Project Technical Report

Dennis Kelly, Robert Steward, Ryan-Arnold Gamilo

Introduction

As it did in many other industries, the COVID-19 Pandemic caused immense disruption in the automotive industry: reducing sales, straining supply-chains, and casting great uncertainty over almost every facet of the business. Even so, automotive manufacturers must still make decisions about the kinds of vehicles to sell, and produce forecasts of future sales for the purpose of managing supply-chains. Moreover, if these decisions and forecasts are going to be reliable, they have to be reached in a principled and quantitatively rigorous way.

Our group adopts a data-driven approach to these questions. We examine historical data to discern the presence of any long-term trends in vehicle sales, as well as any possible associations between sales and external factors such as the price of gasoline. We then look at sales data from 2020 and 2021, to see which car types and which features are most strongly associated with present sales. Taking note of variability between U.S. states in the type of best-selling car, we investigate if this variability is related to differences in the average commute between states. Finally, we look toward the future, assessing the feasibility of using machine learning to produce monthly sales forecasts.

Description of Datasets

We pulled data from several different sources in our research. We give a brief summary of each below:

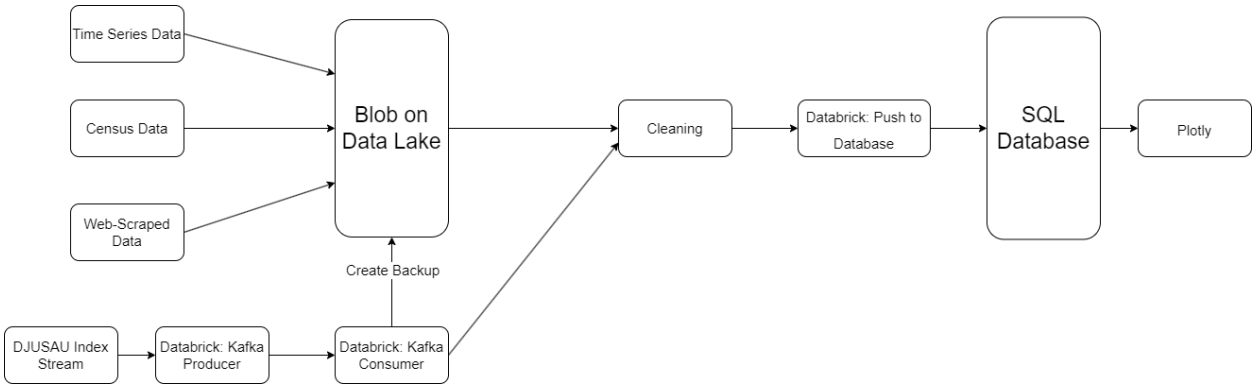
1. US Bureau of Economic Analysis, Supplemental Estimates - Total Vehicle Sales, via FRED
    a. Monthly sales figures for automotives in the US, broken down further into 5 categories: domestic automobiles, domestic light trucks, foreign automobiles, foreign light trucks, and heavy trucks
2. US Census American Community Survey
    a. Data on commuting in the United States, including the number of workers and average travel time to work for all commuters using a particular method in each state
    b. Data on vehicle access by state
3. Cars.com
    a. Web-scraped this site to get information on the characteristics of cars, such as their mpg, horsepower, drivetrain type and many others
4. Carsalesbase.com
    a. Web-scraped this site to get 2020 and 2021 sales data for many makes and models of car, with a particular focus on the best-sellers
5. Edmund's, Most Popular Cars in America
    a. Information on which model of car sold best in each state in 2021

Data Platform Overview

After extracting the data above, we then cleaned it and stored it in a container on an Azure Data Lake. From the Data Lake, we used a Databrick to push the Lake's contents to the database. The Databrick is then integrated into an Azure Data Factory Pipeline, which is triggered when there is an update of any of the blob files.

For the DJUSAU index data, the producer Databrick is integrated into a separate Factory Pipeline from the Consumer Databrick. This is to ensure that the producer and consumer run separately, since the producer will succeed on every trigger, while the consumer will sometimes fail due to the lack of new data from the topic.

This can be seen in the diagram below:

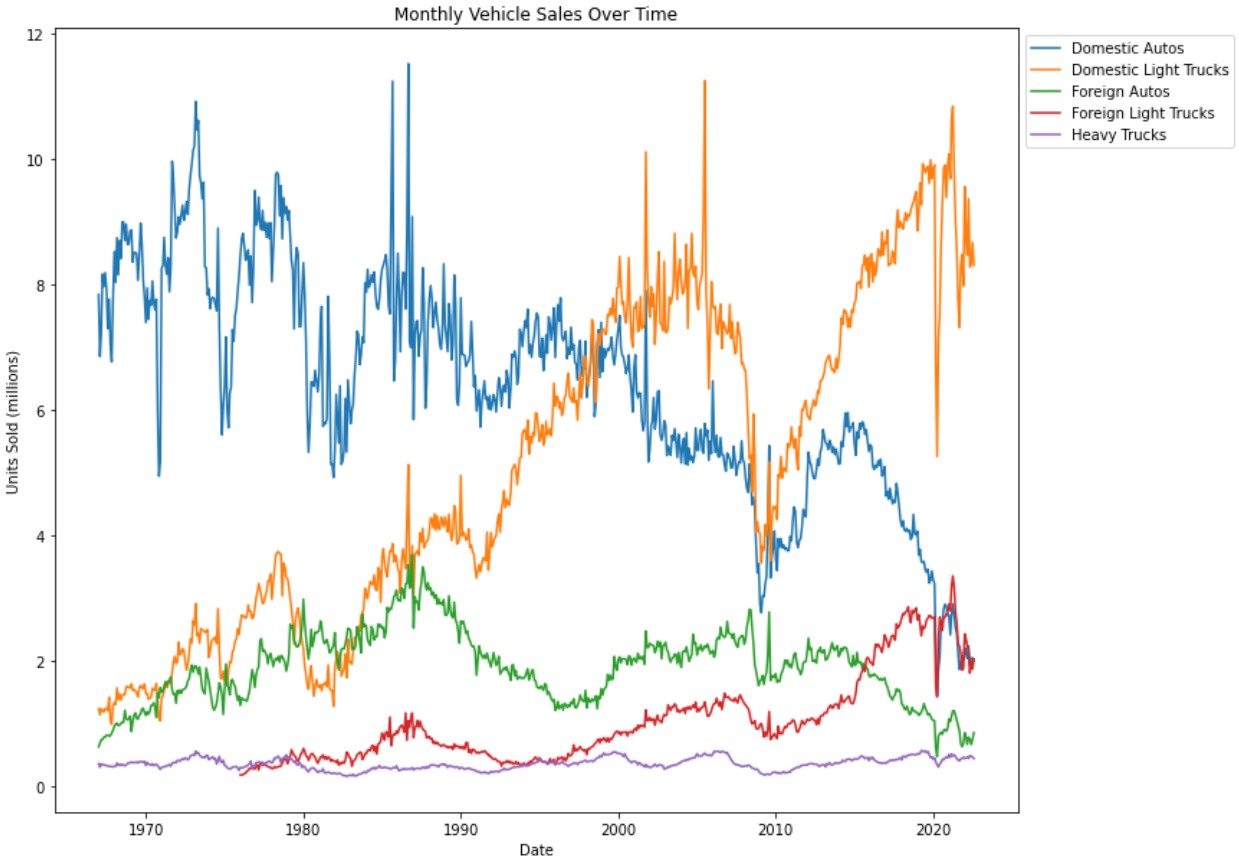


Results

We split this section further into subsections containing the analysis of the historical data, analysis of the 2020 and 2021 sales data, and finally our machine learning results.

Historical Data

The first thing we wanted to analyze in the historical monthly sales data was the presence or absence of any long term trends. To do so, we graphed the values for all five vehicle categories over time and visually inspected the resulting plot. We reproduce our figure below:

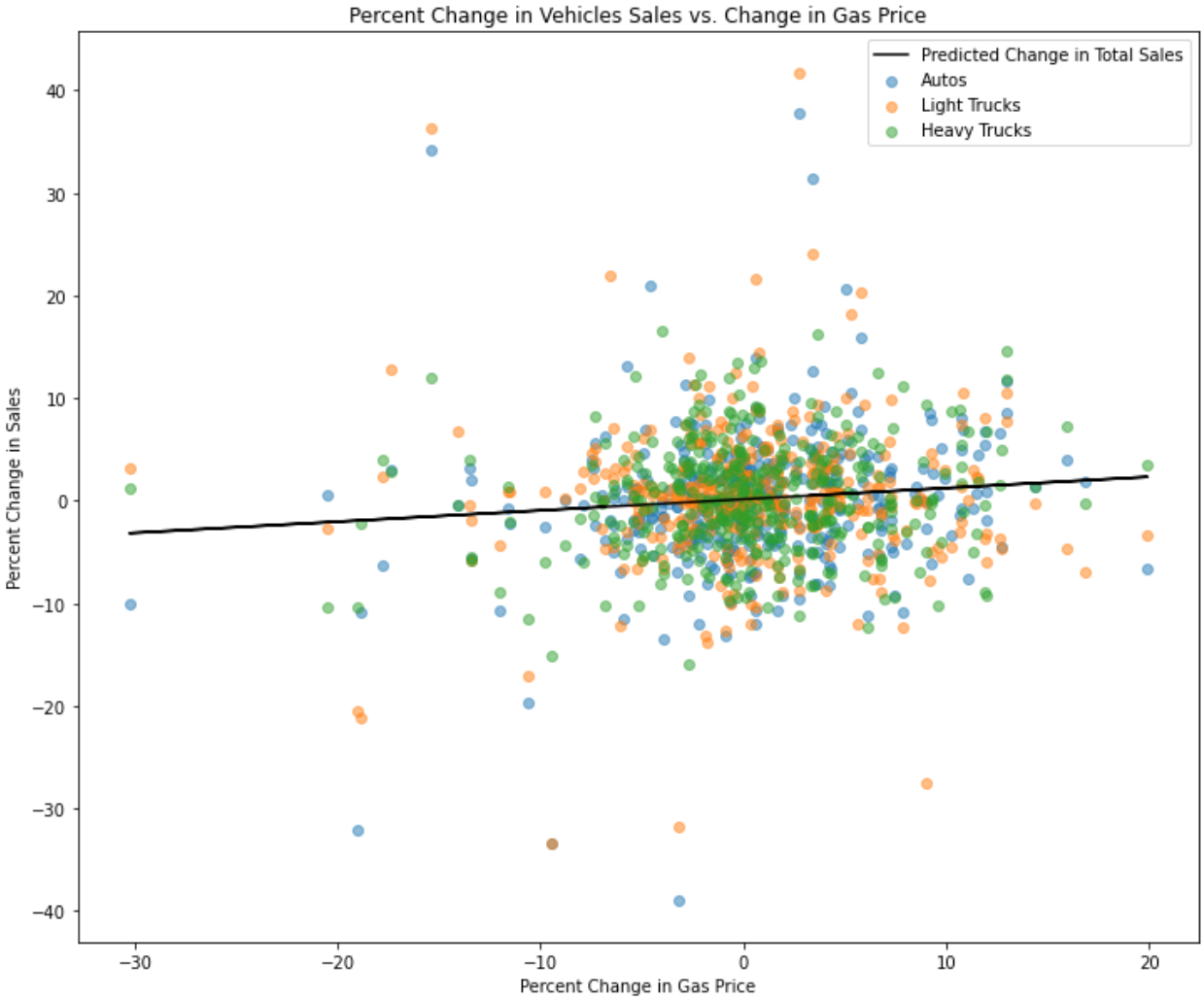


We note several trends very clearly. First, domestic vehicles consistently outsell foreign vehicles, and heavy trucks are almost always the lowest selling category. Second, over time there has been a long-term decline in domestic automobiles, paired with a long-term increase in domestic light trucks. Foreign automobiles and light trucks also exhibit this trend. Heavy trucks, meanwhile, remain at a very stable level over time. Third, we note the presence of market shocks, in particular the Great Recession in 2008 and the COVID pandemic in 2020. Both of these caused sudden and sharp reduction in sales volume.

Next we wanted to investigate whether sales were related in any way to the contemporary price of gasoline. While the obvious thing to do is to check the correlation between the price of gasoline and sales, this methodology is flawed as it does not account for so-called serial correlation. It could be the case that both gas prices and sales evolve over time according to independent trends; in that case, time would be a confounding variable, and we would find an association where non-exists.

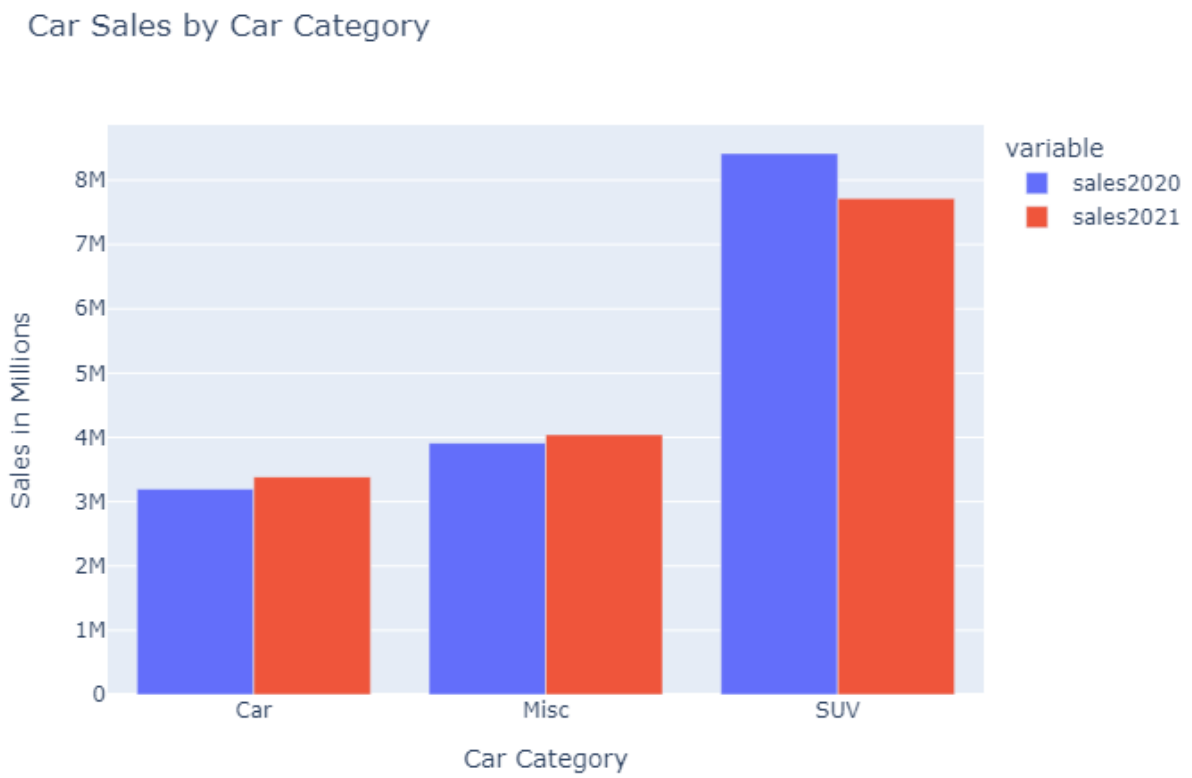
To answer this question then we chose to regress the month-to-month percentage change in total vehicle sales against the corresponding percentage change in gasoline price. This resolves the issue of serial correlation. Our regression coefficient for change in gas price was 0.11, which means that a 1% increase in the price of gasoline in a particular month was associated with a 0.11% increase in the amount of vehicles sold in that same month. We were surprised by the positive sign, as we hypothesized that increasing gas prices would push consumers towards other methods of travel. Moreover, this coefficient achieved statistical significance at the 0.05

level. However, the practical significance of this finding is minimal, as the effect size is very small, and thus of limited explanatory power. The graph below emphasizes this point visually.

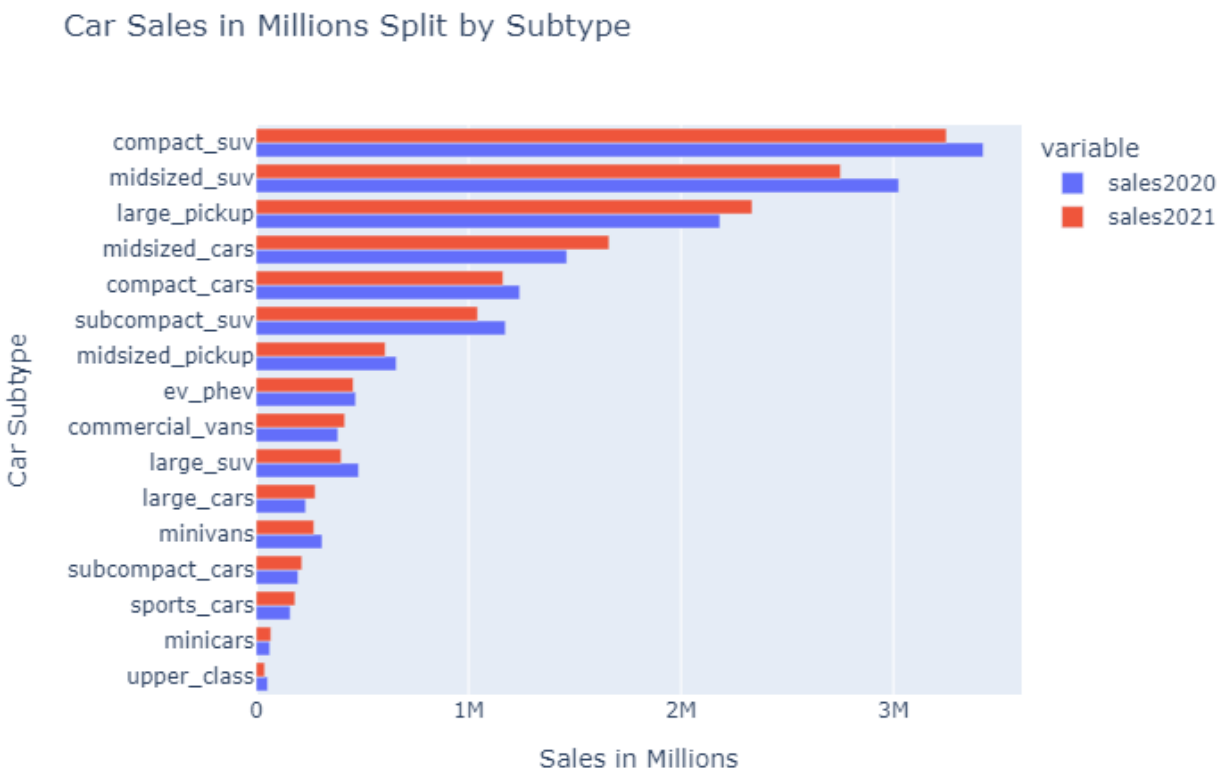


2020 and 2021 Sales Data

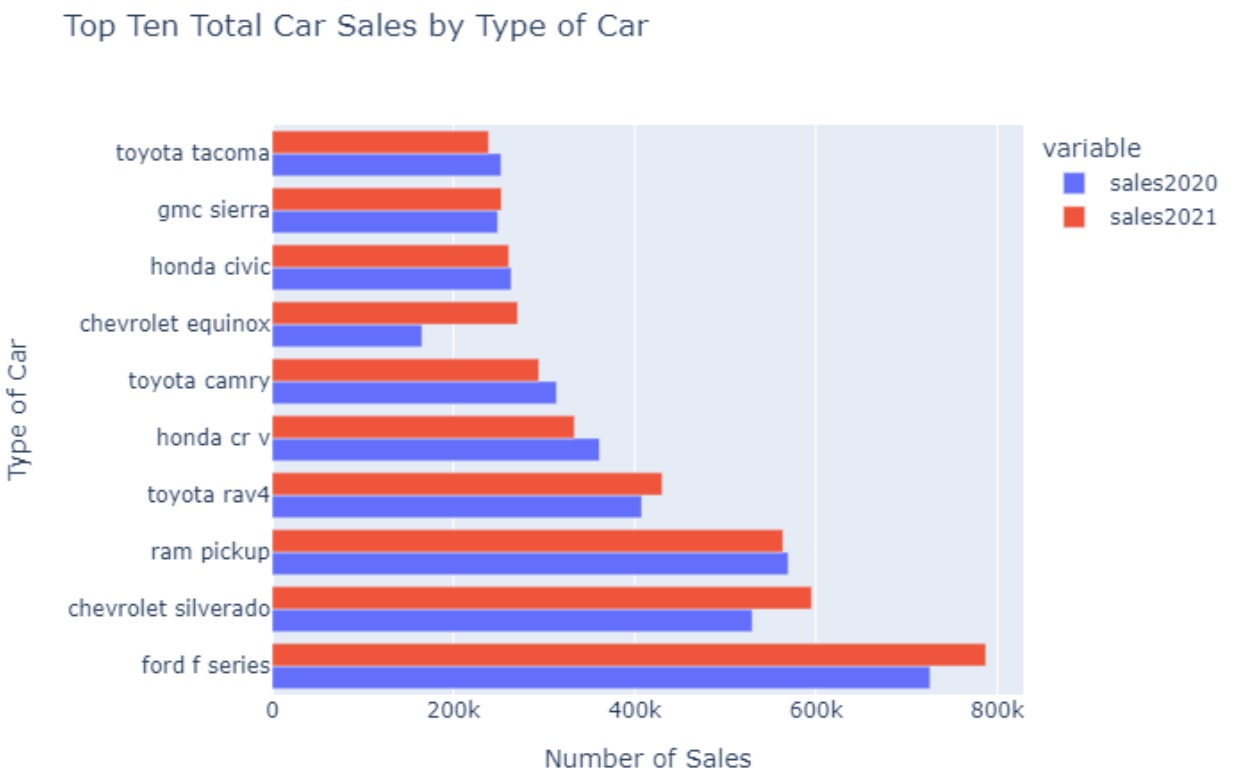
In order to look at data representing modern car sales and architecture, we web scraped data from carsalesbase.com and cars.com. The two sources gave us 2020 and 2021 sales for a large selection of cars within the American market and the specifications for those cars. These cars are generically split into three groups, SUV, Car, and Miscellaneous.



As can be seen the SUVs have dropped in sales between 2020 to 2021 and there's an increase in sales for both of the other categories. We then drill down a bit more into the data to look more at the subtypes within each car category.

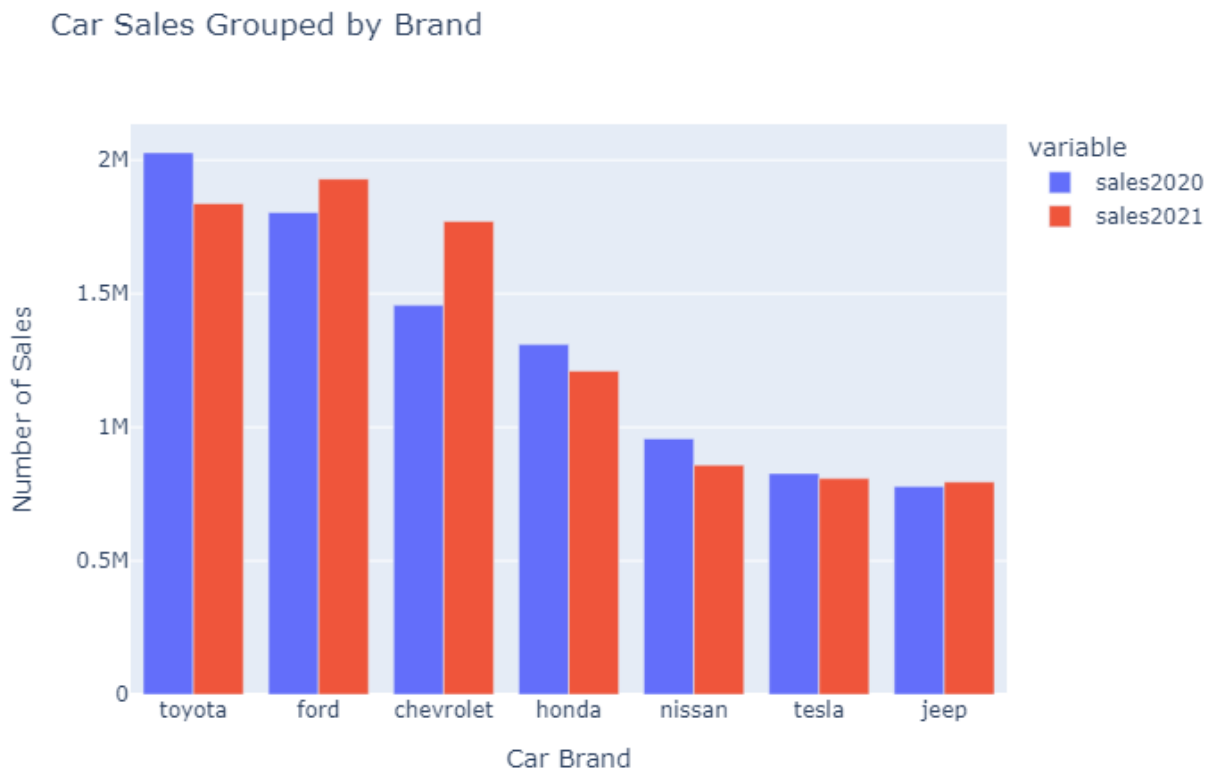


When we drill down into the data, we find that the two groups that increased in sales are large pickups and mid-sized cars. This implies that our pickup trucks lie within the miscellaneous category and obviously the mid-sized cars lie within the cars category. We then drilled down another level to look at the top ten models determined by number of sales.



We find that the Ford F Series and the Chevrolet Silverado have increased in their sales in 2021. These contribute to the gap we see in every graphic that we've drilled down through. It is worth noting that the Chevrolet Equinox has a massive increase in sales going into 2021 and also falls into the SUV category. Yet, despite this large increase in sales, the sale of SUVs overall went down. We now look at the brands to see if there is an interesting change between

2020 and 2021.



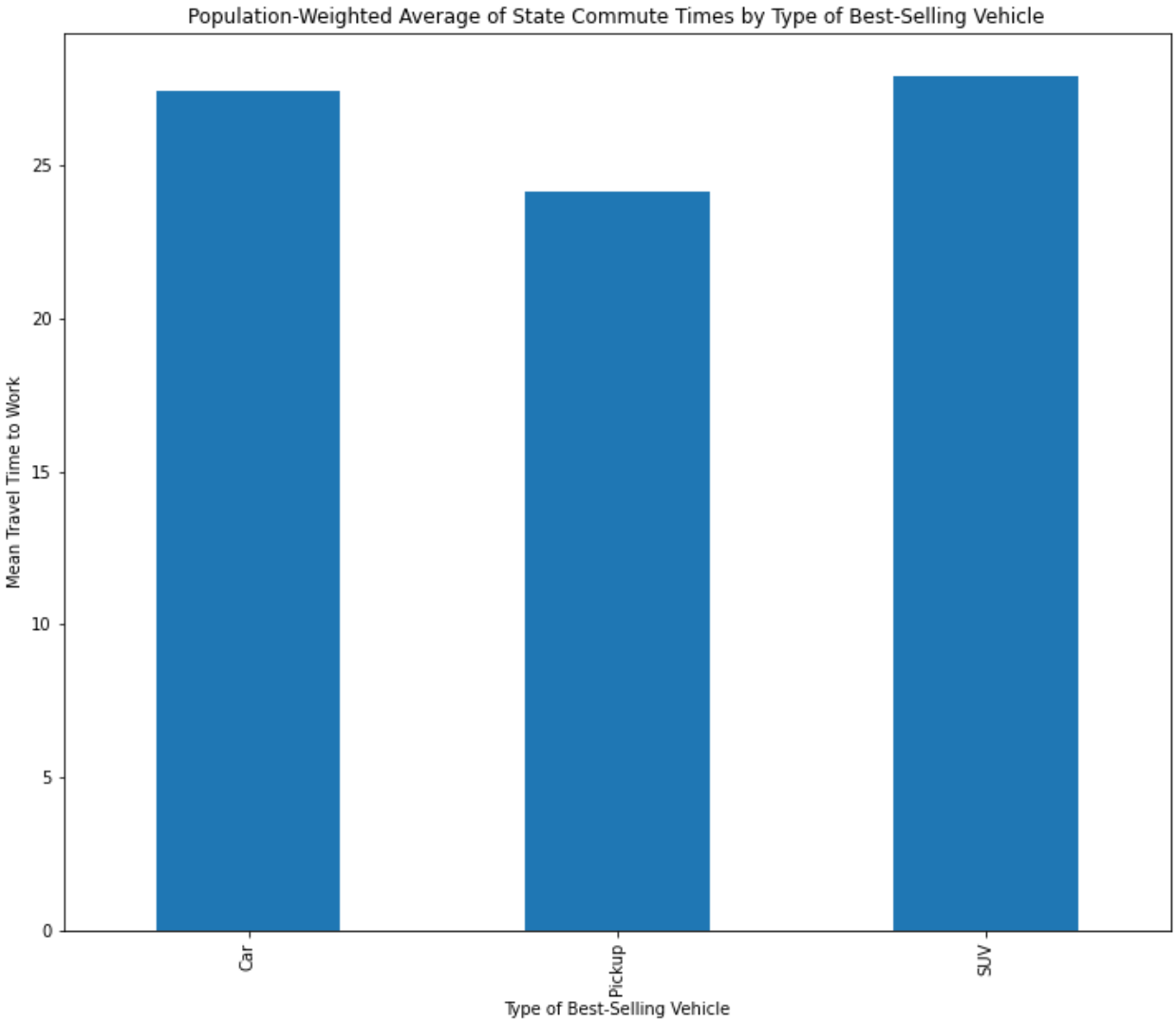
We immediately note that Ford and Chevrolet had the largest increases in sales amongst the top seven brands (by number of sales) for 2020 and 2021. This seems to be a byproduct of the increased sales of pickup trucks in the United States as well as the large increase in sales of the Chevrolet Equinox. Separately, it is shown in the above graphic that of the top seven car brands, all of those that had an increase of sales in 2021 were American brands.

We then look at the various specifications of these cars and calculate the correlation between the specifications and the qualitative functions like brand, and subtype to the number of sales in 2021.

	sales2021
sales2020	0.969572
mpg	0.117514
length	0.155754
height	0.224801
engineDrive_All Wheel Drive	-0.116990
brand_audi	-0.103139
brand_chevrolet	0.121388
brand_ford	0.165703
brand_honda	0.128712
brand_ram	0.185817
brand_toyota	0.141063
subtypeName_compact_suv	0.104792
subtypeName_large_cars	-0.112042
subtypeName_large_pickup	0.505166
subtypeName_upper_class	-0.101461

The above shows the correlation scores that are greater than or equal to the absolute value of one tenth. It's not a surprise to see sales 2020 up there as sales should correlate with sales of the previous year. The most interesting correlation scores show a positive, albeit weak, correlation between miles per gallon, length, and height with the sales in 2021. The strongest correlation is between the large pickup trucks and the sales in 2021. This falls in line with what we saw in the past as pickup trucks have been on the rise for some time. We also see weak positive correlation with several of the brands represented in the top seven brands. In terms of the brands and subtypes, there is nothing particularly surprising outside of the relatively strong correlation between large pickup trucks and 2021 sales. Amongst the things tested with the correlation test, we don't find any correlation for things like engine drive (all wheel drive, front wheel drive, and back wheel drive) and no correlation with things that there probably shouldn't be correlation for such as number of doors or number of seats. For the most part, we also find that of the sixty-seven features checked, only the above have a strength above one tenth and if we get stricter, there's only six features with a strength above .15.

We find evidence of a systematic association between the type of best-selling vehicle in each state and the overall average commute time of workers.



Commuters who live in states where the type of best-selling car is a pickup truck on average have shorter commutes than those living in other states. Those in states where the best-selling car is an SUV, by contrast have longer commutes than average. The overall figures are 24.2 minute for those in pickup states, 27.4 minutes in car states, and 27.9 minutes in SUV states.

We did not perform a statistical test such as ANOVA to assess the statistical significance of these differences, but seeing as there are tens of millions of workers in each of these three regions, they almost certainly are. It is not clear however that commute times exert a causal influence on the type of best-selling car, as there remain numerous demographic and economic differences between these three regions, which we have not accounted for in this analysis.

Machine Learning

To assess the viability of producing monthly sales forecasts with machine learning, we used the historical sales data from the US Bureau of Economic Analysis to fit two different kinds of machine-learning models.

Before we discuss our particular models it is worth mentioning some of the specific difficulties of forecasting. In forecasting generally, as well as in our project in particular, the kind of data used is time-series data, where values are recorded at a series of equally-spaced time steps. Usually, values at a particular time will depend on the values at prior time steps, and this dependence renders many of the statistical theorems underpinning classical methods invalid. Analogues of these can be recovered for series exhibiting a special property called stationarity, where neither the mean nor the noise changes over time, but at the cost of extra variance compared to situations involving independent data.

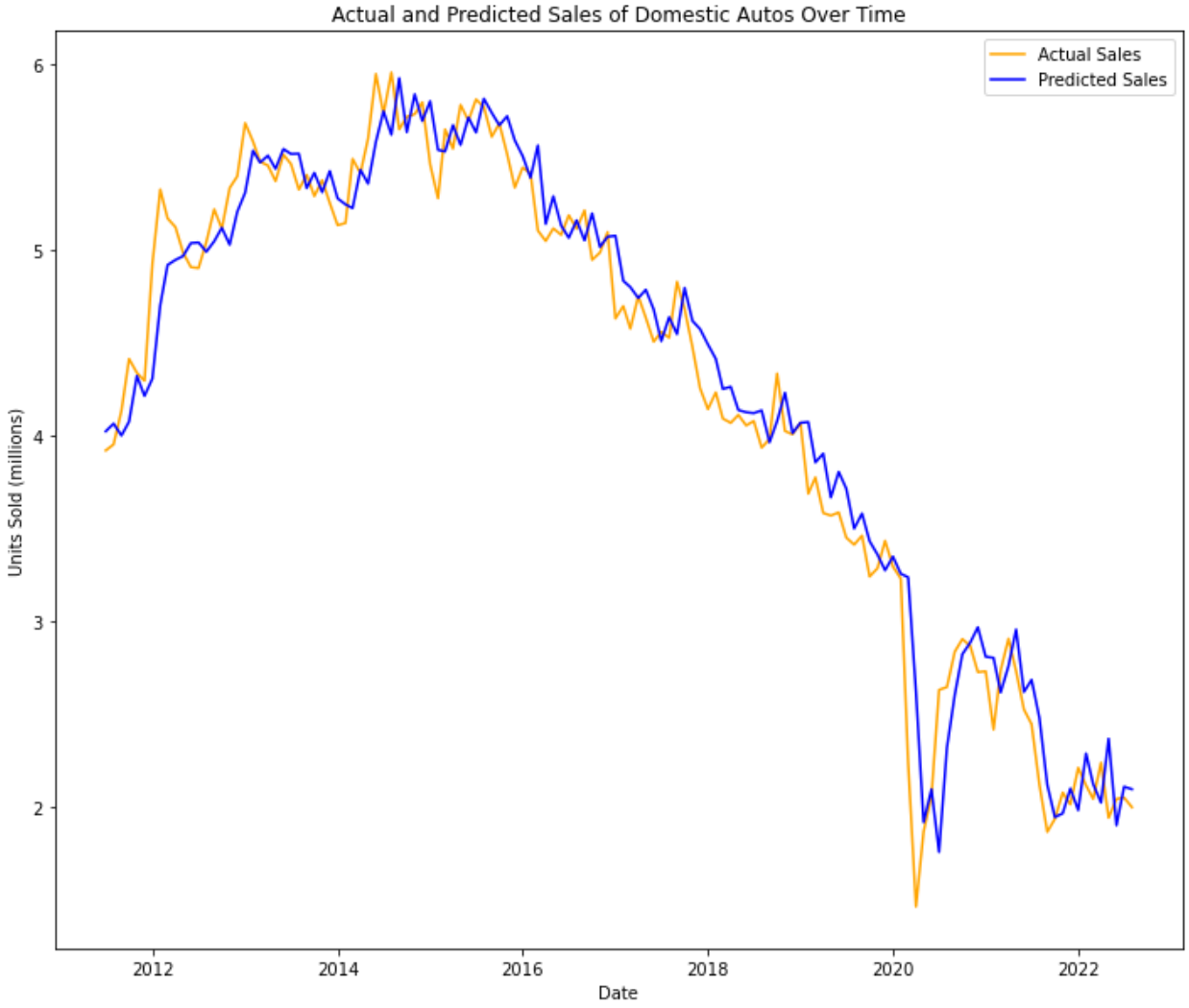
Vector Autoregression Model

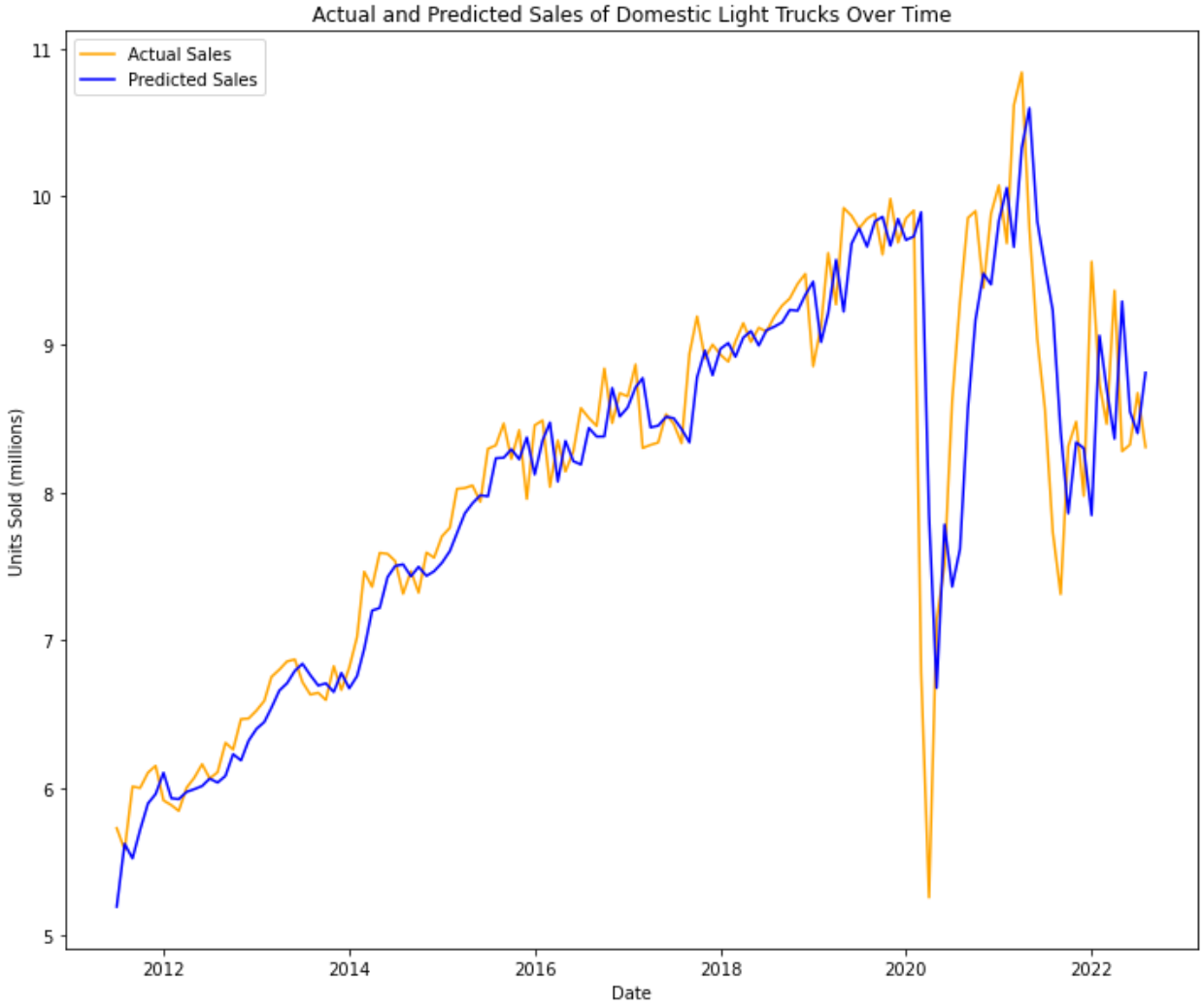
The first model we built was a vector autoregressive model trained on and predicting the values for three series: domestic automobiles, domestic light trucks, and heavy trucks. Autoregressive models predict future values as linear combinations (i.e. weighted averages) of previous values. The model tries to find the values of the coefficients (i.e. the weights) that minimize the mean squared error between the actual and predicted values. Vector autoregression tries to estimate future values of multiple series simultaneously, as linear combinations of not just the past values of each series, but the past values of all of the series available to the model.

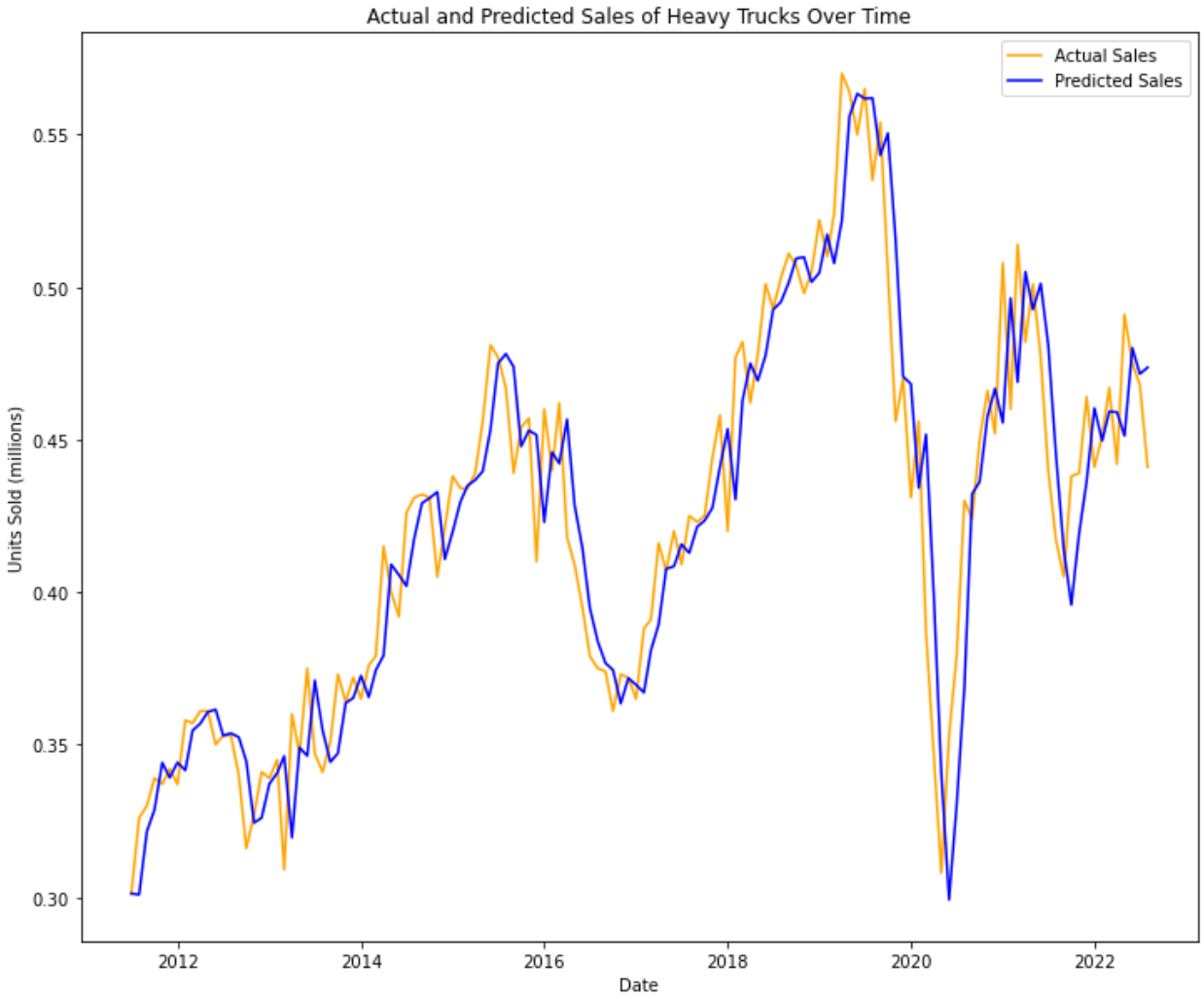
First we split our data into training and test sets, with the former comprising the first 534 observations and the latter the remaining 134. As the series were initially non-stationary we fit the model on the differenced series, where the value in a given month is equal to the number of vehicles sold that month minus the number sold the month prior, as these series of changes are stationary. We undo this transformation after the model has given its predictions. We then determined how many past values to use in the prediction of future values. Many different selection methods exist, but we chose using the Aikake Information Criterion, a metric that balances closeness-of-fit against the number of parameters, obtained on the training data. Based on this we chose a value of four lags. Next we fit our model; the coefficients and training summary can be seen in the *Machine Learning VAR* Jupyter notebook found in the project GitHub repository.

Altogether we have three different equations, one for each of the three series, and each of these has 12 coefficients, four for the past values of the series and the remaining eight the past values of the other series. However many of these cross-series coefficients do not achieve statistical significance, indicating that sales in one category affect only weakly if at all sales in the other two.

Finally, we used the model we just fit to predict values on the test set. We used a one-step ahead prediction framework, where the model was allowed to see all of the previous months' values as well as the current one's values before predicting the next month, and then it would be given the actual value before predicting the month 2 steps ahead and so on. This framework is standard in evaluating time series models. We give plots of actual vs. predicted values for all three series below:


Actual and Predicted Sales of Domestic Autos Over Time
Actual Sales
Predicted Sales
Units Sold (millions)
6
5
4
3
2
2012
2014
2016
2018
2020
2022
Date


Actual and Predicted Sales of Domestic Light Trucks Over Time
Actual Sales
Predicted Sales
Units Sold (millions)
Date
11
10
9
8
7
6
5
2012
2014
2016
2018
2020
2022

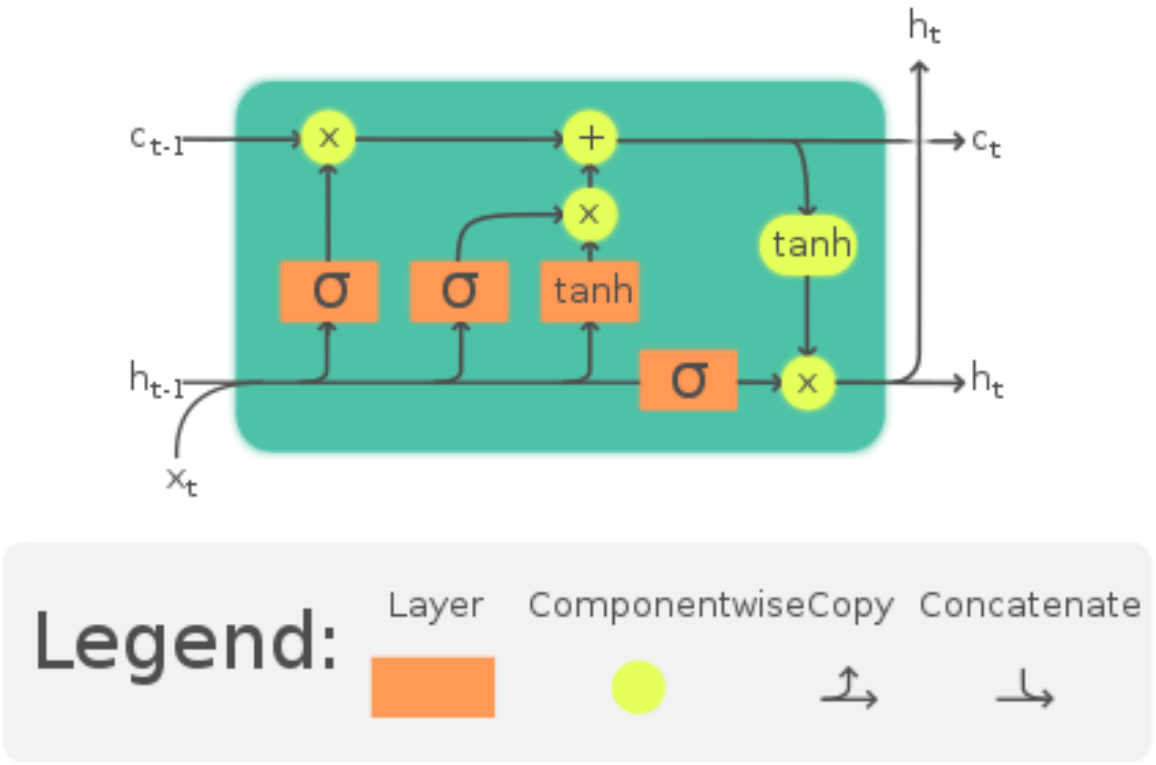


For all three series we see a fairly close correspondence between the predicted values in blue and actual values in orange, although the predicted values do tend to lag behind the actual values. We assess the prediction error quantitatively by examining a metric known as the mean absolute percentage error, or MAPE, which is the average of the absolute value of the percentage difference between the predicted and actual value. The MAPE for both domestic light trucks and heavy trucks was 4.2%, while for domestic automobiles it was 5.2%, yielding predictions that are usable but imperfect.

Neural Networks with LSTM and GRU

The second model goes into the realm of deep learning. Deep learning is a subset of machine learning that utilizes neural networks to “learn” from data. A neural network is a set of algorithms that allow data to be learned in a similar way to how the brain works. As in, data is passed through layers of “neurons” that apply transformations to it such that by the end, the network will have created an output that it “learned” to make from the input. One type of neural network that is particularly useful in our case is the “Recurrent Neural Network”. The “Recurrent Neural Network”, or “RNN” for short, allows for the outputs of a neural layer to be reused in some form as the inputs of that same layer. This is useful for our purposes since our data is represented as a time series, which, as discussed above, can have output of one timestep that depends on the

input of previous timesteps. In this case, an RNN can take the output of one timestep, then feed it into the inputs of other timesteps. In particular, a useful RNN for time series is called “Long Short Term Memory” or “LSTM” for short. This type of RNN not only has transformations for input and output, but also has a transformation that simulates “forgetting”. A variant of this is called a “Gated Recurrent Unit” or “GRU” for short, which has an “update” and “reset” gate. This has one less parameter than the LSTM neural network, which reduces the complexity of models using it.



The structure of an LSTM cell. Source: https://en.wikipedia.org/wiki/Long_short-term_memory

Similar to the VAR and ARIMA models, we use LSTM to predict Domestic Auto Sales. Rather than having to build the neural network completely from scratch, we use the keras API in the tensorflow library. Of note is that unlike the VAR and ARIMA models, the model requires no assumptions to be trained since it can learn complex patterns. However, we apply some preprocessing on the data so that the model might perform better. In particular, we standardize it so that each value is represented as the number of standard deviations from the mean of each column.

We then apply a model to the data consisting of an input layer, an LSTM layer with 64 neurons, a Dense layer with 8 neurons with relu activation, and a Dense output layer with linear activation. After 500 epochs, the mean squared error ends at 0.5, but during training, it hovered between 0.3 and 0.5.

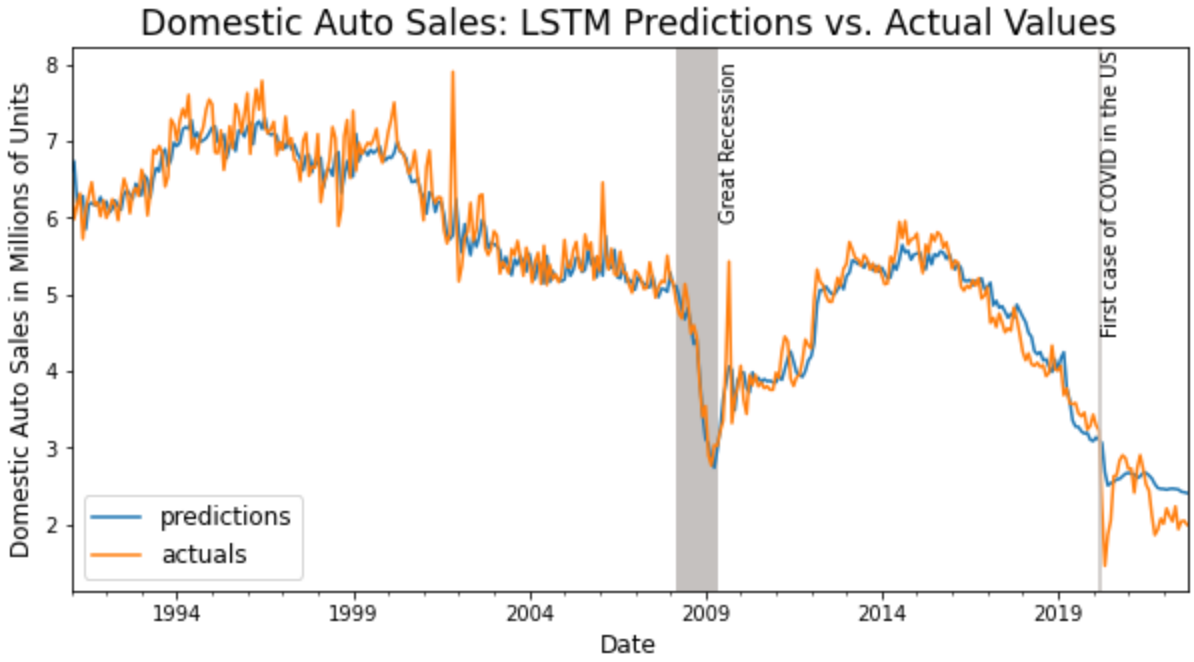
From this model, we tune the model to optimize its performance. In the case of a neural network, the hyperparameters we can tune in order to optimize its performance are the number of epochs, the learning rate, the input window size (analogous to the number of lags in the time series model), and the activation functions used. However, we can also edit the structure of the neural network itself by adding and removing layers.

We also apply feature engineering on the data, so that features not highly correlated with Domestic Auto Sales are not considered in the model.

	Correlation with Domestic Autos
Domestic Light Trucks	-0.302
Foreign Autos	0.382
Foreign Light Trucks	-0.706
Heavy Trucks	-0.048
Gas Price	-0.656
GREATRECESSION	-0.242
COVID	-0.591

The correlation matrix with Domestic Autos. Because of their weak correlations, Domestic Light Trucks, Foreign Autos, and Heavy Trucks are removed from the model.

After a lot of tweaking, we found that the optimal model (or at least the most optimized one we could find) contains an input layer, an LSTM layer with 64 neurons, another LSTM layer with 32 neurons, a Dense layer with 8 neurons with ReLU activation, and another Dense layer with linear activation. The number of epochs was 500 and the learning rate was 0.001. During training, the validation error reached a minimum of 0.09 and the training error hovered around 0.09 as well, suggesting no overfitting issues. When plotted against the actual values, the model fits the data rather nicely:



LSTM predictions plotted against actual values.

However, inspecting the graph further, we notice that after the Great Recession, the model starts to get away from the actual values a little bit until finally it falls apart completely in the months following COVID. A simple explanation is that there are not enough data points after COVID for the model to develop resilience to the volatility of COVID recovery. Another possibility worth exploring is that the periods after the Great Recession and after COVID mark permanent economic changes that the model's features by themselves are simply not sufficient to capture anymore. Whatever the case may be, this makes predictions on the future rather difficult.

Conclusion

We have achieved several of our group goals in this project: we detected a marked shift away from automobiles towards the light truck, now by far the dominant vehicle category in the US; we identified several features correlated with strong sales in the present market; and we built machine-learning models capable of producing reasonably accurate forecasts of monthly-vehicle sales. However, we feel as though we have left certain questions unanswered. We have not discerned, for example, which factors might explain the rise of the light truck. Nor were we able to systematize our analysis of the present market into a statistical or machine-learning model.

The difficult truth is that our data cannot provide the answers to these questions. Data-acquisition is often an arduous and expensive process, and our case proves no exception. Much of the data we envisioned being able to find when we first formulated this project could not actually be found, at least not freely. We wanted, for instance, to examine the relationship between vehicle sales and many other economic factors besides the price of gasoline. The particular constraints of nearly all time-series models would force these to also be month-level

data. For the most part we could not find the data at this level of precision; where we could it was not nearly historically extensive enough to conduct meaningful analyses. Much of the data we did collect, limited as it was, already required intricate and laborious web-scraping to do so.

Considering our results in light of these limitations, we are optimistic that with higher-quality and more specific data our same approach and methods would generate deeper and more insightful conclusions about the automotive industry, both past and present, as well as produce even more accurate machine-learning models. There are a number of firms who collect and sell the kind of data we were looking for, both automotive and economic. To anyone seriously interested in rigorously understanding or forecasting the automotive we recommend, if they have the institutional resources available to them, the purchase of such data. Otherwise we hope you come away, as we have, with a deeper understanding of the automotive industry.

Sources

U.S. Bureau of Economic Analysis, Motor Vehicle Retail Sales: Domestic Autos [DAUTOSAAR], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/DAUTOSAAR, September 22, 2022.

U.S. Bureau of Economic Analysis, Motor Vehicle Retail Sales: Domestic Light Weight Trucks [DLTRUCKSSAAR], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/DLTRUCKSSAAR, September 22, 2022.

U.S. Bureau of Economic Analysis, Motor Vehicle Retail Sales: Foreign Autos [FAUTOSAAR], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/FAUTOSAAR, September 22, 2022.

U.S. Bureau of Economic Analysis, Motor Vehicle Retail Sales: Foreign Light Weight Trucks [FLTRUCKSSAAR], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/FLTRUCKSSAAR, September 22, 2022.

U.S. Bureau of Economic Analysis, Motor Vehicle Retail Sales: Heavy Weight Trucks [HTRUCKSSAAR], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/HTRUCKSSAAR, September 21, 2022.

United States Census Bureau. American Community Survey: S0802: MEANS OF TRANSPORTATION TO WORK BY SELECTED CHARACTERISTICS, September 21, 2022.

United States Census Bureau. American Community Survey: S0501: SELECTED CHARACTERISTICS OF THE NATIVE AND FOREIGN-BORN POPULATIONS, September 21, 2022.

Cars.com. https://www.cars.com/research/, September 21, 2022

Carsalesbase.com. carsalesbase.com, September 21, 2022

Edmunds's. Most Popular Cars in America https://www.edmunds.com/most-popular-cars/, September 21, 2022