clustering

November 21, 2024

[1]:

```
import pandas as pd
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
data = pd.read_excel('/content/EastWestAirlines.xlsx',sheet_name='data')
data
```

[1]:

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

      Bonus_trans  Flight_miles_12mo  Flight_trans_12  Days_since_enroll  \
0               1                  0                0               7000
1               2                  0                0               6968
2               4                  0                0               7034
3               1                  0                0               6952
4              26               2077                4               6935
…           …                …              …                …
3994            4                200                1               1403
3995            5                  0                0               1395
3996            8                  0                0               1402
3997            1                500                1               1401
3998            0                  0                0               1398

      Award?
0          0
1          0
2          0
```

```
3          0
4          1
…      …
3994       1
3995       1
3996       1
3997       0
3998       0

[3999 rows x 12 columns]
```

[2]:

```
data.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
```

[3]:

```
data.isna().sum()
```

[3]:

```
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
```

[4]:
```
data.describe()
```

[4]:
```
               ID#       Balance    Qual_miles    cc1_miles    cc2_miles  \
count  3999.000000  3.999000e+03   3999.000000  3999.000000  3999.000000   
mean   2014.819455  7.360133e+04    144.114529     2.059515     1.014504   
std    1160.764358  1.007757e+05    773.663804     1.376919     0.147650   
min       1.000000  0.000000e+00      0.000000     1.000000     1.000000   
25%    1010.500000  1.852750e+04      0.000000     1.000000     1.000000   
50%    2016.000000  4.309700e+04      0.000000     1.000000     1.000000   
75%    3020.500000  9.240400e+04      0.000000     3.000000     1.000000   
max    4021.000000  1.704838e+06  11148.000000     5.000000     3.000000   

         cc3_miles    Bonus_miles  Bonus_trans  Flight_miles_12mo  \
count  3999.000000    3999.000000   3999.00000        3999.000000   
mean      1.012253   17144.846212     11.60190         460.055764   
std       0.195241   24150.967826      9.60381        1400.209171   
min       1.000000       0.000000      0.00000           0.000000   
25%       1.000000    1250.000000      3.00000           0.000000   
50%       1.000000    7171.000000     12.00000           0.000000   
75%       1.000000   23800.500000     17.00000         311.000000   
max       5.000000  263685.000000     86.00000       30817.000000   

       Flight_trans_12  Days_since_enroll       Award?  
count      3999.000000         3999.00000  3999.000000  
mean          1.373593         4118.55939     0.370343  
std           3.793172         2065.13454     0.482957  
min           0.000000            2.00000     0.000000  
25%           0.000000         2330.00000     0.000000  
50%           0.000000         4096.00000     0.000000  
75%           1.000000         5790.50000     1.000000  
max          53.000000         8296.00000     1.000000  
```

[5]:
```
data['ID#'].nunique()
```

[5]:
```
3999
```

[6]:
```
data.drop(columns='ID#', inplace=True)
```

[7]:
```
data.head()
```

[7]:
```
   Balance  Qual_miles  cc1_miles  cc2_miles  cc3_miles  Bonus_miles  \
0    28143           0          1          1          1          174   
1    19244           0          1          1          1          215   
2    41354           0          1          1          1         4123   
3    14776           0          1          1          1          500   
```

```
4    97752          0          4          1          1        43300

   Bonus_trans  Flight_miles_12mo  Flight_trans_12  Days_since_enroll  Award?
0            1                  0                0               7000       0
1            2                  0                0               6968       0
2            4                  0                0               7034       0
3            1                  0                0               6952       0
4           26               2077                4               6935       1
```

[8]:

```
from sklearn.preprocessing import StandardScaler,MinMaxScaler
min_max_scaler = MinMaxScaler()
norm = min_max_scaler.fit_transform(data)
norm
```

[8]:

```
array([[0.01650773, 0.        , 0.        , ..., 0.        , 0.84374246,
        0.        ],
       [0.01128788, 0.        , 0.        , ..., 0.        , 0.83988425,
        0.        ],
       [0.02425685, 0.        , 0.        , ..., 0.        , 0.84784181,
        0.        ],
       ...,
       [0.0431695 , 0.        , 0.5       , ..., 0.        , 0.16879672,
        1.        ],
       [0.03220189, 0.        , 0.        , ..., 0.01886792, 0.16867615,
        0.        ],
       [0.00176908, 0.        , 0.        , ..., 0.        , 0.16831444,
        0.        ]])
```

[9]:

```
len(norm)
```

[9]:

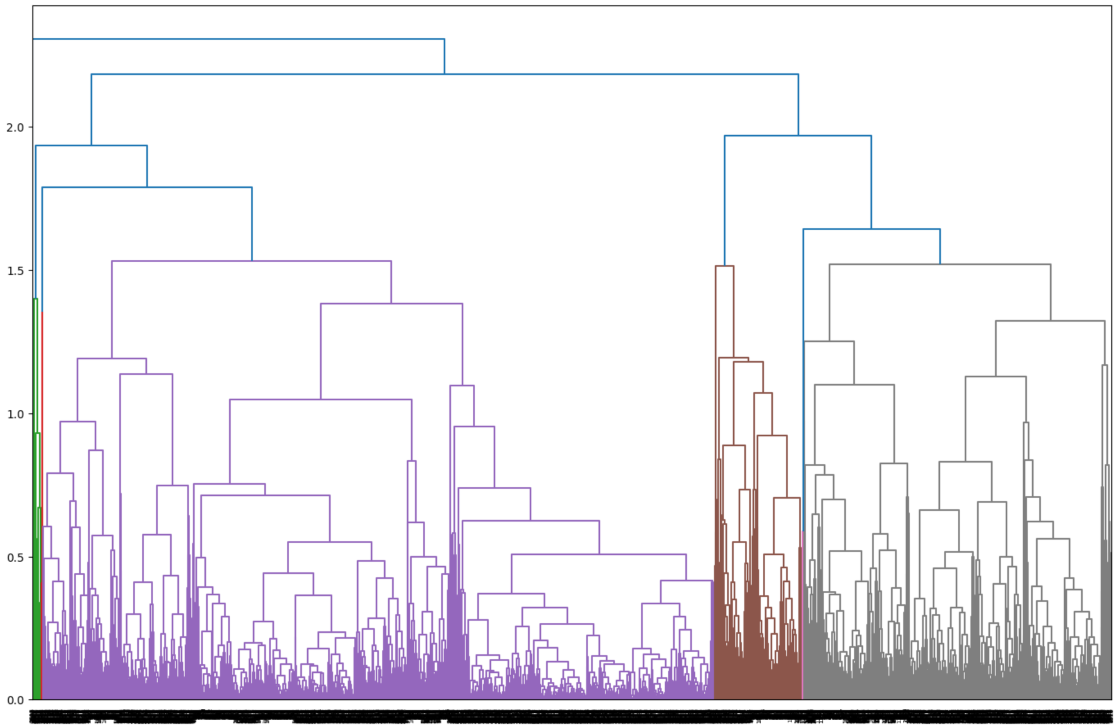
```
3999
```

[10]:

```
import scipy.cluster.hierarchy as sch
from sklearn.cluster import AgglomerativeClustering
```

[11]:

```
plt.figure(figsize = (17,11))
dendrogram=sch.dendrogram(sch.linkage(norm, method='complete'))
```



[15]:
```
hc = AgglomerativeClustering(n_clusters=5, linkage = 'complete')
# Remove affinity parameter as it's not needed when linkage is 'complete'
# For linkage='ward', affinity is fixed to 'euclidean' in older versions
# If you need other distance metrics, consider upgrading scikit-learn
```

[16]:
```
y_hc = hc.fit_predict(norm)
y_hc
```

[16]:
```
array([0, 0, 0, …, 2, 0, 0])
```

[17]:
```
hc.labels_
```

[17]:
```
array([0, 0, 0, …, 2, 0, 0])
```

[18]:
```
data['h_clusterid'] = hc.labels_
```

[19]:
```
data
```

[19]:
```
      Balance  Qual_miles  cc1_miles  cc2_miles  cc3_miles  Bonus_miles  \
0       28143           0          1          1          1          174
1       19244           0          1          1          1          215
2       41354           0          1          1          1         4123
3       14776           0          1          1          1          500
4       97752           0          4          1          1        43300
```

```
…       …         …        …        …        …          …
3994    18476           0          1          1          1         8525
3995    64385           0          1          1          1          981
3996    73597           0          3          1          1        25447
3997    54899           0          1          1          1          500
3998     3016           0          1          1          1            0

      Bonus_trans  Flight_miles_12mo  Flight_trans_12  Days_since_enroll  \
0                1                  0                0               7000
1                2                  0                0               6968
2                4                  0                0               7034
3                1                  0                0               6952
4               26               2077                4               6935
…           …                …              …                 …
3994             4                200                1               1403
3995             5                  0                0               1395
3996             8                  0                0               1402
3997             1                500                1               1401
3998             0                  0                0               1398

      Award?  h_clusterid
0          0            0
1          0            0
2          0            0
3          0            0
4          1            1
…      …          …
3994       1            2
3995       1            2
3996       1            2
3997       0            0
3998       0            0

[3999 rows x 12 columns]
```

[20]:
```
data.h_clusterid.value_counts()
```

[20]:
```
h_clusterid
0    2495
2    1144
1     325
4      31
3       4
Name: count, dtype: int64
```

[21]:
```
data.sort_values("h_clusterid")
```

[21]:

```
      Balance  Qual_miles  cc1_miles  cc2_miles  cc3_miles  Bonus_miles  \
0       28143           0          1          1          1          174
2485    23649           0          1          1          1         3250
2487   169794           0          3          1          1        22824
2488    23781           0          1          1          1           50
2491     5970           0          1          1          1         1000
…         …         …        …        …        …          …
940     44824           0          1          3          1        21107
3959    13942           0          1          2          1         5822
3779    35850           0          1          3          1        17759
1389    49145           0          1          2          1        12755
3191    56624           0          1          2          1        12311

      Bonus_trans  Flight_miles_12mo  Flight_trans_12  Days_since_enroll  \
0               1                  0                0               7000
2485           16                  0                0               3176
2487           19                767                2               3119
2488            1                 50                1               3085
2491            1                  0                0               3091
…             …                …              …                …
940            19               2000                4               5916
3959           12                  0                0               1458
3779           18                  0                0               3439
1389           22               2450                7               5323
3191           14                  0                0               2491

      Award?  h_clusterid
0          0            0
2485       0            0
2487       0            0
2488       0            0
2491       0            0
…        …          …
940        1            4
3959       0            4
3779       0            4
1389       0            4
3191       0            4

[3999 rows x 12 columns]
```

[22]:

```
data[data.h_clusterid==3]
```

[22]:

```
      Balance  Qual_miles  cc1_miles  cc2_miles  cc3_miles  Bonus_miles  \
2015    53232         888          4          1          1        80696
3235   287033           0          1          1          1        26161
3583   160114         500          1          1          1        71954
```

```
3594    27619           0          4          1          1        83726

      Bonus_trans  Flight_miles_12mo  Flight_trans_12  Days_since_enroll  \
2015           65              22100               45               3831
3235           58              12873               53               2272
3583           86              30817               53               1373
3594           68              14050               46               1325

      Award?  h_clusterid
2015       1            3
3235       1            3
3583       1            3
3594       1            3
```

[23]:

```
data.groupby('h_clusterid').mean()
```

[23]:

```
                   Balance  Qual_miles  cc1_miles  cc2_miles  cc3_miles  \
h_clusterid
0             59968.433667   88.883768   1.712224   1.000401   1.011222
1            157084.578462  208.673846   4.661538   1.000000   1.061538
2             80173.963287  248.550699   2.104895   1.009615   1.000874
3            131999.500000  347.000000   2.500000   1.000000   1.000000
4             45515.064516   32.258065   1.000000   2.483871   1.000000

              Bonus_miles  Bonus_trans  Flight_miles_12mo  Flight_trans_12  \
h_clusterid
0            10271.530261     9.105812         227.797194         0.656112
1            70477.086154    23.249231        1019.433846         3.196923
2            16882.864510    13.412587         739.958916         2.236888
3            65634.250000    69.250000       19960.000000        49.250000
4            14618.870968    16.129032         443.225806         1.967742

             Days_since_enroll    Award?
h_clusterid
0                  3825.392786  0.000802
1                  5770.572308  1.000000
2                  4304.383741  1.000000
3                  2200.250000  1.000000
4                  3784.258065  0.193548
```

[24]:

```
data2=data.drop(columns='h_clusterid')
data2.head()
```

[24]:

```
   Balance  Qual_miles  cc1_miles  cc2_miles  cc3_miles  Bonus_miles  \
0    28143           0          1          1          1          174
1    19244           0          1          1          1          215
2    41354           0          1          1          1         4123
```

```
3    14776           0          1          1          1         500
4    97752           0          4          1          1       43300

   Bonus_trans  Flight_miles_12mo  Flight_trans_12  Days_since_enroll  Award?
0            1                  0                0               7000       0
1            2                  0                0               6968       0
2            4                  0                0               7034       0
3            1                  0                0               6952       0
4           26               2077                4               6935       1
```

[25]:

```python
from sklearn.cluster import KMeans
scaler = StandardScaler().fit_transform(data2)
scaler
```

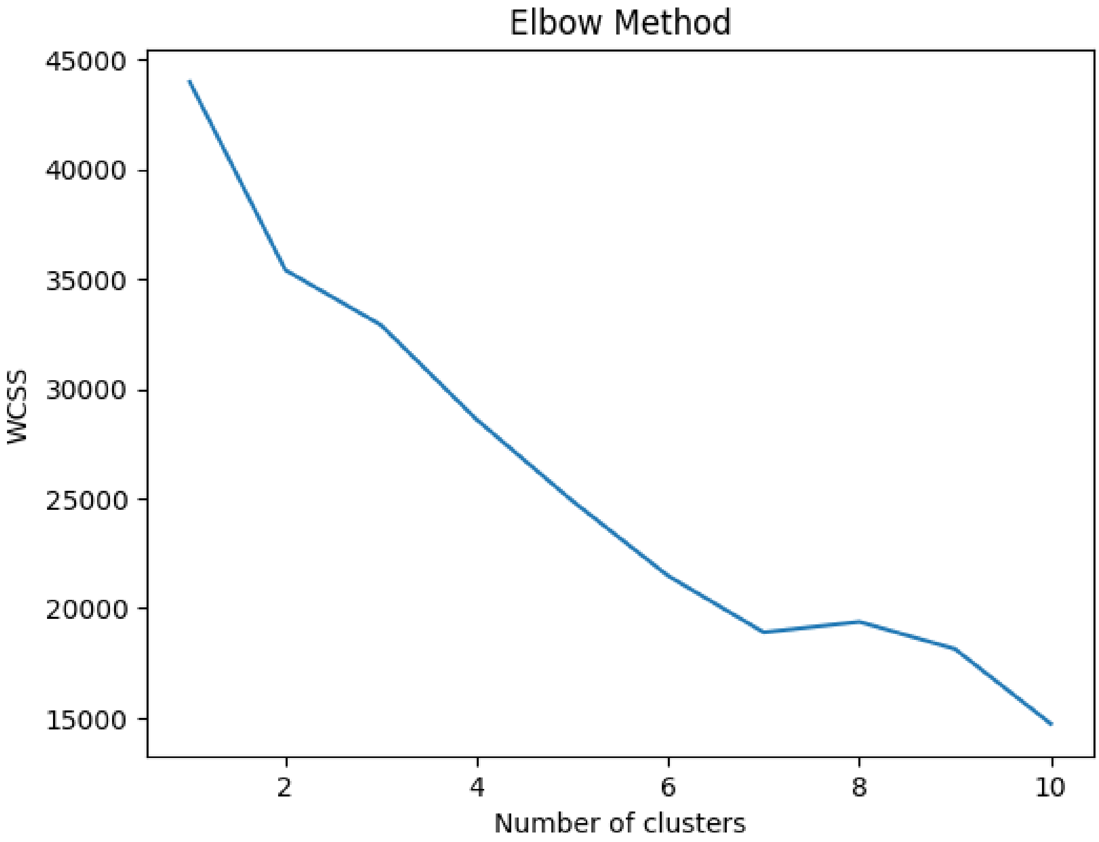
[25]:

```
array([[-4.51140783e-01, -1.86298687e-01, -7.69578406e-01, …,
        -3.62167870e-01,  1.39545434e+00, -7.66919299e-01],
       [-5.39456874e-01, -1.86298687e-01, -7.69578406e-01, …,
        -3.62167870e-01,  1.37995704e+00, -7.66919299e-01],
       [-3.20031232e-01, -1.86298687e-01, -7.69578406e-01, …,
        -3.62167870e-01,  1.41192021e+00, -7.66919299e-01],
       …,
       [-4.29480975e-05, -1.86298687e-01,  6.83121167e-01, …,
        -3.62167870e-01, -1.31560393e+00,  1.30391816e+00],
       [-1.85606976e-01, -1.86298687e-01, -7.69578406e-01, …,
        -9.85033311e-02, -1.31608822e+00, -7.66919299e-01],
       [-7.00507951e-01, -1.86298687e-01, -7.69578406e-01, …,
        -3.62167870e-01, -1.31754109e+00, -7.66919299e-01]])
```

[26]:

```python
wcss = []
for i in range(1, 11):
    kmeans = KMeans(n_clusters=i,random_state=0)
    kmeans.fit(scaler)
    wcss.append(kmeans.inertia_)

plt.plot(range(1, 11), wcss)
plt.title('Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```



[27]:
```
wcss
```

[27]:
```
[43989.0000000001,
 35409.516629861384,
 32901.45804218546,
 28589.20585898687,
 24884.115340435033,
 21470.51454980924,
 18891.750169018418,
 19371.700292851994,
 18134.257221538734,
 14727.658831288609]
```

[28]:
```
clusters = KMeans(6, random_state=42)
clusters.fit(scaler)
clusters.labels_
```

[28]:
```
array([2, 2, 2, …, 0, 3, 3], dtype=int32)
```

[29]:
```
clusters.fit_predict(scaler)
```

```
[29]: array([2, 2, 2, …, 0, 3, 3], dtype=int32)
```

```
[30]: clusters.inertia_
```

```
[30]: 25599.269402878585
```

```
[31]: from sklearn.metrics import silhouette_score
      silhouette_score(data2, clusters.labels_)
```

```
[31]: -0.03481925489833332
```

```
[32]: c2=KMeans(2, random_state=41)
      c2.fit(scaler)
```

```
[32]: KMeans(n_clusters=2, random_state=41)
```

```
[33]: silhouette_score(data2, c2.labels_)
```

```
[33]: 0.3446749880634653
```

```
[34]: data2['clusterid']=clusters.labels_
      data2
```

```
[34]:       Balance  Qual_miles  cc1_miles  cc2_miles  cc3_miles  Bonus_miles  \
      0       28143           0          1          1          1          174
      1       19244           0          1          1          1          215
      2       41354           0          1          1          1         4123
      3       14776           0          1          1          1          500
      4       97752           0          4          1          1        43300
      …         …         …        …        …        …          …
      3994    18476           0          1          1          1         8525
      3995    64385           0          1          1          1          981
      3996    73597           0          3          1          1        25447
      3997    54899           0          1          1          1          500
      3998     3016           0          1          1          1            0

            Bonus_trans  Flight_miles_12mo  Flight_trans_12  Days_since_enroll  \
      0               1                  0                0               7000
      1               2                  0                0               6968
      2               4                  0                0               7034
      3               1                  0                0               6952
      4              26               2077                4               6935
      …           …                …                …                 …
      3994            4                200                1               1403
      3995            5                  0                0               1395
      3996            8                  0                0               1402
      3997            1                500                1               1401
```

```
3998            0                   0                0               1398

      Award?  clusterid
0          0          2
1          0          2
2          0          2
3          0          2
4          1          1
…      …        …
3994       1          0
3995       1          0
3996       1          0
3997       0          3
3998       0          3

[3999 rows x 12 columns]
```

```
[35]: data2['clusterid'].value_counts()
```

```
[35]: clusterid
      3    1185
      2     980
      0     819
      1     800
      4     128
      5      87
      Name: count, dtype: int64
```

```
[36]: data2.groupby('clusterid').mean()
```

```
[36]:                  Balance  Qual_miles  cc1_miles  cc2_miles  cc3_miles  \
      clusterid
      0           56127.758242  194.483516   1.675214   1.019536   1.000000
      1          107510.410000  136.801250   4.180000   1.001250   1.056250
      2           59727.241837   71.454082   1.481633   1.020408   1.001020
      3           34522.096203   93.091139   1.245570   1.011814   1.000000
      4          132067.828125  678.835938   2.132812   1.046875   1.000000
      5          528835.448276  463.931034   3.666667   1.011494   1.034483

                  Bonus_miles  Bonus_trans  Flight_miles_12mo  Flight_trans_12  \
      clusterid
      0          10511.936508    10.586081         417.383394         1.249084
      1          48701.263750    19.910000         340.520000         1.022500
      2           6263.462245     8.532653         175.791837         0.517347
      3           4402.372996     6.589873         140.030380         0.422785
      4          30273.203125    28.906250        5995.664062        17.382812
      5          66229.632184    22.149425        1377.620690         4.816092
```

```
           Days_since_enroll   Award?
clusterid
0                4253.031746  1.00000
1                4817.865000  0.61125
2                5647.725510  0.00000
3                2103.353586  0.00000
4                4427.500000  0.78125
5                6191.137931  0.83908
```

```
[37]: from sklearn.cluster import DBSCAN
      dbscan = DBSCAN(eps=1, min_samples=7)
      dbscan.fit(scaler)
      dbscan.labels_
```

```
[37]: array([0, 0, 0, …, 1, 0, 0])
```

```
[38]: data3=data.drop(columns='h_clusterid')
      data3.head()
```

```
[38]:    Balance  Qual_miles  cc1_miles  cc2_miles  cc3_miles  Bonus_miles  \
      0    28143           0          1          1          1          174
      1    19244           0          1          1          1          215
      2    41354           0          1          1          1         4123
      3    14776           0          1          1          1          500
      4    97752           0          4          1          1        43300

         Bonus_trans  Flight_miles_12mo  Flight_trans_12  Days_since_enroll  Award?
      0            1                  0                0               7000       0
      1            2                  0                0               6968       0
      2            4                  0                0               7034       0
      3            1                  0                0               6952       0
      4           26               2077                4               6935       1
```

```
[39]: data3['clusterId']=dbscan.labels_
      data3
```

```
[39]:       Balance  Qual_miles  cc1_miles  cc2_miles  cc3_miles  Bonus_miles  \
      0       28143           0          1          1          1          174
      1       19244           0          1          1          1          215
      2       41354           0          1          1          1         4123
      3       14776           0          1          1          1          500
      4       97752           0          4          1          1        43300
      …       …          …          …          …          …          …
      3994    18476           0          1          1          1         8525
      3995    64385           0          1          1          1          981
      3996    73597           0          3          1          1        25447
```

```
3997    54899           0          1          1          1          500
3998     3016           0          1          1          1            0

      Bonus_trans  Flight_miles_12mo  Flight_trans_12  Days_since_enroll  \
0               1                  0                0               7000
1               2                  0                0               6968
2               4                  0                0               7034
3               1                  0                0               6952
4              26               2077                4               6935
…           …                …              …                …
3994            4                200                1               1403
3995            5                  0                0               1395
3996            8                  0                0               1402
3997            1                500                1               1401
3998            0                  0                0               1398

      Award?  clusterId
0          0          0
1          0          0
2          0          0
3          0          0
4          1          1
…      …        …
3994       1          1
3995       1          1
3996       1          1
3997       0          0
3998       0          0

[3999 rows x 12 columns]
```

[40]:

```
data3['clusterId'].value_counts()
```

[40]:

```
clusterId
 0    2299
 1    1072
-1     604
 2      11
 4       8
 3       5
Name: count, dtype: int64
```

[41]:

```
data3.groupby('clusterId').mean()
```

[41]:

```
                 Balance   Qual_miles  cc1_miles  cc2_miles  cc3_miles  \
clusterId
-1         177887.369205   896.723510   2.690397   1.077815   1.081126
```

```
 0          52733.796433     3.679426   1.656807   1.000000   1.000000
 1          60530.977612     6.554104   2.591418   1.000000   1.000000
 2          28365.363636     0.000000   1.000000   2.000000   1.000000
 3          51030.000000     0.000000   1.000000   1.000000   1.000000
 4          24545.375000  2401.000000   1.000000   1.000000   1.000000

            Bonus_miles  Bonus_trans  Flight_miles_12mo  Flight_trans_12  \
clusterId
-1         38686.490066    21.649007        2103.266556         6.089404
 0          9001.937364     8.428882         113.070465         0.373206
 1         22723.841418    12.793843         268.355410         0.840485
 2          8825.272727    11.818182          22.727273         0.181818
 3          4737.600000    11.400000        4242.800000        10.400000
 4          2427.750000     5.000000          37.500000         0.250000

           Days_since_enroll    Award?
clusterId
-1               4741.235099  0.677152
 0               3791.968247  0.000000
 1               4503.971082  1.000000
 2               2702.000000  0.000000
 3               2871.600000  0.000000
 4               2042.625000  0.000000
```

[ ]: