Overview

Brought to you by YData (https://ydata.ai/?utm_source=opensource&utm_medium=ydataprofiling&utm_campaign=report)

Dataset statistics

Number of variables	6
Number of observations	815
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	76
Duplicate rows (%)	9.3%
Total size in memory	35.0 KiB
Average record size in memory	44.0 B

Variable types

Text	1
Categorical	2
Numeric	3

Alerts

Dataset has 76 (9.3%) duplicate rows	Duplicates
`Price` is highly overall correlated with `company` and 1 other fields (company, year)	High correlation
`company` is highly overall correlated with `Price`	High correlation
`year` is highly overall correlated with `Price`	High correlation

Reproduction

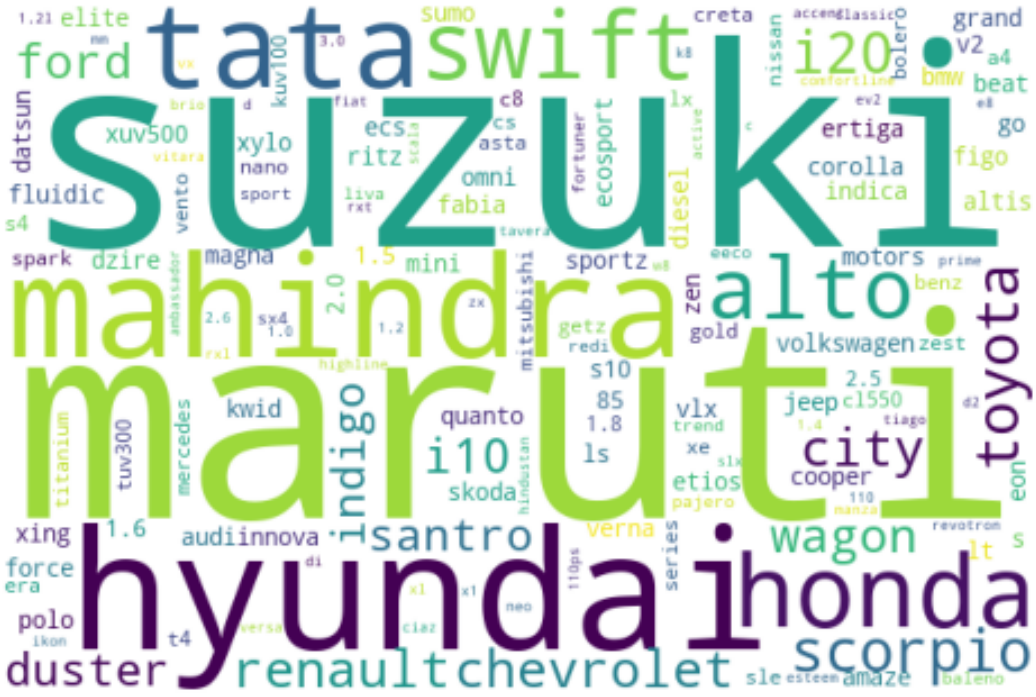
Analysis started	2024-09-25 08:31:22.921533
Analysis finished	2024-09-25 08:31:27.792019
Duration	4.87 seconds
Software version	ydata-profiling vv4.10.0 (https://github.com/ydataai/ydata-profiling)
Download configuration	config.json (data:text/plain;charset=utf-8,%7B%22title%22%3A%20%22Pandas%20Profiling%20Report%22%2C%20%22dataset%22%3A%20%7B%22description%22%3A%20%22%22%2

Variables

Select Columns

name
Text

Distinct	254
Distinct (%)	31.2%
Missing	0
Missing (%)	0.0%
Memory size	12.7 KiB



Length

Max length	28
Median length	24
Mean length	17.03681
Min length	6

Characters and Unicode

Total characters	13885	
Distinct characters	63	
Distinct categories	1 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)	?
Distinct scripts	1 (https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)	?

Distinct blocks	1 (https://en.wikipedia.org/wiki/Unicode_block)	?

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

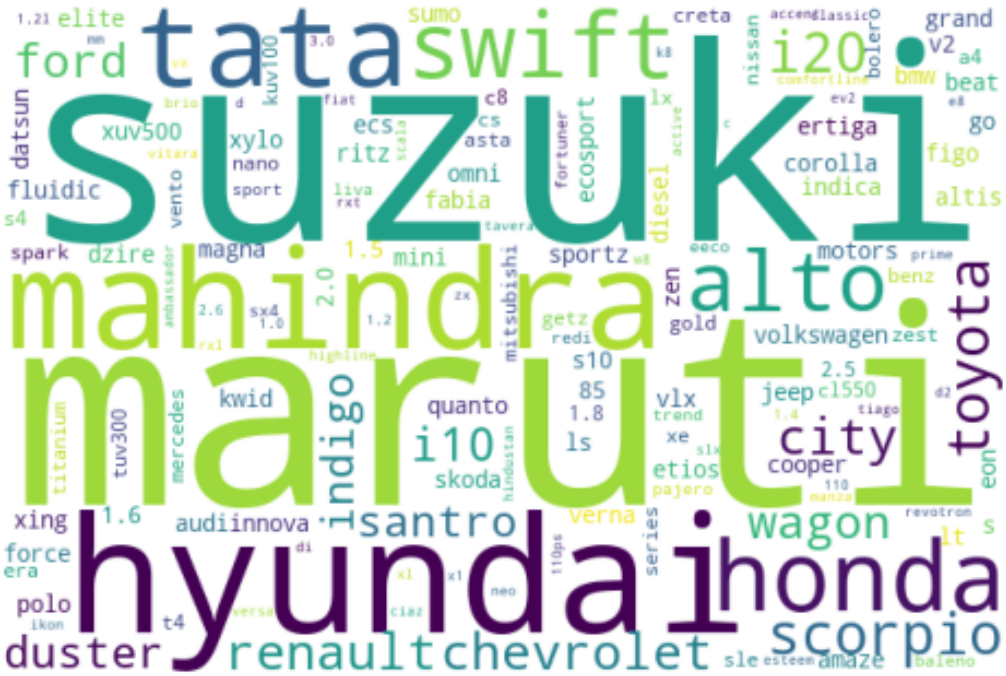
Unique

Unique	124	?
Unique (%)	15.2%	

Sample

1st row	Hyundai Santro Xing
2nd row	Mahindra Jeep CL550
3rd row	Hyundai Grand i10
4th row	Ford EcoSport Titanium
5th row	Ford Figo

Value	Count	Frequency (%)
maruti	224	9.5%
suzuki	221	9.4%
hyundai	139	5.9%
mahindra	97	4.1%
tata	65	2.8%
honda	60	2.5%
swift	51	2.2%
alto	42	1.8%
scorpio	39	1.7%
i20	38	1.6%
Other values (269)	1377	58.5%



Most occurring characters

Value	Count	Frequency (%)
	1538	11.1%
a	1242	8.9%
i	1212	8.7%
u	944	6.8%
t	801	5.8%
o	730	5.3%
r	724	5.2%
n	686	4.9%
e	491	3.5%
S	485	3.5%
Other values (53)	5032	36.2%

Most occurring categories

Value	Count	Frequency (%)
(unknown)	13885	100.0%

Most frequent character per category

(unknown)

Value	Count	Frequency (%)
	1538	11.1%
a	1242	8.9%
i	1212	8.7%
u	944	6.8%
t	801	5.8%
o	730	5.3%
r	724	5.2%
n	686	4.9%
e	491	3.5%
S	485	3.5%
Other values (53)	5032	36.2%

Most occurring scripts

Value	Count	Frequency (%)
(unknown)	13885	100.0%

Most frequent character per script

(unknown)

Value	Count	Frequency (%)
	1538	11.1%
a	1242	8.9%
i	1212	8.7%
u	944	6.8%
t	801	5.8%
o	730	5.3%
r	724	5.2%
n	686	4.9%
e	491	3.5%
S	485	3.5%
Other values (53)	5032	36.2%

Most occurring blocks

Value	Count	Frequency (%)
(unknown)	13885	100.0%

Most frequent character per block

(unknown)

Value	Count	Frequency (%)
	1538	11.1%
a	1242	8.9%
i	1212	8.7%
u	944	6.8%
t	801	5.8%
o	730	5.3%
r	724	5.2%
n	686	4.9%
e	491	3.5%
S	485	3.5%
Other values (53)	5032	36.2%

company

Categorical

HIGH CORRELATION (This variable has a high overall correlation with 1 fields: Price)

Distinct	25
Distinct (%)	3.1%
Missing	0
Missing (%)	0.0%
Memory size	12.7 KiB

Length

Max length	10
Median length	9
Mean length	6.3030675
Min length	3

Characters and Unicode

Total characters	5137	
Distinct characters	35	
Distinct categories	1 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)	?
Distinct scripts	1 (https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)	?
Distinct blocks	1 (https://en.wikipedia.org/wiki/Unicode_block)	?

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	3	?
Unique (%)	0.4%	

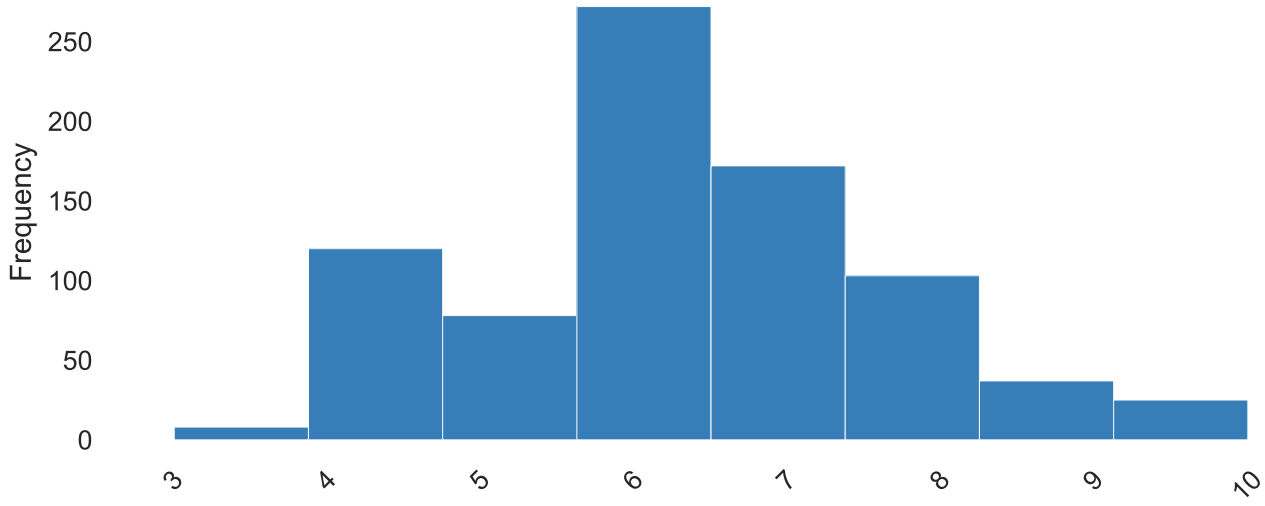
Sample

1st row	Hyundai
2nd row	Mahindra
3rd row	Hyundai
4th row	Ford
5th row	Ford

Common Values

Value	Count	Frequency (%)
Maruti	221	27.1%
Hyundai	139	17.1%
Mahindra	97	11.9%
Tata	65	8.0%
Honda	60	7.4%
Toyota	36	4.4%
Chevrolet	34	4.2%
Renault	33	4.0%
Ford	30	3.7%
Volkswagen	19	2.3%
Other values (15)	81	9.9%

Length



Histogram of lengths of the category

Value	Count	Frequency (%)
maruti	221	27.1%
hyundai	139	17.1%
mahindra	97	11.9%
tata	65	8.0%
honda	60	7.4%
toyota	36	4.4%
chevrolet	34	4.2%
renault	33	4.0%
ford	30	3.7%
volkswagen	19	2.3%
Other values (15)	81	9.9%

Most occurring characters

Value	Count	Frequency (%)
a	870	16.9%
i	515	10.0%
u	422	8.2%
t	409	8.0%
r	394	7.7%
n	376	7.3%
d	360	7.0%
M	346	6.7%
o	234	4.6%
H	202	3.9%
Other values (25)	1009	19.6%

Most occurring categories

Value	Count	Frequency (%)
(unknown)	5137	100.0%

Most frequent character per category

(unknown)

Value	Count	Frequency (%)
a	870	16.9%
i	515	10.0%
u	422	8.2%
t	409	8.0%
r	394	7.7%
n	376	7.3%
d	360	7.0%
M	346	6.7%
o	234	4.6%
H	202	3.9%
Other values (25)	1009	19.6%

Most occurring scripts

Value	Count	Frequency (%)
(unknown)	5137	100.0%

Most frequent character per script

(unknown)

Value	Count	Frequency (%)
a	870	16.9%
i	515	10.0%
u	422	8.2%
t	409	8.0%
r	394	7.7%
n	376	7.3%
d	360	7.0%
M	346	6.7%
o	234	4.6%
H	202	3.9%
Other values (25)	1009	19.6%

Most occurring blocks

Value	Count	Frequency (%)
(unknown)	5137	100.0%

Most frequent character per block

(unknown)

Value	Count	Frequency (%)
a	870	16.9%
i	515	10.0%
u	422	8.2%
t	409	8.0%
r	394	7.7%
n	376	7.3%
d	360	7.0%
M	346	6.7%
o	234	4.6%
H	202	3.9%
Other values (25)	1009	19.6%

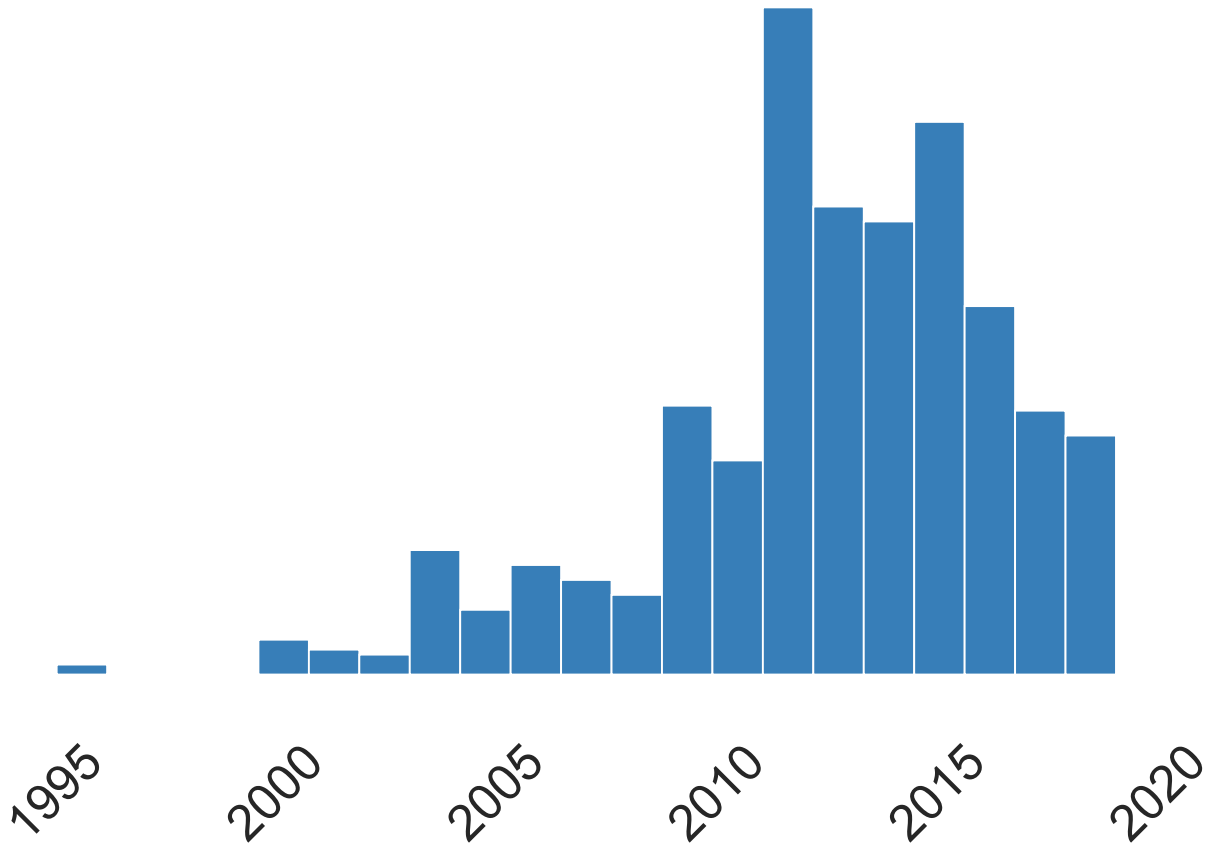
year

Real number (ℝ)

HIGH CORRELATION (This variable has a high overall correlation with 1 fields: Price)

Distinct	21
Distinct (%)	2.6%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	2012.4429

Minimum	1995
Maximum	2019
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	9.6 KiB



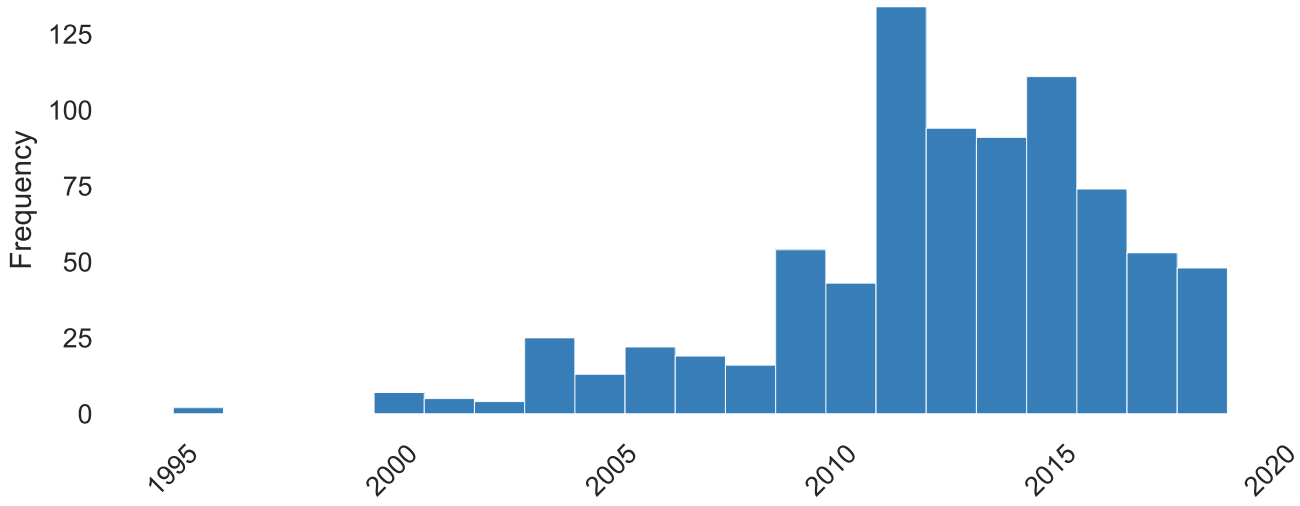
Quantile statistics

Minimum	1995
5-th percentile	2004
Q1	2010

median	2013
Q3	2015
95-th percentile	2018
Maximum	2019
Range	24
Interquartile range (IQR)	5

Descriptive statistics

Standard deviation	4.0050792
Coefficient of variation (CV)	0.0019901579
Kurtosis	1.1376741
Mean	2012.4429
Median Absolute Deviation (MAD)	2
Skewness	-1.0017285
Sum	1640141
Variance	16.04066
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=21)

Value	Count	Frequency (%)
2015	111	13.6%
2013	94	11.5%
2014	91	11.2%
2012	75	9.2%
2016	74	9.1%
2011	59	7.2%
2009	54	6.6%
2017	53	6.5%
2010	43	5.3%
2018	30	3.7%
Other values (11)	131	16.1%

Value	Count	Frequency (%)
1995	2	0.2%
2000	7	0.9%
2001	5	0.6%
2002	4	0.5%
2003	13	1.6%
2004	12	1.5%
2005	13	1.6%
2006	22	2.7%
2007	19	2.3%
2008	16	2.0%

Value	Count	Frequency (%)
2019	18	2.2%
2018	30	3.7%
2017	53	6.5%
2016	74	9.1%
2015	111	13.6%
2014	91	11.2%
2013	94	11.5%
2012	75	9.2%
2011	59	7.2%
2010	43	5.3%

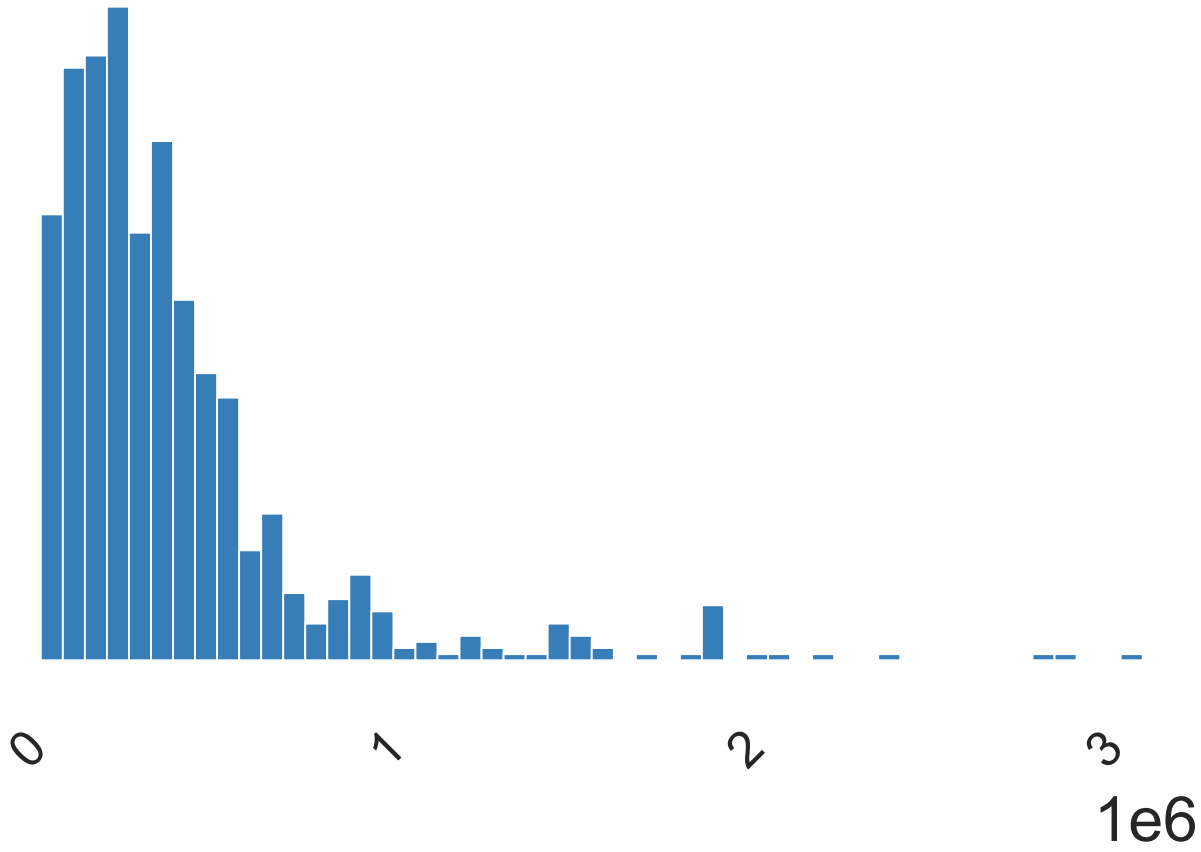
Price

Real number (ℝ)

HIGH CORRELATION (This variable has a high overall correlation with 2 fields: company, year)

Distinct	271
Distinct (%)	33.3%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	401793.34

Minimum	30000
Maximum	3100000
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	9.6 KiB



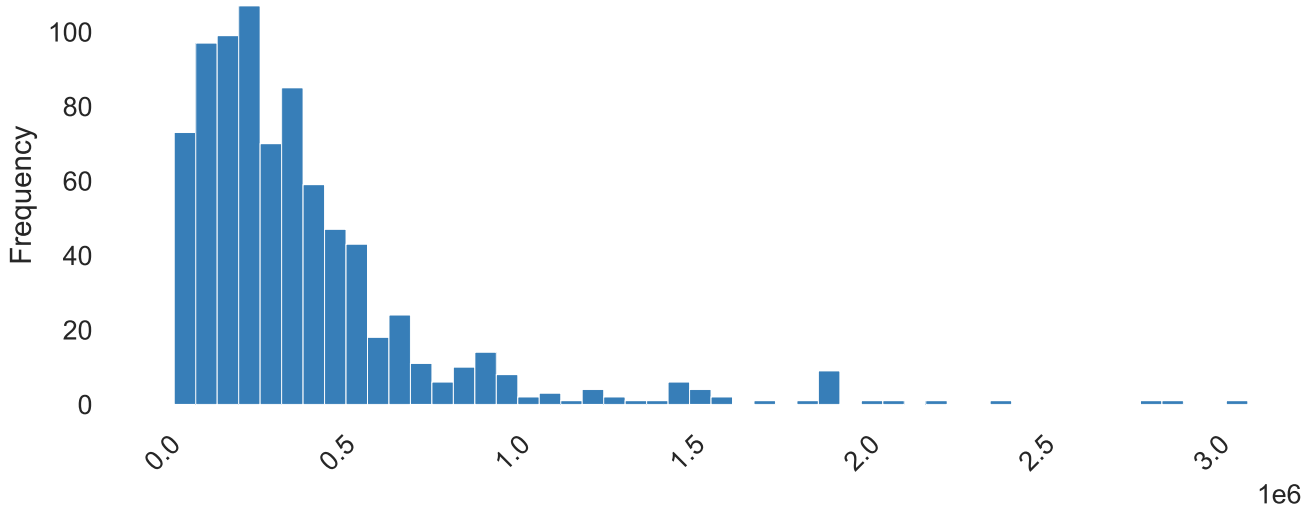
Quantile statistics

Minimum	30000
5-th percentile	74250
Q1	175000

median	299999
Q3	490000
95-th percentile	1074999.3
Maximum	3100000
Range	3070000
Interquartile range (IQR)	315000

Descriptive statistics

Standard deviation	381588.82
Coefficient of variation (CV)	0.94971415
Kurtosis	11.425123
Mean	401793.34
Median Absolute Deviation (MAD)	149999
Skewness	2.8921545
Sum	3.2746157×10^{8}
Variance	1.4561003×10^{11}
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
250000	16	2.0%
350000	12	1.5%
400000	11	1.3%
130000	11	1.3%
210000	10	1.2%
120000	10	1.2%
375000	10	1.2%
90000	10	1.2%
450000	10	1.2%
180000	10	1.2%
Other values (261)	705	86.5%

Value	Count	Frequency (%)
30000	2	0.2%
32000	2	0.2%
35000	1	0.1%
35999	1	0.1%
39999	1	0.1%
40000	3	0.4%
42000	1	0.1%
45000	2	0.2%
48000	1	0.1%
49000	1	0.1%

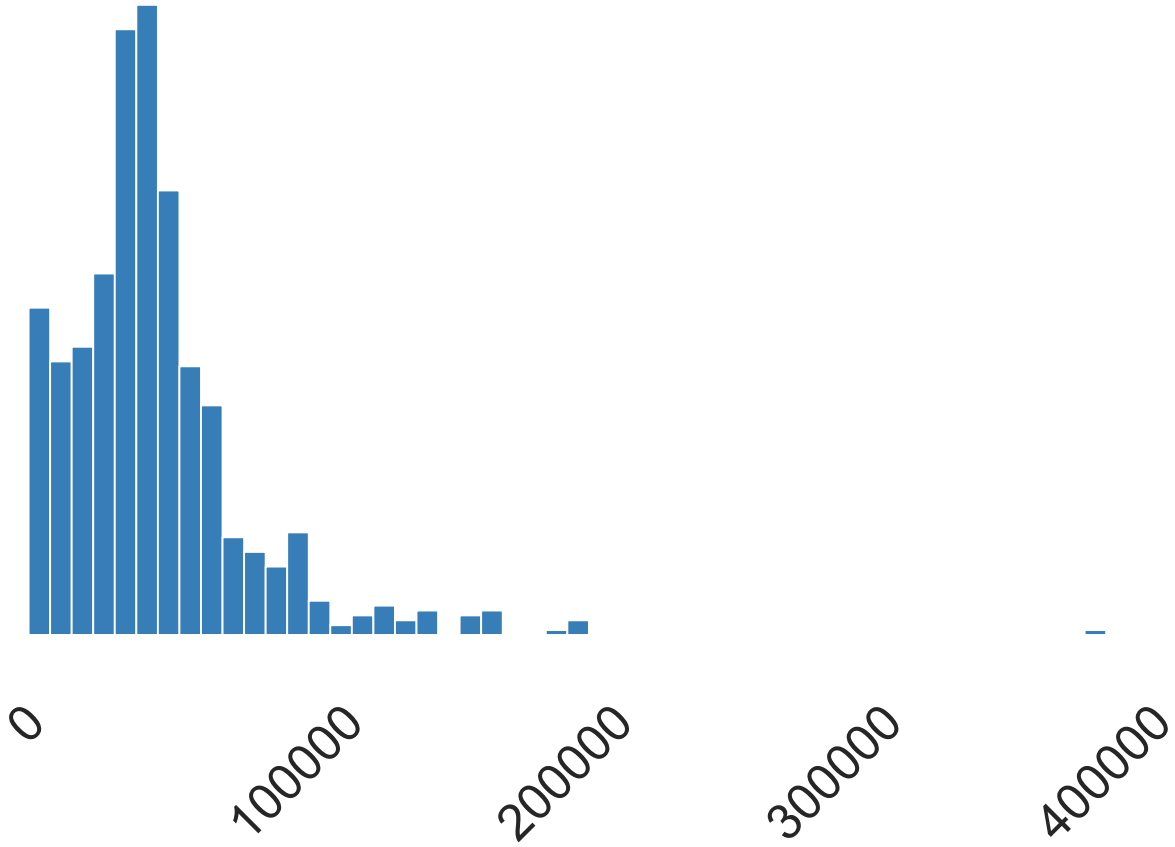
Value	Count	Frequency (%)
3100000	1	0.1%
2900000	1	0.1%
2800000	1	0.1%
2390000	1	0.1%
2190000	1	0.1%
2100000	1	0.1%
2000000	1	0.1%
1900000	1	0.1%
1891111	8	1.0%
1850000	1	0.1%

kms_driven

Real number (ℝ)

Distinct	246
Distinct (%)	30.2%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	46277.097

Minimum	0
Maximum	400000
Zeros	7
Zeros (%)	0.9%
Negative	0
Negative (%)	0.0%
Memory size	9.6 KiB



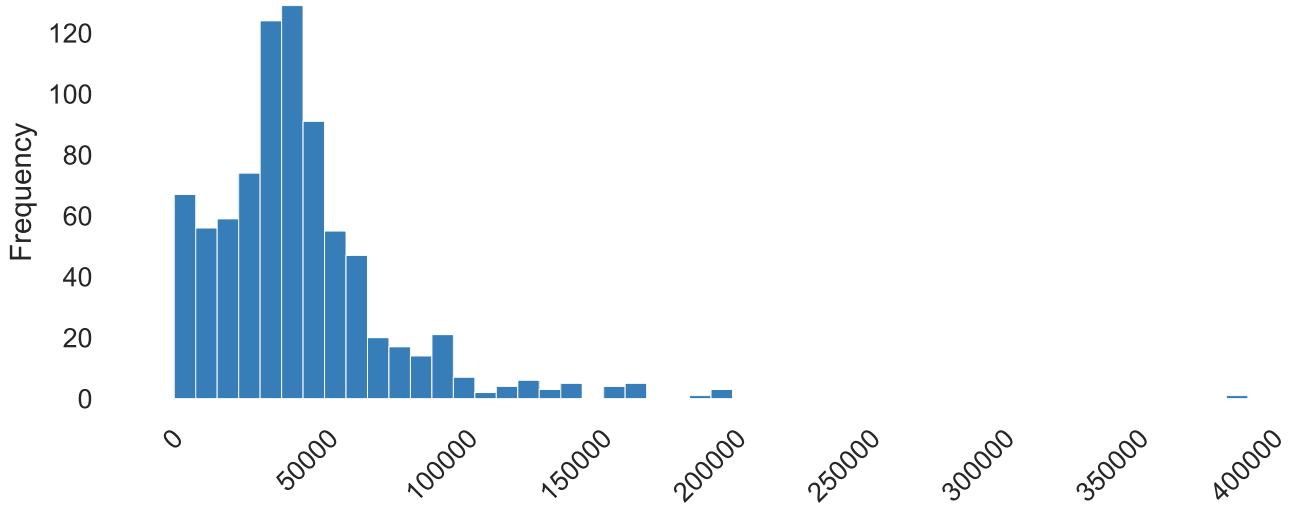
Quantile statistics

Minimum	0
5-th percentile	3455
Q1	27000
median	41000

Q3	56879
95-th percentile	103687.1
Maximum	400000
Range	400000
Interquartile range (IQR)	29879

Descriptive statistics

Standard deviation	34318.46
Coefficient of variation (CV)	0.74158627
Kurtosis	16.470638
Mean	46277.097
Median Absolute Deviation (MAD)	15400
Skewness	2.6532066
Sum	37715834
Variance	1.1777567×10^9
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
45000	29	3.6%
35000	29	3.6%
55000	25	3.1%
50000	23	2.8%
20000	21	2.6%
60000	20	2.5%
30000	19	2.3%
40000	18	2.2%
65000	16	2.0%
41000	16	2.0%
Other values (236)	599	73.5%

Value	Count	Frequency (%)
0	7	0.9%
40	2	0.2%
60	2	0.2%
65	1	0.1%
73	1	0.1%
100	1	0.1%
122	1	0.1%
300	1	0.1%
383	1	0.1%
588	2	0.2%

Value	Count	Frequency (%)
400000	1	0.1%
200000	3	0.4%
195000	1	0.1%
175430	3	0.4%
175400	1	0.1%
170000	1	0.1%
166000	1	0.1%
160000	3	0.4%
150000	3	0.4%
147000	1	0.1%

fuel_type
Categorical

Distinct	3
Distinct (%)	0.4%
Missing	0
Missing (%)	0.0%
Memory size	12.7 KiB

Length

Max length	6
Median length	6
Mean length	5.992638
Min length	3

Characters and Unicode

Total characters	4884	
Distinct characters	11	
Distinct categories	1 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)	?
Distinct scripts	1 (https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)	?
Distinct blocks	1 (https://en.wikipedia.org/wiki/Unicode_block)	?

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0	?
Unique (%)	0.0%	

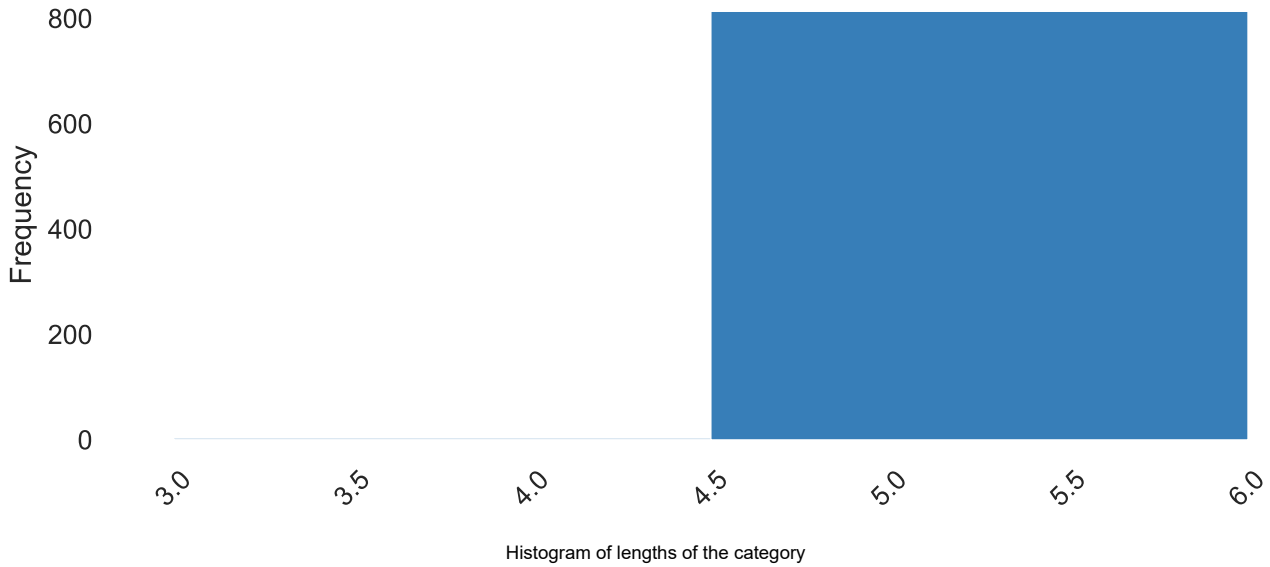
Sample

1st row	Petrol
2nd row	Diesel
3rd row	Petrol
4th row	Diesel
5th row	Diesel

Common Values

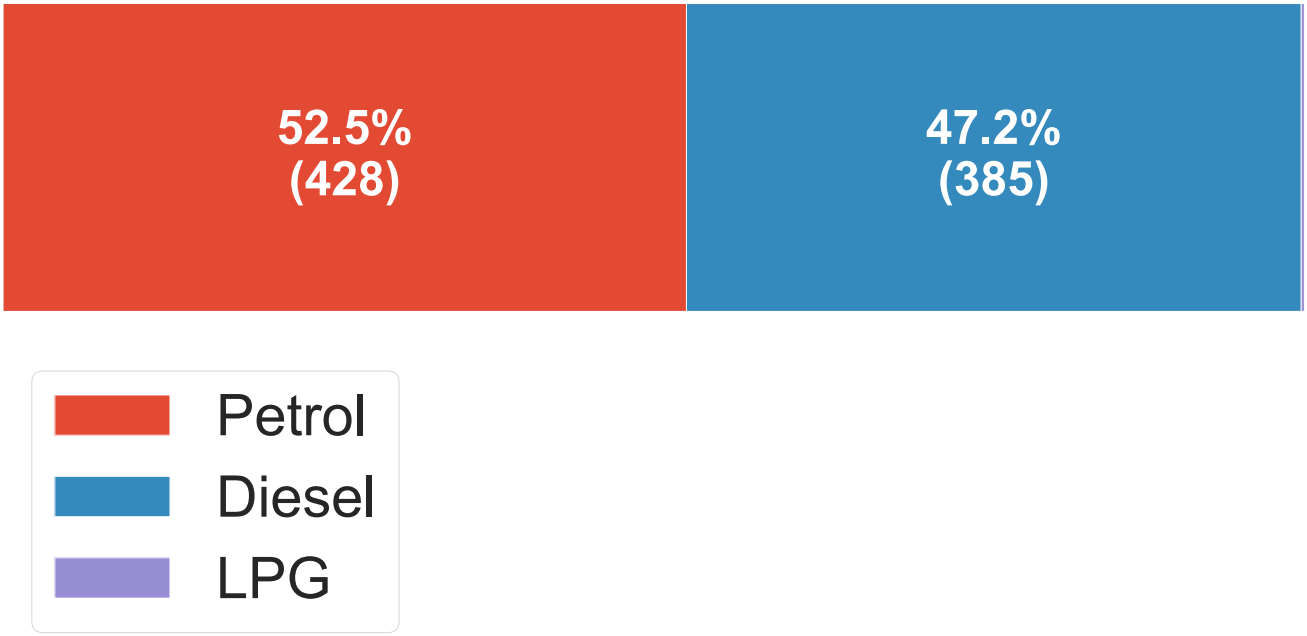
Value	Count	Frequency (%)
Petrol	428	52.5%
Diesel	385	47.2%
LPG	2	0.2%

Length



Histogram of lengths of the category

Common Values (Plot)



Value	Count	Frequency (%)
petrol	428	52.5%
diesel	385	47.2%
lpg	2	0.2%

Most occurring characters

Value	Count	Frequency (%)
e	1198	24.5%
l	813	16.6%
P	430	8.8%
t	428	8.8%
r	428	8.8%
o	428	8.8%
D	385	7.9%
i	385	7.9%
s	385	7.9%
L	2	< 0.1%

Most occurring categories

Value	Count	Frequency (%)
(unknown)	4884	100.0%

Most frequent character per category

(unknown)

Value	Count	Frequency (%)
e	1198	24.5%
l	813	16.6%
P	430	8.8%
t	428	8.8%
r	428	8.8%
o	428	8.8%
D	385	7.9%
i	385	7.9%
s	385	7.9%
L	2	< 0.1%

Most occurring scripts

Value	Count	Frequency (%)
(unknown)	4884	100.0%

Most frequent character per script

(unknown)

Value	Count	Frequency (%)
e	1198	24.5%
l	813	16.6%
P	430	8.8%
t	428	8.8%
r	428	8.8%
o	428	8.8%
D	385	7.9%
i	385	7.9%
s	385	7.9%
L	2	< 0.1%

Most occurring blocks

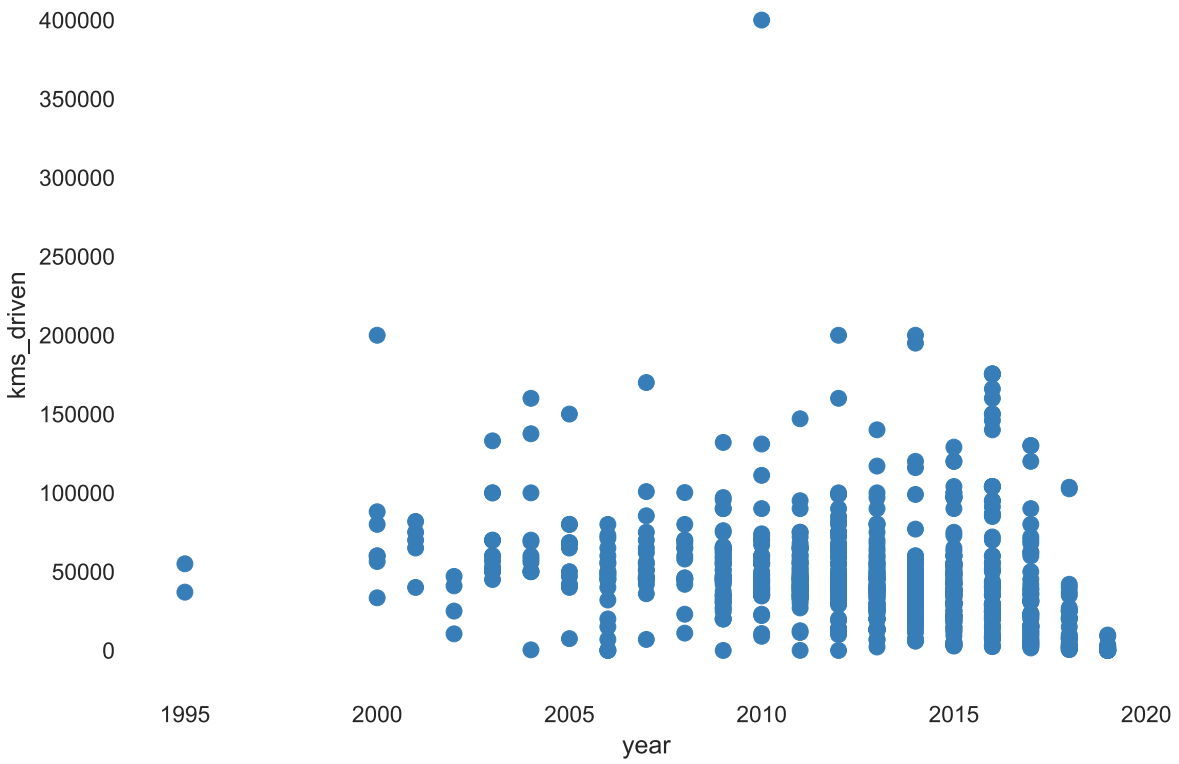
Value	Count	Frequency (%)
(unknown)	4884	100.0%

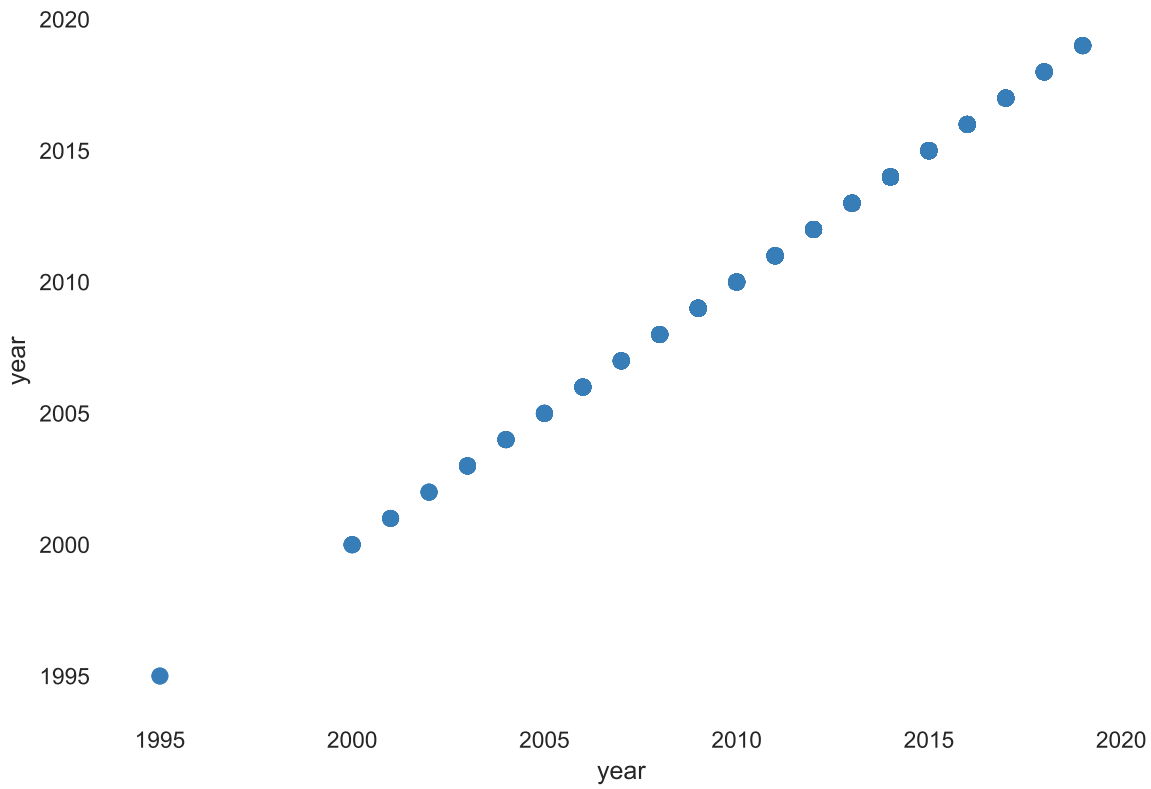
Most frequent character per block

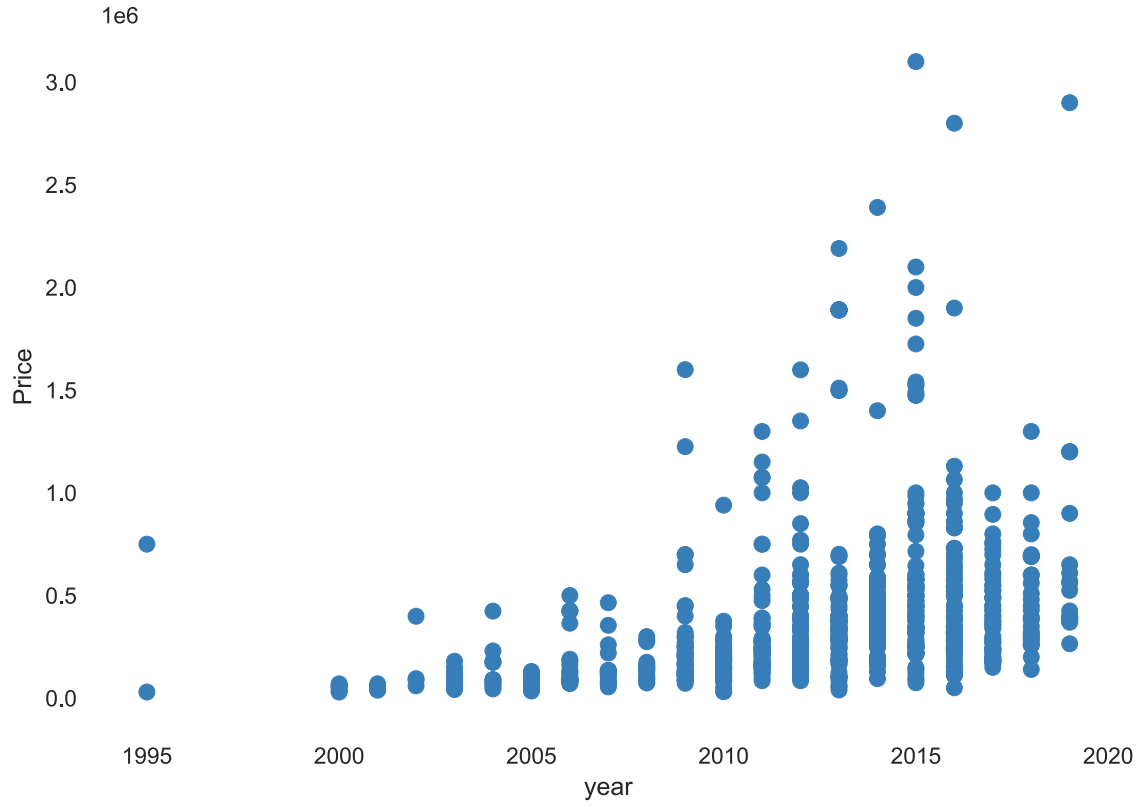
(unknown)

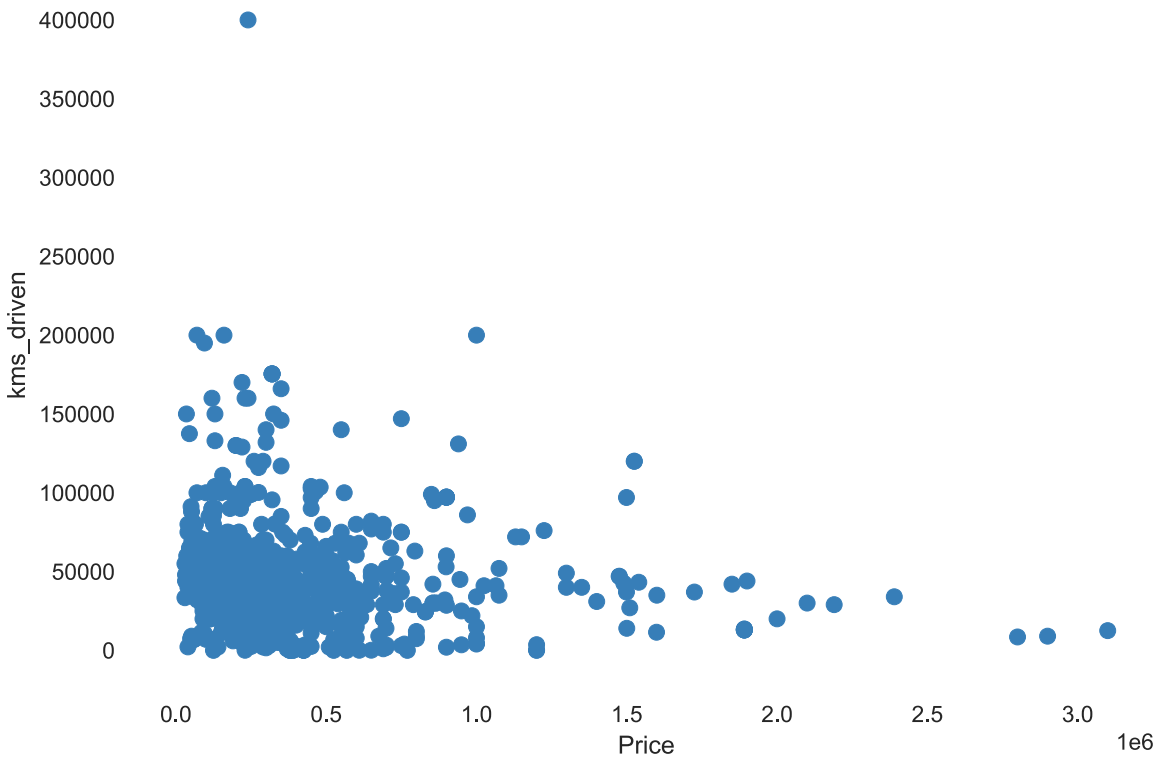
Value	Count	Frequency (%)
e	1198	24.5%
l	813	16.6%
P	430	8.8%
t	428	8.8%
r	428	8.8%
o	428	8.8%
D	385	7.9%
i	385	7.9%
s	385	7.9%
L	2	< 0.1%

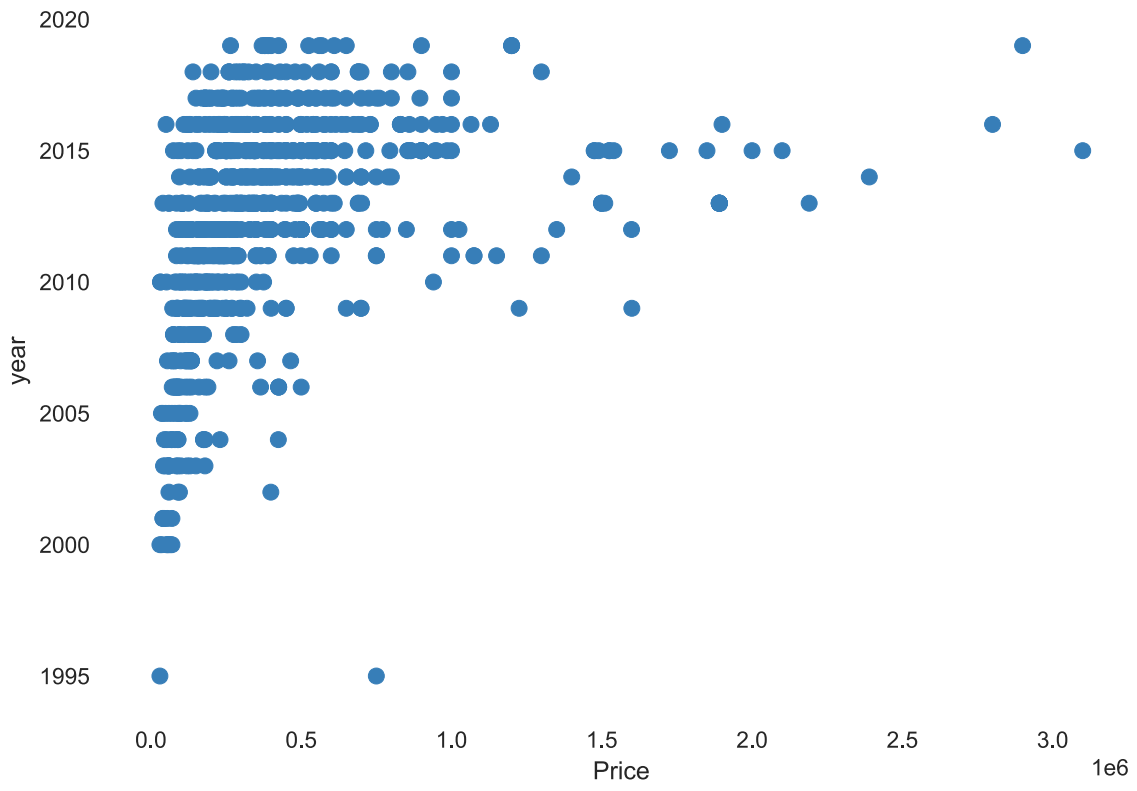
Interactions

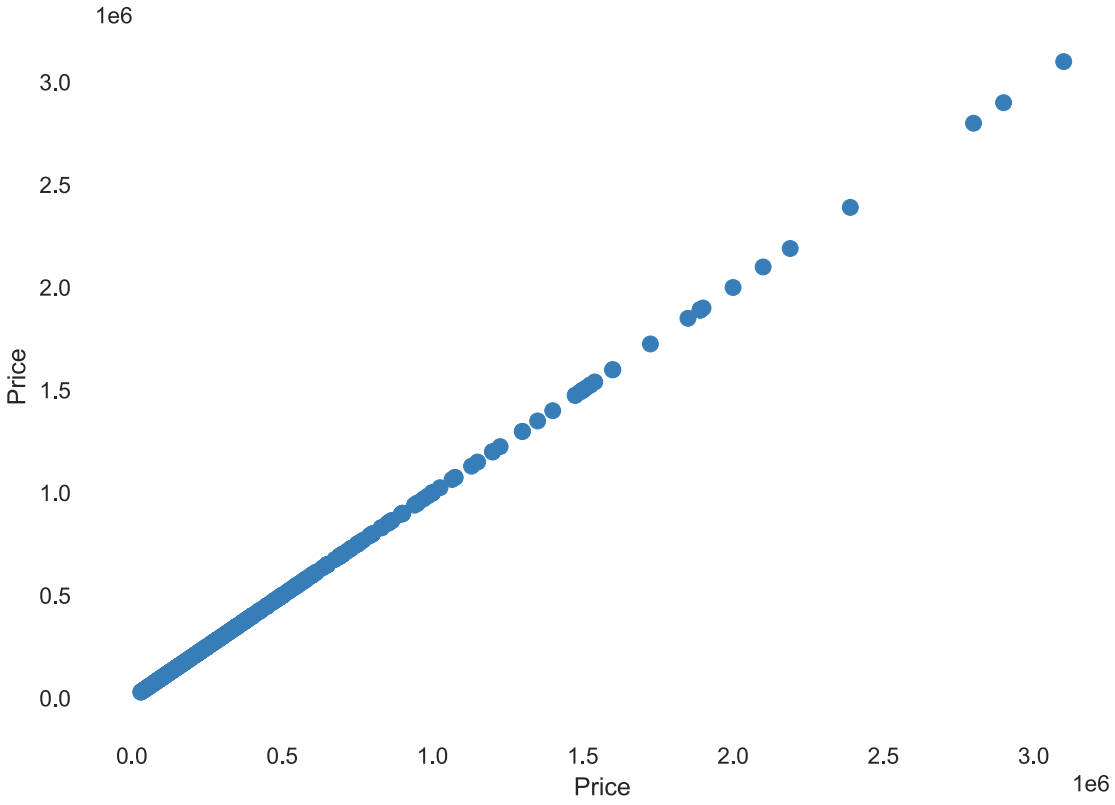


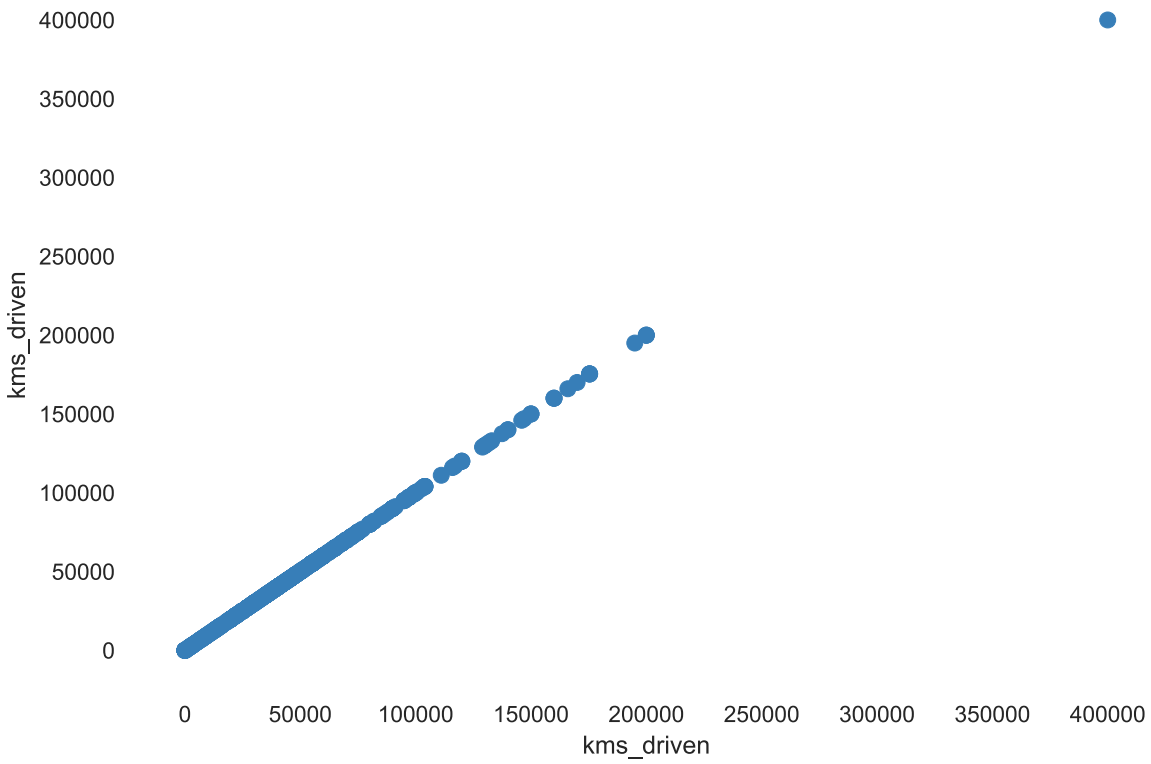

2020
2015
2010
2005
2000
1995
year
1995
2000
2005
2010
2015
2020
year

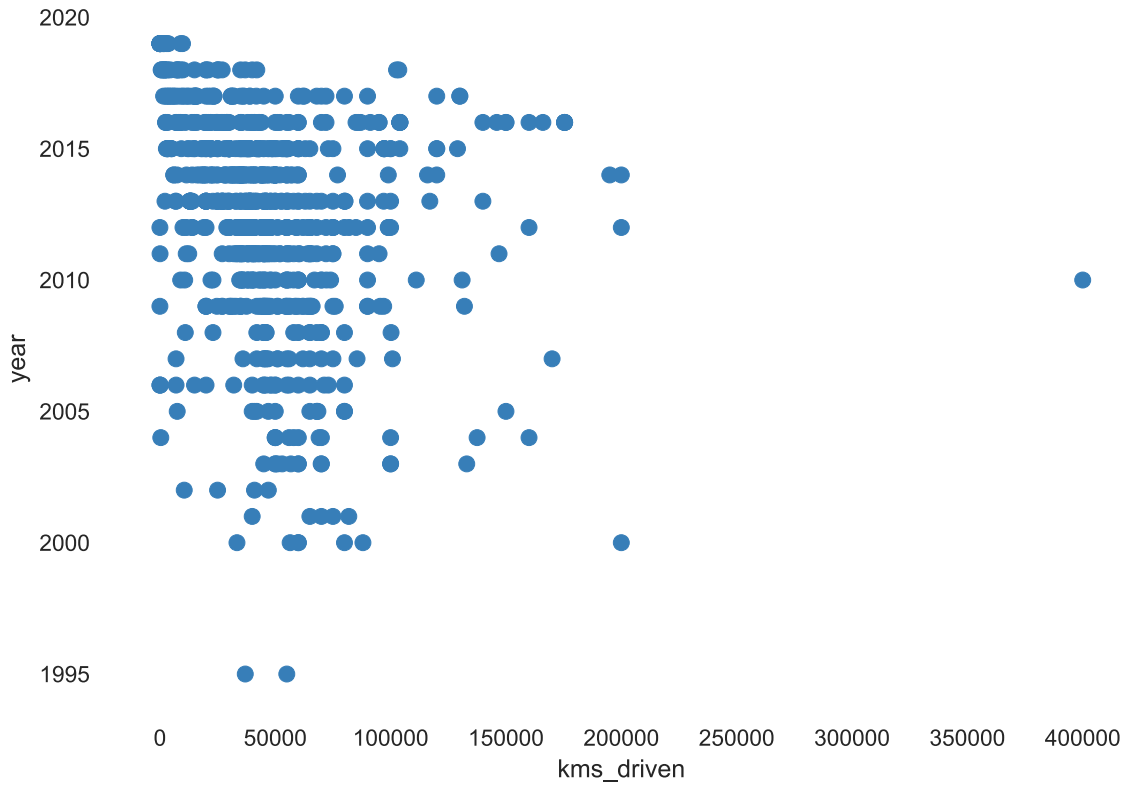

1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Price
1995
2000
2005
2010
2015
2020
year

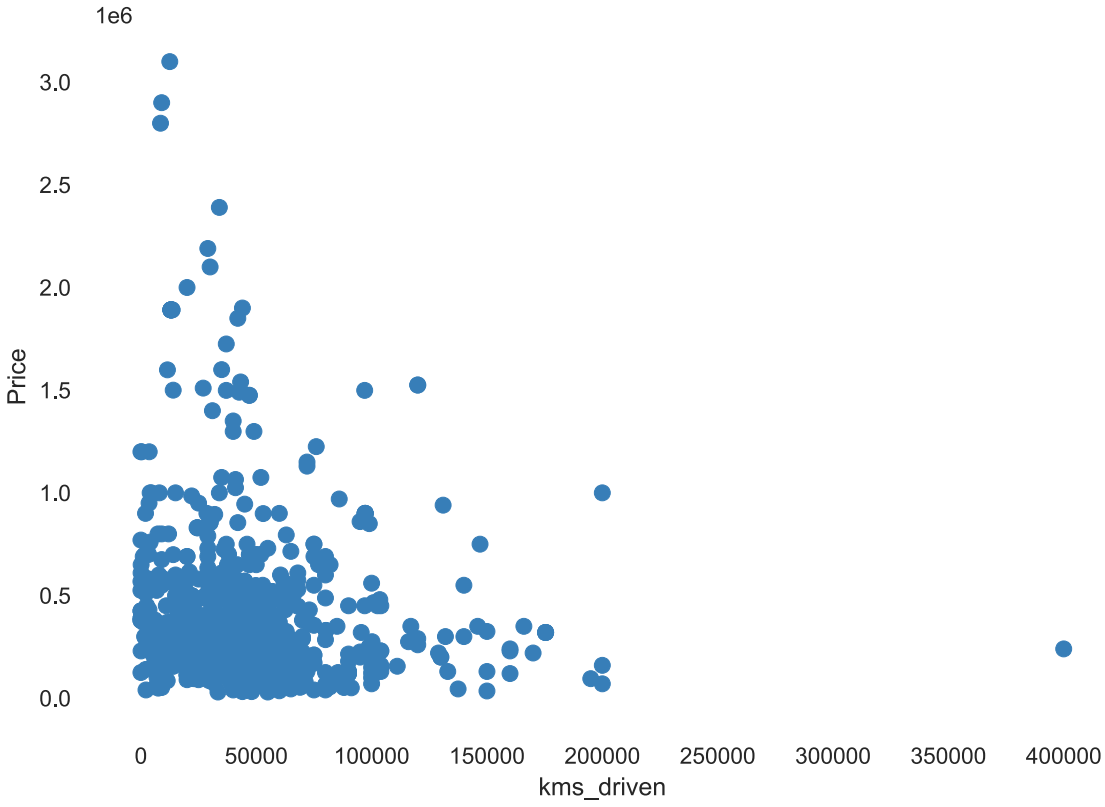

400000
350000
300000
250000
200000
150000
100000
50000
0
kms_driven
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1e6
Price


2020
2015
2010
2005
2000
1995
year
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1e6
Price

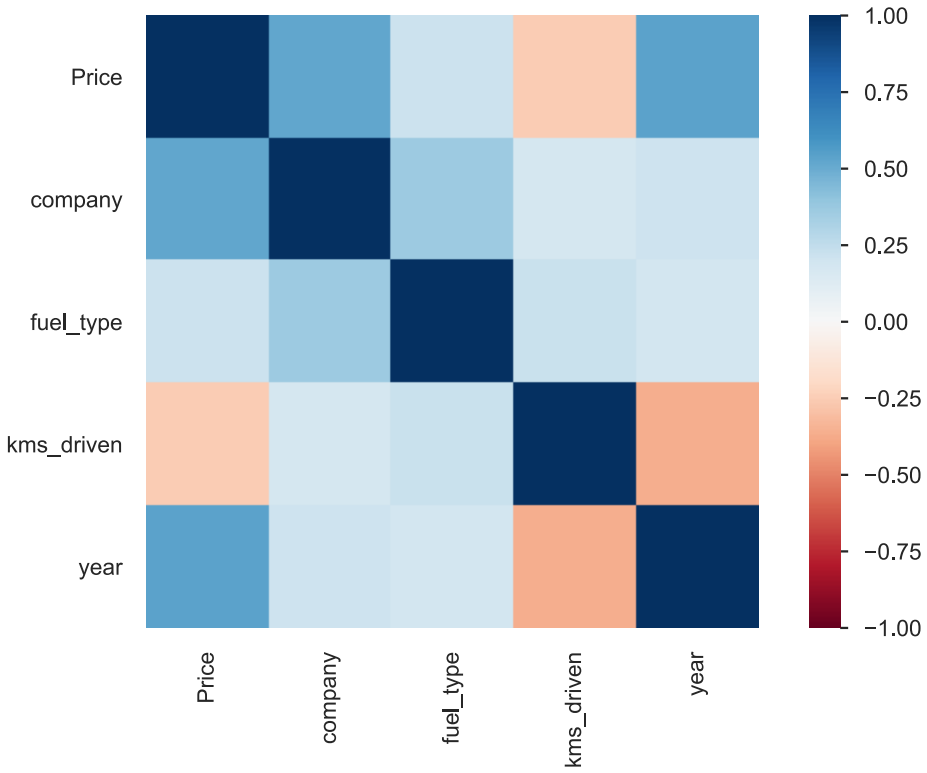

1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Price
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1e6
Price


400000
350000
300000
250000
200000
150000
100000
50000
0
kms_driven
0
50000
100000
150000
200000
250000
300000
350000
400000
kms_driven


2020
2015
2010
2005
2000
1995
year
0
50000
100000
150000
200000
250000
300000
350000
400000
kms_driven

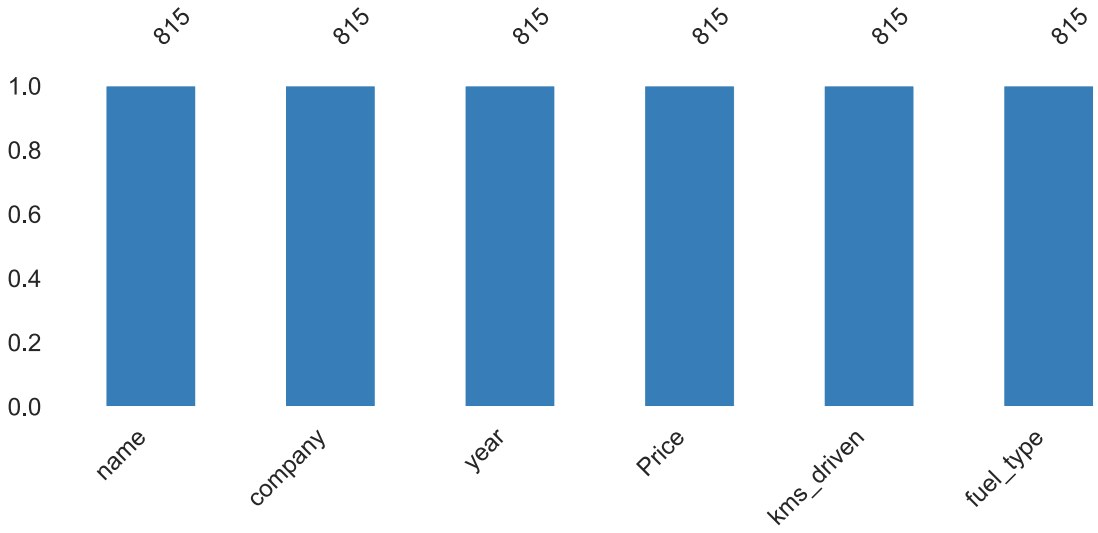

1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Price
0
50000
100000
150000
200000
250000
300000
350000
400000
kms_driven

Correlations

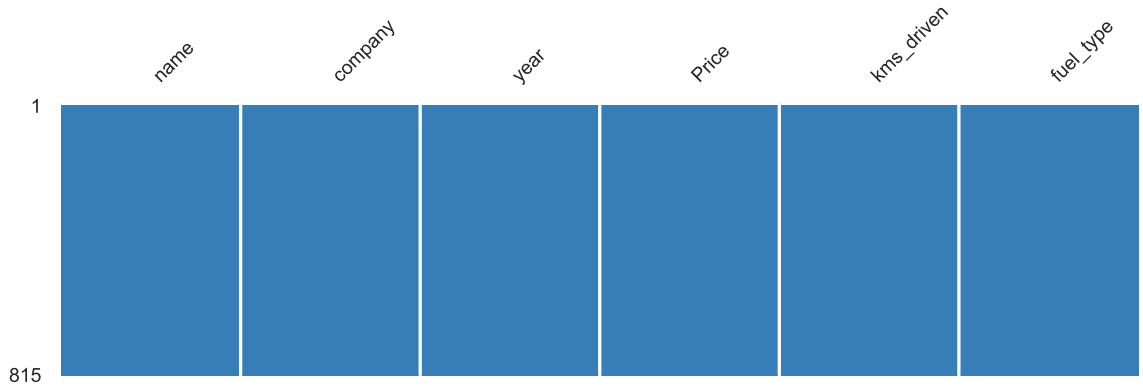


	Price	company	fuel_type	kms_driven	year
Price	1.000	0.521	0.213	-0.253	0.539
company	0.521	1.000	0.364	0.175	0.204
fuel_type	0.213	0.364	1.000	0.220	0.183
kms_driven	-0.253	0.175	0.220	1.000	-0.363
year	0.539	0.204	0.183	-0.363	1.000

Missing values



A simple visualization of nullity by column.



Nullity matrix is a data-dense display which lets you quickly visually pick out patterns in data completion.

Sample

	name	company	year	Price	kms_driven	fuel_type
0	Hyundai Santro Xing	Hyundai	2007	80000	45000	Petrol
1	Mahindra Jeep CL550	Mahindra	2006	425000	40	Diesel
2	Hyundai Grand i10	Hyundai	2014	325000	28000	Petrol
3	Ford EcoSport Titanium	Ford	2014	575000	36000	Diesel
4	Ford Figo	Ford	2012	175000	41000	Diesel
5	Hyundai Eon	Hyundai	2013	190000	25000	Petrol
6	Ford EcoSport Ambiente	Ford	2016	830000	24530	Diesel
7	Maruti Suzuki Alto	Maruti	2015	250000	60000	Petrol
8	Skoda Fabia Classic	Skoda	2010	182000	60000	Petrol
9	Maruti Suzuki Stingray	Maruti	2015	315000	30000	Petrol

	name	company	year	Price	kms_driven	fuel_type
806	Hyundai Getz	Hyundai	2006	125000	80000	Petrol
807	Mercedes Benz C	Mercedes	2006	500001	15000	Petrol
808	Maruti Suzuki Alto	Maruti	2005	95000	65000	Petrol
809	Maruti Suzuki Swift	Maruti	2009	250000	51000	Diesel
810	Skoda Fabia	Skoda	2009	110000	45000	Petrol
811	Maruti Suzuki Ritz	Maruti	2011	270000	50000	Petrol
812	Tata Indica V2	Tata	2009	110000	30000	Diesel
813	Toyota Corolla Altis	Toyota	2009	300000	132000	Petrol
814	Tata Zest XM	Tata	2018	260000	27000	Diesel
815	Mahindra Quanto C8	Mahindra	2013	390000	40000	Diesel

Duplicate rows

Most frequently occurring

	name	company	year	Price	kms_driven	fuel_type	# duplicates
58	Mini Cooper S	Mini	2013	1891111	13000	Petrol	6
32	Mahindra Scorpio S10	Mahindra	2015	900000	97200	Diesel	5
9	Honda Amaze	Honda	2013	284999	46000	Diesel	4
10	Honda Amaze	Honda	2015	344999	22000	Petrol	4
57	Maruti Suzuki Wagon	Maruti	2013	105000	39000	Petrol	4
4	Force Motors Force	Force	2015	580000	3200	Diesel	3
22	Hyundai i20 Sportz	Hyundai	2012	100000	55000	Petrol	3
30	Mahindra Quanto C8	Mahindra	2013	375000	20000	Diesel	3
52	Maruti Suzuki Swift	Maruti	2012	280000	48006	Diesel	3
63	Renault Duster 110PS	Renault	2012	501000	35000	Diesel	3