Predicting the Influence of Local Gas Price on Preferred Vehicle Fuel Efficiency

Team 51: Andrew Getz, Kathy Mirzaei, Matthew Rand, Stefano Romano

Background Information

The United States has had for decades some of the lowest gas prices among developed nations worldwide[4]. In addition, due to the high percentage of the population living in rural areas and urban planning design choices, the United States has the sixth most vehicles per capita in the world and the second-largest vehicle market in the world (Hedges & Company). Because of these factors, many American consumers' impact on global greenhouse gas emissions stems directly from their vehicle usage.

Problem Statement

Our problem statements are as follows:

- To analyze the relationship between gas prices in the United States and the makes, models, types, and fuel economies of cars purchased by Americans.
- To investigate how gas price changes affect car purchasing patterns over time.

Research Questions & Hypotheses

Do Americans purchase more fuel-efficient vehicles when gas prices are higher? If so, how much more fuel-efficient? We hypothesize that Americans would purchase more fuel-efficient cars as gas prices rise, but we are unsure how much more fuel efficient.

Literature Review

The number of cars in the world is increasing. Hedges & Company's report estimates that there will be over 1.4 billion cars in the world by 2023, and the passenger car market in the United States is expected to continue to grow in the future[3]. Transportation is a significant contributor to greenhouse gas emissions and climate change. According to the Environmental Protection Agency (EPA), transportation is the largest source of greenhouse gas emissions in the United States, and carbon pollution from transportation has been increasing steadily over the past few decades[2].

One factor that can impact consumer behavior and the demand for fuel-efficient vehicles is the gas price. A Gallup poll conducted in 2011 found that high gas prices can motivate consumers to consider fuel-efficient cars when making car purchases[1]. Academic research also supports the idea that gasoline prices can impact consumer demand for fuel-efficient vehicles. A study by Langer and Miller (2008) found that gas prices significantly affect consumer preferences for fuel-efficient cars, while the impact of vehicle prices is less pronounced[5].

Impact

Key points of strategic value:

- The United States market had almost 300 million vehicles traveling on American roads in 2022.
- The US vehicle market is massive, with over $500 billion in vehicles sold annually (Statista).
- 27% of US greenhouse gas emissions come from cars (EPA).

If long-term trends for gas prices hold, gas prices will continue to rise, potentially changing consumers' demand by vehicle type and fuel efficiency.

Auto manufacturers need to be able to predict this change in demand years in advance to design and develop more fuel-efficient vehicles now in anticipation of future demand shifts.

The vehicle market is more competitive than ever. Developing a new vehicle costs between $1 - $6B (Shea), with an average profit margin of ~2.63% (Macrotrends). Making accurate predictions about what drives customer purchasing behaviors is vital to ensure new vehicles brought to the market include the right features.

Exploratory Analysis & Preparation

Exploratory Analysis

Initially, our goal was to understand our data quickly. We built an exploratory analysis using the DataExplorer package across each dataset to display: basic statistics, data structure, missing data profiles, univariate distributions, correlation analysis & principle component analysis. This analysis built our understanding of what cleaning was going to be necessary, outlined what analysis we could perform, and centered our discussions.

Data Cleaning

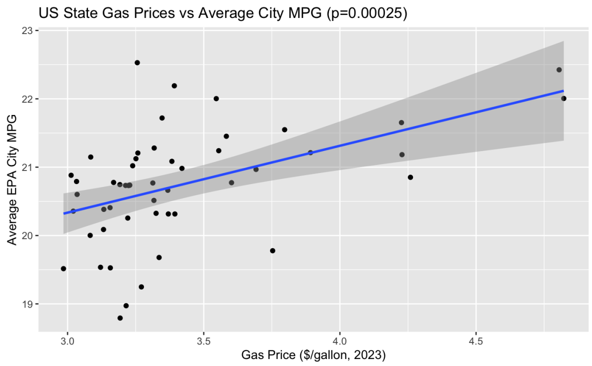
The first challenge to overcome was how to clean the datasets sensibly to reduce the noise and increase the signal. The vehicle-to-mpg and gas price datasets were mostly clean. The vehicle sale dataset was the most unclean. The main task was to trim down the features only to pull in what was necessary for our analysis and eliminate false or confounding information.

Data Joining

One of the main challenges we encountered in the project was joining our three datasets from different sources. For the project, joining the fuel economy dataset to the Craigslist used car dataset via the car's year, make, and model was necessary to have the MPG for each car sold. We also joined the Craigslist used car dataset to the gas prices dataset via state and year. This second join is fairly straightforward; however, for the other join, the Craigslist car dataset has freeform text in the year, make, and model columns for each car. Because of this, Craigslist users can enter whatever data they want, making this data very difficult to join to the fuel economy dataset. We created a fuzzy matching algorithm to overcome this issue, which correctly cleans most of the car years, makes, and models in the Craigslist dataset, allowing us to merge the fuel economy dataset and continue with our project.

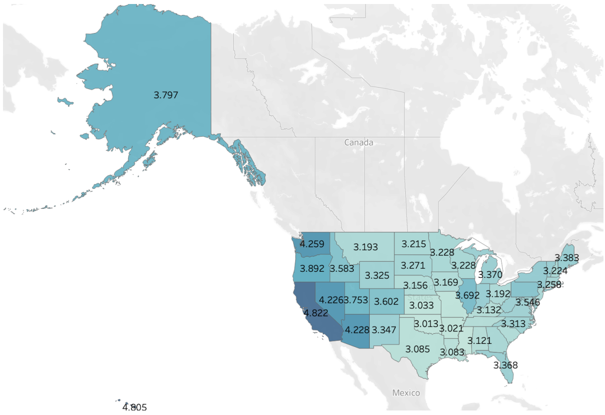
Initial Modeling

Next, we wanted to test our initial base hypothesis to see if there is any relationship between fuel prices that differ by state and the average fuel efficiency of cars purchased in those states, with no other factors considered to start. Using simple linear regression, we regressed average city fuel efficiency on gas price and found a strong positive linear relationship between the two. The independent variable coefficient of 0.97 +/- 0.25 (gas price per gallon) can be interpreted as for every $1 increase per gallon in gas price, the average car in the same state will have a higher city MPG rating by 0.97. Based on the R-squared value, gas prices explain 24% of the variance in average city MPG differences between states. For the full statistical output, please refer to the Appendix section. This strong positive linear relationship is also statistically significant, as indicated by the low p-value (p=0.00025).

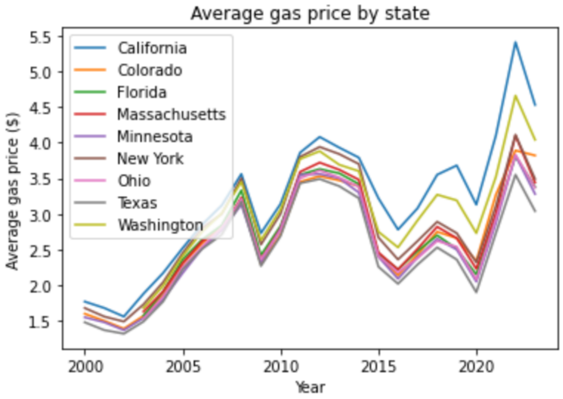


Historical Trends in Gas Prices and Fuel Efficiency

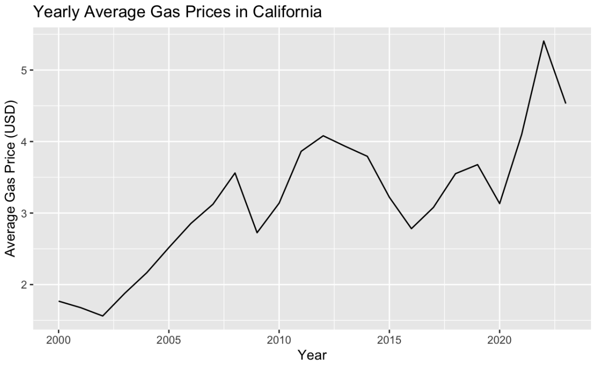
First, we wanted to understand how gas prices have changed in recent years and whether those differences appear to be influenced by their region. It is important to understand that different portions of the country have different consumer preferences and different exogenous market characteristics that may be difficult to capture in the data. Therefore, establishing whether different regions behave differently allows us to determine whether national trends are applicable in each locality.



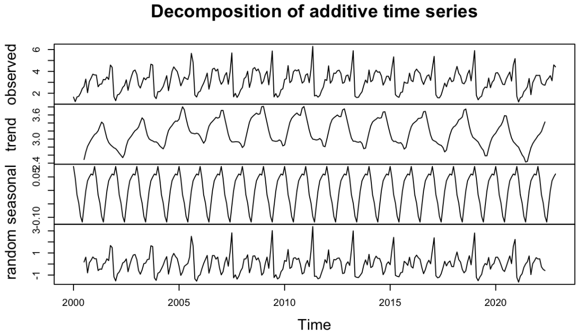
As expected, average gas prices tend to vary across states. The plot below shows that the average gas price by state varies, but they generally follow the same trend. These trends correspond with different economic, political, or environmental events that must be accounted for in our analysis by decomposing the gas price time series over time.



California gas prices are the highest among all the other states. As we can see from the graph below, the trend in the gas price has been upward in California, with the year 2022 showing the highest price of all time.



We then used time-series analysis to understand the underlying patterns and trends in the gas price to make predictions of gas value. In this graph, we display the monthly decomposed gas prices in California per year.

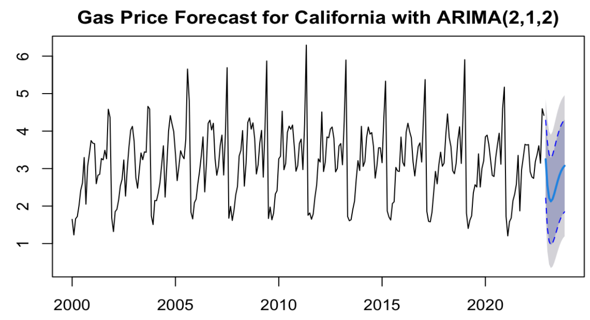


It appears that the seasonal component for each month is relatively constant across all years(2000 to 2022), with a slight variation from year to year; it suggests that the time series has a strong, consistent seasonal pattern over time.

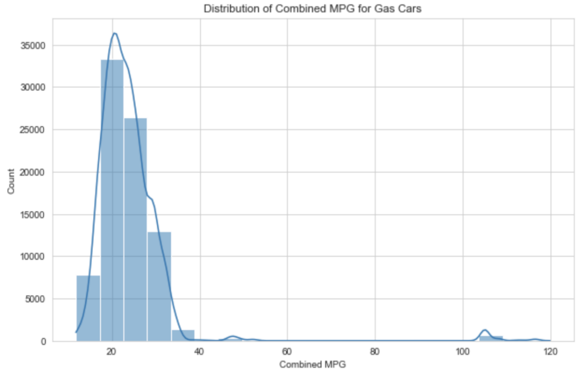
There are some fluctuations in the trend of the data, which shows the general direction of gas prices over time. However, the prevailing gas price has been trending upward over the past years. It's worth noting the other factors that affect gas prices, such as changes in supply and demand, geopolitical conditions, or environmental events.
Nevertheless, the trend in the data suggests that gas prices have generally increased over the years.

We also used Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA) models to perform a time series analysis on gas prices. After experimenting with different hyperparameter values, we improved the performance of our model with ARIMA(2,1,2) and its RMSE of 0.8078608. Therefore, this output suggests that the ARIMA(2,1,2) model performs better than the previous ARIMA(1,1,1) and SARIMA(1,1,1) models, with a more accurate forecast.

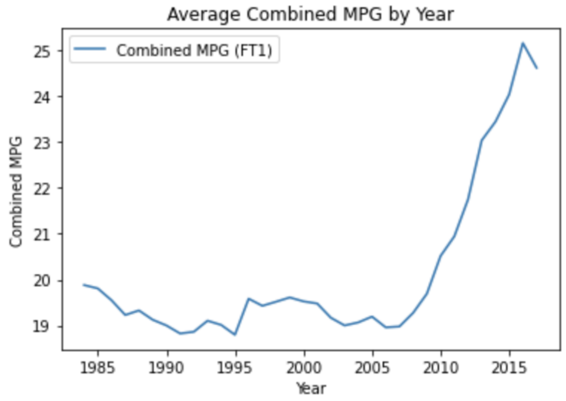
For our analysis, we can contend that trends in gas prices are likely to stay consistent, and therefore their influence on purchasing decisions should follow historical patterns.



We also looked at its distribution to gain some insights into the combined MPG. We can see that there are a few extremely high values. We assume that these values represent electric vehicles, which have been given an effective MPG estimate incorporated into the dataset. This assumption is reasonable given the sharp divide in the dataset: traditional internal combustion engine vehicles follow a normal distribution, while hybrid and electric vehicles have much higher values.



We have further looked into the MPG data to see if we can identify any trends by looking at the years. From the graph, we see a change in trend between 2005-2010 in the combined MPG.



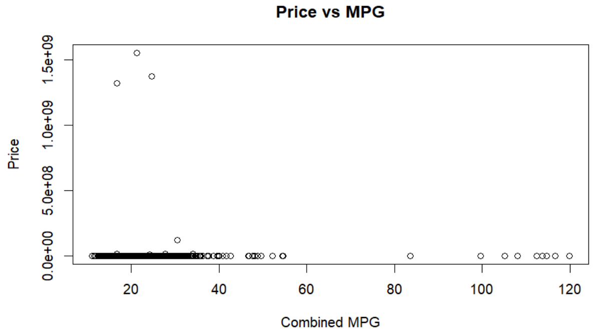
This change in trend further suggests that our analysis may need to control for electric and hybrid vehicles in the dataset. It is unlikely that fuel efficiency increased so dramatically in 2008, and this is more likely due to the addition of high MPG vehicles to the dataset.

This analysis of gasoline prices and average fuel economy by state shows some interesting trends. First, it's clear to see the local influences of gas prices across the country. We will have to account for the location of the vehicle sale when determining whether the price of gasoline influences the price of the vehicle at the time of sale. However, we must determine whether this connection impacts aggregate vehicle sales.

Vehicle Pricing and Consumer Purchasing Patterns

We also wanted to understand what factors and features have the greatest influence on the price of a vehicle. This feature/factor prioritization is critical to understanding consumer considerations when purchasing a vehicle. Using the Craigslist sales dataset, we developed a linear regression model to see whether we could predict the vehicle's sales price.

It was important for us to narrow the focus of our analysis. When determining the price that a consumer is willing to pay for a vehicle, it is important to consider what factors are most important in determining the price. We began by plotting the price against combined MPG to see if we could observe any trends or relationships.

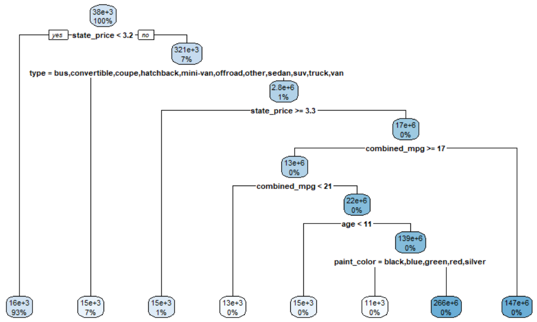


We could not see any trends visually, and we found no relationship when we regressed price solely on combined MPG. However, it's important to note that we did not control for other influential factors; we simply considered price and MPG alone. This comparison is not truly accurate, and we should only consider price versus MPG in the context of other confounding variables.

The dataset also had a few outliers, for example, high-value cars and those with high combined MPG. According to our research, the EPA states that the highest combined MPG in the market for an internal combustion engine is 59 MPG. So, any vehicle with a combined MPG greater than 59 should be excluded. Additionally, since we are looking at whether the price of gas influences purchasing decisions, we have to make the assumption that it will only influence those who care about the price of gasoline. We contend that consumers who can purchase high-value vehicles may not care about the price of gasoline, and therefore we eliminate all vehicles valued over $100,000. We assume that more budget-conscious customers may be more likely to factor gas prices into their purchase decision, which may be more influential in a purchasing decision.

Additionally, we needed to develop a few more variables for the model. First, the year the manufacturer made the car is less influential than the car's age, which was computed by subtracting the manufacturing year from the sale date. The combined MPG was computed using the standard formula, 55% city MPG + 45% highway MPG. All other variables were directly from the manufacturer's data for the vehicle.

Using a decision tree, we wanted to isolate which factors explained the most variance in the model. We found that factors such as the local price of gasoline (state_price), type of vehicle, and combined_MPG are all significant factors to consider. However, this analysis purposefully excluded the make & model of the car, choosing instead only to examine continuous variables. This model only reduced the relative error by 17%, indicating that it did a relatively poor job subsetting the data into logical price prediction groups. This poor performance also means that our chosen factors have very little explanatory power.

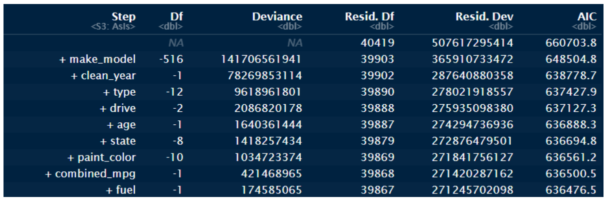


Note*: Larger version is available in the Appendix.*

Next, we began by looking at linear regression to see if we could predict vehicle price. We began with the subset of our data and tried to find a combination of factors that yielded the best result. We used several different parameters in the model. The make & model of the vehicle had the most explanatory power over the sale price. In contrast, factors such as vehicle characteristics (type, drive, paint), location, fuel efficiency, and gas price had much less influence on the model.

We found a few values of note: combined MPG tended to have an additive influence on the sale price, which is interesting, given that more expensive cars tend to have lower MPG. This additive influence suggests that for budget-conscious consumers, combined MPG is a factor they are willing to pay more for. Additionally, we found that the local price of gasoline had a subtractive relationship, meaning that for each additional dollar gas prices went up, the price of a vehicle would fall. All of these factors were significant at a 95% confidence interval.

Before concluding which factors were influential and which were not, we ran a forward stepwise regression to determine which variables best-reduced deviance in the data set. We found that while combined_MPG is included, fuel price does not, which seems to indicate that local fuel prices do not influence the sale price of a vehicle since it does not significantly influence the reduction in residual deviance.



Step <S3: AsIs>	Df <dbl>	Deviance <dbl>	Resid. Df <dbl>	Resid. Dev <dbl>	AIC <dbl>
	NA	NA	40419	507617295414	660703.8
+ make_model	-516	141706561941	39903	365910733472	648504.8
+ clean_year	-1	78269853114	39902	287640880358	638778.7
+ type	-12	9618961801	39890	278021918557	637427.9
+ drive	-2	2086820178	39888	275935098380	637127.3
+ age	-1	1640361444	39887	274294736936	636888.3
+ state	-8	1418257434	39879	272876479501	636694.8
+ paint_color	-10	1034723374	39869	271841756127	636561.2
+ combined_mpg	-1	421468965	39868	271420287162	636500.5
+ fuel	-1	174585065	39867	271245702098	636476.5

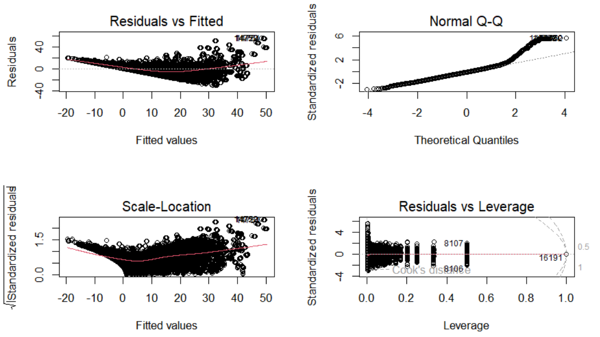
After determining which factors were likely to influence the price of a vehicle, we decided to look at the problem differently. If we flip the problem on its head and ask, "Do gas prices and MPG impact the volume of car sales" we can use the vehicle price as a representative variable for those intangibles. The vehicle's sale price indicates several factors not included in our model: features on the car, the personal attachment the seller has, etc. Looking at it this way allows us to capture that information in our modeling effort.

For this analysis to be meaningful, we need to consider a study where we have enough data on the vehicles we are analyzing to ensure a rigorous analysis. As such, we will begin by eliminating vehicles from the data set with too few data points. A threshold of 30 vehicle sales by make & model was selected as our threshold.

Then, we will build a regression model to predict the number of vehicles sold based on the make_model, time, state, local gas price, vehicle sale price, and the average combined MPG.

We find that location, price, local price of gasoline, and average combined MPG are all significant at the 95% confidence level in determining the sale volume of a vehicle. The make and model could be influential, but many of the variables could have approximated to 0 based on the confidence bands. Our adjusted R^2 = 0.5919, which means the model explains roughly 60% of the variance in the vehicle's sale volume.

However, we found that our dataset still exhibited significant biases, as evidenced by its non-normality and the distribution of its residuals. To improve our model, we wanted to see whether a transformation of our data would improve performance. To test this, we ran a combination of linear and log transforms on price, sale volume, and combined MPG to see if a transformation would improve fit. The transformations did little to improve model performance, with no model outperforming the linear-linear model.

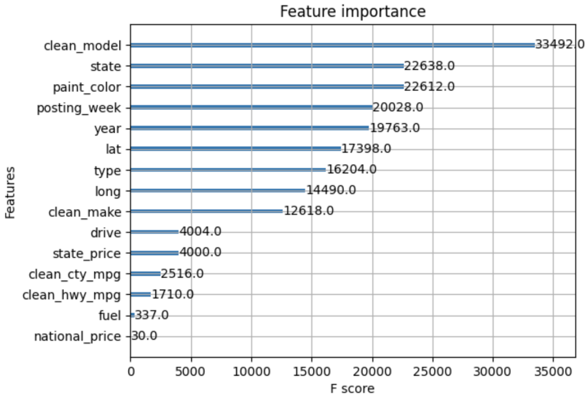


Vehicle Pricing Gradient Boosting and Feature Importances

To achieve higher accuracy, we also tried to predict the car sales price using a gradient boosting model, a more advanced machine learning model. We used the same cleaned Craigslist dataset, featuring 189,000 car sales and 16 independent features for each car, including the year, make, model, and fuel efficiency. We split our data into train and test datasets, then ran the data through XGBoost's gradient boosting regression model. XGBoost, and specifically Python's implementation of XGBoost, were selected because, in newer versions of the package, XGBoost can handle categorical variables natively, which is extremely important for this project due to the categorical nature of most of our independent features. It is also important to note that the gradient boosting algorithm has many benefits versus a traditional multiple linear regression model. It can make more accurate predictions, including that it can handle interaction effects between independent variables and nonlinear effects between the independent and dependent variables without additional feature engineering.

We decided it was important to perform hyperparameter tuning for our gradient boosting model because suboptimal hyperparameter selection with gradient boosting models can lead to poorly fit (underfit/overfit) models. From experience, we know that the most important hyperparameters to tune are n_iterations, the number of iterations for the gradient boosting algorithm to perform, learning_rate, the rate at which the gradient boosting algorithm moves its predictions towards the correct training dependent variable labels, as well as max_depth, the maximum depth of each tree created during each iteration of the gradient boosting algorithm. We selected values of [100, 250, 500, 1000] for n_iterations, [1, 2, 3, 5, 8] for max_depth, and [0.01, 0.05, 0.10] for learning rate both from experience as well as based on research of best practices from online resources. We also decided to use MAPE, or mean absolute percentage error, for our desired metric, as car prices are on a logarithmic scale, so traditional linear metrics such as RMSE (root mean square error) would distort and penalize errors specifically on extremely highly-priced cars.

We performed a grid search over these hyperparameter values and found that the following hyperparameter values minimized the hold-out test set MAPE, giving us the most performant model: n_iterations = 1000, max_depth = 8, learning_rate = 0.10. While our train MAPE was 0.117, our test MAPE was higher at 0.340, meaning the model is likely slightly overfit. Future work could be performed to examine alternate hyperparameter values and alternate hyperparameters to tune to improve the fit of this model.



We show above the feature importance chart for our most performant gradient boosting model. We see that the fuel type of the car, as well as both the city MPG and highway MPG are among the least informative features in our dataset. This would suggest that American consumers are indifferent to fuel efficiency when it comes to the cars they purchase and how much they are willing to pay for them. One possible concern with this analysis could be that MPG is collinear with the model of the car, so the feature importance of MPG could be understated in the above plot.

Conclusions

From the results of the analyses, it's ambiguous to see the exact impact of fuel efficiency and gas prices on the sale price of any given vehicle. Based on the analyses performed, fuel efficiency and the local gasoline price influenced the vehicle's sale price, but not very much. The linear regression models showed that local gasoline prices had a slight deflationary effect, lowering prices by roughly $685 per additional dollar in gas prices. However, it is atypical for gasoline prices to move in full dollar increments, meaning that the price of gasoline is likely negligible in the grand scheme of a purchasing decision. Similarly, the linear regression analysis showed that a vehicle's fuel economy tended to have a positive increase in the price of roughly +$84/MPG, showing consumers valued higher efficiency. Still, the effect was relatively small when considering the average sale price of the vehicle was $14,200.

When considering the effects of the local price of gasoline and fuel efficiency on the volume of vehicles sold, we saw a much stronger correlation. High gas prices tend to decrease the volume of vehicles sold, whereas higher average MPG tends to increase the sales volume. This trend shows that, when controlling for make and model, the sales volume of a particular vehicle depends on the local gasoline price. When gasoline prices are elevated, consumers tend not to purchase as many low-efficiency vehicles as when prices are low.

When considering the impact of this work, it's important to remember that the projected change in gasoline prices followed extremely predictable norms and that there was no indication to show that the established pattern would break. As such, we can conclude that vehicle sales will follow historical patterns and not be influenced by changes in the gasoline market so long as the market maintains historical trends. Thus, the approximations hold true, assuming price actions consistent with historical norms.

Limitations

This work was limited by the data available for analysis. One of the key limitations of this work is that the date range for sales from Craigslist only held a few discrete values, which prevents better-trending approximations over time. The sales data only spanned three years, thus limiting the observed impact of changes in gasoline prices or vehicle efficiency over time.

When aggregating the dataset, it became extremely limiting to run analyses because of the sheer number of make and model combinations. After controlling for all appreciable factors and subsetting the data by those vehicles who had sold more than 30 cars amongst the dates of interest, roughly 5,000 make and model combinations remained for roughly 40,000 data points. This highlights a critical problem with the data and the extrapolatory nature of the analysis.
It is difficult to contend that this analysis holds true for all vehicles when many may not have representative samples large enough to influence the model. For this reason, we must temper our model's output with the expectation that it may not hold true for all make & model combinations.

Lastly, because of the sheer volume of categorical variables, it was difficult to run train/test splits on the data for our linear regressions because it was difficult to ensure that the training set and test set would have sufficient overlap in categorical variable values. Additionally, the Craigslist data set is all second-hand sales; the sale price is not standardized and not influenced by manufacturer recommendations directly, which adds additional variability to the model that may be difficult to capture. Because prices are set based on the seller's willingness to part with the goods rather than the actual book value, the consistency across prices in the data set is inconsistent, leading to a less accurate result.

Future Work

If this work were to be expanded in the future, the key to being able to provide a more thorough and conclusive analysis would include a considerable amount of additional accurate data, some information which may or may not be attainable:

- Vehicle information regarding the vehicle itself, including but not limited to: features, accident history, mileage, or anything else that may correlate to its price on the market.
- Sale information regarding the sales of larger samples of vehicles across the market, including new, used, leased, and rented vehicles.
- Fuel price information at the zip code level going back at least 25 years.

Including the additional information would require an increased amount of computational resources to analyze it and an increased cost. This additional data regarding the vehicle itself would help to explain more of the variation in car price to understand better where MPG ranks in the features that explain the price paid for the use of the vehicle (purchased/rented). With finer-grain gas prices, we will be able to focus more on economic events (recessions, booms/busts) that had an effect on gas prices in the economy and look into consumer purchasing actions/behaviors before and after to evaluate the change, if any.

In the US, there is also the confounding aspect of market behavior around expected fuel price. For example, even if there is a significant increase in fuel prices due to an economic event, fuel prices will return to an expected value in line with historical prices adjusted for inflation. If that is no longer the case, for example, if a tax is put in place that fuel prices will increase each year by $0.50, we could find that the impact of fuel prices on car sales becomes much more significant than what we have seen in the above analyses. Consumers likely purchase cars based on their future expectations of fuel prices over their car's life rather than today's fuel price. As a result, it is likely that the market typically cares less about the MPG of their vehicles for most fuel price changes. If there were to be a significant shift in current or future expected fuel prices, then MPG may quickly become a price driver.

Works Cited

Acobe, Dennis. “In U.S., High Gas Prices May Make Many Get Fuel-Efficient Cars.” Gallup (2011). https://news.gallup.com/poll/147746/high-gas-prices-may-fuel-efficient-cars.aspx.

Environmental Protection Agency. (2022, May 19). Carbon Pollution from Transportation. Retrieved from https://www.epa.gov/transportation-air-pollution-and-climate-change/carbon-pollution-transportation.

Hedges & Company. "How Many Cars are There in the World in 2023?" (2023). https://hedgescompany.com/blog/2021/06/how-many-cars-are-there-in-the-world/.

International Energy Agency. (2021). Global Energy Review 2021: Assessing the effects of economic recovery on energy and emissions. Retrieved from https://www.iea.org/reports/global-energy-review-2021

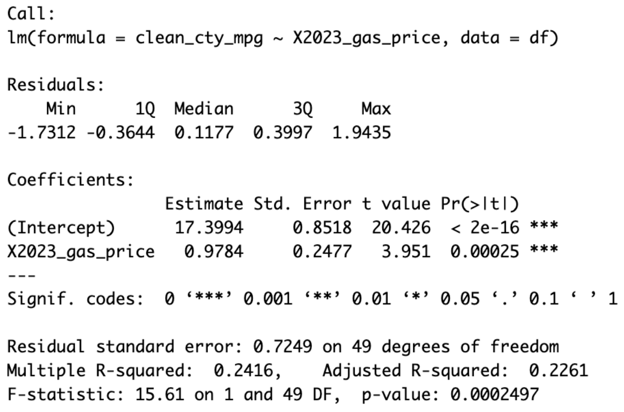
Langer, A. M., & Miller, N. H. (2008). "Automobile Prices, Gasoline Prices, and Consumer Demand for Fuel Economy." Energy Journal, 29(3), 61-90.

Shea, Terry. “Why Does It Cost So Much for Automakers To Develop New Models?” Autoblog (2010). https://www.autoblog.com/2010/07/27/why-does-it-cost-so-much-for-automakers-to-develop-new-models/

Macrotrends. “Cars Profit Margin 2015 - 2022” Macrotrends (2022). https://www.macrotrends.net/stocks/charts/CARS/cars/profit-margins?q=cars+profit+margin

Appendix

This linear model output resulted in the US State Gas Prices vs. Average City MPG mentioned in the current progress section.

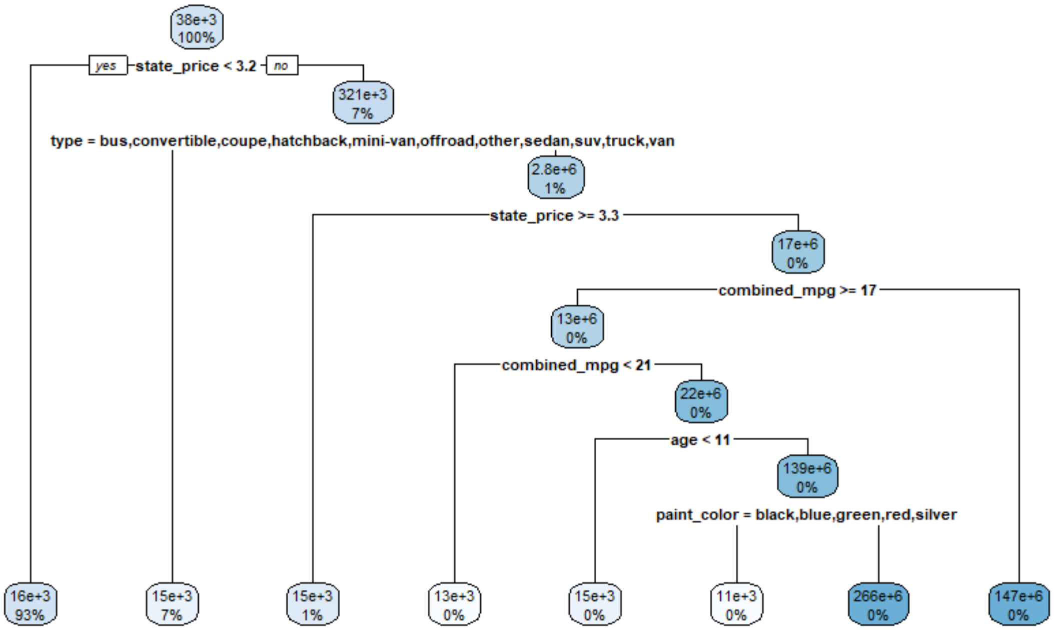


```
Call:
lm(formula = clean_cty_mpg ~ X2023_gas_price, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7312 -0.3644  0.1177  0.3997  1.9435 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      17.3994     0.8518  20.426  < 2e-16 ***
X2023_gas_price   0.9784     0.2477   3.951  0.00025 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7249 on 49 degrees of freedom
Multiple R-squared:  0.2416,	Adjusted R-squared:  0.2261 
F-statistic: 15.61 on 1 and 49 DF,  p-value: 0.0002497
```



Here is a larger version of the Decision Tree model to predict the purchase price of vehicles in the Craigslist dataset.