Project: Forecasting Sales

Complete each section. When you are ready, save your file as a PDF document and submit it here: <https://classroom.udacity.com/nanodegrees/nd008/parts/edd0e8e8-158f-4044-9468-3e08fd08cbf8/project>

## Step 1: Plan Your Analysis

*Look at your data set and determine whether the data is appropriate to use time series models. Determine which records should be held for validation later on (250 word limit).*

*Answer the following questions to help you plan out your analysis:*

1. **Does the dataset meet the criteria of a time series dataset? Make sure to explore all four key characteristics of a time series data.**

**Answer**

Yes, the dataset meets the criteria of a time series dataset. All the following four characteristics of the dataset are met:

1. It's over a continuous time interval
2. There are sequential measurements across that interval
3. There is equal spacing between every two consecutive measurements
4. Each time unit within the time interval has at most one data point
5. **Which records should be used as the holdout sample?**

**Answer**

Since we have to provide a forecast for the next 4 months of sales and report our findings, I will use last 4months data as the holdout sample (from June 2013 to September 2013)

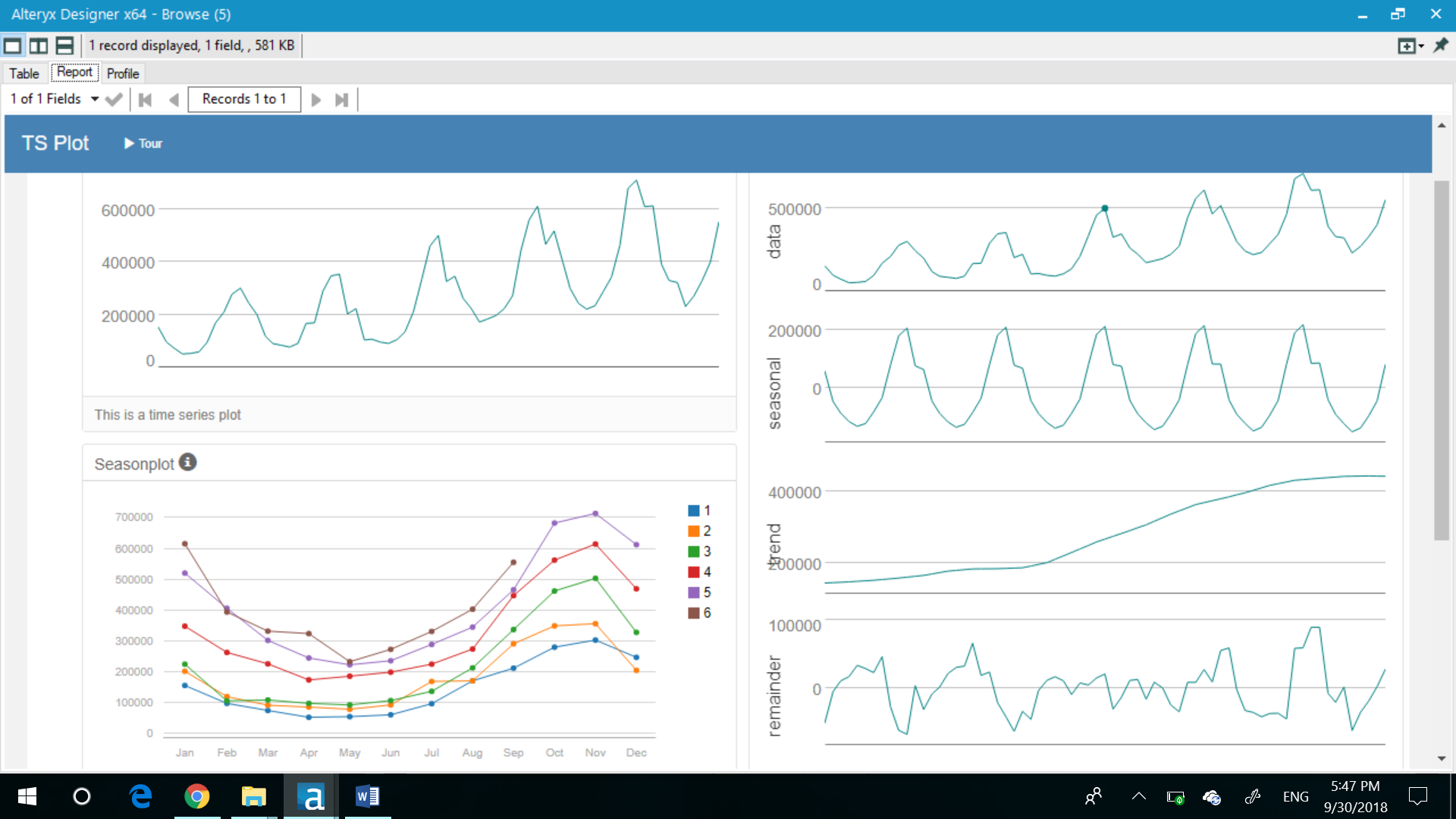
## Step 2: Determine Trend, Seasonal, and Error components

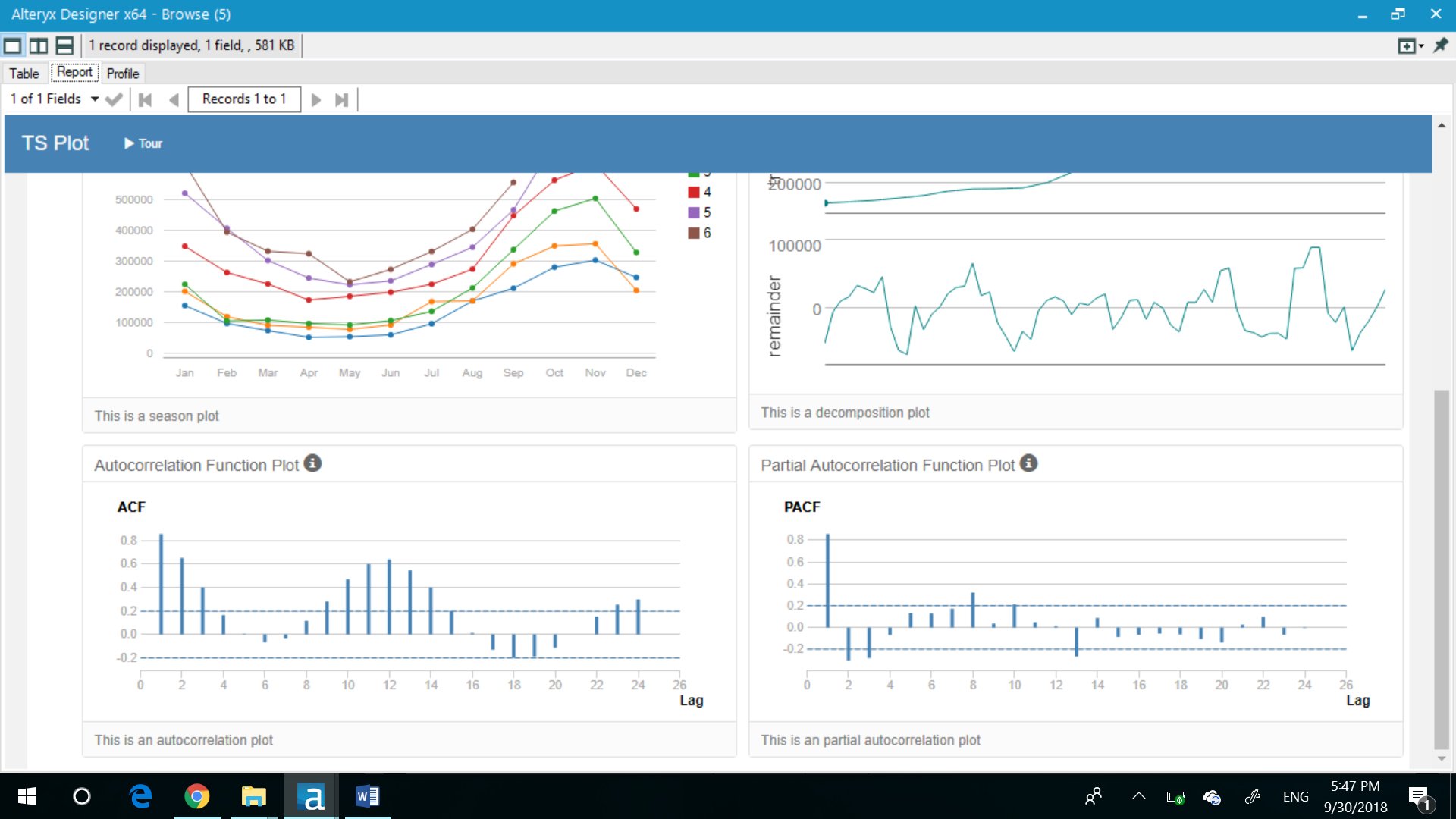
Graph the data set and decompose the time series into its three main components: trend, seasonality, and error.  *(250 word limit)*

*Answer this question:*

1. What are the trend, seasonality, and error of the time series? Show how you were able to determine the components using time series plots. Include the graphs.

The decomposition plot shows the three main components: error, trend, and seasonality of the time series are as given in the plot below (using TS Plot from Alteryx):





Each can be applied either additively, multiplicatively, or not at all.

Trend - The trend plot is linear so we apply it additively (A).

Seasonal – There are peaks and valleys at regular intervals in our data indicating that we have seasonality in our data and need to apply differencing. The size of the seasonal fluctuations tends to increase (very small increments) with the level of time series, hence we apply it multiplicatively (M).

Error -  The error plot is fluctuating between large and small errors over time, hence we apply it multiplicatively (M).

Step 3: Build your Models

*Analyze your graphs and determine the appropriate measurements to apply to your ARIMA and ETS models and describe the errors for both models. (500 word limit)*

*Answer these questions:*

1. What are the model terms for ETS? Explain why you chose those terms.
   1. Describe the in-sample errors. Use at least RMSE and MASE when examining results

We refer the decomposition plot to define our terms for the ETS model:

* Trend line is linear so we will use an additive method.
* The seasonality changes in magnitude each year so a multiplicative method is used.
* The error changes in magnitude as the series goes along so a multiplicative method will be used.

This leaves us with an ETS(M, A, M) model​.

We can use internal and external validation to validate the quality of the model.

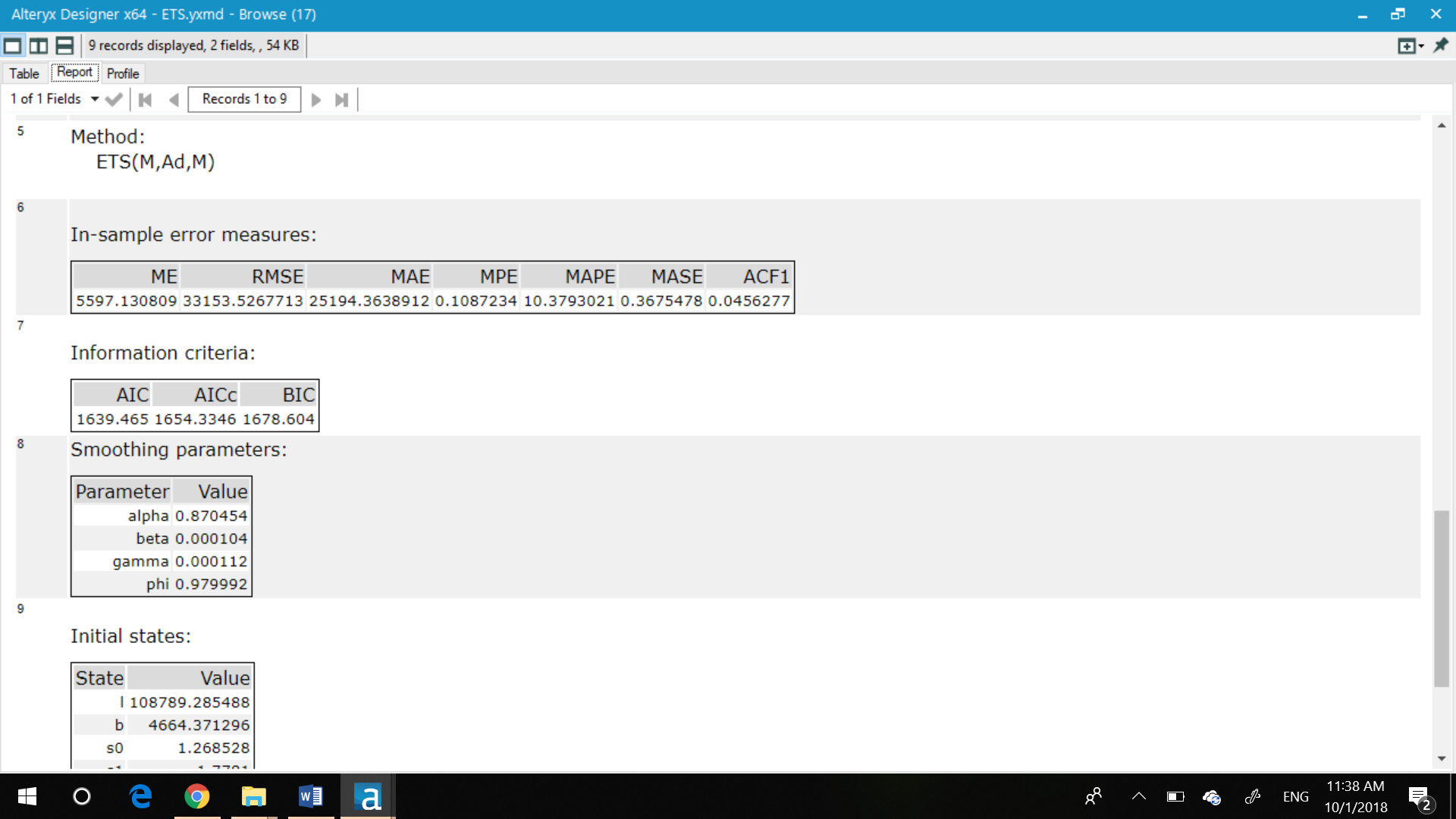
For internal validation we look at in-sample error measures, particularly RMSE (Root-Mean-Square Error) and MASE (Mean Absolute Scaled Error).

RMSE is the in-sample standard deviation of the [residuals](https://www.statisticshowto.datasciencecentral.com/residual/) (prediction errors). It is a measure of how spread out these residuals are.

MASE which can be used to compare forecasts of different models. It is another relative measure of error that is applicable only to time series data. It is the mean absolute error of the model divided by the mean absolute value of the first difference of the series. It measures the relative reduction in error compared to a naive model. Ideally its value will be significantly less than 1 but is relative to comparison across other models for the same series

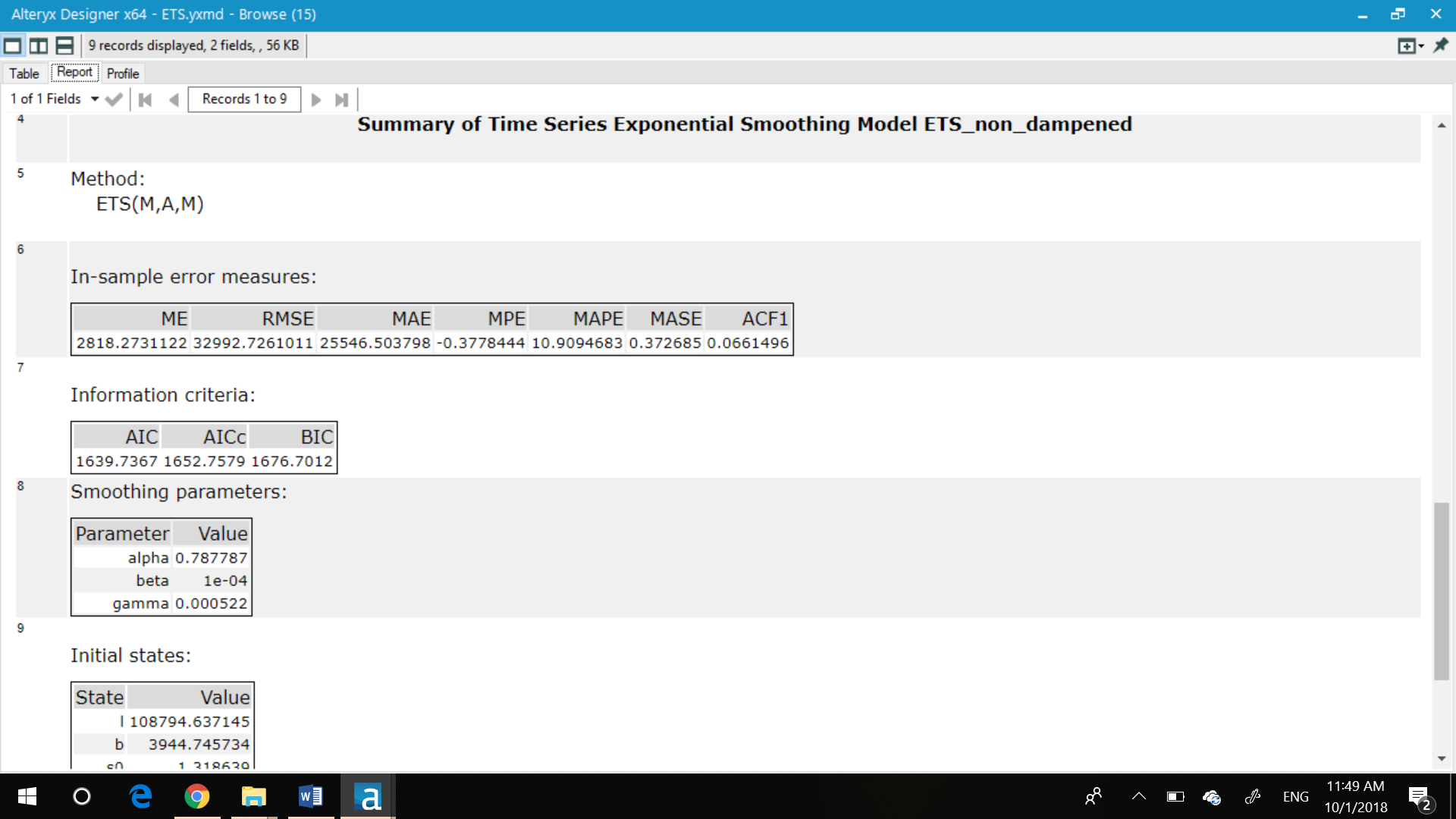
**Dampened ETS model:**

In Dampened ETS model the RMSE is 33153 and MASE is 0.367, less than 1 and is acceptable.



**Non- dampened ETS model**

In non-dampened ETS model the RMSE is 32992 and MASE is 0.372 (less than 1 and hence acceptable).



1. **What are the model terms for ARIMA? Explain why you chose those terms. Graph the Auto-Correlation Function (ACF) and Partial Autocorrelation Function Plots (PACF) for the time series and seasonal component and use these graphs to justify choosing your model terms.**

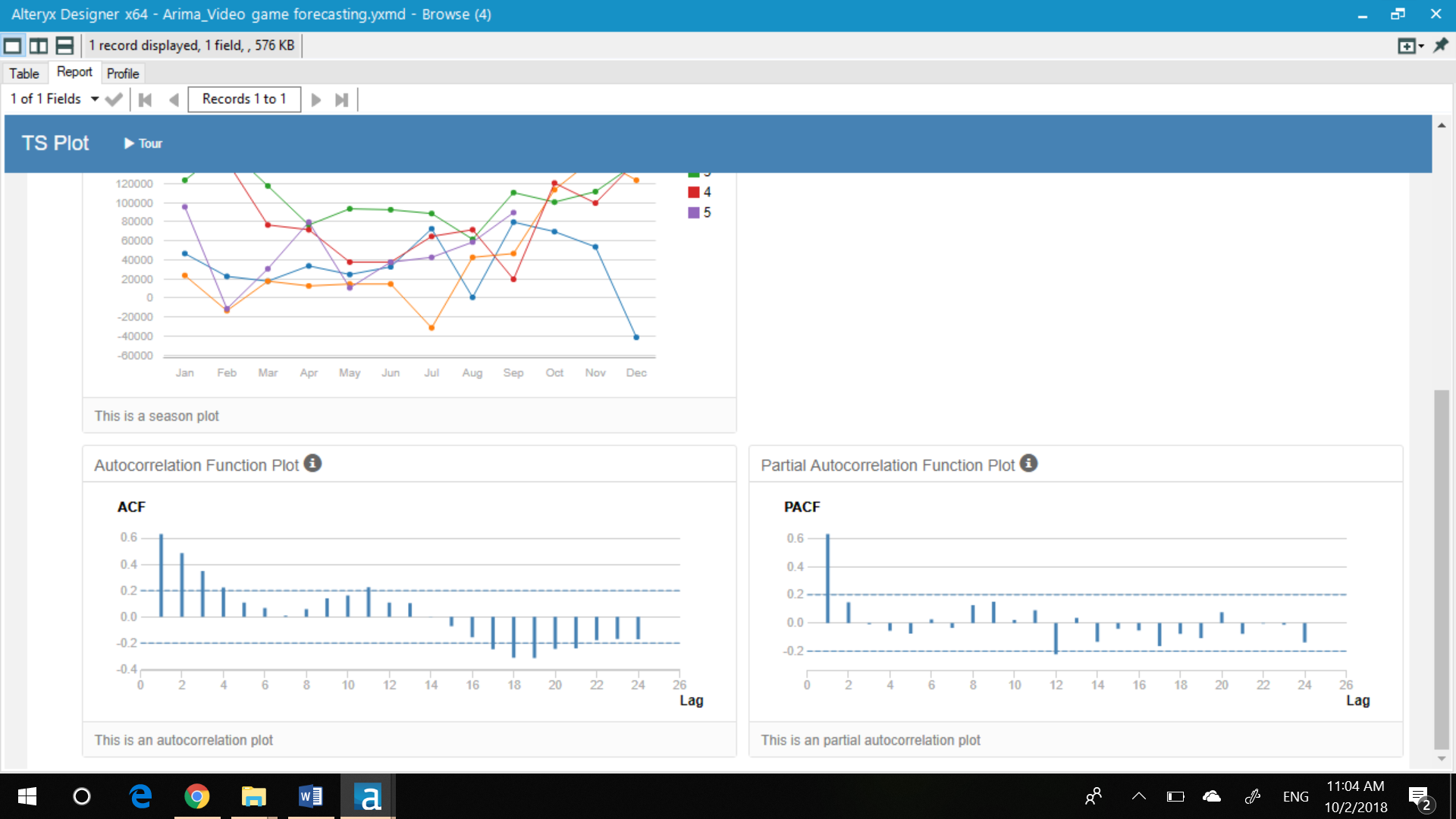
## As seen in the TS plot of Step2 there are seasonal components found in the time series I will use an ARIMA(p, d, q)(P, D, Q)m model for forecasting.

**The graph of Auto-Correlation Function (ACF) and Partial Autocorrelation Function Plots (PACF) for the time series**:



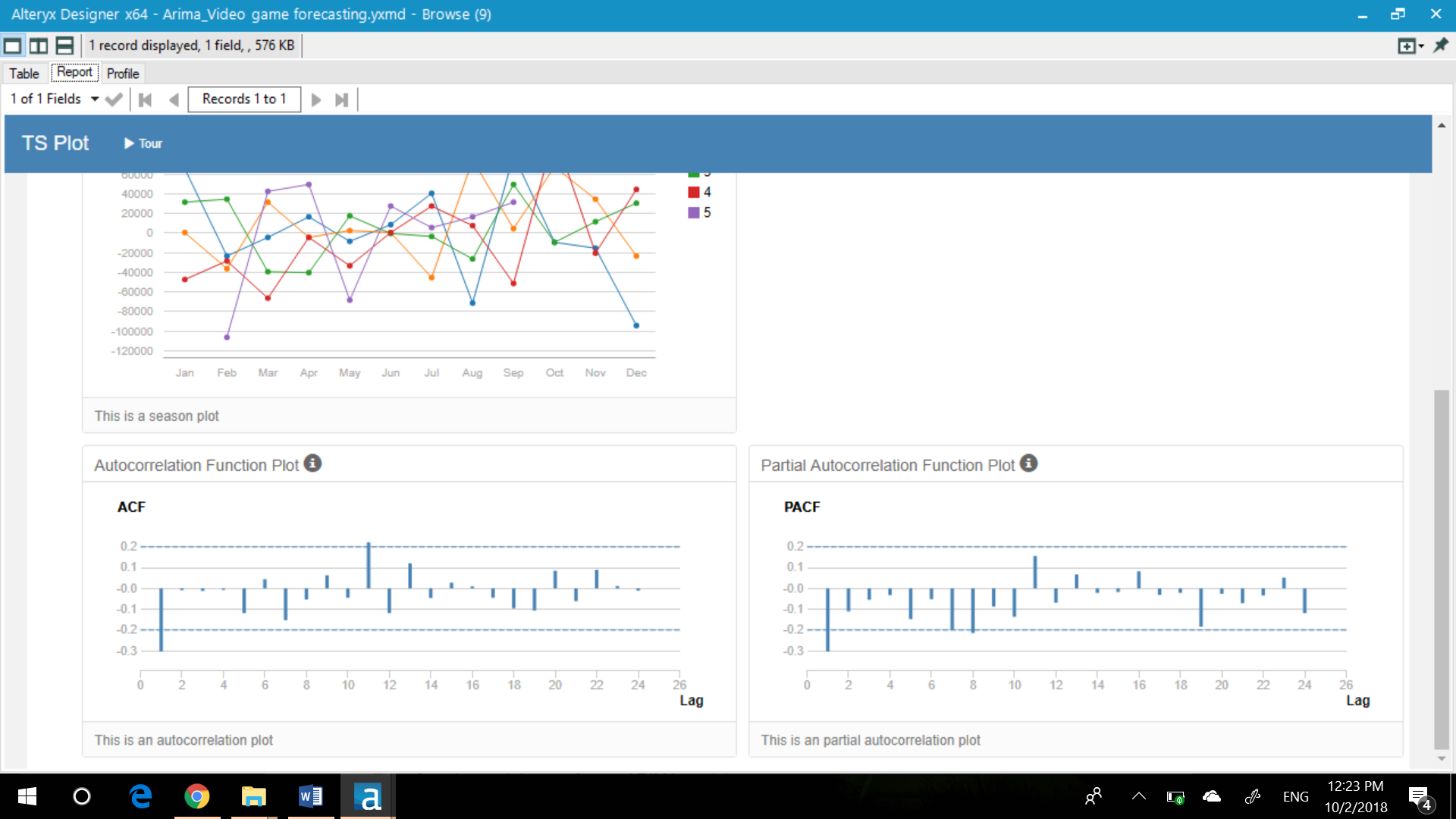
## The ACF presents slowly decaying serial correlations towards 0 with increases at the seasonal lags. Since serial correlation is high I will need to seasonally difference the series.

**Seasonal Difference ACF and PACF**::- After the seasonal differencing we can see that still there is a correlation in the lags and all the lags are significant. Correlation has reduced by seasonal differencing, but it needs to be removed by differencing further.



## Seasonal First Difference:

The seasonal first difference of the series has removed most of the significant lags from the ACF and PACF so there is no need for further differencing. The remaining correlation can be accounted for using autoregressive and moving average terms.



The differencing terms will be d(1) and D(1) for first differencing.

The ACF and PACF plots show that there is a strong negative correlation at lag 1, this suggests that the MA (1) for the non-seasonal component.

Non-seasonal (p,d,q) is (0,1,1)

As our seasonal lags in the ACF and PACF (lag 12, 24, etc.) do not have any significant correlation, we will consider the Seasonal (P, D, Q) as (0,1,0) i.e., there is no need for seasonal autoregressive or moving average terms.

m is 12

## Hence, ARIMA(p, d, q)(P, D, Q)m model for forecasting is ARIMA(0,1,1)(0,1,0)12

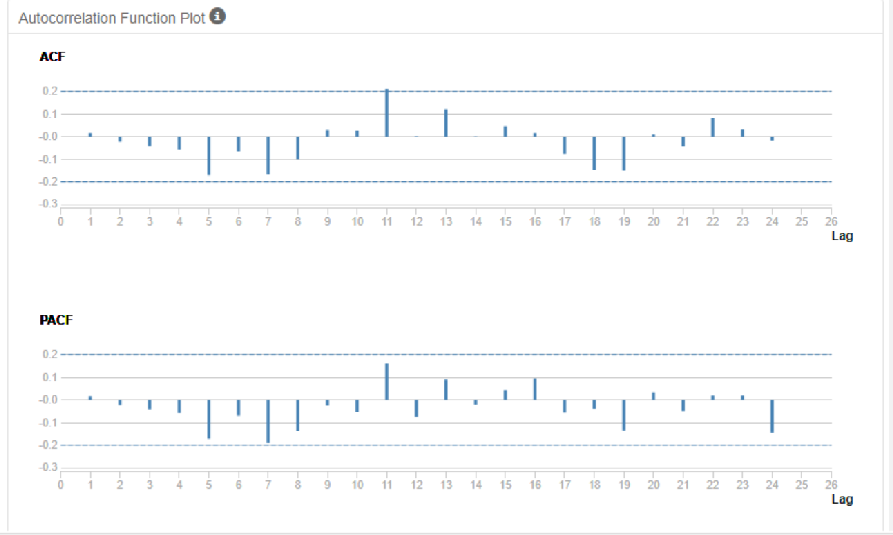
**Question**

* 1. **Describe the in-sample errors. Use at least RMSE and MASE when examining results**
  2. **Regraph ACF and PACF for both the Time Series and Seasonal Difference and include these graphs in your answer.**



The RMSE( in-sample standard deviation) is 36761

The MASE(used to compare forecasts of different models) is 0.36 which is less than 1, and is acceptable.



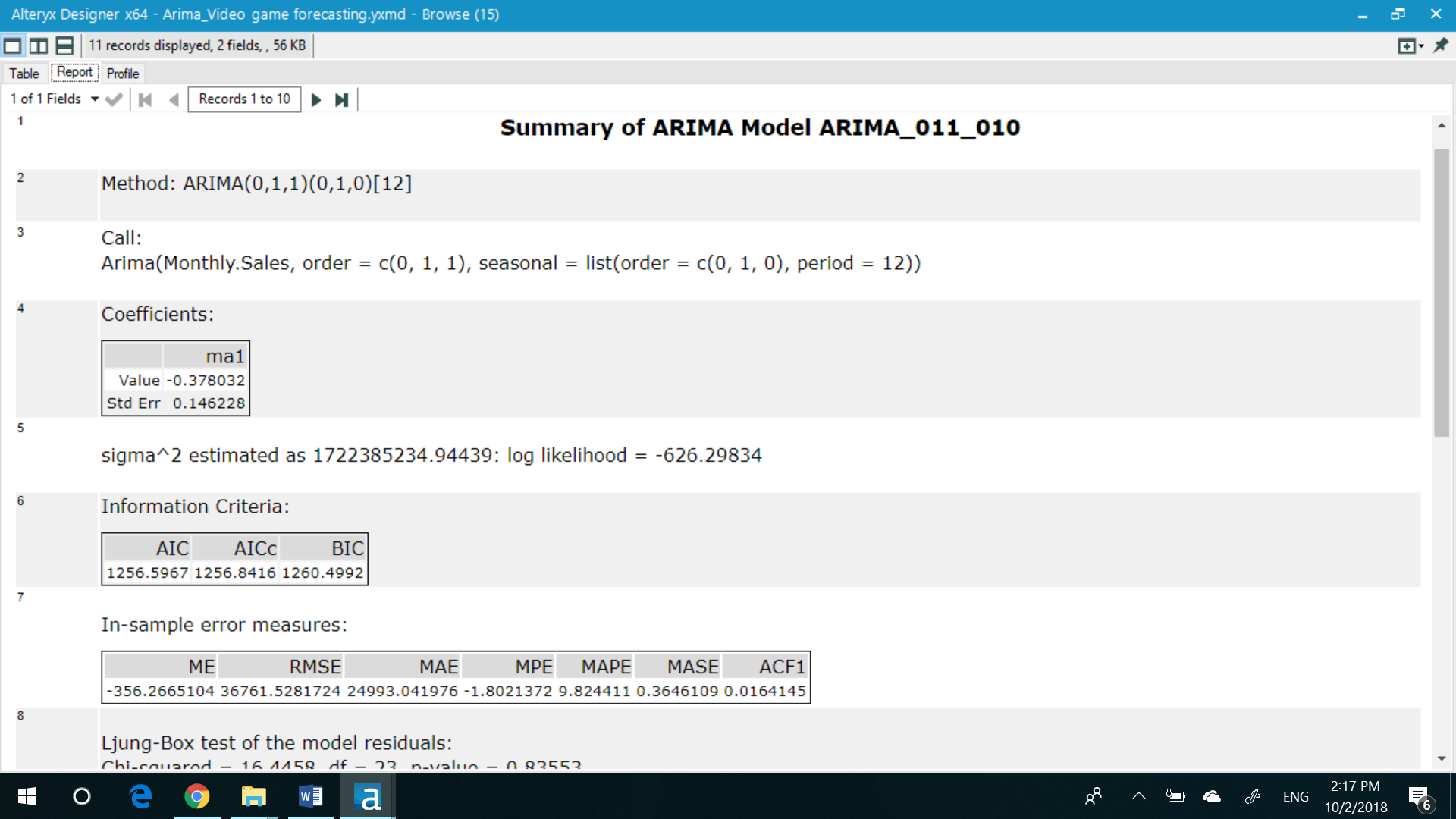
## Step 4: Forecast

*Compare the in-sample error measurements to both models and compare error measurements for the holdout sample in your forecast. Choose the best fitting model and forecast the next four periods. (250 words limit)*

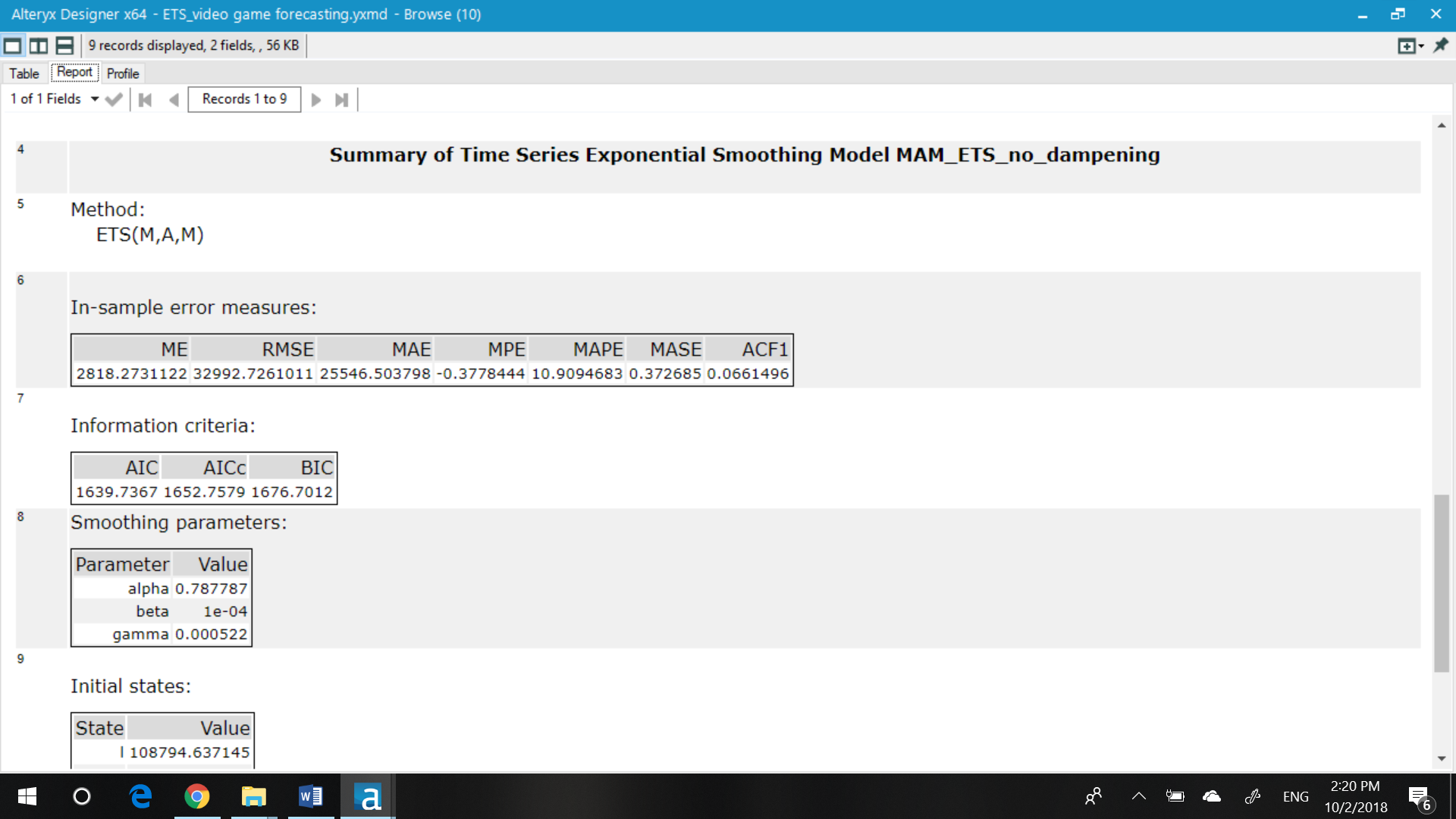
*Answer these questions.*

1. Which model did you choose? Justify your answer by showing: in-sample error measurements and forecast error measurements against the holdout sample.
2. What is the forecast for the next four periods? Graph the results using 95% and 80% confidence intervals.

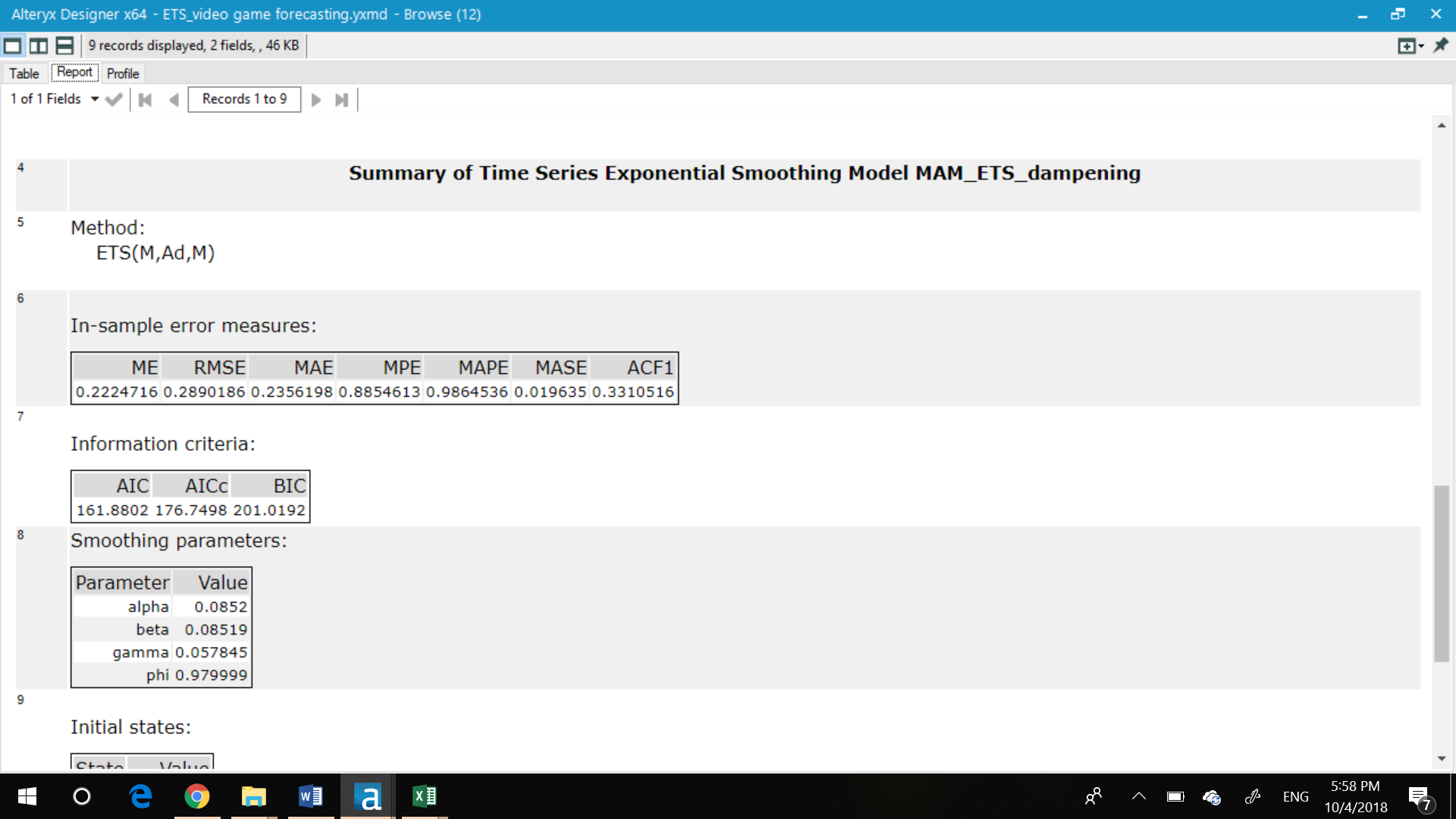
**Summary of Arima Model(In sample errors)**:



Summary of ETS Non-damped Model(In sample errors)



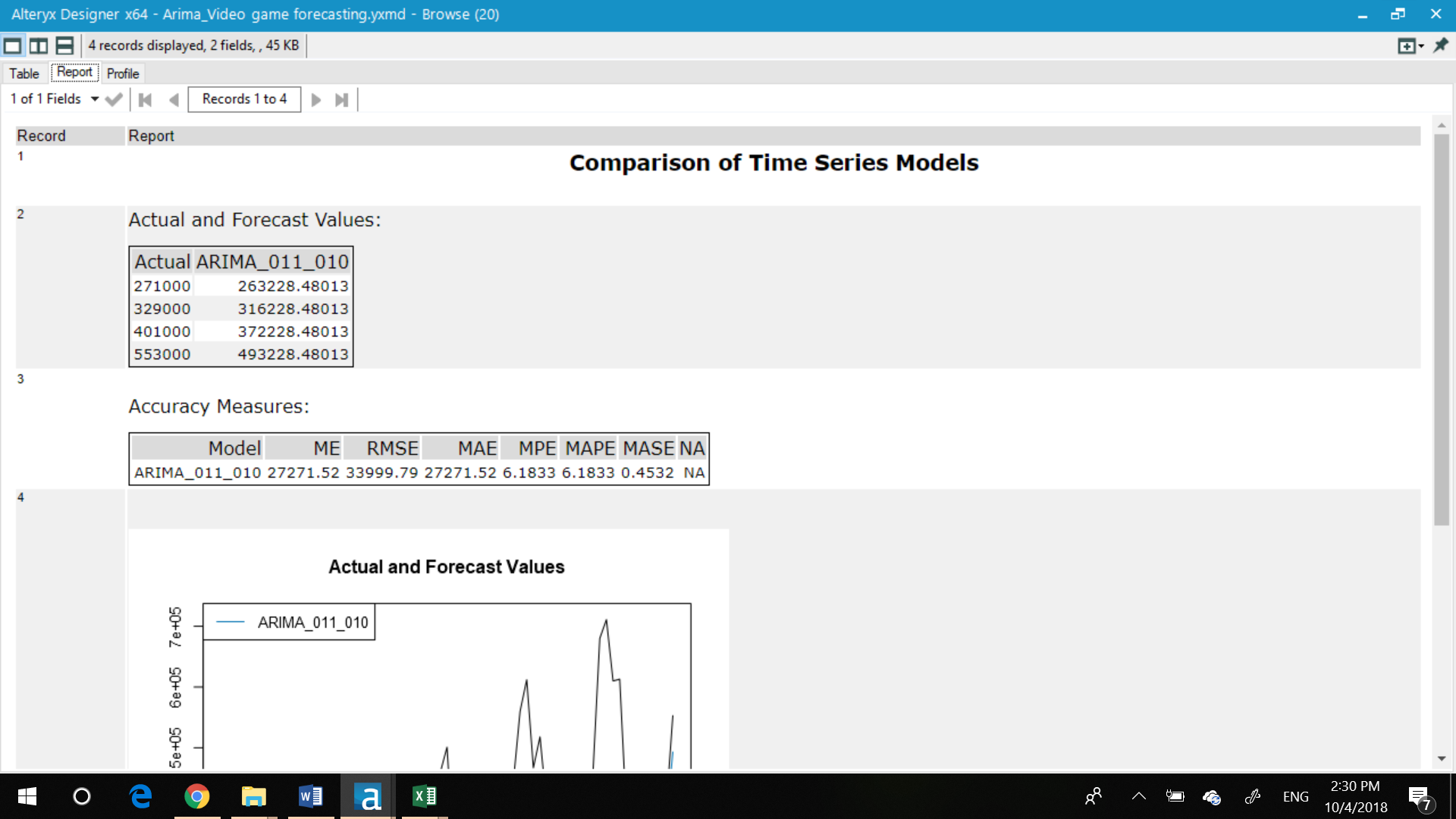
Summary of ETS damped Model (in sample errors)



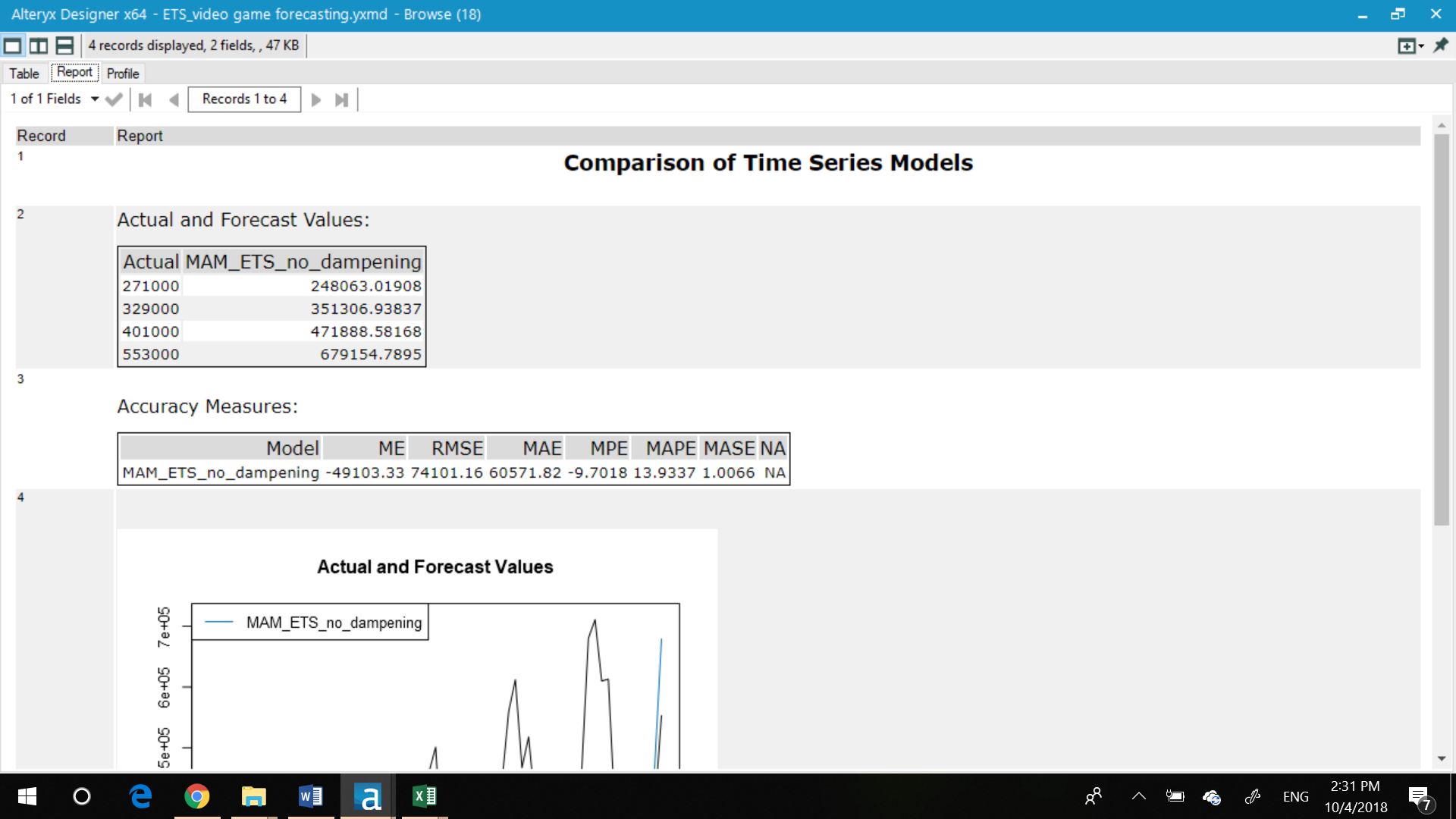
Further investigation shows that the MAPE and ME of the ARIMA model are lower than the ETS(non-damped) and ME is lower than both ETS (damped) and ETS (non-damped). This suggests that, on average, the ARIMA model misses its forecast by a lesser amount.

Now let’s compare the predictability of all the models for hold out sample:

ARIMA Model

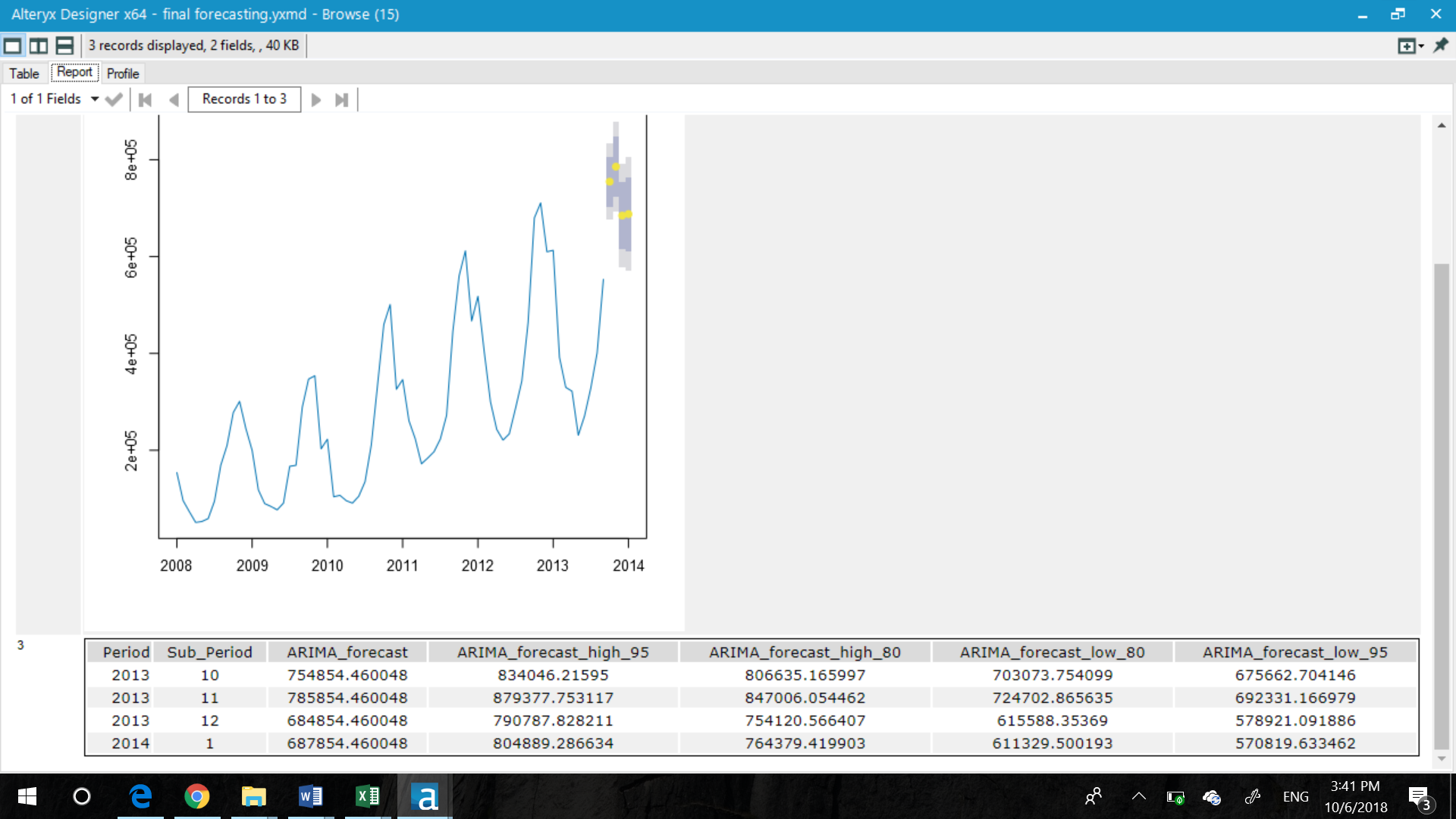


ETS (No Dampening)

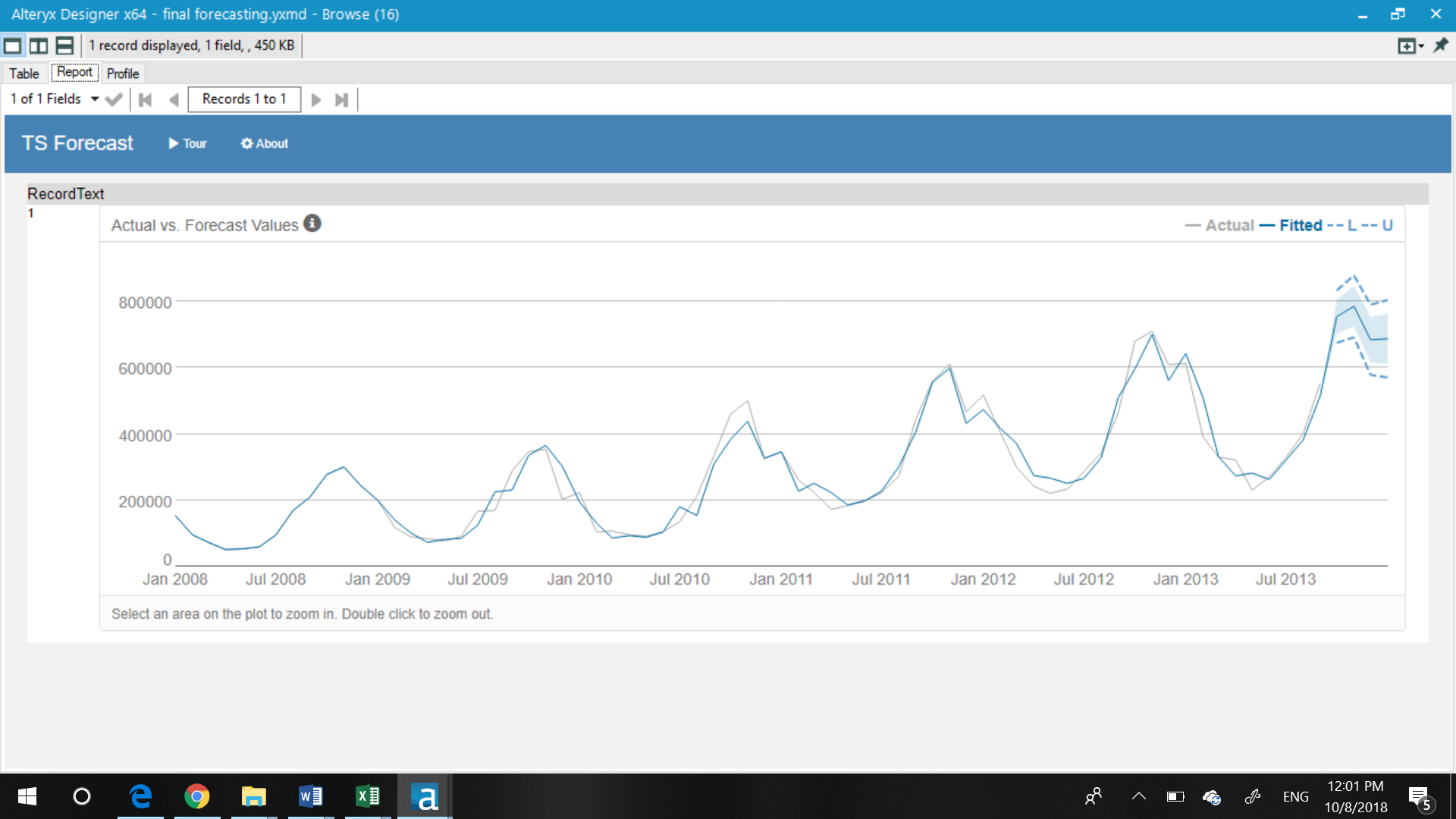


|  |  |  |
| --- | --- | --- |
| **Accuracy measures on hold out sample** | | |
|  | **Arima** | **ETS Non-dampened** |
| **ME** | 27271 | -49103 |
| **RMSE** | 33999 | 74101 |
| **MAE** | 2721 | 60571 |
| **MPE** | 6.18 | -9.7 |
| **MAPE** | 6.18 | 13.9 |
| **MASE** | 0.45 | 1.006 |

we see that the ARIMA model has better predictive qualities in just about every metric. Hence for forecasting we will use ARIMA model.



The graph of actual vs forecast values is ass given below:



Before you Submit

Please check your answers against the requirements of the project dictated by the [rubric](https://review.udacity.com/#!/rubrics/302/view) here. Reviewers will use this rubric to grade your project.