Task 2: Clustering Analysis

In [1]:
```
import warnings
warnings.filterwarnings('ignore')
import pandas as pd
import numpy as np
```

In [2]:
```
data = pd.read_csv("Honda.csv")
data.to_csv('100131001-100131002_T2Org.csv', index=False)
data.head()
```

Out[2]:

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	mileage	engine	max_power	torque	seats
0	Honda Amaze E i-DTEC	2017	500000	70000	Diesel	Individual	Manual	First Owner	25.8 kmpl	1498 CC	98.6 bhp	200Nm@ 1750rpm	5
1	Honda Amaze E i-DTEC	2017	515000	60000	Diesel	Individual	Manual	Second Owner	25.8 kmpl	1498 CC	98.6 bhp	200Nm@ 1750rpm	5
2	Honda Amaze E i-VTEC	2017	475000	57000	Petrol	Individual	Manual	First Owner	17.8 kmpl	1198 CC	86.7 bhp	109Nm@ 4500rpm	5
3	Honda Amaze E Option i-DTEC	2017	550000	120000	Diesel	Individual	Manual	First Owner	25.8 kmpl	1498 CC	98.6 bhp	200Nm@ 1750rpm	5
4	Honda Amaze i-VTEC Privilege Edition	2017	490000	80000	Petrol	Individual	Manual	First Owner	17.8 kmpl	1198 CC	86.7 bhp	109Nm@ 4500rpm	5

In [3]:
```
edit_data = data.drop(['name', 'fuel','seller_type', 'transmission', 'owner', 'mileage', 'engine', 'max_power', 'torque'],axis=1)
edit_data.to_csv('100131001-100131002_T2Mod.csv', index=False)
```

In [4]:
```
from sklearn.preprocessing import StandardScaler
X = StandardScaler().fit_transform(edit_data)
```

Perform any necessary preprocessing steps. Add explanation to your report.

Ans: I dropped the unused features and only keep the features I plan to use, such as year, selling_price, km_driven, seats.

In [5]:
```
from sklearn.cluster import KMeans
kmeans_3 = KMeans(n_clusters=3)
kmeans_3.fit(X)
kmeans_4 = KMeans(n_clusters=4)
kmeans_4.fit(X)
kmeans_5 = KMeans(n_clusters=5)
kmeans_5.fit(X)
print("SSE for k=3: {}\nSSE for k=4: {}\nSSE for k=5: {}".
      format(round(kmeans_3.inertia_, 4), round(kmeans_4.inertia_,4), round(kmeans_5.inertia_, 4)))
print("The least SSE is k=5 with SSE={}".format(round(kmeans_5.inertia_, 4)))
```

```
SSE for k=3: 189.2816
SSE for k=4: 143.4199
SSE for k=5: 112.2917
The least SSE is k=5 with SSE=112.2917
```

In [6]:
```
edit_data['class'] = kmeans_5.labels_
edit_data.to_csv('100131001-100131002_T2Class.csv', index=False)
edit_data
```

Out[6]:

	year	selling_price	km_driven	seats	class
0	2017	500000	70000	5	2
1	2017	515000	60000	5	2
2	2017	475000	57000	5	2
3	2017	550000	120000	5	2
4	2017	490000	80000	5	2
...	...	...	...	...	...
90	2019	1300000	20000	5	3
91	2019	2000000	24857	5	4
92	2019	840000	1500	5	3
93	2019	750000	3100	5	3
94	2019	840000	1500	5	3

95 rows × 5 columns

In [ ]:
```

```