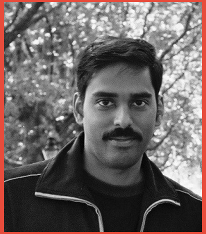
Automobile Data - Exploratory Analysis



Justin Jose

git https://github.com/justinpolackal/eda-automobiles

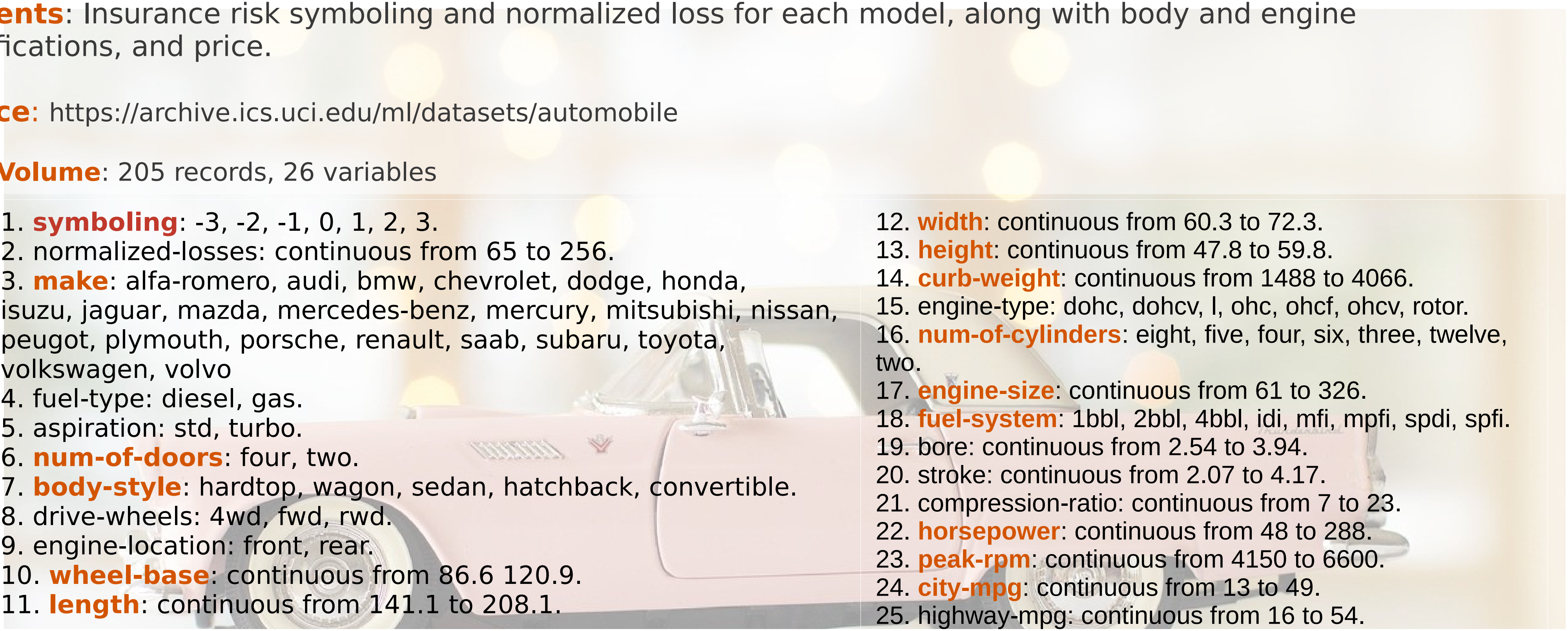
in www.linkedin.com/in/justinpolackal

About the Data

Contents: Insurance risk symboling and normalized loss for each model, along with body and engine specifications, and price.

Source: https://archive.ics.uci.edu/ml/datasets/automobile

Data Volume: 205 records, 26 variables



Attribute Information

1. **symboling**: -3, -2, -1, 0, 1, 2, 3.
2. normalized-losses: continuous from 65 to 256.
3. **make**: alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo
4. fuel-type: diesel, gas.
5. aspiration: std, turbo.
6. **num-of-doors**: four, two.
7. **body-style**: hardtop, wagon, sedan, hatchback, convertible.
8. drive-wheels: 4wd, fwd, rwd.
9. engine-location: front, rear.
10. **wheel-base**: continuous from 86.6 120.9.
11. **length**: continuous from 141.1 to 208.1.
12. **width**: continuous from 60.3 to 72.3.
13. **height**: continuous from 47.8 to 59.8.
14. **curb-weight**: continuous from 1488 to 4066.
15. engine-type: dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
16. **num-of-cylinders**: eight, five, four, six, three, twelve, two.
17. **engine-size**: continuous from 61 to 326.
18. **fuel-system**: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. bore: continuous from 2.54 to 3.94.
20. stroke: continuous from 2.07 to 4.17.
21. compression-ratio: continuous from 7 to 23.
22. **horsepower**: continuous from 48 to 288.
23. **peak-rpm**: continuous from 4150 to 6600.
24. **city-mpg**: continuous from 13 to 49.
25. highway-mpg: continuous from 16 to 54.
26. **price**: continuous from 5118 to 45400.

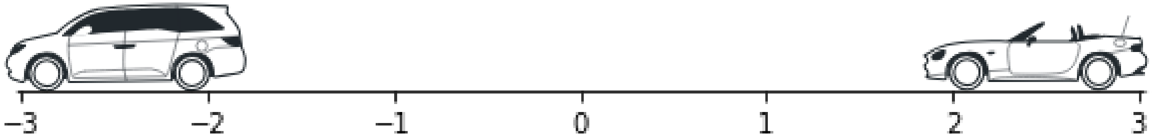
Data Preparation Steps

Step	Details
Variable Identification	Symboling, Price, and other necessary variables to support initial hypothesis
Univariate Analysis	Refer Notebook (GitHub)
Bi-variate analysis	Refer Notebook (GitHub)
Treating missing values	Imputed missing/non-numeric values with mean of the group for continuous and mode of the group for categorical variables
Detecting,analysing and treating outliers	Engine size and horsepower outliers are kept because they represent real world data
Deriving variables	New variables calculated – *isrisky*, *volume* and *sizegroup* for the purpose of analysis.

GitHub Link for Python Notebooks: https://github.com/justinpolackal/eda-automobiles

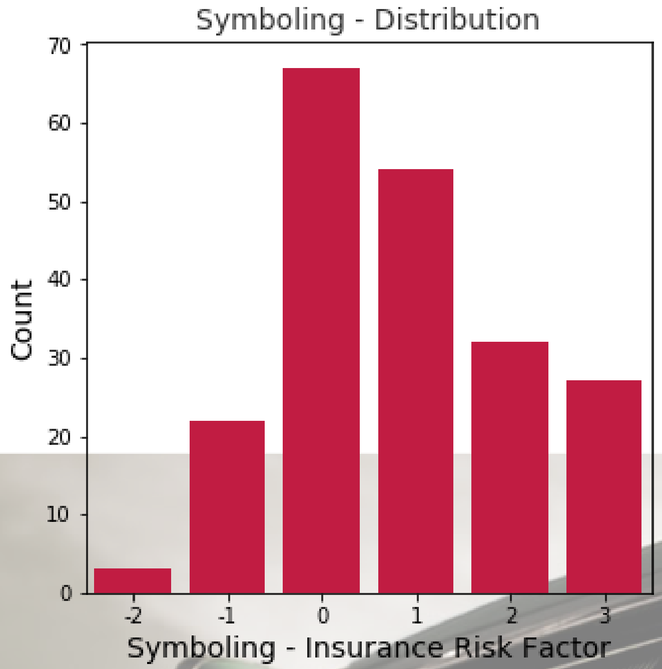
Analysis 1

Does body size influence symboling ?



The study aims to find the relationship between the physical size and styling of the vehicle with symboling.

About Symboling

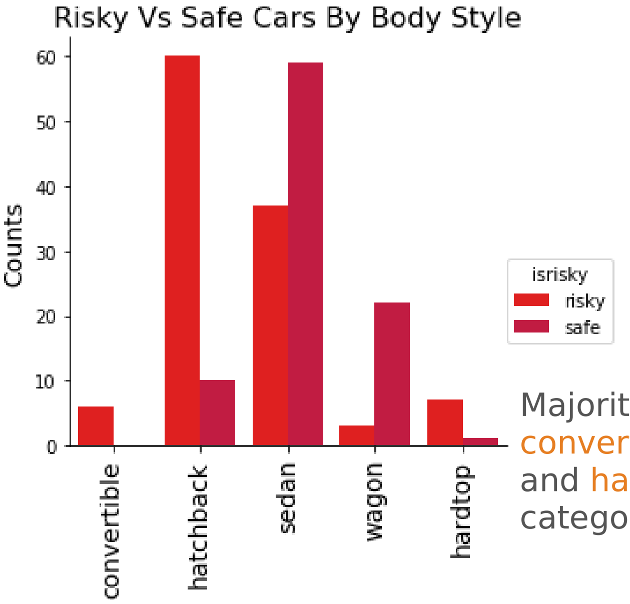


Symboling value shows how risky or safe a vehicle is, from an insurer's perspective. It can range from -3 to +3.

-3 indicates a safe car while +3 denotes a risky one.

*205 vehicle records

Risky / Safe Body Styles

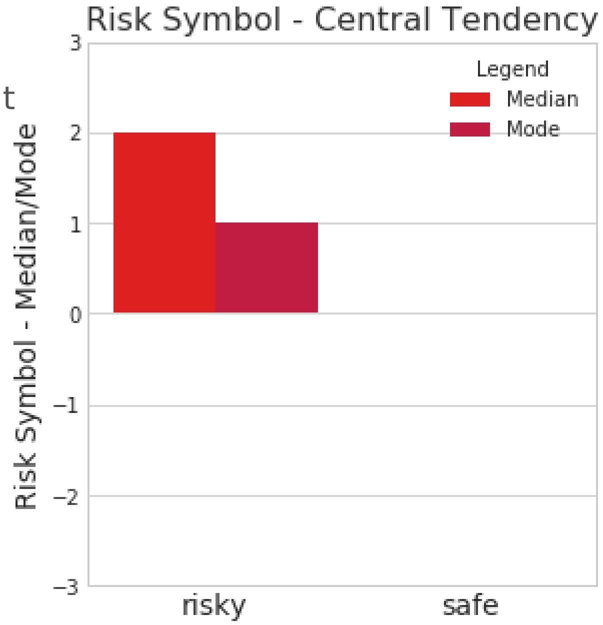


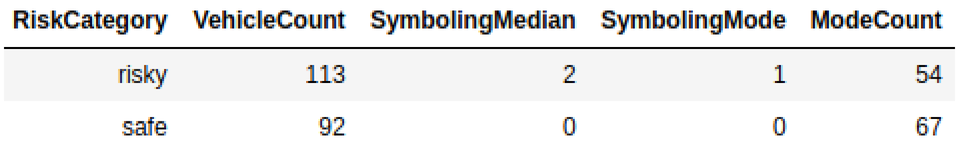
Introduced a new classification - ‘risky’ and ‘safe’

Vehicles falling in the range -3 to 0 are classified as safe, while those in the range +1 to +3 are risky.

Majority vehicles in convertibles,hatchbacks and hardtops are in risky category.

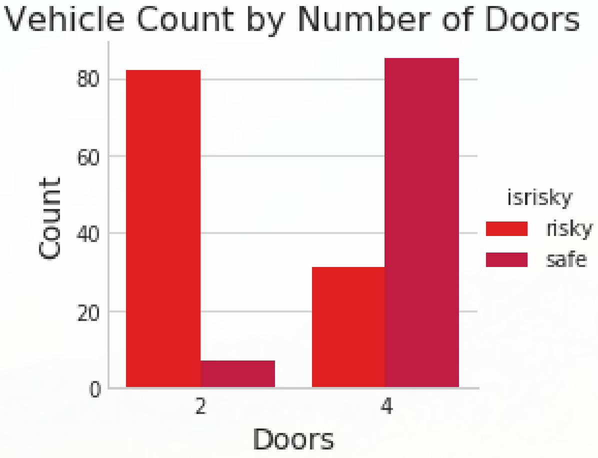
Risky vehicle’s symbol values tend towards 2, while for safe vehicles it is 0





RiskCategory	VehicleCount	SymbolingMedian	SymbolingMode	ModeCount
risky	113	2	1	54
safe	92	0	0	67

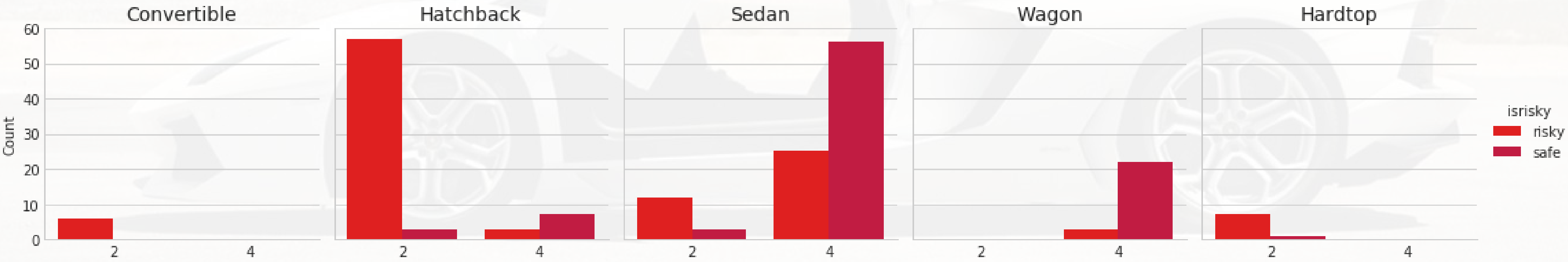
Number of Doors



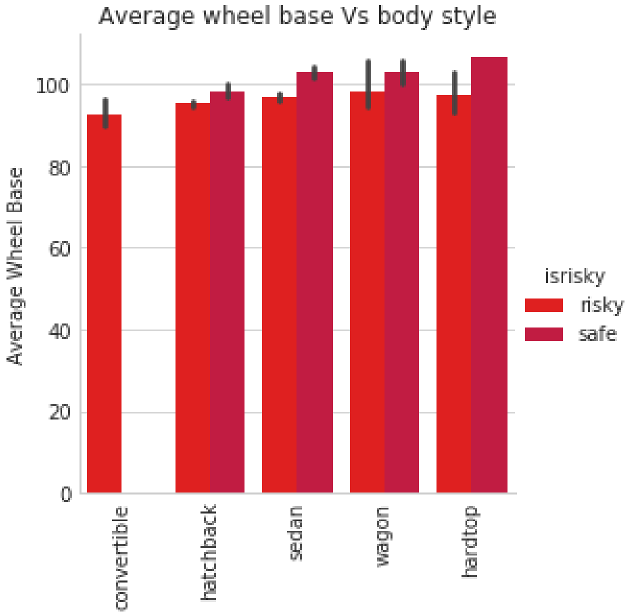
Vast Majority of safe cars have four doors

A two door car is aimed at people who enjoy driving, where as a four door vehicle is meant to carry more passengers. They are more likely to be used to carry families around than a two door version. Needless to say, a car that carries families will be driven with more caution than a driver only car driven by an enthusiast

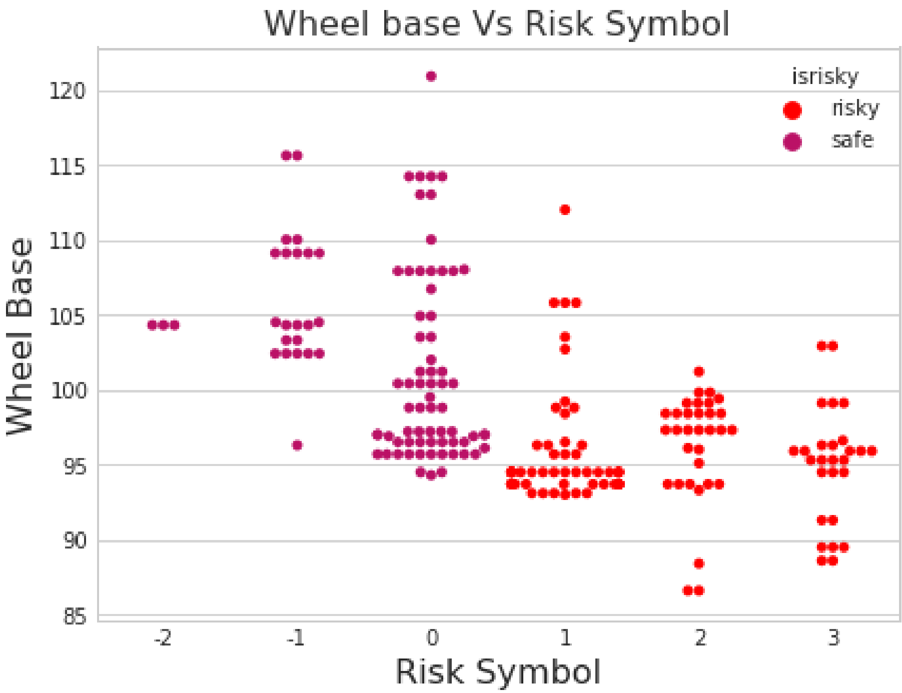
This trend is seen even within each body style.



Wheel Base



As the wheel base decreases, symboling value tends to the risky side.

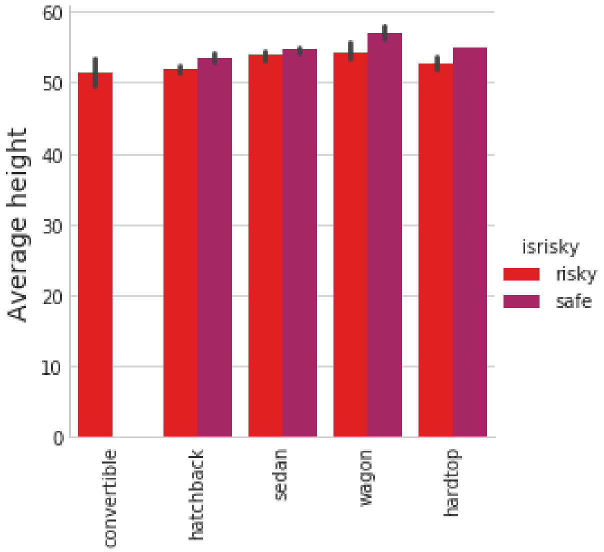


More the wheelbase, more stable the car is, but at the cost of maneuverability.

That means, high speed cornering becomes easier in a shorter wheel base car compared to a longer wheel base one.

Reduced wheel base has a clear correlation with risky cars.

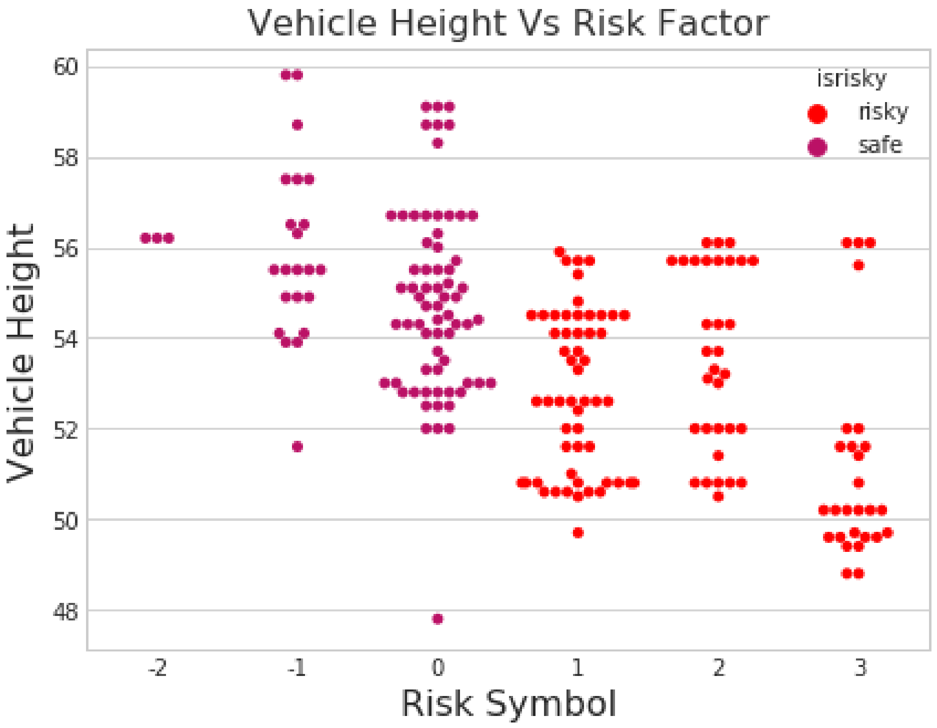
Height



As the height of a car reduces, so does its center of gravity(CG).

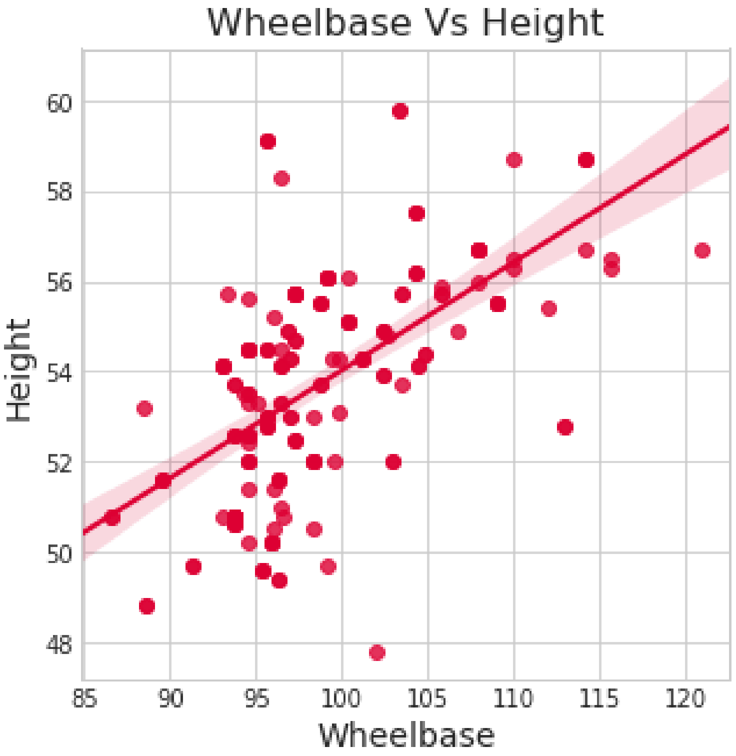
A low CG car stays glued to the tarmac at higher speeds than a taller car with higher CG.

Performance cars aim to keep their CG down, so that drivers can push them to their limits.



Safe vehicles are taller than risky ones.

Wheel base and Height

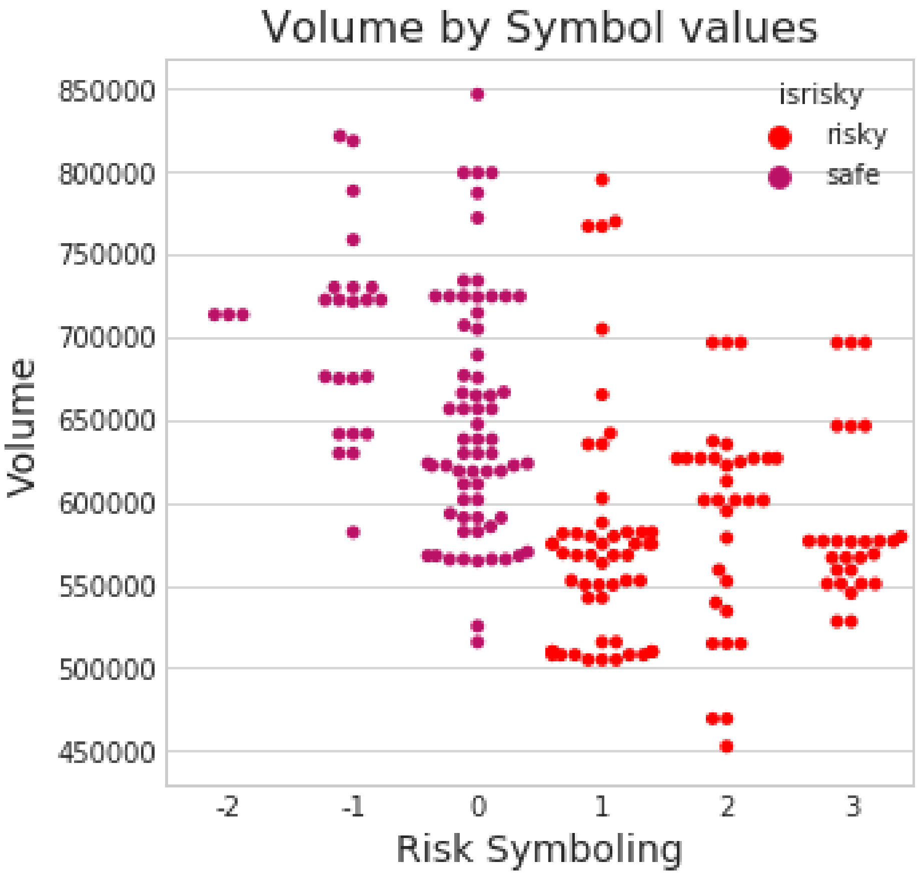


Symbol values tends to risky side as

- Wheelbase reduces
- Height decreases

Wheelbase and Height of a vehicle are correlated.

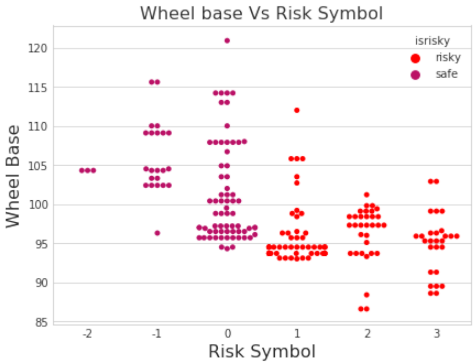
Introducing volume

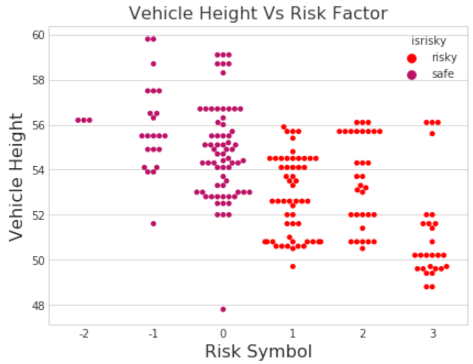


A new variable – volume - is derived from length, height and width of cars.

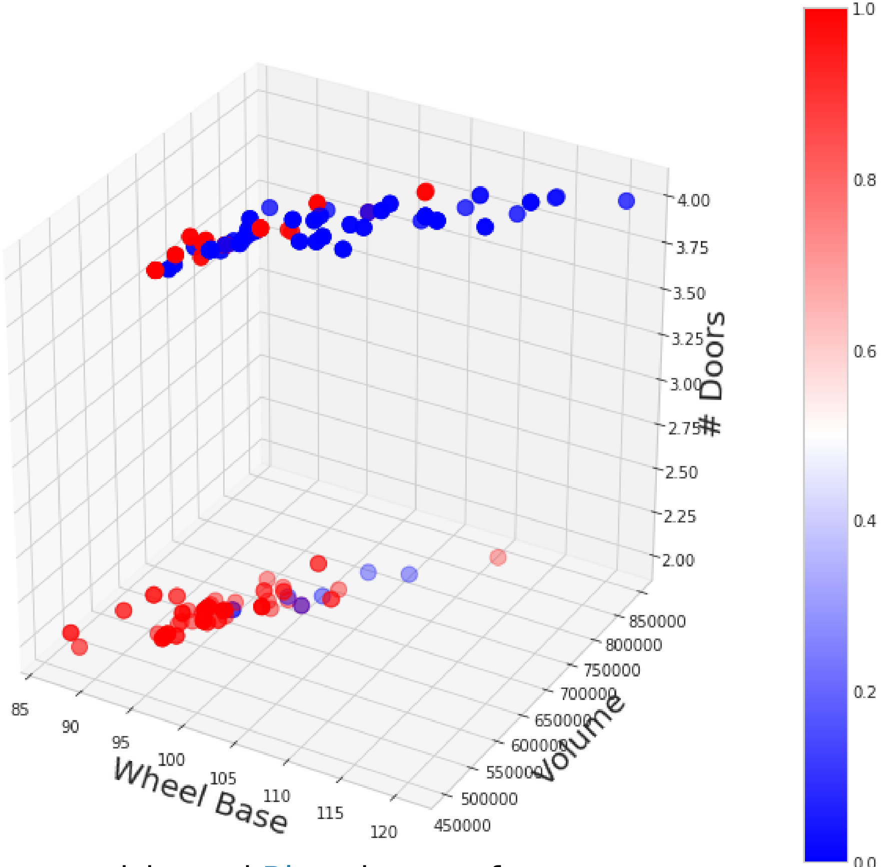
volume=Length x Height x Width

As the vehicle volume reduces, symboling values increases, indicating an increase in risk.





Bringing them all together



Shorter wheel base and lower body height results in lesser volume.

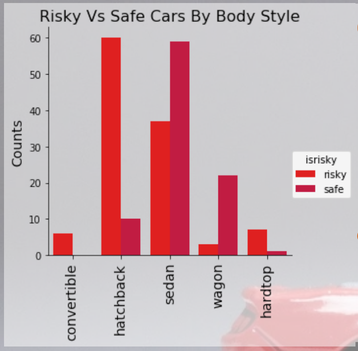
Compared to safer cars, risky ones have:

- Lesser volume
- Two-doors instead of four

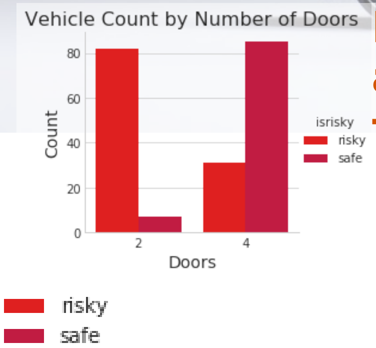
* Red denotes risky and Blue shows safe cars

Analysis 1: Summary

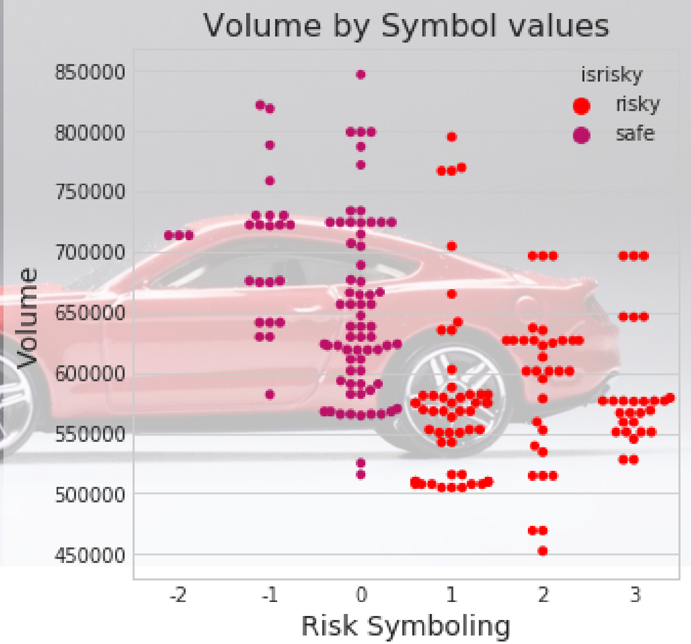
Does body size influence symboling ?



Convertibles, hardtops and hatchbacks are riskier than sedans and wagons



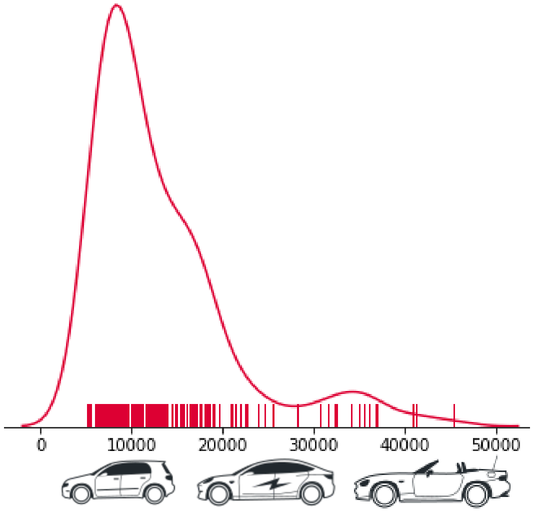
Majority safe cars are found to have four doors.



The volume of a car influences risk symboling. Lesser the volume, higher the risk.

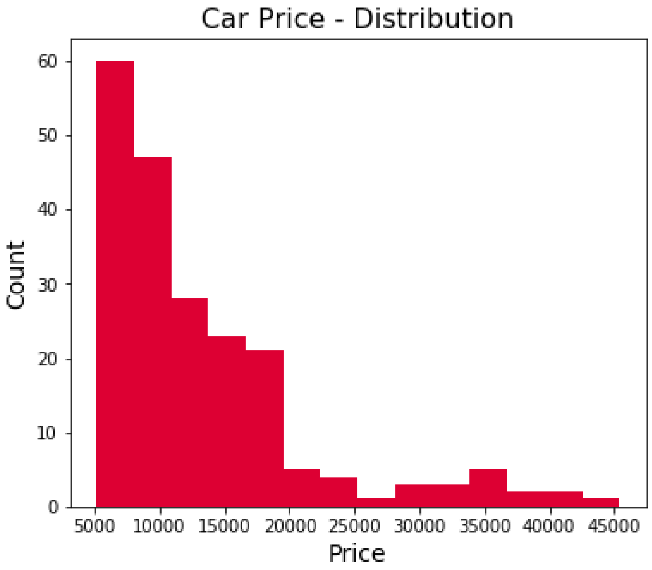
Analysis 2

Do **Body Style, Size and Engine Specs** determine car **prices**?



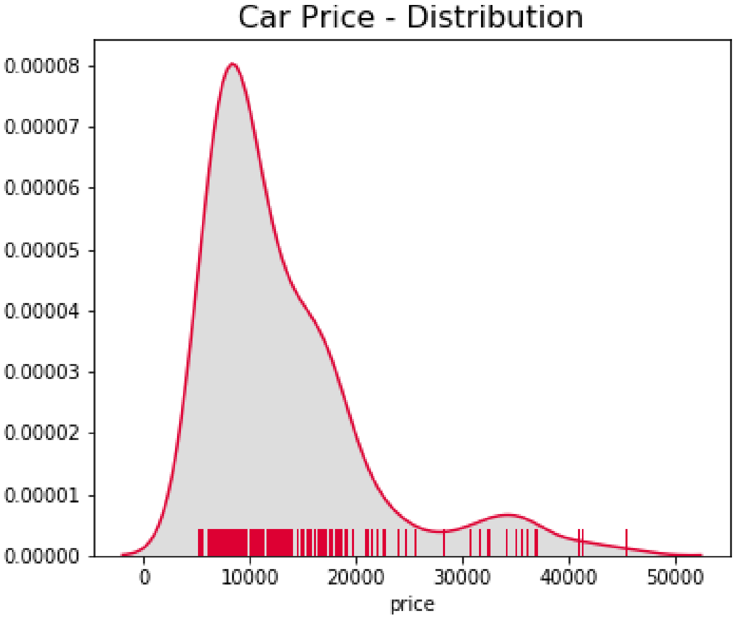
Study aims to find the relationship between car prices and body style, size and engine specs

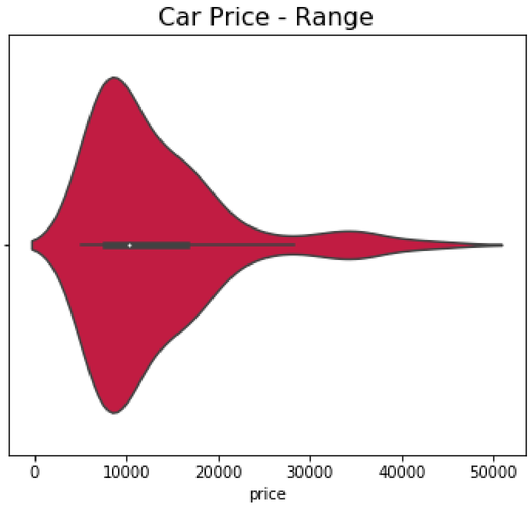
Car Price – Data Distribution



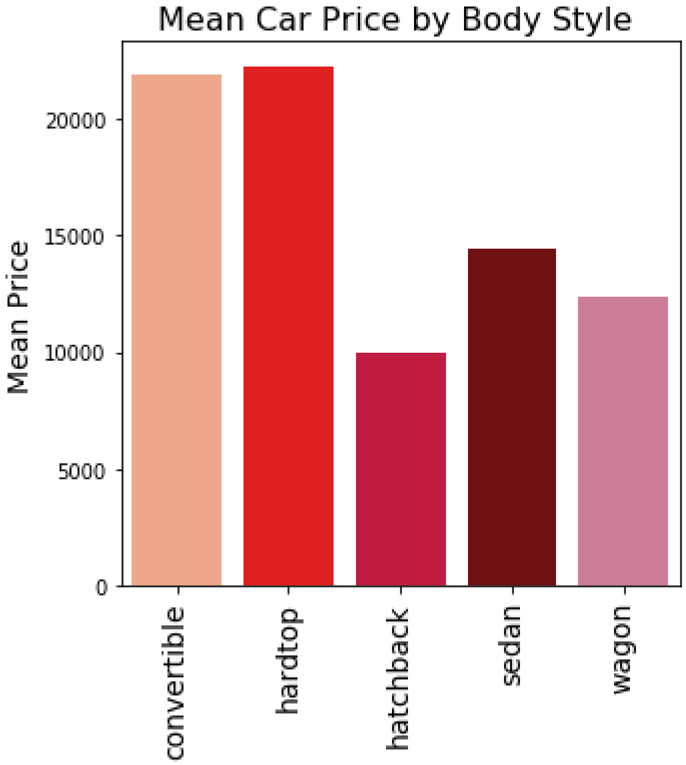
Distribution of Price

Majority of cars belong to the lower price brackets (< 20K) even though there are cars that go up to 45K

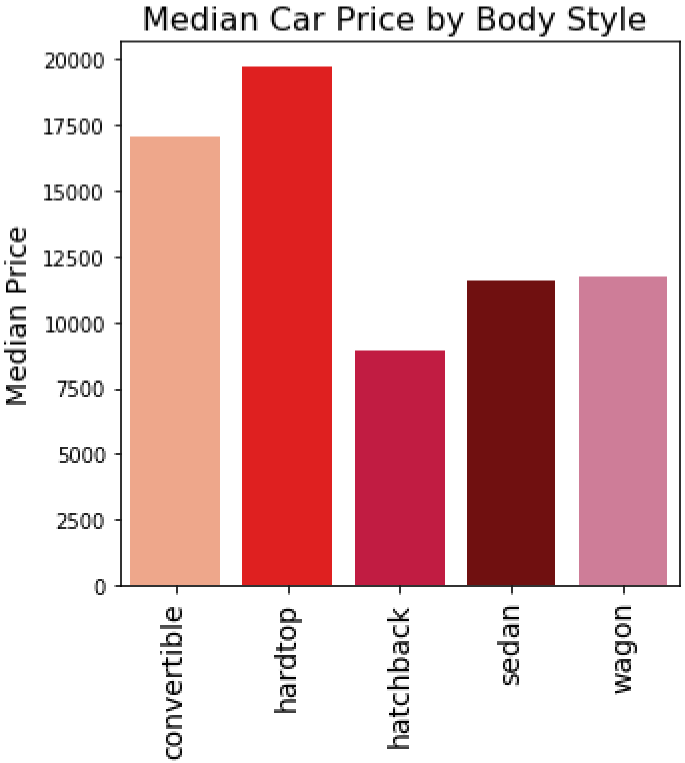




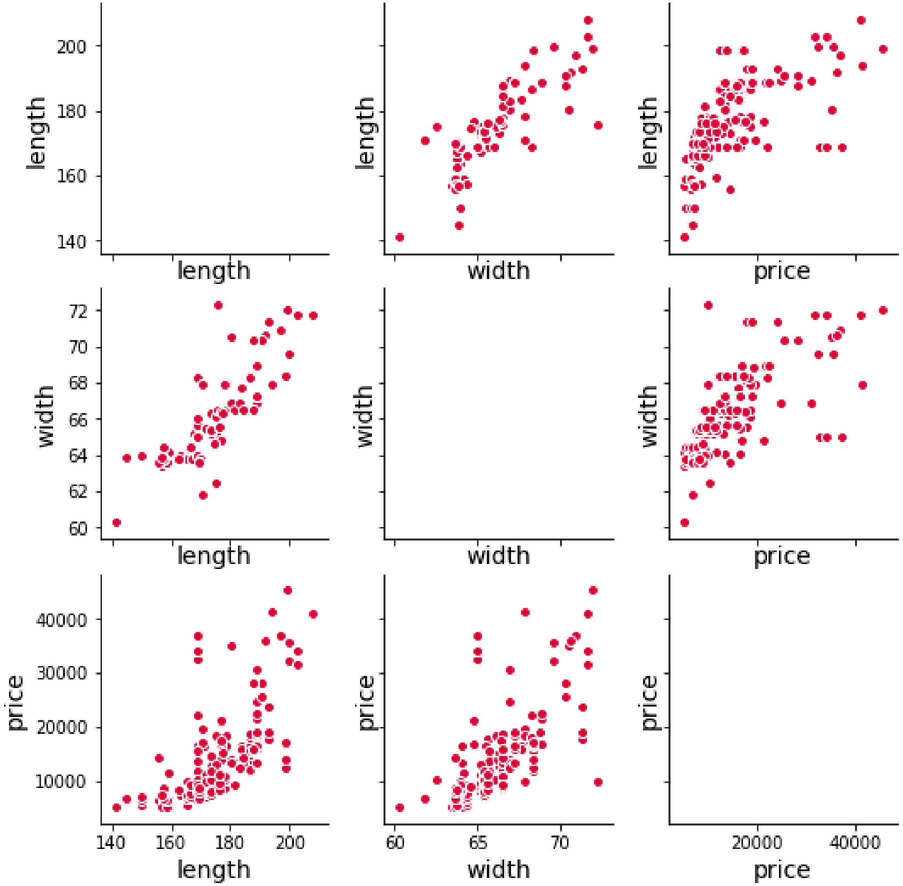
Body Styles and Pricing



Convertibles and hardtops are the costliest car models.



Body Size – Length and Width Vs. Price



Car's length and width have got strong correlations with its price.

US Insurance Institute for Highway Safety Highway Loss Data Institute classifies cars into

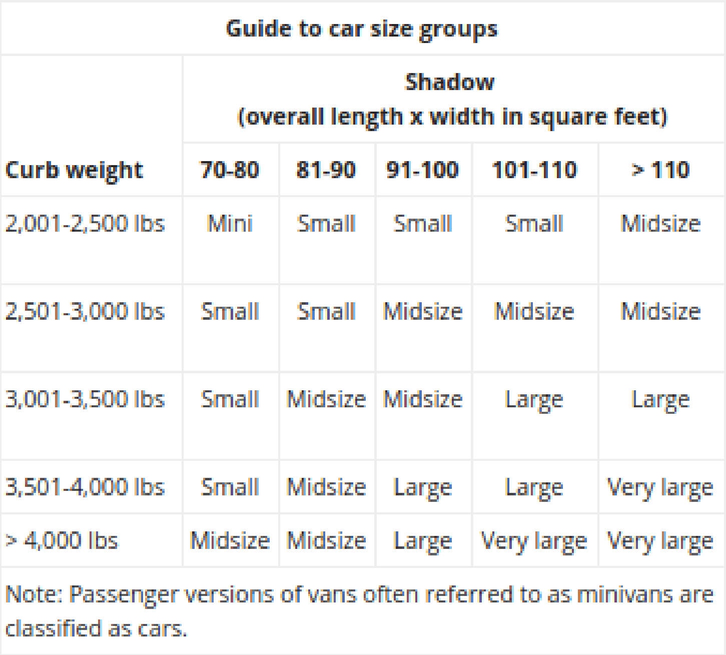
- Mini
- Small
- Midsize
- Large, and,
- Very Large

based on their shadow area (square footage of exterior length × width) and curb weight.

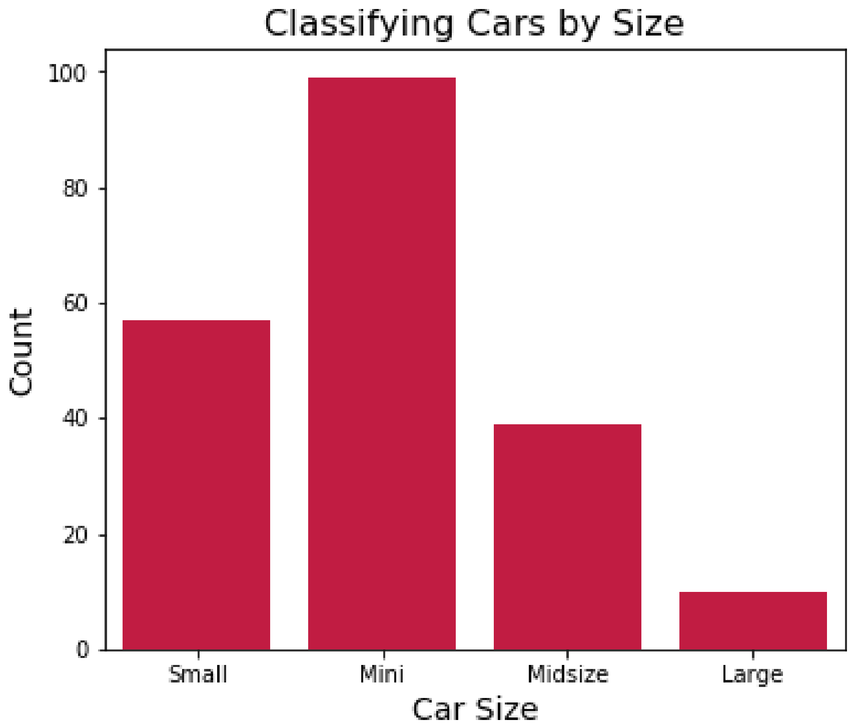
References:
1. https://en.wikipedia.org/wiki/Car_classification
2. http://www.iihs.org/iihs/topics/t/vehicle-size-and-weight/fatalityfacts/passenger-vehicles

Introducing “Sizegroup”

Sizegroup classifies cars based on their shadow area and curb-weight.

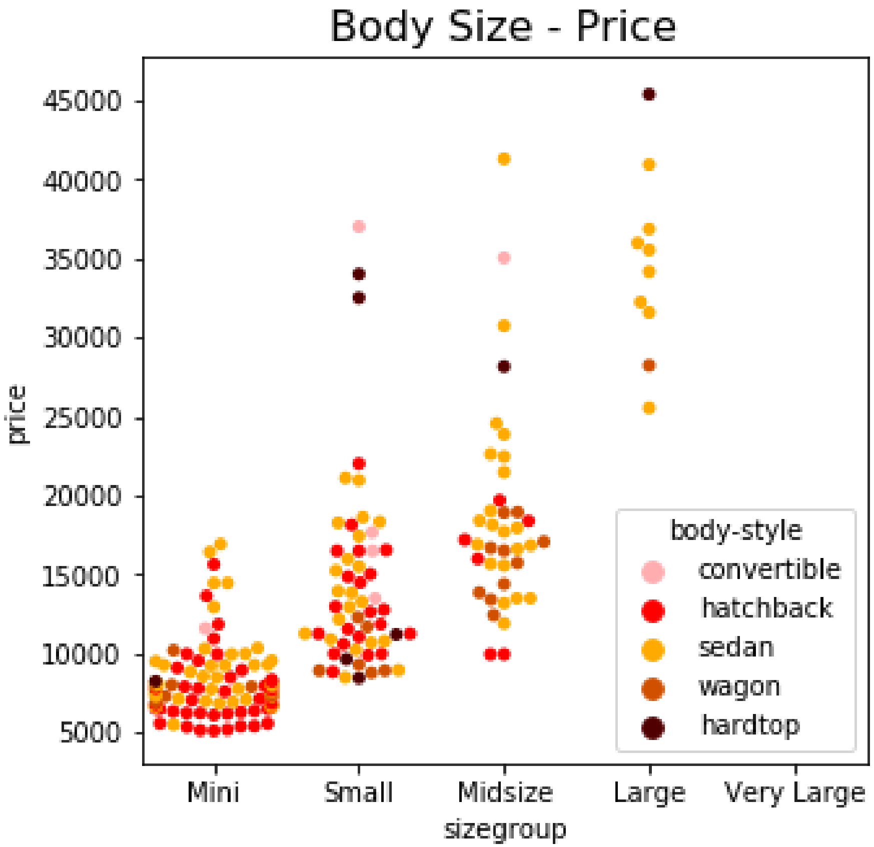


Guide to car size groups					
	Shadow (overall length x width in square feet)				
Curb weight	70-80	81-90	91-100	101-110	> 110
2,001-2,500 lbs	Mini	Small	Small	Small	Midsize
2,501-3,000 lbs	Small	Small	Midsize	Midsize	Midsize
3,001-3,500 lbs	Small	Midsize	Midsize	Large	Large
3,501-4,000 lbs	Small	Midsize	Large	Large	Very large
> 4,000 lbs	Midsize	Midsize	Large	Very large	Very large
Note: Passenger versions of vans often referred to as minivans are classified as cars.					



Source: http://www.iihs.org/iihs/topics/t/vehicle-size-and-weight/fatalityfacts/passenger-vehicles

Body Size and Price



Mini cars are priced between 5K and 10K.

Prices go up along with the size of the car.

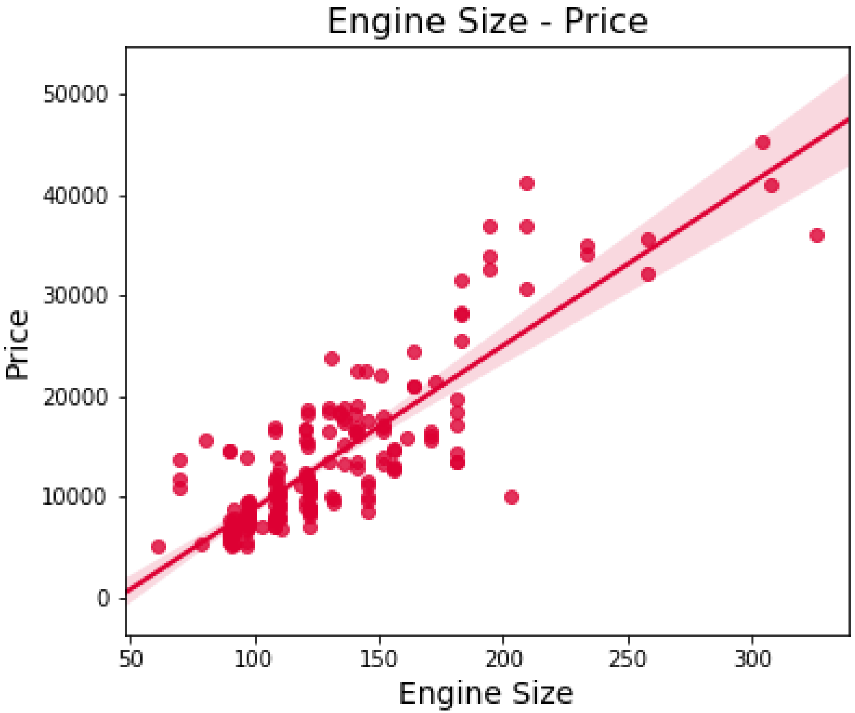
There is substantial price gap between Midsize and Large cars.

Engine Specifications and Price

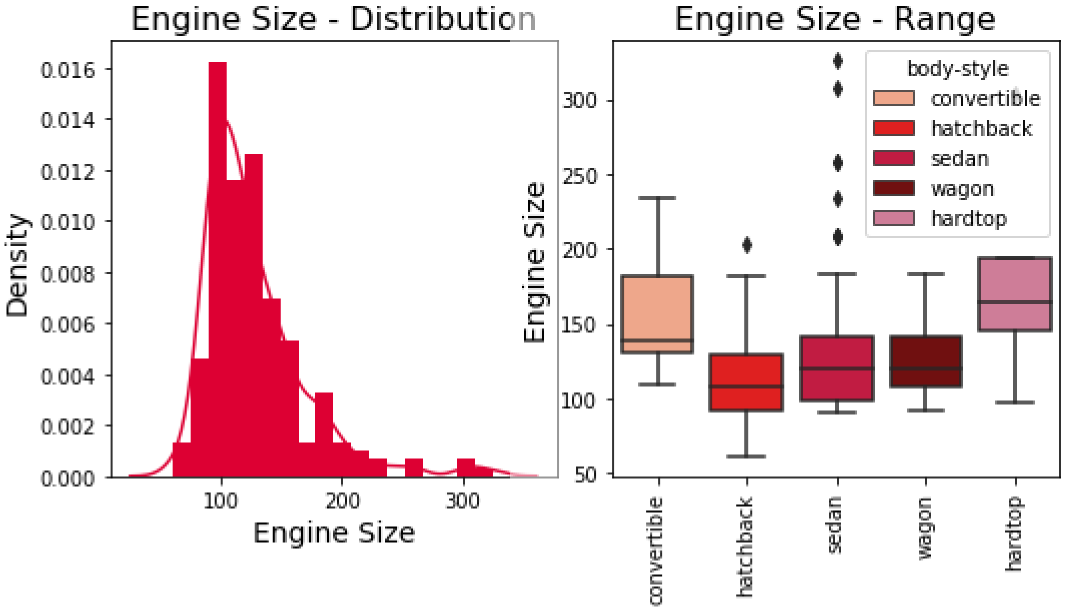
Analyze relationship with car pricing for the following engine parameters:

- **Engine Size**
- **Horsepower**
- **Fuel Efficiency**
- **Number of Cylinders**

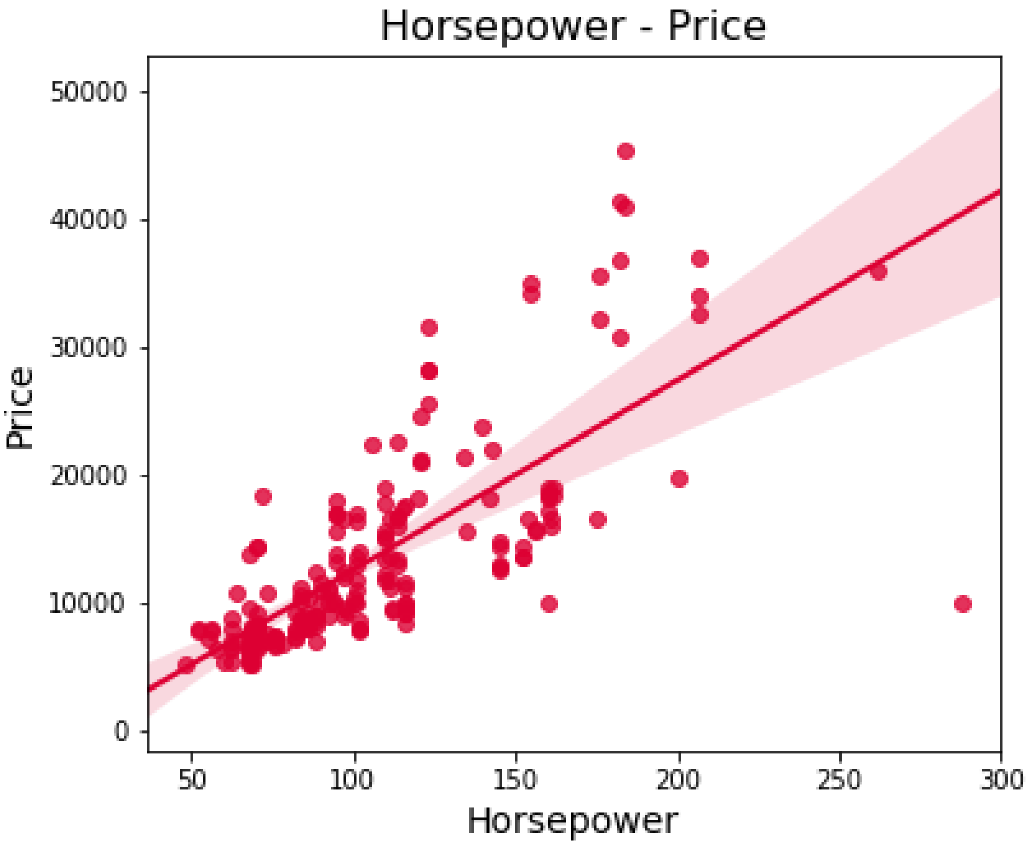
Engine Size



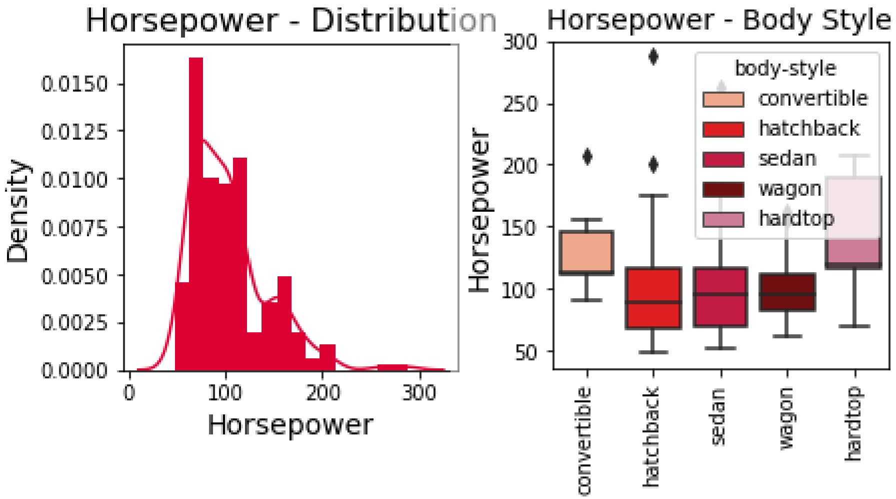
Car pricing maintains strong positive correlation with its engine size.



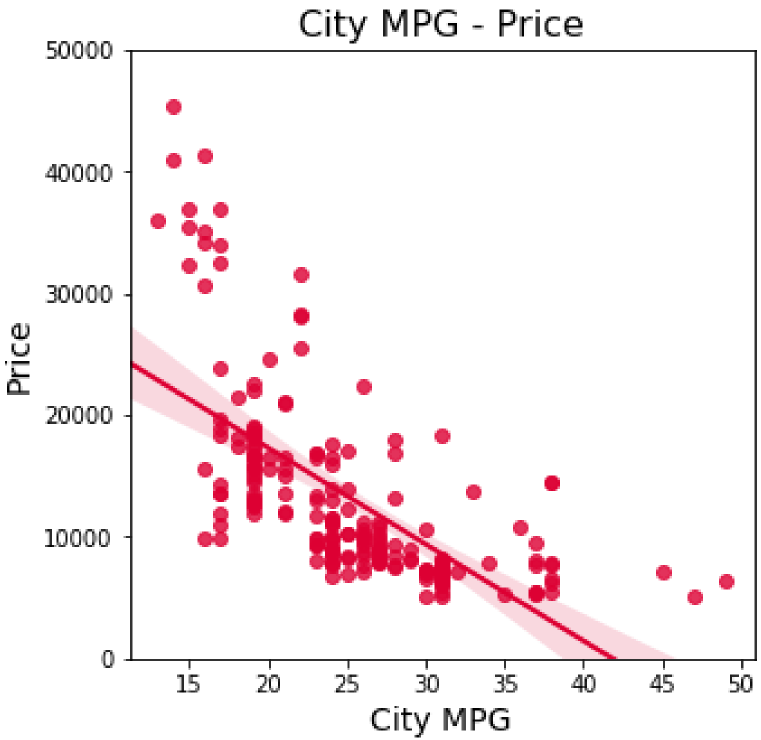
Horsepower



Car pricing maintains strong positive correlation with the engine Horsepower.



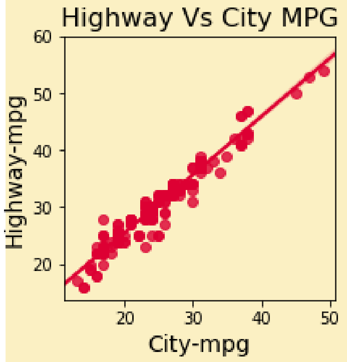
Fuel Efficiency (Miles per Gallon)



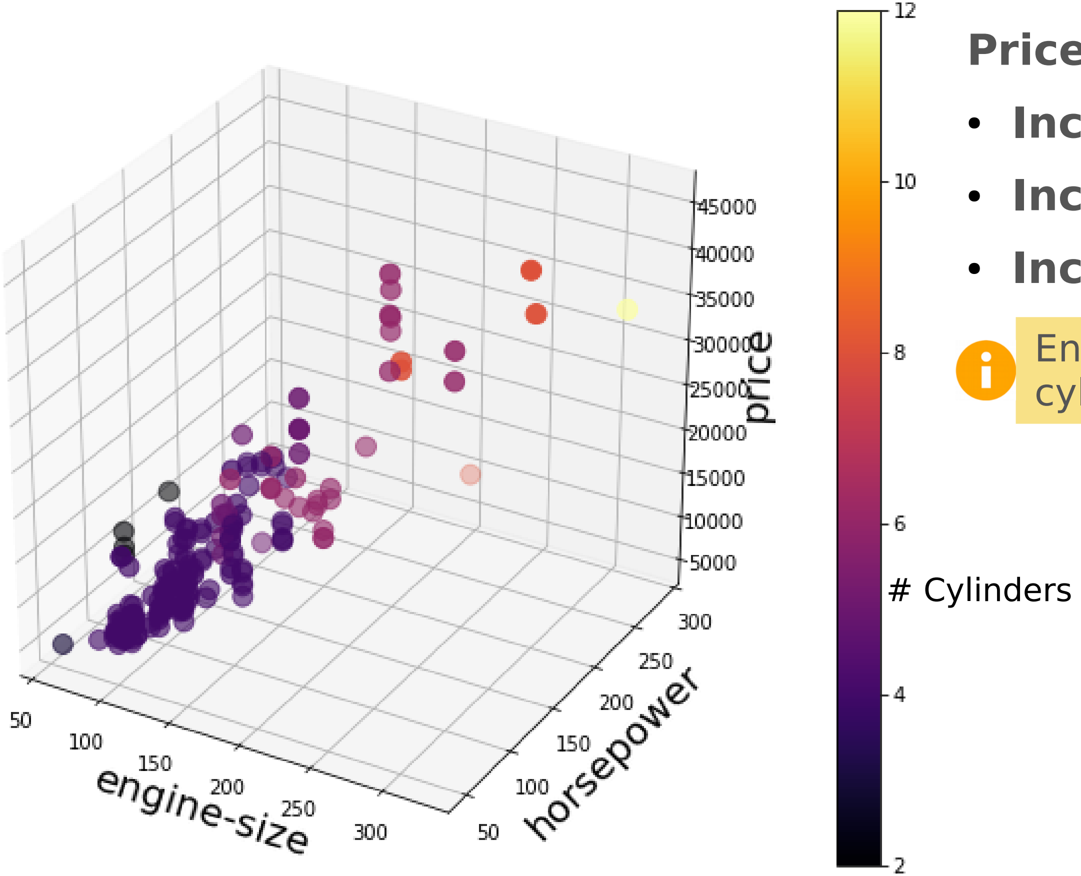
Fuel Efficiency shows negative correlation with pricing.

⚠ One assumption that can be made about the negative correlation is that,
high fuel efficiency cars are bought by budget conscious customers and hence products are made for the lower price bracket.
However, we do not have enough data to support this.

City-MPG is very strongly correlated to Highway MPG and hence only City-MPG is used for analysis purposes



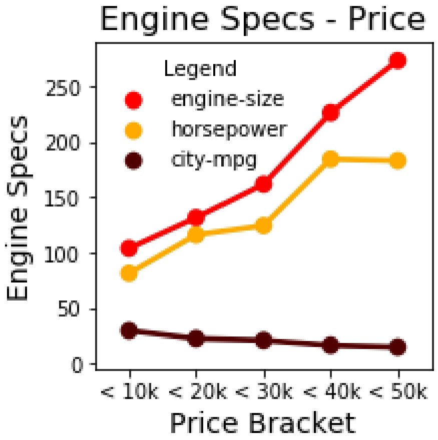
Engine Specs and Pricing



Prices go up along with:

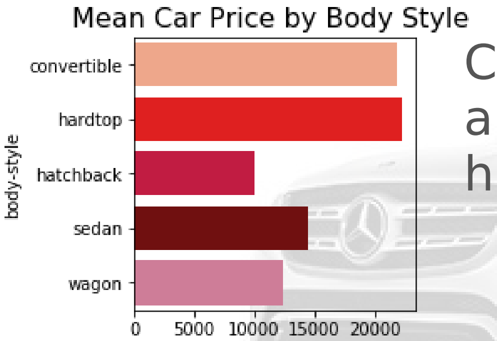
- **Increase in engine size**
- **Increase in horsepower**
- **Increase in number of cylinders**

Engine size, horsepower and number of cylinders are correlated among themselves too.

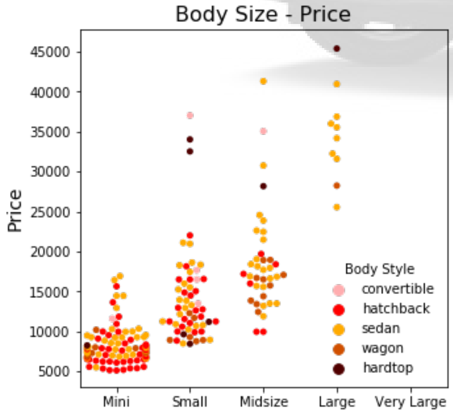


Analysis 2: Summary

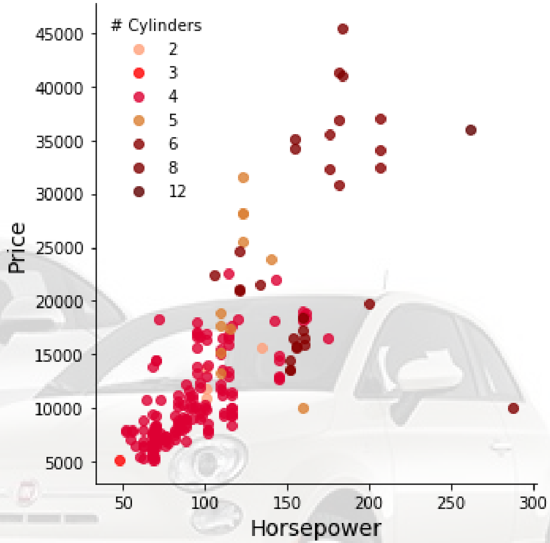
Do Body Style, Size and Engine Specs determine car prices?



Convertibles and hardtops are priced above sedans, hatchbacks and wagons



Bigger vehicles are priced above smaller ones



Cars having bigger engines,more cylinders and more power are priced higher

Thank You

Python Notebooks on GitHub:

1. https://github.com/justinpolackal/eda-automobiles/blob/master/RiskyVsSafe_Analysis.ipynb
2. https://github.com/justinpolackal/eda-automobiles/blob/master/PriceAndCarSpecs_Analysis.ipynb

References:

1. US Insurance Institute for Highway Safety | Highway Loss Data Institute - classification of cars based on vehicle size and weight: http://www.iihs.org/iihs/topics/t/vehicle-size-and-weight/fatalityfacts/passenger-vehicles
2. Car body styles http://www.nadaguides.com/Cars/Body-styles
3. Seaborn statistical data visualization: https://seaborn.pydata.org/
4. OHV, OHC, SOHC and DOHC (twin cam) engine design: https://www.samarins.com/glossary/dohc.html
5. UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/automobile