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.

* Data covers full continues time interval: there are 5.9 years of continue data.
* Equal spacing between every 2 consecutive measures: all data is tracked monthly for all the provided data
* There is at most one data point for every time unit: there is only one record for every month in the full data

1. Which records should be used as the holdout sample?

It has been used last 4 months for holdout to perform the forecast validation.

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

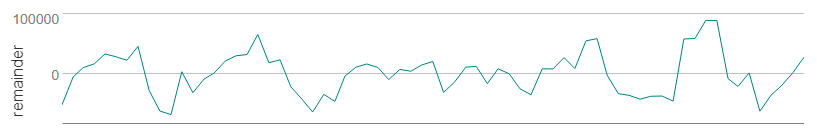
Graph the data set and decompose the time series into its three main components: trend, seasonality, and error.  *(250 word limit)*

*Answer this question:*

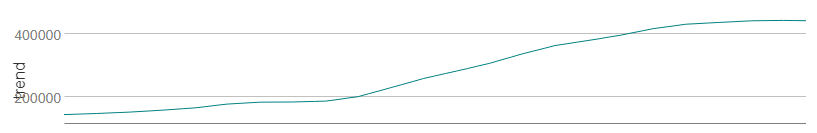
1. What are the trend, seasonality, and error of the time series? Show how you were able to determine the components using time series plots. Include the graphs.

Using the TS Plot has been generated the next graphs for and Error, Trend and Seasonality respectively.

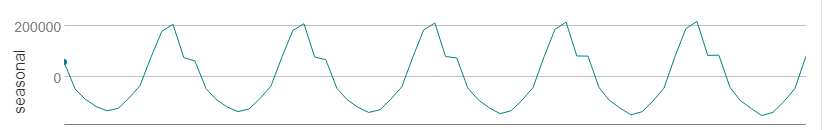
*The graph for remainder or error shows a fluctuation of large and small errors over time, then has been considered as Multiplicatively (M).*



*The trend in the full graph can be considered as linear and therefore the Trend is Additive (A), although after the half of graph, it has been observed that trend grows up more considerably.*



*Per the graph, the fluctuation for every season seems to remains constant, but if observed with more detail, for last 3 seasons, there is a small increase in the peaks. So, Seasonality has been considered as Multiplicatively (M).*



Step 3: Build your Models

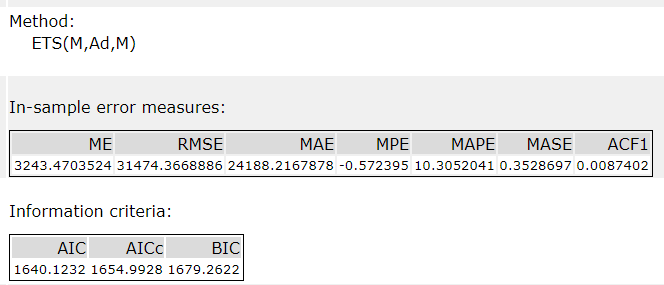
*Analyze your graphs and determine the appropriate measurements to apply to your ARIMA and ETS models and describe the errors for both models. (500 word limit)*

*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

The components set for the ETS are (M, A, M) as result of the analysis from TS Plot graphs for Error, Trend and Seasonality.

The Internal Validation has been done using the RMSE and MASE errors, as well is being considered the AIC value:



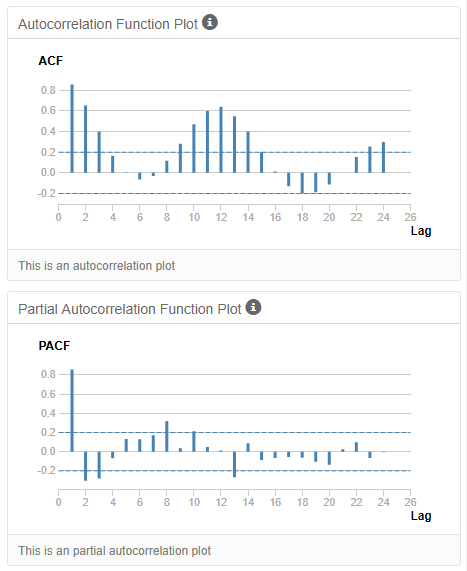
1. What are the model terms for ARIMA? Explain why you chose those terms. Graph the Auto-Correlation Function (ACF) and Partial Autocorrelation Function Plots (PACF) for the time series and seasonal component and use these graphs to justify choosing your model terms.
   1. Describe the in-sample errors.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

After analysis of graphs from the TS Plot has been observed that there is a seasonality. And therefore, has been chosen to configure a Seasonal ARIMA Model.

For the ARIMA time series model has been chosen the next terms ARIMA (p,d,q) (P,D,Q) m considering the graphs for ACF and PACF from the TS Plots using the original data and calculating the seasonal differences. After this analysis, has determined to use the next values for the ARIMA MODEL.

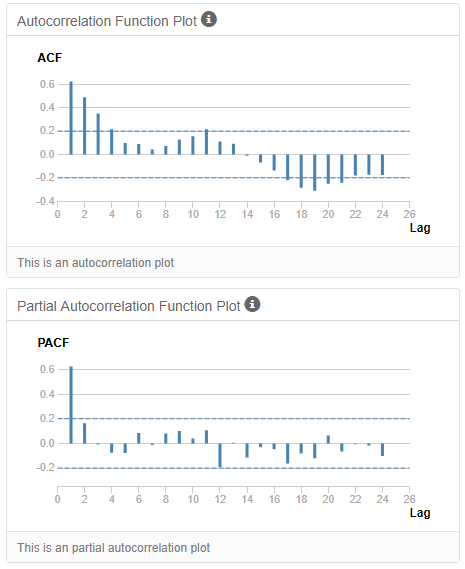
* p = 0
* d = 1
* q = 1
* P = 0
* D = 1
* Q = 0
* m = 12

By the initial ACF and PACF graphs, can be observed that lag 1 is positive and slowly decrease for lags 2,3,4 and 5. And increase again at seasonal lags in 12 and 24. This can be considered as highly correlation and therefore has been required to calculate the seasonal difference and generate ACF and PACF graphs.



**Seasonal Difference for ACF and PACF**

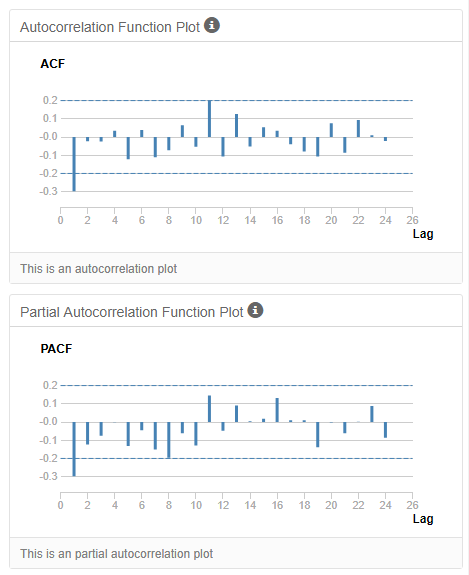
After applying the Season Difference it is observed that ACF and PACF graphs have reduced the value for lag 1 and decreased slowly near to 0 at lag 4. But this is still an indicator or high correlation, and is required to calculate first seasonal difference and plot again the ACF and PACF graphs to determine better the values for the ARIMA termns.



**First Seasonal Difference for ACF and PACF**

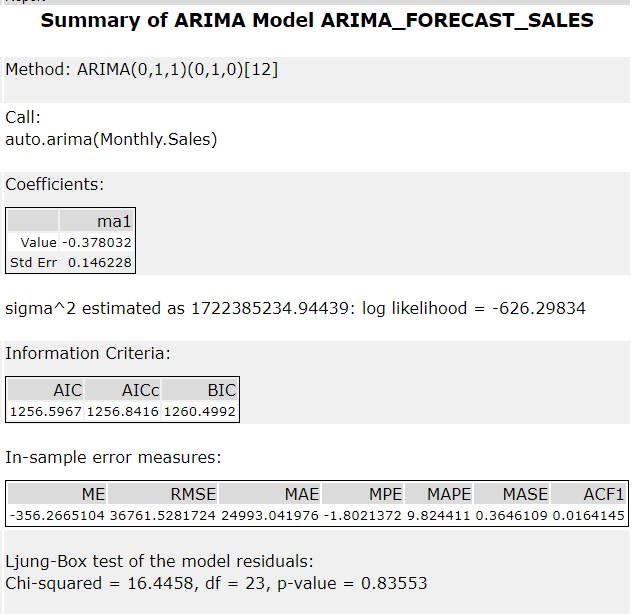
According to the ACF and PACF graphs below can be observed that the seasonal correlation has been eliminated and there is no required to perform any aditional difference. Therefore d and D has been set equal to 1.

Also the graphs for ACF and PACF have a negative correlation at lag 1, and therefore MA has been set to 1.



Internal validation considering the RMSE, MASE

* RMSE: 36,761, it is near to the median value.
* MASE: 0.36, since it is lower than 1, it is a good indicator of a good level of accuracy.



## Step 4: Forecast

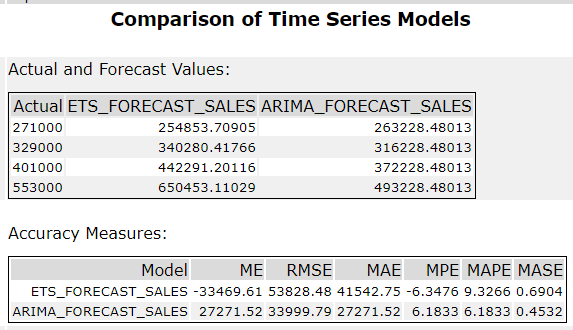
*Compare the in-sample error measurements to both models and compare error measurements for the holdout sample in your forecast. Choose the best fitting model and forecast the next four periods. (250 words limit)*

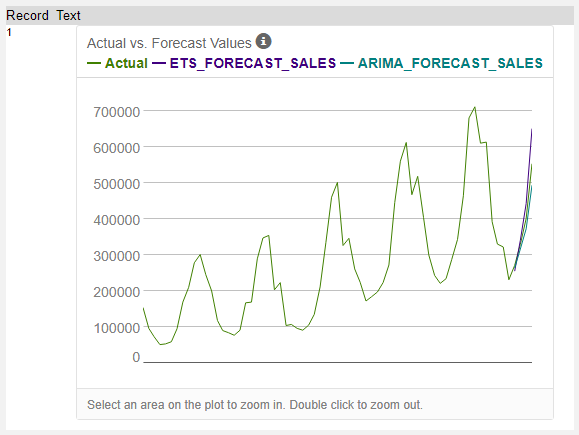
*Answer these questions.*

1. Which model did you choose? Justify your answer by showing: in-sample error measurements and forecast error measurements against the holdout sample.

**The External Validation using the hold out sample with ETS and ARIMA models**

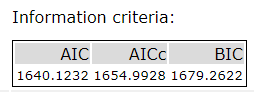
It can be observed that the values for RMSE and MASE are lower for the ARIMA model, which indicates is performing better at making the predictions again the hold out sample. Also, the MASE value can be considered as a good indicator of the accuracy for the forecast prediction, because is lower than 1.

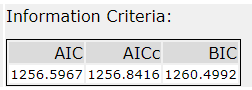




**AIC values for ETS and ARIMA**

Also, has been compared the AIC values for the ETS and ARIMA models respectively, for which can be confirmed that ARIMA model is performing better because it has a lower value.

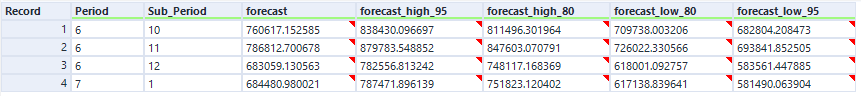




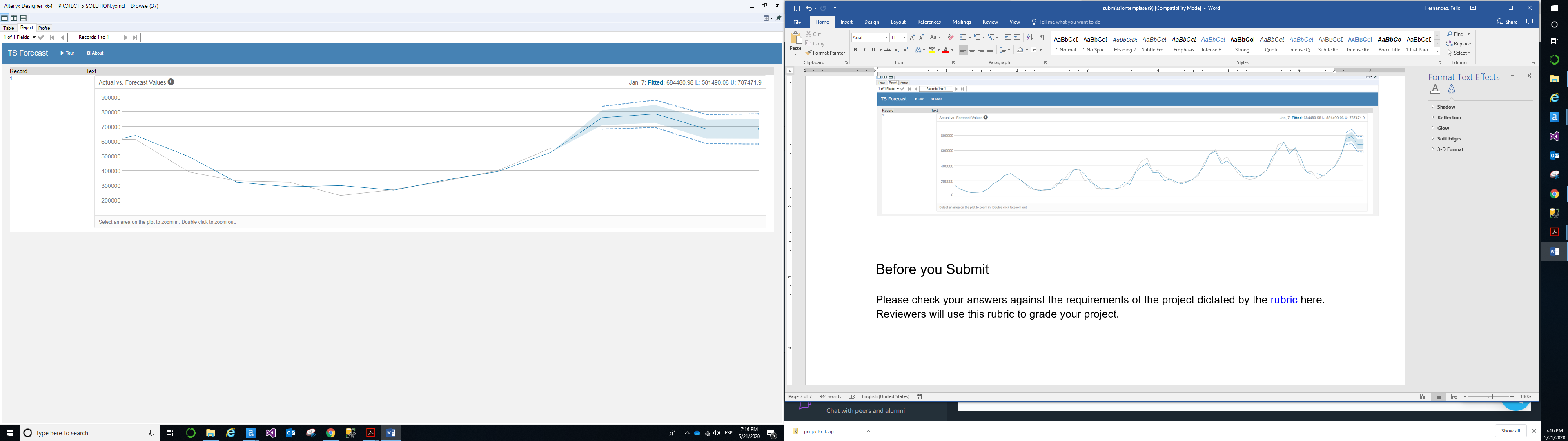
Therefore, has been selected the ARIMA model as the best option, because it has lower RMSE, MASE and AIC values as well it has generated more accurate values for the test forecast.

1. What is the forecast for the next four periods? Graph the results using 95% and 80% confidence intervals.

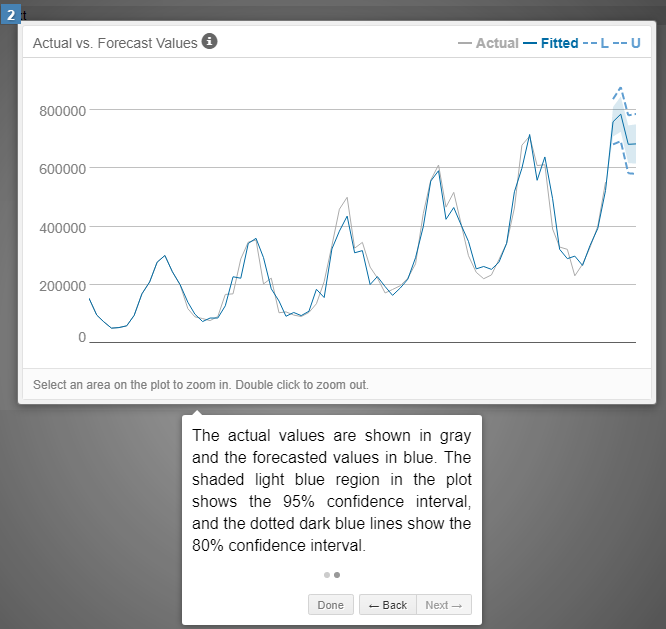
Using the ARIMA model with the 69 periods of data the result for the prediction of the next f4 months are:



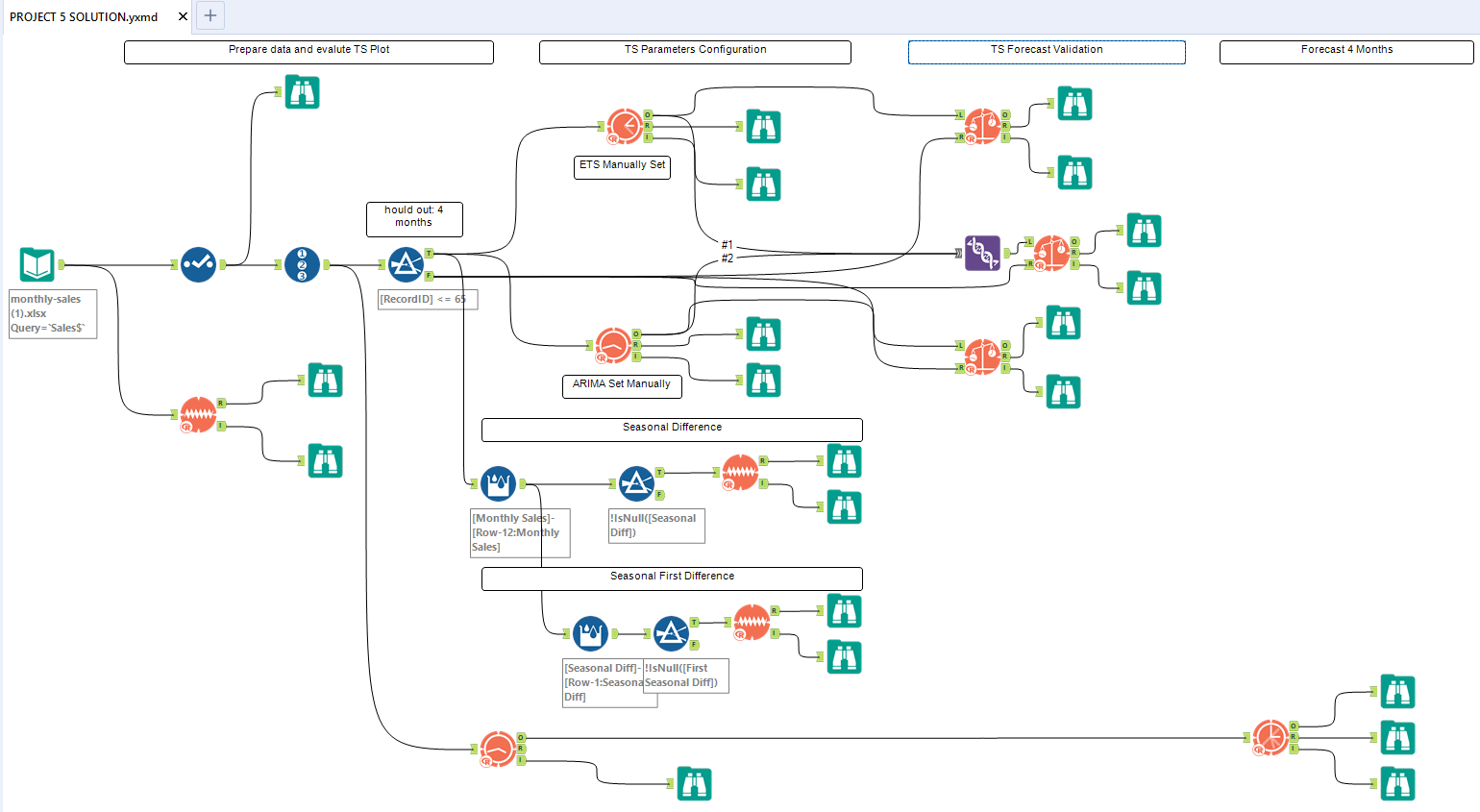
Last year of data including 4 months forecast:



All periods including 4 months forecast:



**Alteryx Workflow**



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.