Om Sharan

10129754

Analyzing CAR ValueS

Table of Contents

[Executive Summary 2](#_Toc26571912)

[Introduction 2](#_Toc26571913)

[Purpose & Methodology 2](#_Toc26571914)

[Background Research 2](#_Toc26571915)

[The Data Set 3](#_Toc26571916)

[Data Cleaning, Transformation and Exploration 3](#_Toc26571917)

[Variable Selection 5](#_Toc26571918)

[Regression 6](#_Toc26571919)

[Findings 7](#_Toc26571920)

[Under-valued and Over-valued Cars 10](#_Toc26571921)

[Conclusion 11](#_Toc26571922)

[References 12](#_Toc26571923)

[Appendices 13](#_Toc26571924)

[Appendix I: Cars grouped by Manufacturer 13](#_Toc26571925)

[Appendix II: Car’s Age vs Log(MSRP) 14](#_Toc26571926)

[Appendix III : Correlation Matrices 15](#_Toc26571927)

[Appendix IV: Final Result from Stepwise Selection Process 15](#_Toc26571928)

[Appendix V: Final Regression Results 16](#_Toc26571929)

# Executive Summary

In this paper, we look at the composition of a car’s value. We evaluate the most valuable features of a car including the engine’s characteristics, car’s dimensions and features, safety features, and additional features that help with the driving experience. We use stepwise selection to find the most relevant variables and then create a model using a regression in order to predict what a car’s true value is. Using this value, we find over-valued and under-valued cars based on the predictions of the model.

The most significant factors which accounted for most of the difference in a car’s value was its engine type, suspension type and net horsepower. The transmission type, drivetrain, passenger capacity, fuel type, fuel economy and steering type did also affect the retail price, but to a lesser extent than the first variables. Additional features such as child safety locks, parking aid and back-up cameras were also statistically significant.

# Introduction

Buying new cars is expensive and leasing or buying used cars has become a growing trend. Most consumers primarily pick their next car based on the aesthetics of the vehicle, “how it feels” and the brand name rather than many of the car’s specs. While these attributes are more popular for most consumers, what attributes are consumers paying for? And what is the best value car in the current market?

# Purpose & Methodology

The goal of this paper is to create a model that first analyzes which attributes have the biggest impact on the retail price of a car model and by how much. The best model based on the AIC criteria from a stepwise selection process and then used in a linear regression. Lastly, the best value cars based on this model will be identified.

This will be helpful for consumers to better understand how much they are paying for each feature of the car. This can be useful as it gives the consumer a tool for evaluating the value of car and ensure that they are being treated fairly. Additionally, they would be able to identify cars with the best value if they are shopping for a car.

# Background Research

The MSRP, or manufacturer's suggested retail price, is the price that the manufacturer suggests that the dealer should sell a specified car model (Autotrader 2019). Although the MSRP is the suggested price, dealers can change this depending on the car’s demand and market demand (Autotrader 2019). Generally, there will be a base price for the model and then separate prices which include various options, destination charges or add-ons (Autotrader 2019).

Poornima Tapas and Rita Dangre find that consumers are now increasingly purchasing luxury cars as they represent “a symbol of power, recognition, independence and status” (A study of consumer preferences for buying passenger cars 2013). More and more customers prefer imported cars and their study shows an increase in market share for international manufacturers in India (A study of consumer preferences for buying passenger cars 2013).

Deloitte’s 2019 Global Automotive Consumer study finds that consumers in Europe and Asia are increasingly moving away from internal combustion engines in favor of electric cars, but North Americans are keeping away due to low fuel prices (Deloitte 2019). Autonomy has had a growing interest, but consumers still want a track record of safety and are currently unsure if fully autonomous vehicles are indeed the future (Deloitte 2019). Consumer opinions are mixed regarding additional features that may enhance their driving experience (Deloitte 2019). Interest in these time-saving features is high, but significant concerns remain over privacy and security of these features (Deloitte 2019). Equipment manufacturers are also facing an uphill battle in creating a demand for these features and getting people to pay for them (Deloitte 2019).

# The Data Set

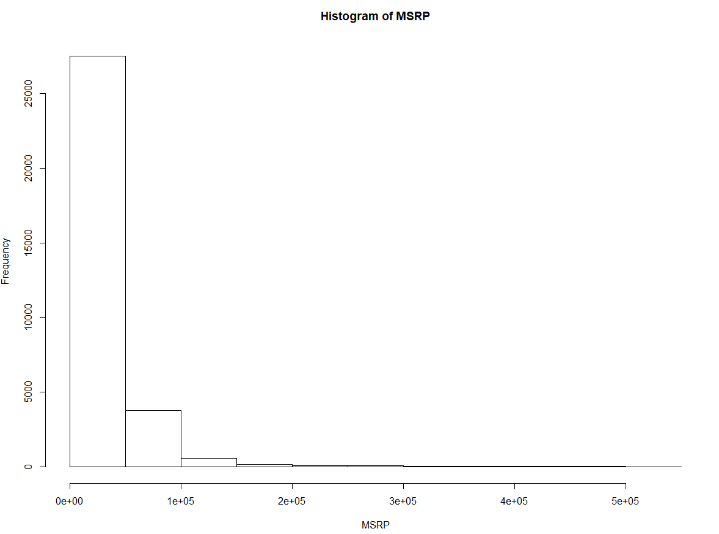
The dataset was scraped by Nicolas Gervais from The Car Connection (The Car Connection 2019) using a scraping method, and the data is available on his GitHub directory (Gervais 2019). The Car Connection is a research website which provides reviews on cars built by 43 different manufacturers that are available in the United States (The Car Connection 2019). They provide specs for most vehicles and have writers who provide reviews for some of the more popular cars (The Car Connection 2019). Additionally, they provide a quick comparison tool for users to compare cars as well as a feature to find used cars nearby (The Car Connection 2019).

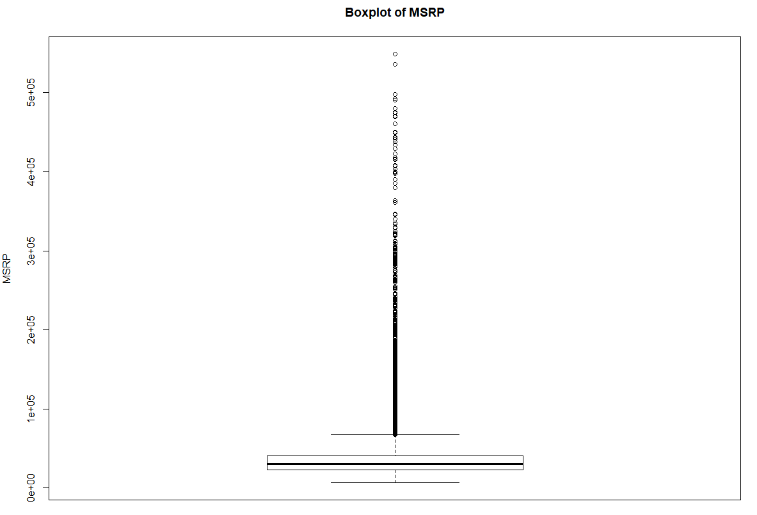
There are 32,316 different car models in the dataset ranging from 1990 to 2019 makes. There are 234 attributes for these cars which include: brand, year, gas mileage, passenger capacity, passenger doors, drivetrain, engine type, body style and many others. The primary response variable will be MSRP (Manufacturer’s Suggested Retail Price) which is in USD. The mean MSRP is $37,707 with a minimum of $6,929 and a maximum of $548,000.

# Data Cleaning, Transformation and Exploration

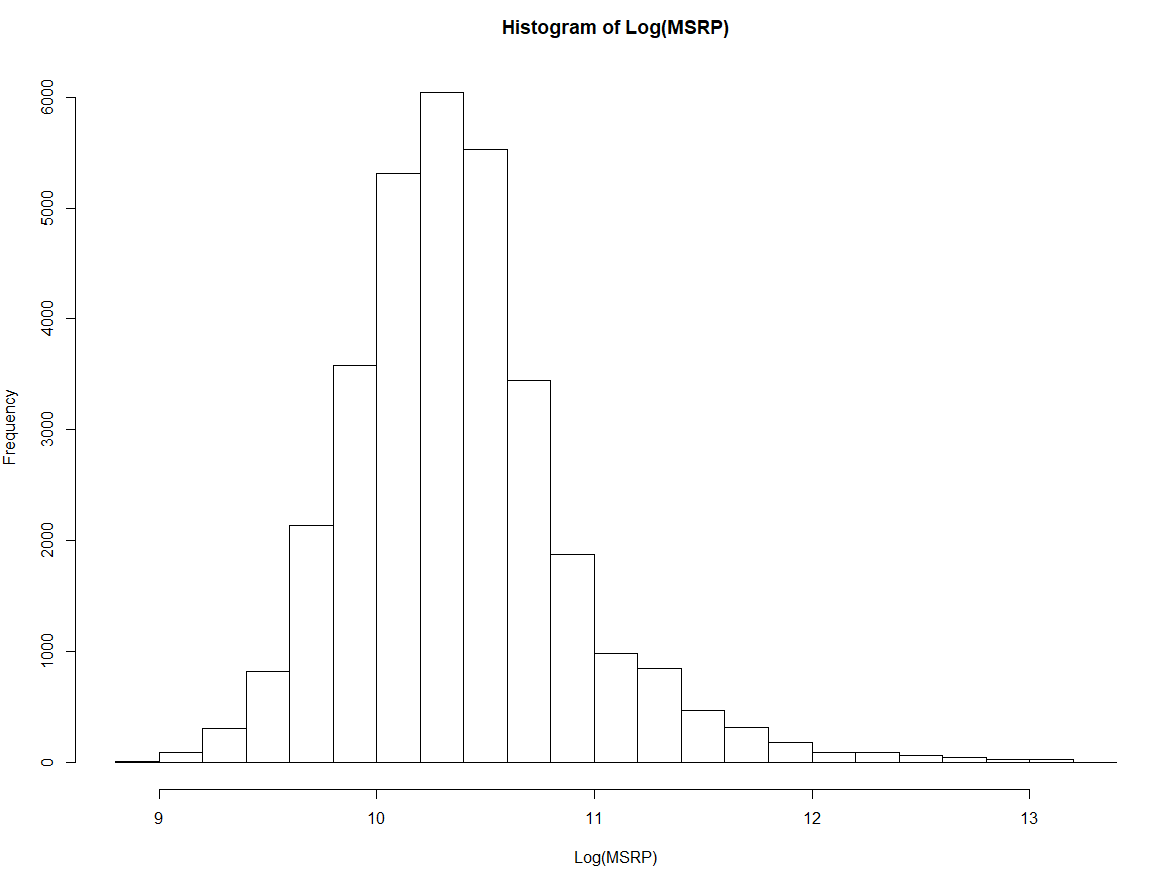
The data was first transposed so that each feature and specification was a column and each vehicle model was a row. Cars with missing MSRP values were omitted since MSRP is the primary response variable. The manufacturer and year were extracted from the vehicle model to create new variables.

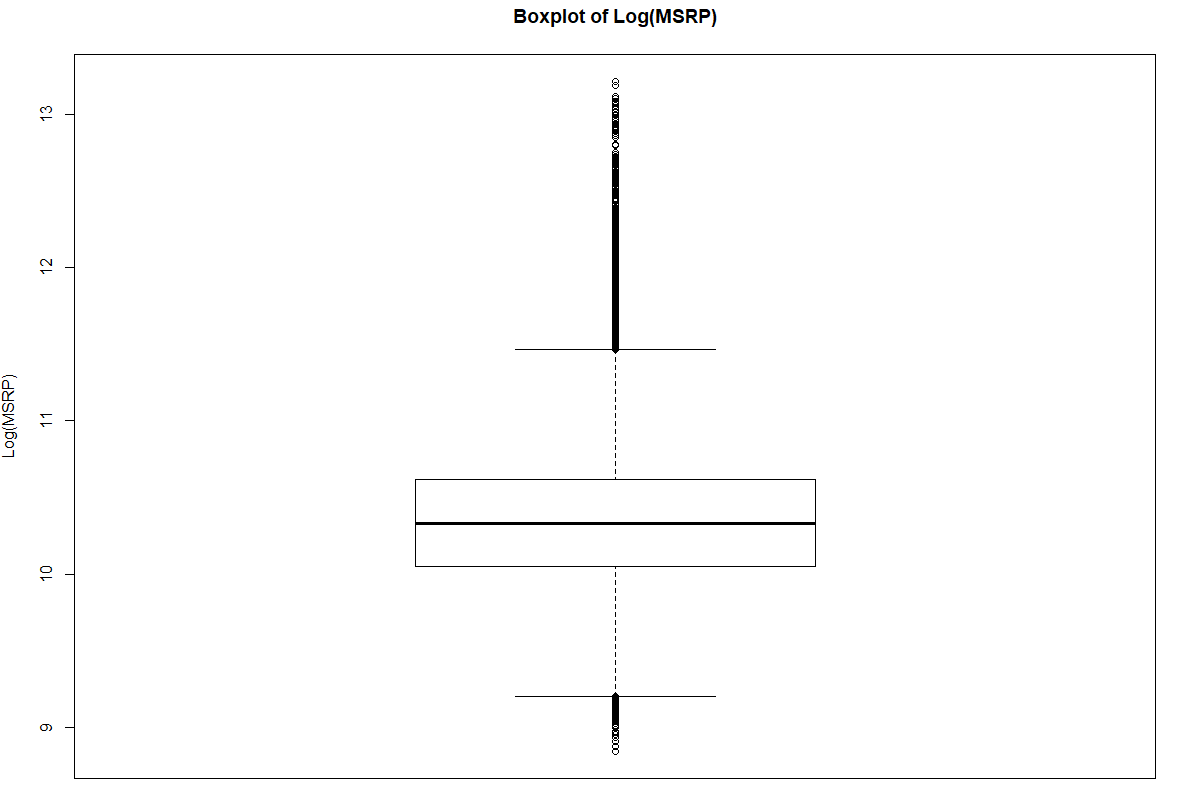
The MSRPs are clearly right-skewed as is evident from the histogram and boxplot of MSRP seen in the two figures below. This makes sense as there are various luxury cars within the dataset that have substantially higher values than most standard vehicles.





Due to the MSRP being right-skewed, the variable was transformed by taking its logarithm. This resulted in a slightly more normalized distribution but was still skewed right. This is evident in the histogram and boxplot of the Log(MSRP) in the graphs below.





There are 43 different manufacturers. A summary of each manufacturer with their mean, median, minimum and maximum MSRPs for their various vehicles can be found in Appendix I: Cars grouped by Manufacturer.

The year of production also naturally affects the price of the car as one would expect older cars to be cheaper than newer ones due to having worse features. The dataset contains cars produced from 1990 to 2019. The negative correlation between production year and the log(MSRP) can be seen in Appendix II: Car’s Age vs Log(MSRP). However, since it will be highly unlikely for a customer to purchase a new car that was not manufactured in the current or previous year, all car models older than 5 years were omitted. This left 10,893 vehicles in the dataset.

Many of the variable values were cleaned up to match and be grouped into categories. This would allow for comparisons between each group but may lose a certain degree of uniqueness for special cases. For example, cleaning up a variable such as the drivetrain (“4-wheel drive” is the same as “four-wheel drive”) can easily be grouped, but there were so many engine types that simply grouping it by configuration and number of cylinders may accidentally be grouping two very different engines together (i.e. not every I-6 engine is the same).

# Variable Selection

There were 234 variables in this dataset which highlighted various specifications of the car. These variables are categorized by the Car Connection as: interior dimensions, exterior dimensions, cargo area dimensions, fuel tank, mileage, engine specs, transmission specs, brakes, steering, tires, wheels, suspension, safety features, warranty, electrical and trailering (The Car Connection 2019).

Naturally 234 variables are too many and was due to the high level of detail that is provided for the car’s specifications. To reduce the number of variables, firstly any variables with more than 50% of missing data were removed. This is simply in order to not lose too much of the sample size. Most of the variables that were omitted in this phase were incredibly specific variables such as the “maximum alternator capacity” or the “dead weight hitch – max tongue weight (lbs.).”

Secondly, correlation matrices were used to identify which numerical variables had high correlations. Some of these correlation matrices can be found in Appendix III : Correlation Matrices. If two variables had a high correlation, the variable that was more sensical while the other removed. For example, the driver’s inches of shoulder room and passenger’s inches of shoulder room had a correlation of 0.93. This makes sense as generally passenger vehicles are symmetrical but might differ for cargo vehicles that are used for transporting goods. Therefore, the passenger’s shoulder room was kept, while the driver’s was removed from the dataset.

Thirdly, some variables did not have any variance now that the dataset only considered vehicles from the past 5 years. These variables such as frontal airbags for the driver may have been relevant when considering all vehicles as the dataset included cars prior to 1998 but were no longer relevant. Legislation passed on September 1st, 1998 made air bags mandatory for frontal airbags meaning that ever car manufactured since then had air bags (National Highway Traffic Safety Administration n.d.). Variables that were no longer relevant were also excluded.

Close to 40 variables were now remaining. Now to further identify relevant variables, the step-wise selection process was used. The smallest possible regression used included no variables, while the largest included all the variables that were remaining. The AIC from the stepwise regression dropped from -7449 from the starting regression to -7879 for the final regression. The stepwise selection process concluded that 21 variables were relevant in explaining the differences in log(MSRP).

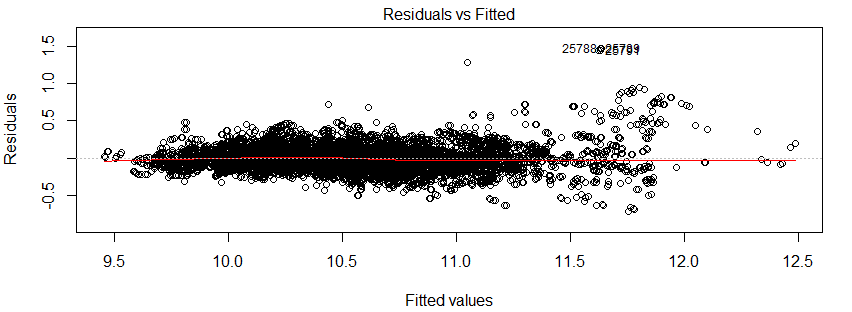
This was still a lot of variables. Using Occam’s Razor, some of these variables were omitted in order to create a more simplified model. This would better reflect the views of a consumer who is interested in purchasing a car.

# Regression

The final variables included: Suspension type, Engine type, Transmission type, Drivetrain, Fuel type and Steering type as categorical variables which represented the vehicle’s specifications. The car’s Passenger Capacity, Fuel Tank Capacity (in Gallons), Estimated Fuel Economy, and Net Horsepower were included as numeric variables. Most vehicle body types were not statistically significant, only “Convertible” was. Therefore, a dummy variable was created to represent if the car was a convertible or not. Vehicle features such as child safety door locks, parking aid, and a camera for backing up were statistically significant dummy variables as well.

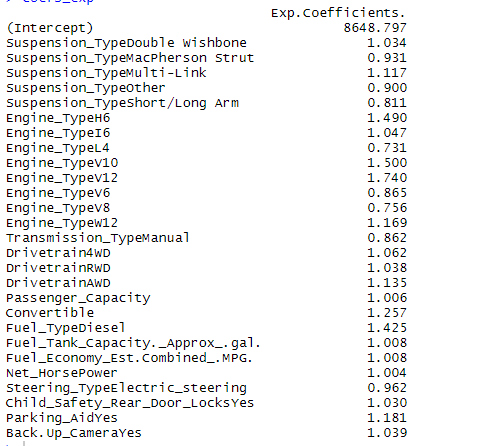
The regression results can be found in Appendix V: Final Regression Results. Each variable is statistically significant to the 95% confidence interval. There are 9454 observations used and the residual standard error is 0.1873. The adjusted R2 is 0.8706 and the F-statistic is 1727 on 37 and 9454 degrees of freedom. This indicates a strong goodness of fit and explanation of the variance by the model.

Below is the residuals vs fitted plot for the regression. It is apparent that the linearity assumption is held as the residuals are evenly distributed along 0. Normality seems to also hold despite the log(MSRP) still have long tails even after the transformation. This is likely due to the large sample size. Independence should also hold. The only assumption that needs to be monitored is of Equal-Variance. While the residuals are generally evenly spread out, it appears that the variance does increase slightly more for fitted values about 11.25. If the assumption doesn’t hold, the least square estimates are still unbiased, but the standard errors may be misleading.



# Findings

Below are the exp(coefficient) values for each of the variables. The intercept is $8,649 and represents the expected retail price for a car with a coil-over suspension, electric engine, automatic transmission, front-wheel drivetrain, no passengers, not a convertible, gas fuel, 0 horsepower, rack-and-pinion steering, and no special features.



Among the suspension types, it appears that the coil-over suspension (intercept) value is the third most expensive suspension type. A double wishbone suspension type would raise the price from the coil-over suspension by 3.4% and the multi-link suspension by 11.7%. The MacPherson strut suspension would drop the price by 6.9%, the short or longa arm suspension by 19.9% and other types generally by 10%. The Magnetic Ride Control, HiPer Strut, and Torsion Bar suspension types were not statistically significant. Coil-Over suspensions are generally used in pick-up trucks and the dataset has only GMC and Chevrolet using them. Suspensions are useful in “maximizing friction between tires and the road surface, in order to provide steering stability and handling” (Harris 2005). Therefore, it appears the Coil-Over is an average suspension system with MacPherson Strut, Short/Long Arm and Other types reducing a car’s value, while Double Wishbone and Multi-Link suspension systems raise a car’s value.

The engine type had the biggest discrepancy out of all the variables in the regression. L4 (-26.1%), V8 (-23.5%) and V6 (-13.5%) engines were the only ones that would drop the car’s vehicle price from an electric engine. I6 engines raised the price over electric engines by 4.7%, W12 engines by 16.9%, H6 engines by 49%, V10 engines by 50%, V12 engines by 74%. It is apparent with this variance in price for engines that they are a crucial element of the car’s value with more powerful engines being substantially more expensive than standard ones. Additional variables which exclusively considered the configuration type as well as the number of cylinders were not selected by the stepwise selection process. This is apparent in the V engine configurations where the V6 and V8 reduce the value of a car while the V10 and V12 raise it substantially. The cylinder count can also be seen to not be significant directly from these coefficients as 6-cylinder vehicles decrease in price when using V6 engines but increase with H6 and I6 engines. This indicates that the configuration type and number of cylinders must be accounted for together when evaluating an engine.

Swapping to a car that is a manual from an automatic would reduce the retail price by 13.8%. Considering that the dataset is taken from the United States, these results do seem to follow general intuition as automatic cars are significantly more common than manual cars (Montoya 2019). Cars that offered both manual and automatic transmissions were not statistically significant, most likely due to a small sample size, and therefore omitted.

Having front-wheel drive is the most common drivetrain on standard cars. This is further backed by the fact that all other types of drivetrains would increase the price from the common drivetrain. Rear-wheel drive increased the price by 3.8%, four-wheel drive by 6.2%, and all-wheel drive by 13.5%. These results do appear to follow general engineering philosophy as cars with four-wheel drive provides power to both front and rear axles whereas the front-wheel drivetrain only provides power to front tires. With more power, there is an increase in performance and therefore an increase in a car’s value. Even more powerful are vehicles with all-wheel drive which use torque vectoring and increase the performance of cars and SUVs in non-off road scenarios, which is more useful to the common customer. It is surprising that the all-wheel drivetrain is more than double than the four-wheel drivetrain as many consumers may not know that there even is a difference (McHugh and Corcoran 2019).

Passenger Capacity was statistically significant, but an increase in one passenger only increases the value of a car by 0.6%. This is likely due to most cars and sedans being standard 5-seater vehicles. The dataset did also have cars with only 2 passengers which were generally luxurious and had an average price of $93,362. On the other side of the spectrum, there also are 9-person Chevrolet Tahoe and GMC Yukon’s whose average models cost $50,797. Passenger capacity does not have a seemingly large impact on price even though it may be a major criterion for consumers. The reason the regression probably doesn’t result in a strong coefficient is because more than 60% of cars were 5-seaters and more than 90% were 4-7 seaters. Discrepancy in vehicle prices normally is not accounted for by passenger capacity.

Out of all the body styles, only the convertible style ended up being statistically significant. Therefore, a dummy variable was created to represent vehicles which were a convertible. Convertibles raised the price over non-convertible cars by 25%. This does seem to follow general intuition as most consumers do consider convertibles to be more expensive than regular cars.

Like the body styles, only the electric steering type was significant. The standard is rack-and-pinion steering which was valued higher than electric steering. Electric steering reduces the price of a car from rack-and-pinion by 3.8%.

Changing from a regular gas fuel type to diesel would increase the car’s value by 42.5%. This is significantly higher than engines that use regular gas and may be surprising to most customers as their prices are relatively similar when pulling up at gas stations. However, apparently cars and small trucks achieve a higher performance when using diesel engines as they get better mileage and fuel efficiency (Bjornstad 2016). Electric and hybrid fuel types ended up not being statistically significant in the regression.

A larger fuel tank would increase the price of the car. For every increase in size of the fuel tank by 1 gallon, the expected price of the car would increase by 0.8%. The smallest fuel tank (1.9 gallons) in the dataset belonged to the BMW i3 series which is a hybrid car and primarily uses its electric engine and only has a small reserve tank for backup. The largest fuel tank (48 gallons) belongs to the Ford Super Duty F-250 pick-up truck. Judging by the magnitude of the coefficient, this does appear to make sense as one would expect the price to go up with a larger fuel tank, but this is not the most significant factor in increasing a car’s value.

Like fuel tank capacity, a car’s fuel efficiency was positively correlated and statistically significant, but not as major of a factor in changing a car’s price in comparison to other variables. A fuel economy increase of 1 mile per gallon resulted in a 0.8% increase in the vehicle’s price.

Net Horsepower was also positively correlated with a car’s retail price. This makes sense as a car with more horsepower is more powerful and therefore has a higher value. For every increase of 100 HP, the price of the car would increase by 40%. The dataset varied from cars with 1600 HP (Mercedes Benz GLE series) to 6500 HP (Ferrari 812). This variable was one of the most significant as it had the biggest t-value of 94.229. One of the reasons this is the case is most likely differentiating the luxury cars from standard cars very well as it has such a high spread in horsepower. Generally horsepower isn’t the only aspect of a car’s power that is evaluated alone as experts also recommend accounting for the rotations per minute (RPM) needed to achieve the max horsepower as well as the torque (Sherman 2016), but those variables were deemed insignificant by the stepwise selection process.

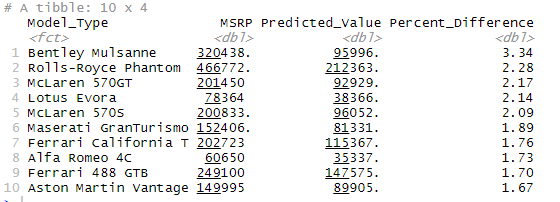
Out of the additional features, only the presence of child safety locks, back-up cameras and parking aids were statistically significant. Inclusion of a child safety lock mechanism would increase the car’s price by 3%. Inclusion of a camera that aids drivers in reversing increased the predicted car price by 3.9%. Inclusion of a parking aid system within the vehicle that assists the driver in parking increased the price of a car by 18.1%. Day-time Running Lights, Fog Lamps, Tire Pressure Monitor, Traction Control and Stability Control were additional features that were not statistically significant. This is most likely due to them being common in modern cars and the sample size of cars which do not include them being low. Rollover Bars and Night Vision were features that were insignificant as well, but these two were most likely due to most cars not having them and the sample size of cars which do include them being low.

Surprisingly, no safety variables ended up being statistically significant. This is likely due to the most common types of Air bags being present in all vehicles as well as Anti-braking systems. The additional air bags were not statistically significant either most likely due to not many cars having side air bags for their front and rear passengers. This, however, is an interesting finding as it somewhat contradicts the findings from the Deloitte Global Automation report (Deloitte 2019). It appears that cars will all have a certain degree of safety and supplementary safety features do not increase a car’s value.

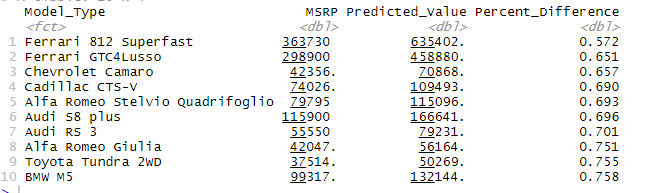
# Under-valued and Over-valued Cars

Using the predicted values for each car from the regression, the percent difference was used to evaluate which cars were the most under-valued or most over-valued based on their specifications. The mean percent difference was an over-value of 5% over the expected with a mean of 1%. The minimum was a 43% under-value and the maximum a 224% over-value.

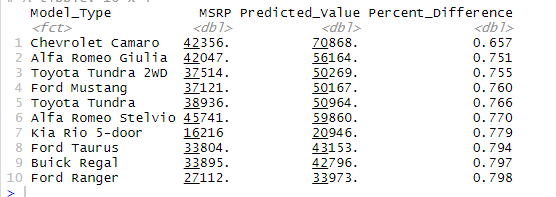
Below is a list of the top 10 most over-valued cars. It is led by Bentley’s Mulsanne model which the model deems is over-valued by 224%. The top-10 cars seem to predominantly be filled with luxury cars as it includes models from Bentley, Rolls-Royce, McLaren, Maserati, Ferrari and Aston Martin. The only cars below $100,000 in MSRP are the Lotus Evora and Alfa Romeo 4C which the model things are still highly over-valued.



Below is a list of the top 10 most under-valued cars. It is led by Ferrari’s 812 Superfast. Both it and the 2nd place Ferrari GTC4Lusso are likely so heavily under-valued due to having some of the highest Horsepower in the dataset for any vehicle. Ferrari and Alfa Romeo are the only manufacturers to have a car model that is in the top 10 in over-valued cars and in the top-10 in under-valued cars.



Most of the cars listed in the top-10s are premium cars and would not be bought by the average consumer. Below is a list of the top-10 under-valued cars with an MSRP less than $50,000. This would be most relevant for most consumers when buying their next car. It is led by Chevrolet’s Camaro which the regression deems is under-valued by 34.3%. Toyota, Alfa Romeo and Ford are manufacturers which two models each in this list.



# Conclusion

In conclusion, we reject the null hypothesis that all features of a car have the same impact on the price of a car in favor of the alternative. The engine type, suspension type and net horsepower have the biggest impacts on a car’s price. The transmission type, drivetrain, passenger capacity, fuel type, fuel economy and steering type did also affect the retail price, but to a lesser extent than the first variables. Additional features such as child safety locks, parking aid and back-up cameras were also statistically significant. No safety features were statistically significant. This likely shows that all cars had a certain standard for safety features and additional safety features would not increase its value.

The most over-valued car was deemed to be the Bentley Mulsanne and the most under-valued the Ferrari 812. The most under-valued car under $50,000 was the Chevrolet Camaro.

# References

2013. "A study of consumer preferences for buying passenger cars." *Researchgate.* March. Accessed December 2019. https://www.researchgate.net/publication/309739444\_A\_study\_of\_consumer\_preferences\_for\_buying\_passenger\_cars.

Autotrader. 2019. *Buying a Car: What's an MSRP?* July. https://www.autotrader.com/car-tips/buying-a-car-whats-an-msrp-228292.

Bjornstad, Erik. 2016. *Diesel vs. Gasoline: Which Engine is a Better Fit for You?* January 21. Accessed December 2019. https://www.bellperformance.com/blog/diesel-vs.-gasoline-which-engine-is-a-better-fit-for-you.

Deloitte. 2019. "2019 Global Automotive Consumer Study." Accessed December 2019. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/manufacturing/2019-global-automotive-consumer-study-americas-report.pdf.

Gervais, Nicolas. 2019. *Predicting Car Price from Scraped Data.* Accessed November 2019. https://github.com/nicolas-gervais/predicting-car-price-from-scraped-data.

Harris, William. 2005. *Car Suspension.* May 11. Accessed December 5, 2019. https://auto.howstuffworks.com/car-suspension.htm.

McHugh, Brian, and T. Arthur Corcoran. 2019. *AWD vs. 4WD: What's the Difference?* August 23. Accessed December 2019. https://cars.usnews.com/cars-trucks/awd-vs-4wd.

Montoya, Ronald. 2019. *Manual vs. Automatic Pros and Cons: Which Is Better?* October 14. Accessed December 2019. https://www.edmunds.com/fuel-economy/five-myths-about-stick-shifts.html.

National Highway Traffic Safety Administration. n.d. *Air Bags.* Accessed December 2019. https://www.nhtsa.gov/equipment/air-bags.

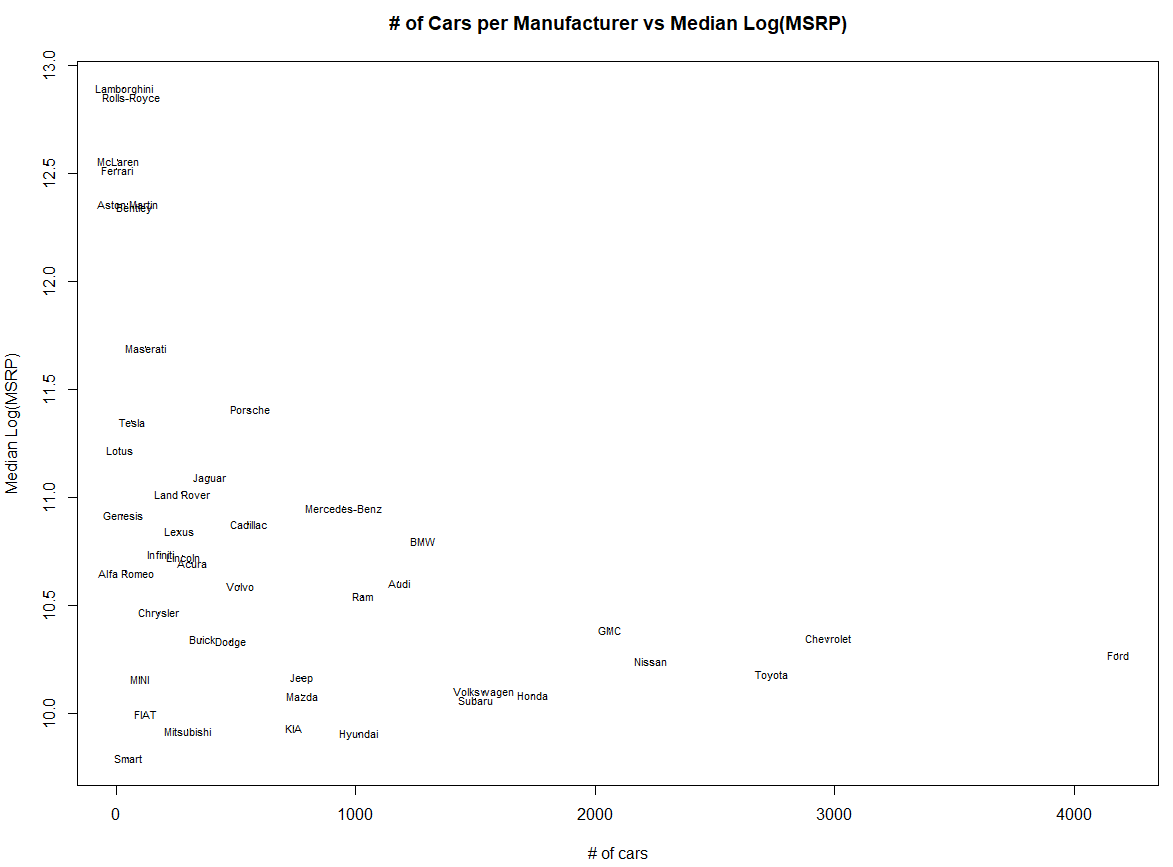
Sherman, Don. 2016. *Horsepower vs. Torque: What’s the Difference?* April 15. Accessed December 2019. https://www.caranddriver.com/news/a15347872/horsepower-vs-torque-whats-the-difference/.

The Car Connection. 2019. December. https://www.thecarconnection.com/.

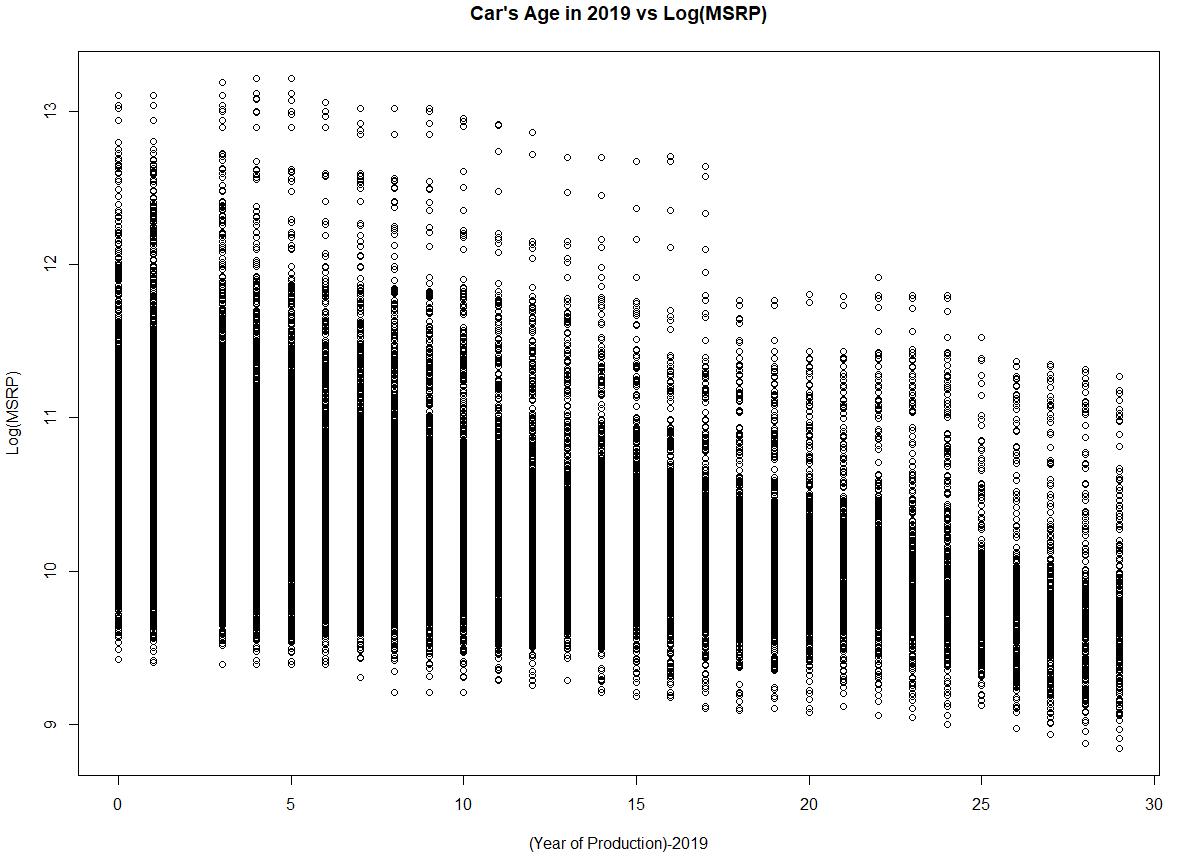
# Appendices

## Appendix I: Cars grouped by Manufacturer

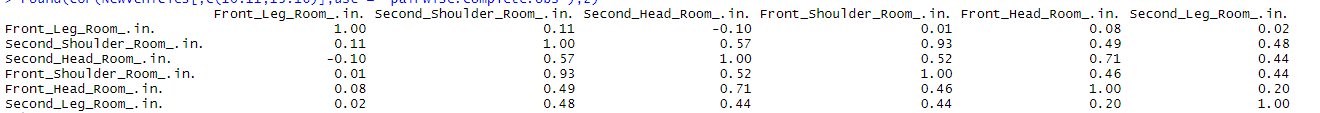


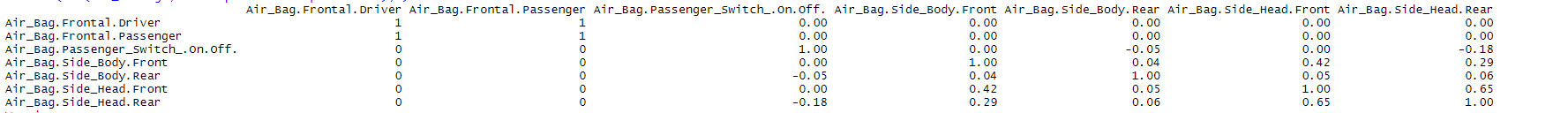


## Appendix II: Car’s Age vs Log(MSRP)

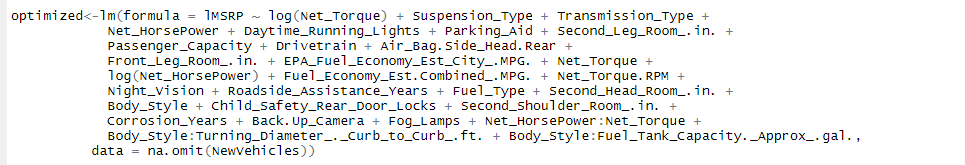


## Appendix III : Correlation Matrices





## Appendix IV: Final Result from Stepwise Selection Process



## Appendix V: Final Regression Results

