

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

In [4]: df=pd.read_excel("C:/Users/HP/Downloads/EastWestAirlines.xlsx", sheet_name='data')

In [5]: df1=df.copy()

In [6]: df1.head()

Out[6]:
```

	ID#	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo	Flight_trans_12	Days_since_enroll	Award?
0	1	28143	0	1	1	1	174	1	0	0	7000	0
1	2	19244	0	1	1	1	215	2	0	0	6968	0
2	3	41354	0	1	1	1	4123	4	0	0	7034	0
3	4	14776	0	1	1	1	500	1	0	0	6952	0
4	5	97752	0	4	1	1	43300	26	2077	4	6935	1

```
In [8]: df1_norm = preprocessing.scale(df1)

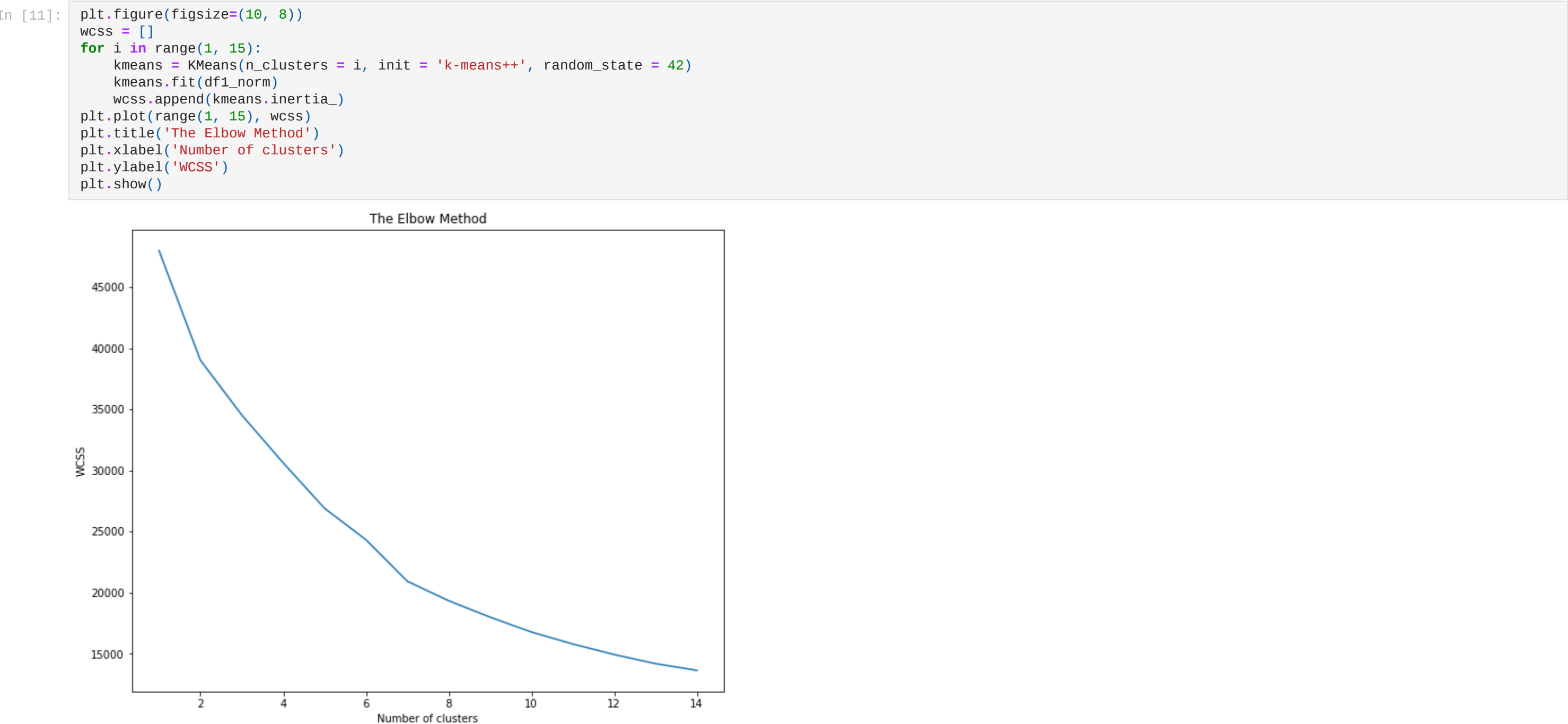
In [9]: #standardise the data to normal distribution
df1_norm = pd.DataFrame(df1_norm)

In [10]: df1_norm.head()

Out[10]:
```

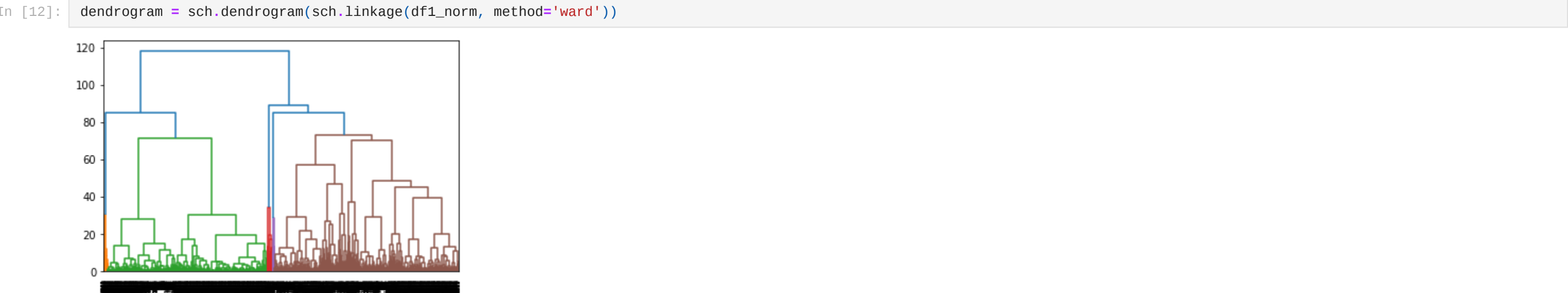
	0	1	2	3	4	5	6	7	8	9	10	11
0	-1.735125	-0.451141	-0.186299	-0.769578	-0.098242	-0.062767	-0.702786	-1.104065	-0.328603	-0.362168	1.395454	-0.766919
1	-1.734263	-0.539457	-0.186299	-0.769578	-0.098242	-0.062767	-0.701088	-0.999926	-0.328603	-0.362168	1.379957	-0.766919
2	-1.733402	-0.320031	-0.186299	-0.769578	-0.098242	-0.062767	-0.539253	-0.791649	-0.328603	-0.362168	1.411920	-0.766919
3	-1.732540	-0.583799	-0.186299	-0.769578	-0.098242	-0.062767	-0.689286	-1.104065	-0.328603	-0.362168	1.372208	-0.766919
4	-1.731679	0.239678	-0.186299	1.409471	-0.098242	-0.062767	1.083121	1.499394	1.154932	0.692490	1.363975	1.303918

Finding out the optimal numbr of clusters



As seen from the elbow graph, the slope changes at 2. However, since splitting the dataset into 2 groups would not be very beneficial, we further evaluate clusters for higher values of k.

H clustering

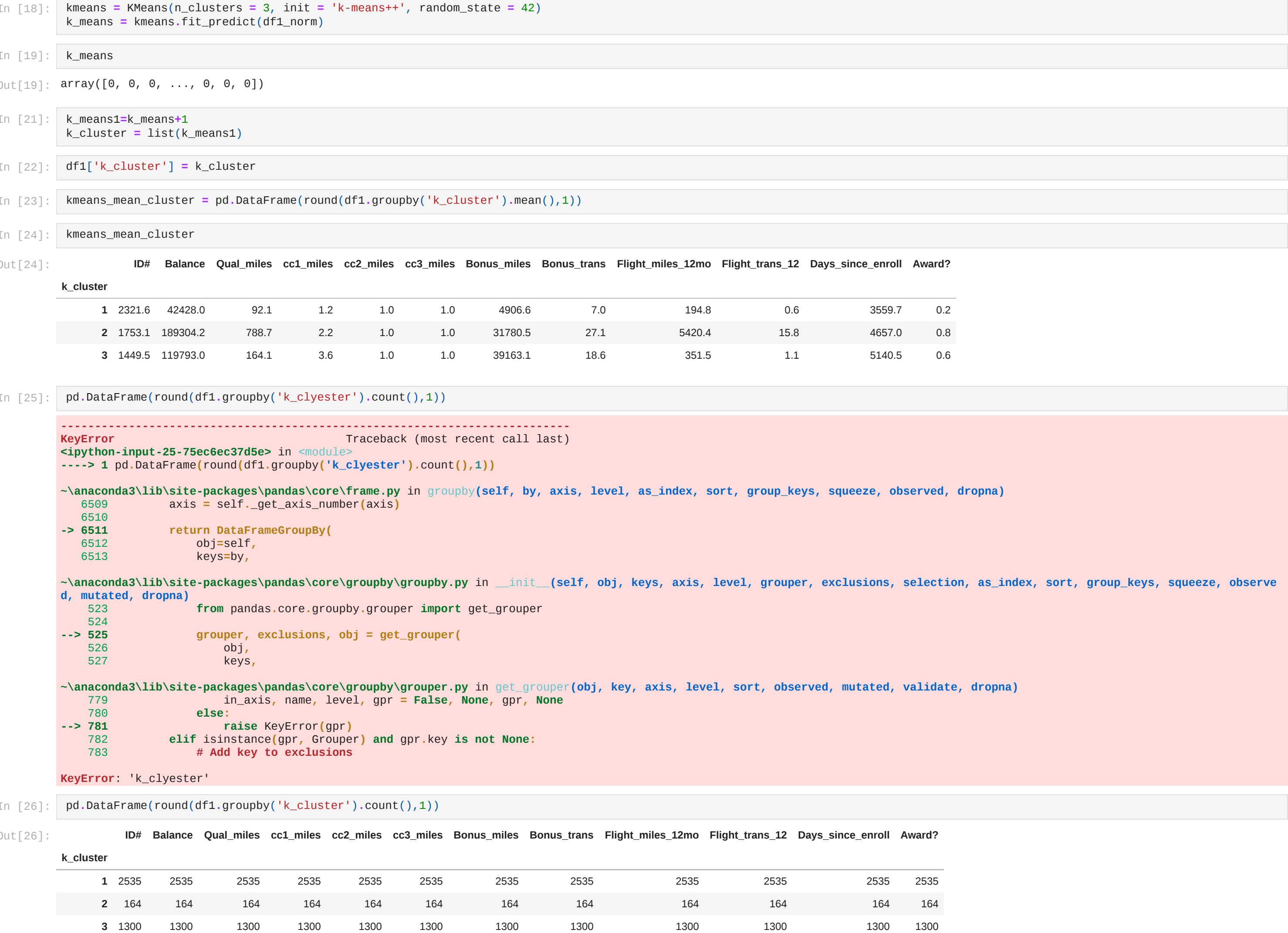


From the Ward method, we see that as the height increases the clusters get grouped together

We decided to cut the tree at height 85 to obtain 3 clusters and then assigned each cluster with its respective observations



5 - K means



Conclusion

From the above data generated from K-Means clustering, we can see Cluster-1 has around 63% total travelers and cluster 2 has 33% of the travelers. We will target cluster 1 & 2. Cluster 1 contains less frequent or first time travellers, by giving them discount provided they travel more than twice or thrice and introduce more offer if they register or take the membership.