The way that we have organized our data and the fact that our index is months of the year shows that we are working with time series. A time series is a sequence where a metric is recorded over regular time intervals. We want to create a model that is able to predict future values based on the performance of the past. This is an extremely powerful tool that could possibly allow companies to reap massive rewards when we are seeing increasing profits but also at the same time show when we are declining numbers. What we need to create is accurate in our forecasts and we will test this accuracy by using the Train/Test method. We will be splitting the data that we already have into two sets: a training set and a testing set. For our purposes, we will be selecting 78% of data for training and the remaining 22% for testing. The reason for 78% is that the date of February 2015 was selected as the start of the testing and everything before would be the training. This means that we will be creating forecasted data after a model is made which will then be compared to the test data which is the actual results already collected. Finally, we will use the mean absolute percentage error (MAPE) to measure the accuracy as a percentage of how accurate the forecast system is.

Since we are dealing with data that is dependent on time as a variable we cannot use linear, logistic, or even polynomial regression for our predictions since we do not two separate dependent and independent sets to test against. We need a prediction model that includes time series in its prediction which is possible using ARIMA modeling as well as its variations SARIMA and SARIMAX. The reason for multiple models is because there are two ways to forecast time series: using previous values of time series to predict future values and using exogenous variables to forecast the time series. ARIMA is the abbreviation of Auto-Regressive Integrated Moving Average and uses its own lags and lagged forecast errors to create an equation to be used to forecast future value. What we discovered previously is that our data fluctuates and resembles seasonal patterns so we will be comparing ARIMA to Seasonal ARIMA model or more simply SARIMA. Finally, we will be looking at the SARIMAX model because we are still analyzing the Movie Playing variable in our dataset which is our exogenous variable. The assumption before we start is that as we go from ARIMA to SARIMA to SARIMAX our model will improve prediction accuracy.

Before building our ARIMA model, we need to find the p,d, and q values. The p value is the order of the Auto Regressive term or the number of lags of Y to be used as predictors. The d value is the number of differences required to make the time series stationary. The q value is the order of the Moving Average term or the number of lagged forecast errors that should go into the model. Finding each value requires running three separate tests that determine what

order each value should have. Fortunately, there is a method in Python that will automatically pick the best p,d, q, and seasonal values by comparing them to all other possible combinations. After the function runs and picks the best model we can now view the residual plots to see how well a fit it is to the data.

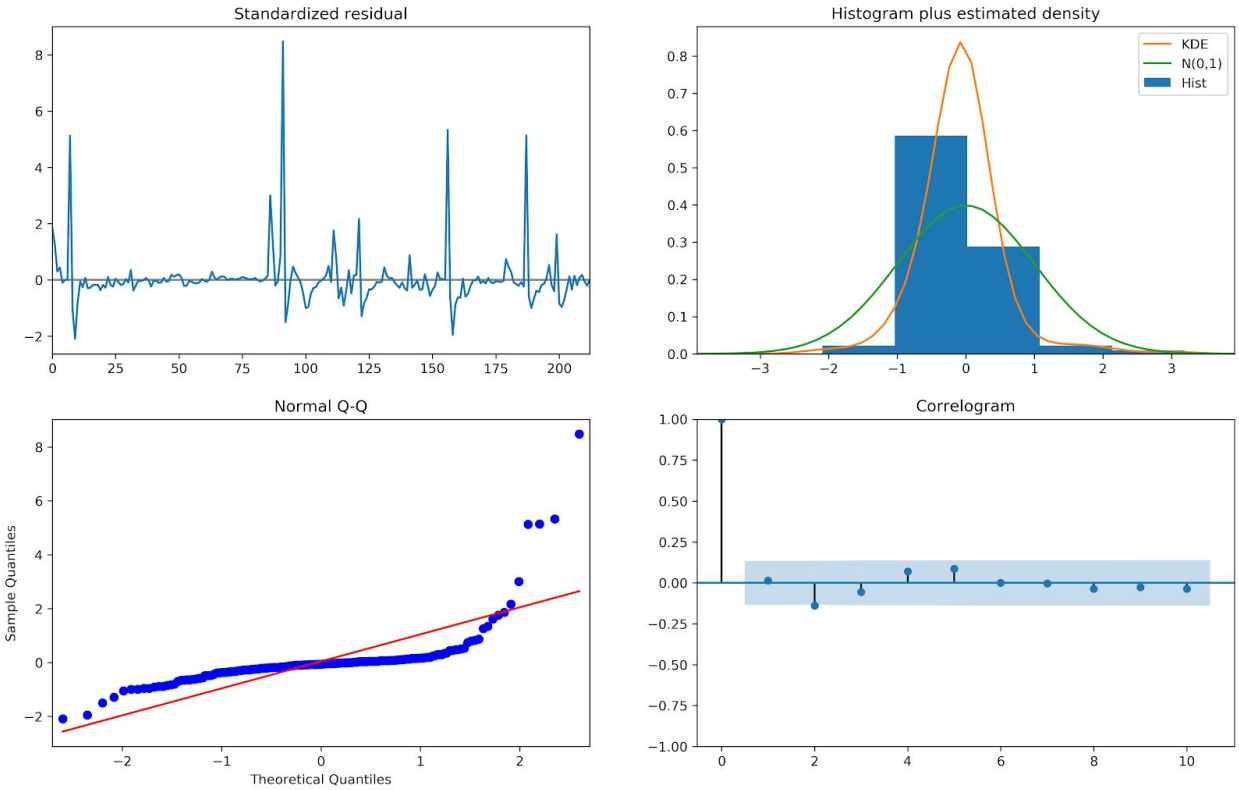
The function plot diagnostics will then give us four separate graphs that coincidentally we would be using if we were trying to find the parameters manually. The four different graphs are Standardized residual, Histogram plus estimated density, Normal Q-Q, and Correlogram. For the Standardized residual we want the values to fluctuate around zero and have a uniform variance. For the Histogram plus estimated density, we want a normal distribution with mean zero as well as having KDE and N(0,1) as close as possible to each other. For the Normal Q-Q, we want the blue dots to be as close to the red line as possible with points farther from the line being considered a skewed distribution. Finally, for the Correlogram, we are checking for randomness in the data and want the values to be as close to zero as possible. When we are running SARIMA and SARIMAX models, we employ seasonal differencing so that instead of just p, d, q values we got for the ARIMA values we now have three additional terms written like SARIMA(p,d,q)x(P.D, Q) to take account for these seasonal patterns. Unlike ARIMA which analyzes previous data from consecutive terms, SARIMA and SARIMAX analyze data from previous seasons and require more computation power to discover the best model. Finally, the SARIMAX model is identical to SARIMA with one key difference, we will be using the Movie Playing variable as the external predictor. In order to use the variable, we need to know the value during the forecast period and since our testing period starts in February 2015 and will conclude in December 2019, we know when movies will be playing for this time period.

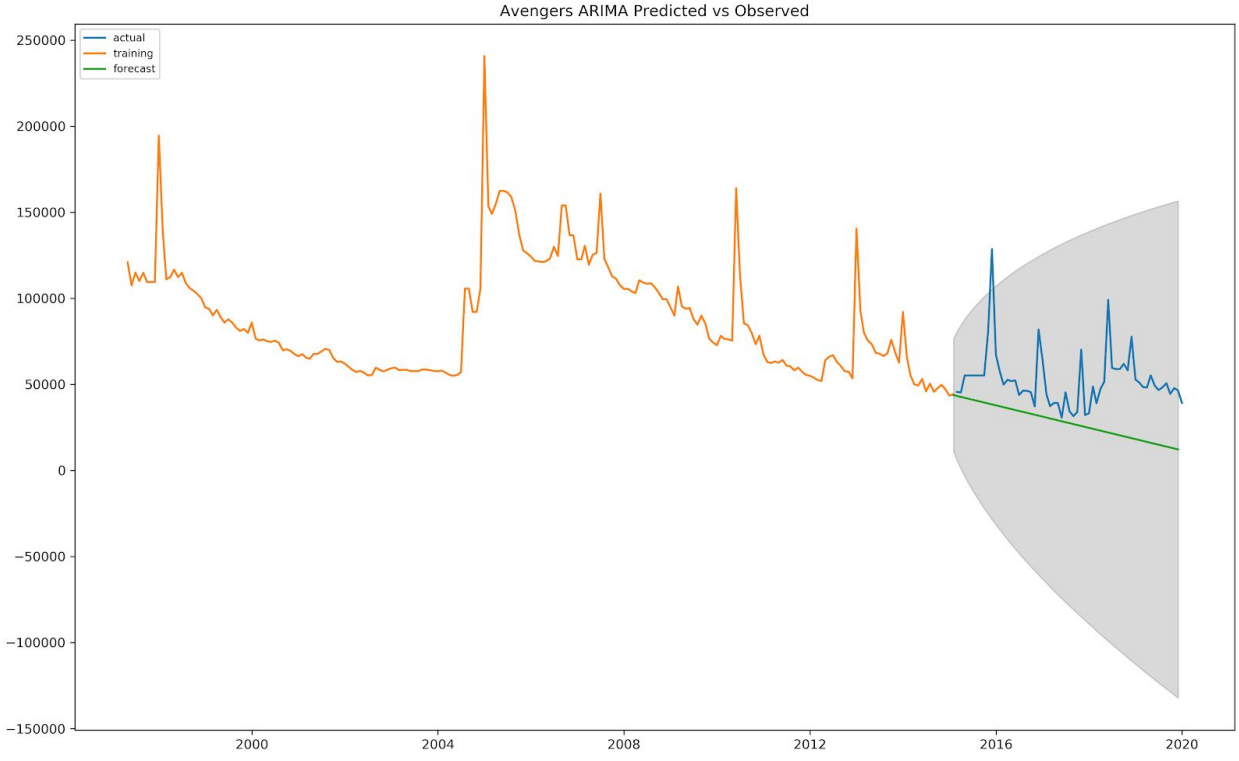
Now that we have gone through the steps of forecasting our data and comparing it to the actual data, we can now analyze the results. The ARIMA models that we have for all five groups all have one thing in common that gives an indication of the shortcomings of the method we have applied. The data that we have collected fluctuates wildly with large spikes in the sales but the forecast that we are getting is very linear. This shows us that ARIMA models can be effective for more simple trends of increasing or decreasing data which is not what we have, unfortunately. It does, however, seem to do a good job of predicting the general direction of the sales without taking into account the fluctuation of our data so it is stepping in the right direction. Also, when we are looking at our Normal Q-Q plots we can see we have a lot more skewed distributions compared to later models. This means that this data exhibits seasonal patterns that were not known at the time. Looking at the predictions for SARIMA and SARIMAX shows

that even these models have a hard time predicting the big spikes and what we are given are small fluctuations.

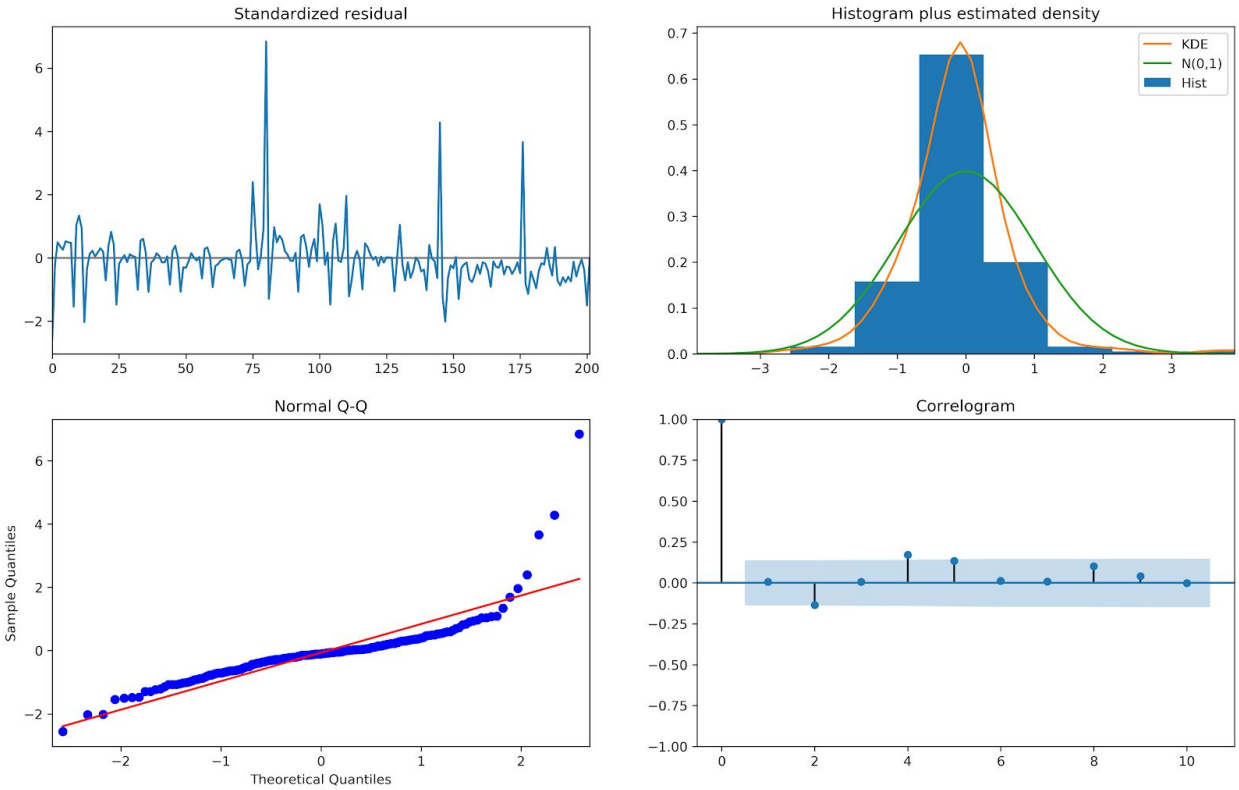
Out of all the five different franchises, the one that sticks out the most is Justice League because our predictions become negative numbers. However, upon closer inspection, we can see that Justice League sales were decreasing at a greater rate compared to other franchises and not leveling out as others did. Accuracy for the models was a mixed bag with Justice League being the most inaccurate due to its extreme prediction for SARIMA and SARIMAX models but Spider-Man SARIMAX model also gave us the most errors as well. What's most surprising is that when we calculate the average of the MAPE across the different models we find that ARIMA is the most accurate even though sales aren't very linear. Avengers and X-Men were also the only franchise that had the SARIMAX model be the most accurate for the group. Finally, the most accurate model was Batman's ARIMA model which even though it had wildly fluctuating data was able to correctly predict the average change of sales over a period of time. It would seem that using Movie Playing as the exogenous variable did not work as well as would be liked. The possible exogenous variables could be comic book relaunches, comic book special events, and specific comic book writers. Further investigation could be done to find how to better model the data but the predictions we received are encouraging.

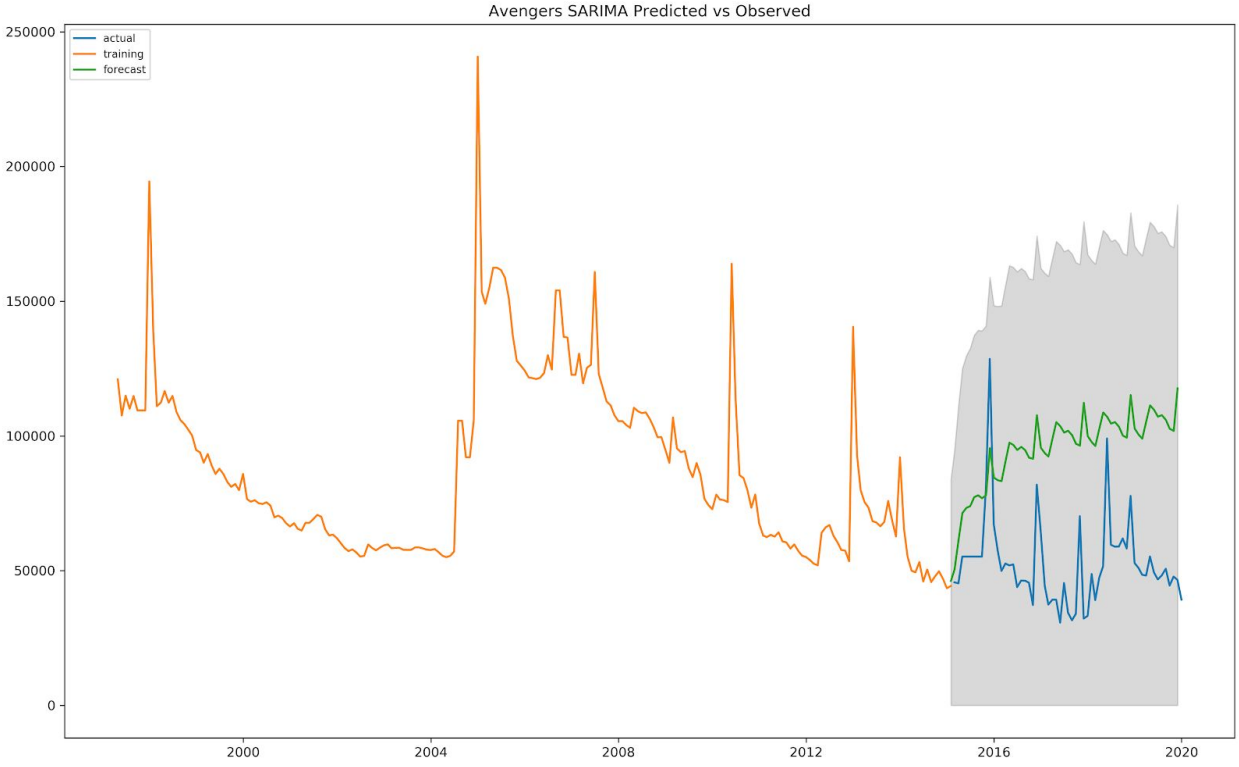
Avengers ARIMA Diagnostics



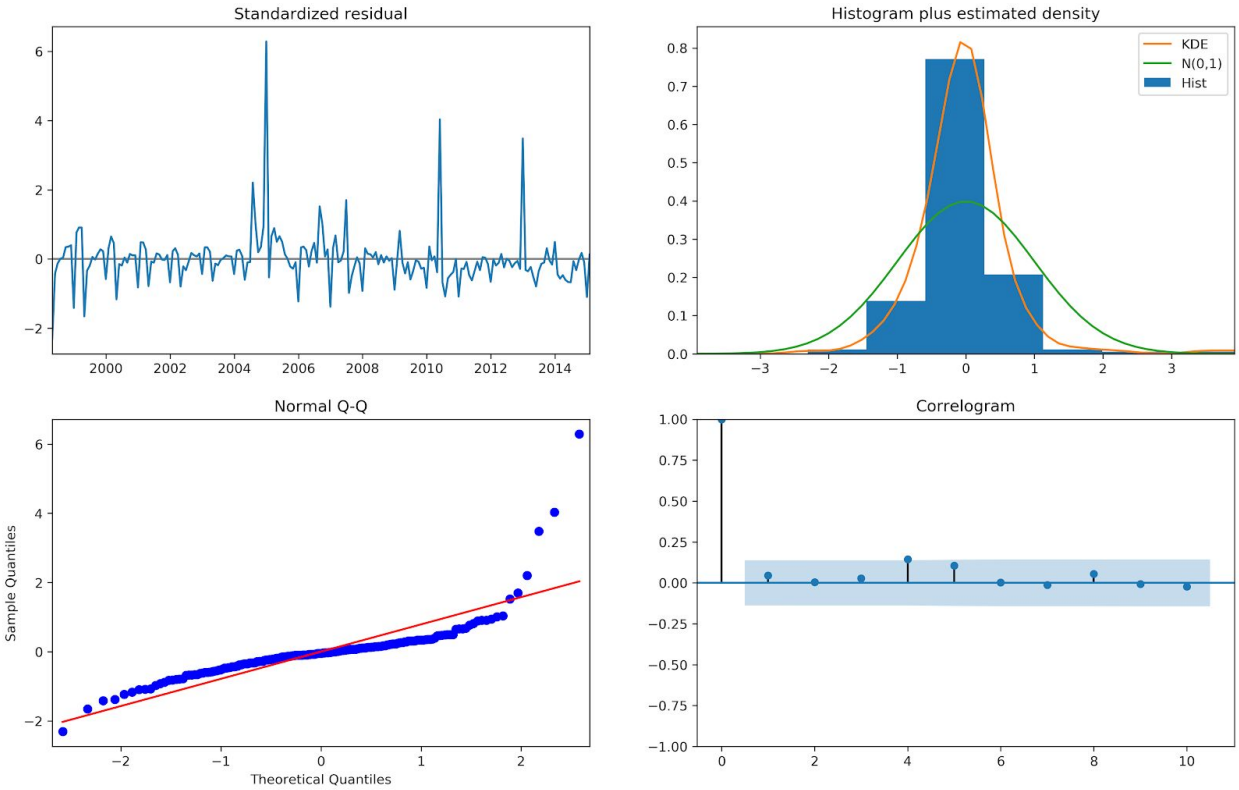


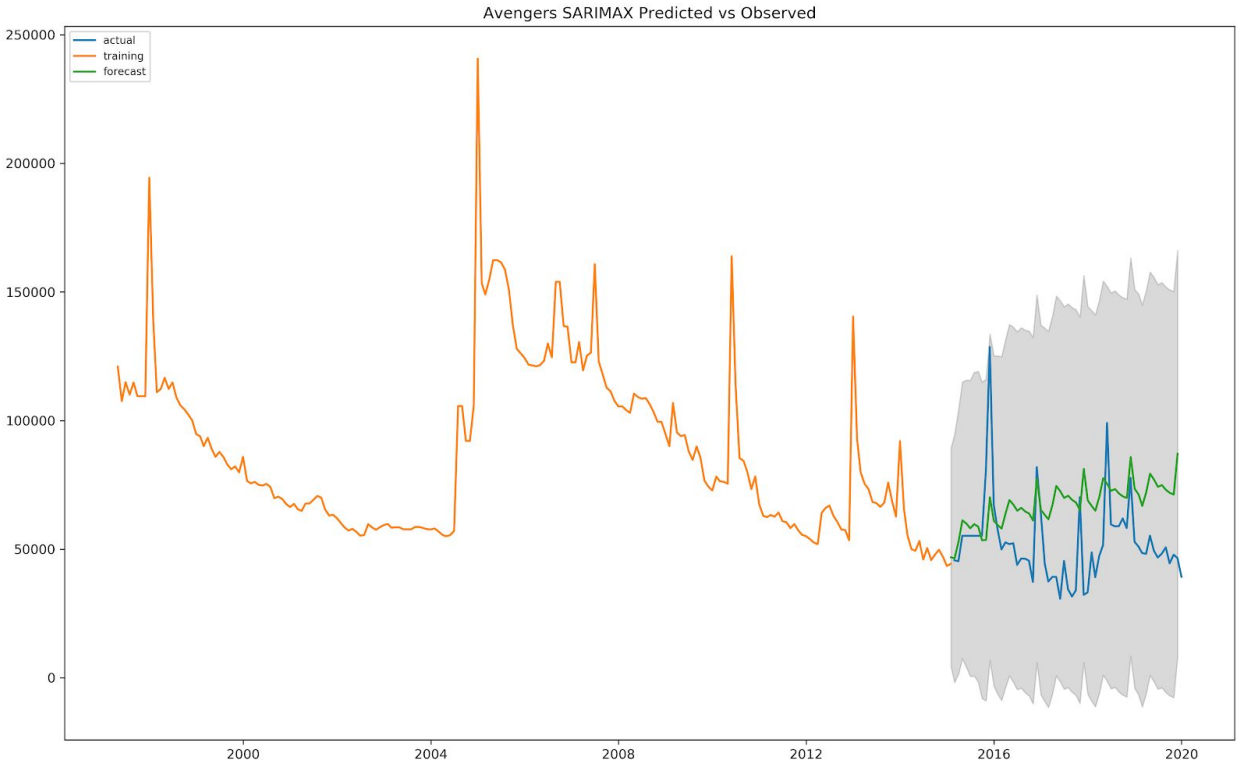
Avengers SARIMA Diagnostics



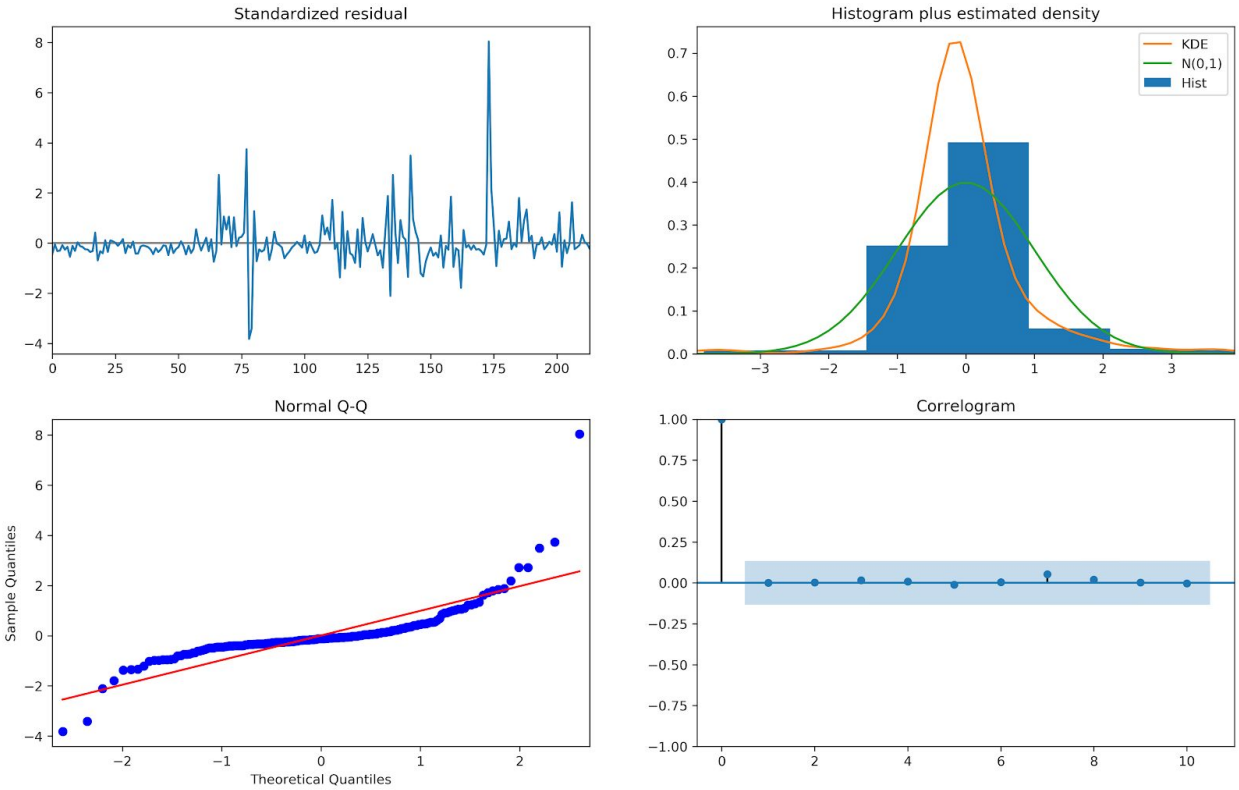


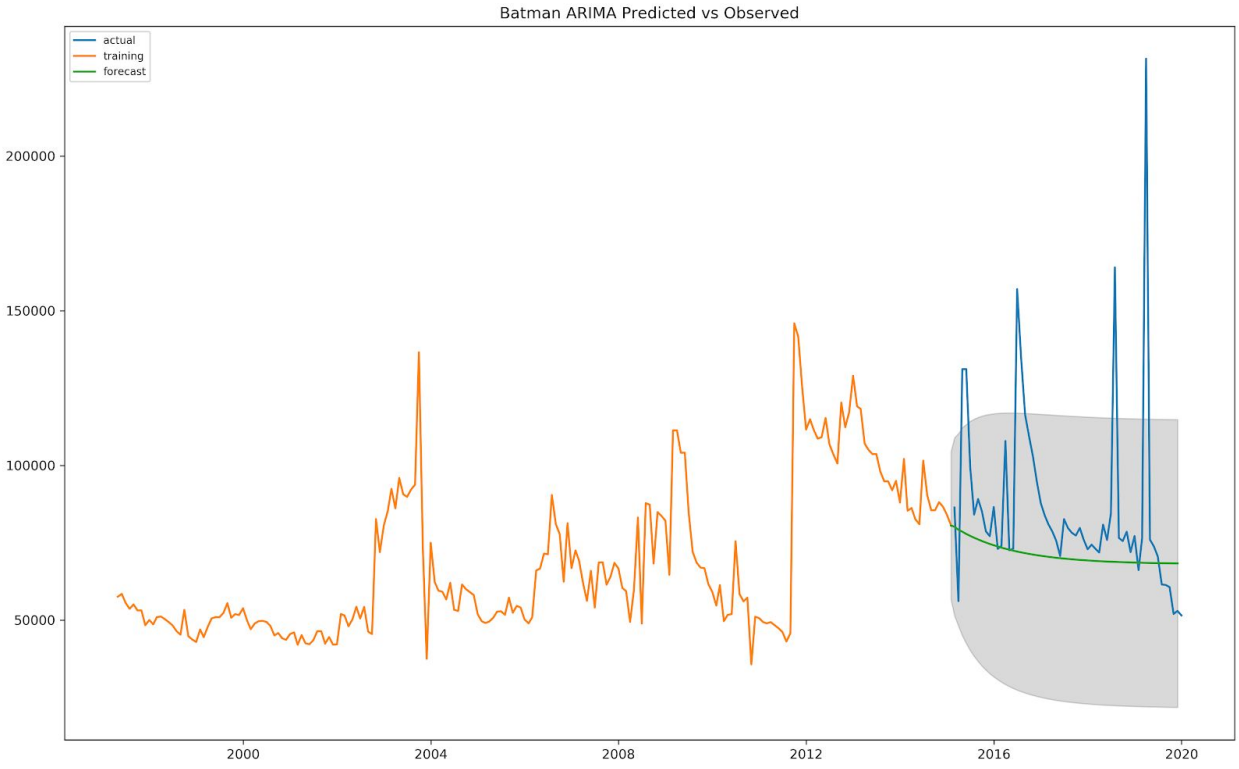
Avengers SARIMAX Diagnostics



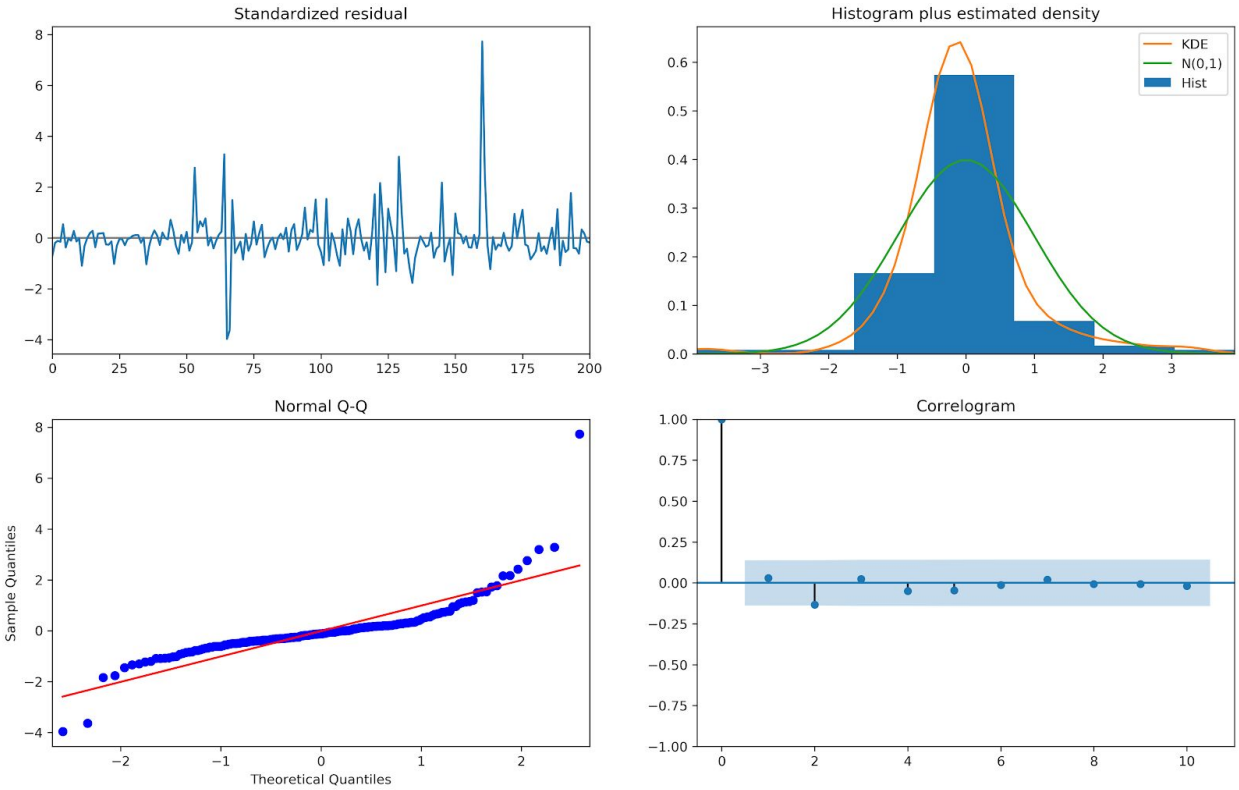


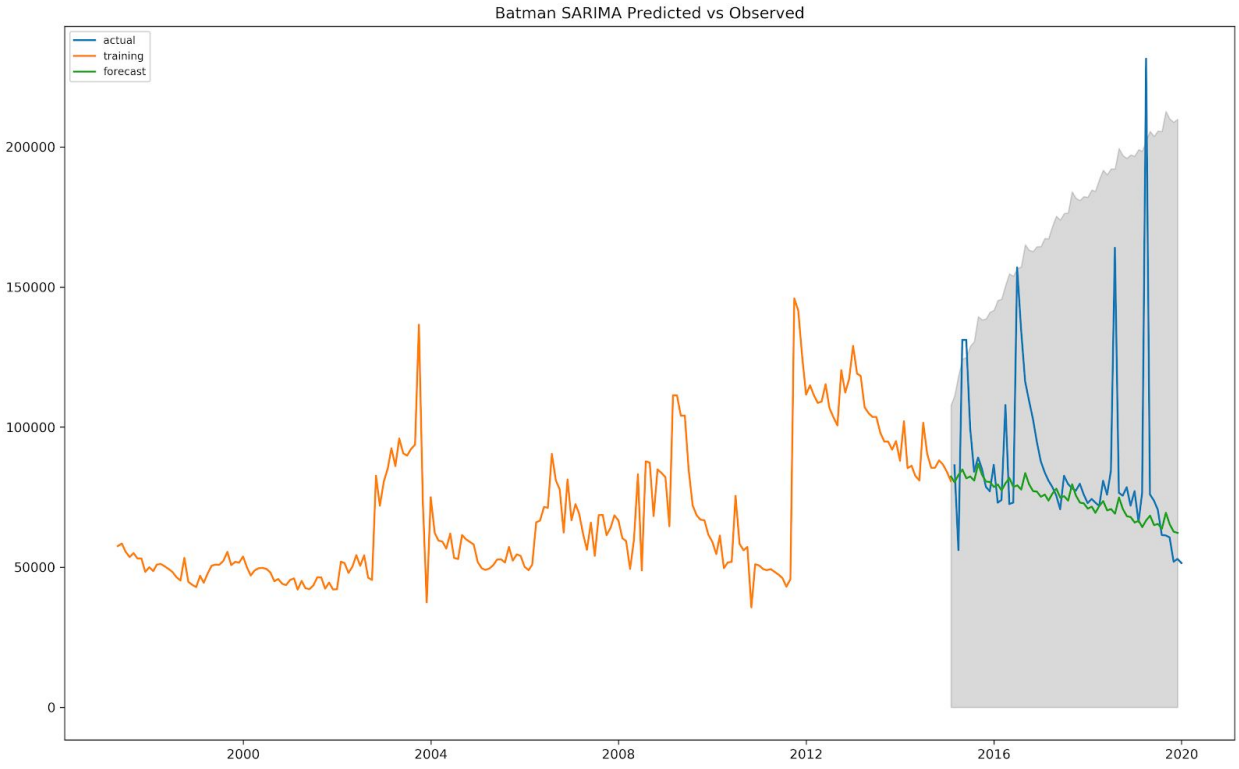
Batman ARIMA Diagnostics



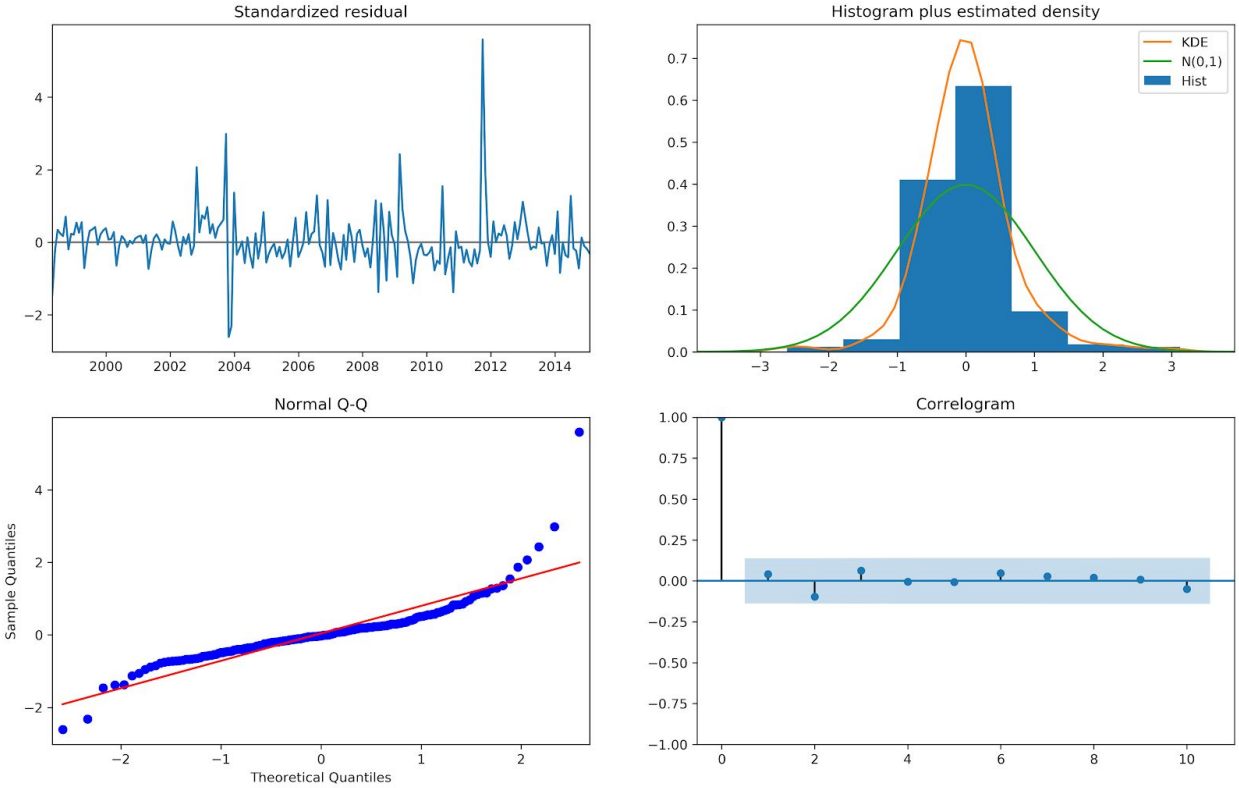


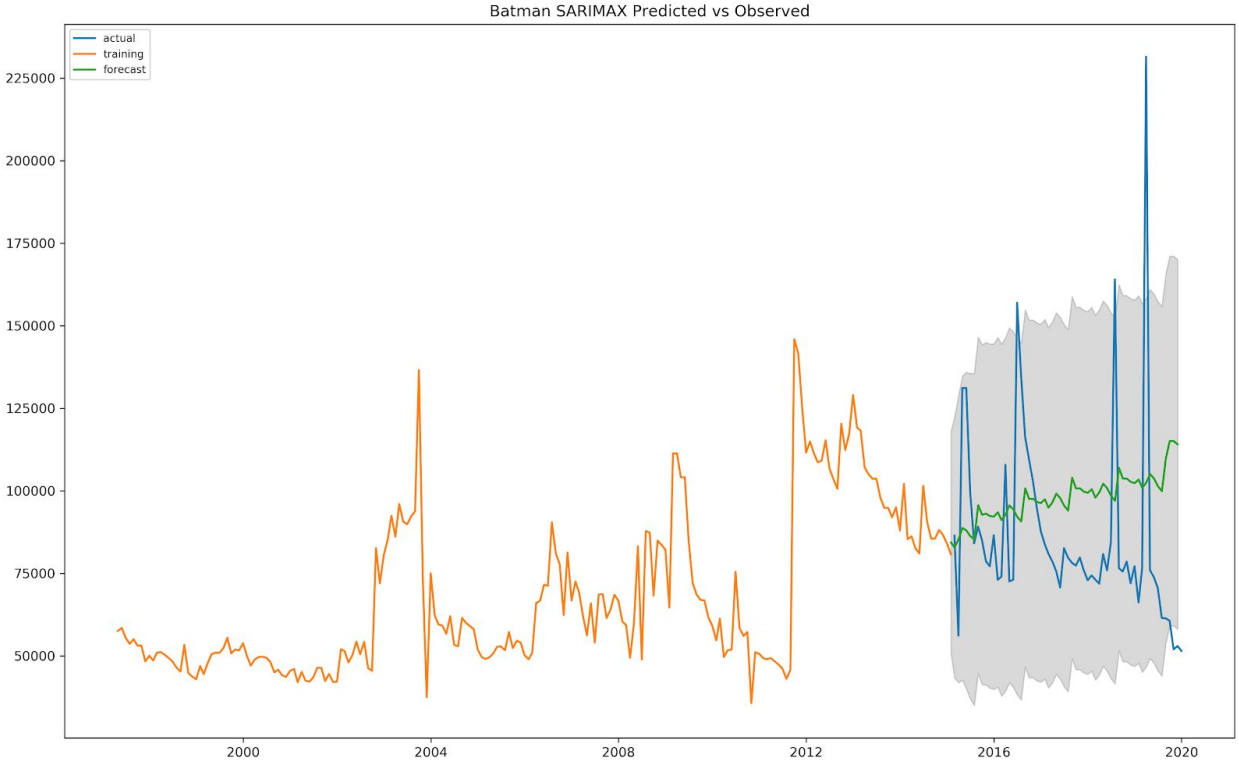
Batman SARIMA Diagnostics



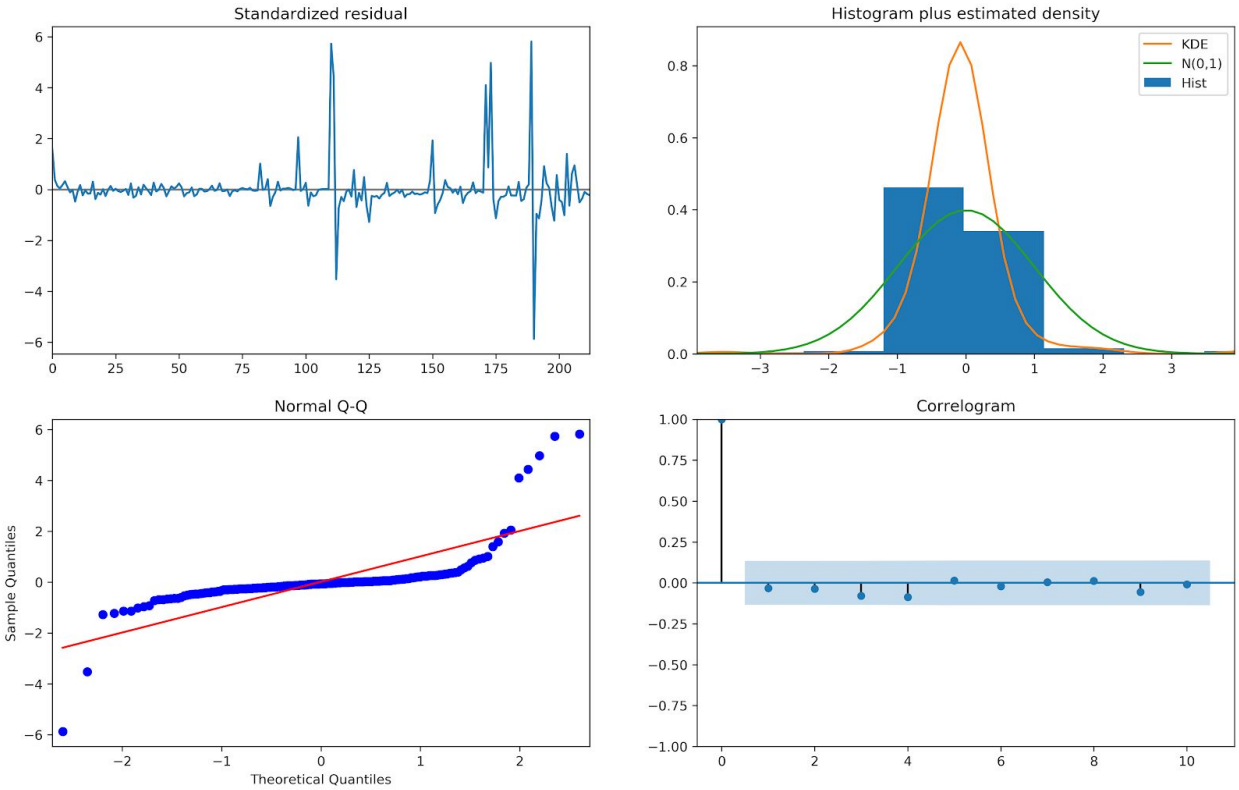


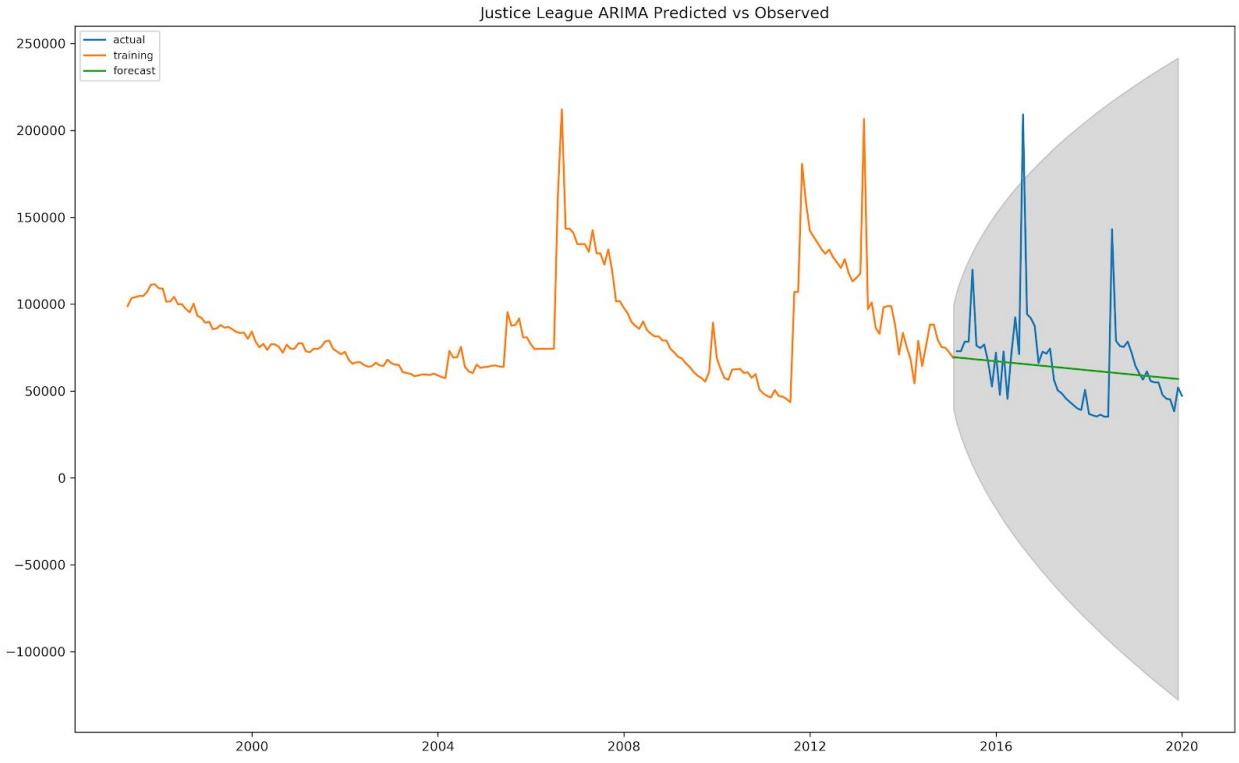
Batman SARIMAX Diagnostics



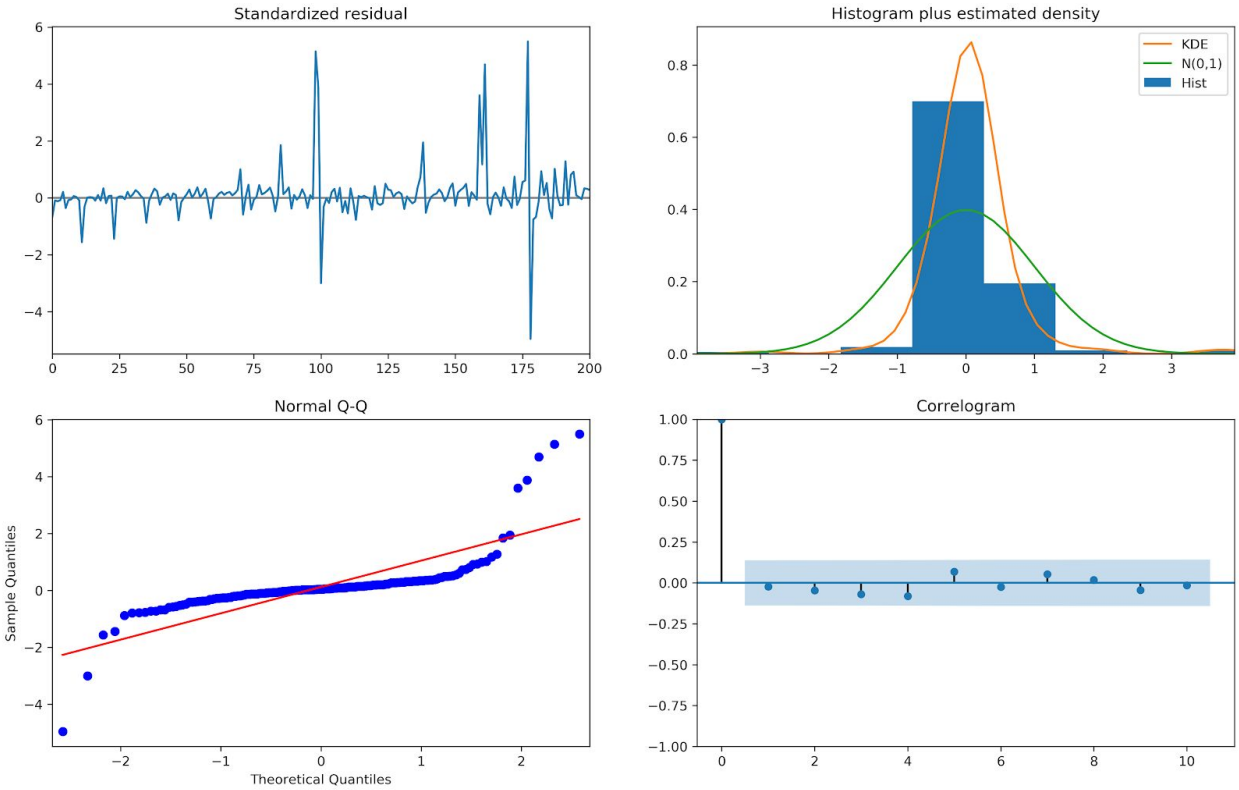


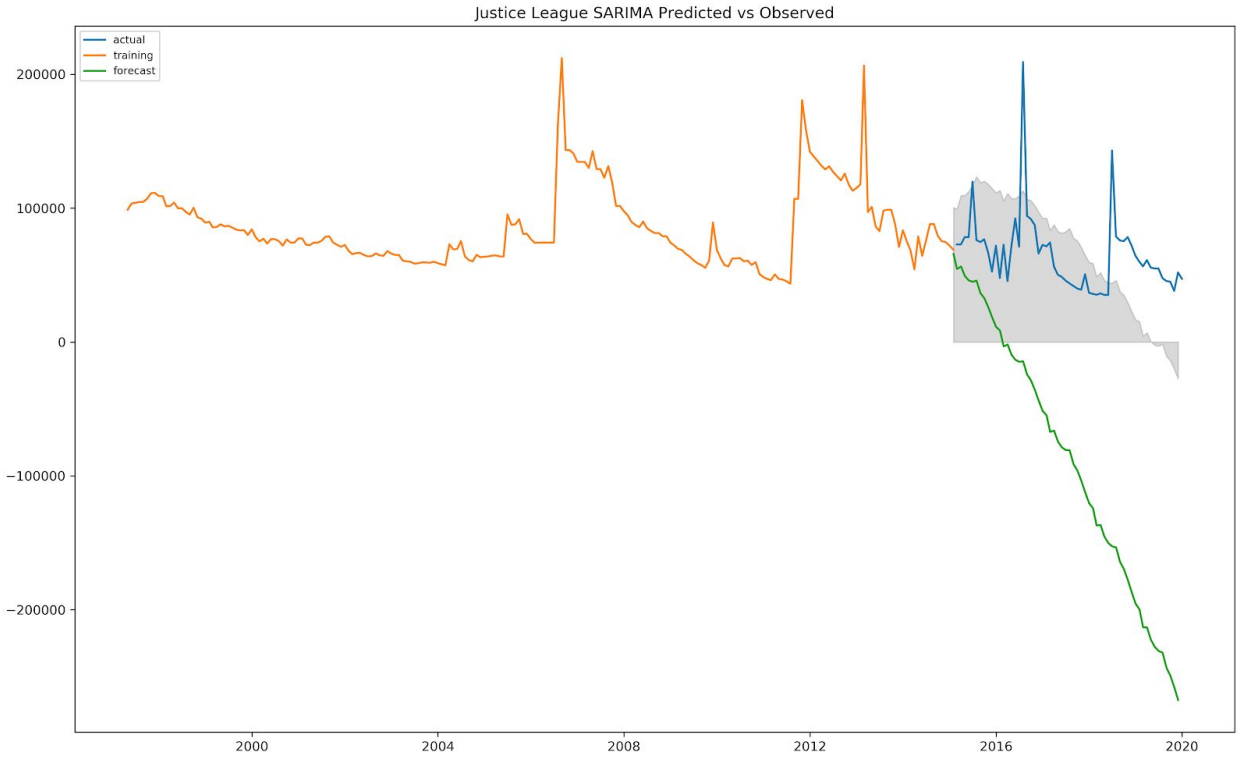
Justice League ARIMA Diagnostics



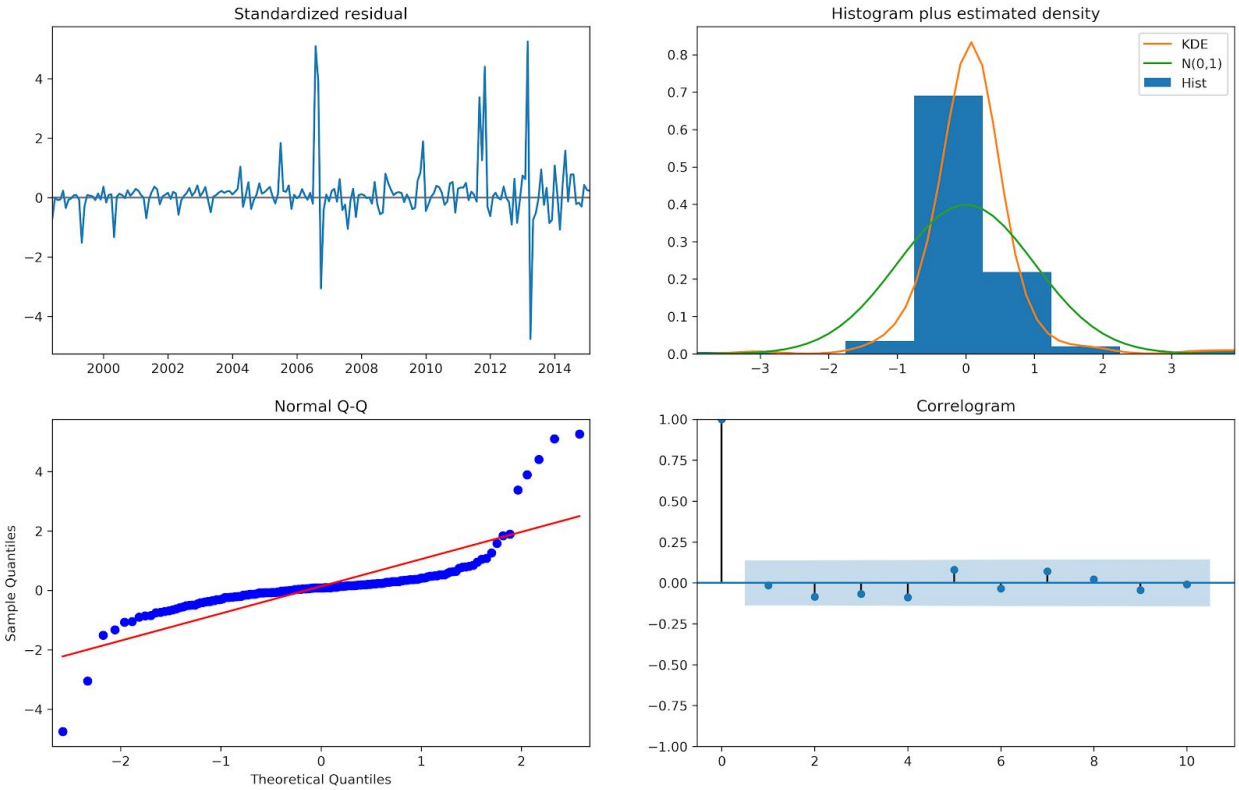


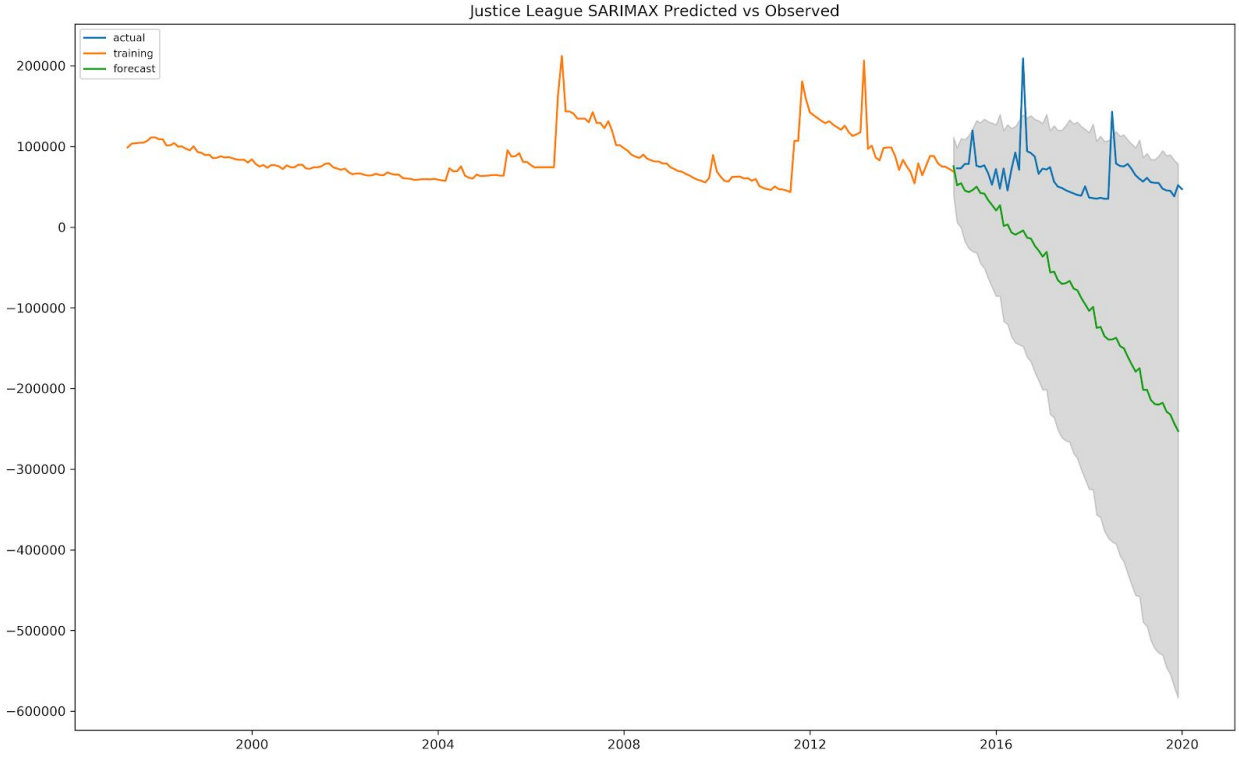
Justice League SARIMA Diagnostics



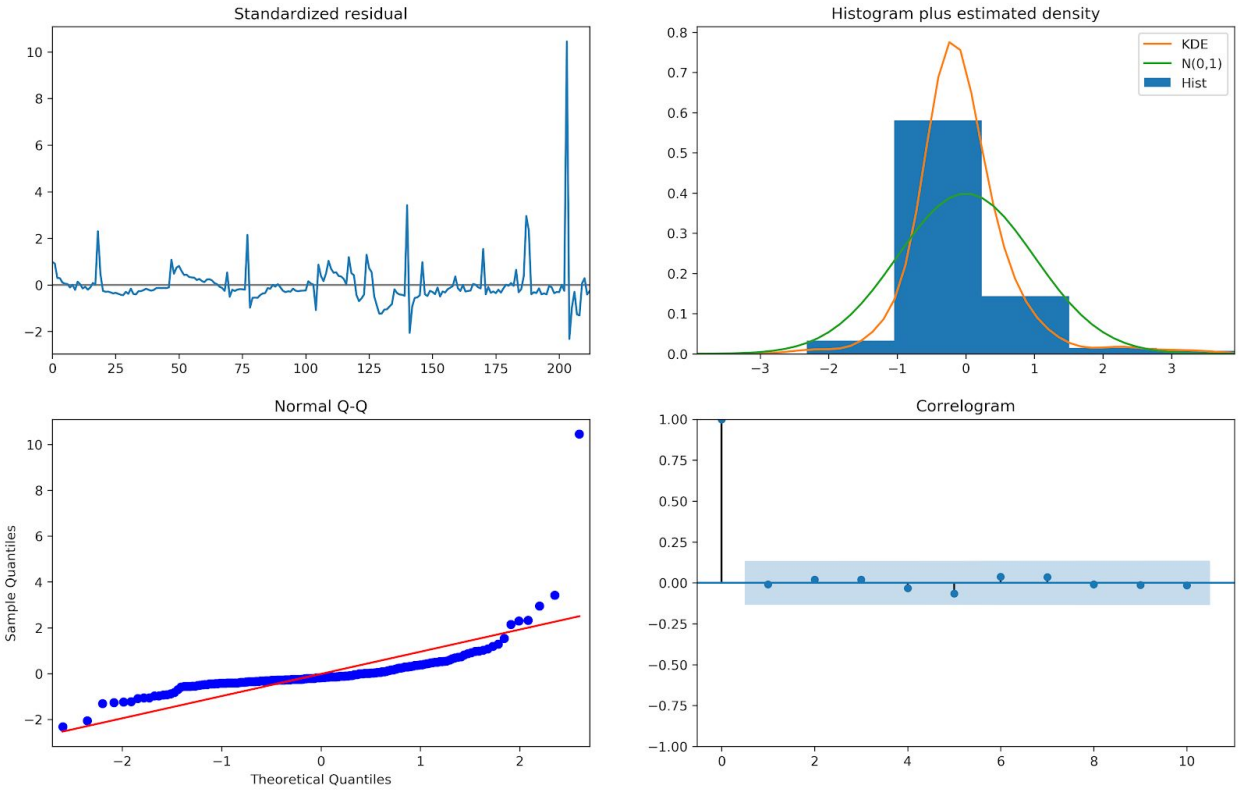


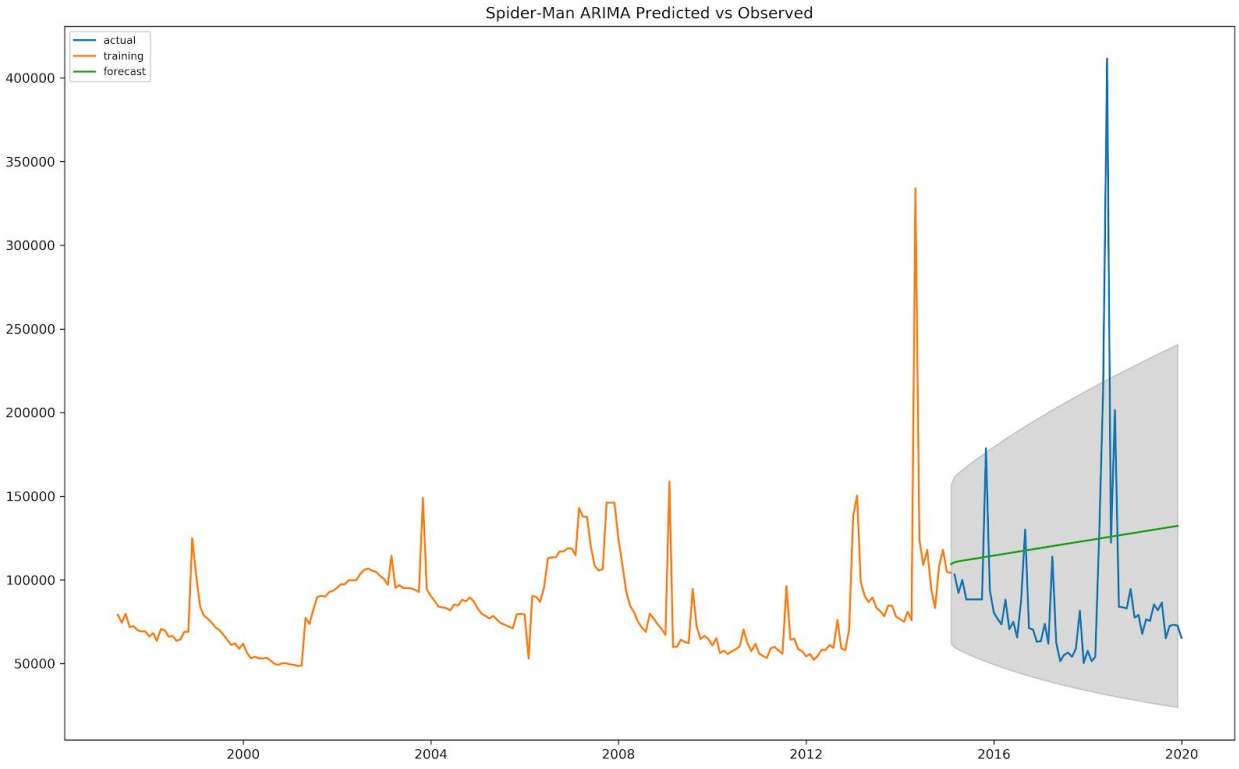
Justice League SARIMAX Diagnostics



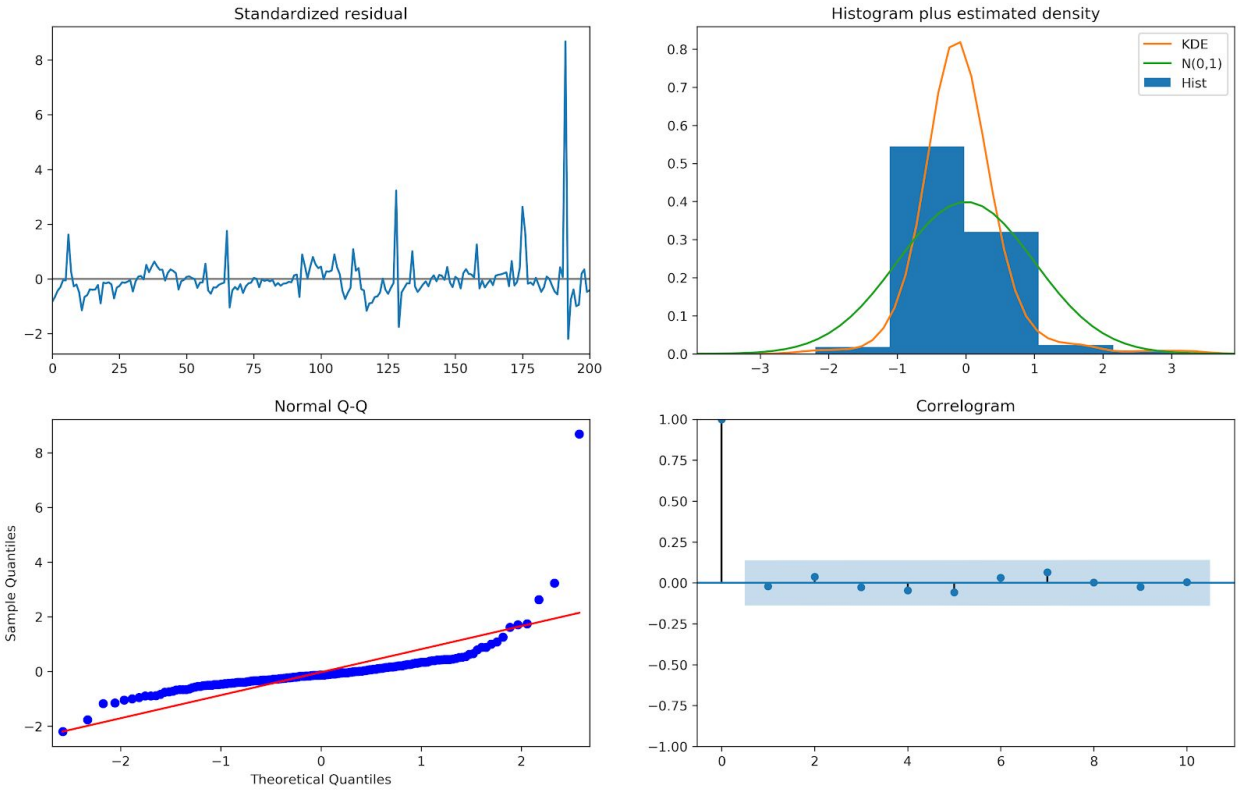


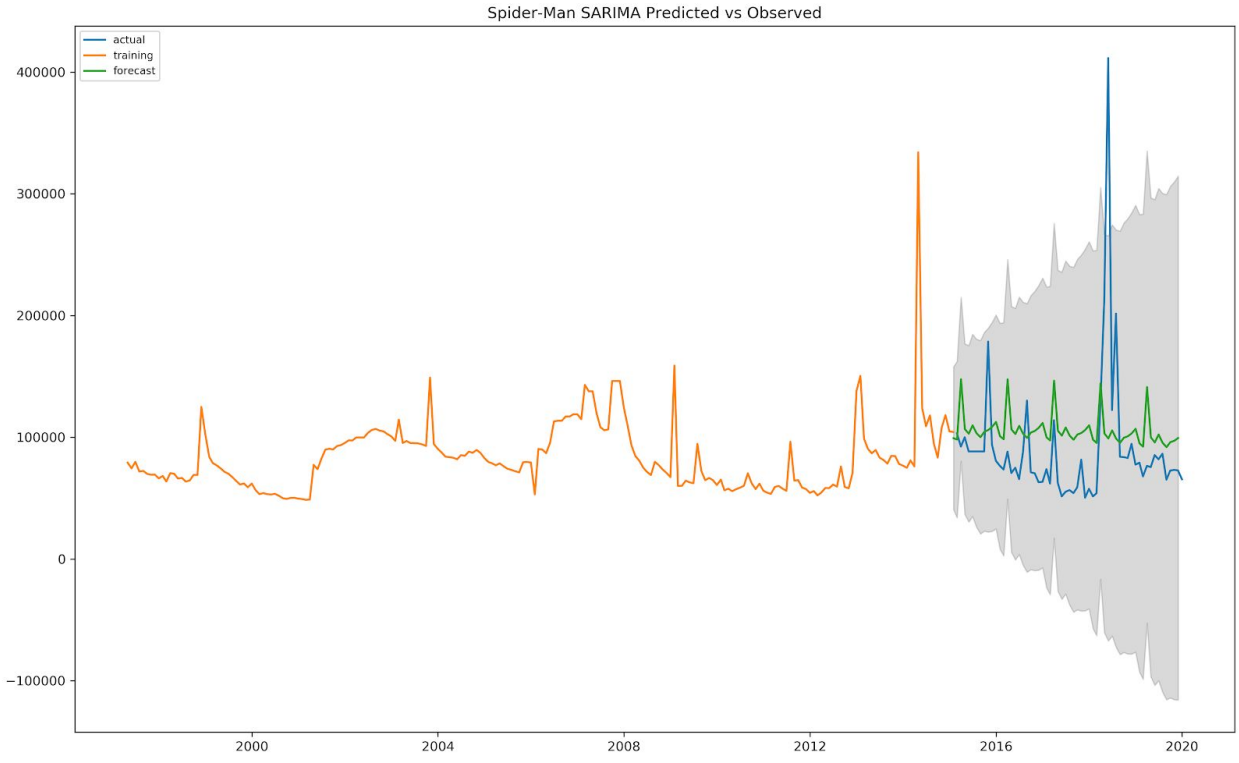
Spider-Man ARIMA Diagnostics



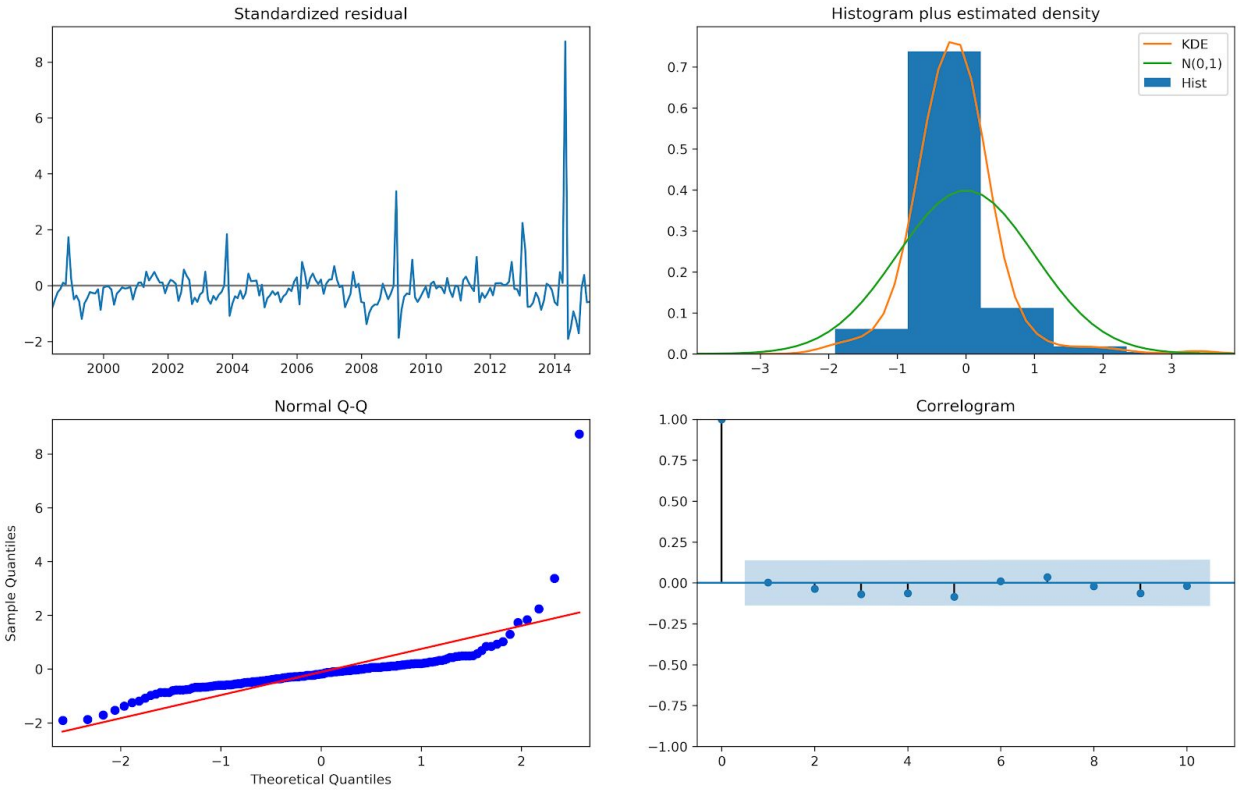


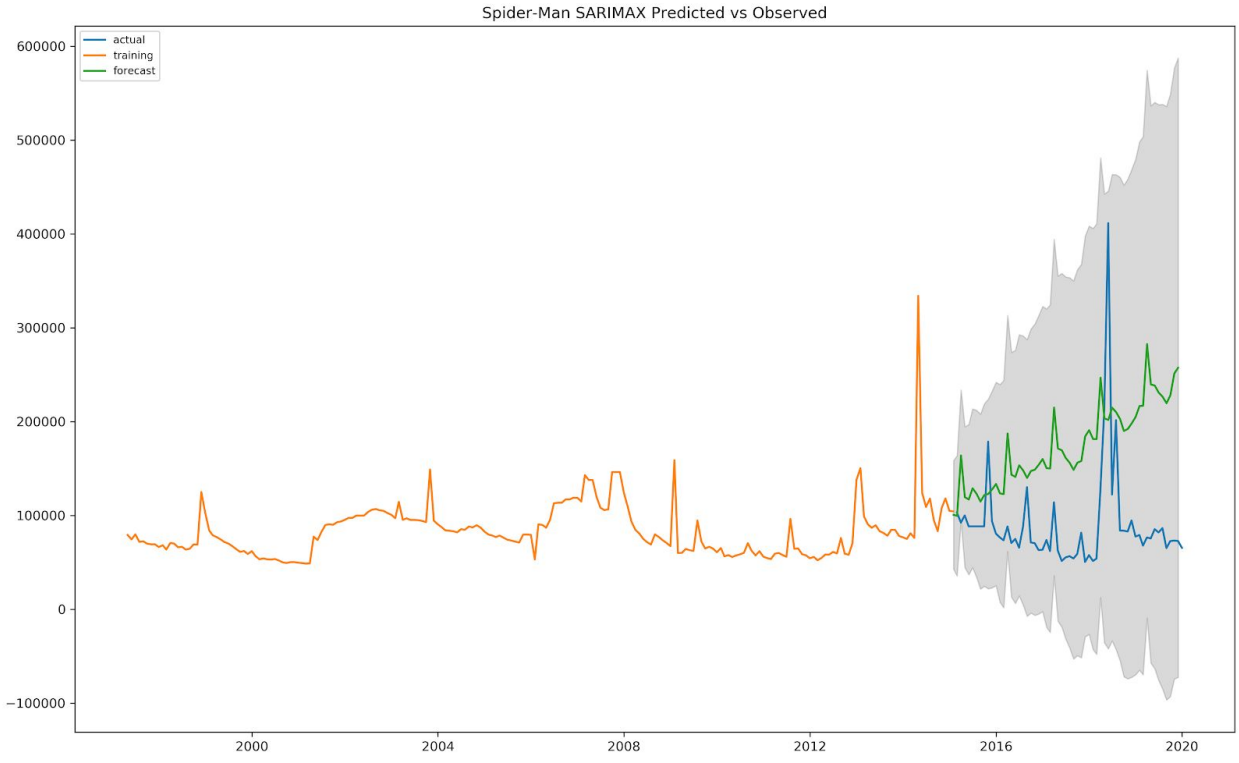
Spider-Man SARIMA Diagnostics



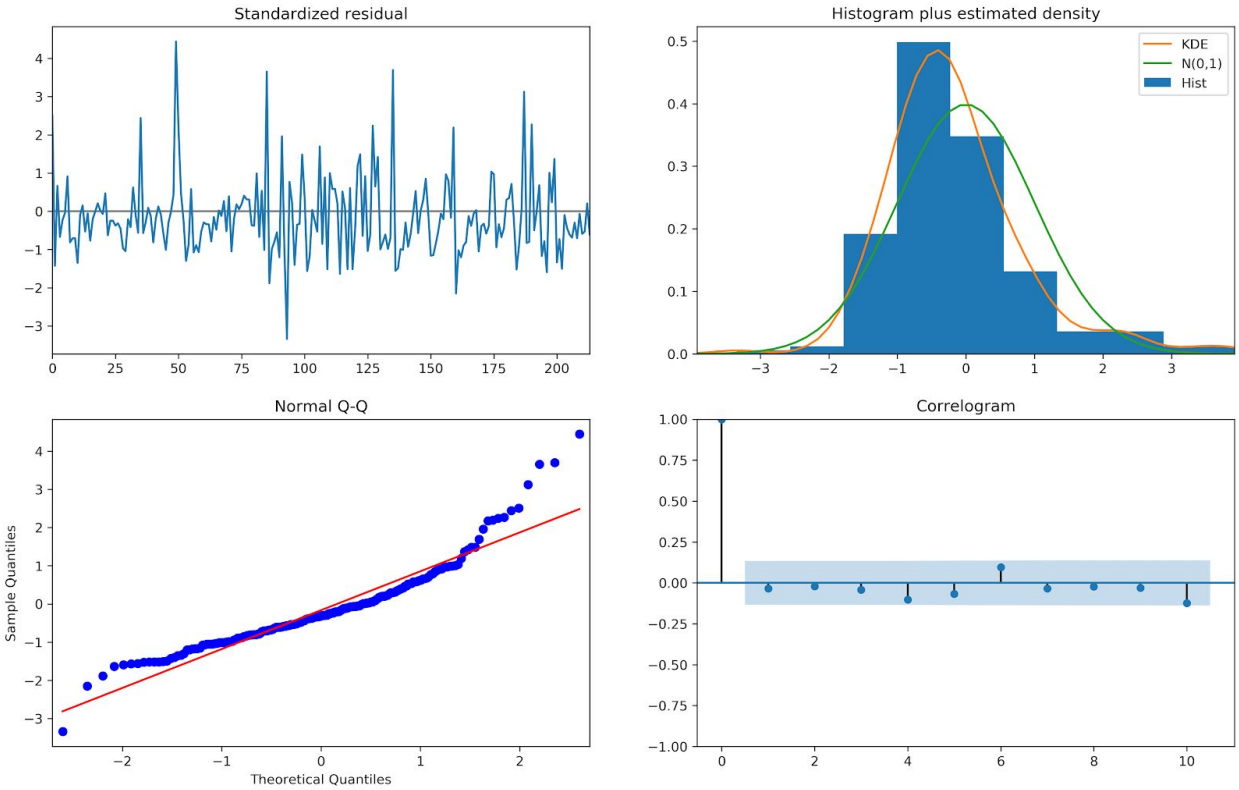


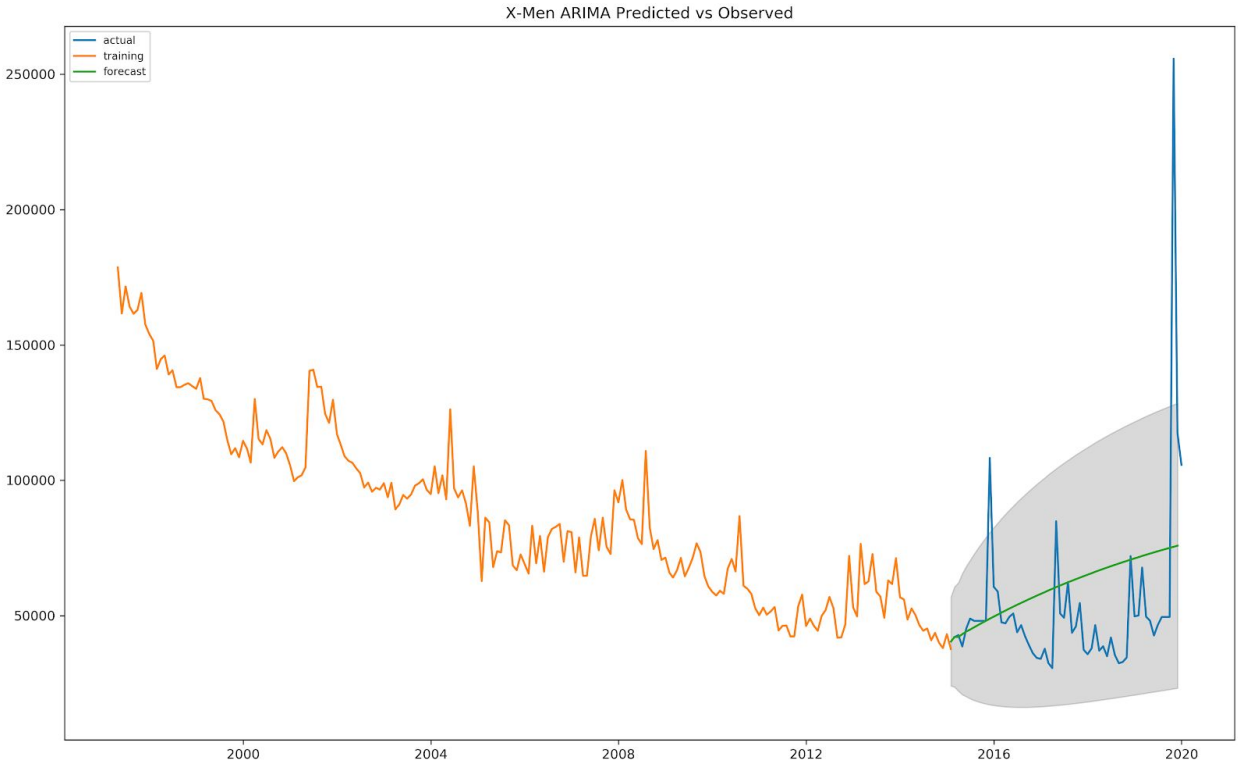
Spider-Man SARIMAX Diagnostics



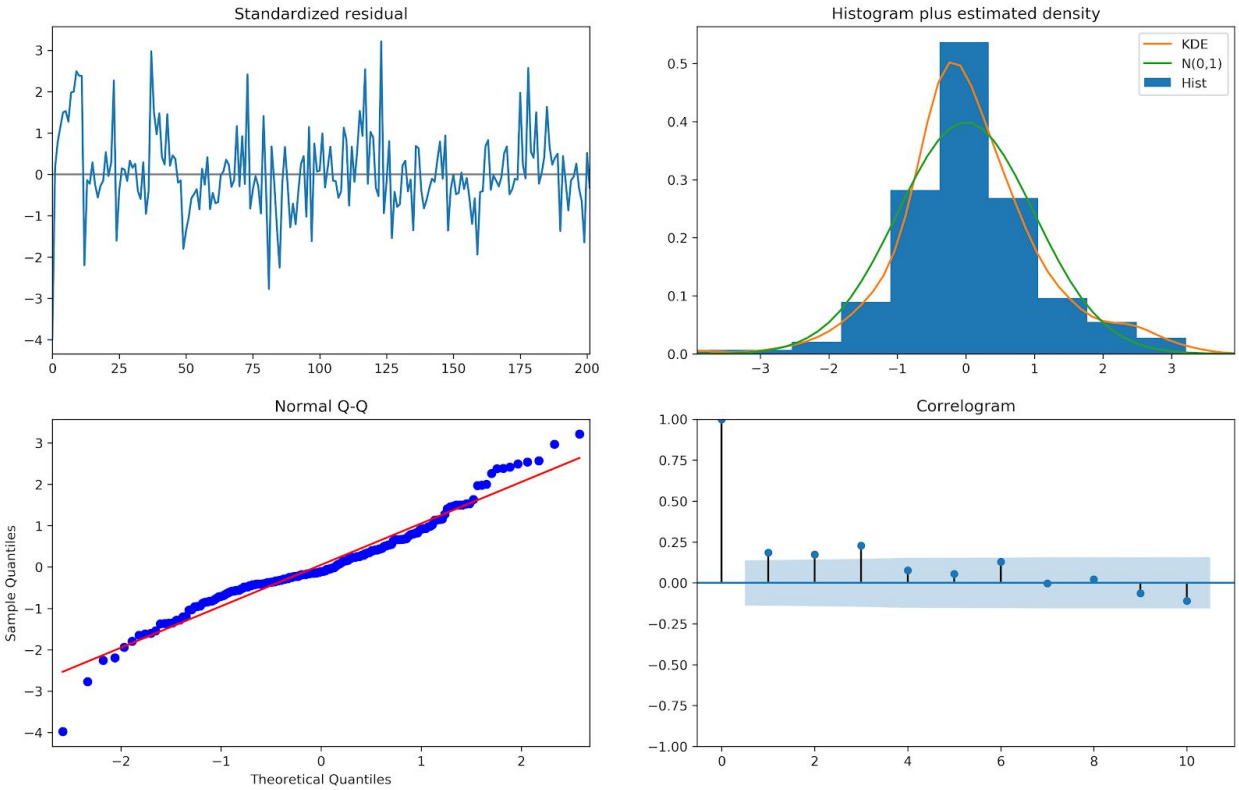


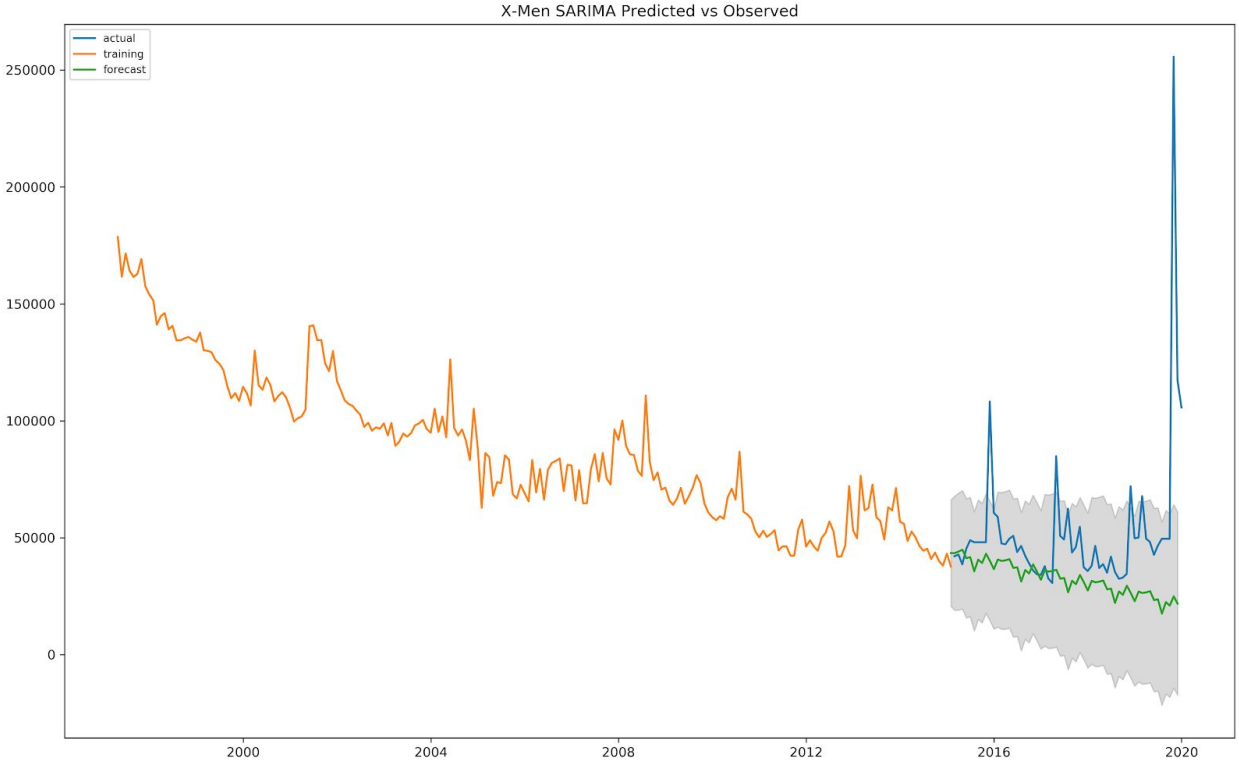
X-Men ARIMA Diagnostics





X-Men SARIMA Diagnostics





X-Men SARIMAX Diagnostics