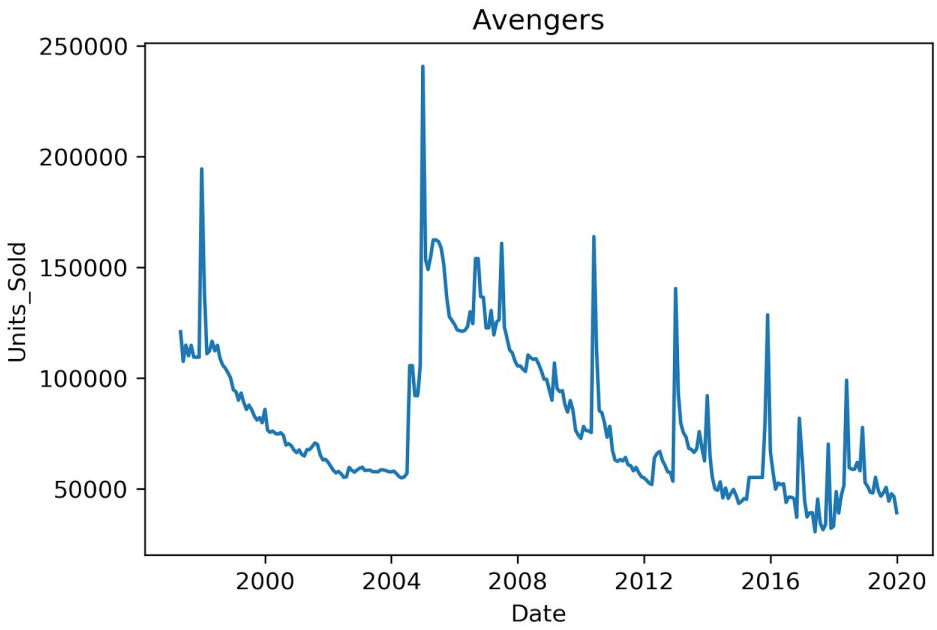
This capstone project's goal is the investigation of comic book sales and the effect movies have on its sales. Superhero comic books have been around since the 1940s with the creation of Superman and Batman by DC Comics as the top sellers but also had a big boom as well in the 1960s with the creation of the X-Men, Spider-Man, and the Avengers by Marvel Comics. It wasn't until the late 90s and the early 2000's that movies were beginning to become huge blockbuster hits that piggybacked off the brand awareness from the comic books. Movies that were based upon comic book characters that were decades-old were reliably breaking 100 million dollars or greater between DC and Marvel Comics. DC struck first with the success of Superman and Batman with the movies being produced by Warner Brothers while Marvel had its characters split up between three different companies. The X-Men franchise was being handled by Fox, Spider-Man by Sony, and the Avengers by Marvel Studios. Sales for these movies are among the best selling movies of all time so being able to produce these movies is a highly profitable venture. Comic books while not as profitable as these big-budget movies are a reliable way of building an audience that can be continually consuming various types of products about the subject that they are reading. While it is pretty obvious that strong comic book sales led to strong movie ticket sales at the box office another useful question to answer is how comic book sales are affected by the movies playing during the same time. While we are analyzing comic book sales and movies based on comic books, we can also relate our findings to any sort of intellectual property that has been created on one medium and being marketed in the future with another medium. For example, examining movies based on video games and books and the effects it has on the original medium.

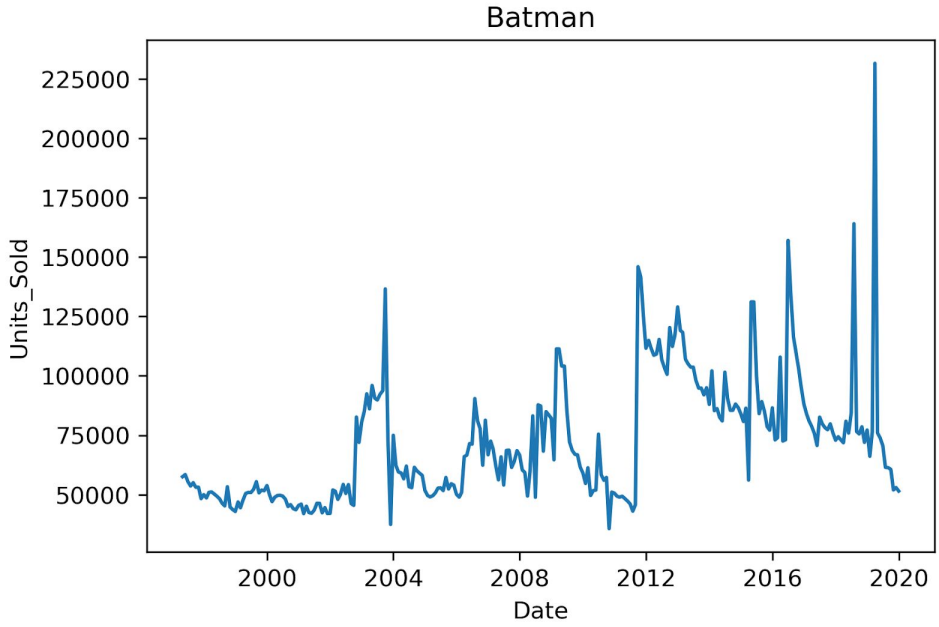
Data obtained for this project was acquired from the website Comichron which has records of physical comic book sales which reliably start in April 1997. The reason for this is due to the fact that was when a single distributor handled the vast majority of sales in stores and data before that had not been reliably collected. Unfortunately, the data itself is organized and maintained only on webpages with embedded spreadsheets and not in a format that we can readable use. In order to use this data, it had to be exported to Excel through the website's server which took a while due to the enormous amount of records that it had. When we are talking about monthly sales of comic books, we are talking about at least 300 different titles for each month for the last 24 years. So while we now had all the sales we needed, we also had had a lot of information we did not need. To figure out what we needed to keep and discard, we first needed to determine which movie franchises to focus on. What we had found so far about

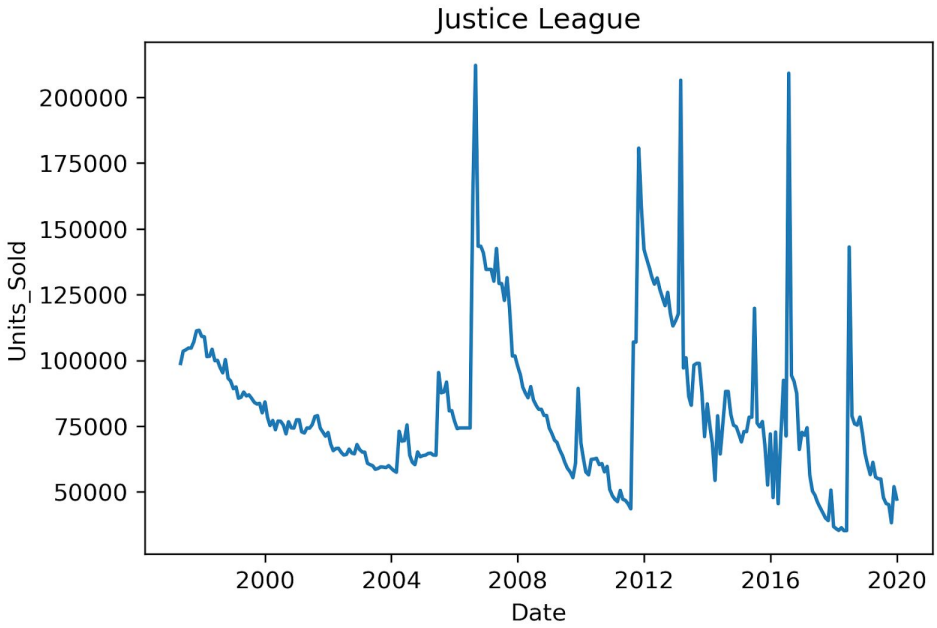
the movie franchise that already exists is that we have four different movie studios producing comic book movies. They are Sony with their Spider-Man movies, Fox who did X-Men movies, Marvel did the Avengers and its spin-offs, and Warner Brothers did both Batman and its newer Justice League spin-off movies. So now we have five different groups to analyze: Batman, Spider-Man, X-Men, Avengers, and Justice League. We will be looking at the main comic titles for these franchises that are released every month and will be excluding any titles that did not release according to this schedule. Also, we will be excluding any spin-off titles that belong to the group to focus on the main ongoing titles that are being released. Based on these criteria, we are now able to filter our current results and create CSV files that will be able to be loaded into Python. For the data itself, we have a ranked list every month that sorts comic books by its unit sales. For our purposes, we will be looking at comic book titles, unit sales, and the month they were released. For each of our five groups, we will have a single CSV file that contains all our data all issues released during that time. With the comic book data now being obtained, we also need to acquire some movie data for when each franchise had a movie playing. Using the website Box Office Mojo, we can get a list of the movies playing for each franchise and the exact months they were playing in the theaters. Adding a column called Movie Playing in our dataset we set each month to either a 0 or 1 value depending on if a movie was playing or not. Now that our data has been loaded in Pandas as a database, we need to clean our data so that it can be usable for our purposes. First, we only want to view Date, Units Sold, and Movie Playing columns for each of our franchises, and our code will be separated into each of these franchises as well. Since we are looking at our data over a monthly basis we want to put the column Date as our index. Finally, since some franchises sometimes had multiple titles running at the same time we will be taking the average for all the titles so that we only have a value for each month in our data. After creating a single record for each month, we discovered that we had null values in our data due to the fact that some months no titles were released. The method of dealing with the null values was initially thought to be their deletion but was later changed with forward fill so that we have an equal number of entries for each month to help our prediction models.

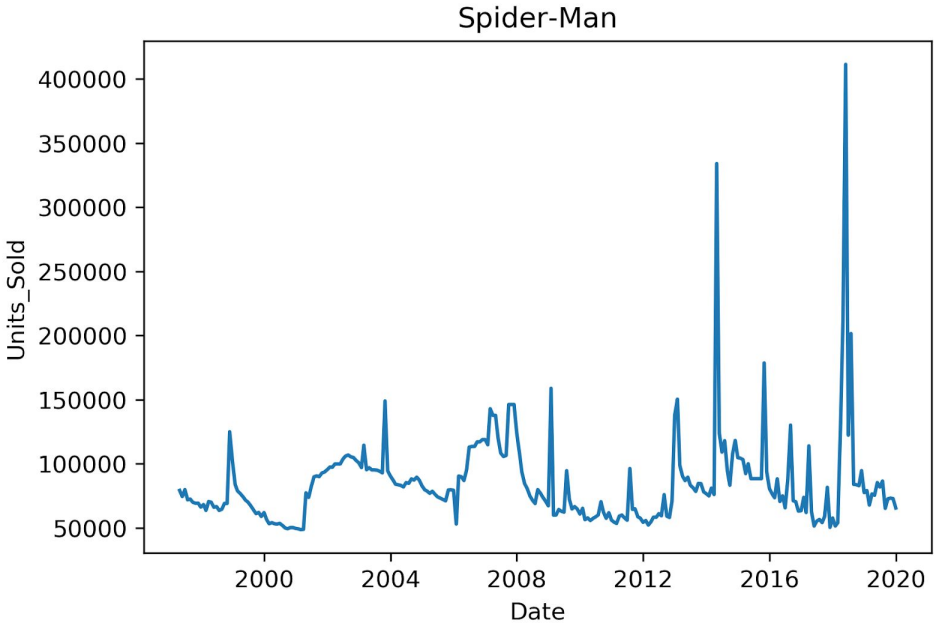
Initial findings for our data starts with creating line graphs that show comic book sales for each of the franchises over time. What we discover is that each graph has no clear, simplistic trend as we go from month to month and seems to fluctuate wildly. The data seems to respond to some type of seasonal trend where comic book sales increase at specific times. The fact that comic book movies usually seem to be released in the summer could be a factor in reasons why

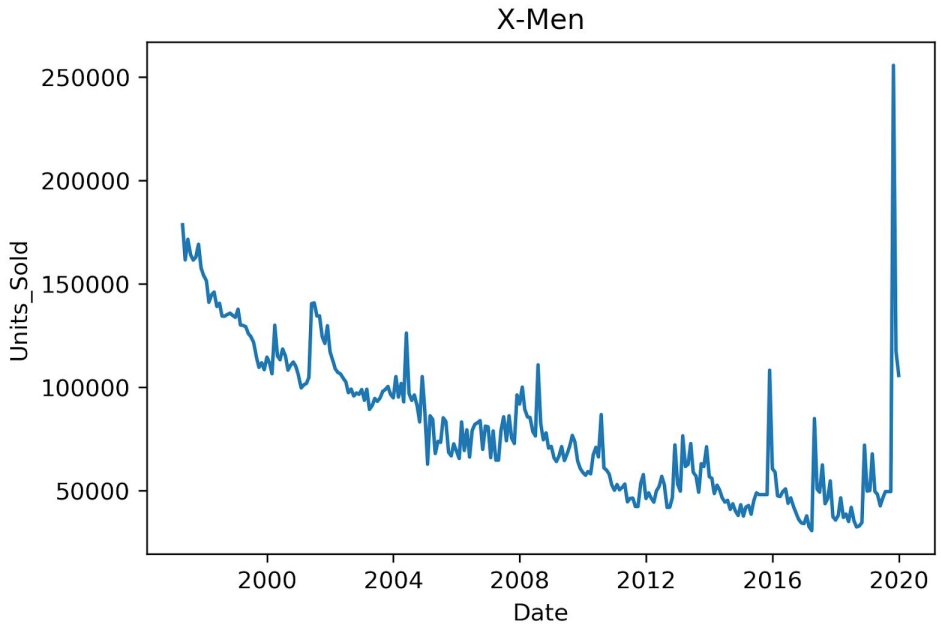
sales differ depending on what time of year it is. When looking at the start of the theatrical run versus at the end we can see that comic book sales are stronger in the beginning and decrease towards the end. Apparently, marketing for movies does seem to increase sales for comic books and gradually decrease over time when the hype for the movie dies down.






Justice League
Units_Sold
200000
175000
150000
125000
100000
75000
50000
2000
2004
2008
2012
2016
2020
Date


Spider-Man
Units_Sold
400000
350000
300000
250000
200000
150000
100000
50000
2000
2004
2008
2012
2016
2020
Date



Our data is now ready to be analyzed by applying inferential statistics techniques to explore the data. The first statistical method we are employing is finding the difference between comic book sales that occurred when a movie versus when a movie was not playing. Running a bootstrap sampling the mean 10,000 times we discover that Avengers, Justice League, and X-Men had more sales when movies were not playing while Spider-Man and Batman had more sales when movies were playing. What's interesting from these results is that not only are Batman and Spider-Man the only franchises that had more sales when movies were playing but they also had the lowest difference in means for movie vs no movie. Batman had 4137 and Spider-Man had a 1090 difference in the mean, while the other franchises had differences of around -20000.

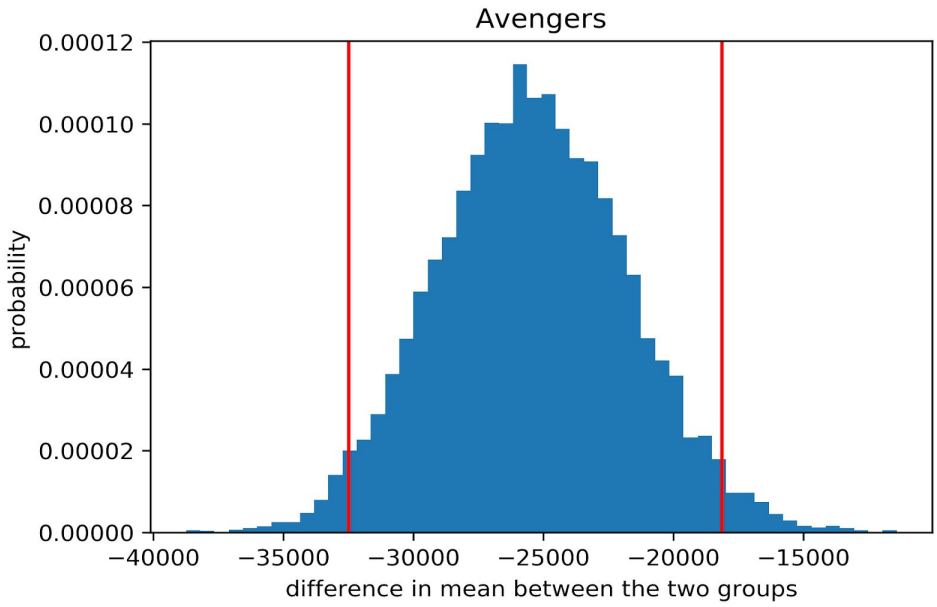
Next, for our statistical analysis, we will be conducting a hypothesis test so that we can either accept or reject based on our findings. There will be two kinds of hypotheses: the null and the alternate. The null hypothesis is the control and is assumed to be true usually with some sort of equal sign while the alternate is the opposite of the null and what we usually want to be proven true. For our purposes, our null hypothesis will be that there is no difference in comic book sales when a movie was playing versus when it was not playing. The alternate hypothesis will be the opposite of this statement in that that is a difference in comic book sales when a movie is playing versus having no movie playing.

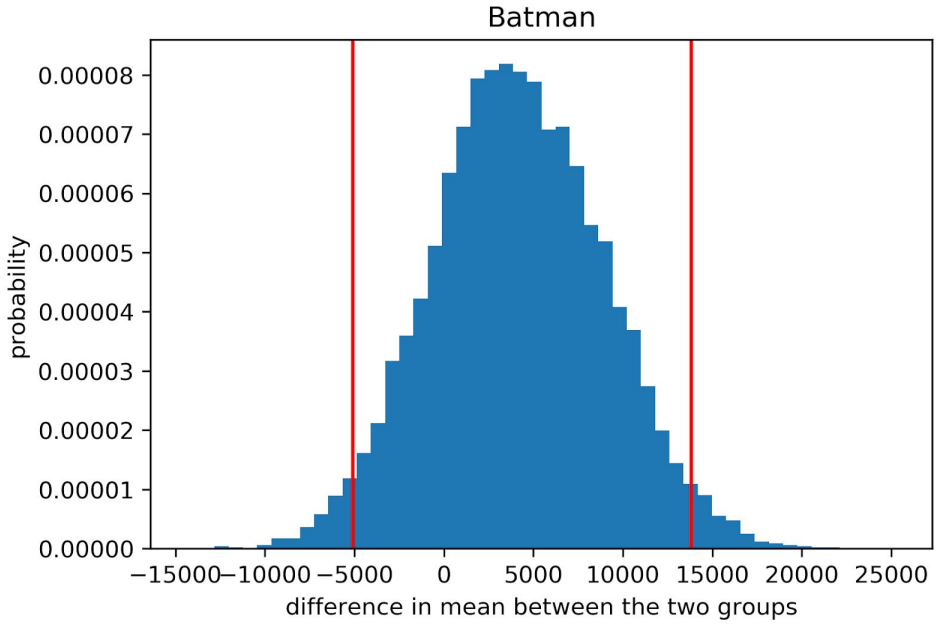
The way that we test this hypothesis is with bootstrapping and confidence intervals that will check if what we are looking for is consistent with what we observe in the sampling distribution. We will take the difference in means we previously calculated and then select values at the chosen percentile for the confidence interval. The chosen percentile is called alpha and we have chosen a confidence interval of 95%. After we have calculated a confidence interval from our bootstrapped difference of means, we make a histogram plot showing the distribution and drawing vertical lines showing where the intervals lie on. What we observe is that even though we know that the difference in mean for Spider-Man was over 1000 when looking at the plot we see that a difference in zero is in the middle of the histogram. Even Batman which had a slightly higher difference of 4000 does not have zero in the middle of the histogram.

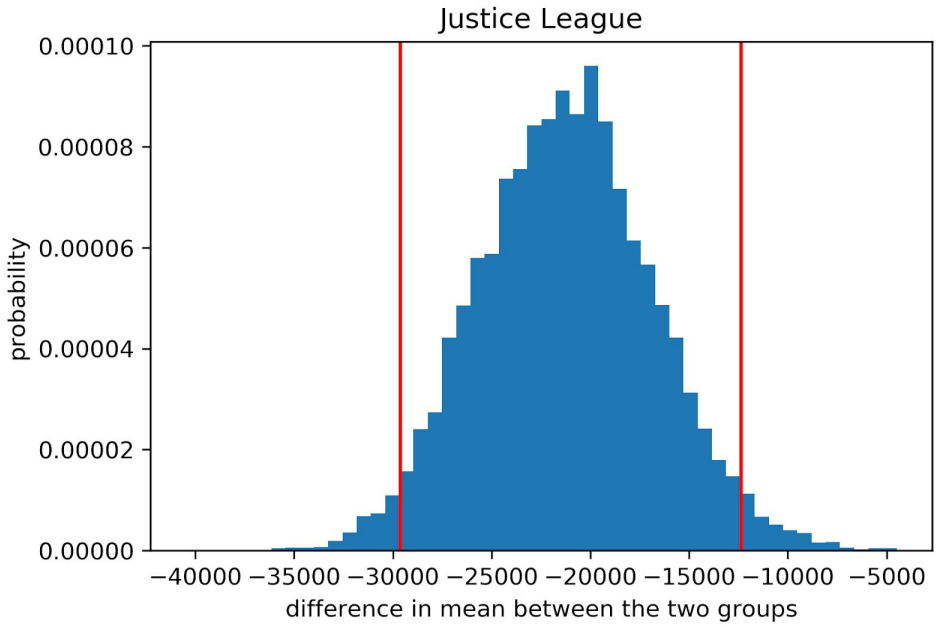
While we have two hypotheses that we are testing for, so far we do not have a way to test whether or not they are true or false. For that, we need a probability value which is more commonly known as the p-value. If the null hypothesis is true then the p-value is the probability of obtaining your sample data. At a certain point, we can make the determination that the null hypothesis is false since the p-value is the probability of our data occurring. This point is called the significance level and is represented by alpha which we already defined with the confidence interval. It is another way of saying how willing we are to committing errors and in our case, we are willing to have a 5% rate. When our p-value goes below the value of 0.05 we can reject our null hypothesis.

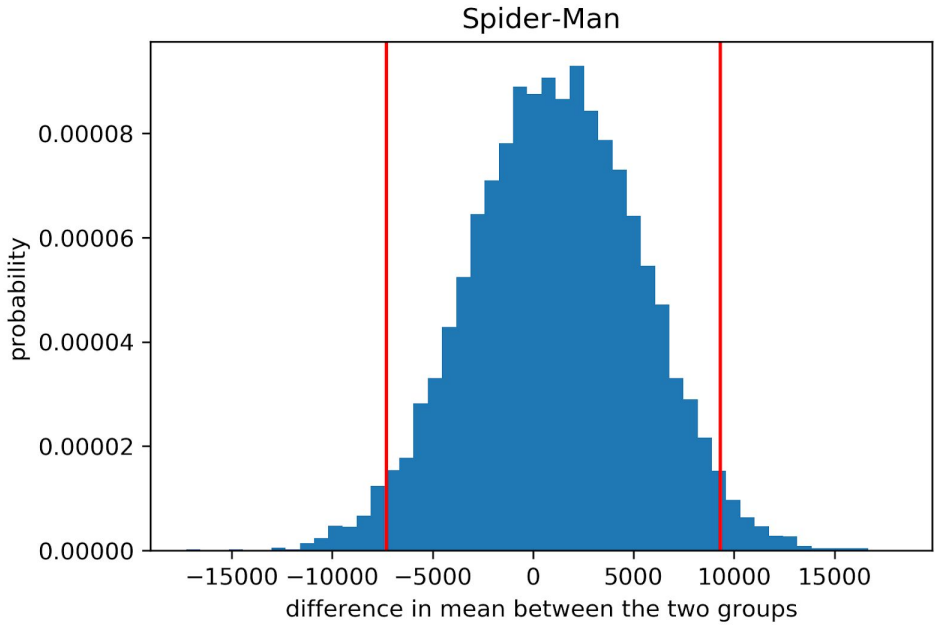
To obtain our p-value we use the SciPy stats T-test comparing the movie and no-movie means. This is a two-sided test for the null hypothesis for two independent samples which we calculate with different means and variances which we discovered previously. We are able to reject the null hypothesis for Avengers, Justice League, and X-Men since their p-value is extremely small. However, we are unable to reject the hypothesis for Batman and Spider-Man since their p-values are so large.

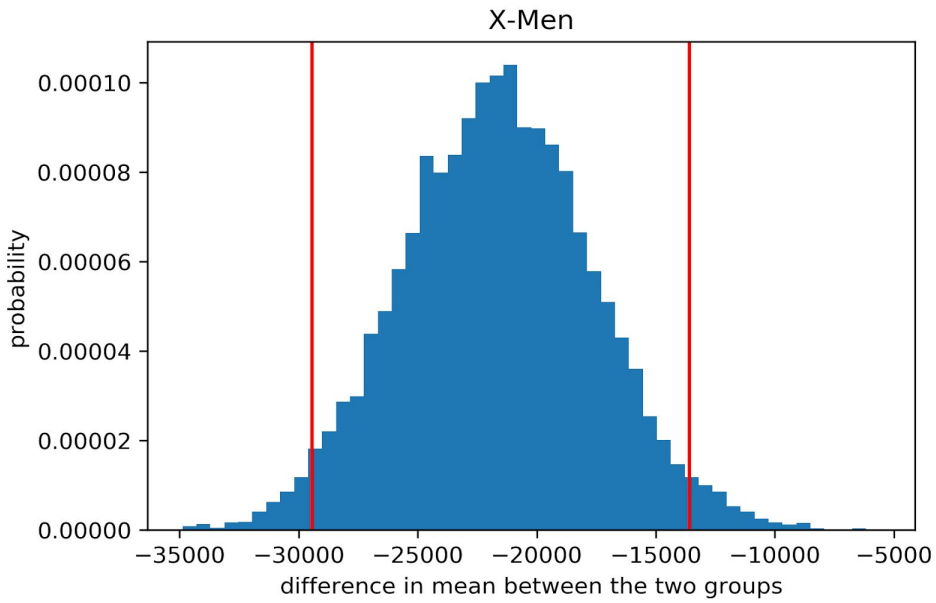
Our results show that depending on the franchise we can have either a difference in means or no difference when looking at the separate groups of comic book sales for when a movie is playing versus no movie playing. This means that when we are modeling our data and creating predictions for them, we should be getting different results for each franchise.


Avengers
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
probability
−40000
−35000
−30000
−25000
−20000
−15000
difference in mean between the two groups


Batman
0.00008
0.00007
0.00006
0.00005
0.00004
0.00003
0.00002
0.00001
0.00000
probability
−15000
−10000
−5000
0
5000
10000
15000
20000
25000
difference in mean between the two groups


Justice League
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
probability
−40000 −35000 −30000 −25000 −20000 −15000 −10000 −5000
difference in mean between the two groups


Spider-Man
0.00008
0.00006
0.00004
0.00002
0.00000
probability
−15000 −10000 −5000 0 5000 10000 15000
difference in mean between the two groups



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.

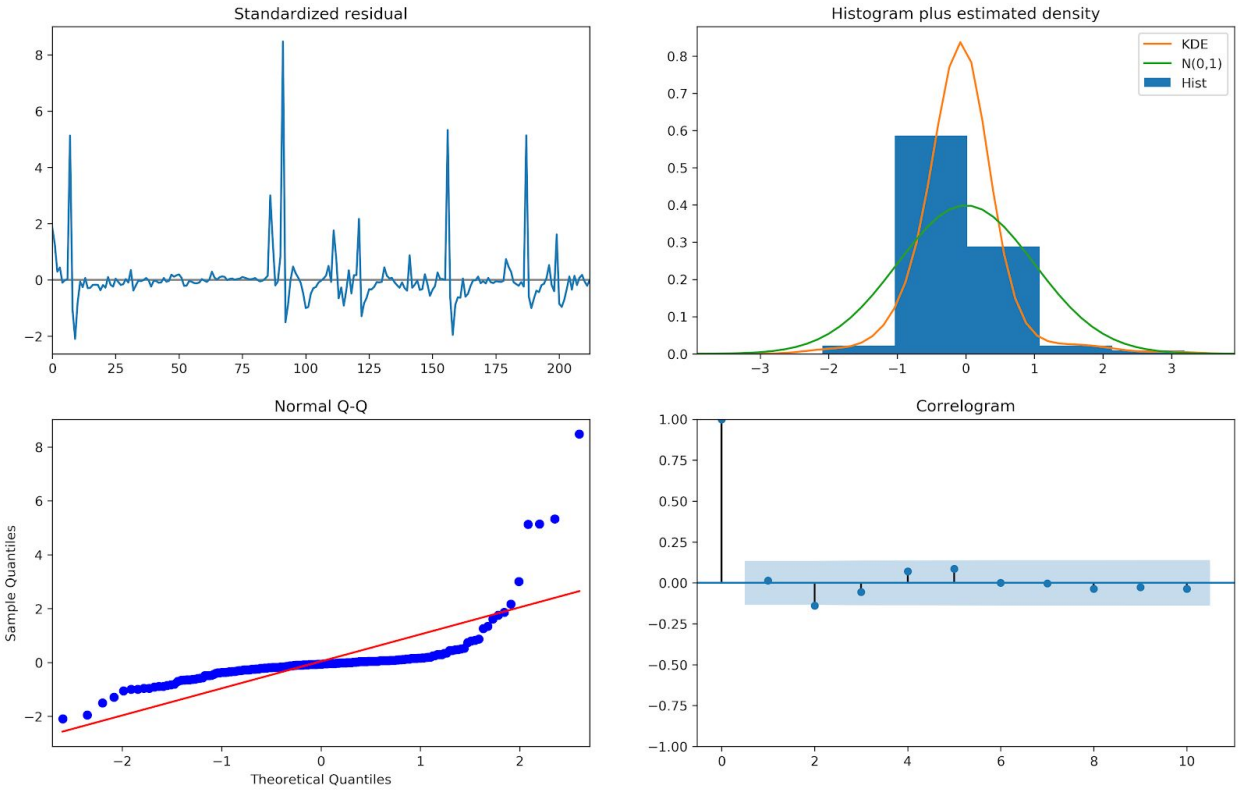
The function plot diagnostics will then give us four separate graphs that coincidentally we would be using if we were trying to find the parameters manually. The four different graphs are Standardized residual, Histogram plus estimated density, Normal Q-Q, and Correlogram. For the Standardized residual 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

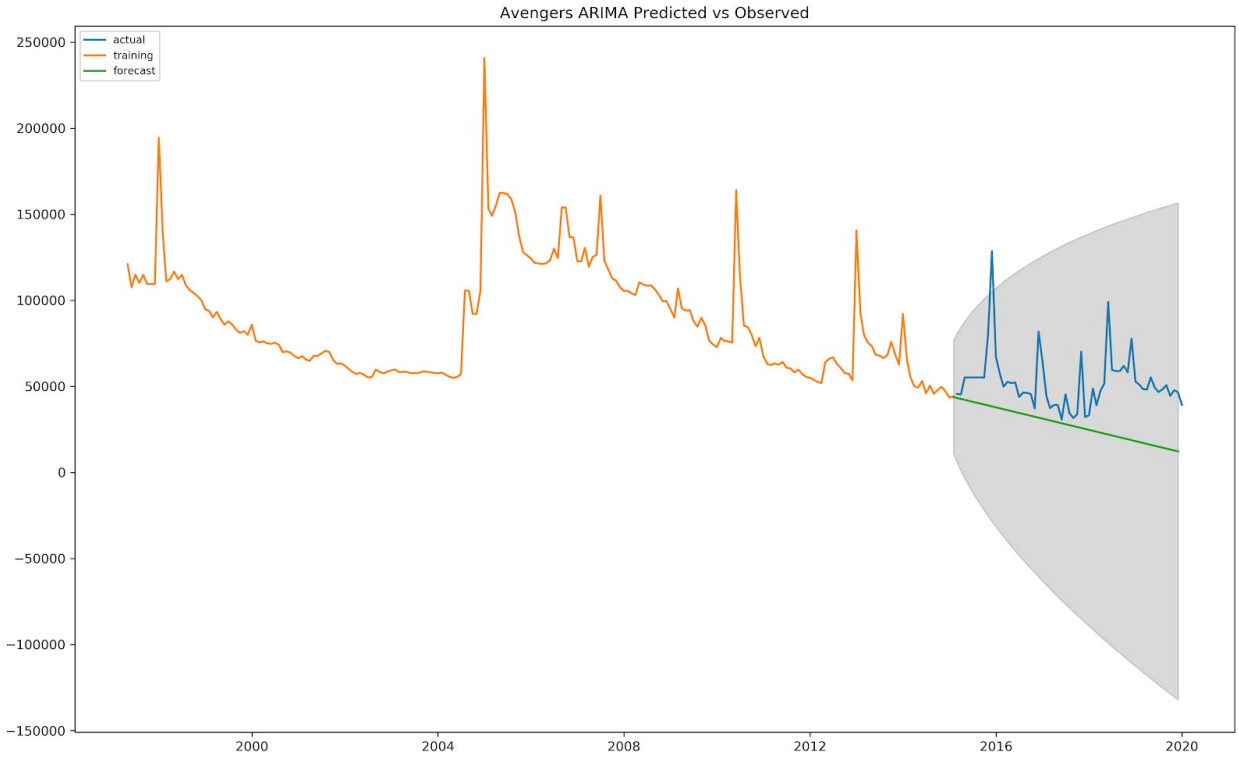
previous seasons and require more computation power to discover the best model. Finally, the SARIMAX model is identical to SARIMA with one key difference, we will be using the Movie Playing variable as the external predictor. In order to use the variable, we need to know the value during the forecast 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.

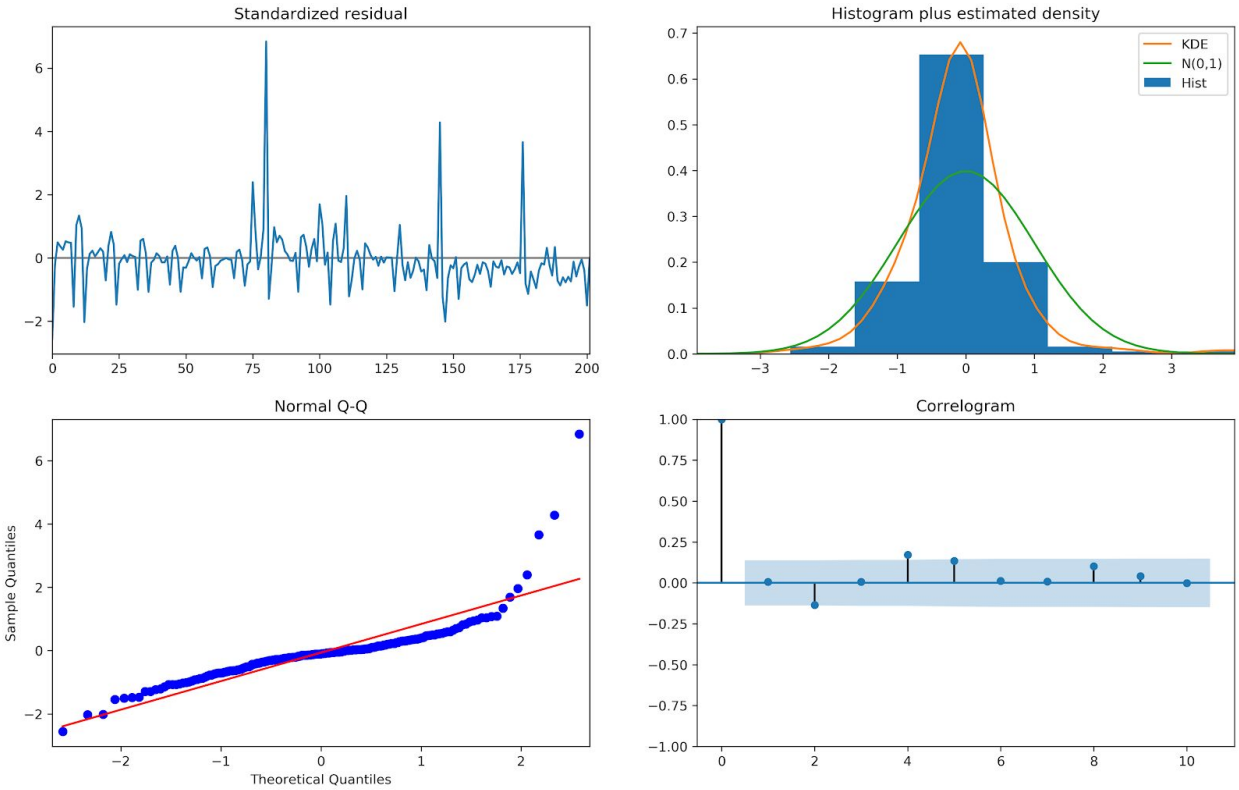
Out of all the five different franchises, the one that sticks out the most is Justice League because our predictions become negative numbers. However, upon closer inspection, we can see that Justice League sales were decreasing at a greater rate compared to other franchises and not leveling out as others 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.

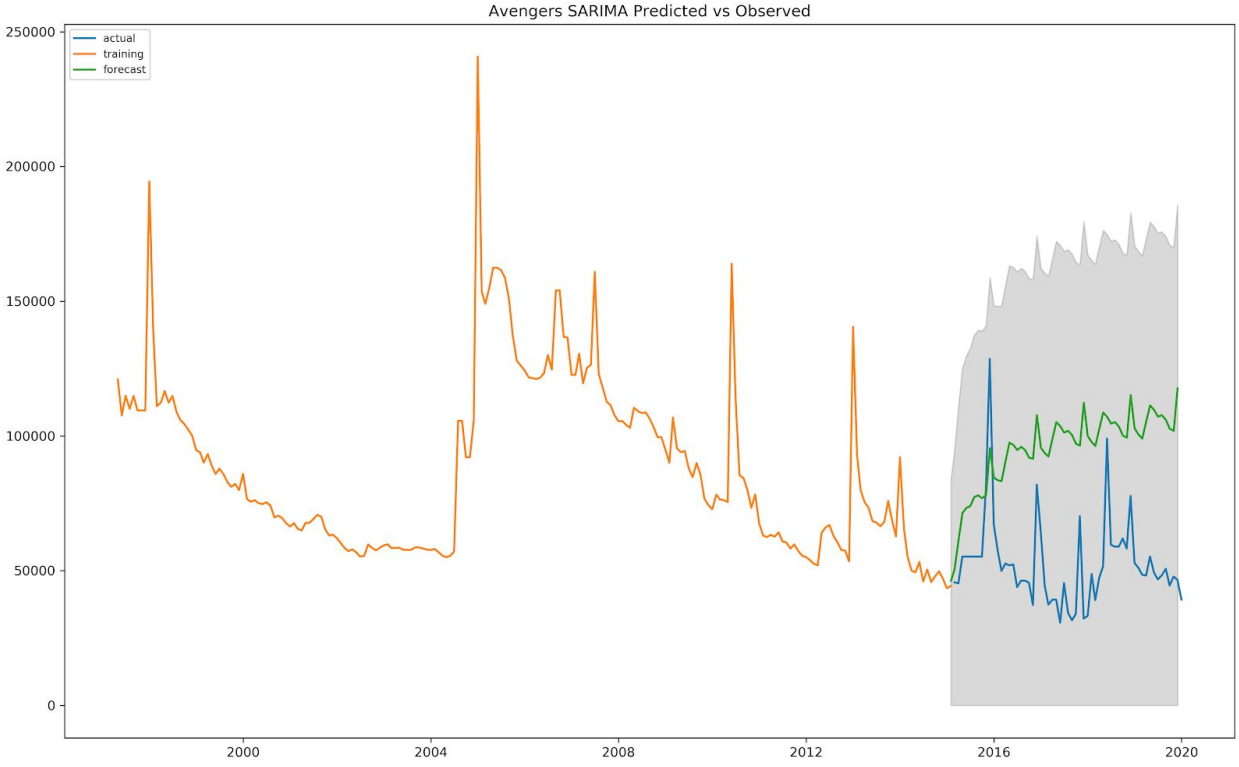
Avengers ARIMA Diagnostics



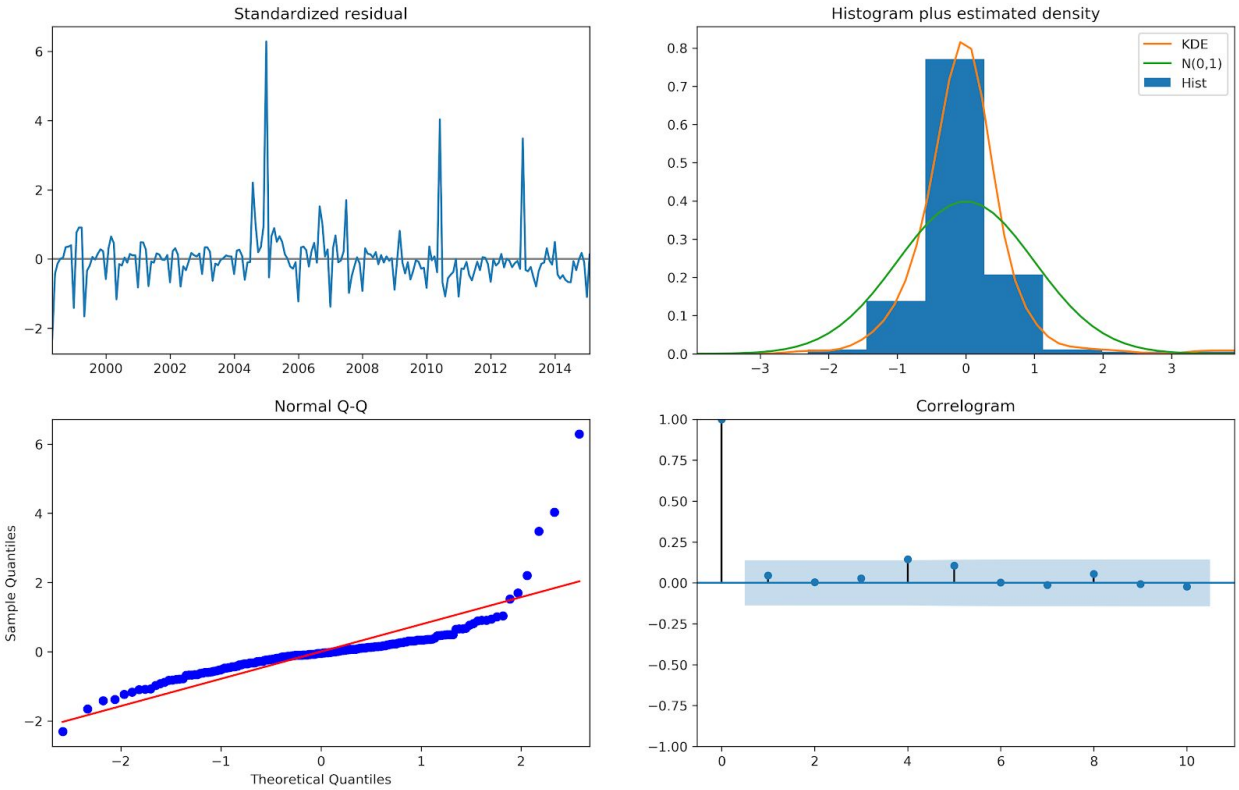


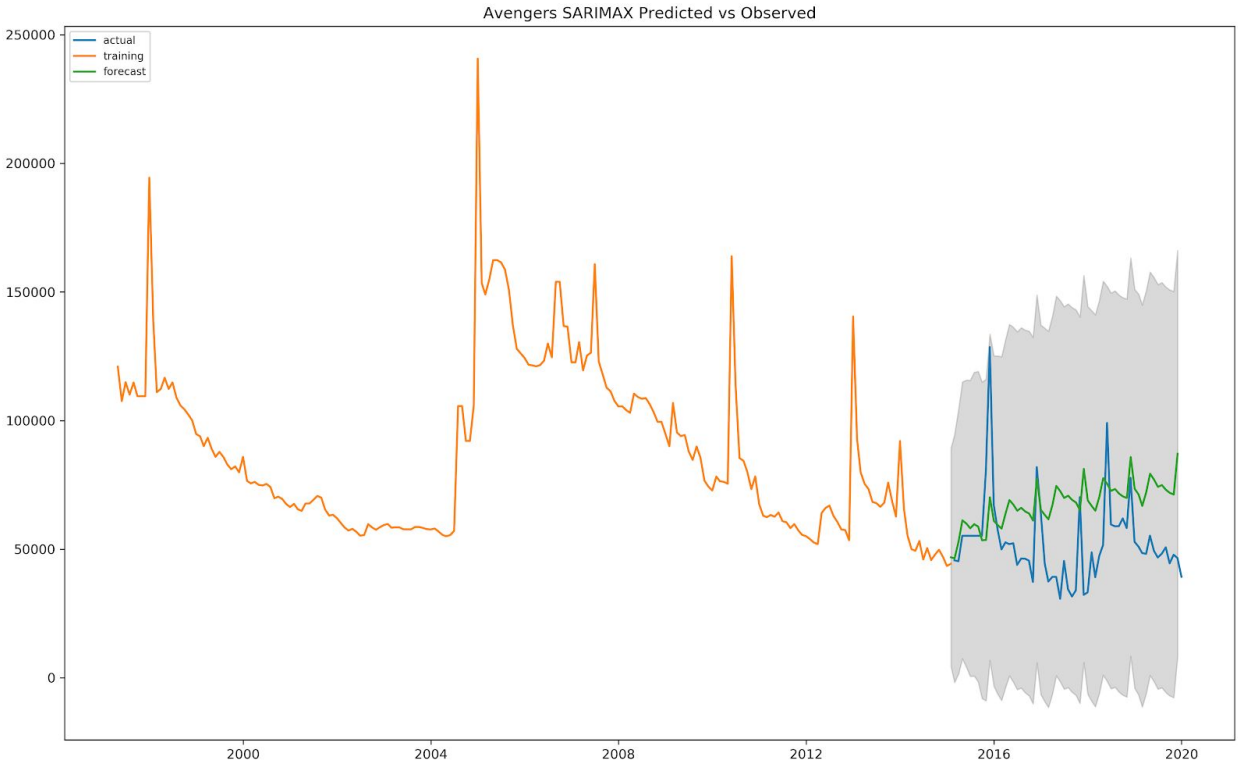
Avengers SARIMA Diagnostics



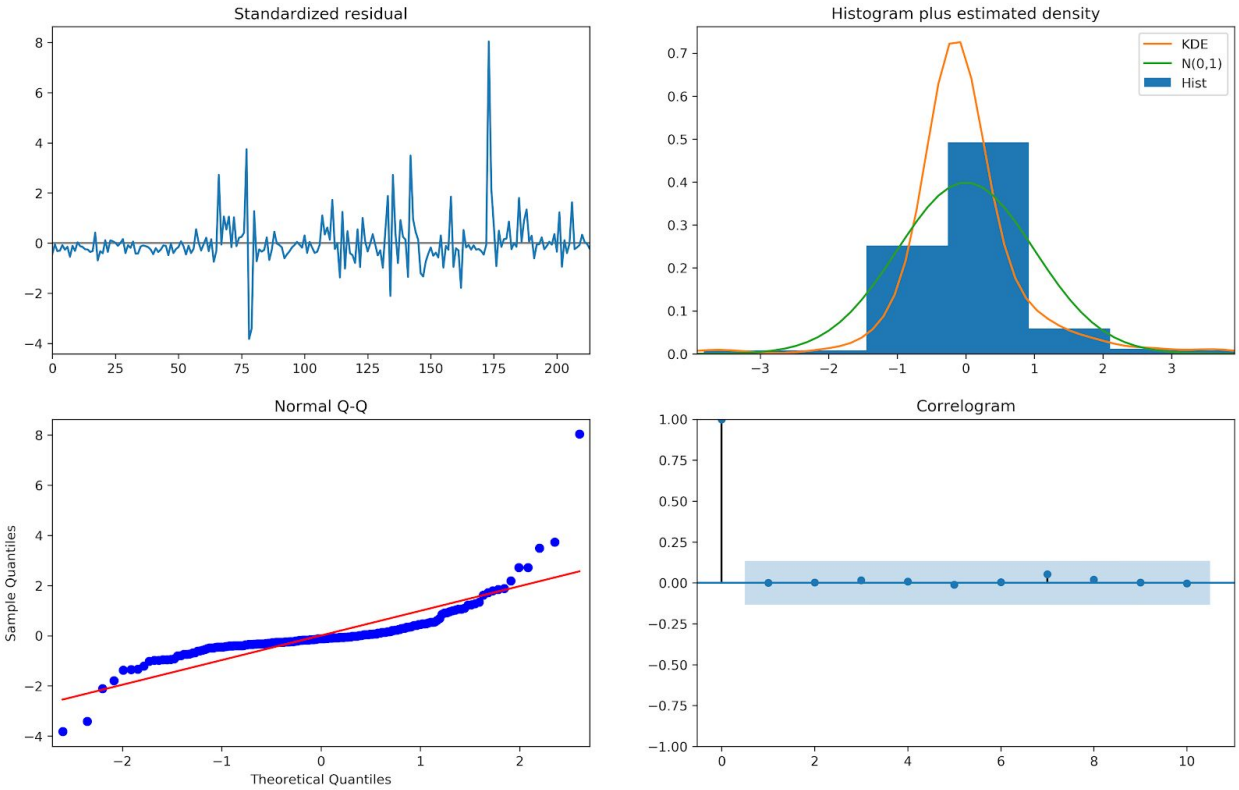


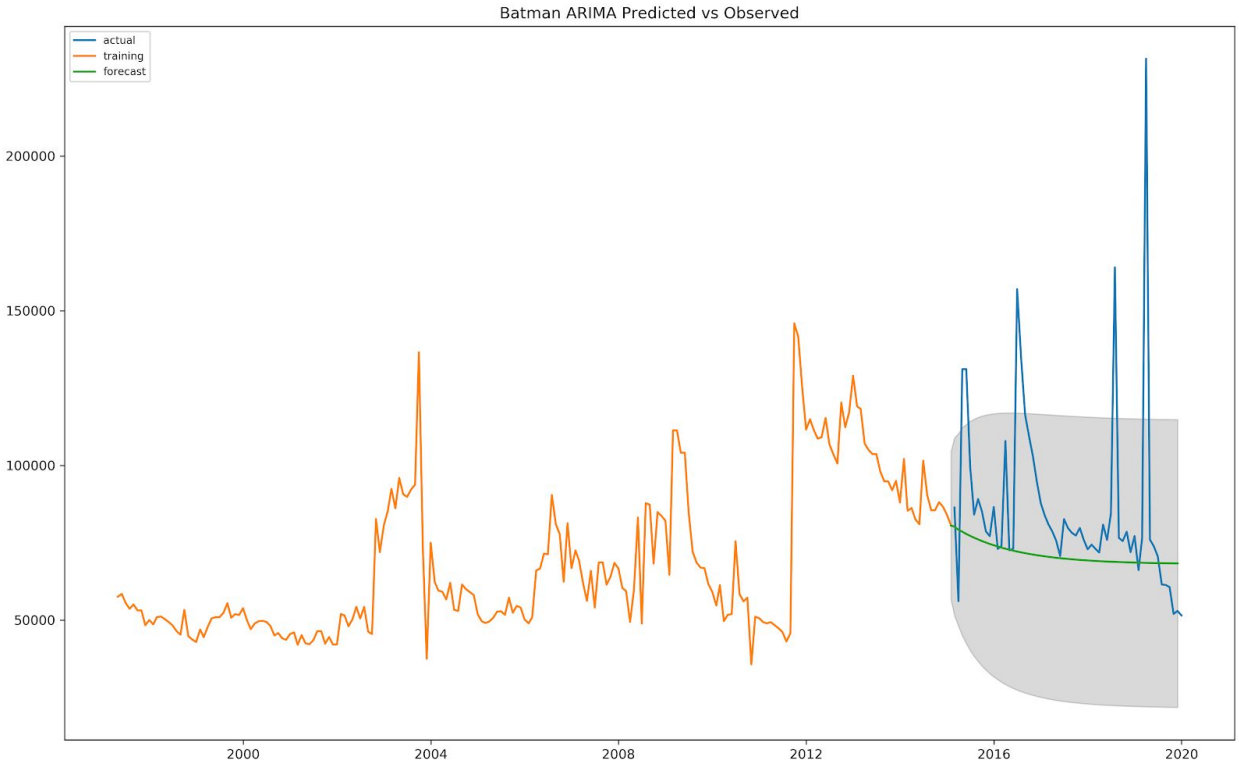
Avengers SARIMAX Diagnostics



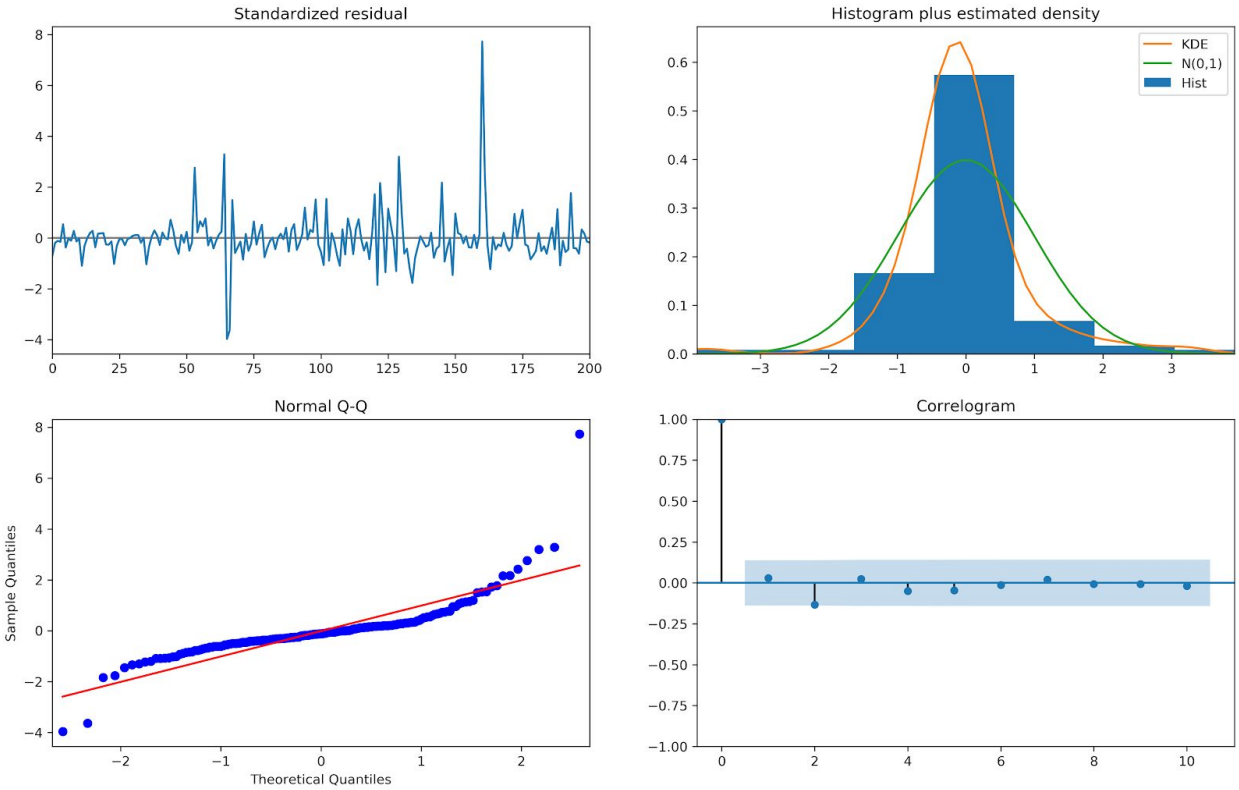


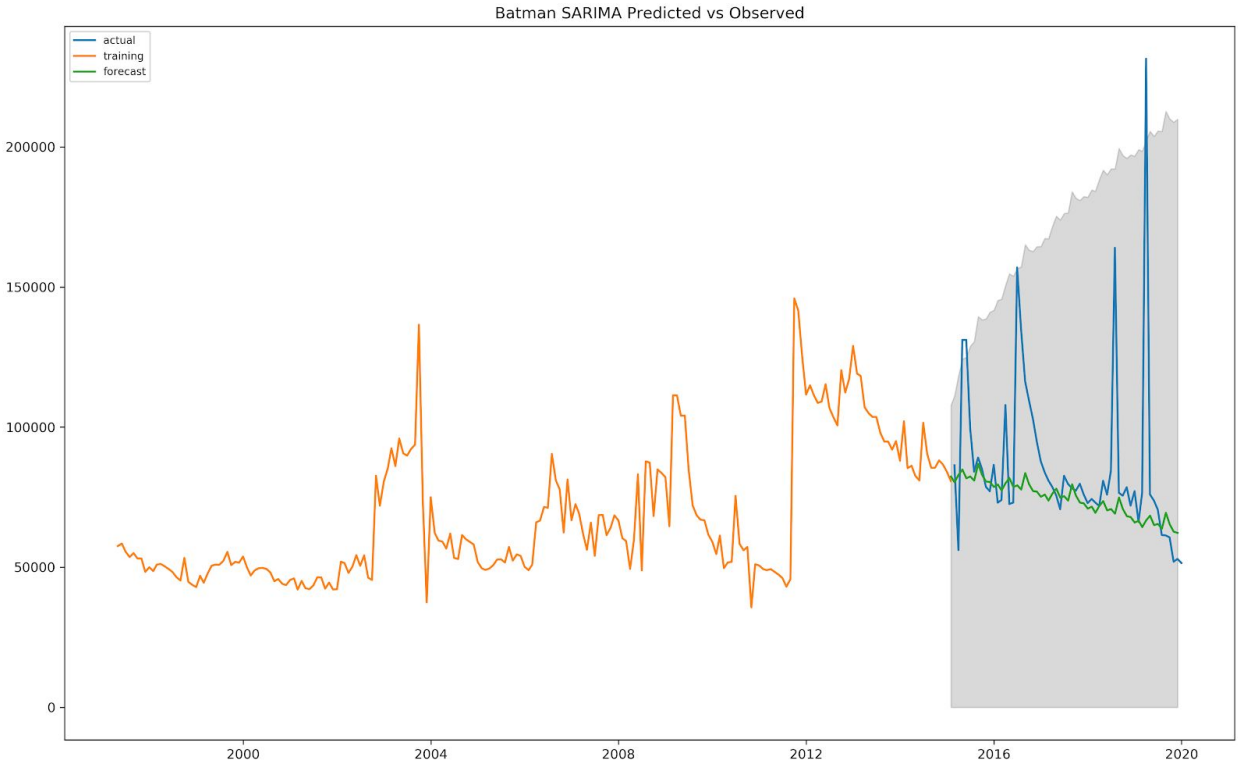
Batman ARIMA Diagnostics



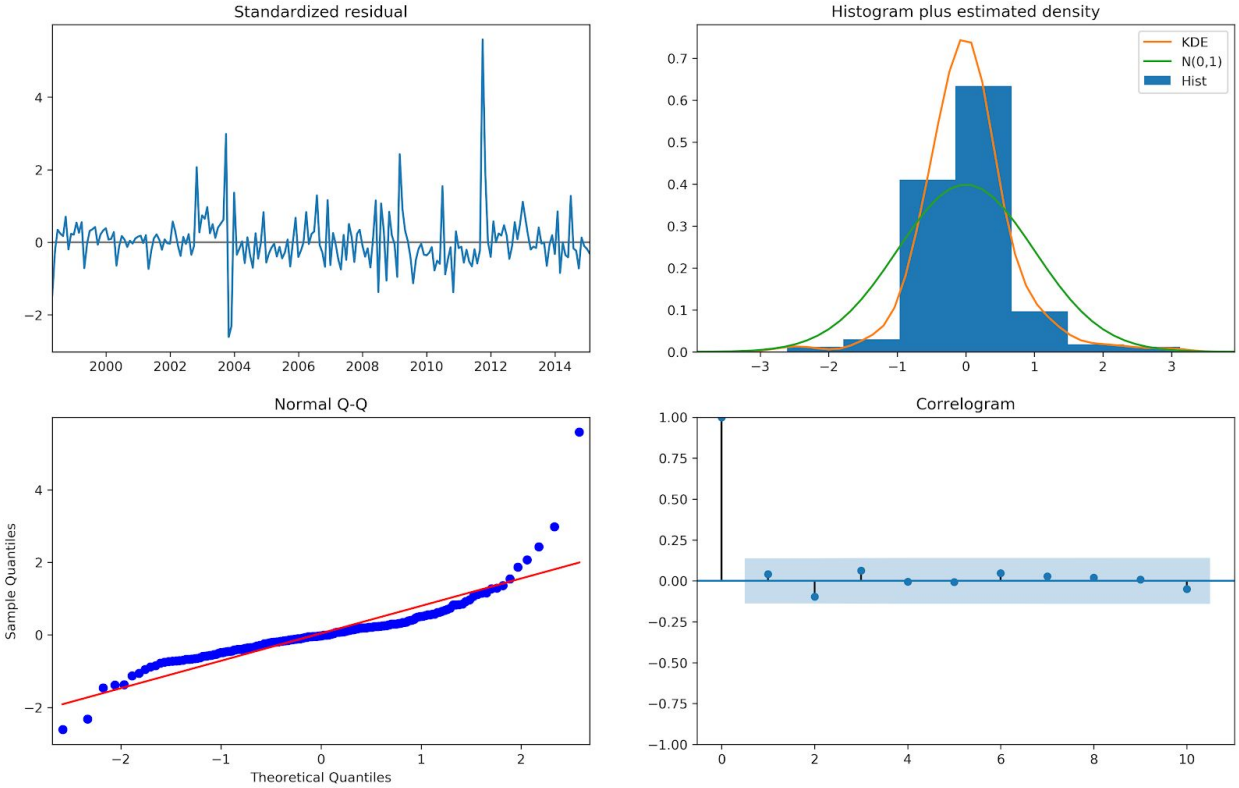


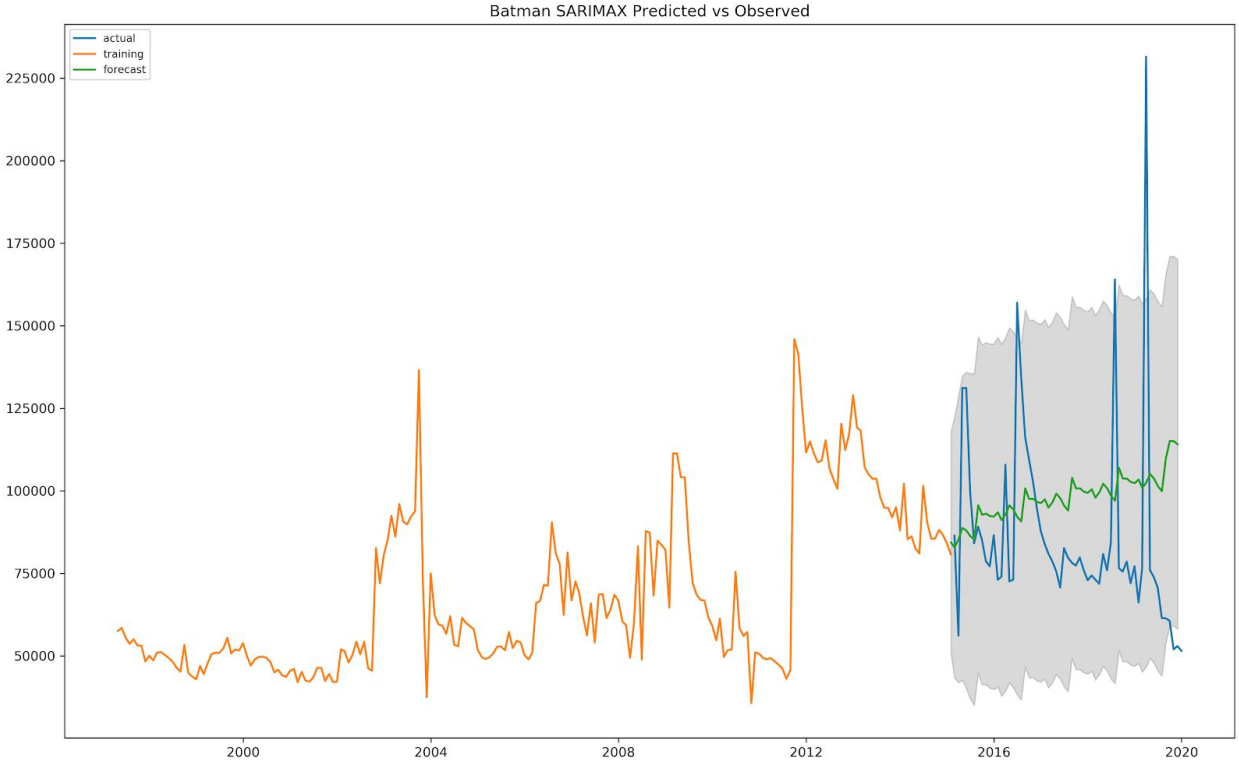
Batman SARIMA Diagnostics



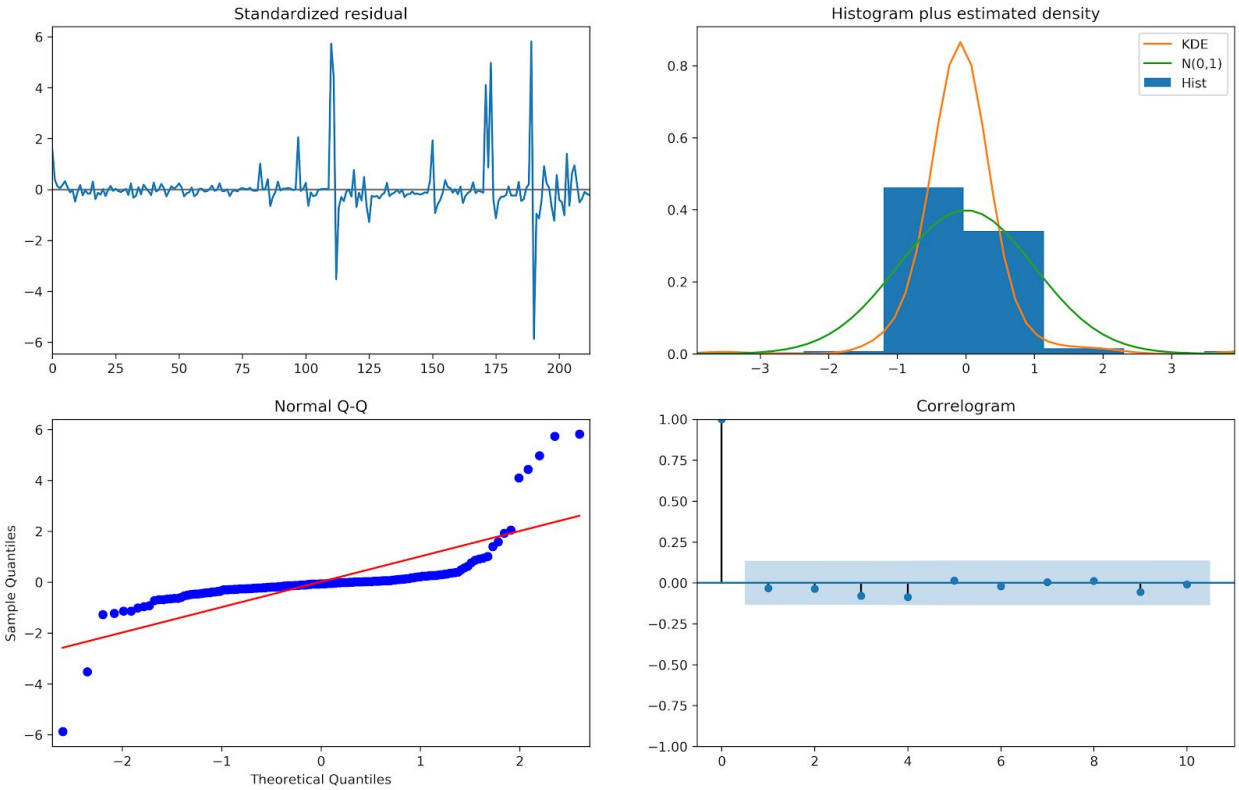


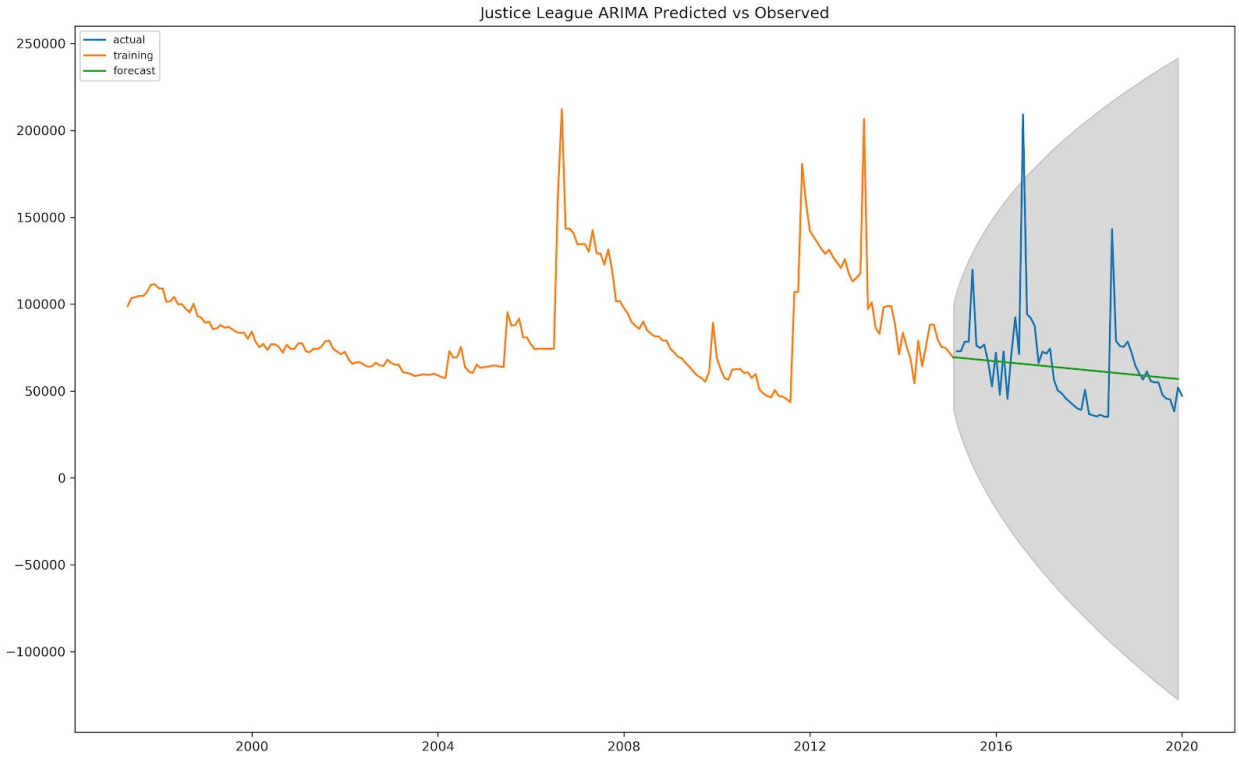
Batman SARIMAX Diagnostics



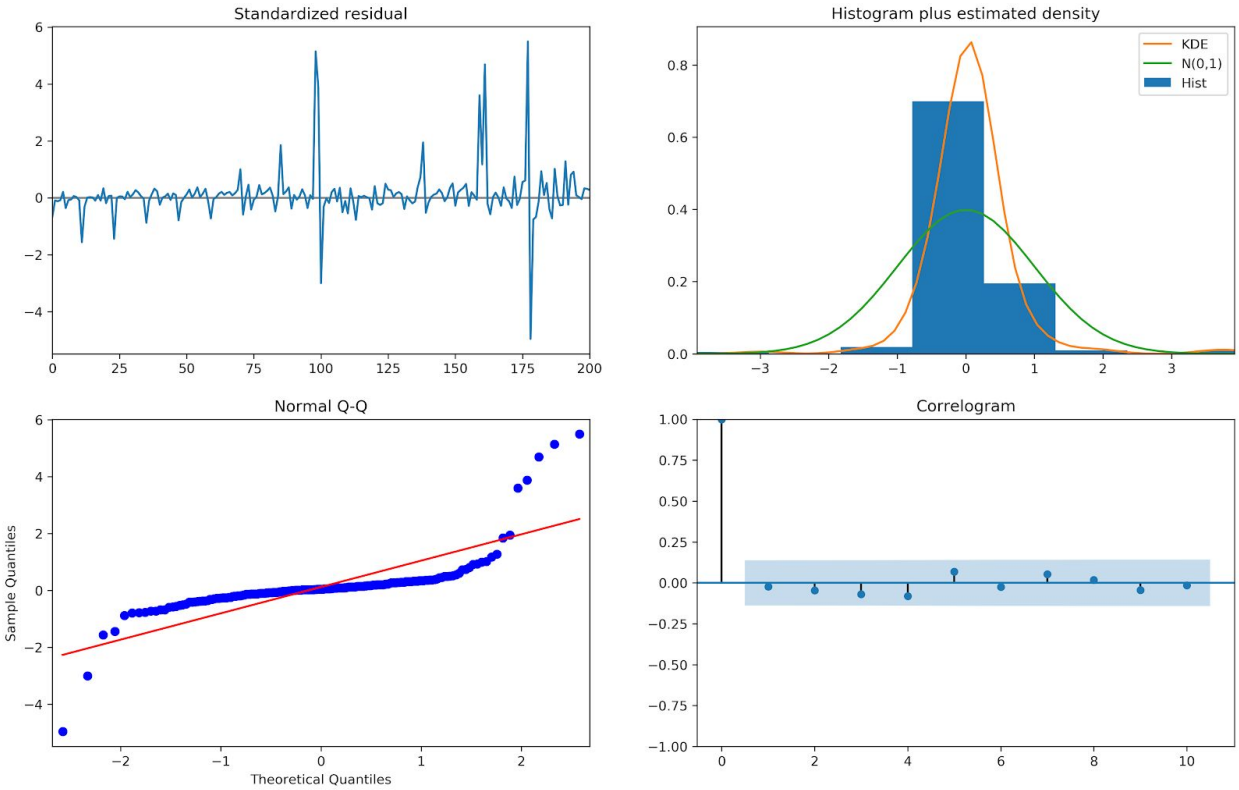


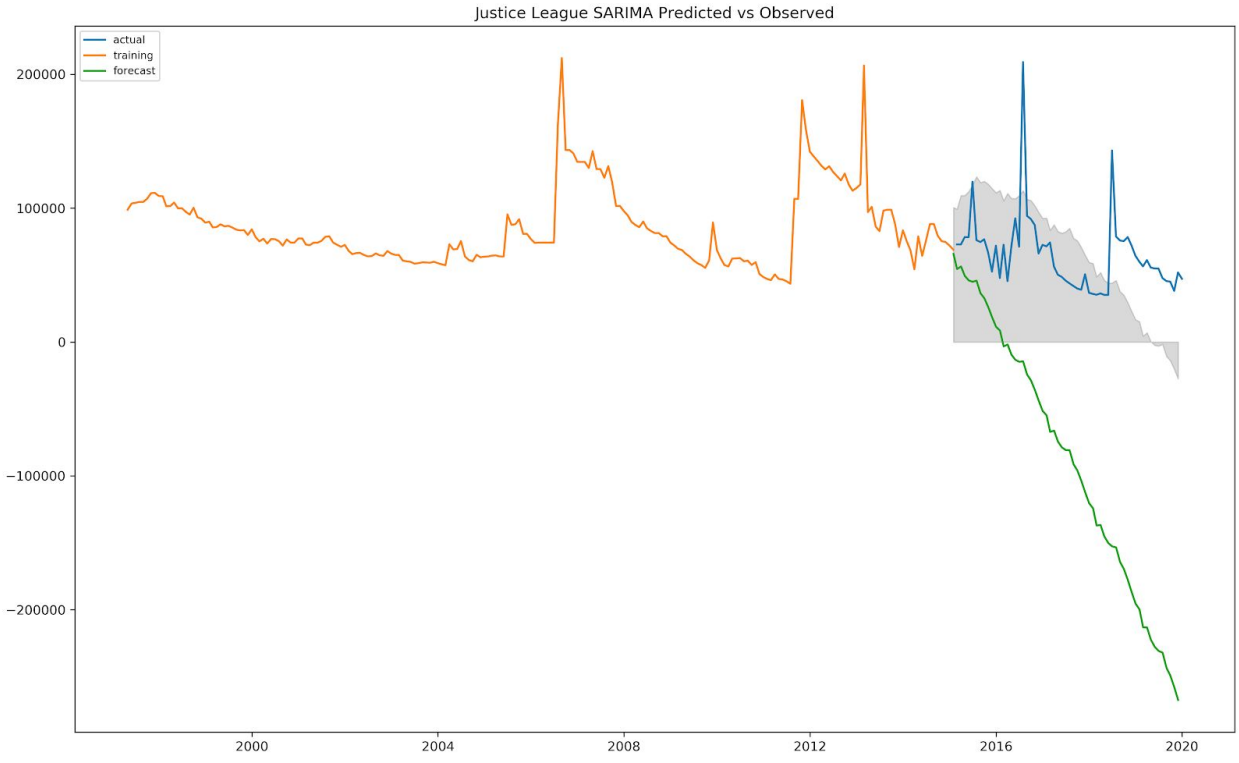
Justice League ARIMA Diagnostics



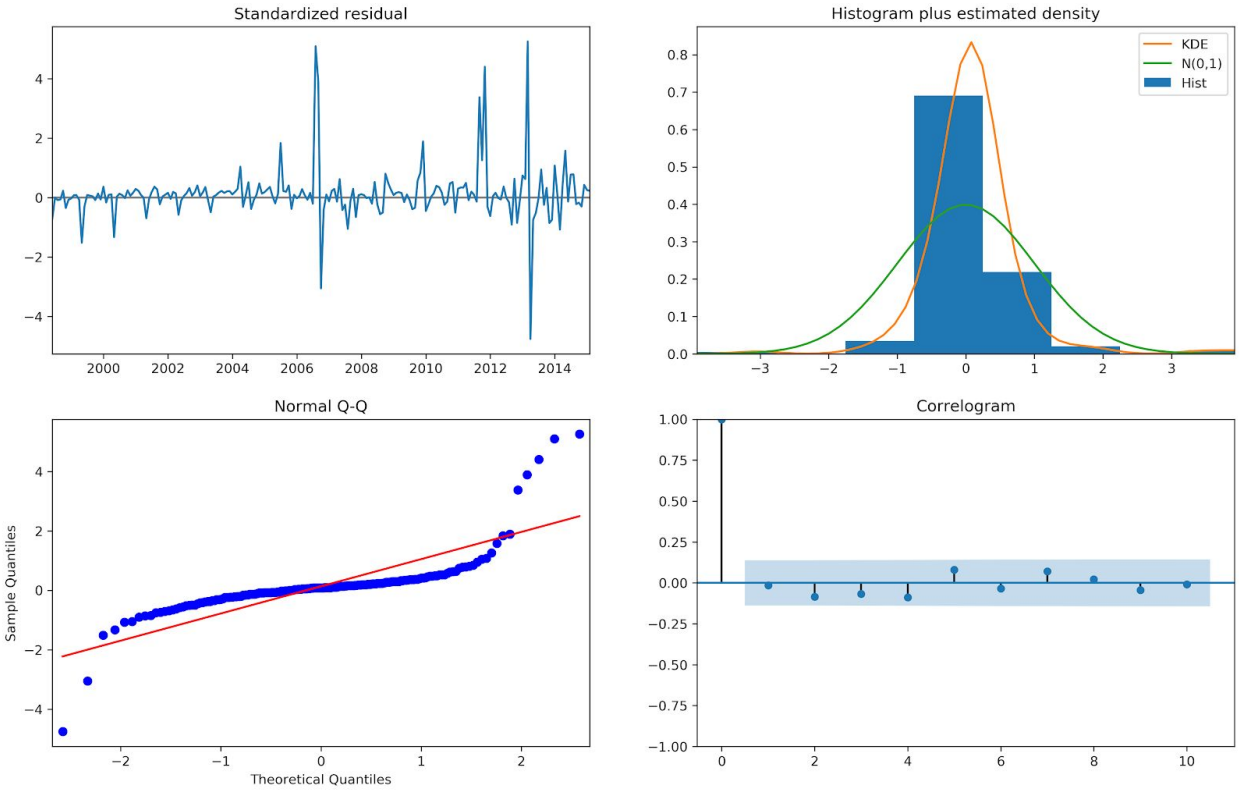


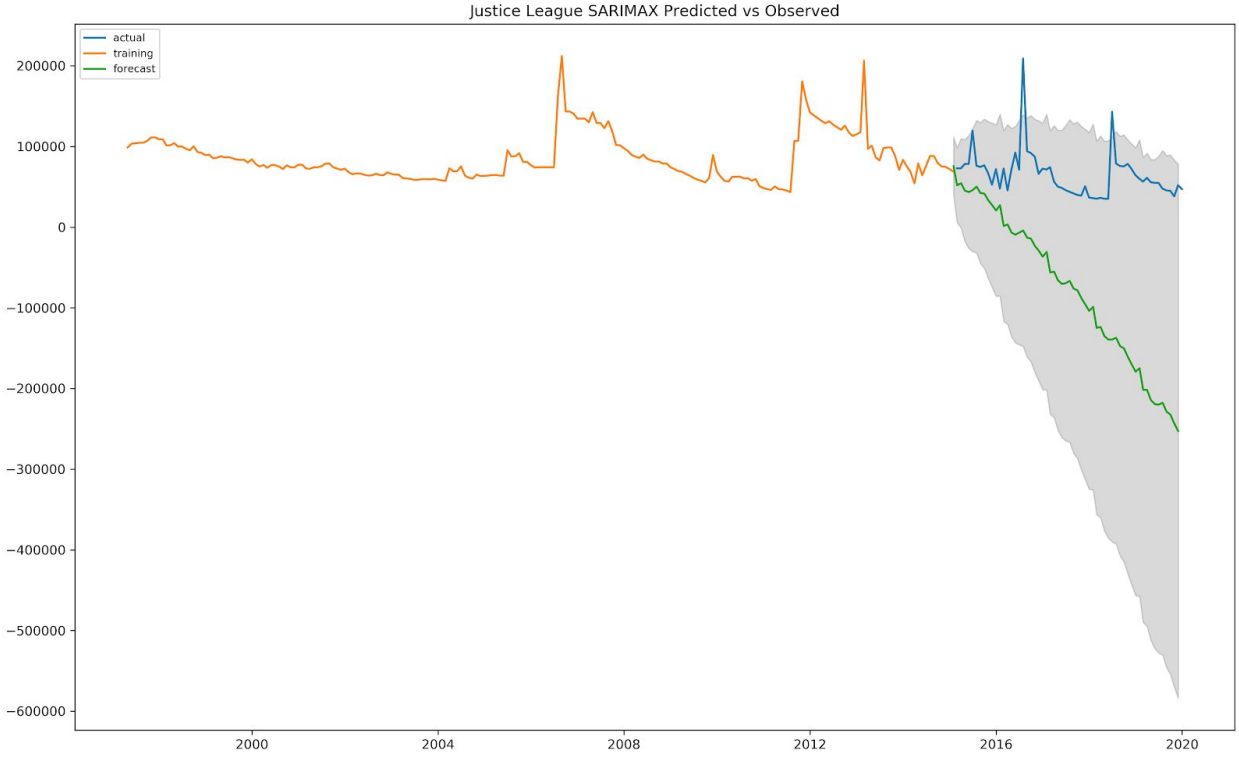
Justice League SARIMA Diagnostics



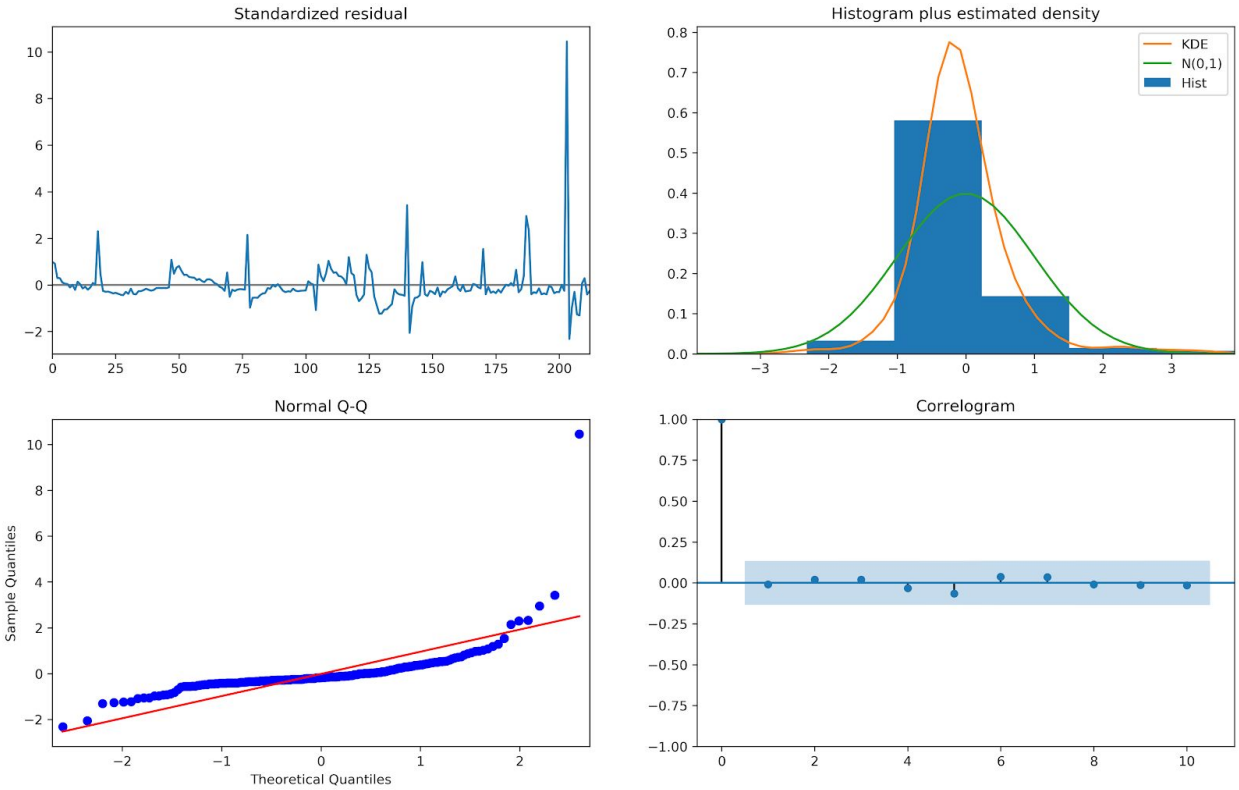


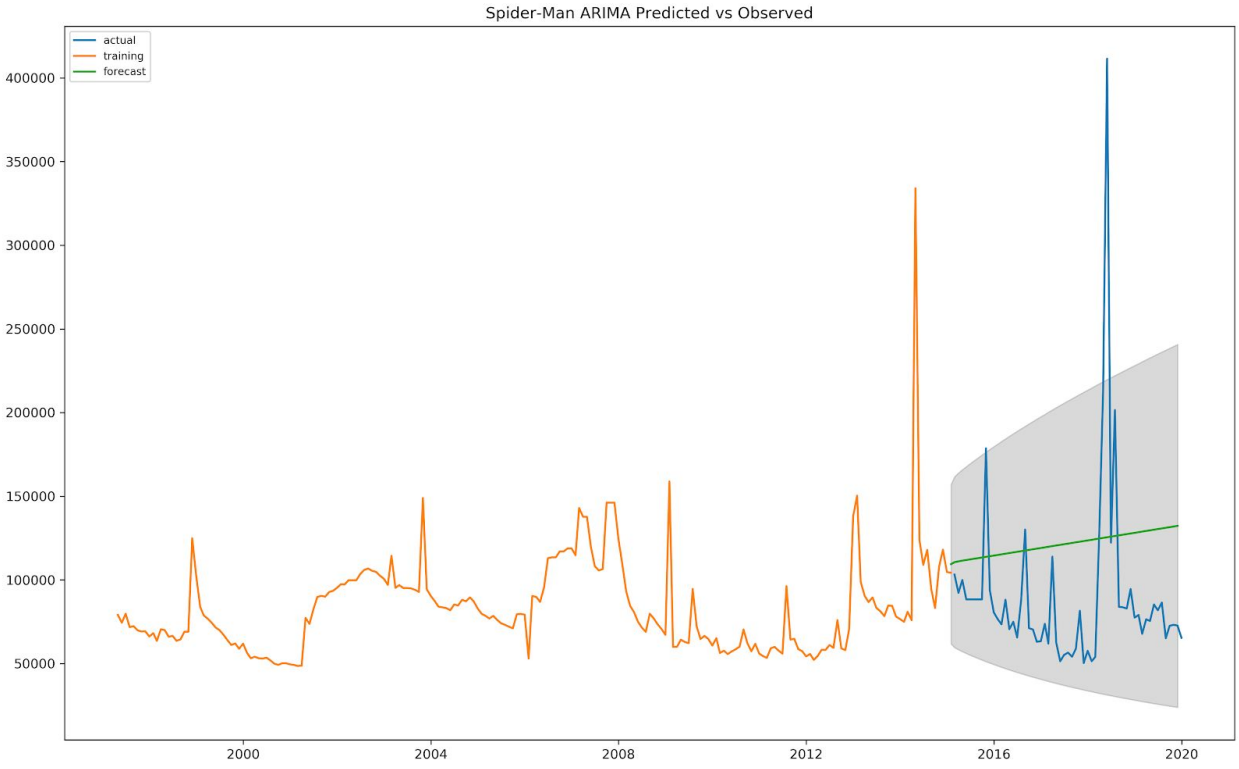
Justice League SARIMAX Diagnostics



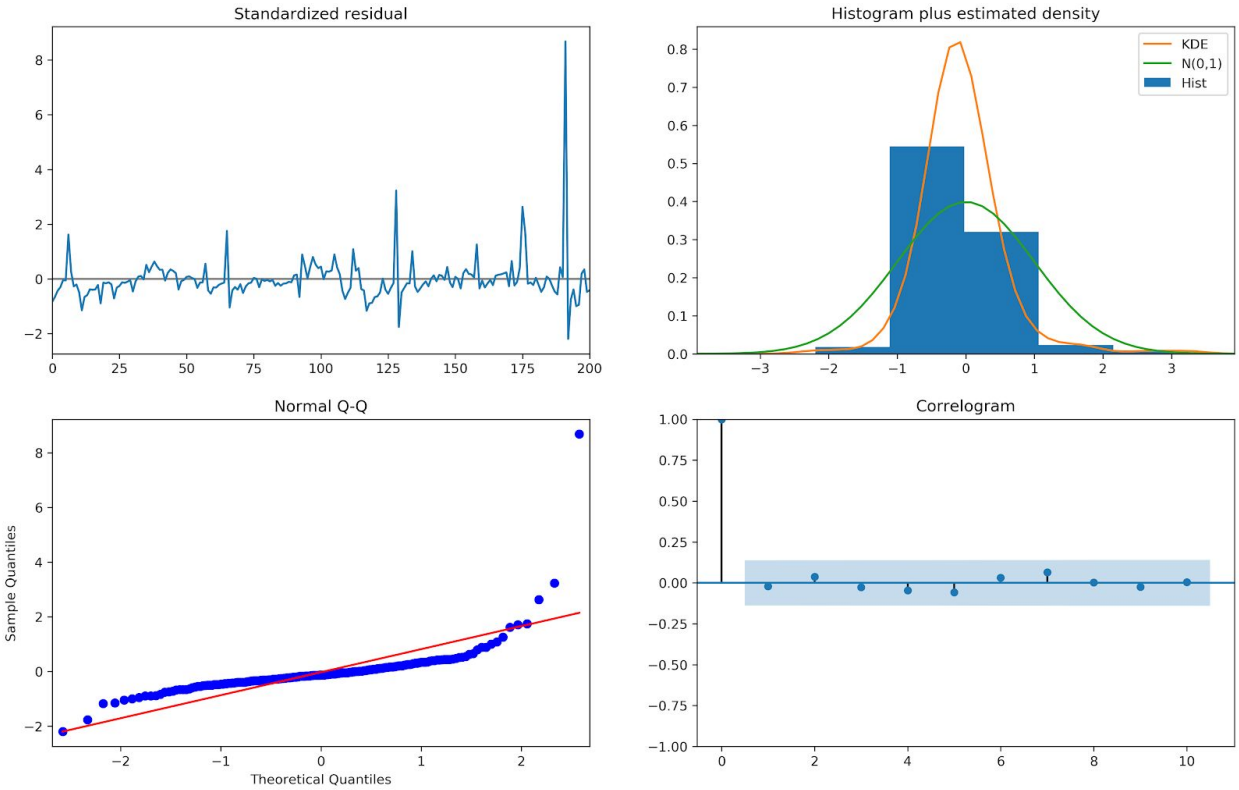


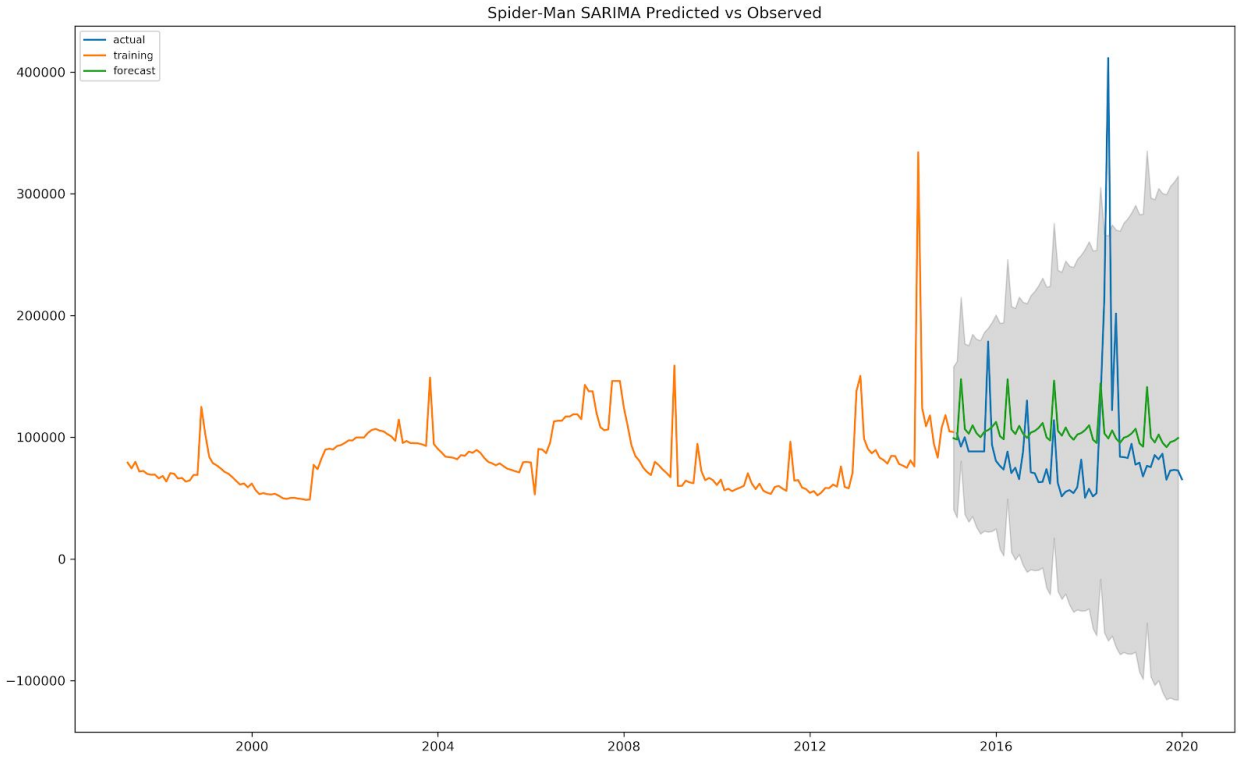
Spider-Man ARIMA Diagnostics



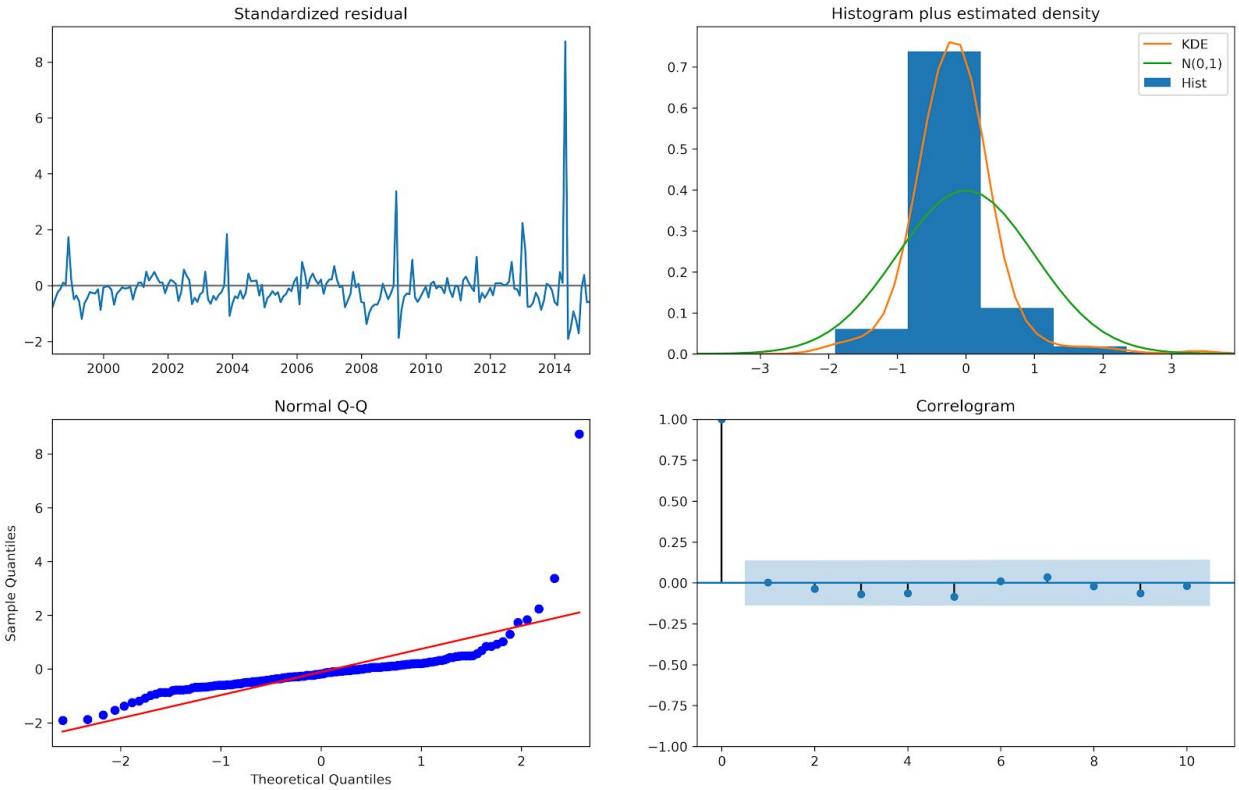


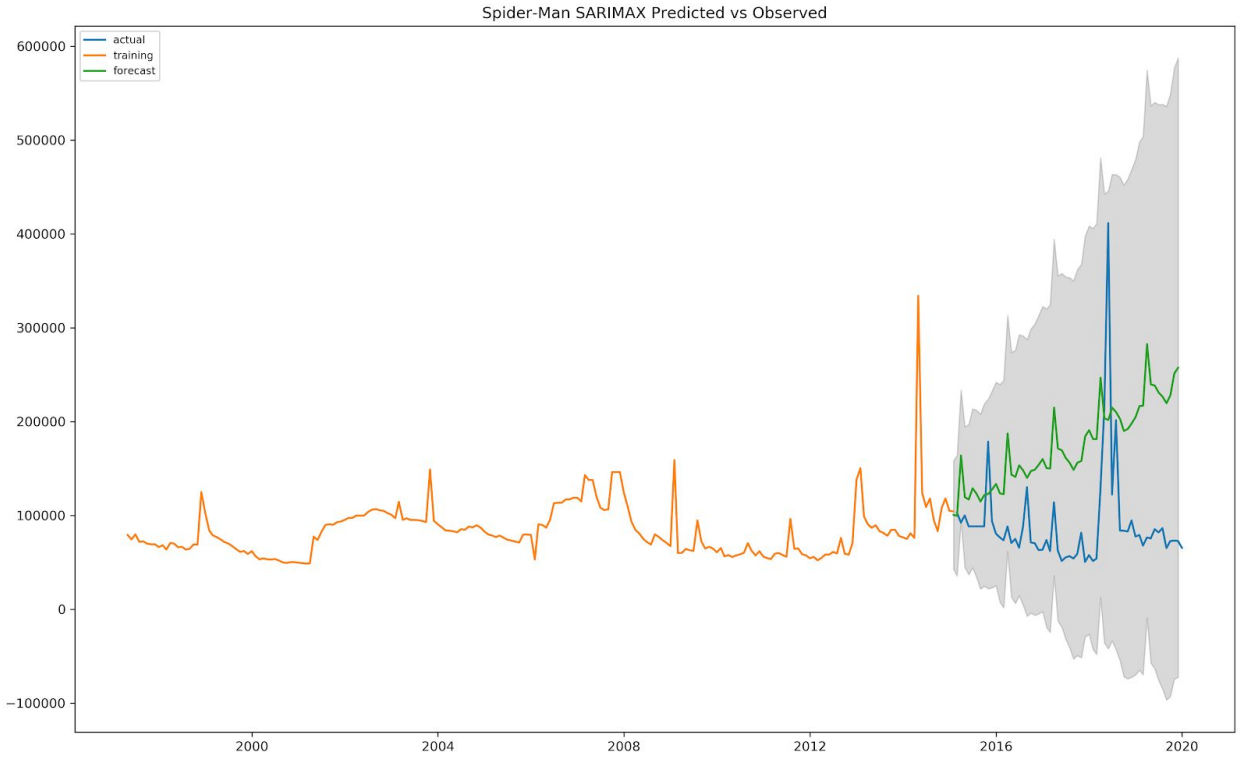
Spider-Man SARIMA Diagnostics



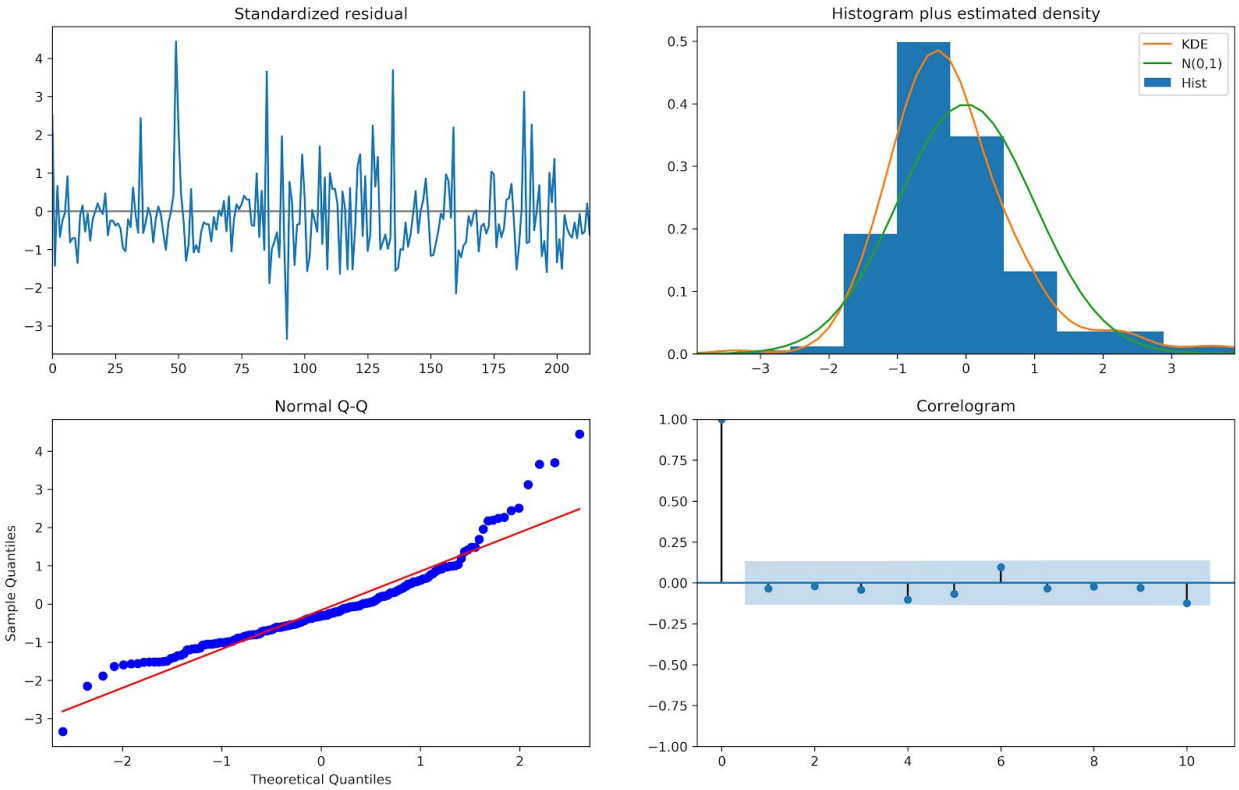


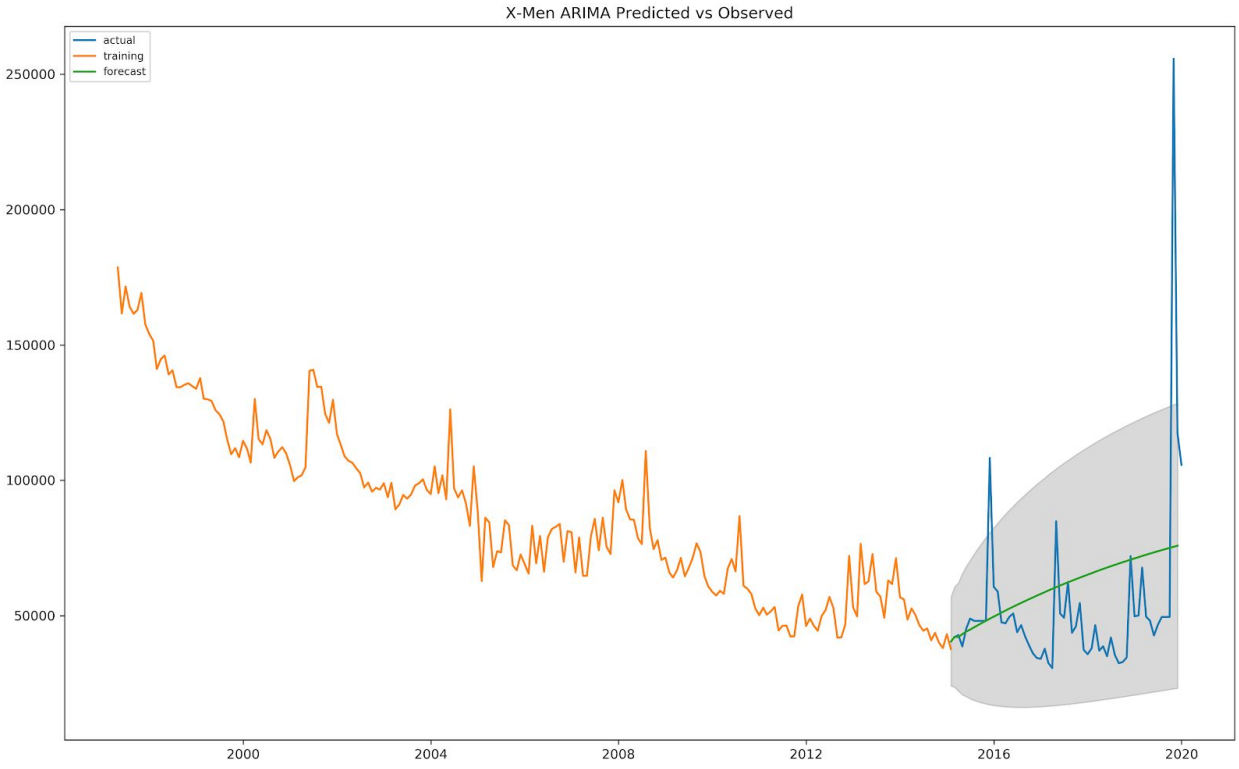
Spider-Man SARIMAX Diagnostics



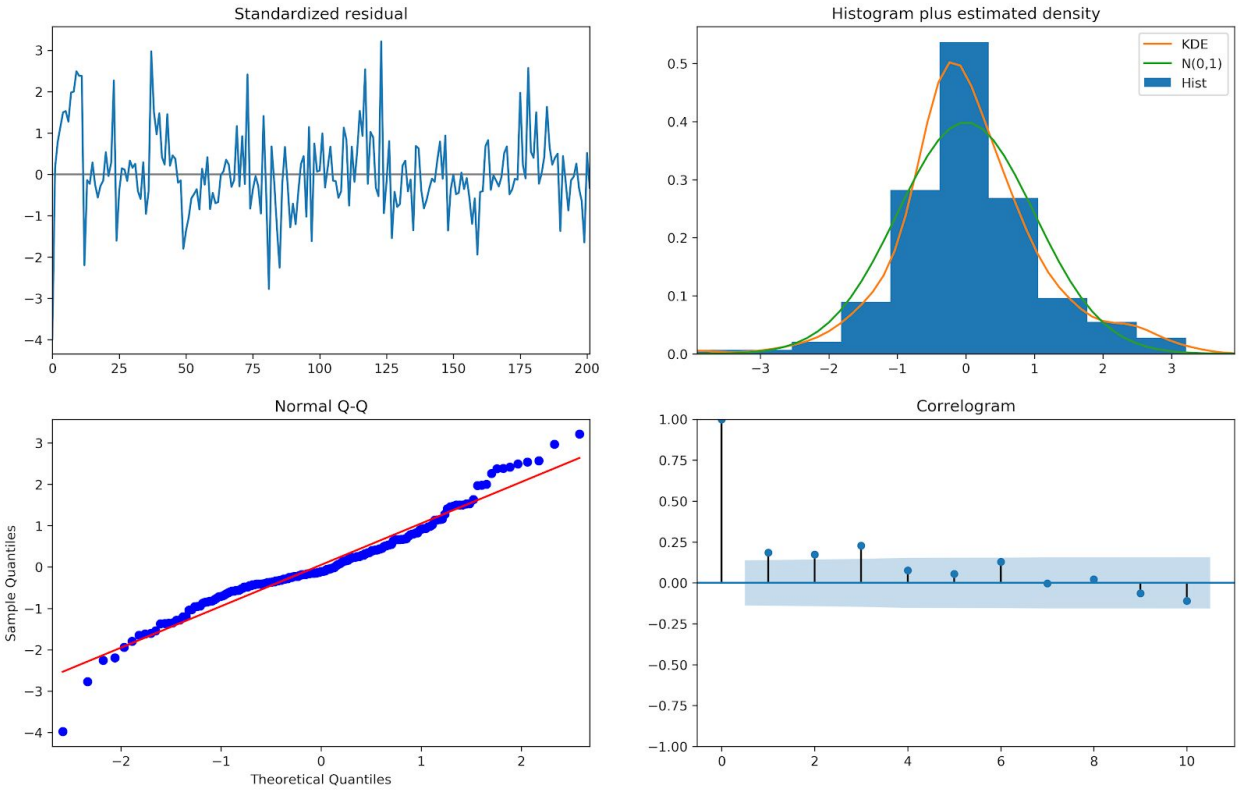


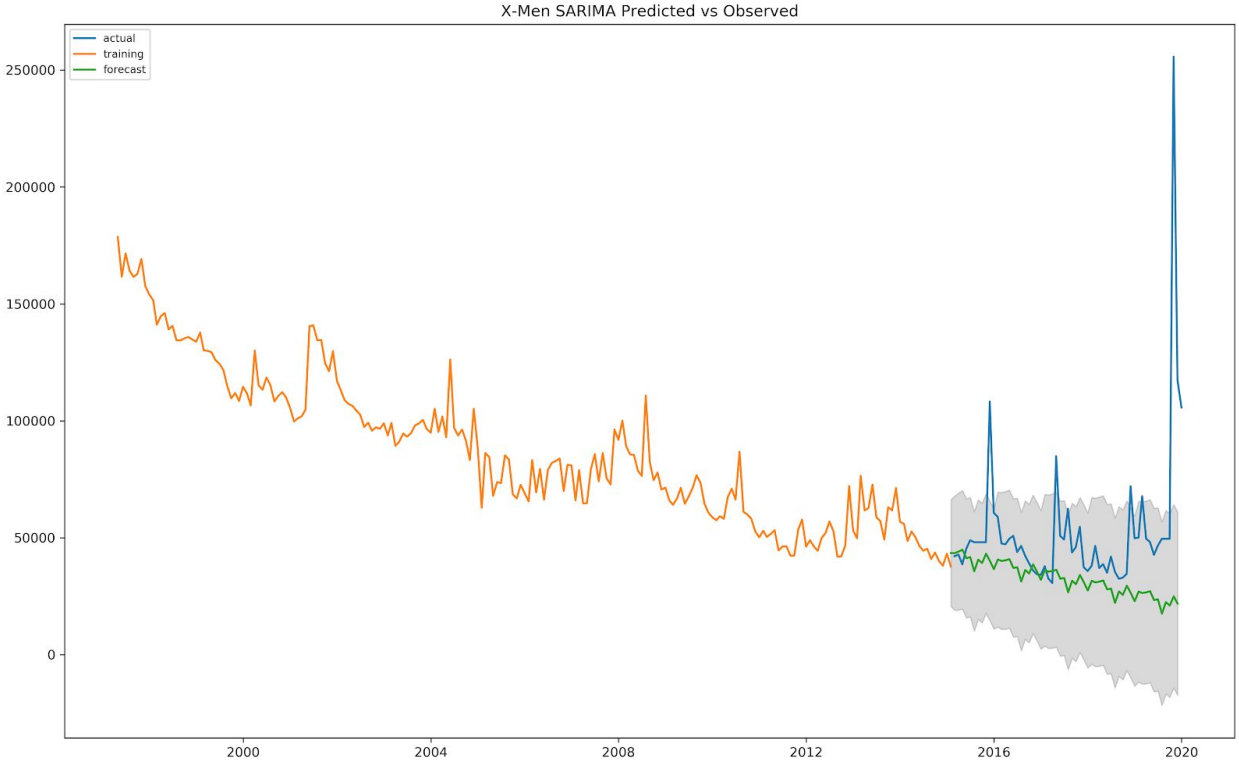
X-Men ARIMA Diagnostics





X-Men SARIMA Diagnostics





X-Men SARIMAX Diagnostics