The George Washington University

Total United States Vehicle Sales Forecast Analysis
Comparing Multiple Forecast Methods

Fernando A. Zambrano

Dr. Reza Jafari

22 April 2020

Abstract:

Explore the change in total monthly United States (US) vehicle sales from January 1976 to January 2020. Compare multiple forecasting methods to derive the most accurate model to forecast future vehicle sales. The methods that will be compared are Holt-Winters Exponential Smoothing (HW), Multiple-Linear Regression (OLS), and Autoregressive Moving Average (ARMA). These methods will be compared to a naïve method as a baseline to further asses the validity of the models. These models and calculations are implemented in Python to render fast and reproducible results.

Introduction:

One of the most important economic sectors of the United States (US) is the automobile industry. This industry represents almost 10 million jobs across 44 states that account for 6% of total US gross domestic product (GDP). With the auto industry playing a significant role in the economy at large keeping track of this industry is important. Furthermore, forecasting future sales will give insight into the current health of the sector. However, developing and choosing a model to forecast this industry requires a thorough assessment of the multiple options available that come with their own set of advantages and disadvantages. Some models will only analyze the total vehicle sales such as Holt-Winters or ARMA models, but sales may be driven by other factors that may fail to take into account. In such cases, a multi-variable model may be more appropriate taking into account factors other than sales. Therefore, this study will aim to resolve the issue and find the most suitable model for the automobile industry.

The first task is to inspect the data and make sure it is suitable for the forecast methods that will be implemented by inspecting variables and underlying statistics of the data. The second task will derive a baseline model that uses a naïve approach. Third, derive models for the three different models that will be compared. Next, compare the significant metrics and forecast accuracy of each model to confirm the best model. Finally, the results of the entire study and models will be inspected and determine the limitations and faults of data and models.

Data Description:

The data for this study comes from the Federal Reserve Economic Data (FRED) database. The data includes eight different variables from January 1976 to January 2020. From these features the variable interest or dependent variable is total monthly sales units in million. The other seven features or independent features which will be used to predict vehicles sales are based on features believed to be strongly associated with sales of vehicles. These features are: Consumer Price Index (CPI) for All Urban Consumers, CPI for All Urban Consumers Less Food and Energy, CPI for Used Cars and Trucks for All Urban Consumers, CPI for New Cars and Trucks for all Urban Consumers, Total Number of Employees Not Including Farming in thousands, US Unemployment Rate as a percent and Capacity Utilization for Automobile and Motor Vehicles as a percent of capacity . Overall, there are 529 records for a total of 4,232 observations without any missing values (NaN). Splitting the data into 80% training and 20% testing sets results with 423 records in training and 106 in testing. The test set will be used to validate and compare the accuracy of the forecasting models.

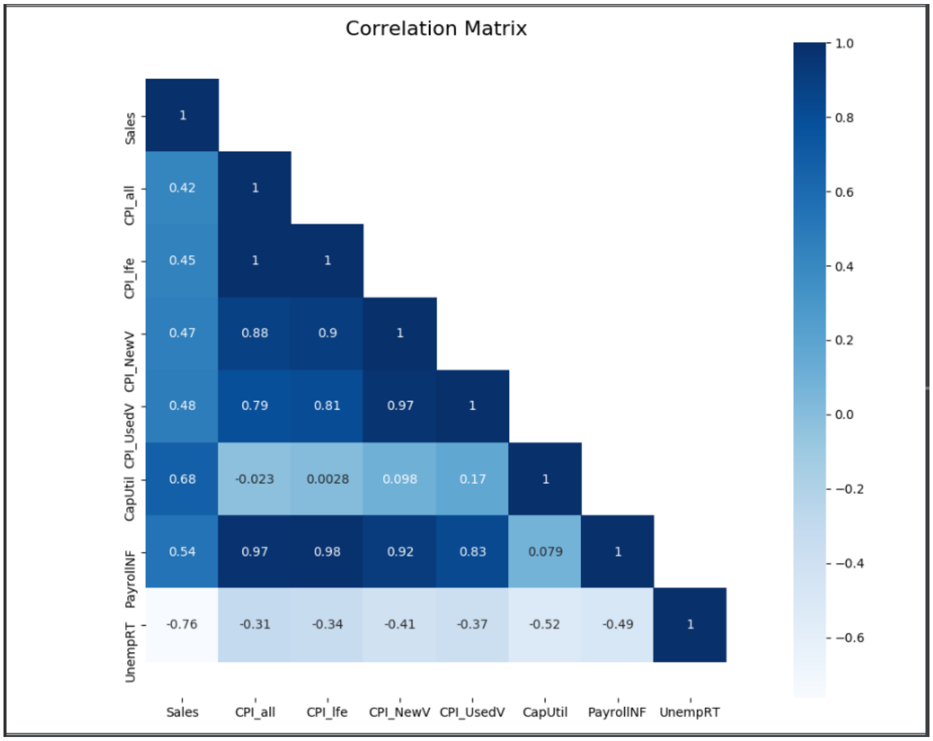
Correlation Analysis:

Correlation, r, measures the strength of a linear relationship between two features. This value ranges from -1 to +1. A correlation value of -1 demonstrates perfect negative linear relationship, that as one feature increases in value the other decreases. A correlation of +1 is a perfect positive relationship, both features decrease and increase together. Correlations between -0.5 and +0.5 are considered weak relationships, with 0 demonstrating the absence of any linear relationship. The equation for correlation is below.

$$r = \frac{\sum(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum(x_t - \bar{x})^2}\sqrt{\sum(y_t - \bar{y})^2}}.$$

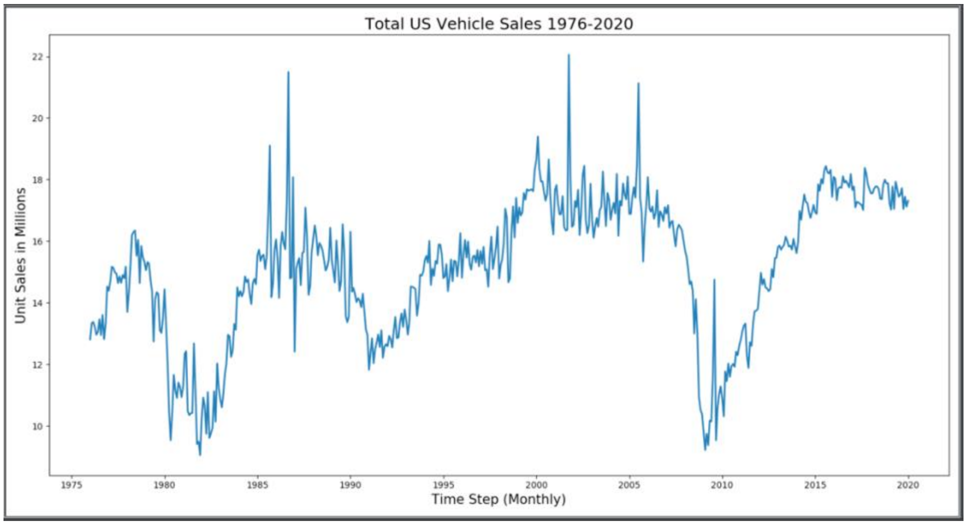
The formula can be separated into two parts. The first part is, deals with the calculation of the numerator. The numerator is the covariance between two variables. Covariance determines how linearly related two features are associated. This is done by taking the product of each feature's variance. The issue with why covariance is not used as the principle statistic to demonstrate the strength of a linear relationship is that it maintains the units of its features. This makes it difficult to draw comparisons between two features that have large differences in their spread. To compare two features without the drawback of dimensionality, the features need to be normalized. This is the second part, deals with the denominator which is the product of the square root of the squared variance of each feature. This process cancels out the dimensions of each feature and forces the resulting value of r to fall between -1 and +1.

Below is the correlation matrix illustrating the linear association between all 8 variables. This matrix will illustrate which variables are estimated to be good predictors for US vehicle sales and show if any independent variables demonstrate signs of collinearity. Ideally, the correlation matrix would show the first column with high (negative or positive) correlation values and the rest of the matrix with low correlation values. This would indicate that the independent variables have a strong association with the dependent variable and the rest of the independent are not correlated. However, the results are not ideal. Only a few features are strongly correlated with sales: capacity utilization (CapUtil), number of employees on payroll (PayrollNF), and unemployment rate (UnempRT). The CPI features are all equally correlated with sales but exhibit high correlation amongst each other and with PayrollNF. The reason for the multicollinearity is most likely due to the fact the CPI features are an index that is calculated with thousands of features and may include payroll information. Hence the features are redundant and will not necessarily help in prediction. Since PayrollNF is not correlated with non-CPI features, it should be kept and CPI_UsedV should also be kept because it is the most correlated with sales and least correlated with other variables. The rest of the CPI features should be removed. The final validity of the variables kept for analysis will be covered in the feature selection section of this report.

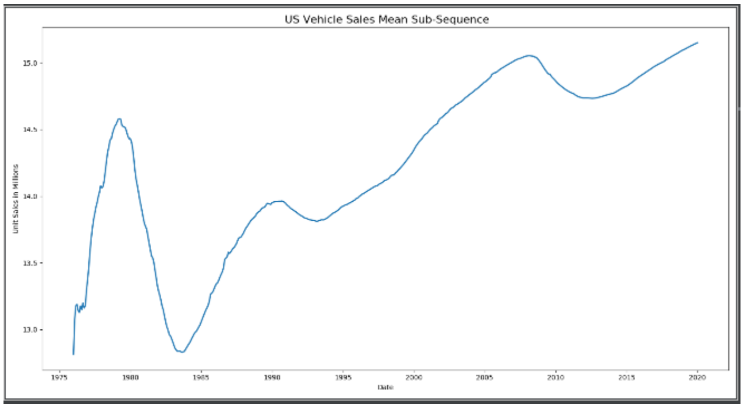


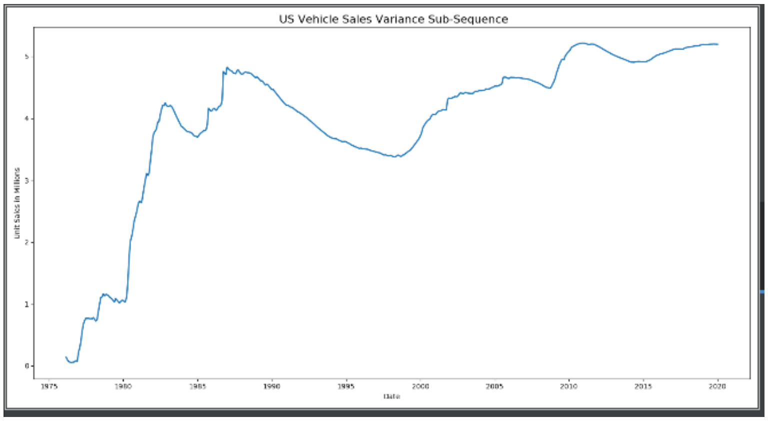
Stationarity:

Inspecting the total vehicle sales across time we can visually generalize the data which will help determine how models will work. Below is the graph of total US vehicle sales.



The graph illustrates that the graph shows a cyclic rather than seasonal behavior because the peaks and troughs of sales occur at inconsistent time intervals. Additionally, there is a slight upward trend in the data. As a result, the data does not look stationary. A dataset is stationary when its mean and variance remain constant over time which produces a predictable pattern. Hence, the predictability that stationarity offers allows for useful and reliable forecasting Sub-sequencing the sales data by inspecting how its mean and variance can further illustrate the characteristics of the data's stationarity. Below are the graphs of mean and variance changing through time.





The sub-sequence graphs visually confirm that vehicle sales data is non-stationary since the mean and variance have an upward trend. If they were, the graphs would show curve flattening out trough time. However, it is not possible to empirically confirm stationarity from visual interpretation.

A popular option to determine if a dataset is stationary is to perform a statistical hypothesis test which determines the likelihood of the dataset being non-stationary. The Augmented Dickey-Fuller (ADF) test, or "unit root test" makes a strong assumption that the dataset in hand is non-stationary. This test sets up two hypotheses:

Null Hypothesis (H0): if failed to be rejected, then the time-series has a unit root and the data is non-stationary.

The Alternative Hypothesis (H1): if the null hypothesis is rejected, then there is no unit root and the data is stationary.

To reject the null hypothesis, there must be strong evidence to suggest that the likelihood of the data being non-stationary is significantly unlikely. This is done by setting high confidence levels which are determined by the predetermined critical values. The standard practice is to have between a 95% - 99% confidence level. A 0.05 and 0.01 critical values correspond with 95% and

99% confidence level, respectively. The ADF test will result in a p-value which is compared to the set critical value. If the p-value is less than the critical value, then there is significantly strong evidence to suggest that the null hypothesis (non-stationary) be rejected in favor of the alternative (stationary). Conversely, if the p-value is greater than the critical value, then there is not enough significantly strong evidence to support rejecting the null hypothesis.

Below are the results of the ADF test on US vehicle sales. The first results are from the original dataset which shows that it does pass the ADF test for stationarity since its p-value (0.264689) is greater than the .05 threshold. Therefore, vehicle sales are non-stationary.

ADF TEST: US Vehicle Sales

ADF Statistic: -2.050952

p-value: 0.264689

Critical Values:

1%: -3.442891

5%: -2.867072

10%: -2.569716

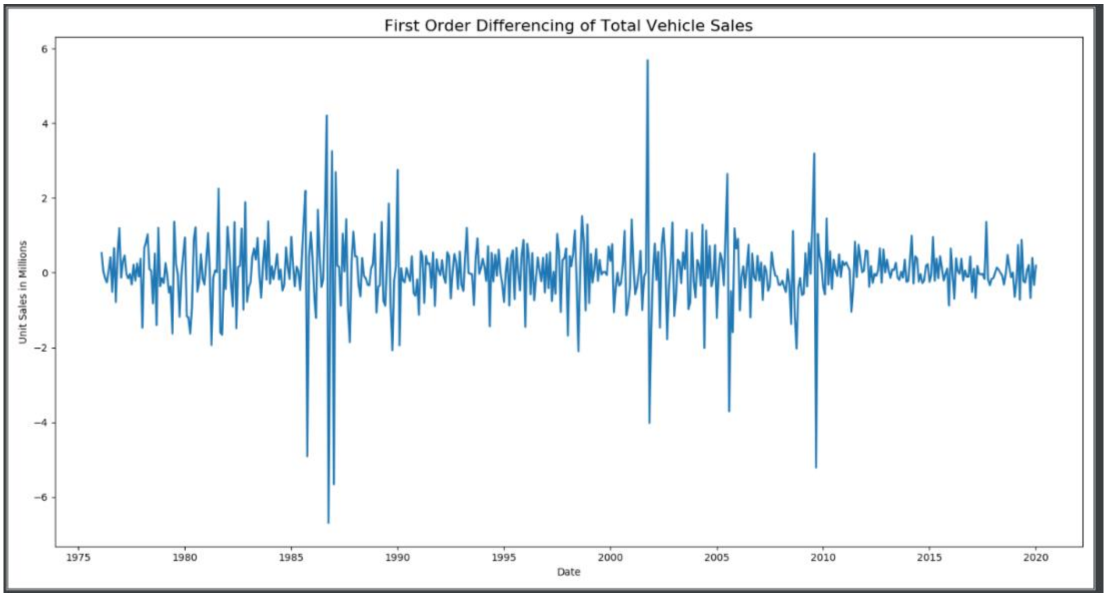
However, there are techniques to transform raw data that is non-stationary into a stationary series. This is accomplished by taking the first-order difference of the data. Differencing is a method that helps stabilize the mean from changing over time. This is done by subtracting the previous observation from the current observation. First-order differencing stabilizes the mean, and second-order stabilizes the variance. Anytime differencing or transformations are applied to data a reverse transformation must also be done. Below are the formulas for first and second-order differencing, respectively. The following graphs show the subsequence of the mean and variance after first order differencing, and the respective ADF test.

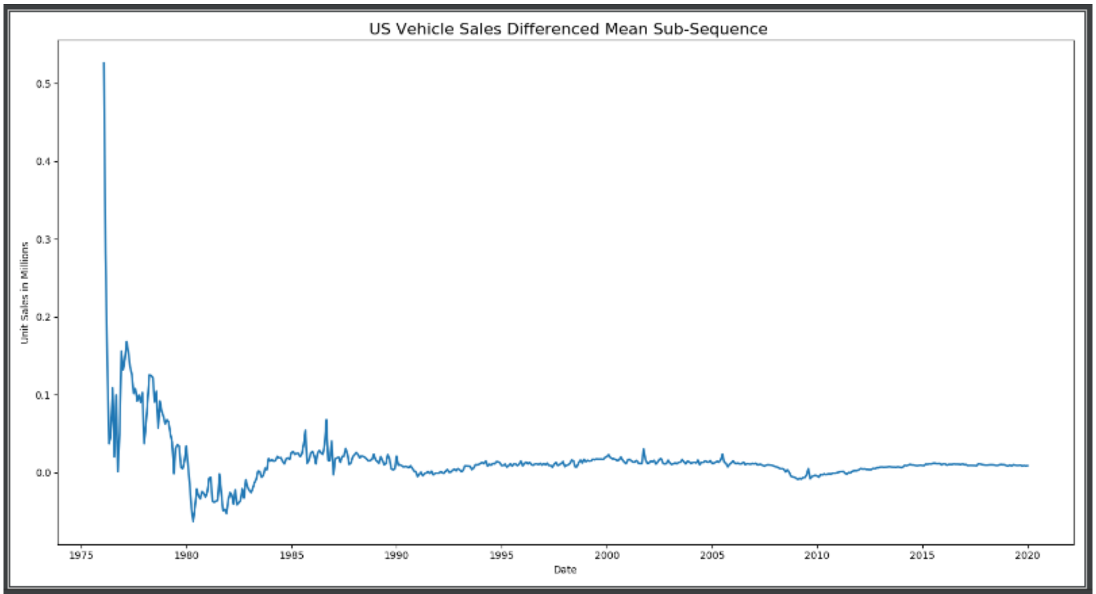
First Order

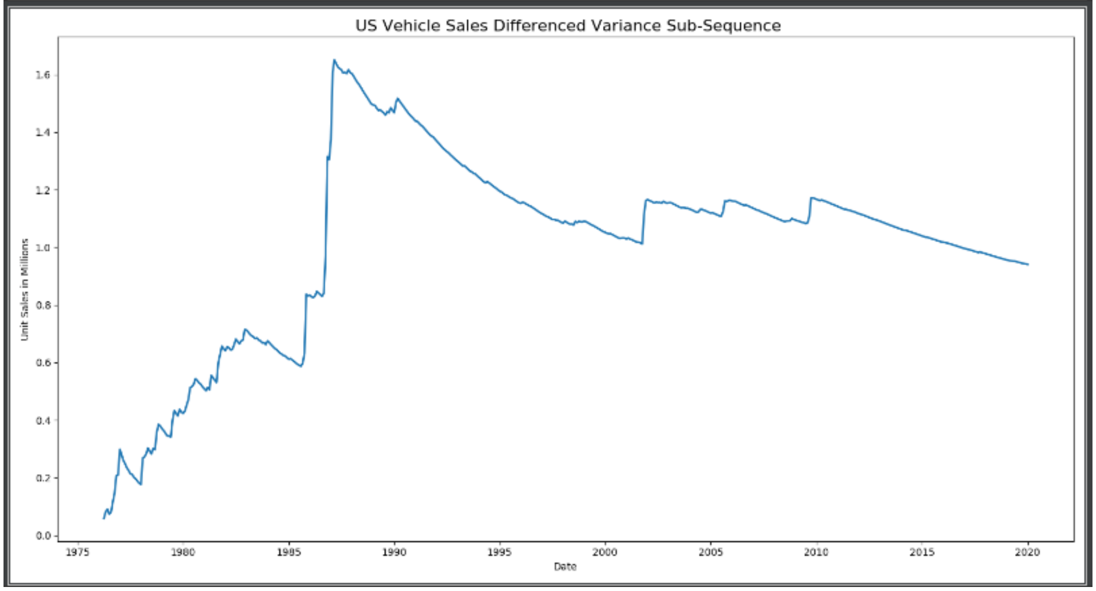
$$\Delta y(t) = y(t) - y(t-1)$$

Second Order

$$\Delta^2 y(t) = y(t) - 2y(t-1) + y(t-2)$$







ADF TEST: US Vehicle Sales (Differenced)

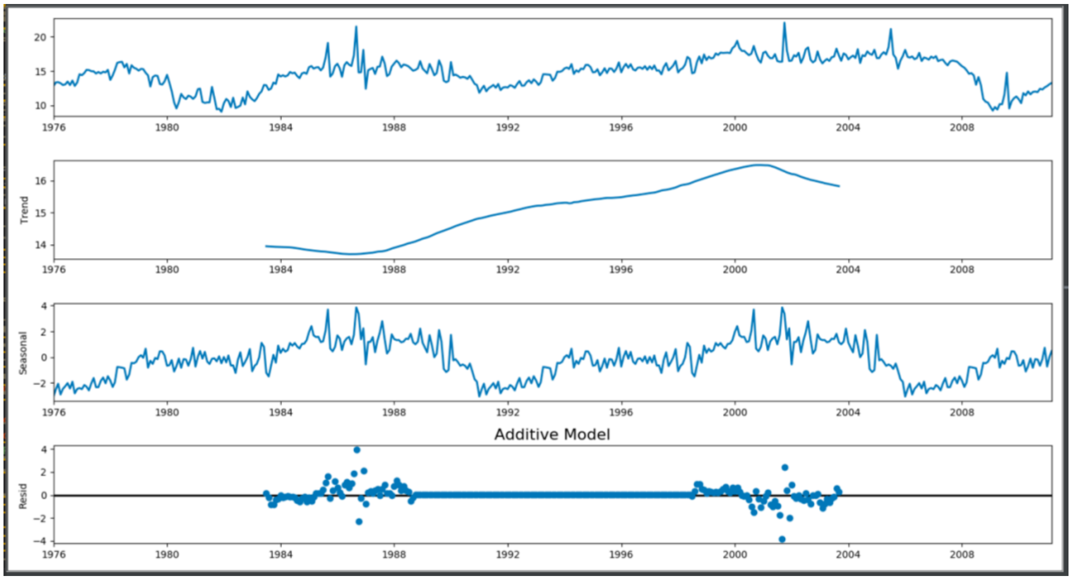
```
ADF Statistic: -16.818233
p-value: 0.000000
Critical Values:
    1%: -3.442891
    5%: -2.867072
    10%: -2.569716
```

The results of first-order differencing indicate that the mean stabilized due to the curve flattening out. However, the variance still showed a clear upwards trend that closely resembles the raw data, but with a smaller scale. If visual interpretation was the only criterion for determining stationarity, the data would not pass this test. Yet, the ADF test considers the data stationary given the p-value (0.00) is less than the critical value threshold of 0.05.

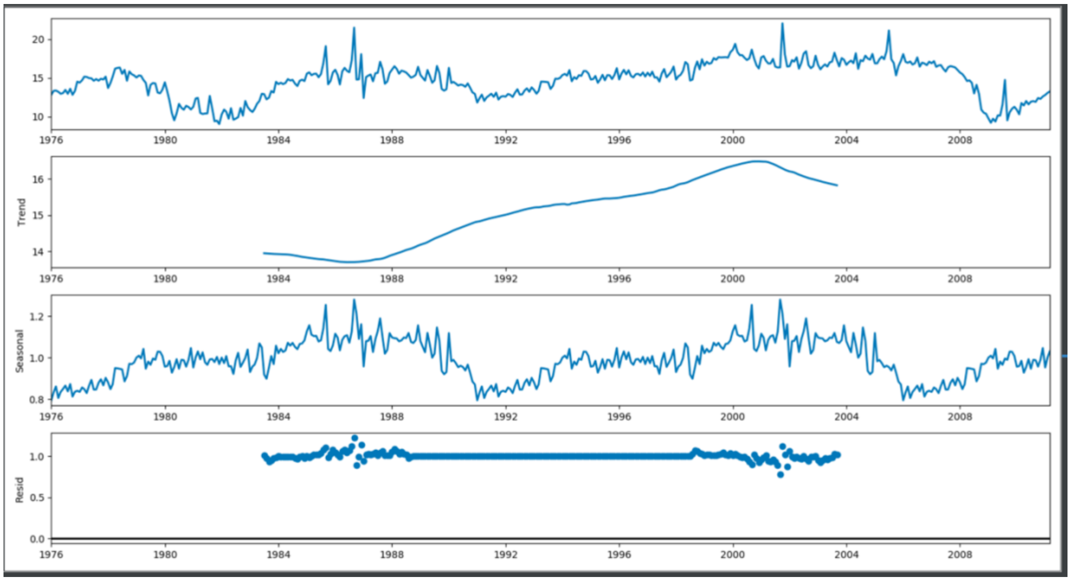
Time Series Decomposition:

Decomposing the US sales vehicle data into three different components, trend, seasonality, and residual will help understand the behavior of the time series. There are two main decomposition models, additive or multiplicative. In the additive model, the three components are added together, whereas in the multiplicative they are multiplied. Below are the graphs of the additive and multiplicative models respectively. The model with the lowest spread of residuals is the best decomposition.

Additive Decomposition



Multiplicative Decomposition

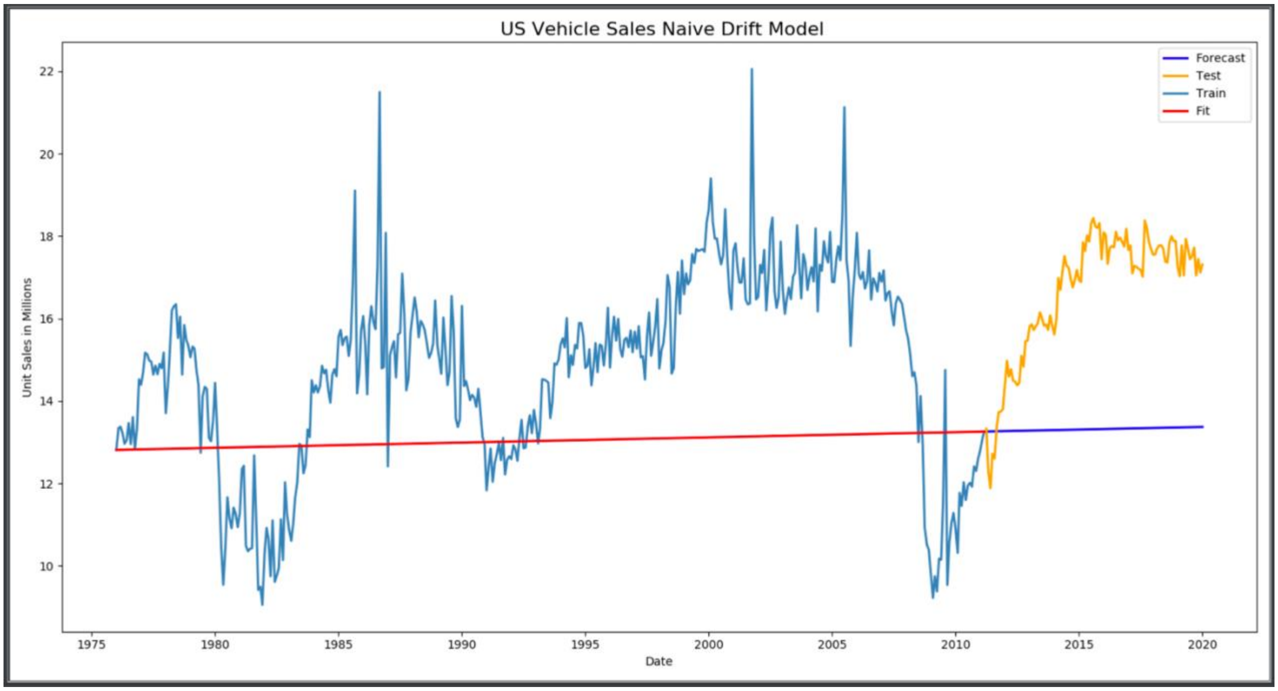


Based on the decomposition graphs, the additive model is the best choice. The residuals are more aligned with 0 whereas the multiplicative residuals are aligned with 1.

Naïve Model:

The primary objective of forecast models is to make predictions that are better than just guessing, averages, or using the last known observations. These methods are referred to as naïve methods. They are simple models that can be compared to more sophisticated models. If the sophisticated model cannot predict better than the naïve model, it is not a good model. The naïve model that will be used for this study is the drift method because it can handle trended data such as US vehicle sales. This method is similar to a line of best fit, but instead of trying to minimize the error for each point, it takes the slope of the line between the first and last observation in the training set. The drift line then forecasts using that slope for h number of observations in the test set. The formula for the drift method is below.

$$\hat{y}_{T+h|T} = y_T + \frac{h}{T-1}\sum_{t=2}^{T}(y_t - y_{t-1}) = y_T + h(\frac{y_T - y_1}{T-1})$$



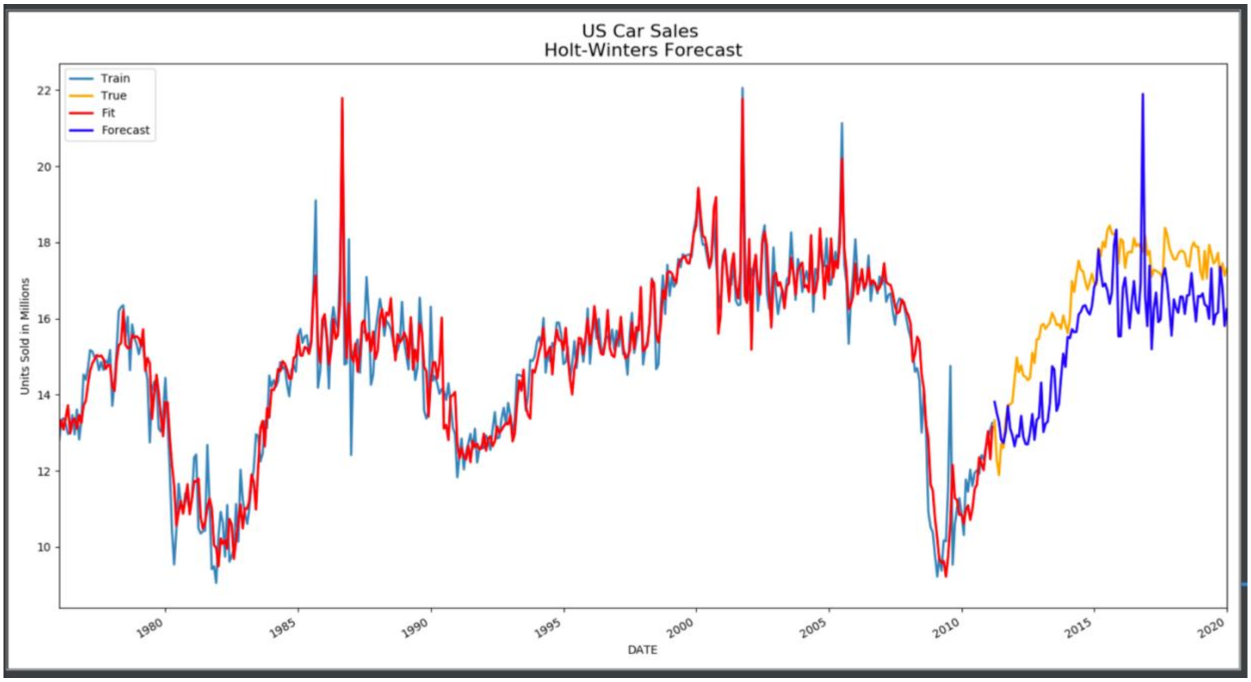
Naïve Drift	Residual	Forecast
SSE	3398.41	1423.20
MSE	8.03	13.42
RMSE	2.834	3.664
VAR	5.034	2.264

Based on the graph, the results of the drift prediction are not very impressive. The sum of square errors (SSE) of the residuals (training data – fitted data) is 3398.41 and a mean square error (MSE) of 8.03. The forecast errors (test data – forecast) are an SSE of 1423.20 and MSE of 13.42. These metrics will serve as the baseline for the more sophisticated models to surpass.

Holt-Winters Method:

The first sophisticated model that will be implemented is the Holt-Winters method (HW). Unlike the naïve drift method, HW can forecast data that has both a seasonal and a trend. This makes the method even more useful as it can apply to non-stationary datasets. HW method through a combination of four different equations, the forecast equation, and three smoothing equations. The smoothing equations are for three components of the data, level(lt), trend(bt), and seasonality(st). Through optimization, in the Stats. Model package in Python the parameters for each equation can be optimized. The equation for HW is below:

$$\begin{aligned}
\hat{y}_{t+h|t} &= \ell_t + hb_t + s_{t+h-m(k+1)} \\
\ell_t &= \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\
b_t &= \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1} \\
s_t &= \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}
\end{aligned}$$



Holt-Winters	Residual	Forecast
SEE	191.95	216.18
MSE	0.45	2.04
ME	-0.003	1.091
RMSE	0.671	1.428
VAR	0.455	0.857

The results of the HW method are a great improvement compared to the naive baseline. As the graph above illustrates, the fitted data represents the training data well. Regarding the forecast, the model was able to predict the upwards trend and also the seasonal component the trend declines. One of the trends that were predicted was a large and short spike in sales in 2016 that has usually happened in the past but did not occur in the test set

Linear Regression (OLS):

One of the most popular methods of forecasting is Ordinary Least Squares (OLS) linear regression. This method can also be referred to as multiple linear regression or multivariate regression if more than one feature is used to predict the dependent variable. The fundamental concept of regression is the idea of a linear model or linear equation that takes the values of certain variables to then predict what the value of the dependent variable should be based upon the inputs. The name of this equation is called the normal equation which is below.

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t$$

The breakdown of this equation is as follows:

Yt is the desired prediction, B0 is the coefficient for not having a predicting variable, so it is just as effective as the mean of the predicted variable. When there are predicting variables, each variable Xi is assigned a coefficient Bi. εt is the amount of deviation between actual points and the fitting line. The smaller these deviations become the lower the expected error will between actual values and predictions. Calculated errors are squared to removed signage and then summed to get the total squared error or SSE. SEE generalizes the total error of the fitted line. The lower the SSE, the more accurate the model. The difficult part about regression is choosing the right coefficients for each variable to minimize SSE. This can be solved by rewriting the linear model as a system of linear equations, where the predicted variable, Y, the predicting variable(s) X, and the coefficients B, are written as matrices and vectors.

Since Y and X are known, the equation can be re-written to solve for B.

$$\boxed{\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}}$$

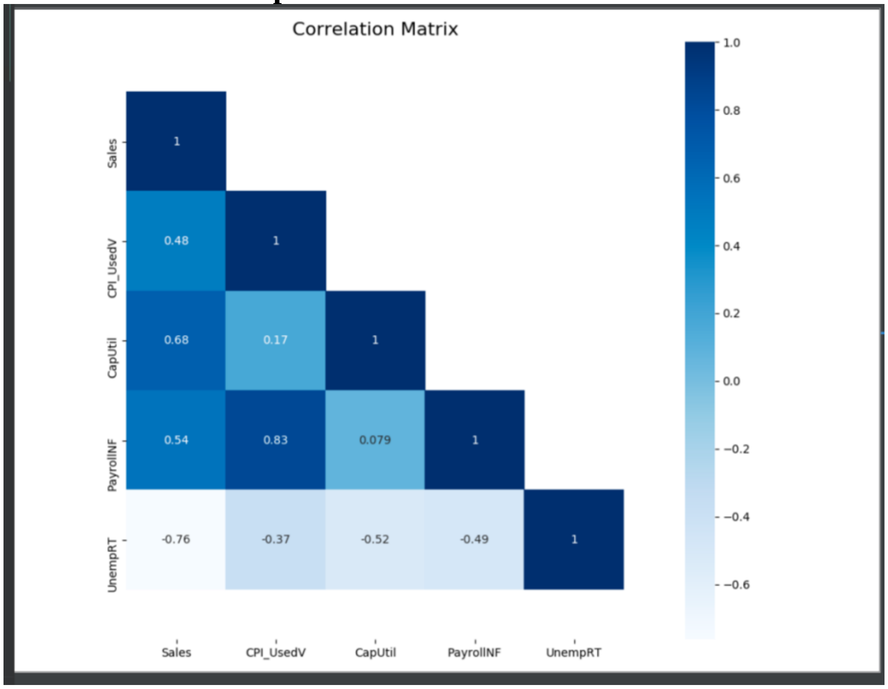
Feature Selection:

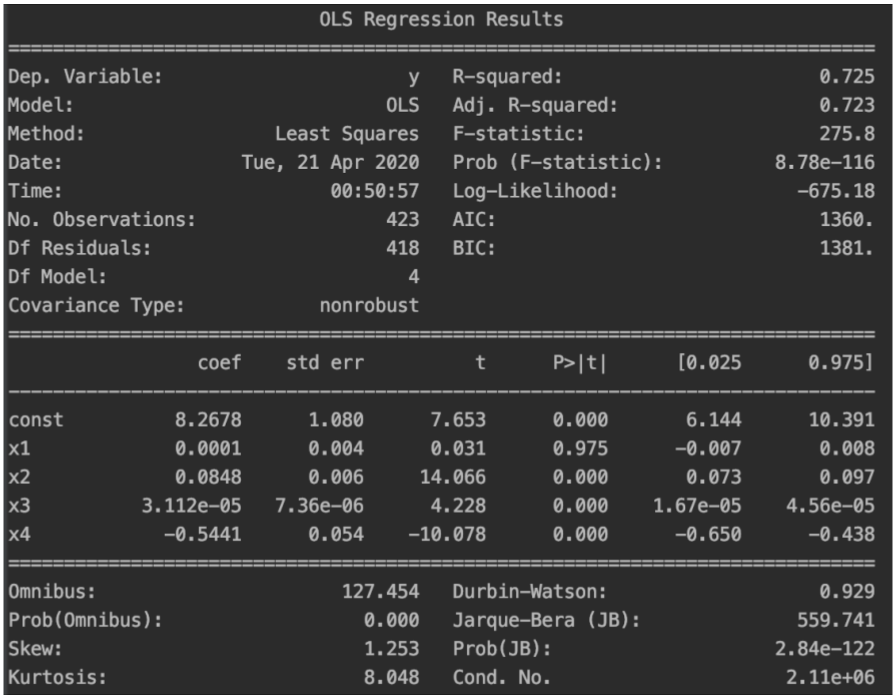
As mentioned before in the Correlation Analysis section, the CPI features and payroll demonstrated multicollinearity because they were all highly correlated with each other. Having multicollinearity predictors is not conducive to accurate and reliable, hence such predictors should be dropped or exempted from the regression analysis. Since CPI of used vehicles demonstrated the highest correlation with sales and least multicollinearity with other predictors it can be kept along with payroll since it was not correlated with the other predictors.

The number of features dropped from 7 to 4. The four remaining features are CPI of used vehicles, capacity utilization, number of employees on payroll, and unemployment rate. Once features have passed multicollinearity, they must also pass a t-test that will determine if that feature is statistically significant in predicting the dependent variable. Like the ADF test, features must pass a 95%

confidence test, in other words, a p-value that is less than 0.05. Using the OLS function in the Stats.Model package in Python will return a model summary of the regression model along with the results of the t-test to confirm which features should be kept or dropped.

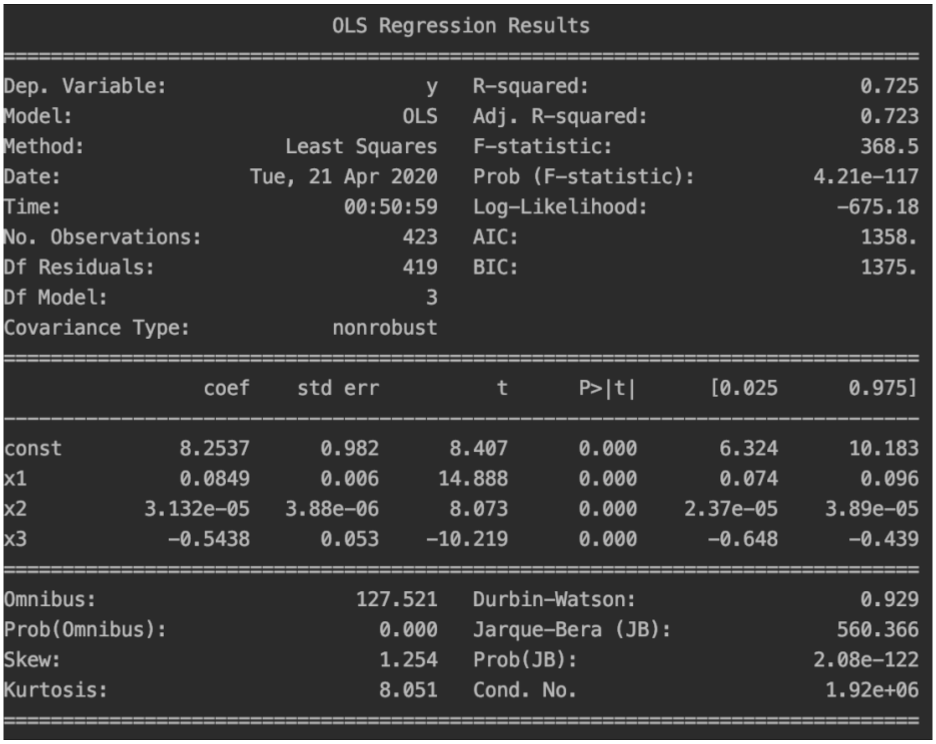
Updated Correlation Matrix





```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.725
Model:                            OLS   Adj. R-squared:                  0.723
Method:                 Least Squares   F-statistic:                     275.8
Date:                Tue, 21 Apr 2020   Prob (F-statistic):          8.78e-116
Time:                        00:50:57   Log-Likelihood:                -675.18
No. Observations:                 423   AIC:                             1360.
Df Residuals:                     418   BIC:                             1381.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          8.2678      1.080      7.653      0.000       6.144      10.391
x1             0.0001      0.004      0.031      0.975      -0.007       0.008
x2             0.0848      0.006     14.066      0.000       0.073       0.097
x3          3.112e-05   7.36e-06      4.228      0.000    1.67e-05    4.56e-05
x4            -0.5441      0.054    -10.078      0.000      -0.650      -0.438
==============================================================================
Omnibus:                      127.454   Durbin-Watson:                   0.929
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              559.741
Skew:                           1.253   Prob(JB):                    2.84e-122
Kurtosis:                       8.048   Cond. No.                     2.11e+06
==============================================================================
```

The results from the model summary show that CPI Used Vehicle (x1) is not statistically significant because its p-value (0.975) is higher than the t-test of the critical value of 0.05. Therefore, CPI Used Vehicles should be removed and rerun the model.
The model summary without CPI Used Vehicles is below.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.725
Model:                            OLS   Adj. R-squared:                  0.723
Method:                 Least Squares   F-statistic:                     368.5
Date:                Tue, 21 Apr 2020   Prob (F-statistic):          4.21e-117
Time:                        00:50:59   Log-Likelihood:                -675.18
No. Observations:                 423   AIC:                             1358.
Df Residuals:                     419   BIC:                             1375.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          8.2537      0.982      8.407      0.000       6.324      10.183
x1             0.0849      0.006     14.888      0.000       0.074       0.096
x2          3.132e-05   3.88e-06      8.073      0.000    2.37e-05    3.89e-05
x3            -0.5438      0.053    -10.219      0.000      -0.648      -0.439
==============================================================================
Omnibus:                      127.521   Durbin-Watson:                   0.929
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              560.366
Skew:                           1.254   Prob(JB):                    2.08e-122
Kurtosis:                       8.051   Cond. No.                     1.92e+06
==============================================================================
```

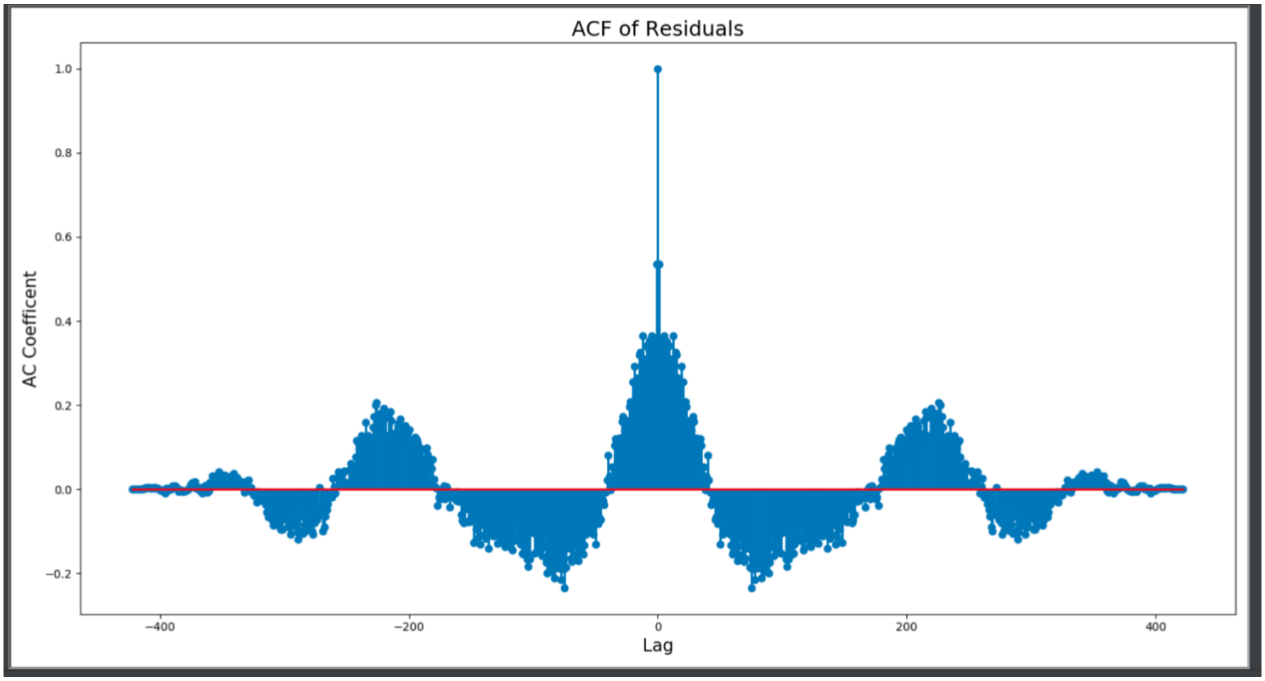
The new model summary shows that all features passed the t-test. However, the F-test certifies if the model as a whole is statistically significant. Similar to the t-test, if the p-value of the F-test is less than 0.05, the model is statistically significant. In the case of this model, the model is statistically significant because the p-value (4.21e-117) is less than 0.05.

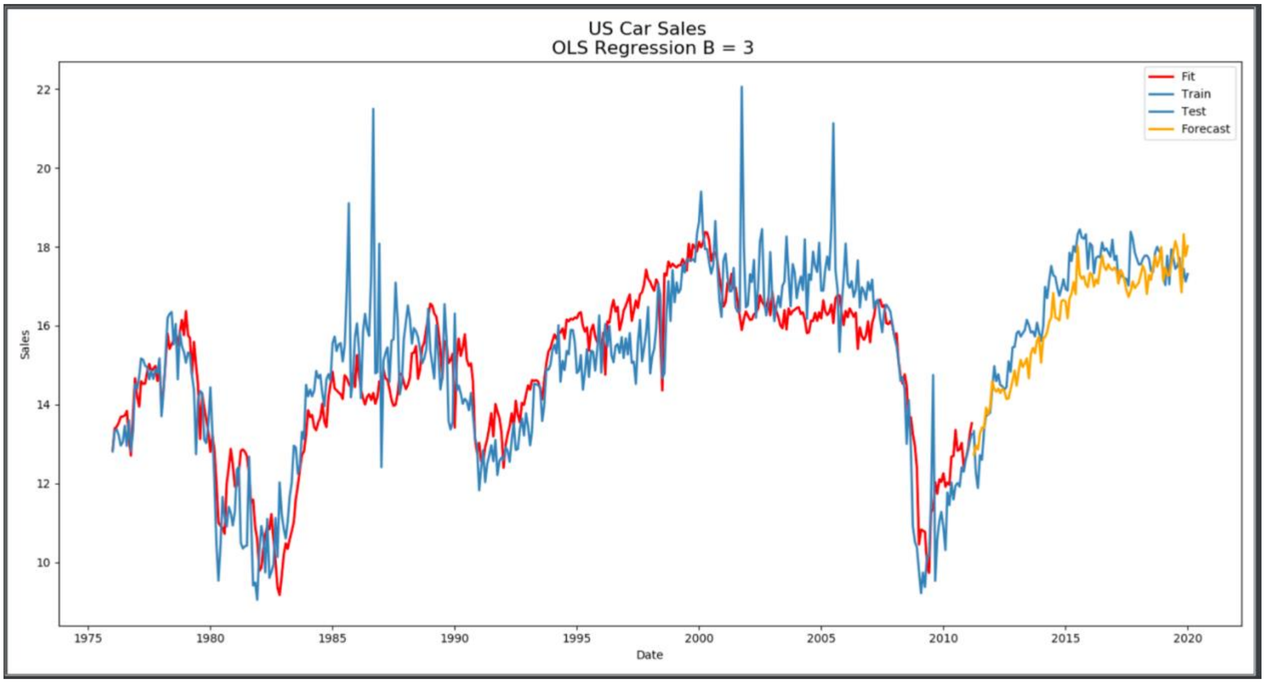
OLS Results:

Other important evaluation metrics on the residuals are r-squared and adjusted r-squared. These metrics demonstrate how well the model fits the training data and how well the predictors explain the variation in sales data. 0.725 and 0.723 are strong scores, but not the best. AIC and BIC are metrics used to compare models with a different number of predictors. This regression model had a slightly lower AIC and BIC than the previous regression model. Based upon the Q-value and the ACF plot below, the residuals are not white which illustrates that the OLS model does not capture the entire relationship between sales and its predictors.

Compared to HW, OLS had a worse SSE and MSE for residuals but had a much more reliable and accurate forecast.

OLS	Residual	Forecast
SEE	602.95	45.51
MSE	1.43	0.43
ME	3.987396e-13	0.39
RMSE	1.19	0.66
VAR	1.43	0.27
Q	2537.160	None





ARMA Model:

The last model that will be implemented is the autoregressive and moving model (ARMA). These models are similar to multivariable linear regression, where the predictors are lag versions of the series with their coefficients. This model can have up to na predictors or coefficients which depends on the number of lags taken into the model. This also determines the order of the model. An AR (2) model will take into account the series lagged at t -1 and t -2 steps. This concept often used to refer to AR models as "AR (na) models." One of the principal requirements of AR models is that they require stationarity – no trend or seasonality. There needs to be a constant level of variance and autocorrelation throughout the entire series. The equation for an ARMA process is below.

AR (*na*) MA(*nb*) Process

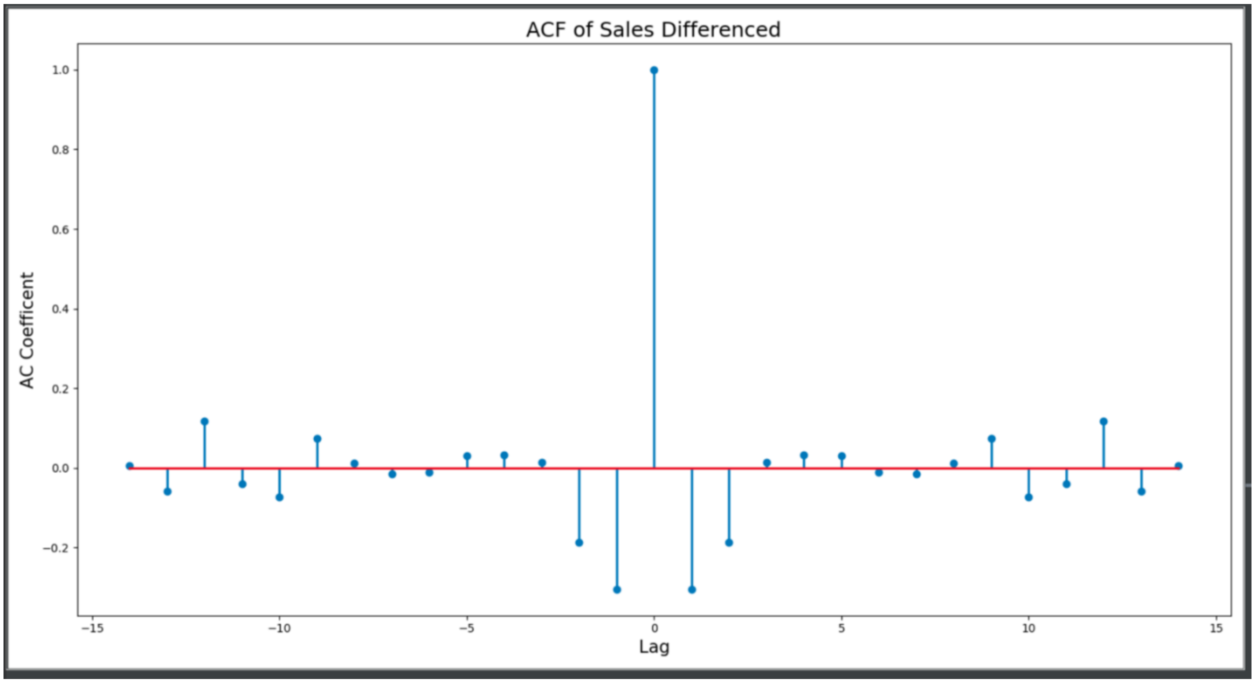
$$y(t) + a_1 y(t-1) + a_2 y(t-2) + \ldots + a_{n_a} y(t-n_a) = \epsilon(t) + b_1\epsilon(t-1) + b_2\epsilon(t-2) + \ldots + b_{n_b}\epsilon(t-n_b)$$

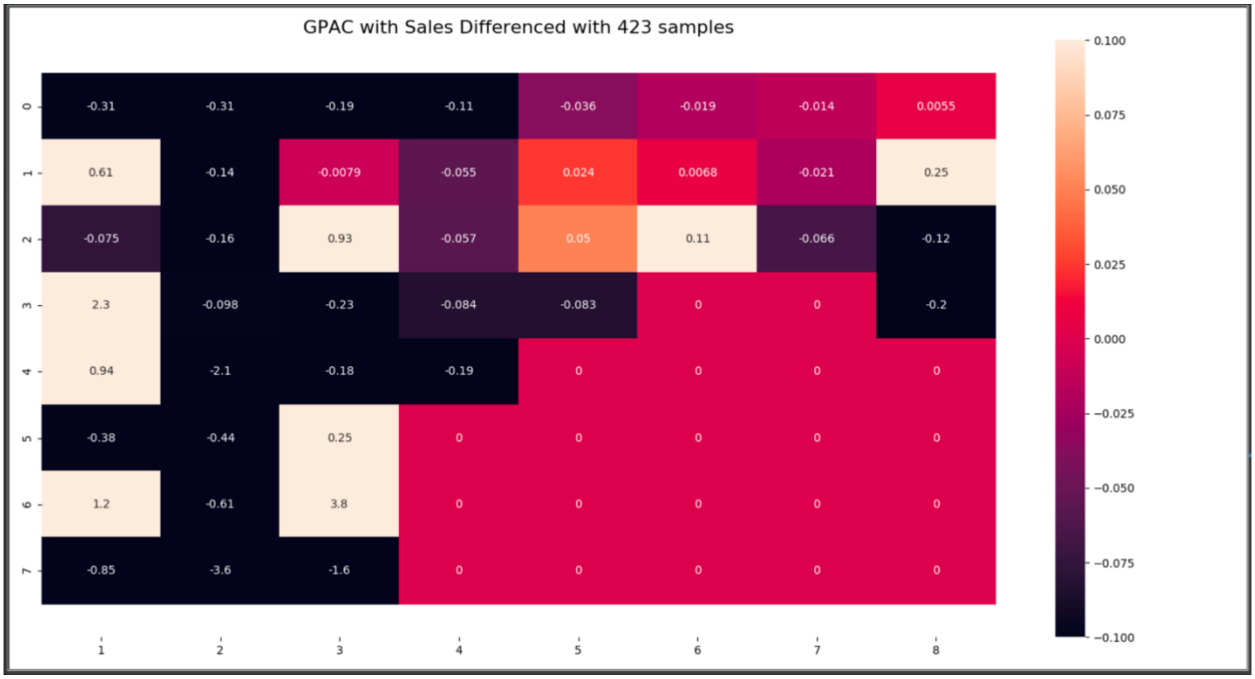
However, one of the shortcomings of using ARMA models for experimental and research purposes is that their performance is dependent on using the proper order or lag value of the AR and MA components for a dataset. This information is not always known nor is it trivial to detect. For this reason, the Generalized Partial Autocorrelation, or GPAC, is used to estimate the best order of the ARMA process for a given dataset. GPAC uses the Autocorrelation Function or ACF of a dataset to estimate the order of the ARMA process by constructing a table with the possible combinations for the ARMA orders. The parameters of this method are j, k, and phi. Ry(j) represents the estimated autocorrelation at lag j. j also denotes the estimated value nb for the MA process. k denotes the estimated value for na for the AR process. Phi is the result of the division between the determinate of the matrices of the given formula below:

Once order determination is established the next step is parameter estimation of theta, which represents the coefficients for AR and MA processes. There are a couple of algorithms that can estimate parameters, but this study will focus on the Levenberg-Marquardt algorithm or LM. The coefficients can be estimated using Maximum Likelihood Estimation or MLE in conjunction with the Levenberg-Marquardt algorithm.

Before the sales data can be implemented into the GPAC table it needs to undergo a first-order transformation because the original raw data is not stationary.

The ACF plot and GPAC table for the differenced US sales data are below. An important caveat of the GPAC table is that it is still an estimation. Even if a clear pattern shows, it may not be the optimal order, but it works as a starting point.





	1	2	3	4	5	6	7	8
0	-0.31	-0.31	-0.19	-0.11	-0.036	-0.019	-0.014	0.0055
1	0.61	-0.14	-0.0079	-0.055	0.024	0.0068	-0.021	0.25
2	-0.075	-0.16	0.93	-0.057	0.05	0.11	-0.066	-0.12
3	2.3	-0.098	-0.23	-0.084	-0.083	0	0	-0.2
4	0.94	-2.1	-0.18	-0.19	0	0	0	0
5	-0.38	-0.44	0.25	0	0	0	0	0
6	1.2	-0.61	3.8	0	0	0	0	0
7	-0.85	-3.6	-1.6	0	0	0	0	0

Based upon the GPAC table two possible ARMA orders are ARMA (4,2) and ARMA (3,2). The number of parameters is passed to the LM algorithm program to estimate their values.

The results of the LM parameter estimates for ARMA (4,2) are below:

Theta	Estimated Theta	STD
a1	-0.344	0.102
a2	-1.140	0.057
a3	0.444	0.071
a4	0.206	0.051
b1	0.118	0.094
b2	-0.834	0.090

95% Confidence Interval

Theta	Theta - 2xSTD	Theta + 2xSTD
a1	-0.547	-0.140
a2	-1.253	-1.026
a3	0.302	0.585
a4	0.103	0.309
b1	-0.071	0.307
b2	-1.015	-0.654

The importance of the confidence interval is to make sure that the estimated parameter is not zero. If the confidence interval includes zero, then the corresponding parameter should be ignored and drop the matching AR or MA order by one. Hence the ARMA (4,2) should be an ARMA (4,1).

The results of the LM parameter estimates for ARMA (3,2) are below:

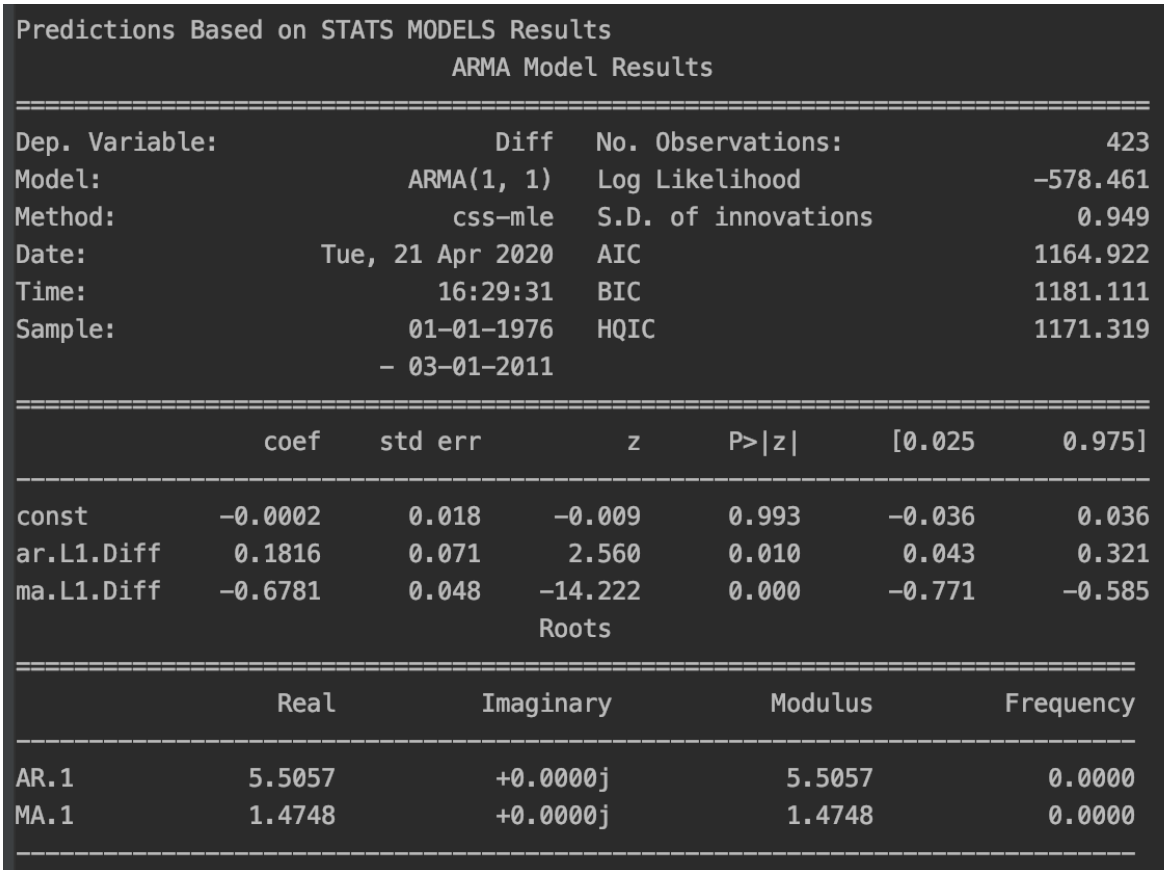
Theta	Estimated Theta	STD
a1	-1.303	0.224
a2	0.148	0.349
a3	0.239	0.137
b1	-0.810	0.230
b2	-0.045	0.229

95% Confidence Interval

Theta	Theta - 2xSTD	Theta + 2xSTD
a1	-1.750	-0.855
a2	-0.549	0.845
a3	-0.035	0.5132
b1	-1.270	-0.349
b2	-0.502	0.413

Only a1 and b1 are relevant parameters since their confidence interval does not include zero. The ARMA process should be ARMA (1,1).

The Stats.Model (SM) package in Python is used to forecast and further validate chosen ARMA processes. The ARMA processes that were chosen and implemented in the LM algorithm failed the SM t-tests. New processes need to be chosen. ARMA (1,1) passed the SM parameter t-test. The results are below.



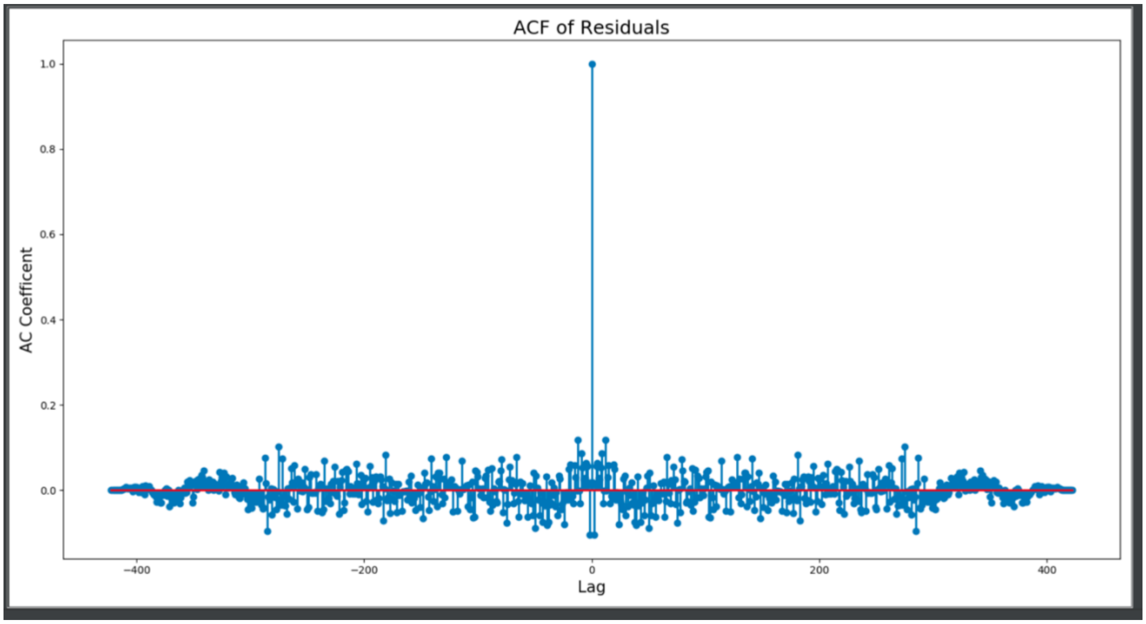
```
Predictions Based on STATS MODELS Results
                              ARMA Model Results
==============================================================================
Dep. Variable:                   Diff   No. Observations:                  423
Model:                     ARMA(1, 1)   Log Likelihood                -578.461
Method:                       css-mle   S.D. of innovations              0.949
Date:                Tue, 21 Apr 2020   AIC                           1164.922
Time:                        16:29:31   BIC                           1181.111
Sample:                    01-01-1976   HQIC                          1171.319
                         - 03-01-2011
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0002      0.018     -0.009      0.993      -0.036       0.036
ar.L1.Diff     0.1816      0.071      2.560      0.010       0.043       0.321
ma.L1.Diff    -0.6781      0.048    -14.222      0.000      -0.771      -0.585
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            5.5057           +0.0000j            5.5057            0.0000
MA.1            1.4748           +0.0000j            1.4748            0.0000
-----------------------------------------------------------------------------
```

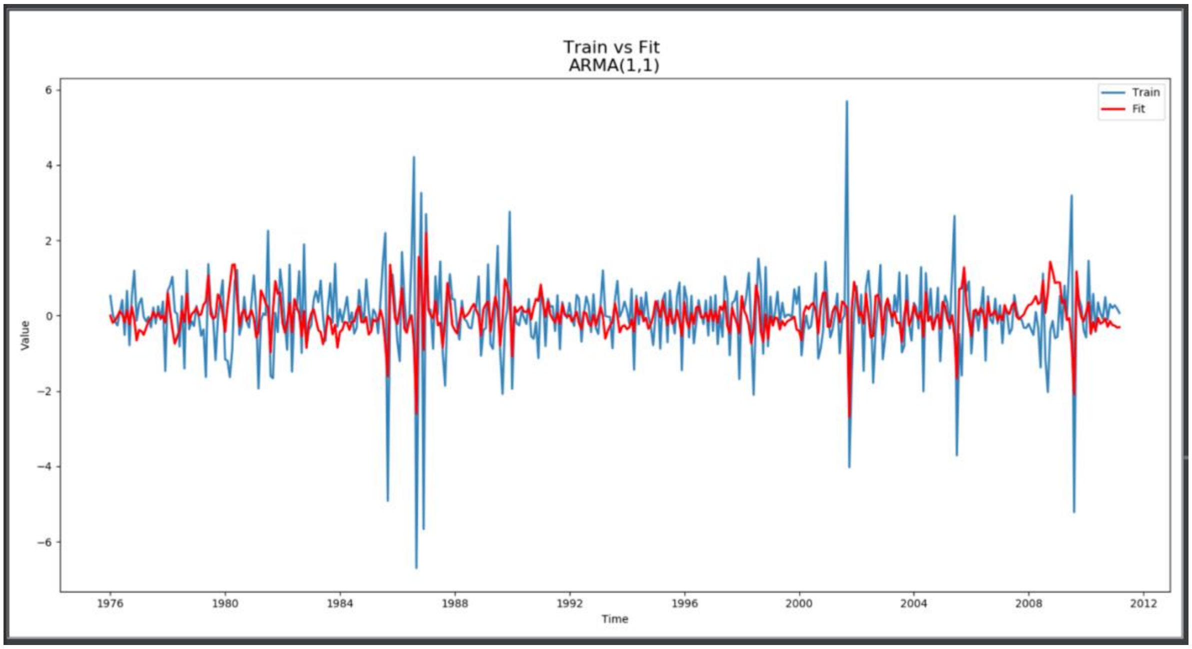
ARMA (1,1)	Residuals	Forecast
ME:	0.001	0.059
SSE:	381.376	38.530
MSE:	0.902	0.363
RMSE:	0.034	0.242
Variance of Error:	0.902	0.360
Q:	182.522 < CHI Critical: 491.430	None
Residuals are white	YES	None

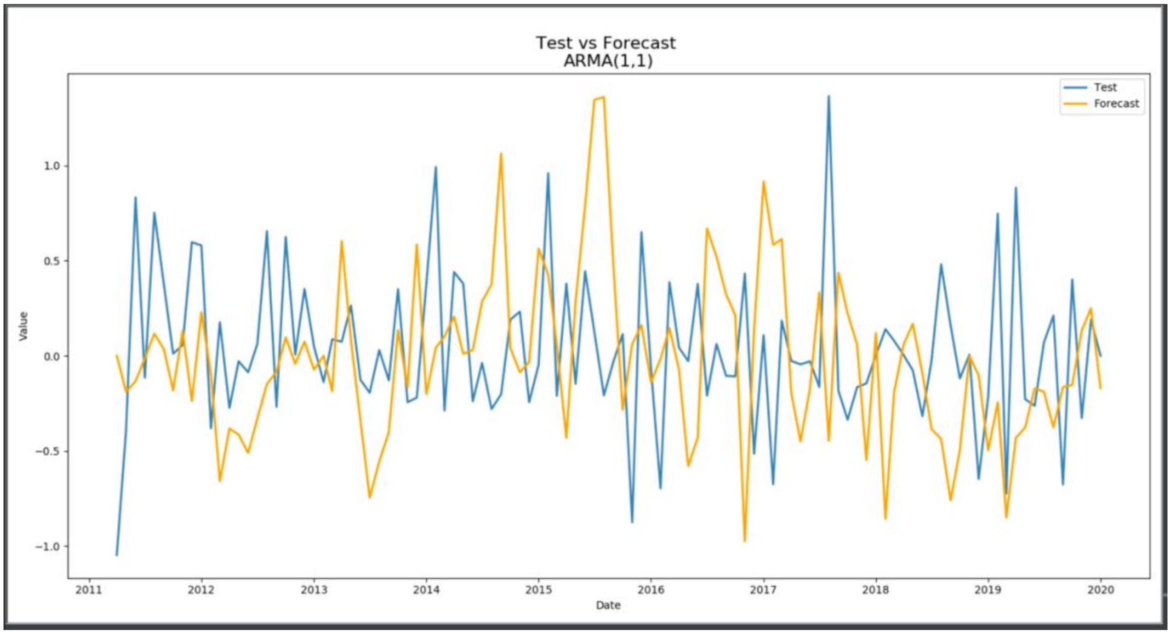
The ARMA (1,1) process needs to pass other key tests beyond the t-test to validate the model's statistical significance. The estimated parameters need to pass the Zeros/Poles root test which makes sure the process does not have AR an MA components with the same roots. In this process, the roots for the AR component are 5.5 and the MA is 1.4 therefore it passes the Zero/Poles Cancelation test. Furthermore, the residuals of the process must be uncorrelated, if they are then the model does not capture the full underlying relationship of the data and the model is unreliable. To pass the Chi-square test Q value of the process must be less than the Chi-square critical value. Q is the product between the number observations and the sum of squares the ACF of the residuals minus the first lag (0).

$$Q = N \sum_{\tau=1}^{h} R_e^2(\tau)$$

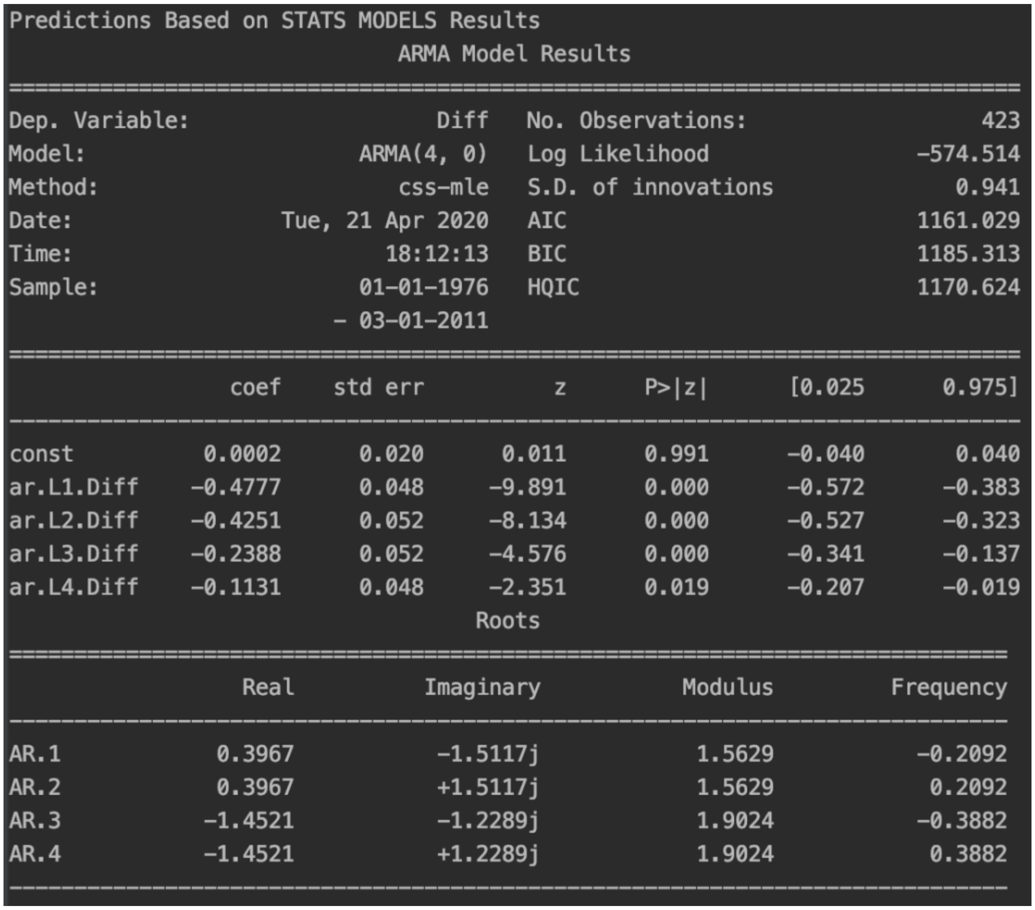
The Chi-square critical value is based on the degrees of freedom (DOF) of the process which is the number of lags – the number of parameters (na +nb) and a set alpha threshold (0.01). Here the value of Q (182.522) is less than the Chi-square critical value of 491.430 which makes the residuals uncorrelated or "white." The graphs for the fitted values and forecast are below.



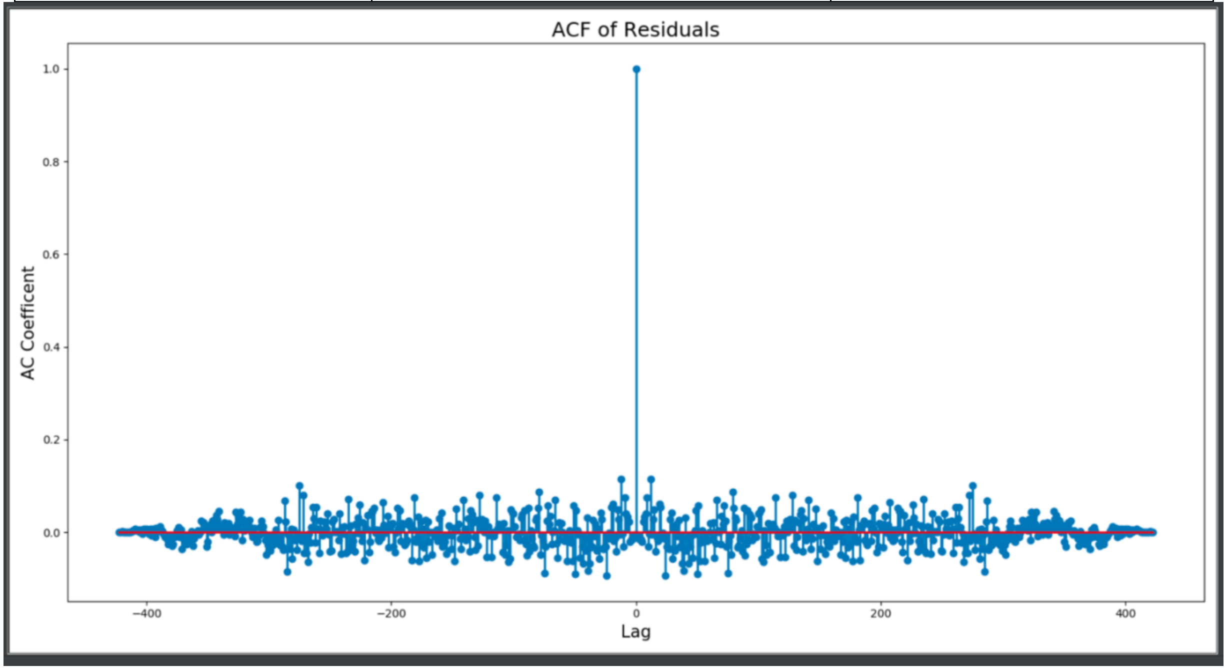


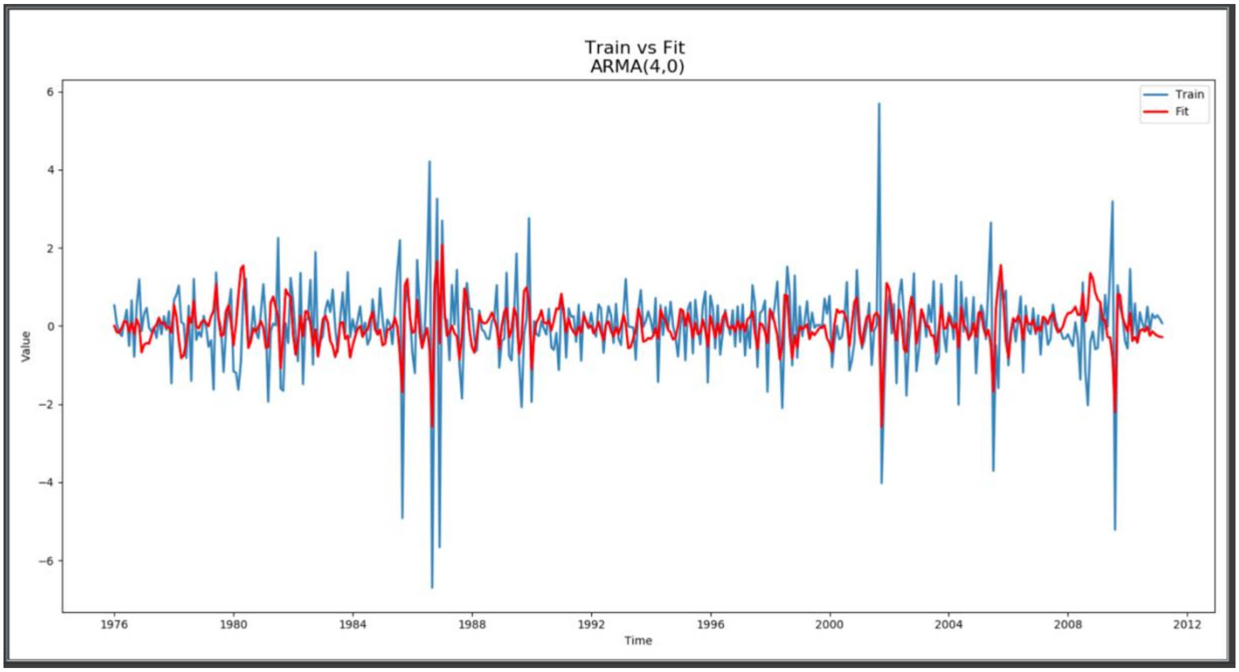


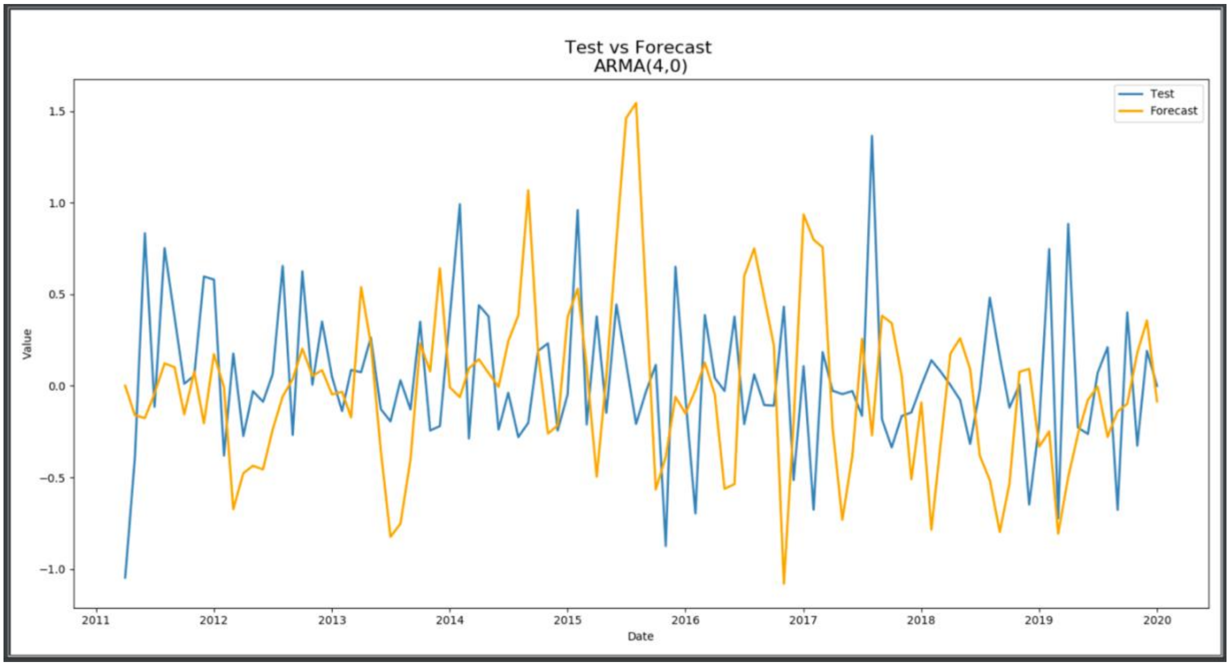
Another ARMA process that passes the relevant tests, t-test, Zeros/Poles, and Chi-square is ARMA (4,0). The model summary and results are below.


```
Predictions Based on STATS MODELS Results
                              ARMA Model Results
==============================================================================
Dep. Variable:                   Diff   No. Observations:                  423
Model:                     ARMA(4, 0)   Log Likelihood                -574.514
Method:                       css-mle   S.D. of innovations              0.941
Date:                Tue, 21 Apr 2020   AIC                           1161.029
Time:                        18:12:13   BIC                           1185.313
Sample:                    01-01-1976   HQIC                          1170.624
                         - 03-01-2011
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0002      0.020      0.011      0.991      -0.040       0.040
ar.L1.Diff    -0.4777      0.048     -9.891      0.000      -0.572      -0.383
ar.L2.Diff    -0.4251      0.052     -8.134      0.000      -0.527      -0.323
ar.L3.Diff    -0.2388      0.052     -4.576      0.000      -0.341      -0.137
ar.L4.Diff    -0.1131      0.048     -2.351      0.019      -0.207      -0.019
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            0.3967           -1.5117j            1.5629           -0.2092
AR.2            0.3967           +1.5117j            1.5629            0.2092
AR.3           -1.4521           -1.2289j            1.9024           -0.3882
AR.4           -1.4521           +1.2289j            1.9024            0.3882
-----------------------------------------------------------------------------
```

ARMA (4,0)	**Residuals**	**Forecast**
ME:	0.001	0.054
SSE:	374.271	41.049
MSE:	0.885	0.387
RMSE:	0.032	0.231
Variance of Error:	0.885	0.384
Q:	176.255 < CHI Critical: 489.269	None
Residuals are white	YES	None

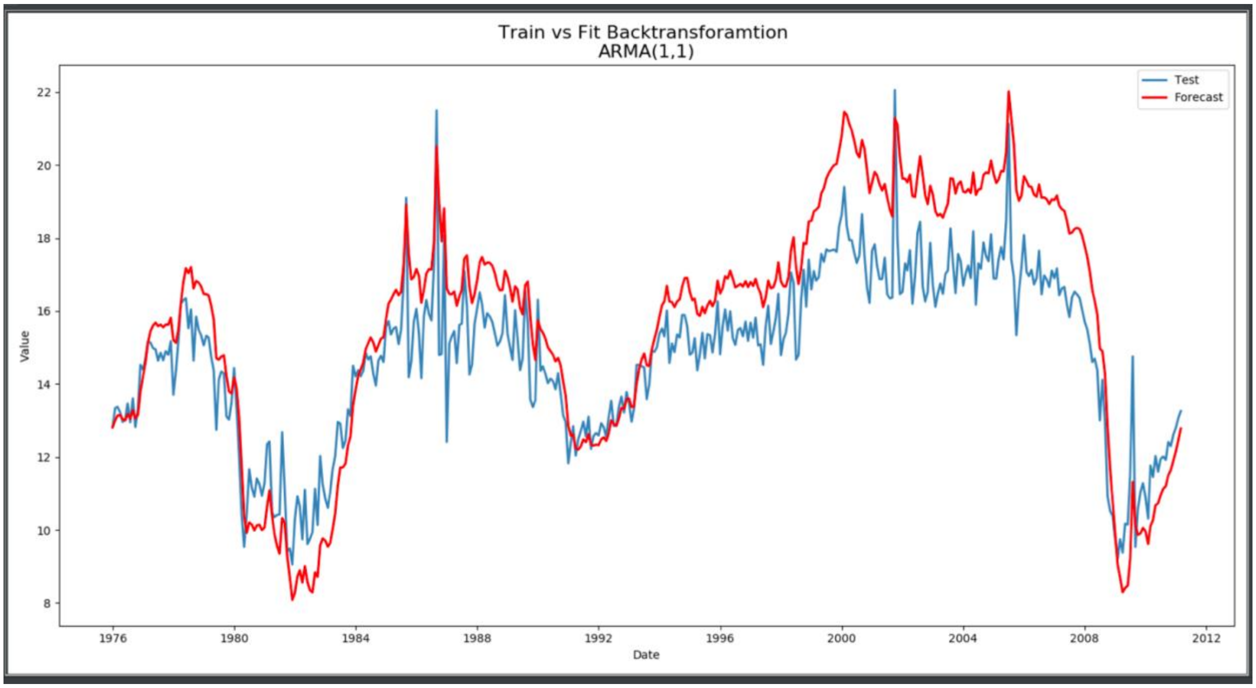


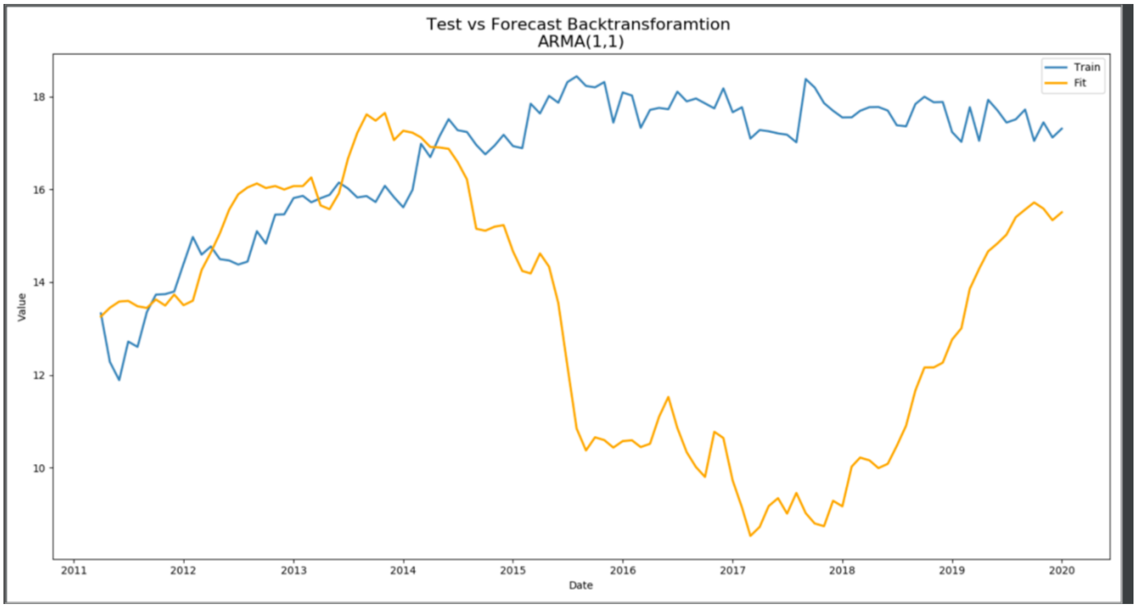




Reserve Transformation:

The previous results and graphs are for the differenced data of the US vehicle sales data, not of the actual data itself. To get the real and useful results the results need to go through "reverse transformation." This is done by adding the results of the differenced data and the original data of the prior observation. yk is the original data and zk is the differenced data.

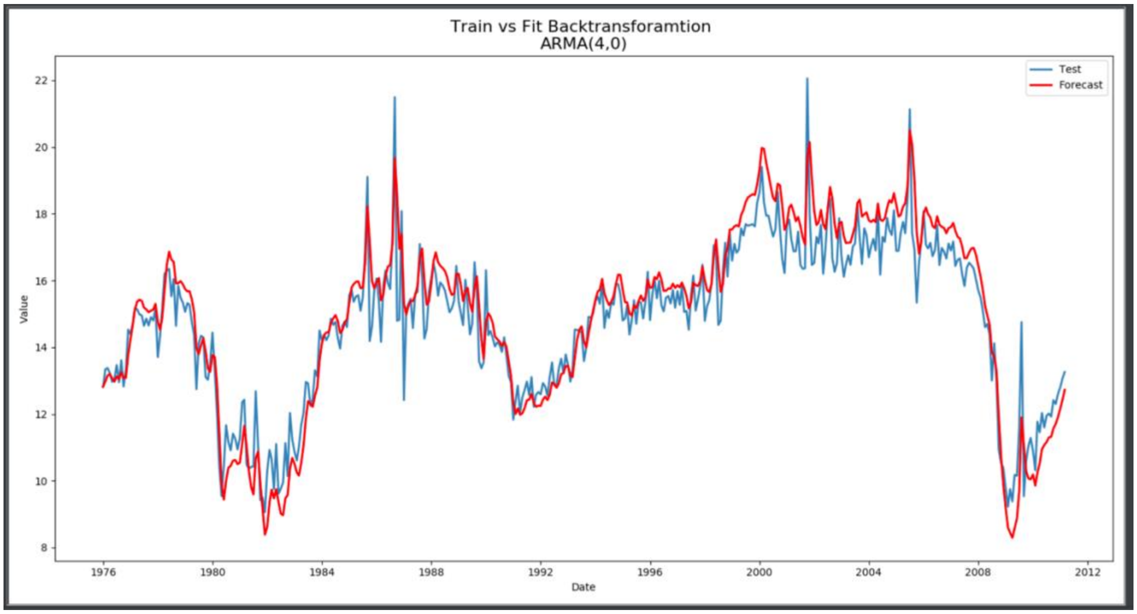


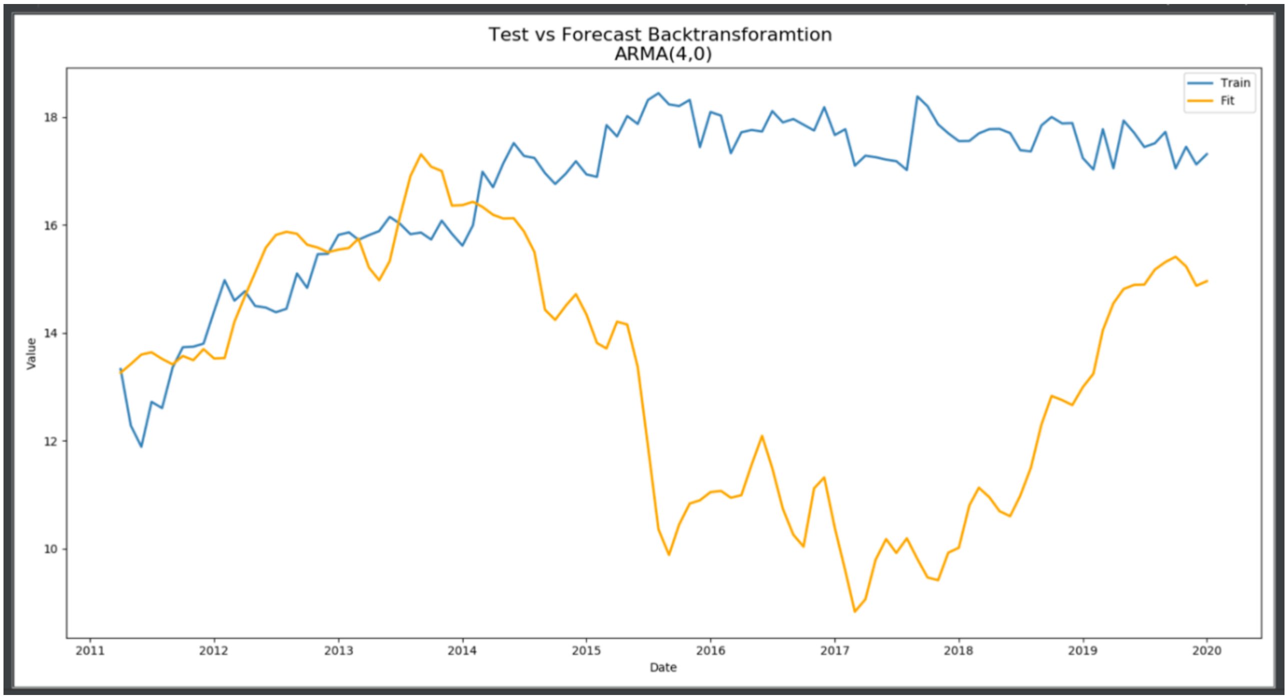


ARMA (1,1) RT	Residuals	Forecast
ME:	-0.895	3.382
SSE:	1058.297	2633.521
MSE:	2.502	24.845
Variance of Error:	1.700	13.405

There is a stark difference in the results between the stationary and reverse transformation of the ARMA (1,1). The reversed transformed data is no longer as accurate concerning residual SSE and MSE

Next are the results of the ARMA (4,0).





ARMA (4,0) RT	Residuals	Forecast
ME:	-0.213	3.382
SSE:	288.688	2362.548
MSE:	0.682	22.288
Variance of Error:	0.637	10.847
Q	1595.088 > CHI-Critical: 603.310	None
Residuals White	NO	None

Similar to the ARMA (1,1) model, the ARMA (4,0) the residual SSE, and MSE all dramatically increased. However, the SSE of ARMA (4,0) is much lower than the SSE of ARMA (1,1). This is clearly shown as the fitted data forms a "tighter" fit to the training data in ARMA (4,0), whereas ARMA (1,1) the fitted data is consistently overestimating. Another key difference that occurred in the reverse transformation was that the models switched from unbiased to biased models. This can be observed in the change in mean error (ME) of the residuals. Unbiased models have a residual ME of zero, whereas biased models have an absolute difference that is much higher. In the case of the two ARMA processes, they had 0.001 ME in the stationary results, but increased to – 0.2, and – 0.8. Overall, the ARMA (4,0) is the better of the two models.

Final Model Selection:

Selecting the best model from various models can be a difficult choice as several criteria need to be evaluated and compared across models. The three main metrics to compare are SSE, MSE, and RMSE. The four models that will be compared are HW, OLS, and ARMA (4,0), and the naïve drift as a baseline. The error stats for each model are below.

Naïve Drift	Residual	Forecast
SEE	3398.41	1423.20
MSE	8.03	13.42
RMSE	2.834	3.66

ARMA (4,0) RT	Residual	Forecast
SSE:	288.688	2362.548
MSE:	0.682	22.288
RMSE	0.825	4.721

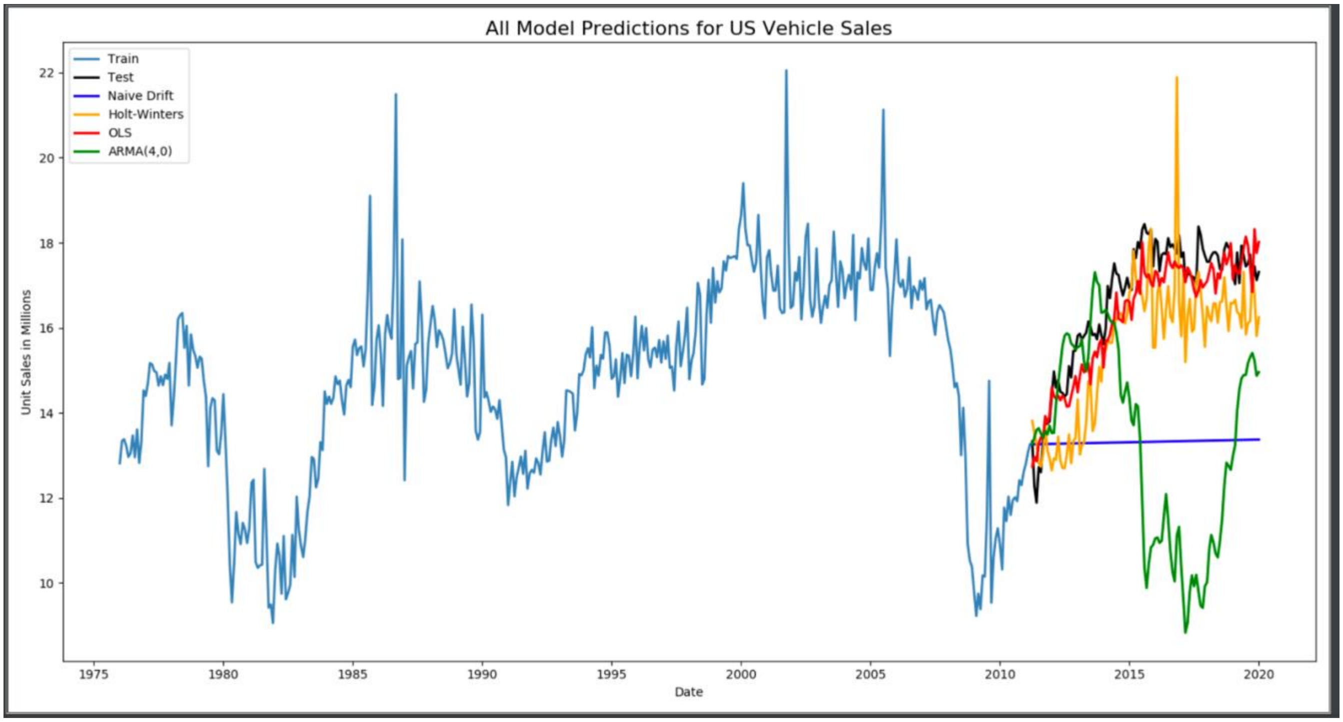
OLS	Residual	Forecast
SEE	602.95	45.51
MSE	1.43	0.43
RMSE	1.19	0.66

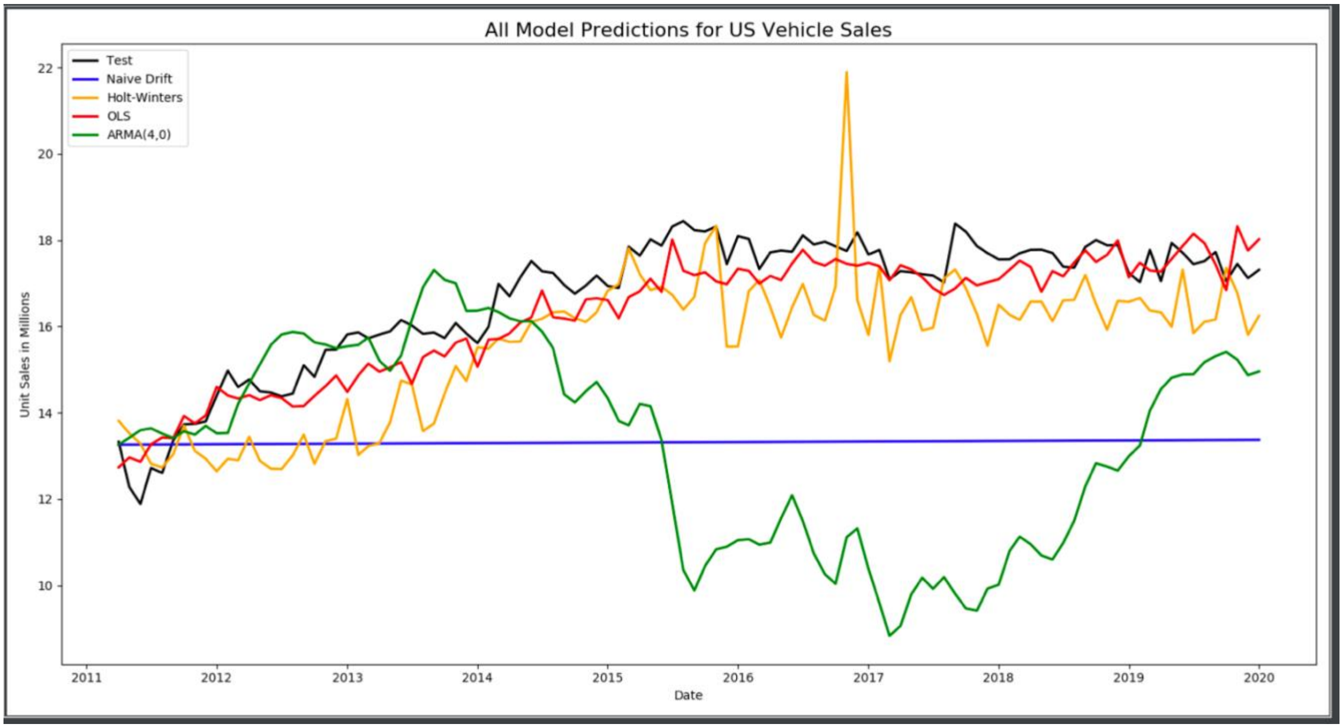
Holt-Winters	Residual	Forecast
SEE	191.95	216.18
MSE	0.45	2.04
RMSE	0.671	1.428

The first step is to see which models did not outperform the baseline line model. The ARMA model fails to outperform the baseline model. The naïve model had about 40% less SSE and MSE forecast error. Choosing between the OLS and HW models comes down to comparing their forecast errors. The OLS had significantly better forecast accuracy than HW with well under 50% less error across all three metrics, despite its much higher residual errors. Therefore, the best model for US vehicle sales is the OLS model.

Prediction:

Even though forecasts have been made for each model, the graphs below will illustrate how each model compares to one another. The first graph includes the training data to show how well the model behaved based on past data. The second graph zooms into the test and forecast data to have a better understanding of scale.


All Model Predictions for US Vehicle Sales
Train
Test
Naive Drift
Holt-Winters
OLS
ARMA(4,0)
Unit Sales in Millions
22
20
18
16
14
12
10
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
Date


All Model Predictions for US Vehicle Sales
Test
Naive Drift
Holt-Winters
OLS
ARMA(4,0)
Unit Sales in Millions
22
20
18
16
14
12
10
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Date

Summary & Conclusion:

Overall, each of the three sophisticated models showed some degree of accurately predicting total monthly US vehicle sales, but the best model was OLS. The OLS model had the best forecasting error metrics in RMSE, MSE, and SSE. Beyond error metrics, examining the forecast graph gives insight into how each model failed or successfully predicted US vehicle sales. The ARMA (4,0) model was able to capture the upward trend in the sales data from 2011 until mid-2013. Afterward, predictions were grossly under forecasted. The pivot in the direction of prediction may be due to the model responding to the cyclical aspect of the data at the wrong time. As a result, the ARMA model was less successful than the baseline drift model. The next best model was the HW model. The HW model had better residual errors than OLS but performed much worse in forecasting. The HW model was able to properly predict the trend and some cyclical aspects as it followed the increasing sales at a decreasing rate similar to OLS. The shortcoming of the HW model was that it consistently underestimated sales more than OLS, and sometimes grossly overestimated such as in the clear spike near 2017. Although the OLS model was the best of three models, it has its shortcomings as well. Reviewing the fitted values of OLS, the model did poorly at fitting to the cyclical peaks and instead fit more linearly by "cutting through" the peaks. If this is the case, then the OLS model will have trouble forecasting further steps beyond the test set available when sales eventually begin to decline. The reason this is not evident in this study is that the test set happens to take the form of a linear trend. It is foreseeable that when more sales data is available, the HW method might prove more appropriate. Therefore, businesses tied to the US auto industry that depend on vehicle sales forecast can expect to have reliable results from the OLS model in the short run, and reassessing model accuracy should be explored gaining the future. Further improvements can be made to the OLS model by including other features that may capture a better relationship with US vehicle sales or taking extra steps validation models with cross-validation and k-folds.

Sources:

Alliance of Automobile Manufacturers; https://autoalliance.org/in-your-state. April 20, 2020.

U.S. Bureau of Economic Analysis, Total Vehicle Sales [TOTALSA], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/TOTALSA, February 25, 2020.

U.S. Bureau of Labor Statistics, Consumer Price Index for All Urban Consumers: All Items in U.S. City Average [CPIAUCSL], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CPIAUCSL, February 27, 2020.

U.S. Bureau of Labor Statistics, Consumer Price Index for All Urban Consumers: All Items Less Food and Energy in U.S. City Average [CPILFESL], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CPILFESL, February 27, 2020.

U.S. Bureau of Labor Statistics, Consumer Price Index for All Urban Consumers: Used Cars and Trucks in U.S. City Average [CUSR0000SETA02], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CUSR0000SETA02, February 27, 2020.

U.S. Bureau of Labor Statistics, All Employees, Total Nonfarm [PAYEMS], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/PAYEMS, February 27, 2020.

U.S. Bureau of Labor Statistics, Unemployment Rate [UNRATE], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/UNRATE, February 27, 2020.

Board of Governors of the Federal Reserve System (US), Capacity Utilization: Durable Manufacturing: Automobile and light duty motor vehicle [CAPUTLG33611S], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CAPUTLG33611S, February 27, 2020

U.S. Bureau of Labor Statistics, Consumer Price Index for All Urban Consumers: New Vehicles in U.S. City Average [CUSR0000SETA01], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CUSR0000SETA01, February 27, 2020.