Auto Production Forecast - Time Series Analysis

Paige Gonzales

January 10, 2024

Contents

Initial Project 2
Project Background 2
Data Source and Integrity 2
Initial Results 2
Project Motivation 2
Data Exploration 2
Data Preparation 3
Data Modeling and Forecasting 5
Project Expansion 8
Goals for Improvement 8
Expanded Results 8
Challenges and Further Work 12
Conclusion **12**
References **13**

Initial Project

I will start by discussing the initial assignment and why we chose our topic for analysis. I will then move on to explain the results from the initial project and what I will re-assess and improve upon. Finally, I will elaborate on my improved analysis and findings and discuss some challenges and future work.

Project Background

This project was completed in Fall 2023 in a course focused on time series analysis and forecasting. I worked with a team of four other people to find a real time series data set, apply time series analysis techniques to model the data, then forecast the future values using the created model. We used R and RStudio for the analysis and wrote a report of our findings in Microsoft Word. Some of the original time series models we created were auto-regressive integrated moving average (ARIMA) and seasonal auto-regressive integrated moving average (SARIMA). Time series modeling can be subjective, so to ensure our group chose an optimal final model, each of the team members worked independently in constructing and evaluating models. We then met up to compare results, ultimately deciding on one model to forecast future values. Once we agreed on a model and forecast, the team collaborated to write a report explaining the step-by-step process of constructing the time series forecast and communicating our results.

Data Source and Integrity

The data set for this project was downloaded directly from the Federal Reserve Economic Data (FRED) website, an online database that collects economic data from various sources (St. Louis, 2023). The source of this time series is from the U.S. Bureau of Economic Analysis (BEA) (Bureau of Economic Analysis (BEA), 2023). At the time of download in Fall 2022, this data set consisted of 357 observations, each representing the thousands of units of automobiles produced monthly in the United States. The data spans from January 1993 to September 2022. This information is publicly available by the government, therefore there is no issues with data privacy.

Initial Results

Project Motivation My team and I chose auto manufacturing as our subject of interest because when we started this project in September 2022, there were a lot of headlines in the news about the shortage of vehicles in the United States (Coffin, 2022). Additionally, there was, and still is, concern about the state of the United States economy post-COVID (Falkenburg-Hull, 2023). The auto industry represents 3% of America's GDP, and around 5% of U.S. jobs, so it seemed like an interesting and useful challenge to forecast auto production (Cutcher-Gershenfeld, 2015) (AAPC, 2020). Considering the importance of this industry to the American economy, having an unknown manufacturing rate poses a problem, decreasing the ability of America to predict the amount of revenue it can expect from the auto industry. Being able to predict this value is also valuable in determining if the auto manufacturing supply will meet the demands for auto products. Thus, forecasting plays a vital role in planning for the general economic outlook and meeting demands.

Data Exploration Figure 1 shows a time series plot representing the domestic auto production values in the United States since 1993. What initially sticks out is the major drop in production beginning in 2008. The recession of 2008 caused a major disruption in production when new vehicle sales declined 40% and several manufacturers were forced into bankruptcy (Dupor, 2020). In the decade following the recession, manufacturing slowly increased, but the COVID-19 pandemic caused a drastic decrease in production yet again. These two isolated occurrences exemplify how the auto industry is deeply entwined with the overall American economy and state of the country. Overall, auto production in the United States has been on a downward trend.

Another glaring point in the time series plot is an extreme drop in production in March and April 2020, which marks the beginning of the COVID-19 pandemic. In fact, the steep decrease in production in March 2020 is the lowest it has been in at least 30 years. When handling outliers in a time series analysis, it is often easiest to assume the outliers are errors, remove them, and continue with modeling (Sandbrink, 2020). My team and I were hesitant to do this because, although the outliers in question show a nearly 99% decrease from previous months values, the production values are a true reflection of output that occurred during the pandemic. In other words, numerically they are outliers, but they are real, nonetheless.

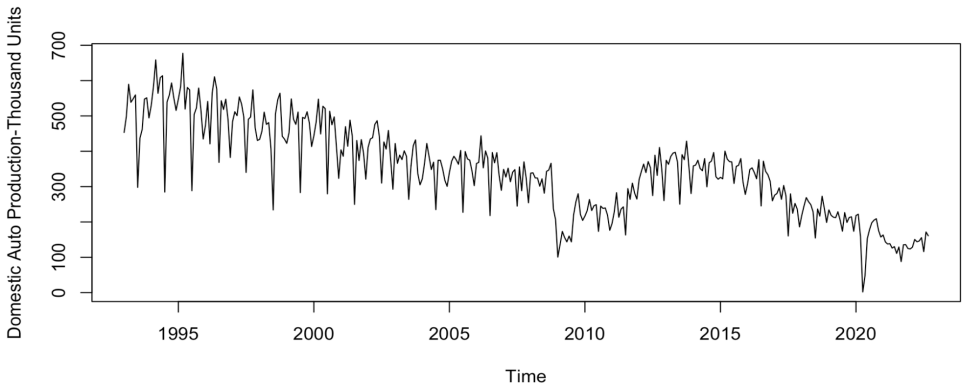

Figure 1: Time Series of Domestic Auto Production from 1993-2022

The pandemic was a historically unusual occurrence. To leave the values in the data set might risk creating a biased model, but to remove the values could cause problems with inaccuracy. After much discussion, we ultimately decided to keep the values in the time series data since they are true events. In addition, at the time we completed the project, the potential for more COVID-19 anomalies seemed more likely, and we didn't want to remove the values during a period of volatility.

Data Preparation One of the major assumptions of the models that will be created for this project is that the data is stationary, therefore, to produce valid models the data needs to meet three criteria (Singh, 2023):

1. Constant mean: the average value of the time series should stay constant over time.
2. Constant variance: the data should show consistency in its ups and downs over time.
3. No trend or seasonality: there should be no upward trends, downward trends, or any other patterns over time.

Looking at Figure 1, it is clear that this time series does not meet these three criteria. There is an obvious downward trend over time, and the variance is sporadic. Variance will look like major spikes in the data, where neighboring points vary widely in how different they are from each other. In Figure 1, you can see that the oscillation between highs and lows range significantly, indicating variance in the data. My team and I chose to stabilize the variance by applying a logarithmic transformation to the production values.

After stabilizing the variance in the data, the time series plot is much smoother, as shown in 2. However, there is a large drop during March and April 2020, as shown in the red box. This is to be expected since these values are so different from the rest of the data points. The next step is to address the trend in the time series.

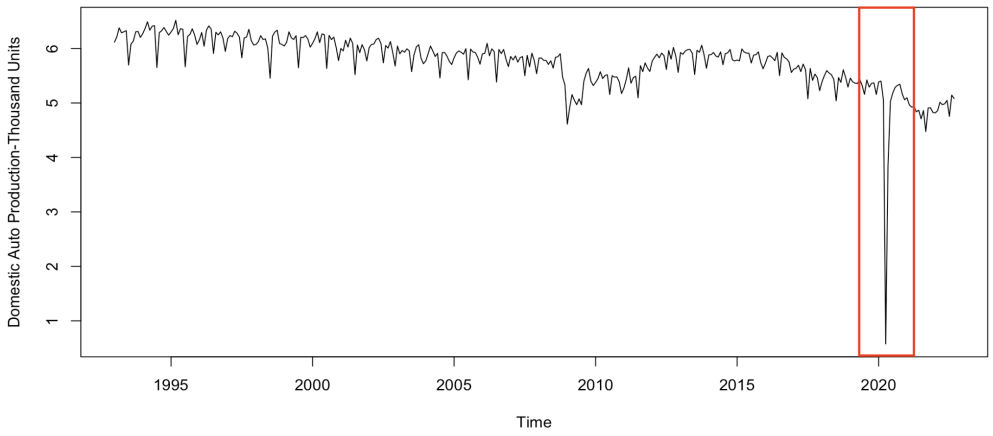

Figure 2: Time Series of Domestic Auto Production After Log Transformation

There are two ways we noticed that the data needed to be de-trended. The first would be the obvious downward trend that's shown in the original time series in Figure 1. However, after stabilizing the variance in the series, it is a little bit more difficult to notice the downward trend. The second method of checking for a trend is to look at the auto-correlation function (ACF), which is a measure of the correlation between a time series with itself at lagged time intervals (Smith, 2023). We want the plot of the ACF to show a sharp decrease as the lag, represented by the x-axis, increases. The plot in Figure 3 indicates the opposite, a slow decay over time, which needs to be corrected before modeling, as this violates one of the assumptions for the model.

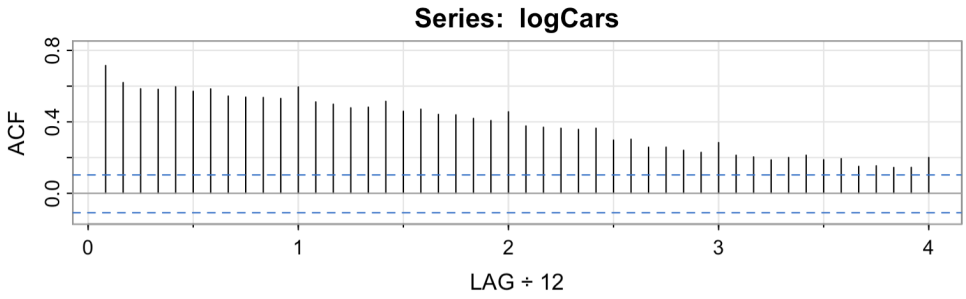

Figure 3: ACF of Domestic Auto Production After Log Transformation

To handle the upward trend in the data, we applied a first order differencing function to the log transformed data. This will remove the trend by subtracting each successive data point from the value before it. The time series shown in Figure 4 reflects the two changes made to the time series thus far. The data now appears quite constant over time, aside from the two 2020 points. Later we will see if this anomaly causes any major issues with the model.

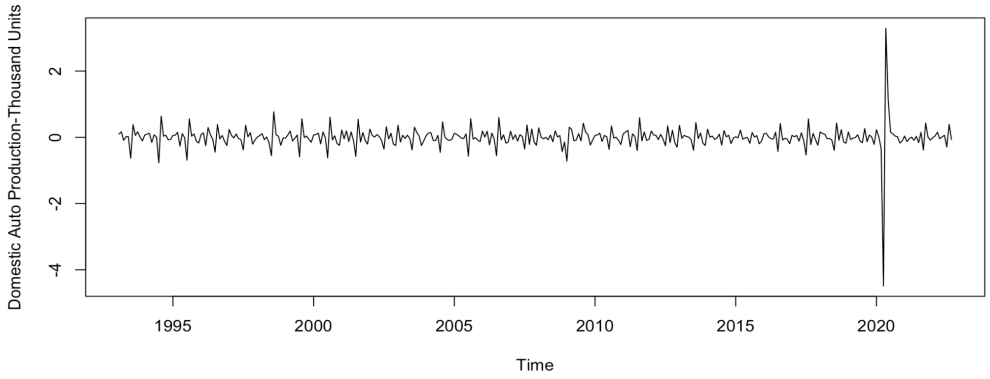

Figure 4: Time Series of Domestic Auto Production After Log Transformation and First Order Differencing

Next, we want to check for any seasonality in the data. Re-plotting the ACF function shows that there is a very clear seasonal component for auto production. This means that we can expect a similar pattern in production rates at certain points in the year. It looks like every year the lowest production occurs in July. There is no obvious month with a high production rate. There also shows a consistent uptick in production from September to October each year, followed by a decrease into November and December. March is a sporadic month, with some years having a very high production and some average or below average. Since there is an obvious seasonal trend, we need to include a component in our model to account for seasonality.

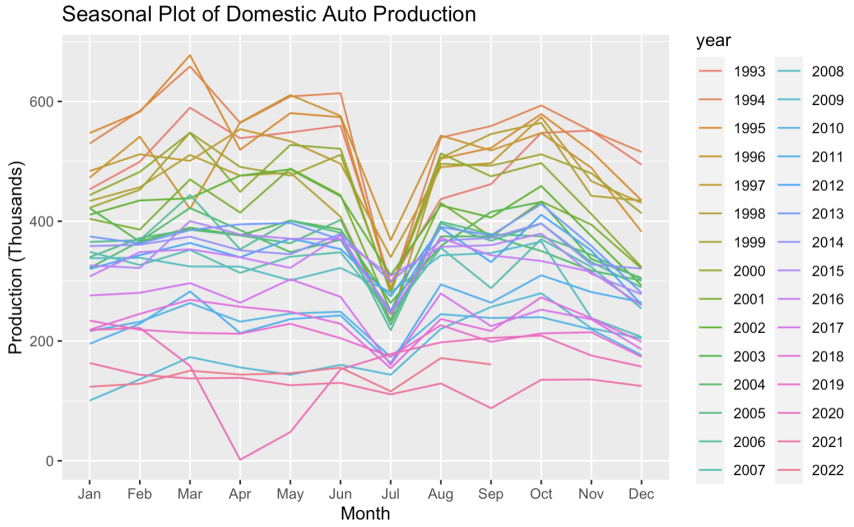

Figure 5: Seasonal Plot of Domestic Auto Production in the United States

Data Modeling and Forecasting The data now meets the three criteria for stationarity and is ready for modeling. The ACF and the partial auto-correlation function (PACF) will help to create an initial model to be evaluated and improved upon (Cahoy, 2022). The ACF and PACF plots are shown in Figure 6. Since it was determined the data has a yearly seasonality, we are applying a twelfth order differencing before plotting the ACF/PACF. There are many complex components and theorems in time series analysis, but for the purpose of explaining the logic in this project's results, I will note a few basic ideas on how to use these plots for model creation. For both the seasonal component, indicated by the higher values every twelve months, and the non-seasonal component, the following applies:

1. If the ACF decays rapidly over time and the PACF cuts off abruptly, then we can infer an optimal model is of the type auto-regressive (AR).
2. If the ACF cuts off abruptly and the PACF decays rapidly over time, then we can infer an optimal model is of the type moving average (MA).
3. If the ACF and PACF both decay rapidly over time, then we can infer the model is of the type (ARMA).

Looking at 6 for the seasonal data, the ACF cuts off abruptly at lag 1, and the PACF decays quickly, so it seems likely the model is of type moving average. For the non-seasonal component, the ACF cuts off abruptly at lag 1 or lag 2, and the PACF decays quickly, so it will also be of type moving average as well. As a reminder, time series analysis can be subjective, and interpretations may differ from analyst to analyst. Following this logic, we chose models $SARIMA(0,1,1) * (0,1,1,12)$ and $SARIMA(0,1,2) * (0,1,1,12)$ to begin with.

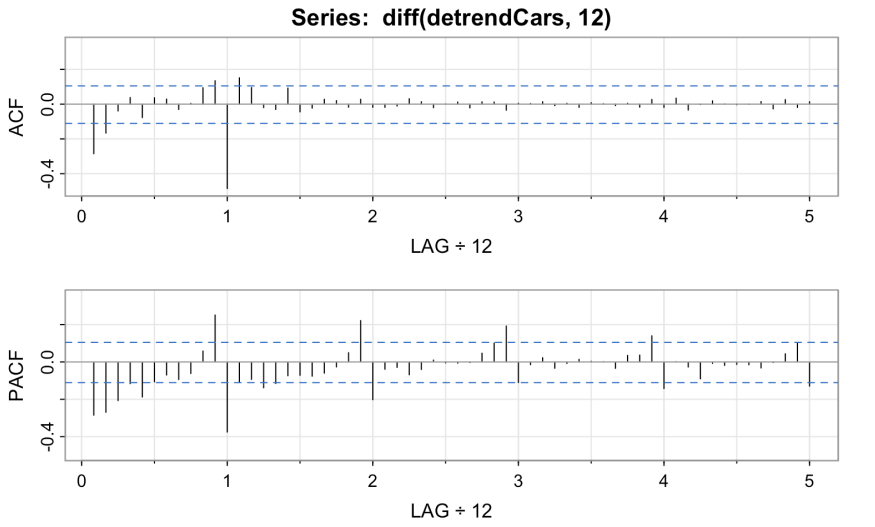

Figure 6: ACF and PACF Plots of Stationary Auto Production Time Series

The final model that my team agreed on was $SARIMA(0,1,2) * (0,1,1,12)$. This model's diagnostic plots, shown in Figure 7, suggest that the model is a good fit for the following reasons:

1. The Ljung-Box statistics, shown in the bottom plot, all had p-values greater than the rejection value marked by the blue dashed line, which means we fail to reject the null hypothesis that the residuals of the time series model are independently distributed. This is an assumption made when creating a model of this type, so it is ideal to fail to reject the null hypothesis (Bobbitt, n.d.).
2. The ACF of the residuals in the middle-left plot resembles white noise (no trend) and shows stationarity.
3. The qqplot on the middle rights shows most of the points are on the qqline, meaning the residuals are nearly normally distributed.

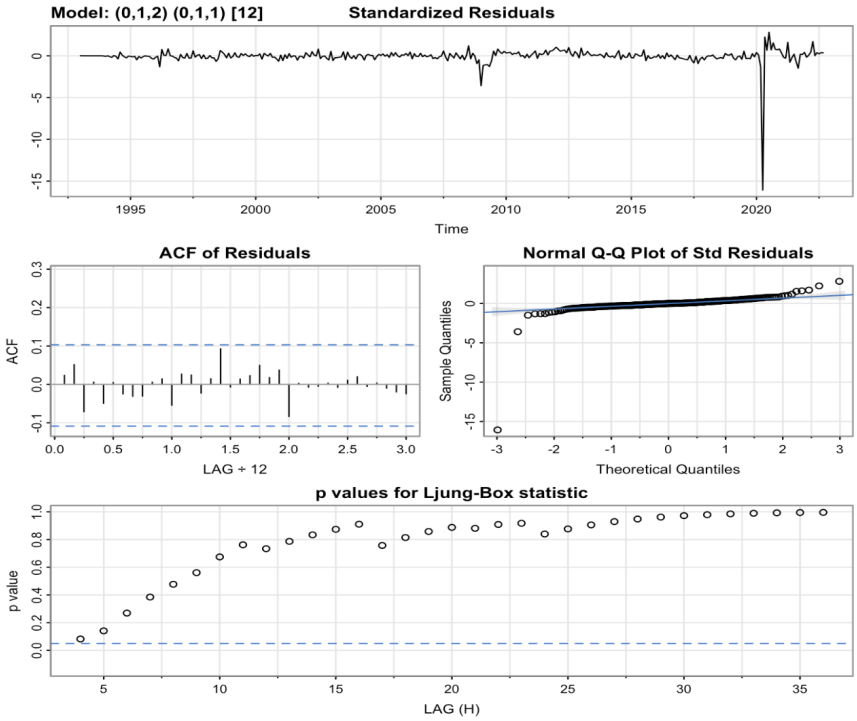

Figure 7: Diagnostics of SARIMA Model SARIMA(0,1,2)*(0,1,1,12)

We then applied the model to forecast the next twelve months, starting with October 2022, as shown in Figure 8. The predictions from October 2022 forward show an overall decrease in production value in the next year.

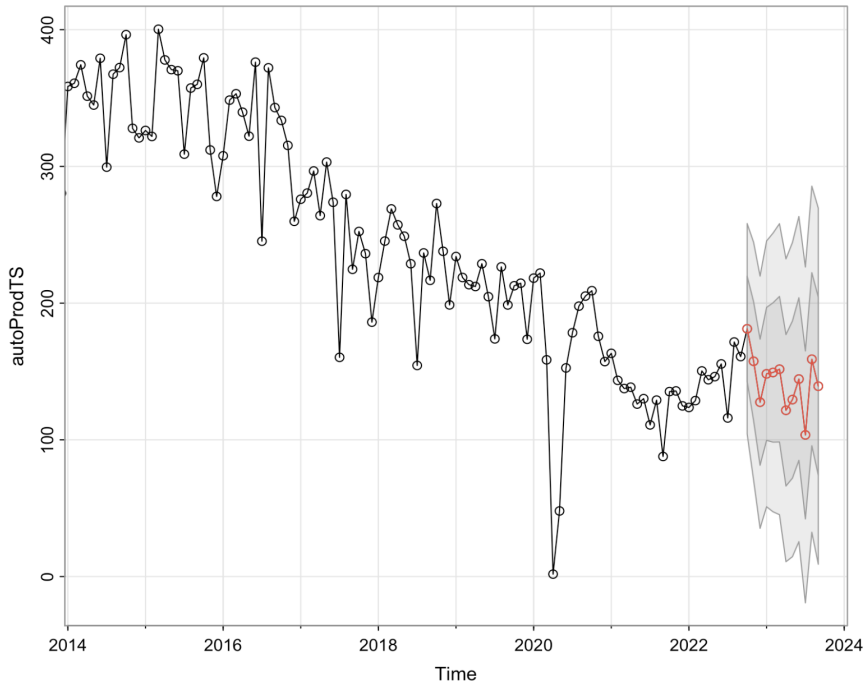

Figure 8: 12-Month Forecast For Auto Production Using Model SARIMA(0,1,2)*(0,1,1,12)

Project Expansion

Goals for Improvement

Reassessing this time series nearly a year later could provide some useful insight into how effective the model we created turned out. In addition, I wanted to test the effects of the major outliers during COVID-19. It has been nearly a year since this project concluded, and I am still curious about whether we handled the outliers in the best way. I will first reassess the original forecast, comparing the true values from October 2022 to May 2023 to see how close our model came to reality. Next, I will handle the COVID-19 outliers by instead smoothing those two data points and remodeling. Finally, I will improve upon the model with the most recent data set and revise the forecast to predict the next twelve months of auto manufacturing.

Expanded Results

I downloaded a newer, updated version of the data set from (St. Louis, 2023), which shows production units up to May 2023, adding an additional eight months to the original project's data set. When comparing the forecasted values to the actual values for October 2022 to May 2023, shown in Figure 9, the first five months are very similar to reality. However, starting in March 2023, the predicted values started to become much lower than the actual values, ranging between 16% to 22% difference. In other words, our model only worked in the short term.

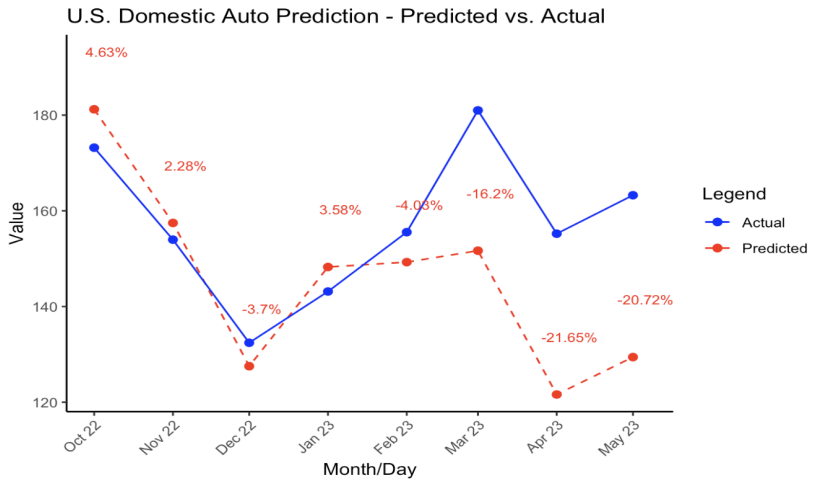

Figure 9: Predicted Versus Actual Production Values From October 2022 to May 2023

I wanted to see if the model would have improved had we removed the two COVID-19 outliers. To verify this, I decided to smooth this portion of the time series, replacing those two values with an average of the nearby values using function *na.approx* in R package *zoo* (RDocumentation, n.d.). In the below Figure 10, the blue line represents the original time series and the red line represents the smoothing.

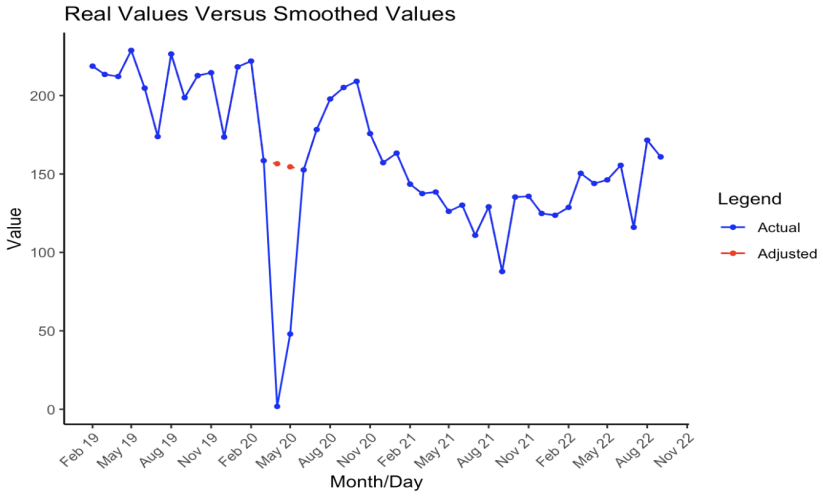

Figure 10: Actual Values Versus Smoothed 2020 Values in Auto Production Time Series Plot

When checking the ACF and PACF of these models, there is no obvious choice of whether the model is AR, MA, or ARMA. This indicates that removing these two values has disrupted the original pattern in the data. When re-running the same model as previously, the diagnostics of the results (Figure 11) are not nearly as promising for several reasons:

1. The Ljung-Box statistic plot, shown in the bottom plot, had all p-values less than the rejection value marked by the blue dashed line, which means we reject the null hypothesis that the residuals of the

time series model are independently distributed. This means the model fails to meet the assumption made when creating a model of this type (Bobbitt, n.d.).

2. The ACF of the residuals in the middle-left plot shows non-stationarity, since there are several ACF values past the blue dashed line.

3. The qqplot on the middle rights shows several points deviating from the qqline, meaning the residuals are not normally distributed.

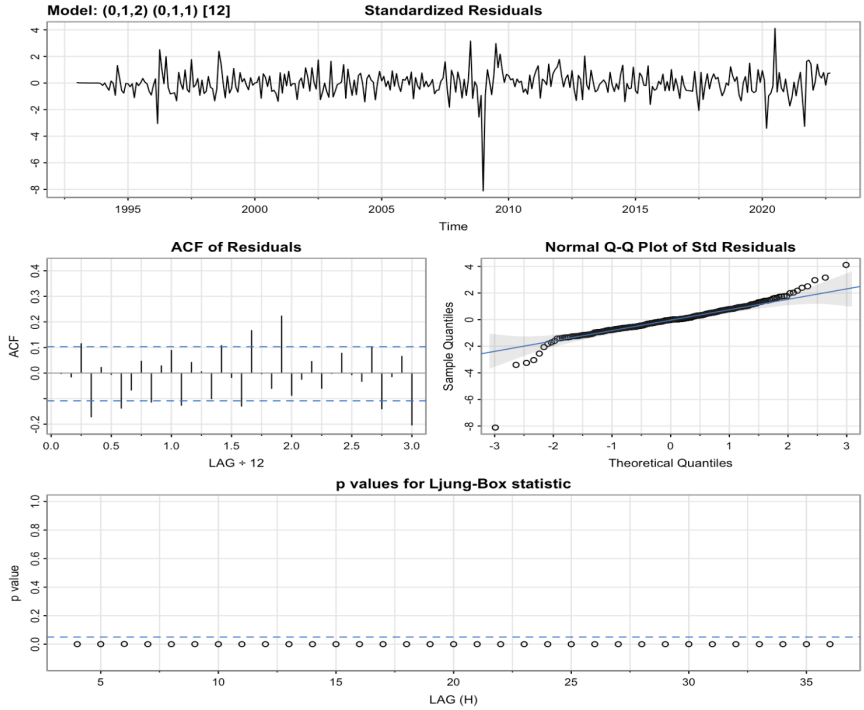

Figure 11: Diagnostics of SARIMA Model SARIMA(0,1,2)*(0,1,1,12) With Smoothing

Although I did not have high hopes for this data considering the diagnostics of the original model, I still tried to find a suitable model for the smoothed data. I ran several different models to try and find a best fit, but all showed sub-par diagnostics.

The fact that no model "fit" is interesting because even though the model diagnostics are poor, the actual predictions resemble the true values much more than the model with acceptable diagnostics, as shown by the green line in Figure 12.

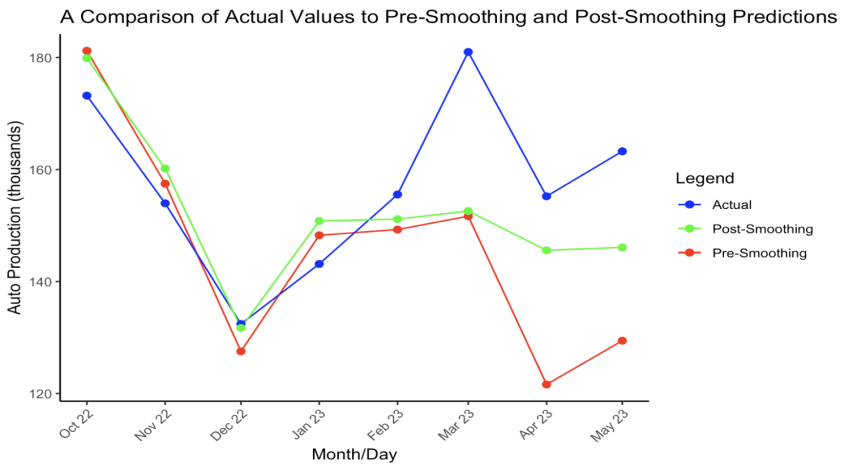

Figure 12: Comparing Pre/Post Smoothing Predictions to Actual Values

My next step was to update the data set with the most recent month's values and re-run the model to see if it is still a good fit or should be adjusted. I did this for both the COVID-19 smoothed version and the original version. Since the smoothed model had poor diagnostics when applied to the original SARIMA model or any other model, I didn't expect there to be any improvement when 8 extra values were added. As predicted, the diagnostics still showed that this model was not a good fit for the smoothed data.

The data set with production in March 2020 and April 2020 kept as their original values shows the chosen SARIMA model as acceptable with nearly all diagnostic plots identical to the original project's model. It is intuitive to conclude that our original model is still optimal, since only eight values have been added to the original 357 data points. At this point in my analysis, I was unsure of which model to choose for forecasting the next twelve months. While the original model using non-smoothed data had ideal diagnostics, the forecast didn't line up with the prediction last year, so it seems like the model wouldn't be the best fit in application .The model with "bad" diagnostics using the smoothing method looked more realistic. After some thought, I decided to use the data with the outliers included. I created a new forecast for the next 12 months, starting with June 2023, shown in Figure 13. The forecast continues to show a decline in production for the next twelve months.

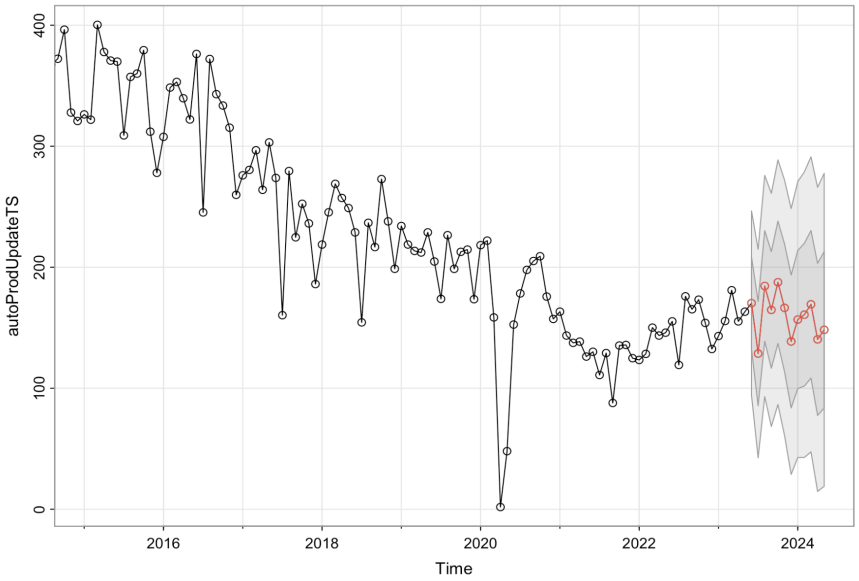

Figure 13: Updated Time Series Forecast Beginning June 2023

Challenges and Further Work

The challenge of this project remains the same as what we originally noted when first completing it nearly a year ago. The COVID-19 pandemic caused significant economic impact. Even looking at auto production only, you can see out of the past 30 years, nothing similar has come close to the economic shock that occurred in 2020. It is difficult to decide how to handle this issue because it isn't something that happens often. When researching this topic, I read about several different options on techniques that can be used to handle an occurrence like this, and I have only touched on one of them. In future work, I would like to try an alternative approach to handling those COVID outliers by applying a regressor (Sandbrink, 2020). I also plan to reassess this time series every few months to continue to optimize model performance.

Conclusion

The three projects in my portfolio were chosen because they showcase the wide range of knowledge gained from the last 18 months spent in the MSDA program. These projects rely on academic understanding of advanced statistical techniques of time series analysis and machine learning algorithms, heavy data visualization skills, and extensive R coding experience. I hope this portfolio can serve as an introduction to what potential employers can gain in hiring me. As always, my work is never done and I plan to further strengthen my skills by self-studying, specifically focusing on data analytics in Python and Tableau. I know much more now that I did when I first began this program, and I am excited to see what the future holds.

References

AAPC, A. A. (2020). *US economic contributions*. https://www.americanautomakers.org/us-economic-contributions

Bobbitt, Z. (n.d.). *Ljung-box test: Definition + example*. https://www.statology.org/ljung-box-test/

Bureau of Economic Analysis (BEA). (2023). *U.S. Bureau of Economic Analysis (BEA)*. https://www.bea.gov/

Cahoy, D. (2022). *Chapter 3. Autoregressive integrated moving average (ARIMA) models*. University of Houston - Downtown.

Coffin, D., David. (2022). *The roadblocks of the COVID-19 pandemic in the u.s. Automotive industry*. https://www.usitc.gov/publications/332/working_papers/final_the_roadblocks_of_the_covid-19_pandemic_in_the_automotive_industry.pdf

Cutcher-Gershenfeld, J. (2015). *The decline and resurgence of the u.s. Auto industry*. https://www.epi.org/publication/the-decline-and-resurgence-of-the-u-s-auto-industry/

Dupor, B. (2020). *Auto sales and the 2007-09 recession*. https://research.stlouisfed.org/publications/economic-synopses/2019/07/05/auto-sales-and-the-2007-09-recession

Falkenburg-Hull, E. (2023). *How the auto industry will be disrupted in 2023, according to experts*. https://www.newsweek.com/2023/04/14/how-auto-industry-will-disrupted-2023-according-experts-1791933.html

RDocumentation. (n.d.). *Na.approx: Replace NA by interpolation*. https://www.rdocumentation.org/packages/zoo/versions/1.8-12/topics/na.approx

Sandbrink, J. (2020). *How COVID-19 messed up our time series*. https://medium.com/@jorritsandbrink/how-covid-19-messed-up-our-time-series-6d84516f80c0

Singh, J. (2023). *Why do time series have to be stationary before analysis?* https://www.tutorialspoint.com/why-do-time-series-have-to-be-stationary-before-analysis#:~:text=The%20model%20will%20not%20be,series%20being%20stationary%20before%20analysis

Smith, T. (2023). *Autocorrelation: What it is, how it works, tests*. https://www.investopedia.com/terms/a/autocorrelation.asp

St. Louis, F. R. B. of. (2023). *What is FRED?* https://fredhelp.stlouisfed.org/fred/about/about-fred/what-is-fred/