

Basic Analysis using numpy and pandas

Vehicle Dataset

To import library

In [1]:

```
import numpy as np
```

In [2]:

```
import pandas as pd
```

To import dataset

In [3]:

```
d=pd.read_csv(r"C:\Users\user\Downloads\vc.csv")
d
```

Out[3]:

	ID	model	engine_power	age_in_days	km	previous_owners	lat	
0	1.0	lounge	51.0	882.0	25000.0	1.0	44.907242	8.6115
1	2.0	pop	51.0	1186.0	32500.0	1.0	45.666359	12.241
2	3.0	sport	74.0	4658.0	142228.0	1.0	45.503300	11
3	4.0	lounge	51.0	2739.0	160000.0	1.0	40.633171	17.634
4	5.0	pop	73.0	3074.0	106880.0	1.0	41.903221	12.495
...	
1544	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1545	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1546	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Null
1547	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1548	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

1549 rows × 11 columns

To get Top 10 record

In [4]:

```
d.head(10)
```

Out[4]:

	ID	model	engine_power	age_in_days	km	previous_owners	lat	lon
0	1.0	lounge	51.0	882.0	25000.0	1.0	44.907242	8.6115598
1	2.0	pop	51.0	1186.0	32500.0	1.0	45.666359	12.241889
2	3.0	sport	74.0	4658.0	142228.0	1.0	45.503300	11.417
3	4.0	lounge	51.0	2739.0	160000.0	1.0	40.633171	17.634609
4	5.0	pop	73.0	3074.0	106880.0	1.0	41.903221	12.495650
5	6.0	pop	74.0	3623.0	70225.0	1.0	45.000702	7.682270
6	7.0	lounge	51.0	731.0	11600.0	1.0	44.907242	8.6115598
7	8.0	lounge	51.0	1521.0	49076.0	1.0	41.903221	12.495650
8	9.0	sport	73.0	4049.0	76000.0	1.0	45.548000	11.549469
9	10.0	sport	51.0	3653.0	89000.0	1.0	45.438301	10.991700

To get last record

In [5]:

```
d.tail(20)
```

Out[5]:

	ID	model	engine_power	age_in_days	km	previous_owners	lat	
1529	1530.0	lounge	51.0	731.0	22551.0	1.0	38.122070	13.3
1530	1531.0	lounge	51.0	670.0	29000.0	1.0	45.764648	8.9
1531	1532.0	sport	73.0	4505.0	127000.0	1.0	45.528511	9.59
1532	1533.0	pop	51.0	1917.0	52008.0	1.0	45.548000	11.5
1533	1534.0	sport	51.0	3712.0	115280.0	1.0	45.069679	7.70
1534	1535.0	lounge	74.0	3835.0	112000.0	1.0	45.845692	8.66
1535	1536.0	pop	51.0	2223.0	60457.0	1.0	45.481541	9.41
1536	1537.0	lounge	51.0	2557.0	80750.0	1.0	45.000702	7.6
1537	1538.0	pop	51.0	1766.0	54276.0	1.0	40.323410	17.5
1538	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1539	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1540	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1541	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1542	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1543	NaN	NaN	NaN	NaN	NaN	NaN	NaN	cc
1544	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1545	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1546	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
1547	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1548	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

Statistics Analysis

In [6]:

```
d.describe()
```

Out[6]:

	ID	engine_power	age_in_days	km	previous_owners	la
count	1538.000000	1538.000000	1538.000000	1538.000000	1538.000000	1538.000000
mean	769.500000	51.904421	1650.980494	53396.011704	1.123537	43.54136
std	444.126671	3.988023	1289.522278	40046.830723	0.416423	2.13351
min	1.000000	51.000000	366.000000	1232.000000	1.000000	36.85583
25%	385.250000	51.000000	670.000000	20006.250000	1.000000	41.80299
50%	769.500000	51.000000	1035.000000	39031.000000	1.000000	44.39409
75%	1153.750000	51.000000	2616.000000	79667.750000	1.000000	45.46796
max	1538.000000	77.000000	4658.000000	235000.000000	4.000000	46.79561

To get row and column

In [7]:

```
np.shape(d)
```

Out[7]:

(1549, 11)

Find Number of Elements

In [8]:

```
np.size(d)
```

Out[8]:

17039

Find Missing Value

In [9]:

```
d.isna()
```

Out[9]:

	ID	model	engine_power	age_in_days	km	previous_owners	lat	lon	price
0	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False
...
1544	True	True	True	True	True	True	True	False	False
1545	True	True	True	True	True	True	True	False	False
1546	True	True	True	True	True	True	True	False	False
1547	True	True	True	True	True	True	True	False	False
1548	True	True	True	True	True	True	True	False	False

1549 rows × 11 columns



To drop the missing value

In [10]:

```
d.dropna(axis=1,how="any")
```

Out[10]:

	lon	price
0	8.611559868	8900
1	12.24188995	8800
2	11.41784	4200
3	17.63460922	6000
4	12.49565029	5700
...
1544	length	5
1545	concat	lonprice
1546	Null values	NO
1547	find	1
1548	search	1

1549 rows × 2 columns

In [11]:

```
d["model"]
```

Out[11]:

```
0      lounge
1        pop
2        sport
3      lounge
4        pop
...
1544      NaN
1545      NaN
1546      NaN
1547      NaN
1548      NaN
Name: model, Length: 1549, dtype: object
```

In [12]:

```
data=d[['km','price']]
data
```

Out[12]:

	km	price
0	25000.0	8900
1	32500.0	8800
2	142228.0	4200
3	160000.0	6000
4	106880.0	5700
...
1544	NaN	5
1545	NaN	lonprice
1546	NaN	NO
1547	NaN	1
1548	NaN	1

1549 rows × 2 columns

In [13]:

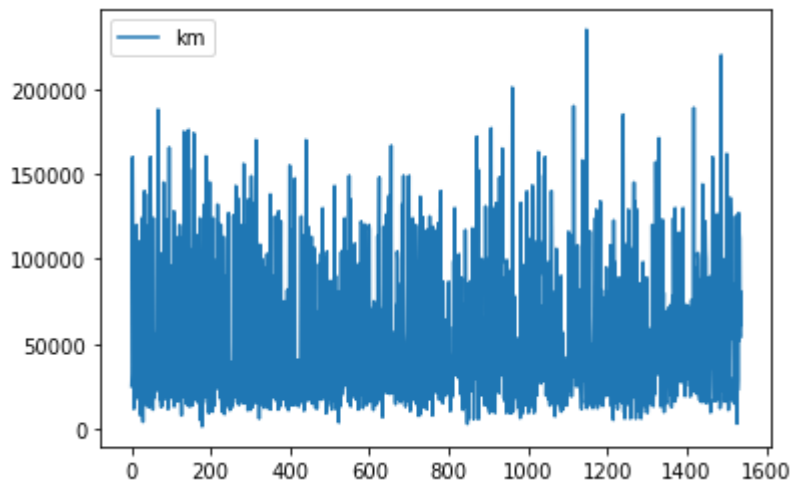
```
import matplotlib.pyplot as pp
```

In [14]:

```
data.plot.line()
```

Out[14]:

<AxesSubplot:>

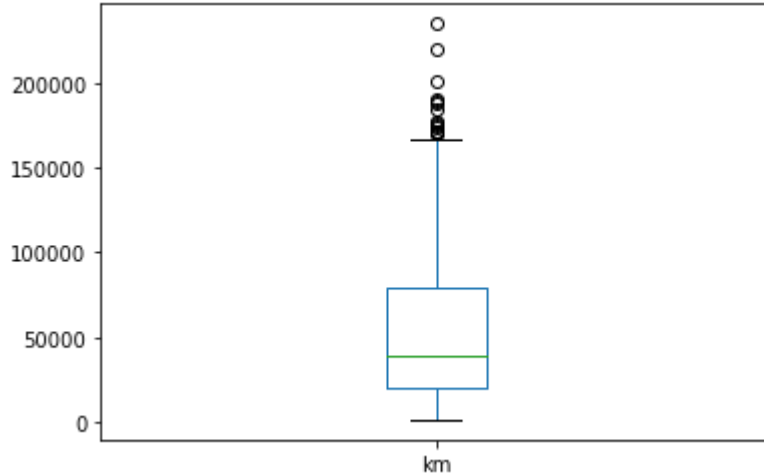


In [15]:

```
data.plot.box()
```

Out[15]:

<AxesSubplot:>

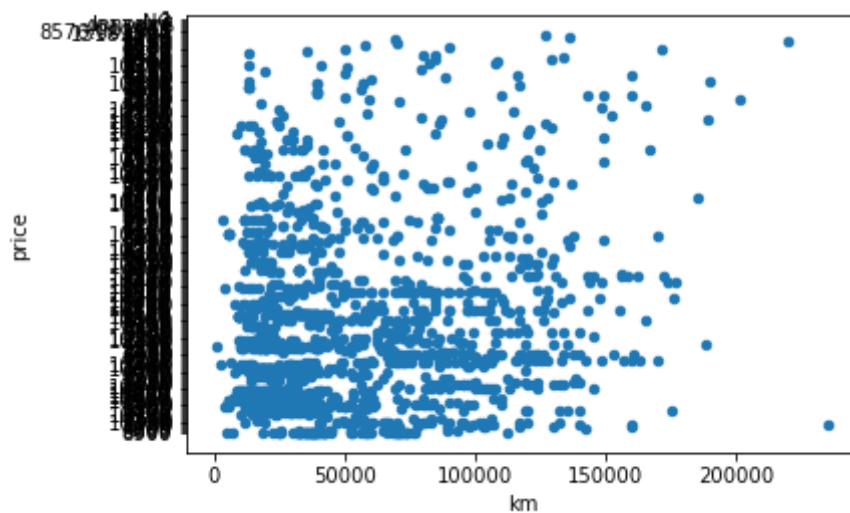


In [16]:

```
data.plot.scatter(x='km',y='price')
```

Out[16]:

<AxesSubplot:xlabel='km', ylabel='price'>

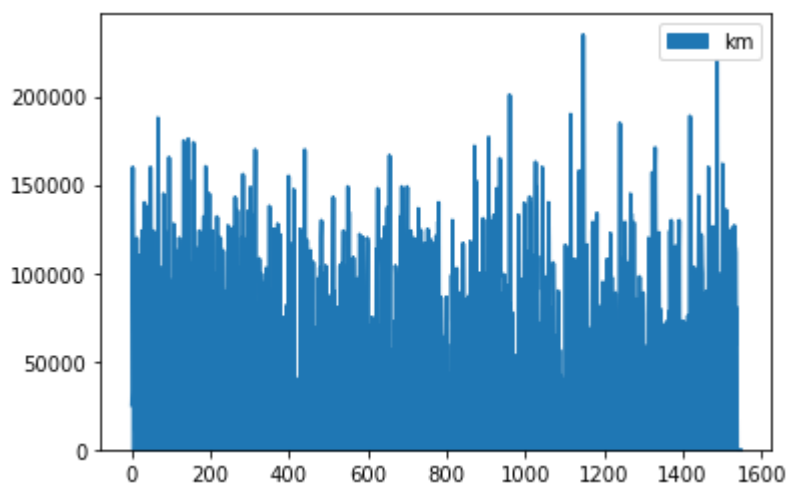


In [17]:

```
data.plot.area()
```

Out[17]:

<AxesSubplot:>

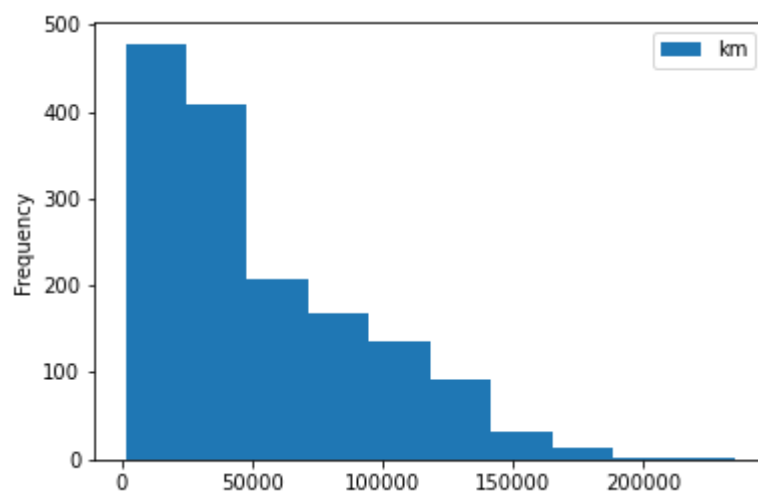


In [18]:

```
data.plot.hist()
```

Out[18]:

<AxesSubplot:ylabel='Frequency'>



In []: