Forecasting assignment

Carlos Montenegro Vela

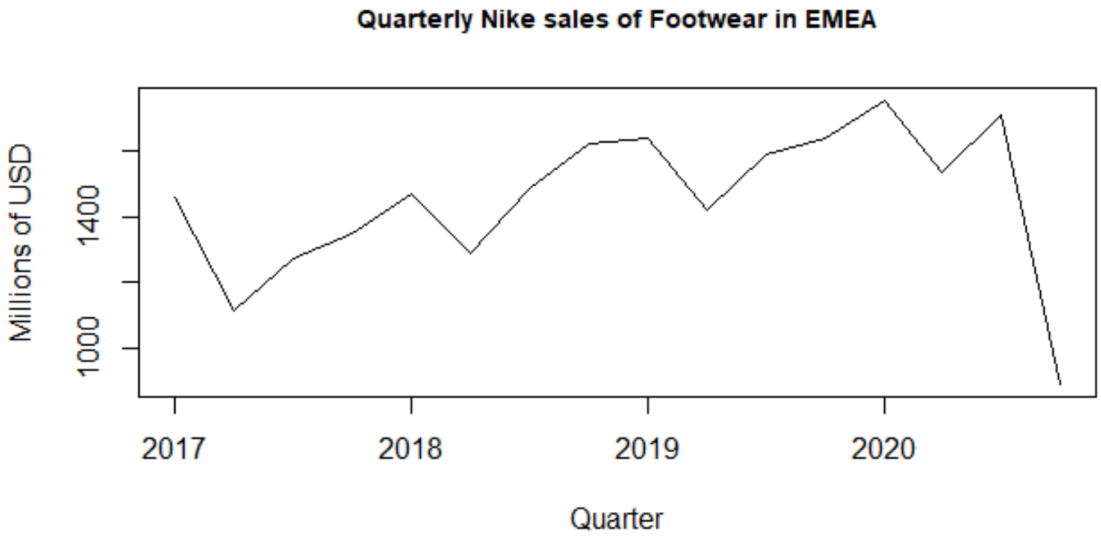
September 2020

Introduction

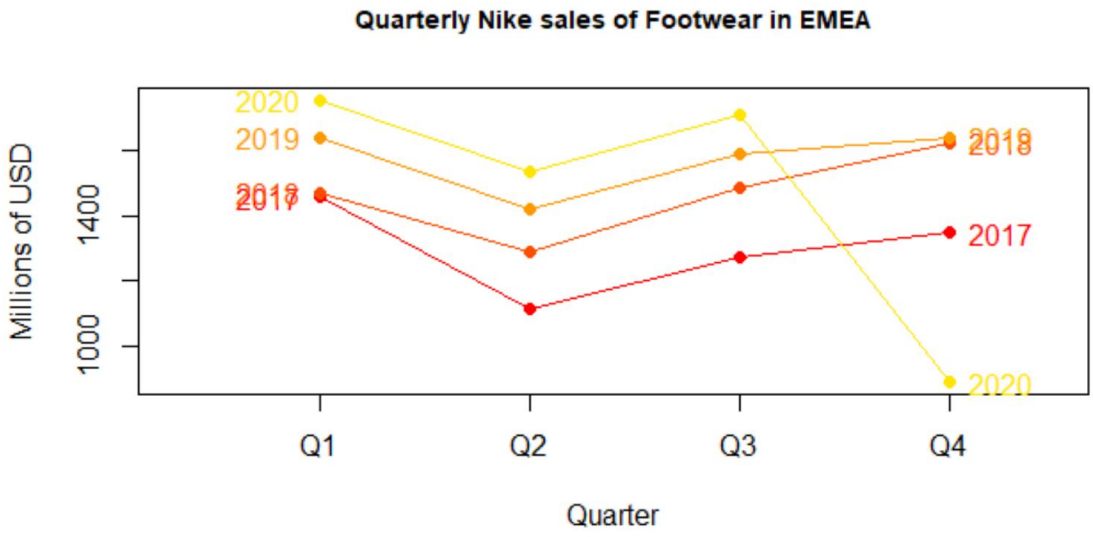
The aim of this paper is to find a recent and relevant time series to forecast, using the techniques that have been discussed during the lectures on Forecasting dictated by Professor Vandenbossche. In my analysis, I set up a selected forecasting process, taking data considerations and implementations issues into account. Finally, I describe my approach and motivate my choices.

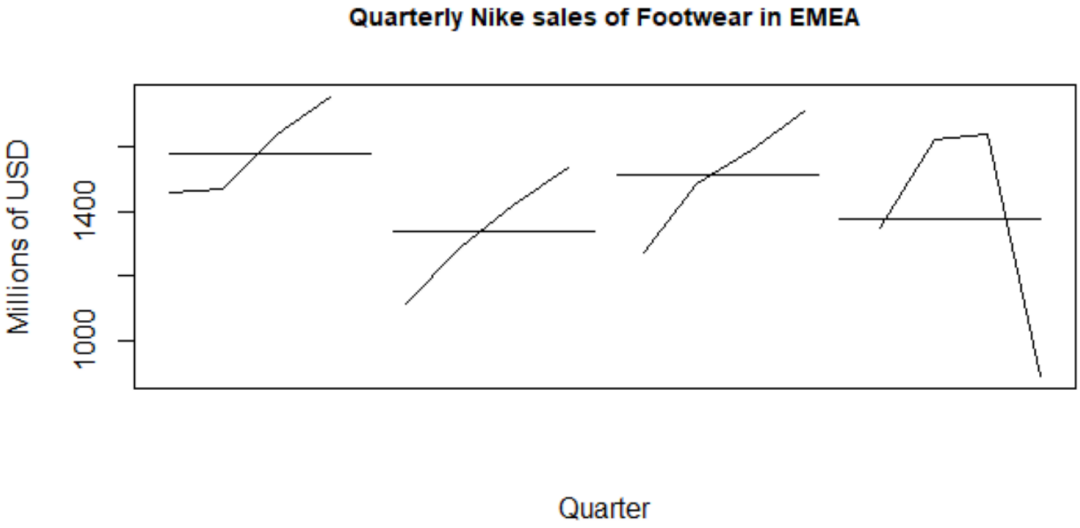
Times Series: Quarterly Nike sales of footwear in 2016 - 2020

Addressing the problem of originality of the data and the way I came up with the idea of using this data was because I was looking to explore the applications of forecasting in the operations of a company present in Europe. I discovered that there was plenty of recent information available in the financial statements of this company, therefore I tried to manually obtain the values of the sales of the last 4 years by looking at the financial statement of each quarter. The reason I decided to choose 4 years instead of more years was because the information from the financial statement of the previous quarters had different categories and therefore the information of the sales was not comparable. Also, 4 years resulted on a optimal number of years because we are trying to analyze the future based on not so old information of the operations of the company.

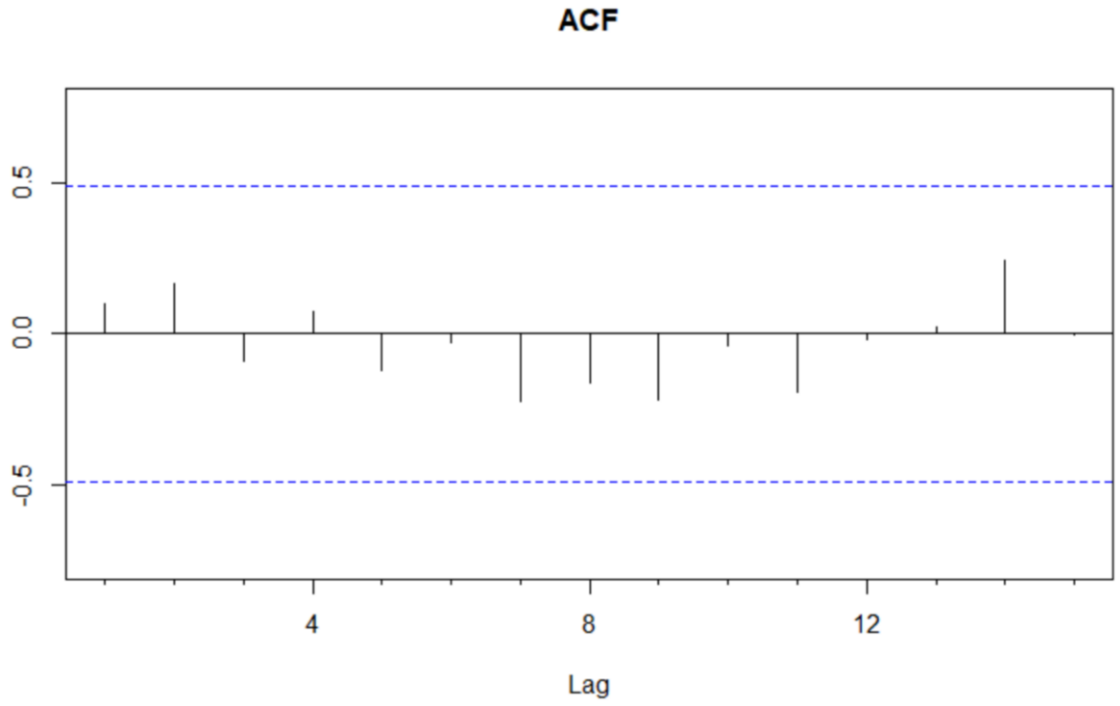


We must point out that the data being analyzed is from the fiscal years of the enterprise and that for example the fiscal year of 2020 started the 1st of June of 2019 and finished the 31st of August of 2019. In specific, the data contains information of Nike sales from the first quarter of 2016 to last quarter of 2020. As we can observe the sales have been increasing and there is not much increase in the variation of the sales along the years. But it has experienced one significant drop: at the end of the fiscal year of 2020.

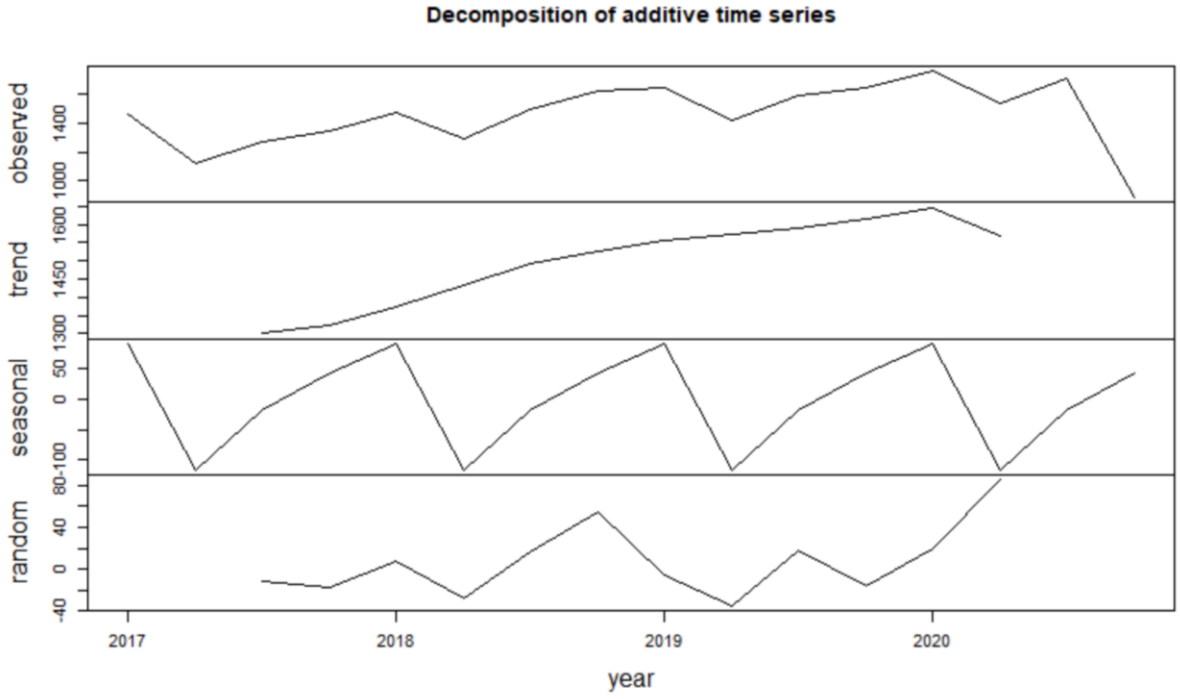




From the graphs above we observe that excluding from the forth quarter of 2020 there is a clear increase in sales along the years and in each of the quarters. There are some periods where the yearly increase in sales is not much but the only period where we observe a decrease in the forth quarter of 2020 (or 31-05-2020) which can be explained by the store closures and supply chains disruptions caused by the current pandemic.



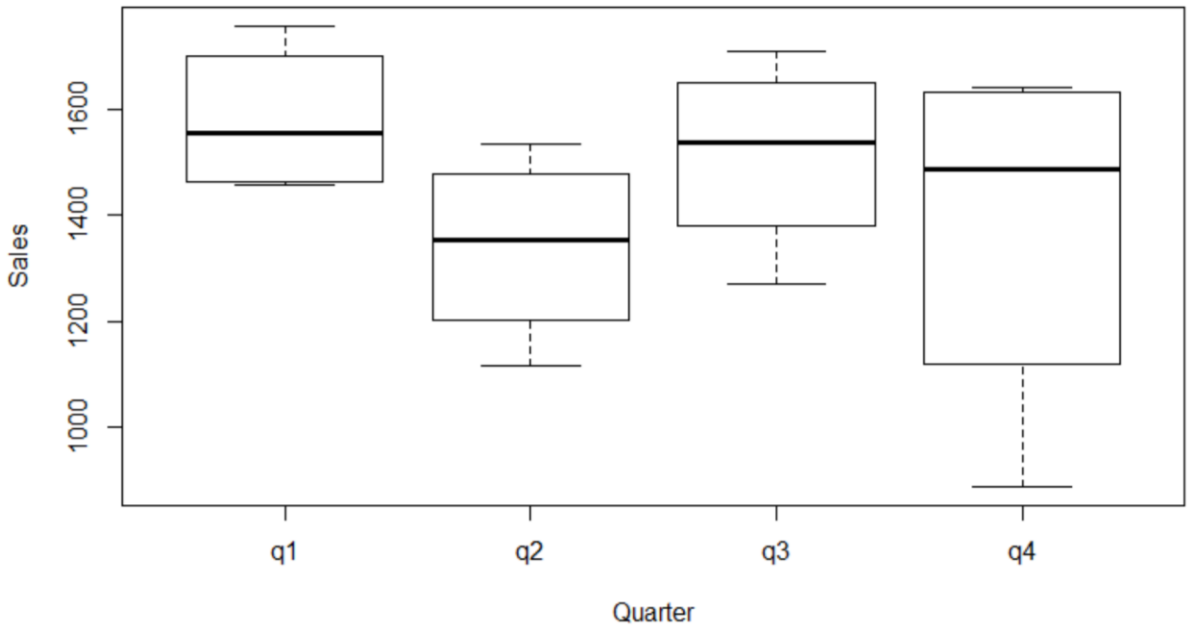
From this graph we can determine that there is no effect of the past lags in the forecast of future values. In the next graph we can observe the seasonal an trend components of the time series, the trend component is increasing over time. We should be careful with the random component in the last period as we need to recognize that this is an external effect, the pandemic.



We also show some summary statistics and the boxplot by month.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
887    1334    1480    1453    1629    1758
```

Looking at the boxplot we can observe that the mean value of each month is a little different but not much, with the exception of the q4 (again, due to the pandemic).



For this data series, it does not seem that we will need to stabilize its variance but we will still use the Bow-Cox transformation to get the optimal value of Lambda:

```
BoxCox.lambda(footwear)
[1] -0.9999242
```

The Box-Cox gives us a value of Lambda close to negative 1, and so I will not apply any transformation in the time series.

Forecast modeling

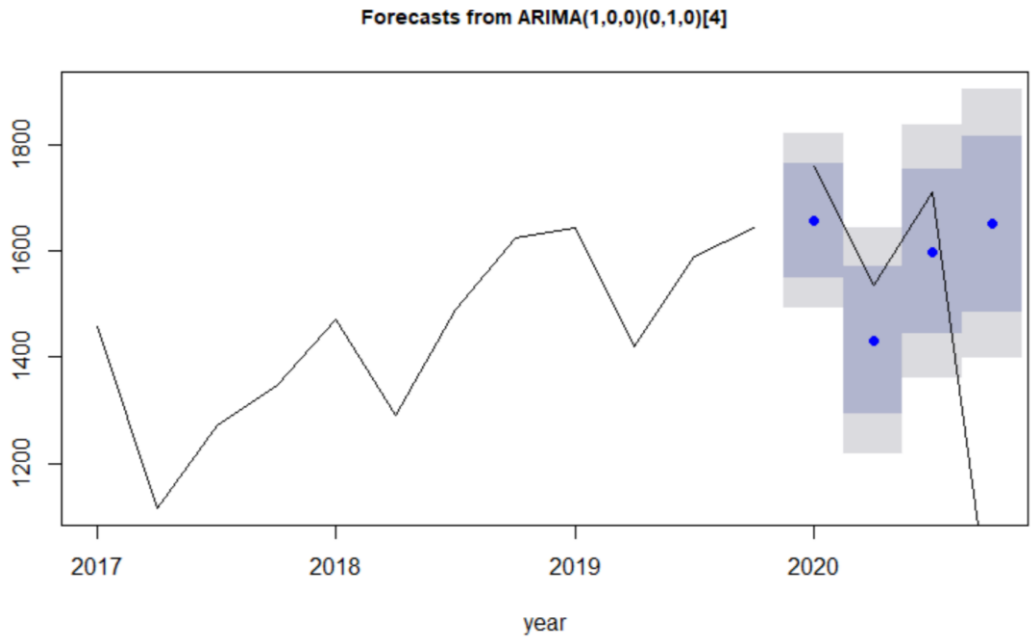
I will forecast the sales using two models: ETS and ARIMA. Later I will compare the performance of each of the different methods by dividing the data in train and test. The train dataset contains information of the sales of footwear of Nike from the first quarter of the fiscal year 2017 to the last quarter of the fiscal year 2019 and the test dataset contains information of the sales of footwear of Nike from the first quarter of the fiscal year 2020 to the last quarter of the fiscal year of 2020.

Forecast: ARIMA

Using the auto.arima function we find the best model based on the AIC criteria.

```
Z test of coefficients :

    Estimate Std. Error z value  Pr(>|z|)
ar1  0.80959    0.14641  5.5297 3.208e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

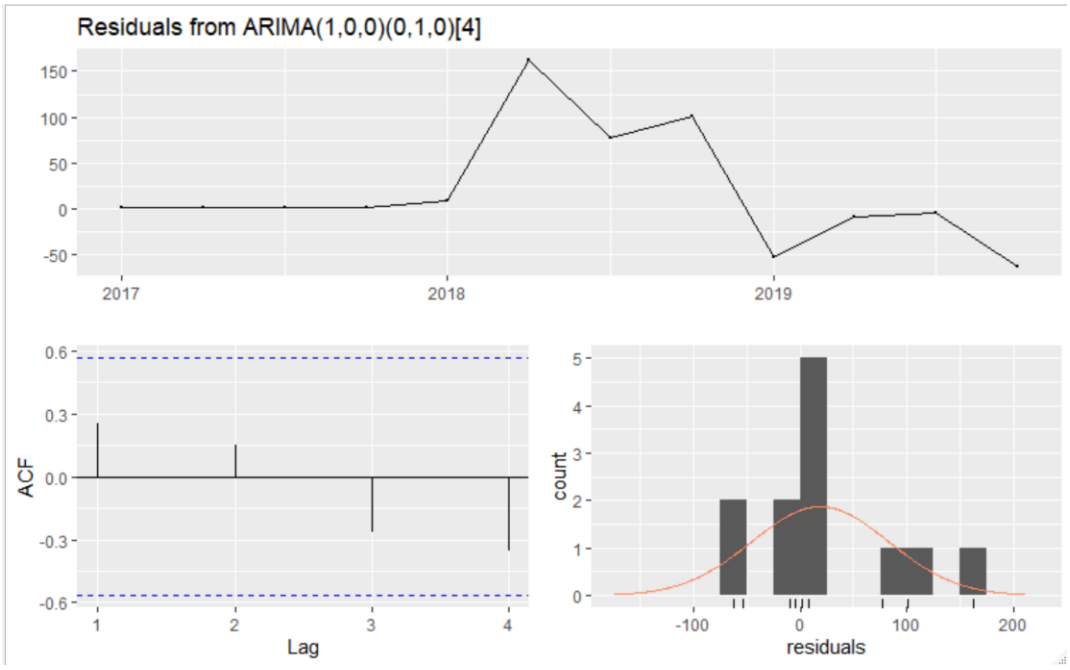


This generated model is an ARIMA (1,0,0)(0,1,0). This model has 1 estimated parameter and is significant. Regarding the accuracy, we have the following results.

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	18.63512	64.22274	40.3169	1.409472	2.746092	0.2929475	0.25286135	NA
Test set	-111.16358	392.83816	270.7027	-16.728052	26.323393	1.9669591	-0.09076737	0.8933524

Checking mainly the MASE but also the other indicators, the model shows a better performance in the training set which is expected as it was build using this information.

Apart of this, we can observe the residuals in the following graph:

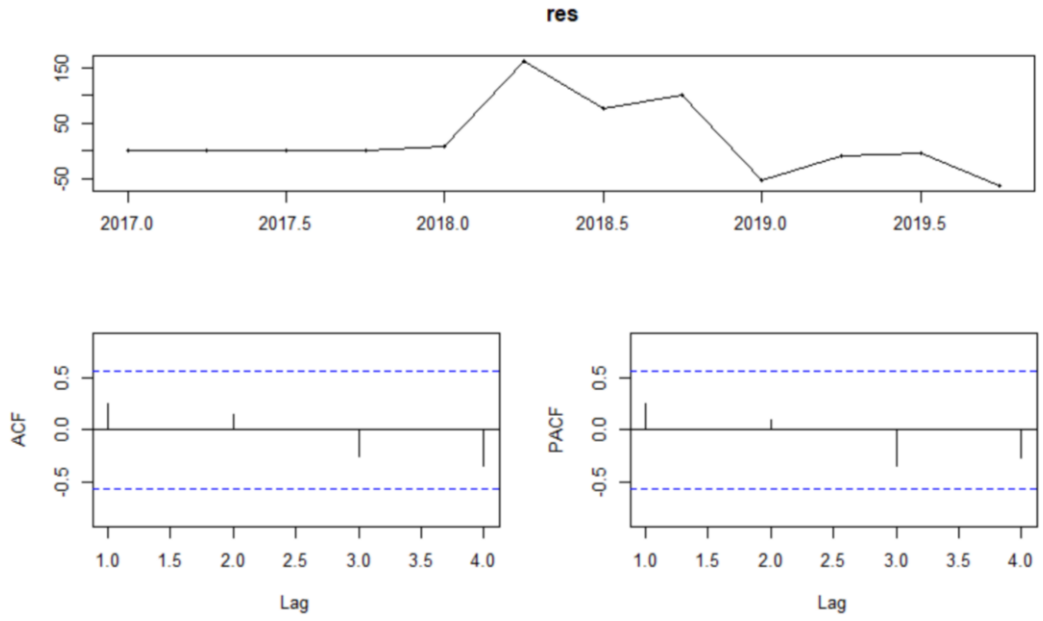



```
        Ljung-Box test

data:  Residuals from ARIMA(1,0,0)(0,1,0)[4]
Q* = 5.3053, df = 3, p-value = 0.1508

Model df: 1.   Total lags used: 4
```

We can suggest that the residuals are not around zero and they do not follow a normal distribution, there are peaks that do not seem to follow any pattern.

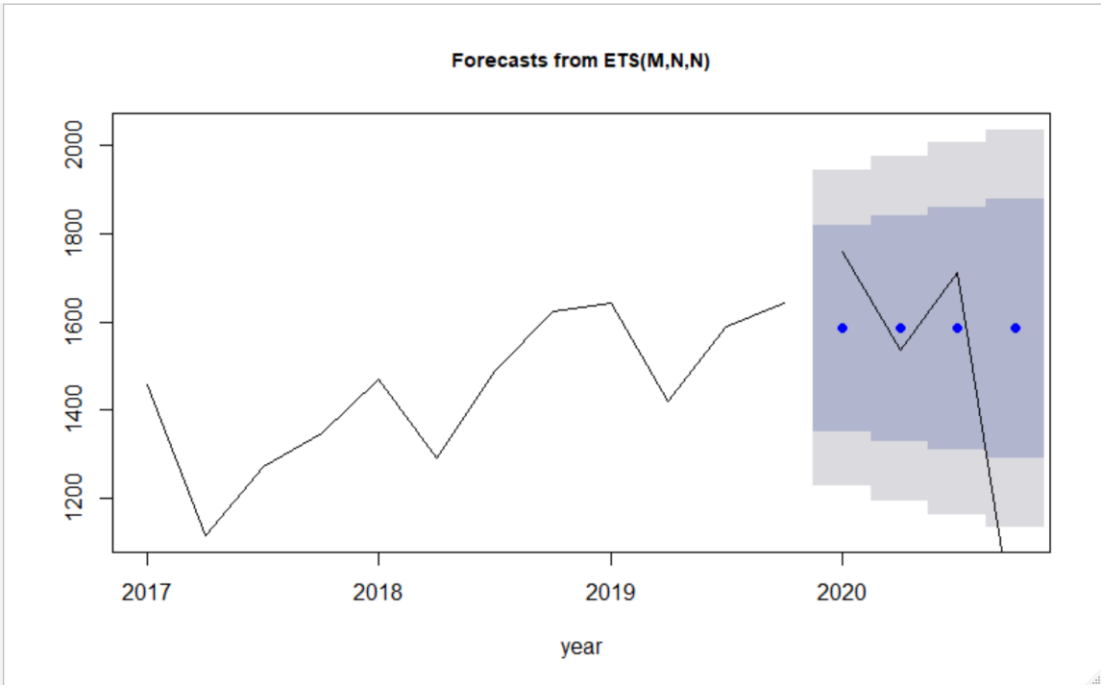


The ACF of the residuals suggests that we do not have more information to consider from previous periods that might affect the forecast. On the other hand, the Ljung Box test below indicates that we can fail to reject the null hypothesis of white noise in the residuals since the p values are higher than 0.05.

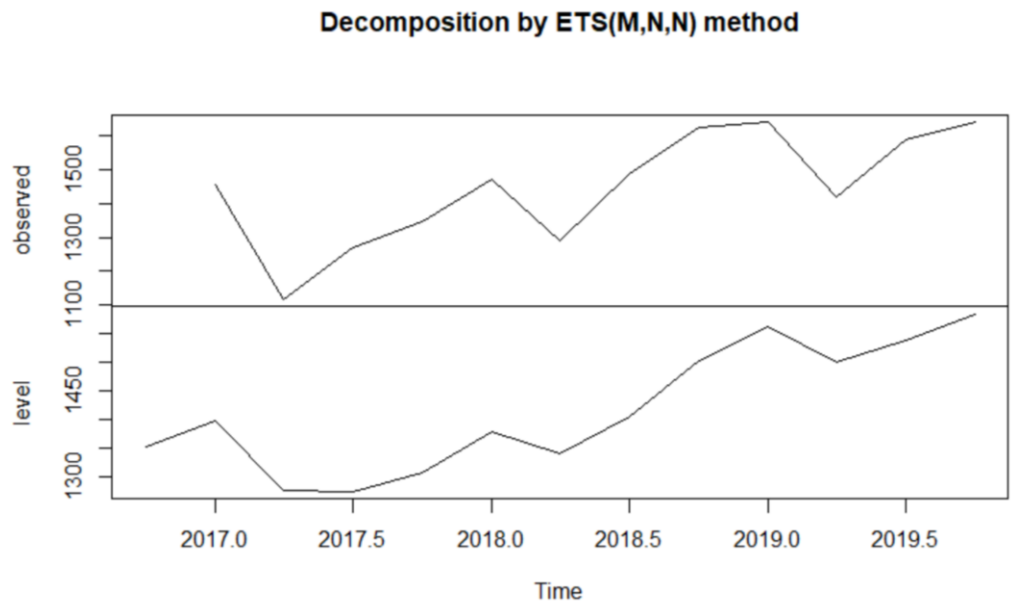
```
lags statistic df   p-value
1 0.9765208  0 0.0000000
2 1.3681300  1 0.2421334
3 2.6559864  2 0.2650086
4 5.3053211  3 0.1507575
5 6.8451418  4 0.1443016
6 8.8792462  5 0.1139786
7 8.9205772  6 0.1780943
8 9.1764511  7 0.2402347
9 9.3210468  8 0.3159409
10 9.4573338  9 0.3961724
11 9.6179731 10 0.4746242
```

Forecast: ETS

We use the et function from R and we can see below the results from this exercise. This function returns the best ETS method based on AIC, MSE, and others. The resulting method is a "M,N,N" or "Multiplicative error, no trend, no seasonal component". The parameter associated to the model is "alpha = 0.4381". Which means that the model is just considering the last observations.



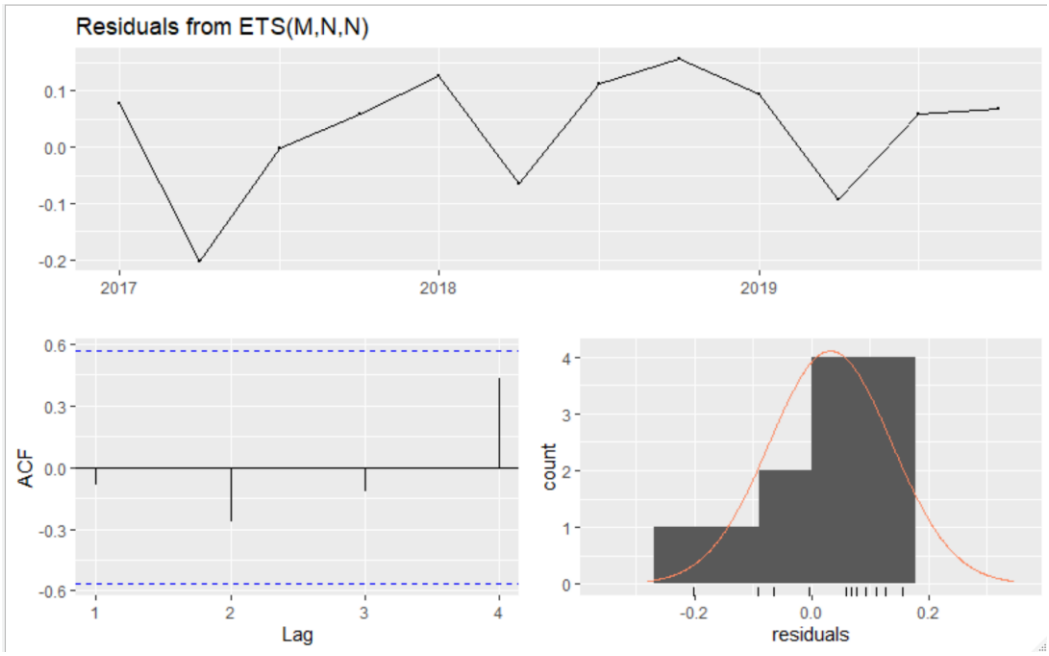
We observe that the level increases along time.



Checking the accuracy, we check again that the MASE and other indicators have a better performance in the train set than in the test set as expected.

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	44.11024	147.5560	130.4825	2.120478	9.216382	0.9481019	-0.08482515	NA
Test set	-111.52261	365.7127	261.5000	-16.134381	24.764331	1.9000908	-0.21952089	0.8161035

On the residuals we have the following results. Even thought the residuals are around 0 they do not strictly follow a normal distribution as we can observe that they create a skewed type of distribution.



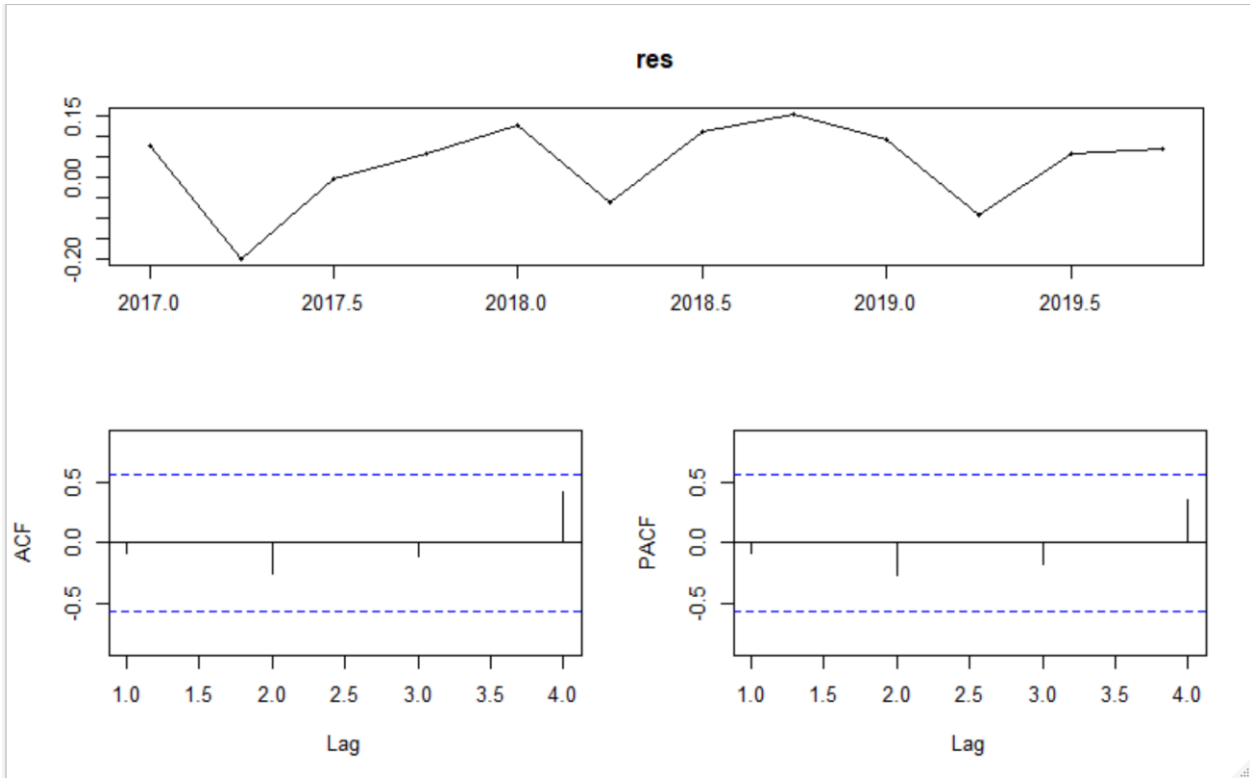
The Ljung-Box tests suggest that we can fail to reject the null hypothesis of white noise in the residuals since we have a p-value higher than 0.05. Which means that the model is correctly capturing g most of the information in the data generation process.

```
        Ljung-Box test

data:  Residuals from ETS(M,N,N)
Q* = 7.7046, df = 3, p-value = 0.05253

Model df: 2.   Total lags used: 5
```

As we observed with the previous method, we do not observe significant spikes in specific lags, meaning that there is no significant information from the past that could affect in the forecasts.




```
lags  statistic df    p-value
   1  0.1053052  0         NA
   2  1.2675620  0 0.00000000
   3  1.5208742  1 0.21748725
   4  5.4015175  2 0.06715454
   5  7.7046162  3 0.05252763
   6  9.6823219  4 0.04613279
   7  9.6826600  5 0.08474279
   8 12.6601318  6 0.04876202
   9 13.3290942  7 0.06448464
  10 13.6177960  8 0.09228762
  11 13.6464567  9 0.13546891
```

Conclusion

We analyzed the methods of ARIMA and ETS to forecast the Sales of Footwear for the future quarters of the fiscal year of Nike. We found out that both of the models perform worse in terms of accuracy in the test dataset. However the test that performed the best is the ETS. Both of the models were able to capture most of the information of Sales.