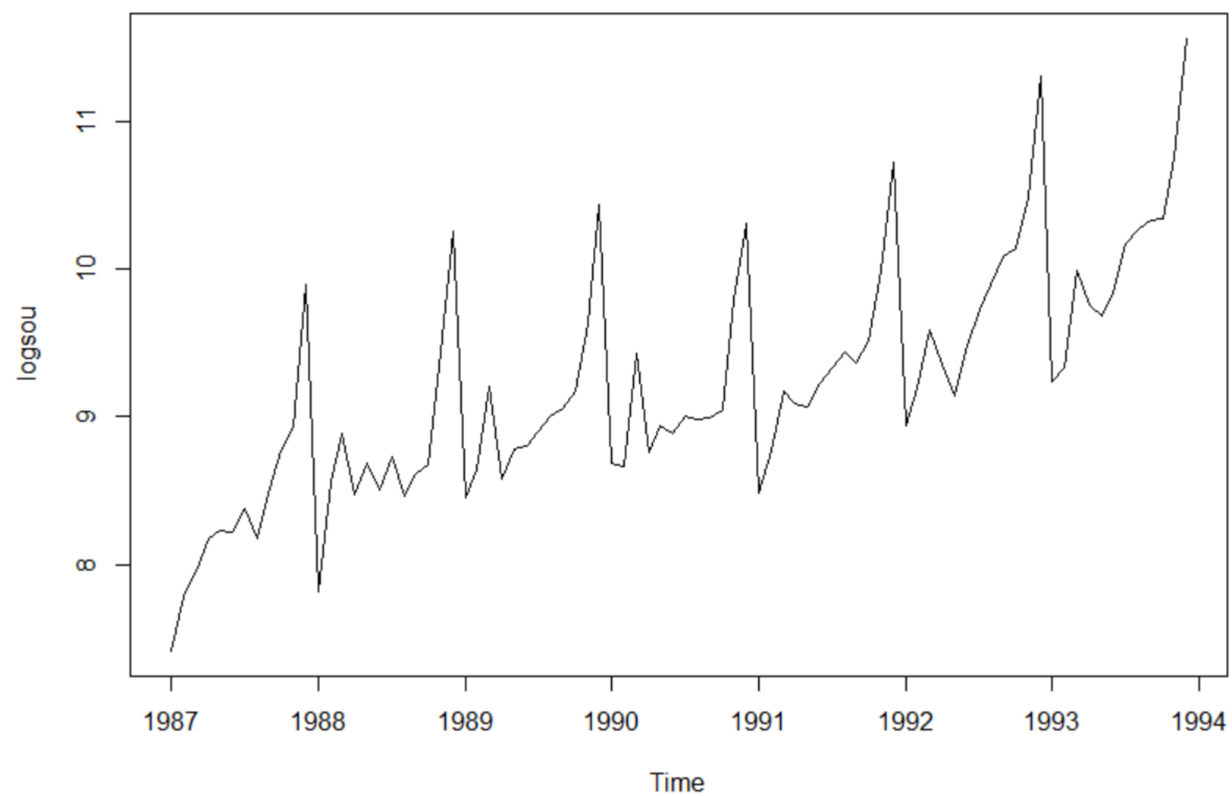

Figure 2: Logarithmically transformed sales over time data plot

After transforming the data, we can see that the exponential trend has disappeared and instead, there is a nearly constant increase. Because we have successfully done a log transformation to the data, we can now apply the three EWMA techniques (Single, Double, and Triple EWMA). The number of the EWMA technique represents the number of factors that are taken into account (level, trend, and season).

1.3 Data Interpretation

Looking at the original and transformed data plots, we can draw some conclusions about the data. It seems like there is a clear pattern to the sales over the years, which seems to suggest that forecasting should work pretty well. As a whole, especially after applying the log transformation, there are few abnormalities from year to year. We can see that there is always a big spike in sales near the end of the year and a drop to a steady level at the beginning of the next year. Over the years, there is a clear increase in sales; the sales spikes are higher and the baseline of sales also rise. It seems like the spike is around 10 times higher than the number of sales at the beginning of the year.

1.4 Model Building

To create models for forecasting the sales for 1993, we will use the data points from the previous six years for model training. To perform the three EWMA analyses, we will use the first 72 log transformed data points to create models, one for each type of EWMA analysis performed (results are given below for each EWMA analysis). Alpha for the values represents level, beta represents trend and gamma represents season.

After fitting to the model, we get the fitted values along with the prediction errors for each model. Using these values we can get a table of predictions alongside their errors which are presented in tables 1, 2, and 3.

Single EWMA:

```
Holt-Winters exponential smoothing without trend and without seasonal component.

Call:
HoltWinters(x = MD, beta = FALSE, gamma = FALSE)

Smoothing parameters:
 alpha: 0.3716002
 beta : FALSE
 gamma: FALSE

Coefficients:
      [,1]
a 10.57251
```

The single EWMA model only accounts for level and not trend nor season. Only the historical data and Holt-Winters exponential smoothing drive the predictions for the model.

Because the alpha value is not too large (alpha = 0.372), more of the past data will be used in the weighted average for forecasting.

			DFMDsfit
[1,]	7.782194	7.417466	0.364728038
[2,]	7.951809	7.552999	0.398810024
[3,]	8.173939	7.701197	0.472742069
[4,]	8.230300	7.876868	0.353431975
[5,]	8.220064	8.008204	0.211860427
[6,]	8.377841	8.086931	0.290910557
[7,]	8.179295	8.195033	-0.015738176
[8,]	8.521548	8.189185	0.332362690
[9,]	8.767715	8.312691	0.455024273
[10,]	8.935982	8.481778	0.454204349
[11,]	9.891223	8.650561	1.240662625
[12,]	7.823970	9.111591	-1.287620938

Table 1: Single EWMA prediction-data table (first 12 rows)

In table 1, we see the comparison of the original data with the fitted model data along with prediction errors in the last column. For the sake of space, we only show the first 12 data rows, but the fitted data comparisons have 71 rows (loses one degree-of-freedom to calculate alpha). For the first 12 data points, some of the errors seem large but others seem to be pretty small.

Double EWMA:

```
Holt-Winters exponential smoothing with trend and without seasonal component.

Call:
HoltWinters(x = MD, gamma = FALSE)

Smoothing parameters:
 alpha: 0.4219034
 beta : 0.1566127
 gamma: FALSE

Coefficients:
        [,1]
a 10.7551734
b  0.1605617
```

The double EWMA model accounts for both level and trend but not season. This model accounts for the same things as the single EWMA model but also includes an overall trend. The alpha value is slightly larger than single EWMA so a little bit less past data will be used in the weighted average for forecasting. Because the beta value is small, the slope of the linear trend is not too large.

```
                                    DFMDdfit
 [1,]  7.951809  8.146922 -0.1951130577
 [2,]  8.173939  8.416439 -0.2425001418
 [3,]  8.230300  8.649940 -0.4196401363
 [4,]  8.220064  8.780977 -0.5609133681
 [5,]  8.377841  8.815348 -0.4375066875
 [6,]  8.179295  8.872876 -0.6935810955
 [7,]  8.521548  8.776537 -0.2549892923
 [8,]  8.767715  8.848393 -0.0806772985
 [9,]  8.935982  8.988460 -0.0524777903
[10,]  9.891223  9.136958  0.7542652566
[11,]  7.823970  9.675662 -1.8516915629
[12,]  8.556075  8.992552 -0.4364764711
```

Table 2: Double EWMA prediction-data table (first 12 rows)

In table 2, we see the comparison of the original data with the fitted model data along with prediction errors in the last column. For the sake of space, we only show the first 12 data rows, but the fitted data comparisons have 70 rows (loses 2 degrees-of-freedom to calculate alpha and beta). For the first 12 data points, some of the errors seem to be a bit larger but others seem smaller so there is no clear conclusion we can draw.

Triple EWMA:

```
Holt-Winters exponential smoothing with trend and additive seasonal component.

Call:
HoltWinters(x = MD)

Smoothing parameters:
 alpha: 0.4379168
 beta : 0
 gamma: 1

Coefficients:
            [,1]
a    10.139495294
b     0.029963187
s1   -0.646470510
s2   -0.416049412
s3    0.005888226
s4   -0.228125708
s5   -0.363832549
s6   -0.101330406
s7    0.017796279
s8    0.051425914
s9    0.038028700
s10   0.009006250
s11   0.386200208
s12   1.159267545
```

The triple EWMA model accounts for level, trend, and season. In addition to the overall trend and historical data, this model also includes seasonal factors derived from past seasonal data. The alpha value is around the same as the double EWMA so a similar amount (moderate amount) of past data will be accounted for in the modeling. Because the beta value is 0, there is no linear trend in this model. The gamma value of 1 is relatively large, so the model depends on the most recent yearly cycle greatly to account for seasonality.

			DFMDtfit
[1,]	7.823970	7.587087	2.368828e-01
[2,]	8.556075	8.426301	1.297741e-01
[3,]	8.885322	8.826618	5.870416e-02
[4,]	8.477627	8.474833	2.793342e-03
[5,]	8.682857	8.694191	-1.133446e-02
[6,]	8.507414	8.507357	5.652839e-05
[7,]	8.728931	8.764110	-3.517918e-02
[8,]	8.466352	8.530939	-6.458651e-02
[9,]	8.611854	8.803730	-1.918757e-01
[10,]	8.671647	8.944285	-2.726383e-01
[11,]	9.441458	8.991612	4.498462e-01
[12,]	10.259122	10.142982	1.161403e-01

Table 3: Triple EWMA prediction-data table (first 12 rows)

In table 3, we see the comparison of the original data with the fitted model data along with prediction errors in the last column. For the sake of space, we only show the first 12 data rows, but the fitted data comparisons have 60 rows (loses 12 degrees-of-freedom because the first year is used to predict the second year). For the first 12 data points, it seems like this model has errors that are much smaller than those of the single EWMA and the double EWMA.

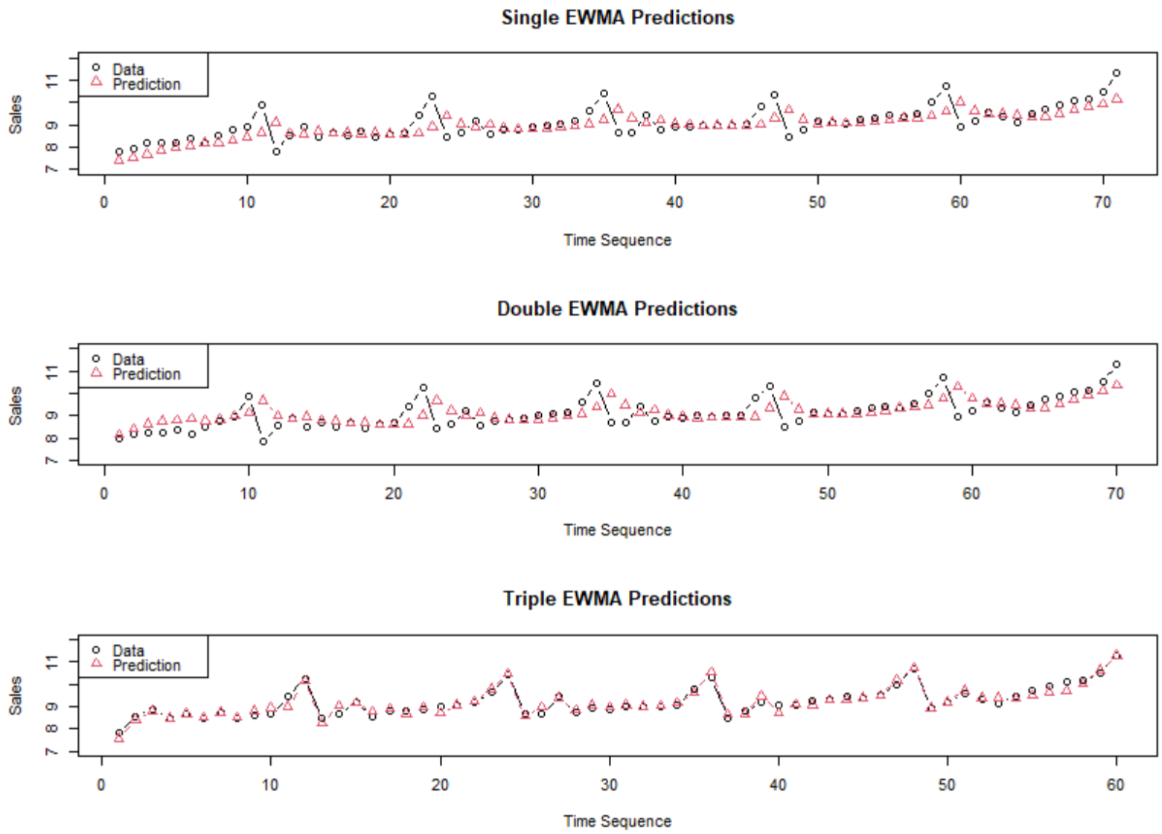

Figure 3: Model Prediction plots for Single, Double, and Triple EWMA

Looking at the prediction plots for the three EWMA models in figure 3, we can see that the Triple EWMA predictions most closely match the actual data. It looks like the triple EWMA model performs the best because considering the season factor, along with the trend and historical data, is important for matching the data used.

Both single and double EWMA have predictions that are a bit far off from the actual data points. While the single EWMA predictions seem to match closely with some intermediate points, it completely misses the peaks and the subsequent drops. For example, the 11th and 12th points have prediction errors of 1.24 and -1.29 respectively which are quite high. In contrast, the prediction errors from the 1st point to the 10th all have prediction errors within +/- 0.5. The double EWMA has the same problems as the single EWMA but has some prediction errors that are slightly more far off. For example, the 11th data point has a prediction error of -1.85 and a few of the points between the 1st and the 10th have prediction errors greater than +/- 0.5. On the other hand, the triple EWMA predictions seem to be very close to the actual data. For the high and low points that single and double EWMA struggle most with, triple EWMA has prediction errors within 4.8% and

1.1% for the 11th and 12th point respectively, which is much smaller than the errors for single and double.

Looking at table 4 for reference, we can see that regarding prediction errors, triple EWMA is the closest matching to actual data followed by single EWMA and double EWMA.

Model	Single EWMA	Double EWMA	Triple EWMA
SSE	20.894	23.329	1.764

Table 4: Sum of Squared Errors for the three models

Model Metric	Single EWMA	Double EWMA	Triple EWMA
MFE	0.120	-0.044	-0.003
MAD	0.397	0.412	0.132
MAPE	4.301	4.531	1.447
MSE	0.294	0.333	0.029

Table 5: Four-Metrics Evaluation Table

- MFE (Mean Forecast Error)
  - The closer to zero the better because the forecasting predictions will be more accurate.

$$MFE = \frac{\sum Y_t - \widehat{Y}_t}{n}$$

- MAD (Mean Absolute Deviation)
  - This metric measures the accuracy of the fitted time series values in the same units as the data. The smaller the value the better.

$$MAD = \frac{\sum |Y_t - \widehat{Y}_t|}{n}$$

- MAPE (Mean Absolute Percentage Error)
  - This metric measures the accuracy of the fitted time series values in a percentage. The smaller value the better.

$$MAPE = \frac{\sum\left|\frac{Y_t - \widehat{Y}_t}{Y_t}\right|}{n}$$

- MSE (Mean Squared Error)
  - A measure of deviation that is more sensitive to large fitting errors and can easily be compared across models. The smaller the value the better.

$$MSE = \frac{\sum(Y_t - \widehat{Y}_t)^2}{n}$$

In table 5, we can see the results of the four-metrics evaluation. The four metrics can help us see the performance of each of the EWMAs. For MFE, the closest value to zero is triple EWMA with -0.003 followed by double EWMA with -0.044 and single EWMA with 0.120. Compared to triple and double EWMA, which have small negative values, the MFE for single EWMA is relatively large and also positive. For MAD, the smallest value is triple EWMA with 0.132 followed by single EWMA with 0.397 and double EWMA with 0.412. For MAPE, the smallest value is triple EWMA with 1.447 followed by single EWMA with 4.301 and double EWMA with 4.531. For MSE, the smallest value is triple EWMA with 0.029 followed by single EWMA with 0.294 and double EWMA with 0.333. Overall, triple EWMA performs best according to all 4 metrics with single EWMA performing second best according to 3 metrics and double EWMA performing second with only 1 metric..

For triple EWMA, the MFE is a very small value close to zero which means that the forecasting predictions are very accurate. The MAD is 0.132 which is in the same units as the data, meaning that the average deviation from the actual data is 0.132 from a given point (i.e. data point 12 with actual value of 10.259). The MAPE is similar to MAD but measures in percentages, so a value of 1.447 means the average prediction error is 1.447% of the actual data. The low MSE value of 0.029 means that there are no large fitting errors for the triple EWMA model. Because MSE can be easily compared between models, single EWMA has an MSE around 10 times larger than triple EWMA and 13.3% smaller than double EWMA.

1.4 Model Forecasting

After performing our model forecasting, we get the following results in figure 4 for single, double and triple EWMA showing the upper bound, lower bound, and the model fitted values. In figure 5, we show a side-by-side comparison of actual values and forecasted values along with error for each of the 3 EWMA methods.

	fit	upr	lwr
Jan 1993	10.57251	11.61697	9.528043
Feb 1993	10.57251	11.68675	9.458261
Mar 1993	10.57251	11.75242	9.392599
Apr 1993	10.57251	11.81461	9.330403
May 1993	10.57251	11.87384	9.271176
Jun 1993	10.57251	11.93049	9.214530
Jul 1993	10.57251	11.98486	9.160154
Aug 1993	10.57251	12.03722	9.107795
Sep 1993	10.57251	12.08777	9.057244
Oct 1993	10.57251	12.13669	9.008326
Nov 1993	10.57251	12.18412	8.960893
Dec 1993	10.57251	12.23020	8.914816

	fit	upr	lwr
Jan 1993	10.91574	12.05205	9.779420
Feb 1993	11.07630	12.34069	9.811908
Mar 1993	11.23686	12.64932	9.824396
Apr 1993	11.39742	12.97590	9.818940
May 1993	11.55798	13.31855	9.797414
Jun 1993	11.71854	13.67567	9.761421
Jul 1993	11.87911	14.04592	9.712296
Aug 1993	12.03967	14.42820	9.651138
Sep 1993	12.20023	14.82161	9.578849
Oct 1993	12.36079	15.22541	9.496175
Nov 1993	12.52135	15.63897	9.403739
Dec 1993	12.68191	16.06176	9.302065

	fit	upr	lwr
Jan 1993	9.522988	9.861859	9.184117
Feb 1993	9.783372	10.153312	9.413433
Mar 1993	10.235273	10.633867	9.836679
Apr 1993	10.031222	10.456544	9.605900
May 1993	9.925479	10.375946	9.475012
Jun 1993	10.217944	10.692225	9.743663
Jul 1993	10.367034	10.863989	9.870079
Aug 1993	10.430627	10.949265	9.911988
Sep 1993	10.447193	10.986644	9.907742
Oct 1993	10.448133	11.007624	9.888643
Nov 1993	10.855291	11.434126	10.276455
Dec 1993	11.658321	12.255877	11.060765

Figure 4: Forecast results for single, double and triple EWMA (left to right)

			DFTDsfore			DFTDdfore			DFTDtfore
[1,]	9.234373	10.57251	-1.3381346	9.234373	10.91574	-1.681362	9.234373	9.522988	-0.28861472
[2,]	9.329623	10.57251	-1.2428852	9.329623	11.07630	-1.746674	9.329623	9.783372	-0.45374953
[3,]	9.990896	10.57251	-0.5816122	9.990896	11.23686	-1.245963	9.990896	10.235273	-0.24437740
[4,]	9.761770	10.57251	-0.8107377	9.761770	11.39742	-1.635650	9.761770	10.031222	-0.26945216
[5,]	9.680206	10.57251	-0.8923020	9.680206	11.55798	-1.877776	9.680206	9.925479	-0.24527282
[6,]	9.830999	10.57251	-0.7415088	9.830999	11.71854	-1.887545	9.830999	10.217944	-0.38694490
[7,]	10.171801	10.57251	-0.4007065	10.171801	11.87911	-1.707304	10.171801	10.367034	-0.19523249
[8,]	10.260691	10.57251	-0.3118173	10.260691	12.03967	-1.778977	10.260691	10.430627	-0.16993615
[9,]	10.325659	10.57251	-0.2468486	10.325659	12.20023	-1.874570	10.325659	10.447193	-0.12153336
[10,]	10.335962	10.57251	-0.2365456	10.335962	12.36079	-2.024829	10.335962	10.448133	-0.11217116
[11,]	10.750093	10.57251	0.1775854	10.750093	12.52135	-1.771259	10.750093	10.855291	-0.10519725
[12,]	11.558479	10.57251	0.9859708	11.558479	12.68191	-1.123436	11.558479	11.658321	-0.09984241

Figure 5: Forecast-data table with errors for single, double and triple EWMA (left to right)

In figure 5, we have combined the data (first column), with the predictions (second column) along with their errors (third column) for single, double, and triple EWMA models from left to right. We can see immediately that double EWMA has the largest forecasting errors with all the predictions having between 10% and 20% forecasting errors. For single EWMA, the largest forecasting error is 14.5% which is much larger than triple EWMA's largest forecasting error of 4.9%. In general, the errors across all the EWMA methods are negative, meaning the predictions usually are greater than the actual values.

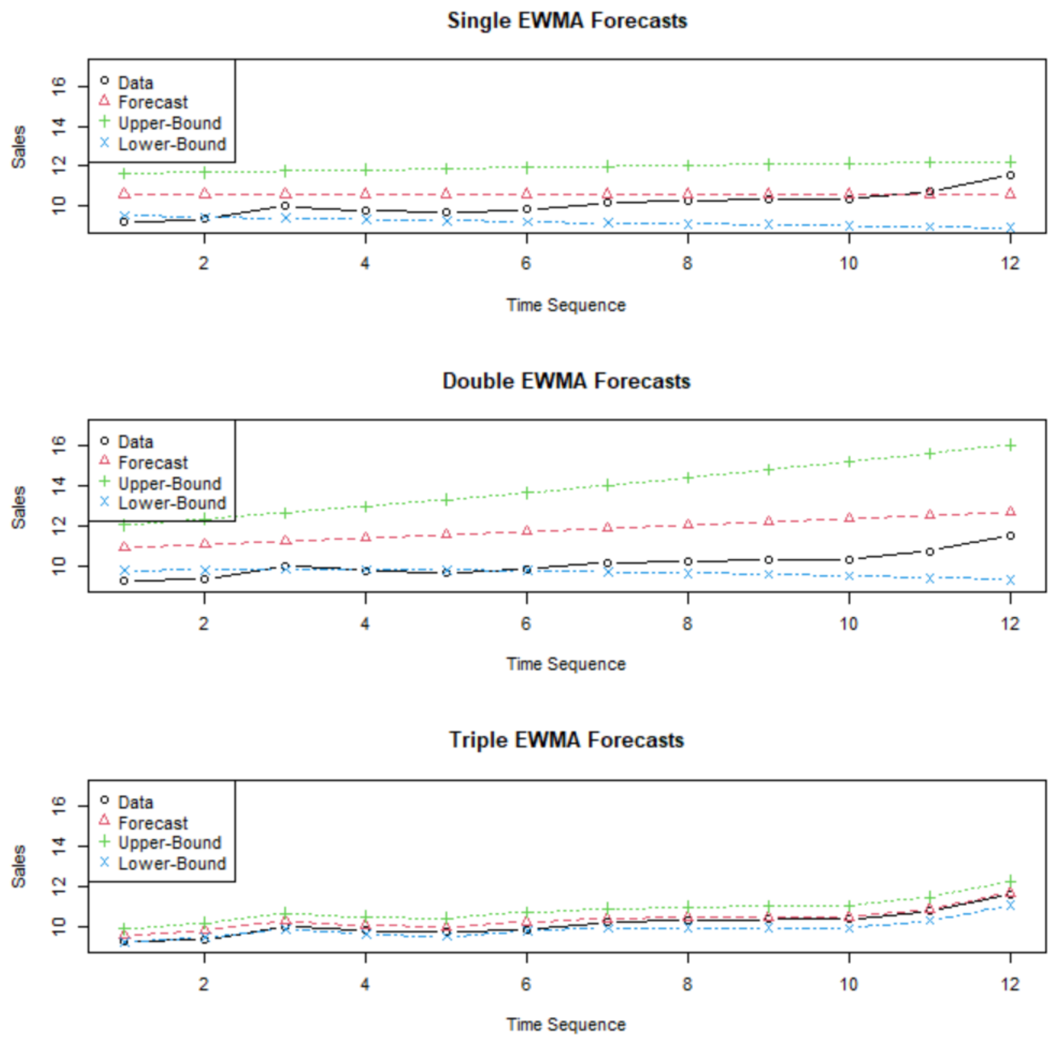

Figure 6: Forecasting plots for the three EWMA methods for the predicted year

In figure 6, we can see the lower-bound, upper-bound, forecast points, and data for the three EWMA models during the forecasting year. We will focus on the forecasting results because we have already commented on the model fitting.

For the single EWMA plot, we can see that the forecasting seems to capture the overall cyclic trend of data, but seems to miss the nuances for each month. We can also see that 10 out of the 12 forecasting points are within the forecasting interval, which means 2 points fell completely outside of the model's expected deviation. For the double EWMA plot, we can see that the forecasting seems to be very far off from the actual data, more than the single EWMA's forecasting. It seems like the forecasting overshoots the actual data by a lot when accounting for the linear upwards trend. We can also observe that the forecasting interval is wider than single EWMA, but misses more points (only 8 out of 12 inside interval). For the triple EWMA plot, we can see that the forecasting seems to match the values very closely, much more closely than both double and single EWMA. It seems to account for both the cyclic and slight upward trends in data. We can see that the forecasting interval is much narrower than both double and single EWMA but includes all 12 of the points inside the interval.

Model / **Metric**	**Single EWMA**	**Double EWMA**	**Triple EWMA**
MFE	-0.470	-1.696	-0.224
MAD	0.664	1.696	0.224
MAPE	6.712	16.901	2.279
MSE	0.588	2.941	0.062

Table 6: Four Metrics for Forecasting Results

Looking at table 6, we can see that once again, triple EWMA has much smaller values than the other two methods. For example, looking at MAPE, triple EWMA's percent error is 2.3% which is almost 3 times lower than single EWMA's 6.7% and around 7 times lower than double EWMA's 16.9%. We can clearly see that tripe EWMA provides the highest quality forecasting results.

1.5 Comparison of Prediction and Forecasting

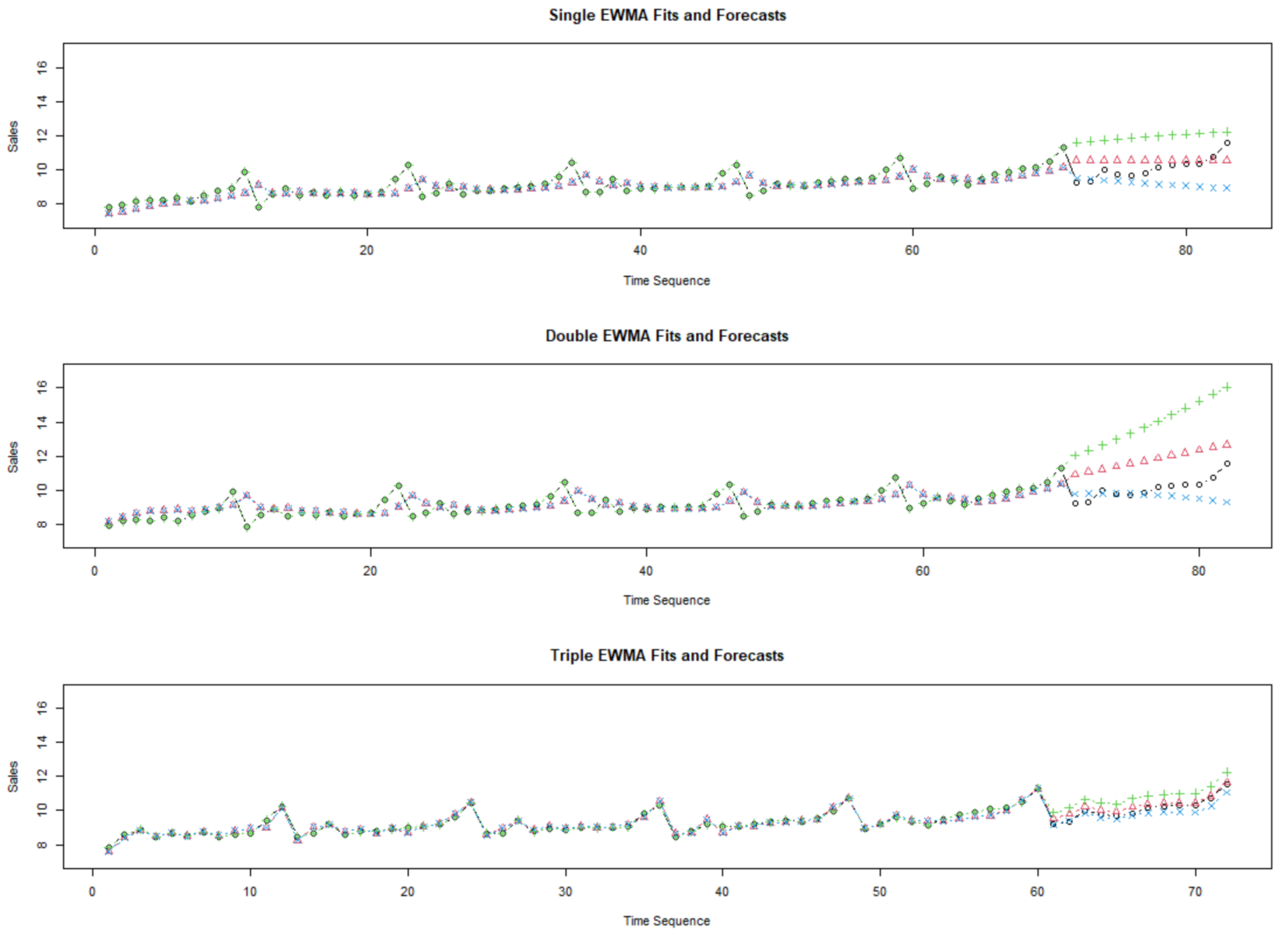

Figure 7: Fitted data and forecasting plots for the three EWMA models

Looking at figure 7, we can see both the fitted data and forecast data for the three EWMA methods. Comparing the quality of the fitted data and the forecasting, we can conclude that in general, for all three methods, the fitted data matches more closely than the forecasted data. This is to be expected because it is easier to build a model around pre-existing data than it is to predict new data points. As a whole, triple EWMA is the better model for both fitted data and forecasted data, followed by single EWMA then double EWMA.

Model Metric	Model Fitting Metrics			Forecasting Metrics		
	Single EWMA	Double EWMA	Triple EWMA	Single EWMA	Double EWMA	Triple EWMA
MFE	0.120	-0.044	-0.003	-0.470	-1.696	-0.224
MAD	0.397	0.412	0.132	0.664	1.696	0.224
MAPE	4.301	4.531	1.447	6.712	16.901	2.279
MSE	0.294	0.333	0.029	0.588	2.941	0.062

Table 7: Comparison of model fitting metrics and forecasting metrics results

In table 7, we see the metrics results for both fitted data and prediction data using single, double, and triple EWMA. Across the board, we can see that the metrics for fitted data are lower than the forecasting metrics. This result reflects what we discussed previously about figure 7. This reflects the fact that our models are more accurate for fitting data than for predicting data. Using the MAPE metric as an example, we see that the single EWMA metric increases by around 1.5 times, the double EWMA metric increases by almost 4 times, and the triple EWMA metric increases by around 1.5 times. Still, we see that in all metrics, triple EWMA is still the best for both model fitting and forecasting.

This project was very helpful for understanding certain elements of EWMA and time-series modeling. Comparing the forecasting results with the model fitting helps me understand how EWMA methods have a dual usage. Also, analyzing the plots and the metrics helps me see how we can deduce certain information from performing a time-series analysis.