Hyperiondev

TASK

Exploratory Data Analysis on the Automobile Data Set

Visit our website

Introduction

Automobile dataset contains 205 rows × 26 columns, which contain the data of price, manufacturing company, fuel efficient, risks, and information related to the components of individual car. The exploratory data analysis (EDA) will focus on the on any variables related to price. Besides, the efficiency and power of engine for the automobiles are also important outcomes the manufacture company would like to know what variables could promote them. **1. price, 2. city-mpg, 3. highway-mpg, 4. horsepower, 5. normalized-losses and 6. symboling** are the variables we would like to explore what other variables could affect them.

There variables are as follows:
symboling [Numerical]: risk rating or insurance rating assigned to a vehicle. It indicates the level of risk associated with the car in terms of its potential for accidents, theft, or other risks. Typically, symboling values range from -3 to +3, where a higher positive value indicates a lower risk rating.
normalized-losses [Numerical]: the average amount paid by an insurance company for claims on a particular vehicle model.
wheel-base [Numerical]: the distance between the centers of the front and rear wheels of a vehicle. A crucial parameter that affects vehicle stability, handling, and interior space.
fuel-system [Categorical]: the mechanism and components involved in supplying fuel to the engine.
bore [Numerical]: the diameter of the cylinders in an internal combustion engine. A crucial parameter that affects the engine's displacement, performance, and power output
stroke [Numerical]: the distance traveled by the piston in a single movement inside the cylinder.Stroke length, along with bore size, determines the engine's displacement and affects its power, torque, and efficiency
compression-ratio [Numerical]: the ratio of the volume of the combustion chamber at the bottom dead center (BDC) to the volume at the top dead center (TDC) in an internal combustion engine. It indicates the degree of compression of the air-fuel mixture before combustion. A higher compression ratio typically results in improved engine efficiency and performance.
peak-rpm [Numerical]: the engine speed or revolutions per minute (RPM) at which the maximum power output is achieved. The highest rotational speed at which the engine can operate efficiently and generate the most power.
make [Categorical]: manufacturing company
fuel-type [Categorical]: fuel type of automobile
aspiration [Categorical]: draws air into the engine cylinders
num-of-doors [Categorical]: number of doors
body-style [Categorical]: shape of automobile
drive-wheels [Categorical]: essentially dictates the traction of the cars
engine-location [Categorical]: engine location
length [Numerical]: length of automobile
width [Numerical]: width of automobile
height [Numerical]: height of automobile
curb-weight [Numerical]: the weight of an automobile without occupants or baggage
engine-type [Categorical]: vehicle states how the engine is assembled or designed in terms of operations of valves and cylinders. :In this dataset we have seven engine types dohc (Dual OverHead Cam),dohcv (Dual OverHead Cam and Valve),l (L engine),ohc (OverHead

Cam),ohcf (OverHead Cam and Valve F engine),ohcv (OverHead Cam and Valve) ,rotor (Rotary engine)
num-of-cylinders [Categorical]: related to how the engine works, where generally more cylinders mean more potential power
engine-size [Numerical]: engine size of automobile
horsepower [Numerical]: horsepower measures the amount of power produced by the engine
city-mpg [Numerical]: fuel consumption in city by mpg (miles per gallon) unit
highway-mpg [Numerical]: fuel consumption in highway by mpg (miles per gallon) unit
price [Numerical]: price of automobile

DATA CLEANING

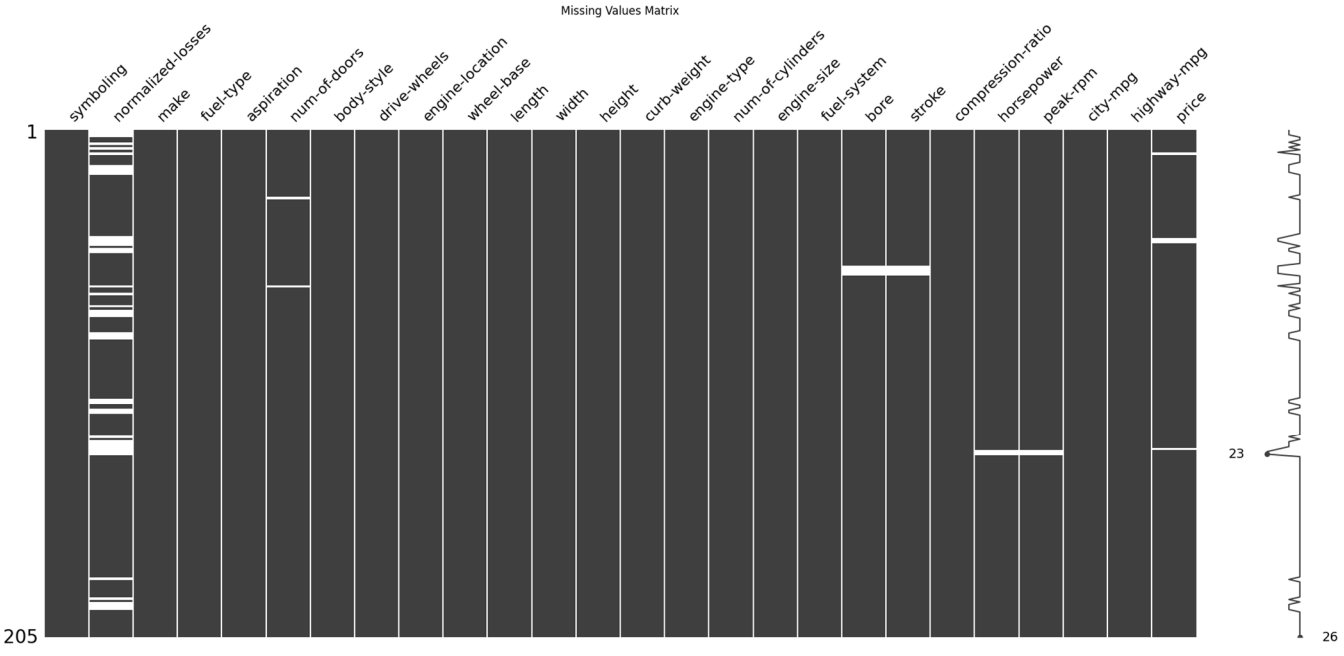
Irrelevant information: None was found in the dataset. All variables were analysed.

Duplicate rows: None was found in the dataset.

Null data/missing values: Unique values of each variable were identified and missing values found to be showed as '?' in the dataset. 41 in normalized-losses, 2 in num-of-doors, 4 in bore, 4 in stroke, 2 in horsepower, 2 in peak-rpm, 4 in price.

Variables were casted and modified to appropriate data types to be analysed.

The missing values were turned to NaN value and the missingno matrix was showed as follow:

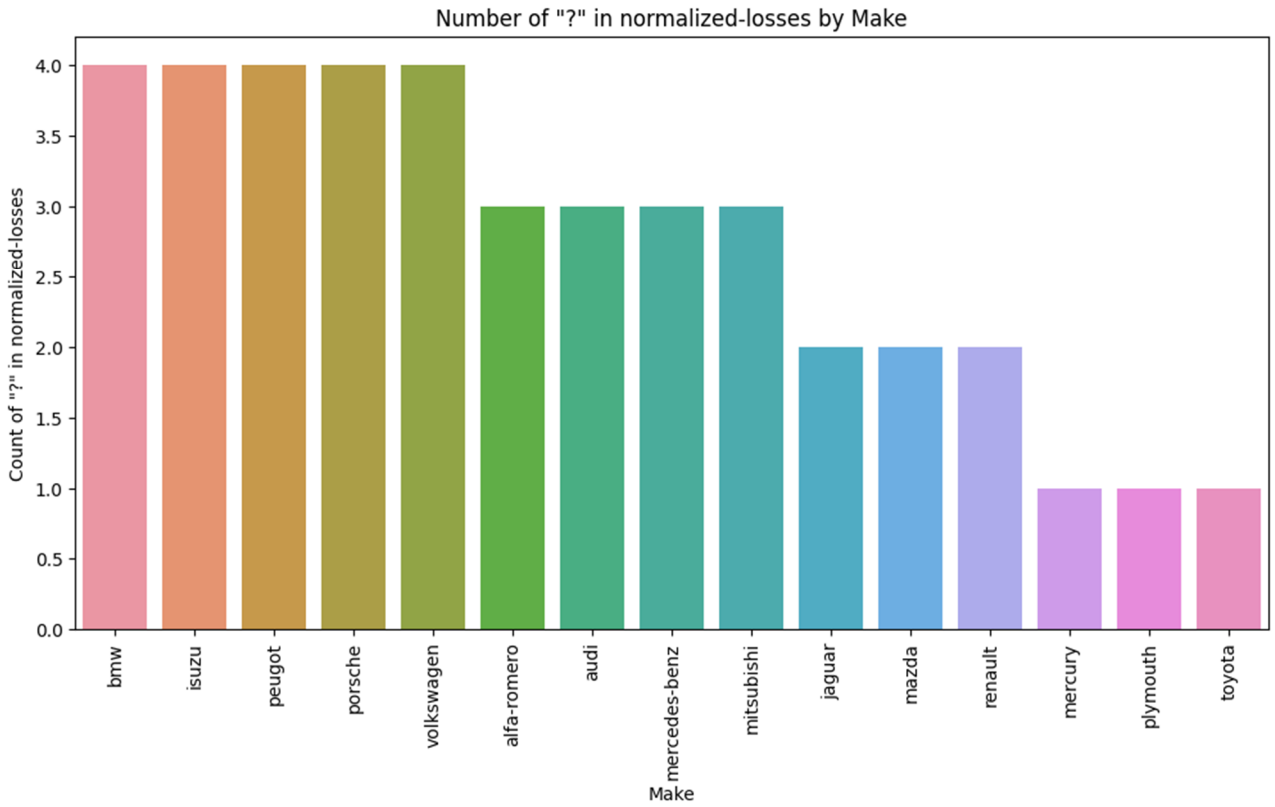


MISSING DATA

Missing values:
normalized-losses: 41

num-of-doors: 2
bore: 4
stroke: 4
horsepower: 2
peak-rpm: 2
price: 4

Investigating any pattern for missing values in 'normalized-losses' related to the information provided by individual manufacturing company:



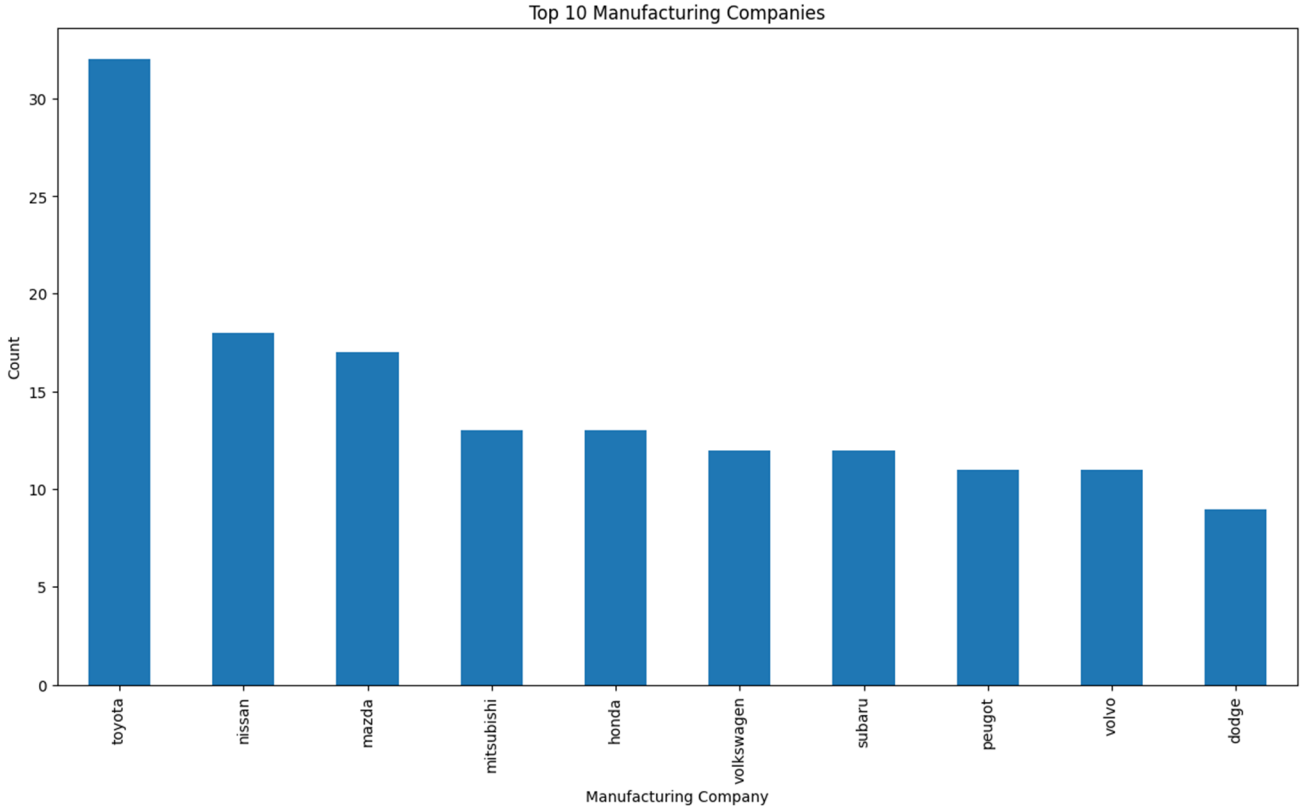
No obvious finding is observed.

The missing values of the variables 'normalized-losses' is about 20% (41/205), which were kept instead of dropping.
The missing values of numerical data were replaced with their median: normalized-losses, bore, stroke, horsepower, peak-rpm and price, while the missing values of categorical data were put in a new group unknown: num-of-doors.

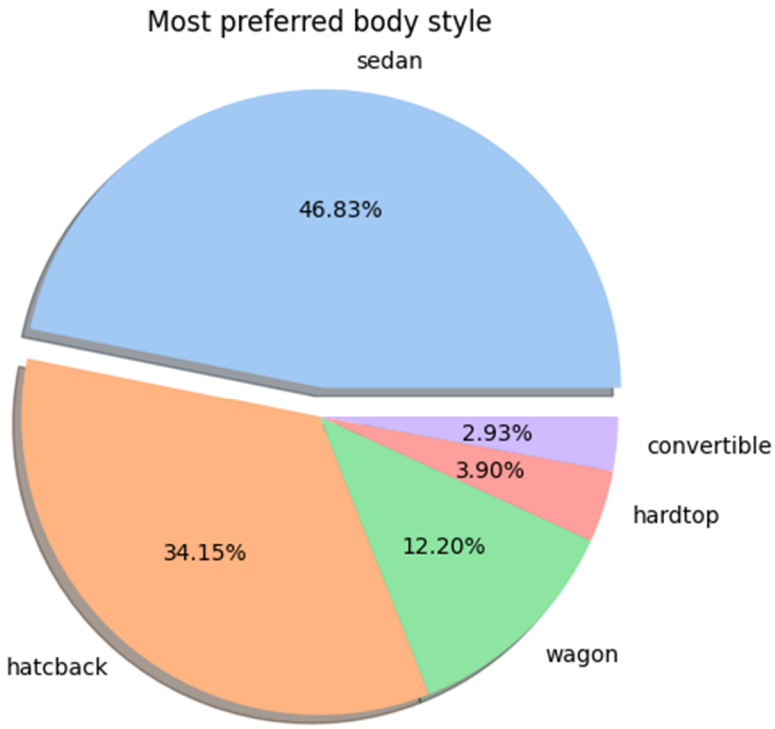
DATA STORIES AND VISUALISATIONS

Exploratory Data Analysis

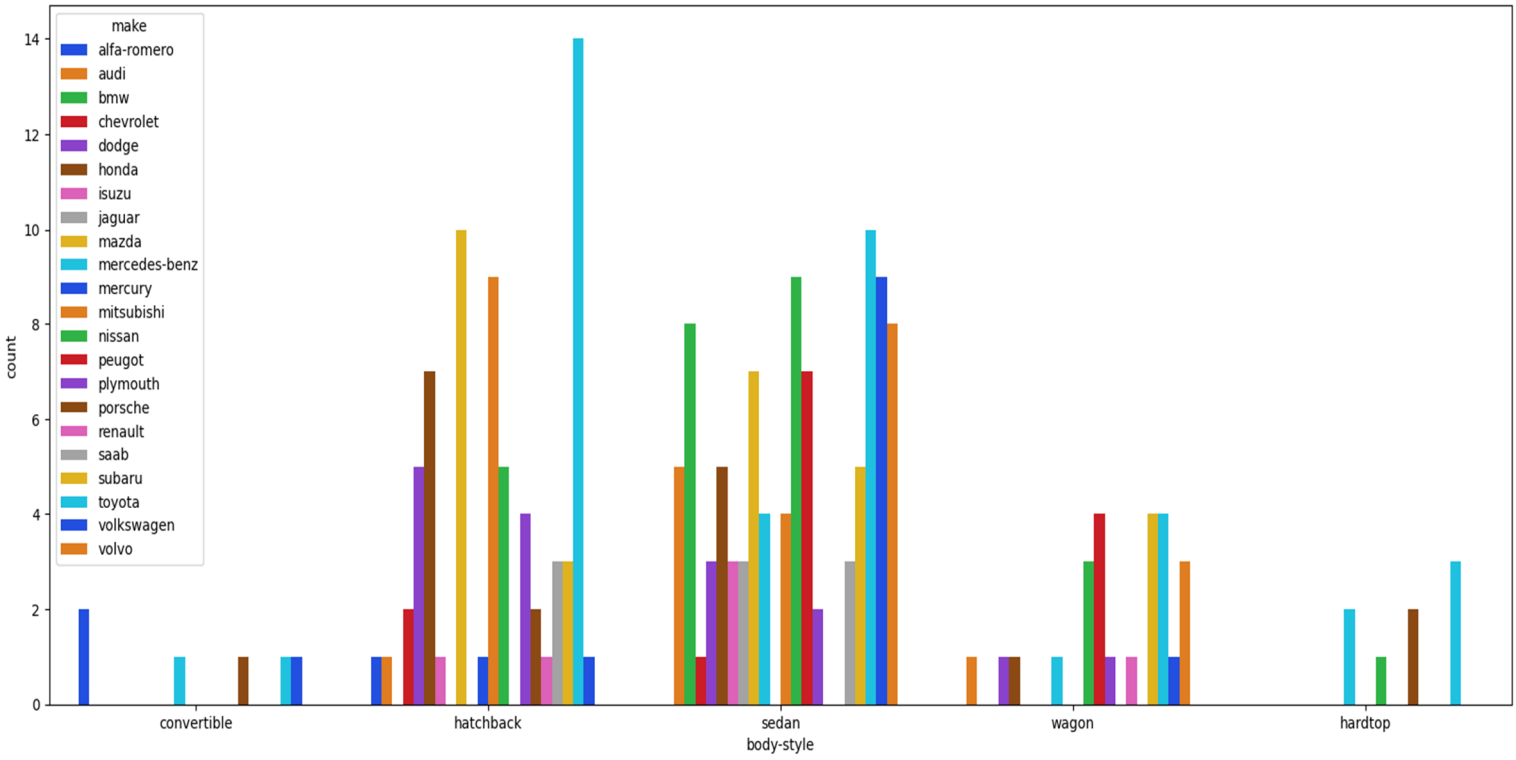
General statistics:



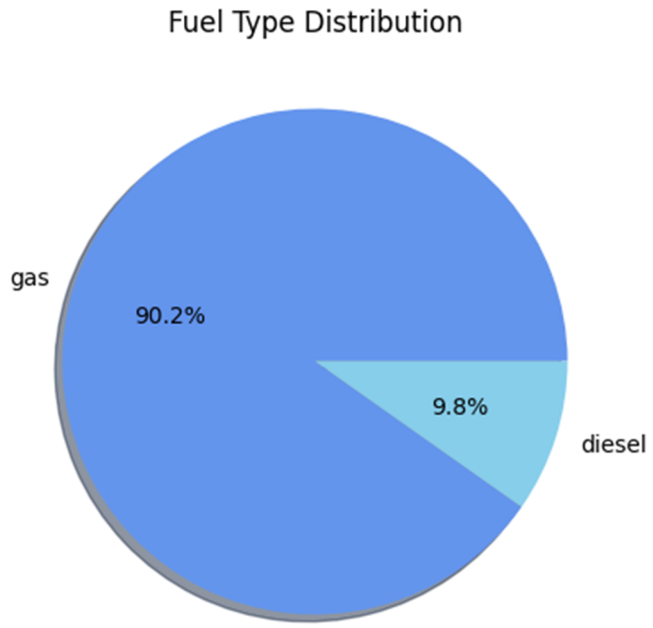
Toyota is the leading manufacturing company for production in the dataset. Nissan and Mazda are the second and third.



Sedan is the dominant body style of the cars in our dataset.

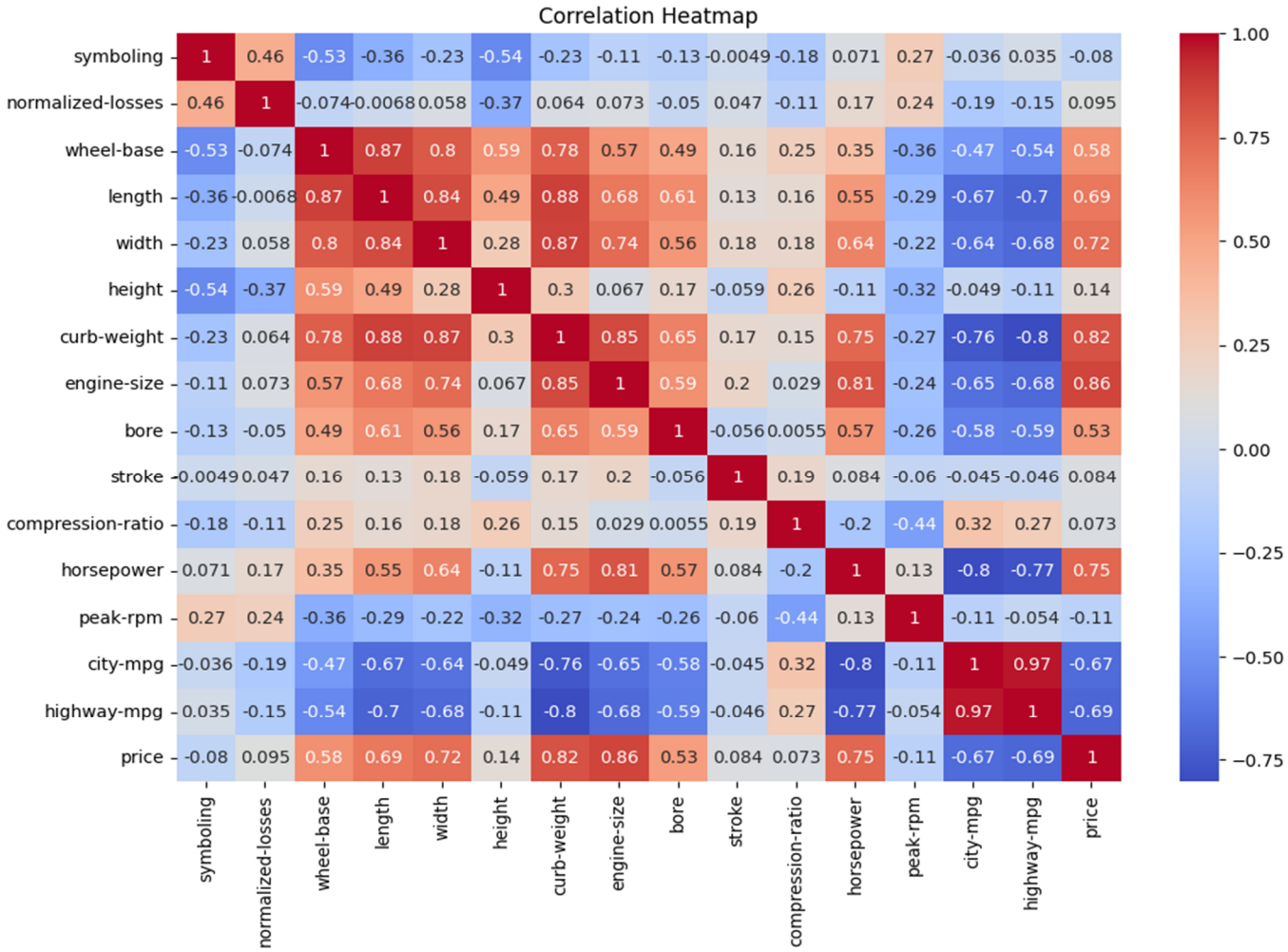


Toyota had the highest number of cars with sedan body style produced in the dataset, followed by Volkswagen and Nissan.



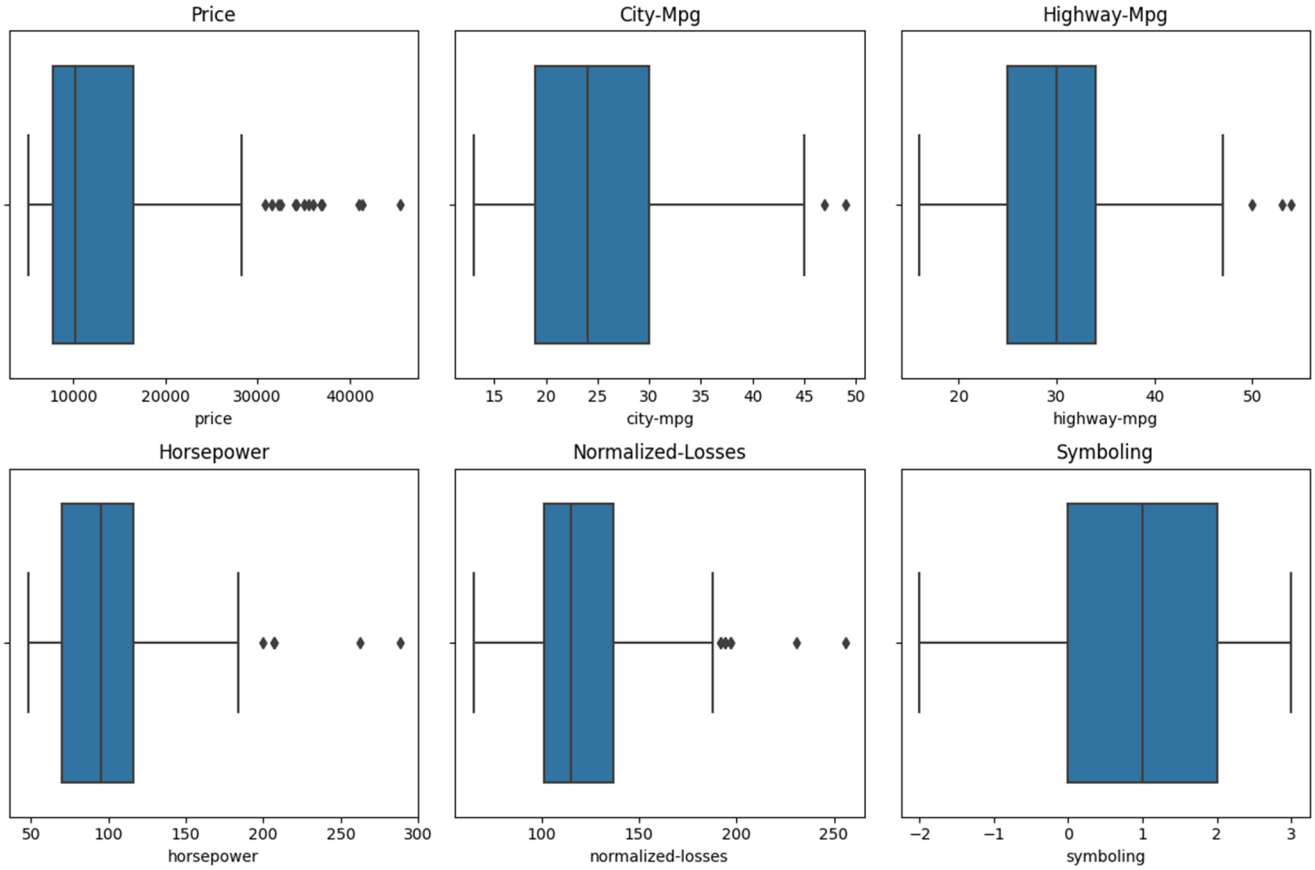
The cars using gas are dominant in the dataset.

Heatmap (Numerical variables):

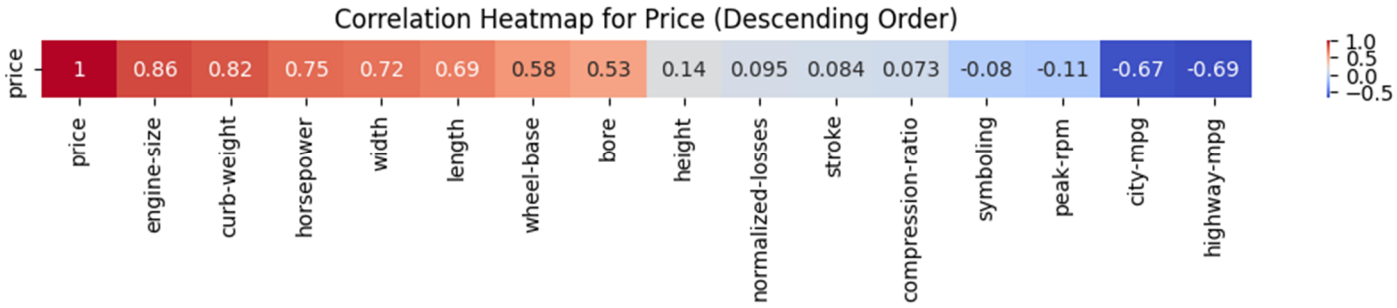


1. price, 2. city-mpg, 3. highway-mpg, 4. horsepower, 5. normalized-losses and 6. symboling are the interested parameters.

Overview of the range and outliers of the variables 1-6:



1. Price



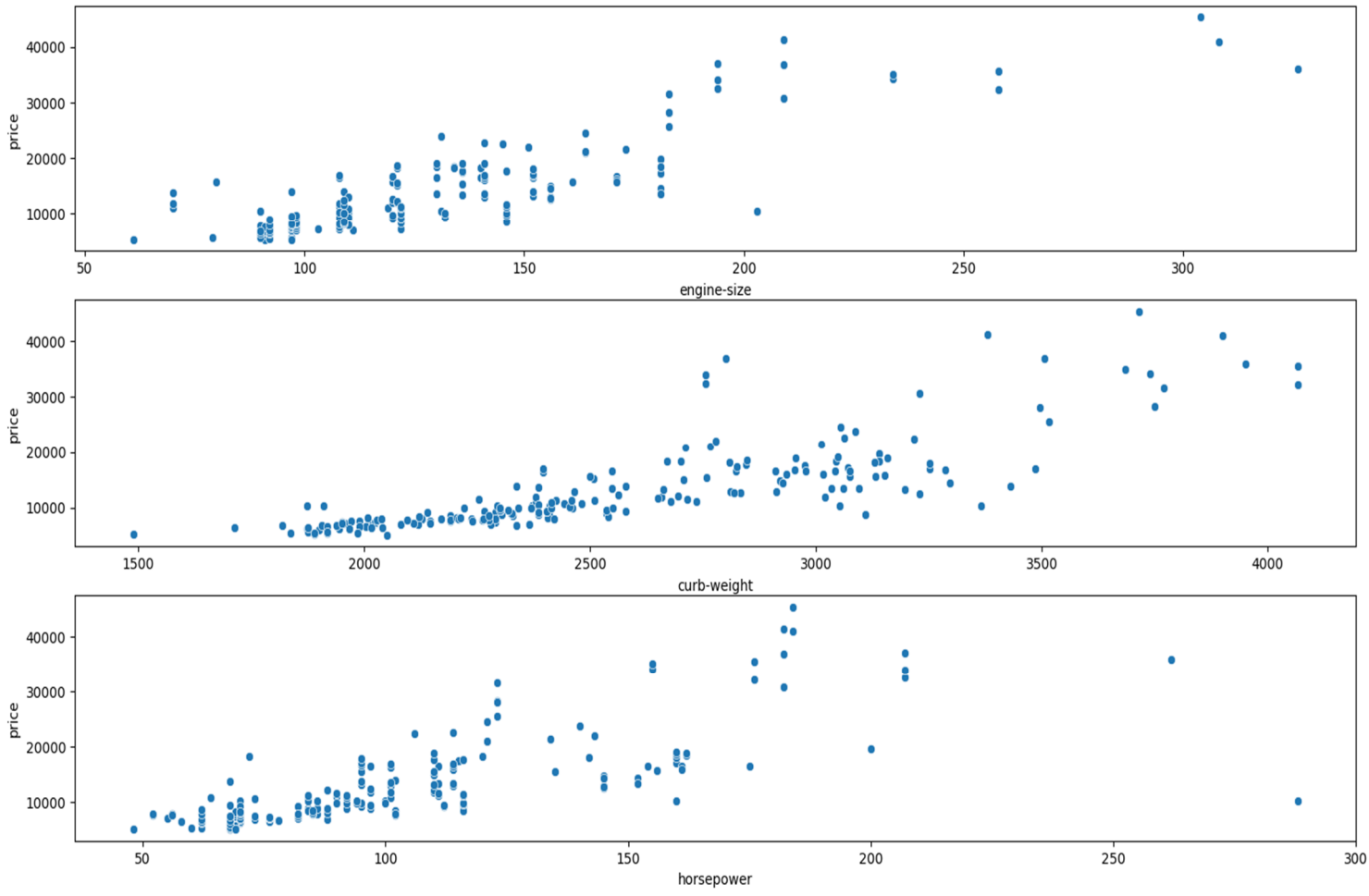
There are strong correlations with **engine size :0.86, curb weight:0.82, horsepower: 0.75, width: 0.72, length: 0.69, highway-mpg: -0.69, city-mpg: -0.67.**
Curb weight, width and length are not the parameters we are interested as they may be varying because of different size of engine (we can see there is strong correlation between them and engine size: 0.85, 0.74, 0.68)

ie: Larger engine needs larger curb weight, width and length of the automobile to install the engine.

Correlation values reference:

	Column	Correlation	p-value	R-squared
0	symboling	-0.080149	2.532966e-01	0.006424
1	normalized-losses	0.095489	1.732144e-01	0.009118
2	wheel-base	0.584847	3.338017e-20	0.342046
3	length	0.686567	6.428844e-30	0.471375
4	width	0.724558	1.178970e-34	0.524985
5	height	0.140439	4.459578e-02	0.019723
6	curb-weight	0.819817	4.794551e-51	0.672099
7	engine-size	0.860343	2.511194e-61	0.740190
8	bore	0.532861	1.964025e-16	0.283941
9	stroke	0.083627	2.332119e-01	0.006993
10	compression-ratio	0.072890	2.989755e-01	0.005313
11	horsepower	0.749919	2.770149e-38	0.562379
12	peak-rpm	-0.107283	1.257478e-01	0.011510
13	city-mpg	-0.668822	6.034975e-28	0.447322
14	highway-mpg	-0.693037	1.131042e-30	0.480301

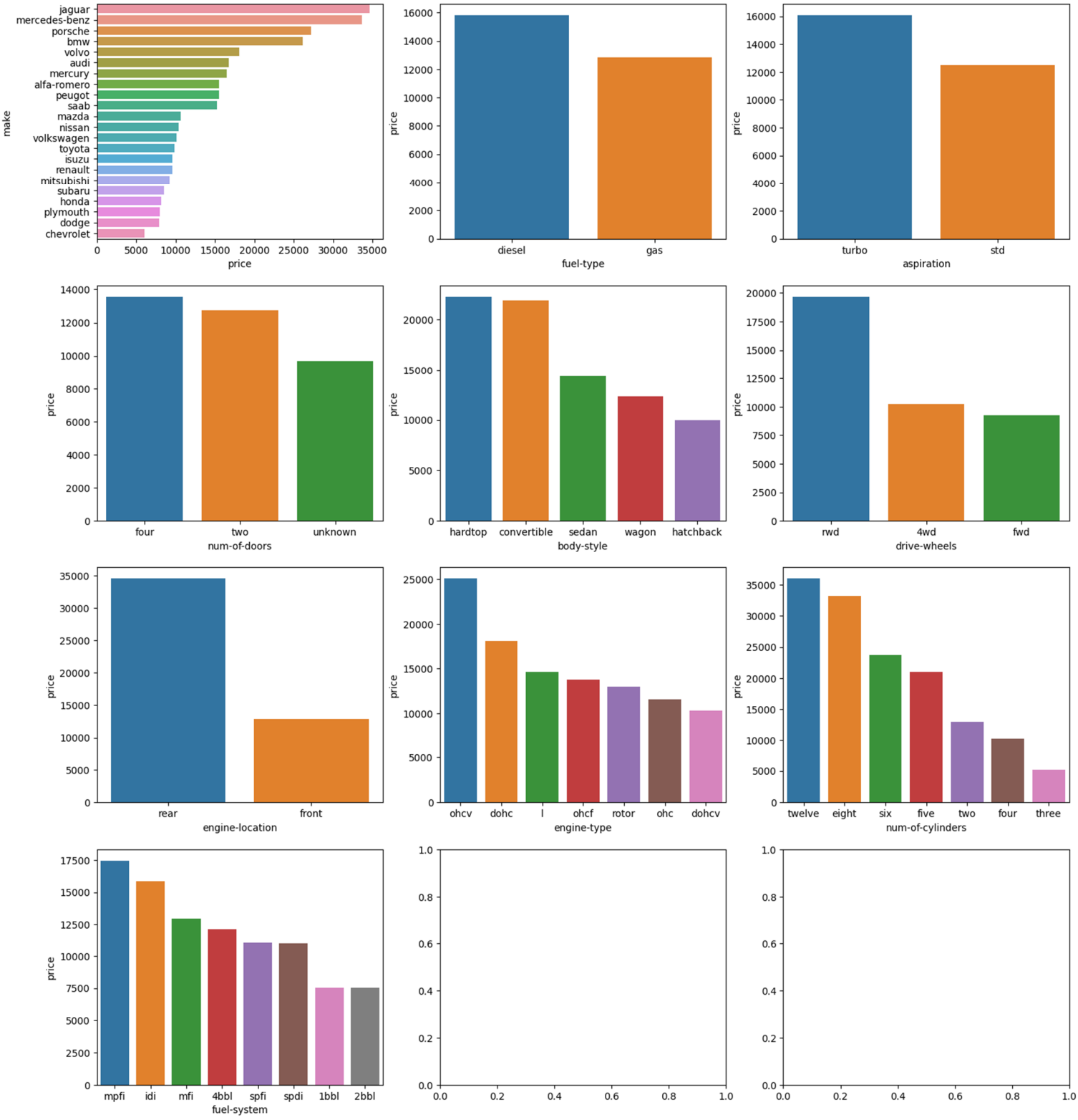
Price of the cars with different engine size, curb weight and horsepower:



The engine size, curb weight and horsepower have strong positive correlation with price. This makes sense, since engine size is directly related to car weight and horsepower as well apart from the width and length mentioned above.

	price
price	1.000000
drive-wheels_rwd	0.633080
make_mercedes-benz	0.525508
fuel-system_mpfi	0.503366
num-of-cylinders_six	0.487449
num-of-cylinders_eight	0.402909
engine-type_ohcv	0.395552
make_jaguar	0.332575
make_bmw	0.332494
engine-location_rear	0.331459
make_porsche	0.282217
num-of-cylinders_five	0.236079
body-style_hardtop	0.232240
num-of-cylinders_twelve	0.203540
body-style_convertible	0.193087
aspiration_turbo	0.175745
engine-type_dohc	0.157548
make_volvo	0.148838
body-style_sedan	0.145989
fuel-type_diesel	0.112439
fuel-system_idi	0.112439
make_audi	0.086796
make_peugot	0.070855
num-of-doors_four	0.055063
engine-type_l	0.046866
make_saab	0.045797
make_alfa-romero	0.036406
make_mercury	0.029865
engine-type_ohcf	0.021030
fuel-system_mfi	-0.001660
num-of-cylinders_two	-0.002339
engine-type_rotor	-0.002339
fuel-system_4bbl	-0.015587
fuel-system_spfi	-0.018727
engine-type_dohcv	-0.025434
body-style_wagon	-0.036906
num-of-doors_unknown	-0.043869
make_renault	-0.044898
num-of-doors_two	-0.046497
fuel-system_spdi	-0.058885
make_isuzu	-0.063618
num-of-cylinders_three	-0.071256
drive-wheels_4wd	-0.079154
make_mazda	-0.095548
make_volkswagen	-0.097484
make_nissan	-0.107944
make_chevrolet	-0.110756
fuel-type_gas	-0.112439
make_plymouth	-0.124081
make_mitsubishi	-0.129462
make_dodge	-0.143810
make_subaru	-0.146220
make_honda	-0.164391
fuel-system_1bbl	-0.169497
aspiration_std	-0.175745
make_toyota	-0.178629
body-style_hatchback	-0.291631
engine-type_ohc	-0.329810
engine-location_front	-0.331459
fuel-system_2bbl	-0.493613
drive-wheels_fwd	-0.587752
num-of-cylinders_four	-0.673478

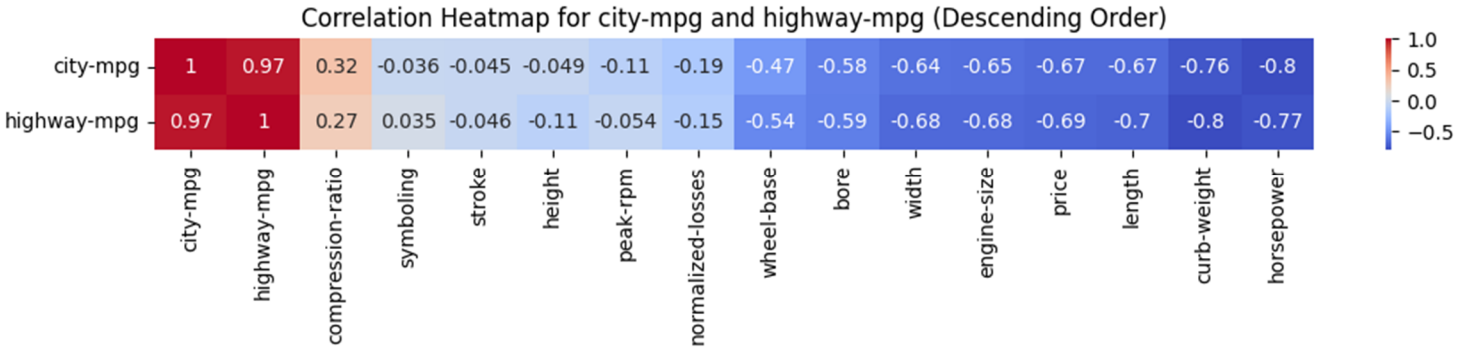
Highlights: **drive-wheels_rwd:0.633080, make_mercedes-benz:0.525508, fuel-system_mpfi:0.503366** made an expensive automobile, while **fuel-system_2bbl:-0.493613, drive-wheels_fwd:-0.587752, num-of-cylinders_four:-0.673478** made a cheaper automobile.



In the graphs we can see the average car price for each categorical variable.
Highlights of highest price in the variables are as follows:

Manufacturing company (top 3): jaguar, mercedez-ben and Porsche
Fuel type: diesel
Aspiration: turbo
Number of the doors: four
Body style: hardtop and convertible
Drive wheels: rwd
Engine location: rear
Engine type: dohcv
Number of cylinders: eight and twelve
Fuel system: mpfi and idi

2. City-mpg and 3. Highway-mpg

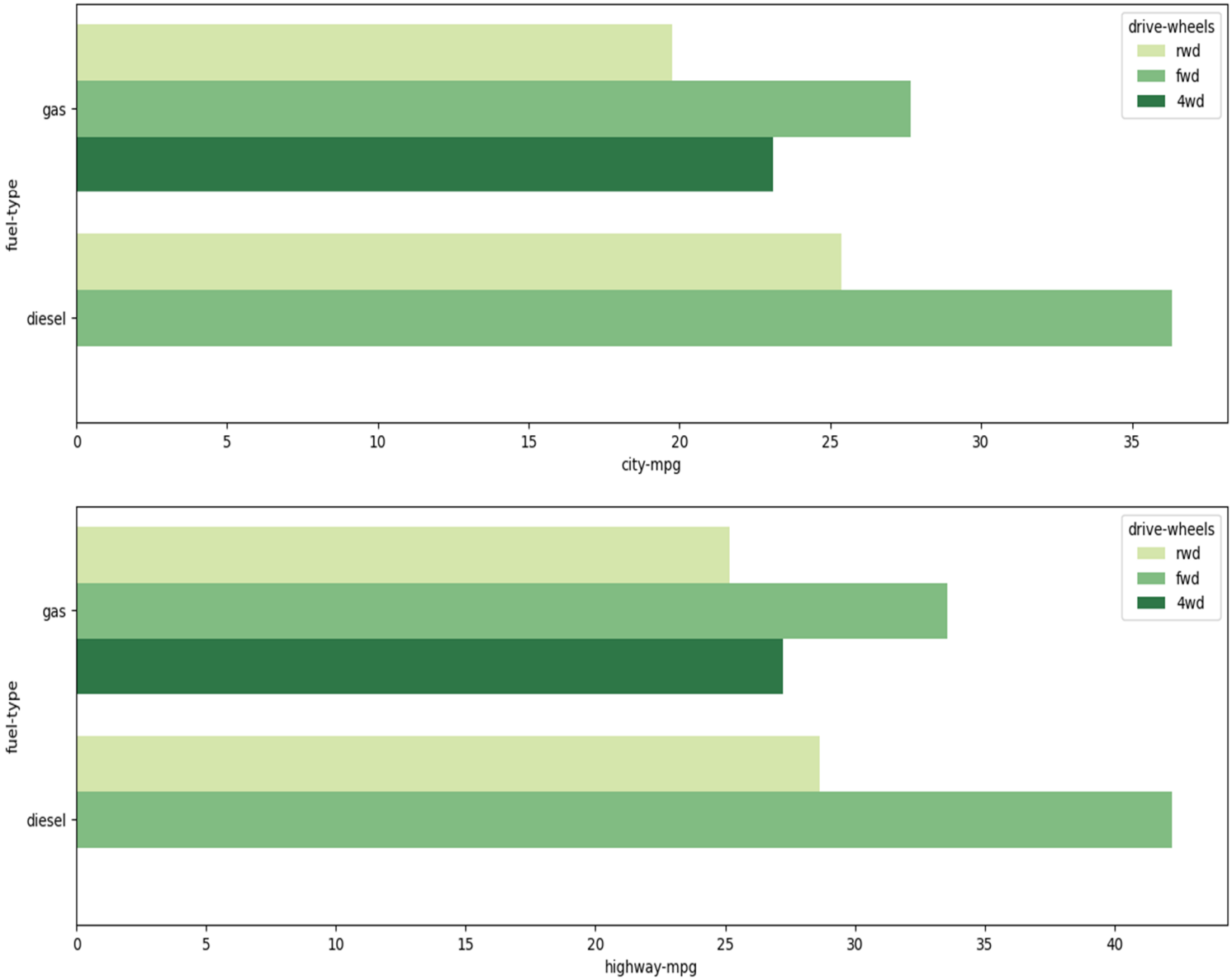


Highlights: The higher the horsepower, curb-weight, size (engine size, length, width, bore, wheel base), price they are, the less fuel efficient they will be (negative correlation)**.**
So **smaller engine (-0.68), smaller curb weight: -0.8, length: -0.7, width: -0.68 even cheaper automobile: -0.69 had higher fuel efficient**.
There is also **positive correlation with compression-ratio: 0.32 and 0.27.** A high compression ratio is desirable because it allows an engine to extract more mechanical energy from a given mass of air–fuel mixture due to its higher thermal efficiency.

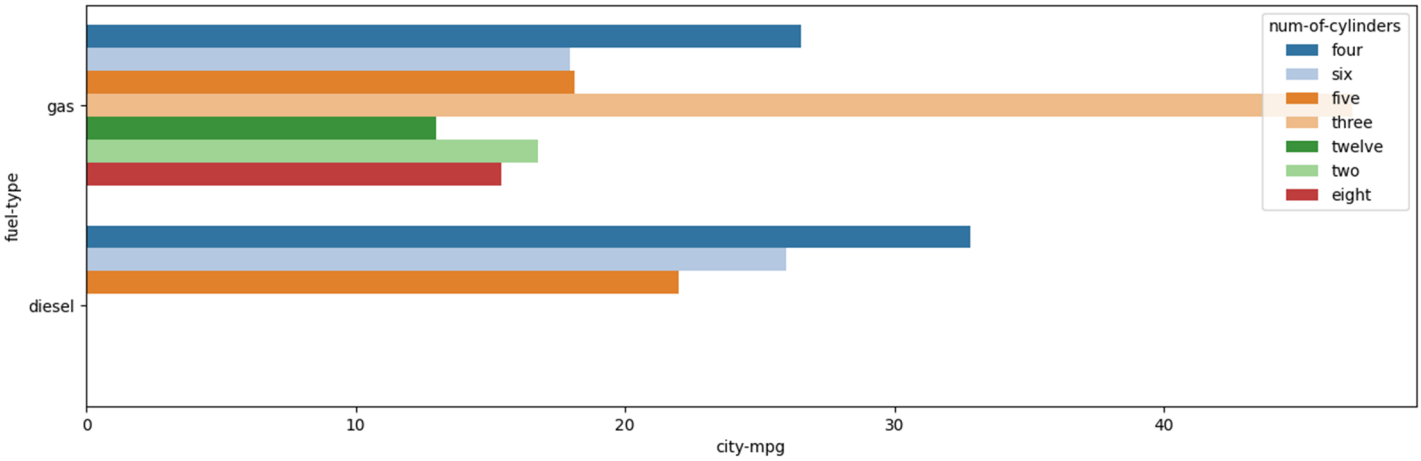
	City MPG	Highway MPG
city-mpg	1.000000	0.971337
highway-mpg	0.971337	1.000000
drive-wheels_fwd	0.563879	0.600828
num-of-cylinders_four	0.541267	0.547326
fuel-system_2bbl	0.520751	0.528009
engine-type_ohc	0.391236	0.426049
make_chevrolet	0.294678	0.276426
fuel-system_idi	0.255963	0.191392
fuel-type_diesel	0.255963	0.191392
num-of-cylinders_three	0.233665	0.226756
fuel-system_1bbl	0.227497	0.200844
make_honda	0.205941	0.178418
aspiration_std	0.202362	0.254416
engine-location_front	0.153487	0.102026
make_toyota	0.150287	0.134919
make_volkswagen	0.128525	0.151196
make_isuzu	0.124951	0.107785
body-style_hatchback	0.120795	0.148868
make_dodge	0.091297	0.104806
make_nissan	0.084644	0.099053
make_plymouth	0.084224	0.092831
num-of-doors_unknown	0.072708	0.075839
make_subaru	0.042557	-0.000044
make_mazda	0.022411	0.052089
body-style_sedan	0.014870	0.011218
num-of-doors_two	0.014271	0.037452
make_mitsubishi	-0.011819	0.015251
fuel-system_spfi	-0.013083	-0.017848
engine-type_l	-0.027491	-0.069615
num-of-doors_four	-0.028621	-0.052365
engine-type_ohcf	-0.032413	-0.047086
make_renault	-0.033757	0.003595
fuel-system_mfi	-0.066724	-0.068807
make_mercury	-0.066724	-0.068807
body-style_wagon	-0.067356	-0.110194
drive-wheels_4wd	-0.069229	-0.110081
engine-type_dohcv	-0.088181	-0.028040
make_alfa-romero	-0.091242	-0.072460
make_peugot	-0.100885	-0.142632
body-style_hardtop	-0.110993	-0.102707
fuel-system_spdi	-0.123954	-0.106615
body-style_convertible	-0.125571	-0.120094
make_saab	-0.130005	-0.086392
num-of-cylinders_twelve	-0.131093	-0.140150
make_volvo	-0.147323	-0.170992
engine-location_rear	-0.153487	-0.102026
fuel-system_4bbl	-0.153487	-0.137506
make_bmw	-0.180469	-0.157709
num-of-cylinders_two	-0.183076	-0.159173
engine-type_rotor	-0.183076	-0.159173
make_audi	-0.183306	-0.180874
make_porsche	-0.189449	-0.109356
engine-type_dohc	-0.199427	-0.214806
aspiration_turbo	-0.202362	-0.254416
make_jaguar	-0.203284	-0.220292
num-of-cylinders_five	-0.207029	-0.237167
make_mercedes-benz	-0.207487	-0.286047
num-of-cylinders_eight	-0.237904	-0.265867
fuel-type_gas	-0.255963	-0.191392
engine-type_ohcv	-0.336926	-0.360616
num-of-cylinders_six	-0.386551	-0.366696
drive-wheels_rwd	-0.545789	-0.566149

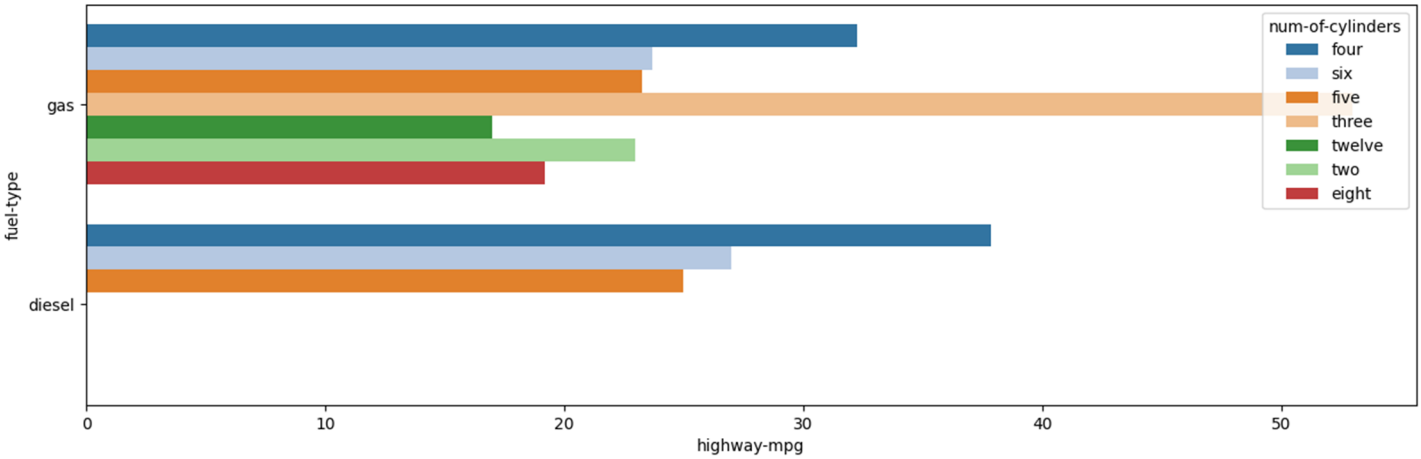
Highlights: **drive-wheels_fwd:0.563879 and 0.600828, num-of-cylinders_four:0.541267 and 0.547326, fuel-system_2bbl:0.520751 and 0.528009, engine-type_ohc:0.391236 and 0.426049** have positive correlation with the fuel efficient. While **engine-type_ohcv:-0.336926 and -0.360616, num-of-cylinders_six:-0.386551 and -0.366696, drive-wheels_rwd:-0.545789 and -0.566149, fuel-system_mpfi:-0.644489 and -0.610813** have negative impact on the fuel efficient.
Drive-wheels,num-of-cylinders,fuel-system,engine-type become the 4 important factors to explore the relationship with MPG.

fwd drive-wheels showed the best fuel efficient on both gas and diesel:

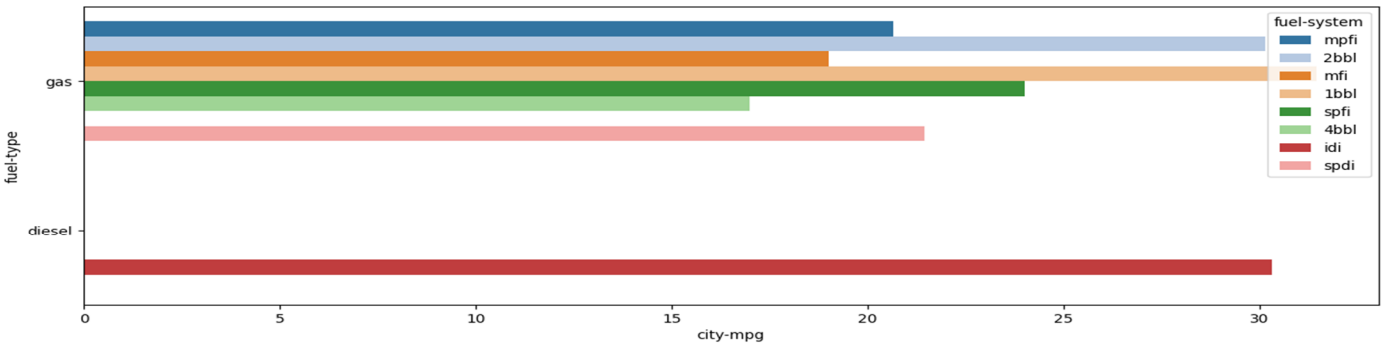


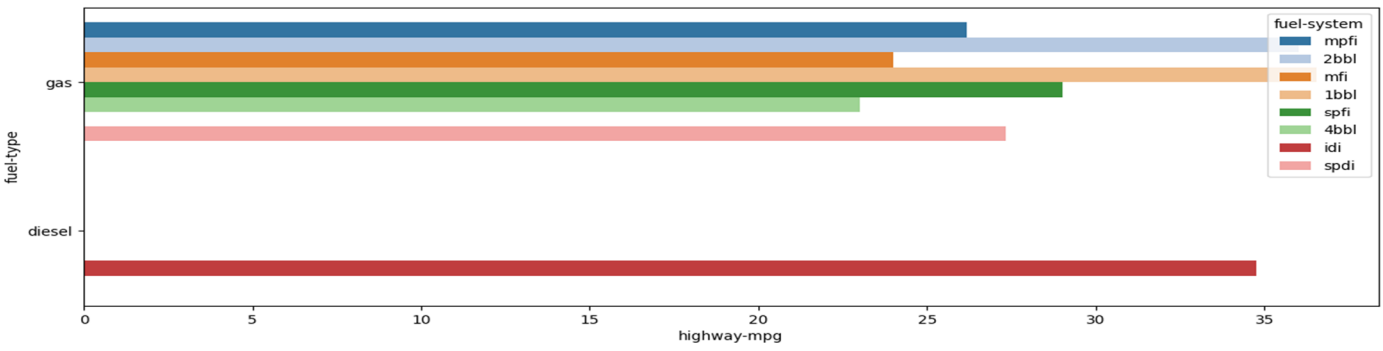
Fiver cylinders showed a high fuel efficient but it only exists for gas, while automobile with four cylinders have the highest efficient:



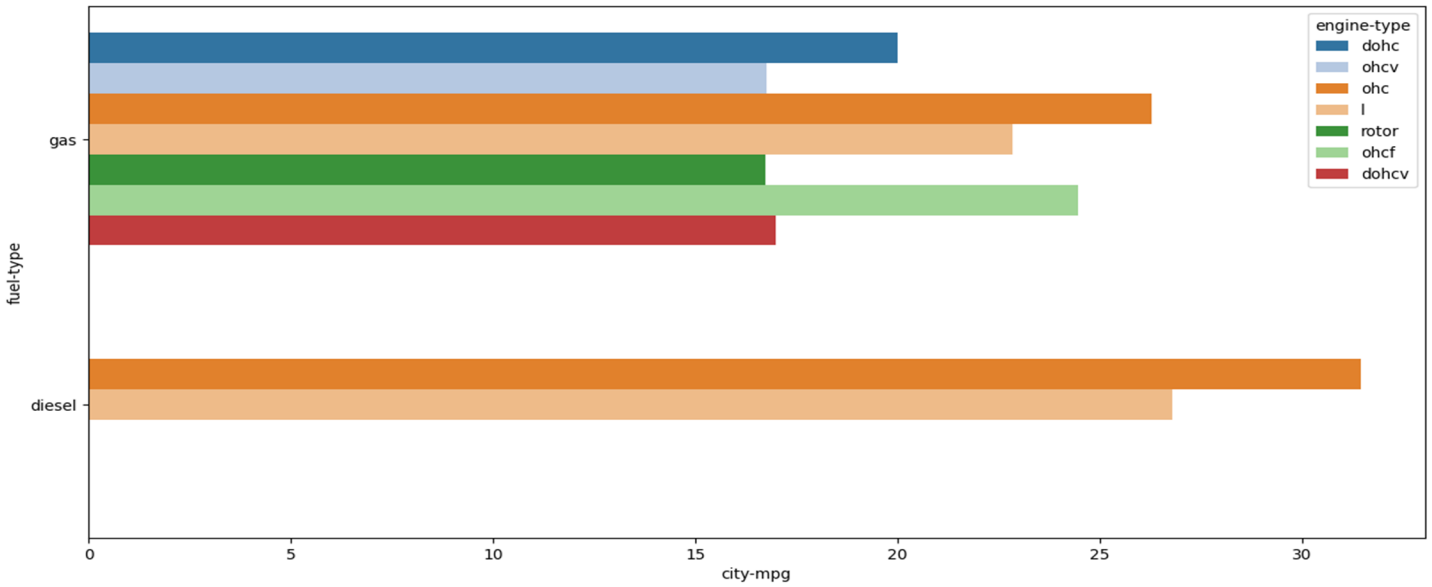


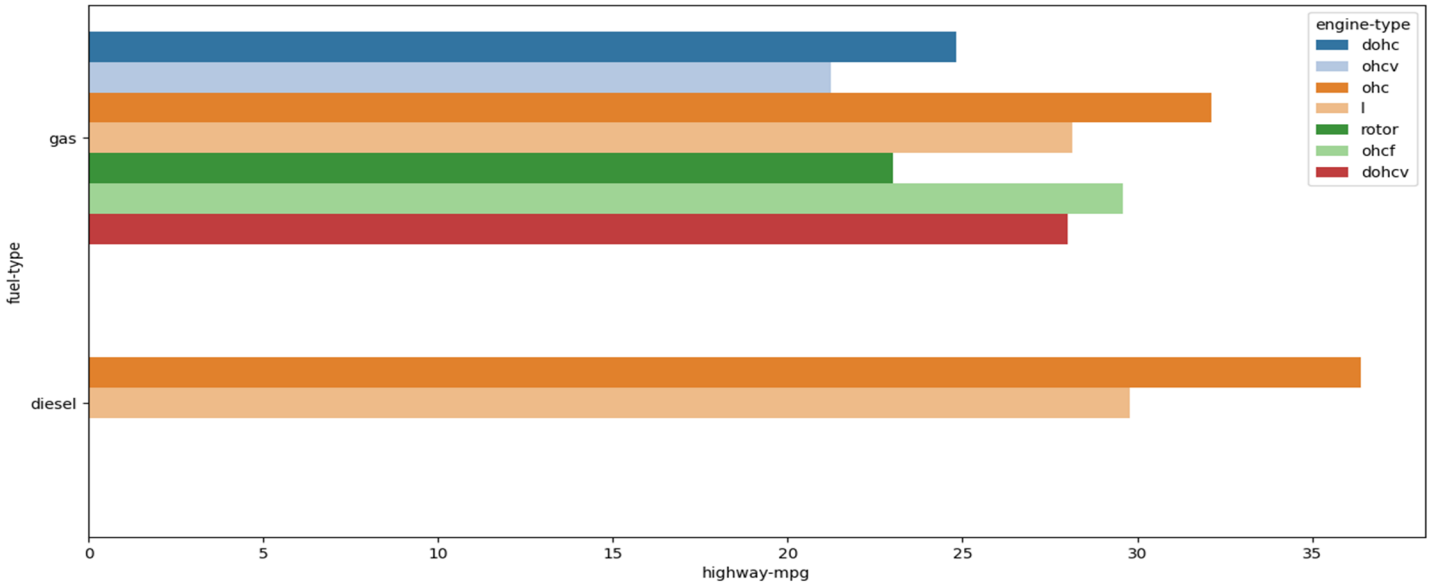
2bbl and 1bbl provided high fuel efficient for the automobile using gas:



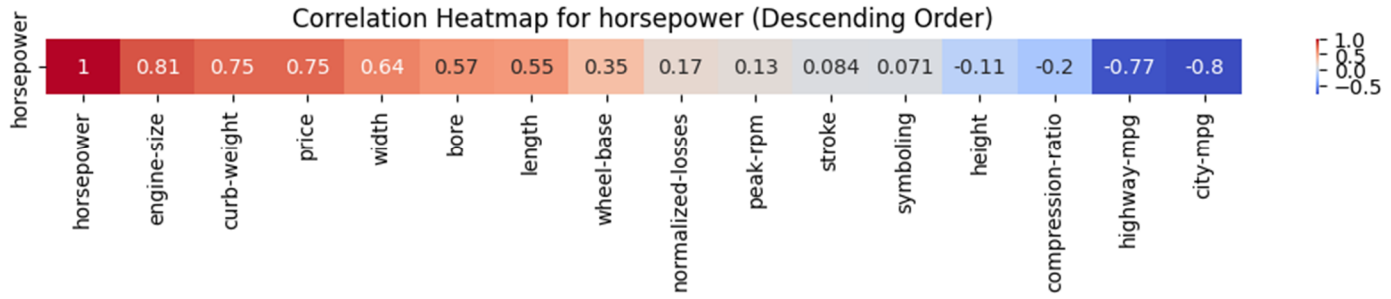


ohc engine provided highest fuel efficient for both gas and diesel automobile:





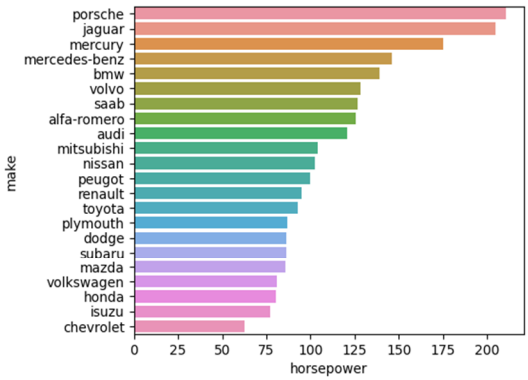
4. Horsepower

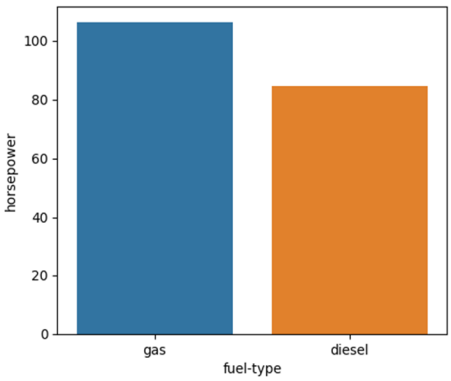


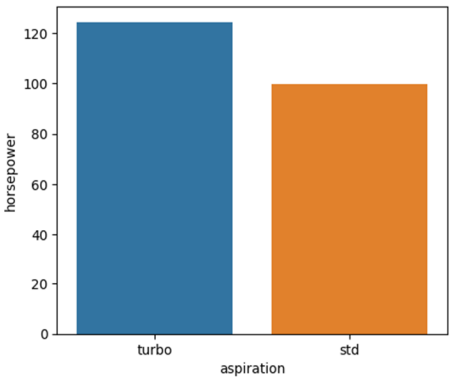
Bigger engine makes higher weight and so as width, length, bore, wheel base.
Highlights: And **bigger engine should have high horse power, which makes sense for the positive correlation:0.81**.
Although the engine size could promote the horsepower, the **fuel efficient declined (negative correlation: -0.77 and -0.8)** with it. It showed smaller engines tend to be more fuel efficient than larger engines as there is less fuel to burn to create power.

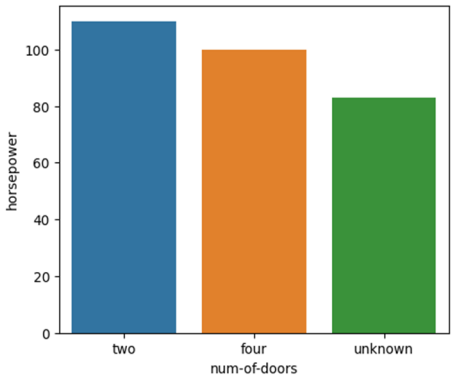
	City MPG	Highway MPG
city-mpg	1.000000	0.971337
highway-mpg	0.971337	1.000000
drive-wheels_fwd	0.563879	0.600828
num-of-cylinders_four	0.541267	0.547326
fuel-system_2bbl	0.520751	0.528009
engine-type_ohc	0.391236	0.426049
make_chevrolet	0.294678	0.276426
fuel-system_idi	0.255963	0.191392
fuel-type_diesel	0.255963	0.191392
num-of-cylinders_three	0.233665	0.226756
fuel-system_1bbl	0.227497	0.200844
make_honda	0.205941	0.178418
aspiration_std	0.202362	0.254416
engine-location_front	0.153487	0.102026
make_toyota	0.150287	0.134919
make_volkswagen	0.128525	0.151196
make_isuzu	0.124951	0.107785
body-style_hatchback	0.120795	0.148868
make_dodge	0.091297	0.104806
make_nissan	0.084644	0.099053
make_plymouth	0.084224	0.092831
num-of-doors_unknown	0.072708	0.075839
make_subaru	0.042557	-0.000044
make_mazda	0.022411	0.052089
body-style_sedan	0.014870	0.011218
num-of-doors_two	0.014271	0.037452
make_mitsubishi	-0.011819	0.015251
fuel-system_spfi	-0.013083	-0.017848
engine-type_l	-0.027491	-0.069615
num-of-doors_four	-0.028621	-0.052365
engine-type_ohcf	-0.032413	-0.047086
make_renault	-0.033757	0.003595
fuel-system_mfi	-0.066724	-0.068807
make_mercury	-0.066724	-0.068807
body-style_wagon	-0.067356	-0.110194
drive-wheels_4wd	-0.069229	-0.110081
engine-type_dohcv	-0.088181	-0.028040
make_alfa-romero	-0.091242	-0.072460
make_peugot	-0.100885	-0.142632
body-style_hardtop	-0.110993	-0.102707
fuel-system_spdi	-0.123954	-0.106615
body-style_convertible	-0.125571	-0.120094
make_saab	-0.130005	-0.086392
num-of-cylinders_twelve	-0.131093	-0.140150
make_volvo	-0.147323	-0.170992
engine-location_rear	-0.153487	-0.102026
fuel-system_4bbl	-0.153487	-0.137506
make_bmw	-0.180469	-0.157709
num-of-cylinders_two	-0.183076	-0.159173
engine-type_rotor	-0.183076	-0.159173
make_audi	-0.183306	-0.180874
make_porsche	-0.189449	-0.109356
engine-type_dohc	-0.199427	-0.214806
aspiration_turbo	-0.202362	-0.254416
make_jaguar	-0.203284	-0.220292
num-of-cylinders_five	-0.207029	-0.237167
make_mercedes-benz	-0.207487	-0.286047
num-of-cylinders_eight	-0.237904	-0.265867
fuel-type_gas	-0.255963	-0.191392
engine-type_ohcv	-0.336926	-0.360616
num-of-cylinders_six	-0.386551	-0.366696
drive-wheels_rwd	-0.545789	-0.566149
fuel-system_mpfi	-0.644489	-0.610813

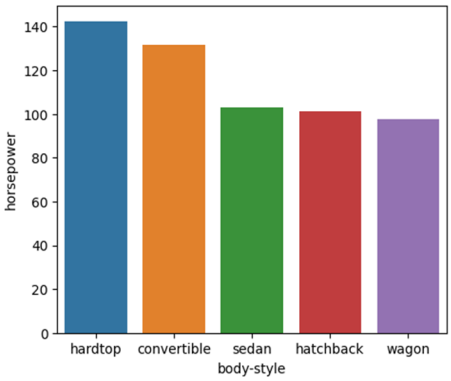
Highlights: **fuel-system_mpfi:0.629946, drive-wheels_rwd:0.574825, num-of-cylinders_six:0.533289, engine-type_ohcv:0.431375, make_porsche:0.425963 make the high horsepower.**

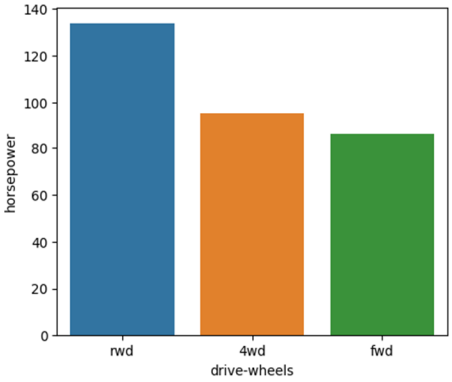


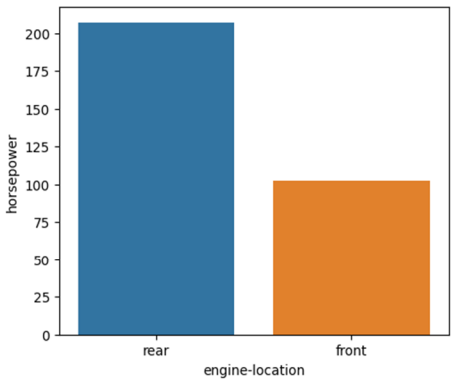


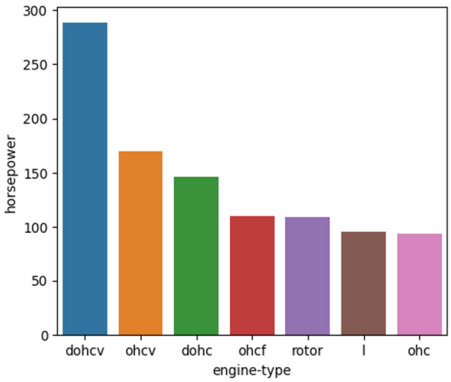


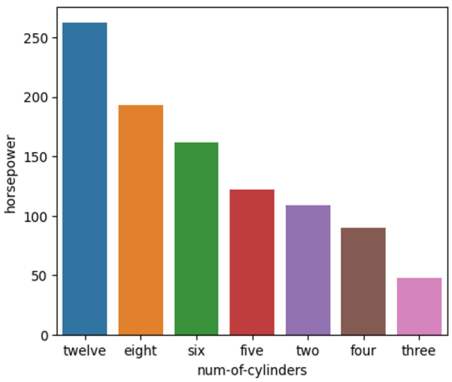


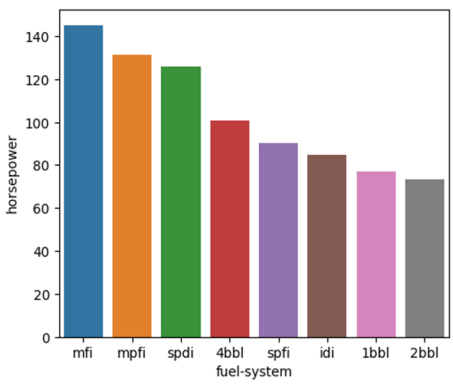


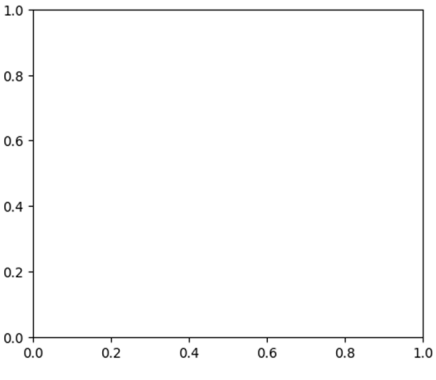


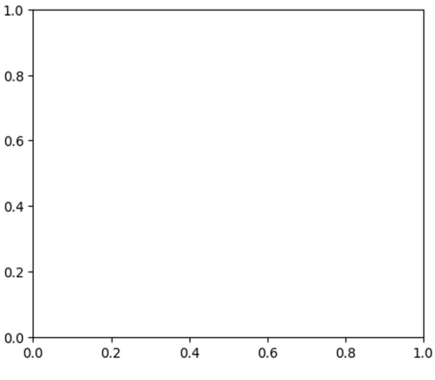




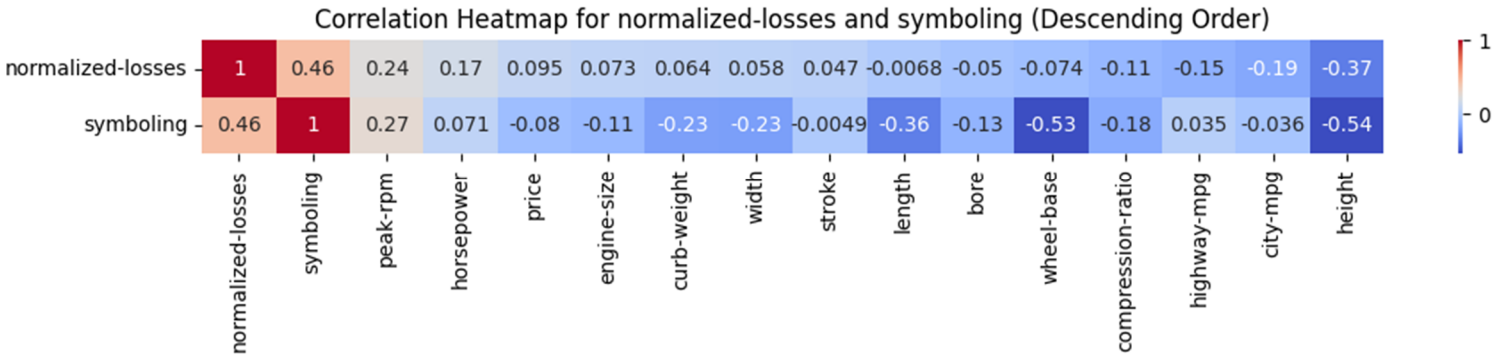








5. normalized-losses and 6. Symbolling



Highlights: Higher risk (symboling) gave higher relative average loss payment per insured vehicle year (normalized-losses): with **positive correlation :0.46**. They are both interestingly **positively related to peak-rpm (0.24 and 0.27)**, which is stronger than the relationship with horsepower (only 0.17 and 0.071). RPM stands for revolutions per minute and is a measure of how fast the engine is spinning. The faster an engine spins, the more power it makes. At a higher RPM, the engine is burning more fuel, so it makes more power and consumes more fuel.

Normalized-losses and symboling are **both negatively related to height (-0.37 and -0.54)**, it may suggest taller automobile drivers are more conscious to drive or the taller automobiles are more stable, which made fewer incidents or claims of insurance.

The **larger the wheel base is also observed to reduce the risks (symboling): -0.53.**

	Normalized Losses	Symboling
normalized-losses	1.000000	0.457484
symboling	0.457484	1.000000
num-of-doors_two	0.348850	0.663595
drive-wheels_rwd	0.273564	-0.076381
body-style_hatchback	0.205236	0.435648
make_bmw	0.202613	-0.074482
num-of-cylinders_six	0.197788	-0.000238
fuel-system_mpfi	0.179458	0.012532
make_peugot	0.177668	-0.159891
engine-type_l	0.170806	-0.133979
engine-type_dohc	0.151813	0.116925
make_mitsubishi	0.150905	0.211978
make_nissan	0.142443	0.041422
num-of-cylinders_two	0.130721	0.245950
engine-type_rotor	0.130721	0.245950
engine-type_ohcv	0.130717	-0.013597
make_audi	0.122589	0.068348
fuel-system_4bbl	0.112927	0.212471
fuel-type_gas	0.104668	0.194311
make_dodge	0.086751	0.028609
fuel-system_mfi	0.053844	0.122067
fuel-system_spdi	0.052231	0.181939
make_porsche	0.042858	0.224755
make_plymouth	0.037928	0.025103
make_saab	0.035026	0.232847
body-style_hardtop	0.034298	0.168845
num-of-doors_unknown	0.034100	-0.026699
make_mazda	0.021632	0.068625
engine-location_front	0.021510	-0.212471
make_jaguar	0.016901	-0.081830
body-style_convertible	0.011311	0.279440
aspiration_std	0.011273	0.059866
num-of-cylinders_three	0.000883	0.065707
num-of-cylinders_eight	-0.000997	-0.004346
num-of-cylinders_five	-0.009961	-0.090188
make_volkswagen	-0.011265	0.167106
aspiration_turbo	-0.011273	-0.059866
fuel-system_spfi	-0.012358	0.065707
num-of-cylinders_twelve	-0.012358	-0.047012
engine-type_dohcv	-0.012358	0.009347
make_mercury	-0.012358	0.009347
make_renault	-0.017519	0.013252
make_alfa-romero	-0.021510	0.147071
engine-location_rear	-0.021510	0.212471
make_isuzu	-0.024899	-0.009555
body-style_sedan	-0.025142	-0.378341
make_chevrolet	-0.079126	0.016270
make_mercedes-benz	-0.083999	-0.135313
fuel-type_diesel	-0.104668	-0.194311
fuel-system_idi	-0.104668	-0.194311
fuel-system_1bbl	-0.122539	-0.037911
fuel-system_2bbl	-0.123927	-0.034069
drive-wheels_4wd	-0.133128	-0.067222
make_toyota	-0.137758	-0.094046
make_honda	-0.144344	-0.045822
engine-type_ohc	-0.156069	-0.082855
num-of-cylinders_four	-0.188104	-0.034161
engine-type_ohcf	-0.210771	0.037513
drive-wheels_fwd	-0.212838	0.102839
make_volvo	-0.218742	-0.403849
make_subaru	-0.222808	-0.067071
body-style_wagon	-0.285199	-0.298243
num-of-doors_four	-0.354755	-0.656712

Highlights of Symboling

Positive correlation: **num-of-doors_two:0.663595 and body-style_hatchback: 0.435648.**

Negative correlation: **num-of-doors_four: -0.656712 and make_volvo: -0.403849**

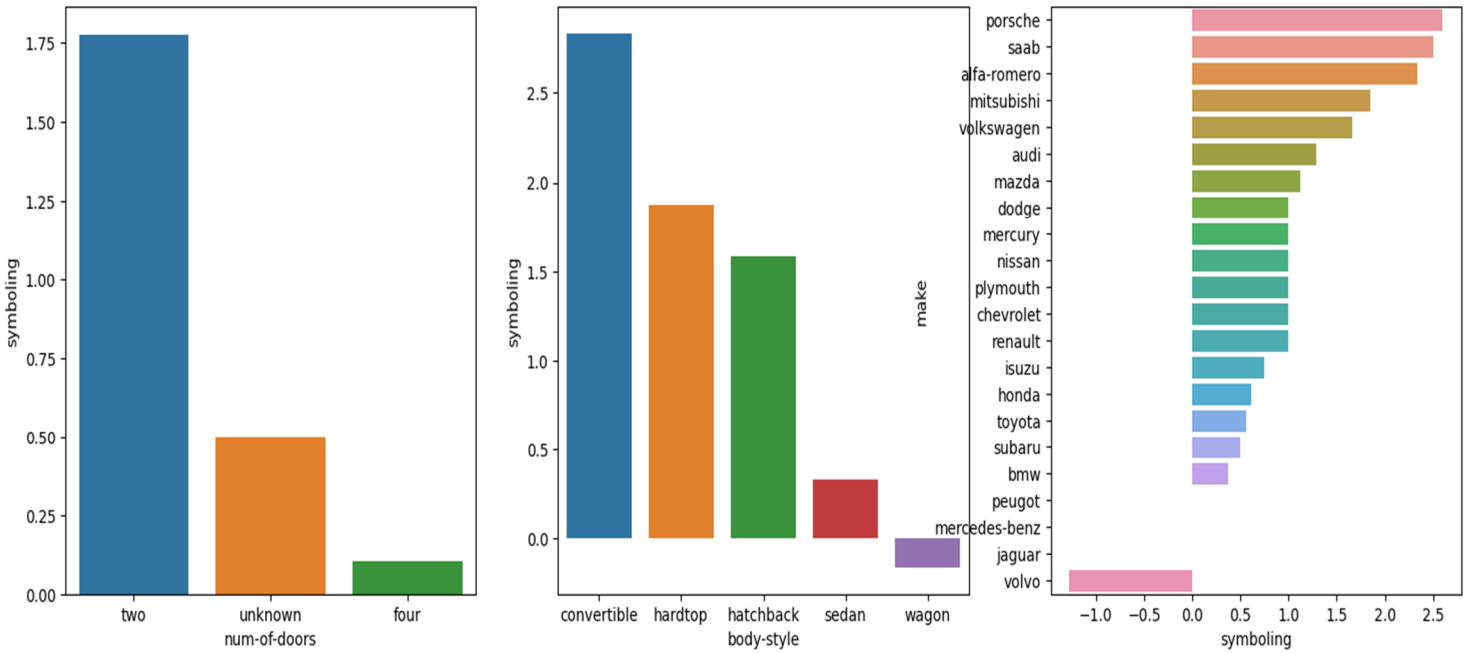
body-style_sedan: -0.378341

Highlights of Normalized Losses

Positive correlation: **num-of-doors_two: 0.348850**

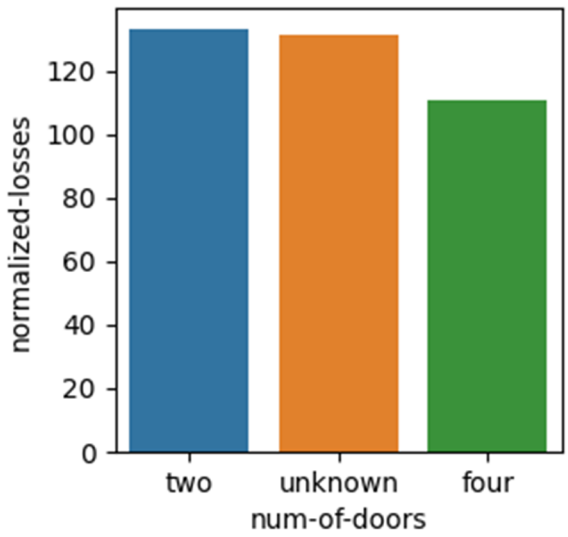
Negative correlation: **num-of-doors_four: -0.354755**

Symboling with different number of doors, body style and manufacturing company:



Volvo is the only manufacturing company getting negative mean symboling.

Normalized Losses with number of doors:



Conclusion:

The basic statistics of the dataset were explored and the EDA involves factors affecting:
1.price, fuel efficient: 2.city-mpg, 3.highway-mpg, 4.horsepower, 5.normalized-losses and 6.symboling
Price: Larger engine size, horsepower, with drive-wheels of rwd, make of mercedes-benz, system of mpfi increased the price.
Fuel efficient: The cars with higher compression-ratio, smaller engine, drive wheels of fwd, 4 cylinders, fuel-system of 2bbl, engine type of ohc, cheaper (lower price) and smaller size (curb-weight, length, width) had higher fuel efficient.
Horsepower: The cars with bigger engine size, fuel system of mpfi, drive wheelsof rwd, 6 cylinders, engine type of ohcv, make of porsche had higher horsepower.
Normalized-losses and symboling: The cars with 2 doors and body style of hatchback reduced the risks but the same time, the cars with 2 doors had higher normalized-losses.

THIS REPORT WAS WRITTEN BY : Chu, Alex Wai Leung Chu