Hyperiondev

TASK

Exploratory Data Analysis on the Automobile Data Set

Visit our website

THIS REPORT WAS WRITTEN BY: AVESHNEE IYER

CONTENTS

INTRODUCTION

Conduct an Exploratory Data Analysis (EDA) on the Automobile dataset to get further insights, clean the data and answer some questions.

Approach

- Finding patterns in data.
- Determining relationships in data.
- Checking of assumptions.
- Drawing conclusions.

What question(s) are you trying to solve (or prove wrong)?

1. Do the Body Size, Style and Engine Specification determine the car price?
2. Which type of cars are better in terms of mileage?
3. Which are highest selling cars based on brand, body style and price?
4. Which are highest normalized loss reported cars based on body style and no of doors?

DATA CLEANING

1. Data types:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 205 entries, 0 to 204
Data columns (total 26 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   symboling          205 non-null    int64
 1   normalized-losses  205 non-null    object
 2   make               205 non-null    object
 3   fuel-type          205 non-null    object
 4   aspiration         205 non-null    object
 5   num-of-doors       205 non-null    object
 6   body-style         205 non-null    object
 7   drive-wheels       205 non-null    object
 8   engine-location    205 non-null    object
 9   wheel-base         205 non-null    float64
 10  length             205 non-null    float64
 11  width              205 non-null    float64
 12  height             205 non-null    float64
 13  curb-weight        205 non-null    int64
 14  engine-type        205 non-null    object
 15  num-of-cylinders   205 non-null    object
 16  engine-size        205 non-null    int64
 17  fuel-system        205 non-null    object
 18  bore               205 non-null    object
 19  stroke             205 non-null    object
 20  compression-ratio  205 non-null    float64
 21  horsepower         205 non-null    object
 22  peak-rpm           205 non-null    object
 23  city-mpg           205 non-null    int64
 24  highway-mpg        205 non-null    int64
 25  price              205 non-null    object
dtypes: float64(5), int64(5), object(16)
memory usage: 41.8+ KB
```

Data volume - 205 rows.
Data columns - 26 columns.
Data Types - float64 (5), int64 (5), object (16).
From the NULL check, no null values found.

2. Data Snapshot (few rows):

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	engine-size	fuel-system	bore	stroke	compression-ratio	horsepower
0	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	111
1	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	111
2	1	?	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	...	152	mpfi	2.68	3.47	9.0	154
3	2	164	audi	gas	std	four	sedan	fwd	front	99.8	...	109	mpfi	3.19	3.40	10.0	102
4	2	164	audi	gas	std	four	sedan	4wd	front	99.4	...	136	mpfi	3.19	3.40	8.0	115

3. Feature Attributes:

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.

4. This data set consists of three sections:

1. The specification of an auto in terms of various characteristics.
2. Its assigned insurance risk rating.
3. Its normalized losses in use as compared to other cars.

5. Summary statistics of the dataframe

	symboling	wheel-base	length	width	height	curb-weight	engine-size	compression-ratio	city-mpg	highway-mpg
count	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000
mean	0.834146	98.756585	174.049268	65.907805	53.724878	2555.565854	126.907317	10.142537	25.219512	30.751220
std	1.245307	6.021776	12.337289	2.145204	2.443522	520.680204	41.642693	3.972040	6.542142	6.886443
min	-2.000000	86.600000	141.100000	60.300000	47.800000	1488.000000	61.000000	7.000000	13.000000	16.000000
25%	0.000000	94.500000	166.300000	64.100000	52.000000	2145.000000	97.000000	8.600000	19.000000	25.000000
50%	1.000000	97.000000	173.200000	65.500000	54.100000	2414.000000	120.000000	9.000000	24.000000	30.000000
75%	2.000000	102.400000	183.100000	66.900000	55.500000	2935.000000	141.000000	9.400000	30.000000	34.000000
max	3.000000	120.900000	208.100000	72.300000	59.800000	4066.000000	326.000000	23.000000	49.000000	54.000000

According to the count, there seems to be no missing values - however on closer inspecti on, the null values have been filled with a question mark (?) and therefore do not show a s missing values in the count.

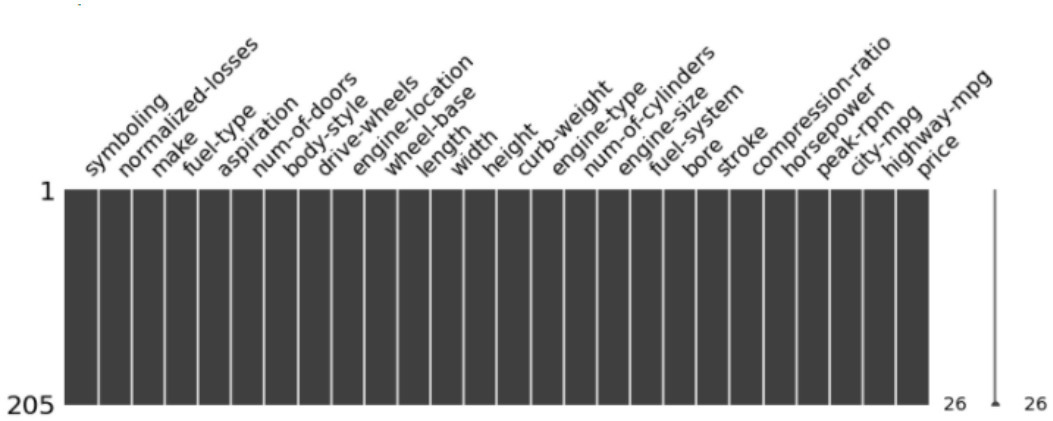
In a normal distribution, about 68% of the scores are within one standard deviation of th e mean and about 95% of the scores are within two standard deviations of the mean. Sta ndard deviations are quite high for Curb-weight implying the variation could be large.

The upper quartile (sometimes called Q3) is the number dividing the third and fourth qu artile. The upper quartile can also be thought of as the median of the upper half of the n umbers. The upper quartile is also called the 75th percentile; it splits the lowest 75% of d ata from the highest 25%

Minimum values - the minus value under symbolling referring to ratings, no other colum ns have negative values.

MISSING DATA

1. Visualised Missing values:



There seems to be no missing values in the visualisation.

2. Missing values data points and count:

symboling	0
normalized-losses	0
make	0
fuel-type	0
aspiration	0
num-of-doors	0
body-style	0
drive-wheels	0
engine-location	0
wheel-base	0
length	0
width	0
height	0
curb-weight	0
engine-type	0

dtype: int64

The above does not reflect any missing values

3. Actual missing values:

From investigation, the missing values have been replaced with a question mark (?) and hence do not show in the missing value counts above.
In order to solve for missing values (?) with NULL value, identify the number of missing values and then apply a median or mean to fill values in.

The following columns have missing values and need to be cleaned:

1. normalized-losses – 41
2. price – 4
3. horsepower – 2
4. bore – 4
5. stroke – 4
6. peak-rpm – 2
7. num-of-doors – 2

4. Process applied for missing data:

1) Normalized-losses:

a. # Cleaning the NORMALISED LOSSES field
b. # Find out number of records having 'NaN' value for normalized losses
c. # Setting the missing value to mean of normalized losses and convert the datatype to integer

2) Price:

a. # Find out the number of values which are not numeric using Boolean
b. # List out the values which are not numeric
c. #Setting the missing value to mean of price and convert the datatype to integer

3) Horsepower:

a. # Cleaning the HORSEPOWER
b. # Checking the numeric and replacing with mean value and convert the datatype to integer
c. # Checking the outlier of horsepower
d. # Excluding the Outlier data for horsepower

4) Bore

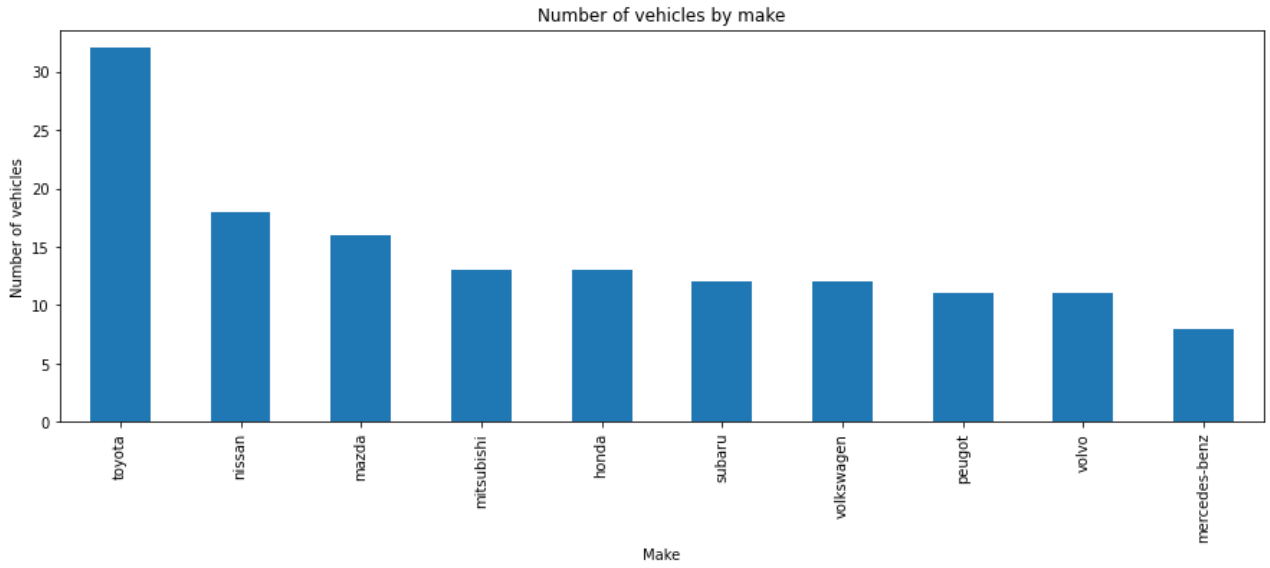
a. # Cleaning BORE
b. # Find out the number of invalid values
c. # Replace the non-numeric value to null and convert the datatype

5) **Stroke**
   a. # Cleaning the STROKE
   b. # Replace the non-number value to null and convert the datatype
6) **Peak RPM**
   a. # Cleaning the STROKE
   b. # Convert the non-numeric data to null and convert the datatype

7) **Num of doors**
   a. # Cleaning the num-of-doors data
   b. # remove the records which are having the value '?'

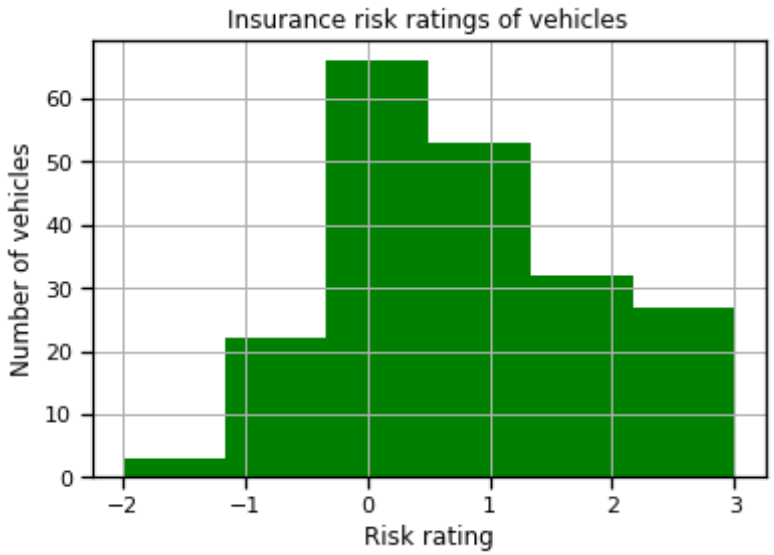
DATA STORIES AND VISUALIZATIONS

1. Vehicle make frequency



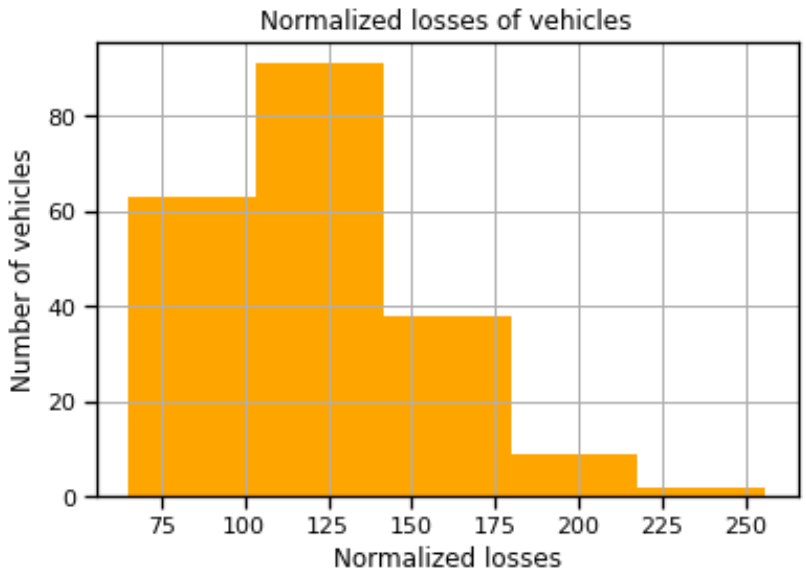
Toyota by far exceeds the other brands on the data set, almost @ 40%. Nissan is the 2nd highest. The lowest is Mercedes probably as it is a more niche vehicle.

2. Insurance risk ratings Histogram



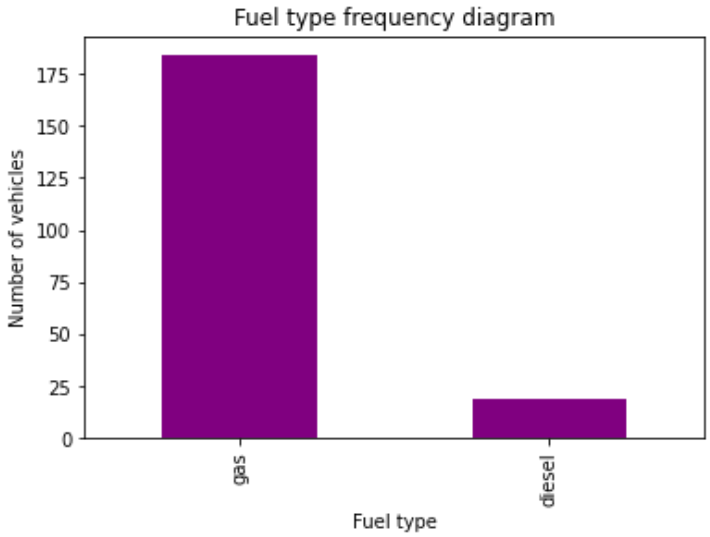
The insurance risk ratings range between -3 and 3. This dataset starts from -2 to 3. More cars fall in the range of 0 and 1 in this case.

3. Normalised losses Histogram



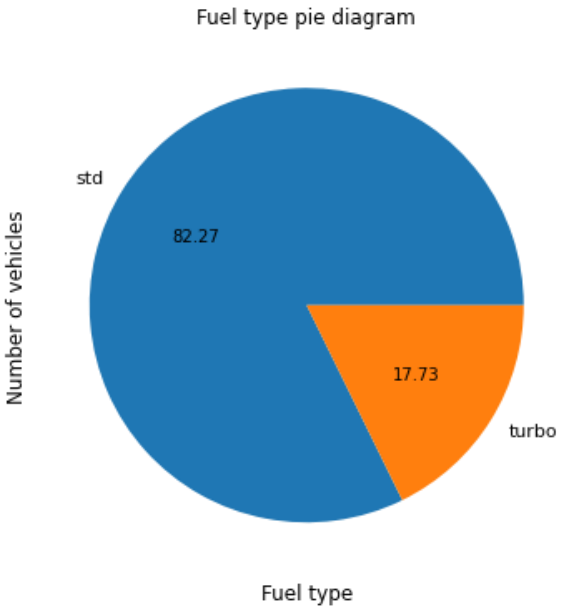
Normalized losses - average loss payment per insured vehicle year – sits mostly between the range 65 and 150.

4. Fuel type Bar chart



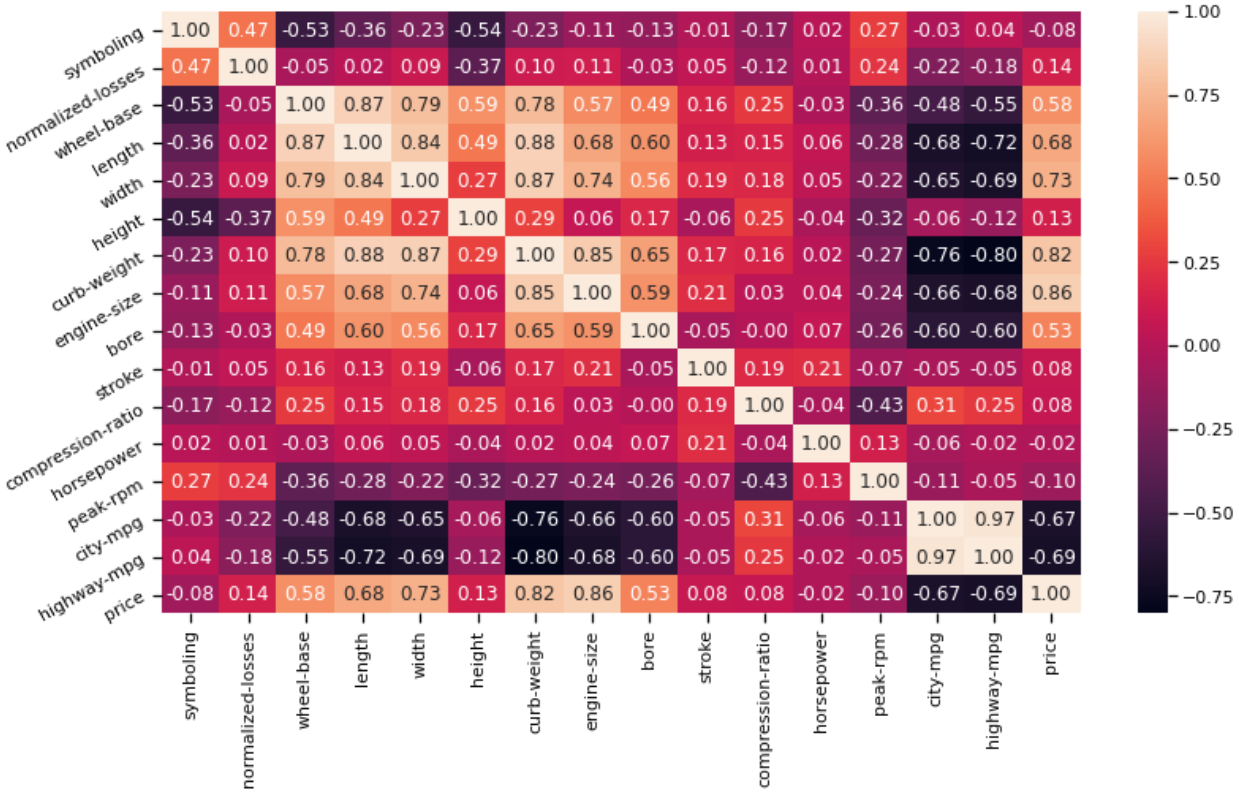
The fuel type is predominately diesel.

5. Fuel type Pie chart



82% of the fuel type is standard versus turbo.

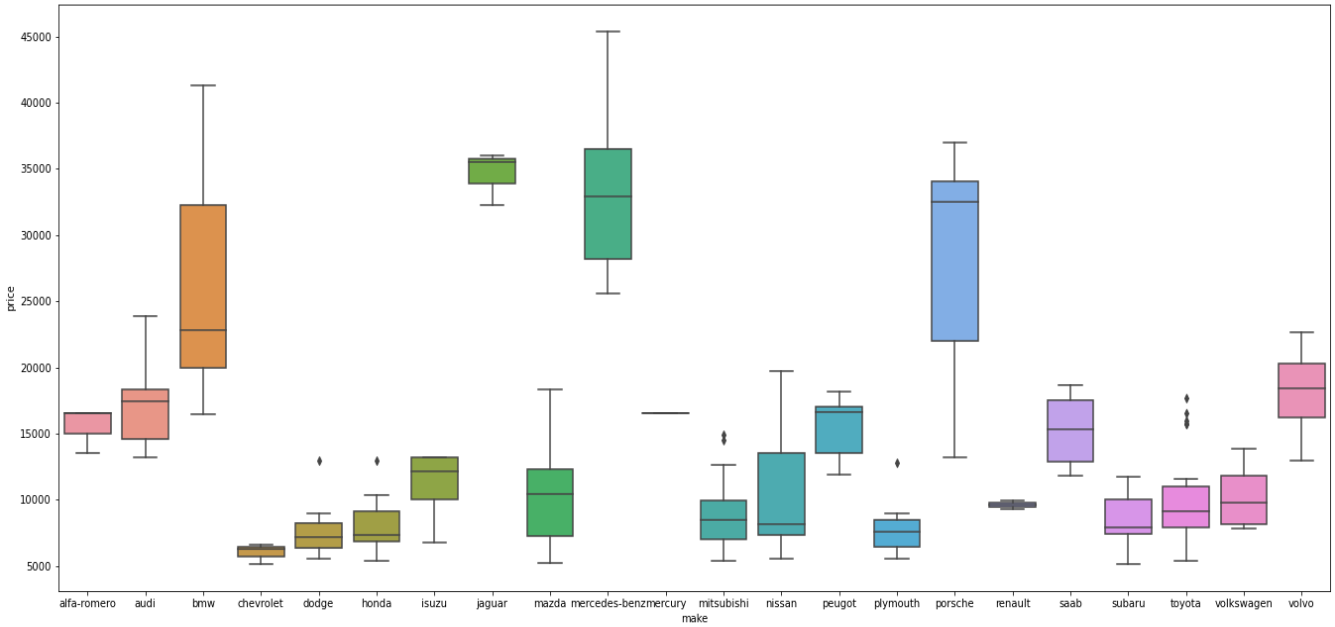
6. Heatmap showing correlation between features



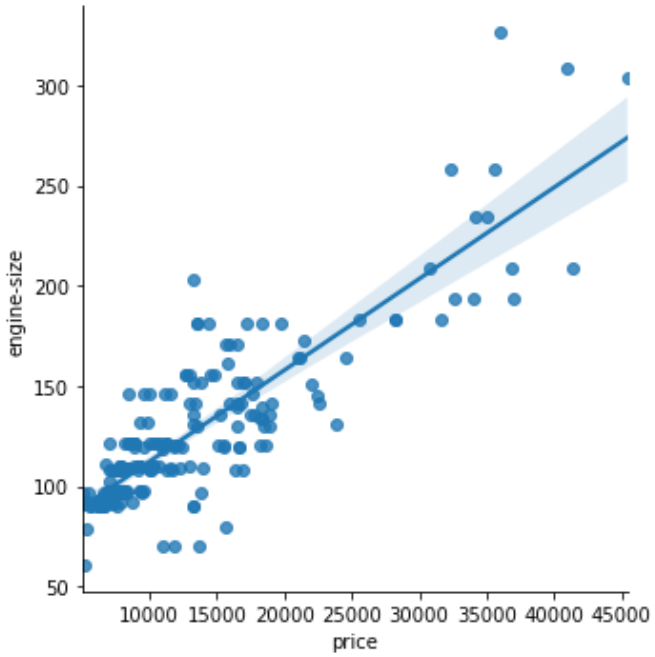
Price is more correlated with engine size and curb-weight of the car.
Curb-weight is mostly correlated with engine size, length, width, and wheel base - which is expected as these add up the weight of the car.
Wheel-base is highly correlated with the length and width of the car.
Symboling and normalized car are correlated more than the other fields.

7. Price and Make Box Plot

The most expensive make is Mercedes Benz and the least expensive is Chevrolet.
The premium cars are BMW, Jaguar, Mercedes Benz, and Porsche.
Less expensive cars are Chevrolet, Dodge, Honda, Mitsubishi, Plymoth and Subaru.
The mid-range has the highest number of cars.

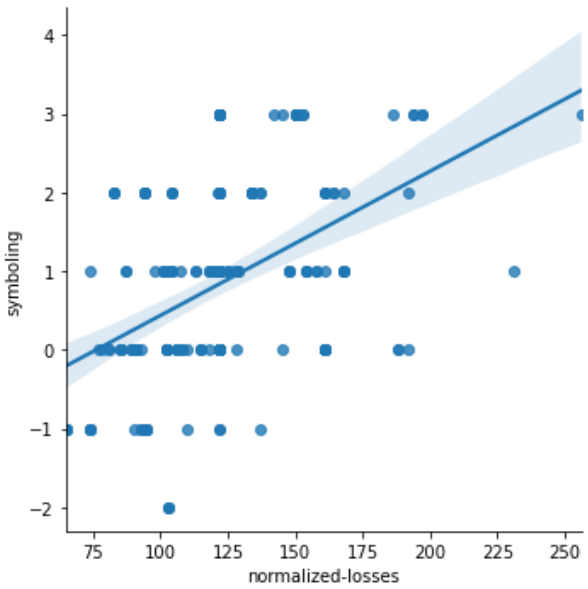


8. Scatter plot of price and engine size



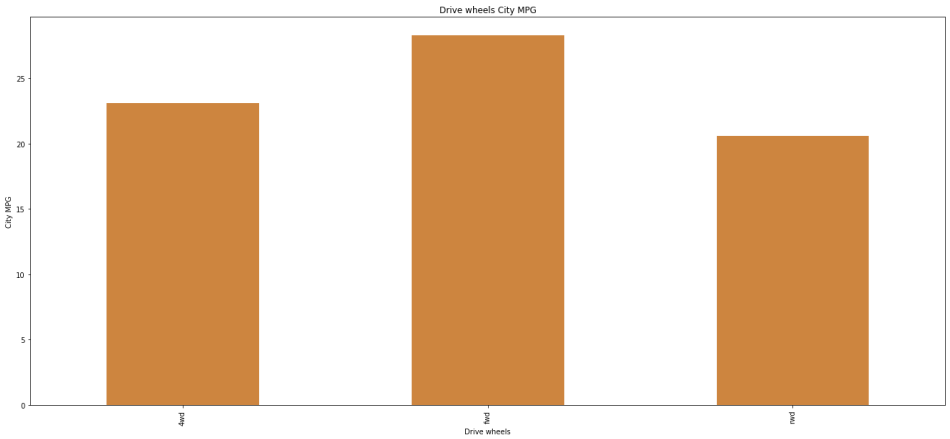
Engine size and price is positively correlated.

9. Scatter plot of normalized losses and symbolling

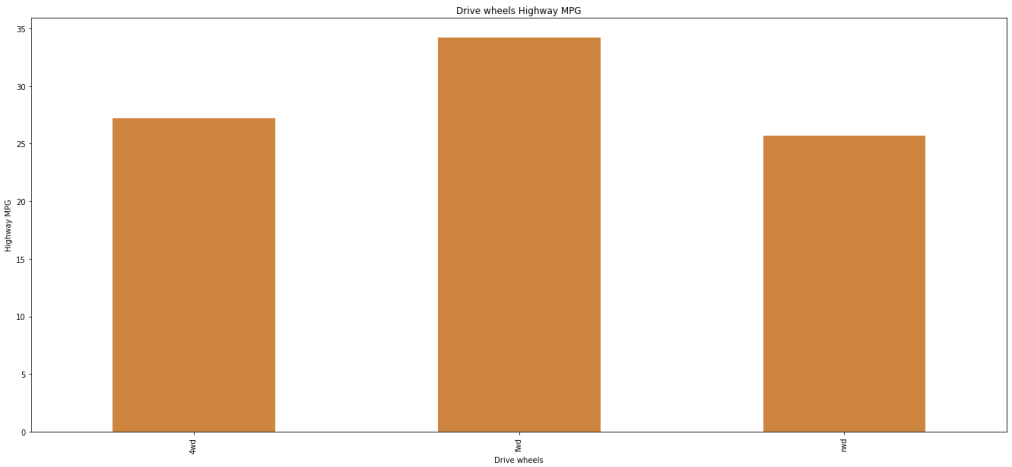


The lower the rating, the lower the normalised loss.

10. Drive wheels and City MPG bar chart

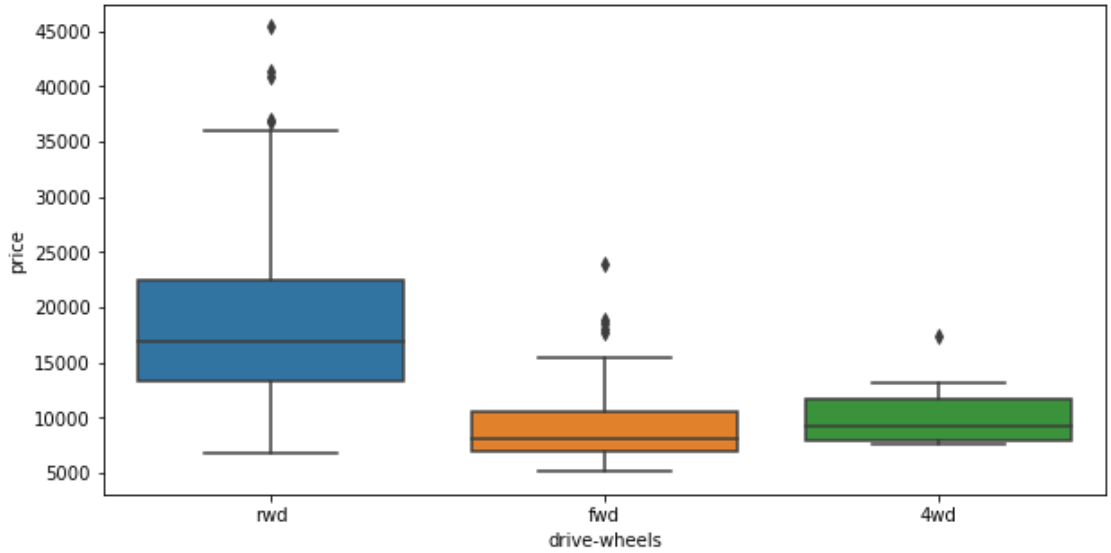


11. Drive wheels and Highway MPG bar chart



12. Boxplot of Drive wheels and Price

Rear wheel are the most expensive and then 4-wheel drives. Front wheel is the least expensive.



13. Normalized losses based on body style and no. of doors

Two-door cars have more losses than four door cars.