Springboard Data Science Career Track
Capstone Project

Finding the Car Price that is Right:
Price Point

Darren Lyles (2018)

Table of Contents

I. Introduction

Buying a car is one of life's key milestones — right up there with getting your driver's license, graduating from school, and buying your first house. It can also be one of life's most expensive endeavors, so it's a decision that should not be taken lightly.

The goal here is to address questions coming from car buyers, marketing analysts, and car manufacturers. For example: I want a car with this set of features. What is my price point? Between Kia and Toyota, which brand has the competitive price edge? Is Mercedes more expensive than its competitor, BMW? I took a data-driven approach, which focuses on a car's features and its pricing. I addressed features such as make; engine horsepower (HP); number of cylinders; transmission type; city and highway miles per gallon (MPG); and number of doors. For the price prediction phase of the analysis, I used a Random Forest regressor model that incorporates all of the features mentioned.

The other question to address is why does predicting a car's price even matter? Customers can determine what price they expect to pay for their desired features, and decide if a car is priced reasonably. One example of this would be: A customer wants a Kia with 150 HP, 4 cylinders, 4-door, automatic transmission, 35 city MPG, and 40 highway MPG. The predicted price is $22,000. The buyer would expect this car in the market would be priced as such and reasonable. Additionally, businesses and market analysts can price cars competitively for their respective markets. Using the previous example, Kia wants to compare a car from Toyota with the exact features, and it turns out that the Kia is priced $20,000 and the Toyota is priced at $23,000.

II. Overview of Data Set

The original dataset, which was acquired from Kaggle.com, contains 11,914 rows and 16 columns. Five of these columns are numeric and 11 are categorical. I wanted to get the most recent year feature in the data set so that I could narrow my focus to new cars instead of used cars. The values of interest were the cars manufactured for 2017; therefore, I extracted a subset of the data set, which only contained rows that had a year feature of 2017. This subset, which contained 1,668 rows and 16 columns, was used to perform data wrangling and exploratory data analysis. We can now refer to this subset is now our dataset.

III. Data Wrangling

When I began cleaning the data, I checked all 16 columns to make sure all entries had the same data type. For example, the make, engine HP, and city MPG columns had data types string, float, and int, respectively. To verify that each entry in a given column was consistent, I iterated through the column and tallied the entries that had the correct data type and the entries that had otherwise. These particular columns had a 100% success rate in terms of data type consistency, with the exception of the engine HP column; this column had 16 NaN values, which resulted in 0.96% of bad data. This column was retained because an overwhelming majority of the entries still contained good data. Additionally, there was one column that had a 77% success rate. This was the marketing category column. The data type that was expected for each entry in the column was of type string; however, string values ended up missing and were denoted as NaN. Since 23% of the values were not of type string, which is quite significant, I dropped the marketing category column to prevent any further complications in the data analysis. The data set is now down to 15 columns for analysis, with the number of rows unchanged.

The next stage of cleaning the data actually occurred when performing EDA. When looking for correlations between MSRP and highway MPG, I noticed that there was a significant but unrealistic outlier in the data. Upon further investigation, I was able to find the row, which, I believed contained erroneous data. It was a 2017 Audi A6 Sedan with a city MPG of 24 and a highway MPG of 354! After fact checking with fueleconomy.gov, I confirmed the highway MPG data was clearly wrong. Instead of changing that particular data point, I removed the entire row and continued on with EDA.

Overall, cleaning the data was pretty straightforward, especially when it came to checking consistency in data types across all columns. However, surprises may start to show up when performing EDA as with what happened with the erroneous MPG data point. Although one takes the necessary steps to make the data as clean as possible initially, the data may not always be completely refined for EDA. It may be necessary to switch back to data wrangling from EDA to refine, clean, and sometimes modify the data appropriately before moving forward.

IV. Initial Findings

When investigating the data set, I wanted to find features interesting and relevant to count. Below is a summary of what I found:

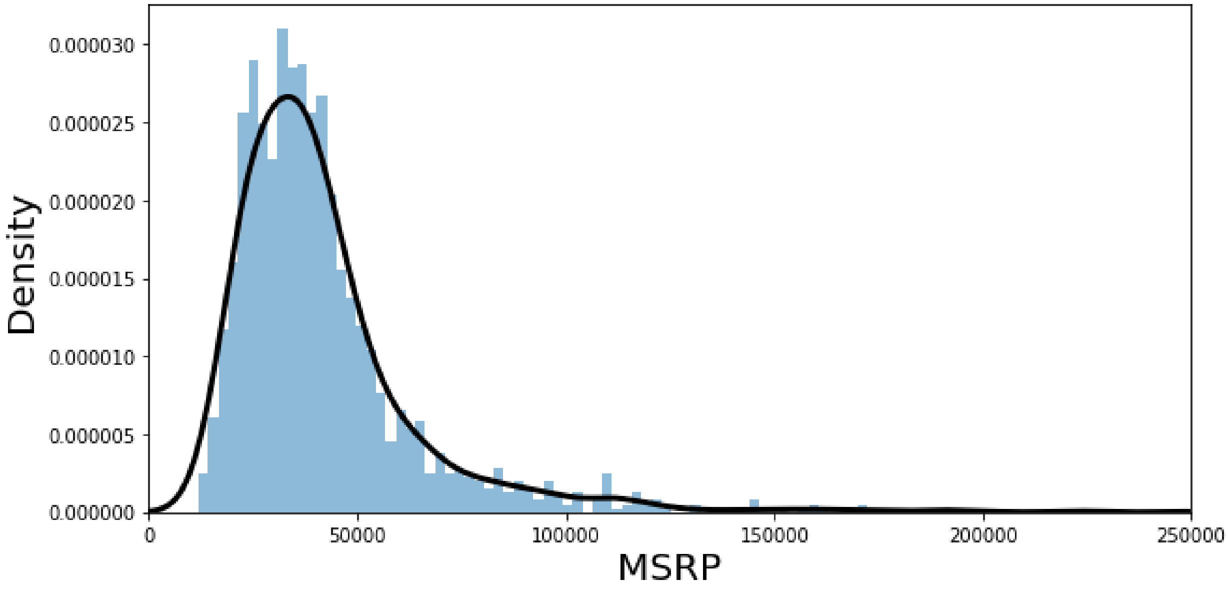
Number of auto brands	30
Cars with manual transmission	205
Cars with automatic transmission	1356
Cars with automated manual transmission	98
Cars with direct drive transmission	9
Number of 2-door cars	276
Number of 3-door cars	16
Number of 4-door cars	1376

These values are interesting since we can better understand the scope and limitations that the data set given has to offer. Our first observation shows that we are limited to 30 car brands out of 50 possible car brands which are sold in the US. Differentiating between cars with manual and automatic transmission is also interesting because it is known that manual transmission is cheaper to implement than automatic. With that in mind, transmission type should have an effect in the pricing of a car. There are also direct drive and automated manual transmission, however the number of vehicles which have these transmission is relatively small.

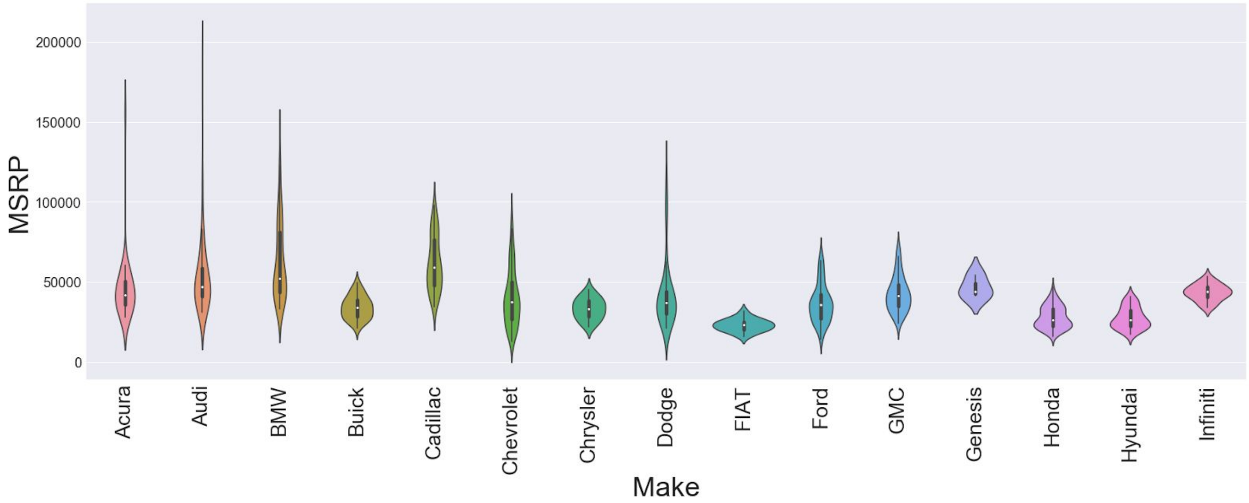
The amount of cars with two, three, and four doors are also something to consider since two door cars can usually be a high performance or sports car. There were 16 three door cars, which was also intriguing and with a little bit of investigation into the data set, these three door cars were found to all be Ford Transit Wagons. The third door for this vehicle is the sliding door on the side which lets multiple passengers enter the vehicle.

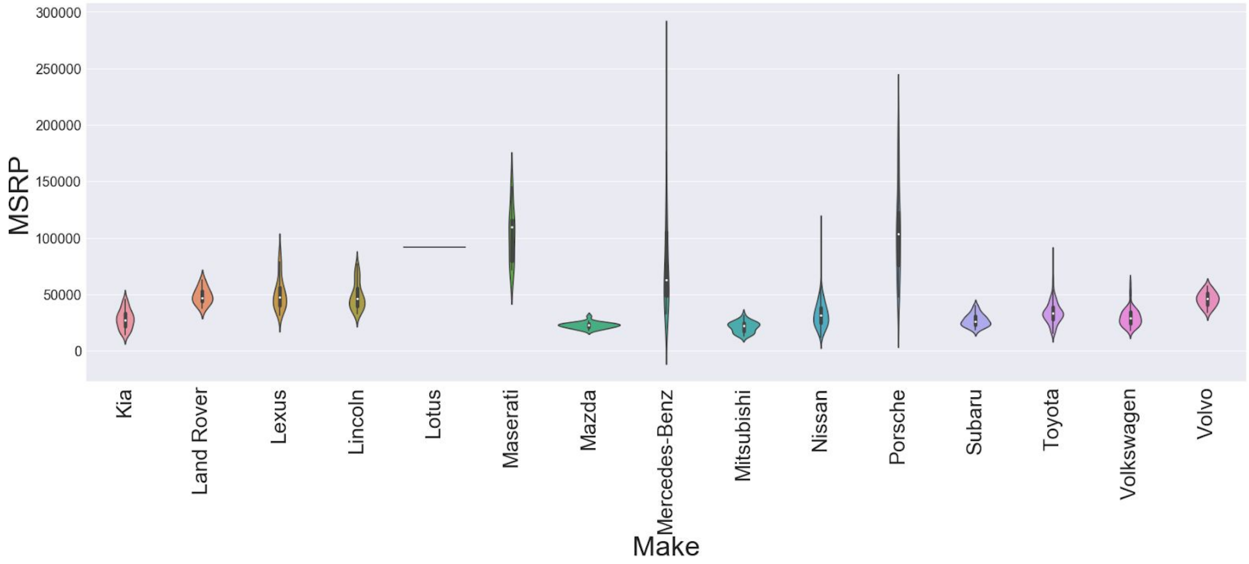
V. Exploratory Data Analysis

Now we will deep dive into the data set and apply more serious EDA. First of, one interesting discovery about the data set is that it is not normally distributed in terms of pricing.



The distribution was heavily skewed to the right; therefore, it fails to satisfy one of the criteria of being a normal distribution. Furthermore, to quantify this claim, I applied a normal test and obtained a p-value of 1.26E-241, which is essentially zero. Since this p-value falls well out of the 95% confidence level, the distribution MSRPs is not normal. The lack of normality implies that there are vehicles which belong in different markets. The long right tail indicates the presence of sports and luxury car markets, while the bell portion of the curve indicates the economy car market. The bell portion of the curve is where most consumers and manufacturers fall under. After analyzing the price distribution, I then took a look at the overall distributions of the MSRP by brand.

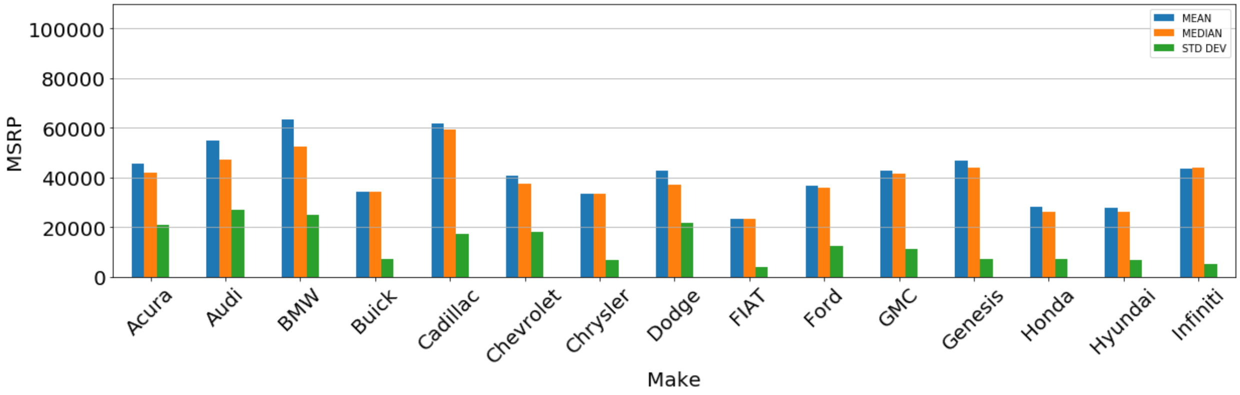


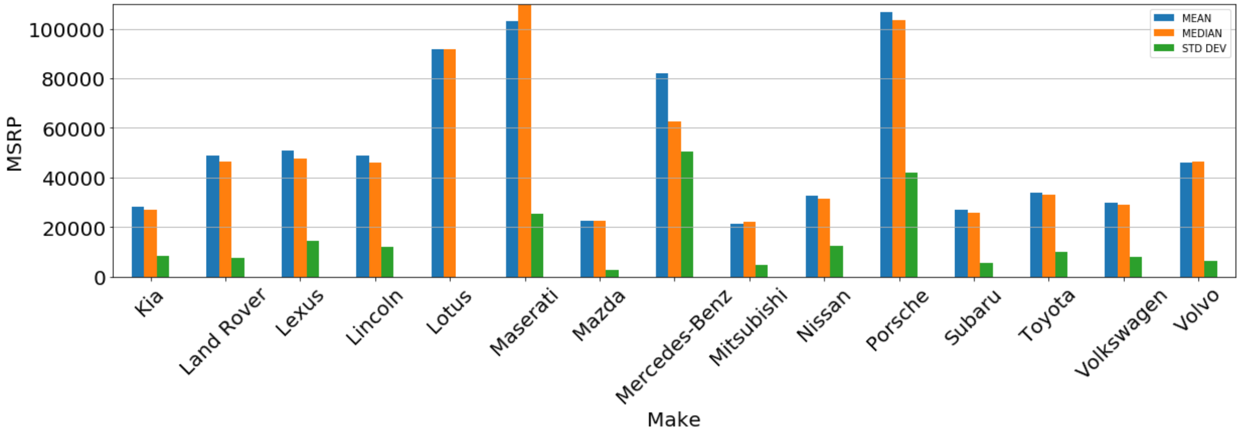


In the violin plots above, I enumerated each car brand in the data set and provide violin plots to better understand the MSRP pricing for each brand. One of the observations we find in these plots is that luxury car brands, namely Acura, Audi, BMW, Maserati, Mercedes Benz, and Porsche, have larger spreads of pricing compared to the other brands listed. This makes sense from a consumer standpoint since these brands have a variety of models that range from being affordable for someone in the middle class and up to costing six figures. From a business perspective, these manufacturers are trying to appeal to wide spectrum of socioeconomic backgrounds to maximize profits.

Another observation is that brands usually associated with being more affordable have a much smaller spread in pricing. Some examples include Fiat, Honda, Hyundai, Kia, Mazda, Mitsubishi, and Subaru. Not only is the spread in pricing small, but the median pricings for these makes is well within the $20000 to $30000 range. From a consumer standpoint, one can argue that if they're looking for a more affordable car, filtering your options by brand name would be a good start! These brands are particularly focused on selling in the economy car market.

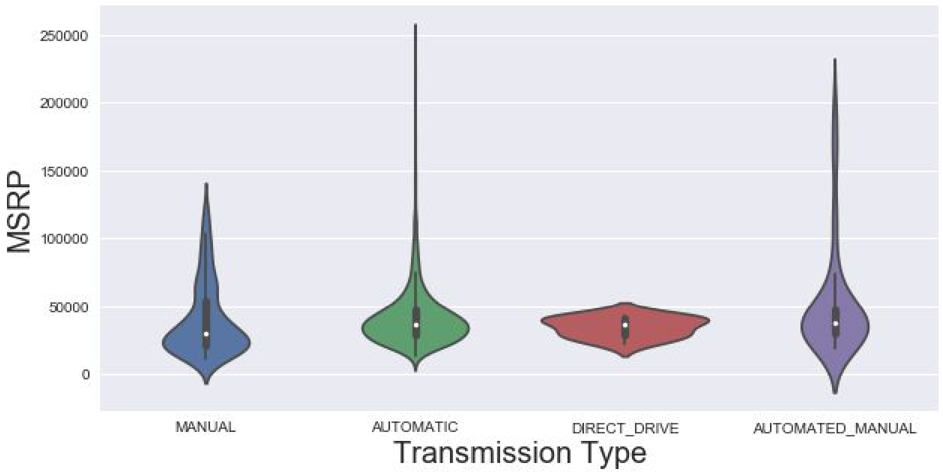
After visualizing of the MSRP distributions by make, I visualized the mean, median, and standard deviation summary statistics as shown below.

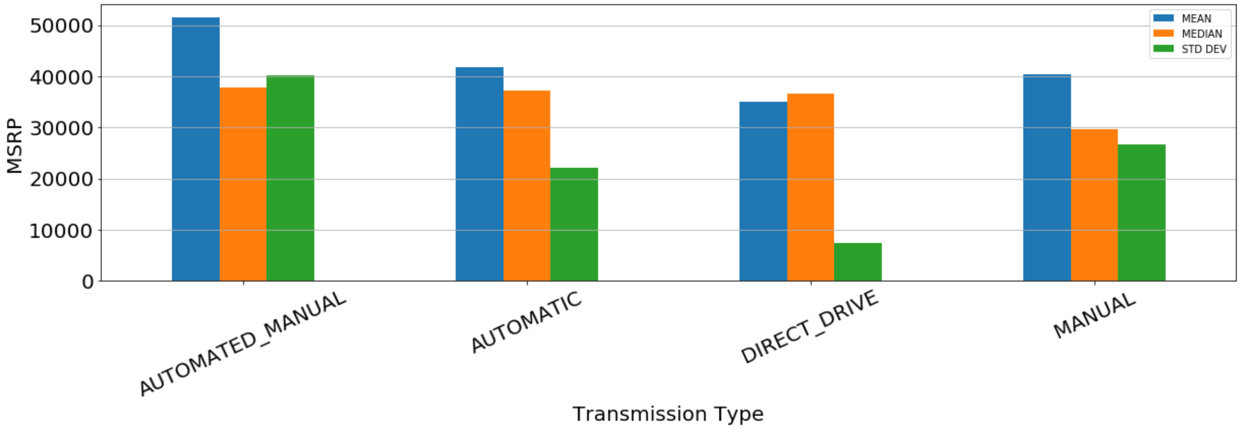




The luxury brand vehicles have a higher mean and median than economy or compact brands. Mercedes has the largest spread, or standard deviation, of all the car brands in the data set. This makes sense because although Mercedes is a luxury car company, it manufactures compact and economy cars, as well as high-end luxury cars.

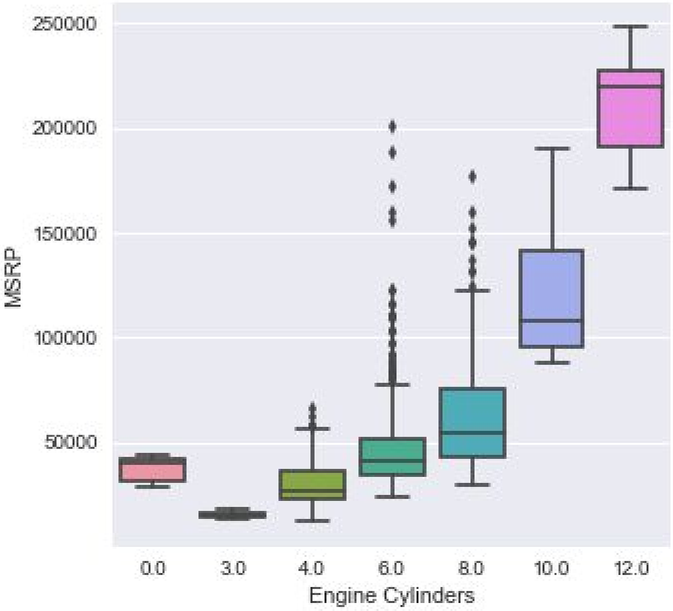
Next, we look at MSRP with respect to transmission type with the following visualizations below.





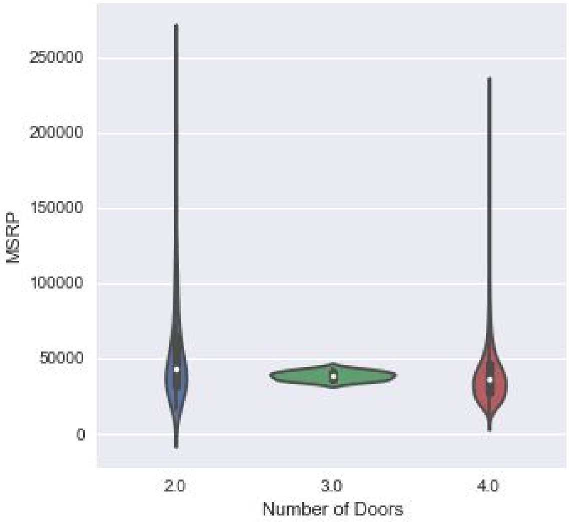
When categorizing MSRP by transmission, we see that manual transmission has the lowest MSRP, with respect to median. It's best to observe the medians with respect to transmission type because there may be outliers (high MSRP vehicles) that can skew the average MSRP. It can be concluded that more sophisticated transmission types tend to raise the median MSRP of a vehicle. Statistically speaking, it may be your best bet to choose a manual transmission vehicle, if you plan to save on your purchase.

Next, we visualize MSRP by number of cylinders as shown below.



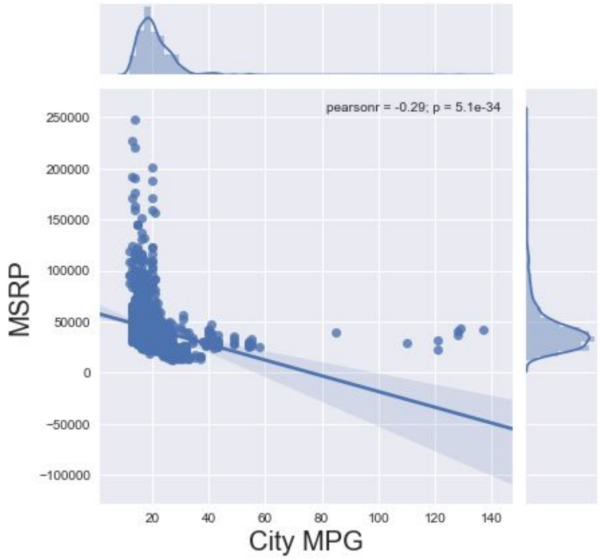
One thing to point out before discussing this plot is that the 0.0 engine cylinder category implies that the vehicles in this distribution are electric. An interesting observation seen here as well is the more engine cylinder a vehicle has, the higher the mean, and median pricing it is with the exception of the electric vehicles. Vehicles with more engine cylinders tend to be your sports and performance vehicles, which is already expensive to begin with. For example, the 10 cylinder vehicles in the data set are exclusively Audi R8 and Dodge Vipers, and the 12 cylinder vehicles are exclusively Mercedes-Benz Maybach, S-Class, and SL-Class. These cars will be a nice accessory if the consumer has a budget well over $100k to spend.

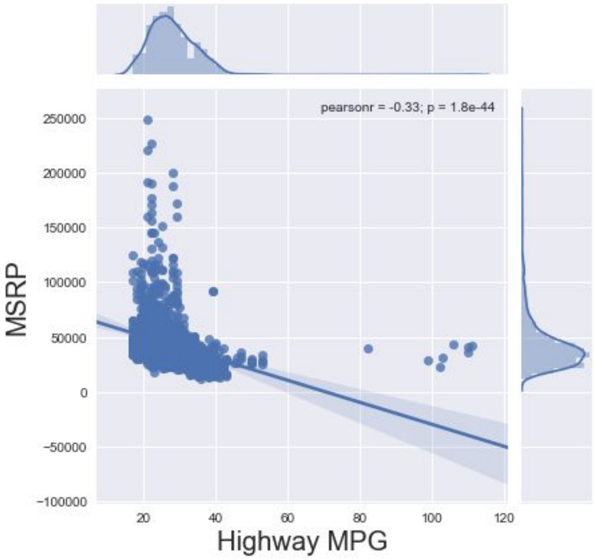
The violin plots shown below summarize the distributions of MSRP with respect to number of doors.

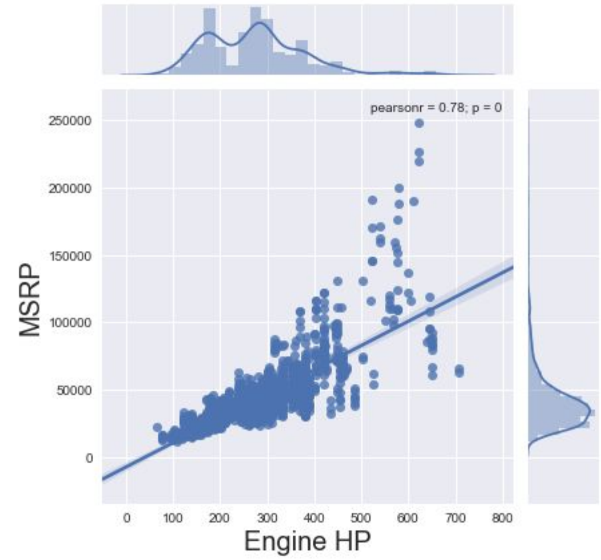


One surprise was the 3-door category. After deep diving into the data set, what was uncovered was the the 3-door car belonged was exclusively the Ford Transit wagon, which has two doors for the driver and passenger seat, and a sliding door on the right hand side for additional passengers. From uncovering this, the narrow distribution in the 3-door category makes sense since a particular make and model is not expected to have too wide of a price range. The 2-door vehicles have a higher median and mean pricing than the 4-door vehicles and also have a much wider spread in pricing. Therefore, it would be recommended to save on MSRP by going for a 4-door vehicle instead of a 2-door.

Next, I found significant correlations between independent numerical values. In particular, MSRP vs. MPG (city and highway) and MSRP vs. Engine HP. I then used the jointplot method provided by Seaborn to show the scatter plot and the regression line.







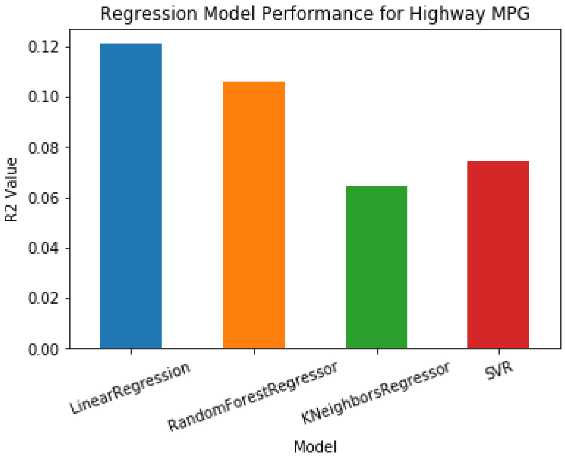
The first row of plots display the relationships between MSRP and city/highway MPG. By looking at the regression lines, there is a weak negative correlation between MSRP and MPG. More specifically:

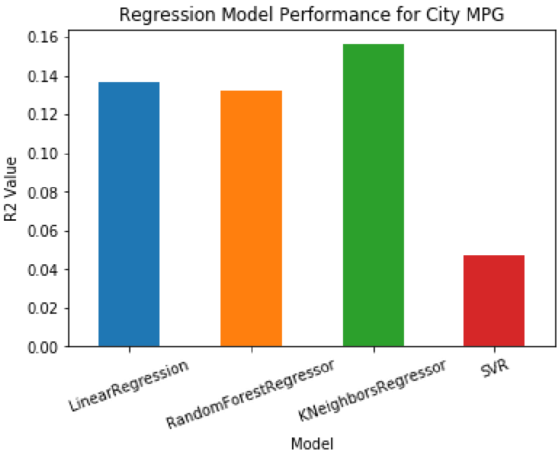
MSRP vs. City MPG	R=-0.29	R^2=0.0841
MSRP vs. Highway MPG	R=-0.33	R^2=0.1089

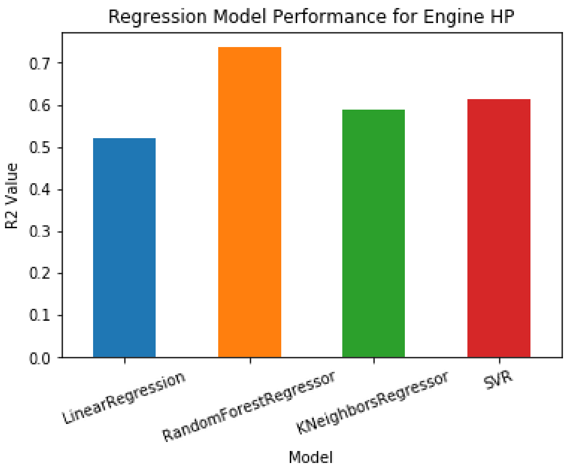
Based off of the correlation coefficients, we can see that correlations are weak. Correlation coefficients come with an associated coefficient of determination (R^2). In this case, our R^2 values are significantly low, which means there is a very low goodness of fit for the given data points on the regression lines. The second row, and only plot on that row, shows the relationship between MSRP and engine horsepower. Unlike the previous relationships, this one has a strong positive correlation with R=0.78 and a goodness of fit of R^2=0.6084. From this, we can infer that a vehicle with low engine horsepower will have a more attractive price than an engine with higher horsepower. The vehicles in the higher horsepower range are usually sports and performance vehicles, which, by nature, are more expensive.

VI. Machine Learning

Next, we dive into the machine learning aspect of analyzing the car dataset. At this stage, we are interested in predicting the price of a car given its features. Before doing so, I removed significant outliers from the set by filtering the data to have entries which only had a highway MPG value of 60 or less. I will be comparing various regression models to see which one will work best for predicting a car's MSRP. Since I chose to use regression to predict car pricing, I narrowed the features down to those that have numerical values. These features are: engine HP, highway MPG, and city MPG.

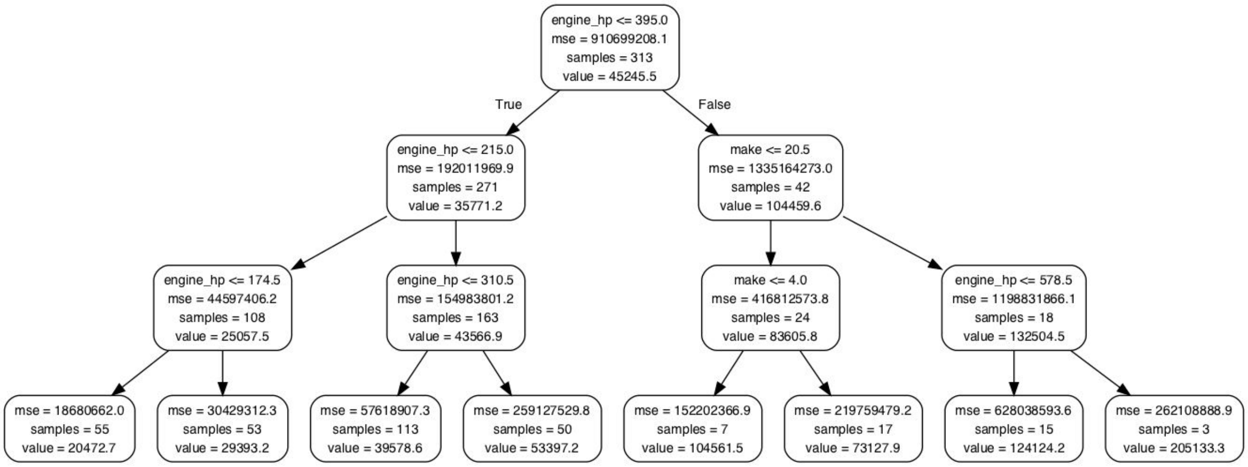




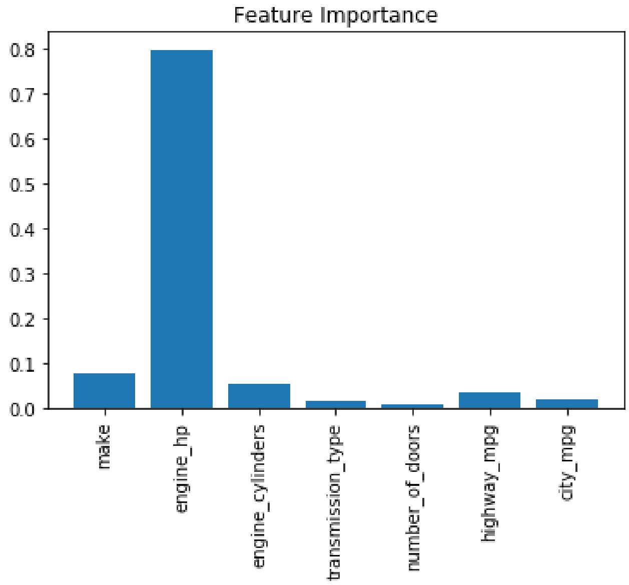


In the bar plots above, I compared the R^2 scores of the following models: Linear Regression, Random Forest Regression, K Nearest Neighbors Regression, and Support Vector Machine Regression. Overall, the Random Forest Regression model has the best overall goodness of fit. The next step now is to apply Random Forest and discover insights the model has to offer. **I have also included the following categorical features into this model: Make, Number of Cylinder, Number of Doors, and Transmission type*

The model I used has 1000 estimators and uses 30% of the data set as training data. This Random Forest Regressor yielded an R^2 value of 0.874, a root mean squared error (RMSE) of $7599.96, and a mean absolute error of $5000 for car pricing. The entire decision tree for the first estimator can be found here, but for a simplistic visualization, I provide the following decision tree here:



This tree comes from a similar Random Forest Regression model, however the only differences are that it has 10 estimators and the maximum depth of the decision trees are constrained to 3. This was conducted only to provide an easy to read decision tree to exemplify our original model's process. Looking at the decision tree from the first estimator, we see that in each node, the condition of a feature determines values for the mean squared error, number of samples, and the value of interest (in our case, the MSRP). A price prediction is determined once a leaf in the decision tree is reached.



This bar plot tells us which features are the most important in making a prediction on MSRP. What we see here is that engine_hp plays the most significant role by far in MSRP prediction.

We will now look at some examples of applying the model to answer the questions mentioned in the beginning of this report:

1) I want a car with this set of features. What is my price point?
2) Between Kia and Toyota, which brand has the more competitive edge?
3) Is Mercedes more expensive than one of its competitors, BMW?

A solution to question 1 will be by simply applying the Random Forest Regressor model as demonstrated below.

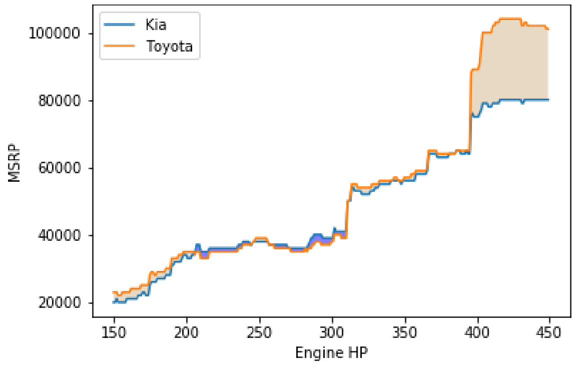
Input:

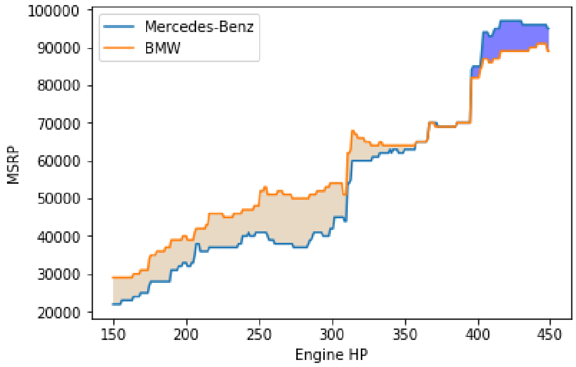
make	=	Kia
engine_hp	=	150
engine_cylinders	=	4
transmission_type	=	AUTOMATIC
number_of_doors	=	4
highway_mpg	=	40
city_mpg	=	35

Output:

$20,000

To answer questions 2 and 3, a single prediction output will not suffice; therefore, we need to have a more comprehensive approach to providing insight. Since engine HP plays the most significant role in predicting the MSRP of a car, we can let this feature vary while keeping the other features at fixed value.





As shown in the two plots above the answers to questions 2 and 3 would be that it depends. In the case of Kia vs. Toyota, Kia has an overall competitive edge against Toyota when it comes to pricing vs HP. There are, however, small intervals where Toyota is priced lower, however it's not sufficiently significant to compete against Kia. Whereas in the case of Mercedes-Benz vs. BMW, Mercedes-Benz is the more overall competitive brand until you compare cars that have an engine HP of over 400. In the 400-450 HP interval, BMW has more competitive options.

If a consumer is interested in knowing what price to expect for a desired car, the features to focus on are: make, transmission, number of cylinders, and engine horsepower. Make is important since various car brands target various markets such as economy, sports, and luxury. Looking at various brands such as Honda, Hyundai, Mitsubishi, Toyota, Kia, and Mazda, we see that they provide pricing that is appropriate for the economy car market. Brands such as Porsche and Maserati are priced to target the sports and luxury car markets. Car manufacturing companies can also use this information to know which competitors are in their market. As for transmission type, consumers and manufacturers would save on average for a manual transmission, which makes sense since they are usually cheaper to implement. As for number of cylinders and engine horsepower, it's recommended to minimize these quantities since they have a strong impact on the pricing unless you are in the sports car market. It is also more intuitive as a buyer to focus on engine horsepower feature rather than number of cylinders. City and highway MPG may intuitively feel as if they affect car pricing, however the influence they have is not significant. Overall, my strongest recommendation to buyers and manufacturers is to look at the make and horsepower of the cars in question. These features have the most influence in determining a reasonable and competitive price.