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.
   Yes, dataset is time series data set:
   - measurments are in constant distances
   - measurments have no gap
   -
   -
2. Which records should be used as the holdout sample?
   Last 4 records, because dataset is grouped monthly and we are requested to forecast 4 months. It make sense to holdout last 4 months for validation of the model.

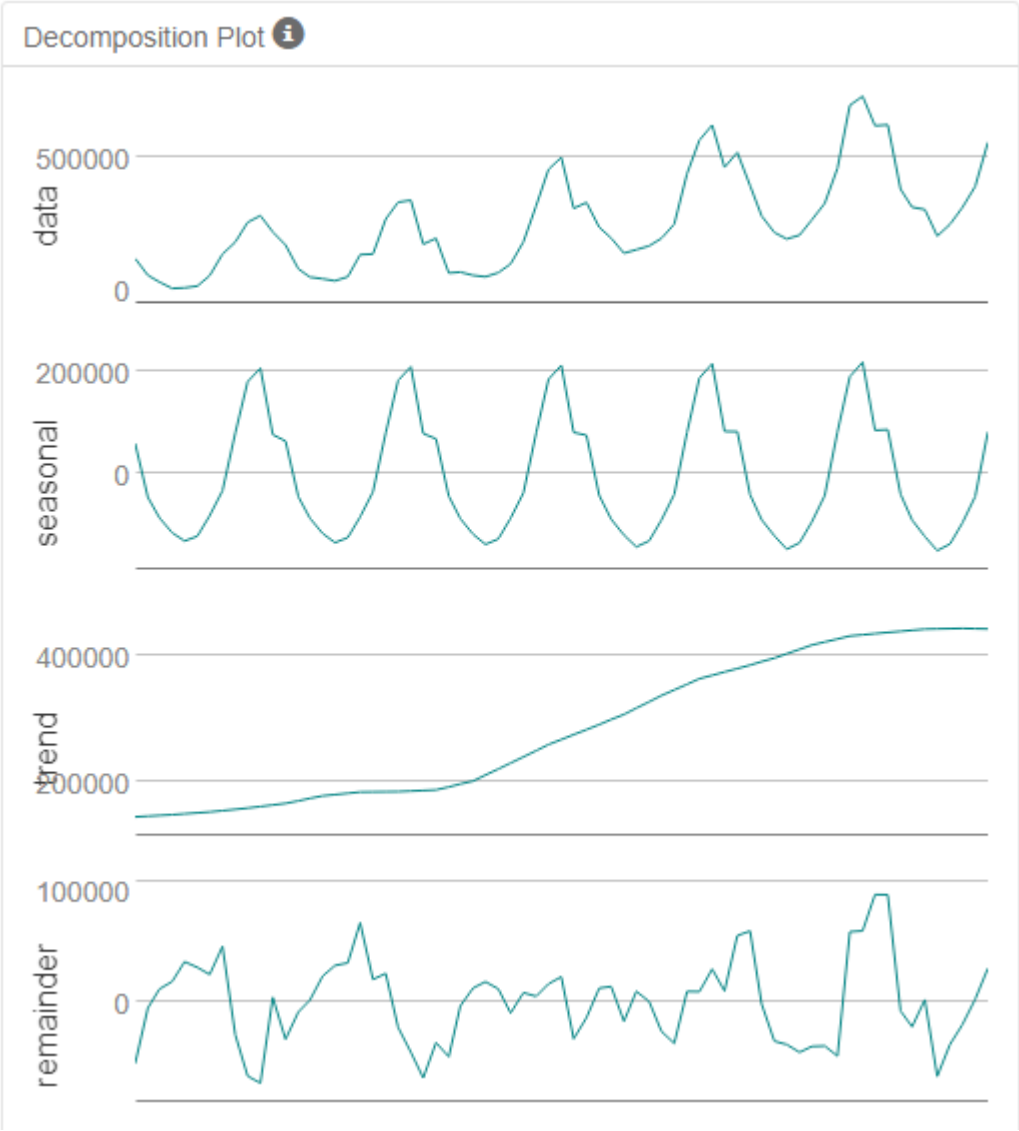
Step 2: Determine Trend, Seasonal, and Error components

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

Answer this question:

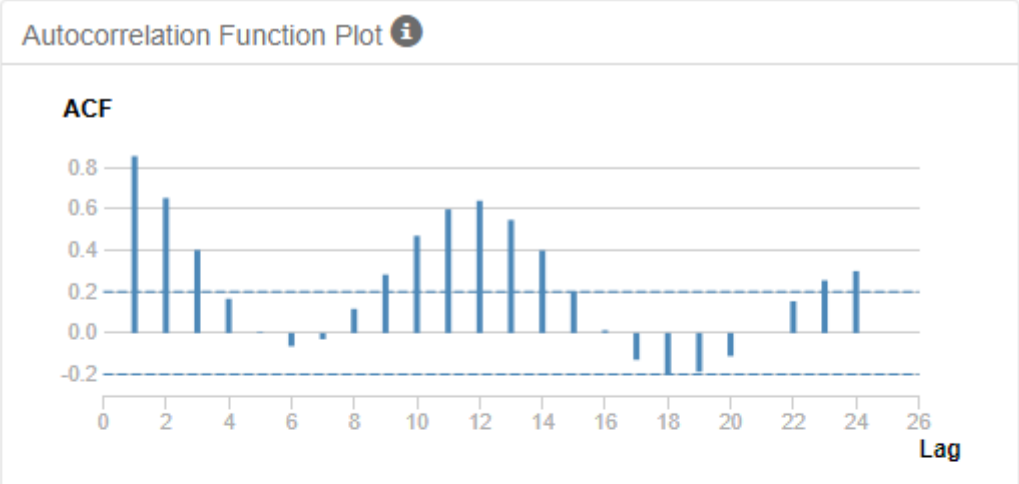
1. What are the trend, seasonality, and error of the time series? Show how you were able to determine the components using time series plots. Include the graphs.

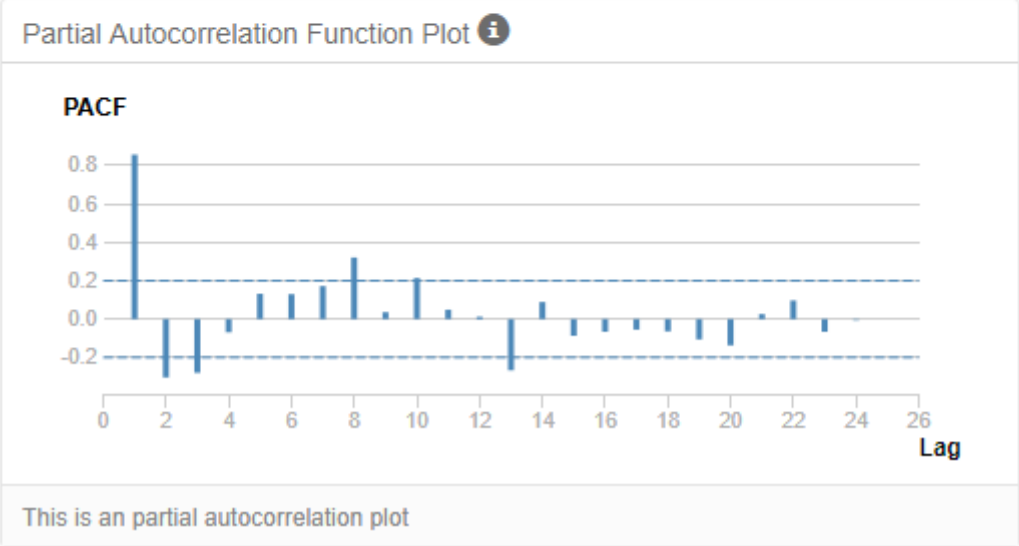
   I have used TS plot to visualize details:



We can see that series has stable and constant seasonality. Trend is increasing in linear manner. Error (remainder) is changing variance over time.

Autocorelation (ACF) and Partial autocorlelation (PACF) plots show that series will need to be stationalized before we can use them in model.





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:

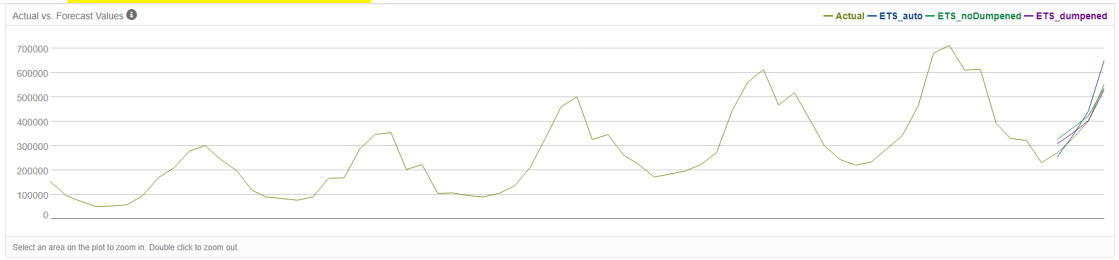
1. What are the model terms for ETS? Explain why you chose those terms.
    a. Describe the in-sample errors. Use at least RMSE and MASE when examining results

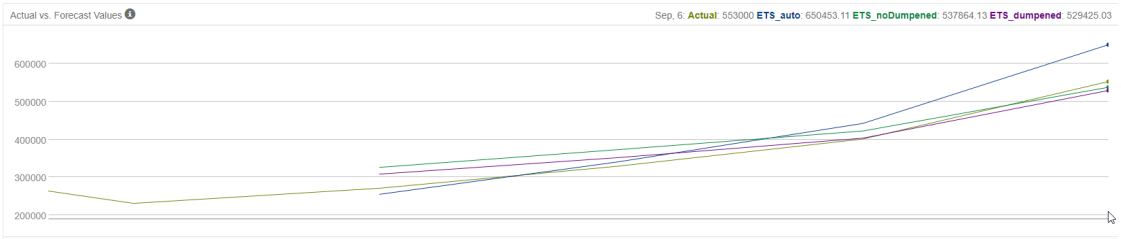
Model terms for ETS:
- trend is linear -> type Additive
- Seasons are constant and not increasing much in magnitude -> additive type
- Errors are changing in variance over time -> multiplicative type

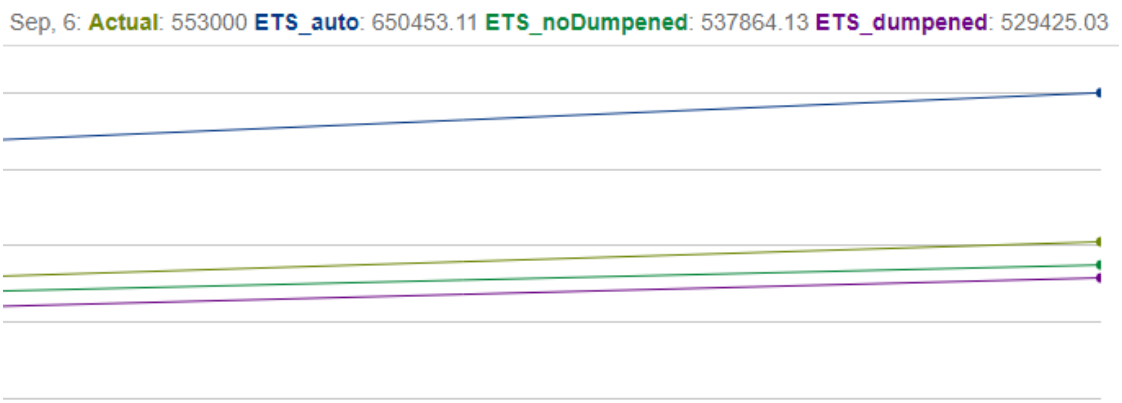
- no trend dampening

Reasoning:
- Prepared 3 ETS models:
--- ETS_auto
--- ETS_noDupened
--- ETS_dumpened

Results of all 3 models:





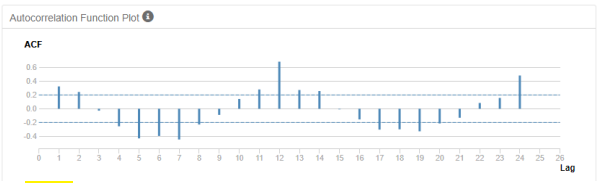


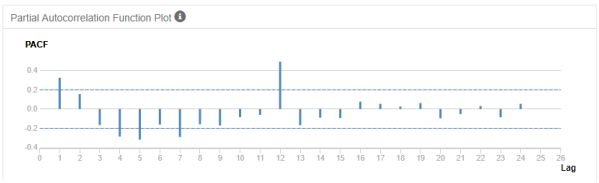
We can see that closest to Actual data are my models.

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

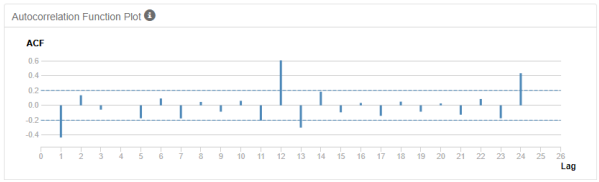
First TS plot shows that ACF and PCAF indicate that series is seasonal AND not stationary. First I needed to stationarized series.

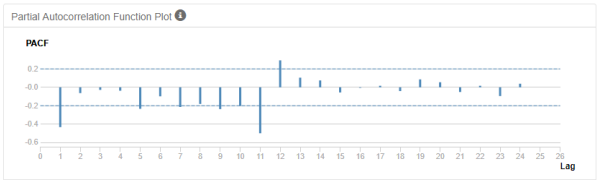
Calculated 1st differences:



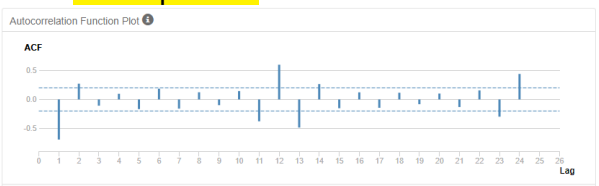


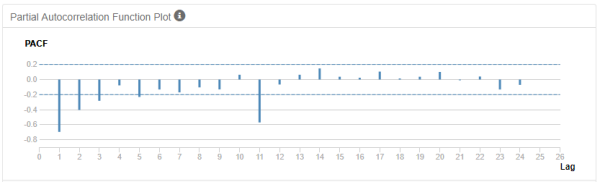
ACF and PACF still show some correlation to the past – therefore I need to calculate 2nd differences:





We can see that correlation dropped and pops out only when new season is starting – this seems acceptable. Although I have calculated 3rd differences to check how things develop there.

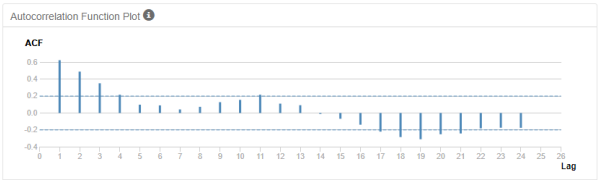




In the 3rd difference charts looks worse than in 2nd.

I have decided that with 2nd difference series is stationarized.
Having that I can select (p,d,q) of ARIMA model:
- p = 0; because ACF 1lag is negative
- d = 2; because we need 2 iterations of differentiation to make series stationary
- q = 1; because ACF 1lag is negative

Now seasonality:
I have calculated differences of current sample vs -12 periods (yearly season) and ploted ACF and PACF:



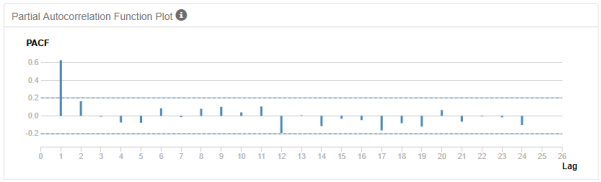
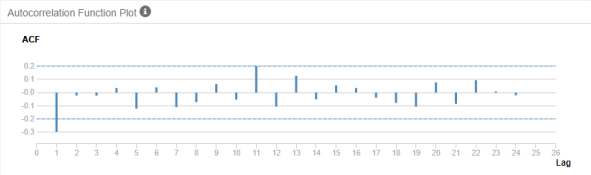
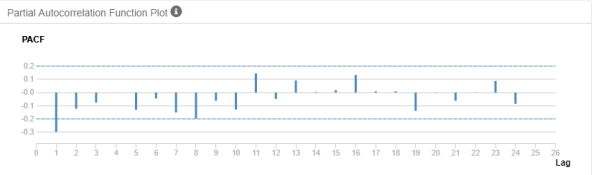


Chart shows that series is not stationarized, therefore I need to calculate 1st difference on seasonal series:





Now seasonal series seems stationary. I can determine (P,D,Q)m parameters of ARIMA.
- P = 0; ACF is negative,
- D = 1; because we needed one round of differentiating
- Q = 1; because ACF is negative and PACF is dropping gradually
- m = 12; as we have 12 periods of the seasonality

My ARIMA model is: (0,2,1)(0,1,1)12

a. Describe the in-sample errors. Use at least RMSE and MASE when examining results

Record	Report
1	**Summary of ARIMA Model ARIMA_021_011_12**
2	Method: ARIMA(0,2,1)(0,1,1)[12]
3	Call: Arima(Monthly.Sales, order = c(0, 2, 1), seasonal = list(order = c(0, 1, 1), period = 12))
4	Coefficients: ma1 sma1 Value -0.998032 -0.191808 Std Err 0.063216 0.203151
5	sigma^2 estimated as 1960404137.95439: log likelihood = -622.06075
6	Information Criteria: AIC AICc BIC 1250.1215 1250.6321 1255.917
7	In-sample error measures: ME RMSE MAE MPE MAPE MASE ACF1 -190.2842049 38442.7251815 26554.3601589 -0.5114379 10.6772968 0.3873881 -0.2387433

My in-sample errors indicate that:
- ME: on average my model is underestimating by $190.

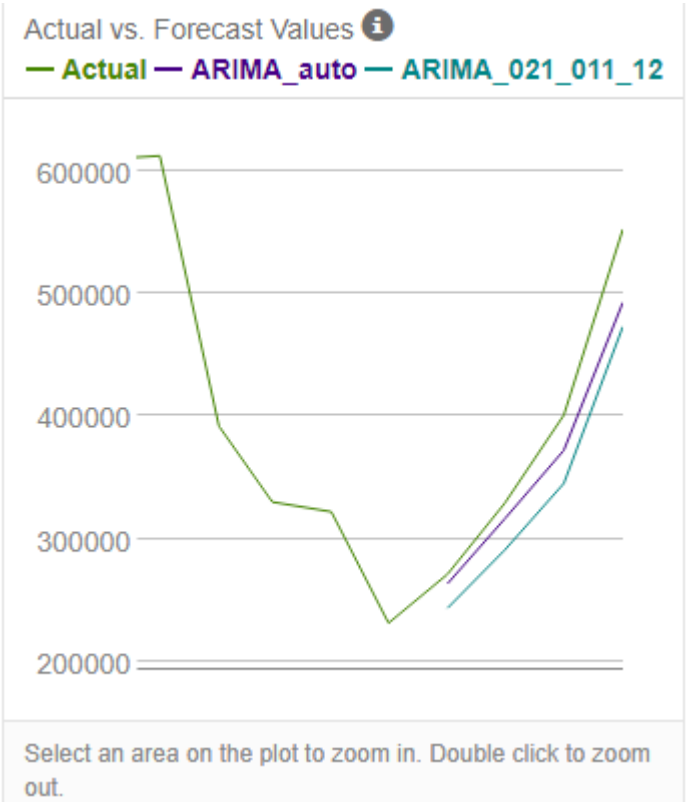
- RMSE: shows that predictions of my model are on average $38442 away from actual ones. Value alone is not saying much, but if we pair it with magnitude of sales we are talking about (300,000 – 550,000) it shows that on average we may be off by something like 10% (which in fact is confirmed by MAPE)

- MAE: is similar to ME but takes ABSOLUTE values. What we get here is that on average error is $26554 (regardless of is it in plus or in minus).
- MPE: is -0.51 which indicates similar like ME that my model is biased toward underestimating.
- MAPE: indicates that my model has on average 10.7% error
- MASE: shows value 0.39, which can be interpreted as much less than 1 and this indicates that this model is much better than naive model of forecast.

Similar like in ETS – I have prepared also AUTO version of ARIMA model to double check my decisions.
ARIMA_auto turned out to be better model ((0,1,1)(0,1,0)[12])

In-sample error measures:

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
-356.2665104	36761.5281724	24993.041976	-1.8021372	9.824411	0.3646109	0.0164145



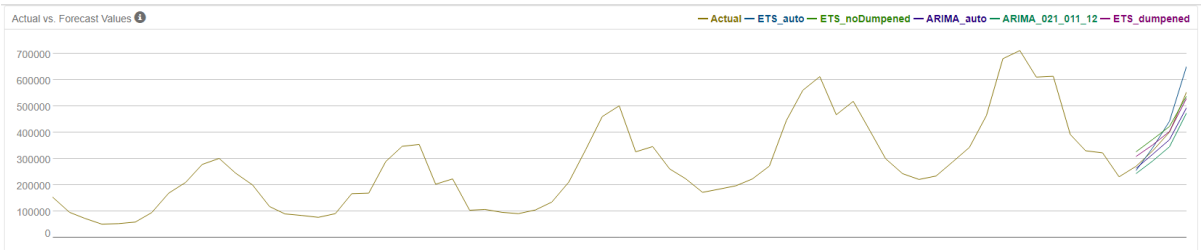
b. Regraph ACF and PACF for both the Time Series and Seasonal Difference and include these graphs in your answer.
Done, see above

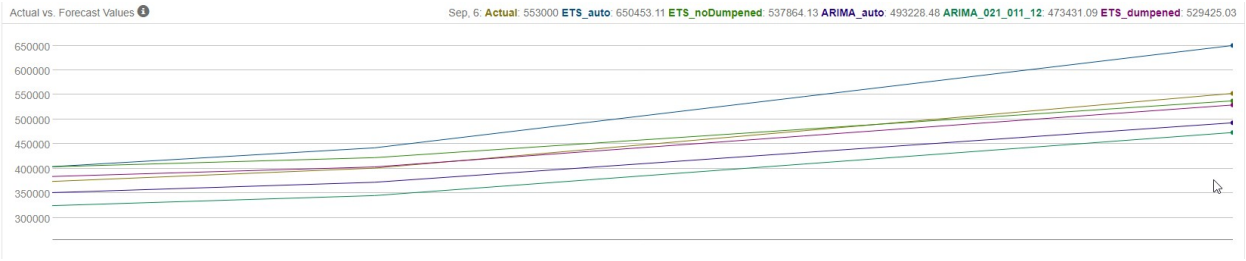
Step 4: Forecast

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.
I have compared all the models:





Record	Report
1	**Comparison of Time Series Models**
2	Actual and Forecast Values:

Actual	ETS_auto	ETS_noDumpened	ARIMA_auto	ARIMA_021_011_12	ETS_dumpened
271000	254853.70905	326115.61258	263228.48013	243251.65147	308342.11249
329000	340280.41766	372831.59175	316228.48013	290717.30878	351709.62935
401000	442291.20116	422306.64685	372228.48013	345165.61813	403424.03551
553000	650453.11029	537864.12937	493228.48013	473431.09227	529425.02729

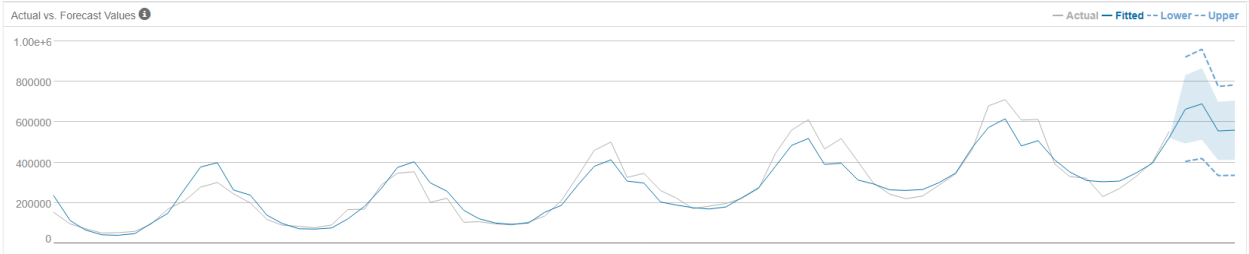
3 Accuracy Measures:

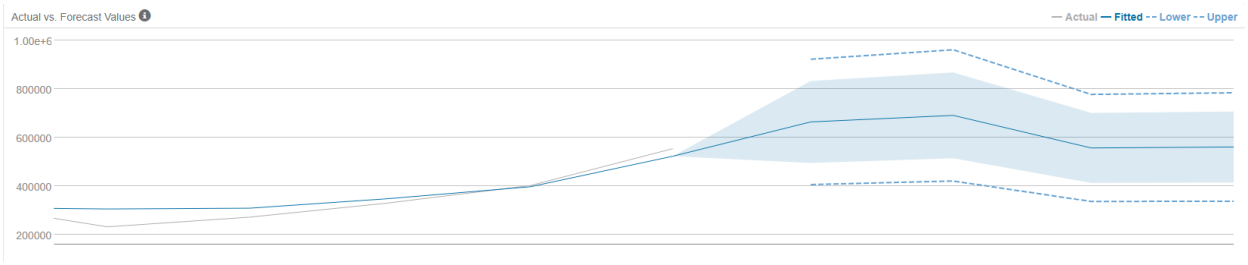
Model	ME	RMSE	MAE	MPE	MAPE	MASE
ETS_auto	-33469.609	53828.48	41542.75	-6.3476	9.3266	0.6904
ETS_noDumpened	-26279.495	37556.65	33847.43	-9.0592	10.4277	0.5625
ARIMA_auto	27271.52	33999.79	27271.52	6.1833	6.1833	0.4532
ARIMA_021_011_12	50358.582	54046.8	50358.58	12.5469	12.5469	0.8369
ETS_dumpened	-9725.201	24858.68	21512.69	-4.2558	6.3874	0.3575

4

Looking at errors (especially MAPE and MASE) we can see that ARIMA_auto and ETS_dumpened are performing similarly and are best ones from all 5 models prepared. ETS_dumpened had tendency to underestimate while ARIMA_auto the opposite. ARIMA_auto MPE and MAPE are the same – it means that this model ALWAYS head overestimated predictions. **ETS_dumpened has the smallest RMSE and that would be my final argument to choose this model.**

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





Period	Sub_Period	forecast_ETS_dumpened_final	forecast_ETS_dumpened_final_high_95	forecast_ETS_dumpened_final_high_80	forecast_ETS_dumpened_final_low_80	forecast_ETS_dumpened_final_low_95
6	10	663108.461584	921263.118605	831906.718831	494310.204336	404953.804562
6	11	689986.705723	960179.780727	866656.461393	513316.950052	419793.630718
6	12	555874.265632	776330.265268	700022.695754	411725.835511	335418.265997
7	1	559828.083032	783142.815814	705845.738286	413810.427777	336513.35025

Period	Sub_Period	forecast_ETS_dumpened_final	forecast_ETS_dumpened_final_high_95	forecast_ETS_dumpened_final_high_80	forecast_ETS_dumpened_final_low_80	forecast_ETS_dumpened_final_low_95
6	10	663108.46	921263.12	831906.72	494310.20	404953.80
6	11	689986.71	960179.78	866656.46	513316.95	419793.63
6	12	555874.27	776330.27	700022.70	411725.84	335418.27
7	1	559828.08	783142.86	705845.74	413810.43	336513.35

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.