In [1]:

```
import pandas as pd
from sklearn.preprocessing import normalize
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.cluster.hierarchy as sch
from sklearn.cluster import AgglomerativeClustering
```

In [2]:

```
Excel = pd.ExcelFile('EastWestAirlines.xlsx')
```

In [3]:

```
airline_data = pd.read_excel(Excel,'data')
airline_data
```

Out[3]:

| | ID# | Balance | Qual_miles | cc1_miles | cc2_miles | cc3_miles | Bonus_miles | Bonus_trans |
|---|---|---|---|---|---|---|---|---|
| **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 [4]:

```
airline_data.head(10)
```

Out[4]:

| | ID# | Balance | Qual_miles | cc1_miles | cc2_miles | cc3_miles | Bonus_miles | Bonus_trans | Flig |
|---|---|---|---|---|---|---|---|---|---|
| **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 | |
| **5** | 6 | 16420 | 0 | 1 | 1 | 1 | 0 | 0 | |
| **6** | 7 | 84914 | 0 | 3 | 1 | 1 | 27482 | 25 | |
| **7** | 8 | 20856 | 0 | 1 | 1 | 1 | 5250 | 4 | |
| **8** | 9 | 443003 | 0 | 3 | 2 | 1 | 1753 | 43 | |
| **9** | 10 | 104860 | 0 | 3 | 1 | 1 | 28426 | 28 | |

In [5]:

```
airline_data.shape
```

Out[5]:

```
(3999, 12)
```

In [6]:

```
airline_data.dtypes
```

Out[6]:

```
ID#                int64
Balance            int64
Qual_miles         int64
cc1_miles          int64
cc2_miles          int64
cc3_miles          int64
Bonus_miles        int64
Bonus_trans        int64
Flight_miles_12mo  int64
Flight_trans_12    int64
Days_since_enroll  int64
Award?             int64
dtype: object
```

In [7]:

```
airline_data.isna().sum()
```

Out[7]:

```
ID#                 0
Balance             0
Qual_miles          0
cc1_miles           0
cc2_miles           0
cc3_miles           0
Bonus_miles         0
Bonus_trans         0
Flight_miles_12mo   0
Flight_trans_12     0
Days_since_enroll   0
Award?              0
dtype: int64
```

In [8]:

```
del airline_data['ID#']
```

In [9]:

```
airline_norm = pd.DataFrame(normalize(airline_data),columns=airline_data.columns)
airline_norm
```

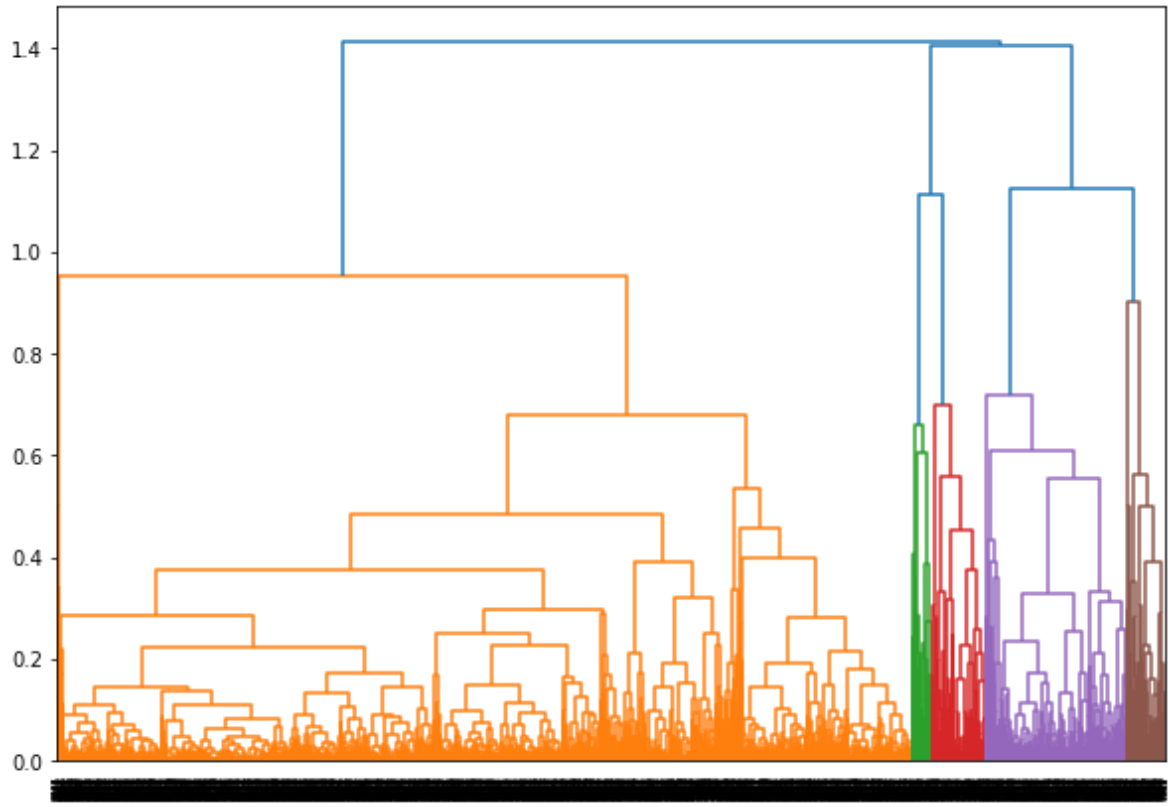
Out[9]:

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

3999 rows × 11 columns

In [10]:

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




In [11]:

```
#create cluster
hc = AgglomerativeClustering( n_clusters=5,affinity='euclidean',linkage='ward')
hc
```

Out[11]:

```
AgglomerativeClustering(n_clusters=5)
```

In [12]:

```
y_hc=hc.fit_predict(airline_norm)
y_hc
```

Out[12]:

```
array([4, 2, 2, ..., 2, 4, 2], dtype=int64)
```

In [13]:

```
clusters=pd.DataFrame(y_hc,columns=['Clusters'])
clusters
```

Out[13]:

| | Clusters |
|---|---|
| **0** | 4 |
| **1** | 2 |
| **2** | 2 |
| **3** | 2 |
| **4** | 3 |
| **...** | ... |
| **3994** | 3 |
| **3995** | 4 |
| **3996** | 2 |
| **3997** | 4 |
| **3998** | 2 |

3999 rows × 1 columns

In [14]:

```
airline_data['clusters'] = clusters
```

In [15]:

```
airline_data
```

Out[15]:

| | Balance | Qual_miles | cc1_miles | cc2_miles | cc3_miles | Bonus_miles | Bonus_trans | Fl |
|---|---|---|---|---|---|---|---|---|
| **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 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | |
| **3994** | 18476 | 0 | 1 | 1 | 1 | 8525 | 4 | |
| **3995** | 64385 | 0 | 1 | 1 | 1 | 981 | 5 | |
| **3996** | 73597 | 0 | 3 | 1 | 1 | 25447 | 8 | |
| **3997** | 54899 | 0 | 1 | 1 | 1 | 500 | 1 | |
| **3998** | 3016 | 0 | 1 | 1 | 1 | 0 | 0 | |

3999 rows × 12 columns

In [16]:

```
airline_data[airline_data['clusters']==0]
```

Out[16]:

| | Balance | Qual_miles | cc1_miles | cc2_miles | cc3_miles | Bonus_miles | Bonus_trans | Flight |
|---|---|---|---|---|---|---|---|---|
| **27** | 8828 | 0 | 1 | 1 | 1 | 0 | 0 | |
| **31** | 10021 | 0 | 1 | 1 | 1 | 0 | 0 | |
| **39** | 2176 | 0 | 1 | 1 | 1 | 0 | 0 | |
| **51** | 1300 | 0 | 1 | 1 | 1 | 370 | 1 | |
| **55** | 14448 | 0 | 1 | 1 | 1 | 1625 | 6 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | |
| **3861** | 3126 | 0 | 1 | 1 | 1 | 100 | 1 | |
| **3876** | 1000 | 0 | 1 | 1 | 1 | 0 | 0 | |
| **3942** | 2131 | 0 | 1 | 1 | 1 | 405 | 3 | |
| **3981** | 1010 | 0 | 1 | 1 | 1 | 0 | 0 | |
| **3984** | 404 | 0 | 1 | 1 | 1 | 550 | 3 | |

229 rows × 12 columns

In [17]:

```
airline_data[airline_data['clusters']==1]
```

Out[17]:

| | Balance | Qual_miles | cc1_miles | cc2_miles | cc3_miles | Bonus_miles | Bonus_trans | Flight |
|---|---|---|---|---|---|---|---|---|
| **15** | 28495 | 0 | 4 | 1 | 1 | 49442 | 15 | |
| **16** | 51890 | 0 | 4 | 1 | 1 | 48963 | 16 | |
| **41** | 10470 | 0 | 4 | 1 | 1 | 38094 | 26 | |
| **58** | 38077 | 0 | 3 | 1 | 1 | 34024 | 8 | |
| **78** | 49238 | 0 | 4 | 1 | 1 | 38037 | 18 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | |
| **3919** | 5000 | 0 | 1 | 1 | 1 | 5000 | 1 | |
| **3924** | 14775 | 0 | 1 | 1 | 1 | 14275 | 9 | |
| **3930** | 40424 | 0 | 4 | 1 | 1 | 44110 | 26 | |
| **3944** | 2124 | 0 | 1 | 1 | 1 | 2324 | 2 | |
| **3978** | 10071 | 0 | 2 | 1 | 1 | 27701 | 16 | |

453 rows × 12 columns

In [18]:

```
airline_data[airline_data['clusters']==4]
```

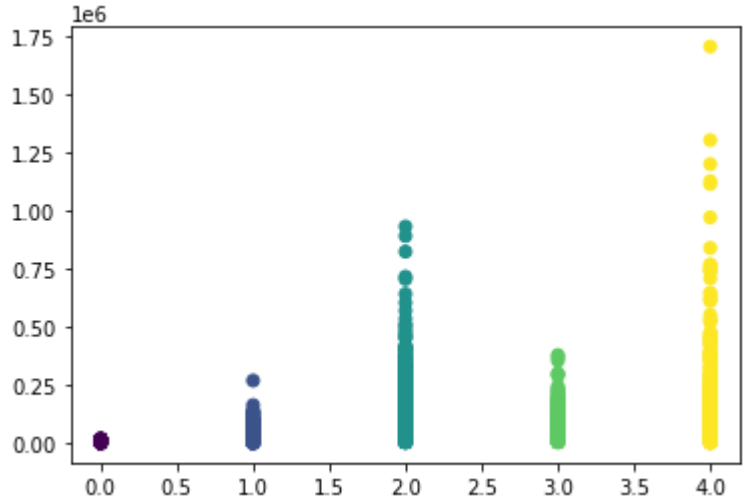
Out[18]:

| | Balance | Qual_miles | cc1_miles | cc2_miles | cc3_miles | Bonus_miles | Bonus_trans | Flight |
|---|---|---|---|---|---|---|---|---|
| **0** | 28143 | 0 | 1 | 1 | 1 | 174 | 1 | |
| **8** | 443003 | 0 | 3 | 2 | 1 | 1753 | 43 | |
| **21** | 185681 | 2024 | 1 | 1 | 1 | 13300 | 16 | |
| **23** | 66275 | 0 | 1 | 1 | 1 | 2533 | 11 | |
| **24** | 205651 | 500 | 1 | 1 | 1 | 4025 | 21 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | |
| **3982** | 11463 | 0 | 1 | 1 | 1 | 339 | 4 | |
| **3983** | 26173 | 0 | 1 | 1 | 1 | 305 | 1 | |
| **3987** | 11933 | 0 | 1 | 1 | 1 | 249 | 3 | |
| **3995** | 64385 | 0 | 1 | 1 | 1 | 981 | 5 | |
| **3997** | 54899 | 0 | 1 | 1 | 1 | 500 | 1 | |

1191 rows × 12 columns

In [19]:

```
#Plot clusters
plt.scatter(airline_data['clusters'],airline_data['Balance'],c=y_hc)
plt.show()
```




## K-Means

In [30]:

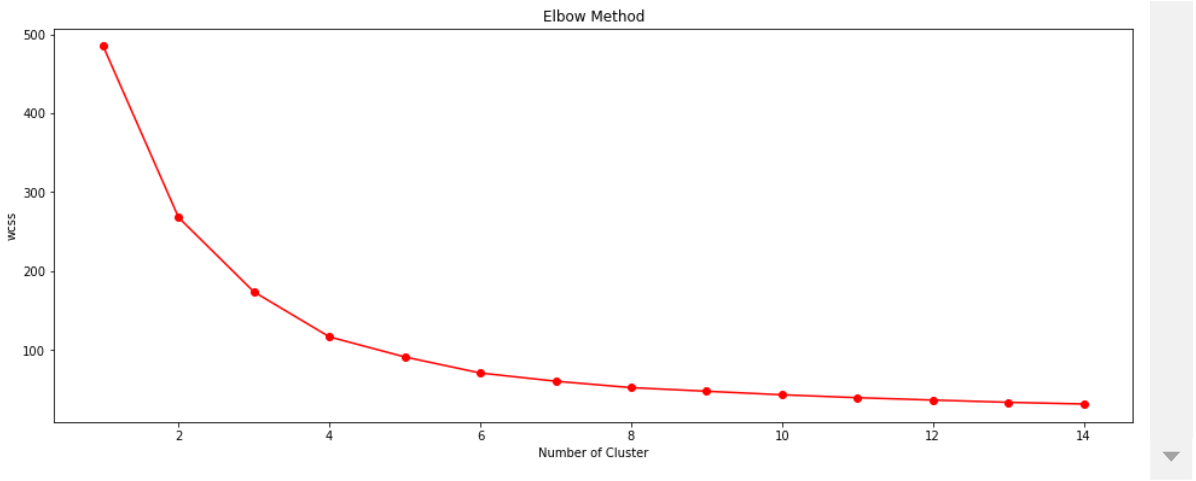
```
from sklearn.cluster import KMeans
```

In [33]:

```
from sklearn.cluster import KMeans
wcss = []
for i in range(1,15):
    kmeans=KMeans(n_clusters=i)
    kmeans.fit(airline_norm)
    wcss.append(kmeans.inertia_)
    print(i,wcss)


plt.figure(figsize=(16,6))
plt.plot(range(1,15),wcss,'ro-')
plt.title('Elbow Method')
plt.xlabel('Number of Cluster')
plt.ylabel('wcss')
plt.show()
```

```
1 [484.85112913072544]
2 [484.85112913072544, 267.597410959919]
3 [484.85112913072544, 267.597410959919, 173.27025625511422]
4 [484.85112913072544, 267.597410959919, 173.27025625511422, 116.32222839064
875]
5 [484.85112913072544, 267.597410959919, 173.27025625511422, 116.32222839064
875, 90.82757155770837]
6 [484.85112913072544, 267.597410959919, 173.27025625511422, 116.32222839064
875, 90.82757155770837, 70.47296409901432]
7 [484.85112913072544, 267.597410959919, 173.27025625511422, 116.32222839064
875, 90.82757155770837, 70.47296409901432, 60.073011804310084]
8 [484.85112913072544, 267.597410959919, 173.27025625511422, 116.32222839064
875, 90.82757155770837, 70.47296409901432, 60.073011804310084, 51.9349244619
6252]
9 [484.85112913072544, 267.597410959919, 173.27025625511422, 116.32222839064
875, 90.82757155770837, 70.47296409901432, 60.073011804310084, 51.9349244619
6252, 47.36862214468709]
10 [484.85112913072544, 267.597410959919, 173.27025625511422, 116.3222283906
4875, 90.82757155770837, 70.47296409901432, 60.073011804310084, 51.934924461
96252, 47.36862214468709, 42.869331732378996]
11 [484.85112913072544, 267.597410959919, 173.27025625511422, 116.3222283906
4875, 90.82757155770837, 70.47296409901432, 60.073011804310084, 51.934924461
96252, 47.36862214468709, 42.869331732378996, 39.17955530793661]
12 [484.85112913072544, 267.597410959919, 173.27025625511422, 116.3222283906
4875, 90.82757155770837, 70.47296409901432, 60.073011804310084, 51.934924461
96252, 47.36862214468709, 42.869331732378996, 39.17955530793661, 36.18858654
810717]
13 [484.85112913072544, 267.597410959919, 173.27025625511422, 116.3222283906
4875, 90.82757155770837, 70.47296409901432, 60.073011804310084, 51.934924461
96252, 47.36862214468709, 42.869331732378996, 39.17955530793661, 36.18858654
810717, 33.33543020380802]
14 [484.85112913072544, 267.597410959919, 173.27025625511422, 116.3222283906
4875, 90.82757155770837, 70.47296409901432, 60.073011804310084, 51.934924461
96252, 47.36862214468709, 42.869331732378996, 39.17955530793661, 36.18858654
810717, 33.33543020380802, 31.167831262978414]
```




In [34]:

```
k_model=KMeans(n_clusters=3)
y_knn=k_model.fit_predict(airline_norm)
```

In [35]:

```
clusters=pd.DataFrame(y_knn,columns=['Clusters'])
clusters
```

Out[35]:

| | Clusters |
|---|---|
| **0** | 0 |
| **1** | 0 |
| **2** | 0 |
| **3** | 0 |
| **4** | 0 |
| **...** | ... |
| **3994** | 0 |
| **3995** | 0 |
| **3996** | 0 |
| **3997** | 0 |
| **3998** | 0 |

3999 rows × 1 columns

In [36]:

```
airline_data
```

Out[36]:

| | Balance | Qual_miles | cc1_miles | cc2_miles | cc3_miles | Bonus_miles | Bonus_trans | Flight_mi |
|---|---|---|---|---|---|---|---|---|
| **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 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | |
| **994** | 18476 | 0 | 1 | 1 | 1 | 8525 | 4 | |
| **995** | 64385 | 0 | 1 | 1 | 1 | 981 | 5 | |
| **996** | 73597 | 0 | 3 | 1 | 1 | 25447 | 8 | |
| **997** | 54899 | 0 | 1 | 1 | 1 | 500 | 1 | |
| **998** | 3016 | 0 | 1 | 1 | 1 | 0 | 0 | |

99 rows × 12 columns

In [37]:

```
airline_data[airline_data['clusters']==1]
```

Out[37]:

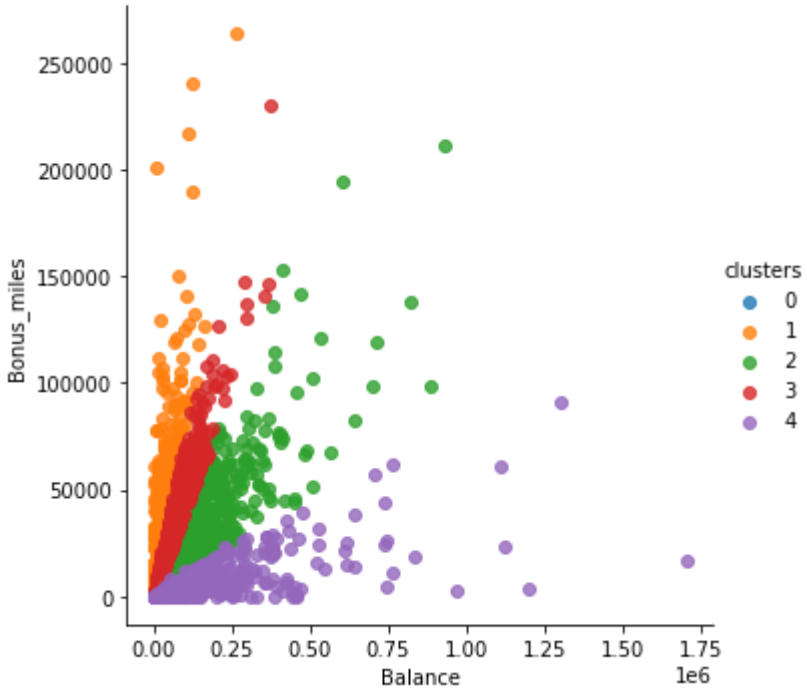
| | Balance | Qual_miles | cc1_miles | cc2_miles | cc3_miles | Bonus_miles | Bonus_trans | Flight |
|---|---|---|---|---|---|---|---|---|
| **15** | 28495 | 0 | 4 | 1 | 1 | 49442 | 15 | |
| **16** | 51890 | 0 | 4 | 1 | 1 | 48963 | 16 | |
| **41** | 10470 | 0 | 4 | 1 | 1 | 38094 | 26 | |
| **58** | 38077 | 0 | 3 | 1 | 1 | 34024 | 8 | |
| **78** | 49238 | 0 | 4 | 1 | 1 | 38037 | 18 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | |
| **3919** | 5000 | 0 | 1 | 1 | 1 | 5000 | 1 | |
| **3924** | 14775 | 0 | 1 | 1 | 1 | 14275 | 9 | |
| **3930** | 40424 | 0 | 4 | 1 | 1 | 44110 | 26 | |
| **3944** | 2124 | 0 | 1 | 1 | 1 | 2324 | 2 | |
| **3978** | 10071 | 0 | 2 | 1 | 1 | 27701 | 16 | |

453 rows × 12 columns

In [40]:

```
sns.lmplot('Balance','Bonus_miles',data=airline_data,hue='clusters', fit_reg=False)
import warnings
warnings.filterwarnings('ignore')
```

```
C:\Users\Asus\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureW
arning: Pass the following variables as keyword args: x, y. From version 0.1
2, the only valid positional argument will be `data`, and passing other argu
ments without an explicit keyword will result in an error or misinterpretati
on.
  warnings.warn(
```




## DBSCAN

In [41]:

```
airline_data.head()
```

Out[41]:

| | Balance | Qual_miles | cc1_miles | cc2_miles | cc3_miles | Bonus_miles | Bonus_trans | Flight_mi |
|---|---|---|---|---|---|---|---|---|
| **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 | |

In [43]:

```
x = airline_data.iloc[:,[2,3]].values
```

In [44]:

```
x.shape
```

Out[44]:

```
(3999, 2)
```

In [49]:

```
from sklearn.cluster import DBSCAN
min_samples=5
```

In [51]:

```
dbs=DBSCAN(min_samples=5,eps=0.2)
clusters=dbs.fit_predict(airline_data.iloc[:,[2,3]])
```

In [52]:

```
clusters
```

Out[52]:

```
array([0, 0, 0, ..., 2, 0, 0], dtype=int64)
```

In [53]:

```
airline_data[airline_data['clusters']==0]
```

Out[53]:

| | Balance | Qual_miles | cc1_miles | cc2_miles | cc3_miles | Bonus_miles | Bonus_trans | Flight_n |
|---|---|---|---|---|---|---|---|---|
| **27** | 8828 | 0 | 1 | 1 | 1 | 0 | 0 | |
| **31** | 10021 | 0 | 1 | 1 | 1 | 0 | 0 | |
| **39** | 2176 | 0 | 1 | 1 | 1 | 0 | 0 | |
| **51** | 1300 | 0 | 1 | 1 | 1 | 370 | 1 | |
| **55** | 14448 | 0 | 1 | 1 | 1 | 1625 | 6 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | |
| **3861** | 3126 | 0 | 1 | 1 | 1 | 100 | 1 | |
| **3876** | 1000 | 0 | 1 | 1 | 1 | 0 | 0 | |
| **3942** | 2131 | 0 | 1 | 1 | 1 | 405 | 3 | |
| **3981** | 1010 | 0 | 1 | 1 | 1 | 0 | 0 | |
| **3984** | 404 | 0 | 1 | 1 | 1 | 550 | 3 | |

129 rows × 12 columns

In [54]:

```
airline_data[airline_data['clusters']==2]
```

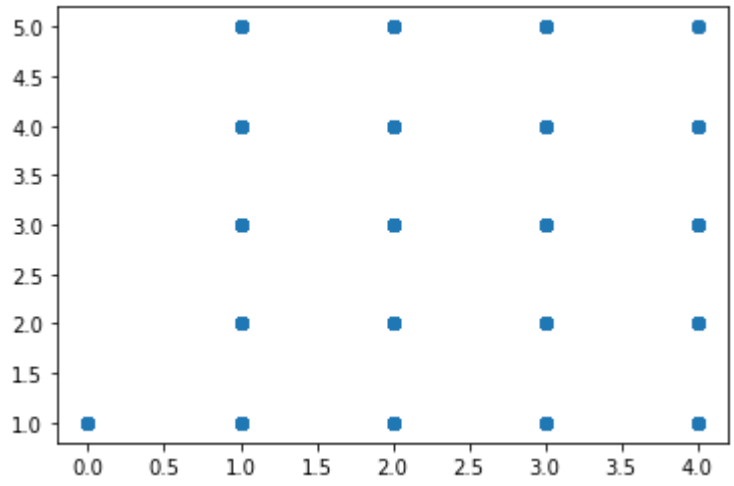
Out[54]:

| | Balance | Qual_miles | cc1_miles | cc2_miles | cc3_miles | Bonus_miles | Bonus_trans | Flight |
|---|---|---|---|---|---|---|---|---|
| **1** | 19244 | 0 | 1 | 1 | 1 | 215 | 2 | |
| **2** | 41354 | 0 | 1 | 1 | 1 | 4123 | 4 | |
| **3** | 14776 | 0 | 1 | 1 | 1 | 500 | 1 | |
| **5** | 16420 | 0 | 1 | 1 | 1 | 0 | 0 | |
| **6** | 84914 | 0 | 3 | 1 | 1 | 27482 | 25 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | |
| **3989** | 2622 | 0 | 1 | 1 | 1 | 1625 | 6 | |
| **3992** | 11181 | 0 | 1 | 1 | 1 | 929 | 12 | |
| **3993** | 3974 | 0 | 1 | 1 | 1 | 365 | 3 | |
| **3996** | 73597 | 0 | 3 | 1 | 1 | 25447 | 8 | |
| **3998** | 3016 | 0 | 1 | 1 | 1 | 0 | 0 | |

1547 rows × 12 columns

In [57]:

```
plt.scatter(x=airline_data['clusters'],y=airline_data['cc1_miles'])
plt.show()
```




In [ ]: