Time Series Modeling Project

Kyle Wu

2023-01-30

```
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'
```

```
## Rows: 889 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): CUSR0000SETA01, DAUPSA, TOTALNSA, CUSR0000SETB01
## dbl  (1): MPRIME
## date (1): DATE
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

R Markdown

New Car CPI

Prime Rate Loans

Total Vehicle Sales

Domestic Auto Production

GAS CPI

Data Application

In the United States, one of the main modes of transportation is the automobile. To the average consumer, it has seemed that the new car has been slowly getting out of reach, with the average price of a new vehicle currently sitting around $49,500 ("No End in Sight: New Vehicle Transaction Prices End 2022 at Record Highs, According to New Data from Kelley Blue Book" 2023). Despite the high prices of vehicles for many Americans a car is not only a luxury, but a necessity, and many citizens find themselves shelling out a large portion of their paychecks for their transportation. Even when individuals opt to purchase used cars, they are still often faced with prices that would have seemed exorbitant not that long ago.

Since this is the case, studying the United States car market over time will allow us to gain useful knowledge that will be of significance not only to the average consumer, but also for economists trying to understand what trends the American auto market may be facing going forwards and what factors influence automotive sales. Past research into the American auto market have been vital to our understanding of the forces driving the auto market. For example, it is well known that the chip shortage that occurred as a result of COVID-19 shutdowns, among other reasons led to a chip shortage that has in many ways created problems for the world economy ("Inflation and the Auto Industry: When Will Car Prices Drop" 2022). Since cars now heavily rely on computers to work, this resulted in many manufacturers around the world decreasing production projections, which decreased vehicle production, and partially led to the rapid rise in vehicle prices. However, if we look at production figures, we can see that the domestic production of cars had been following a decreasing trend since the 90s, so researchers at Federal Reserve Economic Data (FRED) found that it is hard to say if COVID was fully responsible for the decreased production, or if it would have happened regardless ("Long-Term Trends in Car and Light Truck Sales" 2021). Research by FRED also indicated that despite the increase in population since the mid 1970s, the total number of vehicles sold has remained relatively flat aver the past few decades ("What's Been Drivin the Rise in Auto Prices Since COVID" 2022).

Looking at the data offered by the Federal Reserve could allow us to answer even more questions regarding the American auto market. For example, we could try to understand if it is likely that american automakers would have decreased their production numbers even without the disruptions brought about by COVID or if COVID led to new trends. If we take into account other economic factors, such as interest rate or gas prices, we can then try to measure what economic factors may most affect the sale of motor vehicles. Using the data we obtained and after determing factors that determine automotive sales, we can then create a forecast to determine how each factor relating to the automotive industry will change in the future. For example, we can try to answer the question of whether it is likely new vehicles will continue facing inflation or if it might become stable in the near future. Besides looking at the various factors individually, we can also look at the auto market holistically, asking what the future may be in terms of vehicle purchases in the United States and is it likely that vehicle purchases return to pre-COVID levels. For the average consumer, the questions that will be answered will allow them to perhaps better plan for the expenditure that comes with the purchase of a new car.

Furthermore, studying time series data of vehicles can allow us to better understand how or if certain policy changes may change vehicle prices or purchasing behavior. For example, we could potentially find time periods with varying federal funds rates, which influence bank prime loan rates to see if this changed the overall behavior of consumers.

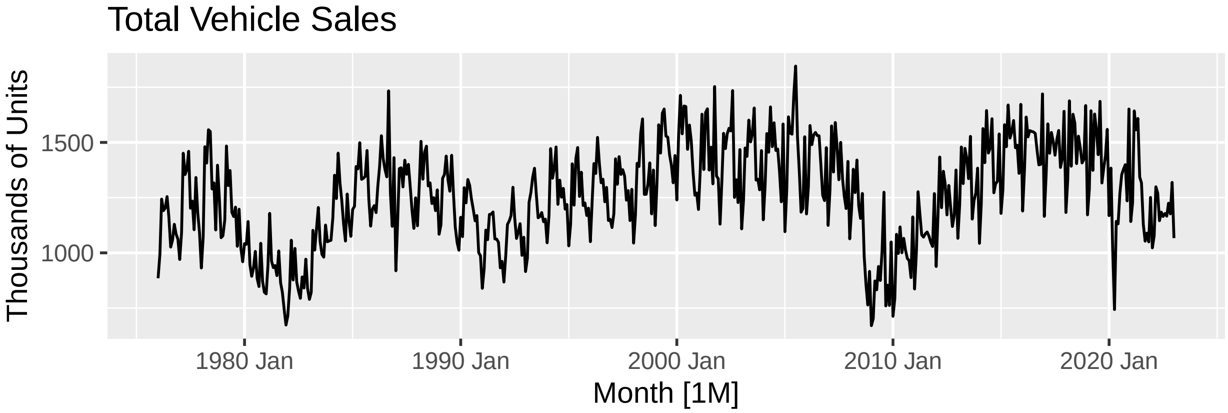
Gathering all this data about the American auto market would then allow us to broadly gain an understanding not only of factors affecting vehicle sales, but also of the health of the American economy due to the fact that vehicles are often the second most expensive possessions of individuals, second only to homes. Increased purchasing of vehicles would indicate that the American has been healthy and following a positive trend, whereas decreased vehicle purchases may indicate that the economy had been following a general downwards trajectory.

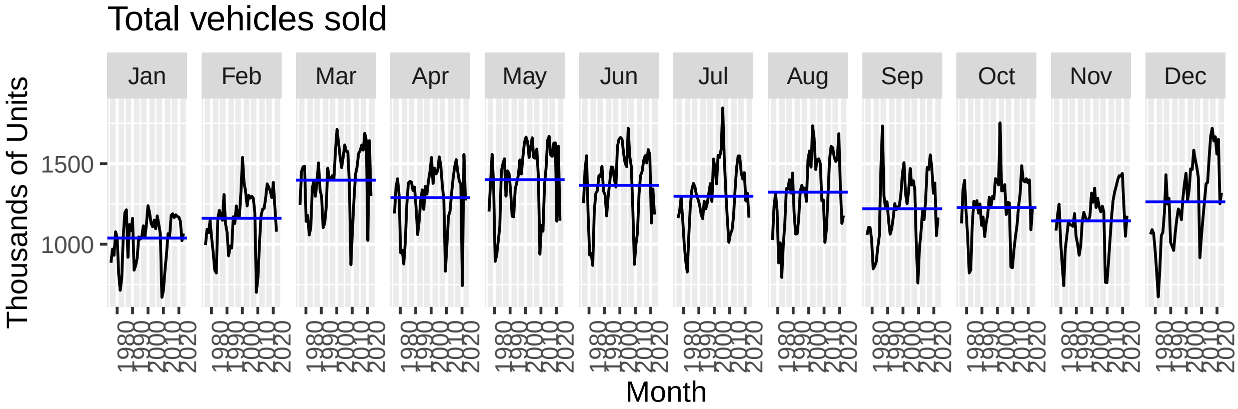
Analysis of Empirical Properties

In all cases, the data I selected came from the Federal Reserve Economic Data database and all variables selected were recorded on a monthly basis and input into a format that was very neat and effective for time series analysis. The variables I have chosen are Total number of vehicles sold, new vehicle consumer price index, domestic auto production, fuel price index, and the bank prime rate. In this study, I will use total number of vehicles sold as the gauge of the american auto market, and the other variables will be used as predictors.

We will first analyze the variables individually before talking about all the factors as they may relate to projecting future car sales.

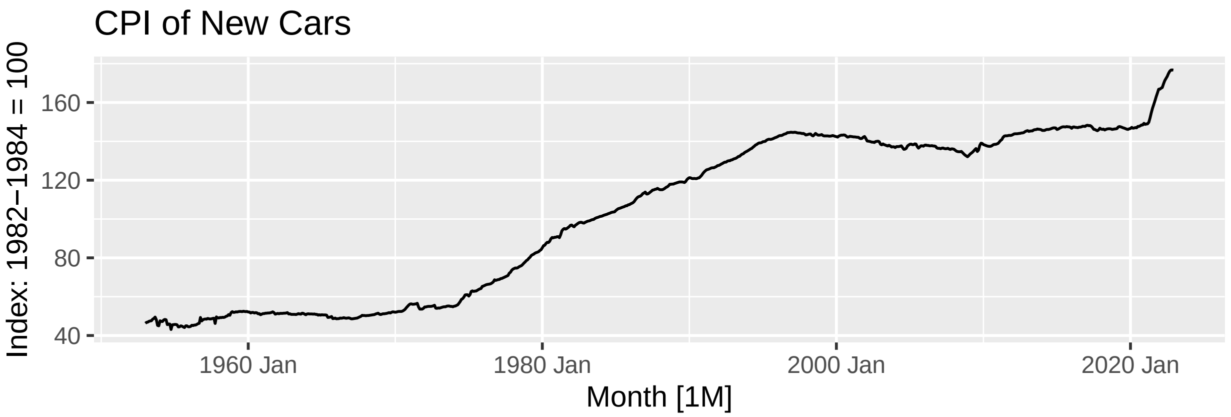
The first variable we will look at is total vehicle sales.

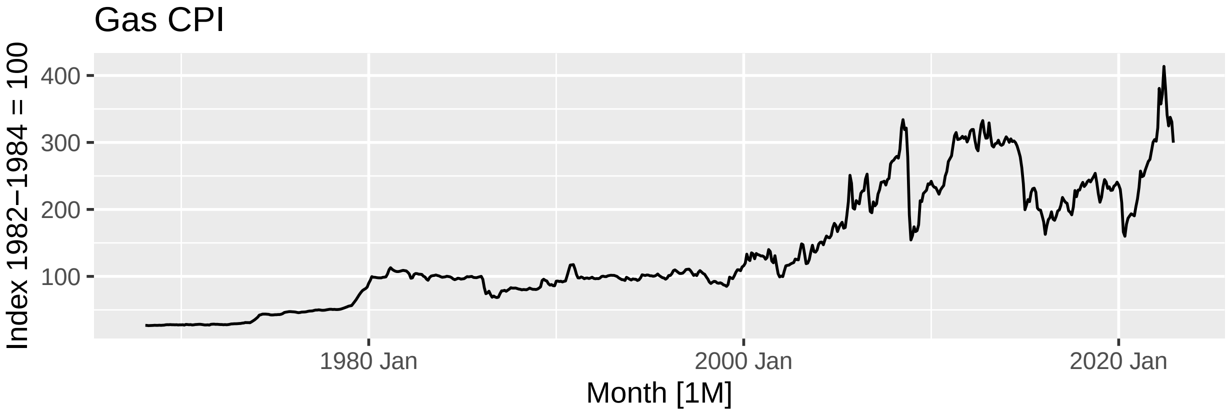




Our data for vehicles sold starts from January 1976, is recorded monthly, and does not have any missing values.The data was collected from the U.S. Bureau of Economic Analysis, which has formatted the data in a easily usable database. From the subseries plot, we see that the data definitely follows a seasonal pattern, as demonstrated by the fluctuating mean lines based on the month of the year. Additionally, from the scalloped shape of the plot, we can see that the data follows a seasonal trend of peaks and troughs every twelve months. Logically this makes sense, as it is well known that vehicle sales usually increase during the summer and tend to decrease during the winter. From the data we can determine the number of vehicles sold over the last 5 decades and whether there has been a general trend in vehicle sales.

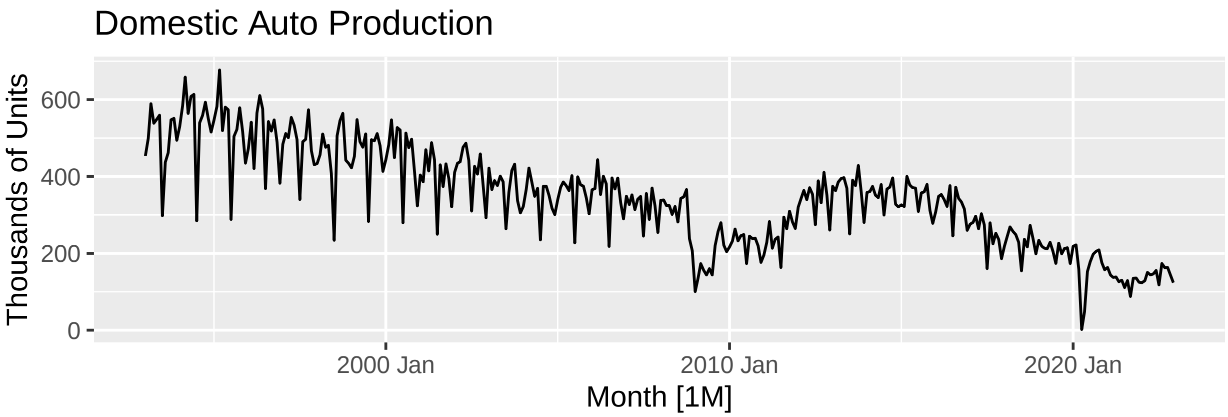
We can now look at the consumer price index of new vehicles in the United States.

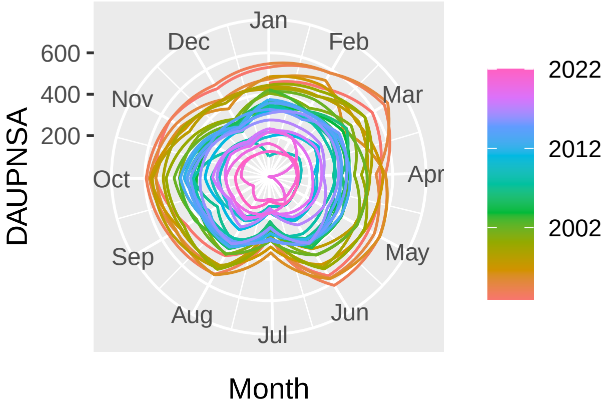




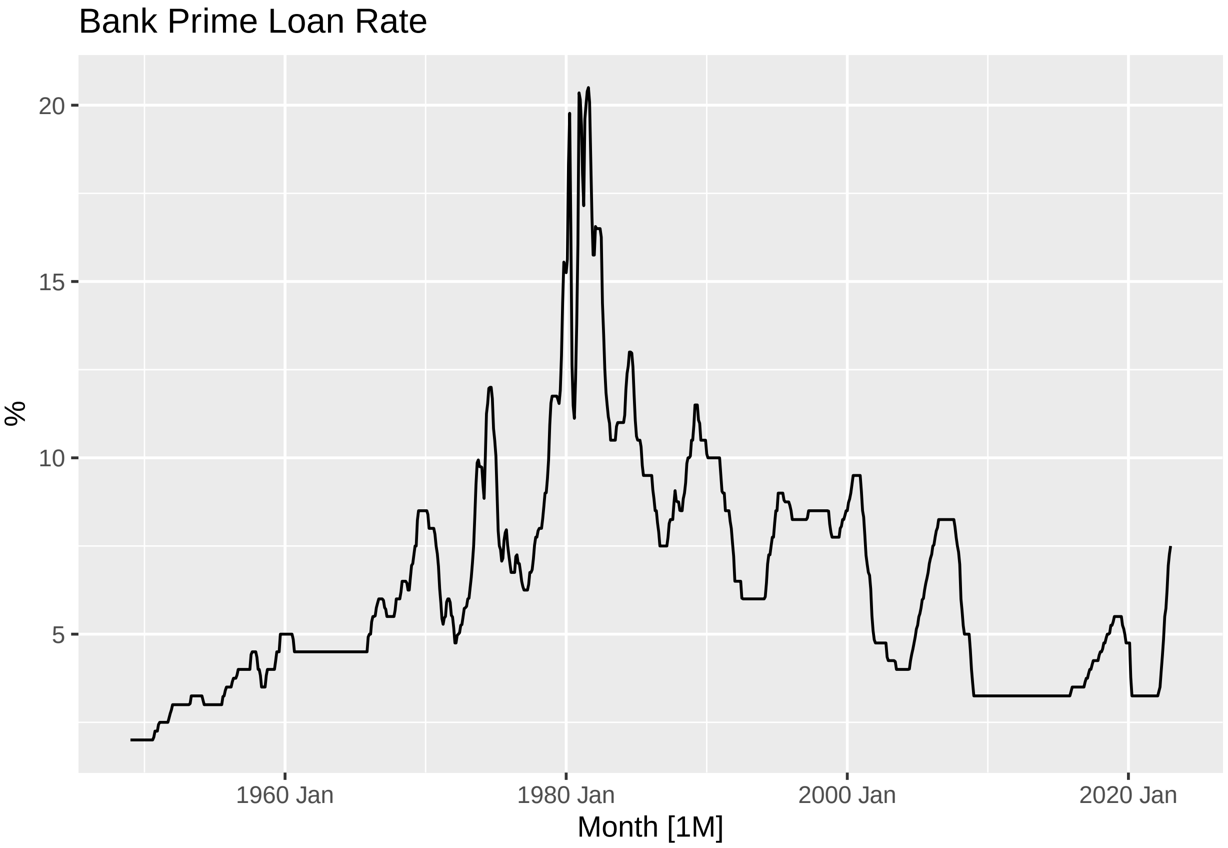
The data for the consumer price index measure of new cars begins in January 1953 and is recorded monthly. This dataset has already been seasonally adjusted so that will not be something that I will have to worry about later. It is worth mentioning that CPI should not be regarded as the price of a vehicle, but is a measure of how much more or less an item cost relative to the base year (1982-1984 in this case). In essence, this CPI measures the spending power consumers have over the good. Overall, there does not appear to be any sharp fluctuations in the data but we do have a couple periods of relatively rapid vehicle price inflation, namely from the 70s to the late 1990s and then again after the arrival of COVID-19 in 2020. It is interesting to see that from the late 1990s to 2020, the CPI value of cars did not change drastically.

Records for the price index of gasoline start from February 1968 and are recorded monthly by the U.S. Bureau of Labor Statistics with no missingness present. From the data, we see that there has been a general increase in the price of gasoline from 1968 until the early 2000s before prices seem to level off, but with high volatility. We also see that following COVID, there was a large spike in gasoline prices, which has since gradually come down. This data will allow us to understand what trends are present in terms of gasoline prices and whether gasoline prices have an effect on the total number of vehicles sold. In other words, do high gasoline prices lead to less consumers buying vehicles?





For Domestic Auto Production, we have monthly recorded data from the U.S. Bureau of Economic Analysis with no missingess present in the data. Overall, we appear to see volatility month-to-month and pretty consistent seasonal patterns with July being a month with consistently low production numbers in relation to other months and March and October having relatively high production numbers. This data alone can answer a couple questions about the United States auto market. First of all, we can answer if the COVID induced supply shortages drastically impacted U.S. vehicle production numbers or if, as suggested by FRED COVID simply highlighted a long-term trend that would have occurred regardless. Using this data will then also allow us to make forecasts on what the future outlook may be for domestic auto production.



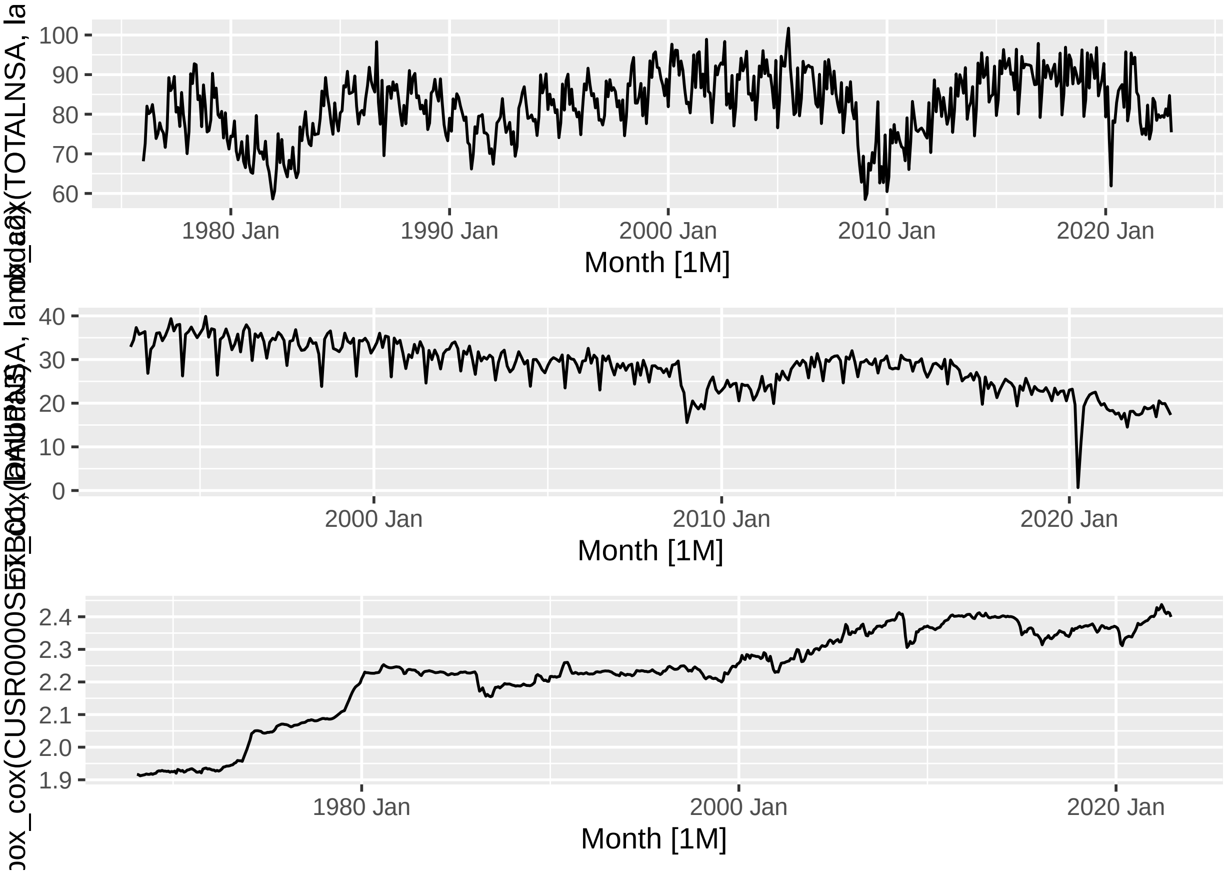
The Bank Prime loan rate is set in relation to the federal funds rate set by the federal reserve and we have data going back to January 1949. The rate is often set based on other determinants of the market so we may not be able to generate too many insights from it individually. However, if we use the bank prime rate in addition to the other factors we discussed earlier, we may be able to create a regression model that will allow us to understand what economic factors influence the car purchasing decisions of Americans. Additionally, studying all the data together will allow us to understand if COVID-19 disruptions altered the purchasing decisions of Americans or if there were vehicle market trends that would have likely occurred regardless of the pandemic.

Models for Data Fitting

Throughout this case study, we will aim to fit models that allow us to analyze and forecast future values of our time series using the Box-Jenkins Methodology. There are three main stages to setting up a Box-Jenkins model. The first step is to examine the data and to see which parts of the ARIMA process appear to be the most appropriate to apply. The second step is to estimate the parameters of the chosen models. Dinally, the third step is diagnostic checking where we examine the residuals from the fitted morel to see it they appear to be sufficient. If the model we originally chose is insufficient, then we should try other models until a sufficient model is found.

The first step to fitting our model is to see if any of our time series need to be transformed or adjusted, which often allows us to analyze simpler time series. Adjusting the data will allow us to make patterns more consistent across our data, which will allow us to better model the data, and will lead to better forecasts. For the metrics I used, the only ones that may need transformations due to variations in seasonality are those for `daupnsa` and `totalnsa`. Since these two series show variations that change with the level of the series, we can use Box-Cox transformations to stabilize the variance. Box-Cox transformations depend on a parameter λ and can be defined as follows:

$$y(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}.$$

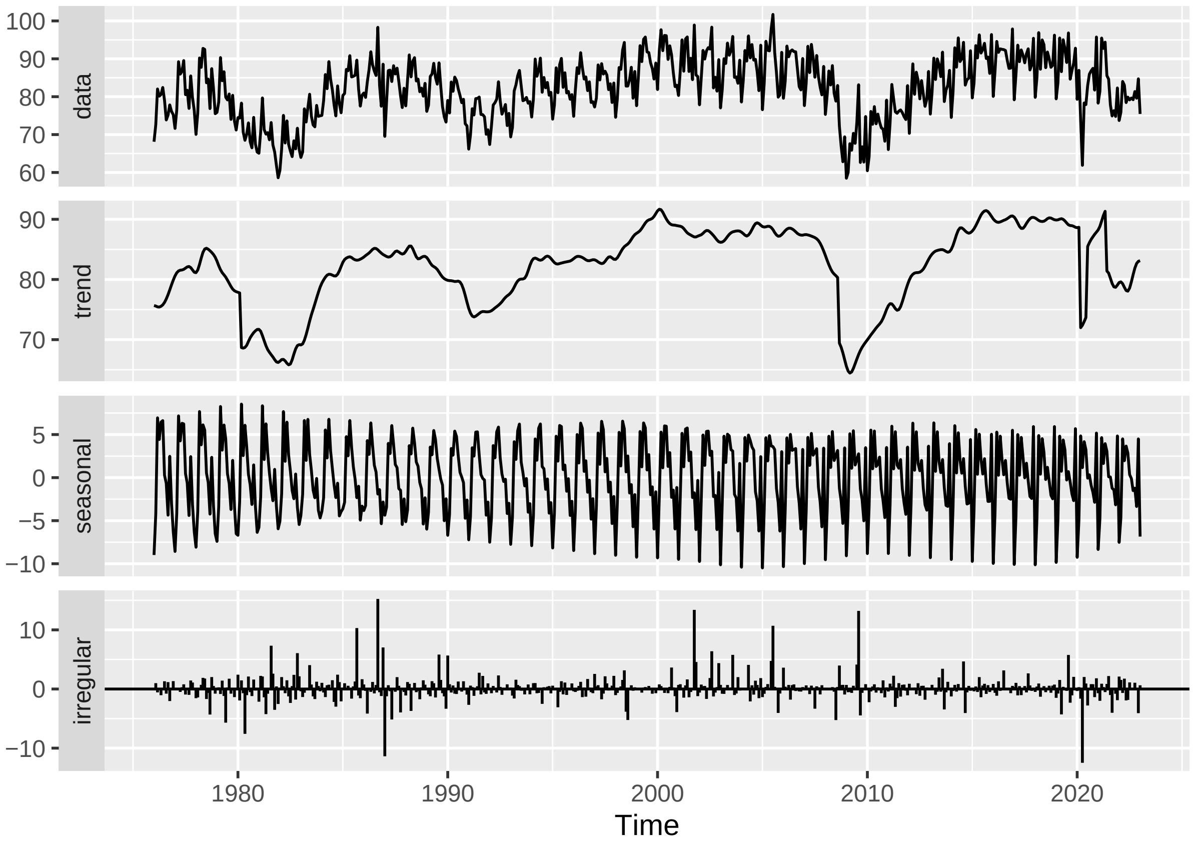


After running the calculations, we find that the appropriate λ value for total auto sales was approximately 0.53, the appropriate λ value for domestic auto production was 0.45, and the appropriate λ for gasoline CPI was -0.36. From the plots that we created above, we also see that we were able to successfully adjust our data since it appears that for the most part, our seasonal effects have been made consistent over the range of out time series.

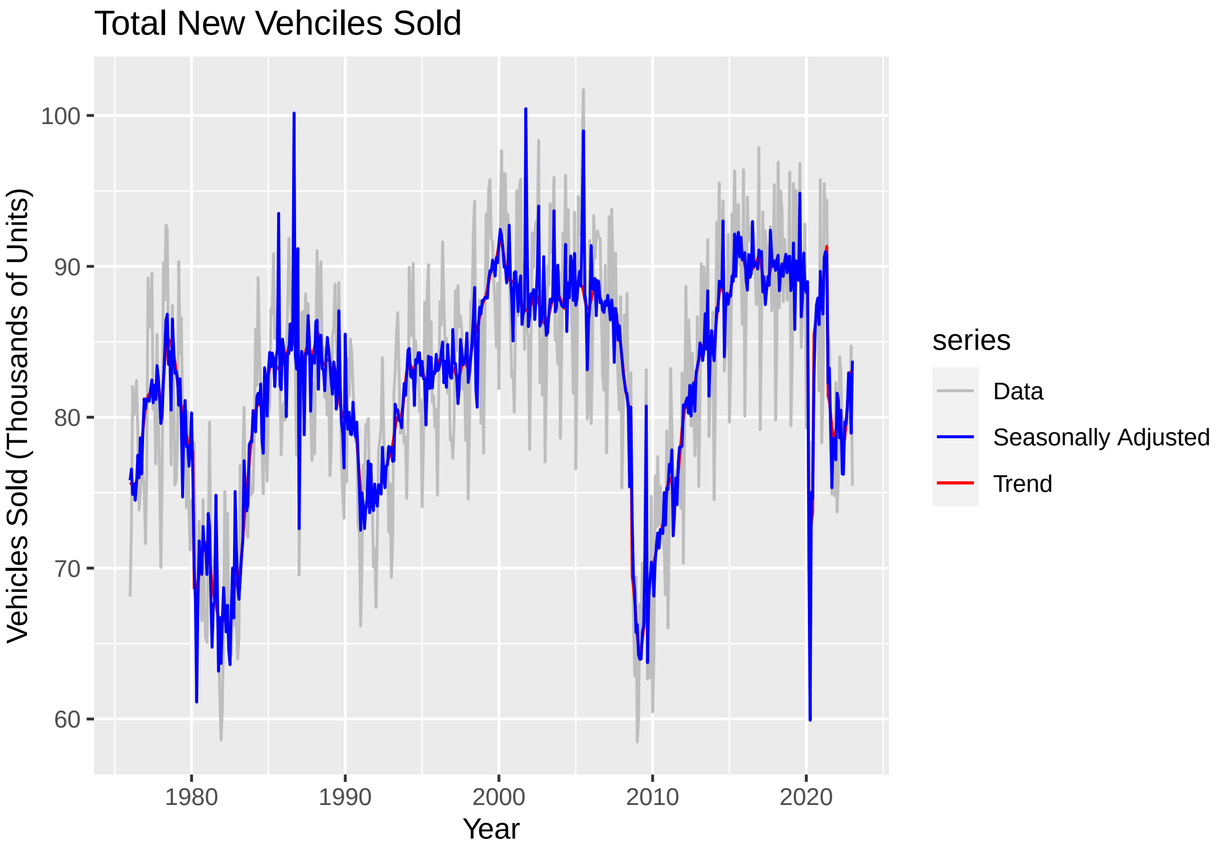
The next step to fitting our model is to decompose our data into the time series' separate components. Depending on the type of data we have, there are two possible ways that we can decompose the data. The first possible case of data is a time series with a trend but no seasonal variations, which follows the form $X_t = \alpha + \beta t + \epsilon_t$, where α, β are constants and ϵ_t is a random error term with mean zero. Since our CPI data for gasoline prices and new car prices are both seasonally adjusted, this is the type of decomposition they will undergo.

The second possible case of data is a time series that contains both a trend, and seasonal variation. In this case, there are two possible cases of decomposition we can take. If we assume additive decomposition, then we can use $y_t = S_t + T_t + R_t$, where S_t is the seasonal component, T_t is the trend component, and R_t is the remainder component. The other possible case involves multiplicative decomposition which follows the form $Y_t = S_t \times T_t \times R_t$. We use the multiplicative case if the variation around the trend-cycle appears to be proportional to the level of the time series ("PSTAT 174/274: Time Series Part IV" 2023). The additive case would be if the magnitude of the seasonal fluctuations don't significantly vary over time. Since we used a box-cox transformation to stabilize our variance, we can use the additive case in our decomposition.

In our case, we will use an X-11 decomposition, which is commonly used by the U.S. Census Bureau. X-11 decomposition works by estimating the trend by a moving average, removing the trend to leave the seasonal and irregular components, and estimating the seasonal component using the moving averages to smooth out irregularities. Due to the fact that seasonality cannot be identified without the trend and a trend cannot be estimated without data beign seasonally adjusted, X-11 models use an iterative approach to come up with the best solution. An example of X-11 decomposition is shown below



From the decomposition of this time series, along with the other time series, we can get a general idea of the trend and seasonality present within the plot, which will then allow us to better estimate and thus forecast the data moving forwards. Additionally, following the decomposition, we can then see how the trend and seasonally adjusted versions of our data fit with the original data.



For this case study, our goal is to eventually forecast future values of each of our time series data through the use of ARIMA models. The goal of ARIMA models is to describe autocorrelations in the data for forecasting (Rob J Hyndman 2021). When attempting to fit ARIMA models, we attempt to first remove any trend or seasonality present within the data so that we can create stationary time series, which will allow us to attempt to model the remaining residuals. Stationary time series are time series whose mean and variance is constant over time (Glenn, n.d.).

One way to convert non-stationary time series into stationary time series is the method of differencing, where we compute the differenece between consecutive observations in a time series and can be written as follows

$$y'_t = y_t - y_{t-1}$$

Differencing time series allows us to stabilize our means because it removes or reduces trend and seasonality. Furthermore, if our data still does not appear stationary, it may require second-order differencing, which follows the form

$$y''_t = y'_t - y'_{t-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2}$$

Yet another way method to convert non-stationary time series into stationary time series is through seasonal differencing. As the name suggests, this method of differencing is applied if strong seasonal patterns are present in the data. Seasonal differences are the difference between an observation and the previous observation from the same season and can be expressed as

$$y'_t = y_t - y_{t-m}$$

where $m =$ the number of seasons.

Before we start differencing our data, we should first test if our data is already stationary, which would then allow us to go straight into further analysis. The test we will use for stationarity is the Kwiatkowski-Phillip-Schmidt_shin (KPSS) test. The KPSS test works by breaking up a time series down into the following decomposition

$$Y_t = r_t + \beta_t + \epsilon_t$$

where r_t is a random walk, β_t is the trend, and ϵ_t is the stationary error. The KPSS test also involves hypothesis testing with

$$H_0 : Y_t \text{ is trend (or level) stationary}$$

$$H_1 : Y_t \text{ is a unit root process}$$

where H_0 and H_1 is the null and alternative hypothesis respectively. Setting our significance level at e can now test each time series to determine whether they will require differencing.

The value of the t-statistic for each test is as follows:

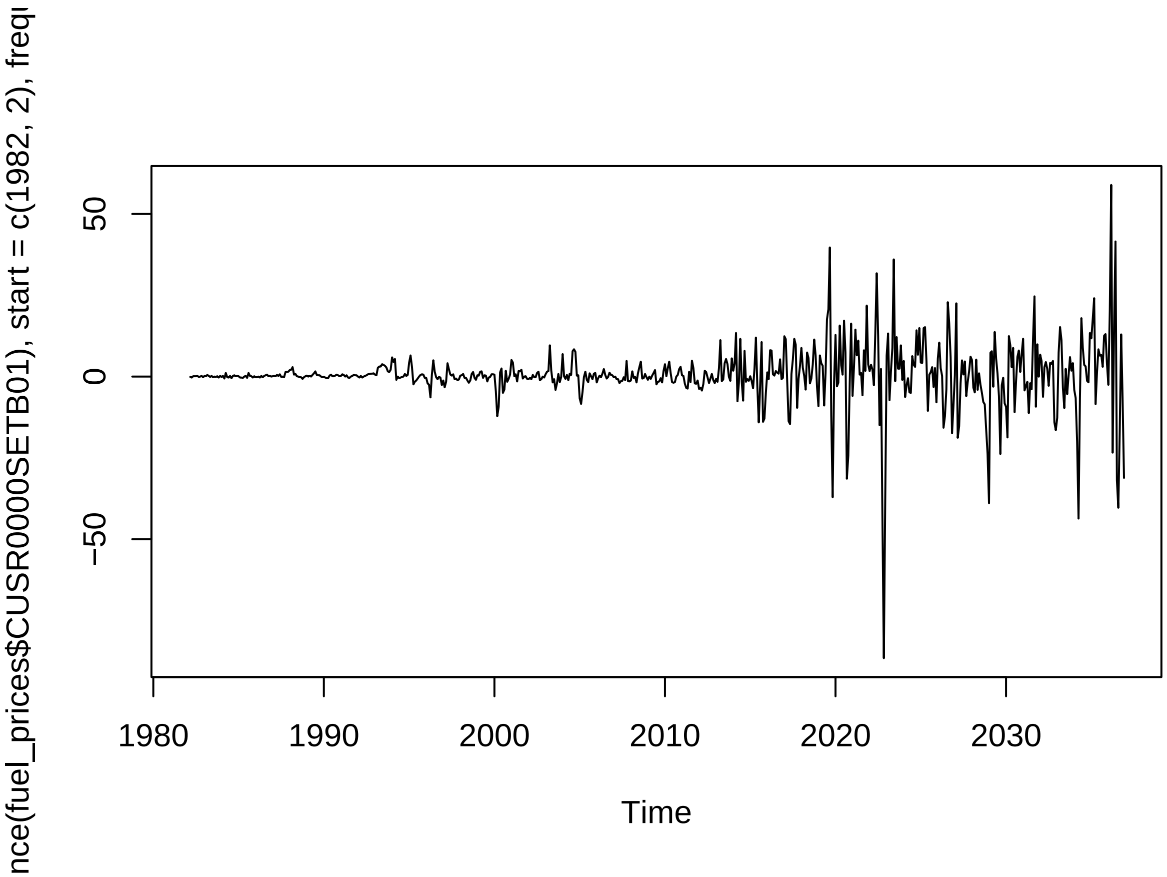
Test Statistic and Significant Values of Time Series

Time Series	Test-Statistic	P-Values
Total Vehicles	1.4887	0.01
Domestic Production	4.7110	0.01
Bank Prime Rate	2.1478	0.01
New Car CPI	11.6058	0.01
Gasoline CPI	7.8250	0.01

Since all of the time series have p-values that are less than 0.05, we can reject the null hypothesis and conclude that it is likely that all our time series are this suggests that differencing will be required in order to make our data stationary. We will demonstrate this process on our time series for gasoline CPI and for domestic auto production in order to show the process on seasonally adjusted and non-seasonally adjusted data respectively, but the process will be conducted on all time series in this case study as needed.

In the case of gasoline CPI the following procedure will first apply a differencing function and then plot the function to see if the differencing allowed the data to become stationary.

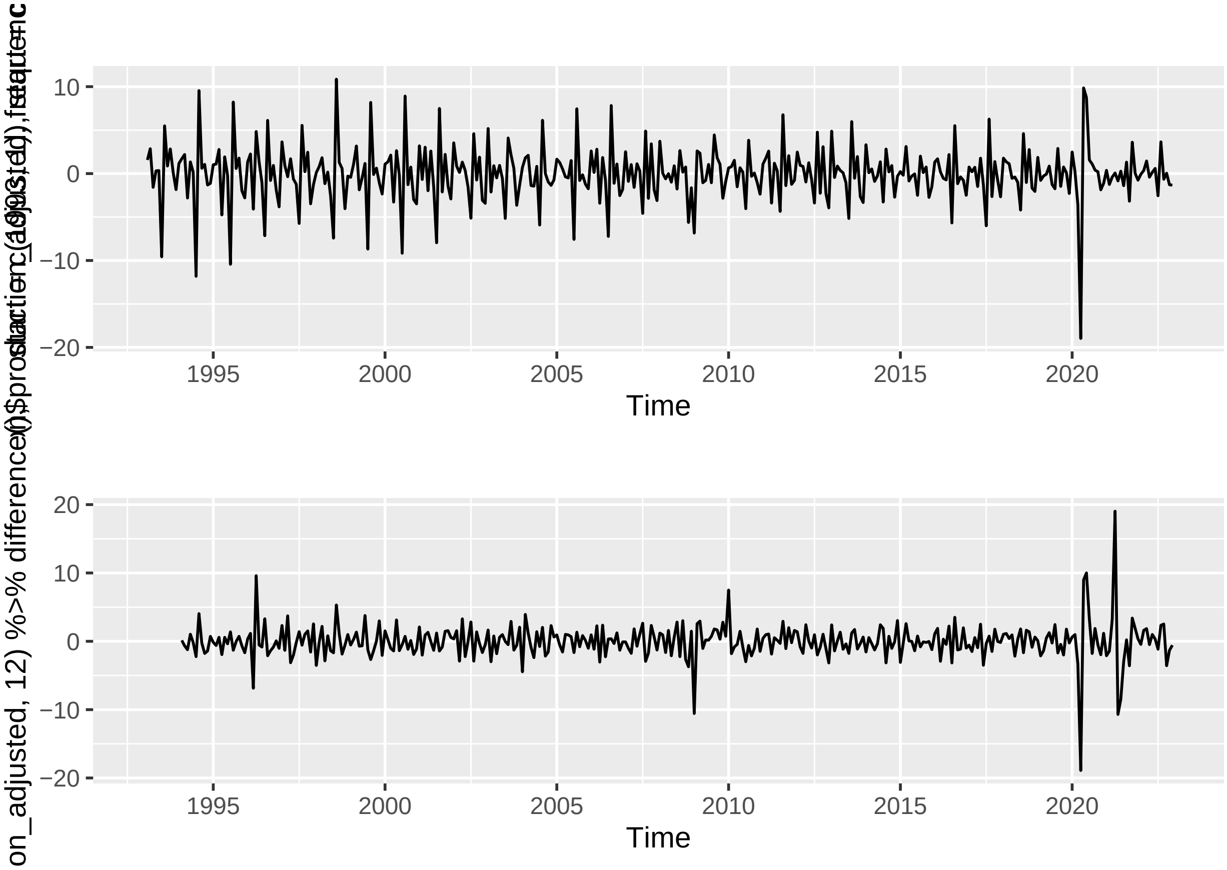
kpss_stat	kpss_pvalue
0.03260776	0.1



From the above plot, we see that there does not appear to be any defined pattern that this data follows, so we see that adding one difference allows the time series to become stationary.

The process for seasonal data is slightly different and the following will show how to create stationarity for data with seasonal patterns.

kpss_stat	kpss_pvalue
0.01439842	0.1



From the output `kpss_pvalue`, we see that we get a p-value greater than 0.1, which means that we can conclude that the differenced data is stationary, which will then allow us to complete further analysis of the data. In this case, we also seasonal differencing is also required because we see in the top plot that there is a strong seasonal pattern still present and seasonal differncing removes the pattern as seen in the lower plot.

After we have successfully differenced our data, we can then begin to consider what types of models to fit to our stationary time series. When considering ARIMA models, there are two model types that we can use: autoregressive models and moving average models.

We will first discuss autoregressive models. In autoregression, our goal is to forecast the variable of interest using linear combinations of previous values of the variable. This means that in an autoregression, the variable is regressed on itself. An model that utilizes autoregression of order p can be denoted by

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

, where ϵ_t denotes white noise. In such a case, we would say that our model is an AR(p) model. For these types of models, we face certain constraints, mainly:

- For an AR(1) model: $-1 < \phi_1 < 1$
- For an AR(2) model: $-1 < \phi_2 < 1, \phi_1 + \phi_2 < 1, \phi_2 - \phi_1 < 1$.

While autoregression models use past values of itself, moving average models use past forecast errors in a regression_like model (Rob J Hyndman 2021). Moving average models take on the form

$$y_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

, where ϵ_t is white noise. In such a case, we would say that our model is an MA(q) model. Each value of y_t can be thought of as a weighted moving average of past forecast errors. MA models also have stationarity constraints which are as follows:

- For an MA(1) model: $-1 < \theta_1 < 1$
- For an MA(2) model: $-1 < \theta_2 < 1, \theta_1 + \theta_2 > 1, \theta_1 - \theta_2 < 1$.

Non-Seasonal ARIMA Models

For our non-seasonal time series we can appropriately use Non-seasonal ARIMA models which follow the form

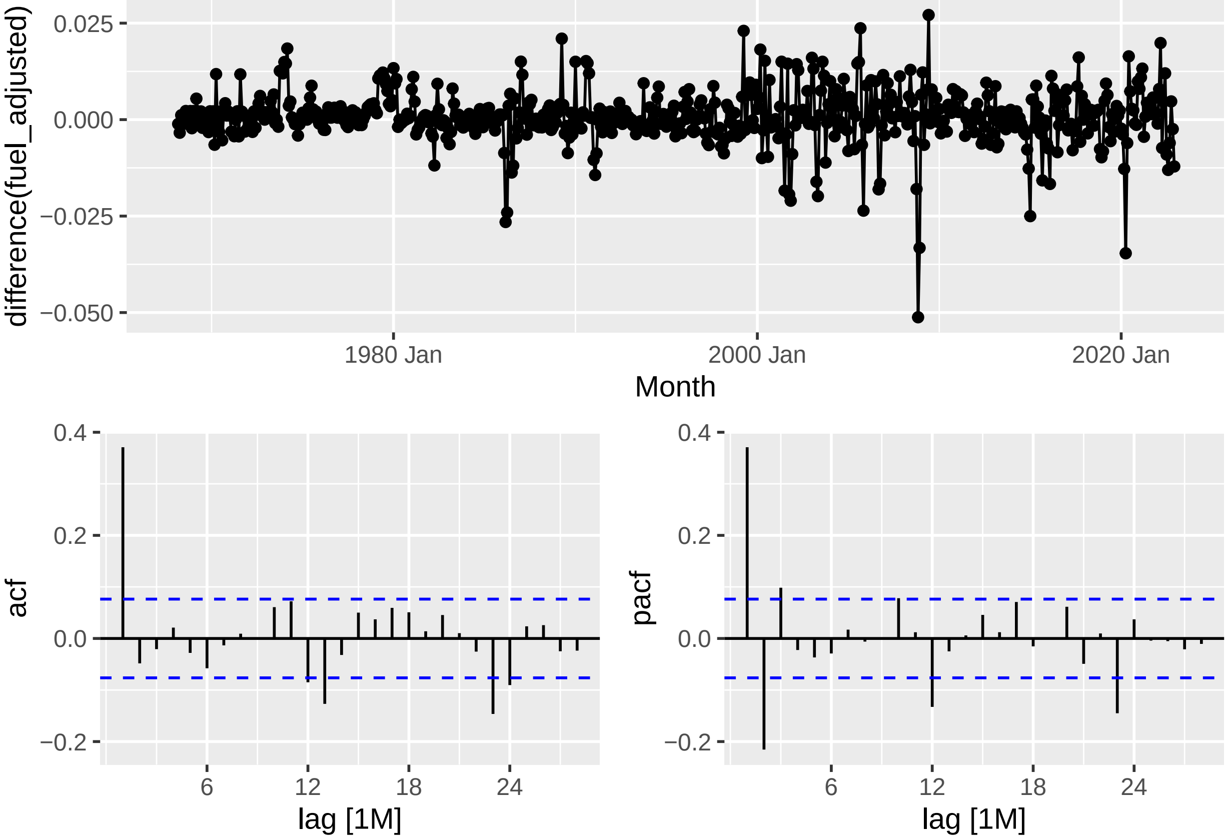
$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_1 \epsilon_{t-q} + \epsilon_t$$

To find the best model, we will have to create ACF and PACF plots so that we may visually decide which ARIMA model may best model our stationary time series, we will also take advantage of functions which will pick the best ARIMA model and then compare the given answers. The ACF will allow us to decide which MA nmodel to use and the PACF will allow us to decide which AR model to use.

In the case of fuel prices:

```
## Warning: Removed 1 row containing missing values (`geom_line()`).
```

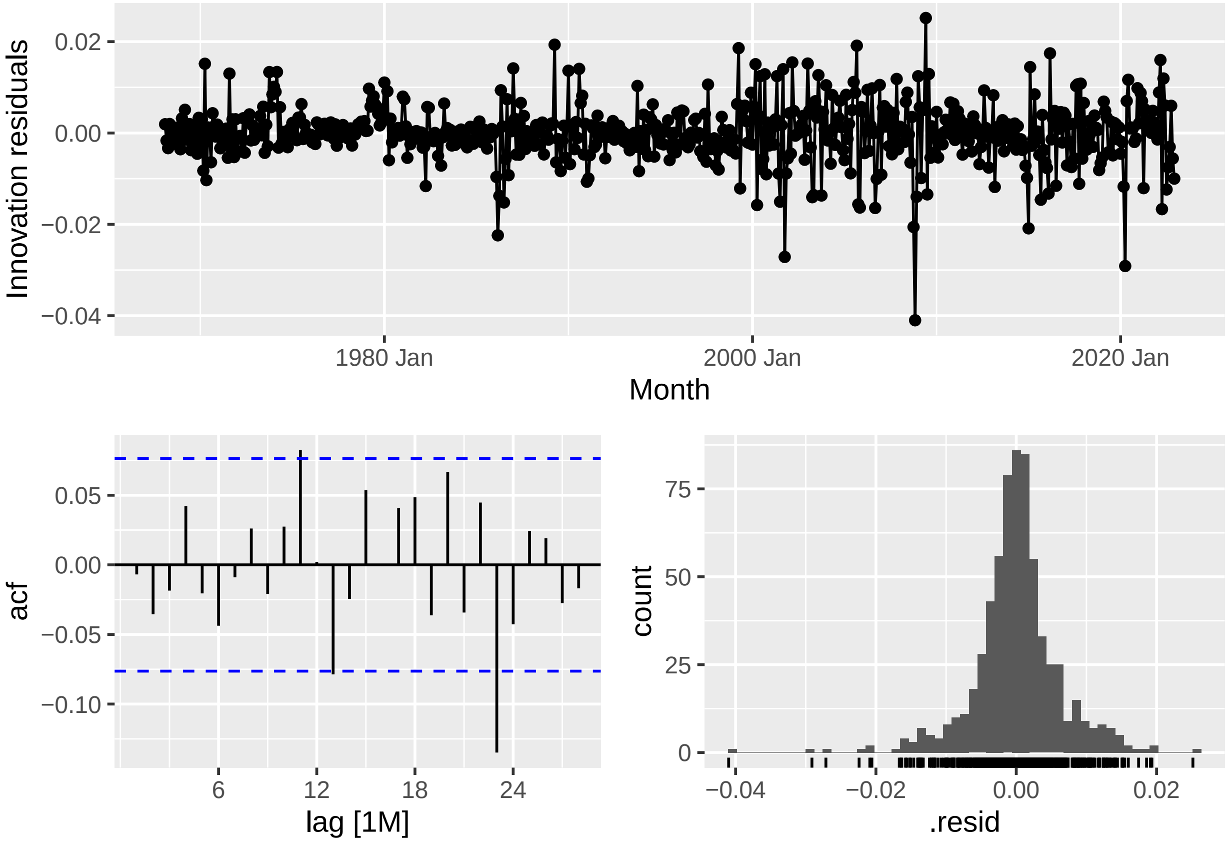
```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



```
## Series: fuel_adjusted
## Model: ARIMA(0,1,1)(0,0,1)[12] w/ drift
##
## Coefficients:
##          ma1     sma1  constant
##       0.4828  -0.1016     7e-04
## s.e.  0.0338   0.0409     3e-04
##
## sigma^2 estimated as 3.862e-05:  log likelihood=2410.9
## AIC=-4813.8   AICc=-4813.74   BIC=-4795.84
```

```
## # A mable: 1 x 4
##                              arima310                           arima011
##                               <model>                            <model>
## 1 <ARIMA(2,1,0)(0,0,1)[12] w/ drift> <ARIMA(0,1,1)(0,0,1)[12] w/ drift>
## # ... with 2 more variables: stepwise <model>, search <model>
```

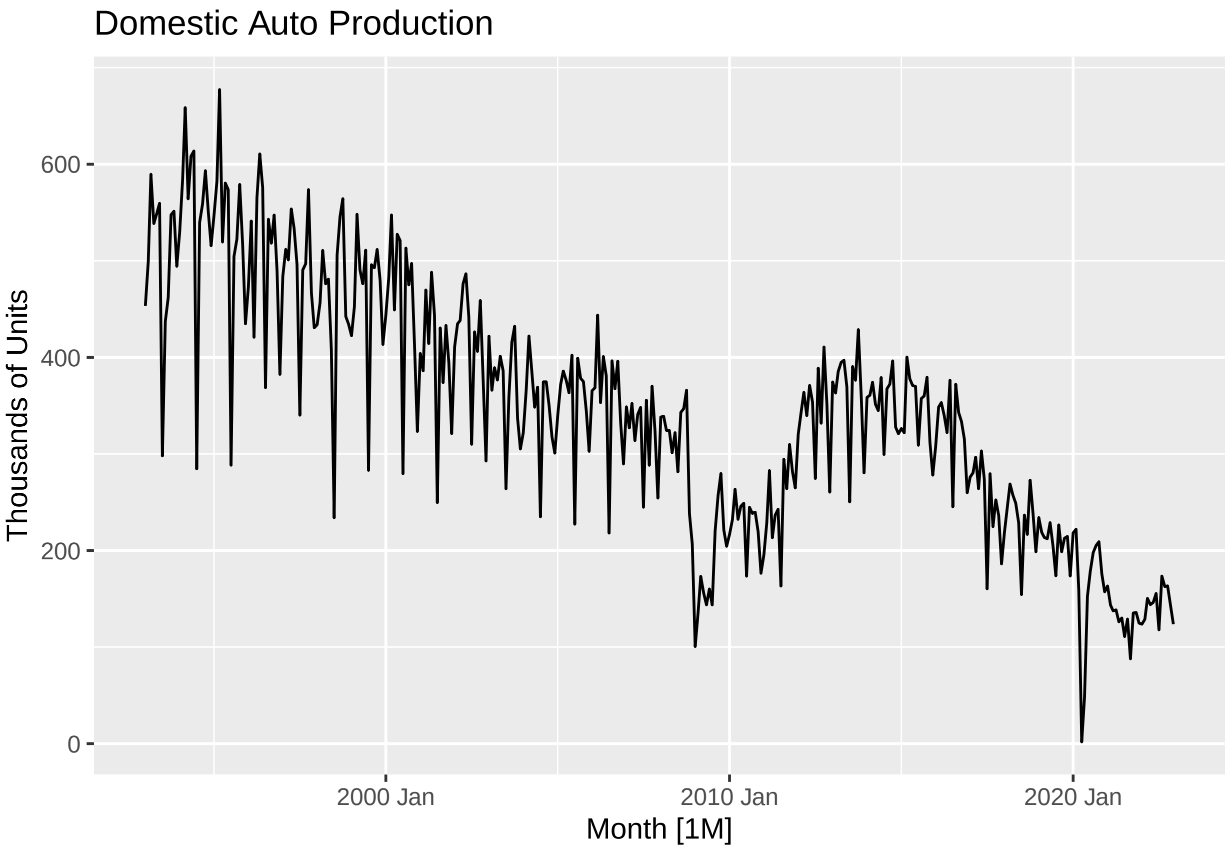
```
## # A tibble: 4 x 6
##   .model      sigma2 log_lik    AIC   AICc    BIC
##   <chr>        <dbl>   <dbl>  <dbl>  <dbl>  <dbl>
## 1 arima011 0.0000386   2411. -4814. -4814. -4796.
## 2 stepwise 0.0000386   2411. -4814. -4814. -4796.
## 3 search   0.0000386   2411. -4814. -4814. -4796.
## 4 arima310 0.0000390   2408. -4807. -4807. -4785.
```



```
## # A tibble: 1 x 3
##   .model lb_stat lb_pvalue
##   <chr>    <dbl>     <dbl>
## 1 search    5.14     0.643
```

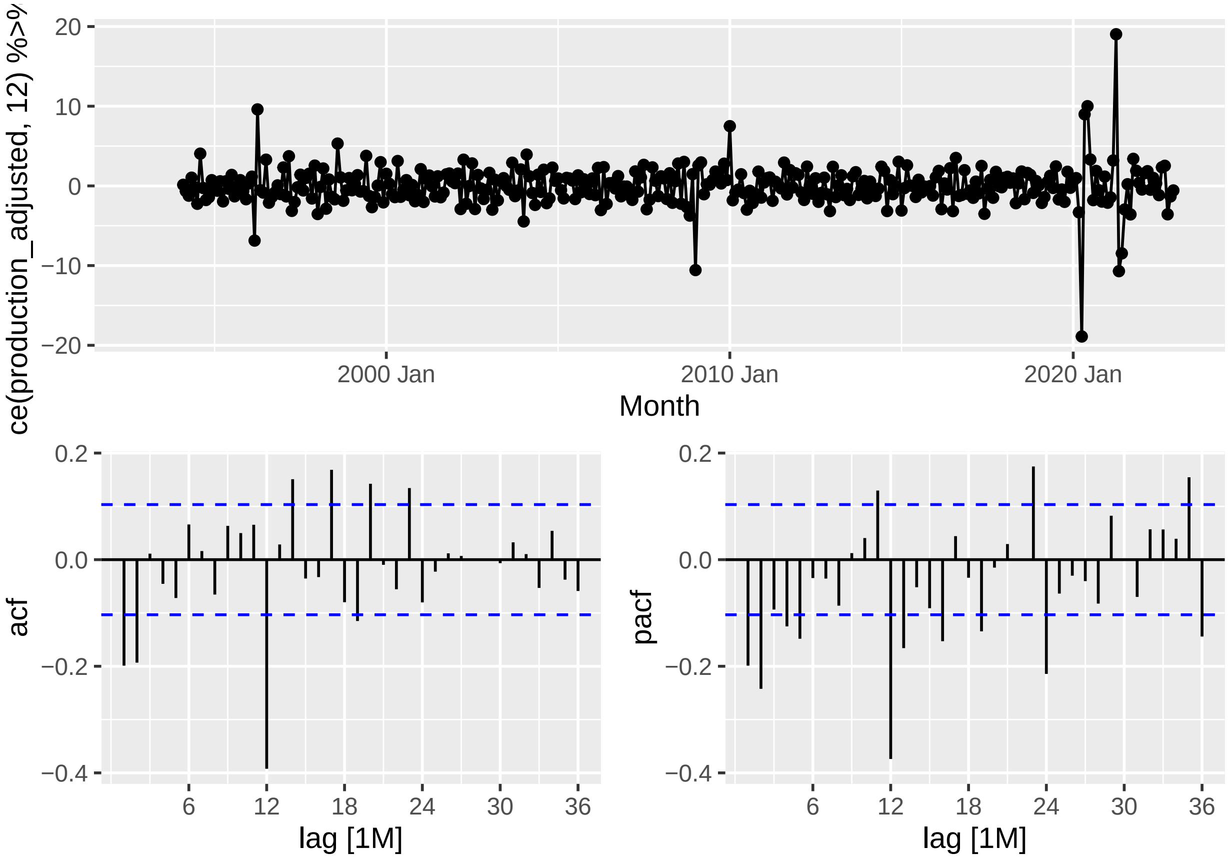
Seasonal ARIMA Models

Combining autoregression and moving average models along with differencing allows us to obtain ARIMA models, which stands for Auto Regressive Integrated Moving Averge. ARIMA models can be written as

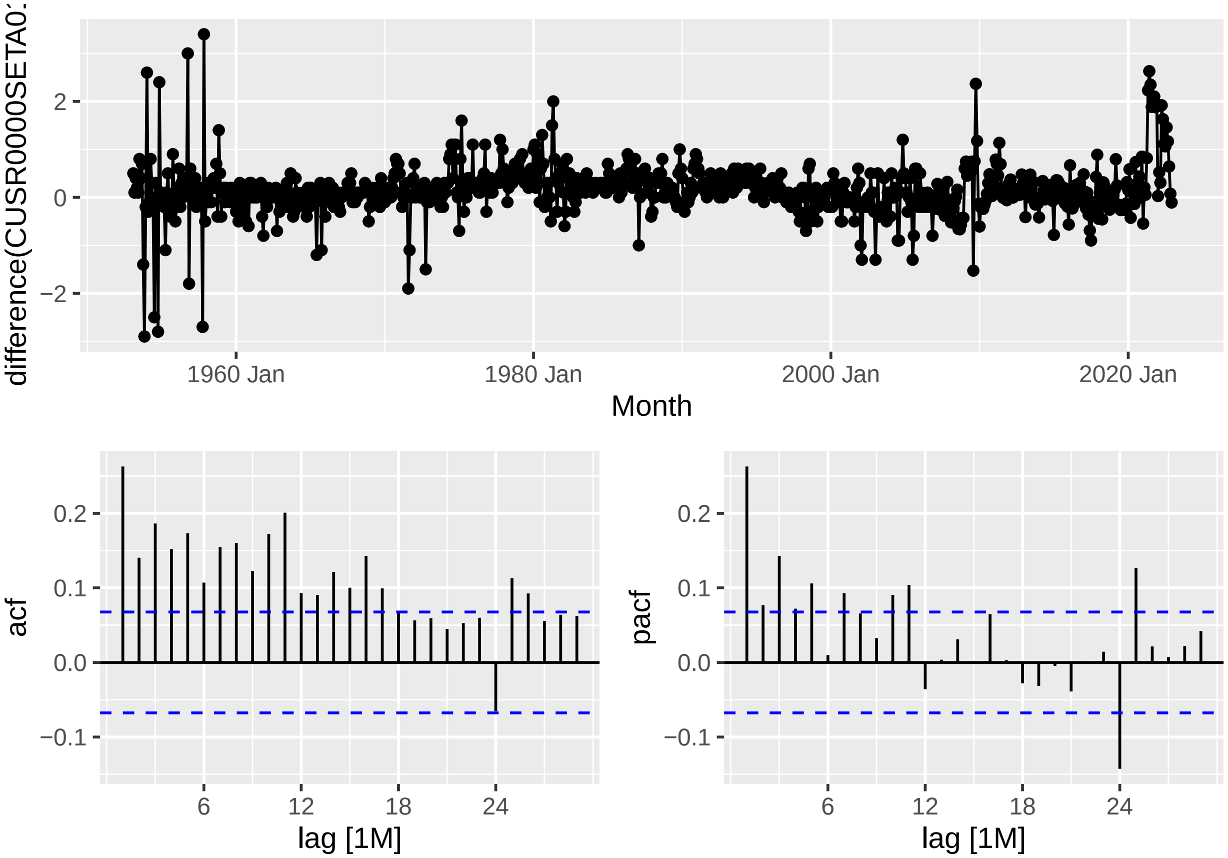


```
## Warning: Removed 13 rows containing missing values (`geom_line()`).
```

```
## Warning: Removed 13 rows containing missing values (`geom_point()`).
```

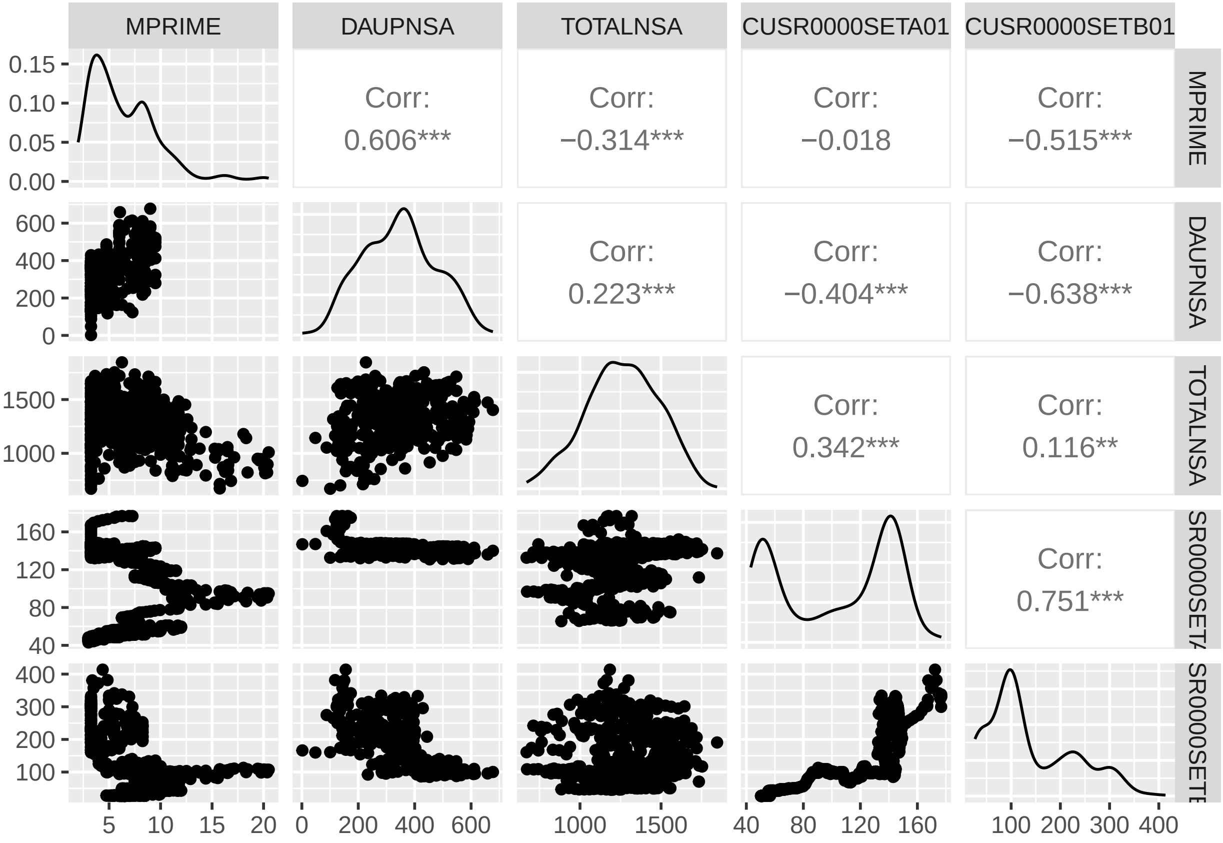


```
## Series: production_adjusted
## Model: ARIMA(1,1,1)(2,0,0)[12]
##
## Coefficients:
##          ar1      ma1    sar1    sar2
##       0.4906  -0.9027  0.4617  0.2643
## s.e.  0.0705   0.0376  0.0520  0.0522
##
## sigma^2 estimated as 4.36:  log likelihood=-775.81
## AIC=1561.62   AICc=1561.79   BIC=1581.04

## Warning: Removed 1 row containing missing values (`geom_line()`).

## Warning: Removed 1 rows containing missing values (`geom_point()`).
```



```
## Series: CUSR0000SETA01
## Model: ARIMA(1,1,1)(2,0,2)[12] w/ drift
##
## Coefficients:
##          ar1      ma1     sar1    sar2    sma1     sma2  constant
##       0.9777  -0.8744  -0.2270  0.0889  0.1772  -0.3306    0.0044
## s.e.  0.0132   0.0283   0.1664  0.1558  0.1596   0.1469    0.0018
##
## sigma^2 estimated as 0.2453:  log likelihood=-598.44
## AIC=1212.88   AICc=1213.05   BIC=1250.74
```

After we have appropriately differenced our models so that they are stationary, we can then begin to consider what type of models to fit to our data.



References

Glenn, Stephanie. n.d. "KPSS Test: Definition and Interpretation." https://www.statisticshowto.com/kpss-test/.

"Inflation and the Auto Industry: When Will Car Prices Drop." 2022. https://www.jpmorgan.com/insights/research/when-will-car-prices-drop.

"Long-Term Trends in Car and Light Truck Sales." 2021. https://fredblog.stlouisfed.org/2022/10/whats-been-driving-the-rise-in-auto-prices-since-covid/?utm_source=series_page&utm_medium=related_content&utm_term=related_resources&utm_campaign=fredblog.

"No End in Sight: New Vehicle Transaction Prices End 2022 at Record Highs, According to New Data from Kelley Blue Book." 2023. https://www.coxautoinc.com/market-insights/kbb-atp-december-2022/.

"PSTAT 174/274: Time Series Part IV." 2023. University of California, Santa Barbara.

Rob J Hyndman, George Athanasopoulos. 2021. "Forecasting: Principles and Practice." Monash University, Australia; OTexts.

"What's Been Drivin the Rise in Auto Prices Since COVID." 2022. https://fredblog.stlouisfed.org/2021/03/long-term-trends-in-car-and-light-truck-sales/?utm_source=series_page&utm_medium=related_content&utm_term=related_resources&utm_campaign=fredblog.