```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

```
train=pd.read_csv('/content/train-data.csv')
train.head()
```

	Unnamed: 0	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Ow
0	0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	
1	1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	
2	2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	

```
train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6019 entries, 0 to 6018
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Unnamed: 0         6019 non-null   int64  
 1   Name               6019 non-null   object 
 2   Location           6019 non-null   object 
 3   Year               6019 non-null   int64  
 4   Kilometers_Driven  6019 non-null   int64  
 5   Fuel_Type          6019 non-null   object 
 6   Transmission       6019 non-null   object 
 7   Owner_Type         6019 non-null   object 
 8   Mileage            6017 non-null   object 
 9   Engine             5983 non-null   object 
 10  Power              5983 non-null   object 
 11  Seats              5977 non-null   float64
 12  New_Price          824 non-null    object 
 13  Price              6019 non-null   float64
dtypes: float64(2), int64(3), object(9)
memory usage: 658.5+ KB
```

```
train.isna().sum()
```

```
Unnamed: 0             0
Name                   0
Location               0
Year                   0
Kilometers_Driven      0
Fuel_Type              0
Transmission           0
Owner_Type             0
Mileage                2
Engine                36
Power                 36
Seats                 42
New_Price           5195
```

```
dtype: int64
```

```
train.describe()
```

	Unnamed: 0	Year	Kilometers_Driven	Seats	Price
count	6019.000000	6019.000000	6.019000e+03	5977.000000	6019.000000
mean	3009.000000	2013.358199	5.873838e+04	5.278735	9.479468
std	1737.679967	3.269742	9.126884e+04	0.808840	11.187917
min	0.000000	1998.000000	1.710000e+02	0.000000	0.440000
25%	1504.500000	2011.000000	3.400000e+04	5.000000	3.500000
50%	3009.000000	2014.000000	5.300000e+04	5.000000	5.640000
75%	4513.500000	2016.000000	7.300000e+04	5.000000	9.950000
max	6018.000000	2019.000000	6.500000e+06	10.000000	160.000000

```
train.shape
```

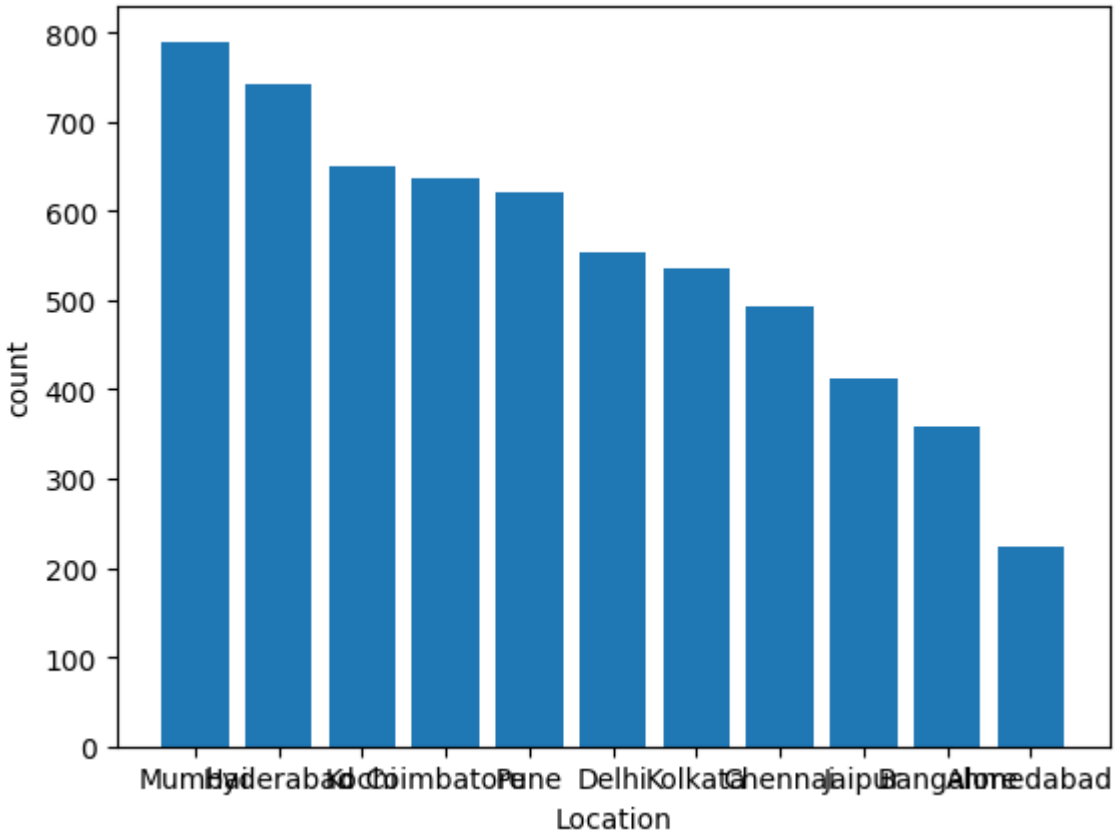
```
(6019, 14)
```

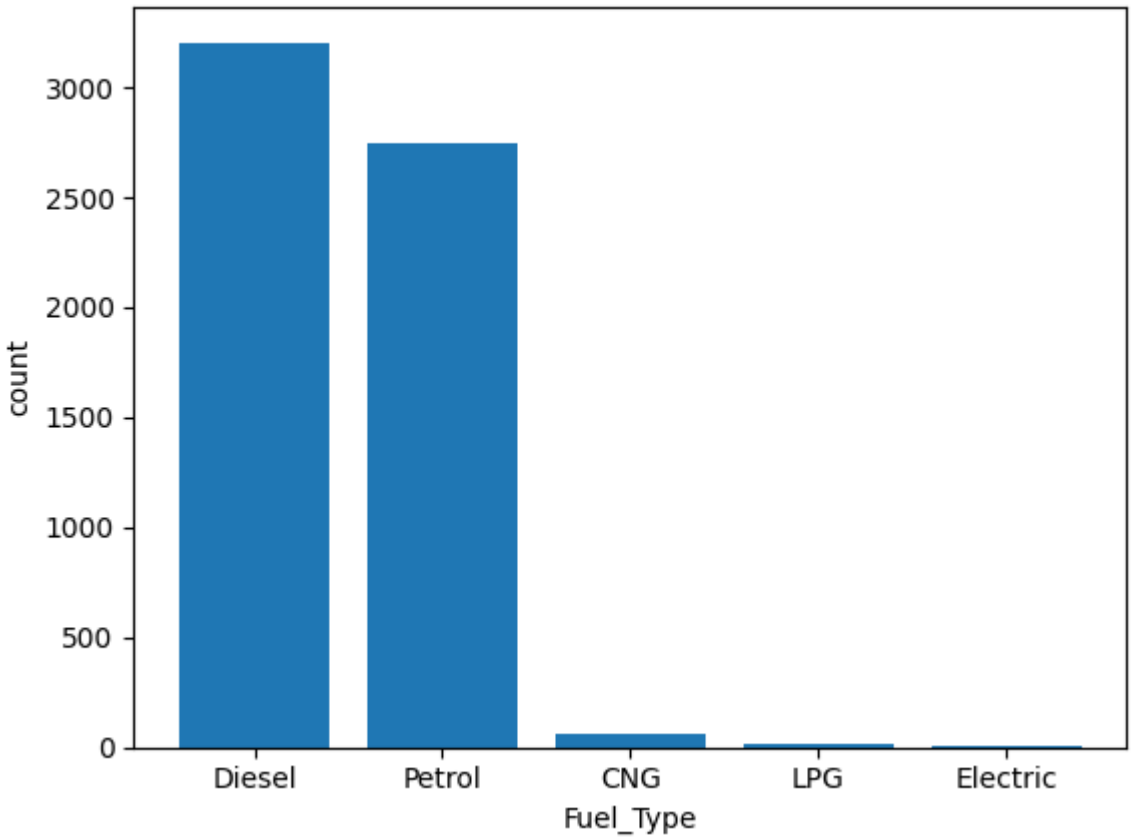
```
ls=['Name','Location','Fuel_Type','Transmission','Owner_Type']

for i in ls:
  count=train[i].value_counts()
  print('column ',i,'have ',len(count),' unique values')
  print(count.index)
  print('*'*100)
```

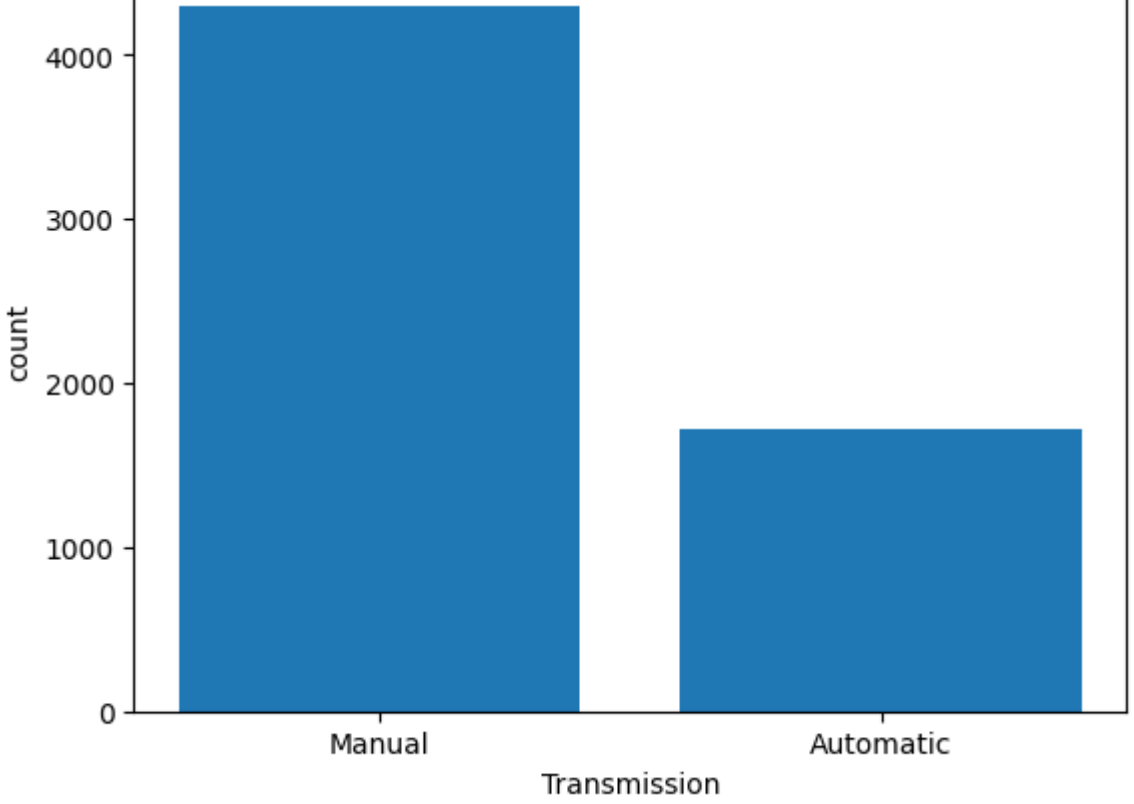
```
column  Name have  1878  unique values
Index(['Mahindra XUV500 W8 2WD', 'Maruti Swift VDI', 'Honda City 1.5 S MT',
       'Maruti Swift Dzire VDI', 'Maruti Swift VDI BSIV', 'Maruti Ritz VDi',
       'Hyundai i10 Sportz', 'Toyota Fortuner 3.0 Diesel',
       'Honda Amaze S i-Dtech', 'Hyundai Grand i10 Sportz',
       ...
       'Mahindra Scorpio SLE BSIII', 'Land Rover Discovery HSE Luxury 3.0 TD6',
       'Hyundai Tucson 2.0 Dual VTVT 2WD AT GL', 'Audi A4 2.0 TFSI',
       'Volvo S60 D4 SUMMUM', 'Ford Fiesta Titanium 1.5 TDCi',
       'Mahindra Scorpio S10 AT 4WD', 'Hyundai i20 1.2 Era',
       'Toyota Camry W4 (AT)', 'Mahindra Xylo D4 BSIV'],
      dtype='object', length=1878)
****************************************************************************************************
column  Location have  11  unique values
Index(['Mumbai', 'Hyderabad', 'Kochi', 'Coimbatore', 'Pune', 'Delhi',
       'Kolkata', 'Chennai', 'Jaipur', 'Bangalore', 'Ahmedabad'],
      dtype='object')
****************************************************************************************************
column  Fuel_Type have  5  unique values
Index(['Diesel', 'Petrol', 'CNG', 'LPG', 'Electric'], dtype='object')
****************************************************************************************************
column  Transmission have  2  unique values
Index(['Manual', 'Automatic'], dtype='object')
****************************************************************************************************
column  Owner_Type have  4  unique values
Index(['First', 'Second', 'Third', 'Fourth & Above'], dtype='object')
****************************************************************************************************
```

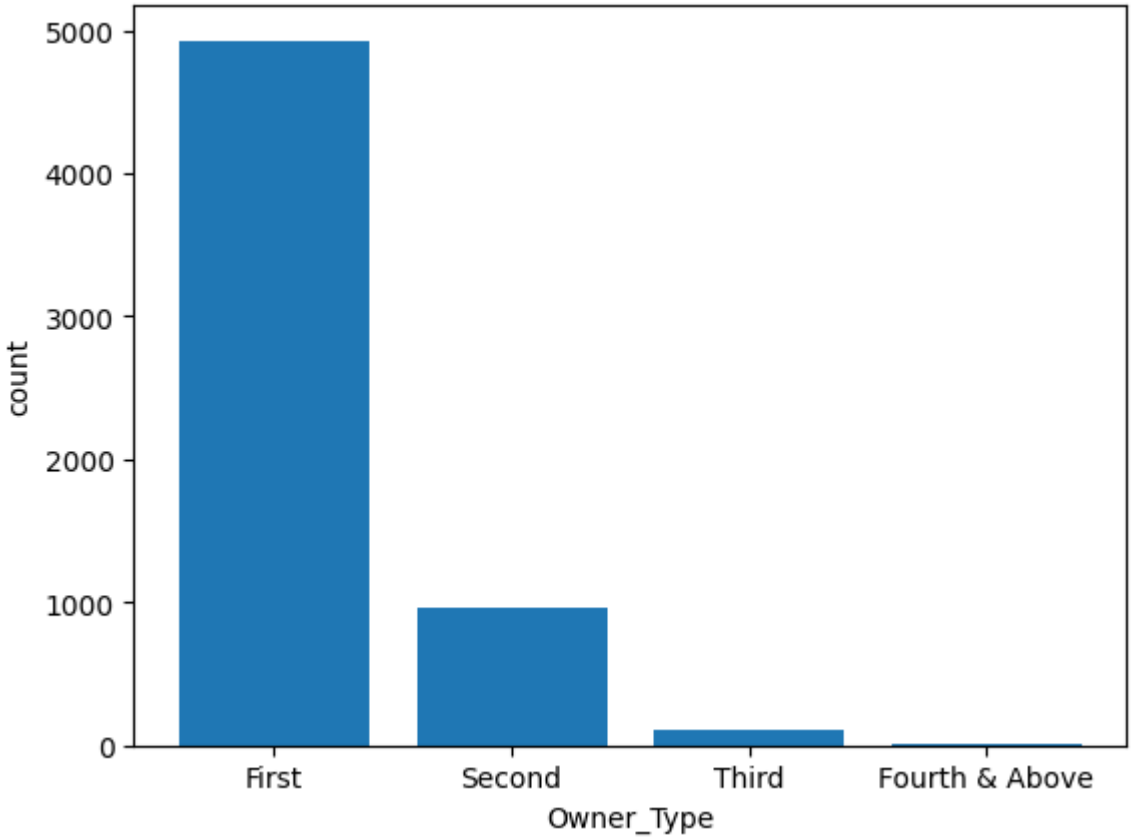
****

```
lst=['Location','Fuel_Type','Transmission','Owner_Type']

for i in lst:
  coun=train[i].value_counts()
  plt.bar(coun.index,coun)
  plt.xlabel(i)
  plt.ylabel('count')
  plt.show()
```










```
# we should drop columns named - unnamed,newprize,name
# newprize column have large missing value and name have large set of unique values


#get dummies encoding
df1=pd.get_dummies(train[['Location','Fuel_Type','Transmission','Owner_Type']],drop_first=T
df1
```

	Location_Bangalore	Location_Chennai	Location_Coimbatore	Location_Delhi	Locat
0	0	0	0	0	
1	0	0	0	0	

2	0	1	0	0
3	0	1	0	0
4	0	0	1	0
...	...	...	...	...
6014	0	0	0	1
6015	0	0	0	0
6016	0	0	0	0
6017	0	0	0	0
6018	0	0	0	0

6019 rows × 18 columns

```
dfe=pd.concat([train,df1],axis=1)
dfe
```

	Unnamed: 0	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission
0	0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual
1	1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual
2	2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual
3	3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual
4	4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic
...	...	...	...	...	...	...	...
6014	6014	Maruti Swift VDI	Delhi	2014	27365	Diesel	Manual
6015	6015	Hyundai Xcent 1.1 CRDi S	Jaipur	2015	100000	Diesel	Manual
6016	6016	Mahindra Xylo D4 BSIV	Jaipur	2012	55000	Diesel	Manual
6017	6017	Maruti Wagon R VXI	Kolkata	2013	46000	Petrol	Manual

6018	6018	Chevrolet Beat Diesel	Hyderabad	2011		47000	Diesel	Manual

6019 rows × 32 columns


```
dfe.columns
```

```
Index(['Unnamed: 0', 'Name', 'Location', 'Year', 'Kilometers_Driven',
       'Fuel_Type', 'Transmission', 'Owner_Type', 'Mileage', 'Engine', 'Power',
       'Seats', 'New_Price', 'Price', 'Location_Bangalore', 'Location_Chennai',
       'Location_Coimbatore', 'Location_Delhi', 'Location_Hyderabad',
       'Location_Jaipur', 'Location_Kochi', 'Location_Kolkata',
       'Location_Mumbai', 'Location_Pune', 'Fuel_Type_Diesel',
       'Fuel_Type_Electric', 'Fuel_Type_LPG', 'Fuel_Type_Petrol',
       'Transmission_Manual', 'Owner_Type_Fourth & Above', 'Owner_Type_Second',
       'Owner_Type_Third'],
      dtype='object')
```

```
# test data file dont have a column named 'Fuel_Type_Electric' therefore we should drop it
dfe1=dfe.drop(['Unnamed: 0','Name','Location','Fuel_Type','Transmission','Owner_Type','New_
```

```
# replace unit from mileage,engine,power

dfe1['Mileage']=dfe1['Mileage'].str.replace('km/kg','')
dfe1['Mileage']=dfe1['Mileage'].str.replace('kmpl','')
dfe1['Engine']=dfe1['Engine'].str.replace('CC','')
dfe1['Power']=dfe1['Power'].str.replace('bhp','')

# there is 'null' in engine,power,mileage given in description

dfe1['Mileage']=dfe1['Mileage'].str.replace('null','0')
dfe1['Engine']=dfe1['Engine'].str.replace('null','0')
dfe1['Power']=dfe1['Power'].str.replace('null','0')

dfe1
```

	Year	Kilometers_Driven	Mileage	Engine	Power	Seats	Price	Location_Bangalore
0	2010	72000	26.6	998	58.16	5.0	1.75	0
1	2015	41000	19.67	1582	126.2	5.0	12.50	0
2	2011	46000	18.2	1199	88.7	5.0	4.50	0
3	2012	87000	20.77	1248	88.76	7.0	6.00	0
4	2013	40670	15.2	1968	140.8	5.0	17.74	0
...	...	...	...	...	...	...	...	...
6014	2014	27365	28.4	1248	74	5.0	4.75	0
6015	2015	100000	24.4	1120	71	5.0	4.00	0
6016	2012	55000	14.0	2498	112	8.0	2.90	0

6017	2013	46000	18.9	998	67.1	5.0	2.65	0
6018	2011	47000	25.44	936	57.6	5.0	2.50	0

6019 rows × 24 columns

```
dfe1.dtypes

# uint8 un directional integer
```

```
Year                            int64
Kilometers_Driven               int64
Mileage                        object
Engine                         object
Power                          object
Seats                         float64
Price                         float64
Location_Bangalore              uint8
Location_Chennai                uint8
Location_Coimbatore             uint8
Location_Delhi                  uint8
Location_Hyderabad              uint8
Location_Jaipur                 uint8
Location_Kochi                  uint8
Location_Kolkata                uint8
Location_Mumbai                 uint8
Location_Pune                   uint8
Fuel_Type_Diesel                uint8
Fuel_Type_LPG                   uint8
Fuel_Type_Petrol                uint8
Transmission_Manual             uint8
Owner_Type_Fourth & Above       uint8
Owner_Type_Second               uint8
Owner_Type_Third                uint8
dtype: object
```

```
# convert datatype of object into int

dfe1['Engine']=dfe1['Engine'].astype(float)
dfe1['Mileage']=dfe1['Mileage'].astype(float)
dfe1['Power']=dfe1['Power'].astype(float)
dfe1.dtypes
```

```
Year                            int64
Kilometers_Driven               int64
Mileage                       float64
Engine                        float64
Power                         float64
Seats                         float64
Price                         float64
Location_Bangalore              uint8
Location_Chennai                uint8
Location_Coimbatore             uint8
Location_Delhi                  uint8
Location_Hyderabad              uint8
Location_Jaipur                 uint8
Location_Kochi                  uint8
Location_Kolkata                uint8
```

```
Location_Kolkata              uint8
Location_Mumbai               uint8
Location_Pune                 uint8
Fuel_Type_Diesel              uint8
Fuel_Type_LPG                 uint8
Fuel_Type_Petrol              uint8
Transmission_Manual           uint8
Owner_Type_Fourth & Above     uint8
Owner_Type_Second             uint8
Owner_Type_Third              uint8
dtype: object
```

```
dfe1.isna().sum()
```

```
Year                          0
Kilometers_Driven             0
Mileage                       2
Engine                       36
Power                        36
Seats                        42
Price                         0
Location_Bangalore            0
Location_Chennai              0
Location_Coimbatore           0
Location_Delhi                0
Location_Hyderabad            0
Location_Jaipur               0
Location_Kochi                0
Location_Kolkata              0
Location_Mumbai               0
Location_Pune                 0
Fuel_Type_Diesel              0
Fuel_Type_LPG                 0
Fuel_Type_Petrol              0
Transmission_Manual           0
Owner_Type_Fourth & Above     0
Owner_Type_Second             0
Owner_Type_Third              0
dtype: int64
```

```
# consider the '0' value we give instead of 'null' as a missing value and replace with NaN
dfe1.loc[dfe1.Engine==0,'Engine']=np.NaN
dfe1.loc[dfe1.Mileage==0,'Mileage']=np.NaN
dfe1.loc[dfe1.Power==0,'Power']=np.NaN
```

```
dfe1.isna().sum()
```

```
Year                          0
Kilometers_Driven             0
Mileage                      70
Engine                       36
Power                       143
Seats                        42
Price                         0
Location_Bangalore            0
Location_Chennai              0
Location_Coimbatore           0
Location_Delhi                0
Location_Hyderabad            0
Location_Jaipur               0
Location_Kochi                0
Location_Kolkata              0
```

```
Location_Mumbai                0
Location_Pune                  0
Fuel_Type_Diesel               0
Fuel_Type_LPG                  0
Fuel_Type_Petrol               0
Transmission_Manual            0
Owner_Type_Fourth & Above      0
Owner_Type_Second              0
Owner_Type_Third               0
dtype: int64
```

```
# filling missing value

dfe1['Mileage']=dfe1['Mileage'].fillna(dfe1['Mileage'].mean())
dfe1['Engine']=dfe1['Engine'].fillna(dfe1['Engine'].mean())
dfe1['Power']=dfe1['Power'].fillna(dfe1['Power'].mean())
dfe1['Seats']=dfe1['Seats'].fillna(dfe1['Seats'].mode()[0])
```

```
dfe1.isna().sum()
```

```
Year                         0
Kilometers_Driven            0
Mileage                      0
Engine                       0
Power                        0
Seats                        0
Price                        0
Location_Bangalore           0
Location_Chennai             0
Location_Coimbatore          0
Location_Delhi               0
Location_Hyderabad           0
Location_Jaipur              0
Location_Kochi               0
Location_Kolkata             0
Location_Mumbai              0
Location_Pune                0
Fuel_Type_Diesel             0
Fuel_Type_LPG                0
Fuel_Type_Petrol             0
Transmission_Manual          0
Owner_Type_Fourth & Above    0
Owner_Type_Second            0
Owner_Type_Third             0
dtype: int64
```

```
x=dfe1.drop(['Price'],axis=1)
y=dfe1['Price']
```

```
# loading test dataset and do the preproccessing

test=pd.read_csv('/content/test-data.csv')
test.head()
```

	Unnamed: 0	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Ow
0	0	Maruti Alto K10	Delhi	2014	40929	CNG	Manual	

		LXI CNG						
1	1	Maruti Alto 800 2016-2019 LXI	Coimbatore	2013	54493	Petrol	Manual	
2	2	Toyota Innova Crysta Touring Sport 2.4 MT	Mumbai	2017	34000	Diesel	Manual	

```
test.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1234 entries, 0 to 1233
Data columns (total 13 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Unnamed: 0         1234 non-null   int64  
 1   Name               1234 non-null   object 
 2   Location           1234 non-null   object 
 3   Year               1234 non-null   int64  
 4   Kilometers_Driven  1234 non-null   int64  
 5   Fuel_Type          1234 non-null   object 
 6   Transmission       1234 non-null   object 
 7   Owner_Type         1234 non-null   object 
 8   Mileage            1234 non-null   object 
 9   Engine             1224 non-null   object 
 10  Power              1224 non-null   object 
 11  Seats              1223 non-null   float64
 12  New_Price          182 non-null    object 
dtypes: float64(1), int64(3), object(9)
memory usage: 125.5+ KB
```

```
test.describe()
```

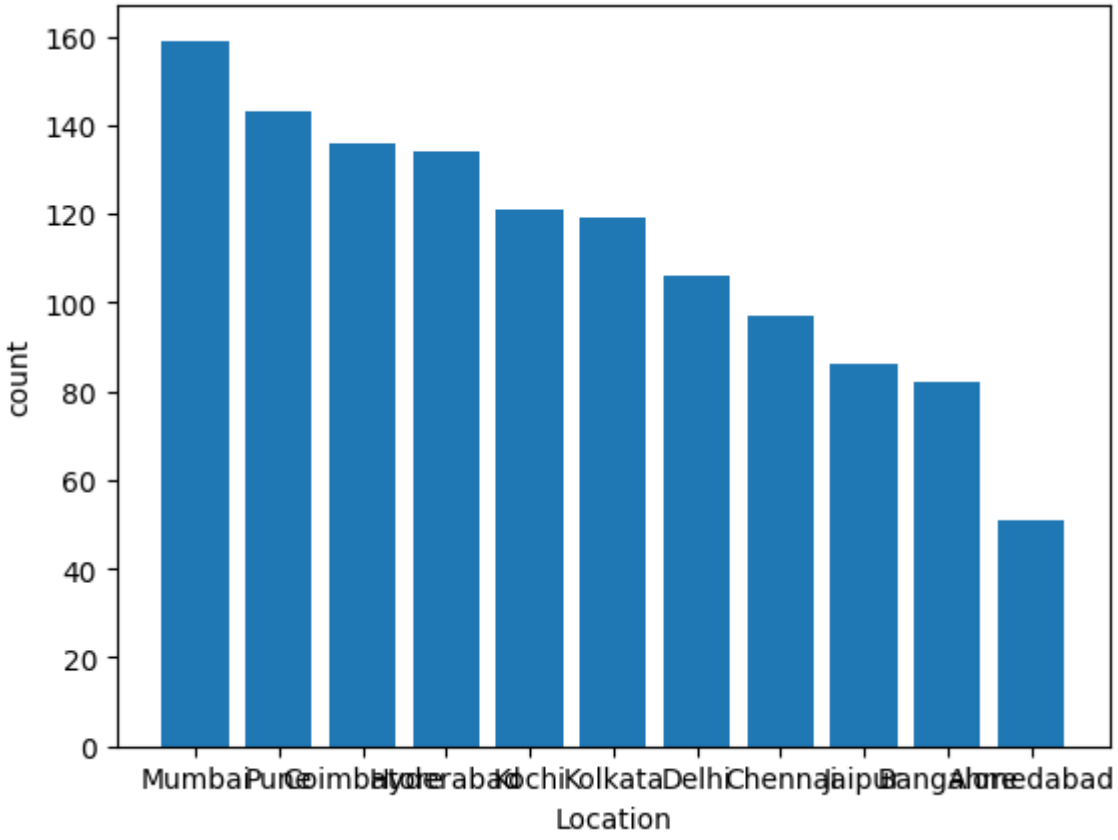
	Unnamed: 0	Year	Kilometers_Driven	Seats
count	1234.000000	1234.000000	1234.000000	1223.000000
mean	616.500000	2013.400324	58507.288493	5.284546
std	356.369424	3.179700	35598.702098	0.825622
min	0.000000	1996.000000	1000.000000	2.000000
25%	308.250000	2011.000000	34000.000000	5.000000
50%	616.500000	2014.000000	54572.500000	5.000000
75%	924.750000	2016.000000	75000.000000	5.000000
max	1233.000000	2019.000000	350000.000000	10.000000

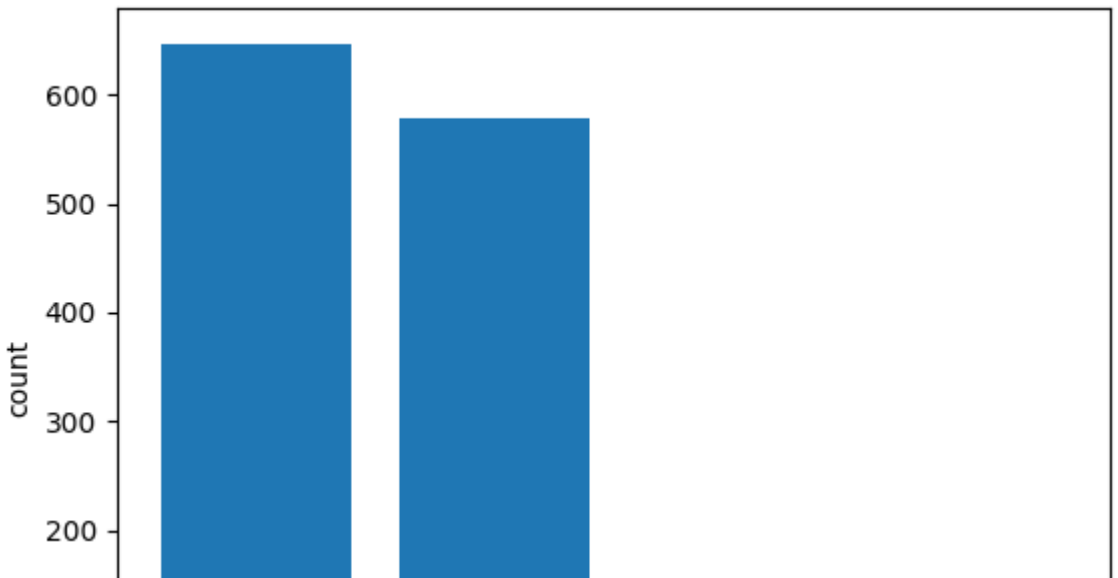
```
test.isna().sum()
```

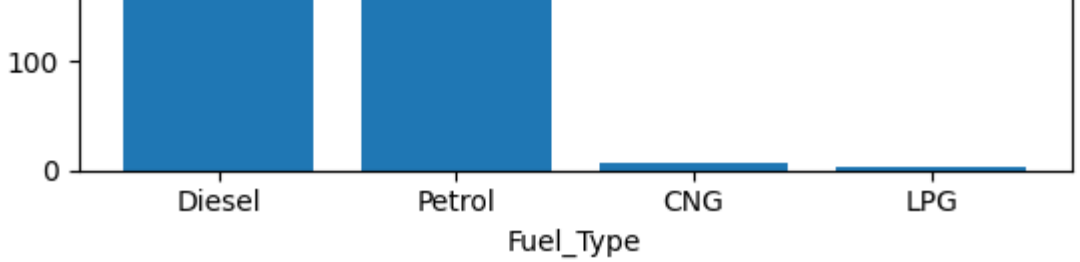
```
Unnamed: 0            0
Name                  0
Location              0
```

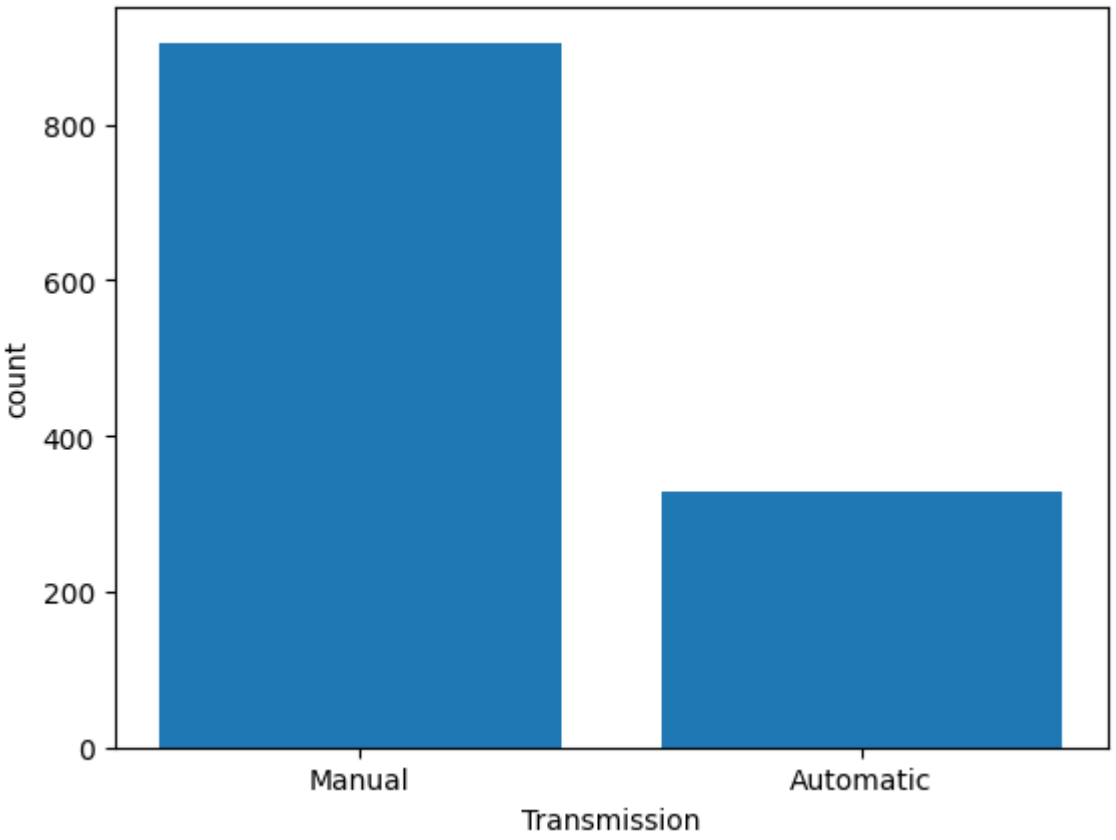
```
Year                    0
Kilometers_Driven       0
Fuel_Type               0
Transmission            0
Owner_Type              0
Mileage                 0
Engine                 10
Power                  10
Seats                  11
New_Price            1052
dtype: int64
```

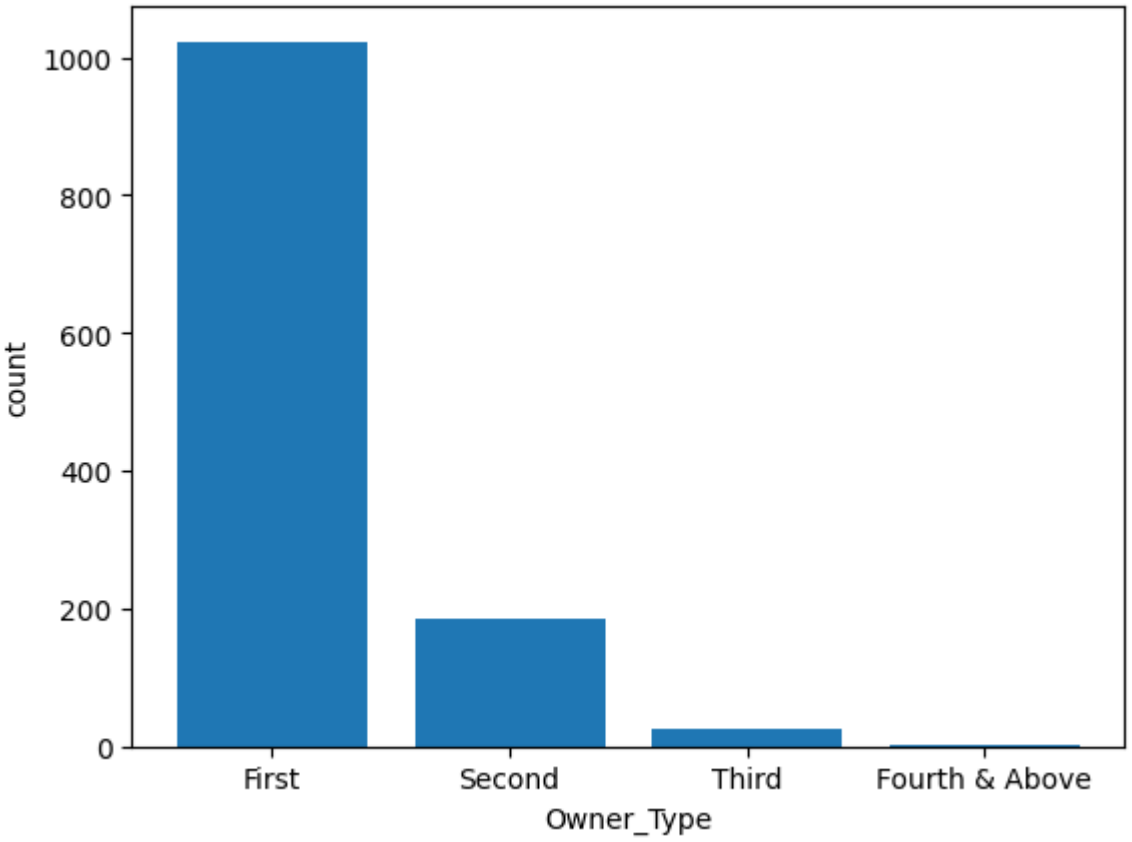
```
lst=['Location','Fuel_Type','Transmission','Owner_Type']

for i in lst:
  coun=test[i].value_counts()
  plt.bar(coun.index,coun)
  plt.xlabel(i)
  plt.ylabel('count')
  plt.show()
```












```
# encoding using get dummies

ts=pd.get_dummies(test[['Location','Fuel_Type','Transmission','Owner_Type']],drop_first=Tru
ts
```

	Location_Bangalore	Location_Chennai	Location_Coimbatore	Location_Delhi	Locat
0	0	0	0	1	
1	0	0	1	0	
2	0	0	0	0	
3	0	0	0	0	
4	0	0	0	0	
...	...	...	...	...	
1229	0	0	0	0	
1230	0	0	0	0	
1231	0	0	0	0	
1232	0	0	0	0	
1233	0	0	0	0	

1234 rows × 17 columns

```
tst=pd.concat([test,ts],axis=1)
tst
```

	Unnamed: 0	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmissio
0	0	Maruti Alto K10 LXI CNG	Delhi	2014	40929	CNG	Manua
1	1	Maruti Alto 800 2016-2019 LXI	Coimbatore	2013	54493	Petrol	Manua
2	2	Toyota Innova Crysta Touring Sport 2.4 MT	Mumbai	2017	34000	Diesel	Manua
3	3	Toyota Etios Liva GD	Hyderabad	2012	139000	Diesel	Manua
4	4	Hyundai i20 Magna	Mumbai	2014	29000	Petrol	Manua
...	...	...	...	...	...	...	.
1229	1229	Volkswagen Vento Diesel Trendline	Hyderabad	2011	89411	Diesel	Manua

1230	1230	Volkswagen Polo GT TSI	Mumbai	2015	59000	Petrol	Automati
1231	1231	Nissan Micra Diesel XV	Kolkata	2012	28000	Diesel	Manua
1232	1232	Volkswagen Polo GT TSI	Pune	2013	52262	Petrol	Automati
1233	1233	Mercedes-Benz E-Class 2009-2013 E 220 CDI Avan...	Kochi	2014	72443	Diesel	Automati

1234 rows × 30 columns

```
tst1=tst.drop(['Unnamed: 0','Name','Location','Fuel_Type','Transmission','Owner_Type','New_
tst1
```

	Year	Kilometers_Driven	Mileage	Engine	Power	Seats	Location_Bangalore	Locat
0	2014	40929	32.26 km/kg	998 CC	58.2 bhp	4.0	0	
1	2013	54493	24.7 kmpl	796 CC	47.3 bhp	5.0	0	
2	2017	34000	13.68 kmpl	2393 CC	147.8 bhp	7.0	0	
3	2012	139000	23.59 kmpl	1364 CC	null bhp	5.0	0	
4	2014	29000	18.5 kmpl	1197 CC	82.85 bhp	5.0	0	
...	...	...	...	...	...	...	...	...
1229	2011	89411	20.54 kmpl	1598 CC	103.6 bhp	5.0	0	
1230	2015	59000	17.21 kmpl	1197 CC	103.6 bhp	5.0	0	
1231	2012	28000	23.08 kmpl	1461 CC	63.1 bhp	5.0	0	
1232	2013	52262	17.2 kmpl	1197 CC	103.6 bhp	5.0	0	
1233	2014	72443	10.0 kmpl	2148 CC	170 bhp	5.0	0	

1234 rows × 23 columns

```
# removing the unit portion from the data

tst1['Mileage']=tst1['Mileage'].str.replace('km/kg','')
tst1['Mileage']=tst1['Mileage'].str.replace('kmpl','')
tst1['Engine']=tst1['Engine'].str.replace('CC','')
tst1['Power']=tst1['Power'].str.replace('bhp','')

# there is 'null' in engine,power,mileage given in description

tst1['Mileage']=tst1['Mileage'].str.replace('null','0')
tst1['Engine']=tst1['Engine'].str.replace('null','0')
tst1['Power']=tst1['Power'].str.replace('null','0')
```

```
tst1.dtypes
```

```
Year                           int64
Kilometers_Driven              int64
Mileage                       object
Engine                        object
Power                         object
Seats                        float64
Location_Bangalore             uint8
Location_Chennai               uint8
Location_Coimbatore            uint8
Location_Delhi                 uint8
Location_Hyderabad             uint8
Location_Jaipur                uint8
Location_Kochi                 uint8
Location_Kolkata               uint8
Location_Mumbai                uint8
Location_Pune                  uint8
Fuel_Type_Diesel               uint8
Fuel_Type_LPG                  uint8
Fuel_Type_Petrol               uint8
Transmission_Manual            uint8
Owner_Type_Fourth & Above      uint8
Owner_Type_Second              uint8
Owner_Type_Third               uint8
dtype: object
```

```
# convert datatype of object into int
tst1['Engine']=tst1['Engine'].astype(float)
tst1['Mileage']=tst1['Mileage'].astype(float)
tst1['Power']=tst1['Power'].astype(float)
tst1.dtypes
```

```
Year                           int64
Kilometers_Driven              int64
Mileage                      float64
Engine                       float64
Power                        float64
Seats                        float64
Location_Bangalore             uint8
Location_Chennai               uint8
Location_Coimbatore            uint8
Location_Delhi                 uint8
Location_Hyderabad             uint8
Location_Jaipur                uint8
Location_Kochi                 uint8
```

```
Location_Kolkata               uint8
Location_Mumbai                uint8
Location_Pune                  uint8
Fuel_Type_Diesel               uint8
Fuel_Type_LPG                  uint8
Fuel_Type_Petrol               uint8
Transmission_Manual            uint8
Owner_Type_Fourth & Above      uint8
Owner_Type_Second              uint8
Owner_Type_Third               uint8
dtype: object
```

```
# consider the '0' value we give instead of 'null' as a missing value and replace with NaN
tst1.loc[tst1.Engine==0,'Engine']=np.NaN
tst1.loc[tst1.Mileage==0,'Mileage']=np.NaN
tst1.loc[tst1.Power==0,'Power']=np.NaN
```

```
tst1.isna().sum()
```

```
Year                          0
Kilometers_Driven             0
Mileage                      13
Engine                       10
Power                        32
Seats                        11
Location_Bangalore            0
Location_Chennai              0
Location_Coimbatore           0
Location_Delhi                0
Location_Hyderabad            0
Location_Jaipur               0
Location_Kochi                0
Location_Kolkata              0
Location_Mumbai               0
Location_Pune                 0
Fuel_Type_Diesel              0
Fuel_Type_LPG                 0
Fuel_Type_Petrol              0
Transmission_Manual           0
Owner_Type_Fourth & Above     0
Owner_Type_Second             0
Owner_Type_Third              0
dtype: int64
```

```
# filling missing values

tst1['Mileage']=tst1['Mileage'].fillna(tst1['Mileage'].mean())
tst1['Engine']=tst1['Engine'].fillna(tst1['Engine'].mean())
tst1['Power']=tst1['Power'].fillna(tst1['Power'].mean())
tst1['Seats']=tst1['Seats'].fillna(tst1['Seats'].mode()[0])
```

```
tst1.isna().sum()
```

```
Year                          0
Kilometers_Driven             0
Mileage                       0
Engine                        0
Power                         0
Seats                         0
Location_Bangalore            0
```

```
Location_Chennai             0
Location_Coimbatore          0
Location_Delhi               0
Location_Hyderabad           0
Location_Jaipur              0
Location_Kochi               0
Location_Kolkata             0
Location_Mumbai              0
Location_Pune                0
Fuel_Type_Diesel             0
Fuel_Type_LPG                0
Fuel_Type_Petrol             0
Transmission_Manual          0
Owner_Type_Fourth & Above    0
Owner_Type_Second            0
Owner_Type_Third             0
dtype: int64
```

```
z=tst1
```

```
# model creation
from sklearn.linear_model import LinearRegression
ln=LinearRegression()
ln.fit(x,y)
ln.predict(z)
```

```
array([ 2.87588492, -1.29344912, 16.1069494 , ...,  0.1378514 ,
        9.27293255, 21.48043251])
```