In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import scipy.cluster.hierarchy as sch
from sklearn.cluster import AgglomerativeClustering
from sklearn.preprocessing import normalize
```

In [2]:
```
airline=pd.read_csv("C:\\Users\\Admin\\Downloads\\assignment 4\\EastWestAirlines.csv")
airline
```

Out[2]:

	ID#	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flight_mile
0	1	28143	0	1	1	1	174	1	
1	2	19244	0	1	1	1	215	2	
2	3	41354	0	1	1	1	4123	4	
3	4	14776	0	1	1	1	500	1	
4	5	97752	0	4	1	1	43300	26	
...	...	...	...	...	...	...	...	...	
3994	4017	18476	0	1	1	1	8525	4	
3995	4018	64385	0	1	1	1	981	5	
3996	4019	73597	0	3	1	1	25447	8	
3997	4020	54899	0	1	1	1	500	1	
3998	4021	3016	0	1	1	1	0	0	

3999 rows × 12 columns

In [3]:
```
airline.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3999 entries, 0 to 3998
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   ID#                3999 non-null   int64
 1   Balance            3999 non-null   int64
 2   Qual_miles         3999 non-null   int64
 3   cc1_miles          3999 non-null   int64
 4   cc2_miles          3999 non-null   int64
 5   cc3_miles          3999 non-null   int64
 6   Bonus_miles        3999 non-null   int64
 7   Bonus_trans        3999 non-null   int64
 8   Flight_miles_12mo  3999 non-null   int64
 9   Flight_trans_12    3999 non-null   int64
 10  Days_since_enroll  3999 non-null   int64
 11  Award?             3999 non-null   int64
```

```
dtypes: int64(12)
memory usage: 375.0 KB
```

In [4]:
```
airline.describe()
```

Out[4]:

	ID#	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles
count	3999.000000	3.999000e+03	3999.000000	3999.000000	3999.000000	3999.000000	3999.000000
mean	2014.819455	7.360133e+04	144.114529	2.059515	1.014504	1.012253	17144.846212
std	1160.764358	1.007757e+05	773.663804	1.376919	0.147650	0.195241	24150.967826
min	1.000000	0.000000e+00	0.000000	1.000000	1.000000	1.000000	0.000000
25%	1010.500000	1.852750e+04	0.000000	1.000000	1.000000	1.000000	1250.000000
50%	2016.000000	4.309700e+04	0.000000	1.000000	1.000000	1.000000	7171.000000
75%	3020.500000	9.240400e+04	0.000000	3.000000	1.000000	1.000000	23800.500000
max	4021.000000	1.704838e+06	11148.000000	5.000000	3.000000	5.000000	263685.000000

In [5]:
```
airline.shape
```

Out[5]:
```
(3999, 12)
```

In [6]:
```
airline2=airline.drop(['ID#'],axis=1)
airline2
```

Out[6]:

	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo
0	28143	0	1	1	1	174	1	
1	19244	0	1	1	1	215	2	
2	41354	0	1	1	1	4123	4	
3	14776	0	1	1	1	500	1	
4	97752	0	4	1	1	43300	26	207
...	...	...	...	...	...	...	...	...
3994	18476	0	1	1	1	8525	4	20(
3995	64385	0	1	1	1	981	5	
3996	73597	0	3	1	1	25447	8	
3997	54899	0	1	1	1	500	1	50(
3998	3016	0	1	1	1	0	0	

3999 rows × 11 columns

In [7]:

```
airline2_norm=pd.DataFrame(normalize(airline2),columns=airline2.columns)
airline2_norm
```

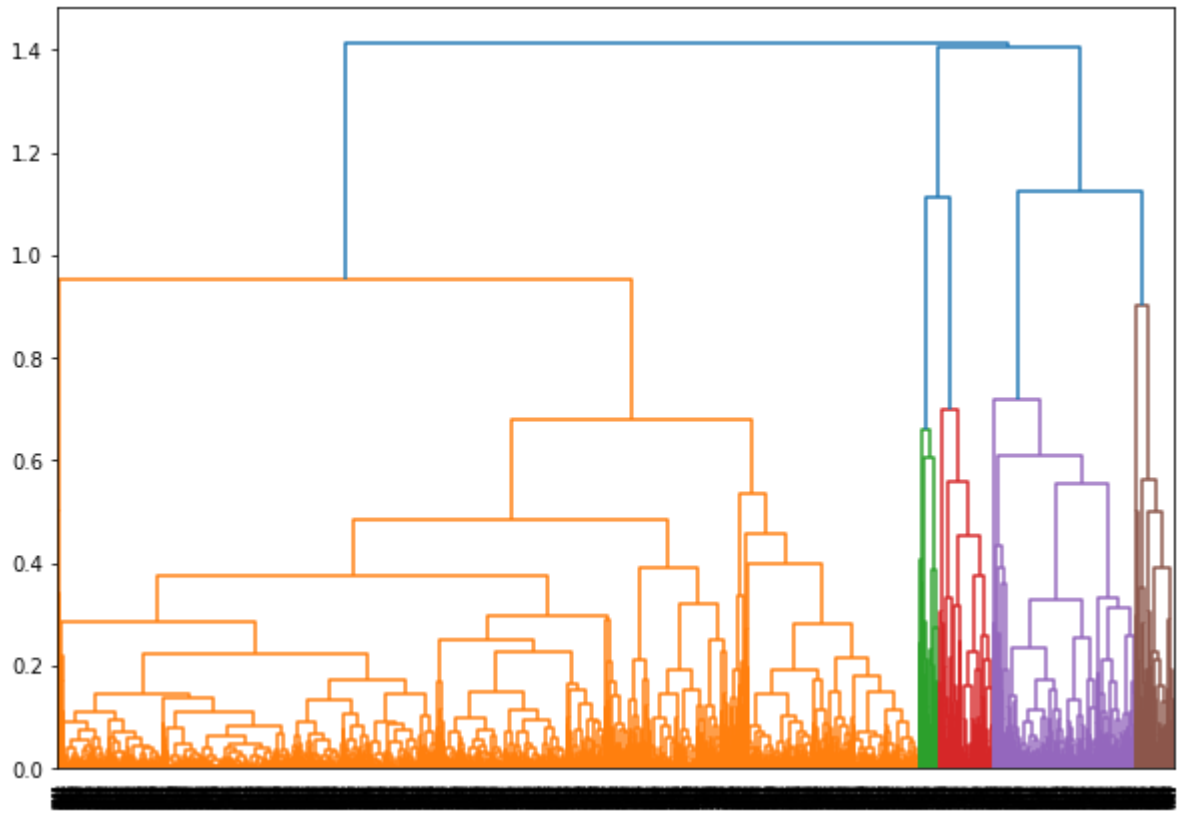
Out[7]:

	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flight_miles_12m
0	0.970414	0.0	0.000034	0.000034	0.000034	0.006000	0.000034	0.00000
1	0.940209	0.0	0.000049	0.000049	0.000049	0.010504	0.000098	0.00000
2	0.981113	0.0	0.000024	0.000024	0.000024	0.097817	0.000095	0.00000
3	0.904428	0.0	0.000061	0.000061	0.000061	0.030605	0.000061	0.00000
4	0.912226	0.0	0.000037	0.000009	0.000009	0.404078	0.000243	0.01938
...	...	...	...	...	...	...	...	
3994	0.905810	0.0	0.000049	0.000049	0.000049	0.417949	0.000196	0.00980
3995	0.999649	0.0	0.000016	0.000016	0.000016	0.015231	0.000078	0.00000
3996	0.944948	0.0	0.000039	0.000013	0.000013	0.326726	0.000103	0.00000
3997	0.999592	0.0	0.000018	0.000018	0.000018	0.009104	0.000018	0.00910
3998	0.907271	0.0	0.000301	0.000301	0.000301	0.000000	0.000000	0.00000

3999 rows × 11 columns

In [15]:

```
plt.figure(figsize=(10,7))
dendograms=sch.dendrogram(sch.linkage(airline2_norm,'complete'))
```



In [9]:
```
hclusters=AgglomerativeClustering(n_clusters=5,affinity='euclidean',linkage='ward')
hclusters
```

Out[9]:
```
AgglomerativeClustering(n_clusters=5)
```

In [16]:
```
y=pd.DataFrame(hclusters.fit_predict(airline2_norm),columns=['clustersid'])
y['clustersid'].value_counts()
```

Out[16]:
```
2    1547
4    1191
3     579
1     453
0     229
Name: clustersid, dtype: int64
```

In [17]:
```
airline2['clustersid']=hclusters.labels_
airline2
```

Out[17]:

	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo
0	28143	0	1	1	1	174	1	0
1	19244	0	1	1	1	215	2	0
2	41354	0	1	1	1	4123	4	0
3	14776	0	1	1	1	500	1	0
4	97752	0	4	1	1	43300	26	2077
...	...	...	...	...	...	...	...	...
3994	18476	0	1	1	1	8525	4	200
3995	64385	0	1	1	1	981	5	0
3996	73597	0	3	1	1	25447	8	0
3997	54899	0	1	1	1	500	1	500
3998	3016	0	1	1	1	0	0	0

3999 rows × 12 columns

In [18]:
```
airline2.groupby('clustersid').agg(['mean']).reset_index()
```

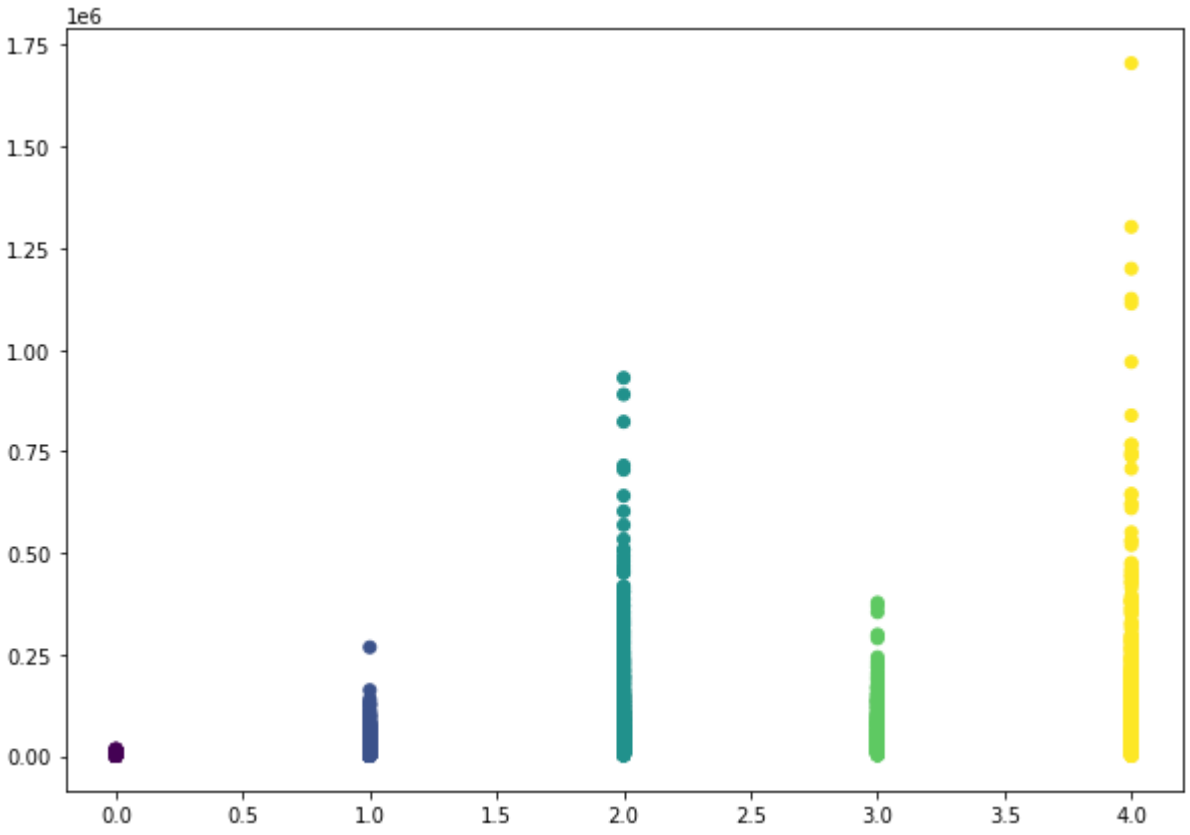
Out[18]:

	clustersid	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flig
		mean	mean	mean	mean	mean	mean	mean	
0	0	5524.222707	8.755459	1.000000	1.000000	1.000000	584.532751	2.401747	
1	1	31066.514349	111.415011	3.200883	1.026490	1.070640	40266.935982	17.289183	
2	2	81201.080802	136.521008	2.115061	1.013575	1.000646	16350.149968	13.574014	
3	3	69569.894646	97.257340	3.326425	1.032815	1.022453	35743.675302	17.784111	

	clustersid	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flig
		mean	mean	mean	mean	mean	mean	mean	
4	4	94957.590260	215.220823	1.141058	1.005038	1.002519	3524.928631	5.640638	

In [19]:
```
plt.figure(figsize=(10,7))
plt.scatter(airline2['clustersid'],airline2['Balance'],c=hclusters.labels_)
```

Out[19]:
```
<matplotlib.collections.PathCollection at 0x1c6c81c9490>
```



In [ ]: