WHAT FACTORS DETERMINE THE PRICE OF A CAR?

General overview:

Group 18: Austin Pesina, Evan Boyle, Changhui Han, Shekinah Liza Jacob, Parth Patel

GitHub URL: https://github.gatech.edu/MGT-6203-Fall-2023-Canvas/Team-18

Objective: *Our primary objective is focused on the manufacturer's suggested retail price (MSRP) of car prices and identifying the main contributing variables of price.*

Background: Initial data set was obtained from a reddit post which had scraped and aggregated various car specifications (e.g., MSRP, cylinder, MPG, car make, car model, car year) into a .csv file. The .csv file contained 32,000+ rows of data for various car makes and models manufactured between 2014 and 2019. There are 238 columns for each vehicle, however, most columns are a binary response for various features and nonessential variables.

Importance: Understanding the intricate relationship between car specifications and Manufacturer's Suggested Retail Price (MSRP) is essential for the average consumer, significantly influencing their decision-making process when faced with a budget limit. By grasping how various specifications impact car pricing, consumers can strategically manage their budget, while being able to gain the most value out of their car purchase. This understanding also enables a more meaningful comparison between different car models, helping consumers identify the best value based on their priorities. An informed approach to car pricing not only streamlines the purchasing process but ensures that the chosen vehicle aligns with the individual's needs, preferences, and budget, enhancing the overall ownership experience.

Initial Approach: We started the data analysis by examining the original dataset and removing any car makes and models with a considerable number of missing values. We continued to identify significant variables that may influence car MSRP and created preliminary visualizations to unveil patterns and any variables displaying multi-collinearity. Focusing on these key variables, we applied diverse modeling techniques, including linear regression, random forest, and XGBoost regression which can help identify the factors that most affect MSRP. Utilizing the results from all three models, we drew conclusive insights into the key contributors to car MSRP prices by combining the outcomes.

Hypothesis: In our initial hypothesis, we determined that highway and city miles per gallon (mpg) would be substantial contributions to the Manufacturer's Suggested Retail Pirce (MSRP). The rationale behind this assumption stems from the expectation that cars with more advanced technology would be capable of higher fuel efficiency, which would drive up price. Furthermore, we anticipated that horsepower would play a significant role, as cars with greater

horsepower are seen as more luxurious and equipped with more expensive engines, therefore influencing higher MSRP.

Overview of Data:

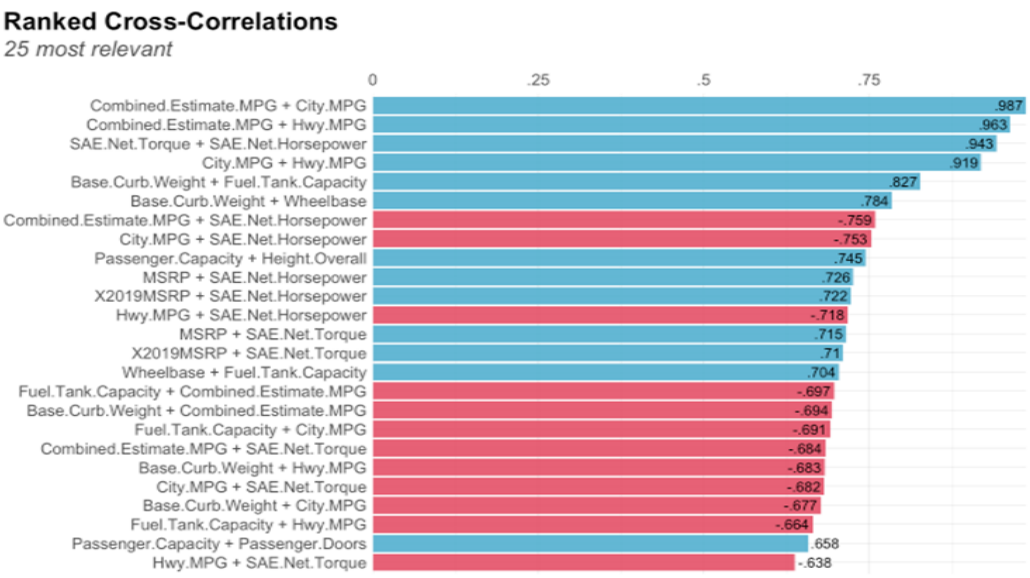
In our data preparation for the car specifications dataset, we encountered a significant issue with missing data across various car makes and models, particularly in categories like MSRP, mileage, body style, and EPA class. Much of the missing data is due to missing data of the initial data scrap from the source (thecarconnection.com). Many of the car makes and models specifications were not complete for the large amount of predictor variables. To address this, we made the decision to remove cars with incomplete data from the testing database, resulting in a reduction of our dataset from the original 9,505 rows to 4,862 rows.

Furthermore, to ensure the data's quality and consistency, we undertook the task of standardizing both categorical and numerical variables. Categorical variables like Brake Type, Steering Type, and Fuel System were standardized to eliminate variations in terminology used by different car manufacturers. For EPA Class, we followed the guidelines provided by the US Environmental Protection Agency to achieve uniformity in class definitions. Additionally, we standardized numerical variables such as net horsepower and net torque, which may have been measured at different revolution speeds depending on the car manufacturers. This standardization process is crucial to ensure that the data is reliable and can be effectively used for analysis and modeling in our project.

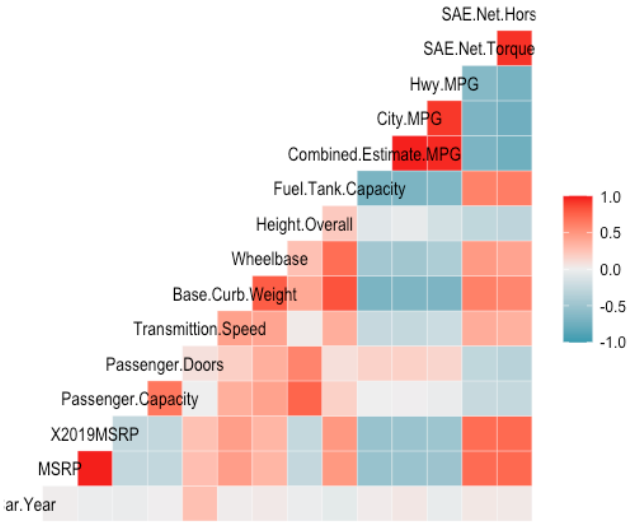
In response to the valuable suggestions provided by the teacher's assistant, we recognized the need to adjust for inflation in the car Manufacturer's Suggested Retail Prices (MSRPs) to ensure that the data remains accurate and relevant over time. To address this concern, we adopted a standardized approach. We established a baseline year of 2019, which would serve as the reference point for MSRP calculations across all car specifications data. To calculate the adjusted MSRPs, we leveraged the inflation calculator provided by the Federal Reserve Bank of Minneapolis. By doing so, we aimed to account for the changes in the value of money and ensure that the MSRPs are comparable and meaningful across different years, thus enhancing the quality and reliability of our dataset for analysis and modeling.

After cleaning the data, we performed exploratory data analysis to identify any initial patterns, relationships, or data distributions. This helped us to understand the main characteristics of our cleaned dataset and to move forward with deciding on which models we were going to use.

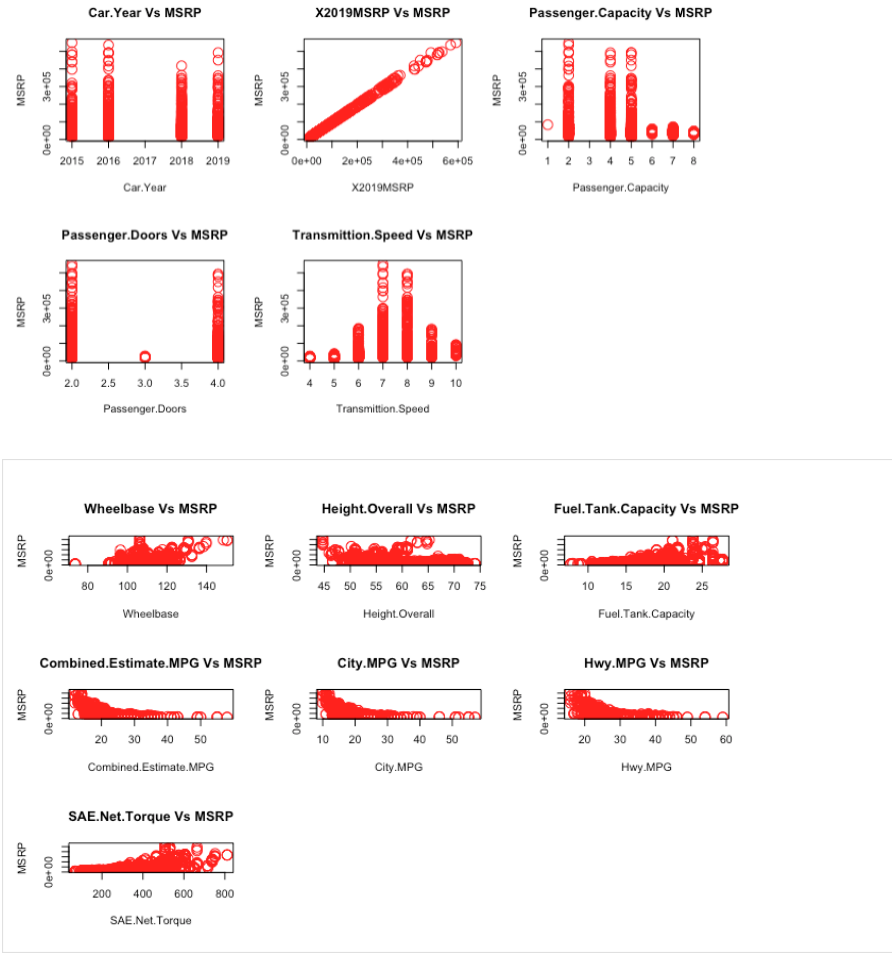
During Visualization, we categorized the variables into numerical variables and character variables. We analyzed the relationship between the numerical variables through a correlation matrix and scatterplots and then converted the character variables to factors. We also utilized the heat map and ranked cross-correlations to understand the dataset in different angles.



Ranked Cross-Correlation of top 25 variables with p-value> 0.05



Correlation Matrix of Numerical Variables

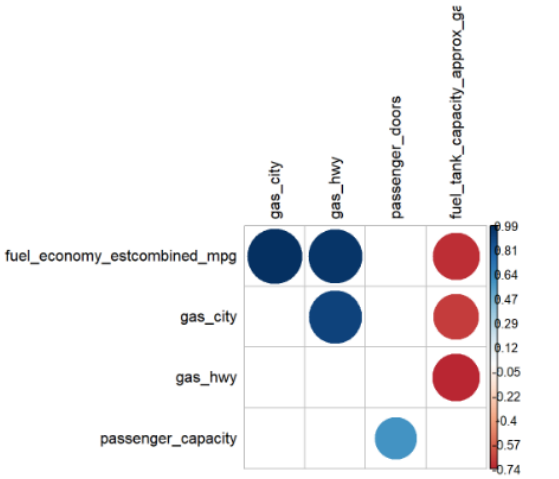


Scatterplots of MSRP with independent variables

We also noticed various instances of multicollinearity in our data during the exploratory analysis phase.

	GVIF	Df	GVIF^(1/(2*Df))
fuel_economy_estcombined_mpg	262.314898	1	16.196138
gas_city	113.896378	1	10.672225
gas_hwy	48.585787	1	6.970351
epa_class	84209.273274	29	1.215962
drivetrain	8.477889	5	1.238308
passenger_capacity	3.954158	1	1.988507
passenger_doors	17.728704	1	4.210547
body_style	78689.905522	16	1.422320
fuel_tank_capacity_approx_gal	4.960368	1	2.227188
parking_aid	1.432587	1	1.196907
tire_pressure_monitor	2.516836	1	1.586454
backup_camera	1.199611	1	1.095268

VIF on initial dataset



Correlation matrix for most highly correlated variables in the initial dataset

Figure shows VIF (Variance Inflation Factor) and multicollinearity analysis on the initial dataset

This meant that we would have to start the process of variable reduction / selection before we were able to run the models. We chose two different methods of variable reduction, stepwise (AIC and BIC) and PCA (Principal Component Analysis) analysis: Stepwise, AIC, and BIC are "*statistical measures used for variable reduction in model selection, with AIC favoring goodness of fit and BIC penalizing model complexity*"[1] PCA analysis: "*PCA(Principal Component Analysis is a dimensionality reduction technique that identifies and projects the most important features (principal components) of a dataset, reducing complexity while retaining data structure.*"[2]

Using these two methods we were able to reduce the number of variables to the top 20 most significant variables (see below).

EPA Class	Fuel System
Drive Train	Engine Type
Passenger capacity	Net Hp
Base Weight	Brake type
Doors	Child locks
Wheelbase	Day lights
Height	Night Vision
Fuel Tank Cap	Roll Bars
Combined MPG	Park Aid
Net Torque	Backup Cam

Overview of Modeling:

Our modeling approach consisted of running a linear regression model, a random forest model, and an XG boost model to identify key relationships between the multiple factors and the MSRP of a given car.

Our first model was a linear regression model which predicts the relationship between two variables by assuming a linear connection between the independent and dependent variables. It seeks the optimal line that minimizes the sum of the squared differences between predicted and actual values. This method is applied in different domains which analyzes and forecasts data trends. Here we used multiple linear regression to predict the value of the variable based on the value of independent variable. We divided the data into 90% training data and 10% testing data. We used 90% training data to fit the linear regression model and found the training error. We also found testing error with the 10% testing data.

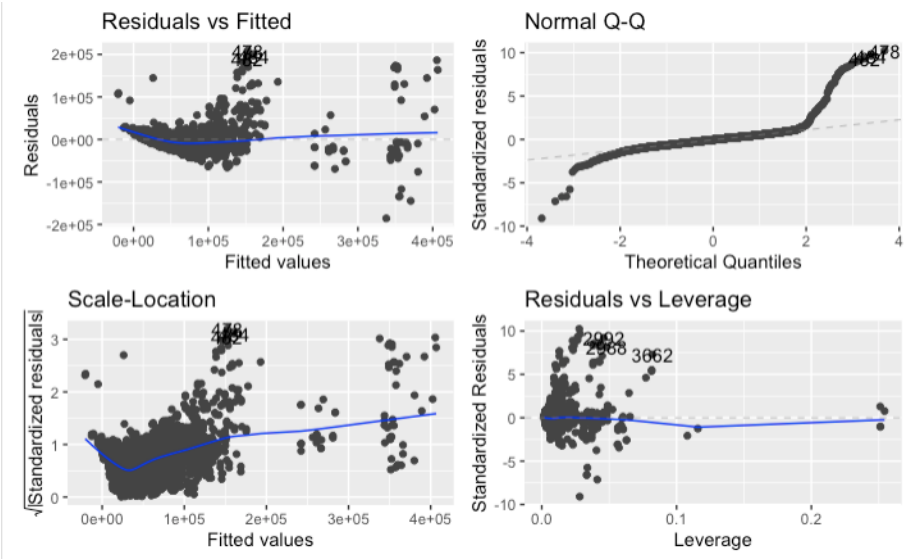
The training and testing errors are shown below:

Training Error – 426763222
Testing Error – 423746366

We also did a Monte Carlo Cross Validation of 100 runs to find the Training Error Mean and Variance and Testing Error Mean and Variance which are shown below:

Training Error Mean- 451041049
Training Error Variance- 1.168546e+16

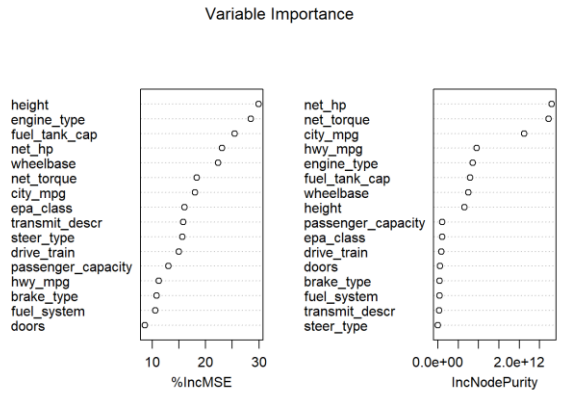
Testing Error Mean- 449947243
Testing Error Variance- 1.183567e+16



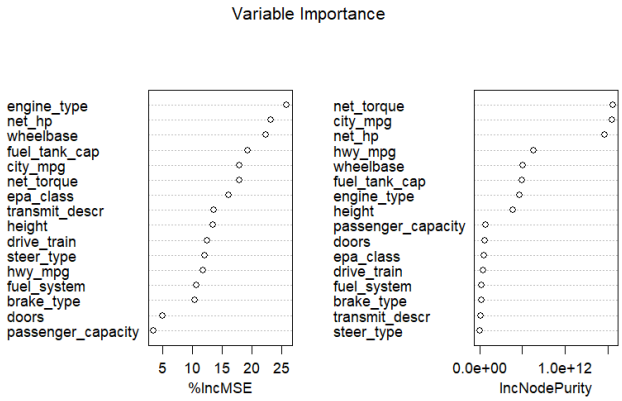
Assumptions of Linear Regression in graphical form

Our second model was a random forest model which we applied twice to the updated and cleaned dataset. The initial model contained all the variables and the second model that we iterated upon only had key variables that we identified through PCA and variable selection. Through this analysis, we determined the variables that were most significant in predicting car prices and evaluated the accuracy of the predictions made by the linear regression models.

Through a random forest model with 500 trees, we were able to get the variable importance graphs shown below. The %IncMSE is the percentage increase in the mean squared error (MSE) of the predictions resulting from the permutation of a particular predictor variable. The higher the %IncMSE, the more important the variable is in predicting the response variable, which is the MSRP for a given car in this case. From the second graph below, predictors like *engine_type, net_hp, wheelbase, fuel_tank_cap, city_mpg,* and net_torque seem to be the most important. On the other hand, predictors like *passaneger_capcity*, *doors, brake_type*, and *fuel_system* do not seem to be as important in the regression models.

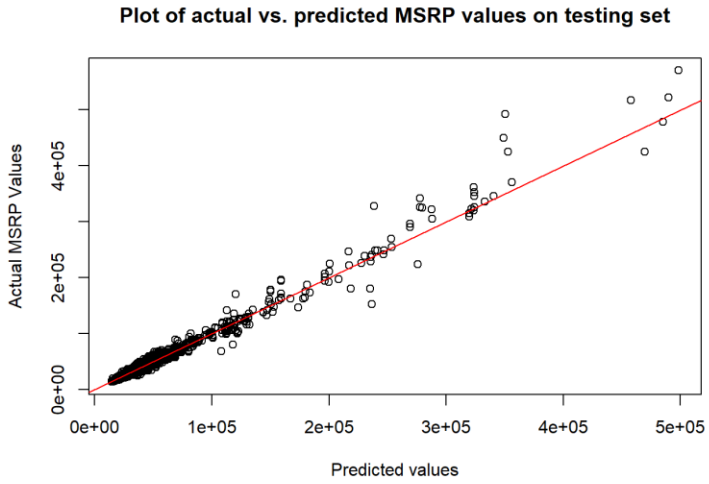


Graphs shows variable importance after running an initial random forest model



Graphs shows variable importance after running a second random forest model with key variables

Additionally, we evaluated the random forest model by splitting the data set into training, validation, and testing subsets. The model achieved an R-squared value of 97.7% on the testing data set which is high. This may be due to overfitting or if there are highly influential predictor variables that have strong relationship with the MSRP of a car. Overall, this might indicate that the results are best understood with a combination of other types of models and further analysis.



Graph shows a plot of actual MSRP values against the predicted MSRP values by the random forest model.

In addition to the Random Forest and Linear Regression models, we built an XG Boost regression model using the same predictors. While training the XG Boost model, it was able to explain most of the variance within the variables, having and R-Squared of 98.4%.

If we were given more time or resources, we would have included the safety data that we obtained and cleaned from a report conducted by the Insurance Institute for Highway Safety (IIHS) for 2017 model year vehicles. This study highlighted the death rate for each vehicle, including deaths that occurred in rollover accidents as well as multiple-vehicle and single-vehicle accidents. We spent a considerable time transforming the data from a .pdf to a .RData in order to run any analysis. However, we soon realized that combining the safety data for the cars that IIHS studied with the cars we obtained from our initial dataset from the Reddit post was not feasible for many reasons. Many of the different car types and their models did not line up or have any safety data that we could use. We would have had to find additional data regarding safety of each car to fully incorporate the safety metrics into our study. For these reasons, we decided to discard the safety dataset and only focus on our main goal of finding key factors that have the most impact on a car's MSRP.

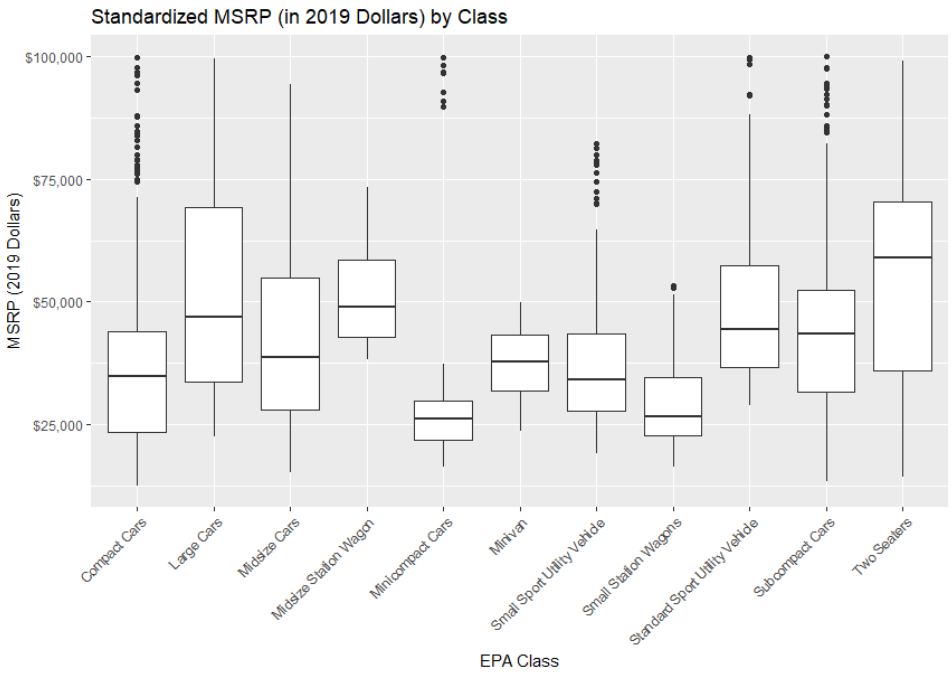
After running all three models on the validation data, the Random Forest and XG Boost models performed similarly well. In terms of prediction error, the Random Forest model had a Mean Absolut Percent Error (MAPE) of 7.7% while the XG Boost model had a MAPE of 7.2%. After narrowing down to these two models, we ran them again on the test data set.

On the test data set, the Random Forest accounted for 97.1% of the variance while the XG Boost accounted for 95.5% of the variance in the data. From a predictive perspective, the models again performed similarly. The Random Forest has a MAPE of 7.9% while the XG Boost had a MAPE of 7.5%. In terms of dollar value, the XG Boost model was $158.05, on average, better than the Random Forest model.

Overall, both models performed equally well and could be used to predict MSRP, in 2019 dollars, of a vehicle.

Overall conclusions and key takeaways:

After analyzing results from all three models, we determined that the following factors have the greatest impact towards car prices in the United States: torque (net_torque), city gas mileage (city_mpg), horsepower (net_hp), and highway gas mileage (hwy_mpg).
While not the most important predictor, vehicle classification type (epa_class) also played a key role in determining MSRP. This would indicate that the type of engine is one of the main factors that a consumer or manufacturer should look at when pricing or buying a car. Torque, mpg, and horsepower are all ancillary results of the type of engine that is in vehicle. We can use these engine variables to determine if a car is overpriced or underpriced relative to the market and can look for "good deals" as a consumer. Those cars would be outliers in our analysis.



Graph shows a boxplot of MSRP values by EPA Class. Vehicles with an MSRP over $100,000 were excluded.

With the results from all three models, it became evident that our initial hypothesis held true, and variables such as highway and city miles per gallon (mpg) did indeed emerge as significant contributors to Manufacturer's Suggested Retail Price (MSRP). Additionally, our findings revealed that torque, another crucial performance metric, played a key role in influencing car prices. Surprisingly, all predictors are associated with the type of engine the car makes and models are equipped with. This aligns with the notion that engine technology, including factors

like horsepower and torque, carries considerable weight in determining the overall pricing of cars. Our results underscore the importance of considering not only fuel efficiency but also the specific attributes tied to the type of engine when comprehensively understanding the pricing dynamics in the automotive market.

Results from this project study can be especially useful for businesses ranging from marketing and product development to insurance industries. For example, we identified certain key factors like gas mileage, vehicle classification, and horsepower to provide insights into the price range of a car that a family of four might be looking to buy. Using this information, manufacturers and dealerships can develop and market vehicles tailored to this specific target audience. Insurance companies can use the results about key crucial factors in cars and their respective impacts on a car's value to adjust premiums and assess risk with more accuracy.

If we had included the safety data in our analysis, we would have been able to show a clear benefit from analyzing which cars are more prone to accidents. This would have better informed consumers before making a purchase, which would have eventually led to safer roads. Manufacturers can use these insights to improve their future cars with additional safety features associated with less accident-prone cars.

Works Cited:

1. Akaike, H (1974). A new look at the statistical model identification. IEEE Transaction on Automatic Control, 19(6), 716-723; Schwarx, G. (1978). Estimating the dimension of a model. The Annals of Statistics 6(2), 461-464
2. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32)
3. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In proceedings of the 22nd ACM SIGKDD International Conference on Knowledge discovery of Data Mining
4. Draper, N.R., & Smith,.H (1998). Applied Regression analysis (3rd ed.). Wiley
5. Federal Reserve Bank of Minneapolis. "Inflation Calculator." Federal Reserve Bank of Minneapolis, https://www.minneapolisfed.org/about-us/monetary-policy/inflation-calculator
6. Jolliffe, I.T. (2002). Principal component analysis. Wiley Online Library
7. U.S. Environmental Protection Agency. "EPA Certifications." EPA, https://www.epa.gov/aboutepa/epa-certifications.