Predicting Car Price

AUTO LEARNING

Dongtao Jiang | Springboard Data Science Career Track | 9/2/2019

Introduction

A customer is always concerned about if the money paid for the product is worth it. Car is a commodity with sophisticated technological wonders built into it. In any kind of modern car, there exist deep knowledge bases in mechanical engineering, electrical engineering, aerodynamics, software engineering, chemical engineering, automation, etc. Nobody has such full understanding and knowledge to make sound judgement of a new car's price. Car value and reputation are usually based on many years user experience, popularity, quality testing, etc. To the majority of car buyers, it would be a great tool to use a data-driven objective model to find the best quality car with all technical features being taken into account. To a customer that doesn't have an engineering background, the following evaluation metrics that are extracted from thecarconnection.com would help.

- Style: Points can be earned or lost based on above- or below-average interior and exterior style; excellent or poor interior or exterior style; and exceptional (or very poor) style.

- Performance: Points can be earned or lost based on powertrain performance; ride and handling performance. Exceptionally quick (0-60 mph in less than 5 seconds) or exceptionally slow (0-60 mph in more than 10 seconds) can earn or lose an additional point. An additional point can be awarded (or lost) for exceptional circumstances, i.e. off-road prowess, or supercar credentials.

- Comfort: Points can be earned or lost based on comfort in the front seats, back seats, or third-row seats (where applicable); good or bad interior storage and cargo capacity; and good fit and finish.

- Safety: Cars with official crash data gain points for a five-star overall rating by the NHTSA, or Top Safety Pick/Top Safety Pick+ status by the IIHS. An additional point is awarded for cars that come standard with full-speed automatic emergency braking. We award points for excellent outward vision and for abundant safety features and options such as parking assistance, surround-view camera systems, or driver-assistance features. Cars with official crash data lose points for a four-star overall rating by NHTSA, any "Marginal" IIHS or three-star NHTSA ratings, for poor outward vision, and when they lack forward-collision warnings and automatic emergency braking.

- Cars without crash data aren't given a rating at all. Cars with only partial ratings may be scored, generally when it improves their score.

- Features: Cars with excellent base equipment earn a point above average. Extra points can be added for exceptional available features, good value, good infotainment systems with screens larger than 7.0 inches, and good warranty or

service programs. Cars may lose points for substandard or expensive features; bad feature packages; poor relative value; or bad warranty or service availability.

- Green: Cars are assigned a rating based on their EPA-estimated highway and combined mileage ratings. Plug-in and battery-electric vehicles start at 9. Electric-only cars with a range of more than 200 miles earn a score of 10. All other vehicles are sorted on a sliding scale based on EPA fuel economy.

Dataset

The new cars dataset was obtained
https://www.reddit.com/r/datasets/comments/b6rcwv/i_scraped_32000_cars_including_the_price_and_115/

It was originally scraped from thecarconncetion.com that provide comprehensive care specs and prices. According to its website, the Car Connection is an automotive property of Internet Brands, which owns and operates the largest network of car buying and financing resources in North America, including CarsDirect, Motor Authority, Green Car Reports, and Auto Credit Express.

Data Wrangling

- Rename column names shorter and regularized
- Removing Extraneous Data
    - Columns with 100% missing values
    - Remove duplicate rows
    - Remove columns with only one unique value
- Prepare target column
    - Drop the few rows that have missing values. Remove '$ 'sign and ','.
- Extract Year and Model from one column
- Define function to convert relevant columns to floats using pd.to_numeric()
- Clean up column 'Displacement'
    - Fill missing data with empty string to avoid error during cleaning operations.
    - replace 'L/...' and strip white spaces right and left
    - Strip leading and trailing white space

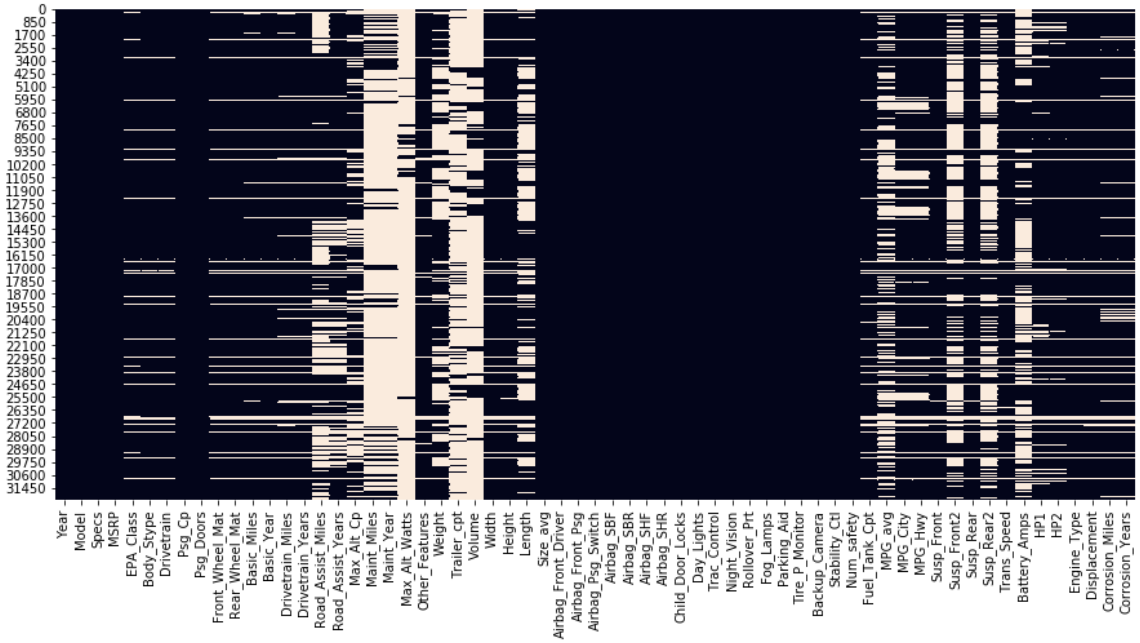
- Use len() to list out other data to be cleaned.
- Romove '/xx'
- Remove ' (152)'
- transform '39.5 Cu.in. Range Extender' to liter
- calculate from cubit inch to liter:

- Two columns are identical. Remove one. ['EPA Class', 'EPA Classification']

Exploratory Data Analysis

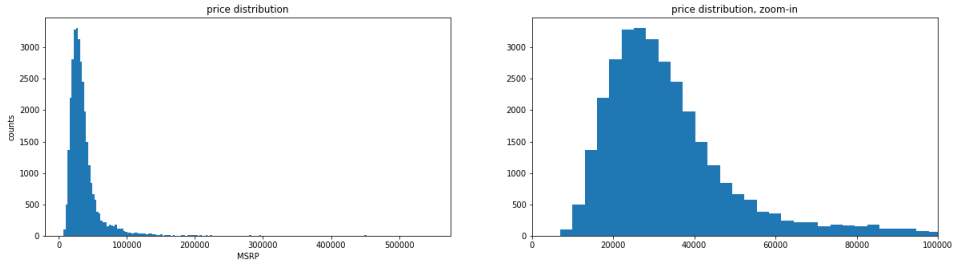
Missing data

How the missing values are distributed in the dataframe can be best visualized by the pair plot function provided by seaborn package. The white area in the figure below represents missing entries. Columns like Volume and Trailer_cpt are mostly missing.

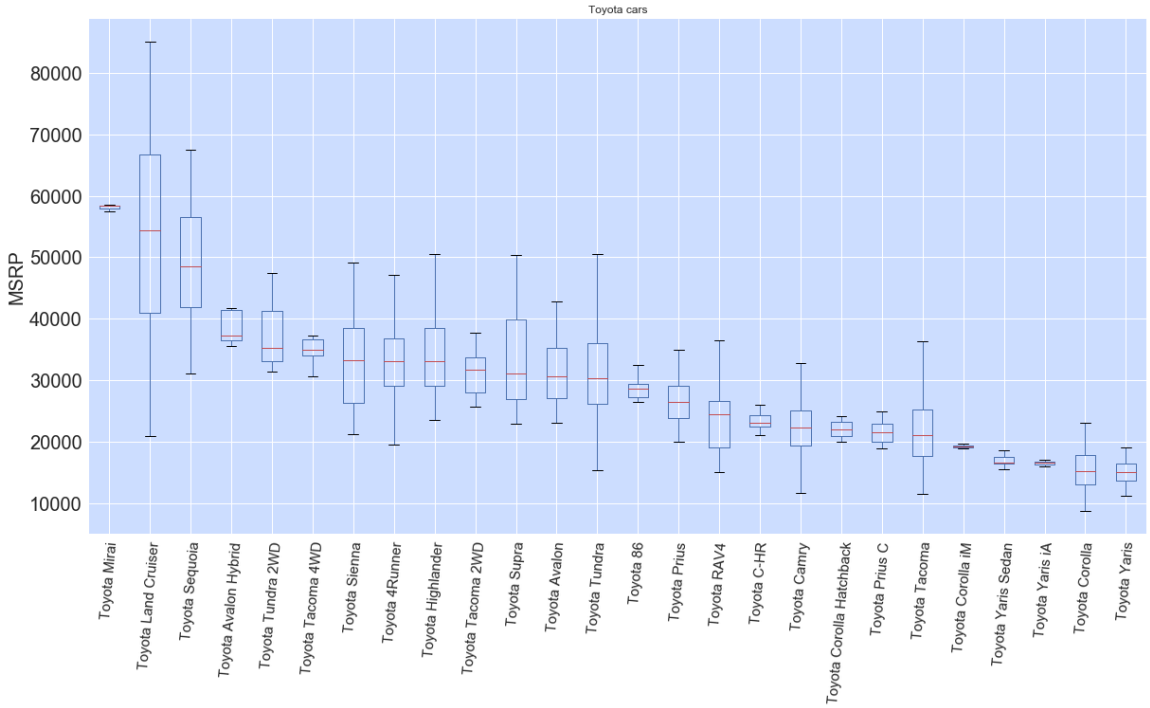


Overall price distribution

The overall price distribution of all car models and years is shown below. The most frequent prices are around $29295.0.

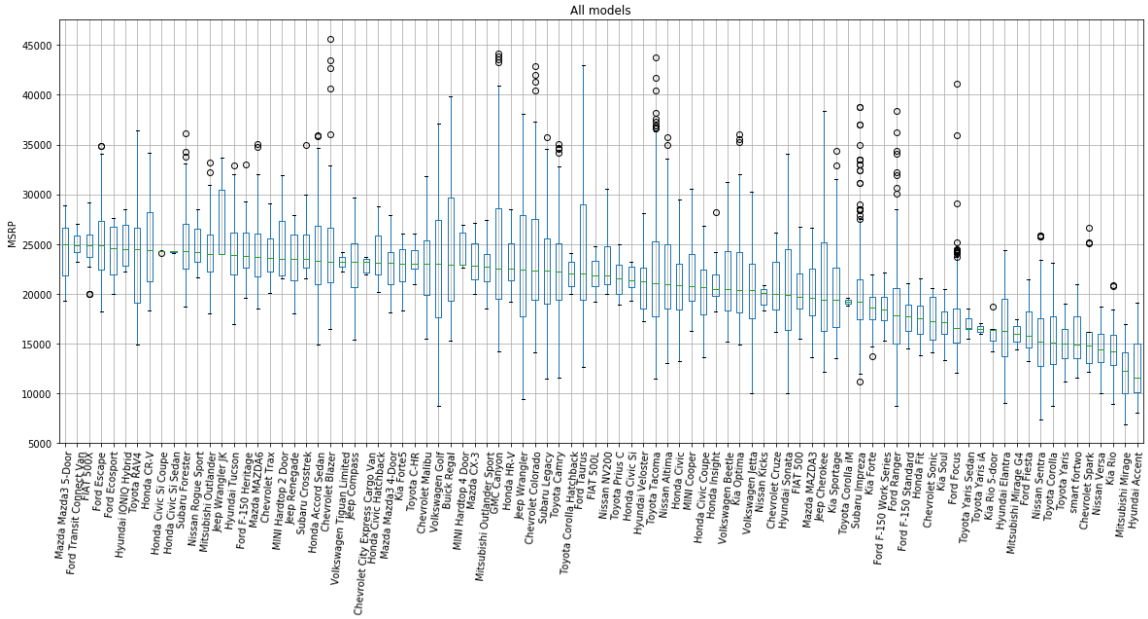


The high-end and low-end car models from Toyota are shown in the figure below with median prices of each model being sorted. Luxury Toyota models include Mirai, Land Cruiser, Sequoia, etc. Low-end models include Yaris, Corolla, etc.



Buying guide for budget-tight customers

For budget-tight customers, the following box plots provide a guide on models the prices are below $25,000. The order is based on the median price of each model.

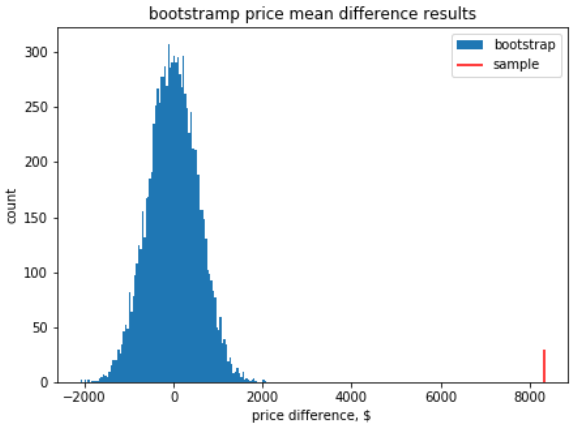


Hypothesis testing on two low-end car models

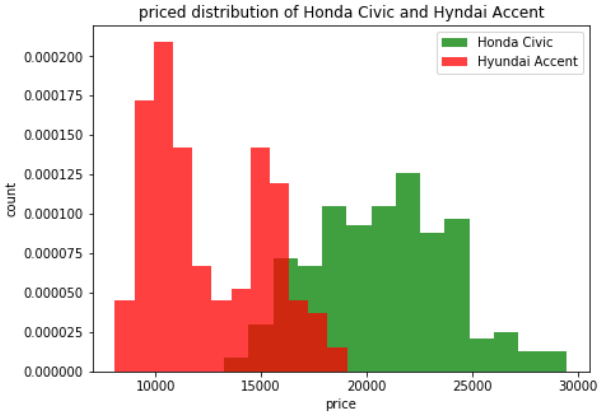
It was told that both Honda Civic and Hyundai Accent are good quality cars with good awesome deals. The box plot above shows the two models have large overlap in thier price range. Is the difference in their mean price significant or negligible? To answer the question, the following hypothesis testing is conducted.

To quantify the difference of mean of the prices of the two models statistically, Bootstrap approach seems to be appropriate.

- Ho: There is no difference in the mean price between Hyundai Accent and Honda Civic.
- Ha: There is obviouse difference in the mean price between Hyundai Accent and Honda Civic.
- alpha: 5%

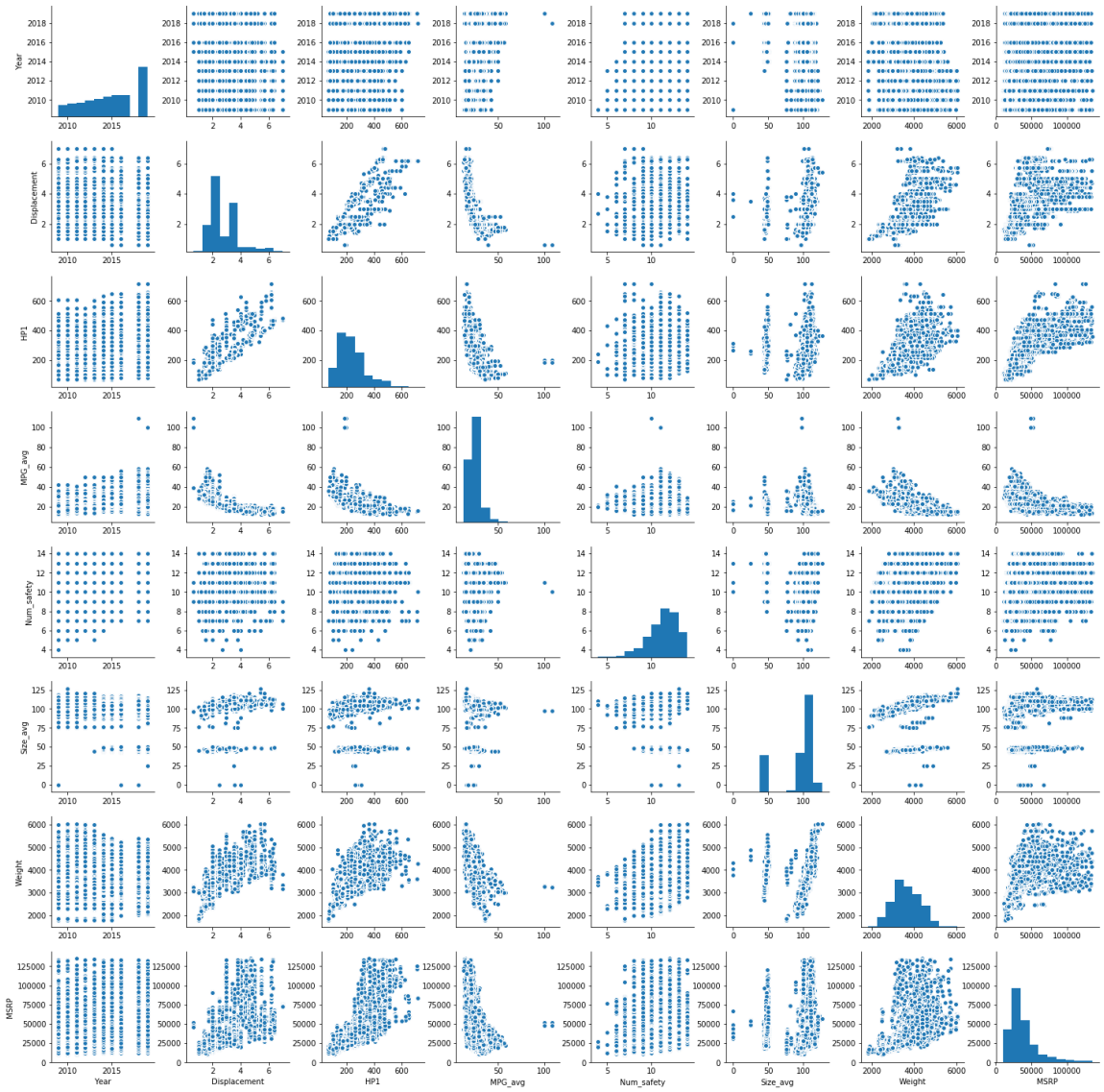


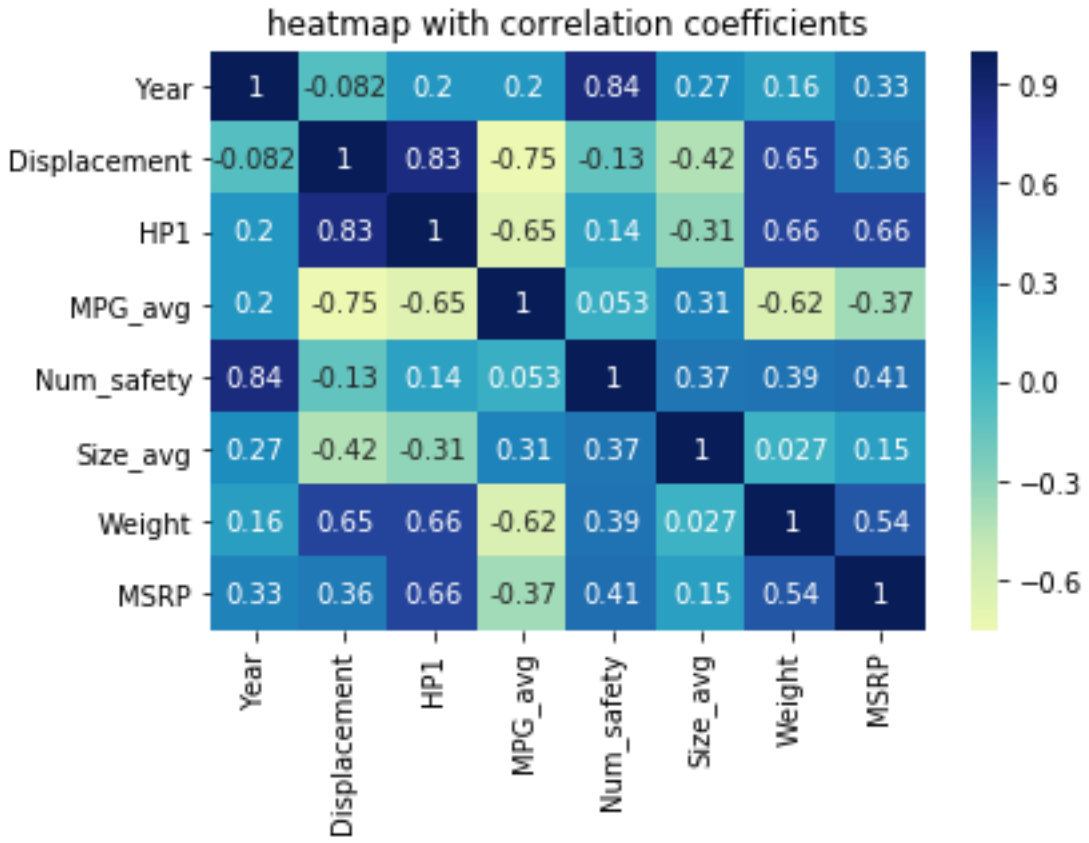
p-value is zero, therefore we reject the null hypothesis and approve the sharp difference in mean. The bootstrap results is plotted in the above figure, strongly supporting there is significant difference in mean. The figure below shows the price distribution of the two models, indicating a clear difference in them.



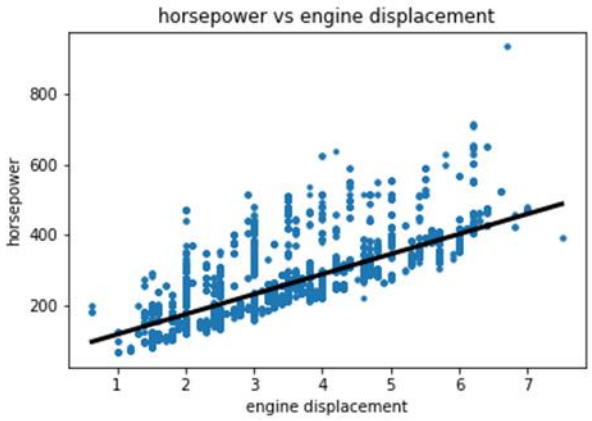
Correlation investigations

Possible correlations between features can be investigated by using the pair plots function couple with heatmap from seaborn, as shown below.

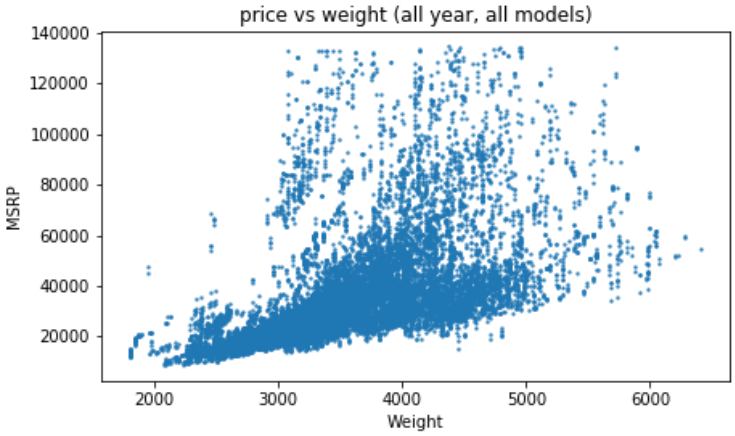

Year
Displacement
HP1
MPG_avg
Num_safety
Size_avg
Weight
MSRP



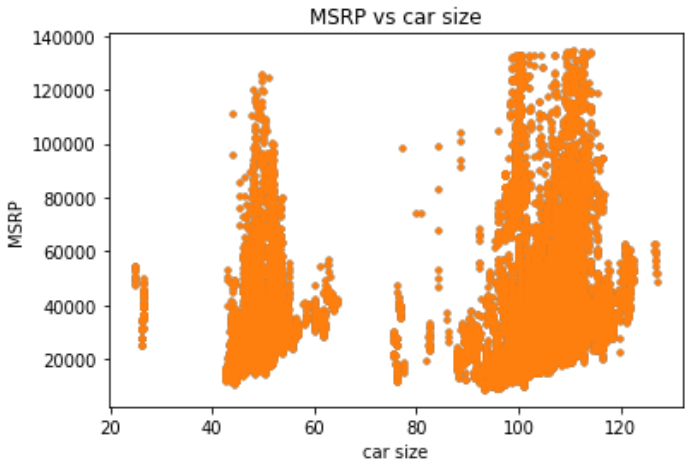
Both the pair plot and heatmap indicate a positive correlation between horsepower and engine displacement, which agrees well with physics and engineering principles. A linear fitting line was is shown in the figure below. Some correlation also exists between size and displacement, horsepower and MPG, price and horsepower, price and weight.



It is interesting to notice that the bottom one in the price range with the same weight is almost linearly proportional to price, as shown in the figure below. ¶



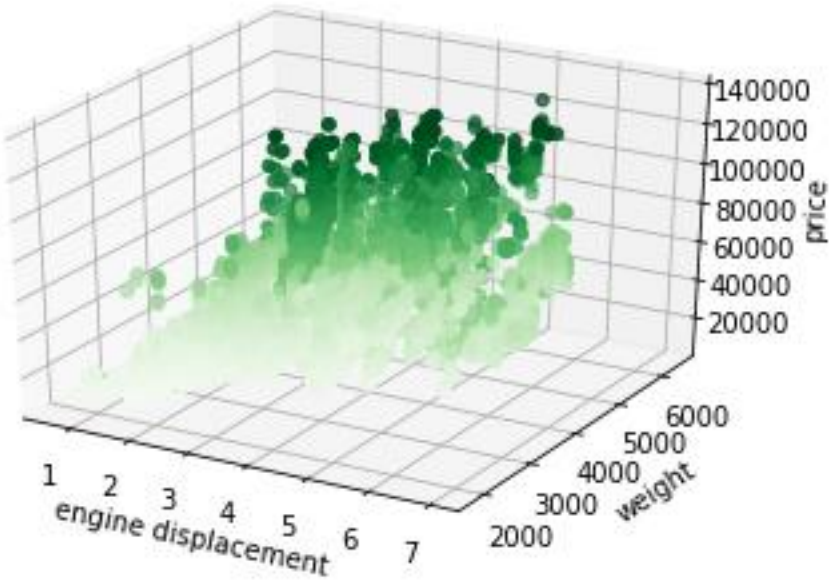
The influence of car sizes on the price are aggregated or clustered. There is no obvious trend for the price related to size.

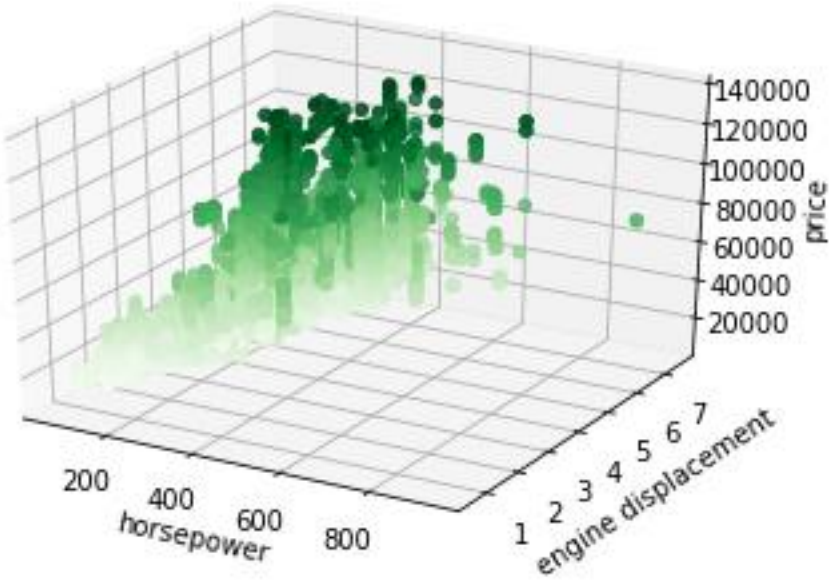


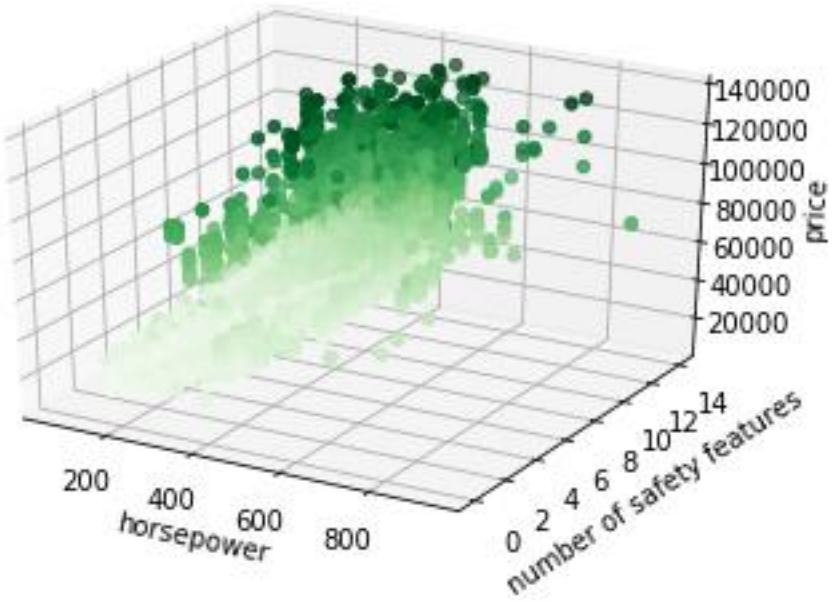
Explore correlations using 3-D Plots

By exploring combination of any two different features to see their effect on MSRP in the 3D plots, it looks like the following features are well correlated to the car price,

- horsepower
- weight
- engine displacement
- number of safety features


140000
120000
100000
80000
60000
40000
20000
price
6000
5000
4000
3000
2000
weight
1
2
3
4
5
6
7
engine displacement


140000
120000
100000
80000
60000
40000
20000
price
200
400
600
800
horsepower
1
2
3
4
5
6
7
engine displacement


140000
120000
100000
80000
60000
40000
20000
price
200
400
600
800
horsepower
0
2
4
6
8
10
12
14
number of safety features