problem statement:

check best fit the Dataset?

In [8]:
```
#importing the libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn import preprocessing,svm
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
```

In [9]:
```
df=pd.read_csv(r"C:\Users\MY HOME\Downloads\used_cars_data.csv")
df
```

Out[9]:

	S.No.	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New
0	0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5.0	
1	1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5.0	
2	2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5.0	8.6
3	3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7.0	
4	4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...
7248	7248	Volkswagen Vento Diesel Trendline	Hyderabad	2011	89411	Diesel	Manual	First	20.54 kmpl	1598 CC	103.6 bhp	5.0	
7249	7249	Volkswagen Polo GT TSI	Mumbai	2015	59000	Petrol	Automatic	First	17.21 kmpl	1197 CC	103.6 bhp	5.0	
7250	7250	Nissan Micra Diesel XV	Kolkata	2012	28000	Diesel	Manual	First	23.08 kmpl	1461 CC	63.1 bhp	5.0	
7251	7251	Volkswagen Polo GT TSI	Pune	2013	52262	Petrol	Automatic	Third	17.2 kmpl	1197 CC	103.6 bhp	5.0	
7252	7252	Mercedes-Benz E-Class 2009-2013 E 220 CDI Avan...	Kochi	2014	72443	Diesel	Automatic	First	10.0 kmpl	2148 CC	170 bhp	5.0	

7253 rows × 14 columns

In [10]:
```
df.isna().sum()
```

Out[10]:
```
S.No.                0
Name                 0
Location             0
Year                 0
Kilometers_Driven    0
Fuel_Type            0
Transmission         0
Owner_Type           0
Mileage              2
Engine              46
Power               46
Seats               53
New_Price         6247
Price             1234
dtype: int64
```

In [5]:
```
#Taking selected attributes for regression
df=df[['Year','Price']]
df
```

Out[5]:

	Year	Price
0	2010	1.75
1	2015	12.50
2	2011	4.50
3	2012	6.00
4	2013	17.74
...	...	...
7248	2011	NaN
7249	2015	NaN
7250	2012	NaN
7251	2013	NaN
7252	2014	NaN

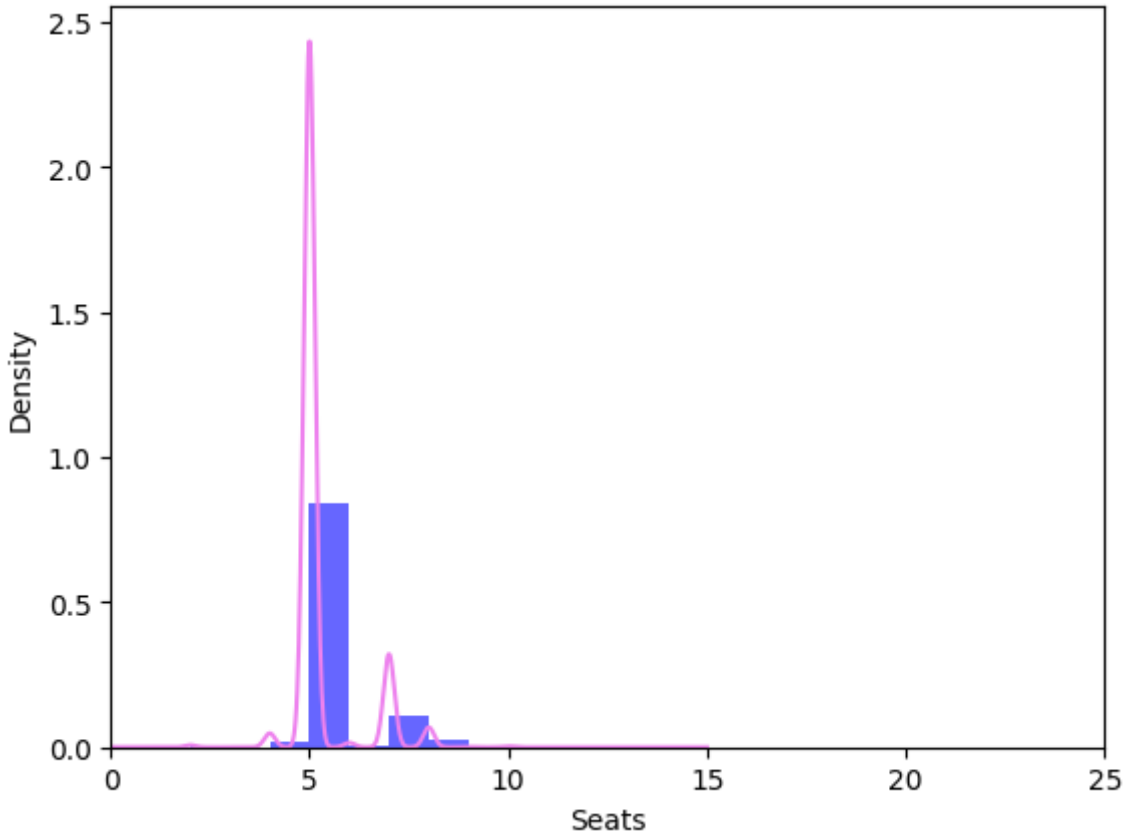
7253 rows × 2 columns

In [6]:
```
df.head(10)
```

Out[6]:

	yr	price
0	2010	1.75
1	2015	12.50
2	2011	4.50
3	2012	6.00
4	2013	17.74
5	2012	2.35
6	2013	3.50
7	2016	17.50
8	2013	5.20
9	2012	1.95

In [24]:
```
ax=df["Seats"].hist(bins=10,density=True,stacked=True,color='blue',alpha=0.6)
df["Seats"].plot(kind='density',color='violet')
ax.set(xlabel='Seats')
plt.xlim(-0,25)
plt.show()
```



In [12]:
```
print(df["Seats"].mean(skipna=True))
print(df["Seats"].median(skipna=True))
```

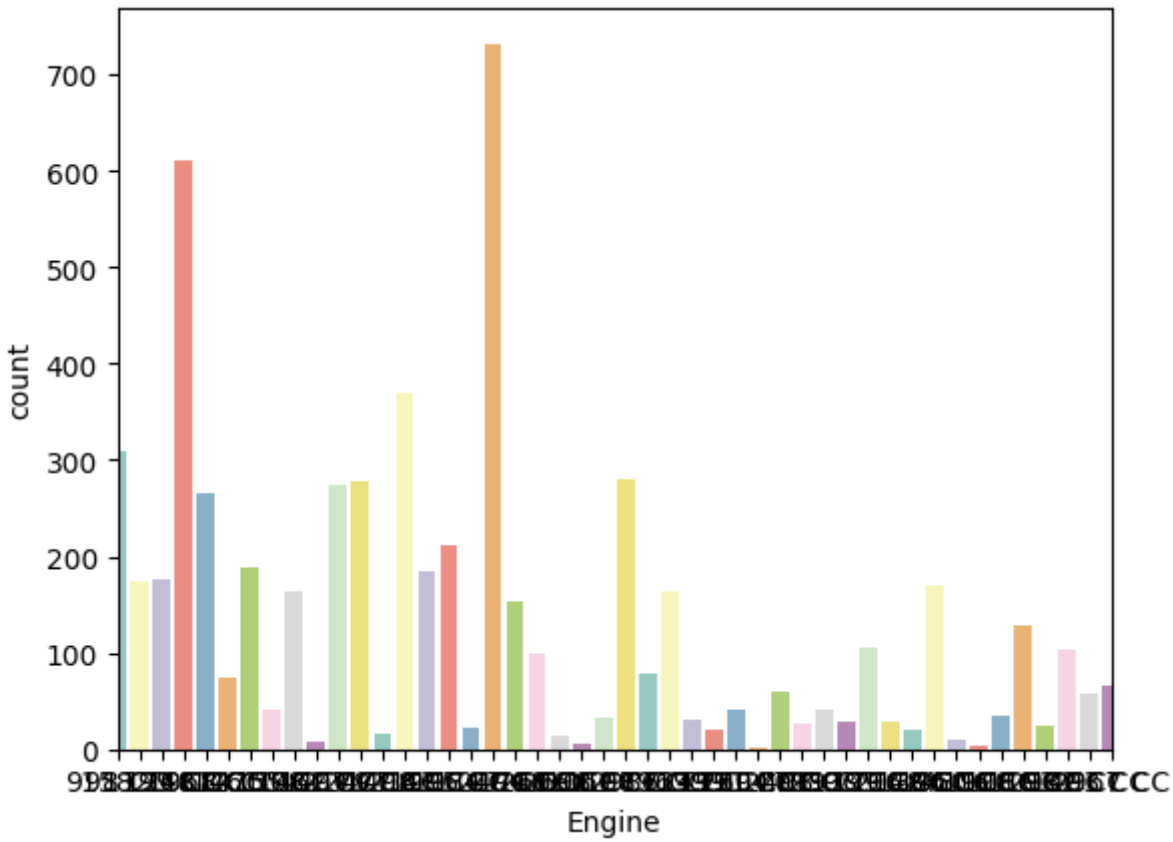
```
5.279722222222222
5.0
```

In [13]:
```
print(df["New_Price"].isnull().sum()/df.shape[0])
print(df["Price"].isnull().sum()/df.shape[0])
print(df["Mileage"].isnull().sum()/df.shape[0])
print(df["Engine"].isnull().sum()/df.shape[0])
print(df["Power"].isnull().sum()/df.shape[0])
```

```
0.8612987729215497
0.1701364952433476
0.0002757479663587481
0.006342203226251206
0.006342203226251206
```

In [14]:
```
print(df['Engine'].value_counts())
sns.countplot(x='Engine',data=df,palette='Set3')
plt.xlim(-0,45)
plt.show()
```

```
Engine
1197 CC    732
1248 CC    610
1498 CC    370
998 CC     309
1198 CC    281
          ...
1489 CC      1
1422 CC      1
2706 CC      1
1978 CC      1
1389 CC      1
Name: count, Length: 150, dtype: int64
```



In [15]:
```
data=df.copy()
data['Seats'].fillna(df['Seats'].median(skipna=True),inplace=True)
data.drop('New_Price',axis=1,inplace=True)
data['Price'].fillna(df['Price'].median(skipna=True),inplace=True)
data['Mileage'].fillna(df['Mileage'].value_counts().idxmax(),inplace=True)
data.drop('Engine',axis=1,inplace=True)
data.drop('Power',axis=1,inplace=True)
```

In [16]:
```
data.isnull().sum()
```

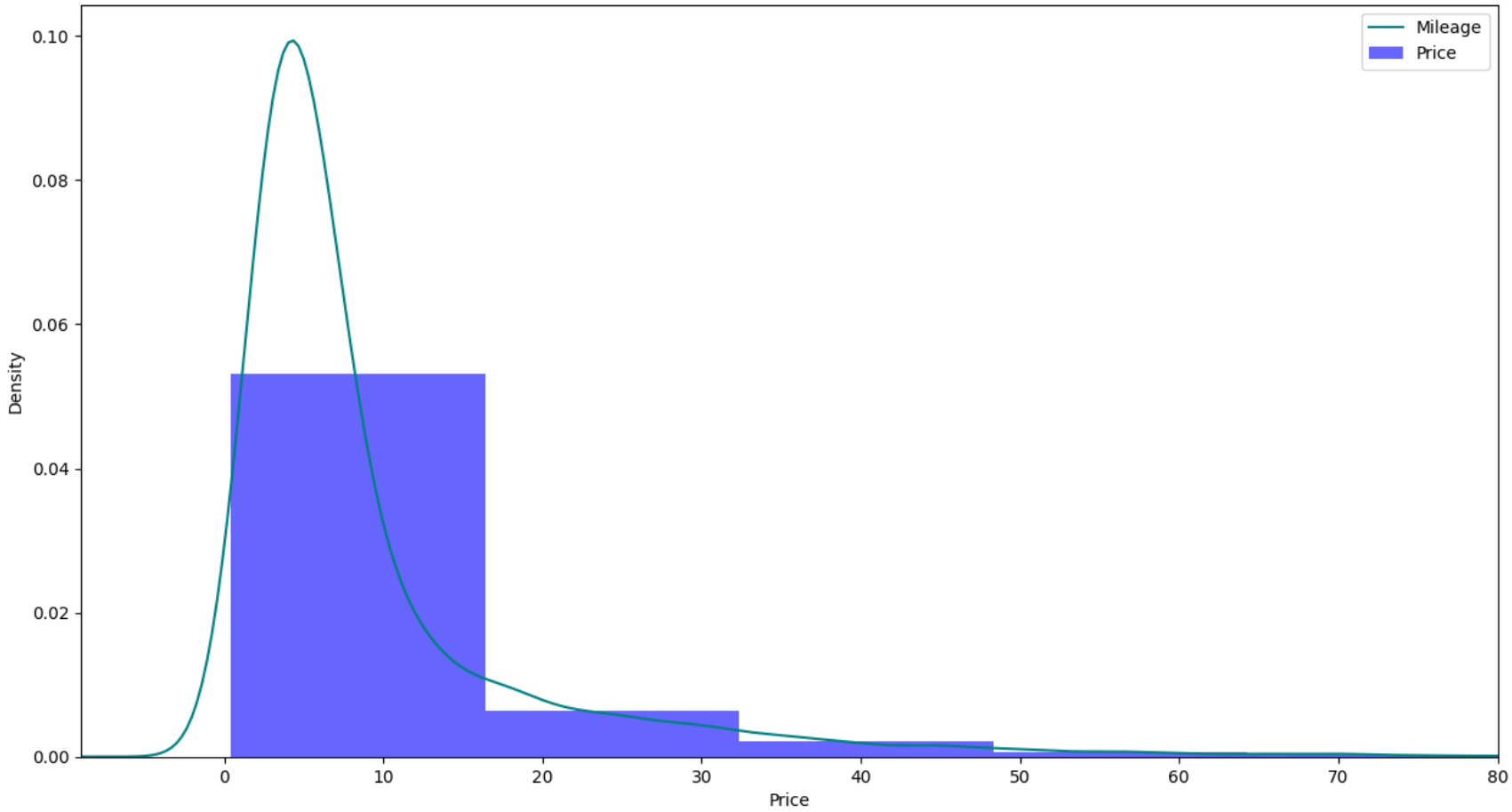
Out[16]:
```
S.No.                0
Name                 0
Location             0
Year                 0
Kilometers_Driven    0
Fuel_Type            0
Transmission         0
Owner_Type           0
Mileage              0
Seats                0
Price                0
dtype: int64
```

In [17]:
```
data.head()
```

Out[17]:

	S.No.	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Seats	Price
0	0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	5.0	1.75
1	1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	5.0	12.50
2	2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	5.0	4.50
3	3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	7.0	6.00
4	4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	5.0	17.74

In [23]:
```
plt.figure(figsize=(15,8))
ax=df["Price"].hist(bins=10,density=True,stacked=True,color='blue',alpha=0.6)
df["Price"].plot(kind='density',color='teal')
ax.legend(['Mileage','Price'])
ax.set(xlabel='Price')
plt.xlim(-9,80)
plt.show()
```



In [19]:
```
training=pd.get_dummies(data,columns=["S.No."])
final_train=training
final_train.head()
```

Out[19]:

	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Seats	Price	...	S.No._7243	S.No._724
0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	5.0	1.75	...	False	Fals
1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	5.0	12.50	...	False	Fals
2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	5.0	4.50	...	False	Fals
3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	7.0	6.00	...	False	Fals
4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	5.0	17.74	...	False	Fals

5 rows × 7263 columns

In [ ]: