In [ ]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import sklearn as sklearn
df = pd.read_csv("autodata.csv")
df.head(5)
```

Out[ ]:

	Unnamed: 0	symboling	normalized-losses	make	aspiration	num-of-doors	body-style	drive-wheels
0	0	3	122	alfa-romero	std	two	convertible	rwd
1	1	3	122	alfa-romero	std	two	convertible	rwd
2	2	1	122	alfa-romero	std	two	hatchback	rwd
3	3	2	164	audi	std	four	sedan	fwd
4	4	2	164	audi	std	four	sedan	4wd

5 rows × 30 columns

In [ ]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 201 entries, 0 to 200
Data columns (total 30 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Unnamed: 0         201 non-null    int64  
 1   symboling          201 non-null    int64  
 2   normalized-losses  201 non-null    int64  
 3   make               201 non-null    object 
 4   aspiration         201 non-null    object 
 5   num-of-doors       201 non-null    object 
 6   body-style         201 non-null    object 
 7   drive-wheels       201 non-null    object 
 8   engine-location    201 non-null    object 
 9   wheel-base         201 non-null    float64
 10  length             201 non-null    float64
 11  width              201 non-null    float64
 12  height             201 non-null    float64
 13  curb-weight        201 non-null    int64  
 14  engine-type        201 non-null    object 
 15  num-of-cylinders   201 non-null    object 
 16  engine-size        201 non-null    int64  
 17  fuel-system        201 non-null    object 
 18  bore               201 non-null    float64
 19  stroke             197 non-null    float64
 20  compression-ratio  201 non-null    float64
 21  horsepower         199 non-null    float64
 22  peak-rpm           199 non-null    float64
 23  city-mpg           201 non-null    int64  
 24  highway-mpg        201 non-null    int64  
 25  price              201 non-null    float64
 26  city-L/100km       201 non-null    float64
 27  horsepower-binned  199 non-null    object 
 28  diesel             201 non-null    int64  
 29  gas                201 non-null    int64  
dtypes: float64(11), int64(9), object(10)
memory usage: 47.2+ KB
```

In [ ]:
```
print("\nData Description:\n")
df.describe()
```

```
Data Description:
```

Out[ ]:

	Unnamed: 0	symboling	normalized-losses	wheel-base	length	width	he
count	201.000000	201.000000	201.00000	201.000000	201.000000	201.000000	201.00
mean	100.000000	0.840796	122.00000	98.797015	0.837102	0.915126	53.76
std	58.167861	1.254802	31.99625	6.066366	0.059213	0.029187	2.44
min	0.000000	-2.000000	65.00000	86.600000	0.678039	0.837500	47.80
25%	50.000000	0.000000	101.00000	94.500000	0.801538	0.890278	52.00
50%	100.000000	1.000000	122.00000	97.000000	0.832292	0.909722	54.10
75%	150.000000	2.000000	137.00000	102.400000	0.881788	0.925000	55.50
max	200.000000	3.000000	256.00000	120.900000	1.000000	1.000000	59.80

In [ ]:
```
print("\n Variable Description:\n")
df.dtypes
```

```
Variable Description:
```

Out[ ]:
```
Unnamed: 0             int64
symboling              int64
normalized-losses      int64
make                  object
aspiration            object
num-of-doors          object
body-style            object
drive-wheels          object
engine-location       object
wheel-base           float64
length               float64
width                float64
height               float64
curb-weight            int64
engine-type           object
num-of-cylinders      object
engine-size            int64
fuel-system           object
bore                 float64
stroke               float64
compression-ratio    float64
horsepower           float64
peak-rpm             float64
city-mpg               int64
highway-mpg            int64
price                float64
city-L/100km         float64
horsepower-binned     object
diesel                 int64
gas                    int64
dtype: object
```

Removing Null Values

In [ ]:
```
print("Data Preprocessing - Missing Values : \n")
df.isnull()
```

```
Data Preprocessing - Missing Values :
```

Out[ ]:

	Unnamed: 0	symboling	normalized-losses	make	aspiration	num-of-doors	body-style	drive-wheels	eng loca
0	False	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	False	
...	...	...	...	...	...	...	...	...	
196	False	False	False	False	False	False	False	False	
197	False	False	False	False	False	False	False	False	
198	False	False	False	False	False	False	False	False	
199	False	False	False	False	False	False	False	False	
200	False	False	False	False	False	False	False	False	

201 rows × 30 columns

In [ ]:
```
print("Data Preprocessing - Sum of missing values:\n")
df.isnull().sum()
```

```
Data Preprocessing - Sum of missing values:
```

Out[ ]:
```
Unnamed: 0           0
symboling            0
normalized-losses    0
make                 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
num-of-cylinders     0
engine-size          0
fuel-system          0
bore                 0
stroke               4
compression-ratio    0
horsepower           2
peak-rpm             2
city-mpg             0
highway-mpg          0
price                0
city-L/100km         0
horsepower-binned    2
diesel               0
gas                  0
dtype: int64
```

In [ ]:
```
print("using Not Null:\n")
df.notnull()
```

```
using Not Null:
```

Out[ ]:

	Unnamed: 0	symboling	normalized-losses	make	aspiration	num-of-doors	body-style	drive-wheels	eng loca
0	True	True	True	True	True	True	True	True	
1	True	True	True	True	True	True	True	True	
2	True	True	True	True	True	True	True	True	
3	True	True	True	True	True	True	True	True	
4	True	True	True	True	True	True	True	True	
...	...	...	...	...	...	...	...	...	
196	True	True	True	True	True	True	True	True	
197	True	True	True	True	True	True	True	True	
198	True	True	True	True	True	True	True	True	
199	True	True	True	True	True	True	True	True	
200	True	True	True	True	True	True	True	True	

201 rows × 30 columns

In [ ]:
```
df['stroke']=df['stroke'].fillna(df['stroke'].mean())
print("Data Processing - sum of missing values after removing null values of str
df.isnull().sum()
```

```
Data Processing - sum of missing values after removing null values of stroke:
```

Out[ ]:
```
Unnamed: 0           0
symboling            0
normalized-losses    0
make                 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
num-of-cylinders     0
engine-size          0
fuel-system          0
bore                 0
stroke               0
compression-ratio    0
horsepower           2
peak-rpm             2
city-mpg             0
highway-mpg          0
price                0
city-L/100km         0
horsepower-binned    2
diesel               0
gas                  0
dtype: int64
```

In [ ]:
```
df=df.drop(columns=['horsepower'])
df.isnull().sum()
```

Out[ ]:
```
Unnamed: 0           0
symboling            0
normalized-losses    0
make                 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
num-of-cylinders     0
engine-size          0
fuel-system          0
bore                 0
stroke               0
compression-ratio    0
peak-rpm             2
city-mpg             0
highway-mpg          0
price                0
city-L/100km         0
horsepower-binned    2
diesel               0
gas                  0
dtype: int64
```

Variable Types

In [ ]:
```
df.dtypes
```

Out[ ]:
```
Unnamed: 0             int64
symboling              int64
normalized-losses      int64
make                  object
aspiration            object
num-of-doors          object
body-style            object
drive-wheels          object
engine-location       object
wheel-base           float64
length               float64
width                float64
height               float64
curb-weight            int64
engine-type           object
num-of-cylinders      object
engine-size            int64
fuel-system           object
bore                 float64
stroke               float64
compression-ratio    float64
peak-rpm             float64
city-mpg               int64
highway-mpg            int64
price                float64
city-L/100km         float64
horsepower-binned     object
diesel                 int64
gas                    int64
dtype: object
```

In [ ]:
```
df['normalized-losses']=df['normalized-losses'].astype(bool)
df.sample(5)
```

Out[ ]:

	Unnamed: 0	symboling	normalized-losses	make	aspiration	num-of-doors	body-style	drive-wheels
131	131	2	True	saab	std	four	sedan	fwd
135	135	2	True	subaru	std	two	hatchback	fwd
24	24	1	True	dodge	std	four	sedan	fwd
110	110	0	True	peugot	std	four	wagon	rwd
139	139	0	True	subaru	std	four	sedan	fwd

5 rows × 29 columns

In [ ]:
```
df['wheel-base']=df['wheel-base'].astype(int)
df.sample(5)
```

Out[ ]:

	Unnamed: 0	symboling	normalized-losses	make	aspiration	num-of-doors	body-style	drive-wheels	en loc
8	8	1	True	audi	turbo	four	sedan	fwd	
158	158	0	True	toyota	std	four	sedan	fwd	
38	38	0	True	honda	std	four	sedan	fwd	
90	90	1	True	nissan	std	four	wagon	fwd	
94	94	1	True	nissan	std	four	wagon	fwd	

5 rows × 29 columns