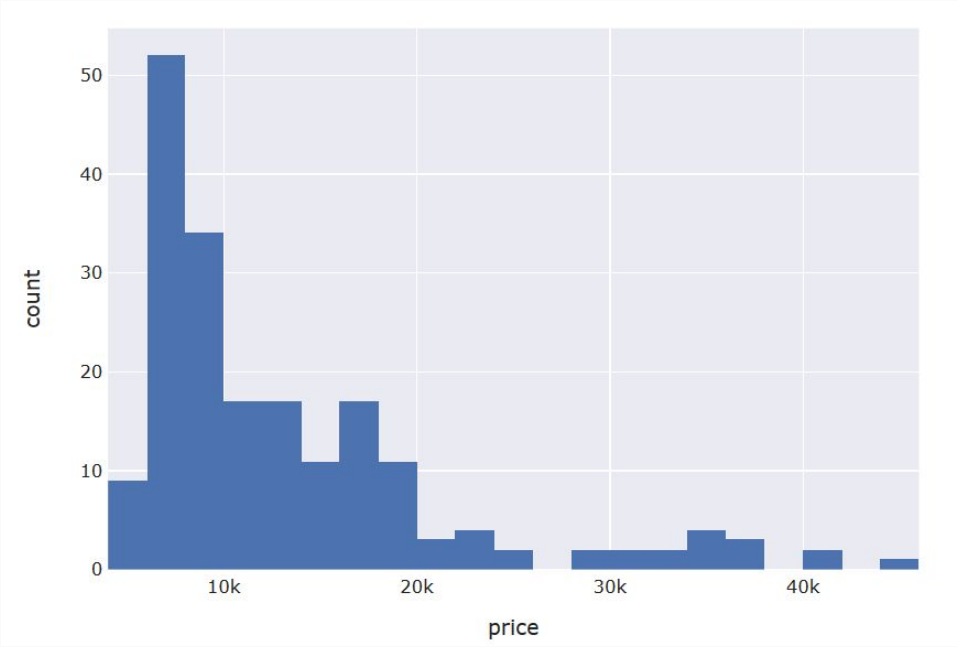
Automobile Pricing

Gaurav Kumar

Automobile Pricing

Automobile Price Spread

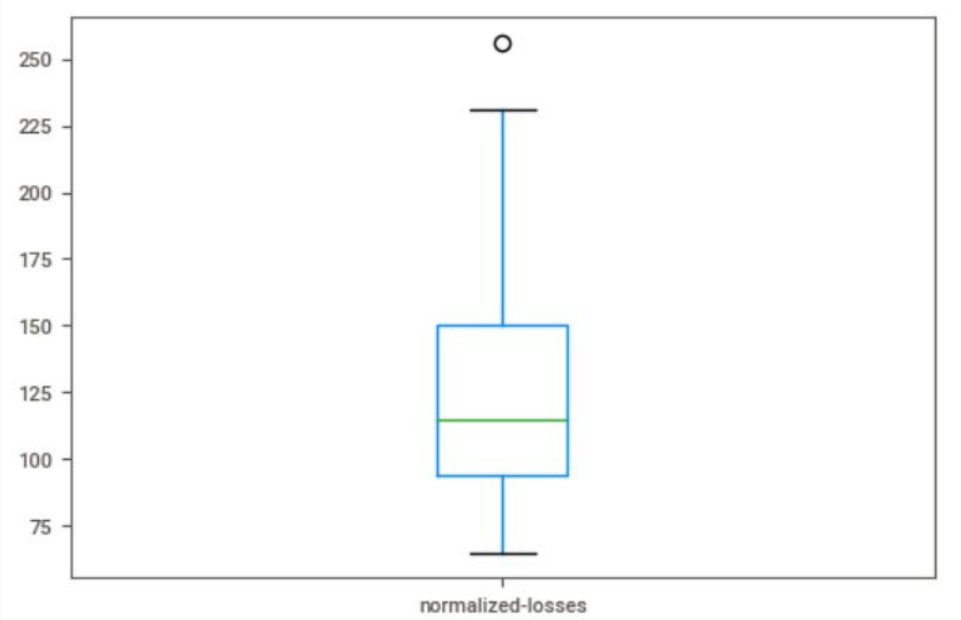


skew: 1.8
kurtosis 3.12

Skewed Data to Right with outliers
ML Model is proposed with log transformation of price

Automobile Pricing

Cleaning [normalized-losses]




```
df['normalized-losses'].value_counts()
```

```
?       41
```

There are 41 invalid entries, which is high number, thus is replaced with median , instead of dropping these Median Used instead of Mean due to outliers, Apart from this column, row entry has been deleted with invalid entry, since number of invalid entries <=4

Automobile Pricing

Cleaning [num-of-doors & num-of-cylinders]

```
four    114
two      89
?         2
Name: num-of-doors, dtype: int64
```

```
4    114
2     89
Name: num-of-doors, dtype: int64
```

```
four      157
six        24
five       11
eight       5
two         4
twelve      1
three       1
Name: num-of-cylinders, dtype: int64
```

```
4      157
6       24
5       11
8        5
2        4
12       1
3        1
Name: num-of-cylinders, dtype: int64
```

Since there is numerical ordering possible , it is converted to numbers directly, and since only 2 rows have invalid entry, it is deleted

Automobile Pricing

Final Features

Categorical

	make	fuel-type	aspiration	body-style	drive-wheels	engine-location	engine-type	fuel-system
0	alfa-romero	gas	std	convertible	rwd	front	dohc	mpfi
1	alfa-romero	gas	std	convertible	rwd	front	dohc	mpfi

Numerical

symboling	normalized-losses	num-of-doors	wheel-base	length	width	height	curb-weight	num-of-cylinders	engine-size	bore	stroke	compression-ratio	horsepower	peak-rpm	city-mpg	highway-mpg
3	115	2	88.6	168.8	64.1	48.8	2548	4	130	3.47	2.68	9.0	111	5000.0	21	27
3	115	2	88.6	168.8	64.1	48.8	2548	4	130	3.47	2.68	9.0	111	5000.0	21	27

Automobile Pricing

Handling Categorical Features

Actual

	make	fuel-type	aspiration	body-style	drive-wheels	engine-location	engine-type	fuel-system
0	alfa-romero	gas	std	convertible	rwd	front	dohc	mpfi
1	alfa-romero	gas	std	convertible	rwd	front	dohc	mpfi

Encoded

	make_alfa-romero	make_audi	make_bmw	make_chevrolet	make_dodge	make_honda	make_isuzu	make_jaguar	make_mazda	make_mercedes-benz	make_mercury	m
0	1	0	0	0	0	0	0	0	0	0	0	
1	1	0	0	0	0	0	0	0	0	0	0	

2 rows × 47 columns

make_mitsubishi	make_nissan	make_peugot	make_plymouth	...	drive-wheels_rwd	engine-location_front	engine-location_rear	engine-type_dohc	engine-type_l	engine-type_ohc	engine-type_ohcf	engine type_ohcv
0	0	0	0	...	1	1	0	1	0	0	0	(
0	0	0	0	...	1	1	0	1	0	0	0	(

etc...

Automobile Pricing

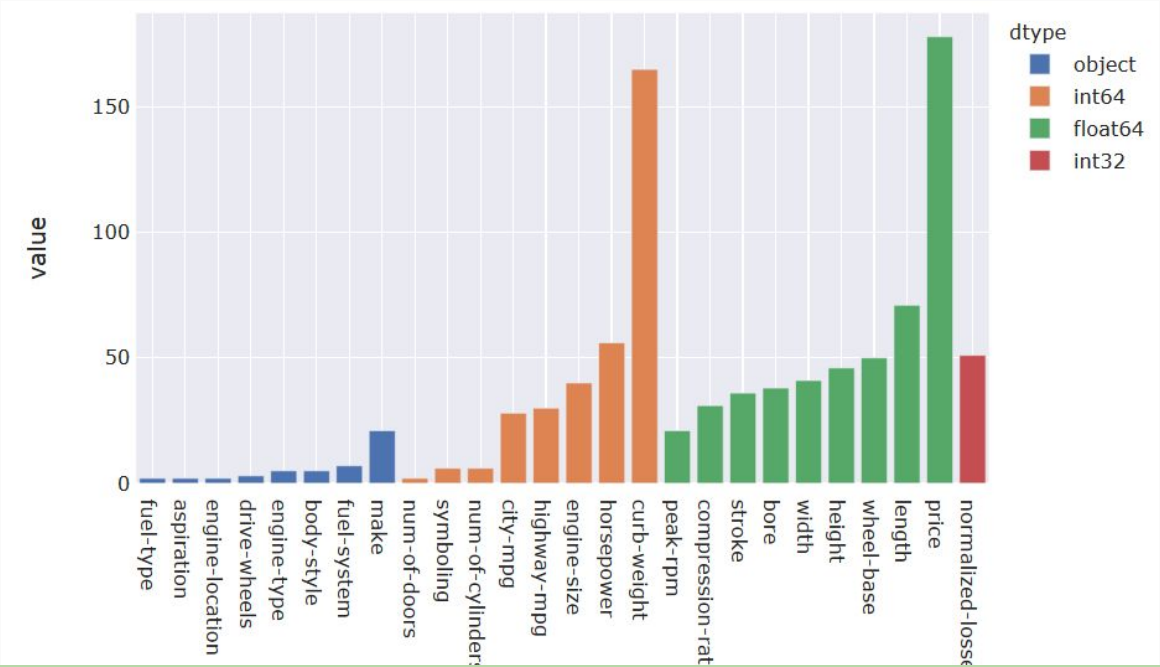
DataFrame for Machine Learning Model

	symboling	normalized-losses	num-of-doors	wheel-base	length	width	height	curb-weight	num-of-cylinders	engine-size	bore	stroke	compression-ratio	horsepower	peak-rpm	...	drive-wheels_rwd
0	3	115	2	88.6	168.8	64.1	48.8	2548	4	130	3.47	2.68	9.0	111	5000.0	...	1
1	3	115	2	88.6	168.8	64.1	48.8	2548	4	130	3.47	2.68	9.0	111	5000.0	...	1
2	1	115	2	94.5	171.2	65.5	52.4	2823	6	152	2.68	3.47	9.0	154	5000.0	...	1
3	2	164	4	99.8	176.6	66.2	54.3	2337	4	109	3.19	3.40	10.0	102	5500.0	...	0
4	2	164	4	99.4	176.6	66.4	54.3	2824	5	136	3.19	3.40	8.0	115	5500.0	...	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
200	-1	95	4	109.1	188.8	68.9	55.5	2952	4	141	3.78	3.15	9.5	114	5400.0	...	1
201	-1	95	4	109.1	188.8	68.8	55.5	3049	4	141	3.78	3.15	8.7	160	5300.0	...	1
202	-1	95	4	109.1	188.8	68.9	55.5	3012	6	173	3.58	2.87	8.8	134	5500.0	...	1
203	-1	95	4	109.1	188.8	68.9	55.5	3217	6	145	3.01	3.40	23.0	106	4800.0	...	1
204	-1	95	4	109.1	188.8	68.9	55.5	3062	4	141	3.78	3.15	9.5	114	5400.0	...	1

193 rows × 65 columns

Automobile Pricing

DataFrame dtype wise unique values

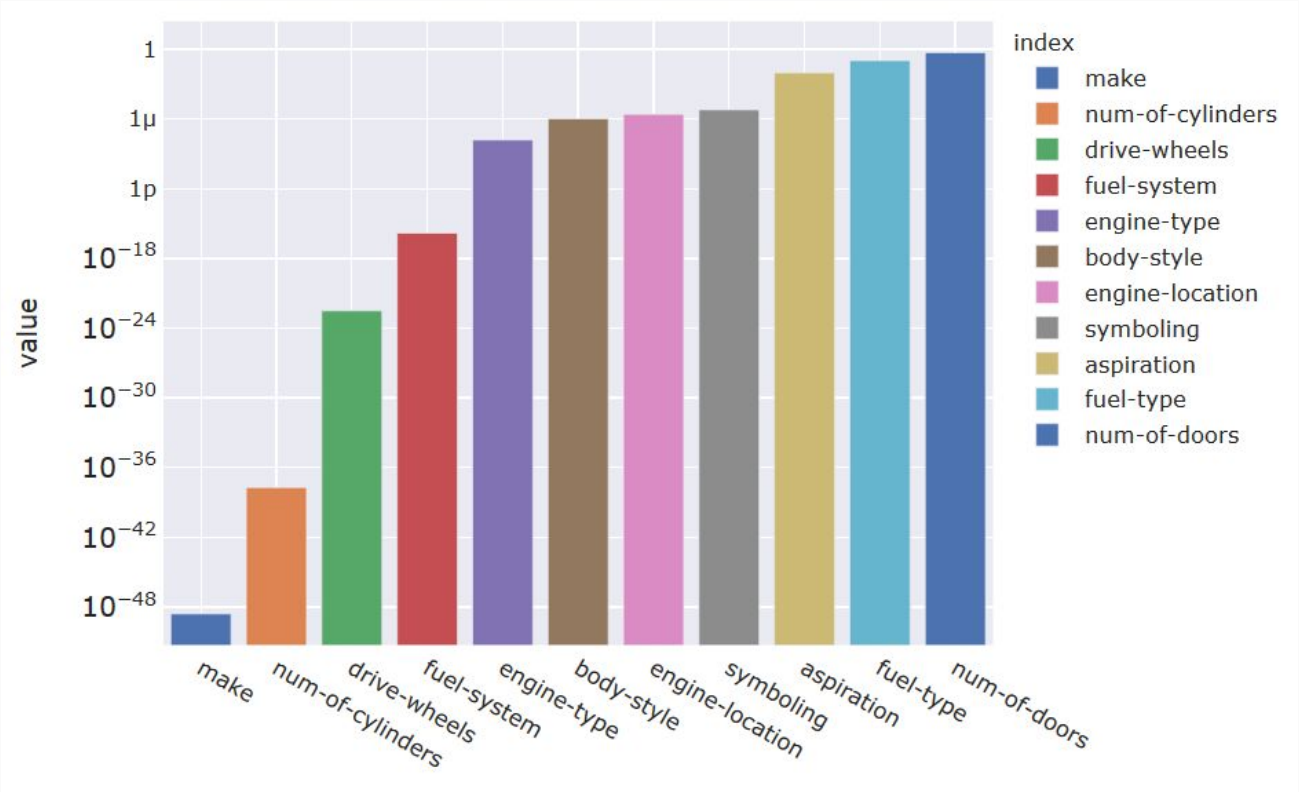


Since total number of columns in ML model is 65, which seems large compared to only 193 rows, in case of overfitting, make can be removed since it alone is contributing one hot 21 columns

Important Categorical Features

Automobile Pricing

Influencing factors using Anova test for relationship



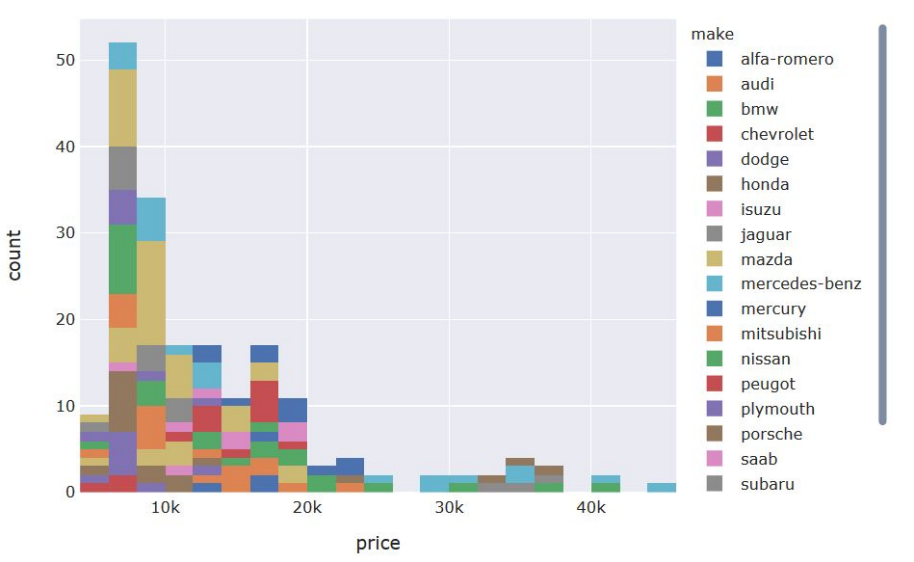
Low Statistical P Values

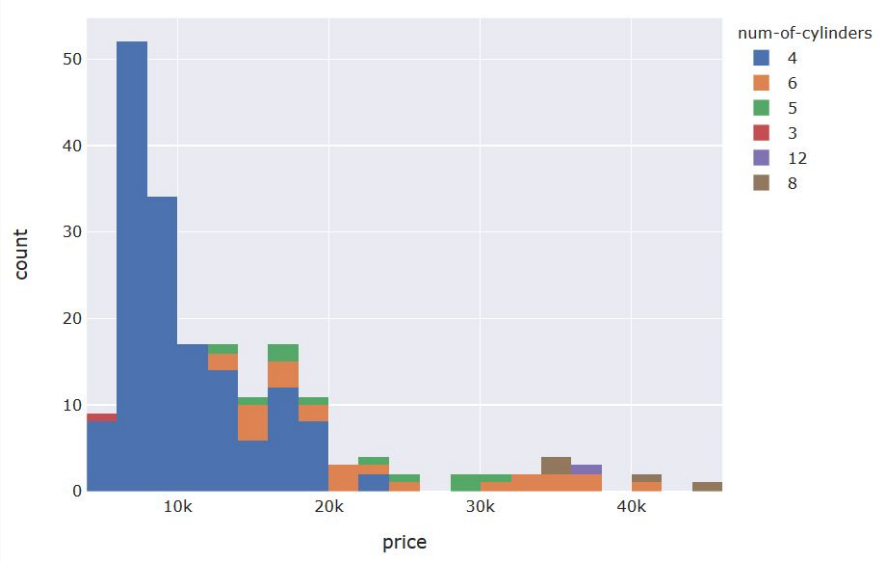
High Possibility Of Correlation

Top 4 features will be studied for inferences

Automobile Pricing

Important Categorical features I & II





Some companies are higher in prices which is also leading to outliers
High number of cylinders is equivalent to costlier vehicle

Automobile Pricing

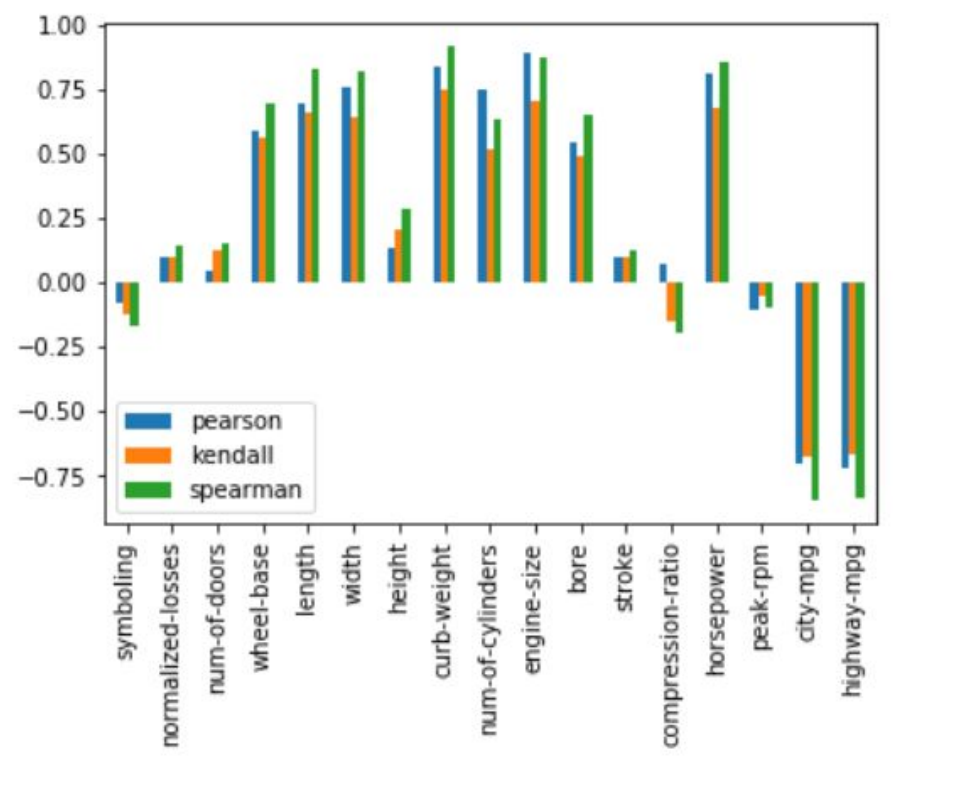
Important Categorical features III & IV

Fwd are relatively cheaper
Some fuel systems are on cheaper vehicle than others

Important Numerical Features

Automobile Pricing

Influencing factors using correlation coeff. for relationship



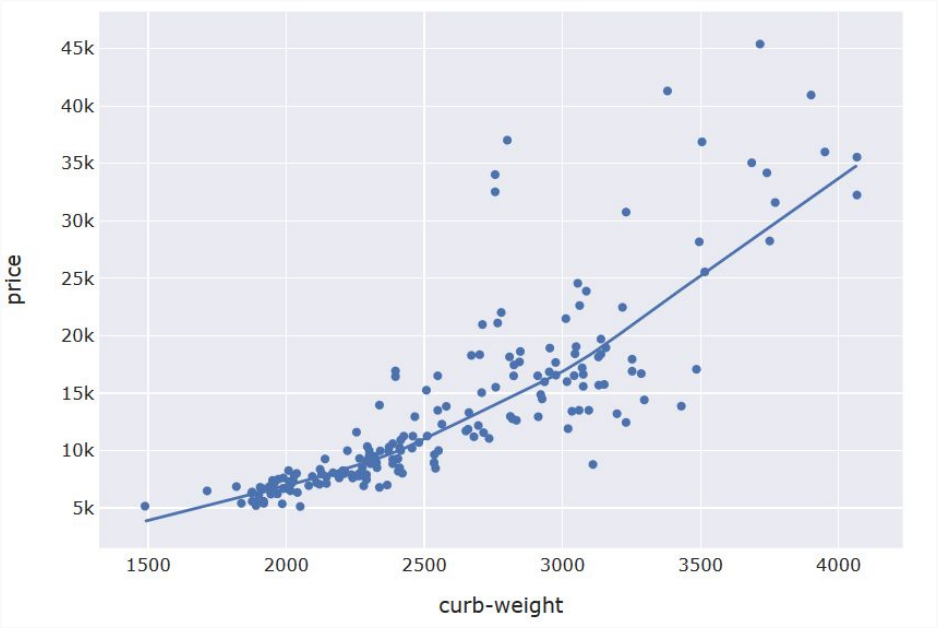
Correlation coeff..

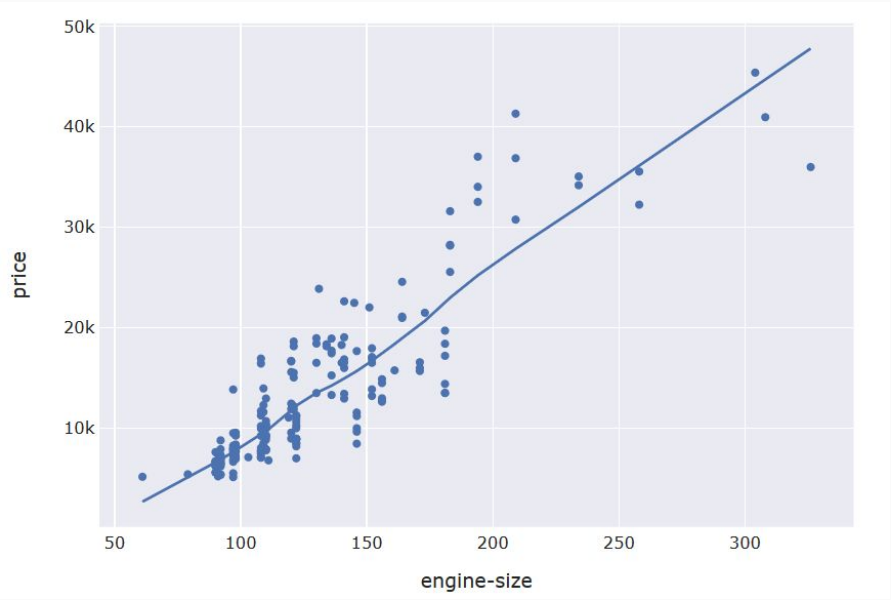
-1	0	+1
Important Features...	Not so Important Features...	Important Features...

Top 4 features will be studied for inferences

Automobile Pricing

Important Numerical features I & II

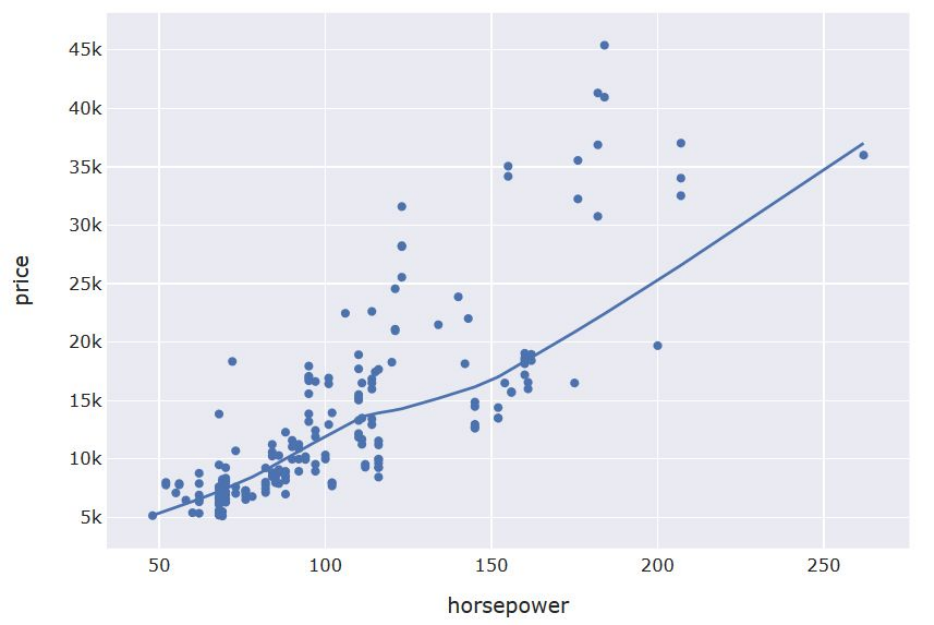


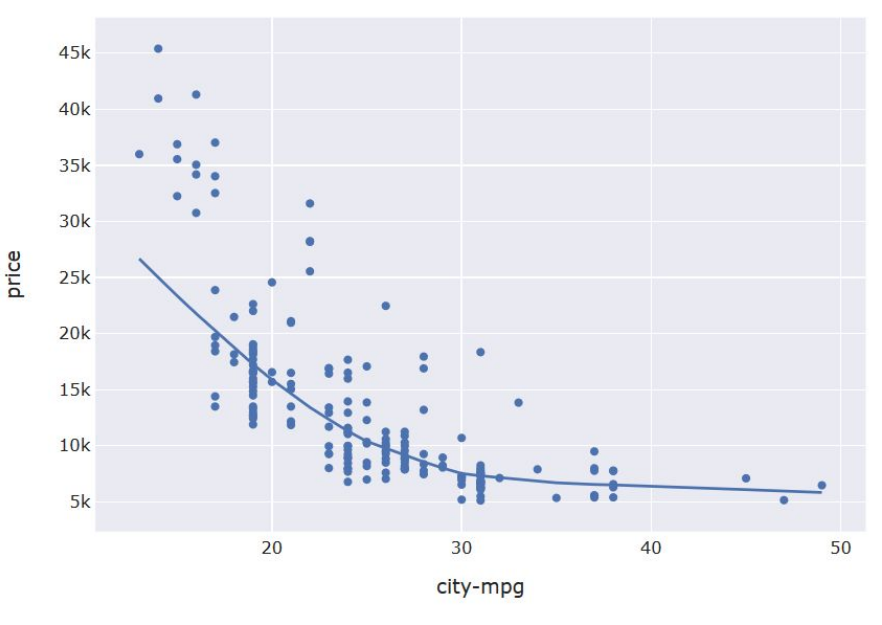


Linear/exponential correlations

Automobile Pricing

Important Numerical features III & IV





Linear/exponential correlations

Automobile Pricing

Interrelated features

city-mpg	highway-mpg	0.971975
length	curb-weight	0.882694
wheel-base	length	0.879307
width	curb-weight	0.867640
length	width	0.857368
curb-weight	engine-size	0.857188
engine-size	horsepower	0.845325
wheel-base	width	0.818465

Some Features are highly correlated among themselves